INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 1

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser—Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 1

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 1 *xlv*

VOLUME 1

Abortion, Demography of: United States 1
Rachel K Jones

Absolutism, History of 6
Cesare Cuttica

Absorptive Capacity (of Organizations) 12
Heiko Gebauer and Hagen Worch

Academic Achievement Motivation, Development of 20
Jacquelynne S Eccles and Allan Wigfield

Academic Careers in Comparative Perspective 26
Hans Pechar and Lesley Andres

Academic Engagement 31
Jennifer A Fredricks

Academic Motivation and Performance: Task Value Interventions 37
Yoi Tibbetts, Elizabeth A Canning, and Judith M Harackiewicz

Academic Performance, Effects of Socio-Economic Status on 43
Brandon L Carlisle and Carolyn B Murray

Academic Research and Employment: Recent Changes in Europe and the United States 49
Elke Park and Hans Pechar

Academic Self-Concept and Achievement 54
Kit-Tai Hau and Herbert W Marsh

Access: Geographical 64
Emily Talen

Acculturation 68
David L Sam

Action, Theories of Social 75
Frank Kalter

Actor-Network Theory 80
Fabian Muniesa

Adaptation, Fitness, and Evolution 85
Ward B Watt

Addams, Jane (1860–1935) 91
Louise W Knight

Addictions: General Considerations 97
Ulrich John

Administration in Organizations 103
Jean-Claude Thoenig

Administrative Law: The United States and Beyond 109
Cary Coglianese

Adolescence, Sociology of 115
Monica K Johnson

Adolescent Health and Health Behaviors 121
Thomas A Wills and Rebecca Knight

Adolescent Sexual Risk 128
Christopher Browning and Jenny C Malave

Adoption, Demography of 136
Jessaca B Leinaweaver

Adoption: Domestic, International and Global Perspectives 142
Janette Logan

Adult Education and Lifelong Learning: The US Experience and Beyond 148
Jennifer A Margrett and Kyu ho Lee

Adult Mortality in Developing Countries 153
Katherine T Lofgren and Haidong Wang

Adult Mortality in Eastern Europe and the Former Soviet Republics 161
France Meslé

Adult Mortality in Industrialized Societies 170
Marc Luy

Adulthood: Dependency and Autonomy 181
Hans-Werner Wahl

Adverbial Clauses 185
Jingxia Lin

Advertising Agencies 189
Brian Moeran

Advertising and Advertisements 194
Linda Scott

Advertising, Control of 201
Jean Boddewyn

Advertising: Effects 208
Thomas C O'Guinn

Advertising: General 213
John Sinclair

Advocacy and Equity Planning 218
Norman Krumholz

Advocacy in Anthropology 223
Leslie E Sponsel

Aesthetic Education 229
Tracie E Costantino

Affect and Emotion, Anthropology of 236
Kathleen Stewart and Elizabeth Lewis

Affect-Regulation Motivation 241
Michaela Riediger

Affirmative Action: Comparative Policies and Controversies 248
Christopher McCrudden

Affirmative Action: Empirical Work on Its Effectiveness 256
Jonathan S Leonard

Affirmative Action, Sociology of 262
Anthony S Chen

Affordable Private Schools in Developing Countries 269
Pauline Dixon

African Legal Systems 272
Gordon R Woodman

African Studies: Culture 276
Jane I Guyer

African Studies: Health 280
John C Caldwell

African Studies: History 286
Frederick Cooper

Africa's Environment: A Stressed Biogeographical and Cultural Landscape 292
Delali Benjamin K Dovie

Africa, Sociocultural Overviews: West Africa 300
Paul Richards

Afrocentric Approaches to Social Work 305
Patricia Reid-Merritt

Age and Crime 310
Deinya Phenix

Age: Anthropological Aspects 318
Paul Spencer

Age Policy 323
Anne-Marie Guillemard

Age, Sociology of 327
Linda K George

Age Stratification 333
Dale Dannefer and Tirth Bhatta

Age Structure 338
Shiro Horiuchi

Agency Theory 344
Stefan Linder and Nicolai J Foss

Agenda Setting, Media Effects on 351
Maxwell McCombs

Agenda Settting, Public Policy in 357
Christoffer Green-Pedersen

Agendas: Political 362
Frank R Baumgartner

Agent Based Modeling, Statistics of 367
David Banks and Jacob Norton

Aggression, Social Psychology of 373
Wayne A Warburton and Craig A Anderson

AGIL, History of 381
Giuseppe Sciortino

Aging and Health in Old Age 394
Christine L Himes

Aging and Memory 398
Lars-Göran Nilsson

Aging and Older People, Geography of 407
Gavin J Andrews and Mark W Skinner

Aging and the Labor Market 413
Alan Barrett

Aging and Work 420
Keith L Zabel and Boris B Baltes

Aging Mind: Facets and Levels of Analysis 428
Shu-Chen Li

Aging, Theories of 435
K Warner Schaie

Agnosia 440
Daniel Tranel and Antonio R Damasio

Agonism 444
Kevin W Ryan

Agraphia (Acquired Dysgraphia) 450
Claudio Luzzatti

Agrarian Political Economy 456
Henry Bernstein

Agreeableness 463
Renée M Tobin and Daniel L Gadke

Agricultural Sciences and Technology 471
Lawrence Busch and Diana Stuart

Agriculture, Economics of 477
Bruce L Gardner

Agroecology and Agricultural Change 484
Charles A Francis and Alexander Wezel

AIDS: Acquired Immune-Deficiency Syndrome 488
Ting Ting Lee and Ian Paul Everall

Albert, Hans (1921–) 495
Eric Hilgendorf

Alcohol Interventions: Disease Models vs. Harm Reduction 501
Thomas Hall

Alcohol-Related Disorders 507
Michael Soyka

Alcohol Use among Young People 513
Rainer K Silbereisen and Karina Weichold

Alcohol Use and Abuse 517
Fiona Measham and Ian Paylor

Alexander, Jeffrey C. (1947–) 523
Jason L Mast

Algorithm 529
Laura Martignon

Algorithmic Complexity 534
Peter Sunehag and Marcus Hutter

Alienation, History of 539
Devorah Kalekin-Fishman and Lauren Langman

Alienation: Psychosociological Tradition 544
Catherine E Ross and John Mirowsky

Alienation, Sociology of 551
Christopher Adair-Toteff

Alliances: Political 556
Ole R Holsti

Allport, Gordon W (1897–1967) 562
Thomas F Pettigrew

Alternative Food Movements 566
Liz Grauerholz and Nicole Owens

Alternative Schools of Economic Thought 573
Peter J Boettke and Solomon M Stein

Altruism and Prosocial Behavior, Sociology of 579
Rafael Wittek and René Bekkers

Alzheimer's Disease: Behavioral and Social Aspects 584
Dieter Ferring

Alzheimer's Disease, Neural Basis of 591
Heiko Braak and Kelly Del Tredici-Braak

American and French Revolutions: Impact on the Social Sciences 597
John C Torpey

The American Conservative Movement 602
Ann M Horwitz

American Revolution, The 609
Gwenda Morgan

American Studies, Twentieth Century History of 615
Wil M Verhoeven

Americas, Sociocultural Overviews: Central America 621
Daniel Reichman

Americas, Sociocultural Overviews: Mexico 626
Paul Liffman

Americas, Sociocultural Overviews: North America, Native Americans 633
Russell Thornton

Americas, Sociocultural Overviews: Caribbean 638
Karen Fog Olwig

Amnesia: General 644
Kristoffer Romero and Morris Moscovitch

Amnesia: Psychogenic 651
Angelica Staniloiu and Hans J Markowitsch

Analysis of Variance and Generalized Linear Models 659
Ronald Christensen

Analytical Marxism 665
Nicholas Vrousalis and Philippe Van Parijs

Analytical Sociology 668
Peter Hedström and Petri Ylikoski

Anaphora 674
Yan Huang

Anarchism 680
Maple Razsa

Ancestors, Anthropology of 688
John Middleton

Ancient Egypt, Archaeology of 691
Marleen D Meyer and Stefanie Vereecken

Ancient Greece and Rome, History of 697
Wilfried Nippel

Animal Rights in Research and Research Application 703
Lesley J Rogers

Annales School 708
Eamon O'Flaherty

Anomie 714
Philipe Besnard

Anomie: History of the Concept 718
Mathieu Deflem

Anthropocene Epoch 722
Anson W Mackay

Anthropological Approaches to the Study of Education: The United States and Beyond 728
Bradley A Levinson, Norma González, and Kathryn Anderson-Levitt

Anthropological Research, Ethics of 734
Robert Albro

Anthropological Writing 740
Alma Gottlieb

Anthropology and History 746
James D Faubion

Anthropology at Home 751
Sophie Chevalier

Anthropology, Genomics, and Human Variation: National Roots 758
Gisli Palsson

Anthropology, History of 765
Thomas H Eriksen

Anthropology: Overview 772
Ulf Hannerz

Antidepressant Drugs 778
Carmine M Pariante, O Benkert, A Szegedi, and MJ Müller

Antisemitism 784
Robert Fine and Glynis Cousin

Antisocial Behavior 790
Stephen Crossley

Antitrust Policy: Lessons from the US 796
Daniel L Rubinfeld

Anxiety and Anxiety Disorders 804
Klaus Wiedemann

Anxiety and Fear Conditioning, Neural Basis of 811
Rheall F Roquet and Marie H Monfils

Aphasia 818
David Caplan

Appeals: Legal 822
Paul D Carrington and Marin K Levy

Appeasement: Political 826
Stephen R Rock

Applied Criminology and Criminal Justice 832
Brian Stout and Kerry Clamp

Applied Demography 839
David A Swanson

Applied Geography: A US Perspective 845
John W Frazier

Applied Social Research, History of 850
Sara M Strickhouser and James D Wright

Applied Sociology 858
Jammie Price and Jeff Will

Apprenticeship and School Learning: Lessons from Germany 861
Peter Noack and Bärbel Kracke

Apprenticeship, Anthropological Aspects 865
Roy Dilley

Apprenticeships and School Learning: General Considerations 870
Hans Gruber and Heinz Mandl

Apraxia 874
Georg Goldenberg

Archaeology and Cultural Memory 881
Cornelius Holtorf

Archaeology, Politics of 885
Lynne Goldstein

Archaeology: Philosophy and Science 891
Timothy Webmoor

Archaeology, Theory in 899
Staša Babić

Archaeometry 905
Alan Watchman

Archival Methods 909
John Markoff

Archiving: Ethical Aspects 916
Josefina J Card and Tamara J Kuhn

Arctic Archaeology and Prehistory 921
Owen K Mason

Arctic: Sociocultural Aspects 927
Peter P Schweitzer

Area and International Studies: Cultural Studies 933
Benjamin Lee

Area and International Studies: Political Economy 938
Ian Robinson

Area and International Studies: Stakeholders 943
Gilbert W Merkx

Area Studies, History of 949
Matthias Duller

Areal Linguistics 955
Lyle Campbell

Arendt, Hannah (1906–75) 961
Antonia Grunenberg

Aristocracy, Nobility, and Gentry, European History of 965
Antoni Mączak

Armed Conflict and Social Work: Significance and Implications 970
Shulamit Ramon

Arms Control 976
Coit D Blacker

Aron, Raymond (1905–83) 979
Panagiotis Christias

This page intentionally left blank

CONTRIBUTORS TO VOLUME 1

Christopher Adair-Toteff
University of South Florida, Tampa, FL, USA; and University of Kent, Canterbury, UK

Robert Albro
Center for Latin American and Latino Studies, American University, Washington, DC, USA

Kathryn Anderson-Levitt
Graduate School of Education and Information Sciences, University of California, Los Angeles, CA, USA

Craig A. Anderson
Iowa State University, Ames, IA, USA

Lesley Andres
University of British Columbia, Vancouver, BC, Canada

Gavin J. Andrews
McMaster University, Hamilton, ON, Canada

Staša Babić
Faculty of Philosophy, Belgrade, Serbia

Boris B. Baltes
Wayne State University, Detroit, MI, USA

David Banks
Duke University, Durham, NC, USA

Alan Barrett
Economic and Social Research Institute, Dublin, Ireland

Frank R. Baumgartner
The University of North Carolina, Chapel Hill, NC, USA

René Bekkers
Center for Philanthropic Studies, VU University Amsterdam, Amsterdam, The Netherlands

O. Benkert
University of Mainz, Mainz, Germany

Henry Bernstein
University of London, London, UK; and China Agricultural University, Beijing, China

Philipe Besnard[†]
Observatoire Sociologique du Changement, Paris, France

Tirth Bhatta
Case Western Reserve University, OH, USA

Coit D. Blacker
Stanford University, Stanford, CA, USA

Jean Boddewyn
Baruch College, City University of New York, New York, NY, USA

Peter J. Boettke
George Mason University, Fairfax, VA, USA

Heiko Braak
University of Ulm, Ulm, Germany

Christopher Browning
The Ohio State University, Columbus, OH, USA

Lawrence Busch
Michigan State University, East Lansing, MI, USA

John C. Caldwell
Australian National University, Canberra, ACT, Australia

Lyle Campbell
University of Hawai'i Mānoa, Honolulu, HI, USA

Elizabeth A. Canning
University of Wisconsin—Madison, Madison, WI, USA

David Caplan
Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Josefina J. Card
Sociometrics Corporation, Los Altos Hills, CA, USA

Brandon L. Carlisle
University of California Riverside, Riverside, CA, USA

[†] Deceased.

Paul D. Carrington
Duke University School of Law, Durham, NC, USA

Anthony S. Chen
Northwestern University, Evanston, IL, USA

Sophie Chevalier
Université de Franche-Comté, Besançon, France

Ronald Christensen
University of New Mexico, Albuquerque, NM, USA

Panagiotis Christias
University of Cyprus, Nicosia, Cyprus

Kerry Clamp
School of Social Sciences and Psychology, University of Western Sydney, Penrith, NSW, Australia

Cary Coglianese
University of Pennsylvania Law School, Philadelphia, PA, USA

Frederick Cooper
New York University, New York, NY, USA

Tracie E. Costantino
Rhode Island School of Design, Providence, RI, USA

Glynis Cousin
University of Wolverhampton, Wolverhampton, UK

Stephen Crossley
Durham University, Durham, UK

Cesare Cuttica
Université Paris 8-Vincennes, Saint-Denis, France

Antonio R. Damasio
University of Iowa, Iowa City, IA, USA

Dale Dannefer
Case Western Reserve University, OH, USA

Mathieu Deflem
University of South Carolina, Columbia, SC, USA

Kelly Del Tredici-Braak
University of Ulm, Ulm, Germany

Roy Dilley
University of St Andrews, St Andrews, Scotland, UK

Pauline Dixon
Newcastle University, Newcastle Upon Tyne, UK

Delali Benjamin K. Dovie
Department of Geography and Resource Development, University of Ghana, Legon, Ghana; School of Animal, Plant and Environmental Sciences, University of the Witwatersrand, Johannesburg, South Africa; and Regional Institute for Population Studies, University of Ghana, Legon, Ghana

Matthias Duller
University of Graz, Graz, Austria

Jacquelynne S. Eccles
University of Michigan, Ann Arbor, MI, USA

Thomas H. Eriksen
University of Oslo, Oslo, Norway

Ian Paul Everall
University of Melbourne, Melbourne, VIC, Australia; and Royal Melbourne Hospital, Parkville, VIC, Australia

James D. Faubion
Rice University, Houston, TX, USA

Dieter Ferring
Research Unit INSIDE, University of Luxembourg, Walferdange, Luxembourg

Robert Fine
University of Warwick, Coventry, UK

Nicolai J. Foss
Copenhagen Business School, Frederiksberg, Denmark; and Norwegian School of Economics and Business Administration, Bergen, Norway

Charles A. Francis
University of Nebraska — Lincoln, Lincoln, NE, USA

John W. Frazier
SUNY, Binghamton, NY, USA

Jennifer A. Fredricks
Connecticut College, New London, CT, USA

Daniel L. Gadke
Mississippi State University, Mississippi State, MS, USA

Bruce L. Gardner
University of Maryland, College Park, MD, USA

Heiko Gebauer
Eawag — Swiss Federal Institute of Aquatic Science and Technology, ESS — Environmental Social Sciences, Dübendorf, Switzerland

Linda K. George
Duke University, Durham, NC, USA

Georg Goldenberg
Bogenhausen Hospital, Munich, Germany

Lynne Goldstein
Michigan State University, East Lansing, MI, USA

Norma González
School of Education, University of Arizona, Tucson, AZ, USA

Alma Gottlieb
University of Illinois at Urbana-Champaign, Urbana, IL, USA

Liz Grauerholz
University of Central Florida, Orlando, FL, USA

Christoffer Green-Pedersen
Aarhus University, Aarhus, Denmark

Hans Gruber
Institute of Educational Science, University of Regensburg, Regensburg, Germany

Antonia Grunenberg
Berlin, Germany

Anne-Marie Guillemard
University Paris Descartes Sorbonne, Paris, France

Jane I. Guyer
Johns Hopkins University, Baltimore, MD, USA

Thomas Hall
University of Central Florida, Orlando, FL, USA

Ulf Hannerz
Stockholm University, Stockholm, Sweden

Judith M. Harackiewicz
University of Wisconsin—Madison, Madison, WI, USA

Kit-Tai Hau
The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

Peter Hedström
Institute for Futures Studies, Stockholm, Sweden

Eric Hilgendorf
Faculty of Law, University of Würzburg, Würzburg, Germany

Christine L. Himes
Syracuse University, Syracuse, NY, USA

Ole R. Holsti
Duke University, Durham, NC, USA; and Salt Lake City, UT, USA

Cornelius Holtorf
School of Cultural Sciences, Linnaeus University, Kalmar, Sweden

Shiro Horiuchi
CUNY School of Public Health and CUNY Institute for Demographic Research, New York, NY, USA

Ann M. Horwitz
University of Maryland, College Park, MD, USA

Yan Huang
University of Auckland, Auckland, New Zealand

Marcus Hutter
Research School of Computer Science, Australian National University, Canberra, ACT, Australia

Ulrich John
University Medicine Greifswald, Greifswald, Germany

Monica K. Johnson
Washington State University, Pullman, WA, USA

Rachel K. Jones
Guttmacher Institute, New York, NY, USA

Devorah Kalekin-Fishman
University of Haifa, Haifa, Israel

Frank Kalter
University of Mannheim, Mannheim, Germany

Louise W. Knight
Northwestern University, Evanston, IL, USA

Rebecca Knight
University of Hawaii Cancer Center, Honolulu, HI, USA

Bärbel Kracke
University of Jena, Jena, Germany

Norman Krumholz
Cleveland State University, Cleveland, OH, USA

Tamara J. Kuhn
ETR, Scotts Valley, CA, USA

Lauren Langman
Loyola University of Chicago, Chicago, IL, USA

Benjamin Lee
Rice University, Houston, TX, USA

Kyu ho Lee
Iowa State University, Ames, IA, USA

Ting Ting Lee
University of Melbourne, Melbourne, VIC, Australia

Jessaca B. Leinaweaver
Brown University, Providence, RI, USA

Jonathan S. Leonard
University of California, Berkeley, CA, USA

Bradley A. Levinson
School of Education, Indiana University, Bloomington, IN, USA

Marin K. Levy
Duke University School of Law, Durham, NC, USA

Elizabeth Lewis
University of Texas at Austin, Austin, TX, USA

Shu-Chen Li
Chair of Lifespan Developmental Neuroscience, Technische Universität (TU) Dresden, Dresden, Germany

Paul Liffman
Centro de Estudios Antropológicos, El Colegio de Michoacán, Michoacán, Mexico

Jingxia Lin
Nanyang Technological University, Singapore

Stefan Linder
ESSEC Business School, Paris, France

Katherine T. Lofgren
Institute for Health Metrics, Seattle, WA, USA

Janette Logan
University of Manchester, Manchester, UK

Marc Luy
Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/OEAW, WU), Vienna Institute of Demography of the Austrian Academy of Sciences, Vienna, Austria

Claudio Luzzatti
University of Milano-Bicocca, Milano, Italy

Anson W. Mackay
University College London, London, UK

Antoni Maczak
University of Warsaw, Poland

Jenny C. Malave
The Ohio State University, Columbus, OH, USA

Heinz Mandl
Department Psychology, Ludwig-Maximilians-Universität, München, Germany

Jennifer A. Margrett
Iowa State University, Ames, IA, USA

John Markoff
University of Pittsburgh, Pittsburgh, PA, USA

Hans J. Markowitsch
University of Bielefeld, Bielefeld, Germany; and Hanse Institute for Advanced Study, Delmenhorst, Germany

Herbert W. Marsh
Institute for Positive Psychology and Education, Australian Catholic University, NSW, Australia; King Saud University, Riyadh, Saudi Arabia; and University of Oxford, Oxford, UK

Laura Martignon
PH Ludwigsburg University of Education, Ludwigsburg, Germany

Owen K. Mason
Institute of Arctic and Alpine Research, University of Colorado, Boulder, CO, USA

Jason L. Mast
Institute of Advanced Study, University of Warwick, Coventry, UK

Maxwell McCombs
University of Texas at Austin, Austin, TX, USA

Christopher McCrudden
Oxford, UK

Fiona Measham
School of Applied Social Sciences, Durham University, Durham, UK

Gilbert W. Merkx
Center for International Studies, Duke University, Durham, NC, USA

France Meslé
INED, Paris, France

Marleen D. Meyer
KU Leuven, Leuven, Belgium

John Middleton
Yale University, New Haven, CT, USA

John Mirowsky
University of Texas, Austin, TX, USA

Brian Moeran
Copenhagen Business School, Copenhagen, Denmark

Marie H. Monfils
University of Texas at Austin, Austin, TX, USA

Gwenda Morgan
University of Newcastle, Newcastle upon Tyne, UK

Morris Moscovitch
University of Toronto, Toronto, ON, Canada

M.J. Müller
University of Mainz, Mainz, Germany

Fabian Muniesa
Mines ParisTech, Paris, France

Carolyn B. Murray
University of California Riverside, Riverside, CA, USA

Lars-Göran Nilsson
Stockholm Brain Institute, Stockholm University, Stockholm, Sweden

Wilfried Nippel
Humboldt-Universität zu Berlin, Philosophische Fakultät I, Unter den Linden, Berlin, Germany

Peter Noack
University of Jena, Jena, Germany

Jacob Norton
North Carolina State University, Raleigh, NC, USA

Eamon O'Flaherty
University College Dublin, Dublin, Ireland

Thomas C. O'Guinn
University of Illinois, Urbana–Champaign, Champaign, IL, USA

Karen Fog Olwig
University of Copenhagen, Copenhagen K, Denmark

Nicole Owens
University of Central Florida, Orlando, FL, USA

Gisli Palsson
University of Iceland, Oddi, Reykjavik, Iceland

Carmine M. Pariante
Institute of Psychiatry, King's College London, London, UK

Elke Park
Institute of Science Communication and Higher Education Research, Alpen Adria Universität, Vienna, Austria

Ian Paylor
Lancaster University, Lancaster, UK

Hans Pechar
Institute of Science Communication and Higher Education Research, Alpen Adria Universität, Vienna, Austria

Thomas F. Pettigrew
University of California, Santa Cruz, CA, USA

Deinya Phenix
St Francis College, Brooklyn, NY, USA

Jammie Price
Journal of Applied Social Science, Wilmington, NC, USA

Shulamit Ramon
Anglia Ruskin University, Cambridge, UK

Maple Razsa
Colby College, Waterville, ME, USA

Daniel Reichman
University of Rochester, Rochester, NY, USA

Patricia Reid-Merritt
The Richard Stockton College of New Jersey, Galloway, NJ, USA

Paul Richards
Wageningen University and Research Center, Wageningen, The Netherlands

Michaela Riediger
Max Planck Institute for Human Development, Berlin, Germany

Ian Robinson
University of Michigan, Ann Arbor, MI, USA

Stephen R. Rock
Vassar College, Poughkeepsie, NY, USA

Lesley J. Rogers
School of Science and Technology, University of New England, Armidale, NSW, Australia

Kristoffer Romero
University of Toronto, Toronto, ON, Canada

Rheall F. Roquet
University of Texas at Austin, Austin, TX, USA

Catherine E. Ross
University of Texas, Austin, TX, USA

Daniel L. Rubinfeld
University of California, Berkeley, Berkeley, CA, USA; and New York University, New York, NY, USA

Kevin W. Ryan
National University of Ireland, Galway, Ireland

David L. Sam
University of Bergen, Bergen, Norway

K. Warner Schaie
University of Washington, Seattle, WA, USA

Peter P. Schweitzer
University of Vienna, Vienna, Austria

Giuseppe Sciortino
Università di Trento, Trento, Italy

Linda Scott
University of Illinois, Urbana, IL, USA

Rainer K. Silbereisen
Center for Applied Developmental Science, University of Jena, Jena, Germany

John Sinclair
The University of Melbourne, Melbourne, VIC, Australia

Mark W. Skinner
Trent University, Peterborough, ON, Canada

Michael Soyka
Ludwig Maximilian University, Munich, Germany; and Privatklinik Meiringen, Meiringen, Switzerland

Paul Spencer
University of London, London, UK

Leslie E. Sponsel
University of Hawai'i, Honolulu, HI, USA

Angelica Staniloiu
University of Bielefeld, Bielefeld, Germany

Solomon M. Stein
George Mason University, Fairfax, VA, USA

Kathleen Stewart
University of Texas at Austin, Austin, TX, USA

Brian Stout
School of Social Sciences and Psychology, University of Western Sydney, Penrith, NSW, Australia

Sara M. Strickhouser
University of Central Florida, Orlando, FL, USA

Diana Stuart
Michigan State University, East Lansing, MI, USA

Peter Sunehag
Research School of Computer Science, Australian National University, Canberra, ACT, Australia

David A. Swanson
University of California Riverside, Riverside, CA, USA

A. Szegedi
University of Mainz, Mainz, Germany

Emily Talen
Arizona State University, Tempe, AZ, USA

Jean-Claude Thoenig
Université Paris-Dauphine, Paris, France

Russell Thornton
University of California, Los Angeles, CA, USA

Yoi Tibbetts
University of Wisconsin—Madison, Madison, WI, USA

Renée M. Tobin
Illinois State University, Normal, IL, USA

John C. Torpey
Graduate Center, City University of New York, New York, NY, USA

Daniel Tranel
University of Iowa, Iowa City, IA, USA

Philippe Van Parijs
UC Louvain, Louvain-la-Neuve, Belgium

Stefanie Vereecken
KU Leuven, Leuven, Belgium

Wil M. Verhoeven
University of Groningen, Groningen, The Netherlands

Nicholas Vrousalis
University of Cambridge, Cambridge, UK; and KU Leuven, Leuven, Belgium

Hans-Werner Wahl
Institute of Psychology, Heidelberg University, Heidelberg, Germany

Haidong Wang
Institute for Health Metrics, Seattle, WA, USA

Wayne A. Warburton
Macquarie University, NSW, Australia

Alan Watchman
James Cook University, Townsville, QLD, Australia

Ward B. Watt
Stanford University, Stanford, CA, USA

Timothy Webmoor
University of Colorado at Boulder, Boulder, CO, USA

Karina Weichold
University of Jena, Jena, Germany

Alexander Wezel
ISARA-Lyon, Lyon, France

Klaus Wiedemann
University Hospital Hamburg Eppendorf, Hamburg, Germany

Allan Wigfield
University of Maryland, Baltimore, MD, USA

Jeff Will
University of North Florida, Jacksonville, FL, USA; and Journal of Applied Social Science, Wilmington, NC, USA

Thomas A. Wills
University of Hawaii Cancer Center, Honolulu, HI, USA

Rafael Wittek
University of Groningen, Groningen, The Netherlands

Gordon R. Woodman
University of Birmingham, Edgbaston, Birmingham, UK

Hagen Worch
Swiss Distance University of Applied Sciences/ Fernfachhochschule Schweiz, Institute for Management and Innovation, Regensdorf, Switzerland

James D. Wright
University of Central Florida, Orlando, FL, USA

Petri Ylikoski
University of Helsinki, Helsinki, Finland

Keith L. Zabel
Wayne State University, Detroit, MI, USA

This page intentionally left blank

A

Abortion, Demography of: United States

Rachel K Jones, Guttmacher Institute, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Abortion is one indicator of undesired fertility. In 2011, there were 1.06 million abortions in the United States, and the abortion rate was 16.9 per 1000 women aged 15–44; one in five pregnancies resulted in abortion. Groups overrepresented among women obtaining abortions and, in turn, having higher abortion rates include: women aged 18–24, black women, cohabiting women, and poor women. Given current abortion rates, it is estimated the 30% of U.S. women will have an abortion by age 45.

Introduction

About half of the 6.6 million pregnancies (including miscarriages) that occurred in the United States in 2008 were unintended (51%). A majority of these pregnancies resulted in unintended births, but 40% ended in abortion (Finer and Zolna, 2014). Thus, abortion is a measure of undesired fertility.

While abortion is one of the most politicized aspects of health care, information on this outcome is essential for several purposes. First, data on abortion are used in the measurement and calculation of the rates of unintended pregnancy (Finer and Zolna, 2014), teenage pregnancy (Kost and Henshaw, 2013), and contraceptive failure rates (Kost et al., 2008). Variations in abortion and abortion rates by characteristics such as age, marital status, and race and ethnicity can suggest groups that have the most difficulties in preventing unintended pregnancy and, in turn, may benefit from targeted public health interventions. This article summarizes the most recent social science and demographic research on abortion in the United States, including abortion incidence, characteristics of women who have abortions, and lifetime incidence of abortion.

Data Sources

The information summarized in this article comes from two national surveys, both conducted by the Guttmacher Institute: the Abortion Provider Census and the Abortion Patient Survey.

The number of abortions performed in the United States, as well as basic information on the availability of abortion services, is obtained from the Guttmacher Institute's Abortion Provider Census, which are periodic surveys of all identifiable U.S. abortion providers. The first of these surveys was conducted in 1974, obtaining data for 1973, the first year in which abortion was legal at the federal level. Since then, the Institute has conducted surveys at variable intervals, up to 5 years apart; estimates are made for years in which data were not collected by interpolation between adjacent years. The most recent Guttmacher provider survey, conducted in 2012 and 2013, collected information on number of abortions performed in 2010 and 2011. More detailed information about the data collection procedures is available elsewhere (Jones and Jerman, 2014).

The abortion incidence data collected by the Guttmacher Institute are considered to be the most comprehensive available (Pazol et al., 2013). While the Centers for Disease Control and Prevention (CDC) collates abortion incidence from state health departments and summarizes this information in an annual report – the most recent CDC Abortion Surveillance Report is for 2010 (Pazol et al., 2013) – the data are incomplete. This is mainly due to the fact that several state health departments, including California, do not collect abortion data; in 2011, California accounted for 17% of the abortions in the United States (Jones and Jerman, 2014).

Information on the characteristics of women who obtain abortion comes from the 2008 Abortion Patient Survey (APS). This survey is the Guttmacher Institute's fourth in a series and uses a design and questionnaire similar to those for the three earlier studies of abortion patients, conducted in 1987, 1994–95, and 2000–01 (Henshaw SK and Kost K 1996; Henshaw and Silverman, 1988; Jones et al., 2002). For the 2008 APS, a random sample of 107 abortion providing facilities – including hospitals, physicians' offices, and clinics – were asked to distribute a 4-page survey to every patient who obtained an abortion during the fielding period (which ranged from 2 to 12 weeks, depending on the facility caseload). The final data set contained information from 9493 women obtaining abortions at 95 facilities. Weights were constructed to correct for any bias produced by deviation from the original sampling plan and nonresponse. In turn, the data are considered to be nationally representative of all abortion patients. To estimate abortion rates by subgroup, information from the 2008 APS was combined with population data from the 2008

Current Population Survey and other national data sources. More detailed information about the study is available in previously published reports (Jones and Kavanaugh, 2011; Jones et al., 2010).

Abortion Incidence and Abortion Rates by Subgroup

In 2011, there were 1.06 million abortions, and the abortion rate was 16.9 per 1000 women aged 15–44 (Jones and Jerman, 2014); put differently, this latter figure means that 1.7% of reproductive aged women had an abortion in 2011. In that year, one in five pregnancies (including births and abortions, but not miscarriages) ended in abortion.

These most recent figures on abortion are more interesting when placed in historical context. The number of legal induced abortions performed annually in the United States increased dramatically in the decade following the legalization of abortion, with the highest number of procedures occurring in the late 1980s, reaching a peak of 1.6 million in 1990 (Table 1). After 1990, even as the U.S. population continued to grow, the number of abortions declined. The incidence of abortion stabilized between 2005 and 2008, but then resumed its long-term decline. The 1.06 million abortions performed in 2011 were the lowest number since 1976. The rate of abortions generally followed the same pattern as abortion numbers with a peak rate of 29.3 abortions per 1000 women aged 15–44 in 1980, to 16.9 abortions per 1000 women in 2011, the lowest since 1973.

While number of abortions and abortion rates are useful national indicators of unwanted fertility, information on the demographic characteristics of women who have abortions, including differences in abortion rates by subgroup, provide a more comprehensive picture.

Public discussions of abortion and the women who have them often focus on adolescents, which may create the impression that most abortion patients are teenagers. However, the majority of women who had abortions in 2008, 58%, were in their 20s; women in their 30s were the second most common age group (22%) (Table 2). Adolescents, or women under the age of 20, accounted for slightly less than one in five abortions (18%). This percentage includes minors, or patients under the age of 18, who accounted for 7% of abortions.

Patterns in abortion rates varied slightly (Table 2). Women in their early 20s had the highest abortion rate, 39.9 per 1000, and the second highest abortion rate was seen among 18–19-year-olds, 34.7 per 1000. Women over the age of 40 had the lowest rate, 3.2 per 1000.

Women's desires to have children, as well as their ability to negotiate the responsibilities of childrearing, may be influenced by relationships with male partners, and abortion varies substantially by union status. Just fewer than half of women having abortions were living with male partners: Married women accounted for 15% of all abortions, and an additional 29% were unmarried, but cohabiting with male partners in the month they became pregnant. Women who were not living with their partners accounted for 56% of all abortions, and most had never been married (45%).

Married women had the lowest abortion rate, 6.6 per 1000, and cohabiting women had the highest, 52.0 per 1000. Never-married women had rates slightly higher than all women, 23.9 per 1000.

Abortion patients are racially and ethnically diverse: non-Hispanic white women comprised 36% of patients, non-Hispanic black women 30%, Hispanic women 25%, and non-Hispanic women of other races for 9%. However, minorities were overrepresented and, in turn, had higher abortion rates. Black women had the highest rate, 40.2 per 1000, followed by Hispanic women, 28.7 per 1000.

In 2008, 16% of women obtaining abortions were foreign-born. This proportion was similar to their representation in the larger population of women and, in turn, the abortion rate

Table 1 Number of reported abortions, abortion rate, and abortion ratio, United States, 1973–2011

Year	*Abortions (in 000s)*	*Rate*[a]	*Ratio*[b]
1973	744.6	16.3	19.3
1974	898.6	19.3	22.0
1975	1034.2	21.7	24.9
1976	1179.3	24.2	26.5
1977	1316.7	26.4	28.6
1978	1409.6	27.7	29.2
1979	1497.7	28.8	29.6
1980	1553.9	29.3	30.0
1981	1577.3	29.3	30.1
1982	1573.9	28.8	30.0
1983	(1575.0)	(28.5)	(30.4)
1984	1577.2	28.1	29.7
1985	1588.6	28.0	29.7
1986	(1574.0)	(27.4)	(29.4)
1987	1559.1	26.9	28.8
1988	1590.8	27.3	28.6
1989	(1566.9)	(26.8)	(27.5)
1990	(1608.6)	(27.4)	(28.0)
1991	1556.5	26.3	27.4
1992	1528.9	25.7	27.5
1993	(1495.0)	(25.0)	(27.4)
1994	(1423.0)	(23.7)	(26.6)
1995	1359.4	22.5	25.9
1996	1360.2	22.4	25.9
1997	(1335.0)	(21.9)	(25.5)
1998	(1319.0)	(21.5)	(25.1)
1999	1314.8	21.4	24.6
2000	1313.0	21.3	24.5
2001	(1291.0)	(20.9)	(24.4)
2002	(1269.0)	(20.5)	(23.8)
2003	(1250.0)	(20.2)	(23.3)
2004	1222.1	19.7	22.9
2005	1206.2	19.4	22.4
2006	(1242.2)	(19.9)	(22.9)
2007	1209.6	19.4	21.9
2008	1212.4	19.4	22.5
2009	(1151.6)	(18.5)	(22.2)
2010	1102.7	17.7	21.7
2011	1058.5	16.9	21.2

Note: Figures in brackets are estimated by interpolation of numbers of abortions and adjustments made to state health department reports.
[a]Abortions per 1000 women aged 15–44 as of July 1 of each year.
[b]Abortions per 100 pregnancies ending in abortion or live birth; for each year, the ratio is based on births occurring during the 12-month period starting in July of that year.
Source: Jones, R. K., Jerman, J., 2014. Abortion incidence and service availability in the United States, 2011. Perspectives on Sexual and Reproductive Health.

Table 2 Number of U.S. abortions, population characteristics, and abortion rates of U.S. women aged 15–44 in 2008, by selected characteristics

	Abortions 2008			*All women 15–44 2008*		*Abortion rate*
Characteristic	*N*	*%*	*95% CI*	*N*	*%*	*2008*
Total	1 212 350			61 918 950		19.6
Age group						
<15	4850	0.4	0.3–0.6	Na		na
15–19	208 520	17.2	16.1–18.3	10 526 220	17.0	19.8
15–17	75 170	6.2	5.6–6.8	6 625 330	10.7	11.3
18–19	133 360	11.0	10.3–11.8	3 838 970	6.2	34.7
20–24	404 920	33.4	32.2–34.6	10 154 710	16.4	39.9
25–29	295 810	24.4	23.4–25.4	10 340 460	16.7	28.6
30–34	163 670	13.5	12.7–14.3	9 597 440	15.5	17.1
35–39	99 410	8.2	7.6–9.0	10 464 300	16.9	9.5
≥40[a]	35 160	2.9	2.5–3.4	10 835 820	17.5	3.2
Union status						
Married	179 430	14.8	13.5–16.2	26 996 660	43.6	6.6
Cohabiting, not married	354 010	29.2	27.6–30.8	6 811 080	11.0	52.0
Never-married, not cohabiting	545 560	45.0	43.0–47.1	22 848 090	36.9	23.9
Previously married, not cohabiting	133 360	11.0	9.9–12.1	5 201 190	8.4	25.6
Race and ethnicity						
Non-Hispanic white	437 660	36.1	31.5–40.9	38 080 150	61.5	11.5
Non-Hispanic black	358 860	29.6	24.6–35.1	8 916 330	14.4	40.2
Non-Hispanic other	113 960	9.4	7.4–11.8	4 396 250	7.1	25.9
Hispanic	301 880	24.9	19.8–31.0	10 526 220	17.0	28.7
Foreign-born						
No	1 013 520	83.6	80.8–86.1	51 330 810	82.9	19.7
Yes	198 830	16.4	13.9–19.2	10 588 140	17.1	18.8
Hispanic and foreign-born	95 776	7.9	6.1–10.2	5 263 110	8.5	18.2
Education[b]						
<High school	122 870	12.3	10.9–13.9	5 242 060	10.2	23.4
High school graduate/GED	282 710	28.3	26.7–30.0	13 310 720	25.9	21.2
Some college/associate degree	394 590	39.5	38.1–40.9	16 754 030	32.6	23.6
≥College graduate	198 800	19.9	18.3–21.5	16 085 920	31.3	12.4
Prior births						
0	474 030	39.1	37.0–41.2	27 182 420	43.9	17.4
1	321 270	26.5	25.2–27.8	10 835 820	17.5	29.6
≥2	418 260	34.5	32.8–36.1	23 900 710	38.6	17.5
Religious affiliation						
Protestant	453 420	37.4	33.5–41.4	29 597 260	47.8	15.3
Born-Again/Evangelical/Fundamentalist	180 640	14.9	12.8–17.2	15 789 332	25.6	11.4
Roman Catholic	341 880	28.2	24.9–31.7	15 355 900	24.8	22.3
Other	84 860	7.0	6.2–7.9	6 625 330	10.7	12.8
None	333 400	27.5	25.6–29.5	10 340 460	16.7	32.2
Family income as % of federal poverty level						
<100	514 040	42.4	39.8–45.1	9 845 110	15.9	52.2
100–199	321 270	26.5	25.4–27.7	11 516 920	18.6	27.9
≥200	377 040	31.1	28.7–33.6	40 494 990	65.4	9.3

[a]Denominator is women aged 40–44.
[b]Among women aged 20 and older.

for this subgroup was similar to that for all women. Foreign-born women had an abortion rate quite similar to US-born women, 18.8 versus 19.7 per 1000. Notably, 8.0% of all abortion patients were foreign-born Hispanic women, and the abortion rate for this group, 18.4 per 1000, was lower than the abortion rate for all Hispanic women.

Education can influence fertility intentions in several ways. For young women in particular, the desire to pursue or complete education can provide motivation to delay childbearing, and both attendance and completion of schooling can provide access to information and resources (e.g., sex education, health care) that better allow them to avoid unintended pregnancies. The overwhelming majority of abortion patients aged 20 and older, 88%, had graduated from high school. Most commonly, 40% had some college or an associate's degree, and one in five women had a bachelor's degree or higher.

One common reason women report terminating unintended pregnancies is that having a baby would prevent them from achieving goals such as pursuing an education. Patterns in abortion rates by educational attainment may be due, in part, to educational goals. For example, the higher abortion rate

among women with some college (24 per 1000) may reflect the fact that this population is currently in school and having a baby would prevent them from achieving their goal. The substantially lower rate among women with college degrees (12 per 1000) may reflect that higher education provides exposure to information about and access to contraception, and, perhaps, increased motivation to avoid unintended pregnancy in the form of job and career opportunities.

Abortion and motherhood are often regarded as opposing interests, and it is sometimes assumed that women who obtain abortions do not want to be mothers because they are unable or unwilling to assume the responsibilities of raising a child. But the majority of women obtaining abortions in 2008, 61%, already had children, including 34% who had two or more. Women with only one prior birth had an abortion rate almost twice as high as women with no children and those with more than one prior birth (29.6 versus 17.4 and 17.5 per 1000, respectively).

A substantial minority of women obtaining abortions in 2008, 42%, reported incomes that qualified them as poor, and an additional 27% were low-income, having family incomes between 100% and 199% of the federal poverty line. Poor women were overrepresented among abortion patients and had one of the highest abortion rates of all the subgroups examined, 52.2 abortions per 1000. As income levels increased, the abortion rate decreased, and women with family incomes at or above 200% of the federal poverty level had a rate slightly lower than half of the national rate at 9.3 per 1000.

Almost three-quarters of women obtaining abortions in 2008 reported a religious affiliation, most commonly Protestant (37%). Catholic was the second most common affiliation (28%) and a similar proportion of abortion patients, 27%, indicated they had no religious affiliation. One in five abortion patients identified themselves as born-again, evangelical, charismatic, or fundamentalist, 75% of whom were Protestant (not shown).

Protestants and women reporting 'other' as their religious affiliation had the lowest abortion rates of the affiliations examined, 12.7 per 1000. Born-again/fundamentalist Protestants had an abortion rate almost half as low as all women, 11.4 per 1000 (not shown). Women with no religious affiliation had a higher abortion rate than those with any affiliation, 32.2 per 1000. Catholics had an abortion rate slightly higher than the national rate.

Lifetime Incidence of Abortion

Data from the 2008 APS were used to determine the lifetime incidence of abortion, based on the proportion of women in each age group who were obtaining first abortions. The cumulative first-abortion rate, or proportion of women estimated to have had an abortion by the time they reach the end of a specified age range, was obtained by multiplying each age-specific first-abortion rate by the number of years in that age group (e.g., 15–17 age group had a multiplier of three), and summing all age groups up to that age group. The lifetime incidence of abortion is the equivalent of the cumulative first-abortion rate for women aged 40–44.

Table 3 Abortion rate, % first abortions, first-abortion rate and cumulative first-abortion rate, 2008

	2008				
		% Obtaining first abortion			
Age at outcome	Abortions per 1000	95% CI		First-abortion rate	Cumulative first-abortion rate
<15[a]	2.4	97.3	83.4–99.6	2.4	2.4
15–17	11.3	85.7	81.6–89.0	9.7	31.5
18–19	34.7	73.4	69.6–76.8	25.5	82.5
20–24	39.9	55.6	53.2–57.9	22.2	193.4
25–29	28.6	39.9	36.8–43.1	11.4	250.5
30–34	17.1	34.6	31.2–38.1	5.9	280.0
35–39	9.5	32.2	28.6–36.1	3.1	295.3
≥40[b]	3.2	34.5	28.5–41.2	1.1	300.9
Total	19.6	50.4	48.2–52.6	9.9	300.9

[a]Denominator is women aged 14.
[b]Denominator is women aged 40–44.

In 2008, there were 2.4 abortions for every 1000 adolescents under the age of 15 (Table 3). Because 97% of these were first-time procedures, the first-abortion rate and cumulative first-abortion rate for this group were also 2.4 per 1000. The overwhelming majority, 86%, of adolescents aged 15–17 were obtaining their first abortion. In turn, the first-abortion rate for this group was slightly lower than their overall abortion rate (10 vs 11 per 1000). By adding the first-abortion rate for women <15 to 3 times the first-abortion rate of women 15–17 (to account for the 3-year size of that age group), their cumulative first-abortion rate was 31.5 per 1000.

The cumulative first-abortion rate increases with age, and women aged 40 and older had a rate of 301 per 1000 women. Put differently, an estimated 30.1% of women aged 15–44 in 2008 will have an abortion by age 45 if exposed to prevailing abortion rates throughout their reproductive lives. Similarly, an estimated 8% of U.S. women would have had an abortion by age 20% and 25% by age 30.

A Note About (Potential) Change

Information on abortion patients collected in 2008 is the most recent available. The abortion rate declined somewhat substantially (13%) between 2008 and 2011 (Jones and Jerman, 2014) and may have increased, further decreased or stabilized since that time. However, it is unlikely that the general patterns in the characteristics of women who obtain abortions changed as dramatically since 2008. Research examining changes in characteristics of women who had abortions between 1974 and 2004 did find substantial demographic shifts – for example, the proportion of abortion patients under the age of 20 dropped by almost half (from 33% to 17%), and the proportion aged 30 and older increased – but, these changes occurred gradually over the three decades; moreover, these 'demographic transitions' occurred even while the abortion incidence was increasing and decreasing (Henshaw and Kost, 2008). Thus it is likely still the case that the majority of abortion patients are in their 20s, that most women having abortions have one or more children, are unmarried, and have some college education; similarly, while

the actual abortion rates for various subgroup may have changed, the patterns in rates are likely similar (e.g., women aged 20–24 probably still have the highest rate compared to women in other age groups).

The one characteristic that has the potential to change somewhat more rapidly is poverty. Between 2000 and 2008, the proportion of abortion patients living below the federal poverty level increased 59%, from 27% to 42% (Jones and Kavanaugh, 2011; Jones et al., 2010). Similarly, the abortion rate for this group increased 18% even while it decreased for the larger population (Jones and Kavanaugh, 2011). To some extent, this was due to the fact that 2008 was in the midst of the economic recession, and there were more poor women in the population (Jones et al., 2010). Still, poor and low-income women have been overrepresented among abortion patients for as long as it has been measured on the APS and this general pattern is likely to remain unchanged.

Similarly, the lifetime incidence of abortion is directly influenced by the abortion rate. Since this declined between 2008 and 2011, it is possible that the lifetime incidence has declined as well.

Conclusion

Abortion is an important measure of unwanted fertility. Some 40% of U.S. unintended pregnancies end in abortion as do 21% of all pregnancies (Finer and Zolna, 2014).

Abortion has largely been declining since the early 1990s, but research has yet to uncover the causal factors behind this trend. The decline is likely due to a variety of factors including increased access to highly effective contraceptive methods, inability of women to access abortion services (and a concomitant increase in unintended births during some parts of the country), and fewer women becoming pregnant (Jones and Jerman, 2014).

Disparities in rates of abortion among different groups of women often reflect vulnerabilities in other aspects of life, including poverty, insurance coverage, access to health care as well as other types of social and economic disadvantage. These conditions can pose structural, logistical, and personal barriers that make it difficult for disadvantaged groups of women to access and effectively use contraception and plan their pregnancies compared to more socioeconomically advantaged women.

Despite recent declines, it is worth noting that the abortion rate in the United States is generally higher than those of Western and Northern European countries. In 2008, abortion rates were substantially lower in Switzerland (7 abortions per 1000 women), Germany (7), the Netherlands (8), and Finland (11) relative to the United States (19) (Sedgh et al., 2011). These comparisons suggest that there is room for improvement to helping U.S. women avoid unintended pregnancies.

See also: Children, Value of; Fertility of Single and Cohabiting Women; Second Demographic Transition; Sex Selective Abortion.

Bibliography

Finer, L., Zolna, M., 2014. Shifts in intended and unintended pregnancies in the United States, 2001–2008. American Journal of Public Health 104 (S1), S43–S48.

Henshaw, S.K., Kost, K., 1996. Abortion patients in 1994–1995: characteristics and contraceptive use. Family Planning Perspectives 28 (4), 140–158.

Henshaw, S.K., Kost, K., 2008. Trends in the Characteristics of Women Obtaining Abortions, 1974 to 2004. Guttmacher Institute, New York.

Henshaw, S.K., Silverman, J., 1988. The characteristics and prior contraceptive use of U.S. abortion patients. Family Planning Perspectives 20 (4), 158–168.

Jones, R.K., Darroch, J.E., Henshaw, S.K., 2002. Contraceptive use among U.S. women having abortions in 2000–2001. Perspectives on Sexual and Reproductive Health 34 (6), 294–303.

Jones, R.K., Kavanaugh, M.L., 2011. Changes in abortion rates between 2000 and 2008 and lifetime incidence of abortion. Obstetrics & Gynecology 117 (6), 1358–1366.

Jones, R.K., Finer, L.B., Singh, S., 2010. Characteristics of U.S. Abortion Patients, 2008. Guttmacher Institute, New York.

Jones, R.K., Jerman, J., 2014. Abortion incidence and service availability in the United States, 2011. Perspectives on Sexual and Reproductive Health 46 (1), 3–14.

Jones, R.K., Susheela, Singh, Finer, L.B., Frohwirth, L.F., 2006. Repeat Abortion in the United States, vol. 29.

Kost, K., Singh, S., Vaughan, B., Trussell, J., Bankole, A., 2008. Estimates of contraceptive failure from the 2002 national survey of family growth. Contraception 77 (1), 10–21.

Kost, K., Henshaw, S.K., 2013. U.S. Teenage Pregnancies, Births and Abortions, 2008: State Trends by Age, Race and Ethnicity. Guttmacher Institute, New York.

Pazol, K., Creanga, A., Burley, K.D., Hayes, B., Jamieson, D.J., 2013. Abortion surveillance — United States, 2010. MMWR Surveillance Summaries 62 (8), 1–44.

Sedgh, G., Singh, S., Henshaw, S.K., Bankole, A., 2011. Legal abortion worldwide in 2008: levels and recent trends. International Perspectives on Sexual and Reproductive Health 37 (2), 84–94.

Absolutism, History of

Cesare Cuttica, Université Paris 8-Vincennes, Saint-Denis, France

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous article by R. Vierhaus, volume 1, pp. 5–9, © 2001, Elsevier Ltd.

Abstract

This article focuses on the highly contested but widely employed historiographical category 'absolutism.' It illustrates the origins of the term and provides an account of its main scholarly interpretations. Together with showing how divided academic opinions are with regard to the meaning(s) of 'absolutism,' the following pages ask whether it is the same thing to speak of absolutism with regard to the practices of early modern European monarchies or with reference to the political ideas of so-called absolutist theorists. In so doing, this article provides a comprehensive account of debates on absolutism and, at the same time, suggests that further work needs to be carried out on its theoretical aspects. In this respect, particular attention is directed to a series of key principles, which are meant to encapsulate the core of an early modern doctrine of absolutist monarchical sovereignty. This will help to delineate the history of absolutism (term, practice, and concept) in its multifarious political aspects and its rich theoretical dynamics.

Absolutism and Its Historiography

All categories ending with the suffix *'ism'* are problematic and often generate endless historiographical debates. This could not be more true than in the case of 'absolutism.' Like a great number of *'ism'* terms, absolutism is of nineteenth century coinage (Höpfl, 1983; Cuttica, 2013). Prior to 1770 'absolutism' in English, German, and French did not exist, but 'absolute monarchy,' '*monarchia absoluta*,' '*pouvoir absolu*' did (Schmale, 1998). This gap between the concept and the thing has caused vibrant – if not vehement – controversies amongst historians. Thus, for the French scholar Georges Durand absolutism is "that empty word into which everyone reads the meaning he wants, that favorite cliché of historians in a hurry" (Durand, 1976). On the other side of the Channel, Nicholas Henshall maintained that the concept of absolutism originated with the French Revolution whose goal was to condemn everything associated with the *ancien régime*. Therefore, to use it corresponds "to accept the propaganda of revolutionaries as historical reality." On the same plane as 'renaissance' and 'enlightenment,' absolutism is employed by historians to mean 'whatever' they 'want it to mean,' so that it turns out to be "an impressive excuse for sloppy thinking" (Henshall, 1992).

Leaving aside this highly polemical vein, we can plausibly argue that the study of absolutism – whose literature is immense – has been informed by three major perspectives. These are the conservative perspective for which it has always existed; the Marxist perspective for which it was an epoch between feudalism and capitalism, contaminated by and connected to both; and the revisionist perspective for which there was no absolutism and, therefore, we should not speak of it. Besides these three consolidated schools, more recently historians have taken positions which can be divided into two broad camps: those who reject the notion of absolutism per se and regard it in terms of social collaboration (between a dependent king and an increasingly powerful and influential nobility) and those who have contested the former approach. Amongst the first can be cited William Beik, Sir George Clark, James B. Collins, Sharon Kettering, Roger Mettam, Julian Swann, Peter Wilson, and Hillay Zmora. As for the second group, John Brewer, Michael J. Braddick, Daryl Dee, Brian M. Downing, Philip Gorski, John J. Hurt, and John A. Lynn can be mentioned (Cuttica and Burgess, 2012).

One important feature of these interpretations is their prevalent focus on the analysis of politics, diplomacy, dynastic structures, and court culture. Absolutism has been studied in its practical (legislative, executive, administrative) dynamics, that is, as an increasingly dominant political, fiscal, bureaucratic, and military entity from the sixteenth to the late eighteenth century in Europe. At a theoretical level, the attempt to define the pivots of absolutist political thought has been combined – not always successfully – with the examination of the relationship between the theories and the diversity of historical arguments, as well as of the interplay between the languages of absolutist theory and the realities of social experience and political practice. Taking Jean Bodin's *Les Six Livres de la République* (1576), and earlier the reworking of Roman law pursued by some French lawyers (Church, 1941) as the foundational moment in the mature development of a theory of absolutism, it has been argued that the latter can be found when "the ruler in any State holds sovereign authority, cannot be actively resisted, can change existing constitutional arrangements in a case of necessity (though he ought otherwise to maintain them), and should be obeyed by his subjects provided that his commands are not contrary to those of God and nature" (Sommerville, 1999; Russell Major, 1994).

Distinguished (more or less plausibly) both from republican polities (e.g., The Netherlands, Venice, the Swiss cantons) and despotic States (e.g., Turkey, Muscovy) at one stage or another throughout the long early modern period (*c.* 1500–1800), absolutism was considered to be the system of government of countries such as Spain, France, England, Denmark, and the Habsburg territories. This seemingly clear picture, however, has not found much agreement within the scholarly mainstream, where not only the practices, but also the principles of the absolutist State are disputed. Whereas Herbert H. Rowen skeptically stated that "absolutism was historically defined to fit the case of Louis XIV" (Rowen, 1969), Roland

Mousnier had no doubt that absolutism existed and concerned both monarchies and other forms of government, being "sovereignty which recognized no superior" (Mousnier, 1971). On his part, François Olivier-Martin singled out four types of *puissance absolue* as articulated by French monarchists since medieval times: the king as defender of the public good against private interests; the king as feudal *suzerain des suzerains* (highest overlord); the king as *Roy Très Chrétien* (protector of the faith); the king as emperor in his realm (Olivier-Martin, 1988). Jean Mesnard maintained that the seventeenth century did not invent the 'divine right monarchy' in that it continued a centuries-old legacy. Its feature was thus the development of absolute monarchy with a clear prominence of profane elements, even though religious language persisted in the description of power (Mesnard, 1996). And if for David Parker "absolutism was always in the making, but never made," and was "no more than a conservative philosophy" aimed at affirming royal authority in a world confronted by rapid change (Parker, 1983), Fanny Cosandey and Robet Descimon acknowledged that absolute monarchy constituted "a vast discourse endowed with historical effectiveness" (Cosandey and Descimon, 2002). Such a discursive mode did not simply convey the idea of a display of authority, but it enunciated a proper system of government. Etienne Thuau distinguished between a philosophy of monarchy (theorists and lawyers) and a religion of monarchy (parties, ceremonies etc.) (Thuau, 1966). Famously, Marc Bloch viewed absolutism as a sort of religion. The prince embodied a lay part and a religious one too: the image of priest and of king coexisted. Likewise, the monarch was commonly represented as *Roi-thaumaturge* (Bloch, 1924).

Another historical trait of absolutism was the tension between the personalization of kingship – whereby the sovereign was depicted as the individual who was superior because of his qualities and, therefore, one who had to be obeyed and adored; and a process of abstraction of kingship where he was associated with the purity and public presence of the State (Keohane, 1980). Although "[t]here was a vast gap between such claims [to kingly absolute power] and political reality," for Henry Kamen the theory "helped to justify subsequent efforts by princes to free themselves from the control not only of their elites but also, most importantly, of the Church and papacy, which had also made far-reaching claims to political authority." *De facto*, 'absolutism' was "an ideal construction, in which thinkers attempted to create order out of the disorder they saw around them": this means that "the theories are a significant attempt to deal with a real problem" (Kamen, 2000). On his part, Helmut G. Koenigsberger thought that "[t]he growth of absolutism […] could be furthered either by the secularization of church property where the Reformation was introduced or by an effective alliance with the old church where it was rejected" (Koenigsberger, 1978).

While such a range of divergent interpretations might leave readers daunted, it should nonetheless make them aware that to study absolutism entails a complex and nuanced analysis of the experiences of absolute power across different geographical and cultural areas of early modern Europe. This warns the reader to be suspicious of conventional views based on an evolutionary model, whereby from a paternal monarchy, absolutism was deemed to have moved on to a 'divine right' monarchy and, subsequently, to a rational and national State. In fact, absolutist parlance and politics were partly products of Renaissance, Baroque, Catholic, and Protestant thinking; they were also part and parcel of the new philosophical and scientific trends, and as such, they may be considered to have actively contributed to the emergence of the loosely defined 'modern world.' Despite the continuity of certain ideas and policies throughout time and across places – be it sufficient to refer to the *absolutist* regimes of Charles I of England, Philip II of Spain, Louis XIV of France, the Russian Tsars, Christian V of Denmark, Charles XII of Sweden, and some small States in the Italian peninsula – the patterns of change were composite and multistranded.

Absolutism in Practice and Theory

Departing from the conviction that absolutism is "an essentially negative term" which explains more about what it *was not* than about what it *was* (Bonney, 1987), it is pivotal to establish a little more clearly what absolutism means and what it is; what thinkers thought absolute power was and what they meant by it. To do this, we need to tease out the presence and status of some relevant and recurrent features of the practice (the acting out) as well as the theory (the arguing out) of absolutism. Although the analysis of early modern absolutism, here undertaken, focuses on the ideal descriptions of how a government should function rather than on how it actually worked, a few words on what factors can be considered as central to any absolutist policy are in order. Amongst the latter are the role of centralization of authority; the creation of a bureaucratic apparatus; the monocratic nature of power; the elaborate ceremonials centered on the sacredness of power (e.g., the *sacre*, the *lit de justice*, funerals); the place of the army and that of the fiscal structure; the influence of the court; the role of diplomacy and dynasty; the function of the theater, the arts, ballet, court festivals, masques, and architecture in the consolidation of power (the theatricality of power and the power of sight); the role of secularization as much as that of religion in the development and expansion of absolute monarchies. Besides, political absolutism was transconfessional in that both Catholics and Protestants (of various denominations) embraced it.

As for theory, absolutist thinkers subscribed to *some* (not indiscriminately to all) of the following principles as well as concepts: the absolute and indivisible power of monarchs; the unconditional obedience of subjects; the advancement of the power of the king (his will) versus the ancient and immemorial liberties, rights of Parliament and privileges of assemblies; a strong criticism of democracy (the people could not govern because they represented an irrational and headless multitude, so that popular government was the worst polity); the rejection of natural rights theories and the idea of the natural freedom of mankind; the derivation of power from God whereby the king was responsible for his decisions and actions to Him only; the formulation of monarchy as the best of all governments, the most perfect and natural (oneness to be found in Nature); notions of monarchical sovereignty as inalienable, hereditary, and not elective, mixed, or limited. Moreover, it was

traditionally stated that the absolute king held for himself some important powers: to make laws without the aid of any intermediate body; to condemn to death and regulate life; to wage war and stipulate peace; to coin money; to appoint ministers and magistrates; to levy taxes and impositions. The king was also often reputed to be above the law. His grace granted the estates or the Parliament the right to propose new legislation or present grievances. His will was infallible, and he could not be held accountable for his deeds to the will of the community, artificially mediated through customary or fundamental laws. The ruler's wisdom and rationality were universal. Therefore, his actions inevitably conformed to ideals of divine and natural justice. In taking decisions, he was protected by the medieval doctrine of the mysteries of State, which pertained exclusively to him as his own business in which subjects or counseling bodies could not meddle (Kantorowicz, 1957; Méchoulan, 1985). Some of these ideas were seized by a series of phrases to which authors frequently resorted to describe the absoluteness of the king's legislative power (*legibus solutus*, *lex loquens*, *quod principi placuit legis habet vigorem*) and his authority (*majestas absoluta*, *summa potestas*, *potestas perpetua*, *potestas legibus soluta*, *nam regia maiestas non moritur*). The king was both the voice and the eye of the kingdom.

Other pivotal traits of the absolutist framework were that: power stemmed from God; the Bible, nature, and history confirmed its divine origin (this frequently generated contrast between monarchies' pretences to absolute power and the Church's); the form of government could not be subject to discussion because it was a monarchy where power was transmitted from the incumbent ruler to the (in some cases strictly male) heir via primogeniture; in contrast to tyrants, kings – fathers and shepherds of their people – were regarded to be in tune with God's commandments; monarchical tyranny was better than anarchy. To these can be added the identification of the king with the monarchy, the State, and the *patria* (fatherland). Indeed, from the late sixteenth century, in countries like England and France, the sovereign became to be portrayed as *patria ipsa*. The king was not only superior, but he was immortal (not his body – *persona personalis*, but his dignity – *persona idealis*).

Making use of biblical teachings, classical models, history, analogical reasoning, Roman law, the fathers of the Church, and modern authors alike, absolutist discourse addressed both the issue of the origins of government and that of the method of government (governance). So, absolutists like Pierre de Belloy, Jean Bédé de la Gourmandiere, François Le Jay, John Cowell, Sir Robert Filmer, Claudius Salmasius, and Thomas Hobbes (Cuttica, 2011) underscored the idea that power was strictly the affair of few people. Accordingly, in many cases they couched their works in the language of the theory of order (where the hierarchical element was predominant); employed the notion of kingly 'absolute' power in the sense of the ruler being both supreme (*legibus solutus*) in his command over Parliament and subjects, and independent from an allegedly superior authority (Pope or Emperor); and resorted to the tenet that the will of the monarch ruled because he was the sole source of power (e.g., Filmer, 1991).

The affirmation of these principles during the seventeenth century signaled a transformation in absolutist reflection. This is to say that theorists depicted the king as the protector of the nation. For this reason, the parliamentary principle that the king was bound to respect the law of the land was rejected since he *was* the speaking law of the land. By making the king the exclusive interpreter of the law, absolute monarchists also dismissed the role of lawyers and jurisconsults. Privileging the political standpoint to the detriment of the legalistic one, they claimed that the absolute king was the only creator of legislation in the kingdom. This explains how some of the political conflicts of the late sixteenth and early seventeenth centuries centered on the question of who was the ultimate holder of the *interpretatio* of the law. Opting for the *interpretatio politica*, absolutists adhered to the concept that the king was totally self-sufficient in ruling the country and that no further legal aid was necessary.

Moreover, if some thinkers (e.g., King James VI and I Stuart) occasionally maintained that monarchs were limited by laws and institutions to which they, or their predecessors, had consented and submitted, others like Filmer and Hobbes whose *Leviathan* (1651) is considered to be the highest and most sophisticated articulation of absolutist doctrine – defended the voluntaristic approach to the law-making process on the basis of which the will of the sovereign had the utmost preeminence in the body politic. They insisted that the king's singleness of will gave order, unity, control, and shaped an unformed and heterogeneous mass (the body politic). Thus, the monarch wielded uncompromising power; was complete in his solitude; had a clear identity; could not be mistaken thanks to the hereditary principle (uncertainty being the faulty outcome of people's ignorance); and was perfect in his majesty and superiority. Likewise, his law had to be unconditional; clear, precise, established, and certain; flawless (Cuttica, 2012). Some seventeenth-century authors went so far as to claim that the king was justified in acting arbitrarily by the necessity to bypass the law when it was too rigid and the safety of the State was at stake. With regard to the monarch's political conduct, the notion of arbitrary power stressed the importance of values such as courage and wisdom. In some cases, this led thinkers to advocate the need to resort to violence to protect the State and the common good (e.g., Guez de Balzac, 1631; Cardin Le Bret, 1632).

These ideas glorifying the majesty of kingship derived from Roman civil law writers of the first half of the sixteenth century, especially in France. Their goal was to place the monarch above all his subjects in a way "resembling a *potestas legibus soluta* in action" (Church, 1941). Many theorists appropriated the third-century jurist Ulpian's formulae *rex solutus legibus est* and *quod principi placuit legis habet vigorem* to define "the monopoly the king exerts over the law" (Daly, 1978). Moreover, the king was considered as 'lex animata' – the breath which regulated the legislative life of the kingdom, and without whose intervention the body politic could not survive. Thus, *potestas absoluta* corresponded to various functions the monarch possessed: suspending, derogating, pardoning, mitigating, and facing unforeseen circumstances. It was a form of 'discretionary power' which shaped the concept of royal prerogative. These principles greatly emphasized the role of the king as national monarch capable of guaranteeing the making of the law and its preservation. To put political thinking on this track were *politiques* such as Jean Bodin, Etienne Pasquier, Pierre de Belloy,

Louis Servin, William Barclay, Pierre Pithou, Guy Coquille, and Christofle de Thou, whose works proved largely influential across Europe. In particular, they shaped the ideas and the policies based on the 'reason of State' paradigm, which Cardinal Richelieu and his supporters employed in the seventeenth century (Cosandey and Descimon, 2002; Church, 1941).

According to the historian Jean-François Courtine, the 'absolutisation' of power in the *État Baroque* also occurred through "a resacralisation of the State ['État']." It was thus a 'crypto-theological' element in the political ideas of the early seventeenth century that conveyed concepts of sovereignty and models of State-practice. In fact, the power of the State and the features theorists attributed to it, derived from papal power and the principles associated with it, especially the concept of pontifical *plentitudo potestatis*. The sovereign appeared as a superior human type. In practice, it was the process of 'sacralisation' and 'deification' that led to the 'self-governance' and 'rationalization' of political monarchy. It followed that the Church was an absolute and rational monarchy founded on a mystical basis, while the State was a mystical monarchy founded on a rational basis (Courtine, 1985). However, it is essential to remember that not all thinkers expounded an idea of the sacralised image of the king. In fact, absolutists often presented the sovereign as an efficient, patriotic, and rational ruler at a time of both internal and external tension. For Joël Cornette the king was a plurality of things. He embodied a series of roles: "overlord, lawmaker, judge, warrior, demigod, healer, minter and obedience-supplier through levies." The État Baroque was the response to a society that was anxious and unstable. It represented the moment when the scientific revolution and the modern prince met to engender a new spirit that opposed traditional ways of thinking and proposed new shared values at the center of which stood reason. This phase corresponded to the creation of an *État*-Leviathan whose pivots were order, authority, and profit (Cornette, 1985).

This model of government served as spearhead with which to attack republican institutions seen as entities that fortified the private interests and individual passions of a minority of powerful nobles. In contrast to eighteenth-century philosophers like Montesquieu, absolutist theorists argued that republics represented a deleterious mixture of aristocratic factionalism and democratic aspirations. On the contrary, the pursuit of the *bien publique* could only be guaranteed in an absolute polity. Linking patriotism and rationality as cornerstones of monarchical government, absolutists reacted to Montesquieu's well-known claim that only republics prompted people to embrace civic virtue and patriotic feelings. They contended that a new 'decision-making' monarchy assured prosperity and safety (*Polizey*) through a very empirical approach to problems in the polity. This 'public safety'-based structure was founded on a balance of forces where the estates were replaced by the individual citizen, and economic rationality was 'accompanied' by 'moral improvement.' This gave a novel – more modern – face to monarchy. Such a situation created the mutual collaboration of an enlightened State-apparatus guaranteeing 'convenience and safety' with an 'enlightened citizenry' patriotically engaged in abiding by the laws and loyal to the government representing the fatherland (Kontler, 2012). In summary, political, theological, metaphysical, and temporal aspects all contributed to define the history of absolutism and to establish the administrative, fiscal, and military form of the modern State.

Besides all of the aforementioned characteristics, to investigate the nature of absolutism in historical perspective means to take into further account the significant changes that occurred in European societies and their political structures during the eighteenth century. In other words, it is important to follow – although not teleologically – the trajectory of both practice and theory from medieval England to eighteenth-century Prussia through Enlightenment France and Naples, and from the vocabulary of indivisible, inalienable, hereditary sovereignty to descriptions of the monarch as the constitutional artisan of a complex administrative apparatus. Hence absolutism and monarchical discourse need to be read in relation to republicanism (especially, its conflict with it); enlightened despotism; patriarchalism; patriotism; liberty (e.g., intellectual liberty and freedom of opinion); property (protection of subjects' property); the state of nature; scientific developments; toleration; iconographic propaganda; and the ethical sphere (both public and private). The advent of the eighteenth century forced absolutists to confront new factors informing *the art of politics*: cameralism, publicity, political secularisation, attention to the civil freedoms of citizens, the politicization of sculpture (as was the case, for instance, under Empress Catherine II of Russia), the new utilitarian spirit. A different era had dawned and with it different ways of conceiving (absolute) power.

Absolutism: Terminology and Typology

It has been observed that three groups of meanings were attached to the term 'absolute': uncompromising, unconditional, complete; positive, certain, decided; faultless, perfect (Daly, 1978). As a result, it seems plausible to identify – at least – seven theoretical cornerstones enabling a configuration of political absolutism that concerns the ruler:

- Ab-soluteness (absolved and above)
- Independence (internal and external)
- Solitude
- Perpetuity
- Exclusive inequality (of the sovereign compared to the equality of the rest of the subjects)
- Individuality
- Effectiveness

This configuration of absolute power helps to separate absolute government from tyranny in that the latter was marred by cruelty and by the search for private interest, while the former aimed at the pursuit of the public good. Tyranny acted in a vacuum (legal, ethical, hereditary), whereas the absolute king intervened in the juridical sphere with his supreme power, in order to mitigate the rigidity of the law and avoid injustice. Although it is often claimed that the degree of absoluteness of a king's power depended on the extent to which the law of nature was thought to legislate, this does not mean that he was held to be limited. In fact, absolutists, of the kind described here, regarded monarchs as capable of restricting the domain of

political life that the law of nature could regulate. Consequently, not only laws but also customs were the outcome of the supreme decision of the *lex loquens* king, who was depicted as the creator of the nation's ethos. This implies paying attention to the following units of meanings as foundations of absolute rulership: arbitration, choice, judgment, decision, will, mastery, and authority.

On the whole, any analysis of absolutism needs to examine what theorists wrote and described, as well as the concepts they expressed in a prescriptive and normative way, according to their theoretical, ideological, and doctrinal objectives. Reality might not have necessarily corresponded to their designs, but what is interesting is precisely the nature of these designs, projects, visions of political society and government, power and authority, sovereignty, and liberty. More focus should be on the conceptual sphere, especially in relation to the idea of sovereignty and, within it, to the modality(ies) of power. In other words, it is necessary to investigate how ideas formed; how they served to think or reshape institutions, polities, and systems of government; how intellectual categories helped communities to reflect on their organization and their life; how concepts legitimized actions; how they gave meanings as well as significance to different narratives of the past. Absolutism should thus be approached as *words in action*. It should also be read as part of humanist philosophy in that it is placed at the center of its reflection, man with his rationality and will.

The foregoing points enable us to identify a series of paradigms – interdependent and mobile rather than fixed and a-contextual ones – that characterize the history of European absolutism. These different strands help not only to comprehend the role of ideas in the practice of absolute policies, but they also show the weight of certain thinkers in their transnational dimension. As a result, a sample of what are the most relevant and – it need be repeated – historically overlapping, nonmutually exclusive but interchangeably connected political, ethical, and intellectual models informing absolutism can be provided:

1. Machiavellian absolutism (Tacitean) >> insistence on prudence and the tradition of the mirror of princes (religion as an instrument in the hands of those in power);
2. Bodinian absolutism >> emphasis on the indivisibility and inalienability of sovereignty, but also on the role of positive laws and property;
3. Patriarchalist absolutism >> Filmer and the Adamite paradigm (family-model; ruler as the founder and shaper of the commonweal's ethos, its customs, and laws);
4. Patriotic absolutism >> French monarchists against Jesuit and Ultramontanist ideas; monarchist discourse in eighteenth-century Denmark and Prussia;
5. Constitutional absolutism (royalism) >> the specific case of seventeenth-century England;
6. Reason of State absolutism >> priority given to the *arcana imperii*;
7. Divine absolutism >> focus on the divine right theory (references to witchcraft and mystical authority). Divine rights of kings primarily concerned obedience;
8. Hobbesian absolutism (*Leviathan*) >> role of the state of nature and totalization of politics;
9. Miraculous absolutism >> royal touch and insistence on the king's physical and moral characteristics (king seen not only as God, saint, and giver of justice, but also as wizard).

These models operate as analytical templates through which to explore and decipher the nature of absolutism in early modern Europe. They present the latter as an intellectual and practical intercourse made of different (at times contrasting) languages, vocabularies, paradigms, functions, and dynamics. They help to go beyond the interpretative impasse whereby absolutism is inevitably the expression of static, archaic, and oppressive political societies. They also prompt to revise two other common assumptions about absolutism: that it was a myth and that it was founded exclusively on the idea of the sacred and the dimension of the divine. Likewise, they do not look to absolutism exclusively as an "ahistorical phenomenon connected with the aggrandizement and the centralization of the state." Instead of proposing an ahistorical image of absolutism as "a strange and dangerous beast" aspiring "to rise above reality" (Kossmann, 1976), they bring back into play the various languages that theorists deployed in disputes and, at the same time, cast light on traditionally antagonistic readings of absolutism as either feudal, bourgeois, or fiscal.

Thus, if we subscribe to the prescription that for the historian to make a legitimate use of an 'ism' it is necessary to have people deriving ideas and finding in a thinker "an authoritative expositor" of views they agreed with and shared (Höpfl, 1983), then absolutism (as we have used it in this article) is plausibly employed. In fact, we have described a core of political, juridical, and doctrinal principles adopted by early modern theorists to portray specific images of kingship as well as to respond to rival visions of the body politic in debates across Europe. In addition to connecting absolutism to historical and social phenomena, this article has also depicted it as the embodiment of particular political traditions, theoretical languages, and rhetorical tropes. In this respect, absolut*ism* needs to be approached as a network of meanings and practices that shaped the history of early modern Europe.

Conclusion

There are two more aspects of the history of absolutism that are worth mentioning. First, it is important to underscore that both as practice and as doctrine, absolutism took different shapes in different contexts at different times. Despite all the common elements highlighted above, it is historically essential to be aware of the contextual varieties surrounding the phenomenon and its ideas. Second, it has to be remembered that there exists a series of concepts which have recurrently (but often wrongly) been associated with absolutism. Amongst these the main ones are: dictatorship, caesarism, autocracy, despotism, and totalitarianism. A few words are in order with regard to the last two. As for despotism, some authors considered it a legal form of government, implying the servile nature of a people where the relation between governor and governed was like that between a master and his servants/slaves. Therefore, it was generally associated with so-called barbarian peoples (Asiatic populations). In the eighteenth century, philosophers such as Boulanger, Helvétius, the physiocracts, Mably, and Rousseau

viewed despotism as the death of the body politic, while Benjamin Constant spoke of 'despotisme de la liberté.'

As for totalitarianism, we need not confuse it with absolutism in that the former is quintessentially a twentieth-century phenomenon (e.g., Nazism, Fascism, Stalinism) corresponding to the total identification of each individual with the whole State. Totalitarianism entails the politicization of life, and of every dimension of it. For Hannah Arendt, it implies the alienation of private existence and the transformation of human nature. An ideology of terror and the encompassing role of propaganda are central factors in the construction of a totalitarian polity. Likewise, a one-party-only structure and the systematic deployment of a brutal police force are pivotal to its functioning. The identification of the enemy, of the often-fictitious 'foe,' represents another crucial trait of this form of authoritarian regime. Given that the use of technology and a general control over the economic sphere inform it, its workings become a pervading threat within society (Arendt, 1951). Totalitarianism asks people not only for their active political support, but their full participation in the life of the State, so that their pastimes are invested too. This explains the presence of an oppressive apparatus of constant surveillance of the private existence and space of the citizens/subjects. The process of industrialization and the related concept of the masses are also considered key aspects of the totalitarian State (Friedrich and Brzezinski, 1956).

It is widely accepted that the nineteenth century with the gradual emergence of democracy, constitutional change, and national awakening in many European countries, as well as the forging of new political ideas/ideals, put an end to absolutism in its more traditional mold. However, any attempt to pursue a much-invoked "reassessment of absolutism" (Bonney, 1987) cannot overlook the fact that contemporary sociologists and politologists speak of 'neo-absolutism,' which in contrast to centralized absolutism, is "post-modern and polycentric, dissipative [dispersed, scattered], omnipresent and yet dislocated. It has too many centres, almost all of which are invisible. It is in the net with innumerable ramifications. Therefore, it becomes "elusive" (and, maybe, even impregnable)" (Revelli, 2010).

If anything, these considerations confirm that absolutism is more than a complex historiographical category. They show that it continues to be a historically fascinating and conceptually rich object of study across multiple disciplines. It can thus be said that the interesting history of absolutism is not over yet.

See also: Bureaucratization and Bureaucracy, History of; Dictatorship in History; Parliaments, History of; Political History: History of Politics; State, History of.

Bibliography

Anderson, P., 1974. Lineages of the Absolutist State. Verso, London.

Arendt, H., 1951. The Origins of Totalitarianism. Schocken Books, New York.

Asch, R.G., Duchhardt, H. (Eds.), 1996. Der Absolutismus – ein Mythos? Strukturwandel monarchischer Herrschaft in West- und Mitteleuropa (ca. 1550–1700). Bölau, Köln.

Bloch, M., 1924. Les Rois Thaumaturges. Etudes sur le caractère surnaturel attribué à la puissance royale, particulièrement en France et en Angleterre. Publications de la Faculté de Lettres de l'Université de Strasbourg, Strasbourg-Paris.

Bodin, J., 1576. Les Six Livres de la République. Paris.

Bonney, R., 1987. Absolutism: what's in a name? French History 1, 93–117.

Burns, J.H., 1986. Absolutism: The History of an Idea, Creighton Trust Lecture. University of London, London.

Church, W.F., 1941. Constitutional Thought in Sixteenth-Century France. A Study in the Evolution of Ideas. Harvard University Press, Cambridge, Mass.

Cornette, J., 1985. Fiction et Realité de l'Etat Baroque (1610–1652). In: Méchoulan, H. (Ed.), L'État Baroque 1610–1652. Vrin, Paris, pp. 7–87.

Cosandey, F., Descimon, R., 2002. L'absolutisme en France. Histoire et historiographie. Seuil, Paris.

Courtine, J.F., 1985. L'Héritage Scolastique dans la Problématique Théologico-politique de l'Age Classique. In: Méchoulan, H. (Ed.), L'État Baroque 1610–1652. Vrin, Paris, pp. 89–118.

Cuttica, C., 2011. Anti-jesuit patriotic absolutism: Robert Filmer and French ideas (ca. 1580–1630). Renaissance Studies 25, 559–579.

Cuttica, C., 2012. An absolutist trio in the early 1630s: Sir Robert Filmer, Jean-Louis Guez de Balzac, Cardin Le Bret and their models of monarchical power. In: Cuttica, C., Burgess, G. (Eds.), Monarchism and Absolutism in Early Modern Europe. Pickering & Chatto, London, pp. 131–145.

Cuttica, C., 2013. To Use or Not to Use. The Intellectual Historian and the Isms: A Survey and a Proposal, Etudes Epistémè 13, Varia, http://revue.etudes-episteme.org/?to-use-or-not-to-use-the.

Cuttica, C., Burgess, G. (Eds.), 2012. Monarchism and Absolutism in Early Modern Europe. Pickering & Chatto, London.

Daly, J., 1978. The idea of absolute monarchy in seventeenth-century England. Historical Journal 21, 227–250.

Durand, G., 1976. What is absolutism? In: Hatton, R. (Ed.), Louis IV and Absolutism. Macmillan, London, pp. 18–36.

Filmer, R., 1991. In: Sommerville, J.P. (Ed.), Patriarcha and Other Political Writings. Cambridge University Press, Cambridge.

Friedrich, C.J., Brzezinski, Z.K., 1956. Totalitarian Dictatorship and Autocracy. Harvard University Press, Cambridge, Mass.

Guez de Balzac, J.-L., 1631. Le Prince. Paris.

Henshall, N., 1992. The Myth of Absolutism: Change and Continuity in Early Modern European Monarchy. Longman, London.

Hobbes, T., 1996. In: Tuck, R. (Ed.), Leviathan. Cambridge University Press, Cambridge.

Höpfl, H., 1983. Isms. British Journal of Political Science 13, 1–17.

Kamen, H., 2000. Early Modern European Society. Routledge, London, New York.

Kantorowicz, E., 1957. The King's Two Bodies. Princeton University Press, Princeton.

Keohane, N.O., 1980. Philosophy and the State in France: The Renaissance to the Enlightenment. Princeton University Press, Princeton.

Koenigsberger, H.G., 1978. Monarchies and parliaments in Early Modern Europe: dominium regale or dominium politicum et regale. Theory and Society 5, 191–217.

Kontler, L., 2012. Polizey and patriotism: Joseph Von Sonnenfels and the legitimacy of enlightened monarchy in the gaze of eighteenth-century state science. In: Cuttica, C., Burgess, G. (Eds.), Monarchism and Absolutism in Early Modern Europe. Pickering & Chatto, London, pp. 75–90.

Kossmann, E., 1976. The singularity of absolutism. In: Hatton, R. (Ed.), Louis IV and Absolutism. Macmillan, London, pp. 3–17.

Le Bret, C., 1632. De la Souveraineté du roy. Paris.

Méchoulan, H. (Ed.), 1985. L'État Baroque 1610–1652. Vrin, Paris.

Mesnard, J., 1996. La Monarchie de Droit Divin, Concept Anticlérical. In: Ferreyrolles, G. (Ed.), Justice et Force. Politiques au Temps de Pascal. Klincksieck, Paris, pp. 111–138.

Miller, J. (Ed.), 1990. Absolutism in Seventeenth-Century Europe. Macmillan, Basingstoke.

Mousnier, R., 1971. The exponents and critics of absolutism. In: Cooper, J.P. (Ed.), The New Cambridge Modern History, vol. iv, first ed., 1970, ch. iii. Cambridge University Press, Cambridge, pp. 104–131.

Olivier-Martin, F., 1988. L'Absolutisme Français. Editions Loysel, Paris.

Parker, D., 1983. The Making of French Absolutism. Arnold, London.

Revelli, M., September 23, 2010. L'Eguaglianza Uccisa Dal Progresso. La Stampa, p. 8.

Rowen, H.H., 1969. Louis XIV and absolutism. In: Rule, J.C. (Ed.), Louis XIV and the Craft of Kingship. Ohio State University Press, Columbus, pp. 302–316.

Russell Major, J., 1994. From Renaissance Monarchy to Absolute Monarchy. French Kings, Nobles & Estates. Johns Hopkins University Press, Baltimore/London.

Schmale, W., 1998. The future of "absolutism" in historiography: recent tendencies. Review Article. Journal of Early Modern History 2, 192–202.

Sommerville, J.P., 1999. Royalists and Patriots. Politics and Ideology in England 1603–1640, second ed. Longman, London/New York.

Thuau, E., 1966. Raison d'État et pensée politique à l'époque de Richelieu. Colin, Paris.

Absorptive Capacity (of Organizations)

Heiko Gebauer, Eawag – Swiss Federal Institute of Aquatic Science and Technology, ESS – Environmental Social Sciences, Dübendorf, Switzerland
Hagen Worch, Swiss Distance University of Applied Sciences/ Fernfachhochschule Schweiz, Institute for Management and Innovation, Regensdorf, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A firm's competitive advantage originates increasingly from absorbing external knowledge. Absorbing external knowledge and the underlying learning processes are referred to as a company's *absorptive capacity*. In this article, we outline research trends on absorptive capacity. We apply a bibliometric analysis to describe the concept's historical development, define the intellectual core of the absorptive capacity concept, and discuss recent conceptualizations. Then, we identify two prominent streams in the absorptive capacity literature and provide a new approach on how to integrate them. Finally, we provide an outlook on possible themes in future research on absorptive capacity.

Introduction

Both practitioners and academics increasingly recognize that competitive advantages no longer rely on internal knowledge alone, but rather originate from absorbing external knowledge. This in turn is based on learning processes, which are directed at exploring, assimilating, transforming, and exploiting external knowledge (Camisón and Forés, 2010; Gebauer et al., 2012; Lane et al., 2006). These learning processes support companies in converting their external knowledge into innovations. The literature refers to the absorption of external knowledge and the underlying learning processes as a company's *absorptive capacity* (Cohen and Levinthal, 1989, 1990).

Absorptive capacity, as an application of external knowledge for commercial purposes, can lead not only to product or service innovation but also to strategic innovation. Strategic innovation aims at reshaping the existing business model, opening up new and uncontested markets, and creating a leap in customer value (Christensen et al., 2002). Thus, absorptive capacity is a crucial channel for creating new and maintaining existing competitive advantages.

This article outlines most important research trends in the theoretical and practical discussion of absorptive capacity. Applying a bibliometric analysis, we start with describing the historical development of the absorptive capacity research. Afterward, we define the intellectual core of the absorptive capacity concept and discuss the recent conceptualizations. We identify two prominent streams in the management literature that are specifically linked to the concept of absorptive capacity and provide a new approach on how to integrate the two streams. The article ends with an outlook into the future of the absorptive capacity research.

Absorptive Capacity

Absorptive capacity was originally defined as a firm's ability to recognize the value of new information, assimilate it, and apply it for commercial purposes (Cohen and Levinthal, 1990). Cohen and Levinthal's (1989, 1990) notion of absorptive capacity refers to the understanding of R&D investments being not only important for creating inventions but also to the ability of a firm to internalize knowledge from external sources (Schildt et al., 2012). Since these early contributions, there have been an increasing number of articles. We applied bibliometric methods for analyzing the existing research. We first tracked the citations of the two seminal contributions by Cohen and Levinthal (1989, 1990). We second identified articles using 'absorptive capacity' as one of the keywords. We conducted the search in Scopus. The Scopus database was considered as more suitable than alternatives such as Web of Science or Google Scholar. The Web of Science database contains fewer journals than Scopus. Google Scholar has no advanced search function. The search results were merged into a master list of articles. By combining both lists, we checked for double entries and inconsistencies such as misspelled names or wrong publication years. The combination yielded 828 articles.

The second bibliographic method is citation analysis, which is an acceptable surrogate for the intellectual core of the absorptive capacity research (Culnan, 1986). We tracked the number of articles published per year, the journals, subject areas, and keywords. As shown in **Figure 1**, research on absorptive capacity is developing into a highly dynamic management discipline. The number of articles has increased steadily, with a late peak in 2012 and 2013, at 117 and 114 articles, respectively

The historical development of the absorptive capacity concept can be distinguished into three phases:

- *Phase 1*: Early emergence of the absorptive capacity concept (1989–2001)
- *Phase 2*: Conceptual foundation and establishment as a research domain (2002–07)
- *Phase 3*: Emergence of an important research domain in business research (2008–)

With the publication of Cohen and Levinthal's (1989, 1990) seminal articles, the concept of absorptive capacity had been formulated. As researchers and practitioners had quickly acknowledged absorptive capacity as a relevant concept to better understand organizational learning, knowledge

 http://dx.doi.org/10.1016/B978-0-08-097086-8.73106-0

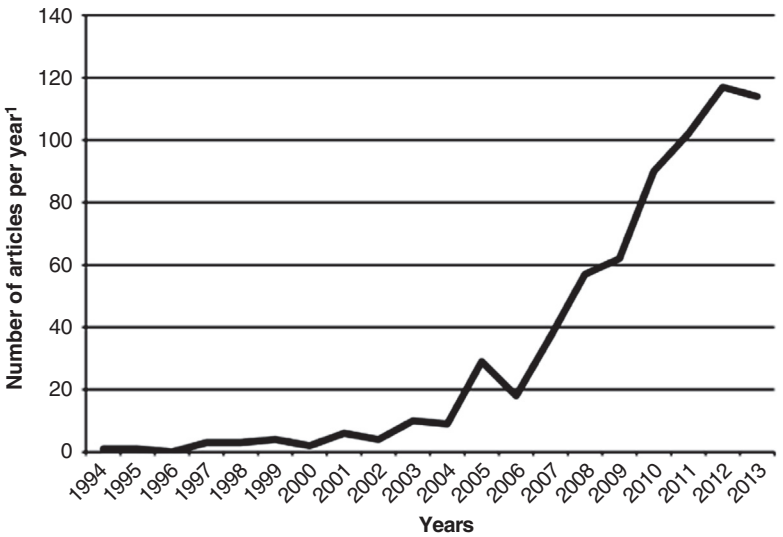

[1]– Number of articles per year using absorptive capacity as a key word. Database: Scopus.

Figure 1 Articles on absorptive capacity in the social science field.

management, R&D, innovation and competitive advantage, the concept was discussed, applied, and replicated in numerous contributions in the following years (Szulanski, 1996; Tsai, 2001). In the first phase, the concept gained momentum within management science.

In the early 2000s, the concept has been further developed with a number of important conceptual refinements and extensions. This period can be interpreted as a second phase, in which the absorptive capacity concept was established as an own research domain within the field of management research and organization theory. Central contributions in this period were among others articles by Zahra and George (2002), Jansen et al. (2005), Lane et al. (2006), and Todorova and Durisin (2007).

The third phase started at the end of the 2000s, when publications on the absorptive capacity concept were increasingly used as an approach to frame empirical research. In this enduring phase, the number of publications on the topic has experienced a substantial increase every year until a peak was reached in 2012 and 2013 with almost 120 annual publications. Various contributions inquired into the relationship between key constructs of the absorptive capacity concept. Other research aimed at gaining a better understanding of what role context conditions and other external factors played for absorptive capacity to have a significant impact on innovation performance. Contributions in this third phase include Lewin et al. (2011), Vasudeva and Anand (2011), Schildt et al. (2012), and Gebauer et al. (2012, 2014).

Five important keywords emerged from the analysis of keywords and subjects of the absorptive capacity articles. These are knowledge management (e.g., Lane and Lubatkin, 1998; Minbaeva et al., 2003; Vasudeva and Anand, 2011), organizational learning (e.g., Schildt et al., 2012), innovation management (e.g., Lichtenthaler and Lichtenthaler, 2009), research and development (Boschma and ter Wal, 2007; Stock et al., 2001), and competitive advantages (e.g., Lewin et al., 2011). Absorbing external knowledge requires adequate knowledge management and organizational learning processes. Having such processes in place to absorb external knowledge is essential for a company's innovation management and research and development. Due to the effect of these processes, it is argued that absorptive capacity is a key driver for gaining a competitive advantage.

All five components are highly relevant for the absorptive capacity concepts. However, as indicated in **Figure 2**, the relative importance of each component has changed across the three phases. Whereas the relative importance of knowledge management, innovation management, and research and development has increased, articles on absorptive capacity put less emphasis on organizational learning and competitive advantage.

Knowledge management, innovation management, and research and development cover increasingly diverse research themes. Innovation management had initially a relative narrow focus, but has now been extended to include discussions on innovation performance and open innovation into the concept of absorptive capacity (e.g., Boschma and ter Wal, 2007; Lichtenthaler and Lichtenthaler, 2009). Knowledge management has diversified into themes such as knowledge transfer, knowledge-based systems, knowledge acquisition, knowledge sharing, and human capital (e.g., Gebauer et al., 2012, 2014; Minbaeva et al., 2003; Shenkar and Li, 1999; Vasudeva and Anand, 2011). Articles on research and development have linked absorptive capacity to technology transfer, patents, and inventions (e.g., Grünfeld, 2003; Stock et al., 2001).

As indicated in **Figure 2**, absorptive capacity is also related to many other topics. In phase 3, for example, there have been 17 articles about the link between absorptive capacity and dynamic capabilities. These articles embedded absorptive capacity in the debate on dynamic capabilities, which originates from the evolutionary theory of the firm (Zollo and Winter, 2002). Dynamic capabilities enable companies to respond to changes in the business environment. A similar argumentation is used for the absorptive capacity concept.

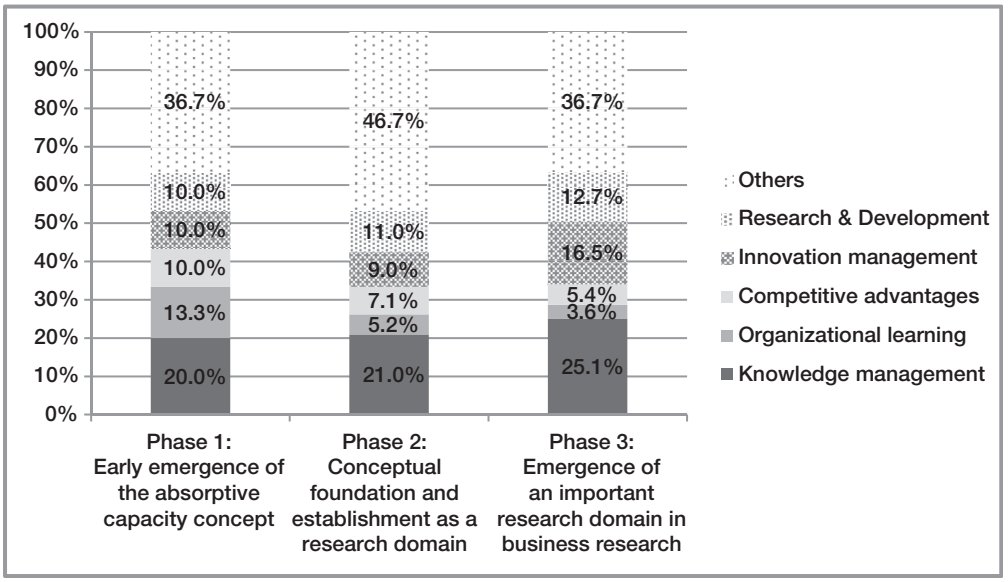

Figure 2 Main themes across the phases of absorptive capacity research.

With its focus on a firm's ability to recognize, assimilate, and apply external knowledge, the concept of absorptive capacity has a strong link to the literature on dynamic capabilities, where the reconfiguring of existing knowledge structures is a defining characteristic.

Among the 828 articles, we selected those with 100 and more citations. Applying the threshold of 100 citations leads to a core set of 26 most influential contributions to the absorptive capacity concept (Intellectual core of absorptive capacity). This core is informed purely by academic papers indicating the strong research interest in the topic. These papers are published in most prominent journals such as Strategic Management Journal, Academy of Management Journal, Academy of Management Review, and Administrative Science Quarterly among others. The 26 articles account for 72.8% of the total number of citations in the entire set of the 828 articles published on absorptive capacity.

Recent Developments in Absorptive Capacity Research

Conceptualization of Absorptive Capacity

Since the original definition, various conceptualizations of absorptive capacity have emerged (e.g., Lane et al., 2006; Lewin et al., 2011; Gebauer et al., 2012). While the early conceptualizations focused on R&D-related issues, later research broadened the concept to developing absorptive capacity at the organizational level (Tsai, 2001). The term knowledge within the absorptive capacity concept subsumes both procedural and declarative knowledge. The latter provides a description of state such as information on customer needs, technological trends, and strategic plans and refers to the notion of know-what. Procedural knowledge describes the current practices inside a firm such as the tools and processes that companies use to determine customer needs, extrapolate technological trends, and formulate strategic responses (Kogut and Zander, 1992). The notion of know-how captures procedural knowledge. There are several dimensions along which the absorptive capacity concept has evolved.

Taking a very general perspective, a number of existing conceptualizations describe absorptive capacity as the independent variable and innovation performance as the dependent variable. Such a view tends to capture absorptive capacity as an aggregated construct. It hardly allows researchers to explain the underlying processes of absorptive capacity at the firm level. Yet, aggregated concepts enable scholars to theorize and empirically test how variances of the firm-level characteristics, specific determinants, and context conditions of absorptive capacity affect innovation performance. Some of these general conceptualizations entail moderators, which can either strengthen or weaken the relationships between absorptive capacity and innovation outcomes. For example, Van den Bosch et al. (1999) propose business strategy as such a moderator. A first-mover strategy yields advantages when it involves building up absorptive capacity. By contrast, a follower strategy requires lower absorptive capacity. Similarly, Tsai (2001) argues that the centrality of a company's position in the knowledge network strengthens the impact of absorptive capacity on performance.

In contrast to the variance models, other research on absorptive capacity has defined the research agenda in this field over the past years. This stream focused on detailing the underlying processes of absorptive capacity. This led to a clearer understanding of the nature of absorptive capacity and resulted in the emergence of various process models. Zahra and George (2002) introduced the distinction of absorptive capacity into potential absorptive capacity and realized absorptive capacity. The former captures knowledge acquisition and assimilation, which refer to a firm's capacity to identify and acquire externally generated knowledge. Realized absorptive capacity refers to the capacity to transform and exploit the knowledge for

commercial purposes. The literature is contradictory on the sequence of knowledge absorption. Zahra and George (2002) and Jansen et al. (2005) conceptualize the sequence as a linear relationship between acquisition, assimilation, transformation, and exploitation, whereas Todorova and Durisin (2007) interpret assimilation and transformation as two parallel elements. Knowledge is assimilated, if the existing cognitive structure of organizational members does not change. Transformation means that new knowledge is interrelated with changing existing cognitive structures.

Independent of the sequence of knowledge assimilation and transformation, potential and realized absorptive capacity are linked through an efficiency factor. A higher efficiency factor leads to greater innovation performance because organizations pursue a course of action in response to their potential knowledge (Zahra and George, 2002). According to the efficiency factor, Winter (2000) suggests that satisficing rather than optimizing guides managers in transforming potential into realized knowledge. According to this argument, knowledge stocks accumulated through potential absorptive capacity function as strategic reference points and aspiration levels.

An important aspect of the process models is that they enable scholars to conceptualize and therefore capture learning processes as one of the central underlying elements of absorptive capacity. With the increasing emphasis on learning processes, the absorptive capacity literature has recently developed close links to two other streams in the management strategy and organization theory literature; on the one hand, the literature on interorganizational knowledge transfer and learning and on the other hand, the literature on dynamic capabilities with its focus on sustained competitive advantage through processes of learning and change. In the following, we present some contributions on absorptive capacity in the light of these two streams in management research and present the implications of these contributions.

Absorptive Capacity and Learning Processes

Both potential and realized absorptive capacities are cumulative and depend on past experiences. Past experience and prior knowledge are key elements in the absorptive capacity concept. Prior R&D investments – and therefore prior knowledge – determine at the firm level the extent and diversity at which an organization is able to recognize and understand, assimilate, transform, and apply knowledge from a variety of external sources. Thus, the starting point for a firm to absorb external knowledge and the kind of knowledge that a firm absorbs are cumulative and path dependent. This has a number of implications.

One implication is that efforts to develop absorptive capacity in one period make it easier to accumulate it in the next period (Cohen and Levinthal, 1990). Accordingly, absorptive capacity is not static, but rather evolves through learning processes (Lane et al., 2006; Todorova and Durisin, 2007). Lane et al. (2006) conceptualize absorptive capacity as a firm's ability to utilize external knowledge through three sequential learning processes: exploratory, transformative, and exploitative. Exploratory learning is about the acquisition of external knowledge and corresponds to the notion of potential absorptive capacity. Through exploitative learning, companies can apply acquired knowledge and therefore combine existing knowledge with newly generated knowledge. Such learning reflects the concept of realized absorptive capacity. Transformative learning links the exploratory and exploitative learning processes. It can span from maintaining and reactivating knowledge to conversion and combination of knowledge (Flatten et al., 2011).

Camisón and Forés (2010) extend that perspective on transformative learning further. The authors describe transformative learning as developing and refining "… the internal routines that facilitate the transference and combination of previous knowledge with the newly acquired or assimilated knowledge. Transformation may be achieved by adding or eliminating knowledge, or by interpreting and combining existing knowledge in a different, innovative way" (Camisón and Forés, 2010: p. 709).

Alternatively, exploratory and exploitative learning processes can be linked through assimilative learning. The term assimilation is more about integrating this knowledge into the organizational knowledge base. Considering the dynamic capability perspective, exploratory and transformative learning are of particular importance in turbulent environments.

Another dimension of interorganizational knowledge transfer and learning is directly linked to the knowledge base on which absorptive capacity is built. In a recent contribution, Vasudeva and Anand (2011) highlight the distinction between latitudinal and longitudinal absorptive capacity. "[L]atidudinal absorptive capacity processes and uses diverse knowledge, and longitudinal absorptive capacity processes and uses distant knowledge" (Vasudeva and Anand, 2011: p. 612). This is an important distinction and a further development of the absorptive capacity concept because it decomposes absorptive capacity into qualitatively different parts. In fact, this distinction specifies the path dependency of absorptive capacity. If the two components are linked to different search approaches, then this has the implication that a firm may not only have too little absorptive capacity in general but may also have an imbalance (or suboptimal) composition of latitudinal and longitudinal absorptive capacity. Consequently, a firm may be restricted regarding the knowledge that it can absorb. This, in turn, determines the breadth of areas, from which a firm can absorb knowledge and subsequently develop new knowledge and innovation. In this sense, the concept of absorptive capacity is closely related to interorganizational learning and knowledge transfer and the internal capacities to recognize and access this knowledge.

Absorptive Capacity and Dynamic Capabilities

The concept of absorptive capacity is also embedded in the debate on dynamic capabilities (Teece et al., 1997; Zollo and Winter, 2002). The debate on capabilities originates from the resource-based view of the firm. Organizational capabilities are firm-specific resources and processes to accomplish strategic goals by utilizing the available know-how and nonfirm-specific resources (e.g., Teece et al., 1997; Zollo and Winter, 2002). One of the most prevalent terms for capabilities is the distinction into operational and dynamic capabilities (Teece et al., 1997).

Operational capabilities comprise the competences and skills that determine a firm's efficiency and effectiveness in executing its current business activities and in accomplishing specific tasks (Zahra et al., 2006). Dynamic capabilities encapsulate the evolutionary nature of resources in firm organizations (Teece et al., 1997; Zahra and George, 2002). Dynamic capabilities enable companies to respond to changes in the business environment and change their operational capabilities accordingly. Teece et al. (1997: p. 516) defined dynamic capabilities as "… the firm's ability to integrate, build, and reconfigure internal and external competences to address rapidly changing environments." Zollo and Winter (2002: p. 340) specify the meaning of dynamic capabilities as being "… a learned and stable pattern of collective activity through which the organization systematically generates and modifies its operating routines in pursuit of improved effectiveness …". In fact, dynamic capabilities define an organization's ability to identify the need for change, formulate an appropriate response, and implement a course of action.

Dynamic capabilities avoid a competence trap, in which competences become irrelevant due to changes in the business environment. Zahra et al. (2006: p. 918) define dynamic capabilities as the managerial ability "to reconfigure a firm's resources and routines in the manner envisioned and deemed appropriate." Dynamic capabilities are beneficial in highly turbulent settings (Zahra et al., 2006).

As seen in more detail below, it is exactly in this context of distinguishing between operational and dynamic capabilities to interpret the relationship between absorptive capacity as learning and knowledge transfer processes and the dynamic capabilities as a higher level concept. As absorptive capacity is defined as a firm's ability to recognize, assimilate, and apply external knowledge, it can be interpreted in the context of dynamic capabilities and specifically in the context of being a concept that underpins the conditions to generate the level of adaptability that is required for adjusting and reconfiguring the absorptive capacity and its corresponding processes at the organizational level to acquire new knowledge from external sources. Some recent contributions in the absorptive capacity literature have highlighted this link (e.g., Lewin et al., 2011; but also Lane et al., 2006).

In sum, there have been significant developments at the conceptual level of absorptive capacity with some recent contributions on interorganizational knowledge transfer and dynamic capabilities. However, both streams lack some aspects. The interorganizational learning approach is relatively quiet about the capabilities required to structure and configure the adequate absorptive capacity processes. The capability approach, in contrast, remains often on a relatively general conceptual level without making the links to the various absorptive capacity processes (i.e., acquisition, assimilation, transformation, exploitation) explicit. In the next section, we provide an approach how to integrate the two streams and provide an interesting new view of absorptive capacity as learning and adaptation processes with a link to dynamic capabilities. **Table 1** summarizes the most prominent conceptualization of absorptive capacity.

Table 1 Most prominent conceptualization of absorptive capacity

Conceptualization of absorptive capacity as…	*Description of the conceptualization*	*References*
Aggregated concept	● Absorptive capacity as independent variable, innovation performance as dependent variable ● Allows inquiring how variances of firm-level characteristics and context conditions of absorptive capacity affect innovation performance ● Business strategy as moderator (Van den Bosch et al., 1999) ● Company's position in the knowledge network as moderator (Tsai, 2001)	Van den Bosch et al. (1999), Tsai (2001)
Learning process	● Absorptive capacity evolves through learning processes ● Typical phases of the learning process are knowledge recognition and acquisition, knowledge assimilation, knowledge transformation, knowledge exploitation ● Prior knowledge as base for learning processes ● Refinement of the absorptive capacity concept by distinguishing various types of knowledge that a company aims to absorb and process: e.g., diverse vs distant knowledge (Vasudeva and Anand, 2011)	Zahra and George (2002), Lane et al. (2006), Jansen et al. (2005), Todorova and Durisin (2007), Vasudeva and Anand (2011)
Learning process configured through dynamic capabilities	● Absorptive capacity evolves through learning processes with firms requiring capabilities (dynamic capabilities) to build, configure, and reconfigure these learning processes ● Dynamic capabilities were initially understood as strategic context conditions driving and influencing the absorptive capacity process (e.g., Lane et al., 2006) ● Recently, dynamic capabilities have been understood as endogenous part of the absorptive capacity concept as interrelated learning processes conceptualized as combinative capabilities (e.g., Gebauer et al., 2012) and Metaroutines (e.g., Lewin et al., 2011)	Lane et al. (2006), Lewin et al. (2011), Gebauer et al. (2012)

Integrating Dynamic Capabilities into the Conceptualization of Absorptive Capacity

One way to bring the two streams together is to conceptualize the capabilities required in various process phases of absorptive capacity as operational capabilities and interpret combinative capabilities as dynamic capabilities. More specifically, the absorptive capacity's learning processes interact with combinative capabilities, which describe how a company systematizes, socializes, and coordinates knowledge (Zollo and Winter, 2002). Systemizing, coordinating, and socializing knowledge can either contribute to or hinder learning processes at the corresponding level of absorptive capacity (Van den Bosch et al., 1999). Coordinating knowledge refers to cross-functional interfaces and participation in decision processes. Knowledge can be systematized by the formalization and routinization of organizational actions. The socialization of knowledge is based on the density of social linkages (structural aspects) and shared social experience (cognitive aspects) in an organization and between the organization and its external partners (e.g., customers, suppliers) (Van den Bosch et al., 1999; Jansen et al., 2005).

This argumentation indicates that increasing the level of external knowledge does not always enhance innovation. More important is how combinative capabilities interact with learning processes of the absorptive capacity process. The key point is that creating an in-depth understanding of the interaction between learning processes and combinative capabilities could also explain why, in similar business environments, some companies achieve greater competitive advantages than others, through converting external knowledge into strategic innovations. That is because companies may differ in the dynamic capabilities they possess to reconfiguring the learning and knowledge transfer in the absorptive capacity process.

Figure 3 presents the new conceptualization. Against this background, past experiences set the reference points for developing absorptive capacity. The degree of strategic innovation is the dependent variable. Learning processes (exploratory, assimilative, transformative, and exploitative learning processes) driving absorptive capacity form the independent variable. Combinative capabilities (systematization, coordination, and socialization) mediate the relationship between learning processes and innovation outcomes, whereas strategy and network position are considered as moderating the evolution of learning processes and combinative capabilities.

Implications and Outlook

The integrated view presented in the previous section has a number of implications. It allows us to address some new aspects, provides an outlook on emerging topics, and lays out some directions for future research. This new approach enriches the theoretical knowledge on relationships between absorptive capacity, learning processes, combinative capabilities, and (strategic) innovations. The general contribution is twofold.

Firstly, using strategic innovation as the dependent variable, the absorptive capacity conceptualization includes the interactions of learning processes and combinative capabilities and as such forms the independent variable. Therefore, this conceptualization departs from the existing preoccupation

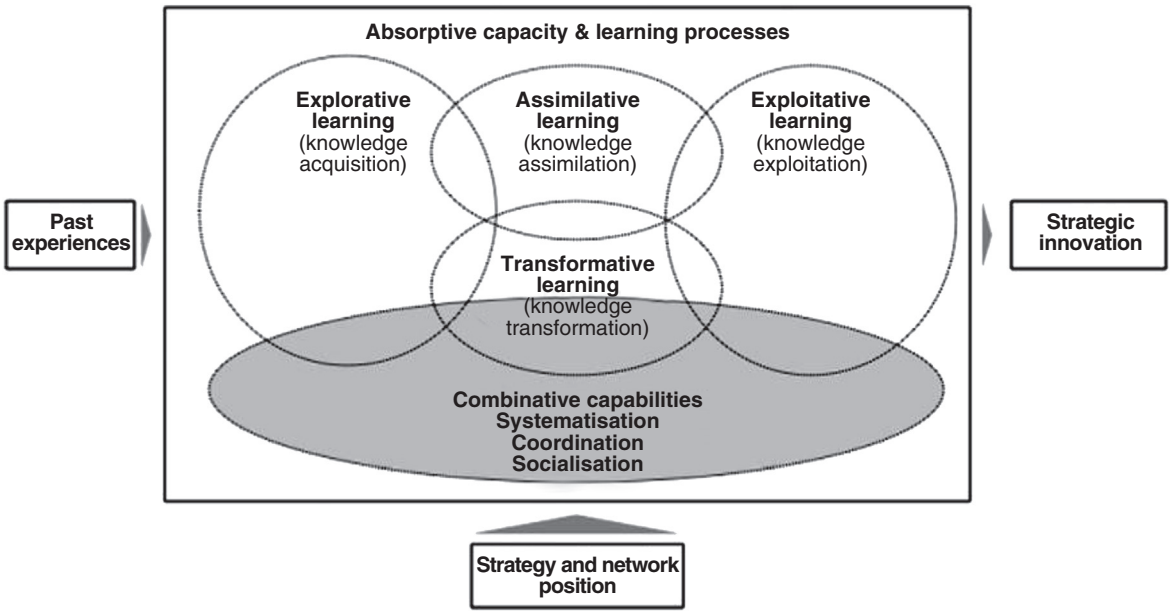

Figure 3 Dynamic capabilities, absorptive capacity, and learning processes: a new conceptualization (Gebauer, H., Worch, H., Truffer, B., 2012. Absorptive capacity, learning processes and combinative capabilities as determinants of strategic innovation. European Management Journal 30 (1), 57–73). Adapted from Lane, P.J., Koka, B.R., Pathak, S., 2006. The reification of absorptive capacity: a critical review and rejuvenation of the construct. Academy of Management Review 31 (4), 833–863; Zahra, S.A.S., George, G.G., 2002. Absorptive capacity: a review, conceptualization, and extension. Academy of Management Review 27 (2), 185–203; Zollo, M., Winter, S.G., 2002. Deliberate learning and the evolution of dynamic capabilities. Organization Science 13 (3), 339–351.

with absorptive capacity having direct innovation outcomes. This insight is in the tradition of Lane et al. (2006), but includes dynamic capabilities as independent variable rather than solely as a context condition.

Secondly, to explain how the necessary knowledge for strategic innovation is absorbed, our approach provides an understanding of how learning processes and combinative capabilities contribute to strategic innovation. The conceptualization suggests that assumptions on strategic behavior, network position, and the sequence of learning processes need to be reconsidered. This new view has implications and allows us to reconsider (and reinterpret) some findings in the absorptive capacity literature.

For example, one implication is that the argument that first-mover strategies have advantages for a company's absorptive capacity cannot easily be transferred as a proposition for strategic innovation (Van den Bosch et al., 1999). Recent findings suggest that follower strategies seem beneficial, where first-mover strategies would generate counter-productive side effects (Gebauer et al., 2012, 2014). The reasons for such side effects are constraints on a firm's combinative capabilities, which are necessary for taking advantage of the exploratory learning processes. First-mover strategies seem to hinder the departure from a strong formalization of knowledge sharing. This, in turn, constrains the development of diverse combinative capabilities such as a broader range of problem-solving skills, a higher cognitive diversity across the management team, and moving away from vertically centralized decision-making authority. Therefore, the new approach suggests that a first-mover strategy can also hinder strategic innovations, whereas a follower strategy could enhance strategic innovation due to a more broadly underpinned – though time intensive – innovation process. Further theory-building and empirical research should elaborate how a first-mover strategy can form rigidities, which constrain strategic innovations.

Similar to the first-mover strategy, Tsai (2001) argues that a more central network position strengthens the relationship between absorptive capacity and innovation performance. Yet recent findings indicate that a central network position and strong ties with network partners tend – under certain circumstances – to constrain the knowledge creation process (Gebauer et al., 2012, 2014). This is because centrality in a network may be important to shape the interpretations and the use of knowledge, but less in perceiving new perspectives that would allow a firm to have new interpretations of its existing knowledge. Therefore, the new approach summarized in **Figure 3** suggests that increasing centrality in the network position tends to constrain strategic innovation. Further theory-building and empirical research should elaborate how centrality in the network position leads to strategic reference points, which limit the knowledge creation process and ultimately innovation performance.

The new approach also sheds light on the discussion of the sequence of the acquisition, assimilation, transformation, and exploitation of knowledge (Jansen et al., 2005; Todorova and Durisin, 2007; Zahra and George, 2002). Assimilation and transformation are not sequential, but also not necessarily parallel. As recent findings show (Gebauer et al., 2012, 2014), if the exploratory learning processes rely predominantly on previous combinative capabilities, they will not only lead to some ideas that are relatively close to the existing knowledge base, but will also assimilate, rather than transform the knowledge. Under such conditions, exploratory learning processes are preceded by an assimilation of the newly acquired knowledge. Assimilation continues with previous combinative capabilities. Other reconfigurations of combinative capabilities within the exploratory learning process may enable transformation, in which the newly acquired knowledge interacts with changes in combinative capabilities. Therefore, this conceptualization suggests that transformative learning plays a key role in strategic innovation and contributes significantly to strategic innovation.

Finally, departing from existing configurations in combinative capabilities suggests that absorptive capacity also involves the unlearning of capabilities. Adding to Todorova and Durisin's (2007: p. 777) argument that "…firms often fail to identify and absorb valuable new external knowledge, because they are hampered by their embedded knowledge base, rigid capabilities, and path dependent managerial cognition…", rigidities exist specifically in how knowledge is systematized, coordinated, and socialized. We propose that firms reconfigure their combinative capabilities. For example, exploratory and transformative learning processes benefit from decreased formalization and more interdisciplinary routines for knowledge systematization, cross-functional interfaces, job rotation, and an umbrella strategy for knowledge coordination, as well as increased cognitive diversity and more dense social linkages. While these reconfigurations essentially depart from past experiences, relying on initial systematization capabilities would promote exploitative learning processes.

The results support the argument that the socialization of knowledge influences all three learning processes (Jansen et al., 2005), and not only the impact of potential on realized absorptive capacity, as proposed by Zahra and George (2002). This line of argumentation is not restricted to the socialization of knowledge, but also includes its systematization and coordination. Furthermore, the findings support the feedback process perspective on absorptive capacity (Lane et al., 2006; Todorova and Durisin, 2007), and depart from the linear approach to absorptive capacity (Jansen et al., 2005; Zahra and George, 2002). Thus, the new approach suggests that combinative capabilities enhance exploratory and transformative learning. Further theory-building and empirical research should elaborate how combinative capabilities facilitate these learning processes and specifically enhance exploratory and transformative learning.

See also: Business Models; Business Platforms; Development: Organizational; Innovation; Learning: Organizational; Modularity and Organizations; Network Paradigm: Applications in Organizational Science; Organizational Emergence and Firm Formation; Strategic Management; Strategizing.

Bibliography

Boschma, R.A., ter Wal, A.L.J., 2007. Knowledge networks and innovative performance in an industrial district: the case of a footwear district in the south of Italy. Industry & Innovation 14 (2), 177–199.

Camisón, C., Forés, B., 2010. Knowledge absorptive capacity: new insights for its conceptualization and measurement. Journal of Business Research 63 (7), 707–715.

Christensen, C.M., Johnson, M.W., Rigby, D.K., 2002. Foundations for growth: how to identify and build disruptive new businesses. MIT Sloan Management Review, 22–31.

Cohen, W.M., Levinthal, D.A., 1989. Innovation and learning: the two faces of R&D. Economic Journal 99 (397), 569–596.

Cohen, W.M., Levinthal, D.A., 1990. Absorptive capacity: a new perspective on learning and innovation. Administrative Science Quarterly 35 (1), 128–152.

Culnan, M.J., 1986. The intellectual development of management information systems, 1972–1982: a co-citation analysis. Management Science 32 (2), 156–172.

Flatten, T.C., Engelen, A., Zahra, S.A., Brettel, M., 2011. A measure of absorptive capacity: scale development and validation. European Management Journal 29 (2), 98–116.

Gebauer, H., Worch, H., Truffer, B., 2012. Absorptive capacity, learning processes and combinative capabilities as determinants of strategic innovation. European Management Journal 30 (1), 57–73.

Gebauer, H., Worch, H., Truffer, B., 2014. Value innovations in electricity utilities. In: Fuglsang, L., Rønning, R., Enquist, B. (Eds.), Framing Innovation in Public Service Sectors. Routledge, New York and London, pp. 85–111.

Grünfeld, L.A., 2003. Meet me halfway but don't rush: absorptive capacity and strategic R&D investment revisited. International Journal of Industrial Organization 21 (8), 1091–1109.

Jansen, J.J.P., Van den Bosch, F.A.J., Volberda, H.W., 2005. Managing potential and realized absorptive capacity: how do organizational antecedents matter? Academy of Management Journal 48 (6), 999–1015.

Kogut, B., Zander, U., 1992. Knowledge of the firm, combinative capabilities, and the replication of technology. Organization Science 3 (3), 383–397.

Lane, P.J., Koka, B.R., Pathak, S., 2006. The reification of absorptive capacity: a critical review and rejuvenation of the construct. Academy of Management Review 31 (4), 833–863.

Lane, P., Lubatkin, M., 1998. Relative absorptive capacity and interorganizational learning. Strategic Management Journal 19 (5), 461–477.

Lewin, A.Y., Massini, S., Peeters, C., 2011. Microfoundations of internal and external absorptive capacity routines. Organization Science 22, 81–98.

Lichtenthaler, U., Lichtenthaler, E., 2009. A capability-based framework for open innovation: complementing absorptive capacity. Journal of Management Studies 46, 1315–1338.

Minbaeva, D., Pedersen, T., Björkman, I., Fey, C.F., Park, H.J., 2003. MNC knowledge transfer, subsidiary absorptive capacity, and HRM. Journal of International Business Studies 34 (6), 586–599.

Schildt, H., Keil, T., Maula, M., February 2012. The temporal effects of relative and firm-level absorptive capacity on interorganizational learning. Strategic Management Journal 1173, 1154–1173.

Shenkar, O., Li, J., 1999. Knowledge search in international cooperative ventures. Organization Science 10 (2), 134–143.

Stock, G.N., Greis, N.P., Fischer, W.A., 2001. Absorptive capacity and new product development. The Journal of High Technology Management Research 12 (1), 77–91.

Szulanski, G., 1996. Exploring internal stickiness: Impediments to the transfer of best practice within the firm. Strategic Management Journal 17, 27–43. Special Issue: Knowledge and the Firm.

Teece, D.J., Pisano, G., Shuen, A., 1997. Dynamic capabilities and strategic management. Strategic Management Journal 18 (7), 509–533.

Todorova, G., Durisin, B., 2007. Absorptive Capacity: Valuing a Reconceptionalization 32 (3), 774–786.

Tsai, W., 2001. Knowledge transfer in interorganizational networks: effects of network position and absorptive capacity on business unit innovation and performance. Academy of Management Journal 44 (5), 996–1004.

Van den Bosch, F.A.J., Volberda, H.W., de Boer, M., 1999. Coevolution of firm absorptive capacity and knowledge environment: organizational forms and combinative capabilities. Organization Science 44, 551–568.

Vasudeva, G., Anand, J., 2011. Unpacking absorptive capacity: a study of knowledge utilization from alliance portfolios. Academy of Management Journal 54 (3), 611–623.

Winter, S.G., 2000. The satisficing principle in capability learning. Strategic Management Journal 21, 981–996.

Zahra, S.A.S., George, G.G., 2002. Absorptive capacity: a review, conceptualization, and extension. Academy of Management Review 27 (2), 185–203.

Zahra, S.A., Sapienza, H.J., Davidsson, P., 2006. Entrepreneurship and dynamic capabilities: a review, model and research agenda. Journal of Management Studies 43 (3), 917–955.

Zollo, M., Winter, S.G., 2002. Deliberate learning and the evolution of dynamic capabilities. Organization Science 13 (3), 339–351.

Academic Achievement Motivation, Development of

Jacquelynne S Eccles, University of Michigan, Ann Arbor, MI, USA
Allan Wigfield, University of Maryland, Baltimore, MD, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 1, pp. 14–20, © 2001, Elsevier Ltd.

Abstract

What explains individual differences in academic achievement motivation? This article outlines the answer to question in terms of three basic questions students ask themselves: Can I succeed? Do I want to do this task? And, Why am I doing this task? To the extent that individuals have positive answers to each of these questions, they will be motivated to achieve. The developmental origins of individuals' answers to these questions are summarized. In addition, specific achievement motivation pathologies (e.g., test anxiety and learned helplessness) are discussed.

Over the years, psychologists have proposed many different components of academic motivation (see Weiner, 1992 for full discussion of history of this field). Historically, this work began with efforts to understand and formalize the role of the basic need of achievement for human drive, the introduction of the idea of competence motivation, and early work on expectancies and social learning. Developmentalists such as Vaugh and Virginia Crandall, Battle, and Heckhausen translated these ideas into a developmental framework for studying the origins of individual differences in achievement motivation (e.g., Battle, 1966; Crandall, 1969; Crandall et al., 1962; Heckhausen, 1968). Sarason and his colleagues elaborated the concept test anxiety, developed measures, and outlined a developmental theory to explain the origins of individual differences in this critical component of academic achievement motivation (e.g., Sarason et al., 1960; Hill and Sarason, 1966).

Through this early period, the focus was on achievement motivation as a drive and need. With the cognitive revolution of the 1960s, researchers shifted to a much more cognitive view of motivation. Largely through the work of Weiner, attribution theory became the central organizing framework (see Weiner, 1992). This article falls in this cognitive tradition.

Eccles et al. (1998) suggested that one could group these various components under three basic questions: Can I succeed at this task? Do I want to do this task? Why am I doing this task? Children who develop positive and/or productive answers to these questions are likely to engage their school work and to thrive in their school settings more than children who develop less positive and/or noneffectual answers.

Can I Succeed?

Eccles and her colleagues' expectancy – value model of achievement-related choices and engagement, (see Eccles et al., 1998) is depicted in **Figure 1**. Expectancies and values are assumed to directly influence performance, persistence, and task choice. Expectancies and values are assumed to be influenced by task-specific beliefs such as perceptions of competence, perceptions of the difficulty of different task, and individuals' goals and self-schema. These social cognitive variables, in turn, are influenced by individuals' perceptions of other peoples' attitudes and expectations for them, by their own interpretations of their previous achievement outcomes, and by their affective memories of, or affective expectations about, similar tasks. Individuals' task-perceptions and interpretations of their past outcomes are assumed to be influenced by socializer's behavior and beliefs, by their own histories of success and failure, and by cultural milieu and unique historical events.

Bandura (1997) proposed a social cognitive model of motivated behavior that also emphasizes the role of perceptions of efficacy and human agency in determining individuals' achievement strivings. He defined self-efficacy as individuals' confidence in their ability to organize and execute a given course of action to solve a problem or accomplish a task. Bandura proposed that individuals' efficacy expectations (also called perceived self-efficacy) are determined by: previous performance (people who succeed will develop a stronger sense of personal efficacy than those who do not); vicarious learning (watching a model succeed on a task will improve one's own self-efficacy regarding the task); verbal encouragement by others, and the level of one's physiological reaction to a task or situation.

Bandura (1997) proposed specific development precursors of self-efficacy. First, through experiences controlling immediate situations and activities, infants learn that they can influence and control their environments. If adults do not provide infants with these experiences, they are not likely to develop as strong a sense of personal agency. Second, because self-efficacy requires the understanding that the self produced an action and an outcome, Bandura argued that a more mature sense of self-efficacy should not emerge until children have at least a rudimentary self-concept and can recognize that they are distinct individuals – which happens sometime during the second year of life. Through the preschool period, children are exposed to extensive performance information that should be crucial to their emerging sense of self-efficacy. However, just how useful such information is likely depends on the child's ability to integrate it across time, contexts, and domains. Since these cognitive capacities emerge gradually over the preschool and early elementary school years, young children's efficacy judgments should depend more on immediate and apparent outcomes than on a systematic analysis of their performance history in similar situations.

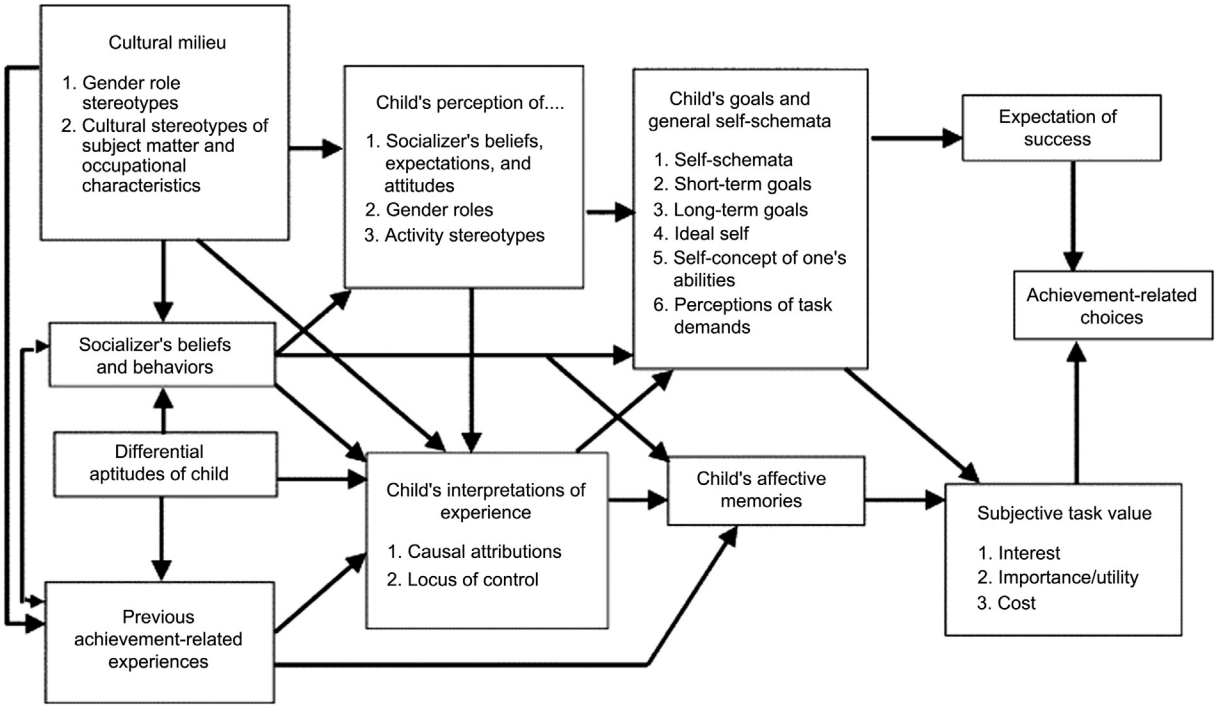

Figure 1 Model of Achievement Goals.

The Development of Competence-related/ Efficacy Beliefs

Changes in Children's Understanding of Competence-related Beliefs

Nicholls asked children questions about ability, intelligence, effort, and task difficulty, and how different levels of performance can occur when children exert similar effort (e.g., Nicholls, 1990). He found four relatively distinct levels of reasoning: Level One (ages 5 to 6) – effort, ability, and performance are not clearly differentiated in terms of cause and effect; Level Two (ages 7 to 9) – effort is seen as the primary cause of performance outcomes; Level Three (ages 9 to 12) – children begin to differentiate ability and effort as causes of outcomes; Level Four – adolescents clearly differentiate ability and effort. They understand the notion of ability as capacity and believe that ability can limit the effects of additional effort on performance, that ability and effort are often related to each other in a compensatory manner, and, consequently, that a successful outcome that required a great deal of effort likely reflects limited ability.

Change in the Mean Level of Children's Competence-related Beliefs

Children's competence-related beliefs decline across the school years (see Eccles et al., 1998). To illustrate, in Nicholls (1979) most first graders (6 years old) ranked themselves near the top of the class in reading ability, and there was essentially no correlation between their ability ratings and their performance level. In contrast, the 12-year-olds' ratings were more dispersed, and their correlation with school grades was 70 or higher. Expectancies for success also decrease during the elementary and secondary school years. In most laboratory-type studies, 4- and 5- year old children expect to do quite well on a specific task, even after repeatedly failing (Parsons and Ruble, 1977). Across the elementary school years, the mean levels of children's expectancies for success both decline and become more sensitive to both success and failure experiences.

These studies suggest that most children begin elementary school with quite optimistic ability-related self-perceptions and expectations, and that these beliefs decline rather dramatically as the children get older. In part this drop reflects the initially high, and often unrealistic, expectations of kindergarten and first- grade children. Other changes also contribute to this decline – changes such as increased exposure to failure feedback, increased ability to integrate success and failure information across time to form expectations more closely linked with experience, increased ability to use social comparison information, and increased exposure to teachers' expectations.

Some of these changes are also linked to the transition into elementary school. Entrance into elementary school and then the transition from kindergarten to first grade introduces several systematic changes in children's social worlds. First, classes are age stratified, making within-age-ability social comparison much easier. Second, formal evaluations of competence by 'experts' begin. Third, formal ability grouping begins usually with reading group assignment. Fourth, peers have the opportunity to play a much more constant and salient role in children's lives. Each of these changes should impact children's motivation. Parents' expectations for, and perceptions of, their children's academic competence are also influenced by report card marks and standardized test scores given out during the early elementary school years, particularly for mathematics (Alexander and Entwisle, 1988).

There are significant long-term consequences of children's experiences in the first grade, particularly experiences associated with ability grouping and within class differential teacher treatment. For example, teachers use a variety of information to assign first graders to reading groups including temperamental characteristics like interest and persistence, race, gender, and social class. Alexander et al. (1993) demonstrated that differences in first-grade reading group placement and teacher-student interactions have a significant effect (after controlling for initial individual differences in competence) on motivation and achievement several years later. Furthermore, these effects are mediated by both differential instruction and the impact of ability-group placement on parents' and teachers' views of the children's abilities, talents, and motivation (Pallas et al., 1994).

Theories Concerned With the Question 'Do I Want to Do This Task?'

Subjective Task Values

Eccles et al. (1983) outlined four motivational components of subjective task value: attainment value, intrinsic value, utility value, and cost. Attainment value is the personal importance of doing well on the task. Intrinsic value is the enjoyment the individual gets from performing the activity, or the subjective interest the individual has in the subject. Utility value is how well a task relates to current and future goals, such as career goals. Finally, they conceptualized 'cost' in terms of the negative aspects of engaging in the task (e.g., performance anxiety and fear of both failure and success), as well as both the amount of effort that is needed to succeed and the lost opportunities resulting from making one choice rather than another.

Eccles and her colleagues have shown that ability self-concepts and performance expectancies predict performance in mathematics and English, whereas task values predict course plans and enrollment decisions in mathematics, physics, English, and involvement in sport activities even after controlling for prior performance levels (see Eccles et al., 1998). They have also shown that values predict career choices.

Development of Subjective Task Values

Eccles and their colleagues have documented that even young children distinguish between their competence beliefs and their task values. They have also shown that children's and adolescents' valuing of certain academic tasks and school subjects decline with age.

Although little developmental work has been done on this issue, it is likely that there are differences across age in which of the components of achievement values are most dominant motivators. Wigfield and Eccles (1992) suggested that interest is especially salient during the early elementary school grades. If so, then young children's choice of different activities may be most directly related to their interests. And if young children's interests shift as rapidly as their attention spans, it is likely they will try many different activities for a short time each before developing a more stable opinion regarding which activities they enjoy the most. As children get older the perceived utility and personal importance of different tasks likely become more salient, particularly as they develop more stable self-schema and long-range goals and plans.

A third important developmental question is how children's developing competence beliefs relate to their developing subjective task values? According to both the Eccles et al. model and Bandura's self-efficacy theory, ability self-concepts should influence the development of task values. Mac Iver et al. (1991) found that changes in junior high school (ages 11 – 13) students' competence beliefs over a semester predicted changes in children's interests much more strongly than vice versa. Does the same causal ordering occur in younger children? Wigfield (1994) proposed that young children's competence and task-value beliefs are likely to be relatively independent of each other. This independence would mean that children might pursue some activities in which they are interested regardless of how good or bad they think they are at the activity. Over time, particularly in the achievement domain, children may begin to attach more value to activities on which they do well, for several reasons: first, through process associated with classical conditioning, the positive affect one experiences when one does well should become attached to the activities yielding success. Second, lowering the value one attaches to activities that one is having difficulty with is likely to be an effective way to maintain a positive global source of efficacy and self-esteem. Thus, at some point the two kinds of beliefs should become more positively related to one another.

Interest Theories

Closely related to the intrinsic interest component of subjective task value is the work on 'interest' (Renninger et al., 1992). Researchers in this tradition differentiate between individual and situational interest. Individual interest is a relatively stable evaluative orientation towards certain domains; situational interest is an emotional state aroused by specific features of an activity or a task. The research on individual interest has focused on its relation to the quality of learning. In general, there are significant but moderate relations between interest and text learning. More importantly, interest is more strongly and positively related to indicators of deep-level learning (e.g., recall of main ideas, coherence of recall, responding to deeper comprehension questions, representation of meaning) than to surface-level learning (e.g., responding to simple questions, verbatim representation of text). The research on situational interest has focused on the characteristics of academic tasks that create interest. Among others, the following text features arouse situational interest: personal relevance, novelty, and comprehensibility.

Developmental Changes in Interest

Several researchers have found that individual interest in different subject areas at school declines continuously during the school years. This is especially true for the natural sciences. These researchers have identified changes in the following instructional variables as contributing to these declines: clarity of presentation, monitoring of what happens in the classroom, supportive behavior, cognitively stimulating experiences, self-concept of the teacher [educator vs scientist], and achievement pressure.

Intrinsic Motivation Theories

Over the last 25 years, studies have documented the debilitating effects of extrinsic incentives on the motivation to perform even inherently interesting activities (Deci and Ryan 1985). This has stimulated interest in intrinsic motivation. Deci and Ryan (1985) argue that intrinsic motivation is maintained only when actors feels competent and self-determined. Deci and Ryan (1985) also argue that the basic needs for competence and self-determination play a role in more extrinsically motivated behavior. Consider, for example, a student who consciously and without any external pressure selects a specific major because it will help him earn a lot of money. This student is guided by his basic needs for competence and self-determination but his choice of major is based on reasons totally extrinsic to the major itself. Finally, Deci and Ryan postulate that a basic need for interpersonal relatedness explains why people turn external goals into internal goals through internalization.

Developmental Changes in Intrinsic Motivation

Like interest and subjective task value intrinsic motivation declines over the school years (see Eccles et al., 1998), particularly during the early adolescent years (which coincide in many countries with the transition into upper-level educational institutions). Such changes lead to decreased school engagement. The possible origins of these declines have not been studied but are likely to be similar to the causes of declines in expectations, ability-related self-confidence and interest – namely, shifts in the nature of instruction across grade levels, cumulative experiences of failure, and increasing cognitive sophistication.

Why Am I Doing This?

The newest area of motivation is goal theory. This work focuses on why the children think they are engaging in particular achievement-related activities and what they hope to accomplish through their engagement. Several different approaches to goal theory have emerged. For instance, Schunk (1991) focuses on goals' proximity, specificity, and level of challenge and has shown that specific, proximal, and somewhat challenging goals promote both self-efficacy and improved performance. Other researchers have defined and investigated broader goal orientations. Nicholls and his colleagues (Nicholls, 1990) defined two major kinds of motivationally relevant goal patterns or orientations: ego-involved goals and task-involved goals. Individuals with ego-involved goals seek to maximize favorable evaluations of their competence and minimize negative evaluations of competence. Questions like 'Will I look smart?' and 'Can I outperform others?' reflect ego-involved goals. In contrast, with task-involved goals, individuals focus on mastering tasks and increasing their competence. Questions such as 'How can I do this task?' and 'What will I learn?' reflect task-involved goals. Dweck and her colleagues provide a complementary analysis distinguishing between performance goals (like ego-involved goals), and learning goals (like task-involved goals) (Dweck and Leggett, 1988). Similarly, Ames (1992) distinguishes between the association of performance (like ego-involved) goals and mastery goals (like task-focused goals) with both performance and task choice. With ego-involved (or performance) goals, children try to outperform others, and are more likely to do tasks they know they can do. Task-involved (or mastery-oriented) children choose challenging tasks and are more concerned with their own progress than with outperforming others.

Development of Children's Goals

To date there has been surprisingly little empirical work on how children's goals develop. Nicholls (1990) documented that both task goals and ego goals are already developed by second graders. However, Nicholls also suggested that the ego-goal orientation becomes more prominent for many children as they get older, in part because of developmental changes in their conceptions of ability and, in part, because of systematic changes in school context. Dweck and her colleagues (Dweck and Leggett, 1988) also predicted that performance goals should get more prominent as children go through school, because they develop a more 'entity' view of intelligence as they get older and children holding an entity view of intelligence are more likely to adopt performance goals.

It is also likely that the relation of goals to performance changes with age due to the changing meaning of ability and effort. In a series of studies looking at how competitive and noncompetitive conditions, and task and ego-focused conditions, influence pre- and elementary-school-aged children's interests, motivation, and self-evaluations, Butler (e.g., 1990) identified several developmental changes. First, competition decreased children's subsequent interest in a task only among children who had also developed a social-comparative sense of ability. Competition also increased older, but not younger, children's tendency to engage in social comparison. Second, although children of all ages engaged in social comparison, younger children seemed to be doing so more for task mastery reasons, whereas older children did so to assess their abilities. Third, whereas, 5, 7, and 10 year-old children's self-evaluations were quite accurate under mastery conditions, under competitive conditions 5- and 7-year-olds inflated their performance self-evaluations more than 10-year-olds.

The Development of Motivational Problems

Test Anxiety

Performance anxiety has been an important topic in motivational research from early on. In one of the first longitudinal studies, Hill and Sarason (1966) found that test anxiety both increases across the elementary and junior high school years and becomes more negatively related to subsequent grades and test scores. They also found that highly anxious children's achievement test scores were up to 2 years behind those of their low anxious peers and that girls' anxiety scores were higher than boys'. Finally, they found that test anxiety was a serious problem for many children.

High anxiety emerges when parents have overly high expectations and put too much pressure on their children (Wigfield and Eccles, 1989). Anxiety continues to develop in school as children face more frequent evaluation, social comparison, and (for some) experiences of failure; to the extent that schools emphasize these characteristics, anxiety become a problem for more children as they get older.

Anxiety Intervention Programs

Earlier intervention programs emphasized the emotionality aspect of anxiety and focused on various relaxation and desensitization techniques. Although these programs did succeed in reducing anxiety, they did not always lead to improved performance, and the studies had serious methodological flaws. Anxiety intervention programs linked to the worry aspect of anxiety focus on changing the negative, self-deprecating thoughts of anxious individuals and replacing them with more positive, task-focused thoughts. These programs have been more successful both in lowering anxiety and improving performance.

Learned Helplessness

Dweck and her colleagues initiated an extensive field of research on academic learned helplessness. They defined learned helplessness 'as a state when an individual perceives the termination of failure to be independent of his responses' (Dweck and Goetz, 1978: p. 157). They documented several differences between helpless and more mastery-oriented children's responses to failure. When confronted by difficulty (or failure), mastery-oriented children persist, stay focused on the task, and sometimes even use more sophisticated strategies. In contrast, helpless children's performance deteriorates, they ruminate about their difficulties, often begin to attribute their failures to lack of ability. Further, helpless children adopt an 'entity' view that their intelligence is fixed, whereas mastery-oriented children adopt an incremental view of intelligence.

In one of the few developmental studies of learned helpless behavior, Rholes et al. (1980) found that younger children did not show the same decrements in performance in response to failure as some older children do. However, Dweck and her colleagues' recent work (Burhans and Dweck, 1995) suggests that some young (5- and 6-year-old) children respond quite negatively to failure feedback, judging themselves to be bad people. These rather troubling findings show that negative responses to failure can develop quite early on.

What produces learned helplessness in children? Dweck and Goetz (1978) proposed that it depends on the kinds of feedback children receive from parents and teachers about their achievement outcomes, in particular whether children receive feedback that their failures are due to lack of ability. In Hokoda and Fincham (1995), mothers of helpless third-grade children (in comparison to mothers of mastery-oriented children) gave fewer positive affective comments to their children, were more likely to respond to their children's lack of confidence in their ability by telling them to quit, were less responsive to their children's bids for help, and did not focus them on mastery goals.

Alleviating Learned Helplessness

There are numerous studies designed to alleviate learned helplessness by changing attributions for success and failure so that learned helpless people learn to attribute failure to lack of effort rather than to lack of ability (see Fosterling, 1985). Various training techniques (including operant conditioning and providing specific attributional feedback) have been used successfully in changing children's failure attributions from lack of ability to lack of effort, improving their task persistence, and performance.

Self-efficacy training can also alleviate learned helplessness. Schunk and his colleagues (Schunk, 1994) have studied how to improve low-achieving children's academic performance through skill training, enhancement of self-efficacy, attribution retraining, and training children how to set goals. A number of findings have emerged from this work. First, the training increases both children's performance and their sense of self-efficacy. Second, attributing children's success to ability has a stronger impact on their self-efficacy than does either effort feedback, or ability and effort feedback. Third, training children to set proximal, specific, and somewhat challenging goals enhances their self-efficacy and performance. Fourth, training that emphasizes process goals (analogous to task or learning goals) increases self-efficacy and skills. Finally, combining strategy training, goal emphases, and feedback to show children how various strategies relate to their performance has a strong effect on subsequent self-efficacy and skill development.

Summary

In this article, a basic model of achievement motivation was presented and discussed. Developmental origins of individual differences in students' confidence in their ability to succeed, their desire to succeed, and their goals for achievement were summarized. To a large extent individual differences in achievement motivation are accounted for by these three beliefs. Most importantly, lack of confidence in one's ability to succeed and extrinsic (rather than intrinsic) motivation are directly related to the two major motivational problems in the academic achievement domain: test anxiety and learned helplessness. Specific interventions for these two motivational problems were discussed. Future research needs to focus on interconnections among the various aspects of achievement motivation. For example, how is confidence in one's ability to master academic tasks related to individuals' desire to master these tasks and to the extent to which the individual is intrinsically motivated to work towards mastery? More work is also needed on the impact of families, schools, and peers on the development of confidence, interest, and intrinsic motivation. Exactly how can parents and teachers support the development of high interest and high intrinsic motivation to work hard to master academic tasks? Finally, we need to know a lot more about the motivational factors that underlie ethnic and gender group differences in academic achievement patterns.

See also: Academic Self-Concept and Achievement; Motivation and Actions, Psychology of; Motivation, Learning, and

Instruction; Motivation: History of the Concept; Personality and Educational Outcomes; School Achievement: Motivational Determinants and Processes; Test Anxiety and Academic Achievement.

Bibliography

Alexander, K.L., Entwisle, D., 1988. Achievement in the first two years of school: Patterns and processes. Monographs of the Society for Research in Child Development 53 (2, Serial No. 218).

Alexander, K.L., Dauber, S.L., Entwisle, D.R., 1993. First-grade classroom behavior: Its short- and long-term consequences for school performance. Child Development 64, 801–803.

Ames, C., 1992. Classrooms: Goals, structures, and student motivation. Journal of Educational Psychology 84, 261–271.

Battle, E., 1966. Motivational determinants of academic competence. Journal of Personality and Social Psychology 4, 534–642.

Bandura, A., 1997. Self-efficacy: The Exercise of Control. Freeman, New York.

Burhans, K.K., Dweck, C.S., 1995. Helplessness in early childhood: The role of contingent worth. Child Development 66, 1719–1738.

Butler, R., 1990. The effects of mastery and competitive conditions on self-assessment at different ages. Child Development 61, 201–210.

Crandall, V.C., 1969. Sex differences in expectancy of intellectual and academic reinforcement. In: Smith, C.P. (Ed.), Achievement-related Motives in Children. Russell Sage Foundation, New York, pp. 11–74.

Crandall, V.J., Katkovsky, W., Preston, A., 1962. Motivational and ability determinants of young children's intellectual achievement behavior. Child Development 33, 643–661.

Deci, E.L., Ryan, R.M., 1985. Intrinsic Motivation and Self-Determination In Human Behavior. Plenum Press, New York.

Dweck, C.S., Goetz, T.E., 1978. Attributions and learned helplessness. In: Harvey, J.H., Ickes, W., Kidd, R.F. (Eds.), New Directions in Attribution Research, vol. 2. Erlbaum, Hillsdale, NJ.

Dweck, C.S., Leggett, E., 1988. A social-cognitive approach to motivation and personality. Psychological Review 95, 256–273.

Eccles, J.S., Wigfield, A., Schiefele, N., 1998. Motivation to succeed. In: Eisenberg, N. (vol. ed.), Demon, U. (series ed.), Handbook of Child Psychology, fifth ed, vol. 3. Wiley, New York, pp. 1017–1095.

Eccles, P., Adler, T.F., Futterman, R., Goff, S.B., Kaczala, C.M., Meece, J.L., Midgley, C., 1983. Expectancies, values, and academic behaviors. In: Spence, J.T. (Ed.), Achievement and Achievement Motivation. H. Freeman, San Francisco, pp. 75–146.

Fosterling, F., 1985. Attributional retraining: A review. Psychological Bulletin 98, 495–512.

Heckhausen, H., 1968. Achievement motivation research: Current problems and some contributions towards a general theory of motivation. In: Arnold, W.J. (Ed.), Nebraska Symposium on Motivation. University of Nebraska Press, Lincoln, NE, pp. 103–174.

Hill, K.T., Sarason, S.B., 1966. The relation of test anxiety and defensiveness to test and school performance over the elementary school years: A further longitudinal study. Monographs for the Society for Research in Child Development 31 (2, Serial No. 104).

Hokoda, A., Fincham, F.D., 1995. Origins of children's helpless and mastery achievement patterns in the family. Journal of Educational Psychology 87, 375–385.

Mac Iver, D.J., Stipek, D.J., Daniels, D.H., 1991. Explaining within-semester changes in student effort in junior high school and senior high school courses. Journal of Educational Psychology 83, 201–211.

Nicholls, J.G., 1979. Development of perception of own attainment and causal attributions for success and failure in reading. Journal of Educational Psychology 29, 94–99.

Nicholls, J.G., 1990. What is ability and why are we mindful of it? A developmental perspective. In: Sternberg, R.J., Kolligian, J. (Eds.), Competence Considered. Yale University Press, New Haven, CT.

Pallas, A.M., Entwisle, D.R., Alexander, K.L., Stluka, M.F., 1994. Ability-group effects: Instructional, social, or institutional? Sociology of Education 67, 27–46.

Parsons, J.E., Ruble, D.N., 1977. The development of achievement-related expectancies. Child Development 48, 1075–1079.

Renninger, K.A., Hidi, S., Krapp, A. (Eds.), 1992. The Role Of Interest in Learning and Development. Erlbaum, Hillsdale, NJ.

Rholes, W.S., Blackwell, J., Jordan, C., Walters, C., 1980. A developmental study of learned helplessness. Developmental Psychology 16, 616–624.

Sarason, S.B., Davidson, K.S., Lighthall, F.F., Waite, R.R., Ruebush, B.K., 1960. Anxiety In Elementary School Children. Wiley, New York.

Schunk, D.H., 1991. Self-efficacy and academic motivation. Educational Psychologist 26, 207–231.

Schunk, D.H., 1994. Self-regulation of self-efficacy and attributions in academic settings. In: Schunk, D.H., Zimmerman, B.J. (Eds.), Self-Regulation of Learning and Performance. Erlbaum, Hillsdale, NJ.

Weiner, B., 1992. Human Motivation: Metaphors, Theories, and Research. Sage, Newbury Park, CA.

Wigfield, A., 1994. Expectancy-value theory of achievement motivation: A developmental perspective. Educational Psychology Review 6, 49–78.

Wigfield, A., Eccles, J.S., 1989. Test anxiety in elementary and secondary school students. Educational Psychologist 24, 159–183.

Wigfield, A., Eccles, J., 1992. The development of achievement task values: A theoretical analysis. Developmental Review 12, 265–310.

Academic Careers in Comparative Perspective

Hans Pechar, Institute of Science Communication and Higher Education Research, Alpen Adria Universität, Vienna, Austria
Lesley Andres, University of British Columbia, Vancouver, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, we focus, from a comparative perspective, on the traditional academic career within a university setting. We commence by describing the various facets relevant to academic careers, including academic training, employment contracts, hierarchies, and the case for and against tenure. We conclude with a discussion of current reforms and policy debates.

The term 'academic career' is slippery to define. Most commonly, it refers to employment at universities by those educated at the doctoral level with the "triumvirate of teaching, research, and service" (Schuster and Finkelstein, 2006) as central employment responsibilities. Academic-like careers can also be found in other venues, including research institutes (e.g., CNRS in France, Max Planck Institutes in Germany, Russian Academy of Sciences), various levels of government, and arm's-length or nongovernmental organizations (e.g., OECD, UNESCO). For the purposes of this article, we focus, from a comparative perspective, on the traditional academic career within a university setting. We commence by describing the various facets relevant to academic careers, including academic training, employment contracts, hierarchies, and the case for and against tenure. We conclude with a discussion of current reforms and policy debates.

Academic Training

An academic career is characterized by an extended training period. Increasingly, the PhD is the entrance requirement for permanent teaching positions, research positions, or both and is now standard at most research universities in OECD countries. However, there is significant variation among academic systems regarding how the training period is conceptualized and interconnected with paid employment.

The range of doctoral training models is defined primarily in terms of the nature of the relationship – and the related degree of dependence – between the student and her or his research supervisory committee and the degree to which programs are structured. Doctoral programs fall on a continuum ranging from almost complete independence (i.e., 'reading' for a degree) where the student works independently with minimal contact and input from a research supervisor and committee and without a prescribed set of courses, to those that are highly structured, in terms of programs of study, and close surveillance by the research supervisory committee (Kehm, 2006). Most doctoral students are required to produce a dissertation that is 'original' theoretically, conceptually, and/or empirically. However, the nature and extent to which the dissertation is assessed – that is, whether examiners external to the supervisory committee, department, and university – varies from program to program (Usher, 2002).

Also, the extent of rigor, in terms of research training, varies across institutions and countries (Nerad and Heggelund, 2008). Programs offering full-fledged research training strive to ensure that, upon graduations, graduates are prepared to embark on research careers, and hence are intended for students who want to pursue this trajectory. Other programs are designed to provide credentials (and the related title of 'Dr') for individuals who seek to use such credentials as a signal in the labor market. As an illustration, in some countries (e.g., Austria, Germany), as late as 40 years ago, the doctorate was the first academic degree earned by university students. Clearly, it served a very different purpose than doctorates earned in highly differentiated university systems (e.g., the United States, UK). For example, Egon Franck and Christian Opitz (2007) demonstrate that 58.5% of CEOs of major German companies have earned doctorates, whereas only 5.5% of CEOs in comparable companies in the United States possess the same credential. The emphasis on labor market signaling can be regarded as problematic because of the potential to dilute the rigor of research training. This concern was the major impetus for a major redesign of doctoral programs in many European countries, in particular Germany, by refocusing the emphasis on research training (European University Association, 2007). However, that universities are accustomed to training academics for the nonacademic labor market can also be regarded as an advantage at a time where there is mismatch between supply and demand within academic labor markets (Nature, 2011; Usher, 2002).

The extent of rigor in doctoral examination policies, procedures, and practices also differentiates types of doctoral training. Efforts exerted by institutions to examine doctoral candidates range from minimal (e.g., dependent on one person, usually the research supervisor) to extensive (e.g., multiple checks and balances including external review of the dissertation, policies preventing conflicts of interest between the research committee and external examiners).

Doctoral training also serves as the interface between education and employment, with variation in the precise design of this interface. In many European countries, doctoral students may be formally employed by the university (e.g., Sweden). It is rare that the majority of doctoral students have an employment contract; usually, the most promising students are employed as assistants to their research supervisors. Those employed as assistants are regarded as 'early stage researchers'

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03001-4

in most European countries as defined by the Bergen Communiqué (European Ministers Responsible for Higher Education, 2005; Williams Committee of Inquiry into Education and Training, 1979). This type of employment is fixed term in nature and thus does not usually hold the promise of a permanent appointment upon successful completion of a doctoral degree. Because of the lack of rigor in some doctoral programs, many European systems require a second thesis, the *Habilitation*. In countries with a *Habilitation* system, fixed-term employment may also be the case while the junior academic works on her or his *Habilitation*. As with the doctoral degree, successful completion of a *Habilitation* does not lead to a permanent position. Rather, it is the opposite; by completing the *Habilitation*, many junior academics reach the maximum years allowed in fixed-term contracts and are required to leave the university (see chain contracts below).

Until recently, doctoral training in North America has been regarded predominantly as part of the education of prospective academics. Many doctoral students are employed as research or teaching assistants. Although they may be considered to be 'early stage' researchers (Bazeley, 2003), they are graduate students (and may belong to student unions) and not employees. Doctoral training in the United States and Canada is more rigorous than in most European higher education systems in terms of length, due to mandatory course work and other requirements such as comprehensive examinations. A doctoral degree is the necessary and sufficient formal entry requirement to an academic career. In North America, a maximum length of time to completion is usually specified, whereas in Europe, a minimal time to completion tends to be stipulated.

Employment Contracts

Employment contracts can be either permanent or fixed term. Senior academics typically have permanent positions. However, the dividing line between junior and senior academics and the criteria that define senior academics vary among academic systems.

In many countries, senior academics have tenure which guarantees higher job protection than does a regular permanent position (Finkin, 1996; Horn, 1999b). Because the case for and against tenure is complex and controversial, we devote a separate section to this topic. Types of tenure vary. Academics in most European countries are civil servants who can be terminated only under special circumstances; hence, they are accorded high job protection. However, in most instances only those holding a Professorship and Chair benefit from such protection. Academic tenure in North America is different in that even in the public universities of the United States and Canada, academics are not civil servants but rather employees of the university. Academic tenure in North America is the outcome of a long 'probation' period followed by rigorous peer review. This review involves peer review both within one's university (usually at the departmental, faculty, and university-wide levels) and externally by seeking expert opinions from academics from other universities. (In some countries, tenure status is not awarded to academics. For example, in the UK, tenure was abolished and replaced with permanent and fixed-term contracts, see Court (1998).)

There are different reasons for fixed-term contracts. First, fixed-term employment is used as a probation mechanism, usually in the early stages of an academic career. Under these circumstances, employees typically can look forward to the prospect of permanent employment if they stand the 'tenure' test. One example for this type of probation is the assistant professor within the North American tenure track system. The assistant professorship is a fixed-term position that concludes in an 'up or out' decision. A positive evaluation is required to achieve a permanent and tenured position.

In many European higher education systems, the term 'assistant' has a different meaning (Busch, 1963). Assistants in the German-speaking higher education systems are usually not on probation; they have fixed-term training positions which allow them to complete either their doctoral degrees or their *Habilitation*. The crucial difference to the tenure track is that successful completion of the training phase does not guarantee a permanent position.

Although there are structural similarities between those undergoing either a period to complete the *Habilitation* or as a tenure track professor (e.g., a period of extension beyond the doctorate; undergoing another level of peer scrutiny through examination in the case of the former or assessment of the tenure file in the case of the latter; the right to teach independently (*venia legendi*)), there are several key differences. In the *Habilitation* system, the training period, called the 'qualification agreement,' does not guarantee employment. The 'assistant' literally assists her or his professor. Tenure track assistant professors do not require a training period. Although their research, teaching, and service skills improve with time and experience, they are expected to perform their duties independently, often with advice and support of senior colleagues, but not under their thumbs. Fixed-term contracts are also granted for special projects that have a beginning and an end. Researchers employed in such projects are usually hired on 'soft money'; that is, funds outside of the regular general operating funds of the university (e.g., a research grant). In order to increase competitiveness, some European countries have strongly amplified this type of research by shifting financial resources from general university funds to the granting agencies; this practice is criticized by Münch (2011). Hence, the peer review process ensures that only 'excellent' researchers are successful in these competitions. As a result, the number of fixed-term postdoctoral positions has escalated.

To what degree such types of employment carry the risk of precarious positions depends on a variety of factors. The element of time is crucial. For doctoral students, fixed-term employment in research projects provides welcome opportunities to gain practical experience in research. The same is the rule for postdoctoral fellows, granted that there is a realistic opportunity for them to embark on regular – that is, tenure track – career trajectories or other forms of permanent employment. However, in this type of employment arrangement, the later stage researcher is at much greater risk in terms of job security. Another critical factor is the legal possibility to continue in renewable fixed-term projects over the long term. In some countries, unions have been successful at legally banning a succession of fixed-term contracts, labeled 'chain

contracts,' that exceed a given number of years. The intention of the unions was to reduce risk for employees by forcing employers to convert fixed-term positions into permanent ones. However, contrary to what was intended, in many European universities systems, rather than strengthening the job security of researchers, this policy has made these positions more vulnerable. Universities either cannot afford or are not willing to give researchers hired to work on specific projects permanent positions. Nor are they legally allowed to renew contracts after a specified number of years. The term *Prekariat* in German and *précariat* in French, defined as a social group that suffers multiple forms of insecurity, has emerged and has triggered debate about the increase of such positions and their associated dangers.

Hierarchies

Academic systems are highly stratified with respect to reputation, salaries, and the working conditions of academics. In the past, small elite systems that had preceded massification of higher education were indeed stratified. However, subsequent expansion has increased the functional differentiation of academic work significantly, and as a result has increased the degree of stratification. During the late nineteenth and early twentieth centuries, most higher education systems adopted, in one way or another, the Humboldtian ideal of unity of teaching and research (although as Kezar (2013) rightly points out, historically the academic profession has been one of continuous change). Massification of higher education, commencing in the second half of the twentieth century, has partly reversed this trend by creating sectors and segments that are devoted primarily to teaching (e.g., the community college sector in North America, the former polytechnics in UK, and *Fachhochschulen* in Germany). The research function is the most important criterion for the stratification of reputation, income, and working conditions. Typically, research universities have higher reputations than teaching institutions. However, some countries have elite teaching sectors, such as the private liberal arts colleges in the United States or the *grand écoles* in France. The former are undergraduate institutions and the latter are professional specialized schools that constitute the elite segment of higher education and have a more esteemed reputation than research universities in France.

Mass higher education systems of today embody different kinds of functional differentiation and reputational stratification. As a rule, national systems that are strongly steered by governmental intervention are differentiated into different sectors. In such systems, reputation, salaries, and working conditions are usually quite different among sectors; however, within each sector, little stratification exists. In most European countries, public research universities are treated equally by governments, irrespective of informal reputational hierarchies (as expressed in rankings). As such, working conditions for academics in each of the sectors are fairly equitable. In countries with more market-oriented systems such as the UK or the United States, the research university sector is highly stratified in terms of, for example, reputation, wealth, faculty access to resources, working conditions, and composition of the student body. Stratification among universities in the United States has greatly intensified over the last 30 years (Geiger, 2010). This shift is evident by the emergence of 'elite clubs' such as the Russell Group in the UK and the 'Big Five' in Canada.

University systems that are divided into sectors have little institutional differentiation within each sector. However, within each institution, the hierarchy between the different academic status groups, between full professors and nonprofessorial staff, between senior and junior academics, is strongly pronounced. In other words, academia in such countries is divided in a national guild of senior academics (with little differentiation among institutions) and various types of nonprofessorial groups. Status differences appear within every institution, with working conditions for each status group fairly equal within each sector. Conversely, status hierarchies in market-oriented systems within each university are comparatively flat; however, status differences between different institutions are strongly pronounced. Depending on the perspective, these patterns in one system may be regarded as more egalitarian than another. Some observers regard European higher education systems as less hierarchical than US higher education because institutional stratification is much more flat. Others claim that the North American pattern is more egalitarian because the hierarchy between senior and junior academics is much less vertically stratified. In fact, hierarchies are strongly pronounced in both academic worlds, albeit in qualitatively different forms.

International data on academic salaries are rare and problematic in terms of comparability. The United States compiles extensive databases at the national level; however, the way that data are aggregated does not reflect substantive wage differentials. The Academic Career Observatory of the European University Institute in Florence (http://www.eui.eu/ProgrammesAndFellowships/AcademicCareersObservatory/CareerComparisons/SalaryComparisons.aspx) provides salary comparisons for selected countries. A study by the European Commission (2007) has attempted to adjust academic salaries to the cost of living of the respective countries. Also, salary progression by experience and gender is provided. An attempt to broaden comparisons beyond OECD countries has been conducted by Philip Altbach et al. (2012).

The Case for and against Tenure

The most common perspectives used to frame the discussion in support of tenure are either that of academic freedom or economics of tenure.

Academic Freedom Perspective

Historically, the hiring of professors was haphazard with little specificity in contracts regarding terms of the employment. Faculty had 'tenure during pleasure' which meant the pleasure of the university board and faculty members could be fired without cause. As early as 1915, the American Association of University Professors (AAUP) addressed the issue of tenure (Metzger, 1987: 167). Following a spate of dismissals and related legal challenges, in 1940 the AAUP produced a document entitled *Statement of Principles on Academic Freedom and Tenure* which stated,

> academic freedom is essential to these purposes and applies to both teaching and research. Freedom in research is fundamental to the advancement of truth. Academic freedom in its teaching aspect is fundamental for the protection of the rights of the teacher in teaching and of the student to freedom in learning. It carries with it duties correlative with rights. Tenure is a means to certain ends; specifically: (1) freedom of teaching and research and of extramural activities, and (2) a sufficient degree of economic security to make the profession attractive to men and women of ability. Freedom and economic security, hence, tenure, are indispensable to the success of an institution in fulfilling its obligations to its students and to society.
>
> American Association of University Professors, 1970

Scholars such as Christopher Jencks and David Riesman (1968) celebrated "the rise to power of the academic profession" (p. xiii) because tenure harnessed the unfettered power of administrators. Academic freedom is a central tenet of tenure (Brown and Kurland, 1990).

Economics of Tenure Perspective

The long-held view is that tenure is appropriate for the type of work for which academics are trained and in which they engage. Preparation for such a career is highly specialized and requires years of training; hence, possibilities for transferring these skill sets are limited. Without tenure, it would be difficult to entice highly talented young people to devote extensive time and financial resources on careers that are considered as high risk (McPherson and Winston, 1988). In addition, without a tenure system, a newly hired faculty member could be perceived to pose a considerable threat to the positions of existing faculty members. Since hiring in academia is unlike other careers because it entails a strong peer review component, in the absence of tenure, faculty may be more inclined to hire less dangerous and hence more mediocre junior colleagues (Carmichael, 1988).

In the 1980s, the concept of tenure was besieged (Chait and Ford, 1982; Finkin, 1996; Horn, 1999a,b). Arguments fueling the debate against tenure included the following: Because it was difficult to terminate nonproductive professors, it promoted the promulgation of 'deadwood' within academic departments. Also, because of particular departmental cultures, tenure was alleged to inhibit, rather than promote, creativity. Also, it was argued that universities were paralyzed by a rigid – rather than flexible – academic labor force. The key counterargument supporting tenure systems was that a better substitute did not exist. Renewable contracts were costly in terms of resources that would not offset the 'deadwood' problem. Also, because tenure goes hand in hand with academic freedom, tenure protects outspoken and intellectually innovative faculty. Over the last three decades, these debates continue (Acker, 2009; Finn, 2009; Horn, 2009; Kingwell, 2009; Soto Anthony and Hayden, 2009). Such attacks have served as ideological support for shrinking of the stock of tenured professors based on economic grounds.

Current Reforms and Policy Debates

Over the last three decades, significant changes have taken place in regard to the structure of academic careers. With tremendous expansion of higher education – in Martin Trow's (1999) terminology, from elite (less than 15%) to mass (i.e., up to 40%) to universal participation (greater than 40%) – concurrently the nature of the academic labor force has been restructured. In most countries, academic posts in higher education and research systems continue to grow (as expansion of higher education has reached unprecedented levels); however, the composition of the academic workforce is undergoing transformation. The balance between permanent and fixed-term positions is shifting toward the latter. The academic core of tenured professors is declining in most OECD countries, not necessarily in absolute numbers, but as a percentage of the total academic workforce (Kezar, 2013; Schuster and Finkelstein, 2006).

In different countries, this general trend is evolving in various ways. On one extreme, the American system has the highest percentage of permanent positions; on the other extreme, in the German system only 10% of the entire academic workforce has permanent positions (Kreckel and Zimmermann, 2014). All others have fixed-term positions in the form of 'professors in waiting' or those who work from contract to contract (soft money) for their entire careers without any guarantee of renewal. The contraction of a highly developed system (e.g., the United States) in order to reorient itself to higher education expansion in the twentieth century is in stark contrast to an ineffective response of an outdated elite system (e.g., Germany) – characterized by escalation of the 'risk' character of academic careers and related dead ends. However, increasing awareness of the challenges facing such systems has led some universities to introduce bold reforms. For example, the Technical University of Munich has introduced a tenure track system (München, 2012).

Ongoing contraction of tenured positions in both systems has severe implications for the career prospects of junior academics and graduate students who are considering academic careers. Because the prospect of embarking on a secure academic position is decreasing, the next generation of prospective academics faces the risk of precarious employment. Two recent reports address this issue (Boston University, 2010; The Modern Language Association of America, 2014).

Currently, two phenomena can be observed. On the one hand, an increasing proportion of teaching duties in higher education systems is carried out by fixed-term and mostly part-time academics, such as adjunct professors in the United States or *Lektoren* in the German-speaking countries. On the other hand, academic research is becoming increasingly project based and is mostly performed by postdoctoral fellows (sometimes doctoral students) who are employed on soft money. Increasingly, these positions are funded by agencies that fund basic research and/or from private contractors.

Both developments point toward fundamental shifts in the normative underpinnings of the academic career structure. Since the Humboldtian reforms in Prussia approximately 200 years ago, the norm for the academic core position has been the combination of teaching and research, mostly supplemented by other duties such as administration or service. This strong teaching/research nexus was tailored originally for elite higher education institutions of the nineteenth century; however, it has been maintained and adapted during the period of transformation from elite to mass higher education. In that course of

massification, such a strong emphasis on both research and teaching has been questioned regarding its appropriateness for the changing nature of higher education. However, the normative power of this model has been forceful enough to keep the traditional academic job description intact. As higher education systems surpass massification and move toward universal participation, the unconditional research/teaching nexus shows signs of erosion. The high cost of the traditional academic model that is increasingly at odds with a more and more diversified system of research and higher learning is the main driver of change. Academic careers that are defined by a strong emphasis on both research and teaching will continue to exist, but most likely the proportion of faculty who hold such positions will be reduced to minority status in academia. This raises the question of how the decoupling of this traditional structure will affect the academic profession in general. Currently, it goes hand in hand with a decrease in permanent positions and an intensified risk of precarious positions for future generations of academics.

Other parallel structures of full-time employment are either developing or under consideration in various jurisdictions. Teaching tenure track positions are emerging to meet the demand for more instructional time which is currently met increasingly through contingent teaching staff (e.g., see the University of British Columbia). Renewable full-time contracts for sessional or adjunct faculty, as specified in university collective agreements, have the potential to provide a higher level of job security and benefits for those who would otherwise face precarious labor market conditions. However, precarious academic appointments, defined by lack of job security, low pay, few or no benefits, poor working conditions, and limited job security remain common in the academic institutions of today (Kezar, 2013).

Academia has always been a 'risk career' (Weber, 1947); however, recent developments are dangerous in that the most talented graduates are increasingly less attracted to this profession. High on the agenda are policy solutions that do justice to the increased diversification of the academic landscape while providing stable career options to future teachers and researchers.

Bibliography

Acker, S., 24–27 October/November 2009. Gender Equity and the Tensions of Tenure. Academic Affairs.

Altbach, P.G., Reisberg, L., Yudkevich, G.A., Pacheco, I.F. (Eds.), 2012. Paying the Professoriate. A Global Comparison of Compensation and Contracts. Routledge, New York.

American Association of University Professors, 1970. 1940 Statement of Principles on Academic Freedom and Tenure with 1970 Interpretive Comments. Retrieved from: http://www.aaup.org/report/1940-statement-principles-academic-freedom-and-tenure.

Bazeley, P., 2003. Defining "early career" in research. Higher Education 45, 257–279.

Boston University, 2010. Report of the Task Force for Non-Tenure-Track Faculty. Boston University.

Brown, R.S.J., Kurland, J.E., 1990. Academic Tenure and Academic Freedom. Faculty Scholarship Series, Paper 2718.

Busch, A., 1963. The vicissitudes of the "privatdozent": breakdown and adaptation in the recruitment of the German university teacher. Minerva 1, 319–341.

Carmichael, L.H., 1988. Incentives in academics: why is there tenure? Journal of Political Economy 96 (3), 453–473.

Chait, R.P., Ford, A.T., 1982. Beyond Traditional Tenure. Jossey-Bass, San Francisco.

Court, S., 1998. Academic tenure and employment in the UK. Sociological Perspectives 41 (4), 767–774. Retrieved from: http://www.jstor.org.ezproxy.library.ubc.ca/stable/1389668.

European Commission, 2007. Remuneration of Researchers in the Public and Private Sectors. European Commission, Brussels.

European Ministers Responsible for Higher Education, May 2005. The European higher education area: achieving the goals. In: Bergen: Communiqué of the Conference of European Ministers Responsible for Higher Education, pp. 19–20.

European University Association, 2007. EUA's Contribution to the Bologna Ministerial Meeting. European University Association, Brussels.

Finkin, M.W., 1996. The Case for Tenure. Cornell University Press, Ithaca, NY.

Finn, P., 22–23 October/November 2009. The Real Case against Tenure. Academic Affairs.

Franck, E., Opitz, C., 2007. The singularity of the German doctorate as a signal for managerial talent: causes, consequences, and future developments. Management Revue 18 (2), 220–224.

Geiger, R., 2010. Postmortem for the Current Era: Change in American Higher Education, 1980–2010. Working Paper No. 3. Pennsylvania State University, University Park, Pennsylvania.

Horn, M., 1999a. Academic Freedom in Canada: A History. University of Toronto Press, Toronto.

Horn, M., 1999b. Tenure and the Canadian professoriate. Journal of Canadian Studies 34 (3), 261–281.

Horn, M., 5–8 October/November 2009. The Case for Tenure. Academic Affairs.

Jencks, C., Riesman, D., 1968. The Academic Revolution. Doubleday, New York.

Kehm, B.M., 2006. Doctoral Education in Europe and North America: A Comparative Analysis. Wenner Gren International Series. Portland Press Ltd, Portland.

Kezar, A., 2013. Changing Faculty Workforce Models. TIAA-CREF Institute, New York.

Kingwell, M., 13–16 October/November 2009. The Tenure Blues. Academic Affairs.

Kreckel, R., Zimmermann, K., 2014. Hasard oder Laufbahn. In: Akademische Karrierestrukturen im internationalen Vergleich. Akademische Verlagsanstalt, Leipzig.

McPherson, M.S., Winston, G.C., 1988. The economics of academic tenure: a relational perspective. In: Breneman, D., Youn, T. (Eds.), Academic Labor Markets and Careers. Falmer Press, London/New York, pp. 174–199.

Metzger, W.P., 1987. Academic profession in United States. In: Clark, B.R. (Ed.), The Academic Profession. National, Disciplinary, and Institutional Settings. University of California Press, Berkeley.

Münch, R., 2011. The rhetoric of functionality in reconstructing the academic world. In: Halvorsen, T., Nyhagen, A. (Eds.), Academic Identities – Academic Challenges? Cambridge Scholars Publishing, Newcastle, pp. 39–63.

München, T.U., 2012. TUM Berufungs- und Karrieresystem. Statut zum Qualitätsmanagement. Technische Universität München. Retrieved from: http://www.google.at/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0CCAQFjAA&url=http%3A%2F%2Fportal.mytum.de%2Fkompass%2Fpersonalwirtschaft_public%2FTUM-Berufungs-und-Karrieresystem-deutsch-englisch.pdf%2Fdownload&ei=g93pU7XEF4LMyAOv4YLQDQ&usg=AFQjCNEALqoUpDS7ZOINYFHgIRsXfxArjQ&bvm=bv.72676100,d.bGQ.

Nature, 2011. The PhD factory. The world is producing more PhDs than ever before. Is it time to stop? Nature 472, 276–279.

Nerad, M., Heggelund, M. (Eds.), 2008. Toward a Global PhD. Forces and Forms in Doctoral Education Worldwide. University of Washington Press, Seattle.

Schuster, J.H., Finkelstein, M.J., 2006. The American Faculty: The Restructuring of Academic Work and Careers. The John Hopkins University Press, Baltimore.

Soto Anthony, J., Hayden, R., 17–21 October/November 2009. Are Tenured Faculty Slackers? Academic Affairs.

The Modern Language Association of America, 2014. Report of the MLA Task Force on Doctoral Study in Modern Language and Literature. The Modern Language Association of America, New York. Retrieved from: http://www.mla.org/pdf/taskforcedocstudy2014.pdf.

Trow, M., 1999. From mass higher education to universal access: the American advantage. Minerva 37, 303–328.

Usher, R., 2002. A diversity of doctorates: fitness for the knowledge economy? Higher Education Research and Development 21 (2), 143–153.

Weber, M., 1947. Science as a profession. In: Gerth, H.H., Mills, C.W. (Eds.), From Max Weber: Essays in Sociology. Kegan, London, pp. 129–156.

Williams Committee of Inquiry into Education and Training, 1979. Commonwealth of Australia, Canberra.

Academic Engagement

Jennifer A Fredricks, Connecticut College, New London, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Engagement is a 'metaconstruct' that includes behavioral, emotional, and cognitive engagement. The research on engagement has grown out of two theoretical traditions including research on school dropouts and research grounded in motivational and psychological theory. Engagement has been shown to be a key contributor of academic achievement, school completion rates, and lower risk behaviors. Prior research has linked family factors and instructional and social dimensions to higher engagement. Methodological advances in research include: (1) conceptualizing engagement as a multidimensional construct, (2) examining developmental trajectories of engagement, (3) testing reciprocal relations between context and engagement, and (4) using person-centered analyses to explore diversity of engagement profiles.

Engagement

There in a growing interest in the construct of engagement among researchers, policy makers, and practitioners. Engagement has been studied at four different nested levels, including the following: (1) engagement with prosocial activities (i.e., families, schools, and community organizations), (2) engagement with school, (3) engagement in the classroom, and (4) engagement with learning activities (Skinner and Pitzer, 2012). Engagement in school refers to involvement in academic contexts and school-based extracurricular activities, such as sports, music, and student government. Student engagement in the classroom includes aspects of curriculum and students' relationships with teachers and peers. Finally, engagement in learning activities refers to specific aspects of the task and activity structure.

An important question in the literature is how engagement differs from motivation. Although some scholars use these terms interchangeably, the majority assume that engagement and motivation are distinct but related constructs. Motivational constructs tend to emphasize individual differences and psychologist processes. In contrast, engagement reflects an individual's interaction with context. An individual is engaged in something (i.e., task, activity, relationship), and their engagement cannot be separated from the environment (Fredricks et al., 2004). Most scholars postulate that motivation is the precursor of engagement. In other words, engagement is the outward manifestations of a motivated student.

What Is Engagement?

What is school or student engagement? The answer to this question has changed over time (Appleton et al., 2008; Christenson et al., 2012; Fredricks et al., 2004). Early scholars defined engagement primarily in terms of observable behaviors such as participation or time on task. Other scholars have proposed a two-dimensional model that includes both behavioral and emotional engagement (Finn, 1989; Skinner et al., 2009). More recently, scholars have suggested a tripartite conceptualization of engagement that includes behavioral, emotional, and cognitive dimensions (Fredricks et al., 2004). Still other scholars have proposed a four-dimensional model, adding academic engagement (i.e., time on task, homework completion) or authentic engagement (i.e., proactive, intention, constructive contribution) to the two- and three-dimensional engagement models (Christenson et al., 2012).

Behavioral engagement has been defined in several ways including the following: (1) positive conduct and absence of disruptive behavior, (2) participation in classroom learning and academic tasks, and (3) participation in school-based extracurricular activities. Emotional engagement has been defined in terms of students' positive and negative reactions, including interest, boredom, happiness, sadness, and anxiety. Other scholars have defined emotional engagement as value and school belonging. Finally, cognitive engagement has been defined as the use of self-regulatory strategies, investment in learning, depth of processing, and preference for challenge (Fredricks and McColskey, 2012).

One commonality across the recent conceptualizations of engagement is that it is a multidimensional construct. However, there is less agreement on the number and types of engagement, which range from two (i.e., behavior and emotion) to four (i.e., academic, behavioral, cognitive, and psychological/affective). In addition, there has been considerable variation in how each subtype is defined and operationalized for measurement purposes. There is also variability in the location and time frame of engagement. For example, scholars differ in whether they focus on engagement at the school or at the classroom level, and whether they focus on moment-to-moment task engagement or longer term engagement. Scholars also differ in whether they view engagement on a single continuum ranging from high to low, or whether engagement and disengagement (i.e., disaffection) are considered distinct and separate constructs (Skinner et al., 2009). Finally, there is a debate about whether it is important to distinguish between facilitators of engagement, or explanatory casual factors such as teacher support, from indicators, or descriptive markers of engagement, such as effort, attention, and interest, in both definitions and measures of engagement (Appleton et al., 2008).

Measurement of Engagement

This past decade has also seen increasing attention to the measurement of this construct.

The most common way to measure engagement is with student self-report measures. This assessment technique is particularly useful for assessing emotional and cognitive engagement which are not directly observable and need to be inferred from behavior. In a recent review of the literature, Fredricks and McColskey (2012) identified 12 self-report survey measures that have been developed to assess engagement, with varying quality of psychometric properties. These measures differed in the types of engagement assessed and whether they were designed to measure school engagement generally or within a specific academic subject area. The majority of these self-report measures assessed general school engagement. One problem with general measures of engagement is that it is difficult to determine the extent to which engagement represents a general tendency and the extent to which it is situation specific.

Another way to measure student engagement is with observational methods. The majority of these observational measures assess on-task behaviors as an indicator of behavioral engagement. These measures use a form of time sampling, in which an observer records whether a predetermined category of behavior is present or absent during a specific time period. One problem with these observational measures is that they provided limited information on the quality of effort, participation, or thinking (Fredricks et al., 2004). For example, some students who are judged to be on-task by observers may not be cognitively engaged, while students who appear to be off-task may actually be deeply thinking about the task.

Why Does Engagement Matter?

There has been an explosion of research on engagement over the past two decades. Why? First, engagement is a key contributor to academic performance. Several studies have linked engagement in school and out-of-school extracurricular activities to higher grades, achievement test scores, school completion, and graduation rates (see Fredricks et al., 2004; Fredricks and Eccles, 2006). Engagement is also correlated with lower delinquency, substance use, depression, and dropout rates (Li and Lerner, 2011). Second, the ideas of engagement and disengagement are easily understood by educators. Many teachers report that disengagement is one of the biggest challenges they encounter in their classrooms. Recent studies indicate that as many as 40–60% of students are showing signs of disengagement (Yazzie-Mintz and McCormick, 2012).

Third engagement is presumed to be malleable and can be influenced by social and instructional factors. In fact, increasing engagement has been the central goal of many school improvement efforts, especially at the secondary level (National Research Council and Institute of Medicine, 2004). Focusing on student and school engagement as an explanatory variable for achievement, deep learning, and school completion offers more insight into intervention and prevention strategies than focusing on unalterable demographic variables like gender, ethnicity, and socioeconomic status. Finally, engagement has intuitive appeal because it is assumed to be a metaconstruct that has the potential to provide a richer characterization of how students' think, act, and feel, than research on any single component can offer. Considering engagement as a multidimensional construct also can provide more specific prescriptions for intervention and prevention efforts.

Theoretical Frameworks – Dropout Literature

Research on engagement has grown out of two separate literature that draw on different theoretical traditions including: (1) research on school dropouts and (2) research grounded in psychological and motivational theory. For example, the participation–identification model (Finn, 1989) is a seminal theory addressing the critical role of school engagement in the process of dropout and school completion. In this two-component model, participation refers to behaviors that engage students in learning (i.e., behavioral engagement), and identification refers to students' feelings of belonging and valuing (i.e., emotional engagement). According to this model, dropping out of school is the result of a long-term developmental process that occurs when students fail to participate in learning activities and fail to identify with school. Prior research supports the assumption that dropout and school completion rates can be predicted from behavioral indicators and attachment to school (Christenson et al., 2012).

The life course perspective is another theoretical model that comes out of the dropout literature. This theoretical perspective views dropping out of school as a process of academic disengagement that can be traced back to student's earliest experiences in school (Alexander et al., 2001). A few longitudinal studies have used a life course perspective to track the long-term educational experiences of low-income children. This research supports the assumption that developmental pathways to school dropout and completion begin with the engagement behaviors and engagement attitudes developed during the early school years.

Theoretical Frameworks – Motivational and Developmental Literature

Much of the research in this field has applied motivational and developmental theories to research on student engagement. For example, one prominent theory guiding the current literature is the self-system motivational model. This mediational model is rooted in self-determination theory and links contextual factors (i.e., classroom structure, autonomy support, and involvement) to patterns of engagement (i.e., engagement vs. disengagement), through self-system processes, or an individual's appraisals of how *competent*, *autonomous*, and *related* he or she feels within particular contexts (Connell and Wellborn, 1991). This model assumes that a supportive classroom, peer, or family context (i.e., optimal structure, autonomy support, and high involvement) promotes positive self-perceptions, which in turn, increases behavioral, emotional, and cognitive engagement. Conversely, a less supportive classroom, peer, or family context undermines self-perceptions which in turn, results in disengagement. There is growing body of literature supporting the links between contextual factors, self-system processes, and behavioral, emotional, and cognitive engagement (Skinner and Pitzer, 2012).

The person–environment perspective (Eccles and Roeser, 2009) is another relevant theoretical model. According to a person–environment fit model, behavior and motivation are influenced by the fit between characteristics individuals bring to their social environments and the characteristics of the social environment. Eccles and her colleagues have used this theory to explain the decline in motivation and engagement during the middle school years, suggesting that there is a mismatch between adolescents' developmental needs and the environment of middle schools. Person–environment fit models could also help to explain the continuity and/or discontinuity between the values and practices espoused by families and schools to support engagement. One explanation for the disengagement that some students experience in schools is a discontinuity between their school and home environments in terms of support and expectations for learning.

Finally, other scholars have used a bioecological model as the overarching theoretical model for studying engagement (Bronfenbrenner and Morris, 2006). This model focuses on four contributors to development, including *person*, *process*, *context*, and *time*. This theoretical model recognizes both the individual characteristics, such as ability, socioeconomic status, and performance, and the contextual factors, such as the home, school, neighborhood, and larger sociocultural context, that impact on engagement.

Classroom Context and Engagement

Teacher and Peer Support

Teacher support in terms of caring, involvement, and encouragement has been linked to indicators of behavioral, emotional, and cognitive engagement across race, ethnicity, and class (Fredricks et al., 2004). A close supportive relationship with a teacher is especially important for students who are academically at risk and showing signs of disengagement. Peer support for learning and the quality of friendships are also linked to behavioral and emotional engagement (Christenson et al., 2012; Kindermann, 2007).

Task Characteristics

The type of academic tasks students do in school is another important determinant of engagement. Many students disengage from learning in school because the classroom tasks are boring and offer limited opportunities for deeper conceptual understanding. Researchers are exploring how to make classroom tasks more intrinsically motivating, interesting, and cognitively engaging to students. Engagement is higher in classrooms with authentic tasks that are situated in meaningful contexts, are cognitively complex, and reflect how learning happens outside of the classroom (Fredricks et al., 2004). The degree of autonomy support, or the amount of freedom a student is given to determine his or her behavior, also influences classroom engagement (see Eccles and Roeser, 2009). For example, autonomy supportive instruction (giving choices, making learning relevant) predicts higher classroom engagement (Reeve et al., 2004).

Classroom Management/Structure

There is a positive relationship between being in a well-managed classroom and exhibiting behavioral engagement as measured through time on task (Fredricks et al., 2004). Both behavioral and emotional engagement are also enhanced in classrooms where teachers establish procedures that help to monitor student progress, ensure work completion, provide adequate structure, and facilitate productive group functioning.

Family Context and Engagement

Parents influence children's engagement through the types of home environments they create, the values they endorse, and the experiences they provide. Prior research has linked parental expectations, parental involvement, and parental social support to indicators of behavioral, emotional, and cognitive engagement (Eccles, 2007; Wang and Eccles, 2012a). Parents also impact on student engagement through their extent of autonomy support, or the degree to which they encourage their children to initiate their own behavior, take an active role in solving their own problems, and express their points of view. For example, strong parental autonomy support predicts behavioral engagement (i.e., time on homework, lower deviance), emotional engagement (i.e., intrinsic motivation, positive attitudes toward school), and cognitive engagement (i.e., persistence in the face of challenge) (Pomerantz et al., 2007).

Engagement as a Multidimensional Construct

In 2004, Fredricks and her colleagues outlined several limitations with the literature on engagement to date. Over the past decade, there is a growing body of research addressing these critiques. One of the central concerns articulated in the 2004 review is that scholars had failed to capitalize on the potential of engagement as a multidimensional construct. Prior to 2004, much of the research had either combined conceptually different engagement measures into one index or focused on only one of components of engagement, usually behavioral engagement. In the last decade, a growing number of scholars have conceptualized engagement as a multidimensional construct. The majority of this research defines engagement in terms of behavioral, emotional, and cognitive dimensions. However, there is still significantly more research on behavioral engagement than the other two dimensions.

Another critique was that few studies used a multidimensional perspective to disentangle differences in the antecedents and consequences of engagement by type of engagement. This critique has been addressed in a few recent studies. For example, Wang and Eccles (2012a) found that the association between social support and engagement differed across the three types of engagement. Other research has shown that the three types of engagement have differential relations to developmental outcomes (Archambault et al., 2009; Wang and Eccles, 2012b). For example, Archambault et al. (2009) found that a measure of global disengagement was associated with dropping out of school. However, when broken into subcomponents, only behavioral engagement predicted school dropout.

Trajectories of Engagement

Another criticism articulated in the 2004 review was the over-reliance on cross-sectional and short-term longitudinal studies. Recently, a few studies have addressed this critique by using longitudinal modeling techniques to chart changes in engagement over longer periods of time. This research shows declines in student engagement during the high school years (Van de gaer et al., 2009; Wang and Eccles, 2012b). In addition, a few studies have used longitudinal modeling techniques to explore the heterogeneity in engagement trajectories during the early grades (Ladd and Dinella, 2009) and during the adolescent years (Archambault et al., 2009, 2008; Li and Lerner, 2011). These studies show diversity in pathways of behavioral, emotional, and cognitive engagement over time. For example, Li and Lerner (2011) identified four engagement trajectories: (1) high stability, (2) moderate stability, (3) transitory decreasing, and (4) decreasing. In turn, these trajectories were linked to different patterns of academic achievement, psychological adjustment, and risk behavior.

Although these longitudinal studies have provided a more nuanced portrait of developmental changes in school engagement, there is still a need for research that examines the trajectories of engagement at different stages of development, particularly through changing school contexts and developmental transitions. In addition, there is a need for research that examines the individual and contextual factors that may help to explain both continuity and change in engagement trajectories. Such research will provide insights into the risk and protective factors associated with the development of each type of engagement. Finally, it is important to further examine the relation between different types of engagement trajectories and student outcomes over time to better understand what are the most adaptive and risky profiles.

Reciprocal Relations

Although most theories assume a reciprocal relation between context and engagement, the majority of research has been based on correlational and cross-sectional studies. This research tends to be interpreted as reflecting context effects on engagement. A few survey studies have tested bidirectional relations between contextual factors and engagement. These studies show that adults respond differently to children depending on their level of engagement or disruptive behavior, and that children are allowed entry into peer groups based on their engagement in school (Kindermann, 2007; Skinner and Pitzer, 2012). This initial research suggests that these bidirectional relations are self-amplifying and magnify initial differences in engagement over time. Higher engaged students get more positive support, while less engaged youth experience less supportive and positive interactions over time.

Although the research has become more methodologically sophisticated over the past decade, there is clearly a need for further longitudinal research that tests the cyclical associations between contextual factors, engagement, and developmental outcomes. Incorporating other measurement techniques may also help scholars of engagement to measure cyclical processes. For example, several measurement techniques have been developed in the self-regulated literature to evaluate students' engagement in the context of specific learning tasks. These measures include: (1) behavioral traces (i.e., overt indicators of student cognition during tasks), (2) direct observations of student behavior, (3) think aloud protocols, and (4) self-regulated learning (SRL) microanalysis (i.e., structured interview of students' attitudes and cognitions before, during, and after a task) (Cleary and Zimmerman, 2012).

Person-Centered Approaches

The vast majority of research on student engagement has used variable-centered techniques to isolate the average relations between contextual factors, engagement, and other developmental outcomes. A complementary approach is to use person-centered techniques to capture patterns of engagement within individuals across time or contexts. These techniques can help to answer questions about how much individuals vary in terms of their behavioral, emotional, and cognitive engagement profiles and whether these different profiles have different implications for individual developmental processes. Failure to consider the diversity of within-person profiles impedes our ability to design targeted intervention for specific groups of students.

A few studies have used person-centered approaches to identify different patterns of student engagement (Janosz et al., 2000; Wang and Peck, 2013). This research has shown that there are different profiles of behavioral, emotional, and cognitive engagement. For example, Wang and Peck (2013) identified five different profiles: highly engaged, moderately engaged, minimally engaged, emotionally disengaged, and cognitively disengaged. These five profiles differed on indicators of educational and psychological functioning. Recently, a few studies have combined person-centered and longitudinal approaches to understand variability in engagement profiles over time (Archambault et al., 2009; Ladd and Dinella, 2009; Li and Lerner, 2011).

Conclusion and Future Directions

Important methodological and theoretical advances have led to increased knowledge about engagement, though it is clear that several gaps in the literature still remain. Part of the appeal of engagement as a construct is that it is multidimensional and comprised of observable behaviors, internal cognitions, and emotions. This makes it a relevant construct in a variety of research areas and appeals to both policy makers and practitioners. However, one problem with engagement being a broad 'metaconstruct' is that it has resulted in considerable variability in definitions both within and across the different types of engagement. In other words, one author's conceptualization of behavioral engagement can be and often is the same as another's operationalization of cognitive engagement (Christenson et al., 2012). For example, value has been classified as an aspect of both emotional and cognitive engagement. Additionally, effort has been used by some to describe compliance with work requirements (behavioral engagement), and by others to reflect the degree of psychological investment

in learning (cognitive engagement) (Fredricks and McColskey, 2012).

One concern is that engagement is being used so broadly that it appears to explain everything related to students' experience and success in school. Although this broad generality has utility in discussions related to policy and practice, it is more problematic in research that is attempting to examine specific hypotheses about the relation between context and engagement. This lack of definitional clarity has also made it difficult to compare findings across studies. Another concern is that definitions of engagement overlap with other educational constructs such as self-regulation, school bonding, belonging, and school climate, which often have much stronger bodies of literature supporting the construct. Researchers need to be clearer about how they are defining engagement, at which level they are measuring, and the 'value added' from studying engagement as opposed to these earlier literature (Fredricks et al., 2004).

Although there is a growing consensus that engagement is a multidimensional construct, we know much less about the extent to which emotional, cognitive, and behavioral engagement have separate and cumulative effects on achievement outcomes, school completion rates, psychological functioning, and risk behaviors, as well as the potential mediational processes involved (Christenson et al., 2012). Another important question is how the different types of engagement influence each other. Theoretical and empirical relations between the three types of engagement are lacking. One example is the participation–identification model, which assumes that behavioral engagement (participation) leads to emotional engagement (identification), which in turn increases behavior (Finn, 1989). Longitudinal research can help to verify the tenets of this model and how the different dimensions are associated with each other.

In addition to definitional challenges, there is a critical need for better measures of student engagement. One problem is that scholars often do not have a strong theoretical or conceptual framework guiding the choice of indicators. Several scholars have argued that a more systematic and thoughtful attention to the measurement of student engagement is one of the most pressing imperatives for future research (Appleton et al., 2008; Fredricks and McColskey, 2012). It is clear that we cannot know if we are improving engagement unless we are able to measure it accurately.

As outlined in Fredricks and McColskey's review (2012), the number of self-report surveys developed to assess engagement has increased dramatically over the past decade. However, the extent of psychometric support for these measures varied dramatically. In addition to self-report instruments, there is need to consider measurement from additional perspectives (i.e., teacher and parent). It is also important to incorporate multiple methods (i.e., qualitative and quantitative) and assess engagement at multiple levels (i.e., school, class, and moment to moment) (Christenson et al., 2012). For example, experience sampling method (ESM) is one innovative method that allows researchers to collect detailed data on engagement in the moment rather than retrospectively. In this methodology, individuals respond to periodic ESM signals with a series of questions about their location, activities, and cognitive and affective responses. Finally, it is important to examine group differences in the measurement of engagement by age, race, ethnicity, culture, and gender. This research can help to determine the extent to which the construct of engagement is universal and the extent to which it varies across demographic factors (e.g., age, gender, ethnicity, culture).

One of the strengths of engagement as a construct is that it is malleable and thus offers insight into contextual factors that can be targeted for interventions. Unfortunately, much of the prior research has made it difficult to test questions of malleability (Fredricks et al., 2004). There is a critical need for research that goes beyond descriptive and correlational studies to determine which aspects of context influence engagement over time. There is also a dearth of intervention studies which examine the efficacy of different interventions in increasing engagement.

Finally, the range of participants in engagement studies needs to be expanded to account for variation in students' backgrounds and experiences. In addition, there is a need for research testing whether contextual factors have a differential effect on engagement across different demographic groups. For example, some research suggests that supportive relationships with teachers may have a greater effect on engagement for minority students and for girls than for Caucasian students and boys (Christenson et al., 2012). Person–environment fit is one theoretical model that can be used to examine the types of contexts that are most effective in increasing engagement for particular types of students. This information can help to develop targeted interventions that are specific to different groups of students.

See also: Academic Motivation and Performance: Task Value Interventions; Affect-Regulation Motivation; Apprenticeship and School Learning: Lessons from Germany; Avoidance and Approach Motivation: A Brief History; Burnout, Psychology of; Flow in Motivational Psychology; Gender and Academic Motivation; Grit; Interest, Psychology of; Leisure Activities Choices among Adolescents; Mastery Learning; Motivation in Australian Aboriginal Populations; Motivation in Youth Sport and Physical Activity: Developmental Perspectives; Motivation, Learning, and Instruction; Motivation: Life Course and Sociological Perspectives; Passion and Motivation; Personal Projects; Race and Academic Motivation; School Achievement: Motivational Determinants and Processes; School Burnout and Engagement: Lessons from a Longitudinal Study in Finland; Schooling: Impact on Cognitive and Motivational Development; Self-Determination Theory; Self-Regulated Learning: Theories, Measures, and Outcomes; Self-Regulation in Adulthood; Sociodigital Revolution: Digital Natives vs Digital Immigrants; Temperament and Motivation; Test Anxiety and Academic Achievement; Vocational Interests, Values, and Preferences, Psychology of.

Bibliography

Alexander, K.L., Entwisle, D.R., Kabbini, J.S., 2001. The dropout process in life course perspective: early risk factors at home and school. Teachers College Record 103, 760–821. http://dx.doi.org/10.1111/0161-4681.00134.

Appleton, J.J., Christenson, S.L., Furlong, M.J., 2008. Student engagement with school: critical conceptual and methodological issues of the construct. Psychology in the Schools 45, 369–386. http://dx.doi.org/10.1002/pits.20303.

Archambault, I., Janosz, M., Morizot, J., Pagani, L., 2009. Adolescent behavioral, affective, and cognitive engagement in school: relation to dropout. Journal of School Health 79, 408–415. http://dx.doi.org/10.1111/j.1746-1561.2009.00428.x.

Archambault, I., Janosz, M., Fallu, J., Pagani, L.S., 2008. Student engagement and its relationships with early high school dropout. Journal of Adolescence 32, 651–670.

Bronfenbrenner, U., Morris, P., 2006. The bioecological model of human development. In: Lerner, R.M., Damon, W. (Eds.), Theoretical Models of Human Development, The Handbook of Child Psychology, sixth ed., vol. 1. Wiley, New York, pp. 793–828.

Christenson, S.L., Reschly, A.L., Wylie, C. (Eds.), 2012. Handbook of Research on Student Engagement. Spring Science and Business Media, New York.

Cleary, T.J., Zimmerman, B.J., 2012. A cyclical self-regulatory account of student, engagement: theoretical foundations and applications. In: Christenson, S., Reschly, A.L., Wylie, C. (Eds.), Handbook of Research on Student Engagement. Spring Science and Business Media, New York, pp. 237–258.

Connell, J.P., Wellborn, J.G., 1991. Competence, autonomy, and relatedness: a motivational analysis of self-system processes. In: Gunnar, M.R., Sroufe, L.A. (Eds.), Self Processes in Development: Minnesota Symposium on Child Psychology, vol. 23. University of Chicago Press, Chicago, pp. 43–77.

Eccles, J.S., 2007. Families, schools, and developing achievement-related motivations and engagement. In: Grusec, J.E., Hastings, P.D. (Eds.), Handbook of Socialization. The Guilford Press, New York, pp. 665–691.

Eccles, J.S., Roeser, R.W., 2009. Schools, academic motivation and stage-environment fit. In: Lerner, R.M., Steinberg, L. (Eds.), Handbook of Adolescent Psychology, third ed. John Wiley & Sons, Hoboken, NJ, pp. 404–434.

Finn, J.D., 1989. Withdrawing from school. Review of Educational Research 59, 117–142. http://dx.doi.org/10.3102/00346543059002117.

Fredricks, J.A., Blumenfeld, P.C., Paris, A.H., 2004. School engagement: potential of the, concept, state of the evidence. Review of Educational Research 74, 59–109. http://dx.doi.org/10.3102/00346543074001059.

Fredricks, J.A., Eccles, J.S., 2006. Is extracurricular participation associated with beneficial, outcomes: concurrent and longitudinal relations? Developmental Psychology 42, 698–713. http://dx.doi.org/10.1037/0012-1649.42.4.698.

Fredricks, J.A., McColskey, W., 2012. The measurement of student engagement: a comparative analysis of various methods and student self-report instruments. In: Christenson, S., Reschy, A.L., Wylie, C. (Eds.), Handbook of Research on Student Engagement. Springer, New York, pp. 763–783.

Janosz, M., LeBlanc, M., Boulerice, B., Tremblay, E.R., 2000. Predicting types of school, dropouts: a typological approach with two longitudinal samples. Journal of Educational Psychology 92, 171–190. http://dx.doi.org/10.1037/0022-0663.92.1.171.

Kindermann, T.A., 2007. Effects of naturally existing peer groups on changes in academic, engagement in a cohort of sixth graders. Child Development 78, 1186–1203. http://dx.doi.org/10.1111/j.1467-8624.2007.01060.x.

Ladd, G.W., Dinella, L.M., 2009. Continuity and change in early school engagement: predictive of children's achievement trajectories for first to eight grade. Journal of Educational Psychology 1, 190–206. http://dx.doi.org/10.1037/a0013153.

Li, Y., Lerner, R.M., 2011. Trajectories of school engagement during adolescence: implications for grades, depression, delinquency, and substance use. Developmental Psychology 47, 233–247. http://dx.doi.org/10.1037/a0013153.

National Research Council and Institute of Medicine, 2004. Engaging Schools: Fostering High School Students' Motivation to Learn. National Academy Press, Washington, DC.

Pomerantz, E.M., Moorman, E.A., Litwak, S.D., 2007. The how, whom, and why of, parents' involvement in children's academic lives: more is not always better. Review of Educational Research 77, 373–410. http://dx.doi.org/10.3102/003465430305567.

Reeve, J., Jang, H., Carrell, D., Jeon, S., Barch, J., 2004. Enhancing students' engagement by increasing teachers' autonomy support. Motivation and Emotion 28, 147–169.

Skinner, E.A., Kindermann, T.A., Furrer, C., 2009. A motivational perspective on engagement and disaffection: conceptualizations and assessment of children's behavioral and emotional participation in academic activities in the classroom. Educational and Psychological Measurement 69, 493–525. http://dx.doi.org/10.1177/0013164408323233.

Skinner, E.A., Pitzer, J.R., 2012. Developmental dynamics of student engagement, copying, and everyday resilience. In: Christenson, S.L., Reschly, A.L., Wylie, C. (Eds.), Handbook of Research on Student Engagement. Springer, New York, pp. 21–44.

Van de gaer, E., Pustjens, H., Van Damme, J., De Munter, A., 2009. School engagement and, language achievement: a longitudinal study of gender differences across secondary school. Merrill-Palmer Quarterly 55, 373–405.

Wang, M.T., Eccles, J.S., 2012a. Social support matters: longitudinal effects of social support on three dimensions of school engagement from middle to high school. Child Development 83, 877–895. http://dx.doi.org/10.1111/j.1467-8624.2012.01745.x.

Wang, M.T., Eccles, J.S., 2012b. Adolescent behavioral, emotional, and cognitive engagement trajectories in school and their differential relations to educational success. Journal of Research on Adolescence 22, 31–39. http://dx.doi.org/10.1111/j.1532-7795.2011.00753.x.

Wang, M.T., Peck, S., 2013. Adolescent educational success and mental health vary across school engagement profiles. Developmental Psychology 49, 1266–1276. http://dx.doi.org/10.1037/a0030028.

Yazzie-Mintz, E., McCormick, K., 2012. Finding the humanity in the data: understanding, measuring, and strengthening student engagement. In: Christenson, S.L., Reschly, A.L., Wylie, C. (Eds.), Handbook of Research on Student Engagement. Springer, New York, pp. 743–762.

Academic Motivation and Performance: Task Value Interventions

Yoi Tibbetts, Elizabeth A Canning, and Judith M Harackiewicz, University of Wisconsin–Madison, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

We review interventions that promote motivation in academic contexts. A variety of interventions have been used to enhance subjective task utility – the belief that a task is relevant to other activities or personal goals, with positive implications for interest and performance. We focus on experimental field studies in high school science and college psychology classes. We also review a randomized intervention in which parents received information about the utility value of math and science for their teens in high school; this intervention led students to take nearly one semester more of science and mathematics, compared with the control group. We discuss the mechanisms through which these interventions work.

What students learn in school does not always appear to have relevance to their own lives. When students encounter subjects that appear to have minimal practical use, they may find themselves disengaged, lacking the motivation that educators and parents strive to inspire. This same apathy may be the cause of a number of disturbing academic trends, including reduced interest in math and science over time (Hidi and Harackiewicz, 2000) as well as marked declines in students' pursuit of STEM (science, technology, engineering, and math) majors when compared to other academic fields (National Center for Education Statistics (NCES), 2011). However, researchers have recently begun to develop interventions that may address these motivational problems by increasing student interest in academic disciplines. Some researchers have tried to promote interest by varying features of academic tasks (Bergin, 1999; Hidi and Baird, 1988; Lepper and Henderlong, 2000), others have tried to influence students' sense of autonomy and competence (Deci and Ryan, 1985; Harter, 1981), and others have worked to change individuals' perceptions of tasks (Harackiewicz and Hulleman, 2010; Wigfield and Cambria, 2010). Because interest is a powerful predictor of important achievement choices such as future course enrollment and choice of major (Harackiewicz et al., 2002) as well as various academic performance outcomes (Ainley et al., 2002; Schiefele, 1991), interventions that prove successful in promoting interest in academic topics represent meaningful strides toward a better educated society.

Interventions designed to promote interest can take many approaches. For example, a teacher might try to make an academic task more interesting by changing features of the task, assigning group projects, or embedding learning activities in games. However, it is not always possible to change the nature of a task or activity. A parent or teacher cannot change the fundamental principles of mathematics or science, but they may be able to change the way students think about these subjects. For example, a parent might be able to promote interest in an academic topic by relating it to their child's recreational interests or career goals. By changing the perception of the activity (i.e., helping students perceive tasks as personally important), it may be possible to influence interest. With various possibilities for influencing student interest, it is important to distinguish interventions that change the structure of an activity (Lepper and Cordova, 1992; Schraw and Dennison, 1994) from those that change an individual's perceptions of an activity (Iyengar and Lepper, 1999). Task-based interventions typically vary collative variables (e.g., complexity, incongruity, novelty, variability) to stimulate attention, arousal, and task engagement (Berlyne, 1960), whereas task-value interventions focus on individuals' perceptions and development of subjective task values (Harackiewicz and Hulleman, 2010), guided by Eccles' expectancy-value model (Eccles et al., 1983).

Expectancy-Value Model

According to this model, perceived expectancies for success and subjective *task values* both determine motivation and performance on achievement tasks. One way to inspire interest and motivation is to increase the perceived expectancy of success, and a large research literature has examined the role of self-efficacy and performance expectations in promoting interest and performance (Harter, 2006; Pajares, 1996). It can be difficult to intervene with respect to students' performance expectations, however, and in this article, we concentrate on interventions focused around task values.

Indeed, it may prove more feasible to influence students' subjective task values for academic tasks. Eccles (Eccles, 2009; Eccles et al., 1983) argued that it is important to consider how individuals perceive and value a task, and identified four subjective task values: *intrinsic value* – the perceived importance of a task because of its inherent enjoyment; *attainment value* – the perceived importance of a task for an individual's identity and self-worth; *utility value* – the perceived importance of a task for accomplishing future goals, relevant to an individual's life; and *cost value* – the negative aspects of engaging in a task (e.g., time consumption, performance anxiety). Her expectancy-value model posits that an increase in intrinsic, attainment, or utility value will lead to greater motivation toward an academic task. Conversely, if these perceived task values decrease, individuals may be less likely to pursue the task.

Of the four task values, Eccles and colleagues consider utility value to be the most 'extrinsic' because it extends beyond the task itself to connections between that task and other tasks, activities, or goals (Wigfield and Eccles, 1992). A person finds utility value in a task if they believe it is useful and relevant

beyond the immediate situation, for other tasks or aspects of a person's life. It involves learning the content as it applies to oneself or real life, rather than learning the content absent any personal applicability. For example, when students encounter plant genetics in their biology class, the content may not seem immediately valuable or applicable to their lives. Learning about DNA methylation and gene coding in plants may not seem to have obvious practical or personal implications. However, if a health-conscious student perceives a connection between what they are learning and how it can enable them to distinguish genetically modified food from organic food at grocery stores, they may become more vested in biology and engage more with the content. These types of external connections to content distinguish utility value from the other more internally regulated task values.

Whereas intrinsic and attainment value are based on the inherent enjoyment of the task itself and the importance of the task for an individual's identity, respectively, utility value is predicated on perceiving connections between the immediate task and a future task or activity. Accordingly, some researchers have claimed that utility value is more 'externally regulated' than other task values (Simons et al., 2004). Extrinsic factors (e.g., rewards, prizes, competitions) have historically been at the heart of controversy in the motivation literature with many researchers arguing against the use of using external motivation to promote task motivation (Harackiewicz, 1979; Lepper et al., 1973). In fact, some of the original theorists in the achievement literature have argued that extrinsic motivation is antithetical to the development of interest (Deci and Ryan, 1985) with particularly deleterious effects for interesting activities (Deci et al., 1999). However, when an individual perceives utility value in a task, they may connect the task to important personal goals and outcomes in an intrinsically regulated way that promotes the development of interest (Vansteenkiste et al., 2006). Thus utility value may have a more positive potential than originally thought.

Moreover, given that utility value is based on perceptions of the usefulness of a task for other goals and applications, it may prove to be the task value most amenable to external intervention. A teacher or parent can point out possible connections, or help students appreciate the importance of a topic for their future goals. For this reason, recent research has examined the effects of perceived utility value within laboratory and classroom settings and focused on developing interventions that promote the perception of utility value. In order to explore how changes in perceived utility value impact interest, however, it is important to understand how interest develops over time, as well as distinguish interventions that trigger the development of interest from those that promote the maintenance or deepening of interest (Hidi and Harackiewicz, 2000).

Interest Theory

Hidi and Renninger (2006) proposed a four-phase model of interest development that charts the transition from situationally based interest to personal interest. They argue that interest is the outcome of an interaction between a person and a particular content. The potential for interest lies within the person, but the content and environment influence the strength and direction of interest as well as its continued development (Hidi and Renninger, 2006; Renninger and Hidi, 2011). The fact that interest is influenced by both content and the context of the situation suggests that interventions have the potential to both trigger interest and promote the maintenance of interest.

The two earlier phases of interest are characterized as varying degrees of 'situational interest.' In phase 1, a trigger is necessary (provided by the content or the environment) to spark a temporary affective and cognitive change that results in a short-term increase in interest. If this triggered situational interest is further supported, typically by external sources, it can develop into a more maintained situational interest (phase 2). The latter two phases of Hidi and Renninger's (2006) model are characterized by a predisposition to seek repeated engagement with the content. In order to develop emerging individual interest (phase 3) and well-maintained individual interest (phase 4), the individual must play a more active role in their interest development. Interest in the latter phases is therefore more self-generated and does not necessarily require external support to develop, suggesting that interventions may be most effective in the early stages of interest development by either aiding the development of situational interest, or promoting the transition to a more internalized interest.

In addition to describing the four phases of interest, Hidi and Renninger (2006) hypothesized that increasing the perception of the value of a task is critical for progressing from situational to individual phases of interest. An increase in the perceived value of a task serves as a motivator for individuals to continue engaging in content or an activity. Thus both the Eccles model and the Hidi and Renninger model predict that increasing perceived task value is a viable way to promote interest and motivation. With these theoretical frameworks in mind, researchers have explored the relationship between interest, motivation, and the task value most amenable to outside intervention: utility value.

Utility Value Interventions

The first studies to examine the effect of perceived utility value on educational outcomes were correlational. For example, research demonstrated that the perceived instrumentality of studying was positively correlated with students' persistence and academic performance (De Volder and Lens, 1982; Van Calster et al., 1987). Malka and Covington (2005) showed that the perceived relevance of college students' schoolwork to their future goals predicted classroom performance. A myriad of other correlational studies indicates that when students perceive value in course topics, they develop more interest, take more advanced courses in those academic disciplines, and perform better (Harackiewicz et al., 2000, 2008; Wigfield, 1994).

For example, Hulleman et al. (2008) measured perceived utility value in an introductory psychology class by asking participants to report, early in the semester, the extent to which they found class material to be useful to their everyday life and future career and then examined the relationship between perceived utility value and measures of interest and course grades. They found that perceived utility value was a significant predictor of both interest (measured at the end of the semester) and course grades. Perceptions of utility value were positively correlated with both of these important educational outcomes,

even after controlling for other important variables such as initial interest and other task values (e.g., intrinsic value).

Building on these correlational findings, Durik and Harackiewicz (2007) measured students' baseline interest in mathematics and then taught them a mental math technique in an experimental laboratory paradigm. They varied both the collative features of the task and the availability of utility value information in a crossed experimental design. They varied 'catch' features (features designed to trigger or 'catch' interest) by presenting the mental math technique with colorful fonts and pictures in the catch condition, compared to a plain black and white control condition. To vary utility value information, they provided some students with information about how the technique could be useful in everyday life (e.g., "You might use mental math to figure out tips at restaurants or to manage your bank transactions.") whereas the control condition did not present utility value information. Results indicated that the 'catch' manipulation was effective in promoting the interest of students who were low in initial interest whereas the utility value intervention increased the interest of students who were high in initial interest. Durik and Harackiewicz reasoned that situational triggers, in the form of enhanced collative features of the task, were necessary to initiate interest development for students low in initial interest, but that the utility value intervention helped high-interest participants develop a more maintained interest. Thus a task-based intervention appeared to trigger interest for low-interest participants, but the utility value intervention was effective in supporting interest for high-interest students.

Hulleman et al. (2010) also tested utility value interventions with experimental methods. In two experiments – a laboratory study and a randomized trial in a college class – utility value was manipulated through a writing intervention in which participants were asked to explain how the topic of the learning activity (math in the laboratory study and psychology in the college class) was relevant to their lives. In other words, participants in the experimental condition were asked to generate their own connections and discover utility value themselves through writing about the task. Those in the control condition wrote a summary of the material. In both experiments, the utility value intervention increased interest, especially for participants who were low in expected (laboratory study) or actual (classroom study) performance, indicating the importance of both utility value and performance expectations in predicting interest. It may be that for students who do not expect to perform well, the process of generating personal connections to an academic discipline is particularly important for triggering interest and promoting engagement in the task. Students who expect to do well may already be aware of the utility value of the activity. This possibility is supported by mediation analyses (Hulleman et al., 2010) showing that the intervention worked for low-expectancy students by promoting perceptions of utility (measured via questionnaire). In other words, the intervention increased perceptions of utility for these low-expectancy students, and perceived utility value was a positive predictor of subsequent interest in the task for all students.

Hulleman and Harackiewicz (2009) tested a similar experimental intervention with high school science students. Students enrolled in 9th-grade biology classes were randomly assigned to utility value or control writing conditions at the beginning of the semester, within classrooms. Teachers were told that the research concerned the effectiveness of writing assignments, but were blind to hypothesis and experimental condition. Students' success expectancies and initial interest in science were measured at the beginning of the semester. At the end of each unit (about every 2–3 weeks), students either wrote about how the material they were studying applied to their own lives (utility value condition) or wrote a summary of the same material (control condition). Students' interest in science and future plans for science-related careers were measured at the end of the semester, and course grades were obtained from school records. Results indicated that the intervention was particularly effective for students with low performance expectations in the class: these students reported more interest in the course topic at the end of the semester and obtained higher grades. Given that students who do not believe they can do well are especially at risk for poor performance and decreased academic interest (Eccles et al., 1983; Renninger, 2000), these are the students most in need of help. The utility value intervention improved performance for these at-risk students by nearly two-thirds of a letter grade, and promoted interest for this group. Moreover, Hulleman and Harackiewicz (2009) showed that interest predicted students' science-related career plans, suggesting that utility value interventions might have long-term effects. These results suggest that this simple intervention aimed at promoting the perception of utility value was powerful in promoting important academic outcomes.

It is interesting to note that in Durik and Harackiewicz's (2007) study in which students were presented with utility value information from an outside source, the utility value intervention was most effective for highly interested students, yet in the Hulleman studies (Hulleman et al., 2010; Hulleman and Harackiewicz, 2009), in which students generated their own utility value connections, the utility value intervention was more powerful for less confident students. This highlights the fact that learners approach tasks with different backgrounds (i.e., varying levels of interest and performance expectations) and that utility value interventions may have differing effects depending on an individual's initial orientation toward the task.

For example, the fact that externally presented utility value information (i.e., telling students why a task is important for them) was particularly effective for high-interest students in the Durik and Harackiewicz (2007) study suggests that for students who are already interested in a topic, utility value information may serve as another meaningful way to connect to content and deepen interest. In this way, utility value information may act as an additional motivator for students to continue pursuing a task. The fact that the self-generated utility value intervention in the Hulleman and Harackiewicz (2009) and Hulleman et al. (2010) study was most effective for students with low success expectancies suggests that the act of identifying personal utility value connections (i.e., writing essays about how the content is personally relevant) may be especially important for triggering interest among students who might otherwise become disengaged with the task. For these students with low success expectancies, discovering how a particular content or task relates to their life may be

a powerful mechanism for initiating the development of interest. Considered together, these results suggest that there may be different routes to the promotion of perceived utility value for students. In fact, recent research has examined the possibility of intervening with the people who likely have the greatest influence over students: their parents.

Parents: An Untapped Resource

Given the impressive potential of utility value interventions for promoting important academic outcomes, more recent research has investigated whether it is possible to influence students' perceptions of utility value in other ways. In fact, according to the expectancy-value model parents play a pivotal role in influencing their children's motivational beliefs (Jacobs and Eccles, 2000). Recent empirical work has corroborated this notion by showing that parents' beliefs in educational domains are closely linked to the beliefs and behaviors of their children, and that parental involvement is a strong predictor of student outcomes (Jodl et al., 2001; Simpkins et al., 2012). This work suggests that parents may be in a good position to influence their children's perceptions of utility for school topics.

Harackiewicz et al. (2012) tested whether it is possible to influence students' perceptions of utility value by intervening with parents in a randomized field study. Specifically, they tested whether intervening with parents could be an effective way of indirectly promoting students' perceptions of utility value, which in turn could increase mathematics and science course enrollment. This issue is particularly important in the United States where we have a disturbing trend of high school students electing not to take advanced mathematics and science courses. Only 12% of high school graduates have taken calculus, and only 29% take physics (National Science Board, 2004). Since the publication of "A Nation at Risk" in 1983 (National Commission on Excellence in Education), reinvigorating interest in STEM subjects has been part of the national agenda. Educators are worried that the increase in demand of STEM jobs in America will outweigh the supply of viable candidates; and if the United States wants to compete in a global market, it is critical that more American students pursue STEM careers (National Research Council, 2007). As recently as April 2013 the Obama administration committed $3.1 billion to improve STEM education nationwide, with $450 million being directed at boosting the number of trained educators and developing programs to inspire students to pursue STEM careers. Given all the emphasis on widening the pipeline of students who go onto STEM careers, interventions that promote STEM course-taking in high school have grown increasingly more important.

Harackiewicz et al. (2012) implemented their utility value intervention with two brochures mailed to parents and a Web site that highlighted the utility value of various STEM courses. The intervention was targeted exclusively at parents, with the hope that parents would then communicate utility value information to their teens. This represents an indirect utility value intervention in which parents were given utility value information, with encouragement to communicate that information to their teens, as well as guidance for how to do so. Participants in the randomized experiment were 188 (88 girls, 100 boys) adolescents from 108 different high schools and their parents. The first brochure, titled "Making Connections: Helping Your Teen Find Value in School," was mailed to parents of 10th grade students in October. The brochure provided information about the utility of mathematics and science in daily life and for various careers. In addition to the utility value information, the brochure included guidance for parents on how to talk to their children about connections between mathematics, science, and their children's lives.

The second brochure, titled "Making Connections: Helping Your Teen with the Choices Ahead," was sent to each parent of 11th grade students separately in January, and included a password-protected Web site titled "Choices Ahead." Similar to the first brochure, the second one emphasized the connections between mathematics and science to people's lives as well as the importance of conveying these connections to students. The second brochure was different from the first in that it placed an increased emphasis on the relevance of STEM courses for preparing students for college and future careers.

The Web site contained clickable links to a wealth of resources about STEM fields and careers in addition to interesting science sites that described the relevance of STEM topics to everyday life. The Web site also presented excerpts of interviews with current college students who highlighted the importance of their high school mathematics and science courses for their college preparation. Parents were also given the option of e-mailing specific links from the Web site to their teens. Parents in the control group did not receive either of the brochures or access to the Web site.

Self-report measures of parents' perceived utility of mathematics and science for their teens (e.g., "Math and science are important for my teen's life") were recorded at two separate points: once when the students were in 9th grade (prior to the intervention) and once when the students were in 11th grade (after the intervention materials were issued). Following 12th grade, students and parents each completed a survey assessing the extent to which parents and teens had engaged in conversations about the importance of mathematics and science, and teens provided self-reports of the perceived utility value of mathematics and science. The main hypothesis was that students whose parents had received the intervention would enroll in more advanced mathematics and science courses.

The results of this relatively simple intervention were dramatic. Harackiewicz et al. (2012) found that students whose parents were in the experimental condition enrolled in significantly more mathematics and science courses in 11th and 12th grade than their peers whose parents were in the control group. The difference was equivalent to nearly an extra semester of mathematics or science over a 2-year period. They also found, consistent with previous findings (Jodl et al., 2001; Simpkins et al., 2006), that parental education was a significant predictor of STEM course-taking in high school. The two effects were independent, and the size of the intervention effect ($\beta = 0.16$) was comparable to the effect of parental education ($\beta = 0.17$) (**Figure 1**).

Additional analyses indicated that the intervention significantly increased mothers' perceptions of STEM utility value for their teens as well as students' reports of conversations with parents about the importance of mathematics and science.

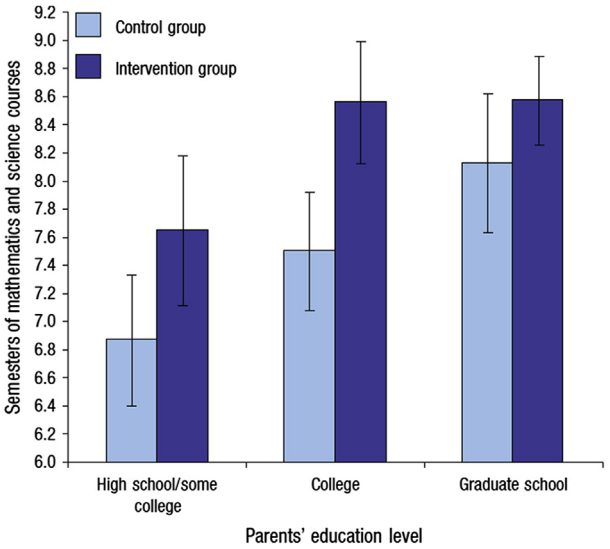

Figure 1 Number of semesters of mathematics and science courses that students took in the last 2 years of high school (as reported on their transcripts) as a function of parents' education level and experimental condition. Error bars represent ±1 Standard Error.

Thus the intervention appears to have been effective in changing parental values and in promoting conversations about the value of STEM disciplines. Process analyses elaborated on these key findings by showing that these direct effects of the intervention on mothers' perceptions of STEM utility value and students' reports of conversations about the importance of STEM were associated with indirect effects indicating that both mother values and teen reports of conversations were significant predictors of students' perceptions of STEM utility value after graduation. Overall, these results suggest that an intervention that targeted parents had direct effects on their teens' STEM course-taking in high school, and indirect effects on their teens' perceptions of STEM utility value. This suggests that parents, a largely untapped resource, can and should be viewed as powerful agents in the promotion of STEM motivation.

Apart from suggesting that future educational intervention research should consider targeting parents, Harackiewicz et al.'s (2012) study indicates that the onus of inspiring motivation does not lie solely on teachers or educational programs. There are numerous avenues for inspiring students, and parents have the ability to play an active role in that process. The fact that perceived utility value is a powerful motivator and a malleable construct suggests that there are numerous interventions that can work to motivate students. While we have addressed several utility value interventions in this article, there are certainly other effective ways to increase perceived utility value that future research should investigate.

Conclusion

The positive effects of utility value interventions reviewed here suggest that perceived utility value is an important construct and deserving of a prominent role in the motivational and educational literature. These findings suggest that educators and parents should try to enhance the perceived utility value of academic disciplines for students. While we know that an increase in perceived utility value has positive educational outcomes, research has yet to determine the *most* effective way of enhancing this value. The answer is likely complicated. As evidenced by the work of Durik and Harackiewicz (2007) and Hulleman and colleagues (Hulleman and Harackiewicz, 2009; Hulleman et al., 2010), different utility value interventions may be more effective for different types of people (Shechter et al., 2011; Durik et al., in press; Durik, et al., 2014). Moreover, the research by Harackiewicz et al. (2012) adds a different twist: in some cases, it may also be important to influence the perceived utility value of the person (e.g., parent) who communicates utility value to the target (e.g., teen, student). One avenue for future research is to investigate the context in which utility value information is presented (classrooms, family conversations, peer conversations) as well as the medium through which it is discovered (information from teachers, personal essays, informal interactions), which may provide insight into the development of even more effective interventions.

See also: Academic Engagement; Expectancy-Value-Cost Model of Motivation; Interest, Psychology of; Motivation and Actions, Psychology of; Motivation, Familial Influences on; Motivation, Learning, and Instruction; School Achievement: Motivational Determinants and Processes; Self-Determination Theory.

Bibliography

Ainley, M., Hidi, S., Berndorff, D., 2002. Interest, learning, and the psychological processes that mediate their relationship. Journal of Educational Psychology 94, 545–561.

Bergin, D.A., 1999. Influences on classroom interest. Educational Psychologist 34, 87–98.

Berlyne, D.E., 1960. Conflict, Arousal, and Curiosity. McGraw-Hill Book Company, New York, NY, USA.

Deci, E.L., Koestner, R., Ryan, R.M., 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin 125, 627–668.

Deci, E.L., Ryan, R.M., 1985. The general causality orientations scale: self-determination in personality. Journal of Research in Personality 19, 109–134.

De Volder, M.L., Lens, W.W., 1982. Academic achievement and future time perspective as a cognitive–motivational concept. Journal of Personality and Social Psychology 42, 566–571.

Durik, A.M., Harackiewicz, J.M., 2007. Different strokes for different folks: how individual interest moderates the effects of situational factors on task interest. Journal of Educational Psychology 99, 597–610.

Durik, A.M., Hulleman, C.S., Harackiewicz, J.M., in press. One size fits some: instructional enhancements to promote interest don't work the same for everyone. In: Renninger, K.A., Nieswandt, M. (Eds.), Interest, the Self, and K-16 Mathematics and Science Learning. American Educational Research Association, Washington, DC.

Durik, A.M., Shechter, O.G., Noh, M., Rozek, C.S., Harackiewicz, J.M., 2014. What if I can't? Success expectancies moderate the effects of utility value information on situational interest and performance. Motivation and Emotion 1–15.

Eccles, J., 2009. Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. Educational Psychologist 44, 78–89.

Eccles, J.S., Adler, T.F., Futterman, R., Geoff, S.B., Kaczala, C.M., Meece, J.L., Midgley, C., 1983. Expectancies, values, and academic behaviors. In: Spence, J.T. (Ed.), Achievement and Achievement Motives. W.H. Freeman and Company, San Francisco, pp. 75–146.

Harackiewicz, J.M., 1979. The effects of reward contingency and performance feedback on intrinsic motivation. Journal of Personality and Social Psychology 37, 1352–1363. http://dx.doi.org/10.1037/0022-3514.37.8.135.

Harackiewicz, J.M., Barron, K.E., Tauer, J.M., Carter, S.M., Elliot, A.J., 2000. Short-term and long-term consequences of achievement goals: predicting interest and performance over time. Journal of Educational Psychology 92, 316–330.

Harackiewicz, J.M., Barron, K.E., Tauer, J.M., Elliot, A.J., 2002. Predicting success in college: a longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. Journal of Educational Psychology 94, 562–575.

Harackiewicz, J.M., Durik, A.M., Barron, K.E., Linnenbrink-Garcia, L., Tauer, J.M., 2008. The role of achievement goals in the development of interest: reciprocal relations between achievement goals, interest, and performance. Journal of Educational Psychology 100, 105–122.

Harackiewicz, J.M., Hulleman, C.S., 2010. The importance of interest: the role of achievement goals and task values in promoting the development of interest. Social and Personality Psychology Compass 4, 42–52.

Harackiewicz, J.M., Rozek, C.R., Hulleman, C.S., Hyde, J.S., 2012. Helping parents motivate adolescents in mathematics and science: an experimental test. Psychological Science 43, 899–906.

Harter, S., 1981. A new self-report scale of intrinsic versus extrinsic orientation in the classroom: motivational and informational components. Developmental Psychology 17, 300–312.

Harter, S., 2006. The self. In: Damon, W., Lerner, R.M., Eisenberg, N. (Eds.), Handbook of Child Psychology: Social, Emotional, and Personality Development, vol. 3. Wiley, New York, pp. 505–570.

Hidi, S., Baird, W., 1988. Strategies for increasing text-based interest and students' recall of expository texts. Reading Research Quarterly 23, 465–483.

Hidi, S., Harackiewicz, J.M., 2000. Motivating the academically unmotivated: a critical issue for the 21st century. Review of Educational Research 70, 151–179.

Hidi, S., Renninger, K., 2006. The four-phase model of interest development. Educational Psychologist 41, 111–127.

Hulleman, C.S., Durik, A.M., Schweigert, S.B., Harackiewicz, J.M., 2008. Task values, achievement goals, and interest: an integrative analysis. Journal of Educational Psychology 100, 398–416.

Hulleman, C.S., Godes, O., Hendricks, B.L., Harackiewicz, J.M., 2010. Enhancing interest and performance with a utility value intervention. Journal of Educational Psychology 102, 880–895.

Hulleman, C.S., Harackiewicz, J.M., 2009. Promoting interest and performance in high school science classes. Science 326, 1410–1412.

Iyengar, S.S., Lepper, M.R., 1999. Rethinking the value of choice: a cultural perspective on intrinsic motivation. Journal of Personality and Social Psychology 76, 349–366.

Jacobs, J.E., Eccles, J.S., 2000. Parents, task values, and real-life achievement-related choices. In: Sansone, C., Harackiewicz, J.M. (Eds.), Intrinsic and Extrinsic Motivation: The Search for Optimal Motivation and Performance. Academic Press, San Diego, pp. 405–439.

Jodl, K.M., Michael, A., Malanchuk, O., Eccles, J.S., Sameroff, A., 2001. Parents' roles in shaping early adolescents' occupational aspirations. Child Development 72, 1247–1265.

Lepper, M.R., Cordova, D.I., 1992. A desire to be taught: instructional consequences of intrinsic motivation. Motivation and Emotion 16, 187–208.

Lepper, M.R., Greene, D., Nisbett, R.E., 1973. Undermining children's intrinsic interest with extrinsic reward: a test of the 'overjustification' hypothesis. Journal of Personality and Social Psychology 28, 129–137.

Lepper, M.R., Henderlong, J., 2000. Turning 'play' into 'work' and 'work' into 'play': 25 years of research on intrinsic versus extrinsic motivation. In: Sansone, C., Harackiewicz, J.M. (Eds.), Intrinsic and Extrinsic Motivation: The Search for Optimal Motivation and Performance. Academic Press, San Diego, pp. 257–307.

Malka, A., Covington, M.V., 2005. Perceiving school performance as instrumental to future goal attainment: effects on graded performance. Contemporary Educational Psychology 30, 60–80.

National Center for Education Statistics, 2011. Digest of Education Statistics, 2010. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011015.

National Commission on Excellence in Education, 1983. A Nation at Risk: The Imperative for Educational Reform: A Report to the Nation and the Secretary of Education. United States Department of Education. The Commission, Washington, DC.

National Research Council, 2007. Rising above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future. The National Academies Press, Washington, DC.

National Science Board, 2004. Science and Engineering Indicators 2004: Elementary and Secondary Education, Mathematics and Science Coursework and Student Achievement (NSB 04-01). Retrieved from: www.nsf.gov/statistics/seind04/.

Pajares, F., 1996. Self-efficacy beliefs in academic settings. Review of Educational Research 66, 543–578.

Renninger, K., 2000. Individual interest and its implications for understanding intrinsic motivation. In: Sansone, C., Harackiewicz, J.M. (Eds.), Intrinsic and Extrinsic Motivation: The Search for Optimal Motivation and Performance. Academic Press, San Diego, pp. 373–404.

Renninger, K., Hidi, S., 2011. Revisiting the conceptualization, measurement, and generation of interest. Educational Psychologist 46, 168–184.

Schiefele, U., 1991. Interest, learning, and motivation. Educational Psychologist 26, 299.

Schraw, G., Dennison, R., 1994. The effect of reader purpose on interest and recall. Journal of Reading Behavior 26, 1–18.

Shechter, O.G., Durik, A.M., Miyamoto, Y., Harackiewicz, J.M., 2011. The role of utility value in achievement behavior: the importance of culture. Personality and Social Psychology Bulletin 37, 303–317.

Simons, J., Vansteenkiste, M., Lens, W., Lacante, M., 2004. Placing motivation and future time perspective theory in a temporal perspective. Educational Psychology Review 16, 121–139.

Simpkins, S.D., Davis-Kean, P., Eccles, J.S., 2006. Math and science motivation: a longitudinal examination of the links between choices and beliefs. Developmental Psychology 42, 70–83.

Simpkins, S.D., Fredricks, J.A., Eccles, J.S., 2012. Charting the Eccles' expectancy-value model from mothers' beliefs in childhood to youths' activities in adolescence. Developmental Psychology 48, 1019–1032.

Van Calster, K., Lens, W., Nuttin, J.R., 1987. Affective attitude toward the personal future: impact on motivation in high school boys. The American Journal of Psychology 100, 1–13.

Vansteenkiste, M., Lens, W., Deci, E.L., 2006. Intrinsic versus extrinsic goal contents in self-determination theory: another look at the quality of academic motivation. Educational Psychologist 41, 19–31.

Wigfield, A., 1994. Expectancy-value theory of achievement motivation: a developmental perspective. Educational Psychology Review 6, 49–78.

Wigfield, A., Cambria, J., 2010. Students' achievement values, goal orientations, and interest: definitions, development, and relations to achievement outcomes. Developmental Review. 30, 1–35.

Wigfield, A., Eccles, J., 1992. The development of achievement task values: a theoretical analysis. Developmental Review 12, 265–310.

Academic Performance, Effects of Socio-Economic Status on

Brandon L Carlisle and Carolyn B Murray, University of California Riverside, Riverside, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The purpose of this article is to review the relationship between socioeconomic status and academic achievement. The complexity of this relationship will be examined by discussing factors such as socioeconomic segregation, school funding, teacher expectations, and academic climate. The effect that socioeconomic status can have on academic achievement is of global importance to educators, researchers, and policymakers, as they continue to address disparities in academic achievement and educational attainment. Efforts to ensure that all students receive an equitable educational experience are also discussed, with particular emphasis placed on the importance of effective schools.

Introduction

Accounting for differences in academic achievement among individuals and social groups has received considerable global attention from educators, researchers, and policymakers. Efforts to understand and account for these differences have involved an examination of individual student characteristics and the characteristics of their school environment (e.g., Coleman et al., 1966; Marjoribanks, 2003; Palardy, 2013). Socioeconomic status (SES) has been identified as one of the most commonly used contextual variables within research on academic achievement (Sirin, 2005). Research on SES and academic achievement has revealed a consistent relationship (Milne and Plourde, 2006); specifically, it is often the case that high SES is associated with greater academic achievement. Furthermore, the educational literature has provided evidence that SES is one of the strongest predictors of academic achievement and educational attainment (Reardon, 2011).

White (1982) outlined the most frequent applications of SES found in the educational literature. First, SES may be used as a covariate within quasiexperimental studies; in some cases, SES may not be of interest, although it relates to variables of primary interest, therefore researchers may choose to statistically control or adjust significant SES differences within a given sample. Second, SES can be used to improve the precision of experimental studies by controlling for it in analyses of covariance. Third, researchers can assess potential interaction effects, such as examining the possibility that a particular teaching method may be effective among high SES students but not among low SES students. Fourth, SES can be utilized to provide additional demographic information, which can facilitate efforts to determine external validity and accurately generalize empirical findings. Finally, SES can be used as a predictor variable when testing causal models of academic achievement and educational attainment.

While most researchers are in agreement regarding the importance of SES, White (1982) points out that SES has been defined and subsequently measured in a variety of ways across an abundance of empirical studies. Common indicators of socioeconomic status include income, occupation, and education (Sirin, 2005; White, 1982). Chapin (1928) defined SES as "the position an individual or family occupies with reference to the prevailing average of standards" (p. 99). Consistent with Chapin's definition, Mueller and Parcel (1981) much later described SES as an individual's or family's position in a societal hierarchy that dictates the degree to which the individual or family has access to wealth and power. Although the relationship between SES and academic achievement has been studied for many decades, the social sciences are continuing to address complex research questions pertaining to student level factors, school level factors, and the nature of the relationship within countries of developing or developed status. The purpose of this article is to review the relationship between SES and academic achievement and to discuss the implications of this relationship on educational attainment.

Country Income

The relationship between SES and academic achievement has been studied at a cross-national level as equitable education persists as an international concern. Researchers have progressively shown a strong interest in international comparative studies (Wiseman and Baker, 2005). For example, Heyneman and Loxley (1982, 1983) conducted a comparative study of high-income and low-income countries in order to further examine the nature of the relationship between SES and academic achievement. In high-income countries, they found evidence of a stronger association between individual student SES and academic achievement and a relatively weaker association between school-level factors (e.g., school and teacher quality) and academic achievement. However, their analyses also indicated that within low-income countries, school-level factors shared a relatively stronger association with academic achievement than did individual student SES.

In the literature, these contrasting results are often referred to as the *Heyneman–Loxley Effect* (HL Effect; Baker et al., 2002). The nature of the HL effect challenged the perception that weaker associations exist between school-level factors and achievement and stronger associations exist between individual student SES and achievement (Baker et al., 2002). Heyneman and Loxley (1983) concluded that the quality of schools and its teachers significantly influence the academic outcomes of students and within poorer countries, this effect seems to be more pronounced.

Several studies have been dedicated to replicating the HL effect but the results have not always been consistent. For

example, Riddell (1989) critiqued the HL effect by addressing an important methodological concern; specifically, analyses that revealed the effect relied on ordinary least squares regression, which did not take into account the hierarchical structure of the data (i.e., students nested within classrooms, classrooms nested within schools). Years later, Riddell (1997) reviewed 16 studies that implemented multilevel modeling in order to study academic achievement in developing countries and did not find evidence of the HL effect. Baker et al. (2002) also did not find evidence of the HL effect but posited that the effect may have diminished over time due to the increased efforts of poorer countries to invest in formal education as an economic strategy.

Heyneman (2005) argued that any failures to replicate the HL effect could be attributed to significant differences in samples of countries across studies. Heyneman also stressed that the influence of student SES is not consistent across societies and that it may differ due to grade level of students, gender composition of students, and how academic achievement is operationalized. While Baker et al. (2002) did not find evidence of the HL effect, a more recent cross-national study by Chiu (2010) found evidence that the association between individual student SES and achievement was stronger in richer countries, which is consistent with the HL effect.

Ethnicity

Ethnicity and SES are closely related; so much so that it can become difficult to disentangle the two constructs (Jussim et al., 1996). Regarding the student characteristics of the studies reviewed by White (1982) and Sirin (2005), it was found that minority status moderated the relationship between SES and academic achievement. Specifically, the mean effect size for Caucasian students was significantly larger than the mean effect size for minority students. This suggests that parental education, income, and occupation may be less related to achievement for minorities than issues such as the environment, the neighborhood and/or the school SES (Sirin, 2005).

When examining the relationship between SES and academic achievement, it is often the case that researchers need to consider the minority status or ethnicity of the student. The educational environment in which students find themselves can be extremely influential in their development. Unfortunately, there is disparate access to affluent educational environments; previous research has indicated that low-income, minority students are often subject to ineffective teachers, low teacher expectations, inadequate teaching aides, and defiant peers (Conchas, 2001). In the US, African Americans and Latinos are more likely to be members of low-SES families and attend low-SES schools (Williams and Collins, 2001); it is also the case that African Americans and Latinos tend to be underrepresented in higher education (Oakes et al., 2004). In Taiwan, Aborigines are a minority ethnic group that holds a disadvantaged socioeconomic position; they also tend to have lower academic achievements in comparison to Taiwanese students (Sung et al., 2013). Findings such as these illustrate the ecological validity of the relationship between SES, academic achievement, and ethnicity.

The ethnicity of students is an important contextual variable that can provide meaningful information about students; such information is relevant to research and application. Ideally, future research and *metaanalytical* studies will be able to code for multiple ethnic groups rather than only code for minority status. For example, it would be useful to determine if a specific ethnicity is associated with significantly different outcomes in comparison to other ethnic groups within a given country.

Immigrant Status

Similar to ethnicity, SES is an important factor when accounting for differences in academic achievement between native and immigrant students. Schnepf (2004) found a positive correlation between SES gaps between natives and immigrants and achievement gaps in reading scores. Relatedly, Buhlmahn (2003, as cited in Sung et al., 2013) found that students of immigrant status had lower academic achievements than native German students. Similarly, in a comparison of native students and low-SES, second-generation students, Alba et al. (2011) presented data that depicts achievement gaps in reading and mathematics within European countries and the US.

Understanding the influence of immigrant status and its relation to SES is a complex matter. The selection of an appropriate measure of SES requires careful consideration. For example, Schnepf (2004) argues that the SES indicator of parental education is problematic when comparing native and immigrant students because the quality of parental education may not be equal across countries. Furthermore, Schnepf also points out that it is important to consider an immigrant's country of origin, migration motives, and integration into the receiving country. Each of these factors may provide important contextual information when understanding the interrelationships between SES, academic achievement, and immigrant status.

Teachers

The quality and effectiveness of teachers can be very influential on student performance. The importance of teachers cannot be overlooked as they can directly influence student achievement. Teacher effectiveness and quality relate to SES because students of low-SES, low achievement, and minority status are less likely to be exposed to highly qualified, effective teachers (Lankford et al., 2002). This is due, in part, to the tendency for more qualified teachers to eventually seek employment in schools that have high-achieving students, greater resources, and strong administrative support (Hanushek et al., 2004; Rumberger and Palardy, 2005).

Research investigating the effects of teacher quality on academic performance has demonstrated that having access to highly qualified teachers can have a significant impact on student achievement. For example, in a cross-national study of teacher quality and mathematics achievement, Akiba et al. (2007) found that countries containing a higher percentage of students taught by highly qualified mathematics teachers (i.e., having earned a degree in mathematics, teacher certification, and at least 3 years experience) achieved a significantly higher national average mathematics score.

They also examined opportunity gaps in access to qualified teachers. Opportunity gaps were calculated by determining the percentage of high-SES students that have access to qualified teachers, determining the percentage of low-SES students that have access to qualified teachers, then calculating the difference between these two values. The largest opportunity gaps between high-SES and low-SES students were found in the countries of Syria, Chile, Taiwan, the US, and Hong Kong. Any opportunity gaps in access to qualified teachers can create a learning opportunity gap; this is an added disadvantage for low-SES students who may already experience a significant resource gap (Akiba et al., 2007).

An additional factor relevant to the issue of teachers is the influence of their expectations on student performance. Rosenthal and Jacobson (1968) demonstrated that teacher expectations could have a definite influence on the academic performance of students; specifically, if teachers have high expectations of their students, this can translate into stronger student performance. To further explore this outcome, Harris and Rosenthal (1985) conducted *metaanalyses* on the effects of teachers' expectations on the intellectual functioning of their students and discovered not only that these interpersonal expectancy effects occur but that their magnitudes are of substantial practical importance as well.

These *metaanalyses* supported the importance of behaviors that teachers may exhibit (e.g., praise, encouragement, long interactions, smiles) that mediate expectancy effects. In addition, these analyses provided support for a theoretical framework involving the mediation of interpersonal expectancy effects. Specifically, teachers who hold positive expectations for a student will tend to display a warmer affect, express more positive feedback, provide more input with regards to the quantity and difficulty of material that is taught, and increase the amount of student participation by offering more response opportunities (Harris and Rosenthal, 1985).

This research is pertinent to the discussion of SES because previous research has indicated that teachers hold higher expectations and more positive attitudes for high-SES students, in comparison to their low-SES counterparts (e.g., Auwarter and Aruguete, 2008). Furthermore, in relation to SES, Jussim et al. (1996) found that students of low-SES backgrounds were more susceptible to expectancy effects. They posited that low-SES students may have "reduced social and psychological resources for combating erroneous teacher expectations" (p. 371).

Academic Climate

The academic climate a student experiences can also have an impact on the relationship between SES and academic achievement. Evaluating academic climate may consist of determining the types of messages students receive from their peers, teachers, and administrators with regard to academic achievement and educational attainment. It can also involve an assessment of school safety, the average number of hours spent completing homework per week, and the average number of college prep courses taken (Rumberger and Palardy, 2005). Additionally, a more comprehensive approach to understanding academic climate has considered factors both in and out of school (e.g., home atmosphere, neighborhood unemployment rate).

Lee et al. (1999) studied how academic achievement is related to student social support and school academic press. They defined student social support as the personal relationships between a student and individuals (e.g., peers, parents, and teachers) who are able to provide them with encouragement and assistance in and out of school. Academic press was defined as the extent to which school members (e.g., students, teachers, and administrators) emphasize academic success and satisfying standards of achievement. Their analyses revealed that social support and academic press were positively correlated with academic achievement in the subject areas of mathematics and reading. Furthermore, the constructs of academic press and social support were related to SES; specifically, they found that students who attended low-SES schools were least likely to experience the ideal combination of high social support and high academic press.

Socioeconomic Segregation

Throughout the world, there are students belonging to affluent families that receive a distinguished education and students belonging to low-income families that receive a substandard education. These disparate educational experiences contribute to the academic achievement gap that has been a source of concern for decades. The existence of socioeconomic segregation should be of concern because of its relationship to individual student performance (e.g., Bankston and Caldas, 1996; Rumberger and Palardy, 2005). In their analysis of cohort data, Duncan et al. (2013) found that the familial income gap between low-SES children and high-SES children accounted for significant amounts of the academic achievement gap. According to Duncan and Murnane's (2011) conceptual model, high-SES children are more likely to have greater access to high quality child care, schools, and settings that enhance the development of important skills related to education while low-SES children may have parents who cannot afford such resources.

In reaction to the unequal educational experiences associated with socioeconomic segregation, socioeconomic integration (e.g., low-status immigrants attending more affluent schools, low-SES students attending high-SES schools) has been discussed as a potential means to reduce the achievement gap and improve equity in education (e.g., Alba et al., 2011; Ryan and Heise, 2002). Rumberger and Palardy (2005) conducted a study to further investigate the effect of socioeconomic segregation on academic performance. Their analyses were conducted on hierarchical data that included students nested within schools as they were followed from Grades 8 through 12. They found that the average SES of a student's school had a near equal influence on achievement as individual student SES. Additionally, school SES had a similar influence on low-SES students and high-SES students.

Further analysis revealed that the effect of school SES on academic outcomes was no longer significant after controlling for school policies and practices (e.g., teacher expectations, academic climate). Based on these results, Rumberger and Palardy concluded that socioeconomic integration may not be

necessary, provided school policies and practices can be reformed. It may be the case that improving low-SES schools in this fashion would be more effective and realistic, in comparison to mandated socioeconomic integration.

Disparate Funding

The inequity created by socioeconomic segregation is related to the amount of funding a school district receives as well as the amount of funding a school receives from their respective district. Governing bodies are faced with the responsibility of ensuring that schools receive sufficient funding so they can provide equitable, effective education to all of its students (Augenblick et al., 1997). This objective can become more challenging for low-income countries; the amount of funding schools receive can depend on the wealth of a nation and how they disperse public expenditures on schooling (Colclough and Al-Samarrai, 2000).

In their review of school funding equity, Augenblick et al. (1997) demonstrate that SES is associated with school funding by providing evidence that school districts with wealthier residents tend to receive significantly higher per-pupil funding. It is also the case that disparities in funding can exist within a school district. Condron and Roscigno (2003) conducted a study that included urban elementary schools of diverse racial and SES composition. They found that within districts, instructional per-pupil expenditure and percentage of students eligible for free or price-reduced lunch were negatively correlated. Further, they found that operations and maintenance per-pupil expenditure were also negatively correlated with the percentage of students eligible for the free or price-reduced lunch program. This provides evidence that it may be insufficient to only address disparate school funding between districts; disparate school funding within districts is also worthy of close attention and resolution.

Augenblick et al. (1997) provided recommendations as to how districts should proceed in funding its schools. These recommendations include: districts establishing a base level of per-pupil funding that is consistent with expectations of academic achievement; governing bodies distributing funding to districts and then districts distributing funding to schools, which should be based on relative needs such as students enrolled in high-cost programs, prevalence of students at risk of failure, and the size of the school; and governing bodies providing equal support for the construction and maintenance of school facilities.

For example, within the United States, a potential solution could involve states being encouraged to ensure more equitable funding through incentives provided by the federal government; the establishment of a more significant federal role in financing education is a matter of collective and political will (U.S. Department of Education, 2013). Among other recommendations, the U.S. Department of Education's Equity and Excellence Commission calls for new federal funding that should be allocated to schools with high concentrations of low-SES students and the expansion of the government's authority to ensure the equity of school financing. Additional funding to schools with high concentrations of low-SES students could facilitate efforts to accrue new educational resources and, if necessary, reform the policies and practices of these schools.

Intervention

Closing the achievement gaps has become an important issue for many countries (Sung et al., 2013). Reform efforts have been discussed and implemented throughout the world; these efforts aim to improve the education practices believed to influence learning for low-performing students, and thus decrease the achievement gap (Feuer et al., 2002; Hargreaves, 2000; Pressley et al., 2004). One such program, developed by the United States is the 'no child left behind' (NCLB) Act, which was implemented in order to aid disadvantaged children. This program expanded the federal role in public education by rewarding schools that set high standards and established measurable goals to improve individual outcomes in education. While the U.S. Department of Education reported that student achievement had improved in reading and math, critics of the program argued that these statistics were misleading. One criticism was that creators of the standardized tests were accused of making the assessments less challenging so that it was easier for schools to significantly improve.

While there is debate surrounding the effectiveness of intervention programs, some interventions seem to have narrowed the achievement gap between high and low socioeconomic groups. For instance, the Finnish government has significantly invested in its country's education and the implementation of educational policies that include comprehensive schooling with no tuition or book fees for its students (Lavonen and Laaksonen, 2009). According to the findings of the Program for International Student Assessment (PISA), the achievement gap in Finland has significantly reduced over time.

Still other interventions have not had a significant impact on academic outcomes and disparities in performance. For example, the French government implemented the Zone of Education Priority program; this intervention established new teaching projects and distributed more resources to schools in low-SES areas (Sung et al., 2013). In their evaluation of the effectiveness of this program, Bénabou et al. (2009) found evidence to suggest that the program did not have a significant impact on student success, which was measured by several measures (e.g., obtaining a diploma, performance on national examinations). In the United Kingdom, the Ethnic Minority Achievement Grant was developed to reduce the achievement gap between immigrant, minority students, and native English students. An evaluation of its effectiveness revealed that the intervention contributed to a partial reduction in the achievement gap but significant differences between ethnic minority achievement and national average achievement remained (Tikly et al., 2005). In South Korea, the government implemented an education welfare action zone policy that was designed to provide educational resources to low-income areas; this intervention was able to increase positive educational activities among participating schools (i.e., increased student usage of school libraries) however their results did not reveal significant changes in academic outcomes among low-SES students (Lee, 2008).

For many intervention programs, the jury is still out. While SES may complicate the ability to establish a standard of education that prepares all students for success, the characteristics of a successful school can be achieved if teachers and administrators are willing to exert the necessary effort. See Edmonds (1979) and other researchers (e.g., Purkey and Smith, 1983) for a discussion. Although these researchers acknowledge that family SES does indeed make a difference, they were able to identify schools where the student populations were comprised of low SES students who were achieving well academically. Edmonds identified the following seven characteristics that these schools had in common: (1) strong administrative leadership, (2) clear school mission, (3) high expectations for student success, (4) frequent monitoring of school performance, (5) focus on basic skills, (6) safe and orderly environment, and (7) a focus on home–school relations. These characteristics can serve as clear tenets for an educational system that needs to remedy disparities in academic experiences and outcomes. Furthermore, if schools and faculty members are committed, it is possible that these characteristics can be met without significant changes to school funding. In the words of one highly successful educator:

> Effective schools do not attempt to eliminate the achievement gap that separates low-SES children from high-SES children: they raise the floor below which nobody falls.
>
> Ronald Edmonds (1986: p. 101)

To date, globally the achievement gap between high- and low-SES students still persists. This should not come as a surprise since none of the countries attempting to reduce the SES achievement gap have implemented all seven characteristics of 'effective schools.'

See also: Academic Achievement Motivation, Development of; Behavioral Problems, Effects of Socio-Economic Status On; Gender and Academic Motivation; Inequality, Social; Language Development: Influence of Socio-Economic Status; Race and Academic Motivation; Socioeconomic Status and Social-Emotional Development of Children from Birth to 36 Months of Age: A Systematic Narrative Review.

Bibliography

Akiba, M., LeTendre, G.K., Scribner, J.P., 2007. Teacher quality, opportunity gap, and national achievement in 46 countries. Educational Researcher 36, 369–387.

Alba, R., Sloan, J., Sperling, J., 2011. The integration imperative: the children of low-status immigrants in the schools of wealthy societies. Annual Review of Sociology 37, 395–415.

Augenblick, J.G., Myers, J.L., Anderson, A.B., 1997. Equity and adequacy in school funding. The Future of Children 7, 63–78.

Auwarter, A.E., Aruguete, M.S., 2008. Effects of student gender and socioeconomic status on teacher perceptions. The Journal of Educational Research 101, 243–246.

Baker, D.P., Goesling, B., LeTendre, G.K., 2002. Socioeconomic status, school quality, and national economic development: a cross-national analysis of the "Heyneman-Loxley Effect" on mathematics and science achievement. Comparative Education Review 46, 291–312.

Bankston, C., Caldas, S.J., 1996. Majority African American schools and social injustice: the influence of de facto segregation on academic achievement. Social Forces 75, 535–555.

Bénabou, R., Krammarz, F., Prost, C., 2009. The French zone d'education prioritaire: much ado about nothing? Economics of Education Review 28, 345–356.

Chapin, F.S., 1928. A quantitative scale for rating the home and social environment of middle class families in an urban community: a first approximation to the measurement of socio-economic status. Journal of Educational Psychology 19, 99–111.

Chiu, M.M., 2010. Effects of inequality, family, and school on mathematics achievement: country and student differences. Social Forces 88, 1645–1676.

Colclough, C., Al-Samarrai, S., 2000. Achieving schooling for all: budgetary expenditure on education in Sub-Saharan African and South Asia. World Development 28, 1927–1944.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfield, F.D., et al., 1966. Equality of Educational Opportunity. U.S. Government Printing Office, Washington, DC.

Conchas, G.Q., 2001. Structuring failure and success: understanding the variability in Latino school engagement. Harvard Educational Review 71, 475–504.

Condron, D.J., Roscigno, V.J., 2003. Disparities within: unequal spending and achievement in an urban school district. Sociology of Education 76, 18–36.

Duncan, G.J., Kalil, A., Ziol-Guest, K.M., 2013. Increasing inequality in parent incomes and children's completed schooling: correlation or causation? Unpublished manuscript presented at the California Population Research Seminar Series. Retrieved from: http://www.hks.harvard.edu/inequality/Seminar/Papers/Duncan13.pdf.

Duncan, G., Murnane, R., 2011. Whither Opportunity? Rising Inequality and the Uncertain Life Chances of Low-Income Children. Russell Sage, New York.

Edmonds, R., 1986. Characteristics of effective schools. In: Neisser, U. (Ed.), The School Achievement of Minority Children: New Perspectives, Lawrence Erlbaum, Hillsadle, pp. 93–104.

Edmonds, R., 1979. Effective schools for the urban poor. Educational Leadership 37, 15–27.

Feuer, M.J., Towne, L., Shavelson, R.J., 2002. Scientific culture and educational research. Educational Researcher 31, 4–13.

Hanushek, E.A., Kain, J.F., Rivkin, S.G., 2004. Why public schools lose teachers. Journal of Human Resources 39, 326–354.

Hargreaves, D.H., 2000. Teaching as a research-based profession: possibilities and prospects. In: Moon, B., Butcher, J., Bird, E. (Eds.), Leading Professional Development in Education. Routledge-Falmer, London.

Harris, M.J., Rosenthal, R., 1985. Mediation of interpersonal expectancy effects: 31 meta-analyses. Psychological Bulletin 97, 363–386.

Heyneman, S.P., 2005. Student background and student achievement: what is the right question? American Journal of Education 112, 1–9.

Heyneman, S.P., Loxley, W.A., 1982. Influences on academic achievement across high and low income countries: a re-analysis of IEA data. Sociology of Education 55, 13–21.

Heyneman, S.P., Loxley, W.A., 1983. The effect of primary-school quality on academic achievement across twenty-nine high- and low-income countries. American Journal of Sociology 88, 1162–1194.

Jussim, L., Eccles, J., Madon, S., 1996. Social perception, social stereotypes, and teacher expectations: accuracy and the quest for the powerful self-fulfilling prophecy. Advances in Experimental Psychology 28, 281–388.

Lankford, H., Loeb, S., Wyckoff, J., 2002. Teacher sorting and the plight of urban schools: a descriptive analysis. Educational Evaluation and Policy Analysis 24, 37–62.

Lavonen, J., Laaksonen, S., 2009. Context of teaching and learning school science in Finland: reflections on PISA 2006 results. Journal of Research in Science Teaching 46, 922–944.

Lee, H.Y., 2008. A study on the effects of education welfare action zone policy in Korea. Educational Research for Policy and Practice 7, 35–45.

Lee, V.E., Smith, J.B., Perry, T.E., Smylie, M.A., 1999. Social Support, Academic Press, and Student Achievement: A View from the Middle Grades in Chicago. Chicago Annenberg Challenge, Chicago, IL.

Marjoribanks, K., 2003. Family background, individual and environmental influences, aspirations and young adults' educational attainment: a follow-up study. Educational Studies 29, 233–242.

Milne, A., Plourde, L.A., 2006. Factors of a low-SES household: what aids academic achievement? Journal of Instructional Psychology 33, 183–193.

Mueller, C.W., Parcel, T.L., 1981. Measures of socioeconomic status: alternatives and recommendations. Child Development 52, 13–30.

Oakes, J., Mendoza, J., Silver, D., 2004. California opportunity indicators: informing and monitoring California's progress toward equitable college access. UCACCORD Public Policy Series, Brekely, pp. 1–28.

Palardy, G.J., 2013. High school socioeconomic status segregation and student attainment. American Educational Research Journal 50, 714–754.

Pressley, M., Duke, N.K., Boling, E.C., 2004. The educational science and scientifically based instruction we need: lessons from reading, research, and policymaking. Harvard Educational Review 74, 30–36.

Purkey, S.C., Smith, M.S., 1983. Source effective schools: a review. The Elementary School Journal. Special Issue: Research on Teaching 83, 426–452.

Reardon, S., 2011. The widening achievement gap between the rich and the poor: new evidence and possible explanations. In: Duncan, G.J., Murnane, R.J. (Eds.), Whither Opportunity? Rising Inequality, Schools and Children's Life Chances. Russell Sage, New York, pp. 91–116.

Riddell, A.R., 1989. An alternative approach to the study of school effectiveness in third world countries. Comparative Education Review 33, 481–497.

Riddell, A.R., 1997. Assessing designs for school effectiveness research and school improvement in developing countries. Comparative Education Review 41, 178–204.

Rosenthal, R., Jacobson, L., 1968. Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development. Holt, Rinehart & Winston, New York, NY, USA.

Rumberger, R.W., Palardy, G.J., 2005. Does segregation still matter? The impact of student composition on academic achievement in high school. Teachers College Record 107, 1999–2045.

Ryan, J.E., Heise, M., 2002. The political economy of school choice. The Yale Law Journal 111, 2043–2136.

Schnepf, S.V., 2004. How Different are Immigrants? A Cross-country and Cross-survey Analysis of Educational Achievement. IZA Discussion Paper Series, No. 1398.

Sirin, S.R., 2005. Socioeconomic status and academic achievement: a meta-analytic review of research. Review of Educational Research 75, 417–453.

Sung, Y.-T., Tseng, F.-L., Kuo, N.-P., Chang, T.-Y., Chiou, J.-M., 2013. Evaluating the effects of programs for reducing achievment gaps: a case study in Taiwan. Asia Pacific Education Review 15, 99–113.

Tikly, L., Osler, A., Hill, J., 2005. The ethnic minority achievement grant: a critical analysis. Journal of Education Policy 20, 417–435.

U.S. Department of Education, 2013. For Each and Every Child – a Strategy for Education Equity and Excellence. Washington, DC.

White, K., 1982. The relationship between socioeconomic status and academic achievement. Psychological Bulletin 91, 461–481.

Williams, D.R., Collins, C., 2001. Racial residential segregation: a fundamental cause of racial disparities in health. Public Health Reports 116, 404–416.

Wiseman, A.W., Baker, D.P., 2005. The worldwide explosion of internationalized education policy. International Perspectives on Education and Society 6, 1–21.

Academic Research and Employment: Recent Changes in Europe and the United States

Elke Park and Hans Pechar, Institute of Science Communication and Higher Education Research, Alpen Adria Universität, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Academic research is increasingly carried out on the basis of externally sponsored research projects. This article focuses on the transformative role of new funding mechanisms and its influence on academic work and careers. First, we will outline the increase in externally funded research at higher education institutions in Europe and the United States over the last decades. Second, we will discuss the impact of changes in the funding regime on academic careers and employment conditions, most notably on the postdoctoral phase. We will present criticism of recent developments and conclude with an outline of the corresponding, emerging governance regime in higher education, network governance.

In the following we address profound, ongoing changes in the structure of academic work and research over the last three decades. These changes correspond to more general transformations in the nature of work also affecting other social subsystems: a turn toward project-based, nonpermanent work contracts and more flexible, network-like institutional environments (see Boltanski and Chiapello, 2005). At universities, fixed-term and part-time labor – often externally funded – is replacing tenured full time positions (Schuster and Finkelstein, 2006). Academic research is increasingly carried out in the framework of temporarily delineated projects. In an ever shifting network-like structure, groups or research teams continuously form anew to apply and compete for project funding. The *New Spirit of Capitalism* (Boltanski and Chiapello, 2005) certainly deeply permeated academe since the onset of New Public Management (NPM) policies in the 1980s and 1990s.

The rise of neoliberalist policies in most Western countries since the 1980s has set strict limits on public spending for education and research. Burton Clark argued that "governments increasingly indicate that they are not prepared to pay the unit costs of mass higher education at the level of elite education" (1997: 247). However, central to the idea of New Public Management, or rather the notion of the managerial or entrepreneurial university, is not only the quantitative restriction of public funding – i.e., increased budgetary restraints – but, more importantly, the introduction of qualitative changes in the way these funds are distributed. Under the premises of accountability, research funds are no longer distributed unconditionally, or untargeted, in the form of general university funds. Rather, higher education institutions act as 'applicants' having to qualify and continuously requalify for funding in a competitive process. The distribution of funds for university research via competitively awarded projects and grants can certainly be regarded as one of the central axioms of NPM-inspired approaches to the governance of higher education (for an overview of NPM policies in higher education, see Ferlie et al., 2008).

The vital importance of education for a 'knowledge-based society' and the growing economic relevance of research which resulted in an unprecedented expansion of higher education and research systems did not necessarily lead to better and more favorable conditions for the respective social subsystems. On the contrary, most constituencies regard the changes of the last decades as a burden. This is not only true for students who complain about increasing private contributions to education – as it is the case in the US – or the decrease in quality of mass higher education in European countries retaining open access policies. Academics bemoan deteriorating working conditions and in particular junior academics complain about the difficulty to embark on a stable career. The increased significance of knowledge resulted in qualitative changes in the relationship between higher education and research and public funding bodies.

The Increase and Impact of External Research Funding on Academic Work

> As universities have transformed into modern research complexes, the organizational character and rhythm of academic work changed to accommodate the **increased centrality of externally sponsored research**.
>
> Gumport, 2011: 367

Recent data from the US and Europe illustrate the steep rise in externally sponsored academic research or project-based funding at universities over the last 15 years. For example, in the US the National Science Foundation's support for university research and development more than doubled from 1997 to 2012.[1] In Germany so-called third-party funds almost tripled during that period, and in Austria – a country which experienced a relatively drastic turn toward NPM-inspired university governance in the mid-2000s – a more than fourfold increase in external, project-based university funding took place.[2] At the same time, general university funds remained comparatively stable. In some instances – for example, the University of California system – general state funds even declined during the same period.[3]

These changes in the funding landscape have had tremendous effects on academic work and academic careers. General university funds which provide the basis for permanent, tenured positions are stagnating or even declining while on the other hand available work on research projects is

increasing. This has led to shifts in the personnel structure at universities.

University research is increasingly following a project logic with a defined beginning and end, and it is carried out in teams. In the social and behavioral sciences as well as in other fields of science it is no longer – if it ever was – a single professor elaborating theories on society in the privacy of his or her office. Instead, the data collection and subsequent analysis of data sets in large empirical research projects require a team effort. In commissioned or sponsored research, project teams generally consist of a Principal Investigator (PI) as the project leader providing the general lines of inquiry – usually a professor or other tenured member of a university – and several postdoctoral scholars and graduate research assistants/PhD students employed on the basis of 'soft money' or external funds. The latter typically do the groundwork and carry out the research, sometimes they are also tasked with writing the proposals. These projects – and thus a large part of university research – require flexible, readily available researchers willing to take on fixed-term work. Clark Kerr has called these academic workers 'the un-faculty' (2001: 49).

The numbers of postdoctoral appointees in science in the US increased rapidly from 1998 to 2011, in the social sciences postdoctoral appointments almost doubled in that period. [4] On the other hand, 'ladder faculty' (tenured or tenure-track positions) has grown at a much smaller rate (see Schuster and Finkelstein, 2006 for the US). On the European continent, the trend is equally, if not slightly more pronounced. At German higher education institutions, externally funded academic staff doubled in only 7 years (2005–12) whereas staff employed on the basis of general university funds only increased by 33%.[5] Also, at Austrian universities, the numbers of 'core staff' have stagnated while the fastest growing group of 'academic workers' were pre- and postdoctoral positions on the basis of external funds or 'project money.'[6]

As a result of the growing relevance of research projects and the increase in numbers of postdocs, the nature of the postdoctoral phase in academic careers is changing.

The Postdoctoral Experience: 'Project Workers' in the US and Europe

The postdoctoral phase is considered the most difficult and high risk, critical phase of an academic career (Brechelmacher et al., 2015). It falls between the completion of the doctorate and entering into permanent, tenured employment. In the US, postdoctoral positions are typically funded on the basis of grant money. They are positions offered in the framework of research projects or external research grants and contracts.[7]

The NSF and the National Institutes of Health (NIH) only recently defined a postdoctoral scholar as an individual who has received a doctoral degree (or equivalent) and is engaged in "a temporary and defined period of mentored advanced training to enhance the professional skills and research independence needed to pursue his or her chosen career path."[8] Postdocs are considered trainees, however, under specific circumstances they can also act as principal investigators. Having the NSF/NIH officially recognize the postdoc position by establishing a definition was an important milestone in the recognition of the contributions of postdocs. Until recently, as Cathee Johnson Philips of the National Postdoctoral Association claims, "postdocs were invisible […]. Most institutions did not know how many postdocs they had, due to the nature of the position."[9]

However, the composition of the academic workforce is changing, as the above numbers illustrate for both the US and Europe. "The […] scientific enterprise has become increasingly dependent upon postdocs to conduct research and maintain its position in the global research enterprise."[10] Today, postdocs account for much of the labor in academic research. Henry R. Bourne (2013) points out for the biomedical sciences that "the laboratory workforce [is] largely made up of PhD students and postdoctoral researchers, mostly supported by research project grants, with a relatively small number of principal investigators leading ever larger research groups." And he concludes provocatively that, today, "trainees equal research workforce."

Universities rely on post- (and pre-)doctoral appointees to carry out their – externally funded – research. The conditional and temporary nature of project funded research takes its toll on those embarking on academic careers. Research projects generally cannot generate permanent jobs as universities are reluctant to create permanent positions on the basis of external funds (see Brechelmacher et al., 2015). If funds run out, even experienced researchers will be asked to leave.

Due to an oversupply of postdocs and a high level of competition, it is becoming harder to secure a position on the tenure track. As a result, young academics are often forced to take on one postdoc position after the other. "In too many instances 'postdoctoral training'—when a young scientist is learning new approaches and techniques towards independence—has turned into 'postdoctoral employment', with the postdoc remaining at the same professional rank with little advancement or additional training" (National Research Council, 2005). However, as *Johnson Phillips* argues, "if a person remains in a postdoc [position] for longer than five years, their chances of having a successful career in science diminishes."[11] Some academic job seekers are already "wondering if there is a stigma attached to holding too many postdoc/soft money positions when looking for a tenure-track position."[12]

While on the one hand, successful postdoctoral work has increasingly turned into a prerequisite for obtaining a tenure-track position or permanent employment, on the other hand, "postdocs are starting to become a substitute for real jobs."[13] One could raise the question if the notion of what constitutes a 'real job' is not in the process of changing.

Further Effects of Project-Based, External Funding in Academe

Beyond the issue of uncertain working conditions and the high-risk nature of project-funded employment, project-based funding schemes affect the way academic research is carried out in other ways as well: the debate centers mostly around the fact that the successful acquisition of 'external' funding has become an indicator for quality on the institutional level as well as

a means to measure the scientific reputation of individuals in academe (see Münch, 2006: 440).

Richard Münch, a critic of the increasing centrality of conditional funding in the German higher education system, maintains that an increasing reliance on external funds favors 'normal science' – standardized mainstream research. Due to limited project duration, the probability of delivering results in a certain timeframe becomes a central factor in funding decisions and thus puts riskier research endeavors at a disadvantage (Münch, 2006: 448). The peer-review process – while mostly able to filter out research proposals of lower quality – can also block new, creative approaches outside of existing 'schools of thought.' In the funding process, both sides strive to reduce uncertainty, as Münch argues, and thus large, mostly quantitative research projects are favored, as well as those building on and extending formerly approved projects: "Serial project funding along the same lines over and over again" hinders innovation and the creation of new knowledge especially in the social sciences (Münch, 2006: 449). To break this cycle, Münch advocates the explicit use of external funding to reward more risky scientific endeavors and research off the beaten track, as 'normal science' is carried out at higher education institutions anyway.

The ability to raise external funds has become a measure of scientific excellence and an indicator of academic potential. As a result, the amount of soft money raised by an academic is increasingly determining his or her 'value' in the academic job market. The acquisition of external funds thus plays a decisive role in the hiring and recruitment processes of academic personnel.

Another consequence of the increased reliance on external funding in university research is a certain realignment of tasks, affecting what it is that professors actually do: less and less do PIs find time to engage in actual research themselves. Their time is less spent on scientific inquiry and increasingly devoted to acquiring funding. Further, their contributions to the projects are becoming more and more managerial. A recent article entitled "Dr. No Money: The Broken Science Funding System - Scientists spend too much time raising cash instead of doing experiments" quotes a 2007 US study according to which university faculty members spend about 40% of their research time navigating the bureaucratic labyrinth, and the situation is not different in Europe.[14]

With tenured faculty or PIs outlining the general direction of research and PhDs and postdocs carrying out the actual research, sponsored projects have taken on a training function in higher education. As Brechelmacher et al., 2015 show for Europe, a predoctoral position on the basis of soft money is often regarded as a 'typical' entry position into an academic career.[15] Research training increasingly takes place in the framework of externally funded projects, and these projects fulfill a vital role in the training of the academic workforce.

Finally, current funding schemes tend to create a dynamic similar to what Merton referred to as the 'Matthew principle' in science whereby formerly successful recipients of research funds will be more likely to receive future monies than others (see Merton, 1968). A concentration process thus occurs (see also Kerr, 2001: 47). Differentiation in the US higher education system has traditionally been high. Around half of total NSF funding in the social sciences went to only 20 of around 200 supported institutions in 2013.[16] At German universities, a concentration process is currently in full swing, and efforts like the so-called Excellence Initiative propel these dynamics even further. Ten out of a total of 100 institutions in the social and behavioral sciences received half of all DFG funding in 2008–10.[17]

Network Governance – Conceptualizing Project Work

The 'project logic' or project orientation of research has deeply permeated and transformed higher education and the way university research is carried out. Boltanski and Chiapello's concept of a 'justificatory regime' of projects – a project-based society or: 'cité' – provides an insightful framework that also applies to higher education and university research (see also Holtgrewe, 2010). Following Boltanski and Chiapelli "the project-oriented systemic logic values adaptability, flexibility, polyvalency, mobility, initiative, autonomy, risk-taking, and openness to new people, possibilities, and information" (Markin, 2008). These are qualities expected of individuals attempting to embark on an academic career today.

Most important, however, is the fact that projects are essentially based on networks: "The project is the occasion and reason for the connection" (Boltanski and Chiapello, 2005: 104). Thus, the concept of 'network governance' emerged as a corresponding governance paradigm. The concept was adapted to higher education, among others, by Ferlie et al. (2008), Paradeise et al. (2009) or Bleiklie et al. (2011). It is defined as a "form of multi-level governance, encompassing a greater range of actors engaged in shared decision-making processes and interactions between various layers of influence" (Campbell, 2013: 206) and refers mostly to "new kinds of policy networks related to external research funding mechanisms, evaluation and accreditation agencies" (Bleiklie et al., 2011: 176). Network governance – power and decision-making processes rooted in shifting networks – can also be viewed as a by-product and consequence of NPM-inspired governance schemes (budgetary restraints, conditional funding, increased level of competition, heavier emphasis on performance (Ferlie et al., 2008).

Network governance in higher education is best exemplified by the process of peer review: crucial decisions in academe – which projects get funded, who gets hired – are no longer taken by the institutions or institutional players (such as the department, the university, or the ministry of education) themselves, but are transferred or outsourced to superimposed networks of experts. Recruitment decisions are no longer or increasingly less taken by the actual organization itself, rather they are determined by external experts in so-called peer-review processes.

Summary and Outlook

The phase of massification of higher education in the second half of the twentieth century saw a steep increase in public research expenditure and with it the rise of the 'research project.' However, growing monetary support for academic research was accompanied by a decrease of trust in higher education

institutions. A decreasing part of research funding is distributed unconditionally to universities via lump-sum agreements. Rather, funding is designated for specific purposes and only the 'best' applicants are rewarded in a competitive process. A network of 'peers' is to determine the quality of the proposals. As there is no guarantee that these monies will flow permanently to institutions, they outsource the risk. Thus, research is increasingly carried out by people who are employed on the basis of external funds, researchers who are 'external' to the institutions they are actually working at, rather than by tenured staff holding permanent positions at the institution. With shrinking general university funds and an increasing amount of academics employed on the basis of 'soft money,' it becomes difficult to reach tenure for the larger part of the research workforce at universities. The question remains if project-shaped work environments – continuous reapplication and revalidation for work/funding – is a specific characteristic of academic work or, rather, a general social trend.

End Notes

1. The same holds true for total federal US support for university science & engineering R&D. Tables on "Federal Support for University R&D by Agency" as well as "University S&E R&D Funding by Source, 1990–2012"; based on data from the NSF's HERD survey available online at http://www.aaas.org/page/rd-colleges-and-universities. Data in constant 2014 dollars.
2. For Germany: An increase of 170%, adjusted by inflation an increase of 130%. Source: Deutsche Forschungsgemeinschaft (DFG), Funding Atlas 2012, p. 29 as well as author's own calculations on the basis of data from the German Federal Statistical Agency/Statistisches Bundesamt (DESTATIS): Bildung und Kultur. Finanzen der Hochschulen 2012 (Wiesbaden 2014). Available online at: https://www.destatis.de/GPStatistik/receive/DESerie_serie_00000119; For Austria: An increase of around 460%, adjusted by inflation an increase of 365%; Source: Austrian Federal Ministry for Education, Hochschulbericht 1999, Vol. 2, p. 13 and Statistisches Taschenbuch 2013, p.89, Table 6.4. Available online at bmwfw.gv.at (Unidata).
3. General university funds in Germany increased by 30% from 1998 to 2012, adjusted by inflation this number sinks to about 10%. By 2012, external funds amount to almost one-third (29%) of university funding in Germany (up from around one-sixth in the late 1990s). Source: see footnote 1. The Austrian general university budget increased by around 60% from 1997–2012, adjusted by inflation 30%. Source: see footnote 1. For the University of California system, state funds decreased from 2519 billion USD in 1998 to 2374 billion USD in 2011, adjusted for inflation a decrease of almost 30%. See http://budget.universityofcalifornia.edu/.
4. Across all disciplines an increase of 49.8% from 1998 to 2011, for the social sciences (excluding Psychology) an increase of 97%. Source. NSF, Graduate Students and Postdoctorates in Science and Engineering: Fall 2011, Table 28, http://www.nsf.gov/statistics/nsf13331/content.cfm?pub_id=4290&id=2.
5. Statistisches Bundesamt Wiesbaden, Fachserie 11, Personal an Hochschulen 2005 and 2012; https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Hochschulen/PersonalHochschulen.html.
6. Data Available at bmwfw.gv.at (Unidata).
7. See NSF, Graduate Students and Postdoctorates in Science and Engineering: Fall 2011, Table 39 (primary source of support) and 40 (primary mechanism of support). http://www.nsf.gov/statistics/nsf13331/content.cfm?pub_id=4290&id=2.
8. http://grants.nih.gov/training/Reed_Letter.pdf.
9. The Current State of the Postdoc Experience, Interview with Cathee Johnson Philips, Executive Director of the National Postdoctoral Association: http://www.higheredjobs.com/HigherEdCareers/interviews.cfm?ID=184.
10. See End note 9.
11. See End note 8.
12. http://academia.stackexchange.com/questions/14964/is-there-a-stigma-in-computer-science-toward-too-many-postdoc-positions.
13. Beryl LieffBenderly, Worrisome Trends for Chemistry Postdocs (May 2013), http://sciencecareers.sciencemag.org/career_magazine/previous_issues/articles/2013_05_24/caredit.a1300112, quoting William F. Banholzer, article last retrieved June 22, 2014.
14. http://www.scientificamerican.com/article/dr-no-money/, April 19, 2011, last retrieved June 22, 2014.
15. Based on the analysis of around 500 qualitative interviews in eight European countries in the framework of the ESF's EUROAC project: http://www.uni-kassel.de/einrichtungen/incher/forschung/wissenschaftlicher-wandel/euroac-academic-profession-in-europe.html.
16. 47.5%; author's own calculations based on data on NSF Budget Internet Information System, http://dellweb.bfa.nsf.gov/.
17. 49% of all funding went to the top 10 institutions. This number compares to around 40%, 10 years earlier (in 1999–2001). DFG Funding Atlas 2012, Table 4.6: Absolute DFG-Bewilligungen für 2008 bis 2010 nach Hochschulen im Fachgebiet Sozial- und Verhaltenswissenschaften. http://www.dfg.de/dfg_profil/foerderatlas_evaluation_statistik/foerderatlas/tabellen_abbildungen/index.html#micro6470262. DFG Förderranking 2003: Tabelle A3-6: DFG-Bewilligungen 1999 bis 2001 je Hochschule und Fachgebiet, http://www.dfg.de/download/pdf/dfg_im_profil/evaluation_statistik/ranking/archiv/dfg_foerderranking_2003.pdf.

See also: Academic Careers in Comparative Perspective.

Bibliography

Bleiklie, I., Enders, J., Lepori, B., Musselin, C., 2011. New public management, network governance and the university as a changing professional organization. In: Christensen, T., Lægreid, P. (Eds.), The Ashgate Research Companion to New Public Management. Ashgate, Farnham, 161–176.

Boltanski, L., Chiapello, È., 2005. The New Spirit of Capitalism. Verso, London/New York.

Brechelmacher, A., Park, E., Ates, G., Campbell, D., 2015. The rocky road to tenure – Career Paths in Academia. In: Fumasoli, T., Goastellec, G., Kehm, B.M. (Eds.), Academic Careers in Europe: Trends, Challenges, Perspectives. Springer, Dordrecht, 13–40.

Bourne, H.R., 2013. A fair deal for PhD students and postdocs. eLife 2, e01139. http://dx.doi.org/10.7554/eLife.01139. Published online Oct 1, 2013.

Campbell, D., 2013. New university governance: how the academic profession perceives the evaluation of research and teaching. In: Teichler, U., öhle, E.A.H. (Eds.), The Work Situation of the Academic Profession: Findings of a Survey in Twelve European Countries. Springer, Dordrecht, 205–228.

Clark, B., 1997. .The modern integration of research activities with teaching and learning. Journal of Higher Education 68, 241–255.

Castells, M., 1996. The Rise of the Network Society. In: The Information Age: Economy, Society And Culture, vol. 1. Blackwell Publishers, Malden, MA.

The Current State of the Postdoc Experience, Interview with Cathee Johnson Phillips, Executive Director of the National Postdoctoral Association, http://www.higheredjobs.com/HigherEdCareers/interviews.cfm?ID=184.

Deutsche Forschungsgemeinschaft, 2012. Förderatlas 2012, Kennzahlen zur öffentlich finanzierten Forschung in Deutschland. Wiley, Weinheim.

Ferlie, E., Musselin, C., Andresani, G., 2008. The steering of higher education systems: a public management perspective. Higher Education 56 (3), 325–348.

Gumport, P.J., 2011. Graduate education and research: interdependence and strain. In: Altbach, P.G., Gumport, P.J., Berdahl, R.O. (Eds.), American Higher Education in the Twenty-First Century: Social, Political and Economic Challenges, third ed., 365–408.

Holtgrewe, U., 2010. Projects under pressure: how networked capitalism both requires and undermines its "new spirit" and how security continues to matter. In: Paper Contributed to the 10th International Labour Process Conference, Rutgers University, 15–17 March 2010.

Kerr, C., 2001. The Uses of the University, fifth ed. Harvard University Press, Cambridge/MA.

Markin, P., 2008. Analysis of Boltanski and Chiapello's (2005) the New Spirit of Capitalism. http://pablomarkin.blogspot.co.at/2007/11/review-of-boltanski-and-chiapellos-2005.html.

Merton, R.K., 1968. The matthew effect in science. Science 159 (3810), 56–63.

Münch, R., 2006. Drittmittel und Publikation. Forschung zwischen Normalwissenschaft und Innovation. Soziologie 35 (4), 440–461.

National Research Council, 2005. Identifying Opportunities for and Challenges to Fostering the Independence of Young Investigators in the Life Sciences. National Academies Press, Washington, DC. http://www.ncbi.nlm.nih.gov/books/NBK22688/#top.

Paradeise, C., Reale, E., Bleiklie, I., Ferlie, E., 2009. University Governance. Western European Comparative Perspectives. Springer, Dordrecht.

Schuster, J.H., Finkelstein, M.J., 2006. The American Faculty: The Restructuring of Academic Work and Careers. The Johns Hopkins University Press, Baltimore.

Weber, Max, 1947. Science as a profession. In: Gerth, H.H., Mills, C.W. (Eds.), From Max Weber: Essays in Sociology. Kegan, London, 129–156.

Academic Self-Concept and Achievement

Kit-Tai Hau, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong
Herbert W Marsh, Institute for Positive Psychology and Education, Australian Catholic University, NSW, Australia; King Saud University, Riyadh, Saudi Arabia; and University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Positive academic self-concept has been considered one of the most important indicators of educational success. Empirical evidence across divergent educational and cultural contexts supports (1) a hierarchical multidimensional model of self-concept, (2) reciprocal causal relations between academic achievement and corresponding self-concepts, (3) the use of both internal and external frames in making self-concept judgment, and (4) the importance of both social comparison and reflected glory effects.

Self-concept is defined as "a person's self-perceptions that are formed through experience with and interpretations of one's environment" (Marsh et al., 2011; see also Shavelson et al., 1976). It is affected in particular by evaluative information from other people (e.g., teachers, parents, classmates), such as outcome feedback, reinforcements, and attributions for one's performance. When a person interprets one's environment subjectively, this process may involve the use of objective measures (e.g., standardized achievement tests) as well as dispositional (e.g., personality; Marsh, 2008) and situational (e.g., average achievement levels of a school) factors.

Thus, self-concept is one of the most important constructs in many fields of psychology. This article concentrates on the notion of academic self-concept, the improvement of which is a major educational goal or outcome given that high self-concept is a positive indicator of educational success. Academic self-concept is also an important educational mediator in that high self-concept leads to other desirable personal or educational outcomes. Its importance was summarized neatly by Branden (1994),

> *I cannot think of a single psychological problem* – from anxiety to depression, to under-achievement at school or at work, to fear of intimacy, happiness or success, to alcohol or drug abuse … – *that is not traceable, at least in part, to the problem of deficient self-esteem*. (p. xv)

Although self-concept is one of the oldest and most important constructs in the social sciences, its theoretical development has been slow, particularly in the days of behaviorism. However, since the 1980s, research on self-concept has progressed on the basis of advancements in psychological measurement instruments, research methodologies, and theoretical refinements (e.g., the multidimensional model; see Marsh, 2007; Marsh et al., 2011).

Hierarchical Multidimensional Structure

The development of the theoretical model of self-concept has differentiated between within-network and between-network research. In the former type of research, scholarly interest focuses on the internal structure, features, and attributes of self-concept, whereas in the latter researchers concentrate on how self-concept is related to other constructs beyond its own conceptual domain.

Global Self-Esteem vs Multidimensional Self-Concept

Typically, within-network research is seen to be more important than between-network research during the early stages of theory building. As Marsh and Scalas (2010: p. 11) noted, "The determination of whether theoretically consistent and distinguishable dimensions of self-concept existed, and their content and structure [within-network research], should be prerequisite to the study of how these dimensions, or overall self-concept, are related to other variables [between-network research]."

Current models of self-concept can be traced back to the multidimensional hierarchical model proposed by Shavelson et al. (1976) (see **Figure 1**). In this hypothetical model, situation-specific (e.g., verbal, numeric) self-concepts are at the base of the hierarchy, self-concepts of broader domains (e.g., social, academic) are in the middle, and the most generalized and global self-concept (known in general as self-esteem) is at the top.

Despite this heuristic multidimensional model, unidimensional instruments that measured a single self-esteem factor (e.g., Coopersmith, 1967) predominated before the mid-1980s, probably because of the lack of strong empirical support for the multidimensional model at that time. Indeed, this issue of whether self-concept is unidimensional or multidimensional has remained the focus of ongoing debate (Marsh and Craven, 1997; Rosenberg et al., 1995; Suls, 1993). With the development of instruments that were much stronger psychometrically, empirical evidence began to show the necessity and usefulness of differentiating between the dimensions that comprise self-concept. A number of new multidimensional instruments that were developed typically contained specific self-concept scales for the academic (e.g., verbal, numeric), social (e.g., with friends), physical (e.g., physical competence, attractiveness), emotional, and global (i.e., self-esteem) domains (see Marsh and Scalas, 2010). Reviews (e.g., Byrne, 1984, 1996; Hattie, 1992; Wylie, 1989) recommended the use of the Self Description Questionnaires (SDQ), which

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92153-6

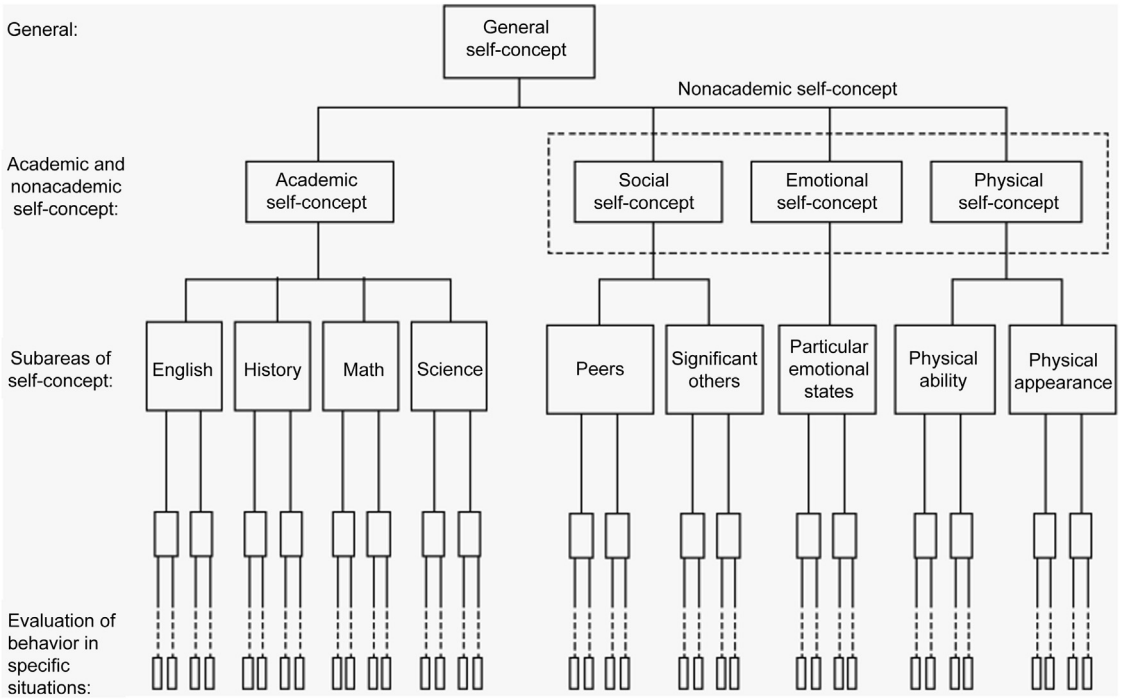

Figure 1 The multidimensional, hierarchical model of the self-concept construct. The box that consists of dashed lines around the nonacademic self-concept factors is used to distinguish these from the academic self-concept factors, but does not imply that there is a single higher order non-academic factor, as is hypothesized for the academic factors. The unlabeled boxes at the bottom of the hierarchy are used to show that the model posits additional levels in the hierarchy and even more domain-specific components of self-concept than those that are presented explicitly (e.g., mathematics self-concept might be divided into different mathematical topics, such as algebra, trigonometry, and calculus, and each of these could be further subdivided into specific components that are relevant to each of the mathematical subjects). Reprinted with permission from Shavelson, R.J., Hubner, J.J., Stanton, G.C., 1976. Validation of construct interpretations. Review of Educational Research 46, 407–441.

targeted specific age groups (e.g., SDQ I, II, and III were designed for use with children, adolescents, and young adults, respectively; Marsh, 2007).

Research has since shown that certain measures of self-concept can even be used with very young children. For example, Marsh et al. (1991, 1998) developed an individually administered procedure for SDQ I for use in younger children (aged 5–8 years), which has acceptable psychometric properties even for younger special education students (Grades 2–6) who have mild intellectual disabilities (intelligence quotient 56–75) (Tracey et al., 2003). Although scores for the different dimensions become more differentiated with increasing age, and their relations with external indicators also strengthen (e.g., increasing correlations with achievement), research suggests that young children can differentiate between multiple dimensions of self-concept at a much earlier age than was believed previously.

Empirical Evidence: Low Correlations among Self-Concept Domains

These multidimensional characteristics of self-concept have been supported in several areas of empirical research. First, the advanced confirmatory factor analyses that have been developed over recent years have shown that academic self-concepts are virtually uncorrelated or even negatively correlated with nonacademic self-concepts and self-esteem. Multitrait multimethod analytical techniques on the ratings from different instruments and by different significant others (teachers, parents vs students) also support the discriminant (i.e., differences across multiple dimensions of self-concept) and convergent (i.e., similarity across instruments and raters) validity of multidimensional self-concept ratings. Further, self-concepts have been found to be much more differentiated than academic achievement (or school grades) in the corresponding domain (Marsh and Craven, 2006).

Second, different domains of academic achievement (e.g., verbal achievement and mathematical achievement) tend to be correlated substantially, as does each domain of achievement with its matching self-concept, for example, verbal achievement with the verbal self-concept. However, different domains of self-concepts (e.g., verbal self-concept and mathematics self-concept) are uncorrelated in general. In other words, academic subjects share a large percentage of common variance (i.e., general achievement or intelligence; they are substantially correlated and hierarchically ordered), whereas their corresponding self-concepts do not.

Third, in structural equation models that relate self-concepts to other constructs, although different domains of academic achievement or school grades have predictable relations with the matching and nonmatching domains of self-concepts, they are unrelated in general to self-esteem.

In light of the foregoing empirical evidence, the multidimensional hierarchical model of Shavelson et al. (1976) was revised in the 1980s. Specifically, although all first-order academic-subject-specific academic self-concepts are grouped under the second-order academic self-concepts in the original hierarchical model, subsequent empirical results showed much weaker (indeed, close to zero) correlations between different first-order domains of self-concepts (e.g., mathematics and verbal self-concepts). The original model of Shavelson et al. was thus revised into the Marsh/Shavelson model, in which the second-order academic factor was divided into domain-specific factors (e.g., verbal and mathematical self-concepts) (Marsh et al., 1988).

Not only did the revised model fit the data much better than the original, but stronger relations were also found when domain-specific self-concepts were used. For example, Marsh and Yeung (1997) showed much stronger effects of these domain-specific self-concepts (vs general self-esteem) on subsequent coursework selection, which demonstrated the importance of domain-specific academic self-concept as compared with that of general self-esteem. These empirical results also suggested that the hierarchical structure was weaker than anticipated originally, whereas multidimensional characteristics appeared to be more salient, useful, and important.

Educational Implications

Empirical findings demonstrate that students are able to evaluate their competencies in a number of academic subjects. Consequently, researchers are encouraged to use scales that measure self-concept in specific academic subjects (e.g., using the mathematics self-concept) – such as by applying the Academic Self Description Questionnaire (Marsh, 1990b) or the SDQ (Marsh, 2007) – in addition to or instead of adopting general academic self-concept scales.

According to the multidimensional perspective, any type of self-concept intervention, if successful, should have positive effects only on specific targeted or closely related facets of self-concept rather than unrelated ones. Indeed, this approach to construct validation has been supported in enhancing physical fitness (Marsh and Peart, 1988), and in Outward Bound program (Marsh et al., 1986) and a wide range of other studies (Haney and Durlak, 1998; Marsh and Craven, 1997; O'Mara et al., 2006). Using the global self-concept approach, Haney and Durlak's (1998) meta-analysis showed the moderate positive effect ($d = 0.51$ over 460 effect sizes) of a self-concept enhancement program. This was much smaller than the effect observed in the subsequent update and reanalysis by O'Mara et al. (2006) ($d = 1.16$), who used the multidimensional approach.

The implication of these findings is that, in educational interventions, teachers must take into consideration the multidimensional characteristic of self-concept in their instructional designs in order to enhance this construct. Further, positive educational feedback must be focused specifically on the targeted academic subject or set of skills. Vague or broadly applied enhancement programs that aim to raise the general level of self-concept in order to improve performance in a specific academic subject might be ineffective.

Causal Ordering of Academic Self-Concept and Achievement

Although the majority of interest in academic self-concept might stem from the belief that students can improve their academic achievement by boosting their level of self-concept in different areas (Byrne, 1984; Marsh, 1990b), there has been a distinct lack of empirical support before the 1990s. In other words, although the specific domains of academic achievements have been shown to be correlated with their respective self-concepts as discussed earlier, there are practical (e.g., in terms of actual educational interventions) and theoretical reasons to differentiate the cause from the outcome.

Self-Enhancement and Skill Development Models

Two hypothetical models have been proposed to explain the relation between academic achievement and academic self-concept, namely, the self-enhancement and skill development models (see Marsh and Martin, 2011). In the former, self-concept is thought to be the predominant cause of subsequent achievement, whereas the reverse is true in the latter. Specifically, researchers would like to understand empirically whether a more positive academic self-concept is the cause of better academic achievement or whether better academic achievement leads to a more positive academic self-concept.

Although an experimental study using control groups might be an appealing approach to solve such a problem of causal ordering, the empirical manipulation of achievement and self-concept might not be easy or practical. For example, it is difficult, if possible at all, to assign students to a high-achievement group and then raise their levels of academic achievement experimentally without inducing other changes. These difficulties in experimental design have led to the use of longitudinal panel data instead, with domain-specific self-concept and achievement data collected several times (at least twice; thrice in **Figure 2**). Logically, prior achievement should positively affect subsequent achievement (e.g., see **Figure 2**, Time 1 achievement (T1-ACH) on Time 2 achievement (T2-ACH)), because students who have higher achievement in the first round of data collection (T1) would tend to have higher achievement in the second round (T2). Similarly, prior self-concept should also influence subsequent self-concept positively, because students who evaluate themselves highly at T1 would also tend to evaluate themselves more positively at T2.

The most important theoretical questions are (1) whether better prior achievement also benefits subsequent self-concept (e.g., T1-ACH → T2 academic self-concept (ASC; **Figure 2**) and (2) whether a more positive prior self-concept also benefits subsequent achievement (e.g., T1-ASC → T2-ACH; **Figure 2**). Specifically, as presented in **Figure 2**, after statistically controlling for the two main effects (T1-ACH on T2-ACH; T1-ASC on T2-ASC, etc.), we are interested in whether cross-lagged paths exist. With the development of stronger assessment instruments, research designs, and analytical methodologies, empirical research and meta-analyses (Marsh and Craven, 2006; Valentine et al., 2004) have supported the reciprocal effects model (REM; Marsh, 1990a), which proposes that both types of effects (path) are statistically significant and important. Although it is accepted widely that

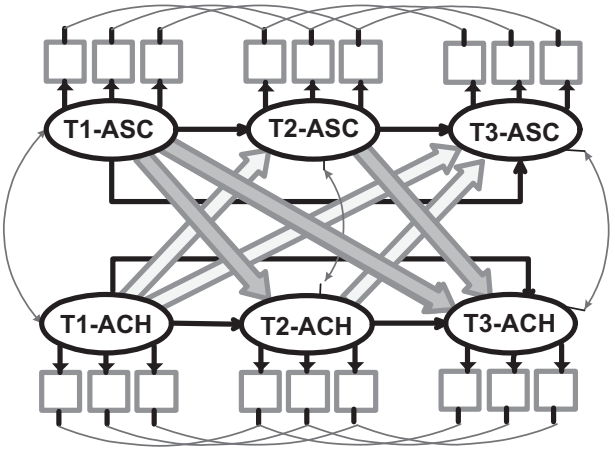

Figure 2 Prototype causal ordering model to test self-enhancement, skill development, and REMs. In this full-forward, multiwave, multivariable model, multiple indicators of ASC and ACH are collected in three successive waves (T1, T2, and T3). Each latent construct (represented by ovals) has paths that lead to all latent constructs in subsequent waves. Within each wave, ASC and ACH are assumed to be correlated. In the first wave, this correlation is a covariance between two latent constructs, and in subsequent waves, it is a covariance between residual factors. Curved lines at the top and bottom of the figure reflect correlated uniqueness between responses to the same measured variable (represented by boxes) collected on different occasions. Paths that connect the same variable on multiple occasions reflect stability (the solid black paths), but these coefficients typically differ from the corresponding test–retest correlations (which do not include the effects of other variables). Light gray arrows reflect effects of prior achievement on subsequent academic self-concept, whereas dark gray arrows reflect the effects of prior academic self-concept on subsequent achievement. Adapted with permission from: Marsh, H.W., 2007. Self-concept Theory, Measurement and Research into Practice: The Role of Self-concept in Educational Psychology. British Psychological Society, Leicester, UK. http://www.bps.org.uk/publications/bps-journals/journals-copyright-authors/journals-copyright-informationInformation.

achievement is a determinant of subsequent self-concept, the authors' main research interest lies in demonstrating the existence of the self-enhancement path (ASC → ACH).

Previous research has also shown that the above-mentioned REM, which relates achievement and self-belief, can be applied to a range of academic subjects (i.e., not just the most widely studied subject of mathematics), as well as nonacademic areas (e.g., physical self-concept with gymnastics and elite swimming) and psychological domains (e.g., self-efficacy) (Guay et al., 2003; Marsh and Craven, 2005; Valentine and DuBois, 2005; see meta-analyses, Valentine et al., 2004). Further, this model is robust across age groups (not only university or high school students but also younger children) and across cultures, for example, Chinese students from Hong Kong (Marsh et al., 2002) and students from East and West Germany at the fall of the Berlin Wall (Marsh and Köller, 2003).

Importance of Multidimensionality and Positive Psychology

With respect to the debate on the significance of self-concept in positive psychology, the importance of considering the multidimensional characteristic of self-concept again becomes apparent. In particular, the recent emphasis on positive psychology has tended to advocate that positive self-belief or self-evaluations are desirable for maximizing life outcomes. In their reviews, Baumeister et al. (2003, 2005) originally seemed to question whether high self-esteem really leads to better performance, happiness, or a healthier lifestyle, concluding that "[positive] self-esteem per se is not the social panacea that many people hoped it was" (2003: p. 38) and that "efforts to boost people's self-esteem are of little value in fostering academic achievement or preventing undesirable behavior" (2005: p. 84). However, in refuting these inferences, Marsh and Scalas (2010; see also Marsh and Craven, 2006) pointed out that Baumeister et al.'s (2003, 2005) conclusions "were based largely on research studies, statistical methodology, and theoretical conceptualizations of self-concept that are no longer current" (p. 665).

Baumeister et al. (2003, 2005) concentrated on self-esteem and self-concept from a unidimensional perspective, whereas Marsh and Craven (2006) drew on more recent research on an explicitly multidimensional model of self-concept to demonstrate convincingly the benefits of positive academic self-concept. Similarly, in their reanalysis of the US nationally representative Youth in Transition database (five waves of data that span 8 years from Year 10), Marsh and O'Mara (2008) showed strong positive reciprocal effects between academic self-concept and grade point average (GPA). They also showed that besides prior achievement, academic self-concept is the best predictor of long-term educational attainment. This finding contrasted with the earlier analyses by Baumeister et al. (2003, 2005), who found weak and inconsistent relations between global self-concept and achievement. In summary, strong empirical evidence supports the notion that academic self-concept, when taken as a multidimensional construct, positively influences subsequent achievement after controlling for the effects of prior achievement.

Educational Implications

An important implication of the REM is that educational practices and interventions must take reciprocal relations into consideration. Educational practices that rely solely on the skill development model (i.e., focus primarily on improving academic skills without enhancing self-concept) or the self-enhancement model (i.e., foster self-concept alone without improving academic skills) are overly simplistic and their effects will not be long lasting. The most effective interventions should improve academic skills and self-concept simultaneously.

Frame of Reference: The Internal/External Model

For at least two reasons, the self-concepts of different academic domains would be expected to be correlated. First, as discussed earlier, achievements in different domains are often moderately correlated (e.g., achievements in mathematics and the verbal domain are typically correlated at 0.5–0.8; Marsh, 2007), whereas academic achievement in each individual domain is

related to its respective self-concept (e.g., mathematics achievement and mathematics self-concept). Hence, it is logical to expect the self-concepts of different domains (e.g., mathematics and verbal) to be correlated, too. Second, according to the proposed hierarchical structure (Marsh et al., 1988), the self-concepts of academic subjects should share some common variance in order to build up this hierarchical tree structure.

Simultaneous Use of External and Internal Frames

Surprisingly, however, empirical research has shown repeatedly that self-concepts in different academic domains are almost unrelated and sometimes even negatively correlated. This hardly reflects the corresponding relations among academic achievement in different domains and reaffirms the conclusion that domain specificity overrides hierarchical structure. These seemingly paradoxical relations were explained by Marsh (2007) using the internal/external (I/E) model. The model proposes that objective accomplishment (e.g., 80% of items correct in a test) is only one of the determinants of self-concept. Indeed, an individual's self-perception cannot be understood adequately without recognizing the prevailing frame of reference (e.g., how other students perform in the same test). Thus, unless this frame of reference is identical for all students (e.g., 80% correct is considered to be excellent by all students), the self-concepts of students will vary according to the specific reference frames that they have adopted.

The I/E model postulates that students simultaneously use both an external (normative) frame and an internal (ipsative-like) frame in forming their self-concepts. In the former, students compare their performances with external information and criteria such as objective indexes in public examination results and the performances of classmates. On the basis of this external frame, students who perform outstandingly in public examinations or better than their classmates tend to have higher levels of self-concept. In the internal frame, students judge their abilities in one academic subject with reference to their own performance in other subjects. Thus, when two students who have similar verbal examination scores are asked to rate their verbal self-concept, students who perform better in mathematics will tend to have a lower verbal self-concept than those who perform worse in mathematics.

Empirical Evidence

Statistically, if the verbal and mathematics achievements and self-concepts of students are measured, the comparison in the external frame would predict that verbal achievement positively affects verbal self-concept. The greater the verbal achievement of students, the higher are their levels of verbal self-concept (see the strong positive path from verbal achievement to verbal self-concept in Figure 3). Simultaneously, the comparison in the internal frame would predict that mathematics achievement negatively influences verbal self-concept and, similarly, verbal achievement negatively affects mathematics self-concept (the moderately negative crossed paths in Figure 3). Thus, the correlation

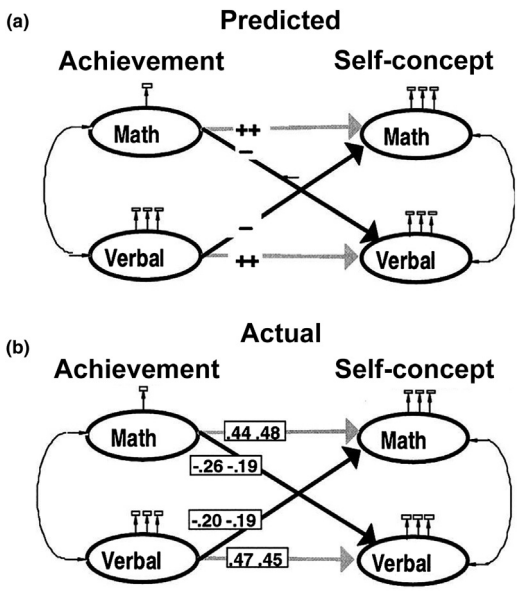

Figure 3 Predicted (Panel a) and actual (Panel b) results based on the I/E frame of reference model. In Panel (a), the horizontal (positive) paths are predicted to be substantial and positive (++), whereas the cross (negative) paths are predicted to be smaller and negative (−). In Panel (b), the actual results, which are based on total group analysis and the multiple group analysis, are consistent with the predictions. Reprinted with permission from Marsh, H.W., Hau, K.T., 2004. Explaining paradoxical relations between academic self-concepts and achievements: Cross-cultural generalisability of the internal-external frame of reference predictions across 26 countries. Journal of Educational Psychology 96, 56–67. http://www.apa.org/about/contact/copyright/index.aspx.

between verbal and mathematics self-concepts would be substantially lower than the typically high values of correlation between verbal and mathematics achievements. Depending on the relative strength of the effects of these internal and external comparisons, this correlation would be close to zero, slightly positive, or slightly negative.

The above statistical predictions have been confirmed and supported in a large number of cross-cultural studies, reviews, and meta-analyses (e.g., Marsh and Hau, 2004; Möller et al., 2009). For example, in the large ($N = 55\ 577$) Organisation for Economic Co-operation and Development Program for International Student Assessment (PISA) study, which was composed of nationally representative samples of 15-year-olds from 26 countries (Marsh and Hau, 2004), the horizontal paths from mathematics achievement to mathematics self-concept and verbal achievement to verbal self-concept were highly positive (0.44 and 0.47, respectively; see Figure 3), whereas the two cross-paths, mathematics achievement to verbal self-concept and verbal achievement to mathematics self-concept, were negative (−0.26 and −0.20, respectively). As predicted, the correlation between verbal and mathematics self-concepts (0.10) was substantially smaller than that between verbal and mathematics achievements (0.78). These findings, which were based on the total sample, were replicated in almost all the separate analyses for each of the 26 countries. Thus, the results were

in line with the conclusion of Möller et al.'s (2009) meta-analyses ($N = 125$, 308 from 69 data sets) on the relations between mathematics and verbal achievements and self-concepts.

In summary, although cross-cultural studies (e.g., the PISA study) and meta-analyses have their strengths and limitations (Marsh et al., 2009), the convergent results from both paradigms provide strong support for the I/E model across divergent educational contexts.

Educational Implications

Theoretically, the I/E model demonstrates once again the importance of the multidimensional nature of self-concept in related research. Practically, teachers and parents who infer the self-concepts of students predominantly from external comparisons (e.g., performance in examinations) should also pay attention to the internal frame adopted by the students (Dai, 2002; Marsh and Craven, 1997). For the brightest students in a class, even though their academic achievements might be better than those of their classmates in most subjects, their self-concepts will vary by subject area. For example, among these brightest students, those who are stronger in mathematics and science subjects and slightly weaker in verbal-related subjects might have much higher levels of self-concept and interest in science and mathematics and hence spend more time and effort on these subjects. By contrast, they might not pursue verbal-related subjects to the same extent, even though their objective abilities and demonstrated performances in these subjects are still much better than those of their classmates. Similarly, for the brightest students who are stronger in verbal subjects and weaker in science subjects, it might be difficult to understand their relative lack of interest and low levels of self-concept in science given their outstanding performance (e.g., objective marks) in these subjects relative to their classmates. These phenomena can be interpreted better by recognizing that self-concepts in different domains are highly differentiated and that some students consider themselves to have a greater aptitude at mathematics or verbal-related subjects on the basis of the internal comparison process.

Students who perform poorly value their worth through the internal comparison process in a similar manner. Although their performances in most academic subjects are weaker than those of their classmates, they still have relatively higher levels of self-concept in their better subjects. Therefore, teachers and parents could aim to build academic interest in these weaker students from the platform of their slightly better subjects (in which they show relatively higher levels of self-concept).

Frame of Reference: The Big-Fish-Little-Pond Effect

The big-fish-little-pond effect (BFLPE) cannot be adequately understood if the standards of comparison and the frames of references people use to evaluate themselves are ignored. It derives from "research on adaptation level, psychophysical judgment, social psychology, sociology, social comparison theory, and relative deprivation theory" (Marsh et al., 2008: p. 321; see also Möller and Marsh, 2013), among others. Marsh (1984) proposed the BFLPE model to explain the frame-of-reference effects that result when students compare their own academic performances and abilities with those of their peers. The BFLPE model postulates that students present higher (lower) levels of self-concept when they compare themselves with less (more) able classmates. Consequently, the BFLPE theory posits that for two students of similar ability, the one who attends a school that has a higher average ability presents a lower self-concept than the attendee of a school that has a lower average ability. Statistically, the academic achievements and abilities of students should have a strong and positive effect on their individual self-concepts, whereas the average abilities of their classmates or of the school as a whole (termed class- or school-average achievement) should negatively affect individual self-concept. Given that high-ability students tend to cluster in similar schools, student achievement should also have a positive effect on school-average achievement.

Empirical Evidence

The BFLPE model has been supported empirically in various surveys, experimental studies, and reviews (Marsh, 2007; Marsh et al., 2008). With the development of more appropriate analytical methodologies, the model has also been supported repeatedly by use of the multilevel modeling approach and in a large number of non-Western cultural and educational settings. For example, among the representative samples of students from 26 countries in PISA (Marsh and Hau, 2003), the effect of school-average achievement on individual self-concept was found to be negative in all 26 countries, reaching statistical significance in 24 countries (mean $= -0.20$, standard deviation $= 0.08$; negative in the remaining two countries but not significant). Similarly, in data analyses of the Trends in International Mathematics and Science Study study, school-average achievement was found to affect students' self-concepts negatively (Chiu, 2012). These results further support the generalizability and importance of the BFLPE in diverse settings.

Other Outcomes and Moderators

In PISA, 2003 (OECD, 2004), Seaton et al. (2009) extended the work done by Marsh and Hau (2003) and replicated their results with samples from 41 countries. Nagengast and Marsh (2011, 2012) also examined the BFLPE in PISA, 2006 (OECD, 2007) for the total international sample, the total UK sample, and each of the four UK countries. In particular, they used the more appropriate doubly latent model (vs the multilevel model with a single observed indicator variable at the class/school level) and a less often studied subject (i.e., science). The results supported the general applicability of the BFLPE with respect to science for the international and UK samples. Importantly, among the 57 countries examined, they found that (1) students' individual achievements were positively related to self-concept (52 countries) and career aspirations (42 countries), (2) the positive effect on career aspirations was mediated through self-concept (54 countries), and

(3) the effect of school-average achievement (BFLPE) on self-concept (50 countries) and career aspirations (31 countries) was negative (Nagengast and Marsh, 2011, 2012). These results attested to the generalizability of the negative effects of the BFLPE and their importance in mediating career aspirations across diverse educational settings.

Indeed, the negative effects of school-average achievement can be profound and long lasting. Marsh (1991) analyzed the High School and Beyond data on 1000 randomly selected high schools in the United States and approximately 30 randomly selected students from each school. This longitudinal study showed that the school-average achievement negatively affected academic self-concept, educational aspirations, general self-concept, coursework selection (e.g., selecting less demanding coursework), school grades, standardized test scores, occupational aspirations, and subsequent attendance at college. Moreover, some of these negative effects continued for at least 2 years after high school graduation and they were significant even after controlling for the effects of intermediate variables for the sophomore or senior year at high school. This finding suggested that the effects of school-average achievement (BFLPE) were negative beyond the already negative effects of the sophomore year at high school, suggesting that the BFLPE is long lasting.

Researchers have also investigated the potential mediators of the BFLPE (Marsh et al., 2011) in order to identify the contextual variables (e.g., classroom atmosphere) or individual differences (e.g., achievement goals) that might counteract or reduce the negative effects of the BFLPE. For example, studies have examined whether the BFLPE would have a negative effect on students of all levels of ability. Statistically, this is equivalent to testing the interaction effect between school-average achievement (i.e., BFLPE) and individual students' abilities (or other variables). However, in general, research does not support this interaction (Marsh and Hau, 2003; Marsh et al., 2011), and the BFLPE has been shown to be consistently negative across all achievement levels and all countries (e.g., all 26 countries in PISA, Marsh and Hau, 2003). In PISA 2003, Seaton et al. (2010; see also Marsh et al., 2008) also examined a wide range of potential moderators (ability, socioeconomic status, learning style, elaboration, memorization, control strategies, extrinsic motivation, intrinsic motivation, self-efficacy, anxiety, competitive preferences, cooperative learning preferences, identification with school, attitudes to school, sense of belonging, student–teacher relations), but found that these interactions were either not significant or of no substantive importance.

Mechanism of the BFLPE

Those who are skeptical about the negative effects of school-average achievement might argue that this shortcoming is only a temporary disadvantage because of grading-on-a-curve practices. In other words, all schools, irrespective of how well or poorly their students perform, award similar distributions of high and low grades to their students (i.e., all schools have similar GPA distributions). For two students who have similar abilities (as measured by public examinations), the one who studies in a competitive school would receive lower school grades and hence have a lower self-concept than the student who attends a school that has a lower school-average achievement.

The effect of school grades, independent of the contribution of school-average achievement, has been examined in various empirical studies (Marsh, 1987; Marsh and Rowe, 1996; Trautwein et al., 2006). This research has demonstrated that although school grades do partially explain and contribute toward the BFLPE, the negative effects of school-average achievement persist in addition to the effect of school grades. Reviews have also shown that these effects remain stable over time – even years after graduation (e.g., Marsh et al., 2007, 2008). For example, Marsh and O'Mara (2010) found that in a representative US sample (followed from Grade 10 to 5 years after high school graduation), the effects of the BFLPE due to school-average ability on academic self-concept, school grades, and educational and occupational aspirations were negative and that these were partially mediated through self-concept on distal outcomes.

Theoretically, there are two counterbalancing effects in attending a school that presents a high school-average achievement. On one hand, attending such schools should boost students' levels of self-concept, because the so-called assimilation effect will be positive (i.e., the 'reflected glory effect'). On the other hand, students in prestigious schools endure constant competition and comparison with high-ability classmates and schoolmates. This 'social comparison effect' is often negative. Thus, the BFLPE is the sum of these two counteracting effects and in general is negative. Given that the BFLPE has been found to be consistently negative across studies and countries, the negative effects of social comparison must be greater than the positive effects of reflected glory across a wide spectrum of educational settings. Indeed, Marsh et al. (2000) demonstrated empirically that when the positive reflected glory effect from attending a high-achieving school is removed, the negative effects of social comparison are much more negative than the generally documented overall BFLPE.

Experimentally, a negative BFLPE due to high school-average achievement can also be demonstrated by comparing the effects that arise from intergroup (positive social glory by comparison with lower average achievement schools) and intragroup (negative social comparison with high-ability classmates) comparisons. Zell and Alicke (2009) manipulated the standards of comparison in a series of studies and demonstrated that the overwhelmed negative intragroup comparisons are the main source of students' self-evaluation.

Educational Implications

Parents, teachers, and students might have believed naively in the benefits that attending a high-achieving school might confer on students' levels of self-concept and achievement, while forgetting that these students might actually begin with higher levels of achievement. When these prior differences are controlled for, students in these schools might actually suffer from the constant negative social

comparisons with classmates. Consequently, teachers in prestigious schools might need to consider deemphasizing intraclass competition and stressing individual growth, progression (comparison with self), or external objective standards. At the very least, teachers and parents should be aware of the potential negative impact on students' self-concept when they attend competitive and elite schools.

The BFLPE can be extended to the two extreme groups with respect to ability, namely, gifted-and-talented classes at the high end and inclusion classes at the low end. Grouping gifted-and-talented students in the same class might lead to a decline in self-concept (i.e., compared with that of similarly gifted students in mixed ability groups) due to the social comparison effect (Marsh et al., 1995). The labeling theory suggests that the self-concept of students who have a learning disadvantage could be hampered if they are placed in special classes with similarly disadvantaged students, a main argument for the inclusive education movement. According to the BFLPE, students with a learning disadvantage would have relatively low levels of self-concept in regular mixed ability classes (i.e., compared with that of similar ability students in special classes for disadvantaged students) because of constant social comparisons with higher ability students. Indeed, empirical evidence for children who have a learning disadvantage and are in mainstream and support classes suggests that the negative BFLPE from social comparison is stronger than the positive benefits of studying in regular classes (Tracey et al., 2003; Marsh et al., 2006a). Research has shown that learning-disadvantaged students in mainstream classes have significantly lower academic self-concepts than their counterparts in special classes, which thus argues against the applicability of labeling theory to the self-concept of learning-disadvantaged students in this kind of mainstream schooling.

Although the above findings do not mean that the movement for inclusive education should be disbanded, teachers and parents should be aware of the potentially negative effects of social comparison when disadvantaged students are placed in mainstream classes. If policy makers decide to adopt mainstreaming, it might be necessary to exert additional effort to reduce the potentially negative effects on students' self-concept, for example, by minimizing intraclass comparisons. Similarly, it is necessary to be equally cautious when segregating gifted-and-talented students into special classes or segregating students by ability within schools (i.e., tracking, see Hattie, 2002), which might reduce rather than enhance students' self-concept.

As shown earlier, the BFLPE is consistent and relatively unaffected by other potential mediators (Marsh et al., 2011; Seaton et al., 2010). The discouraging implication of this fact is that teachers might have little ability to reduce the negative effects of the BFLPE (e.g., creating a cooperative rather than a competitive classroom atmosphere, hoping to reduce the negative BFLPE). If teachers find that the negative BFLPE is having a damaging effect on ability-tracked classes, it is likely that changing to a mixed ability class structure will be more effective than attempting drastic measures to alter classroom culture, such as moving toward ability-segregated classes.

Summary and Conclusion

The results from the multitude of previous empirical studies on self-concept have helped to support, refine, and advance the theories on which such studies were built. Importantly, domain-specific self-concept rather than general self-esteem has been demonstrated and is theorized to be more useful in the interpretation and understanding of various psychological constructs. Interventions such as improving the verbal competencies of students through enhancement of their self-concept must take account of the multidimensional characteristic of self-concept. Further, it is necessary to work on specific domains of self-concept rather than apply traditional enhancement strategies to general self-concept. Similarly, according to the REM, educational interventions must focus on both improving students' levels of achievement and enhancing self-concept in order to be effective and long lasting.

The present article has demonstrated the successful synergy and interplay between theory building and empirical research. The development of instruments and the design of relevant empirical studies can now be based on strong theoretical models. These empirical studies are substantive, theoretically important, and relevant to policy makers because they are grounded on strong theories. In turn, the findings from these studies support the revision and refinement of the theories. Therefore, this methodological/substantive synergy offers crucial construct validity for any research project (Marsh et al., 2006). The research on self-concept demonstrates the successful interplay among the inextricably intertwined notions of instrument development, empirical research, and theory building.

See also: Academic Achievement Motivation, Development of; Ethnicity and Educational Achievement; Gender and Academic Motivation; Internal/External Frame of Reference Model; Motivation, Familial Influences on; Motivation, Learning, and Instruction; Race and Academic Motivation; School Achievement: Motivational Determinants and Processes; Self-Concept: From Unidimensional to Multidimensional and Beyond; Self-Concepts: Educational Aspects; Self-Esteem.

Bibliography

Baumeister, R.F., Campbell, J.D., Krueger, J.I., Vohs, K.D., 2003. Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? Psychological Science in the Public Interest 4, 1–44.

Baumeister, R.F., Campbell, J.D., Krueger, J.I., Vohs, K.D., 2005. Exploding the self-esteem myth. Scientific American 292, 84–92.

Branden, N., 1994. Six Pillars of Self-esteem. Bantam, New York.

Byrne, B.M., 1984. The general/academic self-concept nomological network: a review of construct validation research. Review of Educational Research 54, 427–456.

Byrne, B.M., 1996. Measuring Self-concept across the Life Span: Issues and Instrumentation. American Psychological Association, Washington, DC.

Chiu, M.S., 2012. The internal/external frame of reference model, big-fish-little-pond effect, and combined model for mathematics and science. Journal of Educational Psychology 104, 87–107.

Coopersmith, S.A., 1967. The Antecedents of Self-esteem. W.H. Freeman, San Francisco.

Dai, D.Y., 2002. Incorporating parent perceptions: a replication and extension study of the internal-external frame of reference model of self-concept development. Journal of Adolescent Research 17, 617–645.

Dijkstra, P., Kuyper, H., van der Werf, G., Buunk, A.P., van der Zee, Y.G., 2008. Social comparison in the classroom: a review. Review of Educational Research 78, 828–879.

Guay, F., Marsh, H.W., Boivin, M., 2003. Academic self-concept and academic achievement: developmental perspectives on their causal ordering. Journal of Educational Psychology 95, 124–136.

Haney, P., Durlak, J.A., 1998. Changing self-esteem in children and adolescents: a meta-analytic review. Journal of Clinical Child Psychology 27, 423–433.

Hattie, J., 1992. Self-concept. Erlbaum, Hillsdale, NJ.

Hattie, J., 2002. Classroom composition and peer effects. International Journal of Educational Research 37, 449–481.

Marsh, H.W., 1984. Self-concept: the application of a frame of reference model to explain paradoxical results. Australian Journal of Education 28, 165–181.

Marsh, H.W., 1987. The hierarchical structure of self-concept and the application of hierarchical confirmatory factor analysis. Journal of Educational Measurement 24, 17–19.

Marsh, H.W., 1990a. The causal ordering of academic self-concept and academic achievement: a multiwave, longitudinal panel analysis. Journal of Educational Psychology 82, 646–656.

Marsh, H.W., 1990b. The structure of academic self-concept: the Marsh/Shavelson model. Journal of Educational Psychology 82, 623–636.

Marsh, H.W., 1991. The failure of high ability high schools to deliver academic benefits: the importance of academic self-concept and educational aspirations. American Educational Research Journal 28, 445–480.

Marsh, H.W., 2007. Self-concept Theory, Measurement and Research into Practice: The Role of Self Concept in Educational Psychology. British Psychological Society, London.

Marsh, H.W., 2008. A multidimensional, hierarchical model of self-concept: an important facet of personality. In: Boyle, G.J., Matthews, G., Saklofske, D.H. (Eds.), The SAGE Handbook of Personality Theory and Assessment, Personality Theories and Models, vol. 1. Sage, Thousand Oaks, CA, pp. 447–469.

Marsh, H.W., Bornmann, L., Mutz, R., Daniel, H.-D., O'Mara, A., 2009. Gender effects in the peer reviews of grant proposals: a comprehensive meta-analysis comparing traditional and multilevel approaches. Review of Educational Research 79, 1290–1326.

Marsh, H.W., Byrne, B.M., Shavelson, R.J., 1988. A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. Journal of Educational Psychology 80, 366–380.

Marsh, H.W., Chessor, D., Craven, R.G., Roche, L., 1995. The effects of gifted-and-talented programmes on academic self-concept: the big fish strikes again. American Educational Research Journal 32, 285–319.

Marsh, H.W., Craven, R., 1997. Academic self-concept: beyond the dustbowl. In: Phye, G. (Ed.), Handbook of Classroom Assessment: Learning, Achievement, and Adjustment. Academic Press, Orlando, FL, pp. 131–198.

Marsh, H.W., Craven, R.G., 2005. A reciprocal effects model of the causal ordering of self-concept and achievement. New support for the benefits of enhancing self-concept. In: Marsh, H.W., Craven, R.G., McInerney, D.M. (Eds.), International Advances in Self Research: New Frontiers for Self Research. Information Age, Greenwich, CT, pp. 15–52.

Marsh, H.W., Craven, R.G., 2006. Reciprocal effects of self-concept and performance from a multidimensional perspective. Beyond seductive pleasure and unidimensional perspectives. Perspectives on Psychological Science 1, 133–163.

Marsh, H.W., Craven, R.G., Debus, R., 1991. Self-concepts of young children aged 5 to 8: their measurement and multidimensional structure. Journal of Educational Psychology 83, 377–392.

Marsh, H.W., Craven, R.G., Debus, R., 1998. Structure, stability, and development of young children's self-concepts: a multicohort-multioccasion study. Child Development 69, 1030–1053.

Marsh, H.W., Hau, K.T., 2003. Big-fish-little-pond effect on academic self-concept: a cross-cultural (26 country) test of the negative effects of academically selective schools. American Psychologist 58, 364–376.

Marsh, H.W., Hau, K.T., 2004. Explaining paradoxical relations between academic self-concepts and achievements: cross-cultural generalisability of the internal-external frame of reference predictions across 26 countries. Journal of Educational Psychology 96, 56–67.

Marsh, H.W., Hau, K.T., Kong, C.K., 2002. Multilevel causal ordering of academic self-concept and achievement: Influence of language of instruction (English compared with Chinese) for Hong Kong students. American Educational Research Journal 39, 727–763.

Marsh, H.W., Köller, O., 2003. Bringing together two theoretical models of relations between academic self-concept and achievement. In: Marsh, H.W., Craven, R.G., McInerney, D.M. (Eds.), International Advances in Self Research, vol. 1. Information Age, Greenwich, CT, pp. 17–48.

Marsh, H.W., Kong, C.K., Hau, K.T., 2000. Longitudinal multilevel modelling of the big-fish-little-pond effect on academic self-concept: counterbalancing social comparison and reflected glory effects in Hong Kong high schools. Journal of Personality and Social Psychology 78, 337–349.

Marsh, H.W., Martin, A.J., 2011. Academic self-concept and academic achievement: relations and causal ordering. British Journal of Educational Psychology 81, 59–77.

Marsh, H.W., Martin, A.J., Hau, K.T., 2006a. A multiple method perspective on self-concept research in educational psychology: a construct validity approach. In: Eid, M., Diener, E. (Eds.), Handbook of Multimethod Measurement in Psychology. American Psychological Association, Washington DC, pp. 441–456.

Marsh, H.W., O'Mara, A.J., 2008. Reciprocal effects between academic self-concept, self-esteem, achievement, and attainment over seven adolescent years: unidimensional and multidimensional perspectives of self-concept. Personality and Social Psychology Bulletin 34, 542–552.

Marsh, H.W., O'Mara, A.J., 2010. Long-term total negative effects of school-average ability on diverse educational outcomes: direct and indirect effects of the big-fish-little-pond effect. German Journal of Educational Psychology 24, 51–72.

Marsh, H.W., Peart, N., 1988. Competitive and cooperative physical fitness training programmes for girls: effects on physical fitness and on multidimensional self-concepts. Journal of Sport and Exercise Psychology 10, 390–407.

Marsh, H.W., Richards, G., Barnes, J., 1986. Multidimensional self-concepts: a long-term follow-up of the effect of participation in an outward bound programme. Personality and Social Psychology Bulletin 12, 475–492.

Marsh, H.W., Rowe, K.J., 1996. The negative effects of school-average ability on academic self-concept – an application of multilevel modelling. Australian Journal of Education 40, 65–87.

Marsh, H.W., Scalas, L.F., 2010. Self-concept in learning: reciprocal effects model between academic self-concept and academic achievement. In: Peterson, P., Baker, E., McGaw, B. (Eds.), International Encyclopedia of Education, vol. 6. Elsevier, Oxford, UK, pp. 660–667.

Marsh, H.W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K.T., O'Mara, A.J., Craven, R.G., 2008. The big-fish-little-pond-effect stands up to critical scrutiny: implications for theory, methodology, and future research. Educational Psychology Review 20, 319–350.

Marsh, H.W., Tracey, D.K., Craven, R.G., 2006b. Multidimensional self-concept structure for preadolescents with mild intellectual disabilities: a hybrid multigroup-MIMC approach to factorial invariance and latent mean differences. Educational and Psychological Measurement 66, 795–818.

Marsh, H.W., Trautwein, U., Lüdtke, O., Baumert, J., Köller, O., 2007. The big-fish-little-pond effect: persistent negative effects of selective high schools on self-concept after graduation. American Educational Research Journal 44, 631–669.

Marsh, H.W., Xu, M., Martin, A.J., 2011. Self-concept: a synergy of theory, method, and application. In: Harris, K., Graham, S., Urdan, T. (Eds.), APA Educational Psychology Handbook, vol. 1. American Psychological Association, Washington, DC, pp. 427–458.

Marsh, H.W., Yeung, A.S., 1997. Coursework selection: relations to academic self-concept and achievement. American Educational Research Journal 34, 691–720.

Möller, J., Marsh, H.W., 2013. Dimensional comparison theory. Psychological Review 120, 544–560.

Möller, J., Pohlmann, B., Köller, O., Marsh, H.W., 2009. A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. Review of Educational Research 79, 1129–1167.

Nagengast, B., Marsh, H.W., 2011. The negative effect of school-average ability on science self-concept in the UK, the UK countries and the world: the big-fish-little-pond-effect for PISA 2006. Educational Psychology 31, 629–656.

Nagengast, B., Marsh, H.W., 2012. Big fish in little ponds aspire more: mediation and cross-cultural generalizability of school-average ability effects on self-concept and career aspirations in science. Journal of Educational Psychology 104, 1033–1053.

O'Mara, A.J., Marsh, H.W., Craven, R.G., Debus, R., 2006. Do self-concept interventions make a difference? A synergistic blend of construct validation and meta-analysis. Educational Psychologist 41, 181–206.

OECD, 2004. Learning for tomorrow's world – first results from PISA 2003. OECD, Paris.

OECD, 2007. PISA 2006: Science competencies for tomorrow's world. OECD, Paris.

Rosenberg, M., Schooler, C., Schoenbach, C., Rosenberg, F., 1995. Global self-esteem and specific self-esteem: different concepts, different outcomes. American Sociological Review 60, 141–156.

Seaton, M., Marsh, H.W., Craven, R.G., 2009. Earning its place as a pan-human theory: universality of the big-fish-little-pond effect across 41 culturally and economically diverse countries. Journal of Educational Psychology 101, 403–419.

Seaton, M., Marsh, H.W., Craven, R.G., 2010. Big-fish-little-pond-effect: generalizability and moderation - two sides of the same coin. American Educational Research Journal 47, 390–433.

Shavelson, R.J., Hubner, J.J., Stanton, G.C., 1976. Self-Concept: Validation of construct interpretations. Review of Educational Research 46, 407–441.

Suls, J. (Ed.), 1993. Psychological Perspectives on the Self, vol. 4. Erlbaum, Hillsdale, NJ.

Tracey, D.K., Marsh, H.W., Craven, R.G., 2003. Self-concepts of preadolescent students with mild intellectual disabilities: Issues of measurement and educational placement. In: Marsh, H.W., Craven, R.G., McInerney, D.M. (Eds.), International Advances in Self Research, vol. 1. Information Age, Greenwich, CT, pp. 203–230.

Trautwein, U., Lüdtke, O., Marsh, H.W., Köller, O., Baumert, J., 2006. Tracking, grading, and student motivation: using group composition and status to predict self-concept and interest in ninth-grade mathematics. Journal of Educational Psychology 98, 788–806.

Valentine, J.C., DuBois, D.L., 2005. Effects of self-beliefs on academic achievement and vice-versa: Separating the chicken from the egg. In: Marsh, H.W., Craven, R.G., McInerney, D.M. (Eds.), International Advances in Self Research, vol. 2. Information Age, Greenwich, CT, pp. 53–78.

Valentine, J.C., DuBois, D.L., Cooper, H., 2004. The relations between self-beliefs and academic achievement: a systematic review. Educational Psychologist 39, 111–133.

Wylie, R.C., 1989. Measures of Self-concept. University of Nebraska Press, Lincoln.

Zell, E., Alicke, M.D., 2009. Contextual neglect, self-evaluation, and the frog-pond effect. Journal of Personality and Social Psychology 97, 467–482.

Access: Geographical

Emily Talen, Arizona State University, Tempe, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Access in a geographical context is the quality of having interaction with, or passage to, a particular good, service, facility, or other phenomenon that exists in the spatiotemporal world. Access is a relative concept that varies according to the level of opportunity afforded and in the sense that assessments of access (or lack of access) are made meaningful by comparing access in one zone (or for one type of individual) to access in (or for) another. If goods are spatially specific, geographical access typically involves one or more origins and one or more destinations and the distance between them. Emerging definitions of access are concerned with the role of information technologies, environmental justice issues, and normative views. The role of access in defining spatial equity is also paramount. Measurement of access involves not only the characteristics of origins and destinations, but characteristics of the travel mode and route between them. Also, the characteristics of the individual seeking access affect how access is measured. Accessibility is measured in a variety of ways, and there can be significant variation in the resulting measurement depending on the method used.

Definition and Meaning

Access in a geographical context is the quality of having interaction with, or passage to, a particular good, service, facility, or other phenomenon that exists in the spatiotemporal world. For example, access may be based on measuring the distance or travel time between where residents live (housing units) and the facilities they need (e.g., medical facilities, shops, workplaces). Access is also a relative concept that varies according to the level of opportunity afforded at the destination. Assessments of access (or lack of access) are made meaningful by comparing access in one zone (or for one type of individual) with access in (or for) another. If goods are spatially specific, geographical access typically involves one or more origins and one or more destinations and the distance between them.

Alternative Definitions

While the above definition of access is common, other concepts of access are gaining importance. One emerging view is that the notion of access must be redefined for the information age, whereby transactions take place in virtual as opposed to physical space or some hybrid form. Part of this interest relates to the notion of varying levels of access to information technologies and how this variation affects matters of equity in a wide variety of ways. But there is also an attempt to understand how information technology has changed accessibility patterns by changing the geographic locations of people and the built environment that sustains them. If information technologies affect patterns of land use, for example, such technologies indirectly affect accessibility patterns that are determined by land use configurations. Another complexity is that access does not have to be viewed as a positive phenomenon. Access to goods can also be negative, as in the case of environmentally hazardous areas, dilapidated buildings, or other services and facilities viewed as having negative externalities. When these negative costs associated with access are analyzed, the issue is one of environmental justice and whether there are discriminatory patterns of negative access along racial or economic lines. Studies have shown that access to negative conditions in the environment is often higher among low-income groups (Bowen et al., 1995).

Some views of access are not based on distances between two or more locations in space, but may instead be based on social factors, cultural barriers, or ineffective design. For example, barriers to access may be based on whether or not an individual possesses a certain subjectively defined level of 'citizenship' (Staeheli and Thompson, 1997). In addition to exclusionary practices that prohibit certain groups from 'free' access to a given good, problems inherent to the good, service, or place itself can affect access. For example, public parks may or may not be designed appropriately to deter crime by incorporating defensible space techniques, which in turn may significantly affect access to that space.

Spatial Equity and Access

Access defined on the basis of spatial distributions invokes the concept of spatial equity. The issue is one of who has access to a particular good or service and who does not, and whether there is any pattern to these varying levels of access. Spatial equity can be defined as equality, in which everyone receives the same public benefit (i.e., access), regardless of socioeconomic status, willingness to pay, or other criteria. Alternatively, access equity may vary according to indicators such as poverty, race, or the nature of the service being provided.

In a distance-based analysis, the goal might be to determine if access to a particular good is discriminatory. Such inquiries might entail, more specifically, detecting whether there is a spatial pattern to varying levels of access, and whether that spatial pattern varies according to spatially defined socioeconomic groups (Talen, 1998). For example, do people of minorities have to travel further to gain access to public goods than do others?

An overview by Marsh and Schilling (1994) on the measurement of equity in facility location analysis lists the contributions of political scientists, sociologists, economists, geographers, and management scientists. Many divergent modes

 http://dx.doi.org/10.1016/B978-0-08-097086-8.72002-2

of analysis are discernible. Health-care professionals are often interested in the degree to which access to health care is properly distributed not only to prevent health crises, but to ensure that overall population health is maintained (Guagliardo, 2004).

Access based on need is a matter of matching resources and facilities to target populations. Examples include matching HIV service providers to target populations (Fulcher and Kaukinen, 2005), or connecting food supermarkets to socially deprived areas, or those who could benefit most from having access to food retailers (Apparicio et al., 2007). Savannah, Georgia has analyzed spatial access to facilities to direct city resources into the neighborhoods judged to be 'the worst off' (Toulmin, 1988). Health organizations are often involved, participating in 'mapping audits' of health service provision, now made easy through the use of geographic information systems (GIS). Higgs (2004) reviewed the use of GIS-based measures of access to health care among government entities intent on increasing a 'social justice policy agenda' (p. 119).

In addition to exposing differentials in accessibility, research seeks to discover why certain patterns of access exist. Factors implicated include urban form, organizational rules, citizen contacts, politics, and race. Until recently, spatial inequity has been explained predominantly by the notion of unpatterned inequality (Mladenka, 1980). This is the idea that although inequality in people's access to services and facilities exists, no evidence of a clear discriminatory pattern can be found. In the absence of patterned inequality, some argue, it is difficult to attach blame to those responsible for the existing distributional pattern. Some critiques of this theory (Miranda and Tunyavong, 1994) focus on the failure to take the political process properly into account as well as the problem of variable definition.

The complexities involved in the equitable access of public resources include not only methodology (how can equitable access be measured?), but also a multitude of value judgments about who should benefit, the nature of social justice, and the definition of political consensus. Measures have become more and more complex and nuanced, for example, in the application of indices of 'relative well-being' that are thought to more carefully account for inequalities (Albrecht and Ramasubramanian, 2004: p. 371). The stakes are high: since public resources are, in some sense, part of each individual's income, their spatial distribution directly affects the distribution of public welfare (Pahl, 1971; Harvey, 1973).

Distribution (access) based on demand is a matter of surveying the spatial distributions of specific kinds of facilities and matching those distributions to the specific populations who need access to them. For example, a study of access to the services needed by public housing residents, based on mapping levels of demand and need, concluded that almost half of public housing residents were not having their demands met (Apparicio and Seguin, 2006).

Normative Views of Access

Related to equity considerations, the concept of access has taken on a normative role. Accessibility has been a long-standing component of theories of good urban form (see in particular Jacobs, 1961; Lynch, 1981; Jacobs and Appleyard, 1987). Kevin Lynch (1981) held access as a key component of his theory of ideal urban form and argued that access could be used as a measure of 'settlement performance' by factoring in (1) the feature to which access is being given; and (2) the person receiving access. Increasing access to needed facilities, goods, and services among low-income populations is considered to be an essential part of social policy. The 'geography of opportunity', significantly influences the ability of low-income residents to improve their lives (Briggs, 2005). It is also understood that for locally oriented populations – residents who rely on modes of transport other than the automobile (e.g., the elderly and the poor) – accessibility to urban services may be more important because distance is not elastic (Wekerle, 1985).

Elevating the role of geographical access is closely tied to the view that settlement patterns should increase access between residents, their places of work, and the services they require. In this regard, accessibility is not conceptualized as an issue of private mobility but is generally approached as a community-wide, public problem. Equity in accessibility to resources is therefore tied to the principles of smart growth (Song and Knaap, 2004) and active living environments (Heath et al., 2006; Norman et al., 2006) in which pedestrian access to daily life needs is viewed as especially important. In this same genre, New Urbanists have developed a specific town-planning manifesto based on enhancing access at the level of region (by promoting a variety of transportation alternatives), metropolis (by promoting compact urban form), and neighborhood (by promoting mixed uses and housing density).

Of course, other factors challenge the idea that accessibility is a crucial goal. Notably, there is recognition of the fact that physical distance is lessening in importance, technology is obfuscating the need for centered settlements, and low-density dispersed development is merely the inevitable outcome of the post-Fordist urbanization process (Bruegmann, 2005). With the emphasis on information technology and digital methods of social connection, the idea of access may be downplayed as tangential, despite the importance of pedestrian-oriented access for those who lack mobility (the elderly and disabled, the poor, working parents). Downplaying physical access may mean that access is not deemed relevant or that market forces are deemed too difficult to overcome. In practice, the role of accessibility in transportation planning might be reframed as congestion reduction or efficiency. Others might frame the issue as one of increasing people's land use and transportation choices (Levine, 1998), or emphasizing that changes to transportation systems have profound consequences for environmental justice (Chakraborty et al., 1999).

The Measurement of Access

The majority of studies of geographic access assume that access is a positive phenomenon, and that access is based in part on some measurement of distance in space. If access is being considered as something desirable, the impediments to access – friction or blockage of the opportunity to interact or the right to enter – must be factored in. Right of entry assumes that a transaction must occur between the consumer and the good, service, or facility. More important in terms of definitions of access is that these transactions have a cost associated

with them. Geographical interest in access therefore is often focused on these transaction costs. The interest is often methodological – how can transaction costs be measured? Empirical investigations focusing on, for example, who pays a higher transaction cost for access and why invoke the issue of spatial equity discussed above.

Factors Affecting the Measure of Access

Five classes of factors affect the measure of access. The first two are simply the spatial locations of points of origin and points of destination. Usually, the points of origin refer to housing locations and points of destination involve entities that can be spatially referenced, such as schools or places of employment.

The third factor is the travel route and its distance between an origin(s) and destination(s). This involves not only the distance between two or more points, but the qualities of the route and the mode of travel that occurs on that route. Route factors include topography, design speed, number of lanes of traffic, and mode. For pedestrian access, perceived safety, sidewalk quality, and traffic volumes are important. Measuring the distance along a route can be based on the shortest distance between destination and origin, or can be more complex and involve a variety of spatial networks.

Another factor affecting the measurement of access has to do with the attributes of the people who seek access; these characteristics are usually measured for spatial aggregations such as a census tract or census block (the degree of disaggregation of the spatial unit varies widely). Factors that might affect access include socioeconomic status, age, gender, and employment status. Certain assumptions can be made about the attractiveness or relevance of travel to certain facilities (and the likely mode of travel) based on these characteristics. The frictional effect of the available travel mode is also likely to be predicated on the characteristics of residents. For example, lack of bus service may adversely affect access for low-income individuals but have only a marginal impact on high-income groups.

A final class of factors concerns the size, type, and quality of the destination (e.g., facility). These attributes determine the attractiveness of a destination for consumers and therefore affect how access to it is measured.

Types of Measures

Accessibility can be measured in a variety of ways. At one end of the spectrum, accessibility measures can be based on random utility theory, where access is based on the desirability or utility of a set of destination choices for an individual (Handy and Niemeier, 1997). Another similar approach uses travel diaries of individuals to estimate 'personal accessibility', based on daily movement patterns (Kwan, 1999). At the other end of the spectrum, accessibility is characteristic of place. Ultimately, how access is measured should be based on how the measure is to be used. For the visualization of spatial equity, it may make sense to focus on the measurement of 'place-based' accessibility – a measure that serves as a characteristic of a place. This conceptualization of access carries with it a number of variations.

One of the most widely used measurement methods is known as the 'container' approach, which is simply a count of the number of facilities within a given area such as a census tract, political district, or municipal boundary. Alternatively, a 'minimum distance' approach measures access as the distance to the nearest facility (often used in research on access to health-care services), whereas the 'travel cost' approach calculates the distance (cost) between an origin and all included destinations (often the preferred method for calculating access to amenities such as parks within a city). Using the basic idea of 'proximity as a measure of access', Lindsey et al. (2001) looked at the equality of access between urban greenways and populations groups, finding that 'minorities and the poor have disproportionate access to trails'.

Another option is to use the gravity 'potential' measure, where facilities are weighted by their size (or other characteristic) and adjusted for the frictional effect of distance. Pacione (1989) examined differences in access to secondary schools by compiling mapped indices of access using a gravity-based model, which revealed different 'undulating' surfaces that could be used as a basis for comparing the effects of school closings on access. In addition to visual comparison of spatial variation, Pacione derived an overall coefficient of variation, by which the aggregate effect on the overall access of different spatial patterns of facilities can be compared. The socioeconomic characteristics of the neighborhoods that fared better and those that fared worse, in terms of facility provision, were presented to reveal any underlying patterns of distributional bias. Geertman and Ritsema Van Eck (1995) also showed how to produce maps of 'potential surfaces', using the gravity potential method of measuring accessibility. Although socioeconomic variables were not included, the authors showed how the maps can be used to visually identify, for example, potential building sites with adequate public transport.

Some researchers have compared the results of using multiple access measures. Brabyn and Gower (2004), for example, compared the ratio, least-cost path, and allocation methods to determine how the interpretation of access to medical providers changes. Luo and Wang (2003) synthesized two accessibility measures, a gravity-based method and a floating catchment area method, to analyze accessibility to health care in a GIS-based visual assessment.

A primary question is what characterization of access is most suitable? Largely, the answer depends on how best to characterize distance between the user and the facility. If evaluating access to amenities, the value of the facility to the user declines with distance. If the analysis concerns emphasizing the effect of distance as a deterrent, then the gravity model may be appropriate. Alternatively, it may be appropriate to compare access to facilities as an average of all distances to all facilities (travel cost measure), which treats the resources of a city as a complete package of public goods. If the goal is to assess how to minimize the inequality of nearest distance between origin and destination, then a minimum distance measure is appropriate. Finally, if the goal is to maximize the number of people 'covered', and if it is decided that beyond a given radius, users acquire no benefit, then a covering approach may be warranted.

If the purpose of access measurement is to determine neighborhood service provision, 'covering' and 'minimum distance' are probably the most appropriate. This is because

access at that scale essentially revolves around two concepts of access: how far does one have to walk to reach a destination, and how many urban opportunities are within walking distance? For the first question, minimum distance is measured between some point of origin (such as a census block centroid) and the point location (street address) of the given facility. For the second question, a covering radius is drawn around the point of origin, and the number of facilities within, for example, walking distance, is determined.

For each measurement approach, a number of different measurement factors have to be considered. First, points of origin have to be selected; these could be point locations of individual housing units on a parcel or the centroids of larger geographic units (e.g., census blocks, block groups, or tracts). An obvious issue in this selection is the potential introduction of aggregation error that results when a single point is chosen to represent an entire area. At a minimum, researchers will need to address the problem of ecological fallacy and the modifiable areal unit problem. Hewko et al. (2002) looked at spatial accessibility for neighborhoods and found that the measures used were 'susceptible to numerous methodological problems' (p. 1185), particularly when the amenity being investigated had many locations.

Another measurement issue concerns population characteristics. If the goal of an accessibility study is social equity analysis, analysts will want to consider the socioeconomic characteristics of a population, since such information allows some assumptions to be made about level of need and travel mode availability. For example, it is likely that residents in poorer areas are more dependent on public transit, have lower access to private automobiles, and therefore require greater access than do residents of high-income areas. Even though distance is the most critical factor in determinations of access one must recognize that in low-income areas distance is more difficult for residents to overcome and lower levels of accessibility are likely to be particularly detrimental.

See also: Discrimination and the Law; Discrimination, Economics of; Discrimination: Racial; Information Society, Geography of; Justice, Access To; Location, Absolute and Relative; Spatial Analysis in Geography; Spatial Equity; Spatial Pattern, Analysis of; Transportation Geography.

Bibliography

Albrecht, Jochen, Ramasubramanian, Laxmi, 2004. Journal of Medical Systems 28 (4), 371–384.

Apparicio, Philippe, Seguin, Anne-Marie, 2006. Measuring the accessibility of services and facilities for residents of public housing in Montreal. Urban Studies 43 (1), 187–211.

Apparicio, Philippe, Cloutier, Marie-Soleil, Shearmur, Richard, 2007. The case of Montreal's missing food deserts: evaluation of accessibility to food supermarkets. International Journal of Health Geographics 6 (4). http://www.ij-healthgeographics.com/content/6/1/4.

Bowen, W.M., Salling, M.J., Haynes, K.E., Cyran, E.J., 1995. Toward environmental justice: spatial equity in Ohio and Cleveland. Annals of the Association of American Geographers 85, 641–663.

Brabyn, Lars, Gower, Paul, 2004. Comparing three GIS techniques for modeling geographical access to general practitioners. Cartographica 39 (2), 41–49.

Briggs, X. de Souza, 2005. The Geography of Opportunity: Race and Housing Choice in Metropolitan America. Brookings Institution Press, Washington, DC.

Bruegmann, Robert, 2005. Sprawl: A Compact History. University of Chicago Press, Chicago.

Chakraborty, Jayajit, Forkenbrock, David J., Schweitzer, Lisa A., 1999. Using GIS to assess the environmental justice consequences of transportation system changes. Transactions in GIS 3 (3), 239–258.

Fulcher, Christopher, Kaukinen, Catherine, 2005. Mapping and visualizing the location HIV service providers: an exploratory spatial analysis of Toronto neighborhoods. AIDS Care 17 (3), 386–396.

Geertman, Stan C.M., Ritsema Van Eck, Jan R., 1995. GIS and models of accessibility potential: an application in planning. International Journal of Geographical Information Systems 9 (1), 67–80.

Guagliardo, Mark F., 2004. Spatial accessibility of primary care: concepts, methods and challenges. International Journal of Health Geographics 3. http://www.ij-healthgeographics.com/content/3/1/3.

Handy, Susan L., Niemeier, Debbie A., 1997. Measuring accessibility: an exploration of issues and alternatives. Environment and Planning A 29, 1175–1194.

Harvey, David, 1973. Social Justice and the City. Edward Arnold, London.

Heath, Gregory W., Brownson, Ross C., Kruger, Judy, Miles, Rebecca, Powell, Kenneth E., Ramsey, Leigh T., The Task Force on Community Preventive Services, 2006. The effectiveness of urban design and land use and transport policies and practices to increase physical activity: a systematic review. Journal of Physical Activity and Health 3 (Suppl. 1), S55–S76.

Hewko, Jared, Smoyer-Tomic, Karen E., John Hodgson, M., 2002. Measuring neighbourhood spatial accessibility to urban amenities: does aggregation error matter? Environment and Planning A 34, 1185–1206.

Higgs, Gary, 2004. A literature review of the use of GIS-based measures of access to health care services. Health Services & Outcomes Research Methodology 5, 119–139.

Jacobs, Allan, Appleyard, Donald, 1987. Toward an urban design manifesto. Journal of the American Planning Association (Winter), 112–120.

Jacobs, Jane, 1961. The Death and Life of Great American Cities. Vintage Books, New York.

Kwan, Mei-Po, 1999. Gender and individual access to urban opportunities: a study using space-time measures. Professional Geographer 51, 210–227.

Levine, Jonathan, 1998. Rethinking accessibility and jobs-housing balance. Journal of the American Planning Association 64 (2), 133–149.

Lindsey, Greg, Maraj, Maltie, Kuan, SonCheong, 2001. Access, equity, and urban greenways: an exploratory investigation. The Professional Geographer 53 (3), 332–346.

Luo, Wei, Wang, Fahui, 2003. Measures of spatial accessibility to health care in a GIS environment: synthesis and a case study in the Chicago region. Environment and Planning B: Planning and Design 30, 865–884.

Lynch, Kevin, 1981. Good City Form. MIT Press, Cambridge, MA.

Marsh, Michael T., Schilling, David A., 1994. Equity measurement in facility location analysis: a review and framework. European Journal of Operational Research 74 (1), 1–17.

Miranda, Rowan A., Tunyavong, I., 1994. Patterned inequality? Reexamining the role of distributive politics in urban service delivery. Urban Affairs Quarterly 29, 509–534.

Mladenka, Kenneth R., 1980. The urban bureaucracy and the Chicago political machine: who gets what and the limits to political control. American Political Science Review 74, 991–998.

Norman, Gregory J., Nutter, Sandra K., Ryan, Sherry, Sallis, James F., Calfas, Karen J., Patrick, Kevin, 2006. Community design and access to recreational facilities as correlates of adolescent physical activity and body-mass index. Journal of Physical Activity and Health 3 (Suppl. 1), S118–S128.

Pacione, Michael, 1989. Access to urban services—the case of secondary schools in Glasgow. Scottish Geographical Magazine 105, 12–18.

Pahl, R., 1971. Poverty and the urban system. In: Chisholm, M., Manners, G. (Eds.), Spatial Policy Problems of the British Economy. Cambridge University Press, London.

Song, Yan, Knaap, Gerrit-Jan, 2004. Measuring urban form. Is Portland winning the war on sprawl? Journal of the American Planning Association 70 (2), 210–225.

Staeheli, Lynn A., Thompson, A., 1997. Citizenship, community and struggles for public space. The Professional Geographer 49, 28–38.

Talen, Emily, 1998. Visualizing fairness: equity maps for planners. Journal of the American Planning Association 64, 22–38.

Toulmin, Llewellyn M., 1988. Equity as a decision rule in determining the distribution of urban public services. Urban Affairs Quarterly 23 (3), 389–413.

Wekerle, Gerda R., 1985. From refuge to service center: neighborhoods that support women. Sociological Focus 18, 79–95.

Acculturation

David L Sam, University of Bergen, Bergen, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A 1936 definition of acculturation serves as the point of departure in discussing the concept of acculturation. Psychological acculturation (as individual-level phenomenon) is the main focus of the article. Three main theoretical perspectives that define the field are presented. The article also looks at acculturation as a process (focusing on acculturation strategies), as an outcome (focusing on psychological and sociocultural adaptations), and as the relationship between processes and outcomes. Future directions of the field are identified.

Acculturation as a concept encompasses all the changes that occur when individuals and groups of people belonging to different cultural backgrounds come into contact with each other. The term is commonly used when discussing what happens to immigrants, refugees, and ethnocultural groups (particularly so-called ethnic minorities) when they come to live in another society. In recent years, the term has acquired several meanings as it has become associated with concepts such as *globalization*, *multiculturalism*, *integration*, and *assimilation*. While acculturation is closely related to these concepts (see Sam and Berry, 2006 for a discussion), they should not be confused with them.

Powell (1883) is suggested to have been the first to use the term 'acculturation' in the English language in the 1880s, to refer to psychological changes arising from cross-cultural interactions. McGee (1898) referred to acculturation as the process of exchange and mutual improvement by which societies advanced from savagery, through to barbarism and civilization, to enlightenment. Simons (1901), on the other hand, considered acculturation to be a two-way process of reciprocal accommodation, and equated acculturation to the term assimilation. Assimilation was defined as the process of adjustment or accommodation that occurs between the members of two different races. To avoid the different use of the term, and to enhance theory development and research, the Social Science Research Council of the United States set up a committee in the 1930s to analyze and define the parameters for acculturation as a field of inquiry within cultural anthropology. The work of this committee resulted in defining acculturation as

> those phenomena which result when groups of individuals having different cultures come into continuous first-hand contact, with subsequent changes in the original culture patterns of either or both groups…under this definition acculturation is to be distinguished from cultural change, of which it is but one aspect, and assimilation, which is at times a phase of acculturation. (Redfield et al., 1936: pp. 149–152)

This definition, in spite of its age, is still useful, as it points to a number of issues central to the concept.

Some Central Issues to the Concept

Mutuality of Influence

The expression 'changes in the original culture patterns of either or both groups' in the above definition entails *mutual* or *reciprocal* influences, which imply that both groups influence each other. However, due to power differences (such as economic power, or numerical strength), one group often exerts more influence than the other. Often times, there is a dominant group exerting more influence than another, which is less dominant. This has resulted in the assumption that only the nondominant group undergoes acculturation, and has contributed to the rise of the term *ethnic minority*. The assumption that acculturation involves changes only among ethnic minorities has also contributed to a one-sided view of acculturation, and an abundance of research on changes taking place in nondominant groups.

Continuous First-Hand Contact

Continuous first-hand contact as part of the above definition was meant to encompass such complex interactions as a foreign power colonizing another sovereign country and forcing its rule and way of life on the people colonized, to the somewhat benign situations such as living next door to a foreign national (over a prolonged period). The definition from 1936 did not factor in the rapid development in information technology, and the rise of electronic social networks, which in no doubt affect people's behavior. To what extent should virtual interactions between people who live thousands of kilometers away be considered as acculturation? This is clearly an uncharted area that deserves research attention.

Group versus Individual-Level Phenomenon

Although acculturation as a concept was first proposed as a group-level phenomenon, it was also recognized as an individual-level phenomenon. Graves (1967) coined the term psychological acculturation, which has been defined as "the changes an individual experiences as a result of being in contact with other cultures, or participating in the acculturation that one's cultural or ethnic group is

undergoing" (Berry, 1990: p. 203). A distinction between acculturation as a group-level phenomenon and acculturation as an individual-level phenomenon is important; the kinds of changes that take place at the two levels tend to be different. Moreover, not every group or individual enters into, participates in, or changes in the same way during acculturation. Vast individual differences in psychological acculturation exist, even among individuals who have the same cultural origin and reside in the same acculturative context.

Change

Change is central to acculturation and inherent in change are the notions of *process*, which is dynamic, and *an outcome* that may be relatively stable. These two notions underscore that acculturation research and theory are concerned with *how* change comes about, what changes, and the extent of the change, during acculturation (Berry et al., 2006). While cross-sectional studies may be suitable when addressing 'what changes' and the 'extent of change,' they may be insufficient when dealing with how changes come about. Longitudinal studies may be better suited when dealing with the *how* question.

Until the 1970s, acculturation researchers assumed acculturation took place along a single dimension (one-dimensional view), and this ranged from a preference for fully remaining as a member of one's heritage culture, through to becoming a full member of the dominant society. The one-dimensional view also assumed that one could not be fully embedded in the two cultures in contact; the more one was embedded in his or her heritage culture, the less one could be embedded in the new culture, and vice versa.

In contrast, Berry (1974) proposed that acculturation occurred along two independent cultures (a bidimensional view), which presently dominate the field. However, there are some ambiguities regarding how the two dimensions should be operationalized. According to Berry, the two dimensions are (1) the degree to which people wish to maintain their heritage cultures and identities, and (2) the degree to which people partake in the larger society, and interact with others. Research has identified different ways the second dimension has been operationalized. This has included identification with the larger society; as adoption of the national culture; and as adapting to the larger society. Snauwaert et al. (2003) found that an individual's preference for a particular way of acculturating (referred to as acculturation strategy) varied very much depending on how the second dimension was operationalized.

Framework and Theoretical Perspectives

A framework outlining the links between group- and individual-level acculturation has been proposed by Berry et al. (2011), which argues that any comprehensive study of acculturation needs to take into consideration key features of the two original cultures (depicted as Culture A and Culture B in **Figure 1**) prior to coming into contact with each other. Acculturating individuals and groups bring cultural and psychological qualities with them to the new society, and the new society also has a variety of such qualities. The compatibility in cultural values, norms, attitudes, etc. between the two cultural communities in contact needs to be examined as a basis for understanding the acculturation process that is set in motion. Equally important is the nature of the contact relationships between the two groups, whether it is based on domination of one group over the other, on mutual respect or hostility. There is also the need to understand the resulting cultural changes in both groups that emerge during acculturation. The psychological changes individuals undergo and how they adapt to the new situation

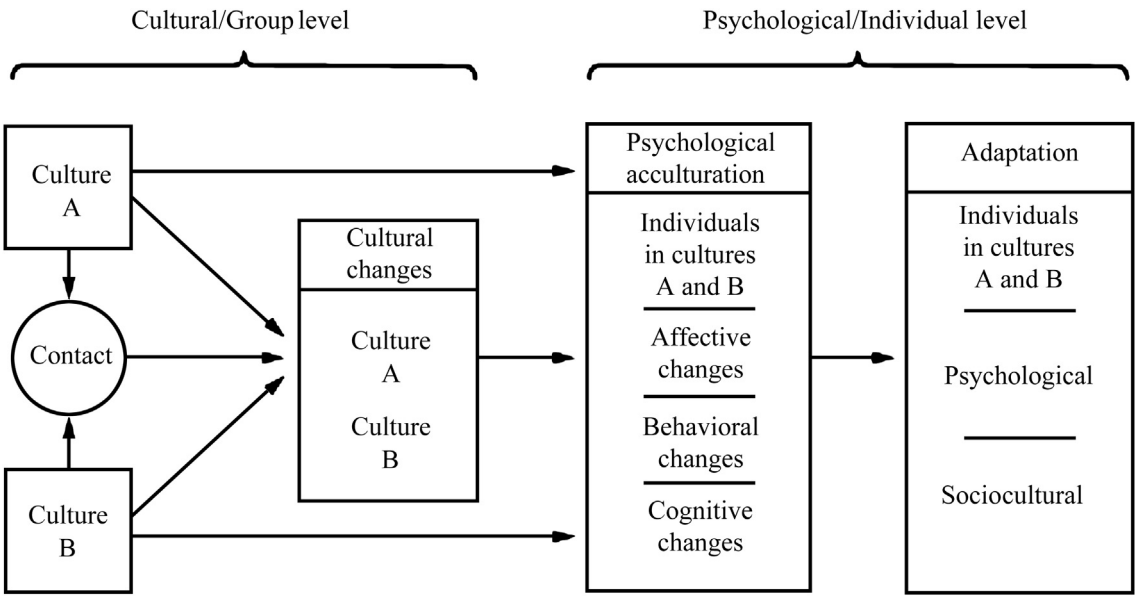

Figure 1 Framework for conceptualizing and studying psychological acculturation.

is also important to look at. Identifying these changes requires sampling a population and studying individuals who are variably involved in the process of acculturation. These changes range from simple behavioral changes (e.g., in ways of speaking, dressing, and eating); cognitive reorganization in terms of how they identify themselves and categorize the individuals and groups they interact with to more problematic outcomes, producing acculturative stress as manifested by uncertainty, anxiety, and depression. These changes at the individual level roughly correspond to what has been referred to as the ABCs of acculturation. Individuals then adapt to these changes, psychologically (e.g., sense of well-being) and/or socioculturally (e.g., acquiring the social skills for daily living).

The ABCs of Acculturation

The ABCs of acculturation is an acronym that refers to the three areas of human life that undergo psychological acculturation, namely, affective, behavioral, and cognitive aspects of acculturation (Ward, 2001). The ABCs are in turn, respectively, linked to different theoretical perspectives that dominate the field: a stress and coping theoretical framework, a culture learning approach, and a social-identification orientation to acculturation.

Affective Changes

Although Oberg's work on culture shock may be seen as the start of affective perspectives, the work of Berry on *acculturative stress* epitomizes these perspectives (see Sam and Berry, 2006). Affective perspectives emphasize the emotional aspects of acculturation and focus on such issues as psychological well-being and life satisfaction. This perspective was inspired by Lazarus and Folkman's stress model where acculturation is likened to a set of major life events that may pose challenges to the individual. These life events may qualify as stressors and provoke stress reactions in an individual, particularly when the appropriate coping strategies and social supports are lacking. When serious challenges are experienced and are appraised to be problematic because one is unable to deal with them easily, such as simply changing one's behavior (i.e., behavioral changes) then acculturative stress results. Acculturative stress is a stress reaction in response to life events that are rooted in the experience of acculturation. In line with Lazarus's stress model, not all acculturation changes result in acculturative stress because there are a number of moderating and mediating factors (both before and during the acculturation) such as personal characteristics including age and gender, and social support that may influence the perception and interpretation of the acculturation experience.

Behavioral Changes

These changes build on culture learning theory, where the central premise is that people in cultural transitions may lack the necessary skills needed to engage the new culture. This, therefore, could result in difficulties managing the everyday social encounters. To overcome these difficulties, individuals are expected to learn or acquire the culture-specific behavioral skills (such as the language) that are necessary to negotiate this new cultural milieu. Cultural learning approaches to acculturation entail gaining an understanding in intercultural communication styles, including its verbal and nonverbal components, as well as rules, conventions, and norms and their influences on intercultural effectiveness. This approach corresponds to the behavioral changes indicated in **Figure 1**. Arguably, second-language proficiency and communication competence form the core of all cultural learning approaches and ultimately sociocultural adaptation.

Cognitive Changes

These changes within the context of psychological acculturation are concerned with how people process information about their own group (in-group) and about other groups (out-groups), as well as how people identify with these categories. Individuals need to belong to a group in order to secure a sense of well-being. There is also the tendency to put others and oneself into categories, and this helps one to associate (i.e., identify) with certain groups and not others. Furthermore, humans compare the group they belong to with others, and there is a tendency to have a favorable bias toward seeing positive qualities of the group to which one belongs, thereby boosting our self-image. When individuals and groups enter into an acculturation situation, they are often faced with the questions: 'who am I?' and 'to which (ethnocultural) group do I belong?' Dealing with these questions sets into motion a process of (re)defining one's identity as an individual, and as part of the ethnic group one belongs to (i.e., ethnic identity) and that of the larger dominant society in which they now reside (i.e., national identity). There is a large body of literature that examines ethnic identity from developmental (e.g., Phinney, 1990) and from sociopsychological (e.g., Verkyten, 2005) perspectives. A new line of research within the cognitive perspective is the 'bicultural identity integration' (see Benet-Martínez and Haritatos, 2005; Wiley and Deaux, 2011).

Developmental Issues

Although not a clearly defined theoretical perspective in acculturation, these issues have come to the fore in recent years because of concerns about the future of children and adolescents from immigrant families in Western industrialized societies (see Mansten et al., 2012). A central developmental issue is that young immigrants simultaneously undergo acculturation and development (see Motti-Stefanidi et al., 2012), and as such acculturation theories should be more cognizant of how the two types of changes influence and impact each other. The issues and areas often identified as relevant for acculturation, and which could complicate normal developmental changes include cultural identity, development of self, family values, and peer relations. Recurring questions from this line of research include whether there are age differences for accomplishing developmental milestones, whether immigrant youth deal with developmental tasks in the same way as their national

peers, and whether their acculturation experiences have special impact on how they resolve developmental tasks.

Personality and Individual Factors

Psychological acculturation is based on the notion that individuals enter and differentially participate in acculturation. As such, identifying individual factors and how they impact the acculturation process and outcome has been of great interest. Existing research has examined different aspects of demographic factors and personal characteristics of the individual (broadly defined as personality). One goal of these studies has been to identify the 'overseas type' who can readily adjust to new cultural environments, by focusing on how certain characteristics of the individual (e.g., ethnocentric tendencies) affected adjustment. The list of personality traits/types that have been examined with respect to acculturation is long, ranging from attachment styles and the Big Five, through cultural intelligence to locus of control and motivation. The studies on acculturation and personality have yielded mixed results, and the explained variance of personality factors has been generally low. One reason for the mixed findings is a combination of factors ranging from the operationalization of the personality trait/type itself, to the adjustment outcome examined, and of course the ethnicity and type of acculturation person (e.g., sojourner, immigrant, etc.) of interest.

Acculturation Processes: How Do Individuals Acculturate?

The process of acculturation is concerned with how individuals deal with the meeting between their original culture and that of the new society. Working from the perspective that acculturation involves individual and group factors, it is important that factors from both levels are taken into consideration. At the group level, the important issues to consider are the policies and the expectations from the larger society (see below for a brief discussion, see also Azzi et al., 2011), as well as the expectations from the ethnocultural group that the individual belongs to.

At the individual level itself, Berry has identified two issues of importance, namely, the two dimensions discussed above. According to Berry, preferences for the two dimensions lead to the adoption of four different acculturation strategies, which he termed assimilation, integration, separation, and marginalization. Acculturation strategies depend on the extent to which an individual balances attitudes and/or behaviors around the two issues or dimensions (see Figure 2).

Assimilation is the strategy used when individuals do not wish to maintain their cultural identity and seek close interaction with other cultures (or in some cases adopt the cultural values, norms, and traditions of the new society). The separation strategy is defined by individuals who place a high value on holding on to their original culture and avoid interaction with members of the new society. The integration strategy is used by individuals with an interest in maintaining one's original culture while having daily interactions with other groups; there is some degree of cultural integrity maintained, while at the same time they seek, as a member of an ethnocultural group, to participate as an integral part of the larger social network. The marginalization strategy is defined by little possibility or lack of interest in cultural maintenance (often for reasons of enforced cultural loss) and little interest in having relations with others (often for reasons of exclusion or discrimination). The four strategies are illustrated on the left-hand side of Figure 2.

The four strategies are neither static nor end outcomes in themselves. They can change depending on situational factors (e.g., in the wake of the 9/11 attack in the United States, many

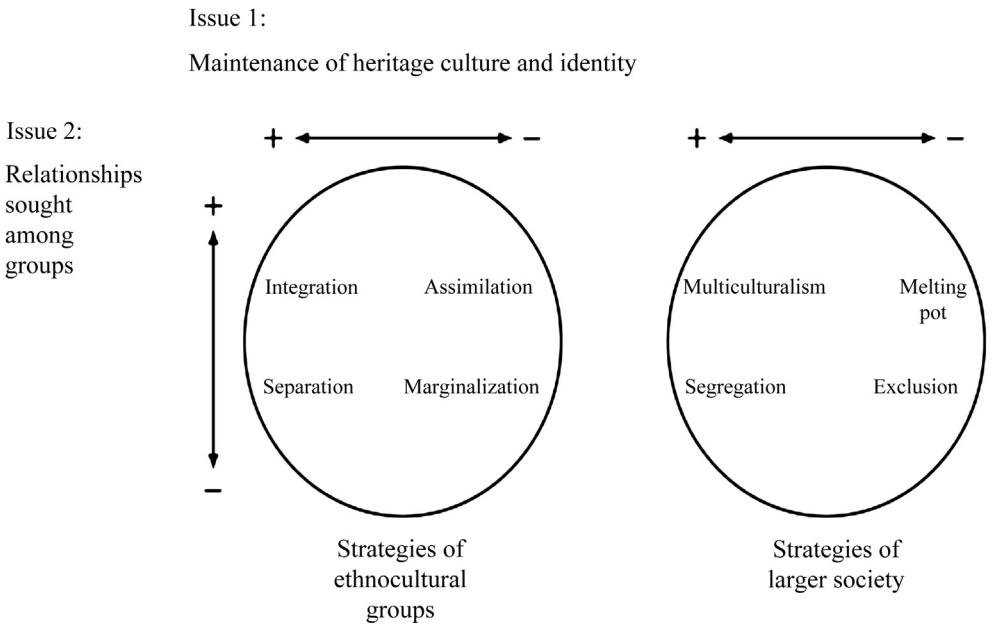

Figure 2 Acculturation strategies in ethnocultural groups and the larger society.

Muslims had to redefine their cultural identities). Moreover, individuals may utilize different strategies for different domains of life (e.g., integration in the area of language, separation in the area of intimate relationships, assimilation in the area of food habits, and marginalization in the area of religion). An individual may live with a constellation of different strategies encompassing different domains of life. This complex constellation may explain the generally low internal consistency in acculturation strategy measures. In a cluster analysis involving five intercultural areas of life – acculturation attitudes, cultural identities, language use and proficiency, peer relationships, and family relationship values among over 30 different ethnocultural groups living in 13 different countries – Berry et al. (2006) found support to the notion that acculturating individuals combine different domains of life into four acculturation profiles, which they referred to as national, ethnic, integration, and diffuse, and these roughly corresponded to the four acculturation strategies of assimilation, separation, integration, and marginalization. The term acculturation profile is used because these clusters encapsulated different aspects of intercultural relations, rather than simply attitudes.

The right-hand side of **Figure 2** illustrates the parallel concepts that are often employed when describing the public attitudes and public policies in the larger society (see Sam and Berry, 2006, for full discussion). The right-hand side of the figure suggests that the larger society may have its own expectations of how individuals should acculturate, and these may not always be consistent with the desires of nondominant individuals. Society's expectation of how individuals of the nondominant group should acculturate (i.e., acculturation expectations) has been the basis of a number of theoretical models including the Interactive Acculturation Model (Bourhis et al., 1997; see also Horenczyk et al., 2013 for discussion of these models).

Acculturation Outcomes: How Well Do People Acculturate?

From a cross-cultural psychological perspective, human behavior is an adaptation to ecological and cultural contexts (Berry et al., 2011). As such, a natural and legitimate question to ask is what happens to people's adaptive behavior when their heritage culture is altered, or comes into contact with another one? A more specific question in this regard is *how well* do people adapt to acculturation demands? A concomitant question is whether there is a relationship between *how* people acculturate and *how well* they adapt.

When discussing how well people adapt during acculturation, the interest is on the long-term outcome of acculturation. As portrayed in **Figure 1**, adaptation is not synonymous with acculturation, but a result of acculturation. In the context of acculturation, adaptation is concerned with well-being, communication competence, self-awareness, stress reduction, feelings of acceptance, and culturally skilled behaviors (see Ward, 2001). Within acculturation, Ward (2001) makes a distinction between *psychological* and *sociocultural* adaptation. Psychological adaptation refers to internal phenomenon characterized by 'feeling well.' This form of adaptation is often assessed by lack of psychological problems (e.g., depression and anxiety), presence of well-being such as self-esteem and satisfaction with life. Psychological adaptation is largely concerned with the affective changes to acculturation, and to some extent the cognitive changes as they bother on self-concept and self-definition. Sociocultural adaptation, refers to the degree to which individuals are competent in carrying out their daily lives in their new social and cultural contexts, such as at school, at work, within the larger society, and is characterized by 'doing well.' Studies of sociocultural adaptation have usually focused on the absence of behavior problems, school achievement, job satisfaction, language skills, interpersonal relations, and social competence. Studies within this domain usually focus on the behavioral changes of acculturation.

How well do people acculturate? Do acculturating people adapt equally well, better, or more poorly than some reference group? The basis for comparisons is an important issue when discussing the adaptation of people undergoing acculturation. Should acculturating people be compared with non-acculturating members of one's own ethnic group; other acculturating groups in the new society; or with members of the new and larger society? Should they be assessed against a psychometrically and standardized instrument measuring adaptation? All three reference groups for comparison as well as using a standardized instrument should ideally be utilized simultaneously. But, quite often, information on the comparison reference groups is not available, nor does the standardized instrument provide unbiased scores. Moreover, considering that everyone in contemporary society is undergoing some form of acculturation, so-called nonacculturating groups as a reference group is nothing more than hypothetical. In view of the different possible groups of people that could be used in making comparisons, it is perhaps not surprising that research findings in terms of how well acculturating groups adapt are mixed. Some studies have found good adaptation outcomes (both psychological and sociocultural) with some acculturating groups doing either better or equally well as their national peers in the society of residence (Berry et al., 2006). However, other studies have found poor adaptation outcomes (Young et al., 2010).

Aside from the issue of different reference groups being used when comparing the adaptation of acculturating individuals, mixed findings have also arisen because of differences in operationalization of acculturation itself (e.g., duration of residence and language proficiency as measures of acculturation), and the adaptation outcome of interest (see Koneru et al., 2007; for a review). Adaptation outcomes that have been studied have ranged from simple psychological problems such as anxiety to more serious ones such as schizophrenia; and from positive adaptation outcomes such as satisfaction with life, self-esteem, to negative behavioral problems. Differences in adaptation outcomes warrant not only a systematic literature review but also a meta-analytic estimate of effect sizes in group difference, and strength of association between acculturation variables and outcome.

Although much attention has been directed to psychological and sociocultural adaptation, some research has found links between acculturation and physical health including different forms of cancer and cardiovascular diseases. However, rather than acculturation resulting in poor physical

health, some studies have suggested that immigrants have better physical health, when compared with their non-migrating peers in the society of emigration. It appears that with increasing acculturation (defined in terms of generation), health status 'migrates' or converges toward the national norm. Closely linked to this observation is what has become known as the *immigrant paradox* (Garcia Coll et al., 2012), where immigrants either show better adaptation outcomes than their national peers, or first-generation immigrants reporting better adaptation than their second-generation peers.

Relationship between How People Acculturate and How Well They Adapt

A number of studies have found that the acculturating strategy that people adopt is related to how well they adapt. The most common finding is that the integration strategy is the most adaptive in several settings and is associated with better psychological and sociocultural adaptation. In a meta-analytic study involving over 23 000 participants, Nguyen and Benet-Martínez (2013) found a strong and positive association between the integration strategy (referred to as biculturalism by the authors) and adjustment (both psychological and sociocultural). Analyses also indicated that the association between integration and adaptation was moderated by the way acculturation was assessed, the adaptation domain, and sample characteristics. Stronger integration–adaptation associations were found when acculturation was conceptualized as a bidimensional as opposed to a unidimensional scale. Marginalization has also been found to be the least adaptive outcome in a number of studies (Berry et al., 2006). One possible reason for why integration results in better adaptation outcome is that it entails a form of double competence and the availability of double resources. These competencies come from one's own ethnic and cultural group and from the new and larger society, and these resources double an individual's ability to cope with cultural transitions. In contrast, marginalization entails little competency in and lack of support from any cultural group; hence, the risks of adaptation difficulties are higher.

Ethnic and racial discrimination is one factor that has been found to be detrimental for positive acculturation outcome, and undermines the acculturation process itself. For instance, people who experience high ethnic discrimination are more likely to prefer separation, whereas those experiencing less discrimination prefer integration or assimilation. This phenomenon has been referred to as reciprocity in mutual attitudes, suggesting that if immigrants experience rejection (e.g., through discrimination) from the society of settlement, then they are more likely to reject (members of) the society in return. Similarly, there is evidence that discrimination is often the most powerful predictor of poor psychological and sociocultural adaptation (Berry et al., 2006). Taking this evidence all together, there appears to be a close (triangular) relationship among acculturation strategies, adaptive outcomes, and discrimination. While it is difficult to ascertain causality because of the use of cross-sectional data, using a structural equation modeling Berry et al. (2006) found a good fit to the assumption that discrimination may be a starting point; high discrimination predicts low preference for integration and poor adaptation, whereas integration predicts positive adaptation.

Conclusions

Worldwide migration and globalization together with increased intercultural relationships have catapulted acculturation into an active and an expanding research field in psychology and many social sciences. Although this article has focused largely on the psychology of acculturation, many of the issues discussed are relevant for the other social and behavioral sciences. As a growing field, many of the issues discussed are still in flux, but some patterns of findings are beginning to crystallize. Contrary to early assumptions, assimilation is not the only way, or the ultimate outcome of acculturation. Neither is it the most desired form of acculturation. Much research has suggested that integration is the most preferred way of acculturating, and also has the best long-term outcomes. Nevertheless, research is still needed to fully understand the antecedents to acculturation preferences, and how these antecedent factors interact with the cultural context in which the acculturation is taking place. To date, much of acculturation research is cross-sectional and quasi-experimental in nature, making it difficult to fully ascertain the impact of the multiple factors. The need for experimental and longitudinal studies is in dire need. Considering that acculturation is taking place across the entire world, research findings to date come from Western industrialized countries. Research from China, India, and Russia where large-scale intercultural contacts are taking place, and among sojourners and refugees settled in various regions of Africa, the Gulf States, and West Asia is very much needed to help develop a broader picture of this phenomenon. Finally, there is the need to expand the mutuality of influence from a simple dominant vs. nondominant group relations, but to several mutual influences. Many contemporary societies are made up of several nondominant ethnocultural groups together with a dominant group. As such, acculturating individuals may be influenced by several nondominant groups, and not only one's ethnic group and the dominant group of the society. Moreover, defining one's identity may be more than simply national and ethnic, and may also include religious identification.

See also: Cross-Cultural Psychology; Cultural Influences on Interpersonal Relationships; Culture and the Self: Implications for Psychological Theory; Ethnic Identity, Psychology of; Ethnicity and Migration in Europe; Identity and Identification, Social Psychology of; Immigration: Social Psychological Aspects; Intergroup Relations; Racism: Social Psychological Perspectives; Self-Categorization Theory; Social Identity in Social Psychology; Xenophobia: Social Psychological Aspects.

Bibliography

Azzi, A., Chryssochoou, X., Klandermans, K., Simon, B., 2011. Identity and Participation in Culturally Diverse Societies. Wiley-Blackwell, Oxford.

Benet-Martínez, V., Haritatos, J., 2005. Bicultural identity integration (BII): components and psychosocial antecedents. Journal of Personality 73, 1015–1050.

Berry, J.W., 1974. Psychological aspects of cultural pluralism. Culture Learning 2, 17–22.

Berry, J.W., 1990. Psychology of acculturation. In: Berman, J.J. (Ed.), Nebraska Symposium on Motivation, 1989: Cross-cultural Perspectives. University of Nebraska Press, Lincoln, pp. 201–234.

Berry, J.W., Phinney, J.S., Sam, D.L., Vedder, P., 2006. Immigrant Youth in Cultural Transitions: Acculturation, Identity, and Adaptation across National Contexts. Erlbaum, Mahwah, NJ.

Berry, J.W., Poortinga, Y.H., Breugelmasn, S.M., Chasiotis, A., Sam, D.L., 2011. Cross-cultral Psychology. Research and Application, third ed. Cambridge University Press, Cambridge.

Bourhis, R.Y., Moise, L.C., Perreault, S., Senecal, S., 1997. Toward an interactive acculturation model: a social psychological approach. International Journal of Psychology 32, 369–386.

Garcia Coll, C., Paton, F., Marks, A.K., Dimitrova, R., Yang, R., Suaraz, G.A., Patrico, A., 2012. Understanding the immigrant paradox in immigrant youth: developmental and contextual considerations. In: Mansten, A.S., Liebkind, K., Hernandez, D.J. (Eds.), Realizing the Potential of Immigrant Youth. Realizing the Potential of Immigrant Youth. Cambridge University Press, Cambridge, pp. 159–180.

Graves, T.D., 1967. Psychological acculturation in a tri-ethnic community. Southwestern Journal of Anthropology 23, 337–350.

Horenczyk, G., Jasinskaja-Lathi, I., Sam, D.L., Vedder, P., 2013. Mutuality in acculturation. Towards an integration. Zeitschrift fur Pschologie, Journal of Psychology 221, 205–213.

Koneru, V.K., Weisman de Mamania, A.G., Flynn, P.M., Betancourt, H., 2007. Acculturation and mental health: current findings and recommendations for future research. Applied and Preventive Psychology 12, 76–96.

Mansten, A.S., Liebkind, K., Hernandez, D.J. (Eds.), 2012. Realizing the Potential of Immigrant Youth. Cambridge University Press, Cambridge.

McGee, W.J., 1898. Piratical acculturation. American Anthropologist 11, 243–249.

Motti-Stefanidi, F., Berry, J.W., Chryssochoou, X., Sam, D.L., Phinney, J.S., 2012. Positive immigrant youth adaptation in context: developmental, acculturation and social psychological perspectives. In: Mansten, A.S., Liebkind, K., Hernandez, D.J. (Eds.), Realizing the Potential of Immigrant Youth. Cambridge University Press, Cambridge, pp. 117–158.

Nguyen, A.-M.T.D., Benet-Martínez, V., 2013. Biculturalism and adjustment: a meta-analysis. Journal of Cross-cultural Psychology 44, 122–159.

Phinney, J.S., 1990. Ethnic identity in adolescents and adults: a review of research. Psychological Bulletin 108, 499–514.

Powell, J.W., 1883. Human evolution: annual address of the President, J. W. Powell, delivered November 6, 1883. Transactions of the Anthropological Society of Washington 2, 176–208.

Redfield, R., Linton, R., Herskovits, M.J., 1936. Memorandum for the study of acculturation. American Anthropologist 38, 149–152.

Sam, D.L., Berry, J.W. (Eds.), 2006. The Cambridge Handbook of Acculturation Psychology. Cambridge University Press, Cambridge.

Sam, D.L., Berry, J.W., 2010. Acculturation: when individuals and groups of different cultural backgrounds meet. Perspectives on Psychological Science 5, 472–481.

Simons, S.E., 1901. Social assimilation. American Journal of Sociology 6, 790–822.

Snauwaert, B., Soenens, B., Vanbeselaere, N., Boen, F., 2003. When integration does not necessarily imply integration: different conceptualizations of acculturation orientations lead to different classifications. Journal of Cross-cultural Psychology 34, 231–239.

Verkyten, M., 2005. The Social Psychology of Ethnic Identity. Psychology Press, Hove.

Ward, C., 2001. The A, B, Cs of acculturation. In: Matsumoto, D. (Ed.), The Handbook of Culture and Psychology. Oxford University Press, Oxford, United Kingdom, pp. 411–445.

Wiley, S., Deaux, K., 2011. The bicultural identity performance of immigrants. In: Azzi, A.E., Chryssochoou, X., Klandermans, B., Simon, B. (Eds.), Identity and Participation in Culturally Diverse Societies. A Multidisciplinary Perspective. Wiley-Blackwell, Chichester-Oxford, pp. 49–68.

Young, C.B., Fang, D.Z., Zisook, S., 2010. Depression in Asian-American and Caucasian undergraduate students. Journal of Affective Disorders 125, 379–382.

Action, Theories of Social

Frank Kalter, University of Mannheim, Mannheim, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R. Boudon, volume 1, pp. 54–58, © 2001, Elsevier Ltd.

Abstract

Social action is a key concept in sociology, and the development of adequate theories that can explain it is a central task. The social sciences suggest very different models of individual behavior without agreeing on a best solution. To understand the controversies and to sketch fruitful pathways, it is helpful to examine the specific role of social action theories within the more general sociological enterprise in more detail and to derive central quality criteria. Against the background of these criteria, a review of influential, traditional 'models of man' reveals the relative advantages and disadvantages of each approach. Major questions are how to deal with this variety of models and how to progress on the front of social action theories.

Theories of Social Action and Their Role in Sociology

Max Weber clearly defined the concept of 'social action' in his classical work 'Economy and Society': "We shall speak of 'action' insofar as the acting individual attaches a subjective meaning to his behavior – be it overt or covert, omission or acquiescence. Action is 'social' insofar as its subjective meaning takes account of the behavior of others and is thereby oriented in its course" (Weber, 1968, p. 4). This definition is commonly shared by sociologists, as is the fact that the concept is of key importance to sociology. Max Weber even regarded it as the constituent subject of the discipline, which is expressed by the sentence preceding the quotation above: "Sociology (…) is a science concerning itself with the interpretive understanding of social action and thereby with a causal explanation of its course and consequences" (Weber, 1968, p. 4). 'Theories of social action,' accordingly, aim at this understanding and explanation and try to answer why certain (types of) actors chose specific alternatives of action.

Superficially, moving theories of social action into the center of sociological interest might seem somewhat questionable: The demarcation of sociology from other social sciences, for example psychology, is usually seen in the fact that it is basically interested in collective phenomena, rather than in the individual actor; in other words, the 'analytical primacy' for a sociologist lies at the macro level (i.e., the level of the society) (Wippler and Lindenberg, 1987). This view has been explicitly expressed by another prominent father of sociological thinking, Emile Durkheim. According to Durkheim, sociology is the science of 'social facts,' which he explicitly viewed as "existing outside the consciousness of the individual" (Durkheim, 1982, p. 51) and "not having the individual as their substratum" (Durkheim, 1982, p. 52).

When it comes to explaining social phenomena, however, the strategy of staying only at the macro level faces at least two severe problems: On the one hand, relations between social phenomena can hardly be regarded as 'sociological laws,' as they are simply not stable enough but are heavily dependent on time- and space-specific conditions (Hedström, 2005). On the other hand, even under well-specified limited conditions, regularities that are stated at the macro level lack a deeper understanding of the underlying mechanisms and leave a black box in any explanatory storyline (Hedström, 2005; Wippler and Lindenberg, 1987). It is here where theories of social action come in. Going down to the micro level and looking at the behavior of individuals can provide the necessary depth and possibly fill this black box. Regularities in social action seem to be more stable and general than those at the macro level, so social action theories can possibly build the 'nomological core' of a sociological explanation. In other words, while the analytical primacy of sociology is at the macro level of the society, there are good reasons to see the 'theoretical primacy' at the micro level of the individual; this is the general position of 'methodological individualism' to which the concept of social action and the striving for social action theories in sociology are intrinsically tied.

These ideas are usually illustrated with the help of the so-called macro–micro–macro scheme (Coleman, 1990; Wippler and Lindenberg, 1987; Hedström and Swedberg, 1998), which allows one to locate the role of social action theories in sociology more closely. Accordingly, any sociological explanation entails three analytically distinct steps. (1) It has to specify how macro-level conditions affect the situation of (typical) individual actors. This macro–micro step is often called the 'logic of the situation,' and the assertions and hypotheses developed in this step are termed 'bridge assumptions.' (2) It has to explain why (typical) individual actors select specific individual actions. This step on the micro level, the 'logic of selection,' is the realm of social action theories. (3) It has to clarify how the actions of typical individual actors interact and give rise to the collective phenomena of interest. This micro–macro step is called the 'logic of transformation' or 'logic of aggregation.'

The reference to the macro–micro–macro model is not only helpful in understanding the specific role of social action theories as one vital step within the more general sociological enterprise, but also crucial for identifying and understanding the more detailed requirements that a social action theory should fulfill in order to be appropriate for sociology (Lindenberg, 1985). Most importantly, it must be able to connect smoothly to the other two steps: the logic of the situation and the logic of transformation. This means, among

others, that the action theory should conveniently allow formulation of bridge hypotheses that link structural conditions to the parameters of actions, without requiring too much knowledge about the individual. The fact that the analytical primacy is on the macro level sets somewhat other priorities for a social action theory appropriate for sociology compared to action theories in a field like psychology, where the analytical primacy is on the micro level (Wippler and Lindenberg, 1987). Nevertheless, social action theories in sociology should certainly not explicitly contradict well-known findings in neighboring disciplines. Thus, in a sense, they can and should be more abstract, but allow one to decrease the level of abstraction if necessary (Lindenberg, 1992). Not least, like all theories, theories of social action should aim for high precision (i.e., information content) and should be well corroborated (Popper, 1959).

Traditional Views on Social Action

In the history of sociological thought, very different views on social action and underlying 'models of man' have been suggested and discussed. Among the most influential are the homo sociologicus, which emerged from Parsons' voluntaristic theory of action, the interpretative paradigm, and the homo economicus of neoclassical economics.

Voluntaristic Theory of Action and Homo Sociologicus

A major milestone in the development of sociological theories of social action is Talcott Parsons' *The Structure of Social Action* (1937), in which he develops his 'voluntaristic theory of action.' This can be understood as an attempt to synthesize ideas on human behavior to be found in the work of various classical sociologists, among them Weber and Durkheim, but also in that of classical economists. The basic elements of Parsons' framework are actors, goals, means, situational conditions, and normative orientations (norms, values, and ideas). Social action should, accordingly, in general be understood as individual actors making decisions about means to achieve goals, whereby the goals and the availability of alternative means are influenced, on the one hand, by situational constraints and, on the other, by values, norms, and ideas that are shaped by processes of internalization and institutionalization.

Accounting for situational conditions and, most importantly, for normative orientations is Parsons' contribution to the logic of the situation. It pays tribute to the legacy of Durkheim and marries his idea of social determination to the methodological individualism of Weber. The fact that, thus, social structure is explicitly integrated into a general model action is certainly a major comparative strength of the 'voluntaristic theory of action.' But there is also a major weakness, which is the lack of information content with respect to the logic of selection. Parsons' focus is clearly on synthesis and conceptualization; his framework does not provide, however, a clear rule for choosing a precise action alternative. The mere reference to values and norms is not sufficient as they tend not to represent coherent systems that give unambiguous advice.

The latter problem is especially visible in the general idea of homo sociologicus, which emerged out of Parsons' work (Dahrendorf, 1968). This model of man underlying structural functionalism assumes that individuals basically act according to internalized expectations that are embedded in systems of social roles. Role conflicts and deviant behavior are only the most obvious difficulties of this view on behavior. When referring to the macro–micro–macro scheme, the more fundamental problem of the homo sociologicus idea is that, in a sense, it tries to relate social structure directly to the actions, neglecting the role of the active and resourceful actor. As a consequence, it seems adequate only under very special and stable conditions (Lindenberg, 1985), and thus it ends up with problems similar to those of a purely macro-theoretical account.

The Interpretative Paradigm

The voluntaristic theory of action has stimulated intensive sociological attention to the problem of finding adequate theories of social action, and homo sociologicus as suggested by structural functionalism has provoked much critique. A very important line of argument has led to the 'interpretative paradigm' (Wilson, 1970), which is intended to be a corrective against the so-called normative paradigm underlying the work of Parsons and Durkheim. The interpretative paradigm can be traced to rather diverse roots, such as European phenomenology (Husserl, 1965; Schütz, 1967) or American pragmatism and behaviorism (Dewey, 1922; Mead, 1934).

The theoretical perspective has been worked out, especially within symbolic interactionism by Herbert Blumer (1969); it targets the passive role of actors in the normative paradigm, in which norms and values seem to treat individuals like puppies on a leash. This general critique crystallizes into two major points. First, homo sociologicus is seen as, at least, severely incomplete, as norms and values only rarely give clear advice for specific actions. Second, and even more importantly, norms and values cannot be understood as externally given by social structure, but they should rather be understood as an endogenous result of the interactions between individual actors.

According to symbolic interactionism, the way that individuals act toward things is fundamentally dependent on the meaning they attribute to them. This meaning is by no means objectively given but must be understood as a subjective accomplishment of the individual actor, an idea that is tellingly described by the famous 'Thomas theorem': "If men define situations as real, they are real in their consequences" (Thomas and Thomas, 1928, p. 572). The meaning results, above all, from permanent social interactions and communications in which symbols and their interpretation play crucial roles. Accordingly, meanings are constantly changing in response to the environment, and interpretation processes are heavily dependent on situational clues.

Symbolic interactionism thus brings the actor back in and emphasizes his or her very active role via processes of interaction and interpretation. In doing so, it contributes specifically to a more adequate and realistic understanding of the definition of the situation. In this attempt, however, the analytical primacy shifts away from the macro level, and correspondingly it is difficult within this approach to develop bridge hypotheses

that link individual interpretations back to social structure; it requires quite a lot of idiosyncratic information about the individual. Moreover, the approach is particularly weak (even weaker than the normative paradigm) with respect to the logic selection: It does not provide clear and information-rich rules about how actors finally choose among different action alternatives.

Neoclassical Economics and Homo Economicus

The theory of behavior used by neoclassical economics has always been a kind of baseline and contrast model against which the sociological search for appropriate social action theories has developed. Basically, homo economicus is assumed to calculate the costs and benefits of available action alternatives and to choose the alternative that best satisfies his or her interests. The underlying core assumptions are thus that social action (1) aims at the realization of preferences, (2) underlies opportunities and restrictions, and (3) is guided by the principle of maximization. These assumptions are seen to constitute the core of so-called rational choice theory (RCT) (Opp, 1999), which comes in a series of more differentiated variants (Goldthorpe, 1998).

Given the inbuilt weaknesses of the genuine sociological approaches sketched here, the economic approach to behavior has also received much attention in sociology. And it has been applied quite successfully: For example, economist Gary Becker received the Nobel Prize in 1992 for proving the microeconomic approach to be extremely fruitful in core sociological fields like discrimination, crime, and the family (Becker, 1976). No less important and impressive, James Coleman demonstrated in his *Foundations of Social Theory* (1990) that RCT can lay the groundwork for a comprehensive treatment of sociology. In general, RCT is one of the most influential theoretical approaches guiding current empirical research in sociology (Kroneberg and Kalter, 2012).

Considering the general criteria for social action theories, RCT provides many comparative advantages: Most importantly, the maximization principle delivers an extremely precise rule of selection, thus overcoming one of the major weaknesses of the normative and interpretative paradigm. It conveniently allows the specification of bridge hypotheses linking action back to social structures, as opportunities and restrictions are crucial parameters of the actor's decision-making process. Viewing actors not only as consumers but also as producers even allows fruitful bridge hypotheses on preferences (Stigler and Becker, 1977). All these aspects foster the main purpose of any explanatory framework, which is to derive precise and testable hypotheses. And all this explains the above-mentioned success of RCT, especially in applied empirical research.

This has led many scholars to the view that homo economicus would in the end best satisfy the needs of explanatory sociology and that the search for alternative models of man might be superfluous and misleading; the term 'economic imperialism' has even been used. The diffusion of RC theorizing in sociology has been likewise accompanied, however, by a growing skepticism: Empirical research has yielded more and more findings that question the core assumptions of RCT (Elster, 1989). Much evidence along these lines stems from experimental research in social psychology, behavioral economics, and behavioral game theory, suggesting that actors are often by no means egoistic maximizers but that behavior is heavily dependent on situational clues, cultural frames, fairness norm, and so on. The evidence that actors thus only show a 'bounded rationality' (Simon, 1957) is overwhelming. Similar doubts on the rational actor model arose from applied research in sociology, with one example being that individuals contribute to the provision of public goods in spite of the incentives of free riding (Opp, 1986). Thus, homo economicus fails to satisfy the basic criterion of delivering an action principle that could claim general validity, and it explicitly contradicts empirical findings in sociology and its neighboring disciplines.

Strategies and Perspectives

The 'Traditional Views' section has shown that the different traditional models of man exhibit different advantages and disadvantages with respect to criteria for an adequate theory of social action. They also emphasize different aspects of human behavior, all of which seem to be indispensable parts of a comprehensive sociological account. This leads back to Max Weber, who in his famous 'Economy and Society' also made a well-known distinction between four main types of social action: action based on instrumentally rational action, value-rational action, affectual action, and traditional action (Weber, 1922, pp. 24–25).

To deal with the different strengths and weaknesses of the different approaches to social action and with the seemingly different types of social action that can be observed in social life, two general strategies can be distinguished: Proponents of a 'pluralistic strategy' claim that sociology simply has to live and work with different models of actors, to keep all of them in a toolbox in peaceful coexistence, and to employ each for different tasks (Schimank, 2000). This, however, remains unsatisfactory for explanatory purposes as long as there is no clear overarching meta-rule specifying exactly under what conditions and for what kind of concrete behavior the one or the other model might be more appropriate; it does not free each of its inbuilt methodological weaknesses, either. Proponents of an 'integrative strategy' therefore seek to work out more refined theories of social action that combine the insights from the different approaches within one overarching model, so that meta-rules are clearly specified and so that the strength of one might heal the weaknesses of the others.

An important development along the integrative line has been the refinement of the narrow model of homo economicus underlying neoclassical economics into 'wide versions of RCT' (Opp, 1999). In particular, these wide versions relax the assumptions that actors are perfectly informed, are purely egoistic, regard only tangible consequences, and react only to objective characteristics of the situation. By allowing, for example, that actors might receive intrinsic utility from fairness, altruism, and compliance to norms, or that they might fear the costs of sanctions, it is possible to marry RCT to many aspects of homo sociologicus. By emphasizing that actors' beliefs are subjective, and can be based on imperfect or even systematically biased information, wide versions of RCT can refute some of the basic objections from the

interpretative paradigm and explicitly account for the importance of processes of interpretation in the logic of a situation. At the same time, however, they keep the main advantage of RCT, which is the clear selection rule. The development from expected value (EV) theory, over expected utility (EU) theory, into the theory of subjective expected utility (SEU) is a major lane in the process of widening (Schoemaker, 1982), and SEU theory has become especially popular in empirical sociological research.

However, the integrative merits and the huge flexibility of wide versions of RCT come at a cost: Releasing the strict assumptions of neoclassical economics leaves the core of the theory almost empty; if everything can in principle be conceptualized as costs and benefits, the theory seems to come close to being a tautology. While this allegation is exaggerated (Opp, 1999), the least one can say, however, is that wide versions of RCT move the explanatory burden from the action theory to the bridge assumptions (i.e., to the auxiliary assumptions made in the step of the logic of the situation, creating new methodological challenges) (Esser, 1998). Consequently, there is widespread skepticism toward 'softening' the RCT approach (Mood, 2009), and there are pleas to stick to narrower versions (if ever, and for as long as possible) (Raub et al., 2011), sometimes leading to an instrumentalist tendency in practice (Hedström, 2005). The 'method of decreasing abstraction' (Lindenberg, 1992) provides a potential recipe in the context of these problems, by emphasizing that it is adequate to work with simplifying assumptions as long as these do not contradict more complex insights and that it must be possible, in principle, to replace the former with the latter if need be.

There are important arguments, however, that even allowing for a much wider range of preferences, expectations, and constraints within RCT might not be enough to enable the necessary theoretical integration, but that a more fundamental revision, touching also – partly at least – its core of the maximization principle, is needed. Accordingly, attempts explicitly try to account for the fact that rationality in actions is conditional and variable. An early example is the contribution by political scientists William H. Riker and Peter C. Ordeshook (1973) to model aspects of bounded rationality in a two-step model, where in the first step actors confronted with information costs decide whether to make a decision at all, thus allowing routines (i.e., 'traditional action' in terms of Max Weber) to take over in many circumstances. Influential two-steps models, in which rational processing is only one possible option, were also suggested in social psychology (Fazio, 1990).

Sociologists working on the front of social action theories have especially tried to bring classical insights about the definition of the situation explicitly and more accurately back into behavioral models. Raymond Boudon (1996) has suggested a 'cognitivist model' that, among others, gives Weber's notion of value-rationality a new fruitful interpretation. Accordingly, actors are often motivated by normative beliefs rather than instrumental incentives, and they are seen to hold these beliefs, because they have good reasons to do so. According to Boudon, this cognitivistic model of action avoids the shortcomings of RCT by providing a more realistic account, at the same time saving RCT's advantage of providing a selection rule within an explanatory framework (Boudon, 1998).

Siegwart Lindenberg (2008) has developed a theory of 'goal-framing' that pays specific attention to the role of situational clues within actions. The basic idea is that actors, rather than considering a wide range of preferences and weighting the corresponding consequences of action (as assumed in traditional RCT), are rather limited and usually regard only one goal that is dominant in a given situation. This goal is in the foreground and mainly frames the definition of the situation (i.e., the attentions, perceptions, and consequently the behavior of actors). Other goals stay in the background, but they might be activated by situational clues and challenge the salience of the dominant 'goal-frame.' Lindenberg identifies some general goals that are shared by all actors, such as the hedonic goal, which is the default in many situations; the gain goal; and the normative goal.

Frames are also the key concept in the 'model of frame selection' that has been brought forward by Hartmut Esser (2009). The model distinguishes several subprocesses in action: 'Mode-selection' determines, as a kind of meta-selection, whether actors behave automatically–spontaneously or reflexively–calculatingly. 'Frame-selection' is crucial to how actors define the situation. 'Script-selection' provides actors with repertoires of action, routines, and programs. 'Action-selection' is the last step. A comparative strength of the model is that it specifies very explicit rules about each of these subprocesses and about how they are interrelated (Kroneberg et al., 2010), thus trying to retain the explanatory power of RCT even while modifying some of its core assumptions. This precision, on the other hand, has made it vulnerable to critique.

In general, these – and other – recent attempts to integrate different lines of sociological thinking about social action are confronted with much skepticism, even more so than against the wide versions of RCT. Many regard these models as much too complex, and especially the concept of frames is suspected to bring black boxes back into explanations (Boudon, 2003). As the analytical primacy of sociology is collective phenomena, the key to the explanation might often lie in the complexity of the aggregation (Hedström, 2005); so one can afford, or is even forced, to work with much more parsimonious models of individual behavior. Proponents of the more complex theories of action would certainly agree, but they would argue that, if necessary, a decrease in abstraction (Lindenberg, 1992) must be possible, and a social action theory must specify in a systematic way how simplifying assumptions, if necessary, can be relaxed.

It is important to recall that whichever kind of simplification or abstractness in the theories of social action is appropriate will always be dependent on the specific explanatory task under consideration.

See also: Coleman, James Samuel (1926–95); Durkheim, Emile (1858–1917); Interactionism, Symbolic; Macrosociology-Microsociology; Methodological Individualism in Sociology; Norms; Parsons, Talcott (1902–79); Popper, Karl Raimund (1902–94); Rational Choice Theory in Sociology; Weber, Max (1864–1920).

Bibliography

Becker, G.S., 1976. The Economic Approach to Human Behavior. University of Chicago Press, Chicago.

Blumer, H., 1969. Symbolic Interactionism: Perspective and Method. Prentice-Hall, Englewood Cliffs, NJ.

Boudon, R., 1996. The 'cognitivist model': a generalized 'rational-choice model'. Rationality and Society 8, 123–150.

Boudon, R., 1998. Social mechanisms without black boxes. In: Hedström, P., Swedberg, R. (Eds.), Social Mechanisms. Cambridge University Press, Cambridge, pp. 172–203.

Boudon, R., 2003. Beyond rational choice theory. Annual Review of Sociology 29, 1–21.

Coleman, J.S., 1990. Foundations of Social Theory. Harvard University Press, Cambridge.

Dahrendorf, R., 1968. Homo sociologicus. In: Dahrendorf, R. (Ed.), Essays in the Theory of Society. Stanford University Press, Stanford, CA, pp. 19–87.

Dewey, J., 1922. Human Nature and Human Conduct. Henry Holt, New York.

Durkheim, E., 1982 [1895] (W.D. Halls, Trans.). In: Lukes, S. (Ed.), The Rules of Sociological Method. The Free Press, New York.

Elster, J., 1989. Nuts and Bolts for the Social Sciences. Cambridge University Press, Cambridge.

Esser, H., 1998. Why are bridge hypotheses necessary? In: Blossfeld, H.-P., Prein, G. (Eds.), Rational Choice Theory and Large-Scale Data Analysis. Westview, Boulder, CO, pp. 94–111.

Esser, H., 2009. Rationality and commitment: the model of frame selection and the explanation of normative action. In: Cherkaoui, M., Hamilton, P. (Eds.), Raymond Boudon: A Life in Sociology, Part Two: Toward a General Theory of Rationality. Bardwell, Oxford, pp. 207–230.

Fazio, R.H., 1990. Multiple processes by which attitudes guide behaviour: the MODE model as an integrative framework. Advances in Experimental Social Psychology 23, 75–109.

Goldthorpe, J., 1998. Rational action theory for sociology. British Journal of Sociology 49, 167–192.

Hedström, P., 2005. Dissecting the Social. On the Principles of Analytical Sociology. Cambridge University Press, Cambridge.

Hedström, P., Swedberg, R., 1998. Social mechanisms. In: Hedström, P., Swedberg, R. (Eds.), Social Mechanisms. Cambridge University Press, Cambridge, pp. 1–31.

Husserl, E., 1965 [1936]. Phenomenology and the Crisis of Western Philosophy. Harper & Row, New York.

Kroneberg, C., Kalter, F., 2012. Rational choice theory and empirical research: methodological and theoretical contributions in Europe. Annual Review of Sociology 38, 73–92.

Kroneberg, C., Yaish, M., Stocké, V., 2010. Norms and rationality in electoral participation and in the rescue of Jews in WWII: an application of the model of frame selection. Rationality and Society 22, 3–36.

Lindenberg, S., 1985. An assessment of the new political economy: its potential for the social sciences and for sociology in particular. Sociological Theory 3, 99–114.

Lindenberg, S., 1992. The method of decreasing abstraction. In: Coleman, J.S., Fararo, T.J. (Eds.), Rational Choice Theory: Advocacy and Critique. Sage, Newbury Park, pp. 3–20.

Lindenberg, S., 2008. Social rationality, semi-modularity and goal-framing: what is it all about? Analyse & Kritik 20, 669–687.

Mead, G.H., 1934. Mind, Self, and Society. University of Chicago Press, Chicago.

Mood, C., 2009. Problems without ends: how rational choice theory escapes its explanatory task. In: Cherkaoui, M., Hamilton, P. (Eds.), Raymond Boudon: A Life in Sociology, Part Two: Toward a General Theory of Rationality. Bardwell, Oxford, pp. 271–287.

Opp, K.D., 1986. Soft incentives and collective action: participation in the anti-nuclear movement. British Journal of Political Science 16, 87–112.

Opp, K.D., 1999. Contending conceptions of the theory of rational action. Journal of Theoretical Politics 11, 171–202.

Parsons, T., 1937. The Structure of Social Action. McGraw-Hill, New York.

Popper, K.R., 1959. The Logic of Scientific Discovery. Hutchinson,, London.

Raub, W., Buskens, V., van Assen, M.A.L.M., 2011. Micro–macro links and micro-foundations in sociology. Journal of Mathematical Sociology 35, 1–25.

Schimank, U., 2000. Handeln und Strukturen. Einführung in die akteurstheoretische Soziologie. Juventa, Weinheim.

Schoemaker, P.J.H., 1982. The expected utility model: its variants, purposes, evidence and limitations. Journal of Economic Literature 20, 529–563.

Schütz, A., 1967 [1932]. The Phenomenology of the Social World. Northwestern University Press, Evanston, IL.

Simon, H., 1957. Models of Man. Wiley, New York and London.

Stigler, G.S., Becker, G.S., 1977. De gustibus non est disputandum. American Economic Review 67, 76–90.

Thomas, W.I., Thomas, D.S., 1928. The Child in America. Behavior Problems and Programs. Alfred A. Knopf, New York.

Weber, M., 1968 [1922]. Economy and Society. Bedminster Press, New York.

William H. Riker, & Peter C. Ordeshook., 1973. An introduction to positive political theory. Prentice-Hall, Englewood Cliffs, NJ.

Wilson, T.P., 1970. Normative and interpretive paradigms in sociology. In: Douglas, J.D. (Ed.), Understanding Everyday Life. Toward the Reconstruction of Sociological Knowledge. Aldine, Chicago, pp. 57–79.

Wippler, R., Lindenberg, S., 1987. Collective phenomena and rational choice. In: Alexander, J.C., et al. (Eds.), The Micro–Macro Link. University of California Press, Berkeley, pp. 135–152.

Actor-Network Theory

Fabian Muniesa, Mines ParisTech, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article summarizes some important aspects of actor–network theory. It provides relevant elements of the historical origins of this line of inquiry, examines some of its most distinctive concepts, and signals its development in science and technology studies and in some other areas of the social sciences. The article emphasizes the position of actor–network theory within the poststructuralist landscape, discusses its particular treatment of constructivism, and examines in particular the philosophical implications of the notion of translation, which plays a crucial role in the articulation of this approach.

The social-scientific line of inquiry known as actor–network theory (ANT) is characterized by a distinctively materialist, radically constructivist approach to social theory and to empirical research. It is often recognizable through the use of a collection of notions (such as 'translation' and 'actor–network') that were developed in the 1970s and 1980s in response to developments in the sociology of science. An influential current in science and technology studies, ANT has also gained prominence in a number of different fields in the social sciences and the humanities, from organization studies to political science, from anthropology to economic sociology. In what follows, the historical context of the formation of ANT is first discussed, a series of central concepts are then clarified, and a philosophical assessment is finally suggested (for other relevant accounts see Law, 2009; Mol, 2010).

Intellectual Foundations and Historical Origins

The expression 'actor–network theory' was developed in the late 1980s in order to refer to an emergent school of thought that started to form in Paris, France, a decade earlier. An early mention to an 'actor–network approach' is found in the introduction to an edited volume on the sociology of technology (Bijker et al., 1987). The editors used that expression as a tag to identify and differentiate the approach defended by Michel Callon, who contributed to the volume with a seminal ANT case study on the attempted development of an electric vehicle in France (Callon, 1987). Callon had already proposed a different expression, 'sociology of translation' (*sociologie de la traduction* in French), in order to specify the singularity of this contribution, most notably in another seminal ANT case study on scallop farming along the coast of Brittany (Callon, 1986; see also Callon, 1980). But, in a somewhat path-dependent way, it was the notion of 'actor–network theory' that prevailed in Anglophone academic circles. The intellectual foundations and historical origins of ANT are located in the concrete institutional site of its inception, in the emergence of French poststructuralism and in debates in the sociology of science.

Engineering and Innovation

ANT's origins are to be found in a very concrete location. The approach was crafted at the Centre de Sociologie de l'Innovation (CSI), a small research center at the École des Mines de Paris, an engineering school that is both an elite higher education institution (a so-called *grande école*) and a research institution distinctively oriented toward industrial innovation. The CSI, created in 1967 with the purpose of studying the social determinants of innovation processes, had progressively shifted to a more catholic interest on the interrelations between science and society. Michel Callon, an engineer and sociologist by training and École des Mines alumnus whose earlier research focused on the interfaces between science and industry, was appointed director of the CSI in 1982. The same year, the CSI recruited Bruno Latour, a philosopher by training who was already known for his pioneering ethnography of a neuroendocrinology laboratory (Latour and Woolgar, 1979), a study that prefigured the most salient tenets of ANT's blend of semiotics. The CSI soon became the crucible in which ANT started to take shape. Slightly younger researchers such as Madeleine Akrich and Antoine Hennion, also École des Mines alumni, and visiting scholars such as John Law, then at the University of Keele, participated extensively in the inception of that movement (Susan Leigh Star and Geoffrey Bowker, then at the University of California, were also among the scholars who visited the CSI in the 1980s). Editorial efforts such as *Pandore*, a bulletin that greatly contributed to the diffusion of science and technology studies in France, and exchanges with sociologists and historians of scientific knowledge, notably based in the United Kingdom and the Netherlands, contributed to the shaping of the early orientation of ANT toward the sociology of science.

This localization of the founding of ANT is not accidental. In a sense, the special circumstances that were prompted at the CSI gave ANT some notable characteristics. First, the CSI's intellectual familiarity and daily contact with engineering (in particular in relation to electric systems and extractive industries) translated into a distinctively industrial and technological understanding of constructivism. For ANT, reality is constructed, but it is constructed in the engineer's sense (solid reality as the outcome of an organized, fragile, and laborious process of material articulation) rather than in the sense usually put forward in standard social sciences (social construction considered in terms of social conventions, belief systems, mental states, or collective representations). Second, the CSI's marginal position in the field of instituted social sciences in France and its relative lack of disciplinary control translated into an interdisciplinary (or rather

adisciplinary) attitude. Consequently, ANT questions the boundaries between disciplines and, more profoundly, between the social and the natural sciences, favoring approaches that are able to cope with heterogeneous realities. Third, the tradition of pursuing contract research and developing applied expertise at the CSI, and the combination of highly theoretical developments with extremely practical research reports, translated into a novel preference for hybrid styles. ANT indeed portrays the researcher in general, and the social scientist in particular, as an entangled participant rather than as an external observer viewing reality from above.

Parisian Poststructuralism

The formation of ANT is, in part, an episode in the history of French poststructuralism and a consequence of the intellectual atmosphere prompted in Parisian academic circles by authors such as Michel Serres (see Dosse, 1999; Serres and Latour, 1995). The constructivist approach to science defended by Serres in the 1970s and his distinctive use of the notion of translation played a pivotal role in the intellectual education of early contributors to ANT. The philosophical take on a materialist understanding of signification defended by Gilles Deleuze and the decisive impact that his collaboration with Félix Guattari had on a number of French intellectuals in the 1970s and 1980s was also at work in the emergent repertoire of ANT, especially with their notions of rhizome and collective assemblage of enunciation. The approach to semiotics constructed by Algirdas Julien Greimas (his actantial model in particular) was crucial in the development of ANT's interpretation of the operations of signs and texts. Other notable influences on the formation of ANT can be located in the works of authors such as Michel Foucault, Gilbert Simondon, François Dagognet, André Leroi-Gourhan, and Michel de Certeau, who were widely read at the CSI in the 1980s.

ANT's early development was definitely closer, intellectually speaking, to the philosophical preoccupations broadly associated with poststructuralism, than to other concerns that also characterized French academic life during that period, such as the novel developments in sociology prompted, for example, by Pierre Bourdieu, Alain Touraine, Michel Crozier, or Raymond Boudon, or the issues on complexity and chaos raised in the natural sciences and in mathematics by authors such as Henri Atlan, René Thom, or Ilya Prigogine. The philosophical penchant of ANT was deeply marked by Bruno Latour's own trajectory and vision. The distinctive touch of ANT lay in a taste for a hybrid disciplinary positioning (definitely not as standard social science), an obsession with the materiality of signification (hence the call for a semiotic approach to technology), and, in a sense, a certain freedom to engage in intellectual experimentation (with no imposed canon). But ANT definitely partook in a general poststructuralist movement that was seeking alternatives to the intellectual machineries defended at that time by influential figures such as Louis Althusser, Jacques Lacan, and Claude Lévi-Strauss.

Constructivism in the Sociology of Science

The formation of ANT is, also in part, an episode in the history of science and technology studies, one mainly characterized by a materialist approach to agency and a constructivist understanding of truth (see Pickering, 1992, 1995). ANT is often presented as a reaction to (and a dialogue with) two threads in the understanding of scientific inquiry: the French tradition of the epistemology of science and the British tradition of the sociology of scientific knowledge. The idea that Nature does not speak for itself and that scientific facts are constructed is almost unanimously acknowledged in the French tradition in the epistemology of science exemplified by authors such as Alexandre Koyré, for whom scientific discovery is determined by structures of thought that are historically constituted. Whether grounded in an intellectual frame of reference or a rational method (in line with Gaston Bachelard or Karl Popper) or in the rules of a social institution or discursive habit (whether in Michel Foucault or Ludwig Wittgenstein's terms), these structures of scientific thought are definitely historically established. Thomas Kuhn's *tour de force* (subsuming all this under the notion of paradigm) bolstered and comforted this viewpoint. But what is ruled out from this perspective, ANT would claim, is the relevance of any empirical account of the concrete operations (experimenting, measuring, calculating, writing, communicating) that are needed in order to obtain scientific facts – precisely the type of inquiry envisaged by ANT. The intellectual movement that characterized the British tradition of the sociology of scientific knowledge was quite different, almost the opposite of ANT's direction of inquiry. Authors such as Barry Barnes and David Bloor (the 'Edinburgh School' in the sociology of scientific knowledge) or Harry Collins and Trevor Pinch (the 'Bath School') were working out a sociological response to a rather empiricist, inductive British tradition in the epistemology of science, quite different from the French tradition. Accordingly, their response emphasized the role of social interests and collective representations in the establishment of scientific truth and falsehood. Seen from this angle, ANT's emphasis on the study of scientific controversies (situations in which scientific truth and falsehood are still in the making) was appreciated, but its focus on the empirical operations of science and on the material conditions of scientific work appeared as a fatal step backward toward scientific realism and positivism. As exemplified in numerous debates (for exemplary studies, position statements and debates, see Knorr-Cetina and Mulkay, 1983; Bijker et al., 1987; Law, 1986; Pickering, 1992), ANT was trying to be constructivist and realist at the same time, by accepting that scientific truth and scientific reality do exist, but not without a painstakingly material process of instauration (Latour, 1987, 1999a).

The elaboration of ANT's constructivist realism found some hospitable terrain in other areas of science and technology studies. Harold Garfinkel's studies in ethnomethodology, as applied in particular to the understanding of laboratory work, shared to some extent with ANT an interest on the ongoing accomplishment of reality, be it social or otherwise. Studies in situated action and distributed cognition, as illustrated by the work of Lucy Suchman or Edwin Hutchins, developed an understanding of agency that was very much in line with ANT. Historians of science who, like Peter Galison, Lorraine Daston, and others, have highlighted the material cultures of scientific objectivity have contributed greatly to arguments that have been actively endorsed by ANT. The development of ANT from the 1980s onward trespassed, however, the perimeters of

science and technology studies. ANT or ANT-oriented perspectives are today present in a multiplicity of areas (organization studies, political science, anthropology, economic sociology, to name a few), although not always straightforwardly linked to issues of science and technology.

Central Problems

ANT has made explicit and attempted at treating a number of problems of social-scientific inquiry, sometimes creating adherence and sometimes sparking controversy. This has usually translated into the production and use of a number of recognizable concepts, precepts, standpoints, and representative case studies on the problem of agency, on the notion of translation and on the constitution of collective realities.

Generalized Agency and the Critique of Modern Reason

ANT is often associated in popular views with an insistence on 'nonhuman agency,' that is, on sources and agencies of action other than purely human, conscious, and intentional. In fact, ANT stands as a reaction to both the downplaying of human agency in accounts of events favored in the natural and formal sciences (an ellipsis of the action of the experimenter in a microbiology laboratory, for example, in reports of findings) and the downplaying of nonhuman agency in accounts of events favored in the social sciences and the humanities (an ellipsis of the actions of bacteria, the medium, and the laboratory instrument). The point is twofold. First, the events that need to be accounted for are the conjunction of all kinds of agencies, the variety of which cannot be subsumed under a simple human/nonhuman or intentional/unintentional divide. For example, in the microbiology laboratory, action is prompted by personal purposes, unconscious motives, physical bodies, social institutions, corporate actors, living organisms, technological devices, gravitational forces, atmospheric conditions, budgetary constraints, collective ideologies, measurement instruments, legal codes, etc. In short, as soon as something happens, there is action to be accounted for, and a good ANT account does not single out any particular form of action, be it social or otherwise. Second, the fact that accounts are usually biased in favor of one form or other of 'purified' agency, is due, ANT claims, to the legacy of modern reason, that is, to a style of thought that is most distinctively characterized by the instauration of a neat divide between nature and culture, or between the physical condition of the world and its political constitution. This is both a bias that needs to be overcome in order to account for the proliferation of hybrid realities that modernity itself prompted (hence the need to develop an appropriate vocabulary) and an object of anthropological inquiry as such, as posited in Bruno Latour's 'anthropology of modern reason' (Latour, 1993, 2013).

On the question of the nature of human agency alone, ANT is known for contributing to a double refutation: that of the modern, liberal ideal of the sovereign, emancipated individual and that of the equally modern, sociological ideal of the dominated, conditioned social agent. The idea of using actors' interests as ultimate explanatory factors was replaced, in early contributions, by the idea of studying processes of 'interessement' and elaborating how they come about (Callon and Law, 1982; Callon, 1986). The pivotal notion of 'attachment' and the notable case studies on drug addiction and music lovers that accompanied its development (Gomart and Hennion, 1999; Latour, 1999b; Hennion, 2005, 2007) allowed ANT to shift attention from the 'doing' to the 'making do,' and from the 'moving' to the 'being moved,' as a constitutive feature of humanity. Through such movement, ANT arrives at a rather anthropological and philosophical reconsideration of what a 'thing' is and of what it does (e.g., Latour, 2000, 2004a, 2007). On the vocabulary of agency in standard social theory, ANT reportedly pointed, from its inception, to the recurrent use of 'nonhuman' parlance in sociology when dealing with the agency of large actors such as 'society,' 'the economy,' 'the corporation,' 'the nation,' or 'the state,' actors whose agency is in part structured with the collaboration of social scientists (Callon and Latour, 1981). The purpose of ANT is not to deny such agency but, precisely, to describe how it is mounted.

Material Semiotics, Inscriptions, Translation, and Networks

The task of developing a 'convenient vocabulary' adapted to this view of generalized agency was a *mot d'ordre* in the early years of ANT (see Akrich and Latour, 1992). The crux of this vocabulary lay in a combination of two crucial ingredients: a semiotic treatment of agency (mainly based on Greimassian semiotics and the notion of 'actant') and a focus on material devices (inspired by the anthropology of technology and by the engineering of sociotechnical systems). The analysis of how a program of action or a script of conduct is inscribed or not into a material artifact, and of how it interacts with other programs or other scripts, stood as one key ingredient of the ANT recipe (Akrich, 1992; Latour, 1992). The semiotic, actantial analysis of scientific publications was accompanied by quite novel developments, at the time, that traced materially the networks formed by texts, for example, through coword analysis and bibliometric maps (Callon et al., 1986). Inventing new social-scientific literary styles altogether was also part of ANT's original endeavor (Latour, 1996).

ANT's particular blend of material semiotics is intrinsically pragmatist. Instead of considering reference or signification from the viewpoint of a binary partition between signs and things (or between mind and matter), ANT considers reference as a materially traceable operation of displacement, transportation, and circulation. As demonstrated by Latour in his minute examination of the scientific power of Louis Pasteur, the ability to refer to the anthrax bacillus within the laboratory setting; the ensuing ability to refer, in entirely new terms, to a disease affecting cattle; and the ability to transform an entire country altogether (its cattle, its farmers, its inhabitants, its bacteria, and its economy) can be traced back to the material organization of referential operations through which all these realities circulate (Latour, 1983; see also Latour, 1988). But this circulation does not happen without transformation, and the key notion here is 'translation.' The idea of translation is indeed the crucial vehicle for the material semiotic approach of ANT. It has been formulated in a number of writings, but Michel Callon's case study on the relationships between fishermen, scientists, and scallops in the Bay of Saint-Brieuc remains a classic (Callon, 1986). The idea of translation

operates as follows: the starting point, the empirical matter on which ANT operates, is best understood as a set of statements, which are embedded in a number of settings and which characterize a series of participating entities. Those statements (e.g., 'scallops do anchor,' 'fishermen want the bay restocked,' etc.) are deemed more or less problematic in the sense that they demand action to be taken, raise issues and concerns, or are prone to controversy. A translation, in ANT's sense, is an operation that transforms one particular problematic statement into the language of another particular statement: tackling problematic statement A in its own terms means in fact shifting terms and therefore tackling problematic statement B. This operation of translation transforms the issues at stake and the configuration of relations that link statements to one another. It also modifies acting capacities (since these statements are claims for conduct, and their strength is synonymous with power to act), and it engenders agency (since an operation of translation is both an empirical act and the condition of the configuration of a capable actor). This vocabulary resonates to some extent with an engineering-inspired problem-solving view, with some sort of a 'function' (in a mathematic sense) that would transform statements into other statements (early ANT formulations have indeed been criticized for this kind of tone, e.g., Star, 1991). But it also contains the theoretical potential of a nondualistic form of metaphysics: all that there is a network of ongoing operations of translation that define the agency of problematic statements. There is no need to revert to a distinction between an inside and an outside, or between actors and actions. The very notion of an 'actor–network' (a notion that spread through what was first labeled 'sociology of translation' but then became 'actor–network theory') conveys this intuition quite well: instead of considering a distinction between actors (that act) and networks (which actors form among themselves, in order, for example, to act collectively), the 'actor–network' is both the 'network-that-acts' and the 'actor-that-springs-from-a-network,' precisely what is constituted through processes of translation.

Constitution of Collective Worlds

Despite its praise for interdisciplinary and its distancing from a modern idea of 'the social' as displayed in standard social sciences, ANT portrays itself as some sort of a sociology (Latour, 2005). It does so in a quite particular sense: as a science of the formation and deformation of collective worlds, worlds that are constituted precisely through the material semiotic processes that ANT sets out to describe. This is 'sociological' not in the sense that it emphasizes 'social factors,' but rather in the sense that it generalizes a political understanding of the constitution of these worlds. ANT indeed is recognizable through the use of words imported from political science (e.g., 'spokesperson,' 'parliament,' 'association,' 'representative') in order to analyze, say, a laboratory and all its constituents. ANT can hence be seen as the generalization of a political frame to the understanding of all types of assemblages, not only human assemblies.

This strand of ANT is particularly important in contributions to political ecology and democratic theory (e.g., Latour, 2004b; Callon et al., 2009; Marres, 2012). These contributions claim that the ecological revolution in all its dimensions (generalization of environmental controversies, radical uncertainty on the externalities of innovation, unsustainability of the modern economic rationale, and proliferation of problematic entities) provide a blatant demonstration of the limitations of the standard social-scientific repertoire. Notions developed in ANT, such as 'hybrid forum' or 'parliament of things,' take into account both the democratic appraisal of issues of a technical nature – issues that provoke the emergence of new entities and new collectives – and the consideration of the constitution of nature as a political endeavor. Aspects of these ideas have had some impact in science and technology policy literature, especially through the topic of responsible innovation.

Insistence on the performative role that technical metrics, accounting methods, statistical techniques, management devices, marketing tools, or economic formulas (all vehicles for operations of translation) play in the constitution of the entities they refer to (the population, the firm, the economy, aggregate demand, profits and losses, capital, etc.) also partake of ANT's contribution to the understanding of the enactment of what standard social sciences call 'society' (Law, 1994, 2004; Power, 1996; Callon, 1998; Desrosières, 1998; Barry and Slater, 2005; Czarniawska and Hernes, 2005; Callon et al., 2007; MacKenzie et al., 2007). Investigations in the terrain of health care and biomedicine also illustrate the possible contributions of ANT to the understanding of the formation of bodily states and medical conditions and their relation to the formation of medical communities and patient organizations (Mol, 2002; Callon and Rabeharisoa, 2008; Keating and Cambrosio, 2003). Architecture and the built environment have also provided fertile ground for ANT explorations in the analysis of the construction of collective worlds (Farías and Bender, 2010; Yaneva, 2009). On these and other terrains, research inspired by ANT has contributed to novel orientations in the understanding of the formation of collective realities – or, in a sense, of society (at large).

Philosophical Developments

In the 1990s and 2000s, and in parallel to progress in a wide variety of empirical fields, proponents and commentators of ANT have engaged in a number of philosophical debates. In particular, ANT has reclaimed affiliation with the pragmatist tradition in philosophy, most notably with authors such as William James, John Dewey, Charles S. Peirce, and Alfred North Whitehead. The breadth of a nondualist approach to knowledge and of a materialist, semiotic approach to reality in pragmatism marks ANT's entanglement with this tradition. Rapprochement with the philosophical viewpoint of Gabriel Tarde obeys a similar logic, with an emphasis on contesting the dominant, modern understanding of society. Latour's philosophical elaborations also have played a pivotal role in the development of new approaches to metaphysics that focus attention on the reality and effectuation of objects (Harman, 2009; Latour et al., 2011).

Two versions of ANT, it can be said, are in circulation in the early first quarter of the current century. One, certainly more attached to the vocabulary developed in the 1980s and to the topics of inquiry that have most distinctively characterized the label 'ANT' through the years, works today as a widely used and acknowledged perspective in the social sciences, particularly

appropriate for the study of complex and controversial situations. Another one, rather experimental and philosophical in nature, shifts in a sense from exploitation to exploration, perhaps at the expenses of the neatness of the label 'ANT,' and engages with a number of serious issues of political philosophy. Perhaps the most crucial issue, which corresponds indeed to recent undertakings of Latour (2013), can be formulated as follows: how can a perspective that emerged as a frontal critique of the intellectual categories of modernity, including the political ones, consider ideals of political constitution (or political insurrection) that are nonetheless suitable for the maintenance and preservation, if not reparation, of a world largely inherited from modernity itself?

See also: Experiment in Science and Technology Studies; Laboratory Studies: Historical Perspectives; Latour, Bruno (1947–); Modernity: History of the Concept; Science and Technology Studies, Ethnomethodology of; Science and Technology, Anthropology of; Science, Sociology of; Science: Constructivist Perspectives, History of; Scientific Controversies; Social Constructivism; Visualization in Science and Technology.

Bibliography

Akrich, M., 1992. The de-scription of technical objects. In: Bijker, W.E., Law, J. (Eds.), Shaping Technology/Building Society: Studies in Sociotechnical Change. MIT Press, Cambridge, MA, pp. 205–224.

Akrich, M., Latour, B., 1992. A summary of a convenient vocabulary for the semiotics of human and nonhuman assemblies. In: Bijker, W.E., Law, J. (Eds.), Shaping Technology/Building Society: Studies in Sociotechnical Change. MIT Press, Cambridge, MA, pp. 259–264.

Barry, A., Slater, D. (Eds.), 2005. The Technological Economy. Routledge, London.

Bijker, W.E., Hughes, T.P., Pinch, T. (Eds.), 1987. The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology. MIT Press, Cambridge, MA.

Callon, M., 1980. Struggles and negotiations to define what is problematic and what is not: the socio-logic of translation. In: Knorr, K.D., Krohn, R., Whithley, R. (Eds.), The Social Process of Scientific Investigation. Reidel Publishing, Dordrecht, pp. 197–221.

Callon, M., 1986. Some elements of a sociology of translation: domestication of the scallops and the fishermen of St. Brieuc Bay. In: Law, J. (Ed.), Power, Action and Belief: A New Sociology of Knowledge? Routledge, London, pp. 196–233.

Callon, M., 1987. Society in the making: the study of technology as a tool for sociological analysis. In: Bijker, W.E., Hughes, T.P., Pinch, T. (Eds.), The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology. MIT Press, Cambridge, MA, pp. 83–103.

Callon, M. (Ed.), 1998. The Laws of the Markets. Blackwell, Oxford.

Callon, M., Lascoumes, P., Barthe, Y., 2009. Acting in an Uncertain World: An Essay on Technical Democracy. MIT Press, Cambridge, MA.

Callon, M., Latour, B., 1981. Unscrewing the big Leviathan: how actors macrostructure reality and how sociologists help them to do so. In: Knorr Cetina, K.D., Cicourel, A.V. (Eds.), Advances in Social Theory and Methodology: Toward an Integration of Micro- and Macro-sociologies. Routledge and Kegan Paul, London, pp. 277–303.

Callon, M., Law, J., 1982. On interests and their transformation: enrolment and counter-enrolment. Social Studies of Science 12 (4), 615–625.

Callon, M., Law, J., Rip, A. (Eds.), 1986. Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World. Macmillan, London.

Callon, M., Millo, Y., Muniesa, F. (Eds.), 2007. Market Devices. Blackwell, Oxford.

Callon, M., Rabeharisoa, V., 2008. The growing engagement of emergent concerned groups in political and economic life: lessons from the French Association of neuromuscular disease patients. Science, Technology, & Human Values 33 (2), 230–261.

Czarniawska, B., Hernes, T. (Eds.), 2005. Actor–Network Theory and Organizing. Business School Press, Malmö, Liber, and Copenhagen.

Desrosières, A., 1998. The Politics of Large Numbers: A History of Statistical Reasoning. Harvard University Press, Cambridge, MA.

Dosse, F., 1999. Empire of Meaning: The Humanization of the Social Sciences. University of Minnesota Press, Minneapolis.

Farías, I., Bender, T. (Eds.), 2010. Urban Assemblages: How Actor–Network Theory Changes Urban Studies. Routledge, London.

Gomart, E., Hennion, A., 1999. A sociology of attachment: music amateurs, drug users. In: Law, Hassard (Eds.), Actor Network Theory and After. Blackwell, Oxford, pp. 220–247.

Harman, G., 2009. Prince of Networks: Bruno Latour and Metaphysics. Re.press, Melbourne.

Hennion, A., 2005. Pragmatics of taste. In: Jacobs, M., Hanrahan, N.W. (Eds.), The Blackwell Companion to the Sociology of Culture. Blackwell, Oxford, pp. 131–144.

Hennion, A., 2007. Those things that hold us together: taste and sociology. Cultural Sociology 1 (1), 97–114.

Keating, P., Cambrosio, A., 2003. Biomedical Platforms: Realigning the Normal and the Pathological in Late-Twentieth-Century Medicine. MIT Press, Cambridge, MA.

Knorr-Cetina, K.D., Mulkay, M., 1983. Science Observed: Perspectives on the Social Study of Science. Sage, London.

Latour, B., 1983. Give me a laboratory and I will raise the world. In: Knorr-Cetina, K.D., Mulkay, M. (Eds.), Science Observed: New Perspectives on the Social Studies of Science. Sage, London, pp. 141–170.

Latour, B., 1987. Science in Action: How to Follow Scientists and Engineers through Society. Harvard University Press, Cambridge, MA.

Latour, B., 1988. The Pasteurization of France. Harvard University Press, Cambridge, MA.

Latour, B., 1992. Where are the missing masses? The sociology of a few mundane artifacts. In: Bijker, W.E., Law (Eds.), Shaping Technology/Building Society: Studies in Sociotechnical Change. MIT Press, Cambridge, MA, pp. 225–258.

Latour, B., 1993. We Have Never Been Modern. Harvard University Press, Cambridge, MA.

Latour, B., 1996. Aramis, or the Love of Technology. Harvard University Press, Cambridge, MA.

Latour, B., 1999a. Pandora's Hope: Essays on the Reality of Science Studies. Harvard University Press, Cambridge, MA.

Latour, B., 1999b. Factures/fractures: from the concept of network to that of attachment. RES: Anthropology and Aesthetics 36, 20–31.

Latour, B., 2000. When things strike back: a possible contribution of 'science studies' to the social sciences. British Journal of Sociology 51 (1), 107–123.

Latour, B., 2004a. Why has critique run out of steam? From matters of fact to matters of concern. Critical Inquiry 30 (2), 225–248.

Latour, B., 2004b. Politics of Nature: How to Bring the Sciences into Democracy. Harvard University Press, Cambridge, MA.

Latour, B., 2005. Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, Oxford.

Latour, B., 2007. Can we get our materialism back, please? Isis 98 (1), 138–142.

Latour, B., 2013. An Inquiry into Modes of Existence: An Anthropology of the Moderns. Harvard University Press, Cambridge, MA.

Latour, B., Harman, G., Erdélyi, P., 2011. The Prince and the Wolf: Latour and Harman at the LSE. Zero Books, Alresford.

Latour, B., Woolgar, S., 1979. Laboratory Life: The Social Construction of Scientific Facts. Sage, London.

Law, J. (Ed.), 1986. Power, Action and Belief: A New Sociology of Knowledge?. Routledge, London.

Law, J., 1994. Organizing Modernity. Oxford, Blackwell.

Law, J., 2004. After Method: Mess in Social Science Research. Routledge, London.

Law, J., 2009. Actor network theory and material semiotics. In: Turner, B.S. (Ed.), The New Blackwell Companion to Social Theory. Blackwell, Oxford, pp. 141–158.

MacKenzie, D., Muniesa, F., Siu, L. (Eds.), 2007. Do Economists Make Markets? On the Performativity of Economics. Princeton University Press, Princeton.

Marres, N., 2012. Material Participation: Technology, the Environment and Everyday Publics. Palgrave, Basingstoke.

Mol, A., 2002. The Body Multiple: Ontology in Medical Practice. Duke University Press, Durham, NC.

Mol, A., 2010. Actor–network theory: sensitive terms and enduring tensions. Kölner Zeitschrift für Soziologie und Sozialpsychologie 50 (1), 253–269.

Pickering, A. (Ed.), 1992. Science as Practice and Culture. University of Chicago Press, Chicago.

Pickering, A., 1995. The Mangle of Practice: Time, Agency and Science. University of Chicago Press, Chicago.

Power, M. (Ed.), 1996. Accounting and Science: Natural Inquiry and Commercial Reason. Cambridge University Press, Cambridge.

Serres, M., Latour, B., 1995. Conversations on Science, Culture and Time. The University of Michigan Press, Ann Arbor.

Star, S.L., 1991. Power, technology and the phenomenology of conventions: on being allergic to onions. In: Law, J. (Ed.), A Sociology of Monsters: Essays on Power Technology and Domination. Routledge, London, pp. 26–56.

Yaneva, A., 2009. The Making of a Building. Peter Lang, Oxford.

Adaptation, Fitness, and Evolution

Ward B Watt, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Adaptation,' a central concept in evolutionary biology, represents the degree of suitedness between organisms' characteristics and their environments. It is the cause of differences in organisms' performances of biological tasks that lead throughout their life cycles to reproductive success, the distinct concept of 'fitness.' Organisms' characteristics need not be adaptive but may be 'neutral' (i.e., they have no consequences for fitness or may be subject to various constraints). Misperceptions of adaptation may result from confusions of terminology. The assumption that all features of organisms are adaptive is common but misleading. Future study of adaptation will include diverse approaches and greater rigor of conceptual analysis.

Introduction

The adaptation, or adaptedness, of organisms to their environments is a central concept in evolutionary biology. It is both a striking phenomenon needing explanation and a basic feature of the mechanisms underlying the patterns of evolutionary stasis and change alike. The organism–environment interaction that the adaptation concept embodies is the causal driver of the process of evolution by natural selection. Its nature, role in the evolutionary concept structure, and limitations must all be understood if a clear view of evolution is to be possible. In particular, adaptations distinctness from and relation to the concept of fitness must be seen clearly. Only thus can evolution by natural selection be properly understood.

Adaptations Identity and Its Distinction from Fitness

If no concept is more central to evolution by natural selection than adaptation, then also none has been more debated. All the basic features of its definition are found in the work of Darwin, but progress in unfolding its full scope and implications continue even at present. Biological evolution, as distinct from cultural evolution (though often interwoven with it; e.g., Cavalli-Sforza and Feldman, 1981), is manifested as change in the genetic composition of populations over time (*see* Microevolution). Therefore, some genetic terminology is needed at the outset. A 'gene' is a functionally coherent sequence of bases in nucleic acid, usually DNA (except for some viruses), determining or influencing some biological structure and/or function. An 'allele' is one possible sequence (a variant) of a gene, determining one alternative state of gene action. Many organisms, including most animals, carry two copies of each gene (and are thus termed 'diploid'). 'Genotype' refers to the whole heritable composition of a creature, whether viewed gene by gene (e.g., carrying two copies of the same allele, hence a 'homozygote,' or one copy each of two different alleles, hence a 'heterozygote') or more broadly up to the whole 'genome' which includes all genes. 'Phenotype' refers to the expressed structure and function of the organism as it develops via interactions of its genotype with the environment in which development takes place (*see* Genotype and Phenotype). Present understanding of the complexities of the evolutionary process requires this terminology to avoid ambiguity and confusion.

Basic Definitions of Adaptation

As a general concept, adaptation or adaptedness is best defined as the extent of matching or suitedness between the heritable features (heritable functional phenotypes) of organisms and the environments in which they occur. In other words, adaptation comprises genotype–phenotype–environment interactions. It finds direct expression in the effectiveness with which organisms perform essential biological tasks (osmoregulation, locomotion, capturing food, evading predators, etc.) in their environments. As such, its states are in principle quantitatively measurable, or at least orderable, rather than only qualitatively organized. This general definition is found in many parts of Darwin's writing, as in the introduction to his *Origin of Species* (1859/1872) where a woodpecker appears equipped "with its feet, tail, beak, and tongue, so admirably adapted to catch insects under the bark of trees." Here, the phenotypic states of these morphological characters, modified as compared to simpler forms found in other birds, are related to their functional performance effects in acquiring food resources that other birds, lacking those specific adaptive phenotypic states, cannot reach.

Adaptation also refers to the process of successively descended, modified phenotypes becoming more suited, 'better adapted,' to environments via natural selection on variation in those phenotypes. Paleontology finds strong evidence for improvement of adaptation over time, as in the escalation of predator and prey attack and defense morphologies in marine invertebrates (Vermeij, 1987). Real-time studies have shown adaptive improvement directly, such as in the evolution of a bacterial stock in novel culture conditions over periods of 10^4 generations: a stock at an early stage in the process, if samples are frozen for later reactivation, is found to be inferior in performance to its better adapted descendants sampled late in the experimental history (Lenski and Travisano, 1994).

This evolutionary refinement of an adaptive state has led to debate over when a phenotypic feature may be called 'an' adaptation and when it may not be (i.e., how far it has been specifically selected for its current functional state). Given that

adaptive states differ quantitatively, any viable phenotype has some level of adaptation, and this debate loses urgency. Recognizing a phenotype as 'an' adaptation only if it is the best available at a given time (as argued by Reeve and Sherman, 1993) would require continuous revision as newer alternatives arise, and it offers no compensating advantage.

Elaborations of the Basic Concept

Gould and Vrba (1982) extended and refined definitions of adaptation in useful ways. In their terminology, 'aptation' describes the primary, historically unmodified relation of suitedness between phenotype and environment – that of any newly arisen variant, positive or negative, in its functional effects. They regarded 'adaptation' as the successive refinement of phenotypic suitedness by selection of newer variants, and coined the term 'exaptation' for the coopting of a phenotypic feature by selection for a new function, as in the modification of skull–jaw joint bones toward the ossicles of vertebrate ears (e.g., Romer, 1955). The exaptation–adaptation distinction poses problems of discrimination (how much change under a new selection pressure is needed before a phenotype of *ex*aptive origin is recognizable as presently *ad*aptive? See Reeve and Sherman, 1993) and also emphasizes that we are dealing with quantitative scales of variation, not alternate qualitative categories. Often the Gould–Vrba terms are not used unless the distinctions are pertinent to the issue at hand, and otherwise 'adaptation' is used as a generally inclusive term.

Another important extension of the adaptation concept is the work of Laland et al. (1996, 1999) on 'niche construction.' This term refers to the active modification of environments by organisms in ways favorable to their own function and fitness, as emphasized by Lewontin (1983). It occurs in diverse ways in different groups: for example, bacteria may release protease enzyme catalysts into their surroundings to aid in foraging upon potential food items, while among multicellular animals beaver lodges and dams are a dramatic case of such activities (aside from the obvious capabilities of humans in this direction). Evolutionary models incorporating niche-constructive feedbacks on organism–environment interactions may have very distinct properties from those not including such active forms of adaptation (Laland et al., 1996, 1999).

The Distinction between Adaptation and Fitness

Alternative states of adaptation are the causes of evolutionary changes through their differences in genotype–phenotype–environment interactions and hence performances of these phenotypes. These performances, minute by minute to year by year, cumulatively alter how long individuals live and how much they reproduce. In short, adaptive differences among phenotypes alter their demographic parameters: survivorship ($= l_x$ of demography, where x denotes time intervals) and male mating success or female fecundity ($= m_x$ of demography). These parameters are components of what, since the advent of mathematical population genetics, has been termed 'fitness' or 'Darwinian fitness' (though Darwin did not use the word in this way): the reproductive success of whole populations or of specific genotypes. Adaptation and fitness, then, are serially related concepts, but are in no sense the same.

In evolutionary genetics, fitness is usually measured as the net replacement rate of organisms, whether an average value for a whole population or more specific average values for particular genotypes. It is defined in 'absolute' terms as $R = \Sigma l_x m_x$ (e.g., Roughgarden, 1979) under simple demographic conditions of nonoverlapping generations and homogeneous reproductive periods (as, e.g., in annual plants or many insects). For complex demography in age-structured populations, the most similar expression is λ (the leading eigenvalue of the demographic 'Leslie matrix'), a number that summarizes complex interactions of age-specific survivorships and fecundities (e.g., Charlesworth, 1994; McGraw and Caswell, 1996). The *concept* of fitness is the same among these cases; what varies is the *measure* of fitness as is proper to each case. If either R or λ, as appropriate, is compared among genotypes by taking the ratio of each value to that of a chosen standard genotype, there result 'relative' genotypic fitnesses, whose value for the standard genotype is 1. Most evolutionary-genetic models use relative fitnesses for symbolic or numeric convenience.

Usage of the terms adaptation and fitness has changed dramatically since Darwin. He, Wallace, and other early evolutionists used 'fitness' as a synonym for 'adaptation,' and by 'survival' they often referred not to the demographers' life cycle variable l_x but to 'persistence over long time periods.' Spencers phrase 'survival of the fittest,' translated, meant 'the persistence through time of the best adapted.' Darwin had (necessarily) a clear view of the concept that evolutionary biologists now denote by the term 'Darwinian fitness,' but he represented it by versions of a stock phrase (for which he had no summary term), "the best chance of surviving and of procreating," in the *Origin of Species* (1859/1872). Failure to recognize these usage changes, and thus blurring of the sharp distinction between adaptation as cause and fitness as within-generation result, has led to much confusion in later literature, including mistaken claims of an alleged circularity of evolutionary reasoning.

The Roles of Adaptation and of Fitness in Darwin's Argument for Natural Selection

From the inceptions of both Darwin's and Wallace's ideas of natural selection, differences in adaptation among heritable variants played the central, causal role in the process (*see* Darwinism). Darwin formalized his argument in Chapter 4 of *Origin*, especially in its first paragraph and its concluding summary, in such a way that it can be cast as a verbal theorem – as Depew and Weber (1995) make clear by judicious editing of Darwin's summary. Here it may be reformulated in modern terms.

To begin the argument, there are three points 'given' by direct observation: (1) organisms vary in phenotype; (2) some of the variants are heritable; and (3) some of these heritable variants can perform their biological functions differently in a specific habitat (i.e., some are better adapted than others to that habitat). Then, Darwin's Postulate is that the better adapted, and hence better performing, variants in a habitat will survive and/or reproduce more effectively over their life cycles (i.e., have higher fitness) than other variants. Demography shows that greater reproduction of variants will maintain or

increase their frequencies in successive generations of a population. One thus concludes that when the Postulate holds, the best-adapted heritable phenotypes will persist and/or increase in frequency over time, realizing evolution by natural selection. This completes Darwin's Theorem.

The distinction between differences in organisms' adaptive performances, minute by minute to year by year, and resulting fitness differences among them over their life spans is simply the difference between cause and effect. Its recognition is essential to keep straight the logic of natural selection and to organize empirical studies of the process (Feder and Watt, 1992; Watt, 1994).

Alternatives to Adaptation in Evolution

Adaptation is not ubiquitous, and natural selection is not all-powerful. 'Darwin's Theorem,' as summarized in this article, is empirically testable and indeed may not hold in some well-defined circumstances. Two main sources of limitation on the scope of adaptation are now considered.

Neutrality

As Darwin wrote in Chapter 4 of *Origin*, "Variations neither useful nor injurious would not be affected by natural selection." The modern concept of neutrality (Kimura, 1983; Gillespie, 1991), which he thus described, is the null hypothesis for testing all causal evolutionary hypotheses. It occurs at each of the recursive stages of natural selection, as recognized by Feder and Watt (1992).

First, at the **genotype → phenotype** stage, genetic variants may differ in sequence but not in resulting function. For example, the 'degeneracy' of the genetic code often means that differences in DNA base sequence lead only to the same amino acid's insertion into a given position of a protein molecule. Alternatively, at least in the case of some positions in proteins, substitution for one amino acid residue by a similar one may sometimes have little effect on the protein's function.

Next, at the **phenotype → performance** stage, functional differences among variants may not lead to performance differences among them, as other phenotypic mechanisms constrain or suppress their potential effects. For example, in the physiological reaction pathway used by bacteria to digest milk sugar (lactose), a twofold range of natural genetic variation in a phenotypic parameter (the V_{max}/K_m ratio) is observed for each of the protein catalysts, or enzymes, catalyzing the first two reactions. When these variants' resulting performances were measured under steady-state growth conditions, variants of the first enzyme in the pathway showed sizable, reproducible differences, but no such effects were seen among variants of the second enzyme despite the similar size of their phenotypic differences – due to system constraints related to the position of the reactions in the pathway, analyzable by the theory of metabolic organization (Watt and Dean, 2000).

At the stage of **performance → fitness**, performance differences may not lead to corresponding fitness differences (e.g., if improved performance has less fitness impact above a threshold value of habitat conditions). For example, performance differences among feeding phenotypes (bill sizes and geometries) of Darwin's finches have little fitness impact when food is plentiful in wet seasons, but have much more impact when it is scarce in dry seasons (Grant, 1986).

Finally, at the stage of **fitness → genotype**, which completes the natural-selective recursion, small population size can allow random genetic drift to override fitness differences, as in the loss from small mouse populations of developmental ('t-system') mutant alleles that should be in frequency equilibrium between haploid gametic selection favoring them and recessive lethality at the diploid developing-phenotypic stage of the life cycle (Lewontin and Dunn, 1960). Because the usual statistical null hypothesis is that no treatment effect exists between groups compared, any adaptive hypothesis of difference between heritable phenotypes is *ipso facto* evaluated against neutrality by statistical testing.

Further, there is a subtler neutral hypothesis, that of association or 'hitchhiking': variants that seem to differ in fitness at a gene under study may be functionally neutral but genetically linked to an unobserved gene whose variants are the real targets of selection. But 'hitchhiking' predicts that fitness differences seen among variants will not follow from any functional differences among them, so it is rejected when prediction from function to fitness is accurate and successful.

Indeed, where substantive adaptive difference exists among genetic variants in natural populations, neutral null hypotheses may be rejected by testing at any of these levels, from phenotypic function to its predictable fitness consequences and the persistence or increase of the favored genotypes. This has been done, for example, for natural variants of an energy-processing enzyme in the 'sulfur' butterflies, *Colias* (Watt, 1992). The explicit test of adaptive hypotheses against neutral nulls gives important rigor to experimental study of natural selection in the wild (Endler, 1986).

Constraint

Gould (1980, 1989) emphasized that many features of organisms may not result from natural selection, but rather from various forms of constraint due to unbreakable geometric or physical properties of the universe at large or of the materials from which organisms are built, or other, more local biological limitations or conflicts of action. Functional or geometric constraints may play a major role in the form or function of organisms (e.g., in snail shells' form) (Gould, 1989), or in the negative interactions in combinations of individually positive amino acid changes, which constrain the paths of positive evolution of antibiotic resistance in bacteria (Weinreich et al., 2006).

Selection among phenotypic alternatives at one time may entail diverse predispositions or constraints at later times. In one such case, the tetrapodal nature of all land-dwelling vertebrate animals (the bipedality of birds, kangaroos, or hominid primates is secondary) follows from the historical constraint that their ancestors, lobe-finned fish, swam with two pairs of oarlike ventral fins having enough structural strength *ab initio* that they could be exaptively (i.e., not through natural selection) modified into early legs (e.g., Gould, 1980; Cowen, 1995). In a more pervasive case, the evolved rules of diploid, neo-Mendelian genetics constrain many evolutionary paths. For example, if a heterozygous genotype is the best adapted,

and hence most fit, in a population, it can rise to high frequency in that population but cannot become the only genotype present because it does not 'breed true.'

Conflicts among different aspects of natural selection may constrain the precision of adaptation in diverse ways. As a case in point, adjustment of insects' thermoregulatory phenotypes may be held short of maximal or 'optimal' matching to average conditions in cold, but highly variable, habitats, because such 'averagely optimized' phenotypes would overheat drastically in uncommon but recurrent warm conditions (Kingsolver and Watt, 1984). This illustrates the general point that environmental variance may sharply constrain adaptation to environmental means.

Misdefinitions of Adaptation or Misconceptions of Its Role

Many misdefinitions of adaptation err by confusing it with fitness in one fashion or another. Much of this may originate in the usage changes, discussed earlier in this article, between the early Darwinians and the rise of evolutionary genetics, such that 'fitness' ceased to be a synonym of adaptation and came to mean instead the 'best chance of surviving and of procreating' (e.g., Darwin, 1859/1872, p. 63). This entirely distinct concept is, as noted above, the cumulative demographic effect of adaptation. Some writers on evolutionary topics have been confused by inattention to these usage changes, but others have erred through conscious disregard or blurring of the adaptation–fitness distinction.

For example, Michod (1999), despite early recognition of the separate nature of adaptation and fitness and of their antecedent–consequent relationship (Bernstein et al., 1983), sought to collapse these concepts into different 'senses' of the single term 'fitness' to be used in different contexts to refer to both 'adaptive attributes' and their consequences in reproductive success. Authors may choose terminology for their own uses within some limits, but this usage is at best an ill-advised source of confusion, and at worst a mistaken conflation of distinct concepts.

Another misconception was asserted by Lewontin (1983) during an otherwise important argument for studying 'niche construction' (cf. above; Laland et al., 1996, 1999). Arguing that the adaptation concept implies 'passiveness' of adapting organisms, he criticized it for allegedly implying that adaptation is like 'filing a key to fit a preexisting lock.' But no such passivity is really in evidence. In Darwin's example above, the woodpecker's feeding 'strategy' actively transforms its environment compared to that experienced by more generally feeding birds, using a resource that those other birds do not even perceive. Further, Darwin's discussions (1859/1872, Chapter 4) of mutualisms between flowers and pollinators also show the constructive nature of those adaptations: pollinators obtain resource rewards from plants and spread their pollen during their foraging, and by supporting the reproduction of their food sources in this way, they increase their own future resource bases. Niche construction is thus *an important form of adaptation*, not distinct from or opposed to it.

Lewontin also misstated the role of adaptation in the evolutionary process, arguing that 'three propositions' – variation, heritability, and differential reproduction – alone were sufficient to explain natural selection, and that the adaptation concept was gratuitously introduced into the argument by Darwin for sociological reasons (e.g., Lewontin, 1984). This claim is wrong and has been widely critiqued (e.g., Hodge, 1987; Brandon, 1990; Watt, 1994; Depew and Weber, 1995). Adaptation is the one element that distinguishes natural selection from artificial selection or sexual selection. Without it, Lewontin's three propositions are sufficient only to define 'arbitrary' selection, wherein we do not know the cause of any differential reproduction of heritable variants. But the adaptive cause is, indeed, central to evolutionary change resulting from natural selection.

Finally, a serious barrier to effective study of adaptation has been a claim by Mayr (1961, 1980) that 'proximate causes' (phenotypic mechanisms) have little to do with 'ultimate causes' (evolution by natural selection) in biology. Since adaptation comprises genotype–phenotype–environment interactions as the driver of natural selection, this 'proximate–ultimate' dichotomy distorts the basic nature of the evolutionary process and discourages the breadth of approach needed for study of evolution (Watt, 2000, 2013; West-Eberhard, 2003; Laland et al., 2011).

Adaptationism and Its Drawbacks

Rose and Lauder (1996) identify adaptationism as "a style of research … in which all features of organisms are viewed a priori as optimal features produced by natural selection specifically for current function." Some, such as Parker and Maynard Smith (1990) or Reeve and Sherman (1993), hail the assumption of adaptiveness as a virtue, while others (e.g., Gould and Lewontin, 1979) indict it as a vice. The question is: is it helpful, or legitimate, to assume that adaptation is ubiquitous?

First, is it true that adaptiveness is often assumed in practice? The usual null hypothesis in statistical testing is that there is no 'treatment effect.' Thus, any statistical test of adaptive difference among character states assumes *ab initio* that there is *no* such difference (i.e., that the character states in question are *neutral*). Only if this null hypothesis can be rejected according to standard decision rules is an effect recognized. All the null models of population genetics itself, beginning with the single-gene Hardy–Weinberg distribution, start with neutral assumptions. Tests of the population genetic consequences of putatively adaptive differences in phenotypic mechanisms may or may not find departure from neutrality, but it is the routine 'starting point.'

Mayr (1988) argued for testing all possible adaptive causes for phenotypes before turning to the explanation of chance (i.e., neutral) origins. But this argument depends on a historicist approach to evolution. If one can analyze a phenotype by testing among neutrality, constraint, or adaptation with present-day experiments, historicism is not needed. Even fossil structures absent in living relatives may be studied functionally by various means (Hickman, 1988). A historical approach may sometimes be indispensable, but it is not the only one available to evolutionary biology.

As Gould (1980) observed, assuming the ubiquity of adaptation discourages attention to structural or constrained

alternative explanations of phenotypes. It is not enough merely to test one specifically adaptive hypothesis about some phenotype against neutrality; one should consider all feasible alternative hypotheses, including the constraint-based one that a phenotype does make a nonneutral difference to performance and thence fitness, but does so in a particular way because no other way is feasible or possible, rather than because it is 'optimized.'

Indeed, the strongest objection to adaptationism may be that using the optimal adaptedness of a phenotypic feature as a null hypothesis (as sometimes suggested by adherents of this view) runs the serious risk of falling victim to 'the perils of preconception.' How can a scientist making such an attempt know that the phenotypic function has been correctly identified, or that an appropriate adaptive hypothesis has been arrived at, to begin with (cf. Gordon, 1992)?

As a cautionary illustration, the behavior of certain 'laterally basking' butterflies in orienting their wings perpendicular to sunlight was at first guessed to be an adaptation to minimize casting of shadows, hence to avoid predators' attention. More careful study shows that it does not do so! Parallel orientation of the closed wings to the solar beam truly minimizes shadow. It was instead shown experimentally, with proper testing against neutral null hypotheses, that the perpendicular solar orientation is adaptive, but in relation to thermoregulatory absorption of sunlight (Watt, 1968).

Some users of adaptationist approaches do recognize these concerns, and they build optimizing models for testing in comparison to possible constraints or other alternative explanations (e.g., Houston and McNamara, 1999). Nonetheless, the intellectual hazards of assuming the adaptiveness of phenotypes outweigh the possible advantages. Certainly, studies of adaptive mechanisms in diverse organisms are routinely carried out, achieving results that are both rigorous and generalizable, without this assumption (e.g., Lauder, 1996; Watt and Dean, 2000).

The Future Study of Adaptation

Mechanistic studies of adaptation in the wild are increasing in diversity and effectiveness, as in the application of biomechanical approaches to the function of morphological adaptations (Lauder, 1996), or of molecular approaches to adaptation in metabolism and physiology (Watt and Dean, 2000). Diverse tools of modern 'genomics' and 'proteomics' may be used in the study of adaptation in complementary ways, either (1) focusing at first on natural variation in genes and processes of known function, and working upward through their genotype–phenotype–environment interactions to their performance and fitness effects in the wild (e.g., Wheat et al., 2006; Barrett and Hoekstra, 2011); or (2) screening organisms' genomes broadly for genetic variation that is associable with biogeographic or other patterns of possibly adaptive biological specialization (e.g., Fournier-Level et al., 2011; Hancock et al., 2011), then working downward to identify those genes and variants' roles in genotype–phenotype–environment interactions and their effects (e.g., Prasad et al., 2012).

At the same time, philosophical ground clearing may reduce misunderstanding or misapplication of the adaptation concept, and lead to better specific work as well as greater possibilities for general insight (Brandon, 1990; Lloyd, 1994; Watt, 2000). There has often been tension between the study of well-known 'model' systems, which can maximize experimental power, and the fascination with diversity that drives the study of evolution for many workers. Both have value for the study of adaptation, and the tension may be eased by the interplay of comparative and phylogenetic studies (Larson and Losos, 1996) with genetics-based experimental or manipulative study of organism–environment interactions and their demographic consequences in the wild. This synergism of diverse empirical and intellectual approaches holds great promise for the widening study of adaptation as a central feature of evolution by natural selection.

See also: Cultural Evolution: Theory and Models; Darwinism; Evolution, History of; Evolutionary Theory, Structure of; Genotype and Phenotype; Microevolution.

Bibliography

Barrett, R.D.H., Hoekstra, H.E., 2011. Molecular spandrels: tests of adaptation at the genetic level. Nature Reviews Genetics 12, 767–780.

Bernstein, H., Byerly, H.C., Hopf, F.A., Michod, R.E., Vemulapalli, G.K., 1983. The Darwinian dynamic. Quarterly Review of Biology 58, 185–207.

Brandon, R.N., 1990. Adaptation and Environment. Princeton University Press, Princeton, NJ.

Cavalli-Sforza, L.L., Feldman, M.W., 1981. Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press, Princeton, NJ.

Charlesworth, B., 1994. Evolution in Age-Structured Populations, second ed. Cambridge University Press, Cambridge, UK.

Cowen, R., 1995. The History of Life, second ed. Blackwell Scientific Publications, Oxford, UK.

Darwin, C., 1859/1872. The Origin of Species, sixth rev. ed. New American Library, New York.

Depew, D.J., Weber, B.H., 1995. Darwinism Evolving. MIT Press, Cambridge, MA.

Endler, J.A., 1986. Natural Selection in the Wild. Princeton University Press, Princeton, NJ.

Feder, M.E., Watt, W.B., 1992. Functional biology of adaptation. In: Berry, R.J., Crawford, T.J., Hewitt, G.M. (Eds.), Genes in Ecology. Blackwell Scientific Publications, Oxford, UK, pp. 365–392.

Fournier-Level, A., Korte, A., Cooper, M.D., Nordborg, M., Schmitt, J., Wilczek, A.M., 2011. A map of local adaptation in *Arabidopsis thaliana*. Science 334, 86–89.

Gillespie, J.H., 1991. The Causes of Molecular Evolution. Oxford University Press, Oxford, UK.

Gordon, D.M., 1992. Wittgenstein and ant-watching. Biology and Philosophy 7, 13–25.

Gould, S.J., 1980. The evolutionary biology of constraint. Daedalus 109, 39–52.

Gould, S.J., 1989. A developmental constraint in *Cerion*, with comments on the definition and interpretation of constraint in evolution. Evolution 43, 516–539.

Gould, S.J., Lewontin, R.C., 1979. The spandrels of San Marco and the Panglossian paradigm. Proceedings of the Royal Society of London B205, 581–598.

Gould, S.J., Vrba, E.S., 1982. Exaptation – a missing term in the science of form. Paleobiology 8, 4–15.

Grant, P.R., 1986. Ecology and Evolution of Darwin's Finches. Princeton University Press, Princeton, NJ.

Hancock, A.M., Brachi, B., Faure, N., Horton, M.W., Jarymowycz, L.B., Sperone, F.G., Toomajian, C., Roux, F., Bergelson, J., 2011. Adaptation to climate across the *Arabidopsis thaliana* genome. Science 334, 83–86.

Hickman, C.S., 1988. Analysis of form and function in fossils. American Zoologist 28, 775–793.

Hodge, M.J.S., 1987. Natural selection as a causal, empirical, and probabilistic theory. In: Kruger, L., Gigerenzer, G., Morgan, M.S. (Eds.), The Probabilistic Revolution, vol. 2. MIT Press, Cambridge, MA, pp. 233–270.

Houston, A.I., McNamara, J.M., 1999. Models of Adaptive Behaviour. Cambridge University Press, Cambridge, UK.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK.

Kingsolver, J.G., Watt, W.B., 1984. Mechanistic constraints and optimality models: thermoregulatory strategies in *Colias* butterflies. Ecology 65, 1835–1839.

Laland, K.N., Odling-Smee, F.J., Feldman, M.W., 1996. The evolutionary consequences of niche construction: an investigation using two-locus theory. Journal of Evolutionary Biology 9, 293–316.

Laland, K.N., Odling-Smee, F.J., Feldman, M.W., 1999. Evolutionary consequences of niche construction and their implications for ecology. Proceedings of the National Academy of Sciences of the United States of America 96, 10242–10247.

Laland, K.N., Sterelny, K., Odling-Smee, F.J., Hoppit, W., Uller, T., 2011. Cause and effect in biology revisited: is Mayr's proximate-ultimate dichotomy still useful? Science 334, 1512–1516.

Larson, A., Losos, J.B., 1996. Phylogenetic systematics of adaptation. In: Rose, M.R., Lauder, G.V. (Eds.), Adaptation. Academic Press, New York, pp. 187–220.

Lauder, G.V., 1996. The argument from design. In: Rose, M.R., Lauder, G.V. (Eds.), Adaptation. Academic Press, New York, pp. 55–91.

Lenski, R.E., Travisano, M., 1994. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. Proceedings of the National Academy of Sciences of the United States of America 91, 6808–6814.

Lewontin, R.C., 1983. Gene, organism, and environment. In: Bendall, D.S. (Ed.), Evolution from Molecules to Men. Cambridge University Press, Cambridge, UK, pp. 273–285.

Lewontin, R.C., 1984. Adaptation. In: Sober, E. (Ed.), Conceptual Issues in Evolutionary Biology. MIT Press, Cambridge, MA, pp. 235–251.

Lewontin, R.C., Dunn, L.C., 1960. The evolutionary dynamics of a polymorphism in the house mouse. Genetics 45, 705–722.

Lloyd, E.A., 1994. The Structure and Confirmation of Evolutionary Theory, second ed. Princeton University Press, Princeton, NJ.

Mayr, E., 1961. Cause and effect in biology. Science 134, 1501–1506.

Mayr, E., 1980. Some thoughts on the history of the evolutionary synthesis. In: Mayr, E., Provine, W.B. (Eds.), The Evolutionary Synthesis. Harvard University Press, Cambridge, MA, pp. 1–48.

Mayr, E., 1988. Toward a New Philosophy of Biology. Harvard University Press, Cambridge, MA.

McGraw, J.B., Caswell, H., 1996. Estimation of individual fitness from life-history data. American Naturalist 147, 47–64.

Michod, R.E., 1999. Darwinian Dynamics. Princeton University Press, Princeton, NJ.

Parker, G.A., Maynard Smith, J., 1990. Optimality theory in evolutionary biology. Nature 348, 27–33.

Prasad, K.V.S.K., Song, B.-H., Olson-Manning, C., Anderson, J.T., and 11 others, 2012. A gain-of-function polymorphism controlling complex traits and fitness in nature. Science 337, 1081–1084.

Reeve, H.K., Sherman, P.W., 1993. Adaptation and the goals of evolutionary research. Quarterly Review of Biology 68, 1–32.

Romer, A.S., 1955. The Vertebrate Body, second ed. Saunders, Philadelphia, PA.

Rose, M.R., Lauder, G.V., 1996. Post-spandrel adaptationism. In: Rose, M.R., Lauder, G.V. (Eds.), Adaptation. Academic Press, New York, pp. 1–8.

Roughgarden, J., 1979. Theory of Population Genetics and Evolutionary Ecology: An Introduction. Macmillan, New York.

Vermeij, G.J., 1987. Evolution and Escalation. Princeton University Press, Princeton, NJ.

Watt, W.B., 1968. Adaptive significance of pigment polymorphisms in *Colias* butterflies. I. Variation of melanin pigment in relation to thermoregulation. Evolution 22, 437–458.

Watt, W.B., 1992. Eggs, enzymes, and evolution – natural genetic variants change insect fecundity. Proceedings of the National Academy of Sciences of the United States of America 89, 10608–10612.

Watt, W.B., 1994. Allozymes in evolutionary genetics: self-imposed burden or extraordinary tool? Genetics 136, 11–16.

Watt, W.B., 2000. Avoiding paradigm-based limits to knowledge of evolution. Evolutionary Biology 32, 73–96.

Watt, W.B., 2013. Causal mechanisms of evolution and the capacity for niche construction. Biology and Philosophy, doi: 10.1007/s10539-012-9335-1.

Watt, W.B., Dean, A.M., 2000. Molecular-functional studies of adaptive genetic variation in prokaryotes and eukaryotes. Annual Review of Genetics 34, 593–622.

Weinreich, D.M., Delaney, N.F., DePristo, M.A., Hartl, D.L., 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312, 111–114.

West-Eberhard, M.J., 2003. Developmental Plasticity and Evolution. Oxford University Press, Oxford, UK.

Wheat, C.W., Watt, W.B., Pollock, D.D., Schulte, P.M., 2006. From DNA to fitness differences: sequences and structures of adaptive variants of *Colias* phosphoglucose isomerase (PGI). Molecular Biology and Evolution 23, 499–512.

Addams, Jane (1860–1935)

Louise W Knight, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Jane Addams (6 September 1860–21 May 1935) was a world-renowned social activist, public intellectual, political leader, author, and lecturer, and the first American woman (and second woman) to receive the Nobel Peace Prize. She was also a brilliant social thinker, at home in the humanities field of philosophy, and the social science fields of social psychology, sociology, anthropology, and political science. As the founder and leader of the first settlement house in the United States, Hull House in Chicago, she is considered to be a founder of the field of social work.

Biography

Born in Cedarville, Illinois, to a leading agricultural businessman and a woman admired for her charitable works, Jane Addams earned her B.A. degree from Rockford Female Seminary (later Rockford College), and aspired to earn an MD degree but abandoned that goal after an initial attempt. She spent much of her twenties ambitious to contribute to society but unsure what to do. She decided to found a settlement house after reading about the world's first settlement house, Toynbee Hall, in London. Following a visit there, she and a friend, Ellen Gates Starr, opened a settlement house in Chicago, Hull House, in September 1889.

Initially, Addams's intention was to offer clubs, classes, concerts, and sports to enrich the lives of her working class, mostly immigrant neighbors. However, they quickly educated her about the challenges they and their children faced due to low wages, dangerous working conditions, and long hours. As a result, while the settlement house remained committed to social and cultural enrichment, Addams's purposes expanded to include seeking ways to work with labor unions and women's organizations, some of them cross-class, to improve the working conditions of working people. Starting in 1893, when she lobbied a legislature for the first time, Addams became a strong advocate for worker rights, especially as they affected women and children. In 1903, she cofounded, along with other settlement leaders and labor organizers, the Women's Trade Union League. She was a member of the National Child Labor Committee and supporter of the 8-h day, the state women's pension laws (and later the federal law Aid to Dependent Children), and the minimum wage.

Eager for working women and women like herself to have the vote so that they could apply more effective political pressure on legislatures, Addams became a vice president of the National American Woman's Suffrage Association in 1911. She was among the nation's most prominent advocates for that reform, giving speeches to unions, clubs, college groups, and churches, and, in 1912, to voters, as part of her campaigning for Theodore Roosevelt as the presidential candidate for the Progressive Party.

Meanwhile, as the field of social work emerged in the 1890s, Addams found herself considered one of its early leaders, though she never referred to herself formally as a social worker. The field, consisting of prison workers, settlement workers (paid and unpaid), and charity workers (paid and unpaid), had disagreed in the 1890s and early twentieth century over the question of nature versus nurture (Addams siding with the 'nurture' position) but Addams's election in 1911 as the first woman president of the National Council of Charities and Correction (precursor to the National Council of Social Work) signaled that a majority of charity and social workers had come to agree that the environment played a critical role in 'causing' poverty.

Addams's path to peace advocacy began with her embrace, in her twenties, of Leo Tolstoy's ideas about Christianity, including his emphasis on the importance of nonresistance to evil. She had been interested in peace as early as 1893, when she became a member of the tiny Chicago Peace Society. Her central conviction, inspired by her interpretation of Tolstoy's Christian writings, was that antagonism of any sort, whether between individuals or nations, could never serve a constructive purpose. In 1898, when the United States engaged in the Spanish-American War, she realized that she would have to become either 'more or less a Tolstoyan' right away (Knight, 2005).

Addams eagerly read the thoughts of other influential authors on peace as well. In her second book, *Newer Ideals of Peace* (1907), she quoted L.T. Hobhouse's view that "universal and permanent peace" was not a vision but "an actual process of history." Both Addams and Hobhouse put their faith in the power of the growing social forces that in time could create the conditions of permanent peace. At the same time she agreed with British anti-imperialist economist John A. Hobson that this process could be derailed by the forces of economic imperialism (Addams, 1907).

She was soon a leader in the National Peace Congress, during a time when the men of high status were also active, but Addams's commitment, unlike theirs, did not change when World War I broke out. Elected in early 1915 as president of the Woman's Peace Party, she attended the International Congress of Women in the Netherlands and was elected president of the new organization formed there, the International Committee of Women for Permanent Peace, later renamed the Women's International League for Peace and Freedom (WILPF). She would remain its president until 1929.

During her presidency she built the organization into a forceful advocacy group for world disarmament, the spread of international law, and for pacifism. In the United States WILPF

and Addams were vilified for these supposedly dangerous views; from 1915 and continuing through the 1920s, she was mostly hated in her own country. The hate subsided somewhat when, in 1931, she was awarded the Nobel Peace Prize. Upon her death in 1935, she was freshly honored as the nation's leading citizen.

For much of the second half of the twentieth century few scholars recognized Addams as a significant intellectual. Indeed, when her ideas were considered she was assumed to hold the views typical of white, prosperous, reform women of her times regarding the inferiority of certain races, the superiority of the prosperous classes, and women's biologically dictated nature. Yet no one familiar with her body of work would make such a claim (Knight, 2005, 2010; Deegan, 2007).

Her Intellectual Development

The Early Years

As a child and young woman, Addams was fascinated with reformers and their ideas: Robert Owens's failed cooperative community, Lucy Stone's campaign for legal equality for women, and John Brown's impatience to see slavery end. She decided medicine offered the best way she could be helpful to poor people. The curriculum at the female seminary had little to offer her in the way of science but she and a friend, Ellen Gates Starr, founded a Science Club and she read Darwin's *On the Origin of the Species* and *The Descent of Man*.

After college, with more time to read, she read not only novels and history but also early works in social science that probed the nature of society and how it might be improved. Her reading list included Jean Baptiste Godin's *Social Solutions*, Ferdinand Tonnies's *Community and Society*, and Herbert Spencer's *Social Statics*. She also became an enthusiast for Auguste Comte's positivism in these years, particularly his idea of 'unity' – the belief that most people desired to feel a universal affection for other human beings.

Once at Hull House, Addams's interests came to include investigations. Already an admirer of Charles Booth's *Life and Labour of the People in London* (1889), she respected the power of statistics to capture the reality of conditions. The residents of Hull House, including Addams, soon became involved in various studies of problems facing their neighbors, including housing conditions, infant mortality rates, and the social value of saloons. Addams describes some of these in her two books about Hull House (Addams, 1910, 1930).

Hull House Maps and Papers

Most of these investigations were known only to Chicagoans, but in 1895, Hull House gained a national reputation for the publication of a book whose authors were officially listed as the 'Residents of Hull House.' The book's long title was *Hull-House Maps and Papers: A Presentation of Nationalities and Wages in a Congested District of Chicago, Together with Comments and Essays on Problems Growing Out of the Social Conditions* (Residents, 1895).

The book's significance has more often been assumed than assessed. For example, it is sometimes credited with setting new standards for social investigation. In fact the work on display was typical of the standards of the early 1890s, as a glance at the 1892 transactions of the American Social Science Association makes clear. And although *Maps* is sometimes described as having pioneered in its interest in women and children, only one of the 10 essays focuses on people in either category.

The maps in the book were a methodological breakthrough, but not an original concept since they were consciously modeled after Charles Booth's wage maps of London, including using the same color codes for wage levels. Deegan argues that the Hull House maps had a different purpose than those of adopted later by Chicago sociologists. The settlement maps were meant to be useful to the community, she claims, while the scholars' maps were for experts (Deegan, 1988).

Maps and Papers is often believed to reflect Addams's social thought, but it actually more reflects the thought of Florence Kelley, an experienced investigator who arrived at Hull House in 1892. The maps, the schedules, and two essays in the book all grew out of Kelley's work for the U.S. Department of Labor collecting Chicago data for its special report on wages, residential patterns, and nationalities of those living in the 'slums' of the nation's four largest cities (Deegan, 1988; Ross, 1998; Sklar, 1998).

The idea of publishing color-coded maps came first and then the residents decided to include expanded versions of five papers that some of them had presented at the Congress on Social Settlements, held at Hull House in conjunction with the 1893 Columbian Exhibition. The five papers included the two by Florence Kelley, on the sweating system and on wage-earning children. The other three were by Addams (on the settlement and the labor movement), Ellen Gates Starr (on art and labor), and Julia Lathrop (on county charity work). Rounding out the collection were five essays written by residents expressly for the book. These were on the maps project, the income and expenses of cloakmakers in Chicago, and three ethnic groups – the Jewish, the Bohemian, and the Italian communities – in Chicago.

Addams was proud of the book, the first with which she was associated, and disappointed when there was not a second printing, but she also did not want *Maps* to mislead the public regarding the main purposes of Hull House. In the Prefatory Note, she carefully stated that the energies of the Hull House residents "have been chiefly directed, not toward sociological investigation, but to constructive work" (Addams, 1895). She also noted that the maps and papers were based on "recorded observations [that were] … the result of long acquaintance" (Addams, 1895). She did not want anyone to think that the residents of Hull House saw their neighbors merely as objects of investigation from whom to extract data.

While *Maps* was not as pioneering as some have assumed, it was possibly more influential than earlier sociological studies because of its association with the famous Jane Addams, the famous Hull House and the important city of Chicago. Deegan (1988) argues that it launched the Chicago school of sociology, because of its use of mapping, its focus on some of the topics that Chicago sociologists would study for decades to come, and because it was about Chicago.

The book had a direct influence on another important study, W.E.B. DuBois's 1899 investigation, *The Philadelphia Negro*. DuBois's assistant for the project, who became an important contributor as well, was Isabel Eaton. She had been

a resident at Hull House. Addams recommended her to DuBois for the position, and Eaton was herself a contributor to *Maps*.

Jane Addams's Epistemology

Because Addams was first and foremost a reformer, her orientation toward knowledge was practical and placed a great value on experience. But she was also an intellectual. Compelled by the power of ideas to change the world, she spent a great deal of time in the company of other intellectuals, both socially and through reading. Thus she was both drawn to the academy and yet had a detached view of scholarly goals and methods.

In 1896 she published her first academic essay based on what she and others had learned from personal experience. 'A Belated Industry,' appearing in the first issue of the *American Journal of Sociology (AJS)*, was an expanded version of a speech on household labor she gave at the Chicago World's Fair of 1893 (Addams, 1896). She revised it again when making it a chapter of her first book, *Democracy and Social Ethics* (Addams, 1902).

Her second scholarly essay, "Trade Unions and Public Duty," was also based on experience. Published in 1899 in the *AJS*, it was a revised version of a paper she gave at the Congress of Social Settlements, combined with parts of a speech she gave about the 1894 Pullman Strike (Addams, 1899a).

In her third academic essay, Addams changed gears and wrote an original piece for a scholarly readership that tackled the question of the place of research in settlement houses and concluded that it did not belong there. Published in 1899 in the *Annals of Political and Social Sciences*, 'A Function of the Social Settlement' set forth Addams's view that the university's role was to generate and disseminate new knowledge, and the settlement house's role was to generate and improve it by experimentation (Addams, 1899b). (The interesting context for her argument is that in 1895 the University of Chicago proposed to make Hull House a part of the university, but Addams rejected the president's offer (Deegan, 1988).) But tucked behind this neat distinction is Addams's real and daring point: that knowledge is pointless unless the search for it is fully engaged with human experience from beginning to end.

Aware of the feathers she would ruffle with this essay, Addams sought protection by citing John Dewey and William James as agreeing with her, but that did not mean that either was her source for this idea. She had first encountered it in the writings of John Ruskin, Thomas Carlyle, and Matthew Arnold, and in the activist bent of the Social Christians (Knight, 2005).

In 'A Function of the Social Settlement,' Addams clearly distances herself (and the settlement house movement) from dry, abstract, specialized knowledge. The settlement, she wrote, "stands for application as opposed to research; for emotion as opposed to abstraction, for universal interest as opposed to specialization." Settlements were created "not with the object of finding clinical material, not to found 'sociological laboratories,' not, indeed, with the analytical motive at all, but rather … with a desire to use [the knowledge they possess] synthetically and directly, to test its validity and to discover the conditions under which this knowledge may be employed" (Addams, 1899b). Though she stated her position as representing the settlement movement, there were many, including her friend settlement house founder Robert Woods, who endorsed the concept of settlements as sociological laboratories (Lengermann and Niebrugge-Brantley, 2002). Addams was staking out her position in the debate.

Addams addressed scholars in person at the 1905 convocation at the University of Chicago in a speech titled, "Recent Immigration: A Field Neglected by the Scholar." Bearding the lion in its den, Addams charged that immigration scholars were relying on old views and traditions and therefore emphasizing false dangers from immigration – the destruction of the American way of life, etc. – rather than bringing their interpretations in line with the nation's 'present motley and cosmopolitan character.' Thus they were neglecting research's 'higher function': to resolve the contradictions between traditional beliefs and new realities. In seeing immigration as a threat, she wrote, "we are remaining constant to a truth we no longer wholly believe … and that is not justified by our latest information." (Addams, 1905).

This passage captures not only Addams's point that scholars tended to reflect the social consensus rather than challenge it, but also the passion she shared with scholars for the truth. At the same time, she was highlighting a problem that scholars rarely considered: the human tendency to resist changing one's mind. As she wrote in 1907, "[I]t is so easy to commit irreparable blunders because we fail to correct our theories by our changing experience" (Addams, 1907).

Addams also differed from most scholars in her method for discovering truth: she thought it could only be found in experience (a category that included observations and factual data but not abstract logical arguments). Addams's friend, George Mead, reviewing her book *Newer Ideals of Peace* in the *AJS* in 1907 caught this aspect of her epistemology when he astutely noted that she was arguing "not [for] the necessity of a deduction, but [for] the necessity of reality."

Addams cited Dewey and James to substantiate her pragmatism, but recent scholarship reveals that she worked out her version of a pragmatist philosophy earlier than either of the two famous male philosophers and that she was an important early influence on Dewey. In particular, in her thinking about democracy and about war, Addams was a pragmatist before Dewey, who did not reject moral absolutism completely until after World War I (Knight, 2005; Ross, 1998; Seigfried, 2011; Knight, 2014). Addams and James agreed on the need for a moral equivalent to war but it was Addams who turned this idea into a whole book, *Newer Ideals of Peace* (discussed below).

Addams believed interpretation based on experience was the essential method for arriving at the knowledge needed for effective reform. This involved both an ability to "interpret sympathetically the motives and ideals [of people whose] rules of conduct … [are] widely different from our own" (Addams, 1902) and an ability to revise one's own beliefs in response. She called that "the hard work of careful … self-examination." (Addams, 1910). Historian of science Dorothy Ross argues insightfully that Addams offered 'an interpretative form of social knowledge' and that it was 'relational in a double sense,' in that it was both based on her own experience in human relations and on her interest in the role of relational experience in understanding society (Ross, 1998).

Jane Addams and the Social Sciences

Jane Addams and Sociology

While Addams drew a bright line between the work of the settlement house and that of the university, she considered academics, particularly those in sociology, to be her professional colleagues. She was a founding member of the American Sociological Society (later the American Sociological Association) in 1905, and presented papers at its meetings. She was friends with most of the sociologists at Chicago, including Albion Small, George H. Mead, George Vincent, Charles Zeublin, Charles Henderson, and W.I. Thomas. In 1913, Small proposed that she take a half-time position in the department teaching graduate students, but she declined (Deegan, 1988). Just how compatible her ideas were with the men of the Chicago school is a separate question. Deegan emphasizes the collegial ties, and addresses ideas secondarily. Dorothy Ross highlights Addams's discomfort with their cautious devotion to the status quo (Ross, 1998).

Addams was influenced by and influenced the field. She had ties with W.E.B. DuBois, Richard T. Ely, Franklin Giddings, and Charles H. Cooley (Deegan, 1988); the syllabus for her 1898 University of Chicago extension course, 'Democracy and Social Ethics,' includes writings by scholar/friends Amos G. Warner, Sidney Webb, and Thorstein Veblen (Addams, 1898b). She was particularly interested in Veblen's notion of the instinct of workmanship (Addams, 1922). Her books were reviewed in the *American Journal of Sociology* and professors adopted her texts in courses (Deegan, 1988). But sociology at that time was pressing for increased scientific objectivity and knowledge for its own sake, which inevitably reduced Addams's influence, given her interest in knowledge's interpretation and application (in addition to the fact that she was female in a male field).

The sociologist whose ideas seemed closest to her own and whom she certainly influenced was Charles Cooley. He cited her work seven times in his 1909 text, *Social Organization*, where he argued for the central role of primary groups (family, friendship groups, etc.) as the source of one's ethics, a view that Addams had long held. Cooley's methods were also Addams's: case studies and empirical observation. Cooley also agreed that society was an experiment that led to enlarging the individual's social experience and to finding new ways to coordinate variety, points Addams first made in the 1890s (Deegan, 1988; Ross, 1998).

Addams wrote one sociology text for the classroom, *Newer Ideals of Peace* (1907). Her argument was that the old-fashioned ethic of militarism (which, drawing from Herbert Spencer, she defined as not just values related to war but the reliance on physical force as the solution to every difficulty and the embrace of hierarchal rankings in society) was too influential in current thinking about government and its roles, and was the cause of government's failures. Militarism should therefore be replaced by the new ethic of humanitarianism which would allow democratic governments to thrive. While her examples were mostly drawn from the United States, the book hints throughout that she had the world as much as her own nation in mind.

Beginning in the late twentieth century, and due to the important efforts of Mary Jo Deegan, Dorothy Ross, Patricia Madoo Lengermann, and Jill Niebrugge-Brantley to integrate early women sociologists into the history of the field, Addams began to receive recognition as an important sociologist, though her views remain controversial, as Ross, Lengermann, and Niebrugge-Brantley emphasize.

The field of social psychology was Addams's particular interest. Her library included the writings of Mary Follett, Normal Angell, and Graham Wallas. And her interest in social relations and the changing status of women are remarkably mirrored in the writings of Lester Frank Ward. John Dewey and William James were also leading social psychologists and both men viewed Addams's ideas with profound respect (Knight, 2010). In the early 1890s, well before James gave his lectures on the "varieties of religious experience," Addams wondered at the varieties of religious experience found in her immigrant, working class neighborhood, and pondered the role of emotions and experience in religious thought (Knight, 2005).

Jane Addams and Anthropology

While Addams's associations and intellectual connections with sociology are recognized, those with anthropology are not. Her intellectual interests were profoundly anthropological; she was a sociocultural anthropologist using methods of participant observation. Her neighborhood was home to immigrants of many different nationalities, many of whom came to Chicago as peasants. She sympathetically observed their encounters with modern, industrial, urban culture and wrote eloquently about their tensions with their children over dress, financial and social independence, and sexuality.

Although she used the terms 'primitive' and 'civilized,' she used them in ways that only careful scholarship can unpack. One of her most radical arguments was that those whom society called 'primitive' were actually the most civilized when it came to their sense of human brotherhood, and that they had much to teach those whom society called 'civilized' (i.e., Anglo-Saxons) (Knight, 2005). One of her most anthropological books was her study, *Spirit of Youth and City Streets* (1909), a sensitive exploration of the ways that youth, with their fundamental instincts for play, adventure, sex, and imagination, responded to the city.

She was also a close student of women and their ways of claiming power in a patriarchical society. Her essay, "A Modern Devil Baby," examined how an ancient fairy tale flamed to life in her neighborhood. Sheprobed how women wielded the rumor of the devil baby's presence at Hull House as a way to intimidate men into being better fathers. That essay was revised to become two chapters in *The Long Road of Woman's Memory* (Addams, 1916), an insightful study of how women's memories functioned to transmute the past, reinterpret the present, disturb conventions, integrate women with modern industrial life, and challenge war.

The works of anthropologists shaped her views on the distinctive place of women in early human history and her belief that the role of government should be based on the family's nurturing role. Otis Mason's *Woman's Share of Primitive Culture* (1894) supplied her with arguments for women as the founders of the family, which he called "parental government," and therefore for women being qualified to engage in modern government. In her unpublished 1911 speech, "The Woman and the State," Addams drew on Mason's and other anthropologists

ideas to argue that centuries of men's dominance in governance had created a situation harmful to human life (Addams, 1911; Knight, 2010).

So far Addams's ideas have apparently left no trace in anthropological scholarship though one groundbreaking anthropologist, Ashley Montagu, has called her "a genius," and her writings, "great literature" (Montagu, 1960). It remains for scholars to study Addams as a sociocultural anthropologist.

Jane Addams and Political Science

Addams was a moral philosopher whose core beliefs about human beings centrally shaped her understanding of how democratic politics worked and how it could potentially work. Her views about political democracy included ideas about citizenship, which she considered a responsibility shared by everyone in a democracy, and about the purposes of government, which she believed were to remove obstacles that prevented citizens from developing their full potential. Much of her work at Hull House and as a political activist was shaped by these convictions.

Her interest in morality in politics was partly compelled by the fact that Hull House was located in a ward represented by Chicago's most powerful and most corrupt alderman, John Powers. The neglect of the neighborhood's municipal services, particularly garbage removal, threatened the health of the ward. Addams took seriously the fact that Powers was popular with his voters and sought to understand it. Her resulting brilliant essay, "Ethical Survivals in Municipal Corruption," is a masterful interpretation of democracy in a mostly immigrant urban ward. Rather than condemn Powers for being immoral, she explored how the ethics of the voters found resonance in Powers' political strategies.

Addams's political actions as much as her writing on these subjects provide a window into her political thought. She was a political activist during the Progressive Era, lobbying in support of child labor, immigrant, and suffrage legislation. In 1889, she had felt political democracy was inaccessible to her as a woman but by 1912 she was fully engaged. In that year she seconded the presidential nomination of Theodore Roosevelt, campaigned for him, and became a leader herself of the Progressive Party. She served as a member of its executive committee for at least 4 years (Knight, 2010).

Although Addams wrote a great deal on the subject of democracy, political philosophers have only recently begun to study her ideas rigorously. This is challenging to do since she does not present her ideas systematically or concisely; important thoughts are strewn about her writings.

Political scientists have been drawn to study her book *Democracy and Social Ethics*, because the title includes the word 'democracy.' The book accurately captures her thinking as it evolved during her first 10 years at Hull House. It does not reflect her more mature understanding, gained from her expanding activities as a political activist. That is captured in her second book, *Newer Ideals of Peace* (1907). The title is misleading, since her subject is really the trend toward, and need for, more democratic government (see above discussion of the book).

Another excellent source for Addams's democratic thinking is a book she published 5 years before she died, *Second Twenty Years at Hull-House* (1930). Long out of print because the family controls the copyright, it is rarely cited but includes insightful chapters about social service and the progressive party, the woman suffrage movement, prohibition, and federal immigration policy.

Political scientists have recently probed Addams's ideas on many subjects. These include democratic citizenship, cosmopolitan citizenship, gendered and ungendered citizenship, and the ethic of care (Nackenoff, 2009; Sarvasy, 2009, 2010; Hanagan, 2013; Tronto, 1993). Though Addams is generally assumed to be a liberal, her ideas challenge many liberal assumptions. In particular, she rejected the individualistic ethic that undergirds liberalism (Knight, 2009).

Addams and Social Work

It is often said that Jane Addams was a founder of the field of social work. This is for several reasons. Settlement leaders, though not Addams, partly created the profession, founding one of the first schools of social work, as they were later called. (It should be noted that leaders of Charity Organization Societies (COS) had a greater influence over the emerging profession. COS leaders in the 1890s were devotees of a highly individualistic, classically liberal philosophy of human self-sufficiency that was diametrically opposed to the settlement house philosophy.) Also, Addams was the most famous and admired settlement house leader, making it natural to consider her a founder of the profession that attracted the interest of settlement movement enthusiasts.

Nonetheless, if founders should have a major influence in the fields they found, then it is difficult to see Addams as a founder. Perhaps the most salient disagreement between Addams and the new social work profession was her discomfort with the idea of professional training for settlement work. She saw herself and the residents as volunteers (being a resident at Hull House was never a paid position during Addams's lifetime, though it was at other settlements) who learned from experience while living in the neighborhood and being good neighbors. Furthermore, social work as a field came to be primarily associated with individual clinical work, which never appealed to Addams as a strategy for transforming lives or improving society. Addams was at core a reformer, a political activist, and she thought it a natural process that settlement volunteers would become engaged in social change around policy issues. This has generally not been the strength of the social work profession, especially once government funding became crucial to the field.

While Addams is credited as a founder of the social work profession and while scholars in the profession have written essays tracing her accomplishments and urging a return to her democratic ideals, Addams is too often viewed as a much beloved historical figure rather than what she also was – an intellectual and reformer with unorthodox ideas.

Conclusion

In the early twenty-first century, women intellectuals have begun to be acknowledged as contributors to the various social and behavioral sciences. It can therefore be hoped that

the past neglect of Addams's ideas in these various fields is coming to an end. The scholars in sociology and political science have led the way. And if, as it is sometimes thought, the social sciences in general are making a turn toward the qualitative, and away from the quantitative, Addams's methods and interest in how experiences and ideas shape human behavior should be of particular interest. Like her friends John Dewey and William James, Jane Addams had large and original ideas that will continue to reward scholarly investigation.

See also: Biography: Historical; Citizenship, Historical Development of; Democracy, History of; Gender History; Intellectual History; Labor History; Social Ecology: The Chicago School; Social History; Urban History.

Bibliography

Addams, J., 1895. Prefatory note. In: Residents of Hull House, Hull House Maps and Papers: A Presentation of Nationalities and Wages in a Congested District of Chicago, Together with Comments and Essays on Problems Growing Out of the Social Conditions. Crowell, New York, pp. vii–viii.

Addams, J., March 1896. A belated industry. American Journal of Sociology 1, 536–550.

Addams, J., April 1898a. Ethic survivals in municipal corruption. International Journal of Ethics 8, 273–291.

Addams, J., 1898b. Democracy and Social Ethics (Syllabus). Anita, M. Blaine Papers. Wisconsin State Historical Society, Madison.

Addams, J., January 1899a. Trade unions and public duty. American Journal of Sociology 4, 448–462.

Addams, J., 1899b. A function of the social settlement. American Academy of Political and Social Science 13, 33–55 (323–345).

Addams, J., 1902. Democracy and Social Ethics. Macmillan, New York.

Addams, J., 1905. Recent immigration: a field neglected by the scholar. University Record 9 (9), 274–294.

Addams, J., 1907. Newer Ideals of Peace. Macmillan, New York.

Addams, J., 1909. The Spirit of Youth and City Streets. Macmillan, New York.

Addams, J., 1910. Twenty Years at Hull House. Macmillan, New York.

Addams, J., 1911. Woman and the state. Speech to the New York Woman's Political Union, 2.2.1911. Jane Addams Papers Microfilm, Reel 7, frame 0096.

Addams, J., 1916. The Long Road of Woman's Memory. Macmillan, New York.

Addams, J., 1922. Peace and Bread in Times of War. Macmillan, New York.

Addams, J., 1930. The Second Twenty Years at Hull House. Macmillan, New York.

Brieland, D., March 1990. The Hull-House tradition and the contemporary social worker: was Jane Addams really a social worker? Social Work 35, 134–148.

Daynes, G., Longo, N.V., 2004. Jane Addams and the origins of service-learning practice in the United States. Michigan Journal of Community Service Learning 11 (1), 5–13.

Deegan, M.J., 1988. Jane Addams and the Men of the Chicago School, 1892–1918. Transaction Books, New Brunswick.

Deegan, M.J., 2007. Jane Addams. In: Scott, J. (Ed.), Fifty Key Sociologists: The Formative Theorists. Routledge, London, pp. 3–8.

Elshtain, J.B., 2002. Jane Addams and the Dream of American Democracy. Basic Books, New York.

Fischer, M., 2009. The conceptual scaffolding of newer ideals in peace. In: Fisher, M., Nackenoff, C., Chmielewski, W. (Eds.), Jane Addams and the Practice of Democracy. University of Illinois Press, Champaign, pp. 165–182.

Hanagan, N., 2013. Democratizing responsibility: Jane Addams's pragmatic ethics. Polity 45 (3), 347–371.

Knight, L., 2005. Citizen: Jane Addams and the Struggle for Democracy. University of Chicago Press, Chicago.

Knight, L., 2009. Jane Addams's theory of cooperation. In: Fisher, M., Nackenoff, C., Chmielewski, W. (Eds.), Jane Addams and the Practice of Democracy. University of Illinois Press, Champaign, pp. 65–97.

Knight, L., 2010. Jane Addams: Spirit in Action. W. W. Norton, New York.

Knight, L., 2014. Jane Addams and John Dewey debate war. In: Clark, G., Jackson, B. (Eds.), Trained Capacities: John Dewey, Rhetoric and Democratic Culture. University of North Carolina Press, Chapel Hill.

Lengermann, P.M., Niebrugge-Brantley, J., 2002. Back to the future: settlement sociology, 1885–1930. The American Sociologist 33 (3), 5–20.

Montagu, A., 1960. Introduction to position of women section. In: Cooper, E. (Ed.), Jane Addams: A Centennial Reader. Macmillan, New York, pp. 100–102.

Nackenoff, C., 2009. New politics for new selves: Jane Addams's legacy for democratic citizenship in the twenty-first century. In: Fisher, M., Nackenoff, C., Chmielewski, W. (Eds.), Jane Addams and the Practice of Democracy. University of Illinois Press, Urbana, pp. 119–142.

Residents of Hull House, 1895. Hull-House Maps and Papers: A Presentation of Nationalities and Wages in a Congested District of Chicago, Together with Comments and Essays on Problems Growing Out of the Social Conditions. Crowell, New York.

Rosiek, J.L., Pratt, S., 2013. Jane Addams as a resource for developing a reflexively realist social science practice. Qualitative Inquiry 19 (8), 578–588.

Ross, D., 1998. Gendered social knowledge: domestic discourse, Jane Addams, and the possibilities of social science. In: Silverberg, H. (Ed.), Gender and American Social Science: The Formative Years. Princeton University Press, Princeton, pp. 235–264.

Sarvasy, W., 2009. A global 'common table': Jane Addams's theory of democratic cosmopolitanism and world social citizenship. In: Fisher, M., Nackenoff, C., Chmielewski, W. (Eds.), Jane Addams and the Practice of Democracy. University of Illinois Press, Urbana, pp. 183–202.

Sarvasy, W., 2010. Engendering democracy by socializing it: Jane Addams's contribution to feminist political theorizing. In: Hamington, M. (Ed.), Feminist Interpretations of Jane Addams. Pennsylvania State University Press, University Park, pp. 293–310.

Schneiderhan, E., 2011. Pragmatism and empirical sociology: the case of Jane Addams and Hull-House, 1889–1895. Theory and Society 40, 589–617.

Seigfried, C.H., 2011. Democracy as a Way of Life: Addams' Pragmatist Influence on Dewey, Society for the Advancement of American Philosophy Conference, 12 March 2011. http://american-philosophy.org/events/saap_2011_program.htm (accessed 13.10.14.).

Sklar, K.K., 1998. Hull House maps and papers: social science as women's work in the 1890s. In: Silverberg, H. (Ed.), Gender and American Social Science: The Formative Years. Princeton University Press, Princeton, pp. 127–155.

Tronto, J., 1993. Moral Boundaries: A Political Argument for an Ethic of Care. Routledge, New York.

Winkelman, J., 2013. A working democracy: Jane Addams on the meaning of work. The Review of Politics 75, 357–382.

Addictions: General Considerations

Ulrich John, University Medicine Greifswald, Greifswald, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A basic understanding of addiction is given. It includes addictive elements and disorders. Addictive elements are the addictive stimulus, behavior, conditions, and reward. Interactions of these elements may lead to addictive and further disorders. Addictive disorders include substance-related and non-substance-related disorders. Further disorders are related to addiction in that they depend on addiction elements but do not necessarily fulfill criteria for addictive disorders. Addiction has relevance for public health since highly prevalent risk behaviors such as tobacco smoking and overeating are involved. It is concluded that addiction is important in understanding and planning of interventions in public health.

Background

Importance

Addiction is a phenomenon that has been demonstrated in humans and animals. It may devastate social ties, behavior, and health. Starting from a historical perspective a basic understanding is presented for substance-related addictive disorders (SRD) and non-substance-related addictive disorders (NSRD). A public health perspective of addiction will be considered before conclusions are drawn.

Addiction at first glance seems to be a psychic phenomenon that increases and maintains 'bad behavior' despite known negative consequences. These include impaired social relations, health risk behaviors, as well as addictive and other chronic health disorders. There is international agreement that chronic disease largely depends on behavior (WHO, 2011). Behavior has been shown to be the single most influential among five domains that determine our health (McGinnis et al., 2002). Worldwide, among death cases, 10.5% have been attributed to tobacco smoking, 4.9% to overweight or obesity due to high caloric intake, and another 4.0% to unhealthy alcohol consumption (WHO, 2011). These behaviors compared to other factors are connected with the highest risk of global burden of disease (Lim et al., 2012). Smoking is one of the risk factors that are associated with the highest proportion of cardiovascular deaths, and smoking itself is related to the highest proportion of cancer death cases in the United States (Danaei et al., 2009). While tobacco smoking remained to be the single most preventable disease risk, there is an increasing overweight epidemic with a caloric intake that is too high and with a sedentary lifestyle (Finucane et al., 2011). Given this evidence, two questions arise: first, why do so many people adopt one or more of these health risk behaviors? Second, why do abundant smokers, alcohol consumers, and overeaters not change that behavior, although they may know about the negative consequence such as family or job loss, detriments on health, or premature death? One main part of answers to these questions is addiction. Addiction is a driving force in the adoption and maintenance of health risk behaviors.

History

Addiction is predominant in a number of terms that in large parts share the same meaning, from inebriety over addiction and dependence to addictive disorders (APA, 2013; Edwards, 2012). In 1964, the World Health Organization came up with the term dependence. In 1976, Edwards and Gross published the alcohol dependence syndrome (Edwards, 2012). It was a tremendous progress since it improved international standardization of diagnoses and provided a definition for research and clinical practice. The alcohol dependence syndrome helped to protect alcohol-dependent persons from moral disgust by the understanding of alcohol dependence as a disease. International classification systems of diseases, the Diagnostic and Statistical Manual of the American Psychiatric Association (DSM), and the International Classification of Diseases of the World Health Organization provided criteria and number of criteria required to diagnose alcohol dependence or abuse. The criteria were based on evidence and consensus of experts in the field.

In the most recent version of the DSM, DSM-5, new progress has been made by combining 'dependence' and 'abuse' into the category 'addictive disorders' that now includes SRD and NSRD. Addiction has been put more into focus in the frame of addictive disorders than was the case before (APA, 2013). The terms 'substance dependence' and 'abuse' have been dismissed. However, the criteria remained for SRD and have been adapted for one single NSRD in DSM-5, i.e., pathological gambling (APA, 2013).

What Is addiction?

Basic Understanding

Addiction is determined by social, behavioral, and physical factors. Addiction may be described by elements and disorders. There are four addiction elements: the external stimulus, the person's behavior, the conditions, and the reward. The elements are active in the formation and progress of disorders. Disorders include addictive and further disorders. Addictive disorders are SRD and NSRD. Further disorders, although not part of SRD or NSRD, are related to addictive elements (**Figure 1**).

Elements

Stimulus

A stimulus is a psychotropic drug or a nondrug stimulus. Psychotropic drugs include tobacco, alcohol, illicit drugs, and

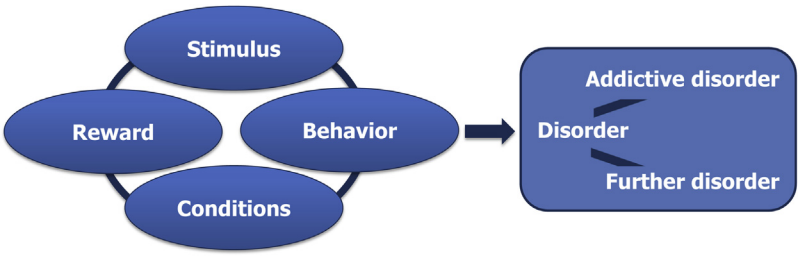

Figure 1 Addiction: elements and disorders.

psychoactive prescription drugs. Psychotropic or psychoactive means that the substance acts on the central nervous system and has the potential to change mood, perception, cognition, or behavior. Clear evidence exists that only very few, approximately 100, out of about 30 million chemical substances, are known to be addictive in the sense that they are estimated as rewarding or pleasurable (Gardner, 2011). To alter mood, a substance requires to produce stimulation or sedation and antinociceptive or pronociceptive effects. Such effects, in co-occurrence with other specific elements, particularly specific conditions, may be experienced as feeling delighted or depressed. In DSM-5, a total of 10 classes of psychoactive drugs are addressed: alcohol; cannabis; hallucinogens; inhalants; opioids, sedatives or hypnotics, or anxiolytics; stimulants; tobacco; other substances; and caffeine. Caffeine differs from the other substances according to disorder groups. Non-substance-related stimuli among others may be games, goods, and the Internet.

Behavior

The following behaviors may be subsumed under the term 'addictive behaviors,' although there may be additional ones: intake of psychoactive drugs, eating food, using games with money, video games, Internet, buying goods, or sexual activities. These behaviors are needed for the development and maintenance of addiction.

A behavior that may become addictive such as gambling often is felt as pleasurable in early stages of practice, and as such the behavior is goal directed. Addictive behaviors develop over time. This process includes, among others, habit formation. Habits are largely automatized actions. Automaticity depends on the response of the individual and the context (Wood and Neal, 2007). Contexts may activate the performance of a response (Wood and Neal, 2007). Habits develop on grounds of associative learning (Wood and Neal, 2007). For the understanding of habits, one important question is how automatized processes develop a behavior that has originally been goal directed. In animal studies, drug use has been shown to increasingly become habitual and inflexible (Pierce and Vanderschuren, 2010). As part of this process, additional brain structures become involved (Pierce and Vanderschuren, 2010). However, addiction is a more complex phenomenon than habit formation. It needs supportive conditions.

Conditions

Conditions include social conditions and personality characteristics. Social conditions are factors at different societal or group level that may have an influence on using psychoactive substances or exerting addictive behaviors. At a societal level, ease of access to the stimulus and stimulus-related social norms are important factors. Stimuli such as cigarettes, alcoholic beverages, or gambling halls may be easily available but may also be restricted in access, as is the case by regulation in the framework of prevention. Social norms about drug use or addictive behaviors may influence the behavior of the individual, both at the social and individual level. The social network of a person may influence the behavior, particularly as long as a disorder has not yet developed. Peer groups may have a strong influence on whether a novice (e.g., in smoking tobacco or cannabis) will judge her or his first use experience as positive or negative. This may determine whether or not the stimulus will be accessed further on. In later stages of use trajectories, social marginalization may occur, as was documented in subpopulations of addicted alcohol consumers.

Personality characteristics include impulsiveness. Impulsiveness or impulse control disorders are one main underlying personality condition of addiction (von Ranson et al., 2013). Impulsiveness has been suggested by data to contribute to the risk of SRD, eating disorders, and gambling disorders (e.g., von Ranson et al., 2013; Karim and Chaudhri, 2012). Evidence suggests that an overlap exists between brain systems that are involved in SRD and brain systems that are involved in food intake (Dagher, 2009). One proposition maintains that palatable food may become addictive in the run of a binge pattern of consumption (Corwin and Grigson, 2009). In animal experiments, bingeing on food was associated with behavioral and neuronal changes that resemble those in drug addiction (Corwin and Grigson, 2009).

Reward

After having performed addictive behavior, reward may elicit the repetition of behavior. This addiction cycle is activated recurrently. The individual may perceive pleasant feelings or relief from unpleasant states, such as aches or agitation, as rewarding. A brain reward system has been shown by neurobiological research (Gardner, 2011). Specific brain areas are involved in self-administration of drugs. To this system belongs the nucleus accumbens and the dopaminergic system (Gardner, 2011).

Reward may be induced by craving for the stimulus and then performing the behavior. Craving has been understood as an urge to consume the psychoactive substance when being abstinent (Skinner and Aubin, 2010). This might also be true for non-substance-related stimuli. Craving may elicit

a motivation to consume the substance or exhibit the desired behavior. Craving is one of the diagnostic criteria of SRD according to DSM-5 (APA, 2013).

Reward and craving are the core underlying processes in one of the most common models of urge to smoke. It is assessed by the Fagerstrom Test for Nicotine Dependence (FTND: Heatherton et al., 1991). The FTND is based on the fact that nicotine is metabolized over time and that an urge to supply the body with nicotine takes place. The urge to smoke varies in strength among smokers. To assess the urge to smoke, question 1 of the FTND seems particularly useful: How many minutes after awaking in the morning pass before the first cigarette is smoked? Further questions follow this rationale.

Disorders

Two kinds of health disorders are related to addiction elements: addictive disorders and further health disorders. Both groups of disorders may include disturbance of social relations, behaviors, as well as mental and physical functioning. Addictive disorders are SRD and NSRD. Further health disorders are those that are affected by addiction elements, irrespective of whether an addictive disorder does or does not exist.

Addictive disorders

SRD and NSRD are apparent at the social, behavioral, and neurobiological levels. According to DSM-5, SRD include substance use disorders (SUD), substance intoxication, and substance withdrawal (APA, 2013). These three disorder groups are valid for the 10 classes of psychoactive drugs in DSM-5 except caffeine. For caffeine, there is intoxication and withdrawal, but no SUD.

SUD and substance withdrawal need chronic application of the substance in an amount that is suited to add to the development of disorder. However, the amount may differ by individuals. SUD is diagnosed using 11 criteria according to DSM-5 (APA, 2013). As part of these criteria, social, behavioral, and neurobiological levels of addictive disorders are represented by social impairment, risky use behavior, tolerance, withdrawal, and impaired control.

Social impairment includes recurrent use of the substance, which results in 'a failure to fulfill major role obligations at work, school, or home' (criterion 5); continued use of the substance 'despite having persistent or recurrent social or interpersonal problems caused or exacerbated by' effects of the substance (criterion 6); and important 'social, occupational, or recreational activities' being given up or reduced in the course of substance use (criterion 7, APA, 2013: p. 491).

Risky use behavior includes current use of the substance under circumstances that include hazards to physical health (criterion 8) and continued use of the substance, although the consumer knows of recurrent physical or psychological problems that may result from substance use (criterion 9, APA, 2013).

Tolerance is understood as a need for markedly increased amounts of the substance to achieve the desired effects or a diminished effect of the same amount of the substance (criterion 10). To withdrawal belong withdrawal symptoms and use of the substance to relieve or avoid withdrawal symptoms (criterion 11). There are eight withdrawal symptoms: autonomic hyperactivity as noticed by sweating or pulse rate above 100 beats per min, tremor, insomnia, nausea, transient hallucinations or illusions, psychomotor agitation, anxiety, and seizures. At least two of these must be met for withdrawal of the substance hours or days after cessation of using it (APA, 2013).

Impaired control includes the substance being taken in larger amounts or longer than intended (criterion 1), a 'persistent desire or unsuccessful efforts to cut down or control' substance use (criterion 2), much time having been spent in obtaining the substance (criterion 3), and craving 'or a strong desire or urge to use' the substance (criterion 4, APA, 2013: p. 491).

NSRD are also known as behavioral addictions or process addictions. According to DSM-5, a diagnosis of gambling disorder is provided if four or more of nine criteria are fulfilled. They include symptoms similar to SRD. Social impairment is indicated by having lost, e.g., social relationships or jobs because of gambling (APA, 2013). Risky use might be indicated by 'chasing' one's losses, i.e., after having lost money returning to gamble again (APA, 2013: p. 585). Impaired control may be seen in needs "to gamble with increasing amounts of money in order to achieve the desired excitement" (APA, 2013: p. 585). Internet gaming disorder is mentioned in DSM-5 as a condition for further study (APA, 2013). Other addictive disorders have been disregarded in DSM-5 because of insufficient evidence.

SRD and NSRD share common features of addictive disorders. The individual accesses the addictive stimulus repeatedly with goal-directed and habitual addictive behaviors. One effect of this is that some temporary satiation occurs before the individual restarts to engage in the behavior. Attention of the individual is increasingly absorbed by the addictive stimulus. Other obligations or interests are neglected more and more in favor of being concerned about the addictive stimulus. Social relationships and mental and physical functioning may become disordered. Control over the stimulus decreases (Sussman and Sussman, 2011).

In detail, SRD and NSRD share 13 common features. Among them, nine are core phenomena of addiction, three are related to the development of addiction, and one is related to treatment. The nine core phenomena are, first, a "failure to resist an impulse, drive, or temptation to perform an act that is harmful to the person or to others" (Grant et al., 2010; APA, 2000: p. 663). Second, for part of the behavioral addictions, the individual "feels an increasing sense of tension or arousal before committing the act and then experiences pleasure, gratification, or relief at the time of committing the act" (APA, 2000: p. 663, Grant et al., 2010). Third, SRD and NSRD seem to have common cognitive (Grant et al., 2010) and, fourth, learning and memory processes involved (Karim and Chaudhri, 2012). Fifth, individuals with SRD and those with NSRD show high impulsiveness (Grant et al., 2010). Sixth, SRD and NSRD seem to share common neurobiological processes such as alterations of dopaminergic pathways (Grant et al., 2010; Karim and Chaudhri, 2012; Ashley and Boehlke, 2012). Seventh, genetic evidence supports that NSRD may share genetic grounds with SRD (Grant et al., 2010). Eighth, there is comorbidity between pathological gambling and SRD (Grant et al., 2010). SRD and NSRD are both comorbid to further mental disorders. Ninth, there seems to be evidence that SRD and NSRD become obvious after shorter time among females than among males (Grant et al., 2010).

Development of both SRD and NSRD includes a decrease in positive mood during repeated behavior or a need to increase behavior intensity (Grant et al., 2010). Second, increasing focusing on, and engagement in, the addictive behavior interferes with other life domains such as role obligations or interests (Grant et al., 2010). Third, trajectories of addictive disorders over the life course seem to be similar among SRD and NSRD, with a beginning at age below 20 in many cases. Furthermore, treatment use is similar among individuals as SRD and NSRD are often treated in a similar way.

Further health disorders

Detriments of health other than addictive disorders belong to addiction irrespective of whether an addictive disorder does exist or not. Further health disorders are related to addiction in that they depend on addiction elements. It may be the case that all four elements of addiction are present in a disorder that has not been subsumed under the category 'addictive disorders.' Obesity from eating more food than is needed by the body may be such a further health disorder. In addition to overeating, relevant behaviors are particularly tobacco smoking, alcohol risk drinking, and sedentary lifestyle. There is clear evidence of co-occurrence of addictive disorders with other disorders of mental and physical health. A variety of psychiatric or physical disorders such as depression or obesity may precede, accompany, or follow an addictive disorder. Main aspects of further disorders following from substance use are the amount of intake and the time since quitting substance use. Evidence suggests that underlying common causal processes might add to such disorders. Addiction elements in part may explain why people maintain behaviors that are known to be health detrimental.

Taken together, addiction includes public-health-relevant elements. Each of them alone and interactions among them may add to the development of such disorders. This understanding is more comprehensive than models according to which addiction itself is disease (American Society of Addiction Medicine, 2013).

Addiction Severity

Addiction severity is important for the understanding of causal relations between addiction elements and SRD, NSRD, or further disorders. In DSM-5, mild SRD except caffeine requires two to three; moderate, four to five; and severe, six or more of the 11 criteria (APA, 2013). In an alternative approach to assess addiction severity, self-statements of symptom frequency are used (John et al., 2003).

Addiction severity may further be assumed from the strength of the single addictive elements and interactions among them. Stimuli might vary in their intensity and differ according to their reinforcing properties. This indicates the addictive power of a stimulus. It might be determined on grounds of relapse, strength of craving, or efforts to access the stimulus. Individuals might differ in vulnerability to these criteria of addictive power. Behavior may differ according to characteristics such as repetitiveness. According to health disorder, one main behavior characteristic is dose, e.g., number of cigarettes smoked per day. Conditions might more or less support the fact that addiction and reward differ in strength of effects on individuals.

Public Health

Relevance of addiction for public health is given by the prevalence of disorders and by approaches to improve public health. Prevalence data are differentiated according to whether the disorder was currently present, i.e., during the last 12 months prior to data collection, or lifetime, i.e., at any time during life. Data that have been collected before DSM-5 were published. In several studies, DSM criteria for dependence or abuse were used, which in the majority are comparable with DSM-5. For the United States, results of three nationally representative surveys among adults are available: the National Comorbidity Survey Replication (NCS-R) (Kessler et al., 2004), the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) (Grant et al., 2004), and the National Survey on Drug Use and Health (Samhsa, 2011). NCS-R data revealed 1.3% current and 5.4% lifetime prevalence for alcohol dependence (Merikangas and McClair, 2012). Alcohol abuse or dependence was found for 8.5% among the general population currently and for 30.3% lifetime in NESARC, and for 4.4% currently and 18.6% lifetime in NCS-R. Drug abuse or dependence prevalence was 2.0% currently and 10.3% lifetime according to NESARC, and 1.8% currently and 10.9% lifetime according to NCS-R (Merikangas and McClair, 2012). The data collection differed in that the NCS-R used the Composite International Diagnostic Interview (CIDI) and the NESARC used the Alcohol Use Disorder and Associated Disabilities Interview Schedule. Also, criteria for SUD diagnosis differed (Merikangas and McClair, 2012). This difference is likely to be a reason for the different in prevalence estimates. Both interviews provide DSM, version IV, diagnoses. Prevalence estimates for residents aged 15 or older in the European Union countries revealed that current alcohol, opioid, and cannabis dependence were among the four most prevalent mental disorders (Wittchen et al., 2011). The CIDI methodology was used in a study for 17 countries from Africa, America, Asia, Europe, and New Zealand. The samples included 85 052 respondents for the face-to-face interview. Findings show lifetime prevalence rates for SUD from 1.3 to 15.0% (Kessler et al., 2007).

The data revealed differences according to gender and age of onset of alcohol consumption. The prevalence rates for current SUD (alcohol or illicit drugs) among males were found to be nearly double (11.9%) that among females (6.1%) in the US National Survey on Drug Use and Health (Merikangas and McClair, 2012). It included a general population sample of residents aged 12 or older. Among them, 8.7% had an SUD. Among those aged 21 or older, the subpopulation that had drunk alcohol at age 14 or less were more than 5 times as likely to have alcohol dependence or abuse compared to those who had their first alcohol consumed at age 21 or more(Samhsa, 2011). SUD are related to a variety of social factors including low education (Schulden et al., 2009).

Eleven addictive disorders, SRD and NSRD including eating disorders, have been considered for one prevalence estimate. According to it, 47% of the US adult population may have signs of addictive disorders (Sussman and Sussman, 2011). As part of this, overweight is a main problem from a public health perspective. Prevalence of overweight increased worldwide (Finucane et al., 2011). Eating disorders such as anorexia nervosa or bulimia have lower prevalence rates. Epidemiologic

research showed consistently high psychiatric comorbidity of SRD and NSRD, particularly with mood, anxiety, and personality disorders (Schulden et al., 2009).

Addiction informs approaches to improve public health. It seems self-evident that preventive efforts should be helpful in extending life expectancy and in reducing disability and low quality of life. Considerable success of public health approaches such as governmental action to control for the risk behavior was demonstrated (Lightwood and Glantz, 2013). Comprehensive programs that include a variety of single preventive measures are most likely to be successful. The single provisions in such programs seem to add to effects at the national level (Lightwood and Glantz, 2013).

One main restriction in preventive action using a comprehensive approach of action is that too few countries so far have spent considerable effort on it. In many countries, there are strong barriers against the implementation of preventive measures that have been proven in practice. However, there are also limitations of our knowledge. We do not know enough about how addictive disorders or further disorders act as barriers against change of addictive behavior. Specific intervention approaches that are most effective in a public health perspective are not sufficiently known yet. A promising preventive strategy is the implementation of specific measures, particularly tax increases, as demonstrated in case of smoking reduction (Lightwood and Glantz, 2013). Curbing tobacco smoking and eating disorders are of high priority.

Conclusion

Evidence from social, behavioral, and neurobiological fields clearly reveals that addiction exists. People exert risk behaviors against barriers and without being forced directly by anyone. Addiction may be characterized by elements and disorders that are interrelated at social and behavioral levels. Social, behavioral, and further aspects have been included in classification systems of diseases. Addiction is a useful element in understanding how to curb the health risks of psychoactive substance use and NSRD.

Some caution is needed when SRD and NSRD are considered. First, there is a risk of 'pathologization' since the threshold of labeling somebody as suffering from 'mild addictive disorder' is critically low. Second, 'addiction' as a term is often used in everyday life for a variety of repetitive and rewarding behaviors. It is a task to transfer knowledge about more restrictive criteria from science to the general population. Third, NSRD may not always be easily detected and defined. In spite of this, addiction is important to be considered in public health, particularly by addiction-related further health disorders, in addition to addictive disorders.

See also: Alcohol Use among Young People; Compression of Morbidity; Drug Use and Abuse: Psychosocial Aspects; Globalization and Health; Health Behaviors; Health Education and Health Promotion; Public Health as Social Science; Smoking and Health.

Bibliography

American Society of Addiction Medicine, 2013. Definition of Addiction (Online). http://www.asam.org/for-the-public/definition-of-addiction.

APA, 2000. Diagnostic and Statistical Manual of Mental Disorders. Text Revision, Fourth ed. American Psychiatric Association, Washington, DC.

APA, 2013. Diagnostic and Statistical Manual of Mental Disorders, Fifth ed. American Psychiatric Association, Arlington, VA.

Ashley, L.L., Boehlke, K.K., 2012. Pathological gambling: a general overview. Journal of Psychoactive Drugs 44, 27–37.

Corwin, R.L., Grigson, P.S., 2009. Symposium overview – Food addiction: fact or fiction? Journal of Nutrition 139, 617–619.

Dagher, A., 2009. The neurobiology of appetite: hunger as addiction. International Journal of Obesity (London) 33 (Suppl. 2), S30–S33.

Danaei, G., Ding, E.L., Mozaffarian, D., Taylor, B., Rehm, J., Murray, C.J., Ezzati, M., 2009. The preventable causes of death in the United States: comparative risk assessment of dietary, lifestyle, and metabolic risk factors. PLoS Medicine 6, e1000058.

Edwards, G., 2012. "The evil genius of the habit": DSM-5 seen in historical context. Journal of Studies on Alcohol and Drugs 73, 699–701.

Finucane, M.M., Stevens, G.A., Cowan, M.J., Danaei, G., Lin, J.K., Paciorek, C.J., Singh, G.M., Gutierrez, H.R., Lu, Y., Bahalim, A.N., Farzadfar, F., Riley, L.M., Ezzati, M., 2011. National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9.1 million participants. Lancet 377, 557–567.

Gardner, E.L., 2011. Addiction and brain reward and antireward pathways. Advances in Psychosomatic Medicine 30, 22–60.

Grant, B.F., Stinson, F.S., Dawson, D.A., Chou, S.P., Dufour, M.C., Compton, W., Pickering, R.P., Kaplan, K., 2004. Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders: results from the National Epidemiologic Survey on Alcohol and Related Conditions. Archives of General Psychiatry 61, 807–816.

Grant, J.E., Potenza, M.N., Weinstein, A., Gorelick, D.A., 2010. Introduction to behavioral addictions. American Journal of Drug and Alcohol Abuse 36, 233–241.

Heatherton, T.F., Kozlowski, L.T., Frecker, R.C., Fagerstrom, K.O., 1991. The fagerstrom test for nicotine dependence: a revision of the fagerstrom tolerance questionnaire. British Journal of Addiction 86, 1119–1127.

John, U., Hapke, U., Rumpf, H.-J., 2003. A new measure of the alcohol dependence syndrome: the severity scale of alcohol dependence. European Addiction Research 9, 87–93.

Karim, R., Chaudhri, P., 2012. Behavioral addictions: an overview. Journal of Psychoactive Drugs 44, 5–17.

Kessler, R.C., Berglund, P., Chiu, W.T., Demler, O., Heeringa, S., Hiripi, E., Jin, R., Pennell, B.E., Walters, E.E., Zaslavsky, A., Zheng, H., 2004. The US National Comorbidity Survey Replication (NCS-R): design and field procedures. International Journal of Methods in Psychiatric Research 13, 69–92.

Lightwood, J., Glantz, S.A., 2013. The effect of the California tobacco control program on smoking prevalence, cigarette consumption, and healthcare costs: 1989–2008. PLoS One 8, e47145.

Lim, S.S., Vos, T., Flaxman, A.D., et al., 2012. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010. Lancet 380, 2224–2260.

McGinnis, J.M., Williams-Russo, P., Knickman, J.R., 2002. The case for more active policy attention to health promotion. Health Affairs (Millwood) 21, 78–93.

Merikangas, K.R., McClair, V.L., 2012. Epidemiology of substance use disorders. Human Genetics 131, 779–789.

Pierce, R.C., Vanderschuren, L.J., 2010. Kicking the habit: the neural basis of ingrained behaviors in cocaine addiction. Neuroscience & Biobehavioral Reviews 35, 212–219.

von Ranson, K.M., Wallace, L.M., Holub, A., Hodgins, D.C., 2013. Eating disorders, substance use disorders, and impulsiveness among disordered gamblers in a community sample. European Eating Disorders Review 21, 148–154.

Samhsa, 2011. Results from the 2010 National Survey on Drug Use and Health, Summary of National Findings. Substance Abuse and Mental Health Services Administration, Rockville, MD.

Schulden, J.D., Thomas, Y.F., Compton, W.M., 2009. Substance abuse in the United States: findings from recent epidemiologic studies. Current Psychiatry Reports 11, 353–359.

Skinner, M.D., Aubin, H.J., 2010. Craving's place in addiction theory: contributions of the major models. Neuroscience & Biobehavioral Reviews 34, 606–623.
Sussman, S., Sussman, A.N., 2011. Considering the definition of addiction. International Journal of Environmental Research and Public Health 8, 4025–4038.
WHO, 2003. WHO Framework Convention on Tobacco Control. World Health Organization, Geneva.
WHO, 2011. Global Status Report on Non Communicable Diseases 2010. World Health Organization, Geneva.
Wittchen, H.U., Jacobi, F., Rehm, J., Gustavsson, A., Svensson, M., Jonsson, B., Olesen, J., Allgulander, C., Alonso, J., Faravelli, C., Fratiglioni, L., Jennum, P., Lieb, R., Maercker, A., Van OS, J., Preisig, M., Salvador-Carulla, L., Simon, R., Steinhausen, H.C., 2011. The size and burden of mental disorders and other disorders of the brain in Europe 2010. European Neuropsychopharmacology 21, 655–679.
Wood, W., Neal, D.T., 2007. A new look at habits and the habit-goal interface. Psychological Review 114, 843–863.

Administration in Organizations

Jean-Claude Thoenig, Université Paris-Dauphine, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Administration in organizations emerged as a specific field of inquiry for social sciences in the middle of the twentieth century. Herbert Simon has defined a program that allows social sciences to move from principles to concepts about action and action taking. Four main perspectives of administration and management are currently influential for further knowledge building as well as for practice-oriented approaches: the art and craft of handling strategic actors and power-based processes, the processes by which moral communities are built and integrated, the cognitive and epistemic dynamics generating cohesion and shared identities inside differentiated organizations, and domination regimes.

Administration in organizations – sometimes referred to as administrative science, or management theory – is a mid-twentieth-century knowledge construct that is considered a scientific revolution. Social sciences such as sociology and political science consider management issues and practices to be a specific field of inquiry that should not be left, as before, to the sole attention of microeconomics or mechanical engineering-inspired principles. It deals with action and action taking in social units that follow some particular purpose: public agencies, firms, and voluntary associations (*see* Bureaucracy and Bureaucratization). How far is it possible within such apparently rational entities to mobilize heterogeneous resources and differentiated logics of action so as to mobilize people and behaviors around similar agendas, and to integrate them so as to induce compatibility and cooperation between them, from top to down, from upstream to downstream, and across hierarchical levels and the inner division of labor?

Generating efficiency in achieving common goals and policies, a recurrent issue in profit-oriented, not-for-profit, and public administration, is explored by theories and based on empirical observations that explain behaviors and outcomes and even make them more predictable. Four approaches derived from social science inquiry shall be presented that play an influential role, whether for current knowledge building or for educational and professional purposes: the organization as an arena for strategic behavior, administration as moral community building, the organization as a cognitive and knowledge set, and administration as the agent of societal dynamics.

From Principles to Concepts

Modern management and organization thinking are rooted in the industrial revolution. What is often called the classical theory was dominant well into the 1940s.

The extensive use of machines and its corollary, specialization, meet competitive requirements to mass-produce low-cost standard products supplied by poorly skilled labor forces. In this setting, a human organization is assumed not to be different from a machine: what basically matters is its formal design. Basic assumptions are that factories exist to accomplish economic goals in accordance with rational criteria of choice, and that there always exists one best way, and only one, to solve a problem. Administration is supposed to be a matter of science. Action guidelines can be derived from universally applicable principles, whatever the type of organization is: centralization of equipment in factories, hyperspecialization of tasks, unity of command, and salaries based on individual financial incentives.

Mechanical engineers are in charge of designing the division of labor and defining the right procedures. They govern by principle because they are knowledgeable. Daily administration relies on two principles. What is expected from the workforce is basically conformity and discipline: foremen are hired to ensure that procedures get enforced. And workers, while considered ignorant and therefore not entitled to suggest alternative administrative approaches, are assumed to be basically motivated by greed: bonuses drive their behaviors to deliver productivity.

Some of the leading theorists are well known, such as Taylor (1911), an American manufacturer; Elton Mayo, a psychology professor at Harvard University; and Fayol (1949), a French engineer. While Fayol was handling the issues of how to manage a firm as a whole, Taylor was defining expertize about how to get the individual worker organized, and Mayo was exploring how closely actual individual behaviors matched the assumptions made by the classical school.

A strong attack was launched after World War II challenging such oversimplistic, mechanistic views of administration, in particular, by scholars trained in sociology and in political science. Simon (1946) emerged as the most prominent rebel.

According to Simon, principles as defined by the classical school are merely proverbs: they are neither true nor false. Simon criticizes in a rather abrupt manner the relevance of such an approach. Rigid task specialization, span of control, and unity of command lead to impasses. They are conflicting and inconsistent with most situations that administration faces. With equal logic, they are applicable in diametrically opposed ways to the same sets of circumstances, as suggested by empirical observation. Therefore, in order to become a really scientific theory, administration in organizations has to substitute concepts for principles and make them operational.

In a subsequent book, Simon (1947) lays the groundwork for administration as a specific field of inquiry. He sketches a conceptual framework, the meaning of which corresponds

to facts or situations that are empirically observable. He questions, for instance, the relevance of the principle of absolute rationality. In organizations, even if they are purposive, individuals do not have the intellectual or informational capacity to maximize, and they are also vulnerable to their surrounding social and emotional contexts. What human beings do when making decisions is to satisfice: they try to find trade-offs between preferences and processes, and they do the best they can where they are. Human and organizational decisions are, therefore, subject to bounded rationalities.

In addition, efficiency is not a goal that is shared the same way by everyone in the organization, including the managers, nor can it be defined *ex ante*. It should be a research question, starting from the hypothesis that the individuals or the organizations themselves carry a specific definition of what is good or correct from an efficiency point of view. In more general terms, contexts vary, and they make a difference. Organizations are local orders with specific characteristics.

Simon extends Max Weber's perspectives: administration belongs to the domain of rational action. Firms or public agencies are organizations driven by purposes. But managers rely upon the mediation of an organized setting in order to implement goals, purposes, or values. Therefore, the organization simultaneously provides a resource and becomes a constraint; managers experience it as a solution as well as a problem. Simon underlines the necessity for social sciences to approach administration as a field aimed at understanding the nature of empirical phenomena. Its primary goal is not to formulate solutions for action but to consider action as a problem under scrutiny. Practitioners could nevertheless rely upon relevant findings and apply such a body of knowledge – or part of it – to enlighten problem solving. Such an agenda is structured around studying the actual functioning of organizations.

In a more specific way, Simon defines decision-making processes or action as the core concern of a scientific discipline of management. Any decision or action can be studied as a conclusion derived from a set of premises by the organization or an individual. Some premises are factually grounded: they link a cause to an effect. Therefore, they are subject to testing in real life. Other premises are of a different nature: they are value grounded, made out of norms or ethical references. In this case, they are not empirically checkable. While both categories are not separable in action, analysts have to separate them and focus upon factual premises only.

Firms and public agencies should also be treated as open organizations. They do not and cannot exist as self-contained islands within society and the market; they are linked to specific environments. The relationships structured between the inside and the outside play very important functions. Where and how an organization is embedded, and what is exchanged, are phenomena that have an impact on an organization's inner functioning as well as on the environment.

Philip Selznick describes how an American federal authority gains support for its programs within local communities where it operates (Selznick, 1949). The concept of cooptation refers to a social process by which an organization brings outside local milieux and their leaders into its policy-making process, enabling such elements to become allies rather than a threat to its existence and its mission. Bringing the environment back in solves a major problem that classical approaches of public administration did not address especially when dealing with public administration. Two of his founders, Woodrow Wilson and Frank J. Goodrow, had been calling for a theory of public administration that made a dichotomy between politics and administration, between the elaboration of the policy by policy makers and the execution of that will by bureaucratic enforcement agencies. Selznick suggests that such a postulate should become a research question. He also proposes that, beside organizational phenomena as such, science should consider institutionalization dynamics, which means how values and norms are diffused and appropriated, and what impacts they have on managerial action taking.

Managing Arenas for Strategic Behavior

Simon's agenda paved the way for a behavioral revolution in the field of administration. During the 1950s, at the Carnegie Institute of Technology and under his leadership, he reviewed with James March studies of bureaucracies authored by social scientists such as Robert K. Merton, Philip Selznick, and Alvin W. Gouldner. They compared various models of bureaucratic behavior (March et al., 1958). In highly proceduralized organizations, whether private or public, individuals, and groups do not remain passive: they reinterpret rules and procedures, they play with and around them, and they use them for secondary purposes of their own such as increasing their autonomy inside the hierarchical line of authority or bargaining their participation with the organization. At an organizational level, administration by rule generates dysfunctional processes. Formalization and centralization induce vicious circles. In order to fight the unintended consequences that such tools generate – such as rigidity or inner-silo effects – organizations administer themselves by using more rules and by increasing centralization, therefore reinforcing rigidity and inner differentiation.

The Carnegie School also criticizes the theory of the firm as defined by orthodox microeconomics. Organizational decision making is the focal point. Is utility maximization the main function that business firms do, in fact, achieve? Cyert and March (1963) study how trade-offs are structured and activated inside a company, at their top as well as middle levels, around action taking and choice processes. Negotiations occur through which coalitions impose their demands on the organizational objective. It is most likely that any organization functions like a polyarchy: single members do not share an identical definition of its mission, goals, and collective interest. Simon's conception is demonstrated as being applicable to economic actors: satisficing is a much more powerful concept to explain their strategic decisions than maximizing economic profit. Such is specifically the case of pricing in an oligopolistic market. In other words, organizational structure and inner political games determine rational behavior.

Conflict is a basic attribute of any organization. Business firms and public agencies are not monolithic entities or integrated machines sharing one common purpose. They behave as pluralistic systems in which differentiated and even antagonistic interests float around, conflict, define acceptable compromises, or induce temporary cooperation. They look like political coalitions between subgroups (March, 1962). From an

action-taking or administrative perspective, organizations require their leaders to develop skills that are less analytical than behavioral. Administrators are close to political brokers, and negotiating and bargaining inside their organizations are crucial tasks to handle. A firm looks like an arena for strategic microbehaviors, a collection of subunits pursuing separate goals. The role of administration is to structure inducements so that each individual subunit identifies its interests with those of the firm and, thereby, contributes to its mission. Executives at the top behave and decide accordingly to roles that are more or less social rites (Mintzberg, 1973).

In the 1960s, the behavioral approach widened internationally and gave birth to a stream of organizational researches about decision making, power, and efficiency.

Allison (1971) studied the same event – the US presidential handling of the 1962 Cuban missile crisis – comparing three different paradigms about decision making. An organizational process model, which is clearly derived from the Carnegie School approach, completed with a political process model, which deals with partisan politics and presidential tactics on the public opinion scene, shows a superior ability compared to a rational actor or classical model to explain how President John F. Kennedy addressed the challenge and which outcomes were elaborated, despite the game theory–based techniques used by the Executive Branch to define rational solutions. Lindblom (1959) too takes a hard look at the rational model of choice. He rejects the notion that most decisions should be made based on total information processes. He suggests that synoptic approaches provide self-defeating strategies for action. Instead, he defines the whole policy-making processes as being dependent upon small instrumental decisions that tend to be made in a disjointed order or sequence in response to short-term political conditions. Such a muddling-through view prescribes managers to make small changes at a time and at the margin, not focusing too much or explicitly about content, and whenever it is possible and if needed making some minor concession – two steps forward and one step backward.

Do formal decision makers in certain contexts really play a relevant part in how decisions are made? A so-called garbage can model of choice suggests that it might well not be the case at all, in particular in organizations that are either highly bureaucratic or, at the opposite, very loosely structured (Cohen et al., 1972). Choices are characterized by ambiguity about goals, intentions, technologies, causation, participation, and relevance. What looks like a problem to actor A looks like a solution to actor B. Formal opportunities for choice look after problems to handle. Decisions are made without being considered by the participants as being made. Such anarchic contexts occur in specific organizational settings such as bureaucracies and very loosely formalized structures such as universities. Nobody is really in control of the process, and decisions are experienced as random-based outcomes. The implications of such a model for top managers are that they should not use quantitative tools as instruments of government or intervene in tactical ways but keep their hands free for what they consider fundamental issues and use two basic vehicles as action tools: the selection of their immediate subordinates and a redesign of their organization's formal structures.

Power phenomena are viewed as key variables for understanding and administering. The French sociologist Crozier (1963) offers a perspective that helps interrelating microprocesses (behaviors of single actors) and macroprocesses (the functioning of the whole organization). Individuals and groups are pursuing rational strategies: they try to fulfill goals and satisfy vested interests that are structured by the specific context within which their act takes place. Asymmetric interdependence relationships link them together: some are more dependent than others to satisfice their stake or goal from the discretion of others rather than from their own goodwill. Those who control a source of uncertainty on which others depend control power bases and are able, in exchange for their goodwill, to set up the rules of the game. In other terms, organizational functioning and change derive from the social regulation processes as induced by the actors who, at various levels of the pyramid, try to make their specific and heterogeneous strategies or logics of action compatible. From a managerial point of view, such a comprehensive framework implies that administration is about the art and skills to reallocate uncertainties and power inside the organization and, therefore, to structure interests inducing the actors to cooperate or not. Bower (1970) applies such a perspective to strategic investment planning in a giant corporation. Allocating capital resources is a process that requires management to identify the various organizational components such as routines, parochialism, attention to issues, and discretionary behaviors of action controlling major uncertainties.

A third major critical contribution made by sociologists deals with rationality, optimality, and efficiency. Landau (1969) argues that redundancy within a firm or a public agency is not a liability – a symptom of waste and inefficiency – but a fundamental mechanism of reliability production. Duplication and overlap provide solutions for action taking in general. The breakdown of one part does not penalize the whole system. The arrogance of a subsystem controlling a monopoly on a problem or a function is diminished. Duplication and overlap may create political conflicts. They also generate conditions for communication, exchange, and cooperation. They lower risks. Organizations are not self-evaluating entities. They tend to substitute their own knowledge for the information generated by their environment. Economic efficiency and optimality as defined by economists are normative enterprises.

Such sociological contributions therefore question three fundamental postulates of the classical theory of administration: the ability of the center or the top, which has the monopoly of authority, to control the agenda of the whole organization by designing formal structures and rules; the fact that the organization functions like a serial or linear machine, in which what happens at the periphery is fully determined by the center; and the expectation that membership makes a difference. In fact, it may well happen that those who have full authority control and exert not much power, while those who have no hierarchical authority in fact have the power and these autonomous actors produce and shape the actual outcomes and policies; in such situations, there is not much continuity between the top and the bottom, and formal limits do not make a difference. Implicit in this view of the organization is the assumption that organizational actors, either persons or subunits, possess preferences and influence

resources that include position or office, functional or professional expertize, side payments, and the like. Because organizations are action arenas driven by power dynamics, administration therefore is the art and craft of generating strategic behaviors. The levers of rewards and sanctions, to align individual behavior and collective goals, rely upon specific incentives, whether social (status, autonomy, etc.), organizational, or financial, and relate to vested interests of various sorts. Designed poorly, incentives may produce subunit conflict and poor firm performance.

Administration as Moral Community Building

While the behavioral revolution still remains a most influential perspective, social sciences have explored alternative perspectives to understand administration processes in organizations. Another view of administration, while still assuming that organizational actors hold resources that may drive decision making, differs from the first view, however, by relaxing the assumption of strong preferences for specific action outcomes.

A series of studies underline the importance of cultural effects (Hofstede, 1980). More specifically, they provide evidence that within the same firm very different patterns are shared across the various countries in which its units or subsidiaries are located as far as values, norms, and attitudes linked to administrative processes and behavioral practices are concerned. For instance, in certain countries, bypassing the hierarchy is considered as a moral sin, while in other countries it is considered as appropriate. Leadership styles considered as acceptable or efficient here are considered as nonappropriate or noneffective there. In other terms, national culture, an exogenous factor, shapes administration in the organization, leaving no room for maneuver except to act in accordance with these cultural macrofactors. Identical phenomena may also be linked to the role played by professions and to the references their members share across organizations, such as MBA-trained managers or the medical professions.

Research nevertheless also shows that organizations as such are not passive but are specific and proactive cultural entities. They carry and mold values, norms, beliefs, and rites even inside the same country. Administration socializes their members as belonging to a specific community that differs from others. Administrators take actions that are designed to help structure more or less plastic preferences. They govern the organization and its members as cultural entrepreneurs. Mechanisms include leadership, especially charismatic leadership; ideology; socialization; recruitment; and environmental constituencies to which individuals have private or professional loyalty. The organization is understood and managed as a moral community. Common to all these mechanisms is the attempt to foster identifications and loyalties, the normative order providing the integrative backbone of the organization. Administration is about forging and changing values, norms, and cognitive characteristics: it may also have to do with preaching and educating.

The role of administrators in structuring preferences is documented by a set of literature on missionary, professional, and community organizations. Institutionalization as studied by Selznick (1949) offers a vehicle to mobilize an organization for meaning and action. Knowledge, or interpretations in action, structures the community. The theoretical roots of such an approach relate to two different social science traditions.

Shils (1975) identifies in each society the existence of a center or a central zone that is a phenomenon of the realm of values, beliefs, as well as action. It defines the nature of the sacred and it embodies and propounds an official religion, something that transcends and transfigures the concrete individual existence, the content of authority itself. The periphery in mass society is integrated through a process of civilization. Anthropologists such as Geertz (1973) demonstrate that culture as a collectively sustained symbolic structure is a means of saying something of something. Through emotions, common cognitive schemes or common meanings are learned: they provide an interpretative function, a local reading of a local experience, which is a story the participants tell themselves about themselves.

More recent contributions have laid down perspectives focused specifically around organizations and their administration. Various processes within firms actually play the role of center: brainstorming sessions, informal encounters, networks linking persons across departments and units, socialization mechanisms of newcomers, and so on. Strong centers are able to create rigidity phenomena in terms of cognitive blindness, which is the firm as a community being unable to catch signals emitted by its environment. Kogut and Zander (1996), for instance, treat firms as organizations that represent the social knowledge of coordination and learning: identity lies at the heart of such social systems, which implies a moral order as well as rules for exclusion.

Action-Oriented Cognition Building

Organizations develop endogenous capabilities to reason, analyze, and think. Daft and Weick (1984) propose a model of organizations as interpretation systems that stresses their sociocognitive characteristics more than the economic ones. Interpretation is the process through which information is given, meaning and actions are selected and fulfilled. They produce knowledge of their own. They are not only power arenas or moral and symbolic orders, but also action-oriented language systems. This perspective emphasizes that administration influences the references mobilized by its members, with such references being actually considered as more or less implicit references and criteria for action taking. It also considers an organization as a collective actor that is not a community, as a homogeneous entity sharing identical values, and as a complex pluralistic, polyarchical, or even heterarchical system (a heterarchy is a system of organization replete with overlap, multiplicity, mixed ascendancy, and/or divergent-but-coexistent patterns of relation) (Stark, 2001). An organization and its administration generate knowledge, knowledge that is not a mere representation of the world and the market, and that is more or less biased. It has consequences. Daily interactions and choices at all levels in the organization carry specific modes of thought.

While sociological perspectives such as ethnomethodology, rational choice and individual methodologism, and reflexivity theory give much attention to individual cognitions, by

comparison fewer studies question the postulate that organizations function like passive social robots, and consider their administration as the art and craft of mental 'bricolage.' Administration in organizations develops abilities to understand what their members do while they do it. They have good reasons to make errors, to the extent that prejudices and beliefs result from ways of reasoning in a logical manner.

Interpretative processes are produced in a collective way (Eden and Spender, 1998). Cognitions generate and diffuse knowledge that is both substantial and procedural. They drive behaviors, providing coordination and cooperation. Specific and evolving languages for action connect the ambitions of the organization and administrative acts, headquarters and operational units, and upstream and downstream functions and tasks (Michaud and Thoenig, 2003). Sharing a common language does not imply that within and across an organization, a uniform and identical set of references is prevalent that is the same for every part or member. Shared means that in interactive situations, use is made of cognitive references and causalities that are compatible (i.e., that link one with another). Such is the case in firms and also in intergovernmental organizations (Schemeil, 2011). Cognitions in action provide implicit decision criteria. They define causal relationships, choice indicators, and time and space horizons.

Two different sources of languages exist. Exogenous languages come from third-party sources, from the exterior of the interactive social body: they are imported as kits, for instance, by hiring professionals such as marketers or controllers. Endogenous languages are codes through which content is constructed by the organization itself. They may not be in line with references or beliefs dominant in the environment. Administration generates practical theories and performative abilities. As a legitimate source, it sets up transitory arrangements between different functional and professional requirements. It defines principles, and it also transgresses them. Administration constructs specific senses, interpretations, and theories of action. Organizations are cognitive and interpretative islands in societies and economies (Koza and Thoenig, 2003).

When Societal Dynamics Impose Administration Regimes

Administration may also be considered as an exogenous rationale whose content is modeled by societal trends and imposed by macroforces. Organizations therefore are not at all local orders of their own. Administration is to be understood as the acts and nonacts of actors who are agents or servants of principals that are out of reach of the organizations: professional communities (engineers, lawyers, doctors, etc.), elites such as Oxbridge and French Grandes Ecoles alumni, administration trends and fads such as modern management and New Public Management, and so on.

In social sciences, this holistic perspective is in line with an old tradition. Such is the case with the concept of the iron cage developed in the very early 1900s by Max Weber (Weber et al., 2002). In this belief, societal, if not global ideologies, for instance, linked to religion or to civil society values, integrate organizations along identical postulates and norms. Administrators adopt the beliefs they induce about action and representation of the world because they prefer to conform rather than to deviate. Therefore, mesolevel-generated change or innovation not only is risky for themselves and for their organization, but also is experienced as a moral sin or social extravaganza.

This sociological tradition overemphasizing deterministic approaches of administration remains quite lively at the beginning of the twenty-first century, partly as an attempt to challenge political economics as a discipline in order to explain how and why economic forms of exchange such as markets operate and evolve, in particular, at the time of rapid globalization of economies and polities.

For instance, sociological institutionalism underlines the importance of constraints such as legitimacy imperatives shared by national teacher associations, with specific local context variations relegated to the background (Meyer and Scott, 1983). Organizational sociologists note the existence of populations of organizations, firms, and nonprofit organizations operating in an identical sector (wineries, newspapers, museums, regulatory bodies, etc.) that adopt identical formal structures, policy ambitions, or inner surveillance rules that are considered as legitimate or as best practices in the field in which they operate (DiMaggio and Powell, 1983). Coercive, normative, and imitative types of isomorphism are driven by the fact that organizations and their administration expect to gain the trust and support of the stakeholders they depend on to achieve their goals. To a large extent, administration as a construction of symbols is linked to messages that are transmitted by certification agencies, global standards, and regulations, when not consulting companies (Brunsson and Jacobsson, 2000). Nevertheless, globalization does not mean that local orders do not matter any longer and that administration at the mesolevel (when not at the microlevel) becomes residual (Sorge, 2005).

A more radical school of thought considers administration at the local order as a mere artifact, as an agent of dominant social classes. Since the end of the twentieth century, a new iron cage is in command in which administration implements postbureaucratic approaches.

This critical school considers that modern management, as a professional reference and as a set of local practices, basically expresses the interests and pressures of neocapitalism's dynamics and pressures. Shareholder value creation is the ultimate goal to satisfy. Business school education and research provide tools and legitimacy to a new set of methods to incentivize and control members of an organization. Short-term and quantitative performance indicators should only be considered, due to the members of the organization being driven by greed, and this is expected to apply from top to down all along the authority hierarchy. Any knowledge supplied by social sciences is considered as irrelevant for organizational administration: individuals and groups behave with a rationality that is close to the postulates on which microeconomics is based.

This radical approach describes administration as a domination regime. Contrary to polyarchic regimes, oligarchic forms of power overemphasize hierarchy of authority and top-down approaches as legitimate governance processes. They make no room for deliberation and discussion. Blind allegiance to the rules is expected from the rank and the file. But this domination

is administered in a soft manner, and not by coercion or raw force. While the discourse keeps making reference to processes such as decentralization, quality of working life or participation, or self-actualization, in practice what is expected is that individuals and groups appropriate the goals and pressures emitted by their hierarchy as being rational and making sense for themselves. Symbolic violence is the name of the game (Courpasson, 2006).

See also: Closed and Open Systems: Organizational; Conflict: Organizational; Learning: Organizational; Management: General; Organizational Behavior, Psychology of; Organizational Decision Making; Organizations, Sociology of; Organizations: Authority and Power; Work and Industry, Sociology of.

Bibliography

Allison, G.T., 1971. Essence of Decision: Explaining the Cuban Missile Crisis. Brown, Boston: Little.
Bower, J.L., 1970. Managing the Resource Allocation Process. Harvard University Press, Boston.
Brunsson, N., Jacobsson, B., 2000. A World of Standards. Oxford University Press, Oxford.
Cohen, M.D., March, J.G., Olsen, J.P., 1972. A garbage can model of organizational choice. Administrative Science Quarterly 17 (1), 1–25.
Courpasson, D., 2006. Soft Constraint. Liberal Organizations and Domination. Liber: Copenhagen Business School Press.
Crozier, M., 1963. The Bureaucratic Phenomenon. University of Chicago Press, Chicago.
Cyert, R.M., March, J.G., 1963. A Behavioral Theory of the Firm. Prentice Hall, Englewood Cliffs.
Daft, R.L., Weick, K.E., 1984. Toward a model of organizations as interpretation systems. Academy of Management Review 2, 284–295.
DiMaggio, P.J., Powell, W., 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. American Sociological Review 48, 147–160.
Eden, C., Spender, J.C. (Eds.), 1998. Managerial and Organizational Cognition. Sage, London.
Fayol, H., 1949. General and Industrial Management. Pitman, London.
Geertz, C., 1973. The Interpretation of Culture. Basic Books, New York.
Hofstede, G., 1980. Culture's Consequences. Sage, Beverly Hills.
Kogut, B., Zander, U., 1996. What firms do? Coordination, identity and learning. Organization Science 7, 502–518.
Koza, M.P., Thoenig, J.C., 2003. Rethinking the firm: organizational approaches. Organization Studies 8, 1219–1229.
Landau, M., 1969. Redundancy, rationality, and the problem of duplication and overlap. Public Administration Review 29, 349–358.
Lindblom, C.E., 1959. The science of 'muddling through'. Public Administration Review 19, 79–88.
March, J.G., 1962. The business firm as a political coalition. Journal of Politics 24, 662–678.
March, J.G., Simon, H.A., Guetzkow, H., 1958. Organizations. Wiley, New York.
Meyer, J.W., Scott, W.R., 1983. Organizational Environments: Ritual and Rationality. Sage, Beverly Hills.
Michaud, C., Thoenig, J.C., 2003. Making Strategy and Organization Compatible. Palgrave Mcmillan, London.
Mintzberg, H., 1973. The Nature of Managerial Work. Harper and Row, New York.
Schemeil, Y., 2011. Dynamism and resilience of intergovernmental organizations in a world of persisting state power and rising non-state actors. In: Bousseau, E., et al. (Eds.), Governance, Regulations and Powers on the Internet. Cambridge University Press, Cambridge, pp. 237–251.
Selznick, P., 1949. TVA and the Grass Roots. University of California Press, Berkeley.
Shils, E., 1975. Center and Periphery. University of Chicago Press, Chicago.
Simon, H.A., 1946. The proverbs of administration. Public Administration Review 6, 53–67.
Simon, H.A., 1947. Administrative Behavior. MacMillan, New York.
Sorge, A., 2005. The Global and the Local. Understanding the Dialectics of Internationalization. University Press, Oxford.
Stark, D.C., 2001. Ambiguous assets for uncertain environments: heterarchy in postsocialist firms. In: DiMaggio, P. (Ed.), The Twenty-First-Century Firm: Changing Economic Organization in International Perspective. Princeton University Press, pp. 69–104.
Taylor, F.W., 1911. The Principles of Scientific Management. Harper & Brothers, New York.
Weber, M., Baehr, P.R., Wells, G.C., 2002. The Protestant Ethic and the Spirit of Capitalism (1904–1905) and Other Writings. Penguins, London.

Administrative Law: The United States and Beyond

Cary Coglianese, University of Pennsylvania Law School, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Administrative law constrains and directs the behavior of officials in the many governmental bodies in every country responsible for implementing legislation and handling governance responsibilities on a daily basis. This field of law consists of procedures for decision making by these administrative bodies, including rules about transparency and public participation. It also encompasses oversight practices provided by legislatures, courts, and elected executives. The way that administrative law affects the behavior of government officials holds important implications for the fulfillment of democratic principles as well as effective governance in society.

Administrative law refers to the body of rules and procedures affecting government agencies as they implement legislation and administer public programs. Yet it is also much more than just rules and procedures. Administrative law applies to the ongoing operation of government bodies and seeks to shape official decisions that impact businesses and citizens throughout society. These decisions include granting licenses, dispensing government benefits, conducting inspections and investigations, imposing sanctions, issuing orders, awarding contracts, collecting information, hiring employees, and even making still further rules and regulations that apply to both governmental and private actors. Administrative law affects all of these varied decisions and addresses fundamental questions about how government authority can and ought to be exercised. It implicates society's most deep-seated political and moral values: democracy, equity, efficiency, privacy, transparency, and justice. And it does so by intervening in complex and diverse organizational environments within which public and private actors face varied, often shifting, motivations, incentives, and constraints. A proper study of administrative law therefore requires immersion in a wide breadth of issues in social science: normative as well as positive political theory; individual as well as organizational behavior; and law as well as politics, sociology, public administration, and economics.

Even when it is just considered as a body of rules, administrative law is complex. It draws its legal pedigree from a variety of sources: constitutional law, statutory law, internal policy, and, in some countries, common law. Government agencies' organizational structures and routines are shaped by provisions derived from both generic procedural statutes (such as, in the US, the Administrative Procedure Act) and statutes addressing specific substantive policy issues such as energy, education, taxation, or welfare benefits. This array of legal sources means that administrative rules and procedures can vary significantly across agencies and, even within the same agency, across discrete policy issues or types of actions.

The social science study of administrative law seeks to make sense of the complexity of administrative law and how it shapes, and is shaped by, the organizational environments within which it operates. Research has proceeded not only to test theoretical propositions about whether and how legal norms and institutions influence administrative behavior but also to identify and help solve applied problems. Administrative law research is characterized in part by prescriptive efforts to design rules that better promote political and social values, and in part by empirical efforts to explain how law influences the behavior of government agencies. Government agencies often possess considerable policy discretion but are staffed by unelected officials, so a key objective for administrative law scholars has been to understand how agency officials are, or can be, held democratically accountable (Lodge and Stirton, 2010). Administrative law places particular emphasis on the empirical understanding of the impact of courts and other oversight bodies – as these entities purport to hold administrators accountable to elected officials and the publics they represent. Although administrative law scholarship has a rich and important tradition of doctrinal analysis, the insights and methods of social science have become essential for understanding how administrative law and legal institutions can affect democratic governance. By drawing on social science methods to understand how legal rules and institutions affect governance, administrative law scholarship aspires both to inform and to improve the outcomes of public institutions.

Administrative Law and Democracy

Administrative agencies make decisions affecting citizens' lives and entire industries – but these agencies are usually staffed by officials who are neither elected nor otherwise directly accountable to the public. A fundamental challenge in both positive and prescriptive scholarship has been to analyze administrative decision making from the standpoint of democracy. This challenge is particularly pronounced in constitutional systems with executive bodies that are formally separate from the legislature and where political party control can be divided between the branches of government, as in the US. But the general challenge applies anywhere because of the enormous discretion afforded to unelected administrative officials. Much work in administrative law aims either to justify administrative procedures in democratic terms or to analyze empirically how those procedures can effectuate democratic values.

A common way to reconcile decision making by unelected administrators with democratic principles has been to consider administrators as mere implementers of decisions made through a democratic legislative process. Under what is sometimes called the 'transmission belt' model of administrative

law, administrators are treated as mere instruments used to implement the will of the democratically controlled legislature (Stewart, 1975). Statutes serve as the 'transmission belt' to the agency, both transferring democratic authority to administrative actors and constraining those actors so that they advance legislatively approved goals.

As a positive matter, the 'transmission belt' model underestimates the amount of discretion held by administrative officials. Statutes are seldom self-executing. They need interpretation and must be applied in myriad concrete circumstances. In interpreting and applying statutes, administrators assume discretion. Statutes do not always speak clearly to the varied circumstances that confront administrators. Not only are many of these circumstances unanticipated by legislators, but elected officials often may lack incentives for making laws clear or precise in the first place, as it can be to their electoral advantage to appear to have addressed vexing social problems only in fact to have passed difficult policy questions and tradeoffs along to unelected administrators. For some administrative tasks, particularly monitoring and enforcing laws, legislators give administrators explicit discretion over how to allocate their agencies' resources to pursue broad legislative goals.

Scholars disagree about how much discretion legislators ought to allow administrative agencies to exercise. Minimalists, emphasizing the electoral accountability of the legislature, have urged that any legislative delegations of authority to agencies be narrowly constructed (Lowi, 1979). Those scholars of a more expansionist bent emphasize administrators' indirect accountability to elected officials and contend that legislatures themselves are not perfectly representative, especially when key decisions are delegated internally to committees and legislative staff (Mashaw, 1985). While the optimal amount of authority to be delegated to agencies remains a subject of analysis (Stephenson, 2008), in practice administrative agencies continue to possess considerable discretion, even under relatively restrictive delegations.

Given that agencies do possess discretion, one aim of administrative law has been to identify procedures that encourage administrators to exercise their discretion in ways that promote both procedural and substantive values. A leading approach has been to design administrative procedures to promote broad public participation, including representation of a wide array of interest groups (Stewart, 1975). Transparent procedures and opportunities for public input give organized interests and ordinary citizens an ability to represent their views in the administrative process. Such procedures include those providing for open meetings, access to government information, hearings, and opportunities for public comment, and the ability to petition the government. Transparency and participation requirements are defended not only on the grounds of procedural fairness, but also because they are expected to deliver more information to administrators before they make decisions. These procedures may also protect against regulatory capture – the much-decried predicament where an agency's decisions come to promote an industry's private interests to the exclusion of the broader public interest (Stigler, 1971).

Although certain requirements, such as the notice-and-comment procedure followed by US agencies when making rules, provide the public with the opportunity to participate in administrative decision making, this does not necessarily mean that any extensive or representative portion of the public actually participates in administrative policymaking. Nor does it mean that public participation has any significant impact on agency decisions. In the US experience, most agency rulemaking proceedings garner only a small number of comments – and in most rulemakings by far the largest number of these are submitted by businesses or other organized groups rather than by what might be considered ordinary members of the public (Coglianese, 2006). On occasion, however, agencies will issue high-salience rules that do garner thousands of comments – typically short, unsophisticated expressions of preferences rather than comments conveying substantive information.

As to whether comments, simple or sophisticated, make a difference, the answer appears to be at most 'sometimes.' Studies find varying degrees of association between arguments presented in comments and changes made to proposed rules (West, 2004; Yackee, 2005; Shapiro, 2008). Formal comments, though, are submitted only after agencies have invested much staff time in developing their proposed rules, a point at which much analysis and decision making has already been completed. For this reason, representatives of organized interest groups often seek to influence administrative policy by making informal contact with officials well before the agency proposes a rule and invites formal public comments (Furlong and Kerwin, 2005). Interest group representatives may also continue to remain involved with the agency after a 'final' decision has been made. Whether through litigation or further discussions, agencies can be persuaded to issue amendments or make other policy changes to otherwise final rules (Coglianese, 1996; West and Raso, 2012).

The widespread use of the Internet has generated interest in so-called e-rulemaking, or the use of information technology to connect the public more closely with the work of administrative agencies. The advent of agency web sites has put much more extensive information about administrative matters at the fingertips of users around the world (provided, of course, individuals can easily navigate through all the extraneous information also taking up space on agency web sites) (Coglianese, 2013). Agencies now routinely accept public comments submitted by email, and many also have a presence on social media sites like Facebook and Twitter (Coglianese, 2013). In the US, the federal government has created a one-stop web site called *Regulations.Gov* which indexes agency regulatory proceedings, houses supporting documents and previously submitted public comments related to new rules, and provides a button for users to submit comments on proposed rules. Although many early advocates of e-rulemaking heralded technology's promise to expand public participation in the regulatory process, to date it appears that the patterns of commenting on agency rulemaking remain largely unchanged (Balla and Daniels, 2007).

It should not be surprising that levels of public participation in rulemaking remain relatively low, as the subject matter of much administrative action remains high in complexity or low in salience – or both. However, technology has undoubtedly made it easier for elites inside and outside of government to monitor what agencies are doing, as well as for scholars of administrative law to study more systematically how

administrative rules and procedures may better serve democratic principles.

Courts and Judicial Review

Concern about democracy also undergirds administrative law's emphasis on judicial review of government action. Under well-accepted legal principles, courts serve as key enforcers both of the substantive laws that government officials are charged with implementing as well as the procedural requirements that these same officials must follow in their implementation of substantive laws. Courts have also imposed their own additional procedures on administrators based on constitutional and sometimes common law principles. A key normative question has centered on how aggressive courts should be when it comes to reviewing the actions of administrative agencies.

Administrative agencies typically possess a greater capacity for making sound technical and policy judgments than do courts. Even in legal systems with specialized administrative courts, not only do agency officials and their staffs possess greater policy expertise than judges but administrators are also often more closely connected to democratic institutions than judges. These considerations have long weighed in favor of judicial deference to administrative agencies, lest judges disregard either the technical expertise or the political legitimacy reflected in many administrative decisions. On the other hand, it is also generally believed that some credible oversight by the courts bolsters agencies' compliance with administrative law and may improve their overall performance. The prescriptive challenge has been to identify the appropriate degree of deference for courts to give to agencies overseeing their decision making.

Sometimes the degree of deference is said to depend on whether agencies are making factual, policy judgments as opposed to making judgments about the meaning of the law. Courts might have grounds for giving more deference to agencies' policy judgments, simply ensuring that they have followed transparent procedures. Yet courts have also been known to take a careful look at policy decisions to see that they are based on a thorough analysis of all relevant issues. The latter approach is sometimes referred to as 'hard look' review in the US, as it calls for judges to probe carefully into the agency's reasoning to ensure that agency officials conducted a thorough analysis of policy options before reaching a decision. Although one might suspect courts would give less deference to agencies' legal interpretations than to their factual judgments, especially when agencies must interpret their own governing legislation, one of the most widely cited US Supreme Court opinions calls upon courts to defer to agency interpretations of ambiguous provisions within the statutes they implement (*Chevron v. Natural Resources Defense Council*, 1984). Prescriptive scholarship seeks to provide analytic guidance for judges on the appropriate level of deference that they should give to both legal and policy choices made by agencies (Zaring, 2010).

The proliferation of prescriptive doctrinal principles in contemporary legal systems gives rise to the question of what impact administrative law has on the actual decision making of judges in deciding cases. After the US Supreme Court issued its *Chevron* decision, lower courts reportedly shifted to deferring more to agency interpretations (Schuck and Elliott, 1990). Yet legal principles, whether articulated by the Supreme Court or reflected in laws adopted by the legislature, are only one factor that may explain how judges make their decisions. Just as administrators themselves possess residual discretion, so too do judges possess discretion in deciding how deferential to be to administrative agencies' policy and legal determinations. As in other areas of law, political ideology also may help explain patterns of judicial decision making in administrative law cases (Revesz, 1997; Miles and Sunstein, 2006).

In addition to analyzing judicial decision making, the field of administrative law has been centrally concerned with the impact of judicial review on the behavior of officials within administrative agencies. Normative arguments about judicial review typically depend on empirical assumptions about the effects courts have on the behavior of administrative agencies. Indeed, much legal scholarship in administrative law builds on the premise that judicial review, if deployed properly, can improve governance (Edley, 1990). The effects often attributed to judicial review include making agencies more observant of legislative mandates, increasing the analytic quality of agency decision making, and promoting agency responsiveness to a wide range of interests. Administrators who know that their actions may be subjected to review by the courts can be expected to exercise greater overall care, presumably making better, fairer, and more responsive decisions than administrators who are insulated from direct oversight.

Notwithstanding these purported beneficial effects from judicial review, scholars have also emphasized courts' potentially debilitating effects on agencies. They have widely accepted, for example, that administrators in the US confront a high probability that their actions will be subject to litigation. Cross-national research suggests that courts figure more prominently in government administration in the US than in other countries (Kagan, 2003). The threat of judicial review purportedly creates significant delays for agencies seeking to develop regulations (McGarity, 1992).

In some cases, agencies have been said to have retreated altogether from efforts to establish regulations. The US National Highway Traffic Safety Administration (NHTSA) is usually cited as the clearest case of this so-called ossification effect, with one major study suggesting that NHTSA has shifted away from developing new auto safety standards in order to avoid judicial reversal (Mashaw and Harfst, 1990). Other research, however, indicates that the threat of judicial interference in agency decision making has been significantly overestimated. Litigation challenging administrative action in the United States occurs less frequently than is generally assumed (Coglianese, 1997), and some research indicates that agencies can surmount seemingly adverse judicial decisions to achieve their policy objectives (Jordan, 2000). Large-sample studies have failed to confirm the view that judicial review significantly obstructs the rulemaking process in the United States (O'Connell, 2008; Yackee and Yackee, 2010).

Concern over excessive adversarialism in the administrative process persists in many countries. Government decision makers have at times pursued collaborative or consensus-based processes as alternative strategies for creating and implementing administrative policies. In the US, an innovation called negotiated rulemaking has been used by some administrative

agencies in an effort to prevent subsequent litigation. In a negotiated rulemaking, representatives from government, business, and nongovernmental organizations work toward agreement on proposed administrative policies (Harter, 1982). In practice, however, these agreements have not reduced subsequent litigation, in part because litigation in the US over agency rules has ordinarily occurred much less frequently than generally assumed (Coglianese, 1997). Moreover, even countries with more consensual, corporatist policy structures experience litigation over administrative issues, often because lawsuits can help outside groups penetrate close-knit policy networks (Sellers, 1995). In pluralist systems such as the US, litigation is typically viewed as a normal part of the policy process, and insiders to administrative processes tend to go to court at least as often as outsiders (Coglianese, 1996).

Overall, the impact of the judiciary on administrative governance has been and will remain a staple issue for administrative law. Empirical research on the meaning and impact of litigation in an administrative setting has the potential for informing prescriptive efforts to craft judicial principles or redesign administrative procedures in ways that contribute to more effective and legitimate governance.

Legislative and Executive Oversight

In addition to the judiciary, other governmental institutions oversee the work of government agencies and may have a significant impact on administrative governance. In the US, given its system of separate branches of government, administrative agencies find themselves on the receiving end of pressure from both legislative and executive officials. Much empirical scholarship on administrative law has investigated oversight mechanisms and how they affect behavior within administrative agencies.

An influential political economy theory treats the procedures imposed by legislative and executive overseers as mechanisms of control deployed to influence agency outcomes (McCubbins et al., 1987). According to this approach, administrative law addresses the inherent principal–agent problem confronting elected officials when they delegate power to unelected administrators. Administrators inevitably face incentives to implement statutes in ways that may stray from the goals intended by the coalition that enacted the legislation. Yet it is difficult for legislators and others to monitor agencies continually and, in any case, a law's original enactors do not remain in power forever. Elected officials therefore have good reason to create administrative procedures with the goal of entrenching the outcomes desired by the original coalition. Empirical research, however, suggests that administrative procedures provide at best only limited tools for locking in the enacting coalitions' preferences (Balla, 1998). Agencies may be less faithful to the enacting coalition's interests because they are more responsive to the politics of the moment than their institutional independence might suggest. Some analysis suggests that agencies are actually better reflective of current public preferences than are legislatures or elected executives (Stephenson, 2008).

An overarching question in research on legislative and executive oversight is whether officials from either legislative or executive bodies exert the greater degree of influence over administrative agencies. One school of thought posits legislative dominance in the oversight of US agencies, whether through the legislation they adopt, their control of agency budgets, or their ability to hold hearings or launch investigations (Weingast and Moran, 1983). Another school of thought holds that presidents exert more influence, whether through their powers to appoint the heads of agencies, direct agencies to comply with internal management and analytical requirements, or take the leading role in negotiations over agency budgets (Moe and Wilson, 1994). Given that agencies operate in a complicated political environment in which they are subject to multiple institutional constraints and pressures from both legislators and executive officials, the existing evidence seems to provide support for both schools of thought. It is clear, in other words, that both presidents and legislative officials exert influence over agencies, even if neither exercises complete control over administrative action.

One way legislatures have sought to influence agencies has been to try to direct their policymaking agendas. Not only can a legislature shape the direction of an agency by how it structures its delegation of substantive authority, but a legislature can also exert influence on the timing of administrative action. Statutes can contain deadlines for agency action, imposing a legal obligation on agencies to develop implementing rules by a specified time. Only a minority – perhaps even only a small fraction – of all regulations in the US are established under the stricture of a statutory deadline (West and Raso, 2012; Gersen and O'Connell, 2007). However, the legislature still prompts the initiation of many more administrative regulatory proceedings in the US than do executive branch officials or the courts (West and Raso, 2012). The imposition of deadlines also appears to speed up the regulatory process, at least modestly (Gersen and O'Connell, 2007).

Once an administrative agency decides to initiate a regulatory proceeding, in many jurisdictions the agency must conduct a regulatory impact analysis that will be reviewed by either a legislative or executive branch oversight body (Wiener, 2013; Radaelli and de Francesco, 2010). In the US, every president since Ronald Reagan has imposed a requirement that agencies develop regulatory impact analyses for their most significant administrative rules. Such mandated analyses must be reviewed by a White House office called the Office of Information and Regulatory Affairs (OIRA), an oversight body that has been extensively debated by administrative law scholars. The dominant theory is that presidents use the OIRA oversight process to coordinate regulatory priorities and resolve the principal–agent problem that exists between the president and those appointees the president selects to head regulatory agencies. As a normative matter, proponents of legislative supremacy decry the encroachment of presidents on the work of agencies that possess authority delegated to them by statute. Presidentialists, on the other hand, favor OIRA review as it offers a mechanism for the one official elected in a nationwide election to oversee the ongoing work of dozens of agencies that issue hundreds of important rules every year. OIRA oversight, based as it is on economic analyses that agencies prepare, has also triggered normative debate over the use of benefit–cost analysis in administrative policymaking. Advocates claim that benefit–cost analysis helps improve regulatory policy, while

opponents claim it only obfuscates decision making and delays much needed rules. Concern also exists that business interests use the OIRA process as a backdoor means of influencing regulatory policy to the detriment of achieving statutory goals or advancing the overall public interest.

Empirical researchers have been motivated by the normative debate over the OIRA process. They have documented that modern presidents have indeed sought to use OIRA review to achieve goals consistent with their policy priorities, even if these may not always comport with the results of benefit–cost analysis (Shapiro, 2005). Researchers have also shown that, in practice, OIRA review manifests itself differently across different administrations, especially in the degree to which interactions between White House and agency staff are cooperative or adversarial (West, 2006; Croley, 2003). Notwithstanding OIRA's prominence, agency staff members continue to report that they retain considerable discretion in framing and making many regulatory policy decisions, even ones formally subject to OIRA scrutiny (Bressman and Vandenbergh, 2006). Furthermore, the economic analysis produced as part of the OIRA review process appears to have much less impact on decision making than many advocates of benefit–cost analysis have hoped (Hahn and Tetlock, 2008) – but also much less of an impact in terms of delaying regulatory output as opponents of such analysis have feared (Coglianese, 2008).

Administrative policymaking occurs within a complex political and legal environment, one in which legislatures and high-level executive officials clearly play important roles. However, even major oversight entities do not possess the high degree of control that their proponents desire or their critics fear. An ongoing challenge for administrative law research remains to explain better the precise effects of legislative and executive oversight under varied conditions.

Administrative Law and Governance

Administrative law lies at several intersections, crossing the boundaries of law and politics, political theory and political science, public law, and public administration. As the body of law governing governments, the future of administrative law rests in expanding knowledge about how law and legal institutions can advance core political and social values. A concern with democratic principles will continue to dominate research in administrative law, as will interest in the role of judicial, legislative, and executive oversight in improving administrative governance. Yet administrative law can and should expand to meet new roles that government will face in the future. Ongoing efforts at deregulation and privatization may signal a renegotiation of the divisions between the public and private sectors in many countries, the results of which will undoubtedly have implications for administrative law. Administrative law also now functions in an increasingly globalized and digital world, with the emerging application of both international administrative institutions and new uses of technology that might advance both public legitimacy and policy effectiveness – or that might undermine or support administrative law institutions. No matter where the specific challenges may lie in the future, social science research on administrative law will continue to be needed to understand the operation of governmental institutions and identify ways to design rules and procedures that can potentially increase social welfare, promote the fair treatment of individuals, and expand the potential for transparent and democratic governance.

See also: Civil Law; Dispute Resolutions in Economics; Disputes, Social Construction and Transformation of; Governments; Judicial Review; Law and Democracy; Litigation; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Occupational Health and Safety, Regulation of; Public Administration: Organizational Aspects; Regulation and Administration; Rule of Law (and Rechtsstaat).

Bibliography

Balla, S., 1998. Administrative procedures and political control of the bureaucracy. American Political Science Review 92, 663–673.
Balla, S., Daniels, B., 2007. Information technology and public commenting on agency regulations. Regulation & Governance 1, 46–67.
Bressman, L.S., Vandenbergh, M.P., 2006. Inside the administrative state: a critical look at the practice of presidential control. Michigan Law Review 105, 47–99.
Chevron, v., 1984. Natural Resources Defense Council, 467 U.S. 837.
Coglianese, C., 1996. Litigating within relationships: disputes and disturbance in the regulatory process. Law & Society Review 30, 735–765.
Coglianese, C., 1997. Assessing consensus: the promise and performance of negotiated rulemaking. Duke Law Journal 46, 1255–1349.
Coglianese, C., 2006. Citizen participation in rulemaking: past, present, and future. Duke Law Journal 55, 943–968.
Coglianese, C., 2008. The rhetoric and reality of regulatory reform. Yale Journal of Regulation 25, 85–95.
Coglianese, C., 2013. Enhancing public access to online rulemaking information. Michigan Journal of Environmental and Administrative Law 2, 1–66.
Croley, S.P., 2003. White House review of agency rulemaking: an empirical investigation. University of Chicago Law Review 70, 821–885.
Edley Jr., C.F., 1990. Administrative Law: Rethinking Judicial Control of Bureaucracy. Yale University Press, New Haven, CT.
Furlong, S.R., Kerwin, C.M., 2005. Interest group participation in rulemaking: a decade of change. Journal of Public Administration Research and Theory 15, 353–370.
Gersen, J.E., O'Connell, A.J., 2007. Deadlines in administrative law. University of Pennsylvania Law Review 156, 923–990.
Hahn, R.W., Tetlock, P.C., 2008. Has economic analysis improved regulatory decisions? Journal of Economic Perspectives 22, 67–84.
Harter, P.J., 1982. Negotiating regulations: a cure for malaise. Georgetown Law Journal 71, 1–118.
Jordan, W.S., 2000. Ossification revisited: does arbitrary and capricious review significantly interfere with agency ability to achieve regulatory goals through informal rulemaking? Northwestern University Law Review 94, 393–450.
Kagan, R.A., 2003. Adversarial Legalism: The American Way of Life. Harvard University Press, Cambridge, MA.
Lodge, M., Stirton, L., 2010. Accountability in the regulatory state. In: Baldwin, R., et al. (Eds.), The Oxford Handbook of Regulation. Oxford University Press, Oxford, pp. 349–370.
Lowi, T.J., 1979. The End of Liberalism: The Second Republic of the United States. W. W. Norton, New York.
Mashaw, J.L., 1985. Prodelegation: why administrators should make political decisions. Journal of Law, Economics, and Organization 1, 81–100.
Mashaw, J.L., Harfst, D.L., 1990. The Struggle for Auto Safety. Harvard University Press, Cambridge, MA.
McCubbins, M., Noll, R., Weingast, B., 1987. Administrative procedures as instruments of political control. Journal of Law, Economics, and Organization 3, 243–277.
McGarity, T.O., 1992. Some thoughts on 'deossifying' the rulemaking process. Duke Law Journal 41, 1385–1462.
Miles, T.J., Sunstein, C.R., 2006. Do judges make regulatory policy? An empirical investigation of Chevron. University of Chicago Law Review 73, 823–881.
Moe, T.M., Wilson, S.A., 1994. Presidents and the politics of structure. Law and Contemporary Problems 57, 1–44.

O'Connell, A.J., 2008. Political cycles of rulemaking: an empirical portrait of the modern administrative state. Virginia Law Review 94, 898–986.
Radaelli, C., de Francesco, F., 2010. Regulatory impact assessment. In: Baldwin, R., et al. (Eds.), The Oxford Handbook of Regulation. Oxford University Press, Oxford, pp. 279–301.
Revesz, R.L., 1997. Environmental regulation, ideology, and the D.C. circuit. Virginia Law Review 83, 1717–1772.
Schuck, P.H., Elliott, E.D., 1990. To the Chevron station: an empirical study of federal administrative law. Duke Law Journal 1990, 984–1077.
Sellers, J.M., 1995. Litigation as a local political resource: courts in controversies over land use in France, Germany, and the United States. Law & Society Review 29, 475–516.
Shapiro, S., 2005. Unequal partners: cost–benefit analysis and executive review of regulations. Environmental Law Reporter 35, 10433–10444.
Shapiro, S., 2008. Does the amount of participation matter? Public comments, agency responses and the time to finalize a regulation. Policy Sciences 41, 33–49.
Stephenson, M.C., 2008. Optimal political control of the bureaucracy. Michigan Law Review 107, 53–110.
Stewart, R.B., 1975. The reformation of American administrative law. Harvard Law Review 88, 1667–1813.
Stigler, G.J., 1971. The theory of economic regulation. Bell Journal of Economics and Management Sciences 2, 3–21.
Weingast, B.R., Moran, M.J., 1983. Bureaucratic discretion or congressional control? Regulatory policymaking by the Federal Trade Commission. Journal of Political Economy 91, 765–800.
West, W.F., 2004. Formal procedures, informal processes, accountability, and responsiveness in bureaucratic policy making: an institutional policy analysis. Public Administration Review 64, 66–80.
West, W.F., 2006. Presidential leadership and administrative coordination: examining the theory of a unified executive. Presidential Studies Quarterly 36, 433–456.
West, W.F., Raso, C., 2012. Who shapes the rulemaking agenda? Implications for bureaucratic responsiveness and bureaucratic control. Journal of Public Administration Research and Theory 23, 495–519.
Wiener, J.B., 2013. The diffusion of regulatory oversight. In: Livermore, M.A., Revesz, R.L. (Eds.), The Globalization of Cost–benefit Analysis in Environmental Policy. Oxford University Press, Oxford, pp. 123–141.
Yackee, J.W., Yackee, S.W., 2010. Is agency rulemaking "ossified"? Testing congressional, presidential, and judicial procedural constraints. Journal of Public Administration Research and Theory 20, 261–282.
Yackee, S.W., 2005. Sweet-talking the fourth branch: the influence of interest group comments on federal agency rulemaking. Journal of Public Administration Research and Theory 16, 103–124.
Zaring, D., 2010. Reasonable agencies. Virginia Law Review 96, 2317–2379.

Relevant Websites

www.e-rulemaking.org – E-rulemaking Resources.
http://new.eur-lex.europa.eu – European Union Law.
www.oecd.org/regreform – OECD Regulatory Reform.
www.regblog.org – RegBlog.
www.regulations.gov – US Official Rulemaking Portal.
www.whitehouse.gov/omb/inforeg – US Office of Information and Regulatory Affairs.
www.reginfo.gov – US Regulatory Information Dashboard.

Adolescence, Sociology of

Monica K Johnson, Washington State University, Pullman, WA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by F.F. Furstenberg, volume 1, pp. 94–97, © 2001, Elsevier Ltd.

Abstract

The sociological study of adolescence seeks to understand it as a socially constructed stage in the life course and to explain the social factors that shape adolescents' behavior, orientations, and socioemotional well-being. As a life stage, adolescence is characterized by semi-independence. The ages at which individuals enter and exit adolescence are historically and culturally situated and the boundaries are thought to be currently shifting, lengthening the period overall. The family, schools, peer group, workplace, and neighborhood are critical institutions and contexts shaping adolescent development, though sociologists also view adolescents as actors that help create the norms and contexts in which they live. Recent scholarship reflects a growing interdisciplinary approach to the study of adolescents, bringing together social science with biological and genetic approaches to understanding human behavior.

Introduction

Psychologist G. Stanley Hall (1904) popularized the term 'adolescence' early in the twentieth century, though it did not become a major area of research for social scientists until midcentury. Early on, the study of adolescence was dominated by the study of the 'problems' of adolescence, and the period itself was viewed as inherently tumultuous. Though problem behaviors (e.g., delinquency) are still of great interest and scholars tend to characterize even other behaviors in terms of their risk (e.g., a conceptualization of sex as inherently risky or only focusing on its riskier behaviors such as contraceptive nonuse), sociologists have also extensively examined prosocial behavior (e.g., volunteerism) and a range of behaviors and orientations theorized to promote status attainment and well-being in adulthood (e.g., academic performance, occupational aspirations). Indeed, adolescence is often characterized as a critical period in the process of adult educational attainment and occupational pathways.

The sociology of adolescence is principally concerned with understanding adolescence as a life stage, as well as explaining the experiences and behaviors of adolescents as social beings. As a life stage, sociologists study cultural notions characterizing adolescence as well as the structural forces defining its nature and its boundaries with contiguous life stages: childhood and adulthood. In this way, adolescence is an interrelated set of ideas that are culturally and historically situated. With respect to the study of adolescents' experiences and behaviors, sociologists primarily examine how institutions and other social factors shape adolescents' behavioral and socioemotional well-being. The ways in which adolescents create and change the social world around them garners less attention, but is also fundamental to a sociological understanding of adolescence.

Adolescence as a Life Stage

Adolescence is generally defined as a period of the life course characterized by semidependency. Adolescents have greater autonomy than children, but not the full rights and responsibilities of adulthood (Crosnoe and Johnson, 2011). As a social construction, it is subject to change over time and varies across societies. Adolescence in the contemporary United States is largely conceived of as a time of identity development and preparation for adulthood.

The beginning of adolescence is often marked by the onset of puberty, a biological process, though the social meaning of it varies across contexts. Its end is usually marked by the assumption of adult roles, which in Western industrialized societies tends to include residential independence, completing schooling, marriage, parenthood, and full-time employment. The development of responsibility for oneself and others is also recognized as defining adulthood and tends to go hand-in-hand with role assumption, though these can occur separately (e.g., a teen parent who has not embraced the responsibilities of the parent role). The study of connections between adolescence and other life stages has been facilitated by longitudinal data collections following children into adolescence and adolescents into adulthood (e.g., Kindergarten Cohort of the Early Childhood Longitudinal Study and the National Longitudinal Study of Adolescent Health in the United States, the 1958 National Child Development Study in Great Britain).

The boundaries of adolescence, particularly its upper boundary with adulthood, are structured by formal and informal norms prescribing and proscribing behavior. Formal norms include laws delineating ages when young people become responsible for their own behavior (e.g., are criminally responsible; contracts can be enforced) can leave school, vote, drive, drink alcohol, marry, and so on. Informal norms help define adolescence as well. For example, norms against nonmarital pregnancy are stronger in relation to adolescence than adulthood, though even for adolescents there is racial/ethnic and socioeconomic status variation in norms (Mollborn, 2009).

Adolescence is also tied to and structured by social institutions that are themselves historically and culturally situated (Shanahan and Macmillan, 2008). For example, factory jobs increasingly available during industrialization allowed earlier economic independence among young people, facilitating entry into adulthood. Likewise, a strong economy and significant public spending in the post–World War II period in the United States (e.g., subsidized higher education, federal backing of home loans, etc.) enabled a much more rapid move into economic self-sufficiency for young people and accompanying movement into residential independence, marriage,

and child rearing at relatively young ages. During periods when educational investments in age-segregated institutions expanded, such as when high schools became more universal, adolescence has been extended into older ages. Similarly, age restrictions on employment and certain job duties have pushed the boundaries of adolescence upward at times.

Currently, scholars suggest that the boundaries of adolescence are again shifting. On the lower end, the age of puberty has dropped over time (Herman-Giddens, 2007). To the extent physical maturation elicits different responses from others, including greater autonomy granting by parents and expectations of 'adolescent' behavior by adults and peers, the beginning of adolescence may be dropping. Advertising trends and the advent of digital media are also topics explored by sociologists considering whether children's transition to adolescence may be shifting to younger ages. The upper boundary is in question as well. With major restructuring of the economy, including the rising wage premium to college and graduate degrees relative to lower education levels, economic self-sufficiency is taking longer to achieve (Settersten et al., 2005). Young people are staying in school longer, and are delaying marriage and full-time work. Those not pursuing higher education are often working in jobs that cannot support a family. Pathways in the transition to adulthood are becoming more diverse as well in terms of the ordering and timing of role transitions as well as involving more frequent reversals (e.g., going back to school, moving back to the parental household). Whether young people are pursuing higher education or working in low-paying jobs, their parents are often contributing substantial financial support during their 20s and into their 30s (Settersten et al., 2005). How these demographic changes observed over the past several decades have been discussed in relation to adolescence as a life stage is a key place of divergence between the fields of sociology and developmental psychology. Whereas developmental psychologists have more readily embraced recognition of a new life stage between adolescence and adulthood, referred to as 'emerging adulthood,' sociologists have characterized these changes as extending adolescence or as lengthening and diversifying the transition from adolescence to adulthood.

Social Institutions and Contexts Shaping Adolescent Development

In addition to the study of adolescence as a life stage, sociologists seek to understand adolescents' behavior, orientations, health and well-being, here referred to under the broad conceptual umbrella of development. Sociological explanations for development focus on the practices, values, and organization of other major social institutions and the larger social contexts in which adolescents live. Most research focuses on single institutions or contexts, including the family, peers, schools, work, and neighborhood. Research at the nexus of these is emerging however.

Family

A key part of adolescence involves a renegotiation of parent and child roles as young people gradually seek and are granted more autonomy. While early scholarship assumed high levels of conflict between adolescents and their parents during this period of growing independence, research consistently demonstrates that adolescents maintain close relationships with parents as they develop. Conflictual relationships remain an exception to the rule. Furthermore, while peers become a more important part of adolescents' lives, parents' behavior and parent–child relationships remain influential (Crosnoe and Cavanagh, 2010; Giordano, 2003).

Research consistently shows that adolescents' whose parents provide high levels of support and avoid harsh punishment, and who monitor their behavior, have better emotional, academic, and behavioral outcomes. These parenting practices have been shown to be beneficial across racial/ethnic groups, though recent research also suggests that what can appear to primarily White scholars as too controlling or harsh parenting may not be as harmful to minority adolescents as it is to Whites (Crosnoe and Cavanagh, 2010).

Parents' marital status is also linked to a range of adolescent behaviors, orientations, and experiences (Amato, 2010; Sweeney, 2010). Significant demographic changes in family life, including the rise of divorce, nonmarital childbearing, and cohabitation, have diversified family life in key ways. Whether parents are never-married singles, cohabiting, married, or divorced is associated with family economic resources, parental time availability, and parenting behaviors including parental support and monitoring (Amato, 2010; Sweeney, 2010). Importantly, static cross-sectional comparisons hide important differences in adolescents' longer term experiences of their parents' marital histories. Apart from marital status at a given time point, family structure instability is associated with deleterious outcomes for adolescents, including greater delinquency, lower academic performance, and instability in adolescents' own romantic relationships. Stress is thought to cumulate across family structure transitions due to changes in family composition and parenting behavior as well as disruptions in school and neighborhood relationships.

Peers

Compared with childhood, friends tend to have greater influence on behavior and orientations in adolescence, and peer relationships are more complex (Crosnoe, 2000). Friends, and peers more broadly, are theorized to influence one another by serving as examples to each other and through setting and reinforcing social norms. Isolating the effect of friends' behaviors or attitudes on adolescents has proven challenging, however, as adolescents often select friends with similar characteristics and behavior as their own. The nature of friendships, that is whether they are close, supportive, conflictual, and so on, also affects adolescent well-being (Giordano, 2003) and may serve to qualify friends' influence. As with other social relationships, stronger attachments are thought to magnify social influence.

Social network approaches to studying friend influence have focused on characteristics of friends in the network as well as friendship network properties themselves. For example, adolescents whose friends are engaged in more delinquent behavior report greater delinquency themselves, but this association is stronger in dense (closely knit) friendship networks than in less cohesive ones, and is stronger for adolescents who

hold a more central place in their network than those whose place is more peripheral (Haynie, 2001).

Peer networks are often studied through friend nominations – in which adolescents identify with whom they are and are not friends. Crowds are a larger grouping of peers that involve both friendships and acquaintances and typically contain multiple friendship networks (Crosnoe, 2000). Frank et al. (2008) also examine peer relationships by identifying clusters of students taking sets of courses that differentiate them from other students. They argue that these groups of students form a social context (which they call local positions) in which students observe, make comparisons, and transmit information and expectations, shaping their academic behavior. For girls, whose math-taking decisions are thought to be more susceptible to peer influence, levels of math taking in their school and in their local position in the baseline year of the study predicted math advancement over the following year. The effect was strongest, and also extended to boys, at the lowest math levels.

Romantic relationships in adolescence are of developmental significance as well, though scholarly attention has come much later than it did for friendships and peer networks (Giordano, 2003). New for adolescents, they generally involve greater emotionality than friendships, and the norms differ as well, including norms of exclusivity (Giordano, 2003). Despite social change in adolescents' attitudes and behaviors regarding sex, gendered norms continue to operate. Girls still pay a price in terms of peer acceptance when they have multiple sexual partners over time, whereas boys with more partners receive higher peer acceptance (Kreager and Staff, 2009). Romantic relationships can also serve as key links between peer groups. Examining romantic partner's drinking, friends' drinking, and partner's friends' drinking behavior, Kreager and Haynie (2011) find significant effects of partner's friends' behavior on dater's future drinking. Their findings support the claim that romantic relationships connect peer networks, facilitating the diffusion of behavior across them.

Whereas research on friendships tends to emphasize the positive, (Giordano, 2003) and research on peer networks considers both the positive and negative (e.g., promoting proacademic norms, encouraging risk behavior), the study of romantic relationships is dominated by a concern with risk. Studies of romantic relationships focus heavily on sex (framed as problematic), dating violence, and whether romantic relationships undermine adolescents' socioemotional well-being.

At the nexus of families and peers, studies show parents influence the nature of adolescents' friendship networks. For example, Knoester et al. (2006) find parenting practices such as selecting neighborhoods for the quality of schools, parental monitoring, and relationship quality shape the levels of problem behavior (i.e., fighting, deviance) and school-related behavior (i.e., grades, college expectations, extracurricular participation) of friends. Problem behavior in the network is lowest and prosocial behavior is highest when adolescents have high quality relationships with their parents and monitoring is relatively high. Peer support can also compensate for the deleterious effects of lower quality relationships with parents. Rodgers and Rose (2002) found the negative effects of low parental support on emotional well-being (a composite of self-esteem, depressive symptoms, and suicidal ideation) was buffered by peer support, though only in divorced, single parent families.

School

School features such as the sector (private vs public), size, racial and socioeconomic composition, level of economic and social resources, and nature of the assignment system (e.g., school choice) are linked to varying degrees with adolescent achievement, and to some extent adolescent health and socioemotional well-being. Internal organization, including level of curricular specialization and systems of tracking are of interest as well. Whereas earlier systems in the United States, tracked students into global, multiyear positions (e.g., vocational track, college-bound track), tracking is now subject-specific, 1 year at a time, and involves more mobility (Lucas, 1999). Nevertheless, informal and formal links are involved in course progression, especially in subjects such as math. Course pathways, or sequences of courses that groups of students tend to follow, are becoming of interest in the absence of identifiable and static track placements. Of primary concern across eras has been the extent to which placement within the curriculum is based on factors other than prior achievement (e.g., race), ease of mobility within the system (i.e., movement across tracks), and whether placement affects ongoing learning. Lower social class background and being Black or Hispanic are disadvantages in curricular placement, even taking into consideration prior levels of achievement (Lucas, 1999).

Another body of work focuses on issues of school climate and school culture. The climate and culture of schools can be shaped by some of the institutional factors above (e.g., size, racial composition), but are also thought to have independent effects on adolescents. Coleman's (1961) depiction of 'adolescent society' was one of the first, and certainly the most influential, studies of school culture. Coleman argued that lengthy, age-segregated systems of education isolated teenagers from adults and allowed development of a largely separate youth culture, varying some across schools, but generally at odds with mainstream adult values and norms. More recent scholarship characterizes school culture as more complex; involving mainstream as well as oppositional elements, varying in significant ways across schools, and varying in its impact on adolescents given their place in the social system of a school. The level of emphasis on academic achievement in school culture, at the heart of Coleman's concerns, is important for understanding adolescent behavior, but so too are nonacademic aspects. As an example, consider a school's cultures surrounding appropriate body size and weight management practices. High school girls' efforts to lose weight are influenced by the body sizes of other girls in their school, as well as other girls' weight control efforts. Interestingly, while the proportion of overweight and underweight girls in a school overall shapes girls' weight control behavior, efforts to lose weight at the school level are not a factor. It is only the weight control efforts of those of a similar body size that is associated with girls' efforts to lose weight (Mueller et al., 2010).

As noted above, adolescents develop within multiple settings and institutions. How schools contribute to adolescent outcomes alongside and in combination with the influence of families demonstrates this idea well. In their review of this

literature, Parcel et al. (2010) conceptualize the resources of both families and schools in three forms of capital: financial (e.g., parents' earnings, school budgets), human (e.g., parents' education levels, teacher training), and social (e.g., shared norms or social support within families or schools). They emphasize this latter form of capital often includes ties between parents and schools. Importantly, family and school resources can interact to produce adolescent outcomes, sometimes compounding (dis)advantage and other times compensating for a lack of resources in one domain. For example, when it comes to attending selective colleges, Kim and Schneider (2005) find that parents' involvement in school programs on college and financial aid matters most when parents have relatively low education levels. Institutional guidance may compensate for lack of some resources at home. Importantly, the effects of resources in more than one context are found to be additive in some studies – each important but operating independent of one another. Which resources reinforce or substitute for one another across institutions and for what outcomes continues to occupy scholars' attention.

Work

Nearly all US teenagers work during the school year at least at some point during high school, and summer employment rates exceed that of the school year. Whereas at early ages employment is often informal and involves a limited time commitment (e.g., yardwork, babysitting), it becomes increasingly formal and time intensive with age (Mortimer, 2010). Rates of employment have been on the decline recently though. An analysis of 16–17 year olds over the period from 1950 to 2010 indicates an overall downward trend in labor force participation since the late 1970s for this age group, marked by notable drops during recessionary periods that never fully recover (Smith, 2011). Employment rates by gender have been fairly comparable in recent years, but adolescents from disadvantaged minority groups are less likely to be employed than non-Hispanic White adolescents due to reduced opportunity in local labor markets and other structural factors.

The consequences of employment while attending school have been a matter of lively and ongoing debate in sociology and other related disciplines including psychology. Whereas research on families, schools, peers, and neighborhoods tends to focus on variation in structure and dynamics, the study of adolescent work is almost exclusively focused on work status and the amount of time spent working. The evidence accumulating from this effort is conflicting. On the one hand, working a limited number of hours in adolescence is associated with higher grades and positive health behaviors, and working more hours a week has been linked to higher earnings and labor force participation after high school. On the other hand, working more hours has also been linked to higher levels of substance use, problem behaviors, and poor health habits. Employment is argued by some scholars to build human capital and help prepare adolescents for adult work roles, while to others it competes with school and exposes adolescents to stress and situations for which they are developmentally unready. Recent research suggests that the risks and benefits of employment may depend on teens' reasons for working, how they use their earnings, their developmental history, and the social context (Mortimer, 2010). Importantly, efforts to address selection effects, wherein adolescents who work (and work in varying amounts) differ prior to employment, have increased substantially in the past 15 years. While the jury is still out, at least some of these studies indicate that the problems associated with adolescent employment are partially or wholly due to selection processes.

Neighborhood

Neighborhoods structure adolescents' lives physically and socially; they are the settings in which families live, they are often the basis for school assignment, and they tie adolescents and their families to other organizations (e.g., religious organizations) and opportunities (e.g., jobs for teenagers). Studies of neighborhood effects on adolescents have expanded greatly since the mid-1990s, much of it motivated by a concern with economically disadvantaged neighborhoods and their relation to risk behavior (Sampson et al., 2002). The key issue is whether and how neighborhood characteristics (e.g., rate of poverty in neighborhoods) affect residents above and beyond individual and familial characteristics (e.g., family income).

Residential stability, racial/ethnic heterogeneity, density, and other characteristics are tied to adolescent outcomes in varying ways, though studies of concentrated disadvantage (and advantage) have dominated the field (Sampson et al., 2002). For example, Harding (2003) finds that those growing up in neighborhoods with higher rates of poverty are more likely to drop out of high school, and girls are more likely to become pregnant as a teenager. Looking at neighborhoods from the other side, that of advantage, Ainsworth (2002) shows adolescents' math and reading test scores are higher when the proportion of high status adults (i.e., those with college degrees or working in managerial or professional occupations) in the neighborhood is higher.

The study of neighborhood effects also considers the overlapping context of adolescents' lives, particularly how neighborhoods and parents jointly shape adolescents. In a study of adolescent sexual behavior, Browning et al. (2005) find evidence of both family and neighborhood effects on the timing of first sex. Attachments to family delays onset of sex, and neighborhood poverty accelerates it. But key aspects of parenting and neighborhoods work in conjunction with one another. Neighborhood collective efficacy (a measure capturing social cohesion and the tendency for adults to know each other and each other's children) only delays sexual onset among adolescents who experience low place monitoring by their parents (i.e., being allowed to wander in public places without supervision). The effects of neighborhood and school collective efficacy can also depend on each other. A recent study by Kirk (2009) reports that school collective efficacy reduces the likelihood of suspension, especially when neighborhood collective efficacy is low. In contrast to the compensatory nature of the relationship for school suspension, an accentuating process was observed for likelihood of arrest with low levels of collective efficacy in both the neighborhood and school most problematic.

Harding's (2010) in-depth study of Black and Latino adolescent boys in poor as compared to working class neighborhoods also highlights how neighborhood conditions structure peer networks. Fear of violence, and in particular

violence that is structured as cross-neighborhood rivalries in the inner city, fosters cross-age interactions among adolescent males. Older boys, who have earned status by successfully navigating the dangers, are sources of protection for younger boys, but are also sources of cross-age socialization. Mostly jobless, these older males introduce alternative cultural models to younger males that exist alongside mainstream cultural models in poor neighborhoods. Adolescent boys are thus exposed to conflicting and confusing models providing information about opportunities, risks, and pathways to success in work, school, and relationships.

Adolescents as Actors

Sociologists recognize adolescents as actors in their own lives, navigating and influencing relationships, selecting contexts in which they interact, and making choices with short- and long-term goals in mind. As already noted, adolescents choose the friends that come to influence them. Adolescent employment can also be viewed as instrumental action in which adolescents invest in employment in ways that fit into their longer term plans. A strategy of steady part-time work combined with schooling, for example, appears to facilitate educational attainment in young adulthood, especially for those showing lower educational promise during high school (Mortimer, 2010). At times these social processes are framed as selection problems – something that must be dealt with in order to isolate the effects of other experiences (e.g., friend behavior, hours of work) on adolescents. Yet they are also a critical aspect of the sociological approach and at times become the central question itself.

One area in which this is clearly evident is in the study of peer culture. Adolescents enact, modify, ignore, or resist ideas and pressures from society in creating peer cultures. In her study of masculinity and sexuality, for example, Pascoe (2007) documents how high school students as well as adults and institutional practices such as school rules all work together to construct adolescent masculinity. A key practice teen boys engage in to affirm their own masculine identities is to call other boys 'fags' as insults and participate in other homophobic rituals in school settings. Various unmasculine acts can elicit the name-calling, but the terms hold racialized as well as sexualized meanings. Black boys, who are hypersexualized in the US culture, have much greater freedom to care about their clothing and appearance without risking the label. In fact, clothing is important to status seeking, particularly among poor Blacks. Dancing, too, is associated with being a fag for White boys, but for Black boys it reinforces community membership and being skilled at dancing is again associated with higher status. Black boys, who use the fag epithet less frequently themselves, get in trouble with school personnel more often when they do use it, however. These practices, alongside heterosexual locker-room talk and formal school rituals emphasizing heterosexual norms (e.g., school dances in which sexual and sexist lyrics are featured and school assemblies featuring competence in hypermasculine behaviors), define masculinities in high school.

Adolescents are not passive recipients of parenting either; their behavior and values can affect parents. Hawkins et al. (2007) demonstrate that engaged parenting among nonresident fathers is a response to, rather than a cause of, adolescent well-being. As another example, parent-adolescent value similarity is generated not only through parental socialization of offspring, but also through adolescents' influence on parents. Pinquart and Silbereisen (2004) find that adolescents' values shape their parents' values, especially when parents are more authoritative – that is, they combine high levels of support with monitoring and involve adolescents in discussion and family decision making. Through these and many other examples, sociologists explain how adolescents are key actors in the processes creating the social world around them.

Interdisciplinary and Multidisciplinary Endeavors

Emerging trends in the sociology of adolescence include bringing together insights and research methods from multiple disciplines. Advances in understanding brain functioning, for example, have helped contextualize the higher level of risk behavior of adolescents compared to children and adults (Steinberg, 2008). Two further examples, discussed in detail, illustrate the potential for combining an understanding of the social world of adolescence with that of biology and genetics.

The first example speaks to the gains made in understanding prosocial development when combining traditional social science data with biological data. Fuligni and colleagues (e.g., Fuligni and Telzer, 2013) have been examining the interaction of adolescents' family lives with biological processes in development. Their UCLA Study of Adolescents' Daily Lives collected data during high school using traditional questionnaires and daily intensive reporting of activities and experiences for 2 weeks, along with blood samples at 18 years of age and functional magnetic resonance imaging (fMRI) brain scans a year later during which participants worked through a decision-making task. They find that adolescents who report helping their families more (e.g., caring for siblings, cooking) do report feeling more burdened by family responsibilities, but also report feeling happier. Feeling like a good son or daughter and a good sister or brother, a form of role fulfillment that results from helping behavior, explains this greater happiness. Consistent with these reports, time spent helping was also associated with two biomarkers at 18 years of age (soluble interleukin 6 receptor and C-reactive protein) that are thought to result from chronic stress, but levels of these two biomarkers were lower for adolescents whose family assistance brought them greater role fulfillment.

Furthermore, the fMRI results indicated that the meaning of family helping behavior was associated with neural activity. Brain scans were taken while participants worked through a task involving decisions in which they or their family gained or lost financially. Adolescents who earlier experienced greater role fulfillment through helping the family experienced greater neural activation in regions of the brain associated with reward when they made decisions that involved personal sacrifice for their family's gain compared to when they gained personally. Hispanic adolescents, who tend to place more importance on helping their families, also demonstrated this pattern under the same scenario. In contrast, non-Hispanic Whites showed greater activation of these regions when they experienced financial gain personally. This example demonstrates that

biological response can depend on the social and cultural meaning of behavior, and suggests that health and well-being depend on a complex interaction between both.

The study of gene–environment interactions has been another growth area in the study of adolescence. Genetics may be one of the factors that aids in understanding individual variation in how social conditions shape behavior. The National Longitudinal Study of Adolescent Health (Add Health), which collected DNA data along with survey data, has greatly facilitated the study of gene–environment interactions in adolescence. Using these data, Shanahan et al. (2008) focus on the DRD2 gene that is thought to increase the likelihood of behaviors dysfunctional in school settings. Consistent with this, the study shows the gene is associated with lower rates of college enrollment among boys. Girls, they hypothesize, may be socialized in ways that damp down the expression of these behaviors even when they share the genetic risk. Among boys, high configurations of social capital (i.e., having higher socioeconomic status parents involved in a high quality school) compensated for the genetic risk of DRD2. Moreover, boys with DRD2 risk were most likely to have configurations of social capital associated with the worst college-going rates. Thus, genetic factors are linked to the likelihood of accessing the resources that would help compensate for genetic risk. Studies such as this highlight the social mechanisms involved in the relationship between genes and behavior. Scholars of adolescence will need to continue to leverage theory and data from other disciplines in order to fully understand adolescent development.

See also: Age Structure; Age, Sociology of; Alcohol Use among Young People; Delinquency, Sociology of; Friendship During Adolescence and Cultural Variations; Gay, Lesbian, and Bisexual Youth; Identity in Childhood and Adolescence; Prosocial Behavior During Adolescence; School Environment for LGBTQ/Sexual Minority Youth; Social Competence During Adolescence Across Cultures; Transition to Adulthood; Youth Culture, Sociology of; Youth Gangs; Youth Movements.

Bibliography

Ainsworth, J.W., 2002. Why does it take a village? The mediation of neighborhood effects on educational achievement. Social Forces 81, 117–152.
Amato, P.R., 2010. Research on divorce: continuing trends and new developments. Journal of Marriage and Family 72, 650–666.
Browning, C.R., Leventhal, T., Brooks-Gunn, J., 2005. Sexual initiation in early adolescence: the nexus of parental and community control. American Sociological Review 70, 758–778.
Coleman, J.S., 1961. The Adolescent Society: The Social Life of the Teenager and Its Impact on Education. The Free Press, New York.
Crosnoe, R., 2000. Friendships in childhood and adolescence: the life course and new directions. Social Psychology Quarterly 63, 77–391.
Crosnoe, R., Cavanagh, S.E., 2010. Families with children and adolescents: a review, critique, and future agenda. Journal of Marriage and Family 72, 594–611.
Crosnoe, R., Johnson, M.K., 2011. Research on adolescence in the 21st century. Annual Review of Sociology 37, 439–460.
Frank, K.A., Schiller, K.S., Riegle-Crumb, C., Mueller, A.S., Crosnoe, R., Pearson, J., Muller, C., 2008. The social dynamics of mathematics coursetaking in high school. American Journal of Sociology 113, 1645–1696.
Fuligni, A.J., Telzer, E.H., 2013. Another way family can get in the head and under the skin: the neurobiology of helping the family. Child Development Perspectives 7, 138–142.
Giordano, P.C., 2003. Relationships in adolescence. Annual Review of Sociology 29, 257–281.
Hall, G.S., 1904. Adolescence. Appleton, New York.
Harding, D.J., 2003. Counterfactual models of neighborhood effects: the effect of neighborhood poverty on dropping out and teenage pregnancy. American Journal of Sociology 109, 676–719.
Harding, D.J., 2010. Living the Drama: Community, Conflict and Culture Among Inner-City Boys. University Of Chicago Press, Chicago, IL.
Hawkins, D., Amato, P., King, V., 2007. Nonresident father involvement and adolescent well-being: father effects or child effects? American Sociological Review 72, 990–1010.
Haynie, D.L., 2001. Delinquent peers revisited: does network structure matter? American Journal of Sociology 106, 1013–1057.
Herman-Giddens, M.E., 2007. The decline in the age of menarche in the United States: should we be concerned? Journal of Adolescent Health 40, 201–203.
Kim, D.H., Schneider, B., 2005. Social capital in action: alignment of parental support in adolescents' transition to postsecondary education. Social Forces 84, 1181–1206.
Kirk, D.S., 2009. Unraveling the contextual effects on student suspension and juvenile arrest: the independent and interdependent influences of school, neighborhood, and family social controls. Criminology 47, 479–520.
Knoester, C., Haynie, D.L., Stephens, C.M., 2006. Parenting practices and adolescents' friendship networks. Journal of Marriage and Family 68, 1247–1260.
Kreager, D.A., Haynie, D.L., 2011. Dangerous liaisons? Dating and drinking diffusion in adolescent peer networks. American Sociological Review 76, 737–763.
Kreager, D.A., Staff, J., 2009. The sexual double standard and adolescent peer acceptance. Social Psychology Quarterly 72, 143–164.
Lucas, S.R., 1999. Tracking Inequality: Stratification and Mobility in American High Schools, Sociology of Education Series. Teachers College Press, New York.
Mortimer, J.T., 2010. The benefits and risks of adolescent employment. Prevention Researcher 17, 8–11.
Mollborn, S., 2009. Norms about nonmarital pregnancy and willingness to provide resources to unwed parents. Journal of Marriage and Family 71, 122–134.
Mueller, A.S., Pearson, J., Muller, C., Frank, K., Turner, A., 2010. Sizing up peers: adolescent girls' weight control and social comparison in the school context. Journal of Health and Social Behavior 51, 64–78.
Parcel, T.L., Dufur, M.J., Cornell Zito, R., 2010. Capital at home and at school: a review and synthesis. Journal of Marriage and Family 72, 828–846.
Pascoe, C.J., 2007. Dude, You're a Fag: Masculinity and Sexuality in High School, first ed. University of California Press, Berkley, CA.
Pinquart, M., Silbereisen, R.K., 2004. Transmission of values from adolescents to their parents: the role of value content and authoritative parenting. Adolescence 39, 83–100.
Rodgers, K.B., Rose, H.A., 2002. Risk and resiliency factors among adolescents who experience marital transitions. Journal of Marriage and Family 64, 1024–1037.
Sampson, R.J., Morenoff, J.D., Gannon-Rowley, T., 2002. Assessing "neighborhood effects": social processes and new directions in research. Annual Review of Sociology 28, 443–478.
Settersten Jr., R.A., Furstenberg Jr., F.F., Rumbaut, R.G., 2005. On the Frontier of Adulthood: Theory, Research and Public Policy. University of Chicago Press, Chicago, IL.
Shanahan, M.J., Macmillan, R., 2008. Biography and the Sociological Imagination: Contexts and Contingencies. W.W. Norton & Co, New York.
Shanahan, M.J., Vaisey, S., Erickson, L.D., Smolen, A., 2008. Environmental contingencies and genetic propensities: social capital, educational continuation, and dopamine receptor gene DRD2. American Journal of Sociology 114, S260–S286.
Smith, C.L., 2011. Polarization, Immigration, Education: What's Behind the Dramatic Decline in Youth Employment? Working Paper in the Finance and Economics Discussion Series. Divisions of Research and Statistics and Monetary Affairs, Federal Reserve Board, Washington, DC.
Steinberg, L.D., 2008. A social neuroscience perspective on adolescent risk taking. Developmental Review 28, 78–106.
Sweeney, M.M., 2007. Stepfather families and the emotional well-being of adolescents. Journal of Health and Social Behavior 48, 33–49.
Sweeney, M.M., 2010. Remarriage and stepfamilies: strategic sites for family scholarship in the 21st Century. Journal of Marriage and Family 72, 667–684.

Relevant Websites

http://www.asanet.org – American Sociological Association.
http://www.s-r-a.org – Society for Research on Adolescence.

Adolescent Health and Health Behaviors

Thomas A Wills and Rebecca Knight, University of Hawaii Cancer Center, Honolulu, HI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by T.A. Wills, volume 1, pp. 105–112, © 2001, Elsevier Ltd.

Abstract

This article considers adolescent health behavior with a focus on the major contributors to mortality in adolescence and young adulthood. We consider substance use, sexual risk behavior, and violence as well as suicide risk. These various risk behaviors tend to occur together and have similar though not identical predictors. Theoretical approaches to vulnerability or protection for these outcomes include attitudinal models, social influence and social perception models, stress-coping models, and recent dual-process models. This article discusses how risk may derive from contributions of environmental variables, family variables, personality characteristics, life stress, and peer social networks. Directions for future research are considered briefly.

This article considers adolescent health behavior with a focus on the major contributors to mortality in adolescence and young adulthood. We consider substance use, sexual risk behavior, and violence as well as suicide risk. These various risk behaviors tend to occur together and have similar though not identical predictors. Theoretical approaches to vulnerability or protection for these outcomes include attitudinal models, social influence and social perception models, stress-coping models, and recent dual-process models. This article discusses how risk may derive from contributions of environmental variables, family variables, personality characteristics, life stress, and peer social networks. Directions for future research are considered briefly.

Definition of Adolescent Health Behavior

This topic comprises two related areas. One area concerns behaviors that may create risk for immediate threats to health during adolescence (e.g., accidents), the other concerns behaviors that place individuals at increased long-term risk for chronic diseases in adulthood (e.g., cancer). Research on adolescent health behavior aims to determine factors related to increased risk for health-compromising behaviors such as cigarette smoking (i.e., risk factors). Protective factors such as family support operate to decrease risk for these behaviors in general and, in addition, may reduce the impact of risk factors (i.e., buffering effects). This article considers factors related to substance use, sexual behavior, violence, and suicide risk (*see* Health Behaviors).

Intellectual Context

The study of adolescent health behavior arose from epidemiological research during the 1950s and 1960s demonstrating that mortality from chronic disease in adulthood, such as heart attack or cancer, was related to factors such as substance use, dietary patterns, and life stress. Studies on longitudinal tracking of behavioral and physiological risk factors gave recognition to the concept that risk status began to develop before adulthood. This suggested a focus on studying health-related behaviors such as cigarette smoking at younger ages so as to determine whether there was empirical support for early preventive approaches.

In the present article our focus is guided by data on causes of mortality. The United States data for 2010 from the Centers for Disease Control show that for adolescents (persons of 15–24 years of age) and young adults (25–34 years of age) the most common causes of mortality in order of frequency are accidents, homicide, and suicide. Cardiovascular disease and cancer only assume prominence at later ages. Though the relative rates for different causes could vary across countries, these statistics draw attention to the fact that accidents, interpersonal violence, and suicide are the leading causes of mortality during adolescence. It is noteworthy that substance use contributes to all these types of mortality. For example, alcohol use is a factor in the majority of fatalities from auto accidents (Taylor et al., 2010) and toxicology studies show that drug use is implicated in both homicide and suicide (Darke, 2010; Sheehan et al., 2013). Thus, inquiring into factors related to substance use is important for understanding the origins of these outcomes.

The factors that place adolescents at risk for substance use and sexual risk behavior span several domains. While demographic and environmental factors may have a role in risk behavior (Gardner et al., 2010), personality factors that contribute to emotional states such as depression and anger are also implicated in risk for substance use and violence (Colder et al., 2010). Sexual risk behavior (e.g., unprotected intercourse, multiple partners), a factor in HIV infection and other sexually transmitted infections, is related to many of the factors that predict adolescent substance use (Salazar et al., 2009). The article focuses on research relevant to these conditions, noting that drug use and sexual risk behavior tend to occur together and personality variables may contribute to their cooccurrence (Cooper, 2006; Wills and Ainette, 2010).

Dominant Theories and Changes in Emphasis over Time

Early Theoretical Models

Cognitive approaches such as the *health belief model*, developed by Rosenstock and Becker, posit that health behavior is based on a reasoned analysis of the perceived benefits of the

behavior, the costs (health, social, or economic) associated with the behavior, and the perceived barriers to engaging in the behavior (*see* Health Behaviour, Psychosocial Theories of). Though these rational models of health are intuitively plausible, knowledge about health consequences has not eliminated cigarette smoking and measures of health knowledge often do not correlate well with behavior. Attitudinal approaches such as the *theory of reasoned action*, developed by Fishbein and Ajzen, posited that favorable attitudes lead to intentions to perform a given behavior, and intentions are then related to occurrence of the behavior. Empirical studies have provided some support for this model but a number of studies have shown the relation between intentions and behavior to be only moderate, and questions have been raised as to how cultural differences in values can limit the global translation of prevention programs based on Western attitudes toward health (DiClemente et al., 2009).

Subsequent theories include the *problem behavior model*, developed by Richard and Shirley Jessor (see Jessor, 1998). This model construes adolescent substance use as a behavior linked to rejection of conventional societal norms for achievement and behavior. This model proposes that tolerance for deviance is produced by factors such as a poor relationship with parents, alienation from conventional routes to achievement (e.g., academic performance), and affiliation with peers who are engaging in deviant behaviors such as fighting, stealing, and substance use. The problem behavior model predicts that individuals who reject conventional values will rebel through adopting multiple deviant behaviors (e.g., heavy drinking, marijuana use, precocious sexuality). This model was the first to account for the observed intercorrelation of various problem behaviors, and prevention programs typically include an aim to reduce deviance-prone attitudes and affiliation with antisocial peers (Smith et al., 2006). The *social influence model* proposes that many behaviors are acquired through observing and modeling the behavior of influential others, and delineates the role of affiliation with peer smokers or drinkers in facilitating the adoption of these behaviors (Hoffman et al., 2006). This approach has been successful in demonstrating that peer affiliations are a final gateway factor for the effects of other variables (Wills et al., 2011) and this model has provided the rationale in prevention programs for teaching skills for resisting peer pressure.

Changes in Focus over Time

Over the last decade, the range of variables shown to predict substance use has led to statistical models emphasizing how risk or protection arises through an interplay of contributions from individual, environmental, and social factors (Wills and Ainette, 2010). Thus, recent research on adolescent substance use has used multivariate analyses that test multiple domains of predictors. One specific change in focus is the conceptualization of social influence processes. Previously it was assumed that adolescent smoking, for example, was caused by explicit pressure applied by smoking peers to unwilling (or ambivalent) targets. However, current research has demonstrated that adolescents' perceptions about the prevalence of peer smoking and drinking are often biased upward and teens perceive more acceptability of substance use than peers actually endorse (Hoffman et al., 2006). These cognitive aspects of social perception make some individuals more likely to adopt smoking, alcohol use, or sexual risk behavior. This knowledge has been incorporated in prevention programs with components to correct erroneous normative perceptions and focus on negative perceptions of teen substance users (Sussman and Ames, 2008).

Recent research has given more prominence to affective factors, as evidence shows that problem behavior tends to occur among adolescents experiencing higher levels of stress or anger (Sinha, 2008). The observation that characteristics measurable in childhood (e.g., irritability, conscientiousness) predict health-related behavior at later ages has led to theories aimed at understanding how early temperament attributes are related to the development of problems with behavioral and emotional regulation at later ages (Wills and Ainette, 2010). Though relations of different aspects of emotion to risk behavior can be complex (Colder et al., 2010), current theories have drawn out the linkages of emotional distress to patterns of peer affiliations and willingness to use substances if an opportunity presents itself. In addition, recent research has shown that difficulty in controlling emotions and behavior is a risk factor for several types of health behavior (Wills et al., 2013a) and has linked poor emotional regulation to suicide risk during adolescence (O'Connor et al., 2012; Wagner, 2009).

Finally, recent theoretical developments have recognized that two different systems are involved in processing information relevant for health behavior. One system is based on deliberative processing of information, with systematic consideration of risks and benefits; the second approach is based more on judgment heuristics and reactions to images. These two modes of processing – sometimes termed the *reflective* and *reactive* systems – have different properties, the first being slower, more controlled, and requiring cognitive effort, whereas the second is quicker, more impulsive, and more automatic in its operation (Gerrard et al., 2008). These concepts have been incorporated in *dual-process theory*, which posits that the reflective and reactive systems make independent contributions to outcomes, so assessing the tendency to use each of these systems may enable better prediction of health risk behaviors (Wills et al., 2013a).

Emphases in Current Research

This section summarizes current knowledge about variables that are related to adolescent health behaviors. Listing a variable as a predictive factor does not indicate that it is the sole cause or even necessarily a strong cause of a behavior. A given variable may not be a major predictor of a behavior from a statistical standpoint, but a combination of variables can have strong predictive power. Also, buffering effects are common in health behavior. For example, a person with high life stress might show few adverse health outcomes because he/she also had a high level of family support. Thus there is no single magic bullet that can be used to tell who is at high risk and who is not. To predict health behavior with confidence, one has to assess multiple variables and consider the balance between levels of risk factors and protective factors. We emphasize also that

there are intercorrelations among the various risk behaviors and a substantial degree of commonality in their predictors. Separate lists of predictors for substance use, injury and violence, and sexual risk behavior would have considerable overlap (DiClemente et al., 2006; Sussman and Ames, 2008; Salazar et al., 2009; Wills and Ainette, 2010). So in the following section we simply give examples of predictors for each of the behaviors.

Demographic Variables

Substance use among US adolescents varies by gender (males typically showing higher rates by 12th grade), ethnicity (higher rates among Caucasians, lower rates among African-Americans), family structure (higher rates in single-parent families), religiosity (lower rates among persons who attend a religious organization), and parental socioeconomic status, with generally higher rates among adolescents from families with lower income and education (Johnston et al., 2013). Socioeconomic effects however are complex, with different patterns for different indices of substance use (e.g., overall alcohol use versus heavy drinking). It is typically found that effects for demographic variables are mediated through risk and protective factors listed subsequently (e.g., academic involvement). We note that demographic effects observed in the US data could differ across countries, so reference to local data is always warranted.

Environmental Variables

The available data suggest that there is elevated risk for substance use in neighborhoods with lower income and higher crime rates, and where the neighborhood is perceived by residents as dangerous and/or neglected by the local government (Gardner et al., 2010). Noting the existence of a relationship, however, fails to characterize the great variability in effects of environmental variables. For example, a large proportion of persons growing up in poverty areas may go through adolescence showing little or no substance use. This is believed to occur because family processes provide protective effects that can offset the risk potential of the environment (Wills et al., 2013b).

Temperament Dimensions

Temperament dimensions evident from early ages such as attentional orientation, the tendency to focus attention on a task, and positive emotionality, the tendency to laugh and smile frequently, are indicated as protective factors. Indicated as risk factors are activity level, the tendency to move around frequently and become restless when sitting still, and negative emotionality, the tendency to be easily irritated and become intensely upset. It appears that these variables act through affecting the development of generalized self-control ability (Wills and Ainette, 2010). Two other temperament-type constructs have been related to substance use, drunken driving, and sexual behavior in adolescence. Novelty seeking reflects a tendency to need new stimuli and situations frequently and to become bored easily. Sensation seeking is a related construct in which high scorers prefer intense sensations (e.g., loud music) and show a preference for engaging in exciting and dangerous activities (Sussman and Ames, 2008).

Family Relationships

A positive relationship with parents is an important protective factor for several health risk behaviors. High supportiveness is present when adolescents feel that they can talk freely with a parent when they have a problem and that parents will provide emotional support or practical assistance when it is needed. Family support is consistently observed to have buffering effects, reducing the impact of risk factors such as negative life events (Wills et al., 2013b). Family support acts through multiple pathways, associated among adolescents with better self-control, higher academic competence, and less affiliation with deviance-prone peers.

A conflictual relationship with parents, involving disagreements and frequent arguments, is a risk factor for various adolescent problem behaviors. It should be noted that some argumentation between parents and children is normal, as teenagers establish autonomy from their parents and work toward their own identity; but a high level of conflict, in the absence of protective factors, is potentially problematic. Young persons who feel rejected by their parents look for other sources of acceptance and approval and tend to gravitate into groups of deviance-prone peers, which can lead to drug use and risk-taking behavior (Jessor, 1998). A high level of arguments with parents can be a significant risk factor and ameliorating family conflict is an important focus for counseling and prevention programs (Smith et al., 2006).

Self-Control and Dysregulation

Behavioral self-control, also termed planfulness, is measured by attributes involved in planning, organizing, and monitoring behavior and dealing with problems through considering alternatives and making a decision about how to solve the problem. Emotional self-control involves soothability, the ability to calm down when one is excited or upset, together with ability to regulate emotional states such as anger and sadness. Self-control is a protective factor for substance use through helping to promote academic competence and ensuring that problems get resolved rather than accumulating over time and producing high life stress (Wills et al., 2011).

Behavioral dysregulation, also termed impulsivity or disinhibition, is manifested as impatience, impulsiveness, and present orientation – the tendency to be reactive to present situational cues without much thought to the future. Emotional dysregulation can manifest as irritability (a sample item being, "There are a lot of things that annoy me"); affective lability, the tendency to change moods quickly; and rumination, a tendency to dwell for long periods of time on remembered episodes of anger or sadness. Dysregulation is related to lower involvement in school and seems to be a risk factor for bringing on negative life events. In later adolescence, poor regulation is associated with perceiving substance use or sexual behavior as useful for coping with life stress, an important factor in high-risk behavior (Cooper, 2006; Wills et al., 2011). Also, self-control buffers the effect of dysregulation, so an individual's level of poor control is less predictive than the

balance of self-control and dysregulation (Wills and Ainette, 2010).

Aggressiveness and Hostility

Whether aggressiveness derives from temperamental attributes (e.g., irritability), is a facet of behavioral and emotional dysregulation, or is a consequence of adverse social conditions (e.g., school failure) could be argued, but in a sense all these conceptualizations are correct. The empirical fact is that verbal or physical aggressiveness is stable over time from childhood onwards, leads to difficulties with parental socialization and peer social relationships, and is a strong correlate of substance use and other problem behaviors. A predisposition to respond aggressively in interpersonal situations is conducive to injury through violent encounters, making this an important factor in adolescents' injuries as well as injury to others (DiClemente et al., 2006). Overt physical aggression is part of a syndrome that also involves impulsiveness and negative affect, and diagnostic studies show that a substantial proportion of youth with antisocial behavior also have high levels of negative mood. The combination of high levels of aggression and depression is a particular risk factor for both substance abuse and suicide in adolescents. However, clinical-level disorder is not a necessary condition for risk, as simple measures of anger predict substance use in adolescence quite well (Wills et al., 2011).

Academic Involvement or Disengagement

Interest in academics and good performance in school is an important protective factor among adolescents, whereas disengagement from school is a notable risk factor for adolescent substance use and other problem behaviors. Disengagement may be reflected in negative attitudes toward school, low grades, and a history of discipline in school. The effect of academic involvement is independent of characteristics such as socioeconomic status and family structure, though it is related to these to some extent (Wills et al., 2011, 2013a). Low involvement in school could be partly attributable to temperamental restlessness or distractibility, which makes it difficult to adjust to the classroom setting (Wills and Ainette, 2010). Disinterest in getting good grades may also derive from a conflictual family that does not socialize children to work toward long-term goals or a social environment that devalues conventional routes to achievement (Jessor, 1998). Adolescents may have one bad year in school and do better in subsequent years, without adverse effect; but a trajectory of deteriorating academic performance could be predictive of subsequent problems such as substance abuse.

Life Events, Emotional Distress, and Coping Motives

An accumulation of negative life events during the previous year has been implicated in substance use and in suicide risk (Sinha, 2008; Wagner, 2009). The life events may be ones that occur to a family member (e.g., unemployment of a parent) or ones that directly involve the adolescent him/herself (e.g., loss of a friend). High life stress has been shown related to increased affiliation with deviant peers, apparently because experiences of failure and rejection dispose the adolescent to disengage from conventional institutions and spend more time with peers who are themselves frustrated and alienated. Life stress also may prime a need for affect regulation mechanisms to deal with the emotional consequences of stressors, hence it is related to coping motives for substance use, a predictor of substance use problems (Wills et al., 2011).

Negative emotional states including depression and anxiety have been linked to adolescent substance use and other problem behavior (Colder et al., 2010). The source of the negative affect may be from recent negative events, from dispositional characteristics (e.g., neuroticism), and/or from living in an unsafe environment; some current evidence supports each of these perspectives. An individual reporting a high level of depressive affect on a symptom checklist is also likely to endorse beliefs that he/she is an unattractive and unworthy person, that current problems are uncontrollable, and that there is no clear purpose or meaning in their present life. Evidence has linked components of this complex (lack of control, pessimism, and perceived meaninglessness) to adolescent substance use and suicide risk (Wagner, 2009). Note that positive affect (which is not simply the absence of negative affect) is a protective factor for various problem behaviors and has buffering effects, reducing the impact of negative affect on substance use (Colder et al., 2010). Thus it is important to assess the balance of positive and negative affect for an individual.

Persons may engage in a given behavior for different reasons, and the reasons have significant implications for adolescent health behavior. Problematic substance use and sexual behavior are prominent particularly among individuals for whom these behaviors are regarded as an important coping mechanism, perceived as useful for coping with negative affect and reducing stress (Cooper, 2006). Coping (as opposed to social) motives distinguish adolescents who use tobacco and alcohol at relatively low rates from those who use multiple substances at high rates and experience negative consequences because of inappropriate use (Wills et al., 2011). Coping (as opposed to social) motives distinguish adolescents who use multiple substances at high rates and experience negative consequences because of inappropriate use from those who use tobacco or alcohol infrequently or not at all (Wills et al., 2011).

Specific Attitudes and Efficacies

Persons who perceive a problem behavior as relatively accepted in their social circle, and/or perceive they are less vulnerable to harmful consequences of use, are more likely to engage in the behavior. Attitudes do not have to be totally favorable in order to create risk status; for example, perceptions of cigarette smokers tend to be negative in the general adolescent population, but those who have relatively less negative perceptions are more likely to begin smoking (Gerrard et al., 2008). It is likely that some variance is attributable to attitudes about substance use communicated by parents (Wills et al., 2013). Influential peers certainly communicate norms about substance use, which tend to differ from those held by parents, and mass media may also communicate favorable images about substance use and sexual behavior (Wills et al., 2013, 2008). Ongoing studies

conducted among US teenagers have shown that rates of substance use vary inversely with attitudes in the population about the harmful effects of cigarettes or marijuana (Johnston et al., 2013). The source of these trends has not been decisively linked to any specific source, but school-based preventive programs and government-sponsored counteradvertising have been suggested as influential (*see* Health Risk Perception).

Specific efficacies are relevant for averting risk behavior. In the area of sexual behavior and contraceptive use, perceived efficacy for discussing contraceptive use with a partner is an important protective factor (Salazar et al., 2009). There are widely varying attitudes about condom use and differences across persons in the degree to which they feel comfortable in communicating with partners about condom use. Thus studies should always make an attempt to address specific beliefs and attitudes about a health behavior. High resistance efficacy, the belief that one can successfully deal with situations that involve temptation for a behavior (e.g., being offered a cigarette at a party), is consistently noted to be a protective factor. Some data have shown efficacy linked to a good parent–child relationship, to good behavioral self-control, and (inversely) to indices of low self-regard (Wills et al., 2007).

Peer Relationships

Across a large number of studies it has been noted that adolescents who smoke, drink, fight, and/or engage in sexual risk behavior tend to have friends who do likewise (e.g., Hoffman et al., 2006). Thus the frequency and extent of peer use is a major consideration for risk status. For example, many teenagers will end up at times in situations where a friend is using some substance, but when many of a person's friends are smoking and drinking frequently and perhaps feeding a growing cycle of alienated beliefs, then concern would mount. Aspects of the context of peer group membership should be considered. While engagement in peer–group activity is normative and desirable for adolescents, it is when a person has high support from peers and low support from parents that substance use is particularly elevated (Wills et al., 2013b). In addition there may be several types of peer networks in a given school, including groups that are academic or athletically oriented as well as subgroups focused around deviant interests (e.g., substance use); hence the frequency of peer activity may be less important than the types of peers and their associated behaviors (Sussman and Ames, 2008). Responding to peer behavior has been a primary focus in prevention programs, which aim to teach skills for communicating well with peers and responding assertively in situations where an opportunity for a problem behavior arises.

Racial/Ethnic Discrimination

In recent years it has been recognized that racial/ethnic discrimination is a significant factor in poor health status. Studies with adolescents are showing that perceived discrimination is related to cigarette smoking and heavy alcohol use as well as to elevations in emotional distress, and discrimination has been linked to cardiovascular disease and other adverse health outcomes in adulthood (Williams and Mohammed, 2009). Explanations advanced for the effect of discrimination on substance use include feelings of resentment and anger and disengagement from conventional institutions (e.g., school). Current studies are examining how discrimination operates to erode self-control and hence make adolescents more vulnerable to a range of disease conditions (Gibbons et al., 2012). However, buffering factors such as strong family support may help to counteract the adverse effects of racism (Wills et al., 2013b).

Obesity, Dietary Intake, and Physical Exercise

A rising health concern globally has been the substantial increase in the prevalence of overweight and obese children and adolescents (Spruijt-Metz, 2011). This health problem has arisen in the United States as well as European countries (Janssen et al., 2005). The public health issues linked to adolescent obesity include adolescent health problems such as type II diabetes and increased risk for cardiovascular disease and some types of cancer in adulthood. Stigmatization and harassment of overweight youth also have significant negative consequences in adolescence (Bucchianeri et al., 2013). There is no single cause for the increased prevalence of obesity, although dietary intake is quite relevant. Dietary patterns have changed substantially in the past few decades. There has been a rise in consumption of soft drinks and candy and a decline of fruits and vegetables in adolescents' diets in many countries, and these have been implicated in the increase in childhood and adolescent obesity (Spruijt-Metz, 2011). In addition to educating children and adults about the causes of obesity, goals of current prevention programs include changing shopping environments so as to increase access to affordable healthy foods, while policy initiatives discourage consumption of sugary drinks and aggressive marketing of children's meals that are high in calories and fat (Spruijt-Metz, 2011).

In the past few decades physical activity patterns in adolescents have also changed. There has been an increase in time spent watching television and using computers, and a decrease in opportunities for physical activity in schools and communities (Iannotti and Wang, 2013). The patterns of physical activity also vary across countries (Janssen et al., 2005). Low physical activity in adolescents is linked concurrently to overweight and lower cognitive performance, and on a long-term basis it has been connected to parameters of the metabolic syndrome, which creates risk for cardiovascular disease (Spruijt-Metz, 2011). While screen-based sedentary behavior (particularly television viewing) has been linked to subsequent adiposity, low physical activity has also been related to adverse psychological outcomes including lower quality of life and quality of family relationships (Iannotti and Wang, 2013). It is important that national governments and international agencies promote physical exercise, provide neighborhood environments that make physical activity easy and safe, and educate the public about the interrelation of exercise, diet, and weight. Furthermore, more research is needed on predictors of high body mass index and low physical activity, in order to understand the factors associated with them.

Future Directions

This article has emphasized that meaningful risk is not predictable from knowledge about a single variable; rather it is the

number of risk factors and their balance with protective factors that is most informative. The concept that health behavior is related to variables at several levels of analysis (environmental variables, personality variables, family variables, and social variables) has been emphasized, and the article has discussed mediating pathways through which these variables operate to produce favorable versus unfavorable health outcomes.

Several future developments can be anticipated in this area. One is increasing use of theories that delineate how different domains of variables are related to problem behavior. For example, epigenetic models suggest how simple temperament characteristics are related over time to patterns of family interaction, coping, and social relationships, which are proximal factors for adolescent substance use and other behaviors (Wills and Ainette, 2010). Research of this type will be increasingly interdisciplinary, involving investigators with expertise in behavior genetics, developmental and social psychology, as well as anthropology and neurobiology.

Another development is the increasing integration of genetic research with psychosocial research. It has been known for some time that parameters relevant for cardiovascular disease (e.g., blood pressure and obesity) have a substantial heritable component, and recent research has shown substantial genetic contributions for liability to cigarette smoking and alcohol abuse/dependence (Korhonen et al., 2012). Although health-related variables have consistently been shown related to genetic characteristics, there is less understanding of the physiological pathways involved. Current investigations are studying genes coding for receptors for neurotransmitters that have been linked to vulnerability to substance abuse and suicide and identifying physiological and behavioral pathways for the effects of genetic variation.

Finally, life span research has indicated that simple temperament characteristics measured at early ages predict health-related outcomes over considerable periods of time (Wills and Ainette, 2010). Such research suggests investigations into whether early temperament characteristics are related directly to physiological pathways from the hypothalamic–pituitary axis, operating to dysregulate metabolic systems so as to create risk for cardiovascular disease and diabetes. Behavioral pathways are also possible in observed longevity effects through liability for smoking or accident-proneness. Integrative research is suggested, using concepts from physiology and behavioral psychology to understand the role of psychosocial processes in premature mortality.

See also: Adolescence, Sociology of; Alcohol Use among Young People; Childhood and Adolescence: Developmental Assets; Coping across the Lifespan; Drug Use and Abuse: Psychosocial Aspects; Health Behavior, Psychosocial Theories of; Health Behaviors; Health Education and Health Promotion; Health Promotion in Schools; Self-Efficacy and Health; Self-Efficacy: Education Aspects; Self-Efficacy.

Bibliography

Bucchianeri, M.M., Eisenberg, M.E., Neumark-Sztainer, D., 2013. Weightism, racism, classism, and sexism: shared forms of harassment in adolescents. Journal of Adolescent Health 53, 47–53.

Colder, C.R., Chassin, L., Lee, M.R., Villalta, A.K., 2010. Developmental perspectives: affect and adolescent substance use. In: Kassel, J.D. (Ed.), Substance Abuse and Emotion. American Psychological Association, Washington, DC, pp. 109–135.

Cooper, M.L., 2006. Does drinking promote risky sexual behavior? A complex answer to a simple question. Current Directions in Psychological Science 15, 19–23.

Darke, S., 2010. The toxicology of homicide offenders and victims: a review. Drug and Alcohol Review 29, 202–215.

DiClemente, R.J., Crosby, R.A., Kegler, M.C., 2009. Issues and challenges in applying theory in health promotion research: adaptation, translation, and global application. In: DiClemente, R.J., Crosby, R.A., Kegler, M.C. (Eds.), Emerging Theories in Health Promotion Practice and Research, second ed. Jossey-Bass, San Francisco, CA, pp. 551–568.

DiClemente, R.J., Gielen, A.C., Sleet, D.A., 2006. Behavioral sciences, injury, and violence prevention. In: Gielen, A., Sleet, A.D., DiClemente, R.J. (Eds.), Injury and Violence Prevention: Behavioral Science Theories and Methods. Jossey-Bass, San Francisco, CA, pp. 485–499.

Gardner, M., Barajas, R.G., Brooks-Gunn, J., 2010. Neighborhood influences on substance use etiology. In: Scheier, L.M. (Ed.), Handbook of Drug Use Etiology: Theory, Methods, and Empirical Findings. American Psychological Association, Washington, DC, pp. 423–441.

Gerrard, M., Gibbons, F.X., Houlihan, A.E., Stock, M.L., Pomery, E.A., 2008. A dual-process approach to health risk decision making. Developmental Review 28, 29–61.

Gibbons, F.X., O'Hara, R.E., Stock, M.L., Gerrard, M., Weng, C.-Y., Wills, T.A., 2012. The erosive effects of racism: reduced self-control mediates the relation between racial discrimination and substance use in African American adolescents. Journal of Personality and Social Psychology 102, 1089–1104.

Hoffman, B.R., Sussman, S., Unger, J., Valente, T.W., 2006. Peer influences on adolescent cigarette smoking: a theoretical review. Substance Use and Misuse 41, 103–155.

Iannotti, R.J., Wang, J., 2013. Patterns of physical activity, sedentary behavior, and diet in US adolescents. Journal of Adolescent Health 53, 280–286.

Janssen, I., Katzmarzyk, P., Boyce, W., Vereecken, C., Mulvihill, C., Pickett, W., 2005. Comparison of overweight and obesity prevalence in school-aged youth in 34 countries and their relationships with physical activity and dietary patterns. Obesity Reviews 6, 123–132.

Jessor, R. (Ed.), 1998. New Perspectives on Adolescent Risk Behavior. Cambridge University Press, Cambridge, New York.

Johnston, L.D., O'Malley, P.M., Bachman, J.G., Schulenberg, J.E., 2013. Monitoring the Future National Survey Results on Drug Use, 1975–2012. Institute for Social Research. The University of Michigan, Ann Arbor, MI.

Korhonen, T., Latvala, A., Dick, D.M., Pulkkinen, L., Rose, R.J., Kaprio, J., Hulzink, A.C., 2012. Genetic and environmental influences underlying externalizing behaviors, cigarette smoking, and illicit drug use across adolescence. Behavior Genetics 42, 614–625.

O'Connor, R.C., O'Carroll, R.E., Ryan, C., Smyth, R., 2012. Self-regulation of unattainable goals in suicide attempters: a two-year prospective study. Journal of Affective Disorders 142, 248–255.

Salazar, L.F., Santelli, J.S., Crosby, R.A., 2009. Sexually transmitted disease transmission and pregnancy among adolescents. In: DiClemente, R.J., Santelli, J.S., Crosby, R.A. (Eds.), Adolescent Health: Understanding and Preventing Risk Behaviors. Jossey-Bass, San Francisco, CA, pp. 275–302.

Sheehan, C.M., Rogers, R.G., Williams, G.W., Boardman, J.D., 2013. Gender differences in the presence of drugs in violent deaths. Addiction 108, 547–555.

Sinha, R., 2008. Chronic stress, drug use, and vulnerability to addiction. Annals of the New York Academy of Sciences 1141, 105–130.

Smith, E.P., Dumas, J.E., Prinz, R., 2006. Prevention approaches to improve child and adolescent behavior and reduce deviant peer influence. In: Dodge, K.A., Dishion, T.J., Lansford, J.E. (Eds.), Deviant Peer Influences in Programs for Youth. Guilford Press, New York, pp. 296–311.

Spruijt-Metz, D., 2011. Etiology, treatment, and prevention of obesity in childhood and adolescence: a decade in review. Journal of Research on Adolescence 21, 129–152.

Sussman, S., Ames, S.L., 2008. Drug Abuse: Concepts, Prevention, and Cessation. Cambridge University Press, Cambridge, New York.

Taylor, B., Irving, H.M., Ranteres, F., Room, R., Borges, G., Cherpital, C., Rehm, J., 2010. A systematic review of how acute alcohol consumption and injury or collision risk increase together. Drug and Alcohol Dependence 110, 108B111.

Wagner, B.M., 2009. Suicidal Behavior in Children and Adolescents. Yale University Press, New Haven, CT.

Williams, D.R., Mohammed, S.A., 2009. Discrimination and racial disparities in health: evidence and needed research. Journal of Behavioral Medicine 32, 20–47.

Wills, T.A., Ainette, M.G., 2010. Temperament, self-control, and adolescent substance use: a two-factor model of etiological processes. In: Scheier, L.M. (Ed.), Handbook of Drug Use Etiology: Theory, Methods, and Empirical Findings. American Psychological Association, Washington, DC, pp. 127–146.

Wills, T.A., Bantum, E.O., Pokhrel, P., Maddock, J.E., Ainette, M.G., Morehouse, E., Fenster, B., 2013a. A dual-process model of early substance use: tests in two diverse populations of adolescents. Health Psychology 32, 533–542.

Wills, T.A., Forbes, M., Gibbons, F.X., 2013b. Parental and peer support: an analysis of their relations to adolescent substance use. In: Scheier, L.M., Hansen, W.A. (Eds.), Parenting and Teen Drug Use. Oxford University Press, New York, pp. 148–165.

Wills, T.A., Murry, V.M., Brody, G.H., Gibbons, F.X., Gerrard, M., Walker, C., Ainette, M.G., 2007. Ethnic pride and self-control related to protective and risk factors: the theoretical model for the Strong African-American Families Program. Health Psychology 26, 50–59.

Wills, T.A., Pokhrel, P., Morehouse, E., Fenster, B., 2011. Behavioral and emotional regulation and adolescent substance use problems: a test of moderation effects in a dual-process model. Psychology of Addictive Behaviors 25, 279–292.

Wills, T.A., Sargent, J.D., Stoolmiller, M., Gibbons, F.X., Gerrard, M., 2008. Movie smoking exposure and smoking onset: a longitudinal study of mediation processes in a representative sample of US adolescents. Psychology of Addictive Behaviors 22, 269–277.

Adolescent Sexual Risk

Christopher Browning and Jenny C Malave, The Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews recent trends and theoretical approaches to understanding variation in exposure to sexual risks during adolescence. Sexual risks can be understood as capturing sexual behaviors and exposures that increase the likelihood of teenage pregnancy, sexually transmitted disease, and sexual violence. We present estimates of the distribution of adolescent sexual risk over time and by key demographic correlates such as age, gender, and race/ethnicity. We then consider research on the etiology of sexual risk, including neuropsychological, social network, and neighborhood/ecological factors. Discussion of new directions for research on sexual risk highlights the considerable potential of multilevel, integrated approaches.

Introduction

Research on 'normative' sexual development during adolescence can be traced back to early-twentieth-century psychoanalytic approaches and the tradition of ontogenic theorizing that followed (Dannefer, 1984). The latter half of the century saw both the apogee of psychoanalytic and related approaches to sexual development and their subsequent decline in the face of fundamental challenges to the notion that 'normal' (and, by implication, deviant) sexual development could be an organizing category for research. Mounting evidence of considerable diversity in sexual behavior, attitudes, and orientation during adolescence led to interest in documenting the complex nature of emerging sexuality. Alongside these debates, the latter half of the twentieth century also saw the increasing prevalence of sexual activity during adolescence and rising sexually related risks such as teen pregnancy, sexual violence, and sexually transmitted infections (including HIV/AIDS). Consensus on the need to understand the etiology of exposure to sexual risks during adolescence has fueled a substantial research literature.

This article emphasizes work on the patterns and origins of adolescent sexual risk. Specifically, the focus of this article will be on examining the various approaches currently utilized within the field of research on adolescent sexual risk, highlighting the need for integrated orientations that incorporate not only individual-level propensities toward risk behavior, but the social network and broader ecological and neighborhood factors that influence risk as well.

We begin with a brief overview of trends in the distribution of sexual risks across time and populations. We then review approaches to the etiology of sexual risk, focusing on three general themes: (1) psychological, biological, and other individual-level predisposing factors; (2) social network contexts; and (3) ecological (neighborhood and institutional) factors that shape sexual risks. In the following sections particular attention will be given to examining the social context in which people are embedded. This article concludes with a brief overview of the strengths of an integrated approach in which individual, network, and ecological factors are incorporated.

Background

The concept of *sexual risk* includes both sexual risk-taking behavior as well as exposure to risks such as sexual violence. Sexual risk taking can include a myriad of behaviors, most commonly defined in terms of inconsistent use of birth control or contraceptives, multiple or concurrent partners, as well as early initiation into sex. These behaviors can lead to a number of negative health outcomes, especially among adolescents. They have been found to increase exposure to sexually transmitted diseases (STDs) (including AIDS/HIV), reduce mental health, and have implications for early childbearing, which can result in uncertain outcomes for both mother and child (Meier, 2003).

Adolescent Sexual Behavior and Risk: Trends and Key Demographic Correlates

We begin with an overview of trends and key demographic correlates of sexual behavior and risk taking. Although reliable data on adolescent sexual behavior have only been available in recent decades, evidence suggests significant changes over time in adolescent sexual risk. Data on the demographic correlates of sexual risk point to the fundamental organizing role of age and gender in the distribution of sexual behavior and sexual risk and demonstrate the unequal distribution of sexual risks across categories such as race and socioeconomic status.

Historical Trends

Self-reports of sexual behavior indicate that sexual activity and teenage pregnancy increased significantly during the 1960s and 1970s (Finer, 2007). These dramatic increases have ebbed in recent years however. The National Survey of Family Growth (NSFG) reports that the percent of unmarried teens between the ages of 15–19 years that have had sexual intercourse has decreased since 1988 for both males and females. Teen pregnancy reached a high point in 1990, but has decreased by approximately 44% since that time. The same is true for birth rates and abortion rates. More recently, between 2008 and 2012, trends have indicated a decrease in abortion and birth rates for teens ages between 15 and 19 years. The 2010 NSFG

finds that the use of contraceptives (including dual use methods, e.g., condoms and hormonal) has increased for both males and females and is likely contributing to the declining rates of teen pregnancy and abortions.

Demographic Correlates

Age is one of the strongest predictors of sexual behavior and risk taking. The probability of sexual onset rises dramatically from early to mid-adolescence, such that between the ages of 15 and 16 years, approximately 30% of youth have experienced sexual intercourse (Finer and Philbin, 2013). On average, males and females experience the first intercourse at the age of 17.1 years (Martinez et al., 2006). Adolescents who engage in sexual activity at younger ages are less likely to use contraceptives, more likely to experience diminished mental health, and more likely to acquire higher numbers of sexual partners throughout the life course (Martinez et al., 2011; Meier, 2003). Additionally, younger teens are more likely to have their first sexual experience with someone with whom they did not have a relationship (Martinez et al., 2011).

The NSFG also shows important differences and similarities in sexual behavior by gender. The percentage of males aged 15–17 years who are sexually active has decreased more rapidly in the past 20 years than for similarly aged females. In 1988, 60% of males in this age group had engaged in sex, as opposed to 42% in 2006–10. Comparatively, female rates have dropped from 51% to 43% over the same period. Extant evidence also points to marked differences across racial/ethnic groups. For example, 25.1% of Hispanic women aged 15–17 years have engaged in sex, as opposed to 30.4% of non-Hispanic white females and 40.8% of non-Hispanic black females (Chandra et al., 2005). Comparatively, 42.6% of Hispanic males, 25% of non-Hispanic white males, and 51.6% of non-Hispanic black males have had sexual intercourse by age 15–17 years (Martinez et al., 2006).

Teen pregnancies exhibit some of the largest discrepancies by race and ethnicity. Among teenagers aged 15–17 years, non-Hispanic whites experienced pregnancy at a rate of 21.6/1000 as compared to 72.8/1000 for non-Hispanic blacks and 69.7/1000 for Hispanics in 2008 (Chandra et al., 2012). African-American youth and young adults are at a higher risk of engaging in sexual activity that will increase the chance of receiving or transmitting an STD than any other racial group (Chandra et al., 2012). A substantially larger percentage of non-Hispanic black men and women aged 15–24 years report having five or more partners within a given year in comparison to non-Hispanic whites and Hispanics (Chandra et al., 2012); 9.5% of non-Hispanic black men and 2.5% of non-Hispanic black women report having had more than five sexual partners in the past year compared to 3.1% of non-Hispanic white males, 3.8% of Hispanic males, 1.6% of non-Hispanic white females, and 1.4% of Hispanic females. Non-Hispanic black women within this age bracket were more likely to seek treatment for an STD than other group: 11% sought treatment compared to 3.8% of Hispanic women and 4.6% of non-Hispanic white women (Chandra et al., 2012). Similarly, research has found that consistent use of condoms appears to be lower for minority adolescents than white adolescents (Brown et al., 1992). Explanations for the disparities in sexual risk by race/ethnicity are varied, but an emerging body of evidence points to the role of challenges in the social environments minority groups face. Below, we review a range of explanations for variation in sexual risk. Although we focus primarily on the social context of adolescent sexual risk, we begin with a brief review of emerging biological and psychological research.

Psychological and Biological Factors

Adolescent predispositions to engage in risky sexual behavior are shaped by a host of psychological and biological factors. The complexity of psychological and biological approaches precludes adequate review here; however, we briefly consider some of the promise and potential challenges associated with research on the neuropsychology of risk. A neuropsychological approach emphasizes the role of variability in the development of brain structure and function in shaping individual-level approaches to risk behavior. Neuropsychological research demonstrates the relevance of temperamental, behavioral, and cognitive processes in predicting participation in risky behavior, including risky sexual activity, during adolescence and young adulthood (Ramrakha et al., 2007). Continuity in the expression of neuropsychological deficits across phases of the life course suggests that a small, largely male subgroup may be at particularly high risk for participating in a range of risky sexual behaviors during adolescence and beyond (Moffitt, 1993).

Neuropsychological explanations for the association between age and participation in risky behavior within adolescence have also been offered with implications for understanding the age distribution of sexual risk. Social neuroscience approaches to adolescent risk, for instance, emphasize the role of dramatic changes taking place within the developing adolescent brain, beginning around the time of puberty. Processes occurring in the prefrontal cortex during adolescence influence reward-seeking particularly when in the presence of peers, resulting in diminished self-regulation. In turn, ongoing development of brain structure and function throughout adolescence eventually results in improved self-regulatory capacity and declining incidence of risky behavior (Robbins and Bryan, 2004). Early onset of puberty has also been linked with earlier occurrence of risky sexual activity (Belsky et al., 2010); changes in brain structure and function may partially account for this link to the extent that the association between pubertal onset and risky sexual activity is explained by biological development (Doremus-Fitzwater et al., 2010).

Neuropsychological predictors of adolescent sexual risk have been the subject of increasing scientific inquiry in recent years as technologies for the assessment of brain structure and functional processes (MRI, fMRI) become more widely available. Although a promising avenue of research, neuropsychological development must be understood as embedded in a larger social context that contributes to both brain structure and function while also constituting the environment in which resulting risk predispositions are manifest.

For instance, early-childhood adversity in the form of exposure to stressors, such as violence and child abuse, may have significant implications for the development of the

prefrontal cortex (Cicchetti, 2002). Traumatic experiences like abuse, in turn, have been linked with the occurrence of risky sexual behavior in adolescence. For example, Wyatt et al. (2002) found that abuse was correlated with more risky sexual behavior and increased risk of HIV across all races and ethnicities. Like low birth weight, however, exposure to early-childhood adversity is structured by a host of economic, demographic, and geographic factors. Thus, increasing evidence of the impact of early-childhood adversity on neurobiological and psychological processes highlights the importance of understanding the social contexts that influence both neurodevelopment as well as other mechanisms that independently contribute to adolescent outcomes. It is these latter processes to which we now turn.

Social Context: Network Factors

A long tradition of research in the social sciences highlights the role of relationships and larger patterns of social ties – captured here by the concept of a *social network* – in organizing the emergence and expression of adolescent sexuality. Despite a considerable literature examining social network influences on adolescent outcomes, the definition of what constitutes a network remains ambiguous. At the individual level, we employ the term 'social network' to encompass the set of relationships maintained by a given person, including intimate relationships as well as more distal social ties and acquaintances. In considering adolescent sexuality, most attention has been paid to the effect of friendship groups or peers, familial networks, and romantic relationships. These ties are characterized by the ability to exert significant influence on the behavior of adolescents.

A network approach to studying adolescent sexual risk can be conceptualized via multiple pathways of influence. First, behavior is shaped by rewards and sanctions conferred through network associates. Friends and family are conduits through which powerful social incentives and disincentives are transmitted to adolescents. Individuals learn which behaviors yield rewards and which result in sanctions – such as social exclusion from the network – through responses to their own behavior and observation of responses to the behavior of key network associates. Second, sexual behavior is shaped by social support – particularly emotional and informational support – that an individual receives from network associates. Lastly, sexual risk behavior is dependent on the availability of social opportunities that facilitate the occurrence of risky behaviors.

Rewards and Sanctions

Vicarious Rewards and Sanctions

Socialization plays a key role in the acquisition of expectations regarding appropriate sexual development and behavior. Behavioral models and sexual norms shape the likelihood of engaging in risk-taking behaviors. In instances where rewards and sanctions are not experienced directly, adolescents perceive what acceptable behavior entails through the behavior of network associates and the subsequent consequence of these behaviors. Reward and sanction processes are especially important to consider during adolescence, as individual participation in sexual behavior is typically only emerging. Simultaneously, the sexual behavior of others in the social environment is highly salient. Social incentives and disincentives to engage in particular types of behavior may vary dramatically by network context – friends and parents, for instance, may model behaviors quite differently. These relationships also vary in dyadic characteristics, such as frequency of time spent together and emotional closeness, as well as embeddedness in larger tie structures, with implications for patterns of influence.

Friendship holds particular importance during adolescence, as the need for social inclusion rises (Berndt, 1982). During this period, adolescents increase the amount of time they socialize with similarly aged peers and tend to distance themselves from parents (Larson and Richards, 1991). Adolescents who find themselves in high-risk networks are increasingly likely to engage in negative health behaviors in order to 'fit in.' For example, an individual is more likely to begin to have sex if there is a high rate of sexual initiation among his/her peer group (Teitler and Weiss, 2000). In general, friends' engagement in sexual activities shapes normative behavior in the peer group, a potentially powerful source of influence during adolescence (Fletcher, 2007).

Low-risk behavior may also be modeled by peers. Studies have shown that perceptions of peer attitudes are especially important to adolescents. Adolescents who perceive that friends hold less favorable views about teen sexual behavior are more likely to delay sexual initiation (Santelli et al., 2004). Perceiving that peers are using contraceptives may encourage individuals to do so as well and has been shown to decrease participation in other high-risk sexual behaviors (DiClemente, 1991; Fisher et al., 1992). Having friends who practice safe sex increases the likelihood that an individual will use similar methods (Berenson et al., 2006).

Similarly, adolescents are also affected by the norms, values, and behaviors of parents. Parents who model more risky behavior are more likely to have children who also engage in sexual risk. Teenagers are more likely to become pregnant if they were born to a teen mother and are more likely to have sex at earlier ages (East et al., 1993). Additionally, low-socioeconomic-status households may consider pregnancy to be acceptable at a younger age, as many markers of adulthood, such as education, are not accessible (Soller and Haynie, 2013; Bettie, 2002). However, parents and other network members also influence sexual risk through *direct* rewards and sanctions.

Direct Rewards and Sanctions

Adolescents experience social control via peers, parents, and significant others. Sexual expression may result in increased popularity, heightened sense of prestige, or alternatively, may be met by shaming, name-calling, or loss of status. Just as acceptable sexual behavior varies widely between school-aged boys and girls (Pascoe, 2007), rewards and sanctions are also organized by gender. Young men are encouraged to express a hypersexual self through bravado, bragging, early sexual experience, and multiple partnering, while young women are typically discouraged from similar behavior (although influences from romantic partners will often counter conventional

sexual expectations for girls) (Muehlenhard, 1988). Participation in sexual activity may increase the popularity or prestige of a given boy among his peers, while girls who act similarly are commonly labeled 'sluts' or 'whores' by boys and girls alike, creating a sexual double standard for adolescents (Aubrey, 2004).

Encouragement from parental networks may not reflect the same values as school-age peer networks. Typically, groups that encourage prosocial norms, like parents, encourage youth to make less risky sexual decisions, while groups like peers and friends encourage more risky behavior. Parents are unique with respect to their ability to enforce sanctions against a child. Parents enforce social control by expressing disapproval of their child's participation in risky behaviors (Dittus and Jaccard, 2000). Subsequent behavioral changes may be explained by a child's unwillingness to jeopardize close relational ties with his/her parents. Children may also perceive that certain behaviors will be rewarded by parents through mechanisms like respect, trust, or praise, which may heighten engagement in or desistence from a particular behavior. Children who feel that they matter to their parents are more likely to avoid engaging in risk-taking behavior than children who lack similar bonds (Elliott, 2010).

For many, adolescence represents the first step into adultlike romantic relationships. Romantic relationships likely represent the first sexual experience that young people will encounter and are therefore categorized by varying degrees of uncertainty. Rewards and sanctions are particularly salient within this setting, as wrong moves may endanger the relationship status. Young women who are dating older men are more likely to engage in high-risk sexual activity, as they lack relative power within the relationship (Males, 1998). As a result, young women with older partners are inhibited in their ability to successfully negotiate sexual interactions (Holland et al., 1992).

Condom negotiation is particularly difficult for young women. Girls commonly cite fear of being rejected by partners as a reason for not discussing birth control (Wingood and DiClemente, 1998). Moreover, many women report that having conversations with partners about safe sex implies distrust in the partner, which can have future consequences (Wingood and DiClemente, 1998). Low self-esteem has also been linked to lack of dialogue between sexual partners with respect to contraception (Sterk et al., 2005; Salazar et al., 2005). Consequently, gender remains a significant predictor of consistent use of contraceptives among adolescents (Ford et al., 2001).

Social Support

Adolescent sexual risk behavior can also be understood as a by-product of social support experienced within an individual's personal social network. A social support framework for the etiology of sexual risk focuses attention on behavioral responses to environmental stressors and the potential moderating influence of social resources. For instance, borrowing from Lazarus and Folkman (1984), one of the ways in which to understand sexual risk taking is as a coping mechanism for dealing with stress. Past studies have linked having multiple partners in a given month to self-reports of having sex to relieve tension (McKusick et al., 1985). Strain can be reduced through the introduction of positive relations that confer social support, specifically emotional and informational support, for turning negative strain into positive outcomes and reducing participation in high-risk sexual behaviors (Folkman et al., 1992).

Emotional Social Support

Emotional support can act as a protective factor that interacts with risk factors to reduce the likelihood of negative health outcomes among adolescents. Social support has been shown to help ameliorate stress caused by poverty (Taylor, 2010), alcohol consumption (Bacharach et al., 2010), and a range of other factors. Parents play a particularly important role in the emotional well-being of adolescents. One study links stronger parental support to lower sexual risk outcomes: Adolescents with stronger mother–child relationships were more likely to delay sexual initiation than those who had weaker bonds to parents (Dittus and Jaccard, 2000). Other studies offer evidence that this relationship is contingent on gender and race: Girls have been found to experience the protective effect of parental support more than adolescent boys (Rose et al., 2005), while white females have been found to experience protective effects of emotional relationships more so than minority females (Miller et al., 1999).

Friends can also play a strong role in providing emotional support during adolescence. Fallon (2010) found that teens were more likely to access emergency contraception if friends accompanied them to the clinic. Friends were found to mitigate much of the social stigma surrounding the use of the morning-after pill by offering emotional support through accompanying their friends to the location. Though research on emotional support within social relations is not conclusive, these studies offer important avenues for exploration. More research is needed to tease out the protective effects of emotional support with respect to sexual risk behavior.

Informational Social Support

Not surprisingly, adolescent sexual risk is also influenced by the informational support an individual receives from their social network. This is particularly true when thinking about how adolescents learn about contraceptives and where to access them. The effect of parent–adolescent communication has yielded mixed results in extant research. While studies show that open communication with parents leads to lower initiation into sex and increased use of contraceptives (Schalet, 2011), other studies have found no significant findings between parental communication and sexual initiation (Jaccard and Dittus, 1993). Studies have shown that many parents fail to have conversations about sex with their children and there is often discordance between adults and adolescence regarding what was discussed (Newcomer and Udry, 1985). In some cases, studies have revealed that having a conversation about birth control may actually increase the odds of an adolescent having sex, albeit protected sex, because they perceive there is parental approval (Jaccard et al., 2000). Others find that children who do have open conversations with

parents have been found to wait longer to initiate sex and to more consistently use contraceptives in a cross-national context (Schalet, 2011).

At the same time, however, social networks can maintain or propagate beliefs that are inaccurate or based on misinformation. Yee and Simon (2010) found that individuals were more likely to trust the anecdotal evidence provided by friends and family than the information provided by a physician. Women were found to reject an effective form of birth control because of the negative experience of a member of their social network, which may ultimately increase exposure to sexual risk (Yee and Simon, 2010). Additionally, parental communications regarding birth control are often found to be inaccurate. Parents who lack proper knowledge about birth control and then transmit this information to their children increase the chances of their adolescent engaging in risky sexual behavior (Jaccard et al., 1999).

Opportunities

Sexual risk behavior is reliant on opportunities that are either created or restricted within social networks. Typically, adolescents who experience early pubertal debut are more likely to engage in higher-risk behavior than later-developing individuals. Pubertal development is linked to social consequences, such as the composition of peer networks, with implications for sexual risk behaviors. Haynie (2003) found that girls who matured at younger ages were more likely to start spending time with older peer groups, especially men, which increased exposure to minor and serious delinquency, encouraged the use of controlled substances, and increased exposure to sexual risks.

Parents may limit exposure to opportunities by restricting access to certain networks like friends or older peers, limiting the amount of time children spend participating in unstructured activities, and supervising the actions of their children. Supervision has been linked to the reduction of delinquency and other problem behaviors (Loeber and Stouthamer-Loeber, 1986). Cohen and Felson (1979) classically posit that deviance occurs only when there is a motivated actor, a suitable object of action, and lack of capable guardianship. If any of the three is missing, action cannot be taken. In this view, the absence of available partners and the presence of guardianship constrain the occurrence of (dyadic) sexual risk behavior.

Many studies have looked at the effect of parental monitoring and have largely found that lack of supervision is linked to increased sexual risk taking. Children who have lower supervision and more unstructured free time are more likely to have earlier sexual initiation (Longmore et al., 2001), multiple sexual partners (Benda, 2002), higher occurrences of risky sexual activity (Huebner and Howell, 2003), and increased frequency of sexual activity (Benda, 2002). Studies have found that having lower levels of supervision around opposite-sex peers is significantly associated with earlier initiation into sex and having ever had sex for adolescents (DiIorio et al., 2004; Perkins et al., 1998).

Adolescents from single-family households are more likely to engage in sexual risk than those living in dual-parent households in part because of lack of supervision (Newcomer and Udry, 1987). Additionally, single-family households are more likely to be of lower socioeconomic status, which may mean the parental figure is working outside of the home. These families may also lack the resources to enroll their child in extracurricular activities or sports, where their child would be monitored by other adults.

Supervision may also be enacted within peer groups with particular norms and values. Specifically, religious groups are likely to hold norms regarding abstinence until marriage. Individuals who are active within church are increasingly constrained by the moral code of the church and are more likely to delay sexual initiation in order to avoid public shaming (Thornton and Camburn, 1989). Religious peers have been found to reduce opportunities in which a person may break their religious commitments by limiting alone time with peers of the opposite sex (Adamczyk and Felson, 2006).

Ecological Factors

Just as adolescents are embedded in networks of social ties, they also spend time in nonhome routine activity locations such as school, places of worship, and neighborhood public spaces. These extraresidential contexts shape adolescent day-to-day experiences and provide the settings in which network interactions are instantiated. This section reviews the expanding literature on the role of ecological contexts in shaping adolescent sexual risk taking.

Neighborhood

Classic perspectives on the role of neighborhoods in shaping adolescent behavior focus on the influence of structural factors such as poverty, residential instability, and race/ethnic heterogeneity in complicating reinforcement of shared community behavioral norms, including those surrounding sexuality (Browning et al., 2005). Indeed, recent literature has pointed to a number of structural features of urban communities in conceptualizing the impact of neighborhood context on adolescent sexual behavior. Growing evidence suggests that neighborhood socioeconomic disadvantage is associated with early sexual debut and related outcomes, such as teenage pregnancy, after individual and family background characteristics are accounted for (Baumer and South, 2001; Browning et al., 2004).

In addition to participation in risky sexual behavior, an expanding literature links communities that have higher levels of residential mobility and economic disadvantage with increased occurrence of sexual assault (Mustaine et al., 2014). For example, Miller (2008) found that within a disadvantaged urban community, sexual assault was common and interventions were unlikely even when the event happened in a public place. This was in part attributed to fear of the criminal justice system and lack of social power among the female residents of the community. Residents were additionally constrained by feelings of powerlessness to change the situation. Thus, sexual victimization often went unreported.

Structurally disadvantaged communities are characterized by compromised social climates along a number of dimensions. Among the most important social processes linked with

structural disadvantage is collective efficacy – or the level of mutual trust, solidarity, and shared values among community residents and associated expectations for prosocial action. While collective efficacy has traditionally been linked to reducing crime-related behavior through informal social control, it has increasingly been utilized as a concept through which to understand community-level variations in a range of risky behaviors, including sexual risk. Shared behavioral expectations at the neighborhood level may lead to collective socialization patterns that limit problem behaviors generally – even those that largely occur in private settings. Browning et al. (2005) found that collective efficacy was negatively associated with early onset of sexual intercourse among urban adolescents. However, this effect was only observed for youth who spent significant amounts of unsupervised time in the neighborhood, suggesting that levels of exposure to neighborhood environments vary across youth and may shape the extent of community influence on behavior. Similarly, evidence suggests that collective efficacy is protective against acquisition of multiple sexual partners during adolescence (Browning et al., 2008).

Other social characteristics more typically associated with disadvantaged neighborhoods have been linked to increases in sexual risk taking among adolescents. Neighborhoods characterized by an absence of recreational options for community members, limited safe environments for socializing with friends and romantic partners, and ineffective parental monitoring have been found to be correlated with increased sexual activity among younger community members (Akers et al., 2011).

Additional evidence suggests that structural disadvantage is associated with access to health-related services that could lower likelihoods of engaging in risky behavior. Neighborhoods with high levels of gang-related violence can negatively impact a resident's ability to access health services (Fleisher, 1998). Early pregnancy may also be considered more normative in a socially disadvantaged neighborhood where pregnancy signals a transition into adulthood for young women (Bettie, 2002).

Institutions

In addition to considering the neighborhood as a residential area, it is important to also consider the institutions embedded within a given space. Institutions, like schools and churches, provide important supervision and socialization contexts. According to De Lamater (1981), social institutions are able to control individual behavior in the following three ways: (1) they define a set of assumptions and norms that define acceptable behavior, (2) those occupying institutional roles will reinforce standards via informal control, and (3) institutions will employ sanctioning systems when norms are violated. Thus, individuals within the institution learn social norms, fear the possible sanctions, and are controlled by the policing of members. Additionally, institutions are able to constrain people in a systematic way so that the ability to engage in or desist from certain patterns is fundamentally altered.

Schools provide an institutional context that may independently influence adolescent outcomes. Some of the most consistent findings within the sexual health literature have been on the positive effect of sexual education. For instance, teens who have had comprehensive sexual education are more likely to effectively utilize birth control than those who were taught abstinence-only education (Kirby et al., 2007). Conversely, with few exceptions (Jemmott et al., 2010), research has suggested that abstinence-only education does not reduce the likelihood of an adolescent engaging in sex; rather, some evidence suggests that it reduces the likelihood of an adolescent knowing how to protect themselves, thus increasing the risk of unwanted pregnancy and STDs (Bearman and Bruckner, 2005).

Participation in religious institutions may also be an important source of influence on adolescent sexual behavior. Most active religions in the contemporary US provide clear dicta against engaging in sexual intercourse prior to marriage. Although the practical emphasis on discouraging sexual behavior among adolescents is more pronounced in certain religious contexts (e.g., Fundamentalist Christian), involvement in religious institutions would typically expose youth to values discouraging sexual activity. Moreover, religious institutions also incorporate youth into activities and social networks that provide other sources of informal social control such as supervised youth groups as well as friendships with youth whose parents are more likely to know one another and share information relevant for controlling behaviors that place youth at risk.

Indeed, some evidence suggests that participation in religious institutions reduces the likelihood of exposure to sexual risk. McCree et al. (2003) found that religion had a negative relationship with risky sexual behavior for African-American females specifically, whereas Bearman and Bruckner (2001) found that religiosity had a negative relationship with the onset of first intercourse among all races other than African-Americans. The effect of religion is stronger for those who actively participate in church or weekly services (Meier, 2003). On the other hand, religious organizations historically focus on abstinence-only teachings. As a result, adolescents who are lacking sexual education at church, home, and school may be unaware of safe sex practices. Recent research looking at the relationship between religion and risky sexual activity has found that religious adolescents who are currently sexually active are less likely to consistently use condoms (Zaleski and Schiaffino, 2000; Bearman and Bruckner, 2005).

Conclusion

The purpose of this article is to synthesize the current literature surrounding adolescent sexual risk-taking behavior. We provide an overview of important etiological factors at the psychological and biological level while paying particular attention to the specific influences of social networks and ecological factors. Past research has largely focused on these explanatory measures as independent of one another, but in fact, these factors may be usefully seen as interacting. Individuals with predispositions to engage in sexual risks are simultaneously embedded in variably influential social networks and ecological contexts. We suggest that future research would benefit from the integration of individual, network, and ecological factors that shape adolescent sexual trajectories.

In moving forward with interdisciplinary and multilevel investigation of sexual risk-taking behavior, researchers should be aware of some of the challenges facing these approaches. First, boundary specification issues plague social network and neighborhood studies alike (Laumann et al., 1989). Determining who is included as part of the network and to what extent certain spaces comprise a given neighborhood are theoretically and empirically thorny problems. Second, future research should consider how to improve data collection efforts. Social network analysis requires relational data, or data containing reference to the relationship among units. Neighborhood analysis requires, at a minimum, information on residential location. Ideally, information on the locations of daily activities should be included as well. Such data is difficult to collect, but will likely yield important findings on the nature of contextual influences on adolescent sexual risk. More generally, new insight into adolescent sexual risk behavior is most likely to come from the intersection of disciplines and substantive domains; thus, future research on this topic should undertake more comprehensive efforts to capture multiple, etiologically relevant domains of influence on human development.

See also: Adolescence, Sociology of; Adolescent Health and Health Behaviors; Defiant Behavior During Adolescence across Cultures; Friendship During Adolescence and Cultural Variations; Friendship Experiences of Lesbian, Gay, Bisexual, and Transgender Individuals; Gay, Lesbian, and Bisexual Youth; Hooking up and Casual Sex; Intimate Partner Violence; Leisure Activities Choices among Adolescents; Self-Regulation During Adolescence: Variations Associated with Individual–Context Relations; Sexual Behavior and Social Networks; Sexual Debut; Sexual Risk Behaviors; Sexuality Over the Life Course; Sexuality and the Internet; Sexuality: Cultural Aspects.

Bibliography

Adamczyk, Amy, Felson, Jacob, 2006. Friends' religiosity and first sex. Social Science Research 35 (4), 924–947.

Akers, Aletha Y., Muhammad, Melvin R., Corbie-Smith, Giselle, 2011. "When you got nothing to do, you do somebody": a community's perceptions of neighborhood effects on adolescent sexual behaviors. Social Science & Medicine 72 (1), 91–99.

Aubrey, Jennifer Stevens, 2004. Sex and punishment: an examination of sexual consequences and the sexual double standard in teen programming. Sex Roles 50 (7–8), 505–514.

Bacharach, Samuel B., Bamberger, Peter, Biron, Michal, 2010. Alcohol consumption and workplace absenteeism: the moderating effect of social support. Journal of Applied Psychology 95 (2), 334–348.

Baumer, Eric P., South, Scott J., 2001. Community effects on youth sexual activity. Journal of Marriage and Family 63 (2), 540–554.

Bearman, Paul, Bruckner, Hannah, 2001. Promising the future: virginity pledges and first intercourse. American Journal of Sociology 106 (4), 859–912.

Bearman, Paul, Bruckner, Hannah, 2005. After the promise: the STD consequences of adolescent virginity pledges. Journal of Adolescent Health 36 (4), 271–278.

Belsky, Jay, Steinberg, Laurence, Houts, Renate M., Halpern-Felsher, Bonnie L., 2010. The development of reproductive strategy in females: early maternal harshness → earlier menarche → increased sexual risk taking. Developmental Psychology 46 (1), 120–128.

Benda, Brent B., 2002. The effect of various aspects of religion and the family on adolescent sexual behavior. Marriage and Family Christian Journal 5 (3), 373–390.

Berenson, Abbey B., Wu, Z., Helen, Breitkopf, Carmen Radecki, Newman, Jennifer, 2006. The relationship between source of sexual information and sexual behavior among female adolescents. Contraception 73 (3), 247–278.

Berndt, Thomas J., 1982. The features and effects of Friendship in early adolescence. Child Development 53 (6), 1447–1460.

Bettie, Julie, 2002. Women Without Class: Girls, Race, and Identity. University of California Press, Berkley, CA.

Browning, Christopher R., Burrington, Lori A., Leventhal, Tama, Brooks-Gunn, Jeanne, 2008. Neighborhood structural inequality, collective efficacy, and sexual risk behavior among urban youth. Journal of Health and Social Behavior 49 (3), 269–285.

Browning, Christopher R., Leventhal, Tama, Brooks-Gunn, Jeanne, 2005. Sexual initiation in early adolescence: the Nexus of parental and community control. American Sociological Review 70 (5), 758–778.

Browning, Christopher R., Leventhal, Tama, Brooks-Gunn, Jeanne, 2004. Neighborhood context and racial differences in early adolescent sexual activity. Demography 41 (4), 697–720.

Brown, Larry K., DiClemente, Ralph J., Park, Teron, 1992. Predictors of condom use in sexually active adolescents. Journal of Adolescent Health 13 (8), 651–657.

Center for Disease Control, 2012. Sexual experience and contraceptive use among female teens — United States, 1995, 2002, and 2006–2010. In: Morbidity and Mortality Weekly Report (MMWR). http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6117a1.htm?s_cid=mm6117a1_e (accessed 09.09.13).

Chandra, Anjani, Martinez, Gladys M., Mosher, William D., Abma, Joyce C., Jones, Jennifer, 2005. Fertility, family planning, and reproductive health of US women: data from the 2002 national survey of family growth. Vital and Health Statistics 25, 1. Series 23, Data from the National Survey of Family Growth.

Chandra, Anjani, Billioux, Veena G., Copen, Casey E., Sionean, Catlainn, 2012. HIV Risk-related Behaviors in the United States Household Population Aged 15–44 Years: Data from the National Survey of Family Growth, 2002 and 2006–2010. US Department of Health and Human Services. Centers for Disease Control and Prevention, National Center for Health Statistics.

Cicchetti, Dante, 2002. The impact of social experience on neurobiological systems: illustration from a constructivist view of child maltreatment. Cognitive Development 17 (3), 1407–1428.

Cohen, Lawrence E., Felson, Marcus, 1979. Social change and crime rate trends: a routine activity approach. American Sociological Review 44 (4), 588–608.

Committee on Appropriations United States Senate, 2004. The Content of Federally Funded Abstinence Only Education. http://www.apha.org/apha/PDFs/HIV/The_Waxman_Report.pdf (Retrieved 29.01.14).

Dannefer, Dale, 1984. Adult development and social theory: a paradigmatic reappraisal. American Sociological Review 49 (1), 100–116.

DeLamater, John, 1981. The social control of sexuality. Annual Review of Sociology 7, 263–290.

DiClemente, Ralph J., 1991. Predictors of HIV-preventive sexual behavior in a high-risk adolescent population: the influence of perceived peer norms and sexual communication on incarcerated adolescents' consistent use of condoms. Journal of Adolescent Health 12 (5), 389–390.

Dilorio, Colleen, Dudley, William N., Soet, Johanna E., McCarty, Frances, 2004. Sexual possibility situations and sexual behaviors among young adolescents: the moderating role of protective factors. Journal of Adolescent Health 35 (6), 528–e11.

Dilorio, Colleen, Dudley, William N., Kelly, Maureen, Soet, Johanna E., Mbwara, Joyce, Potter, Jennifer Sharpe, 2001. Social cognitive correlates of sexual experience and condom use among 13-through 15-year-old adolescents. Journal of Adolescent Health 29 (3), 208–216.

Dittus, Patricia J., Jaccard, James, 2000. Adolescents' perceptions of maternal disapproval of sex: relationship to sexual outcomes. Journal of Adolescent Health 26 (4), 268–278.

Doremus-Fitzwater, Tamara L., Varlinskaya, Elena I., Spear, Linda P., 2010. Motivational systems in adolescence: possible implications for age differences in substance abuse and other risk-taking behaviors. Brain and cognition 72 (1), 114–123.

East, Patricia L., Felice, Marianne E., Morgan, Maria C., 1993. Sisters' and girlfriends' sexual and childbearing behavior: effects on early adolescent girls' sexual outcomes. Journal of Marriage and the Family 55 (4), 953–963.

Elliott, Sinikka, 2010. Parents' constructions of teen sexuality: sex panics, contradictory discourses, and social inequality. Symbolic Interaction 33 (2), 191–212.

Fallon, Deborah, 2010. Accessing emergency contraception: the role of friends in the adolescent experience. Sociology of Health and Illness 32 (5), 677–694.

Finer, Lawrence B., 2007. Trends in premarital sex in the United States, 1954–2003. Public Health Reports 122 (1), 73–78.

Finer, Lawrence B., Philbin, Jesse M., 2013. Sexual initiation, contraceptive use, and pregnancy among young adolescents. Pediatrics 131 (5), 886–891.

Fisher, Jeffrey D., Misovich, Stephen J., Fisher, William A., 1992. Impact of perceived social norms on adolescents' AIDS-risk behavior and prevention. In: DiClemente, R.J. (Ed.), Adolescents and AIDS: A Generation in Jeopardy. Sage, Newbury Park, CA, pp. 117–136.

Fleisher, Mark S., 1998. Dead End Kids: Gang Girls and the Boys They Know. University of Wisconsin Press, Madison, WI.

Fletcher, Jason M., 2007. Social multipliers in sexual initiation decisions among US high school students. Demography 44 (2), 373–388.

Folkman, Susan, Chesney, Margaret A., Pollack, Lance, Phillips, Connie, 1992. Stress, coping, and high-risk sexual behavior. Health Psychology 11 (4), 218.

Ford, Kathleen, Sohn, Woosung, Lepkowski, James, 2001. Characteristics of adolescents' sexual partners and their association with use of condoms and other contraceptive methods. Family Planning Perspectives 33 (3), 100–132.

Haynie, Dana L., 2003. Contexts of risk? Explaining the link between girls' pubertal development and their delinquency involvement. Social Forces 82 (1), 355–397.

Holland, Janet, Ramazonoglu, Caroline, Sharpe, Sue, Thomson, Rachel, 1992. Pleasure, pressure and power: some contradictions of gendered sexuality. The Sociological Review 40 (4), 645–674.

Huebner, Angela J., Howell, Laurie W., 2003. Examining the relationship between adolescent sexual risk-taking and perceptions of monitoring, communication, and parenting styles. Journal of Adolescent Health 33 (2), 71–78.

Jaccard, James, Dittus, Patricia, 1993. Parent-adolescent communication about premarital pregnancy. Families in Society 69 (1), 247–261.

Jaccard, James, Dittus, Patricia, Harold, Litardo, 1999. Parent-adolescent communication about sex and birth control: implications for parent based interventions to reduce unintended adolescent pregnancy. In: Miller, W., Severy, L. (Eds.), Advances in Population Research: Psychosocial Perspectives. Kingsley, London.

Jaccard, James, Dittus, Patricia J., Gordon, Vivian V., 2000. Parent-teen communication about premarital sex: factors associated with the extent of communication. Journal of Adolescent Health 15 (2), 187–208.

Jemmott III, John B., Jemmott, Loretta S., Fong, Geoffrey T., 2010. Efficacy of a theory-based abstinence-only intervention over 24 months: a randomized controlled trial with young adolescents. Archives of Pediatrics & Adolescent Medicine 164 (2), 152–159.

Kirby, Douglas B., Laris, B.A., Rolleri, Lori A., 2007. Sex and HIV education programs: their impact on sexual behaviors of young people throughout the world. Journal of Adolescent Health 40 (3), 206–217.

Larson, Reed, Richards, Maryse H., 1991. Daily companionship in late childhood and early adolescence: changing developmental contexts. Child Development 62 (2), 284–300.

Laumann, Edward O., Marsden, Peter V., Prensky, David, 1989. The boundary specification problem in network analysis. In: Freeman, L.C., White, D.R., Romney, A.K. (Eds.), Research Methods in Social Network Analysis. George Mason University Press, New Jersey, pp. 61–87.

Lazarus, Richard Stanley, Folkman, Susan, 1984. Stress, Appraisal, and Coping. Springer Publishing Company, New York, NY.

Loeber, Rolf, Stouthamer-Loeber, Magda, 1986. Family factors as correlates and predictors of Juvenile conduct problems and delinquency. Crime and Justice. 7, 29–149.

Longmore, Monica A., Manning, Wendy D., Giordano, Peggy C., 2001. Preadolescent parenting strategies and teens' dating and sexual initiation: a longitudinal analysis. Journal of Marriage and Family 63 (2), 322–335.

Males, Mike, 1998. Adult partners and adult contexts of teenage sex. Education and Urban Society 30 (2), 189–206.

Martinez, Gladys M., Chandra, Anjani, Abma, Joyce C., Jones, Jo, Mosher, William D., 2006. Fertility, contraception, and fatherhood: data on men and women from cycle 6 (2002) of the National Survey of Family Growth. Vital and Health Statistics 23 (26).

Martinez, Gladys, Casey, E. Copen, Joyce, C., Abma, 2011. Teenagers in the United States: sexual activity, contraceptive use, and childbearing, 2006–2010 national survey of family growth. National Center for Health Statistics. Vital Health Statistics 23 (31).

McCree, Donna Hubbard, Wingood, Gina M., DiClemente, Ralph, Davies, Susan, Harrington, Katherine F., 2003. Religiosity and risky sexual behavior in African-American adolescent females. Journal of Adolescent Health 33 (1), 2–8.

McKusick, Leon, Horstman, William, Coates, Thomas J., 1985. AIDS and sexual behavior reported by gay men in San Francisco. American Journal of Public Health 75 (5), 493–496.

Meier, Ann M., 2003. Adolescents' transition to first Intercourse, religiosity, and attitudes about sex. Social Forces 81 (3), 1031–1052.

Miller, Jody, 2008. Getting Played: African American Girls, Urban Inequality, and Gendered Violence. NYU Press, New York, NY.

Miller, Kim S., Forehand, Rex, Kotchick, Beth A., 1999. Adolescent sexual behavior in two ethnic minority groups: a multisystem perspective. Adolescence 35 (138), 313–333.

Moffitt, Terrie E., 1993. Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy. Psychological Review 100 (4), 674–701.

Muehlenhard, Charlene L., 1988. "Nice women" don't say yes and "real men" don't say no: how miscommunication and the double standard can cause sexual problems. Women & Therapy 7 (2–3), 95–108.

Mustaine, Elizabeth Ehrhardt, Tewksbury, Richard, Huff-Corzine, Lin, Corzine, Jay, Marshall, Hollianne, 2014. Community characteristics and child sexual assault: social disorganization and age. Journal of Criminal Justice 42 (2), 173–183.

Newcomer, Susan, Udry, J. Richard, 1987. Parental marital status effects on adolescent sexual behavior. Journal of Marriage and the Family 42 (2), 235–240.

Newcomer, Susan, Udry, J. Richard, 1985. Parent-child communication and adolescent sexual behavior. Family Planning Perspectives 17 (4), 169–174.

Pascoe, C.J., 2007. Dude, You're a Fag: Masculinity and Sexuality in High School. University of CA Press, Berkley, CA.

Perkins, Daniel F., Luster, Tom, Villarruel, Francisco A., Small, Stephen, 1998. An ecological, risk-factor examination of adolescents' sexual activity in three ethnic groups. Journal of Marriage and the Family 60 (3), 660–673.

Ramrakha, Sandhya, Bell, Melanie L., Paul, Charlotte, Dickson, Nigel, Moffitt, Terrie E., Caspi, Avshalom, 2007. Childhood behavior problems linked to sexual risk taking in young adulthood: a birth cohort study. Journal of the American Academy of Child & Adolescent Psychiatry 46 (10), 1272–1279.

Robbins, Reuben N., Bryan, Angela, 2004. Relationships between future orientation, impulsive sensation seeking, and risk behavior among adjudicated adolescents. Journal Of Adolescent Research 19 (4), 428–445.

Rose, Allison, Koo, Helen P., Bhaskar, Brinda, Anderson, Karen, White, Gregory, Jenkins, Renee R., 2005. The influence of primary caregivers on the sexual behavior of early adolescents. Journal of Adolescent Health 37 (2), 135–144.

Salazar, Laura F., Crosby, Richard A., DiClemente, Ralph J., Wingood, Gina M., Lescano, Celia M., Brown, Larry K., Harrington, Kathy, Davies, Susan, 2005. Self-esteem and theoretical mediators of safer sex among African American female adolescents: implications for sexual risk reduction interventions. Health Education and Behavior. 32 (3), 413–427.

Santelli, John S., Kaiser, Javaid, Hirsch, Lesley, Radosh, Alice, Simkin, Linda, Middlestadt, Susan, 2004. Initiation of sexual Intercourse among middle school adolescents: the influence of psychosocial factors. Journal of Adolescent Health 34 (3), 200–208.

Schalet, Amy T., 2011. Not under My Roof: Parents, Teens, and the Culture of Sex. University of Chicago Press, Berkeley, CA.

Soller, Brian, Haynie, Dana L., 2013. Structuring the future: anticipated life events, peer networks, and adolescent sexual behavior. Sociological Inquiry 83 (4), 537–569.

Sterk, Claire E., Klein, Hugh, Elifson, Kirk W., 2005. Self-esteem and "at risk" women: determinants and relevance to sexual and HIV-related risk behaviors. Women & Health 40 (4), 75–92.

Taylor, Ronald D., 2010. Risk and resilience in low-income African American families: moderating effects of kinship social support. Cultural Diversity and Ethnic Minority Psychology 16 (3), 344–351.

Teitler, Julien O., Weiss, Christopher C., 2000. Effects of neighborhood and school environments on transitions to first sexual Intercourse. Sociology of Education 73 (2), 112–132.

Thornton, Arland, Camburn, Donald, 1989. Religious participation and adolescent sexual behavior and attitudes. Journal of Marriage and the Family 51 (3), 641–653.

Wingood, Gina M., DiClemente, Ralph J., 1998. Partner influence and gender-related factors association with noncondom use among youth adult African American women. American Journal of Community Psychology 26 (1), 29–52.

Watts, Graham F., Nagy, Stephen, 2002. Sociodemographic factors, attitudes, and expectations toward adolescent coitus. American Journal of Health Behavior 24 (4), 309–317.

Wyatt, Gail E., Meyers, Hector F., Williams, John K., Kitchen, Christina Ramirez, Loeb Tamra, Carmona, Jennifer Vargas, Wyatt, Lacey E., Chin, Dorothy, Presley, Nicole, 2002. Does a history of trauma contribute to HIV risk for women of color? Implications for prevention and policy. American Journal of Public Health 92 (4), 660–665.

Yee, L., Simon, M., 2010. The role of social network in contraceptive decision-making among young, African American and Latina women. Journal of Adolescent Health. 47 (4), 374–380.

Zaleski, Ellen H., Schiaffino, Kathleen M., 2000. Religiosity and sexual risk-taking behavior during the transition to college. Journal of Adolescence 23 (2), 223–227.

Adoption, Demography of

Jessaca B Leinaweaver, Brown University, Providence, RI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Adoption refers to the reassignment or redistribution of parental rights, responsibilities, and practices with respect to a particular child. Adoptions may be informal or formal, private or public, domestic or international, between strangers or kin, transracial or racially 'matching,' and adhering to local norms or 'irregular.' Data on adoptions is difficult to obtain, because these different kinds of adoptions may be recorded differently or not at all. After reviewing this state of affairs, this article challenges the widely circulated narrative that adoption happens because of infertility and changes in supply of children, and explores the relationship between adoption and culture.

Definitions of Adoption

Adoption can be broadly defined as the reassignment or redistribution of parental rights, responsibilities, and practices with respect to a particular child. Scholars and policymakers distinguish between several different kinds of adoption: informal and formal, private and public, and domestic and international. Cross-cutting these categorical distinctions are three other characteristics that are considered salient: whether the adoption is between strangers or kin, whether it is perceived as transracial or racially 'matching,' and whether it adheres to local norms or is 'irregular.'

Formal/Informal

The distinction between informal and formal adoption refers to whether a government is involved in overseeing the transfer of a child. Informal adoption is practiced worldwide and is extremely common. In many parts of the world it is far from unusual for children to spend part or all of their childhoods living with people who are not their biological parents. Anthropological research has found that in places ranging from West Africa to Oceania to Latin America, children may be transferred from one adult to another without the stamp of governmental authority. By contrast, formal adoption is accomplished through courts or government offices. It may also be referred to as legal adoption, although this is not to imply that informal adoption is necessarily illegal or harmful.

Plenary/Simple

Formal adoptions may be plenary, a legal term meaning that they permanently sever a child from his or her family of origin and make him or her officially a child of the new parent or parents in all possible ways (in particular referring to inheritance potential, last name(s), and legal rights). The Hague Convention, which governs international adoptions for all countries that have signed on to it, requires that adoptions conducted under its auspices be plenary (article 2.2). The alternative is that adoptions can be 'simple,' a legal designation meaning that only some parental rights are transferred and a child may still maintain some legal ties to the natal family.

Public/Private

Formal adoptions may be either public or private. Public adoptions refer to adoptions of children who have been declared to be the responsibility of the state. As such, they transfer the responsibility of care from the state to a private family, benefiting the state by reducing its charge. In the United States, children adopted via the public system are already being cared for through foster care, and public agency adoptions are a key way that those children may exit the foster care system. In the U.S. ethnic minority children are overrepresented in the foster care system and the majority of those adopted are more than 5 years old at the time of their adoption (U.S. Department of Health and Human Services, 2009: 7). Overwhelmingly, children adopted from the foster care system are adopted by someone they know: 54% of those adopted in 2007 were adopted by a foster parent and 30% by a relative. By contrast, private adoptions in the United States are accomplished legally but without the same oversight established for public adoptions. They may occur either through the mediation of a private agency, or through a direct arrangement between birthmother and adoptive parent. International adoptions are also private from the perspective of the country the child is being moved to (the 'receiving country,' discussed below), since the children adopted do not move through the domestic foster care system prior to being placed.

Domestic/International

Thirdly, one can also distinguish between domestic and international (also called transnational or intercountry) adoptions. Domestic adoptions occur within the borders of a nation, while international adoptions involve a parent or parents from one country adopting a child from another country. As in migration scholarship, the country from which the underaged migrant comes is referred to as the 'sending country,' and the country to which he or she migrates is called the 'receiving country.' International adoptees are literally child migrants, and much of the paperwork required to formalize their presence in their new home is identical to that completed by migrants. Presently, international adoptions are expected to occur under the auspices of an international Hague Convention, formally entitled the "Convention of 29 May 1993 on Protection of

Children and Co-operation in Respect of Intercountry Adoption" (The Hague, 1993).

Domestic adoptions in the U.S. peaked in 1970, and more recently demand has moved overseas. International adoption formally began as a humanitarian project during midcentury in the context of wars in Europe and Asia (the Spanish Civil War, World Wars I and II, and the Korean War). But there was a veritable international adoption boom at the turn of the twenty-first century. Worldwide, the number of international adoptions increased by 42% between 1998 and 2004; some countries saw much sharper increases, and as one example, during that same period the number of international adoptions to Spain increased by 273% (Selman, 2009: 578).

Kin/Strangers

Three other important characteristics cross-cut these categorical distinctions. First, any of these adoptions may occur between relatives or between strangers. Stepparent or kin adoptions are included in national statistics as well and are difficult to separate out from 'stranger adoptions.' Kin adoptions, also called relative adoptions, may occur in any one of these adoption 'types' whether informal, formal, domestic, international, public, or private.

Race and Adoption

Second, any of them can be defined as transracial or racially 'matching.' Transracial adoptions are those in which a parent and a child physically differ in skin color, hair texture, and other features that are interpreted as racial identity. Most transracial adoptions involve white adoptive parents and minority adopted children, due both to racial hierarchies and ideologies and to the makeup of the foster care system. Since at least the 1950s, white parents have adopted children of color often from a stated position that race should not matter and the children need homes (Dubinsky, 2010).

Some minority communities have mobilized against those assumptions, including two strong statements in the U.S. in the 1970s. In 1972, the U.S. National Association of Black Social Workers issued a 'Position Statement on Trans-Racial Adoption' arguing that white parents are unable to teach African-American children the strategies that they would need to know in order to 'survive in a racist society.' In 1978, the Indian Child Welfare Act established that Native American tribes had specific collective rights and that Indian children's needs should ideally be understood as they relate to the community. (A 2013 Supreme Court ruling in the 'Baby Veronica' case appears to have weakened the ICWA.) Both communities were acting from a long history of having their children removed – African-American children into the foster care system and Native American children into boarding or residential schools.

International adoptions are frequently transracial, and this observation is reflected in current policy in many sending countries, in some cases directly derived from language used in Article 20 of the United Nations Convention on the Rights of the Child, which prioritizes "the desirability of continuity in a child's upbringing and … ethnic, religious, cultural and linguistic background."

Irregular Adoptions

Third, each kind of adoption may be evaluated as to whether it adheres to local norms or is 'irregular.' For example, some informal adoptions are traditional fostering arrangements, and others exploit children for labor. Perhaps more controversially, some international adoptions, while accomplished legally, may also be irregular and some legal scholars have criticized them as akin to traffic (Smolin, 2005) because of the lack of transparency with which poor people's children are deemed to be adoptable.

As is clear from this rapidly complicating picture, adoption is not simple, nor can it be well understood without considering several other related phenomena: informal fostering, public agency foster care, the orphaning or abandonment of children, the institutionalization of children, racial discrimination, the availability of reproductive technologies and their implications for adoption (such as embryo adoptions – when extra embryos, left over from the IVF cycle, are implanted into unrelated women's bodies – or the adoptions that may occur if a surrogate mother gives birth to the child of an intended mother, by contract), and other familiar demographic concerns such as ideal number of children, gender preference of children, and the like. The following section examines the data available for understanding adoption.

Demography of Adoption

Understanding adoption involves several related demographic analyses. Firstly, simply counting adoptions and determining how many occur can be challenging, given the regulatory patchwork and the differences between public and private adoption outlined above. We will also obtain a better understanding of adoption through considering related issues such as fertility rates and changes in rates of unmarried mothers placing their children for adoption.

Data and Collection Issues

Since the 1990s, there have been approximately 118 000 adoptions per year in the U.S., down from an all-time high in 1970 (Davis, 2011: 4). Data from 2000, 2007, and 2008 for the United States indicate that for each of those years, approximately 40% of the year's total adoptions (about 50 000) occur through public child welfare agencies, approximately 13% (about 18 000) were international, and almost half (47%, or just over 60 000) were from other sources: private agencies, stepparents, or tribes, for example (Child Welfare Information Gateway, 2011: 4). Contrast these figures with those from the 1980s and early 1990s when 16% of adoptions originated from public agencies (Flango and Flango, 1995, cited in Child Welfare Information Gateway, 2011: 14). In 2007, of the more than 50 000 children adopted from foster care, 23% were adopted by relatives and 69% were adopted by a foster parent (possibly including relatives) (Vandivere et al., 2009).)

But where does this data come from? The answer is surprisingly complicated. In the U.S., government data is available for public agency adoption. However, much of

private domestic and international adoption cannot be statistically tracked:

> There is no single source for the total number of children adopted in the United States, and there is currently no straightforward way of determining the total number of adoptions, even when multiple data sources are used. No single agency is charged with compiling this information, and agencies that do collect adoption-related data do so for their own purposes and therefore count adoptions differently (e.g., by court cases filed, birth certificates modified, adoptions completed by public agencies), which makes aggregation difficult.
> Child Welfare Information Gateway, 2011: 5

Similarly, several scholars have remarked upon the significant challenges in collecting accurate data on international adoptions (Kane, 1993; Selman, 2002; Weil, 1984). Records, particularly in sending countries, can be incomplete and obtaining data from receiving countries – such as immigration data for the 'orphan visa' in the U.S. – has proven more reliable.

Standardizing Measures

Whether one counts adoption applications, actual adoptions completed, or the population of children in foster care, or whether one measures changes in attitudes toward adoption, all of these methods could yield different figures and might not represent the current state of affairs. Thus, once adequate data has been collected, the problem of comparison also arises.

Peter Selman has proposed some different possibilities for standardizing measures of international adoption in particular. He suggests that international adoptions could be standardized either against the annual number of live births or per 1000 population (Selman, 2002: 207). Such a standardization would allow us to determine whether the countries with the largest overall numbers of international adoptions (the United States holds the top spot) also have the highest *ratios* of international adoptions to live births. In fact, Selman's results – using an 'adoption ratio' of the number of adoptions per 1000 live births – show that European countries had much higher ratios than the United States. In 1998, Norway's adoption ratio was 11.2 (p. 212) and the U.S.'s was 4.2, although the U.S. had 15 774 international adoptions and Norway had 643 that year (p. 213).

A similar measure can be developed to compare numbers of adoptions from specific sending countries. Again, Peter Selman offers two options: considering adoptions in relation to 100 000 population aged 0–4, or standardizing against births. His calculations show, for example, that although the large numbers of children from China are notorious, compared to the total number of under-5 children in China they are less significant than the figures from less populous countries like, in its day, Romania. He also found that standardizing against births allowed him to see that high birth rates do not explain adoption (the 'surplus story'). South Korea, China, and Eastern European nations, which have below-replacement birth rates, nonetheless continue to be significant sending countries (Selman, 2002: 216). Countries like Russia (peak year 2004), China (peak year 2005), Guatemala (peak year 2007), and Ethiopia (peak year 2010) have more recently been the most numerically significant sending countries (Selman, 2010: 7, 8; 2009: 578).

Causes of/Justifications for Adoption

The figures and calculations presented above complicate some of the commonly held assumptions about why children are adopted and lead to new and useful questions. The story about adoption today holds that in Western countries, fertility rates have declined and the population of adoptable infants domestically has decreased, leading to a rise in international adoption (Lovelock, 2000: 908; Selman, 2002: 206). Let us examine these statements more closely.

Infertility

Infertility is said to drive adults' desire to adopt. Certainly postponement of child bearing leads to an increased risk of infertility (*see* Infertility: Demographic Aspects). This story helps to explain the high incidence of international adoption; Selman and his colleagues report that "The proportion of international adopters who are primarily motivated by infertility is usually put at about 75 per cent" (Haworth et al., 2010: 136).

However, low fertility is not the only proximate determinant of a decision to adopt internationally. For example, single people or same-sex couples may not be biologically infertile but may choose to adopt rather than pursue alternative strategies. Evidence from U.S. adopters of Chinese children (Rojewski and Rojewski, 2001) and Catalonian adoption applicants (Font Lletjòs, 2008, cited in Marre, 2009: 231) suggest that infertility may motivate fewer than half of certain subsets of international adoptions. Instead, complex desires for a multicultural family or to help a needy child may be part of the picture (*see* Family Size Preferences).

This is also the case for adoptions that are other than international. The National Survey of Adoptive Parents (2007–08, United States) found that 52% of surveyed parents "noted they were unable to have a biological child, an issue for 18 percent of the intercountry adoptions, 14 percent of the foster adoptions and 20 percent of the private adoptions. Eighty-one percent wanted to give a child a permanent home" (Davis, 2011: 16).

Adoptions from foster care, stepparent, or kin adoptions may be motivated rather by an existing connection with a child. Informal adoptions, too, are not necessarily motivated by infertility. Anthropological research on informal adoptions has found many other possible and overlapping motivations: to strengthen relationships between the birth parent and foster parent, to have company for one's children or more hands to share the chores, or to demonstrate social prestige and generosity.

Thus, increasing rates of infertility are relevant, but not the whole explanation, for what motivates people to adopt. In other words, an infertile person who does not want to have a child will not be motivated to adopt by his or her infertility.

Changing Sources of Adoptable Children

It is also said that prospective adopters currently desire specific kinds of children: young, healthy, and possibly of a particular

race or gender. One reason that is sometimes offered for high rates of international adoption is that such children are no longer available domestically. What would explain a decrease in the available population of infants? The legalization of abortion (1973 in the U.S.) and increase in availability of birth control meant that fewer unwanted children were born. Perhaps more importantly however, changing social mores meant that through the 1970s and 1980s in the U.S. it became much more accepted for a woman to raise a child as a single mother. It was not that single women were no longer having children – it was that they were less likely to place those children for adoption. Prior to 1973, 9% of single mothers placed their children for adoption (19% white and 2% black); from 1973 to 1981 this figure dropped to 4% and from 1982 to 1988 to 3% (Briggs, 2012: 7).

Another complication is that there are actually children available in the public systems of the U.S. and other countries. But these children are not infants, and many of them are minorities. Scholars have identified significant racial disparities in foster care populations. Indigenous children are overrepresented in foster care in the U.S., Canada, and New Zealand, but the problem is particularly acute in Australia, where 24% of the children in foster care but only 5% of the child population overall are indigenous (Tilbury, 2009: 62). In the U.S., African-American children are less than 20% of the child population but almost 50% of the foster care population, and "Once removed from their homes, Black children remain in foster care longer, are moved more often, receive fewer services, and are less likely to be either returned home or adopted than other children" (Roberts, 2002: vi). Over the course of their childhoods, African-American children have an 11% risk of entering foster care, and Native children a 15% risk (Wildeman and Emanuel, 2013).

That is only a partial picture, however, because it is not only that there are fewer infants 'available' domestically but also that we perceive there to be a surplus of infants available overseas. On one hand there is a narrative about war, disaster, or HIV orphans needing homes (the unusual case of China is an outlier here, in which population policy has directly led to the abandonment of numerous female infants; *see* Family Planning Programs: Feminist Perspectives). However, the majority of children adopted from abroad are not orphans, and have living relatives (Graff, 2008). There is also a Malthusian tinge to this narrative: In many developing countries, fertility rates are higher than they are in the 'receiving countries' of international adoption, and in some cases poor parents cannot care for all their children – particularly when war, natural disasters, or HIV are also features of the landscape. As Selman concludes, "It is, however, evident that the major sources [of internationally adopted children] have not been the poorest or highest birth rate countries, that patterns persist long past the 'crisis' and that demand for children is also a key factor" (Selman, 2002: 218, 219).

Thus, changes in the 'supply' of adoptable children domestically may motivate some adults to look overseas for an available child. The demographic makeup of the 'adoptable' minor population is not an explanation for why adoption happens, however. It is rather an attempt to document differences in that population that might explain why people who have already decided to adopt choose to do so through the public system, privately, or internationally.

Since 2004, numbers of international adoptions have fallen substantially, although numbers of applications to adopt have continued to grow (Selman, 2010, 2009). Decreases since 2004 in numbers of adoptions around the world do not appear to reflect a decline in demand but rather a shrinking supply of available children (reductions in China and Russia and a moratorium on adoptions from Guatemala have all had significant impact) (Selman, 2009: 575, 581). One reason this supply is shrinking is because almost half of the 40 primary source countries for international adoptions over the past 15 years – "places such as Belarus, Brazil, Ethiopia, Honduras, Peru, and Romania – have at least temporarily halted adoptions or been prevented from sending children to the United States because of serious concerns about corruption and kidnapping" (Graff, 2008: 60). Additionally, in 2008 the Hague Convention finally entered into force in the United States, the country with the highest numbers of international adoptions worldwide. This voluntary regulation meant adoptions from Guatemala among other places would stop. That same year, the global economic crisis began, and those economic straits are affecting fertility rates in general as well as people's desire (or financial eligibility) to adopt. International adoptions to the United States have dropped substantially since these changes, from 17 000 in 2008 to 11 000 in 2010 (Child Welfare Information Gateway, 2011: 25).

Neither infertility nor changes in the supply of adoptable children explain why adoptions happen today. They would not happen where there is no desire or demand for them. Although the stated purpose of adoption is to provide a family for a child, many iterations of adoption today look more like a demand for children rather than efforts to protect them. One way of resolving this tension is to argue that parents who want a child are the best thing for a child, justifying the high demand for children as a measure of child protection (Leinaweaver, 2013). And indeed, the primary cause of adoption – the proximate determinant – is that an adult feels socially compelled to, or is driven to, submit to the often invasive and lengthy procedures required of prospective adopters. Desire for a child, and the feeling that an adopted child will satisfy that desire, is what motivates adoption (*see* Children, Value of) (Lozano and Kossoudji, 2009: 20).

Adoption and Cultural Transformation

What remains to be understood is the relationship between adoption and culture. This final section will focus on changes in how adoption itself is socially understood and on the social changes that affect how adoptees are perceived and understood.

Despite the dominant narrative that adoption is first and foremost a measure to protect a child, adoption is understood and conceived of as a family-making project by those adults who enter into it. They do so out of a desire to be a parent, and an adoptive parent in particular. To understand this desire, we need to examine cultural features of adopting societies (*see* Fertility and Culture: Anthropological Insights). For the early stages of adoption – the years preceding it and the first several years of a child's life – we must conceptualize adoption itself as a family-making project. What this means has shifted over time. The profile of a desirable adoptable

child has changed along with it. In the nineteenth century, older children were desirable adoptees because they could collaborate with a family work project, such as a farm. But by the 1930s children were valued for their sentimental qualities and thus ought not to be 'useful,' leading to a desire for helpless babies (Zelizer, 1985: 168). Furthermore, the rise of psychological and child development theories about bonding grounded adults' wishes for a younger child who was more impressionable and had had fewer effects of being institutionalized. Along with these changes have gone legal changes that determine the secrecy and openness in adoption. In the United States, adoption was in the nineteenth century an accepted practice, with records available to adult adoptees; but between 1930 and 1980 almost all states closed adoption records. Initially, this was said to protect adoptive families from birth parents who wanted to contact or claim their children. Later, as a greater pressure to openness prevailed in the 1960s and beyond, the justification shifted and sealed records were defended as a promise of secrecy that had been given to birth mothers upon surrender of their children (Samuels, 2001: 370, 371).

Other recent changes have to do with who is permitted to adopt a child. In many countries single persons or gay and lesbian couples are permitted to adopt openly. This, combined with blended families already familiar from divorce and remarriage or stepparent adoption, and interracial adoptions, has all led to a greater diversity in the makeup of families. In some quarters this is a development to be viewed with great enthusiasm: the normalization of family diversity might mean greater social support for families once viewed as unusual, such as those including single mothers or multiracial families. Yet there is also cause for caution. Simultaneously with the greater normalization of adoption and other kinds of nontraditional families, other social changes are occurring whose impacts on adoptive families are as yet unclear. For example, as internationally adopted, transracial adoptees grow up, and international migration has boomed, adoptees may be 'misidentified' as immigrants, something observed in Europe (Hübinette and Tigervall, 2009: 344; Leinaweaver, 2013) (*see* Migration: Cultural Aspects; Assimilation of Immigrants). In this sense, it is possible that a multicultural context is also a risk factor for transracially adopted youth as long as that multicultural context is characterized by discrimination against minorities.

A final, and critical, cultural transformation is to be found in scholarship and policymaking on adoption. Up till now, most research on adoption has been conducted by adoption professionals and in some cases, adoptive parents. The cohort of adopted adults is growing and research that originates from and takes seriously their standpoints and perspectives is likely to transform the way that we think about adoption. It is even rarer to consider the perspective of birth parents (but see Modell, 1986), but similarly essential to expand the kinds of questions that we ask about adoption, its causes, and its consequences.

In conclusion, the one constant in adoption appears to be its diversity over time. Adoption today looks very different than it did even two decades ago, to say nothing of the differences from its origins in nineteenth century civil code or Roman law. And those changes reflect broader transformations that are of great interest to demography. As Davis has observed, "Although adoptions represent a small portion of family growth, from a demographer's point of view it is significant" (Davis, 2011: 3). Demographers are well positioned to track the ways in which supply of children and demand for adoption have shifted over time. And doing so will continue to yield productive findings about changes in family structure, the value of children, and the consequences of poverty and discrimination.

See also: Adoption: Domestic, International and Global Perspectives; Child Protection: International Issues; Child Protection; Children and Law; Children's Geographies; Children, Rights of: Cultural Concerns; Family Policy; Family and Kinship, History of; Fertility and Culture: Anthropological Insights; Infertility: Demographic Aspects; Kinship in Anthropology; Motherhood; Reproduction and Cultural Anthropology; Trafficking of Women and Children in Latin America.

Bibliography

Briggs, L., 2012. Somebody's Children: The Politics of Transracial and Transnational Adoption. Duke University Press, Durham.

Child Welfare Information Gateway, 2011. How Many Children Were Adopted in 2007 and 2008? U.S. Department of Health and Human Services, Children's Bureau, Washington, DC.

Davis, M.A., 2011. Children for Families or Families for Children: The Demography of Adoption Behavior in the U.S. Springer Science+Business Media B.V, Dordrecht.

Dubinsky, K., 2010. Babies without Borders: Adoption and Migration across the Americas. University of Toronto Press, Toronto.

Flango, V., Flango, C., 1995. How many children were adopted in 1992? Child Welfare 74, 1018–1032.

Font Lletjòs, E., 16 March–April 2008. "Perfil de les families adoptants a Catalunya." Infancia: Bulletí dels professionals de la infància I la adolescencia.

Graff, E.J., 2008. The lie we love. Foreign Policy 169, 59–66.

Haworth, G., Selman, P., Way, J., 2010. Infertility and inter-country adoption. In: Crawshaw, M., Balen, R. (Eds.), Adopting after Infertility: Messages from Practice, Research and Personal Experience. Jessica Kingsley Publishers, London.

Hübinette, T., Tigervall, C., 2009. To be non-white in a colour-blind society: conversations with adoptees and adoptive parents in Sweden on everyday racism. Journal of Intercultural Studies 30, 335–353.

The Hague, 1993. Hague Convention on Intercountry Adoption (Online). Available: http://www.webcom.com/~kmc/adoption/law/un/un-ica.html.

Kane, S., 1993. The movement of children for international adoption: an epidemiological perspective. Social Science Journal 30, 323–339.

Leinaweaver, J.B., 2013. Adoptive Migration: Raising Latinos in Spain. Duke University Press, Durham.

Lovelock, K., 2000. Intercountry adoption as a migratory practice: a comparative analysis of intercountry adoption and immigration policy and practice in the United States, Canada and New Zealand in the post W.W. II period. The International Migration Review 34, 907–949.

Lozano, F.A., Kossoudji, S.A., 2009. The Unknown Immigration: Incentives and Family Composition in Intercountry Adoptions to the United States. IZA Discussion Papers, 4547.

Marre, D., 2009. We do not have immigrant children at this school, we just have children adopted from abroad: flexible understandings of children's 'origins'. In: Marre, D., Briggs, L. (Eds.), International Adoption: Global Inequalities and the Circulation of Children. New York University Press, New York, pp. 226–243.

Modell, J.S., 1986. In search: the purported biological basis of parenthood. American Ethnologist 13, 646–661.

Roberts, D.E., 2002. Shattered Bonds: The Color of Child Welfare. Basic Books, New York.

Rojewski, J.W., Rojewski, J.L., 2001. Intercountry Adoption from China: Examining Cultural Heritage and Other Postadoption Issues. Bergin & Garvey, Westport, CT.

Samuels, E.J., 2001. The idea of adoption: an inquiry into the history of adult adoptee access to birth records. Rutgers Law Review 53, 367–437.

Selman, P., 2002. Intercountry adoption in the new millennium; the "quiet migration" revisited. Population Research and Policy Review 21, 205–225.

Selman, P., 2009. The rise and fall of intercountry adoption in the 21st century. International Social Work 52, 575–594.

Selman, P., 2010. Intercountry adoption in Europe 1998–2008: patterns, trends and issues. Adoption & Fostering 34, 4–19.

Smolin, D.M., 2005. Intercountry adoption as child trafficking. Valparaiso Law Review 39, 281–325.

Tilbury, C., 2009. The over-representation of indigenous children in the Australian child welfare system. International Journal of Social Welfare 18, 57–64.

U.S. Department of Health and Human Services, 2009. The AFCARS report: preliminary FY 2008 estimates as of October 2009 (16) (Online). Available: http://www.acf.hhs.gov/sites/default/files/cb/afcarsreport16.pdf.

Vandivere, S., Malm, K., Radel, L., 2009. Adoption USA: A Chartbook Based on the 2007 National Survey of Adoptive Parents. The U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation, Washington, DC.

Weil, R.H., 1984. International adoptions: the quiet migration. International Migration Review 18, 276–293.

Wildeman, C., Emanuel, N., 2013. Cumulative Risks of Foster Care Placement by Age 18 for U.S. Children, 2000–2011. Population Association of America, New Orleans.

Zelizer, V.A., 1985. Pricing the Priceless Child: The Changing Social Value of Children. Princeton University Press, Princeton.

Relevant Websites

http://pages.uoregon.edu/adoption/ – The Adoption History Project. This digital humanities project, an ongoing effort of Professor Ellen Herman at the University of Oregon collects and synthesizes archival sources on the history of child adoption in the U.S.

https://www.childwelfare.gov/ – Child Welfare Information Gateway. The U.S. Department of Health and Human Services offers this web site full of practical information about adoption as well as data useful for adoption professionals.

http://www.adoptioninstitute.org/ – Evan B. Donaldson Adoption Institute. This nonprofit, U.S.-based organization offers statistics, research, educational, and policy information about adoption.

http://www.hcch.net/ – Hague Conference on Private International Law. This web site provides the text of the Hague Convention of 29 May 1993 on protection of children and cooperation in respect of intercountry adoption in more than a dozen languages.

http://aspe.hhs.gov/hsp/09/NSAP/ – The National Survey of Adoptive Parents (NSAP). Results from this survey of adoptive families, conducted in the U.S. in 2007-08, are available online.

Adoption: Domestic, International and Global Perspectives

Janette Logan, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Adoption transcends all cultures and has existed over centuries. This article charts the evolution of adoption from Moses in the Bible to contemporary domestic and intercountry adoption practices. Internationally and globally it has evolved in response to changing political, economic, and social circumstances, acting as a solution to social problems of the time. An adoption order is the most radical order that can be made in the field of family law for it involves the permanent transfer of all legal rights and responsibilities from birth parents to a new family.

Introduction

Adoption is the process by which the legal relationship between a child and his or her birth parents is severed and a new relationship established with adoptive parents. An adoption order is the most radical order that can be made in the field of family law for it involves the permanent transfer of all legal rights and responsibilities from birth parents to a new family. Once an adoption order is made, parentage is completely and irrevocably transferred to the adoptive family. Adoption demonstrates a sense of optimism about society's ability to engineer happy outcomes for children while at the same time, it deprives birth parents of any legal parental status and historically, has rendered them strangers to their children. At best, adoption conjures up promises of family life for adults and children who otherwise would have been without it. At worst, it represents the most Draconian form of state intervention in family life (Smith and Logan, 2004). This article considers the evolution of adoption and makes some suggestions for practitioners working in this field. It also refers mainly to the specifics of Anglo-American societies. Adoption in other parts of the world raises different socioeconomic, political, and cultural issues that require separate consideration, e.g., Latin American and African societies. However, space prevents me from doing so here.

Because adoption involves the raising of children by parents and families other than those to whom they were born, it is a complex and emotional subject. It profoundly changes the course of everyone's life, for the rest of their life. For children, it changes their primary relationship with their birth parents and can raise concerns about why they were given up for adoption with profound curiosity about their origins. The impact on birth parents and adoptive parents can be equally profound as their lives are changed forever.

For this reason, adoption has attracted immense public and political interest. It is a controversial and sensitive subject, touching on issues of love and loss, nature and nurture, identity, ethics, and human rights. Transracial and intercountry adoptions (ICA) and adoptions without parental consent further magnify these issues.

Adoption raises important questions concerning the family and the relationship between biological and social parenting. Traditionally in Western society/culture, the blood bond has served as the basis for the creation and continuation of family kinship systems, and common sense notions of 'blood is thicker than water' and 'our own flesh and blood' prevail to this day. Adoption challenges these assumptions for it represents a remarkable relationship between the legal definition and effects of an adoption order and the way in which adopters and their children develop bonds of love and attachment through the everyday social construction of family life.

Modern adoption has adapted rapidly to social change and throughout history it has acted as a solution to social problems of the time. Adoption today is diverse and complex, and globally, the scope and meaning of adoption varies significantly.

In contemporary Western societies, there are two types of adoption; domestic and intercountry. In the UK, the majority of adoptions are domestic and ICA account for only 4% of the total number. In contrast, domestic adoptions in other European countries are rare and ICA more commonplace. In the US, both types of adoption feature with the US adopting more children both internationally and domestically than the rest of the world combined (Pertman, 2013). US domestic adoptions have much in common with the UK as in both countries the fate of children in the care of the state has become the major policy issue.

The Social Construction of Adoption – A 'Legal Fiction'

The meanings attached to adoption are considerably more complex than a straightforward legal mechanism would suggest. Adoption has to be understood as both a legal reality and a socially constructed phenomenon, which is influenced by the historical, political, cultural, and social context in which it is located. Adoption, like everything else, takes its meaning from the world around it and adoption means different things in different times. This is evident in cross-cultural studies of adoption (Silk, 1980; Terrell and Modell, 1994; Berebitsky, 2006) and examinations of how domestic adoptions in the US and UK have developed.

Since its legal origin in 1926 in the UK (1930 in Scotland), adoption has acted as a barometer for social change, reflecting, and responding to issues of public and political interest, representing something of a litmus test in respect of much larger issues such as: what 'the family' should look like, how should

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28002-1

the state intervene when a parent is unable to look after their child, and when family life is fundamentally a private matter?

In order to understand parenthood that is a contractual arrangement based on law in a culture in which parenthood is created by birth, Judith Modell contends that the constructed relationship must be as much like the biological one as possible (Modell, 1994). Traditional adoption in Western societies was thus constructed to mirror biological kinship – a socially and legally endorsed form of kinship equivalent to blood ties with the particular practice of closed adoptions and secrecy acting to support this perspective. Adoption legislation in the UK and US continue to regulate the construction of adoptive families through 'legal fictions' – eradicating a child's ties to birth kin and incorporating him or her into the adoptive family 'as if' begotten, and the parent 'as if' genealogical (Smith and Logan, 2002, 2004). The Adoption Act, 1976 and the Adoption and Children Act, 2002 in the UK perpetuate the fiction that adopted children have no family other than their adoptive family (Ryburn, 1997; Smith and Logan, 2002). The Hague Convention similarly utilizes legal fictions with regards to ICA.

Critics have objected to the wording contained in these pieces of legislation, claiming they support a construction of adoption that denies the ongoing significance of birth family members in children's lives. In its review of adoption law, the Law Commission of New Zealand (2000: 43–44) recommended that the 'legal fiction' incorporated in their legislation is a 'repugnant and unnecessary distortion of reality' and should be removed from future legislation.

Sir Henry Maine, a British jurist in the nineteenth century, was the first to acknowledge the 'fictive' nature of adoption:

> We must try to regard the fiction of adoption as so closely simulating the reality of kinship that neither law nor opinion makes the slightest difference between a real and an adoptive connection.
> (Maine, 1861: 239, cited in Modell, 1994)

The law is a powerful force reflecting societal values and determining who can and cannot adopt. In an endeavor to reflect biological kinship alongside assumptions of what constitutes a 'normal' family, adoption had historically been the prerogative of heterosexual married couples and legislation had outlawed the adoption of children by unmarried, including gay and lesbian, couples (Logan and Sellick, 2007). Single people were, however, able to adopt. It is only since the introduction of the Adoption and Children Act, 2002 that lesbians and gay men have been allowed to adopt in the UK and it continues to be illegal in some States in the US. Nonetheless, adoption by same sex couples is growing in the US but an uneven legal landscape, due to laws prohibiting same sex marriage, means that adopted children are without the rights and protection extended to children adopted by heterosexual couples (Tavernise, 2011).

The Evolution of Adoption

Adoption transcends all cultures and has existed over centuries. The story of Moses in the Bible represents the first account of an adoption. Forms of adoption are also recorded in ancient Greece and Rome where its roots lie in legal mechanisms for the acquisition of heirs and transmission of property.

A historical examination of adoption practice in the Western tradition reveals a number of distinct periods, each one reflecting the preoccupations of the time, with adoption serving different functions at different times (Triseliotis et al., 1997).

In the nineteenth century period of industrialization, adoptions were concerned with providing security for children orphaned or born illegitimately. Disease, death, and destitution lead to increasing numbers of children placed in workhouses, the conditions of which were so poor there were serious concerns about infant mortality. Consequently, children were 'indentured' into families who would clothe, feed, and educate them in return for helping out the family with its work. These arrangements represented a protoform of adoption (Howe, 2009) and were evident in the UK and US. In some American states they became legalized through specific acts of legislature in the midnineteenth century, and the adoption statute passed in Massachusetts in 1851 became the model for subsequent adoption legislation in the US. In Britain, they remained *de facto* adoptions with no legal security. It was not until 1926 that the first adoption law was introduced in England (1930 in Scotland) – the impetus for which was the large numbers of orphaned or illegitimate children born during and after the First World War. During the period between the two world wars a number of societies were set up in both the UK and US to arrange such adoptions and by 1929 all states in the US had adoption legislation. In order to protect the privacy of the adoptive family and to protect the child from the stigma of illegitimacy these adoptions were 'closed' with no contact between the child's old and new families and the child was issued with a new birth certificate. The emphasis was on a fresh start and a clean break from their families of origin.

For much of its history, adoption has been seen primarily as a solution to the problem of illegitimacy, unwanted pregnancy, and the stigma and shame attached to these. After the Second World War, adoption was viewed as a solution to the problem of married couples' infertility, and the emphasis was on meeting the needs of infertile couples rather than those of the baby. Known as the era of 'the perfect baby for the perfect couple' great emphasis was put on matching the child to the adoptive parents in an attempt to ensure as much similarity as possible between the placed baby and the new family.

In both the US and UK, the 1960s and 1970s witnessed a shift from adult-centered practice to one that focused on the child. A number of factors were influential in this move toward a more child-centered approach: the legalization of abortion, the availability of contraception, the increasing acceptance of single parenthood, and availability of state welfare benefits led to a huge drop in the number of young White children available for adoption. At the same time, the number of couples wanting to adopt was on the increase. It was also becoming increasingly apparent that children who were not able to be looked after by their birth families, were drifting in the public care system, with no long-term plans for their future. The 1980s saw the emergence of the 'permanency movement' in the US and UK, in which adoption became seen as a positive placement of choice for these children, many of whom had

suffered early adversity. As a consequence, older children, children with disabilities, non-White children and sibling groups were increasingly considered for adoption, with an emerging mantra that 'no child is unadoptable'.

Domestic Adoption Today

In the US and UK adoption has now evolved into a means of providing a solution to the needs of children in public care. These children are unable to return to their birth family due to the risk of maltreatment or neglect, resulting from serious parental problems such as drug and alcohol misuse or mental health problems. It has become identified as child centered, long lasting and potentially advantageous over other placement options such as long-term foster care and residential care. Research has demonstrated that it has positive outcomes, particularly for infants and very young children (a discussion of adoption outcomes is complex and outside the scope of this article; see Rushton and Dance, 2003, Selwyn et al., 2006; Steele et al., 2008).

In the UK, the age and profile of children being adopted has changed considerably. In 1971 there were 21 495 adoptions, most of which were baby adoptions. However, domestic adoptions of infants are now rare and the last 40 years has witnessed the growth of adoption as a placement option for older children. Many of these have special needs and histories of abuse and neglect and the adoption of 'looked after' children is now a central feature of the child care system in the UK.

In 2013, the number of children adopted from care in England was 3980 of which only 2% (90) were under 1-year-old. The majority (74% or 2960) were between 1 and 4 with the average age being 3 years 8 months. These adoptions account for only 5% of the total number of looked after children, most of whom are placed in foster care. The majority of adopters are still heterosexual couples; 90% compared to 10% single adopters and 6% same-sex couples (www.baaf.org.uk/res/statengland).

Adoption in the UK is highly regulated with every local authority serving as an adoption agency. The voluntary sector also has an important role, both in facilitating adoptions and in the provision of postadoption support. The adoption process has become an active legal contest between the state and birth parents centering on the culpability of the parents in respect of abuse and neglect and whether they can or cannot change to adequately care for their child. While adoptions with parental consent are legally possible, rarely do parents agree with the local authority that they have abused or neglected their child and in these cases consent is dispensed with by the court as the best interests of the child require it (Simmonds, 2012).

From an ethical and human rights perspective adoption continues to be a controversial subject (Harris-Short, 2008). Dispensing with parental consent and severing the legal tie of children to their birth family through adoption is seen as an infringement of human rights in many European countries and therefore not an option available. In these countries long-term fostering and/or residential care are favored over domestic adoption as placements of choice.

From Secrecy to Openness in Adoption

The most significant change in the nature of adoption concerns the move away from confidential, closed adoptions to a more open approach. More than anything else, the issue of adoptee–birth family contact has had a far-reaching impact on the changing face of adoption today and has dominated adoption policy and practice since the mid-1980s.

For many years the 'secrecy is best' philosophy dominated practice in domestic adoptions in the UK and US. However, it is now acknowledged that this fresh start and clean break approach to adoption naively ignored the complexities that adoption would create for birth families and adoptive families with research and personal testimonies emerging to challenge the simplicity of this approach (Triseliotis, 1973; Winkler and Van Keppel, 1984). Origins, identity, and heredity continue to be important (Howe and Feast, 2000) and it is now acknowledged that many of the psychological problems that may affect adopted people, adoptive parents, and birth parents are directly related to the secrecy or anonymity of the closed traditional form of adoption.

The move toward a more open approach to adoption was also fueled by the 'search movement', initially led by Florence Ladden Fisher in 1971 in the US and rapidly spreading to Canada and the UK. 'Identity politics' and the rise of new social movements at the time informed her ideology of searching based on adoptee rights. However, their inability to gain access to adoption records by claiming constitutional rights led to a second generation of search leaders with an ideology based on psychological need (Carp, 2002). Led by three adoption researchers in the US (Baran et al., 1974) their work emphasized that searching stemmed from "an innate curiosity about their genealogical past" (Baran et al., 1974: 532) and the need to establish a clearer sense of identity. A third generation of the movement emerged in 1996, exemplified in the activism of 'Bastard Nation'. This group challenged mainstream adoption activist organizations with their emphasis on 'psychological need' and the facilitation of unions between adoptees and their birth relatives, and instead, once again, focused on adoptee rights.

In non-Western cultures, open adoption is not a new phenomenon. Adoption arrangements are publicly acknowledged and both sets of parents are well known to each other. This notion of a dual connection is an important aspect of adoption in the Pacific Islands of Oceania (Silk, 1980) and similarly, in the Maori system of New Zealand, adoptions are public knowledge and children's relationships with their birth parents well known and understood (Rockel and Ryburn, 1988). Generally, Black and Asian, Hispanic and Native American families are far more familiar than White communities with the idea of contact and in fact, expect it, whether in short- or long-term child care arrangements.

Postadoption contact arrangements between adopted children and their birth family members are a central feature of most contemporary domestic adoptions in the UK and US and may involve face-to-face meetings with birth relatives (direct) or contact by phone, letters, e-mail (indirect) (Parker, 1999, Neil and Howe, 2004, Grotevant and McRoy, 1998).

However, contact is a complex area of work and practice in openness has evolved to include a number of key dimensions

indicating that 'indirect' and 'direct' contact oversimplify the complexity of openness as it has developed over the last decade or more. Focusing on what form the contact should take (e.g., letters or meetings) neglects the nuances and complexities inherent in both forms of practice. For example, both forms of contact may or may not involve some form of mediation (by the adoptive parents or the placing agency), may vary according to the level of identifying information the parties may hold about each other, and, crucially, may vary in the extent to which the children are involved. New forms of contact emerging through the use of social networking sites such as Facebook, further adds to the complexity (Smith and Logan, 2004). Within the research community 'communicative openness' (Brodzinsky, 2006) – the openness of adoptive parents in thinking and talking about adoption has been identified as equally, if not more important than what kind of contact takes place. Whatever its form, the debate about openness has led to a polarization of views within the adoption community and an adversarial relationship between advocates for, and critics of, openness. While few would challenge the need for openness about children's preadoption history, there is much less certainty about the circumstances in which adoption with some form of continuing contact should be planned. Nonetheless, in the UK there is now a clear expectation that children will have *some form* of contact with their birth relatives (Lowe et al., 1999, Parker, 1999) and the Adoption and Children Act, 2002 specifies that contact issues have to be thoroughly explored prior to the making of an adoption order. Similar issues apply to US domestic adoptions and there is emerging evidence that postadoption contact is becoming a feature of ICA.

Intercountry Adoption – The Globalization of Adoption

ICA feature mainly in the US, the Nordic countries, Spain, and other parts of western Europe. Originally it emerged as a global legal phenomenon involving formal agreements between sending and receiving countries as a response to the needs of orphaned children as a consequence of war. After the Second World War, international adoption was largely about the movement of children from countries in Europe and Japan to the US. These altruistic motives for ICA continued from the mid-1950s with large number of children from the Korean and Vietnamese wars being adopted overseas (Selman, 2009). By the 1970s ICA were well established in the US, mainland Europe, and Australia, and was seen primarily a service for childless couples in the West and the children were victims of poverty rather than war. This period witnessed a drastic decline in the availability of babies for domestic adoptions in Western countries due to increasing access to contraception and abortion and a growing acceptance of single parenthood.

It is estimated that since Second World War, ICA have been a solution for approximately one million children rendered parentless due to political, economic, and social circumstances in less developed countries. The opening of China and Russia in the 1990s saw an 'explosion' of ICA with 410 000 children adopted by citizens of 27 countries between 2000 and 2004. In Russia the break up of the Soviet Union was a significant factor and in China it was the discovery of the impact of the one-child policy, which lead to the abandonment and subsequent institutionalization of mainly female children. Under the Ceausescu regime in Romania, women had to have a minimum of four children, which led to many children entering institutions because their parents could not support them. ICA were seen as a preferable option to the negative consequences of institutional care (for detailed analyses of ICA, see Selman, 2009, 2012a,b).

After decades of steady growth, ICA are now on the decline and the number of ICA dropped by nearly 50% between 2004 and 2010 (Selman, 2012a). Paradoxically, however, the numbers of children sent abroad for adoption from the US is on the increase! These are mainly Black American babies whose birth mothers are choosing to have them adopted – mainly in Canada and Holland, in the belief that they will escape the evils of racism. In 2009, 315 American children were adopted internationally. While considerably fewer than overseas orphans joining US families, this nonetheless represents a threefold increase of American children being adopted overseas since 2004!

The global decline of ICA is not due to fewer orphans worldwide nor less demand from prospective parents but increasing regulation and a growing concern about irregularities in practice. Initially viewed as a humanitarian exercise in rescuing children from countries ravaged by war or natural disaster (e.g., after the earthquake in Haiti in 2010), more recently, ICA have generated concern as a form of Western imperialism, with commercial interests taking priority over the rights and welfare of children or birth parents. Children in ICA are frequently seen as objects of commodification, available for a price that only wealthy parents can afford. The extreme case of commodification is the rumor that children are being stolen from poor countries and sold for their organs in rich ones (Khabibullina, 2009).

The 1993 Hague Convention for the Protection of Children was introduced in an attempt to regulate adoption arrangements in both sending and receiving countries to ensure that policies and practices became child centered. Yet, there have been continuing reports of irregularities in adoption and of adoption as a cover for child trafficking. In the past decade a number of adoptive parents in the US have been found guilty of killing their adopted child, and in 2009, Russia suspended the adoption of Russian children by American families after Artyom Savelyev, a 7-year-old adopted Russian boy was rejected by his American mother and sent back alone to Russia. David Smolin, who discovered that his two adopted daughters from India had been stolen, has written extensively on the topic of irregularities in ICA (Smolin, 2006, 2011). He argues for much stronger regulation of practice, without which, ICA should be abolished. In contrast, advocates of ICA argue that, if reformed, ICA could increase once more and form a key component of globalized child welfare in the same way that adoption has been seen as the placement of choice for abused children in the US and UK (Bartholet, 2011).

Nonetheless, research outcomes demonstrate that ICA can offer new hope to children who might otherwise spend their childhood in institutions (Juffer and van Ijzendoorn, 2009). The challenge is to continue this support for individual children while encouraging the development of adoption and foster care in the child's country of origin.

Race, Ethnicity, Identity, and Belonging: Contemporary Issues in Adoption

Race and ethnicity in adoption bring to the foreground political, social, psychological, and legal issues calling into question our ideas of racial and ethnic boundaries, identity, and belonging (Barn and Kirton, 2012; Lind, 2012; Hubinette and Andersson, 2012). According to the UN Convention on the Rights of the Child (UNCRC, 1989), a child's ethnic, religious, cultural, and linguistic background should be paid due regard in adoption, as should the desirability of continuity in a child's upbringing (Article 20). Article 16b of the Hague Convention on the Protection of Children and Co-operation in Respect of Intercountry Adoption (HCCH, 1993) contains similar wording. Both conventions also stipulate that ICA should only take place after possibilities for placing the child within the country of origin have been explored (UNCRC, 1989, Article 21b; HCCH, 1993, Article 4b).

However, adoption has been associated with violations of the rights of ethnic minority groups. In both the US and the UK, the practices of transracial placement and racial matching within domestic adoptions have a history of fierce debate with proponents for and opponents against transracial adoption. Dating back to the 1960s, several factors form the backdrop of this debate: the disproportionate representation of minority ethnic children in the public care system, the 'unavailability' of suitable minority ethnic adopters, concerns about adoptees' racial/cultural identity and the 'suitability' of White parents to raise racially competent children (Barn and Kirton, 2012).

Research in the 1980s in the UK, established that while children placed transracially successfully met developmental milestones, questions remained about the long-term impact on the child being dislocated from their ethnic and cultural heritage (Gill and Jackson, 1983). The adoption of Black children by White parents was seen as evidence of the colonization of the Black community by removing its children and 'making them White'. As a consequence, a commitment to same race placements became dominant in policy and practice. Only relatively recently has this policy been challenged on a number of grounds: the acknowledgment that society is now a complex mix of ethnic groups, identity is not fixed and singular but changes over time and in different contexts and, particularly significant in the UK, the serious delays experienced by many minority ethnic children in care, which can result in them not being placed for adoption at all (Simmonds, 2009).

In the UK, the legal requirement (under the 1989 Children Act and 2002 Adoption and Children Act) is to give due consideration to the child's religious persuasion, racial origin and cultural and linguistic background with strongly worded guidance that this should not prohibit transracial placements nor lead to delay while an ethnic match is sought. However media headlines suggest that White parents are being denied the opportunity to offer suitable adoptive homes to minority ethnic children in need (for a comprehensive discussion of the UK context, see Barn and Kirton, 2012) and currently the English Government is seeking to establish a policy that will minimize the consideration of a child's ethnicity and, in consequence, promote transracial adoption of minority ethnic children in the UK.

Similar debates prevailed in the US until the Multi-Ethnic Placement Act 1994 and the Inter-Ethnic Adoption Provisions 1996 were passed to introduce legally enforced color blindness. The aim was to remove the barriers to transracial placements though this has not had the desired effect (for a comprehensive discussion of the US context, see McRoy and Griffin, 2012).

When adoption takes place between different countries, questions of identity, belonging, ethnicity, and culture become even more prominent. Despite this, for many years ICA were not considered transracial placements in some countries. Unlike the UK and US, for example, there has been no debate about whether White people should adopt people of color in Sweden. The societal and cultural norm has been one of official antiracism and color blindness and only in the last decade have ICA been identified as transracial adoption, a view resisted by many adoptive parents and agency officials (for a comprehensive discussion of the Swedish experience, see Lind, 2012; Hubinette and Andersson, 2012).

A number of studies have addressed identity in ICA, for example, Volkman in Sweden (2005), Dorow in the US (2006), Howell in Norway (2009), Gray in Australia (2009), and Richards in the UK (2012). All support the argument that ethnicity and culture are fluid and flexible concepts, meaning different things to different people within specific communities and change across space and time. With support, adoptees are able to create their own narratives of belonging and construct their own sense of identity. Nonetheless, this is a complex process and narratives of transnational adoptees highlight the "contingent struggles and ambiguities of intercountry adoption" (Richards, 2012: 110).

Conclusion

This article has outlined the changing nature of adoption. Internationally and globally, it has evolved in response to changing political, economic, and social circumstances. In some form or other, controversy and conflict have featured throughout its evolution and will no doubt continue to mark the development of adoption policy and practice for years to come (Simmonds, 2012).

See also: Child Neglect; Child Protection: International Issues; Child Protection; Children and Families in Social Work; Critical Social Work Practice; Social Work Theory.

Bibliography

Baran, A., Pannor, R., Sorosky, A.D., 1974. Adoptive parents and the sealed record controversy. Social Casework 55, 531–536. Cited in Carp, W.E., 2002. Adoption, blood kinship, stigma, and the adoption reform movement: a historical perspective. Law and Society Review 36 (2), 433–459. This is exactly how it appears in Carp 2002.

Bartholet, E., 2011. International adoption: a way forward. New York Law School Review 55, 687–699. Cited in Selman, P., 2012. The Global decline of intercountry adoption: what lies ahead? Social Policy and Society 11 (3), 381–397. This is exactly how it appears in Selman 2012.

Berebitsky, J., 2006. Family ideals and the social construction of modern adoption: a historical perspective. In: Weger, K. (Ed.), Adoptive Families in a Diverse Society. Rutgers University Press, Rutgers, NJ, pp. 29–42.
Barn, R., Kirton, D., 2012. Transracial adoption in Britain: politics, ideology and reality. Adoption and Fostering 36 (3/4), 38–48.
Brodzinsky, D.M., 2006. Family structural openness and communicative openness as predictors in the adjustment of adopted children. Adoption Quarterly 9, 1–18.
Carp, W.E., 2002. Adoption, blood kinship, stigma, and the adoption reform movement: a historical perspective. Law and Society Review 36 (2), 433–459.
Gill, O., Jackson, B., 1983. Adoption and Race: Black, Asian and Mixed Race Children in White Families. Batsford, London.
Grotevant, H., McRoy, R., 1998. Openness in Adoption: Exploring Family Connections. Sage, London.
Harris-Short, S., 2008. Making and breaking family life: adoption, the state and human rights. Journal of Law and Society 35 (1), 28–51.
Howe, D., Feast, J., 2000. Adoption, Search and Reunion: The Long Term Experience of Adopted Adults. Children's Society, London.
Howe, D., 2009. Nature, nurture and narratives. In: Miller Robel, G., Neil, E. (Eds.), International Advances in Adoption Research for Practice. Wiley-Blackwell, Oxford, pp. 3–16.
Howell, S., 2009. Return journeys and the search for roots: contradictory values concerning identity. In: Marre, D., Briggs, L. (Eds.), International Adoption: Global Inequalities and the Circulation of Children. New York University Press, New York and London, pp. 256–270.
Hubinette, T., Andersson, M., 2012. Between colourblindness and ethnicisation: transnational adoptees and race in a Swedish context. Adoption and Fostering 36 (3 & 4), 97–103.
Juffer, F., van Ijzendoorn, M.H., 2009. International adoption comes of age: development of international adoptees from a longitudinal and meta-analytical perspective. In: Miller Wrobel, G., Neil, E. (Eds.), International Advances in Adoption Research for Practice. Wiley-Balckwell, London, pp. 169–192.
Khabibullina, L., 2009. International adoption in Russia: 'market', 'children for organs' and 'precious' or 'bad' genes. In: Marre, D., Briggs, L. (Eds.), International Adoption: Global Inequalities and the Circulation of Children. New York University Press, New York, p. 174, 175.
Logan, J., Sellick, C., 2007. Lesbian and gay fostering and adoption: prejudice, progress and the challenges of the present. Social Work and Social Sciences Review 13 (2), 35–47.
Lowe, N., Murch, M., Borkowski, M., Weaver, A., Beckford, V., Thomas, C., 1999. Supporting Adoption: Reframing the Approach. BAAF, London.
Lind, J., 2012. 'As Swedish as anybody else' or 'Swedish, but also something else'? Discourses on transnational adopted identities in Sweden. Adoption and Fostering 36 (3 & 4), 85–96.
Maine, H.S., 1861. Ancient Law. J. Murray, London in Modell, J., 1994. Kinship with Strangers: Adoption and Interpretations of Kinship in American Culture. University of California Press, Berkeley, CA.
Modell, J.S., 1994. Kinship with Strangers; Adoption and Interpretations of Kinship in American Culture. University of California Press, Berkeley, CA.
Marre, D., Briggs, L. (Eds.), 2009. International Adoption: Global Inequalities and the Circulation of Children. New York University Press, New York.
McRoy, R., Griffin, A., 2012. Transracial adoption policies and practices: the US experience. Adoption and Fostering 36 (3–4), 38–48.
Neil, E., Howe, D. (Eds.), 2004. Contact in Adoption and Permanent Foster Care. BAAF, London.
Parker, R., 1999. Adoption Now: Messages From Research. John Wiley, Chichester & sons.
Pertman, A., 17 September 2013. Adoption Nation. Cited in Viogt, K., Brown, S., International adoption in decline as the number of orphans grow, CNN.
Richards, S., 2012. What the map cuts up the story cuts across': narratives of belonging in intercountry adoption. Adoption and Fostering 36 (3/4), 104–111.
Rockel, J., Ryburn, M., 1988. Adoption Today: Change and Choice in New Zealand. Heinemann Reed, Auckland.
Rushton, A., Dance, C., 2003. Preferentially rejected children and their development in permanent family placements. Child and Family Social Work 8 (4), 257–267.
Ryburn, M., 1997. The uneven scales of justice: private law contact applications in divorce and adoptions. Adoption and Fostering 21 (3), 23–34.
Selwyn, J., Sturgess, W., Quinton, D., Baxter, C., 2006. Costs and Outcomes of Non Infant Adoptions. BAAF, London.
Steele, M., Hodges, J., Kanuik, J., Hillman, S., Asquith, K., 2008. Forecasting outcomes in previously maltreated children: the use of the AAI in a longitudinal adoption study. In: Steele, H., Steele, M. (Eds.), Clinical Applications of the Adult Attachment Interview. Guilford Press, New York, pp. 427–451.
Selman, P., 2009. Intercountry adoption, research, policy and practice. In: Schofield, G., Simmonds, J. (Eds.), The Child Placement Handbook. BAAF, London, pp. 276–303.
Selman, P., 2012a. 'The global decline of intercountry adoption: what lies ahead'? Social Policy and Society 11 (3), 381–397.
Selman, P., 2012b. The rise and fall of intercountry adoption in the 21st century: global trends in from 2001–2010. In: Gibbons, J., Rotabi, K. (Eds.), Intercountry Adoption. Ashgate, Farnham, pp. 7–28.
Simmonds, J., 2009. Adoption - developmental perspectives within an ethical, legal and policy framework. In: Schofield, G., Simmonds, J. (Eds.), The Child Placement Handbook. BAAF, London, pp. 220–240.
Simmonds, J., 2012. Adoption: from the preservation of the moral order to the needs of the child. In: Davies, M. (Ed.), Social Work with Children and Families. Palgrave Macmillan.
Smith, C., Logan, J., 2004. After Adoption: Direct Contact and Relationships. Routledge, London.
Smith, C., Logan, J., 2002. Adoptive parenthood as a 'legal fiction'. Its consequences for direct post adoption contact. Child and Family Law Quarterly 14 (3), 563–573.
Smolin, D., 2006. How the intercountry adoption system legitimizes and incentivizes the practices of buying, trafficking, kidnapping and stealing children. Wayne Law Review 52 (1), 113–200. http://works.bepress.com/david_smolin/1.
Smolin, D., 2011. The missing girls of China: population, policy, culture, gender, abortion, abandonment, and adoption in East-Asian perspective. Cumberland Law Review 41 (1), 1–665. http://works.bepress.com/david_smolin/9.
Silk, J.B., 1980. Adoption and kinship in Oceania. American Anthropologist 82, 799–820.
Tavernise, S., 13 June 2011. Adoption by Gay Couples Rise Despite Barriers. New York Times.
Terrell, J., Modell, J., 1994. Anthropology and adoption. American Anthropologist 96, 155–161.
Triseliotis, J., Shireman, J., Hundleby, M., 1997. Adoption: Theory, Policy and Practice. Cassell, London.
Triseliotis, J., 1973. In Search of Origins. Routledge and Kegan Paul, London.
Winkler, R.C., Van Keppel, M., 1984. Relinquishing Mothers in Adoption: Their Long Term Adjustment. Monograph no. 3. Institute of Family Studies, Melbourne.

Adult Education and Lifelong Learning: The US Experience and Beyond

Jennifer A Margrett and Kyu ho Lee, Iowa State University, Ames, IA, USA

Published by Elsevier Ltd.
This article is a revision of the previous edition article by S.L. Willis, J.A. Margrett, volume 1, pp. 299–304, © 2001, Elsevier Ltd.

Abstract

The world is poised for a major educational revolution centered on adult education and lifelong learning. By 2050, the number of persons with 65 years of age and older will be greater than children five and under. 'Nontraditional' students nearly equal the number of traditional students enrolled in higher education and adults are seeking educational and learning opportunities for a multiplicity of reasons. The growing and diverse group of adult students and learners presents opportunities and challenges to educators across settings.

Emergent Population of Adult Learners

A steady increase in average life expectancy in recent years, coupled with the aging of the baby boomer generation, has given rise to a growing number of adults with 60 years of age and older in the United States. In the US, the average life expectancy is 75.7 and 80.8 years for men and women, respectively (U.S. Census Bureau, 2012). This changing demographic is not unique to the United States. According to the Organisation for Economic Co-operation and Development (OECD, 2011), which represents a consortium of over 30 countries across North and South America, Europe, and the Asian-Pacific region, a newborn girl in 2008 is expected to live up to 82.0 years, which is 11 years more than a baby girl born in 1960. A baby boy is also expected to live up to 76.5 years, which is 11 years more than the one born in 1960 (OECD). As a result, the proportion of older adults in many societies is growing. In 1960, only 8.5% of the global population was 65 years of age and older, whereas this percentage rose to13.8% in 2005 (OECD, 2007). "Very low birth rates in developed countries, coupled with birth rate declines in most developing countries, are projected to increase the population ages 65 and over to the point in 2050 when it will be 2.5 times that of the population ages 0–4" (Haub, 2011).

The increase in life expectancy over the last century was concomitant with an increase in education attainment. In 1910, 13.5% of adults of 25 years of age and older in the United States attained a high school diploma as compared to 84% in 2000 and during that same time period the percentage of persons achieving four or more years of college rose from 2.7 to 25.6% (National Center for Education Statistics, 2001). This trend is indicative of attainment rates throughout the world which evidence improvement across both advanced and developing countries and are reflected in indicators such as a decrease in a worldwide 'nonschooling' rate of 47.1% in 1950 to 14.9% in 2010 and increase in secondary schooling for children of 15 years of age and older from 5.2 to 25.9% over the same years (Barro and Lee, 2012). The 'educational revolution' of the nineteenth and twentieth centuries centered on formal, consistent schooling of children and is credited with substantial public benefits such as modernization and economic growth (Baker et al., 2011). Currently, we are poised on the brink of a subsequent educational revolution centering on adult education that may also spur economic growth and lead to healthy, active aging throughout adulthood thereby producing positive public health benefits. The demand for adult educators has risen (Wyatt and Hecker, 2006) as the current educational revolution is spurred on by the increasing number of adults and their continued desire to remain engaged in formal and informal educational and learning experiences for a variety of reasons.

Adults' participation in educational opportunities may be the result of professional requirements (e.g., need to master new content or technology, occupational change). Many middle-aged and older adults are working longer and in 'encore careers' (Metlife, 2008). Within the context of an expanded life span, it is estimated that adults will spend an average of 41 years working in a professional capacity and 12 years in 'retirement' assuming an initial starting work age of 25 years (Gallup, 2012). However, it is anticipated that by 2020 the number of workers of 55 years of age and older will increase by 38% (Bureau of Labor Statistics (2013)) and 75% of workers of 50 years of age and older expect to have a retirement job in the future (Brown et al., 2010). With prolonged employment in adulthood comes a need for continuing education due to shrinking opportunities for high school graduates and increasing educational requirements (e.g., degrees, certifications) set forth by employers across occupational settings (Society of Human Resource Management, 2012).

In addition to motivations centered on work, 'nontraditional' adults pursue formal and informal education for diverse personal circumstances including pursuit of personal and leisure interests as well as a desire to learn and connect with others (Lakin et al., 2008). This emerging audience is creating new opportunities as well as challenges for educators and organizations to meet diverse for adult educational content and learning needs. The first steps already underway have been recognition of lifelong educational opportunities and expanding our understanding of how adults learn.

Lifelong Learning and Andragogy: Recognizing the Context of Adult Learners

Operating at a broad conceptual level, 'lifelong learning' refers to active learning throughout the life course and places

emphasis on the learner (Hager, 2011). The conceptual basis of lifelong learning has been and continues to be heavily influenced by 'three progressive sentiments: the individual, the democratic, and the adaptive' (Bagnall, 2001). As a result, individual development and growth throughout the life span, social justice, and cultural change are at the forefront of lifelong learning theory, practice, and advocacy in order to achieve individual- and societal-level benefits of education (Bagnall).

In contrast to lifelong learning, 'adult' or 'lifelong' education is more narrowly focused on specific elements of education, bringing to mind 'curriculum, teachers, teaching modes, and types of educational providers' (Hager, 2011). For almost a century, adult learners have been recognized as qualitatively different from younger learners. Early work on the science of adult learning focused on 'whether or not adults *could* learn' (Merriam, 2001), however, the field expanded to address how adults learn best and ways to support adults' lifelong learning. Knowles bid 'farewell to pedagogy' in 1970 promoting a distinct field of inquiry into adult learning and education called 'andragogy.' Andragogy reflects the science of adult education that aims to distinguish the needs and best practices related to adult education. The field of andragogy continues to advance today and as described by Merriam (2001) is based on several key premises centered on (1) adults' need for self-directed learning that reflects internal motivations, (2) focused learning linked to an applied need that supports the learner's life context and social roles, and (3) accumulated experiences that serve as a resource for the adult learner.

Although andragogy-related research has progressed, practice advances in the field have not been as rapid. As noted by Imel (2001), "Statistically, in the past 2 decades, adult students have changed the face of postsecondary education. What is not clear is the extent to which their presence has changed instructional practices. As a group, adult students have special concerns that should be addressed in the postsecondary classroom. However, these concerns must be addressed in light of the individual needs of each student." At times, 'nontraditional' and 'adult' students have been described separately in the literature; however, there are overlapping characteristics and motivations of these students. Nontraditional students are often identified by the presence of at least one of the following demographic characteristics: full-time employee, part-time student status, 1 year or greater post high school graduation, lacking high school diploma, financial independence from parents, and being parents themselves (Compton et al., 2006). Adult students tend to have select, work-related goals and view their primary identity as an employee rather than a student and as a result adult students have a greater likelihood of participating in a distance education program and programs that have specific vocational outcomes (Compton et al.). Whatever the label, these learners represent a diverse group with varied needs. Adult learners tend to seek formal education following life transitions and are often prompted by workplace and career pressures (Compton et al.). All adult students are faced with multiple roles as a result of seeking adult education (Kasworm, 2008), however, men and women are likely to have differing reasons for postsecondary education (Deutsch and Schmertz, 2011). Women in particular are likely to face greater external factors and multiple life roles that create added hardships (e.g., financial strain, continued child care and household responsibilities; Deutsch).

Lifelong Educational and Learning Benefits and Opportunities

Educational experiences throughout the life span support whole-person health and well-being. Early formal education in particular seems to set the course of key later life outcomes. These outcomes include indicators of health, functional ability, and survival and are hypothesized to reflect several potential mechanisms of influence such as increased cognitive functioning and human capital and ultimately access to resources and improved health decisions throughout the life course (Baker et al., 2011; Latham, 2012).

Educational benefits continue throughout adulthood and recognition of the benefits and need for continuous education throughout the life span have resulted in the emergence of concepts including lifelong education and lifelong learning. It is increasingly acknowledged that adult educational experiences can be of a formal or less formal nature. Lifelong education and learning may occur in a traditional institution of higher education, within distance or online environments, in the workplace, and via informal classes and experiences (e.g., travel). Although sometimes used synonymously, the terms 'lifelong education' and 'lifelong learning' do not necessarily represent conceptual similarity. 'Lifelong education' typically denotes more formal, structured educational experiences, which rise to some criteria, whereas 'lifelong learning' tends to refer to more broad-based learning that occurs across a variety of contexts (Hager, 2011). "Over the last half-century, lifelong learning has transformed from an avocation to an imperative. Lifelong learning has become intertwined with the labor market, making it difficult to find a better job or keep up with technology-driven industry standards without returning to school, potentially several times over the course of a lifetime" (p. 1; Council for Adult and Experiential Learning, 2012).

Once classified as 'nontraditional,' adult learners of 25 years of age and older "have comprised close to 40 percent of the college-going population [for the last two decades], spanning a range of backgrounds and experiences, from Iraq and Afghanistan war veterans and GED credential holders to 55-year-old professionals and skilled workers in career transition" (American Council on Education, 2013). This depiction highlights the diversity of 'adults' seeking educational opportunities. Within formal institutions of higher education, there is a growing recognition of adult students and examples exist of longstanding programs. For instance, begun in 1971, the University Without Walls at the University of Massachusetts Amherst supports 'nontraditional' undergraduate students by providing extensive advising and support services, assessment of prior experiences for credit, as well as enhanced program options and flexibility (University of Massachusetts Amherst (2013)). In a similar fashion, the Resumed Education for Adult Learners (REAL) program offered by Tufts University was originally created in 1970 to support young mothers returning to or beginning college by providing academic and financial resources (Tufts University, 2013). Rather than segregate adult learners, the REAL program purposefully integrates adult students in mainstream undergraduate programs and the program was opened to include men in 1976 (Tufts University).

Less formal options available for older adult learners are noncredit courses offered by programs such as 'Colleges for Seniors' and Osher Lifelong Learning Institutes (OLLI; The Bernard Osher Foundation, 2005). In the case of OLLI there are 116 programs across the United States offering noncredit courses encompassing academic- and leisure-related topics of interest to 'seasoned adults' of 50 years of age and older (The Bernard Osher Foundation). "Emphasis is placed on learning for the joy of learning and on keeping in touch with a larger world" (The Bernard Osher Foundation).

Benefits and opportunities related to lifelong learning and education are not limited to formal classroom settings, and many adult educational experiences are available outside of a traditional classroom. One classic example is the Road Scholar program, also known as 'Elderhostel' (Road Scholar, 2013). Beginning in 1975 and catering to older adults, this program offers 5500 educational and cultural tours across the United States and 150 countries that support whole-person wellness (e.g., cognitive, social, physical). Many educational opportunities for older adults also incorporate volunteerism and mentoring that benefit both the adult participant and members of his/her community. In the US, the Retired Senior Volunteer Program (Corporation for National and Community Service, 2013) and AARP Experience Corps (affiliate of the national AmeriCorps program; AARP, 2013) provide informal opportunities for older adults to engage in educational and enriching projects in their communities often within intergenerational settings. Interestingly, the European Commission recently combined separate age-specific learning programs, including an adult education-focused program, to create one overarching and integrated lifelong program focused on 'UK education, training, youth, and sport opportunities across Europe' (i.e., Erasmus + program; Ecorys and the British Council, 2013) with the intention of strengthening learner 'mobility,' establishing best practices, and securing policy change.

It is increasingly recognized that lifelong cognitive activity and engagement may help curb normative (typical) cognitive changes as well as the growing number of individuals experiencing nonnormative cognitive impairment and dementia that is fast becoming a public epidemic/concern. Brain 'exercise' can help maintain cognitive vitality and health, and formal and informal opportunities are increasing. Formal interventions typically target one type of cognitive skill (e.g., memory), imparting strategies for improvement (e.g., chunking of information) and occur in settings outside the home with other adults. Informal interventions are often technology-based and involve more individual work framed as 'games.' Scientific query into the efficacy of both types of training mechanisms and their impact on everyday functioning is still nascent.

Future Directions

Together, the increasing number of older adults, decreasing population of potential high school graduates to satisfy the job market, and the changing nature of work are contributing to a second educational revolution. Students who are 'nontraditional' in some aspect are now the majority rather than the exception within institutions of higher education (National Center for Education Statistics, 2002). Lifelong learning is becoming a cultural norm (Council for Adult and Experiential Learning, 2012). These realities create both opportunities and challenges for the field of adult education.

Early educational experiences as well as lifelong learning possess the potential to positively impact adult development and aging at the individual level, thereby promoting active aging in later life. These individual-level benefits also translate to group and societal levels as education can contribute to decreased health disparities. To reap these benefits, however, institutions and educators must recognize the potential of adult learners as well as identify methods to break down barriers to adults' educational and learning experiences. Three areas are particularly key: (1) understanding and embracing the diversity within the group of adult learners, (2) promoting whole-person wellness that is embedded in the ideal of continuous life span learning and within communities of learners, and (3) the thoughtful use of technology.

First, adult educators must recognize the diversity within the group of individuals representing adult learners. In general, adults (and particularly older adults) are a heterogeneous group who arrived at their stage in life after a multitude of experiences and environmental influences. Adult learners seek educational opportunities with varying motivations, expectations, skills and backgrounds, content needs, and contexts outside of the learning environment. "Because adults have competing lives, hopes, and realities, each semester of college involvement represents either a renegotiation or adaptation of themselves and their lives" (p. 29; Kasworm, 2008). As outlined by Fairchild (2003), adult learners balance an assortment of roles and associated responsibilities (e.g., family and caregiving responsibilities, career) in addition to being a student or learner. This reality can cause conflict and overload and in addition, adult students may receive less support from significant others in their educational pursuits compared to traditional students. At the most basic level, educational institutions and policy makers must address financial barriers (e.g., tuition, activity fees applied to all students, government aid policies) that constrain adult education; this may be particularly true for women (Deutsch and Schmertz, 2011) and older adults (Council for Adult and Experiential Learning, 2012). One example of a pilot initiative to combat the latter problem is the 'Lifelong Learning Accounts' that represent 'portable, employee-owned, and employer-matched accounts used to finance education and training' (Council for Adult and Experiential Learning). To address other practical barriers, enhanced and flexible student services should be implemented (e.g., web-based services, opportune office hours) and roles should be acknowledged as strengths and incorporated into the student experience to the student's advantage (e.g., draw upon relevant experiences and knowledge within courses; identify adult students for leadership roles; Compton et al.; Fairchild). Special programs can target psychosocial barriers frequently experienced by adult and 'mature' learners (e.g., 'anxieties about starting over,' worries about working with new technology, or 'competing with younger jobseekers'; Council for Adult and Experiential Learning). In general, we need new strength-based (Compton et al.) models acknowledging the contribution of adult learners (Kasworm, 2008), which create a 'sustaining academic community' (Deutsch and Schmertz, 2011) that

actively engages adult learners both inside and outside the classroom.

Second, enhanced integration and empowerment is needed in several respects. Paramount is the concept of lifelong learning throughout the life span rather than segmented learning within discrete age periods thereby leading to greater continuity of educational experiences. Additionally, moving from a single learner approach is valuable and collaboration among learners and the support of a broader supportive learning community can lead to enhanced outcomes not achievable by individuals. Furthermore, integration of formal and informal learning opportunities can expand the nature and method of educational and lifelong learning opportunities. The final aspect of integration is adoption of a whole-person approach that can assist adult educators in understanding the needs of adult learners and developing individualized plans to achieve goals across domains (e.g., social, intellectual; Compton et al.) and across educational and learning environments. For instance, 'prior learning assessments' can pinpoint already mastered material (for credit in formal educational settings; Council for Adult and Experiential Learning, 2012) and individual learning plans can target areas of development, both of which save time, frustration, and money. Together the philosophies underlying lifelong learning (i.e., individual, democratic, adaptive; Bagnall, 2001) can foster innovative and progressive opportunities and programs for lifelong learning and engagement throughout adulthood. An excellent example of this out-of-the box thinking is the 'Intergenerational School' for elementary-aged children, which incorporates learning opportunities for adults, including persons with dementia, with plans to integrate these educational experiences with the 'InterWell' health promotion program (Whitehouse, 2013).

Third, thoughtful use of technology to enhance and deliver adult education is critical. Currently technology may be a dual-edged sword when it comes to adults and use in an educational setting. Adult learners may be more likely to participate in distance education programs, however, these experiences can be isolating and distance programs can exhibit greater dropout rates (Bolliger and Halupa, 2012; Kruger-Ross and Waters, 2013). In addition, a 'digital divide' may be evident on individual and community/societal levels. On a personal level, new technologies can be intimidating for some adult learners (particularly older adults) because of a variety of factors including lack of exposure and familiarity, but perhaps more importantly due to psychological factors including computer-related and aging anxiety and self-efficacy (Jung et al., 2010). On a broader level, groups of individuals may be excluded as a result of resource (e.g., lack of computer access, low income) and practical constraints (e.g., availability of broadband in rural areas, functional disability) thereby imposing limitations on educational technology use (e.g., Choi and DiNitto, 2013). On the positive side, web-based support services (Compton et al., 2006) and instructional technologies (e.g., synchronous and asynchronous print, video, and audio tools; social media) can facilitate diverse adult learning styles, schedules, (Finch, 2011) and functional (e.g., mobility impairment) and sensory needs (e.g., vision or hearing challenges). In addition, unique experiences supporting adults' education and whole-person wellness may be possible through the use of technology such as 'virtual senior centers' (Selfhelp Community Services, Inc., 2013). These types of programs serve to greatly enhance access, as individuals do not need leave their homes. 'Gerontechnology' (technological applications geared toward aging audiences) is a growing field and technology is anticipated to play a larger role in not only adult education and lifelong learning, but also personal health and wellness promotion and intervention, as successive cohorts of adults become more familiar and facile with a range of technological tools.

As outlined in this section, capitalizing on adult learners' expertise and motivation to continue their learning across a variety of contexts will require an interdisciplinary and systemic approach to reduce barriers and increase access to lifelong educational opportunities. Professionals from fields such as education, psychology, gerontology, and computer/technology science are needed to understand and facilitate adult learners' needs. A systems approach is needed to help move theory and research to the realm of best practices and supporting policy. This approach requires input and collaboration from adult learners (e.g., openness to experience, advocacy, persistence), family and support systems (e.g., encouragement, role negotiation), employers (e.g., flexibility, purposeful professional development opportunities), institutions and organizations offering educational opportunities (e.g., enhanced support services, needed content offerings), as well as state and federal governments (e.g., policies, financial assistance) in order to support and promote continued learning throughout adulthood.

See also: Cognitive Development: Mathematics Learning and Instruction; Cognitive Styles and Learning Styles; Distance Education; Education and Learning: Lifespan Perspectives; Health Education and Health Promotion; Higher Education Market; Instructional Design; Learning Theories and Educational Paradigms; Learning and Instruction: Social-Cognitive Perspectives; Lifelong Learning and Its Support with New Technologies; Metacognitive Development: Educational Implications; Self-Efficacy: Education Aspects; Social Media; Tangible User Interfaces in Learning and Education.

Bibliography

AARP, 2013. Experience Corps. Retrieved from: http://www.aarp.org/experience-corps/.

American Council on Education, 2013. Adult Learners. Retrieved from: http://www.acenet.edu/higher-education/topics/Pages/Adult-Learners.aspx.

Bagnall, R., 2001. Locating lifelong learning and education in contemporary currents of thought and culture. In: Aspin, D., Chapman, J., Hatton, M., Sawano, Y. (Eds.), International Handbook of Lifelong Learning. Kluwer Academic Publishers, Great Britain, pp. 35–52.

Baker, D., Leon, J., Greenaway, E., Collins, J., Movit, M., 2011. The education effect on population health: a reassessment. Population and Development Review 37 (2), 307–332.

Barro, R.J., Lee, J.W., 2012. A new data set of educational attainment in the world, 1950–2010. Journal of Development Economics 104, 184–198.

Bolliger, D.U., Halupa, C., 2012. Student perceptions of satisfaction and anxiety in an online doctoral program. Distance Education 33 (1), 81–98.

Brown, M., Aumann, K., Pitt-Catsouphes, M., Galinsky, E., Bond, J., 2010. Working in Retirement: A 21st Century Phenomenon. Families and Work Institute, New York.

Bureau of Labor Statistics, 2013. Labor Force. Retrieved from: ftp://ftp.bls.gov/pub/special.requests/ep/labor.force/labor.xls/.

Council for Adult and Experiential Learning, 2012. Not Too Late for School. Retrieved from: http://www.cael.org/pdfs/Mature-learners—AP.

Choi, N.G., DiNitto, D.M., 2013. The digital divide among low-income homebound older adults: internet use patterns, eHealth literacy, and attitudes toward computer/internet use. Journal of Medical Internet Research 15 (5), e93. http://dx.doi.org/10.2196/jmir.2645.

Corporation for National and Community Service, 2013. Retrieved from: http://www.nationalservice.gov/programs/senior-corps/rsvp.

Compton, J.I., Cox, E., Laanan, F.S., 2006. Adult learners in transition. New Directions for Student Services 114, 73–80.

Deutsch, N.L., Schmertz, B., 2011. "Starting from ground zero:" constraints and experiences of adult women returning to college. The Review of Higher Education 34 (3), 477–504. http://dx.doi.org/10.1353/rhe.2011.0002.

Ecorys and the British Council, 2013. Erasmus + Factsheet – December 2013. Retrieved from: http://www.erasmusplus.org.uk/sites/default/files/Erasmus%2B%20factsheet_0.pdf.

Fairchild, E.E., 2003. Multiple roles of adult learners. New Directions for Student Services 102, 11–16.

Finch, A., Spring 2011. Adult learning styles and technology-driven learning for online students. Academic Leadership Live: The Online Journal 9 (2).

Gallup, 2012. Expected Retirement Age in U.S. Up to 67. Retrieved from: http://www.gallup.com/poll/154178/expected-retirement-age.aspx.

Hager, P.J., 2011. Concepts and definitions of lifelong learning. In: London, M. (Ed.), The Oxford Handbook of Lifelong Learning. Oxford Press, Oxford.

Hager, P., Halliday, J., 2006. Recovering Informal Learning: Wisdom, Judgment, and Community. In: Lifelong Learning Book Series, vol. 7. Springer, Dordrecht, The Netherlands (Paperback edition, 2009).

Haub, C., 2011. World Population Aging: Clocks Illustrate Growth in Population under Age 5 and over Age 65. Retrieved from: http://www.prb.org/Publications/Articles/2011/agingpopulationclocks.aspx.

Imel, S., 2001. Adult Learners in Postsecondary Education. Retrieved from: http://www.calpro-online.org/eric/textonly/docgen.asp?tbl=pab&ID=107.

Jung, Y., Peng, W., Moran, M., Jin, S.-A.A., McLaughlin, M., Cody, M., Jordan-Marsh, M., Albright, J., Silverstein, M., 2010. Low-income minority seniors' enrollment in a cybercafé: psychological barriers to crossing the digital divide. Educational Gerontology 36, 193–212. http://dx.doi.org/10.1080/03601270903183313.

Kasworm, C.E., 2008. Emotional challenges of adult learners in higher education. New Directions for Adult and Continuing Education 120, 27–34.

Knowles, M.S., 1970. The Modern Practice of Adult Education: Andragogy versus Pedagogy. Association Press, New York.

Kruger-Ross, M.J., Waters, R.D., 2013. Predicting online learning success: applying the situational theory of publics to the virtual classroom. Computers & Education 61, 176–184.

Lakin, M.B., Mullane, L., Robinson, S.P., 2008. Mapping New Directions: Higher Education for Older Adults. American Council on Education, Washington, DC.

Latham, K., 2012. Progressive and accelerated disability onset by race/ethnicity and education among late midlife and older adults. Journal of Aging and Health 24 (8), 1320–1345.

Lawson, K., 1982. Lifelong education: concept or policy? International Journal of Lifelong Education 1 (2), 97–108.

Merriam, S., 2001. Andragogy and self-directed learning: pillars of adult learning theory. New Directions for Adult and Continuing Education 89 (1), 3–14. http://dx.doi.org/10.1002/ace.3.

Metlife, 2008. A MetLife Foundation/Civic Venture: Encore Career Survey. Retrieved from: https://www.metlife.com/assets/cao/foundation/Encore_Survey.pdf.

National Center for Education Statistics, 2001. Table 8.—Years of School Completed by Persons Age 25 and over and 25 to 29, by Race/Ethnicity and Sex: 1910 to 2000. Retrieved from: http://nces.ed.gov/programs/digest/d01/tables/PDF/table008.pdf.

OECD, 2011. "Life expectancy at birth", in Health at a Glance 2011: OECD Indicators, OECD Publishing. Retrieved from: http://dx.doi.org/10.1787/health_glance-2011-4-en.

National Center for Education Statistics, 2002. Nontraditional Undergraduates. Retrieved from: http://nces.ed.gov/pubs2002/2002012.pdf.

Organisation for Economic Co-operation and Development (OECD), 2007. Trends in Severe Disability Among Elderly People: Assessing the Evidence in 12 OECD Countries and the Future Implications. Retrieved from: http://www.oecd.org/denmark/38343783.pdf.

Road Scholar, 2013. Adventures in Lifelong Learning. Retrieved from: http://www.roadscholar.org/.

Selfhelp Community Services, Inc., 2013. Virtual Senior Center. Retrieved from: http://selfhelp.net/virtual-senior-center.

Society of Human Resource Management, 2012. What's Next: Future Global Trends Affecting Your Organization Evolution of Work and the Worker. Retrieved from: http://www.shrm.org/about/foundation/shapingthefuture/documents/2-14%20theme%201%20paper-final%20for%20web.pdf.

The Bernard Osher Foundation, 2005. Osher Lifelong Learning Institutes. Retrieved from: http://www.osherfoundation.org/index.php?olli.

Tufts University, 2013. The R.E.A.L. Program: Resumed Education for Adult Learning. Retrieved from: http://uss.tufts.edu/undergradEducation/academics/real/.

University of Massachusetts Amherst, 2013. University without Walls at the University of Massachusetts Amherst. Retrieved from: https://www.umass.edu/uww/.

U.S. Census Bureau, 2012. Expectation of Life at Birth, and Projections: 2010. Retrieved from: http://www.census.gov/compendia/statab/2012/tables/12s0104.pdf.

Whitehouse, P., 2013. The challenges of cognitive aging: integrating approaches from neuroscience to intergenerational relationships. Journal of Intergenerational Relationships 11 (2), 105–117. http://dx.doi.org/10.1080/15350770.2013.782740.

Wyatt, I.D., Hecker, D.E., 2006. Occupational changes during the 20th century. Monthly Labor Review, 35–57.

Adult Mortality in Developing Countries

Katherine T Lofgren and Haidong Wang, Institute for Health Metrics, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Adult mortality measurement is becoming increasingly important as populations globally age and health-policy targets premature deaths in adult ages. Adult deaths are relatively rare events and can present a difficult challenge to estimate. This challenge is especially true in the developing world where vital registration systems to capture deaths in a country are often not well established or well functioning. This article aims to capture the available data sources for adult mortality and the methods to analyze those data sources as well as discussing the current levels and trends around the world.

Introduction

Developing countries around the globe have experienced fast demographic transition in the past six decades. We have witnessed decline in fertility and improvements in mortality. As a result of this impressive change, life expectancy at birth for both sexes combined in the developing world has increased from 42.28 years in 1950–55 to about 65.95 years in 2005–10 ("World Population Prospects: The, 2010 Revision," 2011). Much of this improvement is attributed to the fast decline in child mortality, which is defined as the probability of death from birth to 5 years of age. Billions of dollars have been invested in reducing child mortality among developing countries and great strides have been made since the United Nations Millennium Declaration was made by 189 nations in 2000. Millennium Development Goal 4 is set to reduce child mortality by two-thirds by 2015 from its level in 1990. Although it is true that the majority of the nations would not be able to achieve this goal by 2015 (Lozano et al., 2011; UN Inter-agency Group for Child Mortality Estimation, 2012), child mortality has declined significantly in the past two decades among developing countries thanks to the paramount efforts made by the governments, international organizations, and private donors. Unfortunately, less attention has been paid to monitoring and reducing preventable premature mortality in adult age groups.

In measuring adult mortality, two metrics are widely used to date: probability of death from age 15 to 50 and probability of death from age 15 to 60. The latter has gained more popularity since the publication of Global Burden of Diseases (GBD) Study by the World Health Organization, and later in the organization's annual World Health Report series. This measure succinctly conveys the risk of premature death in adult populations. An alternative is calculating the age-specific and sex-specific mortality rates. Stratifying mortality indictors by age and sex can help highlight inequalities and specific populations where the mortality burden is the highest. In this article, we employ probability of death from age 15 to 60 to measure the adult mortality rate.

Data Sources and Estimates

Three major sources provide estimates of adult mortality for all major populations in the world: United Nations Population Division through its biannual World Population Prospects report, World Health Organization through its World Health Report series, and more recently the Institute for Health Metrics, and Evaluation at University of Washington through its GBD, Injuries and Risk Factors Study.

To estimate adult mortality rate, ideally, raw data would come from civil registration systems that enumerate deaths in a country and record cause of death and other important epidemiologic information. Unfortunately, civil registration systems are expensive and require a large amount of infrastructure to be effective at capturing deaths in a population. Although some of these systems are established in developing countries, for a large portion of the developing world we have to rely on surveys and censuses to approximate mortality levels. Large survey series such as the Demographic and Health Surveys (*Macro International, Inc.*) often include questions aimed at approximating adult mortality in place of civil registration systems. This section aims to give detailed information about what data sources are available for developing countries and some of the considerations that should be taken with each as a source of mortality information.

Vital Registration Systems

Vital registration (VR) systems are a fundamental source of information regarding adult mortality in a population. Mortality rates can be calculated directly from the deaths and population information captured in the civil registration system. Although VR systems have the potential to provide the most reliable and continuous sources of adult mortality data, they often fall short of capturing all the deaths in a population. This undercounting of deaths is especially true of VR systems in the developing world where infrastructure and maintenance of the reporting systems is particularly difficult. VR-derived data are a critical source of reliable information. It is imperative that resources be channeled to better information systems in developing countries to create the most accurate estimates of mortality. Through better evidence and understanding regarding the age/sex patterns and causes of death, better health-policy decisions can be made.

Table 1 shows the developing countries with VR systems (as defined by active reporting within a 10-year window) and the estimated completeness of that system in 2010 as determined by data available in the GBD 2010 Study.

Table 1 Vital registration completeness for developing countries, 2010

GBD 2010 Region	Country	Completeness (%)	
Asia, Central	Armenia	92.8	(77.9–100.0)
	Azerbaijan	87.4	(72.8–100.0)
	Georgia	93.0	(78.4–100.0)
	Kazakhstan	93.1	(78.6–100.0)
	Kyrgyzstan	95.4	(81.4–100.0)
	Mongolia	84.4	(70.1–100.0)
	Tajikistan	79.6	(66.0–94.7)
	Turkmenistan	91.0	(76.1–100.0)
	Uzbekistan	84.2	(69.8–100.0)
Asia, East	Taiwan	95.8	(76.7–100.0)
Asia, Southeast	Malaysia	98.1	(87.9–100.0)
	Maldives	97.5	(86.6–100.0)
	Mauritius	97.4	(86.5–100.0)
	Myanmar	91.4	(78.4–100.0)
	Philippines	97.1	(85.9–100.0)
	Seychelles	99.1	(91.5–100.0)
	Sri Lanka	98.5	(89.2–100.0)
	Thailand	87.9	(75.0–100.0)
Caribbean	Antigua and Barbuda	93.9	(79.5–100.0)
	Bahamas	97.8	(86.6–100.0)
	Barbados	99.4	(92.6–100.0)
	Belize	98.0	(86.9–100.0)
	Cuba	97.6	(85.9–100.0)
	Dominica	98.3	(87.3–100.0)
	Dominican Republic	76.2	(63.8–90.4)
	Grenada	100.0	(100.0–100.0)
	Guyana	91.5	(77.1–100.0)
	Jamaica	98.1	(87.0–100.0)
	Saint Lucia	100.0	(100.0–100.0)
	Saint Vincent and the Grenadines	99.8	(97.6–100.0)
	Suriname	95.5	(82.1–100.0)
	Trinidad and Tobago	99.6	(94.9–100.0)
Latin America, Andean	Ecudar	93.2	(77.8-100.0)
	Peru	75.9	(62.0-91.8)
Latin America, Central	Colombia	92.3	(78.1–100.0)
	Costa Rica	99.3	(92.1–100.0)
	El Salvador	96.8	(84.1–100.0)
	Guatemala	99.2	(91.5–100.0)
	Mexico	98.8	(89.4–100.0)
	Nicaragua	94.2	(80.1–100.0)
	Panama	94.7	(81.1–100.0)
	Venezuela	98.9	(90.0–100.0)
Latin America, Southern	Argentina	98.7	(92.9–100.0)
	Chile	99.6	(95.4–100.0)
	Uruguay	99.4	(94.9–100.0)
Latin America, Tropical	Brazil	99.6	(95.3–100.0)
	Paraguay	81.6	(73.4–90.6)
North Africa/ Middle East	Algeria	94.7	(72.9–100.0)
	Bahrain	78.7	(55.7–100.0)
	Egypt	97.8	(80.5–100.0)
	Iran, Islamic Republic of	81.8	(58.4–100.0)
	Iraq	97.5	(78.9–100.0)
	Jordan	97.9	(81.0–100.0)
	Kuwait	99.9	(100.0–100.0)
	Libyan Arab Jamahiriya	80.6	(57.5–100.0)
	Morocco	46.7	(32.8–64.4)
	Occupied Palestinian Territory	61.3	(42.8–85.2)
	Oman	93.8	(71.3–100.0)
	Qatar	82.3	(58.4–100.0)
	Saudi Arabia	65.7	(46.2–91.0)
	Syrian Arab Republic	99.8	(100.0–100.0)
	Turkey	74.7	(52.8–100.0)
Oceania	Fiji	99.4	(90.5–100.0)
	Marshall Islands	80.1	(53.7–100.0)
Sub-Saharan Africa, Southern	South Africa	99.6	(93.5–100.0)

Completeness is defined as the percent of the deaths in a nation covered in the VR system.

Within the countries that are listed in **Table 1**, there are 35 countries with VR completeness above 95%, meaning that the registration system is estimated to capture 95% of the deaths of people aged 15 and above in the country. For systems that fail to capture all of the deaths in a system, the data must be adjusted to account for the bias using demographic methods. In 15 of the countries in **Table 1**, there is an estimated coverage of adult deaths below 85%. The method used to assess completeness of VR systems in adult populations is referred to as the death distribution method (DDM). Detailed information regarding the application of several specific types of DDMs can be found in Manual X which was published by the United Nations Population Division (United Nations. Department of International Economic and Social Affairs. Population Division and National Research Council (U.S.). Committee on Population and Demography, 1983). Three variants that are often used include the generalized growth balance method (GGB), the synthetic extinct generation method (SEG), and a combined method that employs both GGB and SEG. GGB and SEG were developed in the 1980s by Hill and Bennett & Horiuchi, respectively (Bennett and Horiuchi, 1981; Hill K, 1987). Hill and Choi build upon these methods to propose the combined variant in the mid-2000s (Hill K et al., 2009). Other efforts have been to identify the optimal age trims that should be used in estimating the completeness of a VR system (Murray C.J.L. et al., 2010). **Figure 1** below shows the distribution of VR systems in developing countries, delineated by whether the system captures at least 95% of the deaths in the country.

Of note in the figure is the clear lack of VR systems in the majority of Africa and much of Asia. Although both complete and incomplete VR data can be a useful source of information,

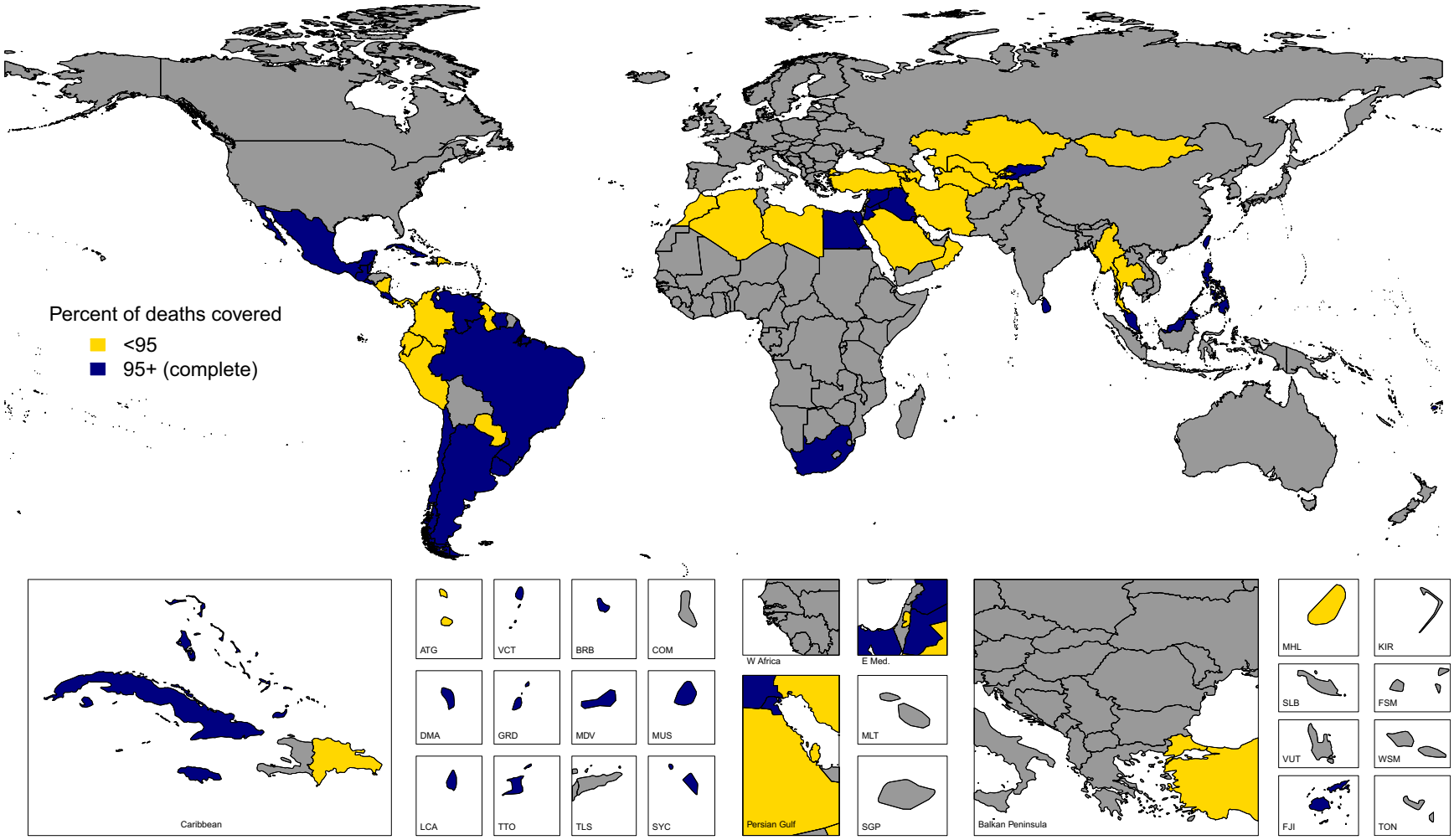

Figure 1 Vital registration systems in developing countries.

it is not available in much of the developing world. Only 65 of the 137 developing countries have operating VR systems in 2010, or 47% of the countries. Further, there has not been progress in establishing new systems; in 1980 there were 67 out of 137 developing countries reporting deaths through VR systems, 2 more reporting countries than we observe in 2010. Lack of functioning VR systems among developing countries presents a problem for researchers worldwide who are interested in the true level of mortality in these countries. Without information from VR systems, supplemental information from censuses and surveys that provide information on health and mortality has been utilized.

Censuses

Censuses are also an important source of mortality information, especially when small area mortality estimates are desired. Censuses are unique in their comprehensive nature and large sample size. Although censuses may include (theoretically) the entire population, they also are not targeted to estimate adult mortality as their primary purpose. Often the only question included for the estimation of adult mortality is the number of deaths in the household over a specific recall period (typically 12–18 months). With data on the entire population captured in the census and the number of deaths in a household, adult mortality can be estimated. It is important to note that asking respondents questions about the deceased age at death often includes recall error and age heaping. Some of the recall error and age heaping can be avoided by survey design, often event calendars, and culturally appropriate survey questionnaires can reduce both of these biases (Aday, 1989; Belli et al., 2001; Mukherjee and Mukhopadhyay, 1988).

Interestingly, census household deaths recall data can be corrected for underreporting in the same manner the VR data can be assessed and adjusted, using DDMs. A full compilation of completeness estimates and adult mortality estimates for census-derived death information is available online for the GBD Study at http://vizhub.healthdata.org/mortality/. Although censuses have large sample sizes, they often occur infrequently – especially in resource-poor settings. In addition to censuses, national and subnational surveys provide critical information on the most current trends in adult mortality.

Surveys and Survey Series

Surveys and survey series are the backbone of adult mortality estimation in developing countries. Surveys offer an opportunity to create targeted survey modules specifically designed to gain accurate information about mortality trends and levels in a community. Although they can provide an economic means to gain information compared to a VR system, they are often flawed. Sampling schemes must be designed correctly to capture a representative sample and often missingness and recall bias make the data unusable. Luckily, there are many well-conducted surveys to pull from and in some cases, a series of surveys conducted with similar modules and sampling frames. Survey series like the Demographic and Health Surveys conducted by Measure DHS (ICF International) and the Reproductive and Health Surveys conducted by the Center for Disease Control and Preventions often provide a consistent data source across countries and over time.

Survey data are typically released first as a report, which may include a few summary estimates of adult mortality. Ultimately, access to the microdata will provide the most information about mortality trends in the sampled population. Surveys can ask about adult mortality through several strategies. Similar to censuses, survey can also ask about deaths in the household in the last 12–18 months. Unlike most full censuses, surveys can also include more specific modules aimed specifically at mortality estimation.

Specific modules aimed at capturing adult mortality include sibling and parental survival histories. In a sibling survival history, a respondent is asked about the survivorship status of his/her siblings. Similarly, in a parental history a respondent is asked about the survivorship status of his/her parents. This information, in addition to the respondent's age, can be used to estimate mortality risks.

Sibling histories build off the well-established survey methods to estimate under 5 years' mortality by asking mothers about the total number of children they have given birth to and the number of children currently surviving (Feeney G, 1991; Myburgh, 1956; Rajaratnam JK et al., 2010; Sullivan, 1972). This mortality estimation technique relies on a mother reporting the mortality experiences of her offspring and was first developed by Brass and has gone through refinement over the subsequent years (Brass and Coale, n.d.; Rajaratnam J.K. et al., 2010). In contrast, sibling history modules in surveys require the respondent to answer mortality questions regarding his/her siblings. This is a useful extension of the summary birth history technique to estimate adult mortality when more reliable data sources are lacking. The respondent's age is used as an approximation of the exposure risk experienced by all of the respondent's siblings, with respondent's answering questions about how many siblings they have and the number of those siblings that are currently alive (Hill and Trussell, 1977). The exact methods to estimate adult mortality using these survey modules can be found in demography reference books including 'Manual X' and 'Demography: Measuring and Modeling Population Processes' (Preston et al., 2001; United Nations. Department of International Economic and Social Affairs. Population Division and National Research Council (U.S.). Committee on Population and Demography, 1983).

Sibling histories although useful can prove to be biased. Often respondents are unable to accurately recall the mortality status of all of their siblings, particularly in high-fertility populations. Further, respondents may legitimately not know the survival status of siblings, especially siblings that are much older and may be out of contact with the respondent. Beyond the obvious issues with recall bias, there is also the potential for high-risk mortality families where all siblings have died to be missed, referred to as zero-survivorship bias. Further, the survey sampling scheme may bias the results as large families with many siblings are more likely to be sampled (and sampled multiple times) compared to smaller families. To account for this sampling issue, weights can be applied to the responses based on the respondent's total number of siblings (Gakidou and King, 2006; Masquelier, 2013). Further refinements and applications of these methods have been made over the years,

making sibling histories a data source that can prove meaningful and enlightening when VR and censuses are lacking (Obermeyer et al., 2010).

Parental histories offer an analogous option to sibling histories described above except that the survivorship status is asked from the survey respondent's mother (or father) rather than their siblings. This can be especially challenging as survey respondents may not recall the survivorship of their parents, introducing recall bias. Further, young respondents cannot reliably identify if their caregivers are their biological parents, limiting the age group that can report on their parent's survivorship to those typically above age 15. One of the central underlying assumptions is that a parent's mortality risk is unassociated with their child's risk. The use of parental survival information to estimate adult mortality was first proposed in 1960 by Louis Henry (Henry, 1960). The methods were formally developed by William Brass and Kenneth Hill in 1973 (Brass and Hill, 1973). Through the following decades there was significant improvements and expansion of the methods, including the addition of time references for mortality estimates developed by Brass and Bamgboye (Brass and Bamgboye, 1981). There have been several applications of the method and attempts to improve survey-instrument methodology (Blacker, 1977; Luy et al., 2011; Timaeus and Chackiel, 1996; Timaeus, 1991). There are also further extensions of this method that require two surveys 5–10 years apart; in this case the mortality estimates specifically reference the intersurvey period.

Parental and sibling histories present similar strategies to estimate adult mortality. Both warrant consideration in data-sparse settings and should be used in combination and comparison when both are available. Although the aforementioned survey-based adult mortality estimation methods can provide estimates of mortality, many researchers have avoided using such methods because of the concerns about different biases inherent in these methods. In countries where only child mortality estimates are available, it is common to estimate adult mortality using a model life table system where both age pattern and level of adult mortality are dictated by the entry parameter, normally probability of death from birth to 5 years of age, and the selected standard life table (*MODMATCH*, n.d.; United Nations. Department of International Economic and Social Affairs. Population Division and National Research Council (U.S.). Committee on Population and Demography, 1983). The INDEPTH Network, a global network of health and demographic surveillance systems, has made significant efforts in collecting population health data in 49 surveillance sites throughout Africa, Asia, and Oceania. Although faced with small numbers (both vital events and exposures), data from these sites provide unique insights into levels, trends and age patterns of mortality for many low- and middle-income countries. Efforts are being made in generating model life table system based on data from the INDEPTH network and especially for populations affected with HIV/AIDS.

Levels and Trends of Adult Mortality in the Developing World, 1950–2010

There are various sources of information on country-level adult mortality risk, the probability of death from 15 to 60 years of age. In this section, the numbers cited are from the GBD Study 2010. More information on the collaborators, the structure, and the aims of the GBD Study is available online (http://www.healthdata.org/gbd). The information here is derived from the all-cause mortality GBD 2010 research of Wang et al. published in *the Lancet* (Wang et al., 2012). Continuous efforts are being made to generate regularly updated versions of the GBD Study, beginning with an update through 2013. For the most up-to-date estimates of adult mortality risk, child mortality risk, and other relevant demographic metrics consult the databases and visualizations available here (http://www.healthdata.org/gbd/data-visualizations).

Other sources of information on adult mortality include the estimates generated by the United Nations Population Division. Estimates from the World Population Prospects 2012 Revision can be found online at (http://esa.un.org/wpp/unpp/panel_population.htm) or through the World Bank at http://data.worldbank.org/indicator/SP.DYN.AMRT.MA.

Heterogeneity in Adult Mortality among Developing Countries

Globally, adult mortality in the developing world has been improving at a fast pace since 1950. The GBD 2010 Study found that female adult mortality, the probability between ages 15 and 60, decreased from around 314 per 1000 in 1950 among all developing countries to about 140 per 1000 in 2010. Similarly, male adult mortality decreased from around 357 per 1000 to 214. In other words, adult mortality rates in the developing world have been declining at about 1.3% and 0.9% annually in women and men, respectively. Such summary measurements mask the great heterogeneity among countries in the developing world. In 1950, the highest estimated female adult mortality in the developing world can be found in Burundi, at 478 per 1000. This means nearly 50% of the 15-year-old adults would not live to their 60th birthday assuming they will experience the same set of age-specific mortality rates observed in 1950. At the same time, the lowest female mortality was observed in Qatar at 131 per 1000, an incredible 265% difference. For males, the highest adult mortality rate can be found in Equatorial Guinea, where it was 565 per 1000. In the same year, Qatar has the lowest male adult mortality rate at 184 per 1000. Over the past six decades, both absolute and the relative differences in adult mortality rate between countries with the highest and lowest values have increased, indicating increasing health disparities among these populations. In 2010, the highest adult female mortality rate is estimated to be in Haiti where an earthquake that claimed over 222 000 lives has driven the mortality rate up tremendously. The second highest female adult mortality was from Swaziland, where it was about 526 per 1000. In the same year, the lowest female adult mortality rate among developing countries was from Cyprus at about 42 per 1000. The relative difference between the highest and lowest female adult mortality rates has increased from about 265% in 1950 between Burundi and Qatar to about 1152%. Similar difference can be found for male adult mortality rate. In 2010, the relative difference between the highest adult mortality from Lesotho and the lowest adult mortality from Qatar was 751%, a much higher value compared to the relative difference of 207% between the aforementioned Equatorial Guinea and Qatar.

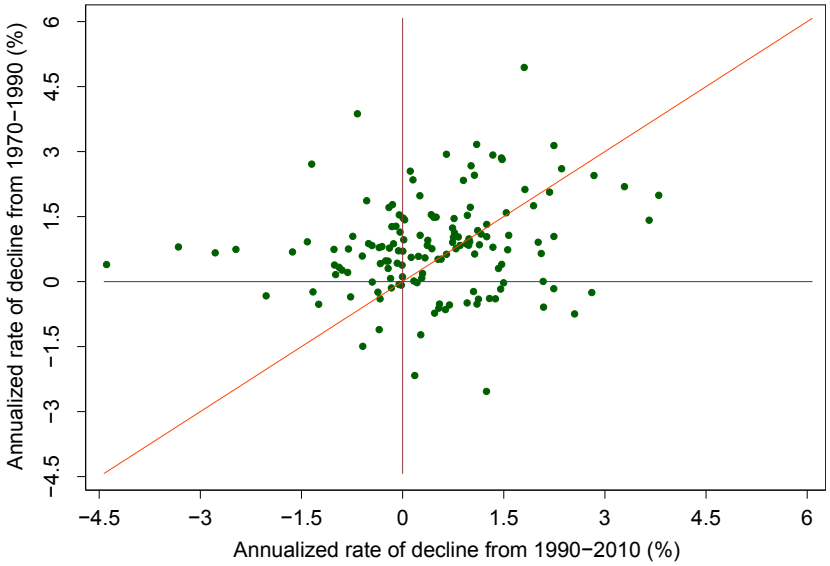

Figure 2 Annualized Rate of Decline in Male Adult mortality, 1970–2010.

There is even greater heterogeneity in the country-level changes in adult mortality over the past six decades based on the GBD 2010 Study. Among men in the developing world, Peru experienced the fastest rate of decline in adult mortality from 1970 to 2010 at an annualized rate of 3.37%. This impressive rate of change helped lower the male adult mortality rate in Peru from 436 per 1000 in 1970 to about 113 per 1000 in 2010. Similar stories of success can be found in South Korea, Singapore, Jordan, Macau, and Maldives where annualized rate of decline in male adult mortality rate exceeded 2.5% for the last four decades. During the same period, 33 developing countries have seen increase in male adult mortality rate, many of which are from the Sub-Saharan African regions. Swaziland, Mozambique, and Lesotho experienced increase in male adult mortality rate at 1–2% annually since 1970. Yet, when we look at periods of 1970–90 and 1990–2010, a different story appears. For all three countries, the increase of adult mortality has been concentrated in years between 1990 and 2010. For Mozambique and Lesotho, adult mortality has been improving from 1970 to 1990. The trend reverted in the late 1980s and mortality started to increase sharply until the early 2000s. This is largely due to the HIV/AIDS epidemic in the region. Take Mozambique as an example; the crude death rate due to HIV/AIDS for men raised sharply from about 1.6 per 100,000 in the late 1980s to about 516 per 100 000 around its peak in 2006. Figure 2 shows the annualized rate of decline in male adult mortality rate for the two periods: 1970–90 and 1990–2010. Based on Figure 3, it is clear that majority of the developing countries have experienced slowdown in improving adult mortality rate during the two periods. Only 57 countries have improved the annualized rate of decline in male adult mortality rate in the period of 1990–2010 comparing to what they have experienced in the period of 1970–90.

On the other hand, the distribution of annualized rate of decline in adult mortality has changed considerably in the past six decades. Figure 3 shows the annualized rate of decline for women in adult mortality rate for the periods of 1950–70, 1970–90, and 1990–2010. It is clear that the heterogeneity in change of adult mortality has increased over the past 60 years. In the period of 1950–70, only five developing countries had deteriorating female mortality rate. The mean annualized rate of decline in adult mortality is well over 1.2% for this period, with a range of −1.17 to 3.96%. Fast forward to the last 20 years; the spread of the distribution is much wider. Annualized rate of decline in female adult mortality rate ranged from −4.62 to 6.1%. Bahrain, China, Maldives, South Korea, Qatar, and Rwanda have made significant strides in reducing female adult mortality rate in this period, all of which have a 3+ percent annualized rate of decline.

At the same time, many countries have suffered increase in adult mortality. Most significant increase in female adult mortality was observed in Haiti, Lesotho, Mozambique, Swaziland, South Africa, and Zimbabwe from 1990 to 2010. All except Haiti (earthquake in 2010) have been affected by the HIV/AIDS epidemic as discussed earlier.

Heterogeneity in Change by Sex

Another phenomenon we have observed is the changing sex mortality differentials among countries. In most country-years, adult mortality rates among men are higher than those among women. There are exceptions though. For example, male adult mortality rates were about 15% lower than female adult mortality rate from 1950 to the mid-1970s. When only focusing on 1970 and 2010, we can find that sex mortality differentials have increased in 111 out of 144 developing countries included in the GBD, Injuries and Risk Factor Study. Sri Lanka, Venezuela, Dominica, and El Salvador have experienced the most significant increase in sex mortality differentials. In Venezuela, male adult mortality rate was about 39% higher than female adult mortality rate in 1970. This relative difference has increased to 153% in 2010. Similarly in El Salvador, the relative difference between male and female adult mortality rates was about 36% in 1970, and this differential increased to about 140% in 2010. It is

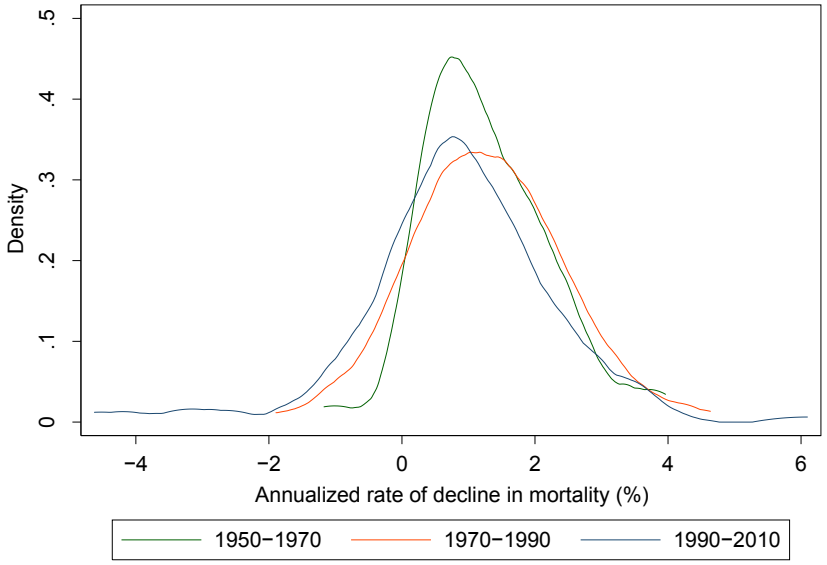

Figure 3 Distribution of annualized rate of decline in adult female mortality.

important to note that for most countries that experienced an increased sex mortality differential, both male and female adult mortality have improved during the same time, with female adult mortality improving at a much faster rate than male adult mortality. In El Salvador, for example, the female adult mortality rate decreased from 192 per 1000 in 1970 to about 104 per 1000 in 2010. At the same time, the male adult mortality rate only managed to improve from 261 per 1000 to 252 per 1000. A full list of the adult mortality results related to the GBD Study is available online with associated visualizations (GBD, 2010: http://vizhub.healthdata.org/mortality/).

Challenges Ahead

It has been shown that adult mortality has made significant progress among developing countries in the past six decades. Yet significant heterogeneity between sexes and among countries still exists. At the same time, long-term improvement masks the heterogeneity in change of mortality by a shorter time period. The rise of the HIV/AIDS epidemic not only threatens the long-term improvement in human mortality but also presents challenges in accurately capturing the true level of adult mortality as the traditional indirect estimate on methods are not well tuned to deal with such situations.

See also: Adult Mortality in Industrialized Societies; Demographic Models; Demographic Techniques: Data Adjustment and Correction; Demographic Techniques: Indirect Estimation; Experimental Methods in Survey Research in Demography; Infant and Child Mortality in the Less Developed World; Life Table; Mortality and the HIV/AIDS Epidemic; Mortality, Epidemiological, and Health Transitions; Mortality: The Great Historical Decline.

Bibliography

Aday, L.A., 1989. Designing and Conducting Health Surveys: A Comprehensive Guide. Jossey-Bass Publishers, San Francisco.

Belli, R.F., Shay, W.L., Stafford, F.P., 2001. Event history calendars and question list surveys: a direct comparison of interviewing methods. Public Opinion Quarterly 65, 45–74.

Bennett, N.G., Horiuchi, S., 1981. Estimating the completeness of death registration in a closed population. Population Index 47, 207–221.

Blacker, J.G., 1977. The estimation of adult mortality in Africa from data on orphanhood. Population Studies 31, 107–128.

Brass, W., Bamgboye, E.A., London School of Hygiene and Tropical Medicine. Centre for Population Studies, 1981. The Time Location of Reports of Survivorship: Estimates for Maternal and Paternal Orphanhood and the Ever-widowed. Centre for Population Studies, London School of Hygiene & Tropical Medicine. University of London, London.

Brass, W., Coale, A. Methods of analysis and estimation, n.d.

Brass, W., Hill, K., 1973. Estimating adult mortality from orphanhood. International Conference on Population 3, 111–123.

Demographic and Health Surveys. Macro International, Inc., n.d.

Feeney, G., 1991. Child survivorship estimation: methods and data analysis. Asian and Pacific Population Forum/East-West Population Institute, East-West Center 5.

Gakidou, E., King, G., 2006. Death by survey: estimating adult mortality without selection bias from sibling survival data. Demography 43, 569–585.

Henry, L., 1960. Mesure indirecte de la mortalité des adultes. Population FRED 15, 457–466.

Hill, K., 1987. Estimating census and death registration completeness. Asian and Pacific Population Forum/East-West Population Institute, East-West Center 1, 8–13.

Hill, K., You, D., Choi, Y., 2009. Death distribution methods for estimating adult mortality: sensitivity analysis with simulated data errors. Demographic Research 21, 235–254.

Hill, K., Trussell, J., 1977. Further developments in indirect mortality estimation. Population Studies 31, 313–334.

Lozano, R., Wang, H., Foreman, K.J., Rajaratnam, J.K., Naghavi, M., Marcus, J.R., Dwyer-Lindgren, L., Lofgren, K.T., Phillips, D., Atkinson, C., Lopez, A.D., Murray, C.J.L., 2011. Progress towards millennium development goals 4 and 5 on maternal and child mortality: an updated systematic analysis. Lancet 378, 1139–1165.

Luy, M., Di Giulio, P., Caselli, G., 2011. Differences in life expectancy by education and occupation in Italy, 1980–94: indirect estimates from maternal and paternal orphanhood. Population Studies 65, 137–155.

Masquelier, B., 2013. Adult mortality from sibling survival data: a reappraisal of selection biases. Demography 50, 207–228.

Mukherjee, B.N., Mukhopadhyay, B.K., 1988. A study of digit preference and quality of age data in Turkish censuses. Genus 44, 201–227.
Murray, C.J.L., Rajaratnam, J.K., Marcus, J., Laakso, T., Lopez, A.D., 2010. What can we conclude from death registration? Improved methods for evaluating completeness. Plos Medicine 7.
Myburgh, C.A.L., 1956. Estimating the fertility and mortality of african populations from the total number of children ever born and the number of these still living. Population Studies 10, 193–206.
Obermeyer, Z., Rajaratnam, J.K., Park, C.H., Gakidou, E., Hogan, M.C., Lopez, A.D., Murray, C.J.L., 2010. Measuring adult mortality using sibling survival: a new analytical method and new results for 44 countries, 1974–2006. Plos Medicine 7, e1000260.
Preston, S.H., Heuveline, P., Guillot, M., 2001. Demography: Measuring and Modeling Population Processes. Blackwell Publishers, Malden, MA.
Rajaratnam, J.K., Tran, L.N., Lopez, A.D., Murray, C.J., 2010. Measuring under-five mortality: validation of new low-cost methods. Plos Medicine 7.
Sullivan, J.M., 1972. Models for the estimation of the probability of dying between birth and exact ages of early childhood. Population Studies 26, 79–97.
Timaeus, I., 1991. Estimation of adult mortality from orphanhood before and since marriage. Population Studies 45, 455–472.
Timaeus, I., Chackiel, J., 1996. Adult Mortality in Latin America. Oxford University Press.
UN Inter-agency Group for Child Mortality Estimation, 2012. Levels & Trends in Child Mortality: Report 2012.
United Nations. Department of International Economic and Social Affairs. Population Division, National Research Council (U.S.). Committee on Population and Demography, 1983. Indirect Techniques for Demographic Estimation. United Nations, New York.
Wang, H., Dwyer-Lindgren, L., Lofgren, K.T., Rajaratnam, J.K., Marcus, J.R., Levin-Rector, A., Levitz, C.E., Lopez, A.D., Murray, C.J., 2012. Age-specific and sex-specific mortality in 187 countries, 1970–2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 380, 2071–2094.
World Population Prospects: The 2010 Revision, 2011. United Nations Department of Economic and Social Affairs Population Division's 1 & 2, Comprehensive Tables.

Relevant Websites

http://www.indepth-network.org/ – INDEPTH Network.
http://ghdx.healthmetricsandevaluation.org/ – Online Compiled Global Health Databank.
http://www.healthdata.org/gbd/data-visualizations – Online Global Burden of Disease Visualizations and Data.
http://childmortality.org/ – Un Inter-agency Group for Child Mortality Estimation.
http://esa.un.org/unpd/wpp/Excel-Data/mortality.htm – United Nations World Population Prospects.

Adult Mortality in Eastern Europe and the Former Soviet Republics

France Meslé, INED, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

From 1965, a strong divergence in life expectancy appeared between Eastern and Western European countries. While life expectancy continued to progress in the West, it stagnated or even reduced in the East. The crisis was especially strong for males at working ages, due to the increase in mortality from circulatory diseases and external causes, with a major role of alcohol. Moreover from the mid-1980s to the late 1990s, countries of the former USSR experienced very large fluctuations of adult mortality, successively due to Gorbachev's antialcohol campaign and its rebound, followed by the deep economic crisis after the transition to market economy. Rapidly after the fall of Berlin Wall, progress resumed in Central European countries, followed by Baltic countries and more recently by Russia and Ukraine. This new progress was mainly due to a decrease in circulatory mortality at all ages. Despite these more favorable trends, the gap in mortality at adult ages between East and West remains important and affects older and older ages.

After World War II, health improved dramatically across Europe with the spread of antibiotics and the large-scale use of immunization. Progress was greatest where life expectancy was lowest. Southern European countries, like Italy and Greece, as well as Central and Eastern European countries, like Russia and Romania, which had lagged far behind in the 1940s almost caught up the most advanced Northern or Western countries (Vallin and Meslé, 2001). In the mid-1960s, all European countries had completed the second stage of the epidemiologic transition as defined by Omran (1971) and had eradicated infectious mortality, especially among young children. They were entering the third stage of man-made and degenerative diseases. In point of fact, mortality from circulatory diseases, traffic accidents, and alcohol abuse was stagnating or increasing almost everywhere at that time. In the West, however, the relative slackening of progress did not last, and by no later than the early 1970s, life expectancy was again on the rise, due both to a reversal of trends in man-made diseases and an accelerated decrease in circulatory diseases. The gains achieved in the West did not spread to Central and Eastern Europe. To the contrary, life expectancy in all Eastern European countries began to plateau or even decrease, especially for males. This huge gap between two parts of Europe continued to widen for three decades, resulting in a yawning divide by the mid-1990s. The evolution of mortality at adult ages, especially working ages, was the main driver of the divergence. During the 1990s, however, unfavorable trends reversed in all Central European countries and in the 2000s progress also started in European republics of the former USSR, thanks to an important decrease of mortality at all ages. Despite this recent progress, mortality at working ages remains very high in most Central and Eastern European countries in comparison with the Western ones.

Health Crisis in Central and Eastern Europe: The Major Role of Adult Mortality

Prior to the mid-1980s, general life expectancy trends were highly comparable in all Eastern and Central European countries (**Figure 1**). Nowhere were they good, but the acuity of the situation differed by country and sex. The worst case was that of males in the republics of the former USSR, where in the five republics represented here (Estonia, Latvia, Lithuania, Russia, and Ukraine), a fall from 1.5 to 3 years occurred in male life expectancy between 1965 and 1984. For females, the decrease was less pronounced but nevertheless there for Russia, Ukraine, and Latvia, while life expectancy was stagnating in Estonia and increasing but slowly in Lithuania. In the Central European countries, Hungarian and Bulgarian males registered the worse results with a decrease of 1.6 and 0.8 years, respectively, while life expectancy was stagnating for males in the other four countries (Czech Republic, Poland, Romania, and Slovakia) and only slowly increasing for females, region-wide.

From the mid-1980s, a sharp distinction appeared between the republics of the former USSR on one side and Central European countries on the other. In the former, trends in life expectancy became very chaotic while remaining very similar. After a period of improvement in the wake of the antialcoholism campaign launched by Gorbachev in 1985 just after coming to power, life expectancy slowly declined as the campaign effects faded. The deterioration accelerated sharply in 1993–94 (or 1995 for Ukraine) as all countries entered the severe economic crisis following the sudden change to a market economy. After the initial shock, life expectancy began rising again, but for Russia and Ukraine, at least, the upturn was short-lived and life expectancy sank again, whereas in the Baltic countries, a continuing rise suggests that they were entering a new stage of progress. Finally in the late 2000s, new progress appeared in Russia and Ukraine.

In Central Europe, the situation appears more favorable as one country after another has resumed its health progress. As early as 1988, trends in the Czech Republic reversed and life expectancy increased very rapidly. Likewise Poland and Slovakia in 1991, Hungary in 1993, Bulgaria and Romania in 1997. As the long-term trends were everywhere more unfavorable to men than to women, the upturn was more spectacular for male than female life expectancy. But females made also substantial gains, with a clear acceleration of the rate of increase in life expectancy in the late 1990s.

The unfavorable long-term trends in Eastern and Central Europe have already been extensively studied (Bourgeois-Pichat, 1985; Meslé, 1991; Okolski, 1993; Meslé and Hertrich, 1997). All these changes in life expectancy were mainly driven by changes in mortality at working ages.

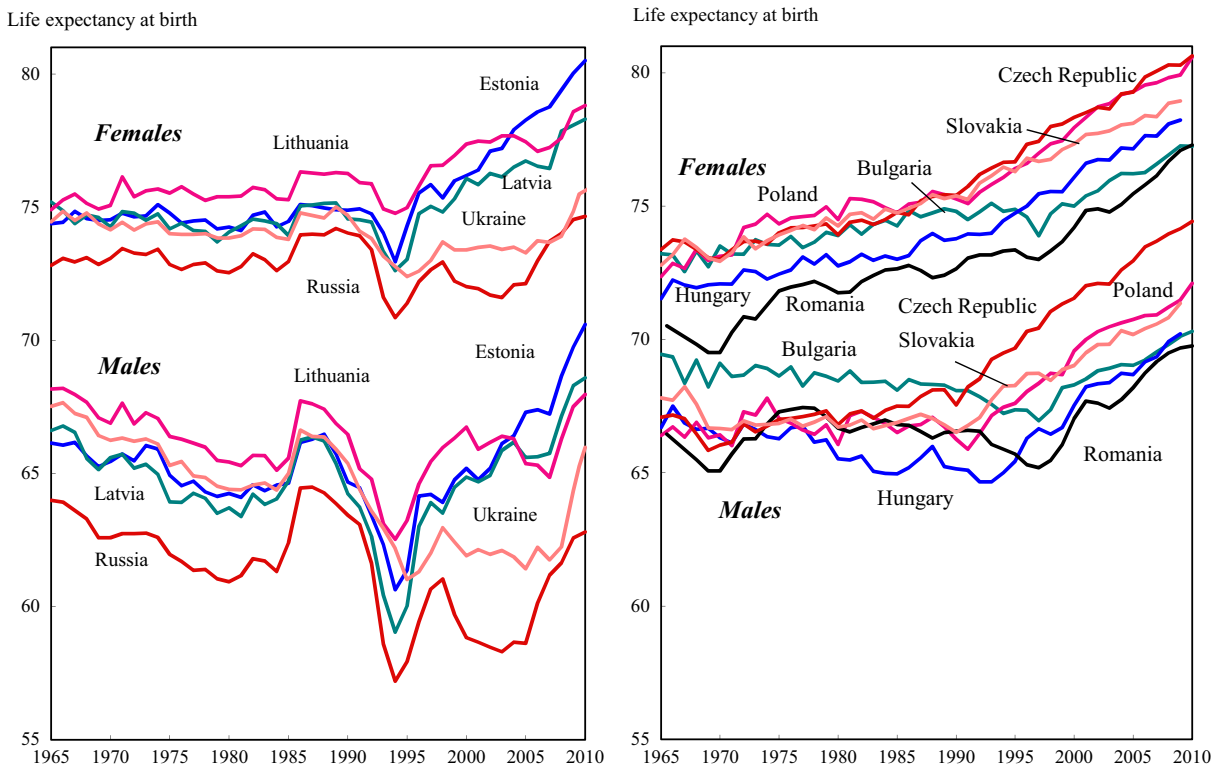

Figure 1 Sex-specific trends in life expectancy in Central European countries and republics of the former USSR since 1965.

Figure 2 displays changes in age-specific male mortality in three countries, Poland, Estonia, and Russia, before and after the upturn. Consequently, the reference year varies from one to another country: 1991 for Poland, 1998 for Estonia, and 2005 for Russia.

In the three countries, the period of decline is characterized by a sharp distortion of the mortality age structure with a significant increase at adult ages, between 25 and 65 (Figure 2). Below age 15, there was a significant reduction of mortality risks while above age 70, death rates stagnated. Whatever the country, the major role played by adult ages in the unfavorable trends appears here clearly. The worsening is especially impressive in Russia where male mortality at age 50 was multiplied by 2.3 in 40 years, but the deterioration was also very important in Poland (a multiplication of 1.7 at age 45 but in 26 years only), or in Estonia (1.9 at age 45 in 33 years).

The changes in death rates, which occurred since the upturn in the three countries, are more equally distributed among all ages. Except old ages, death rates were considerably reduced. Once again the progress could seem less important in Russia, but considering that the reversal occurred in 2005 only, the reduction observed in 5 years is notable. The new stage that these different countries are entering one after the other appears to be one of a general improvement in health at all ages, and is no longer concentrated in a period of life as was the case for the worsening at specifically adult ages.

All these changes in mortality age structure depend on trends in causes of death. To study cause of death trends, we

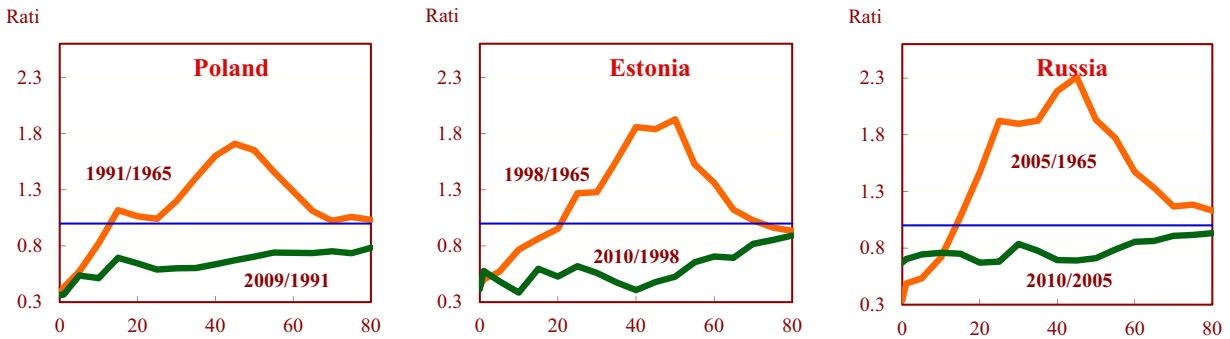

Figure 2 Changes in male mortality age structure during a period of deterioration and a period of improvement in Poland, Estonia, and Russia.

use reconstructed series of deaths by cause that allow to get rid of the ruptures caused by the changes in classification of causes of death (Meslé et al., 1996; Fihel et al., 2010; Meslé et al., 2012). Deaths were gathered in seven large groups of causes.

As for changes in mortality age patterns, the period 1965–2010 is divided into two subperiods that are different for each country and correspond to the decrease and increase of life expectancy at birth, respectively. For each subperiod, we calculated the contributions by age groups of seven groups of causes to changes in life expectancy (Andreev, 1982). **Figure 3** gives the contribution for each age group and each group of causes.

As expected from age pattern changes, the decrease in life expectancy was closely correlated to the rise in mortality at working ages while, everywhere, progress in infant mortality checked the decline (upper panel of graphs in **Figure 3**). In Poland, mortality from circulatory diseases was the main cause of health decline at adult ages. In Estonia and Russia, in addition to the important role played by these diseases, external causes contributed to the life expectancy decrease significantly.

Looking at the period of improvement (lower panel of graphs in **Figure 3**), it comes as no surprise to see that mortality from circulatory diseases was the main determinant of the upturn in life expectancy in Poland, where gains in mortality from this cause account for 54% of the total. In this period, gains are rather equally distributed between adult age groups whereas in the previous period, losses were more concentrated on relatively young age groups. The phenomenon is quite different in Estonia and Russia: in these two countries external causes are mainly responsible for the progress that has been recently observed. This group of causes explains 40% of the gains in Estonia and 43% in Russia versus 36 and 35%, respectively, for circulatory diseases. Furthermore the decrease in mortality from external causes is especially important at young adult age groups. Finally in the two countries of the former USSR, the gains observed in this last period of improvement almost mirrored the losses registered in the previous decades, even if in Estonia, gains due to the decrease in mortality from circulatory diseases are expanding to older ages.

Looking at male trends in Poland, it is clear that this country has now entered a new step of health transition and was successful in the reduction of circulatory diseases, the so-called cardiovascular revolution (Meslé and Vallin, 2006). These results were already clearly shown for all the Central European countries (Meslé, 2004). In Estonia and Russia, however, male trends remain uncertain. Considering the importance of gains at rather young ages for external causes, it could be feared that these countries are just dealing with a new phase of fluctuations, comparable to what was observed from the mid-1980s to the mid-1990s (see Section Fluctuations and Mortality at Working Ages in the Former USSR) and that this recent health improvement will not be sustainable. Indeed the recent favorable trends could be related to new antialcoholic measures, which were taken in the recent years and could be very short lasting (Shkolnikov et al., 2013).

Fortunately, female mortality trends allow to be a little more optimistic for the future (**Figure 4**). Age and cause components of the life expectancy changes in the period of improvement appear completely different of those that prevailed during the period of deterioration. As for males, unfavorable trends were mainly due to a stagnation or even an increase of mortality rates between ages 25 and 60. On the contrary, not only in Poland but also in Estonia, recent progress is much more related to a decrease of mortality at older ages (over age 60), almost totally due to gains against circulatory diseases. Even if this is still less clear for Russia, the process could seem comparable (Grigoriev et al., 2014). In all these countries, the future challenges are now shifting to older ages (see Section An excess mortality shifting to older ages).

Fluctuations and Mortality at Working Ages in the Former USSR

The wide fluctuations in life expectancy that have occurred in the republics of the former USSR since 1985 have also been widely discussed in the literature (Shkolnikov and Nemtsov, 1997; Leon et al., 1997; Shkolnikov et al., 1998; Gavrilova et al., 2001, 2002). Changes in adult mortality were the main determinants of these fluctuations again. The Russian case, taken here as an example is the most emblematic.

In order to obtain a good understanding of the role played by age- and cause-specific mortality in life expectancy changes between the mid-1980s and the late 1990s in Russia, it is important to analyze separately two major events that affected the country:

- 1984–92: the antialcohol campaign and its reversal,
- 1992–98: the 1992–94 economic and social crisis and the postcrisis period.

First, the gain in life expectancy in Russia (1984–1987) was mainly due to the reduction in deaths from injury and poisoning among young adults (**Figure 5**), directly linked, as we know, to the decline in alcoholism (Shkolnikov and Nemtsov, 1997) and, consequently, to the decline in diseases of the circulatory system, which was closely related to the reduction in alcohol consumption. Indeed the campaign involved a reduction in the state-controlled production and sale of alcohol, an increase in the price of alcoholic drinks, a restriction on private distilling, and the development of specialized medical services. However, all these measures were just a flash in the pan, since they were quickly abandoned.

The second change – the fall in life expectancy observed from 1987 to 1992 – fitted a pattern of age- and cause-specific losses that was highly symmetrical to that of the gains in the preceding period. This is because it was dominated by the rise in deaths from external causes among young adults, reinforced by an increase in mortality from diseases of the circulatory system.

This first large fluctuation, resulting from the antialcohol campaign, did much less to disrupt female mortality than male. Even though the same upheavals as for males can still be observed leading in the same direction for females, they result in only modest effects. The antialcohol policy also had consequences for females, but these were much smaller and somewhat different in nature, since they were almost entirely concentrated on diseases of the circulatory system.

The period that followed saw the decline in life expectancy continue, but with a clearly different dynamic in causes of death – one which contradicts the thesis sometimes put forward that the entire fall in life expectancy in the 1990s can be explained by a simple 'counterselection' phenomenon

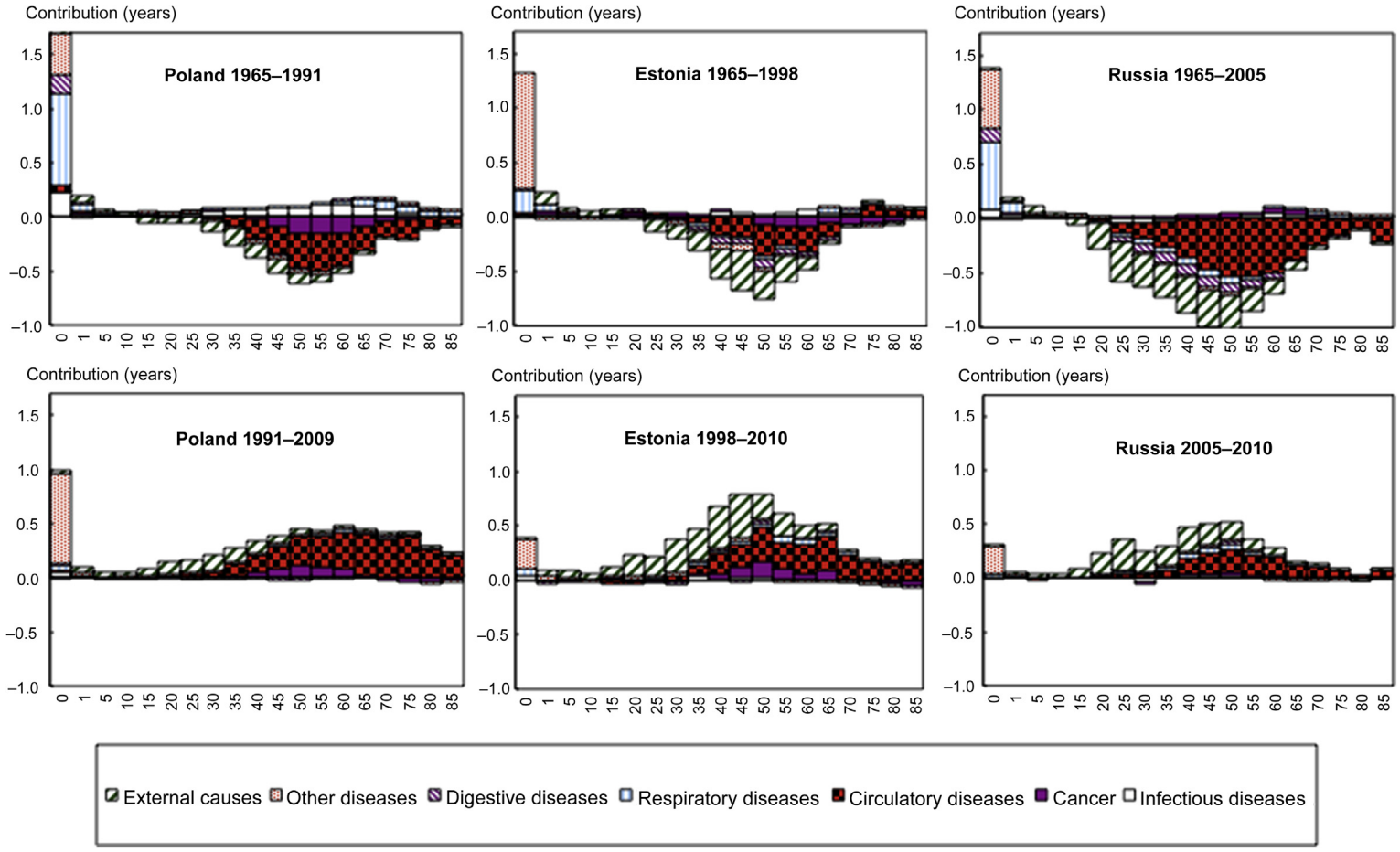

Figure 3 Age and cause components of changes in male life expectancy during a period of deterioration and a period of improvement in Poland, Estonia, and Russia.

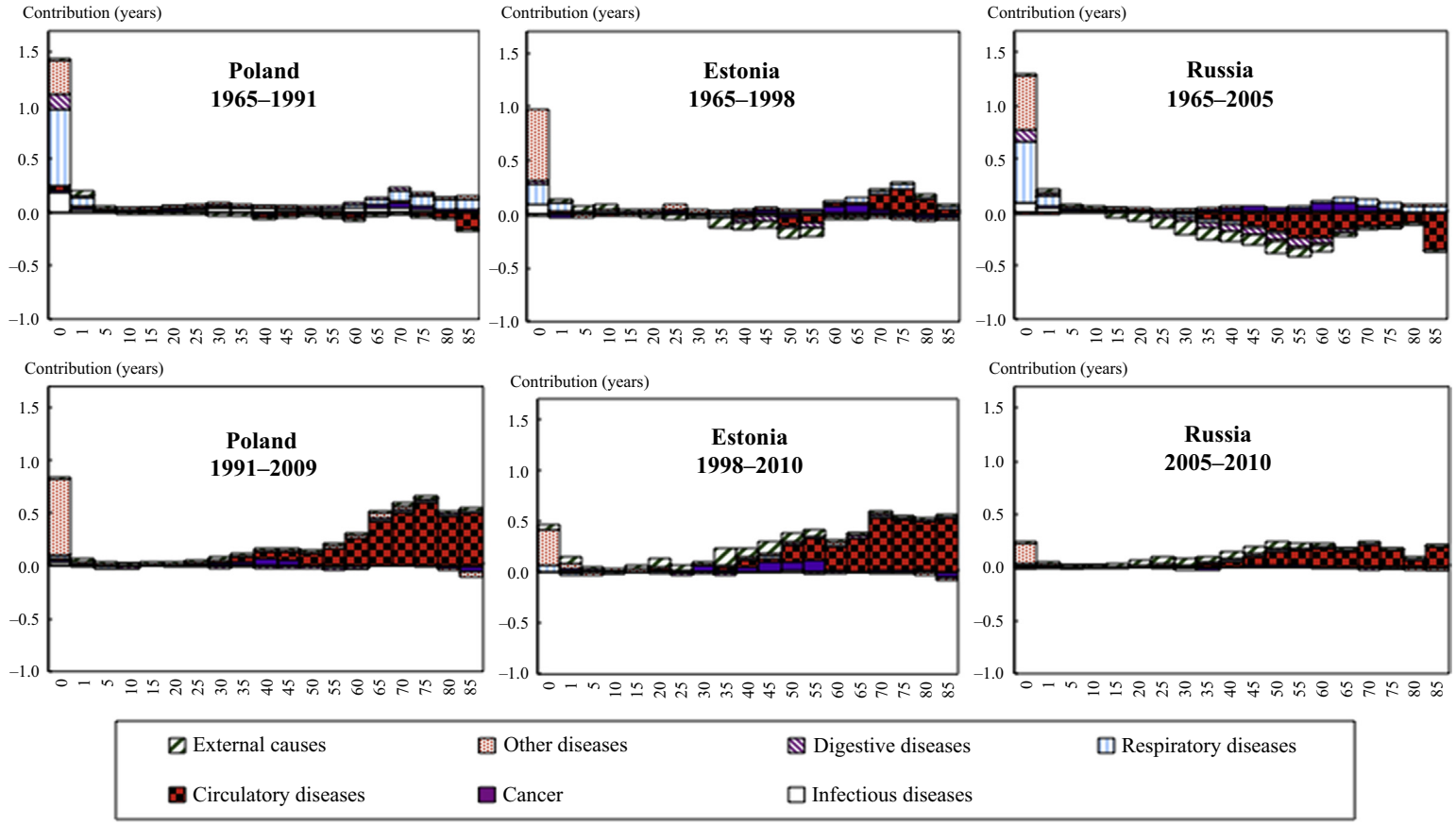

Figure 4 Age and cause components of changes in female life expectancy during a period of deterioration and a period of improvement in Poland, Estonia, and Russia.

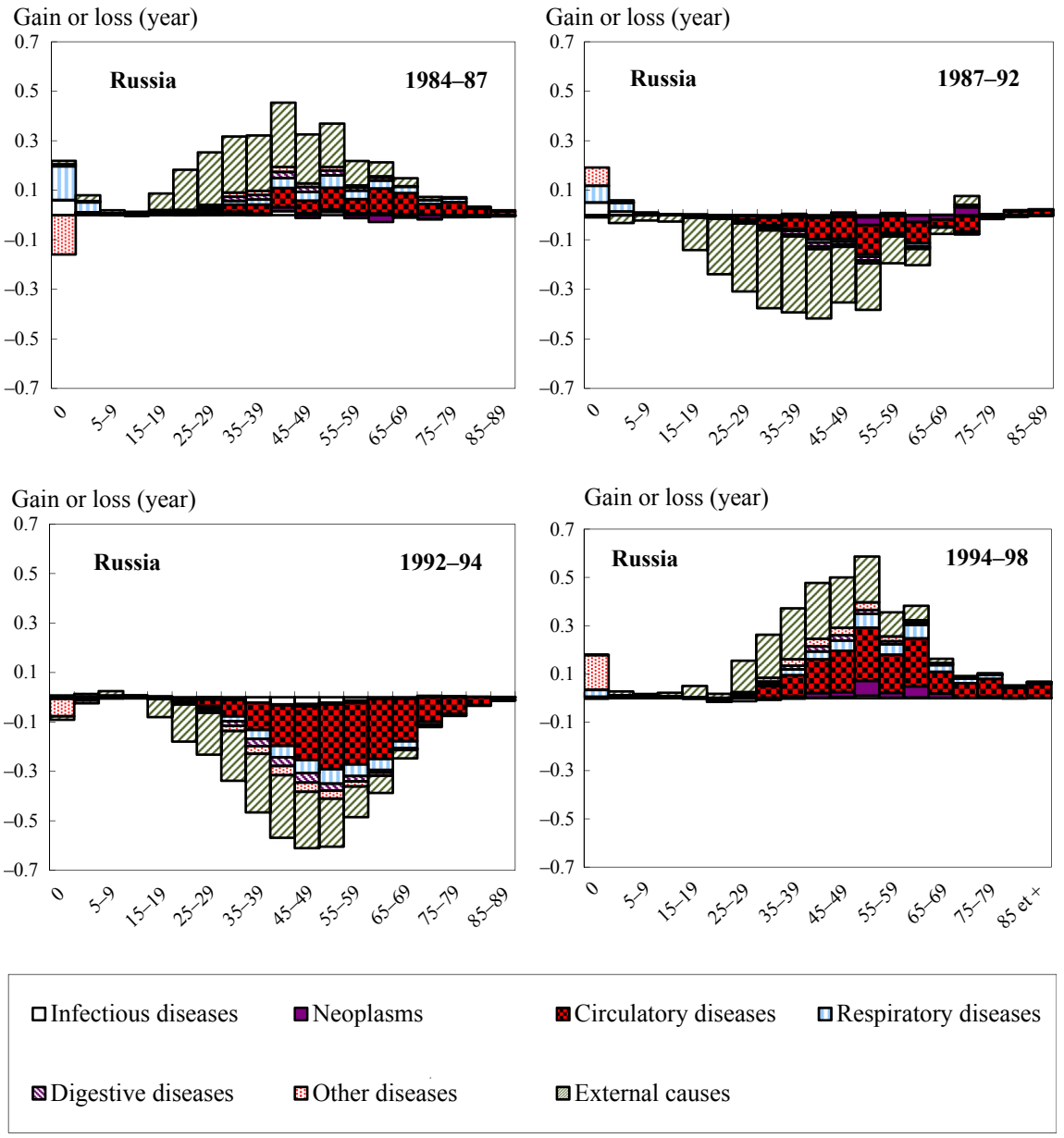

Figure 5 Age and cause components of changes in male life expectancy between 1984 and 1998 in Russia.

(Avdeev et al., 1998; Vishnevskii, 2000). According to this thesis, by leading to an abrupt decline in mortality, the antialcohol campaign delayed the deaths of weaker than average individuals who would otherwise have died. Therefore the population's average state of health deteriorated and, naturally, mortality increased in the years that followed. A cause-specific examination of deaths refutes this thesis in two ways. If it were correct, the rise in mortality immediately following the end of the antialcohol campaign should have resulted from all kinds of causes to which the lives previously saved were more vulnerable. Yet, as we have just seen, this rise was entirely linked to alcohol-related causes (just like the fall in mortality during the campaign). In contrast, from 1992, although mortality had already returned to its precampaign level, the continued rise in mortality resulted from a very different range of causes. Of course, as always in mortality trends for Eastern European countries, the major role played by diseases of the circulatory system and deaths from external causes in the adult age groups is clearly apparent; but these causes did not act alone (all the other groups of causes helped to aggravate the situation) and the losses in life expectancy have an older age profile, even though the fall in infant and child mortality that was also visible in the preceding period gave way to a rise. Far from this new rise in mortality – several years after the end of the antialcohol campaign – being attributable to the deaths of vulnerable individuals temporarily saved by the campaign, it probably relates to a new factor: the economic and social crisis which, at that time, dealt a heavy blow to populations suddenly subjected to the perverse effects of the transition to a market economy, and this blow was the strongest for males

in working ages. After 1995, economic decline continued for a few more years, but at a very reduced pace. Fairly quickly too, after the crisis had passed, the populations involved found the resources to adapt to circumstances, doing more to safeguard their chances of survival. So life expectancy was able to start rising again. In Russia, age- and cause-specific patterns appear almost perfectly symmetrical (Figure 5, lower panel) between the crisis period (1992–94) and the postcrisis period (1994–98).

Like the previous large fluctuation, this second one, linked to the 1992–94 crisis, did much less to disrupt female mortality than male. The sharp movements observed for males produced only modest effects for females, even though still in the same direction. Here too, the really dominant impact was that of diseases of the circulatory system alone, with the effect of deaths from external causes being only secondary in importance (like the other groups of causes).

An Excess Mortality Shifting to Older Ages

Considering long-term trends or fluctuations, it appears that Central and Eastern European women have always done better than men: a less strong worsening, less ample fluctuations, and a more promising trend reversal. These more favorable trends are due to a better health situation at middle ages: women suffered less from alcohol-related mortality and trends in external causes and circulatory diseases were less problematic than for males. In the most recent years, the trend reversal was more driven by progress at older ages, which were more pronounced for women than for men. Does it mean that the difference between Eastern and Western Europe is going to disappear?

Figure 6 displays the age and cause components in the differences in life expectancy between France and Russia, and between France and Poland in 2009 for males and females. The difference appears still very important for males at working

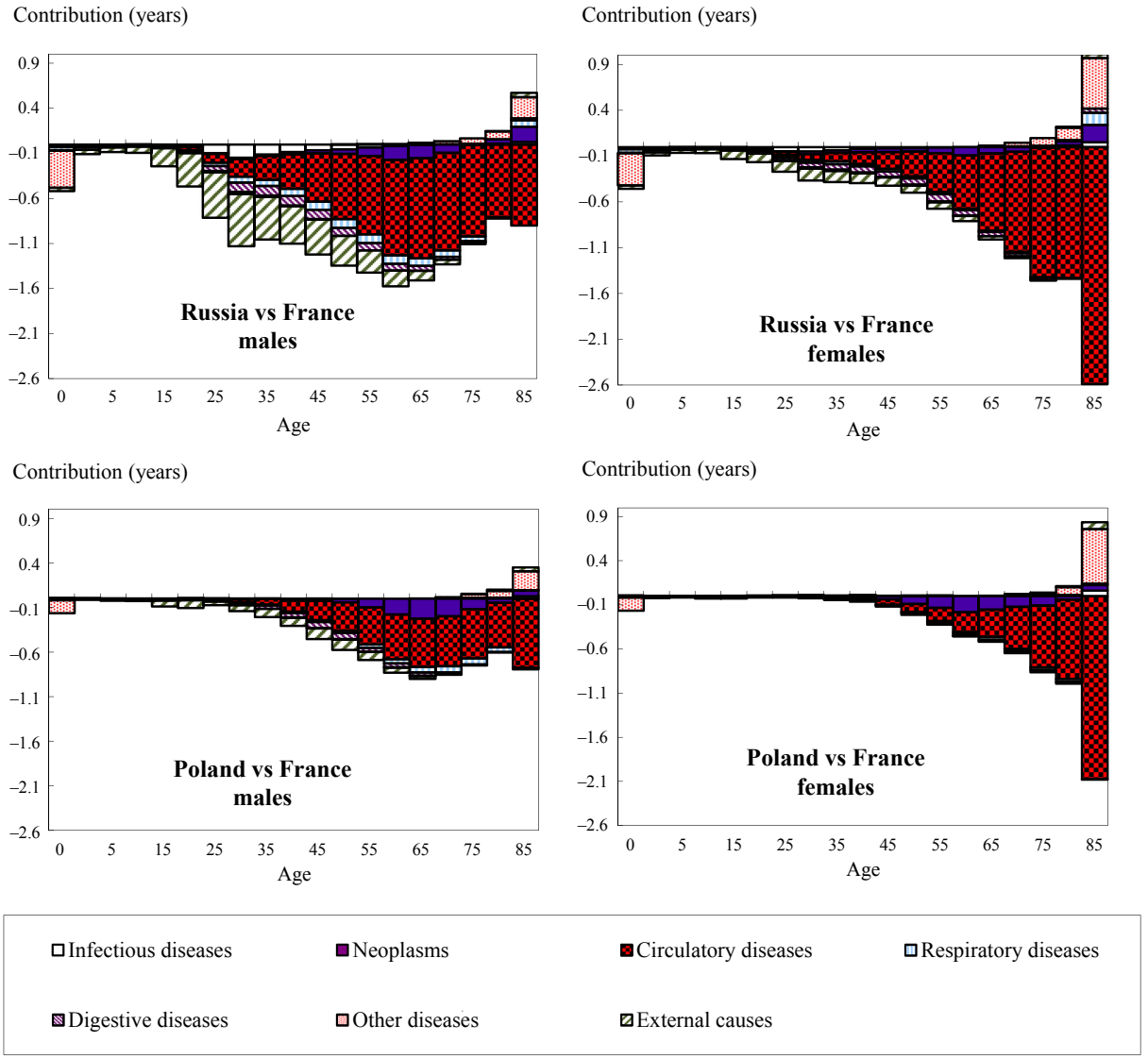

Figure 6 Age and cause components of differences in life expectancy between Russia and France, and between Poland and France in 2009.

ages. For Russia, compared to France, 63% of the total gap (16 years) is due to differences in mortality between age 20 and age 65 and, despite the recent progresses, this part is 48% for Polish males (for a total gap of 7 years with French males). Mortality over age 65 accounts only for 30% of the difference between Russian and French males, and for 48% of that between Polish and French. Even if the recent mortality decrease in Eastern European countries was the most pronounced at relatively young adult ages, there is still a lot to gain at those ages to catch up with Western countries.

On the female side, the situation is altogether different. The gaps in life expectancy observed between Russian and French women (11 years) or between Polish and French women (5.5 years) are mainly related to differences in death rates at older ages (**Figure 6**). Excess mortality over age 65 of Russian females, as compared to French ones, accounts for 57% of the total difference and this part reaches 74% of the French–Polish gap. In the recent years, it is well known that the continuation of life expectancy progress in the most advanced countries was due to a dramatic decrease of mortality at older and older ages (Glei et al., 2010). In that matter, Central and Eastern European countries are far from doing as well as their Western counterparts. In the coming years, it will be a challenging issue for these countries to manage with the health of the elderly population. At age over 85, Russia and Poland females seem to have some advantages for noncirculatory causes of death. Most probably, however, this is only an artifact due to less precise diagnoses among older women. In Eastern European countries, the propensity to attribute deaths at older ages to the circulatory diseases was often stressed. This leads here probably to an overestimation of the part played by circulatory diseases in the difference between East and West, and reversely to the wrong conclusion that Poland and Russia could be in better position for other causes of death. Whatever the differences in the quality of the cause of death statistics, it remains that ages over 85 are now contributing substantially to the differences.

From the mid-1960s, adverse but quite similar mortality trends in Central Europe and the former USSR long contributed to widening the gap between these countries and the rest of Europe. The divergence was mainly due to very different trends in adult mortality at working ages, related to external causes and circulatory diseases. After successfully tackling infectious diseases, industrialized countries had to look for ways of reducing degenerative diseases, and especially circulatory diseases, which were the main factors of mortality, as well as to control behavior-related pathologies (excess alcohol consumption, tobacco smoking). In Western countries, health policies that combined preventative measures with new developments in treatments gradually vanquished these diseases, and their mortality declined substantially. Eastern European countries, by contrast, failed to reverse the increase in circulatory disease mortality that largely contributed to the stagnation or even the decrease of life expectancy (Meslé, 1991; Okolski, 1993; Meslé and Vallin, 2002).

Since the mid-1980s, however, adult mortality trends have developed very differently in Central Europe and the former USSR. Between the late 1980s and the late 1990s, one Central European country after another has reversed long-term negative trends and progress seems well under way, particularly due to a marked decline in mortality from circulatory diseases (Meslé, 2004). Favorable trends, which were observed in Western Europe as long ago as the early 1970s, have now spread to the countries of Central Europe. It is still difficult to assess the main determinants of such a reversal. The progress probably results from the combination of several factors, such as changes in diets, the growth of systematic prevention and screening, the spread of new forms of treatment, and cardiac surgery, in a context of deep reforms of health systems.

At the same time, countries of the former USSR have experienced very marked fluctuations in adult mortality, especially among males, related to the social and economic traumas that these countries have undergone at the end of the twentieth century. After an initial increase in life expectancy following the antialcoholism campaign launched by Gorbachev in 1985, mortality began to rise again as the effects of the campaigns dwindled. The increase accelerated dramatically in 1993–94 when all these countries entered a harsh economic crisis linked to the sudden transition to a market economy. The crisis resulted in a complete dismantling of public services, including the health system, at a time when people were suffering food shortages and psychological stress (Shkolnikov et al., 1998; Gavrilova et al., 2001). Up to the mid-1990s, the trends were completely parallel in all the European republics of the former USSR. From the late 1990s, however, they have begun to diverge. In Baltic countries, mortality trends have more or less mirrored those of Central European countries (Jasilionis et al., 2011). In Russia and Ukraine, the progress was longer to appear. Since the mid-2000s, however, these countries could have also entered a new step of the health transition in succeeding to control alcohol-related pathologies and circulatory diseases (Grigoriev et al., 2014).

If the increase of mortality among young- and middle-age adults was the main factor of the unfavorable trends in life expectancy observed in these different countries, the progress recently observed is driven by a reduction of mortality at all adult ages, including older ages. This phenomenon is especially visible for females. However, when compared to Western countries, the mortality at older ages remains much higher in Central and Eastern Europe. To consolidate the recent favorable trends, the latter need now to take care of the health of their elderly population. As for other advanced countries, this is the main health challenge of the coming years.

See also: Adult Mortality in Industrialized Societies; Aging and Health in Old Age; Mortality of the Oldest-old; Mortality, Epidemiological, and Health Transitions; Mortality: The Great Historical Decline.

Bibliography

Andreev, Evgueni, 1982. Metod komponent v analize prodolzhitelnosti zhizni. Vestnik Statistiki 3, 42–47.

Avdeev, Alexandre, Blum, Alain, Zakharov, Serge, Andreev, Evgeny, 1998. The reactions of a heterogeneous population to perturbation. An interpretative model of mortality trends in Russia. Population: An English Selection 10 (2), 267–302.

Bourgeois-Pichat, Jean, 1985. Recent chances in mortality in industrialized countries. In: Vallin, Jacques, Lopez, Alan (Eds.), Health Policy, Social Policy, and Mortality Prospects. Ordina Éditions, Liège, pp. 507–539, 557 p.

Fihel, Agnieszka, Meslé, France, Vallin, Jacques, 17–19 June 2010. Mortality by Causes of Death in Poland 1970–2007: Preliminary Findings. Third Human Mortality Database Symposium, Paris, 21 slides.

Gavrilova, Natalia S., Evdokushkina, Galina N., Semyonova, Victoria G., Gavrilov, Leonid A., 2001. Economic Crises, Stress and Mortality in Russia. Center on Aging, University of Chicago, Chicago, 28 p. (Paper presented at the 2001 PAA meeting, Washington).

Gavrilova, Natalia S., Semyonova, Victoria G., Evdokushkina, Galina N., 2002. Mortality Crisis in Russia: New Health Threats. Center on Aging, University of Chicago, Chicago, 29 p. (Paper presented at the 2002 PAA meeting, Atlanta).

Glei, Dana, Meslé, France, Vallin, Jacques, 2010. Diverging trends in life expectancy at age 50: a look at causes of death. In: Crimmins, Eileen M., Preston, Samuel H., et al. (Eds.), International Differences in Mortality at Older Ages: Dimensions and Sources. National Academies Press, Washington, pp. 17–67 (coll. Panel on Understanding Divergent Trends in Longevity in High-Income Countries; National Research Council).

Grigoriev, Pavel, Meslé, France, Shkolnikov, Vladimir M., Andreev, Evgeny, Fihel, Agnieszka, Pechholdova, Marketa, Vallin, Jacques, 2014. The recent mortality decline in Russia: beginning of the cardiovascular revolution? Population and Development Review 40 (1), 107–129.

Jasilionis, Domantas, Meslé, France, Shkolnikov, Vladimir M., Vallin, Jacques, 2011. Recent life expectancy divergence in Baltic countries. European Journal of Population 27 (4), 403–431.

Leon, David A., Chenet, Laurent, Shkolnikov, Vladimir M., Zakharov, Sergei, Shapiro, Judith, Rakhmanova, Galina, Vassin, Sergei, McKee, Martin, 9 August 1997. Huge variation in Russian mortality rates 1984–94: artefact, alcohol, or what? The Lancet 350, 383–388.

Meslé, France, 1991. La mortalité dans les pays d'Europe de l'Est. Population 46 (3), 599–650.

Meslé, France, 2004. Mortality in Central and Eastern Europe: long-term trends and recent upturns. Demographic Research, 46–70 (Special Collection 2. Determinants of Diverging Trends in Mortality).

Meslé, France, Hertrich, Véronique, 1997. Évolution de la mortalité en Europe: la divergence s'accentue entre l'Est et l'Ouest. In: Congrès international de la population. Beijing 1997. UIESP, Liège, pp. 479–508, 1532 p.

Meslé, France, Vallin, Jacques, Hertrich, Véronique, 2012. The Baltic Countries: A Delayed Entrance in the Cardiovascular Revolution Workshop "New Measures of Mortality – What Do They Mean?". European Working Group on Health, Tallinn (Estonia).

Meslé, France, Vladimir, Shkolnikov, Véronique, Hertrich, Vallin, Jacques, 1996. Tendances récentes de la mortalité par cause en Russie, 1965–1994, 140 p. + 2 disquettes. (Données statistiques n°2). INED, Paris.

Meslé, France, Vallin, Jacques, 2002. Mortality in Europe: the divergence between East and West. Population-E 57 (1), 157–198.

Meslé, France, Vallin, Jacques, 2006. The health transition: trends and prospects. In: Caselli, Graziella, Wunsch, Guillaume, Vallin, Jacques (Eds.), Demography – Analysis and Synthesis: A Treatise in Population Studies, vol. II. Academic Press, New York, pp. 247–259, 2976 p.

Okolski, Marek, 1993. East-West mortality differentials. In: Blum, Alain, Rallu, Jean-Louis (Eds.), Démographie Européenne. II. Dynamiques Démographiques. John Libbey/INED, Paris, pp. 165–189.

Omran, Abdel R., 1971. The epidemiologic transition: a theory of the epidemiology of population change. Milbank Memorial Fund Quarterly 49 (4), 509–538.

Shkolnikov, Vladimir M., Andreev, Evgeny M., McKee, Martin, Leon, David A., 2013. Components and possible determinants of decrease in Russian mortality in 2004–2010. Demographic Research 28 (32), 917–950.

Shkolnikov, Vladimir M., Cornia, Giovanni, Leon, David, Meslé, France, 1998. Causes of the Russian mortality crisis: evidence and interpretations. World Development 26 (11), 1995–2011.

Shkolnikov, Vladimir M., Nemtsov, Alexander, 1997. The anti-alcohol campaign and variations in Russian mortality. In: Bobadilla, Jose L., Costello, Christine A., Mitchell, Faith (Eds.), Premature Death in the New Independent States. National Academy Press, Washington, DC, pp. 239–261, 404 p.

Vallin, Jacques, Meslé, France, 2001. Trends in mortality in Europe since 1950: age-, sex- and cause-specific mortality. In: Vallin, Jacques, Meslé, France, Valkonen, Tapani (Eds.), Trends in Mortality and Differential Mortality. Council of Europe Publishing, Strasbourg, pp. 31–186, 334 p. (Population Studies n° 36).

Vishnevskii, Anatolii G., May 2000. Подъем смертности в 90-е Годы: факт или артефакт? [The increase in mortality during the 1990s: fact or artefact?]. Население и общество [Population and Society] 45, 1–4.

Adult Mortality in Industrialized Societies

Marc Luy, Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/OEAW, WU), Vienna Institute of Demography of the Austrian Academy of Sciences, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Adult morality in today's industrialized societies is characterized by steady increases of life expectancy with a tendency toward convergence of mortality levels. Exceptions are countries from Central and Eastern Europe and the Arabian Peninsula. The populations differ more in the causes of death than in the age pattern of mortality changes during the past decades. Women outlive men in all populations. However, differences exist in the time pattern of the sex gap. In general, levels and trends of mortality and sex differences as well as the prevailing driving factors reflect to a large extent the development stages of the populations.

Introduction

In this article, industrialized countries are defined as those nations that were classified as 'high-income economies' by the World Bank in 2013 (See http://data.worldbank.org/about/country-classifications/country-and-lending-groups#OECD_members; accessed on 07.10.13.) and had a population of more than 1 000 000 people on 1 July 2010 according to the 2012 revision of the United Nations World Population Prospects (United Nations, D.o.E.a.S.A., Population Division, 2013). (If not stated otherwise, all figures presented in this article are taken from or calculated from this source.) This selection includes 45 countries with a total of 1.26 billion people in 2010, representing 18.2% of the world population. In the quinquennial 2005–10, 56.5 million deaths occurred in these populations, constituting approximately one-fifth of all deaths worldwide. This at-a-first-glance high proportion is due to the high average age of today's industrialized countries. The proportion of people aged 60 years and older is at 20.8%, almost double that of the total world population (11.1%). This becomes even more apparent when one considers only the population aged 80 years and above. While the proportion of individuals older than 80 years is 1.6% in the entire world population, it is 4.1% in the selected 45 industrialized countries.

Adult mortality is the central factor of current levels and trends of overall survival conditions in today's industrialized countries: In the period 2005–10, 98.8% of all deaths occurred in those aged older than 15 years. In the entire world population, the corresponding proportion is 82.7%. Figure 1(a) shows the age-specific death rates for industrialized countries in comparison to the whole world population in 2005–10. It becomes apparent that the largest differences between industrialized countries and the

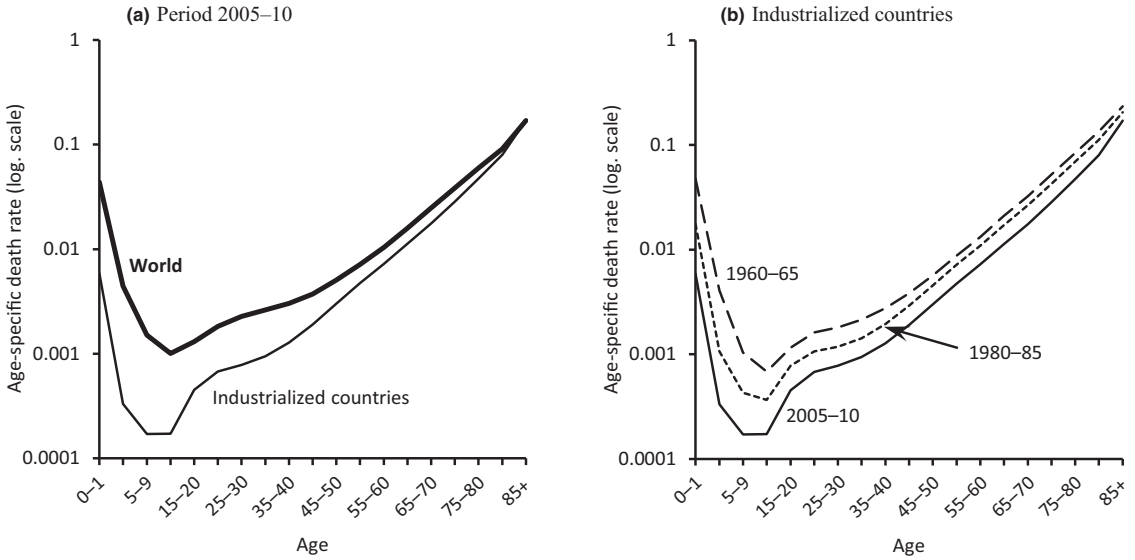

Figure 1 Age-specific death rates in different periods, both sexes. Note: Figures for industrialized countries are unweighted averages. Source: United Nations, D.o.E.a.S.A., Population Division, 2013. World Population Prospects: The 2012 Revision, DVD edition. United Nations, New York; and own calculations with data of United Nations, D.o.E.a.S.A., Population Division (2013).

world occur in younger ages. Between birth and the age of 10 years, death rates in industrialized countries are around 90% below the level of the total world. In adult ages this advantage reduces steadily. Between ages 15 and 35, death rates of industrialized populations are between 58 and 66% below the world average; between ages 40 and 65, it is 30–50%; and above the age of 65, the advantage reduces to 10–25%.

In the course of the past 50 years, industrialized countries experienced a continuous decrease of mortality in all ages. This is illustrated in Figure 1(b), which displays the age-specific death rates in today's industrialized countries in the periods 1960–65, 1980–85, and 2005–10. The graph shows further the shift of improvements from younger to older ages. Between 1960–65 and 1980–85, improvements were larger in younger ages, whereas between 1980–85 and 2005–10 stronger improvements occurred in older ages. The following sections provide overviews of the different levels and trends of adult mortality (both sexes combined) in today's industrialized countries and also describe the underlying age- and cause-specific mechanisms and the differences between women and men. The final section summarizes briefly the driving factors of past and current levels of adult mortality and speculates about future trends. More details about regional or country-specific mortality histories and conditions can be found in the *International Handbook of Adult Mortality* (Rogers and Crimmins, 2011).

Current Levels of Adult Mortality

Table 1 summarizes different indicators of adult mortality in today's industrialized countries for the period 2005–10. Remaining life expectancy is given for the ages 15 (e_{15}), 60 (e_{60}), and 80 (e_{80}) years. Additionally, the table includes the survival probabilities from the age of 15 to 60 years ($_{45}l_{15}$) and from the age of 60 to 80 years ($_{20}l_{60}$). The populations are grouped by seven major geographical areas: America (AMA), Northern Europe (NEU), Western Europe (WEU), Central and Eastern Europe (CEE), Southern Europe (SEU), Western Asia (WAS), as well as Oceania, Eastern, and Southeast Asia (OEA). For the entire population of industrialized countries, e_{15} is 63.7 years and thus approximately 5 years higher than in the total world (58.8 years). However, the differences between the individual industrialized countries are still considerable, amounting to almost 15 years. The highest life expectancy at the age of 15 years is reached in Japan (68.0 years), whereas the lowest can be found in the Russian Federation (53.3 years). Even at higher ages the differences in levels of remaining life expectancy are large, amounting to 8.2 years between Japan and the Russian Federation at the age of 60 years and 5.5 years between Japan and Kuwait at the age of 80 years.

The major geographical areas with the largest heterogeneity among industrialized countries in mortality levels are AMA and CEE where e_{15} varies by 9.1 and 8.9 years, respectively. In AMA, the extent of the span is due to Trinidad and Tobago which exhibits the highest mortality levels according to all five mortality indicators. The other countries, including Canada, Chile, Puerto Rico, the United States, and Uruguay, show considerably better and less differentiated mortality conditions, with Canada being the leader in all indicators of adult mortality. Between these five countries, the level of e_{15} varies by 3.3 years. In CEE, the large range is due to the high mortality level of the Russian Federation. The difference from Lithuania, the country with the second lowest level of life expectancy at age 15 in this geographical area, is 3.7 years. The lowest mortality is prevailing in the Czech Republic, where e_{15} reached 62.2 years in the period 2005–10. The Baltic states, Estonia and Latvia, are closer in their mortality levels to Lithuania, whereas Slovakia and Poland are closer to the level of the Czech Republic. Note that Poland has the highest life expectancy at age 80. This indicates that the Polish population exhibits the lowest mortality in highest ages among the industrialized populations of CEE.

The most homogeneous geographical areas of industrialized countries are NEU, WEU, and OEA. Within these regional units, e_{15} varies only by 2.3–2.4 years. The lowest level of mortality is prevalent in OEA where life expectancy ranges between 65.6 years in the Republic of Korea and 68.0 years in Japan. In NEU these values range from 64.0 years in Denmark to 66.4 years in Sweden. The level of e_{15} in WEU lies in between. Belgium stands at the lower end with 65.0 years, and Switzerland exhibits the highest value of 67.3 years. The homogeneity is most distinct in the two European areas, reflected by the fact that the countries with the highest and lowest mortality levels change between the different indicators for adult mortality. In WEU, for instance, Switzerland shares the first rank with France in e_{60} and falls behind the latter in e_{80}. Belgium holds the last place in the rankings only in life expectancy at ages 15 and 60. In e_{80} the Netherlands is at the end of the ranking, in $_{45}l_{15}$ France, and in $_{20}l_{60}$ Germany. In OEA, Japan and the Republic of Korea hold the first and the last place in all but one indicator. In $_{45}l_{15}$, the highest value can be found in Hong Kong and the lowest in New Zealand.

The level of heterogeneity in SEU and WAS ranges between those of the other regions, being 5.2 and 6.4 years in e_{15}, respectively. In the former, the rather large span is due to the relatively high mortality in Croatia with e_{15} being 61.7 years. In all other countries, life expectancy at the age of 15 years ranges between 64.0 (Slovenia) and 66.9 years (Italy). Consequently, Croatia is lowest in the ranking of all presented indicators of adult mortality. At the top, Italy is replaced by Spain as the leader in life expectancy at ages 60 and 80. The countries belonging to WAS are also characterized by very different mortality levels. The division occurs mainly between the populations of the Arabian Peninsula on the one hand with still relatively high mortality and Cyprus and Israel on the other with low mortality. Among the latter e_{15} is at the level of SEU countries, being 64.4 years in Cyprus and 66.3 in Israel. In the industrialized Arab countries Bahrain, Kuwait, Oman, Saudi Arabia, and the United Arab Emirates, e_{15} ranges between 59.9 (Kuwait) and 61.8 years (Bahrain). Qatar stands out as it has reached a life expectancy level between those of the Arab countries and Israel respective Cyprus (63.4 years). Israel and Kuwait are ranked first and last in four of the five indicators for adult mortality. Only in $_{45}l_{15}$ Cyprus performs best and Oman worst.

Table 1 Main indicators for adult mortality (both sexes) in industrialized countries, 2005–10: life expectancy at age 15 (e_{15}), life expectancy at age 60 (e_{60}), life expectancy at age 80 (e_{80}), survival probability from age 15 to 60 in percentage ($_{45}l_{15}$), and survival probability from age 60 to 80 in percentage ($_{20}l_{60}$)

	e_{15}	e_{60}	e_{80}	$_{45}l_{15}$	$_{20}l_{60}$
World	58.8	19.7	7.8	83.0	49.8
Industrialized countries	63.7	22.8	8.9	88.8	62.1
America (AMA)	64.1	22.8	9.0	89.4	63.4
Canada (CAN)	66.1	23.8	9.4	92.5	67.1
Chile (CHL)	64.4	22.8	9.3	90.4	62.1
Puerto Rico (PRI)	63.6	22.9	9.3	88.3	62.6
Trinidad and Tobago (TTO)	57.0	17.7	7.0	81.5	40.9
United States (USA)	63.9	22.8	9.0	89.1	63.2
Uruguay (URY)	62.8	21.4	8.2	89.1	58.0
Northern Europe (NEU)	65.3	23.0	8.7	92.2	64.7
Denmark (DNK)	64.0	21.9	8.4	91.0	59.8
Finland (FIN)	64.8	23.2	8.7	90.4	66.3
Ireland (IRL)	65.0	22.7	8.4	92.5	63.6
Norway (NOR)	66.0	23.4	8.7	93.2	67.1
Sweden (SWE)	66.4	23.6	8.7	93.7	68.2
United Kingdom (GBR)	65.2	22.9	8.7	92.1	64.3
Western Europe (WEU)	65.7	23.5	8.9	91.9	67.3
Austria (AUT)	65.5	23.2	8.6	92.2	66.9
Belgium (BEL)	65.0	22.9	8.5	91.4	65.5
France (FRA)	66.3	24.5	9.6	91.2	70.7
Germany (DEU)	65.2	22.9	8.5	92.0	65.1
Netherlands (NLD)	65.7	23.0	8.4	93.2	65.3
Switzerland (CHE)	67.3	24.5	9.2	93.8	71.6
Central and Eastern Europe (CEE)	55.5	18.2	6.9	75.7	44.2
Czech Republic (CZE)	62.2	20.6	7.3	89.3	55.7
Estonia (EST)	59.2	19.9	7.6	82.3	52.0
Latvia (LVA)	57.4	18.8	7.1	79.3	47.8
Lithuania (LTU)	57.0	18.8	6.9	78.5	47.5
Poland (POL)	61.1	20.5	7.7	86.1	54.4
Russian Federation (RUS)	53.3	17.3	6.6	71.4	40.2
Slovakia (SVK)	60.4	19.4	7.0	86.4	50.2
Southern Europe (SEU)	66.2	23.7	8.9	93.0	67.9
Croatia (HRV)	61.7	20.0	7.4	89.3	51.6
Greece (GRC)	65.3	22.9	8.3	92.5	65.0
Italy (ITA)	66.9	24.1	9.1	93.8	69.5
Portugal (PRT)	64.2	22.5	8.1	90.9	63.2
Slovenia (SVN)	64.0	22.2	8.3	90.2	62.0
Spain (ESP)	66.7	24.2	9.2	93.1	69.4
Western Asia (WAS)	61.9	20.5	7.8	91.7	49.4
Bahrain (BHR)	61.8	19.1	6.4	92.6	47.2
Cyprus (CYP)	64.4	21.4	7.4	93.8	58.5
Israel (ISR)	66.3	23.6	8.9	93.5	67.2
Kuwait (KWT)	59.9	17.4	4.9	91.4	38.7
Oman (OMN)	61.1	19.6	7.9	88.9	48.4
Qatar (QAT)	63.4	20.7	8.0	93.4	53.2
Saudi Arabia (SAU)	60.8	18.5	6.4	91.3	45.1
United Arab Emirates (ARE)	61.6	19.2	7.1	91.6	48.4
Oceania, Eastern and Southeast Asia (OEA)	67.3	25.0	10.1	93.3	70.5
Australia (AUS)	67.2	24.6	9.4	93.6	71.0
Hong Kong, SAR (HKG)	67.7	24.8	9.8	94.3	70.1
Japan (JPN)	68.0	25.5	10.4	93.4	72.6
New Zealand (NZL)	65.8	23.6	9.0	92.5	67.0
Republic of Korea (KOR)	65.6	23.1	8.7	92.6	64.8
Singapore (SGP)	66.5	23.7	9.7	93.6	65.2

Note: Figures for major areas are weighted by the corresponding national population sizes in age 15+, 60+, and 80+ years in 2005.
Source: United Nations, D.o.E.a.S.A., Population Division, 2013. World Population Prospects: The 2012 Revision, DVD edition. United Nations, New York; and own calculations with data of United Nations, D.o.E.a.S.A., Population Division (2013).

Country-Specific Trends in Adult Mortality

Trends in adult mortality since the mid-twentieth century were characterized by a steady increase of life expectancy in most industrialized countries. These remarkable changes were described intensively during the past decade (Oeppen and Vaupel, 2002; White, 2002; Vallin and Meslé, 2009). Nonetheless, across today's industrialized countries, these trends were heterogeneous and not in every case characterized by a continuous improvement. **Figure 2(a)** illustrates these changes for the eight considered major areas. The most regular trends in e_{15} occurred in OEA and WAS. The trends in NEU, WEU, and SEU are similarly improving since the 1970s, whereas the years before were characterized by stagnating or only mildly improving mortality conditions. AMA differentiates from these areas because improvements since the 1980s lag noticeably. In sharp contrast is the trend in CEE. Between 1950–55 and 1960–65 e_{15} increased as strong as in OEA, followed, however, by a more or less continuous decrease until 2000–05. An interruption of this decline occurred only in the second half of the 1980s when life expectancy increased for some years as consequence of Gorbachev's antialcohol campaign (Murphy, 2011). Whether the increase between 2000–05 and 2005–10 marks a sustained trend change or just another short-term interruption of the long-lasting decrease cannot be assessed to date (see Shkolnikov et al., 2013). The mean annual changes in

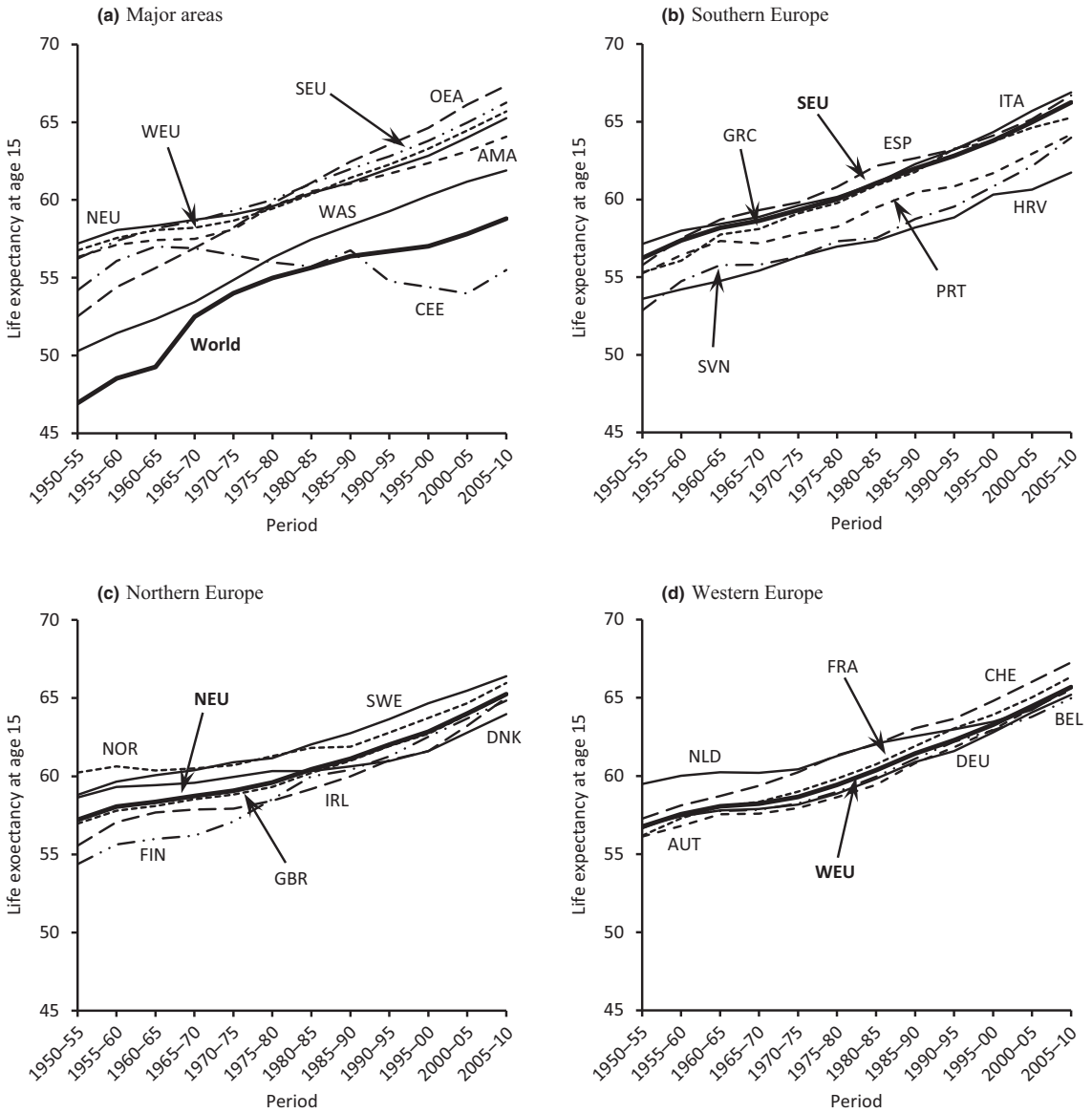

Figure 2 Trends in life expectancy at age 15 (e_{15}), both sexes, 1950–2010. Note: Figures of major areas are weighted averages based on the countries' population sizes at ages 15+ in the first year of the 5-year calendar year periods. Source: United Nations, D.o.E.a.S.A., Population Division, 2013. World Population Prospects: The 2012 Revision, DVD edition. United Nations, New York; and own calculations with data of United Nations, D.o.E.a.S.A., Population Division (2013).

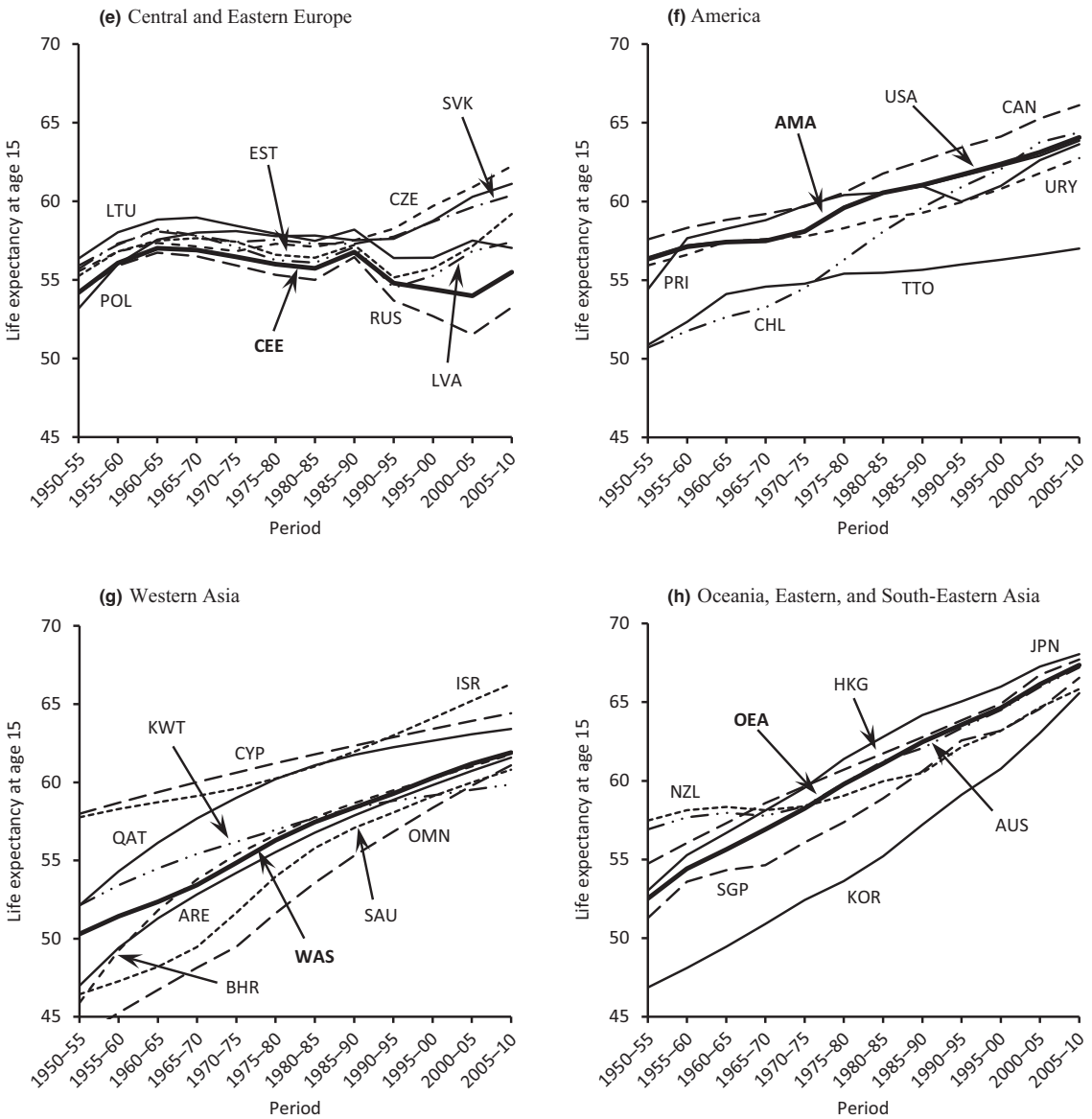

Figure 2 (*continued*).

e_{15} between 1950–55 and 2005–10 were 0.02 years in CEE, 0.14 in AMA, 0.15 in NEU, 0.16 in WEU, 0.18 in SEU, 0.21 in WAS, and 0.27 in OEA. In relative terms, life expectancy at young adult age increased during the observation time by 2.4% in CEE, 13.7% in AMA, 14.0% in NEU, 15.7% in WEU, 17.8% in SEU, 23.1% in WAS, and 28.2% in OEA.

Figure 2(b) to 2(h) provides more details about the trends in e_{15} for the individual industrialized countries. Countries from SEU, NEU, and WEU are characterized by small differences and very regular trends. Industrialized countries from SEU stand out because they show the most stable trends and differences (Figure 2(b)). In contrast, countries from NEU are characterized by a convergence of mortality levels (Figure 2(c)). Notable are the stalling mortality levels of Norway and Denmark for several decades since the 1950s. While a trend changed toward an increase similar to those in the other countries from NEU occurred in Norway in the second half of the 1980s, this happened in Denmark not before the end of the past millennium.

Industrialized countries in WEU (Figure 2(d)) also appear to converge in levels of life expectancy. However, excluding the Netherlands changes this picture. The Dutch population exhibited by far the lowest mortality levels until the early 1970s. Moreover, the overall trend in the Netherlands looks more similar to those of Norway and Denmark than to that of WEU. Switzerland stands out with the most positive trend as the Swiss population did not experience the stall in life expectancy during the 1960s, which was apparent in all other countries.

Also, the populations belonging to the industrialized countries of the CEE area follow similar trends in e_{15} until the end of the 1980s (Figure 2(e)). Their trends are characterized

by an increase of e_{15} until the mid-1960s and a subsequent decrease until the mid-1980s, which then turned into another increase. Lithuania performed best until the end of this time span and the lowest life expectancy was prevalent in Russia. The fall of the iron curtain divided the populations into two groups. In the Czech Republic, Poland, and Slovakia, the increase of e_{15} – which began in the mid-1980s – continued until the most recent period. In contrast, the countries belonging to the former Soviet Union experienced a further decline in life expectancy. Only recently, Estonia and Latvia experienced an increase of e_{15}, which parallels the trends of the Czech Republic, Poland, and Slovakia.

In AMA (Figure 2(f)), Canada exhibits the lowest mortality throughout the observation period with an almost constant advantage compared to the United States. Puerto Rico reached a similar level of e_{15} in the 1970s but then fell behind until the early 1990s when the trend changed again toward increasing life expectancy. Chile shows the most notable trends. Until the early 1970s, e_{15} was lowest among all industrialized countries in AMA, close to the level of Trinidad and Tobago. From then on, life expectancy increased stronger than in any other population.

Industrialized countries from Asia are also characterized by a strong convergence of life expectancy levels. In WAS, Cyprus had the highest life expectancy until the late 1980s and shows the most constant trend throughout the whole observation period. Qatar, Bahrain, United Arab Emirates, and Saudi Arabia exhibited sharp increases of e_{15} until the 1980s, followed, however, by a noticeable decrease in the pace of mortality improvements. While Qatar was able to reach the life expectancy level of Israel in the 1980s and being still close to Cyprus, there is a notable gap to the other Arab countries. Kuwait started on the same level of e_{15} as Qatar in the early 1950s, but the subsequent increases were smaller in all other countries and life expectancy seems to have stalled in the most recent years. Oman stands out for contrasting reasons as it shows a constant increase of e_{15} until today, however, having the lowest life expectancy until the late 1990s.

Today's industrialized countries from OEA also show heterogeneous but converging trends in life expectancy. Japan, Hong Kong, Singapore, and the Republic of Korea experienced the most regular increases. Korea started with the lowest life expectancy in 1950–55 but approached steadily the level of the other countries. Japan, today's world leader in life expectancy, showed the strongest increases until the mid-1980s when mortality improvements started to proceed with a less marked pace. Life expectancy stalled in Australia and New Zealand until the 1970s, as it can be observed in most other Western societies. Since then, e_{15} increased strongly in Australia, but did less so in New Zealand.

Trends in the Age Pattern of Adult Mortality

The two previous sections were based on the life expectancy at age 15. This indicator summarizes the whole mortality from age 15 onward. Consequently, differences in e_{15} can be caused by very different patterns of age-specific mortality. This section provides more detailed information about these age patterns by subdividing adult mortality into the age segments 15–60, 60–80, and above 80 years. Figure 3 shows for the major areas of industrialized populations the trends in ${}_{45}l_{15}$ (Figure 3(a)), e_{60} (Figure 3(b)), ${}_{20}l_{60}$ (Figure 3(c)), and e_{80} (Figure 3(d)). At a first glance, all displayed trends look similar to those for e_{15} (Figure 2(a)), indicating that the different developments of countries and regions do not reflect significant differences in trends of age-specific mortality.

Within Europe, populations from NEU perform comparatively well in young adulthood, but worse at older ages as reflected in the indicators e_{60}, ${}_{20}l_{60}$, and e_{80}. SEU and WEU are similar in all indicators of adult mortality with SEU showing lower mortality in each of them. CEE performs worse in all indicators, but especially in younger adult ages between 15 and 60 where mortality conditions are deteriorating from the 1960s through at least the early 2000s. In the other age segments, mortality in CEE is rather stagnating than decreasing.

Also, AMA shows the strongest disadvantages in younger adult ages compared to the industrialized populations from the other Western areas. This excess in mortality is apparent throughout the displayed period. In all indicators for mortality from age 60 onward populations from AMA are less disadvantaged, and they exhibited even the lowest mortality until the early 1980s. This becomes most apparent in e_{80} where AMA was around 1 year ahead of populations from Europe. In the most recent period, e_{80} was similar in AMA, WEU, and SEU; in e_{60} and ${}_{20}l_{60}$, AMA is close to the level of NEU.

The pictures for the two Asian areas of industrialized countries are very different compared to the other populations. In WAS, the largest improvements occurred in young adult ages. The mortality levels of AMA were reached during the early 1990s and those of WEU and NEU in 2005–10. In all indictors for mortality above age 60, WAS is lagging behind, and in e_{80}, it lies even below the world average in most of the displayed periods. In contrast, OEA shows strong improvements starting from levels clearly below those of AMA and the European populations and ending at the top position in all indicators. The advantages experienced during the most recent decades are gradually increasing with age and are largest in e_{80}. It appears, however, that the improvements in the highest ages start to slow down since the beginning of the new millennium. It might therefore be that mortality conditions between OEA, NEU, WEU, and SEU will converge during the coming decades. Based on the trends of the past 55 years, this seems not likely for AMA, WAS, and CEE. Note, however, that this applies only to the areas as a whole and does not exclude the fact that some individual countries will remain at or catch up to the low mortality levels apparent in the Western populations of Europe and OEA.

Contribution of the Major Causes of Death to Changes in Adult Mortality

This section further develops the description of differences between industrialized countries. Figure 4 shows age- and cause-specific contributions to the changes in life expectancy at age 15 between 1960–65 and 2005–10 for four selected countries. (The data of this section are taken from the WHO Mortality Database, online available at http://www.who.int/healthinfo/mortality_data/en/; accessed on 20.10.13.)

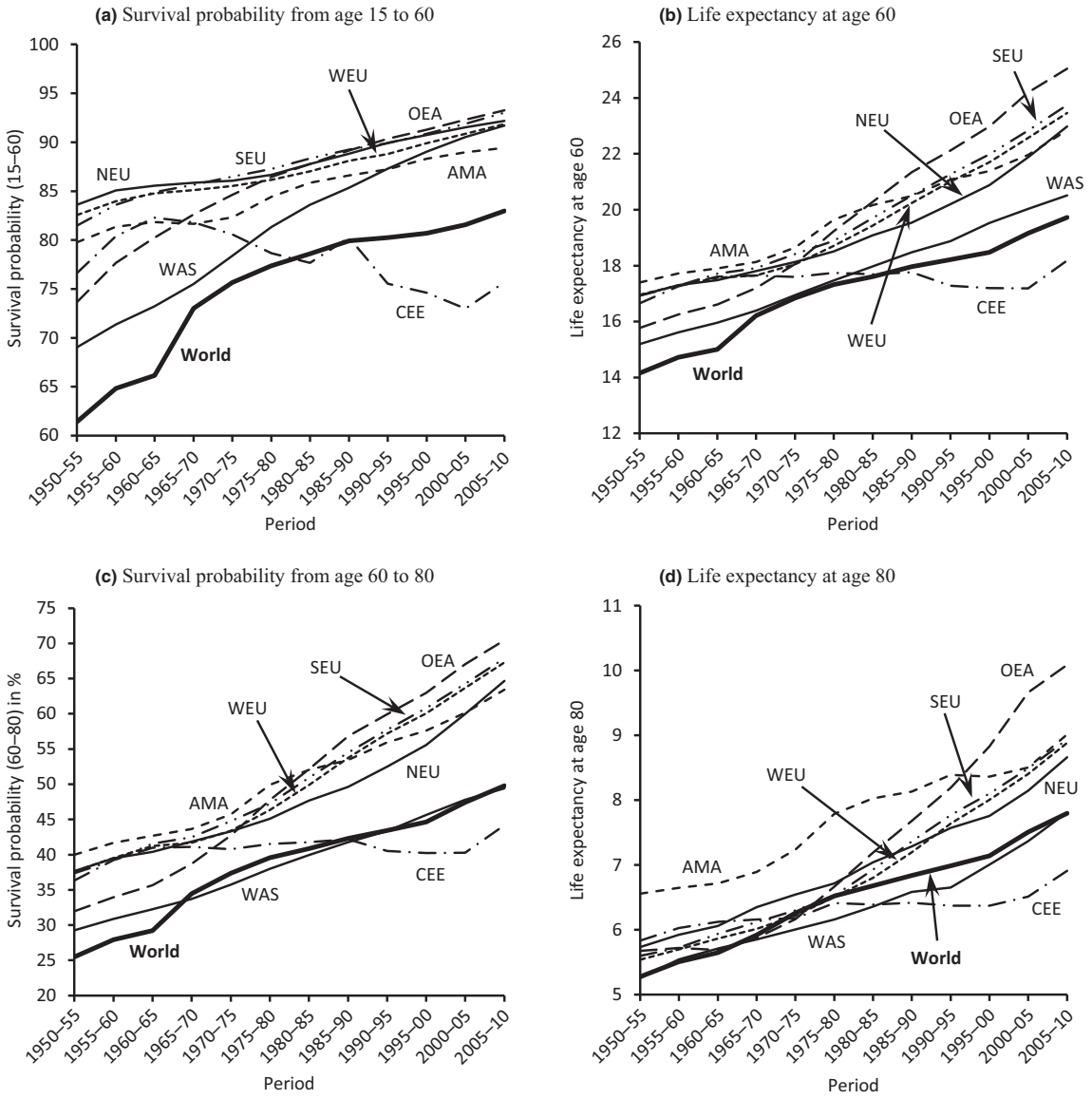

Figure 3 Trends in the survival probabilities from age 15 to 60 and from age 60 to 80 (in percentage) as well as life expectancy at age 60 (e_{60}) and at age 80 (e_{80}), both sexes, major areas, 1950–2010. Note: Figures of major areas are weighted averages based on the countries' population sizes at ages 15+, 60+, and 80+ in the first year of the 5-year calendar year periods. Source: Own calculations with data of United Nations, D.o.E.a.S.A., Population Division, 2013. World Population Prospects: The 2012 Revision, DVD edition. United Nations, New York.

The gains in e_{15} between 1960–65 and 2005–10 were 12.6 years in Singapore, 8.6 years in Italy, 6.6 years in the United States, and 3.4 years in Poland. These gains are decomposed into the contributions of 5-year age groups between ages 15 and 84 and the group aged older than 85 years. These age-specific contributions are further subdivided into the fractions caused by six major causes of death categories – diseases of the circulatory system, malignant neoplasm, diseases of the respiratory system, diseases of the digestive system, infectious diseases, and external causes of death – and a rest group including all other causes. Naturally, the impact of causes of death on adult mortality differs with age. Since in some cases cause-specific changes were unfavorable, some of the bars lie in the negative area, meaning that these specific causes resulted in a loss of life years between 1960–65 and 2005–10. This is true, for instance, for malignant neoplasms in the highest ages in all displayed populations.

Figure 4 reveals some similarities as well as significant differences in the age- and cause-specific changes in adult mortality in the four displayed populations. All have in common the fact that the largest contributions to gains in e_{15} originate close to the retirement age. The modes are located at age 60–64 in Singapore, 65–69 in the United States, 70–74 in Poland, and 75–79 in Italy. The smallest contributions come from the ages 15–49, which caused e_{15} to increase between 1960–65 and 2005–10 by 0.8 years in Poland (23.5% of the total increase in e_{15}), 1.1 years in the United States (16.6%), 1.6 years in Italy (18.6%), and 2.5 years in Singapore (19.8%).

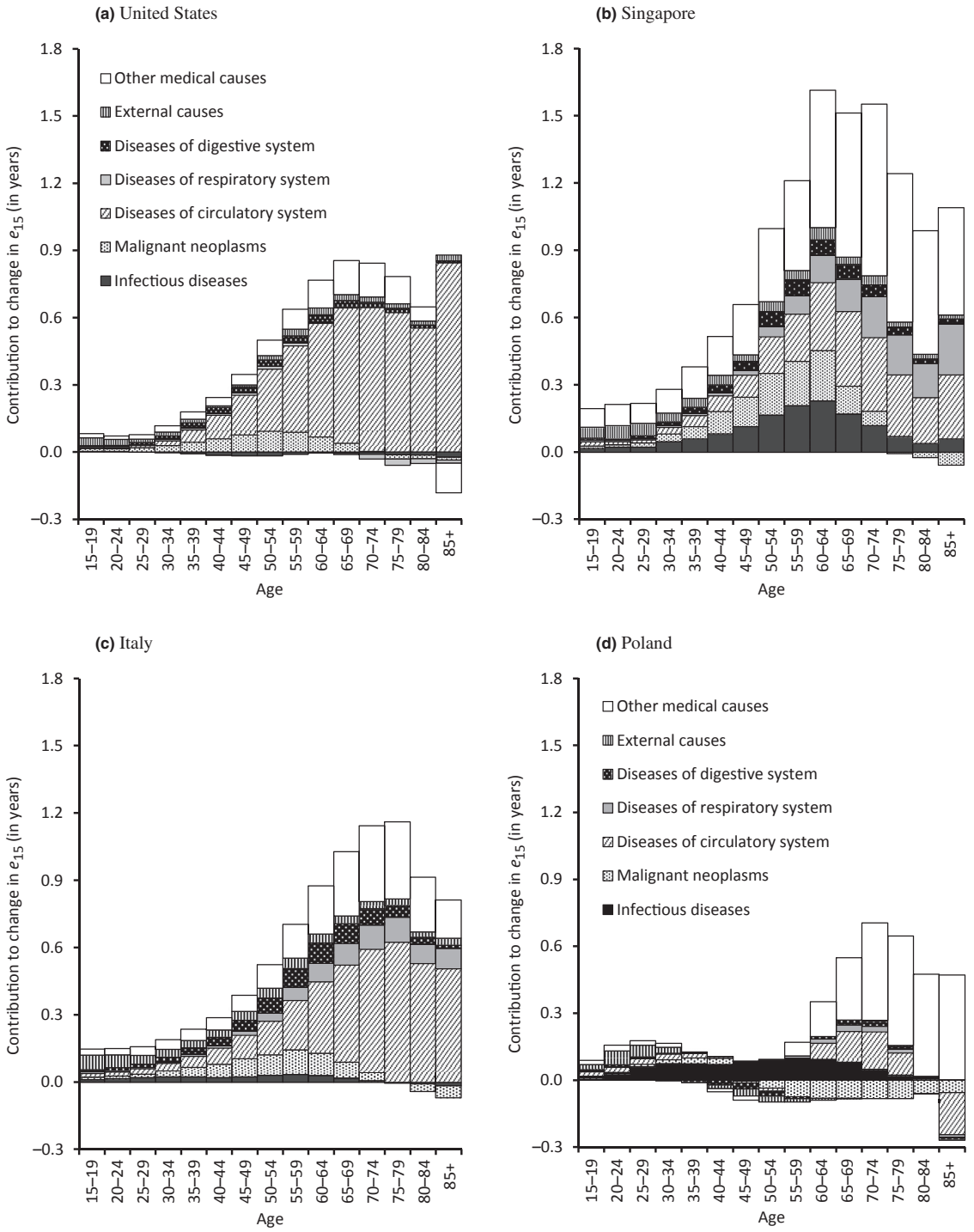

Figure 4 Age- and cause-specific contributions to changes in life expectancy at age 15 (e_{15}) between 1960–65 and 2005–10 and major causes of death, selected countries. Source: Own calculations with data of WHO Mortality Database.

The most regular age pattern in contributions to the gain in e_{15} is prevalent in Italy with a steady increase until the mode and an incremental decrease thereafter. In the United States and Singapore, the group aged older than 85 years contributed stronger than that aged 80–84 years. Poland differs strongest from the others, reflecting to some extent the typical Eastern European pattern of adult mortality. This is characterized by high and stagnating – in some countries even increasing – mortality in some age groups.

In general, the cause of death patterns of mortality changes differ more notably between the four populations than the age patterns. In the United States, the 6.6 years gain in e_{15} between

1960–65 and 2005–10 was mainly due to diseases of the circulatory system (4.8 years or 72.9%). Malignant neoplasms contributed 0.5 years (7.3%), external causes of death 0.3 years (4.9%), diseases of the digestive system 0.3 years (4.4%), and other medical causes 0.8 years (12.6%). The positive and negative contributions of diseases of the respiratory system canceled each other out, and infectious diseases caused a slight decrease of e_{15} amounting to 0.2 years (−2.2%). In Singapore, all cause of death categories contributed positively to the 12.6 years gain in e_{15}: diseases of the circulatory system 2.4 years (19.0%), infectious diseases 1.4 years (11.4%), diseases of the respiratory system 1.2 years (9.7%), malignant neoplasms 1.1 years (8.5%), external causes 0.6 years (4.7%), and diseases of the digestive system 0.5 years (4.2%). The largest contribution originates by the sum of all other medical causes, which contributed 5.4 years (42.5%) to the increases in e_{15}.

It is important to note that this rest category contains ill-defined deaths and thus some part of the other medical causes would belong actually to the main cause of death categories. Ill-defined deaths have several reasons, which hampers the comparability of international comparisons and their extent reflects the quality of the statistical systems (Mathers et al., 2005). The Italian registration system seems therefore less deficient as other causes of death contributed only 25.2% (2.2 years) to the gains of young adults' life expectancy. Diseases of the circulatory system contributed 3.6 years (42.3%) to the total gain in e_{15} of 8.6 years, diseases of the respiratory system 0.8 years (8.7%), diseases of the digestive system 0.7 years (7.6%), external causes 0.6 years (6.8%), malignant neoplasms 0.6 years (6.6%), and infectious diseases 0.2 years (2.8%). Thus, in Italy also all cause of death categories contributed positively to the changes in e_{15}. This is not the case in Poland where malignant neoplasms reduced life expectancy by 0.5 years (13.7%). Also, diseases of the digestive system caused e_{15} to decrease slightly by 0.02 years (0.6%). The increase of e_{15} by 3.4 years between 1960–65 and 2005–10 was mainly due to the sum of other medical causes containing ill-defined deaths (2.4 years or 71.8%). Improvements in infectious diseases increased e_{15} by 0.8 years (24.9%), diseases of the circulatory system by 0.4 years (11.6%), and diseases of the respiratory system and external causes each by 0.1 years (3.7 and 2.4%, respectively).

Trends and Patterns of Sex Differences in Adult Mortality

All previous sections analyzed mortality conditions and trends for total populations without separating by sex. However, women and men of industrialized populations differ significantly in life expectancy as well as age-specific mortality. Across the populations of today's industrialized countries, women's life expectancy at age 15 exceeded those of men by 6.4 years in 2005–10. The sex differences vary greatly between countries, ranging from 1.3 years in Bahrain to 12.6 years in Russia. Figure 5(a) shows the trends of female–male differences in e_{15} for the major areas of industrialized countries and the entire world population from 1950–55 to 2005–10. Most of these area groups exhibit the basic pattern of a reversed U-shape with a regular increase of sex differences until a turning point from when on the differences decrease in a similar steady manner. The trend reversal occurred first in AMA and NEU at the end of the 1970s, followed by SEU and WEU at the end of the 1990s, and then in OEA and CEE at the beginning of the new millennium. In the OEA and CEE, the trend of increasing sex

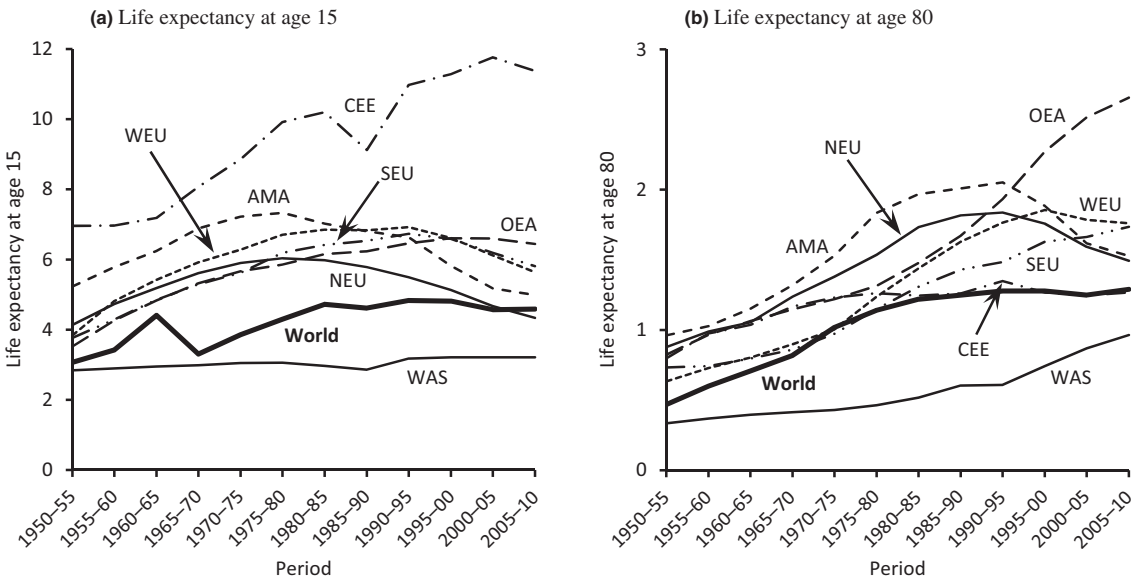

Figure 5 Trends in sex differences in life expectancy at age 15 (e_{15}) and at age 80 (e_{80}), 1950–2010. Note: Figures of major areas are weighted averages based on the countries' population sizes at ages 15+ and 80+ in the first year of the 5-year calendar year periods. Source: Own calculations with data of United Nations, D.o.E.a.S.A., Population Division, 2013. World Population Prospects: The 2012 Revision, DVD edition. United Nations, New York.

differences was shortly interrupted during the late 1980s when the gap decreased as a consequence of periodically decreasing – mainly male – mortality due to external causes of death (Leon et al., 1997). In WAS, sex differences in e_{15} increased only marginal throughout the displayed period without any change in the basic trend. The sex gap was 2.8 years in 1950–55 and 3.2 years in 2005–10.

More notable than the differences in the trend are the differences in the extent of the sex gap in life expectancy. The smallest gap in e_{15} is apparent in WAS. In contrast, male excess mortality is largest in CEE where women outlived men by 7.0 years in 1950–55 and 11.4 years in 2005–10. The other areas lie between these extremes and show less differences in the extent of the sex gap. From the 1950s until the early 1980s, sex differences were higher in AMA than in the Western European populations and OEA. In 2005–10, OEA shows the highest differences of these areas as a consequence of the late peak in the trend of the sex gap.

With regard to e_{80} there are two notable differences (see **Figure 5(b)**). The first concerns CEE where the male disadvantage is much smaller. Since the late 1970s, sex differences in e_{80} are even the lowest among the European areas, being on the level of sex differences of the total world population. This reveals that the extraordinary high male excess mortality in e_{15} is predominantly caused by the high mortality of men in young and middle adult ages. The second significant difference between sex differences in young and old adult ages is prevailing in OEA where the recently observable trend reversal in e_{15} is not visible in e_{80}. In contrast, sex differences at highest ages are sharply increasing even in the most recent observation period.

Within the major areas, the individual industrialized populations show similar trends in the sex gap in e_{15} (data not shown, but available on request from the author). AMA is characterized by a difference between populations from North and South America with the trend change occurring later in the latter. In 2005–10, the sex gap in e_{15} ranged between 4.6 years in Canada and 7.9 years in Puerto Rico. In CEE, the former member states of the Soviet Union show identical 'roller-coaster' trends with similar extents of the sex gap, while sex differences in the populations from Central Europe follow a regular pattern with the peak occurring in the early 1990s. Among these populations, the lowest gap is currently apparent in the Czech Republic with a female surplus of 6.2 years. The single Western European countries follow basically the trends of the defined geographical areas. In the most recent period, 2005–10, sex differences in e_{15} range in WEU from 4.1 years in the Netherlands to 6.9 years in France, in SEU from 5.0 years in Greece to 7.0 years in Slovenia, and in NEU from 4.1 years in Sweden to 6.8 years in Finland.

The areas OEA and WAS show the most heterogeneous trends between countries. Australia, New Zealand, and the Republic of Korea exhibit trends in the sex gap similar to the Western populations of Europe and AMA with the peak occurring in the late 1970s and the early 1980s. However, the extents differ markedly. In 1950–55, sex differences in e_{15} were 3.1 years in the Republic of Korea, 3.9 years in New Zealand, and 5.1 years in Australia. At the time of the peak, the differences amounted to 6.2 years in New Zealand, 6.8 years in Australia, and 8.8 years in the Republic of Korea. Until 2005–10, the sex gap decreased to 4.0 years in New Zealand, 4.4 years in Australia, and 6.7 years in the Republic of Korea. In Japan, sex differences increased continuously until the beginning of the new millennium and started to decrease just recently. The gap was 3.3 years in 1950–55, 6.9 years in 2000–05, and 6.8 years in 2005–10. Very different patterns are apparent in Hong Kong and Singapore. While the former shows the reversed Japanese pattern with the sex difference being 7.4 years in 1950–55, 5.7 years in 2000–05, and 6.1 years in 2005–10, the sex gap developed irregularly in the latter. The female surplus in e_{15} increased from 5.6 years in 1950–55 to 6.3 years 1970–75, then decreased to 4.3 years in 1985–90, and finally increased again to 5.0 years in 2000–05 and 2005–10.

Populations from WAS show even stronger differences in trends of the sex gap. In Saudi Arabia and Cyprus, sex differences in e_{15} increased constantly but slowly from 3.1 and 3.9 years in 1950–55 to 3.5 and 4.1 years in 2005–10, respectively. Israel and Oman exhibit constant but strong increases from 1.4 and 2.3 years in 1950–55, respectively, to 4.0 years in 2000–05. While in Israel the sex gap decreased in the subsequent quinquennial 2005–10 to 3.9 years, it increased further in Oman to 4.2 years. In Kuwait, the sex gap increased between 1950–55 and 1970–75 from 2.0 to 2.5 years, then decreased to 1.7 years in 1990–95 from when on it remained constant until 2005–10. In the remaining countries, the trends in sex differences are characterized by a strong decrease over the observation period. In Bahrain, the female advantage in e_{15} decreased steadily from 4.5 years in 1950–55 to 1.3 years in 2005–10; in the United Arab Emirates, it decreased over the same time from 4.0 to 1.8 years with a recent leveling off tendency; and in Qatar, sex differences decreased from 4.1 years in 1950–55 to 1.2 years in 1985–89, and then they increased again to 1.7 years in 2005–10.

Driving Factors of Past Trends in Adult Mortality and Expectations for the Future

The descriptive comparative trends presented for adult mortality in today's industrialized populations are illustrative. This last section summarizes an overview of driving factors of trends in mortality in low-mortality countries elaborated by Caselli et al. (2014) where more details can be found. Generally spoken, levels and trends of mortality in today's industrialized populations as well as the prevailing driving factors reflect to a large extent the development stages of the populations. For the early developed populations from North America, Oceania, and Europe, trends in life expectancy for the overall and sex-specific populations are most likely due to health behaviors (predominantly tobacco consumption and – above all in CEE – alcohol consumption and violence), health conditions (especially obesity), and structural factors including socioeconomic position, health care access, and health insurance. For instance, compared with Canada, the United States has higher prevalence rates of smoking, obesity, and violence, along with limited access to health care. Moreover, the United States has had a longer history of cigarette consumption, especially among men, which has contributed to lower male life expectancies and larger sex gaps compared with Canada (Pampel, 2005). Smoking-related mortality has also

been shown to be responsible for the less favorable developments in life expectancy in Denmark, the Netherlands, and the United States (Christensen et al., 2010; Preston et al., 2009).

It has been demonstrated that the trend in smoking prevalence follows the spread pattern of an epidemic and involves a diffusion process across socioeconomic strata (Lopez et al., 1994). Smoking starts among groups of high socioeconomic status and then spreads to the rest of the population. Similarly, the decline in smoking prevalence begins among groups of high socioeconomic status, who are typically the first to become concerned with the harmful effects of smoking. These processes can be observed among both sexes, with higher prevalence rates in men and a time lag of two to three decades between men and women. Although the early developed populations are today in later stages of the smoking epidemic, they are still not at the end of this process (Ng et al., 2014). Thus, future trends in life expectancy are likely to be determined by trends in tobacco consumption. For instance, Wang and Preston (2009) predicted for the US population that the changes toward lower smoking prevalence will lead to a faster mortality decline than anticipated by most experts (see also Bongaarts, 2006; King and Soneji, 2011). In many industrialized countries, national health policies are currently conducting intensive antismoking campaigns. Evidence suggests that these measures are effective and thus can be expected to lead to further decreases of smoking rates (e.g., Ádám et al., 2012). The later developed populations among today's industrialized countries are still in earlier stages of the smoking epidemic. It is therefore possible that their future trends in adult mortality will be influenced by increasing smoking prevalence for some decades as observed in the earlier developed populations among men until the 1980s and among women in the more recent past.

Nonetheless, smoking might be the most important driver of recent trends in adult mortality, but its impact differs between populations and it is not the only factor with significant impact on life expectancy levels. How these factors are linked with each other and how they will influence future mortality is heavily debated among demographers, epidemiologists, and researchers from other public health areas. In recent years, two schools of thought have developed that dominate the discussion about the most likely trends of future life expectancy whose supporters are generally referred to as 'optimists' and 'pessimists' (for the most recent dispute see Olshansky and Carnes, 2013; Vaupel and Rau, 2013). Both agree that mortality will continue to decrease, lifespan will increase, and life expectancy records will continue to be broken. Thus, the difference of opinion between the two schools about likely future trends in life expectancy is in fact smaller than generally presumed (Bongaarts, 2006; Wilmoth, 2001). They differ, however, in their expectations regarding the time course and magnitude of these improvements. Today, the majority of researchers tend to follow the optimistic perspective along with the expectation that the still existing differences in mortality between countries are likely to narrow further in the future (see Caselli et al., 2014).

See also: Adult Mortality in Developing Countries; Adult Mortality in Eastern Europe and the Former Soviet Republics; Aging and Health in Old Age; Early Life Influences on Health and Mortality in Adulthood; Infant and Child Mortality in Industrialized Countries; Life Table; Mortality Differentials: Selection and Causation; Mortality of the Oldest-old; Mortality: The Great Historical Decline; Smoking and Health; Socioeconomic Status and Health.

Bibliography

Ádám, B., Molnár, Á., Gulis, G., Ádány, R., 2012. Integrating a quantitative risk appraisal in a health impact assessment: analysis of the novel smoke-free policy in Hungary. European Journal of Public Health 23, 211–217.

Bongaarts, J., 2006. How long will we live? Population and Development Review 32, 605–628.

Caselli, G., Drefahl, S., Wegner-Siegmundt, C., Luy, M., 2014. Future mortality in low mortality countries. In: Lutz, W., Butz, W.P., KC, S. (Eds.), World Population and Human Capital in the Twenty-First Century. Oxford University Press, London.

Christensen, K., Davidsen, M., Juel, K., Mortensen, L., Rau, R., Vaupel, J.W., 2010. The divergent life-expectancy trends in Denmark and Sweden – and some potential explanations. In: Crimmins, E.M., Preston, S.H., Cohen, B. (Eds.), International Differences in Mortality at Older Ages. National Academic Press, Washington, DC.

King, G., Soneji, S., 2011. The future of death in America. Demographic Research 25, 1–38.

Leon, D.A., Chenet, L., Shkolnikov, V.M., Zakharov, S., Shapiro, J., Rakhmanova, G., Vassin, S., Mckee, M., 1997. Huge variation in Russian mortality rates 1984–94: artefact, alcohol, or what? Lancet 350, 383–388.

Lopez, A.D., Collishaw, N.E., Piha, T., 1994. A descriptive model of the cigarette epidemic in developed countries. Tobacco Control 3, 242–247.

Mathers, C.D., Ma Fat, D., Inoue, M., Rao, C., Lopez, A.D., 2005. Counting the dead and what they died from: an assessment of the global status of cause of death data. Bulletin of the World Health Organization 83, 171–177.

Murphy, M., 2011. Adult mortality in the former Soviet Union. In: Rogers, R.G., Crimmins, E.M. (Eds.), International Handbook of Adult Mortality. Springer, Dordrecht, Heidelberg, London, New York.

Ng, M., Freeman, M.K., Fleming, T.D., et al., 2014. Smoking prevalence and cigarette consumption in 187 countries, 1980–2012. Journal of the American Medical Association 311, 183–192.

Oeppen, J., Vaupel, J.W., 2002. Broken limits to life expectancy. Science 296, 1029–1031.

Olshansky, S.J., Carnes, B.A., 2013. Zeno's paradox of immortality. Gerontology 59, 85–92.

Pampel, F.C., 2005. Forecasting sex differences in mortality in high income nations: the contribution of smoking. Demographic Research 13, 455–484.

Preston, S.H., Glei, D.A., Wilmoth, J.R., 2009. A new method for estimating smoking-attributable mortality in high-income countries. International Journal of Epidemiology 39, 430–438.

Luy, M., Wegner, C., Lutz, W., 2011. Adult mortality in Europe. In: Rogers, R.G., Crimmins, E.M. (Eds.), International Handbook of Adult Mortality. Springer, Dordrecht, Heidelberg, London, New York.

Shkolnikov, V.M., Andreev, E.M., Mckee, M., Leon, D.A., 2013. Components and possible determinants of the decrease in Russian mortality in 2004–2010. Demographic Research 28, 917–950.

United Nations, D.o.E.a.S.A., Population Division, 2013. World Population Prospects: The 2012 Revision, DVD edition. United Nations, New York.

Vallin, J., Meslé, F., 2009. The segmented trend line of highest life expectancies. Population and Development Review 35, 159–187.

Vaupel, J.W., Rau, R., 2013. Research versus rhetoric. Gerontology 59, 95–96.

Wang, H., Preston, S.H., 2009. Forecasting United States mortality using cohort smoking histories. Proceedings of the National Academy of Sciences 106, 393–398.

White, K.M., 2002. Longevity advances in high-income countries, 1955–96. Population and Development Review 28, 59–76.

Wilmoth, J.R., 2001. How long can we live? A review essay. Population and Development Review 27, 791–800.

Adulthood: Dependency and Autonomy

Hans-Werner Wahl, Institute of Psychology, Heidelberg University, Heidelberg, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H.-W. Wahl, volume 1, pp. 151–153, © 2001, Elsevier Ltd.

Abstract

Dependency means the ongoing need for external support, but also reflects, to a lesser extent, affiliation, attachment, and bonding to significant others. Autonomy can be defined as a state in which the person is, or feels, capable of pursuing life goals by use of his or her own resources. Based on these definitions, this article discusses the complexities inherent to these terms. Considerations of developmental tasks in middle adulthood lead to the insight that dependency and autonomy normally have fused into interdependence in this life period. Distinctions between structured, physical, and behavioral dependency are used to address the dependency–autonomy dialectic in old age. New research challenges are finally addressed.

Dependency strictly means the ongoing need for external support (e.g., from family members, professionals, state institutions, intensive care units, or assistive devices) in order to fulfill individual or societal expectations regarding what is a 'normal' life. A less strict interpretation of dependency also encompasses human needs for affiliation, attachment, and bonding to significant others, such as to one's partner, children, grandchildren, or close friends (Baltes and Silverberg, 1994). Autonomy can be defined as "a state in which the person is, or feels, capable of pursuing life goals by the use of his or her own resources" (Parmelee and Lawton, 1990: p. 465). Autonomy thus means independent and effective functioning in a variety of life domains ranging from basic activities of daily living to expanded leisure activities, and complex decision-making processes. Constructs related to autonomy are *environmental mastery*, a term suggested by Ryff (1989) as an indication of successful aging, having the feeling that one is the origin of one's own actions (frequently referred to as *primary control*, e.g., Heckhausen et al., 2010), or possessing *self-efficacy* (Bandura, 1977).

Historically, developmental researchers have primarily examined the dynamics between dependency and autonomy from childhood to adolescence. In old age, dependency and loss of autonomy were long viewed as direct consequences of aging worthy of systematic description (e.g., in epidemiological studies) but basically inevitable and irrevocable. It was only in the 1970s that new research findings – based on a more optimistic image of old age and stimulated by a social learning perspective – demonstrated the plasticity of dependency and autonomy in old age (Baltes, 1973). While research in the 1980s and 1990s significantly furthered this insight, middle adulthood, and also the transitions in dependency and autonomy that are experienced as the adult individual matures, attracted so far little scientific interest (Willis and Reid, 1999).

On the Complexities of Dependency and Autonomy in Adult Life

From a life-span perspective, childhood and adolescence are periods when striving toward autonomy and reducing dependency are among the most important *developmental tasks* (Havighurst, 1972). By young adulthood, or at the very latest by middle adulthood, one is normally expected to have accomplished this successfully. Conversely, old age may be characterized, at least to some extent, as a life period that poses the risk of becoming dependent or losing one's autonomy. However, the general assumption that autonomy gradually replaces dependency and then dependency gradually replaces autonomy over the life course is clearly simplistic.

Cultural relativity becomes particularly obvious in the autonomy–dependency dynamics across the life span. For example, while the developmental goal of maintaining autonomy in a wide variety of life domains over the life span is one of the highest values in most Western cultures, one of the most 'normal' elements of many developing countries' or Asian cultures is reliance on children in the later phases of life. Second, although autonomy and dependency play their roles as individual attributes, both should be regarded predominantly as contextual constructs depending strongly on situational options and constraints. For example, the dependent self-care behavior of an 85-year-old man or woman may not reflect physical or mental frailty at all, but may primarily result from the overprotective behavior of family members and professionals (Baltes, 1996). Third, autonomy and dependency should both be regarded as multidimensional, that is, gain in autonomy in one life domain does not automatically lead to reduced dependency in other life domains and vice versa. For example, being able to meet the everyday challenges of life in an independent manner does not necessarily prevent a younger individual from relying strongly on parents or significant others when making a crucial life decision (such as selecting a partner). Fourth, and finally, autonomy and dependency have strong value connotations which shape action. In Western cultures, independent behaviors are generally regarded as positive and highly adaptive, worth supporting by all means, whereas dependency has negative value connotations and should be avoided at all costs. Such global value attributions can be questioned in terms of life complexity and richness. For example, emotional dependency upon another person lies at the heart of mature intimate relationships and may indeed be seen as an important developmental gain associated with aging. Laura Carstensen (2006) argues in her socioemotional selectivity theory that as we move along the life span, we

increasingly invest into bonding processes with intimate others and that such behavior echoes an important motivational shift in the second as compared to the first half of life. Conversely, striving for autonomy may become detrimental when confronted with severe chronic illness, which necessitates help, support, and the delegation of control to the external environment. These differentiations have to be kept in mind as we examine autonomy and dependency in middle and old age more closely.

Dependency and Autonomy – A Challenge in Midlife?

Answers to this question must start with the developmental challenges associated with the life phase that occurs between roughly 45 and 65 years of age. According to life-span theorist Erik Erikson (1950), one central challenge of human midlife is *generativity*, that is, support of and solidarity with the following generations not only in terms of concrete child-rearing activities, but also in the broader sense of transfer of one's own life expertise and experiences to others. Coping with the 'empty nest,' establishing bonds to grown-up children, and becoming a grandparent are other typical challenges to be met in this life phase. With respect to women, the experience of menopause is the most typically age-graded event in this life period, while one of the most challenging nonnormative events of modern middle-aged women (very much less so for men) is taking over the care-giving role for one's father or mother.

What kinds of implications do these midlife challenges have with respect to autonomy and dependency? As Baltes and Silverberg (1994) have persuasively argued, the term *interdependence* best encompasses agency and autonomy *and* attachment and bonding needs in the middle years. Interdependence typically is at work in generativity, where the older generation is dependent on the younger generation in order to feel needed (particularly after children have left home). Generativity particularly means the strong motivation of older adults to teach and mentor younger generations (Erikson, 1950). At the same time, interdependence can strengthen feelings of agency, competence, and autonomy by providing a forum for the transfer of advice, expertise, and 'world knowledge.' The new role of being a grandfather or grandmother further adds to interdependence, particularly in women. As has been argued (Hagestad, 1985), women more easily express interdependence and attachment and thus contribute more to the maintenance of family and kinship. Furthermore, interdependence finds expression in social relations in midlife by helping others and receiving help, i.e., by mutuality and reciprocity, thus contributing to the establishment of a *social convoy* which guarantees social support throughout life (Antonucci, 2001). However, providing care for a chronically ill parent may also negatively impact on well-balanced interdependence: the moral obligation to care for a family member, and also the loss of autonomy one endures in order to provide such care, can lead to severe psychological distress in midlife (Pinquart and Sörensen, 2003).

Driven by the idea of developmental tasks (Havighurst, 1972), autonomy issues in midlife also crystallize in major decision-making processes. One such critical decision important for future personal development is when and how the retirement process is initiated (Löckenhoff, 2012). As has been found, autonomously facing and preparing this decision in a timely manner seems to be associated with how people later adapt to the postretirement phase.

On the Many Faces of Dependency and Autonomy in Old Age

With respect to later life, Margret M. Baltes (1996) has introduced distinctions among structured, physical, and behavioral dependency. *Structured dependency* implies that human worth is determined primarily by participation in the labor force. According to this understanding, society is in a sense producing dependency among certain subgroups (such as the aged) in order to provide opportunity structures for other subgroups (such as younger persons). *Physical dependency* is closely linked to the age-related emergence of chronic physical and mental impairments such as severe mobility loss or Alzheimer's disease. *Behavioral dependency* primarily reflects the effect of the interaction between the older person and his/her social environment, thus emphasizing most strongly of all three forms of dependency, the role of social and behavioral processes in order to explain dependency in the later years. Owing to the understanding of dependency in social and behavioral terms, a whole empirical research program has been conducted by M. Baltes and associates (for an overview, see Baltes, 1996). The major and very robust finding was that social interactions between older adults and social partners are prone to a *dependence–support script*, that is, dependent behavior among elders is reinforced by their social partners (typically by direct helping behavior), whereas autonomous behaviors are typically ignored.

Furthermore, structured, physical, and behavioral dependency interact with each other, that is, macro-social, biological, and psychological processes must be considered in conjunction. Three theoretical approaches have been discussed to address this dynamic interplay and its outcomes. First, dependency in old age may be due in part to *learned helplessness*, that is, an existing noncontingency between action goals and consequences (Seligman, 1975). Macro-social (see again the term structured dependency) and noncontrollable biological impairments may contribute to dependency in this vein. Second, dependency may also be interpreted as a strong instrument for *passive control*. In this regard, the dependence–support script found in the Baltes research also reflects the power of dependent behaviors of elders to provoke positive responses (typically help) from their social environment and thus maintain environmental control as long as possible. The adaptational value of this strategy is particularly suggested by the *life-span theory of control* (Heckhausen et al., 2010). Third, an even more complex interpretation of the dynamics between dependency and autonomy in old age comes into the play when adaptation and the striving for a successful course of aging are viewed in terms of *selective optimization with compensation* (Baltes et al., 2006). In this respect, dependent behaviors in certain life areas (such as basic care needs) may be seen as a powerful compensation tool in the late phase of the human life span in order to prepare the field for further optimization and

development in selected other life domains with high personal priority.

In addition, as seen from a personality developmental view, findings from now available longitudinal research support the view that it would be an oversimplification to assume that dependency is decreasing in all respects. As has for instance been found, dependency-related modes of coping (such as displacement and regression) are decreasing with age, while autonomy related modes (such as intellectualization) are indeed increasing (Diehl et al., 2014). In a sense, thus, although bodily strengths and biological functioning are clearly on the fall as we age, more inner-psychic forms of autonomy may increase. This is also echoed in classic theories of life-span development, such as in Erikson's (1950) concept of ego–integrity in old age as well as more recently in the theory of gerotranscendence (Tornstam, 2005).

Among the major threats of autonomy in old and very old age are dementia-related disorders, that is, pathological impairment of cognitive abilities and the human capacity to exert executive control (Wettstein et al., 2012). In fact, losing one's autonomy in physical and mental functioning and becoming strongly dependent on others is the behavioral core of dementia. But even regarding normative cognitive development in later life, it is important to note that cognitive performance is among the major predictors of autonomous functioning in everyday life (Diehl and Willis, 2004). It is also noteworthy that keeping autonomy in basic activities of daily life (e.g., self-care, meal preparing) is predominantly associated with physical and cognitive functioning, whereas autonomy in expanded activities (e.g., cultural and educational activities, social activities) is more strongly related with personality differences (Baltes et al., 1999). Another major threat to autonomy in later life is chronic vision impairment, affecting about 20% of those who are 65 years and older and about 30% of those who are 80 years and older (Wahl, 2013).

Among the important decisions to be made in old age are residential decisions. Residential decisions are closely interrelated with autonomy and dependency, because different forms of living arrangements may very differently impact on autonomy and dependency. For example, deciding to invest into one's apartment/house in terms of home modification may significantly extend the time period to 'age in place' (Wahl et al., 2009). Aging in place at the same time is a major symbolic clue for many older adults that they 'still' are autonomously operating individuals. On the other hand, moving to a long-term care institution may at the first glance speak to giving up autonomy and risk dependency, but may also echo an important selection, compensation, and optimization strategy with the potential to unfold new developmental pathways for an aging individual (Baltes et al., 2006).

Overall, it seems to be among the major challenges of old and very old people to deal with an ongoing *dialectic* between autonomy and dependency. On the one hand, important resources to maintain autonomy, such as physical functioning, are shrinking in old age, while others operating more in the inner-psychic domain are indeed possibly increasing. At the same time, older adults are becoming more dependent on others, but part of such dependency, e.g., in the emotional realm, seems to be highly adaptive. Such a view is also supported by findings of the Berlin Aging Study that inform the view that the human phase of very old age is characterized by relatively stable maintaining of internal control, while increase in external control beliefs has been observed in parallel (Smith and Baltes, 1999).

In terms of cohort dynamics and new developments, it is also interesting to see that the concepts of autonomy and dependency may find new variations. For example, technology such as the Internet and robotics increasingly is serving older adults toward maintaining their autonomy and indeed a new understanding of what it means to be autonomous may appear based on the role which technology is going to play in future aging (Wahl et al., 2012). At the same time, such increase in the use of technology may also lead to new forms of dependency, for example, when technology fails and by this means elicits new risks.

Conclusions and New Research Directions

The dynamic interplay between dependency and autonomy occurs in different variations across the adult human life span. While the creation and regulation of interdependence is among the major tasks of middle adulthood, the particularities of old age call for the consideration of different pathways to explain the etiology and maintenance of dependency in old age. The insight of the behavioral model of dependency, pointing to the different roles that dependency can take as an adaptive tool in old age, also means a challenge for intervention. Owing to the potential role of dependent behaviors for successful aging, intervention efforts should always be framed within a change philosophy that leaves the older person in control of whether external autonomy-enhancing efforts should be exerted or not. Autonomy in old age, in this sense, may also mean the autonomy to decide for dependence in some life domains, even if this might mean the risk of losing available competencies in the long run by *disuse*.

New research directions should follow at least four avenues. First, research should examine forms of interdependency in midlife, with a view to understanding how they affect dependency in old age. Second, the interaction between social and physical environmental influences and personality traits, and also their effect on fostering dependency or maintaining autonomy, deserves more consideration. Third, more knowledge is needed on the day-to-day balancing and rebalancing of dependency and autonomy and their emotional and behavioral outcomes. For this kind of empirical research, the model of selective optimization with compensation (SOK; e.g., Baltes et al., 2006) and the life-span theory of control (Heckhausen et al., 2010) are probably the best theoretical alternatives to date. The SOK model assumes that an important process of developmental regulation is selection; in addition, compensation is needed in case of reduced competencies and blocked goals. If selection and compensation are orchestrated well, then also optimization is possible at any stage of the life span. The life-span theory of control assumes that more secondary control execution, i.e., compensatory control processes such as replacing no longer achievable goals, is adaptive as we age. Fourth, exploring new forms of autonomy and dependency, particularly those

connected with technology, is an intriguing future research area (Wahl et al., 2012).

See also: Aging and Health in Old Age; Dementia: Psychiatric Aspects; Erikson's Theory of Psychosocial Development; Human Development, Successful: Psychological Conceptions; Self-Efficacy; Self-Regulation in Adulthood; Transition to Adulthood.

Bibliography

Antonucci, T.C., 2001. Social relations. An examination of social networks, social support and sense of control. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, fifth ed. Academic Press, San Diego, CA, pp. 427–453.

Baltes, M.M., 1996. The Many Faces of Dependency in Old Age. Cambridge University Press, Cambridge.

Baltes, M.M., Maas, I., Wilms, H.-U., Borchelt, M., Little, T.D., 1999. Everyday competence in old and very old age: theoretical considerations and empirical findings. In: Baltes, P.B., Mayer, K.U. (Eds.), The Berlin Aging Study. Aging from 70 to 100. Cambridge University Press, Cambridge, pp. 384–402.

Baltes, M.M., Silverberg, S., 1994. The dynamic between dependency and autonomy: illustrations across the life span. In: Featherman, D.L., Lerner, R.M., Perlmutter, M. (Eds.), Life-Span Development and Behavior, vol. 12. L. Erlbaum, Hillsdale, NJ, pp. 41–90.

Baltes, P.B., 1973. Strategies for psychological intervention in old age: a symposium. The Gerontologist 13, 4–38.

Baltes, P.B., Lindenberger, U., Staudinger, U.M., 2006. Life-span theory in developmental psychology. In: Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology, Theoretical Models of Human Development, sixth ed., vol. 1. Wiley, New York, pp. 569–664.

Bandura, A., 1977. Self-efficacy: toward a unifying theory of behavioral change. Psychological Review 84 (2), 191–215.

Carstensen, L.L., 2006. The influence of a sense of time on human development. Science 312 (5782), 1913–1915. http://dx.doi.org/10.1126/science.1127488.

Diehl, M., Chui, H., Hay, E., Lumley, M.A., Grühn, D., Labouvie-Vief, G., 2014. Change in coping and defense mechanisms across adulthood: longitudinal findings in a European-American sample. Developmental Psychology 50, 634–648. http://dx.doi.org/10.1037/a0033619.

Diehl, M., Willis, S.L., 2004. Everyday competence and everyday problem solving in aging adults: role of the physical and social context. In: Wahl, H.-W., Scheidt, R., Windley, P. (Eds.), Annual Review of Gerontology and Geriatrics, Aging in Context: Socio-Physical Environments, vol. 23. Springer, New York, pp. 130–166.

Erikson, E.H., 1950. Childhood and Society. Norton & Company, New York.

Grühn, D., Lumley, M.A., Diehl, M., Labouvie-Vief, G., 2013. Time-based indicators of emotional complexity: interrelations and correlates. Emotion 13 (2), 226–237. http://dx.doi.org/10.1037/a0030363.

Hagestad, G.O., 1985. Continuity and connectedness. In: Bengtson, V.L., Robertson, J.F. (Eds.), Grandparenthood. Sage, Beverly Hills, CA, pp. 31–48.

Havighurst, R.J., 1972. Developmental Tasks and Education, third ed. McKay, New York.

Heckhausen, J., Wrosch, C., Schulz, R., 2010. A motivational theory of life-span development. Psychological Review 117 (1), 32–60. http://dx.doi.org/10.1037/a0017668.

Löckenhoff, C.E., 2012. Understanding retirement: the promise of life-span developmental frameworks. European Journal of Ageing 9 (3), 227–231. http://dx.doi.org/10.1007/s10433-012-0241-9.

Parmelee, P.A., Lawton, M.P., 1990. Design for special environments for the elderly. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, third ed. Academic Press, San Diego, CA, pp. 464–487.

Pinquart, M., Sörensen, S., 2003. Differences between caregivers and noncaregivers in psychological health and physical health: a meta-analysis. Psychology and Aging 18, 250–267. http://dx.doi.org/10.1037/0882-7974.18.2.250.

Ryff, C.D., 1989. Beyond Ponce de Leon and life satisfaction: new directions in quest of successful ageing. International Journal of Behavioral Development 12 (1), 35–55. http://dx.doi.org/10.1177/016502548901200102.

Seligman, M.E.P., 1975. Helplessness: On Depression, Development, and Death. W. H. Freeman, San Francisco.

Smith, J., Baltes, P.B., 1999. Trends and profiles of psychological functioning in very old age. In: Baltes, P.B., Mayer, K.U. (Eds.), The Berlin Aging Study. Aging from 70 to 100. Cambridge University Press, Cambridge, U.K, pp. 197–226.

Tornstam, L., 2005. Gerotranscendence – A Developmental Theory of Positive Aging. Springer, New York.

Wahl, H.-W., 2013. The psychological challenge of late-life vision impairment: concepts, findings, and practical implications. Journal of Ophthalmology 1, 1–11. http://dx.doi.org/10.1155/2013/278135. http://dx.doi.org/.

Wahl, H.-W., Fänge, A., Oswald, F., Gitlin, L.N., Iwarsson, S., 2009. The home environment and disability-related outcomes in aging individuals: what is the empirical evidence? The Gerontologist 49, 355–367. http://dx.doi.org/10.1093/geront/gnp056.

Wahl, H.-W., Iwarsson, S., Oswald, F., 2012. Aging well and the environment: toward an integrative model and research agenda for the future. The Gerontologist 52 (3), 306–316. http://dx.doi.org/10.1093/geront/gnr154.

Wettstein, M., Wahl, H.-W., Shoval, N., Oswald, F., Voss, E., Seidl, U., Frölich, L., Auslander, G., Heinik, J., Landau, R., 2012. Out-of-home behavior and cognitive impairment in older adults. Findings of the SenTra project. Journal of Applied Gerontology. http://dx.doi.org/10.1177/0733464812459373. Online published.

Willis, S.L., Reid, J.D., 1999. Life in the Middle. Academic Press, San Diego, CA.

Adverbial Clauses

Jingxia Lin, Nanyang Technological University, Singapore

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by B. Kortmann, volume 1, pp. 162–167, © 2001, Elsevier Ltd.

Abstract

Adverbial clauses are known from traditional grammar as one of three major classes of subordinate clauses. They are semantically diverse and structurally complex. In addition to modifying main clauses, adverbial clauses can also contribute to discourse cohesion. In light of recent cross-linguistic research, this article discusses adverbial clauses with regard to their structure and distribution, meanings, functions in discourse, and structural properties influencing their interpretation.

Adverbial clauses are known from traditional grammar as one of three major classes of subordinate clauses (the other two being relative and complement clauses). Their major function is that of an adverbial, i.e., they provide information on the (temporal, locative, causal, conditional, etc.) circumstances under which the events depicted in the main clauses take place. Correspondingly, the adverbial clauses in (1) are called temporal, locative, causal, and conditional clauses, respectively.

1. They will meet…
   a. before the sun rises.
   b. where they first made love to each other.
   c. because they need to find a solution.
   d. if we let them.

Given the large spectrum of possible circumstances, adverbial clauses represent the most semantically diverse class of subordinate clauses as well as the most challenging class for interpretation. Also given their subject-predicate structure, adverbial clauses are formally the most complex type of adverbials compared with adverbs (e.g., soon, here, quickly) and adverb phrases (e.g., on Sunday, in the garden, very quickly). Combining with a sentence frame like 'They will meet …' in (1), adverbial clauses yield a complex sentence, whereas adverbs and adverbial phrases still yield a simple(x) sentence. Beyond the complex sentence of which they form a part, adverbial clauses have a crucial function in the creation of a coherent discourse and are thus a prominent feature of texts. It seems that adverbial clauses can be found in all languages of the world even though they are not marked in the same way (Thompson and Longacre, 1985). In light of recent cross-linguistic research, this article discusses adverbial clauses with regard to their structure and distribution (Section The Structure and Distribution of Adverbial Clauses), meanings (Section Semantic Types of Adverbial Clause), functions in discourse (Section Functions of Adverbial Clauses in Discourse), and structural properties influencing their interpretation (Section The Interpretation of Adverbial Clauses).

The Structure and Distribution of Adverbial Clauses

The Structure of Adverbial Clauses

Adverbial clauses are subordinate clauses in the sense that their occurrence usually depends on the main clause. However, not all languages mark the distinction between dependent and independent clauses in the same way. Here are the devices that are commonly found among languages: (i) adverbial subordinators, either in the form of separate words such as 'before' and 'where' in English (1) or verbal suffixes as in Kiowa (a language in central United States) (2); (ii) special verb forms such as infinitives and participles as in English and Latin (3); (iii) special word order as in German where independent clauses usually uses verb-second word order whereas subordinate clauses usually use verb-final word order (4) (Dryer, 2011; Thompson et al., 2007).

2. à-dę̀·k'ɔ́··àl hɔ́n àn à-dę̀·hę̀·m-ô
   1SG-lie-although NEG HAB 1SG-sleep-NEG
   "Although I lie down, I can't fall asleep." (Dryer, 2011 from Watkins, 1984: 242)
3. Inflating her lungs, Fiona screamed.
4. Wir wohn-ten auf dem Lande,wie ich dir schon gesagt habe
   we live-PAST on ART(DAT) land as I you already told have(SG)
   "We lived in the country, as I have already told you." (Thompson et al., 2007: 239)

Adverbial clauses may also be expressed in verbless form, as the English example in (5). Such clauses typically do not specify in which way (temporally, causally, conditionally, etc.) they modify the state of affairs described in the main clause and thus are much more challenging for interpretation than most prototypical adverbial clauses.

5. Alone in his room, she switched on the light.

Subordinators in the form of separate words (such as 'that' and 'if' in English) are the most common way to mark adverbial clauses (Dryer, 2011). Yet the presence of a subordinator does not guarantee that the relevant subordinate clause can be classified as an adverbial clause or a relative or complement clause. As illustrated in (6), the word 'that' is used as an adverbial subordinator in (6a), but a complementizer in (6b) and a relativizer in (6c):

6. a. He talked so fast that most people couldn't follow.
   b. He said that most people couldn't follow.
   c. The talk that most people couldn't follow was given by a colleague of mine.

The sentences in (6) show that even in individual languages there may exist no formal differences between the three major types of subordinate clause. Rather, we need to refer to their

function to determine the type: an adverbial clause is a modifier of the main clause and thus optional; a relative clause is a modifier of a noun (phrase) and thus optional; a complement clause is the core argument of a predicate and thus not omissible (Diessel, 2001).

Even though, a clear-cut classification may still be impossible or depend on one's point of view. The subordinate clauses in example (7) are sometimes called adverbial relative clauses since they "can be paraphrased with a relative clause with a generic and relatively semantically empty head noun" (Thompson et al., 2007: 245); in other words, 'when,' 'where,' and 'as' in (7) can be analyzed as relative adverbs rather than adverbial subordinators and replaced by 'at the time,' 'at the place,' and 'in the way,' respectively. As a matter of fact, it is not difficult to find languages where in particular adverbial clauses of Time, Place, and Manner resemble and share properties with relative clauses (Kortmann, 1997; Diessel, 2001; Thompson et al., 2007).

7. a. She'll leave when John comes.
   b. I forgot the bag where we met last time.
   c. He wrote as his teacher taught him.

The Distribution of Adverbial Coordinators

Adverbial subordinators, by which adverbial clauses are morphologically marked, have two major forms: separate words and verbal suffixes. According to the investigation of 660 languages by Dryer (2011), adverbial subordinators in the form of separate words have three possible positions in a clause, at the beginning as in English, at the end as in Kombai (a Trans-New Guinea language in Indonesia), or inside the clause (e.g., between the subject and the verb) as in Nkore-Kiga (a Bantu language in Uganda); adverbial subordinators in the form of verbal suffixes are mostly found in verb-final languages, so the suffixes are also at the end of the clause (in such cases, the adverbial clause usually precedes the main clause), so that the suffix subordinator is not sentence-final (see Section The Ordering of Adverbial Clauses and Main Clauses).

The number of languages whose adverbial clauses are marked by separate words is much larger than those by verbal suffixes; among the former, the number of languages with clause-initial adverbial subordinators is about four times larger than those with clause-final and clause-internal subordinators (Dryer, 2011). There are also languages with more than one type of subordinators and none of them is the dominant; for instance, the adverbial subordinators in Epena Pedee, a Choco language in Colombia, are found both in words (in clause-final position) and verbal suffixes (Dryer, 2011). However, no language is found yet where its adverbial clauses are marked by verbal prefix.

The Ordering of Adverbial Clauses and Main Clauses

Adverbial clauses either precede or follow the main clauses and are referred respectively as preposed (or initial) and post-posed (or final) adverbial clauses. Studies found that preposed adverbial clauses tend to occur in strong OV languages, so that the sentence ends with the main clause; VO languages and a small number of OV languages are found with both pre- and post-posed adverbial clauses (Diessel, 2001, 2013). The ordering of the adverbial and main clauses also correlates with the position of the adverbial subordinators described in Section The Distribution of Adverbial Coordinators: languages with both pre- and post-posed adverbial clauses tend to have clause-initial subordinators, whereas languages with preposed adverbial clauses only tend to have clause-final subordinators (Diessel, 2001; Thompson et al., 2007). For the former type, studies show that the ordering of the adverbial clauses varies according to their semantic relations with the main clauses: conditional clauses tend to be preposed, temporal clauses exhibit a mixed tendency of both pre- and post-posing, whereas causal and purposive clauses are often found to follow the main clause (Diessel, 2001, 2008, 2013; Diessel and Hetterle, 2011; Schmidtke-Bode, 2009). Pre- and post-posed adverbial clauses also tend to have different pragmatic functions in discourse (see details in Section Functions of Adverbial Clauses in Discourse).

The relations of pre- and post-posed adverbial clauses to the main clauses are not the same either. For example, in interrogative questions, the information carried in a post-posed adverbial clause is usually included in the question, but that in a preposed one is usually not (Diessel, 2013). As illustrated in (8), the post-posed adverbial clause 'while she was still in her office' in (a) can be understood as the focus of the question, that is, whether this is the time when 'you' talked to 'her,' but the preposed clause in (b) is not necessarily an 'integral semantic component' of the main clause (Diessel, 2013: 348).

8. a. Did you talk to her while she was still in her office?
   b. While she was still in her office, did you talk to her? (Diessel, 2013: 348)

In terms of intonation, post-posed adverbial clauses are found relatively more separated from the main clause, i.e., an intonation break may be found in between the two clauses (also see Section The Interpretation of Adverbial Clauses for exceptions); on the contrary, preposed adverbial clauses are usually intonationally bound to the main clause (cf Ford, 1993).

Semantic Types of Adverbial Clause

Traditionally, adverbial clauses are classified according to the semantic relations of the events depicted in different parts of a complex sentence or different chunks of discourse. The exact number and labeling of these semantic relations vary. All languages are reported to use the adverbial clauses (or a subset of the relations) listed in (9) (cf Kortmann, 1997; Givón, 1990; Thompson et al., 2007); other relations found in all European languages include Instrument/Means 'by,' Degree/Extent 'insofar as,' Exception/Restriction 'except/only that,' (Negative) Concomitance 'with(out)' (Kortmann, 1997).

9. Time (Simultaneity Overlap/Duration/Co-Extensiveness 'when/while/as long as,' Anteriority 'after,' Immediate Anteriority 'as soon as,' Terminus 'since,' Posteriority 'before,' Terminus 'until,' Contingency 'whenever');

Place 'where';
Manner 'as, how';
Cause/Reason 'because';
Purpose (Purpose 'in order that,' Negative Purpose 'lest');
Result 'so that';
Condition (Condition 'if,' Negative Condition 'unless,' Concessive Condition 'even if');
Concession 'although';
Contrast 'whereas';
Addition 'in addition to';
Substitution 'instead of, rather than';
Similarity 'as, like'

One would assume that not all of these semantic relations are equally central to human cognition. Indeed, there is evidence suggesting roughly a dozen cognitively most central relations, including simultaneity (overlap 'when,' duration 'while'), place ('where'), similarity ('as'), cause, condition, and concession (Kortmann, 1997). For example, the latter three relations are found in all the European languages that mainly use finite adverbial clauses and these three also have the largest number of adverbial subordinators in the European languages. Moreover, the adverbial subordinators marking the core relations tend to be more reduced morphologically, much more frequently used, and older than those marking the peripheral relations (Kortmann, 1997).

Simultaneity and cause also figure prominently in a large-scale semantic analysis of nonfinite adverbial clauses. In addition, nonfinite adverbial clauses are frequently interpreted in the sense of addition/concomitance ('and at the same time'), as in (10) and exemplification/specification ('e.g., i.e., in that, more exactly'), as in (11).

10. a. There he sat, wearing a white golfing cap.
    b. Sam threw himself to the ground, dragging Frodo with him.
11. a. Shares in Midland were worst hit, falling at one time 42p.
    b. He paid the closest attention to everything Lenny said, nodding, congratulating, making all the right expressions for him.

Therefore, not all semantic relations are equally important in different structural types of adverbial clauses. Likewise, their relative importance as coherence relations depends on the type of discourse. For instance, cause, condition, and concession play a much more important role in academic writing than they do in narrative fiction, where temporal relations as well as, for nonfinite adverbial clauses, addition/concomitance and exemplification/Specification account for a much higher number of adverbial clauses (see Kortmann (1991) for statistics and a discussion of relevant literature).

Functions of Adverbial Clauses in Discourse

Typically adverbial clauses provide background information for what is depicted in the main clause. But they serve additional functions beyond sentence level, contributing cohesion by linking sentences, paragraphs, or even the whole discourse. In interactive communication, adverbial clauses can also modify the speech act of the speaker, rather than the main clause.

The Functions of Adverbial Clauses beyond Sentence

Adverbial clauses help create coherent discourse (Chafe, 1984; Thompson, 1985; Givón, 1990; Ford, 1993; Couper-Kuhlen and Kortmann, 2000; Thompson et al., 2007; Diessel, 2013). Depending on whether they precede or follow their main clause, adverbial clauses produce more global (or textual) coherence or more local coherence, respectively. Preposed adverbial clauses serve a kind of guidepost or scene-setting function for the reader or listener by (1) linking back to what has gone in the (not necessarily immediately) preceding sentence, paragraph, or discourse and (2) preparing the background for what is going to follow in the complex sentence and often even a whole chunk of discourse. By contrast, post-posed adverbial clauses typically have a much more local function, i.e., their scope is restricted to their immediately preceding main clause. They neither reach back into earlier parts of the discourse, nor fore-shadow or prepare for what is going to follow. For example in written English, the preposed purpose clauses usually contain issues from preceding discourse, whereas the post-posed ones often describe purposes of the main clauses (Thompson, 1985). Accordingly, the subject of a post-posed adverbial clause typically is identical with the main clause subject, whereas the subject of a preposed adverbial clause is often identical with that of (one of) the preceding sentence(s).

Speech Act Adverbial Clauses

In addition to the adverbial clauses that link main clauses, sentences, or paragraphs as introduced above, some adverbial clauses function at the level of interactive communication by modifying the speech act of the speaker and are called 'speech act' adverbial clauses (cf Rutherford, 1970; Kac, 1972; Sweetser, 1990; Thompson et al., 2007; Diessel, 2013). Two examples are given in (12). Instead of setting a condition for the event in the main clause, the 'if' clause in (12a) expresses the speaker's attitude that he or she may not be in a position for making the statement in the main clause; similarly, the 'because' clause in (12b) does not state the reason for Harry to be late, but the reason why the speaker states that Harry will be late, i.e., the speaker talked to Harry's wife, through which he or she knows that Harry will be late (Diessel, 2013; Thompson et al., 2007).

12. a. And if I may say so Mr Speaker … they possibly derived some benefit from the presence of the Chancellor …. [International Corpus of English] (Diessel, 2013: 345)
    b. Harry will be late, because I just talked to his wife (Thompson et al., 2007: 267).

The Interpretation of Adverbial Clauses

Problems of interpreting adverbial clauses not only arise for the inherently vague nonfinite adverbial clauses, but also for some finite adverbial clauses. World knowledge or contextually grounded knowledge is important for the interpretation, but

formal (i.e., morphological, syntactic, and prosodic) features may also crucially influence the process of interpretation. For example, polysemy is observed for almost one-third of the adverbial subordinators in the European languages, especially for those with a high text frequency (Kortmann, 1997). This section introduces some of the most important relevant features that may influence the interpretation, always provided that the given language allows for a choice (e.g., has no constraints such as the obligatory use of the subjunctive mood in subordinate clauses or subordinate clauses generally preceding their main clause) (for further discussion see Kortmann (1991) and König (1995)):

i. Choice of tense and/or mood in the adverbial clause and, accordingly, in the main clause. For example, many languages use different tense (and mood) for the three semantic types of conditional: present tense (indicative mood) in factual/real conditional clauses (13a), past tense (or subjunctive mood in some languages) in hypothetical conditional clauses (13b), and past perfect (or a conditional perfect) in counterfactual clauses (13c).

13. a. If she comes home, I will be very happy.
    b. If she came home, I would be very happy.
    c. If she had come home, I would have been very happy.

Tense constraints also hold for the two different meanings (temporal and causal) of English 'since.' 'Since' can have a temporal reading in two situations; being used with past tense is one of them (see the other in (16)). Mood is also adopted in many (e.g., Romance) languages to help differentiate result ('so that') and purpose ('in order to that') that share the same adverbial subordinators: indicative mood leads to a result reading, subjunctive mood to a purpose reading.

ii. Nonsubordinate word order. As introduced in Section The Structure and Distribution of Adverbial Clauses, German usually uses verb-second word order for independent clauses and verb-final word order for dependent clauses. However, the verb-second word order is also typical in spoken German for post-posed *weil*-causals and *obwohl*-concessives clauses. An example is given in (14).

14. Ich hab das mal in meinem ersten Buch aufgeschrieben. Weil dann glauben's die Leute ja. ('I've written that down in my first book. Because people believe it then.')

iii. Intonation. One relevant intonational feature is the presence or absence of an intonation break: only when the complex sentence in (15) is read as a single intonation group does the adverbial clause receive a concessive conditional reading ('even if').

15. "I wouldn't marry you if you were the last man on earth" (Kortmann, 1997: 92 from Haiman, 1986).

iv. The choice of verb forms. The dependent versus independent verb forms were shown to be relevant in (i) already (indicative vs. subjunctive mood). (16) illustrates the impact of the choice between finite and nonfinite form in the adverbial clause on the interpretation of an adverbial subordinator, here 'since': when introducing a free adjunct, 'since' can only receive a temporal reading; a causal one is impossible (cf 'since' in (i)).

16. Since working with the new company, Frank hasn't called on us even once.

v. Constituent order, more exactly the relative order of adverbial and main clauses (cf Kortmann, 1991; König, 1995; Diessel, 2001, 2005, 2008). In English, for example, the great majority of present-participial free adjuncts receive a temporally sequential interpretation relative to the main clause (Kortmann, 1991). That is, the order of events is presented iconically by the relative order of adverbial and main clauses, as illustrated by the minimal pair in (17).

17. a. She uncurled her legs, reaching for her shoes.
    b. Reaching for her shoes, she uncurled her legs.

See also: Conversation Analysis: Sociological; Functional Approaches to Grammar; Grammatical Relations; Iconicity in Linguistics; Linguistic Typology; Semantics.

Bibliography

Chafe, W., 1984. How people use adverbial clauses. Berkeley Linguistics Society 10, 437–449.

Couper-Kuhlen, E., Kortmann, B. (Eds.), 2000. Cause-Condition-Concession-Contrast: Cognitive and Discourse Perspectives. De Gruyter, Berlin.

Diessel, H., 2001. The ordering distribution of main and adverbial clauses a typological study. Language 77, 345–365.

Diessel, H., 2005. Competing motivations for the ordering of main and adverbial clauses. Linguistics 43, 449–470.

Diessel, H., 2008. Iconicity of sequence: a corpus-based analysis of the positioning of temporal adverbial clauses in English. Cognitive Linguistics 19, 457–482.

Diessel, H., 2013. Adverbial subordination. In: Luraghi, S., Parodi, C. (Eds.), Bloomsbury Companion to Syntax. Continuum, London, pp. 341–354.

Diessel, H., Hetterle, K., 2011. Causal clauses: a cross-linguistic investigation of their structure, meaning, and use. In: Siemund, P. (Ed.), Linguistic Universals and Language Variation. Mouton de Gruyter, Berlin, pp. 21–52.

Dryer, M.S., 2011. Order of adverbial subordinator and clause. In: Dryer, M., Haspelmath, M. (Eds.), The World Atlas of Language Structures Online. Max Planck Digital Library, Munich. http://wals.info/chapter/94.

Ford, C.E., 1993. Grammar in Interaction: Adverbial Clauses in American English Conversations. Cambridge University Press, Cambridge.

Givón, T., 1990. Syntax: A Functional-Typological Introduction. John Benjamins, Amsterdam-Philadelphia.

Haiman, J., 1986. Constraints on the form and meaning of the protasis. In: Traugott, E., TerMeulen, A., Reilly, J., Ferguson, C. (Eds.), On Conditionals. Cambridge University Press, Cambridge, pp. 215–228.

Kac, M.B., 1972. Clauses of saying and the interpretation of because. Linguistics 48, 626–632.

König, E., 1995. The meaning of converb constructions. In: Haspelmath, M., König, E. (Eds.), Converbs in Cross-linguistic Perspective. de Gruyter, Berlin, pp. 57–96.

Kortmann, B., 1991. Free Adjuncts and Absolutes in English: Problems of Control and Interpretation. Routledge, London.

Kortmann, B., 1997. Adverbial Subordination: A Typology and History of Adverbial Subordinators Based on European Languages. de Gruyter, Berlin.

Rutherford, W., 1970. Some observations concerning subordinate clauses in English. Language 46, 97–115.

Schmidtke-Bode, K., 2009. A Typology of Purpose Clauses. John Benjamins, Amsterdam-Philadelphia.

Sweetser, E.E., 1990. From Etymology to Pragmatics. Metaphorical and Cultural Aspects of Semantic Structure. Cambridge University Press, Cambridge.

Thompson, S.A., 1985. Grammar and written discourse: initial vs. final purpose clauses in English. Text 5, 55–84.

Thompson, S.A., Longacre, R.E., 1985. Adverbial clauses. In: Shopen, T. (Ed.), Language Typology and Syntactic Description, vol. II. Cambridge University Press, Cambridge, pp. 171–234.

Thompson, S.A., Longacre, R.E., Hwang, S.J., 2007. Adverbial clauses. In: Shopen, T. (Ed.), Language Typology and Syntactic Description, second ed. Cambridge University Press, Cambridge, pp. 237–300.

Watkins, L.J. (with the assistance of Parker McKenzie), 1984. A Grammar of Kiowa. University of Nebraska Press, Lincoln.

Advertising Agencies

Brian Moeran, Copenhagen Business School, Copenhagen, Denmark

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 1, pp. 167–172, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Advertising agencies are the most significant organizations in the development of advertising and marketing worldwide. An advertising agency is an independent service company, composed of business, marketing and creative people, who develop, prepare, and place advertising in advertising media for their clients, the advertisers, who are in search of customers for their goods and services. Agencies thus mediate between three different but interlocking social groups: industry, media, and consumers. The history of advertising is largely the history of the advertising agencies that have served the needs of these three groups. They link industry and media by creating new forms for messages about products and services; industry and consumers by developing comprehensive communications campaigns and providing information thereon; and media and consumers by conducting audience research to enable market segmentation. This article is concerned with the origins, early developments, organization, compensation arrangements, and accounts of contemporary full-service advertising agencies.

An advertising agency is an independent service company, composed of business, marketing, and creative people, who develop, prepare, and place advertising in advertising media for their clients, the advertisers, who are in search of customers for their goods and services. Agencies thus mediate between three different but interlocking social groups: industry, media, and consumers. The history of advertising is largely the history of the advertising agencies that have served the needs of these three groups (*see* Advertising: General). They link industry and media by creating new forms for messages about products and services; industry and consumers by developing comprehensive communications campaigns and providing information thereon; and media and consumers by conducting audience research to enable market segmentation.

Origins and Early Developments

Advertising agencies are the most significant organizations in the development of advertising and marketing worldwide. They came into existence in the United States in the mid-nineteenth century and, later, elsewhere because of the mutual ignorance of the needs of newspaper publishers and would-be advertisers and because of the opportunity for profit provided by both parties' desire for economic assistance. Initially, advertising agents facilitated the purchase and sale of space. In so doing they promoted the general use of advertising, helped advertisers find cheaper and more effective ways of marketing goods, and served to inform the public of the existence of those goods. They acted as a crucial bridge between the activities of selling products and of mass communication at a time when both were undeveloped and expanding rapidly (*see* Journalism; Mass Media, Political Economy of).

The first advertising agents were colonial postmasters in America who accepted advertisements for inclusion in newspapers. In 1843, Volney B. Palmer set up the first independent agency, soliciting orders for advertising, forwarding the copy, and collecting payment on behalf of newspapers that had difficulty in getting out-of-town advertising. This *newspaper agency* was the first of four stages through which the business of the advertising agent proceeded to pass. In the second *space-jobbing* stage, the agent became an independent middleman who sold space to advertisers and then bought space from newspapers to fill his orders, driving a hard bargain with both. In 1865, George P. Rowell initiated the third *space-wholesaling* phase when, anticipating the needs of advertisers, he bought wholesale from publishers large blocks of space that he then resold in smaller lots at retail rates. Finally, in 1867, some agents contracted annually with the publications they represented to buy in advance the entire advertising space of particular newspapers, thereby acquiring the *advertising concession* in a publication. Similar processes of development can be found outside the United States, in countries such as England, France, and Japan, where advertising agencies also owe their early existence to the buying and selling of newspaper space.

Although the early agency business was often a confusing amalgam of these different arrangements, one element was common to them all and has continued to some extent to the present day. Advertising agents found themselves in an equivocal position as they attempted to serve two masters: on the one hand, the *publishers*, on whose behalf they sold advertising space and, on the other, the *advertisers*, to whom they acted as expert and 'impartial' advisors in the placement of their ads. In other words, right from the start, the advertising agency business has invited conflict between the diverse interests of the advertiser, the agent, and the media in which advertising is placed.

It was in order to overcome such a potential conflict of interests caused by space brokering that the idea of an 'open contract plus commission' was introduced. For the first time, in 1875, N. W. Ayer & Son entered into a formal arrangement that enabled a client to use the agency's services exclusively in exchange for payment of a fixed compensation based on the amount of advertising it billed each year. This so-called 'billing' system soon spread and is still widely used by advertising agencies. It is the gross billings posted by advertisers, rather than the actual payments received by agencies from their

clients, which are today used to gauge the relative size of the world's advertising agencies and their industries.

There were two important social consequences of the introduction of a contract between agency and advertiser. First, under the space brokerage system, advertising agents were in constant danger of being squeezed between advertisers and publishers. The open contract, however, permitted the agent to represent the advertiser *as well as* the publisher. Not only this, but it established the agency's relationship to a *client* rather than to a customer which hitherto had given its business to various different agents.

Second, by entering into a contract with an advertiser, an agency's profits were tied to the amount that its clients billed. Its growth was thus linked to the economic well-being of the advertisers it represented, so that even today an agency's prosperity is totally determined by the prosperity of its clients. When the economy is buoyant, advertisers spend and agencies flourish; when there is a recession, advertising appropriations are cut back, and agencies struggle. Such close ties can extend to agencies' organizational expansion. When advertisers' business interests lead them to set up international offices abroad, their agencies often follow. This explains the establishment, growth, and commanding presence of American agencies in most advertising industries worldwide, although English agencies tend to have a strong presence in former British colonies. In Japan, too, a buoyant domestic market has led to the growth of major Japanese advertisers served by Japanese agencies that have resisted American dominance and followed their clients abroad, especially into East and Southeast Asia (*see* International Advertising).

Full Service Agencies

What is known as a full-service agency is an agency that handles the planning, creation, production, and placement of advertising for its clients. It may also handle sales promotion and other related services as needed by individual advertisers. In short, a full-service agency offers its clients a complete range of services beyond the preparation and placement of their advertising.

The development of the full-service agency goes back to the last quarter of the nineteenth century, when mounting competition forced agents to think of other services that they might offer (potential) customers besides the placement of their advertising. One of these was marketing. In 1879, N. W. Ayer & Son conducted its first market survey in a successful attempt to wrest a large advertising account from a competitor. This led to the implementation of advertising planning, as agencies gradually took over the functions of studying the actual and possible markets for the commodities sold by their customers and of devising plans to fit into the general strategy for tapping those markets.

It was at this time, too, that agencies first became involved in what is now known as 'creative' work. N. W. Ayer & Son, and Catkins and Holden were among the first agencies to help advertising become a profession rather than a mere business transaction when they began to write advertisements for their clients, instead of simply placing their finished ads in newspapers. As a result, as well as sending salesmen out to sell space to customers, some advertising agencies began to create, plan, write, and design the copy and artwork that went into every ad. They then produced the necessary plates, and placed them with the publishers.

There were two significant reasons underlying agencies', at first reluctant, move into the creative side of advertising. First, they recognized that the ineffectiveness of a poorly written advertisement (contracted or written by the advertiser) was sure to redound against their fledgling business. Second, they found themselves operating in a highly competitive environment in which increased skill was demanded in the writing of advertising in order to make it stand out effectively. Moreover, new forms of advertising (magazine and outdoor) were being introduced. This necessity for distinction in the preparation of copy was made more urgent by technical advances in the printing and engraving trades. These made possible a more plentiful use of color and illustration, but they also complicated the work of advertising agencies.

Later technological innovations created further opportunities for agencies to develop into new areas of expertise and so come to be seen as essential partners in the advertising endeavor. Radio, for example, offered an entirely new medium of entertainment (and advertising) dependent on sound alone (*see* Radio as a Medium). Within three years of its first transmission, and in spite of the immense technical difficulties involved, agencies had begun to take an active part, not only in the preparation of advertising commercials, but in the planning and production of radio programs. These included continuity drama or 'soaps' which, almost from the very first, were supported by advertising and publicity.

The advent of television, which became popular in most industrialized countries from the 1950s, also obliged agencies to sail further into uncharted territory – this time by learning how to combine sound and images in cinematic form (*see* Television: General). As television advertising became ever more popular, dominating print advertising in many parts of the world, and as agencies continued to be active in the planning and production of media outputs, they found themselves coming into more frequent contact with production companies, celebrities, model agencies, film studios, stylists, photographers, animators, and others in the entertainment world. The latter decades of the twentieth century, therefore, have seen the formation of increasingly close ties among the advertising, media, and entertainment industries (*see* Celebrity; Entertainment).

Besides 'above-the-line' advertising in the four main media of television, newspapers, magazines, and radio, full-service agencies have found themselves moving into 'below-the-line' activities involving all sorts of areas of contemporary life hitherto untouched by advertising. Some – like the ads found on the backs of Australian postage stamps during the reign of Queen Victoria and of British dog licenses in the 1920s and 1930s – have ceased to exist. Most others, however, remain. At this point it seems unlikely that the advertising currently found on airport baggage trolleys, grand prix racing cars, telephone cards, professional sports equipment, and the backs of tickets of all kinds will, short of legislation, fall into disuse.

To further improve the services that they already offer their clients, advertising agencies have in many respects acted as 'cultural intermediaries' between the economy and culture of contemporary societies. This means that they have begun to

move away from the four main media that sustained the advertising industry throughout the whole of the twentieth century, and to invest in new business opportunities, including merchandizing, telemarketing, and e-commerce. The commercial sponsorship of sports, art, and other cultural events that became prominent in the 1980s and 1990s has also been driven by advertising agencies which now look for further opportunities to show their skills in matching their clients' needs for publicity and advertising to the everyday lives of consumers (*see* Mass Media and Sports). This development of 'promotional culture' has led to agencies participating in everything from the development of animation programs in consultation with television networks, supported by product placement and character merchandizing on the part of toy and other manufacturers, on the one hand, to the arrangement of marriages between celebrities, on the other. In many respects, agencies can best be described as the black boxes of cultural continuity and change.

Agency Organization

Many agencies in Europe and the United States have been founded by two entrepreneurs: one a creative person, the other a salesman or account manager. For example, the J. Walter Thompson (JWT) agency was founded by husband-and-wife team of Stanley Resor and Helen Resor who focused on management and copywriting, respectively. While, at first, they may have handled all aspects of their advertising business themselves, eventually they were obliged to take on employees with different areas of expertise and add divisions to handle the basic areas of responsibility in a full-service agency. Although agencies differ in the way they are organized, the main functional areas of a full-service agency are: (1) *account management*, (2) account planning or *research*, (3) *media planning* and buying, (4) *creative* services, and (5) *internal* services (dealing primarily with finance, personnel, and traffic). They are usually led by a chief executive officer (CEO) and one or two vice presidents, who may oversee a board of directors whose members represent different areas of their agency's responsibility.

The exact weight given to the work of each functional area has varied over time and differs also both by agency and by country. During the first half of the twentieth century, emphasis was placed by such famous copywriters as Claude Hopkins on a scientific approach to advertising – an approach supported by the founding of research organizations by A. C. Nielsen, George Gallup, and others. Later, in the 1960s, led by such creative advertising people as 'Marlboro man' Leo Burnett, David Ogilvy, and William Bernbach, art, inspiration, and intuition became the industry's buzzwords. The 1970–90s saw a return to hard-sell advertising in what was called the era of 'accountability.'

How do these different areas of responsibility work in practice? When an agency wins a new account, it generally forms a cross-department account group, or account team, composed of an account supervisor (who is often supported by account executives), account planners and media buyers, as well as a creative team made up of copywriter and art director with, as necessary, a producer (for television and/or radio commercials). The account group proceeds along more or less predetermined lines. First, the account planning department works out and solves a marketing problem by conducting new, and making use of previous, research to find out who the targeted consumers (known as 'prime prospects') are; what their demographic characteristics might be; how the product to be advertised fits into their lifestyle; what they think of this and competing products; what the most suitable advertising medium for the product in question might be; and so on. On the basis of answers to these questions, account planners (who are sometimes simply referred to as 'marketers') formulate a strategy that positions the product in relation to targeted consumers and emphasizes the attribute(s) that will appeal to them.

Once the overall marketing strategy has been determined, the agency's creative team works on an appropriate creative strategy, writes copy, and prepares rough layouts and storyboards. At the same time, the media buyers need to work out a media plan that accords with the marketing strategy, to select an appropriate mix of media, and prepare schedules with costs. These three separate plans are then amalgamated into a single package by the account supervisor and presented to the client, which may or may not get involved in different stages of the agency's preparations along the way. The extent to which an advertiser 'interferes' in its agency's work depends very much on the company concerned. However, evidence suggests that, for reasons that will be explained below, agencies tend to work much more closely with their clients in Japan than they do elsewhere in the world.

One of the main organizational difficulties facing any account group is that it invariably finds itself having to deal with two different audiences whose interests may not be the same. On the one hand, account planners and the creative team focus their attention on how to advertise and sell a product to a particular targeted group of *consumers*. On the other, the account supervisor and his supporting account executives need to liaise between the *advertiser* and the agency, being responsible for interpreting the client's marketing needs to their colleagues, on the one hand, and selling the marketing, creative, and media plans to the client, on the other. Since it is the advertiser who pays the agency for its services, it may disagree with – and occasionally overrule – the consumer-oriented strategy proposed by its agency's account group.

A second organizational difficulty that often emerges in the work of an account group is internal and concerns the different attitudes towards advertising held by account planners, on the one hand, and creative people, on the other. The very *raison d'être* of account planning is to form marketing strategies based on objective, scientific criteria derived from in-depth qualitative and quantitative data. These data then have to be transformed into creative images by copywriters and art directors who usually claim to work according to intuitive, artistic ideas that may have little actual relationship to the expressed marketing aims. Sometimes conflict between the two types of advertising agency employee results and adversely affects the work of the account group which, ideally, needs to blend harmoniously to achieve its client's aims.

Agency Compensation

It was noted earlier that advertising agencies started out as brokers of space on behalf of newspaper publishers who paid

them a commission for their services. Traditionally, this commission has amounted to 15% of the media cost. It may be more for some kinds of advertising (for example, outdoor), or less for certain advertisers whose account an agency, for one reason or another, prizes highly enough to accept a lower commission. In addition to media commissions, agencies may charge a production commission for production work subcontracted (type, photography, and illustration). Here the client is charged the cost of the work involved, plus a 17.65% commission.

As seen above in the actual creation of an advertising campaign, agencies have often found themselves dealing with the conflicting interests of two different partners. The very existence of the commission system has, historically, encouraged the advertiser to try to pay less than the sum demanded by the medium in which it sought to place its advertising. As a result, advertising agencies often found themselves caught between a desire for profit, on the one hand, and a need to retain its business, on the other. This has led, and in some parts of the world still leads, to agencies playing off one partner against the other, as they seek rebates from the media in which they place their clients' advertising, but pass on the full cost to the advertisers concerned.

It is to offset this problem of business ethics that fees have been introduced. This form of compensation is comparable to that used to pay lawyers or accountants. Agency and client agree an hourly fee for the agency's work, to which are added charges for out-of-pocket expenses, travel, and other standard items. No commission is paid on media costs which are billed net of commission. As a rule, advertiser–agency compensation arrangements are negotiated according to the specific requirements of the advertiser and the willingness and ability of the agency to fulfil those needs at a mutually agreeable cost. Over time, the fee system has come to predominate over the commission system in the United States, although in many other countries the commission system, or a combination of fee and commission, is still the norm. It seems likely that in future the fee system will prevail, in which case the traditionally close reciprocal ties between advertising agencies and media organizations will probably weaken to some extent.

The Account System

The sum of money put aside by an advertiser and allocated to an agency for the purpose of selling a particular brand or product group, sometimes through a selected medium, is called an account. It is an advertising agency's job to persuade an advertiser that it is best suited to take on a particular account. To this end it will embark on active solicitation with the aim of being asked to participate in a competitive presentation or 'pitch' in which, together with other agencies, it will present marketing and creative strategies based on the advertiser's initial orientation of its needs. The successful agency will be asked to take on an account, usually for a fixed period, and will then enter into partnership with its new client.

The relationship between agency and client is ideally a professional one, involving a considerable exchange of confidential information which may include, for example, plans for new products or marketing strategies. As a result, when awarding an account, an advertiser will not usually permit its agency to handle companies or products with which it is in direct competition. This is known in Europe and the United States as the *competing account rule*.

The rule about an agency's not handling competing accounts means that, whenever it wins a new account that conflicts with one already being handled, it has to decide which of the two accounts it wishes to keep. Competing accounts are also often a deciding factor when, as has happened frequently since the 1980s, agencies decide to merge, although many such large agencies run their local offices independently. This they do in the hope that their clients will not regard the same type of account being handled in another office as being in competition and thus not raise an objection.

So far as the organization of the advertising industry is concerned, the competing account rule ensures two things: first, that there is a continuous circulation of accounts among agencies; and second, that no agency ever becomes excessively large. As a result, by comparison with other industries, the organizations that make up the advertising industry are quite small.

Although the competing account rule is the norm in almost all countries of the world, an alternative system of distributing advertising appropriations exists in Japan where advertisers split their accounts by product, media, or a combination of product and media, and do not object to the same agency handling a number of competing accounts. Thus large advertisers like Toyota or Kao, for example, contract more than half a dozen different agencies to handle the advertising of their numerous products. These same agencies may also be handling the accounts of rival car or toiletries manufacturers.

The *split account system* has advantages for both agencies and advertisers. Precisely because the overall advertising appropriation is divided, accounts in Japan are not nearly as large as they are elsewhere. This is to an agency's financial advantage when it loses an account, since the sheer number of accounts in circulation means that it can usually make up the financial shortfall and is not obliged, as it might be in the United States or Europe, to lay off staff. In this respect, the split account system contributes to the overall stability of the advertising agency, and of the industry of which it is a part.

Advertisers are able to use the split account system to play off one agency against another since they alone have full knowledge of their overall marketing strategy. This not only enables advertisers to minimize the problem of confidentiality that so worries those working in advertising industries elsewhere. They can use their knowledge strategically to operate a system of divide-and-rule that obliges agencies to keep in constant touch with client personnel, as well as with other agencies and media organizations, in order to gain as much information as possible about their clients' overall plans and activities. In other words, the split account system encourages much closer interaction among the three main players in the advertising industry.

There are disadvantages to the system. One result of the need for Japanese agency account executives to keep in constant touch with their clients is that advertising decisions tend to be based on personal likes and dislikes, rather than on more

objective professional criteria. Another, claimed by agencies, is that they are prevented from producing effective advertising because their clients do not give them access to their *total* marketing strategy. All in all, however, the advantages appear to outweigh the disadvantages within the Japanese advertising industry. In this respect, it is significant that in the autumn of 1999, Procter & Gamble, one of the world's biggest spending advertisers, decided to divide its worldwide advertising appropriation into nine different brands and to distribute these accounts among nine different advertising agencies. It is possible, therefore, that in future the system of split accounts will become more popular with advertisers outside Japan, thereby altering the relationship hitherto developed between an agency and its clients.

Editor's Note

Since Moeran's entry was first published, the key developments in the world of advertising agencies have been globalization of the industry and the emergence and spectacular growth of internet advertising. On the former, an essential reference is Faulconbridge et al., 2011. On the latter, see David and Evans, 2014. Also useful: Sheehan, 2013.

See also: Advertising and Advertisements; Advertising, Control of; Advertising: Effects; Media Effects.

Bibliography

Evans, D.S., 2014. Rivals for Attention: How Competition for Scarce Time Drove the Web Revolution, What it Means for the Mobile Revolution, and the Future of Advertising. Available at SSRN: http://ssrn.com/abstract¼2391833 or http://dx.doi.org/10.2139/ssrn.2391833.

Faulconbridge, J.R., Beaverstock, J.V., Nativel, C., Taylor, P.J., 2011. The Globalization of Advertising: Agencies, Cities and Spaces of Creativity. Routledge, Oxford.

Hower, R., 1939. The History of an Advertising Agency: N. W. Ayer & Son at Work, 1869–1939. Harvard University Press, Cambridge, MA.

Jones, P., 1999. The Advertising Business. Sage, Thousand Oaks, CA.

Leiss, W., Klein, S., Jhally, S., 1990. Social Communication in Advertising: Persons, Products and Images of Well-being, second ed. Routledge, London.

Lien, M., 1997. Marketing and Modernity. Berg, Oxford, UK.

Miller, D., 1997. Capitalism: An Ethnographic Approach. Berg, Oxford, UK.

Moeran, B., 1996. A Japanese Advertising Agency: An Anthropology of Media and Markets. University of Hawaii Press, Honolulu, HI.

Schudson, M., 1984. Advertising: The Uneasy Persuasion. Basic Books, New York.

Sheehan, K.B., 2013. Controversies in Contemporary Advertising, second ed. Sage, Thousand Oaks, CA.

Wells, W., Burnett, J., Moriarty, S., 2000. Advertising: Principles and Practice. Prentice-Hall, London.

Advertising and Advertisements

Linda Scott, University of Illinois, Urbana, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 1, pp. 172–178, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article offers a definition of advertising, followed by brief descriptions of its function from various points of view (traditional economics, Marxist cultural theory, and so on). The forms advertising has taken over time and in various cultures are described and analyzed with an eye to the implications for consumer response. In particular, the claim that advertising is primarily an 'image' genre, is disputed. Then, a few elements that might constitute an 'aesthetics of advertising' are proposed. Lastly, various theories about 'how advertising works' are described and discussed.

Definition

The way the terms 'advertising' and 'advertisement' are defined depends to a great extent on the author's perspective, training, and agenda. Even so, an encyclopedia entry seems to beg for an authoritative definition. This article offers the following: an advertisement is a message paid for, but not delivered, by the sender, that (1) incorporates technologies or forms other than speech, (2) appears in a public forum, and (3) attempts to persuade receivers to behave in a way that brings direct economic benefit to the sender. This definition allows for most of what is customarily considered advertising, while excluding other acts or forms that, while sharing some characteristics with advertisements, are not generally considered such. For instance, this definition would not include a salesman's pitch, but would encompass any brochures he used. Political speeches, whether made in Parliament or during the breaks on *60 Minutes*, would not be 'advertisements' because the expected behavioral outcome is not directly economic. Corporate 'image' messages would count, even when purchase is not the object, as such advertisements ultimately have an economic goal, such as affecting the value of stock by increasing the prestige of the corporation.

Function

The function of advertising likewise will vary in description based upon the education, focus, and motive of a particular author. Economists traditionally have allowed 'advertising' into their models only as a form of market information. Their definition of 'information,' however, is usually limited to prices and availability. To the extent that advertisements incorporate other 'non-information' (such as prestige claims or sex appeals), advertising becomes an 'exogenous variable,' constitutive of 'tastes,' and therefore present in the system only as noise or interference. As long as advertisements serve to notify markets of supply and price, however, the institution of advertising operates neatly within the range of traditional economic models. Indeed, to the extent that advertising supports stable demand, allowing producers to become price-makers rather than price-takers, it becomes an important aid to microeconomic planning. The realization of economies of scale that accrue from such an arrangement, so the argument goes, tends to reduce prices to consumers in the long run.

Critics and consumer groups often charge that advertising merely adds to the prices paid by consumers – sometimes raising them as much as 100% – but such claims do not hold for most goods in the long run. For example, simple bar soap was sold 100 years ago in the US at a price that included advertising expenses equal to direct costs (thus representing about 50% of each unit's price). As consumption of soap grew, however, the advertising expenditures leveled off (advertisement budgets reach a plateau of effectiveness beyond which it makes no sense to spend at higher levels). Consequently, the percentage of each unit's price represented by advertising steeply declined. Since advertising expense is only one of the economies of scale created by mass marketing and production, the total unit price of each bar of soap also dropped dramatically, allowing many US citizens to wash regularly for the first time. Today, it is only the very extraordinary soap that would include an advertising expense more than 3% of its unit price. The same is true for most well-established, mass-produced, and heavily advertised products.

In the Marxist critique of capitalist consumer culture, advertising's most central function is to ensure capitalism continues: by stimulating ever-increasing levels of demand, advertising forces an always-growing array of goods into consumers'homes, regardless of need or utility. Here too, advertisements that inform audiences of availability and prices are not typically thought problematic. However, the 'fetishization' of objects by advertisements that name, sing, or tell stories is seen as a sinister form of industrial 'magic' through which consumers are made to value products for the wrong reasons. By adding this dubious 'exchange value' to the object's 'use value,' advertising serves to rob workers and pad the pockets of capitalists with unearned profits.

The cultural critics' picture of advertising as a guarantee of long-term profits is sharply at odds with the way advertising's function is actually viewed in practice. Though the most obvious reason for a manufacturer to spend money on advertising is to stimulate purchase, advertisers are famously skeptical about the actual impact of media budgets on sales. They tend to feel vaguely robbed by their advertising agencies, forced to participate in a pointless display by social convention. And, indeed, one of the few generally accepted findings of the past 30 years of academic research on advertising effectiveness is that a direct effect on sales is virtually impossible to demonstrate. So many factors influence sales

that it is hard to isolate advertising's function to stimulate demand. And, from case to case, advertising's apparent performance can vary from spectacular to abysmal with little readily discernible logic.

Accounting practices reflect the doubts of the manufacturing community: advertising cannot be treated as an investment in capital (like buying a new plant), but must be treated instead as an annual expense (like office paper). Only in the past 5 to 10 years (and only in Europe) has the possibility of recognizing the long-term benefits of advertising in financial statements been discussed. The reason for reconsidering advertising as an investment is the equity markets' valuation of companies with established brands. Consider that shares in Procter and Gamble, the largest advertiser in the US, go for more than four times book value. The price of the stock reflects the financial community's belief that some value inheres in Procter and Gamble beyond the historical cost of its equipment, patents, and real estate. Probably, the financial markets are imputing a value for Procter and Gamble's many brands, most of which are long-standing leaders in their respective product categories. Category leadership is at least partly attributable to Procter and Gamble's historical practice of consistently advertising each brand at levels that exceed all direct competitors' expenditures.

Historians have offered the view that advertising's function is to weaken, even eliminate, the 'middlemen' in goods distribution by stimulating consumer demand for products. Once consumers want a particular brand, they say, retailers are 'forced' to carry it on their shelves and brokers are 'forced' to provide it to retailers. While it is demonstrably true that the growth of advertising as an institution over, say, the past 150 years has resulted in a 'rationalizing' of distribution systems in a way that is advantageous for manufacturers (and, actually, consumers), the notion that advertising allows makers to control distributors should not be overstated. Distributors and retailers retain a great deal of power in the current system and, in fact, during the last 30 years of the twentieth century, the shift has been back toward the distributor. We see evidence of this in the production of private label brands, the growing power of stores like Wal-Mart, and the broadscale shift of advertising dollars spent by companies like Procter and Gamble into expenditures for trade allowances and the like. Further, while it is certainly plausible that advertising has had the effect of increased control over distribution, arguing that channel management is actually the primary function of advertising conflates 'effect' with 'function.'

The conflation of the terms 'effect' and 'function' is also a problem in works that purport to identify the impact of advertising upon social issues. For example, many feminists have written about the effects of advertising on the status of women, frequently falling into the habit of claiming that advertising's function is to oppress women. While advertisements may indeed contribute to the ideology of gender inequality, arguing that advertising's purpose is the oppression of women seems a misplacement of emphasis.

From the broadest cultural perspective, advertising's function may be more meta-structural than most social critics would describe – or advertisers pay for. Anthropologists argue that the primary function of consumption is cultural, not biological, because goods serve to make the categories of culture visible and, therefore, stable. Since human cognition demands a balance between total heterogeneity and total homogeneity, any system of objects will require an activity that marks items as being appropriate for certain users, places, or occasions. In pre-industrial societies, marking is accomplished through exchange rituals performed in the presence of the entire tribe. The chanting, singing, and naming in such rituals can be uncannily similar in form to modern advertising. Given the proliferation of otherwise homogeneous goods in the mass-production society, one would expect to see some mechanism arise for marking these goods as cultural artifacts. Since mass production was accompanied in history by the emergence of mass communication, it is, perhaps, not terribly surprising that the activity of marking mass-produced goods would occur in the public 'spaces' provided by the mass media. Thus emerges advertising 'as we know it.'

Form

Today, advertising is highly identified with images. Particularly in cultural studies, invoking the dangers of the visual is almost synonymous with advertising criticism. But advertising is not necessarily, nor even distinctively, visual in form. Television commercials that use only the visual dimension are as rare as silence on a sitcom – that is to say, rare indeed. Instead, television advertisements, like most other types of advertising, skillfully integrate more than one symbolic system (speech, writing, numbers, music, gesture, and so on). In radio, where the visual dimension is absent, the sound of speech is often elaborated with rhyme and rhythm or augmented with song or 'sound effects.' Furthermore, the largest category of advertisements, directory and classified advertisements, are usually composed entirely of written text. Even here, however, we see nonverbal dimensions brought to bear. The addition of color, use of stylized typefaces, and strategic placement of written messages in all forms of print emphasize something cultural critics try to ignore: that writing is itself an image; the printed word is, by definition, speech made visible.

Writing systems, furthermore, are not always alphabetic in nature, but are sometimes based on picturing. The first shop signs operated on a kind of pictography, even in alphabetic societies, because only the elites were literate. Thus, the convention of hanging the shape of a shoe outside a cobbler's door, or painting a striped pole to designate a barber, or differentiating taverns by the visual representation of their names (the 'Bull in Boots'). Critics who insist on situating advertising in the industrial period argue that such signs were not a form of advertising because they were merely 'information,' an 'announcement' rather than an attempt to persuade. The range of early signs includes many aggressive, even poetic, pictorial statements, however. In England, the signs became so huge and outlandish that they began to cause accidents by falling or blocking the view – in 1762, they were outlawed.

Today, imagistic advertisements still reach into common linguistic forms – proverbs, colloquialisms, jargon – for the sense of their message. In the advertisement in **Figure 1**, for instance, the association between 'duck' and 'duct tape' has no basis in use: one cannot buy ducks from amazon.com, nor can ducks be customers of amazon.com, nor, for that matter, would duct tape be used on a duck. The association here only 'works' because, in speech, the phrase 'duct tape' sounds like 'duck tape.'

Figure 1 Visual use of speech sound similarity.

Figure 2 Adaptation of familiar image.

The multiform nature of advertising is at least partly a result of the ongoing struggle to garner consumers' attention. In the early years of advertising's growth in the USA, most publishers limited advertisements to ten-point agate type. This policy, known as the 'agate-only rule,' was an attempt to equalize commercial competition and a way of saving paper. Inventive advertisers strained against the agate-only rule by printing their messages in patterns, sometimes getting notice through sheer repetition and other times by using the text to form a picture. When, eventually, publishers allowed advertisers to make use of images, the pressure only escalated. Now, advertisers wanted bigger pictures, more distinctive pictures and, especially, color pictures. The same struggle occurred in aural form during the early days of radio. The competition to stand out, be memorable, or gain credibility resulted in testimonials by leading radio stars, jokes woven into the existing program, jingles put in with the music. The Internet advanced in less than 5 years from a text-only medium to a jumbled screen offering pictures, words, numbers, sound, animation, and even video. While the proliferation of formal features on the Internet was certainly made possible by technological innovation, it was also largely fueled by the demands of the new commercial media, the 'dot.coms,' wanting to coax their reluctant customers, national advertisers.

Advertisements have also adopted the formal features of other genres. Early in the century, for instance, the advice columns of the new national magazines in the USA were quickly imitated by advertisers, who used their own 'experts' to give advice that inevitably involved buying a particular brand. Today, advertisements like the one in Figure 2 reach all the way to fifteenth-century Italian art for their inspiration, leading cultural critics to bemoan the 'pastiche' that seems to crop up so often in advertisements. However, advertisements like this one (and others that adapt genres like movies or comics) also testify to the growth of an elaborate visual vocabulary among the reading populace.

As this visual vocabulary builds, even the images of the past become meaningful to the readers of the present. Thus, we see frequent adaptations of older styles reused to specify a past period. For instance, the typeface, background and headline in Figure 3 all communicate 'the late 1960s.' Yet this advertisement ran on the back of *TV Guide* in the autumn of the year 2000: many of those readers would not be old enough to remember the 1950s or 1960s, but nevertheless are clearly expected (by the advertisers) to know the style associated with that time. Achieving specific reference through subtle variation in the manner of rendering is a hallmark of advanced pictographies. Advertisements like these demand a highly developed form of reading from their viewers, as opposed to the kind of mindless absorption that most critics attribute to consumers of the 'image culture' of advertising.

Though references to the visual past appear frequently in advertising today, it is also true that in our own era – as in others before us – most advertisements struggle to look 'modern' (and here I mean 'hip,' 'new,' 'cutting edge,' etc.) The same is true in other dimensions, such as music: the first Pepsi jingle was in the swing style popular on radio in the 1930s; in the 1960s, Pepsi used Motown artists; in the 1980s, Michael Jackson and Madonna sang for Pepsi. As a result of this propensity to operate right at the horizon of a culture's formal style, advertisements are consistently among the most topical,

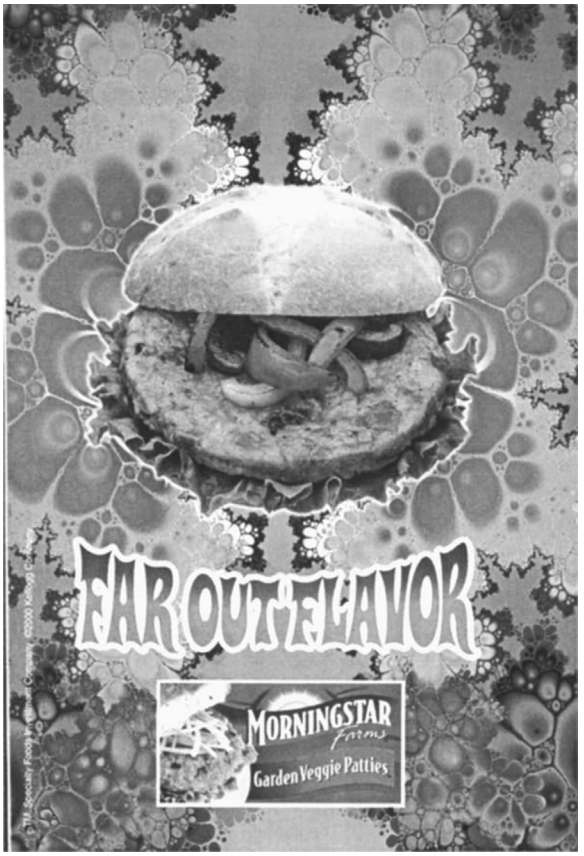

Figure 3 Use of visual style to periodize.

timely, and stylish of cultural texts. Thus, the advertisements of any period tend to be, if nothing else, distinctly 'of their time,' heavily 'periodized.' As Roland Marchand noted, for instance, the advertisements of the 1930s are markedly different from those of the 1920s – more aggressive, less graceful, more pronounced in commercial intent. He attributed the difference to the declining economy. So, not only do the latest visual styles, celebrities, tunes, and dances feature heavily in advertisements, the issues and concerns of the day do as well. This influence may occur in pervasive ways, such as Marchand described, but it also makes itself felt in very specific, identifiable instances. War imagery and patriotic motifs, for instance, are much more common in US advertising between January 1942 and August 1945 than they are in the months either immediately before or immediately after. Fear of disease is the basis for many advertising appeals following the influenza epidemic of 1918; today's advertisements refer to AIDS and breast cancer. Ideology is also reflected: Thomas Frank's *Conquest of Cool* (1997) has eloquently argued that the change in advertising known as the 'Creative Revolution' had its roots in the broader political shift that rumbled through the USA between 1960 and 1970.

If one were to try to theorize an aesthetics of advertising, therefore, a key characteristic might be this tendency to focus on the present, even while evoking the past. Another, which has been noted by authors like Jennifer Wicke, is that advertising is essentially an aesthetics of economy: saying a lot in a small space or a short time is nearly always one of the objectives. Roland Barthes' famous observation that advertising is generally characterizable by its 'frankness' – in the sense that advertisements try to seem natural – is not, in my opinion, supportable. Instead, advertising's need to 'defamiliarize' its propositions has created a corpus of texts notable for whimsy, fancy, outrageousness, and blatant artifice. A good example is a television commercial running now, which refers to the 'animal nature' of a Jeep by having the car shake mud and water from itself like a wet dog, showering its owners in the messy residue. The fundamental impulse to animate an object in the context of exchange – something that is present even in so-called 'primitive' cultures – here runs directly into the expressive capacity of early twenty-first century technology. A simple metaphorical proposition ('a Jeep is like a dog') is transformed, via computer animation, into an extraordinary fiction. The expression of a very basic material proposition in highly imaginative form would also probably be one of the key principles of an aesthetics of advertising.

For all its fictive properties, however, advertising is unlike other art forms in one very mundane way. Advertisements always have 'one foot on the ground' in the sense that they attempt to influence a concrete behavior, usually a purchase. Depending on the particular government under which it is operating, an advertisement may also have to accommodate practical concerns about the product or the status of exchange in some formal way. Commercials for drugs that stimulate weight loss, for instance, often represent attractive possibilities – followed abruptly by disclosure of some rather unattractive side-effects. Many other products, from toys to cigarettes, must include warnings or disclaimers about everything from health risks to whether batteries are included. Regulations on deceptive advertising, at least in the USA, will tolerate a range of fiction and puffery, but the inferences the average consumer makes about the product must be concretely true or the campaign is pulled. What that means, then, is that most US viewers probably do not expect a Jeep to shake like a dog, even after seeing it on TV. In fact, looking at the total range of fancy discernible in advertisements today, viewers must be making, rapidly and on a daily basis, some fairly fine distinctions between the fiction of the advertisements and the facts about the products. This, again, suggests a more skeptical, sophisticated, selective viewer than many who write about the 'manipulation' of the populace by advertisers prefer to imagine.

Raising the specter of an 'aesthetics of advertising' among those who actually make advertisements often revives a long-standing philosophical debate over what 'a good advertisement' is. On one side are those who believe that 'a good advertisement' is one that sells, even if it irritates every member of the audience. Famous admen like Rosser Reeves took this position in the 1950s. Others like Bill Bernbach, who led the Creative Revolution of the 1960s, have argued that 'a good advertisement' is one that is more 'creative' (that is, prettier, funnier, more imaginative, or in some other way a more pleasing work of art), but have also claimed that aesthetically pleasing advertisements actually sell more product anyway. Such positions are hard either to evaluate or defend because of the previously-mentioned difficulty of demonstrating the effect of advertising on sales. These positions are made even more

problematic by the changes over time in (a) what is deemed to be a pleasing advertisement and (b) what is deemed to be a persuasive argument. For instance, in the 1950s, anything that smacked of science gave an aura of 'rationality' to the argument. Today, those same advertisements, with their secret ingredients, graphic cross-sections, and microscopic views, look as irrational as the sci-fi movies of the same period. Several campaigns of the Creative Revolution, like Alka-Seltzer and Volkswagen, still appear on industry 'best campaigns' lists 40 years after they were created. But others, like the Braniff campaign, are a puzzle to contemporary viewers – the Pucci costumes and 'Pop' planes aren't as fresh to today's eye.

Aesthetic notions are always closely connected to a culture's morality. Thus, outside the practical world of advertising, the distinction made between 'good' advertising and 'bad' advertising is often seated in the presumed value of 'information' versus the moral questionability of 'persuasion.' Writers trying to support such divisions among advertisements often use the presence of images as the basis for categorizing: words and numbers tend to be seen as 'informative' or 'rational,' while images are 'persuasive' and 'irrational.' Even an amateur rhetorician, however, can come up with examples in which the display of alphanumeric information (or the withholding of it) is itself an attempt at persuasion. Further, the visual presentation of information – even in the plainest style – affects its persuasiveness. Finally, information is often presented in pictorial form, as in maps and graphs. In many cultures, visual forms are used to record and transmit widely varied types of data. Today, writers like Edward Tufte have made us more aware of the intimate connection between visualization and information, whether our objective is persuasion, 'mere representation,' or facilitating our own understanding. Therefore, the distinction between the informative–written advertisement and the persuasive–imagistic advertisement seems to rest mainly on a naïve understanding of rhetoric and an outdated approach to the pictorial.

How Advertisements Work

There are almost as many theories about 'how advertisements work' as there are advertisements themselves. Several famous admen have published their own treatises on how to make an effective advertisement. There is little agreement among them. Few, however, would subscribe to the theories that have been popular among academics. For instance, Freudian theories popular among cultural critics of the late twentieth century would inspire little more than laughter among those who actually produce advertisements. After a brief flirtation with Freudian notions in the late 1950s (an approach known in the business as 'MR,' which stands for 'motivation research'), the advertising industry abandoned attempts to sell products through subliminal cues, such as phallic symbols or 'hidden' images of genitalia. Such approaches, having proven ineffective, are now thought to be ridiculous. Shirley Polykoff (who wrote the 'Does she … or doesn't she' Clairol advertisements of the 1960s) ridiculed these attempts in her memoirs, published in 1975 (Polykoff, 1975). She recalls that a TV spot introducing a women's shaving cream was changed to accommodate a psychoanalyst's suggestion: after the woman shaved, her legs should rise slowly as she leaned back in her chair, 'a position normally associated with an activity other than shaving.' The product was a bomb: 'It turned out no woman at that time, rising legs notwithstanding, was interested in paying for a shaving cream when she could use her husband's or a bar of soap that was readily at hand'(Polykoff, 1975: p. 98). Cases like this one eventually put the MR people out of business.

Purchase behavior, as it turns out, is better explained by such pedestrian matters as price, practices, and preferences than by phallic imagery. So, in some ways, the subliminal advertising and motivation research fads of the fifties were displaced by a greater understanding of the reasons people consume. In *The Hidden Persuaders*, Vance Packard (1957) accused advertisers of manipulating 'hidden needs' to sell products, sparking a national controversy. Packard's 'eight hidden needs' were then thought to be irrational or even immoral as a basis for buying. Today, these needs – for self-esteem, for safety, for recognition – are no longer particularly 'hidden,' but are popularly accepted as being essential to human mental health. Further, scholars investigating consumption from a more anthropological point of view, like Mary Douglas and Mihalyi Csikszentmihalyi, have now published material emphasizing the complexity of consumption in all cultures, by all humans, and at all times. This literature richly and repeatedly illustrates the myriad of ways that people use material possessions to express themselves, to make connections with others, to protect their families, to enculturate their children, and so on. The tendency in Western post-industrial thought to insist on material life being properly separated from the spiritual or social does not hold up well under the weight of this evidence. A more sophisticated perspective now recognizes that people usually consume for social reasons even when they are satisfying biological needs: we eat dinner with our families because we need both nurture and nutrition. Indeed, the more inquiry is done on the dynamics of consumer behavior, the less currency is held even by the traditional split between needs that are biological (and therefore 'real') versus those that are social (and therefore 'false'). With this in mind, it becomes increasingly difficult to insist that advertising either plays on hidden needs or that it 'makes us buy things we don't need' – because the notion of 'need' becomes utterly redefined.

Elsewhere in academia, however, a different kind of psychology still tries to explain how advertisements work. Experimental psychologists in applied areas like advertising and marketing have tried hard to identify specific formal cues and appeals that will consistently produce sales for advertisers. These researchers design experiments that attempt to determine whether color, happy music, or big type 'work' better to sell us goods than black and white, sad music, or small type. After nearly 30 years of study, not a single generalizable finding has emerged from this research. The new psychological theories of 'how advertisements work,' therefore, are no more likely to win the respect of advertising professionals than the old, Freudian ones.

So, professionals approach the creation of advertisements in much the same way they always have: as a common-sense attempt to persuade their cultural kindred to buy a product.

True, they have a great deal of information about our needs, desires, and dissatisfactions – and they do put that knowledge to use in crafting their messages. To that extent, advertisers may have some edge over the amateur persuaders we meet in everyday life: Girl Scouts who try to sell us cookies, preachers who try to get us to tithe, and offspring who try to make us buy toys. Nevertheless, advertising professionals basically work with rhetorical principles as old as Aristotle: after studying their audience, they offer information and arguments they believe will be convincing. They affect a character they think will match up with the audience's self-image. For 'sweetening' the proposition, they throw in melodies, jokes, and pictures. At the workbench level, this practice, for all the devilry attributed to it, is only a garden variety of rhetoric.

Like any other attempt at rhetoric, some of these practical efforts fail. Some sink dramatically, like Shirley Polykoff's advertisement for shaving cream. Few produce the kind of powerful propellant critics claim for this institution. The public and its pundits, though, continue to believe that advertisers have some secret power to manipulate them. Some of this suspicion is the advertising industry's own fault. In promoting their abilities to press and potential clients, advertising agencies have consistently overclaimed for their own skills. Their bravado, coupled with continued attempts to find magic formulas, makes the public justifiably wary.

Many myths thus continue about the power of advertising, lending fuel to the general discomfort with commercial culture. For instance, the 1950s story of 'subliminal advertising,' in which unsuspecting US movie patrons were persuaded to buy cola and popcorn by flickering messages undetected in the film, is retold today as if it really happened. In actual fact, the whole event was a scam. James Vicary, the man who claimed to have accomplished this feat, was an unemployed market researcher when he reported his story to the press in 1957. While reporters and, soon, Congressmen investigated the validity of his claim, Vicary collected retainers from some of the nation's largest advertisers. But his claim quickly began to fall apart. First, the theater owner said no test had been done on his premises, as Vicary had claimed. Then, Vicary failed to produce the effect for a Congressional committee. Suddenly, less than 6 months after his name first appeared in the papers, Vicary disappeared. His closets and his checking account were left empty. He was never heard from again. Though many researchers have attempted to replicate his study, the purported effect has never been repeated. Even if Vicary was a charlatan, though, the true story does show that advertisers in the USA were willing to pay huge sums to learn how to advertise subliminally. So, however false this cultural myth may be, the suspicion of corporations that lies behind its continued viability is not without foundation.

Conclusion

Advertising seems to touch every aspect of life in the post-industrial world. As both form and institution, advertising is blamed for an array of social ills ranging from the mundane to the millennial. The poor standing advertising has in the global community at the opening of the twenty-first century is not based on the real existence of any secret formula, economic equation, or covert conspiracy. It is certainly not based on any demonstrable effect on prices, social conditions, or even sales. Rather, our attitudes toward advertising are probably a response to the rapid changes in everyday life brought about by industrial commerce, and are intensified by the knowledge that, at least some of the time, advertisers will make extreme attempts to have their way with the public. Nevertheless, the human mind, being far more subtle and sturdy than many theories would suggest, has shown remarkable resistance to all such attempts – the explosion of consumer culture notwithstanding. Our discomfort with advertising is also profoundly situated in ethnocentric prejudices against commerce, material comfort, sensual pleasure, and images. Rigorous thinking on the subject of advertising is rare – at least partly because we are blinded by these very fears and prejudices. Looking forward in a globalizing marketplace, however, should impress upon us the need to break out of such constraints.

Editor's Note

Scott's intriguing entry argues that the effectiveness of advertising is based largely on the power of persuasive rhetoric. A recent extension of this line of argumentation is Tom and Eves, 2014. Much recent literature in the advertising field deals with the effects of advertisements on children, especially food, alcohol and (at one time) tobacco advertisements. See Adams et al., 2012; Kelly et al., 2010; Smith et al., 2014; Hanewinkel et al., 2011.

See also: Advertising Agencies; Advertising: Effects; Advertising: General; Media Effects.

Bibliography

Aaker, D.A. (Ed.), 1993. Brand Equity and Advertising: Advertising's Role In Building Strong Brands. L. Erlbaum Associates, Hillsdale, NJ.

Adams, J., Tyrrell, R., Adamson, A.J., White, M., 2012. Effect of Restrictions on Television Food Advertising to Children on Exposure to Advertisements for 'Less Healthy' Foods: Repeat Cross-Sectional Study. PLoS ONE 7 (2), e31578. http://dx.doi.org/10.1371/journal.pone.0031578.

Barthes, R., 1982. The Rhetoric of the Image, The Responsibility of Forms. Hill and Wang, New York.

Csikszentmihalyi, M., Rochberg-Halton, E., 1981. The Meaning of Things: Domestic Symbols and the Self. Cambridge University Press, New York.

Douglas, M., Isherwood, B.C., 1996. The World of Goods: Towards an Anthropology of Consumption. Routledge, London.

Fox, S., 1984. The Mirror Makers: A History of American Advertising and Its Creators. Random House, New York.

Frank, T., 1997. The Conquest of Cool. University of Chicago Press, Chicago.

Hanewinkel, R., Isensee, B., Sargent, J.D., Morgenstern, M., 2011. Cigarette advertising and teen smoking initiation. Pediatrics 127 (2), 271–278.

Kelly, B., et al., 2010. Television food advertising to children: a global perspective. American Journal of Public Health 100 (9), 1730–1736.

Leiss, W., Kline, S., Jhally, S., 1990. Social Communication in Advertising: Persons, Products, and Images of Well-Being. Nelson Canada, Scarbrough, Ontario.

Marchand, R., 1985. Advertising the American Dream: Making Way for Modernity, 1920–1940. University of California Press, Berkeley, CA.

Packard, V.O., 1957. The Hidden Persuaders. McKay, New York.

Polykoff, S., 1975. Does She … or Doesn't She?: And How She Did It. Doubleday, Garden City, NY.

Presbrey, F., 1929. The History and Development of Advertising. Doubleday, Garden City, NY.

Rogers, S., 1992/1993. How a publicity blitz created the myth of subliminal advertising. Public Opinion Quarterly, 12–17.

Schudson, M., 1984. Advertising, The Uneasy Persuasion: Its Dubious Impact on American Society. Basic Books, New York.

Smith, K.C., Cukier, S., Jernigan, D.H., 2014. Regulating alcohol advertising: content analysis of the adequacy of federal and self-regulation of magazine advertisements, 2008–2010. American Journal of Public Health. e-View Ahead of Print. http://dx.doi.org/10.2105/AJPH.2013.301483.

Tom, G., Eves, A., 2014. The use of rhetorical devices in advertising. In: Blair, K.L., Almjeld, J., Murphy, R.M. (Eds.), CrossCurrents: Cultures, Communities, Technologies. Wadsworth, Boston, pp. 148–155. Chapter 4.

Tufte, E.R., 1990. Envisioning Information. Graphics Press, Cheshire, CT.

Wicke, J., 1988. Advertising Fictions: Literature, Advertisement and Social Reading. Columbia University Press, New York.

Williams, R., 1980. Advertising: the magic system. Problems in Materialism and Culture, 170–195.

Advertising, Control of

Jean Boddewyn, Baruch College, City University of New York, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Like most other human activities, advertising needs to be 'controlled' – that is, its failures must be constrained and its potential to inform people must be optimized. There are two major alternative ways of achieving such control: (1) have the government issue laws and regulations and (2) have the industry develop an enforceable code of conduct. Both approaches have advantages and disadvantages, and the tendency is to blend the two approaches through 'co-regulation.' The co-regulatory systems of advertising control in France and the United Kingdom are analyzed together with technological and social trends.

Advertising is a form of paid communication between a firm and its (potential) customers that uses independent media to communicate positive messages about a good or service. Firms supply it to generate sales and to counter their competitors' advertisements but there is also a demand for advertising because consumers lack information, and ads help lower inevitable 'search costs' – that is, consumers' expenditure of time and money to select what goods to buy.

Advertising Failures

Like other institutions, advertising is prone to natural shortcomings (e.g., information is always incomplete and often asymmetric) and artificial imperfections (e.g., agreements and rules *not* to advertise among professionals such as lawyers). The huge number of laws, regulations, and self-regulatory guidelines reveal that many advertising practices represent actual or potential failures that negatively impact the market system's dependence on adequate information and effective competition, and that impair the functioning of other societal institutions. Consequently, the control of advertising focuses on the use of monopolistic power, consumer misleadingness and deception, unfairness and social irresponsibility.

Monopolistic Power

Producers can erect 'barriers of entry' to better and cheaper products by creating and using meaningless distinctions among brands in their advertisements. Furthermore, they may collude to ban advertising as has often been in the case of legal services but such rarely challenged restraints negatively impact competition and consumer welfare. Even associations that join advertisers, advertising agencies, and the media have been opposed in the United States because they could collude to eliminate certain advertisements.

Consumer Misleadingness and Deception

An advertisement cannot be as lengthy as an instruction manual or even a label but should it include all significant information? This depends on how one defines 'information.' Do people want and need only objective facts such as origin, ingredients, price, performance, and contraindications? Or should information refer to any data or arguments that put forth the attractiveness of a good in the context of a consumer's own buying criteria such as socio-economic background, personality, lifestyle, aspirations, experience, and other factors – whether rational, emotional, or simply habitual? Thus, should government-sponsored lottery ads reveal that the odds of winning are abysmally low, such lotteries pay back the smallest share of any legal game, and the government's take is not fully or mainly used to support education or the arts, as was promised? Or should one accept the appeal of 'All you need is a dollar and a dream' because it reflects the needs of lower-income people? Still, all perspectives on information assume truthful communications although some puffery is tolerated (e.g., The King of Beers), and rules often require that advertisers should be able to substantiate their claims (e.g., The fastest copier).

Unfairness

'Unfair' means 'not sincere, frank, honest, loyal, and right' – all subjective criteria. Unfairness is associated with particular practices such as denigrating through 'comparative advertising' a competing product or firm even if the latter is in fact inferior. Playing on the fears of people may also be considered unfair even when such fears are real (about sickness, death, looks, social status, self-confidence, etc.) and could be alleviated through the advertised products and services (e.g., insurance and cosmetics). Although its legal definition varies across countries, unfairness usually involves: (1) malice (e.g., nobody mentions a competing firm to say good things about it); (2) benefits (e.g., relief) being significantly outweighed by costs (e.g., monetary loss and personal distress), and/or (3) the exploitation of the weaker by the stronger (e.g., large retailers out-advertising smaller ones).

Social Irresponsibility

Advertising should not undermine other institutions such as the state, the family, and the value system. Thus, advertising to children is thought to weaken parental authority and to develop consumeristic attitudes at an age when more important values should be impressed on the young. Larger and

evolving questions about societal welfare and personal happiness are also involved. Thus, as a 'mirror of society,' does advertising reflect and magnify developments in ideologies and lifestyles through new symbols and habits that are detrimental to these goals? Is it even part of a plot to create a 'culture of consumption' essential for the capitalistic machine?

Extent of Advertising Failures

There are no counts of misleading, unfair, or irresponsible advertisements even though as many as 50 million *new* ads may be issued each year in the USA alone, if one includes all national and local media but excludes Internet ads and personal ones – the latter a frequent source of deception (e.g., 'attractive man wants to meet intelligent woman'). The self-regulatory UK Advertising Standards Authority (ASA) once estimated that some 2% of British ads appeared to violate some provision of its extensive Code of Advertising Practice. This is a low rate of malfeasance, compared to other institutions (e.g., the family and the educational system) and considering that much advertising misbehavior results from ignorance and carelessness rather than ill intent. Still, 2% of many millions of advertisements add up to hundreds of thousands of failures worldwide, suggesting a significant need for control mechanisms readily usable by competitors, consumers, governments, and concerned citizens.

Forms of Advertising Control

Theory

Theoretically, the control of advertising is justified by the 'negative externalities' created by false, misleading, unfair, and asocial ads. Its desired outcome is the creation of 'public goods' in the form, for example, of the punishment of infractors, whose benefits are distributed among the populace at large. However, the main problem facing all private-interest groups such as self-regulatory bodies is that of the 'free rider' who benefits from the reputation and services of an association but does not join or participate in it – either by not paying dues or by violating its guidelines. This problem is avoided by the fact that, in the case of advertising, all violations are traceable because all advertisements are visible and nameable by definition so that free riders can always be detected and prosecuted – compared to many other trade violations, which may be hidden as in the cases of collusion (cartels) and the delivery of inferior goods.

The spread of co-regulation (see below) may be associated with the shift from 'imposed' to 'contracted' institutions emerging from below (i.e., the practitioners) instead of the top (i.e., the government) and which are capable of creating social obligations to which powerful actors may be bound outside and beyond their own rational egoism. These contracted rules commit firms to internalize costs – that is, the moral as well as financial costs which it would be beneficial for them to shift unto consumers, the public, and the environment – an evolution currently related to the development of 'corporate social responsibility.' Still, by occasionally constraining the domain of rational-economic action, co-regulation may help forgo opportunities for dynamic marketing innovations as well.

Mechanisms

Most practitioners and organizations behave themselves because they want the esteem of other members of society, they fear losing customers, they feel threatened by the law, and/or they want to lessen uncertainty about their rivals' behavior. The community reacts to advertising's shortcomings by ignoring, discounting, and/or opposing it (e.g., defacing posters and criticizing tobacco ads). Individual and organizational self-discipline, in the forms of personal ethics and company codes of conduct (including media acceptance codes), provides another vehicle for community-based control. Market responses include consumers shunning the goods of unreliable advertisers, and competitors countering ads with their own. State sanctions involve prohibition, restriction, obligation, taxation, and government ads, and they aim to facilitate legal action by customers and competitors. Under industry self-regulation, peers rather than outsiders establish and enforce voluntarily self-imposed rules of behavior although such private governance may be mandated, regulated, and monitored (see below). Since most ads require the use of media (the press, broadcasters, postal services, telephone companies, Internet service providers, etc.), the latter's participation in self-regulation helps put an effective stop to objectionable advertisements.

Respective Strengths and Weaknesses of Alternative Systems

Effective control ultimately requires: (1) developing standards; (2) making them widely known and accepted by practitioners and the public; (3) advising advertisers and agencies before ads are released; (4) pre- and post-monitoring of compliance with the standards through checks and surveys; (5) handling complaints from consumers, competitors, and other affected parties; (6) penalizing bad behavior in violation of the standards, including the publicizing of wrongdoing and wrongdoers, and (7) obtaining legitimacy for the system from the government and the public at large. Somebody has to design, perform, and fund these sizable tasks.

In this regard, community control offers freely shared norms (e.g., against vulgarity) and broadly based pressure (e.g., boycotts) but lacks authoritative means to mobilize social resources beyond what can be obtained voluntarily on account of regard for others. Market competition generates more and better information but is weak on broader social-responsibility issues. Government regulation benefits from universal coverage, compulsion, and legal enforceability although it may also impose high costs, stifle innovation, and be limitedly effective in dealing with complex and evolving issues such as decency and the reach of foreign media. Industry self-regulation can provide customer and competitor protection beyond the minimum standards of the law, and it benefits from the positive commitment of practitioners although industry support of a self-regulatory system is insufficient and unstable in underdeveloped countries.

While the law may engender feelings of personal irresponsibility, effective private rules rest on values and the freedom of those over whom they apply. These rules rely on the autonomy of economic actors that are free and responsible within a moral system which they have created for themselves. However, this

Table 1 Strengths and weaknesses of advertising regulation and self-regulation

Advertising – control tasks	*Government regulation*	*Industry self-regulation*
Developing standards	+ Greater sensitivity to politicized concerns – Difficulty in elaborating standards for taste, opinion, and public decency – Inertia in amending standards	– Greater lag in responding to such concerns – More informed ability to develop and amend standards in these areas + Faster response in amending standards
Making standards widely known and accepted	+ Everybody is supposed to know the law – Compulsory nature of the law generates industry hostility and evasion	– Difficulty in making the public aware of the industry's standards and complaint mechanisms + Greater ability to make industry members respect both the letter and spirit of voluntarily accepted codes and guidelines
Advising advertisers about gray areas before they advertise	– Usually not provided by government	+ Increasingly promoted and provided by industry for free or for a fee, and even mandated in the case of broadcast media and some products (toys, cosmetics, nonprescription drugs, etc.)
Monitoring compliance	± Routinely done or at the prompting of others but often with limited resources	± increasingly done by the industry although restricted by financial resources
Handling complaints	+ Impartial treatment is anticipated + Potential to handle many complaints – Slow and expensive when using the courts – Cannot put the burden of proof on advertisers in criminal cases	– Treatment more likely to be perceived as partial – Limited capacity to handle many complaints in some countries + Faster and less expensive + Usually puts the burden of proof on the challenged advertiser
Penalizing bad behavior, including the publicity of wrongdoings and wrongdoers	+ Can force compliance – Generates hostility, foot-dragging, appeals, etc. – Limited publicity of judgments unless picked up by the media, competitors, and activist groups	– Problems with noncompliers but the media usually refuse to carry ruled-out ads + More likely to obtain adherence to decisions + Extensive publicity of wrongdoings and, to a lesser extent, of wrongdoers
Obtaining legitimacy for the system from the public at large	+ The state is expected to act in the public interest	– Private interests are suspected of favoring industry at the expense of customers + Increasingly perceived as a necessity for self-regulatory control

(+) refers to strengths, and (–) to weaknesses.

system must be open to its external stakeholders' demands for participation, impartiality, objectivity, transparency, and accountability – as backed up by the state, if necessary. However, government participation in rule-making and sanctioning changes the system into one of 'co-regulation' whose equilibrium may be precarious because it challenges its members' voluntary internalization of certain values, participation, and support. Table 1 summarizes the respective strengths and weaknesses of the regulatory and self-regulatory systems which dominate advertising control.

Factors Affecting the Use of Controls

The demand for controls reflects, first, the fact that competitors want to protect their property rights in brands and reputations (e.g., against comparative ads) and to ensure fair competitive rules (e.g., against bigger advertising subsidies granted to large distributors) and; second, that customers and their associations are concerned about consumer deception, advertising unfairness, and social irresponsibility. On the supply side, governments are increasingly tackling such issues through regulation and the support of stronger market competition and self-regulation. Industry has offered self-regulation as a way of preempting or mitigating public regulation and of improving the overall credibility of advertising. The proliferation of national and subnational states as well as of supranational bodies (e.g., the European Union and the World Health Organization) has multiplied mandatory rules and advisory guidelines which have facilitated complaints, the stopping of ads through cease-and-desist injunctions, private suits and even class actions by empowered competitors, consumers, and citizens.

Regulation and self-regulation are often complementary. Thus, all voluntary codes state that "Advertising must be legal" although very few self-regulatory bodies (e.g., the UK Advertising Standards Authority) pursue statutory infractions. Sometimes they specialize, as with the German ZAW body focusing exclusively on taste and decency because the law against unfair competition applies to most other problems. Governments burdened by growing tasks and budget deficits increasingly invoke the principles of 'proportionality' and 'subsidiarity' whereby higher control levels should not deal with what better informed and motivated lower levels can achieve more effectively. Even consumerist groups may agree to this approach, provided they can meaningfully participate in self-regulation.

These developments reflect better knowledge about how advertising works and about the effectiveness of controls. Thus, research has confirmed that: (1) market shares keep shifting when comparative advertising is allowed, with the latter practice helping reduce consumer search costs; (2) government restrictions often result in higher prices and lower quality of

services, and (3) the cost of preventing and correcting all harm to consumers, competitors, and citizens can be excessive. In this regard, the US Supreme Court now requires that, when public policy justifies government controls, the latter be shown to be effective and not disproportionate to the goal to be achieved.

Both regulation and self-regulation are being fostered and assisted by supranational governments (e.g., the European Union's directives on misleading and comparative advertising), international organizations (e.g., the World Health Organization's generation of a global treaty to control tobacco sales and advertising), and industry groups (e.g., the International Chamber of Commerce with its International Code of Advertising Practice, and the European Advertising Standards Alliance (EASA) and its promotion of self-regulation within the European Union).

Still, the existence of some 200 nation-states with varying traditions, values, legal systems (common law vs civil or religious laws), economic resources, and advertising experiences precludes the standardization of controls. Besides, industry conventions and public regulations are culture-specific and not readily transferable to very different political and socioeconomic environments. In particular, if distrust largely prevails among people, self-enforcing voluntary mechanisms are unlikely to work.

Impacts

Self-regulatory bodies handle up to 30 000 complaints a year (with some duplications about the same ad) when they are actively solicited as by the UK Advertising Standards Authority, and the US Federal Trade Commission received 1.8 million of them in 2011 although this number includes many fraudulent and anticompetitive misbehaviors rather than advertising ones. However, the number of adjudicated cases can be relatively low (about 200 a year by the very active US National Advertising Division of the Council of Better Business Bureaus) when these bodies choose to focus on exemplary cases designed to send strong signals or break new grounds. Government agencies are also very selective in their interventions for the same reasons and because of budgetary constraints although courts are increasingly handling advertising cases in developed countries.

Contract law potentially protects customers everywhere against misrepresentation in advertisements. Private suing to stop advertisements or for damages has become common although usually done more by competitors than by consumers. The latter are slowly being empowered to bring class actions against companies whose products cause injury (e.g., cigarette advertisers). Under statutory law, certain practices are forbidden, circumscribed, or mandated (e.g., health warnings) and criminally punishable but public regulatory bodies can also issue civil-law-based cease-and-desist injunctions as well as require corrective advertisements. Self-regulatory codes are being expanded to include more products (e.g., toys and food) and services (e.g., advertising by charities) as well as new media (e.g., Internet advertising) and novel claims (e.g., about environmental effects).

How effective have regulation and self-regulation been in promoting truth, accuracy, fairness, social responsibility and competition? No precise evidence exists, and impacts vary depending on a country's level of development as well as on its priorities. There will always be crooks so that control mechanisms can only aim at improving the overall quality of advertising at a reasonable cost through exhortation, coercion, punishment, denial of access to the media, and other mechanisms. Besides, new problems keep emerging on account of technological innovations (e.g., the Internet and satellite broadcasting that ignore national borders and regulations), new entrepreneurial initiatives (e.g., sports sponsorship and political advertising), and citizen activism (e.g., growing concerns about the environment, children advertising, obesity and privacy invasion).

One can reasonably conclude that such problematic practices as unsubstantiated claims about weight reduction and other cures have been significantly curbed in developed countries through regulation and self-regulation. Besides, the principle that advertisers have a 'prior reasonable basis' before making a claim has been broadly adopted and enforced in many states. More ads about particular goods and in certain media (particularly, broadcasting) must be approved in advance and certain claims (e.g., 'free, low-fat, and environmentally sound') are now restricted while mandated information (e.g., annualized interest rates in financial ads) is growing.

However, controls have also curbed competition and denied valuable information to consumers, as with bans on pharmaceutical, liquor, medical products, and legal services advertising in many countries. Besides, protecting a lengthening list of 'vulnerable groups' (the young, the sick, the recently bereaved, the poorly educated, etc.) shrinks the pool of those targetable by advertising.

Ongoing Developments

Technology

Means of reaching consumers keep multiplying in such forms as satellite broadcasting, international editions of newspapers, magazines and catalogs, and the Internet. However, technology also allows consumers to avoid ads (e.g., with zapping devices) and to access them only voluntarily and selectively, and it facilitates the fast sharing of information among consumers, associations, and governments as well as the filing of complaints. The Internet increasingly links advertisers, customers, and citizens via computers, television sets, and telephones as well as other receivers and processors. Potentially, there will be many more millions of Internet ads that can be quickly reformatted as well as many more advertisers, including individuals. The line between editorial content and advertising message, which have traditionally been separated, will become blurred because there will be many 'banner' ads as well as sponsors linked to entertainment and information programs. Advertising agencies that help screen ads will be less involved to the extent that some of their services are now available from Internet-service providers for free or at a low cost. Surprisingly, the Internet has generated a complementary advertising controller – namely, bloggers who comment on current ads and even recirculate banned ones! Blogging allows people at large to share opinions and to applaud, deride, or criticize advertisers' campaigns.

Another major current issue concerns 'online behavioral advertising' (OBA) which refers to the collection of data from

a particular computer or device regarding Web viewing behaviors for the purpose of using such data to predict user preferences or interests and to deliver advertising to that computer or device, based on what has been inferred from such Web viewing behaviors. Governments and industry are working on what transparent information should be given to consumers about how OBA works and what action they can take to block it if they so choose.

Current Issues

Every day sees the launch of non-governmental-organization (NGO) campaigns and government legislative proposals directed at restricting marketing and advertising practices. Much of it is driven nowadays by concerns about obesity and the perception that advertising plays a negative role in encouraging unhealthy diets and lifestyles. This issue is on the global policy agenda of the World Health Organization, the US government, the European Union, and many other industrialized countries. The pressure is building to find simple regulatory solutions to what are, in fact, complex and wider societal problems. Media visibility of the obesity issue is high as is the pressure on individual food companies. For some firms, it threatens their license to trade.

Yet, it is important to realize that advertising and its control have low 'salience' in the minds of consumers. For example, a 2008 survey conducted by the UK's Advertising Association, which groups all major members of the advertising industry, found out that people are much more concerned about the current cost of living (81% of respondents), crime, law and order (76%), and health and welfare issues (68%) than about greater government regulation of advertising (15%) which is viewed by 64% as a good way to find out about new products. Besides, 73% of people know that advertising is controlled by rules and regulations while 79% realized that there were UK organizations overseeing the industry – that is, that self-regulation existed. Actually, 47% of the respondents had themselves used advertising to find something they were looking for in the last 6 months!

Laws and Codes

Regarding the right to give and receive information, national constitutions usually guarantee freedom of the press because democracies need independent media to challenge governments in 'the marketplace of ideas.' However, does this right cover what is called 'freedom of commercial speech?' Since the 1970s, the US Supreme Court has come to acknowledge such a right also supported by the Council of Europe's Declaration on Freedom of Expression and Information, and by Article 10 of the European Convention on Human Rights. However, this new freedom is limited because it conflicts with growing 'rights' to health, privacy, and sex equality – for example, pharmaceutical advertising remains restricted because it may lead to greater health expenditures by governments on account of greater demand for drugs, which can be incited by ads.

Regulation and self-regulation will proliferate at industry, subnational, national, and supranational levels, thereby making their harmonization more difficult. Such heterogeneity will complicate the creation and delivery of ads although large advertisers can more readily comply with varying rules. It also creates loopholes to advertise from where restrictions are lower, and to sue where litigation is most likely to succeed. Besides, more controls encourage the use of imagery ('Things Go Better with Coke') in lieu of challengeable claims ('Canon prints faster than HP'). When ads emanate from another sovereignty, both regulation and self-regulation face the problem of 'conflict of laws' in terms of what body, domestic or foreign, decides the case, what procedures (e.g., about required evidence) to follow, what law or code is applicable, and how to recognize and enforce decisions. Governments now favor the laws of the ad's country of origin, and so does the self-regulatory EASA as well as international codes for pharmaceuticals and direct marketing.

Regarding the Internet, regulatory control will focus on: (1) criminalizing the sending of certain advertising contents (e.g., pornography and pharmaceuticals); (2) having advertisers and service providers notify ad receivers of private data-collection practices (particularly from children) and of their rights to check these data's accuracy and to be assured that such information is secure, and (3) obligating access providers (including telecommunication companies) to at least respond to complaints since it is impossible for the latter to screen out millions of messages. Hence, the Internet will test and refine the control of advertising in coming years.

The Development of Co-Regulation in France and the United Kingdom

There is a trend in major developed countries toward some form of 'mandated self-regulation' whereby both rule-making and -enforcement are largely left to industry bodies. This system differs from pure voluntary self-regulation insofar as the private regulatory system is officially sanctioned by some government agency which monitors the program and, if necessary, takes steps to ensure its effectiveness. Thereby, nonofficial norms become part of the legislative order by insertion into statutes or by publicly acknowledging in various ways that the product of private negotiations is generally binding on the whole sector of advertising. France and the United Kingdom have made recent moves toward advertising co-regulation although their structures and processes differ significantly.

French Co-Regulation

French legislators have been reluctant to deal with freedom-of-expression issues while their laws bearing on communication and the media have often been ill-conceived and worded so that the French executive government has been happy to delegate rule-making to the self-regulatory system.

However, to be seen as credible and efficient as a 'moralizer of markets,' professional self-regulation must guarantee the effectiveness of its rules through a monitoring system and periodic assessment of its activities. This requirement is particularly important in France where advertising and its practitioners are widely distrusted on account of what are believed to be misleading ads and a shocking lack of values as well as assaults on vulnerable groups (e.g., children). In this context, the private enacting of rules and having them respected by 'professionalized' practitioners must be accepted by the

public – that is, legitimized. This issue manifested itself first in the matter of taste and decency.

The French ARPP (Autorité de Régulation Professionnelle de la Publicité/Professional Advertising Regulation Authority) had faced endless criticisms from consumer and feminist organizations whose ire was particularly directed at ads and commercials where sex was used inappropriately (e.g., being sexist or discriminatory), excessively, or in ways exceeding moral tolerance levels. In response, it developed the criteria of 'the dignity of human beings' and of 'non-discrimination,' it wrote Recommendations on them and it committed itself to report annually to the appropriate ministry about its pursuit of ads violating its guidelines.

Going further, its leadership created in 2008 three new committees including independent members from outside the advertising industry: (1) an advisory Advertising Ethics Council chaired by an independent academic to 'anticipate' new societal developments; (2) an Advertising Parity Council, of which half the members represent consumer and environmental associations and which 'concerts' with industry representatives about the need for new self-regulatory rules, and (3) an independent Advertising Deontology Jury made up exclusively of people with no links to industry or consumer associations in order to solicit and 'sanction' complaints from the public as a complement to the ARPP's monitoring of advertisements. However, new 'Recommendations' have to be approved by a board made up exclusively of industry representatives because, otherwise, they are no longer voluntary and cannot be truly internalized by the ARPP's members.

The Advertising Deontology Jury, which solicits and handles complaints from the public, reported for 2010, 47 valid complaints, 23 of which were related to the protection of human dignity and, in 14 of these cases, the complaint was upheld. Still, public complaining and negative jury decisions are likely to persist because viral advertising on the Internet and word-of-mouth diffusion have created a huge recirculation of advertisements with sexual and violent content.

ARPP sanctions include the publicizing of wrongdoings and wrongdoers ('name and shame') and wide press publicity of its decisions. Members cannot be excluded since adhesion is voluntary but very few culprits challenge the Jury's decisions. Fortunately, most French media belong to the ARPP and, therefore, are committed to stop the diffusion of the proscribed ad after a negative decision of the Deontology Jury which accepts complaints from everybody: consumers, competitors, and members of the public. Besides advising practitioners before ads are issued, the ARPP also monitors large random samples of ads and reports annually its findings to the appropriate ministry. For example, only 0.7% of monitored ads were found to violate its Recommendations in 2009 although the Internet incurred a much higher proportion of violations than other media.

The ARPP system of 'professional regulation' has been publicly recognized in several ways. Thus, a 5 March 2009 law, which transposed into French legislation the recent European Union directive on audiovisual services, officially authorized the government to delegate the preclearance of television commercials to the ARPP. Subsequently, a government agency in charge of controlling the media delegated the task of screening commercials before their diffusion to the ARPP but the French government has also left to it the control of ads in such sensitive areas as taste and decency (see above), ecological claims, and food advertising.

In addition, the Paris Appeals Court stated on 26 October 2010 that "recommendations from the ARPP, even though they have no legal character, are professional practices that the judge must take into account if they do not contradict a legal statutory measure."

Altogether, professional regulation is now acknowledged and accepted by the French government which, through several Commitment Charters, has implicitly agreed *not* to regulate or ban certain practices but requires in exchange an effective system of adequate guidelines as well as an accountability evidenced by periodic and transparent monitoring and reports. These agreements amount to a system of co-regulation between public and private actors who concert and collaborate in the public interest, and help generate a sense of responsibility among advertising professionals now convinced that their industry cannot claim its freedom of commercial speech if it cannot prove its responsibility.

British Co-Regulation

The ASA is the British advertising self-regulatory body whose aim is to insure that ads be 'legal, decent, honest, and truthful to the benefit of consumers, business, and society,' and it is the most active one in the world. Thus, the ASA solicited and received over 30 000 complaints about ads in 2011, including over 10 000 of them on company websites and such online spaces as social networks and video on demand, which these firms control. Some 87% of the complaints pertained to small and medium enterprises – a percentage suggesting that many advertising misbehaviors reflect a lack of knowledge of the standards rather than malevolence. Clearly, the task of making standards known, understood, and internalized is an essential but endless one!

The ASA sponsors twice-yearly independently-run consumer surveys as well as awareness campaigns and focus-group meetings with parents, guardians, and children in order to better understand their concerns which are mainly related to the commercialization and sexualization of childhood. It provides copy advice before advertisements are released – on some 140 000 occasions in 2011 when its website was accessed by 750 000 visitors. All of these services are based on the principle that prevention is better than cure.

The ASA Council is the jury that decides whether advertisements have breeched the ASA Codes. Led by a notable British peer, two-thirds of this Council's 14 members are independent of industry while the remaining members have a recent or current knowledge of advertising or the media. It includes one or two younger members more familiar with contemporary lifestyles but mainly representatives of families, charities, and consumer groups. Handling complaints generally takes from 35 to 85 days and the decisions of the ASA Council can be appealed to an Independent Reviewer who can order the Council to reconsider its decision (close to 50 of them in 2011). The ASA has set itself high standards of service and it reports regularly on such goals as accessibility to members of the public and of the advertising industry, speed in handling complaints, user satisfaction, high-quality professional service, transparency, and accountability.

A large and growing staff assists the various rule-making, advisory, screening, monitoring, sanctioning, and appeal functions of the ASA which is financed by an automatic levy of 0.1% of all advertising expenditures – a sound financing system only emulated in Ireland and the Netherlands, and which has allowed the ASA to cover the vast media of Internet advertising and company websites.

In 2004, after more than 40 years of successful independent self-regulation of nonbroadcast ads, the ASA assumed responsibility for TV and radio ads after the newly formed Office of Communications (Ofcom) took the decision, in a move supported by Parliament, to contract-out responsibility for broadcast (TV and radio) advertising to the ASA system in a co-regulatory partnership. This agreement created for the first time a single regulator for advertising – a one-stop shop for advertising complaints, which is particularly important for members of the public who want a complaints system that is easy to navigate.

While the advertising member associations of the ASA must approve the changes in their main codes – the UK Code of Non-Broadcast Advertising, Sales Promotion, and Direct Marketing and the UK Code of Broadcast Advertising, Ofcom has the right to single-handedly change these rules in the case of broadcast advertising when there is a specific statutory duty to work toward a public-policy goal that markets alone cannot achieve.

It is generally agreed that a state regulator operating with backstop power is necessary for a co-regulatory system to work properly. Besides, third-party advice must be solicited by the ASA from consumer experts, civil-society organizations, and the public at large – just as the French do. Even when control is not completely in the hands of the national government, the inclusion of external stakeholders broadly representative of civil society (consumers, families, women, etc.) as partners in major functions of the self-regulatory system is crucial in terms of avoiding capture of the regulatory state by industry.

In 2007, new Ofcom rules pertaining to the broadcasting of food commercials forbade their use before 9.00 p.m. on television channels devoted to children's programs or those programs likely to be of particular appeal to children below the age of 16 in the case of food or drink assessed as high in fat, salt, and sugar under the principle that ads should not condone or encourage poor nutritional habits or unhealthy lifestyles in children. Advertisers had to comply with these unilaterally imposed rules which resulted in significantly reducing advertising during children's airtime, using ads with mothers and family rather than child appeal, eliminating promotions based on licensed characters (e.g., Tony the Tiger for Kellogg's Frosted Flakes), excluding toy inserts in packages, stopping websites designed to appeal to children, and removing under-16s from company databases. Besides, multimedia campaigns became complicated when what can be done in print advertising cannot be done on television. It is also worth noting that 'children' in other countries usually comprise only those under the age of 11.

Through its member associations, the ASA can have firms expelled from membership, it may impose mandatory pre-vetting (that is pre-checking and approval of ads) on persistent offenders, and it can refer rule violators to the Office of Fair Trading which can seek an injunction from the Court to prevent an ad's further appearance. Independent compliance surveys regularly reveal that more than 97% of ads are in line with the ASA advertising codes.

The British ASA and French ARPP were two of the founders of the EASA located in Brussels, which is the single voice of the advertising industry in the European Union as far as advertising self-regulation is concerned. Among its wide range of operations, EASA acts as a focal point for cross-border complaints so that consumers need to complain only to the self-regulatory organization in their country, no matter where the marketing communication originated. The EASA is a source of information and research on self-regulation, it helps develop new self-regulatory bodies in Europe, and it liaises with the European Union's numerous legislative and executive units. It has generally been rated as very effective by the relevant EU departments in charge of consumer protection.

In conclusion, consumers enjoy advertising's entertainment value and often make use of the information it provides but it is prone to a number of significant failures which public regulation, private self-regulation, and co-regulation have effectively addressed in order to secure high standards for the protection of businesses, consumers, and the general public.

See also: Advertising Agencies; Advertising and Advertisements; Advertising: Effects; Advertising: General; Audience Measurement; International Advertising; International Communication: History; Media Effects on Children; Media Effects; Public Broadcasting; Violence and Media.

Bibliography

Boddewyn, J.J., 1989. Advertising self-regulation: true purpose and limits. Journal of Advertising 18 (2), 19–27.

Boddewyn, J.J., 1992. Global Perspectives on Advertising Self-regulation [38 Countries]. Quorum, Westport, CT.

Boddewyn, J.J., 2011. The control of 'sex in advertising' in France. Journal of Public Policy & Marketing 30 (2), 220–225.

Calfee, J.E., 1997. Fear of Persuasion: A New Perspective on Advertising and Regulation. Agora, Maraz, CH.

European Advertising Standards Alliance, 2010. Advertising Self-regulation in Europe and Beyond. EASA, Brussels.

Gunningham, N., Rees, J., 1997. Industry self-regulation: an institutional perspective. Law and Policy 19 (4), 363–414.

Petty, R.D., 1997. Advertising laws in the United States and European Union. Journal of Public Policy & Marketing 16 (1), 2–13.

Schudson, M., 1984. Advertising, the Uneasy Persuasion: Its Dubious Impact on American Society. Basic Books, New York.

Teyssier, J.-P., 2012. Advertising regulation and self-regulation in France. In: Shaver, M.A. (Ed.), Advertising Regulation and Self-Regulation Around the World. ME Sharpe, White Plains, New York.

Vakratsas, D., Ambler, T., 1999. How advertising works: what do we really know? Journal of Marketing 63, 26–43.

Wotruba, T.R., 1997. Industry self-regulation: a review and extension to a global setting. Journal of Public Policy & Marketing 16 (1), 38–54.

Advertising: Effects

Thomas C O'Guinn, University of Illinois, Urbana–Champaign, Champaign, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 1, pp. 183–188, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

This article provides a review of the advertising effects literature. It begins by putting forward three reasons why so little knowledge about the effects of advertising exists. They are: an anticommercial bias, the inherent difficulties presented by the object of study itself, and inadequate methods. Effects at the economic, sociocultural, and individual level are presented and discussed. These findings are set in a social–historical context.

Introduction

The phenomenal growth of advertising in the twentieth century helped usher in and sustain a defining consumer culture. That is advertising's surest effect. Consumers come into contact with literally hundreds of ads every day. This ubiquity has its effects. People use phrases and other utterances from ads in daily discourse (McKay et al., 2011; Friedman, 1991; Ritson and Elliott, 1999), revealing more than mere mimicry, but also the ad-inspired consumer ethos of contemporary consumer culture. Advertising is now part of the interstitial tissue of daily life on this planet. Yet, the effects of advertising are often anything but clear or easily detected (Spielmann and Richard, 2013).

In fact, the effects of advertising run the gamut from obvious to perplexing and contradictory. On the one hand, most people, most of the time, don't care much about ads. They don't pay attention to them. But every now and then, they do. And even when they don't, the effects may still accrue.

Context

Advertisements have been studied in one way or another for about 100 years, yet we simply don't know much about their effects. There are at least three good reasons for this lack of knowledge. First, scholars have viewed advertising as unworthy of attention. This is the same high culture (and anticommercial) bias which led libraries to cut ads out of magazines before binding them for decades. Many libraries gave space to pornography but not advertisements. It was politically difficult to propose studying something that was regarded as thoroughly unworthy of study. This bias has had a profound effect on this scant scholarly enterprise. Relatively little research has been done, and much of it is more polemic than systematic. Second, the object of study itself presents an inherent difficulty. Advertisements are by their nature textual, generally ephemeral, and encountered as a common social object, situated within layer upon layer of social discourse. They are here and then they are gone. They are commercial rhetoric, actively interpreted by various audiences in various ways. While the general meaning of an advertisement is informed and bounded by cultural forces so as to produce more likely than less likely interpretations, there is no single, objective advertisement to study (Fish, 1980; Scott, 1994). What an advertisement is depends on who is interpreting it and why. This seemingly simple point has thoroughly eluded most advertising researchers. Further, ads exist in their particular social–temporal microclimate and then they become something else (Lee and Joo, 2012). Even the once-legendary 'Man from Glad' is not now what he once was. Typically stripped of context, we have little idea of what advertisements meant in their own milieu.

The third reason for the paucity of work in this field is inadequate method. It is extremely difficult to demonstrate the effects of advertising, and the job is made even more difficult within the confines of a single method, a single literature, a single paradigm. Survey data suffer from the scores of variables that lie between exposure and behavior. Experiments suffer from low ecological validity; they too often have little or nothing to do with how real people experience and interpret real ads in real contexts. Ethnographic work on advertising has been rare, but suffers from its lack of generalizability, the effects of intrusion, and the privileging of informants and text. Textual analysis, and even reader response, suffers from the authority of the one constructing the interpretation. Work by economists with aggregate data suffers from a long and troubling list of assumptions, particularly those positing rational thought. Advertising is a sociotextual phenomenon, which does not travel well outside its natural environs. In their natural space advertisements are a typically trivial background to daily life. Methods sensitive to this reality have yet to be employed to any significant degree.

But advertising certainly must have effects. How could something so widespread, so much a part of contemporary existence, and something which businesses worldwide continue to invest in massively, have no effect? Beyond the reasonable and predictable retort that 'real' effects are not necessarily detectable and demonstrable effects, there must be something to which we can point and say: 'we (probably) know this.' The more robust advertising effects findings are reviewed below.

Economic Effects

Economists have studied the effects of advertising spending on various outcome variables such as GDP, aggregate and

category demand, brand switching, economic concentration, and barriers to entry. Borden (1947) was one of the first to attempt a systematic economic analysis. While acknowledging that advertising might increase 'selling costs,' (p. 881) his study points to the vital need for advertising in growing 'dynamic economies' (p. 881). He sees little downside to advertising. Running counter to this view are both Packard (1960), and Galbraith (1958), who see advertising as the engine of waste and the loss of consumer agency. These are not, however, empirical examinations, but more economic–philosophical theories. They represent the political left's view sans data.

Schmalensee (1972) conducted the next major empirical analysis of the economic effects of advertising. His work showed "that total national advertising does not affect total consumer spending for goods" (p. 85). He argues in a very real and practical sense that sales cause advertising, much more than the other way around. He goes on to say that a major implication of this is that "advertising cannot be all-powerful if it cannot influence the decision to save." He finds national advertising to be particularly inconsequential in driving aggregate demand. In terms of competition, he says: "the effects of national advertising spending by the various consumer goods industries cancel each other out in the aggregate" (p. 86).

The next major exhaustive empirical examination of advertising's economic effects is that of Albion and Farris (1981). Their work revealed a much more complex and nuanced set of findings than previously reported, but were generally supportive of Borden (1947) and Schmalensee (1972). In other words, most of the criticisms of advertising were not borne out by the data. They do, however, find advertising expenditures and industry concentration to be correlated, but argue that other factors, as well as reverse causality, may be at work (Kim and McAlister, 2011). Most significantly, they note the overwhelming importance of context in these effects. Schudson (1984) also makes the case for context and the limited power of national consumer goods advertising. He argues that advertising follows affluence, targeting consumers who already consume large quantities of the goods or services concerned. He gives good examples of both advertising without sales, and sales without advertising (Kim and McAlister, 2011). Luik and Waterson provide one of the more recent (1996) reviews of this literature and find that advertising affects aggregate demand early in a product's life, but as time goes by, has more of an impact on brand share, and less of an impact on aggregate demand. That is, for most mature product categories, advertising seems to impact on brand share, but not total category demand.

Sociocultural Effects

The cultural effects of advertising have largely been the purview of those in the humanities, as well as those in areas of communications, sociology, and anthropology. More recently, a few in marketing have contributed as well. A wide range of orientations and methods has been brought to bear on the search for advertising's effect upon societies and culture. This in fact constitutes the single largest (and most widely accessible) literature on advertising effects.

These works claim that advertising has produced several effects at the sociocultural level. For one, advertising is said to have caused a more materialistic culture (Ewen, 1988; Fox and Jackson Lears, 1983; Pollay, 1986; Richins, 1991). This argument rests on the traditional modernist critique: the real has been replaced with artifice, community with mass society, and a human orientation with one in which material things are more central to human existence. By extension, it is claimed that advertising has placed more emphasis on brand names as symbols than on the actual sum and substance of goods and services. This privileges simulacra over the 'real.' Likewise, advertising is said to have worked as a hegemonic force to the detriment of women, minorities, and other marginalized peoples (Faludi, 1991; Seiter, 1995). These are the major effects claims at the sociocultural level. Although there are many other derivations, they are largely subsumed by those categories outlined above.

The evidence for these claims is, however, rarely as strong as the accompanying rhetoric. Still, there are a few evidence-based conclusions. First, advertising has certainly made consumer culture possible. This may be the biggest advertising effect of all. To understand this, one has to appreciate advertising's role in advancing the practice of branding. Advertising made possible national brands, without which consumer culture would have died on the vine. Advertising projected brands and brand consciousness into national consciousness and into greater social centrality (Muniz and O'Guinn, 2001). Advertising made a brand-oriented society possible – perhaps inevitable. In 1850, few things were branded; by the end of the twentieth century, water and dirt were branded. The whole purpose of a brand is, in economic terms, to create greater inelasticity. A primary predictor of inelasticity is the consumer's belief that there are few acceptable substitutes. For example, when advertising made Ivory more than just soap, demand for Ivory became less elastic than demand for mere 'soap.' The same thing happened with beer, and soft drinks, and thousands of other products. So, clearly advertising lead to a proliferation of brands, which contributed to a more brand-oriented society. The brand is now a defining construct in contemporary consumer culture (Lindenman, 2010). This is an effect of advertising. While societies have always been involved with material culture, the massive extent to which branded culture emerged is quite clear, and not just a matter of degree. This is an advertising effect. Whether or not this has caused individuals to value things more than before is unknown, although it does seem clear that it probably produced more value for the explicitly (and highly marked) commercial object. Whether it caused individuals to value things more in relation to humans than previous generations did is a much harder case to make.

What then of advertising and hegemony? Advertising did and does provide support to modal social structures. Advertising, by its nature, is generally (but not always) a supporter of the status quo. Most researchers do not find advertising to be a great friend to progressive social movements, although it certainly has been from time to time. Unfortunately, many of the arguments regarding the hegemonic dynamics of advertising are hyperbolic and ahistorical (Scott, 2002). Further, the relationship between progressive movement, resistance, consumer agency, and advertising is far more complex than typically held (Muniz and O'Guinn, 2001). Consumer

resistance is now played out by the rejection of one brand (perhaps an ecologically friendly, green brand) over another brand. Belief in a 'star chamber' of corporate elite, or advertisers who control consumer's minds, is being thoroughly discredited (Scott, 2002) by social history research.

Advertising's effects in the areas of race and gender are far from simple, but have generally been negative. There is every reason to believe that narrow, pejorative, and stereotyped social portrayals have had detrimental effects when it comes to people of color (Grady, 2011). The same is no doubt true in the case of women, gays, and lesbians (Taylor et al., 2013). The evidence for this is more inferential than direct, but is consonant with empirical studies in other domains. We know that these groups were absent, stereotyped, or under-represented in typical advertising. Exactly what the effects are on self-perception and other types of perception is less well understood. We can assume it to be anything but positive, however.

In fact, social constructions of reality in general have been shown to be influenced by mass media and advertising portrayals of social reality (O'Guinn and Shrum, 1997). Recently, O'Guinn (unpublished manuscript) has shown that those exposed to a great deal of advertising on television believe there to be higher than actual rates of 'ad-problems' such as gingivitis, athlete's foot, bad breath, etc. There is good reason to believe that advertising has a significant impact in terms of simply making certain things more heavily represented in consumer's constructions of social reality.

Individual Effects

A considerable amount of advertising 'effects' research focuses on the individual. Most has been published in two academic fields: mass communication and marketing, typically from a psychological perspective.

In the 1940s researchers became interested in the effects of mass communication in general. Given the post-World War II consumption boom, advertising was of obvious relevance. The working model of mass communication was at that time a very simple direct effects model: message strikes audience member like a bullet. Experience with propaganda efforts in the war, the ascendancy of psychology in both the academic and larger social world, and outright naiveté led to this very simple notion of hypodermic needle-like effects (O'Guinn and Faber, 1991). There was a great deal of belief in the power of communications at a critical point in the social history of advertising. Exuberant communications researchers were going to wipe out bigotry, inform the electorate, and make things better in all aspects of life, including the realm of modern consumption. This belief was a popular and self-serving one for several groups, including the advertising industry itself.

It is not coincidental that the postwar period is literally drenched in Freudian and neo-Freudian notions. With the mind-control hysteria of the cold war, the prominence of psychology as a science and references to the subconscious mind led to a new exuberance in advertising research in both industry and academia. While the social scientists and public Freudians such as Ernest Dichter had major philosophical differences with one another, they both led to a psychological effects tradition in advertising research. To this day, discussion of advertising effects is influenced by the pervasive (and misguided) beliefs of the period.

There were, however, some exceptions. In communication, Katz (1957) proposed the 'two-step flow' model in which certain social actors and their communication networks were seen as more important than others in the spread of information and influence. This was the first time that any kind of social consciousness was included in the 'effects model' at all. This model held that other social actors mattered as much as the communication message itself, if not more. Since advertising was mass-mediated communication it was assumed to act in the same way. It was at about this time that another new field, marketing, discovered the communication literature. From the late 1950s through the mid-1970s, several marketing researchers went looking for the elusive generalized opinion leader, a person to whom advertising could be effectively targeted. This person would then pass commercial information through interpersonal channels to great effect. But alas, the search was fruitless: the generalized opinion leader could not be found, and most marketing research again eschewed the social for the more psychological 'message in a bottle' approach, at least for a while.

Another critical effects notion came along in the 1960s: the 'hierarchy of effects.' The basic idea was that effects of communication (in particular advertising) occurred along a hierarchy: from exposure to behavior (Colley, 1961). Each step along the way depended upon the step before. This helped to explain why advertising was generally inefficient, why it took so much to move consumers all the way to behavior from awareness. While a vast improvement compared with an undifferentiated bullet theory, it was still bound up in the idea of a linear and fairly inflexible progression of effects. But it inspired a great deal of advertising effects research. Generally speaking, research confirmed that not all ads lead to behavior. This was progress.

A significant trend in advertising effects research began in the mid-1970s. Known as the 'information processing' tradition, its orientation has been to study ads as the bearers of 'information,' which is then 'processed' by audience members. It was assumed that people actually did something with information gained from ads. This too was progress. These researchers held (at least implicitly) that underlying psychological processes would have to be understood in order to ultimately understand advertising's effects, and make them generalizable and predictable. Most of the earliest work in this tradition was heavily influenced by then-contemporary attitude theory (Fishbein and Ajzen, 1975), but later became increasingly more cognitive in orientation. Of all the things advertising does, its ability to get consumers to remember the brand, its name, and something good about it, has the longest research tradition. Interestingly, recall of advertising seems important in some instances, but completely useless in others. Keller (1993) demonstrates that advertisements are much better retrieved, and retrieved in a manner most desirable to the advertiser, when there is a carefully planned congruence between the advertisement itself and the point of purchase. The point of purchase should be rich with memory retrieval cues that are entirely consistent with elements found in the ad itself. This work is very important in that it explicitly links point of encoding and point of retrieval. This helps to explain why recall

of the actual ad seems important sometimes, and at other times, relatively worthless.

One model has dominated research in the psychological realm. It is the elaboration likelihood model (ELM) of Petty and Cacioppo (1986). This model says that there are two routes to persuasion: one peripheral and one direct. The direct route is marked by attention, consideration, engagement; in other words, 'elaboration.' The consumer engages arguments, or attempts at persuasion in a variety of ways, but generally takes them head-on, agrees with some, counter-argues others, etc. The peripheral route is just the opposite; there is little cognitive effort or engagement, and little motivation to process. In addition to these two routes, various cues were also assessed as either peripheral or direct. For example, source credibility, presenter attractiveness, or executional elements such as music, color, etc., are thought to be peripheral. Research emanating from the ELM has been encouraging, and generally supportive of the theory. Peripheral cues are at their best in situations where the consumer is less involved, has less at risk, and attention is low; central processing is most effective when the opposite is true. It is generally thought that long-term advertising effects are greatest when the direct route is engaged. In the long run, getting people to pay attention to ads seems to predict their greatest effect. While it is not exactly clear where emotions fit in the ELM, work by Edell and Burke (1993) demonstrates that emotions (particularly warm and upbeat ones) do influence the overall evaluation of the brand, and are recalled just as well as other types of information.

Another contribution from the effects literature is in the distinction between search attributes and experience attributes (Nelson, 1974; Wright and Lynch, 1995). Search attributes are objective pieces of consumer information: how many miles to the gallon, does the car come in red, how much horsepower does it have, etc. Experience attributes are those things that one can only really know via direct experience: does the engine sound good, did the seats feel comfortable, does the stereo sound good, etc? Wright and Lynch (1995) show that search attributes are actually better learned through exposure to advertising, while experience attributes are better learned through direct contact and use of the product or service. This effect is stronger under low involvement conditions than under high involvement. This finding is important because it has long been assumed that advertising was a fairly weak force compared to direct experience, in almost all cases. This is an important qualification of advertising effects. Similar findings by Edell and Burke (1986) show that consumer's pre-existing attitudes toward the brand have a great deal to do with the attitude-to-the-ad/attitude-to-the-brand connection, particularly as they mediate through motivation to process the ad.

Another interesting research area is the 'third person effect' (Davison, 1981). This work demonstrates that individuals attribute much greater power to advertising when asked about some other (third) person than himself or herself. In other words, they believe advertising to affect others much more than themselves. The effect is particularly strong when those doing the judging are higher in socioeconomic status (Atwood, 1994). So, relatively more affluent and educated people think that others, particularly poorer and less educated people, are more affected by advertising than they are. This effect has considerable implications for effects research in general, and public policy in particular. This work is typically carried out through surveys.

Special Audiences

Two audiences receive special attention in terms of advertising effects: young children (those under 12 years of age) and the elderly. The clearest findings are those reported by Roedder-John and Cole (1986). They find that significant processing difficulties exist in both children and the elderly. These are identified as 'memory-strategy' usage deficits, or strategies for the effective encoding and retrieval of stored information. These strategies are important in knowing how to effectively gain information from advertisements, and to defend oneself against some forms of deception (intentional or not). In addition, young children suffer from knowledge base deficits. They simply do not yet possess the requisite knowledge base to know how to interpret some advertisers' messages. The implication is that the effects of advertising on these populations may be significantly different than for others.

Going Forward

The arguments surrounding the relative worth of this literature are fairly typical philosophy of science ones. Experimentalists argue that the only way one is ever going to be able to understand the effects of advertising is to understand them within the tightly controlled confines of the laboratory, where all other noise is ostensibly screened out, and process is detected and discerned. More naturalistic researchers counter that advertising cannot be divorced from its social context and construction, and thus the pursuit of such effects is a comfortable conceit. Not too surprisingly, both sides make valid points. There are certain processes which would simply be hard, by their definition, to study outside a laboratory, at least where the detection of psychological process is concerned. On the other hand, things closer to actual behavior are more entangled, by their nature, in discourse, and social circumstance. Ethnographic work holds some promise. Actually watching people watch ads (at least on television) makes some degree of sense (Ritson and Elliott, 1999). Yet this work is limited to what people do with advertising content in their daily lives, or the social uses of advertising. We do not typically see ethnographic research at the point of exposure, or have any idea what is going on inside the audience member's head. Furthermore, there are all the familiar criticisms of ethnography: lack of control and ungeneralizable findings. Historical and textual methods also offer promise but it is certain that no single method is going to reveal enough about advertising effects on its own. The phenomenon is simply too large, too layered, and too multifaceted.

One of the problems beyond method is that advertising is simply not very important to most people, most of the time. As Klapper argues in his 1960 limited effects model: mass communication is only effective under two conditions: (1) where the communication is completely consonant and resonates with some social theme, or (2) in a total *tabula rasa* situation. In other words, it does two things very well: delivering

information and increasing knowledge. Beyond that it is an extremely weak force at the micro-individual level. The original VW Bug ads of the late 1950 and 1960s are good examples of the former. The VW advertising rode a developing wave of pop consumer counter-culture. The ads provided information consonant with this social movement. In other instances, literally nothing is known about specific goods or services, and advertising provides that information. A good example is the introduction of the compact disc player. In that instance advertising supplied consumers with essential information about something that was completely unknown to them. Advertising wrote useful information on a blank slate. But this is rare. Most of the time it is the noise in the background while we are preparing dinner, or the things between the segments of our television shows. Its effects are, by advertising's nature, difficult to pin down.

See also: Advertising Agencies; Advertising and Advertisements; Advertising: General; Hegemony and Cultural Resistance; Mass Communication: Normative Frameworks; Media Effects on Children; Media Effects; Media and Child Development.

Bibliography

Albion, M.S., Farris, P.W., 1981. The Advertising Controversy: Evidence on the Economic Effects of Advertising. Auburn House, Boston.

Atwood, E.L., 1994. Illusions of media power: the third-person effect. Journalism Quarterly 71 (2), 269–281.

Borden, N.H., 1947. The Economic Effects of Advertising. Irwin, Chicago.

Colley, R.H., 1961. Defining Advertising Goals for Measured Advertising Results. Association of National Advertisers, New York.

Davison, B., 1981. The third-person effect in communication. Public Opinion Quarterly 47, 1–15.

Edell, J.A., Burke, M.C., 1986. The relative impact of prior brand attitude and attitude toward the ad on brand attitude after ad exposure. In: Olson, J., Sentis, K. (Eds.), Advertising and Consumer Psychology, vol. 3. Praeger Publishers, Westport, CT, pp. 93–107.

Edell, J.A., Burke, M.C., 1993. The impact and memorability of ad-induced feelings: Implications for brand equity. In: Aaker, D.A., Biel, A.L. (Eds.), Brand Equity and Advertising: Advertising's Role in Building Strong Brands. L Erlbaum Associates, Hillsdale, NJ, pp. 195–212.

Ewen, S., 1988. All Consuming Images: The Politics of Style in Contemporary Culture. Basic Books, New York.

Faludi, S., 1991. Beauty and the backlash. In: Backlash: The Undeclared War against Women. Anchor, New York.

Fish, S., 1980. Is There a Text in This Class? Harvard University Press, Cambridge, MA.

Fishbein, Martin, Ajzen, I., 1975. Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research. Addison-Wesley, Reading, MA.

Fox, R.W., Jackson Lears, T.J., 1983. The Culture of Consumption: Critical Essays in American History 1880–1980. Pantheon, New York.

Friedman, M., 1991. A 'Brand' New Language: Commercial Influences in Literature and Culture. Greenwood, Westport, CT.

Galbraith, J.K., 1958. The Affluent Society. Riverside, Cambridge, MA.

Grady, J., 2011. Advertising images as social indicators: depictions of blacks in *LIFE* magazine, 1936–2000. Visual Studies 22 (3), 211–239.

Katz, E., 1957. The two-step flow of communication: an up-to-date report on an hypothesis. Public Opinion Quarterly 21, 61–78.

Keller, K.L., 1993. Memory retrieval factors and advertising effectiveness. In: Mitchell, A. (Ed.), Advertising Exposure, Memory and Choice. L Erlbaum Associates, Hillsdale, NJ, pp. 11–48.

Kim, MinChung, McAlister, Leigh M., July 2011. Stock market reaction to unexpected growth in marketing expenditure: negative for sales force, contingent on spending level for advertising. Journal of Marketing 75 (4), 68–85.

Klapper, J.T., 1960. The Effects of Mass Communications. Free Press, New York.

Lee, W., Joo, J.J., 2012. Understanding the role of culture in advertising. In: Sintaro, O. (Ed.), Handbook of Research in International Advertising. Edward Elgar Publishing, Northampton, MA, pp. 20–41.

Lichtenstein, M., Srull, T.K., 1985. Conceptual and methodological issues in examining the relationship between consumer memory and judgement. In: Alwitt, L.F., Mitchell, A.A. (Eds.), Psychological Processes and Advertising Effects: Theory, Research and Applications. L Erlbaum Associates, Hillsdale, NJ, pp. 113–128.

Lindenman, J., 2010. The Economy of Brands. Palgrave Macmillan, Basingstoke.

Luik, J.C., Waterson, M.J., 1996. Advertising & Markets: A Collection of Seminal Papers. NTC Publications, Oxford.

McKay, N., Mancanda, R.V., Smith, M.c., Huhmann, B.A., 2011. Effects of age, need for cognition and affective intensity on advertising effectiveness. Journal of Business Research 64, 12–17.

Muniz Jr., A., O'Guinn, T.C., March 2001. Brand community. Journal of Consumer Research 35, 737–747.

Nelson, P., July/August 1974. Advertising as information. Journal of Political Economy 83, 729–754.

O'Guinn, T.C., Social anxiety advertising and consumers' perceptions of base rates of product addressed social problems, unpublished manuscript.

O'Guinn, T.C., Faber, R.J., 1991. Mass communication theory and research. In: Kassarjian, H.H., Robertson, T.S. (Eds.), Handbook of Consumer Behavior Theory and Research. Prentice Hall, Englewood Cliffs, NJ, pp. 349–400.

O'Guinn, T.C., Shrum, L.J., March 1997. The role of television in the construction of consumer reality. Journal of Consumer Research, 278–294.

Packard, V., 1960. The Waste Makers. David McKay, New York.

Petty, R.E., Cacioppo, J.T., 1986. Communication and Persuasion: Central and Peripheral Routes to Attitude Change. Springer Verlag, New York.

Pollay, R.W., April 1986. The distorted mirror: reflections on the unintended consequences of advertising. Journal of Marketing 50, 18–36.

Richins, M.L., 1991. Social comparison and the idealized images of advertising. Journal of Consumer Research 18, 71–83.

Ritson, M., Elliott, R., December 1999. The social uses of advertising: an ethnographic study of adolescent advertising audiences. Journal of Consumer Research 26, 260–277.

Roedder-John, D., Cole, C., 1986. Age differences in information processing: understanding deficits in young and elderly consumers. Journal of Consumer Research 13 (3), 297–315.

Schmalensee, R., 1972. The Economic Effects of Advertising. North-Holland, London.

Schudson, M., 1984. Advertising, the Uneasy Persuasion: Its Dubious Impact on American Society. Basic Books, New York.

Scott, L.M., December 1994. The bridge from text to mind: adapting reader response theory for consumer research. Journal of Consumer Research 21, 461–486.

Scott, L.M., 2002. Fresh Lipstick: Redressing Fashion and Feminism. University of Illinois Press, Urbana, IL.

Seiter, E., 1995. Different children, different dreams: racial representation in advertising. In: Dines, G., Humez, J. (Eds.), Gender, Race and Class in Media: A Text Reader. Sage, Thousand Oaks, CA, pp. 99–108.

Spielmann, N., Richard, M., 2013. How captive is your audience? Defining overall advertising involvement. Journal of Business Research 66, 499–505.

Taylor, Kimberly A., Miyazaki, Anthony D., Mogensen, Katherine Beale, 2013. Sex, beauty, and youth: an analysis of advertising appeals targeting U.S. women of different age groups. Journal of Current Issues & Research in Advertising 34 (2), 212–228.

Wright, A.A., Lynch Jr., J.G., March 1995. Communication effects of advertising versus direct experience when both search and experience attributes are present. Journal of Consumer Research 21, 708–718.

Advertising: General

John Sinclair, The University of Melbourne, Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Advertising is an industry that involves a three-way relationship between advertisers, their agencies, and the media. The history of this industry dates from the late nineteenth century in the industrialized countries, while the twentieth century has seen its growth on a global scale. Advertising's high visibility and pervasiveness has attracted much social criticism and successive analytic approaches over this time, with recent attention focusing on advertising's role in the formation of brands and social identities.

This article begins with the concept of advertising, defining it as a cultural industry carried out in agencies which mediate between the whole range of marketing practices pursued by advertisers, and the media which carry advertisements. The functional significance of this relation for the commercial media is then explained. A very brief history of advertising is outlined, with emphasis on the formative stages in the past century, with attention to the forces impelling the internationalization of the advertising industry. This leads to an outline of the rise of social criticism of advertising and its perceived effects, and governmental response in the form of regulation. A brief resume is given of past and present academic approaches to the study of advertising, which includes the analysis of advertisements themselves. The article concludes with some consideration of advertising within an anthropology of consumption and a postmodern commercial esthetic, with particular attention to advertising's role in branding in the age of social media.

Advertising is central to the study of media and the commercial applications of the social sciences. Not only does advertising revenue provide the major source of income for print and broadcast media owners, but also gives those media their characteristic look and sound, and orients their content toward the kind of audiences which advertisers want to reach. Advertising can be defined as a kind of cultural industry which connects the producers of consumer goods and services with potential markets through the diffusion of paid messages in the media. What is loosely referred to as advertising in everyday language is really just the most visible form of marketing, in which particular images become associated with branded goods and services. Within the industry, advertising is seen as just one part of what is now called 'integrated marketing communication.' Yet advertising is distinguished for being the imaginary dimension of what may be called the manufacturing–marketing–media complex of modern societies, the whole institutional structure of production and consumption.

Advertising as a Cultural Industry

Not all media depend upon the sale of their space and time for advertising – films, for example, don't carry advertising as such – and not all advertising depends upon the media for its diffusion. Some companies spend more on 'below the line' or nonadvertising forms of marketing communication, such as direct mail. However, in most countries, both the developed and the developing world, the largest advertisers of consumer goods and services, many of them global corporations, have determined the direction and character of print and broadcast media development because of their demand for the means ('media') which allow them to reach potential consumers. This is true regardless of whether the media are owned by private interests or the state, as was seen in the case of the rapid commercialization of television in Europe in the 1980 and 1990s.

Advertising is cultural in that it deals in images which give public expression to selected social ideals and aspirations, but it is also an industry because of its crucial intermediary role in the manufacturing–marketing–media complex. It was in this sense that Raymond Williams called it 'the official art of modern capitalist society' (1980: p. 184).

The intermediary role is carried out by advertising agencies (*see* Advertising Agencies). These 'agents' produce and place advertising for their 'clients,' the consumer goods or service companies who are the actual advertisers. However, while the clients pay the agencies for their services, agencies also derive income directly or indirectly from the media in the form of a sales commission paid in consideration of the media space and time, traditionally newspaper pages or television 'spots,' which the agencies purchase on behalf of their clients.

This is the crux of the manufacturing–marketing–media complex referred to above. Indeed, this form of dealing in media space and time was the origin of advertising agencies as a form of business in the later nineteenth century, and continues as a fundamental aspect of the industry today. The judicious purchase of media space and time is part of the range of functions routinely performed for clients by 'full-service' agencies, while there have emerged specialized agencies which deal in such media purchasing alone, known as 'media agencies.'

Looking at the complex of relations between advertisers, agencies, and media from the point of view of the media as an institution, it becomes evident that in most countries, the dominant print and broadcast media are 'commercial,' as distinct from government-funded or nonprofit media. This means that the measurement of readerships and audiences becomes an issue for advertisers, as they want to be sure that their advertisements are reaching the kind and range of the prospective consumers they seek (*see* Audience Measurement).

For their part, the media must seek to attract those prospective consumers as readers or viewers, and this becomes the basis for the kind of content they provide.

While this relation is easiest to observe in the case of broadcast television and its constant striving to offer programs which can attract the largest audiences, mostly for the benefit of national advertisers of packaged consumer goods, to many advertisers, the kind of audience can be more important than the size. The agency's expertise in the choice of medium for the client is a key element in their mediation of the process of fitting people to products. Thus, while the broad appeal of television content might reflect a relatively undifferentiated 'mass' nationally-networked audience-market, other media such as prestige newspapers and special interest magazines are supported more by advertisers who are interested in reaching only affluent consumers or niche markets, respectively. Notably, the advent of the internet has given advertisers an unprecedented medium through which to reach ever more closely targeted markets, especially through social networking sites such as Facebook.

All of these connections to the media industries are of most interest to political economists of the media, while social science and cultural studies more generally have been more concerned with the cultural output of advertising agencies, in the form of the marketing campaigns which they devise for their clients, and the actual advertisements which they produce. Here once again, full-service agencies provide various other marketing activities, including market research, a major form in which social science methods are applied for commercial purposes. However, it is the 'creative' side of the agencies' work, including that of specialist creative 'hot shops,' which gives advertising its cultural significance.

The Development of Advertising

Advertising has its origins not just in the sale of space and time, but in its role of providing distinct identities to branded goods and services. It arises in the USA and the UK around the late nineteenth century when packaged household products began to replace generic goods bought in bulk: Pears rather than soap, to name one of the world's very oldest brands. By giving brands a particular character and often a logo or slogan to make them recognizable, advertising contributed to the national, and later the international, expansion of brands such as Lipton, Gillette, Kodak, and Ford. By the 1920s, advertising was forging an alliance with what were then the still fairly new social and behavioral sciences, drawing on sociological techniques for market research, and devising psychological appeals for advertisements themselves.

Also around this time, some of the longest-established US agencies began to expand overseas. For example, J. Walter Thompson upgraded its sales office in London, while McCann-Erickson opened branches in Latin America. In both cases, this expansion was initiated at the request of large clients whose brands they handled in the USA, and that wanted similar advertising services provided for them on a 'common account' basis in the foreign markets they were opening up. US brands such as Coca-Cola thus became known worldwide.

However, the most important period for the internationalization of the advertising industry was the period after World War II. Apart from this being the era in which many US companies transformed themselves into multinational corporations, it was also the time when television was a new medium being adopted by one country after another around the world. Because the US agencies had their common account agreements with their clients from their home market as well as some experience with television, they were able to quickly dominate the advertising industries of those countries they entered.

Indeed, by the 1970s, this movement had provoked some resistance, and US advertising agencies became one of the targets of the rhetoric against 'cultural imperialism' in that decade. Subsequent expansion then had to proceed more through joint ventures in conjunction with national agencies. However, a more elaborate pattern emerged in the 1980s, particularly with the growth of certain British agencies, notably Saatchi & Saatchi. No longer was the worldwide advertising industry defined by a fairly simple conflict between US-based multinationals versus the various national interests, but there were more complex tendencies emerging.

Most significantly, what the British agencies did was to keep their new US acquisitions intact and incorporate them into 'megagroup' structures, a strategy already undertaken by McCann-Erickson following its merger with SSC&B Lintas. This was a way of dealing with 'client conflicts,' for example, British Airways could be handled by Saatchi & Saatchi, while other airline accounts could be assigned to different agencies in the same group without fear of marketing secrets being leaked from one to the other (Sinclair, 2012). Apart from this kind of integrated concentration in the English-speaking world, substantial French and Japanese groups have formed, so that there is a trend to the interpenetration of national markets and of world-regional markets, including joint ventures in culturally difficult markets such as China and India. These trends became more marked with the globalization of the 1990s and 2000s (*see* International Advertising). Also worth noting has been the tendency for the global megagroups to incorporate under their umbrella companies in related marketing communication fields such as public relations, as well as advertising agencies in more specialized fields, notably internet advertising.

The advent of the internet has become a considerable challenge to the traditional, quite settled relationships within the manufacturing–marketing–media complex. In particular, the unprecedented rise of search advertising and the dominance of this field by Google has meant that advertisers can now bypass the agencies and place their advertising directly with Google or the other search engines. Internet advertising has shown very rapid growth, particularly at the expense of print media. The largest advertisers globally, however, continue to invest the bulk of their expenditure in television.

Globalization has had controversial effects upon the content of advertising. While some agencies advocate the standardization of advertising campaigns throughout the world ('one sight, one sound, one sell'), others argue for what the Sony Corporation calls the 'glocalization' of both products and their marketing, that is, adapting them in accordance with the cultural differences evident between the various national

markets. In practice, services such as credit cards or airlines appear to benefit from global advertising, but goods such as packaged foods do not (Sinclair, 2012).

The Critique of Advertising

As the advertising industry has been consolidating itself at its most intensive level of globalization ever, it is ironic that the social critique which accompanied it during its previous decades of expansion has become relatively diminished. However, for most of the later part of the twentieth century, advertising was subject to considerable critique from various quarters of society, and on economic as well as cultural grounds, so it is worth reviewing the terms of the debate and the sources of criticism.

It has been the critics of advertising who have made the most elaborate claims about its economic importance. Marxist and liberal critics alike have argued that advertising creates and controls demand, thus attributing to it a key function in the perpetuation of consumer capitalism. Certainly, modern marketing gives producers a wide range of sophisticated strategies and techniques, but advertising is only one element among them. Furthermore, the fact that the vast majority of new products still fail when they are introduced to the market suggests that there is no one in control of demand.

Another economic criticism of advertising is that it adds to the price of goods and services, because advertising costs are real costs which are passed on to the consumer. The rebuttal is that, because 'advertising is the shortest distance between producer and consumer,' to quote J. Walter Thompson, it rationalizes the distribution process, generating only small costs which are spread easily over the large volumes of production which advertising helps to make possible.

The defenders of advertising argue that it is necessary to maintain competition, and that if producers do not advertise to maintain their market share, they will be driven out by their competitors, creating oligopolies, or markets dominated by very few producers. However, it follows that advertising is also a barrier against the entry of new players, who must necessarily be able to establish high levels of advertising from the beginning. This favors the 'market power' of the existing producers, so advertising would appear to favor oligopolistic markets.

In practice, it is evident that some of the industries which advertise the most heavily, such as packaged foods, or household and personal cleaning products, do indeed have tendencies to oligopolistic concentration. Furthermore, it is worth noting that such oligopolies tend to maintain themselves not by the mass advertising of just one product line, but by product differentiation: for example, different shampoos for different types of hair. Advertising is implicated here because of its crucial role in branding, 'positioning,' and otherwise creating apparent differences between products for different types of consumers.

The connection between the ideological and the cultural dimensions of advertising is achieved most fully in the Marxist critique, where advertising is seen to perform ideological 'functions' which reproduce the capitalist system as a whole. In the USA, Stuart Ewen argued that department store owners and other ideologues of capitalism in the 1920s embarked on a "project of ideological consumerization" (1976: p. 207) to draw working people into compliance with capitalism. More recent theory would see this as part of the 'pact' between capital and labor which stabilized the growth of capitalism in the first part of the century, the age of 'Fordism' (Arvidsson, 2006). In the UK, Marxist critics such as Raymond Williams (1980) were influenced by a British tradition of social critique which denounced advertising on moral grounds, not so much because of its materialist values, but more because of its persuasive appeals and irrational associations. Other Western Marxists took the view that advertising was an agent of 'false consciousness,' lulling the masses into acceptance of their place in an unjust social order.

The Marxists have not been isolated in the critique of advertising, however. Liberal critics of capitalism in the USA in the 1950s, a formative period in the history of the manufacturing–marketing–media complex, were read throughout the English-speaking world; thanks to the innovation and spread of paperback publishing at that time. Particularly in the context of the Cold War and popular fears about ideological 'brainwashing,' exposés of 'motivational analysis,' and the 'manipulation' of symbolic meanings in advertising, notably Vance Packard's *The Hidden Persuaders*, first published in 1957, found a ready audience. The following year, J.K. Galbraith's *The Affluent Society* came out criticizing the US economy for the way it produced the very needs which its goods were intended to satisfy.

The twentieth century's last phase in the social criticism of advertising came from the women's movement, beginning in the 1970s. Advertising was a major public cultural arena which women cited as evidence of the social processes through which sex-roles were represented and reproduced through 'stereotyping.' In particular, advertising was seen to perpetuate women in subordinate domestic roles, at the kitchen sink in detergent advertisements, for example, or alternatively, as sexual objects presenting themselves for the pleasure of the male gaze. Women's criticism of the 'sexism' of their representations in advertising was sustained into the 1980s, often with accompanying activism, even as advertising began more to represent the 'new' independent woman. Since the 1990s, feminist critique has become largely absorbed into a broader academic approach to consumer culture in general, but gender remains an issue, seen in concerns over the impact of advertising on body image, for example.

In addition to the specific economic effects attributed to advertising, as outlined earlier, each of these classic phases of social criticism focuses on one of advertising's alleged cultural effects – that advertising has an ideological role in papering over social inequalities, that it creates false and irrational needs, and that it subordinates and degrades women. As well, there are other social concerns, with a similar perception of its power, which legitimize a public interest in the regulation of advertising – most famously on health issues, notably the question of tobacco advertising, and more recently, advertising's role in childhood obesity (*see* Advertising, Control of).

Beyond academic critique, advertising attracts some form of regulation in all countries, but with great variation, particularly in the ratio of self-regulation by the industry itself, as compared with government-prescribed regulation. Apart from the considerable cultural differences about what is acceptable

in advertising images, advertising regulation typically restricts the type of product which may be advertised, such as alcohol and tobacco; the media which may be used, and at what times; the extent of product claims and invidious 'comparative' advertising; and in some cases, the use of foreign-produced advertisements (Frith and Mueller, 2003).

No doubt the capacity of advertising to exert social influence is greatly overestimated, and its structure and processes much misunderstood, but even in an age of 'deregulation,' most governments are reluctant to surrender too much of their control over advertising. For their part, the advertising interests seek to defend themselves from government intervention by developing voluntary, self-regulatory codes in the most controversial areas. An interesting and current case in point is the environmental movement's critique of advertising, which, they say, promotes consumer demand for products which generate waste and degrade the environment. Advertisers in response have clamored to assert their 'green' credentials, sufficiently to keep governments at bay, although not to convince their critics (Smart, 2010).

Advertising in Social Theory

The social criticisms of advertising referred to in the previous section, arising from liberalism, Marxism, and feminism, are grounded not just in those various social and intellectual movements as such, but also in the considerable academic theorization and research which they have generated in the social sciences.

While it would be less true of the USA than the UK, Canada, and Australia, Marxism provided the dominant paradigm in cultural studies and much social science, particularly sociology, from the mid-1970s until at least the end of the 1980s. This was a 'Western' Marxism in which there were two main trends: one toward 'political economy,' which emphasized the ownership, control, and functioning of the economic structure of capitalism, and the other toward the cultural analysis of role of ideology in maintaining the system as a whole. In both trends, but particularly the latter and more influential way of thinking, advertising was seen to have a crucial role to play in stabilizing a society which would otherwise be torn apart by its own contradictions. Following the French Marxist structuralist philosopher Louis Althusser, Marxist social theory in the 1980s thus shifted its analytical focus away from the economic structure as the basis for capitalist society, and toward ideological reproduction – the representational and signifying practices of capitalist culture, including advertising.

This tendency in cultural Marxism found common cause with semiological structuralism (*see* Semiotics), derived from Ferdinand de Saussure, but mobilized most famously with regard to advertising by Roland Barthes (1977). In this approach, advertisements themselves became the main object of analysis, such that the emphasis was upon how the various meaningful elements, or the 'signifiers' in an advertisement, related to each other so as to produce the meaning of the advertisement as a whole (*see* Advertising and Advertisements). This is a qualitative, interpretive approach which contrasts with the quantitative method of 'content analysis.' The latter, with its roots in US behaviorism, nevertheless can be applied in complementary and productive ways in conjunction with semiological analysis.

Along with Marxist and semiological structuralism there was a significant contribution from anthropological structuralism, stemming from Claude Lévi-Strauss. These strands were all brought together, along with feminism and the psychoanalytic development theory of Lacan, in Judith Williamson's *Decoding Advertisements*, still one of the most definitive books ever on the analysis of advertisements (Williamson, 1978). She provided a coherent fusion of these theories, and applied them in the qualitative analysis of scores of magazine display advertisements, to evince such processes as 'interpellation' and the invocation of ideological 'referent systems' in the interpretation of advertisements. Apart from Williamson's application of Lévi-Strauss's theory of 'totemism' to explain how certain kinds of people become associated with particular products in advertising, such as the 'Pepsi generation,' anthropological structuralism provides a way of understanding how goods become endowed with cultural significance through their position in a total system of meaning. Advertising visibly contributes to this process of giving meaning to goods, but by no means exclusively. The process is socially mediated. For example, of several different international brands of sportswear advertised in similar ways, it will be the peer group that decides which of them is to be the preferred 'cool' brand in any given locality.

The same relational quality of the cultural meaning of goods is also found in the post-structuralist contribution of Jean Baudrillard (1981). In his view, capitalist social structure was the source of needs as well as of the meaning of goods, and like certain of the liberal and Marxist critics of advertising cited above, Baudrillard saw the rise of consumer capitalism as a device by which the system avoided the need to redistribute its wealth. Thus, class differences were concealed beneath an apparent democracy of consumption, a connection which is lost in the bewildering and endless display of signification, he said.

With the advent of post-structuralism and postmodernism, the diversification of feminism, and the eclipse of Marxism, there was much less critical attention paid to advertising as such during the 1990s. Rather, although traditional press and magazine display advertisements, television commercials, and billboards continued to provide the examples of postmodern visual culture which were cited, theory and research from the 1990s onward see advertising in a broader and much more theorized context. This is even wider than the manufacturing–marketing–media complex described above. Thus, for Wernick (1991), advertising was just a part of 'the promotional condition of contemporary culture,' which went beyond the marketing of commercial goods and services to include the mode of public communication now embraced by all major social institutions, from political parties to universities, and also found in the presentation of one's own self.

In ways like this, theory and research were moving beyond the study of advertising as such, and taking a 'cultural turn' toward the analysis of consumer culture in general. Not only has this shift encouraged attention to the role of hitherto-neglected institutions such as the department store and the supermarket, but also to transformations in work, domestic life, and cultural identities, insofar as these have become expressed and commodified in terms of consumer goods. This

agenda in turn has given rise to studies of how specific groups have been constructed, represented, and appealed to by marketing strategies, and of how they have responded. Most recently, because advertising as an institution is so much an agent of commercial culture, it has become an area of particular interest to the 'cultural economy' school of thought in Britain (McFall, 2004).

Such lines of investigation have been a refreshing corrective to the preponderance of attention previously given to the content of advertisements themselves, made without regard to their audiences, which hitherto characterized most theorization and research about advertising. Furthermore, the new approaches have provided some insight into the 'reflexive modernization' with which consuming audiences have been seen to regard media and consumption in the era of globalization (Lash and Urry, 1994). This reflexiveness entails a postmodern esthetic in which consumers express themselves as individual subjects by means of how they mobilize their knowledge of the codes of meaning which goods carry, codes which are partly bestowed by the images in advertising, marketing, and the media, but which become embedded within popular culture. Clearly, this cultural relation of mediated images and their expressive use cannot be understood by analysis of the images alone. Writers such as Klein (2001) have pointed to the way in which advertising observes and exploits movements in popular culture, a process in which advertising creative staff may be personally involved as 'cultural intermediaries' (McFall, 2004). Ethnographic studies, such as that of Nixon (2003) within agencies in Britain, confirm how advertising personnel themselves may be participants in the subcultures which they draw upon in creating their advertisements. Other research shows how advertising agency personnel are also actively engaged in perpetuating the mythic status of advertising itself (Cronin, 2004).

As 'advertising' as an object of study has become more elusive in an age of media fragmentation and audience segmentation, the study of branding in particular has served to refocus upon the relationship between persons and goods in the age of new media. More than uniquely identifying a product or its manufacturer, branding enables products to acquire cultural meanings, such as status positioning; associations with certain lifestyles; and even something like their own personalities (Moor, 2007). Arvidsson has argued that consumers, rather than being the passive dupes of advertisers, actually participate in the making of a brand, albeit unequally: the value of a brand ultimately derives from the 'meaning-making activity' of consumers (2006). Advertisers pick up on these meanings and exploit them by associating them with particular products and services. The advent of the internet has given advertisers a means through which they can tap meaning-making as never before, for instance by drawing upon user-generated content, such as the one that has occurred in the United States for Super Bowl advertising, or by strategically placing videos on the internet in the hope they will 'go viral'; that is, users will want to send them on to each other. On the other hand, the advertisers of brands which are awkward or inauthentic in their appeals to users run the risk of 'culture jamming,' that is, being ridiculed by the same means, such as having lampooned versions of their advertisements being circulated over social media. However, the internet also enables online behavioral advertising, which allows advertisers to target individuals who have readily surrendered information about their tastes and interests on their favorite social networking web sites, thus identifying themselves as having an affinity with particular brands. The quest for individualization and authenticity in consumption is thus seen to be harnessed into forms of 'work' which consumers carry out for advertisers, because of the way in which consumers unwittingly signal lifestyle and subcultural trends to advertisers (Hearn, 2008).

Finally, as far as contemporary studies of the manufacturing–marketing–media complex are concerned, the challenges for theory and research are to keep pace with globalization, to comprehend its complexities, and to monitor the new relationships and business models which are forming between marketing and media in response to the growth of new technologies. Of particular interest is the role of advertising in cultivating huge new markets among the emergent middle classes of countries as diverse as Brazil, Russia, India, and China, and the question of whether those markets will be incorporated into the manufacturing–marketing–media complex as we have known it. On the technology front, the apparent gulf between 'new' and 'old,' or 'social' and 'mass' media, respectively, has been bridged through advertising campaigns increasingly being mounted on a 'cross-platform' basis. Mobile media which give access to the internet, notably 'smart' cellphones and tablets, represent a new frontier, and have opened up the unprecedented targeting of advertising in accordance with a user's precise physical location.

See also: Advertising Agencies; Advertising and Advertisements; Advertising, Control of; Audience Measurement; International Advertising; Semiotics.

Bibliography

Arvidsson, Adam, 2006. Brands: Meaning and Value in Media Culture. Routledge, London.
Barthes, Roland, 1977. Image-Music-Text. Fontana, London.
Baudrillard, Jean, 1981. For a Critique of the Political Economy of the Sign. Telos Press, St. Louis, MO.
Cronin, Anne, 2004. Advertising Myths: The Strange Half-Lives of Images and Commodities. Routledge, London and New York.
Ewen, Stuart, 1976. Captains of Consciousness: Advertising and the Social Roots of the Consumer Culture. McGraw-Hill, New York.
Frith, Katherine, Mueller, Barbara, 2003. Advertising and Societies: Global Issues. Peter Lang, New York.
Hearn, Alison, 2008. 'Meat, mask, burden': probing the contours of the branded 'self'. Journal of Consumer Culture 8 (2), 197–217.
Klein, Naomi, 2001. No Logo. Flamingo, London.
Lash, Scott, Urry, John, 1994. Economies of Signs and Space. Sage, London.
McFall, Liz, 2004. Advertising: A Cultural Economy. Sage, London.
Moor, Liz, 2007. The Rise of Brands. Berg, Oxford, UK and New York.
Nixon, Sean, 2003. Advertising Cultures: Gender, Commerce, Creativity. Sage, London.
Sinclair, John, 2012. Advertising, the Media and Globalisation. Routledge, London.
Smart, Barry, 2010. Consumer Society: Critical Issues and Environmental Consequences. Sage, London.
Wernick, Andrew, 1991. Promotional Culture: Advertising, Ideology and Symbolic Expression. Sage, London.
Williams, Raymond, 1980. Problems in Materialism and Culture: Selected Essays. Verso Books, London.
Williamson, Judith, 1978. Decoding Advertisements: Ideology and Meaning in Advertising. Marion Boyars, London.

Advocacy and Equity Planning

Norman Krumholz, Cleveland State University, Cleveland, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Until recently, most American city planners dealt solely with the physical city. But an alternative planning practice dealing with social equity has existed since the early 1900s, at the margins of the profession. These 'equity or advocacy' planners flourished during times of crisis, which bring forward the reform elements of the American state. They have had an impact on the profession. At the end of the twentieth century, American city planning was more sensitive and collaborative than it had been earlier.

Advocacy or equity planners are professional city planners who deal with the physical city while also trying to redistribute resources, political power, and participation toward disadvantaged urban populations. Their practice goes back to the early twentieth century and includes examples of tenement housing reform, low-income housing, common ownership of land, national resource planning, citizen participation, and linkage arrangements between downtown and poor neighborhoods. Through the years, leading advocacy/equity planners include Benjamin C. Marsh, Catherine Bauer, and Paul Davidoff. Although advocacy/equity planners often struggle against a market-oriented society, they have succeeded to an extent in liberalizing American urban planning. They have not, however, overcome persistent urban poverty, racial segregation, or growing economic disparities between city and suburb.

Until recently, most American city planners dealt solely with the physical city. They designed streets, parks, and boulevards; made plans for the way land was to be used in the community; and prepared regulations to control the use of land. Advocacy or equity planners are those professional planners who not only deal with the physical aspects of the community but who, in their day-to-day practice, also deliberately try to move resources, political power, and political participation toward the lower income, disadvantaged populations of their cities. They are called 'advocacy' or 'equity' planners because they seek greater equity among different groups as a result of their work. Where the work of most city planners is rarely consciously redistributive, advocacy or equity planners often conceive the potential contribution of planning in broad economic and social terms and try to provide for a downward redistribution of resources and political participation in order to create a more just and democratic society.

Many observers place the birth of advocacy and equity planning in the decade of the 1960s when crowds were in the streets of American cities protesting the wholesale demolitions and displacements caused by urban renewal and the interstate highway program. These traumatic events and the antiwar and civil rights movements, which occurred at about the same time, challenged the belief in top–down planning by benign, value-free experts and created a demand for more social planning based on grassroots involvement. The events of the 1960s provided great support for advocacy and equity planning, but actually, an alternative planning practice oriented toward equity considerations had its roots in the turn of the twentieth century.

To begin this exploration into recent history, one must go back to the period between approximately 1880 and 1915, known as the Progressive Era, a time when the respectable urban bourgeois discovered the slum city festering beneath their urban world. At the time, America was rapidly changing from an agricultural to an industrial society. American farmers, forced off their homesteads, joined European immigrants flooding into industrial cities. The largest cities, as centers of manufacturing, exchange, and distribution, had grown most explosively without proper planning or the means to regulate growth. They became choked with slums built to house immigrant workers (Riis, 1971). The slums, in cities like New York and Chicago, became such breeding grounds of disease, crime, and human misery that efforts at reform were introduced by civic and political leaders. It was this revulsion against the slum and the fear of revolutionary social unrest that brought housing reform and the social aspects of urbanism into modern urban planning. Progressive leaders believed that such dismal conditions could not wait on beneficent forces, but could be corrected by diligent and scientific health and housing policies.

The settlement house movement was one of their first efforts at neighborhood improvement. In the poor, immigrant neighborhoods of dozens of major cities, settlement houses like Hull House in Chicago, Henry Street Settlement in New York, and South End House in Boston were established. These were mostly staffed by middle- or upper-class social workers who taught the English language and domestic arts to immigrants, lobbied city government for more neighborhood parks and other public facilities, and pressed for effective tenement housing reform. In the process, the settlement house workers hoped to improve housing and neighborhood conditions, and 'Americanize' the new immigrants. They were not urban planners but, then, advocacy for parks, better housing, and other improvements in the slums helped provide the reform underpinning for the nascent city planning profession.

Three typical progressive era reformers were Mary Kingsbury Simkhovitch, Benjamin C. Marsh, and Alice Constance Austin. Simkhovitch pursued a wide range of social justice issues, including women's suffrage, economic reform, and progressive politics. After establishing Greenwich House in New York City, she worked to improve housing density and building-code

laws. During the 1930s, Simkhovitch performed her most important work when she helped draft the Wagner–Steagall Housing Act of 1937, providing federal participation in low-income housing for the first time in America.

Benjamin C. Marsh was the Secretary of the Committee on Congestion of Population, a prestigious committee formed in 1907 by 37 civic and philanthropic organizations to study, publicize, and promote programs to relieve the problems of excessive massing of people in New York City. Marsh widely traveled and was strongly impressed by European city planning, especially by the ideas of the Englishman Ebenezer Howard (Howard, 1965). Howard, who loathed the industrial city with its filth and overcrowding, proposed a scheme of land development based on population dispersion into a regional pattern of small, self-contained cities. These garden cities would enjoy all the advantages of the core city including nearby jobs in industry, higher wages, and social opportunities, while also enjoying the benefits of the countryside with low rents, fresh air, agricultural gardens, and cooperative arrangements to maintain the land. A central feature of Howard's scheme included the common ownership of land so that the unearned increments in land values could be recaptured for the benefit of the entire community. Howard's sweeping planning ideas were enthusiastically adopted by Marsh, for whom city planning was a holy war against predatory forces, especially real estate speculators. Marsh's energetic career as a skilled organizer and publicist of planning issues made the city planning movement better known nationally and more socially responsive. His small paperback book, *An Introduction to City Planning* (1909), set the stage for the First National Conference on City Planning and the Problems of Congestion convened in Washington, DC, in 1909, which marks the formal birth of American city planning.

Alice Constance Austin also used Ebenezer Howard's new town ideas as examples of good city planning when she assumed the role of city planner and architect in 1915 for the partially built socialist city of Llano del Rio in California (Hayden, 1976). Austin also learned from Patrick Geddes, a Scottish biologist, who drew up dozens of town plans in India and elsewhere, based on a cooperative model of city evolution. Austin's debt to Howard's garden city is reflected in her organization of the city of Llano del Rio, with its 'crystal palace'-like central buildings and its boulevards and street system. Her approach to building design was, however, distinctly feminist. Here, Austin proposed not only private gardens, but also communal kitchens and laundries to liberate women from drudgery.

Early American planners, most of whom were architects and engineers, believed that the way to bring the city under control was to reduce congestion, make physical development more attractive, and control the flow of traffic. They thought planning and zoning examples from Europe, especially from Germany and the United Kingdom, provided good direction. But, although early American planners used European models to define good planning, in one key respect the American approach diverged sharply from the European experience. In France, Germany, and the United Kingdom, the answer to the provision of low-income housing was government support; in the United States, such support was decisively rejected. Lawrence Veiller, author of the New York City tenement-house legislation of 1901 who later founded the National Housing Association, maintained that only local government should concern itself with housing, but only to enforce local building regulations (Scott, 1995). Otherwise, private benevolence could do the job. In Veiller's view, it was proper for local government to clear slums, but not to rehouse the displaced families.

A small group of visionaries rejected the physical determinism of most early planners. The Regional Planning Association of America (RPAA), formed in 1923, included such planning luminaries as Catherine Bauer, Stuart Chase, Benton MacKaye, Lewis Mumford, Clarence Stein, and Henry Wright. They believed in planning entire regions to achieve social objectives. Following the ideas of Howard and Geddes, RPAA members expounded their vision of small, self-sufficient communities scattered through regions in ecological balance with rich natural resources. In the distribution of electric power by regional grids and the speed of the automobile and truck, they saw new tools for rehabilitating declining urban neighborhoods and liberating large cities from congestion and waste.

The thinking and writing of RPAA's members was influential beyond their small numbers. Their ideas on regional planning and environmental conservation led in the 1930s to the creation of the Tennessee Valley Authority (TVA), the Civilian Conservation Corps, the 14-state Appalachian Trail, and the regional studies of the National Resources Planning Board. RPAA member Benton MacKaye, who conceived the Appalachian Trail, was clearly one of the founders of the modern environmental movement.

RPAA's main practical experiment, however, was only a mixed success (Birch, 1980). It was the garden city–inspired development of Radburn, in Fairlawn, New Jersey, constructed in 1928 using Ebenezer Howard's ideas, but adjusted to American customs and laws. The elements of Radburn include the 'superblock,' the cul-de-sac and narrow loop lanes for residential traffic, the clustering of housing around large areas of parkland in common ownership, and the separation of vehicular and pedestrian traffic. Radburn also provided day care for working mothers and similar social services, as well as a community organization to administer commonly held land. Although Radburn has had an important influence on American planning thought and design, it failed to realize its sponsor's hopes of becoming a completely self-sustained community. It has now been swallowed up as part of northern New Jersey's amorphous sprawl.

By the time America emerged from World War I, the Progressive Era had lost much of its momentum. City planning, slum clearance, and housing reform, which had been part of the upsurge of reform, now suffered from a weakening of the liberal impulse. RPAA member Frederick L. Ackerman attacked the profession for no longer being concerned about "the causes which give rise to the existing maladjustment" in cities. America had maladjusted communities, he said, because planners declined to interfere with "the right of the individual to use the community as a machine for procuring individual profits and benefits, without regard to what happens to the community" (Ackerman, 1919).

Urban planning and especially social or equity planning lagged during the 1920s with 'the business of America was business,' but the Great Depression and the New Deal administration of Franklin D. Roosevelt (FDR) brought both back. FDR believed that the power of government should be used to restore misused land, harness wasted water, and revitalize despondent human beings. In his first term, FDR enacted legislation authorizing the widest application of planning yet proposed, including the TVA and the Resettlement Administration. By Executive Order, FDR also established the National Planning Board, later the National Resources Planning Board.

The National Resources Planning Board was America's first effort at national planning. The Board advised the Department of the Interior on regional planning, land and water conservation, and issues of social insurance and poverty. As Secretary of the Interior, Harold L. Ickes, told city planners and members of the American Civic Association in 1933, "long after the necessity for stimulating industry … shall be a thing of the past, national planning will go on as a permanent Government institution." However, the Board was resented by many old-line federal agencies like the Army Corps of Engineers who chafed under the Board's attempts to coordinate their projects, and by conservative Congressmen who feared that the agency was promoting foreign, socialist schemes. As a result, and to wide and dolorous lamentation among planners, Congress abolished the National Resources Planning Board in August 1943.

The Resettlement Administration, established in 1935 under the leadership of Rexford G. Tugwell, suffered a similar fate. Tugwell, who once proposed that the (US) Constitution be changed to include planning as a fourth power of government, immediately embarked on an effort to improve living conditions for working people in metropolitan areas. He seized on the idea of the English suburban garden village, but adapted the concept to the American situation. The towns built by the Resettlement Administration were called Greenbelt Towns. Originally, Tugwell's staff identified eight metropolitan areas where new towns were to be built, but funding limitations narrowed the choices to three areas: Greendale, Wisconsin near Milwaukee; Greenhills, Ohio near Cincinnati; and Greenbelt, Maryland near Washington, DC. The three towns, all examples of high-quality planning, featured small homes in a well-planned garden setting and housed about 2100 families in all. The distinctive feature of all three towns was the surrounding greenbelt protecting the towns from outside encroachment. The residents, largely industrial workers, were provided an opportunity to live in a low-cost, but far superior environment than they were used to. As a demonstration, the Greenbelt Towns made clear that superior alternatives existed to haphazard speculative sprawl. Unfortunately, before the Resettlement Administration could expand the three towns and build two others that were on the drafting table, Congress abolished the agency in June 1938. It was left to post–World War II British planners to build a substantial new towns program, while the United States had to content itself with the three Greenbelt towns and a few commercial initiatives including Reston, Virginia, and Columbia, Maryland.

FDR's New Deal also produced the first major intervention in the field of low-rent housing. Pushed vigorously by RPAA member Catherine Bauer and others, the administration passed into law the Housing Act of 1937, which provided low-rent housing for the deserving working poor. For the first time, the federal government would provide support for the capital costs of public housing construction; tenant rents were to take care of subsequent operating and maintenance costs. Later, in 1949 and 1954, Congress returned to the public housing issue and emerged with a commitment to more low-income housing. But Congress also made a larger commitment to urban renewal, which ultimately demolished more low-income housing close to the city core than was built. The scale of demolition and forced relocation was large; a study by the National Association of Home Builders estimated that the total housing demolition by all public programs between 1950 and 1968 amounted to 2.38 million units. These units were disproportionately the homes of poor and near-poor black families.

Ultimately, a revolt broke out against these excesses. Within an astonishingly short time, this revolt caused an almost complete inversion of almost every basic value in American planning practice. Influential critics Jane Jacobs (1961) and Herbert Gans (1962) spoke for the preservation of older urban neighborhoods, as they pointed out that planners implementing the urban renewal program were thoughtlessly destroying valuable housing and irreplaceable social networks while providing opportunities for profitable real estate investments. The opposition to urban renewal coincided with the civil rights movement, and racial riots tore through the cities revealing just how little the planning process had done for the poor. Opposition to the war in Vietnam, and with it the Pentagon style of planning by top–down, bureaucratic experts, was at its peak. There was now a deep distrust of professional expertise and there was a demand for advocacy and equity planning based on grassroots involvement.

Paul Davidoff, a lawyer, planner, and educator, made the most substantial contribution to the concept of advocacy planning (Davidoff, 1965). He argued that advocacy could reinvigorate planning in three ways: by broadening public debate and participation, by sharpening the skills of planners who would have to defend their choices, and by shifting the focus of planning from the purely physical to social and economic priorities. The essence of advocacy planning is the encouragement of alternative plans by all groups holding special values about the future of their communities. Advocacy planning would supplement the one, official city plan with alternative plans by Liberals, Conservatives, Democrats, and Republicans and other groups, but the emphasis would be on professional planning services for the poor, the black, and the underprivileged. Using legal analogies, the merits of these alternative plans were to be debated and the best plan would emerge from the debate. Davidoff's ideas were taken up by planning practitioners and educators with a strong cumulative effect.

During the 1960s advocacy planning organizations serving minorities, poor, and working-class whites were formed in many American cities (Heskin, 1980). Exemplars were Architects Renewal Committee of Harlem in New York City (1964), Urban Planning Aids in Boston (1967), and the Community Design Center in San Francisco (1967). The effectiveness of these grassroots organizations was limited

given their inadequate resources and often defensive strategies, but they signaled a new diversity. A national organization called Planners for Equal Opportunity was established in the 1960s and in 1975, under the leadership of advocate planner Chester Hartman; it continued as Planner's Network. By 1999, Planner's Network represented over 800 planners and academics concerned with social and economic justice. By the 1970s, the American Planning Association (APA), whose journal carried virtually no discussion of racial issues prior to 1970, had modified its Code of Ethics and Professional Responsibility to reflect concern for vulnerable populations: "A planner shall seek to expand choice and opportunity for all persons, recognizing a special responsibility to plan for the needs of disadvantaged groups and persons …"

How deep the commitment was to rank and file planning practitioners still must be determined, however. In 2005, APA/American Institute of Certified Planners amended its Code of Ethics and Professional Responsibility by moving its clause on "special responsibility to plan for the needs of disadvantaged groups" from the rules of conduct for which planners can be held accountable and subject to disciplinary action to a merely aspirational principle that cannot be the subject of an ethical misconduct charge (APA/AICP, 2005). This change was a major setback. Although the clause was difficult to enforce legally, which was the reason for the change, it served as a guide and inspiration for planners who were trying to shape their practice in a more equity-oriented direction.

Davidoff's writings on choice theory were important intellectual contributions to planning theory and they make him the preeminent symbol of progressive planning. They make perhaps the most persuasive argument on how a planner might reconcile professionalism with political engagement.

But Davidoff's substantial contributions to planning theory and practice were matched by his efforts at institutional innovation. He realized that from 1950 onward, the overwhelming preponderance of growth in population, jobs, and economic investment was not in America's cities but in the suburbs. Because of racial discrimination and exclusionary land use controls, blacks were being shut away from the benefits of suburban growth. In response, Davidoff founded Suburban Action, a nonprofit institute for research, litigation, and advocate planning services within metropolitan regions (Davidoff et al., 1970). Suburban Action used its resources to legally challenge restrictive zoning and land use controls in the suburbs and enlarge suburban opportunities for the black and poor.

Davidoff's special concerns for the disadvantaged were also adopted in some official city planning agencies within local government. One of these was in Cleveland, Ohio, where planning director Norman Krumholz and a core staff of progressive planners developed a system of values and strategies in the 1970s that came to be called 'equity planning.' The central theme adopted by the Cleveland planners was contained in this statement: "In the context of limited resources, first and priority attention should be given to the task of promoting wider choices for those Cleveland residents who have few if any choices" (Krumholz et al., 1975; Krumholz, 1982). This goal made clear that Cleveland's poor and near-poor were to receive priority planning attention.

The Cleveland planners justified the choice of their goal with three arguments. First, they argued that a long-standing, historic moral commitment existed to seek more equity in the social, economic, and political relations among people. Second, building on the ideas of philosopher John Rawls (1971), they used reason as a means of justifying a more equitable society – the kind of society that free, equal, and rational people would establish to protect their own self-interests. Finally, they justified their goal by reality: making explicit the imbalances in income, education, health and other social and economic variables that existed in Cleveland between city and suburb, and white and black citizens (Cleveland City Planning Commission, 1975). Over the 10-year length of the equity planning experiment, and under three different mayors, the efforts of Cleveland's planners resulted in 'fair-share' low-income housing distribution plans for Cuyahoga County, progressive changes in Ohio's property law, improvements in public service delivery, enhancement of transit services for the transit-dependent population, the rescue of lakefront parklands, and many other improvements.

Another outstanding example of progressive politics and equity planning within city government was found in Chicago during the 1980s. Robert Mier, a planning professor who founded the Center for Urban Economic Development at the University of Illinois in Chicago, helped build a political coalition that in 1982 elected Harold Washington, Chicago's first black mayor. When elected, Washington took Mier and some of his associates into City Hall where they explicitly included "redistributive and social justice goals within the governments's policy, planning and implementation frameworks" (Giloth and Moe, 1999).

Mier and his fellow planners wrote the 'Chicago Economic Development Plan, 1984,' a model of equity planning. The plan proposed to use the full weight of the city's tax incentives, public financing, and infrastructure improvements to generate jobs for Chicago residents with emphasis on the unemployed. Specific hiring targets were set for minority and female employment; 60% of the city's purchasing was directed to Chicago businesses; 25% of that was to go to minority and women-owned firms.

The plan also sought to encourage a model of balanced, 'linked' growth between downtown Chicago and the city's neighborhoods. It offered public support to private developers interested in building projects in 'strong' market areas of the city, only if they would agree to contribute to a low-income housing trust fund or otherwise assist neighborhood-based community development corporations to build projects in 'weaker' areas.

Other American urban planners have adopted advocacy/equity approaches because they believe that planning along these lines holds the promise of better lives for the most troubled residents of their cities. In the 1980s and 1990s, equity planning cases were documented in cities such as Denver, Jersey City, San Diego, Berkeley, and Santa Monica (Krumholz and Clavel, 1994).

More recently, planners in some cities have been negotiating linkage deals and Community Benefits Agreements with developers, which offset subsidies or special rights granted by the city with special contributions by developers for low-income housing, job training, public transit, child care, or other

benefits for lower income people and neighborhoods (Wolf-Powers, Laura, 2010).

These cases make clear that there are often political and institutional barriers to equity planning practice, especially in a nation so strongly driven by market forces as the United States, and equity planners who question the status quo may face political reprisals. But obstacles to equity planning practice seem to lie primarily in the areas of the planner's personal confidence, motivation, and will.

A number of lessons seem clear. First, all forms of urban planning, especially advocacy/equity planning, prosper during periods of crisis, which bring forward the reform elements of the American state. It was during the Progressive Era that housing reform and urban planning were initiated. Later, when the Great Depression brought forth FDR's New Deal, planning was introduced for the first time into the federal government; sweeping regional projects like TVA got their start and the federal government involved itself in the provision of low-income housing. During the 1960s, the ferment around civil rights, antiwar protests, and the demolitions of urban renewal and the interstate highway program brought advocacy planners like Paul Davidoff, who combined commitment to social justice and democracy, to the fore.

Although the advocacy/equity planners often struggled against the more powerful currents of a strongly market-oriented society, they succeeded to an extent, and their work has been acknowledged and institutionalized. At the end of the twentieth century, American city planning was much more sensitive and collaborative than it had been earlier. Racial discrimination in housing and in mortgage lending has been prohibited by federal law, and citizens were encouraged to participate in the planning of programs that impacted their lives. Neighborhood-based development corporations have been successfully redeveloping parts of certain older neighborhoods, and some cities have been linking the benefits of downtown growth to their poor neighborhoods.

Also, since the global financial crisis of 2008, there is more interest in equity and advocacy in urban planning. Two cities, Minneapolis, Minnesota, and Portland, Oregon, have established formal offices on equity as part of regular government practice, and many more academics are writing about and studying progressive planning.

What remains troublesome, however, is the persistence of deep poverty in a growing number of urban neighborhoods, the continuance of patterns of racial segregation, the widening gap between rich and poor, the dismantling by conservative national administrations of the social safety net, the chronic competition among American cities that harms the vast majority of their citizens, and the growing spatial separation across metropolitan regions by race and class.

Given these challenges, a new generation of advocacy/equity planners is needed to help ease and resolve the most crucial of these social and physical issues, for what is at stake is nothing less than the future of urban life in America.

See also: Development and Urbanization; Local Economic Development; Neighborhood Revitalization and Community Development; Occupational Health; Planning Ethics; Planning Issues and Sustainable Development; Planning Theory: The Good City and Its Governance; Public Goods: International; Real Estate Development; Strategic Planning for Public and Nonprofit Organizations.

Bibliography

Ackerman, F.L., 1919. Where goes the planning movement. Journal of the American Institute of Architects, 519–520.

APA/AICP, 2005. Code of Ethics and Professional Conduct. Section A 5. American Planning Association, Chicago.

Birch, E.L., 1980. Radburn and the American Planning Movement. Journal of the American Planning Association 46 (4), 424–439.

Cleveland City Planning Commission, 1975. The Cleveland Policy Planning Report.

Davidoff, P., 1965. Advocacy and pluralism in planning. Journal of the American Institute of Planners 31.

Davidoff, P., Davidoff, L., Gold, N., 1970. Suburban action: advocate planning for an open society. Journal of the American Institute of Planners 361, 12–21.

Gans, H., 1962. The Urban Villagers. Free Press, Blencoe, New York.

Giloth, R., Moe, K., 1999. Jobs, equity and the mayoral administration of Harold Washington in Chicago. Policy Studies Journal 27 (1), 129–146.

Hayden, D., 1976. Seven American Utopias: The Architecture of Communitarian Socialism. MIT Press, Cambridge, MA.

Heskin, A.D., 1980. Crisis and response: a historical perspective on advocacy planning. Journal of the American Planning Association 46 (1), 50–62.

Howard, E., 1965. Garden Cities of Tomorrow. MIT Press, Cambridge, MA.

Jacobs, J., 1961. The Death and Life of Great American Cities. Random House, New York.

Krumholz, N., Cogger, J.M., Linner, J.H., 1975. The Cleveland policy planning report. Journal of the American Institute of Planners 41 (5), 298–304.

Krumholz, N., 1982. A retrospective view of equity planning: Cleveland 1969–1979. Journal of the American Planning Association 48, 163–183.

Krumholz, N., Clavel, P., 1994. Reinventing Cities: Equity Planners Tell Their Stories. Temple University Press, Philadelphia, PA.

Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge, MA.

Riis, J.A., 1971. How the Other Half Lives: Studies among the Tenements of New York. Dover, New York.

Scott, M., 1995. In American City Planning Since 1890. APA Press, Chicago, IL, p. 258.

Wolf-Powers, Laura, 2010. Community benefits agreements and local government. Journal of the American Planning Association 76 (2), 141–159.

Advocacy in Anthropology

Leslie E Sponsel, University of Hawai'i, Honolulu, HI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Advocacy and public anthropology are discussed as particular variants of applied anthropology using the term in its broadest sense. Examples of highlighted advocacy and public anthropologists include David Maybury-Lewis of Cultural Survival, Luisa Maffi of Terra Lingua, Robert Borofsky of the Center for Public Anthropology, Paul Farmer of Partners in Health, and Barbara Rose Johnston of the Center for Political Ecology. The main criticisms of such advocacy and public anthropology are identified and discussed. Then some future prospects for these endeavors are suggested as well.

Advocacy Anthropology

Advocacy anthropology is the application of anthropological resources on behalf of the survival, welfare, rights, and self-determination of indigenous communities, ethnic minorities, and other groups. It is a particular variety of applied anthropology advancing the interests of a community, often as a practical plea on its behalf to one or more external agencies. The community is usually indigenes, peasants, an ethnic minority, or refugees, those who are among the most oppressed, exploited, and abused. Advocacy is often connected to human rights, a framework internationally accepted in principle although, not always followed in practice. (Human rights refer to the inalienable universal rights that every human being deserves to have protected and promoted such as access to adequate water, food, shelter, and health care as well as freedom of speech, assembly, and religion.) Advocacy encompasses a broad agenda for social and political activism by promoting cultural survival and identity, empowerment, self-determination, human rights, sustainable economy, and the quality of life of communities. Advocates reject the supposed neutrality of science and adopt a stance on some problem or issue to improve the situation of a community, ideally in close partnership with it. Thus, as an advocate the anthropologist is no longer just observer, recorder, and interpreter (basic research), nor consultant to an external agency (applied), but facilitator, interventionist, lobbyist, or activist for a community (advocacy) (Paine, 1985; Wright, 1988).

Advocacy usually operates from an idealist rather than realist position. Although these terms from John Bodley are problematic, the concepts are valid and useful. Realists accept cultural assimilation and even ethnocide (forced cultural change or extinction) as a natural and inevitable correlate of 'civilization,' 'progress,' and 'development,' a position sometimes linked with social Darwinism. Idealists reject this, viewing ethnocide as a political decision, usually by a government violating human rights. Realists dismiss idealists as romanticists who are trying to preserve indigenes as their private laboratories, zoos, or museums. However, idealists do not argue that culture is static and should be preserved as such; they argue only for self-determination by people to promote their cultural survival, identity, welfare, and rights. Self-determination is seen as the most basic of all human rights, whether referring to an individual, group, or society, and in principle it is limited only to the extent that it does not interfere with the rights of others. The pivotal moral and political issue is the right and power of the state to dominate indigenes or others and in some cases to implement ethnocide. Advocacy attempts to intervene in this power imbalance through demystifying the process, exposing injustices, and offering resistance (Bodley, 2008: 251–290).

Advocacy has a long history. For example, Bartolome de las Casas (1474–1566) – theologian, missionary, and something of a historian and anthropologist – participated in the first decades of European colonialism in the Americas. He chronicled injustice against indigenes and argued in their defense. By the mid-nineteenth century, the Anti-Slavery Society and the Aboriginal Protection Society emerged in England as humanitarian organizations campaigning for just policies toward indigenes and others. During this period, the early anthropological societies of London and Paris debated similar concerns. In the 1960s, organizations focused on advocacy developed, most notably Cultural Survival in Cambridge, Massachusetts; the International Work Group for Indigenous Affairs in Copenhagen; and the Minority Rights Group and Survival International in London (e.g., Corry, 2011). Each publishes newsletters, journals, and/or documents to expose human rights violations and analyzes issues; funds collaborative research with communities; and works to influence national and international governmental and nongovernmental agencies and the public, especially through lobbying, the media, and letter writing. Collaborative ethnography (a close partnership in the planning, implementation, analysis, reporting, and evaluation of ethnographic research with the host community), participatory action research, and needs assessment can be especially relevant for advocacy fieldwork.

Most of the history of advocacy as well as recent cases and issues can be found in the publications of the above advocacy organizations (e.g., *Cultural Survival Quarterly*) and in the journals *Current Anthropology* and *Human Organization*. More recently, special issues of the following periodicals are pertinent to advocacy anthropology: *American Anthropologist* (2006, v. 108, no. 1); *Anthropological Forum* (2012, v. 22, no. 1); *Anthropology in Action* (2003, v. 10, no. 2; 2009, v. 16, no. 2); *Current Anthropology* (2010, v. 51 Supplement 2); *Polar: Political and Legal Anthropology Review* (2005, v. 28, no. 2); *Practicing Anthropology* (2004, v. 26, no. 3; 2010, v. 32, no. 2); and *Urban Anthropology* (2003, v. 32, no. 2). Of historic interest as well is

the initiation of a public anthropology section in the *American Anthropologist* starting in September 2010.

At least two criticisms contributed to the development of advocacy: first, accusations of genocide and ethnocide of indigenes by colonials and neocolonials in 'frontiers' like the Amazon; and second, from outside and within anthropology criticisms of basic and applied research together with calls for increasing social relevance and responsibility, most recently termed engaged anthropology (an omnibus category for various initiatives in socially responsive and responsible anthropology).

In 1969, Vine Deloria, Jr., a Sioux lawyer and interdisciplinary scholar, launched a searing critique of anthropologists engaged in either pure research with natives as objects in their private zoo or applied work for the colonial government; only concerned with advancing their career for self-aggrandizement and money; and irresponsible and unresponsive to the needs and problems of Native Americans. He asserted that anthropologists should obtain informed permission from the host community for research, plan and implement it in close collaboration with them, and focus on their practical needs, problems, and concerns (Biolsi and Zimmerman, 1997).

In 1971, a historically important but neglected conference of mostly Latin American anthropologists developed the Declaration of Barbados which, among other things, criticized anthropology for its scientism, hypocrisy, opportunism, and apathy in the face of the oppression, exploitation, ethnocide, and genocide of indigenes by colonials and neocolonials. For some anthropologists this became a manifesto to join the struggles for liberation and self-determination of indigenes and ethnic minorities through advocacy (Dostal, 1972).

Cultural Survival was launched by David and Pia Maybury-Lewis along with Evon Z. Vogt, Jr. and Orlando Patterson on 2 March 1972. The office was housed in the Peabody Museum of Harvard University. They were responding to the increasing pressure on indigenous societies throughout the world, especially in the Amazon, from external agencies seeking land and resources since the 1960s, often with the rationalization of so-called economic development and progress bolstered by racist and ethnocentric stereotypes. As Prins and Graham (2008: 119) observe: "David recognized that the anthropologist's profession comes with moral obligations that may call for political action in solidarity with those who have become friends, neighbors, and, sometimes, even relatives during fieldwork."

Cultural Survival was established as a partnership working with indigenous communities on behalf of their survival, welfare, sovereignty, self-determination, and human rights. The founders felt that this was a way to genuinely reciprocate for the hospitality of indigenous communities in allowing and facilitating their field research. It was also a dire necessity as indigenes increasingly faced genocide, ethnocide, and ecocide.

David Maybury-Lewis (1929–2007) was a public as well as an advocacy anthropologist. In articles in the *New York Times* and various other newspapers and magazines, he argued in favor of idealism and against realism, as previously distinguished here. His bottom line was that ethnocide through forced acculturation and assimilation is not inevitable, but a matter of political choices. Thus, he advocated an appreciation and tolerance for cultural diversity in a multiethnic society. Furthermore, he hosted the 10 one-hour documentary film series on the British Broadcasting Corporation and in the United States on the Public Broadcasting Service called 'Millennium: Tribal Wisdom and the Modern World' and wrote a companion book by the same title. The film is the product of 10 years of work in 15 countries. It employs first-person narratives with indigenous and Western persons to describe the nature, meaning, and significance of cultural diversity and the underlying universal themes. Thereby it exposed the massive audience of television to new ways of thinking about living and the environment. By now the film has been viewed by more than 50 million people in a dozen countries (Prins and Graham, 2008: 120). The initial showing coincided with the Columbian Quincentenary.

It should be noted that as a pioneering Amazonian ethnographer and a professor of anthropology at Harvard University, Maybury-Lewis was internationally recognized as a distinguished academic scholar and scientist as well as esteemed for his deep and tireless devotion to advocacy. Throughout his career, basic and applied research were mutually reinforcing (e.g., Maybury-Lewis, 2002).

Terralingua is an advocacy organization similar to Cultural Survival, except it is focused on the languages of the world within the framework of the concept of biocultural diversity, encompassing biological, cultural, and linguistic diversity. The organization advocates the study, documentation, and preservation of the linguistic diversity of the world's 6867 ethnolinguistic groups. Luisa Maffi, a linguist, anthropologist, and ethnobiologist, is cofounder and director of Terralingua. The organization was established in 1996 by several dedicated individuals from various backgrounds who had a common concern for the vital importance of biocultural diversity and its maintenance. In association with Terra Lingua, Maffi has published several seminal books in this arena. The organization publishes an e-magazine called *Langscape* (Terralingua, 2012).

Terralingua's work overlaps with public as well as advocacy anthropology. Their website identifies these fundamental beliefs: the diversity of life on earth has biological, cultural, and linguistic components (biocultural diversity); these coevolved in an interdependent and mutually reinforcing manner; there is a rapid, converging crisis of extinction of biocultural diversity and it is imperative to stop this crisis for the survival of all life on the planet; healthy environments, resilient cultures, and vibrant languages are a matter of social justice and basic rights for human societies, including indigenous and local communities; the challenge is to protect, maintain, and restore the earth's biocultural diversity. These beliefs guide the work of Terralingua through innovative initiatives in basic and policy-related research, activism on the ground, and education (Maffi, 2005; Terralingua, 2012).

The San and other peoples of the Kalahari region of southern Africa have been studied extensively by anthropologists from various countries for well over a century. Some of these anthropologists have been particularly conscientious in reciprocating in meaningful ways to the communities who hosted their field research. Most notable is the Kalahari Peoples Fund building on the ethnographic work of Richard B. Lee and others initiated by the Kalahari Research Group of Harvard University. Since 1973, the Kalahari Peoples Fund

has responded to various requests to raise money and to assist with advice and technical matters as major cultural changes spread into the area. Professional anthropologists have volunteered to serve as liaisons with the government for communities advocating in many ways on their behalf such as for health care, natural resource rights, legal matters, education, and emergency situations. The organization also provides information and research on Kalahari peoples for lawyers, students, volunteers, and tourists. Members make presentations in numerous national and international forums to inform academics and the public about the current situation of the Kalahari peoples. Many anthropologists working in the area consider it to be morally impossible not to be engaged on behalf of the people they work with (Biesele, 2003; Lee and Biesele, 2002).

Public Anthropology

During the last decade there has been a marked growth of interest in what may be viewed as a variant of advocacy anthropology; namely, public anthropology. Public anthropology is the application of anthropology for the general public focused on specific public concerns. However, in public anthropology, instead of a local community being the focus, society usually is, especially the anthropologist's own society. The purpose is to apply anthropological knowledge, theories, and methods to better inform the public about contemporary questions, problems, issues, and controversies. Public anthropology is not new, although the term and trend are. Franz Boas (1858–1942), and many of his students such as Margaret Mead (1901–1978), Ruth Benedict (1887–1948), and Ashley Montagu (1905–1999), were public intellectuals who addressed public concerns as engaged scholar activists (Low and Merry, 2010).

In the last two decades, however, Robert Borofsky has been the most vocal proponent advocating public anthropology to counter a discipline that seemed to be increasingly insular and esoteric to a general public and often consumed by internal bickering instead of critical sociopolitical issues. He has pursued advancing public anthropology through organizing sessions and speakers at the annual convention of the American Anthropological Association (AAA), editing more than two dozen books in the California Series in Public Anthropology for the University of California Press at Berkeley since 2001, his website Public Anthropology (2012), and his own publications. For Borofsky, the *sine qua non* of public anthropology is for a far greater percentage of anthropologists to address public problems for the public, thereby fostering transparency and accountability beyond academia. Borofsky pursues a professional ethical stance beyond the fundamental principle of do no harm. His view is that in addition, do some good.

In his exploration of anthropology as a cultural construction which he calls the anthropology of anthropology, Borofsky is challenging the status quo of the establishment, especially the AAA, and in particular, the structures that subvert public engagement. He encourages student activism through the Community Action project. The California Series in Public Anthropology offers positive role models for change. His Public Outreach Assessment is one way that he seeks to draw on political forces beyond academia by ranking doctoral programs in anthropology on the basis of their public outreach. (Information about these and other advocacy initiatives by Borofsky are available on his website Public Anthropology.)

The remarkable work of Paul Farmer for nearly three decades illustrates the overlap in the various manifestations of applied anthropology in general as previously mentioned, and also the importance of applied anthropology being anchored in basic research. (In general, applied anthropology is the use of anthropology to solve practical problems.) A PhD medical anthropologist who is a practicing MD as well, Farmer developed a sophisticated social, clinical, and epidemiological analysis of health and social inequalities with a focus on tuberculosis and HIV/AIDS. Among other countries, he has pursued field research in the squatter settlements of Haiti. Farmer was appointed UN Deputy Special Envoy to Haiti to assist Special Envoy Bill Clinton in improving economic and social conditions.

As a public health advocate for the poorest of the poor, Farmer is the Founding Director of Partners in Health (PIH), an international social justice and health nongovernmental organization. PIH provides free direct health care to the poor, pressures pharmaceutical companies for medicines, and lobbies policy makers. It pursues basic research and advocacy on behalf of the poor in the development of community-based comprehensive primary health care. By now, the organization has expanded beyond its original localities in Haiti to encompass more than 11 000 employees in 49 health centers and hospitals in 11 countries (Partners in Health, 2012).

Grounded in his ethnographic fieldwork on the front lines of poor communities, Farmer's extensive publications focus mainly on the relationship between health and human rights through scrutinizing the role of social inequities in the distribution and outcome of infectious diseases. His work critically challenges conventional thinking about human rights revealing how social, economic, and political injustices influence the suffering and illness of the impoverished and powerless. He argues that 'business as usual' in any venue simply manages inequality instead of confronting structural violence (Farmer, 2002). On 28 November 2011, *Foreign Policy* magazine listed him among the top global thinkers. Farmer is also the subject of a biography by Kidder (2003), which reached the bestseller list of the *New York Times*.

Another outstanding leader in advocacy and public anthropology is Barbara Rose Johnston, Senior Researcher at the Center for Political Ecology (2012). Primarily known as an environmental anthropologist, but overlapping with medical anthropology, Johnston has focused much of her research and publications on the interface between environment and human rights, and in particular, human environmental rights. Many inalienable universal human rights depend on a healthy environment. Environmental problems are not experienced equally; instead inequities render some persons more vulnerable than others.

Johnston's recent research focuses on the biomedical, psychosocial, cultural, and environmental impacts of the US nuclear weapons testing program, encompassing the history and consequences of its classified human radiation experimentation program. Since 1999, Johnston has served as a scientific advisor and expert witness for the Marshall Islands

Nuclear Claims Tribunal. She has pursued recognition and implementation of the right to a healthy environment, environmental equity, and the right to reparation and remedy. She played a very significant role as a researcher and expert witness in the award of over a billion dollars by the Nuclear Claims Tribunal to the victims on Rongelap in the Marshall Islands.

Johnston has published several books as editor or co-author, and was series editor of the eight volumes of regional surveys aimed at high school and college students titled *Endangered Peoples: Struggles to Survive and Thrive in a Globalized World*. She has published in public media such as *Counterpunch*. Her books are readily accessible to the general public as well as solidly grounded in scientific and academic scholarship. Contributing authors to her edited books are among the more prominent individuals in advocacy, public, and engaged anthropology.

Furthermore, Johnston has discussed the complexities, difficulties, and the personal and professional risks of work in advocacy and public anthropology. In particular, the 'anthropology of trouble' (Rappaport, 1988) can bring personal trouble to the anthropologist. It requires serious commitment, courage, persistence, and endurance as well as scientific and scholarly expertise and skill. It may also prevent or jeopardize academic jobs, given the usual criteria for employment, promotion, and tenure commonly focused exclusively on basic research, establishment grants, and publications in standard scientific journals (Low and Merry, 2010). (See Johnston, 2010, for a reflective overview of her work and citations for her publications.)

Another term, engaged anthropology, includes advocacy and public anthropology, but is much broader and more diverse. Engaged anthropology is an omnibus rubric for a broad variety of initiatives pushing for a more socially responsive and responsible anthropology. It includes sharing, support, and collaboration with host communities for research, teaching and public education, social critique, advocacy, and activism. It has attracted increasing attention for more than a decade. Growing engagement is transforming the discipline's approach and relationship with those studied in fieldwork and with the scholarly community as well as the general public (Low and Merry, 2010). There are diverse motivations for a more engaged anthropology, but a major one is simply that the discipline must go public or perish (e.g., Peacock, 1997). However, in their superb overview of engaged anthropology, Low and Merry (2010: S204) identify another motivation: "Personally, the reason we are interested in engaged anthropology is that we are committed to an anthropological practice that respects the dignity and rights of all humans and has a beneficent effect on the promotion of social justice."

Finally, although too complicated to consider in detail here, another arena for the various manifestations of applied anthropology should be at least mentioned; that is, global climate change, the newest agency of ethnocide and ecocide. Its reality can no longer be questioned by any informed person who is honest, given the consensus of hundreds of scientists throughout the world as represented in the Intergovernmental Panel on Climate Change. The only questions are how bad is it going to get, and how will humans adapt? The first book in anthropology to confront this challenge was edited by Susan A. Crate and Mark Nutall as recently as 2009. It is a comprehensive assessment of the anthropology of global climate change with emphasis on witnessing the lived-reality of local impacts on place-based indigenous and other rural communities thereby putting a human face on climate change. It is important to point out that global climate change has a multiplier effect; it exacerbates other problems as well. The authors affirm the veracity and utility of local environmental knowledge and concerns. Furthermore, the authors advocate action with creative suggestions on engaging the public and policy to promote economic and social justice and rights for the vulnerable communities impacted from the material to the spiritual aspects of their unique lifeways from a distance by the poisoning of the atmosphere by more powerful industrial countries obsessed with materialism and consumerism. Many of the anthropologists are convinced that such activism is morally and ethically imperative because they are citizens of the countries primarily responsible for causing the problem (also see Crate, 2011).

Criticisms

There have been criticisms of advocacy and public anthropology from within the profession, mainly for supposedly abandoning scientific objectivity and reducing or abandoning anthropology to some form of social work or political activism. For instance, Peter Elsass (1992) asserts that anthropology rests on criteria of science, objectivity through neutrality, and scholarship for the creation of knowledge; and advocacy on morality and the use of knowledge. He opines that anthropologists are ill equipped to deal with matters such as the politics of state penetration into indigenous areas. He worries that advocacy may make things worse in the community and jeopardize the credibility and prestige of basic research. Elsass believes that fieldwork, moral commitment, a sense of justice, political observation, and anger on behalf of the community are all part of the decision to advocate, but that anthropology as science and scholarship does not lead to advocacy. Carolyn Flueher-Lobban (2006), the foremost student of professional ethics in anthropology, has argued that advocacy is not a professional ethical responsibility, although it may be a personal moral responsibility and choice.

Rebuttal

The response to critics such as Elsass has focused on four things. First, in a general sense most if not all anthropologists are advocates and public intellectuals in some degree and manner. At least since Franz Boas, teaching anthropology advocates the value of the profession, cultural diversity, and cultural relativism; research publications advocate certain arguments, theories, and methods; and both teaching and research may challenge racism and ethnocentrism in favor of valuing equality, justice, and rights.

Second, no science or scientist is totally apolitical and amoral. In particular, when ethnocide, genocide, or other human rights violations occur, it is simply unconscionable and unprofessional for a knowledgeable anthropologist not to act, especially if directly involved with the society or community.

This is implicit in the code of ethics of professional organizations like the AAA, numerous resolutions at its annual meetings over many decades, and the mission and activities of its Committee on Ethics and especially the Committee for Human Rights.

Anthropologists are unavoidably immersed in webs of political relationships (Low and Merry, 2010). Alcida Rita Ramos (1999) argues that doing anthropology anywhere in the world is inevitably a political act, whether or not the anthropologist is aware of or recognizes that fact to any degree, and whether or not the anthropologist likes it. David A. Cleveland (1994: 9) has framed this issue particularly well: "Choosing between science and advocacy is not valid, because it does not serve the goals of either science or advocacy…. We are in fact always advocates—by not making active choices, we are passive advocates for the status quo…. The challenge is to keep up a dialogue between science and advocacy, keeping the two as distinct as possible without separating them completely."

Third, many communities believe that either the anthropologist is part of the solution, or part of the problem. Increasingly, communities are excluding anthropologists, unless they demonstrate social responsibility and relevance. Indeed, if anthropology is not of some genuine relevance to the communities from which research is derived, then some would suspect its credibility as well as utility. How could data collected in a given community be irrelevant to it? Even when a community can readily speak and fight for itself, it usually helps to have an outsider with some special knowledge, resources, and access to speak as well. However, a major dilemma for the anthropologist is dealing with messy inter- and intracommunity factions and conflicts.

Fourth, while avoiding paternalism, many anthropologists are involved in advocacy and public anthropology because they are sincerely concerned with applying knowledge on behalf of the communities who are indispensable for their research, as an expression of genuine reciprocity, and to avoid dehumanizing their hosts and themselves. An outstanding example of such advocacy is the collaboration of Bruce Albert and Gale Goodwin Gomez (1997) in researching and publishing a bilingual ethnomedical guide in Portuguese and Yanomami to facilitate the work of health care workers. Likewise, individuals who commit to the discipline of anthropology as their career and lifestyle, and as conscientious citizens, are often moved to demonstrate its relevance beyond academia to society as a whole through public anthropology (e.g., Johnston, 2010).

Conclusions

Advocacy and public anthropology are manifestations of applied anthropology along with practicing, action, collaborative, and engaged anthropology. There is much overlap among these often ambiguous and/or politicized approaches, the difference being mainly in their focus and clientele. They have all grown exponentially since the 1960s, and especially during the 2000s. It appears most likely that the anthropology of the twenty-first century is going to be quite different in many respects from that of the previous centuries, but grounded in deep historical roots even as it challenges prevailing orthodoxies.

In the future, advocacy needs to more systematically and explicitly develop its foundations and operations in terms of its history, philosophy, theory, methods, ethics, practice, and politics. Also, it should be recognized that there are many ways to contribute to advocacy, such as teaching, even for those critical of its complexities, difficulties, and risks. To be effective, advocacy and public anthropology must be grounded solidly in basic research, but they can also feed into further development of theory and method.

Advocacy, public, and other varieties of applied anthropology are most likely to increase in the twenty-first century because of at least three factors: first, growing population and economic pressures on land and resources with ensuing competition, conflicts, violence, and rights violations; second, increasing encroachment of the state, military, business, industry, and other external forces on indigenous and other communities; and third, insistence by local communities who host field research that anthropologists become more responsible and thereby responsive. Most 'frontier' or 'underdeveloped' zones in the world are in the traditional territory of indigenous societies. There are 350 million indigenous persons in the world, and although they comprise only 5% of the world population, they inhabit 20% of its land mass in 70 countries (Maybury-Lewis, 2002: 7–8). The pressure on indigenous peoples is growing and many conscientious anthropologists are responding accordingly. Advocacy and public anthropology are likely to continue developing well into the future as a significant component of the conscience of anthropology, or at least of some anthropologists.

Ultimately, advocacy anthropology and public anthropology are a matter of personal morality and professional ethics as well as science and scholarship, and, of course, also politics. As Pat Caplan (2003: 3) observes: "Yet the ethics of anthropology is clearly not just about obeying a set of guidelines; it actually goes to the heart of the discipline; the premises on which its practitioners operate, its epistemology, theory and praxis. In other words, *what* is anthropology for? *Who* is it for?"

See also: Anthropological Research, Ethics of; Colonialism, Anthropology of; Conflict: Anthropological Aspects; Cultural Relativism, Anthropology of; Development: Social-Anthropological Aspects; Ethnocentrism; Frontiers in History; Globalization and Health; Globalization, Anthropology of; Human Rights, Anthropology of; Imperialism, History of; Indigeneity; Land Tenure; Primitive Society; Refugees in Anthropology; State: Anthropological Aspects; Third World; Violence in Anthropology; War: Anthropological Aspects, Historical Development of.

Bibliography

Albert, B., Goodwin, G.G., 1997. Saude Yanomami: um manual etnolinguistico. Museu Parense Emilio Goeldi, ORSTOM, and UNICEF-Brasil, Belem-Para, Brasil.

Biesele, M., 2003. The Kalahari peoples fund: activist legacy of the Kalahari research group. Anthropologica 45 (1), 79–88.

Biolsi, T., Zimmerman, L.J. (Eds.), 1997. Indians and Anthropologists: Vine DeLoria, Jr. and the Critique of Anthropology. University of Arizona Press, Tucson, AZ.
Bodley, J.H., 2008. Victims of Progress, fifth ed. AltaMira Press, Lanham, MD.
Caplan, P., 2003. The Ethics of Anthropology: Debates and Dilemmas. Routledge, New York, NY.
Cleveland, D.A., 1994. Can science and advocacy coexist? The ethics of sustainable development. Anthropology News 35 (3), 9–10.
Corry, S., 2011. Tribal Peoples for Tomorrow's World. Freeman Press, Cookhill, Alcester, UK.
Crate, S.A., 2011. Climate and culture: anthropology in the era of contemporary climate change. Annual Review of Anthropology 40, 175–194.
Crate, S.A., Nuttall, M. (Eds.), 2009. Anthropology and Climate Change: From Encounters to Actions. Left Coast Press, Inc., Walnut Creek, CA.
Dostal, W. (Ed.), 1972. The Situation of the Indians of South America. World Council of Churches, Geneva, Switzerland.
Elsass, P., 1992. Strategies for Survival: The Psychology of Cultural Resilience in Ethnic Minorities. New York University Press, New York, NY.
Farmer, P., 2002. Pathologies of Power: Structural Violence and the Assault on Health and Human Rights. University of California Press, Berkeley, CA.
Flueher-Lobban, C., 2006. Advocacy is a moral choice of "doing some good": but not a professional ethical responsibility. Anthropology News 47 (7), 5–6.
Johnston, B.R., 2010. Social responsibility and the anthropological citizen. Current Anthropology 51 (Suppl. 2), S235–S247.
Kidder, T., 2003. Mountains beyond Mountains: The Quest of Dr. Paul Farmer, a Man Who Would Cure the World. Random House, New York, NY.
Lee, R.B., Biesele, M., 2002. Local cultures and global systems: the Ju/'Hoansi-!Kung and their ethnographers fifty years on. In: Kemper, R.V., Royce, A.P. (Eds.), Chronicling Cultures: Long-Term Field Research in Anthropology. AltaMira Press, Walnut Creek, CA, pp. 160–190.
Low, S.M., Merry, S.E. (Eds.), 2010. Engaged anthropology: diversity and dilemmas: an introduction to supplement 2. Current Anthropology 51 (Suppl. 2), S203–S226.
Maffi, L., 2005. Linguistic, cultural, and biological diversity. Annual Review of Anthropology 34, 599–617.
Maybury-Lewis, D., 2002. Indigenous Peoples, Ethnic Groups, and the State, second ed. Allyn & Bacon, Boston, MA.
Paine, R. (Ed.), 1985. Advocacy and Anthropology. Institute of Social and Economic Research, Memorial University, Newfoundland, Canada. (St. John's).
Peacock, J., 1997. The future of anthropology. American Anthropologist 99 (1), 6, 9–17.
Prins, H.E.L., Graham, L.R., 2008. Pioneer in Brazilian ethnography and indigenous rights advocacy: David Maybury-Lewis (1929–2007). Tipiti 6 (1–2), 115–122.
Ramos, A.R., 1999. Anthropologist as political actor. Journal of Latin American Anthropology 4 (2), 172–189.
Rappaport, R.A., 1988. The anthropology of trouble. American Anthropologist 8 (2), 189–191.
Wright, R., 1988. Anthropological presuppositions of indigenous affairs. Annual Review of Anthropology 17, 365–390.

Relevant Websites

Note: All websites accessed 26 June 2012.

http://www.endangeredlanguages.com – Alliance for Linguistic Diversity, 2012. Endangered Languages Project. Catalogue of Endangered Languages (ELCat), University of Hawai'i at Manoa; and The Institute for Language Information and Technology (The Linguist List), Eastern Michigan University, Manoa, HI, and Ypsilanti, MI.

http://anthropologyworldnews.tamu.edu/news/ – Anthropology World News, 2012. College Station, TX.

http://www.centerforpoliticalecology.org – Center for Political Ecology, 2012. San Jose, CA.

http://www.culturalsurvival.org – Cultural Survival, Inc., 2012. Cambridge, MA.

http://www.iwgia.org – International Work Group for Indigenous Affairs, 2012. Copenhagen, Denmark.

http://www.minorityrights.org – Minority Rights Group, 2012. London, UK.

http://www.pih.org – Partners in Health, 2012. Boston, MA.

http://www.publicanthropology.org – Public Anthropology, 2012. Kailua, HI.

http://www.sfaa.net – Society for Applied Anthropology, 2012. Norman, OK.

http://www.survivalinternational.org – Survival International, 2012. London, UK.

http://www.terralingua.org – Terralingua, 2012. Salt Spring Island, British Columbia, Canada.

Aesthetic Education

Tracie E Costantino, Rhode Island School of Design, Providence, RI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R.A. Smith, volume 1, pp. 206–210, © 2001, Elsevier Ltd.

Abstract

While acknowledging the foundational role of the arts in human development, this article will focus on twenty-first century conceptions and manifestations of aesthetic education. This discussion will highlight key theorists John Dewey and Maxine Greene, focusing on their scholarship from the twentieth century that continues to inform aesthetic education and powerfully articulates the ongoing significance of aesthetic education for human civilization. The main portion of this article will review the status of aesthetic education in the early twenty-first century amid an educational culture of accountability, standardization, and global economic competition, with a snapshot focus on the United States.

Introduction

Regardless of the waxing and waning of the presence of the visual and performing arts in formal education throughout the history of Western civilization, the importance of the arts in the development of human culture is undisputed. Ellen Dissanayake (1995) has attributed an ethological function to the arts, asserting that artistic behavior is a fundamentally human proclivity. Cognitive scientist Merlin Donald (2006) has argued that artistic behavior was part of a distributed cognitive system that influenced the evolutionary development of the modern *Homo sapien* brain. This distributed cognitive system – culture – continues to shape human development. With such profound influence on human evolution, the study of the arts is certainly justified. Aesthetic education may, following this argument, be defined as engagement with manifestations of human artistic expression throughout history for mutual cultural education, both local and global. Encountering works of art from one's own society, and the societies of others, educates individuals in the mores and values of diverse cultures, cultivating a tolerance for pluralism that has often aligned aesthetic education with ethics, democracy, and moral education.

While acknowledging the foundational role of the arts in human development, this article will focus on twenty-first century conceptions and manifestations of aesthetic education. This discussion will highlight key theorists John Dewey and Maxine Greene, focusing on their scholarship from the twentieth century that continues to inform aesthetic education and powerfully articulates the ongoing significance of aesthetic education for human civilization. The main portion of this article will review the status of aesthetic education in the early twenty-first century amid an educational culture of accountability, standardization, and global economic competition, with a snapshot focus on the United States. The focus on twenty-first century perspectives on aesthetic education is deliberate in order to underscore the ongoing relevance of aesthetic education, which has been critiqued as an unnecessary leftover from modernism (Tavin, 2007), to contemporary pragmatist, postmodern, poststructuralist, and social reconstructionist aims for education (see chapter by Costantino and chapter by jagodzinsky in Costantino and White, 2010).

A Brief History of Aesthetic Education

While an extensive documentation of the history of aesthetic education is beyond the scope of this article (see White, 2009), it is pertinent to begin with a brief history of the development of aesthetic education since the Enlightenment. The advent of modern aesthetic education may be considered as early as the eighteenth century when the German philosopher Alexander Baumgarten (1714–62) established aesthetics as a field of study within philosophy. Baumgarten's emphasis on aesthetics as a study of sense-perception initiated inquiry into how aesthetic perception may be developed through experience and education. Immanuel Kant (1724–1804) built on the work of Baumgarten to become the leading philosopher of aesthetics into the twentieth century, seeking to establish aesthetics in the realm of reason with his pivotal treatise *The Critique of Judgment*. Indeed, White acknowledges that "much of the contemporary controversy regarding aesthetics emerges from Kant's writings" (2009: p. 27).

Aesthetic education as a component of moral education gained prominence through the seminal work of Friedrich Schiller (1759–1805) *On the Aesthetic Education of Man* and the concept of *bildung* in German philosophy of the nineteenth century. In the first half of the twentieth century the influence of *bildung* continued in the aesthetic theories of hermeneutic philosophers Martin Heidegger (1889–1976) and Hans-Georg Gadamer (1900–2002). In the United States in the early twentieth century aesthetic education as moral education was evident in the picture study movement entailing the distribution of fine art prints to schools for the cultural edification of students. As Efland (1990) explained, "American art educators of that period were anxious to connect art study with the acquisition of American virtues, especially for the children of immigrants" (p. 146).

While the formalization of education through compulsory schooling in the nineteenth century influenced the opportunity for aesthetic education for different groups depending on local resources and priorities, since the advent of the standards and

accountability movement in the United States, there is documented discrepancy in the availability of aesthetic education for different socioeconomic and racial/ethnic groups, to be reviewed below.

What Is Aesthetic Education?

Aesthetic education is the cultivation of aesthetic understanding and the imagination through active engagement with the visual and performing arts. By engaging with works of art from diverse genres and cultures, students develop the ability to interpret and make meaning from an increasing variety of modes of expression, and through arts practice are able to communicate through multiple forms of representation. Elliot Eisner wrote in *Cognition and Curriculum, Reconsidered* (1994) of the importance of providing learners with different forms of representation through which they may communicate meaning. According to Eisner, "forms of representation are the devices that humans use to make public conceptions that are privately held …This public status might take the form of words, pictures, music, mathematics, dance, and the like" (p. 39). Eisner (1994, 2002) wrote extensively on the significance of this multimodal flexibility for cognitive development as students develop productive habits of mind such as attention to relationships, tolerance of ambiguity, resistance to premature closure, and feelings of self-efficacy to name a few. Eisner's ideas about the cognitive potential of arts education was significantly informed by John Dewey's writings on qualitative thought and his aesthetic theory of art as experience.

John Dewey: Training Aesthetic Perception for Democracy

John Dewey (1859–1952), American pragmatist philosopher and educator, sought to return the arts to everyday life in his influential work of aesthetic philosophy *Art as Experience* (1934). Critical of the removal of the arts from communal life into the rarefied "economic cosmopolitanism" (1934: p. 7) of museums and galleries, Dewey situated aesthetic experience as part of the continuity of the process of life. In his theory of art *as* experience, he gave perception a primary role in aesthetic experience. He explained the role of perception as the recognition of relationships. Simple recognition, however, is only the beginning of perception – it is the identification of elements without identifying how they are related. Perception is the interpretation of how various parts of an experience relate to form a unified whole. Discerning this relationship constitutes the meaning-making process: "The action and its consequence must be joined in perception. This relationship is what gives meaning; to grasp it is the objective of all intelligence" (1934: p. 44). The perception of relationships is what gives an experience its satisfying emotional quality. It is what makes an experience aesthetic. Dewey emphasized that the experience of art requires perception in order for it to constitute an aesthetic experience. In other words, Dewey declared that "to be truly artistic, a work must also be esthetic – that is, framed for enjoyed receptive perception" (p. 48). Dewey defined artistic as the relationship between making and perceiving: "The act of producing that is directed by intent to produce something that is enjoyed in the immediate experience of perceiving has qualities that a spontaneous or uncontrolled activity does not have" (p. 48). In essence, art requires the application of intelligence as opposed to random, disassociated thoughts or feelings; it derives from qualitative thought.

Qualitative thought, Dewey explained, is an associative process in which significant relationships are intuitively grasped in an experience "… the immediate existence of quality, and of dominant and pervasive quality, is the background, the point of departure, and the regulative principle of all thinking" (1930/1984: p. 261). Dewey described artistic thinking as an exemplar of qualitative thought. "The logic of artistic construction and esthetic appreciation is peculiarly significant because they exemplify in accentuated and purified form the control of selection of detail and of mode of relation, or integration, by a qualitative whole" (p. 251). Dewey considered the thinking required for artistic creation to be "one of the most exacting modes of thought" (1934: p. 45). Likewise, the act of perceiving a work of art requires the same kind of intelligence. It is a creative act, or an act of "reconstructive doing" (1934: p. 53), that requires cultivation. For Dewey, the appropriate end of training aesthetic perception is a richer everyday life. Dewey's dismantling of the fine versus applied art distinction in *Art as Experience* brings art into the realm of everyday life, thereby democratizing aesthetic experience as something attainable for all. Dewey's (1916) emphasis on education as a democratizing force, and his location of the arts in the midst of civilization was a critical influence for the philosophy of aesthetic education developed by Maxine Greene.

Maxine Greene: Releasing the Social Imagination

Maxine Greene (1917–2014), Professor Emeritus at Teachers College Columbia University and Philosopher in Residence at the Lincoln Center Institute, has defined aesthetic education as "deliberate efforts to foster increasingly informed and involved encounters with art" (1995: p. 138). Maxine Greene (1988, 1995) has written extensively on how experiences in and through the arts can stimulate the imagination to provide openings for thinking of what might be. Greene has explained, "One of the reasons I have come to concentrate on imagination as a means through which we can assemble a coherent world is that imagination is what, above all, makes empathy possible" (1995: p. 3). Her idea of the social imagination, "the capacity to invent visions of what should be and what might be in our deficient society" (p. 5) focuses on using the arts as entry points for dialogue that cultivates a pluralism that is the foundation of democratic communities. Through the empathic relationships constructed via aesthetic experience, schools and classrooms can become communities in which students may imagine the open possibilities of their lives, and both teachers and students may embrace the diverse perspectives and experiences of those they may consider 'other.'

While Greene has emphasized the transformative quality of aesthetic experience in dialogue with works of art (the visual, literary, and performing arts), encouraging teachers to share art with their students as a means of releasing their imaginations of what might be, she also addresses the potential of art making for cultivating the social imagination, for creating an 'in-between' space where community may be built. "In

thinking of community, we need to emphasize the process words: making, creating, weaving, saying, and the like" (1995: p. 39). Indeed, Greene calls for a merging of the pedagogies of art education and aesthetic education so that students and teachers are engaged in both the making and seeing of possibility:

> I would like to see one pedagogy feeding into the other: the pedagogy that empowers students to create informing the pedagogy that empowers them to attend (and, perhaps, to appreciate) and vice versa. I would like to see both pedagogies carried on with a sense of both learner and teacher as seeker and questioner…reflective about his or her choosing process, turning toward the clearing that might (or might not) lie ahead (p. 138).

Greene's work continues to inform and guide contemporary practices in aesthetic education. Stories of encounters with Greene's aesthetic education pedagogy are memorialized in *Dear Maxine* (Lake, 2010), a series of letters to Maxine Greene from those who have been inspired by her work.

Contemporary Practices in Aesthetic Education

Building on the ideas of Maxine Greene, contemporary social reconstructionist curriculum theorists, such as Patrick Slattery (2013), utilize artistic forms in their research, theorizing, and teaching praxis focused on issues of social justice to create 'sudden shocks' of awareness (Greene, 1995: p. 36). Others build on the relational aesthetics of Nicolas Bourriaud (2002) to explore how contemporary art practice may inform K-12 art education (Irwin and O'Donoghue, 2012; O'Donoghue, 2009). Boyd White (2009) has developed a research model, the aesthetigram, informed by art criticism to map participants' interpretive process during aesthetic experience. Aesthetics has also influenced the conduct of qualitative research in education. Drawing on the I–Thou relationship described in the philosophy of Martin Buber (1970 (1923)) and the intimacy of aesthetic experience articulated by John Armstrong (2000), Liora Bresler (2009; and chapter in White and Costantino, 2013) has explored the potential of aesthetic experience to cultivate empathy in her conceptualization of the researcher–participant relationship in qualitative research. While the examples of contemporary practice just given reflect the activity of scholars working in aesthetic and arts education, the aim of aesthetic education is typically geared toward elementary and secondary education.

Status of Aesthetic Education

While most developed nations include the arts and creativity in their national curriculum, or curricular goals in the absence of a national curriculum, the status of the arts and creativity as an educational priority varies. In 21 countries surveyed by the International Review of Curriculum and Assessment Frameworks Archive, most countries emphasize arts practice, especially when the arts subjects of music and visual art are taught as separate disciplines, rather than social/cultural outcomes of arts learning as might be emphasized in aesthetic education. However, when arts education is conceptualized as a general arts course, there is more emphasis on social/cultural outcomes (INCA, 2004) (as was articulated in the aesthetic education philosophies of Maxine Greene and John Dewey).

Singapore's national curriculum reflects the partnership between arts appreciation and arts practice envisioned by Maxine Greene. For example, the two main aims for the curriculum in visual art education at the primary and secondary level are for students to develop visual literacy and to appreciate art. Art practice is encompassed within the aim of visual literacy as students make meaning of what they observe in their environment as well as communicate meaning through art making. The aim of appreciating art is not exclusively for transmitting cultural values and cultural heritage, but for understanding the value of art in students' lives and as a means of developing self-worth. The approach is similar for music education. The conceptualization of arts education in Singapore as reflective of aesthetic education is also made evident in the administration of these programs under the category of aesthetics, health, and moral education in Singapore's Ministry of Education (Ministry of Education, Singapore, n.d.).

Similarly, in the Republic of South Africa, the arts fall under the curriculum category of life skills at the primary and early secondary levels, with the inclusion of craft, however in the upper secondary level the focus is on the relationship between art and culture. However, cultural literacy and creativity are not included as national education aims in the Republic of South Africa (INCA, 2013).

As an example of the arts and creativity as a priority in national curriculum aims, in 2001 the United Kingdom established the Creative Partnerships program to address concerns over a narrowing focus on mathematics and literacy in the National Curriculum. The program created long-term partnerships between schools and creative professionals from the arts, science, and industry. These programs had positive associations with student attendance rates at elementary schools and academic attainment especially at secondary schools (Sharp and Cooper, 2012); despite this the UK government stopped funding the program in 2010 as a result of the economic crisis affecting Europe and the United States. Alternatively, in a report published in 2010 by the Ministry of Education and Culture in Finland, the arts are tied to the creative economy and therefore viewed as a contributor to the growth of Finland's national economy. Finland is highlighted here due to current international focus on the effectiveness of Finland's educational system as reflected in Finnish students' consistent high-level achievement on the Programme for International Student Assessment.

In the United States, the report from the President's Committee on the Arts and the Humanities, *Re-Investing in Arts Education: Winning America's Future Through Creative Schools* (2011), puts creativity in the spotlight with arts education as a primary vehicle for American economic success. This goal is stated clearly in the executive summary:

> In order to effectively compete in the global economy, business leaders are increasingly looking for employees who are creative, collaborative and innovative thinkers. A greater investment in the arts is an effective way to equip today's students with the skills they will need to succeed in the jobs of tomorrow (p. 1).

The report advocates for increased arts integrated curriculum in schools and a greater role for teaching-artists among its recommendations (an oft utilized model in aesthetic education). Whether this report will have an actual impact on federal and state education policy remains to be seen within a climate of Race to The Top that focuses on STEM disciplines (science, technology, engineering, and mathematics) and ongoing standardization through the common core movement, especially as the current provisions for arts education in the United States demonstrate significant inequities across socioeconomic status.

Snapshot: Status of Aesthetic Education in the United States

A significant portion of this research was previously synthesized in the report by T. Costantino, "Educational Equity Challenged: Evidence of Decreased Provisions for Arts Education Since the Enactment of No Child Left Behind" printed in the following conference proceedings: E. DeBray-Pelot (Ed.) (November 2009) *Critical Perspectives on Federally Driven Assessment Policy*, (pp. 63–82). Athens, GA: Education and Evaluation Policy Center, University of Georgia.

In the United States, arts advocates had to lobby forcefully to have the arts included in the Goals 2000 legislation of 1994. One of the main goals of this legislation was the development of a challenging curriculum based on rigorous content standards that states were encouraged to develop. It was only at the eleventh hour that the fine arts were considered a part of the core curriculum in this legislation. Again in 2001, when the No Child Left Behind (NCLB) legislation was enacted, arts advocates had to lobby for the fine arts to remain part of the core curriculum for which content standards were now mandated. When NCLB was signed into law in the United States, there was considerable concern that the accountability emphasis on reading and math would result in a detrimental narrowing of the elementary school curriculum. In a review of research on the influence of this law on the implementation of curriculum in American schools, two trends emerged. The first trend is a narrowing of the curriculum showing decreased instructional time for art and music, and even greater decreases in instructional time for social studies, resulting in minimal cultural education for students. The second significant trend is that these instructional time decreases were more pronounced in schools serving a majority of low-income and minority students and schools designated as failing according to NCLB guidelines. The nonprofit organization, Center for Education Policy (CEP), has published a series of reports tracking NCLB, but theirs is not the only research to demonstrate these trends, including the nonprofit sector (CEP, 2007, 2008; von Zastrow, 2004), the federal government (Government Accountability Office and National Assessment of Educational Progress), and from the arts education arena (Mishook and Kornhaber, 2006). Building on this trend, correlational research has also been conducted examining the relationship between arts engagement and positive social and academic outcomes for adolescents and young adults of low socioeconomic status (Catterall et al., 2012). Beyond the K-12 arena, a survey on arts participation in the United States reinforces the implications of this trend for arts participation (Rabkin and Hedberg, 2011).

Reduction in Instructional Time

In a nationally representative survey of 349 responding school districts conducted in 2006–07, a report by CEP (2008) showed that 16% of responding districts cut art and music by 57 min per week, which represented a 35% decrease in weekly instructional time. However, more districts reduced time for social studies (36%) and for a greater duration at 76 min per week. For those districts that both increased time for math and English language arts, and decreased time for other subjects, these districts cut weekly instructional time for the arts by 35% (from 154 to 100 min).

The Council for Basic Education (CBE) is a nonprofit organization concerned with the status of a liberal arts education for K-12 students in the United States. The CBE surveyed a stratified random sample (urban, suburban, rural, elementary, and secondary) of public school principals in Illinois, Maryland, New Mexico, and New York about the status of liberal arts subjects in their schools' curriculum (von Zastrow, 2004). The survey had a 32% return rate (956 principals). In addition to asking about existing changes in instructional time since 2000, the CBE survey asked principals about their expectations for changes in the next 2 years.

Unfortunately, the report does not specify results according to the elementary or secondary level, so findings are aggregated for these two groups, unless there is a notable difference between the two groups. For all data, the report does present findings according to low or high minority schools, as this reflects an alarming trend that the CBE report identified and intends to highlight. Specifically, the CBE study found that 25% of all responding principals reported decreases in instructional time for the arts, with 36% of high-minority school principals reporting decreases. A large percentage of high-minority school principals anticipated future decreases (42%). For one-third of the high-minority schools reporting decreases, the principals reported large decreases and also expected future decreases to be large. The report summarized, "we found that the greatest erosion of the curriculum is occurring in schools with high minority populations – the very populations whose access to such a curriculum has been historically most limited" (p. 7).

A study by the United States Government Accountability Office (2009) reported similar findings to that of the CEP and CBE, although to a lesser degree. At the request of Congress, the Government Accountability Office (GAO) conducted a study of whether student access to arts education had changed since the enactment of NCLB and if so, how. Congress also wanted to learn about current knowledge of the impact of arts education on student learning outcomes. In other words, how might changes in provisions for arts education in the curriculum affect student learning? Congress also wanted to know what school district officials were doing to provide arts education since NCLB and their challenges in this regard (pp. 1–2). The GAO used data collected by the Department of Education's National Longitudinal Study of NCLB, which surveyed elementary school teachers about changes in instructional time for all subjects between school years 2004–05 and 2006–07. It also surveyed arts officials in state education agencies in 49 states and the District of Columbia and conducted interviews with school district officials and site visits to individual schools in Illinois,

Massachusetts, Florida, and New York. These states were chosen for their large numbers of schools not meeting Annual Yearly Progress and for a diversity of school districts with varying household income levels and rural and urban populations.

The GAO study found that the majority of elementary school teachers surveyed (90%) indicated that instructional time for the arts had not changed from 2004–05 to 2006–07. The teachers came from a range of school demographic characteristics. However, 7% reported a decrease, and there was a statistically significant difference in school characteristics in this group compared to the majority response of no change. "Specifically, teachers at schools identified as needing improvement and those with higher percentages of minority students were more likely to report a reduction in time spent on the arts" (p. 3). Also, the decrease in instructional time for the arts was greater at schools with high percentages of low income or minority students, a 49-min decrease per week compared to a decrease of 31 min weekly. Although the 7% finding is almost half as much as the 16% of districts reporting decreases in instructional time for the arts found by the CEP, the GAO survey studies just a two-year time span, which occurred a few years after the enactment of NCLB. Indeed, the decreased instructional time may have occurred before 2004 and then remained unchanged through 2006–07. The CEP study asked survey respondents to indicate instructional time changes since 2002. Findings from the interviews conducted in the GAO study indicated that changes in instructional time were due to state and district mandated requirements to meet NCLB proficiency standards.

Research conducted within the field of arts education found similar trends. Jacob Mishook and Mindy Kornhaber (2006) were interested in learning how high-stakes accountability was affecting curricular provisions for arts education. Dissatisfied with the snapshot offered by survey studies such as those described above, which do not explain the kind of arts education practice occurring in schools as a result of high-stakes testing, they conducted a qualitative study focused on the state of Virginia, chosen for its strong accountability system. They conducted an interview study centered on 10 arts-focused schools and 8 nonarts-focused schools matched by district, school level (elementary and secondary), and demographics including student income level and minority status, interviewing school principals and occasionally arts coordinators. They found that integration was a prominent strategy for accommodating arts education in the curriculum. They further analyzed the participants' conceptions of arts integration, using Liora Bresler's (1995) categorization of arts integration typically found in schools, the coequal, cognitive approach or the subservient approach (as cited in Mishook and Kornhaber, 2006). In the coequal approach, the art discipline is given equal weight in the lesson with the nonart discipline, maintaining the integrity of art content, skills, and ways of thinking. In the subservient approach, the art discipline is used as a tool in service of the nonart discipline, typically in a superficial manner. Mishook and Kornhaber found that the arts-focused schools were more likely to use a coequal, cognitive approach, while the nonarts schools typically used a subservient approach. Secondly, they found that the schools with a higher student poverty level were more likely to use a subservient arts integration approach, often in the service of test preparation, and that coequal arts integration was less available in high poverty schools. They acknowledged that these findings come from a small sample, but asserted, "If this holds true for a larger sample, this would align with research indicating a greater pressure toward using direct instruction for test preparation among high poverty schools…" (p. 7).

During the 2009–10 school year, the U.S. Department of Education's Institute of Education Sciences and National Center for Education Statistics (NCES) (Parsad and Spiegelman, 2012) conducted a study of the status of arts education (music, visual arts, dance, and drama) in public elementary and secondary schools through surveys to principals, classroom teachers, and arts teachers. While this is a substantial report with comprehensive and extensive findings, due to space restrictions I will highlight only a few of the findings here. The report provides comparisons with the previous NCES study conducted on the status of arts education in the United States a decade earlier (1999–2000). While instruction in music in public elementary schools remained consistent, there was a drop from 87 to 83% for visual art instruction in responding elementary schools. Instruction in dance and drama dropped from 20% of schools in 1999–2000 to 3 and 4% for dance and drama, respectively. For schools in which dance and drama were integrated into other subjects, there was a decrease of 5% for dance and a slight increase for drama. Similarly in secondary schools, the percent of schools providing music instruction remained consistent over the decade, while the percent of schools providing visual arts instruction decreased from 93 to 89%. Among the small percentage of secondary schools providing dance and drama education, there was also a decrease over the decade from 14 to 12% for dance and 48 to 45% for drama. On a positive note, the percent of secondary schools requiring arts coursework for graduation rose from 52 to 57% over the decade.

In relation to the trend of less instructional time in the arts for low-income students described above, the NCES report adds more evidence (Parsad and Spiegelman, 2012). Compared to 94% of public elementary schools without a majority of low-income students offering music instruction, 89% of schools with majority low-income student populations provide weekly music instruction. The comparison is greater at the secondary school level, with 81% of secondary schools with a majority of low-income students offering music instruction compared to 91% for those schools without a high percentage of low-income students. The decrease over the decade, however, is more significant, with a 19% drop in secondary schools with majority low-income student population offering music instruction. For visual arts education in elementary schools, there has been an increase in the percentage of schools with majority low-income students providing art instruction over the decade (from 74 to 80%). However at the secondary school level, there has been a decrease in the percentage of schools with a majority low-income student population providing visual arts instruction over the decade, from 93% in 1999–2000 to 80% in 2008–09. It appears that provisions for arts

instruction have been most impacted at the secondary level across the decade.

Impact of Reduced Provisions for Arts Education

The results of the 2008 National Assessment of Educational Progress for the arts (visual arts and music) indicate a difference in performance among ethnic groups and poverty status. While overall, the aggregate performance in both music and art by eighth grade students was in the midrange on the 0–300 scales used for responding in music and visual art and for creating in visual art, it was less for minority and low-income students. Related to aesthetic education, low-income students scored 28 points lower on the music responding scale and 29 points lower on the visual arts responding scale than those not eligible for free and reduced lunch (National Center for Education Statistics, 2009). The creating scores in the visual arts (creating in music was not assessed) for Black and Hispanic students were lower than the scores of White and Asian students as were the scores lower for low-income students.

In addition to student achievement in the arts, a study by Rabkin and Hedberg (2011) infers what decreased arts instruction at the K-12 level might mean for participation in the arts as adults. Responding to a finding in the 1992 Survey of Public Participation in the Arts (SPPA) that arts education was the strongest predictor of most types of arts participation (excluding arts performance), Rabkin and Hedberg analyzed data from four administrations of the SPPA (from 1982, 1992, 2002, and 2008) to explore that finding. They found that across the four administrations, there was a positive correlation with childhood or adult arts education and attendance at 'benchmark' arts events (classical or jazz music concerts, plays, opera, and dance performances). Arts participation tended to be higher among Americans who also experienced childhood or adult arts education. The analysis also documented a decrease in school-based arts education in childhood from 1982 to 2008, with the peak of the percentage of Americans receiving arts education in childhood in the late 1980s and a sharp decline subsequently. The time period of the decline coincides with the advent of the standards-based and heightened accountability reform movement in the United States. The report also documents the decline in school-based arts education across the four survey administrations to be more pronounced for Black and Hispanic students. The report clearly states that childhood arts education is the most critical for adult arts participation, as most Americans who pursue arts education as an adult had some arts education as a child. They also emphasized that,

> Although there are some free programs and some scholarship programs for needy students, children generally require both a financial commitment from their parents and additional parental support to attend non-school programs. While non-school arts education programs are vital resources in communities all across the country, schools are the only institutions that have the potential to deliver arts education experiences to virtually all children (p. 20).

The significance of this trend is heightened when one examines the correlational research of Catterall et al. (2012) who examined four longitudinal databases to investigate the relationship between arts engagement and positive social and academic outcomes for adolescents and young adults of low socioeconomic status. A sample of their findings include that the majority of youth from low-SES backgrounds but with arts-rich experiences attended college after high school. Also, students from both low and high-SES backgrounds with rich arts experiences are much more likely to show civic-minded behavior such as voting, volunteering, and engagement in local or school politics.

Conclusion

In a democratic country such as the United States that asserts equality as a core value, the uneven provisions of arts education for low-income, Black, and Hispanic students undermines the social reconstructionist philosophies of John Dewey and Maxine Greene that continue to guide aesthetic education in the twenty-first century. While this article provided a focus on aesthetic education in the United States, more research is needed on the status of aesthetic and arts education internationally. Efforts currently underway through the International Society for Education through Art focused on visual art education (http://www.insea.org/), and research agendas presented by the arts education area of the United Nations Educational, Scientific and Cultural Organization will contribute to this effort, presenting cross-cultural comparisons that may provide models of robust aesthetic and arts education for diverse cultural contexts.

See also: Creativity Training; Educational Philosophy.

Bibliography

Armstrong, J., 2000. Move Closer: An Intimate Philosophy of Art. Farrar, Straus and Giroux, New York.

Bourriaud, N., 2002. Relational Aesthetics [Esthetique Relationnelle] (S. Pleasance, F. Woods, M. Copeland, Trans.). Les presses du reel, Paris.

Bresler, L., 1995. The subservient, co-equal, affective, and social integration styles and their implications for the arts. Arts Education Policy Review 96, 31–37.

Bresler, L., 2006. Toward connectedness: aesthetically based research. Studies in Art Education 48 (1), 52–69.

Bresler, L. (Ed.), 2007. International Handbook of Research in Arts Education. Springer, Dordrecht, NL.

Buber, M., Kaufmann, W.A., 1970 (1923). I and Thou. Scribner, New York, (Original work published 1923).

Catterall, J.S., Dumais, S.A., Hampden-Thompson, G., 2012. The Arts and Achievement in At-risk Youth: Findings from Four Longitudinal Studies. National Endowment for the Arts, Washington, DC.

Center on Education Policy (CEP), 2007. NCLB Year 5: Choices, Changes, and Challenges: Curriculum and Instruction in the NCLB Era. Center on Education Policy, Washington, DC. Retrieved from http://www.cep-dc.org.

Center on Education Policy, 2008. Instructional Time in Elementary Schools: A Closer Look at Changes for Specific Subjects. Author, Washington, DC. Retrieved from: http://www.cep-dc.org.

Costantino, T., White, B. (Eds.), 2010. Essays on Aesthetic Education for the 21st Century. Sense Publishers, Rotterdam, NL.

Dewey, J., 1916. Democracy and Education. The Free Press, New York.

Dewey, J., 1984. Qualitative thought. In: Boydston, J. (Ed.), John Dewey: The Later Works, vol. 5. Carbondale and Edwardsville, IL: Southern Illinois University Press, pp. 243–262 (Original work published 1930).

Dewey, J., 1934. Art as Experience. Minton, Balch & Company, New York.

Dissanayake, E., 1995. Homo Aestheticus, second ed. University of Washington Press, Seattle.

Donald, M., 2006. Art and cognitive evolution. In: Turner, M. (Ed.), The Artful Mind: Cognitive Science and the Riddle of Human Creativity. Oxford University Press, New York, pp. 3–20.
Efland, A.D., 1990. A History of Art Education: Intellectual and Social Currents in Teaching the Visual Arts. Teachers College Press, New York.
Eisner, E.W., 1994. Cognition and Curriculum Reconsidered, second ed. Teachers College Press, New York.
Eisner, E.W., 2002. The Arts and the Creation of Mind. Yale University Press, New Haven, CT.
Greene, M., 1988. The Dialectic of Freedom. Teachers College Press, New York.
Greene, M., 1995. Releasing the Imagination: Essays on Education, the Arts, and Social Change. Jossey-Bass Publishers, San Francisco, CA.
International Review of Curriculum and Assessment Frameworks Archive (INCA), 2004. Thematic probe: curriculum and Progression in the arts: an international study. http://www.inca.org.uk/thematic_probes.html.
International Review of Curriculum and Assessment Frameworks Archive (INCA), 2013. Comparative Tables. http://www.inca.org.uk/.
Irwin, R., O'Donoghue, D., 2012. Encountering pedagogy through relational art practices. International Journal of Art and Design Education 31 (3), 221–236.
Lake, R. (Ed.), 2010. Dear Maxine. Teachers College Press, New York.
Ministry of Education, Singapore. Subject Syllabuses—Aesthetics, Health and Moral Education. Retrieved from: http://www.moe.gov.sg/education/syllabuses/aesthetics-health-and-moral-education/, n.d.
Mishook, J., Kornhaber, M., 2006. Arts integration in an era of accountability. Arts Education Policy Review 107 (4), 3–11.
National Center for Education Statistics, 2009. The Nation's Report Card: Arts 2008 (NCES 2009-488). Retrieved from: http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009488.
O'Donoghue, D., 2009. Are we asking the wrong questions in arts-based research? Studies in Art Education 50 (4), 352–368.
Parsad, B., Spiegelman, M., 2012. Arts Education in Public Elementary and Secondary Schools: 1999–2000 and 2009–10 (NCES 2012–014). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC.
President's Committee on the Arts and the Humanities, May 2011. Re-investing in Arts Education: Winning America's Future through Creative Schools. Author, Washington, DC.
Rabkin, N., Hedberg, E.C., 2011. Arts Education in America: What the Decline Means for Arts Participation. National Endowment for the Arts, Washington, DC. Retrieved from: http://www.arts.gov/research/research.php?type=R.
Slattery, P., 2013. Curriculum Development in the Postmodern Era: Teaching and Learning in an Age of Accountability, third ed. Routledge, New York.
Sharp, C., Cooper, L., 2012. Creative partnerships longer-term tracking study: attainment and attendance. In: Research Digest: 2006–2012. Creativity, Culture and Education, Newcastle Upon Tyne, England, pp. 90–94.
Tavin, K., 2007. Eyes wide shut: the use and uselessness of the discourse of aesthetics in art education. Art Education 60 (2), 40–45.
United States Government Accountability Office, 2009. Access to Arts Education (GAO-09-286). Retrieved from: http://www.gao.gov.
White, B., 2009. Aesthetics. Peter Lang Publishing, New York.
White, B., Costantino, T., 2013. Aesthetics, Empathy, and Education. Peter Lang Publishing, New York.
von Zastrow, C., 2004. Academic Atrophy: The Condition of the Liberal Arts in American Schools. Council for Basic Education, Washington, DC.

Relevant Websites

http://www.artsusa.org/ – Americans for the Arts.
http://www.creativitycultureeducation.org/changing-young-lives-2012 – Creativity Culture & Education Research Reports.
http://www.inca.org.uk/thematic_probes.html – International Review of Curriculum and Assessment Frameworks Archive (INCA).
http://www.insea.org/ – International Society for Education Through Art.
http://muse.jhu.edu/journals/the_journal_of_aesthetic_education/ – Journal of Aesthetic Education.
www.lcinstitute.org – Lincoln Center Institute.
http://www.maxinegreene.org/ – The Maxine Greene Center for Aesthetic Education and Social Imagination.
http://www.arts.gov – National Endowment for the Arts.
http://www.unesco.org/new/en/culture/themes/creativity/arts-education – UNESCO/Arts Education.

Affect and Emotion, Anthropology of

Kathleen Stewart and Elizabeth Lewis, University of Texas at Austin, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Attention to emotion and affect is a wide register running through the history of anthropology. Until recently, however, emotions and affect have not often been explicitly articulated as a force in themselves in theories of culture, history, and ways of life. This article traces the history of affect and emotion in anthropology and reveals emergent trajectories for future work. It situates affect and emotion historically and highlights key writings on these themes in cultural, physical, and linguistic anthropology, and in archaeology. Ultimately, it positions affect as a fruitful analytical tool for anthropology and as a generative source of cross-subfield dialogue.

Emotion and Affect as an Analytical Object

In anthropology, attention to emotion and affect is a wide register running through the history of the field. But emotions and affect have not often been explicitly articulated as a force in themselves in theories of culture, history, and ways of life. They have, instead, been taken to be natural and universal states of being, atmospheric backgrounds to life, pervasive but subterranean and coded forces, or the secondary constructed effects of an order of representation or structures of power including class, race, gender, ethnicity, religion, sexuality, and the state. Until the recent emergence of an anthropology of emotion and affect and the replacement of theories of cultural wholes with theories of assemblages, force fields, and micropolitics, emotion and affect were rarely recognized as anthropological objects, but were treated instead as idioms for the then more obviously significant phenomena of structures of power and sociality. Broad affective questions of force, impact, and intensity have long been alluded to in studies of known anthropological objects such as violence, attachment to place and identity, or resistance and revolution. The emotional or affective force of social–cultural orders has been approached through studies of kinship, ritual, hierarchy, exchange, honor, and shame systems, and specific traditions and institutions such as head-hunting and veiling.

These investigations have taken place, for the most part, within a dynamic that pits close contact against distant observation and personal feelings against public knowledge. The idea that perspective and theory are themselves essentially produced through an originary act of distancing has been haunted by affective and emotional marginalia. Raymond Williams's (2001) work on structures of feeling has been long accepted and used to specifically address the qualities of emergence and the residual surrounding reductive models of structure in itself. However, even the relatively recent ethnographic work of anthropologists such as Renato Rosaldo (1993), Michelle Rosaldo (1980, 1984), Jose Limon (1994), and Steven Feld (1982) operated as eruptions in the field that drew strong reactions when they made emotion and affect the explicit object of analysis. These and other scholarly works articulated with both the writing culture critique (Clifford and Marcus, 1986) and feminist interventions that also began to directly document the power of affect and emotions (see, for example, Lutz and Abu-Lughod, 1990).

The generative clash between structure and history, observation and participation, and expert knowledge and the felt impact of everyday life opened divergent trajectories for current anthropology. Anthropology is still driven by a dynamic between emergent phenomena, complex causality, and an attention to the felt presence of historical events and present moments, on the one hand, and an attunement to the clarity of structures, institutions, cultural categories, norms, and conventions on the other. Work on affect and emotions is both symptomatic of this dynamic and itself an emerging field of theory and ethnographic experiment. Most recently, the development of affect studies in cultural studies (Berlant, 2008, 2011; Gregg and Seigworth, 2010) has established a strong affect studies trajectory in anthropology (Mazzarella, 2009; Stewart, 2007, 2010, 2011).

Emotion and Affect Studies in the Four Fields

In cultural anthropology, work on the emotions has long been grounded in cultural relativism, interpretive models, and social constructionism. Questions of emotion and attachment have been central to debates on the role of ethnography and the nature of the ethnographer–research subject relation. Fieldwork has been identified as itself an emotional experience that hinges on efforts to approach the experience of others and the variety of forces that shape the broad outlines of a way of life and remain resonant in daily life and conventions. Approaches to cultural expressivity and form have demonstrated the emotional-affective register of meaning emergent in music, sound, poetry, dance, storytelling, and the performativity of everyday life.

Within cultural anthropology, emotions are often described as either the pliant material upon which acculturative and cognitive forces operate, or the energy which animates otherwise lifeless cultural forms. The explicit attention to emotions as analytical objects began with a treatment of emotions as cultural categories, even a master western cultural category, in the effort to advance a cross-cultural understanding of emotion. Emotions have also often been described as an unconscious normative force. Michelle and Renato Rosaldo's work made emotions central to the theory of the situated subject and situated ethnography. This made way for the predominance not only of native ethnography, but also of new models of positioned epistemology, including emotional thinking, embodied

knowledge (Lutz and White, 1986), and a political imperative to draw on personal experiences, proximity, and intimacy in defining and exploring the subjects and objects of fieldwork. The work of documenting named or categorically represented emotions, such as grief, anger, hope, fear, loyalty, and love, expanded into analysis of structures of feeling and the phenomenology of lived states of being including work on nostalgia, utopianism, religious piety, structures of reciprocity, nationalism, belonging, and visceral attachments.

From an evolutionary perspective, physical anthropologists treat nonhuman emotions and affects as critical components of natural selection. Evolutionary theory, including Charles Darwin's *The Expression of Emotion in Man and Animals*, argued that emotions are a critical component of natural selection as adaptive responses to social and environmental opportunity and threat. It is theorized that social emotions evolved to motivate adaptive social behaviors. Primate behaviorists have long noted a complex range of physical displays of emotion including fear, grief, happiness, anger, and disgust. Emotions and their expressive body language have been taken to be strong evidence of the lines of human–nonhuman primate evolution.

Physical anthropologists have become increasingly interested in the study of emotions in recent decades. New work in neuroscience, primatology, and cognitive psychology has generated approaches to emotional control, social bonding, romantic love, and empathy among primates. Neuroimaging is now used to investigate evolutionary ancient parts of the brain. Neurobiological theories focus on arousal patterns in which neurochemicals (such as dopamine, noradrenaline, and serotonin) step up or step down the brain's activity level, as visible in body movements, gestures, and postures. Such paleocircuits are neural platforms for bodily expression. It is thought that there are certain activities in brain areas related to attention and motivation, especially the limbic system, which includes the hypothalamus, cingulate cortex, hippocampi, and other structures. New theories of the science and biology of emotions have generated innovative understandings of the linkages between behavior and emotions. Some also suggest implications for the ethical underpinning of human life and morality as an essential part of primate nature (De Waal, 2003).

Linguistic anthropology addresses the force and effects of language dominance, loss, and revival, the affective ties that enable language socialization, and the emotional registers of class, gender, race, ethnicity, and other hierarchies and identity markers in language and its performance. The production and comprehension of feelings are basic for language acquisition. The role of emotion in language has been addressed as a communication code including phonological, syntactic, and pragmatic as well as semantic elements. Emotions, like language itself, can be seen as having a semiotic system. Somewhat outside the purview of semiotics, utterances can also be seen as having an affective component, or an impact that moves between and across subjects, expressive forms and situations, and language structures. Linguistic anthropologists have taken up questions of the circulation and the transformative qualities of language forms. James Wilce (2009) advanced an analytical attention to 'feelingful language' as an embodied practice that avoids reduction of feelings to the level of discourse. Paul Kockelman (2011) takes a semiotic approach to theorizing the relationships between ontology, selfhood, and affect, arguing against an analytical separation of these three dimensions.

Archaeological theories of state formation rely on evidence of violence and conflict. They are also attuned to the levels of intensity of historical events and therefore to the study of affect as a register of historical force. With a strong base in the analysis of materials, archaeologists have also launched a powerful critique of radical constructionism; forms of agency and identity are not treated as ciphers through which other social and cultural forces are inscribed, but remain concretely embodied in practices, environments, and strategies of living. Recent theories of prehistorical affective structures and the production of emotions have critiqued naturalized frameworks for feelings and the theoretical shorthands that framework has enabled. One strand of this work calls for much more detailed and precise studies of individual and variable selves, questions of intentionality, and attention to embodied emotional experiences (Meskell, 1996; Tarlow, 2000). Another strand of contemporary archaeological theories of affect and emotion locates them in the material infrastructure of social spaces (e.g., architecture, lighting, and sensuous qualities) and movement (the corporeal staging of particular channels of experience). This work begins with the insight of new materialism that the sensuous quality of things and bodily presence are both, and coconstitutively, productive of perceptions, experiences, and atmospheres (Harris and Sørensen, 2010; Tilley, 1996).

Affect Studies in Anthropology

Emotion studies in anthropology tended to be grounded in questions of culture and cognition and to refer to named states such as anger, grief, love, or shame, which were experienced by subjects and charged with the judgments and norms of a social–political–cultural–representational order. That is, in emotion studies feelings are part of a cultural–political apparatus. Affect studies suspends the assumption of a cultural whole, taking a step back to questions of the intensities and felt impacts of a broad assemblage of circulating forces, which remain disparate though, at times, articulated. Affect studies takes as its object these intensities and impacts and their circulations and articulations. Often strongly influenced by the work of Deleuze and Guattari (1987) and a line of philosophical influence including the work of William James, Nietzsche, and Spinoza, affect studies define affect as the capacity to affect and to be affected. Where emotions are presumed to be a form of social construction – named, value-laden states of being closely tied to a politics of representation – affect is a domain, register, or machinery of intensity, speed, rhythm, and emergent form. It is the force that moves bodies of all kinds – human bodies, bodies of thought, institutions, populations, or bodies of force – into and out of states of being. Affect is not ontologically distinct from or secondary to cognition or meaning. Affect theory assumes not that an abstracted order of meaning creates worlds by instantiating and reproducing itself, or that the mind and its states of consciousness get the world right or wrong, but rather that the processes of emergence or becoming are complex compositions assembled out of bodies, materialities, scenes, events, and the substantial micropolitics

of force fields saturating everything from institutions to collective moods.

The objects of affect theory are simultaneously very abstract and very concrete. They are relational, in motion, and about transversal, transduction, and resonance. An affective event is an increase or decrease in the ability of the body to act. Brian Massumi (2002) theorizes it as the sensation produced in the encounter of bodies and forces and as the tipping point between the potential and the actual – the energy of something moving into being. Its objects are things in the process of becoming something else or forms that are emerging and dissipating.

Affect studies in anthropology pulls into alignment with Foucaultian and Nietzschian articulations of the intimate relations between knowledge and power. These theories focused intellectual projects on the conditions of the emergence of forms built into the conduct of life. The coming into being of forms in the details of daily life moves beyond the oppositions of fabulated/real, structure/agency. Structure takes place as singular events saturated with everyday sensibilities. Politics is not reducible to a communal consciousness or a neatly conceptualized ideology but takes place as intensities of all kinds and in various registers. Agency is not the clear and intentional act of a subject but an energetics. The anthropological objects of affect studies now include, for example, the pressure points of life, the structures of impact and attention being laid down in microbiopolitics, new sensory registers, and the systematic engineering of affect.

Capitalism, in affect studies, is itself a set of constantly performing relays and junctions that are laying down new affective fields of potentiality and closure. Nigel Thrift's (2004, 2011) work carefully outlines the swollen role of affect in the complex dynamics of contemporary capitalism and concomitant shifts in forms of knowledge and experience. In affect studies, systems of power are not flat sets of beliefs and effects but layered, irregularly matched, distinctive capacities that become possible and perhaps only partially legible. Power as system can be seen as the composition of dense loops between TV, cinema, philosophy, neurophysiology, and everyday life that constitute a neuropolitics (Connolly, 2002). This perspective requires new forms of critique and new modes of ethnographic attention. Latour (2005) and other nonrepresentational theorists replace evaluative judgment with an ethical–political attunement that tries to instill generosity toward a world by attending to its infrasensible knowledges.

Politics resides not in the rational subject but in the nonsubjective sensuous mimetic potential of the interlocking patterns, thresholds, and triggers of emergent, transversal bodies (Protevi, 2009). From the perspective of affect as a transductive machinery of relay, new political registers place questions of power and force in the contexts of ways of living through situations and the composition of worlds out of whatever is happening, presenting a fuller view of what is at stake for people in everyday life. Violence and conflict are seen not as exceptional moments of crisis, but as integral parts of a social world comprised of fault lines and edges. Social and cultural ties are not presumed by association with abstracted structures, but have to be tracked as moments of affinity and difference emergent in performance. The questions shift away from tracking the effects of the cultural modeling of a given reality to attending to the processes of social and material life: how places, practices, identities, and forms of cohabitation become lived worlds, how people sense the world they inhabit, how people come to recognize themselves, and their affinities with others and things (Ingold, 2000).

Emergent Anthropological Practices

Affect studies opens avenues to approach enduring anthropological questions: how to trace the forces that traverse the individual and the collective, and person, place, and thing; how to track and model complex articulations and assemblages; how to describe forms of signification that extend beyond code and representation and beyond individual consciousness; how to attend to sensory impacts and forms of knowledge; how to approach questions of force and intensity in ethnographic accounts; and how to build models of meaning grounded in the emergent practices of situations of the present. Affect studies is closely articulated to emergent efforts to address these questions with new paradigms including nature–culture studies, science studies, new materialism, posthumanism, and public culture.

In ethnography, affect studies is among the efforts to retheorize the anthropological object through more nuanced and prismatic approaches to the angles of persons, performances, bodies, atmospheres, rhythms, environments and social, affective, material, and institutional infrastructures. Attention to affect produces an attunement to the potential or emergent forces in a scene and becomes a means of recognizing the not quite visible yet forceful processes at work in even the most seemingly abject, private, or embodied iterations. For example, in *My Cocaine Museum*, Michael Taussig, describes the structure of feeling of a present moment in Colombia when the agribusiness boom has suddenly just stopped:

> … now there is no work at all …. That's all over now. The idea of work work. Only a desperate mother or a small child would still believe there was something to be gained by selling fried fish or iced soya drinks by the roadside, accumulating the pennies. But for the young men *now*, there's more to *life*, and who really believes he'll make it past 25 years of age? … At fourteen these kids get their first gun. Motorbikes. Automatic weapons. Nikes. Maybe some grenades as well. That's the dream. Except that for some reason it's harder and harder to get ahold of, and drug dreams stagnate in the swamps in the lowest part of the city like Aguablanca, where all drains drain and the reeds grow tall through the bellies of stinking rats and toads. *Aguablanca*. White water. The gangs multiply and the door is shoved in by the tough guys with their crowbar to steal the TV as well as the sneakers of the feet of the sleeping child; the *bazuco* makes you feel so good, your skin ripples, and you feel like floating while the police who otherwise never show and the local death squads hunt down and kill addicts, transvestites, gays – the *desechables*, or "throwaways" – whose bodies are found twisted front to back as when thrown off the back of pickups in the sugar cane fields owned by but twenty-two families, fields that roll like the ocean from one side of the valley to the other as the tide sucks you in with authentic Indian flute music and the moonlit howls of cocaine-sniffing dogs welcome you.
>
> (2004: p. 19)

This kind of affective ethnographic writing requires a supple attention and the capacity to imagine trajectories and follow tendencies. When the object of anthropological analysis is the levels of sense, rhythms, aesthetic forms, and energetics, of what

is emerging or falling apart, ethnography takes as its object the generativity of things, ontologies of multiplicity, planes of expressivity, the ways that a tendency takes on consistency, what is being magnetized in a thrown-together thing, what frictions run its machinery, and what happens when things suddenly stop.

Ethnographies of affect have yielded new concepts of ways of living through situations. Affect emerges at the juncture (and disjuncture) of personal experience and public circulation (Gregg and Seigworth, 2010; Mazzarella, 2009). Bodies not only unfold in and inhabit the world through contingent lineages, potentialities, entanglements, and aspirations but they actually acquire their very form through their affective engagements (Ahmed, 2010; Povinelli, 2006; Stewart, 2007, 2010). The entanglements between bodies, discourse, and history are bound up with currents of expectations, landscapes, and histories (Povinelli, 2006); dispositions accumulate (Stewart, 2010); in the charged territory between the potential and the actual, seemingly broken bodies result from a complex constellation of potentialities unfolding over time (Biehl, 2005); precarious ways of living through things that produce prolific strategies and fantasies (Berlant, 2011); affect generates worlds (Stewart, 2010).

Affect is a lens for approaching social worlds and lived experiences as ongoing processes, highlighting partiality, flux, and contingency. It is perhaps particularly well suited to examine the dynamics at work in logics of materiality, cohesion, rupture, or the exhaustion of forces and forms. Cultural anthropologists have used the concept of affective labor to probe the relationships between feelings, work, capitalism and, often times, gender in contemporary life (Hardt, 1999). Elizabeth Povinelli (2011) analyzes the temporality, ethics, and quasi events of affective attachments to probe the ordinariness of isolation and abandonment. William Mazzarella (2009) argues that affect is a necessary moment of any modern institutional practice with aspirations to public efficacy. That is, modernity has always been structurally affective; any social project that is not imposed through force alone must be affective in order to be effective. Lauren Berlant (2011) has elaborated how affective structures shape realities, generating frames in which certain types of possibilities come into being. As work on affect probes the contingent lineages of contact and affinity, gesturing toward moods and collective sensibilities rather than causation, it adds critically important nuance and texture to such traditional frameworks as political economy, structural violence, or biopolitics (Biehl and Locke, 2010). As anthropologists continue to grapple with how to produce theories of large-scale social experience while accounting for the inner worlds and daily lives of individuals, work on affect offers new tools for a simultaneous analysis of both poles (Biehl, 2005; Biehl and Locke, 2010; Das, 2006; Povinelli, 2006).

Conclusion

The long impasse in anthropology of the categorical opposition between biology and culture has played a role in naturalizing emotions to the extent that they are associated with the body. This has also limited approaches to knowledge, agency, and sociality. Where thought and knowledge are stereotypically associated with intentionality and logic, and consciousness with agency, body, and emotion are imagined as both untamed and too tamed, outside the order of representation that stands in as a shorthand for culture and society and also beneath recognition as the banal and all-too-well known. Until the relatively recent surge in interest regarding an anthropology of emotions, they were presumed to be universal aspects of the human experience that were both central to an individual's interior world and also transparently intelligible across cultures (Lutz, 1986). Studies of emotion and affect have produced a middle ground between biology and culture, and the individual and the social. Analyses of affective structures have expanded the reach and depth of studies of nationalism, violence, belonging, and affinity on a global scale and across virtual worlds, parsing out the resonances of distinctly dispersed, yet complexly interconnected, modes of life and practice.

Linking to interdisciplinary work in science studies and nonrepresentational theory, affect theory has been used to posit and produce an alignment between theory and the compositional attunement through which people and things venture out into reals. Reals, in this view, are not the kind of thing that an order of representation simply organizes as truth and dominates, but "transversal arrays of qualities or activities which, like musical refrains, give order to materials and situations, human bodies and brains included, as actions undertaken act-back to shape muscles and hone senses" (Anderson and Harrison, 2010: p. 8). This is not the work of imagination on dead matter but a "mattering (that) is about the (contingent and temporary) becoming-determinate (and becoming-indeterminate) of *matter* and *meaning*" (Barad, 2010: p. 254). This is a world built out of difference and repetition (Deleuze, 1994), composed of potentiality and loss (Berlant, 2011), and leaning toward entities that are both present and absent – atmospheres, affects, virtual memories, hauntings, that are themselves moments of endurance (or not), instants of the holding together of the disparate itself (Doel, 2010). This is a world that is not reducible to its representation, or what we know of it (Anderson and Harrison, 2010). One that is potentially a newly capacious object for the distinctive lines of approach of archeological, linguistic, physical, and cultural thought and method.

See also: Anthropological Writing; Materiality and Culture; Personhood, Anthropology of; Senses, Anthropology of The.

Bibliography

Ahmed, S., 2010. The Promise of Happiness. Duke University Press, Durham, NC.

Anderson, B., Harrison, P., 2010. The promise of non-representational theories. In: Anderson, A., Harrison, P. (Eds.), Taking Place: Non-Representational Theories and Geography. Ashgate, Surrey, England, pp. 1–34.

Barad, K., 2010. Quantum entanglements and hauntological relations to inheritance: dis/continuities, spacetime enfolding, and justice to come. Derrida Today 3, 240–268.

Berlant, L., 2008. The Female Complaint: The Unfinished Business of Sentimentality in American Culture. Duke University Press, Durham, NC.

Berlant, L., 2011. Cruel Optimism. Duke University Press, Durham, NC.

Biehl, J., 2005. Vita : Life in a Zone of Social Abandonment. University of California Press, Berkeley, CA.
Biehl, J., Locke, P., 2010. Deleuze and the anthropology of becoming. Current Anthropology 51 (3), 317–351.
Clifford, J., Marcus, G. (Eds.), 1986. Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley, CA.
Connolly, W., 2002. Neuropolitics: Thinking, Culture, Speed. University of Minnesota Press, Minneapolis, MN.
Darwin, C., 1872. The Expression of the Emotions in Man and Animals. John Murray, London.
Das, V., 2006. Life and Words: Violence and the Descent into the Ordinary. University of California Press, Berkeley, CA.
De Waal, F., 2003. Morality and the Social Instincts: Continuity with the Other Primates, Presented in: The Tanner Lectures on Human Values. Princeton University, Princeton, NJ.
Deleuze, G., 1994. Difference and Repetition (P. Patton, Trans.). Athlone, London.
Deleuze, G., Guattari, F., 1987. A Thousand Plateaus: Capitalism and Schizophrenia. University of Minnesota Press, Minneapolis, MN.
Doel, M., 2010. Representation and difference. In: Anderson, B., Harrison, P. (Eds.), Taking Place: Non-Representational Theories and Geography. Ashgate, Surrey, pp. 117–130.
Feld, S., 1982. Sound and Sentiment : Birds, Weeping, Poetics, and Song in Kaluli Expression. University of Pennsylvania Press, Philadelphia, PA.
Gregg, M., Seigworth, G. (Eds.), 2010. The Affect Theory Reader. Duke University Press, Durham, NCS.
Hardt, M., 1999. Affective labor. Boundary 2 26 (2), 89–100.
Harris, O., Sørensen, T., 2010. Rethinking emotion and material culture. Archaeological Dialogues 17 (2), 145–163.
Ingold, T., 2000. The Perception of the Environment: Essays on Livelihood, Dwelling, and Skill. Routledge, New York.
Kockelman, P., 2011. A Mayan ontology of poultry: selfhood, affect, animals, and ethnography. Language and Society 40, 427–454.
Latour, B., 2005. Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, Oxford.
Limón, J., 1994. Dancing with the Devil: Society and Cultural Poetics in Mexican-American South Texas. University of Wisconsin Press, Madison, WI.
Lutz, C., 1986. Emotion, thought, and estrangement: emotion as a cultural category. Cultural Anthropology 1, 287–309.
Lutz, C., Abu-Lughod, L., 1990. Language and the Politics of Emotion, Studies in Emotion and Social Interaction. Cambridge University Press, Cambridge.
Lutz, C., White, G., 1986. The anthropology of emotions. Annual Review of Anthropology 15, 405–436.
Massumi, B., 2002. Parables for the Virtual: Movement, Affect, Sensation, Post-Contemporary Interventions. Duke University Press, Durham, NC.
Mazzarella, W., 2009. The affective turn: theorizing the social. American Journal of Sociology 114 (6), 1876–1879.
Meskell, L., 1996. The somatisation of archaeology: institutions, discourses, corporeality. Norwegian Archaeological Review 29, 1–16.
Povinelli, E., 2006. The Empire of Love: Toward a Theory of Intimacy, Genealogy, and Carnality. Duke University Press, Durham, NC.
Povinelli, E., 2011. Economies of Abandonment: Social Belonging and Endurance in Late Liberalism. Duke University Press, Durham, NC.
Protevi, J., 2009. Political Affects: Connecting the Social to the Somatic. University of Minnesota Press, Minneapolis, MN.
Rosaldo, M., 1980. Knowledge and Passion: Ilongot Notions of Self and Social Life. Cambridge University Press, Cambridge.
Rosaldo, M., 1984. Towards an anthropology of self and feeling. In: Shweder, R.A., LeVine, R.A. (Eds.), Culture Theory: Essays on Mind, Self, and Emotion. Cambridge University Press, Cambridge, pp. 137–157.
Rosaldo, R., 1993. Culture and Truth: The Remaking of Social Analysis. Beacon Press, Boston, MA.
Stewart, K., 2007. Ordinary Affects. Duke University Press, Durham, NC.
Stewart, K., 2010. Afterword: worlding refrains. In: Gregg, M., Seigworth, G. (Eds.), The Affect Theory Reader. Duke University Press, Durham, pp. 339–354.
Stewart, K., 2011. Atmospheric attunements. Environment and Planning D: Society and Space 29, 445–453.
Tarlow, S., 2000. Emotion in archaeology. Current Anthropology 41 (5), 713–746.
Taussig, M., 2004. My Cocaine Museum. University of Chicago Press, Chicago, IL.
Thrift, N., 2004. Intensities of feeling: towards a spatial politics of affect. Geografiska Annaler 86 (B), 57–78.
Thrift, N., 2011. Lifeworld, Inc. – and what to do about it. Environment and Planning D: Society and Space 29, 5–26.
Tilley, C., 1996. An Ethnography of the Neolithic: Early Prehistoric Societies in Southern Scandinavia. Cambridge University Press, Cambridge.
Wilce, J., 2009. Language and Emotion. Cambridge University Press, Cambridge.
Williams, R., 2001(1961). The Long Revolution. Broadview.

Affect-Regulation Motivation

Michaela Riediger, Max Planck Institute for Human Development, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Affect-regulation motivation refers to how people want to influence their affective experiences. It is typically prohedonic, but can at times also be directed at the enhancement or maintenance of negative, or the dampening of positive affect. This article reviews the current state of research on the following questions: Why are psychologically healthy people occasionally inclined to seek negative, or to dampen positive affective experiences? Does such contra-hedonic motivation require cognitive resources? Are there individual differences in pro- and contra-hedonic affect-regulation motivation? What roles do the individual's age, personality disposition, or cultural background play in this respect?

Affective states – such as feeling angry, downcast, excited, or content – are an integral aspect of human experience. They direct people's attention, guide judgments and behaviors, and provide, when expressed, communicative signals to others that can regulate social interactions. Affective states thus serve essential functions. Being able to influence – or regulate – the experience and expression of affective states, however, is equally important. Terminating negative affective states once their informational, behavioral, or communicative functions have been fulfilled, for example, is necessary for healthy psychological functioning (Larsen, 2000). In addition, many life tasks – from forming and maintaining social relations to successfully pursuing a career – require the ability to influence the intensity and expression of positive and negative affective experiences.

For a long time, the assumption prevailed that affect-regulation efforts are invariably motivated by a desire to enhance the individual's affective well-being. In the past two decades, however, awareness has grown that this is not always the case. Studying affect-regulation motivation can thus contribute to a better understanding of individual differences and within-person fluctuations in affective experiences.

This article reviews the current state of research on affect-regulation motivation, after first clarifying some basic definitional issues.

Some Definitional Issues

Affect, Emotion, and Mood

The term affect is used in this article as an overarching category that subsumes emotions and moods (Gross and Thompson, 2007). Both emotions and moods are experiential entities that are characterized by inner feelings, accompanying physiological processes, and that can be outwardly expressed. While emotions are typically elicited by specific events in the individual's environment or by thoughts about those events, moods tend to have less specific causes and to be related more to the individuals' internal resource capacity (e.g., their level of fatigue). In addition, emotions tend to be more intense than moods, but more short-lived, and to have clearer onset and offset points (e.g., Larsen, 2000). Despite these differences, boundaries between emotions and moods are not clear-cut. This can make empirical differentiation difficult, particularly in studies that assess naturally occurring affective experiences. This explains why many of the studies reviewed below were not able to distinguish between emotions and moods explicitly. In these cases, the umbrella term affective experience will be used.

Pro- and Contra-Hedonic Affect-Regulation Motivation

Affect regulation refers to ways individuals influence the experience and expression of affective states. It often occurs without awareness and intentional effort, but can also be deliberately initiated by the individual. Affect regulation can take many different routes. Avoiding affect-eliciting situations, reappraising their significance, or suppressing outward expressions of inner experiences are but a few examples. Irrespective of their specific manifestation, however, all of these regulatory behaviors are preceded and fundamentally shaped by motivational processes.

Affect-regulation motivation determines the direction of a given affect-regulation effort. It can be directed at the maintenance, enhancement, or dampening of affective experiences and expressions. In everyday life, most people likely do not contemplate their affect-regulation motivation much, if at all. Research has shown, however, that when asked to do so, they can report on it (e.g., Mares et al., 2008; Riediger et al., 2009; Wood et al., 2009). Affect-regulation motivation thus seems to be at least partly accessible to individuals' introspection.

Historically, the assumption long prevailed that motivation of psychologically healthy people in general, and their affect-regulation motivation in particular, are guided by a prohedonic principle. That is, psychologically healthy people were expected to be generally motivated to enhance or maintain positive (pleasant) affective states, and to avoid or dampen negative (unpleasant) affective states (for a discussion of relevant historical positions, see, e.g., Erber and Wang Erber, 2000). Contra-hedonic tendencies were assumed to occur only in mental pathology. Indeed, psychological disorders can involve such symptoms as the deliberate inflicting of injury and pain upon oneself, or a pathological desire to maintain or enhance negative affective states. Recently, however, awareness has grown that contra-hedonic motivation can also occur in nonclinical populations. Although prohedonic affect-regulation motivation is in fact

predominant in psychologically healthy individuals, there can also be occasional exceptions.

Empirical evidence stems, for example, from an experience-sampling study with participants ranging in age from adolescence to old adulthood (Riediger et al., 2009). Participants reported, on average 54 times in 3 weeks, whether they currently wanted to dampen, enhance, or maintain each of six positive and negative affect facets (i.e., feeling angry, downcast, anxious, interested, joyful, and content). The majority of reported affect-regulation motivations were indeed prohedonic, that is, directed at the enhancement or maintenance of positive affect or the dampening of negative affect. Interestingly, however, there were also occasions when participants reported wanting to maintain or enhance negative affect or to dampen positive affect. Across the entire sample, such contra-hedonic motivations were reported, on average, in 15% of the measurement occasions.

Everyday observation also confirms that people are occasionally inclined to forego positive feelings, or to seek or maintain negative feelings. Examples include listening to plaintive music to indulge in sadness, seeking to put oneself into a somber mood before delivering bad news, or trying to work up a state of anger before attempting to assert one's interests in an argument.

The purpose of this article is to review the current state of research on affect-regulation motivation in nonclinical populations. The first part zeros in on the phenomenon of contra-hedonic motivation. It addresses the question why psychologically healthy people might occasionally be inclined to seek negative, or to dampen positive affective experiences, and summarizes research on cognitive resource demands of such contra-hedonic motivation. The second part focuses on individual differences in pro- and contra-hedonic affect-regulation motivation. Evidence of the respective role of individuals' age, traits (personality, self-esteem), and cultural context is reviewed. This article concludes with an outlook on open research questions that need to be resolved in future research.

Possible Reasons for Contra-Hedonic Motivation

Two complimentary explanations for occurrences of contra-hedonic motivation in nonclinical populations are currently discussed. One perspective argues that contra-hedonic motivation may arise when positive experiences accompany apparently negative affective states, that is, when the affective episode is mixed. A complementary account proceeds from the idea that contra-hedonic motivation may occur in situations in which negative affect is useful, or positive affect disadvantageous for the individual. These two positions will be referred to as the mixed-affect and instrumental-affect perspectives below.

The Mixed-Affect Perspective

Affective experiences are not always unequivocally positive or negative in nature. Occasionally, affective episodes may entail a blend of various affective states of opposing valence. For example, people can experience sadness and happiness simultaneously, such as during nostalgic reminiscence; or fear and excitement, such as when engaging in risky activities. Other combinations of affective states of opposing valence can co-occur as well (e.g., disgust and amusement). It is also possible that a given affective state is not accompanied, but followed by a state of opposing valence (e.g., elation followed by embarrassment). The mixed-affect perspective proposes that anticipated or actual mixed affect might motivate individuals to seek or maintain an apparently negative affective state because they associate enjoyable or otherwise positive aspects with it, or, vice versa, to dampen an apparently positive affective state because they associate unpleasant aspects with it as well.

Evidence in line with this perspective stems from the experience-sampling study mentioned above (Riediger et al., 2009). Participants in this study were more likely to report the contra-hedonic motivation of wanting to maintain their momentary negative affect in situations where they experienced mixed affective states, that is, simultaneous positive and negative affect that were both more intense than the individual's respective averages.

Further support comes from a series of studies reported by Andrade and Cohen (2007). These studies showed that students who liked to watch horror movies were more likely to experience both negative and positive affect while watching, whereas persons who avoided horror movies tended to only experience negative affect. Interestingly, the pattern and intensity of negative affect (e.g., fear) while watching did not differ between persons who liked to watch and those who avoided horror movies. This speaks against the possibility that individuals who look for apparently aversive experiences appraise them in a more positive manner. It has also been argued that people may seek negative experiences because of the relief or pleasure that follow in the aftermath. This explanation, however, did not hold in the studies by Andrade and Cohen (2007), where such feelings of relief were greatest in participants who avoided these types of films. Instead, these studies suggest that the apparently contra-hedonic motivation of wanting to view horror movies is associated with positive affective experiences that accompany fear when watching.

The Instrumental-Affect Perspective

The instrumental-affect account of contra-hedonic motivation posits that affective states can be helpful or harmful for an individual's goals, depending on the particular context. This idea is commensurate with evolutionary theories that propose that emotions evolved as multicomponential response systems because they prepared for the fast execution of adaptive behaviors, communicated important information to conspecifics, and thus enhanced chances of survival. A classic illustration of this idea is the instrumental value of fear in threatening situations. The experience of fear, along with the accompanying physiological and cognitive processes – such as enhanced autonomic arousal and attentional narrowing – prepares the individual for rapid engagement in adaptive action, such as flight. The expression of fear furthermore signals danger to others, who can then also get set to seek safety quickly. Fear can thus be helpful in situations that require the avoidance of threats or dangers. In other contexts, however, fear can be detrimental to the individual's interests. When desired objects are available in the absence of immediate danger, for

example, fear might hamper engagement in the approach behavior that would be necessary to obtain the desired item.

Affective states have retained their context-dependent utility until today. Although they may rarely be relevant for immediate survival anymore, they can nevertheless have implications for the attainment of an individual's goals. Central to explanations of contra-hedonic motivations from an instrumental-affect perspective is the observation that there can be situations where negative affect is instrumental for, and positive affect detrimental to an individual's goals. Anger, for example, can help individuals to assert their interests in an interpersonal confrontation; and expressing guilt or shame after wrongdoing can help individuals to be more favorably perceived and better liked by other persons. Joy, on the other hand, can be inappropriate when delivering bad news. Affective states have also been associated with how individuals process information. Although positive affect can facilitate performance in tasks that require creativity, flexibility, or holistically focusing on the overall picture, it is less helpful for chores that require accurate and analytic processing of complex details, or the application of intricate rules. The latter tasks are facilitated more by negative affect of moderate intensity (e.g., Clore et al., 2001). Instrumental-affect accounts therefore propose that contra-hedonic motivation may derive from people's being (consciously or unconsciously) strategic in seeking affective states that are instrumental in a given context.

Research by Erber and colleagues supports this idea. In various studies, either happy or sad mood was induced in participants who were then led to believe that they were about to work on a task either with a stranger, with their own romantic partner, or alone. Prior to the task, participants indicated whether they preferred to read a humorous, sad, or neutral newspaper story. Anticipation of working on a task together with a stranger (vs expecting to work alone) led to an enhanced preference for mood-incongruent newspaper stories, suggesting that participants were motivated to neutralize the induced positive or negative mood. This was not the case, however, when participants expected to work with their romantic partner (Erber and Wang Erber, 2000).

The authors interpreted these findings by arguing that neutralization of positive affective states can be helpful when interacting with strangers, even though it is contra-hedonic, because it unburdens the individuals from possible distractions by their affective state, and thus frees cognitive capacity to attend to the affordances of the unknown social situation. It also reduces the likelihood of having or displaying socially inappropriate affect, which might otherwise lead to unfavorable evaluations by the unknown interaction partner. In other social contexts, such as when interacting with one's romantic partner, in contrast, attenuating positive affect might not be instrumental and may even be detrimental to the relationship (Erber and Wang Erber, 2000).

Various studies by Tamir and colleagues demonstrated that people might seek negative affect when they expect it to be instrumental for the pursuit of their current goals. Participants in two studies reported by Tamir and Ford (2012a), for example, were assigned either a confrontational or a collaboration goal in an interpersonal negotiation task. Participants with the goal to confront (vs collaborate with) their negotiation partner showed a stronger preference for anger-inducing activities in anticipation of the negotiation (e.g., were more likely to prefer recalling past experiences in which they felt angry). Further studies also demonstrated similar contra-hedonic tendencies in other task contexts. For example, enhanced preferences for fear-inducing activities were observed when participants expected to play a computer game in which their goal was to avoid threatening elements (Tamir and Ford, 2009), whereas enhanced preferences for anger-inducing choices were observed when participants expected to play a computer game in which the task was to shoot virtual enemies (Tamir et al., 2008). The authors argued that participants sought negative affect in these studies because of its instrumental value for the pursuit of their assigned goals. For example, in the negotiation studies (Tamir and Ford, 2012a), participants who were set the task of confrontation (vs collaboration) with their negotiation partner were more likely to expect anger to be helpful in their negotiations. Furthermore, selection of anger-inducing activities did indeed lead to more intense feelings of anger during the negotiation, which in turn was associated with better negotiation outcomes in the confrontation condition (similar findings were observed in other task contexts as well, Tamir et al., 2008).

Recent evidence further suggests that knowing when negative affective experiences can be useful for, and when positive experiences can be detrimental to one's goals, is not only associated with better specific goal progress, but also with higher levels of emotional intelligence (Ford and Tamir, 2012) and better psychological adaptation more generally. Student participants rated the extent to which they believed feeling anger or happiness would be useful for them in various confrontational and collaborative situations. The more participants perceived anger as being potentially useful in confrontation situations, the better their psychological well-being, life satisfaction, perceived social support, and subjective health was, and the higher their academic achievement in college was. In contrast, the more participants regarded happiness as potentially helpful in confrontations, the lower they scored on several of these indicators of adaptive functioning. This association pattern was reversed (though overall weaker) for perceptions of useful emotions in collaboration situations (Tamir and Ford, 2012b).

Cognitive Demands of Contra-Hedonic Motivation

Research in the tradition of the instrumental-affect perspective demonstrates that contra-hedonic motivation can be helpful for the individual. This, however, appears to demand cognitive resources. In the experience-sampling study by Riediger et al. (2011), participants not only reported their momentary affect-regulation motivation, but also completed two trials of a working-memory task at each measurement occasion. While prohedonic motivation was only weakly associated with within-person fluctuations in working-memory performance, the association of contra-hedonic motivation and working-memory performance was substantially more pronounced: The more contra-hedonic motivation participants reported, the lower their momentary working-memory performance was. These effects of contra-hedonic motivation

on working-memory performance could not be attributed to lack of effort, or to differences in other individual or situational characteristics.

The authors argued that the reason for these findings could be due to the effort required to overcome one's prevailing prohedonic orientation, which necessitates the allocation of capacity. Consequently, the capacity remaining for the simultaneous processing of another resource-intensive task – storing and manipulating information in working memory – would be reduced in situations with contra-hedonic motivation. Prohedonic motivation, which is the predominant orientation in psychologically healthy individuals, in contrast, might proceed more automatically in comparison and hence require fewer cognitive resources.

Individual Differences in Pro- and Contra-Hedonic Affect-Regulation Motivation

Affect-regulation motivation not only varies across situations, as discussed so far, but also differs substantially between individuals. Below, research findings are summarized that elucidate the roles that people's age, their personality disposition, and their cultural background play in this respect.

Age-Related Differences

Evidence abounds that day-to-day affective experiences differ between individuals from different age groups. Adolescents, for example, typically experience relatively more affective turmoil and relatively more frequent negative affective states than adults. Across adulthood, typical patterns of age-related differences also emerge. Healthy older adults characteristically report more stable and more positive affective states in their everyday lives than younger adult age groups do, and affective well-being appears to decline only at the very end of life (for a review of available evidence, see Riediger and Rauers, 2014).

The psychological mechanisms underlying these age-related differences in affective experiences are not yet fully understood. It has been argued, however, that some of these differences might be brought about by variations in people's affect-regulation motivation (Riediger et al., 2009). According to this argumentation, part of the negative affectivity that is characteristic of adolescence, and part of the positive affectivity that is characteristic of old age, may be actively sought and maintained by the individual, either deliberately or unconsciously.

Empirical evidence supports this assumption. Riediger et al. (2009, 2014), for example, investigated affect-regulation motivation in the everyday lives of participants ranging in age from 12 to 86 years. Contra-hedonic motivation was considerably less prevalent than prohedonic motivation in all investigated age groups. There were, however, pronounced age differences. Adolescents reported contra-hedonic motivation most frequently, namely, in about 25% of the measurement occasions. There was a steep decrease in the prevalence of contra-hedonic motivation between the adolescent and the young adult participants, and a further decline throughout the adult subsamples into old age. Prohedonic motivation, in contrast, showed an opposite prevalence pattern. It was least prevalent among adolescent and young adult participants, and most prevalent in later adulthood.

A corresponding picture of age-related differences emerged in a study where 18- to 82-year-old participants reported how much they sought or avoided several affective experiences in their everyday lives, and how much they regarded these experiences as useful or valuable (Mares et al., 2008). Again, preferences for negative affective experiences were most pronounced among the youngest participants, whereas preferences for positive affect and emotional stability increased throughout adulthood. Further empirical support derives from studies on age differences in entertainment use. For example, younger adults were more likely than older adults to report that they liked movies that elicit sadness or fear, while older adults were more likely to prefer heart-warming movies (Bartsch, 2012; Mares et al., 2008).

What might be the reasons for these age differences in affect-regulation motivation? The relatively high prevalence of prohedonic motivation in older adulthood is in line with the theoretical claim that the shrinking horizon of time-to-live shifts older people's motivations toward wanting to maximize their affective well-being in the here and now (Carstensen et al., 2003). The age-related increase in prohedonic motivation might thus be instrumental for the attainment of older adults' affective goals. Analogously, the relatively higher prevalence of contra-hedonic motivation in adolescence might be helpful for tackling the developmental tasks of that life phase. Repudiating prevailing hedonic conventions, for example, might put adolescents in situations in which they have to deal with negative emotional experiences. This, in turn, might help them to establish emotional autonomy from their parents, affirm a sense of maturity, or refine their self-regulatory competencies. It might also be a means to explore the range and depths of possible experiences, which could help adolescents on their way toward developing their own sense of identity (Riediger et al., 2009).

The mixed-affect perspective can also contribute to an understanding of age-related differences in affect-regulation motivations. The studies by Riediger and colleagues, for example, showed pronounced age-related differences in the prevalence of mixed-affective experiences that followed the same pattern as those of contra-hedonic motivation: Both mixed affective experiences and contra-hedonic motivation were most frequent among adolescents, and least prevalent among older adults (Riediger et al., 2009, 2014). Importantly, this pattern of age-related differences was not restricted to participants' self-report, but was also reflected in implicit representations of affect valence as assessed with implicit association tests. Compared to adults from various age groups, adolescents implicitly associated positive affect least distinctively with pleasantness (vs unpleasantness) and unhappiness least distinctively with unpleasantness (vs pleasantness). Furthermore, the less differentiated people's mental representations of affect valence were, the more likely they were to report mixed affect and contra-hedonic motivation in their everyday lives (Riediger et al., 2014).

Trait-Related Differences

Evidence of trait-related differences in people's affect-regulation motivation has been obtained most systematically with regard to people's personality and self-esteem.

Personality

The two personality traits that are currently best researched in this respect are extraversion and neuroticism. Highly extraverted individuals are strongly oriented toward, and influenced by, their social environment. They actively seek out opportunities to interact with others, are talkative, assertive, and excitable. People scoring high on neuroticism tend to easily perceive situations as threatening or difficult, and to react strongly to them. They are prone to moodiness and irritability.

Numerous studies have demonstrated that extraversion and neuroticism are related to systematic differences in people's affective lives. While extraversion is associated with a higher propensity to experience positive affect, neuroticism is characterized by a susceptibility to experience negative affect (for a review of findings, see Ng and Diener, 2009). It has been argued that part of these affective differences may be driven by variations in people's affect-regulation motivation. In line with this idea, results from various studies suggest that the higher people score on neuroticism, the less motivated they are to ameliorate negative affective experiences. In contrast, the higher individuals score on extraversion, the more motivated they are to reduce or eliminate negative affective experiences, and to seek or maintain positive affect (e.g., Kämpfe and Mitte, 2009; Ng and Diener, 2009).

Tamir demonstrated that these personality differences in affect-regulation motivation are particularly evident in situations of effortful goal pursuit. Across a series of studies, extraversion was predictive of a higher motivation to seek out happiness-inducing experiences in anticipation of effort-demanding tasks (e.g., giving a speech), but not in anticipation of undemanding situations (e.g., listening to music, Tamir, 2009a). Similarly, neuroticism was predictive in various studies of a higher motivation to increase worry in anticipation of effort-demanding (but not in anticipation of undemanding) tasks (Tamir, 2005). The author explained these results by arguing that trait-consistent affective states have instrumental benefits for the individual in effortful situations. Individuals high in extraversion feel rewarded by social acknowledgment and praise, which can be achieved by approach behaviors. Feeling elated and happy is congruent with such an active approach orientation. Individuals high in neuroticism, in contrast, gain rewards from successfully evading threats and punishment. Feeling worried is congruent with such a disposition to avoid negative outcomes. Tamir further proposed that such trait-consistent affect leads to greater task engagement and better task performance. Supporting this idea, individuals scoring high on neuroticism were indeed faster in making valence distinctions when being in a negative as compared to a neutral mood (Tamir and Robinson, 2004), and solved more anagrams after recalling a worrisome versus happy event from their past (Tamir, 2005).

Self-Esteem

Self-esteem refers to an individual's global evaluation of, or liking for, him- or herself. In several studies, people with low self-esteem were less motivated to attenuate negative affect (Heimpel et al., 2002; Wood et al., 2009) and to maintain positive affect (Wood et al., 2003). Enhancing the salience of personal failure or social-rejection experiences further enhanced such contra-hedonic tendencies in participants with low self-esteem (Wood et al., 2009). Evidence within this research tradition further showed that people low in self-esteem felt less deserving of positive affect, and regarded negative affect as more consistent with their self-image than people high in self-esteem. According to self-verification theory, people strive to maintain consistency in their self-views, even if these are unfavorable, because stable self-views enhance the predictability of events and thus provide a sense of control. Striving for self-consistency may thus be among the mechanisms underlying the differences in affect-regulation motivation between individuals with high and low self-esteem (Wood et al., 2009).

Cultural Differences

A frequently used approach to differentiate between cultures is to categorize them as being either individualistic or collectivistic. Members of individualistic cultures tend to define themselves as being unique, to strive for independence and the assertion of their individual interests. Members of collectivistic cultures, in contrast, tend to perceive themselves as being an inseparable part of their social context, to strive for social harmony and the fulfillment of their social duties. Western cultures are typically individualistic, whereas Eastern cultures are often collectivistic.

Affective states likely differ in their instrumental value for the achievement of these culture-specific strivings, which should lead to culture-specific differences in how affective experiences are valued. For example, affect that signals successful approach to personal goals (e.g., pride) should be more important in individualistic cultures, whereas affect that signals failure to meet social obligations (e.g., guilt) should be more important in collectivistic cultures (Eid and Diener, 2001). Empirical evidence indeed suggests cultural differences in the appreciation of affective experiences. Westerners tend to regard positive affect as being more desirable, and negative affect as being more undesirable than Easterners do (Eid and Diener, 2001). Easterners, in contrast, regard finding a dialectical balance between positive and negative affective experiences more valuable than Westerners do (Miyamoto and Ma, 2011).

Similar differences have also been found with regard to the actual affective lives of members from individualistic versus collectivistic cultures. Easterners tend to experience positive affect less frequently than Westerners do (e.g., Mesquita and Karasawa, 2002). Furthermore, positive and negative affect is less negatively correlated (i.e., less mutually exclusive) among Easterners than among Westerners, suggesting a greater likelihood to experience positive and negative affect simultaneously (e.g., Perunovic et al., 2007).

Again, it seems likely that people's affect-regulation motivation is among the mechanisms that mediate these

culture-specific differences in affective experiences. Across a series of studies, Miyamoto and Ma (2011) indeed found that Westerners reported more prohedonic affect-regulation motivation than Easterners did. While Westerners had a clear preference for maximizing positive, and minimizing negative affect, Easterners were more inclined to seek a middle way to balance experiences of positive and negative affect.

Summary and Outlook for Future Research

Affect-regulation motivation refers to how people want to influence their affective experiences. In psychologically healthy individuals, it is typically prohedonic, that is, directed at the enhancement or maintenance of positive affect, or the dampening of negative affect. Occasionally, however, it can also be directed at the enhancement or maintenance of negative affect, or the dampening of positive affect. Such contra-hedonic motivation is accompanied by momentary impairments in the individual's cognitive capacity, which is likely due to the effort that is needed to overcome the prevailing prohedonic orientation.

Two theoretical explanations for occurrences of contra-hedonic motivations have been discussed. The mixed-affect perspective proposes that people occasionally seek negative affective experiences because they are accompanied or followed by positive affect. The instrumental-affect perspective maintains that contra-hedonic motivation may arise when negative affect is helpful for, or positive affect detrimental to the individual's goals. Empirical evidence supports both positions, suggesting that they are complimentary rather than mutually exclusive explanations of contra-hedonic motivation.

Affective experiences vary substantially between individuals. Such differences have been associated, for example, with people's age, their personality, and cultural background. The mechanisms underlying these individual differences in affective experiences are not yet fully understood. Evidence suggests, however, that people's affect-regulation motivations may be one of them. In part, the affective differences between individuals from different age groups or cultural backgrounds, or between individuals with different personality dispositions thus appear to be sought and maintained by the individual actively (although not necessarily consciously). Within the age range from adolescence to old age, for example, contra-hedonic motivation is most prevalent in adolescence, whereas prohedonic motivation is most pronounced among older adults. Furthermore, neuroticism and low self-esteem are associated with less prohedonic and more contra-hedonic motivation, whereas the reverse association pattern has been observed for extraversion. Differences in affect-regulation motivation have also been observed between cultural contexts. Prohedonic motivation, for example, is more pronounced among members of Western, individualistic cultures. Members from Eastern, collectivistic cultures, in contrast, tend to be more inclined to seek a dialectic balance of positive and negative affect.

Among the most important research desiderata in this field of study is the rigorous investigation of the sequence of causality in the observed associations, using appropriate experimental and longitudinal study designs. This will provide a sound empirical basis for distinguishing between antecedents and consequences of different types of affect-regulation motivations. A particularly relevant question in this regard pertains to the link between an individual's affect-regulation motivation and his or her actual engagement in, and success of, affect-regulatory behaviors. The motivation-behavior gap is a well-known phenomenon of human volition, and likely also applies to the specific case of affect-regulation motivation. Having a particular goal does not imply that the individual necessarily engages in its active pursuit, nor does it guarantee that he or she will eventually attain it. An interesting task for future research will therefore be to identify factors that determine whether and how successfully individuals implement their affect-regulation motivation in actual regulation efforts. A related question refers to the longer-term adaptiveness of different types of affect-regulation motivation (cf Tamir and Ford, 2012b). Do certain types of affect-regulation motivation play a role for the individual's longer-term socioemotional adaptation? For example, does contra-hedonic motivation really facilitate the socioemotional adaptation of adolescents, as has been speculated in the literature (Riediger et al., 2009)?

Meticulous investigation of many of these questions would require longitudinal study designs. Such designs would also make it possible to address another important issue regarding the observed age differences in people's affect-regulation motivation: They could help to determine the extent to which these age effects result from differences between birth cohorts rather than from developmental changes within persons growing older.

Affect-regulation motivation has typically been conceived of as a phenomenon that can either be deliberate and intentional or proceed automatically and without conscious awareness. An interesting task for future investigations will be to unravel these different forms of affect-regulation motivation, and to investigate possible consequences that congruence or divergence between deliberate and automatic affect-regulation motivation may have. A related open question refers to the modifiability of affect-regulation motivation. Can individuals, perhaps with the help of an outside intervention, change the way they want to regulate their affective experiences?

Last but not least, types of affect-regulation motivation so far have been primarily distinguished with respect to the valence (i.e., the pleasantness/unpleasantness) of the affective experiences that an individual seeks to regulate. An interesting task for future investigations will be to consider the role of the to-be-regulated affective experience's other characteristics as well; such as its arousal level, or whether it is a relatively short-lived, specific emotion, or a relatively more unspecific and longer-lasting mood.

Taken together, the research reviewed in this article has demonstrated that the systematic empirical inquiry of affect-regulation motivation is a promising tool to further the understanding of affective experiences in general, and of individual differences and within-person fluctuations therein in particular. Many questions, however, still remain to be investigated in the future.

See also: Cross-Cultural Psychology; Culture and the Self: Implications for Psychological Theory; Emotion in Cognition; Emotion, Perception and Expression of; Emotional Regulation; Emotions, Psychological Structure of; Extraversion; Implicit Association Test; Lifespan Development, Theory of; Neuroticism; Personality, Trait Models of; Personality: Historical and Conceptual Perspectives; Self-Esteem; Selfconscious Emotions, Psychology of; Subjective Wellbeing and Culture; Working Memory, Psychology of.

Bibliography

Andrade, E.B., Cohen, J.B., 2007. On the consumption of negative feelings. Journal of Consumer Research 34, 283–300.

Bartsch, A., 2012. As time goes by: what changes and what remains the same in entertainment experience over the life span? Journal of Communication 62, 588–608.

Carstensen, L.L., Fung, H.H., Charles, S.T., 2003. Socioemotional selectivity theory and the regulation of emotion in the second half of life. Motivation and Emotion 27, 103–123.

Clore, G.L., Gasper, K., Garvin, E., 2001. Affect as information. In: Forgas, J.P. (Ed.), Handbook of Affect and Social Cognition. Erlbaum, Mahwah, NJ, pp. 121–144.

Eid, M., Diener, E., 2001. Norms for experiencing emotions in different cultures: inter- and intranational differences. Journal of Personality and Social Psychology 81, 869–885.

Erber, R., Wang Erber, M., 2000. The self-regulation of moods: second thoughts on the importance of happiness in everyday life. Psychological Inquiry 11, 142–148.

Ford, B.Q., Tamir, M., 2012. When getting angry is smart: emotional preferences and emotional intelligence. Emotion 12, 685–689.

Gross, J.J., Thompson, R.A., 2007. Emotion regulation: conceptual foundations. In: Gross, J.J. (Ed.), Handbook of Emotion Regulation. Guilford Press, New York, pp. 3–24.

Heimpel, S.A., Wood, J.V., Marshall, M.A., Brown, J.D., 2002. Do people with low self-esteem really want to feel better? Self-esteem differences in the motivation to repair negative moods. Journal of Personality and Social Psychology 82, 128–147.

Kämpfe, N., Mitte, K., 2009. What you wish is what you get? The meaning of individual variability in desired affect and affective discrepancy. Journal of Research in Personality 43, 409–418.

Larsen, R.J., 2000. Toward a science of mood regulation. Psychological Inquiry 11, 129–141.

Mares, M.-L., Oliver, M.B., Cantor, J., 2008. Age differences in adults' emotional motivations for exposure to films. Media Psychology 11, 488–511.

Mesquita, B., Karasawa, M., 2002. Different emotional lives. Cognition and Emotion 16, 127–141.

Miyamoto, Y., Ma, X., 2011. Dampening or savoring positive emotions: a dialectical cultural script guides emotion regulation. Emotion 11, 1346–1357.

Ng, W., Diener, E., 2009. Personality differences in emotions: does emotion regulation play a role? Journal of Individual Differences 30, 100–106.

Perunovic, W.Q.E., Heller, D., Rafaeli, E., 2007. Within-person changes in the structure of emotion: the role of cultural identification and language. Psychological Science 18, 607–613.

Riediger, M., Rauers, A., 2014. Do everyday affective experiences differ throughout adulthood? A review of ambulatory-assessment evidence. In: Verhaeghen, P., Hertzog, C. (Eds.), The Oxford Handbook of Emotion, Social Cognition, and Everyday Problem Solving during Adulthood. Oxford University Press, New York, pp. 61–82.

Riediger, M., Schmiedek, F., Wagner, G.G., Lindenberger, U., 2009. Seeking pleasure and seeking pain: differences in pro- and contra-hedonic motivation from adolescence to old age. Psychological Science 20, 1529–1535.

Riediger, M., Wrzus, C., Schmiedek, F., Wagner, G.G., Lindenberger, U., 2011. Is seeking bad mood cognitively demanding? Contra-hedonic orientation and working-memory capacity in everyday life. Emotion 11, 656–665.

Riediger, M., Wrzus, C., Wagner, G.G., 2014, August 25. Happiness is pleasant, or is it? Implicit representations of affect valence are associated with contrahedonic motivation and mixed affect in daily life. Advance online publication. Emotion. http://dx.doi.org/10.1037/a0037711.

Tamir, M., 2005. Don't worry, be happy? Neuroticism, trait-consistent affect regulation, and performance. Journal of Personality and Social Psychology 89, 449–461.

Tamir, M., 2009a. Differential preferences for happiness: extraversion and trait-consistent emotion regulation. Journal of Personality 77, 447–470.

Tamir, M., 2009b. What do people want to feel and why? Pleasure and utility in emotion regulation. Current Directions in Psychological Science 18, 101–105.

Tamir, M., Ford, B.Q., 2009. Choosing to be afraid: preferences for fear as a function of goal pursuit. Emotion 9, 488–497.

Tamir, M., Ford, B.Q., 2012a. When feeling bad is expected to be good: emotion regulation and outcome expectancies in social conflicts. Emotion 12, 807–816.

Tamir, M., Ford, B.Q., 2012b. Should people pursue feelings that feel good or feelings that do good? Emotional preferences and well-being. Emotion 12, 1061–1070.

Tamir, M., Mitchell, C., Gross, J.J., 2008. Hedonic and instrumental motives in anger regulation. Psychological Science 19, 324–328.

Tamir, M., Robinson, M.D., 2004. Knowing good from bad: the paradox of neuroticism, negative affect, and evaluative processing. Journal of Personality and Social Psychology 87, 913–925.

Wood, J.V., Heimpel, S.A., Manwell, L.A., Whittington, E.J., 2009. This mood is familiar and I don't deserve to feel better anyway: mechanisms underlying self-esteem differences in the motivation to repair sad moods. Journal of Personality and Social Psychology 96, 363–380.

Wood, J.V., Heimpel, S.A., Michela, J.L., 2003. Savoring versus dampening: self-esteem differences in regulating positive affect. Journal of Personality and Social Psychology 85, 566–580.

Affirmative Action: Comparative Policies and Controversies

Christopher McCrudden, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C.D. Cunningham, volume 1, pp. 210–214, © 2001, Elsevier Ltd.

Abstract

Affirmative action includes measures attempting to increase the participation of particular groups in those contexts in which those groups are regarded as underrepresented. Equivalent policies operate in several other countries in ways that have sometimes borrowed from the United States but mostly taken diverging paths. An overview of the ways in which different jurisdictions approach the issues of affirmative action is provided. A taxonomy of affirmative action measures is developed focusing on five main elements of such policies; six primary justifications for affirmative action are identified; the problems of defining preferred groups are considered; the main areas of political and economic life in which affirmative action tends to be used are described; and the major legal constraints on introducing and continuing affirmative action measures are discussed.

Introduction

The term 'affirmative action' originated in the United States, where it has been both legally and politically controversial at least since the late 1960s. 'Affirmative action' includes measures attempting to increase the participation of particular groups defined in status terms (such as race or gender) in those contexts in which those groups are regarded as underrepresented. Equivalent policies operate in several other countries in ways that have sometimes borrowed from the United States but mostly taken diverging paths (see McCrudden, 2011). Although there has been extensive descriptive and legal analysis of these measures in individual countries, comparisons among a small number of countries (Weisskopf, 2004; Lieberman, 2005; Nesiah, 2000), and several overviews of a range of many countries (Jain et al., 2003; Weiner, 1983; Tomei, 2005; Sabbagh, 2004a; Faundez, 1994), there have been few attempts (Tushnet, 2004; Sabbagh, 2004b; Cunningham, 2002) rigorously to assess such measures comparatively and in a cross-disciplinary perspective.

'Affirmative Action' and Cognate Terms

An initial difficulty in the comparative study of the issues involved with affirmative action thus defined is that different countries have different ways of describing this phenomenon. Some countries use the term 'affirmative action,' borrowing from the US usage, as is the case in Fiji (Cottrell and Ghai, 2007) and Northern Ireland (McCrudden, 1992). 'Positive action' is the term most often adopted to describe these types of measures in Great Britain (Welsh et al., 1994), as well as in the context of European Union law and policy (De Vos, 2007). Other countries and regions tend to use other terms either as well as or instead of the term 'affirmative action,' such as '*parité*' in France (Bird, 2001), 'quotas' in Brazil (Htun, 2004), 'employment equity' in Canada (Abella, 1984) and South Africa (Thomas, 2002), 'reservations' in India (Hooda, 2001), 'preferential policies' or 'special rights' in Malaysia (Haque, 2003), and 'black empowerment' in South Africa (Tangri and Southall, 2008). International human rights law tends to use the term 'special measures' (Bossuyt, 2002). Nevertheless, despite these (and other) important differences, these jurisdictions would generally be viewed as exemplars of the major varieties of affirmative action in the contemporary world.

Those countries that refer to their measures as 'affirmative action' often wish to associate their practices with those of the United States, as was the case in Northern Ireland, South Africa, and Fiji, where the US precedent was regarded as politically significant. In contrast, other countries that have developed their measures since the American debate became so dominant transnationally adopted other terminology, often consciously in order to contrast what that jurisdiction thought it was doing with what it thought affirmative action meant in the United States. One example is Canada, which developed the concept of 'employment equity' partly in order to distance what Canada was doing from the US debate (Abella, 1984). Another is France, where the term 'affirmative active' is often used to distinguish from and contrast with what is regarded as acceptable in the French context (Sabbagh, 2002). In Northern Ireland, some policy makers during the 1980s (largely unsuccessfully) sought to borrow the Canadian terminology of 'employment equity' in order to distance what they proposed from the American concept (Osborne, 1991).

Types of Affirmative Action Measures

The term 'affirmative action' is an 'umbrella' concept, under which many different types of programs shelter. There are five broad categories of measures: positive measures to avoid unlawful discrimination; indirectly inclusionary measures; attracting qualified candidates using advertising and training targeting the underrepresented group; 'tie-break' policies, where individuals from underrepresented group are preferred where equally well qualified; and preferential treatment for members of the targeted group.

Positive Measures to Avoid Unlawful Discrimination

There are three types of positive measures taken to avoid unlawful discrimination that are sometimes classified under

affirmative action: (1) activity preventing and remedying direct discrimination; (2) activity taken proactively to remove indirect, systemic, and institutional discrimination; and (3) reasonable accommodation measures, taken to enable an individual from a particular group to gain access to particular employment or a particular service, for example, changing physical aspects of the work environment to accommodate a person who is physically disabled. The South African Employment Equity Act 1998, section 15, specifies that the types of affirmative action contemplated include 'measures to identify and eliminate employment barriers, including unfair discrimination, which adversely affect people from designated groups,' and 'making reasonable accommodation for people from designated groups.' It is controversial whether reasonable accommodation measures should be regarded as a form of affirmative action, and in the United States, for example, they would be unlikely to be classified as such. These three activities amount to little more than adopting positive measures to avoid what is generally unlawful discrimination, for example, where the employer agrees to remove symbols from the workplace that may well constitute unlawful harassment against a particular group if they remained, or where the system of 'last in, first out' as the basis for making redundancy decisions is replaced by a system with less adverse effects on newly hired minorities, or where an equality body engages in a campaign to urge employers not to discriminate, or where an employer is required to make reasonable efforts to accommodate persons with disabilities, as in the Namibian Affirmative Action (Employment) Act 1998, section 17. More controversially, other measures are sometimes adopted in this context to enable discrimination to be more easily proven, or disadvantage to be more easily identified, such as the collection of statistics by way of monitoring the participation of the targeted group in particular economic activities.

Indirectly Inclusionary Measures

This includes measures such as government assistance targeted at particular inner city areas, for example, resulting in some racial or ethnic groups benefiting disproportionately because they are disproportionately represented in these areas. In the United States, there was much discussion during the 1990s as to the possibility of replacing 'affirmative action' measures that explicitly took race into account, with indirectly inclusionary measures. For example, following the decision of a US Court of Appeals for the Federal Circuit to prohibit race-based affirmative action in student admissions at the University of Texas (*Hopwood v. Texas*, 1996, overturned in *Grutter v. Bolinger*, 2003), the University adopted a system of guaranteeing admission to those students in the top 10% of their class in each High School in the state. Given the highly segregated nature of much of the school system in Texas, this had the effect of producing a significant number of black and Hispanic students eligible for admission to the University. The Fijian Social Justice Act 2001 provides a set of other examples, including the provision of grants targeted at schools with poor academic achievement, and access to training for rural youth. In the Northern Ireland context, a social program devised by government in the early 1990s, termed Targeting Social Need, began as an assistance program targeting geographic areas where need was greatest, with the understanding that this was done in order to reduce the economic disparities between Catholics and Protestants (Hinds and O'Kelly, 2006). These measures sometimes serve as an alternative to other (unacceptable) types of affirmative action.

Advertising and Training Targeting the Underrepresented Group

The third group of affirmative action measures include 'outreach' programs and have traditionally been popular across the political spectrum in several countries (European Commission, 2009). These programs are designed to attract qualified candidates from the previously underrepresented group in two ways: first, by bringing employment opportunities to their attention and encouraging them to apply for permanent employment; and second, by providing training to better equip them for competing on equal terms when they do apply for such employment. These types of measures are sometimes said to operate on the 'supply side' of the equation, meaning that they are intended to affect the behavior of the underrepresented group, rather than on the 'demand side,' referring to measures intended to affect the behavior of employers.

Tie-Break Policies

Tie-break policies operate on the 'demand' side and are more controversial, involving a preference for the person from the underrepresented group, but only where two or more individuals are otherwise equally well qualified. They involve the direct identification of individuals from the targeted group as beneficiaries of the policy, but they have no clear impact on traditional conceptions of 'merit' since traditional conceptions of merit are neither changed nor challenged. Although controversial, they are among the more common types of affirmative action in practice, and have been described by Buchanan (1989) as a 'paradigm of affirmative action.' The US Supreme Court upheld tie-break affirmative action in the context of gender under federal employment discrimination law (*Johnson v. Transportation Agency, Santa Clara County*, 1987). The European Court of Justice held that such tie-break measures operating in favor of women in some of the German *Länder* could be lawful under European Community law (*Marschall v. Land Nordrhein-Westfalen*, 1997). In Britain, the Equality Act 2010 has introduced provisions permitting the use of a tie-break in certain circumstances.

Preferential Treatment for Members of the Targeted Group

Preferential treatment is the classic affirmative action measure and the type of affirmative action measure that is probably most associated in the public mind with American-style affirmative action. These are the affirmative action measures that elicit the most heated debate because they involve the possibility that a less 'well-qualified' person from the underrepresented group would receive the benefit at the expense of the better 'qualified' person. There are, however, important differences between the different types of programs that come under

this broad heading, relating to the type of preference accorded, and the situation in which the preference is accorded. For example, a preference for a particular candidate from the underrepresented group where the preferred candidate is less well qualified, can be distinguished from a general quota in favor of candidates from the underrepresented group. The Fijian Social Justice Act 2001 provides numerous examples, including preferred equity funding for Fijians and Rotumans, educational scholarships designed exclusively for Fijians and Rotumans, provision of special funding to Fijians and Rotumans to buy houses, government assistance only to Fijians and Rotumans to acquire and develop farm land, and quotas for Fijian- and Rotuman-owned businesses to be awarded government contracts.

These methods are usually seen as limiting the operation of 'merit,' but an important variation on this approach involves redefining merit, for example, changing the traditional criteria for admission to university or employment in such a way that the new criteria are more likely to favor candidates from the underrepresented group. Both these types of affirmative action are seen as posing the greatest challenges for traditional conceptions of 'merit' in the sense that in the first traditional 'merit' is overridden in favor of increasing participation, and in the second 'merit' is redefined to include characteristics that the targeted group are more likely to have. For example, a policy may be introduced (as it was in Northern Ireland) requiring a proportion of candidates for the police force to be from a particular ethnic group in order to ensure that policing is acceptable to that group (McCrudden et al., 2004). Another example is provided in the provisions of the Hesse law upheld by the European Court of Justice in *Badeck v. Hessischer Ministerpräsident*, 2000 in which family work and care responsibilities at home were included in a list of suitable qualifications for employment in the public sector, favoring women.

Aims and Justifications for Affirmative Action

The circumstances leading to the introduction of affirmative action in a particular country often have distinct implications for the subsequent development of affirmative action in that jurisdiction. In the European Union, affirmative action began in the context of gender equality, and is voluntary, and has now expanded to include provisions permitting affirmative action in the context of other groups (De Vos, 2007). In India, affirmative action began in the context of caste and is compulsory in the public service, and has now expanded to encompass gender (Nesiah, 2000). Affirmative action that arises from tensions that have contributed to major public order problems, such as Malaysia, will likely take different forms (Haque, 2003). There are also important differences in terms of the size of the target group to benefit, compared to other groups in the country, with consequential effects on the political salience of the issue.

Several different types of justifications are put forward in different jurisdictions, depending in part on the historical context in which affirmative action first developed. At least six separate justifications may be identified that are analytically distinct, although any one jurisdiction may draw on two or more in combination.

Helping to Prevent Current Direct Discrimination and Addressing Stereotyping

This justification is particularly associated with the approach adopted in some of the cases decided by the European Court of Justice regarding positive action for women. The Court adopted an argument that positive action may be justified as a way of addressing stereotyping of women and the resultant segregation of women in particular occupations (see, for example, *Badeck v. Hessischer Ministerpräsident*, 2000).

Compensation for Past Discrimination

Originally, this was probably the most widely used justification for affirmative action in the United States, where affirmative action was often seen as compensation for decades of segregation following centuries of slavery. In the United States, this rationale had two variants: where the past discrimination was carried out by the institution or employer who now is required or wishes to engage in compensatory affirmative action; and where the past discrimination is societal and the particular employer was not directly responsible for this discrimination in the past. The former remains an acceptable basis for limited types of preferential treatment; the latter has long since been held by the US Supreme Court to be an unacceptable basis for preferential treatment (*Regents of the University of California v. Bakke*, 1978).

Egalitarian/Redistributive Argument

Edwards (1995) has identified affirmative action as "one means of effecting some redistribution of resources to minority groups in order to improve their general standing in society and their quality of life and, in the process, of enhancing distributive justice." This is the rationale that is frequently articulated in the decisions of the Indian Supreme Court addressing affirmative action in the context of caste inequalities, as well as remedying past discrimination, to justify affirmative action for *dalits* (*Indra Sawhney v. Union of India*, 1993). Whether or not this rationale is explicitly identified as a basis for affirmative action will often depend on how far this type of affirmative action is seen as contributing to or going against the grain of other aspects of the country's social policies.

Identity and Recognition

Under this justification, affirmative action is advocated, according to Cottrell and Ghai (2007: p. 229), "to maintain the cultural values and institutions of community." They continue: "Affirmative action takes the form of special representation in state institutions, protection from cultural practices even if they may be inconsistent with general national norms, control over land, and forms of self-government." In this context, affirmative action 'positively promotes the value of group identity.' Thus, this type of justification is most commonly used when affirmative action is used in the context of aboriginal peoples.

Diversity Argument

Diversity-based rationales seek to justify affirmative action on the grounds that achieving a greater mix in the composition of

schools, universities, police forces, and employment is good for society and for the organization engaging in affirmative action. Diversity in the classroom is said to improve the educational experience because students' understandings of the world are challenged by exposure to a broader range of perspectives. More diverse organizations are said to make better decisions, because managers are likely to consider a wider variety of possibilities. Organizations that serve a diverse customer base are said to better serve their customers by having a more diverse workforce. The diversity argument is sometimes made as justification for affirmative action in US higher education and has been accepted by the US Supreme Court as constitutional in that context (*Gratz v. Bollinger*, 2003; *Grutter v. Bollinger*, 2003). It has also been applied in other jurisdictions, such as France (Bird, 2001; Sabbagh, 2002). Diversity has also been embraced by human resource managers in several jurisdictions, such as Britain, as the appropriate basis on which to manage personnel (Albrecht, 2001).

Social Cohesion Argument

In some societies where group differences have contributed to such significant divisions that physical conflict has resulted, affirmative action has sometimes been put forward as part of a package of measures aimed at promoting social cohesion and ensuring greater social harmony (Faundez, 1994).

Combinations

In practice, most jurisdictions tend to combine reasons for introducing affirmative action measures, and justifications may well change over time. Occasionally, multiple and changing justifications become problematic within a jurisdiction.

For example, in Northern Ireland, affirmative action began in the context of the public order problems associated with political and economic differences between Catholic Nationalists and Protestant Unionists. Allegations of employment discrimination against Catholics was rife, indeed dating back to the foundation of the state in 1920. Not surprisingly, then, rationales based on past and current discrimination have been the most commonly articulated as justifications for affirmative action directed at advancing the economic position of Catholics. However, such social justice rationales cannot underly the affirmative action provisions in fair employment legislation and practice, which are symmetrical and require Protestant underrepresentation in firms to be addressed as well.

Articulating the justification behind affirmative action for Protestants in Northern Ireland requires another explanation, which may be either pragmatic or principled. The pragmatic reason appears to be that the politics of Northern Ireland is such that if affirmative action were to have been presented simply as a zero-sum game, in which Protestants lost out 'their share' of employment, there might well have been such resistance that it would not have been implemented at all, and thus in order to safeguard against this possibility, Protestants were included as well. A more principled explanation is based on the recognition of identities, that it was seen as important for Protestants to feel that they continued to be an important part of Northern Ireland society, and this meant that they too should be represented across political and economic institutions. Northern Ireland, like Fiji, may therefore be an interesting example of the interaction of these different rationales. If that is the case, then the Northern Ireland example may be a useful test case of a long running controversy between those who see recognition rationales and social justice rationales as mutually antagonistic (Fraser, 2000), and those who see them as mutually supportive (Young, 1997).

How Is the 'Preferred' Group Defined?

Which Groups

The issue of which group (or groups) is defined as the beneficiary of the affirmative action program is another crucial variable. Several jurisdictions have constructed affirmative action programs to benefit women. Where affirmative action is adopted for one group, it is also common for it to be used to benefit other status groups that are disadvantaged in some particular respect, usually economically. Where the 'preferred groups' are women and other groups such as disabled workers, the justifications for affirmative action may be different from those justifying affirmative action in favor of minority racial groups. Differences are also likely to arise between jurisdictions where the beneficiary is part of a 'minority' (as in the United States and Canada; in Canada, the term used is 'visible minorities'), and jurisdictions where the affirmative action is applied to benefit (part of) the majority community (as in Malaysia and South Africa). The proportion of the population that may stand to benefit from preferences, therefore, differs markedly from country to country. Affirmative action may therefore be used to benefit the majority or the minority, and the political and economic cost analysis may vary accordingly.

Aboriginal Groups

Although women, disabled persons, and ethnic minorities have tended to be the groups most likely to have affirmative action measures applied to them, a complicated relationship has emerged between affirmative action measures and aboriginal groups. Here is also where the terminology of 'affirmative action' takes on a strongly political connotation. To term measures that particularly protect aboriginal peoples as 'affirmative action' seems to imply that these groups are in some sense equivalent to women, disabled persons, or ethnic minorities, but many aboriginal peoples resist that equivalence, regarding their moral (and in some cases legal) claims as based on their position as having been in occupation of territory long before the now-dominant group arrived. From this point of view, the rationale for measures protecting these groups differs significantly from the rationales for affirmative action discussed earlier. In the aboriginal context, the rationale for protection is based on the moral or legal claims of original occupation. Special protection measures for aboriginal peoples differ not only in their rationale, but also in the type of measures adopted. Thus, as in Canada and the United States, there is a greater likelihood that such measures involve the transfer of land to the aboriginal group over which the group exercises some degree of 'sovereignty.' Despite these differences, the term 'affirmative action' has been used in some jurisdictions to describe these measures, notably in Canada. In

other countries, too, the measures adopted bear a striking resemblance to those that are termed 'affirmative action.' So, for example, in Taiwan there are provisions establishing preferences for aboriginals in education, procurement, and representation in the legislature.

Symmetrical and Asymmetrical Affirmative Action

Thus far, we have assumed that affirmative action programs are asymmetrical in their targeting of the beneficiary group, and this is indeed the case in many of the jurisdictions that engage in affirmative action. Women (but not men) benefit in the United States, as do black (but not white) Americans, as do native (but not non-native) Malays in Malaysia. In the case of affirmative action used to address disadvantages associated with physical or mental disability, it is usually the case that disabled (not able-bodied) persons benefit from such action. That is not the case in several jurisdictions, however, where the way in which affirmative action is defined means that, at least theoretically, affirmative action could be applied symmetrically. Thus, in Britain, the Equality Act 2010 permits affirmative action to be undertaken not only to benefit women but also men, and black British but also white British. In practice, however, although defined in such a way as to be theoretically symmetrical, in practice these measures are usually *applied* asymmetrically. A unique feature of the Northern Ireland affirmative action program, however, is that it has not only been defined symmetrically to address inequality between communities defined in terms of religion, it has also been applied to target both the majority and the minority communities where there is significant underrepresentation of either group employed in particular firms (McCrudden et al., 2010).

Determining Membership of the Targeted Group

To the extent, as is likely, that the beneficiaries are defined in 'group' terms, the further question is how membership in the group is to be determined. Who is to be regarded, for example, as a member of a 'racial' or 'ethnic' group for affirmative action purposes? This has proven particularly problematic in some contexts. In Brazil, the problem has been that whilst affirmative action has been applied to benefit those who are 'black,' the use of color to designate the beneficiaries has proven difficult in a society where there is every shade along a spectrum from 'white' to 'black' (Racusen, 2004). In India, the issue of who should be regarded as a member of a particular caste has arisen (Jenkins, 2009; Galanter, 1984). In South Africa, the High Court has held that a person who is of South African Chinese origin should be defined as 'black' for the purposes of benefiting from 'black empowerment' measures (*Chinese Association of South Africa v. Minister of Labour*, 2008). In other words, the problem is how to go about identifying individual members of the group once that group is defined as the beneficiary.

Crudely, two methods have been adopted. One method involves 'self-identification' in which an individual claims to be a member of the benefitted group and that claim is accepted by those operating the program. Alternatively, a system may be adopted in which membership of the group is determined 'objectively' either by some authority associated with the group (with the possibility that this reinforces the existing power structures within the community) or by a state authority such as the judiciary (with the possibility that this leads to 'essentialization,' where designated characteristics come to be considered as fixed traits, and variation among group members is discounted or ignored).

The 'Creamy Layer' Issue

In some countries, there is an attempt to ensure that affirmative action is targeted only to that part of the status group that suffers the greatest economic deprivation. In India, for example, one of the benefitted groups (the so-called backward classes) is defined so that those who are economically at the top of a backward class (the 'creamy layer,' as it is called) are excluded from preferences (*Indra Sawhney*, 1993). In these contexts, the relationship envisaged between the group and the individual members of the group becomes quite complicated, and to some extent the application of affirmative action needs to be justified at the level of the individuals. In other jurisdictions, this does not seem to be important (Malaysia, United States).

Who Loses Out?

Another way of viewing the question of beneficiaries is to contrast them with those who may lose out as a result of affirmative action policies, but here too the issue is likely to be complicated. There is likely to be significant debate if there are in fact any losers, who they are, and what the nature of any loss is, particularly when set against the benefits. Those who see themselves as 'losers' may differ from those who are perceived by others as losers, and may differ also from those who may actually lose out in practice. The issue of who loses and how much may well also affect the extent and duration of affirmative action programs. Thus, for example, in the United States, there has been greater unease in affirmative action applying to areas such as layoffs and redundancy, leading to modifications of last-in first-out, for example, in part because those who would lose out are likely to be more identifiable and stand to lose what they regarded as their legitimate expectations as to how they would be treated. The importance of identifying the potential losers is also to identify where the potential resistance to affirmative action may come from, and the nature of that resistance: the application of affirmative action in different sectors is likely to lead to different sets of losers, with differing capacities and opportunities to resist.

In Which Sector Is Affirmative Action Adopted?

Employment

There are different aspects of the employment relationship in which affirmative action may be used, with some programs involving affirmative action measures in the pre-employment stage, with efforts being placed on encouraging individuals from underrepresented groups to apply to firms that they might not previously have applied to, or training individuals in groups that are thought to lack certain skills to enable them to apply successfully for employment.

Affirmative action measures may also extend to hiring, while others extend to promotion and layoffs as well.

There is also an important issue as to whether the employment covered is public only or whether it also extends into the private sector. In several countries, affirmative action in employment extends only to the public sector, since that sector is seen as appropriately reflecting the make-up of the population as a whole. In countries such as India, where employment in the public sector has been a significant proportion of the employment market as a whole, the move to privatization has occasioned considerable debate as to whether affirmative action should begin to apply in private sector employment.

Political Representation

One of the fastest growing affirmative action measures is in the area of political representation, particularly as regards the role of women. In most countries, women are underrepresented in the political life of their country, and there is an increasing move to address this by introducing affirmative action measures of various types. One of the best known is the French *parité* system, in which there is a requirement that the lists of candidates presented to the electorate by political parties must contain a proportion of women (Stetson, 2001; Mazur, 2001), but similar systems have existed in many other countries for some time (Caul, 2001; Matland, 1998). Sometimes, as in the United Kingdom, political parties have been permitted to set up a system to ensure that the shortlist for selection as the Party's candidate for a parliamentary election includes at least one woman (Sim, 2000). In other countries, such as India with regard to the Scheduled Tribes and the Scheduled Castes, representation in the legislature is guaranteed to particular groups.

Education

Affirmative action may be used both at the level of primary and secondary education, and at the level of tertiary education, as in the United States (Bowen and Bok, 1998), Brazil (Htun, 2004), India (Nesiah, 2000), and Fiji (Puamau, 2001). The measures adopted may range from setting aside a set proportion of places for the preferred group in each year's incoming class, or modifying the entry requirements that have to be met to reflect the life experiences of the underrepresented group, or adopting entry requirements that are formulated neutrally (such as being in the top 10% of one's high school class in Texas) but have the effect of including significant numbers of the underrepresented group. A somewhat different method, which is usually not called 'affirmative action,' is to permit the establishment of schools catering specifically for a particular group defined in religio-ethnic, or gender, terms.

Public Procurement Contracts

Several countries, for example, Canada, the United States, Malaysia, and South Africa (McCrudden, 2007), adopt affirmative action preferences for those bidding for public contracts, in the form of preferences for businesses owned or controlled by members of the preferred group. Given that in many countries the expenditure by government on contracting amounts to a significant slice of GDP, this means that the impact of such preferences on the economy may be significant. The use of public procurement in this way needs to be distinguished from another common use of public procurement, which is to use access to public procurement as a method of ensuring the adoption of affirmative action in, for example, employment practices.

Control of Corporations

Several countries attempt to secure more equal distribution of economic wealth by requiring corporations to cede some degree of control to the disadvantaged group (McCrudden, 2007). South Africa, for example, has structured its black empowerment policies to ensure the transfer of corporate ownership to the black population; Malaysia has attempted to ensure that native Malays are represented on the boards of newly privatized companies. Other jurisdictions have sought to use affirmative action to increase the proportion of women on company boards. In Norway, certain public limited liability companies were required to meet gender representation requirements (Tobler, 2007). Israeli law has placed emphasis on both genders being represented on boards of state-owned companies (Izraeli, 2003).

Land and Minerals

Several countries in which the ownership of agricultural land is central to economic power and influence have introduced or contemplated introducing affirmative action measures in an effort to redistribute land from one ethnic group (often the group associated with the former colonial power) to another (often one or more of the groups associated with the move to decolonialization). This has proven particularly controversial in Zimbabwe, where extensive measures have been adopted. In Fiji, the government provides special funding to Fijians and Rotumans to acquire and develop farmland. South Africa has also envisaged the use of affirmative action measures as a tool in land reform but has not, so far at least, activated these measures. South Africa has, however, been at the forefront of using affirmative action as a method of bringing about a radical transformation of mineral development, one that sought to benefit 'historically disadvantaged persons.'

Interactions between Measures

In understanding the potential effect of affirmative action measures taken in these areas, it is important not to underestimate the extent to which there are likely to be interactions between measures taken in different areas. To take an obvious example, affirmative action taken in the area of education would be likely to affect the skills that candidates for employment are likely to be able to present and therefore the likelihood that they will be employed, which may reduce the need for affirmative action at the stage of hiring. However, affirmative action at the hiring stage may create incentives for those from previously excluded groups to undertake the education that would be useful for those jobs. Clearly, too, the greater the prevalence of affirmative action throughout the economy, the more the society will come to be dominated by group identification; the overall effect may be greater, in other words, than the sum of its parts.

Legal Constraints on Affirmative Action Measures

Legal constraints on affirmative action may arise in a constitutional document, a piece of primary or secondary legislation, a judgment from a court, or directions from a government department or other governmental body. Depending on the jurisdiction, one or more types of affirmative action may be legally or constitutionally prohibited, or legally or constitutionally permitted but not required, or legally or constitutionally required, and the difficult issue that arises is often what exactly the law does prohibit or require.

The role of law at the beginning of the process of developing affirmative action needs to be distinguished from the role of law as affirmative action develops. In the latter context, the role of law is a continuing process, and one which is often less visible to nonlegal actors. The domestic legal position is likely to be complicated by the fact that most jurisdictions will have a set of domestic legal requirements that may point in different directions on these issues, with possible tensions or conflicts between different levels of legal authority. Conflicts may arise between constitutional and legislative provisions, or between central and local provisions. One issue that has tended to recur is whether affirmative action is considered to be a permanent feature of domestic policy, or whether it is considered to be a 'temporary' remedial feature. Justice O'Connor in her US Supreme Court judgment in *Grutter v. Bollnger* (2003) viewed it as necessary that race-conscious admissions policies should be 'limited in time.' If 'temporary,' the length of time for which it is permitted needs to be specified (in *Grutter*, it appeared to be in the region of 25 years from the Supreme Court decision), as does the institution responsible for assessing that it should end.

In many jurisdictions, the relationship of affirmative action and the nondiscrimination principle has often proven to be among the most controversial legal issues, and one in which litigation is likely to take place. Here, the crucial issue is whether the affirmative action measures are seen as exceptions to the nondiscrimination principle, or regarded as a way of fulfilling the principle of equality in practice. There is a clear difference between an affirmative action program that is regarded as an exception to the general principle of nondiscrimination (as in the United States, for example), and an affirmative action program that is seen as a fulfilment of the equality principle (as in India, for example). Where affirmative action, in the form of preferences, is seen as an exception to the nondiscrimination principle, a specific exception must be crafted, if preferences are not to be subject to the nondiscrimination requirement and thus unlawful. This approach leads to a much less legally secure basis for affirmative action, although the critical question will often be the intensity of the scrutiny to which the affirmative action program is subject.

Further complications arise where, as is usually the case, a particular jurisdiction not only has domestic legal provisions that are relevant, but also has a set of international legal requirements that must be complied with. But international law does not speak with one voice on affirmative action measures, and there may be an apparent clash between international human rights law, for example, and international economic law. In some jurisdictions, therefore, there is a challenge to affirmative action authorized domestically using international human rights obligations. In other jurisdictions, domestically authorized affirmative action may be challenged using international economic law obligations arising from plurilateral trade treaties, or bilateral investment treaties (as in South Africa, see Schneiderman, 2009). As a result, one area in which there is significant difference from country to country is how much freedom there is to challenge affirmative action provisions, and where there is freedom to do so how much litigation there is in practice seeking to test the limits of permitted affirmative action and the extent of activism by the courts seeking to control the details of affirmative action programs.

See also: Affirmative Action: Empirical Work on Its Effectiveness; Civil Rights; Discrimination and the Law; Equality and Inequality: Legal Aspects; Gender and the Law; Law and People with Disabilities.

Bibliography

Abella, Rosalie S., 1984. Report of the Commission on Equality in Employment. Government of Canada, Ottawa.
Albrecht, Maryann H. (Ed.), 2001. International HRM: Managing Diversity in the Workplace. Blackwell, Malden.
Badeck v. Hessischer Ministerpräsident Case C-158/97, [2000] ECR-I 1875.
Bird, Karen, 2001. Liberté, égalité, fraternité, parité… and diversité? The difficult question of ethnic difference in the French parity debate. Contemporary French Civilization 25 (2), 271–292.
Bossuyt, Marc, 17 June 2002. The Concept and Practice of Affirmative Action: Final Report by the Special Rapporteur, 1998/5, E/CN.4/Sub.2/2002/21.
Bowen, William G., Bok, Derek, 1998. The Shape of the River: Long-term Consequences of Considering Race in College and University Admissions. Princeton.
Buchanan, G Sidney, 1989. Johnson v. Transportation Agency, Santa Clara County: a paradigm of affirmative action. Houston Law Review 26 (2), 229.
Caul, Miki, 2001. Political parties and the adoption of candidate gender quotas: a cross-national analysis. Journal of Politics 63 (4), 1214–1229.
Chinese Association of South Africa v. Minister of Labour, High Court of South Africa (Transvaal Provincial Division), Case No. 59251/2007, 18 June 2008.
Cottrell, Jill, Ghai, Yash, March 2007. Constitutionalising affirmative action in the Fiji Islands. International Journal of Human Rights 11, 227.
Cunningham, Clark D., 2002. Affirmative action: comparative policies and controversies. In: Smelser, Neil J., Baltes, Paul B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Elsevier.
De Vos, Marc, 2007. Thematic Study on Positive Action. European Commission.
Edwards, John, 1995. When Race Counts: The Morality of Racial Preference in Britain and America. Routledge, London, New York.
European Commission, 2009. International Perspectives on Positive Action Measures: A Comparative Analysis in the European Union, Canada, the United States and South Africa. European Communities.
Faundez, Julio, 1994. Affirmative Action: International Perspectives. ILO.
Fraser, Nancy, 2000. Rethinking recognition: overcoming displacement and reification in cultural politics. New Left Review 3, 107.
Galanter, Marc, 1984. Competing Equalities: Law and the Backward Classes in India. The University of California Press, Berkeley.
Gomez, Edmund T., Jomo, Kwame S., 1997. Malaysia's Political Economy: Politics, Patronage, and Profit. Cambridge University Press, Cambridge.
Gratz v. Bollinger, 123 S.Ct. 2411 (2003).
Grutter v. Bollinger, 123 S. Ct. 2325 (2003).
Haque, M Shamsul, 2003. The role of the state in managing ethnic tensions in Malaysia: a critical discourse. American Behavioral Scientist 47 (3), 240–266.
Hinds, Bronagh, O'Kelly, Ciarán, 2006. Affirmative action in Northern Ireland. In: Kennedy-Dubourdieu, Elaine (Ed.), Race and Inequality: World Perspectives on Affirmative Action. Ashgate, pp. 103–126.
Holzer, Harry J., Neumark, David, 2004. Introduction, The Economics of Affirmative Action. In: Holzer, Harry J., Neumark, David (Eds.), The International Library of Critical Writings in Economics. Edward Elgar Publishing Limited, UK.

Hooda, Sagar Preet, 2001. Contesting Reservations: The Indian Experiment on Affirmative Action. Rawat Publications, Jaipur and New Delhi.
Hopwood v. Texas, 78 F.3d 932 (5th Cir. 1996).
Htun, Mala, 2004. From 'racial democracy' to affirmative action: changing state policy on race in Brazil. Latin American Research Review 39 (1), 60.
Indra Sawhney v. Union of India, AIR 1993 SC 477.
Izraeli, Dafna, 2003. Gender politics in Israel: the case of affirmative action for women directors. Women's Studies International Forum 26 (2), 109–128.
Jain, Harish, Sloane, Peter, Horwitz, Frank (Eds.), 2003. Employment Equity and Affirmative Action: An International Comparison. M.E. Sharpe, Armonk.
Jenkins, Laura Dudley, 2009. Identity and Identification in India: Defining the Disadvantaged. Routledge, London.
Johnson v. Transportation Agency, Santa Clara County, 107 S Ct 1422 (1987).
Lieberman, Robert, 2005. Shaping Race Policy: The United States in Comparative Perspective. Princeton University Press, Princeton.
Marschall v. Land Nordrhein-Westfalen, Case C-409/95, [1997] All ER (EC) 865.
Matland, Richard, 1998. Women's representation in national legislatures: developed and developing countries. Legislative Studies Quarterly 23 (1), 109–125.
Mazur, Amy, 2001. Drawing lessons from the French parity movement. Contemporary French Civilization 25 (2), 201–220.
McCrudden, Christopher, 1992. Affirmative action and fair participation: interpreting the Fair Employment Act 1989. Industrial Law Journal 21, 170.
McCrudden, Christopher, 2007. Buying Social Justice. OUP.
McCrudden, Christopher, 2011. A comparative taxonomy of 'positive action' and 'affirmative action' policies. In: Schulze, Reiner (Ed.), Non-discrimination in European Private Law. Mohr Siebeck, Tübingen, pp. 157–180.
McCrudden, Christopher, Ford, Robert, Heath, Anthony, 2004. Legal regulation of affirmative action in Northern Ireland: an empirical assessment. Oxford Journal of Legal Studies 24, 363.
McCrudden, Christopher, Muttarak, Raya, Hamill, Heather, Heath, Anthony, 2010. Affirmative action without quotas in Northern Ireland. Equal Rights Review 4, 7.
Melber, Henning, 2006. Breeding Fat Cats: Affirmative Action, Black Economic Empowerment, and Namibia's Post-colonial Elite Formation. Danish Institute for International Studies. Working Paper 2006/29.
Nesiah, Devanesan, 2000. Discrimination with Reason?: The Policy of Reservations in the United States, India and Malaysia. OUP.
Osborne, Robert D., 1991. Fair employment and employment equity: policy learning in a comparative context. Public Money and Management 12 (4), 11–18.
Puamau, Priscilla Qolisaya, 2001. A post-colonial reading of affirmative action in education in Fiji. Race, Ethnicity and Education 4 (2), 109.
Racusen, Seth, 2004. Making the 'impossible' determination: flexible identity and targeted opportunity in contemporary Brazil. Connecticut Law Review 36, 787–828.
Regents of the University of California v. Bakke, 438 U.S. 265 (1978).
Sabbagh, Daniel, 2002. Affirmative action at Science Po. French Politics, Culture and Society 20 (3), 52–64.
Sabbagh, Daniel, 2004a. Affirmative Action Policies: An International Perspective. United Nations Development Programme, Human Development Report Office. Occasional Paper 2004/12.
Sabbagh, Daniel, 2004b. Elements toward a Comparative Analysis of Affirmative Action Policies. Paper Delivered to 2004 Annual Meeting of the American Political Science Association.
Schneiderman, David, 2009. Promoting equality, black economic empowerment, and the future of investment rules. South African Journal on Human Rights 25 (2), 246–279.
Sebok, Anthony J., 2003. Reparations, unjust enrichment, and the importance of knowing the difference between the two. The New York University Annual Survey of American Law 58, 651.
Sim, Birte, 2000. Gender and Citizenship: Politics and Agency in France, Britain, and Denmark. Cambridge University Press, Cambridge.
Sowell, Thomas, 2004. Affirmative Action around the World: An Empirical Study. Yale.
Stetson, Dorothy McBride, 2001. Democracy and the representation of women: France in comparative context. Contemporary French Civilization 25 (2), 175–192.
Tangri, Roger, Southall, Roger, 2008. The politics of black empowerment in South Africa. Journal of Southern African Studies 34, 699–716.
Thomas, Adèle, 2002. Employment equity in South Africa: lessons from the global school. International Journal of Manpower 23, 237–255.
Tobler, Christa, 2007. Going global in sex equality law: the case of gender representation rules for company boards. In: Monti, Mario, et al. (Eds.), Economic Law and Justice in Times of Globalisation. Nomos, p. 891.
Tomei, Manuela, May 2005. Affirmative Action for Racial Equality: Features, Impact and Challenges. ILO.
Tushnet, Mark, 2004. Interpreting constitutions comparatively: some cautionary notes, with reference to affirmative action. Connecticut Law Review 36, 649.
Weiner, Myron, 1983. The political consequences of preferential policies: a comparative perspective. Comparative Politics 16 (1), 35–52.
Weisskopf, Thomas E., 2004. Affirmative Action in the United States and India. Routledge, London, New York.
Welsh, Colin, Knox, James, Brett, Mark, 1994. Acting Positively: Positive Action under the Race Relations Act 1976. Employment Department, Sheffield. Research Series No. 36.
Young, Iris, 1997. Unruly categories: a critique of Nancy Fraser's dual systems theory. New Left Review 222, 147.

Relevant Websites

US Department of Labor

http://www.dol.gov/dol/topic//hiring/affirmativeact.htmSouthAfrica.info – Portal containing current information on affirmative action obligations of employers and government contractors under federal law, periodically updated.

SouthAfrica.info

http://www.southafrica.info/business/trends/empowerment/ – Provides up-to-date information on the operation of Black Economic Empowerment policies in South Africa, periodically updated.

Quota Project Database

http://www.quotaproject.org/ – The Database provides information on the various types of gender quotas in the area of political representation in existence today, detailing the percentages and targets in countries where they are applicable, and is periodically updated.

Treasury Board of Canada Secretariat

http://www.tbs-sct.gc.ca/reports-rapports/ee/index-eng.asp – This provides access to the annual monitoring reports regarding the implementation of employment equity in the Canadian public service employment, and is periodically updated.

Indian Judgment Information System (JUDIS)

http://judis.nic.in/supremecourt/chejudis.asp – This consists of the Judgments of the Supreme Court of India and several High Courts. In the case of the Supreme Court, all the main reported Judgments concerning the policy of caste reservations are available on this website, and is periodically updated.

Academy of European Law: Gender Documentation: Positive action

http://www.era-comm.eu/oldoku/SNLLaw/Positiveaction.htm – Contains notes by speakers at many recent seminars devoted to positive action for women in European Union law, and is periodically updated.

Equality Commission for Northern Ireland

http://www.equalityni.org/sections/default.asp?cms=employers_registration,%20monitoring%20and%20review&cmsid=129_135&id=135&secid=3 – Provides information on affirmative action in employment requirements applying to employers in Northern Ireland, and is updated periodically.

Affirmative Action: Empirical Work on Its Effectiveness

Jonathan S Leonard, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Affirmative action in the US has proceeded in the shadow of laws barring employment discrimination as well as directly under federal regulation of government contractors. Proactive steps taken by employers in response to laws (such as Title VII of the Civil Rights Act of 1964) barring employment discrimination have contributed substantially to the economic advance of blacks, but not other minorities. Affirmative action under the Contract Compliance regulations have had a more muted impact, increasing the employment share of blacks by less than 1% per year. The impact on females and on non-black minorities appears to have been even smaller. These modest gains have been achieved without the imposition of strict quotas and have not imposed substantial productivity losses. These positive impacts have faded in recent decades.

Introduction

Affirmative action generally refers to a set of public policies meant to redress the effects of past or present discrimination. The term connotes active measures to level the playing field for access to education, to jobs, and to government contracts. For this reason, affirmative action has long been a political lightning rod. To its critics, it is symbolic of quotas unfairly and rigidly imposed. To its proponents, it is symbolic of redressing past wrongs and of leveling the current playing field. An empirical basis for policies to counterbalance current employment discrimination can be sought in (1) continuing judicial findings of systematic employment discrimination, (2) statistical evidence of wage and employment disparities across demographic groups, and (3) audit studies of employers' hiring behavior. For reviews of this evidence, see Altonji and Blank (2000), Blau (1998), Heckman (1998), and Fix and Struyk (1993).

While a wide variety of affirmative action policies have been implemented in different countries, much of the existing research concerns experience with the affirmative action directed at expanding employment opportunities for women and minorities in the United States; those polices are therefore the focus of this article. One defining characteristic of the policy in the United States is its ambiguity. The closest the US Congress has come to explicitly requiring affirmative action in employment is in the Americans with Disabilities Act of 1990, which explicitly requires that employers make 'reasonable accommodations' to hire the disabled. This requires that employers must do more than be blind to differences between the disabled and the able; they must actively invest to overcome these differences. It is worth noting that during a period when affirmative action policies were contentiously debated in the US, a law explicitly requiring affirmative action in employment was swiftly enacted without mention of affirmative action.

Affirmative Action in the Shadow of Title VII of the Civil Rights Act of 1964

Affirmative action in the United States has been implicitly encouraged by judicial interpretation of Title VII of the Civil Rights Act of 1964 (CRA), and explicitly required, but not defined, by Executive Order 11 246 applied to federal contractors. From its inception, the CRA has embodied a tension between words that bar employment discrimination and Congressional intent to promote voluntary efforts to redress discrimination. The courts have struggled with making room for what they sometimes have seen as the intent of Congress within the language of the CRA. For example, in some cases the Supreme Court read Section 703(J)'s statement that 'Nothing in the act shall require numerical balancing' as allowing numerical balancing as a remedy. In other cases, the Court held that the act should not be read to bar voluntary acts to end discrimination. These cases have left considerable, if ambiguous, room for affirmative action. Furthermore, the threat of costly disparate impact litigation under Title VII (following the *Griggs v. Duke Power* case) created considerable incentive for companies to undertake affirmative action. A number of researchers have found that employers responded to the expanding role of Title VII by instituting affirmative action offices, internal labor markets, and due process procedures (Edelman, 1992; Dobbin et al., 1993).

The extent of 'voluntary' (or at least nonjudicially directed) affirmative action taken in response to Title VII can be roughly estimated from existing work on the overall impact of Title VII. Because firms directly subjected to litigation under Title VII represent such a small share of employment, the bulk of the black economic advance credited to Title VII must be due to the statute's indirect effect in promoting 'voluntary' affirmative action. For example, Freeman (1973) shows the overall impact of Title VII in the time series of black–white earnings differentials. From this, the direct impact of Title VII litigation at companies sued for racial discrimination could be subtracted. While these impacts are substantial at the companies incurring litigation, employment at such companies makes up only a small fraction of employment. The impact on market wages from the outward shift in the demand curve for blacks at these companies can only be small given their small share of the market. Most charges of discrimination never result in litigation, and much litigation never reaches a judicial decision, leaving ambiguous interpretation and little trail. It is likely that the bulk of the changes attributed to Title VII are in a broad sense the result of affirmative action taken in response to the indirect threat, rather than the direct act, of litigation. Since the impact of Title VII is at least an order of magnitude greater than

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86003-1

that of the contract compliance program, and only a small share of this could have occurred at companies directly litigated against, the biggest impact of affirmative action, broadly defined, very likely took place in the shadow of Title VII.

One difficulty in evaluating the impact of a national law such as the 1964 CRA is that there is rarely a contemporaneous group unaffected by the law that can be used for comparison. In the case of Title VII, Chay (1998) overcomes this problem by examining the impact of the Equal Employment Act of 1972, which extended the reach of Title VII to smaller establishments. Chay finds that once smaller establishments came under Title VII's coverage, the relative employment and pay of blacks improved at these small establishments. Examining changes in establishment demographics over a longer period, Carrington et al. (2000) find that there was a rapid transition of blacks, and to a smaller extent women, into large establishments in the decade after passage of the 1964 CRA. They also find that this move to larger establishments could account for a narrowing of the racial wage gap – among men – about one-third to one-half the size of that caused by convergence in schooling attainment over the post-1964 period. Again, very little of this is the direct result of litigation, suggesting the importance of indirect pressure under Title VII and through the contract compliance program. Carrington et al. argue that the shift in black and female employment toward larger establishments is a distinctive outcome of such policies.

The motive force of Title VII behind affirmative action may well have weakened recently. In defending class claims under Title VII, divide and conquer has often been an attractive tactic. In Wal-Mart Stores, Inc., v. Betty Dukes, perhaps the most important civil rights case since Griggs, the Supreme Court in 2011 raised the bar to bringing class actions under Title VII. This invites employers to rely on decentralized personnel policies and practices to reduce the size of potential classes (along with statistical precision and potential damages) in class actions.

Affirmative Action through Contract Compliance

In the US, firms that hold federal contracts (and first-tier subcontractors) are required by Executive Order to pursue affirmative action. Because the Executive Order does not apply to all companies, it sets up a contrast that under some assumptions allows us to discern the impact of affirmative action under the contract compliance program by comparing the change in demographics at contractor firms with changes at noncontractors (firms that do not hold federal contracts). The interpretation of these comparisons would be complicated if each employer's demographics or anticipated change in demographics affected its selection or self-selection into contractor status. However, there is no evidence of this. To the contrary, initial demographics are generally not used as a prerequisite to becoming a contractor. As will be explained further below, because of peculiarities in the targeting of compliance reviews during much of the program's existence, and more recently influenced by legal challenges (see, for example, Beverly Enterprises, Inc. v. Herman, 130 F. Supp. 2d 1 (D.D.C. 2000)), firms with unusually low proportions of minorities or females were not likely to come under greater regulatory pressure. Given the regulations as implemented, there has been little reason for firms with low representation of minorities or females to avoid becoming a contractor. Since firms were not held rigidly to whatever promises they may have made to increase the employment of minorities or females, there was also no reason for firms that did not anticipate an increase in female or minority share to avoid contractor status. Given that enforcement creates a generalized pressure rather than a tight link from a firm's demographics to the targeting of enforcement, reverse causation is less of a concern. We can be more confident that the differences in changing demographics between contractor and noncontractor establishment firms are not artifacts of selection or self-selection into or out of contractor status. The research on the contract compliance program allows somewhat stronger conclusions about a weaker program because it relies not just on time-series variation but on a comparison of establishments with and without the affirmative action obligation.

Two mutually inconsistent criticisms of affirmative action have found prominent voice. The first criticism is that affirmative action goals without measurable results invite sham efforts; because these programs do not work very well, the argument goes, they should be eliminated. The second criticism is that an affirmative action 'goal' is really a polite word for a quota; in other words, affirmative action works too well, and therefore should be eliminated. Empirical research into affirmative action programs contradicts both of these criticisms. Affirmative action goals have played a statistically significant role in improving opportunities for minorities. At the same time, those goals have not resulted in quotas (Leonard, 1985b).

Federal Contract Compliance Program Regulations require that every contractor maintain an affirmative action plan consisting in part of a utilization analysis indicating areas of minority and female employment in which the employer is deficient, along with goals and timetables for good-faith efforts to correct deficiencies. To put this program in perspective, it is important to understand that, through most of its history, between 400 and 1000 investigative officers have been responsible for enforcing compliance. The ultimate sanction available to the government is debarment, in which a firm is barred from holding federal contracts. In the history of the program, there have been fewer than 60 debarments. The other sanction typically used is a back-pay award. This sanction too is used infrequently (Leonard, 1985a).

There have been a few studies of the impact of the Contractor Compliance Program (Ashenfelter and Heckman, 1976; Goldstein and Smith, 1976; Heckman and Wolpin, 1976; Smith and Welch, 1984; Leonard, 1984c; Rodgers and Spriggs, 1996). Studies of affirmative action's early years have consistently found that employment of black males increases faster at establishments that are federal contractors than at those that are not, with the possible exception of the period during which the Reagan Administration undercut enforcement. While significant, this effect has always been found to be modest in magnitude. In general, the impact on females and on non-black minorities has been less marked than that on blacks.

Between 1974 and 1980, black male and female employment shares increased significantly faster in contractor

establishments than in noncontractor establishments. These positive results are especially noteworthy, in view of the relatively small size of the agency and its limited enforcement tools. First, consider differences between contractors and noncontractors in the change in female and minority employment share without controlling for other possibly confounding variables. A summary measure, white males' mean employment share, declined by 5% among contractors compared to 3.5% among noncontractors. In other words, the contractors subject to affirmative action reduced white males' employment share by an additional 1.5% over 6 years. For other groups, the comparable difference between contractors and noncontractors in the change in employment share is 0.6% for white females, 0.3% for black females, 0.0 for other minority females, 0.3 for black males, and 0.1 for other minority males (Leonard, 1984c).

Controlling for establishment size, growth region, industry, occupational, and corporate structure, the annual growth rate of black male employment grew 0.62% faster in the contractor sector, and the annual growth rate of white males grew 0.2% slower. Contractor status shifted the demand for black males relative to white males by 0.82% per year. This is less than 1% per year, not a dramatic change, although it can accumulate over time. Studies of earlier periods by Ashenfelter and Heckman (1976) and Heckman and Wolpin (1976) generally find effects on black males of the same order of magnitude. Not surprisingly, growing establishments are better able to accommodate the regulatory pressures. Minority and female employment share increased significantly faster in growing establishments.

In addition to increasing the sheer numbers of minorities employed, the program also has had some success in helping move them up the career ladder. Affirmative action appears to increase the demand for poorly educated minority males as well as for the highly educated. Black males' share of employment increased faster in contractor establishments in every occupational group except laborers and white-collar trainees. Black females in contractor establishments increased their employment share in all occupations except technical, craft, and white-collar trainee. The positive impact of the contract program is even more marked when the position of black females is compared with that of white females (Leonard, 1984b). While some part of this improvement appears to reflect title inflation by employers upgrading detailed occupational categories with a relatively high representation of minorities or females (Smith and Welch, 1984), the occupational advance is accompanied by wage increases incompatible with pure title inflation (Leonard, 1986). The evidence does not support the contention that this is just a program for blacks with skills. To the contrary, it helps blacks across the board. Thus, affirmative action does not appear to have contributed to the economic bifurcation of the black community.

These studies have only sporadically been updated because of the difficulty of accessing administrative data. Rodgers and Spriggs (1996) find slightly larger positive contractor effects on blacks in 1992 than in 1982. Holzer and Neumark (1999) compared establishments that did and did not report using Affirmative Action in hiring using a 1992–94 cross-sectional survey of about 3000 employers in Atlanta, Boston, Detroit, and Los Angeles. They find that the white male's share of total employment was about 15–20% lower in establishments using Affirmative Action.

Stainback et al. (2005) look at the black–white difference in the distribution across occupations within establishments. They find that the black and white occupational distributions were significantly more similar to each other in contractor establishments. Comparing this pattern over time, they find that the contractor effect peaked between 1973 and 1980, becoming progressively smaller in the subsequent two decades.

In a useful combination of survey data with administrative records for 814 firms between 1971 and 2002, Kalev and Dobbin (2006) report that contractor status itself has little effect on minority or female managerial employment (and perversely hurts black females), but that both multiple compliance reviews and litigation under Title VII significantly slow white male representation in management. They separately measure the impact of having an affirmative action plan, and find that this significantly improves the managerial representation of black men and white women, while reducing that of white men. They also report that corporate diversity programs increase women's representation in management and reduce that of white men. Kalev and Dobbin find that the impact of affirmative action compliance reviews fell after the 1970s. Firms that were first reviewed in the 1970s had the largest impact, and that impact persisted into the next two decades. (Firms first reviewed in the 1970s show a 4 percentage point increase in the proportion of white women in management over the period.) In contrast, firms first reviewed in the 1980s or 1990s had smaller or insignificant impacts.

In a recent work, Kurtulus (2011) extends the analysis to the 1973–2003 period. Over these three decades, she finds black and Native-American employment shares grew faster among contractor than noncontractor firms. This result, suggesting effective affirmative action, is robust when controlling for firm size, occupational structure, and industry and firm fixed effects. However, Kurtulus also finds other patterns less suggestive of a well-functioning contract compliance program. For example, the positive impact of the contract compliance program waned over time; the relatively faster growth in minority and female employment shares at contractor relative to noncontractor firms seen in the 1970s faded in the following decades. Kurtulus also finds that, perversely, white male employment shares increased under affirmative action, those of white females fell, and affirmative action failed to increase Hispanic or Asian employment shares. Additionally, the impact of affirmative action changes within a firm as it matures; minority employment share gains are sharpest in the first 4 years after a firm becomes a federal contractor, as firms adjust to their new status. Encouragingly, Kurtulus also finds that employment gains do not dissipate after a firm ceases being a contractor. In a separate study, Kurtulus and Tomaskovic-Devey (2012) present a table showing that the share of females in middle management is significantly smaller in contractor firms, although the difference is tiny. They also find that the cumulative effect of having female top managers on the proportion of female middle managers does not differ significantly by contractor status.

A different way to judge the efficacy of affirmative action is to ask whether contractor firms are less likely to generate claims of employment discrimination for which the Equal Employment Opportunity Commission (EEOC) finds reasonable cause that discrimination occurred. Hirsh and Kornrich (2008) report that the rate of reasonable-cause findings is 32% lower for sex discrimination charges and 74% lower for race charges among federal contractors compared to noncontractors.

The employment goals to which firms agree under affirmative action are not vacuous; neither are they adhered to as strictly as quotas. Affirmative action 'goals' are often considered to be a euphemism for quotas. This appears to overstate what the regulatory authorities actually require: good-faith efforts to meet goals that are in the first instance chosen by the employer and may subsequently be negotiated with the Office of Federal Contract Compliance programs (OFCCP). Under the Contract Compliance Program, firms agree to set goals and timetables. Firms that agree to increase minority employment by 10% will, on average, increase minority employment by about 1%. Employers are not sanctioned for failing to meet their goals. A good-faith effort toward compliance in practice means that firms make it about one-tenth of the way toward the stated goal. Contractor goals do have a measurable and significant correlation with improvements in the employment of minorities and females at the reviewed establishments. At the same time, these goals are not being fulfilled with the rigidity one would expect of quotas (Leonard, 1985b).

Both the contract compliance program and Title VII have been criticized for inducing firms simply to hire a 'safe number' of minorities and women, irrespective of qualifications. According to this argument, whatever 'quota' is imposed should be the same for firms in the same industry and region. Application of such quotas would mean that firms in the same industry hiring out of the same labor market should, over time, start to look more and more like each other in terms of the percent of women and minorities. This does not generally happen, either under Title VII or as a result of affirmative action programs (Leonard, 1990).

Rather than implementing quotas, employers appear to use more sophisticated methods to implement affirmative action. Surveying employers, Holzer and Neumark (2000) found that firms engaging in affirmative action: (1) recruit more extensively by using a larger number of recruitment methods per worker hired; (2) screen more intensively, and pay less attention to characteristics such as welfare recipiency or limited work experience that usually stigmatize candidates; (3) provide more training after hiring; and (4) evaluate worker performance more carefully. Kalev and Dobbin find that while attempts to reduce managerial stereotyping through diversity training do not increase diversity, setting up affirmative action plans and making managers responsible does lead to improvements in managerial diversity. Reducing social isolation among women and minorities through networking and mentoring programs has a modest impact.

As a result, there is some evidence that the general pressures of affirmative action have succeeded in prompting employers to search more widely for new employees, but at the same time have been flexible enough not to impinge upon the economic performance of these firms. Holzer and Neumark (1999) find that although affirmative action employers tend to hire minorities with lower educational qualifications, these hires do not exhibit reduced job performance. For women hired under affirmative action, Holzer and Neumark find neither a performance nor a credential shortfall.

Future Research

Classic economic models of discrimination recognize that discrimination is not only unfair, but also inefficient. The available evidence suggests that affirmative action has helped to promote black employment, and to a lesser degree employment of females and of non-black minorities. But by themselves, these employment gains do not tell us whether affirmative action has reduced discrimination or gone beyond that point to induce reverse discrimination. Two fundamental questions remain: Has affirmative action reduced discrimination and has it helped integrate society?

Whether the employment gains associated with affirmative action and Title VII have reduced discrimination or gone beyond that to induce reverse discrimination can be addressed by asking whether industries that have come under the most pressure from Title VII or affirmative action, or industries that have increased their employment of minorities and women, have suffered in terms of productivity. The earliest paper to use this approach (of comparing relative earnings to relative wages) relied on highly aggregated data and found that minority and female employment gains under affirmative action had not reduced productivity (Leonard, 1984a). Direct testing of the impact of affirmative action on productivity finds no significant evidence of productivity decline, which implies a lack of substantial reverse discrimination. However, these results at the industry-state level are too imprecise to answer this question with great confidence.

Hellerstein et al. (1999) refine this approach of comparing wage differentials to estimated productivity differentials, and develop much more persuasive results with a detailed study of establishments. They find that pay premiums for older workers are roughly matched by productivity increases, but that the female pay penalty in the United States is generally not a reflection of lower productivity. The latter result suggests that far from imposing pervasive reverse discrimination against men, affirmative action for women has not yet succeeded in eliminating labor market discrimination. This direction of research holds great promise to move us beyond the formulaic debates over whether the wage differences that remain between groups after controlling for observable differences in qualifications and preferences are due to omitted human capital or discrimination.

A second approach to this question uses stock-market returns to ask whether investors consider Title VII litigation good for companies because it forces them to use labor more efficiently. Hersch (1991) finds a negative effect on investment returns. This might mean that investors do not believe that litigation will force the firms to become more efficient. Alternatively, investors might take the litigation as news that management is more inept than they thought. Wright et al. (1995) look at the stock market response to announcements by the US Department of Labor that a company is being recognized for having an exemplary affirmative action

program. The stock of these companies jumps immediately after the announcement. In a study of initial public offerings, Welbourne (2000) reports higher stock prices for companies with more women on their management teams.

The political question of whether affirmative action has served to unify or divide is of continuing concern. Sniderman and Piazza (1993) conduct survey-based experiments in the United States and report that just using the term affirmative action is enough to provoke more discriminatory responses.

Affirmative action-type policies in education and in government contracting have drawn increasing attention. Analysis of both education and public contracting face the challenge of measuring the impact of heterogeneous policies with substantial local variation that is difficult to pin down. For higher education, see Holzer and Neumark (2006) for a useful review. Less is known about set-aside programs at the federal, state, and local level (see Bates and Williams, 1996). Chatterji et al. (2010) find that city set-aside programs increase the rates of both black employment and black business ownership. Gains under these programs tend to quickly evaporate when set-asides are suspended (Blanchflower and Wainwright, 2005). The open questions include whether there is evidence of systematic discrimination in public or private contracting, as well as how discrimination and set-asides programs affect the profitability and growth of firms. Given the context of laws designed to limit discretion in public contracting, evidence of discrimination in contracting can be thought of as a miner's canary signaling loopholes in public contracting laws.

Other Mechanisms, Other Countries

One of the most interesting mechanisms for pursuing affirmative action can be found in the German policy to promote employment of the disabled. It is a play or pay policy. Employers face an explicit quota for employment of the disabled. The German government has drawn a bright line clearly stating minimum employment share for the disabled. However, together with this clear standard is a buy-out provision; firms that do not meet their quota, for whatever reason, pay into a special government fund. This mechanism is remarkably different from those used in the United States. First, the quota is explicit and stated in terms of numerical standards rather than abstract rights. This reduces the uncertainty, risks, delays, and costs of enforcement through the courts. There is also little opportunity to argue over the existence of a hidden quota. At the same time, the buy-out provision defuses complaints about the costly burdens imposed by rigid quotas, because employers experiencing the greatest difficulty meeting the quota can buy their way out. The nearest example of a US policy using a similar mechanism is the creation of tradable pollution rights under US environmental law. Both mechanisms are desirable on efficiency grounds. Firms that face the greatest difficulty reaching the legal standard can buy the right to pollute (or to hire fewer disabled) from firms that can more easily meet the standard. The aggregate level of disabled employment can be adjusted by changing the overall standard and/or the price, leaving to a decentralized market mechanism the question of which firms will employ the disabled and which will pay a fine. One additional benefit is that the penalty paid by firms below quota goes into a fund used to support the training and rehabilitation of the disabled.

Whatever its desirability on efficiency grounds, this mechanism appears to be politically impractical as a solution to race and sex discrimination in the United States. Explicitly setting quotas and putting a price on the right to employ fewer minorities or women raises political conflicts. The current policy places only an implicit price on discrimination. But it is apparently politically advantageous to frame the discussion in terms of absolute human rights, rather than to explicitly highlight the relative price society is willing to put on these rights through enforcement budgets and employment standards.

See also: Affirmative Action, Sociology of; Affirmative Action: Comparative Policies and Controversies; Disability: Sociological Aspects; Discrimination and the Law; Discrimination, Economics of; Discrimination: Racial; Employment and Labor, Regulation of; Equality and Inequality: Legal Aspects; Gender and the Law; Law and People with Disabilities; Race and the Law.

Bibliography

Altonji, J., Blank, R., 2000. Race and gender in the labor market. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics. North-Holland, Amsterdam.

Ashenfelter, O., Heckman, J., 1976. Measuring the effect of an antidiscrimination program. In: Ashenfelter, O., Blum, J. (Eds.), Evaluating the Labor Market Effects of Social Programs. Princeton University Press, Princeton, NJ, pp. 46–84.

Bates, T., Williams, D., 1996. Do preferential procurement programs benefit minority business? American Economic Review 86 (2), 294–297.

Blanchflower, D., Wainwright, J., 2005. An analysis of the impact of affirmative action programs on self-employment in the construction industry. National Bureau of Economic Research, Working paper #11793.

Blau, F., 1998. Trends in the well-being of American women, 1970–1995. Journal of Economic Literature 36 (1), 112–165.

Carrington, W., McCue, K., Pierce, B., 2000. Using establishment size to measure the impact of Title VII and affirmative action. The Journal of Human Resources 35 (3), 503–523.

Chatterji, R., Chay, K., Fairlie, R., 2010. The impact of city contracting set-asides on black self-employment and employment. Unpublished paper. UC, Santa Cruz.

Chay, K., 1998. The impact of Federal civil rights policy on black economic progress: evidence from the Equal Employment Opportunity Act of 1972. Industrial and Labor Relations Review 51 (4), 608–632.

Dobbin, F., Sutton, J., Meyer, J., Scott, R., 1993. Equal opportunity law and the construction of internal labor markets. American Journal of Sociology 99 (2), 396–427.

Edelman, L., 1992. Legal ambiguity and symbolic structures: organizational mediation of civil rights law. American Journal of Sociology 97 (6), 1531–1576.

Fix, M., Struyk, R., 1993. Clear and Convincing Evidence. Urban Institute Press, Washington, DC.

Freeman, R., 1973. Changes in the labor market for Black Americans, 1948–1972. Brookings Papers on Economic Activity 1, 67–120.

Goldstein, M., Smith, R., 1976. The estimated impact of the antidiscrimination program aimed at federal contractors. Industrial and Labor Relations Review 29 (3), 524–543.

Heckman, J., 1998. Detecting discrimination. Journal of Economic Perspectives 12 (2), 101–116.

Heckman, J., Wolpin, K., 1976. Does the contract compliance program work? An analysis of Chicago data. Industrial and Labor Relations Review 29 (3), 544–564.

Hellerstein, J., Neumark, D., Troske, K., 1999. Wages, productivity, and worker characteristics: evidence from plant-level production functions and wage equations. Journal of Labor Economics 17 (3), 409–446.

Hersch, J., 1991. Equal employment opportunity law and firm profitability. Journal of Human Resources 26 (1), 139–153.

Hirsh, C., Kornrich, S., 2008. The context of discrimination: workplace conditions, institutional environments, and sex and race discrimination charges. American Journal of Sociology 113 (5), 1394–1432.

Holzer, H., Neumark, D., 1999. Are affirmative action hires less qualified? Evidence from employer-employee data on new hires. Journal of Labor Economics 17, 534–569.

Holzer, H., Neumark, D., 2000. What does affirmative action do? Industrial and Labor Relations Review 53 (2), 240–271.

Holzer, H., Neumark, D., 2001. Assessing affirmative action. Journal of Economic Literature 38, 483–556.

Holzer, H., Neumark, D., 2006. Affirmative action: what do we know? Journal of Policy Analysis and Management 25 (2), 463–490.

Kalev, A., Dobbin, F., 2006. Enforcement of civil rights law in private workplaces: the effects of compliance reviews and lawsuits over time. Law & Social Inquiry 31 (4), 855–903.

Kurtulus, F., 2011. The Impact of Affirmative Action on the Employment of Minorities and Women over Three Decades: 1973–2003. Unpublished paper. University of Massachusetts, Amherst, MA.

Kurtulus, F., Tomaskovic-Devey, D., 2012. Do female top managers help women to advance? A panel study using EEO-1 records. The Annals of the American Academy 639, 173–197.

Leonard, J., 1984a. Anti discrimination or reverse discrimination: the impact of changing demographics, Title VII and affirmative action of productivity. Journal of Human Resources 19 (2), 145–174.

Leonard, J., 1984b. Employment and occupational advance under affirmative action. Review of Economics and Statistics 66 (3), 377–385.

Leonard, J., 1984c. The impact of affirmative action on employment. Journal of Labor Economics 2 (4), 439–463.

Leonard, J., 1985a. Affirmative action as earnings redistribution: the targeting of compliance reviews. Journal of Labor Economics 3 (3), 363–384.

Leonard, J., 1985b. What promises are worth: the impact of affirmative action goals. Journal of Human Resources 20 (1), 3–20.

Leonard, J., 1986. Splitting Blacks? Affirmative action and earnings inequality within and between races. Proceedings of the Industrial Relations Research Association (39th Annual Meeting), pp. 51–57.

Leonard, J., 1990. The impact of affirmative action regulation and equal employment law on Black employment. Journal of Economic Perspectives 4, 47–63.

Rodgers, W.M., Spriggs, W.E., 1996. The effect of federal contractor status on racial differences in establishment level employment shares: 1979–1992. American Economic Review 86 (2), 290–293.

Smith, J., Welch, F., 1984. Affirmative action and labor markets. Journal of Labor Economics 2 (2), 269–301.

Sniderman, P., Piazza, T., 1993. The Scar of Race. Harvard University Press, Cambridge, MA.

Stainback, K., Robinson, C., Tomaskovic-Devey, D., 2005. Race and workplace integration: a politically mediated process? American Behavioral Scientist 48 (9), 1200–1228.

Welbourne, T., 2000. Wall street likes its women: an examination of women in the top management teams of initial public offerings. ILR CAHRS, Cornell University, Working Paper Series, # 99–07.

Wright, P., Ferris, S., Hiller, J., Kroll, M., 1995. Competitiveness through management of diversity: effects of stock price valuation. Academy of Management Journal 38 (1), 272–287.

Affirmative Action, Sociology of

Anthony S Chen, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C.M. Swain, volume 1, pp. 226–229, © 2001, Elsevier Ltd.

Abstract

The sociology of affirmative action is a strongly interdisciplinary subfield within sociology. It is concerned with understanding the causes and consequences of policies that use membership in certain social groups to determine the allocation of resources or opportunities. It is also concerned with understanding the level and sources of public support and opposition to such policies. This article summarizes developments in these areas and identifies the areas that seem most promising for future researchers.

The sociology of affirmative action is a subfield that sits at the intersection of political sociology, race and ethnic relations, work and occupations, sociology of education, historical sociology, and the sociology of gender. It is strongly interdisciplinary, and it is connected with ongoing lines of related research in economics, political science, and history. The subfield has grown robustly in recent years through the contributions of both qualitative and quantitative researchers, who have explored a range of sociologically significant and policy-relevant questions.

What Is Affirmative Action?

The term 'affirmative action' is typically used in connection with a heterogeneous set of public or private policies or programs in which a person's membership in a socially defined group is a factor that determines the allocation of resources or opportunities. The social groups that are targeted by affirmative action include, but are not limited to selected racial and ethnic minorities as well as women. How group membership is identified, is weighed, and enters into the decision-making process can vary substantially from one context or setting to another.

Affirmative action programs originated roughly half a century ago in the United States, where they took root in both the public and private sector in employment, government contracting, and higher education. The phrase 'affirmative action' first appeared in proposals for antidiscrimination legislation in the 1940s, but it was injected prominently into the discourse of civil rights with President Kennedy's issuance of Executive Order 10 925 in 1961.

Affirmative action programs have assumed many different forms and appeared in a wide range of settings over time. Among their numerous incarnations are the 'goals and timetables' for the utilization of women and minorities that federal contractors are required to write out and keep on file; court-ordered or court-approved hiring quotas that were established by municipal police and fire departments trying to resolve employment discrimination lawsuits filed under Title VII of the Civil Rights Act of 1964 (CRA); various local, state, and federal laws that 'set aside' a proportion of contracts or contracting dollars for minority-owned businesses or subcontractors; various types of recruiting, hiring, and promotion programs voluntarily instituted and operated by private companies, sometimes under the aegis of guidelines issued by the Equal Employment Opportunity Commission (EEOC); and a mixed assortment of recruiting and admission programs at colleges, universities, and K-12 schools – both public and private. The most capacious conceptions of affirmative action also include race-conscious policies in legislative districting and jury selection.

Over time, multiple rationales have been offered for the establishment and maintenance of affirmative action, including the compensatory justifications that were commonplace in the early years to the more recent idea that affirmative action is essential to the attainment of racial and ethnic diversity, which is seen as educationally valuable.

Affirmative action programs were not uncommon in the United States, despite the intense political and legal controversy that often accompanied their adoption and operation. Recent years have witnessed a retrenchment in the geographical scope and institutional prevalence of affirmative action as well as a reduction in the amount of weight that group membership is given, but a meaningful array of more circumscribed programs nevertheless remains in place today.

Since the advent of affirmative action in the United States, several other countries have adopted and adapted it in a variety of different forms and a range of different contexts; among them were Brazil, India, South Africa, and Malaysia.

The Sociology of Affirmative Action

Sociologists have shown interest in a number of research questions about affirmative action, but the literature focuses on three distinct but intertwined lines of inquiry.

What Explains the Developmental arc of affirmative Action Programs in Employment and Education?

Advent, Growth, and Persistence in Employment

The most robustly developed strand of the literature is highly historical in orientation, and it explores why affirmative action in employment was initially adopted. Alongside historians and political scientists, sociologists have been mainly concerned with accounting for developments spanning the period from Kennedy through Nixon's first term – especially the series of policies issued under Executive Order 11 246 (1965), which

called upon federal contractors to take 'affirmative action.' These policies assumed definite form in the late 1960s through executive initiatives like the Philadelphia Plan – first worked out tenuously in 1967 under Johnson and then reformulated more securely in 1969 under Nixon. Both versions of the plan required federal construction contractors to write down goals for hiring minority workers. The idea of using written goals to stimulate and gauge progress had actually itself come from a group of large, private companies organized in 1961 under Kennedy's executive order. Members of the group, known as Plans for Progress, had for several years been voluntarily developing establishment-level plans for 'achieving equal employment opportunity through affirmative action.' In 1970, Order Number 4 extended affirmative action to cover all federal contractors, requiring them to set 'goals and timetables' for the utilization of racial minorities in their workforce. And 1 year later, Revised Order Number 4 extended the same requirement to cover women.

There has also been interest in accounting for a series of court rulings that solidified the legal and constitutional foundations of affirmative action. Most of these were handed down during the Burger Court (1969–86), including the decision not to grant *certiorari* in the case of *Contractors Association of Eastern Pennsylvania vs Secretary of Labor*, 442 F.2d 159 (3d Cir. 1971), which had the effect of validating the legality of the Philadelphia Plan and affirmative action in federal contracting more generally. Other key opinions of interest include *Griggs vs Duke Power Company*, 401 US 424 (1971), which among other things held that the EEOC should be given 'great deference' from the courts in interpreting the implementation of Title VII of the CRA, eventually including the issuance of regulations concerning affirmative action; *United Steelworkers of America vs Weber*, 443 US 193 (1979), which cleared the legality of voluntary affirmative action plans put into place by private employers for the purpose of remedying racial imbalances in historically segregated job categories; and *Fullilove vs Klutznick*, 448 US 448 (1980), a plurality opinion that tenuously upheld the constitutionality of a set-aside program established by Congress.

The initial origins of affirmative action were once relatively murky, but several generations of research have done much to illuminate them. A total consensus is far from evident, but there has been a loose convergence of evidence-based perspectives that differ mainly in their points of interpretation and emphasis.

One of the most developed lines of research highlights the significance of pressure from below. There is indeed solid evidence showing that affirmative action emerged in substantial measure as a response to grassroots protests mobilized by northern civil rights activists, who had been long concerned with the slow pace and meager scope of racial change that had been wrought in the skilled trades by liberal antidiscrimination policies such as fair employment practices (FEP) laws. The sense that something more than simple nondiscrimination and desegregation was needed percolated upward slowly from the grassroots all over the country. By 1965, President Johnson himself had seen fit to announce in a speech at Howard University that "freedom is not enough." In a language that would subsequently be taken as a rationale for affirmative action, he maintained that "it is not enough just to open the gates of opportunity. All our citizens must have the ability to walk through those gates."

A related but distinct perspective goes a step further, offering evidence consistent with the idea that affirmative action was not strictly a reactive response to outside protest but also a consequence of regulatory capture. According to this view, affirmative action was instituted at the behest of an increasingly radical civil rights lobby whose redistributive objectives and policy preferences began dominating the agendas of policy-making institutions throughout the executive and judicial branches.

Political culture and partisan politics are the focal points for a different line of research. The most influential study among them points to the crucial significance of the urban riots that shook many American cities in the late 1960s. It provides some evidence that the riots had the effect of legitimating color-conscious policies like affirmative action, making it possible for partisan actors to deploy it for political purposes – such as when President Richard M. Nixon used it to aggravate tension within the New Deal coalition, particularly between African-Americans and organized labor.

Still another line of work highlights the fragmented and decentralized character of political institutions in the United States, where policy-making authority is dispersed horizontally across the separation of powers and vertically across the ladder of federalism. This strand of the literature has adduced evidence that the structure of American political institutions presented a unique set of political opportunities that facilitated the rise of race-conscious policies, which would have faced much stronger headwinds if they had tried to go through Congress instead.

A final perspective points to the persistent failure of Congress to pass a law that would have promoted the equal treatment of individuals in the labor market through the operation of a strong administrative agency with cease-and-desist authority along the lines of the National Labor Relations Board. Liberals had campaigned for such an FEP law since the mid-1940s, but there is evidence that a coalition of business groups, conservative Republicans, and rural and suburban whites joined with southern Democrats to stand in the way of its passage. With the path through Congress block, what resulted was a regulatory vacuum that ironically encouraged the pursuit of race-conscious policies through a combination of executive and judicial action.

If the initial advent of affirmative action in employment has attracted much of the interest from sociologists, there has nevertheless been some interest in understanding the subsequent expansion of affirmative action. Special concern has been devoted to accounting for how and why various types of affirmative action came to include groups other than African-Americans, including Latinos, Asian-Americans, Native Americans, and women.

One salient approach highlights the importance of social protest and legal mobilization around Title VII of the CRA, especially in the case of women. A somewhat different approach focuses on policy elites and their perceptions. It suggests that African-Americans were originally the focus of affirmative action because the urban riots of the late 1960s led them to be seen by policy elites as societally disadvantaged and morally worthy of special treatment. Mexican-Americans,

Asian-Americans, and Native Americans were then swiftly included as beneficiaries of affirmative action because policy elites likened them to African-Americans, who were regarded as a benchmark of sorts. By contrast, it took women some struggle to achieve inclusion because they were seen by policy elites as morally worthy but categorically different than African-Americans. White ethnics were never included despite struggling for inclusion because they were seen by policy elites as morally unworthy and categorically different.

Affirmative action programs also expanded, in the sense that many companies without federal contracts also began instituting them voluntarily. This growth was fostered in large measure by personnel experts, who advised executives that taking affirmative action would convey a signal to legislators and judges that a 'good faith' effort to achieve equality was being made. Many companies heeded the advice, establishing affirmative action offices and appointing affirmative action officers over the course of the 1970s and 1980s. For instance, more than 25% of the companies in a 1986 survey reported having affirmative action officers, five times the percentage in 1972.

Sociological research on the advent and development of affirmative action in other employment-related areas is less abundant, but it touches on many of the same themes. Here too sociologists have argued for the importance of the riots, pointing to evidence supporting the claim that government programs meant to promote minority business were established during the late 1960s largely in the hope that they would stave off further rioting and restore order in American cities. Among these were the Section 8(a) programs of the Small Business Administration (SBA), wherein the SBA served as a prime contractor for the federal government and sought to promote the award of subcontracts to minority-owned businesses. A similar motive was evidently behind Nixon's issuance of Executive Order 11 458 (1969), which established the Office of Minority Business Enterprise to promote minority-business ownership. In both cases, programs initially meant to address the acute circumstances of African-Americans were quickly extended by analogy to racial minorities like Latinos and Native Americans, even though they themselves did not demand government action along such lines. The value of such programs became increasingly attractive to politically minded policy-makers over time. By 1977, there was sufficient momentum behind the idea of promoting minority business that Congress included a provision in the Public Works Employment Act requiring that 10% of certain funds spent under the law to go toward the purchase of goods or services provided by minority-owned businesses. Such programs – along with numerous state and local programs that were created in their image – would endure through the remaining years of the Burger Court before finding themselves on more treacherous legal, constitutional, and political terrain.

Advent, Growth, and Persistence in Higher Education

Sociological research on the establishment and growth of race-conscious affirmative action in higher education is less plentiful than research on employment, but it is growing.

Here too the 'riot hypothesis' has left a substantial footprint in the sociological literature. There is especially strong and granular evidence supporting the hypothesis in the case of Harvard, Yale, and Princeton, which abandoned race-neutral admission practices in their college admissions and witnessed a significant uptick in the enrollment of black students only after the Watts riot in 1965. Related studies suggest the possible significance of violence on campus – at the very least the threat of violence that accompanied student protest. Whatever the source of disruption, it is argued that disruption led school administrators to accept and even promote race-conscious policies, mainly out of the belief that they might serve as a steam valve that would relieve the pressure of discontent building from below.

But evidence supporting the 'riot hypothesis' in the case of college admissions is fairly limited for schools outside the Big Three. One recent study shows that affirmative action in college admissions was actually adopted in two waves. The 'first wave' started before Watts. At schools such as the University of California, Los Angeles, the University of Michigan, and Cornell University, school administrators of different partisan commitments began to experiment with race-conscious admission programs in the early 1960s, often in response to the nonviolent mobilization of the civil rights movement – such as the confrontation in Birmingham and the March on Washington in 1963. Many administrators were motivated by the belief that it was morally incumbent upon them to not just to do what they could to level the metaphorical playing field, but to take special steps to make up for decades of societal discrimination that had left so many African-Americans in a less-than competitive position in so many areas of American life. Indeed, there is mounting evidence to suggest that the initial emergence of affirmative action at many colleges occurred in the early 1960s and should be seen primarily as a northern 'ripple effect' of the campaign against racial inequality and Jim Crow segregation. To be sure, riots played a role during a 'second wave' of affirmative action, which occurred during the late 1960s, when institutions that were closely associated with Protestant elites were finally motivated to affirmatively incorporate race into their admissions criteria. Student protest in the wake of King's assassination also led many schools to intensify their existing affirmative action by putting a greater weight on race. But key numbers of colleges had already begun to incorporate racial considerations into their admission process long before uprisings in Watts, Detroit, or Newark began grabbing headlines.

For a variety of reasons that sociologists are continuing to explore, including the idea that college administrators were seeking to bolster the legitimacy of their own specific institutions as well as the legitimacy of the higher education system more broadly, affirmative action programs spread quickly in the late 1960s and early 1970s. There is also evidence that professional and accreditation associations played a key role in exerting pressure on the admission programs of law and medical schools.

But the programs were never without their critics, and controversy quickly moved into the open. Affirmative action became the subject of litigation in the federal courts, where a number of white plaintiffs filed lawsuits alleging that affirmative action violated Title VI of the CRA and the Equal Protection Clause of the Fourteenth Amendment. The opinion of a splintered Burger Court in *Regents of the University of California vs Bakke*, 438 US 265 (1978) settled some questions and left other questions ambiguous. The plaintiff in the case, Alan Bakke, had failed to gain admission to the medical school at

the University of California, Davis, which had for a time set aside 16 out of 100 seats for members of 'disadvantaged' groups. The court struck down such programs – and by implication racial quotas and separate tracks – but it appeared to accept the use of race as a single, 'plus factor' in a larger system of individualized, holistic admissions, pointing approvingly to the system operated by Harvard College. Moreover, it indicated that the underlying reason for instituting race-conscious affirmative action mattered. Race could not be considered as a way of compensating certain applicants for experiencing 'societal discrimination,' but it could be considered in the service of crafting a racially and ethnically diverse class – which the court accepted as having educational value for all students on campus.

Affirmative action endured at many American colleges and universities after *Bakke*, most obviously because the opinion offered administrators a viable legal and constitutional rationale for the restrained continuation of such programs – even if the rationale required further clarification and invited future litigation. But there is also some evidence from empirical research on certain large, selective public universities that a range of 'isomorphic' forces may have converged to facilitate the persistence of affirmative action, despite a political context that was growing more infelicitous. This included the 'coercive isomorphism' of *Bakke* itself, which imposed constitutional constraints on ways that universities could consider race in admissions. It may have also entailed a kind of 'normative isomorphism' in which diversity became a taken-for-granted organizational norm promoted by admission officers working in a rapidly professionalizing field. Finally, it may have included a kind of 'mimetic isomorphism' in which institutions coped with legal uncertainty in the wake of *Bakke* by copying the practices of the most prestigious or successful schools.

Partial Retrenchment in Employment and Education

Since the time that it was put into place, affirmative action has experienced varying degrees of retrenchment in employment and education, but a sociological literature on the causes of retrenchment is only beginning to develop.

It was during the Reagan–Bush era and the initial years of the Rehnquist Court (1986–2005) that affirmative action began to find itself under serious assault on multiple fronts. By then, it had completed a journey from partisan cipher to partisan touchstone, and the most vocal critics of the policy were frequently Republicans. Among those who were most aggressive in their opposition was President Ronald Reagan, who had campaigned against 'quotas, ratios, and numerical requirements.' His administration sought to limit affirmative action requirements to only the largest contractors, close the federal agency responsible for monitoring affirmative action compliance, and cease the use of goals and timetables by more than 50 cities.

Sociologists have pointed out that Reagan's opposition to affirmative action contributed in certain ways to the rise of diversity management in corporate America as well as the proliferation of the diversity rationale in higher education.

But the administration and other critics at the time did not experience much success in actually rolling back affirmative action. The most significant developments occurred at something of a distance from electoral politics, in the federal courts, where a series of decisions the Rehnquist Court signaled a deepening shift in the way that cases involving racial consideration would be decided. The first of these, *City of Richmond vs J.A. Croson Co.*, 488 US 469 (1989), involved a city set-aside program that required prime contractors to award 30% of the dollar value of their construction contracts to 'minority-business enterprises.' The court held that state and local programs that invoked racial considerations were indeed subject to the highest level of judicial review, known as strict scrutiny. This was essentially the same standard that was applied in but perhaps not fully fleshed out by *Bakke*. It required race-conscious policies to meet a two-pronged test. First, such policies had to serve a 'compelling governmental interest.' Second, they had to be 'narrowly tailored' to serve such an interest. The court was prepared to accept that racial considerations were valid remedies in the case of 'identified discrimination,' but it struck down the set-aside program in *Croson* on the grounds that the City of Richmond had "failed to identify the need for remedial action in the awarding of its public construction contracts."

After 6 years the court would hand down *Adarand Contractors Inv. vs Pena* 515 US 200 (1995), which held that federal programs taking race into account were also subject to strict scrutiny. The court in the next year declined to hear *Hopwood vs Texas* 78 F.3d 932 (5th Cir. 1996), in which a panel of judges from the Fifth Circuit struck down an affirmative action program operated by the University of Texas Law School, partly on the grounds that the attainment of diversity for educational purposes was not a 'compelling governmental interest' and therefore did not pass muster under one prong of strict scrutiny.

During the mid-1990s, critics of affirmative action began to experience success in state politics. In 1995, led by businessman Ward Connerly, an ally of Republican governor Pete Wilson, the Regents of the University of California voted to end affirmative action in university admissions, employment, and contracting. A year later, California voters passed Proposition 209 by a 54–46 margin, buoyed in substantial measure by the votes of non-Jewish, white Californians. The 'California Civil Rights Initiative,' as it was called by supporters, amended the state constitution to say that the "state shall not discriminate against, or grant preferential treatment to, any individual or group on the basis of race, sex, color, ethnicity, or national origin in the operation of public employment, public education, or public contracting." One sociologist points out that the wording of the initiative may have confused voters: An exit poll by the *Los Angeles Times* indicated that a quarter of those polled who voted for the initiative said they actually supported 'affirmative action.' Several states followed in California's footsteps in subsequent years, passing similarly worded laws, initiatives, or amendments. Among these included action by Washington State (1998), Michigan (2006) Nebraska (2008), Arizona (2010), New Hampshire (2011), and Oklahoma (2012). Florida ended affirmative action in 1999 through Governor Jeb Bush's executive order. Only twice did measures calling for the ban of affirmative action fail when put to a vote. The first time came in Houston, where the wording of the initiative favored those who wanted to preserve affirmative action. The second time came in Colorado in 2008, when Barack Obama was elected president.

Concerned about preserving the racial diversity of their undergraduate population, Texas, Florida, and California put

into place 'percentage plans' in lieu of affirmative action. These plans generally guaranteed admission to a state university for high school students who graduated at the top of their high school class – the top 10% for Texas (1997), the top 20% for Florida (1999), and the top 4% for California (1999).

But the retrenchment of race-conscious affirmative action eventually reached a limit in the federal courts. A pair of cases arising out of Michigan defined the legal and constitutional boundaries. In *Gratz vs Bollinger* 539 US 244 (2003), the Supreme Court struck down the admission program operated by the University of Michigan's College of Literature, Science, and the Arts – which automatically granted applicants from 'underrepresented' backgrounds a set number of points in the admission process. The court held that the program was not narrowly tailored and therefore ran afoul of the Equal Protection Clause. The court also found that the program violated Title VI of the CRA. But the court upheld the validity of the admission program operated by the University of Michigan's Law School in *Grutter vs Bollinger* 539 US 203 (2003). With the aim of capturing the educational benefits that it believed flowed from having a racially and ethnically diverse class, officials at the Law School considered race in a more flexible, individualized fashion, and the court found that their program was both narrowly tailored and served a compelling state interest – and therefore did not offend the Equal Protection Clause. The decision was especially notable because the court explicitly stated an expectation that in 25 years time affirmative action would no longer be necessary for the purposes of composing a racially and ethnically diverse class.

At present, legal contestation of affirmative action remains ongoing. In *Fisher vs Texas*, a panel of judges on the US Fifth Circuit Court of Appeals received the case on remand from the US Supreme Court and found the affirmative action program operated by the University of Texas constitutionally acceptable. The full Fifth Circuit declined to rehear the panel decision, but Fisher reportedly plans an appeal.

What Are the Consequences of Affirmative Action?

A second broad line of inquiry in the sociology of affirmative action investigates the effects of such policies – both their establishment and their elimination – on a range of employment and educational outcomes. This research confronts a number of challenges to credible inference, including the need to distinguish the effect of affirmative action from the effect of race-neutral policies as well as the 'nonrandom assignment' of affirmative action policies to unions, firms, and educational institutions. As a result, much of the research is suggestive – sometimes highly suggestive – but rarely if ever conclusive.

This is certainly the case with the literature on affirmative action in employment. But much of the research in multiple fields appears to consistently point in the direction of a positive effect on minority outcomes. One strand of work in economics is based on a comparison of employment outcomes at contractor versus noncontractor establishments, and it has yielded evidence that affirmative action redistributed employment over time from white males to white females and racial minorities as well as evidence of wage gains on the part of white females and racial minorities. Another strand features similar findings; it is based on metropolitan-level data on a broader slice of the labor force. The magnitude of the effect appears to be fairly modest regardless of the data source. A forthcoming paper on city contracting set-asides finds evidence that they significantly increased the business ownership rates of blacks. The impact on employment is not entirely clear and depends upon the nature of assumptions about previous trends. In sociology, one particularly strong study uses original longitudinal data on a fairly large sample of establishments over a 30-year period to estimate models of managerial diversity with both establishment- and year-fixed effects. It finds that affirmative action plans are generally associated with increases in the share of private-sector management positions held by white women, black men, and black women (in certain industries).

More disagreement exists concerning the effects of affirmative action in higher education. The most comprehensive study along these lines – based on the Mellon Foundation's College and Beyond database, which contains information on students who matriculated at 28 selective colleges at three different points in time – indicates that affirmative action has had significant benefits for black students. The authors find that the weight given to race is small, but the benefits are large and meaningful. While they find evidence of academic underperformance by black students, they point out that black students graduate from College and Beyond schools at high rates, attain graduate and professional degrees at high rates, earn high incomes, and become highly involved in civic and community affairs, especially in leadership capacities. But some researchers are highly critical of the study, and they argue that affirmative action actually harms black students because it gives such a large weight to race in the admission process that it catapults academically underprepared black students to selective schools where they find it difficult to perform well, thereby leading them to experience a range of poor academic outcomes. Affirmative action creates an academic 'mismatch' between black beneficiaries and the schools where they are admitted.

The balance of the literature goes against the 'mismatch' hypothesis. Some sociologists have found very little support for it when carefully analyzing data from such surveys as College and Beyond, High School and Beyond, the National Education Longitudinal Study, and the National Longitudinal Survey of Freshmen. Other sociologists have reached similar inferences after exploiting a 'natural experiment' generated by a quirk in the undergraduate admissions process at the University of California in 2004. Still other sociologists argue that academic costs of affirmative action for black students (mainly in terms of lower grades and class rank) are usually outweighed by the benefits of attending a more selective school. Yet certain researchers have drawn different conclusions based on their own empirical analysis. One prominent study analyzes data on law school and lawyers from several sources, including the Bar Passage Study of the Law School Admissions Council, the National Survey of Law School Performance, and the 'After the JD' survey. This line of research has attracted a vigorous response from numerous researchers across the social sciences, including sociologists. In fact, many of them teamed up to criticize it in an *amicus* brief filed on behalf of the *University of Texas* in the *Fisher* case.

What Does the Public Think of Affirmative Action?

A third broad line of inquiry in the sociology of affirmative action examines various facets of public opinion on affirmative action.

What 'affirmative action' is understood to mean can vary substantially from one person to another, and so the level of public support for affirmative action can be challenging to ascertain and characterize. How much there seems to be depend on the race of the respondent, what particular type of policy or program is referenced, and precisely how the question is worded. At the same time, public support for affirmative action does appear to have changed in certain, nontrivial ways over time.

Broadly speaking, sociologists and other researchers studying opinion polls over the last 40 years have usually found low levels of support for affirmative action when racial preferences are invoked by the wording of the question. In a series of polls taken by Gallup from the late 1970s to the early 1990s, only one in ten respondents tended to express support for affirmative action when 'preferential treatment' was explicitly juxtaposed to 'ability' and 'test scores' as the main basis for 'getting jobs and places in college.' In surveys conducted by CBS News and the *New York Times* (*NYT*) during the mid-1980s to the mid-1990s, around three in ten respondents supported affirmative action when they were asked if they believed that "preference in hiring and promotion should be given to blacks today" in situations "where there has been job discrimination against blacks in the past."

But researchers have sometimes found modest and even high levels of support for affirmative action, often when 'affirmative action' is mentioned and 'preferential treatment' is avoided. Some surveys in late 1970s and early 1980s revealed that more than two-thirds of respondents supported 'affirmative action programs' in both employment and higher education 'provided there are no rigid quotas.' Gallup reported in 1995 that nearly three-quarters (see Steeh and Krysan 1996: 134). of respondents in one poll approved of "companies making special efforts to find qualified minorities and women and encouraging them to apply for jobs." In a 2013 CBS News/*NYT* poll, half of respondents indicated that they favored "affirmative action programs for minorities in hiring, promoting, and college admissions," and Pew Research Center reported the next year that nearly two-thirds of respondents in a poll agreed that "affirmative action programs designed to increase the number of black and minority students on college campuses are a good thing."

What is clear is that racial differences in levels of support for affirmative action persist. A majority of white respondents in the General Social Survey (GSS), roughly 60%, has consistently endorsed the broad belief that the government does not have a 'special obligation to help improve' the 'living standards' of black Americans. Black attitudes continue to remain distinct from white attitudes, though there have been some shifts over time. The overwhelming majority that unequivocally supported the idea of a special obligation has shrunk to a large minority today, and a similarly sized minority of black respondents expresses ambivalence about the idea.

Even stronger racial differences are manifest in responses to questions that ask specifically about 'preferences.' The GSS began asking in 1994 whether respondents favored 'preference in hiring and promotion of blacks,' and nine out of ten whites consistently expressed opposition. A substantial minority of blacks opposed preferences in 1994, while today a small majority opposes them. These broad patterns are echoed in other data, including the National Election Survey.

One of the most exciting new lines of research in the area of public opinion and affirmative action uses experiments to explore how ideas about meritocracy – something critically linked to the issue of affirmative action – are not fixed but malleable. For instance, one recent lab experiment shows that white students lower the academic standards (i.e., minimum class rank) that are used to assess white applicants for admission when exposed to a treatment showing that first-year enrollment is trending toward an Asian-American plurality. In comparison, when exposed to the same treatment, white students raise the academic standards (i.e., minimum test score) that are used to assess Asian-American applicants.

Conclusion

The sociology of affirmative action has grown enormously over the last 20 years, but the research agenda remains vibrant and open.

The origins of affirmative action in employment are well understood. But further research seems needed to clarify its expansion in employment and its origins in higher education. Further research also seems needed to explain the contours and content of the partial retrenchment it has experienced. What would be especially useful are integrated accounts that trace the interplay between the state and national politics as well as connect landmark decisions in the federal courts with partisan politics, institutional strategies, and social and legal movements.

There is now a great deal of sociological research on the consequences and impact of affirmative action, but further research seems needed here as well. More studies of affirmative action in employment featuring high-quality identification seem desirable. It also seems valuable for sociologists interested in affirmative action and higher education to investigate exactly what kinds of policies and programs can help deliver the educational outcomes that are thought to flow from having a diverse student body. The source of black academic underperformance is also a topic worthy of additional research.

Finally, sociologists have a valuable opportunity to exploit a recent surge of interest in experimental methods to explore the apparent malleability and variability of public opinion on affirmative action.

Bibliography

Alon, S., Tienda, M., 2005. Assessing the "Mismatch" hypothesis: differences in college graduate rates by institutional selectivity. Sociology of Education 78, 294–315.

Berrey, E.C., 2011. Why diversity became orthodox in higher education, and how it changed the meaning of race on campus. Critical Sociology 37, 573–596.

Bobo, L., et al., 2012. The *real* record on racial attitudes. In: Mardsen, P. (Ed.), Social Trends in American Life: Findings from the General Social Survey since 1972. Princeton University Press, Princeton, NJ.

Bowen, W.G., Bok, D., 1998. The Shape of the River: Long-term Consequences of Considering Race in College and University Admissions. Princeton University Press, Princeton, NJ.

Chatterji, A.K., Chay, K.Y., Fairlie, Robert W., 2013. The Impact of City Contracting Set Asides on Black Self-Employment and Employment. NBER Working Paper No. 18884.

Chen, A.S., 2008. The Fifth Freedom: Jobs, Politics, and Civil Rights, 1941–1972. Princeton University Press, Princeton, NJ.

Dobbin, F., 2009. Inventing Equal Opportunity. Princeton University Press, Princeton, NJ.

Espenshade, T.J., Radford, A.W., 2009. No Longer Separate, Not yet Equal: Race and Class in Elite College Admission and Campus Life. Princeton University Press, Princeton.

Fischer, M.J., Massey, D.S., 2007. The effects of affirmative action in higher education. Social Science Research 36, 531–549.

Garces, L.M., 2013. Understanding the impact of affirmative action bans in different fields of graduate studies. American Educational Research Journal 50, 251–284.

Graham, H.D., 1990. The Civil Rights Era: Origins and Development of National Policy, 1960–1972. Oxford University Press, New York.

Grodsky, E., 2007. Compensatory Sponsorship in higher education. American Journal of Sociology 112, 1662–1712.

Kalev, A., Dobbin, D., Kelly, E., 2006. Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. American Sociological Review 71, 589–617.

Karabel, J., 2005. The Chosen: The Hidden History of Admission and Exclusion at Harvard, Yale, and Princeton. Houghton Mifflin, Boston, MA.

Kurlaender, M., Grodsky, E., 2013. Mismatch and paternalistic justification for selective college admissions. Sociology of Education 86, 294–310.

Lieberman, R.C., 2005. Shaping Race Policy: The United States in Comparative Perspective. Princeton University Press, Princeton, NJ.

Lipson, D., 2011. The resilience of affirmative action in the 1980s: innovation, isomorphism, and institutionalization in University admissions. Political Research Quarterly 64, 132–144.

Lipson, D., 2007. Embracing diversity: the institutionalization of affirmative action as diversity management at UC-Berkeley, UT-Austin, and UW-Madison. Law and Social Inquiry 32, 985–1026.

MacLean, N., 2005. Freedom Is Not Enough: The Opening of the American Workplace. Russell Sage Foundation Books and Harvard University Press, New York, Cambridge, MA.

Massey, D., et al., 2003. The Source of the River: The Social Origins of Freshman at America's Selective Colleges and Universities. Princeton University Press, Princeton, NJ.

Massey D., Mooney M. The effects of America's three affirmative action programs on academic performance. Social Problems 54, 99–117.

Moreno, P.D., 1999. From Direct Action to Affirmative Action: Fair Employment Law and Policy in America, 1933-1972. Louisiana State University Press, Baton Rouge, LA.

Sampson, F.L., 2013a. Altering public university admission standards to preserve white group position in the United States. Comparative Education Review 57, 369–396.

Sampson, F.L., 2013b. Multiple group threat and malleable white attitudes toward academic merit. Du Bois Review 10, 233–260.

Sander, R., 2004. A systematic analysis of affirmative action in American law schools. Stanford Law Review 57, 367.

Skrentny, J.D., 1996. The Ironies of Affirmative Action: Politics, Culture, and Justice in America. University of Chicago Press, Chicago.

Skrentny, J.D., 2002. The Minority Rights Revolution. Harvard University Press, Cambridge, MA.

Steeh, C., Krysan, M., 1996. Poll trends: affirmative action and the public, 1971–1995. Public Opinion Quarterly 60, 128–158.

Stryker, R., Pedriana, N., 1997. Political culture wars 1960s style: equal employment opportunity-affirmative action and the Philadelphia plan. American Journal of Sociology 103, 633–691.

Sugrue, T.J., 2004. Affirmative action from below: civil rights, the building trades, and the politics of racial equality in the urban North, 1945-1969. Journal of American History 91, 145–173.

Stevens, M., 2007. Creating a Class: College Admissions and the Education of Elites. Harvard University Press, Cambridge, MA.

Stulberg, L., Chen, A., 2014. The origins of race-conscious affirmative action in undergraduate Admissions: a Comparative analysis of institutional change in higher education Sociology of Education 87, 36–52.

Swain, C.M., 2001. Affirmative action, legislative history, judicial interpretations, public consensus. In: Smelser, N., Wilson, W.J., Mitchel, F. (Eds.), Racial Trends and Their Consequences. National Academy Press, Washington, DC.

Thernstrom, S., Thernstrom, A., 1997. America in Black and White: One Nation, Indivisible. Simon and Schuster, New York.

Thernstrom, S., Thernstrom, A., 1999. Reflections on the shape of the river. UCLA Law Review 46, 1583–1631.

Welch, S., Gruhl, J., 1998. Affirmative Action and Minority Enrollments in Medical and Law Schools. University of Michigan Press, Ann Arbor, MI.

Affordable Private Schools in Developing Countries

Pauline Dixon, Newcastle University, Newcastle Upon Tyne, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Affordable private schools have burgeoned in poor countries around the world. These schools provide education for the poor in urban and rural areas of developing countries. Children in the affordable schools typically outperform those attending government schools after controlling for selection bias, family background, and innate ability. This is at a fraction of the teacher cost of government schools. A new industry has set up around this market providing products specifically designed to improve quality and access. Aid agencies are now realizing the contribution affordable private schools can make to achieving quality education for all.

Affordable private schools provide education for children in poor urban and rural areas in developing countries (Tooley and Dixon, 2005b). Survey and census data gathered in slums and low-income areas around the world have revealed the presence of a large private school market that is providing education for the poor in countries such as Ghana, Nigeria, Kenya, South Sudan, Liberia, Sierra Leone, India, and Pakistan (Dixon, 2013). Research teams distinguish two types of fee paying affordable private schools. First, schools that operate under any official radar, functioning in an extralegal sector, which are termed unrecognized or unregistered. Secondly, schools that gain recognition from the government by 'supposedly' abiding by the 'on paper' regulations are known as recognized or registered. Although there are multiple rules and laws that have to be adhered to 'on paper,' these are often unattainable 'in practice' by those operating in poor areas owing to expense or unavailability of the requirement, for example, requiring a 4000-square meter playground in a cramped slum or teacher salaries being equivalent to that of government salary provision. The laws and rules generally do not target inputs that improve quality outcomes. In order to gain recognition status, which implies in most countries the school is able to run state examinations, owners pay unofficial payments, that is, bribes to educational officials, to gain recognition irrespective of whether the school meets the 'on paper' criteria (Dixon, 2003). The survey and censuses where researchers map the locations of affordable schools need to be carried out because unrecognized or unregistered schools do not appear on any of the official government or education department lists.

Affordable Fees

Affordable private schools have typically grown organically within the communities themselves; run by local entrepreneurs who have recognized a need by local children for education. These schools are accessible to the poor because the fees charged are affordable to those earning minimum wages. Families that send their children to these schools are typically employed as daily paid laborers, market traders, taxi drivers, construction workers, and fishermen or hold other manual labor positions. Parents may be illiterate even in their own language. In India, affordable private schools charge fees that amount to around 6% of minimum wage. In Africa it is generally higher, at an average of about 10–12%. These affordable private schools typically provide places for the 'poorest' in their communities by offering scholarships or concessionary places for those most in need (Tooley and Dixon, 2005b). The majority of schools are run for profit. Fees from parents are their main source of income. The government, NGOs, and religious organizations do not support this school management type.

Burgeoning Private School Sector

The number and proportion of affordable private schools is large in low-income areas and slum areas of Asia and Africa. In Pakistan, it is estimated that there are more than 47 000 private schools catering for about one-third of primary school-aged children (Andrabi et al., 2007). In urban settings of India affordable private schools can make up between 60 and 80% of school provision catering to around 65–70% of children. In rural India, some estimates imply that between 24 and 40% of children attend affordable private schools (Pratham, 2011). In Africa numbers are similar. In Lagos State, Nigeria, a large majority of children are attending these schools. In three districts of Lagos state, 540 schools were located of which 66% were affordable private schools. It is estimated that at least 75% of children in Lagos state are attending affordable private schools. The Ga region of Ghana is home to some half-a-million inhabitants of whom 70% live on or below the poverty line; 779 schools located during a census in Ga, indicated 75% being affordable private schools and 65% of children attending these schools (Tooley et al., 2005, 2007). Studies from India and Africa have shown more affordable unrecognized schools than government ones.

Longevity of Affordable Private Schools

These schools are not 'fly by night.' In Pakistan, such schools have been operating for more than 20 years where household surveys at the time found in Lahore a high proportion of children enrolled in private schools, even from the poorest families (Alderman et al., 1996). In Haryana, India, around

13 years ago affordable private schools provided poor parents with educational choices for their children (Aggarwal, 2000).

Teacher Salary and Qualifications

Teachers are paid less in the affordable private schools compared to teachers in the government sector. In India, teachers in affordable schools can be paid one quarter of the salary of a government school teacher. To teach in an affordable school, in practice, teachers need not have a teacher qualification or certificate. However, data have shown that in countries such as Pakistan and India the standard teacher resume characteristics have little bearing on student attainment (Aslam and Kingdon, 2007). What are more effective are teacher activity, retention of teaching staff, and the ability to fire those who are not effective.

English Medium in India

Parents in India choose the affordable private schools for their children because many purportedly teach all subjects in English. Most private schools in India are English medium, and parents believe that learning through English will help their children attain better opportunities after school with regard to employment and further education. Being able to communicate in English, parents believe will give their children a better chance to lift themselves out of poverty (Sen and Blatchford, 2001; Mitra et al., 2003).

Children's Achievement

Affordable private schools perform well compared to other school management types. Testing has been carried out with children in mathematics, English, and home language. Data have also been gathered on the pupils' family background, innate ability, as well as school and teacher characteristics in order to control for school choice, selection bias, and other variables that might affect achievement. Various analytical techniques have been used to analyze data, including multilevel modeling and the Heckman Lee procedure. In general, the results show that pupils in affordable private schools outperform those in government schools for a fraction of the teacher cost (Tooley et al., 2010, 2011; Pratham, 2011; Andrabi et al., 2010). In Pakistan, children from government schools perform 0.8–1.0 standard deviations lower on independently administered tests than do 'equivalent children in private schools'(Andrabi et al., 2010). Data from Pakistan show that children in government schools would need 1.5–2.5 years to catch up with children in affordable private schools (Andrabi et al., 2007).

Recent Developments

There have been several major developments in recent years in this space owing to the advancement in knowledge and information around the affordable private school's market. Currently, there are a handful of affordable school chains including the Omega Schools in Ghana, Bridge Schools in Kenya, and Empathy Learning Systems in India. The first Bridge school opened in 2009 and currently has around 70 schools in their network around Nairobi with a view to opening 400 in the chain. Omega schools opened their first school in 2010 and is a for-profit chain currently consisting of 30 schools charging a daily fee of 1 Ghanaian Cedi (around 60 cents). The daily fee allows much more flexibility for parents and includes lunch, uniform, books, and tuitions. Chains of schools benefit from economies of scale and are able to work on curriculum and pedagogy development as well as in devising and evaluating different methods of testing and assessment. Parents also know what they will receive from the 'brand.' The education company Pearson who also own the Financial Times and Penguin books has recognized the potential in chains of affordable private schools. Pearson have launched a fund called the Pearson Affordable Learning Fund which is investing in and providing support to organizations that focus on education solutions for the poor. Impact investors have now recognized the potential in the affordable private school's market. The supply of capital to improve quality and expedite expansion can now be accessed through companies such as Gray Ghost Ventures, Edify, Opportunity International, and the IDP Foundation. All are now investing in the affordable private school space. International aid agencies such as the Department for International Development (DFID) in the United Kingdom are now funding projects that focus on affordable private schools by improving their quality as well as access. Charities such as Absolute Return for Kids (ARK) in London are funding quality inputs as well as targeted vouchers and associated randomized control trials to augment knowledge around the effectiveness and potential of affordable private schools.

See also: Equity and Education; International Organisations and Education; Parents and Teachers as Partners in Education; School Choice; School Effectiveness Research.

Bibliography

Andrabi, Tahir, Das, Jishnu, Khwaja, Asim Ijaz, Vishwanath, Tara, Zajonc, Tristan, The LEAPS Team, February 20, 2007. PAKISTAN: learning and educational achievements in Punjab schools (LEAPS): insights to inform the education policy debate. Executive Summary vi.

Alderman, Harold, Orazem, Peter, Paterno, Elizabeth M., 1996. School Quality, School Cost and the Public/Private School Choices of Low-income Households in Pakistan. Working Paper Series on 'Impact Evaluation of Education Reforms' Paper No. 2. World Bank, Washington, DC.

Aggarwal, Yash, 2000. Public and Private Partnership in Primary Education in India: A Study of Unrecognised Schools in Haryana. National Institute of Educational Planning and Administration, New Delhi.

Aslam, Monazza, Kingdon, Geeta, 2007. What Can Teachers Do to Raise Pupil Achievement? CSAE. Working Paper Series.

Andrabi, Tahir, Bau, Natalie, Das, Jishnu, Khwaja, Asim Ijaz, 2010. Are Bad Public Schools Public "Bads"? Test Scores and Civic Values in Public and Private Schools. Working paper cited with permission from J. Das.

Dixon, Pauline, 2003. The Regulation of Private schools for Low-Income Families in Andhra Pradesh, India: An Austrian Economic Approach. Unpublished PhD. Newcastle University.

Dixon, Pauline, 2013. International Aid and Private Schools for the Poor: Smiles, Miracles and Markets. Edward Elgar, Northampton, MA.

Mitra, Sugata, Tooley, James, Inamdar, Parimala, Dixon, Pauline, 2003. Improving English pronunciation: an automated instructional approach. Information Technologies and International Development 1 (1), 75–84.

Pratham, 2011. Annual Status of Education Report (Rural) 2010. Provisional. Pratham Resource Centre, Mumbai, New Delhi.

Sen, Reena, Blatchford, Peter, 2001. Reading in a second language: factors associated with progress in young children. Educational Psychology 21 (2), 189–202.

Tooley, James, Dixon, Pauline, 2005a. Private Education Is Good for the Poor – A Study of Private Schools Serving the Poor in Low-income Countries. CATO Institute, Washington, DC.

Tooley, James, Dixon, Pauline, 2005b. Is there a conflict between commercial gain and concern for the poor? Evidence from private schools for the poor in India and Nigeria. Economic Affairs 25 (2), 20–26.

Tooley, James, Dixon, Pauline, Olaniyan, Lanre, 2005. Private and public schooling in low-income areas of Lagos State, Nigeria: a census and comparative survey. International Journal of Educational Research 43 (3), 125–146.

Tooley, James, Dixon, Pauline, Amuah, Isaac, 2007. Private and public schooling in Ghana: a census and comparative survey. International Review of Education 53 (4), 389–415.

Tooley, James, Bao, Yan, Dixon, Pauline, Merrifield, John, 2011. School choice and academic performance: some evidence from developing countries. Journal of School Choice: Research, Theory, and Reform 5 (1), 1–39.

Tooley, James, Dixon, Pauline, Shamsan, Yarim, Schagen, Ian, 2010. The relative quality and cost-effectiveness of private and public schools for low-income families: a case study in a developing country. School Effectiveness and School Improvement 21 (2), 117–144.

Relevant Websites

http://ccsindia.org/ccsindia/index.asp.
http://educationnext.org.
http://egwestcentre.com.
http://www.graymatterscap.com.
http://www.omega-schools.com.
http://www.periglobal.org.
http://www.youtube.com/watch?v=gzv4nBoXoZc.

African Legal Systems

Gordon R Woodman, University of Birmingham, Edgbaston, Birmingham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Many customary and religious legal systems were effective in African societies before colonization. The colonial powers added legal systems based on their own laws. Both types have continued to be effective since Independence and the introduction of written constitutions. Post-Independence state laws have been developed to strengthen state institutions and to promote nation-building. Post-Independence customary laws, observed by populations smaller than those of states but sometimes stretching across state boundaries, have adapted to modern circumstances. They are widely observed in the fields of family law, land law, and dispute resolution procedures, and state laws give them varying degrees of recognition.

Terminology

'African legal systems' means, for the purposes of this article, the bodies of interrelated legal norms and accompanying institutions of norm-creation, norm-finding, and norm-enforcement, which have a social existence in Africa. 'Legal norms' are taken to be social norms which are enforced by relatively strong degrees of coercion.

Customary and Religious (Nonstate) Legal Systems

Customary legal systems exist by virtue of the social observance of their norms, and not by the creation of their norms through state institutional processes, such as the enactment of legislation. Customary legal norms are observed because of a continuing but usually tacit understanding within a population that the bodies of norms to which they belong are to be accepted as obligatory. For the present purpose a practice need not have been observed for a very long period to be 'customary'.

Religious legal systems are for the present purpose a variety of customary legal systems, their distinctive feature being their derivation from systems of religious belief. In Africa the primary systems of religious law are varieties of Islamic law. In many communities, especially in North Africa and the Sahel region, the predominant law is *shari'a*, sometimes modified under the influence of other local customary norms (*see* Islamic Law).

The subjects of an African customary legal system (that is, the population that observes a customary law) virtually never coincides with the population that is subject to a particular state law. The population subject to a state law normally contains populations or portions of populations which observe different customary laws, while many groups of subjects of particular customary laws are divided into subjects of two or even more state laws.

Nonstate African legal systems vary widely, and almost any generalization about them is subject to exceptions. Perhaps the most distinctive feature of African customary legal systems is the frequency with which the parties to legal relations are communities of ascribed membership, especially of membership by descent, not individuals or other corporate persons as in western legal systems. Thus, for example, substantial interests in land are often vested in lineages or in communities comprised of several lineages. Marriage is often contracted by agreement between the lineages to which the bride and groom belong, with the consent of the bride and groom themselves, and creates legal relations between the lineages.

A customary law community often has an individual leader, instances of which range from the 'head of family' of the relatively small lineage to the 'chief' or 'king' of a large polity, such as Asante or Buganda, although there are also many acephalous communities. Leaders are determined according to customary legal norms. Although in some circumstances a charismatic leader may temporarily acquire personal authority and wield discretionary power, generally customary law imposes strict limits on the powers of leaders: the rule of customary law prevails widely.

Another common feature of customary legal systems is the nature of the procedures and principles governing the settlement of conflicts of interests and disputes. These are directed to the achievement of social peace and harmony within the community rather than the determination of legal rights. Dispute processes are less frequently instances of adjudication than of mediation or negotiation, in which there are social pressures on the parties to compromise.

State Legal Systems

The modern African state with its governmental and legal institutions is a product of colonization, with the possible exception of Ethiopia. The colonizing powers drew up by agreement between themselves the territorial boundaries of their African colonies, and these boundaries very frequently divided the territories occupied by populations observing particular customary laws. In most parts of Africa there was relatively little immigration from the colonizing countries, and everywhere the indigenous inhabitants remained in the majority. Nevertheless the colonial powers set up systems of government that resembled those of the metropolitan states, except that they did not provide for extensive popular choice of governments or other participation in government. The English common law was imported into the British colonies. Codes largely identical with those of France and Portugal were enacted in their colonies (Allott, 1960, 1970; *see* Sub-Saharan African Legal Traditions).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10111-4

The received laws shaped the institutions of the state legal systems, such as the courts, and provided bodies of law in fields, such as crime, contract, property transactions, and personal injuries. The technical and social knowledge necessary to administer this law in such roles as those of judge, legal advisor, legislator, and police officer were initially lacking in the indigenous populations. The lacuna was filled by colonial officers brought from the metropolis, some of them members of the professional legal community there, and others with some training and socialization in the ways of that community. Members of indigenous communities later obtained the training and assumed the social practices that enabled them to take over these roles. Thus the laws of the colonial powers became African legal systems. At Independence (achieved in most states between the late 1950s and the end of the 1970s) African rulers and administrators saw their aspirations and material interests as dependent on the continued effectiveness of these legal systems.

Those customary and religious institutions and practices that the colonial powers perceived to be inconsistent with colonial domination, such as primary allegiance to chiefs, they attempted to destroy or transform. Their policy toward other components of these indigenous legal systems varied, but often was to tolerate and sometimes even to encourage their continuance as alternatives to those of received law. Whatever the colonial state policy, customary laws continued to be observed concurrently with state laws. In this form of legal pluralism there was no unified hierarchy of norms. Individuals differed according to whether they gave general priority to the observance of state law rather than a nonstate law, or, gave priority to each on different occasions.

Within state law a further type of legal pluralism was formed. State legal systems gave recognition to African indigenous customary and religious legal systems. 'Recognition' here designates the policy whereby state legal systems treat the institutions or norms of customary and religious law as parts of the law of the state, giving effect to them and enforcing them in the same way as the institutions and norms of received law. Policies of recognition were adopted in some British colonies from the beginning, and eventually adopted everywhere. The forms of recognition can be analyzed as falling into two categories.

Institutional recognition of a customary or religious legal system occurred when institutions of that system were incorporated into a state legal system, as for example, when chiefs became administrative officials or judges of the state. This frequently happened in British colonies, in accordance with the policy of indirect rule which sought to govern colonies through native forms of government. With the abandonment of the policy from the 1940s this recognition became less significant, but it continued in a limited form. In other colonies it was adopted as a concession to local opinion, as well as for practical reasons. In all instances it has continued since Independence.

Normative recognition of a customary or religious legal system occurred when norms of that system were incorporated into the body of norms of the state legal system. State law might provide, for example, that rights in land could be transferred by procedures specified by norms of customary law, or that the inheritance of property might be determined by customary and religious norms. The incorporated norms then became enforceable in state courts. This also occurred in most African colonial legal systems, and also has continued to the present.

Customary and religious laws were not and indeed could not be 'incorporated' into state law without radical changes to their nature and content. The state has always excluded from incorporation some portions of customary laws, such as those providing for slavery, which came to be viewed as contrary to the fundamental values of the state once the European states themselves had abandoned the practices. Moreover, the acceptable elements of customary and religious legal systems could not and cannot be accommodated without reformulation. The institution of chieftaincy, for example, is transformed when its authority ceases to be derived from respect for tradition and communal identity, and becomes based upon the threat of coercion by state institutions. The forms of coercion of state institutions, such as the threat of imprisonment supporting an adjudicative decision, differ from the social pressure traditionally exerted upon disputing parties to agree to a compromise. Consequently the effect of normative recognition has generally been to compel behavior of a different type from that which would otherwise have occurred. Finally, the personnel of state institutions are frequently unfamiliar with customary or religious law and have 'recognized' norms which differ from those socially observed. Thus state institutions create new bodies of norms, 'official', 'lawyers', or 'judicial customary laws', which differ from the 'living', 'people's', 'folk', 'indigenous', or 'practiced customary laws' which have continued to be observed outside state administrations. Some legal historians consider lawyers' customary law to be an 'invention', bearing no significant relationship to preexisting social norms (e.g., Chanock, 1985; cf e.g., Woodman, 1985; Bennett, 2004).

Past and Present Trends in Legal Development

Legal development since the institution of the colonial state has reflected the considerable political and social changes which have taken place in Africa.

Changes in state public law, that is, those branches of the state legal system which regulate government, have generally been directed toward strengthening the state and elaborating the functions and powers of state executive bodies. From various times, the earliest being in the late 1940s, the state public laws have also provided for increasing public participation in government. The Independence constitutions contained rules for democratic and constitutional government. In the immediately following decades many states instituted one-party government, or experienced the violent overthrow of their constitutions followed by military rule. More recently internal developments and external, globalizing pressures have produced in most states trends toward constitutionalism, more democracy and accountability in government, and emphasis on environmental protection. These forces, and notably the formation of the African Charter on Human and People's Rights of 1986, have produced constitutional recognition of a wide range of human rights through judicially enforceable legal provisions.

Nation-building has been a major objective of constitutional orders. The boundaries of modern African states, inherited from the colonial powers, contain different ethnic and religious groups, each with its own culture, customary law and language. These boundaries have in very few instances been changed in post-Independence Africa. However, the individual citizen's

loyalty and sense of belonging are often directed primarily to the ethnic or religious group to which they belong. Constitutional provisions, in an effort to create a sense of national identity, have prohibited discrimination on the ground of ethnicity, language, or religion in many activities, including the selection of members of state institutions and political organizations. This has been seen as necessary to economic and social development, as has the principle of indigenization of positions of economic or social power. There have been increasingly determined attempts to improve the legal position of women, who are seen as seriously disadvantaged by traditional laws and practices, and of children. However, the normative recognition given by state law to customary laws in the colonial period has generally been continued, and in some instances even confirmed by constitutional provisions. In contrast, recognition in state criminal law of customary law has never been extensive, and today the state criminal law is usually contained exclusively in state criminal codes, together with other written laws specifying minor offences.

Developments in public law have generally increased the functions of state law as a regulator of social behavior. All of the trends just noted, and the long-term tendency toward western-style governmental systems have been accompanied by a decline in the African state's use of customary institutions, such as chieftaincy, and so of the strength of these institutions. Nevertheless, the state is still not the sole, unchallenged legal authority. It remains notoriously ineffective in regulating the daily lives of most citizens outside its own bureaucratic institutions. Even within these institutions, instances of corruption and abuse of power show that state officials are not punctilious observers of state law. One global policy may also have reduced the scope of state law. The structural adjustment policies which the World Bank and International Monetary Fund set in the late twentieth century as conditions for economic aid tended to reduce state involvement in economic activity in favor of private business institutions, thus strengthening against the state law forces in the international market and global political influences.

In the fields of private law, changes are visible in the content of received law, living customary law, and lawyers' customary law.

Received private law has been continuously changing both in the European countries of origin and in African states. State laws required received law to be adapted to local circumstances by the courts and legislatures. But decision makers have also asserted the value of transnational uniformity in the principles of the common law and civil law, respectively. Decisions and textbooks from the countries of origin of received laws have been followed in African jurisdictions. Even in the legislative reform of received law the best legislative model is sometimes held to be the latest legislation in the country of origin, in fields as diverse as company and divorce law. On the other hand, in legislation concerned with economic development, such as that regulating foreign investment, there has there been significant innovation.

The extent to which received law is observed has increased as a result of some aspects of social change. Policies of 'modernization' have been accompanied by a general official belief that the application of Western laws gives effect to the principles of the market economy (*see* Law and Development). To the same effect has been the growing involvement in the global economy of private individuals and businesses through the export-oriented development of cash crops and extractive industries. Education and technology have been taken from the West as well as being locally developed. The same tendencies have produced and are in turn reinforced by the emergence and growth of national legal professions. Nevertheless, in activities such as the contracting of marriage, the acquisition and use of lineage assets, and land use, most people still choose to act under customary law.

Systems of living customary law have never been static, and have changed and developed to meet new circumstances, first in the colonial periods, and then since Independence. The individual has gained more autonomy, having today the capacity to hold extensive property and enter into weighty transactions independently of their lineage. Legal transactions are today more often determined by market forces, rather than by standard customary terms. Written records are more commonly used, and more norms have been developed to guide relations between members of communities and outsiders. New bodies of customary law have been formed, especially in the vastly expanded urban areas where religious, professional, welfare, and self-help communities with their own customary laws have been created. The primary fields in which living customary laws receive significant observance are those of family law, property (especially land) law, and processes of dispute resolution.

Lawyers' customary law has also changed, in that norms which the state will recognize and enforce as customary law have been progressively elucidated and embodied in judicial decisions, restatements, and textbooks. To some degree the reformulation entailed in these processes has been in the same direction as changes in living customary laws. However, the processes have also had a conservative effect, because a norm embodied in an authoritative pronouncement is not easily changed thereafter in response to changes in social circumstances. The gap between lawyers' customary law and living customary law could thus grow, although measures are being taken in some jurisdictions to avert this trend by developing lawyers' customary law to be a more accurate reflection of living customary law. Another major development in lawyers' customary law has been the extensive amendment to some areas by legislation. Land laws in particular have been amended in attempts to balance competing claims to land use while also promoting economic development and inhibiting environmental damage. Here the extent to which living customary law diverges from state law is a function of the extent to which the state is able to make its legislation effective.

Current and Future Issues in Legal Development, Theory, and Research

Current lawmaking by states is directed toward a number of practical problems. There is a search for constitutional forms which will maintain stability against interethnic and interreligious conflict, and other forms of civil disorder. In some states the relationship between customary and religious legal systems and state law has become an even more acute issue since the constitutional entrenchment of human rights, with which some nonstate laws appear incompatible. Globalization intensifies certain issues. State law may be developed to protect national economic independence against the power of international

businesses and institutions, such as the World Trade Organization. State law may also introduce institutions which can safeguard and assist economic development in the face of new threats from natural and artificial disasters, such as climate change and the AIDS epidemic.

In the past few years international development agencies have increasingly fostered research and discussion to determine whether customary laws should be treated as factors in development, and taken into account in adjusting and improving development programs, rather than simply obstacles to development (Tamanaha et al., 2012).

Research into African legal systems has in the past been carried on within several disciplines. The study of customary law has used the methods of anthropology and sociology, of state law those of legal analysis, and of Islamic law those established specifically within that scholarly tradition. Today there is more interdisciplinary understanding. One consequence is renewed concern with fundamental questions about the concept of law, especially that as to whether customary and religious normative orders are truly law. These debates may also raise doubts as to whether it is appropriate to speak of 'systems' of law.

With indigenization and expansion of the African civil service and academia, most research into African legal systems will in future be conducted in Africa by African scholars. It will be closely concerned with the practical problems just mentioned. However, it is likely that the more fundamental theoretical issues will continue to inspire and to be investigated by an international community of scholars.

See also: African Studies: Culture; African Studies: History; Development and the State; Disputes, Social Construction and Transformation of; Globalization: Legal Aspects; Islamic Law; Law and Development; Law as an Instrument of Social Change; Law, Mobilization of; Law: Anthropological Aspects; Legal Culture and Legal Consciousness; Legal Pluralism; Legal Systems, Classification of; Postcolonial Law; Sub-Saharan African Legal Traditions.

Bibliography

Allott, A.N., 1960. Essays in African Law: With Special Reference to the Law of Ghana. Butterworths, London.

Allott, A.N., 1970. New Essays in African Law. Butterworths, London.

Bennett, T.W., 2004. Customary Law in South Africa. Kluwer, The Hague.

Chanock, M., 1985. Law, Custom and Social Order: The Colonial Experience in Malawi and Zambia. Cambridge University Press, Cambridge.

Doucet, M., Vanderlinden, J. (Eds.), 1994. La réception des systèmes juridiques: implantation et destin (The Reception of Legal Systems: Implantation and Development). Bruylant, Brussels.

Elias, T.O., 1956. The Nature of African Customary Law. Manchester University Press, Manchester.

Fenrich, J., Galizzi, P., Higgins, T. (Eds.), 2011. The Future of African Customary Law. Cambridge University Press, New York.

Gyandoh, S.O. (Ed.), 1990. Third World Legal Studies: Building Constitutional Orders in Sub-Saharan Africa. International Center for Law Development, New York.

Journal of African Law, 1957. Cambridge University Press for the School of Oriental and African Studies. University of London, London.

Journal of Legal Pluralism and Unofficial Law, 1969. Routledge, London.

Reyntjens, F. (Ed.), 1989. Pluralism Participation and Decentralization in Sub-Saharan Africa. Third World Legal Studies, pp. vii–x.

Tamanaha, B.Z., Sage, C., Woolcock, M. (Eds.), 2012. Legal Pluralism and Development: Scholars and Practitioners in Dialogue. Cambridge University Press, Cambridge, pp. 83–94.

Van Rouveroy van Nieuwaal, E.A.B., Ray, D.I. (Eds.), 1996. The New Relevance of Traditional Authorities to Africa's Future. Journal of Legal Pluralism 37/38, 1–38 (special double issue).

Vanderlinden, J., 1983. Les Systèmes Juridiques Africains (African Legal Systems). Presses Universitaires de France, Paris.

Woodman, G.R., 1985. Customary law, state courts, and the notion of institutionalization of norms in Ghana and Nigeria. In: Allott, A., Woodman, G.R. (Eds.), People's Law and State Law: The Bellagio Papers. Foris, Dordrecht, pp. 143–163.

Woodman, G.R., Obilade, A.O. (Eds.), 1995. African Law and Legal Theory. Dartmouth Publishing, Aldershot, p. 474.

African Studies: Culture

Jane I Guyer, Johns Hopkins University, Baltimore, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of African culture follows four phases during each of which the focus, basic material, and interpretive framework has been quite different. Within and between these phases there has been a great deal of contention: in part because scholarship was dominated for a long time by Europeans rather than Africans, and in part because the sources are so vast and so relatively difficult to study. In the earliest phase, before the era of academic disciplines, missionary writings, and travelers' writings developed a contrastive model, where Africa was the antithesis of Europe. The first systematic work was based on a natural historical method of study, involving the collection and classification of cultural items, particularly material and artistic ones. The anthropological approach to culture as the entire lifeworld of a community, integrated by key propositions that reflected the social structure, succeeded the previous approach in the 1920s. Research involved field studies of ideas and materials in use in social life, and thereby moved away from the arts, music, and technical inventories toward the ideas that underlie and give meaning to them. Since the 1960s, study of a current and expanding creative production that relates to the present joys and dilemmas of life has replaced the traditional 'worldview' approach to culture. A great flowering of scholarship has been possible on the basis of new technologies of recording and a broad approach to culture itself. The historical connections between Africa and the rest of the world have replaced the conceptual separations, informing a new scholarship on diasporas. Most recently even the icons of traditionality are being restudied in light of the influences of global history.

The study of African culture follows four phases during each of which the focus, basic material, and interpretive framework has been quite different. Within and between these phases there has been a great deal of contention: in part because scholarship was dominated for a long time by Europeans rather than Africans, and in part because the sources are so vast and so relatively difficult to study. In the earliest phase, before the era of academic disciplines, missionary writings and travelers' writings developed a contrastive model, where Africa was the antithesis of Europe. The first systematic work was based on a natural historical method of study, involving the collection and classification of cultural items, particularly material and artistic ones. The anthropological approach to culture as the entire lifeworld of a community, integrated by key propositions that reflected the social structure, succeeded the previous approach in the 1920s. Research involved field studies of ideas and materials in use in social life, and thereby moved away from the arts, music, and technical inventories toward the ideas that underlie and give meaning to them. Since the 1960s, study of a current and expanding creative production that relates to the present joys and dilemmas of life has replaced the traditional 'world view' approach to culture. A great flowering of scholarship has been possible on the basis of new technologies of recording and a broad approach to culture itself. The historical connections between Africa and the rest of the world have replaced the conceptual separations, informing a new scholarship on diasporas. Most recently even the icons of traditionality are being restudied in light of the influences of global history.

Introduction

The concept of culture is one of the most disputed in the social sciences, and nowhere more so than in African studies. The classic study of culture rested on the study of the main media of expression: music, art, material forms, and (above all) language and the written expression of thought in literature, philosophy, and theology. Texts were central. In the foundational era of African Studies, in the nineteenth century in Europe, Africa was assumed to have no texts at all. Its arts of music, performance, and oral poetics were thought to be different in kind from text-based culture. They were performative and multimedia rather than abstract and specialized. Hence their designation as 'primitive', which survives in the occasional museum and art gallery. The basic collections were, therefore, of sculptural art and material culture, which were relatively accessible even to amateur collectors, and which were assembled during the competitive rush for acquisitions by museums in the late nineteenth and early twentieth centuries. Even for specialists in the study of 'primitive cultures', however, techniques for reproducing their non-written cultural forms – such as tape-recording, photography, and film – were still too clumsy and fragile at that time to produce a solid and reliable corpus for study. The analysis of African cultures has grown and changed, therefore, with changes in thinking about culture in general, in the technologies of scholarship, and in the place of Africa in the world. These changes are discussed in general and then their enduring contributions are summarized.

Historical Overview

Since the late nineteenth century in the systematic study of African culture, scholars and artists have struggled to rework the foundational legacy as new study techniques, approaches, and colleagues appeared on the scene. Three major changes have erased the categorical distinction between literate and oral/performative culture. First, new recording technologies have made performative culture amenable to the kinds of analysis given to texts in the past, and have made primary sources more widely available for secondary analyses. Second, and at the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10112-6

same time, modern cultural studies in Europe has moved closer to African studies by embracing a much wider range of arts and performance than in the past, including multimedia work, and by looking at how different media can draw on the same forms and themes. Still using textual methods, they now include popular culture, film, clothing, and other expressive forms, and they link style, substance, and concepts of authorship to social context far more than was done in the classical mode. Finally, the African library of written texts has expanded greatly in the twentieth century. Far more written forms from the past are now recognized than were acknowledged in the classic period, including voluminous sources written in Arabic script and certain sculptural, geometric, and musical forms that are now realized to encode verbal content (e.g., Roberts and Roberts, 1996). The other source is current artistic production. There is a new corpus of literature, music, and the arts that is not 'traditional', but rather responds to the situation of modern life, including the global marketplace. The new expansion in Africa includes African scholars' studies of their own past and present art forms, working from their own languages, and this has contributed new expertise and a new set of ideas to the international arena.

These changes from the classic, classificatory, study of culture to the current expansive incorporation of varied and linked forms are still being worked out and they are still contentious. At the same time, the sheer enormity of the task of adequate study of African cultures is realized increasingly. Only in the European view of the confident, progressive industrial era could the whole continent and its millennia of history be grouped together into a single category as 'Africa'. Africa is the cradle of humanity; its cultures have very deep histories; its population is also highly diverse. There are about 1000 African languages spoken. More than was originally understood also benefit from written works. Even the popular icons of Africa that have been studied for decades – such as masks, body decoration, 'fetishes', and drumming – are far richer as traditions than the scholarship can yet do justice to.

With these provisos about the completeness of the record, the history of cultural study can be divided schematically into four broad successive phases. The first comprises the early missionary and traveler documentation, up to the mid-nineteenth century. The first systematic efforts at documentation (a second phase, from the late nineteenth century up to World War I) were carried out in the natural history style. Nonliterate cultures were studied by collecting cultural items, and describing and classifying them according to the taxonomy of the collector. The third phase was anthropological, and lasted more or less unchallenged from about 1920 to about 1960 (the time of political independence from European colonial rule). It was devoted to the idea of cultures as coherent systems of thought and representation, where all the elements of the life ways of a community (judged either by their common language or common membership in a political community) could be related to each other. The fourth and present phase (from about 1960 and continuing) illuminates cultural themes and techniques by drawing on the classic and modern humanities. Here culture is seen from within, as a field of imagination and debate. There have been major achievements in all four phases, each of which has posed key questions for comparative knowledge, general theory, and public debate.

Phases in African Cultural Studies

Missionary Understandings and Older Sources

The idea that Africa and Europe 'framed' the human experience – initiated and culminated it, expressed two ends of a spectrum of biological and cultural experience – predates the disciplinary study of the continent and has been extremely persistent in Western thought. Evidence about Africa from before the nineteenth century is both sparse and problematic. Nevertheless, these sources set agendas. Writings from that period, by Europeans in their own languages, are now seen as having 'invented' (Mudimbe, 1988) an Africa that is fundamentally contrastive with the West. This conceptualization still informs the general Western lexicon about Africa, but it has been repeatedly challenged by scholars.

African philosophers and scholars outside of the European mainstream have insisted on a return to older sources, to question the basis for the contrastive model. Most prominent has been the Senegalese thinker Diop (1955), who argued for the Egyptian origins, pan-African connections, and historical unity of African cultures, as against a 'primitive' and racialized conception. Painstaking scholarship in several disciplines continues, linking evidence from the old sources to new findings on, for example, the long history of urban cultures, monies, and trade, the use of Arabic script, and fractal geometric and other mathematical concepts in African design and divination (Eglash, 1999). Archaeology and paleontology examine yet earlier traces of cultural life, not only through fossil remains (for the very distant past) but also in rock paintings, house and burial sites, the remains of iron and copper works, and so on.

It was, however, the prevailing contrastive model that was pushed in new directions by the first disciplinary scholarship about African culture.

Culture and Natural History

A passion for creating encompassing classifications of all of life gained momentum in European intellectual circles during the nineteenth century. Other cultures and their attributes were seen as part of the phenomenal world, to be treated similarly to other life forms. The actual process of collecting involved human and political encounters of a depth, complexity, and problematic nature that is still being studied (Fabian, 2000). Some objects were seized during the punitive expeditions of colonial conquest from the 1880s until about 1910; some were accumulated during Christian conversion and the outlawing of the ritual cycles in which they were used; others were acquired by organized museum expeditions, where the items were bargained over and paid for. Hundreds of thousands of pieces – some of great artistic value in the West and religious value to the populations themselves, and others of technical interest only – arrived by one or other of these routes in American and European museums, where the vast majority remain. Some pieces have been missed gravely; one or two have been returned. To their lasting credit, however, some of the collectors were attuned to the breadth and completeness of their collections and to the originality of the material so that they remain a spectacular resource for scholars from all nations and fields. Some collectors also kept acquisition records, which can still be

gleaned for contextual understanding of both the collecting and the objects themselves (Schildkrout and Keim, 1998). And a few, such as the missionaries George Schwab (1947) with George Harley, wrote books about the cultural lives of the people who created and used the objects.

The emphasis was on material culture, with some notable photography of items in use and of daily life in general. It would be hard to exaggerate the amount and the variety of the African material culture in Western museums: everything from divination trays to fish traps, enough spears to arm small militias, carved spoons, blacksmiths' tools, indigenous currencies, cloth, talismans, and 'fetishes'. Culture, in the natural history mode of thinking, was everything that people produced by virtue of learning, from the most intricately symbolic to the most pragmatically functional. The very large number of works that could be classified as 'art' had a further destination than the natural history museum. Some went into an art market that made its own selective choices of value. There are African works from this period that are now considered masterpieces, valued in the hundreds of thousands of dollars. The asymmetric geometry, the nonnatural figurative representations, the brilliant colors, and the general vitality of these pieces were a major influence on modernism in European art. The power of modernism, in turn, fixed these particular pieces in the mind of the public as icons of African aesthetics.

The recognizability of African icons has turned out to be a very mixed legacy because it has made innovation more of a struggle. During the museum era, the icon of African music became drumming, of art became figurative sculpture and especially masks, of architecture became what is still referred to in popular parlance as the 'hut', of body decoration, scarification, and so on. The following phase of cultural study involved an effort to break out of this straightjacket, and above all to place the elements back into the context of life-as-lived. Culture was not just a product of learning, but a means of articulating a world view.

Culture in Anthropology

The era between World War I and political independence, about 1920–60, was the period of effective colonial rule in Africa. During this time, conditions were settled enough for scientists to stay for long phases of their careers, in some places setting up permanent research institutions. By this time, anthropology had become committed to studying societies and cultures as wholes. Where political and cultural lives were lived through oral and performative media, this meant long-term residence with them: to collect the texts, case studies, genealogies, maps, calendrics, and so on from which such an integrated model could be deduced. Political and medical conditions allowed, and intellectual aspirations promoted, the practice of field research, usually carried out for at least a year.

If cultures are conceptualized as integrated wholes then the key to their understanding must be the general concepts that animate and link the separate elements, domains, and performances. So this era of cultural study moved away from the objects, to place greater emphasis on ideas, symbols, and general principles. Particular emphasis was given to the ideas that motivate social life rather than those expressing aesthetic values: the concepts underlying kinship and political identity, religious symbols and practices, and philosophies of the composition of the lifeworld. In fact, cultural studies in European scholarship about Africa moved away from materiality altogether. The key works were about concepts of causation (Evans-Pritchard's eternal classic entitled *Witchcraft, Oracles and Magic among the Azande*, 1937), symbolic power (Turner, 1968), kin relations (Fortes, 1945), cosmology (Griaule, 1948), and philosophy (Tempels, 1946). Scholars of Africa, such as Douglas (1966) made major contributions to general culture theory, by linking conceptions of what is pure and what is dangerous to specific ambiguities in social structures. In a sense, this period replaced a notion that African cultural 'genius' lay in the figurative arts with an 'African Genius' (Davidson, 1969) that lay in the capacity to organize complex social life without the state domination characteristic of Europe. African cultural studies illuminated the study of society in general.

Culture Today

The rebellion against this approach, both within anthropology and beyond, was a rejection of the implication that any cultures could be studied as if they were 'traditional', in the sense of closed, unchanging, and unchallenged from within. The arts resumed center stage, not as iconic objects but as active creativity: emergent novelty, specific named authorship, audience reception, and constant revision and re-creation. This was culture as a field of imagination and its students included its creators: the new novelists and artists of independent Africa, commenting on the present. The humanities became prominent again. Greenberg (1972) suggested a highly persuasive language classification, which changed the history of the modern peopling of the continent. Vansina (1961) innovated in historiography by promoting oral history, which helped to open up a new effort on orature in general. Scholars of their own languages, such as Abimbola (1976) of Yoruba, Kunene (1979) of Zulu, and Ngugi wa Thiong'o (1986) of Gikuyu, worked on their own language creativity. Novelists, poets, and filmmakers – such as Achebe (1958), Soyinka (1964), and Sembene (1960) – used modern media to express the challenges of modernity as well as writing very influential critical works about culture and national life. New popular art forms, such as juju and highlife music, the Yoruba traveling theater, textile fashion, and hairdressing animated local engagement with the future and became subjects of study in the academy. Hountondji's (1977) attack on 'ethnophilosophy' both opened a new era of African philosophy and symbolized the break in cultural studies from the community-based traditionality of earlier anthropology. The center of gravity moved away from communities and toward the study of expertise, innovation, and urban life.

In the 1990s, the cultural scene in Africa became yet more diverse, including new architecture, jazz, video, and new versions of Christian thought and expression. The vast African diasporas in the Americas and elsewhere, from both the era of the slave trade and new global mobility, are now considered to be part of an African cultural ecumene. In turn, African history has been revisited to show how long and how widely Africa has been connected to the rest of the world, with lasting effects on many aspects of life – from cultigens to religious

expression (Blier, 1995) – that were once thought to express the very 'nature' of Africanity. These social geographical observations are matched by a new appreciation of the openness of cultures within Africa to novelty and originality, whether developed from within or borrowed from without. In a world era when dynamism is in full play, among ordinary people as well as cultural elites, this makes the study of Africa's long history, popular ebullience, and varied multiplicity the unique contribution of African Studies to the study of culture: to an understanding of cultural hybridity, memory, power, and the importance of local texts (newspapers, diaries, early novels, and records) and the vernacular popular arts (Barber, 2007)

Diaspora Studies have been pioneered by peoples of African descent: in Brazil, the Caribbean, Europe, and undoubtedly in a growing crescendo elsewhere in the world (Matory, 2006). The 'retentions' and reconstitutions of African culture by the continent's descendants who had been utterly stripped of every object, every community relationship, and any common African language during the era of the slave trade has remained a challenge to understanding. Herskovits's (1941) pioneering work in the 1930s stood for a long time at a tangent to cultural scholarship, which remained continent-based. The return of African-American cultural forms, such as hip-hop (Shipley, 2013), to present-day Africa became another theme. As South Africa's arts opened up to world appreciation, the designs for life and the market among the ordinary people took on a new profile (Fraser, 2002). Politically and intellectually, global attention to Pan-Africanist political and social thinkers, such as Fanon (Gibson, 1999), grew. In the 1990s, these themes were reopened, and in the process African cultural studies became trailblazing in the study of global networks in artistic and intellectual life, and of memory and continuing creativity in diaspora cultures.

See also: African Studies: History; Central Africa: Sociocultural Aspects; Colonialism, Anthropology of; Diaspora; East Africa: Sociocultural Aspects; Multiculturalism, Anthropology of; South America: Sociocultural Aspects.

Bibliography

Abimbola, W., 1976. IFA: An Exposition of Ifá Literary Corpus. Oxford University Press Nigeria, Ibadan, Nigeria.

Achebe, C., 1958. Things Fall Apart. Heinemann, London.

Barber, K., 2007. The Anthropology of Texts, Persons and Publics: Oral and Written Culture in Africa and beyond. Cambridge University Press, Cambridge.

Blier, S.P., 1995. African Vodun: Art, Psychology and Power. University of Chicago Press, Chicago.

Davidson, B., 1969. The African Genius: An Introduction to African Cultural and Social History. Little and Brown, Boston, MA.

Diop, C.A., 1955. Nations nègres et culture. Éditions Africaines, Paris.

Douglas, M., 1966. Purity and Danger; An Analysis of Concepts of Pollution and Taboo. Routledge and Kegan Paul, London.

Eglash, R., 1999. African Fractals: Modern Computing and Indigenous Design. Rutgers University Press, Piscataway.

Evans-Pritchard, E.E., 1937. Witchcraft, Oracles and Magic among the Azande. Clarendon Press, Oxford, UK.

Fabian, J., 2000. Out of Our Minds: Reason and Madness in the Exploration of Central Africa. University of California Press, Berkeley, CA.

Fortes, M., 1945. The Dynamics of Clanship Among the Tallensi; Being the First Part of an Analysis of the Social Structure of a Trans-Volta Tribe. Oxford University Press, London.

Fraser, C., 2002. Shack Chic: Art and Innovation in South African Shack-Lands. Quivertree Publications, Cape Town.

Gibson, N.C., 1999. Rethinking Fanon: The Continuing Dialog. Humanity Books, Amherst, NY.

Greenberg, J.H., 1972. Linguistic evidence regarding Bantu origins. Journal of African History 13 (2), 189–216.

Griaule, M., 1948. Dieu D'eau: Entretiens Avec Ogotemmêli. Éditions du Chêne, Paris (1965 Conversations with Ogotemmêli: An Introduction to Dogon Religious Ideas). Oxford University Press, London.

Herskovits, M., 1941. The Myth of the Negro Past. Harper and Brothers, New York.

Hountondji, P., 1977. Sur la philosophie africaine: critique de l'ethnophilosophie. Maspero, Paris (1983 African Philosophy: Myth and Reality). Indiana University Press, Bloomington, IN.

Kunene, M., 1979. Emperor Shaka the Great: A Zulu Epic (M. Kunene, Trans.). Heinemann, London.

Matory, J.L., 2006. Black Atlantic Religion: Tradition, Transnationalism, and Matriarchy in the Afro-Brazilian Candomblé. Princeton University Press, Princeton.

Mudimbe, V., 1988. The Invention of Africa: Gnosis, Philosophy, and the Order of Knowledge. Indiana University Press, Bloomington, IN.

Roberts, M.N., Roberts, A.F. (Eds.), 1996. Memory: Luba Art and the Making of History. Museum for African Art, New York.

Schildkrout, E., Keim, C.A. (Eds.), 1998. The Scramble for Art in Central Africa. Cambridge University Press, Cambridge, UK.

Schwab, G., 1947. Tribes of the LiberiaN Hinterland, with Additional Material by George W. Harley. Report of the Peabody Museum Expedition to Liberia. Kraus, New York.

Sembene, O., 1960. Les bouts de bois de Dieu: Banty Mam Yall. Le Livre Contemporain, Paris (1976 God's Bits of Wood). Heinemann, London.

Shipley, J., 2013. Living the Hiplife: Celebrity and Entrepreneurship in Ghanaian Popular Music. Duke University Press, Durham, NC.

Soyinka, W., 1964. Five Plays: A Dance of the Forests, the Lion and the Jewel, the Swamp Dwellers, the Trials of Brother Jero, the Strong Breed. Oxford University Press, London.

Tempels, P., 1946. Bantoe-filosofie. Oorspronkelijke Tekst. De Sikkel, Antwerp (1959 Bantu Philosophy). Presence Africaine Paris.

Thiong'o, Ngugi wa, 1986. Decolonising the Mind: The Politics of Language in African Literature. Vita Books, Wembley, Middlesex, UK.

Turner, V., 1968. The Drums of Affliction: A Study of Religious Processes among the Ndembu of Zambia. Clarendon Press, Oxford, UK.

Vansina, J., 1961. Oral Tradition. A Study in Historical Methodology. Aldine, Chicago, IL.

African Studies: Health

John C Caldwell, Australian National University, Canberra, ACT, Australia

Published by Elsevier Ltd.
This article is reproduced from the previous edition, volume 1, pp. 244–250, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Health has been defined by the World Health Organization as a condition of complete physical and psychological wellbeing. In common usage the term 'good health' is employed to mean that there is present no major manifestation of ill-health, and this is also the most satisfactory way of scientifically determining the situation. The most clearly defined outcome of ill-health (or morbidity) is death (or mortality), because it can be most certainly defined and measured, and because in its irreversibility it is an index of the most extreme ill-health. This article focuses on mainland sub-Saharan Africa plus the large island nation, Madagascar, omitting the smaller islands of the Indian and Atlantic oceans with their mixed historical, ethnic, and cultural backgrounds. All of these have lower mortality than any mainland African country and most are richer. The advances in health research since the mid-twentieth century are outlined. From 1950 to 1985 mortality declined and health improved markedly in the region, if not quite as rapidly as elsewhere in the world. Since 1985, however, sub-Saharan Africa, in marked contrast to all other developing world regions, has experienced a major health crisis. It began with problems in sustaining economic growth, which led to both a fall in health investment and the acceptance of structural adjustment programs where the 'consumer pays' principle was applied to both health and education services. Mortality decline almost ceased. Then, parts of the region, conspicuously East and Southern Africa, were afflicted with an AIDS epidemic of far greater intensity than any other world region. Mortality is now rising throughout East and Southern Africa to such an extent that Botswana and Zimbabwe anticipate a drop of 20 years in their life expectancies between the late 1980s and the first years of the twenty-first century. The improvement in health since the beginning of the twentieth century has been a product of better health services, more universal education, the growth of a market economy, changes in attitudes towards children, and other factors. Many of these factors are at present under threat, and this combination is intensifying the African health crisis.

Health has been defined by the World Health Organization as a condition of complete physical and psychological wellbeing. In common usage the term 'good health' is employed to mean that there is present no major manifestation of ill-health, and this is also the most satisfactory way of scientifically determining the situation. The most clearly defined outcome of ill-health (or morbidity) is death (or mortality) because it can be most certainly defined and measured, and because in its irreversibility it is an index of the most extreme ill-health. This entry will focus on mainland sub-Saharan Africa plus the large island nation, Madagascar, omitting the smaller islands of the Indian and Atlantic oceans with their mixed historical, ethnic, and cultural backgrounds. All of these have lower mortality than any mainland African country and most are richer. The advances in health research since the mid-twentieth century are outlined.

Researching Africa's Health Levels

At mid-twentieth century sub-Saharan Africa was assumed to be the unhealthiest region in the world, in spite of an almost complete lack of data to confirm this view. That confirmation was to be achieved later by reconstructing the situation with data from subsequent research, which showed that as late as 1950–5 the region was characterized by a life expectancy at birth of only 37 years (United Nations Population Division, 1999). The problem in making such estimates was a complete lack of vital registration, and of usable mortality and morbidity information in censuses and demographic surveys. Comprehensive national counts of deaths or illness are still not available anywhere in the region.

This problem of inadequate data has been overcome in three ways:

1. The so-called 'indirect methods' of estimating mortality (and fertility) levels from inadequate data have been invented by William Brass and colleagues associated in later years with the London School of Hygiene and Tropical Medicine, while the stable and quasi-stable population model approach was developed by Ansley Coale and colleagues at the Office of Population Research, Princeton University. At first the 'Brass' methods provided estimates only of child mortality, but later techniques were developed for adult mortality, although the latter results were usually less secure.
2. Censuses and national sample surveys with questions allowing indirect or even direct estimates of vital rates were developed. The first censuses with questions on births and deaths were those in British East Africa in 1948. From the 1960 census round, the United Nations assisted African censuses. From 1954, demographic sample surveys were conducted in most francophone African countries. Subsequently, many African countries participated in the great international survey programs: the World Fertility Survey (WFS) (which contained mortality questions) from 1975 and its successor, the Demographic and Health Surveys (DHS) from 1985. Nevertheless, by the year 2000 there still had been no adequate censuses or surveys of the Congo (Democratic Republic) or Angola, and only preliminary reports had been issued for the DHS surveys of South Africa and Ethiopia.
3. High-intensity surveillance projects had been established in a range of areas. These provided demographic and health

information and were often the sites for intervention projects. They included largely demographic projects such as the Sine-Saloum (1962–6) and later studies in Senegal, and those with greater attempts to investigate health: Pare-Taveta, Kenya and Tanzania (1954–6), Keneba, Gambia (1956 onward), Danfa, Ghana (1969–74), Malumfashi, northern Nigeria (1974–9), Machakos, Kenya (1974–81), Kilombera, Tanzania (1982–8) and Navrongo, Ghana (1993 onward). All programs measured health, mostly by mortality or survival measures, and all involved collaboration between African and outside institutions. African demographic estimation was the most challenging in the world, and the new techniques developed to meet that challenge were employed subsequently in other parts of the developing world and by historical demographers.

The data released from these investigations were employed by a succession of research programs:

1. The African project of Princeton's Office of Population Research analyzed the results from predominantly francophone African demographic surveys and anglophone African censuses from 1961, and published *The Demography of Tropical Africa* in 1968. Its major contribution was its fertility estimates, and those on mortality were largely confined to infancy.
2. The International Union for the Scientific Study of Population (IUSSP) in the mid-1980s commissioned papers on African mortality change for a 1987 conference and published *Mortality and Society in Sub-Saharan Africa* in 1992.
3. The American National Research Council's Committee on Population established a program on the population dynamics of sub-Saharan Africa in 1989 and published *Demographic Change in Sub-Saharan Africa* and five other volumes in 1993.
4. From the late 1980s the World Bank commissioned studies of the health of sub-Saharan Africa and in 1991 published *Disease and Mortality in Sub Saharan Africa*. The first part of the book was constituted by studies of child and adult mortality and child malnutrition drawn from WFS and DHS data, the second part by studies of specific ailments from miscellaneous sources, and the third part by reports on morbidity and mortality from the various surveillance projects.
5. A collaborative analytical program of persons working in African universities and institutions on West and Middle Africa resulted in 1975 in the publication of *Population Growth and Socioeconomic Change in West Africa*. In it, Cantrelle developed his thesis on tropical mortality (discussed below).

Some African research themes have a strong behavioral component. Studies of the impact of parental education, especially maternal education on child survival began in the region (Orubuloye and Caldwell, 1975; Caldwell, 1979; Farah and Preston, 1982), and have since become numerous elsewhere. The AIDS epidemic has brought demographers and other social scientists to the study both of AIDS mortality and of sexual relations and other aspects of HIV transmission (Cleland and Way, 1994; Awusabo-Asare et al., 1997).

Health in sub-Saharan Africa has not been a major feature of social science literature, probably because the data have been elusive and difficult to interpret. It has constituted 7.8% of all articles in *Social Science and Medicine*, 2.5% in *Population and Development Review* and, except for anthropological accounts of traditional healing practices, none in *Africa*. The latter point underscores the fact that the major social science input into African health research has been by demographers. Biomedical researchers have undertaken research on specific diseases. The balance of this article summarizes what the research reveals about African health.

The Health Situation

The first reliable estimates of sub-Saharan mortality were for the late 1940s and early 1950s, and were for infants, and then young children, made by adjusting mothers' reports of their children's deaths. These revealed infant mortality rates (deaths per 1000 births during the first year of life) ranging in most of francophone Africa between 200 and 275 (and probably equivalent to life expectancies ranging from 25 to 35 years). At the district level, Mopti in Mali and Luanda in Angola recorded infant mortality rates of 350 and 329, respectively, showing that one-third of all births resulted in deaths during the first year of life. In contrast, Kenya recorded an infant mortality rate of only 132 (a life expectancy around 45 years) and two of its districts, Central and Rift Valley Provinces, registered rates under 100. Subsequent studies increasingly involving life histories, many summarized by Althea Hill, showed infant and child survival improving nearly everywhere until at least the 1980s. West African mortality was higher than that of East and Southern Africa, but convergence was taking place. As appropriate questions were added to surveys, advances were also made in the study of adult mortality, which fell consistently until the 1980s. Thereafter, the decline became much slower in West Africa, and halted or reversed in East and Southern Africa (Timaeus, 1999).

At the end of the twentieth century (1999) sub-Saharan Africa's life expectancy was 49 years, with male expectancy around 48 years and female 50 years. This compared with 61 years in South Asia, 64 years in North Africa, 65 years in Southeast Asia, 69 years in Latin America, 72 years in East Asia, 73 years in Southwest Asia, and 75 years in industrialized countries. With 9% of sub-Saharan African babies dying in the first year of life, and 15% in the first five years, the region's infant and child mortality was also the world's highest. Sub-Saharan Africa's life expectancy compares with that in Western Europe at the beginning of the twentieth century, suggesting a health lag of about 100 years. Within sub-Saharan Africa, Southern Africa recorded a life expectancy of 56 years, West Africa 52 years, Central Africa 49 years, and East Africa 44 years, the latter's poor performance being a recent product of the AIDS epidemic. National levels ranged from 55 or more years in Ghana, Liberia, South Africa, and Cameroon to 42 years or less in Malawi, Swaziland, Botswana, Zimbabwe, Niger, and Ethiopia. Such low levels as those for the last-mentioned countries are no longer found anywhere else in the world.

Past Mortality Trends

If tropical African population growth was almost stationary before the European partition of the region in the 1880s, then stable population calculations suggest that life expectancy was around 20 years. Alternatively, if, as has been suggested, the European presence on the coast had allowed the adoption of new foodstuffs permitting the denser settlement of the forest and other wetter lands, thus leading to a population growth rate as high as 0.5% per annum, life expectancy may have risen to around 23 years. In any case, mortality was, as traders on the west coast knew, the highest in the world, with malaria, yellow fever and other insect-borne diseases protecting the land from European settlement of the type that had occurred in Latin America. These mortality levels are not much greater than the demographic surveys of the 1950s and early 1960s found persisting in some remote rural areas of the West African savanna.

The Second Half of the Twentieth Century

Mortality estimates spanning the second half of the twentieth century, constructed by the United Nations Population Division (1999) from the research findings described above, are shown in **Table 1**. Such estimates can hardly be exact but they are probably reasonably close to the truth. They are compared with estimates for the world and selected other developing regions.

What emerges from **Table 1** is that sub-Saharan Africa is a healthier place than it was half a century ago. By the end of the twentieth century its life expectancy was above that of the world as a whole at mid-century, and only a little behind that of Latin America at that time. Nevertheless, it had fallen further behind every other world region. The most appropriate comparison is with south Asia, which has a somewhat lower per capita income than sub-Saharan Africa (and, taking purchasing power into account, a lower per capita parity purchasing power) and a slightly higher level of female literacy. In 45 years, Africa's life expectancy rose by 12 years, compared with 23 years in south Asia.

Table 2 explores the African health failure further. Neither sub-Saharan Africa nor North Africa participated to the same extent as Asia and Latin America in the great leap forward in reducing mortality that characterized much of the world during the two decades following World War II. Nevertheless, sub-Saharan Africa's record in health advancement was moderately successful until the 1980s, but since about 1985 there has been little health advance at all.

The relatively limited sub-Saharan African success in health up until the 1980s can be at least partly explained by the fact that ecological and other conditions meant that it was the only major world region that completely failed to reduce the level of malaria. Since the 1980s two other factors have also been important. The first was a slowdown in economic growth that necessitated the acceptance of 'structural adjustment' policies: in many countries, investment in the health sector almost ceased, and charges for government medical services were instituted. The second factor was the arrival of the AIDS epidemic with an intensity experienced nowhere else in the world.

All parts of sub-Saharan Africa have experienced a deceleration in health improvement, but this has been spread unevenly, as **Table 3** shows. The health situation in the region had been described as one improving along a diagonal with the lowest levels in the northwest of West Africa and the highest levels in Southern and East Africa (see Feachem and Jamison, 1991: pp. 31–2). Largely as a result of the AIDS epidemic, this situation has been reversing, and by the 2010s West Africa may have the highest regional life expectancy south of the Sahara.

A very similar experience characterized trends in infant and child mortality rates. Sub-Saharan Africa's infant mortality rate in 1950–5 has been estimated as 176 deaths per 1000 births, with the south Asian rate at 186. By 1995–2000 the sub-Saharan African rate was 93 compared with a south Asian rate of 73. The child mortality rate compared even more badly; in 1995–2000 it was 152 in sub-Saharan Africa (i.e., 15.2% of births resulted in a death before 5 years of age) compared with 96, or less than two-thirds the African level, in south Asia. The explanation is what Cantrelle (1975) called the 'tropical pattern,' and others have termed the 'African pattern.' When Cantrelle was writing, not only was infant mortality in tropical Africa high, but almost as many deaths occurred to each cohort of births between their first and fifth birthdays as during the first year of life. That fraction has now fallen to 63% (although it is still 73% in West Africa) compared with 32% in south Asia. The reasons for very high one-to-four- year-old mortality in tropical Africa are probably the high level of infectious

Table 1 Life expectancy at birth (in years), 1950–2000 and per capita income (in US$), 1997

	Sub-Saharan Africa	*North Africa*	*Asia*	*South Asia*	*Latin America*	*World*
1950–55	37	42	41	39	51	46
1960–65	41	46	48	45	57	52
1970–75	45	51	56	50	61	58
1980–85	49	57	60	55	65	61
1990–95	49	62	64	60	68	64
1995–2000	49	65	66	62	69	65
Per capita income 1997	$520	$1160	$2450	$470	$3950	$5170
Adult female literacy 1995	47%	41%	50%	36%	85%	62%

Source: United Nations Population Division; Population Reference Bureau, 1999. World Population Data Sheet 1999. Washington, DC; World Bank, 1999. World Development Report 1998/99. Oxford University Press, New York.

Table 2 Average annual increases in life expectancy (in years), 1950–2000

	Sub-Saharan Africa	*North Africa*	*Asia*	*South Asia*	*Latin America*	*World*
1950–55 to 1960–65	0.4	0.4	0.7	0.6	0.5	0.6
1960–65 to 1970–75	0.4	0.5	0.8	0.5	0.4	0.6
1970–75 to 1980–85	0.3	0.5	0.4	0.5	0.4	0.3
1980–85 to 1990–95	0.1	0.6	0.4	0.5	0.3	0.3
1990–95 to 1995–2000	0.0	0.5	0.2	0.4	0.2	0.3

Source: As for Table 1. Increases not calculated on the rounded figures of Table 1.

Table 3 Life expectancies and average increases for the major regions of Africa, 1950–2000

	Life expectancy at birth (years)				*Average annual increase in life expectancy since previous period (years)*			
	West Africa	*Middle Africa*	*East Africa*	*Southern Africa*	*West Africa*	*Middle Africa*	*East Africa*	*Southern Africa*
1950–55	36	36	36	44	–	–	–	–
1960–65	40	40	41	49	0.4	0.4	0.4	0.5
1970–75	43	44	45	53	0.3	0.4	0.4	0.4
1980–85	46	48	47	56	0.3	0.4	0.2	0.3
1990–95	49	51	45	59	0.2	0.3	–0.2	0.3
1995–2000	50	50	45	54	0.2	–0.1	–0.2	–0.9

Source: As for Table 1.

disease affecting that age group, poor weaning practices, and unsatisfactory foods available for weaning.

Why Did Health Improve?

It seems likely that life expectancy climbed by about 15 years between the 1880s and the early 1950s, and then by another 11 years in the next three decades. The reasons are complex, and the use of modern medicine is only part of the story.

Much of the explanation for mortality decline in the first period was probably the organization brought about by colonial governments. Inter-ethnic and individual violence probably declined. People were separated from others when plague struck, and from wild animals to curb sleeping sickness. Roads and railways helped to usher in a market economy which distributed food, and medicine, more widely. Capitalism and education increased individualism and made it likely that greater initiatives would be taken to prevent or cure illness. Immunization, led by smallpox vaccination, eventually brought the great epidemic diseases under control. Yellow fever has almost vanished, and cholera levels have declined. Digging drains and oiling stagnant water reduced malaria in towns, and mosquito nets partly protected colonial and local elites.

There was a slow spread of government and missionary hospitals. Although there is debate about the impact of modern medicine on poor, predominantly rural societies, a comparison of two areas of similar socioeconomic levels in Nigeria showed that the one which had possessed for a generation a small, adequately staffed and supplied hospital offering free services was characterized by a life expectancy 12 years greater than the area with no facilities (Orubuloye and Caldwell, 1975). The treatment of water supplies, usually in urban areas and often inadequate, has improved, but is still often woefully bad. Better sanitation and hygiene practices have doubtless also reduced mortality. By the 1990s, Demographic and Health Surveys were reporting that about half of all countries had a majority of children immunized against tetanus, diphtheria, and pertussis (whooping cough), while the coverage against measles (a major killer in the region) was increasing and the incidence of poliomyelitis was declining steeply. But morbidity and mortality from diarrhea and pneumonia were still high. Malaria was almost as bad as ever, although many lives were saved by the use of drugs, and HIV/AIDS was presenting a horrific challenge.

Health at the Dawn of the AIDS Era

By the early 1980s sub-Saharan African life expectancy was 48 years, having increased by seven years in the previous decade, and seeming to promise similar gains to come. This was not to be so, because of problems in funding the health system and the emergence of the AIDS epidemic. In contrast to most of Asia and North Africa, female child mortality was as low as that of males. There were significant mortality differentials by region (with mortality lowest in Southern Africa), ethnic group (even when neighbors), parental education (with mother's education being more important for child survival than father's education), by occupation (with farmers' death rates highest), and by residence (with urban health better). Mortality was highest in countries like Ethiopia and Mozambique where civil unrest and war had disorganized the health and other systems. It was also higher in the drought-prone savanna countries, but no higher than would be expected from their relatively low income and educational levels, which were in turn the product of impoverished agricultural resources. Drought still visited these lands regularly but evidence accumulated that, although it caused much distress and livestock loss, excess human mortality was less than might have been anticipated because of the scale of migration to better-off areas and towns.

Information on the nature of illness and the causes of death is still meager because so few people are seen by doctors or die in hospitals. By the 1980s the great epidemic diseases were largely under control. Campaigns, which mostly proved successful, aimed at eradicating onchocerciasis (river blindness), and some progress was being made against schistosomiasis (bilharziasis). Malaria, particularly its worst form, falciparum malaria, was almost universal, moderated only in the highest parts of East and Southern Africa by lower temperatures and in South Africa by a more temperate climate and successful eradication. A Gambian study revealed malaria to be the dominant cause of illness, except among children under 3 months of age, who were relatively free of it. Research in Tanzania showed that 31% of the sick were suffering from malaria, 13% respiratory infections and 7% diarrhea.

A major World Bank/Harvard University study (Murray and Lopez, 1997) of global mortality estimated that 65% of the region's mortality was still attributed to communicable, congenital, maternal, and nutritional causes (compared with 51% in India, 42% in all developing countries, and 6% in developed countries), 23% by non- communicable disease, mostly cardiovascular and cancer (compared with 40%

in India, 47% in all developing countries, and 86% in developed ones), and 12% by violence and accident (only a little higher than other areas). A World Bank survey (Feachem and Jamison, 1991) found that 47% of deaths were from infectious and parasitic causes (malaria and measles being particularly important), which, together with perinatal problems, accounted for nearly all child deaths, and 16% from circulatory disease and cancer (but their category of 'other,' including indefinable, was 23%). A longitudinal study in Machakos, Kenya (Muller and van Ginneken, 1991), an area where altitude renders malaria a minor complaint, ascribed 16% of deaths to respiratory infection, 13% to congenital factors, 11% to intestinal infections, 8% to measles, 7% to tuberculosis, 7% to other infectious and parasitic diseases, 4% to nutritional and metabolic causes, 4% to problems of the digestive system, 3% to malaria, 11% to diseases of the circulatory system and cancer, and 6% to violence and injuries, while 11% could not be determined. Hepatitis B also presents dangers. Maternal mortality has been estimated at 655 per 100 000 births, the world's highest level and 30 times that of Europe. This, together with deaths arising from miscarriages and abortion, translates, because African women still average six live births, into a lifetime chance for females of dying from maternity causes of about 5%. It is likely that the lifetime risk will fall substantially in the near future as the result of a fertility decline.

The AIDS Epidemic

HIV-2, perhaps the more ancient of the two human immunological retrovirus strains, is present in West Africa, Angola and Mozambique, but, because it is not a major cause of mortality, will not be discussed here. The HIV-1 /AIDS epidemic began in East Africa in the early 1980s, at much the same time as it appeared elsewhere. By the mid-1990s prevalence rates in Southern Africa were higher than they have ever been in East Africa. UNAIDS/WHO (1998) estimates for the end of 1997 indicate that 9.6 million Africans have already died of AIDS (82% of the world total) and a further 21 million are now infected (69% of the world total). More startlingly, there are 15 countries, constituting the 'Main AIDS Belt' and stretching from Ethiopia through East Africa to South Africa, which contain 4% of the world's population and half the world's HIV/AIDS. All of them have adult HIV prevalence rates of at least 9%, but the rate reaches 15% in Malawi and Mozambique, around 20% in Namibia and Swaziland, and 25% in Zimbabwe and Botswana. The latter levels imply more than a doubling of the death rate, with a lifetime expectation of dying of AIDS over 50%. United Nations population projections do not anticipate Zimbabwe regaining its 1985 life expectancy without a gap of 35 years, or Botswana with a smaller gap than 50 years. Sub-Saharan Africa's epidemic is almost entirely heterosexual in primary transmission, with the result that at least as many women as men are infected. Thus, although the region contains only 59% of the infected men in the world, it has 81% of infected women, and, because of higher birthrates, probably at least 85% of infected children. Thus, almost uniquely in the world, all parts of the community are affected. The AIDS epidemic is also catalyzing an increase in tuberculosis levels. There is growing evidence that HIV-positive women are likely to be rendered infertile or subfertile. Campaigns to contain the epidemic have been less intensive than the crisis demands, and largely ineffective.

The African Health System

Government hospitals are found chiefly in towns, with teaching or specialist referral hospitals in the capitals or other large cities. In most countries there are also missionary hospitals, often in rural areas. Nevertheless, most rural people depend on primary health services in the form of village health posts, clinics or dispensaries, and health centers. This system is developed to very different degrees across the region, and the health facilities often have insufficient drugs. Most have now moved to a 'user-pays' system, often characterized by declining attendances and by women losing their ability to take their children straight to the facilities without consulting male relatives (Orubuloye et al., 1991). Research has shown that most clients of health centers come from no further away than the village in which the center is located.

The system is supplemented by private pharmacies and medical stores. A Nigerian study showed that in rural areas the majority of drugs sold are malaria suppressants, worm syrups, and analgesics, although antibiotics are usually also in stock. Private doctors and nurses are becoming of increasing significance in most countries. Traditional practitioners remain important. They do not have an agreed-upon pharmacopeia as in various Asian medical systems, and their medicines are derived from animals as well as herbs. Many also identify evil forces and reveal how they can be nullified.

The great campaigns which eliminated smallpox and contained sleeping sickness, yellow fever, and cholera are largely things of the past, although watch is kept for outbreaks. Efforts continue to reduce the incidence of leprosy and tuberculosis, and to ensure that those people with malarial fever do not die. Immunization of children is now a major health weapon, and recent years have seen sustained attempts to control measles. Public health efforts continue to ensure safe drinking water and sanitation, but it is unlikely that the levels of safety are as high as respondents report to the international survey programs.

Trends in the Research Literature

More articles on the social and behavioral aspects of health are published in the journal *Social Science and Medicine* than anywhere else. In the last quarter of the twentieth century 865 of these papers were on sub-Saharan Africa. Their numbers rose rapidly at first from an annual average of nine during the late 1970s to 24 in the early 1980s and then reached a plateau at just under 50 per year. Many of the articles were derived from empirical studies of specific locations or diseases. The major topics were access to health services (132 papers), the interaction of traditional and modern health beliefs and services (75 papers), HIV/AIDS (68 papers, nearly all in the 1990s), maternal and child health (55 papers), the political economy of health (48 papers), and health economics (39 papers). Some trends were noticeable. The number of papers published on the interaction between traditional and modern health services declined after the 1980s,

those on the political economy of health peaked in the late 1980s as economic structural adjustment policies were applied, and HIV/AIDS papers mostly appeared in the 1990s, with no clear trend during the decade.

The Future

The immediate future is not reassuring. In contrast to the rest of the world, the regional life expectancy is stalled at under 50 years. Most sub-Saharan African countries are unable to do more than maintain their present public health systems. The impact of such systems depends on their level of use, which is largely determined by the population's ability to pay, and by their level of education, which raises the priority given to successful treatment and the skill with which it is administered. The fee-for-service principle is limiting the expansion of use of both the health and education systems. On the other hand, the likelihood of accessing the health system is probably increasing with the transition from subsistence to market agriculture. Fertility decline has now begun in a range of African countries. Smaller families will mean fewer women dying from maternal causes, and probably greater concentration on the education and health of children.

It is now likely that no country in mainland sub-Saharan Africa will, by the year 2000, have attained any of the three criteria established by the World Health Organization to certify the attainment of the 1978 Alma Ata Declaration of 'Good Health for All' by the year 2000: a life expectancy of at least 60 years, and a survival level of 95% of births to 1 year of age and 93% to 5 years of age. Indeed, most of East and Southern Africa will probably by that date be slipping further away from these criteria.

Editor's Note

The Ebola virus, rampant HIV infections, the emergence of multiply drug-resistant strains of tuberculosis, along with the traditional African morbidities (malaria, lung infections, diarrheal diseases, tetanus, meningitis, elephantiasis, leprosy, helminthiasis, and the panoply of STDs) has kept the health of the African population on the minds of health researchers around the world. Indeed, it is not unfair to say that African morbidity and mortality is its own subsection of the public health discipline. So the recent literature on African health and disease is vast (Giles-Vernick and Webb, 2013; Jamison et al., 2006; Crane, 2013; Cooper et al., 2013; Fassin, 2007; Amzat and Razum, 2014; Webb, 2014).

See also: AIDS: Acquired Immune-Deficiency Syndrome; Health Politics and Policy; Health: Anthropological Aspects; Mortality and the HIV/AIDS Epidemic.

Bibliography

Amzat, J., Razum, O. (Eds.), 28 February 2014. Medical Sociology in Africa Hardcover. Springer.

Awusabo-Asare, K., Boerma, T., Zaba, B. (Eds.), 1997. Evidence of the socio-demographic impact of AIDS in Africa. Health Transition Review, 7 (Suppl. 2).

Brass, W., Coale, A.J., Demeny, P., Heisel, D.F., Lorimer, F., Romaniuk, A., van de Walle, E., 1968. The Demography of Tropical Africa. Princeton University Press, Princeton, NJ.

Caldwell, J.C. (Ed.), 1975. Population Growth and Socioeconomic Change in West Africa. Columbia University Press, New York.

Caldwell, J.C., 1979. Education as a factor of mortality decline: An examination of Nigerian data. Population Studies 33, 395–413.

Caldwell, J.C., 1985. The social repercussions of colonial rule: Demographic aspects. In: Boahen, A.A. (Ed.), General History of Africa, vol. 7: Africa under Colonial Domination 1880–1935. University of California Press, Berkeley, CA and Heinemann, London for UNESCO, Paris.

Cantrelle, P., 1975. Mortality: Levels and trends. In: Caldwell, J.C. (Ed.), Population Growth and Socioeconomic Change in West Africa. Columbia University Press, New York.

Cleland, J., Way, P. (Eds.), 1994. AIDS impact and prevention in the developing world: Demographic and social science perspectives. Health Transition Review, 4 (Suppl.).

Coale, A.J., Demeny, P., 1966. Regional Model Life Tables and Stable Populations. Princeton University Press, Princeton, NJ.

Committee on Population, National Research Council, 1993. Demographic Effects of Economic Reversals in Sub-Saharan Africa. National Academy Press, Washington, DC.

Cooper, Andrew F., Kirton, John J., Lisk, Franklyn (Eds.), 28 November 2013. Africa's Health Challenges: Sovereignty, Mobility of People and Healthcare Governance, New edition. Ashgate Pub Co..

Crane, Johanna Tayloe (Ed.), 24 September 2013. Scrambling for Africa: AIDS, Expertise, and the Rise of American Global Health Science, first ed. Cornell University Press.

Farah, A.-A., Preston, S.H., 1982. Child mortality differentials in Sudan. Population and Development Review 8, 365–383.

Fassin, Didier (Ed.), March 14, 2007. When Bodies Remember: Experiences and Politics of AIDS in South Africa (California Series in Public Anthropology), first ed. University of California Press.

Feachem, R.G., Jamison, D.T. (Eds.), 1991. Disease and Mortality in Sub-Saharan Africa. Oxford University Press, New York for World Bank, Washington, DC.

Foote, K., Hill, K.H., Martin, L. (Eds.), 1993. Demographic Change in Sab-Saharan Africa. National Academy Press, Washington, DC.

Giles-Vernick, Tamara, Webb Jr., James L.A. (Eds.), 2013. Global Health in Africa: Historical Perspectives on Disease Control. Ohio University Press, Athens, OH.

Hill, A., 1992. Trends in childhood mortality in sub-Saharan Mainland Africa. In: van de Walle, E., Pison, G., Sala-Diakanda, S. (Eds.), Mortality and Society in Sub-Saharan Africa. Clarendon, Oxford.

Jamison, Dean T., Feachem, Richard G., Makgoba, Malegapuru W., Bos, Eduard R., Baingana, Florence K., Hofman, Karen J., Rogo, Khama O. (Eds.), 2006. Disease and Mortality in Sub-Saharan Africa, second ed. World Bank, Washington, DC.

Muller, A.S., van Ginneken, J., 1991. Morbidity and mortality in Machakos, Kenya. In: Feachem, R.G., Jamison, D.T. (Eds.), Disease and Mortality in Sub-Saharan Africa. Oxford University Press, New York.

Murray, C.J.L., Lopez, A.D. (Eds.), 1996. The Global Burden of Disease. Harvard University Press, Cambridge, MA for World Bank, Washington, DC.

Murray, C.J.L., Lopez, A.D., 17 May 1997. Global mortality, disability, and the contribution of risk factors: Global burden of disease study. Lancet 349, 1436–1442.

Orubuloye, I.O., Caldwell, J.C., 1975. The impact of public health services on mortality: A study of mortality differentials in a rural area of Nigeria. Population Studies 29, 259–272.

Orubuloye, I.O., Caldwell, J.C., Caldwell, P., Bledsoe, C.H., 1991. The impact of family and budget structure on health treatment in Nigeria. Health Transition Review 1, 189–210.

Population Reference Bureau, 1999. World Population Data Sheet 1999. Washington, DC.

Timaeus, I.M., 1999. Mortality in sub-Saharan Africa. In: Chamie, J., Cliquet, R.L. (Eds.), Health and Mortality: Issues of Global Concern. United Nations, New York.

UNAIDS/WHO, 1998. Report on the Global HIV/AIDS Epidemic June 1998. Joint United Nations Program on HIV/ AIDS and World Health Organization, Geneva.

United Nations, Population Division, 1999. World Population Prospects: The 1998 Revision, vol. 1, Comprehensive Tables. United Nations, New York.

Van de Walle, E., Pison, G., Sala-Diakanda, M. (Eds.), 1992. Mortality and Society in Sub-Saharan Africa. Clarendon Press, Oxford.

Webb Jr., James L.A. (Ed.), March 31, 2014. The Long Struggle against Malaria in Tropical Africa Hardcover. Cambridge University Press.

World Bank, 1999. World Development Report 1998/99. Oxford University Press, New York.

African Studies: History

Frederick Cooper, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

African and African-American intellectuals long sought to counter primitivizing ideologies of their times by pointing to narratives of African state building. The real breakthrough in writing African history occurred as colonial rule was crumbling and the quest for a usable past – notably a usable *national* past – attracted young scholars in Africa and beyond. If the 1960s witnessed an emphasis on the particularity of African societies, disillusionment with Africa's present brought about a wider consideration of its past, especially the ways the slave trade and colonization constrained the continent's possibilities. Scholars attempt to document varied forms of imagination and communication, of networks and institutions, in different temporal and spatial frameworks.

The first task African historians had to accomplish was to show that African history actually existed. Scholars in Europe and North America had long treated Africa as the embodiment of the primitive and the traditional. In the early twentieth century, a more sophisticated version of anthropology began to treat Africa as a laboratory of social specificity, where forms of social organization could be compared as if each were bounded and timeless. As Africa's subordination to European powers came under increasing attack in the late 1940s, intellectuals inside and outside Africa began to argue that rethinking Africa's past was a necessary part of its future.

The idea of a usable past – specifically a usable national past – was African history's strength and its weakness in the 1960s. Asserting that history could be studied scientifically was part of a new politics of intellectual inquiry. Pioneering studies stressed the dynamism of precolonial societies; resistance to conquest was a harbinger of nationalist movements. But by the 1970s, a more complicated present was leading to a more complicated past, above all to new ways of thinking about Africa's relationship to the rest of the world and the implications of this relationship for historical writing. Yet the time frame for considering the emergence of African history needs to be pushed back even further, and its spatial dimension – the definition of Africa, the relationship of its constituent social and political units, and the significance of the continent to Africans in the Americas – need examination as well. The period after independence is just beginning to be the focus of historical inquiry.

Africa and the World

Africa was in part an invention of its diaspora, a unit that became of world-historical significance because slave traders – from the sixteenth century – defined it as a place where one could legitimately develop commerce in human beings. Over time, enslaved Africans and their descendants in the Americas began to appreciate the commonality of their fate and many looked to 'Africa' as an almost mythic symbol of their being something more than chattel. Certain nineteenth century African-American religious leaders looked toward 'Africa' and 'Ethiopia' (although few slaves came from that kingdom), and through such language asserted Africans' important place in a universal history – in the unfolding of Christian civilization. As some people of African descent returned to the continent in the nineteenth century – repatriated ex-slaves to the British colony of Sierra Leone, Brazilian traders in the Bight of Benin or Angola – some saw themselves as part of broad, transatlantic 'nations', sharing ancestry and culture but needing Christianity to link a torn-apart past to a reintegrated future (Matory, 1999; Hawthorne, 2010; Peel, 2003; Ferreira, 2012).

By the late nineteenth century, Africans from coastal societies – Christian, Western educated, but integrated into regional social organizations – began to write about their own regions in terms that linked different historical sensibilities. Africanus Horton (1868), Edward Wilmot Blyden (1887), and J. E. Casely-Hayford (1903) countered the primitivizing ideologies of the eras of the slave trade and colonization by describing African societies as complex entities whose traditions of origin defined commonality, whose ideas about kingship and social hierarchy defined political order, and whose interest in commerce with the outside world, in Christianity and Islam, and in Western education marked an open, adaptable attitude to interaction. Africans, they thought, had much to learn from Muslims and Christians, but they brought their own ideas of statecraft and of economic organization to the encounter as well.

The period of escalating European exploration and eventual conquest (from the 1870s to 1910s) brought to African contradictory regards. Some explorers encountered powerful kingdoms whose power they had to respect; others recognized in the impressive mosques of the West African desert edge or the East African coast a history of a long encounter of Africans with the outside world; but many chose to see an unchanging landscape of different peoples ensconced in their particular cultures. The culturally mixed inhabitants of West Africa coastal region were marginalized by colonization; the worldwide connections of African Muslims were played down in favor of their 'tribal' characteristics.

After some efforts toward remaking African societies in a European image – the 'civilizing mission' – colonial regimes began to hitch their legitimacy to the chiefs whose authority toward their subjects colonial rulers needed. With that, the idea of 'tradition' as the essential quality of African life acquired a new salience in colonial ideologies. Colonial regimes – and scholars and intellectuals of that era – were interested in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10116-3

'customary law', in 'folklore', and in 'primitive art'. The growth of African ethnography in the 1920s brought to the continent foreign scholars curious about the diversity of social forms and sympathetic to victims of colonial oppression, but their emphasis on the bounded integrity of each African 'society' was for the most part ahistorical (Conklin, 1997; Sibeud, 2002).

A New Past for a New Future

Even within the ethnic cages of colonial polities, consciousness of the past did not necessarily remain static. Some early mission converts used their literacy in French or English to record genealogies and traditions of 'their' people – using the legitimacy of 'European' writing to articulate indigenous views of the past and to emphasize the integrity of local society (Peterson, 2004). Such histories were sometimes invoked to make claims for collective representation in state-sanctioned councils. Activists linked to international political organizations – pan-Africanist, communist, or otherwise – saw Africa as a once-dynamic continent severely damaged by an oppression they shared with people of color throughout the world, a situation which required liberation movements on a global scale (DuBois, 1946).

The burst of interest in African history after World War II thus has a deeper context. The break, however, was fundamental. In the aftermath of a devastating war against Naziism, European states needed simultaneously to justify their rule over African peoples and to intensify their use of African resources. The challenges to colonial legitimacy and control – from within and outside African colonies – and the increased intrusiveness of regimes into African social and economic life has made this a fascinating period for historical investigation (Cooper, 1996). Even at the time, scholars and intellectuals wondered whether the conception of bounded and static units made sense of the Africa they were observing. Although a few anthropologists had seen in the 1930s that migration, for example, was redefining the nature of social connections, by the 1950s, population movement, cross-cultural interaction, and cultural adaptation demanded scholarly analysis. Meanwhile, the anthropologist Melville Herskovits (1944, 1958) was asking what Africans' experience of governing indigenous kingdoms had to offer to Africa's political future and he was pointing to the influence, via the long history of the slave trade, of African cultures on the New World.

That African polities could be a model for the future was not a possibility for most African political elites – and almost all social scientists – wanted to think about; the 1950s witnessed the escalation of claims from African social movements to be considered part of the 'modern' world and to enjoy the possibilities it entailed, irrespective of race or a past of colonization or enslavement. Social scientists took more interest in where history was supposed to end – 'modernity' – than where people had been. Even Georges Balandier's (1951) insightful analysis of the 'colonial situation' had limited echoes in subsequent years.

Academic historians, the professional custodians of the past, were part of the postwar political and intellectual ferment. The first African PhD in history K.O. Dike of Nigeria was trained by historians of imperial expansion, but he told a different sort of story, one of the interactions among traders, rulers, and warriors, both African and European, in the Niger Delta, and of the adaptation of African social institutions to new forms of competition in the nineteenth century. Dike (1956, based on a 1949 dissertation) insisted that oral sources could be used alongside written ones, although most of his own work was archival. At one level, his work staked itself on using the canonical methods of European history to a different end; at another, his insistence that the locus of history could be found in Africa itself – that interaction was more important than transmission – was a charter for nationalist history.

The next generation of European-trained African historians went further in this direction, and many stressed explicitly that the preconquest past was a precedent for the postindependence future, showing how Africans had brought diverse populations into larger political units, how African initiative in agriculture and commerce had linked ecologically distinct regions with each other and with the outside world, how indigenous religious leaders had built networks that transcended ethnic frontiers, and how Africans had adapted Islam to particular political and cultural purposes (Ranger, 1968).

The quest for a usable past was a strength because it attracted a younger generation of Africans to believe that they could combine international scholarship with a sense of the past they had learned in their own communities and because it allowed Americans and Europeans to come to grips with the way in which the end of colonial empires had forced a reordering of intellectual categories. It was also a weakness because the political context privileged 'state building' over the diverse strategies of women and men, traders and religious figures to live their lives in different ways, and this both diminished a varied past and ratified an increasingly authoritarian present within the new African states. The only aspect of colonial history that fitted the bill was 'resistance'; indeed, some African historians saw the 'colonial episode' as a short and not particularly important interval between an autonomous past and an independent future. A consequence of the new historiography, whatever its limitations, was a heightened interest in methodology, to solve the problem of how to reconstruct historical patterns with a paucity of written documentation, much of that from visitors and conquerors. Jan Vansina (1965) helped to develop rigorous criteria for analyzing oral texts with the same critical eye as employed on written ones, and other scholars explored the use of linguistic and archeological material to chart the movements of people, the evolution of material culture, and the spatial configuration of state building.

African historical scholarship was vibrant in the 1960s and 1970s, and the core of the action was in Africa. Every nation had to have a university, every university a history department. Associations and journals were founded; international congresses were held; and UNESCO brought together African authors to write a comprehensive history of the continent (UNESCO, 1981–93).

Reconnecting Africa and the World

Since the 1970s, African historians have become increasingly conscious of the different ways of rethinking the relationship of

Africa and the rest of the world. They no longer had to prove that Africa had a history. The mounting evidence that decolonization did not free Africa from its problematic relationship to the world economy raised questions about pasts before and during colonization. Why African polities became caught up in the Atlantic slave trade, how the African presence in the Americas contributed to European wealth, and how African societies coped with unequal trade and unequal politics became topics of increasing interest (Rodney, 1972). French Marxist anthropologists (Meillassoux, 1975) argued that such institutions as kinship systems were not so much the product of a peculiarly African culture but of the logic of reproduction of agricultural societies, and they went from there to posit that the resulting modes of production interacted in particular ways with the expanding capitalist system. Other scholars saw that the question appeared differently if instead of asking how a 'society' responded to European traders or rulers one asked how ruling elites reacted. They showed that the strength of internal social groups made rulers look to the manipulation of external relations for power and wealth, and that this made African kingdoms and chiefdoms especially open to external relationships, most disastrously those of the slave trade and later of colonial economies (Peel, 1983). The most recent work on the slave trade emphasizes the intersection of different trading networks – those coming out of African polities seeking external resources with those coming out of Europe that developed to meet the voracious demand for slave labor in the plantations of the Americas from the sixteenth century onward (Green, 2012; Nwokeji, 2010; Heywood and Thornton, 2007).

Historical research allowed not just the use of 'informants' to gain information, but the juxtaposition of different kinds of historical sensibilities and the elucidation of how thinking about the past affects and is affected by political processes in the present (Cohen, 1994; White et al., 2001). Oral history revealed different lines of cleavage within African societies, including gender, age, and status. It described not the actions of fixed social groups against each other, but the flexibility of social arrangements, such as the tendency of people detached from kinship groups – through war or efforts to escape patriarchal authority – to form important but unstable groups of clients of 'big men'. Both oral sources and more nuanced reading of documents reopened the colonial era to historical investigation, not as a history of what white people did for or to Africans but as a dynamic process, in which the limited power of colonizing regimes left numerous fissures where women tried to establish autonomy even as male elders tried to contain it, where labor migrants developed networks linking distant regions, where Christians built independent churches and cults, and where cash-crop producers used their incomes to enhance their kinship groups and chieftaincies. Such histories stand alongside those of expropriation of resources of segregation and discrimination against the most 'Westernized' of Africans, of arbitrary authority, and daily oppression under colonial regimes, in their late, 'reformist' phases as well as brutal early ones. South Africa's path to a racialized form of capitalism – with most Africans deprived of access to land and forced to work part of their lives while being treated as members of 'tribes' – led to a quite distinct economic structure while carrying invidious distinction making (among Africans and between Africans and whites) to an extreme. South Africa's deepening of its form of white domination after 1948, just as Britain and France were staking their legitimacy on development and partial acknowledgment of African political voice, marked a parting of the ways. In the end – but only slowly and painfully – the defense of white power in Portuguese, Africa, Rhodesia, and South Africa proved impossible as independent states offered models and sanctuaries for rebels and as global norms were transformed.

Studying African history has offered the possibility of thinking about 'world' histories in a different way. Rather than seeing Africa as a peculiar place that somehow lacks what made other regions develop more rapidly, one can ask what is particular about each region of the world and what, in the process of interaction, produced inequalities of wealth and power. African historians have been in dialogue with historians of Latin America over economic history and with historians of India over the analysis of colonialism. World history can no longer be seen as a single narrative, but this recognition leaves in place the more difficult question of how is one then to analyze large-scale historical processes. The challenge is to chart the multiple pathways without losing sight of the development of highly unequal political and economic relations on a global scale. Recent work in economic history has cautioned against generalizing about African backwardness, instead emphasizing the unevenness of patterns of economic growth. Africa has experienced growth spurts and periods of retrogression – the 1980s prominent among the latter – and different parts of the continent have followed different trajectories (Jerven, 2010). Such a historical change requires analysis that does not reify 'Africa', or 'Europe', but focuses on complex relationships over 200 years.

The Peculiarities of Disciplinary Divisions of Labor

In the 1930s, Africa not only had no history; it held no interest for political scientists or sociologists either, and for economists only in the sense the colonial economies had spawned 'modern' sectors with certain measurable characteristics. Africa was the domain of anthropologists, the custodians of human particularity, while what was putatively 'European' was assimilated to the universal, and that was where social science found its glory. The dramatic challenges to European power in the 1940s and 1950s shook up this division of labor (Pletsch, 1981). The idea of modernization was espoused within colonial bureaucracies before 'modernization theory' – in the style of the Committee on Comparative Politics of the Social Science Research Council – came into vogue. It allowed officials to think that even if African colonies became independent, they would follow a road charted out to them by the early developers (Latham, 2000). However much modernization reinscribed global hierarchy, eurocentrism, and teleology, it offered African intellectuals and policy-makers a chance to position themselves as mediators between a perceived Euro–American modernity and diverse African particularities; it gave a language to leaders with which to ask industrial countries for the resources they needed to 'develop' and it allowed outsiders to see themselves as forging a world in which people of all races could advance.

American and European political scientists and sociologists flocked for a time to Africa – and their standards of empirical research deserve some of the credit for deflating modernization theory. But after that time, the interest of political scientists, sociologists, and economists in Africa has waned. Anthropologists remain the custodians of African particularity, but now this embraces the complexities of innovation and interaction more than unchanging specificities. Historians, from the 1950s, were the new element in the division of labor, and their work, despite temptations to the contrary, can offer a vision of Africa that stresses change – economic, political, social, and cultural – without confining it to a predetermined route.

Some historians have their version of the modernizing vision. The narrative of African state building from the empire of Sundiata to the empire of Shaka was indeed a selective reading that tied progress to the state. Even the 'underdevelopment' school of the 1970s is as a variant of modernizing themes, for its emphasis on how over the long-term European exploitation retarded growth and transformation in African economies presumes a global narrative of capitalist development which Africa is then held to fall *under*. That Europe's oppression, not Africa's backwardness, is held culpable does not negate the inattention to the myriad innovations, struggles, and changes that have occurred within parts of Africa, the importance of regional, not just overseas, mechanisms of exchange and communication, and the ways in which the actions of Africans limited the actions of Europeans, whatever the latters' intentions (Cooper et al., 1993). Much recent social and economic history bounced off the arguments of the underdevelopment school to develop more varied and more interactive views of change in various periods of African history, including those when European power was seemingly at its height (Berry, 1993).

Multiple Perspectives on a Varied Continent

If Africa is neither a homogeneous space reduced into abject poverty by imperialism nor a series of autonomous societies with their own internal logics, getting a grip on units of analysis is not simple. Rather than take present-day ethnicity as a starting point, some scholars have used concepts of region, of network, and of patron–client relations to see how people in the nineteenth and twentieth centuries constituted actual patterns of relationships, which only sometimes crystallized into groups that maintained boundaries (Ambler, 1988; Glassman, 1995; Barry, 1998). Others have emphasized the long-term interconnectedness of different parts of Africa and with other parts of the world: on trade routes and Islamic networks crossing the Sahara (Lydon, 2009), on further connections to Egypt and Arabia, on efforts to build Islamic states within sub-Saharan Africa, and on the ways in which Islamic networks both accommodated to and challenge colonial regimes after the conquest (Robinson, 2000; Hanretta, 2009). Others have examined how African-American missionaries from the 1890s influenced Christianity in South Africa, and how in the 1920s African ports became nodal points in the spread of Garveyite movements across the Atlantic to large parts of Africa (Campbell, 1995). We have learned not only how European states used various coercive means to push Africans – especially younger males – into periods of wage labor away from home, but how Africans themselves developed migratory networks and social institutions to make life in cities, mines, or distant farms more palatable (Manchuelle, 1997). That Africans educated at missions sometimes fought for more voice or control of churches or sought to forge their own social institutions and norms that reflected what they brought to the missionary encounter has become a focus of inquiry (McKittrick, 2003).

Having for some time now moved beyond the need to recuperate a singular African voice, historians have devoted attention to questions of gender, generation, and class – examining both tensions within African societies and connections across social cleavages. Some have argued that colonial rule enforced a view of male and female roles more rigid than had been the case before colonization, but in any case research has revealed struggles – before, during, and after the colonial period – over questions of the norms governing the roles of men and women, young and old (Thomas, 2003). And if slavery, pawnship, manumission, and emancipation have been themes of historical scholarship for some time, more attention is being paid to the ways in which unequal relationships are pushed and pulled from both above and below (Glassman, 1995).

If we circle back to where we began – how national independence shaped the quest for a usable past – we find some historians turning away from their supposed role as organic intellectuals of the nation to point to forms of political affinity that are both sub- and supranational. The politics of the 1940s and 1950s turns out to be more complicated than a quest for the nation in the face of an obdurate colonialism. If some Asante leaders made 'national' claims in the same term as the movement of Kwame Nkrumah, those claims were for a different nation – a region of Ghana. Almost all leaders of French Africa in the decade after World War II sought to make French citizenship into a meaningful category and to transform colonial empire into a federation or confederation of equal states, including France (Allman, 1993; Cooper, 2014). Nor were colonial regimes fixed targets. In fact, it was their attempt to regain legitimacy and efficiency that set the stage for an escalation of claim making by African political and social movements that colonial regimes, in the end, could not contain at an acceptable cost.

If the nation-state was not necessarily the aim of political movements in the 1950s, it was what they could get, and the new rulers of Africa largely erased the complexities of postwar political history in order to situate themselves as fathers of the nation. But the quest to complement 'flag independence' with projects of national economic development and forging a national culture ran into the problem that elites did not necessarily have the resources to make citizens see the rewards of political participation. Some citizens, particularly youth, were caught between their own desire to participate in a national culture and their interest in transnational cultural tendencies and radical politics that governing elites found threatening (Ivaska, 2011). Scholars are uncovering more diverse and richer forms of imagination and communication

among different parts of African populations than opposition of colonizer and colonized, African and European, white and black would suggest (White, 2000).

Conclusion

The Senegalese novelist and filmmaker Ousmanne Sembene accused historians of being 'chronophages' – eaters of time. They impose, he insists, a one-dimensional, progressive view of time on the unruliness of people's experience. Yet the idea that other notions of past give a more varied picture than professional history depends on aggregating different visions of the past – a quintessentially 'academic' operation. The griot or the lineage elder may also recount the past to serve narrow, presentist concerns. Historical scholarship has complicated unilinear narratives as well as imposed them.

The doing of history introduces a tension between a historian's imagination rooted in the present and the fragments of the past that appear in all their elusive vigor, in interviews, letters, newspaper articles, and court records. The data of history are not created neutral or equal: archives preserve certain documents but not others, and the tellers of oral tradition remember some narratives and forget others.

African history has offered a glorious national past that points to a national future, and it has offered an ethnicized past projected backward in time. But history can also be read as process, choice, contingency, and explanation. Since World War II, historical scholarship has provided a sense of the possibilities which mobilization can open up – and an awareness of the constraints that have made and make it so difficult for African states and societies to make their way in the world.

See also: Africa, Sociocultural Overviews: West Africa; African Studies: Culture; Central Africa: Sociocultural Aspects; Colonialism, Anthropology of; Development: Social-Anthropological Aspects; Diaspora; East Africa: Sociocultural Aspects; Historical Thought and Historiography: Sub-Saharan Africa; Southern Africa: Sociocultural Aspects.

Bibliography

Allman, J., 1993. The Quills of the Porcupine Asante Nationalism in an Emergent Ghana. University of Wisconsin Press, Madison.
Ambler, C., 1988. Kenyan Communities in the Age of Imperialism: The Central Region in the Late Nineteenth Century. Yale University Press, New Haven, CT.
Anderson, D., 2005. Histories of the Hanged: Britain's Dirty War and the End of Empire. Norton, New York.
Austin, G., 2005. Labour, Land, and Capital in Ghana: From Slavery to Free Labour in Asante, 1807–1956. University of Rochester Press, Rochester.
Balandier, G., 1951. La situation coloniale: approche théorique. Cahiers Internationaux de Sociologie 11 (1951), 44–79.
Barry, B., 1998. Senegambia and the Atlantic Slave Trade (A.K. Armah, Trans.). Cambridge University Press, Cambridge, UK.
Berry, S., 1993. No Condition Is Permanent: The Social Dynamics of Agrarian Change in Sub-Saharan Africa. University of Wisconsin Press, Madison, WI.
Blyden, E.W., 1887. Christianity, Islam and the Negro Race. Whittingham, London.
Campbell, J., 1995. Songs of Zion: The African Methodist Episcopal Church in the United States and South Africa. Oxford University Press, New York.
Casely, H.J.E., 1903/1970. Gold Coast Native Institutions. Cass, London.
Cohen, D.W., 1994. The Combing of History. University of Chicago Press, Chicago.
Conklin, A., 1997. A Mission to Civilize: The Republican Idea of Empire in France and West Africa. Stanford University Press, Stanford, pp. 1895–1930.
Cooper, F., 1996. Decolonization and African Society: The Labor Question in French and British Africa. Cambridge University Press, Cambridge, UK.
Cooper, F., 2014. Citizenship between Empire and Nation: Remaking France and French Africa, 1945–1960. Princeton University Press, Princeton, NJ.
Cooper, F., Isaacman, A., Mallon, F., Roseberry, W., Stern, S., 1993. Confronting Historical Paradigms: Peasants, Labor, and the Capitalist World System in Africa and Latin America. University of Wisconsin Press, Madison, WI.
Dike, K.O., 1956. Trade and Politics on the Niger Delta. Clarendon Press, Oxford, UK.
DuBois, W.E.B., 1946. The World and Africa: An Inquiry into the Part Which Africa Has Played in World History. Viking, New York.
Ferreira, R., 2012. Cross-Cultural Exchange in the Atlantic World: Angola and Brazil During the Era of the Slave Trade. Cambridge University Press, Cambridge, UK.
Glassman, J., 1995. Feasts and Riot: Revelry, Rebellion, and Popular Consciousness on the Swahili Coast, 1856–1888. Heinemann, Portsmouth, NH.
Green, T., 2012. The Rise of the Trans-Altantic Slave Trade in Western Africa, 1800–1589. Cambridge University Press, Cambridge, UK.
Hanretta, S., 2009. Islam and Social Change in French West Africa History of an Emancipatory Community. Cambridge University Press, Cambridge, UK.
Hawthorne, W., 2010. From Africa to Brazil: Culture, Identity, and an Atlantic Slave Trade, 1600–1830. Cambridge University Press, Cambridge, UK.
Herskovits, M., 1944. Native self-government. Foreign Affairs 22, 413–423.
Herskovits, M., 1958. The Myth of the Negro Past. Beacon Press, Boston, MA.
Heywood, L., Thornton, J., 2007. Central Africans, Atlantic Creoles, and the Foundation of the Americas, 1585–1660. Cambridge University Press, Cambridge, UK.
Horton, J.A.B., 1868/1969. West African Countries and People. Edinburgh University Press, Edinburgh.
Ivaska, A., 2011. Cultured States: Youth, Gender, and Modern Style in 1960s Dar es Salaam. Duke University Press, Durham, NC.
Jerven, M., 2010. African growth recurring: an economic history perspective on African growth episodes, 1690–2010. Economic History of Developing Regions 25 (2), 127–154.
Latham, M., 2000. Modernization as Ideology: American Social Science and 'Nation-building' in the Kennedy Era. University of North Carolina Press, Chapel Hill, NC.
Lydon, G., 2009. On Trans-Saharan Trails: Islamic Law, Trade Networks, and Cross-cultural Exchange in Nineteenth Century Western Africa. Cambridge University Press, Cambridge.
Manchuelle, F., 1997. Willing Migrants: Soninke Labor Diasporas, 1848–1960. Ohio University Press, Athens, OH.
Matory, J.L., 1999. The English professors of Brazil: on the diasporic roots of the Yoruba nation. Comparative Studies in Society and History 41, 72–103.
McKittrick, M., 2003. To Dwell Secure: Generation, Christianity, and Colonialism in Ovamboland. Heinemann, Portsmouth, NH.
Meillassoux, C., 1975. Femmes, greniers et capitaux. Maspero, Paris.
Mudimbe, V.Y., 1988. The Invention of Africa: Gnosis, Philosophy, and the Order of Knowledge. Indiana University Press, Bloomington, IN.
Nwokeji, G.U., 2010. The Slave Trade and Culture in the Bight of Biafra: An African Society in the Atlantic World. Cambridge University Press, Cambridge.
van Onselen, C., 1996. The Seed Is Mine: The Life of Kas Maine, a South African Sharecropper, 1894–1985. Hill and Wang, New York.
Peel, J.D.Y., 1983. Ijeshas and Nigerians: The Incorporation of a Yoruba Kingdom, 1890s–1970s. Cambridge University Press, Cambridge, UK.
Peel, J.D.Y., 2003. Religious Encounter and the Making of the Yoruba. Indiana University Press, Bloomington.
Peterson, D., 2004. Creative Writing: Translation, Bookkeeping, and the Work of Imagination in Colonial Kenya. Heinemann, Portsmouth, NH.
Pletsch, C., 1981. The three worlds, or the division of social scientific labor, circa 1950–1975. Comparative Studies in Society and History 23, 565–590.
Ranger, T., 1968. Connexions between primary resistance movements and modern mass nationalism in East and Central Africa. Journal of African History 9, 437–453, 631–641.
Ranger, T., 2004. Nationalist historiography, patriotic history and the history of the nation: the struggle over the past in Zimbabwe. Journal of Southern African Studies 30, 215–234.
Robinson, D., 1985. The Holy War of Umar Tal: The Western Sudan in the Mid-Nineteenth Century. Oxford University Press, Oxford.
Robinson, D., 2000. Paths of Accommodation: Muslim Societies and French Colonial Authorities in Senegal and Mauritania, 1880–1920. Ohio University Press, Athens, OH.
Rodney, W., 1972. How Europe Underdeveloped Africa. Bogle-L'Ouverture, London.

Sibeud, E., 2002. Une science imperiale pour l'Afrique? La construction des saviors africanistes en France 1878–1930. Editions de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.
Thomas, L., 2003. Politics of the Womb: Women, Reproduction, and the State in Kenya. University of California Press, Berkeley, CA.
Tilley, H., 2011. Africa as a Living Laboratory: Empire, Development, and the Problem of Scientific Knowledge, 1870–1950. University of Chicago Press, Chicago, IL.
UNESCO, 1981–93. General History of Africa. University of California Press, Berkeley, CA.
Vansina, J., 1965. Oral Tradition: A Study in Historical Methodology. (H.M. Wright, Trans.) Aldine, Chicago.
White, L., 2000. Speaking with Vampires: Rumor and History in Colonial Africa. University of California Press, Berkeley, CA.
White, L., Miescher, S., Cohen, D.W., 2001. African Words, African Voices: Critical Practices in Oral History. Indiana University Press, Bloomington, IL.

Africa's Environment: A Stressed Biogeographical and Cultural Landscape

Delali Benjamin K Dovie, Department of Geography and Resource Development, University of Ghana, Legon, Ghana; School of Animal, Plant and Environmental Sciences, University of the Witwatersrand, Johannesburg, South Africa; and Regional Institute for Population Studies, University of Ghana, Legon, Ghana

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P.D. Little, volume 1, pp. 236–240, © 2001, Elsevier Ltd.

Abstract

Humans are an important component of the biophysical environment in Africa, interacting with its resources through prisms of knowledge structures to harness and manage resources sustainably. A review of the human sphere (e.g., population and migration, land and land use, agriculture, and food security) and the ecosphere (e.g., environmental change, biodiversity, water, and coastal resources) is undertaken. The review suggests an environment which is generally degraded, and blamed on biophysical processes and policy failures. Mainstreaming policies and decision-support systems of different sectors and development processes would help reduce conflicts in resource use toward maintaining the integrity of Africa's environment.

Introduction

Africa has an environment which is endowed with a wealth of natural resources, including minerals, land, biological diversity, wildlife, forests, fisheries, and water, and shaped by various factors that may be classified as human or anthropogenic, and natural. The factors include, but are not limited to, population growth, widespread land degradation and desertification, overfishing, deforestation, loss of biodiversity and loss of arable and grazing land, pollution and deteriorating air and water quality, worsening sanitation, declining soil productivity, and depletion of freshwater resources (UNEP, 2003). These issues have far-fetching implications for human welfare, sustainable development, resilient livelihoods, and efforts to eradicate poverty (WRI, 2005). In Africa, society perceives and interprets its biophysical surroundings through its own way of life and culture, within which the physical environment affects humans. The relationship between humans and the environment has often produced two schools of thought: natural against artificial, human or cultural, based on arguments that the activities of humans are strongly controlled or conditioned or even determined by the biophysical environment (Farina and Napoletano, 2010).

Through processes of adaptation, cultural appraisal, innovation, and experimentation, humans are able to influence their environment. For example, in Africa, people shape much of 'the wild' and many species have been carefully nurtured by people, albeit less intensively than those cultivated (Bell, 1995). Hence the surface of the earth as modified by human action represents a cultural landscape defined as geographic areas in which the relationships between human activity and the environment have created ecological, socioeconomic, and cultural patterns and feedback mechanisms that govern the presence, distribution, and abundance of various life forms (Farina and Napoletano, 2010; Taylor, 2002). A review of Africa's environment as a cultural landscape is pursued from two compartments, namely the human sphere and the ecosphere, toward understanding how these are shaping the African environment.

The Human Sphere

Population, Migration, and the Environment Nexus

Africa's annual population growth rate is estimated at 2.3% (UNFPA, 2011), rapidly growing than in other regions of the world although with known consequences for adaptation and sustainable development efforts. As home to an estimated 900 million people, it is projected to grow to 1.2 billion by 2025 and to 2 billion by 2050. Sub-Saharan African countries would be growing more than any other continent (UNPD, 2011). Over 46% of the 2.4 billion people who are projected to be added to the world by 2050 will come from sub-Saharan Africa, contributing 77% of the total increase in global population by 2100 (UNPD, 2011). The growing population would place increasing strains on all aspects of the environment, thus calling for meshing population and economic growth with environmental protection. Urban areas will experience most of the pressures in the coming years, resulting in pressure on urban managers to expand existing services particularly to the poor and those remote from such services. It is expected that there would be gradual but increased displacement and migration of human population as a result of environmental change (e.g., overpopulation, overfishing, land degradation, deforestation, climate change, land use, and land cover change). Movement of people within the continent has dominated cross-border migration (Dovie et al., 2012).

Traditional arguments portray migration and migrants as threats to human security, survival, and development, while the development community argues that human mobility is a development tool and likely to contribute immensely to development than not (Landau and Vigneswaran, 2007). Land use and land cover changes are major outcomes of migration in Africa, as migrants often converge in fragile environments to adopt means to diversify or self-insure risk involving collective decision-making. Such decisions can also contribute to ensuring that usage of environmental resources is sustainable. Migrant practices such as illegal mining have polluted several rivers, enhanced deforestation, displaced communities and farms, and augmented floods, and led to overfishing and

a source of conflicts. Cross-border movement from land degradation has created basic natural resource scarcity. Remittances remain a significant outcome of migration, and therefore are important social safety nets to cushion the poor and offsetting additional burden on environmental resources (Dovie et al., 2012). The direct spending of remittances by families contributes to minimizing poverty outcomes and environmental degradation. Generally, it is expected that the high human mobility in Africa would result in rapid social and economic change with tremendous effects on the ecologically vulnerable or fragile biophysical environments (Dovie et al., 2012). Changes in rainfall patterns and stream flow will have a direct impact on human settlements, so many of which are already unable to provide reliable, affordable services.

Land and Land Use

Approximately 22% of the total land area of Africa is under forest, 43% is characterized as extreme deserts, and 57% is vulnerable to desertification. While 21% of the total land area is suitable for cultivation, it is estimated that some 500 million hectares of land in Africa has been lost to soil degradation of which 65% could be described as agricultural land. Overgrazing is the leading cause of land degradation in Africa, followed by crop production-related activities and vegetation removal (UNEP, 2003). Some 202 million hectares of land can be cultivated and put to various uses in Africa, thus representing nearly half of the world's usable uncultivated land (Byamugisha, 2013). Human activities (land uses) that directly affect the environment constitute proximate sources of change, connecting the cultural landscape and human purpose or intent which reflect human goals. Agricultural expansion, wood extraction, and infrastructure extension are the leading causes of the change underlain by economic factors, policy, and institutional, technological, cultural, and demographic factors (Geist and Lambin, 2002). It is estimated that 79% of the total traditional energy consumed in developing countries is fuelwood, with over 90% dependence in sub-Saharan Africa (UNEP, 2003).

The unfamiliarity and haphazard appearance of domestic gardens in Africa is often overlooked for the rich complexity of these production systems. Forests and other vegetation formations that extend from the immediate neighborhood of settlements have been assumed to be self-regulating wastelands and subjected to abuse (Bell, 1995). However, because land is central to the existence of humans as a source of major ecosystem services such as food, fiber, and medicines, the sustainable management of land can enhance the health of people through increased access to a wider range of ecosystem services (UNEP, 2013). Yet inappropriate land use practices, resulting in loss of vegetation cover, excessive use of chemical pesticides and fertilizers, salinization, contamination by heavy metals, and soil depletion adversely affect land productivity, subsequently leading to land degradation which impacts negatively on food and nutrition security (UNEP, 2003). Yield losses from soil erosion in sub-Saharan Africa are between 2% and 40%, and some changes in land use are known to alter the ecology of human diseases, making people more vulnerable to infections (Collins, 2001). The Africa Health Strategy 2007–15 (AU – CAMH, 2007) emphasized that environmental factors and degradation are not sufficiently addressed in health, which is adding to disease burden factors in Africa and should be prioritized in health planning on the continent. The mounting costs of floods in Africa are partly due to land use changes that push populations and assets into vulnerable areas. Therefore urbanization and land use planning can also be drivers of water impacts on society and vice versa.

Resource Use, Indigenous Knowledge, and Livelihoods

Human societies have for generations interacted with their environments mediated through basic knowledge structures. One of the results of this interaction is the immense genetic diversity within traditional farming systems today. In Africa, kinship ties for instance form the basis for local level decision-making on aspects of livelihoods and resource management, dictated by knowledge and societal norms (Larson, 1998). This knowledge, which is often indigenous in nature or in origin, is generally poorly documented, leading to fears that the long experiences possessed by many traditional and local communities in the use and management of the cultural landscape such as conservation of natural resources would be lost. For centuries, rural peoples have relied on nature for their livelihoods. Farmers have managed genetic resources for as long as they have cultivated crops, and the historical importance of local people as having the knowledge and skill to manage biodiversity through their biocultural diversity is crucial (Moore et al., 2002). Historically, some of the knowledge is administered through cultural factors such as taboos that placed limitations on resource use. Local and indigenous groups could therefore provide much of the intellectual raw material for the sustainable harvesting and management of natural resources.

Thus in Africa, key livelihoods, which comprise capabilities, assets (stores, resources, claims, and access), and activities required for a means of living (Carney et al., 1999), are a product of knowledge systems, resource tenure, and use. Livelihoods are therefore seen to consist of a range of on-farm and off-farm activities that together provide a variety of procurement strategies for food, cash, and in-kind income. It could also be informal or formal, and comprising entitlements and capabilities. As a result, it has been argued that livelihoods and natural resource use are inextricably linked in Africa, often quite diverse and engaging capabilities, skills, and know-how to offset risks that lead to livelihood security (Dovie et al., 2005). It includes adequate access to food, potable water, health facilities, educational opportunities, housing and time for community participation, and social integration. Natural capital comprising land, water, air, rivers, lagoons, oceans, and forests, and wetlands have been major sources of drivers of livelihoods in Africa, providing short-term coping mechanisms in response to shocks and stresses. They also have become adaptive strategies, including the long-term changes and adjustments people make in their livelihood systems based on various knowledge systems in order to respond to difficult circumstances.

Agriculture and Food Security

The importance of agriculture to provide food security comes in two ways, firstly producing the food people eat, and secondly

(perhaps even more important) providing the primary source of livelihood for 36% of the world's total workforce. In sub-Saharan Africa, two-thirds of the working population still makes their living from agriculture (ILO, 2007). Sub-Saharan Africa and Asia have the most share of food insecurity represented by almost 90% of the undernourished people in the world. Almost 870 million people were described as chronically undernourished in 2010–12, where the vast majority live in developing countries having about 850 million people estimated to be undernourished (FAO, WFP, IFAD, 2012). Agriculture is the largest contributor to Gross Domestic Product (GDP) in Africa as the biggest source of foreign exchange, accounting for about 40% of the continent's foreign currency earnings, and the main generator of savings and tax revenues. About two-thirds of manufacturing value-added is based on agricultural raw materials. Agriculture remains crucial for pro-poor economic growth in most African countries as rural areas support 70–80% of the total population. This means that improvements in agricultural performance have the potential to increase rural incomes for large numbers of people to lift them out of poverty and food insecurity. In spite of efforts in place to reverse the undernourished trend in sub-Saharan Africa, an estimated 25% of the African population still remains undernourished, as average protein supply in Africa is estimated as 20% lower than the world's average (FAO, 2011). Yet, climate change is expected to increase the number of undernourished people by between 35 and 170 million people in 2080, depending on projected development paths (Shah et al., 2008). Overall, approximately 40% of sub-Saharan Africa countries will be at risk of significant declines in crop and pasture production due to climate change (Shah et al., 2008).

Almost three decades ago, it was argued that several years of focusing on commodity crops across the globe had not ensured food and livelihood security for the world's poor, and that the focus on commodity-based agriculture threatened millions of people mostly in the developing world (Bell, 1995). It was shown that official agricultural development policy and research were dominated by global scenarios from the developed world, as a result interrupting the survival strategies of resource-poor regions including Africa. As such, the policy overlooked the role of agriculture within the wider portfolio of household livelihood strategies. Thus, agriculture and food production were seen as independent of the broader landscape rather than a strong connection with the wild through agro-ecosystems to domesticated landscapes and homestead gardens in sub-Saharan Africa, and that, inaccurate estimates of food insecurity still persist in Africa (UNDP, 2003). Several models for implementing food systems represent global scenarios and do not elicit the real local and national issues. Therefore, none of the variety of approaches provided definite analytical frameworks that are specifically designed to establish the implications of differences in the sociopolitical economy that influence or are influenced by food systems in local, rural landscapes in Africa. The global scenarios failed to

- account for small and micro-scale activities, subsistence production, and products without formal markets,
- recognize the impacts of social, economic, and demographic heterogeneity, and
- account for the effects and impacts of global environmental change, thus perpetually depriving the African continent of appropriately sustainable food security strategies.

The Ecosphere

Global Environmental Change and Climate Change

Changes in the environment are a process comprising land use and land cover changes, land degradation, desertification, overgrazing, overfishing, population growth, loss of biological resources, deforestation, greenhouse gas emissions, and atmospheric pollution, among others. Of these, atmospheric pollution has come under intense scrutiny in Africa for its implications for climate variability, climate change, and air quality although Africa's contribution to the global pool of greenhouse gas emissions is still relatively low. It is estimated that Africa contributed to only 3.57% of the world's total carbon dioxide emissions in 1988, yet remains the more vulnerable continent to the impacts of climate change (IPCC, 2007; UNEP, 2003). Climate change is attributed to both natural and human-imposed activities mainly through over-consumption of the earth's resources. The Fourth Assessment of the Inter-Governmental Panel on Climate Change (IPCC) reveals that Africa is warming faster than the global average and that it is likely to warm by an average 3–4 °C this century. Extreme weather events such as droughts and floods are also projected to be more frequent (**Figure 1**), intense, or severe with dire consequences for human life and health (IPCC, 2007).

In Africa, impacts of climate change are most felt as a result of widespread poverty, recurrent droughts, inequitable land distribution, and overdependence on rain-fed agriculture. Both climatic and socioeconomic stress shape the processes influencing local vulnerability to climate variation and change which are inherently dynamic, yet the socioeconomic context is seldom explicitly considered, either in local assessments of vulnerability to climate change or in local climate adaptation strategies (IPCC, 2007). Weak planning taking the form of reactive and sector-based responses has been attributed to improper reflection of stresses and the capacity to adapt to change. The high dependence of large populations on climate-sensitive natural resource sectors has increased Africa's vulnerability. Changes in weather conditions ultimately constitute climate variability which adversely impacts agriculture, often translating into income and food insecurity for least resilient communities (UNEP, 2013). Adaptation to climate change is complicated by limited economic diversification, poorly resourced governments and lead agencies, and the virtual absence of insurance cover for climate-sensitive investments such as agriculture. Thus if agricultural production in the low-income developing countries of Africa is adversely affected by global environmental change (e.g., climate change), the livelihoods of large numbers of the rural poor and the vulnerable will be put at risk and their vulnerability to food insecurity increased.

The Interdepartmental Working Group on Climate Change of the United Nations Food and Agricultural Organization estimates that Africa would face a 2–9% decrease of GDP from agriculture, attributed to climate change (FAO, 2007). The food

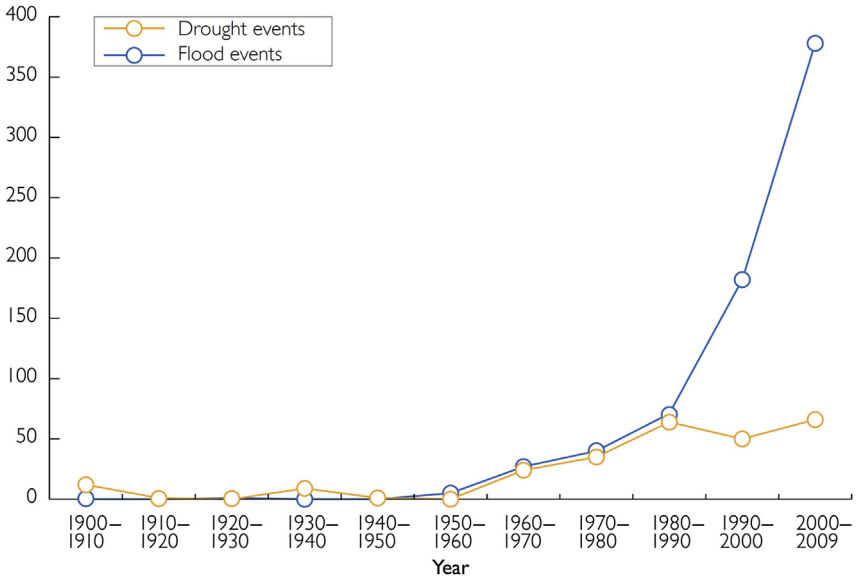

Figure 1 Trends of some extreme weather events associated with climate change. Source: UNEP, 2010. The Fourth Global Environment Outlook: Environment for Development, UNEP-WCMC. UNEP, Nairobi.

security implications of changes in agricultural production patterns and performance as a result of changes in rainfall patterns due to the overreliance on rain-fed agriculture on the African continent are of two kinds:

- First, impacts on the production of food will affect food supply while higher yields in temperate regions would offset lower yields in tropical regions. However, limited financial capacity to trade in many low-income countries and high dependence on own production to cover deficits in food production may not be possible to offset declines in local supply. The outcome would be the increased dependence on food aid.
- Second, the impacts on all forms of agricultural production will affect livelihoods and access to food and that subsistence agriculture in Africa risk having their security undermined, together with the food system.

Human settlements would be affected by changes in rainfall patterns and stream flow, so many of which are already under stress and experiencing shortfalls in the delivery of services. While climate change will impact the capacity to supply potable water, it will also reduce the capacity to dispose of wastes from large urban communities and accompanied by increased costs of appropriate infrastructure and logistics to manage such wastes (Sadoff and Muller, 2009). Groundwater recharge was forecast to decrease by the 2050s, and by more than 70% in Southwestern Africa. The costs of adapting to existing urban water infrastructure in Africa have been estimated at between US$ 1050 million and US$ 2650 million annually (Muller, 2007).

Biodiversity

The cultural landscape created biotic niches that maintain precise relationships with biodiversity and without this diversity the ability to adapt to changing needs and conditions of society and biodiversity would be lost. Africa has a large heritage of biodiversity, forming the basis of the region's natural wealth and development and hosting six of the world's 25 biodiversity hotspots. These resources also have global importance for the world's climate and for the development of agricultural or industrial activities. The region has more than 50 000 known plant species, 1500 species of birds, and 1000 mammals (UNEP, 2003). Africa's rich and varied biological diversity has considerable potential to enhance the productivity of other sectors, contributing immensely to attaining the Millennium Development Goals (MDGs). Africa's ecosystems are a host of a variety of fruits, vegetables, honey, spices, oils, bush meat, fish, edible worms, and mushrooms, contributing to food security on the continent and largely constituting the bulk of non-timber forest products (Dovie et al., 2005).

Recent reports (e.g., the *Africa Environment Outlook*) confirm that the wealth of Africa's biodiversity is being rapidly lost from uncontrolled exploitation and fragmentation of natural habitats largely attributable to rapid population growth and expanded agriculture. Increasing rates of deforestation, overgrazing, soil erosion, and desertification threaten biodiversity and the accompanying benefits to both the environment and humanity (UNEP, 2013). Increased dependence on narrow-range species, overharvesting, and climate change also contribute to ecosystem degradation and biodiversity loss and destabilizing ecosystem processes such as the predator–prey relationships that may create favorable conditions for disease outbreak (UNEP, 2013). Gitay et al. (2002) projected that climate change will halve Africa's biodiversity, affecting individual organisms, populations, species distributions, and ecosystem composition and function. This would be direct (e.g., through increases in temperature and changes in precipitation and in the case of marine and coastal ecosystems also changes in sea level and storm surges) and indirect (e.g., through climate changing the intensity and frequency of disturbances such as wildfires). Thus, significant declines in the

provision of wild foods by a variety of ecosystems, and further impacts can be expected as the world climate continues to change as it is predicted that 81–97% of the suitable habitats will decrease in size or shift owing to climate change. By 2085, between 25% and 42% of the species' habitats are expected to be lost altogether.

The dependence of human populations on ecosystem resources in Africa is well documented, harboring rich areas of important biodiversity including some of the richest tropical rain forests (e.g., the Congo). For example, Dovie et al. (2005) observed in a study of ecosystem resources (i.e., natural resources and agrodiversity including farming) that in the absence of jobs and confronted with high migrant labor, households with open access to natural resources derived more benefits from land-based livelihoods than cash income streams (i.e., 57.5% as against 42.5%). African ecosystems therefore hold key importance for bioprospection even on commercial scales which are significant in terms of livelihoods yet they are not always reflected in market prices (Dovie et al., 2005; Lawes et al., 2004). Forest resources in Africa are therefore important to the livelihoods of many communities, and more specifically in Western, Central, and Eastern Africa where forest resources still abound. On average, forests account for 6% of GDP in Africa and leading the world statistics (UNEP, 2003).

Water Resources

Africa's water resources are characterized by extreme spatial and temporal variability culminating in only about 4% of its renewable freshwater resources being used (UNEP, 2003). Batisani and Yarnal (2010) observe that variations in internal renewable water resources owing to a range of natural and anthropogenic factors such as uneven distribution of rainfall across the continent, its unpredictability, and high evaporation losses tend to influence the natural variation. It is estimated that 25% of the population (approximately 200 million people) are faced with water stress, with more countries expected to experience high risks in the future (Ludi, 2009). This may, in turn, lead to increased food and water insecurity for at-risk populations, and undermining growth and development. Africa's estimated annual per capita water availability stood at 4008 m^3, compared to the global average of 6498 m^3 in 2008 (UNEP, 2010). Water scarcity is projected to increase from 47% in 2000 to 65% in 2025 (Bates et al., 2008). The African continent holds up to 660 000 km^3 estimated groundwater and more than 100 times the estimated annual renewable freshwater resources of the continent (MacDonald et al., 2012). The largest groundwater resources are located in the large sedimentary aquifers in parts of northern Africa and mostly Libya, Algeria, Egypt, and Sudan, which is crucial to relatively dry and arid areas. Of particular concern is the situation in North Africa which faced limitations in renewable freshwater decades ago, while struggling to meet its food requirements based on the available water resources. As one of the driest regions in the world, the region contains only 1% of the world's freshwater resources (Qadir et al., 2007).

The majority of the population of sub-Saharan Africa still lacks safe drinking water, accounting for nearly one-third of the global 884 million people who have no access (UNEP, 2010). By 2010, only 23 countries were on track to meeting the safe drinking water MDG target (**Figure 2**). Although the actual number of people using improved drinking water sources had increased by 11% relative to 1990, only 66% had access to safe water (WHO and UNICEF, 2012). The United Nations Development Programme warns that the progress in human development achieved over the last decade may be slowed down or even reversed by threats to water security from climate change. The impacts of climate change could, by 2080, increase the number of people facing water scarcity by 1.8 billion (UNDP, 2008), with Africa likely to take the bigger share. In some drier areas of sub-Saharan Africa, reductions in stream flow of more than 50% are confidently predicted and many perennial streams would become seasonal, having limited flow and others drying up permanently (Sadoff and Muller, 2009). With water as the major platform for adaptation, climate change could severely impact on human activities through water and food insecurity while bringing about permanent changes in ecosystems through disruption in ecosystem services. Thus a threatened water resource base provides conditions for the disruption in biophysical properties, and social networks and capital on the African continent, as livelihoods of many rural people depend on the vagaries of rainfall. Thus the progress made in social and economic welfare and gains over many years are at risk of being eroded within shorter time periods through water-related stresses. For transboundary water resource networks such as river basins, water supply challenges would have negative consequences for social relations such as conflicts and migration.

Rapid urbanization especially of cities and increasing industrialization, water use, and consumption would become greater and in the process increase competition among the various uses of both human and ecosystem origins. Water use is generally greater than the amount consumed because of conveyance and other inefficiencies such as return flows from agriculture of up to 20–40%. Water usage for agriculture is estimated at 88% contributed largely by 40–60% of the region's irrigation which is lost through seepage and evaporation (UNEP, 2003). Lack of adequate water storage is characteristic of many African countries especially those in arid regions and even where there is significant storage, as in Southern Africa (the Zambezi), the remoteness of settlements undermines water security unless extensive transport and distribution infrastructure is put in place, which would be very costly. This is similar to Eastern Africa's lakes where there is natural storage to store large volumes of water but not easily accessible to the communities that need this water most (Sadoff and Muller, 2009). Problems on access and use of water resources should also be blamed on political, social, and economic factors because they fail to consider the influences of other sectors of the environment such as forestry.

Coastal Zones

Coastal zones contain valuable ecosystems for both humans and the environment, contributing significantly to the cultural landscape. They have become the preferred choice for settlement of humanity, hosting most of the world's 20 largest mega cities and having enormous infrastructure, and

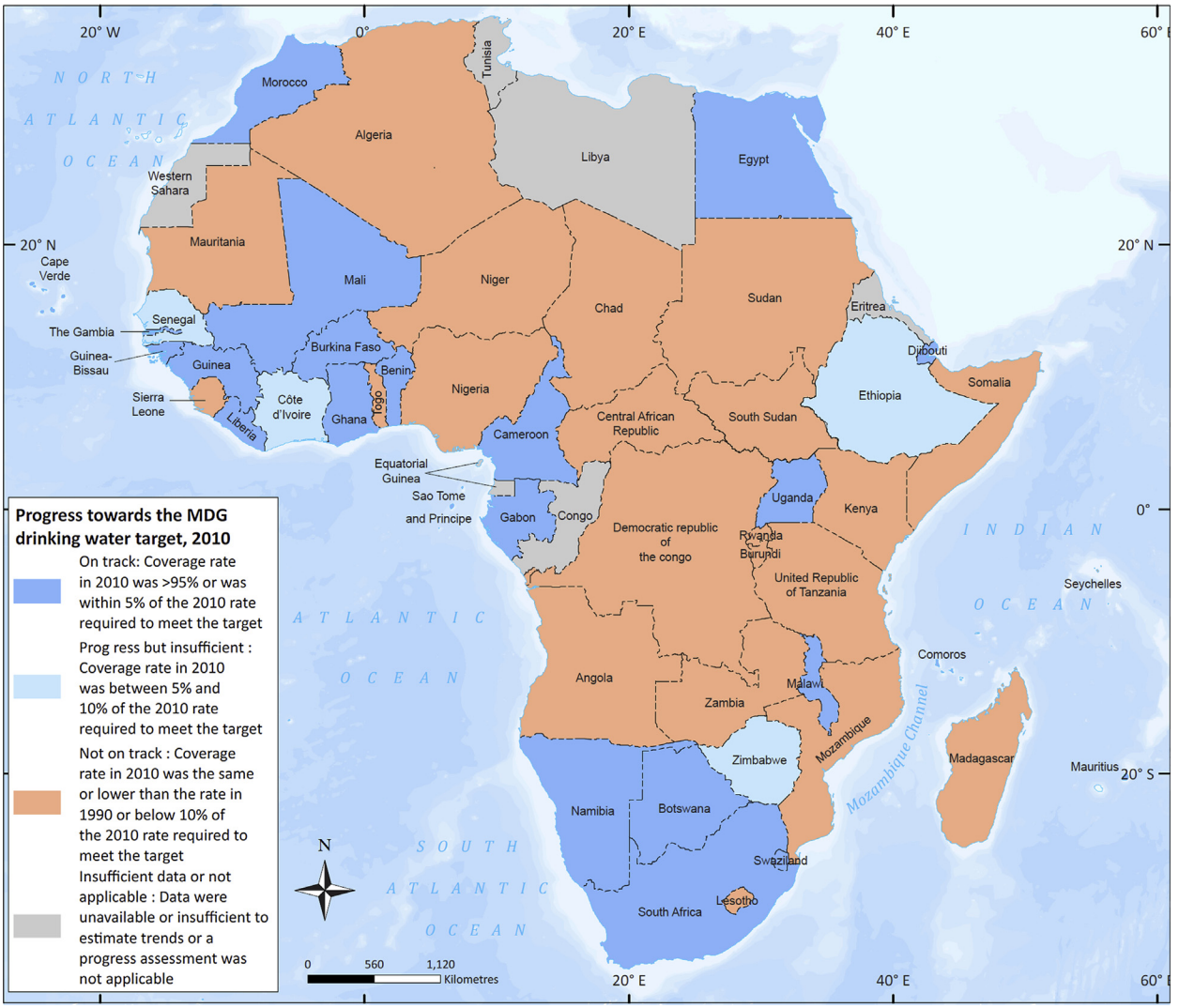

Figure 2 Progress toward improved drinking water. MDGs, millennium development goals. Source: WHO and UNICEF, 2012. Progress on Sanitation and Drinking-Water – 2010 Update. World Health Organization and UNICEF, Geneva.

generate significant amounts of economic activities (Small and Nicholls, 2003). Africa's coastal environments contribute significantly to the economies of many countries, mainly through fishing, tourism, and as industrial destinations. Coastal and marine resources contribute considerably to the revenue of countries of the region through marine fish catch exports and associated employment although facing increasing threat from development-related activities. Up to 38% of the African coastline of 40 000 km, including 68% of marine-protected areas, is considered to be under a high degree of threat. Uncontrolled urbanization and industrialization including the petrochemical industry are a major cause of coastal degradation. There have been cases where marine pollution has even reached toxic levels from industrial pollution, mining, and oil exploration. Although the level of industrial development on the continent is still relatively low compared to other regions of the world, the rate is accelerating along the coastal zones especially to meet the demand from tourism. Significant growth has been expected to take place in heavily populated coastal regions. In Africa in particular, as poverty drives the population to urban coastal areas resulting in significant increases in the stress on coastal resources, there will be increased risk of flooding, and increased frequency and intensity of storm surge events (IPCC, 2010).

The Fourth Assessment Report (2007) of the IPCC predicts "heavy precipitation events, which are very likely to increase in frequency, will augment flood risk" (IPCC, 2007). These floods will affect life and livelihoods in human settlements in all areas, e.g., coastal zones, river deltas, and mountains. Flooding is also increasing in urban areas, causing severe problems for poor people (Bates et al., 2008; IPCC, 2010):

- Most infrastructure, settlements, and facilities located on or near the shore will be affected by sea level rise, coastal erosion, and other coastal hazards, compromising the socioeconomic well-being of island communities and states.

- Accelerated beach erosion and degradation of diversity of ocean lives will all have impacts on incomes from fishing and tourism.

Climate change will influence the occurrence of extreme events and vulnerability with increased risks such as flooding. Coastal areas would face inundation from sea level rise due to climate variability, wiping off development gains. Inundation of many major cities of Africa from coastal erosion and extreme events and the degradation of marine ecosystems in parts of Africa is expected, with costs of adaptation to annual sea level rise estimated to be in the range 5–10% of GDP (IPCC, 2007).

Mainstreaming Environment, Research, Policy, and Development

Research and Development

Knowledge generation from innovative technology and development research remains a huge challenge to the continent of Africa. For the environment, concerns over ecosystem degradation and climate change have undermined political commitment to the total environment, which is reflected in weaknesses to invest in trans-sector and transdisciplinary research and development. Thus,

- Although trans-sector research is important because it considers interrelationships, networks, and linkages in its goals, it is a costly venture to bridge knowledge gaps and development. There is no doubt that the developed world or richer countries have relatively advanced to extensively engage science and technology in environment and development issues.
- In Africa, the complex origins, problems, and livelihoods of the local population are such that technology and development itself is in turn complicated. As a result, environmental dimensions of such technology and development have been of least concern until the past few years after the hosting of the sustainable development summit in Johannesburg in 2002 which also emphasized the MDGs. The focus of advancing technology with sustainable environmental practice remains at the macroeconomic level for lack of pragmatic research. It is therefore not surprising that several technologies have left in their trails environmental concerns that local and poor people have to continuously deal with.
- A great deal of information and innovation will be required to understand environmental concerns to ensure that socioeconomic development matches environmental sustainability to save Africa's environment.

Policy and Policy-Making Processes

Sub-Saharan African countries are known to be least prepared to respond to extreme events, partly because policies on economic development and poverty eradication hardly consider broader impacts of networks of sectors. Individual sector development policies of many countries especially those related to the MDGs are often silent on potential linkages with our changing environments. Policy frameworks are poorly understood due to data impoverishment and nonexistent environmental indicators. Additionally, the restricted use of data to create scenarios for future planning and linkages with growth and development indicators undermines reliable assessments for policy, hence the need to appraise the following:

- Interventions such as UNEP's Africa Environment Outlook report and the African Ministerial Conference on Environment to help guide planning and macroeconomic policy at country levels to reduce environmental degradation.
- There is need for regional sharing of best practices, experiences, and technologies in Africa in priority sectors to raise awareness, engagement of stakeholders, and for informing decision-making processes at all levels of governance.
- Development frameworks such as the Action Plan of the Environment Initiative of the New Partnership for Africa's Development is about Africa's common and shared sustainable development problems and concern. It provides a framework for establishing strong partnership for the protection of Africa's environment and on the commitments contained in the United Nations Millennium Declaration (UNDP, 2003).
- There still remains a lot of human resource capacity to be harnessed to pursue mainstreaming at the macroeconomy level and this need in itself represents a risk to national and regional policy frameworks.

Conclusion

The environment of Africa is continuously experiencing a changing landscape as a result of increasing human consumption and ecosystem health concerns including environmental change, poverty, and livelihood burden. The challenge of the environment in Africa and its outcomes are heightened by enormous gaps in scientific and institutional capacity to fully understand vulnerability of the human sphere and ecosphere. The numerous environmental degradation outcomes occurring on the continent are potential developmental risks. Nationally, growth and development strategies such as those involving poverty eradication do so at the blind side of other sectors, a mix of their impacts that tend to drive environmental change. There is need for scientifically tested activities that seek to bring the issue of environment and sustainability and more recently inclusive green growth to the doorsteps of the poor in society and policy planners. Thus the understanding of the African environment as a social construct which is inextricably linked to the resource base will constitute the fundamental basis for country level and regional development planning. Africa's cultural landscape therefore constitutes a recipe for the continuous assessment of the evolving state of the environment toward creating future scenarios that provide options for environmentally sustainable development outcomes.

See also: Climate Change and Global Poverty; Climate Change: Economics; Climate Change; Environment and Health; Environmental Policy: Protection and Regulation; Food Security and 'Green Revolution'; Geographies of Food: Politics, Place, Power Reprised; Sustainability and Sustainability Science; Sustainable Development; Tourism, the Millennium Development Goals, and Gender Equality; Water Resources and Sustainable Water Management; Water Wars; Water, Health and Social Inequality.

Bibliography

AU – CAMH, 2007. The Africa Health Strategy 2007–2015. African Union, Addis Ababa.

Bates, B.C., Kundzewicz, Z.W., Wu, S., Palutikof, J.P. (Eds.), 2008. Climate Change and Water, Technical Paper of the Intergovernmental Panel on Climate Change. IPCC Secretariat, Geneva.

Batisani, N., Yarnal, B., 2010. Rainfall variability and trends in semi-arid Botswana: implications for climate change adaptation policy. Applied Geography 30, 483–489.

Bell, J., 1995. The hidden harvest. Seedling 12, 23–31.

Byamugisha, F.F.K., 2013. Securing Africa's Land for Shared Prosperity: A Program to Scale Up Reforms and Investments, Africa Development Forum Series. World Bank, Washington, DC.

Carney, D., Drinkwater, M., Rusinow, T., Neefjes, K., Wanmali, S., Singh, N., 1999. Livelihoods Approaches Compared: A Brief Comparison of the Livelihoods Approaches of the UK Department for International Development (DFID). DFID Report, Sussex.

Collins, A.E., 2001. Health ecology, land degradation and development. Special issue: rethinking environment and development in Africa and Asia. Land Degradation and Development 12, 237–250.

Dovie, D.B.K., Codjoe, S.N.A., Anwana, E.O., 2012. Migration (Africa). In: The Berkshire Encyclopedia of Sustainability: Afro-Eurasia: Assessing Sustainability. Berkshire Publishing Group, Great Barrington, pp. 206–211.

Dovie, D.B.K., Witkowski, E.T.F., Shackleton, C.M., 2005. Monetary valuation of livelihoods for understanding the composition and complexity of rural households. Agriculture and Human Values 22, 87–103.

FAO, 2007. Adaptation to climate change in agriculture, forestry and fisheries: perspective, framework and priorities. Inter-departmental working group on climate change. FAO, Rome.

FAO, 2011. The State of the World's Land and Water Resources for Food and Agriculture (SOLAW) – Managing Systems at Risk, Food and Agriculture Organization of the United Nations. Rome and Earthscan, London.

FAO, WFP, IFAD, 2012. The State of Food Insecurity in the World 2012. Economic Growth Is Necessary but Not Sufficient to Accelerate Reduction of Hunger and Malnutrition. FAO, Rome.

Farina, A., Napoletano, B., 2010. Rethinking the landscape: new theoretical perspectives for a powerful agency. Biosemiotics 3, 177–187.

Geist, H.J., Lambin, E.F., 2002. Proximate causes and underlying driving forces of tropical deforestation. Bioscience 52, 143–150.

Gitay, H., Suárez, A., Watson, R., 2002. Climate Change and Biodiversity, IPCC Technical Paper V. WMO, Geneva. UNEP, Nairobi.

ILO, 2007. Employment by sector. In: Key Indicators of the Labour Market (KILM), fifth ed. ILO, Geneva.

IPCC, 2007. Climate Change 2007: Fourth Assessment Report. IPCC Secretariat, Geneva.

IPCC, 2010. Workshop on sea level rise and ice sheet instabilities. In: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Midgley, P.M. (Eds.), Workshop Report of the Intergovernmental Panel on Climate Change. IPCC Working Group I Technical Support Unit. University of Bern, Bern.

Landau, L.B., Vigneswaran, D., 2007. Shifting the focus of migration back home: perspectives from Southern Africa. Development 50, 82–87.

Larson, J., 1998. Perspectives on Indigenous Knowledge Systems in Southern Africa. Discussion Paper No. 3. Africa Group World Bank, Washington, DC.

Lawes, M.J., Eeley, H.C., Shackleton, C.M., Geach, B.S. (Eds.), 2004. Indigenous Forests and Woodlands in South Africa: Policy, People and Practice. University of KwaZulu-Natal Press, Pietermaritzburg.

Ludi, E., 2009. Climate Change, Water and Food Security, Background Note. Overseas Development Institute, Sussex.

MacDonald, A.M., Bonsor, H.C., Dochartaigh, B.E.O., Taylor, R.G., 2012. Quantitative maps of groundwater resources in Africa. Environmental Research Letters 7.

Moore, J.L., Manne, L., Brooks, T., Burgess, N.D., Davies, R., Rahbek, C., Williams, P., Balmford, A., 2002. The distribution of cultural and biological diversity in Africa. Proceedings of the Royal Society of London, Series B 265, 1645–1653.

Muller, M., 2007. Adapting to climate change: water management for urban resilience. Environment and Urbanization 19, 99–113.

Qadir, M., Sharma, B.R., Bruggeman, A., Choukr-Allah, R., Karajeh, F., 2007. Non-conventional water resources and opportunities for water augmentation to achieve food security in water scarce countries. Agricultural Water Management 87, 2–22.

Sadoff, C., Muller, M., 2009. Water Management, Water Security and Climate Change Adaptation: Early Impacts and Essential Responses. TEC Background Papers No. 14. Global Water Partnership.

Shah, M., Fischer, G., van Velthuizen, H., 2008. Food Security and Sustainable Agriculture. The Challenges of Climate Change in Sub-Saharan Africa. International Institute for Applied Systems Analysis, Laxenburg.

Small, C., Nicholls, R.J., 2003. A global analysis of human settlement in coastal zones. Journal of Coastal Research 19, 584–599.

Taylor, P.D., 2002. Fragmentation and cultural landscapes: tightening the relationship between human beings and the environment. Landscape and Urban Planning 58, 93–99.

UNDP, 2003. Millennium Development Goals: A Compact among Nations to End Human Poverty, Human Development Report 2003. UNDP, New York.

UNDP, 2008. Fighting Climate Change – Human Solidarity in a Divided World. UNDP, New York.

UNEP, 2003. Action Plan of the Environment Initiative of the New Partnership for Africa's Development (NEPAD). UNEP, Nairobi.

UNEP, 2010. The Fourth Global Environment Outlook: Environment for Development, UNEP – WCMC. UNEP, Nairobi.

UNEP, 2013. Africa Environment Outlook 3: Our Environment, Our Health. UNEP, Nairobi.

UNFPA, 2011. State of World Population 2011. United Nations Population Fund, New York.

UNPD, 2011. World Population Prospects: The 2010 Revision. United Nations Population Division, New York.

WHO and UNICEF, 2012. Progress on Sanitation and Drinking-Water – 2010 Update. World Health Organization and UNICEF, Geneva.

WRI, 2005. Millennium Ecosystem Assessment – Ecosystems and Human Well-being: Biodiversity Synthesis. World Resources Institute, Washington, DC.

Africa, Sociocultural Overviews: West Africa

Paul Richards, Wageningen University and Research Center, Wageningen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article offers an overview of social worlds, values, and material and nonmaterial cultures of the region south of the Sahara from Mauritania to Cameroon. Attention is paid, in particular, to modes of social solidarity and the cultural dynamics of community formation in West African settings.

This article offers an overview of social worlds, values, and material and nonmaterial cultures of the region south of the Sahara from Mauritania to Cameroon. Attention is paid, in particular, to modes of social solidarity and the cultural dynamics of community formation in West African settings.

The People

West Africa is mainly peopled by Black Africans. Some groups of Berber and Arab descent inhabit the Saharan extremes. A small Lebanese diaspora, distributed mainly along the coast, occupies a role in commerce not unlike the Indian traders of eastern and southern Africa. Malaria deterred European settlement.

Linguistic Diversity

A striking aspect of West Africa is its linguistic diversity (Dalby, 1977). Nigeria, a country of *c.*100–120 million people, covering only about a seventh of the region, is estimated to have between 250 and 400 languages. (The huge expanse of eastern and southern Africa, by contrast, is dominated by closely related Bantu languages.) Language diversity in West Africa mainly reflects *in situ* development of long-settled farming communities, rather than long-distance migration. European languages, especially English and French, became established as public languages during the colonial period.

The Agrarian Base

West Africa divides into three broad environmental and cultural zones – an agro-pastoral desert edge dotted by ancient towns involved in trans-Saharan trade, a coastal rain forest zone shaped by external (sea-borne) commerce (including the Atlantic Slave Trade), and an intermediate zone of drier forest margins and wetter savannas, with a history of endogenous agrarian development dating back several thousand years.

The intermediate zone is the heartland of West African civilization and culture. It is here that local crops, notably white yam (*Dioscorea rotundata*), African rice (*Oryza glaberrima*), and a range of millets ('hungry rice,' pearl millet, and one race of sorghum), were first domesticated. The intermediate zone also provided the oil palm (*Elaeis guineensis*) – the source of oil used in cooking, a mild alcoholic beverage (palm wine), raw material for soap and medication, and ribs, trunks, and leaves from which bags, hammocks, roofs, building frames, and bridges are made.

In the eastern part of West Africa, along the forest–savanna transition and in the river valleys of the savanna, yam–oil palm farming systems prevail. Away from the riverine environments sorghum and pennisetum millet tend to replace yam as staples. The pattern repeats in the western half of West Africa, but with African rice taking the place of white yam (though today African rice is partly replaced by Asian rice, perhaps first introduced by the Portuguese in the era of the slave trade but spread widely only in the last century).

The early rise of a relatively intensive heartland agriculture went hand in hand with other sociotechnical changes, as evidenced by ancient iron working and terracotta art work, suggesting permanent settlement and complex division of labor in the intermediate zone at an early date.

The sacred importance and symbolic significance accorded to intermediate zone crops, even where they have been squeezed out of farming by commercially more productive introductions, is a notable aspect of West African material culture. Wild palms supply a rich red oil that remains an essential ingredient in many rituals, ceremonies, and remedies, greatly favored over the more common plantation substitute. The tasteless white rices of the coastal wetlands are considered no substitute for the nutritionally superior red rices of the interior. From Ghana to eastern Nigeria, yam survives as a preferred staple despite cheaper root crop substitutes. In the *Arrow of God*, Nigerian novelist Chinua Achebe appropriately made the disruption of a new yam festival in eastern Nigeria a striking metaphor for the far-reaching disruption of West African social and cultural values caused by colonial conquest (Achebe, 1964).

Lineage Society: 'Wealth in People'

West African agriculture depends on the hoe. The soils of the wetter savannas and adjacent forests are not suited to plow cultivation. Due to fly-borne disease, pastoralism is well developed only along the desert margins. Elsewhere it is not easy to accumulate wealth in livestock. Successful hoe agriculture depends on abundance of human labor. A central social value in the West African heartland is 'wealth in people.' This has many manifestations. The importance of a chief, for

example, is expressed in terms of wives, children, and domestic dependents, rather than through holdings of land and livestock.

But in hoe agriculture, timing and coordination of labor are as important as numbers of workers, especially in strongly seasonal savanna cultivation systems. Interhousehold labor cooperation, and the social values, crop types, and farming practices conducive to such cooperation figure strongly in West African agrarian social systems. In the old days, the young people of an entire community might pool their labor for farm operations. Today it is more likely that they will form cooperative gangs to tackle backbreaking and time-constrained jobs on a rotational basis. Where agriculture is commercialized labor gangs ply for hire.

Labor cooperation is a practical manifestation of egalitarianism among the young typical of West African heartland social organization. This egalitarianism is frequently fostered through initiation into so-called 'secret' societies, widespread and still powerful in parts of the region. Such societies typically tend to support social values among the young, strikingly in contrast with hierarchical values prevailing in the world of the elders. From Guinea to Cameroon, the senior grades within secret societies form a ladder along which elders compete to advance.

French Marxist anthropologists in the 1960s and 1970s wrote about this typical opposition between youthful egalitarians and ranked elders, as 'lineage' society, or a 'lineage mode of production,' clearly differentiated from feudal relations found in medieval Europe and elsewhere (Seddon, 1978). It was conceived that elderly elites, drawing on the language of kinship and descent for legitimating, reproduced themselves at the expense of the labor of cadets. According to some authors this amounted to a kind of class conflict. The explanation was controversial. An alternative explanation is that typical agrarian societies in the West African heartland are the complex outcome of a number of competing principles of social solidarity, none of which, by itself, can sustain a satisfactory community life over the longer term.

Agrarian Social Solidarity

As suggested, two principles of social solidarity often encountered in West African heartland social formations are egalitarianism and hierarchy. They offer quite distinct approaches to the problems of keeping social groups together (Douglas, 1993). Neither, alone, can sustain a community except in the short term. Compromise is an essential prerequisite for survival.

In a land-abundant setting, hierarchs may be able to do little to prevent youthful egalitarians from forging their own breakaway communities. But breakaway youthful egalitarians quickly run into problems of leadership. Youth labor gangs often have their own 'government,' and spend large amounts of time in dispute settlement, to the detriment of the work in hand. Charismatic leadership sometimes helps the egalitarians to go-it-alone for a time, but as with sectarians everywhere, splits lie around the corner. The more settled authority of the elders offers clear advantages. The typical West African agrarian compromise seems to have allowed for the emergence of spheres of partial autonomy for women and youth within a more complex social totality presided over by elders (and ancestors). Achieving such coordination and compromise without recourse to strong centralized political institutions was often the work of the 'secret societies' (for both men and women) found particularly in the Upper Guinean section of the Atlantic Coast in the western half of West Africa, and more widely in the region. In communities welded by initiation and secrecy, a social contract renewed or reinforced typically manifests itself from time to time in dramaturgical terms (often as masquerade display). The arts of masking are among the greatest manifestations of West African material culture and societal values.

The test of any mode of social solidarity is the extent to which it permits humans to flourish. A central fact about many decentralized ('stateless') societies of the West African heartland is that they often sustained denser populations than adjacent territories with a history of strong (and often strongly oppressive) central institutions (Middleton and Tait, 1959).

Society and Culture at the Margins: Coast and Desert Edge

International trade rather than agriculture was a mainstay of life at the coast and on the desert edge. It is in these regions that we first see most clearly the emergence of a trade-related West African urbanism, with its own distinctive institutions and cultural values. Along the desert edge, long-distance trade was associated with the early spread of Islam, and with Islam came literacy, science, and an enhancement of bureaucratic capacity. A process of modernization and reform took shape among desert-edge trade-related towns and polities in the late eighteenth century, and by the mid-nineteenth century larger regional formations, such as the Sokoto Caliphate (Last, 1967), covered a large part of what is today northern Nigeria. Two interior languages – Hausa and Maninka (Dioula) – developed some importance as media of regional communication, along with written Arabic for letter writing and record keeping.

At the coast, trade was stimulated by European contacts from the fifteenth century. Christianity gained several footholds, and small ports and polities oriented to the Atlantic trade developed, at times ruled by mercantile elites of Afro-European descent. Coastal trade pidgins, drawing first on Portuguese and later on English, developed into a distinctive family of mutually comprehensible English-oriented pidgins and creoles, today perhaps as widely spoken in coastal urban centers from The Gambia to Cameroon as Hausa or Dioula, the trade languages of the savanna, are spoken in interior West Africa. The English-oriented pidgins and creoles have considerable importance in coastal West Africa as means of communication among the nonelite, and in popular music and entertainment (Todd, 1984).

Globalization is no new phenomenon in desert-edge or coastal West African communities. From an early date children of trading elites might be sent to Cairo or Liverpool for their schooling. Nelson's victory against Napoleon's fleet at the Battle of the Nile in 1803 was known in Kumasi within weeks; that the news made sense at all implies a sophisticated world map in the West African interior. The typewriter was in

use among Christian converts in the interior of Yoruba-speaking country (now western Nigeria) within decades of its invention. The colonial telegraphs, and now the modern satellite phone, are among the communication means through which globally dispersed West African families have maintained and developed their wide networks of social contacts in the modern period.

Conversion to the 'world religions,' Islam and Christianity, tells us about growing global awareness even at early dates. Coastal linguistic creolization was part of a larger process of cultural creolization that helps explain the international outlook and mobility of many West Africans today (Hannerz, 1987). Diaspora communities of West Africans, often involved in trade, are found in most world cities today, and not just in the capitals of the former colonial powers. It is not uncommon to find street children or unemployed youth in Dakar, Douala, Freetown, or Lagos with a better grasp of politics and world issues than well-educated teenagers in Britain, America, or France.

The Impact of External Trade on Social Values

Arguments and compromises over social values took different forms under the impact of long-distance trade from those already described for heartland agrarianism. We need to distinguish internal and legitimate trade from the slave trade. Vigorous internal trade, exchanging forest and savanna products (notably cattle, kola nuts, and salt) was a stimulus to social and cultural innovation in a number of communities located on major north–south internal trade routes.

Typical adaptations to long-distance trading (Cohen, 1969) included the development of culturally distinct diasporas (Hausa cattle dealers in western Nigeria, or Fula kola dealers in Sierra Leone for example). At times agrarian elites maintained some control of diaspora trading groups through intermarriage. At other times diaspora communities have been confined to special settlements (Zongo, Sabon Gari) at the edge of villages or towns.

In other cases, women or youth have been free to exploit the potential of semiautonomous social spheres to build up commercial activities. Igbo young men from eastern Nigeria have long adapted the ganglike structures of the labor company to itinerant trading (in, for example, second-hand clothing, cloth, or motor spares). Agrarian labor cooperatives often double, or develop, as rotational savings and credit associations for the purposes of commerce.

Women traders are prominent throughout the West African costal and forest zone, but especially so in Ghana and southern Nigeria. Large-scale female traders, sometimes with little or no formal education, are to be encountered traveling the world, often in groups, perhaps as likely collecting car spares in Taiwan, as exporting a consignment of red palm oil to London. West African women's involvement in commerce typically began with local and regional foodstuffs trade neglected by the colonial commercial houses. The extensive yam trade from the Middle Belt interior to burgeoning colonial coastal cities of Ghana and Nigeria would be an example. In patrilineal communities, women are life-long brokers between kin groups. Taken in conjunction with the gender solidarity fostered by initiation, this helps set up an extrahousehold social space in which women's commerce has tended to flourish.

But we need to be careful not to impute modern ideas about individualism and enterprise to the development of West African trading relationships with the wider world. Frequently overseas or trans-desert trade was the monopoly of rulers who kept a tight rein on contacts and commercial information. Those controlling the trade-based polities of coast and desert edge were more likely to be monopolists and autocrats than agents of an enterprise culture. In the nineteenth century, the spread of firearms allowed trade-oriented rulers extensive use of socially unaccountable force. A not dissimilar militaristic adventurism has reasserted itself in some parts of West Africa (notably in Liberia and Sierra Leone) as the twentieth century ends, prompted by an international political vacuum and a surfeit of cheap weapons from the ending of the Cold War.

The mode of social solidarity that tends to predominate under war-lord autocracy is not individualism but fatalism, a fact illustrated in the 1997 elections in Liberia, where a war-weary population voted for the strongest of the faction leaders from a destructive seven-year civil war, reasoning the most powerful autocrat was the best guarantee of a kind of peace. Fatalism is little appreciated as a basis for social solidarity, but may be preferable to the mayhem of a bush civil war. Here we have an insight into why social life on the West African frontier so often seems to have been characterized by quietism, and a subsistence-oriented conservatism, rather than by the expansive individualism associated with the making of the American frontier (Kopytoff, 1987). The attitude is 'we make no demands other than to be left to get on with our lives.' The old heartland agrarian social mix continues to have a certain basic appeal where the alternative is the postmodern violence of late twentieth century conflict in Liberia and Sierra Leone.

Social Consequences of the Slave Trade and Colonial Rule

Perhaps between 10 and 15 million West Africans were sold into Atlantic slavery from the sixteenth to nineteenth centuries, and 8–10 million into trans-Saharan slavery. In addition to the adverse demographic impact, in a region dependent on labor-intensive hoe agriculture, the conversion of human beings into commodities dealt an enormous blow to the trust essential to the functioning of many social institutions (Inikori, 1982). It is hard to overestimate the continuing consequences of the slave trade for the evolution of social values in West Africa. Fatalism, fear of entrepreneurial individualism, distrust of chiefs, and undue respect for autocrats in return for basic security are among the symptoms of the damage. Distrusts inherited from the era of the slave trade remain a factor in the nihilism of much of the current youth-led violence blighting the region. How ordinary West Africans continue to come to terms with the harsh social legacy of the slave trade is also very apparent in popular religion, in ideas about chiefs as shape-shifting witches, or through divination and possession as ways of addressing old betrayals and coming to terms with disastrous breakdowns in family or community solidarity (Shaw, 1997).

Colonial protection in West Africa arose from initiatives to eliminate the slave trade, beginning with the assumption in 1807, by the British Crown, of administrative responsibility for Sierra Leone, an eighteenth century West Coast settlement for freed slaves founded by British Christian philanthropists. Ending the slave trade became the pretext for a series of moves by industrializing European powers to control resource extraction and guarantee markets for their own manufacturers. By the end of the nineteenth century, West Africa had been carved up by the Germans, French, and British (Crowder, 1968).

Colonial rule at the beginning of the twentieth century was militaristic, often formally racist, and based on a protectionist principle of imperial preference. Although colonial regimes built railways and some roads, and Christian missions under colonial protection introduced school education and Western medicine, colonial autocracy did nothing to repair the adverse social legacy of the slave trade.

Postcolonial Social Trends

Many leaders in postcolonial West Africa adapted colonial autocracy to their own ends, paying more attention to external stimuli – notably Cold War incentives – and less to local accountability. A spate of coups replaced civilian with military autocrats. Economic development was hindered by unwise investments in prestige projects and by corruption. Former colonial powers embarked on political intrigue to control rich natural resources, for example the role of France in the Nigerian civil war (1967–70). The bright hopes of independence dimmed, and the region lost many of its talented professionals through international migration. Inflows of development aid and migrant remittances became essential parts of most West African economies.

From the 1980s the international financial institutions, and Western aid donors, attached conditions to aid grants in the hope of fostering a transition to a market economy, but without taking full account of the way in which adverse social and institutional developments under the slave trade and colonialism had prevented the development in West Africa of the internal checks and balances between state, market and civil society on which a modern market-based society depends.

Some West African rural groups retained access to herds or large expanses of land for shifting cultivation, and could retreat toward heartland subsistence, but others increasingly found that the system of circulatory migration, in which young people migrated to the cities in search of laboring work, in order to save money for a successful return to village life as a respected elder, had begun to fail. Increasingly, young migrants might find that land had been leased to tenant farmers in their absence, or acquired for commercial development by agribusiness. Unable to reintegrate within rural society, young migrants began to question traditional social values and cultural norms. Some became adherents of populist leaders proposing radical change.

The 1990s was a decade of civil war, in which rural insurgency was fed by a youth population bulge. By now, over half of all West Africans was below the age of 18. These insurgencies severely damaged isolated rural communities in the Upper West African forest belt, the Niger Delta and the desert edge, but were unable to make good their promise of profound and sustained change for young people. However, armed conflict was often prolonged by the involvement of international venture capitalists and private military companies, seeking access to lucrative timber and mineral resources, and able to supply cheap arms and military know-how rendered redundant by the ending of the Cold War.

The storm blew itself out in the first decade of the twenty-first century, as peacemakers switched their attention from negotiating with populist mobilizers of youth to sanctioning the diamond dealers and gun runners, and to addressing more basic underlying social issue, notably youth unemployment. Promises of skills training and jobs attracted some of the young fighters out of the bush. Others continued, under the perhaps more durable appeal of political Islam. The Islamist group Boko Haram bolsters its insurgency in northern Nigeria with a rejection of school education, and the world of 'modern' work to which such education promises to lead. Much remains to be done to stabilize a younger generation denied educational opportunities and jobs by earlier economic malaise.

Most recently, the social landscape in West Africa has begun to experience a further transformation under the impact of China's expansion into the region in search of raw materials for its global industrial machine. Being interested in bulk mineral extraction (iron ore, bauxite, mineral oil), China cannot interact with West Africa in the manner of the hit-and-run opportunists of the conflict years in the 1990s. Chinese investment is focused not only on mineral concessions, but also on the infrastructure of transport, power, water, and telecommunications necessary for long-term extraction ventures.

Other investors have been attracted by this Chinese-led upturn in the region's economic fortunes. Some large-scale overseas ventures have attempted to learn from the lessons of the near collapse of large parts of West African society in the 1990s. Land deals – e.g., for biofuel and energy developments – attempt to incorporate a variety of rural stakeholders, in addition to generating welcome employment opportunities for youth.

At the same time, governments have become more democratic, open and accountable. State revenues (boosted by the rapid improvement in economic fortunes) are now better able to provide for social needs such as basic education and health. But there remain major concerns over the sustainability of the current rapid levels of development.

Signs of environmental fragility abound. Rapid urban development is chaotic and uncontrolled, leading to potentially destabilizing future problems of pollution, congestion and epidemic disease. Most of the new employment in extractive industry and infrastructure development is unskilled and low paid; skills training remains a major bottleneck. The undermining of traditional land rights threatens also to dismantle an important safety net for the poorest groups. Material inequality is manifest. Disgruntled youth still have their rallying points – not least militant Islam.

See also: African Studies: Culture; African Studies: History; Agrarian Political Economy; Colonialism, Anthropology of;

Colonization and Colonialism, History of; Creolization: Sociocultural Aspects; Evolution and Language: Overview; First Language Acquisition, Developmental Psychology of; Horticultural Societies; Language Families, Archaeology and History of; Language and Society; Slavery as Social Institution; Slavery: Comparative Aspects; Slaves and Slavery, History of.

Bibliography

Achebe, C., 1964. Arrow of God. Heinemann, London.
Allen, J.A., Martin Keulertz, M., Sojamo, S., Jeroen Warner, J. (Eds.), 2013. Handbook of Land and Water Grabs in Africa: Foreign Direct Investment in Food and Water Security. Routledge, Abingdon, UK.
Brautigam, D., 2009. The Dragon's Gift: The Real Story of China in Africa. Oxford University Press, Oxford.
Cohen, A., 1969. Custom and Politics in Urban Africa. Manchester University Press, Manchester, UK.
Cooper, F., 1996. Decolonization and African Society. Cambridge University Press, Cambridge, UK.
Crowder, M., 1968. West Africa under Colonial Rule. Hutchinson, London.
Dalby, D., 1977. The Language Map of Africa. International African Institute, London.
Douglas, M., 1993. In the Wilderness. JSOT Press, Sheffield, UK.
Hannerz, U., 1987. The world in creolization. Africa 57, 546–559.
Inikori, J.E. (Ed.), 1982. Forced Migration. Africana Publishing Company, New York.
Kabia, J.M., 2008. Humanitarian Intervention and Conflict Resolution in West Africa: From ECOMOG to ECOMIL. Ashgate, Farnham, UK.
Kopytoff, I. (Ed.), 1987. The African Frontier. Indiana University Press, Bloomington, IN.
Last, M., 1967. The Sokoto Caliphate. Longmans, Harlow.
Middleton, J., Tait, D. (Eds.), 1959. Tribes without Rulers. Routledge & Paul, London.
Richards, P., 1996. Fighting for the Rain Forest. James Currey for the International African Institute, Oxford, UK.
Seddon, D. (Ed.), 1978. Relations of Production. Frank Cass, London.
Shaw, R., 1997. Production of witchcraft/witchcraft as production: memory, modernity, and the slave trade in Sierra Leone. American Ethnologist 24 (4), 856–876.
Todd, L., 1984. Modern Englishes: Pidgins and Creoles. Blackwell/Andre Deutsch, Oxford, UK.

Afrocentric Approaches to Social Work

Patricia Reid-Merritt, The Richard Stockton College of New Jersey, Galloway, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores the use of 'Afrocentricity' as a conceptual framework for social work practice and service intervention in African-based communities. As developed and defined by Asante (1980), Afrocentricity is a theory of social change that affirms the importance of African history, culture, and values as a foundation for understanding the contemporary experiences of African descendant peoples throughout the world. This article will examine the emergence of the concept and advocates efforts to link Afrocentricity to social work practice in the African diaspora.

Afrocentricity and Social Work: Early Beginnings

In the field of social work, Africentric, Afrocentric, and African-centered are concepts, often used interchangeably, to describe the efforts of theoreticians and practitioners to integrate and fully utilize an innovative approach to social work intervention, which places the African experience at the center or as the central focus of practice. The growth and expansion of this initiative can be traced directly back to the early beginnings of the Black Social Work Movement (BSWM) in America. The BSWM harnessed the energies of a radically progressive group of Black social work professionals who desired to alter traditional social work practice to fit the needs of the people they served. These visionary Black social workers were committed to the purpose, goals, and mission of the social work profession and embraced it as a viable social change strategy to aid the Black community. From its earliest inception, social work, which can best be described as a profession committed to the enhancement of the quality of life for individuals, families, groups and communities, expressed a particular interest in serving the needs of the poor, socially disadvantaged, and vulnerable populations (NASW, 1977). Social work proved to be a popular choice for many college-educated, career-oriented Black professionals.

The Black Freedom Movement

During the height of the Civil Rights and Black Power movements that rocked the United States of America in the mid-1960s, Black social workers began to question their role in the struggle for Black freedom and liberation. The social turmoil that erupted during this era forced many Black professionals to reexamine their relationships with the Black community. Located at the nexus of social policy implementation and service delivery to the masses of poor Black people, many professionally trained Black social workers questioned the relevance of their knowledge and skills as it related to the uplift and improvement of the quality of life for their clientele. Frustrated with their roles as ineffective 'state guardians of the poor,' Black social workers sought to connect with other disaffected workers to discuss a more radical approach to practice. As the network of social activists expanded throughout the community and, eventually, throughout the nation, the BSWM was born (Reid-Merritt, 2010).

The BSWM emerged as a grassroots movement of highly trained social welfare workers and community activists seeking effective social service intervention with positive outcomes. It was one of numerous organized efforts that contributed to the larger, mid-twentieth century movement for Black freedom (Karenga, 2010). As interest grew, the activists formally organized, and in May 1968, founded the National Association of Black Social Workers, Inc. (NABSW). This organization would lead the national and international efforts to transform social work and the delivery of social work services in Black communities (Reid-Merritt, 2010). For example, in California, Georgia, Illinois, Michigan, New York, New Jersey, Pennsylvania and Tennessee, local chapters affiliated with NABSW challenge child welfare agencies on practices regarding the removal, placement, and adoption of Black children; a national strategy is developed to resist state efforts to regulate the practice of social work through licensing; and social work graduate schools face demands for changes in faculty staff, student population, and curriculum that are reflective of the racial diversity that exists within the surrounding community.

Seeking New Knowledge

The conscious use of race and culture as major factors to be considered in social work intervention in the Black community was initially described by NABSW and other socially progressive practitioners as 'Black social work practice.' LaVerne McCummings (NABSW, 1970), a BSWM early participant, social activist and NABSW leader, described the initiative as a new form of social work practice 'based on liberation rather than adjustment,' Numerous scholarly efforts were made to help articulate this new form of practice. Better (1972) was among the first to argue the benefits of racially conscious Black social work practitioners. She noted that an awareness of their role as Black professionals, employed as agents of social change within their own communities, would enhance their effectiveness and the delivery of social services. Mathis (1972) promoted the concept of 'nation-building,' where the role of race and identity was clearly understood as a primary responsibility of the Black social work activist. The goal of nation-building was the liberation of the Black masses from "the shackles of physical, emotional, economic, and political dependency" (Mathis, 1972: 2). The racially conscious Black social work practitioner could achieve these goals by rooting themselves in an understanding of the extended Black family;

committing themselves to the development of independent Black institutions; freeing themselves from white social work values and concepts; and operating from a value system based on African communalism. Gilbert (1974) concurred, calling for 'a fundamental change' in the primary goal of the social work intervention in the Black community to make it more liberating than oppressive. All acknowledged the need for a new form of 'cultural grounding' for Black social workers, reflective of their identity and heritage as Black people.

Black scholars and practitioners struggled to identify the new knowledge, concepts, and theories needed to enhance the effectiveness of social work practice within Black communities. And, during the first full decade of the BSWM, various theories, philosophies, and value systems were used to buttress the movement toward a more African-centered approach, including, but not limited to, Kawaida and Nguzo Saba (Karenga, 2010), strengths (Hill, 1972) and nigrescence (Cross, 1971, 1980).

Karenga's (2010) Kawaida theory spoke to the need for a cultural revolution for Africans in America, emphasizing the need for Black people to embrace their own African heritage. He (2010) develops a set of seven principles, described as the Nguzo Saba, which guides the transformative process. Identified first in Swahili and translated into English, they include *Umoja* (Unity), *Kujichagulia* (Self-Determination), *Ujima* (Collective Work and Responsibility), *Ujamaa* (Cooperative Economics), *Nia* (Purpose), *Kuumba* (Creativity), and *Imani* (Faith). Hill identified five common strengths that were prevalent in Black families. These included a solid work ethic, high aspirations for their children, broad kinship bonds and support networks, role flexibility between the men and the women, and a strong religious foundation. Cross (1971) spoke in *The Negro to Black Conversion Experience*, which enabled a young, racialized generation of Black to understand their internal, emotional struggles surrounding their identity. Nigrescence, or the process of becoming Black and being able to accept one's Black identity, occurred in five stages: pre-encounter, encounter, immersion, emersion, and internalization. In each stage the individual is confronted with conflicting social images about his/her Black identity until transitioned to the final stage of internalization where one finds a level of comfort.

However, it was Asante's (1980) introduction of the concept of Afrocentricity that changed the language of blackness and allowed the progressive Black social work practitioner to more accurately describe this new approach to practice.

The Afrocentric Approach

In *Righteous Self-Determination: The Black Social Work Movement in America*, Reid-Merritt (2010: 132) posits that Asante's theory of Afrocentricity "entered scholarly discourse at a time when many scholars struggled to reformulate and re-conceptualize their basic orientation to the study of African people." More than any other emerging concept in the scholarly literature, this intellectual contribution aided Black social workers in their search for the development of a culturally grounded and racially specific approach to social work practice. She (Reid-Merritt, 2010: 134) notes further that Asante's theory "provided the language and cultural grounding needed to guide Black social workers toward the development of nontraditional, alternative approaches to understanding the Black community from an African-centered perspective."

In his ground-breaking work, *Afrocentricity: A Theory of Social Change* (1980), Asante provides a foundation for understanding the contemporary African world experience. He argues that African descendant peoples who have lived under the forced cultural, social, political, and economic hegemony of European dominance throughout the world, experience disorientation and dislocation in contemporary societies. The history of invasion of the African continent, the transfer of millions of Africans to the Americas, the Holocaust of Enslavement, the colonization of Africa, and the ongoing European domination that followed resulted in a 'de-centering' of the African mind, body, and soul. A positive, reaffirming form of social and cultural identity had been lost. The process of reclamation required re-centering oneself in the history, values, and culture of an authentic African past. This social and personal repositioning would help to nurture the development of a more viable and life-affirming 'Afrocentric reality' (Asante, 1980, 1990; Reid-Merritt, 2008).

As the contours of the paradigm continued to unfold, Black social workers began to embrace the concept of Afrocentricity as a new framework for practice. Discourse on black social work practice was transformed with the use of the words Afrocentric, Africentric, and African-centered. Nowhere was this more evident than within the BSWM and the national and international gatherings sponsored by NABSW (1970–2010). For example, NABSW traveled, with large groups of black professionals, to the Caribbean, South America, and Africa, hosting international conferences and spreading the word about the commonalities of the African descendant experience throughout the world. And the national conferences were visited by Black social workers from the international community, including the Bahamas, Brazil, England, Germany, Jamaica, and Nova Scotia. There were common themes: racism, oppression, powerlessness, and despair, all met with a spirit of unity and willingness to fight for the uplift of the community. Both in professional practice and social work literature, Black social workers began to discuss the need for an African-centered approach as numerous scholars began to articulate ways in which Afrocentricity could be applied to various forms of social work practice (Harvey, 1997; King, 1994; Swigorski, 1996; Schiele, 1997, 1996; Graham, 1999). But how was this best achieved and integrated within the discipline?

Afrocentric Social Work: Theory and Practice

As a discipline and practice profession, social work requires a knowledge, value, and skill base to help guide its practitioners toward effective social service and human interventions. Its knowledge base, in particular, borrows heavily from other disciplines, including history, psychology, sociology, anthropology, political science, economics, and other social and behavioral sciences, to form a foundation for understanding the human condition. However, recognizing that traditional sociohistorical literature was Eurocentrically oriented and

noninclusive of the experiences of African people, Black social workers sought to build a new intellectual foundation. The existing knowledge base was inadequate; new theories, values, and skills, reflective of an African-centered worldview, were required. Thus, Black social workers began to focus their discourse around the need to replace the traditional social work knowledge base with Afrocentric knowledge. And they were not alone in their pursuit; Black scholars in other traditional disciplines were experiencing a social and intellectual transformation as well.

In the US, for example, in psychology, the widely popular and traditional theories of Freud, Erickson, Jung, and Adler (Beckett, 2002; Miller, 2009) were being replaced by Akbar (1981), Nobles (1986), Kambon (1992), and Myers (1988). These emerging black intellectuals and practitioners worked on developing innovative theoretical approaches to transform discussions of the practice of black psychology into Afrocentric approaches. In sociological theory, discussions on the impoverished, socially maladjusted, culturally deprived black family (Karpas, 1964; Lewis, 1966; Moynihan, 1965) were replaced with a strengths, vitality, and survival orientation that highlighted the adaptive qualities and capabilities of African American families (Hill, 1972; Sudarkasa, 1980, 1981; McAdoo, 1981). These scholars argued the merits of the aforementioned characteristics as enduring qualities that contributed to the flourishing of the culture and the people.

The discourse in social work emphasized the need to incorporate an Afrocentric approach into all areas and fields of practice, including child and family services, educational programs, mental health services, drug abuse/addiction prevention and treatment programs, pretrial/incarceration/re-entry programs, community development, and social policy. And new programs, utilizing the Afrocentric perspective, did emerge. These included rites of passage programs for youth; parenting programs; community-wide educational programs in African and African American history, art, and culture; community-based Kwanzaa celebrations; and the reintroduction of African rituals into traditional African American cultural and social events (Monges, 1990; Harvey and Rauch, 1997; Reid-Merritt, 2010). Practice approaches may have differed in their particular methods and techniques, however, the basic philosophical grounding remained the same: The African past was a central component of the modern-day black experience; African descendant people were a self-determining population with the right to obtain 'cultural agency'; social work interventions must be based on the realities of the African world; and practitioners must start with a fundamental understanding of the history, values, ideals, culture, and heritage of African people.

Black social workers who espoused a belief in the Afrocentric approach may have appeared to some to be the new renegades in the profession, however, recognizing and accepting the unique qualities of human populations was consistent with the core values and principles of social work. And, Black social workers were determined to use all aspects of their professional education and training to better serve their own communities, which were too often negatively described as problematic havens for 'minority' group members.

Decades after the release of Asante's (1980) pivotal work, efforts to reconceptualize social work from an Afrocentric perspective continued to unfold as practitioners and scholars challenged the dominance of the European worldview. The new, emerging literature was replete with the words Afrocentric and Afrocentricity. And while these were the most popular terms used (Schiele, 1990; King, 1994; Mathis, 1994; Daniels, 2001; Graham, 1999; Reid-Merritt, 2010), others (Roberts et al., 2000; Harvey and Rauch, 1997; Clegg, 1997; Crawley, 2001; Harvey and Hill, 2004; Bent-Goodley, 2005) preferred to use 'Africentric' or 'African-centered' in their challenge to Eurocentric approaches. Each of their contributions supported the expansion of the developing paradigm for Afrocentric social work practice.

The examples are plentiful. For Schiele (1996: 285), the Afrocentric approach offered the opportunity to apply 'traditional African philosophical concepts as a foundation for a new social work practice model.' For Graham (1999, 2005: 71), a leading Afrocentric scholar from the UK who now resides in the US, the new Afrocentric approach to social work incorporated the history, values, and beliefs of an African-based culture and embraced "an African-centered worldview as the basis for theories and models of practice." Mathis (1994: 4) added that an African-centered approach required a new set of values as the basis for social work practice in Black communities. Borrowing from Maatian philosophy, which was based on traditional Egyptian religion, thoughts, practices and beliefs, the seven cardinal values of truth, justice, propriety, balance, harmony, reciprocity, and order replaced western values based on individualism, competition, independence/dependence, and materialism.

For micro-level practice intervention, Crawley (2001: 127) utilized the Afrocentric perspective to provide "a framework that allows clients to connect to a vital sphere of their identity – that is, being of African descent." And Harvey and Rauch (1997: 164–8) described a psychosocial, Afrocentric model that utilizes RIPSO or Principles of Afrocentricism. These value principles are identified as: "responsibility, reciprocity, respect, restraint, reason, reconciliation, interconnectedness, interdependence, inclusivity, participation, patience, perseverance, sharing, sacrifice, spirituality, cooperation, discipline and unconditional love." While many scholars and practitioners continued to incorporate an Afrocentric perspective, fully articulated Afrocentric practice models are still in the early stages of development. An African-Centered Approach to Social Work (AFCENA) is an example of one of these emerging models.

AFCENA: An African-Centered Approach to Social Work Practice – A Model for Practice

AFCENA was introduced to the Black social work community by Reid-Merritt (2010) at a national Policy institute sponsored by NABSW in 1995. AFCENA "fused African-centered contemporary intellectual thought with the fundamental principles, goals and purposes of social work and recognized the significant contributions that good social work practice could have on the African American community" (Reid-Merritt, 2010: 193). This model provided a broad, conceptual framework which specified the building blocks needed for Afrocentric social work practice and offered a comprehensive approach, applicable to the Black community as a whole, rather than its fragmented parts. It is applicable to every area

of practice, including clinical intervention, community organization, policy development, education, and research.

The context of the model is bound by a communal framework, i.e., the depth of knowledge, understanding, and appreciation for the socially unique characteristics of African descendant communities, which speaks to the *collective inheritance* of the people. As a prerequisite to any form of practice intervention in the Black community, social workers must be knowledgeable about the historic cultural ties that bind the community together. This collective inheritance focuses on an accurate historical account of past challenges and achievements and includes (1) an acknowledgment of the importance of *ancestral wisdom*, which has been passed down from generation to generation; (2) an acceptance of the pervasiveness of *epic memory*, which provides a deep remembrance of the behaviors, habits and customs of the people; (3) the ability to embrace the community's *cultural legacy*, which provides an active, continuous reflection on how people choose to live their lives; and (4) a belief in the future guided by *victorious visions*, which help to galvanize the community's efforts to strive toward greatness. This historical cultural knowledge base would prove valuable for all professionals, but deemed essential for the Afrocentric social work practitioner. However, a simple question must be raised: How does one gain proficiency in this area? There are limited expectations that this specialized form of knowledge would be found within traditional schools of social work; rather, the Afrocentric social worker must seek out and design a course of study that provides the in-depth knowledge needed for specialized practice in African-descendent communities.

There are five basic components to the AFCENA model, including an Afrocentric perspective, knowledge/experience, values/ethics, skills/behaviors, and spirit/life force. The components of this circular model remain fluid and accessible throughout all stages of the helping process. "They are constant, flowing and interchangeable. The beginning is the end and the end is the beginning of the process" (Reid-Merritt, 2010: 194). Afrocentricity or the Afrocentric perspective provides a common framework for understanding and analyzing the history, philosophies, religious beliefs, and cultural practices of African people. The knowledge base is buttressed by the infusion of theoretical perspectives grounded in the African worldview. This is critically important in the areas of human growth and development, socialization and community development which focus on understanding the human condition. Experience or 'practice wisdom' helps the practitioner verify the relevance and applicability of new theoretical approaches as they are tested in the helping process.

The principles of Nguzo Saba and Maat serve as the value base for Afrocentric practice, and the NABSW Code of Ethics serve as a constant reminder that "I regard as my primary obligation the welfare of the Black individual, Black family and the Black community and will engage in action for improving social conditions. I give precedence to this mission over my own personal interests" (NABSW, 1970). Helping skills, including engagement, observation, listening, questioning, verbal and nonverbal communication, must be those which are valued, appreciated, and understood in African-based communities, and include communication patterns attuned to the call and response and the participatory nature of the culture. And professional behavior is a reflection of one's ongoing commitment to the uplift and empowerment of the people and community, not one's primary commitment to bureaucratic agencies or ineffective government policies. Finally, we are all in possession of a spiritual life force, the very essence of our existence, which provides meaning to life, motivates us to live our life to the fullest, and helps us interpret our relationship with the universe. The Afrocentric social work practitioner remains fully cognizant of the need to help those who have been temporarily dispirited by personal obstacles and/or social oppression to resecure their place on life's continuum.

The conceptual practice models also provide descriptive details on the interactive system, e.g., the place of contact between the social worker and the client system. This is where the work is done, where problems are manifested, or where success is achieved. There are multiple practice settings in the field of social work. And while a small percentage of the profession has chosen to engage in individual private practice, most social work intervention is sponsored by government and private agencies and foundations. AFCENA describes the point of contact between the social worker and clientele/community as the interactive system. The Afrocentric social worker cannot operate in a vacuum. They must be mindful of the major external factors which impact the client system and the delivery of social services to Black communities. These factors include race, gender, class, religion, geographic location, and political ideology. Moreover, the worker must remain cognizant of the mission, purpose, and goals of social work intervention. Resistance to service, which may appear to be more prevalent in oppressed communities, must be countered with optimism and hope for positive change. Social work service intervention outcomes must go beyond the mere return of individuals, families, and communities to survival status; rather, service intervention should lead to personal and social transformations that help individuals, families, and communities share in the abundant resources of society and thrive under new social conditions and arrangements beneficial to all.

Proof

"Afrocentric approaches to social work offer the greatest potential for the development of racially and culturally-specific models of social work practice intervention in the Black community" (Reid-Merritt, 2010: 196). However, the theories, models, and perspectives utilized to support Afrocentric social work practice have and will continue to be challenged by mainstream social work. The profession's movement toward *evidence-based practice* reminds us of the need to provide proof of the effectiveness of these new approaches to practice. Documentation of our success, and failures, must be supported with rigorous investigations, ongoing research, and the publication of our results in the social work literature. This would prove valuable not just to the advocates of an Afrocentric practice, but to all those who seek alternative approaches to social work practice with marginalized, oppressed, and socially disadvantaged communities.

See also: Critical Social Work Practice; Ethnicity, Race, Oppression, and Social Work: The Canadian Case; Liberation Theology in Social Work; Māori Social Work; Oppression; Orientalism, Oppression, and Contemporary Social Work; Radical Social Work; Social Protest Movements and Social Work Practice; Social Work Theory; Social Work with Racialized Groups: Frameworks for Practice.

Bibliography

Akbar, Na'im, 1981. Mental disorders among african americans. Black Books Bulletin 5 (2), 18–25.

Asante, Molefi Kete, 1980. Afrocentricity: A Theory of Social Change. Africa World Press, Trenton, NJ.

Asante, Molefi Kete, 1988. Afrocentric Idea. Africa World Press, Trenton, NJ.

Asante, Molefi Kete, 1990. Kemet, Afrocentricity and Knowledge. Africa World Press, Trenton, NJ.

Beckett, Chris, 2002. Human Growth and Development. Sage Publishing, Thousand Oaks, CA.

Bent-Goodley, Tricia B., 2005. An african centered approach to domestic violence. Families in Society 86, 197–206.

Better, Shirley, November 1972. The role of the black social worker. Black World, 4–14.

Clegg, Legrand, 1997. A brief history of Africentric scholarship. MAAT News 1 (2). www.melanet.com.

Crawley, Brenda, 2001. Effective programs and services for african american families and children: an African-centered perspective. In: Logan, Sadye L.M. (Ed.), The Black Family: Strengths, Self-Help and Positive Change. Westview Press, Boulder, CO.

Cross Jr., William E., July 1971. The negro to black conversion experience. Black World, 13–27.

Cross Jr., William E., 1980. Models of psychological negrescence: a literature review. In: Jones, Reginald L. (Ed.), Black Psychology. Harper and Row, New York, pp. 319–338.

Daniels, Jean, 2001. Africentric social work practice: the new approach for social work practice intervention in the 21st century. International Social Work 44 (3), 301–309.

Gilbert, Gwendolyn C., December 1974. The role of social work in black liberation. The Black Scholar 6 (4), 16–23.

Graham, Mekada, 1999. The african centered worldview developing a paradigm for social work. British Journal of Social Work 29 (2), 252–267.

Graham, Mekada, 2005. Afrocentric social work. In: Asante, Molefi Kete, Mazama, Ama (Eds.), Encyclopedia of Black Studies. Sage Publications, Thousand Oaks, CA, pp. 72–73.

Harvey, Aminifu, 1997. Group work with african american youth in the criminal justice system: a culturally competent model. In: Greif, G.L., Ephross, P.H. (Eds.), Group Work with Populations at Risk. Oxford University Press, New York, pp. 160–174.

Harvey, Aminifu, Rauch, Julia B., 1997. A comprehensive afrocentric rites of passage program for black male adolescents. Health & Social Work 22, 30–37.

Harvey, Aminifu R., Hill, Robert, B., 2004. Africentric youth and family rites of passage program: promoting resilience among at-risk African American youths. Social Work 49 (1), 65–74.

Hill, Robert, 1972. The Strengths of Black Families. Emerson Hall, New York.

Hill, Robert, 1997. Strengths of Black Families Twenty Five Years Later. R and B Publishers, Washington, DC.

Kambon, Kobi, 1992. The African Personality in America: An African-Centered Framework. Nubian Nation Publications, Tallahassee, FL.

Karenga, Maulana, 1980. Kawaida Theory: An Introductory Outline. Kawaida Publications, Inglewood, CA.

Karenga, Maulana, 2008. Kawaida and Questions of Life and Struggle: African American, Pan-African and Global Issues. University of Sankore Press, Los Angles.

Karenga, Maulana, 2010. An Introduction to Black Studies, fourth ed. University of Sankore Press, Los Angeles.

Karpas, Melvin Ronald, 1964. Readings in Cultural Deprivation. Juvenile Delinquency Research Projects, Chicago.

King, Anthony E., 1994. An afrocentric cultural awareness program for incarcerated African- American males. Journal of Multicultural Social Work 3 (4), 20–29.

Lewis, Oscar, October 1966. The culture of poverty. Scientific America.

Mathis, Thaddeus P., 1972. Social work and nation building in the black community. In: Proceedings of Fourth Annual Conference. NABSW, New York, pp. 159–177.

Mathis, Thaddeus P., 1994. Afrocentric Social Work Practice. National Association of Black Social Workers Policy Institute, Phila., PA (Unpublished).

Mazama, Mambo Ama (Ed.), 2008. Essays in Honor of an Intellectual Warrior: Molefi Kete Asante. Menaibuc, Paris.

McAdoo, Harriette (Ed.), 1981. Black Families. Sage Publications, Beverly Hills, CA.

Miller, Patricia A., 2009. Theories of Developmental Psychology, fifth ed. Worth Publishers, New York.

Monges, Miriam, 1990. Engaging community: the North Philadelphia initiative. In: Professional Development Workshop, April. Temple University, Philadelphia, PA.

Moynihan, Daniel, 1965. The Negro Family. Office of Planning and Research, U.S. Department of Labor, Washington, DC.

Myers, Linda James, 1988. Understanding an Afrocentric World View: Introduction to an Optimal Psychology. Kendall/Hunt, Dubuque, IA.

National Association of Black Social Workers, 1970–2010. Conference Brochures. NABSW, New York.

National Association of Social Workers, 1977. Encyclopedia of Social Work, seventeenth ed. NASW Press, Washington, DC.

Nobles, Wade, 1986. African Psychology: Toward its Reclamation, Reascension and Revitalization. Black Family Institute Publishers, Oakland.

Reid-Merritt, Patricia, 2008. The spread of afrocentricity among the people. In: Mazama, Mambo Ama (Ed.), Essays in Honor of an Intellectual Warrior: Molefi Kete Asante. Menaibuc, Paris.

Reid-Merritt, Patricia, 2010. Righteous Self-Determination: The Black Social Work Movement in America. Black Classic Press, Baltimore.

Reid-Merritt, Patricia, 1995. An African-Centered Approach to Social Work Practice. NABSW Policy Institute.

Roberts, Amelia, Jackson, Mary S,, Carlton-LaNey, Iris, 2000. Revisiting the need for feminism and Afrocentric theory when treating African American female substance abusers. Journal of Drug Issues 30, 901–918.

Schiele, Jerome, 1990. Organizational theory from an afrocentric perspective. Journal of Black Studies 21, 145–161.

Schiele, Jerome, 1997. The contours and meaning of afrocentric social work. Journal of Black Studies 27 (6), 803.

Schiele, Jerome, May 1996. Afrocentricity: an emerging paradigm in social work practice. Social Work 41 (3), 284–294.

Sudarkasa, Niara, November/December 1980. African and Afro-American family structure: a comparison. Black Scholar 11, 37–60.

Sudarkasa, Niara, 1981. Interpreting the african heritage in Afro-American family organization. In: McAdoo, Harriette P. (Ed.), Black Families. Sage Publications, Beverly Hills, CA.

Swigorski, M., 1996. Challenging privilege through Africentric social work practice. Social Work 41 (2), 153–161.

Age and Crime

Deinya Phenix, St Francis College, Brooklyn, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Current knowledge about age with regard to crime has a solid foundation amassed with promising theory and research. There is certainty among social scientists about the existence of a relationship between crime and age; that the aggregate age of criminals peaks at late adolescence and early adulthood; and that this 'age-crime curve' exists in both aggregate and individual level data. Scholar perspectives vary on the nature of individual behavior within the curve, patterns, and determinants of criminal activity early in life; the trajectories represented by the aggregate curve; and desistance of criminal activity in middle age or late life.

The life course is often described as a 'pathway,' implying a linear process where a person's life, viewed as a whole, has a direction based on observed behavior and expectations for future behavior. One's life may be said to change direction, 'taking a turn' for the worse or better, based on behavioral changes. The life course has also been described as a series of relatively discrete phases in human development: childhood, adolescence, transitional adulthood, middle adulthood, and the early and late older years. Each stage is associated with distinctive behavioral roles, expectations, and social resources. Related to the concept of stages are the transitions between stages and the timing of such transitions.

These different conceptions of the life course are generally compatible, as both allow social scientists to examine behavior as part of a larger whole of human experience. With regard to the study of crime and deviance, some controversies do arise when considering how much significance the life course has in explaining behavior.

As social roles and expectations change with age, deviant behavior and delinquency have been tied to the life course. The longitudinal study of crime has developed from empirical observation of criminal behavior by age and observations in support or in contrast to the normative expectation that those who participate in deviance, particularly crime, age out of such behavior. The developmental and life course subfield in criminology gained momentum as a result of a series of debates on the importance of taking age into account in theory and research on crime, on the stability of deviance through the life course, and on the types and number of trajectories in a population. These debates are fundamental to the concept of the criminal career, a notion which has itself been debated publicly for a century (Glueck and Glueck, 1940; Lemert, 1951; Erikson, K.T., 1962, 1966).

Dimensions in DLC Criminology

Toward understanding the relationship between age and deviant or criminal behavior, published longitudinal research has examined age of onset, i.e., when the behavior under observation began, usually described in the binary as early or late onset. Prevalence, an aggregate measure of whether or not individuals participate in certain deviant acts, is examined separately from frequency – the incidence of deviant behavior among those individuals who do participate (Blumstein et al., 1985; Petras et al., 2010). Both variables contribute to the relationship between age and crime, and both are influenced by variables demonstrated to affect criminal behavior. Severity, a measure of the extent self or societal harm, can also vary by age.

Other longitudinal dimensions include career length or persistence, also expressed as continuity (e.g., Nagin and Farrington, 1992a; Ezell and Cohen, 2005; Bushway et al., 2009). Desistance, the cessation of the behavior under observation, is of interest to criminologists and other behavioral health scientists such as those studying addiction (Bushway et al., 2009; Petras et al., 2010; Blonigen, 2010; Hser and Anglin, 2011), as is intermittency, gaps in the behavior of interest, which has been shown to increase with age (Baker et al., 2013).

The concepts mentioned so far are generally homotypic patterns, i.e., the continuity or change of one type of behavior. Heterotypic longitudinal patterns, i.e., qualitative shifts in behavior, are also of interest (Laub and Sampson, 2003; Farrington et al., 2006). Nieuwbeerta et al. (2010) found 'an age–diversity curve,' in which there was evidence of diversification in delinquent behavior between early adolescence and young adulthood and then specialization later in adulthood. Additionally, the specialization is observed even as there is a general decline in frequency with age.

The notion of turning points, life events, or transitions that marks major changes in behavior or status is also an important item on the life course research agenda (Laub and Sampson, 2003; Teruya and Hser, 2010). Good marriages and parenthood are associated with decrease or desistance of delinquent behavior (e.g., Laub et al., 1998). Motherhood, to an even greater extent than marriage, is associated with decreased delinquency in disadvantaged communities (Kreager et al., 2010). The notion of turning points also applies to onset, including late onset, of criminal or deviant behavior (Carr and Hanks, 2012).

The notions of age-related cognitive shifts and identity transformation have proven to be key concepts in elucidating the process of desistance (Veysey et al., 2013; Bushway and Paternoster, 2013). Although measurement problems still preclude large-scale quantitative investigation of these internal and personal dynamics, such processes are considered 'most promising' theoretical explanations of desistance, including disengagement from gang affiliation (Pyrooz and Decker,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45043-9

2011; Sweeten et al., 2013; Flores, 2013). Cognitive transformation has been found associated with treatment and policy interventions; and these transformations have been noted as affecting a variety of criminal careers, including sexual offending (Harris, 2014). Thus, consideration of the effect of age on crime also requires consideration of policy-relevant issues, such as intervention, treatment, and selective incapacitation (Hser and Anglin, 2011; Blumstein et al., 1988).

These time- and age-related factors in a person's behavior link deviant behavior to the general knowledge base about human development. Age or developmental stage is thus considered an important variable, often found to be a risk or protective factor in behavior. However, there was not always consensus on the importance of age and the life course.

Age, Crime, and the Debate about Research Focus

The monumental age–crime debate in the early 80s served as growing pains for the subfield of developmental criminology (Hirschi and Gottfredson, 1983; Greenberg, 1985; Land, 1992; Warr, 1993). In 1983, Hirschi and Gottfredson wrote what served as an open letter to the research community on the trend toward investing increasing amounts of talent and fiscal resources on expensive longitudinal studies, which they argued were unnecessary to understanding the causes of crime.

Writing in reaction to criticism launched against control and other theories that failed to account for aging out and other longitudinal effects, Hirschi and Gottfredson wrote most directly in critique of Greenberg's (1977) theoretical analysis, where he proposed using a combination of strain and control theories to account for variation in criminal behavior, including that variation due to age. Hirschi and Gottfredson argued that theories that adequately differentiate between criminals and noncriminals do not have to account for the effects of age or the various dimensions of the criminal career. Crime, they maintain, can be explained by individual differences.

The typical curve, as depicted in the figure below showing the age of arrestees in detailed national crime reports (e.g., Federal Bureau of Investigation, 2012), can be described as a distinct peak in activity in the late teens, a sudden decline in early adulthood, and a continual decline as age increases. This pattern is consistent with the typical age–crime curve discussed in research literature (Gottfredson and Hirschi, 1988; Greenberg, 1985; Blumstein et al., 1988; Moffitt, 1993) from reports of the US crime rates and similar statistics from the British Home Office (Hansen, 2003). This age distribution, as Gottfredson and Hirschi argue based on similar graphs of crime rates from the United States, England, and Wales across several periods, e.g., 1840, 1908, 1965, and 1977, "is invariant across social and cultural conditions," as well as across genders (Steffensmeier et al., 2006; **Figure 1**).

Because it is ubiquitous, Hirschi and Gottfredson state that the age–crime distribution "cannot be accounted for by any variable or combination of variables currently available to criminology" (1983), as these variables are often confounded with age. In causation terms, they stated that the age effect on crime is direct. Given the primacy of self-control and the stability of between-person differences in self-control in their general theory, Gottfredson and Hirschi argued that explanations of crime at one age suffice to explain crime at any other. Self-control, moreover, is usually set by mid-childhood (8–10 years of age). There is 'stability,' they argued, in the different likelihood that individuals would commit crimes.

Greenberg (1985) demonstrated that the age–crime relationship is not universal but shows demographic

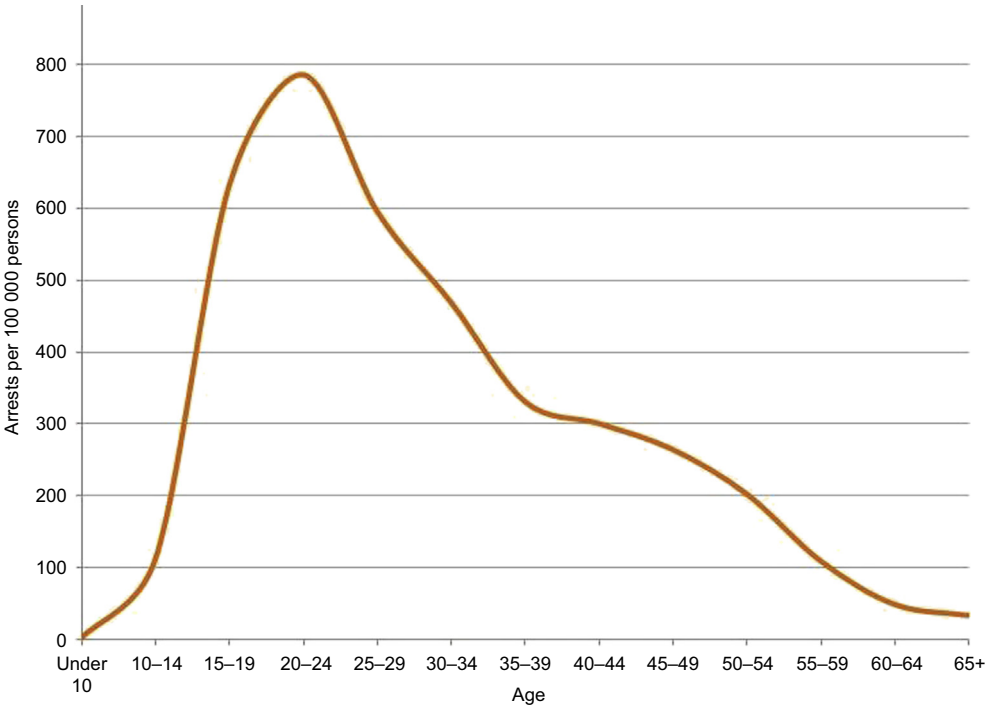

Figure 1 U.S. arrest rates by age. From Federal Bureau of Investigation, 2012. Crime in the United States.

variation and changes over historical periods. He cited several examples of socioeconomic status, gender, and race differences in the age distribution of crime, measured by arrests and survey responses from a variety of settings and locales. He also presented shifts in peak ages across time and evidence of social and demographic effects on desistance.

Warr (1993), responding to the statement that the age distribution of crime cannot be explained by any known variables (Hirschi and Gottfredson, 1983; Gottfredson and Hirschi, 1990), used National Youth Survey data to investigate how differential association theory can explain the age distribution. Exposure to delinquent peers, time spent with peers, and loyalty to peers changes dramatically from 11 to 21 years of age, following the same pattern as crime. Peers have more of an influence in early years, especially the teen years, than later in life. Moreover, the set of predictors of onset at early ages is somewhat different from the set that would predict onset at later ages, and these sets are different from the predictors of persistence in crime (Nagin and Farrington, 1992b).

More recently, Sweeten et al., (2013) demonstrated the relative importance of developmental and criminological concepts in explaining the reduction of criminal behavior over a 7-year period observed in the Pathways to Desistance study. The use of a multilevel model which allowed precise assessment of change in correlates such as psychosocial maturity and social control demonstrated that the effect of age on decrease in criminal behavior in their sample is not direct as argued by Gottfredson and Hirschi.

Developmental criminology, borrowing human development concepts from psychology and expanding on control theory, social learning, strain and symbolic interaction, enhances the notion of criminal career with time-dependent models of change in behavior (Farrington, 2003). Processes such as 'aging out' of crime (Gottfredson and Hirschi, 1986, 1990; Sampson and Laub, 2003) necessitate a longitudinal perspective and the assumption of the possibility of change.

Career Criminals and Population Heterogeneity

Age–crime curves for individuals do not resemble the aggregate curve (Farrington, 1986), and there is an extensive individual variation in the timing of onset, incidence, desistance, and the correlates of these trajectories. The earliest criminologists have noted that some members of the population commit more crimes than others (Lombroso, 2006). While the various crime indicators vary by age, researchers have argued that a certain small segment of the population is responsible for a large percentage of delinquent acts (Blumstein et al., 1985; Wolfgang et al., 1972). Wolfgang et al., (1972) found, in addition to large race and socio economic status (SES) differences in offending, severity, and repeat offending, that 6% of the sample was responsible for 52% of offenses.

Those who commit delinquent acts earlier in life tend to commit more crime later. These early-onset, (and potentially) chronic offenders are the subject of a large amount of research using cohort data sets (e.g., Blumstein et al., 1985). However, Blumstein et al., (1985) took special pains to disambiguate the concepts of *criminal career* and *career criminal*. The latter, they noted, was more sensational, resonating more with policy interests but not as theoretically interesting as the former.

Analyzing the Cambridge Study in Delinquency Development panel of data on 411 boys followed into manhood, Nagin and Farrington (1992a,b) found an inverse association between the age of onset and future conviction. Personal characteristics and family factors, such as the 'daring' nature of the boy, the criminal history and attitudes of his parents, and separation from parents, have some bearing on the boy's onset of criminal behavior and its continuation, lending credence to 'persistent heterogeneity theories' of behavioral change (Gottfredson and Hirschi, 1990; Wilson and Herrnstein, 1985).

The heterogeneity perspective on behavioral change resonates with some of the more classic debates on deviant careers. For instance, this debate was foreshadowed, according to Laub and Sampson (1991), by the debate between Sutherland and the Gluecks, who "sought to understand the age–crime curve in terms of maturational reform" as they investigated their hypothesis regarding the stability of deviance (Laub and Sampson, 1991: 1410), a 'sticking point' generating opposition from those tied to a more sociological perspective, such as Sutherland's differential association. The Glueck focus, in analyzing differences between delinquents and nondelinquents, was 'multidisciplinary' and took a 'multiple causal approach' but also intensely individual – investigating family and other microlevel factors such as "school, opportunities (peers and the use of leisure time), formal sanctions (e.g., arrest, probation, prison), personality development, temperament, and constitutional factors such as body structure (e.g., mesomorphy)" (Laub and Sampson, 1991).

By contrast, Sutherland took a more sociological position and objected to the Gluecks' multiple factor, individual level 'factual search,' devoid of a formal, articulated theory of persistent criminality (Laub and Sampson, 1991). The substantive contention between Sutherland and the Gluecks, as between Hirschi and Gottfredson and other scholars, stemmed from the fundamental relevance of age to crime, which Sutherland dismissed as "slightly more than chance" (quoted in Laub and Sampson, 1991: 1422), and the stability of antisocial behavior over the life course, which he denied.

Another contention was the Gluecks' practical ambition to "influence social policy through the use of their prediction tables"; to improve the process of decision-making by judges, probation officers, parole boards, and military officials; and to identify potential delinquents at school age or perhaps even as early as 2 or 3 years of age to "make possible the application of treatment measures that would be truly crime preventive" (1950, cited in Sampson and Laub, 1993). The Gluecks' ambitions have resonated in much of the applied social research, particularly contemporary study of serious offenders, chronic offenders, selective incapacitation, early intervention, and targeted crime prevention (Blumstein et al., 1985; Blumstein et al., 1988).

State Dependence/Persistence and Desistance

Modern longitudinal research has investigated the notion of state dependence, i.e., the temporal association between prior/current and future behavior (D'Unger et al., 1998;

Skardhamar, 2010). Conceptualized this way, prior offending is still an etiological factor, but the analytic focus is less on a particular age of onset and more focused on the timing of prior incidence. The issue here is one of persistence: current deviants are more likely to remain deviant in the near future, either because their circumstances or social networks are likely to remain the same or worse due to their behavior (Sampson and Laub, 1993), or because latent, time-stable antisocial traits make them more likely to persist (Gottfredson and Hirschi, 1990). Both explanations have implications for differences in short- and long-term trajectories.

Desistance, in another extension of the maturational reform perspective lent to the discipline by the Gluecks, is conceived as a socially nuanced and incremental process (Laub and Sampson, 2003; Bushway and Paternoster, 2013). Ezell and Cohen, in their book based on the analysis of arrest histories in a delinquent population identified by the California Youth Authority in 1981–82, 1986–87, and 1991–92, found some continuity of arrest patterns (state dependence) in shorter rather than longer periods of time. However, their findings included a surprisingly large amount of variation ('heterogeneity') in longer-term longitudinal patterns among delinquents, lending little support to theories of stability.

Laub and Sampson (2003), in a study unparalleled in depth, found continuity in antisocial behavior among delinquents followed over decades and observed via retrospective interviews. However, like others, they found extensive variation in the persistence of deviant and criminal behavior. Desistance, which Sampson and Laub show to be more common than the career criminal model would suggest, is related to social bonds (e.g., family and employment) and the informal social control these bonds generate. Adult attachments are associated with increased social resources, which are, in turn, associated with decreased antisocial behavior.

The acquisition of these social resources led to desistance from delinquent behavior but not certain deviant behavior such as alcohol abuse and violence. Important macrolevel structural factors such as the GI Bill played a part in the directions the interviewee's lives took. Thus, behavioral change over the life course is said to be influenced by "interaction of structure and choice" (2003).

The Sampson and Laub study was not of the same magnitude as the Ezell and Cohen and other large longitudinal studies (e.g., Wolfgang et al., 1987; Bushway et al., 2009) emerging from a delinquent population, or Farrington and West (Farrington et al., 2006), emerging from a general population. However, no longitudinal study has achieved the depth of the Laub and Sampson study in investigating underlying processes in delinquent and deviant behavior over the life course.

Based on evidence of various types (survey research, longitudinal designs, criminal histories, or narrative studies), desistance is shown to be the norm for most offenders (Sampson and Laub, 2005b; Maruna, 1999; Massoglia and Uggen, 2010).

Behavioral Trajectories and Group-Based Analysis

Due to technological and methodological advances in research subject tracking and longitudinal data collection, the longitudinal study of crime has reached paradigmatic status. Following the Gluecks, the research of Wolfgang and others working on the Philadelphia birth cohort reflected a lasting interest in chronic offenders and the debate over what distinguishes chronic offenders. This interest developed into research questions such as whether chronic offenders are a separate type of person, with separate causes for their behavior and a set of characteristics that can be used to identify them.

Moffitt (1993) proposed the existence of two behavioral groups, defined as "qualitatively distinct categories of individuals … each in need of its own distinct theoretical explanation" (Moffitt, 1993: 674). The first of these groups consists of those offenders who temporarily participate in delinquent acts, usually in the teen years, and promptly desist. The second group persists with delinquency throughout the life course. This proposition represented a major advance in the rapidly developing subfield of developmental criminology (Loeber and Le Blanc, 1990; Laub and Sampson, 1993; Nagin and Farrington, 1992a,b), and it subsequently influenced well over 200 articles in sociology, psychology, education, criminology, political science, and criminal justice.

Nagin and Land (1993) made an equally important impact by presenting statistical models that depict population heterogeneity in behavioral trajectories. In their widely read paper, they argued that such heterogeneity can be estimated with mutually exclusive groups based on the observed data. Generally, these models are estimated with an algorithm that 'searches' a longitudinal data set (Moffitt, 2006), posits a user-supplied number of trajectory groups, and then assigns a probability of membership in each group to each case. Analysts can also estimate the proportion of the population in each group.

With the development of group-based modeling procedures in popular statistical software (Jones et al., 2001; Deb, 2008; Muthén, 2004), a growing body of research has emerged using these techniques and a variety of behavioral data (Roeder et al., 1999; Jones et al., 2001; Piquero et al., 2007). A hypothetical set of trajectories, including those proposed by Moffitt, are depicted in the figure below (**Figure 2**).

The key innovation with these methods is the ability to describe "not just the average path but the full distribution of individual trajectories" (Bushway et al., 2009). Thus, a modern feature of longitudinal study is the examination of trajectories – onset, peak, and desistance – as whole units rather than separate dimensions. But the discoveries here were not just methodological; trajectory group presented a new agenda for sociological theory (D'Unger et al., 1998).

Some controversy has followed the group-based trajectory methods. Piquero (2008), who reviewed over 80 studies using group-based trajectory, noted that the source of sample, i.e., offenders versus general population, the length of the observation, and other methodological choices make a difference in the number of groups. In fact, group membership is unstable, especially among 'high-rate chronic' offenders, and highly sensitive to the inclusion of incarceration time (Eggleston et al., 2004). Bauer and Curran (2003) noted a higher risk of falsely inferring groups when using these methods on skewed data.

Group-based discussion, particularly the use of strongly intuitive labels such as 'adolescent limiteds,' 'high-level chronics,' and 'low-level chronics' (Nagin and Land, 1993;

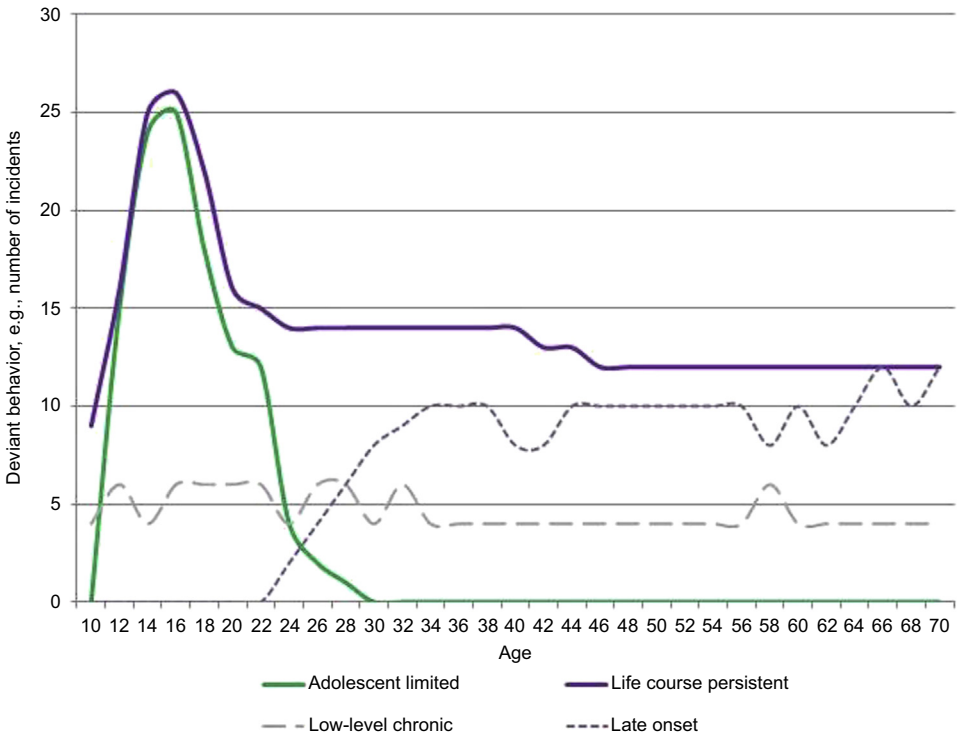

Figure 2 Example of four behavioral trajectories, ages 10 through 70.

Nagin et al., 1995), risks "misappropriation by those predisposed to believe in the idea of a high-rate group" (Sampson and Laub, 2005a, paraphrased by Piquero, 2008). Once groups are 'identified' with statistical models, researchers are inclined to discuss them as *actual groups*, often in the sense that scientists speak of 'natural categories' existing in nature (Skardhamar, 2010). Thus, the policy implications of group-based analysis, particularly the potential for targeting crime prevention and intervention resources (D'Unger et al., 1998; Sampson and Laub, 2005a; Nagin, 2005), are more negative than positive.

Much of the contention around group-based methods questions the assumption of discrete groups. Simulation studies based on artificial data with and without groups (Greenberg and Ezell, 2011; Bauer and Curran, 2003; Skardhamar, 2010) have demonstrated the possibility of finding groups where there were none.

Thus, as Raudenbush (2005) suggested when he wrote, "perhaps we are better off assuming continuously varied growth a priori and therefore never tempting our audience," some recent research has been done with this more relaxed assumption (Telesca et al., 2012; Greenberg and Ezell, 2011; Phenix, 2010). Using visual representation and statistical tests on the distribution of estimated parameters (intercepts and slopes) from multilevel models, these studies estimated the whole population of trajectories in a 19-year panel data of delinquent men (Greenberg and Ezell, 2011) and a 40-year cohort study of men from a general population (Phenix, 2010). Testing the assumption of trajectory groups by first examining the continuum of criminal behavior trajectories, these studies found no strong evidence of trajectory groups. Telesca et al., (2012) also demonstrated the use and importance of more flexible models of deviant behavior based on data from a general population.

All of the studies on behavioral trajectories demonstrate that change over the life course is not universal. However, the variation in how, how much, and when the behavior changes is of continued interest to social scientists. In sum, the shape of the distribution of deviant behavior – and of change in that behavior – is a fundamental question of long-term interest in criminology (Glueck and Glueck, 1940; Sampson and Laub, 1993, 2003; Moffitt, 2003). In investigating how change or the absence of change is connected to and influenced by the social context and the functioning of social institutions at both individual and structural levels, the study of age as related to crime remains an important lens for connecting human behavior and the age-graded set of expectations, resources, and challenges faced by people and the society that they comprise.

See also: Control Theories and Crime; Crime, Desistance, and Persistence; Crime, Sociology of; Crime: Knowledge about and Prevalence; Criminology: Development and Current Status of the Discipline; Data Bases and Statistical Systems: Crime Measurement; Data Bases and Statistical Systems: Crime and Delinquency; Data Bases and Statistical Systems: Sociology; Delinquency, Sociology of; Government Statistics; Growth Curve Analysis; Life Course Criminology; Longitudinal Data; Mixture Models in Statistics; Panel Surveys: Uses and Applications; Quantitative Methods and Crime; Recidivism; Rehabilitation of the Offender.

Bibliography

Aiyer, S.M., Williams, J.L., Tolan, P.H., Wilson, M.N., 2013. Predicting desistance in a high-risk sample: examining the influence of individual and contextual factors. Journal of Community Psychology 41, 408–424. http://dx.doi.org/10.1002/jcop.21545.

Akers, Ronald L., 1998. Social Learning and Social Structure: A General Theory of Crime and Deviance. Northeastern University Press, Boston.

Baker, T., Metcalfe, C.F., Piquero, A.R., January 2013. Measuring the intermittency of criminal careers. Crime & Delinquency. http://dx.doi.org/10.1177/0011128712466382.

Bauer, Daniel J., Curran, Patrick J., 2003. Distributional assumptions of growth mixture models: implications for over extraction of latent trajectory classes. Psychological Methods 8, 338–363.

Bersani, Bianca, Nieuwbeerta, Paul, Laub, John, 2009. Predicting trajectories of offending over the life course: findings from a Dutch conviction cohort. Journal of Research in Crime and Delinquency 46, 468.

Best, Joel, 2004. Theoretical issues in the study of social problems and deviance. In: Ritzer, George (Ed.), Handbook of Social Problems: A Comparative International Perspective. Sage, Thousand Oaks.

Blokland, Arjan A.J., Nieuwbeerta, Paul, 2005. The effects of life circumstances on longitudinal trajectories of offending. Criminology 43, 1203–1240.

Blokland, A.A.J., Nagin, D.S., Nieuwbeerta, P., 2005. Life span offending trajectories of a Dutch conviction cohort. Criminology 43 (4), 919–954.

Blonigen, Daniel, 2010. Explaining the relationship between age and crime: contributions from the developmental literature on personality. Clinical Psychology Review 30, 89–100.

Blumstein, Alfred, Farrington, David P., Moitra, Soumyo, 1985. Delinquency careers, innocents, desisters, and persisters. In: Tonry, Michael, Morris, Norval (Eds.), Crime and Justice: An Annual Review of Research. University of Chicago Press, Chicago, pp. 187–220.

Blumstein, Alfred, Cohen, Jacqueline, Farrington, David, 1988. Criminal career research: its value for criminology. Criminology 26, 1.

Blumstein, A., Nakamura, K., 2009. Redemption in the presence of widespread criminal background checks. Criminology 47, 327–359.

Brame, Robert, Mulvey, Edward P., Piquero, Alex R., 2001. On the development of different kinds of criminal activity. Sociological Methods and Research 29 (3), 319–341.

Brown, J. David, 2008. The professional ex-: an alternative for exiting the deviant career. In: Clarke, Edward J. (Ed.), Deviant Behavior. Worth, New York.

Burraston, Bert. Answering the Career Criminal Debate: Comparing Finite Mixture Modeling with Growth Mixture Modeling Paper Presented at The 2005 Annual Meeting of the American Sociological Association. http://www.allacademic.com/meta/p23019_index.html (accessed 16.01.10.).

Bushway, Shawn, Sweeten, Gary, Nieuwbeerta, Paul, 2009. Measuring long term individual trajectories of offending using multiple methods. Journal of Quantitative Criminology 25 (3), 259.

Bushway, S.D., Paternoster, R., 2013. Desistance from crime: a review and ideas for moving forward. Handbook of Life-Course Criminology 213–231.

Carr, N.T., Hanks, R.S., 2012. If "60 is the new 40," is 35 the new 15? Late onset crime and delinquency. Deviant Behavior 33 (5), 393–411.

Cohlerm, Bertram, Hammack, Phillip, 2007. The psychological world of the gay teenager: social change, narrative, and "normality". Journal of Youth and Adolescence 36 (1), 47–59.

Curran, Patrick J., Willoughby, Michael T., 2003. Implications of latent trajectory models for the study of developmental psychopathology. Development and Psychopathology 15, 581–612.

Deb, Partha, 2007/2008. FMM: Stata Module to Estimate Finite Mixture Models, Statistical Software Components S456895, Boston College Department of Economics. [Computer Files]. Available at: http://ideas.repec.org/c/boc/bocode/s456895.html#provider.

D'Unger, Amy V., Land, Kenneth C., McCall, Patricia L., Nagin, Daniel S., 1998. How many latent classes of delinquent/criminal careers? Results from mixed Poisson regression analyses. American Journal of Sociology 103 (6), 1593–1630.

Eggleston, Elaine P., Laub, John H., 2002. The onset of adult offending: a neglected dimension of the criminal career. Journal of Criminal Justice 30 (6), 603–622.

Eggleston, Elaine P., Laub, John H., Sampson, Robert J., 2004. Methodological sensitivities to latent class analysis of long-term criminological trajectories. Journal of Quantitative Criminology 20, 1–26.

Erikson, K.T., 1962. Notes on the sociology of deviance. Social Problems 9, 307–314.

Erikson, K.T., 1966. Wayward Puritans: A Study in the Sociology of Deviance. Wiley, New York.

Ezell, Michael E., Cohen, Lawrence E., 2005. Desisting from Crime: Continuity and Change in Long-Term Crime Patterns of Serious Chronic Offenders. Oxford University Press, New York.

Farrington, David P., 1986. Age and crime. Crime and Justice 7, 189–250.

Farrington, David P., 2007. Advancing knowledge about desistence. Journal of Contemporary Criminal Justice 23, 125–134.

Farrington, David P., Coid, J.W., Harnett, L.M., et al., 2006. Criminal Careers up to Age 50 and Life Success up to Age 48: New Findings from the Cambridge Study in Delinquent Development, second ed. Home Office Research Study, London. 299.

Federal Bureau of Investigation, 2012. Crime in the United States. Available at: http://www.fbi.gov/about-us/cjis/ucr/ucr-publications#Crime.

Flores, E.O., 2013. God's Gangs: Barrio Ministry, Masculinity, and Gang Recovery. NYU Press.

Glueck, S., Glueck, E., 1940. Juvenile Delinquents Grown Up. Commonwealth Fund, New York.

Glueck, S., Glueck, E., 1950. Unraveling Juvenile Delinquency. Commonwealth Fund, New York.

Glueck, S., Glueck, E., 1968. Delinquents and Nondelinquents in Perspective. Harvard University Press, Cambridge, MA.

Goode, Erich, 2001. Deviant Behavior. Prentce-Hall, Upper Saddle River, NJ.

Gottfredson, Michael, Hirschi, Travis, 1986. The true value of lambda would appear to be zero: an essay on career criminals, criminal careers, selective incapacitation, cohort studies, and related topics. Criminology 24 (2), 213–234.

Gottfredson, Michael, Hirschi, Travis, 1988. Science, public policy, and the career paradigm. Criminology 26 (1), 37–55.

Gottfredson, Michael, Hirschi, Travis, 1990. A General Theory of Crime. Stanford University Press, Stanford, CA.

Greenberg, David F., Ezell, Michael E., 2011. Criminal careers: Discrete or Continuous? Modern Modeling Methods Conference. University of Connecticut.

Greenberg, David, 1977. Delinquency and the age structure of society. Contemporary Crises 1, 189–223.

Greenberg, David, 1985. Age, crime, and social explanation. American Journal of Sociology 91 (1), 1–21.

Greenberg, David, 1991. Modeling criminal careers. Criminology 29 (1), 17–46.

Greenberg, David, 1992. Comparing criminal career models. Criminology 30 (1), 141–148.

Grove, Walter R., December 2004. The career of the mentally ill: an integration of psychiatric, labeling/social construction, and lay perspectives. Journal of Health and Social Behavior 45, 357–375.

Hansen, Kristine, 2003. Education and the crime-age profile. British Journal of Criminology 43, 141–168.

Harris, Danielle A., June 2014. Desistance from sexual offending: findings from 21 life history narratives. Journal of Interpersonal Violence 29 (9), 1554–1578. http://dx.doi.org/10.1177/0886260513511532.

Harry, Joseph, 1993. Being out: a general model. Journal of Homosexuality 26 (1), 25–39.

Hartigan, J.A., Hartigan, P.M., 1985. The dip test of unimodality. Annals of Statistics 13, 70–84.

Hirschi, Travis, Gottfredson, Michael R., 1983. Age and the explanation of crime. American Journal of Sociology 89, 552–584.

Holstein, James A., Gubrium, Jaber F., 2000. Constructing the Life Course, second ed. General Hall, Dix Hills, NY.

Holstein, James A., Gubrium, Jaber F., 2007. Constructionist perspectives on the life course. Sociological Compass 1.

Hendershott, Anne B., 2002. The Politics of Deviance. Encounter Books, San Francisco.

Horney, Julia, Osgood, David W., Marshall, Ineke Haen, 1995. Criminal careers in the short-term: intra-individual variability in crime and its relation to local life circumstances. American Sociological Review 60, 655–673.

Hostetler, Andrew, Cohler, Bertram, 1997. Partnership, singlehood, and the lesbian and gay life course: a research agenda. International Journal of Sexuality and Gender Studies 2 (3), 199–230.

Hser, Yih-Ing, Anglin, M. Douglas, 2011. Addiction treatment and recovery careers. In: Kelly, J., White, W. (Eds.), Addiction Recovery Management: Theory, Research and Practice. Humana Press, New York.

Hser, Yih-Ing, Evans, Elizabeth, Huang, David, Brecht, Mary-Lynn, Li, Libo, 2008a. Comparing the dynamic course of heroin, cocaine and methamphetamine use over 10 years. Addictive Behaviors 33 (12), 1581.

Hser, Yih-Ing, Huang, David, Brecht, Mary-Lynn, Li, Libo, Evans, Elizabeth, 2008b. Contrasting trajectories of heroin, cocaine and methamphetamine use. Journal of Addictive Disorders 27 (3), 13.

Jones, Bobby, Nagin, Daniel S., Roeder, Kathryn, 2001. A SAS procedure based on mixture models for estimating developmental trajectories. Sociological Methods and Research 29 (3), 374–393.

Kazemian, Lila, Farrington, David, LeBlanc, Marc, 2009. Can we make accurate long-term predictions about patterns of de-escalation in offending behavior? Journal of Youth and Adolescence 38, 384–400.

Kerr, Thomas, Fast, Danya, Small, Will, Wood, Evan, 2009. "Coming down here": young people's reflections on becoming entrenched in a local drug scene. Social Science and Medicine 69 (8), 1204–1210.

Kertzner, Robert M., 2001. The adult life course and homosexual identity in midlife gay men. Annual Review of Sex Research 12, 75–92.

Kreuter, Frauke, Muthén, Bengt, March 2008. Analyzing criminal trajectory profiles: bridging multilevel and group-based approaches using growth mixture modeling. Journal of Quantitative Criminology 24, 1–31.

Kreager, Derek A., Matsueda, Ross L., Erosheva, Elena A., 2010. Motherhood and criminal desistance in disadvantaged neighborhoods. Criminology 48 (1), 221–258.

Lemert, Edwin M., 1951. Social Pathology: A Systematic Approach to the Theory of Sociopathic Behavior. McGraw-Hill, New York.

Land, Kenneth, 1992. Models of criminal careers: some suggestions for moving beyond the current debate. Criminology 30 (1), 149–156.

Laub, John H., Sampson, R.J., 1991. The Sutherland-Glueck debate: on the sociology of criminological knowledge. American Journal of Sociology 96, 1402–1440.

Laub, John H., Sampson, Robert J., 1993. Turning points in the life course: why change matters to the study of crime. Criminology 31, 301.

Laub, John H., Sampson, Robert J., 2001. Understanding desistance from crime. In: Tonry, Michael (Ed.), Crime and Justice: A Review of Research. University of Chicago Press, Chicago, IL, pp. 1–69.

Laub, John H., Sampson, Robert J., 2003. Shared Beginnings, Divergent Lives. Harvard University Press, Cambridge, MA.

Laub, John H., Nagin, Daniel S., Sampson, Robert J., 1998. Trajectories of change in criminal offending: good marriages and the desistance process. American Sociological Review 63, 225–238.

Loeber, Rolf, Le Blanc, Marc, 1990. Toward a developmental criminology. Crime and Justice 12, 375.

Loeber, R., Farrington, D.P., Stouthamer-Loeber, M., Moffitt, T., Caspi, A., 1998. The development of male offending: key findings from the first decade of the Pittsburgh youth Study. Studies in Crime and Crime Prevention 7, 141–172.

Lombroso, Cesare, 2006. Selections from Criminal Man. Translated by Mary Gibson, Nicole Hahn Rafter. Duke University Press, Durham.

Maruna, S., 1999. Desistance and development: the psychosocial process of going straight. In: The British Criminology Conferences: Selected Proceedings, vol. 2.

Massoglia, Michael, Uggen, Christopher, 2010. Settling down and aging out: toward an interactionist theory of desistance and the transition to adulthood. American Journal of Sociology 116 (2), 543.

Matza, David, 1964. Delinquency and Drift. Wiley, New York.

Moffitt, Terrie, 1993. "Adolescence-limited" and "life-course-persistent" antisocial behavior: a developmental taxonomy. Psychological Review 100, 674–701.

Moffitt, Terrie E., 2003. Life-course-persistent and adolescent-limited antisocial behaviour: a ten-year research review and a research agenda. In: Lahey, Benjamin B., Moffitt, Terrie E., Caspi, Avshalom (Eds.), Causes of Conduct Disorder and Juvenile Delinquency. The Guilford Press, New York/London.

Moffitt, Terrie E., 2006. Life-course persistent versus adolescent-limited antisocial behavior. In: Cicchetti, Dante, Cohen, Donald J. (Eds.), Developmental Psychopathology, vol. 3. John Wiley, NJ.

Muthén, B., 2004. Latent variable analysis: growth mixture modeling and related techniques for longitudinal data. In: Kaplan, D. (Ed.), Handbook of Quantitative Methodology for the Social Sciences. Sage, Newbury Park, CA, pp. 345–368.

Nagin, Daniel S., 1999. Analyzing developmental trajectories: a semiparametric, group-based approach. Psychological Methods 4, 139–157.

Nagin, Daniel S., 2005. Group-Based Modeling of Development. Harvard University Press, Cambridge, MA.

Nagin, Daniel S., 2007. Overview of a semi-parametric, group-based approach for analyzing trajectories of development. In: Flannery, Daniel J., Vazsonyi, Alexander T., Waldman, Irwin D. (Eds.), Cambridge Handbook of Violent Behavior and Aggression. Cambridge University Press, New York, pp. 740–749.

Nagin, Daniel S., Farrington, David P., 1992a. The onset and persistence of offending. Criminology 30, 501.

Nagin, Daniel S., Farrington, David P., 1992b. The stability of criminal potential from childhood to adulthood. Criminology 30, 235–260.

Nagin, Daniel S., Farrington, David P., Moffitt, Terrie E., 1995. Life-course trajectories of different types of offenders. Criminology 33, 111–139.

Nagin, Daniel S., Land, Kenneth C., 1993. Age, criminal careers, and population heterogeneity: specification and estimation of a nonparametric, mixed Poisson model. Criminology 31, 327–362.

Nagin, Daniel S., Tremblay, Richard E., 2005. Developmental trajectory groups: fact or a useful statistical fiction? Criminology 43, 873–904.

Nieuwbeerta, Paul, Blokland, Arjan A.J., Piquero, Alex R., Sweeten, Gary, 2010. A life course analysis of offense specialization across age: introducing a new method for studying individual specialization over the life course. Crime & Delinquency 57 (1), 3–28.

Orcutt, James D., 1983. Analyzing Deviance. Dorsey Press.

Osgood, D. Wayne, 2005. Making sense of crime and the life course. The Annals of the American Academy of Political and Social Science 602, 196–211.

Petras, Hanno, Nieuwbeerta, Paul, Piquero, Alex R., 2010. Participation and frequency during criminal careers across the life span. Criminology 48 (2), 607–637.

Phenix, Deinya M., 2010. Criminal Careers: The Distribution of Change. Doctoral Dissertation. New York University. ProQuest/UMI Publication No. AAT 3427963.

Piquero, Alex R., Farrington, D.P., Blumstein, A., 2003. The criminal career paradigm. In: Tonry, M. (Ed.), Crime and Justice: A Review of Research, vol. 30. University of Chicago Press, Chicago, pp. 359–506.

Piquero, Alex R., Farrington, David P., Blumstein, Alfred, 2007. Key Issues in Criminal Career Research. New Analysis of the Cambridge Study in Delinquent Development. Cambridge University Press, Cambridge.

Piquero, Alex R., 2008. Taking stock of developmental trajectories of criminal activity over the life course. In: Liberman, Akiva M. (Ed.), The Long View of Crime: A Synthesis of Longitudinal Research. Springer, New York.

Pyrooz, D.C., Decker, S.H., 2011. Motives and methods for leaving the gang: understanding the process of gang desistance. Journal of Criminal Justice 39 (5), 417–425.

Raudenbush, Stephen W., 2005. How do we study what happens next? The Annals of the American Academy of Political and Social Science 602, 131–144.

Roeder, Kathryn, Lynch, Kevin G., Nagin, Daniel S., 1999. Modeling uncertainty in latent class membership: a case study in criminology. Journal of the American Statistical Association 94, 766–776.

Sampson, Robert J., Laub, John H., 1993. Crime in the Making: Pathways and Turning Points through Life. Harvard University Press, Cambridge, MA.

Sampson, Robert J., Laub, John H., 2003. Life-course desisters? Trajectories of crime among the delinquent boys followed to age 70. Criminology 41, 301–339.

Sampson, Robert J., Laub, John H., 2005a. Seductions of method: rejoinder to Nagin and Tremblay's Developmental trajectory groups: fact or fiction? Criminology 43, 905–914.

Sampson, Robert J., Laub, John H., 2005b. A life course view of the development of crime. Annals of the American Academy of Political and Social Science 602, 12–14.

Simon, David R., 2005. Elite Deviance. Allyn and Bacon, Needham Heights, MA.

Skardhamar, Torbjørn, 2009. Reconsidering the theory on adolescent-limited and life-course persistent anti-social behaviour. British Journal of Criminology 49 (6), 863–878.

Skardhamar, Torbjørn, 2010. Distinguishing facts and artifacts in group-based modeling. Criminology 48, 295–320.

Steffensmeier, Darrell, Zhong, Hua, Ackerman, Jeff, Shwartz, Jennifer, Agha, Suzanne, 2006. Gender gap trends for violent crimes, 1980 to 2003: a UCR-NCVS comparison. Feminist Criminology 1, 72–98.

Sweeten, G., Pyrooz, D.C., Piquero, A.R., 2013. Disengaging from gangs and desistance from crime. Justice Quarterly 30 (3), 469–500.

Sweeten, G., Piquero, A.R., Steinberg, L., 2013. Age and the explanation of crime, revisited. Journal of Youth and Adolescence 42 (6), 921–938.

Telesca, D., Erosheva, E.A., Kreager, D.A., Matsueda, R.L., 2012. Modeling criminal careers as departures from a unimodal population age–crime curve: the case of marijuana use. Journal of the American Statistical Association 107 (500), 1427–1440.

Teruya, Cheryl, Hser, Yih-Ing, 2010. Turning points in the life course: current findings and future directions in drug use research. Current Drug Abuse Review 3 (3), 189–195.

Thornberry, Terence P., Moore, Melanie, Christenson, R.L., 1985. The effect of dropping out of high school on subsequent criminal behavior. criminology 23 (1), 3–18.

Tittle, Charles, Ward, David A., 1993. The interaction of age with the correlates and causes of crime. Journal of Quantitative Criminology 9, 3–53.

Uggen, Christopher, Thompson, Melissa, 2003. The socioeconomic determination of ill-gotten gains: within-person changes in drug use and illegal earnings. American Journal of Sociology 109, 146–185.
Veysey, B.M., Martinez, D.J., Christian, J., 2013. "Getting out:" a summary of qualitative research on desistance across the life course. Handbook of Life-Course Criminology 233–260.
Wadsworth, Tim, 2006. The meaning of work: conceptualizing the deterrent effect of employment on crime among young adults. Sociological Perspectives 49 (3), 343–368.
Warr, Mark, 1993. Age, peers, and delinquency. Criminology 31 (1), 17–40.
Weitekamp, Elmar G.M., Kerner, Hans-Juergen, Schindler, Volkhard, Schubert, Axel, 1995. On the 'dangerousness' of chronic/habitual offenders: a re-analysis of the 1945 Philadelphia birth cohort data. Studies on Crime and Crime Prevention 4 (2), 159–175.
West, Donald J., 1982. Delinquency: Its Roots, Careers and Prospects. Heinemann, London.
West, Donald J., Farrington, David P., 1977. The Delinquent Way of Life. Third Report of the Cambridge Study in Delinquent Development. Heinemann Educational Books, London.
Wikström, Per-Olof H., Sampson, Robert J. (Eds.), 2006. The Explanation of Crime: Context, Mechanisms and Development. Cambridge University Press, Cambridge.
Wilson, James Q., Herrnstein, Richard, 1985. Crime and Human Nature. Simon and Schuster, New York.
Wolfgang, Marvin E., Figlio, Robert M., Sellin, Thorsten, 1972. Delinquency in a Birth Cohort. University of Chicago Press, Chicago, IL.
Wolfgang, Marvin E., Thornberry, Terence P., Figlio, Robert M., 1987. From Boy to Man, from Delinquency to Crime: Studies in Crime and Justice. University of Chicago Press, Chicago, IL.

Relevant Websites

http://www.fbi.gov/stats-services/fbi-resources
http://www.icpsr.umich.edu/icpsrweb/NACJD
http://www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/DataPortal/GSOEP.aspx

Age: Anthropological Aspects

Paul Spencer, University of London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Age-related problems (e.g., adolescence, mid-life crisis, or old age) have to be viewed in the context of the life course as a whole. The status of older people is bound up with the evolution of the family. In rural societies, the ambivalent regard for older people is especially associated with family tensions concerning rights in property, where the most senior members are the custodians of the land. Where there is no permanent property associated with the family, as among urban migrants or marginalized groups, there tends to be a loss of status associated with aging. The experience of aging is examined from a number of points of view: women have a different profile of experience from men; aging entails a discontinuous dialectical process whereby mounting contradictions within the family lead to critical changes, highlighted by rites of transition; the experience of aging has to be disentangled from the experience of historical change. In certain societies, age-set systems institutionalize the process of maturation and aging. The sheer complexity of these systems provide a unique opportunity to explore different facets of this process, notably in terms of tensions associated with the succession of generations within the family.

Age is a product of the process of aging, which is partly determined by the social environment. Recognition of this social component has led to various attempts to identify features of the life course historically, portraying them as unique to Western civilization. However, the ethnographic literature has revealed a considerable blurring of some of these stereotypes. Thus Philippe Aries' (1962) influential argument that 'childhood' as opposed to 'adulthood' is essentially a product of the industrial revolution may be challenged with reference to the widespread practice of initiation in other cultures, which frequently marks a transitional point in the life course, distinguishing distinct stages, and is associated with a range of beliefs associated with childhood and personal development (La Fontaine, 1985). Again, G. Stanley Hall (1904) is credited with 'discovering' adolescence as a further category that arose out of the industrial cities in America, leading to a developing interest in this topic; yet deviant subcultures associated with dispossessed youth have been noted in some traditional rural settings and extend even to studies of primate behavior (Spencer, 1965; Pereira and Fairbanks, 1993). Again, Leo Simmons' (1945) survey of the role of the aged in 'primitive' societies is often cited, suggesting that older people were highly respected as compared with the West, where the family has diminishing importance in the process of urbanization (Cowgill and Holmes, 1972). Yet the data for rural societies reveal a varied response to the problem of reconciling respect for the experience of older people with the frustrations arising from their overbearing power within the family and the liability of having to care for them.

The complexity of the problem has led to the development of more refined concepts among sociologists, focusing on particular stages of the life course as prime topics for investigation by specialized subdisciplines. In contrast to this, anthropological studies aim to be holistic, emphasizing that within a particular culture any role or status associated with age has to be viewed in the context of the life course as a whole. Thus, the marginalization of certain age categories – perhaps adolescence or old age – forms part of a wider pattern involving a complex of relations between young and old. Adolescent subcultures, for instance, may be viewed as an alternative to the authoritarian structure of the family, and a milieu where the more open-ended bonds between age peers sharpen their awareness of an alternative experience involving a more creative lifestyle, preparing them for future possibilities within the wider community. Again, some studies have indicated that stress and hardship during adulthood, perhaps a mid-life crisis, appears to reinforce people's ability to cope with the social discomforts of old age in due course. Conversely, a more cushioned life career appears to leave people more vulnerable to the sense of loss and isolation as they grow old.

The Family and Ambivalence toward Aging

Very broadly, ambivalent regard for old age is especially associated with the elaboration of the family as a corporate unit in preliterate agricultural societies. The evolution of the family, reaching a peak in such societies, is entwined with the evolution of age relations and the concept of 'Status' coined by Henry Maine (1861), where social position is ascribed by membership of a family. This is closely associated with power in the hands of the most senior members by age as the legitimate custodians of family tradition. In this milieu, cultural knowledge itself may be treated as a form of property, to be imparted or withheld. In a very pertinent sense, property relations within the family are age relations, creating powerful bonds and tensions (Foner, 1984). A common factor underlying the urban stereotypes of old age and adolescence is the demise of the family as a dominant institution (Maine's 'Status to Contract').

The premise of respect and even fear for older people is especially widespread in rural Africa, where it is often associated with strong patrilineal families and polygyny, which is higher in Africa than any other continent. This highlights the notion that children are the property of the family in 'Status-dominated' societies, giving the senior generation the power to marry off their daughters early and to delay the marriage of sons for perhaps a decade or more. So long as this regime can be maintained, extended bachelorhood facilitates widespread polygyny among older men. The array of life stages is illustrated in **Figure 1** with reference to the Samburu of Kenya, who

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12006-9

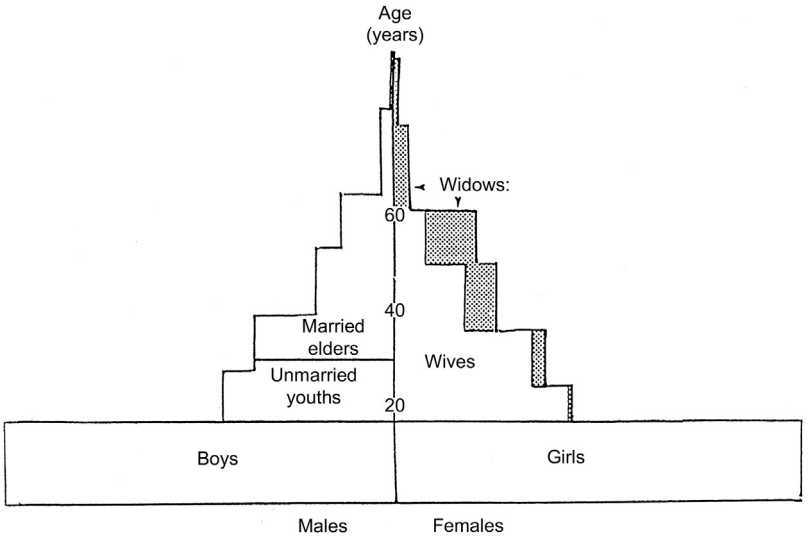

Figure 1 Age, status, and the demographic profile of a polygynous society (Samburu, 1960).

provide a clear-cut example, with no dispensation for younger men to marry early or for widows to remarry. The concave shape of the age distribution is characteristic of preliterate societies, where mortality rates are especially high among the young.

The figure also indicates the contrasting life trajectories of women who are married young to much older men. The depressed status of women in such societies has to be viewed in relation to their total life course. Whereas a boy tends to take the first step from the obscurity of childhood toward a promising adulthood with his initiation, the arranged marriage of a girl transforms her from the obscurity of childhood to an initially obscure role as a young wife with very restricted opportunities and a stranger in her new home. A view of women as the victims of male exploitation is particularly apt at this low point in their careers, and especially in societies where there is a sharp separation between male and female domains. An alternative viewpoint portrays women as active agents who can manipulate their depressed situation to their own advantage. This view becomes increasingly apt as a woman's life course develops and notably once she surmounts the restrictions of her reproductive years and is increasingly independent of her aging husband. In the prime of middle-age, the advantage still rests with men. However, this tends to be reversed beyond this point as men lose the will to assert themselves against younger more competitive successors and are edged to the margins of activity in community affairs. Women are less hampered by aging until they are too frail to play an active role. The process of growing old for women in these circumstances is to free themselves from the domestic routine, but not from their personal networks which focus on their growing family with a freedom to choose how they wish to involve themselves (Amoss and Harrell, 1981).

The Experience of Maturation and Aging

The experience of maturation and aging shapes the perception of time and hence its meaning in a quite fundamental sense. This may be analyzed in several ways. The first involves an autobiographical approach, viewing each major event of the life course as a unique experience that may form a pattern retrospectively, but can only be anticipated within limits. This merges into a view of history that is shared by a whole generation, and an analysis of this shared perception of aging has to be disentangled from history itself. Each generation has a different perspective, and hence a different view of historical change (Mannheim, 1952). Thus, when older people suggest that policemen are getting younger, this may be a sign that they are getting older; but when they suggest that bank managers are getting younger, this may accurately reflect a historically trend. Again, the excessive number of children in Figure 1 was presented in terms of a stable pyramid of aging and of power resting on a base that reflected a high child mortality. But the excessive number of children could also have anticipated a substantial increase in population, and this has been borne out by more recent censuses.

Correspondingly, the extent to which younger people are seen to subvert tradition may be an aspect of a continuous process of adaptation to new opportunities, and 'tradition' itself may shift as a new generation matures and takes over as its custodians. Or youthful subversion may be in part a response to their subordination – an adolescent rebellion that is mounted by each successive generation. In the absence of written records, only longitudinal studies, monitoring the process of change over a period of decades, can distinguish between irreversible historical trends and the tenacity of family structures, perpetuated by the recycling of vested interests and intergenerational strains. To the extent that a popular awareness of social change draws attention to recent innovations and gives an impression of the contemporary scene as a watershed between tradition and modernization, the persistence of age relations embedded in resilient family structures is less evident in short-term studies. Underpinning this resilience is the cumulative nature of privilege associated with age in 'Status-dominated' societies. Those who react against traditional restrictions in their youth become the new custodians as they age. Changes in fashion and outlooks may tell us as much

about rivalries between successive generations as about the long-term development of society.

A more interactive approach to the experience of aging, especially in tight-knit communities, focuses on the accommodation of major life transitions as a discontinuous process: a life crisis theory of aging. The significance of rites of transition is that they involve the wider community, even beyond the family, in a shared experience of irreversible change. In the space between these events or other critical episodes, people age physically, but the configuration of social relations remains unchanged and there is a sense of timelessness as trends leading to the next critical change unfold. When this occurs or is precipitated, the configuration of roles adjusts to a new status quo and in a sense everyone ages. There is a distinct step in time. In dialectical terms, there is a mounting contradiction between inflexible social relations and the unstoppable process of maturation and decay that undermines the array of power relations. From this point of view, the anxieties that accompany these transitions and shifting roles are also anxieties of aging. Critical life events and transitions are key points, both in the experience of aging for the individual and with regard to adaptation and regeneration within the community.

Age Systems and Hidden Knowledge

A third approach to the experience of aging has coined the analogy of a 'cultural clock' that prescribes the appropriate ages for major life transitions within any society, facilitating adjustment, and heightening the awareness of those who forestall or lag the norm (Neugarten, 1968; Souvain-Dugerdil et al., 2005). While this expresses an essentially conformist approach toward aging, it is a particularly appropriate model for societies with age systems. Because members of such societies are very aware of age, they provide an ideal type for examining a range of issues associated with the process of aging. In an age system, those of a similar age are grouped together as an *age-set* (or age group), maturing and aging together. Their position at any stage may be termed an *age grade*, consisting of an array of expectations and privileges. In effect, the age-set passes up an age ladder, rung by rung, through successive age grades in a defined progression, like children passing through school but over a more extensive period in a process that persists into old age. The left-hand side of **Figure 1** illustrates a society with such a system, indicating the extent to which the transitions from boyhood at initiation and to elderhood with marriage are closely associated with age. Each step in the demographic profile broadly represents an age-set, and the successive statuses for married elders could be elaborated to provide a more detailed set of age grades. In general, age-sets tend to involve males only, but they may also define critical aspects of men's relations with or through women, giving women a distinctive role within the age system. Thus, the position of a woman may be highlighted in relation to the age-set of her husband or father or sons, but women are only rarely grouped together by age, except sometimes during the brief period leading up to their marriages.

Age systems institutionalize the cultural premise of respect for age through a form of stratification that in effect inverts stratification by caste. In contrast with the *total immobility* of a caste system, where shared status is determined by birth and persists throughout life, age systems guarantee *total mobility*: a young man is initiated onto the lowest rung of the age ladder as a member of the most junior age-set, and is systematically promoted with his age-set from one age grade to the next toward the top. In stratified age systems, the hierarchy of age-sets is associated with *respect* for age, which younger people have to learn. The nuances of an age system relate to the process of promotion, involving certain pressures from below and resistance from above. It is these pressures, arising from the interplay between a concern for status and the physical process of aging, that provide the mainspring for the 'cultural clock,' which is normally perceived as a recurring cycle of promotions and delays, spanning the interval between successive age-sets.

Elaborate age systems are primarily associated with the pastoralist peoples of East Africa. The relevance of nomadic pastoralism appears to stem from the equality of opportunity in a setting where a mixture of acumen, commitment, and sheer luck determines the success of each stock-owner. Unequal fortunes tend to even out as the more successful convert their surplus into further wives as an investment for the herding enterprise, leading to larger families and the dispersal of this wealth among the sons of the next generation. Correspondingly, the ideals underpinning age-based systems are of equality among age peers associated with the ephemeral nature of wealth and fortune rather than the inheritance of privilege based on birth and accumulated capital. This is complemented by the premise of inequality up the age ladder, regardless of family or wealth, again endowing ultimate moral authority and ritual initiative on the senior age-sets. Age systems were more widespread historically, but have become outmoded by the gathering complexity and inequalities in the process of urbanization.

The position of older men as the repositors of tradition has an affinity with secret societies in West Africa and Papua New Guinea, where esoteric knowledge is acquired by stages individually and is only fully understood in later life by those that are eventually initiated into the higher levels of the organization. It is the hiding of these secrets rather than the elusive knowledge as such that displays power, impeding the progress *of individuals* up the career ladder toward the privileges at the top. Among East African societies with age systems, there is a similar mystique and hierarchy of power associated with age, where careers are controlled by the ritual authority ascribed to those higher up the ladder. However, it is as *age-sets rather than as individuals* that promotion takes place; and to the extent that the pace of promotions is controlled by older men, manipulating the 'cultural clock' to their own advantage, they are playing for time against the inevitability of their own aging. Older women, middle-aged, and even relatively young men may conspire in this, resisting the advancement of their juniors by age. For those that live long enough, the frailty of their aging does not diminish the awe for their great age.

The antithesis of the power of older men is the physical virility of youth, often associated with an alternative lifestyle in 'warrior societies.' This poses a contradiction between the moral advantage that lies with older men and the more

immediate interests of younger men, who may react against traditional restrictions. Historically, situations of political turbulence offered opportunities that favored youth, overriding the constraints of the age system. However, it is also characteristic of the periodic age cycle that certain phases may be associated with greater gerontocratic control over a new age-set. These are paradoxically aspects of the system where privilege is vested heavily in the hands of older men, notably over women and marriage, and this creates a certain power vacuum on the lower rungs of the age ladder, encouraging youthful rebellion. By viewing age systems as interactive enterprises concerned with the distribution of power with age, rather than in terms of 'gerontocracies' or 'warrior societies' as such, their apparent resistance to change relates to involvement in the system at all ages. Those who can claim certain privileges of youth also have a stake in their future as elders.

A holistic approach toward age systems leads one to examine the relationship between principles of age organization and aspects of the family. In polygynous societies generally, competition for wives can give rise to rivalry between brothers, and the delayed marriage of sons can precipitate tension between generations. Age systems can serve to diffuse these strains by imposing restrictions on younger men that bypass the family (Samburu), by creating an alternative and prized niche for younger men (Maasai, Nyakyusa), or by maintaining a disciplined queue toward marriage (Jie, Karimojong).

A clear link between age systems and family structures is illustrated by the extent to which recruitment to an age-set is often complicated by restrictions of generation: the position of the son within the system is determined in part by that of his father or birth order, giving rise to a hybrid 'age/generation' system, rather than one based solely on age (Stewart, 1977). The rules can be highly elaborate leading observers to speculate that these rules are either spurious or have been misunderstood. However, their implications for the distribution of power and authority with age are very specific, and understanding in each instance derives from a wider analysis of relations between old and young within the family and the wider community (Baxter and Almagor, 1978). Younger men have the advantage of physical virility and the rapid accumulation of practical experience. Age systems, and indeed any institutions that endow older people with power, may be viewed in terms of their ability to impose a moral superstructure with a higher authority. The claim of older people to be the true custodians of tradition, and often to have the closest rapport with the ancestors, places society above the brutish forces of nature and inverts the natural process of physical aging in a hidden display of power.

Urban Migration and the Changing Profile of Aging

Given the substantial differences between societies studied by social anthropologists, the patterns of aging in relation to family structure do not translate easily into general statements for rural society. Even Cowgill and Holmes suggestion that the breakdown of the family in urban settings leads to the neglect of older people does not allow for the development of networks and family businesses by astute migrants who maintain and even enhance their position as they age. Again, Maine's generalization concerning the evolution of law from Status to Contract does not allow for the hiatus experienced by rural immigrants seeking work in developing economies with rapidly growing populations.

This growth has precipitated a relentless movement toward expanding urban areas, where older people's control over indigenous knowledge and institutions associated with respect for their age are no longer relevant. Younger people are at the forefront of urban migration, generating growing urban families, and in time a growing sector of urban-born adolescents. With regard to Muslim countries, Huntington (1997: pp. 104, 118) has asserted that population growth 'and particularly the expansion of the 15–24-year-old age cohort, provides recruits for fundamentalism, terrorism, insurgency, and migration. Economic growth strengthens Asian governments, demographic growth threatens Muslim governments and non-Muslim societies.'

While Huntingdon's analysis may be pertinent to the development of Muslim societies, fanned by the divisive loyalties of fundamentalism, the problem of urban population growth is more general. Data for 30 Muslim countries outside sub-Saharan Africa revealed a median annual urban growth rate of 2.4% as compared with 2.0% for 85 other developing countries outside Africa (WDI, 2005). However, these figures are dwarfed by data recorded for 43 countries in sub-Saharan Africa which revealed a median annual urban growth rate of 3.6%. This implies a doubling of urban population every 35 years in the non-African non-Muslim countries, every 30 years in Muslim areas, and every 20 years in sub-Saharan Africa. The general problem of this phenomenon for a growing younger generation clearly has a wider relevance beyond Islam, and notably in sub-Saharan Africa. Generally, there is a massive problem of displacement, from the family on the one hand and from any wider social order on the other. Somewhere between Maine's 'Status' and 'Contract,' there is a void, a lawless anomie in a dispossessed struggle for survival. Clearly this has relevance for all ages among those caught up in this situation. However, it is the younger and most active generation that are best equipped to respond to the opportunities for subversion, creativity, and developing informal networks of trust as investments for an open-ended future. Beyond law, they are the vanguard. Official projections of the balance between youth and aging and growth vary wildly, because of the unreliability of the underlying assumptions. Changing demography by age and by scale will be the major challenge for democratic institutions. As the whole world is squeezed, the profile of aging will be affected in unpredictable ways.

See also: Age Policy; Age Stratification; Age Structure; Age, Sociology of; Generation in Anthropology; Generations in History; Generations, Sociology of; Kinship in Anthropology; Life Course in History; Life Course: Sociological Aspects; Lifelong Learning and Its Support with New Technologies; Lifespan Development, Theory of; Lifespan Development: Evolutionary Perspectives; Youth Culture, Anthropology of; Youth Culture, Sociology of.

Bibliography

Amoss, P.T., Harrell, S. (Eds.), 1981. Other Ways of Growing Old. Stanford University Press, California.

Baxter, P.T.W., Almagor, U. (Eds.), 1978. Age, Generation and Time. Hurst, London.

Cowgill, D.O., Holmes, L.D. (Eds.), 1972. Aging and Modernization. Appleton-Century-Crofts, New York.

Foner, N., 1984. Ages in Conflict: A Cross-Cultural Perspective on Inequality between Old and Young. Columbia University Press, New York.

Hall, G.S., 1904. Adolescence: Its Psychology and Its Relations to Physiology, Anthropology, Sociology, Sex, Crime, Religion and Education. Appleton, New York.

Huntington, S.P., 1997. The Clash of Civilizations and the Remaking of World Order. Touchstone Books, London.

Kertzer, D.I., Keith, J. (Eds.), 1984. Age and Anthropological Theory. Cornell University Press, London.

La Fontaine, J.S., 1985. Initiation. Academic Press, London.

Maine, H.S., 1861. Ancient Law. Murray, London.

Mannheim, K., 1952. 'The Problem of Generations', in Mannheim, Essays on the Sociology of Knowledge. Routledge and Kegan Paul, London.

Neugarten, B.I., 1968. Adult personality: towards a psychology of the life cycle. In: Neugarten, B.I. (Ed.), Middle Age and Aging. University of Chicago Press, Chicago.

Pereira, M.E., Fairbanks, L.A. (Eds.), 1993. Juvenile Primates: Life History, Development and Behaviour. Oxford University Press, New York.

Simmons, L.W., 1945. The Role of the Aged in Primitive Society. Yale University Press, New Haven.

Souvain-Dugerdil, C., Leridon, H., Mascie-Taylor, N. (Eds.), 2005. Human Clocks: The Bio-cultural Meanings of Age. Peter Lang, Bern.

Spencer, P., 1965. The Samburu: A Study of Gerontocracy. Kegan Paul, London.

Spencer, P. (Ed.), 1990. Anthropology and the Riddle of the Sphinx: Theories of Age and Its Transformation. Routledge, London.

Stewart, F.H., 1977. Fundamentals of Age-Group Systems. Academic Press, New York.

WDI, 2005. 'World Development Indicators Database: Urban Population Growth'.

Age Policy

Anne-Marie Guillemard, University Paris Descartes Sorbonne, Paris, France

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A.-M. Guillemard, volume 1, pp. 267–271, © 2001, Elsevier Ltd.

Abstract

With the formation of the modern state and then the welfare state, a 'government by age' of the population has become widespread with, as a consequence, a multiplication of age policies. A few concrete examples of the application of age policies to childhood and old age show how they have institutionalized the life course and assigned functions and identities to the successive stages of life. This trend has, in turn, led to a generalization of age policies with comprehensive, coordinated programs for the old and young. The relevance and effectiveness of age policies are now under question. The advent of a postindustrial society corresponds to a deinstitutionalization and destandardization of the tripartite organization of the life course inherited from the industrial era. Age policies should no longer be the main tool of population management. They should be replaced with life-course policies that are better suited for achieving an 'age-integrated society.'

The formation of modern states has come along with a social and legal construction of the individual around, to a large degree, the universal, neutral criterion of age. What we call 'age policy' has arisen as governments have diversified and expanded their interventions for managing the population. A genuine 'government by age' has been set up; clear evidence of this being the operation of welfare states and their bureaucracies is described below. The implementation of age policies for childhood and old age serves as a concrete example of how these policies have developed into comprehensive, coordinated public policies in behalf of these age groups. Current studies insist, however, that using age as a major criterion in policy-making has perverse effects; it leads to the erection of age barriers and abets age-based discrimination. As a consequence, the relevance of age as a major instrument for state interventions has come under question.

Age Policy, a Product and Major Tool of Public Policies

The Expansion of 'Government by Age' with the Emergence of the Modern State

Along with the modern state has emerged a social and legal construction of the individual. At stake in this process is the creation of conditions that single out the individual for attaining autonomy from family and community bonds. The individual is thus considered "as the prime holder of rights and duties and as the prime target of bureaucratic and administrative acts" (Mayer and Schoepflin, 1989: p. 193). This construction is grounded on a set of universal age-based norms for marking the chronological continuum of life with significant thresholds and organizing it into successive stages.

The legal ages of majority, for voting or joining the army, for example, laid down the threshold for entry into adulthood and full citizenship. In like manner, as social legislation gradually regulated and then forbid child labor or set the age of compulsory schooling, it also defined the bounds of childhood while assigning it specific functions and roles. At the other end of the life course, pension systems and old age insurance funds have, by setting the age of eligibility for a retirement pension, signaled the end of adulthood and worklife, and laid the threshold for entry into old age, when people are to live on pensions without working – even though, given the longer life span, the association between economic inactivity and old age is coming unraveled.

Through these age policies, the state was able to perform its major duties: redistribute income between age groups or generations; maintain social order by assigning roles, statuses, activities, and identities to each age group and individual; and manage human resources throughout the life span. Owing to this government of the population 'by age,' the state distributed social activities, obligations, rights, and resources between the three stages of life: childhood, the worklife (adulthood), and old age. A literal 'age police' was instituted to borrow Percheron's phrase (1991), in which 'police' has the older meaning of government as an act of 'ruling.' This essential aspect of government actions has expanded along with the welfare state. As Smelser and Halpern (1978) have shown, the state oversees the organization of the school–family–work 'triangle.' The life course thus came to be organized in line with an orderly, hierarchical model of three successive, clearly delimited stages. Accordingly, the individual is either a child in school and dependent on the family (a stage of life related to policies targeting children, families, and education); or an adult who is employed and thus benefits, owing to his/her work, from coverage for various social risks; or a retiree eligible for a pension.

Each stage of life was associated with an activity endowing it with a meaning and identity distinct from the other two stages. Under this model, the individual's life revolved around work, which was associated with adulthood. Young people were educated to prepare them for work; and old age was the time for rest and retirement after a life of work. As welfare states arose and set age thresholds, this model was gradually institutionalized; it defined the rights and obligations specific to each stage of life.

Age Policy, the Main Tool for Shaping the Life Course and the Individual's Biography

Universal social entitlements under the welfare state have led to an individualization and periodization of biographies. These entitlements condition clearly distinct situations and have

widened the rift between the worklife and the periods before and after it:

> In the welfare state, the continuous flow of life is transformed into a series of situations all of which have a clear formal definition […] Periodization of life and proliferation of sharp transitions which derive from the social insurance system combine into a lifelong biographical pattern.
>
> Mayer and Schoepflin, 1989: p. 198

Owing to its rules based on the chronological age for receiving benefits, the welfare state standardized life events and gradually institutionalized the life course. All components of the life course were defined: the number of successive stages, their social contents, the timing for passing from one stage to the next, the nature (more or less ritualized) of these transitions, the milestones marking these thresholds, and, not to be forgotten, the forms of solidarity and competition between age groups or generations. This life-course institution conditions the individual's biographical itinerary, sets the timing, and determines his/her plans as a function of 'temporal horizons' (*see* Motivation: Life Course and Sociological Perspectives).

This tripartite life-course model underlaid the development of industrial society and its timing of events. As Castel (1995) has shown, 'industrial wage-earning societies' relied on a strong connection between the dependent economic status of wage earners and an extensive system (the welfare state) of protection against risks. We might add a third dimension to this analysis: this tripartite social organization of the life course placed work at the center of life and placed the individual in a long, foreseeable chronology. This temporality contrasts with what prevailed during the preindustrial era, when only the family (and not the individual) had a 'future,' and the passage from one age of life to the next was blurred and reversible (Hareven, 1986).

Implementing Age Policies: The Cases of Childhood and Old Age

From Policing Ages to Comprehensive, Coordinated Age Policies

Childhood, like old age, was not identified as such till the modern era. During the Middle Ages, it did not exist as an autonomous age of life. It emerged during the eighteenth century, when children were 'placed apart and reasoned' inside institutions that had the duty to educate them (Aries, 1962) (*see* Childhood and Adolescence: Developmental Assets). Likewise, the invention of old age owes much to retirement systems, which, by setting the eligibility age for a pension, have established a threshold and assigned old age the socially uniform meaning of 'pensioned inactivity' (Guillemard, 2000) (*see* Retirement and Encore Adulthood: The New Later Life Course). A wide range of age-based social legislation has been passed regarding these two age groups.

The first laws on childhood regulated child labor; but laws about schooling and the age of compulsory schooling would be the principal vectors of the institutionalization of childhood as a stage of life assigned to educating future citizens and workers. The development of welfare states with their laws on maternity leaves, child care, health care, and family allowances put the finishing touch on this social construction. Comprehensive, coordinated policies targeting childhood gradually took shape in developed countries. This widespread adoption of age-based programs immediately led to the specialization of the professions in charge of managing childhood. As childhood came to be identified as a specific stage of life, infant care, and pediatrics became specialties in the health professions (Heyns, 1988). As much can be said about the age-based segmentation of professions in social work such as professionals in nursery.

The invention of old age as a stage of life regulated by a specific policy occurred at the junction of two age policies: retirement and welfare measures that targeting the aged.

Old age gradually emerged with the generalization of retirement systems, which regulated the transition from the worklife to inactivity on a pension. These systems introduced chronological age as a formal eligibility requirement for putting an end to the worklife. They assigned an identity to this new stage of life whereas, till then, the patriarch endowed with the family's estate had nothing in common with the old people without any estate, who swelled the ranks of the destitute, the poor, the unfit for work, the disabled or handicapped (children as well as adults) who were to be taken in custody by poorhouses or to receive relief while living at home.

The invention and generalization of retirement pensions and old age insurance – definite examples of age policy – have played a decisive part in constructing and consolidating the 'tripartition' of the life course. Kohli (1987), pointing this out, has identified four processes in this social organization of the life course.

First of all, retirement systems have been a major factor in determining the order and ranking of the three stages of the life course – with work at the center as the meaningful activity during adulthood. This stage lies in between youth (devoted to education prior to work) and old age (associated with inactivity and a pension after work). These systems have staked out a life course where the individual's contribution during adulthood in the world of work conditions is the right to rest at the end of life.

Second, retirement systems, along with other social policies (such as education), have given weight to chronological criteria for marking the transition from one stage to the next (*see* Age Stratification). Age policies have thus helped 'chronologize' the life course, marked as it is by the legal ages for starting and leaving school (the latter separating childhood from adolescence) and for going on retirement with a full pension.

Third, this division of the life span into three stages marked by chronological milestones has standardized the life course. At the same age, nearly everyone predictably moves from one stage to the next. At an equivalent level of education, entering the world of work occurs at the same age for nearly everyone. And the retirement age sets the date when most people stop working. Long-term trends over the past 50 years in the age of withdrawal from the labor force provide evidence of this standardization. As pension systems expanded to cover more and more of the population, the average age for going on retirement gradually approached the age of entitlement to a full pension for wage earners (Kohli et al., 1991).

Fourth, pension systems have fostered the making of plans for the future. The individual no longer has the same prospects as in preindustrial society, when the family and private wealth determined the timing of the stages of life and individuals had

no future as such, since they died young. Only the family had a future. The development of old age insurance, along with the longer life expectancy, has led to an individualization and 'temporalization' of the life course. Pensions endow the individual with a future. As a consequence, retirement systems have furthered the change from a society of ascribed statuses (through membership in a family or local group) toward a society of achieved statuses and, thus, of mobility. In this new industrial society, the individual has personal plans. Thanks to old age insurance, security is now based on the person's worklife and/or citizenship rights – and no longer on his property, local bonds, or family status. To ensure this security, old age insurance relates successive generations through forms of reciprocity and statistical, long-term solidarity. It has thus contributed to erecting a new social order in line with the requirements of a society undergoing industrialization, a society of social mobility that exercises control over the future. This illustrates how the state's regulatory interventions instituted age-based norms for 'policing ages.'

As Graebner (1980) has shown in his history of retirement in the United States, the invention of the pension system was more than just a way to insure older workers who were no longer able to work. Companies used this system to control the flow of labor and develop human resource management in line with the principles of Taylorist Scientific Management. They thus rationalized the withdrawal of older workers from the labor force while replacing them with young people who were thought to be more efficient.

Once old age as a definite, foreseeable stage in the life course was instituted, welfare policies were gradually worked out for providing specific coverage to this age group. A range of welfare policies for the elderly provide services and facilities related to care and leisure.

In France, coherent public programs targeting youth and the elderly were drafted as of the 1960s. Their clear intent was to improve the management of these age groups and implement comprehensive and coordinated public action plans with integrated multilevel programs for responding to the specific problems of these groups. French public authorities published the first major report on old age titled – 'Old age policy' – in 1962. It contained a public action plan for integrating the elderly in society. In pursuit of this goal, home care was to be provided instead of institutional custody, which the report declared contrary to its objectives. As an analysis of French old age policy has shown (Guillemard, 2000), this report marked a turning point in diagnosing the problems related to old age. Fighting against poverty through interventions targeting income was now but one way, among others, of achieving social 'inclusion.' Since the main old age problem was said to be the lack of social integration, it was important to act on the environment and lifestyle of the aging population so as to help it – by building adapted housing, expanding home-care services or the neighborhood facilities necessary for helping them remain living at home – continue living an independent life as long as possible. These new policies focused on issues related to the way of life more than on improving living standards. A similar trend can be observed in the United States with the passage of the Older American Act in 1965 (Estes, 1979).

These old age policies emerged along with a segmentation by age of the professions responsible for managing this group, similar to what we have seen in the case of age policy for children. Geriatrics and gerontology became specialties in the health professions (Haber, 1986) and social work. Professional groups were formed around these new specialties, for instance the Gerontological Society in the United States and the Society of Gerontology and Geriatrics in France. They form what Estes (1979) has called 'the aging enterprise.' Later policies were drafted for caring for the frail elderly and the vulnerability of that phase of life, advanced old age (Olson, 1994).

The Perverse Effects of Age Policies

By implementing these age-based policies, public authorities invented and now regulate infancy, childhood, adolescence, old age, and advanced old age. They have assigned social contents and identities to these stages of life. These policies have not, however, always had the effects intended (*see* Population Aging: Economic and Social Consequences).

Research on the old age policies implemented in the United States or Europe since the late 1960s has shown that the measures for providing social services and facilities to help the aged continue living at home and avoid institutional custody (with 'marginalization' and lessened autonomy as consequences) have made these persons dependent (Guillemard, 1983; Walker, 1980). The central argument in studies on these perverse effects is that – despite both the good intentions underlying government interventions and the tangible results for beneficiaries – these programs have, in general, not maintained or developed the targeted group's autonomy. Arrangements for home services have, perversely, turned any physical, mental, or social disability into a form of 'dependence.' Thus emerged a new definition of the senior citizen as the "recipient of services whose extent and nature are decided by others" (Townsend, 1981: p. 19). Moreover, fragmentation in the provision of home-care has led to defining beneficiaries by a long list of needs for health care, social ties, home helpers, cleaning services, etc. A category of professionals has been assigned to satisfy each of these needs; and the beneficiary is the hub of a merry-go-round of services, as helpers, care providers, and professionals successively arrive at his/her bedside, according to their own schedules while the beneficiary lacks any control over the services thus dispensed.

The Rise and Fall of Age Policy

For more than a century now, the state has adopted an age-based management of the population. This form of management seems to have reached its limits.

In her pioneering book, Neugarten (1982) has questioned both the relevance of age as the main lever for handling the social problems of this age group and the efficiency of public policies targeting it. Her call for an 'age-neutral society' suggests organizing government interventions on the basis of needs instead of age. This questioning of the relevance of age is apparently related to current trends in developed societies. The advent of a postindustrial, knowledge-based, or network society (the terms varying according to the theorist) has upset industrial society's tripartite organization of the life course. The life course is becoming flexible (Best, 1981). Biographical itineraries no longer fit into the pattern of three distinct, successive, well-ordered stages. Continuity in careers is coming undone, as

periods of training and inactivity are interspersed throughout the worklife. The new imperative of lifelong learning undermines the idea of youth as a period specialized in education. The tripartite life course, its stages blurred, is being deinstitutionalized and destandardized (Guillemard, 2005).

The spread during the 1980s and 1990s of a wide variety of schemes for withdrawing early from the labor force is evidence of this deinstitutionalization (Guillemard and Van Gunsteren, 1991), since they subvert the regulated transition from work to retirement and wreak havoc in the orderly succession of the three stages of life. Although early exit, or preretirement, seems to signal an event that simply occurs sooner on the retirement calendar and nothing else of significance, these age-based measures have had so great an impact that we can interpret them as increasing flexibility at the end of a life course that is undergoing reorganization (Guillemard, 1997). The most frequently used early exit schemes were developed not under pension systems or old age insurance but instead under disability and unemployment insurance. In all countries, these extremely malleable schemes have continually evolved as a function of employment conditions. We can interpret this as a 'dechronologization' of the life course: the individual can no longer imagine a continuous, foreseeable life marked by chronological age thresholds.

The order of the stages of life and the activities assigned to them are no longer precise, and are even contingent. The timing of definitive exit from the labor market is unforeseeable. No one working in the private sector knows when (at what age) or how (under what conditions) they will definitively stop working. The chronological thresholds at the end of the worklife are no longer clearly visible. We now observe a leveling of age thresholds along with the reversibility of transitions and itineraries. These trends invalidate the relevance and effectiveness of an age-based management of the population, since, given the destandardization and individualization of biographical itineraries, individuals of the same chronological age might be living in very different situations. For example, a wage earner 50 years old might be experiencing late parenthood and still have children in school whereas another of the same age might be facing an 'empty nest,' the children having grown up and left home. The retirement strategies developed by these two persons will differ significantly. A single, standard retirement age cannot meet such different needs. Instead of an age-based management of the population, the destandardization of the life course calls for managing biographical itineraries. Life-course policies capable of providing 'flexisecurity' to individuals with destandardized itineraries must be designed to replace age policies (Guillemard, 2010). Riley et al. (1994) have drawn attention to the 'structural lag' between the expectations of the aging, owing to their long life expectancy, good health and extended capabilities, and the limited 'opportunity structures' offered by public age policies. She has drawn the conclusion that a new formula for combining work, family, and leisure must be worked out to create an 'age-integrated society' where work, family, and leisure activities are woven through all stages of life. This echoes the European Commission's call for a 'society for all ages' (EC, 1999), repeated in the "2012 European year for active ageing and solidarity between Generations" evinces the new interest in fighting against age barriers and age-based discrimination in employment (European Foundation, 1997) and in developing 'age-diversity management.'

See also: Age Stratification; Age Structure; Age, Sociology of; Aging, Theories of; Life Course in History; Motivation: Life Course and Sociological Perspectives; Population Aging: Economic and Social Consequences; Reformation and Confessionalization; Retirement and Encore Adulthood: The New Later Life Course; Retirement and Health; Retirement, Economics of; Welfare Economics; Welfare Policy Comparisons; Welfare.

Bibliography

Aries, P., 1962. Centuries of Childhood. A Social History of Family Life. Random House, New York.

Best, F., 1981. Flexible Life Scheduling. Breaking the Education Work, Retirement Lockstep. Praerger, New York.

Castel, R., 1995. Les métamorphoses de la question sociale. Une chronique du salariat. Fayard, Paris.

Estes, C.L., 1979. The Aging Enterprise. Jossey-Bass, San Francisco, CA.

European Commission, 1999. Toward a Society for All Ages. Promoting Prosperity and Solidarity between Generations, Brussels, Com/99/121.

European Foundation, 1997. Combatting age barriers in employment. European research summary. Luxemburg: Office for Official Publications of the EC.

Graebner, W., 1980. History of Retirement: The Meaning and Function of an American Institution (1885–1978). Yale University Press, New Haven, CT.

Guillemard, A.M. (Ed.), 1983. Old Age and the Welfare State. Sage, London.

Guillemard, A.M., Van Gunsteren, H., 1991. Pathways and their prospects: a comparative interpretation of the meaning of early exit. In: Kohli, M., Rein, M., Guillemard, A.M. (Eds.), Time for Retirement: Comparative Studies of Early Exit from the Labor Force. Cambridge University Press, Cambridge, UK, pp. 362–388.

Guillemard, A.M., 1997. Rewriting social policy and changes within the life course organization: a European perspective. Canadian Journal on Aging 16 (3), 441–464.

Guillemard, A.M., 2000. Aging and the Welfare State Crisis. University of Delaware Press, Newark, NJ.

Guillemard, A.M., 2005. The advent of a flexible life course and the reconfigurations of welfare. In: Andersen, J.G., Guillemard, A.M., Jensen, P., Pfau-Effinger, B. (Eds.), The Changing Face of Welfare. Policy Press, Bristol, pp. 55–74.

Guillemard, A.M., 2010. Les défis du vieillissement. Age, Emploi, Retraite Perspectives internationales. Armand Colin, Paris.

Haber, C., 1986. Geriatric: a specialty in search of specialists. In: Van Tassel, D., Stearns, P.N. (Eds.), Old Age in Bureaucratic Society. Greenwood, Westport, CT, pp. 66–84.

Hareven, T., 1986. Historical change in the social construction of the life course. Human Development 29 (3), 171–180.

Heyns, B., 1988. The Mandarins of Childhood: Toward a Theory of the Organization and Delivery of Children's Services. Basic Books, New York.

Kohli, M., 1987. Retirement and the moral economy: an historical interpretation of the German case. Journal of Aging Studies 1 (2), 125–144.

Kohli, M., Rein, M., Guillemard, A.M. (Eds.), 1991. Time for Retirement. Cambridge University Press, Cambridge.

Mayer, K.U., Schoepflin, U., 1989. The state and the life course. Annual Review of Sociology 15, 187–209.

Neugarten, B.L. (Ed.), 1982. Age or Need? Public Policies for Older People. Sage, Beverly Hills, CA.

Olson, L.K. (Ed.), 1994. The Graying of the World. Who Will Care for the Frail Elderly? Haworth, Binghamton, NY.

Percheron, A., 1991. Police et gestion des âges. In: Percheron, A., Remond, R. (Eds.), Age et Politique. Economica, Paris, pp. 112–139.

Riley, M.W., Kahn, R.L., Foner, A. (Eds.), 1994. Age and Structural Lag. Wiley Interscience, New York.

Smelser, N., Halpern, S., 1978. The historical triangulation of family, economy and education. American Journal of Sociology 84, 288–315.

Townsend, P., 1981. The structured dependency of the elderly. Ageing and Society 1 (1), 5–28.

Walker, A., 1980. The social creation of poverty and dependency in old age. Journal of Social Policy 9 (1), 49–75.

Age, Sociology of

Linda K George, Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Foner, M.W. Riley, volume 1, pp. 275–278, © 2001, Elsevier Ltd.

Abstract

Age is a personal characteristic, a base of stratification, and a characteristic of populations. This article covers the sociology of aging at multiple levels of social organization. Topics reviewed include the age structures of populations, the challenges posed by aging populations, the distinction between aging processes and cohort differences, the effects of social change on cohort differentiation, and the relationships between age and individual well-being.

The sociology of aging is both broad and deep. The breadth of the field can be highlighted in several ways. First, the sociology of aging encompasses investigations of aging as a process of older adults as a group and of old age as a distinctive stage of the life course. Second, aging research is performed at multiple levels of analysis – from macro-level studies of age structure within and across societies, to meso-level studies of labor force participation and family structure, to micro-level investigations of health and well-being. Third, aging research uses the full repertoire of methods that characterize the discipline, including life tables and other demographic methods, survey research, ethnographic methods, and observational studies. The depth of the field results from the accumulation of scientific studies that now span more than three-quarters of a century.

Any attempt to summarize the state-of-the-science in the sociology of aging concisely will inevitably do justice to neither the breadth nor the depth of the field. Here, four major themes in theory and research on aging are reviewed. Selection of these themes is based on a review of appropriate reference works (e.g., handbooks and encyclopedias) and perusal of major journals and textbooks. Nonetheless, it is likely that other scholars would choose other themes, and those considered here are not exhaustive.

Age Structure and Its Implications

Although primarily the province of demographers, both scientific and public interest in aging partially rests on the aging populations characteristic of the modern world. Age structure is based on the sizes of age groups in a given society. In turn, the sizes of age groups are a function of fertility and mortality rates. The *theory of the demographic transition* provides a portrait of the relationships between development (i.e., industrialization, urbanization, and technological advances) and the age structures of societies (Bourgeois-Pichat, 1979). According to this theory, the demographic transition occurs in three stages. In the first, prior to and during the early years of modernization, both fertility and mortality rates are high. The result is an age structure that takes the form of a pyramid, with the largest age group consisting of children and older adults comprising the smallest group (*see* Age Structure). During the process of modernization (the second stage), mortality rates decline, but fertility rates remain high. The result is larger populations in total size and a young population. The dependency ratio, which is the proportion of the population in the labor market relative to the number of children and older adults not in the labor market, is high, but consists primarily of children. As societies achieve modernization (third stage), mortality rates continue to decline, but fertility declines as well. The theory of the demographic transition hypothesized that the third stage would lead to a steady-state population size, in which fertility and mortality rates would be approximately equal, and the age structure would take the form of a cylinder, with all age groups of approximately equal size, and the numerator of the dependency ratio including approximately equal numbers of children and older adults.

The first two stages of the theory of the demographic transition have been supported in empirical studies. Evidence for the hypothesized third stage is much weaker. Although both fertility and mortality rates have declined in modernized societies, fertility rates have declined faster than mortality rates, resulting in populations with disproportionately large numbers and percents of older adults. Two other demographic patterns have exacerbated the aging of the population in some societies. First, in the United States and, to a lesser extent, in Western Europe and Australia, the end of World War II ushered in a decade or more of unusually high fertility, resulting in a 'baby boom,' followed by the expected sizable declines in fertility (*see* Fertility Theory). These unusually large cohorts delayed the predicted declines in fertility, but because of the small cohorts that followed them, escalated the aging of the population. Second, countries in South America, Asia, and Africa have achieved substantial modernization with substantially smaller declines in fertility than expected. An exception to this pattern is China, where the 'one child' policy decreased fertility very rapidly.

The importance of the theory of the demographic transition is not in its accuracy, but rather in its attention to the age structures of societies and the effects of social change (in the form of fertility and mortality rates) on those age structures. Regardless of the indicator used (e.g., proportion of older adults and median age of the population), populations are aging rapidly in the United States, other western societies, and China.

Although initially slow to comprehend the significance of an aging population, both scientific and policy communities are now well aware of the challenges posed by an aging society.

'Young' populations also pose problems for societies, but solutions to those problems (e.g., child welfare and schooling) evolved gradually in modern societies and became institutionally embedded in custom and law. The aging of modern populations occurred more rapidly, resulting in significant structural lag (Riley and Riley, 1994) in institutional responses to the needs of older adults (*see* Population Aging: Economic and Social Consequences).

Demographic research has expanded substantially beyond questions of age structure per se. Most of what is now known about the prevalence, incidence, and course of disability is based on demographic research. Among the contributions of this research are findings that rates of disability among older adults have been gradually declined for approximately 20 years (i.e., in the 1980s and 1990s) and have now leveled off, that the higher rates of disability among women than men result from their greater longevity rather than higher incidence of disability onset, and that, although it is less common than the onset of disability, a sizable proportion of older adults recover or improve in functional status. The distinction between life expectancy (expected years of survival) and active life expectancy (expected years of disability-free survival) is also a high-priority issue. Interest in active life expectancy stems from the concern that the increased longevity characteristic of modern societies has been achieved by prolonging life after the onset of frailty and disability. Evidence suggests that the length of disability prior to death has not increased over the past several decades; however, there also has not been a decrease in the interval between the onset of disability and death. As these topics illustrate, demographers now devote a significant proportion of their efforts to understanding heterogeneity in the older population (*see* Retirement and Encore Adulthood: The New Later Life Course).

A large proportion of sociological research on aging rests on the challenges posed by an aging society, although that impetus is not always explicit. Studies of public and private transfers of money, time, and in-kind services rest in large part on their salience for sustaining an aging population. Studies of health, disability, and quality of life are important not only because they address threats to well-being, but also because they identify factors that keep older adults from excessive reliance on public programs. Even studies of the caregivers of impaired older adults rest not only on concern about the health risks of chronic stress, but also on the desire to enable families to bear as much of the cost of care as possible, thus relieving public programs. Thus, age structure and its social implications is a significant and far-reaching arm of aging research.

Aging as Context: The Significance of Cohorts

Multiple forces, both social and nonsocial, determine the process and experience of aging (*see* Aging, Theories of). Historically, there was a tendency to attribute the aging process and the experience of late life to inherent biological and developmental processes. Most of us are relatively ignorant of the extent to which the process and experience of aging vary across historical time, finding it difficult, for example, to imagine a time when there was no retirement or when the odds of dying were essentially the same during childhood, adulthood, and old age. And yet, retirement as a predictable life--course transition and odds favoring survival to old age both emerged in the twentieth century. The concept of cohort allows us to distinguish conceptually and empirically between inherent components of the aging process and patterns that result from social factors, especially social change and unique historical events or circumstances (*see* Period and Cohort Analysis in Demography).

Although cohort membership can be based on any event, the term is typically used for birth cohorts – for persons born at the same or approximately the same time. Cohorts share more than the timing of their births; they also experience the same historical events and social structures throughout their lives. Cohorts share collective experiences that often differ from those shared by earlier and later cohorts. Thus, there often are sizable cohort differences in the process and experience of old age.

Cohort differences, often observed across cohorts born in relative proximity, can be generated by multiple conditions. First, cohorts can differ substantially in size and composition. Substantial evidence documents that unusually large and unusually small cohorts differ substantially, especially in economic opportunities, with the latter more plentiful in smaller cohorts. Cohort composition can be affected by many factors including excess male mortality during wars, different birth rates across racial/ethnic groups, and changes in immigration policies.

Second, historical events can substantially alter the experiences of cohorts (*see* Life Course: Sociological Aspects). When many cohorts experience the same historical event, effects differ depending on age at the time of the event (e.g., wars most strongly affect young men). Even in the absence of dramatic events or dislocations, historical developments imprint cohorts differently, creating persisting differences (e.g., racial identity among African-Americans before and after the Civil Rights Movement). Substantial evidence suggests that historical events and social change generally affect adolescent and young adult cohorts more than they affect younger and older cohorts.

Third, social change creates cohort differences and the more rapidly the social change occurs, the greater the differentiation across cohorts. Social change, of course, takes many forms, ranging from changes in public policies (e.g., Social Security and Medicare created large cohort differences in economic status during later life), changes in the norms governing social behavior (e.g., norms concerning the acceptability of fertility outside of marriage), and major structural changes such as the shift from an industrial to a service economy. Technological change also generates cohort differences, with varying implications for the older population. For example, devices that save household labor or provide assistance in compensating for disabilities enhance the likelihood that older adults can live independently. Conversely, diffusion of technological changes often takes longer to reach older adults, distancing them from younger cohorts (e.g., personal computers and related use of the Internet).

Finally, older cohorts are inevitably affected by the composition of and changes occurring in younger cohorts. Responses to those changes can create related changes in older

cohorts. For example, the prevalence of custodial grandparents, although still small, increased dramatically over the past two or three decades. This change in older cohorts is a direct result of changes in fertility and childcare practices of younger cohorts.

Cohort comparisons comprise a substantial proportion of sociological research on aging. An issue that receives considerable attention is comparisons of the assets and liabilities that different cohorts bring to late life. To date, research findings paint a rosy picture of this form of cohort change. For more than half a century, successive cohorts have entered old age with higher levels of resources and fewer liabilities than the cohorts that preceded them. This pattern has been especially consistent for health, education, wealth, and the availability of social support, all of which are valuable assets in late life. Recent evidence suggests that this pattern will not continue indefinitely. With regard to health, for example, although early baby boomers are expected to live longer, on average, than their parents, they report higher prevalence of chronic illness and disability than their parents' generation did at the same age (Marin et al., 2009).

The large volume of cohort comparison studies is too large to detail here, but includes issues as diverse as political affiliation and voting behavior, family structure (primarily cohort differences in divorce, remarriage, and single-parent mothers), cognitive abilities, alcohol consumption, and rates of depression. Cohort comparisons also are important for policy planning and analysis. Comparison of cohorts before and after a major policy change, such as the enactment of Medicare, is one of the primary strategies used to evaluate the impact of broad-scale public policies. Cohort comparison studies also remind us that public policies targeted at older adults are not the only policies that create differential advantage or disadvantage across cohorts. Policies that enhance educational attainment during adolescence and young adulthood have long-term benefits, creating cohorts that are more advantaged as they enter old age than were older cohorts. The GI Bill, first made available to World War II veterans, for example, created dramatic increases in the educational attainment and economic resources with which those cohorts entered late life.

At the same time that cohort comparison studies have enjoyed success, a related body of research examines *intracohort variability*. Although there are often large differences across cohorts, cohorts are not homogeneous. Paralleling sociology more broadly, increased attention to heterogeneity has characterized aging research for the past five decades. Intracohort variability has received both conceptual and empirical attention. Two bodies of research have contributed most to this research base.

First, the attention paid to social location, as indexed by basic ascribed and achieved statuses, has increased dramatically. For decades, gender differences, racial/ethnic differences, marital status differences, and socioeconomic (SES) differences typically were examined perfunctorily, if at all. It was not until the mid- to late-1970s, for example, that women's retirement received empirical attention. There is now general consensus that age, race/ethnicity, gender, and SES represent basic social structural categories and are forms of social stratification. Investigation of these multiple forms of stratification has been incorporated in aging research and in the discipline more broadly.

Second, a compelling body of research demonstrates that historical events or conditions do not have uniform effects on cohort members. Some cohort subgroups are strongly affected by historical events; others are largely untouched by them. Several important studies demonstrated that the Great Depression had the strongest contemporaneous and long-term effects on late adolescents whose families experienced the greatest economic deprivation (Elder, 1999); that relatively few young adults participated in the political activism of the 1960s, but that there were persisting differences in the life patterns of those who did and did not (McAdam, 1990); and that veterans' emotional problems in middle and late life were largely a function of amount of combat exposure during World War II (MacLean and Elder, 2007). Broad-based cohort effects also have been observed for these and other historical events or conditions (e.g., the "children of the Great Depression" had greater concerns about financial security than earlier and later cohorts, regardless of the amount of deprivation experienced), but there is also great heterogeneity in the effects of historical events on cohort members (*see* Motivation: Life Course and Sociological Perspectives).

Intercohort comparisons and studies of intracohort variability are arguably the core of sociological aging research. These studies demonstrate that aging is not solely – or even primarily – a biological process, but rather that the aging process and the experience of late life are shaped by social and historical context.

Aging and Well-Being

The vast majority of aging research falls under the general topic of aging and well-being, with well-being broadly defined to include any social asset (e.g., economic resources and life satisfaction) (*see* Aging, Theories of). Social scientific interest in aging was spurred by concerns about the well-being of older adults in both absolute and relative (to other age groups) terms. This is probably not surprising. The history of sociology in general has been driven by concerns about social disadvantage – its prevalence, antecedents, and consequences.

The types of well-being examined in relation to aging are numerous. A partial list of the forms of well-being frequently studied in late life include longevity, physical health, disability, mental health, subjective well-being, economic status, and identity or sense of self.

Self-perceptions during late life is an important and understudied topic relative to studies of physical and mental health and subjective well-being. Two primary dimensions of self-perception are especially important: a sense of self-worth (typically measured as self-esteem or self-acceptance) and a sense of competence (usually measured as self-efficacy, mastery, or sense of control). These self-perceptions are important in their own right – most of us consider adequate self-esteem and sufficient self-efficacy essential components of well-being. In addition, these self-perceptions mediate many of the relationships between social factors and other forms of well-being, including physical and mental health and subjective well-being. For example, self-esteem has been shown to mediate the effects of education on health (Murrell et al., 2003) and of social stress on functional status (Forthofer and Janz,

2001). Self-efficacy mediates the relationship between physical activity and perceived quality of life (Phillips et al., 2013). Traditionally, sociologists tended to view self-perceptions as the province of psychology. There is now plentiful evidence, however, that the antecedents of these self-perceptions are primarily social – and that their distributions are concordant with multiple stratification systems. Adequate sociological explanations of variability in well-being will need to take these psychosocial processes into account.

Three major research strategies underlie most research on aging and well-being. First, some studies examine the well-being of older adults relative to that of younger age groups (e.g., Schieman et al., 2002; Siedlecki et al., 2014). Examples of this kind of research include age comparisons of rates of poverty, chronic physical illness, disability, mental illness, and levels of subjective well-being, self-esteem, and self-efficacy across age groups at a single point in time. These studies are of limited use in understanding the aging process because it is unclear whether differences across age groups are due to age per se or to cohort differences. But comparisons across age groups can be useful for both providing basic descriptive information about the relative status of older adults and for identifying issues important to public policy (e.g., whether it makes sense to target income-maintenance programs at specific age groups). That is, policy makers typically are more concerned about the unmet needs of older adults than they are about disentangling age and cohort effects.

A second strategy for understanding the effects of aging on well-being is to study adults longitudinally as they move from middle age to late life and from being young-old to being old–old (e.g., Lynch and George, 2002; Yang, 2007). The advantage of this strategy, of course, is that age-related changes in well-being are directly observed. These studies focus on the status of the elderly relative to earlier points in their lives, rather than relative to younger age groups. The limitation of these studies is that findings may be cohort-specific, rather than reflecting a consistent developmental pattern. In theory, sampling a large number of cohorts and studying them over long periods of time, investigators can determine whether patterns of change and stability are similar across cohorts. Unfortunately, few data sets are of sufficient breadth in both the number of cohorts studied and the number of measurements over time to permit conclusions about whether patterns of change and stability are generalizable or cohort-specific.

Patterns of change and stability in the multiple forms of well-being that have been studied to date cannot be detailed here. Importantly, however, there is no consistent pattern of age-related decline across all forms of well-being. Declines are the modal pattern for some forms of well-being, such as income and the prevalence and onset of chronic illnesses. Stability is the modal pattern for other types of well-being, including self-esteem and life satisfaction – at least until very late life (i.e., age 80 years and older).

A third strategy is to focus on variability within the older population; that is, to assess variability in well-being among older adults and to identify the antecedents of that variability. Either cross-sectional or longitudinal data can be used to study heterogeneity among older adults, but only longitudinal data permit investigators to establish temporal order between well-being and its presumed antecedents. Note that age is not the independent variable in studies of this type; instead, the independent variables are the presumed causes of variability in well-being.

Compared to studies of age structure and cohort comparisons, the theoretical underpinnings of studies of aging and well-being are typically richer and more complex. A broad range of theories is used in studies of well-being in late life. Most of these theories are imported in whole or in part from other domains of sociology. For example, stratification theories are used to examine both income dynamics and health inequalities in later life, stress theory is used in studies of physical and mental illness, network theory is used to understand older adults' patterns of social support, and aspiration and equity theories are used in studies of subjective well-being. In addition to importing theories from other domains of sociology, a substantial proportion of research on aging and well-being rests on theories developed to highlight the distinctive conditions of later life. Examples include activity theory, which posits that multiple forms of well-being are enhanced in late life by sustaining high levels of activity and engagement – and adding new forms of activity to compensate for losses that often accompany aging; the double-jeopardy hypothesis, which predicts that the combined statuses of being old and a member of a racial/ethnic minority have more damaging effects on health and well-being than their purely additive effects; and socio-emotional selectivity theory (Carstensen et al., 2011), which suggests that declines in social contacts in late life are a purposeful and effective strategy for sustaining high-quality relationships and high levels of subjective well-being. Clearly, the theories used to explain variations in well-being among older adults are rich and varied.

The independent variables used in studies of well-being during late life are typically measures of social status, social context, and social resources. Social context, as indexed by age, race, and gender, is related to economic well-being, physical and mental health, and longevity. As expected, traditional indicators of SES are significant predictors of longevity, physical and mental health, self-esteem and self-efficacy, and subjective well-being. Social resources, especially social integration (e.g., organizational and religious participation) and social support also have positive effects on longevity, health, sense of self, and subjective well-being. As this brief description illustrates, we know a lot about the factors that explain heterogeneity in well-being in later life (e.g., Forthofer and Janz, 2001; Murrell et al., 2003). These same social factors also explain much of the variability in well-being among young and middle-aged adults. As the body of research that compares predictors of well-being across age groups documents, the distinctive feature of well-being in late life is not the specific antecedents of well-being, but rather the distribution of those antecedents across historical and biographical time.

Aging as the Culmination of the Life Course

During the past 40 years, the life-course perspective has assumed increasing influence in sociological research, especially research on aging. The core of the life-course perspective is the proposition that lives unfold over time and that events and conditions at earlier phases of the life course have

persisting effects at later phases – either continuing direct effects or indirect effects via more temporally proximate events and conditions. Taking temporality seriously changes research questions and research methods in multiple ways (Elder et al., forthcoming; George, 2003). Most importantly, perhaps, long-term patterns of change and stability, often conceptualized as trajectories, are the primary focus of analysis. Investigators can study trajectories of independent variables (e.g., marital history), trajectories of dependent variables (e.g., patterns of recovery, remission, and chronicity in depressive symptoms) or both. In addition to the 'shape' of trajectories, duration in states of interest also may be important (e.g., length of time till recovery from disability).

Methodologically, life-course studies require either multiple measurements over long periods of time or retrospective data about earlier phases of the life course. Longitudinal data are more accurate than retrospective data, but most life-course investigators are willing to use retrospective data if they are all that is available. The statistical techniques most frequently used in life-course studies (e.g., latent growth curve analysis) are newer and more complex than those used in cross-sectional or short-term longitudinal studies. Other methodological problems, ranging from substantial attrition in sample size to difficulties in selecting measurement tools that are applicable across adulthood, emerge when using longitudinal data covering long periods of time. Nonetheless, it is the way that research questions are conceptualized that is the hallmark of life-course studies.

Life-course studies provide important information about trajectories of vulnerability and resilience. At their best, they also incorporate the processes by which early events and conditions have persisting effects on outcomes of interest measured decades later. Work by Elder and colleagues on the long-term effects of combat exposure provides a compelling illustration of the knowledge generated by life-course studies (MacLean and Elder, 2007). This research is based on the Terman men, a sample of unusually intelligent males tested on multiple occasions from childhood to early old age. Most of these men participated in World War II, although not all of them were exposed to combat – and among those who were in combat, amount of exposure varied widely. Elder and colleagues demonstrate that combat exposure is a significant predictor of physical and, especially, mental health problems 40 years later, controlling on other known predictors of physical and mental health in later life. They also identified the life-course achievements that allowed some of these combat veterans to avoid or minimize subsequent health problems. Men who achieved greater socioeconomic success, those who sustained contacts with other combat veterans, and especially those who had lasting, high-quality marriages were able to avoid or minimize the health risks posed by combat exposure. This research demonstrates the benefits of life-course research, documenting both the persisting effects of early trauma and the mechanisms that allowed some men to avoid those risks.

In aging research, the life-course perspective has been used most frequently to understand the effects of life-course patterns of SES on multiple forms of well-being in later life. The conceptual framework underpinning most of this research is the *theory of cumulative advantage/disadvantage*. This theory was developed nearly a half-century ago by Robert Merton, as a framework for understanding occupational success among college and university professors (Merton, 1957). Merton observed that, at job entry, assistant professors looked very similar on standard measures of productivity and occupational success, regardless of the status of the institutions in which they were employed. Over time, however, variability increased dramatically in levels of productivity and occupational success among professors, with those employed at resource-rich schools exhibiting patterns of increasing success and those at resource-poorer schools exhibiting steady declines in productivity. Merton referred to the pattern as the Matthew effect. Subsequently, it was and is referred to as cumulative advantage/disadvantage theory (Dannefer and Daub, 2009). In colloquial terms, the theory of cumulative advantage/disadvantage posits that "the rich get richer and the poor get poorer."

In general, what Merton observed among professors is true for SES over the adult life course. That is, SES differences are smallest during young adulthood and largest during late life (Palloni et al., 2009). This is true for income and SES differences are even more dramatic for wealth. Thus, cumulative advantage/disadvantage generates increasing economic heterogeneity over the adult life course.

The life-course perspective also has been valuable in accounting for health inequalities across the life course. As is true for income and wealth, SES differences in health are minimal during young adulthood. By middle age, however, there are large differences in health between the lowest and highest SES quartiles in the United States. Evidence is less clear during old age, with some investigators reporting that SES differences in health continue to widen during late life. Other researchers, however, report that SES differences in health are largest during middle-age and narrow somewhat during old age, although they remain significant (Brown et al., 2012; Dupre, 2007). At some point in late life, SES differences in health are likely to narrow as the result of selective mortality (i.e., the earlier deaths of many lower SES individuals). Some of the mechanisms that account for SES differences in health also have been identified. The benefits of education, occupational status, income, and wealth are, of course, much of the basis of SES differences in health. Other mechanisms that mediate the effects of SES on health include health behaviors; stress, including lifelong accumulation of stressors; and psychological resources such as self-esteem and mastery.

In addition to long-term trajectories of assets and liabilities, the life-course perspective also focuses attention on the persisting effects of early life events and conditions. For example, evidence demonstrates that parental SES predicts health during middle and late life, over and above the effects of individuals' own SES trajectories (Pavalko and Caputo, 2013). Indeed, limited evidence suggests that fetal growth *in utero* plays a substantial role in health during late life (Risnes et al., 2011). In addition to economic resources, lack of emotional support from parents during early childhood increases the risk of both depression and chronic physical illnesses in late life (Shaw et al., 2004). Similarly, a variety of childhood traumas – including child abuse, sexual abuse, persistent poverty, and parental divorce – are known to increase the risk of many mental illnesses decades later, even with other known risk factors taken into account (McLaughlin et al., 2011).

Although the volume of life-course research that extends to late life has increased dramatically during the past three decades, many other topics could be profitably addressed in a life-course framework. For example, although there is compelling evidence that social relationships are powerful predictors of health and well-being in late life, little is known about the relative importance of lifelong patterns of social bonds as compared to the effects of social networks developed in late life. Similarly, religious participation has been demonstrated to be a strong predictor of mortality and morbidity in late life (George et al., 2002). However, this research is based on studies in which *current* religious involvement is measured. Virtually nothing is known about how length of exposure to religious participation affects health. In other words, we don't know how long individuals must participate in religious activities before health benefits are observed.

The life-course perspective also renders the usual distinction between social selection and social causation moot. Studying the life course is the equivalent of studying patterns of selection and causation as they unfold across personal biography. And, of course, the effects of social causation observed earlier in the life course become selection effects for outcomes observed later in life.

The life-course perspective focuses on the complex links between social/historical change and personal biography. In addition, the life-course perspective is ideally suited to linking macro- and meso-level social conditions to individual behaviors and well-being, to tracing the effects of structural opportunities and constraints, highlighting the effects of human agency (i.e., personal choices) over the long haul, and documenting the many ways that the past is indeed prologue to the future. Thus, life-course research is an important and exciting part of the sociology of aging.

See also: Age Structure; Aging and Health in Old Age; Aging and Older People, Geography of; Aging and Work; Aging and the Labor Market; Aging, Theories of; Chronic Illness, Psychosocial Coping with; Fertility Theory; Life Course: Sociological Aspects; Period and Cohort Analysis in Demography; Population Aging: Economic and Social Consequences; Retirement and Encore Adulthood: The New Later Life Course; Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model.

Bibliography

Bourgeois-Pichat, J., 1979. The demographic transition: aging of population. In: Population Science in the Service of Mankind. International Union for the Scientific Study of Population, Liege, pp. 211–239.

Brown, T.H., O'Rand, A.M., Adkins, D.E., 2012. Race-ethnicity and health trajectories: tests of three hypotheses across multiple groups and health outcomes. Journal of Health and Social Behavior 53, 359–377.

Carstensen, L.L., Turan, B., Scheibe, S., Ram, N., Ersner-Hershfield, H., Samanez-Larkin, G.R., Brooks, K.P., Nesselroade, J.R., 2011. Emotional experience improves with age: evidence based on over 10 years of experience sampling. Psychology and Aging 26, 21–33.

Dannefer, D., Daub, A., 2009. Extending the interrogation: life span, life course, and the constitution of human aging. Advances in Life Course Research 14, 15–27.

Dupre, M.E., 2007. Educational differences in age-related patterns of disease: reconsidering the cumulative disadvantage and age-as-leveler hypotheses. Journal of Health and Social Behavior 48, 1–15.

Elder Jr., G.H., 1999. Children of the Great Depression, twenty-fifth anniversary ed. Westview Press, Boulder, CO.

Elder Jr., G.H., Shanahan, M.J., Jennings, J.A. Human development in time and Place. In: Leventhal, T., Bornstein, M. (Eds.), Handbook of Child Psychology and Developmental Science: Ecological Settings and Processes in Developmental Systems, vol. 4, seventh ed. Wiley & Sons, New York, forthcoming.

Forthofer, M.S., Janz, N.K., 2001. Gender differences in the associations of self-esteem, stress, and social support with functional health status among older adults with heart disease. Journal of Women and Aging 13, 19–36.

George, L.K., 2003. What life course perspectives offer the study of aging and health. In: Settersten, R.A. (Ed.), Invitation to the Life Course. Baywood Publishing, Amityville, NY, pp. 161–190.

George, L.K., Ellison, C.G., Larson, D.B., 2002. Explaining the relationships between religious involvement and health. Psychological Inquiry 13, 190–200.

Lynch, S.M., George, L.K., 2002. Interlocking trajectories of loss-related events and depressive symptoms among elders. Journal of Gerontology: Social Sciences 57B, S1170–S123.

MacLean, A., Elder Jr., G.H., 2007. Military service in the life course. Annual Review of Sociology 33, 175–196.

Marin, L.G., Freedman, V.A., Schoeni, R.F., Andreski, P.M., 2009. Health and functioning among baby boomers approaching 60. Journal of Gerontology: Social Sciences 64B, 369–377.

McAdam, D., 1990. Freedom Summer. Oxford University Press, Cambridge.

Merton, R.K., 1957. Priorities in scientific discovery: a chapter in the sociology of science. American Sociological Review 22, 635–659.

McLaughlin, K.A., Breslau, J., Green, J.G., Lacoma, M.D., Sampson, Zaslavsky, A.M., Kessler, R.C., 2011. Childhood socioeconomic status and the onset, persistence, and severity of DSM-IV mental disorders in a U.S. national sample. Social Science & Medicine 73, 1088–1096.

Murrell, S.A., Salsman, N.L., Meeks, S., 2003. Educational attainment, positive psychological mediators, and resources for health and vitality in older adults. Journal of Aging and Health 15, 591–615.

Palloni, A., Milesi, C., White, R., Turner, A., 2009. Early childhood health, reproduction of economic inequalities and the persistence of health and mortality differentials. Social Science Medicine 68, 1574–1582.

Pavalko, E.K., Caputo, J., 2013. Social inequality and health across the life course. American Behavioral Scientist 57, 1040–1056.

Phillips, S.M., Wojcicki, T.R., McAuley, E., 2013. Physical activity and quality of life in older adults: an 18-month panel analysis. Quality of Life Research 22, 1647–1654.

Riley, M.W., Riley, J.W., 1994. Structural lag: past and future. In: Riley, M.W., Kahn, R.L., Foner, A. (Eds.), Age and Structural Lag. Wiley, New York, pp. 15–36.

Risnes, K.R., Vatten, L.J., Baker, J.L., Jameson, K., Sovlo, U., Kajanitie, E., Osler, M., Morley, R., Joketa, M., Painter, R.C., Sundh, V., Jacobsen, G.W., Eriksson, J.G., Sorensen, T.A., Bracken, M.B., 2011. Birthweight and mortality in adulthood: a systematic review and meta-analysis. International Journal of Epidemiology 40, 647–661.

Schieman, S., Van Gundy, K., Taylor, J., 2002. The relationship between age and depressive symptoms: a test of competing explanatory and suppression influences. Journal of Aging and Health 14, 260–285.

Shaw, B.A., Krause, N., Chatters, L.M., Connell, C.M., Ingersoll-Dayton, B., 2004. Emotional support from parents early in life, aging, and health. Psychology and Aging 19, 4–12.

Siedlecki, K.L., Salthouse, T.A., Oishi, S., Jewani, S., 2014. The relationship between social support and subjective well-being across age. Social Indicators Research 117, 561–576.

Thomas, P., 2011. Trajectories of social engagement and limitations in late life. Journal of Health and Social Behavior 52, 430–443.

Yang, Y., 2007. Is old age depressing? Age trajectories and cohort variations in late life depression. Journal of Health and Social Behavior 48, 16–32.

Age Stratification

Dale Dannefer and Tirth Bhatta, Case Western Reserve University, OH, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D. Dannefer, volume 1, pp. 278–283, © 2001, Elsevier Ltd.

Abstract

Age stratification can refer to (1) the age composition of a population, (2) features of social structure, and (3) a 'conceptual framework' for approaching the study of age as a feature of societies as well as individuals. The age composition of the population is produced by the piling up of cohorts (those born during the same time period) at a single point in time, creating what has often been called an age pyramid – a diagram with age categories on the vertical axis and a sex-bifurcated population count on the horizontal axis. The relevance of age as a compositional feature of a population depends on the meaning and normative significance of age in social structure, which has varied substantially across time and across societies. At the same time, the social significance of age depends, to some extent, on the age composition of the population. In attempting to conceptualize the dynamics of age in terms of stratification, scholars have also analyzed the potentials of age as a basis for social solidarity and conflict, and have considered the relationship of age and other bases of stratification. The dynamics involved in these relationships often entail complex interactions between intracohort and intercohort patterns of inequality or other forms of differentiation.

In the human sciences, stratification generally refers to those forms of differentiation that entail ordinal ranking along a defined dimension. Age is an inherently ordinal phenomenon, anchored in the intersection of time and the event of birth. Those individuals born within a given time period comprise a cohort. The defining cohort feature of lived time, or age, defines age difference (Dannefer, 1988). Age strata, however defined, are thus comprised of the continuous succession of living cohorts, glimpsed at a single point in time.

Age Stratification as a Feature of Populations

Figure 1 depicts the age composition of the US population in 1900 and the projected pattern for 2040. The ranked difference is determined by different dates of birth, and succeeding cohorts are essentially piled on top of each other at a single point in time to form a 'snapshot' of the age composition of the population. A comparison of these two figures reflects the dramatic changes that occurred in the intervening 140 years, including population growth, increased life expectancy, recent fertility declines, and the shift of the societal burden of dependency away from youth and toward age. Such changes in age composition result from changes in fertility, mortality, and migration patterns, and have been significantly driven by the maturing baby-boom cohorts (b. 1946–64). Changes in these population processes have resulted from a diverse array of social forces as well as technological and public health improvements (Cutler and Miller, 2005; Anderson and Hussey, 2000).

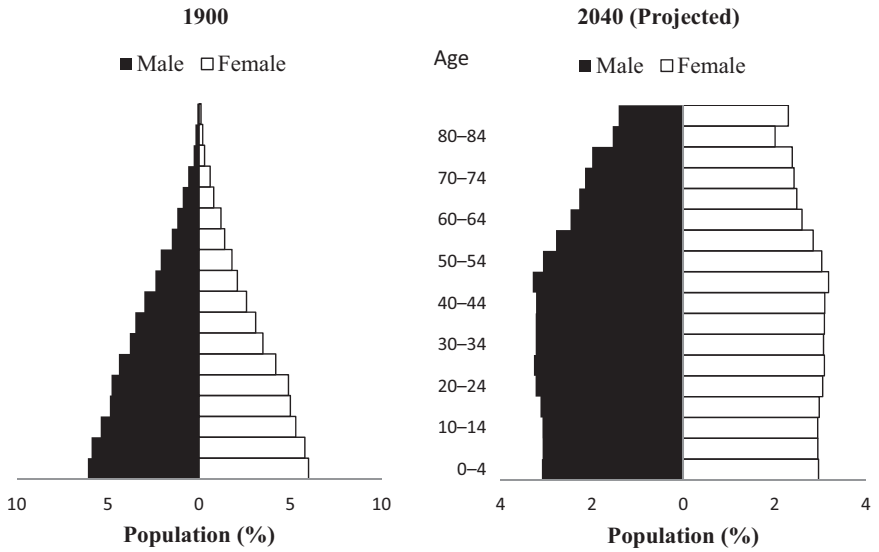

Figure 1 US population in millions by 5-year age categories, 1900 and 2040 (projected). Source: US Census Bureau, 2012. Projections of the Population by Age and Sex for the United States: 2015 to 2060. Population Division, Washington, DC; Klein, H.S., 2004. A Population History of the United States. University Press, Cambridge (adapted).

The shape of the age distribution also has an independent impact on society more generally. Scholars of age recognize immediately the consequences of such shifts in age distribution of a population, which can affect aspects of society as diverse as employment prospects for graduating students, the meaning of old age, marriage markets, and one's economic life chances (Uhlenberg, 1992). Without some knowledge of social context, however, such figures reveal nothing about how age is implicated in the overall structure of society. The effects of age are always contingent on specific institutionalized regimes. For example, a shift in the aged dependency ratio might seem a very different kind of problem in modern welfare states than an agrarian economic system like in Nepal, where the family is expected to share the primary responsibility of providing a safety net for older adults, and where children and many elders may make valued contributions to the household economy. Thus, analysis of age strata inevitably requires a consideration of stratification as a component of social structure.

Age Stratification as a Feature of Social Structure

The Cultural Significance of Age

As a feature of social structure, the boundaries and character of age strata are anchored in the broader social meaning and definition of age and time. The social meaning of age depends upon the political, economic, and cultural significance of a given age stratum. The manner in which societies define age and cohort membership is, of course, not universal but variable. Indeed, the very awareness of age and cohort membership is variable and, ultimately, socially constituted as a more or less integral feature of the larger social order. Societies vary in how they count, in how they measure the passage of time, and in how they identify time of birth. In some societies, the age set – a group of individuals spanning several years – is the central social unit defining life stage and role transition processes (Foner and Kertzer, 1978). For instance, in an ancient Hindu tradition, age was considered meaningful only in relation to four life stages and role transitions (i.e., student stage, householder stage, hermit stage, and wandering ascetic stage) of an individual (Kakar, 1968). In many other traditional contexts, age is similarly defined in descriptive terms – perhaps referenced by memorable events, such as political transitions or natural disasters and without concern for chronometric precision. Indeed, the standardized measurement of both time and age are remarkably recent human practices, driven by economic and political developments (e.g., Baars and Visser, 2007; Bartky, 2007). Despite their widespread taken-for-grantedness, the modern practices of defining age and time by precisely quantified calibrations are thus not essential elements of human nature, but are recent products of historically and socially specific systems of language, ideas, and interests.

Age and Bureaucratic Rationality

Within a single society, the significance of age may undergo dramatic alteration in response to changes in other aspects of social structure. An obvious prototypical case is the transformation to modernity, which generally meant the emergence of age-graded strata supported by newly identified life stages, the legal-bureaucratic reliance on age as an eligibility criterion, and an increase in age consciousness (Chudacoff, 1989). The result has been the institutionalization of the life course (Kohli, 2007). Viewed as a contemporaneous societal cross-section, this same phenomenon comprises the institutionalization of age strata in which the everyday experience of members of the several age strata are differentially organized by legal, governmental, and corporate policies and broader social practices that are explicitly age graded. In age-graded societies, such as modern welfare states, the stratum membership is a quite robust basis for predicting one's likelihood of being socially engaged or 'significant' (Uhlenberg, 1988).

Standard sociological wisdom declares that modern states assign roles based on universalistic criteria that reflect individual abilities rather than sponsorship or ascription. It is thus ironic that these societies have developed criteria (both formal and informal) to govern access to and exclusion from valued social positions based on the ascribed characteristic of age. 'Children,' 'adolescents,' and 'retirees' are examples of life-stage constructs that authoritatively define 'natural' capabilities based on age, even though these 'life stages' were largely unheard of before 1850s. As was true in many other agrarian societies, in the preindustrial USA it was not age but physical functioning and social status (e.g., owning land) that determined one's social position (Chudacoff, 1989; Demos, 1978). The same appears to have been the case in ancient times (Parkin, 1998).

Age has proven to be a useful criterion for bureaucracies faced with the task of managing and processing large populations. It has been used to regulate access to scarce roles and resources, and for attributing competence (or its lack) to individuals based on generalized assumptions of age-related capacities rather than actual abilities. Such practices are taken for granted, in that the rationality of age as an exclusionary criterion is seldom questioned by the rational-bureaucratic logic of the systems that rely upon it, just as ageism a social bias, often goes unrecognized by the most passionate activists for social justice in other areas.

Age Stratification and Social Theory

The Age Stratification Paradigm

The organization of society in terms of an age-stratified social structure must be analytically distinguished from the composition of the population of individuals of different ages who occupy positions in the social structure, and who move through a sequence of age-graded roles as they age. This fundamental sociological distinction between actors and structure, and persons and roles, was articulated for gerontology by Matilda Riley and associates four decades ago, but is still often overlooked (Dannefer et al., 2005; Riley et al., 1972). Confusion can arise if this analytical distinction is not recognized (e.g., when normative age-graded roles become taken for granted and assumed to be part of human nature). The importance of this distinction is a key premise of the pioneering age stratification – later termed 'aging and society' framework – advanced by Matilda Riley and associates (Riley et al., 1972).

One key principle of the aging and society framework is *structural lag* – 'society's failure to provide meaningful roles …,'

for people of all ages (Riley et al., 1994). The archetypal example is the incongruity between (a) a dramatic growth in longevity and late-life vitality among aged individuals and (b) social institutions and policies that are premised on entrenched ageist beliefs in the incompetence and obsolescence of the aged. Riley's work thus provided a sociological foundation that supported or anticipated other scholarly and popular efforts to question the prevailing social organization of age roles.

Some have criticized Riley's approach for its reliance on terminology associated with functionalism. In fact, her approach posed a challenge to the traditional functionalist position, by making visible the human costs and contradictions of age-graded social structures, with the intent of challenging conventional views that prevailed within as well as beyond the social sciences. It is only by making clear, for example, that men and women are not naturally *supposed* to engage in a 40-year stint of uninterrupted work, or that older people must not *inevitably* 'retire,' that it becomes possible to have a basis for a critical analysis of the social organization of age. Allocation implies the existence of an age-stratified opportunity structure with a finite number of roles, within which individuals must find meaningful social engagement. Given the limited elasticity in the number of available age-graded roles, there is considerable potential for an 'imbalance between persons and roles' which can mean an exclusion from desired social participation that is costly for both the individual and the society. Since the cohorts that populate strata are constantly aging, this is especially true when there are sharp discontinuities in the size of adjacent strata (or changes in the aspiration or abilities of stratum members). In stratification terms, such a situation may be called a disordered age structure, a cross-sectional 'snapshot' perspective on Waring's (1975) important concept of disordered cohort flow.

More generally, an issue of ongoing debate in this area concerns the question of whether the potentials of individuals can be realized within an institutionalized role structure imposed by governmental and corporate practices and the resultant norms that regulate access to roles on the basis of age. The systems that produce allocation problems also legitimate and sustain normative age-related expectations, especially those associated with the linear life course consisting of the 'three boxes' – of school, work, and retirement. As a structural alternative to this model, Riley and Riley (1994) proposed a model of an 'integrated life course,' which would imply a dismantling, or as Held (1986) has termed it, "deinstitutionalization" of the life course.

Thus, the terms 'age-integrated life course' and 'deinstitutionalized life course' both envision a similar process and direction of social change. However, the manner in which this process is framed can appear quite differently. On one hand, the idea of an integrated life course within which work, schooling, and leisure time can be experienced in balanced proportions throughout the life course, offering individuals a wider range of life options and opportunities. In this model, control of the design of the individual life-course shifts away from an institutional regime that has tended to stratify individual opportunities and normative possibilities according to age stratum boundaries, and toward an arrangement that affords a greater voice for self-expression.

At the same time, many postindustrial societies with advanced forms of life-course institutionalization (including elaborate education, welfare, and pension benefits) are facing the reality of challenges and in some cases erosion of the institutionalized life course, driven by the daunting economic and demographic challenges, and visible in the midlife disruption of careers once thought to be secure, or in high unemployment rates of young adults, or reduced pension benefits. Such changes may prompt wistfulness for the security of the institutionalized life course, despite ambivalence toward its oppressive aspects.

As examples of other social-structural approaches to age, two general lines of theorizing will be briefly discussed: (1) the effects of cohort size and composition and (2) the potentials for intergenerational conflict.

Effects of Stratum Size

Several scholars have argued that cohort size (and hence, stratum size) has a range of consequences for the lives of cohort or stratum members, including psychological (Waring, 1975) and socioeconomic (Easterlin, 1987) effects. The general argument is that large cohorts are disadvantaged relative to smaller ones, since their members must compete for scarce, age-graded roles and resources. Such ideas have found support in analyses of education and work careers based on aggregated data (Pampel and Peters, 1995; Dannefer, 1988). Of course, all such notions rest on the premise of a rigidly stratified age-role structure: of age as a normative if not legal qualification for occupying desirable educational and occupational positions. To the extent that this is true, the movement of unusually large cohorts through the age structure should mean that the status and resources of the various strata will change over time as the size of strata changes.

The rapid demographic change of the past century has provided a natural laboratory for exploring hypotheses about the relation of the age structure of the population and individual lives.

As increasing life expectancy has produced a graying of the population, elders may also experience disadvantage based on their numbers. For example, in preindustrial USA, the very old were often seen as the experts regarding health and longevity (cf Achenbaum, 1979). They were few, and with no prestigious medical profession, many of these survivors carried a mystique that seemed to imply expertise. Nevertheless, as will be shown in subsequent sections, the status of the aged in nineteenth century USA as in other premodern societies derived from much more than the rarity of nonagenarians. It was broadly rooted in the control of resources by the aged, and the limited options open to young people – circumstances fixed by laws of inheritance and other customary practices that organized the distribution of power in society.

Interstratum Inequality: the Intersection of Age and Other Bases of Stratification

As an ordinally ranked characteristic, age itself is an analytically distinguishable basis of stratification. Yet its significance typically involves its intersection with other dimensions of stratification, especially those involving control of economic, political, or cultural resources, demarcating differentiation or stratification *within* a given age stratum. However, the

interpretation of interstratum differences becomes a central concern. Differences in any measurable characteristic (e.g., health, wealth, and political attitudes; intrastratum inequality) may represent life-course, cohort, or period effects (Yang, 2008; Riley et al., 1972; Dannefer, 1988). In a cross-cultural analysis of age inequality and conflict, Nancy Foner (1984) identified several factors that privilege senior age strata in many societies: control over human and material resources; knowledge, expertise, and experience; prestige and status; community influence; and wisdom or mystical power. Many of these factors also appeared to operate in preindustrial Western societies. With the transformation to modernity, the aged lost both status and economic power, as the venerated characteristics of experience, skill, and wisdom were supplanted by physical strength and endurance, youthful beauty, up-to-date knowledge, and willingness to deal with rapid change.

Some scholars have argued that this structural transformation brought relief to strong intergenerational tensions that were present in many traditional agrarian families in New England and in Europe as, for example, when senior landholders survived until their children were well into or past middle age. When the family was replaced by the firm as the primary unit of economic production, the central familial relationship of economic dependency was removed (Kertzer and Laslett, 1995) – meaning that the relations of adult children and aging parents became more premised on volition and sentiment, rather than economic interests. By contrast, others have proposed that the rapid pace of technological and educational change characterizing modernity tended to create cleavages between cohorts in worldview and values, dramatically increasing the likelihood of interstratum conflict.

In either case, the characterization of modernity as removing the issue of intergenerational economic strain places the entire topic of the prospect of interstratum conflict on subjective grounds. Both arguments focus on values and sentiments. Nevertheless, a resurgence of arguments about economic interest as a source of interstratum conflict occurred in the late twentieth century as concern over the cost of entitlements for the age became a widespread public concern in most modern and postindustrial societies.

Economics and Generational Equity

In the USA, the politicized generational equity controversy has in various guises persisted since the early 1980s. This debate returns economic issues to the foreground of potential age conflict, and illustrates the potential for age to become a basis of interest-group politics (Williamson et al., 1999). The debate is fueled by opposition to the scale of age-based 'entitlements' and by discontent over what is seen as dramatic shifts in the age distribution of resources, especially from children to old people. Even though it can readily be argued that the age-qualified pensions and transfer payments given to older adults have also indirectly benefited younger adults, the proponents of generational equity controversy only highlight special economic consideration in such policies to the senior age strata (Schulz and Binstock, 2008). If being in a large cohort/stratum is disadvantageous for educational and employment opportunities, it may be the reverse for interest group politics, where the senior strata are recognized as comprising an active and growing segment of the electorate.

Grounded in such worldview, the USA continues to see vigorous efforts to pit young versus old (Peterson, 1999). Yet such efforts have yet to gain significant political traction. This failure was essentially predicted by theoretical principles set forth by Anne Foner. One such principle, *age mobility*, sees the inevitability of aging as producing anticipation on the part of midlife adults of their own certain movement into more senior age strata (Foner, 1974). A second principle involves the insulation from direct responsibility for dependent parents that Social Security and Medicare/Medicaid provide to adult offspring. Although intergenerational resource transfers may be more likely to go from parents to children than the reverse, the protection that many midlife adults are afforded by public subsidy of seniors is very real. Indeed, 'downward' resource transfers may depend, in many cases, on the public benefits afforded to the senior generations (Albertini et al., 2007). Other sources of resistance to 'generational equity' may include a belief in intergenerational economic continuity within families, which implies a relatively stable intergenerational reproduction of the structure of economic inequality along familial lines. For example, it may be the affluent neighbors or overpaid supervisor who become the focus of a sense of economic deprivation and limited life chances, and not one's comfortably situated parents.

Interstratum Variation in Intrastratum Inequality

Awareness of nonaged-related economic inequalities that are based on stratification of the general opportunity structures is made even more likely since intrastratum economic disparities appear to be greater in aged strata than in younger ones, as economic inequalities appear to cumulate across the life course of each succeeding cohort (Dannefer, 2003, 1988), creating a picture of an age structure in which inequality in a wide range of outcomes (e.g., socioeconomic status and health) is higher in the older age strata than in younger ones.

Despite characterizations of many traditional societies in gerontocratic terms, it appears that many aged in such societies enjoyed neither wealth nor status. Evidence for this comes from studies of the preindustrial USA (e.g., Demos, 1978) as well as from traditional (Foner, 1984) and ancient (Parkin, 1998) societies. If elders were often the most powerful members of traditional societies, it does not follow that most members of those age strata were affluent or particularly respected. Overemphasis on within-stratum homogeneity and between- or interstratum differences has often come at the cost of not exploring within-group differences or intrastratum inequality. Just as intercohort differences in patterns of intracohort inequality has been a neglected area of life-course research, interstratum differences in intrastratum inequality is a neglected question in analyses of age structure.

Summary

Age stratification brings together the deceptively elusive individual characteristic of age, and the complex and dynamic social phenomena of age as a social construct and as a basis of

stratification. A grasp of society as a systemic structural reality that shapes many aspects of aging has been slow to develop among those who focus on the individual-level characteristics of age and development. Conversely, a grasp that individual social actors – whether citizens, students, parents, or workers – are arrayed in continuously aging and changing cohorts has been underappreciated by sociologists. The study of age stratification has contributed a clear and forceful message concerning the importance of distinguishing age and age-specific subpopulations on the one hand from normatively age-graded social practices and structures on the other. It has contributed principles helpful in understanding how population and social structure jointly impact individual life chances, and has begun to consider the conditions governing the form of the relation of age and other bases of stratification. As issues such as the impact upon individuals of age-graded structures and the question of how socioeconomic stratification and age may interact, new research agendas have begun to take shape around questions of stratification and age.

See also: Age Policy; Age Structure; Age, Sociology of; Aging and Health in Old Age; Aging and Older People, Geography of; Aging and Work; Aging and the Labor Market; Aging, Theories of; Capitalism; Class: Social; Critical Theory: Contemporary; Demographic Measurement: General Issues and Measures of Fertility; Equity in Law; Family Theory: Economics of Intergenerational Relations; Feudalism; Functionalism in Anthropology; Functionalism, History of; Globalization and Health; Health Politics and Policy; Identity in Anthropology; Income Inequality; Industrialization; Inequality, Social; Intergenerational Solidarity; Life Course in History; Life Course: Sociological Aspects; Linked Lives; Modernization and Modernity in History; Modernization, Sociological Theories of; Modernization; Mortality: The Great Historical Decline; Old Age Work Participation; Peasants and Rural Societies in History (Agricultural History); Population Aging: Economic and Social Consequences; Poverty in History; Power in Society; Race: Ethnicity and Health; Social Exclusion, Social Deprivation and Health; Social Inequality and Schooling; Social Inequality in History (Stratification and Classes); Social Justice: Historical and Theoretical Considerations; Social Mobility, History of; Social Stratification; Socioeconomic Status and Health Care; Socioeconomic Status and Health; Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model; Welfare Reform.

Bibliography

Achenbaum, A., 1979. Old Age in the New Land: The American Experience since 1790. Johns Hopkins University Press, Baltimore, MD.

Anderson, G.F., Hussey, P.S., 2000. Population aging: a comparison among industrialized countries. Health affairs 19 (3), 191–203.

Albertini, M., Kohli, M., Vogel, C., 2007. Intergenerational transfers of time and money in European families: common patterns—different regimes? Journal of European Social Policy 17 (4), 319–334.

Baars, J., Visser, H., 2007. Aging and Time: Multidisciplinary Perspectives. Baywood Press, Amityville, NY.

Bartky, I.R., 2007. One Time Fits All: The Campaigns for Global Uniformity. Stanford University Press, Stanford, CA.

Chudacoff, H., 1989. How Old Are You? Age Consciousness in American Culture. Princeton University Press, Princeton, NJ.

Cutler, D., Miller, G., 2005. The role of public health improvements in health advances: the twentieth-century United States. Demography 42 (1), 1–22.

Dannefer, D., Uhlenberg, P., Foner, A., Abeles, R.P., 2005. On the shoulders of a giant: the legacy of Matilda White Riley for gerontology. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 60 (6), S296–S304.

Dannefer, D., 2003. Cumulative advantage/disadvantage and the life course: cross-fertilizing age and social science theory. Journals of Gerontology Series B: Psychological Sciences and Social Sciences 58 (6), S327–S337.

Dannefer, D., 1988. Differential aging and the stratified life course. In: Maddox, G.L., Lawton, M.P. (Eds.), Annual Review of Gerontology and Geriatrics, vol. 8. Springer, New York.

Demos, J., 1978. Old age in early New England. American Journal of Sociology 84, S248–S287.

Easterlin, R., 1987. Birth and Fortune: The Impact of Numbers on Personal Welfare, second ed. University of Chicago Press, Chicago, IL.

Foner, A., 1974. Age stratification and age conflict in political life. American Sociological Review 39, 187–196.

Foner, A., Kertzer, D., 1978. Transitions over the life course: lessons from age-set societies. American Journal of Sociology 83, 1081–1104.

Foner, N., 1984. Ages in Conflict: A Cross-cultural Perspective on Inequality between Old and Young. Columbia University Press, New York.

Held, T., 1986. Institutionalization and deinstitutionalization of the life course. Human Development 29 (3), 157–162.

Kakar, S., 1968. The human life cycle: the traditional Hindu view and the psychology of Erik Erikson. Philosophy East and West 18 (3), 127–136.

Kertzer, D.I., Laslett, P. (Eds.), 1995. Aging in the Past: Demography, Society and Old Age. University of California Press, Berkeley, CA.

Kett, J., 1977. Rites of Passage: Adolescence in America, 1790–1920. Free Press, New York.

Klein, H.S., 2004. A Population History of the United States. Cambridge University Press, Cambridge.

Kohli, M., 2007. The institutionalization of the life course: looking back to look ahead. Research in Human Development 4, 253–271.

O'Rand, A., 1996. The precious and the precocious: the cumulation of advantage and disadvantage over the life course. Gerontologist 36, 230–238.

Pampel, F.C., Peters, H.E., 1995. The easterlin effect. Annual Review of Sociology 21, 163–194.

Parkin, T.G., 1998. Ageing in antiquity: status and participation. In: Johnson, P., Thane, P. (Eds.), Old Age: From Antiquity to Post-Modernity. Routledge Press, New York, pp. 19–42.

Peterson, P.G., 1999. Gray dawn: How the Coming Age Wave Will Transform America—and the World. Times Books, New York.

Riley, M.W., Johnson, M.E., Foner, A., 1972. Aging and Society, vol. III. Russell Sage, New York.

Riley, M.W., Kahn, R., Foner, A., 1994. Age and Structural Lag: Society's Failure to Provide Meaningful Opportunities in Work, Family and Leisure. Wiley Interscience, New York.

Riley, M.W., Riley, J.W., 1994. Age integration and the lives of older people. The Gerontologist 34 (1), 110–115.

Riley, M.W., Riley, J., 1999. The aging and society paradigm. In: Bengtson, V.L., Schaie, K.W. (Eds.), Handbook of Aging Theory. Springer, New York.

Schulz, J.H., Binstock, R.H., 2008. Aging Nation: The Economics and Politics of Growing Older in America. Johns Hopkins University Press, Baltimore.

US Census Bureau, 2012. Projections of the Population by Age and Sex for the United States: 2015 to 2060. Population Division, Washington, DC.

Uhlenberg, P., 1992. Population aging and social policy. Annual Review of Sociology 18, 449–474.

Uhlenberg, P., 1988. The societal significance of cohorts. In: Birren, J.E., Bengtson, V.L. (Eds.), Emergent Theories of Aging. Springer, New York.

Waring, J., 1975. Social replenishment and social change. American Behavioral Scientist 19, 237–256.

Williamson, J., Watts-Roy, D.M., Kingson, E.R., 1999. The Generational Equity Debate. Columbia University Press, New York.

Yang, Y., 2008. Social inequalities in happiness in the United States, 1972 to 2004: an age-period-cohort analysis. American Sociological Review 73 (2), 204–226.

Age Structure

Shiro Horiuchi, CUNY School of Public Health and CUNY Institute for Demographic Research, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Trends in the age structure of a population should be considered in policy development and planning by governments and private businesses, because demands for goods and services, labor force participation rates, health conditions, and people's behavioral patterns vary substantially with age. Age structure is primarily determined by fertility, mortality, and migration during the last several decades. Most developed and developing countries in the world have experienced notable changes in the age structure due to significant declines in fertility and mortality. Various statistical and mathematical tools have been developed for measuring and analyzing age structure trends and their effects.

The age structure (or age composition) of a population is the pattern of its distribution by age. The age distribution is often shown separately by sex and displayed as the *population pyramid*, in which the sizes of age–sex groups are indicated by horizontal bars arranged from the youngest age-group at the bottom to the oldest age-group at the top. Several examples of population pyramid are in **Figure 1**. The graphs are called population *pyramids* because the graphs often exhibit triangular shapes, partly due to the tendency for the number of survivors to decrease with age and partly due to the increasing size of birth cohort in growing populations. The age structure is *young* if the proportion of younger people is high and that of older people is low (e.g., India, 1980, in **Figure 1(c)**). The age structure is considered *old* if the opposite is true (e.g., Sweden, 2011, in **Figure 1(b)**).

It is crucial to consider current and prospective changes in the age structure when planning government policies and forecasting consumer demand and labor supply for private industries (*see* Age Policy; Population Aging: Economic and Social Consequences). Demands for different types of goods and services vary significantly with age. A large youth population implies high demand for school facilities and teachers. A greater elderly population may mean more needs for nursing homes, medical services, and hospital facilities and a higher amount of pension payments. A society with a rapidly increasing number of people in 20s and early 30s may need to create more employment opportunities and provide new housing.

Furthermore, labor force participation is null or low at both the ends of age spectrum and high in between. If the proportion of the working-age population (between the age when people typically start their first full-time job and the typical retirement age) is decreasing, people in the labor force may have a growing burden of economically supporting those outside it.

Underlying these societal consequences of age structure are age variations in health conditions, behavioral patterns, and socioeconomic characteristics of people (*see* Motivation: Life Course and Sociological Perspectives; Life Course: Sociological Aspects). This is partly because humans experience substantial age-associated physiological changes in their lives (physical development at young ages, sexual maturation, and senescent processes at middle and old ages), and partly because age-related social roles have been built on these biological life history patterns.

Population Dynamics and Age Structure

The age structure is determined primarily by the levels, age patterns, and trends of fertility, mortality, and migration during the past few decades (Coale, 1972). A population with higher fertility tends to have a larger proportion of children, and a population with lower mortality has a higher proportion of people who survive to old ages. In a growing population, the sizes of more recent birth cohorts tend to be larger, producing a young age structure, but a decreasing population tends to have an old age structure.

The migration rate tends to be high among young adults. Thus, usually migration reduces the proportion of young adults in countries that send out migrants, and increases the proportion in countries that receive them. The long-term effects of migration on the age structure are complicated if fertility and mortality patterns of the migrants are considerably different from those of the native population, or if there are significant flows of return migration.

During the last one and a half centuries, the age structure has become older in the majority of countries in the world. Significant population aging started during the first half of the twentieth century in most economically and technologically developed countries, and it started during the second half in many developing countries. Figures 1(a) and 1(b) illustrate the historical age structure changes in Sweden, indicating that the age composition of Sweden is noticeably older in 2011 than in 1900. Figures 1(c) and 1(d) show data from India as an example of a developing country. The relative narrowing of the pyramid base from 1980 to 2011 reflects recent fertility declines in India. The difference in the timing of population aging between developed and developing countries is illustrated in **Figure 2** for Sweden and India.

The population aging was due to considerable declines in mortality and fertility. This international demographic trend is called the *demographic transition*, which accompanied the industrialization of economy and related social changes. The age structure of developed countries, where the demographic

 http://dx.doi.org/10.1016/B978-0-08-097086-8.34030-2

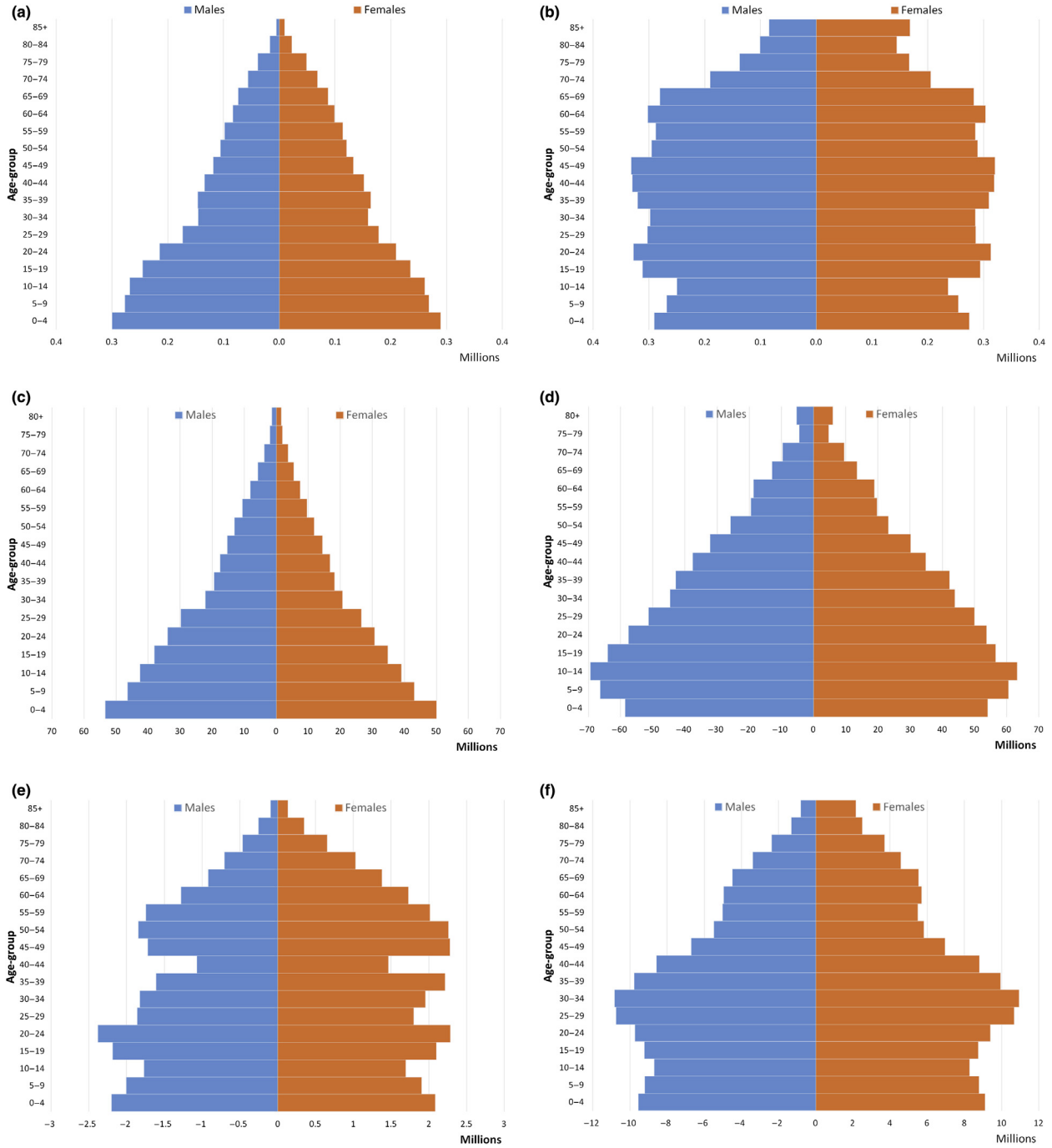

Figure 1 Selected population pyramids: (a) Sweden, 1900; (b) Sweden, 2011; (c) India, 1980; (d) India, 2011; (e) West Germany, 1960; (f) The United States, 1990.
Sources: Sweden, 1900 and 2011; West Germany, 1960; USA., 1990. Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at http://www.mortality.org (data downloaded on 02.04.14.); India, 1980: United Nations (2012). World Population Prospects: The 2012 Revision; India, 2011: Office of the Registrar General and Census Commissioner. Census of India 2011. Available at http://censusindia.gov.in.

transition started earlier, is significantly older than that of developing countries. (For example, compare the population pyramids of Sweden, 2011, in **Figure 1(b)**, and India, 2011, in **Figure 1(d)**.)

Until recently, the major driving force of age structure changes in the demographic transition was fertility declines (Coale, 1972) (*see* Fertility Theory). Although the significant mortality decline started before the significant fertility decline in the majority of countries, the age structure of some populations became younger in early phases of mortality decline, as the pronounced reduction of infant and child mortality increased the number of young children. The decline in the

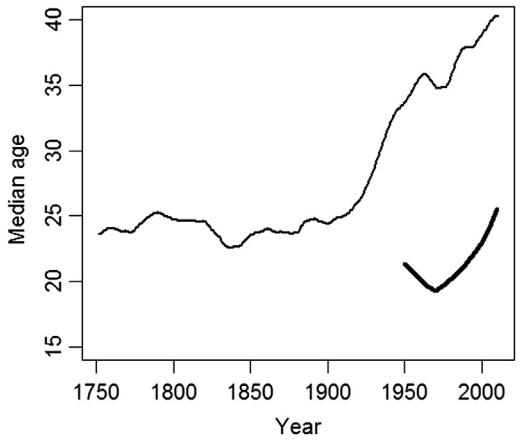

Figure 2 Trends in the median age of population for India, 1950–2010 (thick line) and Sweden, 1751–2011 (thin line).
Source: India: United Nations (2012). World Population Prospects: The 2012 Revision; Sweden: *Human Mortality Database*. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at http://www.mortality.org (data downloaded on 02.04.14.)

median age of India's population in the 1950s and 1960s, seen in Figure 2, may be due to the decrease in youth mortality. However, during the last few decades, reductions of old-age mortality in many developed countries had strong impacts on population aging (Horiuchi, 1991; Preston et al., 1989; Preston and Stokes, 2012; Vaupel and Jeune, 1995).

The age structure may also show significant fluctuations in cohort size, reflecting temporary changes in fertility, mortality, or migration. Events such as famine, drought, epidemic, and war often reduce fertility and disproportionately increase death rates for infants and young children, producing relatively small cohorts who were born during or immediately before the catastrophic period. Wars may also cause considerable attrition of cohorts deeply involved in combat. Figure 1(e) shows the population pyramid for West Germany, 1960, in which the proportions aged 40–44 and 10–14, born during the World Wars I and II, are substantially small. Also the male cohorts aged 35–54 were noticeably smaller than their female counterparts, reflecting gender differences in mortality of young adults during World War II.

Temporary rises in fertility (baby booms) also produce significant cohort size variations. In Figure 1(f), the relatively high proportion of population aged 25–44 in the United States, 1990, is due to the baby boom after World War II. Such cohort size variations may be echoed among later cohorts. For example, Figure 1(f) shows that under 15 years of age, the size of younger cohort is larger, probably because cohort size variations among the baby boom cohorts were repeated by those among their children, though with some diminution.

Measurement of Age Structure

Age structure can be considered as a pattern of C_is, where C_i is the proportion of the population that falls in age-group $i(i = 1, 2, \ldots, n$ and $\sum_{i=1}^{n} C_i = 1)$. With age as a continuous variable x, age distribution can also be represented as the density function $c(x)$ for the proportional distribution of population by age. Then the proportion of population that falls in the age interval between a and b is expressed as $\int_a^b c(x)\mathrm{d}x$, and $\int_0^\infty c(x)\mathrm{d}x = 1$.

Table 1 Estimated summary indicators[a] of age structure for selected countries[b], 2010

	Median age	*C(0–19)*	*C(65+)*	*Age-dependency ratio*
Kenya	18.5	53.1	2.6	125.7
India	25.5	39.9	5.1	81.8
Brazil	29.0	33.9	6.9	68.9
Sweden	40.7	23.4	18.2	71.2
Japan	44.9	18.0	23.0	69.5

[a]C(0–19) is the proportion (in %) of population aged under 20 and C(65+) is the proportion of population aged 20 and over.
[b]Countries are listed in the ascending order of median age.
Source: United Nations (2012). World Population Prospects: The 2012 Revision.

Common summary measures of age structure include the mean age of population, median age of population, proportion of population at young ages (e.g., under age 20), and proportion of population at old ages (e.g., age 65 and above). The age-dependency ratio is the ratio of the population outside the working-age range to the population in the age range. The working-age range is defined as ages between x_1 and x_2 (usually including x_1 but excluding x_2), where x_1 is a typical age at the start of the first full-time employment and x_2 is the typical age of retirement. Since multiples of 5 are preferred, 15 and 20 are widely used for x_1 and 60 or 65 for x_2. The ratio is usually multiplied by 100. It is used as a measure of effects of the age structure on the extent to which people in labor force have to economically support those outside labor force. Also used are the child-dependency ratio, which is the ratio of the population under x_1 to the population between x_1 and x_2, and the old-age dependency ratio, which is the ratio of the population aged x_2 and above to the population between x_1 and x_2 (for a more detailed explanation of measures of age structure, see Hobbs, 2004).

Table 1 shows the median age of a population, proportion under age 20, proportion aged 65 and over, and age-dependency ratio (using 20–64 as the working age) for five selected countries that have different age structures. Kenya has the youngest age distribution among the three developing countries, followed by India and Brazil. The differential age structures of the three countries reflect different timings of their demographic transitions. As for the two developed countries, Sweden started its demographic transition much earlier than Japan, but now the age structure of Japan is older than that of Sweden, because of the steep declines in fertility and mortality in Japan during the last few decades. The age-dependency ratios of Brazil, Sweden, and Japan are fairly close, indicating that populations with different age structures could have similar age-dependency ratios, if the sum of the proportion under the working-age range and proportion above it happen to be close among those countries. Overall, Table 1 suggests that currently there are very large variations in age structure among countries in the world.

The direction and pattern of age structure changes in a population are indicated by the profile of age-specific

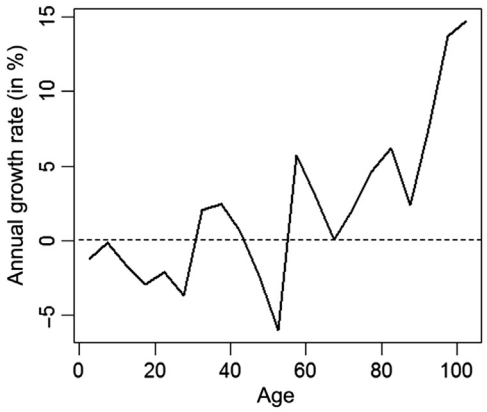

Figure 3 Annual rate of population growth by age for Japan, 2010. Population size estimates by 5-year age-groups were used, and geometric rates of change in percent were computed. The dashed line indicates the growth rate of total population. Source: Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at http://www.mortality.org (data downloaded on 02.04.14.).

growth rates for the population (Horiuchi and Preston, 1988). For example, **Figure 3** shows that although the total population size of Japan remained nearly constant during 2010, the population under age 30 decreased and the population aged 55 and over increased. In particular, the population aged 95 and over grew by about 15% in the year. The proportions of age-groups with higher growth rates than that of the total population increase, and those of age-groups with lower growth rates decrease. Thus, the large age variations of population growth in **Figure 3** suggest that the age structure of the population of Japan was changing fairly rapidly in 2010.

It should be noted that some demographic data on age must be used with caution because self-reported ages in populations with low literacy rates may not be accurate (Hobbs, 2004). Particular attention should be given to ages of old persons and old decedents in historical demographic data collected in those days and data from current developing countries, partly because older cohorts tend to have higher illiteracy rates in many countries, and partly because age inaccuracy tends to produce larger biases in statistics on older adults (Preston et al., 1996).

Age-Standardization

The incidence of various events (such as marriage, childbearing, and death) is strongly associated with age. The crude rate of such an event, usually computed as the ratio of the number of the events in a year to the total midyear population, could be significantly affected by the age structure of the population. Thus, for comparing the rate among populations, it should be adjusted for age structure differences.

For example, the crude death rate (CDR) is the ratio of the number of deaths in a year to the total midyear population, usually multiplied by 1000. CDR is strongly affected by age structure, because it is the sum of age-specific death rates weighted by the proportion of the population in that age-group:

$$\mathrm{CDR} = \frac{D}{N} = \sum_{i=1}^{n} \frac{D_i}{N_i} \cdot \frac{N_i}{N} = \sum_{i=1}^{n} M_i C_i \qquad [1]$$

where D is the total number of deaths in the year, N is the midyear total population, n is the number of age-groups, D_i is the number of deaths in age-group i, N_i is the population of age-group i, M_i is the death rate of age-group i, and C_i is the proportion of total population that belongs to age-group i. Thus, the CDR tends to be high not only in populations with high age-specific death rates but also in populations in which a large proportion of the population belongs to age-groups that have relatively high death rates (e.g., old age-groups).

A widely used method to adjust per capita rates for age structure is age-standardization. The age-standardized death rate (ASDR) for a population is obtained by selecting an actual or hypothetical population as the 'standard' and assuming that the study population has the age structure of the standard population: the ASDR for population j is computed by

$$\mathrm{ASDR}^j = \sum_{i=1}^{n} M_i^j \cdot C_i^s \qquad [2]$$

where M_i^j is the death rate of age-group i in population j and C_i^s is the proportion of total population that belongs to age-group i in the standard population. If the ASDR is calculated for a number of populations using the same standard population, their differences in the ASDR are due to their differences in age-specific death rates only, and completely independent of their differences in age structure. The selection of a different standard does not change ASDR differences significantly unless the age pattern of mortality differs considerably among the study populations. This method is applicable to incidence rates of other age-associated events such as childbearing and migration, as well as proportions of the population with age-related characteristics such as marital status, labor force participation, and morbidity due to a particular disease. (For a more thorough explanation of age-standardization, see Section 2.2 of Preston et al., 2001).

Mathematical Expressions of Age Structure

Relationships between the age structure and demographic rates can be mathematically formulated (Bennett and Horiuchi, 1981; Preston and Coale, 1982) (*see* Population Dynamics: Theory of Nonstable Populations). The density function for population age distribution can be expressed as

$$c(a) = b e^{-\int_0^a r(x)\mathrm{d}x} p(a) e^{\int_0^a \lambda(x)\mathrm{d}x} \qquad [3]$$

where b is the crude birth rate, $r(x)$ is the population growth rate at age x, $p(a)$ is the proportion surviving from birth to age a in the period life table, and $\lambda(x)$ is the rate of net migration (immigration minus emigration) at age x. Note that this is

a relation between the age structure and demographic rates observed in the same period.

It may seem puzzling that the current age structure, which primarily resulted from population dynamics in the *recent past*, can be solely determined by *current* demographic rates. This paradoxical relationship can be understood by recognizing the role of age-specific growth rates in the equation. Age-specific growth rates reflect effects of past demographic changes on the current age structure (Horiuchi and Preston, 1988): the growth rate of population at age a and time t, $r(a, t)$, can be decomposed into the growth rate of the number of births, and cumulated changes in age-specific death rates and age-specific rates of net migration that the cohort aged a at time t has previously experienced:

$$r(a,t) = r_B(t-a) - \int_0^a \frac{\partial \mu(x,u)}{\partial u}\,\mathrm{d}x + \int_0^a \frac{\partial \lambda(x,u)}{\partial u}\,\mathrm{d}x \qquad [4]$$

where $r_B(t)$ is the growth rate of the number of births at time t, $\mu(a, t)$ and $\lambda(a, t)$ are the instantaneous age-specific death rate and instantaneous rate of net migration, respectively, at age a and time t, and the derivatives are assessed at $u = t - a + x$.

Equation [3] is a generalization of the age structure equation in stable populations (Doublin and Lotka, 1925; Lotka, 1998; originally 1939) (*see* Population Dynamics: Theory of Stable Populations, Population Dynamics: Classical Applications of Stable Population Theory). Consider a population in which the growth rate and age structure have been changing up to a certain point of time t. It can be shown that if the population is closed to migration and its age-specific birth rates and age-specific death rates remain unchanged after time t, the population will eventually become a stable population, in which the population growth rate and age structure remain constant over time. Simulations suggest that if the initial age structure of a human population is far from stable, it takes about a century or a little more for the population to become close to stability, but a few centuries to become almost exactly stable.

In a stable population, the density of age distribution at age a is given by

$$c(a) = b\mathrm{e}^{-ra}p(a) \qquad [5]$$

where b is the CDR, r is the growth rate of total population, and $p(a)$ is the proportion surviving from birth to age a, which is determined by age-specific death rates between ages 0 and a. In a stable population, these demographic measures are constant over time.

Equation [3] is used for demographic estimation from incomplete data (e.g., Bennett and Horiuchi, 1981; Preston and Bennett, 1983), eqn [4] is used for analyzing past changes in population growth and age structure (Horiuchi, 1991, 1995), and eqn [5] is used for demographic estimation from data on age distributions of historical populations before the demographic transition and populations in developing countries that are in early stages of demographic transition (United Nations, 1983: Chapter 7).

Age Structure and Population Growth

The age structure has strong effects on the increase or decrease of total population size in the immediate and near future. Even if several populations have the same size and same age schedules of fertility and mortality, a population with a higher proportion of population at peak childbearing ages (usually late 20s) is likely to have a greater number of births in the next several years, and a population with a higher proportion of children tends to have a greater number of births a few decades later when those children reach peak childbearing ages.

An indicator of the extent to which the age structure is conducive to population growth is the ratio of the size of stable equivalent (Keyfitz, 1969), denoted by Q, to the actual population size, N. The stable equivalent is a hypothetical quantity computed assuming that the age schedules of fertility and mortality at t_1 remain unchanged in the future and the population is closed to migration after t_1. Then the population would converge to stability at some time in the future, denoted here by t_2. The stable equivalent at t_1 is given by

$$Q(t_1) = N(t_2)\cdot \mathrm{e}^{-r(t_2-t_1)} \qquad [6]$$

where $N(t_2)$ is the size of the population at t_2. If the ratio of stable equivalent to actual population size at t_1, $Q(t_1)/N(t_1)$, is greater than 1, the population size would increase between t_1 and t_2 at the rates that are higher than the stable population growth rate r, due to the positive effects of age structure on population growth.

Another indicator is the population momentum, which is based on the hypothetical scenario that the fertility at t_1 instantly changes to the replacement level and then the age schedules of fertility and mortality remain unchanged, and there is no in- or out-migration (*see* Population Dynamics: Momentum of Population Growth). The population would eventually become a stationary population, which is a stable population the size of which remains constant ($r = 0$) over time. The population momentum is usually measured as the ratio of the size of the stationary population to the size of population immediately before the start of the replacement-level fertility (Keyfitz, 1971). The ratio may substantially deviate from 1, which indicates that even with the replacement-level fertility, the population may continue to significantly increase or decrease for some length of time, because the population increase or decrease in the near future has already been built in the age structure.

Acknowledgment

I am grateful to Jesica Rodriguez-Lopez for producing graphics for this article.

See also: Age Policy; Age Stratification; Age, Sociology of; Fertility Theory; Life Course: Sociological Aspects; Period and Cohort Analysis in Demography; Population Aging: Economic and Social Consequences; Population Dynamics: Momentum of Population Growth; Population Dynamics: Theory of Nonstable Populations; Second Demographic Transition.

Bibliography

Bennett, N., Horiuchi, S., 1981. Estimating the completeness of death registration in a closed population. Population Index 47, 207–221.

Coale, A.J., 1972. The Growth and Structure of Human Populations. Princeton University Press, Princeton.

Doublin, L.I., Lotka, A.J., 1925. On the true rate of natural increase. Journal of the American Statistical Association 20, 305–339.

Hobbs, F., 2004. Age and sex composition. In: Siegel, J.S., Swanson, D.A. (Eds.), The Methods and Materials of Demography. Elsevier Academic Press, San Diego and London, pp. 125–173.

Horiuchi, S., 1991. Assessing the effects of mortality reduction on population ageing. Population Bulletin of the United Nations 31/32, 38–51.

Horiuchi, S., 1995. The cohort approach to population growth: a retrospective decomposition of growth rates for Sweden. Population Studies 49, 147–163.

Horiuchi, S., Preston, S.H., 1988. Age-specific growth rates: the legacy of past population dynamics. Demography 25, 429–441.

Keyfitz, N., 1969. Age distribution and the stable equivalent. Demography 6, 261–269.

Keyfitz, N., 1971. On the momentum of population growth. Demography 8, 71–80.

Lotka, A.J., 1998. Analytical Theory of Biological Populations. Plenum Press, New York. A.J. Lotka, Trans. (1939) Théorie Analytique des Associations Biologiques. Duuxième Partie: Analyse Dèmographique avec Application Particulière è l'Espèce Humaine. Hermann et Cie, Paris.

Preston, S.H., Bennett, N.G., 1983. A census-based method for estimating adult mortality. Population Studies 37, 91–104.

Preston, S.H., Coale, A.J., 1982. Age structure, growth, attrition and accession: a new synthesis. Population Index 48, 217–259.

Preston, S.H., Elo, I.T., Rosenwaike, I., Hill, M., 1996. African-American mortality at older ages: results of a matching study. Demography 33, 193–209.

Preston, S.H., Heuveline, P., Guillot, M., 2001. Demography: Measuring and Modeling Population Processes. Blackwell Publishing, Oxford and Malden.

Preston, S.H., Stokes, A., 2012. Sources of population aging in more and less developed countries. Population and Development Review 38, 221–236.

Preston, S.H., Himes, C., Eggers, M., 1989. Demographic conditions responsible for population aging. Demography 26, 691–704.

United Nations, 1983. Manual X: Indirect Techniques for Demographic Estimation. United Nations, New York (Sales No. E.83.XIII.2).

Vaupel, J.W., Jeune, B., 1995. The emergence and proliferation of centenarians. In: Jeune, B., Vaupel, J.W. (Eds.), Exceptional Longevity: From Prehistory to the Present. Odense University Press, Odense, pp. 109–116.

Agency Theory

Stefan Linder, ESSEC Business School, Paris, France
Nicolai J Foss, Copenhagen Business School, Frederiksberg, Denmark; and Norwegian School of Economics and Business Administration, Bergen, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Agency theory studies the problems and solutions linked to delegation of tasks from principals to agents in the context of conflicting interests between the parties. Beginning from clear assumptions about rationality, contracting, and informational conditions, the theory addresses problems of ex ante ('hidden characteristics') as well as ex post information asymmetry ('hidden action'), and examines conditions under which various kinds of incentive instruments and monitoring arrangements can be deployed to minimize the welfare loss. Its clear predictions and broad applicability have allowed agency theory to enjoy considerable scientific impact on social science; however, it has also attracted considerable criticism.

Foundations

Delegation and Conflict of Interest

A key tenet of economics is that specialization is productive. On the individual level, it is indeed often beneficial not to engage in a particular task oneself, but to delegate it to another person specializing in the task. Examples are counseling on legal matters, managing another person's funds, diagnosing and advising on illnesses, and precision-grinding a cylinder or machine tool. Delegating such tasks may be beneficial due to lack of time or knowledge of how to best perform the task (Laffont and Martimort, 2001). Yet, truly reaping the benefits of delegating a task to another person (an 'agent') is not always easy in practice. In *The Wealth of Nations*, Adam Smith, while praising the benefits of the division of labor and specialization, also provides what may be the first written account of the problems of delegation that the division of labor gives rise to. Referring to the directors of stock companies who as agents of the owners specialize in the day-to-day handling of the business, he (Smith, 1776: 700) observes that

> being the managers rather of other people's money than of their own, it cannot well be expected, that they should watch over it with the same anxious vigilance with which the partners in a private copartnery frequently watch over their own. … Negligence and profusion, therefore, must always prevail, more or less, in the management of the affairs of such a company.

Thus, Smith describes how different and even conflicting interests may lead to efficiency losses ('negligence and profusion') in the context of a situation of delegation in which principals cannot costlessly enforce the actions they prefer their agents to take, that is, where 'agency problems' exist. However, with few exceptions (particularly the debates of the 1930s on the economics of socialism and the separation of ownership and control in the modern corporation), awareness of agency problems in economics was virtually nonexistent until the end of the 1960. At that time, fundamental advances in economic analysis (economics of uncertainty and information, mechanism design, property rights) paved the way for a rigorous and systematic approach to these problems. Agency theory (or, 'principal–agent theory' or sometimes just 'incentive theory') is an integral part of contemporary microeconomics.

The canonical set-up of the theory is this. Two self-interested individual recognize potential gains from a trade, in which one (i.e., the principal) delegates a physical or mental task to the second (i.e., the agent), whose choice of actions and/or effort level affect the payoffs of both parties (Grossman and Hart, 1983; Jensen and Meckling, 1976; Ross, 1973). The agent is paid by the principal, who assumes the role of residual claimant in the relationship, that is, he holds the claims to net cash flows that result from differences between inflows and promised payments to other claim holders (Jensen and Smith, 1985). The agent and the principal hold different interests. For example, top managers acting as agents of the firms' shareholders may prefer 'empire building,' perks, leisure time, and so on, instead of maximizing shareholders' returns ('conflict in outcome-type preferences') (Holmström, 1979; Jensen and Smith, 1985). Likewise, managers may prefer engaging in capital expenditures that maximize the survival chances of the firm that pays their salaries, while shareholders will likely prefer them to go for high returns, but also typically more risky investments as they can diversify risk ('conflict in risk preferences') (e.g., Grossman and Hart, 1983). Or, yet again, managers may not plan to stay on for a long time with the firm in question and thus differ in the time horizon they consider in decision-making from (assumingly long-term oriented) shareholders ('conflict in time horizon') (e.g., Baker et al., 1994). Many other forms of conflicts of interests between the principal(s) and the agent(s) are obviously conceivable. All of these imply that the agent (or agents) may not act in the principal's best interest. Yet, these conflicting interests only become problematic when combined with information being asymmetrically distributed between the principal(s) and the agent(s).

Asymmetric Information, Hidden Characteristics, and Hidden Action

The delegation of a task is often motivated by the principal lacking knowledge, abilities, or skill to perform the task

 http://dx.doi.org/10.1016/B978-0-08-097086-8.73038-8

themselves, knowledge differences that explain why gains from trade exist in the first place. However, it also implies that the principal faces a disadvantage with respect to judging the agent's true knowledge and/or efforts in carrying out the delegated task. For example, patients typically cannot fully ascertain the physician's knowledge, abilities, and skills. They may rely on the academic degrees the agents hold, their reputation, or the like; however, only the agents know the true state of their knowledge, abilities, and skills. The situation is complicated by self-interested agents potentially deliberately hiding their true qualities (e.g., a lack of appropriate skills for conducting the task at hand) or even falsifying signals (e.g., faking their CV or the degrees they hold) in order to get a job and to earn the related rents. Therefore, when informational asymmetry is present already before the principal hires an agent this gives rise to the so-called 'hidden characteristics' problem. It is one of the core problems in delegation under conflicting interests and information asymmetry between the parties studied by agency theory.

Even if the principal is perfectly informed about the characteristics of the agent(s), she may, nevertheless, face an information disadvantage with respect to the actions taken by the agent(s). More precisely, the disadvantage (i.e., informational asymmetry) concerns which action (or effort level) the agent actually took, whether it was the right one given the circumstances, and exactly which circumstances pertained (Grossman and Hart, 1983; Holmström, 1979). For example, the sales of a salesperson on a local market are influenced by stochastic influences, such as changes to local demand, that may be hard or simply too costly to ascertain by the principal (the firm who employs the salesperson). Of course, agents are likely to hold superior knowledge concerning the true actions or efforts they exert in carrying out the task. Agents may exploit this information asymmetry by engaging in actions not valued by the principal or withholding effort or other forms of 'hidden action.' This is often captured by the notion that the agent experiences 'disutility of labor,' which is somewhat misleading as it implies that agents are lazy. They may be, but the more general interpretation is that agents prefer dedicating effort to activities they choose themselves.

The upshot of the above is that the principal may face informational disadvantages either ex ante (i.e., 'adverse selection,' 'hidden information'), that is, in the contracting stage of the relationship, or ex post ('moral hazard,' 'hidden action') when the agent may carry out the delegated task in a manner or with an intensity diverging from what would maximize the principal's payoffs. Further complicating is the fact that both problems may appear within the same relationship. For example, when looking for legal advice from a lawyer, the client may neither know enough to fully judge alternative lawyers' abilities, knowledge, and skill nor may she/he be able to observe and properly judge the efforts made by the lawyer chosen. Under conflicting interests and when the principal is either not knowledgeable of the task she wishes to delegate and/or cannot observe the agent's characteristics and actions, the principal faces problems of hidden characteristics or hidden action (or both). Hence, both information asymmetry and conflict of interest are necessary ingredients for a 'principal–agent problem' to exist.

Core Insights and Predictions of the Agency Model

The Linear Agency Model

The workhorse model of agency theory, the 'linear model' (Holmström and Milgrom, 1991; see also Holmström, 1979, for a seminal contribution that is also echoed in the following) studies a two-party setting: one principal and one agent. Both are assumed to be self-interested, behave rationally, and maximize their utility. Due to information asymmetry ex post, the principal is unable to contract over the agent's actions (which the principal assumingly cannot observe), but only over the output z of the agent's actions. (In agency theory, all observable variables are contractible.) For convenience, most agency theorists think of the agent's actions as his effort e, that is, alternative actions correspond to different effort levels (e.g., Holmström, 1979) and the terms effort and action are thus used interchangeable in much of the literature. The result z in turn is assumed to depend linearly on the sum of the agent's choice of effort e and some stochastic influence from the environment θ. For simplicity, the basic model assumes this external influence to be standard-normally distributed with an expectation of zero. Both, the distribution and the expectation, are known by both parties, whereas the true effort e exerted by the agent and the actual state of the environment θ are only known to the agent; the principal can only observe the result z caused by both e and θ. The principal is thus assumed to be unable to ascertain whether a good result z is due to the agent's high-effort level e or due to 'luck' (θ); hence, she/he faces a problem of hidden action/moral hazard.

The interaction between the principal and the agent starts with the principal contracting with the agent about a wage W, which is comprised of some proportion β of the result z generated by the agent (potentially plus some fixed salary component W_0). The share of the result z that the principal agrees to pay the agent (i.e., the β) is referred to as 'incentive intensity.' As the true effort e cannot be observed, the agent's wage W must be based on the result z obtained (which is influenced by the stochastic environment). Therefore, any β larger than zero implies that the agent is exposed to risk with respect to his/her remuneration. And here comes another assumption of the (basic) agency model into play: The risk aversion r_A of the agent. While the principal is assumed to be risk neutral (she is able to diversify risk, e.g., by holding a market portfolio of relationships), the agent is assumed to be risk averse as his payoffs are linked to the particular relationship (e.g., Grossman and Hart, 1983). Risk aversion implies that the agent will ask for a premium in order to accept a contract offered by the principal based on the risky result z.

This risk premium reduces the surplus, but does not give the agent extra utility. Thus, the larger the risk premium, the smaller the total value created in the relation (i.e., the sum of the monetary equivalents of the utilities of the principal and the agent). The size of the risk premium depends on both the degree of risk aversion of the agent r_A and the riskiness of the wage (i.e., the variance in the wage due to the environment's impact ($\mathrm{Var}(W) = \beta^2\,\mathrm{Var}(\theta)$)). Thus, β has to solve two tasks: providing incentives to the agent to work hard and insurance against risk (note: literature unfortunately also uses the term 'uncertainty' to denote the stochastic nature of the agent's payoffs, which we do not follow here as it clutters the

differences between randomness and true uncertainty in the sense of Frank Knight, that is, outcomes for which no distribution and expected value are known). It is, therefore, usually impossible under information asymmetry to handle both tasks optimally (Holmström, 1979). Therefore, any solution is inefficient ('second-best') as compared to a hypothetical first-best solution in which the principal would bear all the risk and the agent would be paid on his/her efforts only.

In addition to the (risky) payoffs associated with the relation, agents are expected to consider their costs (Grossman and Hart, 1983). Agents are assumed to have exponentially growing disutility of engaging in effort ($C(e)$). Principals in turn need to take this into account when solving their own optimization problem concerning the optimal β. As Holmström and Milgrom (1991) show, this can be done neatly by starting out from the certainty equivalents of both parties. A certainty equivalent is the certain value that makes an individual indifferent in terms of utility between obtaining the 'risk-free' certain value and a risky 'lottery' value. For the agent, the certainty equivalent corresponds to the wage minus the risk premium minus the costs (i.e., $CE_A = W_0 + \beta \bullet e - ½ \bullet r_A \bullet \beta^2 \bullet \mathrm{Var}(\theta) - C(e)$). Taking the first derivative of this term to e and then setting the expression to zero and solving for β yields the helpful term of $\beta = C'(e)$, that is, the agent's optimal choice of e is where marginal revenue equals marginal cost. The optimizing behavior of the agent implies that his efforts e are a function of β. Based on this insight only a few mathematical optimization steps are now necessary to solve the principal's problem, that is, to get from the task to maximize the total certainty equivalent for both parties (max $CE_{\mathrm{Total}} = P(e(\beta)) - C(e(\beta)) - ½ \bullet r_A \bullet \beta^2 \bullet \mathrm{Var}(\theta)$) to the optimal incentive intensity (denoted as β^*) (for the details of the mathematical derivation see e.g., Milgrom and Roberts, 1992: pp. 222n; and for the original formulation, Holmström and Milgrom, 1991).

The resulting expression $\beta^* = P'(e)/[1 + r_A \bullet \mathrm{Var}(\theta) \bullet C''(e)]$ suggest that the optimal β depends on four factors: (1) The more sensitive the payoff to the principal on changes in the agent's efforts or actions ($P'(e)$), the larger should be β^*; (2) the stronger the agent's risk aversion r_A, the lower the β^*; (3) the larger the riskiness of the results obtained by the agent due to environmental influences $\mathrm{Var}(\theta)$, the lower the incentive intensity; and (4) the stronger the agent's discretion regarding the choice of activities ($\partial e/\partial \beta = 1/C''(e)$), that is, the lower his/her costs of acting in a manner functional for maximizing the principal's payoffs ($C''(e)$), the larger the β^*.

Core Insights

The set of insights what determines β^* has become known in literature under the label of the 'incentive intensity principle.' It is one of the core contributions of agency theory to the understanding of how principals can set incentives and design-monitoring schemes in order to minimize the welfare loss ('agency costs') resulting from conflict of interest and information asymmetry ex post. In short, β^* is determined by incentive elasticity of profits (e.g., there may be tasks where strengthening incentives does not lead to more effort being supplied); the risk tolerance of the agents (e.g., agents that are high in risk aversion will demand a high risk premium); the effort elasticity of incentives (e.g., some agents may have work/leisure tradeoffs such that they are not easily incentivized); and the measurability of outputs (if outputs are very costly to measure, it may be counterproductive to tie pay to outputs).

A key application of the linear model concerns when to use additional information γ in the incentive design, such as, for example, the sales performance of other sales representatives when incentivizing a particular salesperson or the performance of similar firms in the industry when judging top-management's efforts, and so on. The extended model demonstrates that if the additional information gathered is not 'informative' about the focal agent's true effort, that is, if the additional information γ used has a covariance of zero with the state of the environment θ, then including the additional information in the wage contract with the agent is unreasonable. Including it under such conditions only adds noise, which makes drawing conclusions about the agent's true effort level or actions even harder (Milgrom and Roberts, 1992). Therefore, the 'informativeness principle' calls for including additional available information only if it allows reducing noise.

The 'equal compensation principle' stems from yet another extension of the basic model. This extension renders the model more realistic by allowing the focal agent being responsible not only for a single task or single-dimensional task, but for multiple tasks or multidimensional tasks, with performance on some tasks or dimensions easily measureable and others only difficult or prohibitively expensive to assess. Holmström and Milgrom (1991) show that, under such conditions, agents will shift their effort to those tasks/dimensions that are measured (and rewarded) by the principal(s) at the detriment of other tasks (or other dimensions of the task) that are too expensive to assess. For example, the agent may focus on her/his results looking good in terms of the financial performance achieved, but neglect other nonfinancial and (may be only qualitative) aspects, such as performance with respect to customer satisfaction, environmental or social dimensions of a task, and so on. The extended model thus suggests two consequences: (1) grouping easy-to-measure tasks together instead of spreading them over multiple individuals, and to provide a low β to those individuals engaging in the tasks with the hard or impossible to measure dimensions/tasks, and (2) choosing only a low β if the task dimensions cannot be separated organizationally on multiple individuals and agents thus have to carry out multidimensional tasks with some dimensions costly to measure but essential to the principal.

Finally, the 'monitoring intensity principle' highlights the complementarities of incentive provision and monitoring in addressing hidden-action problems. While the basic model treats the variance with which efforts are measured as outside the principal's control, it is in many cases more realistic to assume the principal to be able to improve measurement by spending resources on, for example, more detailed data collection, a lower span of control of supervisors allowing them to closely monitor individual employees, and so on. While all of these actions are costly to the principal, they improve the principal's information about how the agent performs. By including the costs for measurement in the optimization problem, the extended model suggests that more resources should be spent on measurement when the incentive intensity is higher. Hence, high incentive intensity and careful

performance measurement are complementary (Milgrom and Roberts, 1992).

Besides these exemplary extensions, agency theorists have developed a great number of other refinements of the model that relax certain assumptions and render the model more realistic. For example, scholars have refined the model to allow for multiperiod settings, risk-averse principals, risk-neutral agents, nonlinear production functions, subjective versus objective performance measures, or to expand the model to comprise more than two parties – that is, multiple principals and/or agents (e.g., Baker et al., 1994; Milgrom and Roberts, 1992).

The basic agency model (and many of its extensions) provides insights for handling both ex post as well as ex ante information asymmetry. For example, it suggests that principals offering agents a contract with a $\beta > 0$ are likely not only to reduce problems of hidden action, but can also already ex ante trigger that agents with lower abilities and skills prefer not to accept such a performance-contingent contract and to, for example, look for other employment opportunities where they are given a fixed salary. Thus, firms switching from paying their employees a flat wage can be expected to benefit from two effects: First, an increase in the effort level exerted by the employees already employed by the firm ('incentive effect') and second, a beneficial change in the pool of new applicants for positions with the firm (i.e., the 'sorting effect' of variable incentive schemes).

Principals may also rely on increased information gathering already ex ante about potential agents in the contracting stage to reduce information asymmetry. Job interviews are one example of such information gathering or screening activities ex ante that aim at reducing the hidden-characteristics problem. Agents themselves may take the initiative by signaling their true ('good') characteristics to the principal(s), that is, reducing information asymmetry between them and the principal(s) through such instruments like, for example, providing warrantees for used cars that they wish to sell in the example of car dealers, or investing into acquiring academic degrees in case of agents seeking a job.

Exemplary Applications and Contributions

Conflict of interest and asymmetric information are widespread phenomena in cooperative activities among individuals (Holmström, 1979; Jensen and Meckling, 1976; Ross, 1973). Thus, it is not surprising to see agency theory being applied to the study of a wide range of relationships within labor market (i.e., labor market economics) and business contexts (i.e., personnel economics), as well as in sociology or political science research. Discussing the many applications of agency theory is beyond the scope of this article; in fact, merely scratching the surface of this huge body of literature is a daunting challenge.

By looking at relationships between two or more cooperating individuals in terms of a 'nexus' of explicit or implicit contracts (e.g., Alchian and Demsetz, 1972; Jensen and Meckling, 1976), agency theorists were among the first scholars who opened up the 'black box' of firms and organizations, and contributed the explaining rationales and workings of other kinds of economic organization (i.e., various kinds of nonfirm contracting arrangements). An early influential stream (Alchian and Demsetz, 1972) conceptualized firms as efficient responses to agency problems caused by 'team production' in which individual productivity is costly to observe, but the team's output is not. In this situation, it makes sense to appoint a monitor who holds residual income rights to the team's output, because this means that he will undertake an efficient level of monitoring. Thus, the monitor is also the owner of the firm and holds right to discipline team members – a rudimentary theory of why there are firms in a market economy (see Holmström, 1982, for important refinements and extensions).

These contributions represent marked scientific progress over the previously dominant perspective of firms as production functions with their own preferences and decision-making presumably corresponding to a single individual. Agency theory has profoundly changed how economics and business research think about firms and organizations in general in fields such as corporate governance, human resource management (in particular: remuneration), strategy implementation, performance measurement, organizational control, accounting as well as phenomena affecting intersections of multiple of these fields.

Agency theory (and derivations thereof) sees application to a growing number of research fields within political science and sociology (see Kiser, 1999, for a detailed review). Scholars in political science have applied (often more or less adapted versions) of agency theory to a wide array of topics, with particular emphasis on studying legislatures and policy implementation. In an early application, Rose-Ackerman (1975), for example, uses agency theory to study corrupt dealings in the government contracting process. Others have focused on how 'red tape' can serve a monitoring function to achieve that policy decisions by unelected bureaucratic officials are responsive to the preferences of citizens. Adams (1996) in turn provides an example of the theory's use within sociology to study the relationship between patrimonial states and colonial trading companies in the Netherlands and England.

Criticisms and Limitations

Agency theory has attracted considerable criticism from various authors and scholarly fields. In particular, numerous authors have criticized the assumptions underlying the standard agency model as too restrictive, that is, as not being generalizable to the overwhelming part of humans, but as rather being particular to just a subset of individuals.

Given and State Independent Utility/Risk Functions

Empirical and experimental research suggests that the von Neumann and Morgenstern utility functions underlying agency theory are likely not to be as generally applicable, as was originally hoped. For example, Prospect Theory advanced by Daniel Kahneman and Amos Tversky in response to empirical evidence questioning the von Neumann and Morgenstern utility theory suggests that individuals will consider their current wealth when evaluating how to act, implying that the same individual may sometimes be risk averse, risk neutral, or even risk seeking depending on the state of his/her personal

wealth. So far, these insights have not yet been incorporated into agency theory.

Complete Contracting

Agency theory assumes complete rationality of the parties involved, implying that the contracts between the principal(s) and the agent(s) will contain all available information and that the terms stipulated in the contract will consider all possible future situations (the so-called 'complete contracts'). Hence, while principal(s) and agent(s) are assumed to differ in terms of the information about the characteristics of the agent(s) or their actions, the theory assumes that each of the parties makes full use of the information available to it in designing the contract and in deciding on how to act, respectively. Agency theory thus – as opposed to transaction cost economics – abstracts from potential costs for incorporating information into the contracts as well as from the possibility that contracts may be incomplete simply due to insufficient knowledge about all potential future situations or actions possible. While Ross (1973: 135) submits that under the assumption of complete contracts "the problem is considerably simplified but much of the interest does remain," agency theory's underlying assumption thus, nevertheless, hampers its usefulness for the study of a number of highly important real-world phenomena. Entrepreneurship, for example, while undoubtedly a core economic phenomenon, is hard to study when allowing only for risk – that is, probabilistically known future states that hence can be included in a complete contract, but not uncertainty in the sense of Frank Knight, which defies standard probabilistic accounts, but which is part of the very nature of entrepreneurial activities. Hence, such fundamental questions of what the principal should want the agent to do and how the principal actually became the principal in the first place have been rather sidelined in agency theory.

Self-Interest/Opportunism

Some authors have pointed out that agency theory may paint a too dark picture of human nature by assuming that individuals behave opportunistically. Perrow (1986) even went as far as claiming that agency theory is 'not only wrong but dangerous' (p. 11) – a theme echoed by a host of writers after the onset of the current financial crisis claiming that agency theory reflects a misguided libertarian ideology and would prompt the adoption of an overly cynical view of human nature. They warn that when this view is generalized, for example, through the teachings of economists in universities and business schools, there is a risk that important prosocial behaviors are crowded out or that a self-fulfilling prophecy might result from assuming individuals to be opportunistic. Moreover, some scholars have pointed out that the theory does not devote sufficient attention to the potential consequences of the principal's opportunism, in the sense that he may try to cheat on the agent in the performance evaluation or reward. However, given that the theory assumes complete contracting, i.e., that the agent might appeal to a third party – e.g., a court – for enforcement of the contract, this criticism seems rather misguided.

Neglect of Intrinsic Motivation

The phenomenon of a 'crowding out' of task-autonomous motivation (often called 'intrinsic motivation') largely associated with the pioneering works by Edward L. Deci and Richard M. Ryan has attracted massive interest among scholars in psychology, education research, and lately also in economics, and management research. While it seems rather well established within education research that administrating incentives that are perceived as controlling undermines the preexisting task-related autonomous motivation of the agent to engage in the task – for example, due to enjoyment of carrying out the task or due to fully internalized norms and values of the agent – the jury is still out when it comes to the phenomenon's existence and magnitude in common business settings. Several studies in economics (e.g., Lazear, 2000) reported increases in employee's engagement in routine or mundane activities – such as, for example, the installation of auto glass – when providing performance-contingent incentives for these activities. However, the existing evidence on motivation crowding suggests that it is likely that the phenomenon pertains to nonmundane tasks in business situations. Existing agency models thus are likely to require adaptation to take the interaction of incentives (and monitoring) with task-autonomous motivation of the agent(s) for nonmundane tasks explicitly into account. This might lead to more nuanced versions of the incentive and monitoring intensity principles.

Dominance of Linear Models with Small Number of Players in Agency Models

The common linear models used in much of agency theory provide great tractability. Yet, linear contracts may not be optimal contracts and the generality of the findings may thus be limited. Similarly, limiting the models to only a small number of principals and/or agents is necessary to keep model complexity to a manageable level. However, real-world social behavior, in particular in groups of 'agents,' is likely to be much more complex than suggested by models focusing only on a small number of players.

Practical Usefulness of Agency Insights

A number of scholars have questioned the practical applicability of (some of) the insights obtained from agency models. First, in practice, managers and HR professionals often lack the information assumed available in principal–agent models, such as, for example, the marginal effect on the principal's payoffs of an increase in the agent's effort level by one unit, risk tolerance, or the agent's responsiveness to incentives. Measurement of these variables, in practice, is still a thorny issue, reducing many of the highly-interesting theoretical insights generated by agency theory to mere rough guidelines from a practitioner's perspective. Second, contrary to what agency theory typically assumes, employment contracts and many other contracts are subject to external restrictions beyond the principal's control, such as labor laws, generating outcomes different from those analyzed in the theory. Third, contracting may, in practice, not necessarily correspond to the situation assumed in agency theory with the principal proposing

contracts on a take-it-or-leave-it basis. Rather, some bargaining between the principal(s) and the agent(s) will take place. Thus, it may not be wise for principals, in practice, to propose to their agent(s) the optimal contract right away, but rather to propose one that will – considering the bargaining to take place – lead to the optimal contract (or one that is close enough). As agency theory does not provide guidance here, insights from bargaining theory may be necessary and valuable complements to practitioners.

In Defense of Agency Theory

While agency theory's assumptions have attracted considerable criticism from many sides, they do possess a particular strength: their explicit nature. This allows systematically relaxing them, that is, adapting them to assumptions deemed more 'realistic.' Thus, while the assumptions underlying the basic agency model are likely not representative of humans in general, the theory provides for a framework that allows modeling a large number of more- or less-diverging assumptions – and testing the need for, benefits of, and shortcomings of alternative governance modes, as well as incentive and monitoring schemes under these assumptions. The explicit nature of agency theory's assumptions thus is an asset facilitating cumulative knowledge growth and continuous refinement of the recommendations developed by scholars.

The use of the theory itself outside economics and business administration has triggered criticism, sometimes sparked outcries of 'economics imperialism.' Yet, it is important to note that agency theory, while first gaining popularity in economics, is conceptually closely linked in many ways to earlier work in sociology, such as Max Weber's works. Moreover, agency theory has not only strongly influenced research in political science and sociology, but has seen itself being adapted and transformed to better fit their different disciplinary contexts (Kiser, 1999). This holds particularly true for the variant of agency theory typically employed in sociology, which can be seen as an amalgam of Weberian sociological insights and the economic agency model, which implies a much broader conception of both the micro- and the macrolevels (see Kiser, 1999, for an excellent discussion of this topic).

Agency theory's popularity among scholars and, in particular, its seemingly simple 'lesson' that agency problems can be largely cured by relying more on performance-contingent incentives instead of fixed salaries, have contributed significantly to the marked trend toward merit-based payment and promotion schemes in recent years. The 'lesson' that one needs to 'pay for performance' in order to obtain superior results made it into MBA curricula, consultants' recommendations, and ultimately into management practice at most larger firms in North America and Europe. Hence, agency theory can be seen to have had a tremendous practical impact – which, as mentioned, has spawned criticism of the theory. However, a closer look at the scientific debate about agency problems and the ways to address them suggests that the incentive practices used are at best a bad copy of what agency theory recommends doing. Thus, already the basic linear model shows that strong variable incentives (a high β) are by no means recommendable under all circumstances. And the extensions of the basic model introduced into the literature over the course of the past 30 years further strengthen the conditionality under which 'high-powered' incentives are optimal – and when they promise to fail (e.g., Baker et al., 1994; Holmström and Milgrom, 1991). Surprisingly, however, these insights seem to have had little impact on compensation practices in banking and the public debate about them. Overall, thus agency theory, while having inspired practice, has unfortunately not seen a more widespread application of one of its most important insights: the conditions necessary for using high-powered incentives.

Conclusion and Outlook

Agency theory studies the problems and solutions linked to delegation of tasks under information asymmetry and conflicting interests between two or more parties. It assumes rationality and self-interest of the parties involved and deals with both problems of ex ante ('hidden characteristics') as well as ex post information asymmetry ('hidden action'). Agency models provide a number of very important recommendations for designing contracts, such as the incentive intensity and the monitoring intensity principles. The theory's broad applicability (Holmström, 1979; Jensen and Meckling, 1976; Ross, 1973) allows agency theory to enjoy tremendous scientific impact both within economics and management research as well as beyond. At the same time, it also has attracted considerable criticism. Most of this criticism focuses on the assumptions underlying agency theory, and in particular, those underlying simple models. These assumptions are often very restrictive to foster tractability of the problems in mathematical terms. Some of the polemical criticisms of it are however misguided. Recent years have witnessed considerable effort in economics and management research addressing some of the theory's major limitations, and some of the criticisms do not apply to the 'derivatives' of economic agency theory found in sociology and political science (Kiser, 1999).

See also: Boundaries and New Organization Forms; Corporate Governance; Hierarchies and Markets; Management: General; Pay, Compensation, and Performance, Psychology of; People in Organizations; Personality and Risk Taking; Personnel Selection, Psychology of; Rational Choice Explanation: Philosophical Aspects; Rational Choice Theory in Sociology; Rational Choice and Organization Theory; Self-Determination Theory; Stockholders' Ownership and Control; Transaction Costs and Property Rights; Uncertainty: History of the Concept; Work Motivation.

Bibliography

Adams, J., February 1996. Principals and agents, colonialists and company men: the decay of colonial control in the Dutch East Indies. American Sociological Review 61, 12–28.

Alchian, A., Demsetz, H., December 1972. Production, information costs, and economic organization. American Economic Review 62, 777–795.

Baker, G., Gibbons, G.R., Murphy, K.J., 1994. Subjective performance measures in optimal incentive contracts. The Quarterly Journal of Economics 109, 1125–1156.

Grossman, S.J., Hart, O.D., 1983. An analysis of the principal-agent problem. Econometrica 51 (1), 7–45.

Holmström, B., 1979. Moral hazard and observability. The Bell Journal of Economics 10 (1), 74–91.
Holmström, B., 1982. Moral hazard in teams. The Bell Journal of Economics 13 (2), 324–340.
Holmström, B., Milgrom, R., 1991. Multitask principal agent analyses: incentive contracts, asset ownership, and job design. Journal of Law, Economics & Organization 7, 24–52.
Jensen, M., Meckling, W., October 1976. Theory of the firm: managerial behavior, agency costs, and capital structure. Journal of Financial Economics 3, 305–360.
Jensen, M.C., Smith, C.W., 1985. Stockholder, manager, and creditor interests: applications of agency theory. In: Altman, E.I., Subrahmanyam, M.G. (Eds.), Recent Advances in Corporate Finance. Dow-Jones Irwin, Homewood, IL, pp. 93–131.
Kiser, E., 1999. Comparing varieties of agency theory in economics, political science, and sociology: an illustration from State Policy Implementation. Sociological Theory 17 (2), 146–170.
Lazear, E.P., 2000. Performance pay and productivity. American Economic Review 90, 1346–1361.
Laffont, J.-J., Martimort, D., 2001. Theory of Incentives I: The Principal-Agent Model. Princeton University Press, Princeton.
Milgrom, P., Roberts, J., 1992. Economics, Organization & Management. Prentice Hall, Englewood Cliffs.
Perrow, C., 1986. Economic theories of organization. Theory & Society 15, 11–45.
Rose-Ackerman, S., 1975. The economics of corruption. Journal of Public Economics 4 (2), 187–203.
Ross, S.A., 1973. The economic theory of agency: the principal's problem. American Economic Review 63 (2), 134–139.
Smith, A., 1776. An Inquiry into the Nature and Causes of the Wealth of Nations. Strahan and Cadell, London.

Agenda Setting, Media Effects on

Maxwell McCombs, University of Texas at Austin, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Agenda setting explicates Walter Lippmann's thesis in *Public Opinion* that the news media are a major link between the world outside and the pictures in our heads. In the decades since the seminal Chapel Hill study, evidence of this media influence has been found worldwide for public issues, perceptions of public figures, and other topics. This evidence also explicates the underlying psychology of these agenda-setting effects and the consequences of these media effects for attitudes, opinions, and behavior.

The news media have a major influence on the focus of public attention, the specific issues that members of the public regard as most priority at any moment. Beyond this influence, the news media also direct our attention to specific aspects of these issues. This combined influence of the news media on the public's attention to and learning the key details of the major issues of the day is referred to as the agenda-setting role of the news media.

The intellectual father of agenda-setting theory is Walter Lippmann, whose book, *Public Opinion* (1922), began with a chapter titled 'The world outside and the pictures in our heads.' The central thesis of the book is that the news media are the primary bridge between the vast array of events in the external world and the truncated views of these events in our minds. As Lippmann noted, our opinions and behavior are a response to these pictures in our heads, not the larger outside world.

Nearly half a century later, during the 1968 US presidential election, Lippmann's observations were the basis of the seminal Chapel Hill study (McCombs and Shaw, 1972) that introduced the theory of agenda setting. The core concepts of this theoretical metaphor are a media agenda, a public agenda, and the transfer of salience of the items on the media agenda to the public agenda. In agenda-setting theory, salience refers to the prominence and perceived importance of items in the news. The term 'agenda' as it is used here is a neutral descriptive term, quite the opposite from its meaning in the phrase 'to have an agenda.' The transfer of salience from the media agenda to the public agenda is the inadvertent by-product of the necessity of the news media to focus their attention on a small number of topics at any particular point in time. In other words, agenda-setting effects are an incidental result of people's use of the news media. However, as we shall see, the strength of these agenda-setting effects can vary considerably. The public are not a tabula rasa waiting to be programmed by the media.

Comparing Media and Public Agendas

To test the central assertion of agenda-setting theory that the media agenda sets the public agenda, the Chapel Hill study compared the salience of five major issues defining the media agenda with the salience of these issues on the public agenda of undecided voters. The focus on undecided voters in Chapel Hill was a strategic decision about the most likely setting for media effects because the conventional wisdom of the time held that the media had little impact on the public due to their selective exposure to compatible political messages (Klapper, 1960). McCombs and Shaw reasoned that selective exposure should be minimal among undecided voters and a more likely setting for media effects.

In the Chapel Hill study and subsequent research, the media agenda is defined by the pattern of news coverage over a period of several weeks. Systematic content analysis reveals which issue receives the most coverage, which the second most coverage, etc. In other words, the issues in the news can be rank-ordered according to the amount of coverage that they receive. The public agenda most often is determined by the long-standing Gallup Poll question, 'What is the most important problem facing this country today?' The aggregate responses to this question or similar measures of the perceived importance of the issues of the day also can be rank-ordered from most to least frequent.

In Chapel Hill, McCombs and Shaw compared the salience of five major issues defining the media agenda with the public agenda among undecided voters and found a near-perfect match in their rank-order (+0.97, where the maximum value of this correlation coefficient used to index the strength of agenda-setting effects is +1.0). The empirical correlations among general populations are somewhat lower. A year-long study during the 1976 US Presidential campaign found a peak correlation of +0.63 between the television agenda and the public agenda during the spring primaries (Weaver et al., 1981). In the 1995 local elections in Pamplona, Spain (McCombs, 2014), there were substantial matches between the public agenda and the agendas of both local newspapers (+0.90 and +0.72, respectively) and television news (+0.66). Wanta and Ghanem's (2000) meta-analysis of agenda-setting studies found a typical correlation of +0.55 between the media agenda and the public agenda.

There are now hundreds of empirical studies worldwide documenting agenda-setting effects. These studies have examined the presentation of a wide variety of public issues – and a handful of other topics – by various combinations of newspapers, television, and other communication media and the public's response to these media agendas in both election and nonelection settings in Asia, Europe, Australia, and South America, as well as in the USA and Canada. For rigorous tests of

agenda-setting theory's assertion of a causal link between the media agenda and the public agenda – that is, the media agenda sets the public agenda – these effects also have been produced in controlled laboratory experiments (Althaus and Tewksbury, 2002; Conway and Patterson, 2008; Iyengar and Kinder, 1987).

In the new media landscape, a wide variety of Internet channels have joined the chorus of media voices and demonstrate agenda-setting effects among the public similar to those found over the decades for newspapers and television (Tran, 2014). With this vast expansion and transformation of the communication landscape, some observers have predicted the diminution, if not the actual disappearance, of agenda-setting effects on the scale that we have observed them over the past half century (Chaffee and Metzger, 2001). Despite the popularity of speculation on this possibility, the overwhelming preponderance of the evidence to date suggests that the agenda-setting role of the communication media endures.

Although there is divergence in the contemporary media-use patterns among different generations, two statewide surveys in the US found little difference in agenda-setting effects among the younger, middle, and older generations (Coleman and McCombs, 2007). Greater attention to the Internet and much less attention to traditional media among younger adults had little impact on agenda setting. Particularly compelling is the comparison in Louisiana of the issue agendas of low and high Internet users to the issue agenda of the state's major newspapers. For low Internet users, the correlation with the newspaper agenda is +0.90. For high Internet users, the correlation is +0.70.

Both the strength of agenda-setting effects in past decades and their continuing strength in the contemporary media setting result from long-standing patterns of behavior in the media and among the public. The high degree of homogeneity among media agendas found in the original Chapel Hill investigation continues in contemporary settings (Boczkowski, 2010). Among the public, strong agenda-setting effects across the population result from civic osmosis, the continuous exposure to a vast sea of information from many channels of communication (McCombs, 2012; Stromback and Kiousis, 2010; Webster and Ksiazek, 2012). For most people, this exposure ranges from habitual and deliberate attention to some news channels to incidental exposure to other news channels in the course of daily life. In tandem with the homogeneity of these news channels, the outcome is a high degree of consensus on the major issues of the day.

A Second Level of Agenda-Setting Effects

Agenda-setting theory focused initially on the objects defining the media and public agendas. The term 'object' is used here with the same meaning as the term 'attitude object' in social psychology. In agenda setting, the objects most frequently studied are public issues and political figures. However, any set of objects that is of interest can be analyzed, such as institutions, corporations, or brands of goods. Moving beyond the focus on objects, media messages about public issues and other objects include descriptions of these objects. In abstract terms, objects have attributes. Just as these objects vary in salience, so do the attributes of these objects. When the media present an object – and when the public thinks about and talks about an object – some attributes are emphasized. Others are mentioned less frequently, some only in passing.

Just as there is an agenda of objects, there is an agenda of attributes for each of these objects, an agenda on which the attributes can be rank-ordered according to their frequency of appearance. The influence of the media on the relative salience of these objects among the public is the first level of agenda setting. The influence of the media on the relative salience of these objects' attributes is the second level of agenda setting.

Images of political leaders among the public afford examples of attribute agenda setting (McCombs, 2014). In the 1994 mayoral election in Taipei, Taiwan, the correlations ranged from +0.59 to +0.78 for six comparisons between voters' images of the three candidates – their attribute agenda – and news coverage of these attributes – the attribute agenda in two major daily newspapers. Here the substantive attributes of these candidates included personal qualities, such as ability, experience, and political style, as well as nonpolitical attributes, such as personality, integrity, and speaking ability.

Beyond the substantive attributes of a candidate or other object in the news – the specific traits or characteristics that describe the object of attention – there is a second dimension of the attribute agenda, the affective tone associated with each substantive attribute. When the media and the public describe an object in terms of its substantive attributes, these attributes are presented in positive, negative, or neutral terms.

In the 1996 Spanish general election, both the media attribute agendas and the public attribute agendas for these candidates were organized in terms of both their substantive and affective dimensions. In this demanding test of attribute agenda-setting effects, there was substantial correspondence between the news coverage in seven news media of the three major candidates and their images among Pamplona voters. In the analysis, the combination of five substantive categories and three levels of affect resulted in a 5×3 matrix describing each candidate's attribute agenda. For six comparisons of the voters' attribute agendas for each of the three candidates with the attribute agendas of two local newspapers, the median correlation was +0.70. For six comparisons with two national newspapers, the median correlation was +0.81, and for six comparisons with two national TV news services it was +0.52.

Attribute agenda setting also occurs with public issues (McCombs, 2014). Some aspects of issues are emphasized in the news and in how people think and talk about issues. Other aspects are less salient. News coverage in Japanese newspapers about global environmental problems in the months prior to the 1992 United Nations Rio de Janeiro conference resulted in a steady increase in public agreement with the media agenda. In February, the match was +0.68 and by April, +0.78. A similar pattern was found during a 3-week period prior to a local tax election in the USA. Correspondence between the voters' attribute agenda, the relative salience of various aspects of the issue, and the local newspaper's framing of the local tax increased from +0.40 to +0.65. The match with the political advertising on the issue increased from +0.80 to +0.95. A longitudinal study of a Swiss referendum on changing the country's political asylum law also found strong attribute agenda-setting effects evolving among persons reporting heavy

reliance on newspapers and TV for public affairs information. By the third wave of interviewing, the match between the media's attribute agenda and the public's attribute agenda was +0.92. The news media have a significant influence on both object salience and attribute salience.

Exploring a Third Level of Agenda-Setting Effects

Returning to Lippmann's phrase "the pictures in our heads," the first level of agenda setting – the transfer of object salience from the media agenda to the public agenda – answers the question "What are the pictures about?" The second level of agenda setting – the transfer of attribute salience from the media agenda to the public agenda – answers the question "What are the principal characteristics presented in these pictures?" The emerging third level of agenda setting brings us closer to answering the question "What are the pictures?" Beyond the transfer of salience for objects and attributes taken individually, can the news media transfer the salience of a more integrated image, a more comprehensive picture of an object and its attributes?

In the research to date on the first and second levels of agenda setting, the elements that have been investigated are discrete objects and attributes. That is, the objects have been disaggregated from their larger context and rank-ordered according to their frequency of appearance in news stories and in survey respondents' answers to the most important problem (MIP) question or a similar measure of object salience. At the second level of agenda setting, the attributes describing political figures, issues, or other objects have been disaggregated from their larger context and rank-ordered according to their frequency of appearance in news stories and in survey respondents' answers to questions measuring attribute salience. However, in the actual news stories and in respondents' descriptions, these elements are bundled together. Attributes are linked to objects and typically more than one attribute appears in a news story or a respondent's reply.

The third level of agenda-setting effects theorizes that the salience of these bundled relationships among objects and attributes are also transferred from the news media to the public. Theoretically central to this perspective is an associative network model of memory. Rather than conceptualizing our mental representations as a hierarchical or linear structure as implied in the traditional understanding of agenda-setting theory, this associative network model holds that the representation operates pictorially, diagrammatically, or cartographically (Braddon-Mitchell and Jackson, 2007; Cummins, 1996). In this network model, individuals' cognitive representations are presented as a network-like structure where each element is connected to numerous other elements.

To explore whether the media can bundle a variety of elements and make them salient in the public's mind simultaneously, Guo and McCombs (2011a) reanalyzed data from a previous study that had found strong attribute agenda-setting effects based on a traditional analysis of discrete sets of political candidates' attributes. In the original study of the images held by voters of four statewide political candidates in the USA (Kim and McCombs, 2007), the overall correspondence between the media attribute agenda and the public attribute agenda was +0.65. Similar correlations were found when the attribute agendas of each of the four candidates were analyzed separately. Using the statistical technique of network analysis to analyze the pattern of bundled attributes on the media agenda and the public agenda – that is, an investigation of third-level agenda-setting effects – the correlation was +0.67, a result that is statistically very similar to the original analysis. Of course, the networked representations of these agendas are a much richer picture of these attributes. These results from the bundled 2002 Texas election data were replicated with new data collected during the 2010 gubernatorial election and yielded a correlation of +0.71 (Guo and McCombs, 2011b). In the 2012 US presidential election, Vargo et al. (2014) compared the correspondence of Twitter network issue agendas from mainstream media, partisan media, and the supporters of Barack Obama and Mitt Romney.

Explaining Agenda-Setting Effects

News reports are a limited portrait of our environment and, as Lippmann noted in *Public Opinion*, create a pseudo-environment to which the public responds. Often there is little correspondence between the pattern of news coverage and underlying historical trends.

Shortly after the Chapel Hill study, Funkhouser (1973) compared the responses to the Gallup Poll's MIP question across the entire decade of the 1960s with the pattern of coverage on those issues in the news magazines and found a high degree of correspondence, a correlation of +0.78. Most importantly, he added a third set of data, statistical indicators of the 'reality' of those issues across the 1960s. Funkhouser (1973: p. 72) noted that "the patterns of media coverage did not have a one-to-one relationship to the realities of any of the issues." While the media and public agendas were strongly linked, both had an arm's-length relationship with the historical trends of the decade.

Similar disconnects between 'the world outside' and the agendas of the media and the public have been found over the years for a wide variety of issues. From 1970 to 1990, the public responded to the increasing coverage of environmental problems in the face of decreasing air and water pollution (Ader, 1995). In the 1970s, public concern was aroused by news reports about the availability of petroleum in Germany when there was no real evidence of any shortage (Kepplinger and Roth, 1979). The salience of crime as an issue appears to be particularly susceptible to news coverage. Gordon and Heath (1981) found significant differences in concern about crime in Chicago, Philadelphia, and San Francisco between readers of newspapers with flamboyant crime coverage vs newspapers with conservative crime coverage. In the 1990s there was a similar public response in Texas to an increase in news coverage of crime during a time when there was a decreasing trend in actual crime (Ghanem, 1996). More recently, Gross and Aday (2003) found that exposure to local TV news with its heavy crime coverage was the major predictor of naming crime as an important issue in Washington, DC.

The public's response in all these situations is reminiscent of the phenomenon of 'alarmed discovery,' the initial stage of public response to a new issue on the agenda that is described

in Downs's (1972) theory of the 'issue attention cycle.' The media's presentation of the issues just discussed also can be characterized as 'alarmed discovery' because the news began to emphasize each of these issues at a time that nothing out of the ordinary was occurring in the real world. In effect, these were natural experiments in a real-world setting that yield compelling evidence of the agenda-setting impact of the news on the public.

These agenda-setting effects of the mass media occur worldwide, wherever there are reasonably open political and media systems. Under these circumstances, the public frequently turns to the news media for orientation on the major issues of the day, especially those issues beyond the ken of personal experience. Even in many cases where personal experience creates high salience for an issue, people turn to the media for additional information and perspective.

The concept in agenda-setting theory explaining this behavior is need for orientation (NFO), the cognitive equivalent of the physical science principle that nature abhors a vacuum. People are psychologically uncomfortable in unfamiliar situations, such as elections with a plethora of candidates and issues, and frequently turn to the media to satisfy their NFO (Weaver, 1977). This psychological concept, which is defined in terms of relevance and uncertainty, explains, for example, the strong agenda-setting effects found in 1968 among Chapel Hill undecided voters. Obviously, both relevance and uncertainty were high for these voters, the condition defining the highest level of NFO.

Matthes' (2006) recent reconceptualization of the measurement of NFO takes account of new developments in agenda-setting theory, particularly, attribute agenda setting, and expands the measurement to three dimensions: the NFO toward issues, toward facts, and toward journalistic evaluations. Chernov et al.'s (2011) experimental study found that both the traditional and new Matthes' NFO scales are reliable tools for predicting first-level agenda-setting effects, and that they are significantly correlated with each other. The traditional NFO scale performs better than the new NFO scale when predicting first-level agenda-setting effects. However, when only the first subdimension of the new NFO scale is used (i.e., NFO toward issues), both the traditional and new scales perform about equally.

With increased levels of media use, there is also increased agreement about the most important issues of the day among disparate demographic groups, such as men and women or those with high and low education. These patterns of social consensus have been found in Spain, Taiwan, and the USA (McCombs, 2014). Consensus also is facilitated by the limited capacity of the aggregate public agenda. Typically, no more than three to five issues are individually able to garner a constituency of 10% or more of the public who regard that single issue as the most important issue of the day, and the public agenda is best characterized as a zero-sum game (McCombs and Bell, 1996).

Sources of the Media Agenda

Although the majority of empirical research on agenda setting has examined the relationship between the media agenda and the public agenda, scholars also have asked "Who sets the media agenda?" Influences shaping the media agenda range from the external activities of major news sources to the internal dynamics of the media system (Dearing and Rogers, 1996; McCombs, 2014). Examination of the *New York Times* and *Washington Post* across a 20-year period found that nearly half of the news stories were based substantially on press releases and other direct inputs by news sources, such as press conferences and background briefings. News coverage of Louisiana government agencies was substantially based on information provided by their public information officers to the state's major newspapers. Across an 8-week period the correspondence between the agenda originating with the press information offices and all news stories about those agencies was +0.57.

Political campaigns make a concerted effort to influence the news agenda. In the 1993 British general election, a series of comparisons between the three major parties' agendas and seven news media, both newspapers and television, found a median correlation of +0.70. American political parties do not fare as well at the national level. A comparison of television news coverage during the 1996 New Hampshire Presidential primary, the inaugural primary in the lengthy US election year, with the candidates' speeches found only a moderate correspondence (+0.40) in their agendas. However, at the local level, in an election for Governor of Texas the combined agendas of the Democrat and Republican candidates shaped the issue agenda of both the local newspaper (+0.64) and the local television stations (+0.52) in the state capital. The Texas election also reflected intermedia agenda setting, the influence that one news medium has on another. In Austin, the correspondence between the local newspaper agenda and subsequent television news coverage of public issues was +0.73. A similar comparison in Pamplona, Spain, of two local newspapers with local television news found correlations of +0.66 and +0.70. In the USA, the *New York Times* is regarded as a major agenda-setter among the news media. A case study of the drug issue during the 1980s found that the *New York Times* influenced subsequent coverage by the national television networks, news magazines, and major regional newspapers.

Consequences of Agenda Setting

The agenda-setting role of the media has consequences beyond the focusing of public attention. These consequences encompass both attitudes and opinions and observable behavior. In terms of attitudes and opinions, it is important to distinguish two aspects, first the strength of opinion, beginning with the fundamental point of whether an opinion even exists. Strength of opinion also distinguishes between weakly and strongly held opinions regardless of whether those opinions are positive or negative. Second is the widely measured direction of opinion, whether some object or attribute is regarded in a positive or negative light.

To begin with the strength of opinion, there is a fundamental link between the first level of agenda setting, the transfer of object salience from the media agenda to the public agenda, and the formation of opinions by members of the public. For example, Kiousis and McCombs (2004) found a strong

relationship (+0.81) between the pattern of news coverage during the 1996 presidential election and the percentage of persons who held an opinion about each of 11 political figures. The greater the amount of news coverage, the greater the salience of that person among the public and the greater the likelihood of having an opinion about them. A strong relationship also was found between the pattern of news coverage and the strength of these opinions. Comparisons of the frequencies of presidential campaign stories mentioned in the *New York Times* and three network news broadcasts with a time series from the 2004 National Annenberg Election Study also found that increases in media coverage of the election were negatively related to the refusal rate in surveys about the election (Stroud and Kenski, 2001).

Turning to the direction of opinions, positive opinions during 1992 and 1993 about the overall performance in office by Hong Kong's last British Governor were significantly primed by the pattern of news coverage on his proposals to broaden public participation in local elections. Exposure to this news coverage significantly increased the importance of these proposals in Hong Kong residents' overall approval of the Governor's performance. By calling attention to some matters while ignoring others, the news media influence the criteria by which public officials subsequently are judged (Iyengar and Kinder, 1987). This process of priming is a major consequence of first-level agenda-setting effects, a process in which the salience of an issue among the public becomes a significant factor in opinions about a public figure associated with that issue. Priming effects on opinions about the overall performance of the president of the United States have been found for many presidents on a wide variety of issues. In short, there is a major link between the first level of agenda setting, the transfer of issue salience from the media agenda to the public agenda, and the direction of opinions expressed by members of the public.

At the second level of agenda setting, the tone of news reports and their attribute agenda-setting effects on the public are also related to the direction of opinions expressed by members of the public (Fishbein and Ajzen, 2010). In Germany, shifts in the tone of news stories about Helmut Kohl preceded shifts in public opinion from 1975 to 1984 (Kepplinger et al., 1989). Daily observations during the final 3 months of the 1992 and 1996 US Presidential campaigns found that the positive and negative tone of television news about key campaign events influenced voters' opinions about the candidates (Shaw, 1999). The pattern of negative headlines about the US economy over a 13-year period influenced both subsequent measures of consumer sentiment and major statistical measures of the actual economy (Blood and Phillips, 1997).

Conclusion

The history of agenda-setting theory in the decades since the 1968 Chapel Hill study is rich and productive. In particular, the theoretical expansion to three levels of effects and the consequences of these effects for attitudes and opinions require the revision of Bernard Cohen's (1963) seminal observation that the media may not tell us what to think, but are stunningly successful in telling us what to think about. His distinction between the affective and cognitive effects of the media was an important precedent for research on first-level agenda setting. In turn, the expansion to a second level of effects, attribute agenda setting, and its consequences reinvigorated the consideration of media effects on attitudes and opinions. This expanding perspective, now joined by a third level of effects, also is a rebuttal of the criticism that agenda setting has focused narrowly on the initial stages of the communication and public opinion process.

Agenda-setting theory details the range of effects on the public that result from the news media's inadvertent focus on a small number of topics and their attributes. To the extent that the news agenda is set by social forces external to the news media, the role of news institutions is important, but neutral, as a transmission belt. To the extent that the news media exercise autonomy in defining the public's news diet, they are in themselves a powerful social force.

See also: Attitudes, Political and Public Opinion; Media Effects; Political Communication; Public Opinion: Social Attitudes.

Bibliography

Ader, C., 1995. A longitudinal study of agenda setting for the issue of environmental pollution. Journalism & Mass Communication Quarterly 72, 300–311.

Althaus, S.L., Tewksbury, D., 2002. Agenda setting and the 'new' news: patterns of issue importance among readers of the paper and online versions of the New York Times. Communication Research 29, 180–207.

Blood, D.J., Phillips, P., 1997. Economic headline news on the agenda: new approaches to understanding causes and effects. In: McCombs, M., Shaw, D., Weaver, D. (Eds.), Communication and Democracy. Erlbaum, Mahwah, NJ, pp. 97–114.

Boczkowski, P.J., 2010. News at Work: Imitation in an Age of Information Abundance. University of Chicago Press, Chicago, IL.

Braddon-Mitchell, D., Jackson, F., 2007. Philosophy of Mind and Cognition, second ed. Blackwell, Cambridge, MA.

Chaffee, S., Metzger, M., 2001. The end of mass communication? Mass Communication and Society 4, 365–379.

Chernov, G., Valenzuela, S., McCombs, M., 2011. An experimental comparison of two perspectives on the concept of need for orientation in agenda-setting theory. Journalism & Mass Communication Quarterly 88, 142–155.

Cohen, B.C., 1963. The Press and Foreign Policy. Princeton University Press, Princeton, NJ.

Coleman, R., McCombs, M., 2007. The young and agenda-less? Age-related differences in agenda-setting on the youngest generation, baby boomers, and the civic generation. Journalism & Mass Communication Quarterly 84, 299–311.

Conway, M., Patterson, J.R., 2008. Today's top story? An agenda-setting and recall experiment involving television and Internet news. Southwestern Mass Communication Journal 24, 31–48.

Cummins, R., 1996. Representations, Targets, and Attitudes. MIT Press, Cambridge, MA.

Dearing, J.W., Rogers, E.M., 1996. Agenda-Setting. Sage, Thousand Oaks, CA.

Downs, A., 1972. Up and down with ecology: the "issue-attention cycle." The Public Interest 28, 38–50.

Fishbein, M., Ajzen, I., 2010. Predicting and Changing Behavior: The Reasoned Action Approach. Psychology Press, New York.

Funkhouser, R., 1973. The issues of the sixties. Public Opinion Quarterly 37, 62–75.

Ghanem, S., 1996. Media coverage of Crime and Public Opinion: An Exploration of the Second Level of Agenda Setting. Unpublished doctoral dissertation. University of Texas at Austin, Austin, TX.

Gordon, M., Heath, L., 1981. The news business, crime and fear. In: Lewis, D. (Ed.), Reactions to Crime. Sage, Beverly Hills, CA.

Gross, K., Aday, S., 2003. The scary world in your living room and neighborhood: using local broadcast news, neighborhood crime rates, and personal experience to test agenda setting and cultivation. Journal of Communication 53, 411–426.

Guo, L., McCombs, M., 2011a. Network agenda setting: a third level of media effects. Paper Presented to the ICA Annual Conference, Boston, MA.
Guo, L., McCombs, M., 2011b. Toward the third level of agenda setting theory: a network agenda setting model. Paper Presented to the AEJMC Annual Conference, St. Louis.
Iyengar, S., Kinder, D.R., 1987. News that Matters: Television and American Opinion. University of Chicago Press, Chicago, IL.
Kepplinger, H.M., Roth, H., 1979. Creating a crisis: German mass media and oil supply in 1973–74. Public Opinion Quarterly 43, 285–296.
Kepplinger, H.M., Donsbach, W., Brosius, H.B., Staab, J.F., 1989. Media tone and public opinion: a longitudinal study of media coverage and public opinion on Chancellor Kohl. International Journal of Public Opinion 1, 326–342.
Kim, K., McCombs, M., 2007. News story descriptions and the public's opinions of political candidates. Journalism & Mass Communication Quarterly 84, 299–314.
Kiousis, S., McCombs, M., 2004. Agenda-setting effects and attitude strength: political figures during the 1996 presidential campaign. Communication Research 31, 36–57.
Klapper, J., 1960. The Effects of Mass Communication. Free Press, New York.
Lippmann, W., 1922. Public Opinion. Macmillan, New York.
Matthes, J., 2006. The need for orientation towards news media: revising and validating a classic concept. International Journal of Public Opinion Research 18, 422–444.
McCombs, M., 2012. Civic osmosis: the social impact of media. Comunicacion Y Sociedad 25, 7–14.
McCombs, M., 2014. Setting the Agenda: Mass Media and Public Opinion. Polity Press, Cambridge, UK.
McCombs, M., Bell, T., 1996. The agenda-setting role of mass communication. In: Salwen, M., Stacks, D. (Eds.), An Integrated Approach to Communication Theory and Research. Erlbaum, Mahwah, NJ, pp. 93–110.
McCombs, M., Shaw, D., 1972. The agenda-setting function of mass media. Public Opinion Quarterly 69, 176–187.
Shaw, D., 1999. The impact of news favorability and candidate events in presidential campaigns. Political Communication 16, 183–202.
Shaw, D., McCombs, M. (Eds.), 1977. The Emergence of American Political Issues: The Agenda Setting Function of the Press. West, St. Paul, MN.
Stromback, J., Kiousis, S., 2010. A new look at agenda-setting effects – comparing the predictive power of overall political news consumption and specific news media consumption across different media channels and media types. Journal of Communication 60, 271–292.
Stroud, N., Kenski, K., 2001. From agenda setting to refusal setting: survey non-response as a function of media coverage across the 2004 election cycle. Public Opinion Quarterly 71, 439–559.
Tran, H., 2014. Online agenda setting: a new frontier for theory development. In: Johnson, T. (Ed.), Agenda Setting in a 2.0 World: New Agendas in Communication. Routledge, New York, pp. 205–229.
Vargo, C.J., Lei, G., McCombs, M., Shaw, D., 2014. Network issue agendas on Twitter during the 2012 U.S. presidential election. Journal of Communication 64, 296–316.
Wanta, W., Ghanem, S., 2000. Effects of agenda-setting. In: Bryant, J., Carveth, R. (Eds.), Meta-Analysis of Media Effects. Erlbaum, Mahwah, NJ, pp. 37–52.
Weaver, D., 1977. Political issues and voter need for orientation. In: Shaw, D., McCombs, M. (Eds.), The Emergence of American Political Issues. West, St. Paul, MN, pp. 107–119.
Weaver, D.H., Graber, D., McCombs, M., Eyal, C., 1981. Media Agenda Setting in a Presidential Election: Issues, Images and Interest. Praeger, New York.
Webster, J., Ksiazek, T., 2012. The dynamics of audience fragmentation: public attention in an age of digital media. Journal of Communication 62, 39–56.

Agenda Settting, Public Policy in

Christoffer Green-Pedersen, Aarhus University, Aarhus, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

It is not until recent decades that an extensive empirical research tradition has emerged in relation to policy agenda-setting. The starting point for this research has been the agenda-setting models of Kingdon as well as Baumgartner and Jones. The concepts and ideas of these two models have led to empirical investigations of many aspects of policy agenda-setting. In recent years, the policy agenda-setting literature has developed a broader focus which includes questions relating to the traditional comparative public policy literature like the role of political parties.

The agenda-setting tradition is one of the most established research traditions in public policy studies (*see* Public Policy). It is often referred to as 'policy agenda-setting theory' (Baumgartner et al., 2006), a description that distinguishes it from other agenda-setting traditions, which do not directly study public policy, for instance, studies of how the media influence the public agenda or studies how control of the formal agenda provides opportunities to influence the outcome of voting processes in parliament (*see* Agendas: Political).

The core claim of the policy agenda-setting tradition is that changes in attention generate policy changes. If we want to understand why and how some policy changes – and why it does not – it is crucial to study what issues or policy questions the decision makers pay attention to and what makes them change their attention. A fundamental idea of the tradition is that attention is a scarce commodity in politics and the existence of problems and solutions does not by itself generate decisions. You need the scarce thing called the actors' attention.

The following describes how this very general understanding of policy making has been unfolded into empirical research. We outline the intellectual origins of the tradition and the research tradition with focus on the two books that have shaped it: John Kingdon's (1984, 1995) *Agendas, Alternatives and Public Policies*, and Frank R. Baumgartner and Bryan D. Jones' (1993, 2009) *Agendas and Instability in American Politics*. (Both have appeared in several editions. The new editions are primarily updates of the empirical material from the first editions. The 1995 version of John Kingdon's book and the 2009 version of Baumgartner and Jones' book are used here.) We also outline some of the findings from a number of recent comparative studies which show how an understanding of agenda-setting dynamics is crucial for explaining comparative differences in public policy.

The Roots of Policy Agenda-Setting Research

The intellectual roots of the policy agenda-setting tradition are two seminal pieces in political science. One is Bachrach and Baratz's article on the *Two Faces of Power* (1962), where the second face of power points to the crucial role of agendas (*see* Power). The other and probably the most important one is Schattschneider's (1960) *The Semisovereign People*. Many of the central insights from the book like: 'Organization is the mobilization of bias,' 'The definition of alternatives is the supreme instrument of power' and 'A conflict is likely to change profoundly as it becomes political' are core insights in policy agenda-setting research today. Or, put simply, current research within the policy agenda-setting tradition can be considered attempts to turn Schattschneider's ideas into empirical research.

Though the intellectual roots of the policy agenda-setting tradition are 50 years old, policy agenda-setting research as an empirical research tradition did not take off until the 1970s and has not really flourished until the last 15–20 years. Part of the reason is probably that the founding works of the tradition, i.e., Schattschneider and Bachrach/Baratz, focused more on presenting the importance of agenda-setting as a critique of the pluralist view of the openness of the political system. Thus, the starting point for the tradition was focused on highlighting the importance of keeping issues away from the agenda rather than – empirically – researching the issues that did receive political attention.

Studies of Policy Agenda-Setting

The First Studies

The first example of more 'positive' and empirically oriented policy agenda-setting research was Cobb and Elder's book (1972, 1983) *Participation in American Politics*. (The book was published in a second edition in 1983. References are made to the second edition.) Besides being the first book-size example of empirical studies of agenda-setting, the book raised at least two theoretical issues that have been central in the tradition since. One is the question 'What is an agenda?' In the policy agenda-setting tradition, the political agenda is broadly understood as the set of issues that are debated and considered for decision in a political system at a given time (Baumgartner, 2001: p. 288). Though broadly agreed on, such a definition is too general as guidance for empirical research. Thus Cobb and Elder (1983: pp. 14–16) introduced a distinction between a systemic and an institutional agenda. The systemic agenda is issues discussed in society, whereas the institutional agenda is the agenda of a particular political institution. The distinction opened a question about the relationship between the issues debated in society and in the institutions in the political system

as well as a discussion about the relationship between the agendas of different institutions in the political system.

The other theoretical question was the importance of issue characteristics. In the policy agenda-setting tradition issues are mostly understood as relatively broad policy issues like energy, the economy, health, or education, but sometimes also more narrowly as particular policy questions like abortion, nuclear power, or inflation. No matter how broadly issues are defined, their policy content varies enormously. From a policy perspective, the economy or inflation is very different from transportation or railways. By focusing on issue characteristics like technical complexity and the size of the affected public, Cobb and Elder (1983: pp. 94–109) offered the first attempt to theorize how agenda-setting processes were affected by an issue's characteristics. Though Cobb and Elder's book contained many questions for future agenda-setting research, few were actually taken up.

Simultaneously with Cobb and Elder's book, Downs (1972) published the article 'Up and down with ecology. The issue attention cycle.' The article has been influential with its analysis of the public or systemic agenda and how attention to a particular issue – in this case the environment – may suddenly be triggered, but disappear again as people realize that problems like the environment cannot be solved. The idea of an issue attention cycle thus highlights how attention is very volatile if it is not in some way institutionalized as it rarely is when we study the public or systemic agenda.

Kingdon's Model of Agenda-Setting

The next major study is Kingdon's seminal book, *Agendas, Alternatives and Public Policies*, first published in 1984, which presents a theory or a framework for understanding how attention affects political decision making that is still central within the policy agenda-setting tradition. A useful starting point for presenting Kingdon's agenda-setting model is his distinction between what he labels the governmental agenda and the decision-making agenda (1995: pp. 3–4). The governmental agenda is basically the issues that government officials and the actors surrounding them (experts, lobbyist, politicians, and journalists) pay attention to, whereas the decision-making agenda is the issues on the governmental agenda that are actually up for decision. To put it simply, Kingdon's model of agenda-setting can be considered a model of how issues move from the governmental agenda to the decision agenda or what it takes to move issues from just being considered by political actors toward actually being decided on. This also implies that Kingdon when compared with for instance Cobb and Elder focuses on a subsequent aspect of the policy making process. The question of how the governmental agenda relates to the systemic agenda is not Kingdon's central focus. Questions relating to the systemic agenda are mainly considered when they are important for understanding the relationship between the governmental agenda and the decision-making agenda. Kingdon's model of agenda-setting also draws extensively on organizational and decision-making theory, especially the garbage can model (1995: pp. 84–86).

Key concepts in Kingdon's model are the three streams of problems, alternatives and politics. Problems (1995: pp. 90–115) are basically societal problems. Alternatives (1995: pp. 116–144), or 'the policy primeval soup,' consist of the different policy solutions that actors within the policy communities surrounding most issues are trying to promote. Finally, the politics stream consists of political factors like the public mood, election results, and changes of administration – or government outside the US context (1995: pp. 145–164). According to Kingdon, the streams are normally not connected. Thus the mere existence of problems and alternatives, i.e., solutions, does not move issues from the governmental to the decision-making agenda, nor does the existence of politics besides problems and alternatives. Further, there is no logical connection between the three streams. For instance, actors in the alternatives stream may be actively looking for problems with which to connect their preferred alternatives.

The streams are connected by policy windows (Kingdon, 1995: pp. 165–195), which may appear in the problem stream or the political stream. In the problem stream, focusing events – unpredictable events like earthquakes or terrorist attacks – may open policy windows, but policy windows that open in the problems stream are not necessarily unpredictable. Measuring societal conditions is an important aspect of defining them as political problems and the release of new measures, for instance unemployment statistics, constitutes an often predictable policy window. In the political stream, windows may open both predictably and unpredictably as well. A change of administration is predictable, but other aspects like the retreat of a minister and the appointment of a new one can be very unpredictable. Policy windows rarely open from the alternatives stream. New alternatives of course emerge, but that does not in itself open a policy window. Rather, new alternatives emerge in the policy primeval soup and then have to wait for a policy window to open where they might be connected with a problem. The coupling of the independent stream is not an automatic process even when a policy window has opened. The coupling of the three streams takes a 'policy entrepreneur' (pp. 172–183) who can use a political platform to couple the streams and thus move issues from the governmental to the decision-making agenda.

As Mucciaroni (2012) has pointed out, Kingdon's model is widely cited, but rarely directly tested and used as foundation for systematic empirical research. This may be an indication of the model's weaknesses. A frequent critique of Kingdon is that the model portrays the policy process as basically unpredictable, which of course makes it difficult to use it as a basis for predicting policy decisions. Kingdon's model clearly stresses the unpredictability of the agenda-setting process. No actor controls the process. However, Kingdon also points to many predictable aspects of the process. Policy windows can often be predicted by actors and certain actors are much more likely to be successful policy entrepreneurs than others. For instance, the US president has a privileged position (1995: pp. 23–26). Kingdon has also been central in developing a literature on focusing events (cf Birkland and DeYoung, 2012). That being said, the strength of his model is that it describes *when and how* rather than *why* issues move from the government agenda to the decision-making agenda.

Baumgartner and Jones' Model of Agenda-Setting

Kingdon's model of agenda-setting is an important part of the intellectual foundation of Baumgartner and Jones' work on agenda-setting, which has been pivotal in the development of the policy agenda-setting tradition over the past 20 years. Compared to Kingdon's agenda-setting model, Baumgartner and Jones' work initiated by the 1993 book has a somewhat different focus. Kingdon's model focused on how issues move onto the decision-making agenda during the short periods when policy windows open. Baumgartner and Jones' work focuses on long-term patterns of attention to policy issues and its effect on policy making. Whereas Kingdon – and Cobb and Elder – based their agenda-setting models on case studies, Baumgartner and Jones base their work on systematic long-term – often decades – tracking of attention to particular issues on the entire agenda. Likewise, attention was tracked on the systemic agenda – media attention and public opinion – and on institutional agendas like Congress, the Presidency, and the Supreme Court; see www.policyagendas.org.

The long-term perspective is a central feature of Baumgartner and Jones' work because it has been the foundation of the punctuated equilibrium model. The core idea of this model is that attention and public policy are characterized by long periods of stability and short periods of dramatic change. Such a pattern can only be observed when attention or policy is tracked over long periods. To understand what lies behind the punctuated equilibrium pattern, Baumgartner and Jones (1993) have developed an agenda-setting model that explains both stability and change. Stability or equilibrium is generated by the existence of what Baumgartner and Jones describe as 'policy monopolies' (2009: pp. 6–12) (*see* Issue Networks: Iron Triangles, Subgovernments, Policy Communities, Policy Networks). Such monopolies will develop around policy issues like nuclear power or tobacco. They contain both a venue like a congressional committee where actors interact, and a 'policy image' or a shared understanding of the particular policy problem. A policy image consists of both a causal understanding of the policy problem and a positive or negative view of it (pp. 25–38). One example is nuclear power in the US, which for a long time enjoyed a positive policy image. Nuclear power was seen a clean and efficient way of solving energy problems; it was linked to a causal understanding that nuclear power was safe and without major negative side effects like waste (Baumgartner and Jones, 2009: pp. 60–82).

Policy monopolies – consisting of venues and policy images – generate stability or equilibrium. If new information – for instance information about nuclear waste problems – challenges the policy image, the policy monopoly will able to ignore or diffuse the information. Such 'negative feedback mechanisms' ensure a policy monopoly long-term stability. However, nothing lasts forever and a central point in Baumgartner and Jones' agenda-setting model is that equilibriums are always 'local.' The political system in general is never in equilibrium, which implies that it is the interaction with the surrounding political system that is potentially destabilizing for a policy monopoly. There are always actors in the political system surrounding a policy monopoly who want to destabilize it, and one option is to use focusing events.

Such a 'Schattschneider mobilization' where attention to a policy issue is expanded beyond an existing policy monopoly may also be related to what Baumgartner and Jones (2009: pp. 86–102) label 'venue shopping.' A policy monopoly is linked to a particular venue and generating attention in another venue is a way to challenge the monopoly. The policy monopoly surrounding US tobacco policy – which was based on tobacco as an exportable agricultural product – was also linked to the agricultural committees in Congress. This monopoly has been overthrown by a process that started at the US state level. States like California had no interest in tobacco as an agricultural product, but had to bear the health care costs of smoking. US states thus began to regulate smoking and took the issue to the courts in order to reclaim health care costs. This expanded the tobacco issue to the broader political system and led to the breakdown of the original tobacco monopoly (2009: pp. 264–281).

Policy monopolies thus create local equilibriums, which may last long, but in the long run they will not stay isolated from the surrounding political system and may thus be overthrown. According to Baumgartner and Jones such periods are characterized by positive feedback processes where challenges to monopolies become self-reinforcing and policy monopolies that have existed for a long time may quickly disappear. Baumgartner and Jones call these short periods of changes 'punctuations.'

The idea of policy monopolies raises a number of questions; for example, what happens when new policy issues emerge, i.e., when there is prior policy monopoly? Policy monopolies do not emerge automatically, but are generated by what Baumgartner and Jones label a 'Downsian mobilization,' referring to Downs' analysis of the emergence of the environment on the systemic agenda. Downs' issue attention cycle implied that attention to issues disappears again. However, periods of intense attention to an issue leaves behind policy monopolies as actors are able to institutionalize their preferred policy image linked to certain policy venues. Periods of sparks in attention to new issues thus have long-term implications.

Baumgartner and Jones' work as first presented in the 1993 book has been pivotal in the flourishing of the policy agenda-setting tradition during the last decades. The most direct application of Baumgartner and Jones' (1993) book is testing the idea of punctuated equilibrium through studies of the distribution of changes in both attention and budgets. Punctuated equilibrium entails the idea of either stability or few dramatic changes, but few gradual changes. If we study the distribution of changes, we should thus see a picture of many very small changes, i.e., stability, and a few substantial changes. This distribution can be compared with a normal distribution and we can test whether it is leptokurtic, i.e., has a higher peak and more slender shoulders than a normal distribution. Changes in attention and budgets have been found to exhibit this pattern across different political systems (Baumgartner et al., 2009; Jones et al., 2009), but the pattern has also been found to be more pronounced the further we move down the policy cycle (*see* Policy Cycle). Policy decisions – budgets – are more leptokurtic than inputs to the political process like elections or media attention, which are closer to a normal distribution. The universal finding of punctuated equilibrium of course raises the question of what can explain it. Here Jones and Baumgartner (2005) point to the idea of friction in terms of information processing at both the individual level and the institutional level. Individuals and institutions tend to either over- or underattend problems.

Agenda-Setting and Comparative Public Policy

The work of Baumgartner and Jones has shaped the policy agenda-setting tradition fundamentally. In recent years, the tradition has thus been characterized by a 'comparative turn' based on the establishment of major datasets on attention in a number of countries similar to one developed by Baumgartner and Jones on the US, see www.comparativeagendas.info.

One the one hand, the development of these datasets have broadened the scope of policy agenda-setting theory beyond public policy to traditional comparative politics questions about party competition, federalism, coalition government, and the functioning of political systems (Green-Pedersen and Walgrave, 2014). One the other hand, the development of these comparative datasets has also moved policy agenda-setting literature closer to more traditional comparative public policy literature. In terms of interacting with comparative public policy literature, the policy agenda-setting tradition has traditionally been limited by its very strong US focus. The work of Kingdon and Baumgartner and Jones has only been based on studies of US policy making. With the comparative expansion of the policy agenda-setting tradition, the empirical foundation for interacting with the comparative public policy literature exists and there are a number of examples of this.

One example is Engeli et al.'s (2013) study of cross-national differences in policy permissiveness with regard to morality issues (abortion, euthanasia, stem cell research, etc.). The article argues that to understand such comparative differences, we must first understand whether the issues are politicized, i.e., are part of party competition. Whether or not morality issues are politicized can again be explained by whether or not there is a preexisting conflict in the party system between confessional and secular political parties (*see* Church and State). In the affirmative, morality issues are seen as yet another example of this conflict, which brings it into party competition. In countries where morality issues are politicized, permissive policy decisions are passed by secular governments, whereas confessional – often Christian Democratic – governments try to delay permissive decisions. In countries where morality issues are not politicized due to the absence of a conflict between confessional and secular parties, which parties govern is not decisive for permissive policy decisions. These decisions are rather generated by policy entrepreneurs like individual MPs or interest groups who manage to push the issues to a decision, often without much attention or political resistance.

Another example is Green-Pedersen and Krogstrup (2008) study of attention and policy change in relation to immigration in Denmark and Sweden (*see* Immigration Policy). Both countries had liberal immigration policies in the 1980s, but from the 1990s and especially during the 2000s, Danish immigration policy became considerably more restrictive. The explanation for this should be found in the in differences in agenda-setting process and party competition in the two countries. In Denmark, the major right-wing parties, the Conservatives and the Liberals, increased their attention to the issue when in opposition from 1993 to 2001 supported by the radical right-wing Danish People's Party. This lead to increasing political attention to immigration and caused the Social Democratic led government to tighten immigration policy. After returning to government in 2001, the Liberals and Conservatives implemented the very restrictive immigration policy they had argued for when in opposition. In Sweden, the right-wing bloc has always been dominated by center-right parties and the major right-wing party, the Conservatives, have always been forced to cooperate toward the center rather than toward the radical right. Therefore, the party has never tried to make immigration a central issue on the political agenda in Sweden by advocating a restrictive immigration policy. Immigration has thus never been central on the political agenda in Sweden and the country has stayed with its liberal immigration policy.

These comparative turn within policy agenda-setting research has also implied that theoretical debates within the comparative public policy literature are engaged more directly like for instance the 'do politics matter' debate (*see* Ideological Constraint: History and Current Status of the Concept). Seeberg's (2013) study of law and order in Denmark for instance show that politics do not only matter for public policy in terms of the color of the government. The opposition may also, if it is successful in bringing issues high on the political agenda, cause the government to move policy in the opposite direction of what the government would have preferred. Increasing attention to law and order in Denmark generated by the right-wing opposition thus caused the Social Democratic led governments in the 1990s to implement policies like longer sentences, which the Social Democrats had before resisted.

A final example of the comparative turn draws more directly on the concept of venue shopping developed by Baumgartner and Jones (2009). Political institutions function not only as arenas or venues of decision making but also as venues of attention. Filing a court case or scheduling a hearing is a way to draw attention to an issue whether or not a decision is made. Further, no political arena or venue is neutral. Tobacco-related questions are treated differently in a health care committee compared to an agricultural committee; and a court is very different from a congressional committee or a parliament. This idea has especially inspired studies of the European Union, where its 'complex' institutional structure makes the idea of venue shopping applicable. Guiradon (2000) study of immigration policy in the European Union and Sheingate's (2000) comparative study of agricultural policy in the European Union and the US are examples of this.

Recent Trends in Policy Agenda-Setting Research

Politics in Western countries have over the last decades being characterized by such developments as increased mediatization and rising importance of issue voting. This has made the question of agenda-setting of increasing importance. Today, it is less given which issues should dominate the political agenda and how they should be framed. The increased use of communication strategies and communication experts by political actors is one sign of this.

For our understanding of public policy, this makes a theoretical understanding of policy agenda-setting of increased importance. Political attention is still scarce, so most policy decisions are likely to be made with relative limited political attention from a closely defined group of actors – what Baumgartner and Jones describe as 'subsystem politics.'

However, the chances of issues and the related policy decision being subject to macropolitical attention have increased substantially. The agenda-setting literature has shown that attention is both scarce and consequential. If issues are brought out of the sub-system context and into macropolitics, the policy processes and decisions are likely to change.

Despite its long tradition, the policy agenda-setting literature has only begun to develop a profound theoretical understanding of agenda-setting dynamics and many aspects needs much further investigation. One aspect that is particularly in need of this is the question of framing or issue definition. 'Policy images' are a key element in a policy monopoly and more broadly points to the central importance of how policy problems are defined or framed. Baumgartner and Jones (2009 [1993]) were influenced by Deborah Stone's (1988) seminal book *Policy Paradox and Political Reason*, which highlighted how policy problems can be defined or constructed differently in terms of 'causal stories' with very different implications for policy solutions. The most recent and extensive empirical example of the importance of 'causal stories' or framing is Baumgartner et al.'s study of framing of the death penalty in the US (2008). The study is an ideal example of how framing can be studied empirically as it traces the long-term evolution of the framing of the issue in the media. It is also an ideal example of how powerful a reframing of an issue can be. Whereas the death penalty was long seen and supported as a morally justified sentence in case of particularly brutal crimes, it was increasingly framed as a policy tool that often led to wrongful convictions with huge costs for society. DNA tests proved the innocence of convicted and sometimes executed prisoners and raised serious questions about the death penalty as a reliable policy instrument. Further, the costs of long trials with numerous appeals often followed by long periods on death row for convicts were linked to widespread skepticism toward government in the US. The death penalty was framed as yet another example of an expensive government that makes too many mistakes. Support has declined; the number of executions has dropped; and many states have abolished the death penalty. This development came about because the issue was reframed as yet another example of government failure – not because public belief in death penalty as morally justified changed. This study clearly demonstrates how consequential a reframing of a policy question can be, but it is also a unique example and many more of such studies are needed.

The importance of agenda-setting dynamics for democratic political system has been recognized for decades. However, developments like mediatization and increased importance of issue voting have made it increasingly important to study agenda-setting process. In recent years the agenda-setting tradition has gained momentum and started to engage more directly with research within both the comparatively public policy tradition like the 'do politics matter' question and political science more broadly. Thus as agenda-setting is becoming more important, the research tradition has broadened its theoretical platform and has expanded its empirical basis much beyond the US.

See also: Agendas: Political; Church and State; Ideological Constraint: History and Current Status of the Concept; Immigration Policy; Issue Networks: Iron Triangles, Subgovernments, Policy Communities, Policy Networks; Policy Cycle; Power; Public Policy.

Bibliography

Bachrach, P., Baratz, M.S., 1962. Two faces of power. American Political Science Review 56, 947–952.

Baumgartner, F.R., 2001. Political agendas. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of Social and Behavioral Science. Political Science. Elsevier Science and Oxford: Pergamon, New York, pp. 288–290.

Baumgartner, F.R., Jones, B.D., 2009[1993]. Agendas and Instabilities in American Politics. University of Chicago Press, Chicago.

Baumgartner, F.R., Green-Pedersen, C., Jones, B.D., 2006. Comparative policy agendas. Journal of European Public Policy 13, 959–974.

Baumgartner, F.R., de Boef, S., Boydstun, A., 2008. The Decline of the Death Penalty and the Discovery of Innocence. Cambridge University Press, Cambridge.

Baumgartner, F.R., Breunig, C., Green-Pedersen, C., Jones, B.D., Mortensen, P.B., Nuytemans, M., Walgrave, S., 2009. Punctuated equilibrium in comparative perspective. American Journal of Political Science 53, 603–620.

Birkland, T., DeYoung, S.E., 2012. Focusing events and policy windows. In: Araral Jr., E., Fritzen, S., Howlett, M., Ramesh, M., Wu, W. (Eds.), Routledge Handbook of Public Policy. Routledge, London, pp. 175–188.

Cobb, R.W., Elder, C.D., 1972, 1983. Participation in American Politics. John Hopkins University Press, Baltimore.

Downs, A., 1972. Up and down with ecology. The issue attention cycle. Public Interest 28, 38–50.

Engeli, I., Green-Pedersen, C., Larsen, L.T., 2013. The puzzle of permissiveness. Understanding policy processes concerning morality issues. Journal of European Public Policy 20, 335–358.

Green-Pedersen, C., Krogstrup, J., 2008. Immigration as a political issue in Denmark and Sweden. How party competition shapes political agendas. European Journal of Political Research 47, 610–634.

Green-Pedersen, C., Walgrave, S., 2014. Agenda Setting, Policies, and Political Systems: A Comparative Approach. University of Chicago Press, Chicago.

Guiradon, V., 2000. European integration and migration policy: vertical policy-making as venue shopping. Journal of Common Market Studies 38, 251–271.

Jones, B.D., Baumgartner, F.R., 2005. The Politics of Attention: How Government Prioritizes Problems. University of Chicago Press, Chicago.

Jones, B.D., Baumgartner, F.R., Breunig, C., Wlezien, C., Soroka, S., Foucault, M., François, A., et al., 2009. A general empirical law of public budgets. A comparative analysis. American Journal of Political Science 53, 855–873.

Kingdon, J.W., 1984, 1995. Agendas, Alternatives and Public Policies. Harper Collins, New York.

Mucciaroni, G., 2012. The garbage can model and the study of the policy making process. In: Araral Jr., E., Fritzen, S., Howlett, M., Ramesh, M., Wu, W. (Eds.), Routledge Handbook of Public Policy. Routledge, London, pp. 320–328.

Schattschneider, E.E., 1960. The Semi-Sovereign People. A Realist's Guide to Democracy in America. Holt, New York.

Seeberg, H., 2013. The opposition's policy influence through issue politicisation. Journal of Public Policy 33, 89–107.

Sheingate, A., 2000. Agricultural retrenchment. Issue definition and venue change in the United States and European Union. Governance 13, 335–363.

Relevant Websites

www.comparativeagendas.info.
www.policyagendas.org.

Agendas: Political

Frank R Baumgartner, The University of North Carolina, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The political agenda is the set of issues that are the subject of decision making and debate within a given political system at any one time. The concept of conflict expansion is key to the idea of agendas. Issues are typically handled in a routine manner within specialized policy subsystems until they emerge on the agenda. Issue-definition, or framing, plays a key role here, as do the roles of individual policy entrepreneurs, social movements, and political leaders. Studies of agenda setting have blossomed in the past 10 years as new data resources have become available.

The political agenda is the set of issues that are the subject of decision making and debate within a given political system at any one time. Cobb and Elder (1972), in the first book-length treatment of the political agenda, noted the difference between the systemic agenda, defined as the group of issues that were under discussion in society, and the institutional agenda, or the set of issues being discussed in a particular government institution. Since then, scholars have variously written about the public agenda, the media agenda, the legislative agenda, and any number of other agendas as they have focused on different political institutions. Studies of agenda setting have often focused on the question of issue definition, which has led scholars to note the importance of social movements, professional communities, and strategically minded politicians and policy entrepreneurs in affecting how we think about important social issues, and how these social issues become political issues in the first place. Whereas the questions of agenda setting were once taken for granted or ignored, the field today is home to a vibrant intellectual community with many studies being conducted each year. Newly available quantitative indicators of public attention, media coverage, and the like have encouraged larger-scale studies of the agenda-setting process.

The political agenda is the set of issues that are the subject of decision making and debate within a given political system at any one time. Significant research specifically on the topic of agenda setting, as opposed to decision making, dates mostly from the 1960s. Early studies of agenda setting were quite controversial because they were often presented as critiques of the pluralist studies of the 1950s and 1960s. Truman (1951) mostly ignored the issue of who set the agenda of political debate. Dahl (1956) discusses the matter in mentioning that ensuring that no group has control over the range of alternatives discussed within the political system is a requisite for democracy. In his study of New Haven he explicitly raises the question of agenda setting, noting that with a permeable political system virtually all significant issues would likely come to the attention of the elites. "Because of the ease with which the political stratum can be penetrated, whenever dissatisfaction builds up in some segment of the electorate party politicians will probably learn of the discontent and calculate whether it might be converted into a political issue with an electoral pay-off" (Dahl, 1961: p. 93). In Dahl's view, then, any issue with a significant potential following in the public would likely find an elite-level champion, though he also notes that issues with no large-scale electoral pay-off might never enter the agenda.

Conflict Expansion

E.E. Schattschneider (1960) focused attention on how political debates often grow from the conflict of two actors, the more disadvantaged of whom may have an incentive to 'socialize' the conflict to a broader political arena. Of course, the more advantaged disputant strives to 'privatize' the conflict. Schattschneider was one of the first to note that the composition of the political agenda was itself a fundamental part of the political process, and he was the first to give it a prominent role in his view of the political system. By around 1960, then, scholars had firmly noted the importance of the study of the political agenda as an important area of research.

After the critique of Schattschneider (1960), scholars were less willing to take the composition of the agenda for granted. Peter Bachrach and Morton Baratz (1962) provided one of the most telling critiques of pluralism when they noted that studies of decision-making, power, and influence were misleading. Their aptly titled article, 'The two faces of power,' noted that the 'first face' of power, the authority to choose between alternatives, may be less important than the 'second face' of power, the ability to control what alternatives are under discussion in the first place. Whereas Dahl and others saw this as a relatively open process, where any social group with a legitimate problem that could potentially be converted into votes in an election could gain access to the political agenda, others saw the process in a decidedly more negative light. Following Bachrach and Baratz, many scholars attempted to study not just governmental decision making, as the pluralists had done, but also nondecisions, or agenda control, as well. For example, Matthew Crenson (1971) noted that air pollution was rarely discussed in public or government in one city despite a very serious pollution problem. In another similar city with much less pollution, however, public and governmental leaders discussed it often and took steps to combat it. The reason behind the difference in the behavior between the two cities appeared to be the ability of powerful economic interests to control the agenda. John Gaventa (1980) followed this study with an analysis of poverty-stricken Appalachian towns and the 'quiescence' characterizing the demobilized populations there.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.93003-4

These agenda theorists argued that power was most evident when objective conditions of suffering were not the subject of debate. Bachrach and Baratz (1962), Crenson (1971), and Gaventa (1980) raised important issues and directly challenged the relatively optimistic views of the pluralists but did not convince all, because of the difficulty of discerning exactly what would be a neutral political agenda. In other words, it was hard to know what findings would demonstrate elite control and what findings would demonstrate democratic openness; in this situation two scholars looking at the same findings could disagree forever (and they did; see Baumgartner and Leech, 1998: Chapter 3, for a discussion of these issues relating to the community power studies of the 1950s and 1960s; see also Polsby's (1980) treatment of these methodological issues). For example, Dahl (1961) had already noted that politicians might become aware of a grievance but see no electoral value in acting on it. While this was perhaps portrayed as reflecting the lack of gain, it could also have been because of an expected push-back or objection from those interests who might be harmed if action were taken. For Gaventa and Crenson, emphasis was on this former possibility: that the economically and socially powerful would, through their closer connections to those in power, effectively keep important issues – especially those affecting the distribution of wealth – from ever reaching the agenda for debate. Dahl saw the system as relatively open for those with a significant grievance; Gaventa and Crenson suggested that perhaps this was not so, or true only within certain bounds of acceptability.

The Development of a Literature

Roger Cobb and Charles Elder (1972), in the first book-length treatment of the political agenda, noted the difference between the systemic agenda, defined as the group of issues that were under discussion in society, and the institutional agenda, or the set of issues being discussed in a particular government institution (see also Cobb et al., 1976). Since then, scholars have variously written about the public agenda, the media agenda, the legislative agenda, and any number of other agendas as they have focused on different political institutions.

More recent studies of agenda setting have moved away from the concepts of nondecisions and power because of the difficulties inherent in designing rigorous research on the topic. Instead, scholars have focused on the rise and fall of issues in the public or institutional agendas and how decision making during high salience periods differs from the more routine decision making that takes place when an issue is low on an agenda. Jack Walker (1977) provided one of the first statistically based studies in the area with his analysis of the US Senate's agenda. He noted that issues often rose on the Senate's agenda following heightened levels of discussion within professional communities.

John Kingdon's (1984) treatment of the public agenda set the stage for much of our current understanding of where issues come from. He emphasized the separate sources of policy problems from the solutions that may be offered to them. Government programs, he noted, come about when a given solution is attached to a particular problem, and his analysis of health care and transportation policies in the USA showed just how unpredictable these couplings can be. Political actors' search for popular issues, windows of opportunity open and close, stochastic events such as natural disasters or airplane crashes momentarily focus public attention on an issue. The confluence of many unrelated factors, often serendipitous, helps explain why a given policy is adopted, according to his study. Kingdon's (1984) was the first major book-length study on the topic since Cobb and Elder's (1972), and it was based on hundreds of interviews with government and other policymakers in the 1970s and 1980s. (Polsby, 1984 also reached many of these conclusion in a book that appeared in the same year as Kingdon's.)

Frank Baumgartner and Bryan Jones (1993) provided the next major treatment of political agendas in their analysis of nine different policy areas over a 40-year period. Utilizing publicly available sources such as media indices and records of congressional hearings, they noted how particular issues rose and fell on the agenda over the entire post-World War II period. They developed a punctuated equilibrium model of policy change in which episodic periods of high agenda status typically were related to dramatic and long-lasting policy changes. During these high-salience periods, institutional procedures were often created or altered. The subsequent ebbing of the issue from the public agenda enabled the newly empowered political institutions and policymakers to settle into stable routines of behavior persisting for decades at a time. Agenda setting was related to dramatic changes, often upsetting long-standing routines of behavior and power by replacing them with new ones.

Policy Communities

The literature on agenda setting is difficult to read outside the context of the literature on policy communities, issue-networks, and the roles of specialized communities of policy experts. This literature, dating at least to the early work of Bentley (1908), focuses on the idea that public policies develop complicated and specialized communities of experts who may gain some degree of autonomy through their shared knowledge of a given policy domain. Because the government deals simultaneously with hundreds or thousands of different public policies, at any given time the vast majority of these issues are 'off the agenda' and when they are, this means that specialized communities of experts are operating with relatively little political or journalistic oversight. The literature on 'iron triangles,' 'whirlpools,' 'subgovernments,' 'policy communities,' and 'issue-networks' focused on the potential for such groups to gain substantial autonomy from the broader political system (see for example Lowi, 1964, 1969; Berry, 1989; Browne, 1998; Heinz et al., 1993; Peterson, 1993; this work also carried over into comparative politics – Hall, 1993; Heclo, 1974, 1978; international relations – see Haas, 1992; Sikkink and Keck, 1999). The question of interest and one of the key motivators, following from Schattschneider's analysis of conflict expansion, was what happens when one set of political values dominates among the experts but another obtains in the larger political arena. Influential analyses of the US Congress were also based on a discussion of the 'Giant Jigsaw Puzzle' which involves granting relatively

significant influence to 'policy specialists' serving on the various committees, knowing that these specialists might be 'high demanders' for the services – such as aid to farmers or support for the Defense Department – if they were allowed to seek their committee assignments based on such factors as heavy constituency reliance on the industry in question (see Shepsle, 1978). Agenda setting is also the analysis of the dynamics of the relations among experts and generalists in the political arena, and of the political ramifications of devolving power to experts in different policy domain each working in parallel but with little centralized oversight.

Issue Definition

Schattschneider brought our attention to the dynamics of conflict expansion but he did not explicate a theory of how conflict expanders ask for help. Issue definition and framing are the keys here. Studies of agenda setting have often focused on these questions because of this. Echoing a major theme in Baumgartner and Jones (1993), David Rochefort and Roger Cobb (1994) brought together a number of essays showing how public understanding and media discussion of a given issue can change over time, often quite dramatically. Deborah Stone (1988) also discussed this in her analysis of 'causal stories.' Policy entrepreneurs frame issues by explaining the causes of a given problem with a narrative justifying a particular governmental response. Book-length studies of the issues of child abuse (Nelson, 1984), pesticides (Bosso, 1987), health care reform (Hacker, 1997), and various natural and human-made disasters (Birkland, 1997) have shown the impact of changing issue definitions and of focusing events in pushing an issue on to the public agenda. Roger Cobb and Marc Howard Ross (1997) brought together a series of essays on the rarely studied topic of 'agenda denial,' whereby political actors keep threatening issues off the agenda.

William Riker (1986, 1988, 1993, 1996) showed the importance of two related issues: the ability of strategically minded politicians to alter the terms of debate by skillfully manipulating issue definitions, and the power of formal agenda control. A voluminous literature in formal and game theory suggests that the controller of a formal agenda can affect the outcomes in a voting situation by altering the order in which alternatives are considered. Riker used game theory to illustrate how formal agenda control can affect such things as votes in a parliamentary setting, and case studies and historical illustrations to show how political leadership could be even more powerful through the means of altering issue definitions. Political leaders can utilize a combination of formal agenda control and informal debating skills to achieve their ends, according to Riker.

Social Movements and the Media

A number of scholars have noted that social movements have often successfully brought new issues onto the public agenda. Thomas Rochon's (1998) analysis of the peace movement in various Western countries fits in this tradition, as does the work of Douglas McAdam (1988), whose study of the Mississippi Freedom Summer documented the success of civil rights activists in putting the issue of racial equality on the national political agenda during the mid-1960s.

Studies of the media agenda have been legion, largely following from the early work of Max McCombs and Donald Shaw (1972); for a review of this literature, see Rogers and Dearing (1988). Bernard Cohen (1963) noted famously that while the media cannot tell the public what to think, they can have a great impact on what the public think about. Within political science, several authors have picked up on the issue of media effects on public opinion (Iyengar, 1991; Iyengar and Kinder, 1987). James Stimson (1991) noted the changes in a broadly measured national mood based on public opinion surveys; John Kingdon (1984) also put considerable emphasis on the national mood in his study of agenda setting in government. As policymakers consider what issues to spend their time on, Kingdon (1984) noted they often make reference to the idea of a national mood.

Studies of the political agenda have been remarkable in political science for their integrative character: rather than focusing on any particular institution of government, scholars have traced the sources of agenda setting in the public, in the roles of interest groups and social movements, by noting the roles of policy entrepreneurs, and by looking at the government in very broad terms. Of course this does not mean that political leaders play an insignificant role. From the work of Richard Neustadt (1960) onwards students of the US Presidency have noted the need for presidents to focus their energy on a few issues (see Light, 1982; for a similar study of congressional leadership see Bader, 1996). Studies of the Supreme Court have noted the extremely tight control that the Court maintains over its agenda, as well as the characteristics of the cases that it is most likely to take. The Court, of course, is unusual among political institutions in that its agenda is reactive rather than proactive. Congress or the President can reach out to discuss whatever issues appeal to them; the Court can only choose from the issues that are presented for its decision (see Perry, 1984; Caldeira and Wright, 1988).

Recent Extensions

Agenda-setting studies have been strongly affected by the creation of the US-based Policy Agendas Project (www.policyagendas.org) which makes available comprehensive databases of government actions from 1947 through the present. Studies using these datasets have allowed new types of policy analysis. For example, Jones and Baumgartner's *Politics of Attention* (2005) shifted focus from the political agenda per se to the limits of human cognition, arguing that institutional designs capable of impressive serial processing (see Simon, 1985) were nonetheless constrained by a 'bottleneck of attention' reflected in the limited agenda of the centralized political leadership (see also Jones, 1994, 2001). This, in turn, has led to a growing literature on budgetary distributions which has shown in every case that budget changes over time are characterized by a distribution with a strong tendency toward recreating the status quo (e.g., many very small changes) at the same time as they have very

'fat tails' (e.g., many high-percentage changes; these findings have been replicated for the US federal budget, in states and municipalities, school districts, and in a number of western countries (see Breunig and Koski, 2006; Breunig et al., 2010; John and Margetts, 2003; Jones et al., 2009; Jordan, 2003; Mortensen, 2005; Robinson, 2004)). These studies have also led to analyses of how institutional decision costs (e.g., 'friction') relate to the efficiency of the resulting decisions (Baumgartner et al., 2009; Jones et al., 2003).

Studies based on the quantitative agendas approach have also led to replications in many other countries where scholars have replicated the methodology of the US agendas project. While these are too numerous and diverse to enumerate here, the comparative agendas project website (www.comparativeagendas.info) and several publications give an overview (see Green-Pedersen and Walgrave, 2014) or single country focus (see John et al., 2013).

Conclusion

In sharp contrast to two generations ago, research on political agendas is vibrant and promising today. Though much of the work has been done within the context of US politics, comparative studies have become more common (see Hogwood, 1987; Baumgartner, 1989; Reich, 1991; Zahariadis, 1995; John, 1998). New sources of quantitative data on public attitudes, government archives, and media coverage promise more systematic studies covering a greater range of issues over a longer time period than was typically possible in the past. Studies of political agendas are now firmly established as an important part of the field of political science now some 50 years after the concept was first discussed.

See also: Issue Evolution.

Bibliography

Bachrach, P., Baratz, M., 1962. The two faces of power. American Political Science Review 56, 947–952.

Bader, J.B., 1996. Taking the Initiative: Leadership Agendas in Congress and the 'Contract with America. Georgetown University Press, Washington, DC.

Baumgartner, F.R., 1989. Conflict and Rhetoric in French Policymaking. University of Pittsburgh Press, Pittsburgh, PA.

Baumgartner, F.R., Jones, B.D., 1993. Agendas and Instability in American Politics. University of Chicago Press, Chicago.

Baumgartner, F.R., Leech, B.L., 1998. Basic Interests: The Importance of Groups in Politics and in Political Science. Princeton University Press, Princeton, NJ.

Baumgartner, F.R., Breunig, C., Green-Pedersen, C., Jones, B.D., Mortensen, P.B., Neytemans, M., Walgrave, S., July 2009. Punctuated equilibrium in comparative perspective. American Journal of Political Science 53 (3), 602–619.

Bentley, A.F., 1908. The Process of Government. University of Chicago Press, Chicago.

Berry, J.M., 1989. Subgovernments, issue networks, and political conflict. In: Harris, R., Milkis, S. (Eds.), Remaking American Politics. Westview, Boulder, pp. 239–260.

Birkland, T.A., 1997. After Disaster: Agenda Setting, Public Policy, and Focusing Events. Georgetown University Press, Washington, DC.

Bosso, C.J., 1987. Pesticides and Politics: The Life Cycle of a Public Issue. University of Pittsburgh Press, Pittsburgh, PA.

Breunig, C., Koski, C., 2006. Punctuated equilibria and budgets in the American States. Policy Studies Journal 34 (3), 363–379.

Breunig, C., Koski, C., Mortensen, P.B., 2010. Stability and punctuations in public spending: a comparative study of budget functions. Journal of Public Administration Research and Theory 20 (3), 703–722.

Browne, W.P., 1998. Groups, Interests, and US Public Policy. Georgetown University Press, Washington, DC.

Caldeira, G.A., Wright, J.R., 1988. Organized interests and agenda setting in the U.S. Supreme Court. American Political Science Review 82, 1109–1127.

Cobb, R.W., Elder, C.D., 1972. Participation in American Politics: The Dynamics of Agenda Building. Allyn and Bacon, Boston.

Cobb, R.W., Ross, M.H. (Eds.), 1997. Cultural Strategies of Agenda Denial. University Press of Kansas, Lawrence, KS.

Cobb, R.W., Ross, J.-K., Ross, M.H., 1976. Agenda building as a comparative political process. American Political Science Review 70, 126–138.

Cohen, B.C., 1963. The Press and Foreign Policy. Princeton University Press, Princeton, NJ.

Crenson, M.A., 1971. The Un-politics of Air Pollution. The Johns Hopkins University Press, Baltimore, MD.

Dahl, R.A., 1956. A Preface to Democratic Theory. University of Chicago Press, Chicago.

Dahl, R.A., 1961. Who Governs? Yale University Press, New Haven, CT.

Gaventa, J., 1980. Power and Powerlessness: Quiescence and Rebellion in an Appalachian Valley. University of Illinois Press, Urbana, IL.

Green-Pedersen, C., Walgrave, S. (Eds.), 2014. Agenda Setting, Policies, and Political Systems: A Comparative Approach. University of Chicago Press, Chicago.

Haas, P.M., 1992. Introduction: epistemic communities and international policy coordination. International Organization 46, 1–35.

Hacker, J.S., 1997. The Road to Nowhere: The Genesis of President Clinton's Plan for Health Security. Princeton University Press, Princeton, NJ.

Hall, P.A., 1993. Policy paradigms, social learning, and the state: the case of economic policymaking in Britain. Comparative Politics 25, 275–296.

Heclo, H., 1974. Modern Social Policies in Britain and Sweden: From Relief to Income Maintenance. Yale University Press, New Haven, CT.

Heclo, H., 1978. Issue networks and the executive establishment. In: King, Anthony (Ed.), The New American Political System. American Enterprise Institute, Washington, DC, pp. 90–121.

Heinz, J.P., Laumann, E.O., Nelson, R.L., Salisbury, R.H., 1993. The Hollow Core: Private Interests in National Policymaking. Harvard University Press, Cambridge.

Hogwood, B.W., 1987. From Crisis to Complacency? Shaping Public Policy in Britain. Oxford University Press, New York.

Iyengar, S., 1991. Is Anyone Responsible? How Television Frames Political Issues. University of Chicago Press, Chicago.

Iyengar, S., Kinder, D.R., 1987. News that Matters: Television and American Opinion. University of Chicago Press, Chicago.

John, P., 1998. Analyzing Public Policy. Pinter, London.

John, P., Bertelli, A., Jennings, W., Bevan, S., 2013. Policy Agendas in British Politics. Palgrave Macmillan, Basingstroke.

John, P., Margetts, H., 2003. Policy punctuations in the UK. Public Administration 81, 411–432.

Jones, B.D., Sulkin, T., Larsen, H., 2003. Policy punctuations in American political institutions. American Political Science Review 97, 151–170.

Jones, B.D., 1994. Reconceiving Decision-Making in Democratic Politics: Attention, Choice, and Public Policy. University of Chicago Press, Chicago.

Jones, B.D., 2001. Politics and the Architecture of Choice. University of Chicago Press, Chicago.

Jones, B.D., Baumgartner, F.R., 2005. The Politics of Attention: How Government Prioritizes Problems. University of Chicago Press, Chicago.

Jones, B.D., Baumgartner, F.R., Breunig, C., Wlezien, C., Soroka, S., Foucault, M., François, A., Green-Pedersen, C., John, P., Koski, C., Mortensen, P.B., Varone, F., Walgrave, S., October 2009. A general empirical law for public budgets: a comparative analysis. American Journal of Political Science 53 (4), 855–873.

Jordan, M., 2003. Punctuations and agendas. Journal of Policy Analysis and Management 22, 345–360.

Kingdon, J.W., 1984. Agendas, Alternatives, and Public Policies. Litttle Brown, Boston.

Light, P.C., 1982. The President's Agenda. The Johns Hopkins University Press, Baltimore, MD.

Lowi, T.J., July 1964. American business, public policy, case-studies, and political theory. World Politics, 677–715.

Lowi, T.J., 1969. The End of Liberalism. Norton, New York.

McAdam, D., 1988. Freedom Summer. Oxford University Press, New York.

McCombs, M.E., Shaw, D.L., 1972. The agenda-setting function of the mass media. Public Opinion Quarterly 36, 176–187.

Mortensen, P.B., 2005. Policy punctuations in Danish local budgeting. Public Administration 83 (4), 931–950.
Nelson, B.J., 1984. Making an Issue of Child Abuse. University of Chicago Press, Chicago.
Neustadt, R.E., 1960. Presidential Power. John Wiley and Sons, New York.
Perry Jr., H.W., 1984. Deciding to Decide: Agenda-Setting on the US Supreme Court. Harvard University Press, Cambridge, MA.
Peterson, M.A., 1993. Political influence in the 1990s: from iron triangles to policy networks. Journal of Health Politics, Policy and Law 18, 395–438.
Polsby, N.W., 1980. Community Power and Political Theory, second ed. Yale University Press, New Haven, CT.
Polsby, N.W., 1984. Political Innovation in America: The Politics of Policy Initiation. Yale University Press, New Haven, CT.
Reich, M.R., 1991. Toxic Politics: Responding to Chemical Disasters. Cornell University Press, Ithaca, NY.
Riker, W.H., 1986. The Art of Political Manipulation. Yale University Press, New Haven, CT.
Riker, W.H., 1988. Liberalism against Populism. Waveland Press, Prospect Heights, IL.
Riker, W.H., 1996. The Strategy of Rhetoric. Yale University Press, New Haven, CT.
Riker, W.H. (Ed.), 1993. Agenda Formation. The University of Michigan Press, Ann Arbor, MI.
Robinson, S., 2004. Punctuated equilibrium, bureaucratization, and budgetary changes in schools. Policy Studies Journal 32 (1), 25–39.
Rochefort, D.W., Cobb, R.W. (Eds.), 1994. The Politics of Problem Definition: Shaping the Policy Agenda. University Press of Kansas, Lawrence, KS.
Rochon, T.R., 1998. Culture Moves. Princeton University Press, Princeton, NJ.
Rogers, E.M., Dearing, J.W., 1988. Agenda-setting research: where has it been, where is it going?. In: Anderson, J.A. (Ed.), Communication Yearbook, 11. Sage, Newbury Park, CA, pp. 555–594.
Schattschneider, E.E., 1960. The Semi-Sovereign People. Holt, Rinehart and Winston, New York.
Shepsle, K.A., 1978. The Giant Jigsaw Puzzle. University of Chicago Press, Chicago.
Sikkink, K.A., Keck, M.E., 1999. Activists beyond Borders. Cornell University Press, Ithaca.
Simon, H.A., 1985. Human nature in politics: the dialogue of psychology with political science. American Political Science Review 79, 293–304.
Stimson, J.A., 1991. Public Opinion in America: Moods, Cycles, and Swings. Westview Press, Boulder, CO.
Stone, D.A., 1988. Policy Paradox and Political Reason. Scott, Foresman, Glenview, IL.
Truman, D.B., 1951. The Governmental Process: Political Interests and Public Opinion, first ed. Alfred A. Knopf, New York.
Walker, J., 1977. Setting the agenda in the U.S. Senate. British Journal of Political Science 7, 423–445.
Zahariadis, N., 1995. Markets, States and Public Policy: Privatization in Britain and France. University of Michigan Press, Ann Arbor, MI.

Agent Based Modeling, Statistics of

David Banks, Duke University, Durham, NC, USA
Jacob Norton, North Carolina State University, Raleigh, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Agent-based models (ABMs) have become an important simulation tool for understanding certain categories of complex phenomena. They are widely used in epidemiology, ecology, transportation research, social networks, and other applications in which the global behavior is determined by local behavior (which is usually quite simple, and can be represented through a concise set of rules). However, unlike many other models, such as linear regression, the statistical issues associated with ABMs are largely undeveloped.

Introduction

Agent-based models (ABMs) are a simulation strategy that is especially useful when the phenomenon of interest is complex, not continuously dependent on underlying parameters, and can be described in terms of local actions. In Europe and in some academic disciplines, such as ecology, ABMs are commonly referred to as 'individual-based models.'

Popular applications of ABMs include:

- Weather forecasting, in which each agent is a cubic kilometer of atmosphere, and the local interactions are the exchange of pressure, temperature, and moisture (cannot find this one).
- Auctions, as in Yahoo! or Google, to determine which ads are shown to users (Charles et al., 2013).
- Traffic flow models, as in TRANSIMS, where agents (drivers) space themselves according to the actions of other nearby drivers, and make route choices based on congestion avoidance (Smith et al., 1995).
- Genetic algorithms, in which the agents are primitive algorithms, which interact so as to evolve more successful algorithms (Chatterjee et al., 1996).

ABMs are commonly used in epidemiology, economics, social networks, and many other fields. Often the primary question of interest concerns emergent behavior generated by the cumulative actions of distinct entities, each making choices that satisfy its own requirements.

As a simple example to fix the basic concepts, assume that a safety engineer wants to determine how long it would take to evacuate an office building. The ABM approach would start with a virtual representation of the building, with the rooms and doors and stairs, and then place a typical number of people (agents) at random within the building. Each agent would follow two rules: (1) When the firebell rings, go to the nearest exit; and (2) If the exit is blocked, go to the closest exit that is unblocked. There would be additional constraints on how much crowding is possible in, for example, a stairwell or doorway. In this framework, the safety engineer might run the simulator 100 times, and create a histogram of how long was needed to complete the evacuation. One particularly attractive feature of the ABM approach is that the engineer can play 'what if?' scenarios, such as blocking certain stairwells or adding agents who act as fire marshals and direct the traffic. A second attractive feature is that the engineer can examine the rule set to see if it makes sense – perhaps something more complicated is needed, such as allowing agents who are away from their office to return and pick up their laptops before exiting.

This example indicates a number of features that are common to most ABMs. First, there is a 'geography' in which the agents operate. In this example, it was the virtual building; in TRANSIMS it is the road network; and with weather forecasting it is a three-dimensional grid of cubic kilometers. However, geography is not essential – the auction example and the genetic algorithms example do not require one.

Additionally, there are agents, which follow specific rules. The agents may be of multiple types, such as the employees and the fire marshals in the building evacuation example. Alternatively, the agents may be identical up to a prespecified parameter; for example, in virtual auctions different bidders respond to different keywords, have different amounts of money, and so forth, but all have the same objective.

The rule sets are the most critical common element. The rules are usually simple, but very flexible, and can be readily examined for plausibility. In particular, the rules may allow very unsmooth simulations; for example, in the building evacuation example, an agent may pick among separate exit paths.

Often, an ABM evolves over time, and there is interest in some ensemble behavior that is determined by the interaction of the agents. In the evacuation example, the ensemble behavior is how long it takes to empty the building, and this depends on crowding and the decisions of the agents in the building who only have local information about which exits are blocked. In the TRANSIMS example, the relevant behavior may be some measure of congestion, and in the weather forecasting application it may be temperature or rainfall prediction.

The following sections will elaborate on these issues. Section The History of ABMs traces the history of ABMs, with special emphasis on three influential applications. Section Limitations of ABMs lays out the research challenges in the ABM methodology, and points up some of the ways in which these are being addressed.

The History of ABMs

ABMs grew up with the era of modern computing. Their inception is rooted in cellular automata, a special case of ABMs

invented by John von Neumann and Stanislaw Ulam while pioneering the computing era at Los Alamos in the 1940s (Wolfram, 2002).

Cellular automata are ABMs in which the geography is a grid, and the grid points or the cells formed by the grid interact according to well-chosen rules. The most famous of these is Conway's Game of Life (Gardner, 1970), but the field is rich.

Stephen Wolfram's *A New Kind of Science* is an extensive survey of cellular automata, and argues that these represent a critical frontier for science.

Cellular automata have been generalized in many ways. In probability, interacting particle systems extend cellular automata to situations in which time is continuous rather than discrete and randomness plays a larger role (Liggett, 1985). But the main development has been increasingly rich characterizations of the cells and the actions that are taken; instead of cells in a space that are assigned colors according to some rule set, they have become agents, whose locations may change, rules may evolve, and which can respond adaptively to their environment.

The following subsections trace the historical growth in the sophistication of ABMs through three highly influential examples. Kauffman's model introduced agents that interact through a random network rather than a grid, and led to new insights in biology and new mathematical problems. Artificial societies created by Epstein and Axtell opened the door to a wide range of ABM applications in the social sciences, and the analysis of the spread of rabies in raccoon populations introduced the use of hierarchical statistical models in ABMs.

Kauffman's Random Networks

Stanley Kauffman wanted to understand how the same DNA could produce all of the different tissue types found in organisms. To study this, he used an ABM described in his seminal paper, 'Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets' (Kauffman, 1969).

In Kauffman's model, each agent is a gene. Each gene is either off or on, 0 or 1, depending upon the inputs it receives from other genes. Genes receive inputs from other genes, calculate a Boolean function, and send the result to other genes. In terms of the previous framework, the rule set is the Boolean function or functions, and the geography is the network through which inputs are received and outputs are transmitted.

In Kauffman's ABM, a large number (n) of agents are connected in a network at random, subject to the constraint that each agent must receive inputs from k agents (possibly including itself) and send output to k agents. Given its inputs, each agent uses a Boolean function to produce its binary output. Different agents may have different Boolean functions, and Kauffman assigned those functions to agents at random (excluding two degenerate functions, which always produce 0 or 1). The property of interest was the number of stable cycles that such a system could produce.

If a gene receives k inputs, then there are 2^k possible vectors of inputs (since each of the k inputs may be 0 or 1). And for any given vector of inputs, there are two possible outputs, 0 or 1. So the number of possible functions that map $\{0,1\}^k \rightarrow \{0,1\}$ is 2^{2^k}. Table 1 shows 3 of the 16 possible Boolean functions when $k = 2$.

Table 1 The first table corresponds to Boolean operator AND, the second is OR, and the last is tautology, one of the two degenerate operators that Kauffman excluded. The operator names derive from truth tables used in formal logic

Input		Output
0	0	0
0	1	0
1	0	0
1	1	1

Input		Output
0	0	0
0	1	1
1	1	1
1	0	1

Input		Output
0	0	1
0	1	1
1	0	1
1	1	1

Figure 1 shows how Kauffman's model enabled study of the number of stable cycles in a randomly connected Boolean network. The first panel shows five randomly connected agents (nodes). The agents are either OR or AND operators, as indicated. The second panel shows the transitions between states for the largest component; i.e., if the five agents are initialized as (0, 1, 0, 0, 1) (i.e., the value in the lowest corner of the panel), then the next state is (1, 0, 0, 1, 1), then (1, 0, 1, 1, 1), then (1, 1, 1, 1, 1). This last state is an absorbing state; if the system reaches that state, it does not leave that state. The last panel shows the three other components that are possible. If the system starts in the (0, 1, 0, 0, 0) state, then it evolves to the absorbing state (0, 0, 0, 0, 0). If it starts in the (0, 0, 1, 0, 1) state, it evolves to a stable cycle with three states: (1, 0, 0, 0, 0) going to (0, 0, 0, 1, 0) going to (0, 0, 0, 0, 1) which returns to (1, 0, 0, 0, 0), and so forth.

Epstein and Axtell's Artificial Societies

Epstein and Axtell (1996) brought the ABM perspective strongly into the field of social science. Specifically, they showed that simple rules could lead agents to display many of the complex behaviors found in human societies, including population dynamics, hunter-gatherer migration, division of labor, and a barter economy.

The Epstein and Axtell ABM was based on a planar lattice, which they called a 'sugarscape.' At each intersection of the lattice a resource, 'sugar,' grew at a constant rate. Initially, a fixed number of agents were placed randomly at the intersections of the lattice, with the rule that they were to consume sugar (at a constant rate greater than sugar's growth rate) until the supply was exhausted, and then move to the nearby lattice point and continue consuming. The result was that agents tended to move in large circles whose circumference matched the rate of growth of sugar, so that when the agent returned to its starting point, the sugar was fully replenished. This mirrors the migratory patterns of hunter-gatherer societies, in which phenology drives movement. Figure 2 shows one time point in a sugarscape simulation.

Next, Epstein and Axtell added gender and reproduction. When there was sufficient food and agents of opposite gender were on adjacent lattice intersections, they would have a child. This led to population pyramids, carrying capacity limits to growth, and many other features found in population dynamics. If the rules were extended so that families preferred to stay near each other, tribalism emerged.

Additional rules allowed pollution, diffusion of pollution, accumulation of wealth, the evolution of genetic traits, the spread of disease, specialized labor, cultural tags (memes) that could be shared or defended, trade, and combat. In all, 17 rules

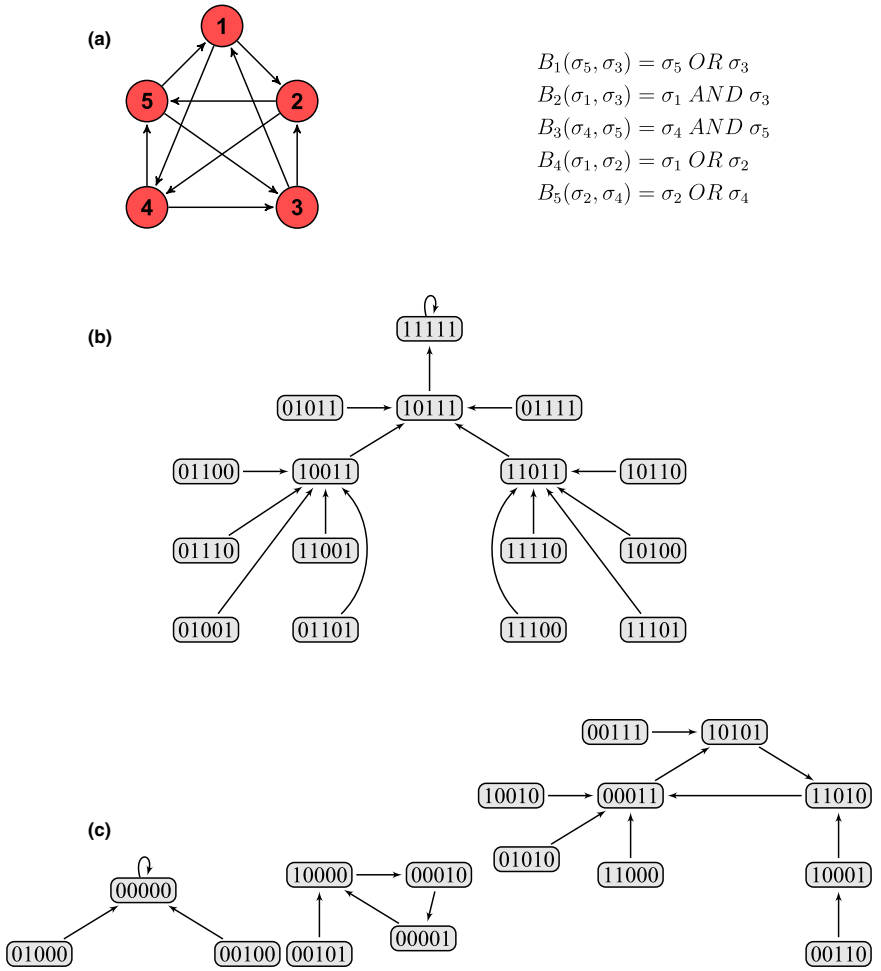

Figure 1 This figure shows how the randomly connected Boolean network, under various initializations, transitions to different stable behaviors, where each stable behavior is an absorbing state or cycle. (a) A network of five agents (genes) in which each agent receives input from two agents and transmits its output to two agents. (b) The largest confluent of states. (c) The remaining three state confluents. Notice that two are, at their center, state cycles.

were sufficient to produce a rich range of social behavior. To illustrate these rules, consider three of them (taken verbatim from Epstein and Axtell, 1996):

1. *Sugarscape growback*: At each lattice position, sugar grows back at a rate of α per time interval up to the capacity of that position.
2. *Agent movement*: Look out as far as vision permits in each of the four lattice directions, north, south, east, and west:
   a. Considering only unoccupied lattice positions, find the nearest position producing maximum welfare;
   b. Move to the new position;
   c. Collect all the resources at that location.
3. Agent mating:
   a. Select a neighboring agent at random;
   b. If the neighboring agent is of the opposite sex *and* if both agents are fertile *and* at least one of the agents has an empty neighboring site *then* a newborn is produced by crossing over the parents' genetic and cultural characteristics;
   c. Repeat for all neighbors.

Note that the first rule includes a tunable parameter; there are many such cases in the full list; and this is common in ABMs in general.

The sugarscape rules are simple to program and easily interpretable in the context of the model. But they do not lend themselves to mathematical expression or analysis.

Rabies in Raccoons

Hooten and Wikle (2010) introduced Bayesian hierarchical models into ABM research. Their technique does not apply to all of the very wide range of ABM formulations, but it is useful for spatiotemporal processes with fairly simple structure.

Their motivating application is the spread of rabies in raccoon populations in Connecticut between 1991 and 1995. On a gridded map representing the townships in the state of Connecticut, they represented the presence or absence of rabies by a binary random variable whose distribution depended upon the states in the neighboring townships at the preceding time period, as well as covariates (which could also vary in time).

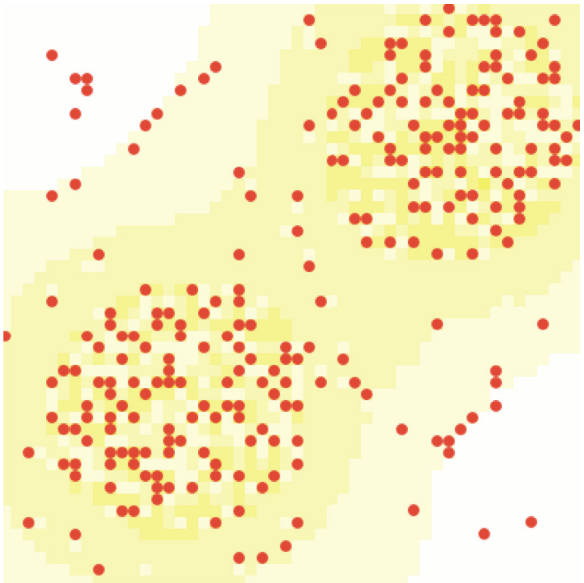

Figure 2 A snapshot of the Sugarscape model described in Epstein, J., Axtell, R., 1996. Growing Artificial Societies: Social Science from the Bottom up. Brookings Institution Press, MIT Press, Cambridge/Washington, DC. Agent locations reflect geographical variation in the rate of growth of 'sugar.' The image was captured from the NetLogo model described in Li, J., Wilensky, U., 2009. NetLogo Sugarscape 3 Wealth Distribution Model. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL. http://ccl.northwestern.edu/netlogo/models/Sugarscape3WealthDistribution.

Let $\boldsymbol{u}_t = (u(1,t), u(2,t), \ldots, u(m,t))'$ denote the binary vector showing the presence or absence of rabies at time t for each of the m townships and let $\boldsymbol{X}_t = (\boldsymbol{x}(1,t), \boldsymbol{x}(2,t), \ldots, \boldsymbol{x}(m,t))$ denote a matrix of corresponding covariates, such as population density, adjacency to the Connecticut River, and so forth. Define the neighborhood for township i by N_i; this is a set of townships. Then the basic model for the spread of the disease is

$$u_{i,t} \big| u_{N_i,t-1} \sim \left[u_{i,t} \big| h\left(u_{N_i,t-1}, x_{N_i,t-1}\right) \right]$$

where $h(\cdot,\cdot)$ is a very general updating function, the subscript N_i indicates the townships relevant to the disease spread at township i (i.e., its neighboring townships), and the bracket notation indicates that the presence or absence of rabies is a random variable with parameters that depend on the conditioning within the bracket. The only substantive difference between this model and a Gaussian state space model is that the random variables need not be Gaussian (which generally precludes closed-form solution, putting this in the realm of ABM simulation).

This model is flexible, and enables disease spread to be anisotropic (i.e., directional, e.g., along the Connecticut River). It enables probabilistic statements about the posterior probability of disease in a particular township, but usually requires Markov chain Monte Carlo (cf Robert and Casella, 2004) to evaluate. It does not apply to all ABMs (e.g., genetic algorithms or the evacuation of a building), but when it does apply, it permits more explicitly statistical inference on the behavior of the ABM.

Limitations of ABMs

ABMs are popular and will be used for the foreseeable future. The primary reasons for this popularity are that they are relatively simple to code and straightforward to validate, at the least at a basic level. Not all problems are amenable to ABM representation, but for those that are, ABMs are generally easier to conceptualize and communicate than models based on complex stochastic processes or other mathematical representations.

Nonetheless, ABMs are problematic, because there is no robust theory of statistical inference. An ABM is a model, just as a linear regression is a model. Statisticians know how to fit linear models, how to assess fit, how to make predictions from linear models with quantified uncertainty, and so forth. But there is virtually no principled theory yet for ABMs.

Verification and Validation

Verification pertains to determining whether the code in an ABM is error-free. Validation asks whether the ABM is sufficiently faithful to reality. Verification lies outside the scope of this article, other than to acknowledge that it is a significant problem, especially in complex ABMs that may entail many thousands of lines of code.

Regarding validation, there are different approaches. Some of these are more rigorous than others, and some of these apply better to some situations than others (cf Louie and Carley, 2008). But none of these is fully satisfactory.

Physics-based validation is commonly used in the hard sciences. One builds a simulation that incorporates all of the physical laws and interactions that are appropriate, and then feels confident that their model is faithful to reality. This can work on smaller problems where all mechanisms are fully understood. Examples, where it is arguably successful, include planetary motions (cf Miller and Page, Section 6.6., 2007), flight simulators, and perhaps virtual mock-ups of semiconductor manufacturing processes. But it tends to break down as the stochasticity and complexity of the problem increase. Also, physics-based modeling often constructs the process in more detail that is actually required for reasonable fidelity, and thus takes a very long time to run.

A second approach might be termed intelligent design. This is the most common validation protocol, and is probably used in all but the most critical applications. Domain experts think through the simulation carefully, building in or approximating all the effects, which they think are substantial. Then they hope that they have been smart enough and wait to see if their results work well enough. Intelligent design can handle more complicated situations than the physics-based models, but it is not a true validation. The review process is more like a careful check of the thinking behind the model.

Face validity is a true validation protocol. The designer tests the ABM by using selected inputs to explore the output behavior. Often the selected inputs are chosen according to a statistical design, such as a Latin hypercube (cite), which increases efficiency greatly when the dimension of the input space is not too large. Alternatively, the designer can select values for the inputs that correspond to expected behaviors or regions in which predictive accuracy is especially important. Face validation is insufficient when the parameter space is large

and there are many interactions. But, to varying degrees, it is used for systems such as TRANSIMS and the battlefield simulations produced at the Defense Modeling and Simulation Office (Davis and Anderson, 2004).

A stronger validation protocol is based upon comparison to another, independently derived, model. This is not done often enough, but it has the potential to be a powerful tool. The advantages are that one can better explore the full range of model behavior. The disadvantage is that it often requires duplication of effort and a great deal of more development expense. But sometimes different scientific teams have developed different approaches to the same problem, as in the weather forecasting example, for which stochastic partial differential equation approximations are also used. In that case, comparison of the models can highlight the strengths and deficiencies of both.

The strongest form of validation occurs when one compares ABM outputs to the historical record for the phenomenon of interest. In principle, this is possible for weather forecasting, epidemic modeling, and TRANSIMS. However, all of these examples are noisy systems, so one expects divergence between the ABM prediction and the actual historical data. One must decide whether the error is unbiased and its variance matches the application of interest, and this requires a great deal of historical data and many runs of the ABM. Although comparison to real-world data is the strongest form of validation, it is still inadequate. One does not have confidence in the fidelity of the simulation in regimes that have not been previously observed, and this is often the context of greatest interest.

Inference

Given an arbitrary ABM, there is no clearly formulated inferential procedure (as is available, say, for a linear regression model). One would like to determine how to tune the parameters in an ABM to fit a given data set, or how to decide which covariates are actually important to the behavior of interest in the ABM.

There are two possible strategies for improving statistical inference in situations for which one cannot write out the likelihood function. The first is based upon emulators. These are Gaussian process approximations to complex systems, where Bayesian methods allow one to combine real-world data with multiple runs of the ABM to estimate tuning functions that provide the best possible fit, and to identify regions of the input space for which the emulator offers a poor approximation to the ABM. Emulators were proposed by Kennedy and O'Hagan (2001), and have been subsequently elaborated by many researchers.

The second possible strategy is Approximate Bayesian Computation (ABC). The method was conceptually proposed by Rubin (1984), but realized in its modern form by Tavaré et al. (1997). ABC starts with a prior over the parameter space of the ABM. It generates a realization of those parameters, runs the ABM, and produces a simulated data set. That data set is compared to real-world data; if it is close with respect some metric appropriate to the research domain, then the random parameters that generated the sample are accepted, and that point in the parameter space has increased posterior probability. The ABC process repeats until one has an estimate of posterior density function.

Both emulators and ABCs are ongoing areas of research, and their strengths and weaknesses, especially in the context of high-dimensional applications, are not fully understood. In particular, it is not known how to decide which of these is most useful in a particular ABM application.

Distances between Models

A third issue in the statistics of ABMs concerns the comparison of two models, when at least one of the models is an ABM. For example, consider the problem of estimating the spread of an epidemic. One person might build an ABM that included a geography based on a city network, where agents have rule sets that move them around the city, and when an infected agent meets an uninfected agent, there is a chance of transmitting the disease. But a second person might create a similar ABM, but with more and different detail, such as higher transmission rates in day care centers or periodic crowding such as church services or smaller time steps that allow more opportunity for people to meet. There is no clear strategy for deciding what degree of elaboration is needed, nor when one model is a proper subset of another. This is complicated by the fact that the input parameters for one model may be the same as, entirely different from, or partially overlapping the input sets for the other.

A similar issue arises when deciding between an ABM and a differential equation model. For example, again in the context of epidemiology, a mathematician might be tempted to use a Kermack–McKendrick model (Kermack and McKendrick, 1927), in which the change in the numbers of susceptibles (S), infected (I), and recovered (R) is described by a system of coupled differential equations:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI \quad \frac{\mathrm{d}I}{\mathrm{d}t} = \beta SI - \gamma I \quad \frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I$$

Qualitatively, the differential equations should produce similar dynamics to those obtained from an elaborate ABM. But there is no formal procedure for deciding how close these two models are, nor whether one is substantially better than the other in terms of the emergent behavior of interest. Bagni et al. (2002) discuss comparison of such epidemiolgical models in more detail.

See also: Hierarchical Models: Random and Fixed Effects; Social Simulation: Computational Models.

Bibliography

Bagni, R., Berchi, R., Cariello, P., 2002. A comparison of simulation models applied to epidemics. Journal of Artificial Societies and Social Simulation 5 (3).

Charles, D., Chakrabarty, D., Chickering, M., Devanur, N.R., Wang, L., 2013. Budget smoothing for internet ad auctions: a game theoretic approach. In: Proceedings of the Fourteenth ACM Conference on Electronic Commerce. ACM, New York, pp. 163–180.

Chatterjee, S., Laudato, M., Lynch, L.A., 1996. Genetic algorithms and their statistical applications: an introduction. Computational Statistics & Data Analysis 22 (6), 633–651.

Davis, P., Anderson, R., 2004. Improving the composability of DoD models and simulations. The Journal of Defense Modeling and Simulation: Applicatins, Methodology, Technology 1, 5–17.

Epstein, J., Axtell, R., 1996. Growing Artificial Societies: Social Science from the Bottom up. Brookings Institution Press, MIT Press, Cambridge/Washington, DC.

Gardner, M., 1970. Mathematical games the fantastic combinations of John Conway's new solitaire game"life". Scientific American 223, 120–123.

Hooten, M., Wikle, C., 2010. Statistical agent-based models for discrete spatio-temporal systems. Journal of the American Statistical Association 105, 236–248.

Kauffman, S.A., 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of Theoretical Biology 22, 437–467.

Kennedy, M., O'Hagan, A., 2001. Bayesian calibration of computer models. Journal of the Royal Statistical Society, Series B 63, 425–464.

Kermack, W.O., McKendrick, A.G., 1927. A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society A 115, 700–720.

Li, J., Wilensky, U., 2009. NetLogo Sugarscape 3 Wealth Distribution Model. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL. http://ccl.northwestern.edu/netlogo/models/Sugarscape3WealthDistribution.

Liggett, T.M., 1985. Interacting Particle Systems. Springer, New York.

Louie, M., Carley, K., 2008. Balancing the criticisms: validating multi-agent models of social systems. Simulation Modeling: Practice and Theory 16, 242–256.

Miller, J.H., Page, S.E., 2007. Complex Adaptive Systems. Princeton University Press, Princeton, NJ.

Robert, C., Casella, G., 2004. Monte Carlo Statistical Methods, second ed. Springer-Verlag, New York.

Rubin, D.B., 1984. Bayesianly justifiable and relevant frequency calculations for the applies statistician. Annals of Statistics 12, 1151–1172.

Smith, L., Beckman, R., Baggerly, K., 1995. TRANSIMS: Transportation Analysis and Simulation System (No. LA-UR–95–1641). Los Alamos National Lab, NM.

Tavaré, S., Balding, D.J., Griffiths, R.C., Donnelly, P., 1997. Inferring coalescence times from DNA sequence data. Genetics 145, 505–518.

Wolfram, S., 2002. A New Kind of Science, vol. 5. Wolfram Media, Champaign, IL, p. 876.

Wolters, B., Steffens, T., 2008. Learning agent-behavior for agent-based simulation using genetic algorithms. In: Proceedings of the European Simulation and Modeling Conference 2008. Le Havre, France, pp. 284–288.

Aggression, Social Psychology of

Wayne A Warburton, Macquarie University, NSW, Australia
Craig A Anderson, Iowa State University, Ames, IA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by L. Berkowitz, volume 1, pp. 295–299, © 2001, Elsevier Ltd.

Abstract

For over seven decades social psychological theories advanced understanding of aggressive behavior. The most recent major model – the General Aggression Model (GAM) – integrates prior theories, thereby encompassing the broadest range of aggressive phenomena. GAM is built on research about factors within a person that predispose them to aggression; factors from the environment that trigger aggression; and the underlying biological, neurocognitive, and psychological processes. This article summarizes historical and modern social psychological theories of aggression, key research methodologies and findings, and challenges of studying violence and aggression in society. It concludes by noting areas for future social psychological research of aggression.

Human aggression is a social behavior, and whilst it has been studied from many perspectives, it is theoretical models and empirical research from the field of social psychology that have provided the strongest framework from which to understand it. This article focuses on the contribution of social psychologists to the understanding of human aggression, providing first some key definitions, then major theories (both classic and contemporary) and a brief summary of social psychological approaches to the study of aggressive behavior. An overview of research findings is presented, including those describing factors within a person that increase the likelihood they will aggress, situational cues that can trigger aggression, internal psychological processes that underlie an instance of aggressive behavior, and processes that increase trait aggressiveness. We conclude by suggesting a 'risk factor' framework for understanding societal violence and noting directions for future research.

Definitions and Characteristics of Aggression

Definitions

There are three key issues with defining human aggression. First, it is hard to interpret research findings and theories about aggression without a clear definition. Historically, however, many different definitions have been used. As a result, many studies of aggressive behavior are hard to meaningfully compare. More recently, definitions of aggression among social psychologists have converged around the notion that aggression is any behavior enacted with the intention to harm another person who is motivated to avoid that harm (e.g., Anderson and Bushman, 2002; Bushman and Huesmann, 2010). Such a definition is wide enough to capture the full range of aggressive behaviors, and to make allowance for activities that can 'hurt' a target person but to which the target of the hurt willingly consents (such as undergoing surgery or engaging in sadomasochistic sex).

The second issue is that many laypersons and misinformed professionals use the term aggression interchangeably with related but conceptually distinct phenomena such as anger, hostility and competitiveness. There is no question that in the field of psychology, aggression refers only to a behavior, and not to a mindset or an emotional state. Feelings such as anger, attitudes such as wishing the worst for another, and motivations such as the desire to win or control one's environment may contribute to a person behaving aggressively but are not aggression per se. To study aggression effectively, such factors need to be clearly differentiated from aggression and from each other.

A third definitional issue involves the common practice of using the term 'violence' interchangeably with the term 'aggression.' Treating these as synonymous creates miscommunications and confusion among researchers, public policy-makers, and the general public. Among most social psychologists, violence is a subtype of aggression. More precisely, 'violence' is aggression that is intended to cause harm extreme enough to require medical attention or to cause death. Many social psychologists extend this definition to include causing severe emotional harm. Thus, all violent behavior is aggression, but most aggression is not violence. Note that this definition of violence is not synonymous with 'violent crime,' which is a legal term, not a scientific one.

Types and Characteristics of Aggression

When considering the many ways in which one human can harm another, it is useful to distinguish between different forms of aggression, and between the different functions that aggression can perform. Different forms of aggression include physically harming another (i.e., *physical aggression* such as hitting, biting, kicking, clubbing, stabbing, shooting), hurting another with spoken words (i.e., *verbal aggression* such as yelling, screaming, swearing, name calling), or hurting another's reputation or friendships through what is said to others verbally or digitally (i.e., *relational aggression*). Aggression may also be *direct* (with the victim physically present) or *indirect* (enacted in the absence of the victim; for example, smashing someone's property or spreading rumors about them).

Aggression also differs by function. It may involve a relatively pure intent to punish/hurt the target person, as in

reacting aggressively to provocation (i.e., *reactive, affective, hostile, hot, impulsive*, or *retaliatory* aggression) or it may involve a considered and deliberate plan to harm another to gain a desired outcome (i.e., *instrumental, proactive, planned*, or *cold* aggression). Aggression may be an *automatic* response driven by hard-wired self-protection mechanisms (e.g., fight or flight) or involve a script for aggressive behavior that is so commonly enacted that the response is no longer thought-through. Of course such distinctions can be problematic. What about a person whose rage drives them to carefully plan the death of another? Such instances do not fit any of these traditional categorical or dichotomous distinctions. A viable alternative approach to understanding the function of aggression is to locate aggressive acts on three dimensions – the degree to which the goal is to harm the victim versus benefit the perpetrator; the level of hostile or agitated emotion that is present; and the degree to which the aggressive act was thought-through (Anderson and Huesmann, 2003).

Social Psychological Theories of Aggression

Theories Outside Social Psychology

This article is focused on the social psychology of aggression, but must be considered as complementing research from other spheres of psychology. Most notably, biological psychology provides many relevant findings, including links to genetic predispositions, hormones, malformation, or damage of brain structures and levels of cortical and nervous system arousal. Psychodynamic approaches and animal psychology have emphasized aggressive drives, and evolutionary- and animal psychology have focused on aggression in terms of factors related to reproductive success and survival (e.g., dominance and resource-holding potential). Social psychological approaches have tended to include such biological, genetic and personality factors as 'person' factors in their models, but typically have not explored detailed interrelationships among these and related social factors.

Early Social Psychological Theories

For more than 70 years, social psychology has provided a variety of frameworks from which hypotheses about the causes and consequences of aggression could be derived and tested. These theories, although distinct, have also tended to overlap as new knowledge has extended an existing framework of aggressive behavior. The earliest influential theory from social psychology was the frustration-aggression hypothesis.

The Frustration-Aggression Hypothesis

During 1939, partially in response to World War II and partially in response to the spreading influence of psychodynamic theories in the US, Dollard et al. (1939) proposed the first systematic theory of aggression. Using assumptions from psychoanalytic theory, they focused on the frustration caused when a goal is blocked, and suggested that "the occurrence of aggressive behavior always presupposes the existence of frustration," and that "the existence of frustration always leads to some form of aggression" (p. 1). Although this theory enjoyed some empirical support, it quickly became obvious that frustration does not always lead to aggression, and that not every act of aggression can be traced back to frustration. Frustration-aggression theory was revised to incorporate the possibility that frustrations can elicit responses other than aggression (e.g., to escape or to find another way to achieve a goal), and that the inclination which will be acted upon is the one that best reduces frustration. In this revised formulation, people learn through experience to respond to frustrations with aggressive or nonaggressive responses.

Learning Theories

The earliest theory of learning in modern psychology explains behavior in terms of *classical conditioning* – learning to associate one thing with another. Pioneered by Pavlov, this approach suggests that once people mentally pair things together, they become 'conditioned' to expect those things to always occur together. This theory was later supplemented with theories of *operant conditioning* developed by Thorndike and Skinner, which suggest that people are more likely to repeat a behavior that has been rewarded and less likely to repeat a behavior that has been punished. In aggression research it has been shown that children can be taught to behave aggressively through rewarding aggressive behavior (*positive reinforcement*) or removing a painful consequence after aggression (*negative reinforcement*). In addition, children learn to discriminate between situations where aggression has a desirable consequence and when it does not, and to generalize this knowledge to new situations. Although such research demonstrates that aggression can be learned through conditioning (e.g., Eron et al., 1971), it was clear by the 1960s that such processes could not explain the acquisition of all learned aggression.

Bandura proposed that social behaviors, including aggression, could be learned through observing and imitating others (i.e., via *observational learning*). In his classic experiments, children observed a film of an actor hitting a 'Bobo Doll' in several novel ways. The children later imitated the behavior in the absence of any classical or operant conditioning. Bandura also developed the concept of *vicarious learning* of aggression, and showed that children were especially likely to imitate models that had been rewarded for behaving aggressively. In *social learning theory* (later called *social cognitive theory*), Bandura hypothesized that the way people mentally construct their experiences is crucial. People may see one person hit another, but will also decide how competent they feel to do the same, and will make assumptions about what constitutes a normal way to respond when someone provokes you. In this way, making inferences about observed aggression not only increases the likelihood of imitating it, but also expands the range of situations to which that response might be generalized (see Bandura, 1986). There is considerable research support for social cognitive explanations of aggression. People sometimes imitate aggressive models, especially if the aggressive behavior is rewarded or carried out by a person who is heroic, admired, of high status, attractive, or similar.

Arousal: Cognitive Labeling and Excitation Transfer

The emergence of cognitive psychology inspired a plethora of new approaches to aggression by social psychologists. Early in this period, researchers explored the way people make meaning of physiological arousal, a known precursor to aggression.

Researchers such as Schacter found that when people are aroused, they look for cues in the environment to help them attribute the cause of their arousal. For example, Schacter and colleagues found that if aroused people were exposed to another person who was angry, they tended to *cognitively label* their arousal as being angry themselves. Zillmann (1979) extended this concept with *excitation-transfer theory* (ETT). Physiological arousal, however produced, dissipates slowly. ETT posits that if two arousing events are separated by a short amount of time, arousal from the first event will add to arousal from the second. However, the cognitive label given to the second event will be misattributed as being relevant to all of the arousal experienced, thus producing an inappropriately strong response (e.g., becoming angry to a level far greater than might be expected for a minor provocation). Because the cognitive label (or attribution) is crucial in determining behavior, strong anger related to excitation transfer may persist long after the arousal itself has dissipated.

Mainstream Cognitive Theories

Information Processing and Script Theories

The confluence of computer availability and the growing dominance of cognitive approaches to psychology in the 1980s heralded a major change of direction in social psychological aggression research. For the first time, researchers started to conceptualize the acquisition of social behavior in terms of computerlike processes – inputs, outputs, and the processing of information. Two key theories of aggression emerged – the Social Information Processing (SIP) theory of Dodge (1980) and Script theory from Huesmann (1982). SIP theory emphasized the way people perceive the behavior of others and make attributions about their motives. A key construct in SIP theory is the *hostile attributional bias* – a tendency to interpret ambiguous events (such as being bumped in a corridor) as being motivated by hostile intent. This bias has been extensively studied and has been found to reliably predict aggressive behavior.

Script theory emphasizes the acquisition of scripts for behavior (much like an actor's script) through either direct experience or observational learning. Once encoded in semantic memory, *scripts* define particular situations and provide a guide for how to behave in them. In script theory, a person faced with a particular situation first considers a script relevant to that situation, assumes a role in the script, assesses the appropriateness or likely outcome of enacting the script, and if judged appropriate, then behaves according to the script. If a person habitually responds to conflict by using scripts that include behaving aggressively, these scripts may become more easily brought to mind (i.e., chronically accessible), become automatic, and generalize to other situations, increasing the likelihood of aggression in a growing number of spheres of life.

Cognitive Neoassociation Theory

Cognitive Neoassociation Theory (CNA) reformulated the frustration-aggression hypothesis within the framework of emerging knowledge about neural connectivity. Assuming that concepts, emotions, memories, and action tendencies are interconnected within the brain's associative neural network, Berkowitz (1989) posited that aversive events such as frustrations, provocations, or unpleasant physical environments produce negative affect, which is neurally linked to various thoughts, feelings, and behavioral tendencies that are themselves linked to both fight and flight tendencies. Depending on the characteristics of the person and the situation, one response set will eventually dominate, with dominant 'fight' responses linked with anger and being more likely to elicit aggression. Importantly, higher-order processes such as making attributions about another's motives or thinking through the consequences of an aggressive response may cause a person to moderate an aggressive impulse in this model.

Current Theories

The General Aggression Model

The General Aggression Model (GAM; Anderson and Bushman, 2002) is the most recent and broadest theory of aggression processes to date. It is a biosocial-cognitive model designed to account for both short- and long-term (developmental) effects of an extensive range of variables on aggression. GAM can explain the widest range of aggressive behaviors, including those not based around aversive events or negative affect. In addition, it is arguably the model that has the most empirical support. GAM unifies previous major models of aggression from the field of social psychology into a single framework, but also incorporates knowledge from other disciplines in psychology.

The model itself is deceptively parsimonious. Every instance of aggression involves a person, with all their characteristics (e.g., biology, genes, personality, attitudes, beliefs, behavioral scripts), responding to an environmental trigger such as a provocation, an aversive event, or an aggression-related cue (lower portion of **Figure 1**). These *person* and *situation* variables influence the person's *present internal state* – cognitions, affects, and physiological arousal. Depending on the nature of activated knowledge structures (which include affect), and on how

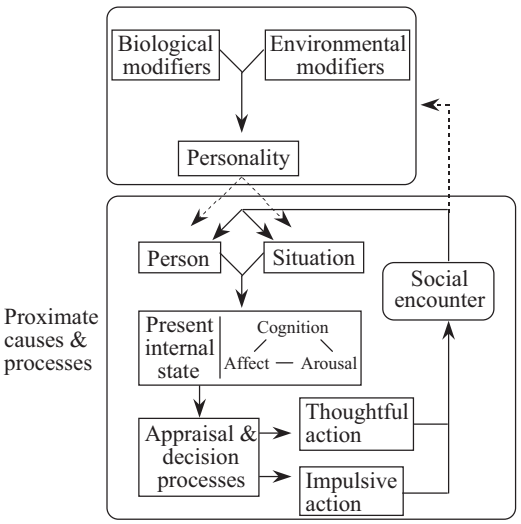

Figure 1 General Aggression Model. From Anderson, C.A., Anderson, K.B., 2008. Men who target women: specificity of target, generality of aggressive behavior. Aggressive Behavior 34, 605–622. Reprinted by permission.

aroused the person is, the person's immediate response may be an impulse to aggress. The person may act on this impulse, but if they have the time and cognitive resources to do so, and if the immediate response is undesirable, a period of appraisal and reappraisal will follow. Consequences are then thought-through, alternate responses considered, and a considered response made. The resulting behavioral action may or may not be aggressive, but in any case all actions feed back into the immediate situation and also influence the person's psychological make-up (i.e., their personality).

Underlying the GAM are detailed assumptions that take into account a myriad of within-person factors, a range of possible triggers for aggression, known internal psychological processes, and the means by which behavior is reinforced and learned. In terms of the latter, knowledge structures such as schemas (a grouping of knowledge, feelings, memories, perceptions and notions about typical behavior that is centered around a particular theme) and scripts (knowledge about how people typically behave in a given situation such as during conflict) are person factors that can not only impel a person to be aggressive in the moment, but also change to reflect our experiences (upper portion of **Figure 1**). Thus, experience leads to changes in the type, content, and accessibility of knowledge structures, which are seen as the basis of personality.

Together, these features of GAM can be used to explain short- and long-term aggression across a range of forms and functions, including the three key dimensions already noted: degree of hostile/agitated affect; degree of automaticity versus conscious thought; and degree to which the goal is to harm the victim versus benefit the perpetrator. Phenomena as different as sexual and nonsexual aggression against women (e.g., Anderson and Anderson, 2008), personality effects on violent crime (Hosie et al., 2014), and dozens more are well explained by GAM.

Aggression Research Methodologies in Social Psychology

As a social behavior, aggression has been primarily studied using methodologies from social psychology. Each of these methodologies has specific strengths and weaknesses, but, importantly, the shortcomings of each methodology can be overcome with the strengths of another. This allows aggression researchers to be strongly confident of an effect where findings converge across methodologies (Warburton, 2013).

Laboratory Assessments of Aggression

Laboratory experiments provide the strongest evidence that a particular factor may play a causal role in aggression. This is because that factor can be manipulated whilst all other factors are (in theory) held constant (e.g., all participants may have an identical experience in the laboratory except for watching a violent or a nonviolent movie clip). Aggression experiments typically measure short-term increases in mild forms of aggression or in known precursors such as aggressive thoughts and feelings. For example, researchers might measure whether aggression-related thoughts are more activated in one group of participants compared with another by testing reaction times to identify aggression-related (hit, blood) versus neutral (sew, rose) words. Aggressive feelings are typically measured by having participants rate the degree to which they feel emotions such as anger, antagonism, and unfriendliness.

Measuring aggressive behavior itself has a long history involving ethical, reliability, and validity concerns. For ethical reasons, serious harm cannot be used as an aggression measure in laboratory experiments. However, numerous valid and reliable aggression measures have been developed, usually involving a contrived laboratory situation that allows participants to behave in a way that they *believe* will harm another, but in which no person is actually hurt. Early measures included counting the number of aggressive acts a child would make toward a target, and the willingness of an adult to deliver a (fake) electric shock to another person purportedly being tested for their ability to memorize stimuli under conditions where they would be 'punished' for mistakes. More recent methods include measuring the duration and/or loudness of aversive 'noise blasts' delivered to an opponent in a competitive reaction time (CRT) game, the amount of hot chili sauce assigned for eating by a stranger known to dislike hot foods, and the number of difficult puzzles that require solving by another person in order to win a reward. Although such measures have been criticized for being unlike 'real-world' situations and subject to biases such as the desire to please (or displease) the experimenter, well-designed modern experiments overcome such problems using careful cover stories and scripts, and have been shown to predict real-world aggression.

Nonexperimental Research about Aggression

Nonexperimental research has the distinct strength that it can examine a wide range of 'real-world' aggressive phenomena and can be used to examine longer-term effects such as the development of a more aggressive personality. *Longitudinal studies* (in which key variables are measured at multiple points in time) are particularly valuable, because they can measure the development and change of aggression over time within individuals, and examine long-term effects of wideranging factors such as home environment, personality, and media violence exposure. In addition, the logical impossibility of a later-introduced factor causing an earlier-mentioned behavior, along with use of recent advances in statistical techniques, allow some causal inferences to be drawn.

Cross-sectional studies (in which all variables are measured once) are also valuable, but require cautious interpretation. Causal inferences are risky, because of the possibility that not all relevant factors were measured and taken into account. Nevertheless, such research has contributed substantially to theory testing and development by providing the opportunity to test causal theory-derived hypotheses and alternative explanations to the causal theory.

Observations of Aggression in Social Psychology

Some of the earliest and most powerful social psychological research of aggression was conducted through observational research, some in the laboratory (e.g., Bandura's Bobo Doll experiments) and some in the field (e.g., studies of aggression on playgrounds). Such studies have the distinct advantage of

observing and recording actual rather than self-reported aggressive behavior, often in the participants' natural environment. Importantly, issues that sometimes arise from self-report questionnaires (e.g., biased responding, lack of self-awareness or capacity to report thoughts and feelings) are not relevant, and populations unsuitable for other forms of research (such as young children) can be examined. However, aggressive behaviors often have a low incidence in observed environments, and ratings of aggression can be somewhat subjective. Researchers overcome the latter issue by creating clear and comprehensive guidelines, detailing behaviors that should be coded (e.g., pushing, shoving, hitting, name calling), and thoroughly training the raters. Also, self-reports may be supplemented by reports of relevant others (e.g., parents, teachers and peers). Indeed, converging data from multiple sources often provides the strongest evidence (Anderson et al., 2007; Warburton, 2014).

Brain Scanning Techniques to Study Aggression

Social psychologists are now using brain-scanning techniques to study aggression, most notably in the field of media violence. Such techniques have the advantages that they can be used on many types of participants, participants cannot 'fake' their responses, and participants do not have to be self-aware to provide valid responses. Brain scans are particularly valuable for assessing factors difficult to measure using other methods such as desensitization to violence, fear responses, and emotional arousal. Functional magnetic resonance imaging (fMRI) studies identify brain activity by measuring changes to blood flow, but are accurate only to a few seconds across time. Brainwave activity measured by electroencephalography (EEG) and magnetoencephalography (MEG) are extremely accurate in terms of the timing of changes to brainwaves, but cannot give accurate locations within the brain. Thus, using both techniques to study the same hypotheses leads to better understanding. Brain-scanning studies have some drawbacks – they generally use small samples because of the cost involved, they need to average images using sophisticated software and sometimes their data are hard to meaningfully interpret.

In typical aggression studies, participants are scanned whilst experiencing one or more stimuli (such as playing a violent or nonviolent video game) or doing various tasks, such as rating different types of pictures or making decisions. Thus, researchers can compare activation patterns to determine whether changes (such as desensitization) occur over time, one type of stimulus has different effects than another, or different groups (e.g., high vs low media violence consumers) typically respond differently.

Research Findings: Determinants of Aggression

Development and Stability Over Time

Scholars studying social development have shown that the frequency of physical aggression typically peaks in the toddler years and then decreases across the life span. Importantly, the degree to which one person is aggressive relative to others of the same age is fairly stable across the life span. Aggressive children tend to become adolescents and adults who are more aggressive than their peers (Bushman and Huesmann, 2010).

Person Factors

Numerous factors in a person's make-up have been shown to increase the likelihood of aggressive behavior. Not all are studied directly in social psychology, but all are taken into account in current social psychological models of aggression.

Gender Differences in Aggression

Overall, males are generally more aggressive than females, and this applies from early in childhood through the life span. This is especially true for physical aggression and violent behavior, although women are as physically aggressive as men when strongly provoked. Men are more likely than women to use direct forms of aggression, but the reverse is true for women, who are more likely to use forms of indirect aggression, including relational aggression.

Within intimate relationships, however, women are somewhat more likely to use physical aggression than men, though for different purposes and with different results. For example, men are much more likely to strike with a fist (women with an open slap), which is one reason why intimate partner violence yields many more women requiring medical attention than men.

Trait Anger

Trait anger reliably predicts an aggressive predisposition. It is characterized by extreme sensitivity to provocation and a considerably increased inclination to respond with aggression once provoked.

Callous Unemotional Personality Traits

There are three personality styles under this umbrella – psychopathy, Machiavellianism, and narcissism. All three are linked with high levels of aggression, lack of empathy, and curtailed emotional responding. Individuals of all three types routinely use aggression instrumentally to obtain desired goals, but narcissists and psychopaths are also prone to reactive aggression. Narcissists often respond aggressively when they feel threatened (particularly by insults, humiliations, or other threats to their inflated ego), or when they fear that their flaws may be exposed. Psychopaths, particularly those with secondary psychopathy characteristics, are often impulsive, fearless, and unconcerned about negative consequences to themselves or victims – a potent mix for a person already predisposed to aggression. Machiavellians most typically use instrumental aggression to achieve their goals and feel little or any remorse for harmful consequences to others. They do, however, consider potential consequences to themselves, and are thus more likely to aggress indirectly so that there is little likelihood of being held responsible for their actions.

Impulsivity, Executive Control, and Self-Control

Impulsivity is a temperament variable often noticeable from early infancy, and is a reliable predictor of aggression, presumably because impulsive people have difficulty curbing aggressive impulses. In contrast, people are less aggressive if they have greater control over their emotions, greater

self-control, and a stronger capacity to inhibit their impulses (Moffitt et al., 2011).

Intelligence

There is not a great deal of research on IQ and aggression, but some studies have found links between low IQ and higher levels of aggression in children, particularly in children with poor verbal intelligence and/or with low self-control.

Personality Traits – The 'Big Five'

Research on the 'Big Five' personality traits and aggression has generally found that people low in agreeableness and high in neuroticism are more aggressive and violent. Furthermore, both of these dimensions are associated with aggressive emotions, and low agreeableness is also associated with greater aggressive thinking (Barlett and Anderson, 2012).

Hormones

The hormone most consistently linked with aggression is testosterone. Males have around 10 times as much testosterone as females, and levels are much high in older teenagers and young adults than in older men. Interestingly, when people dominate others, their testosterone levels typically increase, along with their levels of aggression. There also is evidence that testosterone's effect on aggression is a by-product of its effect on dominance. There also may be links between low levels of estrogen and progesterone and aggression, although results are mixed. Finally, emerging evidence suggests that low levels of oxytocin may be linked with increased aggression.

Genetic Predispositions

Although aggressive behavior has a considerable learned component, studies show that inherited characteristics account for perhaps a quarter to a third of an aggressive predisposition (Tuvblad et al., 2009). More than a dozen genetic markers have been linked with aggressive and antisocial behavior, although links are rarely direct. Typically, genetic predispositions more directly relate to temperament variables such as impulsivity, which are themselves linked with greater aggression. The two most widely studied genetic markers of aggression are a polymorphism in the promoter of the monoamine oxidase A gene (MAOA) and a variation in the 5-HT serotonin transporter gene. Crucially, in line with the emerging field of epigenetics, the MAOA gene polymorphism seems to interact with a child's early environment, so that aggression and antisocial behavior are most likely in those who have this genetic trait *and* also experience childhood maltreatment (Kim-Cohen et al., 2006).

Factors from the Environment and Cues for Aggression

Provocation

Perhaps the single greatest trigger for aggression is provocation by another person (Bettencourt et al., 2006). However, provocation does not need to be direct. People can be provoked to aggression by social exclusion, having rumors spread about them and a range of other 'indirect' provocations.

Weapons

Weapons are one stimulus that almost all people conceptually link with aggressive behavior. Research consistently shows a 'weapons effect' whereby people who view a real or virtual weapon tend to have aggression-related cognitions primed in semantic memory, and become more likely to behave aggressively. Interestingly, this effect varies by type of weapon and hunting experience (see **Figure 2**).

Violent Environment

According to social cognitive models, people who are exposed to a lot of violence, virtual or real, will have an associative neural network with a lot of aggression-related knowledge structures, including aggressive behavioral scripts. This is borne out by research demonstrating that people from violent environments, whether homes, neighborhoods, or war-torn countries, have a greater predisposition to be aggressive (e.g., Aguilar et al., 2000).

Violent Media

The same principle applies to exposure to violent media. It is one of the most studied phenomena by social psychologists, and several hundred studies converge across all major research methodologies in finding that violent media exposure increases the likelihood of aggressive behavior and causes desensitization to violence in both the short- and long-term (Warburton, 2014). In addition, greater exposure to media violence has been linked to hostile biases in thinking, increases in aggressive thoughts and feelings, and decreases in empathy and prosocial behavior (see Anderson et al., 2003; Krahe et al., 2012, for reviews).

Environmental Stressors

A variety of environmental stressors can increase the tendency to aggress. The most notable are physical pain, bad-smelling odors, loud or aversive noises, and hot temperatures. Importantly, it seems that aggression is most likely when the individual has no control over those environmental stressors.

Anonymity

Anonymity in some circumstances increases the likelihood of aggressive behavior. It is much easier to hurt another

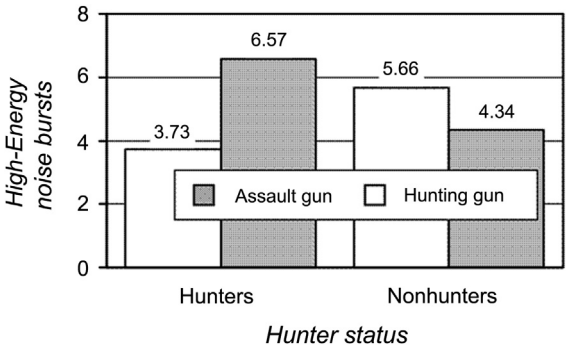

Figure 2 Aggressive behavior (number of high-energy noise bursts directed at the opponent) as a function of hunter status and weapon prime. From Bartholow, B.D., Anderson, C.A., Carnagey, N.L., Benjamin, A.J., 2005. Interactive effects of life experience and situational cues on aggression: the weapons priming effect in hunters and nonhunters. Journal of Experimental Social Psychology 41, 48–60. Reprinted by permission.

if an individual believes there will be no consequences, and anonymity allows a person to experience 'deindividuation' – a lessening of the restraints on antisocial behavior normally accorded to people perceived as being 'individuals.'

Social Rejection

Humans have a fundamental need to feel socially included and to have supportive and enduring relationships. When this need is thwarted through social exclusion or rejection, people sometimes behave more prosocially to facilitate reinclusion. However the dominant response to such rejection is to aggress, especially when the person can do so without significant social reprisals (e.g., Warburton et al., 2006).

Substances

Alcohol intoxication consistently causes both men and women to behave more aggressively inside and outside the laboratory, and is linked with a substantial proportion of murders, assaults, rapes, and incidents of intimate partner violence. Importantly, this increase is due to the aggressor experiencing a diminished ability to inhibit their aggressive impulses. Thus people who are predisposed to behave aggressively are most affected (Giancola, 2000). Aggression has also been linked with other substances that cause disinhibition and/or an increase in physiological arousal, such as stimulants, amphetamines, and methamphetamines.

Research Findings: Factors that Mediate Aggression

The previous section examined factors within the person and in the environment that can trigger or increase the likelihood of aggression. This section deals with the three key types of internal processes noted in GAM that can increase or decrease the likelihood of aggression.

Emotion/Affect

Early models of aggression placed considerable emphasis on the role of negative emotions in causing aggressive behavior, and research has linked several emotions to an increased likelihood of aggression, most notably anger, shame, jealousy, and frustration. Of these, anger is the most researched. Although anger can precede aggression, the pathway is far from a simple cause and effect. Anger increases aggression primarily through reducing inhibitions, narrowing attentional focus to cues for aggression, and alerting people to cues for potential threats (see Anderson and Bushman, 2002). Shame has also been linked to increases in aggression, primarily when the shamed person feels their personal flaws have been exposed. Jealousy has also been linked with aggression and in particular with intimate partner violence.

Recent research suggests that the anticipation of how one will feel in the future can be as important as how one currently feels in determining whether a person will be aggressive.

It should be noted that some emotions can be a protective factor for aggression. For example, empathy (taking another person's perspective and having concern for them) is consistently related to lower aggression.

Cognition

Recent models of aggression have focused increasingly on the cognitions that may underlie aggression. These include attitudes, beliefs, expectations, perceptions, ideas, and concepts as well as aggregated cognitions such as schemas and scripts. It is clear that a variety of external triggers can increase the accessibility of aggressive cognitions in semantic memory. These cognitions may be activated but below the threshold of awareness, or activated to the point of conscious awareness. In either case they can elicit an aggressive behavioral tendency through the priming of aggression-related action-tendencies, the activation of aggressive scripts for behavior, or through the influence of hostile biases on the interpretation of cues from the environment.

Arousal

Physiological arousal and emotional arousal are both linked with increased aggression, and this is true regardless of what caused the arousal in the first instance. This may be due to one or more of several factors. First, arousal increases the likelihood that a person will act on an aggressive action tendency or impulse rather than think through the consequences of an aggressive action. Second, excessive levels of arousal feel unpleasant, and can elicit aggression in the same way as other unpleasant experiences. Third, arousal may be part of a fight or flight response system that bypasses rational thought and impels aggressive action. Fourth, arousal may be cognitively labeled as resulting from anger, thus causing the person to feel and act angry. Excitation transfer may compound this effect, leading to a disproportionately aggressive response. Finally, low levels of arousal may facilitate aggression if people lack the energy and motivation to inhibit aggressive impulses.

Societal Aggression and Violence – a 'Risk Factor Approach'

It is one thing to know the types of factors that increase the likelihood of aggression, but quite another to understand aggression and violence in wider society. No single factor described in this article is either sufficient or necessary to elicit violent or other extreme forms of aggression. These only occur when there is a confluence of 'risk factors' for aggression (such as those detailed in this article) and insufficient 'protective factors' to inhibit aggression. The greater the number of risk factors and the stronger their influence, the more likely it is an individual will behave aggressively, especially when protective factors are few or of little impact (Anderson et al., 2007). The problem for researchers is that the greater the number of risk factors they need to consider when studying aggression and violence, the more difficult it is to determine how the factors interact with each other, and to ascertain the relative impact of each on an act of aggression or violence. However, this is the task that faces social psychologists as they try to make sense of mass killings, school shootings, and societal violence. Importantly, the more that is known about risk factors and protective factors, the greater the ability of psychologists to understand

and prevent societal violence, and indeed aggression in everyday life.

Conclusions

Human aggression has been researched by social psychologists and others for many decades. The result is a large body of knowledge about the factors within people and from the environment that increase the likelihood of aggression, along with a more detailed understanding of the processes that occur in the mind and brain during an instance of aggression. Well-validated models such as GAM have been built around these findings. Less is known about the ways in which risk factors for aggression and violence interact with each other and with protective factors. Still, much is known about how to decrease the likelihood of societal violence. Yet, this knowledge is, in our view, not sufficiently used in society at large.

One ongoing world crisis illustrates this problem. Global climate change as a result of human activity is now a widely (though not universally) accepted fact. In combination with work from a variety of biological, sociological, historical, and anthropological, findings, GAM suggests that there are at least three ways that such rapid global warming will increase violence worldwide. The most direct is the simple heat effect mentioned earlier. A second way is that increased poverty, malnutrition, and family disruption will increase the proportion of children who develop into aggression-prone adolescents and adults. The third way involves ecomigration, the movement of populations from ecological disasters to other regions, and the intergroup conflicts that will result (Anderson and DeLisi, 2011). We hope that the knowledge gained by decades of aggression research in social psychology will be put to better use in the future.

See also: Attitudes and Behavior; Authoritarian Personality; Prosocial Behavior and Empathy; Social Cognition; Social Psychology; Tyranny.

Bibliography

Aguilar, B., Sroufe, L.A., Egeland, B., Carlson, E., 2000. Distinguishing the early-onset/persistent and adolescence-onset antisocial behavior types: from birth to 16 years. Developmental Psychopathology 12, 109–132.

Anderson, C.A., Anderson, K.B., 2008. Men who target women: specificity of target, generality of aggressive behavior. Aggressive Behavior 34, 605–622.

Anderson, C.A., Berkowitz, L., Donnerstein, E., Huesmann, L.R., Johnson, J., Linz, D., Malamuth, N., Wartella, E., 2003. The influence of media violence on youth. Psychological Science in the Public Interest 4, 81–110.

Anderson, C.A., Bushman, B.J., 2002. Human aggression. Annual Review of Psychology 53, 27–51.

Anderson, C.A., DeLisi, M., 2011. Implications of global climate change for violence in developed and developing countries. In: Forgas, J., Kruglanski, A., Williams, K. (Eds.), The Psychology of Social Conflict and Aggression. Psychology Press, New York, pp. 249–265.

Anderson, C.A., Gentile, D.A., Buckley, K.E., 2007. Violent Video Game Effects on Children and Adolescents: Theory, Research, and Public Policy. Oxford University Press, New York.

Anderson, C.A., Huesmann, L.R., 2003. Human aggression: a social-cognitive view. In: Hogg, M.A., Cooper, J. (Eds.), Handbook of Social Psychology. Sage Publication, London, pp. 296–323.

Bandura, A., 1986. Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice-Hall, Englewood Cliffs, NJ.

Barlett, C.P., Anderson, C.A., 2012. Direct and indirect relations between the Big 5 personality traits and aggressive behavior. Personality and Individual Differences 52, 870–875.

Bartholow, B.D., Anderson, C.A., Carnagey, N.L., Benjamin, A.J., 2005. Interactive effects of life experience and situational cues on aggression: the weapons priming effect in hunters and nonhunters. Journal of Experimental Social Psychology 41, 48–60.

Bettencourt, B.A., Talley, A., Benjamin, A.J., Valentine, J., 2006. Personality and aggressive behavior under provoking and neutral conditions: a meta-analytic review. Psychological Bulletin 132, 751–777.

Berkowitz, L., 1989. Frustration-aggression hypothesis: examination and reformulation. Psychological Bulletin 106, 59–73.

Bushman, B.J., Huesmann, L.R., 2010. Aggression. In: Fiske, S.T., Gilbert, D.T., Lindzey, G. (Eds.), Handbook of social psychology, fifth ed., vol. 2. John Wiley & Sons, Hoboken, NJ, pp. 833–863.

Dodge, K.A., 1980. Social cognition and children's aggressive behavior. Child Development 51, 620–635.

Dollard, J., Doob, L., Miller, N., Mowrer, O., Sears, R., 1939. Frustration and Aggression. Yale University Press, New Haven, CT.

Eron, L.D., Walder, L.O., Lefkowitz, M.M., 1971. The Learning of Aggression in Children. Little Brown, Boston.

Giancola, P.R., 2000. Executive functioning: a conceptual framework for alcohol-related aggression. Experimental and Clinical Psychopharmacology 8, 576–597.

Hosie, J., Gilbert, F., Simpson, K., Daffern, M., 2014. An examination of the relationship between personality and aggression using the general aggression and five factor models. Aggressive Behavior 40, 189–196.

Huesmann, L.R., 1982. Information processing models of behavior. In: Hirschberg, N., Humphreys, L. (Eds.), Multivariate Applications in the Social Sciences. Erlbaum, Hillsdale, NJ, pp. 261–288.

Kim-Cohen, J., Caspi, A., Taylor, A., Williams, B., Newcombe, R., Craig, I.W., Moffitt, T.E., 2006. MAOA, maltreatment, and gene-environment interaction predicting children's mental health: new evidence and a meta-analysis. Molecular Psychiatry 11, 903–913.

Krahé, B., Berkowitz, L., Brockmeyer, J.H., Bushman, B.J., Coyne, S.M., Dill, K.E., Donnerstein, E., Gentile, D.A., Huesmann, L.R., Kirsch, S.J., Möller, I., Warburton, W.A., 2012. Report of the media violence commission. Aggressive Behavior 38, 335–341.

Moffitt, T.E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R.J., Harrington, H., Houts, R., Poulton, R., Roberts, B.W., Ross, S., Sears, M.R., Thomson, W.M., Caspi, A., 2011. A gradient of childhood self-control predicts health, wealth, and public safety. PNAS Proceedings of the National Academy of Sciences of the United States of America 108, 2693–2698.

Tuvblad, C., Raine, A., Zheng, M., Baker, L.A., 2009. Genetic and environmental stability differs in reactive and proactive aggression. Aggressive Behavior 35, 437–452.

Warburton, W.A., 2013. Aggression: definition and measurement of. In: Eastin, M. (Ed.), Encyclopedia of Media Violence. Sage, Thousand Oaks, CA, pp. 10–14.

Warburton, W.A., 2014. Apples, oranges and the burden of proof: Putting media violence findings in context. European Psychologist 19, 60–67. doi:10.1027/1016-9040/a000166.

Warburton, W.A., Williams, K.D., Cairns, D.R., 2006. When ostracism leads to aggression: the moderating effects of control deprivation. Journal of Experimental Social Psychology 42, 213–220.

Zillmann, D., 1979. Hostility and Aggression. Erlbaum, Hillsdale, NJ.

AGIL, History of

Giuseppe Sciortino, Università di Trento, Trento, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article chronicles the history of AGIL, the best-known conceptual scheme associated with the work of Talcott Parsons. It starts reviewing the theoretical elements of Parsons' early work that have played a role in his design of the AGIL scheme in the early 1950s. It describes the various development of the AGIL scheme in his subsequent work. It concludes with an assessment of its contemporary relevance.

Introduction to Parsons and AGIL

Both in textbooks and in disciplinary lore, even the most cursory reviews of the work of Talcott Parsons never fail to mention AGIL, an acronym composed of the initials of what he claimed where the four analytic dimensions of any action system: adaptation, goal attainment, integration, and latent pattern maintenance. It is difficult to underestimate the importance Parsons attributed to his invention. For him, AGIL was a generalized template for sociological models. For more than 25 years, he tried, whenever a problem was posed, to show that it could be addressed through the construction of a model satisfying the AGIL requirements (Fararo, 2001). He insisted AGIL was a consistent conceptual scheme, able to highlight the multidimensional nature of social life, clarify disciplinary boundaries, and trigger comparative work across all the social and life sciences. He declared that the development of AGIL had made possible a theory of long-term social change, energized by the pressure of a plurality of independent problems.

Parsons felt AGIL was the very embodiment of his lifelong search for multidimensional analysis (Alexander, 1984). He claimed AGIL was able to overcome any opposition between material and ideal factors and to dispense with the limitations of any base–superstructure model. Unsurprisingly, Parsons applied the AGIL scheme – usually in the graphic form of four-squared boxes, each of them further subdivided by four – to an astonishing range of analytical levels and empirical issues (Parsons, 2007). In his writings, the reader can find AGIL of families, university systems, world societies, civilizations, churches, and small groups. He even applied AGIL to the mysteries of the human condition itself (Parsons, 1978).

Equally unsurprising, many of his critics have used AGIL as their primary foil. Some have seen in it the proofs of Parsons' endemic tetra mania; some others have argued that its formal elegance has been just the cover for lousy arguments held together by elimination, implication, and superficial analogies; still others have claimed AGIL needlessly restricted the possibilities of theoretical thinking in the social sciences (Mulkay, 1971; Luhmann, 1988; Stinchcombe, 1975). Whatever the judgment on AGIL's intrinsic worth, it is difficult to deny that it is one of the most wide-reaching and technically complex conceptual schemes ever developed in the social sciences.

Before AGIL: Action, Situation, and Normative Patterns

Given the core location of AGIL in Parsons' work, it may be a surprise that AGIL is actually a late product of his career. Many of the major works traditionally associated with Parsons (roughly from *The Structure of Social Action* to *The Social System*) have actually been published before the appearance of the scheme. AGIL is consequently a mature development, which both strengthen and innovate Parsons' earlier approach. If Parsons felt AGIL was of particular importance, it is precisely because he identified in it a solution to a cluster of theoretical problems that had shaped his previous work.

Parsons' sociological theory centers on the question of how a shared system of social expectations may regulate and sustain the interaction among a plurality of independent actors. His starting point is the critique of previous social theories for their failure to develop a sociological view of the problem, falling back to extra-sociological factors such as biological heredity or environmental constraints. In his view, on the contrary, actors have no predetermined – biologically or environmentally – mechanisms of coordination (Parsons, 1937). A satisfactory understanding of human action has to assume the *analytical* independence among the personality of the actor, the cultural values, and the requirements of social interaction. Parsons adopts a stance of *institutionalized individualism*, where the actors' agency is not a theoretical given, but rather the outcome of a to-be-investigated socialized growth process where social relationships and cultural templates play a crucial role (Bourricaud, 1977). If actors are independent, any interaction presents a certain degree of uncertainty and risk for all those involved (Parsons, 1951). How can a plurality of independent actors achieve – and take for granted in their everyday activities – an adequate level of coordination and reciprocal understanding? Parsons think there is only one way to contemporize individual agency and social order: the existence of a shared system of normative expectations (Parsons, 1968). This possibility is rooted in the fact that social action – albeit voluntary, goal oriented, and subject to multiple selective criteria – is not idiosyncratic. The meaning of each action – for the actor herself and for any other observer – is contingent upon the matching between the motivations of actors and the (shared) perceived meaning of their interactional objects (or partners). Parsons distinguishes three kinds of perceptive and evaluative choices: cognitive, cathectic

(i.e., emotionally charged), and evaluative (Parsons and Shils, 1951). Any action requires learning about the potential objects and their relationships, feeling them as desirable or undesirable, evaluating them one against the other drawing upon (normative) assessment of their importance. Such subjective judgments are not solipsistic: they have to be understood by interactional partners, who have no access to the internal states of the actor. They rely upon (respectively) cognitive, cathectic, and moral norms, functioning as shared expectations on how the interaction partners (and the personality of the actor) will receive the action itself (Parsons, 1968).

Parsons summarizes the choices faced by the actors through what he calls 'pattern variables,' a small set of five polarized alternatives that constitute the meaning of the situation from the point of view of the actor. Three of the pattern variables refer to individual motivations and values:

- *Affectivity versus affective neutrality*. This includes the choice between immediate gratification and postponement, as well as the choice between charging the object with an emotional meaning or assessing it for its consequences for other purposes.
- *Universalism versus particularism*. This refers to the choice between treating the object as the member of an abstract, generalized category, or in terms of the specificities of its relationship to the actor.
- *Self-orientation versus collectivity orientation*. This is the choice between approaching the object as an individual or as a member of a collectivity the actor is member of (**Figure 1**).

If actors have no predetermined wishes and wants, social objects have no predetermined meanings and usages. According to Parsons, two further pattern variables are consequently needed in order to reduce the ambiguities over the nature of the social objects:

- *Quality versus Performance* (often also labeled as *Ascription vs Achievement*). This is the choice between quality and usability, between seeing the object as valuable in itself or as instrumental for other purposes.
- *Specificity versus diffuseness*. This is a choice between allowing objects to expect a specific set of behaviors or to entrust them with an open-ended set of expected dealings (**Figure 2**).

Pattern variables are elements that structure, in different ways and from different angles, the orientation of personality systems, the value patterns of cultural systems, and the normative structures of social systems (Parsons, 1960b). The degree to which actors assume appropriate selections among the (above mentioned) choices is what guarantees a certain level of mutual predictability and reasonable chances of cooperation among the actors, without – at the same time – precluding the pursuit of independently elaborated goals. Parsons identified two social processes through which actors' orientations could be coordinated: socialization, through which cultural patterns become (or fail to do so) selectively incorporated in personality systems and institutionalization, through which the same cultural patterns are selectively embedded (or fail to do so) in the distribution of actual social rewards (Parsons and Shils, 1951).

The economists use perfect competition and market equilibrium as natural benchmarks against which to study all other market structures. Parsons decided to proceed along similar lines, starting with the case of the (ideal) social situation where socialization and institutionalization are consistent and perfectly tuned. In such an ideal case, that Parsons calls "complete institutional integration of individual motivation," the actors desire and pursue socially desirable goals having access to socially prescribed means. The structure of their interaction, moreover, is such that their actions bring positive assessments by the other interactional partners of their performance, thus ensuring satisfactory outcomes. Parsons stressed that the "complete institutional integration of individual motivation," as the subsequent AGIL equilibrium, was neither an empirical description nor a normative ideal. It was just an abstract point of reference, a comparative criterion for assessing the relative distance of the various empirical contexts. Through his emphasis on socialization and institutionalization, Parsons was able to define the social system as having an integrative function for the whole action system, as it was social interaction that tied together and made compatible social, cultural, and personality elements of any (analytically understood) social action (**Figure 3**).

Until the end of the 1940s, however, Parsons had not yet developed a theory of the social system. While he was able to discuss analytically the role it played in the regulation of individual actors, he had no systematic theory of its articulation. He was not able to describe the processes through which it could maintain a minimal strength and consistency of the normative structure of expectations. Above all, he was not able to anchor systematically the cluster of pattern variables with an independent vision of social structures (Parsons, 2010). In *The Social System*, his most well-known book (and in many previous empirical essays), Parsons distinguished only two functional problems. Any social system – from a couple to a world society – has to deal with (1) the allocation of resources among the various units and (2) the compatibility among its various institutions, through methods of social control and procedures for managing disputes and strains. The first has to do with scarcity: any social system may distribute less facilities and rewards than it would be (ideally) necessary to satisfy all individual motivations of participants (Parsons, 1951). The second has to do with diversity: actors have goals and expectations independent from the context; are loyal to different groups; and they are able to interpret social values in different ways. The integration of a social system requires consequently the existence of processes able to regulate their goals (and loyalties, and interpretations) in ways compatible with the maintenance of the social system as a solidary collectivity. Doing so, however, Parsons incurred in the same problems he had criticized in the early functionalists: the provision of an indefinite list of functions justified by their empirical significance rather than by their inclusion in a consistent analytical framework. This usage, Parsons stressed, had favored the proliferation of new functions on ad hoc basis, thus weakening the appeal of functional analysis (Sciortino, 2008).

The Early AGIL

The roots of AGIL lye in several collaborative projects Parsons was involved at the end of the 1940s and for all of the early

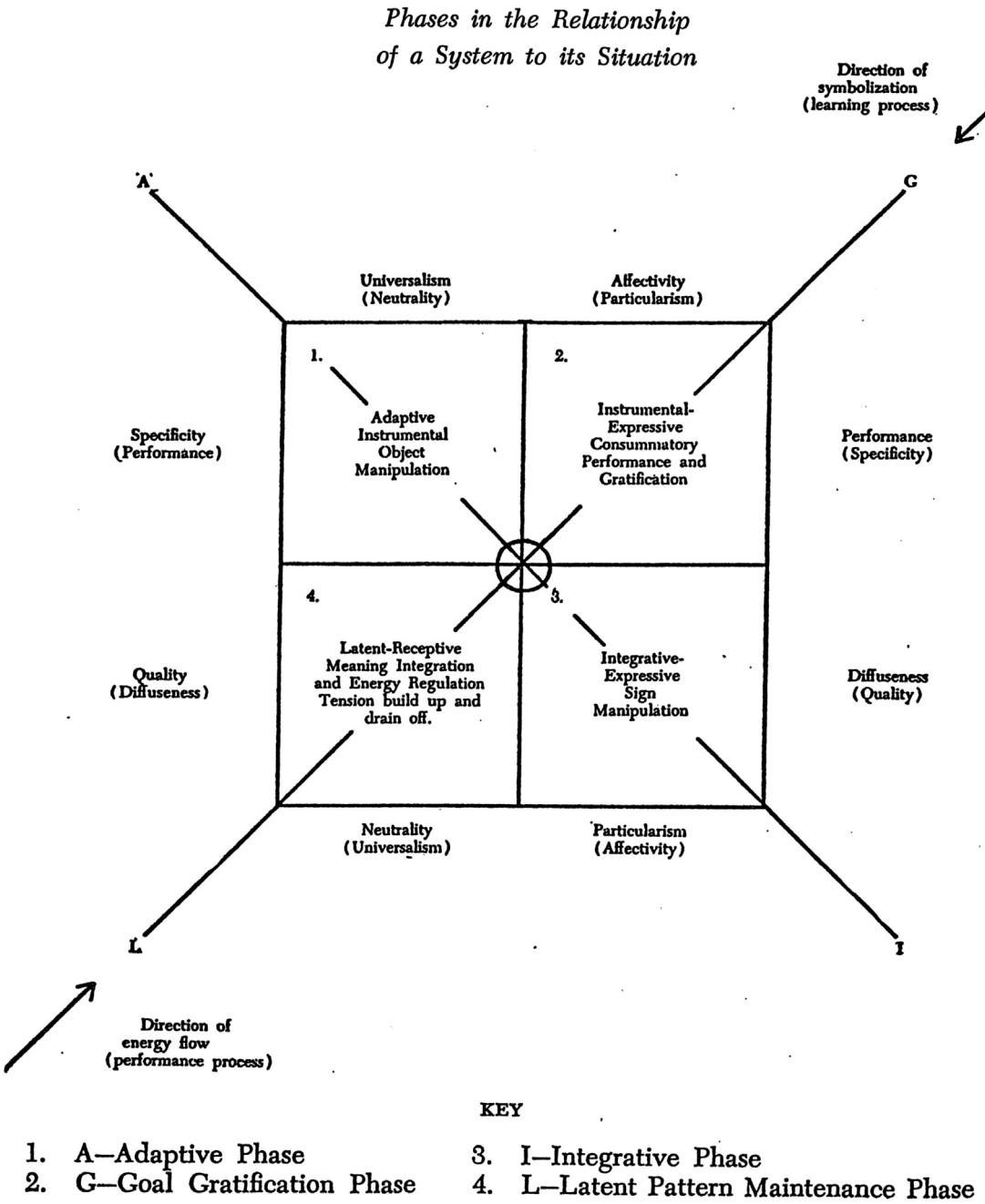

Figure 1 Parsons T., Bales R., Shils E., 1953. Working Papers in the Theory of Action. The Free Press, New York, p. 182.

1950s. Two key figures in the early development of AGIL were Robert Bales, the social psychologist who had pioneered the systematic observation of group processes, and Edward Shils, who was then working on the role of symbolism in the regulation of social interaction (Parsons and Shils, 1951; Parsons et al., 1953). Other colleagues – among them Gordon Allport, Clyde and Florence Kluckhohn, James Olds, and Samuel Stouffer – provided important contributions (Parsons, 1977).

The starting point for the design of AGIL was the IPA framework devised by Bales to account for his observations of interaction in small groups (Bales, 1950). He clustered the problems encountered by the groups into four functional categories, deemed independent from the contents of the various exchanges within the group. Any human group had to face the pressure deriving from (1) the adaptation to external conditions; (2) the instrumental controls over the goals of members necessary to the management of collective tasks; (3) the expression of the motivational energies of members; and (4) the maintenance of its coordination mechanisms (Parsons et al., 1953).

Bales' analysis was important to Parsons for several reasons. First, Bales' model did not postulate the primacy

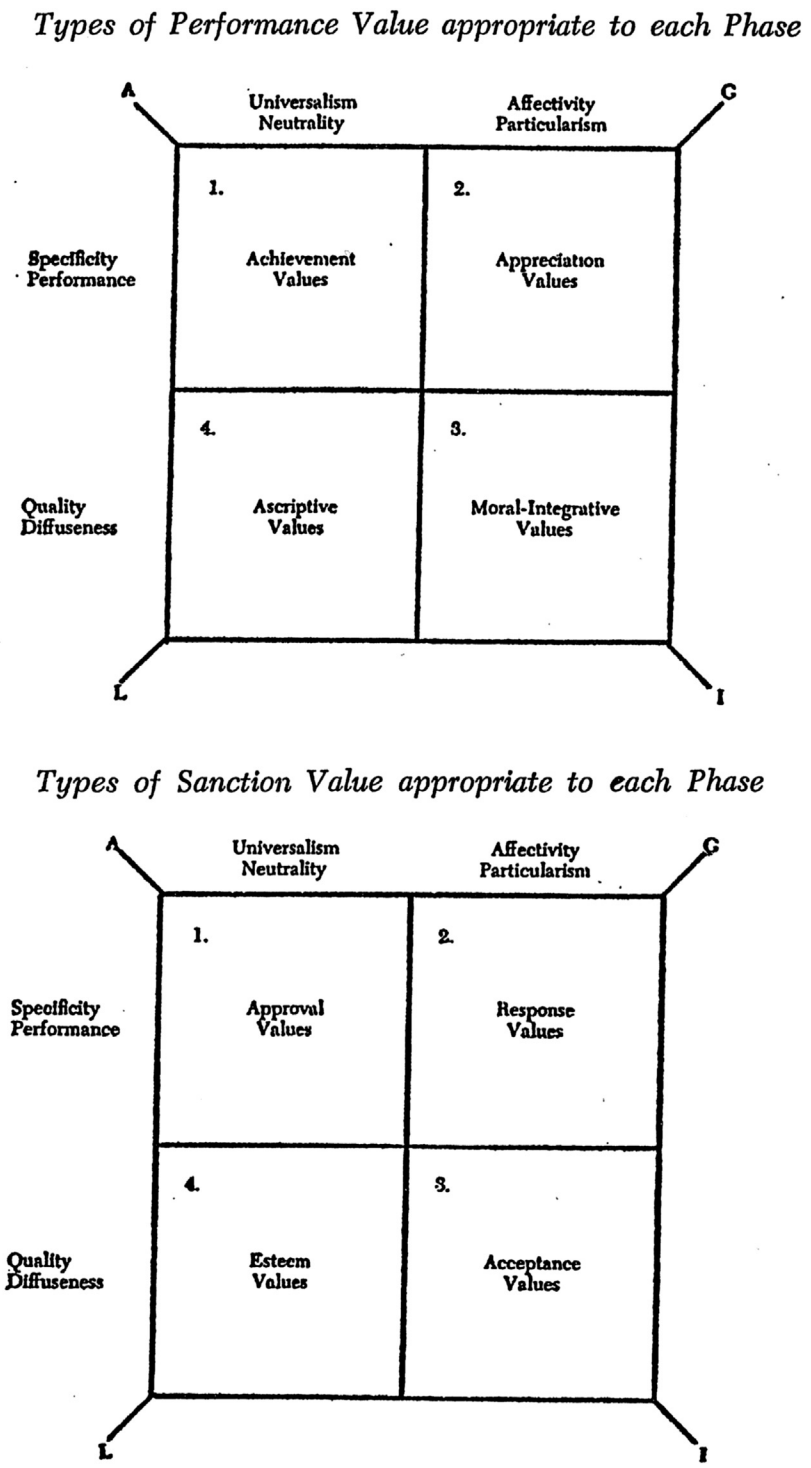

Figure 2 Parsons T., Bales R., Shils E., 1953. Working Papers in the Theory of Action. The Free Press, New York, p. 203.

of certain problems over others, not even 'in the last instance.' Second, it showed how external pressures did not dictate the reaction of the groups, as they also had an internal dynamic that made possible a variety of outcomes. Third, it was compatible with what Parsons had tried to achieve with the pattern variables, as it linked analytically the motivation of members with the structural dynamics of the group.

Bales, Parsons, and Shils produced a collection of papers in 1953, *Working Papers in the Theory of Action* (Parsons et al., 1953). In his fifth, and conclusive, chapter the reader finds the very first version of the AGIL scheme. AGIL is presented as

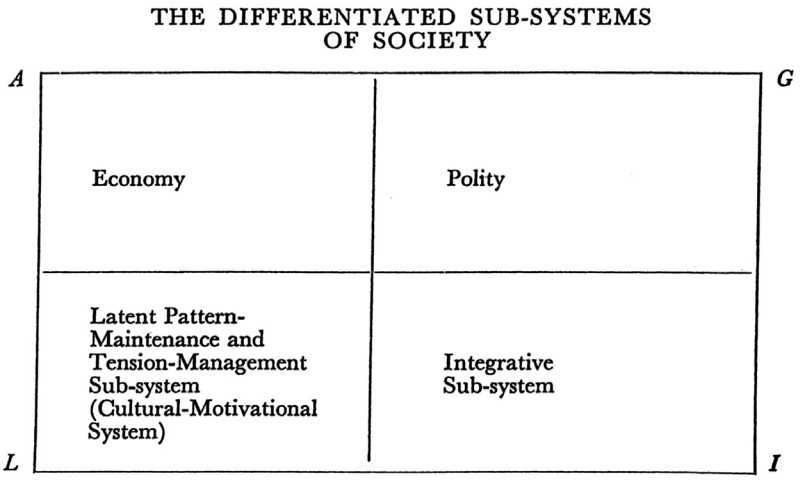

Figure 3 Parsons T., Smelser N.J., 1956. Economy and Society, Routledge and Kegan Paul, London, p. 53.

a descriptive device, designed to summarize the temporal phases in the empirical development of a group. Ironically, the scheme that has often been criticized for its alleged conservative and static biases was born as a dynamic description of a social sequence. Graphically, the four 'problems' are placed along two axes: a vertical one linking the maintenance of psychological and cultural preconditions with the attainment of collective goals and a horizontal one linking external adaptation and integration.

After 1953, all subsequent works by Parsons have taken place within the idea that any action system (both individual action and social systems) has to deal with the following problems:

- *Adaptation* (A): the control and transformation of nonsocial resources;
- *Goal attainment* (G): the management of concerted action by the social units involved for collective purposes;
- *Integration* (I): the adjustment of relationship among the units of the system, the management of conflicts, and the settling of disputes;
- *Latent pattern maintenance* (L): the generation of long-term commitment to shared values and identities (Figures 4 and 5).

Parsons has also maintained across his subsequent work the special importance of the two axes, the vertical one linking L with G and the horizontal one linking A to I. Finally, Parsons maintained the assumption that no action system may satisfy all these 'problems' at once: therefore, there are always tensions and strains that require management and control. In other words, any social system has the same built-in tensions between actual and possible, between conditional and normative, that Parsons had previously placed at the center of his analysis of individual action (Fararo, 2001).

Since 1953, Parsons and his colleagues were increasingly outspoken in their claim that the AGIL scheme was potentially significant for any action system, regardless of size and complexity. What is more important, they claimed its discovery made possible to develop a structural analysis of social systems fully compatible with the subjective orientation of action. As some specific combinations of pattern variables could be shown to match a (independently generated) list of functional problems, this would take away from functional analysis the arbitrariness in the definitions of what counts as functionally important (Parsons, 1960a). They consequently linked some combinations of pattern variables to each of the identified 'problems' or dimensions of the social system:

- *Adaptation* is defined by specific and affectively neutral dispositions, and by a valuation of objects in terms of universalism and performance;
- *Goal attainment* (Bales' instrumental control) is defined by affectionate and specific dispositions, and by a valuation of objects in terms of performance and particularism;
- *Integration* is defined by affectionate and diffuse dispositions and by a valuation of objects in terms of quality and particularism;
- *Latent pattern maintenance* (at the time labeled, in a rather anodyne way, *Latent-Receptive Meaning Integration and Energy Regulation Tension Build-up and rain Off*) is defined by diffuse and affectively neutral motivations and by a valuation of objects in terms quality and universalism (Figures 6 and 7).

AGIL and the Cybernetic Hierarchy

During the 1950s, Parsons experimented with the AGIL scheme, applying it in many sociological analyses. He used AGIL to guide his, rather controversial, analysis of the modern American family (Parsons and Bales, 1955), as well as his, equally controversial, exploration of the relationships between economics and sociology, and between economy and society (Parsons, 1991). He also felt the need to revise his theory of social stratification along the line of the four-problem schema (Parsons, 1954). In his economic sociology, he identified the economy, in terms of the four-function model, as the adaptive subsystem of a society. He also defined the input–output categories traditionally treated in economic theory (as the factors of production

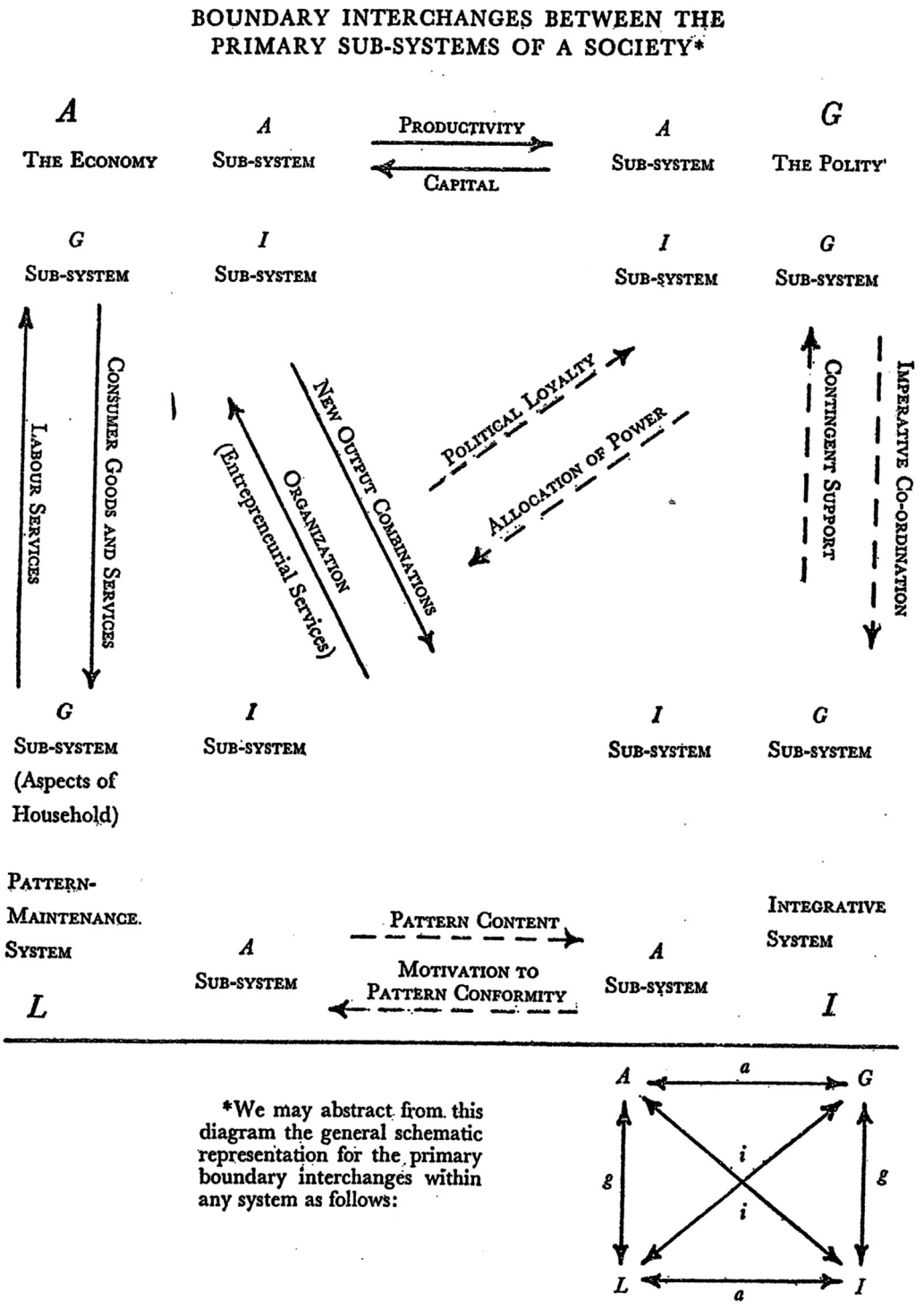

Figure 4 Parsons T., Smelser N.J., 1956. Economy and Society. Routledge and Kegan Paul, London, p. 68.

and the shares of income) as categories of *relation* between the adaptive subsystem and the other functional systems of the society (Parsons and Smelser, 1956). AGIL acquired then the nature of an exchange model, where its subsystems are analytically autonomous but organically interdependent as they provide each other with specialized resources. The exchange dimension of AGIL will become increasingly dominant in the subsequent elaborations by Parsons.

The first full-fledged technical discussion of the AGIL scheme appeared only in 1960. As the original one, it is centered on the relationships between the pattern variables and

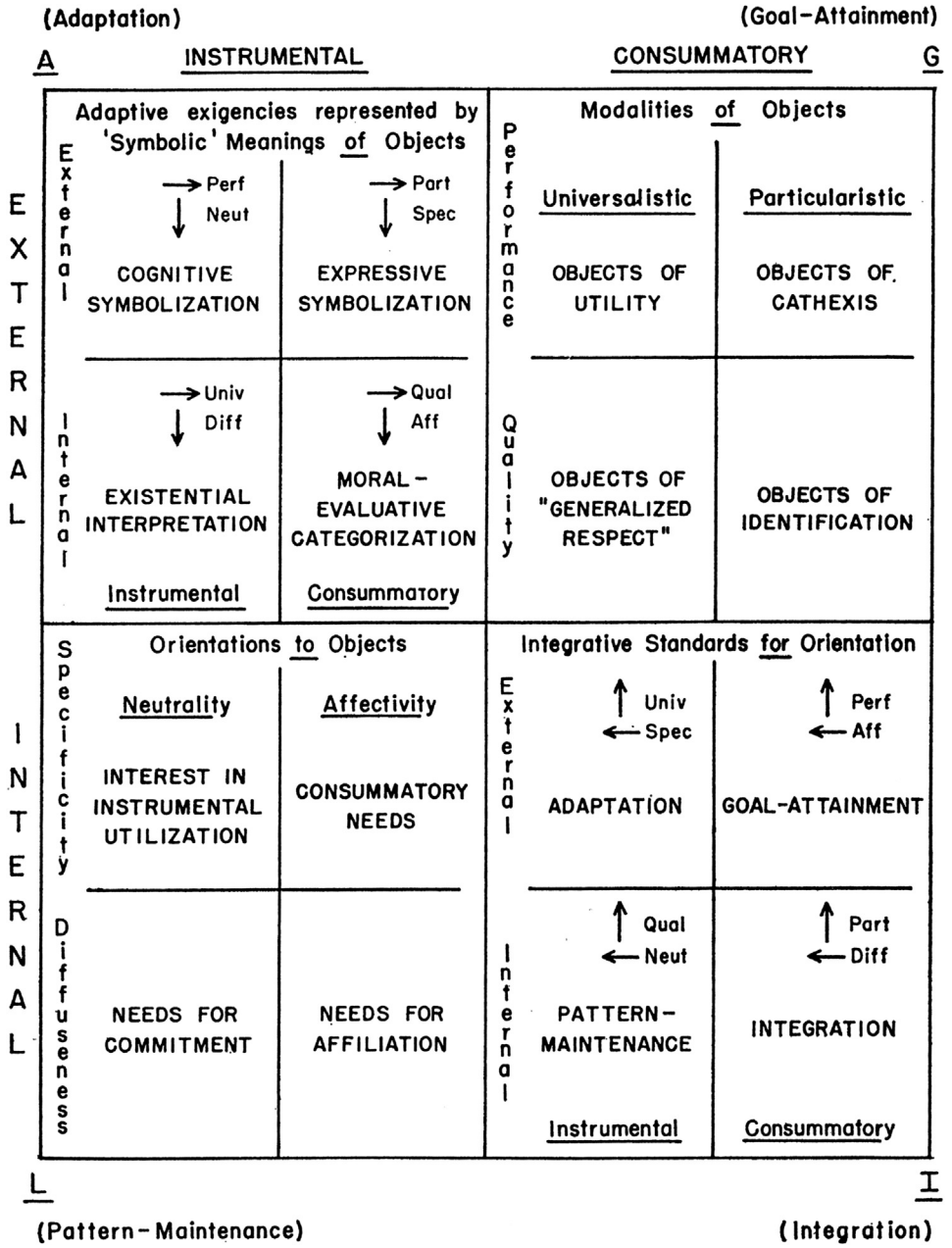

Figure 5 Parsons T., 1960. Pattern variables revisited: a response to Robert Dubin. American Sociological Review 24, 467–483, p. 470.

the AGIL schema. Reacting to several critics, Parsons declares inadequate AGIL's early characterization as a phase model. He presents it as a dimensional template for the analyses of both action and social systems. He also finds a larger congruence between the pattern variables and AGIL: the 1953 combinations are casted as defining the subsystem of to the integrative system, and he proposes further combinations of them as appropriate for the psychological and cultural subsystems of action (Parsons, 1960a).

The original dynamic reading of AGIL does not, however, completely disappear. Parsons reformulates it through the claim that pressures and strains are managed, within any system, according to a hierarchy of control expressed in the form of a 'cybernetic hierarchy' (Parsons, 1960b). Such hierarchy – that later Parsons summarized as the inverse of AGIL, as LIGA – suggested that the more information-rich subsystem (L) would ordinarily enjoy a higher degree of autonomy from the net of structural interdependencies, thus being able to 'control' the changes taking place in them (Parsons and Platt, 1973). Parsons was keen to repeat that, in his view, cybernetic control did not imply domination in every respect nor the absence of feedback from lower, 'conditioning,' subsystems. Still, the lack of a convincing analytical justification for the notion has paved the way for the critique of the cybernetic hierarchy being a cover for his idealistic bias (Alexander, 1984).

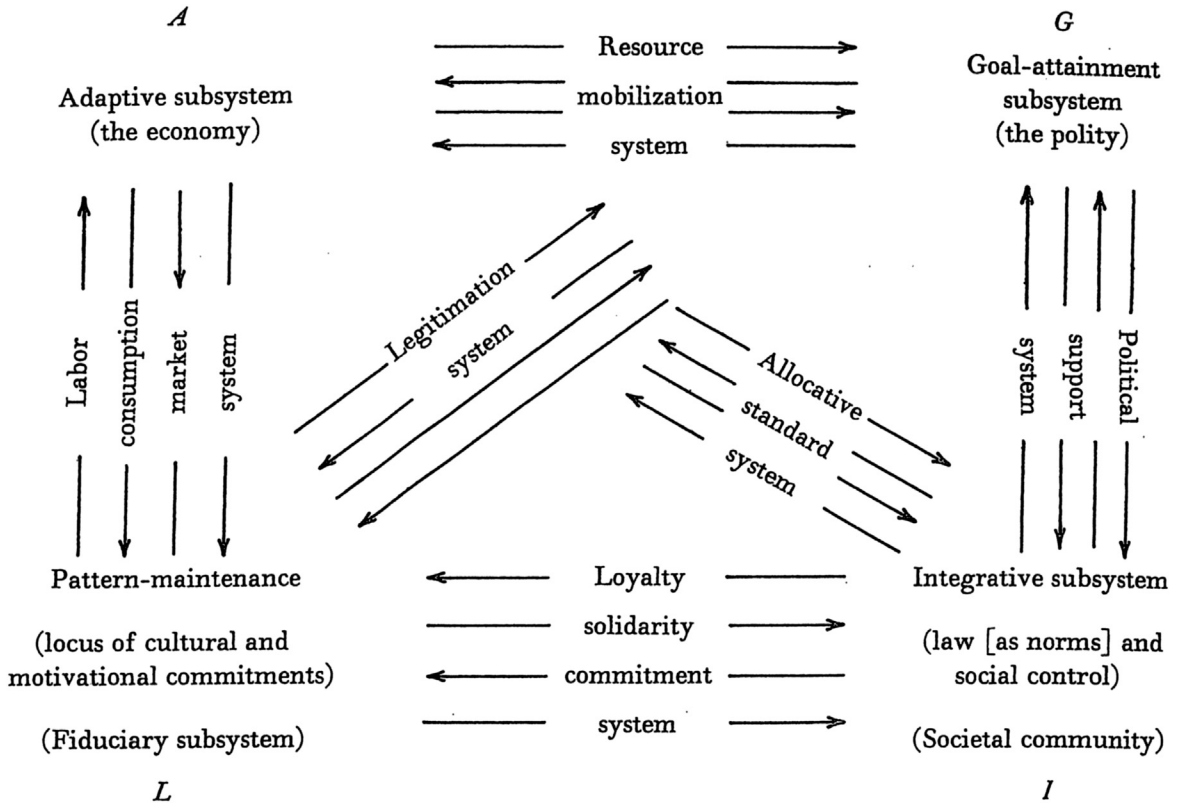

Figure 6 Parsons T., Platt G., 1973. The American University. Harvard University Press, Cambridge, p. 426.

AGIL and the Study of Societies

Starting from the early 1960s, Parsons has developed AGIL in two directions: as a scheme for the analysis of the analytical structures of the social system, including modern societies, and as a template for the media of interchange, linking individual action and social structure. These are the best-known exemplars of Parsons' application of AGIL, although Parsons always stressed that 'societies' are a very specific and extreme form of the social system (Parsons, 1966).

When applied to contemporary societies – a term Parsons assumed, albeit with some perplexities, as closely approximated by the nation-state – the AGIL scheme identifies four distinct subsystems. Parsons sees modern society as structurally differentiated. Albeit all societies, as all social systems, are analytically differentiable in four subsystems, only in modern society such subsystems are 'structurally visible,' more closely – albeit never perfectly – associated with specialized structural units (Fox et al., 2005). The existence of clusters of functionally differentiated, largely autonomous, institutions and collectivities providing specialized contributions is what distinguishes modern society from previous forms of social organization (Parsons, 1966).

Parsons, drawing on his previous analysis, had already associated the modern economic institutions (broadly conceived) with the adaptive subsystem of a modern society, specialized in the development and allocation of fluid resources for a variety of goals (Parsons and Smelser, 1956). The next step was extending his line of reasoning to what he called 'the political system' – made of governmental bodies, as well as of nonpublic organizations – he associated with goal attainment (Parsons, 1969). He subsequently focused on the integrative subsystem. He coined the term 'societal community' for the subsystem specialized, following Durkheim, in the coordination between mechanic and organic forms of solidarity. He associated the integrative subsystem of modern society to the definition of membership, the management of the rights and duties attached to the system of social statuses, and the settling of disputes among competing loyalties (Parsons, 2007). In the same years, he explored what he called the fiduciary system, the latent pattern maintenance (L) subsystem of a society. The fiduciary system provides the cultural and motivational commitments necessary to a highly complex and highly pluralistic world. Parsons associated to such functions the institutions and collectivities crucial in the transmission and development of societal culture (Parsons, 1978). Each societal subsystem was further broken down by Parsons in its constituent parts (Parsons and Platt, 1973, p. 423–451). In all his AGIL reading of contemporary societies, Parsons described them as a complex web of conditional and normative elements, where a network of flexible interdependencies keeps together differentiated institutions and systems of complex solidarities.

From the mid-1960s, moreover, Parsons applied the AGIL schemes to problems of long-term social change. His vision of societal evolution can be described as a combination of Weberian and Durkheimian themes: through his writings, the master trends of social evolution are the breaking of the

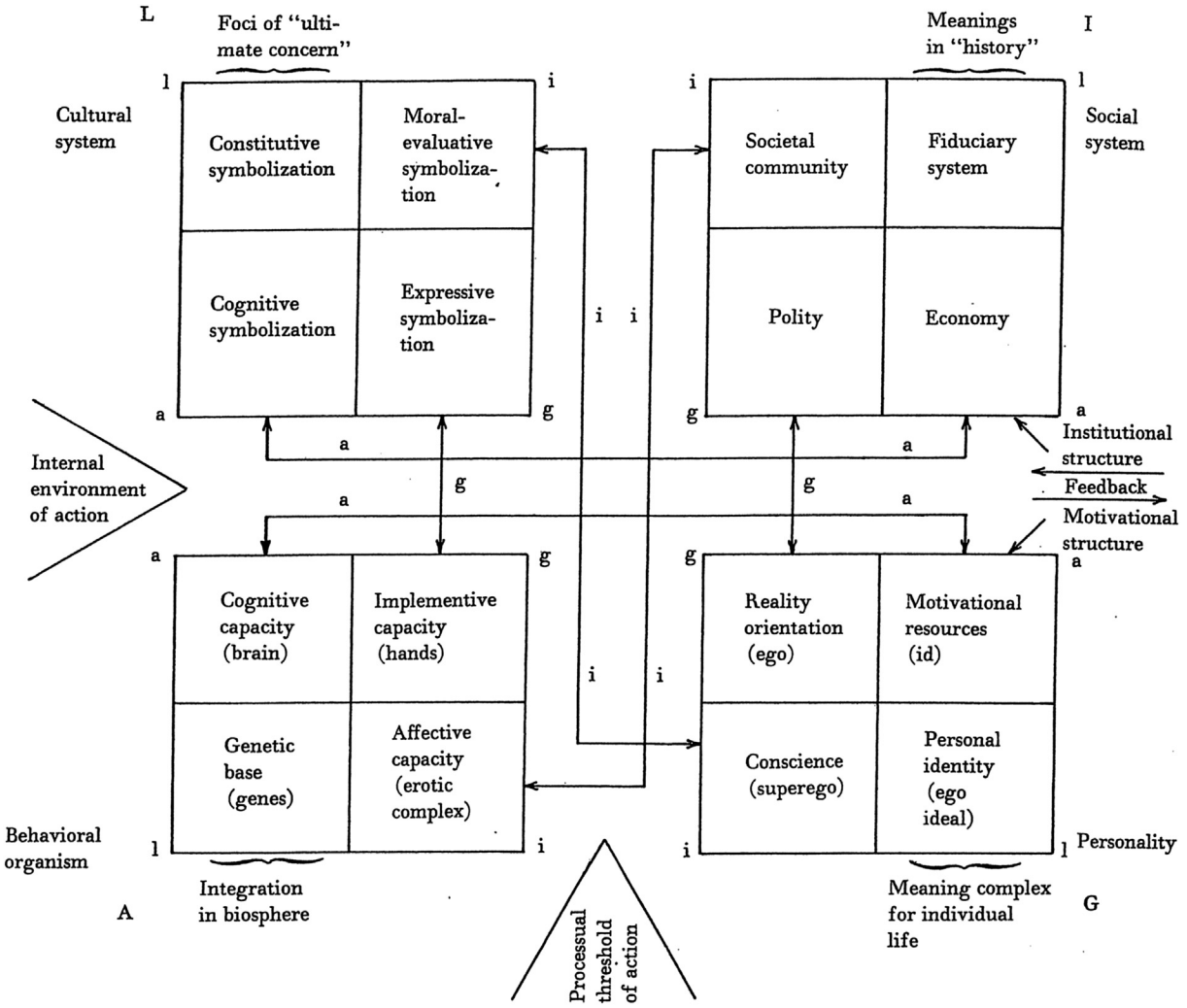

Figure 7 Parsons T., Platt G., 1973. The American University. Harvard University Press, Cambridge, p. 436.

matrix of ascriptive relationships (a process that implies culturally the defamiliarization of the world) and the increasing adaptive capacity vis-à-vis the external environments. In his studies of social change, Parsons has applied AGIL both spatially and temporally (Parsons, 1971). Geographically, Parsons identified modernity with the birth of a system of modern societies, where different spaces and forms of social organizations play differentiated roles in the functioning of the international system. As for the second, Parsons accounted for modernity in terms of a (still uncompleted) sequence of changes in each of the subsystems. The transition to modern society initiates with the industrial revolution, centered on the economy (A). Such a differentiation process, however, requires broad changes in the institutions of collective decision-making process, what Parsons calls the democratic revolution (G). Such changes, however, trigger the possibility of a redefinition of societal membership in more inclusionary terms. With what Parsons calls the social revolution (I), a process broadly associated to the development of Western citizenship and the welfare state – modern society achieve a more inclusive societal community (Parsons, 2007; Sciortino, 2010). The everyday egalitarian interaction among members, however, requires an increased acceptance of diversity and pluralism. Such acceptance places pressures to the fiduciary system to detach the value system from any strictly defined tradition, thus further increasing what Parsons called 'value generalization.' Parsons thought that the main challenges to contemporary differentiated societies were rooted in the consequences of the inclusion processes (in the societal community) and the value generalization processes (in the fiduciary system). He claimed two further revolutions – the Educational and Expressive revolutions (L) – were underestimated but crucial events for contemporary societies (Turner, 1993, 2005).

AGIL and the Generalized Media of Interchange

Another direction in which Parsons developed the AGIL scheme was in linking structurally action and social systems through a set of mechanisms – that he labeled 'generalized media of interchange' – that would translate the outcome of social

interaction into the functioning of societal subsystems (and vice versa). He also entrusted such media with the task of providing the means for the interchange among the four societal subsystems. The blueprint for such set of mechanisms was money, the media that he had already analyzed in his economic sociology (Parsons and Smelser, 1956). He subsequently added power (G), influence (I), and value commitment (L) thus mirroring the four function systems with four function media (Parsons, 1969). Through such addition, he planned to show how the functional analysis of a social system was a necessary step for the development of a scientifically viable voluntarism. The base for the concept of media is a classification of the ways in which, given a double contingency interaction among independent actors, an actor may try to bring about a change in what the actions of other units would otherwise have been (Parsons, 1968). In other words, he focused on social coordination of voluntarist actors, within the same framework in which the analysis of socialization and institutionalization evolved. To classify the means available to the actor, Parsons selected two dimensions: the type of sanctions available to Ego in order to obtain Alter Ego compliance (positive *vs* negative sanctions) and the channel he can use to bring about such compliance (acting on the situation *vs* acting on Alter Ego intentions). He concluded that Ego could obtain Alter Ego's compliance in four ways: inducement, coercion, persuasion, and activation of value commitments.

Each of them has a symbolic element: even the most brutal coercion is meaningless without a previous communication of a contingent threat (Parsons, 1969). The sanction itself, moreover, may be symbolic, as in the case in which we transfer money or property titles. Parsons' emphasis on this symbolic element serves to bring attention to the fact that the interaction-level sanctions and rewards are interdependent with the four structural dimensions of societal systems. In Parsons' view, this symbolic element allows the use in the interaction of more resources than actually materially available in the environment at any given point. The capacity to emit a credible threat makes possible the control of many more actions and contexts than it would be possible in the case Ego had to back each threat with the actual use of physical force. In the same vein, the possibility to take for granted certain symbolically generalized commitment of Alter Ego, instead of making a detailed investigation any time, make possible a degree of trust far higher than otherwise. Parsons sees such expansion of interactional means as made possible by the structural anchoring of the interactional means in the institutionalized media related to the various subsystems (Parsons, 1969). In the modern economic order, the use of money in any interaction relies on the existence of an institutional order that simultaneously constrains and enables Ego in his choices. At the same time, the aggregate composition of Ego's spending decisions has effects on the state of the economic system in many ways, notably in determining processes of inflation and deflation. In a series of complex essay, Parsons argued that the same applies to all the other subsystems, though in different ways linked to the different natures of the media involved. The trick here is that what from the point of view of the actor is a means to further his own interests is a set of stability conditions for the institutional order (Parsons, 2007).

AGIL as a Tool for the Analysis of Action Systems, up to the Human Condition

Parsons had originally devised AGIL for the study of social actions. Starting from the end of the 1960s, after a long period marked by an overall concern with the analysis of social systems, Parsons returned to the original focus, now labeled general action system (Parsons and Platt, 1973; Parsons, 1977, 1978). He used the AGIL to summarize many of his lifelong claims on the multidimensionality of action and the positive interplay of individual agency and institutional regulation (Alexander, 1984). To reframe his earlier claims within the AGIL template, Parsons replaced the classification of culture, social system, and personality with a scheme of four primary subsystems. He had previously added, for the adaptive function, the behavioral organism, constituted of aspects of the individual organism that provide resources for physical behavior (Parsons, 1960b). He now aligned the personality with the goal attainment function because of the importance of psychological patterns of motivation to purposeful agency (Fox et al., 2005). He confirmed the role of the social system as integrative, as it provided the normative structures that regulated actors into common relationships and situations. He also confirmed culture as having primarily a pattern-maintenance function (Parsons, 1977). Each of them, in his most complex statement, was associated with a specialized media of interchange: intelligence (A), performance capacity (G), affect (I), and definition of the situation (L). Exactly, as for the social system, Parsons argued that each subsystem should be modeled according to the same AGIL template (Parsons and Platt, 1973, p. 436; Parsons, 2007).

In the last decade of his life, Parsons was also involved in an attempt to further generalize the scope of the AGIL scheme, in order to explore the connections between action and social systems – both made of meaningful actions – and the physical and biological bases of human societies. Focusing on what he defined as the 'human condition,' Parsons claimed that action systems could be seen as the integrative subsystem of a broader human condition involving also the physical environment (A), the organic, the human environment (G) and a telic, or transcendental environment (L). As in the previous efforts, he labored to provide such last scheme with its own media and interchange categories (Parsons, 1978; **Figures 8–11**).

Post-Parsonian Developments

Parsons' death in 1979 abruptly ended his efforts to explore and formalize the AGIL scheme further. In the subsequent decades, the evaluation of his overall legacy has become increasingly positive. Parsons has been increasingly recognized as a classic figure, able to open large issues, propose insightful and suggestive answers, and yet leaving significant problems unresolved and awaiting future contributions (Fox et al., 2005). The

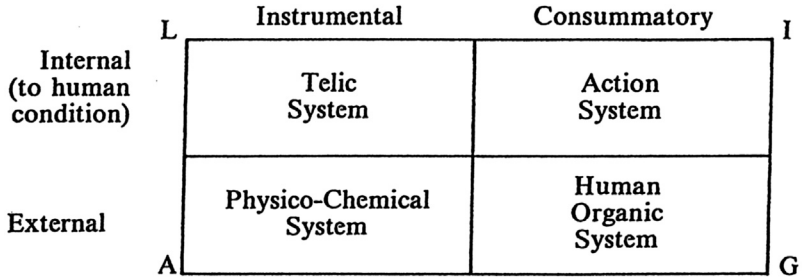

Figure 8 Parsons T., 1978. Action Theory and the Human Condition. The Free Press, New York, p. 361.

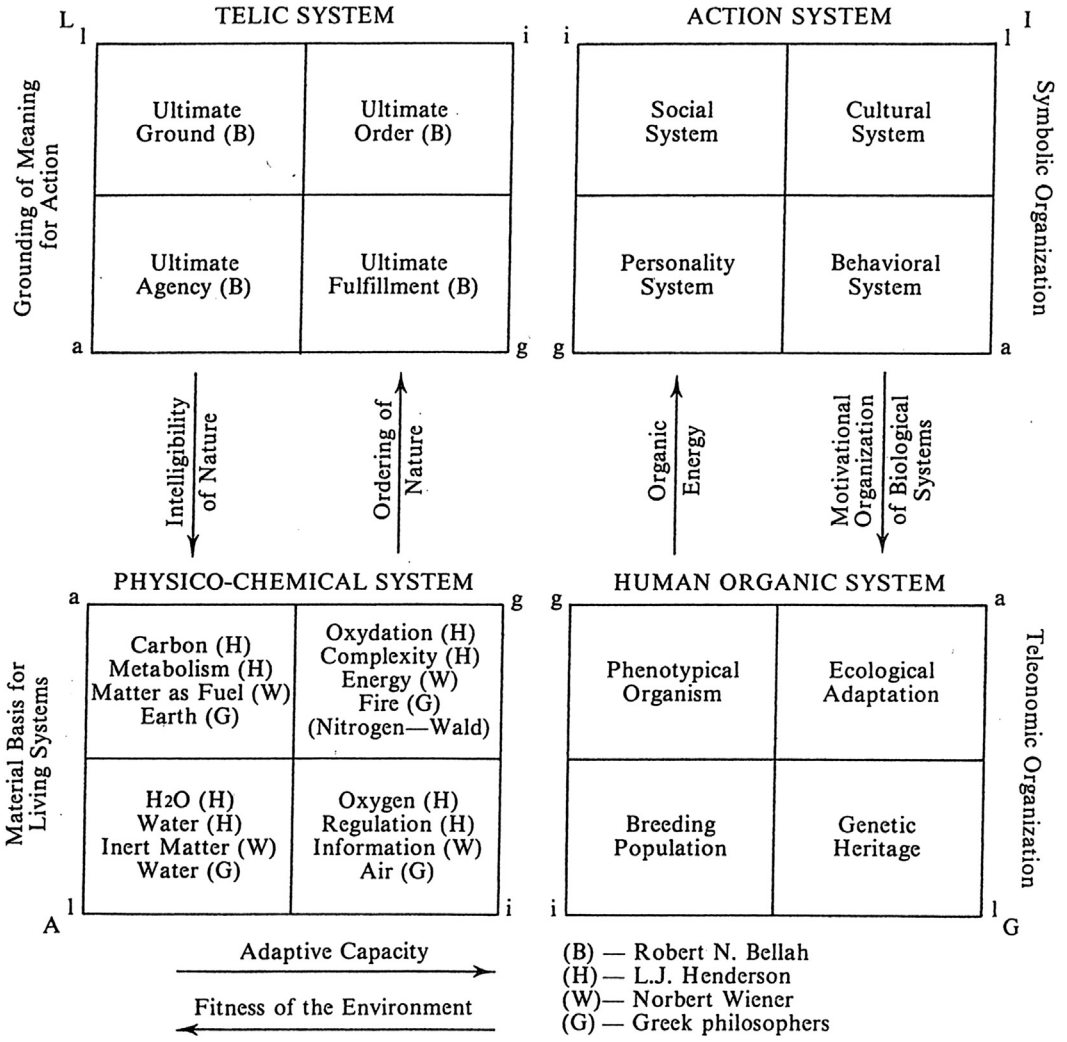

Figure 9 Parsons T., 1978. Action Theory and the Human Condition. The Free Press, New York, p. 383.

AGIL scheme, however, has remained strictly associated with his work, failing to become a conceptual tool and theoretical technique used beyond the boundary of the circle of scholars interested in action theory. There have been some important attempts to increase its formalization of AGIL and to develop critically its theoretical potential (Brownstein, 1982; Loubser et al., 1976; Fararo, 2001). They have been unable to trigger a wider interest of the social science community. Even among the scholars who have done more to restore Parsons to a central position in contemporary social theory, there have been few attempts to use AGIL as little more than a source of broad inspiration (Alexander, 1984; Turner, 2000).

A search on Google Scholar with the key words *AGIL* and *Talcott Parsons* returns more than 1000 items for the period

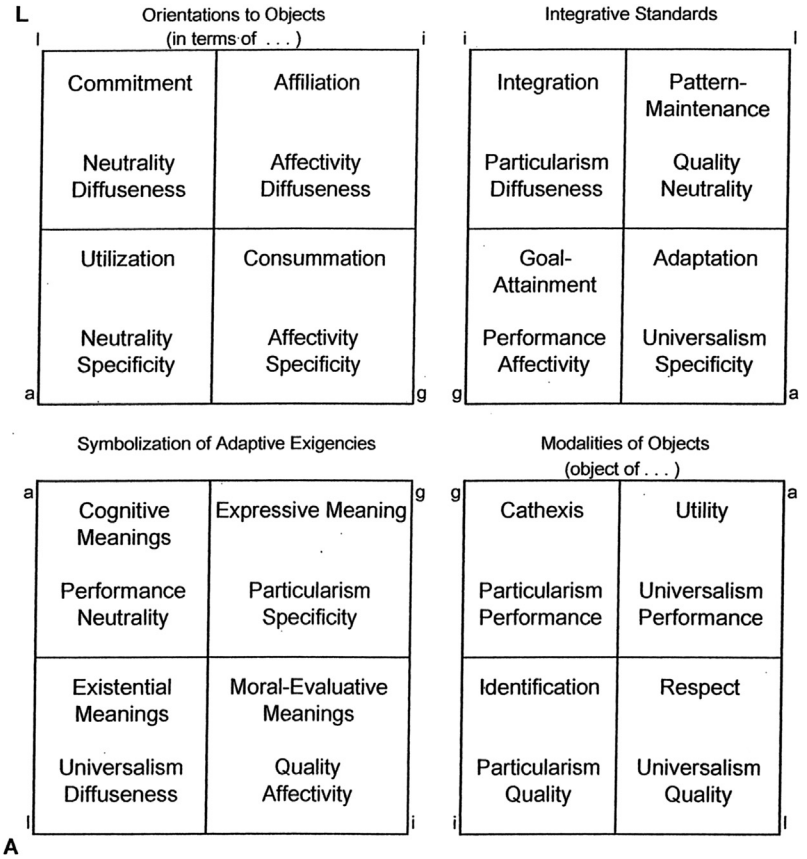

Figure 10 Parsons T., 2007. American Society. A Theory of the Societal Community, Paradigm Publishers, Boulder, p. 88.

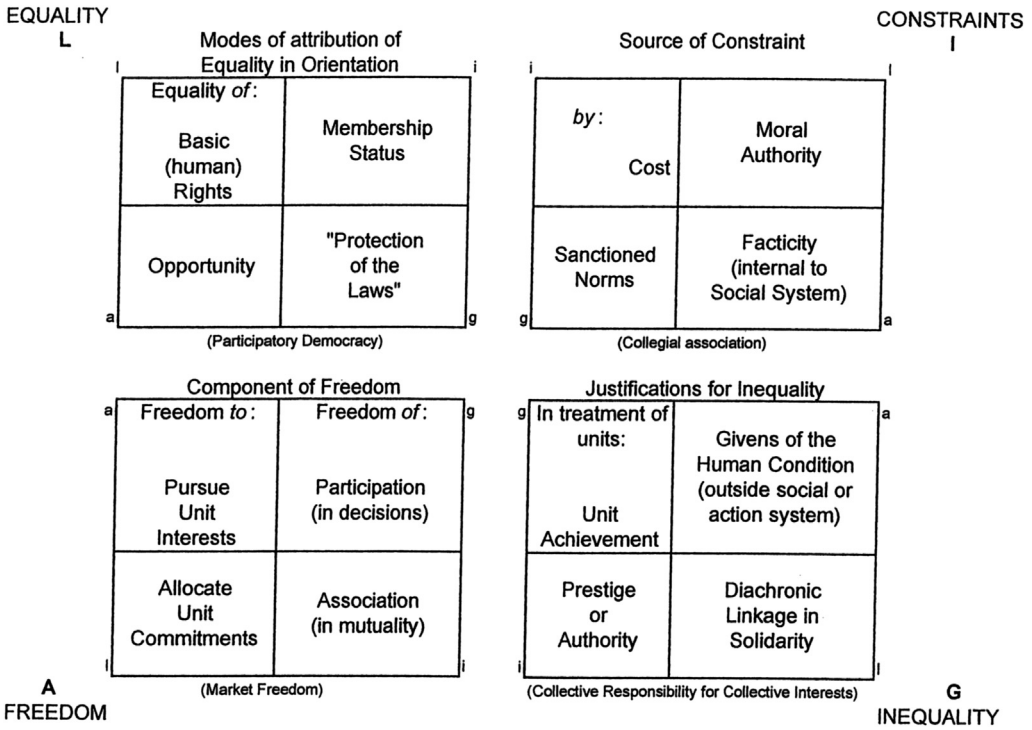

Figure 11 Parsons T., 2007. American Society. A Theory of the Societal Community, Paradigm Publishers, Boulder, p. 173.

2010–14. The scheme is, consequently, far from being forgotten. At the same time, most of the returned items deal with AGIL mostly in term of the history of sociological thought, or as a part of more broader discussions of Parsonian theory. Fresh applications of the AGIL scheme to new social science problems are a minority of the overall output. Is AGIL mainly a historical relic? As Zhu EnLai is said to have concluded about the French revolution, it may be too early to say. In fact, there are at least two ways in which AGIL maintains a contemporary significance in social theory.

First, AGIL is still a reference point for the few contemporary attempts to develop conceptual schemes for broad analyses of large-scale social systems. It is likely that AGIL has marked a point of no return, that has shaped, and shapes, the subsequent conditions for these intellectual projects. Its lasting achievement is the substitution of 'last instance' explanations – rooted in structure/superstructure assumptions – with analysis of interdependencies among analytically irreducible elements. After AGIL, comparisons among 'total' types of societies, previously common, have been increasingly rare, substituted with a nested level of analysis. The search of prime movers has also given way to inquiries into the modalities of coordination of the different.

Second, the AGIL scheme still plays a role in contemporary social theories in at least two further ways. First, there has been a looser understanding of the AGIL not as a technical scheme, but rather as a 'metaphor,' a source of inspiration for the possibilities of a nonrationalist, nonreductionist understanding of modern society. As an example, the neofunctionalist movement has adopted this nontechnical, metaphorical use of the AGIL scheme (Alexander, 1985, 1998). Second, some theorists have refused AGIL as a generalized template, but have adopted its relational logic in their own work (Donati, 2011; Habermas, 1981; Sciulli, 2010). Harrison White himself has also recently claimed the existence of interesting similarities between the AGIL scheme and contemporary work in network theory (White, 2008, p. 76). New encounters with AGIL may become more frequent in the future under the influences of two factors. The demographic change implies the emergence of a new generation of theorists who are emotionally detached from both the era of functionalism's centrality and from the 1960s' polemical reactions to it. For many of them, Parsons will be a classical theorist among many others, rather than a symbol of sanctity or pollution. Moreover, the contemporary theoretical agenda has again at its center some issues – such as the cultural dimension of social action, the need to integrate a variety of social coordination mechanisms, the key role played by societal pluralism, the new debates on modernity – that have inspired the construction of the AGIL scheme.

See also: Functionalism in Anthropology; Functionalism, History of; Luhmann, Niklas (1927–98); Parsons, Talcott (1902–79).

Bibliography

Alexander, J.C., 1984. The Modern Reconstruction of Classical Thought: Talcott Parsons. University of California Press, Berkeley.

Alexander, J.C. (Ed.), 1985. Neofunctionalism. Sage, Los Angeles.

Alexander, J.C., 1998. Neofunctionalism and After. Blackwells, Oxford.

Bales, R.F., 1950. Interaction Process Analysis: A Method for the Study of Small Groups. Addison-Wesley, Cambridge.

Bourricaud, F., 1977. L'individualisme institutionnel. Presses Universitaires de France, Paris.

Brownstein, L., 1982. Talcott Parsons' General Action Scheme. Schenkman, Boston.

Donati, P., 2011. Relational Sociology. Routledge, London.

Fararo, J., 2001. Social Action Systems: Foundation and Synthesis in Sociological Theory. Greenwood, Westport.

Fox, R., Lidz, V., Bershady, H., 2005. After Parsons. Russell Sage, New York.

Habermas, J., 1981. Theory of Communicative Action. Beacon Press, Boston.

Loh, W., 1980. AGIL-Dimensionen im Spatwerk von T. Parsons und Kombinatorik, Kölner Zeitschrift für Soziologie und Sozialpsychologie 32 (1), 130–143.

Loubser, J.J., Baum, C.R., Effrat, A., Lidz, M.V., 1976. Explorations in General Theory in Social Science. The Free Press, New York.

Luhmann, N., 1988. Warum AGIL? Kölner Zeitschrift für Soziologie und Sozialpsychologie 40, 127–139.

Mulkay, M.J., 1971. Functionalism, Exchange and Theoretical Strategy. Routledge, London.

Parsons, T., 1937. The Structure of Social Action. The Free Press, Glencoe.

Parsons, T., 1951. The Social System. The Free Press, New York.

Parsons, T., 1954. Essays in Sociological Theory, The Free Press, Glencoe, IL.

Parsons, T., 1960a. Pattern variables revisited a response to Robert Dubin. American Sociological Review 25, 467–483.

Parsons, T., 1960b. An outline of the social system. In: Parsons, T., Shils, E., Naegele, K., Pitt, J. (Eds.), Theories of Societies, vol. I. The Free Press, Glencoe, pp. 30–79.

Parsons, T., 1966. Societies: Evolutionary and Comparative Perspectives. Prentice-Hall, Englewood.

Parsons, T., 1971. The System of Modern Societies. Prentice-Hall, Englewood.

Parsons, T., 1968. Interaction-social interaction. In: Sills, D. (Ed.), The International Encyclopedia of the Social Sciences, vol. 7. The Free Press, Glencoe, pp. 429–441.

Parsons, T., 1969. Politics and Social Structure. The Free Press, New York.

Parsons, T., 1977. Social System and the Evolution of Action Theory. The Free Press, New York.

Parsons, T., 1978. Action Theory and the Human Condition. The Free Press, New York.

Parsons, T., 1991. The Marshall Lectures – the integration of economic and sociological theory, Sociological Inquiry 61 (1), 10–58.

Parsons, T., 2007. American Society. A Theory of the Societal Community. Paradigm Publishers, Boulder.

Parsons, T., 2010. In: Lidz, V., Staubmann, H. (Eds.), Actor, Situation and Normative Pattern: An Essay in the Theory of Social Action. LIT Verlag, Munchen.

Parsons, T., Bales, R., Shils, E., 1953. Working Papers in the Theory of Action. The Free Press, New York.

Parsons, T., Bales, R., 1955. Family, Socialization and Interaction Process. The Free Press, Glencoe.

Parsons, T., Shils, E. (Eds.), 1951. Toward a General Theory of Action. Harper & Row, New York.

Parsons, T., Smelser, N.J., 1956. Economy and Society. Routledge and Kegan Paul, London.

Parsons, T., Platt, G., 1973. The American University. Harvard University Press, Cambridge.

Sciortino, G., 2008. Functionalism and system theory. In: Turner, B.S. (Ed.), The New Blackwell Companion to Social Theory. Blackwell, London, pp. 106–123.

Sciortino, G., 2010. 'A single societal community with full citizenship for all': Talcott Parsons, citizenship and modern society, Journal of Classical Sociology 10 (3), 239–259.

Sciulli, D., 2010. Theory of Societal Constitutionalism: Foundations of a Non-Marxist Critical Theory. Cambridge University Press, New York.

Stinchcombe, Arthur L., 1975. A parsonian theory of traffic accidents, Sociological Inquiry 45 (1), 27–30.

Turner, B.S., 1993. Talcott Parsons, universalism and the educational revolution: democracy versus professionalism. British Journal of Sociology 44, 1–24.

Turner, B.S. (Ed.), 2000. The Talcott Parsons Reader. Blackwell, London.

Turner, B.S., 2005. Talcott Parsons's sociology of religion and the expressive revolution. Journal of Classical Sociology 5, 303–338.

White, H., 2008. Identity and Control. A Structural Theory of Social Action. Princeton University Press, Princeton.

Aging and Health in Old Age

Christine L Himes, Syracuse University, Syracuse, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by K.G. Manton, volume 1, pp. 304–310, © 2001, Elsevier Ltd.

Abstract

Trends in mortality, disability, and chronic disease at older ages are described. Over the past several decades the years of life added as life expectancy increased resulted in more years of good health for the older population. In recent years, this trend appears to be reversing. Reasons for these observed trends in health indicators are described.

Over the last 50 years, life expectancy has steadily increased for both men and women in the United States. For much of that period, those years of added life were years free from disability. In the most recent period, however, that trend toward improvement may be stalling or perhaps even reversing. Understanding the changing pattern of the trend in health in later life has implications for individuals, families, and public spending. Changes on a number of dimensions of health in the elderly population (i.e., persons aged 65+) have been documented in recent US surveys. Similar to Manton (2001), I review the magnitude and substance of the changes in mortality, disease prevalence, and functioning and disability, and consider such of those factors that are shaping the current patterns and their implications for health and healthcare.

Introduction: Dimensions of Health

The definition of what is meant by 'health' is an important part of any discussion of health status. For many decades, mortality was assumed to be an adequate measure of the health of the elderly. Rapid mortality declines since the 1960s, especially for women, however, raised the question of whether these years of added life were years of good health. Answering this question has involved a considerable effort by researchers for the past several decades. One of the important early discussions of this topic outlined the link between different dimensions of health – mortality, disease prevalence, and disability. Crimmins (1996, 2004) discussed the interrelationship of the dimensions of health status and pointed out that due to differences in the processes driving each, we should not expect all measures to move in the same way over time. An examination of the health of the older population, and particularly the trends in health status, needs to look at all dimensions and take into account their differing influences.

Of most interest to many of those interested in the well-being and care of the elderly is the extent to which health problems limit activity or create dependency. Commonly referred to as 'disability' this is an important indicator of the health and well-being of the older population. Verbrugge and Jette (1994) coined the term disablement process to describe the interrelationship of disease, functioning, and disability. In this model, diseases are seen as creating impairments that lead to limitations in physical functioning. For instance, arthritis may create joint stiffness which results in difficulty bending or grasping an object. When limitations are severe they may interfere with the ability to carry out specific activities, resulting in a disability. Environmental factors are an additional component of the process. Individuals may be able to adapt to their environment, or their task, to accommodate their limitations. In those cases, they may delay the onset of a disability. In this framework, physical limitations are seen as coming before disabilities and more closely related to the underlying disease process.

There are three different models describing how disability may change in the US population. The failure of success model, expressed by Gruenberg (1977) argued that with improvements in the treatment of some chronically disabling diseases the US would enter a period of a 'pandemic' of chronic diseases and disability. That is, it was expected that persons with chronic diseases, and the profound disabilities they can generate, would survive many more years raising the prevalence of chronic disability and the average amount of lifetime that could be expected to be lived in an impaired state (Verbrugge, 1984).

A second perspective, due to Fries (1980), postulated that a 'compression of morbidity' would occur as life expectancy reached a maximum (estimated to be about 85 years). As this survival limit was approached the age range at which deaths occurred would become smaller, leading to a rectangularization of the survival curve. However, Fries suggested, at the same time the age of chronic disease onset could become compressed. As a result, most of life would be spent in an unimpaired state.

The third model suggested that the times of chronic disabilities and diseases onset could be in a dynamic equilibrium with survival (Manton, 1982). In this model, specific diseases may be affected by healthcare advances in differing ways. These differences would affect the relationship between mortality and disability so that both total life expectancy and disability-free life expectancy could be increased. This type of change would decrease the average amount of time spent in disabled states.

While these theories attempted to explain population differences in healthy life expectancy, new developments in the field of aging and health have focused on the role of genetics in affecting healthy survival. It has been commonly observed that longevity clusters within families, those with long-lived parents and siblings are more likely to have a long life themselves. More recent studies have shown that the children

of long-lived individuals, as well, have more favorable risk factor profiles and fewer age-related diseases (Newman et al., 2011). Despite the allure of finding a single 'longevity' gene, it is likely that several genes contribute to this effect, operating in many different ways. For instance, some genes may slow the aging process overall, while others may protect an individual from a particular disease.

Results from twin registry studies (Hjelmborg et al., 2006) and the Framingham Heart Study (Murabito and Lunetta, 2011) show that the genetic influences on longevity operate primarily above age 60. The heritability estimates for overall longevity range from 20 to 30%. The longevity phenotype, however, is very heterogeneous, including a wide range of physical and cognitive functioning levels. An examination of individual aging traits in the Framingham study shows varied heritabilities (Murabito et al., 2012). Physical frailty, for instance, showed a modest heritability, about 19%. But some less complex outcomes have higher heritability; measures of walking time and handgrip strength have substantial genetic contributions with heritabilities of nearly 40%.

Both individual and population measures of health provide important clues to the overall well-being of the older population and the outlook for future changes in health and functioning.

Mortality

Life expectancy itself is one important indicator of health status in a population and reductions in death rates are often used as an indication of improvements in health and well-being. Life expectancy at birth is 78.7 years in 2011; 81.1 years for US women; and 76.3 years for US men (Hoyert and Xu, 2012). The difference in life expectancy between men and women, now just under 5 years, has been declining since 1979, when the gap was nearly 8 years. There is a persistent racial gap in life expectancy as well, blacks have life expectancy of 75.3 in 2011 compared with 79.0 for whites.

More important as an indicator of late life health, however, is the life expectancy at age 65. For the total US population, life expectancy at age 65 has increased from 17.2 years in 1990 to 19.2 years in 2009 (Arias, 2012). White women have the longest expected survival, 20.3 years and black men the shortest, 15.8 years. In 2009 for the first time life expectancy at age 75 exceeds 10 years for all demographic groups, men and women, whites and blacks.

These gains in life expectancy reflect improvements in the underlying age-specific mortality rates. Between 1935 and 2010 death rates for persons aged between 65 and 74 fell by 62% (Hoyert, 2012). In the most recent decade, 2000 to 2010, the largest reductions in mortality occurred for males, with non-Hispanic black men experiencing the largest declines. All three major causes of death, heart disease, cancer, and stroke showed declines (Minino, 2013). Of growing importance for later-life health is the significance of Alzheimer's disease (AD) as a cause of death. Mortality from AD has steadily increased during the last 30 years, becoming the fifth leading cause of death for people aged 65 and over (Tejada-Vera, 2013). Alzheimer's disease accounts for between 5 and 15% of all deaths in older people. Some of this increase is due misclassification of AD into other categories leading to a significant undercounting of deaths attributable to AD. Changes in the coding of cause of death, in place since 1999, have allowed for estimation of a more consistent trend (Moschetti et al., 2012). Since Alzheimer's disease is often preceded by many years of disability, the impact of this increase on health care and family costs is an important consideration for the future.

Disability and Functioning

Since disability is a major driver of health and personal care costs, tremendous effort in recent years has been expended on describing and documenting trends in disability levels in the older population. Disability is most often defined relative to the ability to live independently and take care of one's personal care needs. The inability to perform the most basic activities, referred to as activities of daily living (ADLs), is considered the most severe form of disability. This standard set of activities includes bathing, dressing, eating, toileting, transferring from bed and chairs, and walking around inside. Less severe disability is measured by the inability to perform instrumental activities of daily living (IADLs), routine household tasks shopping, cleaning, cooking, and going places.

In his earlier discussion, Manton (2001) describes the US declines in functional disability first documented in the 1982 to 1989 National Long Term Care Surveys (NLTCS). The NLTCS are sets of longitudinally related surveys designed to assess changes in functional status, social conditions, and Medicare and LTC service use in the US elderly population. Through the 1990s, other US population studies found even larger declines than that first reported by Manton et al. (1997). Freedman and Martin (1998) using the 1991–1996 Survey of Income and Program Participation reported a decline which existed at the higher levels of disability and at advanced (85+) ages. Waidmann and Liu (1998) also found confirmation of declines in the 1993–1996 Current Medicare Beneficiary Survey. Crimmins et al. (1997) found evidence for declines in the 1984 Supplement on Aging and the LSOA (longitudinal study of aging) from 1986 to 1990. Evidence of declines has been found in the 1985 and 1995 Supplements on Aging to the National Health Interview Survey (Crimmins and Saito, 2000). Evidence for declines in European countries was found in Waidmann and Manton (1998).

These various studies, while finding relatively consistent evidence of declines in limitations in IADLs and in functional limitations, produced mixed results on the trend in the most severe type of disability, limitations in ADLs. In an effort to summarize and harmonize the results of these studies a technical working group was formed in 2002, drawing together researchers with experience from five national data sets covering the US population: the Health and Retirement Study (HRS), the Medicare Current Beneficiary Survey (MCBS), the NHIS, the National Long Term Care Survey (NLTCS), and the Supplements on Aging (SOAs). The working group concluded that while evidence of a decline in activity limitations was mixed in the 1980s, by the mid-1990s there was a clear downward trend in reported limitations across surveys, despite differences in question wording and study design (Freedman et al., 2004). While this trend was most pronounced in limitations in

IADLs and in functional limitations, from the mid- to late 1990s the proportion of the older population reporting difficulty in performing ADLs declined steadily, too.

This promising news, described as one of the most significant advances in the health and well-being of Americans in the past quarter-century (Schoeni et al., 2008), has been tempered by more recent evidence. In the most recent decades, since 2000, these trends appear to be changing. A recent summary (Freedman et al., 2013), carried out by many of the same researchers involved in the technical working group, compares results from four of the same national data sets covering the US population with the substitution of the National Health and Nutrition Examination Survey (NHANES) for the SOAs. All of these surveys cover the period 2000–08 and the NLTCS and MCBS allow analysis of both community and institutional populations. Carefully controlling for differences in question wording, coverage, and coding, trends in ADL and IADL limitations were estimated from each study. In contrast to the steady downward trend observed in the 1980s and 1990s, the more recent years of data show no change in the proportion of the older population reporting difficulty in performing these personal care and household activities.

More alarming than the apparent stalling in improvement in functioning at older ages is evidence that health at younger ages, those aged 50–64, is not improving and, perhaps, declining. Analysis of the NHIS (Lakdawalla et al., 2004) found that between 1984 and 1996 there was an increase in needs for help among those in middle age, although the numbers were small. Using the Behavioral Risk Factor Surveillance Survey (BRFSS), Zack et al. (2004) found an increase from 1993 to 2001 in those in middle age reporting poor or fair health. Worsening functioning and health were observed from 1992 to 2004 in the HRS (Soldo et al., 2007) for younger cohorts. Examining a variety of health measures, Martin et al. (2009) conclude that trends for the Baby Boom generation vary by measure, period, and age group. For mortality and self-reported health, baby boomers are doing better, while trends for functional limitations and the need for IADL help are stable. Other analyses find that the proportion of people aged 50–64 who report needing help with personal care, while small, increased between 1997 and 2007 (Martin et al., 2010b).

Disease Prevalence

During the period while mortality rates were falling and disability was clearly declining, prior to 2000, the prevalence of most chronic diseases increased. Between 1984 and 1994 increases in self-reports of arthritis, heart disease, cancer, diabetes, stroke, and osteoporosis were observed in the population aged 70 and older (Crimmins and Saito, 2000; Freedman and Martin, 2000). Manton et al. (1995) observed that some conditions decreased from the 1980s to the 1990s (arthritis, circulatory, and cerebrovascular conditions), while others increased (pneumonia, bronchitis, and diabetes).

In more recent time periods, after 2000, the prevalence of chronic diseases continued to increase in the older population. Importantly, musculoskeletal conditions, like arthritis, which have strong links to functional impairments and disability have shown increases. During the time period, 1997–2004, continuing increases in the prevalence of arthritis, cancer, diabetes were reported in the NHIS, although no significant change was observed for heart disease, stroke, or lung conditions (Freedman et al., 2007). This trend is observed in younger cohorts, too. Among those aged 50–64 an increase in musculoskeletal conditions are the most commonly cited causes of disability (Martin et al., 2010b). Unclear is the effect of cognitive impairments on the disability trends. While the prevalence of Alzheimer's disease has increased, the contribution of this condition to overall disability is harder to measure in our large national surveys due to proxy respondents and the greater likelihood of institutionalization.

Not only has chronic disease prevalence increased, but the likelihood of reporting two or more chronic conditions has increased in recent years. Analysis of data from the NHIS shows that among those aged 65 and over the proportion of those reporting two or more chronic conditions (defined as hypertension, heart disease, diabetes, cancer, stroke, chronic bronchitis, emphysema, asthma, or kidney disease) increased from 37 to 45% between 2000 and 2010 (Fried et al., 2012). The most common combination was hypertension and heart disease (21%) but one of the biggest increases was seen in the percentage with both hypertension and diabetes; an increase from 9 to 15% over the 10-year period. These increases occurred for both men and women and for all racial groups.

Causes of Disability Trends

The reasons for the decline in disability observed during the close of the twentieth century and the reasons for the more recent stagnation in improvement remain unclear. The increased prevalence of chronic diseases in tandem with disability declines indicates some possible explanations (Schoeni et al., 2008). One theory is that the earlier diagnosis of chronic conditions has led to higher disease prevalence at a stage before impairments are evident. Furthermore, earlier treatment of chronic conditions may postpone the debilitating effects of the disease. Better medical treatments and improved drug therapies may make diseases less disabling.

The use of assistive devices and accommodations is another possible explanation for the seeming paradox of increased disease and decreased disability. To the extent that disability is a function of both personal abilities and the environmental context in which activities are performed, changes in the environment alone may reduce reported disability. The IADLs, which saw the largest decline, could have been influenced by changes in the socioeconomic environment which allow changes in socially defined gender roles (e.g., men doing more grocery shopping or laundry).

The reversal of the trend in disability improvement is very new and likely to engage the attention of researchers in the coming decade. Trends in obesity are often cited as a primary driver of the reversal and may operate through diabetes, arthritis, and heart conditions (Sturm et al., 2004), but others attribute less of the increase directly to obesity (Martin et al., 2010a). The past improvements may have reflected a change in the disablement process, particularly the use of technology

and assistive devices (Schoeni et al., 2008). Future declines in disability may require even newer technologies if we have reached a saturation of the use of these devices or run up to the limits of their ability to compensate for functional changes. The lack of health improvements in the cohorts soon to reach advanced age, coupled with the larger size of these age cohorts, may signal a costly reversal of past trends in the older populations need for care.

See also: Adult Mortality in Industrialized Societies; Mortality Crossover; Mortality Differentials: Selection and Causation; Mortality of the Oldest-old; Mortality, Biodemography of; Mortality, Epidemiological, and Health Transitions; Population Aging: Economic and Social Consequences.

Bibliography

Arias, E., 2012. United States life tables, 2008. National Vital Statistics Reports (61) 3. National Center for Health Statistics, Hyattsville, MD.

Crimmins, E.M., 1996. Mixed trends in population health among older adults. Journal of Gerontology: Social Sciences 51, S223–S225.

Crimmins, E.M., 2004. Trends in the health of the elderly. Annual Review of Public Health 25, 79–98.

Crimmins, E.M., Saito, Y., 2000. Change in the prevalence of diseases among older Americans: 1984–1994. Demographic Research 3 (9). Available online at: http://www.demographic-research.org/volumes/vol3/9/3-9.pdf.

Crimmins, E.M., Saito, Y., Reynolds, S.L., 1997. Further evidence on recent trends in the prevalence and incidence of disability among older Americans from two sources: the LSOA and the NHIS. Journals of Gerontology Series B—Psychological Sciences and Social Sciences 52 (2), S59–S71.

Freedman, V.A., Martin, L.G., 1998. Understanding trends in functional limitations among older Americans. American Journal of Public Health 88 (10), 1457–1462.

Freedman, V.A., Martin, L.G., 2000. Contribution of chronic conditions to aggregate changes in old-age functioning. American Journal of Public Health 90 (11), 1755–1760.

Freedman, V.A., Crimmins, E., Schoeni, R.F., et al., 2004. Resolving inconsistencies in trends in old-age disability: report from a technical working group. Demography 41 (3), 417–441.

Freedman, V.A., Schoeni, R.F., Martin, L.G., Cornman, J.C., 2007. Chronic conditions and the decline in late-life disability. Demography 44 (3), 459–477.

Freedman, V.A., Spillman, B.C., Andreski, P.M., et al., 2013. Trends in late-life activity limitations in the United States: an update from five national surveys. Demography 50, 661–671.

Fried, V.M., Bernstein, A.B., Bush, M.A., 2012. Multiple Chronic Conditions among Adults Aged 45 and over: Trends over the Past 10 Years. NCHS Data Brief, No. 100. National Center for Health Statistics, Hyattsville, MD.

Fries, J.F., 1980. Aging, natural death, and the compression of morbidity. NEJM 303, 130–135.

Gruenberg, R., 1977. The failure of success. Milbank Quarterly 55, 3–24.

Hjelmborg, J.V., Iachine, I., Skytthe, A., et al., 2006. Genetic influence on human lifespan and longevity. Human Genetics 119 (3), 312–321.

Hoyert, D., 2012. 75 Years of Mortality in the United States, 1935–2012. NCHS Data Brief No. 88. National Center for Health Statistics, Hyattsville, MD.

Hoyert, D.L., Xu, J.Q., 2012. Deaths: Preliminary Data for 2011. National Vital Statistics Reports (61) 6. National Center for Health Statistics, Hyattsville, MD.

Lakdawalla, D.N., Bhattacharya, J., Goldman, D.P., 2004. Are the young becoming more disabled? Health Affairs 23 (1), 168–288.

Manton, K.G., 1982. Changing concepts of morbidity and mortality in the elderly population. Milbank Quarterly 60, 183–244.

Manton, K.G., 2001. Aging and health in old age. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Elsevier, Oxford, pp. 304–310.

Manton, K.G., Stallard, E., Corder, L., 1995. Changes in morbidity and chronic disability in the U.S. elderly population: evidence from the 1982, 1984, and 1989 National Long Term Care Surveys. Journals of Gerontology B: Psychological and Social Sciences 50, S194–S204.

Manton, K.G., Corder, L., Stallard, E., 1997. Chronic disability trends in elderly United States populations 1982 to 1994. Proceedings of the National Academy of Sciences of the USA 94, 2593–2598.

Martin, L.G., Freedman, V.A., Schoeni, R.F., Andreski, P.M., 2009. Health and functioning among baby boomers approaching 60. Journal of Gerontology: Social Sciences 64B (3), 369–377.

Martin, L.G., Schoeni, R.F., Andreski, P.M., 2010a. Trends in health of older adults in the United States: past, present, future. Demography 47 (Suppl.), S17–S40.

Martin, L.G., Freedman, V.A., Schoeni, R.F., Andreski, P.M., 2010b. Trends in disability and related conditions among people ages fifty to sixty-four. Health Affairs 29 (4), 725–731.

Minino, A.M., 2013. Death in the United States, 2011. NCHS Data Brief, No. 115. National Center for Health Statistics, Hyattsville, MD.

Moschetti, K., Cummings, P.L., Sorvillo, F., Kuo, T., 2012. Burden of Alzheimer's disease-related mortality in the United States, 1999–2008. Journal of the American Geriatrics Society 60 (8), 1509–1514.

Murabito, J.M., Lunetta, K.L., 2011. Genetics of human longevity and health aging. In: Newman, A.B., Cauley, J.A. (Eds.), The Epidemiology of Aging. Springer, New York.

Murabito, J.M., Yuan, R., Lunetta, K.L., 2012. The search for longevity and healthy aging genes: insights from epidemiological studies and samples of long-lived individuals. Journal of Gerontology: Medical Sciences 67A (5), 470–479.

Newman, A.B., Glynn, N.W., Taylor, C.A., et al., 2011. Health and function of participants in the long life family study: a comparison with other cohorts. Aging 3 (1), 63–76.

Schoeni, R.F., Freedman, V.A., Martin, L.G., 2008. Why is late-life disability declining? Milbank Quarterly 86 (1), 47–89.

Soldo, B.J., Mitchell, O.S., Tfaily, R., McCabe, J.F., 2007. Cross-cohort differences in health on the verge of retirement. In: Madrian, B., Mitchell, O.S., Soldo, B.J. (Eds.), Redefining Retirement: How Will Boomers Fare? Oxford University Press, New York, pp. 138–158.

Sturm, R., Ringel, J.S., Andreyeva, T., 2004. Increasing obesity rates and disability trends. Health Affairs 23 (2), 199–205.

Tejada-Vera, B., 2013. Mortality from Alzheimer's Disease in the United States: Data for 2000 and 2010. NCHS Data Brief, No. 116. National Center for Health Statistics, Hyattsville, MD.

Verbrugge, L., 1984. Longer life but worsening health? Trends in health and mortality of middle-aged and older persons. Milbank Quarterly 62, 475–519.

Verbrugge, L., Jette, A.M., 1994. The disablement process. Social Science and Medicine 38 (1), 1–14.

Waidmann, T.A., Liu, K., 1998. Disability Trends among the Elderly and Implications for Future Medicare Spending. Joint Statistical Meetings, Dallas, TX.

Waidmann, T., Manton, K.G., 1998. International Evidence on Disability Trends among the Elderly. Final Report for the Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation.

Zack, M.M., Moriarty, D.G., Stroup, D.F., et al., 2004. Worsening trends in adult health-related quality of life and self-rated health-United States, 1992–2001. Public Health Reports 119, 493–505.

Aging and Memory

Lars-Göran Nilsson, Stockholm Brain Institute, Stockholm University, Stockholm, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article first summarizes the current state of the art regarding research on how five different memory systems change as a function of age. These systems are procedural memory, perceptual representation system, semantic memory, working memory, and episodic memory. Working memory and, especially, episodic memory are the memory systems most affected by increasing age. The second part of the article deals with two types of data that have been used as the basis for theorizing about aging and memory. These are cross-sectional data and longitudinal data. Most of the data that have been collected so far in this field of research are cross-sectional data. It is claimed in this article that such data say very little about the individual development of memory in adulthood and old age. It is claimed that most of the variance in such data are due to cohort differences rather than to the development of memory. The third part of the article is about the large variability in performance in the aging population. Most of the research in aging and memory has conceived of the aging population as a homogeneous group and typically just means and standard deviations for the whole group have been reported. Recent research, however, has demonstrated that there is much important information to be gained when considering subgroups of the whole age group. Three such subgroups have been suggested as one possibility. These three groups are decliners, average, and maintainers. Future research will demonstrate whether there are other subdivisions that are more appropriate.

Research on aging and memory has increased enormously in amount and quality during the last few decades. According to Web of Science, regarding the quantity of contributions, there were six papers published in this field in 1970, 14 in 1980, 83 in 1990, 1797 in 2000, and 5381 in 2010. A major reason for this development is, of course, that researchers in this field have come to realize that there are many important issues to cover, not only because of the benefit for the aging individuals themselves, but also for the society in general in taking care of the aging population. Additionally, researchers are, of course, also interested in obtaining knowledge growth in an interdisciplinary sense about the underlying features of individual development and decline in cognitive processing in adulthood and old age.

There is a large variability in cognitive functioning in old age. Some old people show no or very little cognitive decline and continue a life up to very old age in what can be characterized as successful aging (Rowe and Kahn, 1987). Others remain stable in cognitive functioning at an average level up to old age, and then there is a decline in performance sometimes leading to dementia or other forms of neurodegenerative disorders. Still others start to show cognitive decline relatively early and will eventually develop signs of dementia and some will get the dementia diagnosis.

An issue that has been brought up in recent years on the research agenda in research laboratories is when age-related cognitive decline begins (e.g., Salthouse, 2009). Is dementia inevitable is another question that has started to be discussed in the literature (e.g., Gondo et al., 2014). Still another question in this research is whether cognitive signs and biological markers can be found early in life for persons who later will develop dementia (e.g., Nilsson, 2014). The general question here is of great societal interest as to whether psychological and medical treatments can be offered to these people, who show early signs as risk factors for dementia long before diagnosis of dementia. In hopeful quarters, the time of detection of early markers is the critical factor for the success of treatment. When the dementia development has gone on too long and a person has been diagnosed as demented, it is too late. The degeneration has gone too far and the existing treatments are not of much use. These and other issues are intensively studied in many laboratories around the world at present, contributing to the great increase in the number of published papers in this field. The present article will cover some of these domains of research.

The most primary research areas in the literature previously covered how different memory systems develop and age differently, and how different cognitive domains differ in sensitivity in detecting a decline in performance. These issues are still of great interest in the scientific community. However, in recent years, there has been a change in focus of this research. Previous research was focused on the theoretical emphasis on the development of different memory systems and cognitive domains for the sake of interdisciplinary knowledge growth, whereas recent years of research in this general field have also been keen on providing solid data on early signs of dementia and other neurodegenerative diseases for a possible application of research to individuals, who are at risk for dementia and possible treatments, and for the sake of society and the costs for societies, when life expectancy increases in most countries around the world and more and more people become demented or acquire other neurodegenerative disorders. One important reason for this development is, thus, that life expectancy is increasing in most countries in the world, partly because of improvements of health care in these countries, and an increased awareness that people can affect their own life expectancy by a healthy life style including concerns about exercise, social networks, intellectual activities, and engagement in what is happening in their close surroundings with relatives and friends and more generally in the surrounding world.

Another general change in research in this field of science has been the insight from many researchers that cross-sectional data are not optimal when trying to explore the development of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.51045-9

cognitive aging across the life span (e.g., Nilsson et al., 2009; Schaie, 2009). Instead, it is argued among many that longitudinal studies should be conducted, since cross-sectional studies are primarily picking up cohort differences among participants in such studies. There are also difficulties with longitudinal studies (e.g., test–retest effects and attrition effects). However, these effects can be accounted for and controlled in carefully planned longitudinal studies (e.g., Josefsson et al., 2012; Rönnlund et al., 2005).

Still another new orientation of the research on aging and memory is the interest in trying to combine classical research on cognitive psychology, with research on brain imaging and the role of genetics for understanding how memory, and cognition in general, changes as people become older.

As already mentioned, there has also been an increasing interest in recent years in this research field to take into account the great variability in performance in memory tasks and in cognitive tasks in general, especially so when observing that this variability increases with age. Previously, means and standard deviations were presented, without any closer interest in what the variability in performance across the life span might mean. As argued by many researchers (e.g., Baltes, 1968; Schaie, 1996) for a long time, this variability in performance might carry considerable important information about the development of memory and other cognitive functions across the life span. For example, as some researchers have tried, is it of interest to study whether people develop differently, when they are classified as average in performance, performing at a higher level than average, or performing at a lower level than the average participant in a study. Although this classification of participants is crude, it may carry important information about variables related to each of these categories of participants in a study of aging and memory.

Outline of the Article

Following the traditional presentation of the state of the art of research on aging and memory, I will continue with a section on the difference in results shown in cross-sectional studies and longitudinal studies in research on aging and memory. Closely related to this is the topic of the great variability in performance on memory tasks and cognitive tasks in general.

State of the Art Regarding Current Knowledge on Memory Systems and Cognitive Functioning in Relation to Age

In a seminal paper, Bäckman et al. (2001) described the state of the art of knowledge on this topic. The essence of this article was that different memory systems, according to the terminology proposed by Tulving (1983), age at different rates.

The oldest memory system, according to Tulving (1983), both phylogenetically and ontogenetically, is procedural memory. This memory system is responsible for well-acquired motoric abilities like walking, dancing, swimming, bicycling, and brushing teeth. In general, it is about the acquisition of motor, perceptual, or cognitive abilities. These abilities develop gradually as an effect of practice (Nyberg and Tulving, 1996). This memory system does not age in the same way as other memory systems when persons grow older. If you have learned to go on a bicycle, or to swim, you will never forget this. To the best of my knowledge, there is nothing in the recent literature that has provided evidence for a change in this view. It should be acknowledged though that surprisingly little research has been conducted regarding how this memory system develops as a function of age.

The next memory system in Tulving's terminology is that of perceptual representation system (PRS). This memory system is generally responsible for navigating in the surrounding world. In the laboratory, this system is usually tested by means of word fragment completion tasks or word stem completion tasks. In a word-fragment completion task, participants are first presented with a list of words and asked to classify these, for example, as abstract or concrete words. Nothing is said in the instructions about a forthcoming memory test. Usually after having completed some other cognitive tasks, the experimenter returns to this task and tells the participant that he/she will now see fragments of the words that were shown previously. The task of the participant is to say the first word that comes to mind when each fragment is shown. Interestingly, participants manage this task very well independently of whether they are brain injured or healthy. If, however, the experimenter instead says that the task of the participant is to use the fragment presented as a cue to remember each word that was previously presented, the result pattern changes dramatically. Participants with a brain injury or some other brain disorder (e.g., dementia) do not manage the task at all, whereas healthy controls manage it quite well. In this latter task, participants are demanded by the instructions to use a conscious recollection of a previous episode, i.e., the previous presentation of the list of words. Brain-damaged people fail this task, whereas healthy people manage it reasonably well. This basic finding appears still to hold true. In addition to priming tasks (e.g., word fragment completion and word stem completion tasks), a fair amount of research on this topic of age differences in PRS has been conducted using a paradigm that aim at differentiating an automatic component of implicit retrieval and another component of controlled, deliberate, and explicit retrieval (Jacoby, 1991). Several researchers, who have used this procedure in cognitive aging studies (e.g., Jennings and Jacoby, 1993; Titov and Knight, 1997), have demonstrated a significant age-related deficit in the controlled, explicit retrieval component together with no age differences in the automatic, implicit component. In another experimental paradigm developed by Gardiner and Java (1993), the basic aim was the same, namely, to differentiate between the contribution of a component that is based on conscious recollection of a previously encountered event (these responses are called 'remember' responses) and other responses that are not consciously recalled from a previous event, but rather merely recalled on the basis of the participants having a feeling of familiarity of the event (these responses are usually referred to as 'know' responses in this paradigm). Research has been reported that age differences are substantial for remember responses, but minimal and nonsignificant for know responses (e.g., Mäntylä, 1993).

A third memory system in Tulving's terminology is semantic memory. This system is about general knowledge, e.g., what is the chemical designation of regular table salt?, or what is the

capital of Spain?, or what is the opposite of warm? There is a general consensus in this literature that there is no evidence available for a lower performance of such general knowledge tasks for older than for younger participants. Going all the way back to the time when the term semantic memory was coined (Collins and Quillian, 1969; Collins and Loftus, 1975; Quillian, 1966), there has been a basic assumption that the basis for this memory system is an internal lexicon. It is furthermore assumed that this lexicon is organized hierarchically in a top-down structure with concepts, terms, words, and associations in networks of nodes at different levels. At a top level, one category might be 'flowers,' with wild flowers and garden flowers, respectively, at a lower level. At still another lower level flowers may be categorized as red flowers or yellow flowers. At still another lower level both rose and tulip might be represented as red flowers according to one categorization and yellow flowers according to another possible categorization.

With regard to cognitive aging, nice and convincing data were presented by Laver and Burke (1993) that the internal lexicon remains stable across the life span. Young and old individuals are quite similar in the word associations they generate, they access categorical associations in a similar way, and vocabulary scores remain stable or even increase as a function of age at least until very late in old age (see Bäckman and Nilsson, 1996; for an overview). However, there are reports of age differences, such that older people have greater difficulties in word-finding tasks, e.g., proper names (e.g., Crook and West, 1990), and generating items in tasks of word fluency (e.g., Bäckman and Nilsson, 1996). This finding by Bäckman and Nilsson (1996) brings about the question of to what extent cognitive tasks usually used in this research are process pure. When this study was conducted some 20 years ago (Bäckman and Nilsson, 1996), fluency tests, even letter-fluency tasks, were regarded as adequate tasks to assess semantic memory. Later, however, such letter-fluency tasks have come to be regarded as tasks measuring working memory and executive functions. This has been shown by means of brain imaging techniques showing frontal lobe activity when generating words in a letter-fluency task.

A fourth memory system is working memory. This system was previously called short-term memory or primary memory. Baddeley and Hitch (1974) and Baddeley (1986) coined the term working memory to include not only the passive maintenance of information in mind, but also the more active processing of information in working memory, including executive functioning like planning, switching, and updating. Brain imaging studies have demonstrated that prefrontal regions of the brain are typically activated when participants are solving working-memory tasks (see Cabeza and Nyberg, 2000; for a review). Working memory performance does not in general show any age differences when the task demands are simple like maintaining information in memory (e.g., Corey-Bloom et al., 1996; Gregoire and Van der Linden, 1997). However, as the cognitive demands increase in the tasks used to assess working memory, involving more executive functioning, older participants show a lower performance than younger participants. This relationship was convincingly demonstrated in an early meta-analysis by Verhagen et al. (1993). When the task requires simultaneous storage and processing of information, age differences are in general robust (e.g., Brébion et al., 1995). This conclusion seems to hold true even today (e.g., de Frias et al., 2005, 2010). These findings suggest that working memory is not a unitary concept. Neural support for such a conclusion was reported by Rypma and D'Esposito (2000), when demonstrating no age-related differences in the ventrolateral prefrontal cortex during working memory performance, although there were age-related decreases in the dorsolateral prefrontal cortex.

Episodic memory is the fifth memory system according to Tulving's (1983) terminology. This memory system deals with encoding and retrieval of information in a certain context of time and place. According to Tulving, it is the youngest memory system, both phylogenetically and ontogenetically. Episodic memory is easily and strongly affected by many variables, including age (Nilsson et al., 1997). Episodic memory is different from other memory systems in that it requires a conscious recollection of a particular study episode and it has been suggested that it is this conscious travel backward in time that makes it so vulnerable for a decrement in performance as a function of age.

In trying to explain the age differences in episodic memory that have been empirically demonstrated, several different approaches have been attempted. Early in the history of cognitive aging research, age-related differences involved references to specific memory processes operationalized by the experimental paradigms used (e.g., levels of processing, organization, elaboration). This research and these theoretical attempts were nicely reviewed by Craik (1977; Craik and Jennings, 1992). It was often claimed that old persons show a lower performance than young persons, because they fail in spontaneously adopting a deep semantic encoding, or they do not organize the to-be-remembered materials as efficiently as young persons. Such process-oriented explanations are still proposed in research articles in the field every now and then. However, soon a different approach was attempted in trying to explain the observed age differences in episodic memory tasks. More general constructs were proposed to be the reason for memory deficits in old age. Processing speed was one such construct that was proposed. It was claimed that cognitive processing gets slower as people become older. Salthouse (1996) was probably the most pronounced advocate for this view. Salthouse and his collaborators published several papers in support of this view. Others have argued against this view and the discussion is still going on. For example, Sternäng et al. (2008) argued that this view may hold for cross-sectional data, but does not hold for data based on narrow-cohort age designs or longitudinal designs.

Working memory capacity is another general construct suggested to be responsible for a declining performance level as a function of age of participants in studies trying to find an explanation for age differences in episodic memory. It has been demonstrated that statistical control of working memory capacity attenuates or even eliminates age-related differences in episodic memory using a correlational approach rather than an experimental approach (Hultsch et al., 1998). This meditational approach of explaining age differences in episodic memory has certainly been refreshing the understanding of age differences in episodic memory and potential explanations of these. However, it has also been

claimed (e.g., Zacks et al., 1999) that the links between these constructs and performance in complex episodic memory tasks are as of yet relatively poorly understood. More research is needed regarding these links to episodic memory in order to reach consensus about the reasons for these age differences. Recently, Del Missier et al. (2013) presented a new approach to examine the relationship between working memory and episodic memory.

Still another approach to explain the age differences in episodic memory is based on the assumption that there are multiple conditions with effects on brain functioning that could be involved. Among these conditions are demographic factors, lifestyle factors, health-related factors, and genetic factors that might influence episodic memory performance with main effects or in interactions. This approach to an understanding of age differences in episodic memory is a very complex one. It is not at all understood how these factors might interact. These factors are presently being explored in many studies with respect to main effects of each individual factor. Progress is being made, but when evaluating the empirical results, one should be cautious in interpretation, because the influences of interactions may completely invalidate the interpretation of the results.

With respect to demographic variables affecting episodic performance in late life, education and sex are two very important factors. Regarding education, several studies have shown that there is a strong relationship to memory performance. In one relatively recent study, we (Habib et al., 2007) demonstrated that education is the single most important factor characterizing individuals, who we classified as successfully aged on the basis of cognitive performance in a large test battery. We used the number of years in formal education as a measure of the education attained. We realize that this is a very crude measure that does not take into account, the quality of the schooling and it does not take into account the preparation for this schooling for each individual pupil when starting school. It would be interesting to have early data available of grades or evaluations by teachers about the intellectual level of each individual pupil, who become a participant in a longitudinal study on aging and memory covering a large section of the life span. Such data could serve many purposes. For example, such data could serve as an important starting point for what the intellectual level is before further development in school, university, and professional life. To the best of my knowledge, no study on cognitive aging has reported such data. However, the collection of such data is in progress within a longitudinal study on memory, aging, and health, called the Betula Study (Nilsson et al., 1997, 2004, 2006).

Sex is also a demographic variable that has a strong effect on episodic memory performance. It has been demonstrated in many studies that there is a female superiority in episodic memory performance (e.g., Herlitz et al., 1997). Both biological factors (e.g., Collaer and Hines, 1995) and social factors (Eagly and Wood, 1999) have been proposed, but there is still no consensus in the research literature about which is the most likely explanation of this female superiority in episodic memory performance.

Lifestyle factors have also been proposed as a potential explanation for age differences in episodic memory. In general, the picture is that a healthy lifestyle with physical and cognitive exercise is good for cognitive aging (e.g., Christensen et al., 1996) and unhealthy lifestyle with lack of social, cognitive, and physical activities leads to a lower performance in episodic memory tasks (e.g., Pushkar et al., 1999). Substance use has also been examined in relation to cognitive aging, but the picture is quite mixed. When effects of substance use have been found, in general, the effects have been relatively small. Some researchers (e.g., Hill, 1989) have reported negative effects of substance use on cognitive performance, whereas others (e.g., Elias et al., 1999) have found positive effects.

Stress is another factor that has been proposed as a risk factor for cognitive deficits in old age (e.g., Johansson et al., 2010; Norton et al., 2009; Persson and Skoog, 1996). Detrimental effects of stress on cognitive performance have not been found across the board (e.g., Fountoulakis et al., 2011; Motomura et al., 1998). In an ongoing study in our own laboratory with Betula data, we also failed to demonstrate any negative effects of stress (Sundström et al., 2014). The lesson to learn from this is perhaps that the effects of stress, if they exist, are relatively small, the overall picture of data patterns is rather mixed, and perhaps there is a large risk for false positives in many studies.

Cross-Sectional Data versus Longitudinal Data for Exploring Adult Life Development in Cognitive Aging

Except for the enthusiastic and inspiring attempts by Karl W. Schaie to bring this topic forward to an important issue of the research agenda of this field, there are very few other cognitive-aging researchers, who have brought up this for discussion. In recent years, many researchers have realized the importance of Schaie's claim that longitudinal data are what is needed, when the research question is to seek to understand how individuals develop cognitively in adulthood and old age.

Schaie's basic claim was that there are pros and cons with all research designs (cross-sectional or longitudinal). It should be noted already here that the majority of the studies that have been carried out on aging and memory are cross-sectional in design. Relatively few studies published in this area have used a longitudinal design. It will be argued here that cross-sectional designs primarily pick up cohort effects, when comparing old individuals who perhaps were born in the beginning of the twentieth century with those who were born in the middle of the twentieth century or even later. Some studies can illustrate excellently what these cohort differences are contributing to how these individuals are doing cognitively, when tested several years later. In two recent studies, we have demonstrated how cognitive functions change across generations as a result of cohort differences (Rönnlund and Nilsson, 2006, 2009). In these two studies, we demonstrated that these so-called Flynn effects (Flynn, 1984, 1987) could be attributed to three factors: nutrition, education, and sibship size. These factors accounted for almost all variance in performances on episodic memory, semantic memory, and visuospatial ability. Nutrition (with body height as a proxy), education (with number of years in formal schooling as a proxy), and sibship size (number of siblings in the family was used as a proxy) accounted for

99.6% of the variance in episodic memory tasks, 99.9% of the variance in semantic memory tasks, and 94.6% of the variance in tasks of visuospatial ability.

We have examined performance in episodic memory tasks and semantic memory tasks in one study (Rönnlund et al., 2005), in another study on performance in a Tower of Hanoi task (Rönnlund et al., 2008), and on performance in a visuospatial task in still another study (Rönnlund and Nilsson, 2006) by comparing the data patterns when using a cross-sectional design and when using a longitudinal design. The data we used in these two studies emanate from a longitudinal, cohort study on memory, health, and aging. The study has been going on since 1988, with a total of about 4500 participants, some of whom have been tested six times, 5 years in between test occasions; some have been tested five times; participants in one sample have been tested three times; and the participants of three samples have been tested once (Nilsson et al., 1997, 2004, 2006). Participants have been selected randomly from the population registry of the city of Umeå in northern Sweden. Data for about 2000 variables each time a participant is tested are collected. These variables include an extensive cognitive test battery; a thorough health examination, including blood and saliva sampling; and a series of questionnaires on demographic variables, health history for the participant, parents and siblings to the participant, medication, leisure activities, exercise, and social networks of the participants. From the blood sampling, we extract DNA and have typed for several genes in a candidate gene approach, and for two large-scale genome-wide association studies. We are also doing functional magnetic brain imaging (MRI) and structural MRI on a subsample of about 367 participants (Nyberg et al., 2010).

To illustrate the differences in data patterns in cross-sectional and longitudinal analyses, the study by Rönnlund et al. (2005) will briefly be summarized.

Needless to say, the data from these two designs look quite different. An example of this on the basis of data from Rönnlund et al. (2005) is given in **Figures 1** and **2**. As can be seen from the cross-sectional data presented in **Figure 1**, the decrease in performance for episodic memory from the age of 35 years to the age of 85 years is approximately linear. This linear pattern has also been observed in several other cross-sectional studies (e.g., Schaie, 1996; Park et al., 2002), and in some cases, the decrease in performance has been observed to start as early as at 20 years of age.

Based on the same participants in the Betula Study, longitudinal data for the same episodic memory tasks are presented in **Figure 2**. As can be seen, the data pattern is radically different as compared to the cross-sectional data.

The longitudinal time frame of these data is 10 years measured at three points. In the longitudinal data, control has been made for attrition effects and test–retest effects by means of subtracting the gain in performance, when participants are tested a second or a third time. The size of this gain can be estimated by comparing the performance when participants are tested the first time and when they are tested for the second or the third time. For example, at T3, participants in S4 are tested the first time, whereas participants in S2 and S3 are tested the second time, and participants in S1 are tested for the third time. Thus, when doing this type of control for longitudinal data, the decline in episodic

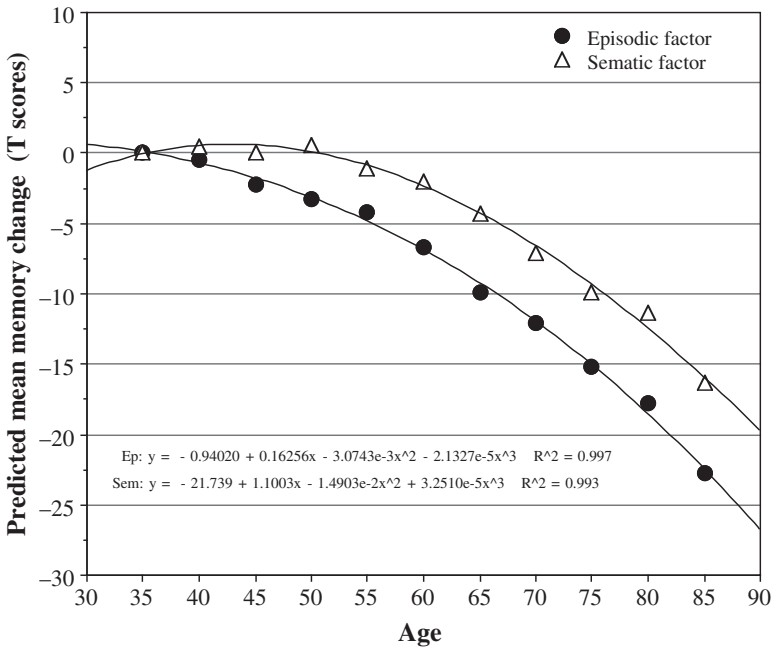

Figure 1 Estimated memory changes across age (T scores) for episodic and semantic memory on the basis of cross-sectional data. Reproduced with permission from Figure 1 in Rönnlund, M., Nyberg, L., Bäckman, L., Nilsson, L.-G., 2005. Stability, growth, and decline in adult life-span development of declarative memory: cross-sectional and longitudinal data from a population-based sample. Psychology and Aging 20, 3–18.

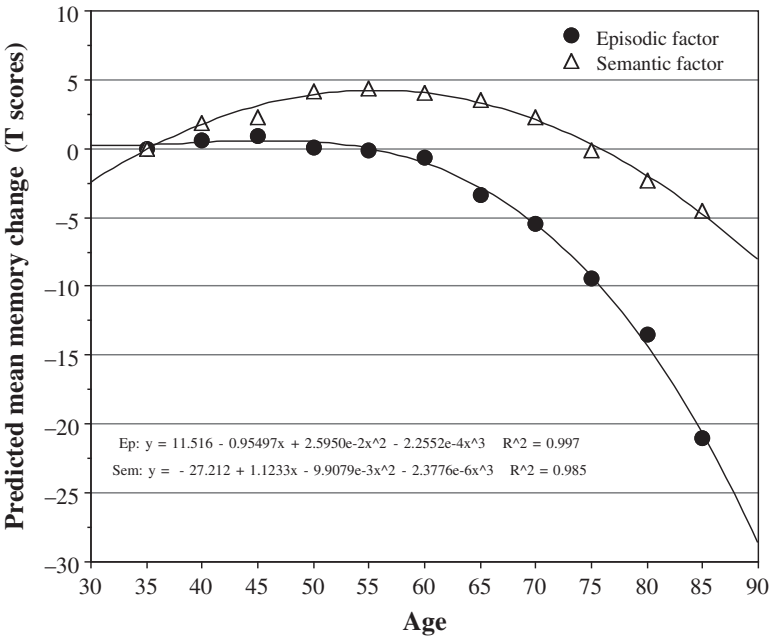

Figure 2 Estimated memory change across age (T scores) for the episodic and the semantic factors on the basis of practice-adjusted longitudinal data. Reproduced with permission from Figure 3 in Rönnlund, M., Nyberg, L., Bäckman, L., Nilsson, L.-G., 2005. Stability, growth, and decline in adult life-span development of declarative memory: cross-sectional and longitudinal data from a population-based sample. Psychology and Aging 20, 3–18.

memory performance is not observed until much later as compared to that in the cross-sectional data. Salthouse (2009) has argued that lack of power typically seen in longitudinal studies is a serious concern for such studies, but the time-related improvements seen in such studies when participants are tested a second or a third time are clearly difficult to reconcile with cross-sectional data (Rönnlund et al., 2005).

It should also be mentioned in this context that there are differences in brain-imaging data between cross-sectional and longitudinal analyses as in the behavioral data just described. It has been consensus in this field for quite some time that increased brain activity in the frontal lobes is associated with increasing age of the participants. The mechanism for this has been believed to be compensation. As people in such studies grow older, they start using the frontal lobes in both hemispheres, whereas young subjects use the frontal lobes in one hemisphere for the same task. In these earlier studies, a cross-sectional design has been used. In a recent study (Nyberg et al., 2010), we also found such a data pattern when we used a cross-sectional design. However, when we analyzed the data by means of a longitudinal design, we found that the activity in the frontal lobes decreased with increasing age. When analyzing the cross-sectional data closer, we found that the increase in frontal lobe activity was due to cohort effects. The increase in this activity only occurred in subjects with high education. However, in the longitudinal analyses, the frontal lobe activity even for these subjects decreased with increasing age.

In summary, the data presented here, both behavioral and functional brain-imaging data, show that cross-sectional results and longitudinal results are very different. When cohort differences, test–retest effects, and attrition are controlled for, longitudinal data give a more valid picture of the individual development of cognitive function in adulthood and old age than cross-sectional data, which generally are contaminated by cohort differences.

Large Variability in Cognitive Performance in the Aging Population

In most of the previous studies on aging and memory, different age groups have been treated as they are relatively homogeneous in the cognitive status within each age cohort. A mean for each age group and a standard deviation have been reported for each age cohort, usually based on cross-sectional data. Sometimes, it has been noted that the standard deviation is larger for the older groups, but most often this information has been left without comments.

In a recent paper, we defined three groups of participants (Josefsson et al., 2012) about this variability in cognitive performance among old people. They referred to those who were performing very well as successfully aged, those who were average, and those who were decliners. This is a new interesting concept in the area of cognitive aging and it is expected to expand.

In cross-sectional studies, the identification of individuals to different groups (e.g., decliners, average, maintainers)

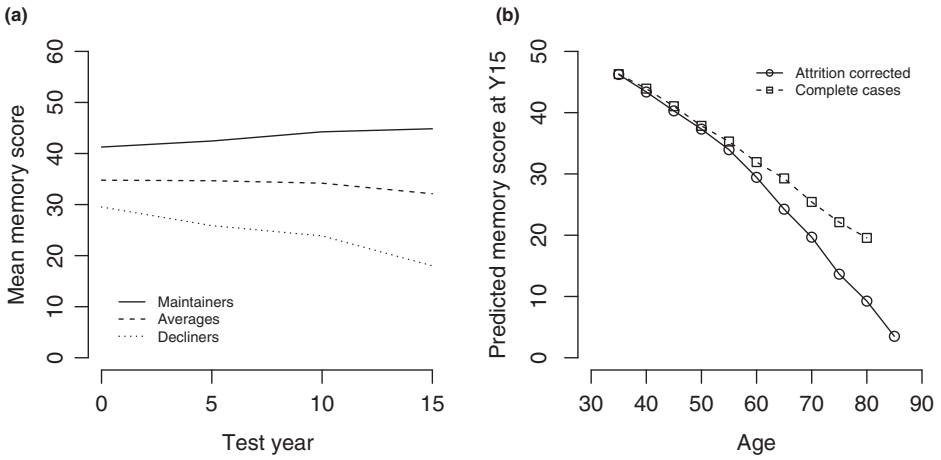

Figure 3 (a) Average test scores for 1558 participants: maintainers, average performers, and decliners. Missing values after dropout are predicted using individual baseline test score and average slope for the memory performance group they belong to. (b) Comparison of predicted episodic memory score at year 15 for complete cases, with and without attrition correction. Reproduced with permission from Figure 1 in Josefsson, M., de Luna, X., Pudas, S., Nilsson, L.-G., Nyberg, L., 2012. Genetic and lifestyle predictors of 15-year longitudinal change in episodic memory. Journal of American Geriatric Society (JAGS) 60, 2308–2312. doi:10.1111/jgs.1200.

misses the point of examining the extent to which the persons have preserved their performance in the same group or whether they have changed to another group, for example, whether a person in the declining group has decreased in performance from the average group. Longitudinal studies, on the other hand, offer a direct assessment of preserved or changed functioning (Josefsson et al., 2012). In this article by Josefsson et al., we divided the participants into three groups (decliners, average, maintainers). The data pattern for these three groups is presented in **Figure 3(a)**.

However, a difficulty of great concern in longitudinal studies is how to deal with attrition. In previous longitudinal studies, it has been assumed that attrition is random. There is evidence that attrition causes bias (Cooney et al., 1988; Habib et al., 2007; Yaffe et al., 2009) because those who remain in the study typically show better cognitive performance at baseline than those who drop out. When participants with a faster decline drop out earlier than those who maintain their performance and remain in the study, attrition is not random and cannot be ignored. The paper by Josefsson et al. (2012) has proposed a statistical model showing how to account for attrition in longitudinal research. **Figure 3(b)** depicts the difference between attrition-corrected data and complete cases using this method.

Final Remarks

Aging and memory is a very active area of research. Much progress has been made, although the findings from earlier studies on the effect of aging on separate memory systems and separate cognitive functions are still holding up. The new and interesting twist to this research area is the interest it has gained from neuroscientific approaches, including research on brain imaging and genetic trends. Much new news is to be seen in the field in the future integrating cognitive psychology, neuropsychology, brain imaging, and genetic analyses in studies we can see ahead.

See also: Aging Mind: Facets and Levels of Analysis; Aging and Health in Old Age; Autobiographical Memory; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Cognitive Aging; Dementia: Psychiatric Aspects; Language Processes in Delirium and Dementia; Semantic Memory; Short Term Memories, Theories of; Working Memory, Neural Basis of; Working Memory, Psychology of.

Bibliography

Bäckman, L., Nilsson, L.-G., 1996. Semantic memory functioning across the adult life span. European Psychologist 1, 27–33.

Bäckman, L., Small, B.J., Wahlin, Å., Larsson, M., 1999. Cognitive function in very old age. In: Craik, F.I.M., Salthouse, T.A. (Eds.), Handbook of Cognitive Aging, vol. 2. Erlbaum, Mahwah, NJ, pp. 499–558.

Bäckman, L., Small, B.J., Wahlin, Å., 2001. Aging and memory: cognitive and biological perspectives. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of Cognitive Aging, fifth ed. Academic Press, London, pp. 349–377.

Baddeley, A.D., 1986. Working Memory. Clarendon Press, Oxford, UK.

Baddeley, A.D., Hitch, G., 1974. Working memory. In: Bower, G.H. (Ed.), The Psychology of Learning and Motivation, vol. 8. Academic Press, New York, pp. 47–90.

Baltes, P.B., 1968. Longitudinal and cross-sectional sequences in the study of age and generation effects. Human Development 11, 145–171.

Brébion, G., Ehrlich, M.-F., Tardieu, H., 1995. Working memory in older subjects: dealing with ongoing and stored information in language comprehension. Psychological Research 58, 225–232.

Cabeza, R., Nyberg, L., 2000. Imaging cognition II: an empirical review of 275 PET and fMRI studies. Journal of Cognitive Neuroscience 12, 1–47.

Christensen, H., Korten, A., Jorm, A.F., Henderson, A.S., Scott, R., Mackinnon, A.J., 1996. Activity levels and cognitive functioning in an elderly community sample. Age and Ageing 25, 72–80.

Collaer, M., Hines, M., 1995. Human behavioral sex differences: a role for gonadal hormones during early development. Psychological Bulletin 118, 55–107.

Collins, A.M., Loftus, E.F., 1975. A spreading activation theory of semantic processing. Psychological Review 82, 407–428.

Collins, A.M., Quillian, M.R., 1969. Retrieval time for semantic memory. Journal of Verbal Learning and Verbal Behavior 8, 240–247.

Cooney, T.M., Schaie, W., Willis, S.L., 1988. The relationship between prior functioning on cognitive and personality dimensions and subject attrition in longitudinal research. Journal of Gerontology 43, 1217.

Corey-Bloom, J., Wiederholt, W.C., Edelstein, S., Salmon, D.P., Chan, D., Barett-Connor, E., 1996. Cognitive and functional status of the oldest old. Journal of the American Geriatric Society 44, 671–674.

Craik, F.I.M., 1977. Age differences in human memory. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging. Van Nostrand Reinhold, New York, pp. 384–420.

Craik, F.I.M., Jennings, J.M., 1992. Human memory. In: Craik, F.I.M., Salthouse, T.A. (Eds.), Handbook of Aging and Cognition. Erlbaum, Hillsdale, NJ, pp. 51–110.

Crook, T.H., West, R.L., 1990. Name recall performance across the adult life span. British Journal of Psychology 81, 335–349.

Del Missier, F., Mäntylä, T., Hansson, P., Bruine de Bruin, W., Parker, P., Nilsson, L.-G., 2013. The multifold relationship between memory and decision making: an individual-differences study. Journal of Experimental Psychology: Learning, Memory & Cognition 39, 1344–1364. http://dx.doi.org/10.1037/a0032379.

Eagly, A.H., Wood, W., 1999. The origins of sex differences: evolved dispositions versus social roles. American Psychologist 54, 408–423.

Elias, P.K., Elias, M.F., D'Agostino, R.B., Silberhatz, H., Wolf, P.A., 1999. Alcohol consumption and cognitive performance in the framingham heart study. American Journal of Epidemiology 150, 580–589.

Flynn, J.R., 1984. The mean IQ of Americans: massive gains from 1932 to 1978. Psychological Bulletin 95, 29–51.

Flynn, J.R., 1987. Massive IQ gains in 14 nations: what IQ tests really measure. Psychological Bulletin 101, 171–191.

Fountoulakis, K.N., Pavlidis, I., Tsolaki, M., 2011. Life events and dementia: what is the nature of their relationship? Psychiatry Research 30 (190), 156–158.

de Frias, C.M., Annerbrink, K., Westberg, L., Eriksson, E., Adolfsson, R., Nilsson, L.-G., 2005. Catechol O-Methyltransferase Val^{155}Met polymorphism is associated with cognitive performance in nondemented adults. Journal of Cognitive Neuroscience 17, 1018–1025.

de Frias, C.M., Marklund, P., Eriksson, E., Larsson, A., Öhman, L., Annerbrink, K., Bäckman, L., Nilsson, L.-G., Nyberg, L., 2010. Influence of COMT gene polymorphism on fMRI-assessed sustained and transient activity during a working-memory task. Journal of Cognitive Neuroscience 22, 1614–1622.

Gardiner, J.M., Java, R.I., 1993. Recognizing and remembering. In: Collins, A.F., gathercole, S.E., Conway, M.A., Morris, P.E. (Eds.), Theories of Memory. Erbaum, Hillsdale, NJ, pp. 163–188.

Gondo, Y., Yukie, M., Hiroki, I., Hirose, N., 2014. How do we measure cognitive function in the oldest old? A new framework for questionnaire assessment of dementia prevalence in centenarians. In: Nilsson, L.-G., Ohta, N. (Eds.), Dementia and Memory. Psychology Press.

Gregoire, J., Van der Linden, M., 1997. Effects of age on forward and backward digit span. Aging, Neuropsychology, and Cognition 4, 140–149.

Habib, R., Nyberg, L., Nilsson, L.-G., 2007. Cognitive and non-cognitive factors contributing to the longitudinal identification of successful older adults in the Betula study. Aging, Neuropsychology and Cognition 14, 257–273.

Herlitz, A., Bäckman, L., Nilsson, L.-G., 1997. Gender differences in episodic memory. Memory & Cognition 25, 801–811.

Herzog, C., Hultsch, D.F., Dixon, R.A., 1999. On the problem of detecting effects of lifestyle on cognitive changes in adulthood: reply to Pushkar et al. (1999). Psychology and Aging 14, 528–534.

Hill, R.D., 1989. The residual effects of cigarette smoking on cognitive performance in normal aging. Psychology and Aging 4, 251–254.

Hultsch, D.F., Hammer, M., Small, B.J., 1993. Age differences in cognitive performance in later life: relationships to self-reported health and activity lifestyle. Journal of Gerontology: Psychological Sciences 48, 1–11.

Hultsch, D.F., Herzog, C., Dixon, R.A., Small, B.J., 1998. Memory Change in the Aged. Cambridge University Press, Cambridge.

Jacoby, L.L., 1991. A process dissociation framework: separating automatic from intentional uses of memory. Journal of Memory and Language 30, 513–541.

Jennings, J.M., Jacoby, L.L., 1993. Automatic versus intentional uses of memory: aging, attention and control. Psychology and Aging 8, 283–293.

Johansson, L., Guo, X., Waern, M., Östling, S., Gustafson, D., Bengtsson, C., Skoog, I., 2010. Midlife psychological stress and risk of dementia: a 35-year longitudinal population study. Brain 133, 2217–2224.

Josefsson, M., de Luna, X., Pudas, S., Nilsson, L.-G., Nyberg, L., 2012. Genetic and lifestyle predictors of 15-year longitudinal change in episodic memory. Journal of American Geriatric Society (JAGS) 60, 2308–2312. http://dx.doi.org/10.1111/jgs.1200.

Laver, G.D., Burke, D.M., 1993. Why do semantic priming effects increase in old age: a meta-analysis. Psychology and Aging 8, 34–43.

Mäntylä, T., 1993. Knowing but not remembering: adult age differences in recollective experience. Memory & Cognition 21, 379–388.

Motomura, N., Ohkubo, F., Tomota, Y., Akagi, H., Asano, A., Seo, T., 1998. Premorbid psychosocial behavior in demented patients. The International Journal of Neuroscience 95, 167–172.

Nilsson, L.-G., 2014. Introduction. Is Alzheimer's disease predictable long in advance on the basis of memory performance. In: Nilsson, L.-G., Ohta, N. (Eds.), Dementia and Memory. Psychology Press.

Nilsson, L.-G., Bäckman, L., Nyberg, L., Erngrund, K., Adolfsson, R., Bucht, G., Karlsson, S., Widing, G., Wilblad, B., 1997. The Betula prospective cohort study: memory, health, and aging. Aging, Neuropsychology and Cognition 4, 1–32.

Nilsson, L.-G., Adolfsson, R., Bäckman, L., de Frias, C., Molander, B., Nyberg, L., 2004. Betula: a prospective cohort study on memory, health and aging. Aging, Neuropsychology and Cognition 11, 134–148.

Nilsson, L.-G., Adolfsson, R., Bäckman, L., Cruts, M., Nyberg, L., Small, B.J., Van Broeckhoven, C., 2006. The influence of *APOE* status on episodic and semantic memory: data from a population-based study. Neuropsychology 20, 645–657.

Nilsson, L.-G., Sternäng, O., Rönnlund, M., Nyberg, L., 2009. Challenging the notion of an early-onset of cognitive aging. Neurobiology of Aging 30, 521–524.

Norton, M.C., Østbye, T., Smith, K.R., Munger, R.G., Tschanz, J.T., 2009. Early parental death and late-life dementia risk: findings from the cache county study. Age and Ageing 38, 340–343.

Nyberg, L., Tulving, E., 1996. Classifying human long-term memory: evidence from converging dissociations. European Journal of Cognitive Psychology 8, 163–183.

Nyberg, L., Salami, A., Andersson, M., Eriksson, J., Kalpouzos, G., Kauppi, K., Lind, J., Pudas, S., Persson, J., Nilsson, L.-G., 2010. Longitudinal evidence for diminished frontal-cortex function in aging. PNAS 107, 22682–22686. http://dx.doi.org/10.1073/pnas.1012651108.

Park, D.C., Lautenschlager, G., Hedden, T., Davison, N., Smith, A.D., Smith, P.K., 2002. Models of visuospatial and verbal memory across the adult life span. Psychology and Aging 17, 299–320.

Persson, G., Skoog, I., 1996. A prospective population study of psychosocial risk factors for late onset dementia. International Journal of Geriatric Psychiatry 11, 15–22.

Pushkar, D., Etezadi, J., Andres, D., Arbuckle, T., Schwartzman, A.E., Chaikelson, J., 1999. Models of intelligence in late life: comments on Hultsch et al. (1999). Psychology and Aging 14, 520–527.

Quillian, M.R., 1966. Semantic memory (Ph.D. dissertation). Carnegie Institute of Technology.

Rönnlund, M., Nilsson, L.-G., 2006. Adult life-span patterns in WAIS block design performance: cross-sectional versus longitudinal age gradients and relations to demographic predictors. Intelligence 34, 63–78.

Rönnlund, M., Nilsson, L.-G., 2008. The magnitude, generality, and determinants of Flynn effects on forms of declarative memory and visuospatial ability: time-sequential analyses of data from a Swedish cohort study. Intelligence 36, 192–209.

Rönnlund, M., Nilsson, L.-G., 2009. Flynn effects on sub-factors of episodic and semantic memory: parallel gains over time and the same set of determining factors. Neuropsychologia 11, 2174–2188.

Rönnlund, M., Nyberg, L., Bäckman, L., Nilsson, L.-G., 2005. Stability, growth, and decline in adult life-span development of declarative memory: cross-sectional and longitudinal data from a population-based sample. Psychology and Aging 20, 3–18.

Rönnlund, M., Lövdén, M., Nilsson, L.-G., 2008. Cross-sectional versus longitudinal age gradients of Tower of hanoi performance: the role of practice effects and cohort differences in education. Aging, Neuropsychology and Cognition 15, 40–67.

Rowe, J.W., Kahn, R.L., 1987. Human aging: usual and successful. Science 237, 143–149.

Rypma, B., D'Esposito, M., 2000. Isolating the neural mechanisms of age-related changes in human working memory. Nature Neuroscience 3, 509–515.

Salthouse, T.A., 1996. The processing-speed theory of adult age differences in cognition. Psychological Review 103, 403–428.

Salthouse, T.A., 2009. When does age-related cognitive decline begin? Neurobiology of Aging 30, 507–514.

Schaie, K.W., 1996. Intellectual Development in Adulthood: The Seattle Longitudinal Study. Cambridge University Press, New York.

Schaie, K.W., 2009. "When does age-related cognitive decline begin?" Salthouse again reifies the "cross-sectional fallacy". Neurobiology of Aging 30, 528–529.
Sternäng, O., Wahlin, Å., Nilsson, L.-G., 2008. A population-based longitudinal analysis with Narrow Age Cohort design of the processing speed hypothesis. Scandinavian Journal of Psychology 49, 419–428.
Sundsröm, A., Rönnlund, M., Adolfsson, R., Nilsson, L.-G., 2014. Stressful life events are not associated with the development of dementia. International Journal of Psychgeriatrics 26, 147–154.
Titov, N., Knight, R.G., 1997. Adult age differences in controlled and automatic memory processing. Psychology and Aging 12, 565–573.
Tulving, E., 1983. Elements of Episodic Memory. Oxford University Press, New York.
Verhagen, P., Marcoen, A., Gossens, L., 1993. Facts and fiction about memory aging: a quantitative integration of research findings. Journal of Gerontology: Psychological Sciences 48, 157–171.
Yaffe, K., Fiocco, A.J., Lindquist, K., et al., 2009. Predictors of maintaining cognitive function in older adults—the Health ABC Study. Neurology 72, 2029–2035.
Zacks, R.T., Hasher, L., Li, K.Z.H., 1999. Human memory. In: Craik, F.I.M., Salthouse, T.A. (Eds.), Handbook of Cognitive Aging, vol. 2. Erlbaum, Mahwah, NJ, pp. 689–703.

Aging and Older People, Geography of

Gavin J Andrews, McMaster University, Hamilton, ON, Canada
Mark W Skinner, Trent University, Peterborough, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article summarizes the interests of and perspectives taken in the geographical study of aging. Dealing first with research engagements with space and then with place, the article concludes by highlighting some important and emerging areas of critical inquiry both in terms of theoretical and empirical interests.

Introduction

The geography of aging is an established field of academic scholarship focused on the spatial aspects of age and aging. Although broad and multidisciplinary in its composition – involving, for example, sociology, environmental psychology, demography, migration studies, urban planning and architecture, nursing, and other professionally focused research – its primary/core disciplines are human geography and social gerontology, hence the increasingly used title '*geographical gerontology*' (i.e., gerontology done geographically) and the lesser used alternative '*gerontological geography*' (i.e., geography done gerontologically). Indeed, these two disciplines have informed the field greatly in terms of its evolving empirical interests and theoretical and methodological innovations for over 40 years (see Andrews et al., 2007, 2009).

One potential reason for the continued and arguably increasing interest in the geography of aging in recent years has been a 'spatial turn' across both the health and social sciences (Andrews et al., 2013). Indeed, while the spatial turn in the social sciences has been theoretically primed by a desire to understand how humans are 'emplaced' in the world around them and the consequences of their presence (i.e., the interweaving of social, psychological, and physical life), the spatial turn in the health sciences has been associated with the increasing importance of the social model of health (i.e., an acknowledgment that health and well-being are strongly rooted in factors that lie outside the production and consumption of medicine, in communities and the broader environment). This said, however, of note is a limited yet significant critique emerging from within human geography itself on the whole idea of the geography of aging. This is based on dissatisfaction with a perceived disproportionate attention attributed in the parent discipline to people and populations at the extreme chronological ends of the human age-spectrum. Hence, as an alternative, certain scholars have promoted the 'relational geographies of age' which, rather than focusing on single demographic cohorts, would instead focus on intergenerationality and lifecourse to help illuminate, more fully, the meaning of age in relation to a greater number of people and situations (Hopkins and Pain, 2007). Although these arguments have gained traction among certain geographers, they have not been powerful or accepted enough to restrict the continued expansion and development of the geography of aging as an academic field; a majority of scholars have recognized that older people and populations have distinct characteristics, challenges, and circumstances that are distinctly geographical in nature and demand dedicated academic consideration (Andrews et al., 2007).

Space and Older People

At one level, drawing on a positivistic 'spatial science' philosophical tradition, space has been understood in the geography of aging as an underlying stage for human agency – an isotropic surface on top of which life plays out. While this understanding might portray an image of space itself as relatively abstract and neutral, when 'things' are located in space, such as older people, facilities and services for them or political boundaries that affect them it becomes more obvious how space might represent substantial features of, and challenges in life (Andrews et al., 2013). This is because space thereafter becomes mathematically distinguishable. *At* the points where things are located, rates, volumes, and other measures emerge and become calculable. *Between* these points, times, distances, movements, and differences do also. The assumption behind thinking about space in this way is that older age has fundamental, underlying, collective geometrical patterns that can be discovered and represented (Andrews et al., 2013).

Based on these understandings, research in the geography of aging has been able to trace how older people are positioned in and move across space, for example, in a substantial volume of demographic and migration research focused at subnational, national (Moore et al., 1997), and international (Kinsella and Velkoff, 2001) scales, and also between them. Notably, not only are older people themselves mapped across space, so are their circumstances and qualities, perhaps the most frequent being their life expectancy, health events, and status. This, for example, has been undertaken at global, continental (Warnes, 1999), and subnational (Pickle et al., 1996) levels, particularly in relation to social intersectionality (variables such as ethnicity, gender, affluence, and access to health and social care). Indeed, more generally within this tradition, scholars have identified the spatial patterning and characteristics of demographic transitions, including declining fertility (due to social change) and increasing life expectancy (due to improvements in diet, living conditions, and healthcare) that underpin demographic aging in particular areas. Notably, although much research here is published in academic

journals, its broad public and policy relevance means that it equally arises as 'gray' literature published widely by governments and NGOs. As Andrews et al. (2013) posit, often in relation to specific policy and administrational contexts, studies have also considered how resources for older people are distributed either by state allocation or by the private sector (Cartier, 2003), often in relation to patterns of utilization (see Nemet and Bailey, 2000). A well-known example of this is research on residential care at national and local scales (Andrews and Phillips, 2002; Ford and Smith, 2008). The motivating belief behind much of this type of research is that it is possible to enhance the spatial design of health and social care systems so that the needs of older people and populations are better met. A utilitarian motivation, which is somewhat idealistic and almost impossible to achieve practically, also emerges here: that where an older person resides should not disadvantage their health, welfare, and longevity – and that state systems should thus work toward the optimum spatial allocations of resources for them.

As Andrews et al. (2013) argue, it would be unwise to classify all of the aforementioned research as a purely 'spatially obsessed' science (a mathematical exercise in mapping for mapping's sake), because treatments of space in gerontology can be far more thoughtful. Indeed, acknowledging the unpredictability and complexity of human behavior, certain research articulates why spatial patterns are not always regular, and/or focuses on the consequences of patterns and/or attempts to critically understand the individual, social, economic, and political processes that drive them. These approaches can be seen, for example, in many of the aforementioned studies of population aging (Davies and James, 2011), the movement of older people (Warnes and Williams, 2006; Newbold, 2008), and services for older people (Rosenberg and Everitt, 2001). Moreover many studies have investigated important social contexts and the consequences of spatial movements; for example, in studies of family networks, the pushes and pulls to and from different locations in migration decision-making and in studies of the outcomes of migrations on individuals, populations, and health and social care systems. Research has considered these issues in the context of short-term and semi-permanent migrations between countries and with respect to short-term and semi-permanent migrations within countries. Notably, a very focused example of an approach that considers the consequences of space can be found in research that builds on studies of the proximity of older people to kin and carers and seeks to understand how the 'challenge of distance' impacts on the nature of personal interactions and relationships (Joseph and Hallman, 1998; Hallman and Joseph, 1999). Indeed, a number of studies here describe the factors that have led to the increasing physical distancing of older people from their adult children, such as social attitudes and economic circumstances. Whatever the reason for increased distancing, this research identifies a decay in immediate interactions and social and care support between families. Studies note though, that distance does not necessarily have to lead to a breakdown of communication between families, highlighting the concept and lived reality of 'intimacy at a distance'.

Also with regard to space, studies have highlighted the substantial challenges that lie ahead for governments and NGOs in meeting even the most basic needs of older populations in developing world countries, which are both growing and aging most rapidly. Indeed, research has highlighted the need for better policies and systems so that care and welfare services are funded and provided to minimum levels and standards. At its most critical, research in this area has also deconstructed globalization processes and the involvement of, and consequences for older individuals and populations (see Harper, 2006). There remains an undeniable 'first-world' bias in much geographical gerontology, a situation that does not reflect the circumstances and needs of the world's aging population (Andrews and Muzumdar, 2010).

Beyond this, other research has understood that space is more dimensional, complex, and problematic than the traditional image of it as a flat surface area. Recognizing instead that space is actively tied to human agency, capacity, and identity, scholars have developed the concept of 'social space' – meaning space as personally used and experienced by older people (Andrews et al., 2013). The idea of social space can be traced in early groundbreaking research by psychologists on environmental press and adaptation (whereby older people, facing physical and mental deterioration and being increasingly challenged by their environment, both reduce their use of space and adjust their relationships to it) (Lawton & Nahemow, 1973). Since these early studies, social space has been taken up far more broadly, now relating to broad-built, social, and other environmental contexts that create particular experiences, opportunities, and challenges for older people (Walsh and O'Shea, 2008; Skinner et al., 2009; Walsh and Gannon, 2011). More specifically, social space is also thought of as space within and between collections of settings that older people frequent and move between in daily, weekly, and monthly rhythms, such as their own homes, their friends' and family members' homes, consumer environments, and formal care environments (Wiles et al., 2009). Meanwhile, at a more intimate level, social space is also about the intricacies of what goes on between people and objects within specific settings, such as health or social care facilities or homes. At one level, the design and functionality of settings is a longstanding research consideration (Barnes, 2002; Gitlin, 2003). At another level, how different groups such as carers, family, and older people connect to, divide, and contest settings is an emerging focus of research (Dyck et al., 2005; Herron and Skinner, 2013). Otherwise, aligning social space with the idea of personal space (the very intimate spatial proximities between people) within settings is considered. These proximities are physical, yet they overlay and impact on interpersonal conversational, emotional, and moral proximities (see Andrews and Peter, 2006).

As described by Andrews et al. (2013), there has been a longstanding research interest in the use of technologies that potentially maintain, recapture, or even extend older peoples' mobility (and thus social space), including vehicles (Rosenbloom, 2001), personal assistive devices, and related interventions (Auger et al., 2010). Underpinning much of this research is a supposition that reduced social space is one factor that contributes to isolation, loneliness, and unhappiness in older age (Scharf and de Jong Gierveld, 2008). Most recently, studies have considered emerging hi/digital technologies and associated systems such as the internet – which compress time

and space as a social and physical barrier for some older people, – leading to what some have termed the 'death of distance' and the creation of replacement 'virtual social spaces' where physical copresence is not a prerequisite to experience and interaction. Of particular concern is how hi-technologies impact upon the form, experiences, and outcomes of older peoples' interpersonal communications and relationships (Milligan et al., 2011). Importantly, this research acknowledges that although some older people have access to, and experience substantial benefits from hi-technologies, others are excluded from them or experience substantial difficulties with their operation (Andrews et al., 2013).

Place and Older People

Places arise, in a very general sense, in much of the aforementioned space-focused research whether they be, for example, particular continents, nations, regions, towns, or settings. Moving beyond such surface engagements however, the facets, processes, and impacts of places have become important research interests in the geography of aging (see Andrews et al., 2013). At one level, a growing interest in place-effects on older peoples' health and welfare has emerged, often at the neighborhood and community level, exploring the impact of social 'composition' on health and welfare (the characteristics of older people in places including affluence, class, and family composition); the impact of social 'context' (the resources available to older people locally such as primary care facilities, various local shops, and services); and the impact of social collectivity (the ways in which people work together and benefit through mobilizing their social capital). At another level, associated with the more humanistic interpretation of space noted above, has been an emerging realization in research that places are much more than coordinates and locations. Indeed, the understanding has developed that peoples' lives unroll in places as complex social and cultural phenomena that are both acted out and felt. It is understood that at one level, because of people and things that occur *in situ*, places possess a basic agency and this agency can be helpful to older people or act as a barrier or challenge. At another level, however, it is understood that underlying this basic agency, far more personal processes are at work. Because of older peoples' experiences of places (the 'embedded knowledge' they gain from frequenting them), places acquire intentionality (i.e., they become about certain things) and essences (i.e., they feel a certain way) (Andrews et al., 2013). Thus, over time, they are ascribed meaning, and might, for example, become centers of attachment. In this way, through these processes places make and affect older people, but at the same time, older people play a part – with other people and things – in the making of places.

Wiles (2005) and later Andrews et al. (2013) summarize how, in line with these ideas, places have been conceptualized in gerontology in six ways. First, as simultaneously physical, social, and symbolic in character, such as a day care facility being subject to design, as social setting, and meaning certain things to society at large. Second, as a process and an essential part of social relationships (including family life and caring relationships). Third, as subject to continued renegotiation, such as when older people move between places or their needs change while in a place. Fourth, as expressing and facilitating power relationships, such as professional hierarchies in clinical settings or intergenerational relationships in home care. Fifth, as fought over and contested, such as between the different groups that occupy a setting where caring takes place. Sixth, as interrelated and interdependent such as a home where care takes place to the neighborhood in which it is located, the neighborhood to a town, and so on.

Andrews et al. (2013) note that across these six categories, the geography of aging has engaged with a wide range of empirical topics. Some issues tackled by recent studies include the role of places in supporting self-regard, self-determination, and self-identity in older age (Hockey et al., 2001; Wiles et al., 2009); the processes behind transnational lifestyles in aging and caring (McHugh and Mings, 1996; Walsh and O'Shea, 2010); the nature and challenges of rurality and aging (Walsh and O'Shea, 2008; Joseph and Skinner, 2012); and urbanicity and aging (Laws, 1993; Walsh and Gannon, 2011). Other well-trodden areas of research inquiry include the dynamics between place and experiences and practices of life and/or caring, particularly in homes and other community settings (Milligan, 2009; Skinner and Joseph, 2011). Given the geographical ideas explicit in them, research attention has also been paid to 'place-based policies' such as 'age-friendly cities' (Plouffe and Kalache, 2010) and, particularly, 'aging-in-place' (Cutchin, 2003; Gilleard et al., 2007; Wiles et al., 2012). Also with regard to place has been a recent engagement with nature and natural landscapes. Drawing on the concept of 'therapeutic landscapes' for example, studies have sought to examine the potentially positive healing and restorative qualities of particular places and activities, such as allotments and gardening (Milligan et al., 2004). Otherwise, with regard to nature, geographers have considered the challenges of weather for older people and their service providers (Skinner et al., 2009; Joseph et al., 2013). Most recently, in an allied development, scholars have attempted to articulate more thoroughly the importance of place to professional healthcare practices. Indeed, with respect to gerontological nursing for example, studies have attempted to identify the spatial aspects of practice that have the potential to affect the illness and recovery experiences of older people and their health outcomes (Andrews et al., 2005). Meanwhile, other studies have attempted to connect geographical concepts and ideas to some of the more progressive concepts and ideas that drive medicine and healthcare such as client-centeredness, intentionality, autonomy, and quality of care. Here, for example, a geographical perspective has helped to rethink practices associated with the physical and chemical restraint of older people with dementia. Collectively, this emerging research makes a reasonably direct contribution to conversations on evidence-based healthcare.

Emerging Critical Themes

An increasing proportion of the geography of aging, whether it be conceptually focused on space, place, or both, is 'critical' in its orientation. This critical orientation aligns very much to the recent critical/cultural traditions in the two parent disciplines of human geography and social gerontology, and involves

a number of core principles and approaches. The first of these is to challenge social and institutional norms and power relationships. Indeed, the idea is that researchers should not simply accept that the ideas, policies, and initiatives of institutions are appropriate simply because they are provided by those with power and authority. One should question them to expose unequal power relationships and their consequences. The second is to question and interpret dominant models of thinking; the concepts and ideas that underpin policy and practice. Each will have relative merits, success or failures, and specific consequences. The third is to give voice to subjects and issues that are neglected or marginalized in mainstream health and social care practice. Certain people, conditions, and situations – often involving the most vulnerable people and groups – 'fall off the map,' and it is up to researchers to expose and challenge neglect, vulnerability, inequality, and oppression. The fourth is to use qualitative or mixed methodologies to gain detailed and in-depth insights. Human circumstances surrounding lives and care are inherently complex and a range of qualitative research methodologies and associated theoretical and analytical approaches are well suited to investigate, interpret, and convey these. The fifth is to draw on philosophy and social theory to theoretically inform and enhance research. This helps to make sense of empirical observations and articulate some of the underlying meanings and processes involved (see Harper and Laws, 1995). The sixth is to involve communities as partners in research, sometimes taking an activist or advocacy or 'public' approach. Scholars arguably need not neglect this important aspect of their work and might work with various 'publics' to directly encourage reform and change (Andrews and Muzumdar, 2010).

One can observe development and progress in critical research, particularly in a number of theoretical and empirical areas which might be further developed in future. The first of these areas is emotion. Studies in this vein have illustrated not only how public, domestic, and social spaces act to facilitate or constrain the expression of a range of emotions in later life, but also the importance of older people's emotional attachment to place, and the often negative impact of transitions *between* places (Milligan, 2005). The second area is the body. Moving away from treating older people as 'objects' of study, several studies engage with aging and the concept of embodiment, to explore their perceptions of health and aging and their negotiations and changing relationships with spaces and places (Mowl et al., 2005). Specifically studies have reflected on body image and ageist discourses and cultural constructions of aging. The third area is image and stereotypes. Notably, here, research has sought to challenge stereotypes that connect older people solely to 'sick' and 'sad' places (see McHugh, 2003) and examine how and why certain places are, instead, associated with independence, resilience, and well-being in older age. These include, for example, 'unremarkable' places routinely passed through in daily life and specialist places associated with active lifestyles and emerging consumer cultures in aging (Laws, 1995; Mansvelt, 1997; McHugh and Larson-Keagy, 2005). An associated strand of recent research has begun to deconstruct the 'imagined landscapes' of aging identity. In these studies, geographical metaphors are explored, including life's late journeys, paths traveled, autumn years, and the sunsets of lives. The fourth related area is positivity. Commentators have also considered the more positive places and activities of healthy aging – the new landscapes of later life. Examples here include retirement communities, RVing, and Universities of the Third Age. Indeed studies have identified the commodification of healthy aging and as part of this, the radical fracturing of traditional spaces of aging and the common understanding of them. In addition scholars have tended to engage positive aging through a policy critique of issues, such as pensions, lifelong learning, education, physical and social activity, and active citizenship (see Minkler and Holstein, 2008). The fifth and final area is community development. Uncovering the complex, interdependent, and place-embedded local dynamics of older people's experiences of aging and the development trajectories of their communities is seen as crucial for understanding how societies are coping with, and responding to the challenges and opportunities of population aging. Within this theme, the importance of voluntarism is noted as a process through which older people are actively being supported and are actively participating/leading community responses (Joseph and Skinner, 2012; Skinner et al., 2013).

Emerging Critical Theory

While, as suggested, the adoption of critical social theory has emerged in the geography of aging under a post-structuralist paradigm – including engagements with thinkers such as Foucault, Goffman, Deluze, and others – arguably the next step is to come up to speed in other ways with the parent discipline of human geography in understanding and approaching space and place. With this in mind, Andrews et al. (2013) flag relational geography and nonrepresentational theory as two areas worthy of further consideration. Associated with the adoption of post-structuralist theories and perspectives in twenty-first-century human geography has been a 'relational turn.' Basically put, thinking space and place 'relationally' involves rather than envisioning a space as a discrete, bounded area (such as a mapable administrative unit) and a place as a fixed center of meaning (such as a lone house, town, neighborhood, or city), an image of spaces and places as produced through their connections with other spaces and places (Andrews et al., 2013). These ideas have potentially far-reaching implications for the geography of aging. Although the field has long been orientated around the relationship between geographical scales (such as individual, population health, and welfare), new relational ways of thinking potentially extend, and more explicitly theorize, inter-scale relationships in later life. Moreover, the new spatial imaginary of relationality potentially challenges the way scholars approach the fundamental question in the geography of aging: of how and why 'space and place matter to older people'? Namely, it complicates the conventional assumption that there exist 'intrinsic' qualities of spaces (for example, with regard to distance and proximity) and single places (for example, with regard to facets/content and meaning) that affect them (Andrews et al., 2013).

The paradigm of nonrepresentational theory, on the other hand, acknowledges that a sizable portion of the world – what happens 'out there' in everyday life – has been drained from,

and remains unrepresented by much conventional qualitative geographical research, due to traditional philosophical commitments involving theoretically driven interpretative searches for meaning and significance. In contrast, nonrepresentational understandings do not posit an external, material world waiting to be represented and explained away by a researcher. The idea, instead, is that they obtain an understanding of the lived world by engaging with it and understanding it as a performative achievement. Thus, nonrepresentational theory moves the focus of research onto the many subtle and unspoken performances and practices involved in the production and reproduction of life. As Thrift (2007) explains, it conveys the geography of what 'takes place' – the 'bare bones' of occasions and the flow of life. Moreover, because nonrepresentational theory does not seek to explain the human subject, it is, instead, interested in the human body's coevolution with things and relatedness to the world (Thrift, 2007). This involves an attention to the expressiveness, responsiveness, and rhythms of human practice (including performativity, embodiment, senses, and expressions) (Cadman, 2009). As the above explanation indicates, nonrepresentational theory is not really a single theory in itself; rather it is a general way of understanding the active world and undertaking research on it. It is the next stage for scholars to work out what this new paradigm of scholarship means for the geography of aging, and how it fits within, and contributes to the wider discipline of social gerontology (Andrews et al., 2013).

See also: Age, Sociology of; Age: Anthropological Aspects; Aging and Health in Old Age; Aging and Work; Aging and the Labor Market; Aging, Theories of; Alzheimer's Disease: Behavioral and Social Aspects; Childhood and Adolescence: Developmental Assets; Cultural Geography; Differential Aging; Elder Care; Emotions and Aging; Health, Geography of; Medical Geography; Migration, Ethnicity, Aging and Social Work Practice; Old Age and Centenarians: The Human 'Warranty Period'; Old Age: Definitions, Theory, and History of the Concept; Social Geography.

Bibliography

Andrews, G.J., Cutchin, M., McCracken, K., Phillips, D.R., Wiles, J., 2007. Geographical gerontology: the constitution of a discipline. Social Science and Medicine 65 (1), 151–168.

Andrews, G.J., Milligan, C., Phillips, D.R., Skinner, M., 2009. Geographical gerontology: mapping a disciplinary intersection. Geography Compass 3 (5), 1641–1659.

Andrews, G.J., Evans, J., Wiles, J., 2013. Re-spacing and re-placing gerontology: Relationality and affect. Ageing and Society 33 (8), 1339–1373.

Andrews, G.J., Muzumdar, T., 2010. Rethinking the applied: public gerontology, global responsibility. Journal of Applied Gerontology 29 (2), 143–154.

Andrews, G.J., Phillips, D.R., 2002. Changing local geographies of private residential care for older people 1983–1999: lessons for social policy in England and Wales. Social Science and Medicine 55 (1), 63–78.

Andrews, G.J., Holmes, D., Poland, B., Leheux, P., Miller, K.-L., Pringle, D., McGilton, K., 2005. "Airplanes are flying nursing homes": geographies in the concepts and locales of gerontological nursing. International Journal of Older People Nursing 14 (S2), 109–120.

Andrews, G.J., Peter, E., 2006. Moral geographies of restraint in nursing homes. Worldviews on Evidence-Based Nursing 3 (1), 2–7.

Auger, C., Demers, L., Gélinas, I., Miller, W.C., Jutai, J.W., Noreau, L., 2010. Life-space mobility of middle-aged and older adults at various stages of usage of power mobility devices. Archives of physical medicine and rehabilitation 91 (5), 765–773.

Barnes, S., 2002. The design of caring environments and the quality of life of older people. Ageing and Society 22 (6), 775–789.

Cadman, L., 2009. In: Kitchen, Thrift (Eds.), Nonrepresentational Theory/Nonrepresentational Geographies. International Encyclopaedia of Human Geography Elsevier.

Cartier, C., 2003. From home to hospital and back again: economic restructuring, end of life, and the gendered problems of place-switching health services. Social Science and Medicine 56 (11), 2289–2301.

Cutchin, M., 2003. The process of mediated aging-in-place: a theoretically and empirically based model. Social Science & Medicine 57 (6), 1077–1090.

Davies, A., James, A., 2011. Geographies of Ageing: Social Processes and the Spatial Unevenness of Population Ageing. Ashgate, Farnham.

Dyck, I., Kontos, P., Angus, J., McKeever, P., Poland, B., 2005. The home as a site of long-term care: meanings and management of bodies and spaces. Health and Place 11 (2), 173–185.

Ford, R.G., Smith, G.C., 2008. Geographical and structural change in nursing care provision for older people in England 1993–2001. Geoforum 39 (1), 489–498.

Gilleard, C., Hyde, M., Higgs, P., 2007. The impact of age, place, aging in place, and attachment to place on the well-being of the over 50s in England. Research on Aging 29 (6), 590–605.

Gitlin, L., 2003. Conducting research on home environments: lessons learned and new directions. The Gerontologist 43 (5), 628–637.

Hallman, B.C., Joseph, A.E., 1999. Getting there: mapping the gendered geography of caregiving to elderly relatives. Canadian Journal on Aging 18 (4), 397–414.

Harper, S., 2006. Addressing the implications of global ageing. Journal of Population Research 23 (2), 205–223.

Harper, S., Laws, G., 1995. Rethinking the geography of ageing. Progress in Human Geography 19 (2), 199–221.

Herron, R.V., Skinner, M.W., 2013. The emotional overlay: Older people and carer perspectives on negotiating aging and care in rural Ontario. Social Science & Medicine 91, 186–193.

Hockey, J., Penhale, B., Sibley, D., 2001. Landscapes of loss: spaces of memory, times of bereavement. Ageing and Society 21 (6), 739–757.

Hopkins, P., Pain, R., 2007. Geographies of age: thinking relationally. Area 39 (3), 287–294.

Joseph, A.E., Hallman, B.C., 1998. Over the hill and far away: distance as a barrier to the provision of assistance to elderly relatives. Social Science and Medicine 46, 631–639.

Joseph, A.E., Skinner, M.W., 2012. Voluntarism as a mediator of the experience of growing old in evolving rural spaces and changing rural places. Journal of Rural Studies 28, 380–388.

Joseph, G., Skinner, M.W., Yantzi, N.M., 2013. The weather-stains of care: interpreting the meaning of bad weather for front-line health care workers in rural long-term care. Social Science & Medicine 91, 194–200.

Kinsella, K., Velkoff, V.A., 2001. An Aging World: 2001. US Census Bureau, Series P95/01-1. US Government Printing Office, Washington, DC.

Laws, G., 1993. "The land of old age": society's changing attitudes toward urban built environments for elderly people. Annals – Association of American Geographers 83 (4), 672–693.

Laws, G., 1995. Embodiment and emplacement: identities, representation and landscape in Sun City retirement communities. International Journal of Aging and Human Development 40, 253–580.

Lawton, M.P., Nahemow, L., 1973. Ecology and the aging process. In: Eisdorfer, C., Lawton, M.P. (Eds.), Psychology of Adult Development and Aging. American Psychological Association, Washington, DC, pp. 657–668.

Mansvelt, J., 1997. Working at Leisure: critical geograhies of ageing. Area 29 (4), 289–298.

McHugh, K.E., 2003. Three faces of ageism: society, image and place. Ageing and Society 23 (2), 165–185.

McHugh, K.E., Mings, R.C., 1996. The circle of migration: attachment to place in aging. Annals, Association of American Geographers 86 (3), 530–550.

McHugh, K.E., Larson-Keagy, E.M., 2005. These white walls: the dialectic of retirement communities. Journal of Aging Studies 19 (2), 241–256.

Milligan, C., Gatrell, A., Bingley, A., 2004. 'Cultivating health': therapeutic landscapes and older people in northern England. Social Science & Medicine 58 (9), 1781–1793.

Milligan, C., 2009. There's No Place like Home: People, Place and Care in an Ageing Society. Ashgate Aldershot.

Milligan, C., 2005. 'From home to 'home': situating emotions within the care-giving experience. Environment and Planning A 37 (12), 2105–2120.
Milligan, C., Roberts, C., Mort, M., 2011. Telecare and older people: who cares where? Social Science & Medicine 72 (3), 347–354.
Minkler, M., Holstein, M.B., 2008. From civil rights to… civic engagement? Concerns of two older critical gerontologists about a "new social movement" and what it portends. Journal of Aging Studies 22 (2), 196–204.
Moore, E.G., Rosenberg, M.W., McGuiness, D., 1997. Growing Old in Canada: Demographic and Geographic Perspectives. Statistics Canada and ITP Nelson, Ottawa and Toronto.
Mowl, G., Pain, R., Talbot, C., 2005. The ageing body and the homespace. Area 32 (2), 189–197.
Nemet, G.F., Bailey, A.J., 2000. Distance and health care utilization among the rural elderly. Social Science and Medicine 50 (9), 1197–1208.
Newbold, K.B., 2008. Interprovincial migration and retirement income transfers among Canada's older population: 1996–2001. Environment and Planning A 40, 1501–1516.
Pickle, L.W., Mungiole, M., Jones, G.K., White, A.A., 1996. Atlas of United States Mortality, Centers for Disease Control and Prevention. National Center for Health Statistics, Hyattsville, Maryland.
Plouffe, L., Kalache, A., 2010. Towards global age-friendly cities: determining urban features that promote active aging. Journal of Urban Health 87 (5), 733–739.
Rosenberg, M., Everitt, J., 2001. Planning for aging populations: inside or outside the walls. Progress in Planning 56 (3), 119–168.
Rosenbloom, S., 2001. Driving cessation among older people: when does it happen and what impact does it have? Transportation Research Record 1779 (13), 93–99.
Scharf, T., de Jong Gierveld, J., 2008. Loneliness in urban neighbourhoods: an Anglo-Dutch comparison. European Journal of Ageing 5 (2), 103–115.
Skinner, M.W., Yantzi, N.M., Rosenberg, M.W., 2009. Neither rain nor hail nor sleet nor snow: provider perspectives on the challenges of weather for home and community care. Social Science & Medicine 68 (4), 682–688.
Skinner, M.W., Joseph, A.E., 2011. Placing voluntarism within evolving spaces of care in ageing rural communities. GeoJournal 76, 151–162.
Skinner, M.W., Joseph, A.E., Herron, R.V., 2013. Spaces of resistance or acquiescence? Learning from media discourses on the evolving role of voluntarism in ageing communities. Environment and Planning A 45, 438–450.
Thrift, N., 2007. Non-representational Theory: Space, Politics, Affect. Routledge, London.
Walsh, K., Gannon, B., 2011. Perceived neighbourhood context, disability onset and old age. The Journal of Socio-Economics 40 (5), 631–636.
Walsh, K., O'Shea, E., 2008. Responding to rural social care needs: older people empowering themselves, others and their community. Health and Place 14 (4), 795–803.
Walsh, K., O'Shea, E., 2010. Marginalised care: migrant workers caring for older people in Ireland. Journal of Population Ageing 3 (1–2), 17–37.
Warnes, A.M., 1999. UK and western European late-age mortality: trends in cause-specific death rates, 1960–1990. Health and Place 5 (1), 111–118.
Warnes, A.M., Williams, A., 2006. Older migrants in Europe: a new focus for migration studies. Journal of Ethnic and Migration Studies 32 (8), 1257–1281.
Wiles, J.L., 2005. Conceptualising place in the care of older people: the contributions of geographical gerontology. International Journal of Older People Nursing 14 (s2), 121–129.
Wiles, J.L., Allen, R., Palmer, A.J., Hayman, K.J., Keeling, S., Kerse, N., 2009. Older people and their social spaces: a study of well-being and attachment to place in Aotearoa New Zealand. Social Science and Medicine 68 (4), 664–671.
Wiles, J.L., Leibing, A., Guberman, N., Reeve, J., Allen, R.E.S., 2012. The meaning of 'ageing in place' to older people. The Gerontologist 52 (3), 357–366.

Aging and the Labor Market

Alan Barrett, Economic and Social Research Institute, Dublin, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Population aging will impact on the labor market in at least two ways. First, the number of people in the standard working ages (20–64) will fall. This is generally accepted as being a cause for concern because, all else equal, a smaller pool of workers will have to support a larger number of dependents. Second, the proportion of older workers will rise. This is often thought to be a concern based on the assumption that older people are less productive. The evidence for this turns out to be weak. The relative importance of the two effects suggests that policy should focus on extending work lives.

Introduction

The phenomenon of population aging has been identified as one of the great challenges facing the modern world. The joint occurrence of reduced mortality and lower fertility was initially a feature of the developed world, but increasingly there is evidence that this demographic transition is occurring in less developed countries also. The result is that the age structure of many populations is changing, with an increase in the size of the older population relative to those of younger ages.

Before discussing aspects of the challenges associated with population aging, it is important to say that reduced mortality and hence increased longevity are among the great achievements of humanity. The improvements in mortality in the very recent past (relative to the vast span of human history) have been extraordinary and are now leading us to question concepts such as retirement ages and the nature of retirement itself. While accepting these positives, we also have to accept that the large changes in the age structure of the population do present challenges and will continue to do so as the process unfolds in many countries. The challenges which come to mind most readily and which have received the most attention relate to the costs associated with older age groups and the increase in the ratio of older people to those of working age (the old-age dependency ratio). The costs typically in question are pensions, healthcare, and long-term care. With a relatively smaller number of people of working age, the tax and contribution base falls relative to the increased need. All else equal, these processes combine to put upward pressure on tax or contribution rates and/or downward pressure on benefit and care levels.

It is frequently argued that one of the possible routes through which the challenges of population aging can be solved, or at least eased, is though higher rates of employment including higher rates of participation among older age groups. It is also argued that productivity increases can contribute to meeting the aging challenge, with productivity increases among older workers again forming part of the overall strategy. Such productivity increases could come about through an increase in the capital/labor ratio but also through increases in human capital.

Given the link between the labor market and some proposed solutions to the challenges presented by population aging, it is important to understand how population aging and labor market processes interact. In this article, I will discuss these interactions by drawing on the economic literature on various aspects of the topic. I will begin by presenting some projections on the size of the workforce in a range of countries in the coming years, in terms of the number of people in the standard working age group, 20–64. Benchmark projections are typically based on existing patterns of participation and so can lead to stark findings on likely trends. In particular, labor forces in many countries are projected to shrink. This in itself is not a problem if the numbers outside of the labor force were also falling. However, with the changing population structure and an increasing proportion of older people, the old-age dependency ratio worsens and it is this which gives rise to problems of affordability and sustainability for public services, including pensions.

The labor force projections can then be used to assess what increases in participation would be needed to maintain the labor force at certain levels. In this way, we are provided with a sense of the challenge which policies aimed at increasing participation must address. In this context, the next issue that I will consider is retirement and the factors which determine it. If labor force participation rates among older people are to be increased, it would be desirable in the first instance to reduce early exits from the labor force. To the extent that 65 represents some notion of the 'normal' or 'standard' retirement age, reducing early exits translates into keeping people in the labor force until that age. However, the labor force projections suggest that more will be needed and so extending worker lives beyond 65 would be desirable too. In this situation, the need arises to encourage but also to facilitate active working by people as they move in to the late sixties and beyond. Considerations which arise in this context include the need for flexibility such as part-time work or the option of reduced levels of responsibility.

I noted earlier that in addition to increasing the size of the labor force, it would also be desirable to increase productivity, including that of older workers. For this reason, I will discuss what the literature says on productivity in the later working years. It is worth noting at this point that there often seems to be an assumption within economics and in other subjects that productivity falls as people age and that this inevitably means that productivity will fall in an aging population. As will be seen, the evidence does not support this view and so thought can be devoted to increasing the

productivity of older workers as opposed to merely ensuring that it does not fall too far. The issues which arise here include life-long learning and efforts to avoid skill obsolescence and also the re-organization of work to facilitate older workers. In discussing this area, one issue that arises is the possibility that employers and employees will react to the aging of the workforce without any prompting from government through policy levers. For example, if employers see that the pool of potential employees is aging, they may engage in proactive efforts to make their workplaces more attractive to older workers.

The factors which will lead people to remain in work beyond the normal retirement age and the factors which will increase their productivity are not mutually exclusive. For this reason, the discussion across the topics in the sections later will be interconnected. In the concluding section, I will bring the lessons together and discuss the policy implications. Referring back to the paragraph earlier, the importance of putting in place policies which will make a positive contribution to meeting the challenges posed by aging cannot be understated. In the absence of success in extending working lives and increasing productivity, including among older workers, the challenges posed by population aging will be more difficult to address and so the full benefits of increased longevity will not be realized.

Labor Force Projections

European Commission (2012) provides detailed forecasts of the populations and labor forces of all EU members out to 2060 as part of a larger exercise to examine the implications of population aging for the public finances. Here, I will focus on just two outputs from that report because they capture neatly the issues under discussion. (Projections of this sort are highly uncertain as they require assumptions on a range of trends such as mortality, fertility, and labor force participation rates. Assumptions for migration flows are also needed and these can be the most difficult of all.)

In **Table 1**, I re-produce one of the tables from the European Commission (2012). In this table, the numbers of people aged 20–64 in employment in 2010 are listed for all countries. Based on the projections of population and on assumptions for variables such as rates of participation and unemployment, it is possible to generate projections for employment for all countries out to 2060, for this specific age group. Total employment is, and will be, higher because people under the age of 20 and over the age of 64 also work. However, by focusing in on the 20–64 year olds, we can get a neater picture of what will happen to the size of the group of standard working age. In addition to the value for 2010, the table also shows when the employment of those aged 20–64 will reach its peak and its trough.

Table 1 Projected peaks and troughs in the numbers employed aged 20–64

	Employment 20–64 (in millions)						
	2010 value	*Peak value*	*Year*	*% Change 2010 peak*	*Trough value*	*Year*	*% Change peak trough*
BE	4.4	4.9	2060	11.7	4.4	2010	−10.5
BG	3.1	3.1	2012	1.1	1.9	2060	−38.8
CZ	4.8	4.8	2012	1	4	2060	−17.8
DK	2.5	2.6	2025	3.7	2.5	2010	−3.5
DE	37.2	37.9	2012	1.9	26	2060	−31.3
EE	0.6	0.6	2012	7	0.4	2060	−24.4
IE	1.8	2.4	2060	37.1	1.7	2015	−28.4
EL	4.5	4.7	2024	5.8	4.2	2060	−12
ES	18.2	22.4	2033	22.7	18.2	2010	−18.5
FR	26.4	28.6	2060	8.5	26.4	2010	−7.8
IT	22.5	24.5	2024	9	22.3	2060	−9.2
CY	0.4	0.5	2044	29.6	0.4	2010	−22.8
LV	0.9	1	2012	5.1	0.6	2060	−35.9
LT	1.3	1.4	2012	6.5	1	2060	−29.9
LU	0.2	0.3	2060	22.6	0.2	2010	−18.5
HU	3.8	4	2027	4.5	3	2060	−23.3
MT	0.2	0.2	2033	5.2	0.1	2060	−16
NL	7.8	7.9	2015	2	7	2060	−11.5
AT	3.9	4	2018	3	3.6	2060	−9.3
PL	16	16.3	2014	1.5	10.8	2060	−33.9
PT	4.6	4.8	2028	4	4	2060	−16
RO	8.7	8.8	2012	0.6	5.2	2060	−40.9
SI	0.9	0.9	2020	0.7	0.8	2060	−17
SK	2.3	2.3	2012	1.5	1.7	2060	−26.4
FI	2.4	2.4	2016	1.3	2.3	2060	−5.2
SE	4.3	4.9	2050	14.4	4.3	2010	−12.6
UK	27.3	31.9	2060	16.7	27.3	2010	−14.3
NO	2.3	2.7	2060	18.2	2.3	2010	−15.4
EU27	210.9	217.6	2022	3.2	195.6	2060	−10.1
EA	138.1	143.9	2024	4.2	129.8	2060	−9.8

If we look toward the end of the table at the row EU27, we can get an overall sense of the impact of population aging on the numbers in employment of those aged 20–64. In 2010, just under 211 million of this group were employed across the EU. According to the projections, this number will increase to 217.6 million in 2022 but this is the peak level. This number is then projected to fall continuously through to 2060 when it will reach 195.6 million. This will amount to a peak-to-trough fall of 10.1%. All else equal, this would imply a significant fall in potential output. Hence, the need for increases in employment from other sources, such as through delayed retirement, is clear.

Looking at the countries within the table, there are a number of interesting points to observe. Germany is among the countries with the largest projected declines. Employment levels for the standard working age group were projected to peak in 2012 at 37.2 million. This level was projected to fall to 26 million in 2060, a fall of 31.3%. The country with the largest projected peak-to-trough decline is Romania (−40.9%). Some countries, such as France and the UK, see employment levels among the 20–64 year olds as being at their lowest in 2010 and so the situation facing them appears to be less stark, at least based on the assumptions used in the projections. In the case of the UK, they expect the population aged 20–64 to peak in 2060. In the case of France, the population aged 20–64 is projected to reach its trough in 2038, and not 2060, thereby providing scope for employment increases among this group between 2038 and 2060. In spite of the projections from the UK, France, and some other countries, the broad pattern to emerge from Table 1 is of a shrinking level of employment among those aged 20–64.

In Table 2, I present a different view of the impact of aging on the labor market, again drawing on European Commission (2012). The table shows the employment of those aged 55–64 as a share of total employment among those aged 15–64 in 2010 and projected values for 2060. Looking toward the lower part of the table, it can be seen that for the EU27 in 2010, 13.2% of employment among those aged 15–64 was made up of people aged 55–64. In 2060, this proportion is projected to rise to 18.8%, an increase of 5.5% points. Italy is projected to experience the largest increase in this share – from 12% in 2010 to 23.7% in 2060.

Table 2 Share of older (55–64) employees in total employment, 2010 and 2060 (projected)

	Ch 10–60	*2010*	*2060*
BE	2.7	11.2	13.9
BG	2	15	16.9
CZ	4.5	14.5	19
DK	3.1	15.5	18.6
DE	6.3	14.9	21.3
EE	1.7	15.7	17.4
IE	4.7	12.6	17.3
EL	7.7	13	20.7
ES	9.3	12	21.3
FR	4.9	12.1	17
IT	11.6	12	23.7
CY	4	13.5	17.5
LV	3.6	13.5	17.1
LT	4.9	13	17.8
LU	3	9.7	12.7
HU	7.8	12.1	20
MT	6.9	11.2	18
NL	2.2	14	16.2
AT	5.7	10	15.7
PL	4.1	10.7	14.8
PT	7.8	13.5	21.3
RO	5.8	12	17.8
SI	6.7	9.9	16.6
SK	4.4	11.8	16.2
FI	−1.4	18.4	17
SE	−1.3	18.9	17.6
UK	2.1	14.7	16.8
NO	0.2	16.6	16.9
EU27	5.5	13.2	18.8
EA	6.6	13.1	19.7

Retirement and Extending Working Lives

The decreasing number of those aged 20–64 suggests that participation among those aged 65 and above will have to increase if the numbers employed is to be maintained. In order for this to happen, we need to have a good understanding of why people retire. In this section, I discuss what the literature on this point has said.

Before getting into the literature, it is useful to consider the trend in retirement ages in recent decades. In Table 3, I show data from the OECD on the average effective age at which older male workers withdraw from the labor force – labeled the effective age of retirement. Looking first at the data on the US, it can be seen that in 1970, the age was 68.5 years. By the turn of the century, the age had fallen to 64.7. This was a large change over a 30-year period and was certainly a cause for concern in the context of population aging. However, the effective age of retirement has moved up again in the first decade of this century and was 65.5 in 2010.

Turning to the European Union, the effective age of retirement was 68.4 years in 1970 and so essentially the same as in the US. By 2000, this had fallen to 61.5 years and so the fall over the 30-year period 1970 to 2000 was stronger in the EU countries compared to the US. Like the US, the EU saw an increase in the effective age of retirement between 2000 and 2010 – from 61.5 to 62.5 years. While both the US and the EU experienced an increase in the effective age of retirement between 2000 and 2010, the gap between the two in 2010 is striking – 65.5 in the US and 62.5 in the EU.

Japan provides another interesting comparison. Its effective retirement age was substantially higher than those of both the

Table 3 Average effective age of retirement: men

	1970	*1980*	*1990*	*2000*	*2010*
France	67.6	63.5	60.0	58.8	59.4
United Kingdom	67.7	66.0	62.8	62.4	64.1
Japan	72.3	70.7	70.4	70.1	70.1
United States	68.5	66.4	64.7	64.7	65.5
EU27	68.4	65.1	63.1	61.5	62.5

EU and the US in 1970 (72.3 years). While it has fallen since then, at 70.1 years in 2010, it remained well above the US and EU values.

A number of broader points can be distilled from **Table 3**. First, in many countries the effective retirement age is below the 'normal' retirement of 65. Second, there is diversity across countries in retirement ages. In France in 2010, the age was below 60; as we have seen, in Japan in that year, it was over 70. These observations suggest that there is a range of factors which determine retirement ages, covering financial incentives and health and cultural norms.

Looking first at financial incentives, some of the most influential work can be found in Gruber and Wise (1999, 2004). Focusing on the 1999 volume, the goal of that study was to analyze the incentives for retirement that were generated by the social security systems in 11 countries and to link these incentives to observed retirement patterns. In order to have a common analytical device across all 11 countries, the following approach was adopted. For each country 'social security wealth' (SSW) was calculated which is the present discounted value of the future stream of social security benefits from when the person retires to when they die. By calculating SSW at different ages of retirement, it is possible to see what incentives are generated by the social security system.

To see how a system can generate incentives toward early retirement, consider the following example which is not atypical of the systems examined in Gruber and Wise (1999). Suppose a social security system allows people to draw benefits at age 62 and suppose that there is no increase in benefits if the person chooses to retire later than 62. In this case, SSW will be lower at age 63 compared to age 62 because the person will draw benefits for 1 year less (assuming that the timing of death is not affected by the timing of retirement). The fall in SSW can be viewed as a tax on working longer. If the person works beyond 62, (s)he will have earned income but if the SSW-related tax is sufficiently large, there will be a clear incentive to retire at 62. Even if the social security benefits are increased between 62 and 63, there will be an incentive for early retirement if the adjustment to the benefits is not actuarially fair.

The country-specific chapters in Gruber and Wise (1999) uncover evidence of incentives toward early retirement inherent in the design of many social security systems. In the case of some countries, the issues relate to payments other than old-age pensions; unemployment and illness/disability payments also have eligibility conditions which facilitate early retirement. Generally, the studies find patterns of retirement that are consistent with the incentives, notably, that a retirement spike occurs when benefits first become available. They also show that a second spike typically occurs around 65, the normal retirement age. This second spike cannot be readily explained in terms of financial incentives and suggests that cultural norms play a role in determining retirement choice. Generally, one of the striking outcomes from the Gruber and Wise study is the similar reactions to financial incentives across countries.

As noted with respect to **Table 3**, effective retirement ages had been falling in many countries up to 2000. Combined with fears over the sustainability of their systems (and possibly recognizing the design flaws as highlighted by Gruber and Wise, 1999), many governments implemented reforms in the last two decades (for a fuller discussion see Martin and Whitehouse, 2008). It is tempting to conclude that these reforms have been successful based on the increases in the effective retirement ages observed since 2000. Coppola and Wilke (2010) suggest that reforms in Germany have been successful in terms of raising the expected ages of retirement. Barrett and Mosca (2013) suggest that reforms in Ireland were less successful in achieving that goal, at least in the immediate aftermath of the announcement of the reform.

While financial incentives clearly play a role in retirement decisions, it seems likely that health would also play a role. Prior to the widespread availability of panel data, generating evidence on this point was difficult. As discussed in Bazzoli (1985), papers based on cross-sectional data which showed a strong positive link between health status and retirement may have suffered from 'ex post rationalization' on the part of survey respondents. By this, she meant that poor health could be used by respondents to explain the fact that they were currently retired, even if poor health had not prompted the retirement.

With the arrival of longitudinal datasets, such problems are reduced (although not necessarily eliminated). Emmerson and Tetlow (2006) use the first two waves of the English Longitudinal Study on Aging (ELSA) (2002/3 and 2004/5) to explore the differences in health status across people who exited the labor market and those who remained in it. Looking at people aged between 50 and 65, they show that people who reported that their health was either poor or fair were about twice as likely to exit the labor force compared to people who reported that their health was very good or excellent. This tendency toward exit for those in poor health was more prominent for those employed in manual occupations. Almost a third of male manual workers who reported fair health or worse at the first wave had left the labor force by the second wave; the corresponding figure for non-manual workers was just under a fifth.

McGarry (2004) looks at the link between health and retirement in a different way. She poses the question of whether or not retirement expectations are affected by health. The data which she uses, the US Health and Retirement Survey (HRS), include information arising from the following question:

> Thinking about work generally and not just your present job, what do you think are the chances that you will be working full-time after you reach age 62?

Using responses to this question as the explanatory variable, she relates health to retirement expectations, along with other variables including those related to finances. She finds that health does affect retirement expectations and also that the size of the effect is larger than that of the financial variables. She also finds that *changes* in health status give rise to *changes* in retirement expectations.

Before leaving this discussion of the impact of health on retirement, it is worth noting that researchers, including economists, have also explored whether retirement has an impact on health. One recent example is Behncke (2012) and her results are striking. Using the first three waves of the English Longitudinal Study on Aging, she finds that retirement

increases the risk of severe cardiovascular disease and cancer. Retirement also appears to increase risk factors for such conditions including BMI, cholesterol, and blood pressure. More generally, retirement seems to reduce self-assessed health. These findings lead Behncke to conclude that delaying the age at which the state pension is paid could be a policy tool for improving public health.

Productivity

Economists have long been interested in the relationship between age and productivity, but this interest was often motivated by an interest in the relationship between age, pay, and productivity together. Specifically, economists have theorized that the gap between productivity and pay is negatively related to age, and this seems to be related in part to an assumption that productivity declines with age whereas wages are assumed not to fall. If this is true, it has implications for the employability of older workers. Another possible implication is the need for mandatory retirement, as explained by Lazear (1979). In his model, younger workers are paid less than their marginal product at the early stages of their careers but they are 'promised' pay increases if they remain with the firm. This leads to higher degrees of attachment to the firm and the resulting lower rates of employee turnover allows for greater productivity. However, beyond a certain age, pay exceeds marginal product and this gives rise to the need for mandatory retirement.

Van Ours (2009) is among a group who have tried to look for direct evidence that productivity declines with age. He begins with a rather limited perspective and uses data on runners to assess whether physical capacity declines with age. He then uses data on the publication rates of academic economists to see if cognitive capacity diminishes. He finds evidence of declines with respect to his measure of physical capacity but not with respect to cognition.

He expands the scope of the analysis by using a linked employer–employee data set from the Netherlands to see whether the pay–productivity gap is related to age. He does not find evidence of a relationship. While his results are interesting, his conclusion is perhaps more striking. He says that the potential negative impacts of population aging on productivity should not be underestimated, but he also cautions that they should not be overestimated either.

Estimating the link between productivity and age at the level of the firm is a complicated task. Studies typically look at the share of employees of different ages and relate this to productivity levels across firms with different age structures. This basic formulation gives rise to many complicated methodological issues. For example, older employees may work in longer-established firms where the capital stock is older. In this case, a naive view of the age–productivity relationship would show older workers to be less productive but this would not be true. It could also be the case that outputs and labor input are determined simultaneously by firms and so a host of endogeneity issues arise.

Some studies have attempted to grapple with these issues. Aubert et al. (2006) explore the issue of the pay–productivity and age relationship, using a linked employer–employee dataset from France which includes information on 70 000 firms over the period 1994–2000. They find that productivity increases up until the age of 40 and then remains stable. This pattern is found to hold in each of the three sectors covered: manufacturing, trading, and services. They also find that the age–productivity profile is similar to the age–wage profile, and so their evidence does not support the notion of a divergence between productivity and wages at later ages.

Göbel and Zwick (2009) use German linked employer–employee panel data and they focus on estimating the relationship between age and productivity. They find that productivity at the firm level increases with the age of the employees up to the age of 55. Thereafter, there is only a slight decrease. Although they find little evidence of productivity falls on average, they also point out that there are large variations across firms in their sample. This implies that productivity falls with age in some settings, which in turn could be related to the nature of the work being undertaken. Göbel and Zwick also suggest that the maintenance of productivity levels in some firms may be related to the personnel policies operated by different firms.

Göbel and Zwick (2013) explore whether or not workplace policies aimed specifically at older workers can help to maintain the productivity levels of those older workers. They use a rich German dataset which includes information on the age structure of the employees of a large sample of firms and also output measures. The survey also includes questions on whether specific human resource measures for older workers were implemented. Although seven such measures were listed in the relevant question, Göbel and Zwick collapse them into the following five categories: specific equipment in workplaces; reduced working hours for old employees; age-specific jobs for old employees; mixed-age working teams; and training for old employees.

A mixed pattern of results emerged. Productivity levels among older workers were found to benefit from specific equipment in the workplace and from age-specific teams. Training for older workers and reduced working time were not found to have an impact. Possibly the most interesting result was with respect to mixed-age teams. In this case, the productivity of both older and younger workers was increased. Göbel and Zwick explain this result through the positive spillovers that arise when different skill sets and perspectives are brought to bear on an issue. This line of thinking has a parallel in the literature on immigration, where it is increasingly argued that the diversity in the workplace which results from immigration can also yield positive spillovers (Ottaviano and Peri, 2006). This is an important result because it recognizes the potential complementarities between younger and older workers and the scope this provides for productivity increases.

A different perspective on the age–productivity relationship is explored by Aubert et al. (2006) and Behaghel et al. (2014). The issue in these papers is whether technological change and innovative workplace practices reduce the demand for older workers, which can be taken to imply a fall in productivity relative to younger workers. Aubert et al. (2006) point out two routes through which effects could operate. First, to the extent that technological change leads to skills obsolescence, this is more likely to affect older workers who acquired their

skills in the past. The impact of technological change on lower skilled workers has been discussed extensively in the economics literature and has been proposed as an explanation for increasing income inequality (Levy and Murnane, 1992). Given this, it is unsurprising that the impact of technological change on older workers would have become a source of interest and concern. The second route through which innovation in the workplace could affect older workers is through less adaptability on their part. If new forms or work organization or new production processes require significant level of adaptation, this could be more challenging for workers who have operated in a particular setting for longer.

Based on a dataset of French firms, Aubert et al. (2006) found that the wage bill of older workers fell in the more innovative firms and that it rose for younger workers. This suggests that the demand for older workers fell relative to younger workers in these settings. The precise mechanism underlying the observation differed depending on whether the innovation related to technology or organizational change. In the case of technological introductions, the hiring of older workers fell. In the case of organizational innovation, rates of exit among the young fell.

Behaghel et al. (2014) expand this analysis. Using the same data set, they explore whether the negative effects on the share of older workers that resulted from innovation at the firm level can be offset by training. Their results show that while training for older workers can mitigate the effects shown by Aubert et al. (2006), such training does not offset the effect completely.

On balance, the evidence would seem to suggest that some of the assumed impacts on productivity of population are overstated. However, specific issues such as technological change do, and will, present specific challenges.

Conclusion

In this article, I have identified two routes through which population aging will impact the labor market. As the labor force is drawn primarily from those aged 20–64, the decrease in the number of this age group in many countries will lead to a fall in the labor force, all else equal. As this reduced labor force would have to support a larger number of older people, again all else equal, this is clearly a potential problem. In addition to the projected fall in labor forces, population aging will also result in an increase in the share of older workers. While many assume that this will lead to lower levels of productivity, based on the notion that older people are less productive, the evidence presented earlier suggests that any decreases in productivity will probably be limited (see Börsch-Supan, 2003, for further evidence on this).

This line of reasoning leads to the conclusion that the dominant policy concern in the aging/labor market intersection is with maintaining the size of the labor force and the old-age dependency ratio through delaying retirement. The first step on this road should be the elimination of financial incentives for early retirement, whether in the pensions system or in other elements of the welfare system such as unemployment or disability payments (Gruber and Wise, 1999). The second step should be the development of the working environment so that it facilitates and encourages delayed retirement (Göbel and Zwick, 2013). It could be that firms will choose to develop their workplaces in an 'age-friendly' response to the changing age structure of the population. Nevertheless, policy makers should still monitor this aspect of the workplace and consider active policy measures if progress is slow.

Better health is a contributing factor to population aging and also to delayed retirement. For this reason, we may witness (and may already be witnessing) a process whereby population aging sees a parallel extension in working lives. This is one of the potential processes that allow us to be less concerned about the potential impacts of population aging.

As a final note, the observation that retirement can be bad for your health is a potentially powerful message. As retirement is a decision which is difficult to reverse, there is less scope to 'experiment' than would be the case with other life choices. This means that people may be unaware of the fact that retirement is not necessarily 'bliss' (to quote Börsch-Supan, 2013) and so may be encouraged to remain in work if the possible impacts of retirement were more fully understood.

See also: Adulthood: Dependency and Autonomy; Age, Sociology of; Aging and Health in Old Age; Aging and Older People, Geography of; Aging and Work; Aging, Theories of; Life Course: Sociological Aspects; Population Aging: Economic and Social Consequences; Retirement and Encore Adulthood: The New Later Life Course; Retirement and Health; Retirement, Economics of; Welfare Reform.

Bibliography

Aubert, P., Caroli, E., Roger, M., 2006. New technologies, organisation and age: firm-level evidence. Economic Journal 116, F73–F93.

Aubert, P., Crépon, B., 2006. Age, Wage and Productivity: Firm-Level Evidence. Working Paper INSEE, Paris.

Barrett, A., Mosca, I., 2013. Increasing the state pension age, the recession and expected retirement ages. Economic and Social Review 44, 447–472.

Bazzoli, G., 1985. The early retirement decision: new empirical evidence on the influence of health. Journal of Human Resources 20, 216–234.

Behaghel, L., Caroli, E., Roger, M., 2014. Age-biased technological and organisational change, training and employment prospects of older workers. Economica 81, 368–389.

Börsch-Supan, A., 2013. Myths, scientific evidence and economic policy in an aging world. Journal of the Economics of Ageing 1-2, 3–15.

Börsch-Supan, A., 2003. Labour market effects of population aging. Labour 17 (Special issue), 5–44.

Behncke, S., 2012. Does retirement trigger ill health. Health Economics 21, 282–300.

Coppola, M., Wilke, C.B., 2010. How Sensitive Are Subjective Retirement Expectations to Increases in the Statutory Retirement Age? The German Case. Working Paper Mannheim Research Institute for the Economics of Aging.

Emmerson, C., Tetlow, G., 2006. Labour market transitions. In: Banks, J., Breeze, E., Lessof, C., Nazroo, J. (Eds.), Retirement, Health and Relationships of the Older Population in England: The 2004 English Longitudinal Study on Ageing (Wave 2). ELSA, London.

European Commission, 2012. The 2012 ageing report: economic and budgetary projections for the EU-27 member states (2010–2060). European Economy 2.

Göbel, C., Zwick, T., 2013. Are personnel measures effective in increasing productivity of old workers? Labour Economics 22, 80–93.

Göbel, C., Zwick, T., 2009. Age and Productivity – Evidence from Linked Employer Employee Data. ZEW Discussion Paper No. 09–020.

Gruber, J., Wise, D. (Eds.), 2004. Social Security Programs and Retirement Around the World. University of Chicago Press.

Gruber, J., Wise, D. (Eds.), 1999. Social Security and Retirement Around the World. University of Chicago Press.

Lazear, E., 1979. Why is there mandatory retirement? Journal of Political Economy 87, 1261–1284.

Levy, F., Murnane, R.J., 1992. US earnings levels and earnings inequality: a review of recent trends and proposed explanations. Journal of Economic Literature 30, 1333–1381.

Martin, J.P., Whitehouse, E., 2008. Reforming Retirement-Income Systems: Lessons from Recent Experiences of OECD Countries. OECD Social, Employment and Migration Working Paper No. 66, OECD, Paris.

McGarry, K., 2004. Health and retirement: do changes in health affect retirement expectations? Journal of Human Resources 39, 624–648.

Ottaviano, G.I., Peri, G., 2006. The economic value of cultural diversity: evidence from US cities. Journal of Economic Geography 6, 9–44.

Van Ours, J.C., 2009. Will you still need me? When I'm 64? De Economist 157, 441–460.

Aging and Work

Keith L Zabel and Boris B Baltes, Wayne State University, Detroit, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Even though we tend to think of and study age as one's age in calendar years, other conceptualizations of age exist, such as how long one has been in their current job or organization. Empirical research suggests that the way age is defined can impact the way it impacts a variety of outcomes in the workplace, such as absenteeism, innovative behavior, and safety behavior. In addition, empirical research suggests the negative stereotypes that are held about older workers typically are not supported with data. In fact, data often suggest the relationships are very different than the negative stereotype suggests.

The empirical study of aging in the workplace is increasing in importance, as changes in demographic trends (e.g., decline in birth rates and increasing life expectancy) have caused economic challenges (e.g., employees need to work longer to support the changing infrastructure of retirement systems), technical challenges (e.g., training older workers for changing technologies), and require greater cultural awareness and sensitivity due to increased globalization (Truxillo and Fraccaroli, 2013). The purpose of this article is to review how aging and its numerous different conceptualizations impact performance and functioning in the workplace. Although this article will almost entirely examine age effects on workplace outcomes as developmental, it is important to briefly describe two other types of aging effects.

Age-Related Effects

One difficulty in studying aging in the workplace is distinguishing between three types of aging effects, including developmental, cohort, and period (Parry and Urwin, 2011). In the context of the workplace, developmental effects refer to how workplace outcomes change as individuals' age or mature, whereas cohort effects refer to how historical events shape the workplace for differently aged employees. Generational cohorts have been defined as "an identifiable group that shares birth years, age location, and significant life evens at critical development stages" (Kupperschmidt, 2000: p. 66). Although the major cohorts have been operationally defined using many different year ranges, the most common conceptualization is that the Baby Boomer generation includes those born between 1946 and 1964, Generation X includes those born between 1965 and 1980, and Generation Y includes those born between 1981 and 1999 (Meriac et al., 2010). Finally, period effects refer to how the environment and values of the environment shape the workplace (Parry and Urwin, 2011). This article focuses almost exclusively on aging as developmental effects, though we recognize that cohort effects and period effects may also impact important workplace outcomes. Focusing on developmental effects, there are several different conceptualizations of age that can impact the relationship between age and workplace outcomes.

Different Conceptualizations of Age

Even though the majority of empirical research and popular press articles conceptualize age as chronological age, or one's age in calendar years, several other conceptualizations of age exist, including *functional/performance-based age*, *psychosocial/subjective age*, *organizational age*, and the *life span concept of age* (Kooij et al., 2008; Sterns and Doverspike, 1989). Functional age refers to one's age in terms of performance, regarding aspects such as cognitive ability and health.

Psychosocial age refers to how old one feels, thinks, acts, or desires to be. Organizational age captures one's age from an organizational perspective in terms of concepts like job tenure (length of time in one's current role), organizational tenure (length of time at one's employing firm), and career stage. Job tenure refers to the length of time one has been in their current functional role or job, whereas organizational tenure refers to the length of time one has been in their current organization. Finally, the life span concept of age refers to how aging processes might differ across one's life span, taking into consideration changes across life stages such as caring for young children or aging parents (Kooij et al., 2008). Even though there is substantial overlap between the different conceptualizations of age (e.g., between chronological age and organizational tenure), this article will illustrate how often, the *magnitude* of age predicting workplace outcomes depends upon the definition of age one uses.

Because more research has examined how chronological age, job tenure, and organizational tenure impact workplace outcomes relative to other conceptualizations of age, this article will place more focus on the chronological and organizational conceptualizations of age. Given the preponderance of workplace stereotypes about age, we begin each section by describing the stereotype and then discuss the extent to which empirical research supports or refutes the stereotype.

Age and Workplace Outcomes

One of the most common negative stereotypes of older workers is that older workers have lower levels of job performance than younger workers (Ng and Feldman, 2012). Three meta-analyses conducted by Ng and Feldman (2008, 2010b, 2012) have

examined how chronological age, organizational tenure, and job tenure, respectively, impact job performance. Meta-analytic results suggested the strongest age–job performance relationship was between organizational tenure and performance, with organizational tenure positively predicting core task performance (Ng and Feldman, 2010a). Even though organizational tenure positively predicted job performance, it only accounted for slightly over 1% of its variance. Therefore, the practical significance of the relationship is low. In addition, the positive relationship between organizational tenure and core task performance goes against the conventional negative stereotype that younger workers are better performers than older workers (Posthuma and Campion, 2009). On the contrary, these results suggest that the socialization process and experience workers have from working in an organization actually help improve their task performance, although the effect is small. When considering chronological age, however, meta-analytic results suggest there is no relationship between age and job performance (Ng and Feldman, 2008) or job tenure and job performance (Ng and Feldman, 2013a).

One explanation for the weak relationship between different conceptualizations of age and job performance is the presence of moderators. This proposition is supported by meta-analytic results that suggest the organizational tenure – job performance relationship was stronger for younger workers (those under age 37) compared to older workers (those over age 37; Ng and Feldman, 2010a). This finding is in line with the popular term 'Peter Principle,' which states that employees are promoted until they reach their own level of incompetence (Peter and Hull, 1969). To the extent to which it is more likely older workers have reached their level of incompetence relative to younger workers, this finding seems reasonable. Meta-analytic evidence also suggests that the organizational tenure-job performance relationship may be curvilinear in nature. Specifically, meta-analytic evidence suggests the organizational tenure–job performance relationship is significantly stronger at 3–6 years of organizational tenure relative to 0–3 years, 7–10 years, 11–14 years, and greater than 14 years (Ng and Feldman, 2010a).

Even though meta-analytic evidence suggests the age–job performance relationship is rather weak, research suggests the negative *stereotype* that older workers have lower levels of job performance than younger workers can impact behavior in the workplace. For example, research using a laboratory study found that the older target received more severe punishment recommendations than the younger target for making a mistake (Rupp et al., 2006). In addition, this research found that individuals with high levels of agism, or prejudice toward older individuals, gave significantly more severe punishment recommendations to the older target compared to the younger target (Rupp et al., 2006).

Similarly, research on the perceived productivity of workers across the life span revealed that employees under the age of 49 rate the perceived productivity of employees over 50 years old significantly lower than employees over 50 years old (Van Dalen et al., 2010). Perhaps more importantly, the aforementioned finding was found to be even stronger using a subset of the sample who self-identified as owners, heads of human resources departments, managing directors, and plant managers (Van Dalen et al., 2010). This finding suggests the negative stereotype that older workers are less productive than younger workers is strongest at the highest levels of organizations, thereby increasing the likelihood of litigation and increasing the likelihood that a culture of negative stereotypes toward older workers will permeate to the rest of the organization.

Given all the meta-analytic evidence, it seems the negative stereotype that a negative relationship exists between age and job performance is not warranted. That being said, findings suggest the negative stereotype does seem to affect how individuals perceive older workers when they make mistakes, and may impact the selection of qualified older workers for jobs when competing with qualified younger workers. Finally, results suggest years in one's organization are a better predictor than one's chronological age of job performance. To the extent to which one better learns how to navigate corporate policy and the culture of their organization after more time with that organization, this finding is warranted.

Training Performance

The negative stereotype also exists that older workers are less likely to seek out training and development or learning opportunities to grow in their career or position (Posthuma and Campion, 2009). Research regarding this negative stereotype is largely mixed. Specifically, one meta-analysis that examined the relationship between chronological age and training mastery found a moderate negative correlation, signifying that as individuals age, their training mastery decreases (chronological age accounted for 6% of the variance in training mastery; Kubeck et al., 1996). More recently, meta-analytic evidence found no effect of chronological age on performance (Ng and Feldman, 2008) or participation (Ng and Feldman, 2012) in training programs. Even though more evidence is needed to determine whether chronological age dampens training performance or participation, chronological age does seem to impact motivations that occur during the training process. Results of a meta-analysis found a small negative relationship between chronological age and both motivation to learn and learning self-efficacy, with chronological age accounting for 2–3% of the variance in those outcomes.

In summary, research regarding the effects of aging on training performance and related outcomes are rather mixed. Even though there does not seem to be an impact of chronological age on training participation, research does suggest chronological age negatively impacts training performance, motivation to learn, learning and learning self-efficacy. Given these findings, organizations would be wise to carefully select older workers for training sessions, and try to determine ways to increase their motivation to complete the training task.

Career Development

Another negative stereotype is that older workers have lower levels of career development motivation and behavior relative to younger workers (Ng and Feldman, 2012). Similar to motivation to learn and learning self-efficacy, meta-analytic results indicate as individuals' age chronologically, they tend

to have slightly lower career development motivation, as well as engage in slightly fewer career development behaviors (Ng and Feldman, 2012). Interestingly, findings suggest this negative relationship between chronological age and career development behaviors is stronger when the ratings were other-ratings as opposed to self-ratings, with chronological age of the target employee accounting for nearly 5% of the variance in career development behaviors rated by others (Ng and Feldman, 2012).

One possible explanation for the negative relationship between chronological age and career development behavior is that younger employees have greater access to career development within their organization than older employees. Recent research found that employees aged 26 years or younger reported greater opportunities for career development compared to those aged 36–42 years. Similarly, employees with less than 3 years of organizational tenure had greater access to learning and development opportunities compared to those with 3–10 years or greater than 10 years of organizational tenure (Pitt-Catsouphes et al., 2009).

In summary, results suggest a slight decrease in career development motivation and behavior as employees' age, both chronologically and within their organization, although the effect is rather small. Even though the relationship between chronological age and career development behaviors/motivation is rather weak, it can be augmented by the negative stereotype that older workers have much lower career development aspirations compared to younger workers. While the finding that career development behaviors decrease with age has been well supported by previous research, it is rather unclear whether this occurs because organizations focus their career development effort and resources more on younger employees, or whether career development interests decline as individuals age within the organization and chronologically.

Innovation/Resistance to Change

Another negative stereotype states that older workers are more resistant to change and have lower levels of innovation compared to younger workers (Posthuma and Campion, 2009). Recent empirical research found that contrary to this popular negative stereotype, younger workers were more resistant to change than older workers (Kunze et al., 2013). In addition, meta-analytic results found a weak positive relationship between chronological age and innovative behavior (Ng and Feldman, 2013b). Qualifying the main effect, findings suggest the relationship between chronological age and innovative behavior is curvilinear in nature, such that the relationship is stronger at less than 35 years of age compared to those over 35 years of age (Ng and Feldman, 2013b). This finding indicates the impact of age on innovation is greatest before age 35, and that after employees reach that age, there is no impact of age on innovative behavior.

Results are even more striking when examining the relationship between job tenure and innovative behavior. A recent meta-analysis found a rather strong positive relationship between job tenure and innovative behavior, with job tenure accounting for 6–12% of the variance in innovative behavior, depending on whether the ratings were self or supervisor ratings of innovative behavior, respectively (Ng and Feldman, 2013b). Qualifying the main effect, meta-analytic results suggest the relationship between job tenure and innovative behavior is curvilinear in nature, such that the relationship is strong at less than 10 years of job tenure, but no relationship exists at over 10 years of job tenure (Ng and Feldman, 2013b). This finding suggests that it may be easier to innovate with more experience on the job. However, after 10 years at the same job, the link from experience to innovation becomes weaker.

Even though meta-analytic evidence suggests that older workers actually tend to have higher levels of innovation and less resistance to change than younger workers, the negative stereotype that younger workers engage in higher levels of innovative behavior than older workers is real, and has consequence for older workers. For example, one study employing a laboratory design found that younger targets were rated more positively than older targets in regards to a number of dimensions of adaptive performance, including solving problems creatively, learning work tasks, cultural adaptability, interpersonal adaptability, and physical adaptability (DeArmond et al., 2006). In summary, contrary to the popular stereotype, innovative behavior actually increases with chronological age. In addition, job tenure positively predicts innovative behavior until age 35. Organizations would be wise to develop age diverse teams, both in terms of chronological age and job tenure to in order to maximize innovative behavior.

Organizational Citizenship Behavior/ Counterproductive Work Behavior

The positive stereotype also exists that older workers are more dependable, honest, and trustworthy than younger workers (Posthuma and Campion, 2009). One reasonable manifestation of this positive stereotype is the proposition that older workers have higher levels of organizational citizenship behaviors and lower levels of counterproductive work behaviors relative to younger workers. Organizational citizenship behaviors are extra, unexpected behaviors that better one's immediate work group or environment (e.g., helping a coworker), whereas counterproductive work behaviors are behaviors that intentionally harm business (e.g., stealing, using inappropriate language; Ng and Feldman, 2008). Meta-analytic evidence partially supports this positive stereotype. Older workers and those with higher levels of organizational tenure do tend to engage in slightly higher levels of organizational citizenship behavior and lower levels of counterproductive work behavior relative to their younger counterparts, and those with lower levels of organizational tenure, respectively, although the effect is relatively small (Ng and Feldman, 2008, 2010a). It is important to note that a stronger relationship was found between chronological age and organizational citizenship behavior directed at tasks (e.g., displaying extra effort on the job), with chronological age accounting for over 4% of the variance (Ng and Feldman, 2008). Thus, older workers are more willing to persist to try hard to improve personal or group performance relative to younger workers. In addition,

meta-analytic results suggest a positive relationship exists between job tenure and organizational citizenship behaviors, and that no relationship exists between job tenure and counterproductive work behaviors (Ng and Feldman, 2013a).

Moderators of the chronological age- counterproductive work behavior and organizational tenure–organizational citizenship behavior relationship exist. Specifically, recent research found the negative relationship between chronological age and counterproductive work behaviors is strongest for those over age 40 and lowest for those under age 25 (Ng and Feldman, 2008). In addition, meta-analytic results suggest the organizational tenure-organizational citizenship behavior relationship is stronger for younger workers (those under age 37) compared to older workers (those over age 37; Ng and Feldman, 2010a). Meta-analytic evidence also suggests that the organizational tenure–organizational citizenship behavior relationship may be curvilinear in nature. Specifically, meta-analytic evidence suggests the organizational tenure – organizational citizenship behavior relationship is significantly higher at 0–3 years of organizational tenure relative to 3–6 years, 7–10 years, 11–14 years, and greater than 14 years (Ng and Feldman, 2010a).

In summary, there seems to be a small positive relationship between three conceptualizations of age (chronological age, job tenure, and organizational tenure) and organizational citizenship behaviors. Given the organizational tenure–organizational citizenship behavior relationship is stronger at both younger ages and lower levels of job tenure, it is difficult to tell whether it is the organizational tenure or age that is specifically driving the moderation. However, to the extent to which organizational citizenship behaviors may increase one's chances of being promoted or receiving other positive workplace outcomes, it is more likely that years of tenure moderates the relationship between organizational tenure and organizational citizenship behaviors.

Absenteeism

Another positive stereotype of older workers is that older workers have lower levels of absenteeism relative to younger workers (Ng and Feldman, 2012). This stereotype seems to be based on fact since the strongest impact of age on performance outcomes can be found with absenteeism. Supporting the positive stereotype that older workers are more dependable than younger workers (Posthuma and Campion, 2009), meta-analytic results suggest that chronological age has the strongest impact on employee tardiness and absenteeism. Indeed, chronological age accounts for 7% of the variability in tardiness and absenteeism, with younger workers showing greater levels of tardiness and absenteeism than older workers (Ng and Feldman, 2008). A similar, albeit weaker trend, has been found in regards to organizational tenure, such that workers with higher levels of organizational tenure display lower levels of absenteeism (Ng and Feldman, 2010a). Contrary to chronological age and organizational tenure, job tenure has no impact on absenteeism (Ng and Feldman, 2013a). In summary, supporting the positive stereotype toward older workers, chronological age is a robust predictor of employee tardiness and absenteeism.

Safety Behavior

Similar to absenteeism, another positive stereotype of older workers is that older workers have higher levels of safety behavior relative to younger workers (Ng and Feldman, 2012). Safety behavior can generally be defined as the extent to which employees obey safety rules (Ng and Feldman, 2008). One reasonable manifestation of this stereotype is that older workers engage in greater levels of safety behavior or compliance relative to younger workers. Meta-analytic evidence supports this proposition, as findings suggest a negative relationship exists between organizational tenure and safety behavior, with organizational tenure accounting for over 3% of the variance in safety behavior (Ng and Feldman, 2010a). Meta-analytic evidence suggests a positive relationship exists between chronological age and safety compliance, and that older workers have lower incidences of workplace injuries relative to younger workers (Ng and Feldman, 2008), although the effects are lower than those accounted for by organizational tenure. Findings suggest there is no relationship between job tenure and safety compliance (Ng and Feldman, 2013a). In summary, older workers and workers with higher levels of job tenure tend to have slightly higher levels of safety behavior relative to younger workers and less tenured employees, respectively. These findings have important ramifications for organizations, which may be well advised to target their safety training on young employees and newly hired employees.

Age and Workplace Attitudes

In addition to safety behavior, the positive stereotype exists that older workers have higher levels of organizational commitment and lower levels of turnover intentions relative to younger workers (Ng and Feldman, 2012). Unlike many of the other workplace outcomes, much research has examined the existence of possible developmental and cohort effects on workplace attitudes. In regards to cohort effects, a recent meta-analysis examined the relationship between cohort and workplace attitudes including job satisfaction, organizational commitment, and turnover intentions (Costanza et al., 2012). Results suggested that older cohorts (e.g., Baby Boomers and Generation X) tended to have slightly higher levels of job satisfaction and organizational commitment than younger cohorts, although there was a wide level of variability in the findings of primary studies. In regards to turnover intentions, results suggested that Generation Y tends to have higher levels of turnover intentions than Generation X and Baby Boomers (Costanza et al., 2012). However, since the differences between the cohorts on workplace attitudes were rather small, Costanza et al. (2012) concluded that the practical significance of cohort differences in workplace attitudes was rather low.

In regards to developmental effects, a recent meta-analysis examined the relationship between chronological age and three broad types of job attitudes, including task-based attitudes (e.g., job satisfaction), attitudes toward others (e.g., coworker support), and organization-based attitudes (e.g., organizational commitment; Ng and Feldman, 2010b). In regards to task-based attitudes, meta-analytic analyses suggested a small

positive correlation between chronological age and job satisfaction, with chronological age accounting for over 3% of the variance in job satisfaction. On the other hand, analyses found a strong negative correlation between chronological age and satisfaction with promotion, with chronological age accounting for nearly 10% of the variance in satisfaction with promotion. Analyses also revealed small positive correlations between chronological age and job involvement and intrinsic motivation, as well as small negative correlations between chronological age and role overload and role ambiguity (Ng and Feldman, 2010b).

Meta-analytic results also suggested the relationship between attitudes toward others and chronological age was largely nonexistent, although results did suggest a small positive relationship between age and satisfaction with coworkers/supervisors, although chronological age only accounted for 1% of the variance in that outcome. In regards to organization-based attitudes, results suggested a small positive correlation between chronological age and organizational commitment and its facets (e.g., affective commitment, normative commitment), with age accounting for 4–6% of the variability in the facets of organizational commitment (Ng and Feldman, 2010b). It is worth noting that while controlling for job tenure did decrease the strength of the relationship between chronological age and the aforementioned attitudes, the relationships still existed.

In addition to the aforementioned attitudes, research has examined how age impacts retirement intentions and decisions. As one might expect, chronological age positively predicts retirement intentions (van Solinge and Henkens, 2009), although perceived life expectancy is a stronger predictor of retirement intentions than chronological age. Indeed, employees who feel they will live longer tend to have lower retirement intentions (van Solinge and Henkens, 2009). This research also found that the life span concept of age impacted retirement intentions, as employees with partners had lower levels of retirement intentions relative to those without partners. Meta-analytic evidence also suggests the performance-based age conceptualization of age predicts retirement behavior, as poor health is a weak, but significant predictor of making the decision to retire (Topa et al., 2009). The specific health markers that have the greatest impact on retirement attitudes and behavior, as they increase with age, seem to be blood pressure and cholesterol. Age accounts for 10% of the variability in blood pressure and 4% of the variability in cholesterol (Ng and Feldman, 2012).

One organizational intervention that seems to impact both task-based attitudes (e.g., job satisfaction) and retirement intentions is the use of flexible work arrangements (e.g., flextime, compressed work weeks, telecommuting; Baltes et al., 1999). Indeed, research suggests flexibility and having input on one's schedule positively predicts perceptions of supervisor effectiveness, which in turn positively predicts job satisfaction of older workers (James et al., 2007). Research also suggests that flexible work solutions lead to increased volunteering on the part of older workers, which in turn leads to decreased turnover (Havens and McNamara, 2007). As a specific example of an organizational intervention, the CVS Snowbird program allows US employees working for CVS to relocate for seasonal employment to a different CVS location in another area of the country, while keeping their current job (The Sloan Center on Aging & Work, 2012). With this program, CVS has decreased turnover and increased the percentage of older employees part in their workforce (The Sloan Center on Aging & Work, 2012). You can keep update to date on the most recent research around aging a multitude of issues in the workplace such flexible work arrangements by visiting the Web site of the Sloan Center on Aging & Work (http://www.bc.edu/research/agingandwork/).

In summary, results suggest that generally, there are weak relationships between age and workplace attitudes. One notable exception is satisfaction with promotion, as older employees have significantly lower satisfaction with promotion. It is worth noting that developmental effects and cohort effects were very similar across the two meta-analyses. However, in both cases, results suggested the effect of increased job satisfaction for older workers or Baby Boomers was rather weak. In addition, older employees, and especially those employees without partners or with health problems are likely to have higher turnover intentions than their counterparts. One type of organizational intervention that can successfully decrease retirement and/or turnover intentions and increase job satisfaction for older employees is the use of flexible work schedules. Given the success of flexible work schedules to increases employee satisfaction, more organizations would be wise to consider this organizational intervention when possible.

Workplace Motives

Another negative stereotype of older workers is that they place more importance on work relative to family (work–family conflict) than younger workers (Ng and Feldman, 2008). Recent research found the relationship between work–family conflict and age may be curvilinear in nature, as work–family conflict was lowest for younger and older workers, and highest for workers in the middle age group (Huffman et al., 2013). One explanation for this finding is that the life span conceptualization of age may be more important for predicting work–family conflict than chronological age. Indeed, a review of the work–family conflict literature found that generally, work–family conflict decreases as individuals move from having preschool children to have adolescent children (Baltes and Young, 2007). Given the curvilinear nature of the chronological age – work–family conflict relationship, it is not surprising that a meta-analysis found no relationship between chronological age and work–family conflict (Ng and Feldman, 2012).

In addition to importance of work relative to family, much previous research has examined how motives for working change as individuals move across the life span. Specifically, Kooij et al. (2011) found no effect of chronological age on social motives (motives that include interacting with others as a focal point). Results suggested a small positive correlation between chronological age and intrinsic motives for working, although the effect was rather small (accounted for less than 1% of the variability in the outcomes). Results suggested small negative correlations between chronological age and

growth motives, as well as security motives and extrinsic motives for work, although the effect was rather small (accounted for less than 1% of the variability in the outcomes). Meta-analytic results also found a moderate positive correlation between chronological age and autonomy needs and a moderate negative correlation between chronological age and need for promotion (chronological age accounted for between 5 and 7% of the variance in the outcome; Kooij et al., 2011). Further moderator results suggested the relationship between age and several workplace motives may actual be curvilinear in nature. For example, results suggested the relationship between chronological age and social motives was negative for younger workers, positive for middle-aged workers, but not significant for older workers (Kooij et al., 2011).

In summary, chronological age impacts several motives for working, including autonomy and promotion needs. In addition, the life span concept of age factors such as age of youngest child and marital status are meaningful predictors of the importance one places on work relative to family (e.g., work–family conflict). These findings have important implications for ways that organizations motivate employees across the life span. For example, older employers are less likely to be motivated by promotions and more likely to be motivated by being given a choice on how and/or where to perform their work. In addition, organizational interventions such as flexible work arrangements may be best utilized for those employees with young children in order to minimize work–family conflict.

Conclusions

In conclusion, age (and the way it is defined) impacts a variety of work outcomes that matter to individuals (e.g., job satisfaction), organizations (e.g., job performance), and societal structures (e.g., importance of work relative to family). As seen in Table 1, chronological age seems to have the greatest impact on health, absenteeism, career development behaviors rated by others, need for autonomy, and organizational commitment. As seen in Table 2, organizational tenure seems to have the greatest impact on innovation. As seen in Table 3, job tenure has the greatest impact on absenteeism and safety behavior.

Even though by definition, there is a large amount of overlap between these three definitions of age, the magnitude of their relationships with important workplace outcomes differ to a greater extent than seems to be typically recognized in the literature. It also seems that several conceptualizations of age are notably understudied in the applied literature, including performance-based age and psychosocial age. Given that the magnitude of the relationship between age and workplace outcomes depends on how you define age, it would be worthwhile to examine these underutilized definitions of age are related to important workplace outcomes. Finally, future research should also place increased attention to the life span concept of age and use more complicated designs (e.g., longitudinal). The many curvilinear relationships shown throughout Tables 1–3 suggest life span concept of age variables such as marital status and youngest age of child play an important role in predicting workplace outcomes.

Table 1 Effects of chronological age on workplace outcomes

Outcome	*Finding*	*Magnitude*
Job performance	● No relationship	● None
Training performance	● Mixed for training performance	● None
	● (−) For motivation to learn and learning self-efficacy	● Small
Career development	● (−) For career motivation and behaviors	● Small
	● (−) For career behaviors rated by others	● Moderate
Innovation/resistance to change	● (+) For resistance to change	● Small
	● (+) For innovation	● Small
	• Curvilinear	• Strongest for those under age 35 compared to over age 35
OCB[a]/CWB[b]	● (−) For counterproductive work behaviors	● Small
	• Curvilinear	• Strongest for those over 40 and lowest for those under 25
	● (+) For organizational citizenship behaviors	
	• Directed at tasks	● Small
		• Moderate
Absenteeism	● (−) For tardiness and absenteeism	● Moderate
Safety behavior	● (+) For safety compliance	● Small
Attitudes	● (+) For job satisfaction, job involvement, intent to stay	● Small
	● (+) For organizational commitment	● Moderate
	● (−) For satisfaction with promotion	● Strong
Motives	● No effect for social motives	● None
	● No effect for importance of work relative to family	● None
	• Curvilinear	• Work–family conflict highest for middle age compared to younger and older workers
	● (+) For intrinsic motives	● Small
	● (+) For autonomy needs	● Moderate
	● (−) For growth, security, and extrinsic motives	● Small

[a]Organizational citizenship behaviors.
[b]Counterproductive work behaviors.

Table 2 Effects of job tenure on workplace outcomes

Outcome	*Finding*	*Magnitude*
Job performance	● No relationship	● None
Innovation/resistance to change	● (+) For innovation • Moderation	● Strong • Stronger for those under 10 years of tenure compared to over 10 years
OCB[a]/CWB[b]	● No relationship for counterproductive work behaviors ● (+) For organizational citizenship behaviors	● None ● Weak
Absenteeism	● No relationship for tardiness or absenteeism	● None
Safety behavior	● No relationship for safety behavior	● None

[a]Organizational citizenship behaviors.
[b]Counterproductive work behaviors.

Table 3 Effects of organizational tenure on workplace outcomes

Outcome	*Finding*	*Magnitude*
Job performance	● (+) For job performance • Curvilinear • Moderated by chronological age	● Weak • Strongest for those at 3–6 years of tenure compared to 0, 6–10, 11–14, and over 14 • Stronger for workers under age 37 compared to over age 37
Career development	● Greater access for career development for those at under 3 years tenure compared to 3–10 years or over 10 years	
Innovation/resistance to change	● Curvilinear	● Curvilinear
OCB[a]/CWB[b]	● (−) For counterproductive work behaviors • Moderation ● (+) For organizational citizenship behaviors • Curvilinear	● Small • Strongest for those under 37 compared to over 37 ● Small • Stronger for those at 0–3 years of tenure compared to 3–6, 6–10, 11–14, and over 14
Absenteeism	● (−) For tardiness and absenteeism	● Moderate
Safety behavior	● (+) For safety behavior	● Moderate

[a]Organizational citizenship behaviors.
[b]Counterproductive work behaviors.

See also: Aging Mind: Facets and Levels of Analysis; Aging and the Labor Market; Aging, Theories of; Differential Aging; Ecology of Aging; Industrial–Organizational Psychology: Science and Practice; Job Analysis and Work Roles, Psychology of; Job Satisfaction; Organizational Citizenship Behavior; Organizational Commitment; Population Aging: Economic and Social Consequences; Retirement and Encore Adulthood: The New Later Life Course; Wellbeing and Burnout in the Workplace, Psychology of; Work Motivation.

Bibliography

Baltes, B.B., Briggs, T.E., Huff, J.W., Wright, J.A., Neuman, G.A., 1999. Flexible and compressed workweek schedules: a meta-analysis of their effects on work-related criteria. Journal of Applied Psychology 84, 496–513.

Baltes, B.B., Young, L.M., 2007. Aging and work/family issues. In: Adams, G., Schultz, K. (Eds.), Aging and Work in the 21st Century. Lawrence Erlbaum Associates, Mahweh, NJ, pp. 251–275.

Costanza, D.P., Badger, J.M., Fraser, R.L., Severt, J.B., Gade, P.A., 2012. Generational differences in work-related attitudes: a meta-analysis. Journal of Business and Psychology 27, 375–394.

DeArmond, S., Tye, M., Chen, P.Y., Krauss, A., Rogers, D.A., Sintek, E., 2006. Age and gender stereotypes: new challenges in a changing workplace and workforce. Journal of Applied Social Psychology 36, 2184–2214.

Havens, J., McNamara, T.K., 2007. Civic engagement: volunteering dynamics and flexible work options. The Sloan Center on Aging & Work 7, 1–9.

Huffman, A., Culbertson, S.S., Henning, J.B., Goh, A., 2013. Work-family conflict across the lifespan. Journal of Managerial Psychology 28, 1–35.

James, J.B., Swanberg, J.E., McKechnie, S.P., 2007. Responsive workplaces for older workers: job quality, flexibility, and employee engagement. The Sloan Center on Aging & Work 11, 1–10.

Kooij, D., de Lange, A., Jansen, P., Dikkers, J., 2008. Older workers' motivation to continue to work: five meanings of age: a conceptual review. Journal of Managerial Psychology 23, 364–394.

Kooij, D.T.A.M., de Lange, A.H., Jansen, P.G.W., Kanfer, R., Dikkers, J.S.E., 2011. Age and work-related motives: results of a meta-analysis. Journal of Organizational Behavior 32, 197–225.

Kubeck, J.E., Delp, N.D., Haslett, T.K., McDaniel, M.A., 1996. Does job-related training performance decline with age? Psychology and Aging 11, 92–107.

Kunze, F., Boehm, S.A., Bruch, H., 2013. Age, resistance to change, and job performance. Journal of Managerial Psychology 28, 1–32.

Kupperschmidt, B., 2000. Multigeneration employees: strategies for effective management. Health Care Manager 19, 65–76.

Meriac, J.P., Woehr, D.J., Banister, C., 2010. Generational differences in work ethic: an examination of measurement equivalence across three cohorts. Journal of Business and Psychology 25, 315–324.

Ng, T.W.H., Feldman, D.C., 2008. The relationship of age to ten dimensions of job performance. Journal of Applied Psychology 93, 392–423.

Ng, T.W.H., Feldman, D.C., 2010a. Organizational tenure and job performance. Journal of Management 36, 1220–1250.

Ng, T.W.H., Feldman, D.C., 2010b. The relationships of age with job attitudes: a meta-analysis. Personnel Psychology 63, 677–718.

Ng, T.W.H., Feldman, D.C., 2012. Evaluating six common stereotypes about older workers with meta-analytical data. Personnel Psychology 65, 821–858.

Ng, T.W.H., Feldman, D.C., 2013a. Does longer job tenure help or hinder job performance? Journal of Vocational Behavior 83, 305–314.

Ng, T.W.H., Feldman, D.C., 2013b. A meta-analysis of the relationships of age and tenure with innovation-related behaviour. Journal of Occupational and Organizational Psychology 86, 585–616.

Parry, E., Urwin, P., 2011. Generational differences in work values: a review of theory and evidence. International Journal of Management Reviews 13, 79–96.

Peter, L., Hull, R., 1969. The Peter Principle. William Morrow, New York.

Pitt-Catsouphes, M., Matz-Costa, C., Besen, E., 2009. Workplace flexibility: findings from the age & generations study. The Sloan Center on Aging & Work 19, 1–21.

Posthuma, R.A., Campion, M.A., 2009. Age stereotypes in the workplace: common stereotypes, moderators, and future research directions? Journal of Management 35, 158–188.

Rupp, D.E., Vodanovich, S.J., Crede, M., 2006. Age bias in the workplace: the impact of ageism and causal attributions. Journal of Applied Social Psychology 36, 1337–1364.

Sloan Center on Aging & Work. http://www.bc.edu/research/agingandwork/(accessed 13.11.13.).

The Sloan Center on Aging & Work, 2012. CVS Caremark Snowbird Program. Chestnut Hill, MA. Retrieved from: http://capricorn.bc.edu/agingandwork/database/browse/case_study/24047.

van Solinge, H., Henkens, K., 2009. Living longer, worker longer? The impact of subjective life expectancy on retirement intentions and behaviour. European Journal of Public Health 20, 47–51.

Sterns, H.L., Doverspike, D., 1989. Aging and the retraining and learning process in organizations. In: Goldstein, I., Katzel, R. (Eds.), Training and Development in Work Organizations. Jossey-Bass, San Francisco, CA, pp. 299–332.

Topa, G., Moriano, J.A., Depolo, M., Alcover, C., Morales, J.F., 2009. Antecedents and consequences of retirement planning and decision-making: a meta-analysis and model. Journal of Vocational Behavior 75, 38–55.

Truxillo, D.M., Fraccaroli, F., 2013. Research themes on age and work: introduction to the special issue. European Journal of Work and Organizational Psychology 22, 249–252.

Van Dalen, H.P., Henkens, K., Schippers, J., 2010. Productivity of older workers: perceptions of employers and employees. Population and Development Review 36, 309–330.

Aging Mind: Facets and Levels of Analysis

Shu-Chen Li, Chair of Lifespan Developmental Neuroscience, Technische Universität (TU) Dresden, Dresden, Germany

Published by Elsevier Ltd.

Abstract

Basic cognitive mechanisms, such as the abilities to briefly maintain, focus, and process information decline gradually as people get older. These phenomena have attracted much public and research attention. Many subfields of cognitive aging research have been advancing rapidly, but mostly in an independent disintegrated fashion until the late 1990s. The first decade of the twenty-first century has seen much more integrated research efforts that help advanced out understandings about the links between the aging mind and aging brain. This article reviews various aspects of aging cognition occurring at the behavioral, information processing, and neurobiological levels, and describes general theoretical considerations aimed at integrative research across different fields.

A Zeitgeist in Search for Interdiscipline Integration

Throughout most of the last century, much of the basic research on cognition has progressed in a rather segregated fashion, with differences in experimental paradigms, methodological, and theoretical orientations together with traditional discipline boundaries setting the dividing lines. Disintegrated research pursuits as such are common, as most endeavors in early stages of research development are first devoted to the discoveries of unique new phenomena and the constructions of competitive theoretical interpretations. As the fields progress with ever-increasing empirical data and theories, integration is then necessary to provide a comprehensive understanding of the accumulated information.

Proposals to Integrate the Studies of Brain, Cognition, and Behavior

The need for developing overarching integrations across the many subfields of cognitive psychology and cognitive science is evident in the last decade of the twentieth century. Approaches for integrating the studies of brain, cognition, and behavior have been independently proposed by researchers of different specialization (**Figure 1** shows a summary diagram). For instance, researchers in the area of artificial intelligence have proposed *cross-domain integration* aiming at constructing comprehensive models to capture different domains of cognitive and behavioral functioning, such as perception, memory, learning, decision-making, emotion, and motivation (e.g., Newell, 1990). There is also the cognitive and computational neuroscience approach of *cross-level integration*, which aims at integrating empirical regularities and theories of cognition across the behavior, information processing, and biological levels (see Gazzaniga et al., 2009 for review). Others, built on Burnswik's and Gibson's earlier emphases on the embeddedness of behavior and cognition in the environmental context, have suggested a *human-ecology integration* stressing that functional adaptivity arising from the human–environment interaction must be considered en route to discoveries of universal principles of behavior and cognition (e.g., Gigerenzer et al., 1999; Shepard, 1995). In order to better capture dynamic exchanges between environmental support and biological resources across the life span, developmental psychologists (e.g., Baltes et al., 1999) have advocated a life span integrative approach to study behavior and cognition. Although these approaches differ in the questions they address, they complement, rather than exclude, each other, with the first two approaches focusing on different domains and levels of cognition and behavior within a person, and the last two focusing on the person–environment interactions and the evolutionary-ontogenetic dynamics that are the backbones of biocultural coconstruction of human development (Baltes et al., 2006; Li, 2003).

Cognitive Aging Phenomena Been Studied at Various Levels

Couched within the broader research context, the field of cognitive aging had also gone through a period of disintegrated research and is now orienting toward integration (*see* Cognitive Aging). Since the 1920s when the first studies on adult age differences in mental abilities were published, studies on cognitive aging have mostly been carried out independently by individual difference and cognitive experimental psychologists and neuroscientists at the behavioral, information processing, and biological levels (**Figure 2** gives an overview). Designs and results from animal neurobiological studies are not always readily testable in human cognitive studies, and vice versa. Therefore, until the recent advances with neuroimaging techniques (Cabeza, 2001), data and theories of cognitive aging have been mostly confined within their respective levels. The goal of this article is, thus, to review evidence of age differences in intelligence and basic cognitive processing in ways that highlight the many facets of aging mind and point out some recent attempts that have been undertaken since the 1990s to link previously less integrated areas of research.

Adult Age Differences in Intelligence

At the behavioral level, psychologists interested in how aging might affect individual differences in intelligence have taken the *psychometric approach*, which has a long tradition (dating back to classical works by Spearman, Galton, and Binet in the

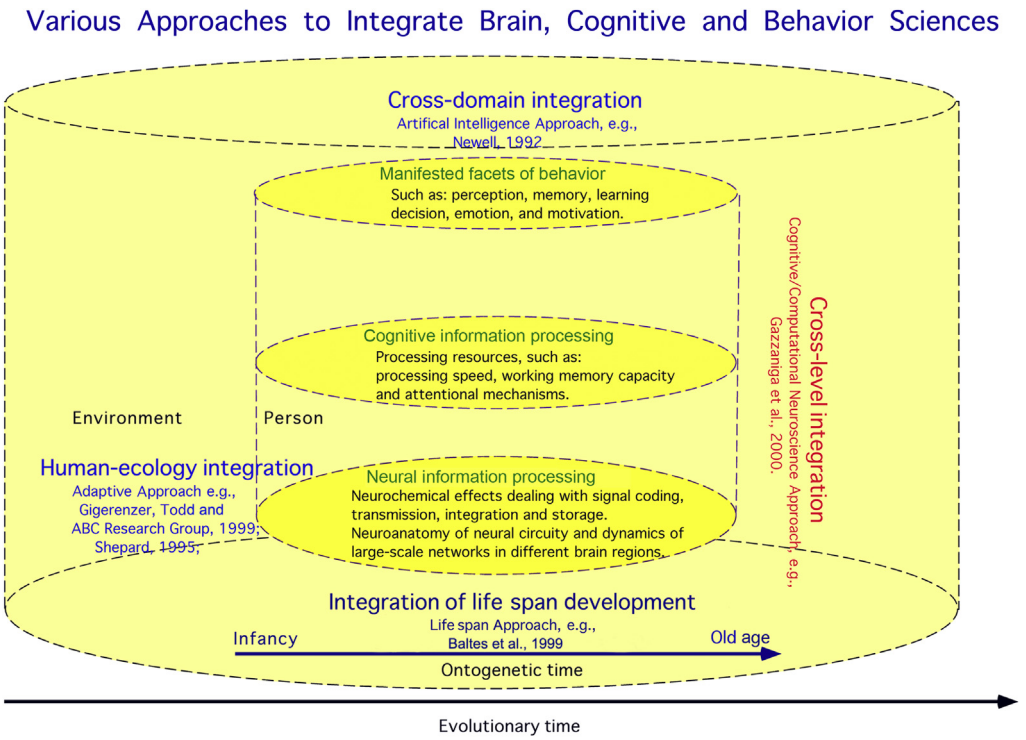

Figure 1 A summary diagram of different approaches proposed in the 1990s for integrating the various fields of brain, cognitive, and behavioral sciences.

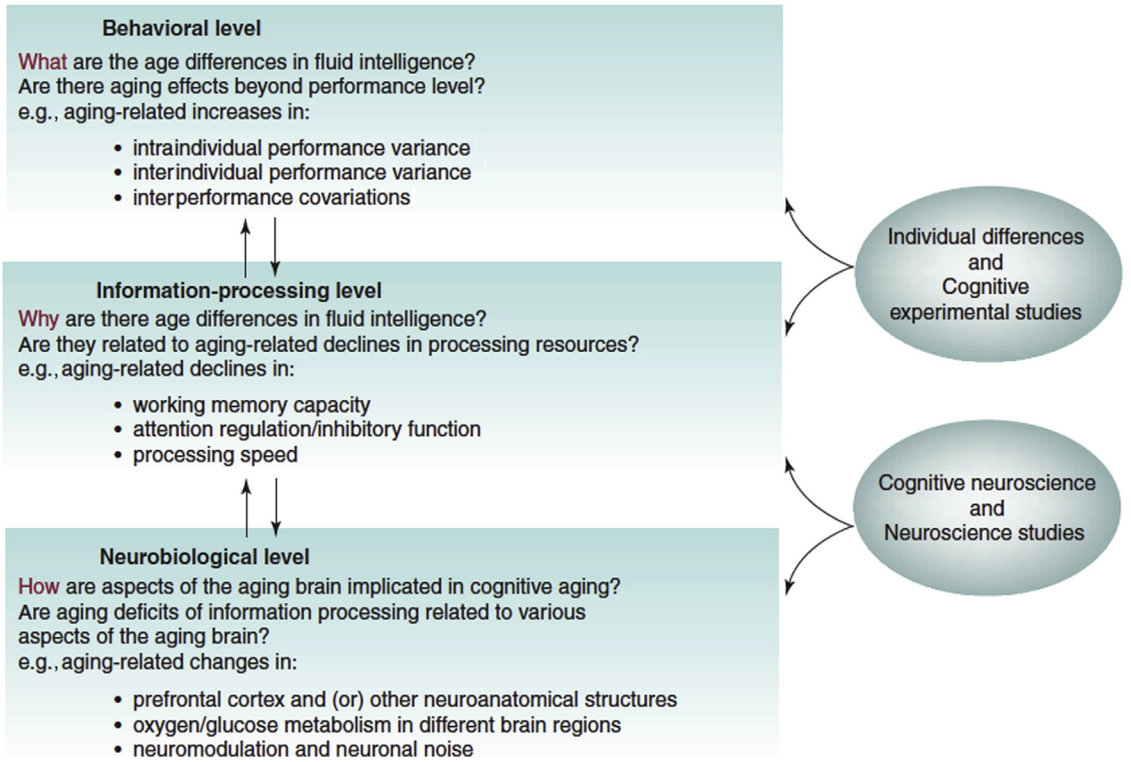

Figure 2 A summary diagram of various issues of aging mind addressed by researchers of different specialization at various levels. Figure adapted with permission from Li, S.-C., Lindenberger, U., Sikström, S., 2001. Aging cognition: from neuromodulation to representation. Trends in Cognitive Sciences 5, 480, Copyright 2001 Elsevier.

1880s and early 1900s), and focused on the *measurement* of age differences in intellectual abilities. The existing psychometric data to date indicate that intellectual aging is multifaceted. Furthermore, aging effects can be observed at three aspects of the behavioral data, namely, performance level, variability, and covariation.

Differential Age-Gradients of Cognitive Mechanics and Pragmatics

Traditionally, two-component models of intelligence distinguish between fluid intelligence reflecting the operations of neurobiological 'hardware' supporting basic information-processing cognitive mechanics and crystallized intelligence reflecting the culture-based 'software' constituting the experience-dependent cognitive pragmatics (Baltes et al., 1999; Horn, 1982). **Figure 3** shows that the fluid mechanics, such as reasoning, spatial orientation, perceptual speed, and verbal memory, show gradual age-related declines starting at about the 40s; while other abilities indicating the crystallized pragmatics, such as number and verbal abilities, remain relatively stable up until the 60s (e.g., Schaie and Willis, 1993). Furthermore, there have also been some recent ongoing theoretical and empirical efforts devoted toward expending the concepts of cognitive mechanics and pragmatics. In addition to the efficacy of information processing, cognitive mechanics also encompasses the optimal allocation of cognitive resources (e.g., Li et al., 2001b). The cognitive pragmatics has been expanded to include many other general as well as person-specific bodies of knowledge and expertise associated with the occupational, leisure, and cultural dimensions of life (e.g., Blanchard-Fields and Hess, 1996). One example is *wisdom*, the "expert knowledge about the world and fundamental pragmatics of life and human affair" that an individual acquires through his or her life history, that also includes an implicit orientation toward maximizing individual and collective well-being (Baltes and Staudinger, 2000).

Age-Related Increase in Variability and Covariation

In addition to age differences in the performance levels of the cognitive mechanics, behavioral data also point to age-related increases of performance variations within a person (e.g., Hultsch et al., 2000) and differences between individuals (for review see Nelson and Dannefer, 1992; *see* Age Stratification). Furthermore, much cross-sectional data show that as people age performances of different subscales of intelligence tests become more correlated with each other (e.g., Babcock et al., 1997), which has been taken as indications of less differentiated ability structure in old people.

Deficits in Basic Information-Processing Mechanisms

In light of age-related declines in psychometrically measured cognitive mechanics, the *information-processing approach* emerged from the rise of information theory and computers in the 1940s was advanced to *explain* age differences in fluid intelligence by identifying age differences in basic information-processing mechanisms. Thus far, age-related declines have been found in three main facets of information processing: people's abilities to keep information in mind, attend to relevant information and process information promptly are compromised with age.

Working Memory

Working memory (WM) refers to people's ability to simultaneously hold information in immediate memory while operating on the same or other information. Age-related decline in WM capacity has been obtained on a variety of experimental tasks including backward digit span, sentence span, and several types of computational span (e.g., Park et al., 1996; see **Figure 4(a)**). Age-related decline in WM capacity plays a role in many other cognitive activities where WM is implicated, ranging from long-term memory encoding and retrieval, syntactic processing, language comprehension, and reasoning (for review see Zacks et al., 2000).

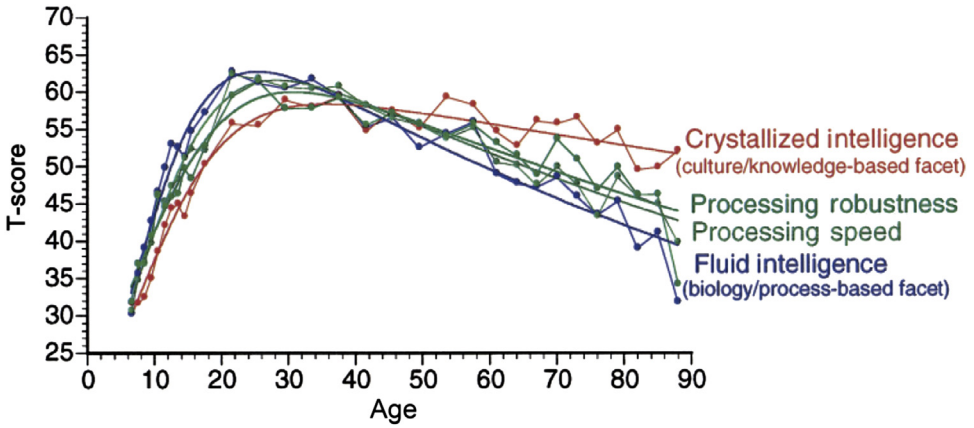

Figure 3 Differential trajectories of fluid mechanic and crystallized pragmatic intelligence as well as processing speed and robustness across the life span. Abilities were assessed with 3–4 different tests, and were scaled in a T-score metric. Figure adapted from Li, S.-C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., Baltes, P.B., 2004. Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. Psychological Science, 15, 158 with permission, Copyright 2004 Association for Psychological Science.

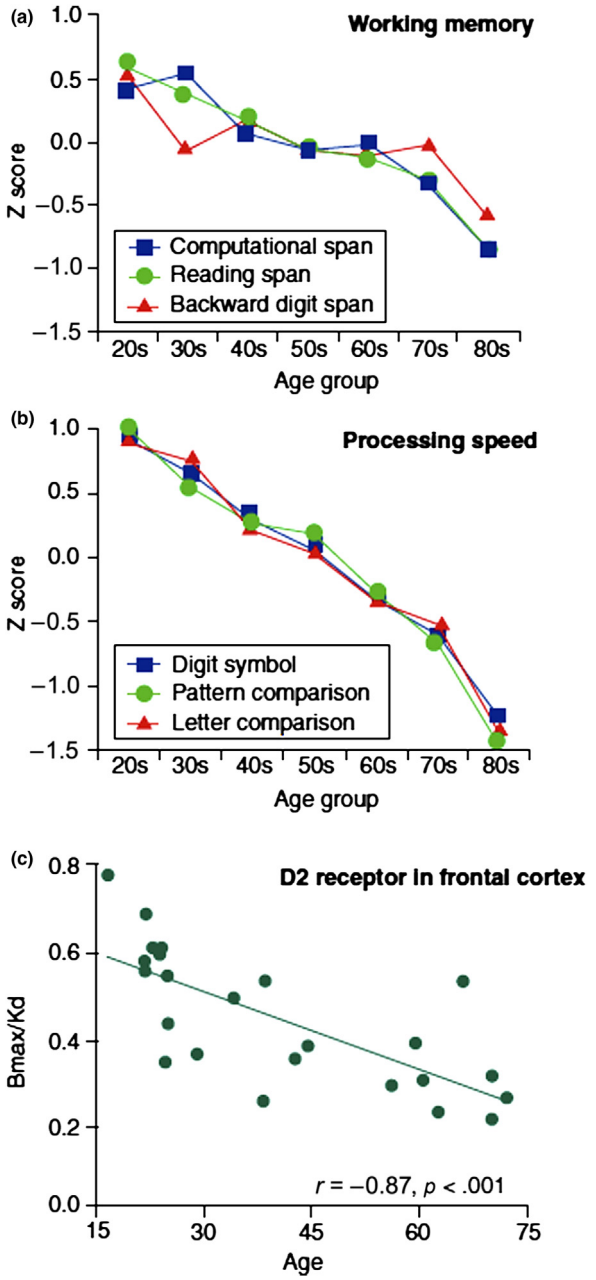

Figure 4 Age-related declines in keep aspects of information processing (WM and processing speed) and neurotransmission. (a) Negative adult age differences in WM (measured by three types span test, and were scaled in Z-score metric). (b) Negative age differences in processing speed (measured by three perceptual speed tests). (c) Age-related decline in dopamine D2 receptor in the frontal cortex. Figure adapted with permission from Li, S.-C., Lindenberger, U., Sikström, S., 2001. Aging cognition: from neuromodulation to representation. Trends in Cognitive Sciences 5, 481, Copyright 2001 Elsevier; Data sources in (a) and (b) based on Park, D.C., et al., 1996. Mediators of long-term memory performance across the lifespan. Psychology and Aging 4, 621–637, Copyright 1996 American Psychological Association, adapted with permission; Data in (c) adapted with permission from Kaasinen et al., 2000. Neurobiology of Aging, vol. 21, Copyright 2000 Elsevier.

Attentional and Inhibitory Mechanisms

Empirical data abound showing that old people have more problems attending to relevant information and ignoring irrelevant information. Negative age differences have been found in various selective and focus attention tasks along with the Stroop and proactive interference tasks (e.g., Lair et al., 1969). Age-related declines in attentional and inhibitory mechanisms have functional consequences on language comprehension, memory, problem solving, and other daily activities such as driving (see McDowd and Shaw, 2000 for review).

Processing Speed and Variability

Speed is a ubiquitous aspect of information processing. All information processing takes time, however fast it is. There is abundant evidence showing that older people are slower in responding comparing to young adults in almost every cognitive task in which processing speed is measured (see Figures 3 and 4(b)). Many correlational analyses showed that the observed age differences in fluid intelligence is greatly reduced or eliminated after controlling for individual differences in processing speed (see Salthouse, 1996 for review). Furthermore, other than processing speed, cognitive aging research over the last decade has consistently shown that aging-related decline in processing variability (fluctuations in trial-by-trial reaction times) adds additional unique effects in predicting individual differences in cognitive declines (Lövdén et al., 2007). Neurocomputational theory (Li et al., 2001b) as well as empirical evidence (see MacDonald et al., 2009, for review) suggest that aging-related declines in dopaminergic modulation (see **Figure 4(c)**) is also one contributing factor to age-related increase in processing variability.

Resource-Reduction Account

Given clear age-related declines in these fundamental aspects of information processing, the most prominent account for cognitive aging deficits thus far has been the general conceptual framework of age-related reduction in processing resources that are indicated by WM capacity, attention regulation, and processing speed (see Salthouse, 1991 for review). However, two major difficulties limit the resource-reduction theory. First, the different aspects of processing resources are *not* independent of each other. Second, the account itself is circular in nature: old people's lower proficiency in cognitive performance is assumed to be caused by a reduction in processing resources, and at the same time, poor performance is taken to be the indication of reduced processing resources. One way to avoid such circularity is to establish better correspondence between the proposed processing resources and their potential neurobiological underpinnings. Lest this be viewed only as reductionistic, it should be mentioned that psychometric data showing stronger trends of age-related decline in the biology-based fluid intelligence motivates the search for biological correlates. Experimental evidence of age-related decline in basic facets of information processing helps focusing the studies of brain aging on those aspects relevant to the affected information-processing mechanisms. Recent developments in cognitive and computational neurosciences have opened new revenues for studying the

functional relationships between behavioral manifestations of the aging mind and the biology of the aging brain.

The Aging Brain of the Aging Mind

At the neurobiological level, brain aging involves both neuroanatomical and neurochemical changes. Anatomically, there are structural losses in neurons and synaptic connections and reductions in brain atrophy (see Raz, 2000 for review). Neurochemically, there is evidence for deterioration in various neurotransmitter systems (Schneider et al., 1996 for review). However, progressive neuroanatomical degeneration resulting from cell death and reduced synaptic density is primarily characteristic of pathological aging such as Alzheimer's disease, and there is now evidence suggesting that milder cognitive problems occurring during normal aging are mostly due to neurochemical shifts in still-intact neural circuitry (Morrison and Hof, 1997; *see* Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences).

Attenuated Neuromodulation

Among different neurotransmitter systems, the catecholamines, including dopamine (DA) and norepinephrine (NE), are important neurochemical underpinnings of age-related cognitive impairments for several reasons. First, there is consensus for age-related decline in catecholaminergic function in the prefrontal cortex (PFC) and basal ganglia. Across the adult life span, dopaminergic function in the basal ganglia decreases by 5–10% each decade (see Schneider et al., 1996). Furthermore, many DA pathways in the basal ganglia are interconnected with the frontal cortex through the frontal-striatal circuits (Graybiel, 1990), hence are in close functional association with the PFC cognitive processes. Second, research over the last two decades suggests that catecholamines modulate the PFC's utilization of briefly activated cortical representations of external stimuli to circumvent constant reliance on environmental cues and to regulate attention to focus on relevant stimuli and appropriate responses (see Arnsten, 1998 for review). Third, there are many findings indicating, specifically, functional relationships between age-related deficits in the dopaminergic system and deficits in various aspects of information processing. For instances, reduced dopamine receptor density in old rats' nigrostriatum decreases response speed and increases reaction time variability (MacRae et al., 1988). Drugs that facilitate dopaminergic modulation alleviate WM deficits of aged monkeys who suffer from 50% dopamine depletion in their PFC (see Arnsten, 1998 for review). In humans, age-related attenuation of dopamine D2 receptor's binding mechanism is associated with declines in processing speed and episodic memory (see Bäckman et al., 2006 for review).

Reduced Hemispheric Asymmetry

In addition to changes in the aging brain's neurochemical environment, recent neuroimaging evidence suggests that cortical information processing in different regions of the brain becomes less differentiated as people age, phenomena that parallel the behavioral findings of less differentiated ability structure in old people. In comparison to the more clearly lateralized cortical information processing in young adults, people in their 60s and beyond showed bilateralized (bihemispheric) activity during retrieval (e.g., Cabeza et al., 1997; Cabeza, 2001) and during both verbal and spatial WM tasks (Reuter-Lorenz et al., 2000).

Integrating the Facets and Levels of the Aging Mind

Faced with the various facets of the aging mind across the different levels, the various subfields of cognitive aging research are ever more inclined to and in need of overarching frameworks for integration (cf Stern and Carstensen, 2000). Some integrative research undertakings along the four general approaches for integrating the studies of brain, cognition, and behavior sciences have already been underway.

With respect to better integrating the human–environment exchange and the evolutionary-ontogenetic dynamics, at a macro level some researchers embedded issues of cognitive aging within a metatheoretical framework of biological and cultural coevolution for studying life span human development (*see* Lifespan Development: Evolutionary Perspectives). While the benefits of evolutionary selection and the efficacy of neurobiological implementations of the mind decrease with aging, the needs for environmental-cultural support increases. In this systemic functional framework it is important for future research to investigate how declines in cognitive resources may be compensated by the individual's more selective allocations of these resources to different task domains and by cultural-environmental supports such as cognitive training (e.g., Dixon and Bäckman, 1995; K.Z.H. Li et al., 2001a). At a more specific level, other researchers have also suggested an environmental support perspective for understanding age differences in episodic memory and attentional mechanisms. Better environmental stimulus and contextual supports are helpful in overcoming age-related deficits in effortful self-initiated processes implicated in various memory and attentional tasks (e.g., Craik, 1986; Park and Shaw, 1992).

Regarding better integrating different domains and levels of behavior and cognition within the person, some researchers have started to work toward bridging the gaps between age-related declines in basic memory and attentional processes and higher-level cognitive function, such as language comprehension (e.g., Light and Burke, 1988; Burke, 1997). Regarding cross-level integration, there actually have been a few classical proposals trying to relate individual differences in the performance level, variance and covariation of intellectual functioning with individual differences in general brain energy (Spearman, 1927) and to link age-related cognitive aging deficits with increased neuronal noise (e.g., Welford, 1965). However, these long-range brain-behavior links could not be specified with more details in early research. It has only recently become possible to investigate these links more explicitly in cognitive and computational neurosciences. There is now some consensus for associations between PFC dysfunctions and aging-related cognitive impairments (West, 1996). However, details of the functional relationships between PFC impairments, aging attenuated neuromodulation, the distribution of information processing across different neural circuitry, and

various behavioral manifestations of cognitive aging deficits await further explications. Recently, one computational neuroscience approach has been undertaken to explore the links between age-related declines in neuromodulatory mechanisms innervating the PFC, noisier neural information processing, and adult age differences in episodic memory, interference susceptibility, performance variability, and covariation (e.g., Li et al., 2001b; Li, 2002). Empirical evidence accumulated during the first decade of the twenty-first century, support this theory and indicates that aging-related declines in neuromodulation, particularly the catecholamines, contribute to various aspects of cognitive deficits, ranging from episodic memory (e.g., Li et al., 2013; Papenberg et al., 2013), WM, and processing speed and fluctuations (see Bäckman et al., 2006, 2010 for reviews).

These integrative research orientations have different advantages and disadvantages. While theoretical considerations about environmental and evolutionary impacts on the aging mind at the metatheoretical level have the strength in providing overarching organizations, they need to be complemented by more information processing and neurobiologically oriented approaches to generate predictions that are more amenable for direct empirical validations. In the process of coevolving a range of related fields, there may not be a 'right' level for integration; rather there is the task to supplement and balance the weaknesses and strengths of different approaches.

Conclusion

The average life expectancy in most industrialized countries has been increasing from an average of about 45 years in 1900 to about 75 years by the end of the twentieth century (Oeppen and Vaupel, 2002). According to the United Nation's 2011 report on world population prospects, the number of people aged 65 years or older will outnumber children under age of 5 years before 2020 (WHO, 2011). A major task for cognitive aging research in the twenty-first century is to identify the causes of and methods to minimize or compensate these cognitive declines, so that the blessings of improved physical health and extended life expectancy in old age could be accompanied by sound aging mind. Attempts to achieve this challenging task require collective contributions of studies from the various subfields of cognitive aging research, ranging from individual difference-based psychometric and behavioral experimental studies to cognitive and computational neurosciences in the future. Furthermore, research on the aging mind necessarily entails an applied orientation; therefore, future research also needs to include more specific focuses on identifying age-relevant knowledge, aging-friendly social and environmental contexts, and aging-rectifying training programs to help old people better allocate and compensate their declining cognitive resources.

See also: Age Stratification; Age Structure; Age, Sociology of; Aging, Theories of; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Cognitive Aging; Ecology of Aging; Lifespan Development: Evolutionary Perspectives; Population Aging: Economic and Social Consequences.

Bibliography

Arnsten, A.F.T., 1998. Catecholamine modulation of prefrontal cortical cognitive function. Trends in Cognitive Science 2, 436–447.

Babcock, R.L., Laguna, K.D., Roesch, S.C., 1997. A comparison of the factor structure of processing speed for younger and older adults: testing the assumption of measurement equivalence across age groups. Psychology and Aging 12, 268–276.

Bäckman, L., Ginovart, N., Dixon, R., Wahlin, T., Wahlin, A., Halldin, C., Farde, L., 2000. Age-related cognitive deficits mediated by changes in the striatal dopamine system. American Journal of Psychiatry 157, 635–637.

Bäckman, L., Nyberg, L., Lindenberger, U., Li, S.-C., Farde, L., 2006. The correlative triads among aging, dopamine, and cognition: current status and future perspectives. Neuroscience and Biobehavioral Reviews 30, 791–807.

Bäckman, L., Lindenberger, U., Li, S.-C., Nyberg, L., 2010. Linking cognitive aging to alterations in dopamine transmitter functioning: recent data and future avenues. Neuroscience and Biobehavioral Reviews 34, 670–677.

Baltes, P.B., Staudinger, U.M., 2000. A metaheuristic (pragmatic) to orchestrate mind and virtue toward excellence. American Psychologist 55, 122–136.

Baltes, P.B., Staudinger, U., Lindenberger, U., 1999. Lifespan psychology: theory and application to intellectual functioning. Annual Review of Psychology 50, 471–507.

Baltes, P.B., Reuter-Lorenz, P., Rösler, F. (Eds.), 2006. Lifespan Development and the Brain: The Perspectives of Biocultural Co-constructivism. Cambridge University Press, Cambridge, UK.

Blanchard-Fields, F., Hess, T.M. (Eds.), 1996. Perspectives on Cognitive Change in Adulthood and Aging. McGraw-Hill, New York.

Burke, D.M., 1997. Language, aging, and inhibitory deficits: evaluation of a theory. Journals of Gerontology Series B-Psychological Sciences & Social Sciences 52B, P254–P264.

Cabeza, R., Grady, C.L., Nyberg, L., McIntosh, A.R., Tulving, E., Kapur, S., Jennings, J.M., Houle, S., Craik, F.I.M., 1997. Age-related differences in effective neural connectivity. Neuroreport 8, 3479–3483.

Cabeza, R., 2001. In: Cabeza, R., Kingstone, A. (Eds.), Functional Neuroimaging of Cognitive Aging. Handbook of Functional Neuroimaging of Cognition. MIT Press, Cambridge, MA.

Craik, F.I.M., 1986. A functional account of age differences in memory. In: Klix, F., Hagendorf, H. (Eds.), Human Memory and Cognitive Capabilities: Mechanisms and Performances. Elsevier, Amsterdam, pp. 409–422.

Dixon, R.A., Bäckman, L. (Eds.), 1995. Compensating for Psychological Deficit and Declines: Managing Losses and Promoting Gains. LEA, Hillsdale, NJ.

Gazzaniga, M.S., et al. (Eds.), 2009. The Cognitive Neuroscience, fourth ed. MIT Press, Cambridge.

Gigerenzer, G., Todd, P., the ABC Research Group, 1999. Simple Heuristics that Make Us Smart. Oxford University Press, New York.

Graybiel, A.M., 1990. Neurotransmitters and neuromodulators in the basal ganglia. Trends in Neurosciences 13, 244–253.

Horn, J.L., 1982. The theory of fluid and crystallized intelligence in relation to concepts of cognitive psychology and aging in adulthood. In: Craik, F.I.M., Trehub, S. (Eds.), Aging and Cognitive Processes. Plenum Press, New York, pp. 237–278.

Hultsch, D.F., MacDonald, S.W.S., Hunter, M.A., Levy-Bencheton, J., Strauss, E., 2000. Intraindividual variability in cognitive performance in older adults: comparison of adults with mild dementia, adults with arthritis, and healthy adults. Neuropsychology 14, 588–598.

Lair, C.V., Moon, W.H., Kausler, D.H., 1969. Associative interference in the paired-associative learning of middle-aged and old subjects. Developmental Psychology 5, 548–552.

Li, K.Z.H., Lindenberger, U., Freund, A.M., Baltes, P.B., 2001a. Walking while memorizing: a SOC study of age-related differences in compensatory behavior under dual-task conditions. Psychological Science 12, 230–237.

Li, S.-C., Lindenberger, U., Sikström, S., 2001b. Aging cognition: from neuromodulation to representation. Trends in Cognitive Sciences 5, 479–486.

Li, S.-C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., Baltes, P.B., 2004. Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. Psychological Science 15, 155–163.

Li, S.-C., Papenberg, G., Nagel, I.E., Preuschhof, C., Schröder, J., Nietfeld, W., Bäckman, L., 2013. Aging magnifies the effects of dopamine transporters and D2 receptor genes on backward serial recall. Neurobiology of Aging 34, 358e1–358e10.

Li, S.-C., 2002. Connecting the many levels and facets of cognitive aging. Current Directions in Psychological Science 11, 38–43.

Li, S.-C., 2003. Biocultural orchestration of developmental plasticity across levels: the interplay of biology and culture in shaping the mind and behavior across the life span. Psychological Bulletin 129, 171–194.

Light, L.L., Burke, D.M., 1988. Patterns of language and memory in old age. In: Light, L.L., Burke, D.M. (Eds.), Language, Memory and Aging. Cambridge University Press, New York, pp. 244–272.

Lövdén, M., Li, S.-C., Shing, Y.L., Lindenberger, U., 2007. Within-person trial-by-trial variability precedes and predicts cognitive decline in old and very old age: longitudinal data from the Berlin Aging Study. Neuropsychologia 45, 2827–2838.

MacDonald, S.W.S., Li, S.-C., Bäckman, L., 2009. Neural underpinnings of within-person variability in cognitive functioning. Psychology and Aging 24, 792–808.

MacRae, P.G., Spirduso, W.W., Wilcox, R.E., 1988. Reaction time and nigrostriatal dopamine function: the effect of age and practice. Brain Research 451, 139–146.

McDowd, J.M., Shaw, R.J., 2000. Attention and aging. A functional perspective. In: Craik, F.I.M., Salthouse, T.A. (Eds.), The Handbook of Aging and Cognition. LEA, Mahwah, NJ, pp. 221–292.

Morrison, J.H., Hof, P.R., 1997. Life and death of neurons in the aging brain. Science 278, 412–429.

Nelson, E.A., Dannefer, D., 1992. Aged heterogeneity: facts or fictions? The fate of diversity in gerontological research. Gerontologist 32, 17–23.

Newell, A., 1990. Unified Theories of Cognition. Harvard University Press, Cambridge, MA.

Oeppen, J., Vaupel, J.W., 2002. Broken limits to life expectancy. Science 296, 1029–1031.

Papenberg, G., Bäckman, L., Nagel, I.E., Nietfeld, W.N., Schröder, J., Bertram, L., Heekeren, H.R., Lindenberger, U., Li, S.-C., 2013. Dopaminergic gene polymorphisms affect long-term forgetting in old age: further support for the magnification hypothesis. Journal of Cognitive Neuroscience 25, 571–579.

Park, D.C., Shaw, R.J., 1992. Effects of environmental support on implicit and explicit memory in younger and older adults. Psychology and Aging 7, 632–642.

Park, D.C., et al., 1996. Mediators of long-term memory performance across the lifespan. Psychology and Aging 4, 621–637.

Raz, N., 2000. Aging of the brain and its impact on cognitive performance: integration of structural and functional findings. In: Craik, F.I.M., Salthouse, T.A. (Eds.), The Handbook of Aging and Cognition. LEA, Mahwah, NJ, pp. 1–90.

Reuter-Lorenz, P.A., Jonides, J., Smith, E., Marshuetz, C., Miller, A., Hartley, A., Koeppe, R., 2000. Age differences in the frontal lateralization of verbal and spatial working memory revealed by PET. Journal of Cognitive Neuroscience 12, 174–187.

Salthouse, T.A., 1991. Theoretical Perspectives of Cognitive Aging. LEA, Hillsdale, NJ.

Salthouse, T.A., 1996. The processing-speed theory of adult age differences in cognition. Psychological Review 103, 403–428.

Schaie, W.K., Willis, S.L., 1993. Age difference patterns of psychometric intelligence in adulthood: generalizability within and across ability domains. Psychology and Aging 8, 44–55.

Schneider, et al., 1996. Handbook of the Biology of Aging, fourth ed. Academic Press, New York.

Shepard, R.N., 1995. Mental universals: toward a 21st century science of mind. In: Solso, R.L., Massaro, D.W. (Eds.), The Science of the Mind: 2001 and Beyond. Oxford University Press, New York, pp. 50–64.

Spearman, C.E., 1927. The Abilities of Man. MacMillan, New York.

Stern, P.C., Carstensen, L.L., 2000. The Aging Mind: Opportunities in Cognitive Research. National Academy Press, Washington, DC.

Welford, A.T., 1965. Performance, biological mechanisms and age: a theoretical sketch. In: Welford, A.T., Birren, J.E. (Eds.), Behavioral, Aging and the Nervous System. Thomas, Springfield, IL, pp. 3–20.

West, R.L., 1996. An application of prefrontal cortex function theory to cognitive aging. Psychological Bulletin 120, 272–292.

WHO (Ed.), 2011. Global Health and Aging. NIH Publication No. 11-7737. http://who.int/ageing/publications/global_health/en/.

Zacks, R.T., Hasher, L., Li, K.Z.H., 2000. Human memory. In: Craik, F.I.M., Salthouse, T.A. (Eds.), The Handbook of Aging and Cognition. LEA, Mahwah, NJ, pp. 293–357.

Aging, Theories of

K Warner Schaie, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by K.W. Schaie, volume 1, pp. 317–322, © 2001, Elsevier Ltd.

Abstract

Theories of aging described include Erikson's stage theory, Baltes' selection, optimization, and compensation theory, Schaie and Willis' stage theory of cognition, as well as the coconstructive perspective. Current views on successful, normal, and pathological aging are also described.

Introduction

There have been few comprehensive theories of psychological development that have fully covered the period of adulthood (Schaie, 2013; Schaie and Willis, 1999). The broadest approaches have been those of Erik Erikson's stage theory (1982; Erikson et al., 1986), and the dialectical life span theory of Paul Baltes (1993). Erikson's stage theory has also been expanded for the period of adulthood and old age by Schaie and Willis (1999, 2000; Willis and Schaie, 2006).

Baltes' selection, optimization, and compensation (SOC) theory argues that gains and losses occur at every life stage, but that in old age losses far exceed the gains. Baltes considers evolutionary development to be incomplete for the very last stage of life, during which societal supports no longer fully compensate for declines in physiological infrastructure and losses in behavioral functionality (see Baltes, 1987; Baltes and Smith, 1999; Baltes et al., 1999). SOC, however, can also be seen as strategies of life management, and thus may be indicators of successful aging (Baltes and Freund, 2003). For a fuller exposition of SOC theory and review of relevant empirical studies, see Riediger et al. (2006).

The SOC theory has recently been expanded to a coconstructionist biosocial theory (Baltes and Smith, 2004; Willis and Schaie, 2006; see below). Theoretical models limited to the domain of cognition have also been proposed by Schaie and Willis (1977–78; 2000; Willis and Schaie, 2006), and by Sternberg (1985). I will here describe more fully, as examples, the Eriksonian, and the Schaie and Willis stage theories, as well as the more recent coconstructive theory.

Finally, I will then briefly discuss the distinction between successful, normal, and pathological aging, which is currently exerting a major influence on theories and research in life span developmental psychology.

Erikson's Stage Model

Traditional psychodynamic treatments of the life span have been restricted primarily to the development of both normal and abnormal personality characteristics. With the exception of some ego psychologists (e.g., Loevinger, 1976), however, Erik Erikson remains the primary theorist coming from a psychoanalytic background who has consistently pursued a life span approach. Although Erikson's most famous concept, the identity crisis, is placed in adolescence, the turmoil of deciding 'who you are' continues in adulthood, and identity crises often recur throughout life, even in old age (Erikson, 1979). Moreover, Erikson (1982) takes the position that "human development is dominated by dramatic shifts in emphasis."

In his latest writing, Erikson redistributed the emphasis on the various life stages more equitably. He argued that the question of greatest priority in the study of ego development is "how, on the basis of a unique life cycle and a unique complex of psychosocial dynamics, each individual struggles to reconcile earlier themes in order to bring into balance a lifelong sense of trustworthy wholeness and an opposing sense of bleak fragmentation" (Davidson, 1995; Erikson et al., 1986; Goleman, 1988).

The *intimacy crisis* is the primary psychosocial issue in the young adult's thoughts and feelings about marriage and family. However, recent writers suggest that this crisis must be preceded by identity consolidation, which is also thought to occur in young adulthood (cf Pals, 1999).

The primary issue of middle age, according to Erikson, is *generativity versus stagnation* (see McAdams and de St. Aubin, 1998; Snarey et al., 1987). Broadly conceived, *generativity* includes the education of one's children, productivity and creativity in one's work, and a continuing revitalization of one's spirit that allows for fresh and active participation in all facets of life. Manifestations of the generativity crisis in midlife are career problems, marital difficulties, and widely scattered attempts at 'self-improvement.'

Successful resolution of the generativity crisis involves the human virtues of caring, giving, and teaching in the home, on the job, and in life in general. In Erikson's view of ego development, the final years of life mark the time of the *integrity versus despair crisis*, when individuals look back over their lives (Haight et al., 1994) and decide that they were well ordered and meaningful (integrated) or unproductive and meaningless (resulting in despair).

Those who despair approach the end of life with the feeling that death will be one more frustration in a series of failures. In contrast, the people with integrity accept their lives (including their deaths) as important and on the whole satisfying. In a sense, ego integrity is the end result of the life-long search for ego identity, a recognition that one has coped reasonably successfully with the demands of both the id and society (Erikson, 1979, 1982; Whitbourne, 1996). Once old age is reached, it may be most advantageous for the person to rigidly maintain this identity (Tucker and Desmond, 1998).

The final stage of life includes an exploration of personal grounds for faith. Erikson points out that the aged share with infants what he calls the 'numinous' or the experience of the 'ultimate other.' For the infant, this particular experience was provided by its mother. By contrast, the experience of ultimate confidence is provided for the older person by the confirmation of the distinctiveness of their integrated life and by its impending transcendence (Erikson, 1984).

A formal investigation of the progression through the Eriksonian stages from young adulthood into midlife has been conducted by administering an inventory of psychosocial development to three cohorts of college students, followed up after 11 and 22 years (Whitbourne et al., 1992). This study showed not only inner psychological changes as postulated by Erikson, but also effects of exposure to particular historical, cultural, and social realities of the environment.

As attainment of higher stages was observed, there also appeared to be further resolution of the earlier stages of development, suggesting a process of continuous reorganization, beyond the stage-specific issues confronted by a given individual. The latter study also raises the possibility that the sequencing of Erikson's stages may not be unidirectional, and it further suggests cohort differences that imply less-favorable resolution of the ego integrity versus despair crisis during the decade of the 1980s (Whitbourne and Connolly, 1999).

Schaie and Willis' Stage Theory of Cognition

Schaie and Willis (1999, 2000) use findings from their research on adult intellectual development to formulate seven adult stages. They argue that while Piaget's childhood stages describe increasing efficiency in the acquisition of new information, it is quite doubtful that adults progress beyond the powerful methods of science (formal operations) in their quest for knowledge. Therefore, if one is to propose adult stages, they should not be further stages of acquisition; but, instead, such stages should reflect different uses of intellect.

In young adulthood, for example, people typically switch their focus from the acquisition to the application of knowledge, as they use what they know to pursue careers and develop their families. This is called the *achieving* stage. It represents most prominently the application of intelligence in situations that have profound consequences for achieving long-term goals. The kind of intelligence exhibited in such situations is similar to that employed in educational tasks, but it requires careful attention to the possible consequences of the problem-solving process.

Young adults who have mastered the cognitive skills required for monitoring their own behavior and, as a consequence, have attained a certain degree of personal independence, which will next move into a stage that requires the application of cognitive skills in situations involving social responsibility. Typically, the *responsible* stage occurs when a family is established and the needs of spouse and offspring must be met. Similar extensions of adult cognitive skills are required as responsibilities for others are acquired on the job and in the community.

Some individuals' responsibilities become exceedingly complex. Such individuals – presidents of business firms, deans of academic institutions, officials of churches, and a number of other positions – need to understand the structure and the dynamic forces of organizations. They must monitor organizational activities not only on a temporal dimension (past, present, and future), but also up and down the hierarchy that defines the organization. They need to know not only the future plans of the organization, but also whether policy decisions are being adequately translated into action at lower levels of responsibility. Attainment of the *executive* stage, as a variation on the responsibility stage, depends on exposure to opportunities that allow the development and practice of the relevant skills (Avolio, 1991; Smith et al., 1994).

In the later years of life, beyond the age of 60 or 65, the need to acquire knowledge declines even more, and executive monitoring is less important because frequently the individual has retired from the position that required such an application of intelligence. This stage, *reintegration*, corresponds in its position in the life course to Erikson's stage of ego integrity. The information that elderly people acquire and the knowledge they apply becomes a function of their interests, attitudes, and values. It requires, in fact, the *reintegration* of all of these. The elderly are less likely to 'waste time' on tasks that are meaningless to them. They are unlikely to expend much effort to solve a problem unless that problem is one that they face frequently in their lives. This stage frequently includes a selective reduction of interpersonal networks in the interest of reintegrating one's concern in a more self-directed and supportive manner (cf Carstensen, 1993; Carstensen et al., 1997).

In addition, efforts must be directed toward planning how one's resources will last for the remaining 15–30 years of postretirement life that are now characteristic for most individuals in industrialized societies. These efforts include active planning for that time when dependence upon others may be required to maintain a high quality of life in the face of increasing frailty. Such efforts may involve changes in one's housing arrangements, or even one's place of residence, as well as making certain of the eventual availability of both familial and extrafamilial support systems. The activities involved in this context include making or changing one's will, drawing up advanced medical directives and durable powers of attorney, as well as creating trusts or other financial arrangements that will protect resources for use during the final years of life or for the needs of other family members.

Although some of these activities involve the same cognitive characteristics of the responsible stage, these objectives involved are far more centered upon current and future needs of the individual rather than the needs of their family or of an organizational entity. Efforts must now be initiated to reorganize one's time and resources to substitute a meaningful environment, often found in leisure activities, volunteerism, and involvement with a larger kinship network. Eventually, however, activities are also engaged in to maximize quality of life during the final years, often with the additional objective of not becoming a burden for the next generation. The unique objective of these demands upon the individual represent an almost universal process occurring at least in the industrialized societies, and

designation of a separate *reorganizational* stage is therefore warranted.

The skills required for the reorganizational stage require the maintenance of reasonably high levels of cognitive competence. In addition, maintenance of flexible cognitive styles are needed to be able to restructure the context and content of life after retirement, to relinquish control of resources to others and to accept the partial surrender of one's independence (Schaie, 1984, 2005).

Many older persons reach advanced-old age in relative comfort and often with a clear mind albeit a frail body. Once the reintegrative efforts described above have been successfully completed, yet one other stage is frequently observed. This last stage is concerned with cognitive activities by many of the very old that occur in anticipation of the end of their life. This is a *legacy-creating* stage that is part of the cognitive development of many, if not all, older persons. This stage often begins by the self- or therapist-induced effort to conduct a life review (Butler et al., 1998). For the highly literate and those successful in public or professional life, this will often include writing or revising an autobiography (Birren et al., 1995; Birren and Schroots, 2006).

There are also many other more mundane legacies to be left. Women, in particular, often wish to put their remaining effects in order, and often distribute many of their prized possessions to friends and relatives, or create elaborate instructions for distributing them. It is not uncommon for many very old people to make a renewed effort at providing an oral history or to explain family pictures and heirloom to the next generation. Last, but not least, directions may be given for funeral arrangements, occasionally including donation of one's body for scientific research, and there may be a final revision of one's will.

The Coconstructive Perspective

Both neurobiological and sociocultural influences on development have long been recognized. Coevolutionary theorists (Dunham, 1991; Tomasello, 1999) suggest that both biological and cultural evolution has occurred and that recent, cohort-related advances in human development in domains such as intelligence can be attributed largely to cumulative cultural evolution. Cultural activities impact the environment, thereby influencing mechanisms such as selection processes, and thus allow humans to codirect their own evolution (Cavalli-Sforza and Feldman, 1981; Dunham, 1991). Baltes et al. (1997), Li (2003), Li and Freund (2005) Baltes' and his colleagues (1997; Li, 2003; Li and Freund, 2005) coconstructionist approach imposes a life span developmental perspective on coevolutionary theory and provides principles regarding the timing of the varying contributions of neurobiology and culture at different developmental periods and across different domains of functioning. Three principles have been proposed regarding the relative contributions of biology and culture influences across the life span:

1. Beneficial effects of the evolutionary selection process occur primarily in early life and are less likely to optimize development in the later half of life.
2. Further advances in human development depend on ever-increasing cultural resources. From a historical perspective, increases in cultural resources have occurred via cumulative cultural evolution and have resulted in humans reaching higher levels of functioning. At the individual level, increasing cultural resources are required at older ages for further development to occur or to prevent age-related losses.
3. The efficacy of increasing cultural resources is diminished in old age, due to decline in neurobiological functions.

Li (2003) proposes a triarchic view of culture involving three aspects of culture that are related to the coconstructionist perspective: resource, process, and developmental relevancy. Culture as social resources involves the knowledge, values, and material artifacts accumulated by a society and transmitted to future generations; these resources continue to develop and change through cumulative cultural evolution (Tomasello, 1999). Expanding upon Li's triarchic view of cultural domains, Willis and Schaie (2006) view accumulated cultural resources as being represented by structural variables such as educational level, occupational status, and ability level. These variables reflect the individual's acquisition and accumulation of cultural knowledge and skills primarily during the first half of adulthood.

Culture as ongoing social process involves the routines, habits, and performances of the individual in daily life that take place in the individual's proximal developmental context and that are shaped by the momentarily shared social reality (Li, 2003). The third component of developmental relevancy suggests that the impact of particular cultural resources and processes on an individual is partially determined by the individual's developmental stage, which has also termed the 'developmental niche' (Gauvain, 1998; Super and Harkness, 1986). The coconstructive perspective may be particularly useful in understanding the interplay risk and protective factors that influence cognitive aging.

Successful, Normal, and Pathological Aging

It is readily apparent that there are vast individual differences in patterns of psychological changes from young adulthood through old age. Scrutiny of a variety of longitudinal studies of psychological aging (cf Schaie and Hofer, 2001) suggest that four major patterns will describe most of the observed aging trajectories, although further subtypes could, of course, be considered (Schaie, 2006). These patterns would classify individuals into those who age successfully (the supernormals), those who age normally, those who develop mild cognitive impairment, and finally those who become clinically diagnosable as suffering from dementia.

The most common pattern is what we could denote as the *normal aging* of psychological functions. This pattern is characterized by most individuals reaching an asymptote in early midlife, maintaining a plateau until the late 50s or early 60s, and then showing modest decline on most cognitive abilities through the early 80s, with more marked decline in the years prior to death (cf Bosworth et al., 1999). They also tend to become more rigid and show some changes on personality

traits in undesirable directions (Schaie et al., 2004). Among those whose cognitive aging can be described as normal, we can distinguish two subgroups. The first include those individuals who reach a relatively high level of cognitive functioning who even if they become physically frail can remain independent until close to their demise. The second group who only reach a modest asymptote in cognitive development, on the other hand, may in old age require greater support and be more likely to experience a period of institutional care.

A small subgroup of adults experience what is often described as *successful aging* (Fillit et al., 2002; Rowe and Kahn, 1987). Members of this group are often genetically and socio-economically advantaged, they tend to continue cognitive development later than most and typically reach their cognitive asymptotes in late-midlife. While they too show some very modest decline on highly speeded tasks, they are likely to maintain their overall level of cognitive functioning until shortly before their demise. They are also likely to be less neurotic and more open to experience than most of their age peers. These are the fortunate individuals whose active life expectancy comes very close to their actual life expectancy.

The third pattern, *mild cognitive impairment* (MCI; Petersen et al., 1999), includes that group of individuals who, in early old age, experience greater than normative cognitive declines. Various definitions, mostly statistical, have been advanced to assign membership to this group. Some have argued for a criterion of one *SD* of performance compared to the young adult average, while others have proposed a rating of 0.5 on a clinical dementia rating scale, where 0 is normal and 1.0 is probable dementia. Earlier on, the identification of MCI required the presence of memory loss, in particular. However, more recently, the diagnosis has been extended to decline in other cognitive abilities. There has also been controversy on the question whether individuals with the diagnosis of MCI inevitably progress to dementia, or whether this group of individuals represents a unique entity; perhaps one could denote them as the *unsuccessful aging* (cf Petersen, 2003).

The final pattern includes those individuals who in early or advanced old age are diagnosed as suffering from *dementia*. Regardless of the specific cause of the dementia, these individuals have in common dramatic impairment in cognitive functioning. However, the pattern of cognitive change, particularly in those whose diagnosis at postmortem turns out to be Alzheimer's disease, is very different from the normal aging. When followed longitudinally, at least some of these individuals show earlier decline, perhaps starting in midlife.

Psychologists and members of the health- and health-related professions in general, have often assumed that age-related cognitive losses as well as other behavioral deficits inevitably accompany the aging process. Although there is some controversy over the possibility of disease-free aging, it is still important to distinguish between biological changes that occur in many individuals with increasing age and the presence of specific physiological or psychological pathologies that may not be age-related even though they occur with greater frequency at advanced ages (cf Solomon, 1999). The fact remains that chronic disease, often associated with aging, may have their origin in genetic predispositions, and become clinically relevant over a wide age range. While cognitive decline affects significant portions of the elderly population, its symptomatology is often confused with that associated with metabolic disturbances, and/or sensory and metabolic disturbances. Indeed, there is no compelling reason to believe that selective age-related ability-declines will inevitably lead to dementia, nor that they cannot be addressed by cognitive training or psychotherapy in many individuals (cf Willis, 1996).

See also: Age, Sociology of; Aging Mind: Facets and Levels of Analysis; Alzheimer's Disease, Neural Basis of; Cognitive Aging; Coping across the Lifespan; Ecology of Aging; Human Development, Theories of; Lifespan Development, Theory of; Old Age and Centenarians: The Human 'Warranty Period'; Retirement and Encore Adulthood: The New Later Life Course.

Bibliography

Avolio, B.J., 1991. Levels of analysis. In: Schaie, K.W. (Ed.), Annual Review of Gerontology and Geriatrics, vol. 11. Springer Publishing Co, New York, pp. 239–260.

Baltes, P.B., 1987. Theoretical propositions of life-span developmental psychology: on the dynamics between growth and decline. Developmental Psychology 23, 611–626.

Baltes, P.B., 1993. The aging mind: potentials and limits. Gerontologist 33, 580–594.

Baltes, P.B., 1997. On the incomplete architecture of human ontogeny: selection, optimization and compensation as foundation of developmental theory. American Psychologist 52, 366–380.

Baltes, P.B., Freund, A.M., 2003. The intermarriage of wisdom and selective optimization with compensation: two meta-heuristics guiding the conduct of lives. In: Keyes, C.L.M., Heidt, J. (Eds.), Flourishing: Positive Psychology and the Life Well Lived. American Psychological Association, Washington, DC, pp. 249–327.

Baltes, P.B., Smith, J., 1999. Multi-level and systemic analyses of old age: theoretical and empirical evidence for a fourth age. In: Bengtson, V.L., Schaie, K.W. (Eds.), Handbook of Theories of Aging. Springer Punishing, New York, pp. 153–172.

Baltes, P.B., Smith, J., 2004. Lifespan psychology: from developmental contextualism to developmental biocultural co-constructivism. Research on Human Development 1, 123–144.

Baltes, P.B., Staudinger, U.M., Lindenberger, U., 1999. Life-span psychology: theory and application in intellectual functioning. Annual Review of Psychology 50, 471–507.

Birren, J.E., Kenyon, G.M., Ruth, J.E., Schroots, J.F., Swensson, T. (Eds.), 1995. Aging and Biography: Explorations in Adult Development. Springer Publishing Co, New York.

Birren, J.E., Schroots, J.J.F., 2006. Autobiographic memory and the narrative self over the life span. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, sixth ed. Elsevier, San Diego, CA, pp. 477–499.

Bosworth, H.B., Schaie, K.W., Willis, S.L., 1999. Cognitive and socio-demographic risk factors for mortality in the Seattle Longitudinal. Journal of Gerontology: Psychological Sciences 54B, P273–P282.

Butler, R.N., Lewis, M., Sunderland, T., 1991. Aging and Mental Health: Positive Psychosocial and Biomedical Approaches, fourth ed. Macmillan, New York.

Carstensen, L.L., 1993. Motivation for social contact across the life span: a theory of socioemotional selectivity. In: Jacobs, J. (Ed.), Nebraska Symposium on Motivation: Developmental Perspectives on Emotion. University of Nebraska Press, Lincoln, NE, pp. 209–254.

Carstensen, L.L., Gross, J.J., Fung, H.H., 1997. The social context of emotional experience. In: Schaie, K.W., Lawton, M.P. (Eds.), Annual Review of Gerontology and Geriatric, vol. 17. Springer, New York, pp. 325–352.

Cavalli-Sforza, L.L., Feldman, M.W., 1981. Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press, Princeton, NJ.

Davidson, F., 1995. On Old Age II: A Conversation with Joan Erikson at 92. Davidson Films, San Luis Obispo, CA.

Dawkins, R., 1989. The Selfish Gene, second ed. Oxford University Press, Oxford, UK.

Dunham, W.H., 1991. Co-evolution: Genes, Culture, and Human Diversity. Stanford University Press, Palo Alto, CA.

Erikson, E.H., 1979. Reflections on Dr. Borg's life cycle. In: Erikson, E.H. (Ed.), Adulthood. Norton, New York.

Erikson, E.H., 1982. The Life Cycle Completed: A Review. Norton, New York.
Erikson, E.H., 1984. Reflections on the last stage—and the first. Psychoanalytic Study of the Child 39, 155–165.
Erikson, E.H., Erikson, J.M., Kivnick, H.Q., 1986. Vital Involvement in Old Age. Norton, New York.
Fillit, H.M., Butler, R.N., O'Connell, A.W., Albert, M.S., Birren, J.E., Cotman, C.W., Greenough, W.T., Gold, P.E., Kramer, A.F., Kuller, L.H., Perls, T.T., Sahagan, B.G., Tully, T., 2002. Achieving and maintaining cognitive vitality with aging. Mayo Clinic Proceedings 7, 681–696.
Gauvain, M., 1998. Cognitive development in social and cultural context. Psychological Science 7, 188–192.
Goleman, D., June 14, 1988. Erikson, in his own old age, expands his view of life. New York Times. C1, 14.
Haight, B.K., Coleman, P., Lord, K., 1994. The linchpins of a successful life review: structure, evaluation and individuality. In: Haight, B.K., Webster, J. (Eds.), The Art and Science of Reminiscing: Theory, Research, Methods and Applications. Taylor and Francis, Washington, DC.
Li, S.-C., 2003. Biocultural orchestration of developmental plasticity across levels: the interplay of biology and culture in shaping the mind and behavior across the life span. Psychological Bulletin 129, 171–194.
Li, S.-C., Freund, A.M., 2005. Advances in lifespan psychology: a focus on biocultural and personal influences. Research in Human Development 2, 1–23.
Loevinger, J., 1976. Ego Development: Conception and Theory. Jossey-Bass, San Francisco, CA.
McAdams, D.P., de St Aubin, E.S. (Eds.), 1998. Generativity and Adult Development: How and Why We Care for the Next Generation. American Psychological Association, Washington, DC.
Pals, J.L., 1999. Is personality adaptively patterned? A controversy—Identity consolidation in early adulthood: relations with ego-resiliency, the context of marriage and personality change. Journal of Personality 67, 295–329.
Petersen, R.C., 2003. Conceptual overview. In: Petersen, R.C. (Ed.), Mild Cognitive Impairment: Aging to Alzheimer's Disease. Oxford University Press, New York, pp. 1–14.
Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., Kokmen, E., 1999. Mild cognitive impairment: clinical characterization and outcome. Archives of Neurology 5, 303–308.
Riediger, M., Li, S.-C., Lindenberger, U., 2006. Selection, optimization, and compensation as developmental mechanism of adaptive resource allocation: review and preview. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, sixth ed. Academic Press, San Diego, CA, pp. 289–314.
Rowe, J.W., Kahn, R.L., 1987. Human aging: usual and successful. Science 237, 143–149.
Schaie, K.W., 1977/78. Toward a stage theory of adult development. International Journal of Aging and Human Development 8, 129–138.
Schaie, K.W., 1984. Midlife influences upon intellectual functioning in old age. International Journal of Behavioral Development 7, 463–478.
Schaie, K.W., 2005. Developmental influences on adult intelligence: The Seattle Longitudinal Study. Oxford University Press, New York.
Schaie, K.W., 2006. Societal influences on cognition in historical context. In: Schaie, K.W., Carstensen, L.L. (Eds.), Social Structures, Aging and Self-regulation in the Elderly. Springer Publishing Co, New York, pp. 13–32.
Schaie, K.W., 2008. A lifespan developmental perspective of psychological aging. In: Laidlaw, K., Knight, B.G. (Eds.), The Handbook of Emotional Disorders in Late Life: Assessment and Treatment. Oxford University Press, Oxford, UK, pp. 3–32.
Schaie, K.W., 2013. Developmental Influences on Adult Intelligence: The Seattle Longitudinal Study, second ed. Oxford University Press, New York.
Schaie, K.W., Hofer, S.M., 2001. Longitudinal studies in research on aging. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, fifth ed. Academic Press, San Diego, CA, pp. 55–77.
Schaie, K.W., Willis, S.L., 1999. Theories of everyday competence. In: Bengtson, V.L., Schaie, K.W. (Eds.), Handbook of Theories of Aging. Springer Publishing Co, New York, pp. 174–195.
Schaie, K.W., Willis, S.L., 2000. A stage theory model of adult cognitive development revisited. In: Rubinstein, R., Moss, M., Kleban, M. (Eds.), The Many Dimensions of Aging: Essays in Honor of M. Powell Lawton. Springer Publishing Co, New York, pp. 175–193.
Schaie, K.W., Willis, S.L., Caskie, G.I.L., 2004. The Seattle longitudinal study: relation between personality and cognition. Aging, Neuropsychology and Cognition 11, 304–324.
Smith, J., Staudinger, U.M., Baltes, P.B., 1994. Occupational settings facilitating wisdom-related knowledge. Journal of Consulting and Clinical Psychology 62, 1989–1999.
Snarey, J., Son, L., Kuehne, V.S., Hauser, S., Vaillant, G., 1987. The role of parenting in men's psychosocial development: a longitudinal study of early adulthood infertility and midlife generativity. Developmental Psychology 21, 568–584.
Solomon, D.H., 1999. The role of aging processes in aging-dependent diseases. In: Bengtson, V.L., Schaie, K.W. (Eds.), Handbook of Theories of Aging. Springer Publishing Co, New York, pp. 133–150.
Sternberg, R.J., 1985. Beyond IQ: A Triarchic Theory of Human Intelligence. Cambridge University Press, New York.
Super, C.M., Harkness, S., 1986. The developmental niche: a conceptualization at the interface of child and culture. International Journal of Behavioral Development 9, 545–569.
Tomasello, M., 1999. The Cultural Origins of Human Cognition. Harvard University Press, Cambridge, MA.
Tucker, D.M., Desmond Jr., R.E., 1998. Aging and the plasticity of the self. In: Schaie, K.W., Lawton, M.P. (Eds.), Annual Review of Gerontology and Geriatrics, vol. 17. Springer Publishing Co, New York, pp. 266–281.
Whitbourne, S.K., 1996. Psychosocial perspectives on emotion: the role of identity in the aging process. In: Magai, C., McFadden, S.H. (Eds.), Handbook of Emotion, Adult Development, and Aging. Academic Press, San Diego, CA, pp. 83–98.
Whitbourne, S.K., Connolly, L.S., 1999. The developing self in mid-life. In: Willis, S.L., Reid, J.D. (Eds.), Life in the Middle: Psychological and Social Development in Middle Age. Academic Press, San Diego, CA, pp. 25–45.
Whitbourne, S.K., Zuschlag, M.K., Elliot, L.B., Waterman, A.S., 1992. Psychosocial development in adulthood: A 22-year sequential study. Journal of Personality and Social Psychology 63, 260–271.
Willis, S.L., 1996. Methodological issues in behavioral intervention research with the elderly. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, fifth ed. Academic Press, San Diego, CA, pp. 78–108.
Willis, S.L., Schaie, K.W., 2006. A co-constructionist view of the third age: the case of cognition. Annual Review of Gerontology and Geriatrics 26, 131–152.

Agnosia

Daniel Tranel and Antonio R Damasio, University of Iowa, Iowa City, IA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 1, pp. 322–326, © 2001, Elsevier Ltd., with updates by the Editor in Chief.

Abstract

Agnosia is a fascinating condition in which, as a consequence of acquired brain damage, patients lose the ability to recognize familiar stimuli, despite normal perception of those stimuli. For example, when encountering the faces of familiar persons such as family members or close friends, a patient with agnosia is unable to identify those persons, or even to recognize that they are familiar. Agnosia can affect recognition of stimuli in any sensory modality, including visual, auditory, and tactile, although visual agnosia is the most common form of the disorder. Careful scientific study of agnosia has provided many important new insights into the manner in which the human brain acquires, maintains, and utilizes various types of knowledge.

Agnosia is a fascinating condition in which, as a consequence of acquired brain damage, patients lose the ability to recognize familiar stimuli, despite normal perception of those stimuli. For example, when encountering the faces of familiar persons such as family members or close friends, a patient with agnosia is unable to identify those persons, or even to recognize that they are familiar. A patient may look at pictures of entities such as animals or tools, and have no idea what the stimuli are. Or a patient may hear well-known sounds, such as a fire siren or a ringing phone, and not be able to identify the sounds or understand their meaning (despite being able to hear the sounds normally). Agnosia is a rare condition, and its clinical presentation borders on the bizarre; nonetheless, careful scientific study of agnosia has provided many important insights into the brain mechanisms important for learning, memory and knowledge retrieval.

Types of Knowledge and Levels of Knowledge Retrieval

Before discussing agnosia, it is important to explain some crucial differences in the types of knowledge that are processed by the brain, and how different task demands influence the mechanisms the brain uses to retrieve knowledge. To begin with, there is a dimension of specificity: knowledge can be retrieved at different levels of specificity, ranging from very specific to very general. Consider the following example: Knowledge about a unique horse ('Little Buck,' a sorrel roping horse) is specific and unique, and is classified at the subordinate level; less specific knowledge about horses (four-legged animals that gallop, used by cowboys; of which Little Buck is an example) is classified at the basic object level; and even less specific knowledge about living things (things that have life, of which horses and Little Buck are examples) is classified at the superordinate level.

Pragmatically, the level at which knowledge is retrieved depends on the demands of the situation, and those demands are different for different categories of entities. In everyday life, for example, it is mandatory that familiar persons be recognized at the unique level – e.g., that's 'President Clinton,' or that's 'my father Ned.' It is not sufficient, under most conditions, to recognize such entities only at more nonspecific levels – e.g., that's a 'world leader,' or that's 'an older man.' For other types of entities, recognition at the basic object level is sufficient for most purposes – e.g., that's a 'screwdriver,' or that's a 'stapler'; here, there is no need to recognize individual, unique screwdrivers and staplers in order for practical interactions with the entity to be productive.

One other critical distinction is between recognition, on the one hand, and naming, on the other. The two capacities are often confused. It is true that recognition of an entity, under normal circumstances, is frequently indicated by naming (e.g., 'stapler'; 'Little Buck'; 'siren'). However, there is a basic difference between knowing and retrieving the meaning of a concept (its functions, features, characteristics, relationships to other concepts), and knowing and retrieving the name of that concept (what it is called); moreover, this difference is honored by the brain. For example, brain damage in the left inferotemporal region can render a patient incapable of naming a wide variety of stimuli, while leaving unaffected the patient's ability to recognize those stimuli (Damasio et al., 1996). For the examples of 'Little Buck' and 'siren' cited above, the patient may produce the descriptions of 'that's my sorrel roping horse that I bought 2 years ago and now lives on my dad's ranch,' and 'that's a loud sound that means there's an emergency; you should pull your car over to the side of the road.' Both responses indicate unequivocal recognition of the specific entities, even if their names are never produced. In short, it is important to maintain a distinction between recognition, which can be indicated by responses signifying that the patient understands the meaning of a particular stimulus, and naming, which may not, and need not, accompany accurate recognition (Caramazza and Shelton, 1998; Gainotti et al., 1995; Pulvermuller, 1999).

The Term 'Agnosia'

The term 'agnosia' signifies 'lack of knowledge,' and denotes an impairment of recognition. Traditionally, two types of

agnosia have been described (Lissauer, 1890). One, termed associative agnosia, refers to a failure of recognition that results from defective retrieval of knowledge pertinent to a given stimulus. Here, the problem is centered on memory: the patient is unable to recognize a stimulus (i.e., to know its meaning) despite being able to perceive the stimulus normally (e.g., to see shape, color, texture; to hear frequency, pitch, timbre; and so forth). The other type of agnosia is termed apperceptive, and refers to a disturbance of the integration of otherwise normally perceived components of a stimulus. Here, the problem is centered more on perception: the patient fails to recognize a stimulus because the patient cannot integrate the perceptual elements of the stimulus, even though those individual elements are perceived normally. It is important to emphasize that the nuclear feature in designating a condition as 'agnosia' is that there is a recognition defect that cannot be attributed simply or entirely to faulty perception.

The terms associative and apperceptive agnosia have remained useful, even if the two conditions do have some overlap. It is usually possible to classify a patient with a recognition impairment as having primarily a disturbance of memory (associative agnosia), or primarily a disturbance of perception (apperceptive agnosia). Not only does this classification have important implications for the management of such patients (e.g., what rehabilitation should be applied), but it also maps on to different sites of neural dysfunction. For example, in the visual modality, associative agnosia is strongly associated with bilateral damage to higher-order association cortices in the ventral and mesial occipito-temporal regions, whereas apperceptive agnosia is associated with unilateral or bilateral damage to 'earlier,' more primary visual cortices. This being said, though, the fact remains that separating associative and apperceptive agnosia can be difficult, which underscores the fact that the processes of perception and memory are not discrete. Rather, they operate on a physiological and psychological continuum, and it is simply not possible to demarcate a specific point at which perceptual processes end and memory processes begin (Damasio et al., 1990; Tranel and Damasio, 1996).

In principle, agnosia can occur in any sensory modality, relative to any type of entity or event. In practice, however, some types of agnosia are considerably more frequent. Visual agnosia, especially agnosia for faces (prosopagnosia), is the most commonly encountered form of recognition disturbance. The condition of auditory agnosia is rarer, followed by the even less frequent tactile agnosia.

Visual Agnosia

Definition

Visual agnosia is defined as a disorder of recognition confined to the visual realm, in which a patient cannot arrive at the meaning of some or all categories of previously known nonverbal visual stimuli, despite normal or near-normal visual perception and intact alertness, attention, intelligence, and language. Typically, patients have impairments both for stimuli that they learned prior to the onset of brain injury (known as 'retrograde' memory), and for stimuli that they would normally have learned after their brain damage (known as 'anterograde' memory).

Subtypes

Prosopagnosia

The study of face processing has remained a popular topic in neuropsychology for many decades, dating back to the pioneering work of Bodamer, Hecaen, Meadows, and others (for historical reviews, see Benton, 1990; De Renzi, 1997). Faces are an intriguing class of stimuli (Damasio et al., 1982; Young and Bruce, 1991). They are numerous and visually similar, and yet we learn to recognize individually as many as thousands of distinct faces during our lifetime; and not only can we learn many individual faces, but we can recognize them from obscure angles (e.g., from the side), attended with various artifacts (e.g., glasses, hockey helmet), after aging has radically altered the physiognomy, and under many other highly demanding conditions. Also, faces convey important and unique social and emotional information, providing clues about the emotional state of a person, or about potential courses of social behavior (e.g., approach or avoidance) (see Darwin, 1955 [1872]/1955 [1872]; Adolphs et al., 1998). And there are a number of remarkable cross-cultural and cross-species consistencies in face processing (cf Ekman, 1973; Fridlund, 1994), which underscore the crucial and fundamental importance of this class of stimuli.

The inability to recognize familiar faces is known as prosopagnosia (face agnosia), and it is the most frequent and well established of the visual agnosias (Damasio et al., 1990; Farah, 1990). The face recognition defect in prosopagnosia typically covers both the retrograde and anterograde compartments; respectively, patients can no longer recognize the faces of previously known individuals, and are unable to learn new ones. They are unable to recognize the faces of family members, close friends, and, in the most prototypical instances, even their own face in a mirror. Upon seeing those faces, the patients experience no sense of familiarity, no inkling that those faces are known to them, i.e., they fail to conjure up consciously any pertinent information that would constitute recognition. The impairment is modality-specific, however, being entirely confined to vision. For example, when a prosopagnosic patient hears the voices of persons whose faces were unrecognized, the patient will instantly be able to identify those persons accurately.

As noted above with regard to agnosia in general, prosopagnosia must be distinguished from disorders of naming, i.e., it is not an inability to name faces of persons who are otherwise recognized as familiar. There are numerous examples of face naming failure, from both brain-injured populations and from the realm of normal everyday experience, but in such instances, the unnamed face is invariably detected as familiar, and the precise identity of the possessor of the face is usually apprehended accurately. Consider, for example, the following common type of naming failure: you encounter someone whom you recently met, and cannot remember that person's name: you can remember when and where you met the person, who introduced you, and what the person does for a living – in short, you recognize the person normally. In prosopagnosia, the defect sets in at the level of recognition.

The recognition impairment in prosopagnosia occurs at the most subordinate level, i.e., at the level of specific identification of unique faces. Prosopagnosics are fully capable of recognizing faces as faces, i.e., performance is normal at more superordinate, nonspecific levels. Also, most prosopagnosics can recognize facial emotional expressions (e.g., happy, angry), and can make accurate determinations of gender and age based on face information (Humphreys et al., 1993; Tranel et al., 1988). These dissociations highlight several intriguing separations in the neural systems dedicated to processing different types of conceptual knowledge, such as knowledge about the meaning of stimuli, knowledge about emotion, and so on. In fact, these neural systems can be damaged in reverse fashion: for example, bilateral damage to the amygdala produces an impairment in recognizing facial emotional expressions, but spares the ability to recognize facial identity (Adolphs et al., 1995).

Although the problem with faces is usually the most striking, it turns out that the recognition defect in prosopagnosia is often not confined to faces. Careful assessment often reveals that the patient cannot recognize other visual entities at the normal level of specificity. The key determinants of whether other categories of stimuli are affected are (1) whether those stimuli are relatively numerous and visually similar, and (2) whether the demands of the situation call for specific identification. Whenever these conditions exist, prosopagnosics will tend to manifest deficits. For example, patients may not be able to identify a unique car, or a unique house, or a unique horse, even if they are able to recognize such entities at the basic object level, e.g., cars as cars, houses as houses, horses as horses. Similar to the problem with faces, they are unable to recognize the specific identity of a particular car, or house. These impairments underscore the notion that the core defect in prosopagnosia is the inability to disambiguate individual visual stimuli. In fact, cases have been reported in which the most troubling problem for the patient was in classes of visual stimuli other than human faces! For example, there was a farmer who lost his ability to recognize individual dairy (e.g., Holstein) cows, and a bird-watcher who became unable to tell apart various subtypes of birds (Assal et al., 1984; Bornstein et al., 1969).

Patients with face agnosia can usually recognize identity from movement. For example, upon seeing a distinctive gait of a familiar person, the patient can identify that person accurately, despite not knowing that person's face. This means not only that their perception of movement is intact, but also that they can evoke appropriate memories from the perception of unique patterns of movement. Conversely, patients with lesions in superior occipitoparietal regions (whose recognition of identity from form is normal, and hence do not have impaired face recognition) have defective motion perception and recognition. These findings underscore the separable functions of the 'dorsal' and 'ventral' visual systems, the dorsal one being specialized for spatial placement, movement, and other 'where' capacities; and the ventral one being specialized for form detection, shape recognition, and other 'what' capacities (Ungerleider and Mishkin, 1982). In prosopagnosia, the dysfunction is in the 'what' system.

One of the most intriguing findings to emerge in this area of research is that despite an inability to recognize familiar faces consciously, prosopagnosic patients often have accurate nonconscious (or covert) discrimination of those faces. This phenomenon has been studied using a psychophysiological index (the skin conductance response [SCR]) to measure nonconscious discrimination (Tranel and Damasio, 1985). SCRs were recorded while prosopagnosic patients viewed a series of face stimuli. The stimulus sets included faces that were well known to the patients, mixed in random order with faces the patients had never seen before. While viewing the faces, the patients produced significantly larger SCRs to familiar faces, compared to unfamiliar ones. This occurred in several experiments, using different types of familiar faces: in one, the familiar faces were family members and friends, in another, the familiar faces were famous individuals (movie stars, politicians), and in yet another, the familiar faces were persons to whom the patients had had considerable exposure after the onset of their condition, but not before. In sum, the patients showed nonconscious discrimination of facial stimuli they could not otherwise recognize, and for which even a remote sense of familiarity was lacking. These findings suggest that some part of the physiological process of face recognition remains intact in the patients, although the results of this process are unavailable to consciousness. The fact that the patients were able to show this type of discrimination for faces to which they had been exposed only after the onset of their condition is particularly intriguing, as it suggests that the neural operations responsible for the formation and maintenance of new 'face records' can proceed independently from conscious influence.

Category-Specific Visual Agnosia

Agnosia can develop for categories of stimuli other than faces, at levels above the subordinate, for example, at basic object level. For instance, patients may lose the ability to recognize animals or tools. This is generally referred to as visual object agnosia. The condition rarely affects all types of stimuli with equal magnitude (Farah and McClelland, 1991; Forde and Humphreys, 1999; Tranel et al., 1997; Warrington and Shallice, 1984). In one common profile of visual object agnosia, there is a major defect in categories of living things, especially animals, with relative or even complete sparing of categories of artifactual entities (e.g., tools and utensils). Less commonly, the profile is reversed, in that the patient cannot recognize tools/utensils but performs normally for animals (Tranel et al., 1997; Warrington and McCarthy, 1994). It has been shown that lesions in the right mesial occipital/ventral temporal region, and in the left mesial occipital region, are associated with defective recognition of animals; whereas lesions in the left occipital-temporal-parietal junction are associated with defective recognition of tools/utensils (Tranel et al. 1997).

Concluding Comment

Despite their relative rarity, agnosias have proved to be important 'experiments of nature,' and they have assisted with the investigation of the neural basis of human perception, learning, and memory. Careful study of agnosic patients over many decades, facilitated by the advent of modern neuroimaging techniques (computed tomography, magnetic resonance) and by the development of sophisticated experimental

neuropsychological procedures, has yielded important new insights into the manner in which the human brain acquires, maintains, and uses various of knowledge.

Editor's Note

Research into the disorders of perception following brain damage has advanced considerably since this article was first published. Current research focuses on the contribution of imaging methods to the precise definition of structural and functional changes in acquired and congenital visual agnosia; the specificity of the disorder of face recognition in prosopoagnosia; and the role of subtle deficits of primary visual processing; and the characteristics of agnosia in the tactile and auditory modality. Readers seeking entry to current literature on these topics will find the following useful: Avidan et al. (2014), Bridge et al. (2013), Cavina-Pratesi et al. (2010), Konen et al. (2011), Rezlescu et al. (2014), Saygin et al. (2010), Serino et al. (2014), Veronelli et al. (2014).

See also: Amnesia: General; Face Recognition Models: Computational Approaches; Face Recognition, Psychological and Neural Aspects; Neural Representations of Objects; Prosopagnosia.

Bibliography

Adolphs, R., Tranel, D., Damasio, H., Damasio, A.R., 1995. Fear and the human amygdala. Journal of Neuroscience 15, 5879–5891.

Adolphs, R., Tranel, D., Damasio, A.R., 1998. The human amygdala in social judgment. Nature 393, 470–474.

Assal, G., Favre, C., Anderes, J., 1984. Nonrecognition of familiar animals by a farmer. Zooagnosia or prosopagnosia for animals. Revue Neurologique 140, 580–584.

Avidan, G., Tanzer, M., Hadj-Bouziane, F., Liu, N., Ungerleider, L.G., Behrmann, M., 2014. Selective dissociation between core and extended regions of the face processing network in congenital prosopagnosia. Cerebral Cortex 24, 1565–1578.

Benton, A., 1990. Facial recognition. Cortex 26, 491–499.

Bornstein, B., Sroka, H., Munitz, H., 1969. Prosopagnosia with animal face agnosia. Cortex 5, 164–169.

Bridge, H., Thomas, O.M., Minini, L., Cavina-Pratesi, C., Milner, A.D., Parker, A.J., 2013. Structural and functional changes across the visual cortex of a patient with visual form agnosia. The Journal of Neuroscience 33, 12779–12791.

Caramazza, A., Shelton, J.R., 1998. Domain-specific knowledge systems in the brain: The animate-inanimate distinction. Journal of Cognitive Neuroscience 10, 1–34.

Cavina-Pratesi, C., Kentridge, R., Heywood, C., Milner, A., 2010. Separate channels for processing form, texture, and color: evidence from fMRI adaptation and visual object agnosia. Cerebral Cortex 20, 2319–2332.

Damasio, A.R., Damasio, H., Van Hossen, G.W., 1982. Prosopagnosia: Anatomic basis and behavioral mechanisms. Neurology 32, 331–341.

Damasio, A.R., Tranel, D., Damasio, H., 1990. Face agnosia and the neural substrates of memory. Annual Review of Neuroscience 13, 89–109.

Damasio, H., Grabowski, T.J., Tranel, D., Hichwa, R.D., Damasio, A.R., 1996. A neural basis for lexical retrieval. Nature 380, 499–505.

Darwin, C., 1955 [1872]. The Expression of the Emotions in Man and Animals. Philosophical Library, New York.

De Renzi, E., 1997. Prosopagnosia. In: Feinberg, T.E., Farah, M.J. (Eds.), Behavioral neurology and neuropsychology. McGraw-Hill, New York, pp. 254–255.

Ekman, P., 1973. Darwin and Facial Expression: A Century of Research in Review. Academic Press, New York.

Farah, M.J., McClelland, J.L., 1991. A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. Journal of Experimental Psychology 120, 339–357.

Farah, M.J., 1990. Visual agnosia. The MIT Press, Cambridge, MA.

Forde, E.M.E., Humphreys, G.W., 1999. Category-specific recognition impairments: A review of important case studies and influential theories. Aphasiology 13, 169–193.

Fridlund, A.J., 1994. Human Facial Expression: An Evolutionary View. Academic Press, New York.

Gainotti, G., Silveri, M.C., Daniele, A., Giustolisi, L., 1995. Neuroanatomical correlates of category-specific semantic disorders: A critical survey. Memory 3, 247–264.

Humphreys, G.W., Donnelly, N., Riddoch, M.J., 1993. Expression is computed separately from facial identity, and it is computed separately for moving and static faces: Neuropsychological evidence. Neuropsychologia 31, 173–181.

Konen, C.S., Behrmann, M., Nishimura, M., Kastner, S., 2011. The functional neuroanatomy of object agnosia: a case study. Neuron 71, 49–60.

Lissauer, H., 1890. Ein Fall von Seelenblindheit nebst einem Beitrage zur Theorie derselben. Archiv für Psychiatrie und Nervenkrankherten 21, 222–270.

Pulvermuller, F., 1999. Words in the brain's language. Behavioral and Brain Sciences 22, 253–336.

Rezlescu, C., Barton, J.J., Pitcher, D., Duchaine, B., 2014. Normal acquisition of expertise with greebles in two cases of acquired prosopagnosia. Proceedings of the National Academy of Sciences 111, 5123–5128.

Saygin, A.P., Leech, R., Dick, F., 2010. Nonverbal auditory agnosia with lesion to Wernicke's area. Neuropsychologia 48, 107–113.

Serino, A., Cecere, R., Dundon, N., Bertini, C., Sanchez-Castaneda, C., Ladavas, E., 2014. When apperceptive agnosia is explained by a deficit of primary visual processing. Cortex 52, 12–27.

Tranel, D., Damasio, A.R., 1985. Knowledge without awareness: An autonomic index of facial recognition by prosopagnosics. Science 228, 1453–1454.

Tranel, D., Damasio, A.R., 1996. The agnosias and apraxias. In: Bradley, W.G., Daroff, R.B., Fenichel, G.M., Marsden, C.D. (Eds.), Neurology in Clinical Practice, 2nd edn. Butterworth, Stoneham, MA, pp. 119–129.

Tranel, D., Damasio, A.R., Damasio, H., 1988. Intact recognition of facial expression, gender, and age in patients with impaired recognition of face identity. Neurology 38, 690–696.

Tranel, D., Damasio, H., Damasio, A.R., 1997. A neural basis for the retrieval of conceptual knowledge. Neuropsychologia 35, 1319–1327.

Ungerleider, L.G., Mishkin, M., 1982. Two cortical visual systems. In: Ingle, D.J., Goodale, M.A., Mansfield, R.J.W. (Eds.), Analysis of Visual Behavior. MIT Press, Cambridge, MA, pp. 549–586.

Veronelli, L., Ginex, V., Dinacci, D., Cappa, S.F., Corbo, M., 2014. Pure associative tactile agnosia for the left hand: clinical and anatomo-functional correlations. Cortex 58C, 206–216.

Warrington, E.K., McCarthy, R.A., 1994. Multiple meaning systems in the brain: A case for visual semantics. Neuropsychologia 32, 1465–1473.

Warrington, E.K., Shallice, T., 1984. Category specific semantic impairments. Brain 107, 829–853.

Young, A.W., Bruce, V., 1991. Perceptual categories and the computation of 'grandmother. European Journal of Cognitive Psychology 3, 5–49.

Agonism

Kevin W Ryan, National University of Ireland, Galway, Ireland

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E.S. Barratt, volume 1, pp. 326–329, © 2001, Elsevier Ltd.

Abstract

As a way of thinking about politics and 'the political,' the concept of agonism emphasizes struggle, contest, and plurality. Originating in the *agōn* of ancient Greece, it has recently come to characterize a distinct branch of democratic theory. Combining insights derived from Nietzsche and Arendt with a poststructuralist ontology, contemporary agonists have staged a critical debate with normative theories that prioritize rational deliberation and consensus as the means and ends of democratic politics. This article surveys the ancient *agōn* as well as contemporary forms of agonism, using the former to pinpoint three axes of contingency that combine as an agonistic politics, and which continue to focus debates in democratic political theory today.

Overview

At once very old and very new, agonism originates in the athletic and oratorical contests of ancient Greece, while more recently it has emerged as a distinct branch of political theory. This article begins with an account of the ancient *agōn*, focusing on how it began as an aristocratic practice, which was later democratized, culminating in a type of contest, which was stretched between distinction and equality, thus creating a tension within the space of the *agōn* itself. Section The *Agōn* (Un)Bound draws on the work of Hannah Arendt, who identifies a further tension in the way the *agōn* was bounded as a threshold between an 'inside' and an 'outside,' which is also the point at which agonistic contest gives way to violence. Out of these tensions emerge what might be described as three axes of contingency, which will serve the purpose of mapping the contours of contemporary agonism. Section Staging the Contest surveys the context that has seen agonism reemerge as a way of staging critical encounters with normative theories that prioritize rational deliberation and consensus as the means and ends of democratic politics. Finally, Sections The Expressive Agon: Pluralization against Pluralism and The Pragmatic Agon: Agonism against Antagonism examine two of the main currents in contemporary agonistic political thought: the 'expressive' agonism of William E. Connolly and the 'pragmatic' agonism of Chantal Mouffe (Schaap, 2009).

The *Agōn* as 'Democratic Narcissus'

Contemporary agonists have borrowed the concept from thinkers such as Hannah Arendt and Friedrich Nietzsche, but, as Andreas Kalyvas points out, they have yet to engage with the *agōn* of the ancients, which is curious, because there is much to learn from this almost-forgotten history, particularly with respect to the tenuous, which is to say, contingently articulated relationship between agonism and democracy (2009: 16). Furthermore, beyond offering important insights into the concept itself – how it was constituted and configured – it also provides a way of pinpointing certain resonances between past and present, which can be used to explore the kinds of questions that the agonists of today are wrestling with.

In reconstructing this history, Kalyvas acknowledges that the task is hampered by the fragmented and partial nature of the evidence, but the evidence does suggest that the archaic Hellenic *agones*, or athletic games of the eighth and ninth centuries BC, probably retained links to older religious rituals and festivals (2009: 18). This was a world described by Nietzsche as one of 'combat and cruelty,' where 'combat is salvation and deliverance,' and where "the cruelty of the victory is the pinnacle of life's jubilation" (1994: 175). As a test of strength, courage and skill, the *agones* were a way of selecting the best among the best, with the process of selection itself restricted to a select few, for the contest was a "competitive aristocratic practice" (Kalyvas, 2009: 18).

Crucial to this staging of games was the public nature of the contest: the *agones* were public gatherings where the best prevailed by engaging in fair and open competition. Victory was rewarded in the form of recognition, praise and admiration, with heroic feats immortalized in poetry and sculpture. Conversely, to lose and suffer defeat was to endure shame and humiliation. The stakes were high but the prize was worth the risk, because the game was staged in front of an audience who would bestow adulation upon the victor, which leads Kalyvas to conclude that the 'psychic structure' of the archaic agon resembles the myth of Narcissus. In other words, the actions of the athlete were mirrored by the audience, and the act itself would be nothing without the presence of this assembly. Just as Narcissus' self-love is mediated by his reflection, so the self-love of the athletes was refracted through, and affirmed by, the shouts, gestures, and encouragement of spectators. As Kalyvas puts it, the *agōn* of the ancients was a "strategy for inducing others to love you so that you can love yourself through their love for you" (2009: 23).

So the *agōn* originates as a contest among equals in public space. But as noted already, it was a very select group of equals who gathered together in staging the games. Between the seventh and sixth centuries this changed as the *agōn* became more diffuse and diverse, no longer confined to athletic contests, and no more the preserve of a nobility. It is in this sense that Kalyvas charts the democratization of the *agōn*, which was staged in a variety of forms, including artistic, legal, rhetorical, philosophical, and musical contests. As more people from a diversity of social spheres began to participate in such

games, so the quasimartial displays of physical prowess gave way to public contests of words, a process described by Kalyvas as the "steady pacification of the agon" (2009: 23). The exemplary technique was rhetoric, whereby skilful speech takes the place of physical demonstration.

Without losing sight of the fact that the freedom of the Greek *polis* was accessible only to citizens, thereby excluding women and slaves, the democratization of the *agōn* also saw the meaning of excellence enlarged. The light of public adoration that had once illuminated the hero now had to radiate outwards to the *polis* itself, so that to excel in the space of agonistic contest was to contribute to the greatness, prosperity, and security of one's city and political community (Kalyvas, 2009: 24). Furthermore, unlike the spatial division between athlete and audience, which had characterized the archaic *agones*, now every spectator was – at least potentially – a player, and because the contest had become a more regular feature of public life, so losers could rise more easily to challenge again.

This transformation did not erase the original psychic structure of the *agōn*, however, which had but sedimented in the practice itself, which is why Kalyvas coins the phrase 'democratic narcissus.' As a public performance in pursuit of excellence, the democratic agon fused political equality to the aristocratic desire for distinction, with the result that citizens – driven by envy, enmity, and ambition – often resorted to deception, manipulation, and lies in order to succeed (Arendt, 2005: 16; Nietzsche, 1994). It was this inbuilt tension that gave rise to what Kalyvas describes as a 'counternarcissistic apparatus of power,' which was intended to protect democracy against its own excesses. One instrument of this apparatus was ostracism: to remove men who became too great and walked like gods among mere mortals. Another was tyrannicide, which allowed for the assassination of people deemed enemies of the public (Kalyvas, 2009: 28–30). Both aimed at protecting the agon. In the case of ostracism, against those who excelled to such an extent that they made the outcome of contest a foregone conclusion, thus rendering the spirit of the agon – a contest among equals – obsolete. As Nietzsche observed, the 'strange institution' of ostracism was a 'stimulant: the preeminent individual is removed to renew the tournament of forces' – there must always be more than one genius so that they 'incite each other to action' and "keep each other within certain limits" (1994: 178). In the case of tyrannicide, the agon was to be defended against tyrants who negated the spirit of fair and open contest by fixing the outcome or circumventing the rules. The purpose of these practices was to keep the contest alive by purging the *polis* of its internal enemies. But there was something else, something corrosive built into the very constitution of the Greek conception of agonistic struggle, and this, argues Arendt, would help to bring an end to the Hellenic world.

The *Agōn* (Un)Bound

Arendt uses Homer's poetry to construct a diagram of political struggle, which enables her to contrast two exemplary forms of political struggle – the *polis* and the *res publica* (republic), and to distinguish politics from violence. 'It would appear,' writes Arendt, that "the Greeks separated struggle – without which neither Achilles nor Hector would ever have made his appearance and been able to prove who he was – from the military world of war, in which brute force has its original home, and in so doing turned struggle into an integrating component of the *polis* and the political sphere" (Arendt, 2005: 171–172). This was the great achievement of the Greeks: the combat between Hector and Achilles was continually played out in the daily lives of citizens as they met in the *agora*, which was the hub of public life – the place where citizens gathered, assembled, and exchanged opinions. This was politics: struggle and rivalry whereby each is confronted by the presence of others.

But if the *agora* was the beating heart of the ancient *polis*, the *polis* itself was bounded in such a way that its contests could only be directed inwards, leading to an insular politics whereby struggle – in the agonistic sense – could not extend beyond the walls of the city. The oratorical contests staged in the *agora* involved a freedom that was not 'irrevocably bound' to a person's own point of view, and this, Arendt reminds us, is *phronēsis*, or the 'insight of the political man,' meaning "the greatest possible overview of all the possible standpoints and viewpoints from which an issue can be seen and judged" (2005: 168). And yet this freedom to see the world in its 'many-sidedness' was spatially delimited (2005: 170). Politics, as a contest of words among equals, begins and ends in the bounded space of the *polis*. For the Greeks, war could not be political because it presupposes brute force, command and obedience, and even negotiations between warring states or city-states could not be political because these are not conducted in the manner of the *agōn* – debate and persuasion – but in the manner of war, i.e., cunning and deception (Arendt, 2005: 164). For the Greeks, the *agōn* was an integrating practice seated in the *polis*, while beyond the *polis* enemies were engaged with through wars of annihilation. Arendt explains that the Greeks 'became themselves' through struggle and war, but in the case of war, once battle had ceased, they retreated inside the walls of the city 'to be with themselves and their glory' (2005: 178).

For the Greeks, the founding law, or *nomos*, was applicable only to the space it constituted, enclosed, and delimited, i.e., the particular *polis*. Furthermore, the law was in effect pre-political, or at least, disconnected from political life. A lawgiver need not be a citizen, but could be commissioned, much like a sculptor or architect, 'to supply what the city required' (Arendt, 2005: 179). The consequence of this way of constituting and bounding the realm of politics was that it was spatially constricted. According to Arendt, though life in the *polis* was 'an intense and uninterrupted contest of all against all,' with citizens 'ceaselessly showing themselves to be the best,' it created an 'agonal spirit' that made alliances with rival states and city-states all but impossible (2005: 181). In short, the agonism that integrated the *polis* could not extend to a politics of interstate rivalry.

Following the thread of Homer's poem, Arendt suggests that the Romans 'were the twin people of the Greeks,' partly because they derived their origin from the same event: the Trojan War (2005: 173). Historical cousins perhaps, but for the Romans "politics grew not between citizens of equal rank within a city, but rather between alien and unequally matched peoples who first came together in battle" (2005: 178). Unlike the Greek *nomos*, the founding law of Rome – the Twelve Tables – "was not the work of one man, but rather a contract between two warring factions, the patricians and the plebes, that required

the approval of the entire populace, the *census omnium*" (2005: 179). In contrast to the insularity of Greek politics, Arendt explains that the original meaning of the Roman *lex* (law) is 'lasting tie,' which eventually came to mean 'contract,' so that laws (and treaties) tied human beings together, not coercively, but through mutual agreement. Roman law was enacted as an intrinsic part of political process, with the result that political struggle – including interstate rivalry – could give rise to "new political arenas" (2005: 178). Most importantly (Carthage being the exception), these were not wars of annihilation, because the Romans understood struggle as "the means by which they recognised both themselves and their opponents" (2005: 178).

For the Greeks, war brought an end to speech (and thus politics) by silencing the contest of words and deploying violence in their stead. For the Romans, war was a prelude to speech, which carried politics beyond the boundaries of the established state. There is a pragmatism to this, which anticipates discussion of Mouffe below, but we should first note, from Arendt, that the Roman agon should not be seen as an ethical enlargement of the political. The Roman way was to vanquish *and* spare an adversary, not out of compassion, but for the sake of expanding the reach of Rome.

There are three things of note to be distilled from the above discussion on the ancient *agōn*, i.e., three axes of contingency, which are constitutive of, and constituted by, an agonistic politics: how the space of politics is internally configured, how it is bounded against an outside, and how these dimensions of political order combine in modulating conflict, struggle, and violence. Together, these can be taken as the focus of contemporary agonism.

Staging the Contest

In a review article titled 'Political Theory and the Agony of Politics' (2007), Andrew Schaap opens with a very appropriate example of agonistic political thought: the contest between liberals and communitarians, which was to the fore of Anglo–American political theory during the 1980s. Schaap notes that the liberals are widely considered to have been the victors in that particular debate, but liberal political theory was itself transformed as a result of the encounter, which can be seen in how John Rawls' seminal, *A Theory of Justice* – which was very much the focus of communitarian critique – was reworked as a theory of *Political Liberalism* (1993). In the earlier work, Rawls used rational self-interest in place of a hypothetical social contract, building his theory around a figure of solitary introspection who was placed behind a 'veil of ignorance,' thus ensuring impartiality in the choice of principles that would 'define the appropriate distribution of the benefits and burdens of social cooperation.' In his *Political Liberalism*, Rawls' retains the original position, but it is now an "undistorted point of view from which fair *agreement* can be reached" (1993: 23, emphasis added), and here it is clear that the earlier figure of solitary introspection has entered into dialogue with others, and does so out of allegiance to a shared 'comprehensive doctrine' of values and beliefs. Furthermore, Rawls' defense of 'reasonable pluralism' insists that those who subscribe to this comprehensive doctrine be willing to engage in dialogue with others who do not, and in this way will endeavor to achieve an 'overlapping consensus' on matters of public import. This shift toward a distinctly dialogic liberalism coincides with developments in democratic theory, which pivot around the concept of 'deliberation,' which is the idea that disagreements can, and ought to be resolved through a process of public debate whereby participants reach collectively binding decisions, not through coercion and violence, but on the basis of what Jürgen Habermas calls the "force of the better argument" (1996). There are certain idealizations at work in the theory of deliberative democracy: that disagreements can be resolved through the public exchange of reasons, that participants are open to persuasion, and that each is willing to recognize the best argument as having validity for all. Habermas would add that such idealizations are not mere abstractions; that they can be reconstructed from how people actually communicate in their everyday lives, and that communication has an inbuilt *telos*, which is to reach mutual understanding and agreement, which together add up to his way of thinking about consensus. It is this ideal of 'consensus' – everything it assumes, implies and, perhaps most importantly, *insists* upon – which is perceived to be so troublesome by those who describe themselves as agonists.

Bonnie Honig (1993) places this dialogic/deliberative trend in political theory under the heading of 'virtue theory,' by which she means thinkers who equate politics with juridical, administrative, and regulative tasks geared to the objective of securing consensus and stabilizing intercommunal relations. For Honig (echoing Arendt), there is a misguided assumption shared by virtue theorists that politics, if done correctly, can be concluded by freeing us from conflict and instability. From an agonistic perspective, this is seen to 'close down the agon' by "duplicitously participating in its contests while pretending to rise above them" (Honig, 1993: 2). Similarly, Schaap insists that political theory must question "liberalism's continuing disavowal of its own political exclusions" (2007: 57). This is also what Honig refers to when she writes of the disruptive and disorderly 'remainders' of virtue theory, which in the world of actual politics carries implications for those perceived to be other. It is this abundance of being – the impossibility of creating order without also constituting remainders – that Honig has in mind when she sets *virtù* into play against virtue, and she does this by drawing on the work of Arendt and Nietzsche so as to "disrupt the closures of moral and political system" and "create new spaces of possibility" (1993: 4).

Abundance, remainders, exclusion, difference, and disruption – these are among the key ideas, which are brought into play against ideals of consensus, order, unity and stability, and together they gesture toward contest and struggle as the lifeblood of 'the political.' While contemporary agonists share a grounding in poststructuralism, they also exhibit important differences in how these ideas are deployed in the service of an agonistic politics, as can be seen by contrasting the expressive agonism of Connolly with the pragmatic agonism of Mouffe.

The Expressive Agon: Pluralization against Pluralism

There is a certain ambiguity in Arendt's discussion of the ancient agon. She is clearly endeared to the Greek *agora*, and

there are reasons to be wary of the Roman mode of conquest, yet there is a promise that emerges from the interweaving of these two ways of practicing an agonistic politics. For Arendt, 'the world' arises *between* people. Freedom and power cannot be the property of an individual, but can only exist among people in their plurality, in their living together, by their participating in political struggle, by 'showing themselves' to others through speech, and in this way coming to recognize both oneself and one's opponent as bearers of a freedom, which is granted to each other.

Some of the key coordinates of the expressive agon can be identified in this conception of political life: a vibrant public space where people voice their opinions and disagreements; where citizens come to see the 'many-sidedness' of things; and where, ultimately, they act in concert to produce earthly 'miracles' – meaning that when people interact politically, something is born from their intrinsic differences that could not have been foreseen at the outset. For Arendt, this is how politics must be understood. It is also the kind of freedom Arendt has in mind when she uses words like 'action' and 'power,' which meet in the idea of acting in 'concert.' And yet her account of the ancient agon offers a compelling example of how concerted action can lead to an insular political realm, which is constituted through exclusion. Connolly's version of the expressive agon can be seem to radicalize the Arendtian conception of politics by challenging what is, arguably, its core value: pluralism as a way of constituting and bounding the political.

According to Connolly, 'conventional' pluralism claims to celebrate diversity, but this is a diversity that exists only within 'settled contexts of conflict and collective action,' so that the question of diversity is resolved, and the scope of diversity itself bounded, by the territorial state, by the notion of a normal individual, and by monotheistic and monosecular conceptions of morality. As a result of such constrictions, the 'pluralist imagination' remains "too stingy, cramped, and defensive for the world we now inhabit" (1995: xii–xiii). For Connolly, plural*ism* must be continually unbound, so that the space of political struggle becomes an unending process of plural*ization*. This argument, which has been formulated through a sustained critical engagement with conventional pluralism, pivots around the notion of contestability, which is the focus of Connolly's work on *The Terms of Political Discourse* (1993a). Building on W.B. Gallie's essay on 'essentially contested concepts,' Connolly effects a subtle revision to Gallie's thesis by suggesting that the terms of political discourse may well be contest*able*, but whether they are contested or not has much to do with the 'politics of discourse.' The argument is that political discourse is saturated with a contingency that runs like a fault line through shared concepts and judgments, but that the contestability of political discourse is itself contingent. An attempt to contest the settled meaning of 'authority,' for example, might fall on deaf or hostile ears, negated by those who deploy a universalist discourse – whether God's will, or reason, or nature – to protect what Connolly calls 'provincial practices,' i.e., practices that anchor a particular form of life (1993a: 226). Contestation can be blocked by those who see themselves as having 'transcended the politics of discourse,' and while this might take the form of fundamentalist beliefs, it can also take the form of reification in the academic world of political theory and political science.

If, as Connolly argues, all practices are provincial, and if the terms of political discourse are contestable, then nothing is fundamental, and politics goes all the way down to being itself. Yet even here, at the threshold between politics and the political, Connolly refuses to be bound by the *logos* of ontology, which presupposes something fundamental, whether a logic or design, which can be encapsulated in a stable and complete representation of reality or existence. For Connolly, and whether we look to Marx or Hegel or some other variant, modern ontology "projects the possibility of drawing all otherness into the whole it endorses," so that "any otherness which persists will be interpreted as irrationality, irresponsibility, incapacity or perversity. It can never be acknowledged as that which is produced by the order it unsettles" (1993b: 132). Rejecting the *logos* but acknowledging that political interpretation necessarily 'invokes a set of fundaments,' Connolly offers us 'ontopolitical interpretation' as a means of staging agonistic political thought (1995: 1). Derived from Nietzsche, this is a concern with the abundance of being, or rather with the abundance that overflows settled conceptions of being. In ontopolitical terms, one of the tasks involved in critical thinking is to strip away the self-evidence of judgments that claim political or procedural neutrality, so that otherness is not so easily subdued by declaring it to be the manifestation of vice, abnormality, or madness. Connolly thus seeks to draw us into an ontopolitics of struggle and contestation where the paradox of identity/difference is not encountered as a puzzle to be solved, but as the basis of practicing 'agonistic respect' and cultivating an 'ethics of critical responsiveness' (1995). Connolly describes critical responsiveness as "an indispensible lubricant of political pluralization" (1995: xvii). As a mode of engaging with others who are perceived to be different, it does not attempt assimilation, but instead commences from an awareness and acknowledgment of identity and difference as a mutually constitutive relation, i.e., that identity is possible only because of difference, while difference is in turn dependent on identity. Furthermore, a relational field of identities is never more than provisionally stable, because difference can disturb the field in ways that transform its constituent elements. When it comes to the more practical question of practicing an ethics of critical responsiveness, Connolly explains that this is about turning a disturbance of what one is into a critical responsiveness to what one is not, so that the experience of disturbance is both intrasubjective and intersubjective (1995: xviii; Wenman, 2008: 202–206). Further to this, the willingness to engage with others through agonistic respect, whether this is extended to a person or a creed, must be 'cultivated,' which is the labor or effort involved in pluralizing the pluralist imagination. The overall objective then is to ensure that the boundaries of ontopolitical enclosures and identities retain a porosity that keeps them in a state of receptiveness and openness to encounters with difference.

The Pragmatic Agon: Agonism against Antagonism

In contrast to the expressive agon, Mouffe's theory of agonistic democracy has been described as 'pragmatic' (Schaap, 2009: 1). Mouffe does not build a normative theory that pivots around an absent presence: an immanent counterfactual such as

Habermas' ideal speech, or an ethical step beyond what currently exists, as in Connolly's pluralization. Mouffe does not imagine and celebrate a time-to-come, which will be hospitable to difference. The pragmatic agon is instead a response to what, for Mouffe, is an inescapable fact of human coexistence: there can be no *demos*, no *polis*, no political unity of any kind without exclusion. According to Mouffe, the paradox of identity/difference takes the form of a hegemony, which is constituted by an 'outside': an excluded surplus or excess, which can never be fully incorporated, and which is both the product of past antagonisms and the source of further (potential) conflicts. In short, struggle and conflict are ineradicable features of political life.

For Mouffe, democratic political thought exhibits a pervasive blindness – a naïve or blinkered understanding of human sociality, which is founded on what she describes as an 'optimistic anthropology': the idea that people are 'moved by empathy and reciprocity,' which leads on to the misguided assumption that rational deliberation can take us beyond enmity and violence. The mistake, argues Mouffe, is to overlook the salience and centrality of 'passions' in political life, and thus to miss the ways in which desire and fantasy – as affective dimensions of collective identifications – can be mobilized (2005: 5–6). For Mouffe, such passionate attachment is not, as liberals and deliberative democrats would have it – a threat to democracy. Instead, it is the inescapable 'how' of politics, and unless this is recognized, then democracy can only sow the seeds of its own destruction, because agonistic struggle will be repressed only to reemerge as violence.

There is an obvious affinity here with Arendt's distinction between politics/power and violence. Mouffe also seems to share Arendt's concern that equating freedom with a post-political vision of harmonious existence – whether among citizens or between nations and states – is to misunderstand the nature of the political. The comparison to Arendt must not be overstated, however, for Mouffe does not model her understanding of politics on the *agora*; it is instead the strategic politics of the Roman way which offers a more suitable analogue. According to Mouffe, antagonism *is* 'the political,' so that the task for democratic theory, as much as for a democratic politics, is to modulate antagonism in such a way that it takes the form of struggle among adversaries rather than a war of annihilation between enemies. If this is what, in Arendt's account of the ancient agon, distinguishes the Romans from the Greeks, it is also what distinguishes the agonism of Mouffe from that of Connolly: the pragmatic agon is not configured as an *ethics* but as a *politics* (Mouffe, 2000: 107n). It might also be noted that while Mouffe seems to converge with Connolly in her commitment to what she calls 'agonistic pluralism,' her way of theorizing this is entirely consistent with her social ontology (antagonism), which is derived from her engagement with Carl Schmitt. Given Schmitt's well-known "compromise with Nazism" (Mouffe, 2005: 4–5), Mouffe is aware that she is making a controversial move, but it is one that entails using Schmitt *against* Schmitt (2000: chap. 2). The key insight derived from Schmitt is his conception of 'the political,' which, he says, "must … rest on its own ultimate distinctions," and what sets this apart from other dimensions of human existence – the moral, the aesthetic, and the economic – is the distinction between *friend* and *enemy*. For Schmitt, it is this distinction that articulates all political relations of "union and separation, association and dissociation" (Schmitt, 1996[1932]: 26–27). For Mouffe, this essential relation of enmity, which is framed in such absolute terms by Schmitt, is the "necessary starting point for envisioning the aims of democratic politics" (2005: 13–14). But what Mouffe proposes, against Schmitt, is to configure the friend–enemy relation in such a way that struggle can accommodate "the pluralism, which is constitutive of modern democracy" (2005: 14). Schmitt insisted on a fundamental opposition between liberal morality and the democratic ideal: between the liberal conception of equal moral worth and democratic equality, which is based on a *political* distinction between those who can be counted among 'us' and those who do not belong. For Schmitt, the relationship between liberalism and democracy is one of mutual negation; they cannot coexist. Moving against the grain of this diagnosis of political modernity, Mouffe argues that we should understand the relationship between the liberal and democratic political traditions as a contingent historical articulation – a fortuitous accident, which has resulted in a paradox. But the 'democratic paradox' (2000) is not a puzzle to be solved so much as a way of staging agonistic political strategies. In short, if opponents meet as adversaries rather than enemies, then the democratic contest will prevail. This is the distinction that Mouffe makes between 'antagonism' and 'agonism': an agonistic democracy is one where "confrontation is kept open, power relations are always being put into question, and no victory can be final" (2000: 15).

For Mouffe, a democratic politics is more than the established laws and institutions of the democratic state. It is also the ideas and meanings that shape the practice of politics, which in the case of liberal democracy, have been forged from what she describes as distinct conceptual 'grammars.' Competing ways of interpreting ideals such as 'liberty,' 'solidarity,' 'equality,' 'rights,' and 'justice' (Connolly would say the essential contestability of such concepts) means that, for example, in a situation where a politics of 'we the people' threatens to overwhelm difference and diversity, then it is possible to use liberal discourse as a way of staging a counterhegemonic strategy. Moving in the opposite direction, in a neo-liberalized world, democratic ideals of solidarity can be invoked to struggle for distributive justice, and here there is an interesting overlap with Kalyvas' democratic narcissus: the way in which the ancient agon was stretched between desire for personal glory and serving the *polis*. There are echoes of this in how Mouffe interprets the pitfalls and possibilities of present day politics: a tension or paradox between individualized competitions to separate the best from the rest and agonistic contests geared toward the public good. While the former are likely to give rise to zero-sum relations of power, the latter dovetail with the positive-sum contests that Arendt had in mind when she equated politics with concerted power.

Concluding Reflections

Writing from within the frame of agonism, Andrew Schaap has suggested that "theorists of agonistic politics today are presented with a choice between Arendt and Schmitt" (2007: 60).

One might add that such a choice would resemble a decision as to whether politics should be modeled on the agonism of the ancient Greeks or ancient Rome. But to accept the inevitability of such a choice is in itself a way of closing down the agon. Perhaps it is not a choice between the expressive vs the pragmatic, so much as a question of strategy: of how to enact these agonistic games through specific engagements, encounters, and problematizations. The question of how to stage such strategies surely cannot be settled in advance, i.e., before struggle has commenced, and thus before the parameters and contours of particular struggles have entered into the realm of experience. But whatever form struggle takes, it is perhaps worth remembering, from the ancients, that the practice of agonism is shaped by the ways in which the space of politics is internally configured and bounded against an outside. It is the political context in question, which is likely to determine the likelihood of either an ethics of critical engagement or a counter-hegemonic political strategy succeeding, i.e., in accordance with the agonistic understanding of pluralism.

See also: Alliances: Political; Conflict and Consensus; Conflict and War, Archaeology of: Weapons and Artifacts; Deliberation and Democracy; Democratic Theory; Ethnic Conflicts; Hegemony and Cultural Resistance; Liberalism: Political Doctrine and Impact on Social Science; Military, War, and Politics; Peace; Pluralism; Political Communication; Politics and Participation; Power; War and Democracy; Warfare in History.

Bibliography

Arendt, H., 2005. The Promise of Politics. Schocken, New York.
Connolly, W.E., 1995. The Ethos of Pluralization. University of Minneapolis Press, London and Minneapolis.
Connolly, W.E., 1993a. The Terms of Political Discourse, third ed. Blackwell, Oxford.
Connolly, W.E., 1993b. Political Theory and Modernity. Cornell University Press, Ithaca and London.
Habermas, J., 1996. Between Facts and Norms. MIT Press, Cambridge, Massachusetts, MA.
Honig, B., 1993. Political Theory and the Displacement of Politics. Cornell University Press, Ithaca and London.
Kalyvas, A., 2009. The democratic narcissus: the agonism of the ancients compared to that of the (post)moderns. In: Schaap, A. (Ed.), Law and Agonistic Politics. Ashgate, Farnham, England and Burlington, VA, pp. 15–42.
Mouffe, C., 2005. On the Political. Routledge, London and New York.
Mouffe, C., 2000. The Democratic Paradox. Verso, London and New York.
Nietzsche, F., 1994. Homer on competition. In: Ansell-Pearson, K. (Ed.), On the Genealogy of Morality and Other Writings. Cambridge University Press, Cambridge, pp. 174–181.
Rawls, J., 1993. Political Liberalism. Columbia University Press, New York.
Schaap, A., 2007. Political theory and the agony of politics. Political Studies Review 5, 56–74.
Schaap, A., 2009. Introduction. In: Schaap, A. (Ed.), Law and Agonistic Politics. Ashgate, Farnham, England and Burlington, VA, pp. 15–42.
Schmitt, C., 1996[1932]. On the Concept of the Political. University of Chicago Press, Chicago and London.
Wenman, M.A., 2008. Agonism, pluralism, and contemporary capitalism: an interview with William E. Connolly. Contemporary Political Theory 7, 200–219.

Agraphia (Acquired Dysgraphia)

Claudio Luzzatti, University of Milano-Bicocca, Milano, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Over the last 150 years, neuroscientists and cognitive neuropsychologists have focused their attention on the acquired writing impairments that may follow a left-hemisphere brain damage (*agraphia*, or *acquired dysgraphia*). During this period, the most effective method to gain understanding of the mental processing of written language and its neurobiological foundation has been the observation of patients suffering from acquired writing deficits. Recently, a massive quantity of additional information regarding the neural writing substrate has been obtained from neuroimaging data. The present article also describes the different types of writing deficits that occur after brain damage and clarifies the principles used in their diagnosis.

Agraphia is an acquired writing impairment generally caused by left-hemisphere brain damage in individuals with normal premorbid literacy acquisition. The term was coined in 1869 by John William Ogle (1824–1905). As this article will describe acquired writing impairments caused by focal brain damage, a preliminary terminological clarification is helpful. The classical neurological authors of the nineteenth and twentieth centuries used the term *agraphia* to denote acquired, mostly isolated, spelling impairments, whereas the term *dysgraphia* was adopted to denote the developmental deficits involving the acquisition of writing abilities. However, the cognitive neuropsychologists of the second half of the twentieth century used the term *dysgraphia* to denote any spelling and writing impairment, irrespective of whether it was acquired or had a developmental origin. In line with this usage, for the purposes of this article the term *agraphia* will only be employed when referring to the classical clinical case descriptions of the nineteenth and early twentieth centuries, while the term *dysgraphia* will be used to describe all types of spelling and writing disorders, consistently with the contemporary cognitive neurolinguistic terminology.

Writing deficits are mostly associated with aphasia; in other words, acquired language impairments caused by left-hemisphere brain damage usually involve all linguistic units (i.e., the phonological, lexical, and morpho-syntactic abilities) and modalities (production and comprehension; oral and written language). Consequently, most aphasic patients suffer from proportionate damage also involving written language (acquired *reading and writing disorders*). Less frequently, focal brain damage may involve almost exclusively the processing of written language, and of writing in particular, which may depend on a spelling deficit, i.e., impaired orthographic processing or be specific to the retrieval of grapho-motor information, i.e., in handwriting only (*pure motor agraphia* and *dyspraxic agraphia*).

General Premises: Oral and Written Language

Written language differs quite radically from oral language, particularly, as far as the mechanisms underlying its acquisition are concerned, both from a philogenetical and onthogenetical point of view. The acquisition of oral language is genetically predetermined and automatic, i.e., a child acquires his/her mother tongue without any effort in a relatively fixed critical period of his/her development; the acquisition of written language, on the contrary, is effortful, and requires specific training. Moreover it cannot be considered as being genetically based, since its use was only introduced approximately 6000 years ago, a very short period of time in terms of natural evolution processes, and even in the Western hemisphere the use of the written language has only become widespread during the last few generations. Furthermore, a child born into an environment where historically there has never been any exposure to a written language acquires this competency with the same degree of skill and effort as a child born into a society where it has been present. This would indicate that written language does not have an independent genetic foundation, i.e., a foundation based on specific and innate neural basis.

Acquired Impairments of Written Language

Written Language Deficits in Classical Aphasiology

Several functional and neuroanatomical models of written-language processing were proposed by the classical neuropsychologists of the nineteenth century, mostly based on clinical observation of neuropsychological patients and on the assumption of lexical stores containing input and output word representations (e.g., Lichtheim, 1885; Charcot, 1883; see **Figures 1 and 2**). There were two main issues with these models, the first being whether written-word processing is dependent or independent from oral processing and the second regarding the mutual independence of the written input lexical store from the output lexical store.

In his seminal work on the neural basis of language, the Austrian physiologist Siegmund Exner (1881) (see **Figure 3**) identified a left-hemisphere area in the middle frontal gyrus (just above Broca's area and anterior to the primary motor control area of the hand), which he found to be critically involved in several patients suffering acquired writing impairments. Based on his anatomo-clinical correlative study, he concluded that this frontal area would play a specific role in regulating written expression. The French neurologist Albert

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54023-9

Ludwig Lichtheim, 1845–1928

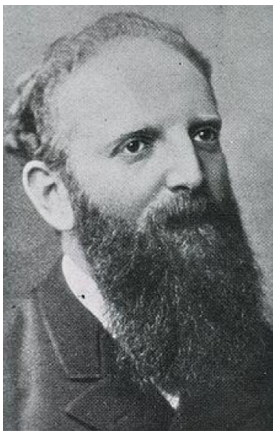

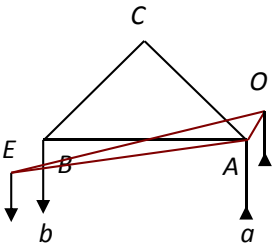

Figure 1 Ludwig Lichtheim and his diagram applied to the processing of written language. A = center of the auditory images of words; B = center of the motor images of words; C = conceptual representation of objects; a = auditory analysis; b = speech motor programs; O = center of visual engrams (word reading); E = center of motor engrams of the hand (handwriting). Modified from Lichtheim, L., 1885. On aphasia. Brain 7, 433–484.

Pitres (1884) found a lesion in the same area to be the cause of pure motor agraphia with a relatively spared ability to spell out words. On the strength of this observation, he suggested that the motor memory of written letters and words is fully independent of the visual memory for letter shapes and letter combinations and the auditory memory of word sounds. However, on the basis of his observation of two patients, one suffering from isolated reading impairment (*pure alexia*, or *word blindness*) and one with reading and writing deficits (*alexia with agraphia*) but no oral production and comprehension deficit, Joseph-Jules Dejerine (1892) assumed a single store of written-word knowledge in the *left angular gyrus*, which must be accessed both in reading and writing (see Figure 4).

Dejerine's model of written language and Exner's writing center continued to be the major functional and neuroanatomical references for the neural foundations of written language processing, and of its impairment after brain damage, until the second half of the twentieth century.

The Cognitive Neuropsychological Approach to Written Language and Its Impairments

The classical functional and neuroanatomical models of language described so far were not able to account for certain

Jean-Martin Charcot, 1825–1893

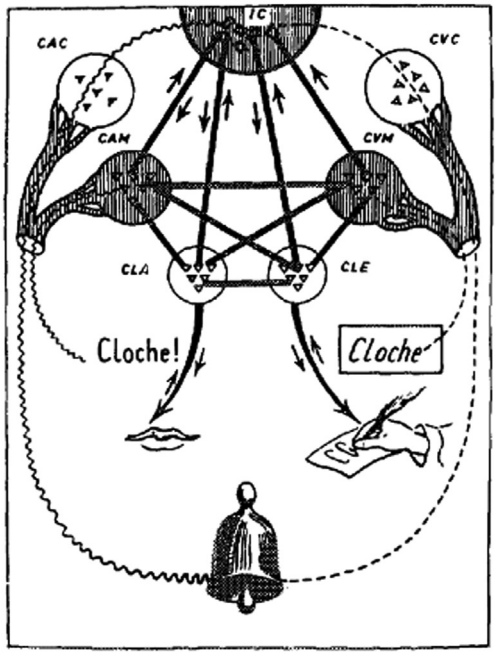

Figure 2 Jean-Martin Charcot and his diagram describing the lexical processing of oral and written language. CAC = common auditory center; CVC = common visual center; CAM = center of auditory memory of words; CVM = center of visual memory of words; IC = ideational center; CLA = center of language articulation (motor memory of words); CLE = center of written language (graphic memory of words). Modified from Charcot, J.M., 1883. Le Differenti Forme di Afasia. Vallardi, Milan.

critical phenomena occurring in patients suffering from acquired reading and writing deficits such as semantic errors (*semantic paralexia* and *semantic paragraphia*), *imageability* effects (concrete words may be read or spelled better than abstract words), or *grammatical class* effects (nouns may be read or spelled more accurately than verbs, and the latter better than function words). Furthermore, different degrees of impairment were found in certain patients when processing words with differing orthographic aspects, such as regular words, irregular words (e.g., YACHT, /jɒt/ or COLONEL, /cornel/) or when processing *nonwords* (plausible phonological or orthographic

Sigmund Exner, 1846–1926

Figure 3 Siegmund Exner.

strings with no meaning like DOUGE). In fact, literate individuals are able to read or spell words with irregular orthography as well as nonwords, but this ability may be selectively disrupted after brain damage, with the result that disproportionate deficits for either irregular words or nonwords, either in reading or in spelling may ensue. This suggests that in literately proficient adults reading and spelling proceed along two parallel routes: a lexical route for processing regular and irregular words whose orthography had already been acquired, and a subword-level routine, which may be employed for processing both regular word and nonwords.

Not all languages that employ an alphabetic script system have the same degree of regularity; for instance, French and English have a relatively opaque transcription system, i.e., the pronunciation and orthography rules regulating these languages have a vast number of exceptions. In English, for example, the orthographic string EA may be read as /i:/ in VEAL, as /ε/ in HEAD, as /Λ/ in HEART, or as /ei/ in STEAK, and the diphthong /ei/ may be spelled as AI in BRAIN or A + consonant + E in FRAME. In other languages, like Italian or Spanish, exceptions are very few, and most words can be spelled and read aloud along both the sublexical and the lexical route.

Processing Models of Writing and Acquired Dysgraphia/Agraphia

Thus a model to account for the acquired spelling ability of a literate adult has to include two independent processing routes: the lexical and the sublexical procedure (see Figure 5).

The *sublexical route* permits the transcription of strings of sounds into the corresponding letter strings. It consists of a set of four sequential processing units: first of all the auditory stimulus is analyzed (*visual analysis*), after which the perceived phonetic string passes through an *auditory-to-phonological*

Joseph-Jules Dejerine, 1849–1917
Augusta Dejerine-Klumpke, 1859–1927

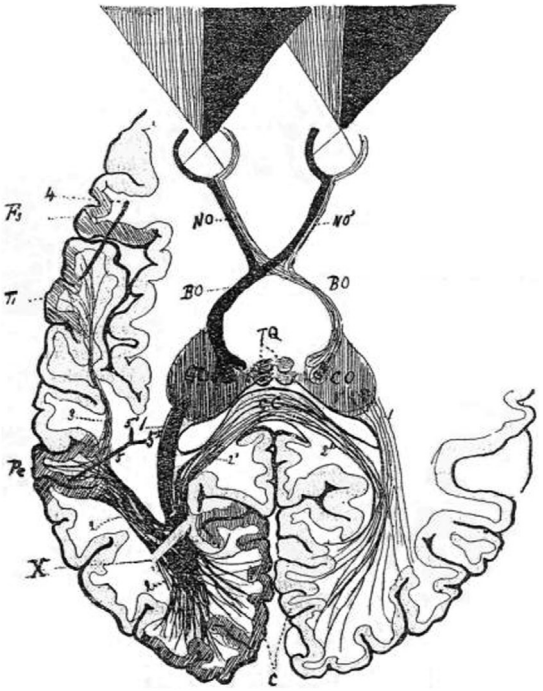

Figure 4 Above: Joseph-Jules Dejerine and his wife Augusta Dejerine-Klumpke; below: anatomo-functional diagram of the processing of written language. Dejerine, J.J., 1892. Contribution à l'étude anatomo-pathologique et clinique des différentes variétés de cécité verbale. Comptes Rendus Hebdomadaires des Séances et Mémoires de la Société de Biologie (Paris) 4, 61–90.

conversion process, by which it is segmented and sequenced into its corresponding phonemic string; the single phonemes are then stocked temporarily in the *phonological buffer*, a working memory store that can transmit them serially to the *phonological-to-orthographic conversion* unit; graphemic representations are then arranged serially in the *graphemic buffer*, a short term memory store in which the orthographic string can

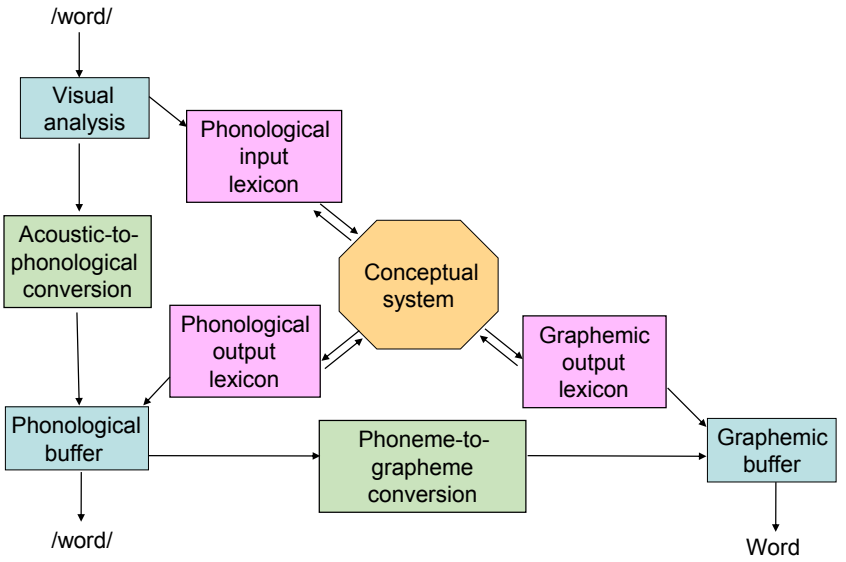

Figure 5 Dual-route model of spelling and writing: lexical and subword-level routes. Modified from Patterson, K.E., 1986. Lexical but nonsemantic spelling. Cognitive Neuropsychology 3, 341–367.

be produced for use by one of the three major vehicles for the final output, i.e., handwriting, typing on a keyboard, or the serial retrieval of letter names (oral spelling). The sublexical route is used for writing acquired regular words whose orthography had been learned in the past (or not), and nonlexical phonological strings (nonwords). Irregular words cannot be spelled correctly by this route; in fact, when it is used regularization errors inevitably occur, i.e., substitution of the correct orthographic sequence with a phonologically plausible orthographic string.

Spelling along the *lexical route* is based on three lexical and semantic stores, i.e., the phonological input, the conceptual, and the orthographic output representations of words that have already been learned. After a preliminary auditory analysis, the known word is recognized in the *phonological input lexicon* through a mapping process by which it is matched with the corresponding stored phonological representation. This, in turn, is matched with the corresponding conceptual knowledge stored in the *cognitive system* (or *semantic system*) and activates the corresponding orthographic representation stored in the *orthographic output lexicon*, which contains the orthographic information regarding words whose correct orthography has been previously learned. The lexical route permits the spelling of regular words and is the only procedure available for spelling irregular words, but it cannot be used for spelling nonwords and words whose orthography has not been previously acquired.

Central and Peripheral Dysgraphias

Cognitive psychologists classify acquired spelling impairments, distinguishing between central and peripheral dysgraphias (see Figure 6). The central forms are the result of damage to either spelling routes and to the underlying functional components (*phonological*, *surface*, and *deep dysgraphia*). Peripheral forms, on the contrary, are caused by damage located below the graphemic buffer, and include callosal agraphia *(apraxic agraphia)*, *left-hand agraphia*, and *allographic dysgraphia*.

Central Dysgraphias

The dual-route spelling model predicts two independent forms of dysgraphia. Damage to the *phoneme-to-grapheme conversion* route results in a spelling disorder that is usually called *phonological dysgraphia*. Patients suffering this impairment are unable to spell nonwords, while they can access the correct orthography of regular and irregular words effortlessly.

Damage to the *lexical route* results in a spelling disorder that is usually known as *surface dysgraphia* (or *lexical dysgraphia)*. Patients suffering from this type of dysgraphia regularize the spelling of irregular words (phonologically plausible errors), i.e., they spell COUGH as KOFF, GENUINE as JENUWEN and CASTLE as CASSELL. Errors of this type clearly indicate that the patient is spelling along the sublexical procedure. However the spelling of regular words and nonwords is relatively preserved (see Figure 6).

Deep dysgraphia. The characteristics of this spelling deficit are very similar to those of phonological dysgraphia. Besides the deficit in spelling nonwords, deep dysgraphic patients also generate *semantic substitutions* (e.g., DOG for HOUND or TREE for FOREST), *morphological errors* (SING for SINGING), lexical effects such as *word frequency* (lower accuracy for low frequency words), *grammatical class* (nouns are spelled better than verbs, and verbs are spelled better than function words), and *imageability* effects (concrete words are spelled better than abstract words). Semantic substitutions are usually explained as the consequence of an instability of the semantic system causing incongruent activation of the orthographic output representation of words that are semantically related to a target word, in the absence of a direct lexical or subword-level mediated output control.

Graphemic buffer dysgraphia. When a spelling disorder originates at the graphemic buffer level a length effect may be

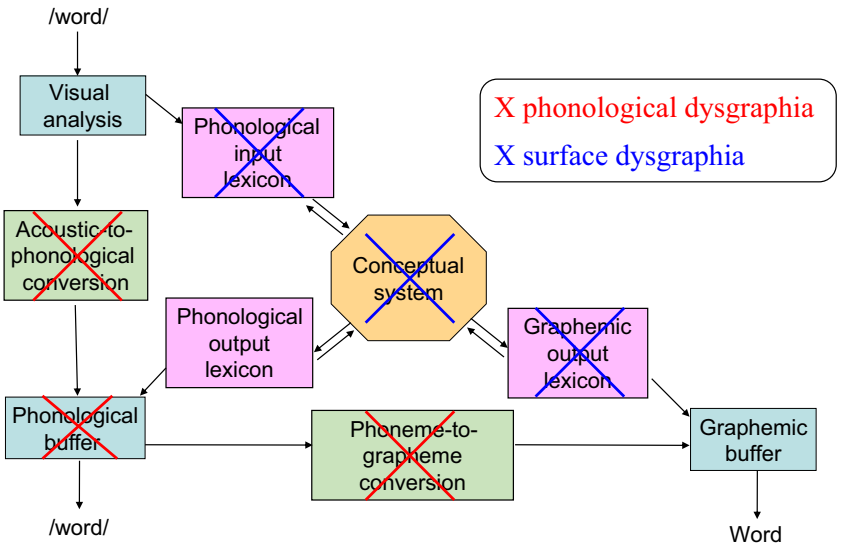

Figure 6 Major types of dysgraphia after acquired brain damage caused by disproportionate damage to either route: functional damages causing phonological and surface dysgraphia. Modified from Patterson, K.E., 1986. Lexical but nonsemantic spelling. Cognitive Neuropsychology 3, 341–367.

observed and graphemic transpositions, omissions, and substitutions predominate among error types. The syllabic structure of the target words is usually preserved, with conservation of the consonant-vowel structure (Glasspool and Houghton, 2005) so that vowels usually substitute other vowels, consonants other consonants, consonant clusters other consonant clusters. The impairment also involves, in parallel, the spelling of words as well as nonwords. These phenomena indicate that transpositions and substitutions originate at a later phase of the spelling process, i.e., at the point when the orthographic syllabic structure has already been encoded.

Peripheral Dysgraphias

These types of acquired dysgraphia originate at various processing levels located below the graphemic buffer. In this phase of the writing performance, the writing process consists of the translation of allographic symbols – the representations of the various letter shapes (e.g., A, a, *a*) – into the corresponding writing movements.

Apraxic agraphia is a writing deficit caused by the inability either to retrieve the letter shapes from a visual or motor memory store, or to produce their graphic motor pattern. This impairment does not result from primary limb or constructional apraxia, since these disorders are not necessarily associated with apraxic agraphia and, vice versa, apraxic agraphia may develop in the absence of either limb or constructional apraxia. Since the impairment consists in an inability to produce the graphic form of letters, and not in a loss of orthographic knowledge, writing on a keyboard and oral spelling are usually unimpaired.

Callosal agraphia (left-hand agraphia) is further type of impairment that affects the writing ability of the left hand only: this form of agraphia is usually caused by damage to the anteromedial tract of the corpus callosum, which results in the left-hand motor centers (right hemisphere) being disconnected from the orthographic knowledge located in the left hemisphere.

Another writing deficit is *allographic dysgraphia*, so-called because it impairs the various allographic handwriting codes (*allographs*): capital letters/italics; uppercase/lowercase. As in the case of the other peripheral dysgraphias, the allographic writing disorder differs from central dysgraphia in that it does not affect oral spelling and writing on a keyboard and handwriting is relatively fluent and letters are well formed. Finally, certain patients are unable to distinguish between capital letters and italics, uppercase and lowercase allographs, and mix different formats in the same text (e.g., d*r*OmE*da*Ry).

Principles in the Assessment of Writing Impairments

Acquired writing impairments are assessed in order to identify the processing units damaged by a cerebral lesion; an assessment is an essential prerequisite for the preparation of a cognitively oriented treatment program (Luzzatti et al., 2000).

In line with the principles outlined by the contemporary cognitive neurorehabilitation standards, a writing assessment (*writing from dictation*) has to contain lexical (regular words, irregular words) and nonlexical (nonwords) stimuli (e.g., Luzzatti et al., 1998).

The integrity of the lexical route can be assessed through performance on irregular words, that of the sublexical route by performance on nonwords. Target stimuli must be selected from different grammatical classes (nouns, verbs, function words) and must vary in length, concreteness, and word frequency. The integrity of the orthographic output lexicon should also be assessed with a written naming task (usually picture naming). When diagnosing peripheral dysgraphia, it is useful to compare the performance on *handwriting* with that on *oral spelling*. With regard to the English language, the Psycholinguistic Assessment of Language Processing in Aphasia assessment battery (PALPA: Kay et al., 1992) responds

appropriately to most of these principles, assessing the spelling of concrete and abstract words, high- and low-frequency words, content and function words, nonwords, as well as written naming.

Neuroanatomy of Writing

As stated in the introductory remarks, there are no grounds for assuming that the processing of the written language is based on a genetically independent neurofunctional circuitry. Consistently with this assumption, studies that have tried to define the neural correlates of the dual-route models of reading and writing have not obtained univocal results.

Anatomo-clinical correlative studies indicate an association of phonological dysgraphia with damage to the supramarginal gyrus and/or the insula, sparing the angular gyrus, while surface dysgraphia with damage to the left angular gyrus (e.g., Rapcsak and Beeson, 2002; Roeltgen and Heilman, 1984), thus supporting Dejerine's hypothesis concerning the importance of this area both for reading and writing (and in particular along the lexical route). However, other anatomo-clinical correlative studies (Rapcsak and Beeson, 2004) and certain functional neuroimaging studies (Beeson et al., 2003) have indicated that the fusiform gyrus and the inferior temporal cortex may have a relevant role as possible stores of lexical orthographic knowledge that need to be accessed both in word reading and in spelling. Furthermore, neuroimaging data indicate a functional overlap between the neural circuits involved in spelling and those involved in spoken language, and between the functional circuits underlying lexical and subword-level spelling performances (e.g., Norton et al., 2007).

See also: Aphasia; Classical Tests for Speech and Language Disorders; Cognitive Neuropsychology, Methods of; Dyslexias, Acquired and Agraphia; Neural Basis of Reading; Neuropsychological Testing; Reading Skills, Acquisition of: Cultural, Environmental, and Developmental Impediments; Writing Systems, Psychology of; Writing Systems.

Bibliography

Beeson, P., Rapcsak, S., Plante, E., Chargualaf, J., Chung, A., Johnson, S., Trouard, T., 2003. The neural substrates of writing: a functional magnetic resonance imaging study. Aphasiology 17, 647–665.

Charcot, J.M., 1883. Le Differenti Forme di Afasia. Vallardi, Milan.

Dejerine, J.J., 1891. Sur un cas de cécité verbale avec agraphie suivi d'autopsie. Mémoires de la Société de Biologie 3, 197–201.

Dejerine, J.J., 1892. Contribution à l'étude anatomo-pathologique et clinique des différentes variétés de cécité verbale. Comptes Rendus Hebdomadaires des Séances et Mémoires de la Société de Biologie (Paris) 4, 61–90.

Exner, S., 1881. Untersuchungen über die Localisation der Functionen in der Grosshirnrinde des Menschen. Braumüller, Vienna.

Glasspool, D.W., Houghton, G., 2005. Serial order and consonant-vowel structure in a graphemic output buffer model. Brain & Language 94, 304–330.

Kay, J., Lesser, R., Coltheart, M., 1992. PALPA: Psycholinguistic Assessment of Language Processing in Aphasia. Lawrence Erlbaum Associates, Hove, UK.

Lichtheim, L., 1885. On aphasia. Brain 7, 433–484.

Luzzatti, C., Colombo, C., Frustaci, M., Vitolo, F., 2000. Rehabilitation of spelling along the sub-word-level routine. Neuropsychological Rehabilitation 10 (3), 249–278.

Luzzatti, C., Laiacona, M., Allamano, N., De Tanti, A., Inzaghi, M.G., 1998. Writing disorders in Italian aphasic patients: a multiple single-case study of dysgraphia in a language with shallow orthography. Brain 121, 1721–1734.

Norton, E.K., Kovelman, I., Petitto, L.A., 2007. Are there separate neural systems for spelling? New insights into the role of rules and memory in spelling from functional magnetic resonance imaging. Mind, Brain and Education 1, 48–59.

Ogle, J.W., 1869. Aphasia and agraphia. Report of the MRC of St George's Hospital, London 2, 83–122.

Patterson, K.E., 1986. Lexical but nonsemantic spelling. Cognitive Neuropsychology 3, 341–367.

Pitres, A., 1884. Considérations sur l'agraphie. Revue de Médecine 4, 855–873.

Rapcsak, S.Z., Beeson, P.M., 2002. Neuroanatomical correlates of spelling and writing. In: Hillis, A.E. (Ed.), Handbook of Adult Language Disorders. Psychology Press, Philadelphia, PA, pp. 71–99.

Rapcsak, S.Z., Beeson, P.M., 2004. The role of left posterior inferior temporal cortex in spelling. Neurology 62, 2221–2229.

Roeltgen, P.D., Heilman, K.M., 1984. Lexical agraphia: further support for the two-system hypothesis of linguistic agraphia. Brain 107, 811–827.

Agrarian Political Economy

Henry Bernstein, University of London, London, UK; and China Agricultural University, Beijing, China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Agrarian political economy in its current sense originated in the 1960s, although it has longer intellectual antecedents. It can be defined as investigation of "the social relations and dynamics of production and reproduction, property and power in agrarian formations and their processes of change, both historical and contemporary". Its substantive research and debates have centered on the class dynamics and gender dynamics of (1) precapitalist agrarian formations in different parts of the world; (2) paths of agrarian change in transitions to capitalism in the now-developed countries; (3) agrarian change in experiences of colonialism; and (4) subsequent processes of change in the moment of (state-led) 'developmentalism' following political independence and the current moment of market-driven ('neoliberal') globalization.

Agrarian political economy in its contemporary usage originated in the 1960s. It is defined by *The Journal of Agrarian Change* as investigation of "the social relations and dynamics of production and reproduction, property and power in agrarian formations and their processes of change, both historical and contemporary".

Its antecedents as a field of study are the classical political economy of eighteenth century England, Scotland, and France, centered on issues of agrarian capitalism according to David McNally's potent 'reinterpretation' (1983), and the scattered writings of Marx and Engels on capitalist agriculture. Of particular note among the latter is Marx's account of 'primitive accumulation': the process by which the conditions of capitalism were established first in the English countryside through the dispossession of peasants (to become proletarians or wage workers), the conversion of land to a commodity, and of farming to market-oriented production, hence investment to make profit (Marx, 1976). Other intellectual antecedents available to 'rediscovery' by agrarian political economy also included the great debates from the 1880s among "Russia's leading economists, statisticians, sociologists, and agricultural experts…[who] provided the richest analytical literature we have on the peasant economy of any country in the period since the Industrial Revolution" (Thorner, 1966: xii), not least Lenin's *Development of Capitalism in Russia* (1964, first published 1899) and Chayanov's *The Theory of Peasant Economy* (1966, first published 1925); Karl Kautsky's *The Agrarian Question* (1988, the first full English translation of a book published in 1899); as well as the work of earlier radical intellectuals from Latin America and colonial Asia and Africa.

Agrarian political economy quickly became a lively, diverse, and contentious field of study and debate across a wide terrain: from the functioning of household farm production in different places at different times (Bernstein, 2010) thru the place of agriculture in different paths of (national) economic development in modern history (Friedmann and McMichael, 1989; Byres, 1991, 1996) to its place in the formation and development of a capitalist world economy (Wallerstein, 1974; Friedmann, 1978, 1982; Schwartz, 2000). The diversity of the field is, in various parts, an effect of (1) its interdisciplinary character, encompassing important contributions from Anthropology, Economics, Geography, History, Law, Political Science, and Sociology; (2) the applications, and clashes, of different paradigms within and across such disciplines; (3) the wide range of its historical and contemporary concerns; and (4) differences between scholars committed to the methods of materialist (Marxist) political economy. Because of such diversity of objects of study and approaches, historical periods, and interpretations, it is possible only to select and illustrate briefly some of the areas encompassed by agrarian political economy in its dynamic recent career, with this review limited to the anglophone academy.

A Founding Moment

One major stimulus, and an enduring preoccupation, was the effort to understand better the problems and prospects of economic and social development of poorer countries (only recently independent of colonial rule in most of Asia and Africa), in which "the peasant is a very essential factor of the population, production and political power" as Engels (1970: 457) had remarked of France and Germany some 80 years earlier.

A second and connected factor, in addition to its intrinsic interest, was the commitment to exploring and testing the possible contributions to such understanding of knowledge of (1) precapitalist agrarian formations in different parts of the world; (2) paths of agrarian change in transitions to capitalism in the now-developed countries; and (3) the dynamics of agrarian change in Latin American, Asian, and African experiences of colonialism, and their legacies for subsequent processes of development and underdevelopment (Bernstein and Byres, 2001).

Key works in the 'founding moment' were Eric Wolf's textbook on *Peasants* (1966) and his *Peasant Wars of the Twentieth Century* (1969); Barrington Moore Jr.'s *The Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World* (1967); the first English translation of a major work by the great Russian scholar A.V. Chayanov, *The Theory of Peasant Economy* (1966); major historical studies of the Russian peasantry before and following the Bolshevik revolution by

Lewin (1968) and Shanin (1973); and the first of a series of influential studies by James C. Scott, *The Moral Economy of the Peasant* (1976). Wolf, Barrington Moore Jr., and Scott pursued new questions beyond the mostly European focus of 'classic' Marxism. Wolf's work was informed by his studies of Mexico, and Latin America and the Caribbean more widely, and his book on peasant wars comprised case studies of Mexico, Russia, China, Vietnam, Algeria, and Cuba. Barrington Moore's great comparative study of the centrality of agrarian class struggle to different paths of state formation in the modern world encompassed the 'classic' European instances of England and France, plus the USA, and Japan, China, and India (further informed by his interest in the historical trajectories of Prussia/Germany and Russia/USSR). Scott's main region of fieldwork is Southeast Asia.

Wolf (1966) and Shanin (1973) were especially relevant to consideration of peasant social structure, Barrington Moore and Scott to peasants and politics (as was Wolf, 1969; Shanin, 1973), and Chayanov (1966) to the nature and logic of peasant agriculture. The distinctiveness of Chayanov's model consisted in its combination of a claim for peasant economy as a general (and generic) 'type' of economy, and staking that claim on a marginalist analysis of the behavior of the peasant household as a unitary farming enterprise, especially in terms of the 'demographic cycle' of generational reproduction.

What was missing from Chayanov's model of peasant economy, as from other forerunners of contemporary agrarian political economy, was any consideration of gender relations within farming households, and indeed how gender relations affect who owns what (notably but not exclusively land), who does what (divisions of labor within household farming), who gets what (divisions of the fruits of labor), what they do with it (various reproduction costs), and who decides such matters. These kinds of issues, and the significance of very different social forms of rural households and of the different ways in which they manage production and reproduction, were impressed on the attention of agrarian political economy by the impact of new feminist scholarship in the social sciences (Young et al., 1981; Kandiyoti, 1985; Agarwal, 1994; Razavi, 2003). 'Gender-blind' notions of the farm household, whether as a unit of 'utility maximization' as in Chayanov or established in other theoretical traditions (for example, petty commodity production in Marx and Lenin), were no longer tenable, and the gender as well as class differentiation of farmers (and the interconnections of gender and class) was gradually incorporated in agrarian political economy, if unevenly so.

Important institutional expressions of the field of agrarian political economy as it emerged and consolidated include, notably, the 'Peasants Seminar' of the University of London convened by Terence J. Byres from 1972 to 1989, which led to the founding of the *Journal of Peasant Studies* in 1973 (Byres, 2001); the *Agrarian Studies Program* at Yale University established by James C. Scott in 1991; the founding of the *Journal of Agrarian Change* in 2001; more recently the *Initiatives in Critical Agrarian Studies* program at the Institute of Social Studies in The Hague, the Netherlands, established by Saturnino M. Borras Jr.; and the new journals *Review of Agrarian Studies* (from 2011) and *Agrarian South; Journal of Political Economy* (from 2012) based in India.

Historical Explorations

Precapitalist Agrarian Formations

This can only be illustrated briefly from a vast literature on agrarian civilizations before capitalism. Whether the characteristics of European feudalism were found in other regions of the world, and especially South Asia, was addressed in a collection edited by Byres and Mukhia (1985), and also impacted on debate of 'tributary modes of production' to characterize the great agrarian civilizations of the East (Haldon, 1993; Banaji, 2010). Bray (1986) argued that the historic rice economies of East and Southeast Asia suggest a model of development of contemporary relevance very different from standard views of 'modern' agriculture derived from Western Europe and North America, and centered on highly capitalized farming. Debate of a transition to feudalism in Europe has been further stimulated by the *magnum opus* of Wickham (2005), subject to critique by Banaji (2010), also the author of a major materialist study of agrarian change in late antiquity (Banaji, 2001).

Part of the interest of such works of scholarship is how they employ key concepts and methods of political economy concerning class analysis and state formation to investigate agrarian pasts, and also how their findings bear on the conditions in which agrarian capitalism emerged and developed – or failed to do so.

Transitions to Capitalism

Schematically, there are two approaches in Marxist and *Marxisant* debate of the origin of capitalism (both of which can claim support from Marx's writings). One locates it in the formation of a 'world system' from the fifteenth or sixteenth centuries. The other approach centers on the transition from feudalism to capitalism in Northwest Europe from the fourteenth century or so, with the search for the 'prime mover' in the transition revisited in a seminal essay by Brenner (1976; Aston and Philpin, 1985; Wood, 2002).

Brenner explained the initial emergence of capitalist farming as a specific and conjunctural outcome of class struggle between (feudal) lords and peasants in England in contrast to France and Germany. Once established, however, capitalist agriculture – with its competitive compulsions of 'market dependence' – both paved the way for England's subsequent 'first industrial revolution' in the late eighteenth century, and generated systemic effects elsewhere. For Brenner, only capitalism was able to generate "a process of self-sustaining economic development characterized by rising labor productivity in agriculture" that overcame two obstacles that had confronted all previous agrarian civilizations: the long-term tendency of population to outrun food supply and the inability of urban population, and nonagricultural labor, to grow beyond a highly limited proportion of total population (Brenner, 2001: 171–172). Brenner's theoretical approach continues both to inspire other historical studies (e.g., Post 2011, on the agrarian origins of US capitalism) and to draw trenchant criticism (e.g., Heller, 2011).

Colonial Experiences

The colonial projects of European countries, and their various places, timings, impulses, and forms – in Latin America, Asia,

and Africa from the late fifteenth to the twentieth centuries – involved controlling and (re)organizing the labor of the colonial subjects of agrarian societies (including importing African labor to supply slave plantations in the Americas, and indentured labor from South Asia and China for sugar plantations around the world). This required intervening in their institutions and practices of land allocation and use, sometimes destroying them, sometimes modifying them. In effect, the making of colonial economies involved the breaking and remaking of precolonial modes of peasant subsistence and of landlordism, rent, and tribute (in agrarian class societies), with various effects, unintended as well as intended.

This then is another vast and diverse historical terrain, explored by agrarian political economy with respect to the formation of new frontiers and forms of agriculture linked to different phases, and mutations, of world markets that supplied an industrializing Europe and North America. Key foci of such research are (1) forms of agricultural production in the colonies, and (2) the labor regimes they deployed. The former ranged in scale from the *haciendas* of temperate Latin America and the later settler farms of east, central, and southern Africa and 'industrial plantations' of South and Southeast Asia, and tropical Latin America, to reconfigurations of peasant farming to produce crops for export and for new domestic markets, and to supply labor to plantations, settler estates, and mines. In some cases, indigenous landed or other economically and politically powerful classes were able to establish profitable locations for themselves within the agrarian economies of colonial capitalism. Various labor regimes deployed, and sometimes combined, slavery, indentured, and other forms of coerced labor (as noted), especially seasonal wage labor of various degrees of 'freedom', and peasant labor both in household farming and employed in other kinds of enterprise, agricultural and nonagricultural like the regional migrant labor system of the South African mines (Bernstein, 2010).

Much debate in agrarian political economy concerns the precise social character and effects of farming shaped by colonial rule, especially how and how much it became capitalist, or represented various amalgams or 'articulations' of capitalist and precapitalist elements, involving a set of issues that are continuing concerns. These center on questions concerning the 'persistence' or 'survival' of peasant production and community, and whether this can be explained as beneficial to capitalism in some way (e.g., by lowering costs of both agricultural commodities and rural migrant labor) or by the resilience of peasant 'resistance' to market integration ('commodification') and the states that promoted it during colonialism and since.

Also much explored and debated is the question of class differentiation of the peasantry, tabled by Lenin (1964) in late nineteenth century Russia to extend, in effect, Marx's account of primitive accumulation in England and its peasant dispossession through enclosure of land as private property. Lenin's model proposed a tendency to dissolution of the peasantry into rich peasants engaged in expanded reproduction, accumulating land and capital to become emergent capitalist farmers; a diminishing number of middle peasants able to maintain simple reproduction; and poor peasants unable to reproduce themselves through their own farming, hence subject to proletarianization, the condition of 'free' landless workers in Marx's ironic use of the term, that is, 'freed' from access to means of production, hence free to work for wages or otherwise starve. Interestingly, a prominent (non-Marxist) scholar of colonialism, possibly influenced by agrarian political economy, argued that classes of rich peasants/emergent capitalist farmers were key to the political settlement of independence from colonial rule in much of Asia and Africa (Low, 1996).

Such issues carry forward to, and connect with, a third central theme: the effects of the extremely diverse forms of agrarian economy and producers created by varied colonial experiences for the ambitions and projects of 'national development' that marked the moment of political independence in Asia and Africa from the 1940s, with echoes of the significantly earlier period of independence in most of Latin America.

Development and Underdevelopment (1950s–70s)

The newly independent countries of Asia and Africa emerged from colonialism still largely agrarian societies but now committed to 'national development', as were most Latin American countries which were generally more industrialized. Modernizing agriculture was usually a central element of ideas about 'national development', if often subordinated to the desire for industrialization, seen as the principal economic basis of prosperity, modernity, and sovereignty. Giving it priority could mean substituting domestic grain production with cheap wheat imports from the USA (Friedmann, 1990), or 'postponing' agricultural modernization until the development of national industry could provide farmers with modern inputs, the dominant view in India for the first 20 years of independence before the 'Green Revolution' was launched.

During the peak period of 'developmentalism' – the pursuit of state-led development – from the 1950s to 1970s, a wide range of policy measures was adopted and applied by governments in the South to 'modernize' their agriculture. Agricultural policy was also used to try to resolve some of the social tensions and contradictions inherited from their colonial histories, no less in Latin America than in Asia and Africa. Thus, for example, land reforms, of very different kinds, were widespread in this period, as was government-imposed resettlement of rural populations (a familiar colonial practice), for example, in parts of Africa and Southeast Asia. The 'integrated rural development programs' of the 1970s, a comprehensive 'package' of education and health as well as economic services to the countryside, was promoted especially strongly by the World Bank and USAID (the US Agency for International Development), which some interpreted as their response to the success of a peasant-based and communist-led war of national liberation in Vietnam.

In this period, agricultural and more broadly rural development policies exhibited a lot of institutional variety and frequent 'paradigm shifts' or, more simply, changing fashions, as they do today. Despite their variety, policies and programs of modernization shared a core logic: promoting a more productive agriculture based in deepening commodity relations, whether through 'smallholder' development or larger scale farming, public and private. This was often pursued by governments in the South in 'partnership' with the

World Bank, bilateral aid donors, notably the USA, Britain, and France, and private agribusiness capital (national and international), all of which supplied designs for modernization.

'More productive' addresses the technical conditions of farming, through improved varieties and cultivation methods, and greater fertilizer use, together with 'soft' credit and technical advice to farmers (extension services). This was typically done on a crop basis, whether for export crops or food crops, most famously the Green Revolution from the 1960s and its high yielding variety (HYV) seeds of the 'big three' grains of wheat, rice, and corn (or maize, the original 'Green Revolution' crop in the USA). The 'package' combined HYV seeds with fertilizers, requiring substantial irrigation to produce larger harvests.

'Deepening commodity relations' involves greater integration of farmers in markets, in which they specialize in producing particular commodities for sale, as well as buying and using greater quantities of means of production ('modern' inputs) and means of consumption, which often include food.

It is difficult to generalize about the effects of agricultural modernization efforts during the moment of 'developmentalism', because of the variety of policy measures, of their technical and institutional 'packages' and of government capacities in delivering them; and the even greater variety of ecological conditions and types of farming to which they were applied. In fact, assessing the impact of policies – a sizable profession in itself – is always challenging, because agricultural 'performance' is affected by many other factors too, from weather to the effects of macroeconomic policies (for example, and notably, concerning exchange rates of currencies and interest rates), to the vagaries of markets and prices, locally and internationally.

Agrarian political economy engaged with, and contributed to, the analysis of agrarian change and its policy debates in the moment of 'developmentalism'. This partly drew from research on precapitalist agrarian formations in the South, paths of agrarian transition in now-developed countries, and colonial experiences of agrarian change and policies to promote it, all of which could be relevant to consideration whether capitalist agrarian transition was occurring and, if so, whether it was complete. Debate of agrarian change in the early decades of independence in Asia and Africa also highlighted a central tension between, on one side, those for whom modernization was a necessary component of agricultural growth, and its role in economic development more widely, requiring "peasant elimination" (Kitching, 2001) and, on the other side, "taking the part of peasants", in Williams' succinct expression (1976). This resonates a tradition of agrarian populism as long as the histories of capitalist agriculture and industrialization (Kitching, 1982). Agrarian populism declares the virtues of peasant or family farmers and identifies with their struggles against those who threaten their reproduction and well-being, from merchants and banks, capitalist landed property, agrarian capital and agribusiness, to projects of state-led 'national development' centered on industrialization, in all their capitalist, nationalist, and socialist variants, of which the Soviet collectivization of agriculture in the 1930s was the most potent landmark. Its modern versions draw on the legacy of Chayanov, himself a victim of Stalin's purges, whose vision of a future 'peasant utopia' combined household farming with cooperation to achieve economies of scale (Bernstein, 2009, and references therein).

This historic, indeed almost constitutive, tension in agrarian political economy is expressed in debates over, for example, the character and effects of land reforms that claim to redistribute 'land to the tiller' (e.g., de Janvry, 1981; Byres, 2004); the political obstacles to taxation of capitalist farmers and rich peasants to generate an accumulation fund for industrialization ('rural bias' or at least class bias; e.g., Mitra, 1977); and in opposition to the latter the argument of 'urban bias' as the principal barrier to stronger growth by smallholder farmers, hence overcoming rural poverty (Lipton, 1977). In turn this was criticized as a (neo)populist 'myth' by Byres (1979). These kinds of questions, and attendant disagreements, carried over, and have intensified if anything, as the moment of state-led 'developmentalism' gave way to the 'neoliberalism' of market-driven doctrines and practices of development in the context of globalization since the 1970s (see below).

Anti-Imperialism and Transitions to Socialism

Two of the defining global moments of the 1960s and early 1970s, the founding moment of contemporary agrarian political economy, were the Vietnamese war of national liberation (the stimulus to Wolf, 1969) and the 'Great Proletarian Cultural Revolution' and its aftermath in China.

Are there socialist alternatives to capitalist agriculture (including the contributions of agriculture to industrialization)? Curiously, while Russia from the last quarter of the nineteenth century to Stalin's sudden and dramatic collectivization of peasants remains a topic of intense interest in agrarian political economy, it has had little to say about the long period of Soviet agriculture from the 1930s to early 1990s. While there was once a keen interest in China's communes as an 'alternative' to both capitalism and Soviet state socialism, this withered with the dismantling of the communes from the 1970s. Most work on China since then has come from agricultural economics and anthropology rather than from agrarian political economy. Recent exceptions include characterizations of current agrarian change in China as "agrarian capitalism with Chinese characteristics" (Zhang and Donaldson, 2008) and "capitalization without proletarianization" (Huang et al., 2012).

Globalization and Development (1970s Onward)

This is another immensely wide-ranging area of research and both academic and public debate, which often contains impassioned positions that are far ahead of empirical knowledge and its judicious assessment. A list of highly topical issues concerning globalization and agrarian change includes:

1. trade liberalization, shifts in global trade patterns of agricultural commodities and associated battles within and around the World Trade Organization (WTO);
2. the effects on world market prices, especially for food, of futures trading in agricultural commodities, that is,

speculation spurred by financial firms (banks, hedge funds, etc.);
3. the removal of subsidies and other forms of support to small farmers in the South as 'austerity' measures required by neoliberalism, together with reduction of government and aid budgets for most farming in the South;
4. the increasing concentration of global corporations in both agricultural input and output industries, marked by mergers and acquisitions and the economic power of fewer corporations commanding larger market shares;
5. new organizational technologies deployed by these corporations along commodity chains from farming thru processing and manufacturing to retail distribution, e.g., the 'supermarket revolution' in the global sourcing of food and market shares of food sales, now extending to some parts of the South;
6. the combination of these organizational technologies with corporate economic power to shape and constrain the 'choices' of farmers and consumers;
7. the push by corporations to patent intellectual property rights in genetic plant material, under the provisions of WTO on Trade-Related Aspects of Intellectual Property Rights, and what its critics term 'biopiracy';
8. the technical frontier of engineering plant and animal genetic material (GMOs or genetically modified organisms) that, together with specialized monoculture, contributes to the loss of biodiversity;
9. the profit frontier of biofuel production, dominated by agribusiness corporations supported by public subsidies in the USA and Europe, and its effects for world grain production for human consumption, as well as for so-called land grabbing in the South by corporations, financial firms, and the sovereign wealth funds of particular countries (in the Gulf, China, and elsewhere), an example of "accumulation by dispossession" (Harvey, 2003) as, in effect a new wave of 'primitive accumulation';
10. the health consequences, including rising levels of toxic chemicals in 'industrially' grown and processed foods, nutritional deficiencies of 'fast food', the growth of obesity and related illness together with continuing hunger and malnutrition in parts of the South;
11. the environmental costs of all the above, including levels of energy use and their carbon emissions, in the continuing 'industrialization' of food farming, processing, and sales; for example, the distances over which food is trucked, shipped, and air-freighted from producer to consumer; and
12. connected issues of the sustainability or otherwise of the current global food system and its continued growth.

How do these topical concerns connect with, or depart from, the concerns of the 'classic' agrarian question (in Kautsky's term, 1988) concerning the origins of capitalism and its subsequent spread, principally in Western and Eastern Europe? While modern agrarian political economy extended the geographical and historical scope of its research, as noted, in investigating agrarian change in the once colonial and now-independent countries of the South it initially followed the model of the 'classics'. This usually implied an 'internalist' focus on the growth (or 'blocking') of agrarian capitalism (Bernstein, 1996) in the countrysides and rural class structures of individual countries: what was happening with agricultural production? Was agrarian capitalism developing? How, and how much? Was agriculture contributing to industrialization or not? These are the kinds of questions that informed Lenin's (1964) study of Russia in the late nineteenth century, and that were transposed, for example, to India (see the Indian 'modes of production' debate of the 1960s and 1970s collected in Patnaik, 1990). Ironically, while colonialism – and the world economy of which it was part – provided an essential international dimension to arguments of the 'development of underdevelopment' in the South, including 'exploitation' of its peasantries, with political independence attention was often much more focused on the social landscapes 'internal' to its rural areas.

A 'world system' approach at least signaled the importance of the international dimension and its effects for farming and farmers in the South (and indeed the North). Its most significant expression is Friedmann's theorization of 'international food regimes' (IFRs) and her analysis of the two principal IFRs to date in the history of modern world capitalism: the first from the 1870s to1914, and the second from the 1940s to 1970s, with debate whether a third IFR has emerged in the context of globalization since the 1970s. Friedmann and McMichael (1989) outlined a global division of labor in agricultural production and trade that they argue was established from the last third of the nineteenth century.

If 'internalist' and 'world system' perspectives ran in parallel for some decades it seems more difficult to keep them separate now with the reshaping of the world agricultural and food economy in the period of globalization, indicated by the themes listed above and the challenges they present to agrarian political economy in the second decade of the twenty-first century.

Corporate Power and Technology

Of the list of 12 themes above, no fewer than seven directly concern corporate agribusiness (the first, and fourth to ninth) and the other themes are often connected by its critics with global corporate power and its technologies and business strategies – material, organizational, and symbolic (e.g., branding of food commodities). The dangers of corporate power are also closely associated with growing awareness of, and concern with, environmental sustainability as well as social justice.

Ecology

Environmental sustainability has become a major concern in recent decades, generating new fields of ecological economics and political ecology. The former is centered on 'the unavoidable clash' between economy, not least the industrialization of farming, and the environment (Martinez-Alier, 2002: ix). This has had a major impact on agrarian political economy, if unevenly so (as in the case of feminist scholarship earlier, noted above). A major landmark in this respect is the remarkable project of Jason W. Moore to construct a materialist theory and history of capitalism as 'world ecology' (e.g., Moore, 2010).

The idea of the industrialization of farming draws attention to the growing energy use of highly modernized agriculture (mechanization and 'chemicalization') and its "accelerating biophysical contradictions" (Weiss, 2010). For Martinez-Alier (2002), as for many others, the solution is to be sought through political ecology: study of the environmentally more friendly, energy 'efficient', practices of small-scale farmers (albeit at much lower levels of labor productivity). The 'environmentalism of the poor' is proposed as the alternative to the destructive technologies and practices of large-scale, highly capitalized agriculture – comprising food (and biofuel) processing and distribution as well as farming – characteristic of North America, much of Western Europe, Australia, and large agricultural exporters in the South like Argentina and Brazil.

In short, in present circumstances of globalization and its discontents, not least the concern with environmental sustainability, the long-standing tension between advocates of large- and small-scale farming ('taking the part of peasants') is as pertinent as ever and as highly charged as in the past. Contributions to this ongoing debate include arguments for taking an 'actor's perspective' to understand and appreciate better the beliefs and practices of small farmers (Long, 2001), and that 'new peasantries' are emerging, in Europe as well as in the South, who can successfully negotiate their integration with markets and choices of technology from a commitment to 'autonomy' (van der Ploeg, 2008).

Fates of the 'Peasant'/Small Farmer (Once More)

All this suggests that for champions of 'new peasantries', small farmers have not been 'eliminated' by global capitalism but can continue to adapt and to thrive, if typically in constrained circumstances, and, as noted, that they provide the key to any prospects of 'agrarian futures' that are sustainable environmentally as well as more socially just. Moreover, these kinds of claims inform the politics of a range of new agrarian social movements that mobilize regionally, nationally, and in several cases transnationally, of which the best-known example is *La Vía Campesina* ('the peasant way'), which originated in Central America (Borras et al., 2008).

Critics suggest that such claims continue to embody a romantic (populist) vision, that (1) their use of a unitary category of 'small farmers'/'peasants' obscures profound class and gender differentiation and inequality in the countryside; (2) many of those described as 'small farmers' in fact gain most of their livelihood from wage work, often as migrants (e.g., 60% plus of rural people in India are unable to reproduce themselves from their own farming); (3) this reinforces the need to examine agrarian change in the South today in terms of key rural–urban linkages, and agricultural–industrial linkages too (as in the North); (4) that small farmers using relatively simple technologies are unable to feed the rest of the world's population, now so much larger than in past eras of 'peasant' societies and now more than 50% urban.

See also: Agriculture, Economics of; Agroecology and Agricultural Change; Alternative Food Movements; Colonialism, Anthropology of; Colonization and Colonialism, History of; Food Security and 'Green Revolution'; Imperialism, History of; Peasantry in the Twenty-First Century; Peasants and Rural Societies in History (Agricultural History); Peasants in Anthropology; Rural Sociology; Sustainable Agriculture; World Systems Theory.

Bibliography

Agarwal, Bina, 1994. A Field of One's Owen: Gender and Land Rights in South Asia. Cambridge University Press, Cambridge.

Aston, T.H., Philpin, C.H.E. (Eds.), 1985. The Brenner Debate. Agrarian Class Structure and Economic Development in Pre-Industrial Europe. Cambridge University Press, Cambridge.

Banaji, Jairus, 2001. Agrarian Change in Late Antiquity: Gold, Labour and Aristocratic Dominance. Oxford University Press, Oxford.

Banaji, Jairus, 2010. Theory as History: Essays on Modes of Production and Exploitation. Brill, Leiden and Boston.

Barrington Jr., Moore, 1967. Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World. Allen Lane the Penguin Press, London.

Bernstein, Henry, 1996. Agrarian questions then and now. In: Bernstein, Henry, Brass, Tom (Eds.), Agrarian Questions: Essays in Appreciation of T.J. Byres. Frank Cass, London, pp. 22–59.

Bernstein, Henry, 2009. V.I. Lenin and A.V. Chayanov: looking back, looking forward. Journal of Peasant Studies 36, 55–81.

Bernstein, Henry, 2010. Class Dynamics of Agrarian Change. Fernwood, Halifax, NS.

Bernstein, Henry, Byres, Terence J., 2001. From peasant studies to agrarian change. Journal of Agrarian Change 1, 1–56.

Borras Jr., Saturnino M., Edelman, Marc, Kay, Cristóbal (Eds.), 2008. Transnational Agrarian Movements Confronting Globalization. Journal of Agrarian Change 8, 2–3 (special issue).

Bray, Francesca, 1986. The Rice Economies: Technology and Development in Asian Societies. Basil Blackwell, Oxford.

Brenner, Robert, 1976. Agrarian class structure and economic development in pre-industrial Europe. Past and Present 70, 30–75.

Brenner, Robert, 2001. The low countries in the transition to capitalism. Journal of Agrarian Change 1, 169–241.

Byres, Terence J., 1979. Of neo-populist pipe-dreams: daedalus in the third World and the myth of urban bias. Journal of Peasant Studies 6, 210–244.

Byres, Terence J., 1991. The agrarian question and differing forms of capitalist agrarian transition: an essay with reference to Asia. In: Breman, Jan, Mundle, Sudipto (Eds.), Rural Transformation in Asia. Oxford University Press, Delhi, pp. 3–76.

Byres, Terence J., 1996. Capitalism from Above and Capitalism from Below: An Essay in Comparative Political Economy. Macmillan, London.

Byres, Terence J., 2001. The peasants seminar of the university of London, 1972–1989: a memoir. Journal of Agrarian Change 1, 343–388.

Byres, Terence J. (Ed.), 2004. Redistributive Land Reform Today. Journal of Agrarian Change 4, 1–2 (special issue).

Byres, Terence J., Mukhia, Harbans (Eds.), 1985. Feudalism and Non-European Societies. Journal of Peasant Studies 12, 2–3 (special issue).

Chayanov, A.V., 1966. In: Thorner, Daniel, Kerblay, Basile, Smith, R.E.F. (Eds.), The Theory of Peasant Economy. Homewood, Illinois: Richard Irwin for the American Economic Association.

Engels, Friedrich, 1970. The Peasant Question in France and Germany. In: Marx, Karl and Engels, Frederick, Selected Works, vol. 3. Progress Publishers, Moscow pp. 457–476.

Friedmann, Harriet, 1978. World market, state and family farm: social bases of household production in the era of wage labour. Comparative Studies in Society and History 20, 545–586.

Friedmann, Harriet, 1982. The political economy of food: the rise and fall of the post-war international food order. American Sociological Review 88 (Annual Suppl.), S248–S286.

Friedmann, Harriet, 1990. The origins of third world food dependence. In: Bernstein, Henry, Crow, Ben, Martin, Charlotte, Mackintosh, Maureen (Eds.), The Food Question. Monthly Review, New York, pp. 13–31.

Friedmann, Harriet, McMichael, Philip, 1989. Agriculture and the state system: the rise and decline of national agricultures, 1870 to the present. Sociologica Ruralis 29 (2), 93–117.

Haldon, John, 1993. The State and the Tributary Mode of Production. Verso, London.
Harvey, David, 2003. The New Imperialism. Oxford University Press, Oxford.
Heller, Henry, 2011. The Birth of Capitalism: A Twenty-First Century Perspective. Pluto Press, London.
Huang, Philip C.C., Yuan, Gao, Peng, Yusheng, 2012. Capitalization without proletarianization in China's agricultural development. Modern China 38, 139–173.
de Janvry, Alain, 1981. The Agrarian Question and Reformism in Latin America. Johns Hopkins University Press, Baltimore, MD.
Kandiyoti, Deniz, 1985. Women in Rural Production Systems. UNESCO, Paris.
Kautsky, Karl, 1988. The Agrarian Question (Pete Burgess, Trans.). Zwan Publications, London.
Kitching, Gavin, 1982. Development and Underdevelopment in Historical Perspective. Methuen, London.
Kitching, Gavin, 2001. Seeking Social Justice through Globalization. Pennsylvania State University Press, University Park, PA.
Lenin, Vladimir I., 1964. The Development of Capitalism in Russia. The Process of the Formation of a Home Market for Large-Scale Industry. Progress Publishers, Moscow.
Lewin, Moshe, 1968. Russian Peasants and Soviet Power. A Study of Collectivization. George Allen & Unwin, London.
Lipton, Michael, 1977. Why Poor People Stay Poor: A Study of Urban Bias in World Development. Temple Smith, London.
Long, Norman, 2001. Development Sociology: Actor Perspectives. Routledge, London.
Low, Donald A., 1996. The Egalitarian Moment: Asia and Africa 1950–1980. Cambridge University Press, Cambridge.
Martinez-Alier, Joan, 2002. The Environmentalism of the Poor. Edward Elgar, Cheltenham.
Marx, Karl, 1976. Capital, vol. 1. (Ben Fowkes, Trans.). Penguin, Harmondsworth.
McNally, David, 1983. Political Economy and the Rise of Capitalism: A Reinterpretation. University of California Press, Berkeley.
Mitra, Ashok, 1977. The terms of trade, class conflict and classical political economy. Journal of Peasant Studies 4, 181–194.
Moore, Jason W., 2010. The end of the road? Agricultural revolutions in the capitalist world-ecology, 1450–2010. Journal of Agrarian Change 10, 389–413.
Patnaik, Utsa (Ed.), 1990. Agrarian Relations and Accumulation: The 'Mode of Production' Debate in India. Sameeksha Trust, Bombay.
Post, Charles, 2011. The American Road to Capitalism: Studies in Class Structure, Economic Development and Political Conflict, 1620–1877. Brill, Leiden and Boston.
Razavi, Shahra (Ed.), 2003. Agrarian Change, Gender and Land Rights. Journal of Agrarian Change 3, 1–2 (special issue).
Schwartz, Herman M., 2000. States versus Markets: The Emergence of a Global Economy, second ed. Palgrave, London.
Scott, James C., 1976. The Moral Economy of the Peasant. Yale University Press, New Haven.
Shanin, Teodor, 1973. The Awkward Class. Political Sociology of Peasantry in a Developing Society: Russia 1910–1925. Clarendon Press, Oxford.
Thorner, Daniel, 1966. Chayanov's concept of peasant economy. In: Thorner, D., Kerblay, B., Smith, R.E.F. (Eds.), A.V. Chayanov, The Theory of Peasant Economy. Richard Irwin for the American Economic Association, Homewood, IL, pp. xi–xxiii.
van der Ploeg, Jan D., 2008. The New Peasantries: Struggles for Autonomy and Sustainability in an Era of Empire and Globalization. Earthscan, London.
Wallerstein, Immanuel, 1974. The Modern World-system I: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century. Academic Press, New York.
Weiss, Tony, 2010. The accelerating biophysical contradictions of industrial capitalist agriculture. Journal of Agrarian Change 10, 314–341.
Wickham, Chris, 2005. Framing the Early Middle Ages: Europe and the Mediterranean 400–800. Oxford University Press, Oxford.
Williams, Gavin, 1976. Taking the part of peasants. In: Gutkind, Peter, Wallerstein, Immanuel (Eds.), The Political Economy of Contemporary Africa. Sage, Beverly Hills, CA, pp. 131–154.
Wolf, Eric, 1966. Peasants. Englewood Cliffs. Prentice Hall, NJ.
Wolf, Eric, 1969. Peasant Wars of the Twentieth Century. Harper & Row, New York.
Wood, Ellen M., 2002. The Origin of Capitalism: A Longer View. Verso, London.
Young, Kate, Wolkowitz, Carol, McCullagh, Roslyn (Eds.), 1981. Of Marriage and the Market. Women's Subordination in International Perspective. CSE Books, London and New York, pp. 88–111.
Zhang, Forrest, Donaldson, John, 2008. The rise of agrarian capitalism with Chinese characteristics: agricultural modernization, agribusiness and collective land rights. The China Journal 60, 25–47.

Relevant Websites

http://www.sagepub.in/browse/journal.asp?Journalid=93&Subject_Name=&SubSubjectName=&mode=1 – Agrarian South: Journal of Political Economy.
http://www.future-agricultures.org/ – Future Agriculture Consortium.
http://www.iss.nl/research/networks_and_projects/critical_agrarian_studies_icas/ – Initiatives in Critical Agrarian Studies, Institute of social studies, The Hague.
http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1471–0366/issues – Journal of Agrarian Change.
http://www.tandfonline.com/toc/fjps20/current – Journal of Peasant Studies.
http://www.yale.edu/agrarianstudies/real/ashome.html – Program in Agrarian Studies, Yale University.
http://www.ras.org.in/ – Review of Agrarian Studies.

Agreeableness

Renée M Tobin, Illinois State University, Normal, IL, USA
Daniel L Gadke, Mississippi State University, Mississippi State, MS, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Arguably the largest of the dimensions of the five-factor approach to personality, agreeableness is the motive to maintain smooth interpersonal relationships. Longitudinal research suggests stability in agreeableness across the lifespan. It is theorized to have developmental origins in the temperamental process of effortful control and has been linked to positive social behaviors including helping, cooperation, and emotion regulation. It has also been connected inversely to maladaptive social behaviors including aggression, conflict, and prejudice. Application of the Opponent Process Theory is used to explain the relations between agreeableness and social behaviors.

This article focuses on agreeableness as a dimension of personality. In it, we define agreeableness and detail its measurement methods. We also review developmental and theoretical accounts of this individual difference as well as the dimension's relations to adaptive and maladaptive social behaviors. Finally, we provide a theoretical account to aid in understanding the links between agreeableness and diverse social behavior.

Personality is the collection of stable characteristics that constitute an individual. Social scientists, especially psychologists, have long been interested in determining the core individual differences that comprise personality. Although there is some dissent about the structure of personality, the Five-Factor Model of Personality, also known as the Big Five, is widely regarded as the most accurate representation of the dimensions of human personality (McCrae and John, 1991). These five factors include agreeableness (e.g., generous, trustful, giving), extraversion (e.g., active, talkative, gregarious), conscientiousness (e.g., organized, ethical, thorough), neuroticism (e.g., anxious, self-defeating, unstable), and openness to experience (e.g., curious, introspective, intellectual; McCrae and John, 1991).

Based on factor analyses of six major data sets, Digman and Takemoto-Chock (1981) found that "friendly compliance vs. hostile noncompliance" was the first and largest factor of personality to emerge. This factor is labeled agreeableness. Agreeableness is described as the motive to maintain smooth interpersonal relations with others (Graziano and Eisenberg, 1997). "More formally, agreeableness is defined as a superordinate summary term for a set of interrelated dispositions and characteristics, manifested as differences in being likable, pleasant, and harmonious in relations with others" (Graziano and Tobin, 2013: p. 347). Research indicates relative stability in agreeableness, such that individuals who are kind early in development tend to be seen as warm, empathic, and cooperative as adults.

In 1997, Graziano and Eisenberg authored a groundbreaking chapter on agreeableness, linking it to theory and research on altruism, prosocial behavior, and motivation. When the chapter was published, little research had been conducted on this major dimension of personality; however, since its publication, research on agreeableness has increased significantly. A PsycINFO search conducted in October 2013 revealed a total of 459 peer-reviewed papers when agreeableness was entered as the keyword. Of that total, only 11 were published before 1997. These articles about agreeableness range in topic, focus, and method, with the majority reporting correlational outcomes for studies designed to examine something other than this personality dimension. Fortunately, beyond these correlational studies, an empirical literature on agreeableness has also emerged in which scientists have focused explicitly on understanding the processes underlying agreeableness and how this individual difference maps onto overt behavior. The primary goal of this section is to provide readers with a summary of some of the main findings related to individual differences in agreeableness over the last 15 years.

Measuring Agreeableness

Examination of the literature on agreeableness reveals that most studies rely on verbal self-reports of personality. Many different measures of the Big Five are available in many different languages, and these self-reports show remarkable convergence. The most popular commercial product assessing the five-factor approach to personality is the NEO, now on its third edition (Costa and McCrae, 1988). This self-report questionnaire provides five-factor scores with six facets each for individuals between the ages of 12 and 99 years. For agreeableness, the six facets are trust, modesty, compliance, altruism, straightforwardness, and tender-mindedness. Relative to the overall factor of agreeableness, less is known about these smaller facets. In addition to the NEO, a similar measure is available in the public domain through the International Personality Inventory Pool (Goldberg et al., 2006). Other self-report measures with strong psychometric properties are also available to measure the agreeableness dimension (e.g., Big Five Inventory; John and Srivastava, 1999). Beyond self-report, agreeableness (and the other dimensions of the Big Five) has been assessed using ratings from others including spouses (Costa and McCrae, 1988), employment supervisors (Hogan et al., 1996), and teachers/childcare supervisors (e.g., Digman and Takemoto-Chock, 1981; Tobin and Graziano, 2011). It is also possible to measure agreeableness in youth by obtaining

parents' free descriptions of their children (e.g., Kohnstamm et al., 1998) or through parental sorting of statements reflecting the extent to which each describes the child (e.g., Abe, 2005).

Correlations between agreeableness and other dimensions of personality vary by method and measure. Adjective marker methods (e.g., Goldberg, 1992) tend to generate stronger correlations between agreeableness and conscientiousness, whereas questionnaires with sentence formats (e.g., Big Five Inventory; John and Srivastava, 1999) tend to yield stronger correlations between agreeableness and extraversion. Agreeableness is typically inversely related to neuroticism. Studies have been mixed in terms of the relation between agreeableness and gender, with some researches finding no evidence of a relation and others finding higher agreeableness in female participants. Non-American samples and samples of young children tend to yield lower internal consistency scores for agreeableness than do American and European samples (e.g., Soto et al., 2008). Some studies also suggest that less satisfactory internal consistency scores are obtained for agreeableness when abbreviated measures are employed.

Social Desirability Artifact?

One of the most challenging issues facing agreeableness researchers is combating claims that this dimension of personality simply reflects socially desirable responding (SDR). SDR is behaving in a manner that makes the respondent appear more favorable (Graziano and Tobin, 2002). Given that items reflecting higher agreeableness tend be more favorable (e.g., 'has a trusting nature'; 'is kind and considerate of others') than those indicating lower agreeableness (e.g., 'tends to find fault with others'; 'is cool and aloof'), social desirability seems a reasonable concern. Perhaps because of the perception of social evaluation, agreeableness is often mistaken simply for social desirability. As previously discussed, agreeableness is most often measured via self-report measures (e.g., NEO, International Personality Item Pool (IPIP), Big Five Inventory (BF)I, etc.), allowing for SDR to influence self-ratings. To address this concern, Graziano and Tobin (2002) designed three studies.

Graziano and Tobin (2002) empirically investigated the potential link between SDR and agreeableness in a multimethod study. In the first study, the researchers explored the zero-order correlations between agreeableness and three different forms of SDR with a sample of 316 participants: impression management, self-deception, and self-monitoring. They found that agreeableness correlated with SDR, particularly impression management; however, correlations were also found across conscientiousness and neuroticism, suggesting that if there is a problem with agreeableness and SDR, it is one that is shared, perhaps to a greater a degree, with other traits of the Big Five.

In the second study, a round-robin approach was used with 348 research participants. Graziano and Tobin (2002) collected self-reports of agreeableness, extraversion, and SDR. Participants were then randomly assigned to one of three instruction groups: (1) *desirable-to-be-agreeable*, (2) *undesirable-to-be-agreeable*, and (3) a *control* condition. Depending on their group, participants were provided with different instructions about how best to behave. Triads were then formed containing a member of each group. The triads were expected to complete a group decision-making task and then were asked to rate the other members of their group on agreeableness and extraversion. Their results, in short, indicated no evidence of differences in self- or other-ratings across conditions, despite manipulation checks of the instructions. Furthermore, they found no evidence of significant differences across groups based on the three forms of SDR.

In the final study, groups of three research participants were asked to rate the quality of 11 different conflict resolution tactics across a series of vignettes. After practicing with their group on a vignette, groups were randomly assigned to three different social desirability groups: (1) *bad-to-be-agreeable*, (2) *good-to-be-agreeable*, and (3) *control*. The first two groups were given instructions that mapped onto their group placement; the third group was not provided with any instructions. Research participants were also asked to complete measures of agreeableness and SDR. Results were similar to those found in Study 2. That is, there was no evidence that group membership influenced ratings of agreeableness. Taken together, the data across the three studies lend little evidence to suggest that SDR accounts for ratings of agreeableness. Graziano and Tobin (2002) note that agreeableness is not "easily manipulated or distorted by SDR" (p. 696).

Theoretical Accounts of Agreeableness

Conceptually, agreeableness is often defined in terms of social motivation (i.e., the desire to maintain smooth interpersonal relations with others; Graziano and Tobin, 2013). It has also been theorized that agreeableness has developmental roots in the regulatory process of effortful control (Ahadi and Rothbart, 1994). As defined by Ahadi and Rothbart (1994), the temperamental process of effortful control is the ability to suppress a dominant response in favor of a subdominant one. This temperamental process is considered the foundation of regulation within adult personality, particularly individual differences in agreeableness in dealing with persons and conscientiousness in dealing with tasks or objects.

The theoretical foundation of effortful control has been substantiated in the empirical literature from preschool through emerging adulthood. In general, research indicates that effortful control is positively related to agreeableness (Abe, 2005; Cumberland-Li et al., 2004; Jensen-Campbell et al., 2002). Cumberland-Li et al. (2004) found that teacher and parent ratings of effortful control were related to teacher ratings of agreeableness in preschool-aged children. In a longitudinal study, Abe (2005) found that parental Q-sorts (i.e., sorting of descriptive terms into piles that apply more or less to the child) of agreeableness at age 3.5 years was related inversely to displays of negative affect at age 5, and to parent ratings of impulsive-hyperactive and conduct problems in adolescence, suggesting that early ratings of children's motives to maintain positive interpersonal relationships map onto immediate and distal regulation. Tobin and Graziano (2011) found that school-aged children high in agreeableness regulated their negative emotional reactions to disappointment more than their peers did.

Similar relations are found later in development. For example, Tobin et al. (2000) found that college students who were higher in agreeableness experienced stronger negative emotions, but they also made greater efforts to regulate these emotions when presented with emotionally evocative, negatively valenced images. Similiarly, Jensen-Campbell et al. (2002) found that college students who were high in agreeableness performed better than their peers on neurological tests associated with control behavior (e.g., Stroop test and Wisconsin Card Sorting Task), suggesting that effortful control is a common developmental substrate of agreeableness. Haas et al. (2007) examined self-reports of agreeableness as a predictor of right lateral prefrontal cortex activation, the part of the area in the brain thought to control the conscious regulation of negative affect. Using functional magnetic resonance imaging, these researchers found a direct positive relation between agreeableness and the activation of the right lateral prefrontal cortex, suggesting that individuals who are high in agreeableness are wired to regulate emotion more effectively than their peers when presented with certain emotionally evocative stimuli (i.e., fearful faces).

In a longitudinal study, Laursen et al. (2002) examined the relations between agreeableness and regulation-related behaviors. Specifically, they collected teacher and peer reports of aggression, compliance, and self-control beginning at age 8. They found that these same variables distinguished adults as high and low in agreeableness at age 33. Based on profile analysis, these researchers found two behavioral types in childhood and two personality types in adulthood, with considerable continuity in the composition of these groups over time. Fewer disobedience and concentration problems were present in high-agreeable childhood types than low-agreeable childhood types. Among boys, better school grades and fewer behavior problems were present in high-agreeable childhood types than their low-agreeable counterparts. Less alcoholism and depression, fewer arrests, and more career stability characterized the high-agreeable adulthood types relative to the low-agreeable adulthood types. From these patterns, Laursen et al. concluded that these differences in agreeableness are stable from childhood to adulthood, and are linked to important regulatory processes that lead to advantages to those children, relative to their peers, over time.

Linking Agreeableness to Social Behaviors

The literature is rich with examples of the predictive utility of agreeableness. As mentioned previously, it is arguably the largest of the five factors (i.e., Digman and Takemoto-Chock, 1981), and it is directly related to an individual's ability to control his or her emotions and related behaviors across development (e.g., Cumberland-Li et al., 2004; Tobin and Graziano, 2011; Tobin et al., 2000). Additionally, the empirical support for agreeableness as a predictor variable for overt behavior extends to include, but is not limited to, behaviors such as cooperation, helping, and conflict resolution. This section is not intended to be an exhaustive review of the literature linking agreeableness to overt behavior; rather, it provides a survey of selected behaviors with considerable research support.

Friendship

Not surprisingly, agreeableness has been directly linked to peer acceptance and friendship. In a multicomponent study, Jensen-Campbell et al. (2002) investigated agreeableness as a predictor of initiating and maintaining friendships in groups of fifth and sixth grade participants. In their initial study, participants were asked to complete measures of personality and self-esteem. Additionally, the researchers used peer nominations, asking the participants to list their five best same-sex friends in their grade and school. These reports were used to generate composite peer acceptance and friendship scores. As expected, agreeableness was directly related to higher composite scores for both peer acceptance and friendship such that participants high in agreeableness had higher composite scores of both.

The second study built on the first using a modified version of the peer nomination form. On the new peer assessment form, participants rated their classmates across several questions that mapped on to the superordinate domains of victimization, physical strength, externalizing behaviors, internalizing behaviors, and prosocial skills. Participants completed this measure in both the fall and spring to allow for the calculation of change in ratings over time as they related to personality. The results revealed that participants rated as high in agreeableness were initially given lower ratings of victimization. This remained true when reevaluated in the spring. Additionally, high agreeableness scores were associated with lower behavioral risk factors over the course of the year (e.g., weakness, internalizing behaviors). Taken together, these studies suggest that high-agreeable students are more likely to be identified as friends (i.e., higher friendship ratings), be less victimized, and overall have less behavioral risk factors as identified by their peers. This comes as no surprise given Graziano and Eisenberg's (1997) hypothesis that agreeableness evolved out of a need for groups to get along and survive together.

Cooperation and Competition

Agreeableness has been linked to differential responding to cooperative and competitive tasks. In a multicomponent study, Graziano et al. (1997) explored the relations among agreeableness, competitiveness, and behavior during games with different goal structures. In the initial study examining expectations, Graziano and his colleagues found that individuals high in agreeableness anticipated problems associated with goal-oriented competitive situations relative to their peers. In a follow-up study, the researchers placed participants in same-sex triads to examine their behavioral responses to a tower-building task. The triads were comprised either two individuals rated as high in agreeableness and one individual rated low in agreeableness, or of two individuals low in agreeableness and one individual high in agreeableness (i.e., A+A+A− or A−A−A+). The triads completed 12 trials of a tower-building task as a group and behavior was examined at the individual and group level. Triads were randomly assigned to complete the task in either a promotive or contrient goal structure condition. In the promotive condition, participants equally benefited from the triad's collective performance. In

contrast, triads in the contrient condition competed with one another, with only one member of the triad receiving a reward for each trial. At the study's conclusion, each member was asked to evaluate him- or herself and the other group members. Results of the second study indicated that individuals low in agreeableness display more competitive behaviors than individuals high in agreeableness. Further, the researchers found that competitiveness acts a mediator between agreeableness and group cooperation, such that low-agreeable individuals do not view themselves as interdependent with the group and are more competitive in response to goal structures relative to high-agreeable individuals. Additionally, overall results of Graziano et al.'s (1997) studies suggest that agreeableness is related to individual differences in competitive group behavior.

Building on this work, Tobin et al. (2002) conducted a similar study with school-aged children and found consistent results. Using a within-subjects design, Tobin and colleagues found individuals high in agreeableness foster cooperation among the group even in competitive situations. They also found greater cooperation in groups as group-level agreeableness increased. In a web-based study of adults, Hilbig et al. (2013) also found that agreeableness was related to non-retaliatory cooperation. Taken together, these studies provide evidence of the connection between agreeableness and cooperative behaviors across age groups.

Helping

Agreeableness has been linked to prosocial behaviors such as helping and empathy. Across a series of four studies, Graziano et al. (2007a) explored the links between prosocial motives, helping, empathy, and personality. In the first study, participants provide ratings of their likelihood of helping in two situations: everyday helping (i.e., being late for work to help a person whose car had broken down on the side of the road) and extraordinary helping (i.e., entering a burning house to save an individual trapped inside), depending on whether the person in need of help was a sibling, friend, or a stranger. Results indicated that individuals high in agreeableness are generally more likely to help others than their peers are. When relationships are taken into consideration, individuals high in agreeableness are more likely to help a friend or a sibling than their peers are in everyday helping situations. In extraordinary helping situations (i.e., entering a burning building), high-agreeable individuals were more likely to help a stranger than their low-agreeable peers, but there was no difference between the two groups regarding their likelihood in saving a friend or a sibling. Thus, in everyday situations high-agreeable individuals may be more likely to help people they know, but in extreme circumstances, high-agreeable individuals are also more likely to potentially risk their life to save a complete stranger.

In the second study, participants listened to a radio broadcast featuring a female college student who had recently lost her parents and was struggling to take care of her younger siblings. Prior to listening to the broadcast, participants were randomly assigned to adapt one of two listening perspectives, empathetic (i.e., focusing on the emotional aspects of the broadcast) or technical (i.e., focusing on the production aspects of the broadcast) perspective. Participants were also randomly assigned to an in-group versus out-group condition: the student attended the same university or a different university as the participant. Following the broadcast, researchers informed the participants that the broadcast would not be shared with the public and then provided an opportunity to help the student. Results indicated that participants high in agreeableness were more likely to help the girl when she was part of the out-group (i.e., attended a different university) than were low-agreeable peers; however, there was evidence of a difference in willingness to help when the girl was part of the in-group. Additionally, individuals high in agreeableness in the technical perspective condition were more likely to provide help to the girl than individuals low in agreeableness in that condition. Finally, consistent with other studies, participants high in agreeableness were more likely to rate the student as similar to themselves than their peers were. Together, these results suggest that high-agreeable individuals may be likely to help members of different groups than their own, express empathy and helping, and are more likely to take a similar perspective and identify individuals in need than low-agreeable individuals are.

Graziano et al.'s (2007b) third study used an identical paradigm to that of Study 2, with the addition of an emotional reaction questionnaire. The questionnaire asked participants to rate their own emotions, empathic concern, and personal distress after hearing the broadcast, but prior to knowing they would be able to offer help. Results indicated that the effects of agreeableness were only noteworthy on willingness to help for those who expressed empathic concern, but this difference was only detected in the technical perspective condition.

In a fourth and final study, Graziano et al. (2007b) used this paradigm again, with the addition of a cost-of-helping variable. Participants were randomly assigned to a high cost of helping or a low cost of helping condition. In the high cost of helping, participants were told that if they volunteered it had to be for a minimum of 5 h and they could only increase their amount of time in 5-h increments. In the low cost of helping condition, individuals were able to choose to volunteer anywhere between 1 and 20 h. Results indicated that individuals high in agreeableness did not differ in their willingness to help regardless of the cost to their time. There was, in contrast, a notable difference in the helping behavior for individuals low in agreeableness. In the empathy-focused condition, when the cost of helping was low, low-agreeable individuals helped more than when the cost was high. Furthermore, low-agreeable individuals helped less in the empathy perspective condition in comparison to those in the technical perspective condition when costs were high. Finally, in the technical perspective condition only, individuals low in agreeableness helped more when the cost was high than when the cost was low. Although the results may be surprising, they suggest that individuals low in agreeableness are willing to help less when cued to empathetic concerns or markers, particularly if the cost is high. Graziano and colleagues (Graziano and Habashi, 2010; Graziano and Tobin, 2009, 2013) explain this pattern of findings by connecting it to the Opponent Process Theory detailed later in this article. Overall, this collection of studies suggested that the foundation of prosocial motivation underlying agreeableness is related to the experience of empathy, and in turn, to the expression of helping behaviors.

Conflict resolution

Agreeableness has also been linked to conflict resolution. In a multimethod study, Graziano et al. (1996) evaluated the use of 11 different conflict resolution tactics across five different types of relationships (i.e., parents, siblings, roommates, significant others, and friends) using college-aged research participants. The 11 different tactics mapped on 3 primary categories: power assertion, negotiation, and disengagement. Results from the initial study indicated that regardless of a participant's level of agreeableness (i.e., high vs low), the use of negotiation tactics was rated as the most effective means of resolving conflict. Additionally, individuals low in agreeableness endorsed power assertion tactics as more effective at resolving conflict than did their high-agreeable peers. Finally, individuals low in agreeableness were more likely to respond differentially about the appropriateness of specific conflict resolution tactics based on the relationship (e.g., parents, siblings) than were individuals high in agreeableness. For example, individuals low in agreeableness endorsed power assertion tactics when dealing with siblings and friends, but endorsed negotiation for parents and significant others. In contrast, individuals high in agreeableness endorsed negotiation as best across all relationship types.

In the second Graziano et al.'s (1996) study, participants were placed in dyads to resolve social conflicts. Findings suggest that individuals high in agreeableness perceived less conflict, liked others more, and were more likely to give their partners positive ratings following the interaction. In contrast, individuals low in agreeableness actually elicited more conflict from partners than their peers did. This finding was particularly strong for male participants who were low in agreeableness, as they were more likely to perceive, engage in, and elicit conflict from their partners than were males who were high in agreeableness.

These relations are also found earlier in development. Building on their previous work, Jensen-Campbell and Graziano (2001) used a similar paradigm with children in middle school. Importantly, they also examined these relations using a diary methodology that allowed for a more complete understanding of these processes in daily life. In this study, school-aged participants were trained to record their interactions throughout the day using a standardized recording instrument. Participants rated every interaction that lasted more than 10 min and every conflict, regardless of length, on a specified day. Conflicts included everything from a simple disagreement to physical altercations. These interactions were rated by the participants on a scale of 1 ('not angry at all') to 7 ('very angry'). Their responses were categorized by researchers as constructive tactics, destructive tactics, negative affect, and outcomes (i.e., problem solved and continue to interact with conflict partner). Overall, children rated as higher in agreeableness on self and teacher reports were more likely to endorse the use of negotiation than their peers low in agreeableness, who were more likely to endorse power assertion tactics. Specific outcomes from the diary collection indicated an inverse relation between agreeableness and anger and hurt feelings. Agreeableness was also positively related to interpersonal adjustment, as measured by the diary data.

Finally, Jensen-Campbell et al. (2003) examined the relation between agreeableness and conflict in a multimethod study. The first study involved the use of questionnaires, including one on which participants rated the appropriateness of similar conflictresolution tactics. Consistent with earlier findings, agreeableness was positively related to the use of constructive tactics and inversely related to destructive tactics. In the second study, Jensen-Campbell and her colleagues placed children in dyads to play a board game together. In these dyads, each member was supplied with a conflicting set of rules to promote a conflict interaction between the two. After the game, participants rated their perception of the game and their partner in efforts to measure their perception of conflict. Consistent with previous studies, children high in agreeableness, regardless of the rules they were given or with whom they were partnered, were less likely to perceive conflict and endorse destructive tactics than their low-agreeable peers.

Aggression

The link between agreeableness and aggression has also been explored in recent years. Gleason et al. (2004) investigated agreeableness as a predictor of aggression in middle school children using a multimethod study. In their first study, the researchers measured self-reported direct and indirect aggression. In the second study, they examined the relation between social cognitions associated with aggression and peer reports of direct aggression. Results from the first study revealed an inverse relation between agreeableness and aggression. Specifically, adolescents rated as high in agreeableness reported lower rates of both direct and indirect aggression. They found the strongest relation between agreeableness and direct aggression, suggesting that individuals high in agreeableness try to avoid acts of direct aggression. The second study substantiated these findings further, replicating that agreeableness was indirectly related to direct aggression as reported by peers. Finally, the second study also found that agreeableness was indirectly related to aggressive social cognitions. That is, participants rated as high in agreeableness were more concerned about the use of aggressive tactics than were their peers low in agreeableness.

Consistent with these findings, Tremblay and Ewart (2004) found a significant inverse relation between agreeableness and both physical and verbal aggression in a correlational study of 246 undergraduate students. Using a different approach, Ode et al. (2008) conducted three studies to systematically examine agreeableness as a moderator of the relation between neuroticism and aggression. Across these three experimental studies, these researchers found that the neuroticism-aggression link was reduced at higher levels of agreeableness. Overall, research indicates that agreeableness is inversely related to aggressive thoughts and behaviors.

Prejudice

Agreeableness is also related to differential reactions to others. In a multistudy article, Graziano et al. (2007a) investigated the relation between agreeableness and prejudice toward a female who is overweight. First, these researchers examined agreeableness as a predictor of reactions toward over 100 potential targets of prejudice, finding that agreeableness was related

inversely to personal endorsements of prejudice. Importantly, they found no evidence that agreeableness was related to perceived social norms. That is, individuals high in agreeableness did not differ from their peers in terms of understanding the social norms about the acceptability of holding prejudiced feelings toward these groups, but they did differ in their personal endorsement of such prejudice. Specifically, individuals high in agreeableness reported less negative reactions to most groups, including traditional targets of prejudice (e.g., homosexuals, Jews, Hispanics) relative to their peers.

In subsequent studies, these researchers examined potentially prejudicial reactions to specific interaction partners. In the second study, participants were randomly assigned to view the photograph of either a typical or overweight female partner with whom they were scheduled to interact over the telephone. Following conversations, participants completed a social distance measure about their partners. Results indicated that male participants low in agreeableness responded with the most prejudicial reactions, but only when partnered with an overweight woman. In the third study, the researchers replicated and extended these findings by showing that these prejudicial reactions also led to discriminatory behaviors. That is, when given an opportunity to change interaction partners, only men low in agreeableness wanted to switch partners, and they only did so when paired with an overweight female partner.

Graziano et al. (2007a) then examined whether situational differences would influence the prejudicial behavior of individuals high in agreeableness. They found that overall individuals low in agreeableness expressed more prejudice than those high in agreeableness, but that some conditions seemed to elicit these types of reactions from individuals high in agreeableness as well. Specifically, they found that individuals high in agreeableness would express prejudice toward a woman who is overweight when she is thought to have engaged in counter-normative behavior (e.g., expressing negative views of the university) or she creates more work for the participant. In contrast, individuals low in agreeableness expressed more negative reactions to their partners regardless of her expressed beliefs or the cause of additional work. Taken together, these findings suggest that agreeableness is linked to more positive reactions to others, including those who are considered targets of prejudice.

Opponent Process Theory

Graziano and colleagues (Graziano and Habashi, 2010; Graziano and Tobin, 2009, 2013) have used the Opponent Process Theory (Solomon, 1980) to help explain the processes underlying agreeableness. This model, shown in **Figure 1**,

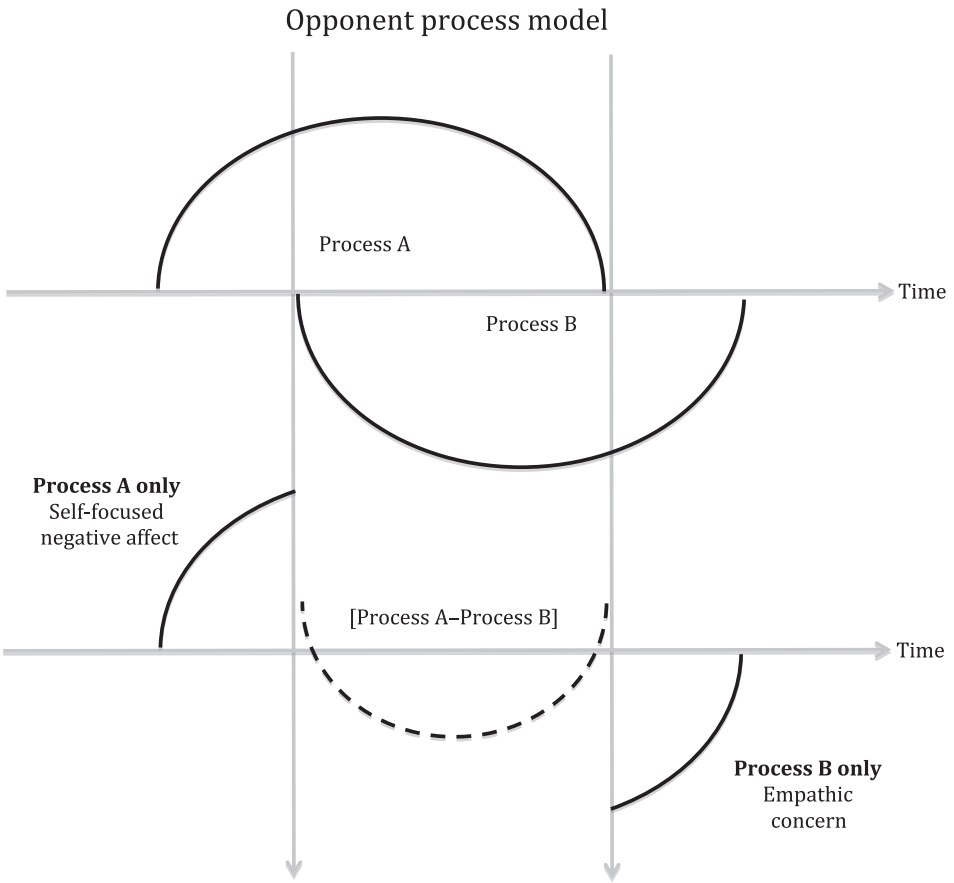

Figure 1 Opponent process model of motivation. Adapted from Solomon, R.L., Corbit, J.D., 1974. An Opponent-Process Theory of motivation: I. Temporal dynamics of affect. Psychological Review 57, 119–145.

specifically describes how two processes are activated and unfold over time. In this model, Process A is considered dominant and is characterized by almost automatic responding. Process B, although less dominant than the automatic Process A, activates shortly after and persists beyond Process A. The persistence of process B beyond A leads to the behavioral manifestation of Process B. Although Solomon (1980) used this model to describe cycles of addictive behavior, it also works well in explaining the behavioral manifestations of agreeableness, such as helping and prejudicial behaviors (Graziano and Habashi, 2010; Graziano and Tobin, 2009, 2013). For example, in the case of helping behavior, Process A is distress or self-focused negative affect and Process B is empathy. When confronted with a person in need of help, Process A (self-focused negative affect) is immediately activated, leading to the desire to escape (identified by Dijker and Koomen, 2007, and others, as the *flight-fight system*). If the individual does not escape the situation, Process B (empathy) will be activated (mapping onto the *care system*; Dijker and Koomen, 2007). Graziano and colleagues hypothesize that for individuals high in agreeableness, the onset of Process B is faster than it is for those low in agreeableness; thus, they more readily engage in empathic concern, and helping behavior follows. It appears that Process B is slower to initiate in individuals low in agreeableness relative to those high in agreeableness, leading to less helping behavior.

As mentioned earlier, these authors have also applied this model to prejudicial behavior, suggesting its usefulness in explaining a range of behaviors. Given the nature of agreeableness, it is not surprising that the motive of social accommodation would take on many forms. Metaphorically, rather than a single switchblade, agreeableness is more like a Swiss Army knife with multiple components available to handle any social situation as needed to maintain good relations. Future research on agreeableness would benefit from explorations of these relations in controlled studies that manipulate (or at least measure) aspects of these processes in fine-grained detail over time.

See also: Aggression, Social Psychology of; Big Five Factor Model, Theory and Structure; Conscientiousness; Cooperation and Competition; Emotional Regulation; Emotions and Intergroup Relations; Empathy During Early Childhood Across Cultures, Development of; Extraversion; Five Factor Model of Personality, Assessment of; Five Factor Model of Personality, Facets of; Gender Differences in Personality and Social Behavior; Helping Intergroup Relations; Longitudinal Analyses of Sexual Development through Early Adulthood; Negotiation and Conflict, The Psychology of; Personality Assessment: Overview; Personality Changes During Adolescence Across Cultures; Personality Development: Systems Theories; Personality and Adaptive Behaviors; Personality, Evolutionary Models of; Personality, Trait Models of; Personality: Historical and Conceptual Perspectives; Prejudice and Discrimination; Prosocial Behavior During Adolescence; Self-Regulated Learning; Self-Regulation During Adolescence: Variations Associated with Individual–Context Relations; Self-Regulation During Early Childhood Across Cultures, Development of; Social Competence During Adolescence Across Cultures; Temperament Development, Theories of; Temperament and Human Development; Temperament.

Bibliography

Abe, J.A., 2005. The predictive validity of the five-factor model of personality with preschool age children: a nine-year follow-up study. Journal of Research in Personality 39, 423–442.

Ahadi, S.A., Rothbart, M.K., 1994. Temperament, development, and the Big Five. In: Halverson Jr., C.F., Kohnstamm, G.A., Martin, R.P. (Eds.), The Developing Structure of Temperament and Personality from Infancy to Adulthood. Erlbaum, Hillsdale, NJ, pp. 189–207.

Costa, P.T., McCrae, R.R., 1988. Personality in adulthood: a six-year longitudinal study of self-reports and spouse ratings on the NEO personality inventory. Journal of Personality and Social Psychology 54, 853–863.

Cumberland-Li, A., Eisenberg, N., Reiser, M., 2004. Relations of young children's agreeableness and resiliency to effortful control and impulsivity. Social Development 13, 193–212.

Digman, J.M., Takemoto-Chock, N.K., 1981. Factors in the natural language of personality: re-analysis, comparison, and interpretation of six major studies. Multivariate Behavioral Research 16, 149–170.

Dijker, A.J.M., Kooman, W., 2007. Stigmatization, tolerance, and repair: An integrative psychological analysis of responses to deviance. Cambridge University Press, New York.

Gleason, K.A., Jensen-Campbell, L.A., Richardson, D.S., 2004. Agreeableness as a predictor of aggression in adolescence. Aggressive Behavior 30, 43–61.

Goldberg, L.R., 1992. The development of markers of the Big Five factor structure. Psychological Assessment 4, 26–42.

Goldberg, L.R., 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In: Mervielde, I., Deary, I., De Fruyt, F., Ostendorf, F. (Eds.), Personality Psychology in Europe, vol. 7. Tilburg University Press, Tilburg, The Netherlands, pp. 7–28.

Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R., Gough, H.C., 2006. The International personality item pool and the future of public-domain personality measures. Journal of Research in Personality 40, 84–96.

Graziano, W.G., Bruce, J.W., Sheese, B.E., Tobin, R.M., 2007a. Attraction, personality, and prejudice: liking none of the people most of the time. Journal of Personality and Social Psychology 93, 565–582.

Graziano, W.G., Eisenberg, N., 1997. Agreeableness: a dimension of personality. In: Hogan, R., Johnson, J., Briggs, S. (Eds.), Handbook of Personality Psychology. Academic Press, San Diego, pp. 795–824.

Graziano, W.G., Habashi, M.M., 2010. Motivational processes underlying both prejudice and helping. Personality and Social Psychology Review 14, 313–331.

Graziano, W.G., Hair, E.C., Finch, J.F., 1997. Competitiveness mediates the link between personality and group performance. Journal of Personality and Social Psychology 73, 1394–1408.

Graziano, W.G., Habashi, M.M., Sheese, B.E., Tobin, R.M., 2007b. Agreeableness, empathy, and helping: a person × situation perspective. Journal of Personality and Social Psychology 93, 583–599.

Graziano, W.G., Jensen-Campbell, L.A., Hair, E.C., 1996. Perceiving interpersonal conflict and reacting to it: the case for agreeableness. Journal of Personality and Social Psychology 70, 820–835.

Graziano, W.G., Tobin, R.M., 2002. Agreeableness: dimension of personality or social desirability artifact? Journal of Personality 70, 695–727.

Graziano, W.G., Tobin, R.M., 2009. Agreeableness. In: Leary, M.R., Hoyle, R.H. (Eds.), Handbook of Individual Differences in Social Behavior. Guilford, New York, pp. 46–61.

Graziano, W.G., Tobin, R.M., 2013. The cognitive and motivational foundations underlying agreeableness. In: Robinson, M.D., Watkins, E.R., Harmon-Jones, E. (Eds.), Handbook of Cognition and Emotion. Guilford Press, New York, pp. 347–364.

Haas, B.W., Omura, K., Constable, R.T., Canli, T., 2007. Is automatic emotion regulation associated with agreeableness?: a perspective using a social neuroscience approach. Psychological Science 18 (2), 130–132.

Hilbig, B.E., Zettler, I., Leist, F., Heydasch, T., 2013. It takes two: honesty-humility and agreeableness differentially predict active versus reactive cooperation. Personality and Individual Differences 54, 598–603.

Hogan, R., Hogan, J., Roberts, B.W., 1996. Personality, measurement, and employment decisions: questions and answers. American Psychologist 51, 469–477.

Jensen-Campbell, L.A., Gleason, K.A., Adam, R., Malcolm, K.Y., 2003. Interpersonal conflict, agreeableness, and personality development. Jounral of Personality 71 (6), 1059–1086.

Jensen-Campbell, L.A., Graziano, W.G., 2001. Agreeableness as a moderator of interpersonal conflict. Journal of Personality 69 (2), 323–362.

Jensen-Campbell, L.A., Rosselli, M., Workman, K.A., Santisi, M., Rios, J.D., Bojan, D., 2002. Agreeablenss, conscientiousness, and effortful control processes. Journal of Research in Personality 36, 224–251.

John, O.P., Srivastava, S., 1999. The Big Five trait taxonomy: history, measurement, and theoretical perspectives. In: Pervin, L.A., John, O.P. (Eds.), Handbook of Personality: Theory and Research, second ed. Guilford, New York, pp. 102–138.

Kohnstamm, G., Halverson, C., Mervielde, I., Havill, V. (Eds.), 1998. Parental Descriptions of Child Personality. Erlbaum, Hillsdale, NJ.

Laursen, B., Pulkkinen, L., Adams, R., 2002. The antecedents and correlates of agreeableness in adulthood. Developmental Psychology 38, 591–603. http://dx.doi.org/10.1037//0012-1649.38.4.591.

McCrae, R.R., John, O.P., 1991. An introduction to the five-factor model and its application. Journal of Personality 62, 175–215. http://dx.doi.org/10.1111/j.1467-6494.1992.tb00970.x.

Ode, S., Robinson, M.D., Wilkowski, B.M., 2008. Can one's temper be cooled? A role for agreeableness in moderating neuroticism's influence on anger and aggression. Journal of Research in Personality 42, 295–311.

Solomon, R.L., 1980. The opponent-process theory of acquired motivation: the costs of pleasure and the benefits of pain. American Psychologist 35, 691–712.

Solomon, R.L., Corbit, J.D., 1974. An opponent-process theory of motivation: I. Temporal dynamics of affect. Psychological Review 57, 119–145.

Soto, C.J., John, O.P., Gosling, S.D., Potter, J., 2008. The developmental psychometrics of Big Five self-reports: acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. Journal of Personality and Social Psychology 94 (4), 718–737. http://dx.doi.org/10.1037/0022-3513.94.4.718.

Tobin, R.M., Graziano, W.G., 2011. The disappointing gift: dispositional and situational moderators of emotional expressions. Journal of Experimental Child Psychology 110, 227–240. http://dx.doi.org/10.1016/j.jecp.2011.02.010.

Tobin, R.M., Schneider, W.J., Graziano, W.G., Pizzitola, K.M., February 2002. Nice Kids in Competitive Situations. Poster presented at the meeting of the Society of Personality and Social Psychology, Savannah, GA.

Tobin, R.M., Graziano, W.G., Vanman, E.J., Tassinary, L.G., 2000. Personality, emotional experience, and efforts to control emotions. Journal of Personality and Social Psychology 79, 656–669.

Tremblay, P.F., Ewart, L.A., 2004. The Buss and Perry Aggression Questionnaire and its relations to values, the Big Five, provoking hypothetical situations, alcohol consumption patterns, and alcohol expectancies. Personality and Individual Differences 38, 337–346.

Relevant Websites

http://ipip.ori.org/.

http://www.psychologicalscience.org/index.php/publications/observer/2012/july-august-12/the-power-of-agreeableness.html.

http://sciencereview.berkeley.edu/the-importance-of-being-agreeable/.

Agricultural Sciences and Technology

Lawrence Busch and Diana Stuart, Michigan State University, East Lansing, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The agricultural sciences include the plant, animal, and food sciences; soil science; agricultural engineering and entomology; as well as related fields such as agricultural economics, rural sociology, nutrition, forestry, fisheries, and home economics. Historians, economists, sociologists, and philosophers have all studied agricultural sciences and technology, although there are few links among these four approaches. Current studies include those of biotechnologies, nanotechnologies, relations between science and colonialism, returns to agricultural research, and the relations between public and private research in agriculture.

Introduction

The agricultural sciences and technology are usually seen to encompass the plant, animal, and food sciences; soil science; agricultural engineering; and entomology. In addition, in many research institutions related fields such as agricultural economics, rural sociology, human nutrition, forestry, fisheries, and home economics are included. The agricultural sciences have been studied by historians, economists, sociologists, and philosophers.

Most of the early work in Science and Technology Studies (STS) focused on physics, said to be the model for the sciences. Unlike the agricultural sciences, theoretical physics appeared disconnected from any clear social or economic interests. Indeed, one early study of the agricultural sciences described them as deviant in that they did not follow the norms found in physics (Storer, 1980[1961]). (Much later, further studies in STS revealed that even physics did not follow the norms claimed for it and was strongly influenced by military and industrial needs.)

Prior to the 1970s, studies of the agricultural sciences tended to be apologetic and uncritical. Then, critical historical, economic, sociological, and philosophical studies of the agricultural sciences began to emerge. These studies built on earlier work that was not within the purview of what usually is called STS. Moreover, despite attempts to incorporate perspectives from this field, it would be an exaggeration to say that studies of the agricultural sciences form an integrated body of knowledge. Indeed, fragmentation has been and remains the rule with respect to theoretical frameworks, research questions, and methods employed.

History

Recent historical studies have challenged the hagiographical approach of official histories, demonstrating how the organizational structure of agricultural science encouraged particular research strategies and products. Of particular import were studies of the role of the state-sponsored botanical and zoological gardens, and later the agricultural experiment stations, in the colonial project as well as within Europe and North America. In particular, historians began to document the close relations between the rise of economic botany as a discipline and the creation of botanical gardens in the various European colonies in the seventeenth century (Drayton, 2000). Such gardens served simultaneously to further the classification of botanical species and the colonial project by identifying plants of economic value that might serve to valorize the new colonies. Coffee, tea, cocoa, rubber, sugar, and other crops were developed as plantation crops and soon thrived in regions far from their locations of origin. In so doing, they provided revenue for colonial governments and profits for the emerging large colonial trading companies.

In the late nineteenth century, botanical gardens were superseded by agricultural experiment stations in most industrialized nations and their colonies. Until the World War II, the agricultural experiment stations were *the* model and often the sole recipient of government support for nonmilitary scientific research. The experiment stations focused on increasing yields of food crops in Europe and North America, thereby keeping industrial wages down through cheap food and avoiding feared Malthusian calamities. At the same time, the experiment stations in the colonies focused their efforts on increasing yields of exports so as to provide a steady supply of raw materials to European industries. For example, the Gezira scheme in the Sudan combined science, commerce, and irrigation so as to provide long staple cotton for the Lancashire mills (Barnett, 1977).

In contrast to most plant and animal research, mechanical and chemical technologies (e.g., farm equipment, fertilizers, pesticides) in agriculture were developed by private companies. Of particular note is that World War I and especially World War II resulted in technical innovations that made cheap nitrogen fertilizer available in much of the industrial world. This was the case because after the war demand for ammonium nitrate for bombs declined; with relatively small investment, the same facilities could be used to produce fertilizer. As a result, over the past century, the agricultural sciences and technologies played an important role in increasing agricultural productivity per hectare and per hour of labor. However, some argue that this was only accomplished by displacing vast agrarian populations and increasing environmental degradation.

In contrast, recent work suggests that European nations were successful in developing technologies that improved the lives and yields of small farmers in the early twentieth century (Harwood, 2012). Apparently, considerable successes were achieved in Germany and Austria among other places by directly involving farmers in plant breeding and making traits desired by farmers central to plant breeding. However, these successes were gradually supplanted by more 'scientific' approaches to breeding, separating users from scientists institutionally and distinguishing scientific from farmer knowledge and practices. As a result, when Green Revolution technologies were introduced in the developing world in the 1960s and the 1970s, a more top-down approach was taken. Only in the mid-1970s did serious criticism of the Green Revolution become institutionalized in the form of participatory plant breeding (e.g., Trouchea et al., 2012), farming systems research (e.g., Darnhofer et al., 2012), and agroecology (e.g., Gliessman, 2013), although those involved in these developments were rarely aware that many of the same approaches had been used nearly a century earlier. Importantly, each of these changes was led by social scientists, often leading to internal disputes within agricultural research organizations, and most certainly challenging the dominance of biological scientists (especially plant breeders) in the agricultural research hierarchy. However, by the late twentieth century, such approaches had once again been relegated to the sidelines, as the new biotechnologies took center stage.

Economics

While economic studies as early as the 1930s celebrated the products of agricultural research and emphasized 'adjustment' to the new technologies by farmers, the newer literature has focused primarily on the social rates of return to agricultural research. Many such studies have been used to lobby for additional public support for research. Recent work suggests that rates of return may have been overestimated, due to misspecifications and lag times between research and development. That said, returns are still estimated to be as much as 10% per annum (Alston et al., 2011). In contrast, critics of this approach have argued that only the benefits are estimated, while many of the costs are excluded as they are difficult to measure (e.g., acute pesticide poisoning, environmental damage, and research programs yielding few or no results).

Another strand of research in economics is based on the 'technology treadmill' theory, developed initially by Willard Cochrane (1993). Proponents of the theory argue that the market structure for undifferentiated agricultural commodities (e.g., wheat, maize) is such that farmers are always price takers. In practice, this means that they cannot raise prices beyond the market price that is set by demand from a few large buyers (a monopsony). Accordingly, the only way to increase one's income is to adopt new technologies that increase yields without raising production costs. Early adopters are often able to benefit from the use of such technologies in terms of higher returns on their investment. Moreover, they can use those returns either to expand their farms (thereby reducing the number of farms and farmers) or to invest in further cost-reducing innovations. However, as most farmers adopt the new technologies, prices decline accordingly. Hence, farmers find themselves on a technology treadmill. Adopting each technical advance allows them to produce more, but as the majority of farmers adopt the new technologies, the prices they receive decline once again. This, in turn, forces them to find yet newer technologies that will allow them to maintain their rate of profits. Those who cannot manage to keep up on the treadmill are forced out of the market, resulting in consolidation and fewer agricultural producers.

More recently, the theory of 'induced innovation' has been developed as an explanation for the directions taken in agricultural research (Hayami and Ruttan, 1985). Proponents of this framework argue that innovations are induced by the relative scarcity of land, labor, and capital. Thus, in Japan, where land is scarce, research has focused on increasing yields per unit of land. In contrast, in the United States, where labor is scarce, research has focused on increasing yields per unit of labor. The theory further asserts that agricultural research is responsive to demands of farmers as voiced in the political sphere, since much agricultural research is publicly funded. However, critics have argued that this is only likely to be true in democratic regimes (Burmeister, 1988; Plucknett et al., 1990).

Others have examined the allocation of research support across commodities, directing particular concern to what has come to be known as the problem of spillover. Since much public agricultural research has focused on the creation of products and practices that are not protectable by patents or copyrights, research completed in one nation or region can often be used with only minor adaptation in other locales. Economists have concluded that developing nations should not engage in research on commodities with high spillover, such as wheat, but should instead rely on other nations and international agricultural research centers for such materials. They argue that such nations would be served better by investing in research on commodities not grown elsewhere.

Others have argued successfully for the formation of international networks for research on particular commodities so as to spread the costs of research over several nations with similar agroecological conditions (Plucknett et al., 1990). Such networks have been effectively used to exchange information and materials as well as to foster collaboration. However, with the rise of stronger intellectual property rights in agriculture over the past several decades (including plant variety protection as well as utility patents for life forms), there is some evidence that spillovers may be on the decline. In addition, materials transfer agreements are now required to engage in certain kinds of research; often these agreements limit who can use the materials and for what purposes. Indeed, bringing many products of biological agricultural research to market now often requires the permission of numerous patent holders, thereby increasing both time and costs involved.

Another area of interest to economists has been the division of labor between public and private financing of agricultural research. While some economists have argued that biological (as contrasted with chemical and mechanical) agricultural research is by definition a public good, others argue that recently enacted stronger intellectual property rights make it possible for the private sector to shoulder most of the burden for such research. They have attempted to build the case that stronger intellectual property rights create incentives for private

firms to invest in biological research (e.g., plant and animal biotechnologies, seed production), leaving only research in the social sciences and natural resource management to the public sector. Indeed, some nations (e.g., the United Kingdom) have privatized part or all of their agricultural research with varying degrees of success. However, recent evidence from the United States, where the private sector now outspends the public sector, suggests that private sector research tends to focus on commodities and technologies with the highest expected private returns (e.g., maize, soybeans, cotton), leading to investment in a smaller range of research topics and greater emphasis on short-term gain (King et al., 2012). The growth in public–private partnerships tends to exacerbate this trend, as these partnerships tend to emphasize commodities of interest to the private sector. In consequence, declining public funding means that many critical research issues now go unaddressed.

Finally, related to the division of labor is the issue of alternative public funding mechanisms. Traditionally, agricultural research has been funded institutionally, based on annual lump-sum, formula-based appropriations. However, in recent years, there has been a shift toward more project-based grant programs in which scientists compete to receive grants. Proponents of competitive grants argue that this approach ensures that the best research is conducted by the most competent scientists. In contrast, proponents of institutional funding argue that agriculture is fundamentally place based, necessitating that investigations be distributed across differing ecological zones. They also note that competitive grants tend to be funded over just a few years, while much agricultural research requires a decade or more to complete and that the competition process itself is costly, especially when success rates are low. Furthermore, when competitive and formula funding are compared with respect to their impact on agricultural productivity, formula funds appear to have a much greater impact (Huffman and Evenson, 2006).

Sociology

In sociology, adoption–diffusion theory was the dominant approach through the 1960s (Rogers, 1995). Diffusion theorists accepted the products of agricultural research as wholly desirable. Hence, their work focused almost exclusively on the fate of innovations designed for farm use. They employed a communications model adopted from engineering in which messages were seen to be transmitted from the sender to the receiver, later adding the engineering term 'feedback' to describe receivers' responses to the messages sent to them. They argued that adoption could be best understood as based on the social psychological characteristics of adopters and non-adopters. Early adopters were found to be more cosmopolitan, better educated, less risk averse, and more willing to invest in new technologies than late adopters, pejoratively labeled 'laggards'. This perspective fit well with the commitments of agricultural scientists to transforming agriculture, making it more efficient and more modern. However, it ignored the characteristics of the innovations. Often they were large, costly, and required considerable skill to operate and maintain. Not surprisingly, those who rejected the innovations lacked the capital and education to use them effectively.

Later studies challenged the diffusion theorists. First, critics of the Green Revolution asked questions about the appropriateness of the research undertaken (Perkins, 1997). They noted that, although inexpensive in themselves, Green Revolution varieties were often parts of packages of innovations that required considerable capital investment well beyond the means of the average farmer. While acknowledging that yields increased, they documented the considerable rural upheaval created by the Green Revolution: growing farm size, displacement of both small farmers and landless laborers to the urban slums, declining status of women, declining water tables due to increased irrigation, impaired air quality, and contamination of ground water from agricultural chemicals.

Others asked how agricultural scientists chose their research problems (Busch and Lacy, 1983). They noted that science and commerce were necessarily intimately intertwined in agriculture, in the choice of research problems, in the institutional relations between the public and private sectors, and in the value commitments of scientists (often from farm backgrounds) and wealthier farmers (who often have a disproportionate influence on research directions). They challenged the engineering model of communication, seeking to substitute for it one drawn from the hermeneutic–dialectic tradition. Drawing on philosophers such as Jürgen Habermas and Hans-Georg Gadamer, they asserted that communications between scientists and the users of the products of agricultural research had to be able to debate fundamental assumptions about what constitutes a desirable future for agriculture as well as specific technical details.

Sociologists have also studied agricultural commodity chains, i.e., the entire spectrum of activities from the production of seed through to final consumption (Friedland et al., 1978). Such studies have examined the complex interaction between scientists and engineers involved in the design of new seeds and equipment and various constituent groups. Unlike the diffusion and induced innovation theorists, proponents of this approach have engaged in detailed empirical analyses of new technologies, challenging the assumptions of the designers. For example, both the tomato harvester and the hard tomato needed to withstand mechanical harvesting were built on the initiative of scientists and engineers in the public sector rather than to meet any need articulated by growers. Together, these technologies transformed tomato production in many parts of the world by reducing the number of growers and farm workers while increasing farm size. Given the limited employment opportunities of those displaced, critics question whether this was an appropriate investment of public funds.

In recent years, sociologists have devoted considerable attention to the new agricultural biotechnologies (e.g., gene transfer, plant tissue culture), especially those involving transformations of plants. It is argued that these new technologies have begun to transform the creation of new plant varieties by (1) reducing the time necessary for breeding; (2) reducing the space necessary, from large fields to small laboratories, to test for the incorporation of new traits; and (3) making it possible in principle to incorporate any gene into any organism. However, analysts note that the vast sums of private capital invested in this sector stem as much from changes in intellectual property rights as from any advantages claimed for the new technologies. In particular, they point to the advent of plant

variety protection (a form of intellectual property right), the extension of utility patents to include plants, and the imposition of Western notions of intellectual property on much of the rest of the world.

Before these institutional changes, most plant breeding was done by the public sector. Private breeding was not profitable, as seeds are both means of production and reproduction. Thus, farmers could save seed from the harvest to use the following year or even to sell to neighbors. Put differently, each farmer was potentially in competition with the seed companies (Kloppenburg, 1988). In contrast, once the new intellectual property regimes began to be implemented, it became possible to prohibit the planting of purchased seed developed using the new biotechnologies. Suddenly, the once barely profitable seed industry became a potential source of profits. Agrochemical companies rapidly purchased all the seed companies capable of engaging in research in hopes of cashing in on the new opportunities. The result has been a shift of plant breeding research for major crops to the private sector and strong prohibitions on replanting saved seed. One unforeseen consequence of this shift is the decline in the number of persons receiving education in plant breeding, as the public sector trains nearly all such persons. As a result, seed companies are having some difficulty in staffing their operations.

Another emerging area of interest is the development and use of nanotechnologies in agriculture and food (David and Thompson, 2008). Nanotechnologies take advantage of the novel physical and chemical properties of compounds displayed at the nanolevel. Potential uses here include the use of nanosensors to monitor plant and animal health as well as to promote greater food safety, the nanoencapsulation of certain compounds so as to change or enhance the flavor of foods, nanofilters to remove harmful compounds from water and other liquids, and various forms of packaging that employ nanofilms.

However, at the same time, there is considerable scientific debate over the effects of airborne nanoparticles on human health. Moreover, the unexpected consumer resistance found with respect to food and agricultural biotechnologies has made private actors wary of nanotechnologies. As a consequence, they have simultaneously begun research on their use, bringing some technologies to market, even as they have tended to label these new processes and products in ways that obscure their nanolevel properties. Hence, some observers have argued that the private sector is setting itself up for a repeat of the problems with biotechnologies.

Another line of work has examined particular agricultural scientific practices and institutions. These include precision agriculture methods, which apply new technologies and tools to reduce inputs, costs, and the environmental impacts of agriculture. For example, precision tools can be used to increase input efficiency and thereby reduce fertilizer, pesticide, and water application. Limited adoption has been blamed on the lack of exposure and information; however, the high costs of new tools such as sensors, modified farm equipment, and computers has also been shown to hinder application (Napier et al., 2000).

Sociologists and others are also beginning to examine the rapidly developing application of robotic technology that has started to reshape agricultural production systems. For example, Automatic Milking Systems (AMSs) were introduced in 1992 and have been adopted throughout Western Europe, Canada, and the United States. Milking robots have reshaped relationships between humans, dairy cows, and technology – reducing the need for human labor, allowing cows to determine when they will be milked, and providing a wealth of new data collected and analyzed by sensors and computers. While AMS continue to be adopted by dairy producers, engineers are also working to develop other uses for robots in agriculture such as 'robotic farmhands' that can test soil and water conditions and apply pesticides and fertilizers as needed. Sociologists continue to examine how new innovations in agriculture, such as AMS, change the relationships between humans, technology, and the other biotic organisms (e.g., livestock, crops, bacteria) in production systems.

Some research has also been done on the role of extension services in using technoscience to resolve problems faced by farmers, usually in ways that require little change from the status quo. Hence, extension agents are generally able to grapple better with underproduction than with overproduction. The former requires new, more productive material technologies that have social technologies concealed in them, while the latter usually requires new overtly social technologies – something that extension agents generally try to avoid. Hence, extension agents tend to promote new technologies in part by presenting them as solutions to problems faced by farmers who they see as their primary clientele. In contrast, technologies that might aid farm workers or consumers generally receive less attention. As a result, extension services have played a part in the gradual reduction of the farm population to less than 2% of the total in the United States, even as they have enhanced productivity largely along industrial and commodity lines.

A final area of growing interest to sociologists (and also to economists and philosophers) is the rapid growth of standards, certifications, and accreditations of food and agricultural products and processes. These standards are usually promulgated and enforced by private companies, industry organizations, or private voluntary associations as means of promoting their respective objectives; as such they are alternatives to or supplements for governmental regulations. Often, these standards and certifications are designed to measure technoscientific properties of products and processes (e.g., food safety, Hazard Analysis and Critical Control Points) or use science-based measures to certify that products meet certain standards (e.g., Brix levels in grapes, acceptable colors in fruits and vegetables). Among the many studies in this emerging area are those concerned with (1) particular standards-setting bodies, such as the Codex Alimentarius (Lee, 2009); (2) the consequences of particular standards and/or certifications (Daviron and Ponte, 2005); and (3) the implications of the shift from public to private modes of governance in the agrifood sector (Kimura, 2013).

Philosophy

In recent years, applied ethicists have begun to take an interest in the agricultural sciences, asking a variety of ethical questions about the nature of the research enterprise, its relation to larger

environmental issues such as the conservation of biological diversity, and the distribution of the products generated by the agricultural sciences. Some have asked whether it is even possible to engage in applied science without considering the ethical issues raised by the research agenda. Two interdisciplinary professional societies emerged out of that interest: The Agriculture, Food and Human Values Society and the European Society for Agricultural and Food Ethics. Each now regularly publishes a journal that addresses these issues.

In a major contribution to the field, Thompson asserts that agriculture is dominated by what he calls the 'productionist ethic', the belief that production is the sole metric for ethically evaluating agriculture (Thompson, 1995). From this perspective, derived from the philosophical work of John Locke, land not in cultivation is wasted. Agricultural scientists, themselves often from farm backgrounds, have posited this as self-evident. This was combined with a positivist belief in the value-free status of science and a naive utilitarianism that assumes that all new technologies adopted by farmers are ethically acceptable. From this vantage point, all distributive issues are to be resolved by making the pie bigger. Similarly, environmental problems are defined as arising from inadequate technologies. In contrast, Thompson proposes an ethic of sustainability in which agricultural production is embedded in environmental ethics. He links this to Aldo Leopold's land ethic, which stresses that our notion of community needs to be expanded so as to include the land, place, and other living inhabitants. That said, he notes that the quest for sustainable systems will be filled with both ironies and tragedies (Thompson, 2010).

Indeed, within the agricultural sciences, ethical concerns have a higher profile than they have had in the past. In most industrialized nations, there has been greater recognition of the need for including ethics and public policy issues in agricultural scientific education and research. Moreover, the challenges to the focus on production from within the agricultural sciences have increased receptivity to ethical and policy questions. For example, agronomy, once the province only of scientists concerned with enhancing annual yield, has become more fragmented as those concerned with sustainable agriculture and with molecular biology have entered the field. Thus, questions of the goals and practices of research, previously ignored, have moved closer to center stage (McIntyre et al., 2009). Indeed, a recent book talks of 'political agronomy' (Sumbert and Thompson, 2012).

The application of technologies to support productionist agricultural goals fosters what Thompson (2001) calls 'optimization' – the balancing of benefits and harms. However, in many cases, industrial companies reap the benefits (profits) while externalizing harms such as environmental degradation and food-borne illness. 'Technophilia' in agricultural production systems continues to shape organizational design and responses to emerging problems. This raises important ethical questions regarding unwavering loyalties to productionist ideologies, profitability, and technological solutions to problems. Short-term technological fixes may only temporarily reduce the symptoms of larger systemic problems in the food system.

Philosophers have also examined ethical aspects of the new agricultural biotechnologies. They note that in addition to nutritional qualities, food is always associated with cultural norms and rituals; hence, it occupies a special place with respect to ethics. Among the many issues of relevance is that of informed consent. In brief, it is often argued that consumers have the right to know what is in their food and to make decisions about what to eat on the basis of that information. From this vantage point, those nations that do not label biotechnologically altered foods violate important ethical norms. This appears to be poorly understood by the scientific community as evidenced by the recent decision by the American Association for the Advancement of Science to argue against such labeling.

In addition, critics of biotechnology have raised questions about the ethics of the use of animals in laboratory experiments, the development of herbicide-resistant crops, the use of bovine somatotropin to enhance milk production in dairy cows, the insertion of toxins from *Bacillus thuringiensis* to create insect resistance in maize and potatoes, and the establishment of intellectual property rights in plants and animals.

Future Directions

As environmental concerns have taken on greater significance for the general public, studies of agricultural science have begun to merge with environmental studies. For example, with growing concerns over climate change, there has been nascent interest in agricultural economics, history, sociology, and philosophy in examining its likely consequences for agriculture and agricultural research. Climate change is expected to shift cultivation patterns, crop and livestock populations, and farm practices as some regions become hotter, others become colder, and variability in weather increases. For example, Canadian prairie farmers are now shifting from wheat to maize production as longer, warmer summers permit production of maize with its much higher yields. In contrast, farmers in East Africa may shift from maize to more drought-tolerant sorghum production as rainfall variability increases. These shifts will require major changes in research programs and projects; they will also open new fields of research on agricultural research itself. In addition to research to facilitate climate change adaptation, scientists are also exploring ways to support climate change mitigation on agricultural lands through increasing carbon sequestration and reducing greenhouse gas emissions. This area of research is likely to expand and involve a variety of academic disciplines.

There is little evidence that the fragmentation that has plagued studies of the agricultural sciences in the past is coming to an end, although the entry of genetics and genomics into the plant and animal sciences has created new links with fields outside the traditional agricultural sciences. Nevertheless, disciplinary boundaries between the relevant academic fields remain relatively impermeable. Moreover, there are rigid institutional boundaries that still separate academic agricultural sciences from other fields of research. In particular, agricultural research and education tend to be found in specialized institutions, partly because of the specialized activities in which they engage, and partly because of the high cost of animal herds and experimental fields. Furthermore, those who study the

agricultural sciences often do so from within the confines of schools and colleges of agriculture. In some institutions, this puts restrictions on what topics are considered appropriate for research.

See also: Agrarian Political Economy; Automation: Organizational Studies; Biotechnology; Colonization and Colonialism, History of; Commodity Chains; Development: Social-Anthropological Aspects; Environmental Sciences; Nanotechnology: Social and Cultural Aspects; Patenting in the Biological Sciences; Rural Sociology; Science Communication; Science and The State.

Bibliography

Alston, J.M., Andersen, M.A., James, J.S., Pardey, P.G., 2011. The economic returns to U.S. public agricultural research. American Journal of Agricultural Economics 93, 1257–1277.

Barnett, T., 1977. The Gezira Scheme. Frank Cass, London.

Burmeister, L.L., 1988. Research, Realpolitik, and Development in Korea. Westview Press, London.

Busch, L., Lacy, W.B., 1983. Science, Agriculture, and the Politics of Research. Westview Press, Boulder, CO.

Cochrane, W., 1993. The Development of American Agriculture. University of Minnesota Press, Minneapolis.

Darnhofer, I., Gibbon, D., Dedieu, B. (Eds.), 2012. Farming Systems Research into the 21st Century: The New Dynamic. Springer, Dordrecht.

David, K., Thompson, P.B. (Eds.), 2008. From Bio to Nano: What Can Nanotechnology Learn from Biotechnology. Academic, New York.

Daviron, B., Ponte, S., 2005. The Coffee Paradox: Global Markets, Commodity Trade and the Elusive Promise of Development. Zed Books, London.

Drayton, R.H., 2000. Nature's Government: Science, Imperial Britain, and the 'Improvement' of the World. Yale University Press, New Haven.

Friedland, W.H., Barton, A.E., Thomas, R.J., 1978. Manufacturing Green Gold: Capital, Labor, and Technology in the Lettuce Industry. Cambridge University Press, Cambridge.

Fuglie, K., Ballenger, N., Day, K., Klotz, C., Ollinger, M., 1996. Agricultural Research and Development: Public and Private Investments under Alternative Markets and Institutions. Report 735. USDA, Washington, DC.

Gliessman, S., 2013. Agroecology and food system transformation. Agroecology and Sustainable Food Systems 37, 1–2.

Harwood, J., 2012. Europe's Green Revolution : The Rise and Fall of Peasant-Friendly Plant Breeding. Routledge, New York.

Hayami, Y., Ruttan, V.W., 1985. Agricultural Development: An International Perspective. Johns Hopkins University Press, Baltimore.

Huffman, W., Evenson, R., 2006. Do formula or competitive grant funds have greater impacts on state agricultural productivity? American Journal of Agricultural Economics 88, 783–798.

Kimura, A.H., 2013. Hidden Hunger: Gender and the Politics of Smarter Foods. Cornell University Press, Ithaca.

King, J., Toole, A., Fuglie, K., 2012. The Complementary Roles of the Public and Private Sectors in U.S. Agricultural Research and Development. Economic Research Service, US Department of Agriculture, Washington, DC.

Kloppenburg Jr., J.R., 1988. First the Seed: The Political Economy of Plant Biotechnology, 1492–2000. Cambridge University Press, New York.

Lee, R.P., 2009. Agri-food governance and expertise: the production of international food standards. Sociologia Ruralis 49, 415–431.

McIntyre, B.D., Herren, H.R., Wakhungu, J., Watson, R.T. (Eds.), 2009. Agriculture at a Crossroads: International Assessment of Agricultural Knowledge, Science and Technology for Development. Island Press, Washington, DC.

Napier, T.L., Tucker, M., Robinson, J., 2000. Adoption of precision farming within three midwest watersheds. Journal of Soil and Water Conservation 55 (2), 135–141.

Perkins, J.H., 1997. Geopolitics and the Green Revolution. Oxford University Press, New York.

Plucknett, D.L., Smith, N.J.H., Ozgediz, S., 1990. Networking in International Agricultural Research. Cornell University Press, Ithaca.

Rogers, E.M., 1995. Diffusion of Innovations. Free Press, New York.

Storer, N.W., 1980[1961]. Science and Scientists in an Agricultural Research Organization: A Sociological Study. Arno Press, New York.

Sumbert, G., Thompson, J., 2012. Contested Agronomy: Agricultural Research in a Changing World. Routledge, New York.

Thompson, P.B., 1995. The Spirit of the Soil: Agriculture and Environmental Ethics. Routledge, London.

Thompson, P.B., 2001. Risk, consent and public debate: some preliminary considerations for the ethics of food safety. International Journal of Food Science and Technology 36, 833–843.

Thompson, P.B., 2010. The Agrarian Vision: Sustainability and Environmental Ethics. University Press of Kentucky, Lexington.

Trouchea, G., Lançon, J., Acuña, S.A., Briones, B.C., Thomas, G., 2012. Comparing decentralized participatory breeding with on-station conventional sorghum breeding in Nicaragua: II. Farmer acceptance and index of global value. Field Crops Research 126, 70–78.

Relevant Websites

http://www.ers.usda.gov/ – Economic Research Service, USDA.

http://www.ifr.ac.uk/waste/ – Food Chain Sustainability.

http://www.fao.org/ag/againfo/themes/animal-welfare/en/ – Gateway to Farm Animal Welfare.

http://www.egfar.org/gcard – Global Forum on Agricultural Research.

http://agassessment.org/ – International Assessment of Agricultural Knowledge, Science and Technology for Development.

http://www.nationalaglawcenter.org/research/#wtobiotech – National Agricultural Law Center.

http://www.pesticideinfo.org/ – PAN Pesticide Database.

Agriculture, Economics of

Bruce L Gardner, University of Maryland, College Park, MD, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 1, pp. 337–344, © 2001, Elsevier Ltd.

Abstract

Agriculture is predominant in primitive economies, but in industrial countries now it accounts for as little as 2% of economic activity. Malthusian worries have been forestalled as food has become very cheaper in real terms, though at the cost of dislocation of redundant farmers and economic problems for remaining agricultural producers. Productivity gains fueled by advances in technology have been the main driving force. Governments have promoted technological changes through investments in research, infrastructure, and institutions fostering market development, and have had to respond politically to situations of both food scarcity and surplus. In many instances, especially in developing countries, agriculture has been regulated and taxed in the pursuit of nonagricultural political interests, while in industrial countries agriculture is substantially subsidized. Agricultural economists have made progress in explaining many aspects of the interactions among technology, markets, farm incomes, and both domestic and international trade policies involving agricultural commodities and rural people, but many key determinants and trends remain only imperfectly understood.

Agriculture is distinguished from other sectors of the economy by virtue of its production processes (biological), its economic organization (on farms), and its products (food and fiber). The importance of these distinctions for economic analysis is not always evident, but they have been sufficient to make agricultural economics a separate subdiscipline of economics, with its own journals and professional organizations.

Agriculture's Primacy in Economic Development

In most of the world historically, and in much of the world today, the economics of agriculture is the economics of subsistence: the effort to wrest the food necessary for survival from productive but fickle resources. The essential economics concerns how individuals carry out such efforts and how families, villages, or other social entities organize their members for doing so. Economic development begins when agriculture generates production in excess of local requirements. Until the mid-nineteenth century the majority of the labor force in most countries of Europe was employed in agriculture. By the end of the twentieth century this percentage had been reduced to less than five in the richest countries. Similar patterns have emerged since 1950 in much of Latin America and Southeast Asia. Nonetheless, the World Bank (1997) estimates that 72% of the world's poor live in rural areas, and the prospects for economic development in agriculture remain a matter of worldwide concern.

A long-debated issue is whether agriculture is best viewed as an engine of growth, with investment in the sector an important source of economic progress or as an economically stagnant source of labor to be mobilized more productively elsewhere as the economy grows. 'Dual economy' models, in which agriculture is economically distinct from the nonagricultural sector, can accommodate both views, depending on how they treat mobility of labor and capital between the sectors, and the processes of technical change and investment in each. Such models can account for the observation of huge outmigration from agriculture together with wage and income levels in rural areas rising toward urban levels after falling behind in the early stages of industrialization. But they do not provide useful empirical guidance for fostering economic development in areas of the world where it still is the most needed. For those purposes, attention to microeconomic and sectoral detail is necessary. For reviews of economists' work on microlevel and aggregate questions, respectively, see Strauss and Duncan (1995) and Timmer (2002).

One of the most striking and still, to some extent, controversial finding about the economics of traditional agriculture is the wide extent to which farmers in the poorest circumstances in the least developed countries act consistently with basic microeconomic principles. They follow economic rationality in the sense of getting the most economic value possible with the resources at hand; but the innovation and investment that would generate economic growth are missing (Schultz, 1964). What is needed is to break out of the poor but efficient equilibrium by means of 'investment in high income streams,' mainly physical capital and improved production methods embodying new knowledge, and investment in human capital that would foster innovation in technology and effective adoption of innovations by farmers. Events such as the 'green revolution' that boosted wheat yields in India in the 1960s showed promising trends that have been sustained in many areas, but agriculture remained moribund through the 1990s in many places, notably in Africa and the former Soviet Union. Agricultural economics faces no more important task today than explaining and finding remedies for this stagnation.

Farms

Farms range from individuals working small plots of land with only primitive tools to huge commercial enterprises. Every operating farm embodies a solution to problems of product choice, production technique, mobilization of inputs, and marketing of output. Many of the choices to be made involve

nonmarket, household activities. Consequently, the economic analysis of farm households has become a major area of empirical investigation, calling upon developments in population and labor economics as well as the theory of the firm. Using these tools, agricultural economists have attempted to understand alternatives that have arisen in the economic organization of agriculture: family farms, cooperatives, plantations, corporate farming, state farms and so on (see for example Binswanger and Rosenzweig, 1986).

Organization of Production

A basic decision is whether to specialize or to diversify production among a number of products. The trend is strongly toward specialization. For example, 4.2 million US farms (78% of all farms) had chickens in 1950, but by 1997 specialization had gone so far that only 100 000 did (5% of all farms). Linked with specialization is the issue of economies of scale in farming. Throughout the developed economies there is a general tendency for farm size to increase over the last century. Data on farmers' costs indicate that a primary reason is economies of size. Yet there are many instances of very large farms failing. Collective farms in the former Soviet Union, employing thousands of workers on thousands of hectares, became paradigms of inefficiency. And in some developing country contexts there is evidence that small farms use their resources more efficiently than large ones. Optimal economic organization with respect to both specialization and scale depend on technical and institutional factors, most importantly the following.

Land Tenure

Land is a valuable asset and is necessary for farming. Yet farmers in many countries are poor. Institutional arrangements have evolved to enable farmers to cultivate and claim returns from land they do not own. The main ones are cash rental and sharecropping. Cash rental encounters several problems: the tenant may lack the means or access to credit for payment in advance, bears all the risks of crop failure or low prices, and has an incentive to use the land in ways that increase current output at the expense of future fertility of the land.

Under sharecropping, a common practice in both developing and industrial countries, the tenant pays after the harvest in the form of a fraction of the crop harvested. The share paid to the landlord varies widely, generally between one-fourth and one-half of the crop, depending on the quality of the land; the labor intensity of the crop; and the value of nonland inputs, if any, contributed by the landlord. In addition, the literature on optimal land contracting finds that shares depend on agency costs, production efficiency of tenants and landlords, and how risk averse each party is. Sharecropping divides production and price risk between landlord and tenant, and obviates the need for payment in advance. But it retains the principal–agent problem in lacking incentives to maintain future land quality, adds a disincentive for tenant effort in that the tenant receives only part of the tenant's marginal product, and adds an incentive for the tenant to underreport output and/or price received so as to reduce the rent paid. Such problems can be dealt with through monitoring, but that is costly. For a comprehensive review of the issues, see Deininger and Feder (2001).

The problems and costs of land rental increase the attractiveness of owner-operated farms, even if they have to be smaller. However, in many countries the institutions for private land ownership are not fully developed, nor are credit markets that would enable people with few initial assets to become landowners. In developed economies, land rental functions as a mechanism through which farmers can mobilize the land resources needed to achieve the least-cost scale of production. In the US in 1997, only 21% of cropland was on farms fully owned by their operators.

Agricultural Labor

About one-half the world's labor force works in agriculture, as either a farmer or a hired worker. (For data by country, see World Bank, *World Development Indicators*, 1998, **Table 1**.) Hired labor is common even on family farms. Hired farm laborers in both developing and industrial countries are among the least well paid and most economically precarious workers. Seasonal workers live under especially difficult conditions in that they often dwell in temporary quarters and are minorities or immigrants, sometimes with dubious legal status, which makes them ripe for exploitation. The plight of migrant workers in many countries has led to legislative and regulatory attempts to limit their numbers and improve their condition, and adds to a general sense that policies should be undertaken to enable landless laborers to gain access to land of their own and become farmers themselves. Nonetheless, hired labor remains a substantial fraction of the farm labor force in both rich and poor countries.

Credit and Input Markets

As agriculture modernizes, increased share of resources used are purchased seeds, fertilizers, chemicals, energy, and capital equipment. Farms that cannot invest become unable to compete effectively. If modern agriculture is to be undertaken by farmers other than those who already possess substantial assets, well functioning credit markets are essential. A major problem for agriculture in many countries is limited access to either purchased inputs or credit. Recent thought about credit markets has emphasized the problems that arise because of asymmetric information between lender and borrower leading to credit rationing or missing markets. If potentially productive loans do not get made, farmers and the rural economy are unable to grow to their potential. This reasoning has led many countries to provide subsidized credit to farmers, but the informational problems that cause market failure have not

Table 1 Annual costs and benefits of protection of agriculture in the EU, US, and Japan

	EU	*US*	*Japan*
Consumer costs due to higher prices	35	6	7
Taxpayer costs of subsidies	12	10	0
Gains to producers	31	12	12

been overcome with government involvement. In addition, the political provenance of these programs causes new problems.

Price Determination and Marketing

A recurrent complaint worldwide is farmers' lack of market power as compared to those who buy from and sell to them. Farmers typically have only a few alternative outlets for their products, and inputs they buy, but the extent of monopsony or monopoly power that results remains unclear. In many countries farmers have established marketing and purchasing cooperatives to increase their market power. In the United States, the first half of the twentieth century saw far-reaching governmental attempts to reduce the market power of meat packers, grain traders, railroad owners, food wholesalers and retailers, and banks through antitrust action and governmental regulatory agencies. The developed countries of the world are today replete with such efforts, and developing countries have followed suit as appeared technically and politically feasible. It is, nonetheless, unclear whether the economic problems of farmers have ever been principally attributable to their lack of market power, or that cooperatives or regulatory institutions have increased farm incomes appreciably.

Important recent developments in marketing involve contractual arrangements between farmers and processors that take some input provision and marketing decisions out of the hands of the farmer. Such changes have gone furthest in broiler chicken production in the United States, where the processor is an 'integrator' who supplies the baby chickens, feed, veterinary and other services, technical information, and perhaps credit. The farmer (or 'grower') receives a payment schedule, contracted for in advance, consisting of a fee per pound of chicken delivered that is adjusted for an efficiency indicator as compared to other growers (but not changes in the market for chicken) in return for the grower's effort in feeding and managing the flock and providing the properly equipped chicken house. Virtually all broilers in the country are now produced under some variant of this type of contract. Under these arrangements productivity indicators of output per unit input have grown far faster for broilers than for any other livestock product, and the US price per pound (live basis) of chicken relative to beef has declined from a ratio of 1.7 in 1940 to 0.5 in 1995. Similar production arrangements are increasingly prevalent for other meat animals.

Risk Management

In subsistence agriculture, crop failures or livestock deaths place the farmer at risk of starvation. In commercial agriculture, fixed costs of crops sown and interest on debt means that losing even a portion of the crop, or receiving low prices, can easily generate negative cash flow. Steps that a farmer can take to manage such risk include savings, diversification of enterprises, emergency borrowing, and purchase of hazard insurance against output risk, or some form of forward pricing against price risk. It remains open to question, however, how risk averse farmers are. Basic evidence that risk aversion is important is farmers' willingness to pay for insurance and their interest in pricing their output in advance. Observations that give pause about the importance of risk aversion are the many farmers who do not buy even subsidized crop insurance and who do not attempt to lock in a price for their output, even when contractual means for doing so are available. Nonetheless, evidence from developing countries suggests risk aversion of a magnitude that could readily impair farmers' willingness to invest in new production methods even when innovation would pay in expected value (see Moschini and Hennessy, 2001).

Production and Technology

The evolution of world agriculture over the long historical record is tied principally to changes in technology. Throughout the developed world, a large and sustained record of growth in agricultural productivity has been achieved. In the US, after 50 years of steady but unspectacular growth, agricultural productivity accelerated markedly after 1940 to a pace of about 2% annually, well above the rate of productivity growth in manufacturing. Moreover, that rate of growth has been maintained for 60 years, with little evidence of the productivity stagnation that plagued manufacturing in the 1970s and 1980s (**Figure 1**).

Economists have devoted much effort to measurement and analysis of productivity changes and farmers' decisions about input use. Nerlove (1958) developed a method of estimating both short-run and long-run output response to product prices. Empirical work using many variations on his approach over the last four decades has estimated generally small short-run effects of price. But in many cases the long-run effects are substantial. Griliches (1957a) provided the first fully developed economic analyses of the adoption of technology in his study of hybrid corn in the United States. Technical change and supply response have been merged in studies of 'induced innovation.' The chief causal factors identified in both supply and productivity growth have been advances in knowledge, improved input quality, infrastructure development, improved skills of farmers, and government policies. But the relative importance of these factors, even for the most-studied countries, still remains in doubt. Sunding and Zilberman (2001) is a comprehensive review of these factors.

Demand and Markets

The world's population tripled in the twentieth century from the two billion of 1900. Agricultural production grew sufficiently not only to feed an additional four billion people, but also to provide the average person with a substantially improved diet. And, the incidence of famine and starvation among the world's poor has been greatly reduced. This capability was not evident 200 years ago when Malthus formulated the proposition that the earth's limited production capacity, coupled with the propensity of population to grow whenever living standards rose above the subsistence level, meant the inevitability of increasing food scarcity (and worse) over the long term. One of the most notable facts about the twentieth century is the failure of Malthusian pessimism to materialize. Nonetheless, the plausibility of elements of this view – basically the fixity of natural resources in the face of

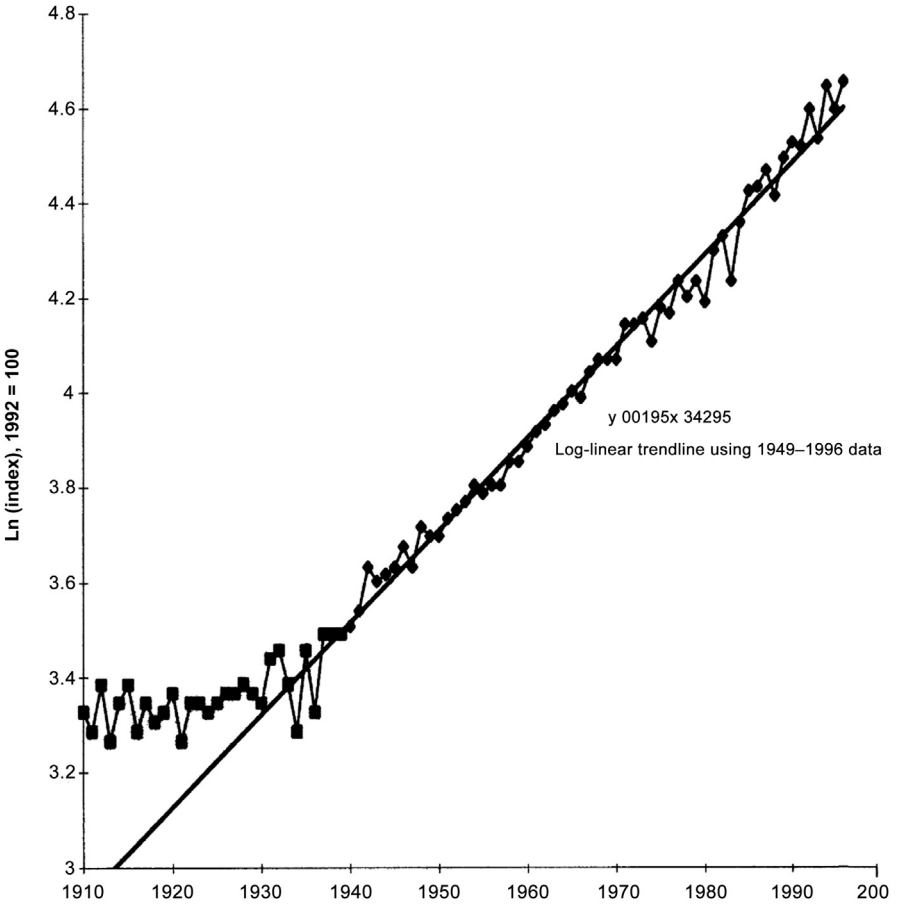

Figure 1 US farm total factor productivity index.

increasing population – is sufficient that the Malthusian worry resonates to the present day. It is therefore important to establish the circumstances under which food scarcity has ceased to be a salient social problem as well as the situations in which scarcity and famine remain a major cause of distress, and to understand why supply and demand have conspired to work out predominantly in the counter-Malthusian direction.

The single best indicator of food scarcity is the real price of staple commodities: cereals and other basic foods. Despite price spikes in wartime and the 1970s, the trend is for ever cheaper commodities. This trend primarily reflects lower real costs of production, a consequence of the productivity trends illustrated in Figure 1. While all acknowledge the uncertainty of any forecast, expert participants in a recent comprehensive assessment of world food prospects were in broad agreement that the trend of lower real prices of staple food commodities is most likely to continue in the twenty-first century (Islam, 1995).

An important factor in food demand is Engel's Law: the share of income spent on food decreases as consumers' incomes rise. The general rise of real incomes over the last century, coupled with the growth in agricultural productivity have meant an inexorable decline in agriculture's economic importance, and have been a source of chronic downward pressure on the economic returns of farmers. In many developing countries, especially former colonies whose economies became attuned to exports of primary products, declining commodity prices have been a key part of a bigger story of economic disappointment. Economic problems of farmers in both developing and industrial countries have kept agriculture firmly on the policy agenda almost everywhere.

Government and Agriculture

Political responses to problems of agriculture have generated a wide variety of government action. Four areas of activity warrant discussion: regulation of commodity markets, rural development policy, food policy, and resource and environmental policies.

Commodity Programs

Government intervention in agricultural commodity markets has been pervasive throughout recorded history. The primordial form of this intervention is taxation. With urbanization, implicit taxation of agriculture has arisen in many countries in the form of regulations intended to keep food prices from rising in times of scarcity. A sharp divide exists between the developing world, in which agricultural output is generally

taxed, and the industrial world, in which agriculture is generally subsidized. This pattern of taxation and subsidy has had the unfortunate consequence of encouraging overproduction in industrial countries and discouraging investment in agriculture in developing countries, many of which have a comparative advantage in agriculture. Contrary to what one might have expected, the share of world agricultural exports accounted for by industrial countries increased from 30% in 1961–63 to 48% in 1982–84, with a corresponding decrease in developing countries (World Bank, 1986: p. 10).

The protection of agriculture in industrial countries harm agriculture in developing countries, and also each industrial country's protection makes it more costly for other industrial countries to maintain protection. The Common Agricultural Policy (CAP), created with the establishment of the European Economic Community in 1958, is notorious in this respect. The main policy instruments of the CAP go back to Britain's Corn Laws of the ninetieth century tariffs that maintain protection against imports by rising when world prices fall ('variable levies') and export subsidies to dispose of domestic surplus production (see Ritson and Harvey, 1997). In the first two decades of its existence the CAP moved its members from being net importers to net exporters of wheat, rice, beef, and poultry meat. Other grain-growing countries that also desired to maintain support prices for their producers introduced or accelerated export promotion and subsidy programs of their own, notably the US Export Enhancement Program of the 1980s. The subsidy competition exacerbated a worldwide decline in commodity prices in the 1980s, increasing the costs of US 'deficiency payments' that made up the difference between legislated 'target' prices and market prices for grains. This in turn triggered massive acreage-idling programs; in 1985–87 about a fourth of US grain-growing land was idled.

The World Bank (1986: p. 121) assessed the annual costs and benefits of agricultural protection in the largest Organisation for Economic Co-operation and Development countries as shown in **Table 1** (in billions of dollars).

Note that the costs to consumers and taxpayers together far outweigh the producer (more specifically, landowner) gains, with the sum for the EU, US, and Japan being a net welfare loss of US$25 billion.

Accurate measurement of these gains and losses is difficult, but virtually all analysts estimate substantial net losses in the industrial countries and to producers in developing countries during most of the post–World War II period, and accelerating losses in the 1980s. This situation provided the stimulus for agricultural policies, after lengthy and contentious negotiations in 1986–93, to be subjected to internationally agreed disciplines that began to be implemented in 1995 under the auspices of the World Trade Organization.

Individual countries have also initiated moves toward less market distorting intervention in the commodity markets in the 1990s. In the developing world, substantive steps in deregulating commodity markets were taken in many countries of Latin America and East Asia; and in Africa many countries reformed and/or abolished marketing boards and related interventions. Most radical of all, beginning in the late 1980s (and before the breakup of the Soviet sphere in 1989), a renunciation of state control of farm enterprises occurred in China and throughout Eastern Europe and the former USSR. But the reforms have as yet achieved nothing near complete liberalization in developed, developing, or transition economies, with the exception of New Zealand.

Rural Development Policy

A broader agenda of governments in promoting economic growth in agriculture and rural areas has more widespread support. Economists have generally concluded that provision of certain public goods and infrastructure investment has been crucial in the economic development of agriculture and that the absence or deficiency of such governmental support is an obstacle, perhaps an insuperable obstacle, to economic growth in agriculture in countries where it has not yet occurred. The World Bank has taken a strong role in urging market liberalization in developing countries and at the same time proposing a broad program of public investment in pursuit of rural development (World Bank, 1997).

Legal Institutions

The most fundamental economic service the State can provide is a system of law governing property and contracts, and protection from lawbreakers. This requirement is not of course peculiar to agriculture, but must be mentioned because legal institutions in rural areas are not weak in many transition and poor economies, especially regarding use and control of farmland and water resources. In industrial countries, too, these institutions have to evolve in response to changes in technical and social realities, most notably, in the 1990s, the establishment of property rights and contractual procedures that bear on innovations in biotechnology.

Agricultural Research, Extension, and Education

Even with well-established institutions fostering private sector research and development, research and information dissemination are likely candidates for public funding, and have long been so funded in many countries. Griliches (1957b) pioneered methods of estimating the costs and benefits of publicly supported research. Since then hundreds of studies in both developing and industrial countries have replicated his finding of extraordinarily high rates of return to public investment in research and the dissemination of knowledge through extension activities (Evenson, 2001).

Since the work of Schultz (1964: Chapter 12) investment in schooling has been seen as a cornerstone of what is needed to improve the economic well-being of farm people, and of increasing agricultural productivity. Solid empirical evidence of the effects of education on farming has been hard to come by, however. Even so, there is widespread support for improved education in rural areas, recently with particular attention to the education of women. Evidence is strong that schooling improves peoples' earning capacity, so it is a promising remedy for rural poverty even if it causes its recipients to leave agriculture.

Rural Infrastructure

Governments in industrial countries have made major investments in roads, railways, shipping channels, and ports to provide remote areas with cheaper access to markets. Lack

of such infrastructure is a major impediment to agricultural development in many parts of the world today. But we have nothing like the studies of returns to research to provide evidence on the rate of return to such investments, and the anecdotal evidence is replete with failure as well as success stories. Even more controversy swirls around investments in water projects. Irrigation was important in facilitating fertilizer response to the new grain varieties that triggered the green revolution of the 1960s, and is essential for opening up arid areas for production. At the same time, dams and irrigation works have been heavily criticized in recent years. Many cases of low or negative returns to large investments have been cited, and the environmental costs of lost habitat for endangered species and reduced water quality have been emphasized. Recent work by agricultural economists has emphasized improving institutions for water pricing and assignment of use rights more than further investment in large projects.

Food Policy

A chief source of governmental discrimination against agriculture in developing countries is a desire to keep food prices low for urban consumers. In industrial countries, too, attempts have been made via price ceilings and export restrictions to keep a lid on food prices in periods when they have risen sharply, as occurred in the commodity boom of the 1970s. More important ongoing policies address the regulation of food quality and safety, food assistance for poor people at risk of undernutrition, and famine relief. Chemical residues on food and the use of genetically modified organisms in agriculture were especially contentious issues in the 1990s.

An important bifurcation of countries today is between those in which food security remains a pressing national issue and those in which assurance of an adequate diet is the problem of only a small minority of the population. International food aid has become a permanent policy in industrial countries, particularly famine relief. Mobilization of funds for such efforts can be difficult except in cases of well-publicized disasters, but the more salient analytical issues have involved the nature of famines and the effectiveness of alternative approaches to remedy the suffering and death they cause. It has become apparent that in most famines the problem is not so much physical unavailability of food as a lack of income with which to acquire food. This may seem a distinction without a difference but the implications are profound for the most effective administration of aid. It has been argued, for example, that it can be counterproductive simply to ship food products to be disbursed by local governments. The undesired result is distribution that is too often negligible to those who most need the goods, and at the same time a depression of commodity prices and hence of the incomes of local farmers who produce goods that compete on the same market as goods brought in but lack income for an overall adequate diet. Generally, international donors have to be careful not to take actions harmful to local coping mechanisms, which in many poor areas are well developed from long and bitter experiences. See Barrett (2002) for a comprehensive review.

Resource and Environmental Policy

The relationships between agriculture and water quality, soil and other resource depletion, wildlife habitat, and chemical contamination have become front-burner policy issues in industrial countries, and are beginning to get attention in developing countries. A difference from the regulation of industrial polluters is that agricultural pollution sources are typically small, scattered, and difficult to monitor. Certain agricultural pesticides have been banned in industrial countries, but reasonably good substitutes have so far been available. Nonintensive uses of erodible or otherwise environmentally sensitive lands have been fostered in the US and Europe by paying farmers to undertake recommended practices.

Nonpolluting and resource saving practices for developing countries have been promoted by international agencies as conducive to 'sustainability' of their productive resources. However, countries have resisted some of these ideas, such as restraints on opening up new land or eschewing large new dams and irrigation projects. The debate is difficult because of a lack of documentation in which the loss of forests and conversion of other lands to agricultural purposes at rates now occurring is a mistake that will come to be regretted. In Europe and North America, an issue that has become prominent in recent decades is the conversion of farmland out of agriculture and into residential and commercial use in suburban areas, not so much out of concern about lost food production but rather because of the loss of open space and other community amenities that farming provides. Land use regulation and agricultural subsidies of various kinds have been introduced, most extensively in Europe.

Agricultural Politics

Why has agriculture been widely discriminated against in developing countries, and supported in developed countries? Evidence that the explanation is not country specific is that countries that have grown sufficiently rich to move from the developing to the developed category, largely in East Asia, have moved from taxing agriculture to subsidizing it. A large body of recent work has attempted to explain the strength and resilience of farmers' political clout in the richest countries, especially in Western Europe, Japan, and the North America. It is particularly notable in that this strength has been maintained even as the farm population has declined from one-fourth to one-half of the total population 50 years ago to 2–10% today. It is also instructive that some commodities are protected much more heavily than others within each country. Many reasonable hypotheses on these and related matters have been advanced, generally linked to interest group lobbying and democratic politics. Knowing more about agricultural politics is important because a governmental role is essential in many aspects of agricultural and rural development, yet governmental action in commodity support programs, trade restrictions, and other regulatory areas have imposed large social costs that are notably resistant to reform. The goal is governmental institutions that provide the services that contribute to sustainable development and that reform wasteful policies. The goal is far from being realized.

See also: Agricultural Sciences and Technology; Agroecology and Agricultural Change; Economic Geography; Food Production, Origins of; Food Security and 'Green Revolution'; Rural Geography; Rural Planning: General; Sustainable Agriculture.

Bibliography

Barrett, C., 2002. Food security and food assistance programs. In: Gardner, B., Rausser, G. (Eds.), Handbook of Agricultural Economics. Elsevier Science, Amsterdam.

Binswanger, H.P., Deininger, K., 1997. Explaining agricultural and agrarian policies in developing countries. Journal of Economic Literature 35, 1958–2005.

Binswanger, H.P., Rosenzweig, M., 1986. Behavioral and material determinants of production relations in agriculture. Journal of Development Studies 22, 503–539.

Deininger, K., Feder, G., 2001. Land institutions and land markets. In: Gardner, B., Rausser, G. (Eds.), Handbook of Agricultural Economics. Elsevier Science, Amsterdam, pp. 287–331.

Evenson, R., 2001. Economic impact studies of agricultural research and extension. In: Gardner, B., Rausser, G. (Eds.), Handbook of Agricultural Economics. Elsevier Science, Amsterdam, pp. 573–627.

Gardner, B.L., 1992. Changing economic perspectives on the farm problem. Journal of Economic Literature 30, 62–101.

Griliches, Z., 1957a. Hybrid corn: an exploration in the economics of technical change. Econometrica 25, 501–522.

Griliches, Z., 1957b. Research costs and social returns: hybrid corn and related innovations. Journal of Political Economy 66, 419–431.

Islam, N. (Ed.), 1995. Population and Food in the Early Twenty-first Century. International Food Policy Research Institute, Washington, DC.

Johnson, D.G., 1991. World Agriculture in Disarray, second ed. St. Martin's, New York.

Moschini, G., Hennessy, D., 2001. Uncertainty, risk aversion, and risk management by agricultural producers. In: Gardner, B., Rausser, G. (Eds.), Handbook of Agricultural Economics. Elsevier Science, Amsterdam, pp. 87–153.

Nerlove, M., 1958. The Dynamics of Supply. Johns Hopkins University Press, Baltimore, MD.

Ritson, C., Harvey, D., 1997. The Common Agricultural Policy and the World Economy, second ed. CAB International, New York.

Schultz, T.W., 1964. Transforming Traditional Agriculture. Yale University Press, New Haven, CT.

Strauss, J., Duncan, T., 1995. Human resources: empirical modeling of household and family. In: Behrman, J., Srinivasan, T. (Eds.), Handbook of Development Economics, vol. 3b. Elsevier Science, Amsterdam, pp. 1883–2024.

Sunding, D., Zilberman, D., 2001. The agricultural innovation process. In: Gardner, B., Rausser, G. (Eds.), Handbook of Agricultural Economics. Elsevier Science, Amsterdam, pp. 207–261.

Timmer, C.P., 2002. Agriculture and economic development. In: Gardner, B., Rausser, G. (Eds.), Handbook of Agricultural Economics. Elsevier Science, Amsterdam.

World Bank, 1986. World Development Report. Oxford, UK.

World Bank Group, 1997. Rural Development: From Vision to Action. Environmentally and Socially Sustainable Development Studies and Monographs Series 12. World Bank, Washington, DC.

Agroecology and Agricultural Change

Charles A Francis, University of Nebraska – Lincoln, Lincoln, NE, USA
Alexander Wezel, ISARA-Lyon, Lyon, France

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G.D. Stone, volume 1, pp. 329–333, © 2001, Elsevier Ltd.

Abstract

Agroecology has been defined as the marriage of agriculture and ecology, and has evolved to embrace whole food systems including production, economics, environmental impacts, and social implications of change. This field now broadly addresses the *ecology* of food systems. Agroecology has developed in response to awareness of declining resource availability and negative impacts of conventional chemical agriculture. With people's growing concern of importance of whole systems research and development, agroecology has now an emerging focus on food production and food systems, and is a vital engine for conceptualizing and implementing responsible agricultural change.

Introduction

Agroecology grew from concerns about long-term sustainability of food production, as many people recognized diminishing supplies of fossil fuels, fresh water, phosphorus, and other key resources and the impacts of these changes on agriculture. Another current reality is the loss of land with productive soils for food production, as soils are increasingly degraded, topsoil is lost, fields accumulate salts, and large areas are appropriated for other human pursuits. Future scarcities of fossil fuel energy and irrigation water for agriculture are accelerating due to demands from other sectors, but less well known is the change in land use to purposes other than food production. In addition to land dedicated to biofuel production instead of food, thousands of hectares are converted each year to industrial use, housing, roads, and other urban infrastructure (Olson and Lyson, 1999). The challenge of producing much more food per hectare by mid-twenty-first century to feed a growing world population will require an intensification of production on available farmland, as well as methods to rehabilitate land that is degraded or lost. We must solve major problems related to food storage and distribution, access to food for poor people, and changes in diets to make everyone more healthy.

Green revolution technologies have provided impressive gains in food production. The higher yields achieved over the second half of the twentieth century have contributed to lower food prices and improved human well-being in many parts of the world. They could only be sustained if research reveals more efficient ways to use scarce resources and achieve resilience in the face of looming changes in climate, especially greater incidence of major storm events that interrupt industrial monoculture agricultural practices. Increased yields of crops have indeed been one of the 'modern miracles' from application of practical science in agriculture, but limitations are on the horizon (Hazell, 2002).

Major yield increases have been achieved by higher crop genetic production potential, applications of high levels of chemical fertilizers and pesticides, and spread of irrigated farming practices. To achieve high yields, farmers mainly use practices that depend on heavy reliance on inputs from off the farm, and that increasingly are based on specialization and large areas in monoculture crops. In more fertile areas, this technology-driven change to an industrial model of farming has produced global grain harvests that have almost kept up with increasing human population and demand for food. Yet there is serious concern that we are reaching the genetic plateaus for yields of major cereal crops, e.g., maize, wheat, and rice (Grassini et al., 2013), limiting future potential for increased yields of these essential crops. Advances through science have contributed to confidence that new technologies will always overcome resource shortages, and have created an attitude of complacency that diverts thinking from resource scarcity and limits to growth. Thoughtful evaluation of resource investment in ever more intensive and expensive technologies and reliance on a singular industrial strategy in agriculture raises concerns about food sustainability. Agroecology addresses these concerns and provides a platform for seeking alternatives.

There are growing concerns with how to counter the often negative environmental impacts of intensive use of fertilizers and pesticides that are integral to industrial technologies. There is further concern that current improved technologies have reached only those farmers with adequate financial resources to access the fruits of science. In reality nearly half the food available today globally is produced by small farmers with traditional methods, and this is unlikely to change in the near future (IAASTD, 2008). This report from the United Nations included contributions of 500 scientists over a period of 5 years, who concluded that agroecological methods that improve intensive systems used by small farmers can provide one solution to food challenges for many rural families. Agroecology combines methods from biophysical and social sciences to look at how and where food is produced, how it follows a chain of transformations to reach the consumer, and how this contributes to healthy diets and human well-being.

Development experts also recognize that many present challenges are not due to inadequate food production, but rather result from economic and political priorities. Current inequities in food availability within countries and regions suggest that our present system is not meeting human needs. Conservative estimates are that 30–40% of food produced does not reach the table due to losses in harvest, storage, processing, and the marketing system as well as being wasted in consumers'

homes (IMechE, 2013). In developing countries there is often inadequate or faulty storage capacity, while in developed countries much food is discarded because of blemishes or excessive caution about food safety. Agroecology promotes inquiry into these questions that can only be answered through analysis of the whole food system including who has access to food and who benefits. Therefore, not only agronomic and ecological, but also many economic and social factors have to be taken into account to develop the whole human production and consumption system – from 'field to the plate.' Appropriate analyses include socioeconomic factors that are closely integrated into the biophysical processes involved in agriculture, clearly a human activity system.

Background

Holistic systems planning has been central to the thinking of farmers for centuries, as they make decisions based on available land and production resources, balance land use for crops, livestock, and meeting needs of the family while producing additional food for sale or barter. History of the term 'agroecology' to describe complexity in agricultural systems and their interactions with the environment was described by Gliessman (2007 and prior edition) and Wezel and Soldat (2009). Broadening use of this term to describe the ecology of food systems – including production, processing, marketing, food consumption – is a recent event (Francis et al., 2003). Wezel et al. (2009) have further summarized and categorized agroecology as a science, as a set of practices, and as a movement. Multiple interpretations and diverse uses of the term agroecology complicate understanding and communication. What is agreed is the importance of current and potential production of food, economics of systems that make them profitable for farmers, environmental impacts of this human activity system that puts at risk quantity and quality of food production in the long term, and social consequences of different types of agricultural systems and how they influence food security and food sovereignty (Francis et al., 2013).

Science of Agroecology

Growing appreciation of the complexity of challenges facing those who produce, process, and market food has led to new approaches to study whole systems. Many of the early advocates of linking ecology with agriculture came from the sciences, and first use of the term was by Bensin (1935, and prior citations) in U.S.A. who talked about ecology related to commercial agriculture. The German ecologist Tischler (1950) used agroecology to discuss problems in pest management and soil biology, and recognized the importance of interactions in complex systems. Altieri (1983) and Gliessman (2007) have pursued the science theme and popularized agroecology as an integral part of the science discourse.

Early application of the concepts of integrated systems was adopted in plant protection and Integrated Pest Management (IPM), in large part a response to Rachel Carson's landmark book *Silent Spring* (Carson, 1964). Development has been especially active in California, U.S.A., with IPM described as a complex method designed to reduce chemical application and improve the environment (http://www.ipm.ucdavis.edu). Definitions of IPM include ecosystem approaches to crop protection that combine multiple management strategies that minimize the use of chemical pesticides (Abrol, 2014). Researchers in organic farming systems are now providing science-based recommendations for improving production (Francis, 2009). One unique application of ecological principles is the research on perennial polycultures, with the objective of creating a multispecies system that produces cereals and legumes for human food as well as forage for grazing livestock (Jackson, 1980). Such systems are designed to produce their own soil fertility and protect against pest damage through functional biodiversity.

Acceptance of agroecology has been less enthusiastic in economics. Most production cost analysis at the farm level concerns short-term expenses and benefits, with focus on cash flow and need to maintain a profitable whole farm operation to stay in business. On a larger scale, we recognize that macroeconomic analyses using national data and contemporary methods that examine consequences of systems in the short-term are inadequate to understand large inequities in current food distribution and availability within a country, for example. Such measures as Gross National Product also disguise substantial differences in distribution of income, and thus access to food, and human welfare. The simplistic and inadequate measurement methods used today to assess economic conditions compel those in research and development to examine new models and metrics in the food system, and suggest that more complex indices could be developed to adequately measure and report economics in terms of human welfare. The emerging perspectives and methods used in agroecology research and education provide useful approaches to look at whole food systems. It should be noted that agroecology is still used in a narrow sense by many scientists to describe applications of ecology in agriculture. Research has provided the foundation for practices described in the next section.

Agroecological Practices

Practical applications of ecological principles include many ways of introducing both spatial and temporal biodiversity into design of cropping and farming systems. When animal traction was important to land preparation, cultivation, and harvest there was extensive planting of hay crops and pastures in small fields that provided species, farm-scale, and landscape diversity. With mechanization, removal of livestock from most farms, and wide planting of a few crops in monoculture, there is far less diversity at all levels of spatial scale and limited use of rotations in present industrial systems. Moving toward greater diversity with some practices that more closely mimic natural ecosystems is one strategy toward generating more fertile soils and protecting against pests without using chemicals.

Crop rotations of cereals and legumes, summer annuals with winter annuals, and perennial species with annual crops all provide diversity in types of root systems, lengths of growing cycle, and residues from crops that contribute to more efficient

nutrient and water use as well as soil fertility. Maize–soybean rotation is a common practice in the U.S. Midwest, but one that has lost favor as price of maize for multiple uses has increased. Likewise, low production costs and ready market for export have caused a large conversion to continuous soybean production in the Pampas Region of Argentina. More complex multiyear rotations that include these important summer crops with winter wheat or other cereals, sunflower, and alfalfa can introduce greater diversity and resilience to cropping systems. A traditional cow/calf grazing system (4–6 years) in rotation with annual crops (3–4 years) provided enough soil nutrients and plant protection to keep this sequence profitable and sustainable for many years in Argentina (Rotolo et al., 2007), although this has been replaced by short-term rotations or continuous cropping of species suitable for processing and export. Studies of more complex rotation systems show that they can be more efficient in resource use and more sustainable in energy terms, and provide examples that can support policies to return to diverse cropping and crop/animal systems (Altieri, 2009).

Agroforestry, sometimes called permaculture, is another system of practices that combines annual crops with perennial plantings and includes a strategy that will maximize internal cycling of nutrients within the system, and in some cases also incorporation of animals (Mollison, 1988). Multiple cropping systems (intercropping) are also valuable strategies used by smallholder farmers throughout tropical and temperate regions (Francis, 1986; Wezel et al., 2014). These combine several crops in strips (strip cropping, relay cropping, alley cropping) or in random plantings that include species diversity and different intensities of interspecific competition in the field. Biodiverse systems may include as many as 15–20 species in the same field during a year, for example, in the humid forest zone of Nigeria or the subtemperate south of Taiwan. Although these systems are not convenient in large fields or with industrial-size equipment, they can be highly productive if managed well, and these systems currently produce a substantial proportion of the world's food (IAASTD, 2008).

Agroecology as Movement

Largely in Latin America, agroecology has been seen as a social movement with the goal of transforming agriculture from an industrial and large-scale exploitation, highly dependent on fossil fuels and outside resources, to a more sustainable strategy that builds on human ingenuity and renewable resources that are internal to the farm. One key person in this movement has been Miguel Altieri, who was a founder of the Latin American Scientific Society of Agroecology (SOCLA), and also an active author who has linked science with practices and organized movements in sustainable agriculture (e.g., Altieri, 2009). This movement is promoted internationally by Via Campesina, an organization that publicizes the efficiencies of small-scale agriculture and promotes a social agenda of peasants' rights to land and a goal of food sovereignty (http://viacampesina.org/en/). Brazil has been at the forefront of grassroots activities by small farmers, and one of the early advocates José Lutzenberger (1976) was Special Secretary for the Environment to the President of Brazil from 1990 to 1992.

Agroecology in Teaching

Although the principles of ecology seem logical to apply in agriculture to reduce pesticide use, and to find biological methods to generate or liberate the nutrients needed to maintain crop yields, there has been slow acceptance of the term and academic courses in most agricultural universities. This may be due in part to its association with organic and biodynamic agriculture, or alternative agriculture, or to the promotion of agroecology as a social movement by activist groups perceived as a challenge to mainstream institutions. The strength of agroecology as a valuable practical direction to maintain production and profitability, reduce negative environmental impacts, and meet social goals such as food sovereignty is found in its transdisciplinary and holistic approach to identifying challenges and solutions (Lieblein et al., 2012). As a result, agroecology curricula became more popular in the last decade of the twentieth century at universities in Europe and in the U.S.A.

Recent Changes in Agriculture

Agriculture continues to evolve to meet new economic, political, and social challenges that are part of larger changes in the global society and environment. Design of production systems today is impacted by growing costs of fossil fuel based inputs, by rapidly changing technologies, by climate change, and by emerging power of consumers in food systems. Scarcity of fossil fuels and other key production inputs such as phosphorus is shaping a research agenda to make agriculture more efficient in use of limited resources. Genetic improvement for nitrogen and water use efficiency is high on the research agenda. Agroecologists approach this challenge by seeking systems less dependent on nonrenewable inputs and design of strategies that are based on local and renewable resources such as sun and rainfall, nitrogen fixed by legumes, management of pests using crop diversity, and integration of biodiversity. New technologies such as equipment guidance systems and precision application of fertilizers and chemicals are now used to reduce inputs and put them where they are most effective. Agroecologists propose this location-specific management through application of agroecological practices, better knowledge and observation of fields, and alternatives to purchased inputs. Transgenic crop varieties are now widely used in some parts of the world, but their safety and economic returns are the subject of continuing debate.

Climate variability is increasing, with greater frequency of major disruptive storm events and change of rainfall amounts and patterns that are beyond control of the farmer. Highly diverse agroecological systems help to build in resilience in farming systems by having several crops at different stages of development, and multiple potential uses for grain and forage that can buffer unfavorable weather events. Including a range of crops and animals in the system builds on integration efficiencies of enterprises, such as livestock that consume crop residues, and provides a range of products for the market to buffer changes in prices as one useful strategy. Direct marketing schemes such as farmers markets, farm sales shops, community supported agriculture, and pick-your-own operations provide

direct links with consumers and raise awareness of how and where food is grown. Growing demands for organic foods and locally produced products act as a countercurrent to the global food economy, and people are becoming more concerned about animal welfare as well as labor conditions where food is produced. These are among the changes that are evaluated in agroecology, in a quest to make food systems and supplies more transparent and encourage consumers to make informed choices about their food. The perspectives and methods of agroecology are increasingly recognized as leading to creative solutions for improving small farm agriculture (IAASTD, 2008) by taking these methods and scaling them up to impact national food production and human well-being (Ecumenical Advocacy Council, 2012).

Conclusions

Agroecology has emerged in the early years of the twenty-first century as an organizing philosophy that is guiding research and education in holistic and integrated agricultural and food systems. While recognizing the technological advances and production successes of the green revolution that brought more food and better well-being to many disadvantaged people, there have been unintended environmental consequences of wide-scale use of chemical fertilizers and pesticides. To produce food and adequate diet for 9 billion people, requires new thinking and innovative strategies for positive change in agriculture.

Scientists and development specialists are now recognizing that adequate food for all people depends on a wide array of social and economic challenges to be overcome. These include use of food crops to produce biofuels rather than changes in current systems to make them more energy efficient, loss of food in the chain from harvest in the field to consumption at home, and pervasive inequities in control of production resources and adequate incomes to allow many poor people to access food. The complexity of these challenges can best be evaluated and systems improved through transdisciplinary understanding and responsible action by public, private, and nongovernmental organizations working together. Agroecology as the 'ecology of food systems' provides a platform for study of agriculture and food systems from production to the consumer, also considering the importance of food distribution, diets, and healthy food. The major goal is moving toward agricultural systems that depend on renewable resources, conserve biodiversity, and that can be sustainable for the long term. In a globally connected world, the importance of local systems and specific solutions for unique places will be part of the future of agriculture and food systems.

See also: Energy and Society; Food Security and 'Green Revolution'; Land Change: The Merger of Land Cover and Land use Dynamics; Sustainability and Sustainability Science; Water Resources and Sustainable Water Management.

Bibliography

Abrol, D.P., 2014. Integrated Pest Management: Current Concepts and Ecological Perspectives. Academic Press, San Diego, CA.

Altieri, M.A., 1983. Agroecology. University of California Press, Berkeley, CA.

Altieri, M.A., July–August 2009. Agroecology, small farms, and food sovereignty. Monthly Review 61 (3).

Bensin, B.M., 1935. Agroecological Exploration in the Soto La Marina Region. Geographical Review, Mexico, 285–297.

Carson, R., 1964. Silent Spring. Houghton-Mifflin Publ, New York.

Ecumenical Advisory Alliance, 2012. Nourishing the World: Scaling up Agroecology. Switzerland, Geneva.

Francis, C. (Ed.), 1986. Multiple Cropping Systems. Macmillan Publishing Co., New York.

Francis, C. (Ed.), 2009. Organic farming: the ecological system. Agronomy Monograph, vol. 54. American Society of Agronomy, Madison, WI.

Francis, C., Lieblein, G., Gliessman, S., Breland, T.A., Creamer, N., et al., 2003. Agroecology: the ecology of food systems. Journal of Sustainable Agriculture 22 (3), 99–118.

Francis, C., Miller, M., Anderson, M., Creamer, N., Wander, M., et al., 2013. Food webs and food sovereignty: research agenda for sustainability. Journal of Agriculture, Food Systems, and Community Development 3 (4), 95–101. http://dx.doi.org/10.5304/jafscd.2013.034.010 (accessed 31.05.14.).

Gliessman, S.R., 2007. Agroecology: The Ecology of Sustainable Food Systems. CRC Press/Taylor & Francis, Boca Raton, FL.

Grassini, P., Eskridge, K.M., Cassman, K.G., 2013. Distinguishing between yield advances and yield plateaus in historical crop production trends. Nature Communications 4. http://dx.doi.org/10.1038/ncomms3918. Article number 2918.

Hazell, P.B.R., 2002. Green Revolution: Curse or Blessing? International Food Policy Research Institute, Washington, DC.

IMechE, 2013. Global Food: Waste Not, Want Not. Institute of Mechanical Engineering, London.

International Assessment of Agricultural Knowledge, Science and Technology for Development (IAASTD), 2008. Agriculture at a Crossroads. International Assessment of Agricultural Knowledge, Science and Technology for Development. New York: United Nations.

Jackson, W., 1980. New Roots for Agriculture. North Point Press, Berkeley, CA.

Lieblein, G., Breland, T.A., Francis, C., Ostergaard, E., 2012. Agroecology education: action-oriented learning and research. Journal of Agricultural Education and Extension 18 (1), 27–40.

Lutzenberger, J., 1976. Fim Do Future. Manifesto Ecologico Brasileiro. L&PM Editores, Porto Alegre, Brazil.

Mollison, B., 1988. Permaculture: A Designer's Manual. Tagari Publications, Tyalgum, Australia.

Olson, R.H., Lyson, T.A., 1999. Under the Blade: The Conversion of Agricultural Landscapes. Westview Press, Boulder, CO.

Rotolo, G., Rydberg, T., Lieblein, G., Francis, C., 2007. Emergy evaluation of grazing cattle in Argentina's Pampas. Agriculture, Ecosystems and Environment 119, 383–395.

Tischler, W., 1950. Ergebnisse und Probleme der Agrarökologie. Schrift. Landwirtschaft. Facultat Kiel 3, 71–82.

Wezel, A., Bellon, S., Doré, T., Francis, C., Vallod, D., et al., 2009. Agroecology as a science, a movement and a practice. A review. Agronomy for Sustainable Development 29, 503–515. http://dx.doi.org/10.1051/agro/2009004 (accessed 31.05.14.).

Wezel, A., Casagrande, M., Celette, F., Vian, J.F., Ferrer, A., Peigné, J., 2014. Agroecological practices for sustainable agriculture. A review. Agronomy for Sustainable Development 34 (1), 1–20.

Wezel, A., Soldat, V., 2009. A quantitative and qualitative historical analysis of the scientific discipline agroecology. International Journal of Agricultural Sustainability 7 (1), 3–18.

AIDS: Acquired Immune-Deficiency Syndrome

Ting Ting Lee, University of Melbourne, Melbourne, VIC, Australia
Ian Paul Everall, University of Melbourne, Melbourne, VIC, Australia; and Royal Melbourne Hospital, Parkville, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by L.S. Zegans, volume 1, pp. 344–350, © 2001, Elsevier Ltd.

Abstract

The AIDS epidemic of the 1980s shone a dark light on the assumptions of contemporary biomedicine. The world was confronted with a disease which it did not understand, could not treat, and which often attacked previously healthy young men. It had echoes of the great plagues of medieval times that decimated the cities of Europe and Asia. Theories abounded about its cause, until it was discovered that a previously unknown retrovirus instigated the destruction of human immune cells. What was particularly challenging about AIDS was its impact on society that extended beyond the usual concerns of biological medicine. Western medicine prided itself on its research methodology and development of new instruments and techniques to study organic processes. What medicine did not do particularly well at the time was to attend to the psychological, social, political, and ethical dimensions of the illness. The onslaught of AIDS altered all of that and forced the biomedical world to broaden its conception of illness and consider elements that went beyond the physical basis of pathology. Every aspect of this illness compelled medicine to ask new questions, challenge old assumptions, and build new collaborative relations with the patients. This article deals with the history of the HIV infection, its clinical manifestations and course, and issues in understanding its virology and immunology. It considers contemporary treatment options and the psychological and economic issues that they generate. It also discusses the neuropsychiatric complications of HIV disease, the psychosocial stressors accompanying the infection, and ethical and policy issues regarding diagnosis and treatment, as well as special problems confronted by minorities, women, care givers and their families, and people living with HIV from resource-limited countries.

With AIDS, the world was confronted with a disease that it did not understand, could not treat, and that often attacked previously healthy young men. Theories abounded about its cause until it was discovered that a newly emerged retrovirus instigated the destruction of human immune cells. Western medical science had recently been comfortable in the belief that infectious diseases were coming under scientific control and that the major health problems of the age lay in the understanding and treatment of chronic diseases and those of aging.

What was particularly challenging about AIDS was that its impact on society extended beyond the usual concerns of biological medicine. Medical science was at the time enjoying groundbreaking discoveries in the arenas of genetics and cellular and molecular biology. It was coming to better understand the mechanisms of diseases and was devising innovative methods of diagnosis and treatment. What medicine did not do particularly well at the time was to attend the psychological, social, political, and ethical dimensions of illness. AIDS altered that, and forced the biomedical world to broaden its conception of illness and consider elements that went beyond the physical basis of pathology. The old authoritarian models of the doctor–patient relationship are gradually being replaced by a process where patients are empowered to gather more information about their condition, play a greater role in making treatment decisions, and become more self-directive in maintaining their health. All of this has taken place in an environment where new discoveries about virology, immunology, and treatment strategies are rapidly occurring.

The story of human immunodeficiency virus (HIV) infection and its culmination in the disease that is called AIDS starts in the USA in the early 1980s (CDC, 1981). Surprised physicians treating young gay men noted that they were falling ill and dying in increasing numbers. There was no clear diagnostic reason for this phenomenon, and a variety of guesses was ventured as to the cause. The number of reported incidents of this strange illness began to increase. Soon it was discovered that patients receiving transfusions for hemophilia and also intravenous drug users were being reported with this disease. Reports began coming in from Africa documenting that heterosexual men and women were likewise falling ill in alarming numbers. It became evident that the primary site of the pathology was the destruction of an important component of the immune system, the CD4 helper T cells that are responsible for mounting a critical defense against infectious agents (Cooper et al., 2013). As a result of the loss of immune competency, an array of 'opportunistic' infections attacked the infected individuals and instigated a wide array of illnesses that infiltrated many body organs, chiefly the lungs and central nervous system (CNS).

Since the first cases in 1980s, around 75 million people have become infected with HIV globally. Statistics also revealed that there were 35.3 million people living with HIV in 2012 (UNAIDS, 2013). AIDS became the leading cause of death globally in the past decades. With the increased awareness and more people accessing treatment over the years, people living with HIV began to live longer; new HIV infection and AIDS-related deaths have also decreased by over 30% since its peak in the new millennium.

The Etiology, Immunology, and Clinical Pathology of HIV/AIDS

There has been some confusion in the public mind about the use of the meaning of the concepts of HIV and AIDS. The term 'HIV' refers to a person's infection with HIV. This virus can be

transported from one person to another through the exchange of certain body fluids (blood, semen, or vaginal discharge). It enters certain immune cells in the body and can progressively destroy them over time. The presence of this virus can be detected by testing antibody titers that reveal the body's response to its presence. A person can be HIV+ and be asymptomatic for long periods of time, or show only mild flu-like symptoms after acquiring the virus, a condition known as a *seroconversion* illness. This virus belongs to a *retrovirus* group called *cytopathic lentiviruses*. These are RNA-based viruses that convert their RNA to DNA and insert that into a cell's genetic pool through a process called *reverse transcription*; they have high mutation rate because of the high error-reading in the *RNA transcriptase*, an enzyme that transcribes viral RNA to viral DNA (Domingo et al., 1996; Roberts et al., 1988).

There are two types of HIV that have been shown to cause AIDS in humans: HIV-1 and HIV-2 (Peeters et al., 2014). These viruses share similar molecular structures and cause similar pathological disruptions. Currently, HIV-1 causes the majority of cases of AIDS throughout the world, while HIV-2 is found mostly in Africa. There are many different subtypes among the HIV-1 strain. These viruses can rapidly change their molecular structure within the body, making them difficult for the immune system to destroy. The mutability of these viruses also makes them difficult targets for preventive immunization strategies.

The term *AIDS* refers to a critical stage of the HIV infection when a large number of CD4 helper T lymphocytes have been destroyed, and the body is not able to mount an effective immune defense against secondary opportunistic pathogens or the toxic effects of HIV itself (Gong, 1984). The first symptomatic signs of immune breakdown may not occur for a number of years after the acquisition of the virus; latency might vary between 4 and 10 years. HIV can provoke diseases that involve the lymph nodes, lungs, brain, kidneys, and the abdominal cavity. There is no universal sequence of clinical symptoms, but the most frequent presentation can include enlarged lymph glands, pneumonia-like symptoms, decline in blood counts, and fungal infections, as well as diarrhea, weight loss, bacterial infections, fatigue, and disorders of the central nervous system. The Centers for Disease Control in 1993 compiled a list of conditions that defined AIDS. This included such infections as toxoplasmosis of the brain, tuberculosis, Kaposi's sarcoma, candidiasis infection of the esophagus, cervical cancer, cytomegalovirus (CMV) retinitis with loss of vision, and progressive multifocal leukoencephalopathy (PML), a viral infection that destroys white matter in the brain. Conditions on the list include opportunistic infections, or diseases that would otherwise not occur in the presence of a healthy immune system.

In order to best understand this devastating disease and its treatment, it is important to understand the relationship of the virus to the immune system. HIV is known as a retrovirus, which means that it can alter the flow of genetic information within a cell. It uses the enzyme *reverse transcriptase* for utilizing viral RNA as a template for producing viral DNA (Roberts et al., 1988). In most cells, DNA produces RNA as a genetic messenger. The sequence of infection after the entry of the virus through exchange of body fluids is the following: (1) the virus attaches to a host immune cell using the CD4 and CCR5 receptors (**Figure 1**); (2) the virus sheds its molecular coat and once it enters the cell, it begins the processes of reverse transcription; (3) viral DNA is integrated into the host cell, causing transcription and translation of the viral genetic code into viral protein, and inducing the host cell into producing more copies of the invading virus; (4) the newly formed viruses are released into the bloodstream, with the death of the host cell and the reinvasion of new immune host cells.

The helper T cells have the CD4 surface receptor that has a high affinity for the HIV surface protein gp120 (Wilen et al., 2012). However, HIV can attach to other cells including macrophages and monocytes (other immune cells), as well as cells in the intestines, uterine cervix, and Langerhans cells of the skin. Researchers have found that chemicals called cytokines also play a significant role in facilitating viral invasion. Interestingly, there are certain mutant strains of cytokine genes that actually prevent the entry of HIV into the target CD4 cell. Studies found that C–C chemokine receptor type 5 (CCR5) delta-32 homozygosity, which leads to a truncated protein, has a protective effect against HIV infection in that the virus cannot

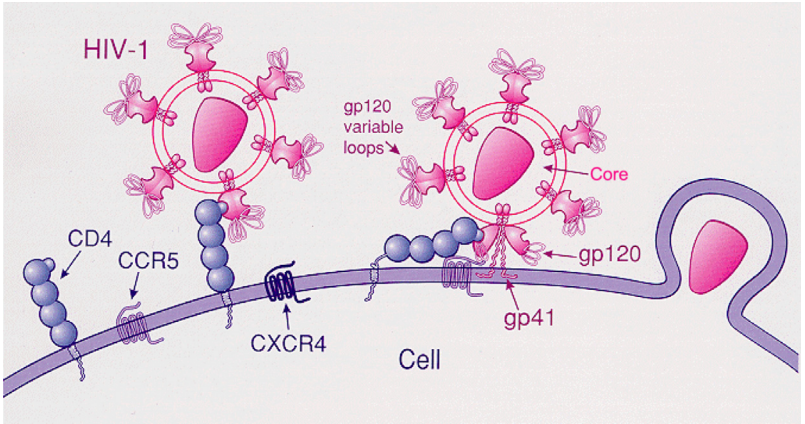

Figure 1 HIV envelope protein gp120 loop attaches to the CD4 receptor and the coreceptor (CCR5 or CXCR4), and while another viral protein gp41 fuses with the cell membrane in order to initiate viral entry into the host cell.
Source: US National Institutes of Health – National Institute of Allergy and Infectious Diseases.

bind to the truncated receptor (Taborda-Vanegas et al., 2011; Estrada-Aguirre et al., 2013). The presence of these genes in certain people may explain why some individuals remain resistant to HIV infection despite repeated exposure to virally loaded body fluids and why some people show slower rates of disease progression. There is also evidence that a patient from Berlin who received stem cell transplantation with this mutation has been cured of HIV (Allers et al., 2011).

Not all infected cells immediately produce new viral copies that destroy the cell and infect many others. Some cells may remain dormant or latent without producing new viral copies. There may be adjuvant factors that convert latent infected cells into ones that produce viral copies. It is speculated that coexisting infections with their additional antigen load may have a facilitating effect, as may certain drugs, possibly stress, fatigue, and so on. Anything that reduces immune competency may become a cofactor. Many individuals who are infected try to find ways to strengthen the immune system. These attempts may include meditation, exercise, herbs, dietary supplements, prayer, and support groups. Whether these have any long-lasting ameliorative effects is uncertain, but they may give the person a greater sense of self-maintenance and control over a devastating illness.

Issues of Testing for HIV and Stages of the Infection

When HIV infection was first recognized in the early 1980s, it was most prevalent among young men who had sex with other men and were living in large metropolitan cities. Prevention programs and the availability of tests for the infection have significantly decreased the rate of new HIV cases among this population. The major routes of infection with HIV have been shown to be through fluid-exchanging sexual behavior; use of HIV-contaminated needles; from mother to infant during pregnancy, delivery, or breast feeding; and through HIV-contaminated blood products passed during transfusions. The development of accurate testing procedures for the infection became important both to assist in the diagnosis of the disease and to protect the nation's blood supply.

Various means of testing for the presence of HIV or its effects have been developed since the discovery of the disease (Crews et al., 2009). The virus can be cultured directly from the blood, but the most efficient means of diagnosis has come from detecting antibodies to the virus. Such testing can determine whether the body has come into contact with the virus and has mounted an immune defense. However, antibody testing can produce both false positive and negative results, and cannot be used during the window period (usually 3–6 months) where the level of antibodies is too low to be detected. Most tests take a few days to receive results, but more rapid same-day tests are available. The usual procedure is for the patient (blood, urine, or saliva sample) to be tested first using an enzyme-linked immunosorbent assay (ELISA), which is very reactive to HIV antibodies. An HIV RNA test allowed early detection within 2 weeks after a suspected infection, therefore could be utilized during the window period of HIV antibody test. A positive result on the first test must be replicated in a western blot antibody test before a diagnosis can be made. It has a high level of specificity for detecting protein antigens of HIV. Rapid point-of-care HIV tests have also been developed to deliver the health care service to resource-limited area (Glynn et al., 2014; Chamie et al., 2014).

Although there are reasonably accurate and fast methods of testing for HIV, not all people who suspect that they have been exposed to the virus seek testing, and not all who do arrange for it early in the progress of the infection. There are a variety of reasons why this is so, some people are simply not aware of the risks of infection and the procedures of arranging for testing. Others have been properly educated, but because of lack of perceived risk, stigma, denial, fear of the outcome, and concern about loss of privacy and public exposure, avoid learning if they have been infected (Ances et al., 2008).

Often when a person has received a reliable diagnosis of HIV infection, they experience a variety of negative psychological responses. They can become anxious, and or depressed, may feel guilt about their past behavior, and are worried about physical deterioration and eventual death. If they have observed friends experience the progress of the disease, they may have frightening images of what may happen to them. Suicidal thoughts are often present immediately after a diagnosis is made and the patient may become sensitive to any physical symptoms. It is vital that such persons receive comprehensive, sensitive, and accurate medical and mental health care upon receiving news of a positive test. Many people at this point in the illness feel isolated, shamed, and unable to talk about their fears and questions. Since the onset of the epidemic, many outstanding clinical and community programs providing vital information and emotional support have been established. Since significant advances in therapies have been developed, with a marked increase in survival rates, early detection and humane medical and psychological interventions are essential for all.

The Clinical and Immunological Progression of the Illness

The finding of a positive antibody test for HIV does not predict an inevitable disease course for all people. Some people exhibit only mild symptoms initially and then remain totally asymptomatic throughout their lifetime. Others demonstrate severe symptoms shortly after the virus is detected and others show no sign of illness for 8–10 years after being diagnosed as positive. AIDS, now termed HIV disease, has uncertain symptoms, long quiescent periods for many, and in the past often lead to a devastating end state.

There are many reasons to explain this variability. Some may have a genetic resistance to the virus, a robust (innate and adaptive) immune system and have been exposed to a less virulent strain of the virus (Kaul et al., 2005). For others, diagnosis may occur long after they were exposed to the virus and their immune system may be severely compromised.

When told that they may develop HIV disease, people face many uncertainties, complex medical and lifestyle decisions, and the need to adapt to changing medical conditions throughout their life. These decisions involve trusting communications with health care workers, friends and family, social support systems, and work associates. For some, the complexity of these decisions may prove overwhelming, and

they may make many poor health and lifestyle decisions. For others, the illness may reactivate existing psychiatric problems or create new neuropsychological symptoms. Indeed, one of the most provocative complications of the disease is the emergence of HIV-related cognitive disabilities (Almeida, 2013).

As noted earlier, the harm caused by HIV is instigated by its entry into the CD4 helper T lymphocyte and the cells subsequent capture of HIV genetic material along with the use of the host cell's machinery to produce many new viral copies. When a full-fledged infection is in progress, millions of new viral particles can be found in the human bloodstream. This causes a drop in the number of CD4 cells, which usually have a level of at least 800 cells/cubic millimetre of blood. Counts below 500 are considered serious and below 200 cells/mm^3 will set the stage for dangerous opportunistic infections (Young et al., 2012). Periodically testing the concentration of CD4 cells has been an important means of measuring and predicting the course of the disease. More recently, tests have been developed to measure the viral load directly in the bloodstream and body fluids. These sensitive tests are considered a most reliable measure of the progress of the infection.

The Progression of the Illness

Acute Infection or Seroconversion

Following the entrance into the immune cells, there is a dramatic drop in CD4 counts and usually the onset of symptoms that can resemble flu or mononucleosis. These can include fever, fatigue, enlarged lymph glands, headaches, rashes, and muscle aches. These symptoms usually resolve in few weeks, as the immune system resists the viral spread. It does so by triggering CD8 cytotoxic cell responses that destroy infected cells and by stimulating an antibody response to the virus (Demers et al., 2013). This reaction binds and removes many HIV particles from the blood.

Viral Latency

These body defenses reduce the viral load but rarely eliminate the virus from the body. In many cases, there tend to be an equilibrium established between the immune defense and the viral level. This so-called 'set point' can be quite different from patient to patient. Whether a patient will become seriously symptomatic depends upon the balance between viral activity and immune competence. When viral activity gains the upper hand over immune defense, serious symptoms will occur. Depending upon when the diagnosis was first made, the set point can vary from 4 to 10 years. Some patients, called 'non-progressors,' may never show serious symptoms and maintain good CD4 levels.

HIV Disease or AIDS

Individuals who have compromised CD4 counts, have a bout of opportunistic infections, and yet be asymptomatic thereafter. Once the CD4 count drops below 200 cells//mm^3, people usually develop complications of HIV disease. The immune system is no longer able to contain the viral spread, and organisms that it usually can control begin producing dangerous infections. Most common are pneumocystis carinii, pneumonia, and toxoplasmosis. However, many other organisms can also be activated and cause multiple organ damage. When the brain is affected in the end stages of the disease, delirium, dementia, and a variety of motor impairments can occur. In other countries where this disease is common, there may be a different array of opportunistic infections found. In places such as Haiti and Africa, one sees more candida infections and crytococcal meningitis. Intestinal disorders with diarrhea and wasting are also common. Both abroad and in the USA, tuberculosis associated with HIV has at times become a major issue.

The Treatment of HIV Infections

The combined antiretroviral therapy (cART) has been highly beneficial for people living with HIV, delaying disease progression and improving quality of life significantly. The introduction of cART has contributed to a sharp decline in the mortality related to HIV infection; however, it needs to be taken for life. The cART does not cure HIV and people living with HIV can still progress to develop HIV-associated neurocognitive disorders (Passaes and Saez-Cirion, 2014). Despite various public and private funded HIV prevention and treatment initiatives, such as the Bill & Melinda Gates Foundation Fund, the Global Fund, and the Ryan White Comprehensive AIDS Resources Emergency (CARE) Act, a significant proportion of people still have restricted access to the ART due to financial limitations. This has created serious national and international concerns about creating a two-class partition concerning distribution and access to these medications. Such questions will inevitably arise as new, expensive, and sophisticated treatments are developed to treat other chronic conditions.

Since the introduction of these new treatment regimes, the death rate from AIDS has dropped steadily since the peak in 2005, and the number of HIV-related hospitalizations has been significantly reduced (UNAIDS, 2013). This was accomplished through a more detailed understanding of the molecular activity of the virus as it enters a human cell. To date, there are 28 Food and Drug Administration (FDA) approved drugs for the treatment of HIV infection in five different classes, each targeting a specific molecular mechanism. These are (1) nucleoside-analog reverse transcriptase inhibitors, (2) non-nucleoside-analog reverse transcriptase inhibitors, (3) integrase inhibitors, (4) protease inhibitors, and (5) fusion/entry inhibitors (Kinch and Patridge, 2014). The new classes of drugs operate either by blocking viral replication or by inhibiting the HIV reverse transcriptase and protease. The original drug against HIV was developed in 1987 and was called *Zidovudine* (*AZT*) (Kinch and Patridge, 2014). This drug is a nucleoside-analog reverse transcriptase inhibitor, functions by preventing the completion of the viral DNA strand in the human cell. Later, non-nucleoside reverse transcriptase inhibitors were developed which also inhibited retroviral activity. The protease inhibitors were subsequently developed, which prevents the division of newly produced HIV proteins. The fusion/entry inhibitors prevent viral entry by blocking the HIV-binding coreceptors CCR5 and CXCR4. More recently, a new class of

integrase inhibitors was introduced to block the action of integrase, a HIV enzyme that facilitate the integration of viral genome into host cell. The usual prescription is for a combination of two or more classes of drugs.

The development of these powerful drugs is important because research has shown that the HIV reproduction is robust early in the disease, but remains stable because of the immune response that produces high numbers of CD4 cells. A strong initial response of CD4 cells facilitates the body's subsequent production of a CD4 subset that reacts selectively to HIV. Medical research has determined that the level of the viral load in the body is highly correlated with ultimate prognosis. If the viral level can drop to an almost undetectable level, the likelihood of developing opportunistic infections and other complications of HIV disease declines (Dembek et al., 2012).

Important as the development of these effective medications have been, they still further complicate the lives of people living with HIV and create a myriad of complex decisions for them. Previously, the medications must be taken on a rigorous schedule and sometimes in extremely large doses. Failure to maintain the timing of a dosage of the drugs may result in their being ineffective, and possibly producing viral resistance. Even when taken correctly, the drugs may not be effective, leading to disappointment.

The recent development of fixed-dose drug combinations (FDCs) represent a major advancement in cART, where the 'cocktail' regimen (up to 16 pills daily) has been simplified to the new once daily, single-tablet regimen (STR) (Aldir et al., 2014). This novel treatment therapy has been found to significantly enhance treatment adherence among people living with HIV, therefore improving their quality of life. With the great benefits, the STR has now been approved and adopted in many countries as the first-line treatment for people living with HIV.

Although a vast improvement in drug development and financial assistance is available, other difficulties such as negative side effects remain challenging. These may encompass anemia, neuropathy, headache, diarrhea, rashes, and hepatitis. The complexity of this medication approach requires a close collaboration between the treating physician and the patient. It underscores the necessity of having the patient being an active participant in his or her treatment.

The latest discovery of the 'Berlin patient' who was allegedly found cured of HIV, followed by two patients from Boston who underwent a similar situation, initiated research searching for HIV cure (Cannon et al., 2014). These patients received a stem cell transplant with CCR5 delta-32 homozygosity mutation from the donors, as part of their cancer treatment, and were later found to have non-detectable viral load. These findings marked the beginning of a new era of biomedical research for treatment and prevention of HIV infection, at the same time provided people living with HIV with a hope for cure.

Neuropsychiatric and Psychosocial Issues

Although HIV infection can cause a wide number of dangerous medical complications, none are more feared by people living with HIV than the neuropsychiatric disorders. HIV can cause damage to the CNS itself and open the door to a myriad of opportunistic brain infections. The possible end-stage of delirium and dementia with loss of personality and body function control is a grim vision of their future.

HIV infection can not only introduce a number of CNS disorders, but it can also cause the reactivation of previous psychiatric illness (Almeida, 2013). There are a great number of people living with who also have a previous psychiatric history. Often these people may be less capable of taking reasonable precautions with regard to unsafe sex and intravenous drug use, rendering themselves vulnerable to infection.

HIV invasion of the CNS has previously been found in more than 50% of people who are HIV + but asymptomatic, and over 90% on autopsy of all AIDS patients showed evidence of neuropathology (Gray et al., 1996). More recently in the era of cART, 78% of postmortem brain samples from individuals with advanced HIV disease showed some degree of neuropathology, which is associated with HIV-associated neurocognitive disorders (HAND) (Everall et al., 2009). These brain abnormalities include inflammation such as HIV encephalitis (HIVE), white matter and subcortical damage, as well as neurodegeneration such as synaptodendritic damage and neuronal loss (Gendelman et al., 2012). Inevitably, patients with involvement of the CNS show symptoms of cognitive impairment, movement problems, and behavioral difficulties. These difficulties may include a change in personality, withdrawal and apathy, inappropriate emotional responses, sharp mood swings, mania or suicidal impulses, and hallucinations. Often the brain involvement leads to an inability of the patient to carry out activities of daily living and requires home assistance. There is belief among some clinicians that early diagnosis and vigorous drug treatment can delay or even prevent later HIV-associated dementia (HAD). This will depend upon the extent of CNS damage existing at the time of initiation of treatment. Lithium and neuroleptic medications have been used to systematically treat people who become manic or agitated (Schifitto et al., 2009). There is also evidence that lithium may be neuroprotective against HIV and could potentially be used in conjunction with cART to improve HIV-associated neurocognitive impairment (Letendre et al., 2006; Ances et al., 2008). The exact means by which the HIV causes damage in the CNS is not certain. There are two possible routes of HIV damage in the CNS: a direct insult where neurotoxicity is triggered by HIV viral protein such as gp120, tat, and nef; while an indirect damage can occur when macrophages and microglia release pro-inflammatory cytokines in response to viral infection and prolonged oxidative stress (Kaul et al., 2005).

Although the presentations of AIDS-related opportunistic infections and tumors have become uncommon due to the efficiency of cART, the milder HAND remain prevalent, potentially due to people living longer with HIV infection (Chan and Brew, 2014). The variation in the severity of HAND between people living with HIV could potentially be due to different CNS penetration properties of drugs administered, or genetic protective factors mentioned earlier (Eisfeld et al., 2013). Nevertheless, treatment with high CNS penetrating antiretroviral (ARV) does not necessarily predict a better outcome of HAND as revealed by a study that reported neurotoxicity features of ARV (Akay et al., 2012).

Both the organic and social stresses of HIV/AIDS are associated with the emergence of psychological distress and

psychiatric symptoms. The most common psychological disorder associated with HIV infection is an adjustment disorder with features of anxiety and depression. Major depression is often observed among people living with HIV (Fellows et al., 2013). This is most common among those with a previous history of depression and those who are isolated and have little social support. A feeling of hopelessness and lack of personal control over the development of the disease is found among these depressed patients. People living with HIV disease experience, during the course of their illness, a variety of losses and other stresses that increase their vulnerability to psychiatric disorders. In the past, these included loss of employment, death or illness of friends, disengagement of family members, financial losses, loss of sexual partners, abandonment of future goals, reduction of physical function, and failing cognitive abilities. Fortunately, with the successful HIV treatment these significant losses are not as common today. Although there are a generation of people living with HIV who survived the pre-cART era that are still coping with the detrimental effect of those historic losses on their lives today.

The occurrence of depression, anxiety, somatization disorders, suicidal ideation, and substance abuse can be traced to the reaction to such losses, or the fear of them. Such psychological distress may by itself compromise the immune system. With the advancement in the field of pharmacology, there is an array of antidepressants and antianxiety medications available today to help manage these conditions. Positive mental attributes acquired through therapy, support groups, prayer, and so on, may also modulate the psychological pain of the patient and help to improve the quality of daily living. Research has not yet determined whether developing more adequate coping mechanisms and enhanced self-esteem will effectively alter the immunological course of the illness.

Like so much else in this disorder, there is no common psychological pathway that all patients follow. Their previous psychiatric history, adaptive responses to the virus, social support, financial resources, access to good medical care, response to medication, will all play a role in helping people living with HIV achieve a positive mental equilibrium. Supportive therapy can help patients deal with fear, uncertainty, and a sense of self-recrimination. Such help can occur in a professional setting, through community groups, religious counseling, and in a variety of other innovative venues. The important consideration is to help people to feel that they are not worthless, socially shunned, or without something valuable to contribute to friends and society.

Policy and Ethical Issues

This epidemic has emphatically raised the question of what is the responsibility of governments, pharmaceutical companies, and insurance plans for protecting and treating people of very limited financial means. HIV disease or AIDS involves populations of individuals who are often out of the spotlight of public attention or who have been morally condemned because of certain behavior characteristics. The most striking example of an ignored population is the many millions of heterosexual individuals who have fallen ill with AIDS in Africa. This disease has disrupted families, created national economic disaster (particularly in agriculture), and has overtaxed the medical resources of extremely poor countries. Yet in the richest industrial and technological countries in the world, there is little concern or knowledge about the problems on this continent. The ethical responsibility for providing the benefits of modern medicine and pharmacology to people in a distant and largely unknown continent is now beginning to emerge in the public's awareness.

In Europe, Australia and other developed countries, people living with HIV are entitled to free or subsidized access to cART at the point of health care delivery. The USA's record in responding to the needs of minorities and underprivileged people with the infection even in its own country is not a cause for optimism. Medical care and treatment for people with AIDS is expensive and as a chronic disease, its costs mount over time. Infected people often lose their insurance, are no longer able to work, and exhaust their financial resources quickly. Yet at the beginning of the twenty-first century, there is at best a patchwork of initiatives to finance the development of drugs and make them available to those in need. Currently available funding of treatment is only sufficient for a small proportion of people living with HIV globally; numerous people are still in need of treatment and care, especially those in the resource-limited areas. Contemporary policy must also take into account the many women and, at times, their children who are infected with HIV. They require many additional social and educational services to deal with their medical and social problems. Attention must also be given to the special needs of adolescents who are at greater risk for contracting this disease.

Compounding the problems of money are issues of protecting both the privacy of people with HIV infection, while also safeguarding the public's health. Generally, it has been felt that well-conceived educational programs can play a major role in both prevention and helping infected persons to make ethical decisions about disclosing their condition to others. Related to this question are issues regarding blood bank testing, protection of health care workers, and disclosure to prospective and current sexual partners.

In the USA, diseases that are sexually transmitted have become a metaphor for troubling issues about 'moral' behavior. It raises questions about the values of the society, parents' control over the behavior of their children, and the images that are conveyed by the media to the public. Among some groups, sexually transmitted diseases are seen as a punishment for immoral behavior. Much of the public's response to AIDS, even after two decades of familiarity with the disease, is shaped not by its medical and biological characteristics, but by US social and cultural attitudes toward the behaviors associated with contracting the illness. People's willingness to help those who are afflicted is molded by their social perspectives. If the public disapproves of the people who have contracted this disease, they are reluctant to provide the medical care, drugs, shelter, and the social support that they need.

The HIV/AIDS epidemic has also raised questions about the right to access to medications that have not yet met Food and Drug Administration standards for testing and release to the public. Should drugs that have not yet proven their safety be given to people who might otherwise die?

What has emerged, however, is the belief that patients and their advocates have a right to be at the table where scientific and public health decisions are being made. Patients must participate in decisions regarding the initiation and termination of treatment, rights of privacy concerning their condition, informed consent, and access to new forms of treatment. Of particularly interest is the AIDS Coalition to Unleash Power (ACT UP), an international group of advocates devoted to promote and protect the well-being of people living with HIV and ultimately end the disease. The notable achievements of the group in the early days include the significant price reduction of Zidovudine manufactured by Burroughs Wellcome, and the ban of placebo use in HIV treatment trials.

The lessons learned in understanding AIDS are reshaping views of the roles of doctor, patient, family, and community. Hopefully, such new knowledge will provide a more humane and comprehensive attitude for the care of patients with all diseases. Illness is not an event that happens in just one person's body. Its consequences are part of the social fabric. Ethical consideration, as well as advances in biological expertise, must inform future policy and health care decisions.

See also: Clinical Psychology, Psychiatry and Homosexuality; Depression, Pessimism, and Health; Depression; HIV Risk Interventions; Mortality and the HIV/AIDS Epidemic; Sexual Risk Behaviors; Sexually Transmitted Infections: Social Network Analysis.

Bibliography

Akay, C., Lindl, K.A., Shyam, N., Nabet, B., Goenaga-Vazquez, Y., Ruzbarsky, J., Wang, Y., Kolson, D.L., Jordan-Sciutto, K.L., 2012. Activation status of integrated stress response pathways in neurones and astrocytes of HIV-associated neurocognitive disorders (HAND) cortex. Neuropathology and Applied Neurobiology 38, 175–200.

Aldir, I., Horta, A., Serrado, M., 2014. Single-tablet regimens in HIV: does it really make a difference? Current Medical Research and Opinion 30, 89–97.

Allers, K., Hutter, G., Hofmann, J., Loddenkemper, C., Rieger, K., Thiel, E., Schneider, T., 2011. Evidence for the cure of HIV infection by CCR5Delta32/Delta32 stem cell transplantation. Blood 117, 2791–2799.

Almeida, S.M., 2013. Cognitive impairment and major depressive disorder in HIV infection and cerebrospinal fluid biomarkers. Arq Neuropsiquiatr 71, 689–692.

Ances, B.M., Letendre, S.L., Alexander, T., Ellis, R.J., 2008. Role of psychiatric medications as adjunct therapy in the treatment of HIV associated neurocognitive disorders. International Review of Psychiatry 20, 89–93.

Cannon, P.M., Kohn, D.B., Kiem, H.P., 2014. HIV eradication-from Berlin to Boston. Nature Biotechnology 32, 315–316.

CDC, 1981. Kaposi's sarcoma and pneumocystis pneumonia among homosexual men – New York City and California. MMWR Morbidity and Mortality Weekly Report 30, 305–308.

Chamie, G., Kwarisiima, D., Clark, T.D., Kabami, J., Jain, V., Geng, E., Balzer, L.B., Petersen, M.L., Thirumurthy, H., Charlebois, E.D., Kamya, M.R., Havlir, D.V., 2014. Uptake of community-based HIV testing during a multi-disease health campaign in rural Uganda. PLoS One 9, e84317.

Chan, P., Brew, B.J., 2014. HIV associated neurocognitive disorders in the modern antiviral treatment era: prevalence, characteristics, biomarkers, and effects of treatment. Current HIV/AIDS Reports 11 (3), 317–324.

Cooper, A., Garcia, M., Petrovas, C., Yamamoto, T., Koup, R.A., Nabel, G.J., 2013. HIV-1 causes CD4 cell death through DNA-dependent protein kinase during viral integration. Nature 498, 376–379.

Crews, L., Patrick, C., Achim, C.L., Everall, I.P., Masliah, E., 2009. Molecular pathology of neuro-AIDS (CNS-HIV). International Journal of Molecular Sciences 10, 1045–1063.

Dembek, C.J., Kutscher, S., Allgayer, S., Russo, C., Bauer, T., Hoffmann, D., Goebel, F.D., Bogner, J.R., Erfle, V., Protzer, U., Cosma, A., 2012. Longitudinal changes in HIV-1-specific T-cell quality associated with viral load dynamic. Journal of Clinical Virology 55, 114–120.

Demers, K.R., Reuter, M.A., Betts, M.R., 2013. CD8(+) T-cell effector function and transcriptional regulation during HIV pathogenesis. Immunology Review 254, 190–206.

Domingo, E., Escarmis, C., Sevilla, N., Moya, A., Elena, S.F., Quer, J., Novella, I.S., Holland, J.J., 1996. Basic concepts in RNA virus evolution. FASEB Journal 10, 859–864.

Eisfeld, C., Reichelt, D., Evers, S., Husstedt, I., 2013. CSF penetration by antiretroviral drugs. CNS Drugs 27, 31–55.

Estrada-Aguirre, J.A., Cazarez-Salazar, S.G., Ochoa-Ramirez, L.A., Acosta-Cota Sde, J., Zamora-Gomez, R., Najar-Reyes, G.M., Villarreal-Escamilla, P., Osuna-Ramirez, I., Diaz-Camacho, S.P., Sanchez-Zazueta, J.G., Rios-Tostado, J.J., Velarde-Felix, J.S., 2013. Protective effect of CCR5 Delta-32 allele against HIV-1 in Mexican women. Current HIV Research 11, 506–510.

Everall, I., Vaida, F., Khanlou, N., Lazzaretto, D., Achim, C., Letendre, S., Moore, D., Ellis, R., Cherner, M., Gelman, B., Morgello, S., Singer, E., Grant, I., Masliah, E., 2009. Cliniconeuropathologic correlates of human immunodeficiency virus in the era of antiretroviral therapy. Journal of NeuroVirology 15, 360–370.

Fellows, R.P., Byrd, D.A., Morgello, S., 2013. Major depressive disorder, cognitive symptoms, and neuropsychological performance among ethnically diverse HIV+ men and women. Journal of the International Neuropsychological Society 19, 216–225.

Gendelman, H.E., Grant, I., Everall, I.P., Fox, H.S., Gelbard, H.A., Lipton, S.A., 2012. The Neurology of AIDS. Oxford University Press, Oxford, UK.

Glynn, M.T., Kinahan, D.J., Ducree, J., 2014. Rapid, low-cost and instrument-free CD4+ cell counting for HIV diagnostics in resource-poor settings. Lab on a Chip 14, 2844–2851.

Gong, V., 1984. Acquired immunodeficiency syndrome (AIDS). American Journal of Emergency Medicine 2, 336–346.

Gray, F., Scaravilli, F., Everall, I., Chretien, F., An, S., Boche, D., Adle-Biassette, H., Wingertsmann, L., Durigon, M., Hurtrel, B., Chiodi, F., Bell, J., Lantos, P., 1996. Neuropathology of early HIV-1 infection. Brain Pathology 6, 1–15.

Kaul, M., Zheng, J., Okamoto, S., Gendelman, H.E., Lipton, S.A., 2005. HIV-1 infection and AIDS: consequences for the central nervous system. Cell Death and Differentiation 12 (Suppl. 1), 878–892.

Kinch, M.S., Patridge, E., 2014. An analysis of FDA-approved drugs for infectious disease: HIV/AIDS drugs. Drug Discovery Today 19 (10), 1510–1513.

Letendre, S.L., Woods, S.P., Ellis, R.J., Atkinson, J.H., Masliah, E., Van Den Brande, G., Durelle, J., Grant, I., Everall, I., 2006. Lithium improves HIV-associated neurocognitive impairment. AIDS 20, 1885–1888.

Passaes, C.P., Saez-Cirion, A., 2014. HIV cure research: advances and prospects. Virology 454-455, 340–352.

Peeters, M., D'arc, M., Delaporte, E., 2014. Origin and diversity of human retroviruses. AIDS Reviews 16, 23–34.

Roberts, J.D., Bebenek, K., Kunkel, T.A., 1988. The accuracy of reverse transcriptase from HIV-1. Science 242, 1171–1173.

Schifitto, G., Zhong, J., Gill, D., Peterson, D.R., Gaugh, M.D., Zhu, T., Tivarus, M., Cruttenden, K., Maggirwar, S.B., Gendelman, H.E., Dewhurst, S., Gelbard, H.A., 2009. Lithium therapy for human immunodeficiency virus type 1-associated neurocognitive impairment. Journal of NeuroVirology 15, 176–186.

Taborda-Vanegas, N., Zapata, W., Rugeles, M.T., 2011. Genetic and immunological factors involved in natural resistance to HIV-1 infection. Open Virology Journal 5, 35–43.

UNAIDS, 2013. Global Report: UNAIDS Report on the Global AIDS Epidemic 2013.

Wilen, C.B., Tilton, J.C., Doms, R.W., 2012. Molecular mechanisms of HIV entry. Advances in Experimental Medicine and Biology 726, 223–242.

Young, J., Psichogiou, M., Meyer, L., Ayayi, S., Grabar, S., Raffi, F., Reiss, P., Gazzard, B., Sharland, M., Gutierrez, F., Obel, N., Kirk, O., Miro, J.M., Furrer, H., Castagna, A., De Wit, S., Munoz, J., Kjaer, J., Grarup, J., Chene, G., Bucher, H., 2012. CD4 cell count and the risk of AIDS or death in HIV-Infected adults on combination antiretroviral therapy with a suppressed viral load: a longitudinal cohort study from COHERE. PLoS Medicine 9, e1001194.

Albert, Hans (1921–)

Eric Hilgendorf, Faculty of Law, University of Würzburg, Würzburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Hans Albert is one of the most influential philosophers of science and social philosophy in the German-speaking world. In the 1960s, he participated in the German 'positivist dispute' on the same side as Karl Popper against T.W. Adorno and Jürgen Habermas, helping to reestablish analytic thinking in German social philosophy. He defended Max Weber's postulate that scientific statements should be free of personal values against its critics. In ethics, Albert has argued for a noncognitivist position. As an outspoken proponent of naturalism, he has repeatedly challenged theology on the ground of its lack of adherence to the rules of science.

Hans Albert, born on 8 February 1921, is one of the most influential philosophers of science and social philosophy in the German-speaking world. He is considered to be one of the main proponents of critical rationalism, a branch of analytic philosophy founded by Karl Popper during his debates with philosophers of the Vienna Circle.

Albert has identified three principal features of the philosophical position he represents: (1) *logical fallibilism*, which is the conviction that all human attempts at problem solving are fundamentally fallible and may need to be revised in the light of new information; (2) *methodological rationalism*, also called the method of critical examination, according to which two steps can be distinguished in any attempt at problem solving: the development of new proposals for solving problems based on experience, imagination, and intuition, and then testing based on reality; and finally, (3) *critical realism*, which includes the belief that although knowledge of reality is always fallible, nevertheless, in principle, it is possible.

Biography

After attending high school in Cologne and serving in the army during the Second World War, Albert commenced his studies as an economics major in the winter semester of 1946 at the Faculty of Social Sciences and Economics of the University of Cologne. The intellectual climate at the university was marked by ideological disillusionment and fervid study; the German sociologist Helmut Schelsky later referred to these students as the *skeptical generation*. Albert, who had already been intensively interested in Oswald Spengler while he was in high school, studied numerous philosophical works in an unsystematic manner, in addition to his pensum of business and economics literature. The thesis for his Diploma was supervised by the sociologist Leopold von Wiese, who also agreed to supervise his PhD studies.

Even at this early age, the very strong influence of Max Weber can be seen, which shaped Albert's thinking for the rest of his life. He summarized his experiences in his own words as follows:

> Soon I was filled with a deep sense of mistrust against a science where – in contrast to Max Weber's principle of objectivity – hidden value judgments seemed to abound and beyond that the real validity of postulated scientific laws hardly seemed to be regarded as a major problem. This discipline seemed to be the perfect example of an ideology, much in the sense which the sociologist Theodor Geiger later characterized as pseudo-scientific statements lacking all empirical basis. Since no one presented any contrary arguments in discussions either inside and outside the university, which were likely to dispel my concerns, I felt that my critical perspective was reinforced again and again.
>
> Albert, 1977/1984: 10

In the summer of 1952, Albert became a postdoc in the Department of Social Policy at the Research Institute for Social and Administrative Sciences, headed by Gerhard Weisser, of the University of Cologne. Weisser was a Kantian from the Leonhard Nelson School, who was intensively interested in fundamental questions. At the Weisser Institute, Albert became acquainted with the works of modern logic and logical positivism, particularly with those of Victor Kraft, whose theories on objectivity and conventionalism Albert soon learned to appreciate. Albert sympathized with the Vienna Circle philosophers, retaining however an existential approach in ethics questions.

In 1955, Albert submitted the dissertation for his Habilitation (second PhD), entitled 'National Economy as Sociology of Commercial Relationships.' The goals of the dissertation were, first, a critical analysis of the modern 'pure' Economics from the perspective of the philosophy of science, and second, the development of the foundations for sociologically oriented empirical Economics, which traces economic processes to the behavior of individuals in certain social roles, thereby taking institutions in society into account. At first, the faculty refused to accept the dissertation because, on the one hand, the novel epistemological approach raised concerns and, on the other, for political reasons, some faculty members suspected Albert of secretly sympathizing with Marxism. Albert was finally able to get his habilitation in 1957 for his other work on the subject of social policy. He taught as an associate professor until 1963 when he was given a new chair in Mannheim for 'sociology and philosophy of science.' He remained there until he became emeritus professor in 1989.

At the suggestion of Ernst Topitsch, Albert started visiting the Alpbach European Forum in 1955. It was there that he got to know, among others, Paul Feyerabend, Alf Ross, and

(in 1958) Karl Popper. Based on his impressions from this meeting, he wrote the comprehensive article 'The critical rationalism of Karl Raimund Popper,' which was published in 1960 in the *Archiv für Rechts-und Sozialphilosophie*. The article anticipated many of the elements present in Albert's main work, the *Treatise on Critical Reason*. In Alpbach, he met in 1956 Margarete von Pacher, whom he married in the winter of 1957. Today Hans Albert lives in Heidelberg.

The German 'Positivist Dispute'

In October 1961, the German Sociological Association (DGS) held a conference in Tübingen to discuss the methodological conflict which had been simmering for some time between empirical sociologists, represented in particular by René König and the Frankfurt School, which was influenced by Marxist ideas. The two main speakers were Karl Popper, who presented a paper on the logic of the social sciences, and T.W. Adorno, who was asked to be the cospeaker. The two rival speakers, both of whom had been forced to emigrate by the Nazis, treated each other graciously despite their significant ideological differences. The dispute became more acerbic when Habermas published an article in 1963, in which he portrayed Popper as a naive positivist (which was the philosophical movement that Popper had distanced himself from in the paper he had presented in Tübingen). Albert replied, which provoked a new rebuttal by Habermas. These and other articles were later published together and repeatedly reissued as an anthology entitled *The Positivist Dispute in German Sociology*.

The positivist dispute quickly created a sensation which went far beyond the simple methodological issues. Many people regarded it as a political dispute between critical rationalism and the Frankfurt school, whereby one side took up the banner of Max Weber and the other the banner of Karl Marx. Today, with 50 years hindsight, one can say that the critical rationalists won on all the major issues. Even the former protagonists in the dispute, particularly Jürgen Habermas, have moved away from their previous neo-Marxist positions, and in substance (but not in vocabulary or style of expression) have moved closer to the positions held by critical rationalism. As its final result, the positivist dispute caused a realignment of the German social sciences. It led to a rehabilitation of analytic approaches in social philosophy, which had been banished from the German-speaking world by the Nazis, as well as to a clear convergence with the scientific style of the Anglo-American world.

Criticism of Whether It Is Possible to Prove an Ultimate Truth: The 'Ultimate Justification Problem'

In 1968 in his book *Treatise on Critical Reason*, Albert grappled with the classic problem of justification, i.e., whether it is possible to know something for certain, much in the style of Descartes' search for a foundation for knowledge. Albert's answer was that it is impossible to get a justification which is certain. In order to make his position plausible, he introduced the so-called *Münchhausen Trilemma*, which, to this day, is still closely associated with the name Hans Albert:

> If one requires a justification for everything, one must also require a justification for the knowledge used to obtain the justification. This leads to a situation where there are three alternatives, all of which appear unacceptable. This is a trilemma, which I want to call the *Münchhausen Trilemma*, because the problem we seek to solve is analogous to the problem Münchhausen faced. Namely, one only has a choice between:
>
> 1. an infinite regression, which appears because of the necessity to go even further back, but is not practically feasible and does not, therefore, provide a certain foundation;
> 2. a logical circle in the deduction, which is caused by the fact that one, in the need to found, falls back on statements which had already appeared before as requiring a foundation, and which circle does not lead to any certain foundation either; and finally,
> 3. a break of searching at a certain point, which indeed appears principally feasible, but would mean a random suspension of the principle of sufficient reason.
>
> Albert, 1968/1991: 15, translation from German slightly modified

Reference to the Münchhausen Trilemma is supposed to show that a certain 'final' justification is not possible. This idea reflects the logical fallibilism which according to Albert should be observed not only in theoretical philosophy, but also in social philosophy and legal philosophy. The trilemma is certainly not limited to deductive conclusions, but rather it is relevant whenever for the justification of one thesis, a different thesis has to be cited.

The Relationship Between Natural Sciences and Social Sciences and the Problem of Objectivity

The three pillars of critical rationalism – namely logical fallibilism, the idea of critical examination, and the approach of critical realism – do apply not only in the natural sciences but also in the social sciences and humanities. Albert rejects the Humanities' claims for autonomy, as have been traditionally expressed in Germany, for example by the Hermeneutics around Gadamer, or the Frankfurt School around Adorno, Horkheimer, and Habermas. The standard model of scientific explanation, which goes all the way back to Carl Gustav Hempel, is, according to Albert, applicable in the social sciences and humanities. Advocates of a special status for the social sciences and humanities argue that human language and other human artifacts cannot be explained, but only *understood*. In contrast, Albert pointed out that 'explaining' and 'understanding' take place in different contexts, namely understanding in the context of discovery and explaining in the context of justification. In his later works, Albert modified this position, arguing there could be a 'technology of understanding,' like hermeneutics in the eighteenth century. Albert emphasizes the methodological unity of the social sciences, which links him with a clear commitment to a naturalistic understanding of the world.

In particular, Weber's *Postulate of Objectivity* is applicable, according to Albert, just as much in the natural sciences as it is

in the social sciences and humanities. This means that researchers must distinguish clearly between facts and their own particular political or moral judgments, so as not to deceive the public about the status of statements made.

The *Werturteilsstrei*t (value judgment dispute), which took place in the 1920s between Max Weber and his critics, has played a prominent role to this day in the German social sciences. The positivist dispute, mentioned above, can be interpreted as a renewal of the earlier value judgment dispute. An understanding of these conflicts is complicated, however, by the fact that to a certain extent, widely different issues have been confounded. In order to remedy that, Albert distinguished the following subquestions:

1. The *logical problem* concerns the question of what meaning to give value judgments. What is important here is primarily the question of what is the difference between statements of fact and value judgments.
2. The *methodological problem* concerns the issue of whether value judgments are necessary in Science, or expressed another way, what goals would make it necessary to use value judgments in Science.
3. The *definitional problem* is whether a language system which includes value judgments can be regarded as 'scientific.'
4. Finally, there is the moral problem whether scientists, not only as private persons but also as teachers in science, can or should express value judgments. This problem is called the *Problem of Kathederwertung* (lectern value judgment) in German.

Statements of fact like 'the picture is 100 years old' and value judgments like 'the picture is beautiful' possess of course the same surface grammar, but differ significantly in their logical content. Albert follows here the results of the Anglo-Saxon metaethics (Stevenson, Ayer), which he adopted quite early and helped spread in the German language area. Statements of fact describe reality. They can be true or false. In contrast, value judgments express, in addition to their descriptive contents, also an assessment, which means a personal opinion of the person making the value judgment. This personal opinion can be political, moral, or, as the example above shows, aesthetic in nature.

The methodological problem of whether value judgments are necessary in the social sciences, Albert answers firstly, by differentiating three questions, namely (1) Can value judgments be the object of social science research and predictions? (2) To what extent do scientific propositions in the social sciences contain such value judgments? (3) To what extent do such value judgments form the basis for scientific propositions in the social sciences? Albert's answer to the first question is unreservedly 'yes.' It is one of the main tasks of the social sciences to examine value judgments. Albert answers the second question, following Max Weber, saying that the social sciences should limit themselves to making available knowledge. It is not the task of the social sciences to make political or moral value judgments. Albert answers the third question, about the extent to which value judgments form a basis for social science research, with affirmation. Even the decision to conduct a scientific enquiry contains a value judgment. The same is true for the area of the research, the research subject, and the methodology chosen.

In summary, Albert felt that it was unnecessary or even dangerous to confound factual statements from the social sciences with moral or political value judgments, because such statements would confuse the general public, which would not be able to distinguish which were facts and which opinions. An academic activity, in which hidden value judgments and other prejudices abound, should not be referred to as a science. It was right so that Albert considered it to be misguided for university professors to mix together factual statements and value judgments when conducting their lectures, without drawing the audiences' attention to what they were doing.

Albert's rigorous defense of the postulate of scientific value freedom has often been criticized as 'conservative,' particularly by representatives of the Frankfurt School such as Adorno or Habermas, who in the late 1960 and 1970s frequently combined academic research with political postulates from neo-Marxist ideology. This led quickly to criticism. The philosopher Martina Plümacher wrote:

> The idea that Science could explain 'facts' and political positions like demands for the fulfillment of basic human needs, but should not preemt political or cultural decisions as a substitute for open public discussions …is fundamentally a radically democratic idea, the roots of which lie in the convictions of the Enlightenment, that all human beings are fundamentally able to think and to come to reasonable, responsible decisions. It reflects a belief in rational solutions to conflicts using democratic institutions and assumes the existence of active political participation by citizens in a democratic state.
>
> Plümacher, 1996: 145

National Economy and Law

National Economy (today usually referred to as Economics) and Law possess in Germany a common disciplinary background in the subject 'Political Economy,' even though the subjects have diverged significantly since the mid-twentieth century. Economics has always worked mainly with idealized assumptions about reality and sees itself as a theory of rational behavior. Mathematical models continue to be used in place of empiric research.

Early on, Albert criticized the methodology employed in Economics as '*model platonism*,' advocating that Economics makes use of the research results of empirical sociology to a far-greater extent than ever before. An overly mathematical Economics was in danger of immunizing itself from real experience.

Albert also expressed opinions on the subject of Law several times. Within the methodology of continental legal science, the view prevails that jurisprudence is a *dogmatic normative* and *hermeneutic* discipline, which means that it assumes that the law is found in statutes, and it attempts, by interpreting the statutes, to derive solutions to resolve concrete legal disputes. In several works, Albert subjected continental jurisprudence's understanding of itself to detailed criticism. Subsequently he made a case to Rudolf von Jhering for a teleological or 'socially technological' understanding of the law. American legal realists, especially Karl Llewellyn, advocated a very similar model. Statutes and the interpretation of statutes are then a means to achieve certain agreed-upon goals.

This model seems obvious today for statute law: Parliament passes laws in order to achieve its legislative goals. The utility of a particular statute, then, lies in its ability to achieve the goals intended by the legislature. Thus, an empirical, truly scientific criterion is used which allows scientifically supported statements to be made about the suitability of statute laws to achieve certain specified purposes.

The situation is very similar, according to Albert, with respect to legal practice. A lawyer works out potential interpretations of a statute, which may be more or less suitable, to achieve the legislative objectives of the statute. When choosing the 'right' construction of a statute, empirical knowledge plays a role. In more recent legal discourse, the term 'result oriented' judicial decision making has come into use.

Ethics

A question, which is one of the most renowned traditional problems in Ethics, is whether moral judgments are either a question of *cognition* or a question of *decision making*. At first glance, statements like 'Murder is wrong' or 'Thou shalt not murder' seem to reflect an established legal or moral order. Whether, for example, the rule exists in the social morality of England or Germany, that one should not murder, would then be a question of cognition. Nevertheless, the question of whether the precept 'Thou shalt not murder' is right or wrong is generally not understood with reference to a particular system of social morality. Rather, the question which is asked is whether it is right or wrong in general, regardless of whether it occurs in one system of social morality or another.

Moral and social philosophers have found two types of answers to questions of this kind: *Cognitivists* are of the opinion that we can recognize what is morally right or wrong. This position presumes that for mankind what is right or wrong has been previously established in some way; for example, the same way that reality determines our responses to questions of fact. Prominent cognitivist ethical systems include Platonist values, as found in Plato's works, traditional natural law, and Kant's moral philosophy. Such positions, however, have always met with criticism. Skeptics consider them to be refuted. They prefer noncognitivist moral systems, in which norms and values have not been prescribed for mankind, but rather came into existence as the product of human decision making. The most important proponents today of this so-called noncognitivism come either from the logical empiricist or from existential philosophy camps. A single statement by Wittgenstein characterizes both of these movements: "We feel that even if all possible scientific questions are answered, the problems in our lives remain untouched" (Ludwig Wittgenstein, *Tractatus logico-philosophicus*, statement 6.52).

Hans Albert takes what is essentially a noncognitivist position in this dispute, but he considers the radical differentiation between cognition and decision making to be exaggerated. Of course ultimate justification is in his opinion impossible not only in the area of factual knowledge but also in the normative area. The Münchhausen Trilemma is applicable in both areas. On the other hand, our knowledge already contains numerous determinations and therefore decisions, for example, with respect to the selection of objects of enquiry and fixing the meaning of the vocabulary used. Decision is therefore also an aspect of our experience.

According to Albert, the idea of critical examination, and the methodologies derived from it of construction and criticism, can also be successfully applied to normative questions. The *unity of reason* links all kinds of human problem solving. Value axioms and normative systems based on these axioms can be analyzed, for instance, for their logical consistency. A second approach which links cognition and decision making is the identification of the factual consequences resulting from the acceptance of certain values and norms. These examples show that the idea is incorrect that in the field of morality no possibility exists for rational assessment.

Morality is not invented at the drawing board. Rather, all people are born into a certain social morality, the values and standards of which appear to them as natural. Albert refers illustratively to *value Platonism* of the natural orientation of the world. Moral philosophy has the task of critically examining established morals:

> For a critical moral philosophy its main task is not to provide the morality prevailing at any time with dubious justifications, in order to anchor it more firmly in the consciousness of the people and the social order. Its job is rather to critically examine that morality, uncovering its weaknesses, and developing ways to improve it.
>
> Albert, 1968/1991: 90

A fundamental principle of any critical moral philosophy is that it is impossible to derive norms from statements of fact (or vice versa). In philosophy, this principle is typically referred to as 'Humes' Law.' Whoever nevertheless purports to derive a 'should' from an 'is,' or an 'is' from a 'should,' is committing a logical error which is usually referred to as the 'naturalistic fallacy.'

Albert proposed the so-called *bridge principles*, in order to overcome the logical barrier between statements of fact and values/norms. In this way, normative statements would be made accessible to evidence-based criticism. The first bridge principle he calls the principle 'should implies can.' Based on this principle, norms which require the impossible can be criticized, and, under certain circumstances, rejected. Albert also calls this the *principle of feasibility*.

Another bridge principle suggested by Albert is called the 'congruence postulate': Norms or values, which, to make sense assume the existence of scientifically unknowable factors or entities, should be rejected. Our normative constructs must, therefore, if we wish to comply with the congruence postulate, be consistent with the scientific world view.

Albert did not complete his formulation of bridge principles, but rather left the development of further principles for scientific research to be developed by others. Taken all together, he draws a positive balance on Ethics:

> It is therefore possible to overcome the positivist resignation in questions of moral philosophy, without falling into the existentialist cult of engagement, which replaces rational discussion of such problems with irrational decisions. The critical philosophy, which provides us with this possibility, also has in itself moral content. Whoever assumes this approach for himself has not decided for an

> abstract principle void of existential meaning, but rather for a way of life. An immediate ethical consequence of critical philosophy is that unshakeable faith, which is rewarded in some religions, and is inaccessible to rational argument, becomes a vice rather than a virtue.
>
> Albert 1968/1991: 94 f.

Criticism of Religion

Since the 1970s, Albert has repeatedly occupied himself with the subject of Theology, subjecting the arguments of prominent theologians like Hans Küng or Josef Ratzinger to critical analysis. In doing so, Albert wanted, among other things, to show that no rational solution exists to the *theodicy problem*: how can suffering in the world be reconciled with the idea of an infinitely good and infinitely powerful God?

Albert sees the world today as a world in religious crisis. People are no longer prepared to accept uncritically the dogmas and privileges of the great institutionalized churches, i.e., the Protestant and Catholic Churches, and they are turning to new religious movements which compete with the traditional churches for members. Albert argues for an ethics of 'practicing charity' in the tradition of Albert Schweitzer, but he rejects the traditional belief in God. He considers life after death and rewards in the hereafter for suffering in this life to be illusions. Instead, he argues for an – admittedly highly fragile – autonomous meaning of life of the individual:

> There cannot be a guarantee that our finite lives on earth have meaning. Such a life can in any case only have meaning, to the extent that we fill it with meaning, by facing tasks, the completion of which we find important, by devoting ourselves to activities which are intrinsically satisfying or which help us to fulfill our chosen tasks. The fact that all the happiness on earth is temporary, is not in itself a reason to see this happiness as worthless, and the transience of all things, also those, which we cherish most, does not make the value we have given them in any way illusory.
>
> Albert, 1984. Der Sinn des Lebens ohne Gott. In Hoerster, N. (Ed.), Religionskritik. Arbeitsbuch für den Unterricht, Stuttgart, p. 120

Appraisal

As a philosopher of science and social scientist, Albert made substantial contributions to establishing in Germany (again) an empirically oriented style of thinking, directed toward clarity of thought and rational verifiability, as has prevailed in the social sciences of Anglo-Saxon countries for many years. It is not an exaggeration to say that this style of thinking also clearly dominates the German-speaking world today, some 60 years after Albert's first scientific publication.

Albert's efforts were also influential in raising awareness of the importance of scientific knowledge based on factual evidence in subjects such as economics and law. In any case, it appears that a rethinking has started in the field of Economics in the wake of the world financial crisis, which all the mathematical models were unable to predict, let alone contain. There are also more and more voices in the field of Law, calling for a convergence with Albert and American legal realism. The notion that laws are not an expression of some higher reason, but are simply the means at the disposal of the legislature to bring about intended social changes, should in the meantime be commonplace.

Hans Albert has succeeded in reestablishing in the German-speaking world the tradition of critical social thought, which had been banished here in the 1930s and early 1940s. If one wanted to describe him and his work in a single phrase, *Defender of the Enlightenment* would probably be most apposite.

See also: Critical Rationalism; Critical Theory; Enlightenment; Ethics and Values; Habermas, Jürgen (1929–); Hermeneutics; Logical Positivism and Logical Empiricism; Naturalism; Popper, Karl Raimund (1902–94); Weber, Max (1864–1920); Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Selected Bibliography of the Works of Hans Albert

a) Monographs

Ökonomische Ideologie und Politische Theorie, Göttingen, 1954. Ökonomische Theorie als politische Ideologie. Das ökonomische Argument in der ordnungspolitischen Debatte. published as, 3rd edition. Tübingen. 2009.

Traktat über kritische Vernunft, 5th ed, 1968. Tübingen. 1991 (English edition published as: Treatise on Critical Reason, Princeton 1985).

Träumereien, Transzendentale, 1975. Karl-Otto Apels Sprachspiele und sein hermeneutischer Gott. Hamburg.

Traktat über rationale Praxis, 1978. Tübingen.

Das Elend der Theologie, 1979. Kritische Auseinandersetzung mit Hans Küng. Hamburg, 3rd extended edition Aschaffenburg 2012.

Die Wissenschaft und die Fehlbarkeit der Vernunft, 1982. Tübingen.

Freiheit und Ordnung, 1986. Zwei Abhandlungen zum Problem einer offenen Gesellschaft. Tübingen.

Kritik der reinen Erkenntnislehre, 1987. Das Erkenntnisproblem in realistischer Perspektive. Tübingen.

Rechtswissenschaft als Realwissenschaft, 1993. Das Recht als soziale Tatsache und die Aufgabe der Jurisprudenz. Baden-Baden published in Würzburger Vorträge zur Rechtsphilosophie, Rechtstheorie und Rechtssoziologie, vol. 15.

Das Ideal der Freiheit und das Problem der sozialen Ordnung. Friedrich von Hayek-Vorlesung, 1994. Freiburg i.Br. 1994.

Kritik der reinen Hermeneutik, 1994. Der Antirealismus und das Problem des Verstehens. Tübingen.

Kritik des transzendentalen Denkens, 2003. Von der Begründung des Wissens zur Analyse der Erkenntnispraxis. Tübingen.

Erkenntnislehre und, Sozialwissenschaft, 2003. Karl Poppers Analyse sozialer Zusammenhänge. Vienna.

Rationalität und, Existenz, 2006. Politische Arithmetik und Politische Anthropologie (first published as a dissertation in 1952). Tübingen.

Kontroversen verstrickt, 2nd ed, 2007. Vom Kulturpessimismus zum kritischen Rationalismus, Berlin. 2010.

Josef Ratzingers Rettung des Christentums, 2008. Beschränkungen des Vernunftgebrauchs im Dienste des Glaubens. Aschaffenburg.

Macht und Gesetz, 2012. Grundprobleme der Politik und der Ökonomik. Tübingen.

b) Anthologies

Marktsoziologie und Entscheidungslogik, 1998. Ökonomische Probleme in soziologischer Perspektive, Neuwied & Berlin 1967, new edition published as Marktsoziologie und Entscheidungslogik. Zur Kritik der reinen Ökonomik, Tübingen.

Plädoyer für kritischen Rationalismus, 4th ed, 1971. München. 1975.

Konstruktion und Kritik, 1972. Aufsätze zur Philosophie des kritischen Rationalismus, 2nd ed. Hamburg. 1975.

Aufklärung und Steuerung, 1976. Aufsätze zur Sozialphilosophie und zur Wissenschaftslehre der Sozialwissenschaften. Hamburg.
Kritische Vernunft und menschliche Praxis, 1977. Mit einer autobiographischen Einleitung. Stuttgart.
Rationalismus, Kritischer, 2000. Vier Kapitel zur Kritik illusionären Denkens. Tübingen.
Hans Albert Lesebuch. Ausgewählte Texte, 2001. Tübingen.
Kritische Vernunft und rationale Praxis, 2011. Tübingen.
Kritik des theologischen Denkens, 2013. Berlin.

c) Works edited by Hans Albert

Theorie, Realität, 1964. Ausgewählte Aufsätze zur Wissenschaftslehre der Sozialwissenschaften, 2nd ed. Tübingen. 1972.
Adorno, T.W., Dahrendorf, R., Habermas, J., Pilot, H., Popper, K.R., 1969. Der Positivismusstreit in der deutschen Soziologie (many editions), with. Neuwied & Berlin.
Werturteilsstreit, Darmstadt, 3rd. Ed, 1971. 1991 (published in Wege der Forschung, vol. 175), with E. Topitsch.
Theorie und Erfahrung, 1979. Beiträge zur Grundlagenproblematik der Sozialwissenschaften. Stuttgart (with K. Stapf).

d) Miscellaneous writings

Baum, W., 1997. In: Feyerabend, Paul, Albert, Hans (Eds.), Briefwechsel. Frankfurt a.M.
Morgenstern, M., Zimmer, R., 2005. In: Albert, Hans/Popper, Karl (Eds.), Briefwechsel 1958–1994. Frankfurt a.M.
Zimmer, R., Morgenstern, M. (Eds.), 2011. Gespräche mit Hans Albert, Berlin.
For a detailed bibliography. http://www.opensociety.de/Web1/Albert/albert.htm.

Selected literature about Hans Albert

Becker, W., 1989. Kritischer Rationalismus oder Kritizismus? Zur Frage der Übertragbarkeit der kritisch-rationalen Grundidee auf die Politik. In: Salamun, K., Karl, R. (Eds.), Popper und die Philosophie des Kritischen Rationalismus. Zum 85. Geburtstag von Karl R. Popper, Amsterdam & Atlanta, pp. 203–220.
Bohnen, A., Musgrave, A. (Eds.), 1991. Wege der Vernunft. Festschrift zum 70. Geburtstag von Hans Albert. Tübingen.
Dahms, H.J., 1994. Positivismusstreit. Die Auseinandersetzungen der Frankfurter Schule mit dem logischen Positivismus, dem amerikanischen Pragmatismus und dem kritischen Rationalismus. Frankfurt a.M.
Ebeling, G., 1973. Kritischer Rationalismus? Zu Hans Alberts Traktat über kritische Vernunft. Tübingen.
Engel, G. (Ed.), 2001. Schwerpunkt: Hans Alberts Kritischer Rationalismus. Aufklärung und Kritik, Sonderheft 5, Nürnberg.
Gadenne, V., Wendel, H.J. (Eds.), 1996. Rationalität und Kritik. Tübingen.
Haltmayer, S., 1978. Hans Albert oder Keine Alternative. Wiener Jahrbuch für Philosophie 11, 153–179.
Hilgendorf, E., 1997. Hans Albert zur Einführung. Hamburg.
Hilgendorf, E. (Ed.), Wissenschaft, Religion und Recht. Hans.

Alcohol Interventions: Disease Models vs. Harm Reduction

Thomas Hall, University of Central Florida, Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The disease model of alcoholism is based on the assumption that alcoholism is an incurable, progressive disease. Abstinence from any alcohol use is prescribed within the disease model. In recent years, harm reduction has emerged as an alternative to the disease model. Harm reduction is based on a public health approach to alcohol abuse. The public health model recognizes the benefit of progressive steps toward decreasing alcohol use and associated physical and emotional harm. Evaluation research and survey data indicate that a public health approach is a viable alternative to the disease model.

Introduction

Typically a lag exists between dissemination of scientific innovation and clinical practice. This is especially true of alcohol-dependency treatment. Applied social and behavioral studies provide both researchers and clinicians opportunities to meet the challenges of responding to contemporary norms, attitudes, and beliefs about alcohol dependency. Evaluation of intervention strategies allows for the development of best practices.

The examination of alcohol abuse treatment efforts starts with a brief review of three treatment modalities and is followed by a description of three types of peer-led support groups. The treatment modalities were the basis of the Project MATCH and COMBINE study interventions. Peer support groups are discussed because of the confusion that exists in the general population regarding the difference between treatment and support. Each peer-led support strategy reviewed is based on one of the treatment modalities examined.

The Project MATCH and COMBINE studies were benchmark studies funded by the National Institute of Alcohol Abuse and Alcoholism (NIAAA) (Willenbring, 2008; Bühringer and Pfeiffer-Gerschel, 2008; Anton et al., 2003). The design of each study allowed researchers to determine the efficacy of a range of treatment alternatives and a philosophy of care. Both studies led to innovation in the delivery of substance abuse treatment. The evaluation literature of non-disease-based treatment modalities was scant prior to both Project MATCH and COMBINE (Willenbring, 2010).

In addition to Project MATCH and COMBINE, the results and implications of the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) impacted the delivery of alcohol-dependency treatment. The NESARC was a longitudinal survey designed to estimate the prevalence of alcohol use disorders and assess changes over time. The first wave of interviews was conducted in 2001–2002 and the second wave in 2004–2005.

A brief history of research and evaluation of alcohol dependence follows. In addition, the disease model of alcoholism is compared to the public health model. Harm reduction is an important component of the public health model, but it is not universally accepted. Critics contend that harm reduction does more harm than good. Typical treatment modalities are also reviewed, along with aftercare or peer support groups. Lastly, the impact of Project MATCH, COMBINE, and NESARC on the treatment of alcohol use disorder is reviewed.

The Center for Alcohol Studies

After the repeal of prohibition, scientists were eager to study the medical characteristics of alcohol use. The Center for Alcohol Studies at Yale University was the first academic research group created to study alcoholism (Freed, 2011). The center evolved from an interest group in the Applied Physiology department at Yale University in 1935. The aims of the center were the investigation of alcoholism and the development of medical treatments to treat alcohol use disorder. During the 1940s and 1950s, the center was the primary home for alcohol use research and treatment in the United States. In 1961, the center relocated to Rutgers University where it remains to this day.

E. Morton Jellinek, a biostatistician and founding member of the center, was the first to systematically study behavioral traits of alcoholism (Candon, Ward and Pandina, 2014). He surveyed a convenience sample of Alcoholics Anonymous (AA) members in New England. Based on his findings, he hypothesized that alcoholism was a result of 'disease process.' Jellinek proposed that the difference between alcohol use and the disease of alcoholism was a loss of control over drinking. Despite methodological flaws, his work framed alcoholism research and treatment for almost two decades.

The Disease Model of Alcoholism

The disease model was officially recognized by the American Medical Association in 1958. The American College of Physicians recognized alcoholism as a disease in 1969. Early on, the disease model was imprecise in identifying those who suffered from this relatively new diagnosis. Alcoholism could be diagnosed based on concerns about problematic drinking reported by family members or close associates or by clinical observations.

Defining alcohol as a disease implies there is a cure or at least some palliative to be found through medical treatment. For most, treatment for alcoholism required abstinence from alcohol use and participation in self-help groups, such as AA.

The concept of alcoholism as a disease replaced the stigma of moral censure and criminal punishment. Stanton Peele, a critic of the disease model, has noted, "Perhaps the most dire consequence of the disease model of addiction is that it has encouraged the abdication of individual responsibility for outrageous conduct" (Peele, 1989; p. 21). He concluded, "creating a world of addictive diseases may mean creating a world in which anything is excusable" (Peele, 1989; p. 16).

Legal, religious, and behavioral or medical 'experts' defined alcoholism as a physical condition that must be socially constrained and medically treated. Addictions therapists were granted 'expert' status to treat alcoholism. Historically, behavioral health treatment in America was divided into two camps, mental health providers and addiction providers. Mental health providers typically earned terminal degrees, whereas addiction providers often did not. Substance abuse treatment was the purview of recovering 'addicts,' while behavioral health providers were located on the periphery. This schism between researchers and addiction providers exists today. Addiction providers are typically cautious about adopting new approaches to treatment (Walters and Rotgers, 2012).

Harm Reduction and the Public Health Model

In the late twentieth century, the public health model of alcohol intervention changed the narrative from curing disease to promoting health (Russell et al., 2011). Prevention and treatment of alcohol abuse included non-recovering addiction providers. Harm reduction strategies now aimed to manage symptoms and decrease risk versus curing a disease. Harm reduction influenced the shift from the disease model to the public health model of prevention and treatment. There was a great deal of angst among treatment-as-usual addictions providers such that non-recovering researchers, psychiatrists and physicians, sociologists, and psychologists did not understand the implications of harm reduction. Perhaps lost in this conversation was an acknowledgment that the founders of the Center for Alcohol Studies were also non-recovering scientists. Their development of the disease model was, at the time, an effort to get people to look on alcoholism as a disease as opposed to a moral failing or character flaw.

As early as the 1970s, Alan Marlatt, Linda, and Mark Sobell challenged assumptions that alcoholism was an incurable disease (Sobell and Sobell, 1978; Marlatt and Dillworth, 2005). They were key figures in speaking out for change. They advocated for the inclusion of moderate drinking as an alternative to the prevailing abstinence-only injunction. Harm reduction was based on a continuum of alcohol consumption that ranged from heavy drinking to abstinence. Alcoholic patients were taught how to reduce the quantity and frequency of their alcohol use as well as associated negative consequences. Marlatt encouraged individuals to choose the goal that best fits their needs.

In the 1990s concerns related to college drinking led to several interventions. Marlatt developed health promotion programs for students based on teaching how to drink responsibly (Baer et al., 1989). His interventions taught college students harm reduction tactics and controlled drinking strategies. His campus harm reduction programs remain germane to college health in the twenty-first century (Dimeff et al., 1999; Whiteside et al., 2010).

Managing consequences associated with alcohol abuse was a significant conceptual shift. Harm reduction encourages goals of safer drinking, reduced drinking, moderate drinking, or abstinence. Critics of harm reduction point out that tolerating risky or illegal conduct sends the message that risk-laden behaviors are acceptable. They also characterize harm reduction as ineffective. They assume it does not reduce harm and over time may increase negative consequences.

Project MATCH Intervention Strategies

Project MATCH was the first multisite federally funded longitudinal study of alcoholism. The study was initiated in 1989 and continued through 1997. The cost of the study was over 27 million US dollars. Project MATCH evaluated the effectiveness of twelve-step facilitation (TSF), cognitive behavioral theory (CBT), and motivational enhancement theory (MET). The study was designed to evaluate the relevance of matching patients with specific treatment modalities. Study investigators tested the assumption that patient outcomes were related to the fit between treatment modalities and individual patient characteristics.

Cognitive Behavioral Therapy

CBT is based on both Pavlov's theory of classical conditioning and B.F. Skinner's theory of behavior modification (Tatarsky, 2002). Behaviorism could not account for all types of behavior; hence CBT was developed to integrate abstract thoughts and feelings with concrete behaviors. CBT has been typically described as problem focused and action oriented. Individual choice is an important construct. Clients are accountable for their cognitive approach to 'problems' and their choice of behaviors. Action is located in the here and now (Tatarsky, 2002).

CBT does not ignore past trauma, shame, or guilt; however, action and accountability cannot occur in the past or future. Action is possible only in the present. Within the cognitive behavioral paradigm, all behavior is conceived of as changeable if clients are motivated to change. Clinicians structure session content to increase interpersonal competence and decrease interpersonal vulnerabilities. Substance abuse intervention utilizes cost/benefit analyses, decisional balancing, and relapse prevention planning. These techniques require the client to engage actively in the present through the use of behavior logs. For example, drink-monitoring cards record the elapsed time and number of standard drinks an individual consumes in a given day and the social context of their drinking behavior. Clinicians process their daily drinking logs with patients to assess any manifestation of cognitive dissonance.

Motivational Enhancement Theory

MET was designed to encourage a feeling of optimism among stigmatized individuals (Scholl and Schmitt, 2009; O'Leary Tevyaw and Monti, 2004; Miller and Rose, 2009). MET

focuses on the level of insight or awareness individuals exhibit regarding the risk associated with their alcohol use. Stein et al. (2009) found that integrating MET techniques in the earliest phases of treatment had positive effects on patient engagement over the course of treatment. Motivational interviewing strategies include the use of open-ended questions, reflective listening, affirmations, and summarizing. MET evaluates patient 'readiness for change' on a continuum that begins with little insight (pre-contemplation) and moves progressively toward self-determination (action and maintenance).

Twelve Step Facilitation Therapy

TSF therapy is based on the first five traditions of AA. TSF assumes that alcoholism is a progressive illness that affects the body, mind, and spirit for which the only effective remedy is abstinence from the use of alcohol (Donovan et al., 2013). TSF posits that without treatment, patients do not accept the reality of their loss of control over drinking. Alcoholism impairs coping mechanisms and interpersonal relationships and ultimately leads to alienation from self and others.

TSF encourages patients to accept that they are victims of a progressive, incurable disease and to surrender their will to a 'higher power.' Kelly et al. (2012) found that the main effect of TSF was the changing of patients' social network and relationships with family members and associates. In addition to alcohol abuse treatment provided by professional counselors, participation in peer-led 12-step support groups improved treatment outcomes. Dawson et al. (2006b) found that 12-step participation posttreatment doubled the chance of abstinence as compared with attendance in formal treatment and no 12-step participation. Dawson asserted that participation in 12-step programs was noteworthy for achieving and maintaining abstinence. Streifel and Servanty-Seib (2006) surmised that participation in AA provided needed social support to manage anger, loss, and grief associated with sobriety.

Aftercare: Selected Peer-Facilitated Support Groups

Support groups are an important adjunct to formal treatment (Kelly et al., 2006). Support groups provide opportunities for individuals in recovery to develop coping skills to manage their sobriety. An important factor in the success of support groups is a belief that recovery from alcoholism is possible. Listening to the successes and setbacks of others who are alcohol dependent can inspire hope for change. Recovery becomes a real experience, separate from the prescribed therapeutic factors of professionally led treatment programs. Twelve-step support groups are the most widely recognized peer-facilitated support groups.

Other peer-facilitated groups have been successful in supporting abstinence or reductions in substance use. SMART Recovery is a non-faith-based alternative to AA (Horvath, 2000). Both SMART Recovery and AA support groups are based on abstinence from alcohol or any other illicit substance use. SMART Recovery is grounded in CBT and offers a 'tool kit' for members to assist with managing cognitive distortions or irrational beliefs that lead to maladaptive coping mechanisms.

The Harm Reduction, Alcohol Abstinence and Management Support (HAMS) is also based on CBT (Anderson, 2010). However, HAMS is not an abstinence-based support group. HAMS incorporates stages of change theory and harm reduction, and meets individuals 'where they are' as opposed to where others believe they need to be. Harm reduction-inspired support groups remain controversial as critics contend they lead to relapse and do more harm than good.

Alcoholics Anonymous

The 12-step support group AA is the largest of all the 12-step programs. Narcotics Anonymous is the second largest and accommodates members who are recovering from addiction to alcohol and other drugs. However, there are over 30 different 12-step support groups covering a broad range of addictions. Each group is unique in its behavioral manifestation of addiction, yet all have the same recovery goals: physical, mental, emotional, and spiritual well-being.

The 12-step programs are grounded in the disease model. The physical dimension has been described as an allergen. Those who are afflicted with the disease of alcoholism cannot drink moderately. A core belief has been the 'powerlessness' of an individual over substance abuse. This was a key premise in Jellinek's disease concept. These groups advocate that the admission of powerlessness within a nonjudgmental social support system creates a context where letting go of shame and guilt is possible. Spiritual components of the program are included to instill hope for change. The process of working through universal 'steps' is intended to replace self-centered beliefs and actions with an awareness of the feelings and needs of others.

SMART Recovery

SMART Recovery was developed as a peer-facilitated support group based on cognitive behavioral strategies. MET and recognition of individual readiness to change underlie the SMART Recovery goals. Self-sufficiency and self-confidence are fundamental features of SMART meetings. SMART Recovery promotes self-directed change.

SMART Recovery advocates for the use of psychological treatments and physician-prescribed medication that manage urges to drink or mediate the physical effects of drinking. SMART Recovery adapts its educational material to incorporate new-scientific innovations. SMART Recovery supports (1) individual progression toward motivation to remain sober; (2) strategies to cope with urges to drink; (3) dealing with attitudes, beliefs, and behaviors; and (4) mindfulness. Similar to AA, SMART Recovery promotes abstinence as the best outcome for managing alcoholism.

Harm Reduction, Abstinence, and Moderation Support

Similar to SMART Recovery, HAMS also provides peer-group support that does not promote faith-based norms. However, unlike SMART Recovery, HAMS is not abstinence based. HAMS

is informed by harm reduction within this paradigm; harm reduction is defined as a nonjudgmental approach that provides individuals with support to make desired changes in their behavior. HAMS attracts individuals who are interested in reducing alcohol-related harm or reducing alcohol use.

HAMS does not prescribe 'acceptable' alcohol use. HAMS accepts that for some, alcohol intoxication is not stigmatized, and the support groups teach individuals how to reduce harms. One of the concerns associated with AA is the loss of individual autonomy connected to powerlessness. A tolerance for and acceptance of individual choice distinguishes HAMS from other peer-facilitated support groups. The focus of HAMS is alcohol harm reduction; however, individuals who use alcohol and other drugs are welcomed and supported whether they choose safer use, reductions in use, or quitting altogether.

Implications of Project MATCH and COMBINE Research

Epidemiologic studies, such as Project MATCH and COMBINE, highlight the need for a change of focus in addiction treatment (Roizen and Fillmore, 2001). Project MATCH and COMBINE results demonstrated that effective treatment was multidimensional. No single strategy accounted for decreased substance use and associated harm. Project MATCH validated the efficacy of three types of behavioral interventions. It also challenged existing beliefs about effective treatment. The belief that some individuals would benefit from one treatment while others would benefit from another was discredited.

The Project MATCH study found that the patient–treatment matching did not significantly influence study outcomes. Treatment techniques did not predict treatment effectiveness (Moos et al., 1999; Babor, 2008). However, Project MATCH results also raised questions about the efficacy of any substance use treatment. The study results left some evaluators with two assumptions. The first assumption indicated that multiple treatment interventions work; one intervention was not more effective than the other. The second assumption indicated that none of the treatments made a difference. While some found the study results reassuring and interpreted these to be proof of the benefits of existing strategies, others suggested that none of the treatment modalities were notably effective.

Critics claimed that treatment success had less to do with treatment modality and more to do with subject perseverance and motivation. If methodical differences between interventions did not influence treatment success, then what did? Cutler and Fishbain (2005) suggested that the study results were a result of motivation. Study subjects who choose to seek treatment were more likely to reduce their drinking, and those who reduce their drinking are more likely to continue in treatment. Cutler and Fishbain noted the similarities between Project MATCH study results and those seen in studies of depression. The act of enrolling in a study sometimes leads to improvement in symptoms prior to the initiation of treatment (Cutler and Fishbain, 2005).

Stanton Peele (1998) challenged the relevance of Project MATCH study conclusions. Similar to Cutler and Fishbain's comments, Peele suggests that the study design was flawed because of the lack of a nontreatment control group. Peele also questioned why drinking frequency and quantity reductions were considered positive treatment outcomes for all treatment modalities. The TSF treatment manual stated intervention goals were consistent with AA abstinence goals; therefore, he asks if any effect other than abstinence should be coded as treatment failure for the TSF facilitation group. Alan Marlatt observed that the study left ample room for anyone to use the results to bolster their ideological point of view (Bower, 1997).

The COMBINE Study

The results of the COMBINE study challenged traditional definitions of effective treatment for alcohol dependence (Willenbring, 2010). The study was the first to compare the effects of pharmacotherapies and behavioral or psychosocial intervention on treating alcohol abuse and dependency. The study is the largest alcohol-use-related pharmacotherapy research project to date. The study evaluated the efficacy of medical management, psychopharmacology, and a Combined Behavioral Intervention (CBI). The COMBINE study randomly assigned alcohol-dependent patients to one of nine different treatment conditions. Eight of the groups received a combination of medical management, pharmacotherapy, and CBI. One group received CBI only.

The efficacy of two oral medications (naltrexone and acamprosate) was tested. The study indicated medical management using naltrexone was beneficial for the treatment of alcohol abuse and dependence; however, acamprosate was not (Anton et al., 2006). This result was surprising given the results of previous studies that reported acamprosate was effective (Willenbring, 2010). The study found that naltrexone reduced relapse rates after treatment. Naltrexone is designed to act as an opioid receptor antagonist and is used primarily in the management of alcohol dependence and opioid dependence. However, the primary use of naltrexone has been for the treatment of alcohol dependence. In addition to decreasing relapse among alcohol-dependent patients, heavy drinking patients who chose to moderate their drinking also benefitted from naltrexone.

The COMBINE study indicated that cognitive behavioral interventions used in tandem with naltrexone yielded the most favorable results. However, placebo pills and meeting with a health-care professional also had a positive effect (Anton et al., 2006). The efficacy of naltrexone led to increased numbers of alcohol-dependent patients being treated in ambulatory medical settings, thus increasing opportunities for patients to be served outside of substance abuse treatment facilities. The COMBINE study provided a framework for medical management in health-care settings.

Results of the COMBINE study led a new generation of physicians to consider pharmacology as well as behavioral health strategies for the treatment of substance abuse. A better understanding of the neurobiological effects of alcohol and other drugs has led to the development of new pharmacotherapies. The COMBINE study found that the best treatment outcome included medical management, medication, and behavioral counseling. The assertion that traditional behavioral counseling alone is the least effective treatment has been controversial. Abstinence-only treatment programs that do not endorse pharmacotherapy dispute this finding.

National Epidemiologic Survey on Alcohol and Related Conditions

The NESARC is the most comprehensive alcohol-use survey to date. Data were collected regarding individual and family history of addiction and treatment, as well as comorbid psychological and psychiatric disorders. The results of this survey indicate most people who meet the criteria for alcohol dependence never seek professional help (Willenbring, 2010). These survey results led researchers to examine under what circumstances people recover from alcohol dependence without professional assistance. Spontaneous remission or natural recovery is not a new concept; however, prior to the NESARC estimates of natural recovery, prevalence rates were unknown. Natural recovery has been a topic in substance abuse treatment since the 1980s. Prior to the NESARC, natural recovery was often dismissed as an anomaly or spurious conclusion.

Dawson et al., (2006a) report findings from the NESARC to suggest that one in four participants who met the criteria for problem drinking received professional help for alcohol problems. Klingemann et al. (2010) found natural recovery to be more common among individuals who rejected treatment as usual. The characteristics of subjects who report 'spontaneous' or natural recovery are germane to substance-dependence research.

Characteristics of individuals who reported natural recovery include satisfaction with work, financial security, committed relationships, and less social pressure to quit drinking (Bischof et al., 2001; Dawson et al., 2006b). Russell et al. (2001) found that when drinking interfered with activities of daily living, those who naturally recovered reported less self-examination of their alcohol-use-related problems. They reported fewer disclosures with associates and little pressure to stop alcohol consumption from close associates, as compared to those who reported they were in treatment or had completed treatment. Bischof et al. (2007) found that for these individuals, social support better accounted for their sobriety. Supportive relationships with peers and family, as well as participation in self-help groups, were characteristic of natural recovery (Dawson, et al., 2006a). Granfield and Cloud (1996) found that middle-class subjects were more likely to recover without the help of professional intervention. Increased anxiety and concern about loss of income and class status were also important characteristics associated with natural recovery.

Conclusion

The foundation of scientific inquiry about alcoholism began in the middle of the twentieth century. During this time the Center for Alcohol Studies at Yale University was created. The realization that alcoholism cut across almost every field of human knowledge led the center to assemble a group of professionals that included a physician, a psychiatrist, a statistician, a psychologist, a sociologist, a lawyer, and an economist (Page, 1997). The aim of this group was to understand the mechanisms of addiction and develop effective treatment modalities.

Initial treatment models were based on Jellinek's research on a group of men who attended AA meetings in New England. Jellinek attempted to identify and classify the characteristics of alcoholics. His study was the most comprehensive study of alcoholism at the time. He sought to use the best-available science to improve the lives of those suffering from alcoholism.

In the twenty-first century, scientific inquiry related to genomics, pharmacology, and neuroscience continues to challenge our understanding of alcohol dependency. Because of science, there exists a better understanding of the interplay between genetic, social, psychological, and environmental factors on substance dependence. Understanding the interplay between social and environmental factors better explains how individuals may recover from substance dependence with or without any formal intervention.

Project MATCH and COMBINE contributed to the adoption of new strategies for treating alcohol dependence. The use of pharmacotherapy as an adjunct therapy for alcohol dependence is promising and increasing. The COMBINE study demonstrated the efficacy of psychopharmacological treatments. A greater understanding of the benefit of medical management and psychotherapy resulted from these research studies. The advantage of prescribing naltrexone for patients who have previously failed to respond to psychotherapy alone is encouraging. The success of various treatment modalities and inclusion of medical management indicate that previous assumptions about alcohol treatment are open for debate. Since the 1940s the AA model has dominated the narrative regarding alcohol-dependence treatment. The disease model is almost 75 years old and in many treatment programs the model continues to persist.

Evaluation and dissemination of alcohol use education, intervention, and public policy lead to best practices. Innovation in medical treatment has created opportunities for research and evaluation with regard to evidence-informed care of alcohol-dependence patients. The public health paradigm shifted the emphasis to include education, clinical intervention, and policy development. However, despite the dissemination of research on effective treatment strategies, the majority of community treatment programs in the United States continue to utilize psycho-education interventions and referral to AA as their primary treatment modality (Willenbring, 2010).

See also: Addictions: General Considerations; Alcohol Use among Young People; Alcohol Use and Abuse; Alcohol-Related Disorders; Cognitive Behavioral Therapy; Health Outcomes, Assessment of; Health Risk Perception; Health Social Work; Public Health as Social Science.

Bibliography

Anderson, K., 2010. How to change your drinking: A Harm Reduction Guide to Alcohol, second ed. HAMS Harm Reduction Network, New York.

Anton, R., Randall, C., Couper, D., et al., 2003. Testing combined pharmacotherapies and behavioral interventions in alcohol dependence: rationale and methods. Alcoholism-Clinical and Experimental Research 27 (7), 1107–1122.

Anton, R.F., O'Malley, S.S., Ciraulo, D.A., Cisler, R.A., Couper, D., Donovan, D.M., Zweben, A., 2006. Combined pharmacotherapies and behavioral interventions for alcohol dependence: the COMBINE study: a randomized controlled trial. Journal of the American Medical Association 295 (17), 2003–2017. http://dx.doi.org/10.1001/jama.295.17.2003.

Babor, T.F., 2008. Treatment for persons with substance use disorders: mediators, moderators, and the need for a new research approach. International Journal of Methods in Psychiatric Research 17 (Suppl. 1), S45–S49. http://dx.doi.org/10.1002/mpr.248.

Baer, J.S., Kivlahan, D., Fromme, K., Marlatt, G., 1989. A comparison of three methods of secondary prevention of alcohol abuse with college students: preliminary results. In: Løberg, T.T., Miller, W.R., Nathan, P.E., Marlatt, G.A. (Eds.), Addictive Behaviors: Prevention and Early Intervention. Swets & Zeitlinger Publishers, Lisse, Netherlands, pp. 157–171.

Bischof, G., Rumpf, H., Hapke, U., Meyer, C., John, U., 2001. Factors influencing remission from alcohol dependence without formal help in a representative population sample. Addiction 96 (9), 1327–1336.

Bischof, G., Rumpf, H., Meyer, C., Hapke, U., John, U., 2007. Stability of subtypes of natural recovery from alcohol dependence after two years. Addiction 102 (6), 904–908.

Bower, B., 1997. Alcoholics synonymous. Science News 151, 62–63.

Bühringer, G., Pfeiffer-Gerschel, T., 2008. [Commentary] COMBINE and MATCH: the final Blow for large-scale black box randomized controlled trials. Addiction 103 (5), 708–710. http://dx.doi.org/10.1111/j.1360-0443.2008.02162.x.

Candon, P., Ward, J., Pandina, R., 2014. The journal of studies on alcohol and drugs and the Rutgers center of Alcohol studies: a history of the evolution of alcohol research. Journal of Studies on Alcohol and Drugs 8–17.

Cutler, R.B., Fishbain, D.A., 2005. Are alcoholism treatments effective? The Project MATCH data. BMC Public Health 5, 1–11. http://dx.doi.org/10.1186/1471-2458-5-75.

Dawson, D., Grant, B., Stinson, F., 2006a. Maturing out of alcohol dependence: the impact of transitional life events. Journal of Studies on Alcohol 67 (2), 195–203.

Dawson, D., Grant, B., Stinson, F., Chou, P., 2006b. Estimating the effect of help-seeking on achieving recovery from alcohol dependence. Addiction 101 (6), 824–834.

Dimeff, L., Baer, J.,S., Kivlahan, D.,R., Marlatt, G.A., 1999. Brief Alcohol Screening and Intervention for College Students: A Harm Reduction Approach. Guilford Press, New York.

Donovan, D., Ingalsbe, M., Benbow, J., 2013. 12-step interventions and mutual support programs for substance use disorders: an overview. Social Work in Public Health 28 (3–4), 313–332. http://dx.doi.org/10.1080/19371918.2013.774663.

Freed, C.R., 2011. Teaching and learning guide for: in the spirit of Selden Bacon: the sociology of drinking and drug problems. Sociology Compass 5 (1), 121. http://dx.doi.org/10.1111/j.1751-9020.2010.00347.x.

Granfield, R., Cloud, W., 1996. The elephant that no one sees: natural recovery among middle-class addicts. Journal of Drug Issues 26, 45–61.

Horvath, A.T., 2000. Smart Recovery®: addiction recovery support from a cognitive-behavioral perspective. Journal of Rational-Emotive and Cognitive-Behavior Therapy 18 (3), 181–191.

Kelly, J., Stout, R., Zywiak, W., Schneider, R., 2006. A 3-year study of addiction mutual-self help group participation following intensive outpatient treatment. Alcoholism: Clinical and Experimental Research 30, 1381–1392.

Kelly, J.F., Hoeppner, B., Stout, R.L., Pagano, M., 2012. Determining the relative importance of the mechanisms of behavior change within alcoholics anonymous: a multiple mediator analysis. Addiction 107 (2), 289–299. http://dx.doi.org/10.1111/j.1360-0443.2011.03593.x.

Klingemann, H.H., 2001. Natural recovery from alcohol problems. In: Heather, N., Peters, T.J., Stockwell, T. (Eds.), International Handbook of Alcohol Dependence and Problems. John Wiley & Sons Ltd, New York, pp. 649–662.

Klingemann, H., Sobell, M., Sobell, L., 2010. Continuities and changes in self-change research. Addiction 105 (9), 1510–1518.

Marlatt, G., Dillworth, T.M., 2005. Controlled drinking, part 2: cheers to 30 years of moderation in action. Psyccritiques 50 (33). http://dx.doi.org/10.1037/05161022.

Miller, W., Rose, G., 2009. Toward a theory of motivational interviewing. American Psychologist 64 (6), 527–537.

Miller, J., 2008. 12-Step treatment for alcohol and substance abuse revisited: best available evidence suggests lack of effectiveness or harm. International Journal of Mental Health and Addiction (4), 568.

Moos, R.H., Finney, J.W., Ouimette, P.C., Suchinsky, R.T., 1999. A comparative evaluation of substance abuse treatment: 1. Treatment orientation, amount of care and 1-year outcomes. Alcoholism: Clinical and Experimental Research 23, 529–536.

O'Leary Tevyaw, T., Monti, P.M., 2004. Motivational enhancement and other brief interventions for adolescent substance abuse: foundations, applications and evaluations. Addiction 99 (Suppl. 2), 63–75. http://dx.doi.org/10.1111/j.1360-0443.2004.00855.x.

Page, P.B., 1997. E.M. Jellinek and the evolution of alcohol studies: a critical essay. Addiction 92 (12), 1619–1637.

Peele, S., July–August 1989. Ain't misbehavin': addiction has become an all-purpose excuse. The Sciences 14–21.

Peele, S., 1998. All wet. Sciences 38 (2), 17.

Pettinati, H., Anton, R., Willenbring, M., 2006. Brief report. The COMBINE study – an overview of the largest pharmacotherapy study to date for treating alcohol dependence. Psychiatry (1550–5952) 3 (10), 36–39.

Roizen, R., Fillmore, K., 2001. Some notes on the new pardigmatic environment of 'natural remission' studies in alcohol research. Substance Use & Misuse 36 (11), 1443–1465. http://dx.doi.org/10.1081/JA-100106959.

Russell, C., Davies, J., Hunter, S., 2011. Predictors of addiction treatment providers' beliefs in the disease and choice models of addiction (English). Journal of Substance Abuse Treatment 40 (2), 150–164.

Russell, M., Peirce, R.S., Chan, A.K., Wieczorek, W.F., Moscato, B.S., Nochajski, T.H., 2001. Natural recovery in a community-based sample of alcoholics: study design and descriptive data. Substance Use & Misuse 36 (11), 1417–1441. http://dx.doi.org/10.1081/JA-100106958.

Scholl, M.B., Schmitt, D.M., 2009. Using motivational interviewing to address college client alcohol abuse. Journal of College Counseling 12 (1), 57–70.

Sobell, M.B., Sobell, L.C., 1978. Behavioral Treatment of Alcohol Problems: Individualized Therapy and Controlled Drinking/Mark B. Sobell and Linda C. Sobell. Plenum Press, *c*.1978, New York.

Stein, M.D., Herman, D.S., Anderson, B.J., 2009. A motivational intervention trial to reduce cocaine use. Journal of Substance Abuse Treatment 36 (1), 118–125. http://dx.doi.org/10.1016/j.jsat.2008.05.003.

Streifel, C., Servanty-Seib, H., 2006. Alcoholics anonymous: novel applications of two theories. Alcoholism Treatment Quarterly 24 (3), 71–91. http://dx.doi.org/10.1300/J020v24n03_05.

Tatarsky, A., 2002. In: Tatarsky, Andrew (Ed.), Harm Reduction Psychotherapy: A New Treatment for Drug and Alcohol Problems. Jason Aronson, *c*.2002, Northvale, NJ. With forward [i.e. foreword] by Alan Marlatt.

Walters, S.T., Rotgers, F., 2012. Treating Substance Abuse: Theory and Technique, third ed. Guilford Press, New York.

Whiteside, U., Cronce, J., Pedersen, E., Larimer, M., 2010. Brief motivational feedback for college students and adolescents: a harm reduction approach (English). Journal of Clinical Psychology 66 (2), 150–163.

Willenbring, M., 2008. New research is redefining alcohol disorders. Addiction Professional 6 (5), 12–19.

Willenbring, M.L., 2010. The past and future of research on treatment of alcohol dependence. Alcohol Research & Health 33 (1/2), 55–63.

Alcohol-Related Disorders

Michael Soyka, Ludwig Maximilian University, Munich, Germany; and Privatklinik Meiringen, Meiringen, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Alcohol consumption has been shown to be causally related to numerous medical conditions and is a significant cause for psychiatric and somatic morbidity and premature death worldwide. While the repeatedly reported moderate mortality reduction in individuals with low alcohol consumption with a lower risk coronary heart disease and ischemic stroke compared to abstainers is controversial, increased alcohol consumption is related to increased risk for liver disorder and many other medical conditions, as well as injuries or accidents, mostly in a rather dose–response relation. The central nervous system is especially vulnerable to alcohol's effects. Brain damage, epileptic seizures, wernicke-korsakoff syndrome, and alcohol psychoses are frequent in individuals with heavy alcohol intake. Numerous other severe disorders are associated with increased alcohol intake. A brief overview about pharmacological and neurobiological actions of alcohol and its effect in brain and body are given.

Alcohol affects the function of nearly all organs in the body and has numerous effects on metabolism, cell function, and physiologic processes. Acute effects of alcohol (e.g., alcohol intoxication) must be differentiated from chronic effects or the results of harmful use or even dependence. Alcohol significantly affects brain function with numerous clinical consequences both on the neuropsychiatric and behavioral level such as impaired memory, cognitive decline, blackouts, seizures, or others. Chronic alcohol consumption can lead to severe organ damage most frequently of the liver but numerous other parts of the body may also be affected. A brief update on the metabolism and pharmacology of alcohol and its general effects in the body is given before more distinct effects on cognitive function, mental processes, body function, and health are outlined. Finally, distinct neuropsychiatric disorders and somatic complications of alcohol dependence are described.

Alcohol (ethanol, C_2H_5OH) is a relative simple molecule, which interacts with numerous transmitters and receptors in the body and brain and also changes structure and function of cells and cell membranes, among others. Virtually, every organ is affected by acute or chronic alcohol intake. It is difficult to define definite cutoffs for risky alcohol consumption. The British Medical Association in 1995 considered 20 g of alcohol for women and 30 g for men as the upper limit for nonrisky alcohol use. More recently, the recommended upper limits for alcohol consumption are about 24 g for men and 12 g for women. Acute effects of alcohol, for example, on blood pressure, circulation, or brain function must be differentiated from more chronic ones like liver dysfunction or withdrawal. Since alcohol's effects in the body are so complex, only a brief overview on its basic mechanisms are given before addressing clinically relevant disorders associated with alcohol consumption.

Alcohol – Metabolism and Pharmacology

Alcohol is quite rapidly absorbed after oral ingestion in the stomach. About 95% of alcohol is oxidized in the liver by the enzyme alcohol dehydrogenase (ADH) to acetaldehyde, which in return is rapidly metabolized by the enzyme acetaldehyde dehydrogenase (ALDH) to acetic acid, which is also rapidly converted to carbon dioxide and water. Only 5% of alcohol is excreted unchanged in the urine, sweat, and breath. There is a genetic polymorphism for both enzymes with different isoenzymes. While most (>90%) of the Caucasian population have 'regular' ALDH isoenzymes, other – especially Asian populations have so-called ALDH-deficient isoenzymes (30–50%) with significant acetaldehyde levels in the blood after alcohol intake. In these individuals, alcohol consumption rapidly results in aversive reactions, so-called 'flush reaction.' Alcohol is usually metabolized at a rate of 0.1–0.15 (or 0.2) mg l^{-1} per hour.

Genetics

There is substantial evidence from a number of family, twin, and adoption studies for a genetic transmission of alcoholism. The risk for alcoholism is significantly increased in first-degree relatives of alcoholics. Some adoption studies have shown an up to four times increased risk for alcoholism for sons of alcoholics even if they were raised apart from their biological parents. Although the heritability of alcoholism is a topic of numerous biological and genetic studies on the genetic and molecular/biological level, no vulnerability marker or gene for alcoholism is definitely identified yet. It seems most likely that alcoholism is not transmitted by a single but a number of genes. Alcoholism appears to be a polygenic disorder. Also no definite genetic marker for alcoholism is found yet. Numerous genome-wide scans to identify genes mediating the risk for alcoholism have been initiated in recent years, with conflicting results. The most robust findings come from chromosomes encoding for alcohol-metabolizing enzymes, especially genetic variations of ADH and ALDH. Other candidate genes are those mediating the pharmacogenetic response to alcohol, such as mu opioid or gamma-aminobutyric acid (GABA) receptor genes.

There are marked differences not only in alcohol metabolism but also in tolerance. Experimental and follow-up studies have shown that high-risk individuals (children of alcoholic parents) usually tolerate alcohol much better than other individuals. This in part explains the increased risk for alcoholism.

General Effects of Alcohol

Brain (Central Nervous System Effects)

Different from other psychoactive substances like opioids there is no special alcohol receptor in the brain. A number of neurotransmitters are involved in mediating alcohol's effects including GABA, glutamate, dopamine, opioids, serotonin, and noradrenalin, among others. Alcohol is a psychotropic agent that depresses the central nervous system (CNS) basically via enhancement of GABAergic neurotransmission. GABA is the most important inhibitory neurotransmitter in the brain. Acute alcohol intoxication results in enhancement of inhibitory neurotransmitters (GABA) and antagonization of excitatory neurotransmitters (glutamate, dopamine, etc.) while the neurotransmitter function in alcohol withdrawal is the opposite (increased activity and release of excitatory, inhibition of inhibitory neurotransmitters). Thus alcohol withdrawal results in an increased excitatory state in the brain, possibly leading to seizures or delirium.

The rewarding, psychotropic effects of alcohol are in part mediated by dopamine, opioids, GABA, glutamate, and serotonin. There seems to be a special addiction memory in the brain, which in part involves brain structures, which are of relevance for physiologic reward processes and controlling of food and fluid intake and sexuality. One of the key structures in the brain mediating these reward effects is the dopaminergic mesolimbic system including the nucleus accumbens. Activation of this system leads to positive reinforcement. Alcohol but also other psychotropic drugs are believed to act predominantly by interactions with neurons in these brain areas.

Alcohol also directly acts on neurons in the CNS. It not only alters the properties of lipids in the membrane of neurons but also has direct neurotoxic effects at least in higher concentrations. Chronic alcohol intake may result in cell damage and destruction of neurons in the brain but also in other regions of the body. Other alcohol-related factors such as vitamin deficiencies or malnutrition in general may contribute to the neurotoxic effects. Although alcohol-related cell loss can be found in all brain areas, the forebrain and the cerebellum are mostly affected in chronic alcoholics. To some extent cell losses in the CNS can be visualized *in vivo* by modern neuroradiological techniques such as cranial computer tomography scans or NMR.

Effects on Cognitive Function and Mental Processes

Modest alcohol intake may cause a number of emotional changes such as sadness, anxiety, or irritability that predominantly occur at peak or with decreasing blood alcohol concentration (BAC). Alcohol in higher doses can cause following psychiatric syndromes: (intense) sadness and anxiety, auditory hallucinations, and/or paranoia without clouding of sensorium. These syndromes can be classified as organic brain syndromes or alcohol psychosis. The former is characterized by mental confusion and clouding of sensorium, which can be found during alcohol intoxication usually at a BAC over 1.5 mg l^{-1}, withdrawal, or as a consequence of alcohol-related disorders.

Behavioral Changes

They depend on age, weight, sex, prior experience with alcohol (e.g., the individual's drinking history), and age. Symptoms of alcohol intoxication are described below.

Tolerance

There are marked differences in alcohol tolerance between individuals, partially due to genetic variances in alcohol metabolism. For a number of not fully understood reasons tolerance in men is usually better than in women. Women have less water in their body so alcohol is less diluted and has greater effects in the tissue. The individual alcohol history (heavy or regular vs sporadic consumption), liver function, organic brain syndromes, or other disorders have marked impact on alcohol tolerance, which is usually increased in heavy drinkers and alcohol dependents, except for those late-stage drinkers with severe physical (liver!) or mental impairment. Some studies in high-risk individuals (offspring of alcoholic families) have shown that alcohol tolerance is usually better in individuals with positive family history for alcoholism and also to some extent predictive for later alcoholism.

Physical Dependence

Alcohol Dependence

According to modern psychiatric classification systems such as ICD-10 and DSM-IV, alcohol dependence is defined as a cluster of physical, psychological symptoms, and social consequences of alcohol consumption (Schuckit, 2005). The categorical distinction between abuse (harmful use) and dependence has been given up by the very recent DSM-5 toward a dimensional concept. Alcohol use disorders are defined by a problematic pattern of alcohol use leading to clinically significant impairment in different domains. In DSM-5, 11 symptoms are given. Presence of 2–3 symptoms indicates a mild, 4–5 a moderate, 6 or more severe disorder.

In contrast, patients who meet ICD-10 diagnosis of alcohol dependence must display three of the following six symptoms:

1. a strong desire or compulsion to drink;
2. tolerance;
3. withdrawal;
4. loss of control;
5. progressive neglect of alternative activities; and
6. persistent drinking despite evidence of harm.

Physical Withdrawal

Many but not all alcoholics develop physical dependence and experience physical and psychological withdrawal symptoms after cessation of alcohol consumption. A number of physiological mechanisms are involved in the development of the

syndrome. Basically the development of withdrawal symptoms can be explained by a number of adaptive mechanisms resulting from long-term alcohol intake. While alcohol enhances the neurotransmission of inhibitory neurotransmitters (GABA) and blocks excitatory neurotransmitters (glutamate, etc.) during alcohol withdrawal, there is an increased excitability in the CNS and an autonomic nervous system dysfunction with an excess release and turnover of excitatory neurotransmitters.

Symptomatology of alcohol withdrawal covers a wide range of symptoms, which develop few hours after the last drink with a peak on day two or three, which usually subsides within 4 or 5 days. While alcohol withdrawal is usually mild in some cases a severe withdrawal syndrome can develop. Key symptoms are tremor, insomnia, malaise, anxiety, inner restlessness, sweating, increase in heart and respiratory rate, mild elevations in temperature, gastrointestinal symptoms such as anorexia, nausea, and vomiting, and psychological or emotional symptoms such as anxiety or sadness. A broad number of other symptoms may also be prevalent, depending on the patient's physical condition. In more severe cases, seizures (5–10% or more of patients) or hallucinations may complicate the clinical course. The most severe variant of alcohol withdrawal is alcohol withdrawal delirium.

Depending on the clinical course and symptomatology, inpatient or outpatient detoxification can be necessary. Pharmacological treatment includes fluid intake, substitution of vitamins and minerals, and sedatives, predominantly benzodiazepines or clomethiazole (in Europe only).

Effects on the Body and Health

Alcohol in light to moderate doses may have a slight beneficial effect in decreasing the risk for cardiovascular disease by increasing high-density lipoproteins (HDLs) although this issue is still controversial. In any case, this effect is far outweighed by the health risks in individuals with heavy alcohol consumption. Mean effects of alcohol in the body are as follows:

1. Cardiovascular and cerebrovascular system: hypertension, heart inflammation, or more often myocardiopathy, arrythymia.
2. Brain: intracerebral hemorrhage. Other data indicate that mild to moderate alcohol consumption ($<50\ \mathrm{g\ day^{-1}}$) may have some protective effect on the cerebrovascular and cardiovascular system, possibly by effects on blood lipids (inhibition of elevated low-density lipoprotein (LDL) cholesterol, increase of HDLs) and some antiatherogenic and antithrombotic effects.
3. Neuromuscular system: polyneuropathy, myopathy, autonomic disorders.
4. Digestive system: increased rate of gastritis and ulcer disease, pancreatitis (possibly followed by diabetes), abnormal functioning of the esophagus including esophagitis.
5. Liver: fatty liver, acute alcoholic hepatitis, chronic active hepatitis, and, finally, cirrhosis. From a chronic alcohol intake of 20 g for women and 30–40 g for men the risk for liver damage is already increasing.
6. Blood cells: the production of all types of blood cells is decreased. Red blood cell anemia (macrocytosis), decrease of white-cell production and function, and decreased production of platelets and clotting factors are the result. Function of thymus-derived lymphocytes, which are essential for immune function, is also impaired.
7. Sexual functioning and hormonal changes in men testicular atrophy, hypogonadism, decreased sperm production and motility, decreased testosterone production, and sometimes impotence are typical results of chronic alcoholism. In women, menstrual irregularities are of relevance, as are effects on the fetus.
8. Other endocrine and metabolic effects of alcohol include impaired thyroid and parathyroid function with an increased risk for osteoporosis and bone fractures. The glucose and carbohydrate metabolism is affected in complex ways. A frequent complication of alcohol is also diabetes, mostly due to pancreatitis.
9. Skin: a number of dermatological conditions can be provoked or worsened by alcohol: porphyries, psoriasis vulgaris, rosacea, cancer of the oral mucosa, pellagra, and others.
10. Increased risk for cancer of the mouth and digestive tract (pharynx, larynx, esophagus, stomach, liver), head and neck, lungs, and breast. A number of variables contribute to this phenomenon: alcohol toxicity, comorbid nicotine dependence (smoking), malnutrition, decrease in immune function, among many others.
11. Malnutrition, vitamin deficiency, and electrolyte changes. Typically chronic alcoholism is associated with some form of malnutrition. Typical effects of chronic alcoholism are zinc deficiency, hypokalemia, deficiencies of vitamin B ($B_{1,3,12}$), vitamin C, and folic acid, among many others.

Morbidity and Mortality

The mortality and morbidity of individuals with heavy alcohol consumption, harmful use, or alcohol dependence are significantly increased compared to the general population. Reasons are multifactorial such as there are numerous somatic and neurologic disorders related to alcohol, also the risk for accidents and suicide is much higher in alcoholics. In addition, the alcoholic's lifestyle (nicotine consumption, low-protein/high caloric food, etc.) also contributes to the increased morbidity and reduced life expectancy in alcoholism. Some studies indicate that 5% of all deaths are related to alcohol consumption.

Alcohol Embryopathy (Fetal Alcohol Syndrome)

A tragic and often underestimated result of drinking during pregnancy is alcohol embryopathy. Key features are intrauterine growth retardation, microcephaly, moderate to severe mental retardation, and relatively typical dysmorphologic facial malformation. Many other symptoms may also be present such as internal and genitourinary malformations, especially congenital heart defects, among many others. The degree of alcohol embryopathy is correlated to the stage of maternal alcohol illness, not to the maternal alcohol consumption.

Psychiatric Complications

A number of distinct neuropsychiatric disorders are caused by chronic alcohol intake including delirium, psychosis, anxiety, and depression and an increased risk for suicide/delinquency. There is a high comorbidity of alcoholism with depression, schizophrenia, antisocial personality, and other personality disorders, anxiety, and addiction to other substances and drugs including tobacco. Alcoholism sometimes develops prior to the psychiatric disorder but in many cases it is secondary, worsening the clinical course and symptomatology.

Driving Ability and Accidents

Alcohol has major effects on driving ability as well as the risk for accidents. These factors contribute significantly to the increased morbidity and mortality in alcoholics. Even at a BAC of 0.15 mg l^{-1} the ability to operate a motor vehicle is significantly impaired. The risk, especially for more severe accidents, rises dramatically with increasing BAC. There are marked differences concerning permitted BAC while operating a motor vehicle in different countries. Some countries do not allow any alcohol intake, while in most states 0.5 or 0.8 mg l^{-1} BAC are the upper limits tolerated. The risk for other accidents (home, workplace, sports) also rises with increasing BAC.

Nervous System

Alcohol Intoxication

Probably, the most frequent alcohol-related disorder is alcohol intoxication. Although there is no strict correlation between blood alcohol level and behavioral or motor impairment the symptomatology depends on BAC, individual alcohol tolerance, and a number of confounding factors such as physical and psychiatric status, intake of other substances, sleep deprivation, among others. Alcohol intoxications can be classified as mild (BAC 0–1 mg l^{-1}), moderate (BAC 1–2 mg l^{-1}), and severe (BAC over 2–2.5 mg l^{-1}). The higher the BAC, the more pronounced are the CNS depressant, sedative effects, and the behavioral dysfunction. Light to moderate alcohol intoxication is usually associated with relaxation and feelings of euphoria as well as impaired coordination. Higher BAC results in severe cognitive, perceptual, and behavioral impairment. This includes blackouts, insomnia, hangover. On the physical level alcohol intoxication is associated with hypertension, cardiac arrhythmia, gastrointestinal symptoms such as vomiting, diarrhea, abdominal pain, nausea, anorexia, gastritis, and hepatitis. Neurological symptoms include ataxia, fainting, and blackouts. Trauma and accidents (traffic safety!) are of special relevance. On the psychological and behavioral level insomnia, anxiety, depression, sexual problems, and inappropriate, aggressive or impulsive behavior may occur. Severe cognitive impairment, clouding of sensorium, disorientation, and amnesia can be found in higher BAC. Mortality is significant in BAC >4 mg l^{-1} due to respiratory paralysis, heart failure, and coma.

Pathological Intoxication (Alcohol Idiosyncratic Reaction)

In some individuals (although extremely rare), mild to moderate alcohol intoxications can be associated with severe aggression and violence or psychotic reactions lasting for a few hours and usually followed by a more or less complete amnesia. The psychiatric and behavioral symptoms are very marked and cannot be explained by the comparatively low BAC level. This syndrome, which is basically of forensic interest, is very controversial among clinicians. Possible predisposing factors are hypoglycemia or other metabolic disorders, organic brain syndromes, or intake of other psychotropic drugs such as stimulants.

Delirium

Delirium usually starts during the first 4–7 days after cessation of alcohol consumption. Key features of delirium are clouding of sensorium, disorientation, severe confusion, fear and agitation, visual and sometimes acoustic hallucinations, delusions of persecution, or others. Delirium is a very serious medical disorder, which is more common than other alcohol psychosis (prevalence rate about 1%) and has a significant mortality if untreated. Symptoms found in alcohol withdrawal can also be seen in alcohol delirium but are usually more severe. The clinical condition is characterized by a severe overactivity of the autonomic nervous system (increased pulse rate and respiratory rate, marked elevation in blood pressure and body temperature). Frequent complications are seizures, cardiac arrhythmia, and many other medical disorders. Patients need substantial medical support and psychopharmacological treatment, usually sedatives such as benzodiazepines.

Alcohol Psychosis

Chronic alcohol consumption can result in different alcohol psychoses. In some cases, a more or less chronic state with suspiciousness or more pronounced paranoid delusions can develop. This disorder is referred to as alcoholic paranoia or alcohol-induced psychotic disorder. The prototype of this psychosis is a delusional jealousy syndrome nearly exclusively found in male alcoholics who believe their spouse to have an extramarital relationship. Sometimes without the slightest evidence the alcoholic is convinced about his spouse's infidelity. Predisposing factors for the development of this syndrome are impotence or other sexual dysfunction, cognitive impairment, and a low self-esteem. The delusions often persist into abstinence. Delusional jealousy is a dangerous disorder with the patient often attacking or even killing his spouse.

The other more prevalent alcohol-induced psychosis is alcohol hallucinosis, which is characterized by vivid predominantly acoustic, sometimes visual hallucinations, delusions of reference or persecution, and fear. Other psychotic symptoms may also be prevalent. Different from alcohol withdrawal delirium the sensorium is usually clear and there is no amnesic syndrome for the psychosis. The psychopathology of alcohol hallucinosis closely resembles paranoid schizophrenia but there is no evidence for a common genetic basis. Alcohol hallucinosis, like alcohol paranoia, can develop during heavy

drinking or more frequently within a few days or weeks of the cessation of drinking. In abstinent patients, the prognosis of alcohol hallucinosis is usually good, but in 10–20% a chronic, schizophrenia-like psychosis can develop. Psychopharmacological treatment in alcohol psychosis (neuroleptics, sedatives) is recommended.

Organic Brain Syndrome, Encephalopathy, and Dementia

While some form of cognitive impairment can be found in up to 75% of chronic alcoholic patients approximately 9% of them have clinically manifest organic brain syndrome. Alcohol itself, but also alcohol-related disorders such as malnutrition including vitamin deficiencies as well as indirect consequences of alcoholism, such as head trauma, hypoglycemia, or other metabolic disturbances can cause cognitive dysfunction, mental confusion, and clouding of sensorium. Serious confusion can be seen during alcohol intoxication and withdrawal, as a result of vitamin deficiency (e.g., thiamin), head trauma, extra- or intracranial hematoma, stroke, hypoglycemia, or simply as a result of long-term alcohol intake, or a combination of these factors.

Wernicke encephalopathy, a dramatic, very acute neurologic syndrome with high mortality, is characterized by a classical symptom trias: ataxia, ophthalmoplegia, and mental disorder (clouding of consciousness). Thiamin deficiency is essential for the development of the syndrome. Patients are disoriented or confused, somnolent or even in coma, show oculomotor abnormalities and gait ataxia. There are distinct symmetric punctuate hemorrhagic lesions in certain brain areas. Rapid thiamin substitution is essential for therapy.

Wernicke encephalopathy in many cases is followed by Korsakoff syndrome (alcohol-related amnesic syndrome), which may also develop without prior Wernicke symptomatology. Key features are anterograde and retrograde amnesia, memory loss, and other cognitive impairment. Apathy, passivity, and confabulations are common symptoms. Prognosis is poor.

Other patients show symptoms of a more gradual cognitive decline and other dementia symptoms without distinct neurological symptoms. Alcohol dementia is a difficult diagnosis. A broad number of other dementia forms including Alzheimer's disease have to be excluded before diagnosis can be made. Chronic hepatic encephalopathy also goes along with cognitive impairment but other neurological symptoms can also be found as follows: frontal release signs, hyperreflexia, pyramidal signs, or others. Organic brain syndromes can also be found as a result of other alcohol-related disorders.

Seizures

Epileptic seizures are the most frequent neurological sequelae with prevalence estimates of 15% or more. The exact pathophysiological basis is unclear. Electrolyte imbalances and neurotransmitter dysfunction (GABA, glutamate) are of special relevance. This disorder is independent from duration of alcoholism and there is no evidence for a genetic risk for seizures in these patients.

Seizures usually occur within the first 24 to a maximum 48 h of abstinence and nearly exclusively are of tonic–clonic (grand mal) type. The clinical and neurological status is usually normal. Other seizure types, especially focal seizures indicate a probable focal brain injury (trauma, hemorrhage, etc). Electroencephalogram and cranial computer tomography may help excluding other reasons than alcohol for seizures but are otherwise usually normal. Prognosis for seizures in abstinent alcoholics is good, otherwise the risk for recurrent seizures is high.

Polyneuropathy

This is a frequent complication of alcoholism (prevalence 9–30%). Beside diabetes, alcohol is the most common cause for polyneuropathy. A number of peripheral nerves with sensory, motoric, or autonomic fibers are affected. The sensoric input and in more severe cases, the motoric system and muscle function are impaired. Typical complaints are symmetric burning or stabbing pain in the feet and mild to more severe weakness of the limbs. Polyneuropathy usually gradually develops and the prognosis in abstinent patients is often positive.

Myopathy

Alcohol has myotoxic effects both on skeletal and cardiac muscles. The more dramatic acute myopathy, which can be accompanied by sometimes extended muscle necrosis, hypokalemia, and secondary renal failure has a prevalence of 0.8–3.3%. Chronic myopathy often with subclinical symptomatology is much more common (23–66%). Myopathy can also be secondary to polyneuropathy. While the acute form goes along with painful muscle swelling, tenderness, and muscle cramps, in the chronic form extended weakness in the muscles is reported. A rare subtype is a myopathy related to hypokalemia.

Autonomic Disorders

Alcohol can also affect autonomic nerves and cause various autonomic dysfunctions (both parasympathicus and sympathicus) such as dysphagia, esophageal dysfunction, abnormal pupillary reflexes, impotence, impaired thermoregulation, among many others. Autonomic disorders are seldom isolated but usually accompanied by other alcohol-related disorders.

Cerebellar Atrophy

Up to 30% of chronic alcoholics show some clinical or neuroradiological symptoms of cerebellar atrophy. Histologically, a degeneration of Purkinje cells in the anterior and superior vermis is seen as well as in the cerebellar cortex. The disorder does not correlate with lifetime consumption of alcohol. Other factors such as vitamin deficiency seem to be of relevance. Cerebellar atrophy develops slowly. Key symptoms are dysarthria, gait and stand ataxia, tremor, and nystagmus. Lower limbs show more impairment than upper limbs. Severe forms of cerebellar atrophy cause astasia and abasia. Symptoms are often at least partially reversible in cases of abstinence and vitamin substitution.

Cerebral Vascular Diseases

There is an increased risk for intracerebral and subarachnoidal hemorrhages in chronic alcoholism with severe neuropsychiatric symptomatology depending on location and size. The association with ischemic stroke is less clear. A more frequent complication is chronic subdural hematoma, often preceded by some sometimes minor head trauma. Symptoms can initially be very mild or even missing. Headache is the most frequent symptom.

Central Pontine and Extrapontine Myelinolysis

A rare complication of alcoholism. A very rapid substitution of hyponatremia, a common electrolyte imbalance in alcoholism, seems to be of special relevance for the development of the demyelination in the pons or some other areas of the brain. Following clinical symptoms are severe with a high mortality: tetraparesis, cerebellar ataxia, bulbary symptoms, paresis of eye muscles, and central fever. The extreme form is a locked-in-syndrome with complete tetraplegia.

Marchiafava–Bignami Syndrome (Corpus Callosum Atrophy)

Another extremely rare disorder with poor prognosis and uncertain pathophysiology. In some chronic alcoholics, especially red wine drinkers in the Mediterranean, a necrosis of the corpus callosum and sclerosis of the cerebral cortex can lead to confusion, clouding of sensorium, seizures, and other neurological symptoms, coma, and death. If the patient survives, dementia is the most frequent outcome.

Tobacco–Alcohol Amblyopy

A bilateral affection (demyelination) of the optic nerve, chiasma opticum, and tractus opticus can lead to blurred or loss of vision. This rare syndrome can predominantly be found in heavy-smoking alcoholics with malnutrition. Tobacco smoke contains cyanides, which cannot sufficiently be detoxified in patients with severe liver dysfunction. They are believed to affect the optic nerve by free cyanides. Prognosis is rather poor.

Alcohol-Related Myelopathy

An extremely rare disorder with good prognosis. Alcohol myelotoxicity, malnutrition, and chronic liver damage can cause a progressive myelopathy with spastic paraparesis, neurogenic bladder dysfunction, and paresthesia.

Movement Disorders

Occasionally extrapyramidal symptoms, similar to Parkinson's disease, or dyskinesias can be seen in chronic alcoholics. The prognosis in abstinent patients is usually good. The more frequent essential tremor can be suppressed by small amounts of alcohol. This syndrome is not a result of chronic alcoholism.

Sleep Disorders

Alcohol consumption has major impact on the sleep architecture. Acute intake can lead to decreased latency to sleep onset, increased slow wave sleep, and decreased rapid eye movement sleep during the first half of the night. Insomnia is frequent during alcohol withdrawal and can persist long into abstinence. Other sleep disorders, e.g., sleep apnea syndrome, is usually worsened by alcohol.

See also: Alcohol Use among Young People; Drinking, Anthropology of.

Bibliography

Boyle, P., Bofetta, P., Lowenfels, A.B., Burns, H., Brawley, O., Zatonski, W., Rehm, J. (Eds.), 2013. Alcohol. Science, Policy, and Public Health. Oxford University Press.

Grant, B.F., Stinson, F.S., Dawson, D.A., Chou, S.P., Ruan, W.J., Pickering, R.P., 2004. Co-occurrence of 12-month alcohol and drug use disorders in the United States. Results from the National Epidemiological Survey on Alcohol and Related Conditions. Archives of General Psychiatry 61, 361–368.

Kessler, R.C., Berglund, P., Demler, O., Jin, R., Merikangas, K.R., Walters, E.E., 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry 62, 593–602.

Noronha, A.B.C., Cui, C., Harris, R.A., Crabbe, J.C. (Eds.), 2014. Neurobiology of Alcohol Dependence. Academic Press, London.

Soyka, M., Kranzler, H.R., Berglund, M., Gorelick, D., Hesselbrock, V., Johnson, B.A., Möller, H.J., 2008. World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for biological treatment of substance use and related disorders, part 1: alcoholism. World Journal of Biological Psychiatry 9, 6–23.

Soyka, M., 2013. Update Alkoholabhängigkeit. Unimed Verlag, Bremen.

Spanagel, R., 2009. Alcoholism: a systems approach from the molecular physiology to addictive behavior. Physiological Reviews 89, 649–705.

Spanagel, R., Vengeliene, V., 2013. New pharmacological treatment strategies for relapse prevention. Current Topics in Behavioral Neurosciences 13, 583–609.

Schuckit, M.A., 2005. Drug and Alcohol Abuse, sixth ed. Plenum, New York. A Clinical Guide to Diagnosis and Treatment.

Whiteford, H.A., Degenhardt, L., Baxter, A.J., et al., 2013. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. Lancet 382, 1575–1586.

Alcohol Use among Young People

Rainer K Silbereisen, Center for Applied Developmental Science, University of Jena, Jena, Germany
Karina Weichold, University of Jena, Jena, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R.K. Silbereisen, volume 1, pp. 122–125, © 2001, Elsevier Ltd.

Abstract

Alcohol use in normative samples becomes more prevalent and frequent across adolescence and early adulthood. This progression points to the facilitating role of consumption in resolving developmental tasks. In contrast, abuse is a relatively rare behavior and often has its roots in adaptation problems in childhood. Given the overall moderate or time-limited use among the majority, most consequences are immediate and circumscribed. Universal prevention efforts attempt to reduce the harm, or attempt to increase skills. In the cases where alcohol use is part of an enduring pattern of maladaptation, it is necessary to implement earlier selective intervention.

Alcohol use is prevalent among people beyond childhood and shows an intriguing association with age. Consumption increases rapidly across adolescence, shows a peak in the early twenties and declines gradually thereafter, once the major developmental tasks of emerging adulthood are resolved. Whereas young children disapprove of drinking, from adolescence on alcohol consumption is most often seen as signifying one's growing social maturity (i.e., in cultures with no absolute prohibition of alcohol use, e.g., for religious reasons).

The developmental-psychological perspective chosen departs from these observations and explains the emergence of alcohol use among the majority of young people as embedded in the normative psychosocial challenges of adolescence (Silbereisen and Eyferth, 1986). This period of the life span is characterized by growing attempts to find a particular place in life, which involves dealing with new social expectations and personal aspirations. The increasing interest at this time in novel and risky activities, and the unsupervised environments associated with them, probably also has neurobiological underpinnings related to the increase in dopamine input to the prefrontal cortex and limbic brain regions during early adolescence (Spear, 2013). Taken together, both viewpoints justify treating alcohol use among young people as a separate issue, distinct from alcohol use in general.

Abuse of alcohol is a relatively rare form of use, characterized by: (1) consumption over extended periods of time in situations which require clarity of perception and judgment; (2) drinking of even small amounts if decision making is impaired, due to developmental immaturity; (3) increasing the level of alcohol in order to compensate for declining psychoactive effects or to avoid malfunctioning; and (4) all forms of consumption which impair health or adequate mastery of normative exchanges with the environment (Newcomb and Bentler, 1989). Only a small subset of young people meets the clinical criteria for severe substance use disorders (e.g., APA, 2013; *see* Sexual Risk Behaviors).

Consumption Prevalence and Trends across Age

According to representative school surveys, such as the Monitoring the Future study in the USA (Johnston et al., 2013), the lifetime prevalence of alcohol use among 12th graders is of the order of 70% (in contrast, episodic heavy drinking (five drinks or more in a row) amounts to 24%). Even higher prevalence rates can be observed in many European countries (Hibell et al., 2012). Concerning frequency, two-thirds of 12- to 17-year-olds in a large German representative sample (BZgA, 2012) reported drinking less than once a month, about one-fourth up to several times a month, and the remaining 15% reported consuming alcohol on a regular basis, i.e., at least once a week (lifetime prevalence of alcohol use 73%). With regard to the quantity, these adolescent consume about 31 g pure alcohol per drinking occasion through alcoholic beverages. About 2% of the 12- to 15-year-olds and 13% of the 16- to 17-year-olds show a risky drinking pattern (more than 12 g or 24 g pure alcohol consumed on average at daily basis). These figures are in line with those that have been reported by the European School Survey Project on Alcohol and Other Drugs (ESPAD; Hibell et al., 2012) that regularly collects comparable data on substance use among 15- to 16-year-old students in nearly 40 European countries. For this age group, this study finds a prevalence of alcohol use for the past 30 days of 57% (87% lifetime prevalence) and an average of 45 g pure alcohol consumed on the last drinking day. In general, gender differences in consumption prevalence among the young are small but males are more likely than females to consume alcohol regularly and more often in higher quantity.

Beginning with the teen years entailing new freedoms and challenges, frequency and amount of consumption increase rapidly. According to a metaanalysis of more than 20 longitudinal studies (Fillmore et al., 1991), and more recent longitudinal studies (Chen and Jacobson, 2012; Paavola et al., 2004), the increase in frequency and quantity peaks in the early twenties, followed by a similarly sharp decline, particularly in frequency, which seems to be triggered by a general age-related trend toward conventionality (Jessor et al., 1991) and growing incompatibilities between consumption and new responsibilities as partner, parent, and worker.

Whereas countries like the US, Canada, and the UK share relatively moderate consumption, some Mediterranean and Eastern European countries rank much higher. In a longer perspective, consumption in industrialized countries increased dramatically after World War II, reaching unprecedented peaks

in the 1970s and 1980s, followed by stable or slightly declining figures thereafter (Silbereisen et al., 1995). During the past 20 years, prevalence rates of alcohol consumption decreased by about 20% in the US-American and European adolescents (Johnston et al., 2013; Hibell et al., 2012).

Immediate Negative Consequences for Well-Being

Due to the overall moderate and/or time-limited alcohol consumption among adolescents, most of the consequences for well-being are immediate. In general, a substantial minority sometimes experience discomfort, including feeling dizzy, hangovers, and headaches, primarily as a result of intoxication. Due to cultural differences in drinking habits (e.g., regarding a general acceptance of a 'culture of intoxication,' reflected, for instance, in speed drinking with peers in pubs), this experience is more common in Nordic countries in Europe (i.e., higher drunkenness rates, low risk perception of heavy use), in spite of higher consumption figures in the South.

According to data from the UK (Miller and Plant, 1996), between 5 and 30% of young people in midadolescence report problems associated with alcohol use in areas of social functioning such as personal adversities (reduced performance in school), social relationships (tensions with friends), sexuality (unwanted sexual encounters), and delinquency (trouble with police). The alcohol-related problems most often experienced by adolescent binge drinkers were accidents or injuries, physical fights, problems with friends or parents, and risky sexual behavior. More serious consequences (such as liver cirrhosis), in contrast, are very rare. However, especially early and/or high and frequent consumption can lead to physical or mental health problems (e.g., Boys et al., 2003), and about 44% of male and 17% of female adolescent regular drinkers met DSM criteria of alcohol abuse until age 34; 21% of men and 7% of women were classified as alcohol dependent (Wittchen et al., 2008). Alcohol is not a gateway drug, but it is certainly true that most users and abusers of other psychoactive substances begin with (and often maintain) the use of alcohol; however, they are also characterized by additional risks, such as family problems or deviant peer contacts (Kandel, 2002).

Risky sexual behavior and alcohol use correlate. This is probably not so much due to being uninhibited under the influence of alcohol but is rather rooted in common situational encounters, often concentrated in small subgroups, which may also share other risk factors such as mental disorders (APA, 2013; *see* Sexual Risk Behaviors).

As far as legal consequences are concerned, in spite of public concerns about the easy accessibility to alcohol for minors, there is little attempt at prosecution. In some countries the legal age for driving is considerably lower than that for alcohol drinking and purchase, which may exacerbate the problem of young people's reckless driving under the influence of alcohol. Recent empirical findings demonstrate that the rate of motor vehicle crashes among young drivers in the US had increased during the past decades. Risk of motor vehicle crashes among 16- and 17-year-olds was increased by driving at night, without adult supervision, together with passengers, being of male gender, and using alcohol (Rice et al., 2003). However, recently, alcohol use has lost its role as the main risk factor for motor crashes among teens, but not for drivers over the age of 20.

Differential Perspectives and Roles in Normative Psychosocial Development

Following an approach put forward by Moffitt (1993), a deeper understanding of the age trends, associations with biographical transitions, and immediate consequences of alcohol consumption can be achieved by distinguishing two sets of developmental antecedents and motives (APA, 2013). With regard to the adolescence-limited trajectory, which is characteristic of the absolute majority, alcohol use emerges because almost all adolescents must wait for the status and privileges of adults, despite their physical maturity, for several years (due to ever-expanding schooling, this gap is growing historically). Once they have resolved these issues, the frequency and intensity of problem behaviors, including alcohol, will vanish due to the influence of new environments that entail fewer opportunities and provide more deterrents concerning use. The life-course-persistent trajectory, in contrast, maintains consumption beyond the normative transitions to adulthood and is rooted in long-lasting problems of adaptation, starting in early childhood and encompassing neurological problems, attention deficit, impulsivity, and the like.

Moffitt's (1993) model is supported by various empirical findings. However, the risk factors for normative vs problematic trajectories of substance use across adolescence may vary by gender (e.g., Weichold et al., 2014). In addition, the model proposed by Moffitt matches well with more elaborate distinctions in the literature on alcohol and alcoholism where one of the subtypes is described as genetically influenced, with early behavioral malfunctioning, and embedded in a long-lasting antisocial personality disorder (e.g., Fitzgerald and Zucker, 2006; Tarter et al., 1999). Moreover, it also enables the remarkable covariation among alcohol use and other drugs, particularly externalizing problem behaviors to be understood. These problem behaviors, such as reckless driving or unprotected sexual activities, signify status for the young but are deemed inadequate by the community due to their precocity. Our general notion that it is the maturity gap which channels the alcohol use of the vast majority of adolescents sounds rather negative. Note, however, that most adolescents perceive alcohol as a means to ease social contacts and improve feelings in such contexts. Only a small minority drinks alcohol with the purpose of mood regulation when facing problems or coping with stress and boredom, whereas many adults do so (BZgA, 2012).

Alcohol is used by teens with regard to the formation of peer and romantic friendships, which are major developmental tasks in the second decade of life. Moderate consumption among those on the adolescent-limited trajectory corresponds prospectively to higher status and better cohesion within one's peer group, and is associated with a higher likelihood of romantic involvement. Moreover, adolescents seem to select leisure settings that offer opportunities for friendship contacts and provide alcohol in the right quantity and environment, such as discotheques, quite deliberately (Silbereisen et al.,

1992). In a nutshell, alcohol consumption may have constructive functions in healthy psychosocial development; for instance, by being a catalyst for the solution of developmental tasks in the majority of adolescents. For a minority, however, it may foreshadow a part of longer-termed maladaptation and problems in coping with challenges during adolescence (see also Schulenberg et al., 2000).

Universal and Selective Prevention Attempts

Given the almost normative use of alcohol among young people in many cultures, efforts aimed at prevention typically target responsible, self-controlled, and health-conscious use rather than abstinence. The multifunctional role of alcohol in the resolution of developmental tasks during adolescence and emerging adulthood represents the major pivot for primary prevention. Appropriate measures need to be undertaken early enough, that is, in late childhood/early adolescence, parallel to the first attempts to actually utilize the possible roles of alcohol consumption in the negotiation of the adolescent challenge.

Concerning measures on the proximal environmental level, one needs to reduce contexts, which entail schedules known to provoke the habituation of drinking, such as the episodic availability of large quantities in seducing locales (like 'binge' drinking in fraternity settings). Efforts most remote to the individual concern attempts to reduce national levels of per capita consumption in general, but curbing heavy drinking seems to affect consumption among the adult population, not the young. More specific measures try to minimize the harm by enforcing controls on the drinking settings, such as the establishment of licensing hours or by training bartenders to refuse serving alcohol to drivers in the UK.

The family is the proximal environment for most adolescents, which represents a major source of risk factors for drinking, such as parental modeling, inconsistency in rule setting, and a lack of developmental challenge. However, very few attempts at prevention on the family level exist to date (e.g., Mason and Spoth, 2012).

Concerning prevention at the individual level, targeting adolescents at school is the rule. Given the role of alcohol use in response to normative developmental difficulties, prominent universal programs targeting entire cohorts address general life skills, such as adequate self-perception, empathy with others, critical thinking, decision making, communication, sociability, affect regulation, and coping with stress (Botvin, 1996). Such life skills are supposed to facilitate the solution of everyday challenges related to the completion of developmental tasks (WHO, 1997).

In addition, as revealed by recent metaanalyses of evaluation studies (e.g., Hansen et al., 2010), the most successful programs are characterized by a combination of skills and attitude development, promotion of bonding to norm-transmitting contexts, and the promotion of resistance skills, positive within-classroom interactions, and positive peer relationships within a context of high social support. In addition, within successful programs, adolescents are actively involved, and they practice skills; for instance, in role-plays representing age-typical challenges (e.g., how to resist unwanted offerings of alcohol by peers). These characteristics are best reflected by the life skills approach. Empirical studies using such programs demonstrated short-, and, at times, longer-term positive effects (e.g., Weichold, in press). Thereby, it is important not to expect sustainable effects unless intervention takes place repeatedly at major milestones during adolescence and beyond.

Concerning the life-course-persistent trajectory of alcohol use, prevention as described would begin too late and is sometimes inefficient (and may even hurt those on the adolescence-limited trajectory due to heightened contacts with negative role models). Rather, prevention would need to start at a much earlier age, and would need to address directly the associated early childhood problems such as impulsivity.

See also: Addictions: General Considerations; Adolescent Health and Health Behaviors; Health Behaviors; Health Education and Health Promotion; Health Promotion in Schools; Sexual Risk Behaviors.

Bibliography

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). American Psychiatric Publishing, Arlington.

Botvin, G., 1996. Substance abuse prevention through life skills training. In: DeV Peters, R., McMahon, J. (Eds.), Preventing Childhood Disorders. Substance Abuse and Delinquency. Sage, Newbury Park, CA, pp. 215–240.

Boys, A., Farrell, M., Taylor, C., Marsden, J., Goodman, R., Brugha, T., Bebbington, P., Jenkins, R., Meltzer, H., 2003. Psychiatric morbidity and substance use in young people aged 13-15 years: results from the child and adolescent survey of mental health. British Journal of Psychiatry 182, 509–517.

Bundeszentrale für gesundheitliche Aufklärung (BZgA), 2012. Die Drogenaffinität Jugendlicher in der Bundesrepublik Deutschland 2011. Teilband Alkohol. Bundeszentrale für gesundheitliche Aufklärung, Köln.

Chen, P., Jacobson, K.C., 2012. Developmental trajectories of substance use from early adolescence to young adulthood: gender and racial/ethnic differences. Journal of Adolescent Health 50 (2), 154–163.

Fillmore, K.M., Hartka, E., Johnstone, B.M., Leino, V., Motoyoshi, M., Temple, M.T., 1991. The collaborative alcohol longitudinal project. A meta-analysis of life course variation on drinking. British Journal of Addiction 86, 1221–1268.

Fitzgerald, H.E., Zucker, R.A., 2006. Pathways of risk aggregation for alcohol use disorders. In: Freeark, K., Davidson, W.S. (Eds.), The Crisis in Youth Mental Health: Critical Issues and Effective Programs, Issues for Families, Schools, and Communities, vol. 3. Praeger Publishers/Greenwood Publishing Group, Westport, CT, pp. 249–271.

Hansen, W.B., Derzon, J., Dusenbury, L., Bishop, D., Campbell, K., Alford, A., 2010. Operating characteristics of prevention programs: connections to drug use etiology. In: Scheier, L.M. (Ed.), Handbook of Drug Use Etiology. American Psychological Association, Washington, DC, pp. 597–616.

Hibell, B., Guttormsson, U., Ahlstrom, S., Balakireva, O., Bjarnason, T., Kokkevi, A., Kraus, L., 2012. The 2011 ESPAD Report: Substance Use among Students in 36 European Countries. The Swedish Council for Information on Alcohol and Other Drugs (CAN). EMCDDA; Council of Europe, Stockholm.

Jessor, R., Donovan, J., Costa, F. (Eds.), 1991. Beyond Adolescence. Problem Behavior and Young Adult Development. Cambridge University Press, Cambridge, UK.

Johnston, L.D., O'Malley, P.M., Bachman, J.G., Schulenberg, J.E., 2013. Monitoring the Future National Survey Results on Drug Use, 1975–2012. In: Secondary School Students, vol. I. Institute for Social Research, The University of Michigan, Ann Arbor.

Kandel, D.B. (Ed.), 2002. Stages and Pathways of Drug Involvement: Examining the Gateway Hypothesis. Cambridge University Press, New York.

Mason, W.A., Spoth, R.L., 2012. Sequence of alcohol involvement from early onset to young adult alcohol abuse: differential predictors and moderation by family-focused preventive intervention. Addiction 107 (12), 2137–2148.

Miller, P., Plant, M.A., 1996. Drinking, smoking and illicit drug use among 15 and 16 year olds in the United Kingdom. British Medical Journal 313, 394–397.

Moffitt, T., 1993. Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy. Psychological Review 100, 674–701.

Newcomb, M., Bentler, P., 1989. Substance use and abuse among children and teenagers. American Psychologist 44, 242–248.

Paavola, M., Vartiainen, E., Haukkala, A., 2004. Smoking, alcohol use, and physical activity: a 13-year longitudinal study ranging from adolescence to adulthood. Journal of Adolescent Health 35 (3), 238–244.

Rice, T.M., Peek-Asa, C., Kraus, J.F., 2003. Nighttime driving, passenger transport, and injury crash rates of young drivers. Injury Prevention 9 (3), 245–250.

Schulenberg, J., Maggs, J., Steinman, K., Zucker, R., 2000. Developmental matters: taking the long view on substance abuse etiology and intervention during adolescence. In: Monti, P., Colby, S., O'Leary, T. (Eds.), Adolescents, Alcohol, and Substance Abuse: Reaching Teens through Brief Interventions. Guilford Press, New York, pp. 19–57.

Silbereisen, R.K., Eyferth, K., 1986. Development as action in context. In: Silbereisen, R.K., Eyferth, K., Rudinger, G. (Eds.), Development as Action in Context: Problem Behavior and Normal Youth Development. Springer, New York, pp. 3–16.

Silbereisen, R.K., Noack, P., von Eye, A., 1992. Adolescents' development of romantic friendship and change in favorite leisure contexts. Journal of Adolescent Research 7, 80–93.

Silbereisen, R.K., Robins, L., Rutter, M., 1995. Secular trends in substance use: concepts and data on the impact of social change on alcohol and drug abuse. In: Rutter, M., Smith, D. (Eds.), Psychosocial Disorders in Young People: Time Trends and Their Origins. Wiley, Chichester, pp. 490–543.

Spear, L., 2013. The teenage brain: adolescents and alcohol. Current Directions in Psychological Science 22 (2), 152–157.

Tarter, R., Vanyukov, M., Giancola, P., Dawes, M., Blackson, T., Mezzich, A., Clark, D., 1999. Etiology of early onset substance use disorder: a maturational perspective. Development and Psychopathology 11, 657–683.

Weichold, K. Translation of aetiology into evidence-based prevention: the life skills program IPSY. New Directions for Youth Development, in press.

Weichold, K., Wiesner, M., Silbereisen, R.K., 2014. Childhood predictors and adolescent correlates of developmental trajectories of alcohol use among male and female youth: results from a German study. Journal of Youth and Adolescence 43 (5), 698–716.

Wittchen, H.U., Behrendt, S., Höfler, M., Perkonigg, A., Lieb, R., Bühringer, G., Beesdo, K., 2008. What are the high risk periods for incident substance use and transitions to abuse and dependence? Implications for early intervention and prevention. International Journal of Methods in Psychiatric Research 17, S16–S29.

World Health Organization (WHO), 1997. Life Skills Education in Schools. World Health Organization, Geneva, Switzerland.

Alcohol Use and Abuse

Fiona Measham, School of Applied Social Sciences, Durham University, Durham, UK
Ian Paylor, Lancaster University, Lancaster, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Alcohol is consumed widely in almost all except Islamic countries of the world. Consumption levels vary widely by region and within individual countries, although generally higher levels of consumption can be found in more developed, higher income, and European countries. The negative consequences of alcohol 'misuse' can include short- and long-term health and social problems for drinkers, and for their families and wider society. Such problems are further shaped by multiple socioeconomic and contextual factors. The widespread yet complex nature of these problems means that a basic understanding of both alcohol misuse and brief interventions is essential for social workers.

Introduction – The Cultural Context to Consumption

The consumption of psychoactive drugs has been a feature of almost all societies, cultures, and tribes across humanity, although the types of substances, motivations for use, and the social controls that surround their use have varied considerably between different societies, historical periods, and cultural groups (Coomber and South, 2004). For some substances their use is condemned and criminalized, whereas for others there are degrees of social or legal acceptance. Motivations for use have ranged from the religious to the recreational. Alcohol, along with caffeine and tobacco, is a drug which has been widely used and socially accepted in most societies around the world for centuries, and is at the heart of many religious ceremonies, festivities, and social celebrations. The brewing and fermentation of alcoholic beverages from plants can be traced back to 5000 BC and the distillation of spirits to the AD ninth century. While alcohol use is outlawed in a handful of Islamic countries, in most countries it is legal to consume alcohol, although restrictions exist relating to aspects of the manufacture, promotion, sale, and consumption of alcoholic drinks, discussed further below. As Rehm et al. note (2003: 154) "with the exception of Islamic regions, alcohol is ubiquitous in the modern world." This article examines alcohol consumption, the problems following its use, and concludes by raising issues for social work practice.

Global Patterns of Consumption

There are enormous differences in alcohol use both within countries and between countries. The World Health Organization produces Global Status reports on alcohol and health which provide a comprehensive picture of patterns of international, national, and regional alcohol consumption along with health consequences and policy responses. One key way of comparing consumption levels around the world is to calculate the average per capita consumption of alcohol, usually measured in terms of liters of pure alcohol or ethanol consumed per year. In the most recent Global Status report (WHO, 2011), worldwide consumption of alcohol by people aged 15 and over was calculated to be 6.13 liters in 2005, ranging from 0.65 liters in the Eastern Mediterranean up to 12.18 liters in Europe.

Pakistan, an Islamic country where alcohol sale and use is illegal, has one of the lowest per capita consumption levels of 0.06 liters per annum, with drinking concentrated in higher income and secular groups, where it is seen as aspirational. By contrast, Russia has one of the highest per capita consumption levels at 15.7 liters, with two-thirds of alcohol consumption being of spirits and particularly vodka. By excluding abstainers from this total, per capital consumption among Russian men is estimated to be 35.38 liters of pure alcohol per annum (WHO, 2011). In the UK recorded per capita consumption was estimated at 10.2 liters of pure alcohol per person aged 15 years and over in 2010 (Gilmore et al., 2013), a slight fall on earlier years. If an estimate of 10% for unrecorded consumption is included (such as duty-free purchases and home-brewed alcohol) and further accounting for at least 10% abstainers in the UK, consumption increases to 12.3 liters per drinker, in line with the European average and about twice the global average. This translates to 1217 units of alcohol per year, or 23 units per person per week, which is higher than the recommended upper limits for low-risk consumption both for women and men (Shield et al., 2013).

Global alcohol consumption levels have been relatively stable since 1990, with a decline in the 1970–90s in some wine drinking countries and an increase in the 2000s in some Islamic regions. However, there are large differences between countries in terms of abstention rates, with higher numbers of drinkers in higher income countries in Europe and North America, and lower numbers of drinkers in North Africa and South Asia where there are greater numbers of Muslims in the population. Abstention is also higher among women than men. Given the large differences in the percentage of regular drinkers in countries around the world – from only a third in Jamaica up to 90% in some European countries (WHO, 2011) – the calculation for average per capita consumption can be modified by excluding abstainers to more accurately reflect average national consumption rates among drinkers.

Per capita consumption is further considered an underestimate of individual drinking because in many regions,

particularly in developing countries, the unrecorded production of homemade or illicit alcohol makes a significant contribution to total alcohol consumed, particularly if controls on legally produced alcoholic beverages are perceived to be too stringent. As the studies in Haworth and Simpson's (2004) international collection explore, the primary driver in the production and consumption of illicit alcohol is poverty. Illicit alcohol can have significant health risks resulting from its variable strength, lack of quality control, and potential contaminants, further compounding the health risks of legal alcoholic beverages (discussed below), although Haworth and Simpson suggest that high profile media coverage has sensationalized what can be exceptional events.

Types of Alcoholic Beverages

There are regional variations in types of alcoholic drinks consumed which reflect the climatic, geographical, and topological features conducive to cultivation of the various fruits and grains utilized in brewing and distillation processes. In general, just under half of total recorded alcohol consumed is of spirits, particularly favored in the Middle East and Asia, a third of alcohol consumed is beer, favored in the Americas, and under one tenth is wine, particularly in Europe. Another tenth takes the form of fermented grains and fortified wines, mostly in African countries (WHO, 2011).

Within Europe, drinking patterns have been characterized as 'dry' in northern European/Protestant countries and 'wet' in southern Mediterranean/Catholic countries. In the north, a pattern of heavy episodic drinking of beer and spirits with friends in bars at weekends and celebrations has predominated whereas in the south, alcohol consumption is characterized as more moderate, frequent, and integrated into daily family life, with wine consumed with meals throughout the week at home. Historically, it is the northern and north-western European and Scandinavian countries that have been associated with excessive consumption and public drunkenness (Jarvinen and Room, 2007). These traditional patterns of consumption have been changing in recent years, however, with growing popularity of European-style lagers and spirits among young people alongside a reduction in wine production and consumption, counterbalanced by a move to idealize southern European wine drinking cultures as producing more integrated and less problematic consumption, supported by various licensing initiatives and urban regeneration programs at local and national level (Jayne et al., 2008). Room and Makela (2000) have further elaborated on this 'wet' versus 'dry' dichotomy in their typology of drinking cultures. They identified four types of drinking cultures – abstinent, constrained ritual drinking, banalized drinking, and fiesta drunkenness – based on combining frequency of drinking with drunkenness. A more recent study of European drinking patterns by Makela et al. (2006), however, suggests that there were no clear differences between countries or regions, a situation further complicated by the influences of gender and age, so that regional variations are becoming less pronounced among younger generations with no country exhibiting an 'ideal' type of drinking culture.

Understanding Alcohol Abuse

The 'use-misuse-abuse' typology is a spectrum which characterizes the degrees of social acceptability of consumption of any psychoactive substance and distinguishes between legitimate, medically or socially sanctioned use (for example, medication from a recognized health practitioner); misuse where consumption is unsanctioned or illegitimate; and abuse where problematic use results in harmful consequences for the individual or society. Therefore, the term 'misuse' recognizes that (moderate) 'use' can be socially acceptable and prosocial in function, yet also allows for the possibility of unsanctioned or problematic use. In this article, a broad definition of alcohol abuse is utilized to cover the full range of drinking problems including harmful and dependent drinking.

The negative consequences of alcohol 'abuse' can include short- and long-term problems for drinkers, and for their families, communities, and wider society. Such alcohol-related problems are further shaped by multiple socioeconomic and contextual factors which, as young people grow up and start to drink, include the influences of parents, siblings, peer networks, and levels of family cohesion and support, as well as factors relating to the marketing and promotion of alcohol (Velleman, 2009).

While in many societies, moderate drinking is seen as both legally and socially acceptable, and for individual drinkers it can be associated with pleasurable and sociable leisure time experiences, for a small minority of drinkers problems can develop. How we define, measure, and explain such alcohol-related problems are themselves influenced by the historical and cultural context within which people are living, as well as differences within socioeconomic, cultural, religious, and demographic groupings. However, some broad observations are possible on the consequences of alcohol and abuse, relating to alcohol-related problems for individual drinkers and wider society.

Health Problems

The World Health Organization describes alcohol as 'one of the world's leading health risks' (2011: 20) linked to over 60 major diseases, accounting for 4.5% of the global burden of disease and injuries, and resulting in the premature death of 2.5 million people (4% of all deaths) each year. Among men aged 15–59, alcohol is the leading risk factor for death. Alcohol-related health problems can be both acute and chronic and are directly affected by not only the average volume of alcohol consumed but also patterns of drinking (Rehm et al., 2010). Acute harms result from intoxication produced by heavy episodic drinking rather than overall consumption levels, resulting in injury, road traffic accidents, poisoning, and sudden cardiac death (Rehm et al., 2003). Unintentional injuries alone account for 30% of all alcohol attributable deaths globally (WHO, 2011). However, the majority of health problems relate to chronic harm, produced by the cumulative effect of repeated drinking beyond recommended levels, resulting in liver cirrhosis (17% of alcohol attributable deaths), increased risk of other serious illnesses including various cancers (22% of attributable deaths), and cardiovascular

diseases (14% of attributable deaths). Up to half of all liver cirrhosis is attributed to alcohol (WHO, 2011). Dependent drinking – defined as repeatedly drinking beyond the sensible drinking guidelines recommended by health advisors and resulting in signs of physical or psychological dependence such as daily drinking, a compulsion to drink, increased tolerance, and a recognized withdrawal syndrome if attempting to stop – is particularly linked to these long-term health problems.

The extent to which drinking alcohol in moderation may offer cardiovascular protection has been debated intensely in recent years, with health professionals and alcohol industry representatives pitted against each other regarding the evidence base for such claims. There are exceptions of course. For example, it is recommended that pregnant women avoid alcohol as there is risk to the fetus. Drinking prior to operating machinery or driving a vehicle is actively discouraged and there is growing concern about the impact on not only the adolescent but also the young adult brain. The physical implications of alcohol misuse are well documented and include increased risk of liver cirrhosis, hepatitis, gastritis, pancreatitis, various cancers, nutritional deficiencies, obesity, diabetes, gout, cardiomyopathy, raised blood pressure, cardiac arrhythmia, strokes, brain damage, neuropathy, acute alcoholic poisoning, and impotence as well as insomnia, depression, anxiety, and amnesia, to name just a few. Damage can occur on both a physical and mental level. Indeed, there is a close link between alcohol dependency and mental health problems, including incidents of suicide. However, alcohol is part of a complex causal web of health and social problems, which makes the causal role of alcohol difficult to disentangle (Babor et al., 2003).

What does seem to be accepted, however, is that the global burden of alcohol-related disease will increase. This is because firstly, increased volumes of alcohol are being consumed in the two largest countries in the world (China and India) associated with their rapid economic development and social change; and secondly, globalization and alcohol advertising is leading to a convergence of drinking patterns. There has been a decline in the southern Mediterranean model of moderate drinking with meals and an increase in heavy episodic drinking as part of a young adult partying lifestyle which characterizes the northern European model (Rehm et al., 2003).

Social Problems

Alongside the health problems associated with alcohol abuse, there is a significant global burden associated with the social harms of alcohol. Social problems which can be attributed directly to alcohol include public drunkenness, disorder, and aggression, although alcohol's role in other social problems, as with many health problems, may be contributory rather than causal (Babor et al., 2003). Both heavy episodic drinking and dependent drinking can have a negative impact upon the health and well being of families, partners, and the wider social sphere of a problem drinker, with alcohol a contributory factor in stranger violence (for both offenders and victims), domestic violence or partner abuse (Hall, 2011), and child neglect and abuse. Intentional injuries account for 12% of all alcohol attributable deaths globally (WHO, 2011). There is also the indirect negative impact of alcohol dependency on significant others. It has been estimated that the stress, pain, and suffering resulting from drug and alcohol addiction impacts upon approximately 100 million family members worldwide (Orford et al., 2013). Finally, the economic impact of alcohol and abuse, measured in terms of lost productivity and absenteeism from work due to accidents, illness, and injury, has been calculated at £6.4 billion in the UK alone (PMSU, 2004).

Heavy Episodic Drinking: From Benders to Binges

At a global level, just over one in ten drinkers have had heavy drinking episodes on a weekly basis during the previous year (WHO, 2011). Heavy episodic drinking tends to be higher in higher income countries but is not necessarily directly related to high alcohol per capita consumption levels. So, for example, there are lower alcohol consuming countries such as some African and South Asian countries (e.g., India and Malawi) where there are high levels of heavy drinking among those who do drink. By contrast, there are some higher alcohol per capita consumption countries such as European wine drinking countries with regular but moderate consumption patterns (e.g., France) where heavy episodic drinking is relatively low.

While the term 'binge' was originally applied in clinical practice to refer to extended drinking resulting in the suspension of the drinker's usual daily activities for a day or more (also called a 'bender'), in the early 1990s American researchers applied the term to heavy consumption during a limited time period (Wechsler and Austin, 1998). In the two decades since then concerns have mounted regarding heavy episodic consumption. Indeed, as noted above (e.g., Rehm et al., 2010), heavy episodic drinking is associated with a range of negative health and social outcomes. However, there are considerable international variations in health advice regarding what could be considered a 'safe' or 'sensible' limit to alcohol consumption, either per drinking session or per week, and what could be considered a 'binge.' For example, in the UK a 'binge' is usually defined as drinking more than double the recommended maximum daily 'sensible' consumption levels in a single drinking session (Department of Health, 1995) of four units of alcohol for men and three units of alcohol for women, known as the 8/6 unit measure. Thus, a man consuming 64 g of alcohol and a woman consuming 48 g of alcohol would be classified as binge drinking. By contrast, in the USA a man would need to drink 70 g of alcohol and a woman 56 g to be classified as binge drinking.

While overall consumption levels have remained relatively stable since 1990 (WHO, 2011), a key change in this period has been an escalation in the quantity of alcohol consumed in a drinking episode, particularly by young people, and an associated increase in drinking to intoxication resulting in changing alcohol-related attitudes as well as behaviors, with a 'normalization of determined drunkenness' in the early twenty-first century (Measham, 2004: 321). This change in consumption patterns has been noted across most continents of the world, but is particularly pronounced in Europe (Jarvinen and Room, 2007). The reasons for this increase in 'binge' drinking can be linked to developments in the alcohol industry, as well as broader changes in the nature of adolescence, young adulthood, and leisure lifestyles in late modern

societies. In relation to young adults of drinking age, it has been suggested that a 'new culture of intoxication' emerged from the 1990s onward (Measham and Brain, 2005) which was linked to changes in licensed leisure venues and the associated expansion of the 'night time economy' (Hobbs et al., 2003).

Female Drinkers

More recently, concerns about drinking, public drunkenness, and health problems have shifted from young people toward women, particularly professional women. While heavy episodic drinking is much more common in men than women (WHO, 2011), and in boys than girls (Hibell et al., 2009), nevertheless in many countries women's consumption has increased at a faster rate than men's, if not yet on a par. This has led to suggestions that we are seeing a gender convergence at least in some countries (Bloomfield et al., 2001), driven by socioeconomic change including women's increased educational and employment attainments and associated increased disposable income, alongside advertising campaigns, promotions, and broader media coverage directed at female drinkers. This can be illustrated in the narrower gender differences in northern Europe than southern Europe, for example, with the frequency of heavy episodic drinking by northern European young women and young men showing the least marked gender difference of the various drinking variables under consideration (Makela et al., 2006).

Responses to Alcohol and Abuse: Population Level Policy Responses

Pricing

At population level, a strong evidence base supports the existence of a relationship between price and consumption, including influential modeling work conducted by researchers at the University of Sheffield (Meng et al., 2012) and from recent Canadian research (Zhao et al., 2013) which concluded that each percent increase in price results in a decrease in alcohol-related problems. Similarly, Ogilvie et al.'s (2005) synthesis of 30 reviews, including seven systematic reviews, concluded that a 10% increase in price in licensed premises reduced demand for beer by about 5% and by about 10% in off licenses. It is clear that the most efficient way to reduce alcohol-related problems at population level is to increase the taxation of alcoholic drinks. However, due to tensions between the interests and preferences of state, industry, and individual drinkers, alternative suggestions have included minimum unit pricing and graded differential taxation on alcohol by volume strength of alcoholic beverages, as well as differential taxation by beverage type to encourage the production of regional and heritage products, such as cider in the UK.

Availability

Restrictions on availability – such as restrictions on trading hours, density of outlets, types of alcoholic beverages sold, and particularly minimum purchase age – can be effective in reducing alcohol-related problems at population level. Brand et al. (2007) found that in general the stronger the alcohol control policy in any given country, the less people seemed to drink, although this was also mediated by the cultural context within specific countries. Other interventions at population level of lesser efficacy include restrictions on advertising and promotional practices, with impact dependent on both the degree of enforcement and whether voluntary industry self-regulation or state regulation. As with other areas of behavioral change such as banning smoking in public places or the introduction of mandatory seat belts for motor vehicle passengers, a key issue is effective enforcement.

Responses to Alcohol and Abuse: Health and Social Services and Individual Behavior Change

Early and brief interventions in primary care with problem drinkers can be effective. For some people at certain points in their lives, treatment for drinking problems can save lives and help facilitate change (NICE, 2013), while others are able to change their drinking behavior without any formal help (Klingemann, 2004). Indeed, some researchers suggest that individual drinkers 'grow out' of problematic use, with periods of 'excessive appetite' followed by the reassertion of self-control and 'normal' everyday life (Orford, 1985). Such change can occur quickly or slowly (Laudet and White, 2008); as part of the maturation process or triggered by changes in a person's life (Barrie, 2012); and without planning or as part of a carefully formulated strategy (Bischof et al., 2000).

A negative drinking experience or a reappraisal of the risks can lead to a rearrangement of the cognitive framework through which a person explains his/her drinking to him-/herself. The emergence of new perspectives which challenge previously held assumptions becomes a forerunner to change (Raistrick et al., 2006). This process can cause cognitive dissonance – an uncomfortable state of mind where a drinker is wrestling with two conflicting ideas or a discrepancy between his/her self-image and his/her behavior (Hohman, 2012).

While people with drinking problems may vacillate at the prospect of change, such ambivalence is normal although it can reduce motivation (Wagner and Ingersoll, 2013). Change holds the promise of benefits but it also involves loss and therefore it can provoke conformist 'comfort-zone' impulses (Miller and Rollnick, 2013). Understanding that ambivalence toward change is normal helps to make sense of the apparent reluctance of people with alcohol problems to act in ways which seem to be manifested in their best interests. Change may bring better health, the promise of a job, and avoidance of the criminal justice system but it can also mean losing a specific social life (Barrie, 2012). For some, change will bring the anxiety-provoking prospect of dealing with physical withdrawal. Where a person's identity is firmly bound up with drinking, a change in direction may entail a profound shift in their sense of self (White, 2012).

Prochaska and DiClemente's (1984) cycle of change suggests that people progress through six distinct stages:

- Precontemplation: the person either does not consider that they have a problem or has no intention of addressing it at this point;
- Contemplation: the contemplator has yet to take action but has begun to give serious consideration to the need to do so. Ambivalence is the hallmark of people in this stage;

- Preparation: the person has committed themselves to the idea that action is necessary and will be formulating what they should do;
- Action: the person alters their drinking behavior in significant ways and makes environmental and other changes which support this;
- Maintenance: this is not a static phase but one where the person acts to sustain the changes made and to avoid relapse;
- Relapse: a return to the previous behavior.

For health and social services, this cycle of change provides a relatively simple framework to address alcohol misuse (Thompson, 2009). However, the idea that people move consecutively through discreet stages has been questioned and some people make change without appearing to pass through the preparatory stages. Despite such criticisms, however, the popularity of the cycle of change suggests that it strikes a chord with alcohol services and service users alike (Paylor et al., 2012). While its limitations indicate that rigid application should be avoided, it provides a common sense framework which brings together a number of key themes including the centrality of motivation, ambivalence, and commitment to change.

Social Work Practice

More broadly, responding to alcohol use is a necessary part of many social workers' jobs, whatever their specialist area of practice. Social workers' knowledge of alcohol does not have to be comprehensive, but they do have to have enough to feel confident to ask the right questions and offer basic advice or brief interventions. Alcohol use cuts across areas of social work practice and does not discriminate according to gender, age, social class, ethnicity, or disability. Service users have a right to a social work service that is prepared for and committed to helping them change their problematic alcohol use.

Increasingly in many countries, however, it is specialist third sector organizations rather than the statutory sector that address alcohol-related problems, some of which have restricted or partisan approaches to what can be a politically and socially contentious problem. Most recently there has been an ideological shift in interventions and treatment in developed countries with the growth of 12-step fellowship programs and the development of a 'recovery movement' resulting in a shift from harm reduction to abstinence-based recovery. The ideological appeal of abstinence-based recovery programs and zealousness of some participants, as well as the contested evidence base for their efficacy, has led to fierce contemporary debate. In response to this, the concept of 'recovery' now focuses not only on overcoming alcohol or drug dependence, but is a much wider and more holistic notion which encompasses the achievement of positive outcomes in relation to health and well-being, educational and employment achievements, as well as personal relationships, and finances (ACMD, 2013).

Conclusion

Alcohol is a legal substance, widely available, not prohibitively expensive and contributes significantly to state economies in the form of employment and taxation receipts. Its use is not only condoned by most societies, but regarded as an important feature of family and social life. We have seen changes in the amount consumed, the consumers, and the products which they consume. There are various definitions of alcohol use and misuse ranging from those that are more physiologically based to those that are socially based. Different studies use different terminology, including 'heavy drinking,' 'dependence,' 'misuse,' and 'abuse' as well as 'problem drinking,' and 'binge drinking.' Such terms frequently overlap in the literature, with no common definition and minimal explanation of the criteria used to measure the behavior that gives cause for concern. However, there does seem to be consensus that what is being described and studied is the consumption of alcohol that warrants attention because it repeatedly, seriously, and negatively affects the drinker, their family, and wider society.

It is clear that alcohol abuse affects not only the individual but also those people around them. Problem drinking is an issue that cuts across many boundaries and impacts upon a wide range of agencies. The treatment of alcohol depends upon the philosophical perspectives of the individual and indeed the agency providing the assistance. Some suggest that problem alcohol use is a disease with a genetic disposition. Those that take this physiological approach would argue that the likelihood of becoming an 'alcoholic' is much more determined by genetics than by social environment. Others see periods of alcohol misuse as part of a maturation process and a symptom of the wider struggle between greater and lesser self-control within 'normal' everyday adolescent and adult life. Still others view drink problems as essentially social, cultural, and psychological in origin. This approach seeks to educate and empower the problem drinker by concentrating on the identification of triggers, insight into high-risk situations, the development of relapse prevention strategies, and positive self-talk. They suggest that five key stages exist, though not necessarily in any order (Precontemplation, Contemplation, Preparation, Action, and Maintenance), and believe that it is possible for health and social service professionals to intervene at any stage, provided the stage has been properly identified. Regardless of these various approaches to the problem, what is clear is that the scale and ubiquity of alcohol-related problems means that many social workers will encounter alcohol misuse in their workloads and therefore an understanding of both alcohol misuse and the potential of brief interventions and motivational interviewing to support change is essential for contemporary social work practice.

See also: Addictions: General Considerations; Alcohol Interventions: Disease Models vs. Harm Reduction; Alcohol Use among Young People; Alcohol-Related Disorders; Drinking, Anthropology of; Drug Use and Abuse: Psychosocial Aspects; Drugs: Illicit Use and Prevention.

Bibliography

Adfam, 2012. 'Challenging Stigma': Tackling the Prejudice Experienced by the Families of Drug and Alcohol Users. Available at: http://www.adfam.org.uk/docs/adfam_challenging_stigma.pdf (accessed 30.10.13.).

Advisory Council on the Misuse of Drugs, 2013. What Recovery Outcomes Does the Evidence Tell Us We Can Expect? Second Report of the Recovery Committee. ACMD, London.
Babor, T., Caetano, R., Casswell, S., et al., 2003. Alcohol: No Ordinary Commodity—Research and Public Policy. Oxford University Press, Oxford.
Barrie, K., 2012. Alcohol. Dunedin Academic Press, Edinburgh.
Best, D., Rome, A., Hanning, K., White, W., Gossop, M., Taylor, A., Perkins, A., 2010. Research for Recovery: A Review of the Drugs Evidence Base. Available at: http://www.scotland.gov.uk/Publications/2010/08/18112230/0 (accessed 30.10.13.).
Bischof, G., Rumpf, H., Hapke, U., Meyer, C., John, U., 2000. Maintenance factors of recovery from alcohol dependence: a cluster analysis approach. Addiction 98, 1737–1746.
Bloomfield, K., Gmel, G., Neve, R., Mustonen, H., 2001. Investigating gender convergence in alcohol consumption in Finland, Germany, The Netherlands, and Switzerland: a repeated survey analysis. Substance Abuse 22 (1), 39–53.
Brand, D., Saisana, M., Rynn, L., Pennoni, F., Lowenfels, A., 2007. Comparative analysis of alcohol control policies in 30 countries. PLoS Medicine 4 (4), e151. http://dx.doi.org/10.1371/journal.pmed.0040151.
Coomber, R., South, N., 2004. Drug Use and Cultural Contexts 'Beyond the West': Tradition, Change and Post Colonialism. Free Association Books, London.
Department of Health, 1995. Sensible Drinking: The Report of an Interdepartmental Working Group. Department of Health, London.
Gilmore, I., Anderson, W., Bauld, L., Bellis, M., Brown, K., Drummond, C., et al., 2013. Health First: An Evidence-based Alcohol Strategy for the UK. Published by the University of Stirling, Scotland, UK. A report published in association with the British Liver Trust and the Alcohol Health Alliance UK. ISBN 978-1-908063-12-0. Available at: http://www.stir.ac.uk/media/schools/management/documents/Alcoholstrategyupdated.pdf.
Hall, P., 2011. 'Intimate violence: 2009/10 British crime survey'. In: Smith, K., Coleman, K., Eder, S., Hall, P., (Eds.), Homicides, Firearm Offences and Intimate Violence 2009/10 Supplementary Volume 2 to Crime in England and Wales 2009/10 (second ed.). Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/116512/hosb0111.pdf (accessed 06.12.13.).
Haworth, A., Simpson, R. (Eds.), 2004. Moonshine Markets: Issues in Unrecorded Alcohol Beverage Production and Consumption. Brunner-Routledge, New York.
Hibell, B., Guttormsson, U., Ahlström, S., Balakireva, O., Bjarnason, T., Kokkevi, A., Kraus, L., 2009. The 2007 ESPAD Report: Substance Use Among Students in 35 European Countries. The Swedish Council for Information on Alcohol and Other Drugs, Stockholm.
Hobbs, D., Hadfield, P., Lister, S., Winlow, S., 2003. Bouncers: Violence and Governance in the Night-time Economy. Oxford University Press, Oxford.
Hohman, M., 2012. Motivational Interviewing in Social Work Practice. The Guildford Press, London.
Humphreys, K., 2006. The trials of alcoholic anonymous. Addiction 101, 617–618.
Jarvinen, M., Room, R. (Eds.), 2007. Youth Drinking Cultures: European Experiences. Ashgate, Aldershot.
Jayne, M., Valentine, G., Holloway, S., 2008. Fluid boundaries – british binge drinking and european civility: alcohol and the production and consumption of public space. Space and Polity 12 (1), 81–100.
Klingemann, H., 2004. Natural recovery from alcohol problems. In: Heather, N., Stockwell, T. (Eds.), The Essential Handbook of Treatment and Prevention of Alcohol Problems. Wiley, Chichester.
Laudet, A., White, W., 2008. Recovery capital as prospective predictor of sustained recovery, life satisfaction and stress among former poly-substance users. Substance Use and Misuse 42 (1), 27–54.
Makela, P., Gmel, G., Grittner, U., Kuendig, H., Kuntsche, S., Bloomfield, K., Room, R., 2006. Drinking patterns and their gender differences in Europe. Alcohol and Alcoholism 41 (Suppl. 1), i8–i18. http://dx.doi.org/10.1093/alcalc/agl071.
Measham, F., 2004. The decline of ecstasy, the rise of 'binge' drinking and the persistence of pleasure. Probation Journal 51 (4), 309–326.
Measham, F., Brain, K., 2005. Binge' drinking, British alcohol policy and the new culture of intoxication. Crime, Media, Culture: An International Journal 1 (3), 263–284.
Meng, Y., Hill-MacManus, D., Brennan, A., Meier, P., 2012. Model-based Appraisal of Alcohol Minimum Pricing and Off-licensed Trade Discount Bans in Scotland Using the Sheffield Alcohol Policy Model (V. 2): Second Update Based on Newly Available Data. ScHARR: University of Sheffield.
Miller, W.R., Rollnick, S., 2013. Motivational Interviewing, third ed. Guildford Press, New York.
NICE, 2013. Alcohol-use Disorders - Preventing Harmful Drinking (PH24). Available at: http://guidance.nice.org.uk/PH24 (accessed 30.10.13.).
Ogilvie, D., Gruer, L., Haw, S., 2005. Young people's access to tobacco, alcohol, and other drugs. British Medical Journal 331, 393. http://dx.doi.org/10.1136/bmj.331.7513.393.
Orford, J., 1985/2001. Excessive Appetites: A Psychological View of Addictions. Wiley, Chichester.
Orford, J., Velleman, R., Natera, G., Templeton, L., Copello, A., 2013. Addiction in the family is a major but neglected contributor to the global burden of adult ill-health. Social Science & Medicine 78 (1), 70–77.
Purshouse, R., Brennan, A., Latimer, N., Meng, Y., Rafia, R., Jackson, R., 2009. Modelling to Assess the Effectiveness and Cost-effectiveness of Public Health Related Strategies and Interventions to Reduce Alcohol Attributable Harm in England Using the Sheffield Alcohol Policy Model Version 2.0: Report to the NICE Public Health Programme Development Group. University of Sheffield, Sheffield.
Paylor, I., Measham, F., Asher, H., 2012. Social Work and Drug Use. Open University Press, Maidenhead.
Prime Minister's Strategy Unit, 2004. Alcohol Harm Reduction Strategy for England. Strategy Unit, Cabinet Office, London.
Prochaska, J., DiClemente, C., 1984. The Transtheoretical Approach. Dow Jones-Irwin, IL.
Raistrick, D., Heather, N., Godfrey, C., 2006. Review of the Effectiveness of Treatment for Alcohol Problems NHS National Treatment for Substance Misuse. Available at: http://www.nta.nhs.uk/uploads/nta_review_of_the_effectiveness_of_treatment_for_alcohol_problems_fullreport_2006_alcohol2.pdf (accessed 30.10.13.).
Rehm, J., Rehn, N., Room, R., Monteiro, M., Gmel, G., Jerniganh, D., Frick, U., 2003. The global distribution of average volume of alcohol consumption and patterns of drinking. European Addiction Research 9, 147–156. http://dx.doi.org/10.1159/000072221.
Rehm, J., Baliunas, D., Borges, G., Graham, K., Irving, H., Kehoe, T., Parry, C., Patra, J., Popova, S., Poznyak, V., Roerecke, M., Room, R., Samokhvalov, A., Taylor, B., 2010. The relation between different dimensions of alcohol consumption and burden of disease: an overview. Addiction 105 (5), 817–843. http://dx.doi.org/10.1111/j.1360-0443.2010.02899.x.
Room, R., Makela, K., 2000. Typologies of the cultural position of drinking. Journal of Studies on Alcohol and Drugs 61 (3), 475–483.
Shield, K., Rylett, M., Gmel, G., Gmel, G., Kehoe-Chan, T., Rehm, J., 2013. Global alcohol exposure estimates by country, territory and region – a contribution to the comparative risk assessment for the 2010 global burden of disease study. Addiction. http://dx.doi.org/10.1111/add.12112.
Thompson, N., 2009. Practicing Social Work. Palgrave Macmillan, Basingstoke.
Velleman, R., 2009. Influences on How Children and Young People Learn about and Behave towards Alcohol: A Review of the Literature for the Joseph Rowntree Foundation (Part One). Joseph Rowntree Foundation, York. Available at: www.jrf.org.uk.
Wagner, C., Ingersoll, K., 2013. Motivational Interviewing in Groups. Guildford Press, New York.
Wechsler, H., Austin, S., 1998. Binge drinking: the five/four measure [Letter]. Journal of Studies on Alcohol 59, 122–123.
White, T., 2012. Working with Drug and Alcohol Users. Jessica Kingsley Publishers, London.
World Health Organization, 2011. Global Status Report on Alcohol and Health. WHO, Geneva.
Zhao, J., Stockwell, T., Martin, G., Macdonald, S., Valance, K., Treno, A., Ponicki, W., Tu, A., Buxton, J., 2013. The relationship between changes to minimum alcohol price, outlet densities and alcohol-related death in British Columbia, 2002–2009. Addiction. Available at: http://onlinelibrary.wiley.com/doi/10.1111/add.12139/pdf (accessed 27.02.13.).

Relevant Websites

www.findings.org.uk/index.php – Findings.
www.icap.org – International Centre for Alcohol Policies (ICAP).
www.swalcdrugs.com/ – Social Work, Alcohol and Drugs.
www.who.int – World Health Organization.

Alexander, Jeffrey C. (1947–)

Jason L Mast, Institute of Advanced Study, University of Warwick, Coventry, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Jeffrey C. Alexander is a social theorist and cultural sociologist. This article describes Alexander's intellectual developments and professional achievements, from his undergraduate days at Harvard University through to the present. Four threads of Alexander's intellectual pursuits are discussed, including his contributions to sociological theory, his development of the strong program in cultural sociology, his engagement in political and social theory and the publication of *The Civil Sphere*, and, finally, his recent contributions on the topics of social performance and iconicity.

Jeffrey C. Alexander is a social theorist and cultural sociologist. With the publication of his doctoral thesis as a four-volume series, titled, *Theoretical Logic in Sociology*, he established himself as a promising new voice in American sociological theory, and took a considerable step toward becoming one of the most influential social theorists in the world. Since the publication of these volumes, Alexander has helmed intellectual movements in neofunctionalism and cultural sociology, theorized and empirically reconstructed the American civil sphere, and engaged in a series of interventions into subjects ranging from social performance, to cultural trauma, to iconicity, and to events as varied as US presidential elections and Egypt's performative revolution. His work is characterized by theoretical sophistication and philosophical musicality, elegant and authoritative prose, an adventurer's openness to interdisciplinarity, and a commitment to maintaining the centrality of meaning in representations of social life.

Alexander was born on 30 May 1947 in Milwaukee, Wisconsin, to Frederick Charles Alexander and Esther Leah Schlossman Alexander. His father worked in advertising, and his mother, who earned a Master of Arts in Social Work from the University of Wisconsin, Milwaukee, stayed at home to raise the couple's two sons and later worked for the Visiting Nurses Association of Los Angeles. Jeffrey's younger brother, Charles 'Skip' Alexander, earned a PhD in psychology from Harvard in 1982, and was a leading figure and published extensively on the intersections of developmental psychology and meditative philosophies and practices.

Jeffrey attended public high school in Los Angeles, California, and in 1965 entered Harvard as an undergraduate in the university's Committee on Degrees in Social Science program. Alexander's intellectual trajectory and writing acumen were formed in the milieu of an interdisciplinary education in political theory, literature and drama studies, and moral and political philosophy. He contributed regularly to the *Harvard Crimson*, participated in New Left study groups, and participated in movement organizations like the Students for a Democratic Society.

A self-described Sixties communard, Alexander's formative intellectual identity took inspiration from representatives of the cultural Marxism of the day, and he identified more with the Marcuse, Gramsci, and Sartre types than voices of the more orthodox, economistic variety (2005: 40–42). Actively engaged in student political culture, writing for the university newspaper, and studying literary, political, and critical theory, Alexander accumulated intellectual ingredients and scholarly practices that would become hallmarks of his future academic career: a command over a broad range of intellectual movements, a disciplined approach to writing, and a commitment to emphasizing and honing analytical tools for apprehending the meaning structures that constitute and order social existence.

Despite having taken but a few social science courses at Harvard, which included a sociology course in his final year taught by the towering figure, Talcott Parsons, Alexander decided to pursue a PhD in sociology and entered the University of California, Berkeley's program in 1969.

Alexander continued to engage in political activism during his first few years at Berkeley. Yet with sectarianism splitting the radical Left, Nixon's backlash politics winning the American political center, and with his interests shifting toward intellectual pursuits, Alexander turned to his studies and to more diverse, or classical and modern, sources for illumination than the Marxism that had shaped his early intellectual self.

Alexander gravitated toward theory courses at Berkeley. He took a year-long sociological theory seminar helmed by Neil Smelser, twice, and immersed himself in Durkheim's work in a course taught by Leo Lowenthal. Robert Bellah had just published his influential essay, 'Civil Religion in America' (1967), and this article as well as Bellah's interests in symbolism, Durkheim, and Weber, and his friendship with anthropologist, Clifford Geertz, shaped Alexander. In addition to contributing to Alexander's knowledge of sociological theory, these three figures, Bellah, Lowenthal, and Smelser, served as members on his doctoral thesis committee. Lowenthal represented a connection to Alexander's early interest in cultural Marxism; Bellah cultivated Alexander's interest in symbolic realism; and Smelser, a leading role in extending Parsonian functionalism, and the personification of theoretical precociousness and youthful prodigiousness, left a mark on the young Alexander as well.

Alexander started an Assistant Professorship in sociology at the University of California Los Angeles (UCLA) in 1976 and received his PhD from Berkeley in 1978, and the *Theoretical Logic* volumes were published in 1982–83. He was promoted directly to a full professorship at UCLA in 1981. He remained at the university for the next 20 years, during which he served in multiple capacities, including as Department Chair from 1989 to 1992. An institutional entrepreneur, Alexander helped build

the UCLA sociology program into one of the most diverse, eclectic, and highly ranked programs in the United States during his tenure.

Also while at UCLA, beginning in 1984, Alexander formed a discussion group with graduate students interested in investigating culture in society, which adopted the moniker the 'Culture Club.' Alexander and an impressive series of graduate students took the cultural turn together and, over the subsequent three decades, built a body of publications that constitute the core of a perspective that Alexander named (invoking David Bloor and colleagues' strong programme in the sociology of scientific knowledge) the 'strong program in cultural sociology.' Alexander moved to Yale University in 2001, where he became Lillian Chavenson Saden Professor of Sociology, and where he resumed institution building. He served as Department Chair from 2002 to 2005 and helped revitalize the program, which was undergoing a challenging period of transition. Also upon arriving at Yale, Alexander formally institutionalized the 'Culture Club' by launching the Center for Cultural Sociology, which he codirects with Yale sociology colleagues Ron Eyerman, Philip Smith, and as of 2013, Frederick Wherry.

A public sociologist in many senses of the phrase, Alexander has deep connections in the field and a vast network of colleagues on whom he has drawn for intellectual, collegial, and personal inspiration and support. In addition to contributing to Alexander's intellectual and professional development, Neil Smelser and the recently deceased Robert Bellah played central roles in the middle and later stages of Alexander's career, and Alexander cites both as representing important exemplars of academic, intellectual, and personal integrity to him over the duration of his career. Alexander has also acknowledged Nicholas Entriken, Ron Eyerman, Roger Friedland, Bernhard Giesen, Steven Seidman, Piotr Sztompka, and Kenneth Thompson as colleagues who have challenged, inspired, and supported him over the course of his career.

Since the mid-1970s, Alexander has been extraordinarily prolific, and discernible threads run through the articles and books he has produced during the four decades of this most productive and creative of careers.

Theory

In the earliest thread Alexander wove into what would become his oeuvre, he adopted an unpopular and antagonistic stance by making the centerpiece of his ambitious intellectual project the then besieged and decidedly unfashionable legacy of Talcott Parsons and the Parsonian structural functionalism that dominated mid-century American sociology. A backlash against Parsonianism was in full sway in 1960s and the 1970s American sociology, propelled by critical theorists arguing that Parsons had neglected conflict, power, and a pervading instrumental rationality, on the one hand, and by microsociologists or symbolic interactionists arguing that Parsons's structuralism neglected the micro, interactional sources of social order and the creative, strategic dimensions of individual actors, on the other. While Alexander appreciated the Parsonian project, his aims in this initial phase transcended Parsons's, and were motivated toward legitimating a postpositivist understanding of theory and social science, advocating for multidimensionality in social scientific endeavors, and in rescuing voluntarism from theory's determinisms.

Alexander opens the *Theoretical Logic* volumes by offering a schematic of the structure of scientific thought. The elemental components of this thought compose a continuum, ranging from the abstract, theoretical pole on one end, to the empirical, 'factual' pole on the other. On the theoretical side lie presuppositions and ideological orientations that characterize assumptions about the originating sources of social action and order, while the empirical pole is composed of issues of empirical testing and measurement, methods, and observations of the factual world. Between these two poles lie intermediate levels of scientific frameworks such as formulations of concepts, definitions, and laws.

From this model Alexander makes several points about the practice of social science: these poles cannot and do not operate independently of one another. Be it deductive, theory-driven science or inductive, model building from observation and facts, scientific representations demonstrate assumptions about both poles on the continuum. "Theory cannot be built without facts," he states in a recapitulation of his argument in *Twenty Lectures* (1987: 5), "but it cannot be built only with them either." Alexander's version of postpositivism argues that social scientists should be reflexive regarding how each pole informs, shapes, and constitutes the other, as well as the dimensions that lie between.

Alexander also argues that all social scientific theories contain, and indeed must contain, assumptions, if not explicit statements, regarding two core dimensions of the social; namely, they must contain characterizations about the nature, origins, and sources of both social action and social order. Presuppositions about social action demonstrate assumptions that people act according to one of two logics, or that social actions are born primarily of rational or nonrational sources and motives. Rational models of action portray actors as generally selfish, as acting to maximize efficiency, and as responding to forces outside of one's self. Nonrational models, by contrast, emphasize that people are motivated by internal sources that are normative and/or moral in nature, or based on emotions and unconscious desires. In terms of the production of social order, or the regularity of patterns in social life, sociologists presuppose either individualistic sources or collectivist sources. Individualistic formulations of order acknowledge that supraindividual patterns appear to exist, but these formulations emphasize that they result from individual choice and negotiation, and are based on considerations made in real time. Collective formations, in comparison, focus on patterns and structures that exist regardless of and prior to any specific individual experiencing them.

In advocating for multidimensional approaches to doing sociology, Alexander argues that tight adherence to either side of the action and order dichotomies produces reductionistic, one-dimensional interpretations whose explanatory powers are weakened because they sacrifice the influence and power of the alternative positions. One-dimensional explanations produce residual categories – or sites, phenomena, or processes – that, since they do not fit into the logic of the dominant explanatory model, must be addressed in ad hoc and oftentimes contradictory ways.

Alexander concludes by dismissing individualistic approaches to order, arguing that they make the very problem

of order itself a residual, or unexplained, category; he does acknowledge, however, that individualistic approaches have revealed important dimensions of some empirical processes. Social action, he concludes, neither should be reduced to either instrumental or normative characterizations, nor should it be conceived as ordered by either internal or external structures; social action should be conceived as composed and motivated by all these sources (1982a: 123).

Over the course of the next three volumes, Alexander analyzes and reinterprets Marx, Durkheim, Weber, and Parsons's bodies of work in terms of their approaches to the problems of action and order. Durkheim and Marx, both with collectivist approaches to order, move in opposite directions over the courses of their intellectual careers, Alexander argues. Action in Marx's mature writing became rigidly rational, mechanistic, instrumental, and strictly materialist in contrast to the more voluntaristic and nonrationalist formulations that characterized writings from his earlier, more Hegelian-inspired output. While the young Durkheim's response to the questions of action was more ambiguous, it nonetheless demonstrated strong materialist assumptions. As he matured, however, Durkheim's formulations of action became increasingly idealist and voluntaristic, and matched the mature Marx in terms of his adherence to a one-dimensional approach to the issue. Weber, on the hand, represented an important while incremental and inconsistent move toward multidimensionality. While Alexander sees instances of multidimensionality in Weber's writings on religion, particularly in his treatment of the protestant ethic, as well as in his methodological essays, far more often Weber's work is characterized by one-dimensional moves into either the materialist or idealist camp, and in Alexander's view, it is most often grounded in materialist formulations (e.g., Weber's political writings in *Economy and Society*). It is in the work of Talcott Parsons, Alexander finds that sociological theory has realized its most faithfully multidimensional framework to date.

The publication of the *Theoretical Logic* volumes drew enormous attention, and produced a considerable number of both critical and laudatory reviews. All acknowledged the enormous ambition, scope, thoroughness, and creativity of Alexander's achievement, however, and these publications established Alexander as one of the leading sociological theorist of his generation.

Under the rubric of neofunctionalism – named, in effect, to suggest an alternative to the era's neo-Marxism – Alexander, frequently collaborating with one of his first graduate students, Paul Colomy, spent over a decade critically reenvisioning Parsons's work, acknowledging its weaknesses and contradictions, and articulating theoretical avenues for navigating beyond its limitations. Even after declaring that the neofunctionalist project had come to an end (1998: 8), Alexander remained deeply engaged in theoretical sociology, producing what has been interpreted as a fifth volume of *Theoretical Logic* in his book, *Twenty Lectures* (1987), which examines postwar social theoretical developments and their presuppositional orientations to the problems of social action and order. Likewise, in a series of essays that examine how intellectuals have narrated, coded, and explained their epochs (modern, anti, post, and neo), Alexander's orientation to the field became increasingly metatheoretical, cultural, and more purely social theoretical. Characteristic of these works, which represent and powerfully reiterate the rationale for his postpostivist challenge to future theorizing, Alexander declares, "there is always an eschatology, not merely an epistemology, in theorizing about social change" (1995: 10; see also Alexander, 2013, for essays from this period analyzing evil, domination, and the decivilizing aspects of modernity).

Strong Program in Cultural Sociology

A second thread in Alexander's work began to take clear form in the mid-1980s and grew into programmatic maturity over the remainder of the decade. During a year at the Institute for Advanced Study near Princeton, in 1985–86, Alexander wrote drafts of two papers in the language of a strong cultural sociology: "Culture and Political Crisis: Watergate and Durkheimian Sociology" (1988a), and "The Promise of a Cultural Sociology: Technological Discourse and the Sacred and Profane Information Machine" (1993). By the turn of the century, 'cultural sociology' would be synonymous with Alexanderian sociology.

One critical ingredient in fomenting Alexander's turn to culture came through one of his earliest engagements with the German sociology scene, a practice of intellectual exchange that he has engaged in regularly throughout his career. In 1983, Alexander (1988b: xi) participated in a conference on "Emile Durkheim's Sociology of Religion" at the Werner Reimers Stiftung Foundation in Bad Homburg, Germany. The exchange spurred in Alexander (1986a,b) a thorough reinvestigation of Durkheim's intellectual trajectory from materialism to subjectivism, as well as his own turn to interdisciplinarity, through which Alexander found scholars in neighboring academic disciplines who had extended and refined the fertile terrain of symbolic classifications that Durkheim had so powerfully elaborated in his late work, *The Elementary Forms of Religious Life*. In the following years, Alexander (1988c) incorporated into his theorizing of cultural sociology Ferdinand Saussure's semiotics; elements of structuralism from the anthropology of Levi-Strauss, Mary Douglas, Marshall Sahlins, and Clifford Geertz; a model of flow from Victor Turner's work on social dramas; Roland Barthes's literary and narrative theory; Foucault's work with discourse; and his mentor Robert Bellah's development of symbolic realism.

In one of his first empirical investigations into demonstrating the purchase of this cultural framework, "The Promise of a Cultural Sociology: Technological Discourse and the Sacred and Profane Information Machine," Alexander (1993) challenges the thesis that the modern world is increasingly being dominated by a creeping technical rationality, a legacy of Weber's iron cage metaphor and Marx's assertion that all human relations will be rendered into commodity form. Using popular representations of the computer's rise and increasing ubiquity in American society, in which the technological artifact is folded into cultural frameworks of salvation and apocalypse, Alexander identifies the "omnipresent shaping of technological consciousness by discourse," and concludes that it is only through understanding this process of cultural constitution can we "gain control over technology in its material form" (192).

Likewise, in his early writings on domestic American politics, two familiar sources of motivation informing Alexander's investigations into how the cultural world shapes social life can be seen: a theoretical challenge to overcome contradictions and conflations within Parsons's approach to reconciling values and culture with social system integration, on the one hand, and

explaining and understanding Nixonian America and the Watergate scandal, on the other. Alexander wrote three versions on the topics, and over the course of these drafts, the emergence of a strong cultural sociology can be seen: Parsons dominates the first version (1984), while Durkheim's 'religious sociology,' and a dramaturgical sensibility interspersed with concepts from anthropologist Victor Turner's work on social dramas, dominate the last (1988a).

Empirically Watergate continued to prove fertile terrain as well. It was in 1985, while watching 50 hours of condensed video footage of the *CBS Evening News with Walter Cronkite*, footage selected for containing the term 'Watergate,' that Alexander first conceived of 'a noncontingent binary structure' that shapes interpretations of social and civil life in the American collective imagination (2006: 572–573, note 10). In two publications in the early 1990s Alexander pursued this idea and produced a theory of the binary codes of civil society, a framework that would critically inform much of his subsequent writings on democracy and the civil sphere. In "Citizen and Enemy as Symbolic Classification," Alexander (1992) asserts that in as much as it is an institutional realm or the province of elites' interests, civil society is also a "structured, socially established consciousness, a network of understandings that operates beneath and above explicit institutions and the self-conscious interests of elites." This consciousness, he argues, is a set of distinctive symbolic codes that define the basic categories of citizens' understandings that distinguish the pure and the impure as well as the democratic and the counterdemocratic and establish the differences between liberty and repression. This binary code, Alexander states, constitutes "the very sense of society for those who are within and without it" (290).

The theory of the binary codes of civil society elaborated in Alexander's "Citizen and Enemy as Symbolic Classification" (1992) would become programmatic. The theory is most frequently associated, however, with the argument's second – revised and empirically extended – iteration, namely, in "The Discourse of American Civil Society" (1993), which Alexander cowrote with his former graduate student and periodic coauthor, Philip Smith. In it the authors further develop Alexander's theory of the binary codes of civil society, and substantiate it with an extended empirical investigation. Examining American political crises and scandals from the 1830s through the 1980s, Alexander and Smith demonstrate that an enduring binary code, or a cultural framework of understanding, has been repeatedly invoked by antagonists in these events in their efforts to portray themselves in positive, democratic hues, and their enemies as motivated by profane, counterdemocratic tendencies. Through this investigation into conflicts throughout the history of American civil society, Alexander and Smith identify "the durability and continuity of a single culture structure over time," a structure "which is able to reproduce itself discursively in various highly contingent contexts" (1993: 198).

In an early and succinct manifesto for a cultural sociology, in an American Sociological Association (ASA) Culture section newsletter contribution titled, "Cultural Sociology or Sociology of Culture? Toward a Strong Program" (1996), Alexander draws a strict line distinguishing 'cultural sociology' from the 'sociology of culture,' and argues that the latter routinely frames culture as a dependent variable to be explained by something entirely outside the realm of meaning itself. To do cultural sociology, rather, means one must subscribe to the idea that "[e]very action, no matter how instrumental and reflexive vis-à-vis its external environments … is imbedded in a horizon of meaning (an internal environment) in relationship to which it can be neither instrumental nor reflexive" (1996: 3; see also Alexander and Smith, 2002).

Alexander designed the strong program and adumbrated its contours. In no small measure it has been his eagerness to work with graduate students over the past two decades that has helped this perspective into develop into a fully formed intellectual movement. Many of these students went on to academic careers, and representatives of the strong program in cultural sociology are found in sociology departments across the United States, Europe, and Asia. Early contributors to the program's body of literature include former members of the Culture Club at UCLA such as Elaine Chan, Laura Edles, Ronald Jacobs, Anne Kane, Agnes Ku, Eric Rambo, Andy Roth, Steve Sherwood, and Philip Smith. Contributors from more recent generations include Dominik Bartmanski, Elizabeth Breese, Shai Dromi, Rui Gao, Nadya Jaworski, Jason Mast, Lisa McCormick, Matthew Norton, Eyal Rabinovitch, and Isaac Reed. Also a testament to the movement's vitality, Alexander launched with Ronald Jacobs and Philip Smith *The American Journal of Cultural Sociology*, a peer-reviewed journal dedicated to publishing new developments in meaning-centered sociology.

The Civil Sphere and Civil Society

Alexander published his magnum opus, *The Civil Sphere*, in 2006. It is substantial in size and scope, ambitious and creative in its theorizing, and thorough and illuminating in its empirical detail. It is also the culmination of over two decades of work in democratic, social, and political theory; it is representative of his cultural sociology; and it incorporates important dimensions of his recent interventions in social performance, trauma, and iconicity. It is the thread to which the other three threads in this article all join and form the representative image of Alexander's overarching project.

Alexander received a fellowship from the John Simon Guggenheim Memorial Foundation in 1979 to pursue a project titled, "Watergate and the Crisis of Civil Society." Civil repair was a pressing concern in the wake of the Watergate crisis, and it was at this time that Alexander (1980) started pursuing an interest in social solidarity, and in-group/out-group, inclusion/exclusion processes in modern society. But it was in the mid-1980s that his grand vision of a theory of civil society began to take fuller form. In addition to turning to cultural theory during this time (discussed above), Alexander received an invitation from political philosopher Michael Walzer to spend a year (1985–86) at the Institute for Advanced Study in Princeton, New Jersey. Walzer had influenced Alexander during the latter's formative, undergraduate years, when they were both at Harvard. It was during his second meeting with Walzer at the Institute of Advanced Study (IAS) that Alexander turned more directly to engaging classical and contemporary democratic theory.

Shortly after his stay at the IAS, civil society would become a central issue not only in academic discourse but also in popular discourse. In the introduction to his edited volume, *Real*

Civil Societies (1998), Alexander recounts making his first return to Eastern Europe in 1990, following the initial phases of the Soviet Union's dissolution. He witnessed Hungary and Poland celebrating the return of civil society, and it was at this time, Alexander reports, that he experienced civil society not just as an 'ideal' but as a 'real' force in society; that civil society was, indeed, a social fact (1).

If it was during his trip to Hungary that his new theory of the civil sphere 'crystallized in its final form,' then it was over the course of the following decade that Alexander began laying the foundations for what would become *The Civil Sphere*.

For instance, in "Bringing Democracy Back In" (1991), Alexander argues that democracy is not just a formal arrangement, and politics is not just the self-interested pursuits of elites and oligarchs. While they are, in part, these things, they are also about a medium of communication and are constituted by a symbolic code (160). With the ideal of theoretical multidimensionality shaping his argument, Alexander concludes that, distinct from the state, civil society is a "realm of solidarity and cultural universalism but also … an institutionalized and differentiated social sphere" (171). Alexander then turned to exploring the paradoxes of civil society (1994), and to articulating a new theory of social movements (1996) in which he argues that they represent social devices that "construct translations between the discourse of civil society and the institution-specific process of a more particularist type." Social movements succeed, he concludes, to the extent that they are able to "employ the civil metalanguage" that will relate the movement's "practical problems to the symbolic centre of society and its utopian premises" (229). From 1998 to 2001, Alexander's theory of the civil sphere assumes its full form as he identifies the three ideal typical forms of theorizing about it (1998), argues that the contemporary world is characterized by partial and plural utopian visions that are oriented toward difference and the utopia of civil repair (2001a,b), and theorizes the three modes of incorporation, namely, assimilation, hyphenation, and multiculturalism (2001c).

The Civil Sphere appears in print in 2006 with Alexander having built on and extending this theoretical framework, incorporating his newly formed theories of social performance (2004a) and cultural trauma (2002, 2004b), and having produced in-depth analyses of empirical cases on gender and motherhood (Part III), the civil rights movement (Part III), and the American Jews (Part IV). Upon publication, *The Civil Sphere* gained enormous attention, sparked considerable debate, and won awards such as the Culture Section of the American Sociological Association's Mary Douglas Prize for Best Book (2008), and the American Publishers Award for Professional and Scholarly Excellence in Sociology and Social Work (2006).

Cultural Theories of Action and Materiality

Approaching his fourth decade as an academic, Alexander remains as productive and imaginative in his interests and writings as at any time during his career. Creating a macrodramaturgical theory of action, one that resists subsuming meaning to practice, in "Cultural Pragmatics: Social Performance Between Ritual and Strategy," Alexander (2004a) specified a conceptual framework designed for analyzing events in complex, differentiated societies. Alexander's argument offers a historical reconstruction of the development and professionalization of theater and criticism and illustrates how over time social formations have moved from fused conditions organized by rituals, to the conditions of defusion, which are organized by 'ritual-like,' performative processes (see also Alexander et al., 2006).

Demonstrating the theory's purchase, Alexander (2010) examines Barack Obama's first campaign for the presidency. Alexander argues that the challenge before those who aspire to power includes refusing the elements of performance, and performing meaningful scripts that engage the audience's understandings of national identity and democracy. Would-be leaders must 'work the binaries,' or stage performances in which they cast themselves as representative of the sacred, democratic code of civil society while casting their opponents as polluted, profane, counterdemocratic characters. They must also 'walk the boundaries' that separate the civil and uncivil spheres in contemporary American society; they must perform the delicate task of identifying themselves with more particularistic spheres like race, class, gender, religion, and ethnicity, while signaling that if they win office they are capable of containing these particularistic dimensions and assuming the more universalist persona of the ideal civil actor.

Alexander has also turned to iconicity to create a cultural theory of materiality (2008a,b; see also Alexander et al., 2011). Iconic consciousness, he states, "occurs when an aesthetically shaped materiality signifies social value" (2008: 782). Iconic experience, on the other hand, is triggered when an artist successfully shapes a surface in such a manner that it draws the viewer deeper into contact with iconic meaning. Alexander describes how in this process "the aesthetic object becomes a symbol" or collective representation, the viewer's concerns about the present or the conditions of the object's production recede, the object's specificity is replaced by a sense of "all 'such things,'" and inside the observer is stirred "a process of typification" (6). Shifting from the materiality of art to that of everyday life, Alexander asserts that people are "drawn into the experience of meaning and emotionality by surface forms," forms that "we 'feel' in our unconscious minds and associate with other ideas and things … ideas and things that are simultaneously personal and social" (6).

See also: Civil Society, Concept and History of; Democracy, History of; Democracy: Normative Theory; Durkheim, Emile (1858–1917); Geertz, Clifford (1926–2006); Hermeneutics; Modernization and Modernity in History; Multiculturalism; Parsons, Talcott (1902–79); Pluralism and Tolerance; Political Thought, History of; Public Sphere: Eighteenth-Century History; Public Sphere: Nineteenth and Twentieth-Century History; Social Movements, History of: General; Weber, Max (1864–1920).

Bibliography

Alexander, Jeffrey C., 1980. Core solidarity, ethnic outgroup, and social differentiation: a multidimensional model of inclusion in modern societies. In: Dofny, J., Akiwowo, A. (Eds.), National and Ethnic Movements. Sage, Beverly Hills, pp. 5–28.

Alexander, Jeffrey C., 1982. Theoretical Logic in Sociology. In: Positivism, Presuppositions, and Current Controversies, vol. 1. University of California Press and Routledge Kegan Paul, Berkeley, CA.
Alexander, Jeffrey C., 1984. Three models of culture and society relations: toward an analysis of watergate. Sociological Theory 3, 290–314.
Alexander, Jeffrey C., 1986a. Rethinking Durkheim's intellectual development, I: on 'Marxism' and the anxiety of being understood. International Sociology 1 (1), 91–107.
Alexander, Jeffrey C., 1986b. Rethinking Durkheim's intellectual development, II: working out a religious sociology. International Sociology 1 (2), 189–201.
Alexander, Jeffrey C., 1987. Twenty Lectures: Sociological Theory since World War II. Columbia University Press, New York.
Alexander, Jeffrey C., 1988a. Culture and political crisis: watergate and Durkheimian sociology. In: Durkheimian Sociology: Cultural Studies. Cambridge University Press, New York.
Alexander, Jeffrey C., 1988b. Durkheimian Sociology: Cultural Studies. Cambridge University Press, New York, pp. 1–21.
Alexander, Jeffrey C., 1988c. Introduction: Durkheimian sociology and cultural studies today. In: Durkheimian Sociology: Cultural Studies. Cambridge University Press, New York, pp. 1–21.
Alexander, Jeffrey C., 1990. Analytical debates: understanding the relative autonomy of culture. In: Alexander, J.C., Seidman, S. (Eds.), Culture and Society: Contemporary Debates. Cambridge University Press, New York, pp. 1–27.
Alexander, Jeffrey C., 1992. Citizen and enemy as symbolic classification: on the polarizing discourse of civil society. In: Fournier, M., Lamont, M. (Eds.), Cultivating Differences: Symbolic Boundaries and the Making of Inequality. Chicago University Press, IL, pp. 289–308.
Alexander, Jeffrey C., 1993. The promise of cultural sociology: technological discourse and the sacred and profane information machine. In: Smelser, N., Munch, K. (Eds.), Theory of Culture. University of California Press, Berkeley, CA, pp. 293–323.
Alexander, Jeffrey C., 1995. Modern, anti, post, and neo: how intellectuals have coded, narrated, and explained the 'New world of our time,'. In: Fin de Siècle Social Theory: Relativism, Reduction and the Problem of Reason. Verso, New York, pp. 6–64.
Alexander, Jeffrey C., 1996. Cultural sociology or sociology of culture? Towards a strong program. ASA Newsletter Culture 10 (3/4), 1, 3–5.
Alexander, Jeffrey C., 1998. Neofunctionalism and after. Basil Blackwell, Malden, MA.
Alexander, Jeffrey C., 2001a. Robust utopias and civil repairs. International Sociology 16 (4), 579–591.
Alexander, Jeffrey C., 2001b. The long and winding road: civil repair and intimate injustice. Sociological Theory 19 (3), 371–400.
Alexander, Jeffrey C., 2001c. Theorizing the 'modes of incorporation': assimilation, hyphenation, and multiculturalism as varieties of civil participation. Sociological Theory 19 (3), 237–249.
Alexander, Jeffrey C., 2002. On the social construction of moral universals: the 'holocaust' from mass murder to trauma drama. European Journal of Social Theory 5 (1), 5–86.
Alexander, Jeffrey C., 2004a. Cultural pragmatics: social performance between ritual and strategy. Sociological Theory 22 (4), 527–573.
Alexander, Jeffrey C., 2004b. Towards a theory of cultural trauma. In: Alexander, et al. (Eds.), Cultural Trauma and Collective Identity. University of California Press, Berkeley, CA, pp. 1–30.
Alexander, Jeffrey C., 2005. The sixties and me: from cultural revolution to cultural theory. In: Sica, A., Turner, S. (Eds.), The Disobedient Generation. University of Chicago Press, pp. 37–47.
Alexander, Jeffrey C., 2006. The Civil Sphere. Oxford University Press, New York.
Alexander, Jeffrey C., 2008a. Iconic experience in art and life: beginning with Giacometti's 'standing woman'. Theory, Culture and Society 25 (5), 1–19.
Alexander, Jeffrey C., 2008b. Iconic consciousness: the material feeling of meaning. Environment and Planning D 26, 782–794.
Alexander, Jeffrey C., 2010. The Performance of Politics: Obama's Victory and the Democratic Struggle for Power. Oxford University Press, New York.
Alexander, Jeffrey C., 2013. The Darkside of Modernity. Polity, Malden, MA.
Alexander, Jeffrey C., Bartmanski, Dominik, Giesen, Bernhard, 2011. Iconic Power: Materiality and Meaning in Social Life. Palgrave Macmillan, New York.
Alexander, Jeffrey C., Giesen, Bernhard, Mast, Jason L., 2006. Social Performance: Symbolic Action, Cultural Pragmatics, and Ritual. Cambridge University Press, Cambridge.
Alexander, Jeffrey C., Smith, Philip, 1993. The discourse of American civil society: a new proposal for cultural studies. Theory and Society 22 (2), 151–207.
Alexander, Jeffrey C., Smith, Philip, 2002. The Strong program in cultural theory: elements of a structural hermeneutics. In: Turner, J. (Ed.), Handbook of Sociological Theory. Kluwer Academic, New York.

Relevant Websites

http://ccs.research.yale.edu/.
http://sociology.berkeley.edu/jeffrey-c-alexander-1969.
http://sociology.yale.edu/people/jeffrey-alexander.

Algorithm

Laura Martignon, PH Ludwigsburg University of Education, Ludwigsburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

An algorithm is an instruction formulated as a finite set of rules to be performed sequentially or in parallel for obtaining a solution or an approximation to the solution of a well-specified problem. It may contain subroutines that are algorithms in themselves. An algorithm can be strictly deterministic or stochastic in nature. Fundamental features of an algorithm are its complexity and its amenability to implementation.

Introduction

An algorithm can be seen as a mathematical recipe, consisting of a finite set of rules to be performed systematically that is an outcome to the solution or an approximation of the solution of a well-formulated problem. The word 'algorithm' stems from the name of the Persian mathematician Abu Ja'far Mohammed ibnMûsâ al-Khowârizm, who lived in the ninth century. Around the year 1825 al-Khowârizm wrote an influential mathematical textbook entitled *Kitab al jabrw'almuqabala*. The term 'algorithm,' or in its more original form 'algorism,' is directly derived from the last part of the author's name.

In today's scientific usage algorithms can be sequential or parallel. In sequential algorithms the steps forming the algorithm are ordered and should be performed one after the other. In parallel algorithms some of the rules are to be performed simultaneously. Algorithms can be graphically represented by flowcharts composed by arrows and boxes. The boxes contain the instructions and the arrows indicate transitions from one step to the next. Algorithms were known much earlier than the eighth century. One of the most familiar, dating from ancient Greek times (*c*.300 BC), is the procedure now referred to as Euclid's algorithm for finding the highest common factor of two numbers.

Algorithms often depend on subalgorithms that can be seen as instructions pertaining to subtasks to be tackled as parts of the process. For instance, the algorithm for obtaining the highest common factor of two numbers relies on the algorithm for finding the remainder obtained when dividing two numbers a and b. Dividing two (natural) numbers a and b, is something we learn at school after having learnt the multiplication tables by heart. Usually we perform a division by reducing it to a sequence of multiplications and subtractions. Yet assume for the moment that to perform division we cannot rely on multiplication, nor can we express numbers in base 10. Both of these operations will require additional algorithms. We represent numbers in a very primitive form by simply writing a sequence of dots. Thus •••••, for instance, represents the number 5.

The remainder algorithm works as follows (using $17 \div 5$ as an example):

(a) Write the sequence of dots for the dividend (17 in this case).
(b) Erase as many dots as correspond to the divisor (5 in this case).
(c) If the remaining number of dots is larger than or equal to the divisor go back to (a).
(d) If the remaining number of dots is smaller than the dividend print out this number.
(e) STOP.

Figure 1 represents the flow chart of the algorithm 'find the greatest common divisor' of a and b. The algorithm just described contains a loop. Observe that for any pair of numbers the algorithm produces the answer in a finite number of steps. Applied to our example of $17 \div 5$, the algorithm performs the following steps:

Current state	*Operation*
START	Write 17 dots
••••••••••••••••••	Erase 5 dots
••••••••••••	Erase 5 dots
•••••••	Erase 5 dots
••	Print 2 dots
STOP	

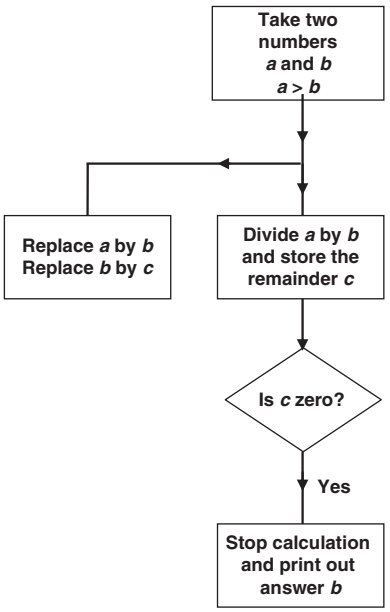

Figure 1 Flow chart of the algorithm 'find the greatest common divisor' of a and b.

The computational algorithm for finding the remainder of an integer when divided by another integer can be used as a subalgorithm of the decision algorithm for the problem 'Does *b* divide *a*?' (the answer is 'yes' if the remainder is zero). Repeated application of these algorithms produces the answer to the decidable question 'Is *a* a prime?' (the answer is 'no' if *a* is divisible by any smaller number besides 1).

An algorithm or a machine is deterministic if at each step there is only one possible action it can perform. A nondeterministic algorithm or machine may make random choices of its next action at some steps. An algorithm is called a decision algorithm if it leads to a 'yes' or 'no' result, whereas it is called a computational algorithm if it computes a solution to a given well-defined problem.

Despite the ancient origins of specific examples of algorithms, the precise formulation of the concept of a general algorithm dates only from the last century. The first rigorous definitions of this concept arose in the 1930s. The classical prototype algorithm is the Turing machine, defined by Alan Turing to tackle the *Entscheidungsproblem* or decision problem, posed by the German mathematician David Hilbert in 1900, at the Paris International Congress of Mathematicians. Hilbert's dream was to prove that the edifice of mathematics is a consistent set of propositions derived from a finite set of axioms, from which the truth of any well-formulated proposition can be established by a well-defined finite sequence of proof steps. The development and formalization of mathematics had led mathematicians to see it as the perfect, flawless science.

Algorithms and the *Entscheidungsproblem*

In 1931, the foundation of mathematics suffered its most crushing blow from a startling theorem proven by the Austrian logician Kurt Gödel. Gödel showed that any mathematical system powerful enough to represent arithmetic is incomplete in the sense that there exist propositions that cannot be proven true or untrue in a finite sequence of steps. Such propositions are said to be undecidable within the given system. Turing had been motivated by Gödel's work to seek an algorithmic method of determining whether any given proposition was undecidable, with the ultimate goal of removing undecidability as a concern for mathematics. Instead, he proved in his seminal paper 'On computable numbers, with an application to the *Entscheidungsproblem*'(1937) that there cannot exist any such universal method of determination.

The question of establishing whether the number of steps required for a given problem is finite or infinite is called the halting problem. Turing's description of the essential features of any general-purpose algorithm, or Turing machine, became the foundation of computer science. Today the issues of decidability and computability are central to the design of a computer program – a special type of algorithm – and are investigated in theoretical computer science.

The question whether intelligent problem solving can be described in terms of algorithms was extensively examined by Herbert Simon in the late 1940s and early 1950s. Newell and Simon proposed the first computer programs for problem-solving algorithms as well as the first programs for algorithms that prove theorems in Euclidean geometry, thus founding the new discipline of Artificial Intelligence (*see* Artificial Intelligence: Genetic Programming).

Around the same time McCulloch had developed a formal model of a neuron (McCulloch and Pitts, 1943), proving that artificial neurons are capable of performing logical operations. An artificial neuron is a device that produces an output that is a function of its inputs if the sum of the inputs exceeds a threshold and otherwise produces no output. The subdiscipline of computer science known as neural networks deals with systems of artificial neurons firing in sequence and/or in parallel, in analogy to the operation of biological neurons.

The Complexity of an Algorithm

One important feature of an algorithm is its complexity. A number of definitions of complexity have been put forward with the most common being time complexity, or the length of time it takes for an algorithm to be executed. Clearly, algorithms with low time complexity are to be preferred to ones with higher time complexity that solve the same problem. The question of establishing a formal definition of complexity was answered and treated formally in theoretical computer science (*see* Algorithmic Complexity).

One possibility is to count the number of operational steps in an algorithm, express this number as a function of the number of free parameters involved in the algorithm, and determine the order of complexity of this function. The order of complexity of a function f is denoted by O(f), where O( ) is usually called the Landau symbol, and is defined as follows: Given two functions $F(n)$ and $G(n)$ defined on the set of natural numbers, we say that F is of the order of G, and write $F = O(G)$, if there exists a constant K such that:

$$\frac{F(n)}{G(n)} < K$$

for all natural numbers n. Thus, for instance, the function $F(n) = 3n+1$ is of the order of $G(n) = n$ *with* $K = 4$. Every polynomial in n, that is, every linear combination of powers of n, is of the order of its highest power of n. Because what is being counted is the number of steps in a sequential process, which is common to view the resulting O( ) criterion as the time complexity of the algorithm, where n denotes the length of the given input.

The notion of algorithm complexity led to a fundamental classification of problems. They are classified as belonging to P or NP, where, as we shall see, this 'or' is not exclusive. P is a shorthand for 'polynomial time' and stands for the set of all problems that can be solved by a deterministic algorithm in polynomial time, which means that the number of operations required by the algorithm can be expressed in terms of a linear combination of powers of n, where n is the number of free parameters. NP is the class of all problems that can be solved by nondeterministic algorithms in polynomial time. The strength of nondeterminism lies precisely in the machine's freedom to search through a large space of possible computations by means of a nondeterministic or stochastic process. Clearly, the relation $P \subseteq NP$ holds. One of the unsolved

problems in computer science is to establish whether P = NP (Garey and Johnson, 1979). Problems for which low complexity algorithms exist (i.e., problems in class P) are called tractable; other problems are called intractable.

Monte Carlo Methods

An important category of nondeterministic algorithms is the class of Monte Carlo methods, introduced around 1940 by Metropolis, von Neumann, and Ulam. The term derives from the gambling games that are an economic mainstay of the city of that name, and originated during World War II as a code name for stochastic simulations associated with nuclear research. The most common application of Monte Carlo methods is to approximate intractable integrals. If $X_1, \ldots, X_n$ are random draws from a probability distribution with density function $f(x)$, then $S_n = \frac{1}{n}\sum_i h(X_i)$ has an expected value,

$$E[S_n] = \int_x h(x)f(x)\mathrm{d}x \qquad [1]$$

and satisfies the central limit theorem. Thus, a difficult integral can be approximated by representing it in the form (1) and executing the following nondeterministic algorithm:

(a) Simulate a random number X_n from the distribution $f(x)$.
(b) Compute the average $S_n = \frac{1}{n}\sum_{i=1}^{n} h(x_i)$.
(c) Estimate whether sufficient accuracy has been achieved.
(d) If yes, STOP; otherwise return to (a).

A common strategy for increasing accuracy in a problem that does not admit exact solutions is to decompose it into subproblems such that the exact algorithms for tractable subproblems can be combined with Monte Carlo estimates for intractable subproblems to obtain accurate approximations to the whole problem. Monte Carlo simulations are important tools in fields such as physics, statistical analysis, stochastic neural networks, and optimization.

Algorithms for Learning

Since the inception of formal algorithms, an important question has dealt with the conditions under which algorithms can be designed to learn from experience. The idea is that an algorithm that learns from experience becomes more and more adequate for solving the problem in question. The last two decades of the twentieth century witnessed an explosion of algorithms that can perform learning tasks and a sound theoretical understanding of learning machines is beginning to emerge. The theoretical and empirical study of learning algorithms is called Machine Learning. One type of learning machine is typical of neural networks that modify their weights by adapting them to feedback information on their performance. Another type of learning machine is provided by what is known as a 'genetic algorithm,' where there is a natural selection between algorithmic procedures and the most effective algorithm arises by a form of survival of the fittest.

Yet another type of learning can be materialized as a search through a large set of possible algorithms for a given problem (*see* Algorithmic Complexity). Here the type of algorithm is fixed from the beginning and what is searched for is the optimal realization of this algorithm, that is, a realization that fits the given data best. Thus learning becomes a special form of model selection. For many problems, including search across models, exact solutions are intractable yet tractable heuristics (or rules-of-thumb) work well for a large class of problem instances. As an example, consider the problem of finding an optimal decision tree algorithm to represent a given decision rule. A decision tree is a graphical representation of a rule for making a categorization decision. The graph for a decision tree consists of nodes and edges pointing from nodes (called parents) to other nodes (called children). One node, called the root node, has no parents. Each other node has exactly one parent. At the bottom of the tree are the leaf nodes, which have no children. With each nonleaf node is associated a rule which selects a child to visit depending on information passed from the parent node.

With each leaf node is associated a rule for computing a value for the node. The decision tree algorithm proceeds as follows:

(a) Begin at root node.
(b) Execute rule to decide which arc to traverse.
(c) Proceed to child at end of chosen arc.
(d) If child is a leaf node, execute rule to compute and print value associated with node and STOP.
(e) Otherwise, go to (b).

It is frequently desired to find an optimal decision tree for a given classification problem and a given criterion of optimality. There are exponentially many possible decision trees for a set of n pieces of information and thus, searching for optimal trees is often an intractable problem. Tractable heuristics have been suggested in machine learning that help find trees good enough for given problems. One such type of classification tree has been introduced by Breiman et al. (1993) and is known as CART, which is an abbreviation of classification and regression tree.

The study of algorithms has been highly influential in psychology. The behaviorist view which emphasizes the relationship between sensory inputs and motor outputs has given way to the cognitive view, which emphasizes the role of internal cognitive states in producing behavior. Information processing models of cognition have thus become a mainstay of cognitive psychology. Hypotheses about the manner in which cognition influences behavior are encoded as algorithms and implemented in a computer software that is used to simulate behavior. Simulated results are then compared to the behavior of human or animal subjects to validate or refute the hypotheses upon which they are based. One of the important issues in recent developments of cognitive psychology has been to establish how the mind deals with classifications and decisions under uncertainty.

Simple Algorithms for Classification Previous to Decision Making: Fast and Frugal Trees

While some schools in the psychology of decision making have defended the approach of classical rationality for which

the mind functions by means of full logical and probabilistic algorithms, a recent development has viewed the unaided mind as relying on very simple heuristics for inference, which are amenable to being modeled by means of algorithms. These *fast and frugal* algorithms, as proposed by Gigerenzer et al. (1999), are actually elementary, fast in construction and execution, and robust in prediction. The most useful among them are the fast and frugal classification trees (*see* Decision and Choice: Heuristics).

Decision making often requires previous classification: Possible actions are assigned to options or alternatives, which emerge as the result of classification. Patients are classified as having or not having a disease by means of cues, which may be symptoms or the result of tests; based on these classifications, decisions are made on possible actions to be taken, like surgery or intake of medicine. Two questions arise: (1) What are the properties of good cues for this type of classification that underlies decision making? (2) What are the algorithms used for this type of classification? In order to answer the first question a concrete example from medical decision making can be useful. In the case of one cue (e.g., a medical test like the mammogram), its *sensitivity* is given by the proportion of patients with the disease (e.g., breast cancer) that tests positive. This measure is important for establishing the goodness of the test but is not sufficient for making decisions on possible treatments, given that the patient is classified as having the disease. What the doctor definitely needs to know is the positive predictive value, also called the validity of the test, assessed as the proportion of patients with a positive value on it, who do have the disease. As has been demonstrated empirically (Gigerenzer and Hoffrage, 1995; Brase, 2008) *natural* representations of information foster good assessments of the predictive value of a test based on knowledge on the sensitivity and vice versa, by facilitating Bayesian reasoning. A positive mammogram, though, is not sufficient for classifying a patient as having breast cancer. It is just the first of a set of tests. Other tests, like ultrasound test and tumor markers and even ambulant biopsies, belong to the medical routine for classifying a patient as requiring global mastectomy or some alternative local treatment.

The answer to the second question is more subtle and there is a plethora of models proposed by both statistics and artificial intelligence for classification. A survey of the literature on these models for classification can be found, for instance, in Berretty et al. (1999). Bayesian classification becomes cumbersome already for three cues and is computationally intractable when several cues have to be taken into account. Even simplifications of the full Bayesian procedure based on natural frequencies become utterly complicated when several cues are considered. There exist more synthetic Bayesian procedures, such as Bayesian networks, but these are also NP-complete (Cooper, 1990).

How do people perform binary classifications based on several cues, when in hurry and when both time and information are costly? A concrete example may illustrate this situation: how do doctors establish, for instance, that a patient should or should not have surgery, based on the results of a set of tests? Or, more concretely, how do doctors classify a patient with severe chest pain as one that should be treated in the Coronary Care Unit? Two researchers of the University of Michigan P. Green and A. Mehr (Green and Mehr, 1997) investigated this question from an innovative perspective that caused a paradigm change in the theory of human classifications and decision making. This is the situation they analyzed: in a Michigan hospital, doctors used to rely on their clinical judgment in order to decide whether a patient with intense chest pain has to be assigned to the coronary unit, right away, or whether he should instead rest in a regular hospital bed for treatment. The doctors ended up sending about 90% of the patients to the coronary care unit. This 'defensive' strategy led to excessive costs, reduced the quality of care provided and introduced new risks. The two researchers from the University of Michigan were called to solve this problem.

Green and Mehr investigated the effect of an intervention, in the course of which physicians were trained to use logistic regression rather than to rely on their clinical judgment (Green and Mehr, 1997). The *Heart Disease Predictive Instrument* (Pozen et al., 1984) consisted of a chart with more than 50 probabilities and a logistic regression formula with which the physician can compute the probability that a patient has acute ischemic heart disease. If the probability is higher than a critical value, then the patient is classified as being in a high risk and sent to the coronary care unit, otherwise not. But physicians were not happy using this and similar systems. They did not understand it. It is not transparent and appears to suggest that clinical judgment should depend on a pocket calculator. The physicians seemed to be using some of the information in the charts – such as what symptoms are diagnostic and which symptoms are more diagnostic than others – but ignored the quantitative aspects of the information in terms of exact probabilities. That opened up a third alternative: the probabilities and computations of the logistic regression may not matter much at all for achieving high accuracy. Based on this observation Green and Mehr (1997) constructed a heuristic by using three building blocks: (1) ordered search, (2) fast stopping rule, and (3) one-reason decision making. The resulting classification heuristic is shown in **Figure 2**. It asks only a few yes-or-no questions and has an exit at every level. If a patient has a certain anomaly in his electrocardiogram – the so-called ST segment change – he is immediately classified as being at a high risk and is assigned to the coronary care unit. No other information is searched for. If that is not the case, a second question is asked: is the patient's primary complaint chest pain? If this is not the case, he is classified as low risk and put to a regular nursing bed. No further information is considered. If the answer is yes, a final question, namely whether any of those four remaining cues not associated with chest pain, is asked to classify the patient and decide on the appropriate treatment.

A surprising result concerning the performance of these fast and frugal trees is that these trees are overall more accurate in classifying actual heart attack patients than both the physicians' intuition and the logistic regression. The fast and frugal tree maximizes the proportion of patients correctly assigned to the coronary care (as measured by a subsequent myocardial infarction). At the same time it has a comparatively low false alarm rate.

Fast and frugal trees have been proposed for a variety of classification tasks (Martignon et al., 2008). Another interesting

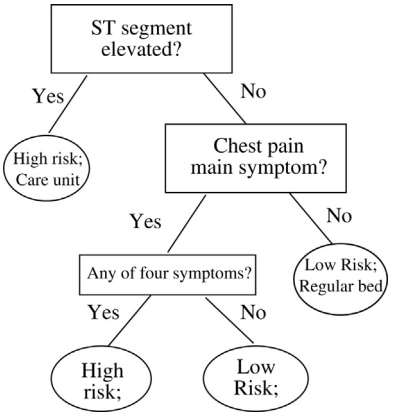

Figure 2 A fast and frugal classification tree: it classifies patients with complains of heart pain into *high risk* or *low risk*.

example is, for instance, deciding whether, or not, to prescribe antibiotic treatment to young children suffering form community acquired pneumonia (Fischer et al., 2002). Fischer et al. (2002) proposed a very simple fast and frugal tree for making this decision. The tree uses at most two cues, whose values can be easily obtained. Initially, it is asked whether the child is older than 3 years. If the answer to this is 'no,' it is immediately concluded that macrolides should not be prescribed. Only if the answer to the first question is 'yes,' is it asked in the second level of the tree whether the child has had fever for more than 2 days. If the answer to the second question is 'no,' it is concluded that macrolides should not be prescribed and if the answer is 'yes,' it is concluded that macrolides should be prescribed. Interestingly, this tree can also be expressed as a rule easy to memorize and quick to apply: 'Prescribe macrolides only if the child is older than three years and the child has had fever for more than 2 days.' Additionally, the fast and frugal tree does not sacrifice accuracy. When evaluated and double-checked on real data, it is classified as 'high risk of microstreptococal pneumonia infection' for 72% of those children who actually were at high risk, while a scoring system based on logistic regression identified 75% of them.

Conclusion

Algorithms are the core element of thinking devices. The last century witnessed the birth of highly intelligent algorithms that replicate, often with great accuracy, some of the important achievements of the human mind. Yet, one of the most important discoveries concerning algorithms is that mathematics cannot be produced by an algorithmic procedure. This discovery, which was experienced as a crisis, can be viewed as liberating because it demonstrates the power of mathematics. Replicating intellectual, symbolic, and even basic common-sense activities by means of algorithms turned out to be daunting tasks, still mostly exceeding the grasp of computable representations.

The strategies that humans use for classification and decision are simple heuristics that can be modeled by means of very simple algorithms.

See also: Algorithmic Complexity; Artificial Intelligence: Combinatorial Search; Decision and Choice: Heuristics; Mathematical Psychology.

Bibliography

Berretty, P.M., Todd, P.M., Martignon, L., 1999. Categorization by elimination: using few cues to choose. In: Gigerenzer, G., Todd, P.M., The ABC research group (Eds.), Simple Heuristics That Make Us Smart. Oxford University Press, New York, pp. 235–254.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1993. Classification and Regression Trees. Chapman and Hall, New York.

Brase, G.L., 2008. Pictorial representations in statistical reasoning. Online verfügbar unter http://www.k-state.edu/psych/research/documents/2009ACP.pdf (accessed 16.12.11.).

Cooper, G.F., 1990. The computational complexity of probabilistic inference using Bayesian belief networks. Artificial Intelligence 42 (2–3), 393–405.

Fischer, J.E., Steiner, F., Zucol, F., Berger, C., Martignon, L., Bossart, W., Altwegg, M., Nadal, D., 2002. Using simple heuristics to target macrolide prescription in children with community-acquired pneumonia. Archives of Pediatric and Adolescent Medicine 156 (10), 1005–1008.

Garey, M.R., Johnson, D.S., 1979. Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman, San Francisco.

Gigerenzer, G., Hoffrage, U., 1995. How to improve Bayesian reasoning without instruction: frequency formats. Psychological Review 102 (4), 684–704.

Gigerenzer, G., Todd, P., The ABC Group, 1999. Simple Heuristics That Make Us Smart. Oxford University Press, New York.

Green, L., Mehr, D.R., 1997. What alters physicians' decisions to admit to the coronary care unit? Journal of Family Practice 45 (3), 219–226.

McCulloch, W.S., Pitts, W.H., 1943. A logical calculus of the idea immanent in nervous activity. Bulletin of Mathematical Biophysics 5, 115–133. Reprinted in: McCulloch, W.S., 1965. Embodiments of Mind. MIT Press, MA.

Martignon, L., Katsikopoulos, K., Woike, J., 2008. Categorization with limited resources: a family of simple heuristics. Journal of Mathematical Psychology 52, 352–361.

Pozen, M.W., D'Agostino, R.B., Selker, H.P., Sytkowski, P.A., Hood Jr., W.B., 1984. A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease. New England Journal of Medicine 310, 1273–1278.

Turing, A., 1937. On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society (Ser. 2) 42, 230–265; 43, 544–546.

Algorithmic Complexity

Peter Sunehag and Marcus Hutter, Research School of Computer Science, Australian National University, Canberra, ACT, Australia

© 2015 Marcus Hutter. Published by Elsevier Ltd. All rights reserved.

Abstract

Algorithmic complexity provides a mathematical formal notion of string complexity. Building on this, one arrives at mathematical 'gold standard' (though incomputable) definitions of randomness, induction, similarity, and even intelligence. These definitions can be turned into practical algorithms by using common compressors to approximate the universal solutions. One can consider the theories as idealized cognition with respect to which one can aim to describe actual biological cognition by listing biases and limitations that need to be defined relative to some normative reference.

Introduction

Our world contains ubiquitous examples of phenomena relying on compact representations. Some examples are the genome (Krakauer, 2002), human cognition (Chater, 1996, 1999; Chater and Vitányi, 2003), and natural languages (Zipf, 1949; Piantadosi et al., 2011) which all involve efficient use of resources. They all involve a compact code and a method of translating the code into a result. The field of algorithmic complexity (Solomonoff, 1964; Kolmogorov, 1965; Chaitin, 1969), sometimes called Kolmogorov complexity (Li and Vitányi, 2008) or Algorithmic Information Theory (Hutter, 2007a), provides a mathematical theory for such phenomena and also answers some fundamental theoretical questions about randomness (Martin-Löf, 1966) as well as telling us how to ideally make predictions about the future (Solomonoff, 1964; Hutter, 2007b; Rathmanner and Hutter, 2011). The translation mechanism mentioned above is called the reference machine and is formalized as a universal Turing machine (UTM) (Turing, 1936). The codes are programs for this machine in the form of a binary string. The result is an output of another binary string. The length of the shortest program that leads to a certain output string is called the Kolmogorov complexity of this string and can be interpreted as a measure of information content as well as of simplicity. A string with lower Kolmogorov complexity than another is considered simpler with respect to the used reference machine since it can be described more succinctly. If a string can be represented in a substantially shorter form than its naive representation, it is called compressible, otherwise incompressible. Incompressible strings look random and incompressibility is now the established definition for what it means for an individual sequence to be random. Furthermore, algorithmic complexity provides a way of defining a priori probabilities (Solomonoff, 1964) for different strings, where simpler strings are deemed more likely. From these a priori probabilities, one can derive conditional probabilities for the future as observations are made. These predictions will converge toward the true probabilities under mild assumptions on the generating process (Solomonoff, 1978; Hutter, 2007b). The assigned probabilities depend on the choice of reference machine since the Kolmogorov complexity differs by up to an additive constant for any pair of machines. This constant can make a big difference initially while the effect vanishes asymptotically. Algorithmic probability has also been combined with reinforcement learning and sequential decision theory to define an intelligent agent denoted AIXI, thereby introducing the field of universal artificial intelligence (Hutter, 2005). The agent follows a policy that maximizes expected utility with respect to the algorithmic probabilities and the utilities are defined as a sum of discounted rewards.

Algorithmic complexity has the drawback of being incomputable, though successful algorithms have been devised using practical compressors to compute crude approximations of the true complexity (Li et al., 2004; Cilibrasi and Vitányi, 2005; Veness et al., 2011). We will mention these as we outline the basic theory in the next section and then we discuss the possibility of using algorithmic complexity to provide an idealized theory of cognition. It has been argued that this model of learning resolves the 'poverty of stimulus' problem with human language acquisition (Solomonoff, 1964; Perfors et al., 2006; Hsu et al., 2013). The article ends with a Section Summary.

Basic Definitions and Results

In this section, we present the formal definitions and basic results in algorithmic complexity, algorithmic randomness, algorithmic probability, and universal artificial intelligence. We refer to Hutter (2005) and Li and Vitányi (2008) for comprehensive studies.

Algorithmic Complexity

Given a Universal Turing Machine (UTM) U, we define the Kolmogorov complexity $K(x)$ of a string x to be the length of the shortest program that makes U output x. Formally

$$K(x) := \min\{\ell(p) : U(p) = x\}$$

where $\ell(p)$ is the length of p and $U(p) = x$ means that U outputs x when given program p on the input tape. That U outputs x can be given several precise meanings. For example, it can mean that x is printed on the output tape and then the machine halts or that x is first printed and then the machine continues. K is defined by demanding that U halts after x is printed but the other mentioned variant as well as several others differs from K by at most $O(\log K)$. We here state all properties (equalities and inequalities) only up to such

a logarithmic term and they, therefore, are true for a whole range of variations on K.

We also define the conditional complexity $K(x|y)$ which is the length of the shortest program to output x given y as side information. For example, if y is a string that only differs from x by missing the last bit, then the program only has to add that bit and such a program is very short even if x is a long complex string by itself. $K(x|y)$ tells us how much information there is in x that is not in y. Furthermore, we define the joint complexity $K(x, y)$ as the length of the shortest program to output first x and then y. This can be much shorter than the sum of the shortest programs to output the individual strings since they can have much in common. It can, however, not be much longer than the sum of those since one can always create a program that first runs p_1 that outputs x and then runs p_2 that outputs y. The extra length is the length of the extra code needed for the command to first run one program and then another and is at most a constant that does not depend on x, y, p_1, or p_2. A profound result that connects all the mentioned concepts is

$$K(x,y) = K(x) + K(y|x) = K(y) + K(x|y)$$

where $=$ should, as everywhere in this article, be interpreted as being up to a logarithm and not as exact equality. This result can be rewritten as a property that is called 'symmetry of information,' stating that

$$K(x) - K(x|y) = K(y) - K(y|x)$$

This relation informally means that the amount of information y has about x ($K(x)-K(x|y)$) is the same as what x has about y. The involved quantity provides one measure of how similar they are but it is not formally a distance. However, one can transform it to

$$d(x,y) := 1 - \frac{K(x) - K(x|y)}{K(x,y)}$$

which satisfies the distance measure properties and still says that two strings are similar if they have a large amount of mutual information. Another alternative to defining a distance is $D(x, y)$ which is the length of the shortest program that converts x to y and y to x. It can be proven (Bennett et al., 1998; Mahmud, 2009) that within a constant term

$$D(x,y) := \max\{K(x|y), K(y|x)\}$$

which means that it is about as hard converting both ways as the hardest of the two directions. Both d and D are universal distance functions in the sense that if two strings are close for any computable distance function (satisfying some mild properties) then they are also close as measured by d or D. In other words, d and D pick up any possible computable regularity between any two given strings.

A normalized version of D called the normalized compression distance

$$\tilde{D}(x,y) := \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}$$

was introduced by (Li et al., 2004) where it was also proven to be universal. Furthermore, Li et al. (2004) introduced practical approximations using common compressors, namely they defined the normalized compression distance by

$$\mathrm{NCD}(x,y) := \frac{Z(xy) - \min\{Z(x), Z(y)\}}{\max\{Z(x), Z(y)\}}$$

where $Z(x)$ is the size of the compression of x. By using NCD as a similarity measure, Li et al. (2004) inferred an evolutionary tree from DNA sequences as well as the language tree from the text of the UN declaration of human rights in 52 languages. Other applications include clustering of music as well as finding the phylogenetic tree of chain letters (Bennett et al., 2013; Cilibrasi and Vitányi, 2005). This method of clustering is sometimes referred to as algorithmic clustering or clustering by compression.

Algorithmic Randomness

It is clear that $K(x) \leq \ell(x)$, where $\leq$ is up to a logarithm as always (though here true for a constant) and $\ell(x)$ is the length of the string, since one can write a program that says Print(x). The extra constant length is the Print command. If there is no shorter program than this, i.e., $K(x) \geq \ell(x)$ we say that x is incompressible and random. The notion of randomness based on incompressibility is also called Martin-Löf randomness (Martin-Löf, 1966) and has been found to be similar to human perception of randomness (Griffiths and Tenenbaum, 2003). Intuitively something is random if no regularities can be found in the string.

Algorithmic Probability

Based on a chosen reference machine U, the algorithmic a priori probability of a string x is

$$M(x) := \sum_{p:U(p)=x} 2^{-l(p)}$$

which can be interpreted as the probability that the string x is the output from U if zeros and ones are placed by independent balanced coin flips on the input tape. The sum of terms $2^{-l(p)}$ for all programs p (in a prefix free set of programs) is upper bounded by 1, according to Kraft's inequality. They do not sum to exactly one and M is actually just a semimeasure though it can be renormalized to a proper probability measure. Given any string x, one would then say that the probability of the next bit being a one is

$$M(1|x) := \frac{M(x_1)}{M(x_0) + M(x_1)}$$

$M(x)$ is closely approximated by $2^{-K(x)}$. This is because much of the mass of $M(x)$ comes from the shortest program. Furthermore, one can prove (Li and Vitányi, 2008) that if one samples a string from M, one will with high probability get a string of low complexity. This is in accordance with the principle of favoring simplicity, which is often called Occam's razor. Algorithmic complexity and algorithmic probability provide a mathematical formalization of this principle.

Sequence Prediction

Ray Solomonoff developed a theory of sequence prediction using M, including his 'Prediction Error Theorem' (Solomonoff, 1978) stating that

$$\sum_{x\in\{0,1\}^*} \mu(x)(M(0|x)-\mu(0|x))^2 \le K(\mu)\frac{\ln(2)}{2},$$

for any computable measure μ. This implies, by rewriting the left-hand-side as $\sum_{j=0}^{\infty}\sum_{\ell(x)=j}\mu(x)(M(0|x)-\mu(0|x))^2$, that with μ-probability 1

$$M(x_n|x_1x_2\ldots x_{n-1})\to\mu(x_n|x_1x_2\ldots x_{n-1})$$

as $n\to\infty$ and that using M we can learn any computable measure. How long it takes depends on the complexity of the true measure.

Induction

If one is interested in finding the explanation for the data and not only about predicting the future one needs to choose a program. The measure M can be understood as a mixture of all computable sequences by choosing prior probability $\Pr(p)=2^{-l(p)}$ for the program p generating the sequence. The minimum message length (Wallace and Boulton, 1968) (or minimum description length (Rissanen, 1978)) approach chooses the program

$$p^* := \operatorname{argmin}\{\ell(p) : U(p) = x\}$$

This is the least complex hypothesis that fits the observed data. The choice can also be understood as maximizing the posterior probability $\Pr(p|d)=\Pr(d|p)\Pr(p)/\Pr(d)$. This latter expression also extends to having a prior over classes of stochastic hypothesis where a trade-off between how well data fit the hypothesis (i.e., $\log\Pr(d|\nu)$) and the prior probability $\log\Pr(\nu)$ (that can be based on algorithmic complexity (Hutter, 2005; Li and Vitányi, 2008)) is made by choosing

$$\nu^* := \arg_\nu\min\{-(\log\Pr(d|\nu)+\log\Pr(\nu))\}$$

Intelligent Agents

In this section, we consider an agent (Russell and Norvig, 2010; Hutter, 2005) that interacts with an environment through performing actions a_t from a finite set $\mathscr{A}$ and receives observations o_t from a finite set $\mathscr{O}$ and rewards r_t from a finite set $\mathscr{R}\subset[0,1]$. Let $\mathscr{H}=(\mathscr{A}\times\mathscr{O}\times\mathscr{R})^*$ be the set of histories. A function $\nu:\mathscr{H}\times\mathscr{A}\to\mathscr{O}\times\mathscr{R}$ is called a deterministic environment. It is called a stochastic environment if instead the resulting (o, r) is stochastic. A function $\pi:\mathscr{H}\to\mathscr{A}$ is called a (deterministic) policy or an agent. The value function V is defined based on geometric discounting by $V_\nu^\pi(h_{t-1})=\sum_{i=t}^{\infty}\gamma^{i-t}r_i$ where the sequence r_i are the rewards achieved by following π from time step t onward in the environment ν after having seen $h_{t-1}\in\mathscr{H}$. If the environment is stochastic we define V using an expectation.

Given a countable class of environments $\mathscr{M}$ and strictly positive prior weights w_ν for all $\nu\in\mathscr{M}$, the a priori environment ξ is defined by letting $\xi(\cdot)=\sum w_\nu\nu(\cdot)$ and a rational agent (Hutter, 2005) is defined by following the policy

$$\pi^* := \arg_\pi\max V_\xi^\pi(\varepsilon)$$

where ε is the empty initial history. In particular the AIXI agent (Hutter, 2005) is based on the same construction of a Solomonoff prior M as in the sequence prediction case. Though the agent is incomputable, practical approximations exist, e.g., MC-AIXI-CTW (Veness et al., 2011).

Cognition and Algorithmic Complexity

Human cognition is naturally very complex as it is an artifact of evolution. However, it is still meaningful to try to identify some of the fundamental principles that are involved with some consistency. After all, cognition operates with limited resources and has evolved to support survival. Cognitive science has discovered many systematic biases (Kahneman et al., 1982) but a question is biases from what ideal. The ideal is usually assumed to be rationality (Chater and Oaksford, 1999; Tenenbaum et al., 2011) in the decision theoretic sense of the word, i.e., expected utility maximization with respect to some of the priori probabilities. Another general principle is a preference for simplicity (Chater, 1996, 1999; Chater and Vitányi, 2003). Starting with rationality and a preference for simplicity, algorithmic complexity, as we have described in Section Basic Definitions and Results, provides a formal 'gold standard' theory for sequence prediction, choosing hypotheses, judging similarity, deciding what is random and acting rationally.

Language Acquisition

Noah Chomsky has famously argued (Chomsky, 2005) that the 'poverty of stimulus' makes the language learning observed in children impossible without an innate grammar. The argument is centered around the lack of negative examples. It has been argued that a rational simplicity-biased approach, formalized using information theory, can in fact perform this task (Solomonoff, 1964; Perfors et al., 2006; Hsu et al., 2013). This simplicity bias is replacing the need for an innate grammar but can possibly be viewed as a soft version of the same, i.e., a preference for certain syntax, based on what is judged to be simple given the reference machine.

Sequential Decision Making

The AIXI agent (Hutter, 2005) is formalizing the same combination of simplicity and rationality as in the sequence prediction setting for the sequential decision-making setting. The sequential decision setting, where you make a sequence of decisions that does affect the environment, is far more complicated than performing inference from gathered data. In the sequential setting you are gathering data as you act in the environment and to gather useful data is important as well as achieving high immediate rewards. The data you gather at one time point allows you to make better decisions later on and receive more reward then. This is known as the exploration-exploitation dilemma. A common method to deal with this when building reinforcement learning agents is to introduce optimism into the agent (Szita and Lörincz, 2008; Sunehag and Hutter, 2012a). Optimism is also a principle that applies broadly in human decision making (Weinstein, 1980) and a case can be made that it should be considered another

fundamental principle rather than a bias to overcome. The AIXI agent can be modified (Sunehag and Hutter, 2012b) to incorporate optimism in a manner consistent with multiple-prior expected utility (Gilboa and Schmeidler, 1989), though reversing pessimism to optimism. These agents have better guarantees for asymptotic optimality, consistent with the empirical observation that the practical MC-AIXI-CTW agent needs more exploration heuristically added for good performance.

Two other issues are the choice of discount function and the choice of reference machine. In this article we describe AIXI based on geometric discounting while the fully general formulation can be found in Hutter (2005). Humans often display a more hyperbolic than geometric discount scheme (Laibson, 1997). The geometric discount is often considered preferable because it is time consistent (Lattimore and Hutter, 2014).

Finally, it has been discussed (Li and Vitányi, 2008) what a natural choice of reference machine would be. The choice will matter as long as we are considering a limited amount of data while asymptotically it does not. An agent based on cellular automata might have a very different sense of what is simple than a human does. The concept of a natural Turing machine could be interpreted as a reference machine that provides a notion of simplicity that agrees with human judgment. The computer programming languages that we create to make it easier for us to program computers capture to some extent a human notion of simplicity. The attempts at finding a more objective notion of a natural Turing machine have so far failed and such a choice might not exist (Müller, 2010). A recent alternative suggested by Sunehag and Hutter (2014) is to learn a reference machine for which Occam's razor is true as a proposition about the relevant world.

Defining and Measuring Intelligence

Given a reference machine and the corresponding universal mixture environment ξ, the quantity V_ξ^π has been considered as a universal intelligence measure (Legg and Hutter, 2007). It has much in common with more narrow standard human IQ tests where a person's task is to see patterns in, e.g., a sequence of numbers and guess their continuation. In principle, one can argue for any next number and find intricate mathematical patterns motivating very different answers. What makes the answers well defined is that there is an implicit simplicity bias. The continuation should be based on the simplest pattern. If one can find the simplest program explaining what has so far been observed from an unknown environment and take actions as if that is the true environment, one receives high reward in expectation and hence, a high intelligence score when measured by V_ξ^π. This universal measure can be seen as a more complex generalization of human IQ tests which have actually been successfully attacked by machine intelligence (Sanghi and Dowe, 2003). This achievement has lead to the conclusion that current IQ tests are too narrow to measure general machine intelligence which can be tailored to the task (Dowe and Hernndez-Orallo, 2012). An approximation to the universal test has recently been introduced (Legg and Veness, 2011) and applied to test a number of agents and even more recently a large class of challenging environments has been created using Atari games (Bellemare et al., 2013).

Summary

Algorithmic complexity provides a formalization of the notions of simplicity and complexity. It leads up to a formal universal theory of sequence prediction, induction, similarity, and randomness based on simplicity-biased rationality, principles suggested as idealized human cognition. Further, for the more difficult sequential decision-making setting algorithmic complexity can again together with rationality be used to define a universal intelligent agent as well as a universal measure and definition of intelligence. For the sequential decision-making setting, there are indications that further principles like optimism should be considered fundamental and not unwanted biases. Practical approximations of the otherwise incomputable theories discussed have been built based on common compression algorithms and have been successfully used for a range of practical applications including clustering of sequence data like DNA, text, and music. Further, algorithmic complexity has also simplified proofs in mathematics and resolved philosophical problems including 'problems of induction' in the philosophy of science and the problem with 'Maxwell's demon' in statistical mechanics.

See also: Algorithm; Computational Approaches to Model Evaluation; Social Network Algorithms and Software; Statistical Clustering.

Bibliography

Bennett, C.H., Gacs, P., Li, M., Vitányi, P.M.B., Zurek, W.H., July 1998. Information distance. Information Theory, IEEE Transactions on 44 (4), 1407–1423.

Bennett, C., Li, M., Ma, B., June 2013. Linking chain letters. Scientific American.

Bellemare, M.G., Naddaf, Y., Veness, J., Bowling, M., 2013. The arcade learning environment: an evaluation platform for general agents. Journal of Artificial Intelligence Research 47, 253–279.

Chaitin, G., 1969. On the length of programs for computing finite binary sequences: statistical considerations. Journal of the ACM 13, 547–569.

Chater, N., 1996. Reconciling simplicity and likelihood principles in perceptual organization. Psychological Review 103, 566–581.

Chater, N., 1999. The search for simplicity: a fundamental cognitive principle? The Quarterly Journal of Experimental Psychology Section A 52 (2), 273–302.

Chomsky, N., 2005. Rules and Representations. Columbia Classics in Philosophy. Columbia University Press.

Chater, N., Oaksford, M., 1999. Ten years of the rational analysis of cognition. Trends in Cognitive Sciences 3 (2), 57–65.

Chater, N., Vitányi, P., 2003. Simplicity: a unifying principle in cognitive science? Trends in Cognitive Sciences 7 (1), 19–22.

Cilibrasi, R., Vitányi, P., 2005. Clustering by compression. IEEE Transactions on Information Theory 51 (4), 1523–1545.

Dowe, D., Hernndez-Orallo, J., 2012. IQ tests are not for machines, yet. Intelligence 40 (2), 77–81.

Gilboa, I., Schmeidler, D., 1989. Maxmin expected utility with non-unique prior. Journal of Mathematical Economics 18 (2), 141–153.

Griffiths, T., Tenenbaum, J., 2003. Probability, algorithmic complexity, and subjective randomness. In: Proceedings of the 25th Annual Conference of the Cognitive Science Society. Erlbaum, Mahwah, pp. 480–485.

Hsu, A., Chater, N., Vitányi, P., 2013. Language learning from positive evidence, reconsidered: a simplicity-based approach. Topics in Cognitive Science 5 (1), 35–55.

Hutter, M., 2005. Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability. Springer, Berlin.

Hutter, M., 2007a. Algorithmic information theory. Scholarpedia 2 (3), 2519.

Hutter, M., 2007b. On universal prediction and bayesian confirmation. Theoretical Computer Science 384, 33–48.

Kolmogorov, A.N., 1965. Three approaches to the quantitative definition of information. Problems of Information Transmission 1, 1–11.

Krakauer, D.C., 2002. Evolutionary principles of genomic compression. Comments on Theoretical Biology 7, 215–236.

Kahneman, D., Slovic, P., Tversky, A., 1982. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, New York.

Laibson, D., 1997. Golden eggs and hyperbolic discounting. Quarterly Journal of Economics 112 (2), 443–477.

Li, M., Chen, X., Li, X., Ma, B., Vitányi, P., 2004. The similarity metric. IEEE Transactions on Information Theory 50 (12), 3250–3264.

Legg, S., Hutter, M., December 2007. Universal intelligence: a definition of machine intelligence. Minds and Machines 17 (4), 391–444.

Lattimore, Tor, Hutter, Marcus, 2014. General time consistent discounting. Theoretical Computer Science 519, 140–154.

Li, M., Vitányi, P., 2008. An Introduction to Kolmogov Complexity and its Applications. Springer.

Legg, S., Veness, J., 2011. An approximation of the universal intelligence measure. In: Algorithmic Probability and Friends, Volume 7070 of Lecture Notes in Computer Science. Springer, pp. 236–249.

Martin-Löf, P., 1966. The definition of random sequences. Information and Control 9, 602–619.

Mahmud, M.M.H., 2009. On universal transfer learning. Theoretical Computer Science 410 (19), 1826–1846.

Müller, M., 2010. Stationary algorithmic probability. Theoretical Computer Science 411, 113–130.

Perfors, A., Regier, T., Tenenbaum, J., 2006. Poverty of the stimulus? A rational approach. Proceedings of the Annual Conference of the Cognitive Science Society, 663–668.

Piantadosi, S., Tily, H., Gibson, E., March 1, 2011. Word lengths are optimized for efficient communication. PNAS 108 (9), 3526–3529.

Rathmanner, S., Hutter, M., 2011. A philosophical treatise of universal induction. Entropy 13 (6), 1076–1136.

Rissanen, J., 1978. Modeling by shortest data description. Automatica 14, 465–471.

Russell, S.J., Norvig, P., 2010. Artificial Intelligence: A Modern Approach, third ed. Prentice Hall, Englewood Cliffs, NJ.

Sanghi, P., Dowe, D.L., 2003. A computer program capable of passing I.Q. tests. In: Proceedings of the Joint International Conference on Cognitive Science (ICCS/ASCS-2003), pp. 570–575.

Sunehag, P., Hutter, M., 2012a. Optimistic agents are asymptotically optimal. In: Proceedings of the 25th Australasian AI Conference. Springer.

Sunehag, P., Hutter, M., 2012b. Optimistic AIXI. In: Proceedings of the Fifth Conference on Artificial General Intelligence, Volume 7716 of Lecture Notes in Computer Science. Springer, pp. 312–321.

Sunehag, P., Hutter, M., 2014. Intelligence as inference or forcing Occam on the world. Proceedings of the seventh conference on Artificial General Intelligence. Lecture Notes in Artificial Intelligence 8598, Springer, pp. 186–195.

Szita, I., Lörincz, A., 2008. The many faces of optimism: a unifying approach. In: ICML, Volume 307 of ACM International Conference Proceeding Series, ACM, pp. 1048–1055.

Solomonoff, R., 1964. A formal theory of inductive inference. Part I and II. Information and Control 7 (1, 2), 1–22, 224–254.

Solomonoff, R., 1978. Complexity-based induction systems: comparisons and convergence theorems. IEEE Transactions on Information Theory 24, 422–432.

Tenenbaum, J., Kemp, C., Griffiths, T., Goodman, N., 2011. How to grow a mind: statistics, structure, and abstraction. Science 331 (6022), 1279–1285.

Turing, A., 1936. On computable numbers, with an application to the Entscheidungs problem. Proceedings of the London Mathematical Society 42, 230–265.

Veness, J., Ng, K.S., Hutter, M., Uther, W., Silver, D., 2011. A Monte-Carlo AIXI approximation. Journal of Artificial Intelligence Research 40, 95–142.

Wallace, C., Boulton, D., 1968. An information measure for classification. Computer Journal 11, 185–194.

Weinstein, N., 1980. Unrealistic optimism about future life events. Journal of Personality and Social Psychology 39 (5), 806–820.

Zipf, G., 1949. Human Behaviour and the Principle of Least-Effort. Addison-Wesley, Cambridge, MA.

Alienation, History of

Devorah Kalekin-Fishman, University of Haifa, Haifa, Israel
Lauren Langman, Loyola University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of alienation has a long and complex history as it has been used in Judeo–Christian theology, Roman law, medieval concepts of insanity, German Idealism especially in Hegel, in the dialectical materialism of Marx, and more recently, in Frankfurt School Critical Theory and in existentialism. As adopted and adapted to different disciplines, alienation serves to convey an understanding of voluntary and involuntary withdrawal from the full realization of capabilities. Thus, alienation bears a variety of meanings, ranging from its usage as a term for the general malaise of the human condition to its usage as a detailed description of the specifics of oppression through mechanisms of wage labor. In structuralist theories, alienation is understood as an inevitable outcome of the social order whether or not individuals are aware of their situation. With this assumption, researchers collect data in the widest possible economic, political, and cultural frameworks. In analyses that focus on psychological evidence, it is often assumed that people's feelings are indicators of levels of alienation. Following this assumption, research is usually based on formal questionnaires the scores of which correlate positively or negatively with broad social orientations. Recent social science researches which center, among others, on social exclusion, inegalitarianism, racism, and intersectionality refine the meanings of the term by making aspects of alienation explicit according to context. Although 'alienation' has often been contested because of its adaptability to different applications, the legacy of its many rich meanings endures. Theoretical explorations of alienation continue to demonstrate its explanatory power in relation to modernity, as disclosed in its consequences and in the procedures of its discontents from social withdrawal to social mobilizations.

When we chronicle uses of the term 'alienation' we find a set of descriptions of the human condition which are not necessarily congruent, and the mix is likely to be bewildering. In some writings, alienation is defined as identical to the oppressive aspects of human life. In others, the term points to explorations of a higher good. And in the writings of still others, the same term is deployed to describe a type of interaction which may have either implication. Each description has different consequences for the design of research and different implications for intervening in social processes. Because of the diversity, alienation has often been denounced for its inappropriate 'flexibility' and plurality of meanings. Still, the construct repeatedly emerges with renewed vigor from the ashes of fiery criticism. Its usefulness has been highlighted in diverse disciplines, among them law, theology, and medicine; it has also played an important role in philosophical, sociological, and social psychological analyses. In this article, we present a more or less chronological reading of the history of alienation, its meanings, and its uses in research, thus underlining the multiple ways in which it has figured in Western thought.

The root of the term alienation is the Latin adjective for 'other' – alienus, -a, -um. In Ancient Rome, the noun 'alienatio' was primarily a legal term related to handing over property (Buckland, 2007/1921: 228f.). In Christian theology, alienation has referred to several ways of 'othering'; most important is the perception that alienation from the Divinity is irremediable because of the primordial sin: eating the fruit of the Tree of Knowledge of Good and Evil. In more recent times, the term alienation has referred to deviation from correct doctrine (Ladner, 1983: 969). Among some theologians alienation has also been adopted as the term that describes the sinful writing of 'fiction,' that is to say, all types of writing which justify the existence of worlds other than that created by the Deity (Haidu, 1977).

Alienation, with mixed implications, has been important in different schools of philosophy. In the integrated philosophical system that Georg Wilhelm Friedrich Hegel (2007 [1770–1831]) developed, alienation figures in two different ways. As *Entfremdung*, alienation is recognized as holding a negative meaning, the process of becoming and remaining isolated, i.e., a stranger. But Hegel also recognized alienation as the positive phenomenon of externalization (*Entäusserung*). As the externalized consciousness of individuals, alienation is necessary because it turns into a collective spirit which enables the construction of *a shared culture in civil society* (Ice, n.d., www.ilovephilosophy.com). In the existentialist philosophy that evolved in the nineteenth and twentieth centuries, alienation was a central concern. For Sören Kierkegaard (1813–55), the average person, living a life of banal conformity, is alienated from his/her self-realization and transcendence. Later, Martin Heidegger (1889–1976) describes *Mann* who embraces an inauthentic life of conformity and conventionality, because he is alienated from his contingency (throwedness) and his finitude (being unto death). Later, Jean-Paul Sartre (1956), building on Hegel, shows that alienation, which is universal and inevitable, appears as a condition that may be both positive and negative. In his explication, alienation involves recognition that everyone has to get along with an 'alien Me,' i.e., with the understanding that part of one's existence is beyond subjective experience. But according to Sartre, this objectification of self and others ultimately reveals a fundamental truth, and thus alienation is not negative. As philosopher and theologian, Tillich (1886–1965) sought correlates between theology's concern with the nature of human life and the concerns of existential philosophy. He describes the 'tragic universality' of 'Entfremdung' as a negation of humanity. The isolation that he describes is a state of alienation that suffers

from a sense of unbelief (rather than faith), a desire for concupiscence (not love), and a drive for *hubris* and self-elevation rather than surrender to the omnipotent immortal (Williamson and Cullingford, 1997).

In social philosophy as well, alienation has served to indicate both positive and negative aspects of human existence. Jean-Jacques Rousseau (1712–78), for example, defined alienation as people's *voluntary* renunciation of power for the greater good of the collective. In this view, alienation is a necessary foundation of democratic social life. By contrast, Adam Smith (1723–90) described alienated individuals as those doomed to primitive surroundings whether in farming or in urban factories and thus cut off from modern culture. In theories that developed from the nineteenth to the twenty-first century, the negative implications of alienation have been beyond doubt. In philosophical anthropology, alienation is described as a social condition that thwarts human possibilities for community and the self-realization that can enable a fulfilling life (Schacht, 1992). Debates center on whether these limitations are due to capitalism, modernity, or the human condition. Major critiques of the meanings of alienation, explanations of its causes, and strategies of amelioration have had an enduring impact (see, for example, Israel, 1971; Mészáros, 1970).

Echoes of the approach of social philosophy are to be found in attempts to describe medical practices in terms of the social context. Denoting mental illness as alienation, physicians diagnosed pathological deviations from normal behavior as contingent on the social context (Jacob and Beveridge, 1995). During the century of the revolutions in France (1789, 1830, 1848, and 1871), 'alienists' attributed what was seen as a rising rate of mental illness to the excess of passion in the state of France. This perception had practical consequences in conclusions that the 'alienated' have to be treated in separate institutions, rather than in general hospitals, as had been the practice till then.

Interestingly, in literature, alienation has been hailed as both technique and topic. Thus nineteenth-century poets such as Charles Baudelaire (1821–67) and Arthur Rimbaud (1854–91) talked to the alienated condition of humankind as did Louis-Ferdinand Celine (1894–1961) in his novels in the twentieth century. From a somewhat different standpoint, writers who committed themselves to the 'new novel' in twentieth-century France sought ways of describing alienation as it unfolds in banal nonevents. Taking up a position of distance from human feeling, they saw writing as a kind of service to the possibilities embedded in the French language. Thus, at a distance, in alienation, they elaborated detailed descriptions of the surround of events rather than the events themselves (Marguerite Duras (1914–96), Alain Robbe-Grillet (1922–2008), Nathalie Sarraute (1900–99)). In research into literature, alienation is frequently found to be a fruitful entry point into the style and goals of authors.

In drama, alienation has been especially important. Sartre depicted poignant situations of alienation in didactic dramas that demonstrated implications of existentialist philosophy for action. Bertolt Brecht (1898–1956) not only wrote about the alienated condition of modern humanity under capitalism, but famously insisted that for his political dramas he wanted the audience to experience *Verfremdung* (usually translated as 'alienation'). He meant for spectators to remain at a distance from what was happening on the stage. Since such an audience can pay rational attention to how the situations are constructed, they are likely to be free to think about the political implications of the drama. Thus, in the theater, alienation is considered the foundation for generating desirable social action.

Alienation in the Social Sciences

All the meanings of alienation noted above appear at some point in the social sciences.

Like Adam Smith, Harriet Martineau (1802–76) decried the exploitation of workers which prevents them from enjoying satisfactions that accrue from work. This is especially true of women whose domestic work is a kind of total enslavement.

Describing the alienating effects of the division of labor in industrial society, Karl Marx (1818–83) moved the Hegelian dialectic from the realm of ideas to its material foundations in political economy (Marx [1844] 1977). Thus, he formalized, contextualized, and refined the meaning of alienation. In his theory, capitalist relations of production turned men into extensions of machines, estranged workers from their selves, cut them off from community support, and from the ability to control their lives effectively. In this context, alienation was a kind of summary of a complex chain of injustices. With land scarce, workers migrated to cities where they sold their labor power to factory owners who determined what to produce and how. Workers have to sell their labor and as with all commodities, the value of labor is in constant flux. Marx developed an extensive critique of capitalist political economy on the basis of his dialectical concept of alienation as the process of oppression most significantly, and the term for the outcomes, acute distress primarily economic but also psychological (Marx and Engels, 1867; Williamson and Cullingford, 1997).

Although alienation is not a term they used, Max Weber (1864–1920) and Emile Durkheim (1858–1917) also wrote about forms of alienation when they noted the adverse consequences of working conditions in modern society. For Weber (1958), alienation is concretized in rationality, the dominant value of modernity. Rational principles of calculability are applied to every domain of modern life including the work of all the 'free professions'; it shapes production and bureaucratic administration in industry as well as in politics. Calculation structures human relations and infiltrates subjectivity, incorporating every aspect of personality within an alien, and alienating system which dehumanizes people and imprisons them in 'iron cages.'

In his writings, Durkheim (1984/1893) describes both negative and positive aspects of alienation. On the one hand, he saw rationality in the 'organic' division of labor that makes industrialization possible as a basis for renewing social solidarity. On the other hand, the manifold varieties of unforeseen situations, the dynamics of industrial change, could lead to a weakening of social ties and a fragmentation of consciousness. The failure of normativity, difficulty in deciding how to apply familiar rules and how to find what new rules of behavior to adopt signals a specific condition of alienation, which was, for Durkheim, 'anomie' (1997/1897).

Conceptualizing Alienation in Critical Sociology

Basing their initial ideas on the Marxian arguments against capitalism, researchers of the school of critical sociology developed more shaded meanings of alienation (Horkheimer and Adorno, 1972). Gyorgy Lukacs (1923) explored the impact of the rational division of labor on consciousness. He described alienation as reification – a misguided perception of social relationships as concrete 'things' in a steady state. Moreover, through the embedding of reification in ideology, a distorted understanding of the social system is ensured. Because people lack a 'dialectical conception of the totality of social relations within the historical process' (Twining, 1980: 419), they remain unaware of the contradictions in the capitalist system which enforce alienation.

Members of the Frankfurt School, who relied on Lukacs' analyses of alienation for a cultural reading of Marx, and on Marx's 1844 *Manuscripts* (unknown until the 1920s), expanded on the understanding of reification (distorted readings of reality) as a means of domination. For Herbert Marcuse (1969), alienation could be positive as the ground for a total revolution. But this was only likely to emerge among those outside the systems of production, e.g., students, artists and writers, the long-term unemployed, and ethnic minorities. In the meantime, conditions of alienated labor keep people deceived, distracted, and powerless either to understand their lives or to mobilize for progressive change. Even the apparent liberation of erotic desire in late capitalism alienates persons from their 'real needs' in order to incorporate them into the 'administered society' (Marcuse, 1969). Archibald (2009) shows alienation in a different light when he reasons that workers no longer care enough about work to feel deprived by its alienating effects. Instead they channel their aspirations to consumption and leisure. Still, with consumption commercialized, mass produced, and less subject to individuals' control, aspirations can easily be frustrated. Thus, alienation imposes itself in a different way.

Additional Conceptualizations of Alienation

In various guises, the topic of alienation appears in the work of most sociologists in the twentieth century. Early in the century, alienation can be recognized in Georg Simmel's stranger who, coming from one location and staying, sometimes for a lifetime, in another, will never fully belong. Paradoxically, as an outsider who suffers from remaining a stranger, he often makes positive contributions to the collective in the new location (see Simmel in Wolff, 1950). Robert E. Park (1928: 893) extended Simmel's insight specifically to a view of the advantages of migration and its effects in creating 'marginal men' whose alienation is a positive factor for the 'new' society. While acknowledging the 'moral turmoil,' the anomie that Durkheim identified as the root of discontent, Park sees the marginality that is consequent on mobility – social or geographical – as a key to advancement and growth.

In discussing 'anomia,' Robert K. Merton (1938) presented a model of how alienation is implemented in continuities and discontinuities between culturally approved goals and institutionally available means. His typology of 'modes of adaptation' shows the outcomes of acceptance or rejection of approved goals and means. In functioning societies, most people adjust to culturally approved goals (values) by accepting institutionalized means (norms) for attaining them. This is a pattern where culturally approved goals are taken for granted and institutionally available means are adopted. Other combinations are deviant: Accepting culturally approved goals while rejecting normatively available means leads to 'innovation,' which may be beneficial for society, but this pattern may also describe role adjustments such as resorting to criminal behavior. Adherence to legitimate means when one no longer accepts culturally approved goals is alienation as a type of 'ritualism.' Rejection of both goals and means were seen as 'retreatism,' cutting oneself off from the social milieu.

A recent compilation of writing on alienation (Langman and Kalekin-Fishman, 2005) follows the Frankfurt School view that alienation is inevitable under capitalist production and that the class system shapes every domain of social life. In that volume, theoretical contributions by Harry F. Dahms, Matthew David, Lauren Langman, and David Norman Smith detail how structures of late capitalist society insinuate themselves into the realms of culture causing (mis)understandings of the world and of one's self. In the family, too, people are implicated in the capitalist mode of production as consumers (Mészáros, 1970).

A different kind of definition of meaning was proposed in Melvin Seeman's (1959) landmark literature review which departs from concerns of viewing the inevitable albeit unwitting effects of social arrangements on consciousness and on behavior. Drawing upon the ideas of Marx, Mannheim, Weber, and Durkheim, Seeman highlights five different social psychological consequences of alienation: *powerlessness*, the feeling that one is incapable of dealing with problems that arise; *meaninglessness*, the feeling that one cannot know the outcome of one's actions; *social isolation*, the feeling of not being a part of any social group; *self-estrangement*, the lack of an inability to identify where one's true interests lie. In this set, he identifies anomie, *normlessness*, the state of bewilderment about what norms are applicable in unforeseen situations, as a particular type of alienation. Later, he added a dimension that he felt was distinct, *cultural alienation*.

By the middle of the twentieth century, then, theoreticians proposed several mutually exclusive types of definition for alienation. Embedded in differently structured theoretical approaches, the various definitions play an important role in different kinds of research. In a sense, one may say that in its adaptability to different kinds of research lies the secret of the vitality of alienation as a construct in the social sciences.

Alienation Research: Approaches and Objects of Investigation

Research based on a structural view of alienation is primarily conducted holistically and contextualized historically. Although researchers of this School often look at what people say and how people act, the importance of statements and actions is taken to be derived from macrosocial trends (Braverman, 1974). As noted above, it is taken for granted that the conditions which cause alienation are likely to be well-hidden from people's consciousness and that research has the extrascientific task of shaping awareness to the point of provoking action. Thus, in this type of research there is a focus on describing the relations of production in the macro, in order

to draw conclusions about likely consequences for the conditions that people have to confront. Among studies which examine the immediate social context of more intimate situations are researches on work and organizational behavior, education, everyday life as well as gender, community, and ethnic relations. In the latter part of the twentieth century, theories of alienation were applied increasingly to studying responses to the mass media and consumerism. In sum, structural approaches to alienation research are realized at the macro-, the micro-, and the mesosocial levels.

At the same time, however, in the mid-twentieth century, in the wake of Seeman's (1959) survey of the literature, researchers showed increasing interest in the assumption that alienation can be described in full by people who suffer from it. Researchers could thus use closed questionnaires to measure the frequency and intensity of experiencing relevant dimensions of alienation. In them, respondents are asked to report on their feelings of alienation at work, in the family, among friends, and even in relation to their 'selves.' For several decades (beginning in the 1950s), many efforts were made by American researchers to produce reliable and valid scales of alienation. The objects of investigation were equally abstract as they attempted to track correlations of alienation with constructs such as liberalism, tolerance, acceptance of democracy, and life satisfaction among different groups. Gwynn Nettler's (1957) 'measure of alienation' tested the degree to which people felt alienated from their own society and its culture. Leo Srole (1956) elaborated a measure of 'powerlessness' and examined its correlation with a readiness to countenance messages about interethnic tolerance. After reducing the five dimensions delineated by Seeman (1959) to three, social isolation, powerlessness, and meaninglessness, Dwight G. Dean (1961) constructed several scales to measure alienation. He proposed that alienation could be identified in social apathy, authoritarianism, conformity, cynicism, hoboism, political apathy, political hyperactivity, personalization in politics, prejudice, privatization, psychosis, regression, and suicide. In work that assessed correlations of vegetativeness (apathy), nihilism, and adventurousness with powerlessness, significant correlations were found among all subscales (Maddi et al., 1979). This approach has not exhausted all the possibilities. Taking still a different view Davids (1955) examined alienation as related to egocentricity, distrust, pessimism, anxiety, and resentment.

The now canonic study by Blauner (1964) is a vivid example of how questionnaires could be shaped to deal with research on the structural grounds for alienation and on feelings of alienation. By circulating questionnaires to workers in four industries each of which demanded a different level of skill from employees, Blauner expected to find negative correlations between levels of skill and levels of perceived alienation. Findings confirmed his hypothesis that the lower the level of skill required for any given job, the higher the score for alienation. But Blauner also examined structural aspects of the workers' lives and ultimately could show that workers' feelings of alienation were mitigated by community solidarity.

During the youth rebellion of the 1960s, Kenneth Keniston (1971) calculated an 'alienation syndrome' that could be explained by socialization in families with a leftist orientation and by an innate 'protest-prone personality.' Scales measuring alienation are currently important in research on childhood, on maladjustment, on depression and mental illness, as well as in organizational research, in research on race and class, and in educational research of different kinds. Not all the research in this vein is of the same kind. Although almost all the scales that measure alienation were compiled in the USA, they are often adapted to research that is conducted in Africa and Asia as well as in Europe.

To this day, there are numerous researches into evidence of different dimensions of felt alienation in hospitals and prisons, as well as in political juggling and in the labor market. Much of the contemporary research is oriented to refining the categories Seeman proposed and not always by means of validated scales. In a study of teachers' feelings of alienation, for example, Jeffreys Brooks et al. (2008) carried out interviews over a two-year period. Content analysis of the data showed that, among the interviewees, not only were conceptualizations of the dimensions interrelated, but the salience of different dimensions shifted under changing circumstances. It has also been shown that dimensions of alienation could serve in analyses of popular jokes (Mauldin, 2008).

Both quantitative and qualitative analyses have been applied to studying the nature of the possible spread of alienation through the expanding uses of new digital technologies. Douglas Kellner (2005) has argued that there are connections between alienation and responses to digital technologies that range from 'technophobia' to 'technophilia.' Contact with digital technologies has been analyzed structurally to show that they are likely to foster new forms of alienation. But understood dialectically, they can also provide new realms of agency and meaning. Research of these possibilities often makes use of questionnaires (see Gunter et al., 2002 for an evaluation of the possibilities; Pietilä, 1970 for an early study of alienation and the mass media). More specifically, research of reactions to digital media has led to revised meanings of alienation as the degradation of political literacy, i.e., losing the capacity for independent, critical thought, and being likely to agree with the latest thing read or heard (Hassan, 2003).

Since the end of the twentieth century, there has been increasing interest in understanding alienation so as to derive meaningful ways of promoting dis-alienation. This research focuses on detailed examinations of extremes of consumption to the point of carnivalization – modes of intentional self-liberation which, paradoxically, are likely to lead to further enslavement (Langman, 2011). A more hopeful line of research in that direction is research into the structural foundations of social movements (Langman, 2013). At the same time, research based on a perception of alienation as a state of self-awareness is pursued in the investigations of teaching and learning in education, in ethnic and race relations, and in relation to persistent problems in the organization of work.

See also: Authoritarian Personality; Counterculture: 1960s and Beyond; Counterculture: The Classical View; Critical Psychology; Immigration: Social Psychological Aspects; Intersectionality; Racism, Sociology of; Social Inequality in Cultural Consumption Patterns; Social Inequality in History (Stratification and Classes); Youth Culture, Sociology of.

Bibliography

Archibald, P., 2009. Marx, globalization and alienation: received and underappreciated wisdoms. Critical Sociology 35 (2), 151–174.

Blauner, R., 1964. Alienation and Freedom. University of Chicago Press, Chicago, IL and London.

Braverman, H., 1974. Labor and Monopoly Capitalism. Monthly Review Press, New York.

Brooks, J.S., Hughes, R.M., Brooks, M.C., 2008. Fear and trembling in the American high school: educational reform and teacher alienation. Education Policy 22, 45–62.

Buckland, W.W., 2007 [1927]. A Text-book of Roman Law: From Augustus to Justinian, third ed. Cambridge University Press, Cambridge. revised by P Stein.

Davids, A., 1955. Alienation, social apperception and ego structure. Journal of Consulting Psychology 19 (1), 21–27.

Dean, D.G., 1961. Alienation: Its meaning and measurement. American Sociological Review 26 (4), 753–758.

Durkheim, E., 1984 [1893]. Division of Labor in Society. Trans. W.D. Hall. The Free Press, New York.

Durkheim, E., 1997 [1897]. Suicide. The Free Press, New York.

Gunter, B., Nicholas, D., Huntington, P., Williams, P., 2002. Online versus offline research: implications for evaluating digital media. Aslib Proceedings 54 (4), 229–239.

Haidu, P., 1977. Repetition: modern reflections on medieval aesthetics. MLN 92 (5), 875–887.

Hassan, R., 2003. The MIT media lab: techno-dream factory or alienation as a way of life? Media, Culture and Society 25 (1), 87–106.

Hegel, G.W.F., 2007 [1807]. Philosophy of the mind. In: Trans. Wallace, W., Miller, A.V., Inwood, M.J. (Eds.), Oxford: Clarendon.

Horkheimer, M., Adorno, T.W., 1972. Dialectic of Enlightenment. Trans. J. Cummings. Seabury, New York.

Ice T., n.d. Alienation and estrangement in Hegel's Phenomenology of the Spirit. Symposia – The Online Philosophy Journal. Available at: www.ilovephilosophy.com.

Israel, J., 1971. Alienation: From Marx to Modern Sociology: A Macro-sociological Analysis. Allyn and Bacon, Boston, MA.

Jacob, F., Beveridge, A., 1995. Madness and politics: French nineteenth-century alienists' responses to revolution. History of Psychiatry 6, 421–429. http://dx.doi.org/10.1177/0957154X9500602401 (accessed 15 July, 2013).

Kellner, D., 2005. New technologies and alienation: some critical reflections. In: Langman, L., Kalekin-Fishman, D. (Eds.), The Evolution of Alienation: Trauma, Promise and the Millennium. Rowman and Littlefield, Lanham MD, pp. 47–67.

Keniston, K., 1971. Youth and Dissent: The Rise of a New Opposition. Harcourt Brace Jovanovich, New York.

Ladner, G.B. (Ed.), 1983. Images and Ideas in the Middle Ages, Vol. II. Edizioni di Storia e Letteratura, Roma.

Langman, L., 2011. Alienation, entrapment and inauthenticity: carnivalization to the rescue. In: Braun, J., Langman, L. (Eds.), Alienation and the Carnivalization of Society. Routledge, New York, pp. 54–75.

Langman, L., 2013. Capitalism, crises, and "great refusals": critical theory, social movements, and utopian visions. Radical Philosophy Review 16 (1), 349–374.

Langman, L., Kalekin-Fishman, D. (Eds.), 2005. The Evolution of Alienation: Trauma, Promise and the Millennium. Rowman and Littlefield, Lanham MD.

Lukacs, G., 1923. History and Class Consciousness. Trans. A. Blunden. Available at: www.marxists.org/archive/lukacs/works/history/hcc05.htm.

Maddi, S.R., Hoover, M., Kobasa, S.C., 1979. An alienation test. Journal of Humanistic Psychology 19, 73–76.

Marcuse, H., 1969. An Essay on Liberation. Beacon Press, Boston, MA.

Marx, K., 1977 [1844]. The economic and philosophical manuscripts. In: Tucker, R. (Ed.), The Marx-Engels Reader. WW Norton and Company, New York, pp. 66–126.

Marx, K., Engels, F., 1867. Das Kapital. Available at: http://ciml.250x.com/archive/marx_engels/german/kapital1.pdf.

Mauldin, R.D., 2008. Alienation: a laughing matter. Humor – International Journal of Humor Research 21 (3), 313–345.

Merton, R.K., 1938. Social structure and anomie. American Sociological Review 3 (5), 672–682.

Mészáros, I., 1970. Marx's Theory of Alienation. Merlin, London.

Nettler, G., 1957. A measure of alienation. American Sociological Review 22 (6), 670–677.

Ollman, B., 1976. Alienation: Marx's Conception of Man in Capitalist Society, second ed. Cambridge University Press, Cambridge.

Park, R.E., 1928. Human migration and the marginal man. American Journal of Sociology XXXIII (6), 881–893. Available at: www.jstor.org/stable/2765982/seq=1 (accessed 24 June, 2013).

Pietilä, V., 1970. Alienation and the use of the mass media. Acta Sociologica. 13, 237–252.

Sartre, J.-P., 1956. Being and Nothingness. Gallimard, New York.

Schacht, R., 1992. Hegel, Marx, Nietzsche and the future of self-alienation. In: Geyer, F., Heinz, W. (Eds.), Alienation, Society and the Individual: Continuity and Change in Theory and Research. Rutgers, NJ, pp. 1–16. Transaction Books.

Seeman, M., 1959. On the meaning of alienation. American Sociological Review 24, 783–791.

Simmel, G., 1950. The stranger. In: Wolff, K. (Ed.), The Selected Writings of Georg Simmel. The Free Press, New York and London, pp. 402–408.

Smith, A., 1986. The essential Adam Smith. In: RL Heilbroner with LJ Malone. WW Norton, New York.

Srole, L., 1956. Social integration and certain corollaries: an exploratory study. American Sociological Review 32 (4), 418–420.

Twining, E.J., 1980. Alienation as a social process. The Sociological Quarterly 21 (3), 417–428.

Walliman, I., 1977. Estrangement, Alienation and Exploitation: A Sociological Approach to Historical Materialism. Columbia University Press, New York.

Weber M (1958) From Max Weber, Trans. HH Gerth, H.H., Mills, C.W. (Eds.), Galaxy Press, New York.

Wexler, P., 1996. Holy Sparks: Social Theory, Education and Religion. St Martin's Press, New York.

Williamson, I., Cullingford, C., 1997. The uses and misuses of 'alienation' in the social sciences and education. British Journal of Educational Studies 45 (3), 263–275.

Zhu, H., Zhou, Y., 2013. Reductionism and alienation of science and technology. http://dx.doi.org/10.3863/j/issn/1674-5094.2013.01.019 (accessed 31 May, 2013).

Alienation: Psychosociological Tradition

Catherine E Ross and John Mirowsky, University of Texas, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Seeman, volume 1, pp 385–388, © 2001, Elsevier Ltd.

Abstract

Subjective alienation is the sense of being separate from oneself or others. Perceived powerlessness is the main type of alienation because it is a separation from important outcomes in one's own life. Self-estrangement is a type of powerlessness related to one's work, and the closest conceptually to Marx's ideas of alienation. Others include social isolation, normlessness, and meaninglessness. Mistrust is a recently identified type of alienation. It is a type of social alienation more profound than social isolation. We focus on perceived powerlessness and mistrust: the concepts and measures, social causes, and emotional consequences. The major social causes include socioeconomic status (education, income, and work); gender; age; and neighborhoods. We conclude with directions for future research on contextual causes of mistrust and powerlessness.

Subjective Alienation

Subjective alienation is the sense of detachment or separation from oneself and society (Seeman, 1959, 1983; Fischer, 1976: Mirowsky and Ross, 2003). In 1959, Seeman identified five types of subjective alienation: powerlessness, self-estrangement, social isolation, normlessness, and meaninglessness. Over the years, Mirowsky and Ross expanded and modified these initial components to include powerlessness, self-estrangement, social isolation, normlessness, meaninglessness, and mistrust. Powerlessness is the essence of subjective alienation because it is separation from oneself. People with a sense of powerlessness think that they have little control over meaningful events and circumstances in their lives. Self-estrangement is a type of powerlessness related to work. People who are estranged from work do not see their work as an expression of themselves, but rather something given over to the control of others. Normlessness is the perceived gap between goals and legitimate means to achieve them, with the resulting dissociation from socially approved norms governing behavior. Social isolation is the sense of separation from others. It is the sense of not having anyone who is someone to you and not being someone to anyone. Meaninglessness is the sense that one's life has no meaning or purpose. Mistrust is suspicion of others.

Powerlessness and normlessness both are signal gaps in personal achievement: the cognitive awareness of a discrepancy between one's goals and one's ability to achieve them. Powerlessness is a more profound sense of alienation than is normlessness, since normlessness is the perceived gap between goals and legal, approved and normative means whereas powerlessness is the perceived gap between goals and *any* available means. A person who feels powerless feels helpless to achieve desired ends through any means, whereas a person who feels normless thinks that socially unapproved behaviors could be effective in achieving goals.

Social isolation and mistrust both signal social alienation, or a sense of separation from others. People who are socially isolated are detached from networks of affection, aid, and reciprocity. The opposite of isolation is commonly called social support, which is the sense of being cared for and loved, esteemed and valued as a person, and part of a network of communication and obligation. This concept of isolation and support differs radically from Seeman's original definition of isolation as assigning low value to mainstream society's goals or beliefs, sometimes called 'cultural estrangement.' Although Seeman originally excluded "the warmth, security, or intensity of an individual's social contacts" from his definition of isolation, it is precisely this meaning that has greatest currency today, and that is most important in research.

Mistrust goes beyond a sense of separation from others to one of suspicion of others. It is the cognitive habit of interpreting the intentions and behavior of other people as unsupportive, self-seeking, and dishonest. Mistrusting individuals do not have faith in other people; they think others are out for their own good and would willingly harm them in pursuit of their goals. They therefore believe it is safer to keep their distance from others. Although Seeman did not include mistrust as a type of subjective alienation, Mirowsky and Ross consider mistrust a profound form of social alienation.

Subjective alienation has social causes and consequences for well-being. Most empirical research examines consequences for psychological distress, defined as painful psychological or emotional symptoms such as depression, anxiety, and anger; and for physical well-being. Theoretically, subjective alienation is the cognitive link between reality and distress. It is the subjective reflection of social conditions of objective powerlessness, structural inconsistency (or a gap between goals and means to achieve them), an inability to achieve goals, alienated labor, dependency, threat, and the absence of supportive, trusting, nonpredatory relationships.

Perceived powerlessness, or a separation from oneself, is the essence of subjective alienation. Most research focuses on powerlessness (including some on self-estrangement) and social isolation, with the newest research focusing on mistrust; less research measures or examines normlessness or meaninglessness. Thus, we next go into some detail about the concept of powerlessness.

Perceived Powerlessness

The Names and Faces of Powerlessness

In sociology, the concept of perceived powerlessness vs control is the core type of subjective alienation (Seeman, 1959).

Powerlessness is the separation from important outcomes in one's own life; or an inability to achieve desired ends. Perceived powerlessness is the awareness of this gap between goals and means. Perceived powerlessness is the learned and generalized expectation that one has little control over meaningful circumstances in one's life. On the other end of the continuum, the sense of personal control is the belief that outcomes are contingent on one's own choices and actions; it is the cognitive awareness of a link between efforts and outcomes. (Mirowsky and Ross, 2003; Seeman, 1983). The importance of perceived control vs powerlessness is recognized in a number of social and behavioral sciences, where it appears in several forms with various names, notably mastery (Pearlin et al., 1981), personal autonomy (Seeman, 1983), the sense of personal efficacy (Downey and Moen, 1987; Gecas, 1989), self-directedness (Kohn and Schooler, 1982), agency (Thoits, 2006), and instrumentalism (Wheaton, 1980), and at the other end of the continuum, external locus of control (Rotter, 1966), fatalism (Wheaton, 1980) and perceived helplessness (Elder and Liker, 1982).

Locus of Control

In cognitive psychology, the concept of perceived powerlessness vs control can be traced to social learning theory, and Rotter's concept of locus of control (1966). Belief in an external locus of control is a *learned*, generalized expectation that outcomes of situations are determined by forces external to one's self such as powerful others, luck, fate, or chance. The individual believes that he or she is powerless and at the mercy of the environment. Belief in an internal locus of control (the opposite) is a learned, generalized expectation that outcomes are contingent on one's own choices and actions.

Both Rotter (1966) and Seeman (1959) recognized that perceived powerlessness – the major form of subjective alienation – and external locus of control were related concepts. In fact, Rotter derived the concept of locus of control from the sociological concept of alienation, stating "the alienated individual feels unable to control his own destiny" (1966: p. 263).

The Sense of Personal Control

Based on these traditions, Mirowsky and Ross (1991, 2003) clarified and developed the measure of the sense of personal control vs powerlessness. The sense of personal control corresponds to the personal control component of Rotter's locus of control scale, which includes questions like "when I make plans I can make them work" or "I have little influence over the things that happen to me." The concept of personal control refers to *oneself*, not others, and it is *general*, not realm specific. Thus, unlike Rotter's locus of control scale, it excludes beliefs about the control others have over their lives and realm specific control, like political control, from the concept. For instance, they do not consider questions from the Rotter scale like "the average citizen can have an influence in government decisions" or "there will always be wars" to be measures of the sense of personal control since they do not refer to oneself, and they are realm specific.

Responses to questions about personal control capture the concept of interest, and two other cross-cutting concepts – the tendency to agree and self-defense. Some people tend to agree with statements irrespective of content. Agreement tendency can make it appear as if internal and external controls are uncorrelated. Some people are more likely to believe that they control the good outcomes in their lives than that they control the bad ones (self-defense); others take more responsibility for their failures than for their successes (self-blame). Agreement tendencies and the tendency toward self-defense or self-blame cross-cut the concept of interest and bias measures unless they balance agreement and defense (Mirowsky and Ross, 1990a, 1991, 1996). Thus, the Mirowsky–Ross measure of personal control vs powerlessness balances defensiveness and agreement tendencies to achieve an unbiased measure. It is a two-by-two index that balances statements about control with those about lack of control, and statements about success (good outcomes) with those about failure (bad outcomes). The Mirowsky–Ross sense of personal control vs powerlessness scale (1991) asks respondents how much they agree with the following statements:

Internal

Control over good:
1. "I am responsible for my own successes"
2. "I can do just about anything I really set my mind to"

Control over bad:
3. "My misfortunes are the result of mistakes I have made"
4. "I am responsible for my failures"

External

Powerless over good:
5. "The really good things that happen to me are mostly luck"
6. "There's no sense planning a lot – if something good is going to happen it will"

Powerless over bad:
7. "Most of my problems are due to bad breaks"
8. "I have little control over the bad things that happen to me"

Learned Helplessness and Self-Efficacy

Two other traditions from behavioral psychology and attribution theory in psychology are also related to the sense of powerlessness vs control: learned helplessness and self-efficacy. The behavior of learned helplessness results from exposure to inescapable, uncontrollable negative stimuli and is characterized by a low rate of voluntary response and low ability to learn successful behaviors (Seligman, 1975). Although learned helplessness is a behavior, not a belief, in humans, there is a link between external control (a cognitive orientation) and learned helplessness (a conditioned response): the *perception* that reinforcement is not contingent on action.

Self-efficacy focuses on the individual's belief that he or she can (or cannot) effectively perform a specific action in a particular context (Bandura, 1986). The sense of personal control, in comparison, is more general, with universal application to many contexts. The degree to which people think they can or cannot achieve their goals, despite the specific nature of the actions required, has applies to many circumstances. More importantly, the sense of personal control may be the root of self-efficacy. A person with a high sense of personal control will likely try other actions if their current repertoire of behaviors is not working. New behaviors may successfully

obtain desired goals, which may in turn increase the perceived ability to shape other events and circumstances in life.

Emotional Consequences of Powerlessness

Compared to the belief that outcomes are determined by forces external to oneself, belief in personal control is associated with low levels of psychological distress (Pearlin et al., 1981; Mirowsky and Ross, 1990b, 2003; Wheaton, 1980). In fact, of all the beliefs about self and society that might affect distress, belief in control over one's own life may be the most important (Mirowsky and Ross, 2003). Perceived powerlessness is significantly associated with depression, anxiety, and anger, but its influence on depression is greatest. If one cannot influence conditions and events in one's own life, what hope is there for the future? In addition to its direct, demoralizing effect, the sense of not being in control of the outcomes in one's life diminishes the will and motivation to solve problems or avoid them, so disadvantages accumulate.

Heuristics in Psychology and Sociology

Ideally, social psychologists who study the links between social structural conditions, perceptions of control, and emotional outcomes, will synthesize the strengths of psychology and sociology – as did Rotter and Seeman–while avoiding the pitfalls. Each discipline has a heuristic, or working assumption, which greatly simplifies reality to provide a base from which to proceed with research. In the extreme, psychology assumes that beliefs come out of people's heads without reference to social conditions, whereas sociology assumes that there is nothing *but* social structure. Sociologists too often discount the ways in which perceptions mediate the effects of social position on well-being; psychologists too often discount the influence of social structure on perceptions.

Social Structural Causes of Perceived Control vs Powerlessness

In the United States, average levels of perceived control are high, and they vary systematically with positions of objective power (Mirowsky et al., 1996).

Objective Power and Perceived Control

Belief in external control is the learned and generalized expectation that one has little control over meaningful events and circumstances in one's life. As such, it is the cognitive awareness of a discrepancy between one's goals and the means to achieve them. Theoretically, social structural positions indicative of objective powerlessness, including dependency, disorder, structural inconsistency, and alienated labor, increase the probability of this discrepancy and thus increase perceived powerlessness (Avison and Cairney, 2003; Mirowsky and Ross, 2003).

Beliefs about personal control are often realistic perceptions of objective conditions. An individual learns through social interaction and personal experience that his or her choices and efforts are usually likely or unlikely to affect the outcome of a situation. Failure in the face of effort leads to a sense of powerlessness, fatalism, or belief in external control, beliefs that can increase passivity and result in giving up. Through continued experience with objective conditions of powerlessness and lack of control, individuals come to learn that their own actions cannot produce desired outcomes. In contrast, success leads to a sense of mastery, efficacy, or belief in internal control, characterized by an active, problem-solving approach to life (Wheaton, 1980, 1983; Mirowsky and Ross, 1983, 2003).

Sociological theory points to several conditions likely to produce a belief in external control. First and foremost is powerlessness. Defined as an objective condition rather than a belief, it is the inability to achieve one's ends or, alternatively, the inability to achieve one's ends when in opposition to others. The second is structural inconsistency, which is a situation in which society defines certain goals, purposes, and interests as legitimate and desirable and also defines the proper procedures for moving toward the objectives but does not provide adequate resources and opportunities for achieving the objectives through legitimate means. The third is alienated labor, a condition under which the worker does not decide what to produce, does not design and schedule the production process, and does not own the product. The fourth is dependency, a situation in which one partner in an exchange has fewer alternative sources of sustenance and gratification than the other. The fifth is role overload, a situation in which expectations of others imply demands that overwhelm the resources and capabilities of the individual. The sixth is threat – the potential for danger and harm in one's environment such as a neighborhood disorder. Although these conditions are not exhaustive, they all point to the generative force of various forms of social power. In looking for the sources of perceived powerlessness, researchers have looked for variables associated with conditions of powerlessness, structural inconsistency, alienated labor, dependency, role overload, and threat.

Among the major sociodemographic correlates of the sense of personal control are socioeconomic status, including education, income, work, and unemployment; neighborhood disadvantage and disorder; and gender and age.

Socioeconomic Status

Most research on the social structural correlates of perceived control vs powerlessness looks at socioeconomic status. Looking at specific components of socioeconomic status (SES) separately, education, family income, personal earnings, economic hardship, job status, and control at work each decrease the sense of powerlessness, adjusting for the other components (Mirowsky and Ross, 2003).

Education develops effective capacities on many levels. First, in formal education one encounters and solves problems that are progressively more difficult, complex, and subtle. The process of learning builds confidence and self-assurance, even if the things learned have no practical value. Those things, however, often do have practical value. Education also instills the habit of meeting problems with attention, thought, action, and persistence. It develops the general habits and skills of communication and analysis, plus those which are tailored to an occupation. Education develops the ability to solve problems on all these levels, and the ability to solve problems increases control over events and outcomes in life. Finally, education serves as an avenue to good jobs and high incomes. Thus it marks the social power

that helps provide control over circumstances of life (Mirowsky, 1995; Mirowsky and Ross, 1998, 2007; Schieman, 2001; Schieman and Plickert, 2008).

In school, academic achievement as measured by grades and test scores increases the sense of control between 8th and 12th grades, and the sense of control in turn boosts academic achievement (Ross and Broh, 2000). In contrast, although success in school also increases self-esteem, self-esteem, in turn, has no significant impact on academic achievement. Young adults who stay in school see an increase in the sense of personal control between the ages of 14 and 22 that persists at least to age 35, whereas those who drop out of high school do not (Lewis et al., 1999). Finally, young adults whose parents are well educated have higher levels of personal control and mastery than the children of poorly educated parents (Lewis et al., 1999).

Education acts like a stored resource, enhancing the development of personal control in young adulthood and slowing its erosion in old age (Mirowsky and Ross, 2007). Compared with the college educated, persons with a high school degree have a lower sense of control which peaks at a lower level and at a younger age, and then declines rapidly in older age. Among those with a college degree, the sense of control starts high, rises in young adulthood, stays high in middle age, and declines less in old age. The most disadvantaged life course trajectory of control is that among persons who have not finished high school. Their sense of control starts low, and never increases in young adulthood. In fact, it begins declining early, stays low, shows a little upturn around retirement (maybe from oppressive jobs), and then declines again to the lowest level of any group (Mirowsky and Ross, 2007).

Jobs are important for a number of reasons. Low-status jobs produce a sense of powerlessness because the job, and the opportunities and income it provides, are seen as barriers to the achievement of life goals (Wheaton, 1980). Jobs that are substantively complex (especially in work primarily with information and people rather than with things) increase the sense of personal control and psychological self-directedness (Kohn, 1976; Kohn and Schooler, 1982). Jobs that provide autonomy – freedom from close supervision and participation in decision-making – increase the sense of personal control (Bird and Ross, 1993; Ross and Mirowsky, 1992). Creative, nonroutine work that allows people to solve problems, learn new things, and engage in a variety of tasks also increases the sense of control (Ross, 2000). Creative and autonomous work each have independent effects on personal control, and, further, they explain the effect of working with people and data on the sense of control (Ross, 2000). Together substantively complex, nonroutine, creative, and autonomous work signals control over one's own work, which Kohn and his colleagues call occupational self-direction. Occupational self-direction – rather than ownership of the means of production or control over the labor of others – increases psychological self-direction (Kohn, 1976; Kohn and Schooler, 1982; Kohn et al., 1990). Job latitude, like occupational self-direction, includes autonomous decision-making and nonroutine work, and it significantly increases perceived control (Seeman et al., 1988). Job disruptions such as being laid off, downgraded, fired, or leaving work because of illness decrease the worker's sense of mastery, partly by lowering income and economic hardship (Pearlin et al., 1981).

Self-Estrangement and Alienated Labor

Self-estrangement is a specific type of powerlessness related to work. It is the sense that one's work is separate from oneself. In sociology, it can be traced to Marx's discussion of alienated labor. Marx was concerned about the impact of wage labor and factories on the quality of life. He argued that in working for someone else the laborer belongs to another person and not to himself. The work is a means of satisfying other needs rather than satisfying in itself; external rather than part of his or her nature; imposed rather than voluntary. Workers deny themselves in work rather than fulfill themselves. Marx argued that alienated labor makes the worker physically exhausted and mentally debased rather than developing his or her mental and physical energies. Thus, alienated labor produces a sense of misery rather than well-being. Alienated labor is the condition in which the worker does not decide what to produce, does not design the production process, and does not own the product. Self-estrangement is the sense of being separate from that part of one's thoughts, actions, and experiences given over to the control of others; of work being foreign to oneself rather than an expression of oneself. Work is seen as drudgery. There is no intrinsic gratification in work, and any rewards lie outside the activity itself (such as pay). There is no pride in work because the work is not a creative expression of oneself; thus the opposite may be called creative work (Mirowsky, 2011; Mirowsky and Ross, 2003).

Gender

Theory suggests that women have a lower sense of control over their lives than men as a result of social stratification and inequality: economic dependency, restricted opportunities, role overload, and the routine nature of housework and women's jobs. Women are more likely to do unpaid domestic work; men are more likely to work for pay. Compared to not working for pay, employment is associated with status, power, economic independence, and noneconomic rewards, for both men and women (Bird and Ross, 1993; Gove and Tudor, 1973). Both economic dependency and the disconnection between work and rewards decrease perceived control among unpaid domestic workers compared to paid workers (Ross and Mirowsky, 1992; Bird and Ross, 1993). Elder and Liker (1982) found that elderly women who had taken jobs 40 years earlier, during the Great Depression, had a higher sense of self-efficacy and lower sense of helplessness than women who remained homemakers.

Women's paid work is also more alienating than men's. Ross and Wright (1998) examined gender differences in four dimensions of work alienation – routine, nonautonomous, estranged, and isolated work; and the ways in which work alienation shapes subjective alienation, measured as the sense of powerlessness vs control. Women's work is more objectively alienating than men's, and these work characteristics shape people's sense of personal control. Nonroutine work, including task variety and problem-solving; autonomous work, including decision-making autonomy and freedom from supervision; fulfilling work; and nonisolated work all significantly positively affect the sense of personal control.

Work and Family Interactions

Overall, the employed have significantly higher average perceived control than do homemakers. Not all jobs are alike, however; nor are all household contexts of employment. Critical combinations of low pay, nonautonomous working conditions, and heavy family demands (conditions faced disproportionately by women) may negate the positive influence of employment on control (Ross and Mirowsky, 1992). Job autonomy, earnings, responsibility for household work, and other family income combine to make the association between employment and the sense of control greater for most men than for most women. Men have higher autonomy and earnings, less responsibility for household work, and lower amounts of other household income. Because of the differences in these factors, employment increases the expected sense of control most for married males, followed by nonmarried males, then nonmarried females, and finally married females. For married women, the typical combination of low pay, low autonomy, high responsibility for household chores, and high family income other than personal earnings nearly negates the positive association between employment and the sense of control.

The gender gap in perceived control is much greater for older persons than for younger. Compared with their male counterparts, older women face more educational, employment, economic, and health disadvantages than do younger women. Older people have lower levels of personal control than do younger, and this is especially true for women. Older women are more disadvantaged than younger because they experienced a lifetime of disadvantages in work and economic conditions that cumulated with age, and because women in older generations had fewer employment and educational opportunities and less equality in unpaid work at home than do younger women. The dynamics of aging, the differences among cohorts, and the link between the two combine to predict more gender equality among young adults than older.

Age

Research using large representative samples of persons across the full range of adult ages, from 18 to 90, shows that older adults have a lower sense of control than do younger adults, and that perceived control decreases with age at an accelerating rate (Mirowsky, 1995; Schieman, 2001; Turner and Lloyd, 1999). Slight increases in young adulthood seen in cross-sectional studies are actually larger when cohort differences are taken into account: control increases in young adulthood and then declines at an accelerating rate as people age into their sixties, seventies, and eighties (Mirowsky and Ross, 2007). Perceived powerlessness in old age results from increasing levels of physical disability and poor health, widowhood, and retirement, all of which are associated with aging (Mirowsky, 1995; Ross and Drentea, 1998). In addition to the effects of aging, older cohorts have lower levels of education than younger (Mirowsky, 1995; Mirowsky and Ross, 2007; Wolinsky and Stump, 1996; Wolinsky et al., 2003).

Neighborhood Disadvantage and Disorder

Neighborhood disadvantage and disorder are subjectively alienating (Ross and Mirowsky, 2009). Socioeconomically disadvantaged neighborhoods where few residents have college degrees, poverty, and mother-only families are common, home ownership is low, and unemployment is high, are prone to high levels of neighborhood disorder. Neighborhood disorder signals the potential for harm. Signs of disorder in one's neighborhood signify collective threat. Graffiti, noise, vandalism, dirty streets, public drinking, trash on the streets, and run-down buildings are cues. These cues tell residents about themselves and the world they live in. They indicate the breakdown of social control and order. Through exposure to threatening social conditions residents learn that they are relatively powerless. Lack of real control over desired outcomes – like a safe and nonthreatening place to live – leads to the perception that one is powerless to control one's life. (Geis and Ross, 1998; Ross, 2011; Ross and Mirowsky, 2009).

Mistrust

Neighborhood Disorder and a Structural Theory of Mistrust

Research on neighborhood disorder led to a new emphasis on mistrust as an important element of subjective alienation. Neighborhood disorder fosters mistrust. Mistrust arises from social conditions where resources and opportunities are scarce; and disorder, crime, threat, and danger are common. When resources are scarce, everyone competes for an inadequate pool, some individuals will try to take whatever they can get at other people's expense, and the consequences of losing what little one has will be devastating. Scarcity tempts individuals to victimize and exploit others. People with few resources compete for a limited pool. They generally seem willing to exploit others for personal gain. Victimization and exploitation appear inevitable and common, which lead to mistrust.

Trust is a belief in the integrity of other people. Trusting individuals have faith and confidence in other people and expect that they can depend on others (Rotter, 1980). Mistrust, the opposite of trust, is the cognitive habit of interpreting the intentions and behavior of others as unsupportive, self-seeking, and dishonest (Mirowsky and Ross, 1983). Suspicion of other people is the central cognitive component of mistrust. Trust and mistrust express inherently social beliefs about relationships with other people. They embody learned, generalized expectations about other people's behaviors that transcend specific relationships and situations (Barber, 1983; Kramer, 1999; Rotter, 1971).

In neighborhoods where residents report high levels of disorder, people face a threatening environment characterized by danger, crime, and incivility, which promotes mistrust. Residents may come to view those around them with suspicion, as enemies who will harm them rather than as allies who will help them. Mistrust makes sense where dangers abound (Ross et al., 2001). Mistrust, in turn, reduces social support and increases psychological distress, including depression, anxiety, and anger.

Social Support and Trust

Social support depends on trust. Social support is the opposite of social isolation – a type of alienation from others.

Social support is the commitment, caring, advice, and aid provided in personal relationships, the sense of being cared for and loved, esteemed and valued as a person, and part of a network of communication and obligation (Ross and Mirowsky, 1989; Mirowsky and Ross, 2003). Trust allows pairs of individuals to establish cooperative relationships whenever doing so is mutually beneficial. Coleman (1990) emphasizes trust as an element of social capital, because trusting social relationships help produce desired outcomes. People who trust others form personal ties and participate in voluntary associations more often than do mistrusting individuals (Paxton, 1999). Trusting individuals are themselves more trustworthy and honest and are less likely to lie and harm others so that they create and maintain environments of trustworthiness. They enter relationships with the presumption that others can be trusted until they have evidence to the contrary (Rotter, 1971, 1980). Because trusting individuals can form effective associations with others, the presumption of trust can be an advantageous strategy, despite the fact that expecting people to be trustworthy is risky.

Pathways Linking Neighborhood Disorder to Distress

Mistrust and powerlessness are the two big links between disorder and distress. Two other types of subjective alienation – normlessness and social isolation – are not. Neighborhood disorder promotes normlessness, but normlessness has little effect on distress. Social isolation has trade-offs at all stages. Social isolation is associated with increased depression, but decreased anger and anxiety. Furthermore, apart from a small indirect effect on isolation by way of mistrust, neighborhood disorder is associated with more social ties with others, not fewer. Powerlessness and mistrust consistently link neighborhood disorder to distress. Of the two, mistrust forms the larger link.

Mistrust was not originally included as a type of subjective alienation by Seeman despite the fact that it clearly fits the definition of subjective alienation as separation from oneself or others. Instead, social psychologists originally considered mistrust as an element of authoritarianism (along with inflexibility), but its importance for psychological distress may come from the deep alienation it represents (Mirowsky and Ross, 2003). Mistrusting individuals are not simply isolated from other people, they are deeply suspicious of others. Because they think other people are likely to harm them, they feel they cannot depend on others, turn to others when they need help, confide in others, or rely on them. To mistrusting individuals, any social interaction carries risk.

What are the sources of mistrust? Mistrust arises when conditions are threatening and resources are scarce. Neighborhood disorder indicates collective threat, which produces mistrust. Furthermore, disadvantaged individuals tend to live in disadvantaged neighborhoods with high levels of disorder. In addition to the collective effects, disadvantaged individuals with fewer resources have higher levels of mistrust. High levels of economic hardship, low levels of education, not being employed or married, and having been a victim of crime are all significantly associated with mistrust. The only exception to this pattern relates to gender. Men typically have more resources than women, yet men are significantly more mistrusting. In general, though, the powerless and disadvantaged are more mistrusting than the powerful and advantaged.

Powerlessness and Mistrust: Directions for Future Research

Mirowsky and Ross concluded that a low sense of control over one's own life creates the primary link between psychological distress and individual disadvantaged social statuses such as low education, economic hardship, unemployment, unstable employment, and poorly paid, tedious, routine, and oppressive jobs. Perceived powerlessness links individual socioeconomic disadvantage to distress. It is possible that mistrust creates a major link between psychological distress and collective disadvantage. Perceived powerlessness is the cognitive awareness of an inability to achieve individual goals. It is a separation from self – from important outcomes in one's own life. Mistrust represents social alienation – a separation from others. A promising direction for future research is the idea that neighborhood context is linked to distress by way of social alienation more than alienation from oneself. Social cues of disorder like people hanging out on the streets, drinking or taking drugs, making trouble, and heightening the sense of danger are cues about other people in one's neighborhood. Even the physical cues of neighborhood disorder like graffiti, vandalism, run-down buildings, and trash on the streets are ultimately cues about the activities of people in one's neighborhood, because a person broke the window or spray-painted the graffiti. The collective threat implied by disorder is created by other people, and may lead to the social alienation of mistrust more than to the personal alienation of powerlessness. Since the research by Ross and Mirowsky (2009) is the first attempt to link distress to neighborhood disorder through of a number of different types of subjective alienation future research should attempt to confirm or disconfirm this idea.

See also: Aging and Health in Old Age; Deprivation: Relative; Inequality, Social; Marxism in Contemporary Sociology; Medical Sociology; Neighborhood Effects; Power in Society; Social Psychology, Sociological; Social Stratification; Social Structure; Work, Sociology of.

Bibliography

Avison, W.R., Cairney, J., 2003. Social structure, stress, and personal control. In: Exert, Steven H., Pearlin, Leonard I., Schaie, K. Warner (Eds.), Personal Control in Social and Life Contexts. Springer, New York, pp. 127–164.

Bandura, A., 1986. Social Foundations of Thought and Action. Prentice-Hall, Englewood Cliffs, New Jersey.

Barber, B., 1983. The Logic and Limits of Trust. Rutgers University Press, New Brunswick, NJ.

Bird, C.E., Ross, C.E., 1993. Houseworkers and paid workers: qualities of the work and effects on personal control. Journal of Marriage and the Family 55, 913–925.

Coleman, J.S., 1990. Foundations of Social Theory. Harvard University Press, Cambridge, MA.

Downey, G., Moen, P., 1987. Personal efficacy, income and family transitions: a longitudinal study of women heading households. Journal of Health and Social Behavior 28, 320–333.

Elder, G.H., Liker, J.K., 1982. Hard times in women's lives: historical influences across forty years. American Journal of Sociology 88, 241–269.

Fischer, C.S., 1976. The Urban Experience. Harcourt Brace Jovanovich, New York.

Gecas, V., 1989. The social psychology of self-efficacy. Annual Review of Sociology 15, 291–316.

Geis, K.J., Ross, C.E., 1998. A new look at urban alienation: the effect of neighborhood disorder on perceived powerlessness. Social Psychology Quarterly 61, 232–246.

Gove, W.R., Tudor, J., 1973. Adult sex roles and mental illness. American Journal of Sociology 78, 812–835.

Kohn, M.L., 1976. Occupational structure and alienation. American Journal of Sociology 82, 111–130.

Kohn, M.L., Naoi, A., Schoenbach, C., Schooler, C., Slomczynski, K.M., 1990. Position in the class structure and psychological functioning in the United States, Japan, and Poland. American Journal of Sociology 95, 964–1008.

Kohn, M.L., Schooler, C., 1982. Job conditions and personality: a longitudinal assessment of their reciprocal effects. American Journal of Sociology 87, 1257–1286.

Kramer, Roderick M., 1999. Trust and distrust in organizations: emerging perspectives, enduring questions. Annual Review of Psychology 50, 569–598.

Lewis, S.K., Ross, C.E., Mirowsky, J., 1999. Establishing a sense of personal control in the transition to adulthood. Social Forces 77, 1573–1599.

Mirowsky, J., 1995. Age and the sense of control. Social Psychology Quarterly 58, 31–43.

Mirowsky, J., 2011. Wage slavery or creative work? Society and Mental Health 1, 73–88.

Mirowsky, J., Ross, C.E., 1983. Paranoia and the structure of powerlessness. American Sociological Review 48, 228–239.

Mirowsky, J., Ross, C.E., 1990a. Control or defense? Depression and the sense of control over good and bad outcomes. Journal of Health and Social Behavior 31, 71–86.

Mirowsky, J., Ross, C.E., 1990b. The consolation prize theory of alienation. American Journal of Sociology 95, 1505–1535.

Mirowsky, J., Ross, C.E., 1991. Eliminating defense and agreement bias from measures of the sense of control: a 2×2 index. Social Psychology Quarterly 54, 127–145.

Mirowsky, J., Ross, C.E., 1996. Fundamental analysis in research on well-being: distress and the sense of control. The Gerontologist 36, 584–594.

Mirowsky, J., Ross, C.E., 1998. Education, personal control, lifestyle and health: a human capital hypotheses. Research on Aging 20, 415–449.

Mirowsky, J., Ross, C.E., 2003. Social Causes of Psychological Distress, second ed. Aldine Transaction, N.J.

Mirowsky, J., Ross, C.E., 2007. Life course trajectories of perceived control and their relationship to education. American Journal of Sociology 112, 1339–1382.

Mirowsky, J., Ross, C.E., Van Willigen, M., 1996. Instrumentalism in the land of opportunity: socioeconomic causes and emotional consequences. Social Psychology Quarterly 59, 322–337.

Paxton, P., 1999. Is social capital declining in the United States? A multiple indicator assessment. American Journal of Sociology 105, 88–127.

Pearlin, L.I., Menaghan, E.G., Lieberman, M.A., Mullan, J.T., 1981. The stress process. Journal of Health and Social Behavior 22, 337–356.

Ross, C.E., 2000. Occupations, jobs and the sense of control. Sociological Focus 33, 409–420.

Ross, C.E., 2011. Collective threat, trust, and the sense of personal control. Journal of Health and Social Behavior 52, 287–296.

Ross, C.E., Broh, B.A., 2000. The roles of self-esteem and the sense of personal control in the academic achievement process. Sociology of Education 73, 270–284.

Ross, C.E., Drentea, P., 1998. Consequences of retirement activities for distress and the sense of personal control. Journal of Health and Social Behavior 39, 317–334.

Ross, C.E., Mirowsky, J., 1989. Explaining the social patterns of depression: control and problem-solving – or support and talking. Journal of Health and Social Behavior 30, 206–219.

Ross, C.E., Mirowsky, J., 1992. Households, employment, and the sense of control. Social Psychology Quarterly 55, 217–235.

Ross, C.E., Mirowsky, J., 2009. Neighborhood disorder, subjective alienation, and distress. Journal of Health and Social Behavior 50, 49–64.

Ross, C.E., Mirowsky, J., Pribesh, S., 2001. Powerlessness and the amplification of threat: neighborhood disadvantage, disorder, and mistrust. American Sociological Review 66, 568–591.

Ross, C.E., Wright, M.P., 1998. Women's work, men's work and the sense of control. Work and Occupations 25, 33–55.

Rotter, J.B., 1966. Generalized expectancies for internal vs. external control of reinforcements. Psychological Monographs 80, 1–28.

Rotter, Julian B., 1971. Generalized expectancies for interpersonal trust. American Psychologist 26, 443–452.

Rotter, Julian B., 1980. Interpersonal trust, trustworthiness, and gullibility. American Psychologist 35, 1–7.

Schieman, S., 2001. Age, education, and the sense of control: a test of the cumulative advantage hypothesis. Research on Aging 23, 153–178.

Schieman, S., Plickert, G., 2008. How knowledge is power: education and the sense of control. Social Forces 87, 153–165.

Seeman, M., 1959. On the meaning of alienation. American Sociological Review 24, 783–791.

Seeman, M., 1983. Alienation motifs in contemporary theorizing: the hidden continuity of classic themes. Social Psychology Quarterly 46, 171–184.

Seeman, M., Seeman, A.Z., Budros, A., 1988. Powerlessness, work, and community: a longitudinal study of alienation and alcohol use. Journal of Health and Social Behavior 29, 185–198.

Seligman, M.E.P., 1975. Helplessness. Freeman, San Francisco.

Thoits, P.A., 2006. Personal agency in the stress process. Journal of Health and Social Behavior 47, 309–323.

Turner, R.J., Lloyd, D.A., 1999. The stress process and the social distribution of depression. Journal of Health and Social Behavior 40, 374–404.

Wheaton, B., 1980. The sociogenesis of psychological disorder: an attributional theory. Journal of Health and Social Behavior 21, 100–124.

Wheaton, B., 1983. Stress, personal coping resources, and psychiatric symptoms: an investigation of interactive models. Journal of Health and Social Behavior 24, 208–229.

Wolinsky, F.D., Stump, T.E., 1996. Age and the sense of control among older adults. Journal of Gerontology: Social Sciences 51B, S217–S220.

Wolinsky, F.D., Wywrich, K.W., Babu, A.N., Kroenke, K., Tierney, W.M., 2003. Age, aging, and the sense of control among older adults: a longitudinal reconsideration. Journal of Gerontology: Social Sciences 4, S212–S220.

Alienation, Sociology of

Christopher Adair-Toteff, University of South Florida, Tampa, FL, USA; and University of Kent, Canterbury, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by F. Gyer, volume 1, pp. 14549–14554, © 2001, Elsevier Ltd.

Abstract

Alienation is the feeling of an unnatural separation or a sense of radical disconnectedness. It is a feeling of homelessness and lacking roots and foundation. It is a feeling that one has little or no worth; it is a belief that life has no meaning. While the sense of alienation is found throughout history, it is found predominately in the West as a result of industrialization and the rise of modern capitalism.

Alienation: An Introduction

Alienation is the sense of unnatural separation or radical disconnectedness; it is a sense of homelessness and uprootedness. An alienated individual often believes that he or she lacks meaning or worth. Marx is usually credited with providing the most influential account of alienation, but its antecedents are found in Hegel. Moreover, Marx rejected the idealistic versions and instead developed a more economy-oriented account of self-alienation. Ferdinand Tönnies and especially Georg Simmel developed significant sociological explications of the concept of alienation and Karl Jaspers contributed important contemporary cultural analyses of it. Having been mentored by Jaspers (as well as by Martin Heidegger), Hannah Arendt returned to a more philosophical account of world alienation. Each of these twentieth century thinkers explored its crucial impact on modern culture.

Philosophical Alienation

It was in the 1960s that the concept of alienation became a popular subject. In his important and lengthy introduction to Richard Schacht's indispensable book *Alienation*, Walter Kaufmann explains how the notion of alienation came to be so popular. He suggested that it was a Cold War intersection between existentialism and Marxism and he traces it back to Erich Fromm's translation of Marx's *Economic and Philosophical Manuscripts of 1844*. These were first published in 1932 and were immediately recognized as providing a humanist conception of Marx, but World War II effectively made much discussion of it impossible. By the 1950s and 1960s the time was finally ripe for scholarly examinations of Marx's early writings. Kaufmann also points out that the young Marx was heavily influenced by Hegel and nowhere is this as evident as with the concept of alienation (Kaufmann, 1970: xvii–xxiii).

Georg Wilhelm Friedrich Hegel was the first scholar to utilize the notion of alienation, which he did in his *Phenomenology of Spirit*. This work is both brilliant and puzzling; it is brilliant in that Hegel sets himself the task to trace the forms of spirit throughout history. Hegel was not a historian, so he was not interested in historical accuracy per se. Instead, he sought to provide an ideal account of spirit in its historical manifestations. It is a puzzling work because of Hegel's audacity in writing it, and even more so in his choice of words. Translated into English Hegel's terminology verges on being strange; in German they seem somewhat better. This is because Hegel uses words in new and different ways; for example, he uses the term 'aufheben' to mean two almost contradictory processes. First, there is the sense of 'to cancel' and second, there is the sense 'to lift up.' By itself, the term seems not to make much sense; but when one understands that Hegel regards the process of the forms of the unfolding spirit as dialectical that it becomes more comprehensible. Accordingly, there is a process in which one form is countered by its opposite, and then both the one and its counterpart cancel each other out and then are together lifted up to the next stage in the process. Hegel's dual use of words is also found with respect to his notion of 'alienation' ('Entfremdung'), which is one of the most central terms of the *Phenomenology*. Richard Schacht notes that Hegel uses 'alienation' in two ways: first as 'separation' and second as 'surrender' (Schacht, 1970: 38, note 1; 43–44). Taken as a whole, the entire *Phenomenology* could be read as a treatise on alienation – the various stages in which humankind is alienated from itself. But, Hegel also devotes two specific sections to the notion of alienation.

In the lengthy section entitled 'Self-Alienated Spirit. Culture' Hegel writes of the alienation of the 'objective' essence and the 'subjective' self (Hegel, 1977: 295). By recognizing that there is something other than the subjective realm of faith, the self becomes alienated. The 'objective' realm of knowledge confronts the subjective belief in faith, which leads not only to doubt but also to the alienation of oneself as a natural being (Hegel, 1977: 297–298). On a more general level there is another process of alienation and opposition between the objective and 'self-less' state and the 'subjective' and 'selfish' striving for wealth (Hegel, 1977: 306, 313). While it seems obvious which side is 'good' and which side is 'bad,' Hegel suggests that this is only an apparent opposition and is part of 'pure culture' (Hegel, 1977: 316–317, 321). This opposition mirrors the opposition between faith and knowledge and leads to 'delusion and deception'; however, faith and enlightenment 'reconcile' at a certain level. This apparent reconciliation then provides the starting stage for the next stage of Hegel's dialectical process (Hegel, 1977: 335, 346).

In the *Phenomenology* there is an earlier section, which not only offers an explication of alienation, but one that will provide considerable material for Marx. This is the much

smaller section entitled 'Lordship and Bondage.' There are two 'individuals' who confront each other. This encounter is problematic because while they must acknowledge that there is another, this recognition leads to a struggle – who recognizes whom? This struggle is the struggle for the certainty of one's self – Hegel writes that only by staking one's life is freedom won (Hegel, 1977: 114). Of course, there can be only one 'winner,' the 'master' while the loser becomes the 'slave.' The master now has consciousness of one's own self, thus seems superior to the slave, who must now be conscious of the master. However, this superiority is illusionary; the slave engages in work, and by working on things he becomes aware of himself as a self-conscious being. In contrast, the master who does not work knows only the desire for things; thus the master is dependent on the slave. It is the activity of work that defines and raises up the slave, while the lack of activity and the burden of desire subjugates the master. It is Hegel's notion of work that will become important for Marx's notion of man.

Hegel's conceptions of alienation are intriguing but they suffer from difficulties. His dual senses of alienation cause confusion – from Hegel's general philosophical point of view they seem to be correct, but when viewed specifically they seem to be contradictory. More importantly, Hegel could not solve the problem of alienation because he abstracted and generalized both 'work' as well as 'needs' (Löwith, 1941: 350–352). Despite the importance of Hegel's account of alienation, it remained largely neglected until 1932 with the publication of Marx's *Economic and Philosophical Manuscripts* (Schacht, 1970: 73).

Economical Alienation

Karl Marx is best known for his writings on capitalism and labor; he specialized in discussing material conditions and modes of productions, of class, and value. However, as Karl Löwith pointed out, Marx was not interested in general conditions. Nor was he interested in theory and contemplation; rather, his concern was with practice and action. This is especially true regarding Marx's opinion of Hegel's concept of alienation. While he acknowledged its importance, Marx's major criticism was that it was too philosophical and too abstract (Löwith, 1952: 71). In contrast to Hegel's concern with abstract spirit, Marx's focus is on man as such. Bertell Ollman also noted that Marx's major concern was with the nature of man, and in Ollman's view, that meant dwelling on the nature and effects of alienation (Ollman, 1976: xi). The concept of alienation is, as István Mészáros has suggested, a key concept that can be found throughout Marx's writings (Mészáros, 1972: 66). However, as Allen Wood has argued, it has two distinctly different roles to play. In Marx's early writings, the concept of alienation is crucial for three reasons: Marx discusses the irrationalities of modern life; he argues that one of the most problematic of these irrationalities is found in modern labor; and this type of labor is historically determinant and also transitory. Alienation is part of the path of human development. Wood goes on to suggest that while Marx's early conception of alienation is important, it lacks persuasion because it is "too vague and metaphorical to perform the explanatory function that Marx tries to assign it" (Wood, 1981: 71). Wood argues further that Marx's later conception of alienation succeeds because he no longer tries to use alienation as an explanation; rather, he uses it as a description. Whether these accounts fail or succeed, is not the issue here; the issue here is simply to provide an overview of Marx's conceptions. Marx's early account of alienation can be broken down into four different variations of alienation: between the individual and one's labor; between the individual and one's products; between the individual and one's fellow human beings; and between the individual and one's species. Marx believed that, unlike political economy accounts, his account would explain man's condition. And, unlike the accounts of political economists, which begin with a 'legendary primordial condition' Marx will begin his account with a *contemporary* fact (Marx, 1964: 120–121).

Marx's notion of labor is crucial to his entire thinking because it is a fundamental fact of human life: humans must produce (Schacht, 1970: 87; Mészáros, 1972: 76–77, 80). Because labor is one of the most important facets of human life, Marx suggests that alienated labor is the most important of all of the types of alienation.

Marx begins with an economic fact: the more wealth the worker produces, the poorer the worker becomes. In other words, the human world is devalued as the value of the world of things increases. This is because the labor of the worker not only produces goods; but it also produces the worker as a commodity (Marx, 1964: 121, 129). This may not say much by itself, but Marx clarifies this by insisting that the worker produces the object, which is an alien one – one which is independent and opposed to the worker. This is the 'objectification of labor,' which makes the laborer subservient to the object – hence, the alienation (Marx, 1964: 121–122). Marx emphasizes the growing alienation; the greater the worker's activity the less he possesses, the greater the product the more he is diminished. The object exists independently of the worker who made it; it is now an alien and autonomous power (Marx, 1964: 123–124). In a moving passage Marx writes: "the more the worker produces the less he has to consume; the more value he creates the more worthless he becomes; the more refined his product the more crude and misshapen the worker; the more civilized the product the more barbarous the worker …." (Marx, 1964: 124).

Marx moves from the type of alienation that results from the product to the type that comes from the process; that is, from the activity of the work itself. Many scholars writing on capitalism contend that 'formally free labor' was a necessary prerequisite for the development of modern capitalism; which would have been impossible to develop with 'slave labor.' Marx takes the opposite view; that in capitalism, work is not voluntary but is imposed – it is 'forced labor.' The worker must work for another, not for himself. The worker works only in order to satisfy needs and in work the worker feels alienated. It is only during those few hours at home that the worker has the sense of leisure and feels less alienated (Marx, 1964: 125–126).

The third and fourth types of alienated labor can be examined together because they no longer pertain to just the individual and its object; now it is concerning the species and other members of the human race. Marx's basic point is that humans are not isolated creatures but exist in a social setting. Accordingly, the alienation of one's self is reflected in the alienation to other human beings (Marx, 1964: 129). Ollman explains

this: first that this form of alienation diminishes, if not eliminates, the human advantage over other animals; and second it leaves 'man as a rump of himself.' Some commentators contend that alienation causes people to lose the capacity to be fully human, while others maintain that alienation causes the worker to lose one's entire sense of being human (Ollman, 1976: 151–152; Mészáros, 1972: 146–149).

Marx traces the problems of alienation to two things: money and private property (Marx, 1964: 121, 131). Money is a problem because it produces a whole system of alienation – which includes monopoly and competition, labor and private property, and "the whole system of money" (Marx, 1964: 121). Private property is a problem because it is a direct result of alienated labor. But, it is not only the basis and cause of alienated labor; alienated labor becomes the cause of private property (Marx, 1964: 132). The elimination of private property will emancipate the workers, thus eliminating alienation. Marx does not expand upon this topic; the manuscript suddenly ends (Marx, 1964: 133–134). We do know that Marx was convinced that there will be a revolution and the result will be socialism. In this socialist paradise there will be no state, no classes, and even no labor: in "the realm of freedom," "labor will be abolished" (Arendt, 1968: 22, 24). And, with labor having been abolished, alienation will also disappear.

There is considerable disagreement about the nature of Marx's conception of alienation and his notion of how it will be eliminated; but there is virtually no disagreement about its massive influence. Many scholars would disagree about how rigorous this conception is; but they would probably all agree that every subsequent account of alienation is in some way indebted to that of Marx.

Sociological and Cultural Alienation

Although Ferdinand Tönnies had engaged in an intensive study of Marx, he never developed a theory of alienation. Nonetheless, his sociological writings often indicate a great awareness of the problem of alienation in modern culture. That is why his distinction between 'Gemeinschaft' ('community') and 'Gesellschaft' ('society') is so important. In *Gemeinschaft und Gesellschaft* Tönnies contrasts the healthy, essential, and organic older community with the unhealthy, arbitrary, and artificial modern society. The community was an 'economic–communistic arrangement' but the society is based exclusively on money and rational exchange (Tönnies, 1991: 31, 39, 44). Tönnies quotes Marx's claim that in the capitalistic system everyone must possess an encyclopedic knowledge of goods; but he also quotes Adam Smith's observation that a merchant "is not necessarily a citizen of any particular country" (Tönnies, 1991: 48). Everything and everyone in society is reduced to a calculable system of intentions, means, and goals (Tönnies, 1991: 93). Modern society is an alienated society; everyone is cut off from one another. It resembles Hobbes' state of nature where there is a "war of all against all" (Adair-Toteff, 1995: 60). There is little question that Tönnies approves of the old community and disapproves of modern society: the great metropolis is the place of capital, money, and artificial society, but the country village is the home for family and friends (Tönnies, 1991: 212). The village is the place for closeness and togetherness; the large city is the place for alienation and cold power (Tönnies, 1991: 215–217). The village is the place of the 'security blanket,' but it lacks the precision that makes progress possible. Moreover, it is not as idyllic as Tönnies makes it seem: its 'conformity and uniformity' could be just as stifling and alienating as the modern city (Segalman, 1981: 43–47).

Where Tönnies recognized alienation as being the problem for one of the two major sociological groups, Georg Simmel saw alienation as being the great problem for the individual. Simmel regarded modern life as being full of conflict and tragedy (Simmel, 1918: 14, 35). In the past, the age-old problem of alienation from God had been minimized to some extent by religious faith and tradition; Simmel writes of the mystical union of a Meister Eckhart: "my eye and God's eye is one eye" (Simmel, 1989: 14). However, modern culture had lost the power of belief and rejected the weight of tradition; instead, it rushed into a restless rationality (Simmel, 1986: 199). Simmel was one thinker who truly captured the tragic spirit of oppositions – of being and becoming, of subject and object, of form and content, of reality and illusion (Simmel, 1986, 1989: 195). And, he was certainly one of the most poetic thinkers to express the feeling of helplessness, of being lost, and of the boundless sense of strangeness (Bauer, 1962: 20). Simmel described the modern epoch as one, which had become 'overripe and tired,' which was filled with discrepancies and discordances – what he called a "tragedy of culture" (Simmel, 1986: 203, 208, 210, 215).

Simmel offers a more specific account of alienation toward the end of *The Philosophy of Money*, as helpfully noted by Jorge Arditi. Unfortunately, instead of focusing on this he concentrates on Simmel's emphasis on irrationality (Arditi, 1996: 94–95). What Arditi does do, is to hint at the prominence that the notion of distance plays in Simmel's thinking (Arditi, 1996: 990). It is found in the section of *Soziologie* devoted to the stranger; it is the stranger's burden to be physically removed from others. Because of distance, he does not share our roots, and by implication he does not share our values. In short, the stranger is not one of us (Simmel, 1971: 145). It is also found in the short work on cities; while the city is crowded with people, each suffers in isolation. This 'mutual strangeness and revulsion' arises because of the infinite complexity of city life (Simmel, 1971: 331). This pervasive 'conflict with human beings' is rooted in the relentless pursuit of money; one becomes a mere 'cog' (Simmel, 1971: 326–327, 337). Above all, the notion of distance plays the most significant role in Simmel's discussion of alienation in the *Philosophie des Geldes*. In the chapter *Style of Life*, Simmel emphasizes that calculation is the defining characteristic of modern culture. Everything is measured, weighed, and exactingly calculated which leads to a growing discrepancy between individual and object. The growing separation between worker and the means of production is mirrored in the increasing distance between the 'creating personality' and the 'created artifact' (Simmel, 1900: 472, 478, 486). Simmel regards this emphasis on 'form, function, and the quantitative' as part of the growing objectification of culture, with the resulting decrease in the importance of the person (Simmel, 1900: 505).

Friedrich Nietzsche did not set out to write explicitly about alienation; however, he traced the inversion and alienation of natural values, which he believed led to the perversion and decline of Western civilization. Nietzsche's unusual task as set out in the *Genealogy of Morals* is to examine the 'value of values' (Nietzsche, 1999c: 253). He examines how the terms 'good' and 'bad' were transformed into 'good' and 'evil' (Nietzsche, 1999c: 261–265). Originally, it was the natural domination of the strong over the weak, the elites over the masses. As he described it in *Beyond Good and Evil*, the noble ones sought danger and believed in risk, but, with the 'slave uprising' of the Jews, followed by Christian belief in salvation, and by the platonizing of Christianity, these values were inverted (Nietzsche, 1999b: 82). Now, what was valued was 'modesty' and 'obedience' (Nietzsche, 1999c: 281, 287). And, now the operative words were 'guilt,' 'conscience,' and 'duty' (Nietzsche, 1999c: 300–306, 318–321, 330–332). Instead of the natural and instinctive enjoyment of the body and the senses, Christianity introduced the ascetic denial and denunciation of all things worldly; as Nietzsche says 'living against life' (Nietzsche, 1999c: 342–344, 361, 365). This is an alienating sickness of weakness, a sickness of suffering against all that is high and noble. It is the herd instinct against the individual (Nietzsche, 1999c: 370–373, 384). In the preface to *Zarathustra*, Zarathustra speaks of man's overcoming his limitations and becoming 'der Übermensch,' but he is met with laughter and derision by the crowd. They reject the idea that God is dead; instead they believe in a 'herd without a shepherd.' To the 'last men' the world has become smaller and warmer, and they believe that they have found 'happiness' (Nietzsche, 1999a: 20). What they have found is the belief in inverted values and alienated life.

Nietzsche was not content to diagnose the problems with the Judeo-Christian beliefs; he also wanted to point to a way to overcome these self-inflicted problems. The main source for Nietzsche's path to health is found in *Thus Spake Zarathustra*. Nietzsche points to the coming of the 'Übermensch,' the one who 'overcomes' one's self. In the section on *The Stillest Hour*, a 'voice without a voice' approaches Zarathustra and urges him to 'become like a child' and to accept what is in him (Nietzsche, 1999a: 187–190). Later, Zarathustra is told 'who he is and who he must become': the teacher of the doctrine of the eternal recurrence (Nietzsche, 1999a: 275). It is the rejection of the illusory promise of the afterlife and is the embracing of the here and now, in its greatness and its smallness (Nietzsche, 1999a: 275–277).

In *Reason and Existence*, Karl Jaspers points to the most important point in Nietzsche's philosophy – "God is dead" (Jaspers, 1947: 15). With the death of God we are free to reject alienation; to become who we are. It is not surprising that Jaspers wrote one of the most significant books on Nietzsche (Jaspers, 1981). He not only brought a philosopher's way of examining Nietzsche's theories, but his training as a psychiatrist helped him understand Nietzsche as a troubled individual. In turn, this helped Jaspers to comprehend the individual's sense of alienation in the modern world. Following his friend and mentor Max Weber, Jaspers believed that rational calculability and modern technology defined the twentieth century (Jaspers, 1999: 4). People were rarely only individuals; instead they belonged to the masses. Unfortunately, as Jaspers indicated, when a person is part of a mass, he or she is no longer a person. As part of the masses, there is no real existence and there is no real faith (Jaspers, 1999: 36–37). Instead, modern existence is dominated by advertisements; everything must be a 'sensation.' There is the constant overcoming of daily life through newspapers, radio, film, and travel. There is also the permanent separation between work and enjoyment (Jaspers, 1999: 42–43). One is ripped from one's roots and has lost one's essence. The masses are dominated by the faceless bureaucracy, which is hidden in a 'ghostlike invisibility.' In such a life one has mostly lost the sense of real responsibility. What one does have is fear and anxiety ('angst') for the future; and this 'angst' manifests itself in physical and psychical illnesses (Jaspers, 1999: 45–47, 55–60). One either becomes apolitical or one becomes blindly obedient; but all have lost the sense of value and worth (Jaspers, 1999: 73, 83). Whereas science once commanded respect, modern technology offers only manipulation. While philosophy had once promised understanding of humans, modern sociology now provides some comprehension of individuals and their interrelations (Jaspers, 1999: 128–130, 141). This answer hints at Jaspers' belief in an answer to the 'radical crisis.' But, it is not Marxism, psychology, or racial theory – it is Jaspers' own "philosophy of existence" (Jaspers, 1999: 146–148). Jaspers acknowledges his philosophical indebtedness to Nietzsche with the phrases like "becoming who we are" and 'self-becoming' (Jaspers, 1999: 152, 166). But, his method for overcoming alienation is his own; it is reseeing the world in a new light and rediscovering true friendship (Jaspers, 1999: 171, 177). One cannot escape the modern situation by moving away from oneself; instead, one must learn to live with oneself. Jaspers' answer to the problem of alienation in the modern world is to take up the task to "remember what it is to be human" (Jaspers, 1999: 183, 194).

The picture that Hannah Arendt painted of human beings and the human condition is often one of totalitarianism and terror. In works such as *The Origins of Totalitarianism*, *Eichmann in Jerusalem*, and in *Past and Future*, Arendt spelled out in great detail the failings of humankind. These all lead to a type of 'world alienation' in which we are cut off from our past and have no firm standards from which to judge. Arendt often wrote on the pressing topics of the day: the trial of Adolf Eichmann, the Vietnam War, and Watergate. But, like her mentor Jaspers, she looked to the great thinkers of the past, not in order to copy their thoughts and beliefs, but to see how they could provide a foundation for our own thinking and judgments. She understood that people tended to seek refuge from painful reality, but she was convinced that we have the obligation to ourselves and to others to confront the facts. In one of her last lectures she said: "When the facts come home to roost, let us try to make them welcome. Let us try not to escape into some utopias – images, theories, or sheer folly." (Arendt, 2003: 275). She may very well had Marx in mind when she wrote that; while she regarded him highly as a thinker who broke with tradition, she rejected his promotion of what she believed to be self-illusions (Arendt, 1968: 24, 28–30). We may be alienated from the world and from our labor, but the most personal form of alienation is the intentional alienation from ourselves.

Conclusion

This account of alienation was necessarily a selective one. There are some scholars who believe that forms of alienation are found in ancient and medieval times, but most agree that it is a modern phenomenon. There are many factors for this, but they would include the increasing rationality and decreasing religious beliefs. They would also include the increase in urbanization and most certainly the massive technological advances. These and other factors undoubtedly help human beings in many physical ways, but there is also no question that they increase the sense of distance from our 'natural' selves. With technical progress comes human alienation.

This account has focused almost exclusively on the conception of alienation; but the theme of alienation is also found in theology (Eric Fromm, Paul Tillich), psychology (Karen Horney, Franz Fanon), and existential literature (Albert Camus, Jean-Paul Sartre). Marx gave us his conception of alienation more than 150 years ago, but it still serves as one of the most definitive and compelling accounts of alienation in the modern age. Alienation can be regarded as "the plight of modern man" (Schacht, 1970: 1).

See also: Labor, Division of.

Bibliography

Adair-Toteff, Christopher, 1995. Ferdinand Tönnies: utopian visionary. Sociological Theory 13 (1), 58–65.

Arditi, Jorge, 1996. Simmel's theory of alienation and the decline of the nonrational. Sociological Theory 14 (2), 93–108.

Arendt, Hannah, 1968. Between Past and Future. Eight Exercises in Political Thought. Penguin Books, New York.

Arendt, Hannah, 2003. Responsibility and Judgment. Edited with an Introduction by Jerome Kohn. Schocken Books, New York.

Bauer, Isadora, 1962. Die Tragik in der Existenz des modernen Menschen bei G. Simmel. Duncker & Humblot, Berlin.

Hegel, Georg, Wilhelm, F., 1977. Hegel's Phenomenology of Spirit. Oxford University Press, Oxford. Translated by A.V. Miller with an Analysis of the Text and Foreword by J.N. Finlay.

Jaspers, Karl, 1947 (1935). Vernunft und Existenz. Johs. Strom Verlag, Bremen.

Jaspers, Karl, 1981 (1936). Nietzsche. Einführung in das Verständnis seines Philosophierens, Vierte unveränderte Auflage. Walter de Gruyter, Berlin und New York.

Jaspers, Karl, 1999 (1931). Die Geistige Situation der Zeit. Walter de Gruyter, Berlin und New York.

Kaufmann, Walter, 1970. Introductory essay. In: Schacht (Xv-lviii).

Löwith, Karl, 1941 (1981). Von Hegel bis Nietzsche—Der revolutionäre Bruch im Denken des 19. Jahrhunderts. In: Löwith, Karl (Ed.), Von Hegel bis Nietzsche, Band 4. J.B. Metzlersche Verlagsbuchhandlung, Stuttgart, pp. 1–489.

Löwith, Karl, 1952 (1988). Man's self-alienation in the early writings of Marx. In: Löwith, Karl (Ed.), Hegel und die Aufhebung der Philosophie—Max Weber, Band 5. J.B. Metzlersche Verlagsbuchhandlung, Stuttgart, Sämtliche Schriften, pp. 70–93.

Marx, Karl, 1964. Karl Marx. Early Writings. McGraw Hill Book Company, New York/Toronto/London. Translated and edited by T.B. Bottomore with a new Foreword by Erich Fromm.

Mészáros, István, 1972. Marx's Theory of Alienation. Harper Torchbooks, New York.

Nietzsche, Friedrich, 1999a. Also Sprach Zarathustra. In: Colli, Gorgio, Montinari, Mazzino (Eds.), Kritische Studienausgabe, Band 4. Deutscher Taschenbuch Verlag de Gruyter, München.

Nietzsche, Friedrich, 1999b. Jenseits von Gut und Böse. In: Colli, Gorgio, Montinari, Mazzino (Eds.), Jenseits von Gut und Böse/Zur Genealogie der Moral, Band 5. Deutscher Taschenbuch Verlag de Gruyter, München, pp. 10–243.

Nietzsche, Friedrich, 1999c. Zur Genealogie der Moral. Eine Streitschrift. In: Colli, Gorgio, Montinari, Mazzino (Eds.), Jenseits von Gut und Böse/Zur Genealogie der Moral, Band 5. Deutscher Taschenbuch Verlag de Gruyter, München, pp. 245–412.

Ollman, Bertell, 1976. Alienation. Marx's Conception of Man in Capitalist Society, second ed. Cambridge University Press, Cambridge.

Schacht, Richard, 1970. Alienation. Double Day/Anchor, Garden City, New York.

Segalman, Ralph, 1981. On the use and misuse of gemeinschaft conceptualization. In: Claussen, Lars, Pappi, Franz Urban (Eds.), Ankunft bei Tönnies. Walter G. Mühlau Verlag, Kiel, pp. 42–53.

Simmel, Georg, 1900. Philosophie des Geldes. Duncker & Humblot, Leipzig.

Simmel, Georg, 1918. Der Konflikt der modernen Kultur. Ein Vortrag von Georg Simmel. Duncker & Humblot, München und Leipzig.

Simmel, Georg, 1971. On Individuality and Social Forms. Edited with an introduction by Donald N. Levine. University of Chicago Press, Chicago and London.

Simmel, Georg, 1986. Der Begriff und die Tragödie der Kultur. In: Habermas, Jürgen (Ed.), Philosophische Kultur. Gesammelte Essais. Verlag Klaus Wagenbach, Berlin, pp. 195–218.

Simmel, Georg, 1989 (1910). Hauptprobleme der Philosophie, 9, unveränderte Auflage. Walter de Gruyter, Berlin/New York.

Tönnies, Ferdinand, 1991 (1887). Gemeinschaft und Gesellschaft. Wissenschaftliche Buchgesellschaft, Darmstadt.

Wood, Allan W., 1981. Karl Marx. Routledge & Kegan Paul, London, Boston and Henley.

Alliances: Political

Ole R Holsti, Duke University, Durham, NC, USA; and Salt Lake City, UT, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by O.R. Holsti, volume 1, pp. 397–401, © 2001, Elsevier Ltd.

Abstract

Alliances are formal agreements, open or secret, between two or more nations to collaborate on national security issues. This article addresses several enduring concerns about alliances, including alliance formation, performance and persistence, international and national effects, and their prospects in the new millennium.

Prominent explanations for alliance formation include realist balance-of-power theories and game-theoretical theses postulating that alliances will increase only to the point of a minimum winning coalition. Still other theories look into various characteristics of members. The article then focuses on alliance performance and the various attributes associated with effectiveness or ineffectiveness, including purposes, similarities of members, size, alliance leadership, and the advent of nuclear weapons. The article then reviews a number of studies on the international and national impact of alliances. Although the 'age of alliances' – the third quarter of the twentieth century – has passed, there is strong reason to believe that alliances will continue to be among the more important instruments of statecraft.

Alliances are formal agreements, open or secret, between two or more nations to collaborate on national security issues. They have been variously considered as techniques of statecraft, international organizations, or regulating mechanisms in the balance of power. This article addresses several of the enduring concerns about alliances, including alliance formation; their performance and persistence; their effects on the international system and domestic politics; and their prospects in the new millennium.

Alliance Formation

Among the oldest explanations of alliances are those derived from balance-of-power theories in which the emphasis is on the external environment, including the structure, distribution of power, and state of relations between units of the system. These are often closely linked to the 'realist' approach to international politics (Morgenthau, 1959; Waltz, 1979; Gilpin, 1981). Balance-of-power politics are almost unthinkable in the absence of alliances. According to Morgenthau, "The historically most important manifestation of the balance of power, however, is to be found not in the equilibrium of two isolated nations but in the relations between one nation or alliance of nations and another alliance" (Morgenthau, 1974: p. 188). Nations join forces as a matter of expediency in order to aggregate sufficient capabilities to achieve certain foreign policy goals, to create a geographically advantageous position – for example, to create buffers or isolate a potential adversary – or to prevent any nation of combination of countries from achieving a dominant position. Alliance partners are thus chosen on the basis of common goals and needs, not for reasons of shared values, shared institutions, or a sense of community. According to balance-of-power theories, nations should be more likely to join the weaker alliance to prevent formation of a hegemonic one ('balancing') rather than to join the dominant one in order to increase the probability of being on the winning side ('bandwagoning') (Waltz, 1979). Several case studies found support for the 'balancing' hypothesis: that nations form alliances to balance threats rather than power (Walt, 1987). But permanent alliances are themselves detrimental to the system and must therefore be avoided. As Nicholas Spykman (1942: pp. 103–104) noted: "He who plays the balance of power can have no permanent friends. His devotion can be to no specific state but only to balanced power. The ally of today is the enemy of tomorrow. One of the charms of power politics is that it offers no opportunity to grow weary of one's friends."

Balance-of-power formulations are a useful point of departure for understanding some of the motivations underlying alliance policies, at least for some periods or regions. But the number of deviant cases is sufficiently large to raise some serious doubts that these formulations can provide a general theory of alliances (Dawson and Rosecrance, 1966). Alliances may also play functions other than balancing power or threats. They can serve as instruments for binding, restraining, and managing relations with other members of the alliance (Schroeder, 1976; Osgood, 1968; Ikenberry, 2001). Although it is a considerable oversimplification, there is more than a grain of truth in Lord Ismay's observation that North Atlantic Treaty Organization (NATO) served three important functions: to keep the Americans in, to keep the Soviets out, and to keep the Germans down. It is hard to imagine that the French or other victims of Hitler's aggression would have accepted a rearmed Germany that was not firmly embedded within NATO.

The 'size principle,' according to which 'coalitions will increase in size only to the minimum point of subjective certainty of winning,' is drawn deductively from 'game theory' (Riker, 1962: pp. 32–33). Coalition theories share a number of characteristics with balance-of-power models. Both place heavy emphasis on calculation of advantage, adequate information, and rationality in alliance formation. They also tend to assume that national attributes other than power are of little importance in the calculations leading to alliance formation. But on one point they appear to differ: whereas an important goal of 'ideal' balance-of-power systems is to prevent the rise of a dominant nation or group of nations, the primary motivation in game approaches is to form just such a coalition, with enough partners to ensure victory but without any additional

ones who would claim a share of the spoils. Although intuitively appealing, a number of difficulties arise when this approach is applied to international politics. If war offers the prospect of winning territories or other divisible rewards, there are advantages to forming an alliance no larger than is necessary to gain victory, but even in redistribution alliances the territorial interests of partners must be complementary, permitting a noncompetitive division of rewards. Moreover, when an alliance is formed for purposes of defense or deterrence, success is measured by its ability to prevent conflict, not by the territorial or other gains derived from successful prosecution of a war; hence, bigger is better. Another potential difficulty is the demanding requirement that leaders be able to measure capabilities with sufficient precision to define a minimum winning coalition. While doing so may be relatively easy in a legislature or political party convention, it is far more difficult in the global arena. Any propensity to use 'worst case scenario' reasoning in foreign policy planning will likely result in a much larger alliance.

In contrast to the balance-of-power and minimum winning coalition theories are those that emphasize attributes other than national interest or power. These approaches do not deny that national interests or power calculations influence alliance formation, but they also assume that we cannot treat nations as undifferentiated units if we wish to understand either their propensity to use alliances as instruments of foreign policy – as opposed to such alternatives as neutrality – or in their choice of alliance partners (Holsti et al., 1973). The common denominator is a suspicion that foreign policy orientations reflect motives that draw upon many sources, some of which may be related to national attributes.

Political stability is sometimes associated with a propensity to join alliances, and instability has been seen as an impetus to go beyond nonalignment and pursue a policy of 'militant neutralism.' On the other hand, leaders faced with domestic instability may actively court allies in the hopes of gaining external support for a tottering regime. By utilitarian criteria, small, poor, or unstable nations are relatively unattractive alliance partners, but such nations have often been sought as allies, and they have even become the focal point of acute international conflicts and crises; twentieth-century examples include Bosnia (1908–09), Serbia (1914), Cuba (1961–62), and Vietnam (1965–73).

Given a propensity to seek allies, does the choice of partners reflect a discernible pattern of preferences? Affinity theories focus on the similarities among nations as an element in their propensity to align. The premise that nations are likely to prefer partners with whom they share common institutions, cultural and ideological values, or economic interests is intuitively appealing, and NATO can be cited as an example, but it has received limited support in studies of pre-1945 alliances (Russett, 1971; Holsti et al., 1973; Walt, 1987). The initial expansion of NATO following disintegration of the USSR did, however, give priority to three former members of the Warsaw Pact which had most effectively moved toward political and economic reform– Hungary, Poland, and the Czech Republic. These three countries had most effectively undertaken political and economic changes that moved them away from the Stalinist institutions of the Cold War era prior to the Soviet collapse.

Alliance Performance

British statesman Lord Palmerston once noted that nations have neither permanent enemies nor allies, only permanent interests. Although the longevity of most alliances can better be measured in years than decades or centuries, these observations do not explain why some alliances are cohesive and effective whereas others are not. The alliance concluded between Great Britain and Portugal in 1703 has survived owing to a shared interest in naval security concerns in the waters around the two countries. In contrast, the 1955 Central Treaty Organization (CENTO) linking Great Britain, Iran, Iraq, Pakistan, Turkey, and, later, the United States was never very effective. It was called the Baghdad Pact until Iraq's withdrawal in 1959, and it died two decades later when Iran withdrew following the revolution that overthrew the Shah's regime. A range of theories has been advanced to explain alliance performance and longevity: some emphasize the attributes of the member nations; some the character of the alliance itself; and others the nature of the international system.

The relative merits of open (democratic) and closed (authoritarian) polities have often been debated, but without consensus. To some, open polities are inferior allies on two counts: they experience more frequent changes in ruling elites, with the consequence that commitments to allies may also change, and the demands of domestic politics may take precedence over alliance requirement. But when democracies are under attack they are more likely to expand alliance functions and to turn alliances into communities of friendship; they are also less likely to renege on commitment by seeking a separate peace (Liska, 1962: pp. 50, 52, 115). Machiavelli noted almost five centuries ago, "Alliances are broken from consideration of interest; and in this respect republics are much more careful in the observance of treaties than princes" (Machiavelli, 1950: p. 268).

Political instability is often associated with poor alliance performance. Unstable regimes may experience radical changes in elites, which in turn result in shifting patterns of alignment, and they may also be more willing to run high risks on behalf of their own interests but not those of allies. A classic example is Austria–Hungary in 1914. Fearing that the multinational Dual Monarchy would be torn apart by demands for autonomy from various national groups, Vienna pursued a punitive policy toward Serbia, initially with the full support of Germany, but later with little attention to advice emanating from Berlin. Some of the most dramatic cases of realignment or alliance defection, however, have occurred in nations with well-entrenched elites. Nazi–Soviet relations offer numerous illustrations. The Nazi–Soviet Pact of 1939 was inconsistent with the Franco-Soviet alliance of 1935 and the Anti-Comintern Pact, of which Germany was a leading member. In the ultimate act of perfidy, Hitler unleashed Operation Barbarossa against his erstwhile Soviet ally in June 1941.

Differences in national bureaucratic structures and processes may be an important barrier to coordination of alliance strategies. If national security policy is the product of constant intramural conflict within complex and varied bureaucracies, even close allies may fail to perceive accurately the nuances in the bureaucratic politics 'game' as it is played abroad, and how the demands of various constituents may

shape and constrain alliance policies (Neustadt, 1970; Allison, 1971). Recognition that bureaucratic conflict may affect foreign policy choices may usefully supplement perspectives that regard alliance policies solely as a matter of rationality and calculation.

Alliances differ in many ways, including the circumstances under which their provisions become operative, the type of commitment, the degree of cooperation, geographic scope, ideology, size, structure, capabilities, and quality of leadership. The ANZUS Treaty (Australia, New Zealand, United States) called only on its members 'to meet the common danger in accordance with its constitutional processes' in case of an attack in the Pacific area on any one of them, and it had rather minimal bureaucratic apparatus. NATO, on the other hand, commits its members through Article V to joint military action in case of an attack on any one of them and, unlike the ANZUS Pact, it has an elaborate political and military infrastructure. Article V was invoked for the first time following the terrorist attacks on New York and Washington in September 2001.

Presumably all nations prefer to join alliances that offer them an effective role in determining goals, strategy, and tactics; a 'fair' share of the rewards without undue costs; and the maximum probability of success in achieving their goals. Redistribution alliance might be expected to achieve fewer successes and to break up more easily than those whose primary motives are deterrence and defense. Failure to achieve goals has a disintegrative effect and, other things being equal, it is easier for an alliance of deterrence to succeed because it is sufficient to deny the enemy a victory or to maintain the status quo. A redistribution alliance, on the other hand, must not only be able to avert defeat; it must also achieve victory if it is to be successful. The rewards of an offensive alliance can be shared noncompetitively only when the goals of each nation are limited and do not overlap with those of allies. In contrast, the more security an alliance of deterrence provides for one nation, the more secure are the others likely to feel. It remains to be seen whether an alliance such as NATO can survive the disintegration of the threat that gave rise to its existence. A leading 'realist' theorist famously declared that 'NATO's days are not numbered, but its years are' (Waltz, 1993: p. 76). Since then, however, NATO has not only expanded, but it has played an important role in Bosnia, Kosovo, Afghanistan, and other conflicts. Waltz pronounced validation for his prediction by the somewhat dubious assertion that NATO was no longer an alliance after its chief adversary collapsed, but had become 'a means of maintaining and lengthening America's grip on the foreign and military policies of European states' (Waltz, 2000: p. 20).

Ideology may play a role in sustaining or dissolving alliance bonds. Even those who minimize the importance of ideology in alliance formation tend to agree that a shared ideology may ensure that issues are defined similarly and facilitate intra-alliance communication. A common ideology may sustain alliances, but only as long as its tenets do not themselves become an issue. When they do, alliances in which doctrinal questions are highly salient are less likely to endure than those in which they are of marginal importance. Concrete alliance problems such as the level of military contributions can almost always be resolved by compromise, but it is a good deal more difficult for 'true believers' to split their differences on questions of 'fundamental truth.' Competing claims by Chinese and Soviet leaders about whose doctrines and policies more faithfully represented Marxist–Leninist ideology exacerbated some of their more concrete differences, including the legitimate boundaries in some contested areas between the two countries that had formed an alliance in February 1950.

Large alliances are usually less cohesive than small ones. The larger the alliance, the smaller the share of attention that nations can give to each ally, and as the size of the alliance increases, the number of relationships within the alliance rises even faster. When two nations ally, theirs is the only relationship; if three more nations join them, the number of two-way relationships jumps to 10; and when NATO increased its membership from 16 to 28, the number of bilateral relationships jumped from 120 to 378. Problems of coordination increase and so do opportunities for dissension. Finally, the larger the alliance, the less important the contributions of any single member, and the easier it is for any partner to become a 'free rider' by failing to meet its alliance obligations.

Alliances differ widely in the kinds of political and administrative arrangements that govern their activities. Allies may undertake wide-ranging commitments to assist each other and yet fail to establish institutions and procedures for communication and policy coordination. This tended to characterize the alliance between Germany and Austria–Hungary during the years leading up to the First World War (Craig, 1965; Williamson, 1969). The Warsaw Pact may be used to illustrate the other extreme. Basic alliance decisions were made in Moscow, all key positions in the alliance were filled by Russians, and even discussions of defection were severely punished, as Hungary discovered in 1956. In contrast, France's partial withdrawal from NATO in 1966 provoked little response from its allies.

Some suggest that hierarchical and centralized alliances are likely to be more cohesive and effective because they can mobilize their resources better and respond more quickly to threats and opportunities. According to others, pluralistic and decentralized alliances are likely to enjoy greater solidarity and effectiveness. In pluralistic alliances, even differences on central concerns are likely to remain confined to the single issue rather than metastasize to all issues (Holsti et al., 1973).

The proposition that influence within an alliance is proportional to strength seems intuitively correct but it may not always be valid. Under some circumstances, weakness may actually be a source of strength in intra-alliance diplomacy. The stronger nation is usually the more enthusiastic partner, it has less to gain by bargaining hard, and it can less credibly threaten to reduce its contributions. The weaker nation may also enjoy disproportionate influence within an alliance because it can commit its stronger ally, which may be unable to accept losses resulting from the smaller partner's defeat (Olson and Zeckhauser, 1966).

Each state brings both assets and liabilities to the alliance. The conventional manner of assessing capabilities is to sum the assets of the members. The more realistic view is that the capabilities of an alliance rarely equal the sum of its parts. Given close coordination, similar equipment, skillful leadership, and complementary needs and resources, economies of scale may be achieved. This may sometimes be the case for alliances of deterrence, but if it becomes necessary to carry out

military operations, alliance capabilities are probably less than those of the individual nations combined. Wartime operations may reveal or exacerbate problems arising from poor staff coordination, mistrust, incompatible goals, logistical difficulties, dissimilar equipment, and organization. This is what Napoleon had in mind when he stated, "If I must make war, I prefer it to be against a coalition" (Padelford and Lincoln, 1967: p. 402), and the Duke of Wellington grumbled that having enemies is nothing like as bad as having allies. A century later Marshal Foch remarked, "My admiration for Napoleon has shrunk since I found out what a coalition was" (Rothstein, 1966: p. 125).

The capabilities of alliance leaders are of special importance. A powerful and wealthy bloc leader can offer side payments – private goods to supplement public ones – to smaller partners, which may in turn render the alliance more effective. According to the economic theory of 'collective action,' alliances that supplement public benefits (those that are shared by all members) with private or noncollective ones are more cohesive than alliances that provide only collective benefits (Olson and Zeckhauser, 1966). Except in pure conflict situations, there are always some tensions between the requisites of alliance cohesion and broader global concerns. If the alliance is to cope effectively with these tensions, it requires effective leadership from its leading member or members. Alliance relationships involve several kinds of tensions and fears. Members face fear of abandonment when their own interests are at stake, and of entrapment whose of others are threatened (Snyder, 1997). The ultimate test of alliance leadership comes during an international crisis when the leading partner may be forced to resolve serious tensions between alliance management (intra-alliance diplomacy) and crisis management (inter-alliance diplomacy).

This may entail a number of critical tasks:

- To convey to the adversary (1) evidence of a determination to maintain existing commitments; and (2) sufficient reassurance to reduce motivations for significant military escalation and, possibly, to permit bargaining on crisis issues.
- To convey to allies (1) assurances that their vital interests will not be sacrificed in seeking accommodation with the adversary (assuaging fears of abandonment); (2) assurances that they will not be drawn into a general war over issues of peripheral interest to the alliance as a whole (assuaging fears of entrapment); and (3) a resolve not to be drawn into a strategy of escalation on behalf of an ally which is willing to run unduly high risks in the pursuit of purely parochial interests.

Virtually every alliance faces the problem of ensuring the credibility of its commitments. The problem is most serious for alliances of deterrence. If adversaries possess serious doubts on this score, the alliance may serve as an invitation to attack. Equally important, if members have doubts about the assurances of their partners, the coalition is unlikely to be effective. 'Irrational' alliance commitments may be undertaken as part of an overall strategy of increasing the credibility of deterrence by conveying to the adversaries – and to other allies – that if the alliance leader is willing to expand vast resources to protect areas of little strategic values, then it should be clear that an even greater effort will be made to defend other allies (Maxwell, 1968: p. 8). Buttressing credibility by 'irrational' commitments may also entail risks – as the United States discovered in Vietnam – forcing the alliance leader to choose between unpalatable alternatives: reducing the commitment under threat, thereby seriously eroding one's credibility in the future; or backing the promise to the hilt, with the possibility of becoming a prisoner of the ally's policies.

Although it is widely asserted that alliance cohesion depends upon an external threat and it declines as the danger is reduced, this generalization must be qualified. Such threats may give rise to divisions if only part of the alliance membership feels threatened, if the threat strikes at the basis of alliance consensus, or if it offers a solution that sets off the interests of one ally against those of another. Occasionally, intense disputes may even break out between allies, as illustrated by the Cyprus issue between NATO members Turkey and Greece. Large alliances such as NATO are almost certain to experience circumstances that are perceived as a threat by only one or two members. The Suez crisis of 1956, the Yom Kippur War of 1973, and the US invasion of Iraq in 2003 are good examples. In the latter case, the Blair government in London joined the United States, but important NATO allies such as France, Germany, and Canada refused to do, citing doubts about the evidence presented by the Bush administration – doubts that proved to be completely valid – regarding Iraq's alleged weapons of mass destruction and complicity in the 2001 terrorist attacks on Washington and New York. The Polish uprising in 1956, the offshore islands crisis of 1958, and the Sino-Indian border conflict of 1962 were situations in which Moscow and Beijing differed rather sharply on their estimates of the threat. But the consequences of such divisive events were quite different for NATO and the Sino-Soviet alliance. Similarly, unless the external danger creates an equitable division of labor among alliance members, cohesion and effectiveness are likely to suffer. Although wartime alliances face considerable external threat, they may sometimes experience tensions because military operations rarely result in equal burdens. As long as the threat calls for a cooperative solution, cohesion will probably be enhanced, but should a solution that favors one ally at the expense of others appear, the alliance may lose unity or disintegrate. Finally, an alliance confronting an external threat for which there is no adequate response may also experience reduced cohesion or disintegration.

Expanding membership in the 'nuclear club' has generated an extensive debate on the likely consequences of nuclear weapons for alliances. Most analysts stress their negative effects. One reason is that contemporary military technology has permitted some nations to gain preponderant power without external assistance. Another strand of the argument is that nuclear deterrence will be less credible if it takes the form of one alliance member providing a 'nuclear umbrella' – extended deterrence – for the others. Put in its starkest form the question is: what nation will risk its own annihilation by using its nuclear capabilities as a means of last resort to punish aggression against its allies? When justifying their acquisition of nuclear weapons, the French asked, "Would Washington commit suicide to save Paris?" The other side of the argument is the fear that one may unwittingly become a nuclear target as a result of an ally's quarrels. The latter point is illustrated by New Zealand's decision in 1986 to deny port facilities to ships

carrying nuclear weapons. As a result, New Zealand was excluded from the ANZUS mutual defense pact.

Another aspect of the problem focuses not only on the *existence* of nuclear weapons, but on their spread to junior alliance members. The view that nuclear proliferation will result in disintegration of coalitions appears grounded in a strictly utilitarian view of alliances. The opposing thesis seems to follow from the position that because other considerations may be important in the life of an alliance, sharing nuclear capabilities will add to the sense of equality and cooperation among allies. Advocates of the first position can point to the *de facto* defections of France and China from their alliances; it is not clear which came first, however, the decision to acquire an independent nuclear deterrent or strains between Paris and its NATO partners and between Moscow and Beijing. The latter viewpoint receives support from Great Britain's continued adherence to NATO after becoming the world's third nuclear power.

International Effects of Alliances

The foregoing discussion has been concerned solely with the causes and conditions of alliances and their performance. We now turn our attention to their *effects* on the international system and on the nations that join or shun them.

Alliances are intended to play a central role in maintaining the balance of power. They provide the primary means of deterring or defeating nations or coalitions that seek to destroy the existing balance by achieving a position of hegemony. Even advocates of balance-of-power diplomacy attach a number of important qualifications. Alliances among great powers contribute to instability if they are strong enough to destroy the existing balance, or if they reduce the number of nations that may act in the role of 'balancers,' which, according to the theory, are supposed to remain uncommitted until there is a threat to the balance.

Alliance critics assert that they merely breed counter-alliances, thereby leaving no nation more secure, while at the same time contributing to polarization and international tensions. For this reason, Quincy Wright (1965: p. 774) suggested that international stability is better served by efforts to break up dangerous coalitions rather than merely offsetting them with counter-alliances. A further criticism is that alliances are incompatible with collective security, which requires every nation automatically to assist the victim of aggression; alliance commitments can thus create a conflict of interest. It does not necessarily follow, however, that in the absence of alliances nations will develop an effective collective security system. Indeed, alliances may arise from disappointed hopes for collective security. Thus, elimination of alliances may be a necessary condition for collective security, but it is not sufficient. Critics of alliances also suggest that they act as conduits to spread conflict to regions previously free of it. They can point to historical instances in which alliances had this result, as can those who stress that alliances serve a deterrent function. But Liska (1962; see also Weitsman, 2004) has come to the conclusion that alliances neither cause nor prevent conflict, nor do they expand or limit it. Another set of studies of all alliances during the period between the end of the Napoleonic Wars and the outbreak of the Second World War incorporates two premises. First, the greater the number of independent units, the more peaceful the international system, because nations can concentrate less of their attention to any other country (Deutsch and Singer, 1962). Second, maximum freedom of interaction between nations enhances international stability because pressures on each country tend to be cross-cutting rather than reinforcing, with the result that bargaining rather than war is a more likely strategy to resolve differences. The Deutsch–Singer assumptions have an evident relationship to several prominent alliance theories. They are quite consistent with theories of integration which postulate that cooperative interaction on one issues tends to nurture cooperation in others. They are also related to classical balance-of-power theories, the central argument of both being that entangling commitments tend to reduce the freedom of action necessary to maintain system stability.

Results based on a system-level analysis of 118 alliances indicate that alliance formation and international conflict are positively correlated for the twentieth century, but the relationship for the nineteenth century is negative (Singer and Small, 1968; Bueno de Mesquita and Singer, 1973). In an analysis of 265 conflicts in the 1815–1965 period, Siverson and Tennefoss (1984) found that equality of power, supplemented by an alliance between a major power and weaker nations, tended to restrain the likelihood of escalation. Kegley and Raymond (1982) concluded that peace is best preserved when the structure of alliances is moderately flexible and when alliance commitments are considered binding. But in some cases the data have yielded contradictory conclusions. Whereas Zinnes (1967) found a positive correlation between the level of alliance commitment in the system and the outbreak of war, Ostrom and Houle (1978) came to the opposite conclusion. Moreover, we may not yet be in a position to rule out the possibility of a spurious correlation between alliances and war. That is, could alliance commitments and international conflict be the product of some common underlying factor?

Another strand of theory, not necessarily incompatible with either of the above, sees alliances as a possible step in the process of more enduring forms of integration. This view begins with the premise that effective cooperation among units in one sphere of endeavors gives rise to collaboration in others and, in the long run, to institutionalization of the arrangements. Presently available evidence appears to support those who doubt the integration potential of alliances. A study of 10 political communities in the North Atlantic area revealed that alliances were neither necessary nor sufficient for integration (Deutsch et al., 1957). A detailed study of NATO supports these findings. Not even in areas thought to be especially promising – armaments, infrastructure, and science – was there evidence of integration (Beer, 1969). Others have similarly concluded that cooperation in security-related matters has a lower potential for 'spillover' than collaboration in issue areas such as economic relations (Etzioni, 1962).

Finally, alliances may be related to change in the structure of the international system. For example, the more cohesive the alliances on matters of substance and strategy, the more polarized the international system. But propositions of this form are of fairly limited interest as they tend to be true by definition; that is, a polarized system is usually defined as one

in which alliances are cohesive and rigid. Thus, we are impelled to ask further questions, such as: under what circumstances are alliances in a balance-of-power or loose bipolar system likely to become rigid, thereby transforming the nature of the system?

National Effects of Alliances

Alliance benefits may include enhanced security, reduced defense expenditures, and possible side benefits such as economic aid and prestige. A strong ally may be a necessary condition for survival. But alliances may be a net drain on national resources, they may distort calculations of the national interest if allies become wedded to 'inherent good faith' models of each other, and they may lead to the loss of decision-making independence. And, if alliances merely trigger off counter-alliances, they may leave everyone poorer and less secure.

Prospects

The third quarter of the twentieth century – 'the age of alliances' – was ushered in by the formation of NATO in 1949 and the Sino-Soviet Security Treaty in 1950. Within a decade the Warsaw Pact, Southeast Asia Treaty Organization, CENTO, and a multitude of other bilateral and multilateral alliances were formed, but many of them dissolved before the end of the Cold War. Contrary to the expectations of some 'realist' theorists, however, NATO has outlived the threat that brought it into existence. Although many Cold War alliances failed to survive into the new millennium, as long as the international system is characterized by independent political units, alliances are likely to persist as a major instrument of statecraft. According to Snyder (1997: p. 78), alliance formation is "the product of systemic anarchy, strength inequalities, and conflict and common interests among states." Because those conditions are likely to characterize the global system as far as one can foretell, alliances are likely to remain among the most important means by which states pursue their interests.

See also: Balance of Power, History of; Globalization: Legal Aspects; Imperialism, History of; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; Peace-Making in History; Peace; War, Sociology of; Warfare in History; Wars among Nation-States: Patterns and Causes.

Bibliography

Allison, G.T., 1971. The Essence of Decision: Explaining the Cuban Missile Crisis. Little Brown, Boston.
Beer, F.A., 1969. Integration and Disintegration in NATO. Ohio State University Press, Columbus.
Bueno de Mesquita, B., Singer, J.D., 1973. Alliances, capabilities, and war: a review and synthesis. In: Cotter, C.P. (Ed.), Political Science Annual: An International Review, vol. 4. Bobbs-Merrill, Indianapolis, IN.
Craig, G., 1965. The World War I alliance of the central powers in retrospect: the military cohesion of the alliance. Journal of Modern History 37, 336–344.
Deutsch, K.W., et al., 1957. Political Community in the North Atlantic Area: International Organization in the Light of Historical Experience. Princeton University Press, Princeton, NJ.
Deutsch, K.W., Singer, J.D., 1962. Multipolar power systems and international stability. World Politics 16, 390–406.
Dawson, R., Rosecrance, R.N., 1966. Theory and reality in the Anglo- American alliance. World Politics 19, 21–51.
Etzioni, A., 1962. The dialectics of supranational unification. American Political Science Review 56, 927–935.
Gilpin, R., 1981. War and Change in World Politics. Cambridge University Press, New York.
Holsti, O.R., Hopmann, P.T., Sullivan, J.D., 1973. Unity and Disintegration in International Alliances. Wiley, New York.
Hunt, Peter, 2010. War, Peace and Alliance in Demosthenes' Athens. Cambridge University Press, New York.
Ikenberry, G.J., 2001. After Victory. Princeton University Press, Princeton, NJ.
Kegley, C.W., Raymond, G.A., 1982. Alliance norms and war: a new piece in an old puzzle. International Studies Quarterly 26, 572–595.
Liska, G., 1962. Nations in Alliance: The Limits of Interdependence. Johns Hopkins University Press, Baltimore, MD.
Machiavelli, N., 1950. The Prince and the Discourses. Modern Library, New York.
Maxwell, S., 1968. Rationality in Deterrence. International Institute for Strategic Studies, London.
Morgenthau, H.J., 1959. Alliances in theory and practice. In: Wolfers, A. (Ed.), Alliance Policy in the Cold War. Johns Hopkins University Press, Baltimore, MD.
Morgenthau, H.J., 1974. Politics among Nations, fifth ed. revised. Knopf, New York.
Neustadt, R.E., 1970. Alliance Politics. Columbia University Press, New York.
Olson, M., Zeckhauser, R., 1966. An economic theory of alliances. Review of Economics and Statistics 48, 266–279.
Osgood, R.E., 1968. Alliances and American Foreign Policy. Johns Hopkins University Press, Baltimore, MD.
Ostrom, C.W., Houle, F.W., 1978. Alliances and war revised: a research note. International Studies Quarterly 22, 215–236.
Padelford, N.J., Lincoln, G.A., 1967. The Dynamics of International Politics, second ed. Macmillan, New York.
Riker, W.H., 1962. The Theory of Political Coalitions. Yale University Press, New Haven, CT.
Rothstein, R.L., 1966. Alignment, non-alignment, and small powers, 1945–1965. International Organization 20, 397–418.
Russett, B.N., 1971. An empirical typology of international military alliances. Midwest Journal of Political Science 15, 262–289.
Schroeder, P.W., 1976. Alliances 1815–1945: weapons of power and tools of management. In: Knorr, K. (Ed.), Historical Dimensions of National Security Problems. Regents Press of Kansas, Lawrence, KS.
Singer, J.D., Small, M., 1968. Alliance aggregation and the onset of war. In: Singer, J.D. (Ed.), Quantitative International Politics: Insights and Evidence. Free Press, New York.
Siverson, R.M., Tennefoss, M.R., 1984. Power, alliance, and the escalation of international conflict, 1915–1965. American Political Science Review 78, 1057–1069.
Snyder, G.H., 1997. Alliance Politics. Cornell University Press, Ithaca, NY.
Spykman, N., 1942. America's Strategy in World Politics. Harcourt Brace, New York.
Walt, S.M., 1987. The Origin of Alliances. Cornell University Press, Ithaca, NY.
Waltz, K.N., 1979. Theory of International Politics. Addison-Wesley, Reading, MA.
Waltz, K.N., 1993. The emerging structure of world politics. International Security 18, 44–79.
Waltz, K.N., 2000. Structural realism after the Cold War. International Security 25, 5–41.
Weitsman, Patricia, 2004. Dangerous Alliances: Proponents of Peace, Weapons of War. Stanford University Press, Stanford, CA.
Williamson Jr., S.T., 1969. The Politics of Grand Strategy: France and Britain Prepare for War, 1904–1914. Harvard University Press, Cambridge, MA.
Wright, Q., 1965. A Study of War, second ed. University of Chicago Press, Chicago.
Zinnes, D.A., 1967. An analytical study of the balance of power theories. Journal of Peace Research 3, 270–288.

Allport, Gordon W (1897–1967)

Thomas F Pettigrew, University of California, Santa Cruz, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Gordon Allport, a major American psychologist of the mid-twentieth century, made significant contributions to both personality and social psychology. His 1937 book, *Personality*, helped to establish personality as an essential component of academic psychology. His 1954 book, *The Nature of Prejudice*, became a classic in social psychology and shaped the study of intergroup relations for half a century. In both areas, Allport's work is characterized by three interwoven features: a broad, eclectic balance of the many sides of psychology; the specification of central problems with initial approaches to their solution; and a uniquely consistent, seamless, and forceful perspective.

Gordon Willard Allport was a major American personality theorist and social psychologist of the mid-twentieth century. His legacy in psychology is twofold. He helped to establish personality psychology as a science and as an integral part of the discipline of academic psychology. In addition, his applied work in social psychology, particularly on intergroup prejudice, furthered the practical value of the discipline's work on important social problems.

The son of a Scottish American medical doctor, Allport was born on 11 November 1897 in Montezuma, Indiana, USA, and grew up in Cleveland, Ohio. His early life strongly shaped his later ideals and achievements (Elms, 1993; Nicholson, 2003). Allport's half-century affiliation with Harvard University began in 1915 when he enrolled as an undergraduate. His student career foretold his adult interests in science and social issues. He majored in both psychology and social ethics and was impressed by his first teacher in psychology, Hugo Münsterberg. He spent much of his spare time in social service activities. This convergence of interests took institutional form later, when he helped to establish both the Society for the Psychological Study of Social Issues and Harvard's Department of Social Relations.

Upon graduation in 1919, Allport taught English and Sociology at Robert College in Constantinople (then part of Greece, now Bogazici University in Istanbul, Turkey). In addition to German, he remained fluent in Modern Greek throughout his life. Returning to Harvard in 1920, he completed his PhD in psychology in 2 years. His dissertation title again reflected his dual commitment to science and social concerns: An Experimental Study of the Traits of Personality: With Special Reference to the Problem of Social Diagnosis. In addition, he assisted his older brother, Floyd, in editing the *Journal of Abnormal and Social Psychology*, the start of more than four decades of association with this core publication of the discipline.

Harvard then awarded him a coveted 2-year Sheldon Traveling Fellowship. He spent the first year in Germany, where the new *Gestalt* school and its emphasis on cognition fascinated him. Indeed, he became a partial Gestaltist – partial because he could not accept the Gestaltists' assumptions about the inflexible genetic basis of cognitive processes (Pettigrew, 1979). He spent his second Sheldon year at Cambridge University, where the experimental psychologists coolly received his report on *Gestalt* developments.

In 1924, Allport became a Harvard instructor in social ethics and taught what was probably the first personality course offered by a North American university. Two years later he temporarily severed his connection with Harvard to accept an assistant professorship in psychology at Dartmouth College. Yet, even during his brief 4 years away, he returned repeatedly to teach in Harvard's summer school. In 1930 he came back to Harvard to stay.

Allport's unique contributions to psychology are best described by three interwoven features of his work. First, he offered a broadly eclectic balance of the many sides of the discipline – holding to William James's contention that there were "multiple avenues to the truth." Second, he formulated the central future explanatory problems of the discipline and proposed original approaches to them. Finally, his entire body of scholarly work presents a consistent, seamless, and forceful perspective.

Broadly Eclectic Balance

Allport sought an eclectic balance for both methods and theory. His two most famous volumes on personality illustrate this dominant feature of his work (Allport, 1937, 1962; Pettigrew, 1990). He urged, for example, the use of both ideographic (individual) and nomothetic (universal) methods. Since he thought the discipline relied too heavily on nomothetic approaches, he sought greater use of ideographic techniques. In an age of indirection, Allport insisted that if you want to know something about a person, why not first ask him? Considered scandalously naive at first, his position helped to right the balance in assessment.

Typically, his was an expansionist, not an exclusionist, view. He simply sought a reasonable trade-off between accuracy and adequacy. Allport thought the two foci together would make for 'a broadened psychology.' Indeed, his own empirical efforts ranged from personal documents, such as *Letters from Jenny* (Allport, 1965), to two popular tests on ascendance-submission and personal values. He also developed ingenious experimental procedures to study eidetic imagery, expressive movement, radio effects, rumor, the trapezoidal window illusion, and binocular rivalry.

His theoretical efforts also sought a balance. He vigorously advocated an open system theory of personality with emphases on individuality, proaction, consciousness, maturity, and the unity of personality. As such, in Rosenzweig's (1970: p. 60) view, Allport served as the field's 'ego' in contrast to Henry

Murray's 'id' and Edwin Boring's 'superego.' Yet Allport did not hold his emphases to be the only matters of importance for a psychology of personality. Rather, it was because he believed the discipline was granting ego concerns too little attention.

In this sense, Allport was, to borrow from boxing, a counterpuncher. He opposed what he regarded as excessive trends in psychology that threatened his conception of an open, balanced discipline. In 1937, in his first personality book, he saw as the major threat the too-rigid applications of experimental psychology. By 1961, in his second personality volume, he saw as the major threat the too-loose applications of psychoanalysis. "Although much of my writing is polemic in tone," he confessed in his autobiography, "I know in my bones that my opponents are partly right." The key word is *partly*. Allport opposed excess, "the strong aura of arrogance found in … fashionable dogmas" (Allport, 1968: pp. 405–406). So he held fast to the open middle ground as he perceived it, and aimed his punches at the 'fashionable dogmas' that existed in each period. Modern readers may miss the significance of his arguments if they are unaware of which dragons he is attempting to slay.

Formulating Central Explanatory Problems and Offering Original Solutions

Psychology recognized Allport throughout his career as a source for specifying the discipline's central problems with explaining behavior. Typically his initially proposed solutions to the problems – such as functional autonomy – won limited acceptance. Nonetheless, many of Allport's initial proposals for addressing basic problems now exist in psychology with new labels and enlarged meanings. Allport only loosely sketched out his innovative ideas. Later work accepted the problem and expanded the ideas. Consider the much derided concept of functional autonomy. The notion that motives can become independent of their origins was considered heretical in 1937. Slowly, psychology came to accept the phenomenon if not the formulation. Today, social psychologists reconceptualize the process in interactionist terms. Motives, established and functional in one situation, lead individuals to new situations where the same motives persist but assume new functions.

Similarly, Allport's conception of personality traits has often been criticized. What critics attack is the mistaken notion that he held a static view of traits as pervasive, cross-situational consistencies in behavior. But Allport advanced a far more dynamic conception of traits. In fact, he was "an interactionist in the sense that he recognized behavior is determined by the person and situation" (Zuroff, 1986: p. 993).

Indeed, Allport broke early ground for many ideas that are now fully developed and accepted. Thus, he provided in his 1937 volume a social constructionist interpretation of identity. And his insistence on multiple indicators and methods offered an initial statement of Campbell and Fiske's (1959) multitrait–multimethod approach.

A Consistent, Forceful Perspective

Above all, Allport's contributions to psychology flowed from a consistent and forceful perspective presented in graceful prose. One reviewer of his 1937 *Personality* book wrote, "One has all the way through it a distinct feeling that 'This is Allport'" (Hollingworth, 1938: p. 103). This pointed observation holds true for all his writing.

Allport's perspective remained consistent but not static throughout his career. He acquired a comprehensive knowledge of the psychological literature from his long years as a meticulous editor. Quite literally, a large proportion of North America's personality and social psychological literature in mid-century crossed his editor's desk. His mastery of the literature also reflected his open-ended view of theory – a view more Popperian than the strict Vienna circle positivism that held sway throughout most of his career.

Yet Allport held to his perspective forcefully. Blunt prose and forthright, critiques characterize his writing style. As one disgruntled reviewer of an Allport book put it, 'There is something in it to irritate almost everyone' (Adelson, 1956: p. 68). Hall and Lindzey (1957) suggest that Allport's many years of teaching led to his expressing his views in a uniquely salient and provocative style.

Allport on Prejudice

This directness is also apparent in his applied volume, *The Nature of Prejudice* (Allport, 1954). This influential book again brought together the combination of science with social action that Allport expressed early in his career. He deemed it his proudest achievement because he thought it "had done some good in the world." Indeed, in its paperback edition, it became one of the best selling social psychological books ever published.

The Nature of Prejudice displays the special characteristics of Allport's contributions to psychology. The volume offers a broad, eclectic perspective with a lens model that ranges from history to the psychological effects of prejudice on its victims. Its open view of the many sides of the phenomenon again demonstrates the distinctive quality of Allport's wide-ranging thought. This book is another seamless work that is 'Allportian' from start to finish. Deftly crafted in a simpler style than his other writings, Allport strove to make it accessible to a wider, nonacademic audience. In addition, it served to structure the entire study of prejudice for the next half a century. This is true for intergroup research in North America; and, in its many translated versions, it has also proven highly influential throughout the world. Its enormous impact is indicated by the existence of a volume with 44 diverse authors that marked the 50th anniversary of the publication of *The Nature of Prejudice* (Dovidio et al., 2005). Indeed, Allport's many predictions, though largely shaped by American intergroup data, have typically been confirmed in studies of intergroup prejudice throughout the world.

The Nature of Prejudice's useful definition of prejudice – "an antipathy based upon a faulty and inflexible generalization" (Allport 1954: p. 9) – stressed both affective and cognitive components, but it wisely left the complex link between prejudice and behavior an open, empirical question. This definition has served the field well, and only recently have social psychologists advanced expansions of its terms.

To attain his distinctive quality of balance, Allport again assumed the role of counterpuncher against prevailing dogmas.

He gave full recognition to the importance of the psychoanalytically inspired authoritarian personality syndrome, which he also had uncovered during the 1940s. But he challenged the Freudian formulation of aggression with a rival theory. Instead of the psychoanalytic steam boiler model of aggression and catharsis, Allport proposed a feedback model with dramatically different implications for prejudice and its remediation. Aggression, he argued, feeds on itself. That is, the acting out of aggression, rather than leading to less aggression, actually increases the probability that further aggression will be expressed. Armed with this insight, *The Nature of Prejudice* proceeds to advocate policies that have indeed reduced levels of prejudice throughout the world – including areas with long histories of intergroup conflict such as Northern Ireland and South Africa.

The book's most important theoretical contribution to addressing prejudice was the intergroup contact hypothesis. He argued that such contact was crucial for improving intergroup relations and that four conditions of the intergroup interaction were critical: equal status in the situation, common goals, no intergroup competition, and authority sanction. Extensive contact research throughout the world has supported his contentions; and these four conditions have proven to be important facilitating, though not essential, factors in reducing intergroup prejudice (Pettigrew and Tropp, 2011). His treatment of prejudice also countered the then-fashionable assumption that group stereotypes were simply the aberrant cognitive distortions of prejudiced people. Instead, Allport held that the cognitive components of prejudice were natural extensions of normal processes, a view now firmly established and universally accepted. Stereotypes and prejudgment, he concluded, were not aberrant at all but unfortunately all too human.

Allport's foresight into the many advances that prejudice research has achieved over recent decades, including in research on group stereotypes, can be traced to his early *Gestalt* training (Pettigrew, 1979). He devoted 10 of *The Nature of Prejudice*'s 31 chapters to cognitive factors and psychology joined the cognitive revolution just after the volume was published. In social psychology, social cognition veered in the largely *Gestalt* direction that had molded Allport's perspective on prejudice. Thus, the influences that shaped the study of prejudice in general and stereotypes in particular from 1960 on, were influences that had earlier guided Allport's thought.

The Nature of Prejudice also presents a host of original hypotheses on specific topics that have stood the test of time. For instance, one now popular theory of prejudice reduction emphasizes recategorization through identity with larger, more inclusive groups. Allport (1954) advocated the same mechanism. Drawing concentric circles with family in the center and humankind at the periphery, he argued that "concentric loyalties need not clash" and that prejudice is minimized by inclusive group membership.

The volume also devotes an entire chapter to the link between religion and prejudice. A devout Episcopalian, Allport was disturbed that research routinely finds nonbelievers far less prejudiced on average than members of organized religions. He proposed a critical distinction between an 'institutionalized' religious outlook and an 'interiorized' one. The more numerous institutionally religious, he argued, are the highly prejudiced. Those of the interiorized type, who have internalized their religious beliefs, are held to be far less prejudiced. In his last empirical publication (Allport and Ross, 1967), Allport presented evidence in support of his hypothesis.

Later work by Batson and his colleagues has extended this view (Batson and Stocks, 2005). They find that Allport's two types exist as independent dimensions of religiosity but not as polar opposites. Further, they add a third, independent religious orientation – 'quest' – in which people face directly such difficult religious issues as the meaning of life and personal mortality in all their complexity. And how these three types relate to prejudice is dependent on whether the prejudice is religiously proscribed or not and whether it is measured overtly or covertly. The institutionalized religious do indeed tend to be more prejudiced on both types of measures when the prejudice is proscribed. The interiorized religious, however, tend to be more prejudiced when the prejudice is not proscribed and is measured covertly. Only those with the quest orientation tend to be less prejudiced regardless of whether the prejudice is proscribed or not and how it is measured.

Allport addressed his prejudice volume primarily to his own in-group – White, Protestant, American males. The examples of prejudice cited throughout involve anti-Black, anti-Jewish, anti-Catholic, and antifemale sentiments. He was clearly lecturing to 'his own kind.' It is safe, easy, and politically expedient to attack the prejudices of out-groups who hold negative views of your in-group. It is quite a different matter to attack the prejudices of your own in-group toward others. So yet another remarkable feature of *The Nature of Prejudice* is its target audience.

Allport as Teacher and Mentor

Though intellectually confident, Allport was a shy and modest man – though his manner was sometimes mistaken for aloofness. In his gentle and supportive way, he was a master teacher and mentor. Consistent with his conception of the uniqueness of personality, he encouraged students to follow their own interests. Hence, though he produced dozens of well-known psychologists, he never developed a 'school' of followers.

One way Allport handled his social shyness was to prepare carefully for occasions in advance. Before giving his Hoernle Lecture at South Africa's leading Afrikaner University at Stellenbosch in 1956, he studied Afrikaans with a tutor for 6 months. He then skillfully gave the introduction to his lecture in Afrikaans, gracefully apologizing for not delivering his entire address in Afrikaans. The Afrikaner audience, accustomed to even other South Africans not knowing their language, reacted with surprise and delight. The audience arose as one with applause.

The Lasting Contribution

The forcefulness and integrity of his work can be seen in Allport's many professional honors. He was elected president of both the American Psychological Association (1939) and the Society for the Psychological Study of Social Issues (1944). He

received the Gold Medal of the American Psychological Foundation in 1963, a few years prior to his death on 9 October 1967 in Cambridge, Massachusetts. The eclectic nature of his work, along with his essential contributions to psychological explanation, is apparent in his legacy to both modern personality and social psychology. Specialists in personality have expanded his conception of traits and now use the concept as a major tool (e.g., Lilienfeld, 2005; Piekkola, 2011). They also draw on Allport's writings to support a resurgence of interest in the uniqueness of individuals (Barenbaum, 2005; Lamiell, 2009). Social psychologists constantly cite his classic on prejudice. Indeed, according to the counts of Google Scholar, Allport's the *Nature of Prejudice* is cited twice as often as his two popular personality texts.

See also: Attitudes and Behavior; Cattell, Raymond Bernard (1905–98); Jung, Carl Gustav (1875–1961); Köhler, Wolfgang (1887–1967); Personality Theory and Psychopathology; Personality and Adaptive Behaviors; Prejudice in Society: Sociological Perspectives; Racial Relations; Relativism, Cognitive; Social Psychology.

Bibliography

Adelson, J., 1956. On man's goodness [Review of becoming: basic considerations for a psychology of personality]. Contemporary Psychology 1, 67–69.

Allport, G.W., 1937. Personality: A Psychological Interpretation. Holt, Rinehart & Winston, New York.

Allport, G.W., 1954. The Nature of Prejudice. Addison-Wesley, Reading, MA.

Allport, G.W., 1961. Pattern and Growth in Personality. Holt, Rinehart & Winston, New York.

Allport, G.W., 1962. Prejudice: is it societal or personal? Journal of Social Issues 18, 120–134.

Allport, G.W., 1965. Letters from Jenny. Harcourt, Brace & World, New York.

Allport, G.W., 1968. The Person in Psychology: Selected Essays. Beacon Press, Boston.

Allport, G.W., Ross, J.M., 1967. Personal religious orientation and prejudice. Journal of Personality and Social Psychology 5, 432–443.

Barenbaum, N.S., 2005. Four, two, or one? Gordon Allport and the unique personality. In: Schultz, W.T. (Ed.), Handbook of Psychobiography. Oxford University Press, New York, pp. 223–239.

Batson, C.D., Stocks, E.L., 2005. Religion and prejudice. In: Dovidio, J.F., Glick, P., Rudman, L.A. (Eds.), On the Nature of Prejudice: Fifty Years after Allport. Blackwell, Malden, MA, pp. 413–427.

Campbell, D.T., Fiske, D.W., 1959. Convergent and discriminate validation of the multitrait–multimethod matrix. Psychological Bulletin 56, 81–105.

Dovidio, J.F., Glick, P., Rudman, L.A. (Eds.), 2005. On the Nature of Prejudice: Fifty Years after Allport. Blackwell, Malden, MA.

Elms, A., 1993. Allport's 'personality' and Allport's personality. In: Craik, K.H., Hogan, R., Wolfe, R.N. (Eds.), Fifty Years of Personality Psychology. Plenum Press, New York, pp. 39–55.

Hall, C.S., Lindzey, G., 1957. Theories of Personality. Wiley, New York.

Hollingworth, H.L., 1938. Review of personality. Psychological Bulletin 35, 103–107.

Lamiell, J.T., 2009. Reviving person-centered inquiry in psychology: why it's erstwhile dormancy? In: Valsiner, J., Molenaar, P.C.M., Lyra, M.C., Chaudhary, N. (Eds.), Dynamic Process Methodology in the Social and Developmental Sciences. Springer, New York, pp. 31–43.

Lilienfeld, S.O., 2005. Longitudinal studies of personality disorders: four lessons from personality psychology. Journal of Personality Disorders 19 (5), 547–556.

Nicholson, I.A.M., 2003. Inventing Personality: Gordon Allport and the Science of Selfhood. American Psychology Association, Washington, DC.

Pettigrew, T.F., 1979. The ultimate attribution error: extending Allport's cognitive analysis of; prejudice. Personality and Social Psychology Bulletin 5, 461–476.

Pettigrew, T.F., 1990. A bold stroke for personality a half-century ago: a retrospective review of Gordon Allport's Personality: a psychological interpretation. Contemporary Psychology 35, 533–536.

Pettigrew, T.F., Tropp, L.R., 2011. When Groups Meet: The Dynamics of Intergroup Contact. Psychology Press, New York.

Piekkola, B., 2011. Traits across cultures: a neo-Allportian perspective. Journal of Theoretical and Philosophical Psychology 31 (1), 2–24.

Rosenzweig, S., 1970. E. G. Boring and the Zeitgeist: Eruditione gesta beavit. Journal of Psychology 75, 59–71.

Zuroff, D.C., 1986. Was Gordon Allport a trait theorist? Journal of Personality and Social Psychology 51, 993–1000.

Alternative Food Movements

Liz Grauerholz and Nicole Owens, University of Central Florida, Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

We review several alternative food movements (AFMs) that have emerged in response to changing food production during the twentieth century. The shift from small-scale, family-owned operations to massive, corporatized factories has resulted in concerns about human health, environmental degradation, and animal welfare; and it is around these issues that AFMs have emerged. The rise in AFMs seems to signal a growing desire for more personal connections between consumers and producers, and between producers and their products. Here, we discuss some of the established AFMs as well as newly emerging ones.

Introduction

New social movements surrounding food systems, especially the production of food products and their consumption, have grown dramatically over the past several decades. These movements emerged primarily as a response to a food production system that has become increasingly industrialized, standardized, and impersonal since the mid-twentieth century, when food production shifted from mostly small-scale, family-owned operations to large, corporatized farm factories, motivated by the same profit principles found in any other type of business. Outputs dramatically increased with these shifts but so did some disturbing outcomes (Lyson, 2004). As a result, concerns about the health – of humans, other animals, and the environment – as well as the ethics and risks of contemporary food production practices have been raised. It is around these issues that alternative food movements (AFMs) have emerged over the past 30–40 years.

Background

Since the mid-twentieth century, farming has transitioned from family-owned, diversified, and independent farms that raise a variety of crops and a few animals to farm factories that produce one crop or animal in intensive, confined space and are owned by a few corporations rather than families (Pew Commission on Industrial Animal Farm Production, 2009). This change in food production systems was prompted primarily by the 'Green Revolution' during the 1940s–1960s that relied on "a regime of genetic selection, irrigation, and chemical fertilizers and pesticides" (Pew Commission on Industrial Animal Farm Production, 2009: p. 3). The Green Revolution resulted in substantial increases in agricultural productivity. Crop yields doubled or tripled between 1940 and 1980. Similarly, meat, milk, and egg production increased dramatically, thanks in part to the low cost of feed grain, genetic selection, and chemical additives to feed (Pew Commission on Industrial Animal Farm Production, 2009).

During the latter part of the twentieth century, agriculture was further transformed by the 'Biotechnology Revolution,' as genetically modified (GM) plants and animals were introduced, allowing producers to breed for specific qualities in plants and animals (Lyson, 2004). Both the Green Revolution and Biotechnology Revolution promised and succeeded in generating greater output of animals and plants at lower costs, which has been a boon for many food producers as well as consumers around the world who need affordable food.

AFMs have emerged in large part in response to the Green Revolution and to a lesser but significant extent, the Biotechnology Revolution. The current popularity of farmers' markets (FMs), inner-city markets, community supported agriculture (CSA), resistance to mass-produced and biotech foods, and so on, signal a return of food production and marketing practices common prior to the Green Revolution. Until the emergence of large retail chain grocery stores (supermarkets) in the early twentieth century, both urban and rural dwellers often produced their own food and bought or bartered the rest from local farmers and specialized vendors (e.g., butchers). What characterized these food markets or practices, or direct agricultural markets, were face-to-face interactions between consumers and producers (farmers) (Hinrichs, 2000). With the emergence of technological advances in farming, chain groceries, and urbanization, these direct agricultural markets and practices were pushed to the margins, severing the personal connection between producers and consumers. In almost all cases, the rise in AFMs signals a return to more personal connections between consumers and producers and between producers and their products.

The Local Food Movement

The local food movement privileges local, seasonal foods over exotic, durable ones (Hinrichs, 2000). Although 'local food' has many meanings, these basic principles have had a major impact on contemporary food movements. They have forged new connections between producers and consumers and those between individuals and the land, and have reshaped the social, economic, and physical landscapes of thousands of communities. Lyson (2004: p. 1) calls the local food movement 'civic agriculture', drawing attention to the tight link between these farming practices and a community's social and economic development.

The local food movement has multiple origins but one of the most important is environmental. Over the past 50 years, food from around the world has become widely available throughout the year – bottled water from Fiji, apples from Chili

in springtime, watermelons from China in winter – are featured in virtually any grocery stores catering to middle- and upper-class shoppers. As a consequence, food may travel hundreds or thousands of miles from field to plate, requiring enormous amounts of nonrenewable resources to process, package, and transport, making food production and marketing two of the most environmentally unfriendly businesses in the world. Here, we discuss the major players in the local food movement – FMs and CSAs – as well as more recent trends that have taken on a do-it-yourself quality that includes urban farming and gardening.

FMs emerged as a major player in the contemporary local food movement in the 1970s. In the USA, there were 1755 FMs in 1994; that number increased to 8144 in 2013 (USDA, 2002, 2013b). FMs are found worldwide; some of the largest operate in Japan and Australia. These FMs are supported by consumers who desire greater freshness and quality of products than typically found in stores, hold idealized images of 'local farmers' and less conventional agricultural practices, and who want to help local economies and farmers and feel more connected to their communities (Ostrom, 2006; Smithers et al., 2008).

In theory, FMs involve producers who "sell directly to retail customers … selling legal goods produced by small scale firms outside of the mass distribution system" (Tiemann, 2004: p. 54). But as the popularity of FMs has grown, so too has their diversity. There exist 'indigenous markets' that operate with no formal rules yet adhere to selling only local, seasonal product, as well as 'experience markets' governed by written, formal rules (Tiemann, 2004). In British Columbia, for example, 'make-bake-grow' restrictions at many markets prohibit resale of items (Wittman et al., 2012). FMs may vary from just a few vendors to massive markets that no longer resemble what most think of as an FM. For example, the Los Angeles' Original Farmers Market began in 1934 with 18 farmers and merchants who paid 50 cents a day to park their trucks on vacant land to sell their products to local shoppers; today it offers not only local products but chain restaurants, gourmet groceries, and even online shopping.

FMs are likely to continue to grow because they benefit all parties. It is less costly for farmers to sell directly to consumers, and the reduction of transportation, storage, and processing costs and time means that consumers have access to fresher, healthier foods, usually at lower costs than found in grocery stores. Unquestionably, part of the popularity of FMs stems from their ability to bring communities together and create distinctive cultures, especially for those who feel that urban areas have become too impersonal and commercialized.

Another component of the local food movement is CSA. Although there is wide variability in how CSAs operate, typically consumers purchase an initial membership or subscription and in return, receive a 'share' of the farmer's bounty – seasonable yield – throughout the growing season. What makes CSA unique is the concept of 'shared risk.' Consumers shoulder some of the burden typically carried by farmers alone. In a good year, when yield is plentiful, consumers may get more lettuce and rhubarb than expected (or desired). In other years, if crops are destroyed by disease, disaster, or weather, consumers may get very little in return for their investment.

CSAs can be traced back to community-based farming systems in Japan and Switzerland during the 1960s and 1970s. In 1965, a group of Japanese mothers created the first CSAs, or 'Teikei' (Fairshare CSA Coalition, 2013). In the USA, the first two CSAs began in 1986 but quickly spread. The USDA reported that 12 549 farms marketed their products through a CSA in 2007 (USDA, 2007). According to the Rodale Institute (2013), CSA exists in Australia, Hungry, India, Hong Kong, France, Denmark, Holland, Germany, England, Brazil, Argentina, and Venezuela. In Japan, the Teikei movement is still thriving.

In addition to promoting deeper ties between farmers and consumers, CSAs often provide important benefits to the wider local community. In a survey of CSA farms (Center for Integrated Agricultural Systems, 2013), nearly three out of every four farms held social and educational events for members and the nonmembers, and more than half provided programs for low-income families and donated unclaimed food-to-food banks.

Urban gardening, which has existed since the 1890s and was especially popular during economic depressions and wartime (Lawson, 2005), has resurfaced as one of the more recent trends within the local food movement. While both FMs and CSAs center on the connection between farmer and consumer, urban gardening eliminates the 'middleman' altogether and turns consumers into producers. These include backyard and community gardens, and even small-scale farming of animals.

According to Lawson (2005), neighborhood community gardens are the most common type of urban gardening today. Typically, individuals sign up for a plot of land on which they grow anything from staples to feed their families to gourmet edibles not available in stores (Lawson, 2005). Community gardens have sprung up in large urban areas, slums, suburbs, and small villages and towns. Several countries, including China, Japan, and Cuba, have well-established, successful community gardening programs (e.g., Havana, Cuba has nearly 200 city farms) (Shaikh, 2013). On the other side of the world in a Kenyan slum where land and food is scarce, residents are learning how to grow food in large bags of compost (Shaikh, 2013).

Two more recent trends in urban farming and gardening are worth noting. The edible frontyard has generated considerable attention in recent years in the US. Though 'edible landscaping' has been popular for years and common in many parts of the world, these frontyard gardens typically replace entire lawn with garden beds, a practice guaranteed to offend some neighbors and city officials. In fact, frontyard gardeners have been ordered to remove their gardens or risk fines and jail (Kurutz, 2012).

Perhaps even more controversial is the recent trend in urban chicken keeping, a practice that is gaining greater acceptance in the USA. According to a USDA (2013c) survey, about 44% of those surveyed were accepting of chickens in their communities, 39% said they would not object to their neighbors raising chickens, and 4% of households planned to have chickens. Interestingly, chicken keeping has been found to restructure the family division of labor in some parts of the world by bringing children and women back into family production. In Botswana, for example, where women have assumed the lead in raising chicken for food and profit, their status has increased as

they come to be viewed as independent, productive, and entrepreneurial (Hovorka, 2012).

The local food movement has clearly become part of the food system and is likely to continue to do so, especially since it overlaps with consumers' desire for healthier, plant-based, organic products. Whether the newer manifestations of the local food movement such as backyard chicken coops and frontyard gardens survive remains to be seen and may well depend on the ability of movement activists and gardeners to frame these practices in terms of broader social issues such as children's and families' health and sustainability. But there are other challenges to the local food movement. Local farming and gardening are seasonal activities and the types of produce grown are limited. Today's consumers have come to expect and appreciate the diversity and availability of nonlocally grown foods, and few families today have the time or skill to preserve foods. It is unlikely that families can survive on local yield alone, thus local foods are likely to supplement rather than replace conventional food production and consumption.

Plant-Based Diets

Individuals who transition or convert to plant-based diets usually cite concerns about health, animal welfare, and the environment (Beardsworth and Keil, 1992). Although individuals may see these concerns as distinct, in reality they are closely related – largely stemming from the mass production of animals as food that have characterized meat production in the US and to a lesser extent, around the world, for several decades.

Although the percentages of individuals who consider themselves to be vegetarian and vegans are relatively small (5 and 2%, respectively; Gallup, 2012), many more individuals claim to incorporate plant-based foods in a deliberate way. According to one study, 40% say that when they eat out they often order a vegetarian dish, and 16% say that more than half their meals are vegetarian (The Vegetarian Resource Group, 2011). An NPR (National Public Radio)-Truven Health Analytics Health Poll (2013) also found that 39% of those surveyed said they eat less meat than they did even 3 years ago. In the spirit of supply and demand, plant-based options are more and more likely to appear on restaurant menus; Rowe (2013) cites an 11% increase in a 3-year period. The use of 'meat alternatives' (faux-meats) is increasing and vegan foods (i.e., containing 'no animal ingredients') saw the biggest gains (200%) in new products marketed between 2008 and 2012 (Food Product Design, 2013).

Although the rates of vegetarianism have not been well documented over time (compounded by the fact that there are no shared definitions of 'vegetarian' among those practicing it), these numbers almost certainly reflect a growing awareness of the benefits of adopting more plants into one's diet. Indeed, public messages promoting plant-based diets have intensified and stem from multiple sources – for example, animal rights activists and health advocates – thereby reaching more individuals and gaining momentum.

If a social movement's impact can be measured by how well its message has been incorporated into mainstream culture, the plant-based diet movement has been highly successful. For instance, the Academy of Nutrition and Dietetics, the world's largest organization of nutrition and food professionals, states that vegetarian diets have a wide range of health benefits including reduced risk of heart disease, cholesterol levels, blood pressure, hypertension, diabetes, obesity, and cancer. The 'Meatless Mondays' campaign is associated with prestigious universities including Johns Hopkins, Columbia, and Syracuse, and it works with schools, universities, government organizations, businesses, hospitals, and media around the world. According to Barclay (2013) about 43% of Americans were aware of the Meatless Monday campaign and of those, 36% said it caused them to reduce meat consumption or consider doing so. Its impact is not just felt in the United States; their Web site boasts that 29 countries have active Meatless Monday programs.

In addition to health issues, concerns about animal treatment and environmental degradation associated with current farming practices have also spurred interest in plant-based diets. These concerns stem mostly from the practice of producing animals in Concentrated animal feeding operations (CAFOs). CAFOs are a type of Animal Feeding Operation distinguished by the number of animals, conditions of confinement, and waste management utilized. These facilities "congregate animals, feed, manure and urine, dead animals, and production operations on a small land area" (EPA, 2012). Although CAFOs produce inexpensive food in greater abundance, the high concentration of animals in industrialized facilities has resulted in serious health concerns, environmental degradation, and animal mistreatment.

The most critical health issue stems from the amount of manure CAFOs produce (Hribar and Schultz, 2010). Although manure in small quantities is beneficial to farms, the quantities produced by CAFOs far exceed manageable levels. The Environmental Protection Agency (EPA) acknowledges that livestock animals in the US produce about 1.2–1.37 billion tons of waste every year – far more than that produced by humans. For example, a single facility with 800 000 pigs produces about 1.6 million tons of waste annually – more than one and a half times that produced by the city of Philadelphia, Pennsylvania (Hribar and Schultz, 2010). Unlike human waste, livestock waste is not treated or well contained (Hribar and Schultz, 2010), posing serious health hazards. Animal manure contains more than 150 pathogens that potentially affect human health, including nitrates and *Escherichia coli*, which can leach into the ground and surface water (Hribar and Schultz, 2010). CAFOs also affect air quality, as they produce high levels of "ammonia, hydrogen sulfide, methane and particulate matter, all of which have varying human health risks," including respiratory problems and death (Hribar and Schultz, 2010: p. 5). In addition, the extensive use of antibiotics in animal feed contributes to antibiotic-resistant bacteria, and inbreeding for desirable traits reduces biological and genetic diversity that threatens food security (Pew Commission on Industrial Animal Farm Production, 2009).

Since the primary goal of CAFOs is profit, not animal welfare, it is no surprise that the treatment of animals at all stages of production – rearing, transport, and slaughter – has

raised ethical and moral concerns. Increased awareness of both routine treatment and conditions within CAFOs (e.g., inability of animals to exhibit natural behaviors, physical alteration of animals without pain relief, close confinement) and excessive cruelty and neglect of captive animals (often revealed by undercover investigations) have led consumers to demand changes in animal treatment. One study found that 75% of the public would like the US government to impose mandates for basic animal welfare measures (cited in Pew Commission on Industrial Animal Farm Production, 2009).

Individuals who adhere to plant-based diets may frame their choices differently. Some focus mostly on the consumption angle, focusing on their own health; others focus on the production angle, wishing to avoid practices that harm the environment and create animal suffering. They also do not necessarily share interests with those engaged in other AFMs. That is, vegetarians are not necessarily opposed to intensive farming practices of crops or GM crops. Despite these differences, collectively they have succeeded in raising public awareness that reduction or elimination of meat in favor of plants is beneficial to human health, the environment, and animals.

Genetically Modified Organisms: Emergence and Resistance

Genetically modified organisms (GMOs) refer to plants or animals whose genetic materials have been altered by adding, removing, or changing their DNA sequences in ways that are not naturally occurring (European Food Safety Authority, 2013). Commonly, GMOs involve the transfer of DNA from one species (e.g., bacteria, virus) into another in order to introduce desirable traits such as pest resistance, higher nutritional value, or enhanced growth. Because these traits are introduced into a plant's or an animal's genetic makeup, they are transferable to the next generation. Although both plant- and animal-based food sources have been genetically engineered, the influx of GMOs into the food system has resulted mostly from plants, primarily maize, soybean, and rapeseed.

GMOs have been deemed the "fastest adopted crop technology in the history of modern agriculture" (James, 2011). The first GM crops (corn and soybean) were introduced into the US market in the mid-1990s and between 1996 and 2011, there was a 94-fold increase in hectares planted with GM crops worldwide (1.7 and 160 million ha, respectively) (James, 2011). According to Holst-Jensen et al. (2012), there are dozens of GM species in the market or in the pipeline, and there has been a large increase in the number of countries growing GM crops, the diversity of traits being introduced and global trading of GM products.

There has been intense debate, especially in the decade and a half since 2000, between those advocating the advancement of GMOs and those opposed to either the production of GMOs or their marketing (or both). Those advocating for widespread use of GMOs argue that GM foods pack both environmental and health benefits. In theory, yields from GM crops are better and farmers could use fewer chemicals if plants were pest resistant. If GM fish or livestock are treated to mature faster or produce larger edible parts, the environmental destruction caused by overfishing or factory farming may be reduced. In terms of health benefits, GM foods can be engineered to be more nutritious and plentiful, thereby helping to alleviate world hunger and malnourishment.

Those opposed to GMOs argue that such outcomes have not been realized. Rather than requiring fewer chemicals, insects, viruses, and bacteria become resistant, resulting in the need for new chemicals to fight new diseases and continuous genetic reengineering, often involving multiple ('stacked') transgenetic traits (Buiatti et al., 2013). Of major concern is gene flow between GM crops and non-GM crops, and the resulting reduction of biodiversity. Inevitable gene flow also has threatened conventional and organic farmers (the biotech corporation Monsanto sued over 100 farmers in North America for patent infringement because Monsanto-patented seeds drifted into these farmers' fields); this is especially problematic for organic farmers because a US ruling determined that certified organic products cannot contain GMOs. From a health perspective, there has been increasing concern about the possibility of gene transfers into humans such as allergens, unhealthy or even fatal bacteria in the gastrointestinal tract, or into human cells (Buiatti et al., 2013). The introduction of GM foods into the world market has also not made a serious dent in world hunger or malnutrition. While crop productivity has increased, the USDA acknowledges that this was the result of shifts in industrialized farming practices, not GMOs (Buiatti et al., 2013: p. 265). Further, GMO crops tend to be produced and consumed in the richest, not poorest, countries.

Worldwide, there are widely different approaches to the introduction of GMOs in the market. At one end, there are countries such as the United States where GM crops have been widely adopted and introduced into a wide variety of food products with no labeling and no reliable method available to consumers for tracking GMOs in their foods. At the other end are countries such as France that have approached GMOs with considerable skepticism, banning certain GM crops from production and all GM animals from market (EFSA, 2013). Although few countries ban imports or cultivation of GM foods, as of 2013, 64 countries including China, Japan, Brazil and all of the EU members required labeling of GM foods (Center for Food Safety, 2013).

Permaculture

A growing movement surrounding food production is *permaculture*. The concept behind permaculture – permanent agriculture or permanent culture – is developing self-sustaining systems that operate according to natural laws. Despite its roots in natural systems that existed for millions of years, the practice of permaculture marks a radical shift from other AFMs and certainly conventional farming practices. For example, just as one is unlikely to observe monocrops in nature, in a permacultural landscape, animals, plants, and humans exist cooperatively. Rather than operating from the view that nature is an enemy, a view common in modern agriculture (e.g., 'weeds' and 'pests'), permaculturalists work with nature rather than against it to build a sustainable

ecosystem that "uses and melds the best features of whatever is available to it," even waste (Hemenway, 2009: p. 5). In permaculture, waste products become resources, as in collecting water runoff from an air-conditioning system to water plants or using garbage to create fences.

Although the concept originated in the early twentieth century, its recent emergence can be traced to the work of Bill Mollison and David Holmgren in Tasmania. Their observations of the rainforest, one of the richest ecosystems on earth, led to the proposition that human communities can be developed to be as sustainable, productive, and rich as those found in the natural world. Permaculture is about design rather than any particular method or system. Hemenway (2009: p. 5) suggests that "If we think of practices like organic gardening, recycling, natural building, renewable energy, and even consensus decision-making and social-justice efforts as tools for sustainability, then permaculture is the toolbox that helps us organize and decide when and how to use those tools."

Permaculture has been adopted on small scales such as backyard landscapes, as well as regions or countries. Cuba, for example, turned to permaculture after the collapse of the Union of Soviet Socialist Republics (USSR) (its main trading partner) and the United States tightened embargos on food imports to Cuba. According to Roberto Perez (The Permaculture Research Institute, 2013), faced with a hunger crisis, Cuba turned to sustainable farming and now leads the world in permaculture. In other cases, deserts in the Middle East have been greened through basic permaculture techniques. The success of such programs and their focus on sustainability are likely to make permaculture one of the most important global food movements of the future, spurred on by grassroots activists working to help communities ravaged by war, poverty, and environmental degradation, or cut off from the land by urban development and industrial agriculture.

Food Justice

Many AFMs have been criticized for ignoring racial and class inequalities, and failing to recognize that disadvantaged groups and communities are most likely to be harmed by current practices (Alkon and Agyeman, 2011). For instance, CAFOs are disproportionately located in areas populated by low-income and non-White residents, and these facilities further depress local economies by concentrating profits into the hands of a few rather than spreading them throughout the community (Kelley et al., 2007). Poorer communities have limited geographical access to healthy foods and are less able to afford them. Farm workers are disproportionately people of color while those who benefit from the system (owners and managers) are predominantly White (Carrasquillo, 2011; USDA, 2013d). The 'food justice movement' attempts to bring these inequalities to light, and places emphasis on fair trade, localism, affordability, equity, and justice (Alkon and Agyeman, 2011).

According to Wekerle (2004), there are three strategies used by the food justice movement to address these injustices. First, it emphasizes the local- and community-based initiatives, drawing on local food movement practices (community gardens/kitchens, youth education, etc.), in order to promote food security from the ground up. Second, it seeks to engage in local, regional, and national policy making and planning to promote access to healthier foods for all. Third, it seeks to create a 'networked food justice movement' that is fluid yet able to exist with limited funding and structure (Wekerle, 2004: p. 381).

There are many such groups, each with its own agenda, messages, and goals. For example, the *Growing Food and Justice Initiative* is a multiracial coalition that works to create sustainable food systems by and for people of color, such as a group of Latino/Latina gardeners in South Central Los Angeles who grow foods indigenous to their native cultures, such as corn, beans, and squash, and participate in transnational seed exchanges (Alkon and Agyeman, 2011). Other food justice groups may target different problems or populations. For instance, *Fair Trade USA* works to ensure that workers earn decent wages and that farmers receive a fair price for their goods (Fair Trade USA, 2013). *Food Not Bombs* cooperate with local groceries and bakeries to collect unsold vegan foods and prepare and serve these to individuals in need. Despite its range of focus, what identifies the food justice movement is empowerment of and coalition building among those groups most negatively affected by contemporary practices.

Challenges and Considerations

Because food is so central to identity, culture, history, and health, and conventional agricultural practices are widely understood as unsustainable, AFMs will continue to grow in numbers and influence. There are clear challenges, however, most notably the AFMs' ability to address "issues of equity, food security, class, gender, and race" (Goodman, 2000: p. 217). Many of the people who are able to participate in alternative food practices (obtaining local, organic, and/or non-GMO food at FMs, CSAs, or from their own lawn and community gardens) are not those most affected by the industrialized food system. Many involved in food movements emphasize educating people which can result in 'telling others what to eat,' rather than focusing on access, affordability, and structural inequality that make AFMs unsustainable and difficult to develop (Allen et al., 2003; Guthman, 2008: p. 436). Furthermore, while participation in alternative food has grown rapidly and receives major support across the global North, alternative food is only beginning to be explored in the global South (Freidberg and Goldstein, 2010). Whether alternative movements will transform the way food is produced and consumed both locally and globally through sustainable, local, fair, and equitable practices is yet to be seen.

A central concern is whether the growing world's population can be fed through nonintensive farming methods and therefore whether methods promoted by AFMs are realistic alternatives to conventional 'factory farming.' Populations with resources and opportunities to participate in alternative food are more able to sustain themselves using nonintensive farming methods, while those less privileged must often rely on food that is produced from industrialized agriculture, such as GM foods. Thus, focusing on eliminating larger social disparities is key to feeding the world through sustainable, ethical farming methods, and

social movements, like AFMs, have been central to addressing these social problems. In this sense, what might make them successful – elimination of structural inequality rather than promotion of any specific farming method or consumption pattern – would also lead to their demise.

AFMs are not likely to replace or seriously challenge conventional, industrialized methods of food production (Lyson, 2004). However, industrialization and alternative food practices do not necessarily have to be mutually exclusive, on opposing sides of a political agricultural continuum. Jarosc (2008: p. 242) argued, "globalization of food is part of the development of local food systems." For example, urbanization is accompanied by more FMs and residents demanding access to local food (Jarosc, 2008). Also, AFMs have had enough impact on the economy that even conventional grocery stores carry organic, non-GMO, and local products. In the US alone, the USDA reported that annual local food sales were dollar;4.8 billion in 2008 (USDA ERS, 2012) and organic food sales were estimated at dollar;27 billion in 2012 (USDA ERS, 2013).

Despite the challenges AFMs face, they are likely to have an even stronger and visible presence in food production and distribution for years to come, particularly if they adequately address the biggest concerns and criticisms concerning their commitment to social justice and sustainability. As technological advancements in food production continue, the conventional food system will be faced with new challenges, prompting even more AFMs to emerge and confront changes brought on by these innovations.

AFMs will persist, as will questions about how these movements will operate and influence society in the coming decades. One area ripe for scholarship is the way technology will transform food systems and movements. One of the most interesting recent developments is laboratory-raised meat, meat that is grown by using stem cells. Since no animals are killed through this process, this development has implications for the environment, social movements, and identities. For instance, would introduction of lab meat substantially reduce global warming and devastation caused by CAFOs and fulfill the food revolutions' promise to reduce world hunger? How might this development affect the agendas and effectiveness of animal rights groups? Can one still be a 'vegetarian' if one eats meat raised in labs?

Other topics for research are the tensions between corporate and public interests surrounding health and food issues, and how these play out on national and global stages. For example, in 2014 General Mills announced that it would no longer use GM sugar and corn in Cheerios™, a decision most likely prompted by food advocates, and potentially forcing other corporations to follow suit, similar to that which occurred around the organic food movement. Further, if some countries adopt stricter laws concerning biotech foods, what impact will this have on corporations and markets in more lax countries such as the United States? It will also be interesting to analyze the shifting political economies surrounding food movements, including the roles governments play in supporting or opposing AFMs (and public health) and the power AFMs have (or do not have) to reshape the political economies at local, national, and global levels. Beyond the political and economic realms, AFMs may help transform social relations both globally and locally. As the risks associated with contemporary food systems intensify, we may see more cooperation and communication between the Global North and Global South around shared interests in safe foods. Thus, an area of concern is not only whether AFMs can address inequalities within wealthiest countries, but can and will they transform the relationships between wealthy and poorer countries?

See also: Climate Change and Global Poverty; Consumption, History of; Food Deserts in US Cities; Food Preferences, Psychology and Physiology of; Food Production, Origins of; Food Security and 'Green Revolution'; Food in Anthropology; Food, Sociology of; Geographies of Food: Politics, Place, Power Reprised; Slow Food as a Social Movement; Smart Cities; Social Movements, History of: General.

Bibliography

Allen, P., FitzSimmons, M., Goodman, M., Warner, K., 2003. Shifting plates in the agrifood landscape: The tectonics of alternative agrifood initiatives in California. Journal of Rural Studies 19 (1), 61–75.

Alkon, A.H., Agyeman, J., 2011. Cultivating Food Justice: Race, Class, and Sustainability. The MIT Press, Cambridge.

Barclay, E., 2013. Meatless Monday Movement Gets More Veggies on Your Plate [online] National Public Radio. http://www.npr.org/blogs/thesalt/2013/10/21/239126655/meatless-mondaymovement-gets-more-veggies-on-the-menu (accessed 7.4.14.).

Beardsworth, A., Keil, T., 1992. The vegetarian option: varieties, conversions, motives, and careers. The Sociological Review 40 (2), 253–293.

Buiatti, M., Christou, P., Pastore, G., 2013. The application of GMOs in agriculture and in food production for a better nutrition: two different scientific points of view. Genes & Nutrition 8, 255–270.

Carrasquillo, N., 2011. Race and ethnicity from the point of view of farm workers in the food system. Race/Ethnicity 5 (1), 121–131. http://www.cata-farmworkers.org/Nelson%20Article.pdf (accessed 20.12.13.).

Centers for Disease Control and Prevention, 2013. Nutrition for Everyone [online] http://www.cdc.gov/nutrition/everyone/fruitsvegetables/index.html (accessed 20.12.13.).

Center for Food Safety, 2013. Genetically Engineered Food: The Labeling Debate. http://www.centerforfoodsafety.org/files/labeling_fact-sheet_35682.pdf.

Center for Integrated Agricultural Systems (CIAS), 2013. Community Supported Agriculture across the Nation: Findings from the 1999 and 2001 CSA Surveys [online] http://www.cias.wisc.edu/economics/csa-across-the-nation-findings-from-the-1999-and-2001-csa-surveys/ (accessed 15.10.13.).

European Food Safety Authority (EFSA), 2013. Genetically Modified Animals [online] http://www.efsa.europa.eu/en/topics/topic/gmanimals.htm (accessed 10.10.13.).

Fair Trade USA, 2013. Vision Statement [online] http://fairtradeusa.org/about-fair-trade-usa/mission (accessed 20.12.13.).

Fairshare CSA Coalition, 2013. History of CSA [online] http://www.csacoalition.org/our-farms/about-csa/history-of-csa/ (accessed 15.10.13.).

Food and Agriculture Organization of the United Nations, 2013. About FAO [online] http://www.fao.org/about/en/ (accessed 20.12.13.).

Food Product Design, 2013. More Non-vegetarians in U.S. Consuming Meat Alternatives [online] www.foodproductdesign.com (accessed 17.12.13.).

Food Tank, 2013. 40 Organizations That are Shaking up the Food System [online] http://foodtank.org/news/2013/05/forty-organizations-that-areshaking-up-the-food-system (accessed 20.12.13.).

Freidberg, S., Goldstein, L., 2010. Alternative food in the global south: reflections on a direct marketing initiative in Kenya. Journal of Rural Studies 27 (1), 24–34.

Goodman, D., 2000. Organic and conventional agriculture: materializing discourse and agro-ecological managerialism. Agriculture and Human Values 17 (3), 215–219.

Gallup, 2012. In U.S. 5% Consider Themselves Vegetarians [online] http://www.gallup.com/poll/156215/consider-themselves-vegetarians.aspx (accessed 7.4.14.).

Guthman, J., 2008. Bringing good food to others: investigating the subjects of alternative food practice. Cultural Geographies 15 (4), 431–447.

Hemenway, T., 2009. Gaia's Garden: A Guide to Home-Scale Permaculture. Green Publishing, White River Junction, VT.

Hinrichs, C.C., 2000. Embeddedness and local food systems: notes on two types of direct agricultural markets. Journal of Rural Studies 16 (3), 295–303.

Holst-Jensen, A., Bertheau, Y., de Loose, M., Grohmann, L., Hamels, S., Hougs, L., Morisset, D., Pecoraro, S., Pla, M., Van den Bulcke, M., Wulff, D., 2012. Detecting un-authorized genetically modified organisms (GMOs) and derived materials. Biotechnology Advances, [e-journal] 30 (6), 1318–1335.

Hovorka, A.J., 2012. Women/chickens vs. men/cattle: insights on gender-species intersectionality. Geoforum 43 (4), 875–884.

Hribar, C., Schultz, M., 2010. Understanding Concentrated Feeding Operations and Their Impact on Communities [pdf]National Association of Local Boards of Health, Bowling Green, OH. www.cdc.gov (accessed 17.12.13.).

James, C., 2011. Global Status of Commercialized Biotech/GM Crops: 2011 [pdf] International Service for the Acquisition of Agri-Biotech Application, Ithaca, NY. http://www.isaaa.org (accessed 19.12.13.).

Jarosc, L., 2008. The city in the country: growing alternative food networks in metropolitan areas. Journal of Rural Studies 24 (3), 231–244.

Kelley, J.D., Wing, S., Osterberg, D., Flora, J.L., Hodne, C., Thu, K.M., Thorne, P.S., 2007. Community health and socioeconomic issues surrounding concentrated animal feeding operations. Environmental Health Perspectives 115 (2), 317–320.

Kurutz, S., 2012. The Battlefront in the Front Yard. The New York Times. http://www.nytimes.com/2012/12/20/garden/gardeners-fight-with-neighbors-and-city-hall-over-their-lawns.html?pagewanted=all&_r=0.

Lawson, L., 2005. City Bountiful: A Century of Community Gardening in America. University of California Press, Los Angeles.

Lyson, T.A., 2004. Civic Agriculture: Reconnecting Farm, Food, and Community. Tufts University Press, Lebanon, NH.

Non GMO Project, 2013. GMO Facts [online] http://www.nongmoproject.org/learn-more/ (accessed 20.12.13.).

Ostrom, M., 2006. Everyday meanings of 'local food': views from home and field. Community Development 37 (1), 65–78.

Oxfam America, 2013. Grow: Food. Justice. Planet [online] http://www.oxfamamerica.org/campaigns/food-justice (accessed 20.12.13.).

Permaculture Research Institute, 2013. How Cuba Leads the World in Permaculture [online] http://permaculturenews.org/2013/09/09/how-cuba-leads-the-world-in-permaculture-podcasts-parts-i-ii/ (accessed 24.11.13.).

Pew Commission on Industrial Animal Farm Production, 2009. Putting Meat on the Table: Industrial Farm Animal Production in America [online] www.ncifap.org (accessed 18.12.13.).

Rodale Institute, 2013. The History of Community Supported Agriculture, Part II [online] http://newfarm.rodaleinstitute.org/features/0204/csa2/part2.shtml (accessed 15.10.13.).

Rowe, M., 2013. Consumer Demand for Veggies Growing: Apparently those USDA Guidelines are Sinking In [online] Restaurant Hospitality. http://restaurant-hospitality.com/consumer-trends/consumer-demand-veggies-growing (accessed 7.4.14.).

Shaikh, T., 2013. Urban Farms Herald Green City 'Revolution' [online] CNN International. http://edition.cnn.com/2010/WORLD/europe/04/08/urban.farming.city.growing.food/index.html (accessed 17.10.13.).

Smithers, J., Lamarche, J., Joseph, A.E., 2008. Unpacking the terms of engagement with local food at the farmers' market: insights from Ontario. Journal of Rural Studies 24 (3), 337–350.

The Vegetarian Resource Group, 2011. How Many Adults are Vegan in the U.S. [online] http://www.vrg.org/blog/2011/12/05/how-many-adults-are-vegan-in-the-u-s/ (accessed on 7.4.14.).

Tiemann, T.K., 2004. American farmers' markets: two types of informality. The International Journal of Sociology and Social Policy 24 (6), 44–57.

United States Department of Agriculture (USDA), 2002. U.S. Farmers Markets – 2000 A Study of Emerging Trends [online] http://agmarketing.extension.psu.edu/ComFarmMkt/PDFs/emerg_trend_frm_mrkt.pdf (accessed 17.10.13.).

United States Department of Agriculture (USDA), 2007. Census of Agriculture [pdf] http://www.agcensus.usda.gov/Publications/2007/Full_Report/Volume_1,_Chapter_2_US_State_Level/st99_2_044_044.pdf (accessed 20.12.13.).

United States Department of Agriculture, 2013a. The National Institute of Food and Agriculture [online] http://www.nifa.usda.gov/ (accessed 20.12.13.).

United States Department of Agriculture (USDA), 2013b. News Release: USDA Celebrates National Farmers Market Week, 4–10 August, Confirms Growth and Sustainability in Farmers Markets [online] http://www.usda.gov/wps/portal/usda/usdahome?contentid=2013/08/0155.xml (accessed 17.10.13.).

United States Department of Agriculture (USDA), 2013c. Urban Chicken Ownership in Four U.S. Cities [online] http://www.aphis.usda.gov/animal_health/nahms/poultry/downloads/poultry10/Poultry10_dr_Urban_Chicken_Four.pdf (accessed 20.10.13.).

United States Department of Agriculture (USDA), 2013d. Farm Workers [online] http://www.ers.usda.gov/topics/farm-economy/farm-labor/background.aspx#demographic (accessed 20.12.13.).

United States Department of Agriculture Economic Research Service (USDA ERS), 2012. Direct and Intermediated Marketing of Local Foods in the United States. [online] http://www.ers.usda.gov/publications/err-economic-research-report/err128.aspx (accessed 28.11.13.).

United States Department of Agriculture Economic Research Service (USDA ERS), 2013. Consumer Demand Drives Growth in the Organic Food Sector. [online] http://ers.usda.gov/data-products/chartgallery/detail.aspx?chartId=35003&ref=collection (accessed 28.11.13.).

U.S. Environmental Protection Agency (EPA), 2012. What is a CAFO? [online] www.epa.gov (accessed 18.12.13.).

United States Environmental Protection Agency, 2013. Agriculture and Food Supply [online] http://www.epa.gov/climatechange/impacts-adaptation/agriculture.html (accessed 20.12.13.).

University of California Berkeley Library, 2006. International Organizations, Food & Agriculture: Digital Libraries and Databases [online] http://www.lib.berkeley.edu/doemoff/govinfo/intl/gov_agriculture.html (accessed 18.12.13.).

Wekerle, G.R., 2004. Food justice movements: policy, planning, and networks. Journal of Planning Education and Research 23 (4), 378–386.

Wittman, H., Beckie, M., Hergesheimer, C., 2012. Linking local food systems and the social economy? Future roles for farmers' markets in Alberta and British Columbia. Rural Sociology 77 (1), 36–61.

Alternative Schools of Economic Thought

Peter J Boettke and Solomon M Stein, George Mason University, Fairfax, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Alternative schools of economic thought are defined relative to the current mainstream of economic theory. This article examines the path of mainstream economics during the twentieth century and the challenges presented to the mainstream by various alternative schools of economic thinking, such as the desire for relevance of American institutional economics, the emphasis on market process and calculation given by Austrian economics, the role of gendered theorizing examined by feminist economics, and the treatment of uncertainty offered by post-Keynesian Economics.

Introduction

Economic thought, contrary to the timeless picture presented in an introductory textbook, is a complex and heterogeneous set of ideas. At any given time, some of those ideas will be ascendant among economists, while other ideas and methods will be out of step with the economic profession at large. Alternative schools of economic thought arise from a desire to explore those out-of-favor ideas, either with a view to repair or replace the currently fashionable. What is considered alternative depends on what is considered to be in the current mainstream, which has varied in scope widely between the turn of the twentieth century and the present.

Mainstream Economics

Boettke et al. (2008) document the changes in the composition of the mainstream, arguing that economics entered the twentieth century with a relatively broad set of ideas and methods in the mainstream, drawing upon the origins of economics in the work of classical political economy. (This core of classical political economy, a focus on exchange and the institutions within which exchanges takes place within the subset of economics that Boettke, Leeson, and Smith call the 'mainline' and trace back to Adam Smith and continued through the classical political economists.) Beginning in the interwar years, the view of economics as an element of that broader political economy tradition that included other sorts of social science and philosophy fades in favor of an emphasis in pursuit of the logic and operation of formal equilibrium theory. The questions and approach to economics characterized in this work came to define the mainstream, with the synthesis of Keynesian macroeconomics and neoclassical, and general-equilibrium microeconomics. The dominance of this approach in mid-century academic economics necessitated in turn the rediscovery of the broader set of questions of political economy, often in the context of particular fields or approaches, such as transaction cost economics, New Institutional Economics, or game theory. These approaches, along with other challenges both to the Keynesian macroeconomic paradigm and to the exclusive focus on general equilibrium in microeconomics, led over time (Perhaps, 'funeral by funeral.') to a mainstream inclusive again of broader approaches and to a lesser extent, a broader set of acceptable methods. As Boettke and Prychitko (2011) recount, the culture of economics in the 1980s was one that reflected this re-expansion of the scope of mainstream economic scholarship.

With this pattern of a broader set of questions and methods condensing into the neoclassical-Keynesian synthesis and then gradually broadening back out to find room for emerging (or perhaps, returning) fields in the later part of the twentieth century, it may be unsurprising that many of the well-developed 'schools' of alternative approaches to economics arise during the period when there was comparative openness to acceptance of new ideas in general; both those approaches that existed alongside what became the mid-century mainstream and found themselves suddenly on the outside looking in, such as the Austrians and American Institutionalism, or those approaches which represent views that, rather than being absorbed into the broader mainstream of the 1980s onward, remain self-consciously opposed to elements of current mainstream practice. The rest of this article focuses upon these groups individually, identifying their origins and how they see themselves as distinct from the current mainstream. While this is by no means an exhaustive list of heterodox thinkers or their ideas, it should give a good indication about the puzzles alternative approaches offer to the mainstream regarding theory and method.

American Institutionalism

Institutional economics refers to a group of American economists in the early twentieth century who to greater or lesser extent dissented from the state of neoclassical economic theory. The most prominent thinker now associated with this tradition is Thorsten Veblen, best known for *The Theory of the Leisure Class* (Veblen, 1899). While Veblen is the most prominent thinker associated with Institutionalism, as Rutherford (2011) recounts, his work was eccentric and considered inspirational, rather than central, to the larger community of economists interested in the institutional program. This community was critical of the neoclassical theory for being primarily focused upon equilibrium statics and resting upon unrealistic (that is, primarily hedonic) psychological assumptions. Instead of this 'value theory,' institutional economists saw themselves as concerned with the role of real-world institutions in shaping

economic reality, understanding institutional (particularly legal) evolution, and operating using more solid 'scientific' foundations both in terms of psychology and methodology.

In the paper that definitively established 'institutional' economics as the label attached to this school, Hamilton (1919) outlines five criteria on which he thought any set of ideas that could be called 'economic theory' ought to be judged. His criteria and why he saw value theory as falling short of their accomplishments present a succinct statement of the critical view of neoclassical theory held by institutional economists. Firstly, Hamilton (1919: p. 312) argued that "economic theory should unify economic science," by which he seems to mean theory should be able to include each field in which economists do research, arguing that 'value theory' does not reach into the majority of applied fields. Secondly, Hamilton (1919: p. 312) argues for relevance in economic theory, particularly to what he calls "the modern problem of control" in comparison to "a mere search for hypothetical truth" which he sees as being central to the abstract theoretical concerns of neoclassical economics. In his interpretation, theoretical development both in the classical economists' time and in his own was driven by the desire to answer questions of policy relevance, such as Ricardo's interest in monetary theory or (as we will see later in this article) the concern of the Austrian economists with the socialist calculation debate. Similar to those concerns, he argues that the relevant world-state which economic theory ought to be developed to answer is the "general demand for control" (Hamilton, 1919: p. 313). His third argument is that, given the two above claims, institutions are the appropriate level of analysis for economic theory. For Hamilton, this means an understanding of the particular forms of industrial society such as the particulars of corporate organization to understand the theory of the firm (in contrast to the more abstract view of neoclassical price theory) and places a strongly constructive emphasis: these particulars "must be recognized by one who would direct, or, if he be an economic theorist, watch, industrial development" (Hamilton, 1919: p. 314). Fourth, economics ought to see these institutions as a process, capable of conscious direction. (As we will see later, this view of process as meaning that it is amenable to rational construction provides an interesting contrast with the implications of the market as a process for Austrian economics.) Finally, Hamilton (1919: p. 316) calls for a grounding of economic theory in "an acceptable theory of human behavior," that is one built on what was considered to be the state of the art in psychological theory at the time. (This call for psychological realism closely resembles those made for the modern development of behavioral economics. Interestingly, the decline of institutional economics was based at least in part on a conception of them as being devoid of a systematic alternative theory to neoclassical economics, and excessively ad hoc in their introduction of psychological or institutional factors, a critique similar to that raised at behavioral economics by critics such as Posner (1998).)

In developing their alternatives, these considerations led many institutional economists to emphasize psychological theories such as instinct and habit psychology or Skinnerian behaviorism, often to critique marginal utility theories of consumption and to favor instead psychological theories of consumption formation. Institutional economists were also concerned with what Rutherford (2011: p. 47) describes as "a conception of the industrial economy as dominated by pecuniary or business interests" in a way that definitively shaped economic activity and thus provided a superior frame of reference than neoclassical price theory. Particular focus was given by economists working in the institutional tradition to the impact of ongoing technological change during the interwar period, such as J.M. Clark's work on the impact of overhead costs (Clark, 1923) and extensive work on labor unions (Commons, 1913, 1919).

One other significant commonality within the institutional movement was a focus on the role of economics as a *tool* of social control, (Over and above the idea that the questions of economics ought to deal with questions of control as suggested by Hamilton, this meant that the purpose of examining these problems was to actively apply the results to policy.) particularly within a turn-of-the-century progressive framework. Institutional economists saw social control as central to the project of economics in their day, and their ideas of what constituted a scientific approach to solving such problems was primarily empirical and quantitative and were central in the founding of institutions devoted to collecting the data needed to pursue this project, most notably the National Bureau of Economic Research (NBER), where Wesley Mitchell, a prominent institutionalist, was appointed Director of Research upon its founding in 1920. However, by the middle of the 1930s, the appeal of the institutional program had faded, and it was increasingly criticized as being unable to offer a systematic alternative paradigm to replace the neoclassical value theory they saw as unsatisfactory. The dominant view of the institutional program in retrospect, advanced by historians of thought such as Blaug and Stigler, reflected this view of a group unified only by dissent, but that view has been challenged recently by Hodgson (2004) and Rutherford (2011).

The Austrian School

The Austrian school of economics was founded in 1871 with the publication by Carl Menger of his *Principles of Economics*. Menger's goal in the publication of his *Principles* was twofold: it contained both an exposition of economics based upon marginal utility theory, which along with the work of Jevons and Walras would lead to the marginal revolution in economic theory, along with defending the idea that economic theory was a universally applicable approach to understanding human behavior. While the scope of economic investigation has been a subject of debate both within economics (For a summary of the ideas of economists on this point over time, see Kirzner (1960).) and between economists and other social scientists, there has seldom been an environment where Menger's statement could have been more controversial than the German academy of his day, where the dominant view was that of the German Historical School, which thought that economic behavior was historically contingent, and that particular epochs of historical development contained their own internal economic logic.

Menger and his Historical School colleagues became embroiled in a debate known as the *Methodenstreit*, or dispute

over method, during which Gustav Schmoller gave Menger and his followers the label 'Austrian School' as a means of belittling them relative to the 'German' historical school, which has stuck since. With respect to other marginalist economists, however, Austrians during those early years saw themselves as part of the broader economic tradition. During the interwar years, key debates within economics saw the Austrians, led by Ludwig von Mises and F.A. Hayek, develop as an increasingly distinct school of thought. The two most prominent of these debates were the debate over Socialist economic calculation and the debate between Hayek and Keynes over monetary policy. In both cases, in the eyes of the economics profession, the Austrian position was seen to be the losing one. Austrian economists in the wake of these debates (and, as a consequence of the Second World War, no longer in Austria) had to reconsider their relationship to the broader economics profession and found themselves increasingly out of step with the mathematically formalist and Keynesian mainstream of the 1940s and 1950s.

Interest in Austrian economics surged in the early 1970s, spurred by the publication of Israel Kirzner's *Competition and Enterpreneurship* in 1973, followed by Hayek's receipt of the Nobel Prize in Economics in the subsequent year. This Austrian revival began as centered on those scholars, as well as the work of Murray Rothbard, Ludwig Lachmann, and others. Along with those scholars, modern Austrian economists draw inspiration from many modern figures in economics who they see as working to return the mainstream of economics toward the mainline, such as Armen Alchian, James Buchanan, Ronald Coase, Harold Demsetz, Elinor Ostrom, Axel Leijonhufvud, Douglass North, Mancur Olson, Vernon Smith, Gordon Tullock, Leland Yeager, and Oliver Williamson.

Boettke (2008) presents ten propositions that broadly capture the substantive beliefs of Austrian economists about economic method and theory. They are as follows:

1. Only individuals choose:

This captures the idea that economics is what Mises (1948) called a *Praxeological* science, that is, it studies the behavior of acting persons, as only individuals make choices. Economics renders these choices intelligible and seeks to understand the unintended consequences of choice.

2. The study of the market order is fundamentally about exchange behavior and the institutions within which exchanges take place:

Within Praxeology, the science of human action, is the study of exchange, *cattalactics*, which places a focus away from what Buchanan (1964) called the allocative paradigm which sees the maximization problems of scarce means and competing ends as the primary focus of economic investigation, and toward the exchange paradigm which focuses more on the process of market behavior and the institutions within which exchange takes place.

3. The 'facts' of the social sciences are what people believe and think:

Since the subject matter of economics is human action, attempts to eliminate the subjectively understood meaning that actors place on their activities seems to the Austrian approach to remove the important content from activities. Austrian economists see the measuring stick for progress in economic science as rendering behavior intelligible, rather than purely by predictive power, rejecting the argument of Friedman (1953).

4. Utility and costs are subjective:

While the proposition that utility is subjective is one of the core elements of the neoclassical theory of value, Austrians reject the Marshallian position that costs are in some way objective. Since supplies of goods are based upon the alternative uses of scarce resources, and those alternative uses are subjectively determined at the moment that those resources are allocated to one use over an alternative, cost is just as subjective as value.

5. The price system economizes on the information that people need to process in making their decisions:

As most clearly stated by Hayek's 1945 article 'The Use of Knowledge in Society,' the price system functions as a method of information transmission, in which the price of any given item communicates knowledge regarding the conditions of the market in that item without any given actor needing to acquire any knowledge besides the price. This includes even items of local or tacit knowledge that could not be acquired by other means.

6. Private property in the means of production is a necessary condition for rational economic calculation:

Stemming from the socialist calculation debate of the 1920s, Austrian economists see private property and profit and loss as necessary conditions for economic resources to be rationally allocated. Mises's argument was that, absent this private ownership of capital goods, there would be no market in those goods, resulting in no money prices, without which there could be no way to determine the exchange ratios between capital goods required for rational calculation.

7. The competitive market is a process of entrepreneurial discovery:

Rather than the notion of competition as a static condition of a state of affairs, Austrian economics focuses upon the process of competition. The competitive market is characterized by constantly changing plans, as entrepreneurs search for new combinations of resources that will yield them additional profit. The feedback of actually realized gains and losses from these speculative activities brings the market closer to a situation of full coordination, while errors by those entrepreneurs create the opportunity for future correction.

8. Money is non-neutral:

Money, the commonly accepted medium of exchange, is a central component of the market process, since it functions as the unit by which prices and profit and loss are calculated. When government policy distorts the monetary unit, the whole system of prices, and thus the communication of information across the economy, is distorted along with it. Although changes in the monetary unit eventually lead to adjustments in prices to account for the change in monetary policy, these changes do not occur instantly or evenly throughout the economy, so money cannot be neutral.

9. The capital structure consists of heterogeneous goods that have multispecific uses that must be aligned:

Austrian economists stress the importance of dealing with capital as heterogenous and with multiple uses, which must be aligned into a structure of production. Production requires investment now for an uncertain future, and the value of those capital goods is determined by their ability to satisfy the final goods they will eventually be used to produce. The heterogenous nature of capital goods and their imperfect substitutability means that any given system of prices renders some alignments of capital profitable, and if revealed to be in error, cannot be costlessly reversed. Instead, resources will be lost in the meantime as costly realignments of capital take place.

10. Social institutions often are the result of human action, but not of human design:

Austrian economists view institutions as a central element of human activity, and emphasize the role of processes in economics, but in contrast to the emphasis on those ideas in American institutional economics that saw those facts as making economic activity amenable to social control, Austrian economists stress that many of these institutions are not the products of conscious direction, but rather emerge out of the unintended interaction of all the various human actions undertaken. The market is the paradigmatic case of such an unintended emergent result of individual action: no actor sets out to create the price system or to coordinate all economic activity, but rather to remove their felt uneasiness or improve their own position via exchange behavior. These 'invisible hand' phenomena are not restricted to the market, but have been fruitfully used as a way to understand law, language, scientific progress, and numerous other social phenomena.

These ten propositions provide a useful view into the propositions regarding the nature of economic theory and how to understand examination of human behavior that Austrian economists tend to emphasize. The degree to which this emphasis is seen as compatible with existing mainstream approaches varies significantly, leading to a lively internal debate on the role of even fundamental concepts, such as equilibrium, in praxeological approaches to economics. (For a defense of the usefulness of equilibrium, see Kirzner (1992), while for a defense of a non-equilibrium (in contrast to disequilibrium) economic approach see Buchanan and Vanberg (1991).)

Feminist Economics

The most recent of the alternative schools considered here, feminist economics only developed into a school of thought within economics during the early 1990s, after having been an approach in other social sciences for several decades. As feminism became a distinct movement and approach in the academy, scattered efforts were made to present a feminist critique of neoclassical theory. However, feminist economics as an organized alternative school emerged out of a conference at the American Economic Association in 1990, resulting in the formation of the International Association for Feminist Economics and the founding of the journal *Feminist Economics*.

Julie Nelson's 1995 article titled 'Feminism and Economics' in the *Journal of Economic Perspectives* presents an overview of the emerging field, outlining the challenges that a feminist approach to economics (as opposed to the perhaps more common feminist critiques of economics from outside the discipline) presents to economic theory. Drawing upon feminist arguments that the history of philosophy of science is dominated by a gendered perspective whereby the values of science have been associated with masculinity, feminist economics as presented by Nelson would try to move beyond this gendered set of judgments that associate 'masculine' economics with the preferable state of economic investigation. She outlines four areas where a more balanced, feminist view of economics would improve economic investigation: modeling, method, topic selection, and teaching.

In model development, Nelson argues that, at least for the most part, economic models reflect a particular conception of autonomous individuals who are always rational choosers, and never undergo periods where they require care for others, such as infancy or old age, and incorporates motivations such as fairness and dependence. She is critical of the defenses of traditional economic theory against these additional behavioral assumptions by economists such as Gary Becker or Bob Lucas on the grounds that their restriction of economic theory to models of *homo economicus* "are not, as they are often taken, statements of demand for high rigor, but rather are demands that androcentric biases be indulged" (Nelson, 1995: p. 137).

Nelson is similarly critical of the importance placed on mathematical formalism (a recurring theme in most alternative approaches to economic theory) and the logical purity of mathematics instead of the 'messy' empirics of less sophisticated statistical techniques or survey data, seeing the discomfort with these supposed losses of objectivity in data as stemming from a flawed, Cartesian notion of what constitutes objectivity and truth in science. Instead, to the feminist perspective being able to outline one's own framework at the outset and then proceed with the analysis is preferable to purporting to a false standard of detachment and objectivity.

With regard to topic selection, Nelson laments the general absence of families and activities generally undertaken outside of the market from the consideration of economic analysis, and notes that many feminists are skeptical of the value of the notable analysis of the family undertaken by Gary Becker. This neglect of the family is also evident in the composition of GDP, where predominantly female-undertaken domestic production is omitted, or in the treatment of human capital as predominately regarding the decisions of young adults considering the marginal personal investment in education, omitting previous decisions that shape human capital. Finally, a feminist approach to economics suggests an alternative pedagogical style with more of an emphasis on participatory, applied teaching which could contribute to a climate which induces more women to enter economics as a discipline, as opposed to being 'selected out' by androcentric theories that suggest that "economics is not for (or about) me" (Nelson, 1995: p. 146).

Along with presenting a feminist perspective on the state of mainstream economic approaches, feminist economists tend to be concerned with topics where they perceive the mainstream as neglecting such as the aforementioned domestic or

care work, or of general importance to feminists such as the impact of economic development and labor market inequalities, as well as examination of the gender distribution of economics itself.

Post-Keynesian Economics

The beginnings of post-Keynesian economics rest in the dissatisfaction with some economists who considered themselves followers of Keynes' theories from *The General Theory* with the neoclassical Keynesian synthesis that had become (as discussed earlier) the mainstream approach to understanding and applying Keynes' ideas. Building upon the work of Joan Robinson, Roy Harrod, Nicholas Kaldor, Michal Kalecki, and others, Eichner and Kregel (1975) presented post-Keynesianism as an alternative economic paradigm, citing as its unique features the role of history and time, placing theoretical importance on the distribution of income as contributing to economic outcomes, and more faithfully retained the approach to understanding a money economy contained in Keynes and in its acceptance of alternate (generally less restrictive) microeconomic foundations relative to neoclassical price theory. *The Journal of Post-Keynesian Economics* followed soon after in 1978.

Accompanying the substantive positions outlined by Eichner and Kregel and the extension of their project to present a more authentic Keynesian system, (For a survey of the state of the art in post-Keynesian modeling, see Asensio et al. (2012).) post-Keynesians have been vocal critics of neoclassical theories of macroeconomics, especially endogenous growth models and New Keynesian macroeconomics, as well as other elements of neoclassical theory based on their omission of the role of fundamental uncertainty on the real world.

Paul Davidson's *Reality and Economic Theory* (Davidson, 1996) presents a good summary of the post-Keynesian approach to uncertainty, as well as the view of reality post-Keynesian economists believe to be held by alternate approaches. "Most mainstream economists today," Davidson argues, "accept as a universal truth the existence of a predetermined reality that can be fully described by unchanging objective conditional probability functions" (Davidson, 1996: p. 479). This renders future outcomes (stochastically) predetermined and enables rational decisions by assuming (correctly) that a sample of the past and present represents also a sample of the future. In contrast to that model, Davidson presents Keynes as developing a model wherein individuals recognize some creative element to the future. For Davidson, almost all theories outside of the work of Keynes, American institutional economics, and post-Keynesianism rely upon an immutable reality at some point, if only in the long run. Acceptance of the (post-)Keynesian conception of a transmutable reality renders intelligible for Davidson the purpose of accumulating liquidity one does not intend to expend as a hedge against the permanently uncertain future. Uncertainty of this type also leads to the presence of what G.L.S. Shackle termed "crucial decisions," which create effects that render the original decision non-repeatable, the making of which is to Davidson the key entrepreneurial function. Far from being purely methodological concerns, for Davidson these distinctions have important relevance for economic policy, with the transmutability of the world implying an absence of a tendency to return to full employment, requiring Keynesian demand policies, which in turn would require some form of anti-inflation policies to stabilize the system.

Of the alternate schools considered here, Post Keynesians are perhaps the most self-consciously heterodox (although elements of the Austrian school also embrace heterodoxy) although recent work on the relationship of heterodoxy published in post-Keynesian journals (notably Colander et al., 2007, responding to Dequech, 2007) has urged self-consciously heterodox economics to focus less upon defining themselves with respect to 'neoclassical economics' and focus on commonalities with economics in general.

Conclusion

All the schools considered here share two defining features that, to a greater or lesser extent both by individual author within the school and over time, keep them from becoming part of the mainstream of economic thought. The first is that they are in general critical of at least some aspects of the language of mainstream economics itself, making adaptation of their arguments into the conventional style of the mainstream more difficult than for other approaches which, despite differences of method or topic nevertheless have become part of the mainstream. Secondly, at least at the level of the ideal-type of a school, each of these alternative approaches are (or were) seen as advocates for some particular set of policy prescriptions. While policy relevance is not itself a disqualification from the mainstream, the combination of a distinctive policy association and criticism of the mainstream approach can lead to a perception of these schools as advocacy as against a (perhaps illusory) objective approach to scientific investigation.

Alternate schools of economic thought such as those discussed earlier are distinct not only from the mainstream but also from alternative theories about economic topics, in that even in their criticism of various aspects of mainstream economic theory, even on fundamental issues these schools are still attempting an 'economic' approach to understanding a particular issue, in contrast to a non-economic approach to an issue within the economy. (While Austrians, post-Keynesians, Institutional Economics, Feminist Economics, and other alternative schools might disagree with the mainstream and with each other over the relevance of power relationships to labor demand, for example, none are attempting to see labor through some other category, as one might do viewing it as a Hegelian dialectic or considering labor markets from a religious or ethical perspective.) Finally, despite an attempt to present each school's subjectively perceived element of distinctiveness with respect to the mainstream, each of these groups of scholars is a highly non-homogenous and non-static classification, just as the 'mainstream' of economics represents a complex and dynamic body of methods, topics, and theories. Within each of these schools exists a lively debate on where exactly the fundamental aspects of their critique of the current

mainstream lie within their body of theory, the appropriate margins to repair or reject mainstream approaches, and the strategy by which to accomplish their goals, both academically and socially oriented.

See also: Behavioral Economics; Economics, History of; Experimental Economics; Gender, Economics of; Law and Economics; Macroeconomics, History of From 1933 to Present; Old and New Institutionalism in Economics; Public Choice Economics.

Bibliography

Asensio, A., Lang, D., Charles, S., 2012. Post Keynesian modeling: where we are, and where are we going to? Journal of Post Keynesian Economics 34 (3), 393–412.

Boettke, P.J., 2008. Austrian school of economics. In: Henderson, D. (Ed.), The Concise Encyclopedia of Economics, vol. 10. Library of Economics and Liberty. http://www.econlib.org/library/Enc/AustrianSchoolofEconomics.html. May 2013.

Boettke, P.J., Leeson, P., Smith, D., 2008. The evolution of economics: where we are and how we got here. The Long Term View 7 (1), 14–22.

Boettke, P.J., Prychitko, D., 2011. 1985: A defining year in the history of modern economics. The Review of Austrian Economics 24 (2), 129–139.

Buchanan, J., 1964. What should economists do? Southern Economic Journal 30 (3), 213–222.

Buchanan, J., Vanberg, V., 1991. The market as a creative process. Economics and Philosophy 7 (2), 167–186.

Clark, J.M., 1923. Studies in the Economics of Overhead Costs. University of Chicago Press, Chicago.

Colander, D., Holt, R.P.F., Rosser, J.B., 2007. Live and dead issues in the methodology of economics. Journal of Post Keynesian Economics 30 (2), 303–312.

Commons, J.R., 1913. Labor and Administration. Macmillan, New York.

Commons, J.R., 1919. Industrial Goodwill. Macmillan, New York.

Davidson, P., 1996. Reality and economic theory. Journal of Post Keynesian Economics 18 (4), 479–508.

Dequech, D., 2007. Neoclassical, mainstream, orthodox, and heterodox economics. Journal of Post Keynesian Economics 30 (2), 279–302.

Eichner, A., Kregel, J., 1975. An essay on post-Keynesian theory: a new paradigm in economics. Journal of Economic Literature 13 (4), 1293–1314.

Friedman, M., 1953. Essays in Positive Economics. University of Chicago Press, Chicago.

Hamilton, W., 1919. The institutional approach to economic theory. The American Economic Review 9 (1), 309–318.

Hayek, F.A., 1945. The use of knowledge in society. The American Economic Review 35 (4), 519–530.

Hodgson, G., 2004. The Evolution of Institutional Economics: Agency, Structure, and Darwinism in American Institutionalism. Routledge, London.

Kirzner, I., 1960. The Economic Point of View. Sheed and Ward, Kansas City.

Kirzner, I., 1992. The Meaning of Market Process. Routledge, New York.

Mises, L., 1948. Human Action: A Treatise on Economics. Fox & Wilkes, San Francisco.

Nelson, J., 1995. Feminism and economics. Journal of Economic Perspectives 9 (2), 131–148.

Posner, R.A., May 1998. Rational Choice, Behavioral Economics, and the Law. Stanford Law Review 50 (5), 1551–1575.

Rutherford, M., 2011. The Institutionalist Movement in American Economics, 1918–1947: Science and Social Control. Cambridge University Press, New York.

Veblen, T., 1899. The Theory of the Leisure Class: An Economic Study of Institutions. Macmillan & Company, New York.

Altruism and Prosocial Behavior, Sociology of

Rafael Wittek, University of Groningen, Groningen, The Netherlands
René Bekkers, Center for Philanthropic Studies, VU University Amsterdam, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.A. Piliavin, volume 1, pp. 411–415, © 2001, Elsevier Ltd.

Abstract

Prosocial behavior involves costs for the self and results in benefits for others. Altruistic acts confer benefits to others, but net costs to the self. Different types of prosocial behavior are distinguished, depending on whether it is enacted by an individual or as part of a group effort, and whether it is first order (direct contributions) or second order (sanctioning) behavior. Six theoretical approaches are outlined (evolutionary, rationalist, structural, institutional, situational, and individual differences). They explain the evolution of prosocial behavior and its variation across groups, contexts, and situations. Avenues for future research are discussed.

Introduction

Prosocial behavior has intrigued scholars for centuries, including prominent academics in the early days of economics, political science, anthropology, and sociology. In the *Theory of Moral Sentiments*, Adam Smith (1759) asked "which principles in his nature interest him in the fortune of others?," answering that sympathy for the misery of others was the answer. Alexis de Tocqueville (1835) suggested the hypothesis that the possibility to have influence through local democracy motivated the community spirit of Americans. More generally, research on prosocial behavior attempts to explain (1) why and how prosocial behavior evolved in humans and (2) why and how the incidence and forms of prosocial behavior vary across individuals, groups, and situations. The present article presents some of the more recent attempts to answer these questions. Since this article focuses on *sociological* explanations, particular attention is paid to the level of the group or larger collectives as potential antecedents, consequences, or contexts for prosocial behavior and altruism. We will first address some definitional issues and subsequently elaborate on the wide variety of different types of prosocial behavior. This is followed by an overview of different theoretical approaches to prosocial behavior. The article concludes with a digression on open research questions in the sociology of prosocial behavior.

Definitions

Prosocial behavior is a broad class of behavior defined as involving costs for the self and resulting in benefits for others. Where prosocial behavior results in *net* benefits for both the beneficiary and the person committing the prosocial act, it is referred to as *mutualism*. Where prosocial behavior benefits others but confers *net* costs to the person committing it, prosocial behavior becomes *altruism* (Bowles and Gintis, 2011). A common distinction is made between pure altruism and impure altruism. Pure altruism is "acting with the goal of benefitting another" (Piliavin and Charng, 1990); prosocial behavior solely motivated by concern for others, also denoted as 'altruistic motivation.' Impure altruism is prosocial behavior motivated by more self-centered motives or even by narrow self-interest, potentially in addition to concern for others. Note that the definition of prosocial behavior is purely behavioral, while altruism combines behavioral and motivational elements.

Types of Prosocial Behavior

Prosocial behavior can take many different forms, depending on the context in which it takes place; it may benefit specific others, known or unknown and at varying degrees of social distance, ranging from close kin to total strangers; it may benefit social groups of which the actor is a member or not, as well as society as a whole or abstract ideals.

Much specialized research on prosocial behavior can be found, focusing on prosocial behavior in specific societal and institutional contexts. Since this research often uses different labels to denote the phenomenon, reviews of research findings face serious challenges, even within a single discipline. For example, within the field of the sociology of organizations, one specific form of formal prosocial behavior is organizational citizenship behavior, defined as prosocial extra-role behavior by employees of an organization benefiting colleagues or the employer (Podsakoff et al., 1997). Another aspect of prosociality recently studied in a specific organizational setting – public administration – is public service motivation, defined as "an individual's orientation to delivering services to people with the purpose of doing good for others and society" (Perry et al., 2010). Another challenge results from the fact that the analytical distinctions between the general theoretical constructs used to describe prosocial behavior are not always clear-cut, and seem to vary between disciplines; concepts like cooperation, solidarity, cohesion, collaboration, support, or help are often used as synonyms.

The following more general distinctions are particularly relevant for theory and empirical research. First, prosocial behavior can differ depending on whether it is enacted by an individual or as part of a group effort, and the effects it has on the beneficiary as well as on third parties. Four types of prosocial behavior can be distinguished based on these two dimensions. First, there are prosocial acts carried out by a single individual, benefitting a specific other individual. This ideal typical form is the implicit assumption of most analyses of prosocial behavior. Examples are passers by donating money to a beggar or doing the grocery shopping for your sick neighbor.

Second, there are prosocial acts carried out by a single individual, benefitting a collective or 'generalized other.' Examples are donations by an individual to a charitable organization, or a soldier holding a lost position to enable his squad to retract, but also less dramatic acts like paying one's taxes. Third, there are prosocial acts carried out by a collective in order to benefit a single individual. Examples are actions targeted to liberate political prisoners by organizations like Amnesty International. Finally, there are prosocial acts carried out by a collective or group, with the intention to help another collective. This kind of prosocial behavior can also be directed to benefit the collective itself. An example would be workers joining a wildcat strike at the risk of being fired (Zetka, 1992).

While benefitting a specific individual or group, prosocial behavior can also have positive or negative consequences for third parties. For example, forms of oppositional solidarity like strikes may benefit the participants, but harm the organization. Conversely, prosocial behavior favoring a specific individual or group may also benefit other groups, as in the case of strikes by a small group leading to pay raises for all employees.

A second important distinction differentiates between formal and informal prosocial behavior. Actions that involve an intermediary organization, like charitable giving, volunteer work for nonprofit organizations, or blood and organ donation are examples of formal prosocial behavior. Social and emotional support to friends and family members, caring for parents and children, or helping strangers are examples of informal prosocial behavior.

Finally, in settings of collective good production, first and second order prosocial behavior needs to be distinguished (Heckathorn, 1996). First order prosocial behavior consists of acts that directly benefit another individual or group. Second order prosocial behavior consists of acts sanctioning behavior that violates rules about prosocial behavior (Fehr and Gächter, 2002). In both cases, an individual can decide to cooperate or defect, but in the second order case, a third option is to oppose sanctioning of others. In situations of collective good production, the combination of first and second order decisions results in different types of prosocial behavior. Individuals who act prosocially by contributing to first-level collective good production can either be *full cooperators* (if they also sanction free riders who do not contribute to the collective good), *private cooperators* (if they refrain from sanctioning free riders), or become part of the *compliant opposition* by defending "the rights of others to refuse to contribute" (Heckathorn, 1996: 254). Individuals who do not contribute to the collective good at the first level can nevertheless decide to behave prosocially at the second level by sanctioning those who also did not contribute, thereby becoming *hypocritical cooperators*. They can also decide to actively oppose attempts of control by others (*full opposition*), or simply also not sanction free riders (*full defection*).

Theories of Prosocial Behavior

Several broad classes of theories explaining prosocial behavior can be distinguished. They advance different mechanisms to explain why prosocial behavior emerged in humans, and why it varies between individuals, groups, contexts, and situations.

Evolutionary approaches have demonstrated that prosocial behavior toward close genetically related family members could be the result of natural selection, because it may ultimately help to spread the benefactor's genes. Some scholars, therefore, suggest referring to prosocial behavior toward close relatives not as altruism, but as mutualism (Bowles and Gintis, 2011). A challenge for evolutionary approaches consists in the fact that prosocial behavior often extends far beyond the circle of close relatives. This observation is at the basis of an ongoing debate about human nature, i.e., the degree to and conditions under which human behavior is motivated by selfishness. Five evolutionary mechanisms are assumed to have made humans a cooperative species (Nowak, 2006): kin selection, direct reciprocity, indirect reciprocity, network reciprocity, and group selection. These mechanisms would also explain why we so frequently observe altruistic punishment; since it is an effective instrument safeguarding altruists from being abused, punishment of free riders became an objective in its own right. Along these lines, recent scholarship points to the hard wired character of *social preferences*, i.e., the fact that humans care for the well-being of others. Humans feel guilt and shame if they defect, feel obliged to help, and derive psychological benefits like pleasure and satisfaction from committing prosocial acts (Aknin et al., 2013; Konow and Earley, 2008), even if – or because – this comes at considerable costs. This holds not only for first order prosocial behavior (Jacquet et al., 2011), but also for its second order version; a series of influential experiments has shown that humans not only tend to develop negative emotions against free riders, but are actually willing to incur considerable costs to sanction them (Fehr and Gächter, 2002). Given this cumulative evidence, evolutionary approaches increasingly turn their attention to the question why altruistic preferences became so widespread and 'outcompeted' unrestricted and selfish gain maximization as the major motive guiding human behavior.

Rationalist approaches focus on the puzzle why and under which conditions selfish individuals engage in prosocial behavior. Classical rational choice theories prefer explanations that do not need to assume social preferences, but that stick to the core assumption of gain maximizing individuals. In this perspective, most prosocial acts fall under the category of 'mutualism,' i.e., they benefit both the sender and the target. This is usually the case if the sender and the receiver of the prosocial act are in some form dependent on each other for the acquisition or joint production of valued goods or services. Since the realization of mutual gains often takes place in a sequence of exchanges, the first player acting prosocially usually faces a trust problem, since the benefactor may decide not to reciprocate. We discuss two important ways to solve such trust problems: selective incentives and signaling.

Selective incentives are rewards and punishments designed to induce prosocial behavior. They can have a formal and an informal basis. For example, firms try to elicit prosocial behavior (i.e., extra-role organizational citizenship behavior) through selective incentives in the form of performance-contingent reward schemes, which are assumed to align the interests of the organization with the interest of the employee. But also in primordial social groups, like families, friendship cliques, or other kinds of sharing groups interdependence creates the need for informal norms (Lindenberg, 1997), because individuals have an interest to make sure that the other party acts prosocially toward themselves. Here, selective

incentives may take the form of remedial norms and emerge in a bottom-up process (Ellickson, 1991). Hence, rather than referring to norms as an external cause of prosocial behavior, rational choice theories prefer to 'endogenize' norms by pointing toward structural interdependencies as an underlying root cause (see also the section structural explanations below).

Signals are traits and behaviors that reliably predict someone's trustworthiness (Gambetta, 2009). Signaling is a key mechanism in situations of information asymmetry between two or more parties. Such asymmetries arise where both parties have at least partially conflicting interests, and where reliable and valid information about the quality (e.g., skills, performance) and/or the intentions (e.g., motivation) of one actor is unavailable or difficult to observe for another actor. *Relational signals* are cues about the signaler's intention to initiate or maintain a mutually rewarding relationship with the receiver (Lindenberg, 2003; Wittek, 2003). Individuals who are effective in successfully producing such relational signals will be more likely to attract prosocial acts from other players. The efficacy of a signal depends on a variety of factors (Connelly et al., 2011), in particular their *observability* (intensity, strength, clarity, visibility) and cost. Whereas some signals are easy to observe (e.g., a philanthropist's donation to a humanitarian fund) others are more difficult to decipher (e.g., evolutionary accounts of depression have argued that signs of depression reduce aggressiveness of exchange partners). The *cost* of a signal (e.g., the size of a donation in relation to a philanthropist's capital) has long been considered as a major condition for its efficacy, but more recent accounts deemphasize the importance of costs and suggest that less-costly signals may also be relevant (e.g., a prime minister's apology for atrocities of previous governments). Other aspects of signals are (Connelly et al., 2011: 52) their *fit* ("the extent to which the signal is correlated with the unobservable quality"), their *frequency* ("the number of times the same signal is transmitted"), and their *consistency* ("agreement between signals from one source").

Institutional approaches emphasize the importance of formal rules or informal norms and shared meaning as a root cause of prosocial behavior. In this framework, prosocial behavior does not require some ulterior payoff to the sender, but it can be the result of deeply ingrained cultural schemas, beliefs, values, or identities. Several mechanisms are often associated with institutional explanations of prosocial behavior. First, many scholars consider the *reciprocity* norm as a fundamental trait, which can be found in all human cultures (Mauss, 1954[1923]). The reciprocity norm requires that 'gifts'trigger the obligation to be repaid. This means that individuals receiving a gift from someone will be inclined to behave prosocially toward this person at a later stage (Gouldner, 1960). Reciprocity can take several forms, ranging from direct reciprocity (in which the beneficiary is the individual who provided the gift), to indirect or generalized reciprocity (where the beneficiary may be a third party). Second, institutional approaches consider *learning and socialization* as powerful triggers of prosocial behavior. Third, shared social identities may trigger feelings of in-group obligation and loyalty, which reduce the incentive to free ride (Ellemers et al., 2002).

Structural approaches: Two interrelated strands of research can illustrate this type of explanation. *Social dilemma* studies focus on the structure or 'anatomy' of the cooperation problem itself, pointing to *negative interdependence* and *scale* as two fundamental structural conditions that systematically affect the incidence of cooperation (Ostrom et al., 1999). The first condition of negative interdependence creates social dilemmas, such as, for instance, the prisoner's dilemma or the 'tragedy of the commons' (Hardin, 1968); everybody will be better off if everyone contributes, while it is individually rational to defect. Second, scale exacerbates free-rider problems, for the higher the *number of participants* (Olson, 1965), the smaller the impact of individual contributions will be on the collective good. Small-scale settings are often considered ideally suited to cooperation, because they guarantee frequent face-to-face contact and stimulate adherence to the social norm (Ellickson, 1991).

Social network researchers have studied how repeated interaction can stimulate individuals to act prosocially toward others. This will be more likely if the parties concerned do have a history of previous exchanges and the prospect of future interactions (Buskens and Raub, 2013). Repeated interaction allows for conditional cooperation and helps to build reputations. Conditional cooperation is considered the simplest solution in a situation of repeated interaction. Reputations about trustworthiness can be built within a dyad, but can also be gathered from third parties through gossiping (Sommerfeld et al., 2007; Ellwardt et al., 2012). Though embeddedness in dense social networks can trigger prosocial behavior (e.g., Flache, 2002), this effect is highly context dependent. For example, self-managed work teams characterized by a high level of trust (i.e., first order prosocial behavior) may score low on second order prosocial behavior, i.e., exhibit a tendency to avoid monitoring and sanctioning noncompliance (Langfred, 2004).

Situational approaches emphasize the power of cues as a trigger for prosocial behavior. Such cues can have their roots in (changes in) the physical or environment an individual is exposed to at a given moment. According to *goal framing theory*, even subtle cues can have strong effects because they activate different types of goal frames, which push either gain, normative, or hedonic motivations into the foreground. For example, a series of field experiments testing the 'broken windows hypothesis' showed that prosocial behavior declines in settings where individuals are exposed to visible signs of rule violations by others, like graffiti on a wall with a 'no graffiti' board (Keizer et al., 2008).

Individual difference approaches argue that independently of the specific institutional context or situation they are in, an individual will consistently cooperate or defect, depending on his or her individual predispositions or personality traits. For example, an experimental study showed that in prisoner's dilemma games individuals with a high internal locus of control, high self-monitoring, and high sensation seeking are more likely to act prosocially than individuals scoring low on these personality traits (Boone et al., 1999). Hence, though confronted with the same situational and institutional setting, there is nevertheless a lot of variation between individuals with regard to their inclination to act prosocially.

Future Directions

The past decade has seen a proliferation of research on prosocial behavior, and much progress has been made in

understanding its sociological implications. We see at least three promising domains for future research.

First, a question that so far received comparatively little attention in the different approaches to prosocial behavior concerns the conditions for *sustainable* cooperation. Institutional approaches still produced limited insights into the conditions needed to create stable sanctioning regimes, or how cooperation is sustained when formal and informal monitoring systems are imperfect. Social preference explanations do not specify under what conditions certain preferences become salient, and why they will remain so. Similarly, experimental studies of iterated N-person dilemmas found that cooperation decreases significantly over time (Ledyard, 1995), starting to wane almost from the start. A stronger focus on the sustainability of cooperative relations and prosocial behavior would also have implications for research designs and evidence-based policy making. For example, with the increasing availability of longitudinal network data and the related statistical models for the analysis of network dynamics (Snijders, 2013), more explicit theorizing on the interrelation between changing social contexts and prosocial behavior becomes possible. Similarly, cumulating evidence on the social basis of human rationality provide new challenges but also opportunities for policy makers facing the challenge to design institutions fostering sustainable cooperation in a variety of settings, like organ donation or tax evasion (Thaler and Sunstein, 2008).

Second, with the growing importance of the World Wide Web for almost all domains of social life, more insight is needed into its effects on prosocial behavior. Internet-based technologies have the potential to reduce coordination and monitoring costs to a minimum, thereby taking away some of the major obstacles that hampered the initiation and maintenance of (large-scale) cooperation. Indeed, growing global interconnectedness is thought to have increased the willingness to contribute to the production of public goods, strengthening as it does a cosmopolitan attitude, while weakening the influence of ethnicity, locality, and national identity (Buchan et al., 2009). Nevertheless, very little is known about such secular trends, and to what degree this technological change affected the emergence of new forms of prosocial behavior and joint production, indeed.

Finally, though the science of cooperation made tremendous progress, sociology can certainly benefit from a more active reception of advances in related fields, particularly those focusing on behavioral microfoundations, like social psychology, behavioral economics, and the cognitive neurosciences. But the same holds for the reception of sociological insights by other fields, particularly for those cases in which physiological or neural processes are presented as ultimate 'causes' of prosocial behavior that are insensitive to changes in social contexts. A nice example is the proliferation of studies on the role of the 'cuddle hormone' or 'love drug' oxytocin as an indiscriminate trigger of prosocial motivations in humans. Instead, a careful recent study, which took the social context serious, found that "oxytocin creates intergroup bias because it motivates in-group favoritism and, to a lesser extent, out-group derogation" (De Dreu et al., 2011: 1). As this and other examples show, there is still much to be gained from more dedicated interdisciplinary work on the sociology of prosocial behavior.

See also: Charity and Philanthropy: Overview; Control, Social; Cooperation and Competition; Cooperation: Sociological Aspects; Evolutionary Sociology; Gift: History of the Concept; Groups, Sociology of; Human Cooperation, Evolution of; Institutions; Interests in Sociological Analysis; Norms; Prosocial Behavior During Adolescence; Prosocial Behavior During Childhood and Cultural Variations; Prosocial Behavior During Infancy and Early Childhood: Developmental Patterns and Cultural Variations; Prosocial Behavior and Empathy; Prosocial Behavior, Effects of Parenting and Family Structure on; Social Dilemmas, Psychology of; Solidarity, Sociology of; Trust, Sociology of; Voluntary Associations, Sociology of.

Bibliography

Aknin, L.B., Dunn, E.W., Whillans, A.V., Grant, A.M., Norton, M.I., 2013. Making a difference matters: impact unlocks the emotional benefits of prosocial spending. Journal of Economic Behavior and Organization 88, 90–95.

Boone, C., De Brabander, B., Van Witteloostuijn, A., 1999. The impact of personality on behavior in five prisoner's dilemma games. Journal of Economic Psychology 20 (3), 343–377.

Bowles, S., Gintis, H., 2011. A Cooperative Species: Human Reciprocity and Its Evolution. Princeton University Press.

Buchan, N., Grimalda, G., Wilson, R., Brewer, M., Fatas, E., Foddy, M., 2009. Globalization and human cooperation. Proceedings of the National Academy of Sciences of the USA. 106 (52), 4138–4142.

Buskens, V., Raub, W., 2013. Rational choice research on social dilemmas: embeddedness effects on trust. In: Wittek, R., Snijders, T.A.B., Nee, Victor (Eds.), The Handbook of Rational Choice Social Research. Stanford University Press, Palo Alto, pp. 113–150.

Connelly, B., Certo, S., Ireland, R., Reutzel, C., 2011. Signaling theory: a review and assessment. Journal of Management 37 (1), 39–76.

De Dreu, C.K., Greer, L.L., Van Kleef, G.A., Shalvi, S., Handgraaf, M.J., 2011. Oxytocin promotes human ethnocentrism. Proceedings of the National Academy of Sciences 108 (4), 1262–1266.

Ellemers, N., Spears, R., Doosje, B., 2002. Self and social identity. Annual Review of Psychology 53, 161–186.

Ellickson, R.C., 1991. Order without Law: How Neighbors Settle Disputes. Harvard University Press, Cambridge, MA.

Ellwardt, L., Labianca, G.J., Wittek, R., 2012. Who are the objects of positive and negative gossip at work? A social network perspective on workplace gossip. Social Networks 34 (2), 193–205.

Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415 (6868), 137–140.

Flache, A., 2002. The rational weakness of strong ties. Failure of group solidarity in a highly cohesive group of rational agents. Journal of Mathematical Sociology 26, 189–216.

Gambetta, D., 2009. Signaling. In: Hedström, P., Bearman, P. (Eds.), The Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford, pp. 168–194.

Gouldner, A.W., 1960. The norm of reciprocity: a preliminary statement. American Sociological Review, 161–178.

Hardin, G., 1968. The tragedy of the commons. Science 162 (3859), 1243–1248.

Heckathorn, D.D., 1996. The dynamics and dilemmas of collective action. American Sociological Review, 250–277.

Jacquet, J., Hauert, C., Traulsen, A., Milinski, M., 2011. Shame and honour drive cooperation. Biology Letters 7 (6), 899–901.

Keizer, K., Lindenberg, S., Steg, L., 2008. The spreading of disorder. Science 322 (5908), 1681–1685.

Konow, J., Earley, J., 2008. The hedonistic paradox: is homo economicus happier? Journal of Public Economics 92, 1–33.

Langfred, C.W., 2004. Too much of a good thing? Negative effects of high trust and individual autonomy in self-managing teams. Academy of Management Journal 47 (3), 385–399.

Ledyard, J., 1995. Public goods: a survey of experimental re- search. In: Kagel, J.H., Roth, A.E. (Eds.), The Handbook of Experimental Economics. Princeton University Press, Princeton, NJ, pp. 111–194.

Lindenberg, S., 1997. Grounding groups in theory: functional, cognitive, and structural interdependencies. Advances in Group Processes 14, 281–331.

Lindenberg, S., 2003. Governance seen from a framing point of view: The employment relationship and relational signaling. In: Nooteboom, B., Six, F.E. (Eds.), The Trust Process in Organizations, Empirical Studies of the Determinants and the Process of Trust Development. Elgar, Cheltenham and Northampton, pp. 37–57.

Mauss, M., 1954[1923]. The Gift: Forms and Functions of Exchange in Archaic Societies (No. 378). WW Norton & Company.

Nowak, M.A., 2006. Five rules for the evolution of cooperation. Science 314 (5805), 1560–1563.

Olson, M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.

Ostrom, E., Burger, J., Field, C.B., Norgaard, R.B., Policansky, D., 1999. Revisiting the commons: local lessons, global challenges. Science 284 (5412), 278–282.

Perry, J.L., Hondeghem, A., Wise, L.R., 2010. Revisiting the motivational bases of public service: twenty years of research and an agenda for the future. Public Administration Review 70, 681–690.

Piliavin, J.A., Charng, H.W., 1990. Altruism: a review of recent theory and research. Annual Review of Sociology, 27–65.

Podsakoff, P.M., Ahearne, M., MacKenzie, S.B., 1997. Organizational citizenship behavior and the quantity and quality of work group performance. Journal of Applied Psychology 82 (2), 262.

Rudd, M., Aaker, J., Norton, M.I., 2014. Getting the most out of giving: concretely framing a prosocial goal maximizes happiness. Journal of Experimental Social Psychology 54, 11–24. http://dx.doi.org/10.1016/j.jesp.2014.04.002.

Smith, A., 2002/1759. In: Hanley, R. (Ed.), The Theory of Moral Sentiments. Penguin, New York.

Snijders, T.A.B., 2013. Network dynamics. In: Wittek, R., Snijders, T.A.B., Nee, Victor (Eds.), The Handbook of Rational Choice Social Research. Stanford University Press, Palo Alto, pp. 252–279.

Sommerfeld, R.D., Krambeck, H.J., Semmann, D., Milinski, M., 2007. Gossip as an alternative for direct observation in games of indirect reciprocity. Proceedings of the National Academy of Sciences 104 (44), 17435–17440.

Thaler, R.H., Sunstein, C.R., 2008. Nudge: Improving Decisions About Health, Wealth, and Happiness. Yale UP, New Haven.

Tocqueville de, Alexis, 2000/1835. In: Mansfield, H., Winthrop, trans, D. (Eds.), Democracy in America. University of Chicago Press, Chicago.

Wittek, R.P.M., 2003. Norm violations and informal control in organizations: a relational signaling perspective. In: Six, F., Nooteboom, B. (Eds.), The Trust Process in Organizations: Empirical Studies of the Determinants and the Process of Trust Development. Elgar, pp. 168–195.

Zetka Jr., J.R., 1992. Work organization and wildcat strikes in the US automobile industry, 1946 to 1963. American Sociological Review, 214–226.

Alzheimer's Disease: Behavioral and Social Aspects

Dieter Ferring, Research Unit INSIDE, University of Luxembourg, Walferdange, Luxembourg

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E. Miller, volume 1, pp. 423–425, © 2001, Elsevier Ltd.

Abstract

This article gives an overview of the behavioral and social aspects associated with Alzheimer's disease (AD). It first offers an outline of the phenomenology and diagnosis of the disease, followed by systematizations, and guidelines to be adopted in the study of psychosocial aspects of AD. These comprise a socioecological viewpoint and a life span developmental perspective underlining the significance of adaptive efforts of both the concerned individual as well as of the family and further significant others. The psychosocial consequences of AD are then described for different stages of the disease elaborating different topics and tasks for individual adaptation.

Introduction

Within the category of neurodegenerative diseases, Alzheimer's disease (AD) takes a prominent role given that an estimated 24 million people worldwide have dementia, the majority of them assumed to have AD (Ballard et al., 2011). Ferri et al. (2005) report that the number of people living with dementia will almost double every 20 years, from worldwide 24.3 million to 42.3 million in 2020 and to 81.1 million in 2040. Based on their Delphi consensus study, the authors estimate that the total number of 4.9 million people suffering from AD in Europe in 2009 will increase to 6.9 million in 2020 and to 9.9 million in 2040.

It comes as no surprise, therefore, that AD has become a 'popular' concept in such a way that the term has already entered everyday language with people using it to explain a wide variety of memory problems or other kinds of non-successful cognitive performance. AD poses a threat to the individual and the society and even though research on this disease has been promoted during the last three decades and several risk factors have been identified, the specific causes of AD and, hence, reliable prevention and treatment remain elusive. A significant aspect of the threat that AD poses to both the individual and the society thus contains uncertainty: neither the individual knows if he or she will be affected nor does society know which impact the disease will have on public expenditures, distribution of resources, and 'resource fairness' between generations.

The reception of AD by the life and social sciences as well as other science disciplines is illustrated by the number of dedicated journals that have appeared during the last decades (including 'Alzheimer's and Dementia – The Journal of the Alzheimer's Association'; 'Alzheimer Disease and Associated Disorders'; 'Alzheimer's Research and Therapy'; 'American Journal of Alzheimer's Disease and Other Dementias'; 'International Journal of Alzheimer's Disease'; 'Journal of Alzheimer's Disease'). In addition to this growing number of journals, one finds specific publications concerning the various aspects associated with AD in high-impact journals of geriatric medicine.

In order to develop a first idea of the behavioral and social aspects that describe the public reception of AD, one may look at the 'World Alzheimer Reports' published since 2009 by Alzheimer Disease International (ADI). ADI represents the international federation of Alzheimer associations around the world combining 80 national associations. Since 2009, this organization has published five reports focusing on different topics associated with the disease. Specifically, the published reports have dealt with the phenomenology of the disease (2009), its economic impact (2010), the possibilities and options of an early diagnosis (2011), the notion of stigma associated with the disease (2012), and long-term care (2013). (All reports can be downloaded at: http://www.alz.co.uk/research/world-report.) Following this, to talk about AD signifies at first an understanding of the phenomena defining the disease. The second notion describes the imminent costs associated with the disease that amounted to US$ 604 billion in 2010; of these 70% appeared in Western Europe and North America (Wimo et al., 2010). Stigma and stereotypes around AD represent an important social aspect of the disease implying different consequences for the affected person and his/her family.

I will address some of these issues in the following starting with a short description of the phenomenology and diagnosis of AD including the options for an early diagnosis since this has consequences for psychosocial effects that the disease will have on all concerned persons. In the second part, I will introduce guiding principles and perspectives that will be helpful to systematize the study of the psychosocial aspects of the disease, and the third part, will focus on the psychosocial aspects by reviewing the effects that the disease has at the personal and social family level.

AD – Phenomenology and Diagnosis

"*Ich habe mich sozusagen selbst verloren* – I lost myself" is an often-cited sentence by Auguste D. – the first person diagnosed with the disease that the world came to know by the name of the psychiatrist and neuropathologist Alois Alzheimer who described this disorder in 1906. The quotation basically describes the loss of self-consciousness and a state of confusion that defines one aspect associated with AD. In 1901, Auguste D. was brought by her husband to the *Institution for the Mentally Ill and for Epileptics* in Frankfurt (Germany) after showing symptoms that are nowadays considered typical for AD. She showed memory loss, delusions, and sleep problems; she also screamed for hours in the middle of the night. Tests performed by

Dr Alzheimer showed further cognitive deficits. (Retrieved at the site of the German Alzheimer Association, May 2014: http://www.deutsche-alzheimer.de/unser-service/archiv-alzheimer-info/auguste-d.html#c767.)

More than 100 years later, the term AD stands for a brain disorder usually starting in late-middle age or old age and including several symptoms such as memory loss, confusion, reduced problem solving and thinking, and changes in language, behavior, and personality. According to the *Diagnostic and Statistical Manual of Mental Disorders Fourth Edition*, AD is defined by multiple cognitive deficits causing significant changes in social or organizational functioning characterized by a gradual onset and continuing cognitive decline. Multiple cognitive deficits comprise memory impairment and at least one or more additional cognitive disturbances (i.e., aphasia, apraxia, agnosia, disturbances in executive functioning). The *International Classification of Diseases-Version 10* (ICD-10) differentiates four forms of 'dementia in Alzheimer's disease' (F00): Dementia in AD with early onset (F00.0), with late onset (F0.01), atypical or mixed type dementia in AD (F00.2), and an unspecified form of the disease (F00.9).

AD is first described as a neurodegenerative disease that affects the central nervous system by continuously destroying neurons and neuronal activity. In particular, the neurodegenerative process involves nerve cells in the brain to stop working by interrupting information transfer between synapses, to lose connections with other nerve cells, and to finally die. Specific for this situation is the presence of amyloid plaques consisting of beta amyloid – a toxic protein peptide – and of neurofibrillary tangles (Brion, 1998). Biochemical processes underlying the production of these proteins have led to the discussion of a genetic component of AD. Here, research has focused on mutations of selected genes on the chromosomes 21, 14, and 1 (Castellano et al., 2011). Moreover, mutations found within the two genes – presenilin-1 and presenilin-2 – may cause early onset AD also by harmfully influencing the processing of beta-amyloid (Haass, 1996).

In 2011, the National Institute on Aging and the Alzheimer's Association elaborated new criteria for the diagnosis of AD. These new criteria differentiate three stages, namely, preclinical AD, mild cognitive impairment (MCI) due to AD, and dementia due to AD (see Jack et al., 2011). In contrast to the categories defined in the ICD-10 comprising mild/early stage, moderate/midstage, or severe/late stage, these recommendations state that AD begins before the mild/early stage implying that specific disease-related changes can be diagnosed even before this stage. Thus, this stage is considered as preclinical, followed by MCI due to AD. The third stage comprises all three stages from mild/early stage to severe/late stage. Mean survival time after diagnosis is estimated between 4 and 8 years in people over 65 years; there is also evidence of people surviving up to 20 years after the diagnosis (see Helzner et al., 2008).

Psychosocial Effects of AD: Adaptation in a Socioecological Context

In order to describe the psychosocial effects that a neurodegenerative disease like AD has on the individual and his or her social environment, one may start with the observation that there are multiple effects on several life domains for all persons involved. A first systemization of these effects starts by distinguishing descriptive levels, and here one may use the socioecological model proposed by Bronfenbrenner (1979) as a heuristic tool. He elaborated this model as description of the developmental contexts in childhood and adolescence but it may also be used to describe the micro- and the macrosystemic interactions that constitute the living context of persons affected by AD.

Bronfenbrenner (1979) places the person at the center of microsystems, which are defined by him as "a pattern of activities, roles, and interpersonal relations experienced by the developing person in a given setting with particular physical and material characteristics" (p. 22). Microsystems comprise the interactions in the family setting and the closer social networks of a person and thus exert a direct influence on the individual. They constitute greater social entities called mesosystems, which are described by the interaction between several microsettings such as family, school, and work. In addition, exosystems constitute settings in which the person does not directly participate but which, nevertheless, have an indirect influence on the individual; such systems comprise, for instance, the place of employment of one's partner and family members, educational institutions of one's children, or the residential background. Changes in these settings will affect a person as well. The macrosystem finally comprises the social, political, and cultural background of a society at a given historical time, and this sets the frame for further influences on the life situation of a person affected by AD.

From this point of view, AD impacts several individual domains of functioning: the affective-motivational, the behavioral, and the cognitive domain. A second differentiation highlights the impact on the social interactions of a person affected with AD at the microlevel, i.e., in his or her partnership, family, and further social networks. This is then followed by the impact that changes in these specific systems will have on the meso- and exolevel. An informal carer, such as a partner or child, may no longer be able to work full time (or at all), they will not be able to participate in former social activities, and finally, informal care has an impact on the whole family system and its relation to settings at the meso- and exolevel, which may then lead to repercussions on the individual and family. The employer may question the efficiency of a family carer, and social partners and networks may minimize the contact. Finally, AD challenges the macrolevel and thus society as well, which is reflected by an ongoing discourse at various public levels. Prominent among these, one finds the sociopolitical debate of meeting the challenges that dementia and neurodegenerative diseases place on society, especially with respect to solutions assuring the financing of individual care in case of dependency. In fact, several European countries have come up with national strategies to deal with AD, and various practices and policies are implemented across Europe with respect to this question. At present, the existence of national Alzheimer's associations in more than 80 countries worldwide underlines the urgency of public debate and the necessity of representing the interests and needs of Alzheimer patients and their families.

A second systematization that is offered here derives from life span developmental theory and captures the notion of

development as a bio-sociocultural coconstruction comprising the influences of biogenetics, culture, and personality dynamics on human behavior. In order to conceptualize human behavior in the context of AD, one has to take into account the complex pattern of epigenetic interactions between personality dynamics and biogenetics in a given cultural context (see Zittoun et al., 2013). In such a view, human behavior is influenced and formed by normative and nonnormative events and tasks that require adaptation across the life span (Boll and Ferring, 2013). Such a perspective includes a focus on the timing and chronology of events challenging the individual. If one takes the diagnosis of AD as a marker event or significant indicator of life change, one can thus differentiate between antecedent, process, and effect variables that describe the process of adaptation to the changed life situation. Persons affected by AD differ in their biographies and, thus, show differing profiles of personal and contextual characteristics prior to the diagnosis. These will shape the effects that AD will have on the person and his or her network and the subsequent process of adaptation.

A third point concerns the concepts of adaptation and coping of the person affected with AD as well as of the family system to the changed life situation. (Here we use the term 'coping' synonymous to the term of adaptation since both capture the meaning of coming to terms with a given life situation involving various criteria of adaptation.) It is a well-known phenomenon in gerontology that individuals facing comparable demands (e.g., multimorbidity, handicaps) may differ greatly in subjective indicators of adaptation (e.g., self-esteem, subjective well-being). The paradox of well-being describes the phenomenon that a 'bad' life situation is judged as good; this is taken as an indicator of the effective use of adaptive processes in old age allowing for a regulation of the self-perception and subjective well-being (see Ferring and Boll, 2010). When considering the effects that AD may have on the individual and family life, one should be aware that comparable levels of demands, impairments, and strain might not result in comparable outcomes. These results will depend on the already mentioned contextual and personal resources that a person and family may rely on, but also on the adaptive processes underlying the regulation of the self and subjective well-being.

Several models specify such processes at the individual level assuming, for instance, that the elderly person may actively select goals, optimize existing capacities, and compensate losses (e.g., Baltes, 1997). The dual process model proposed by Brandtstädter and coworkers (e.g., Brandtstädter and Greve, 1994) supposes – in line with Piaget – that persons in general assimilate experiences into their cognitive schemata thus confirming their view of the self and the world; if this is no longer possible they immunize against self-threatening experiences (e.g., by selective comparisons) or accommodate their cognitive schemata. It goes without saying that all these efforts require a certain level of awareness and cognitive functioning that will be progressively reduced in the context of AD. The conditions for applying these models to the person affected by AD may no longer be given in the course of disease progression, thus challenging at the individual level the conceptual significance of the term adaptation itself.

As a last point, the process of adaptation is, in itself, not only to be conceptualized as an individual adaptive effort but rather as a systemic and historical process in which a concerned person's behaviors (i.e., the diagnosed person as well as the members of his or her social environment) have an impact on one another. One may describe this interplay of adaptive efforts as a type of 'polyadic coping' thus underlining the reciprocity of behavior in social systems.

Psychosocial Effects at the Individual and Family Level

In the following, from a life span developmental perspective, I conceptualize the diagnosis of a life threatening disease as a nonnormative event that confronts the person and his or her family with a series of new tasks that will require adaptation in several domains. A distinction will be made here between the early phases of disease onset and the time course of disease progression and consider phenomena at the individual as well as the family level thereby including involved significant others in the sequence of events.

Onset of AD – A Stage of Insecurity and Emerging Losses

AD is often diagnosed in the mild dementia stage since memory functioning and daily activities are apparently and observably impaired. The diagnosis of AD is considered here to mark the onset of the disease although the indicators of a weakened performance in the cognitive and functional domain certainly are registered, before this 'official' marker, during the pre-dementia stage. The concerned person, the partner, as well as friends and family members may register behavioral changes mostly indicated by weak memory performance and mood- and motivation-related phenomena linked to depression (see Berger et al., 1999). Comparable to other diseases and as has been argued elsewhere (Ferring and Filipp, 2000), one may characterize this stage as 'weakly scripted,' which signifies – in the terminology of cognitive psychology – that the concerned persons have no scripts and, related to this, are lacking the specific behavioral routines to deal with the situation at hand. Moreover, whereas other negative life experiences (e.g., loss of a loved one) are embedded into culturally shaped ways of responding and routines that often facilitate the coping process, the initial diagnosis of AD clearly lacks this external facilitation – there is no 'ritual' or 'normed' response to this.

The time leading to the final diagnosis of AD is often experienced as even more challenging than the diagnosis of other chronic and life threatening diseases. This is due to the already observed and expected damage to cognitive functioning that, in most cases, has led to the diagnosis and that will deprive the concerned person of their capacity to act in a self-responsible way in the long run (see Bourkel et al., 2012). The capacity to make decisions in a self-responsible way comprises the knowledge about the consequences of one's actions as well as the intention to willingly achieve these actions. Self-responsibility in the context of health behavior comprises the responsibility for health self-management even if the causes for the disease may be beyond one's control. With the progression of the disease, a person with AD will become dependent from significant others that will take over in making 'advocatory' decisions concerning his or her life

(see also 'models of helping and coping' according to Brickman et al., 1982).

Sabat (2002) first raised the question of the convergence of awareness of Alzheimer-specific deficits between AD patients and AD researchers and diagnosticians. Moreover, he questioned what this signifies for the epistemological issues involved in social situations where deficits due to AD are assessed. Specifically, he elaborates that by attributing a lack of awareness about one's condition to the person with AD, one may fail to understand "that what the person with AD does or does not do (such as admit to or discuss fully his or her deficits) is itself affected powerfully not only by the organic effects of the disease, but also by (a) the person's reaction to the disease's effects; (b) the behavior and reactions of others; and (c) the reactions of the person with AD to the behavior of others in his or her social world" (p. 281). In this statement, which transcends beyond the organic effects, the author links the persons affected by AD to their social partners and the social situation.

Using this reasoning one may construct a scenario, occurring especially in the early phases of the disease, where both the affected person and his or her social context may be permanently challenged by the question of whether a specific behavior is already indicating a decline that cannot be hindered. Awareness of disease-related changes plays a crucial role here, and high levels of awareness have been related to dysthymia and anxiety of the person with AD (Aalten et al., 2005). Anosognosia is a concept that describes discrepancies in disease awareness between the persons with AD and their social partners (Maki et al., 2012). Recently, van Vliet et al. (2013) reported that intact awareness of the disease was associated with depressive symptoms both in young-onset and late-onset AD. de Boer et al. (2010) conducted interviews with early-stage AD patients, and they reported that "the experiences of the participants appeared less negative than the ideas 'healthy' elderly often have about a life with dementia," thus indicating a more positive evaluation of the disease-related changes.

Within the context of individual adaptation, the potential discrepancy between external views and the views of the AD patients themselves provides us with an additional domain for study. Most models of cognitive adaptation start with the notion that a person perceives an external event, evaluates this event with respect to its significance for individual needs and – depending on this evaluation – initiates coping behavior that may serve problem- and/or emotion-focused goals (e.g., Lazarus and Folkman, 1984). Such a view of goal-directedness presumes a certain 'rationality' underlying individual behavior since the specific behavior will serve a specific need. This assumption is, however, clearly invalid in the situation of a person with AD. Though there is evidence that AD patients may be aware of their situation and it is worsening (van Vliet et al., 2013), it is well established that the capacity for information processing and the regulation of individual attention will be influenced by neurodegenerative processes.

Assuming that attention depends largely on our current state of goals, motives, and needs, a logical question concerns the motives that guide the perceptive structuring of one's environment, as brain function deterioration progresses. Moreover, perception and evaluation both rely on intact memory given that top-down processing plays a significant part in self-regulation allowing us to recognize and predict events and persons. Starting with the reduced capacity of working memory and the progressive impairment in the functioning of semantic (e.g., agnosia, aphasia), procedural (e.g., aphasia), and episodic memory (e.g., inability to recognize formerly well-known persons), it becomes evident that persons with AD are not able to process information in the instrumental and/or rational way necessary for serving individual adaptation. AD thus not only destroys individual information processing and executive functions but also challenges theoretical models of adaptation and the underlying concept of human functioning.

Disease Progression and the Necessity of Care

Caregiving, with all its diverse aspects, represents a central task for the patient's partner and family when the disease progresses. A first and obvious differentiation concerns the role of formal and informal care. Specific cost calculations show, for instance, that there will be considerable challenges to national households due to AD care and that informal care will be necessary to account for this (Wimo et al., 2010). With respect to formal assistance, one may discern between professional home care services, day care units, and nursing homes for meeting the growing care needs of AD patients. Depending on the solutions found in specific national contexts, families will have to decide how they can afford the costs associated with this and – as this is paralleled on the larger scale in political household decisions – responsible persons have to decide where and for whom the family money will be invested. For example, will there be enough money to finance the further education of one's children, and how can this be arranged with the needs of a person who will be in progressive need of care? Here, the notions of 'resource fairness' and intergenerational justice are advanced to describe the underlying question of how to distribute, in a just way, the means available to the family, which certainly will be a distressing task if these means are limited (Ferring, 2010).

Considering the *tasks involved in informal caregiving*, one may start with the notion that the specific literature highlights primarily the stress and strain associated with such tasks for the caregiver and his/her family while the potential gains in such a situation are rarely discussed, if at all (Boll and Ferring, 2013). The caregiver situation and the associated caregiver career certainly represent various demands and challenges where the concerned person – in most cases – cannot rely on specific knowledge or behavioral routines and coping expertise. This concerns not only the person who is identified as the caregiver by the family system and who is, in most cases, female (Di Rosa et al., 2011), but also all other microsystems within this context, including partners, children, siblings, and further kin as well as friends and neighbors. Classification of strain experienced in caregiving starts with the differentiation between primary and secondary stressors (see Boll and Ferring, 2013). The first ones comprise the tasks associated with the concrete tasks of delivering care (see Wilkins et al., 2009). These include *assistance in daily living routines* (e.g., personal hygiene, toilet hygiene, eating, dressing), and in *medical treatment* (e.g., medication, change of dressings and bandages). Further domains comprise *household activities* (e.g., house work, shopping), and dealing with *financial and administrative issues* (e.g.,

health insurance). For the prevention of risky behavior and situations, Alzheimer care also includes the supervision of the AD patient, for example, to prevent the person from running away or to deal effectively with irritable, aggressive behavior. A considerable part of the strain associated with care may also comprise the emotional support to be provided in cases of multiple suffering (Monin and Schulz, 2010). The patient may, for example, experience physical pain, emotional strain, episodes of anxiety and depression, as well as spiritual crises. As these examples illustrate, family carers may face extremely demanding situations, which make adaptability very important.

Primary stressors affect several life domains (e.g., job, leisure time) of carers and thus result in secondary stressors. Dividing one's resources between the person to be cared for and other family members, having to deal with reduced work time or quitting one's job, facing financial shortcomings, and making ends meet are examples of these secondary stressors (Pearlin and Aneshensel, 1994). Primary and secondary stressors add up to the total strain of caregivers. The subjectively experienced strain is, however, not a uniform phenomenon across carers because several factors moderate the impact of these. Research has shown that there is a multitude of tasks associated with caregiving and that there are several moderating factors helping to explain and predict the differing ways persons come to terms with these tasks (for an overview see Pinquart and Sörensen, 2007, 2011).

Moreover, 'coming to terms' can be measured with respect to different criteria of adaptation. One may, for instance, consider indicators of subjective well-being (both cognitive and affective; see Ferring and Boll, 2010), or rely on more objective criteria such as indicators of physical strain (e.g., physical symptoms, numbers of sick days, medication, and life expectancy of caregivers), financial means, or social participation of caregivers. To add even more to the complexity of these findings, coping and adaptation do not take place in a social vacuum but always involve social partners directly or indirectly (i.e., via others) by the different ways a person may cope with caregiving. Specific coping behavior of family members thus always has an impact on the other members within this system (i.e., 'polyadic' coping, see above). In addition, the process of adaptation has to be projected on a time line that covers the growing number of experiences shared by the caregiver and their social networks including the changing tasks associated with disease progression.

This complexity can be systematized by identifying a universal task of individual coping efforts. It is assumed that the *construction of meaning* underlying the diverse experiences that the caregiver and their network undergo constitutes this task. The use of the term 'construction' conveys that this is a process of changing mental representations of the world and the self via accommodative efforts. Meaning construction is thus considered a cognitive effort of self-regulation on the side of the partner and significant others when confronted with a person with AD. Several theoretical ways of conceptualizing meaning in the context of adaptation exist (see Hicks and Routledge, 2013), and here I would like to highlight conceptualizing meaning under a teleological perspective, which involves the construction and identification of goals associated with the experience of care. Identifying goals implies that the specific experiences may serve a function and are thus not meaningless; following such a reasoning, goals may also be considered as gains originating from the experience of caring (Emmons et al., 1998). The assumption builds on theoretical models elaborating that there is a genuine need for the experience of personal control in order to avoid feelings of help and hopelessness (see Ferring and Filipp, 2000). Meta-analytical findings underline this, for instance, by demonstrating that self-efficacy – as opposed to helplessness – differs significantly between caregivers and non-caregivers and that these differences are especially pronounced between dementia caregivers and non-caregivers (Pinquart and Sörensen, 2003).

The Last Stage of the Disease – Accepting Losses

Making meaning of the changed life situation also applies to the last stage of the disease for the family and significant others dealing with the person suffering from AD. This phase marks the severe late stage, which begins 4–10 years after diagnosis, although longer delays have also been reported. During this final stage, AD patients need daily assistance with personal care and experience a decline in physical activities (e.g., unable to sit, rigid muscles, and abnormal reflexes) as well as complete memory loss. Moreover, persons can no longer communicate coherently, are no longer responsive to social interaction, and thus no features of the former personality seem to remain. Patients may reach a vegetative state due to advanced neurodegenerative processes. Death may come from pneumonia, for instance, caused by impaired swallowing that allows food or beverages to enter the lungs and start an infection. Other common causes of death include complications from urinary tract infections and falls. (Retrieved, May 2014, at: http://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers-stages/art-20048448.)

An adult daughter described the apparent loss of personality of her mother and the inability to communicate in end-stage AD as follows: "I don't have a mother anymore. My mother as I knew her has gone – I became aware of this when l looked into her eyes and got no form of any response. There is nobody in there anymore." This statement marks the permanent experience of loss that family members have to cope with when the person with AD reaches a vegetative state. Coping with loss is the central topic of this last stage. These losses are irreversible since the cognitive capacity cannot be reestablished nor can brain degeneration be reversed. The awareness that the disease has progressed and that the person will soon die develops as the changes that indicate the terminal stage are registered. Horowitz (1982) has proposed that coping with traumatic events may be described as a sequence of stages starting with an initial outcry, followed by phases of intrusion and avoidance that may end in the final phase of completion. This process is, in itself, not a linear one but shows oscillations between the stages, and persons may move forward and backward between the specific stages. This model is used here as illustration of individual adaptive efforts that concern the family and other significant others who know the person – now in a state of severe dementia – from the

former shared life. Even at this stage, the interaction with the person with AD will still trigger mental associations and habitual reactions based on a common history on the side of the partner or the children. Nevertheless, the appearance and behavior of the person in a stage of progressed AD can no longer be considered as indicating participation in reciprocal interactions. If one imagines how people rely on the responsiveness of their interaction partners on the one side as well as the routinized habitual behavior guiding social interactions on the other, it becomes evident that the loss of response of the AD patient may represent a strong challenge to individual adaptation.

Moreover, other losses are at hand. Severe dementia may lead informal caregivers to rely more strongly on professional care and one may lose the option of caring. The 'patient' may be admitted to inpatient care and one will lose further opportunities to be close to them. The process of dying, particularly when accompanied by palliative treatment, may lead to further insecurities. The AD patient may experience severe pain and may no longer be able to communicate, which may further increase the family's feelings of insecurity. Finally, death of the person will place new demands on significant others, e.g., mourning and sharing the sentiment of mourning while at the same time trying to achieve a level of daily functioning. Coming to terms with loss means, in our understanding, to integrate these experiences in one's view of the self and the world. The person who is capable of constructing meaning out of the specific experiences in such a way of identifying gains and growth will certainly achieve this goal.

Conclusion

Psychosocial aspects of AD and especially its associated coping demands are comparable to those of other chronic and life threatening diseases; yet there are also aspects, which are disease specific. Decline in cognitive functioning, dramatic changes in motivation and emotion, and bodily decline up to a vegetative state confront the partner, family members, and others with questions for which no routines are available or where available routines may not function. Adapting to these changes and demands is a crucial if not the most crucial question for families and social networks of persons with AD. This statement brings us back to the fundamental dynamics underlying human development. The study of the psychosocial effects of AD is thus a study of the adaptive process with its manifold and multifaceted tasks resulting when a family member is diagnosed with AD. Depending on the outcomes of the adaptive process, differential effects may be expected with respect to the quality of life of persons affected with the disease and their family members. A final word should be mentioned here about the impact that AD will have within the context of an aging society. In addition to the challenge of sustaining public expenditures in various domains, one may also expect a discourse on how to distribute social resources in a 'just' way. People with AD will need advocates who will represent their rights for care in an ever-increasingly competitive society.

See also: Bereavement; Caregiver Burden; Chronic Illness, Psychosocial Coping with; Retirement and Health; Social Epidemiology.

Bibliography

Aalten, P., van Valen, E., Clare, L., Kenny, G., Verhey, F., 2005. Awareness in dementia: a review of clinical correlates. Aging & Mental Health 9 (5), 414–422.

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders, fourth ed. Author, Washington, DC.

Ballard, C., Gauthier, S., Corbett, A., Brayne, C., Aarsland, A., Jones, E., 2011. Alzheimer's disease. The Lancet 377 (9770), 1019–1031.

Baltes, P.B., 1997. On the incomplete architecture of human ontogeny. Selection, optimization, and compensation of developmental psychology. American Psychologist 52 (4), 366–380.

Berger, A.K., Fratiglioni, L., Forsell, Y., Winblad, B., Bäckman, L., 1999. The occurrence of depressive symptoms in the preclinical phase of AD: a population-based study. Neurology 53 (9), 1998–2002.

Boll, T., Ferring, D., 2013. Family caregiving for older people from a life-span developmental point of view. In: Albert, I., Ferring, D. (Eds.), Intergenerational Relations. European Perspectives on Family and Society. Policy Press, Bristol, UK, pp. 223–240.

Bourkel, E., Ferring, D., Weber, G., 2012. Perceived rights of and social distance to people with Alzheimer's disease. GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry 25 (1), 25–32.

Brandtstädter, J., Greve, W., 1994. The aging self: stabilizing and protective processes. Developmental Review 14 (1), 52–80.

Brickman, P., Rabinowitz, C.V., Karuza, J., Coates, D., Cohn, E., Kidder, L., 1982. Models of helping and coping. American Psychologist 37 (4), 368–384.

Brion, J.P., 1998. The role of neurofibrillary tangles in Alzheimer disease. Acta Neurologica Belgica 98 (2), 165–174.

Bronfenbrenner, U., 1979. The Ecology of Human Development. Harvard University Press, Cambridge, MA.

de Boer, M.E., Droes, R.M., Jonker, C., Eefsting, J.A., Hertogh, C.M.P.M., 2010. The lived-experiences of early-stage dementia and the feared suffering: an explorative survey. Tijdschrift voor Gerontologie en Geriatrie 41 (5), 194–203.

Castellano, J.M., Kim, J., Stewart, F.R., Jiang, H., Demattos, R.B., Patterson, B.W., Holtzman, D.M., 2011. Human apoE isoforms differentially regulate brain amyloid-β peptide clearance. Science Translational Medicine 3 (89), 89ra57.

Di Rosa, M., Kofahl, C., McKee, K., Bien, B., Lamura, G., Prouskas, C., Döhner, H., Mnich, E., 2011. A typology of caregiving situations and service use in family carers of older people in six European countries: the EUROFAMCARE study. GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry 24 (1), 5–18.

Emmons, R.A., Colby, P.M., Kaiser, H.A., 1998. When losses lead to gains: personal goals and the recovery of meaning. In: Wong, P.T.P., Fry, P.S. (Eds.), The Human Quest for Meaning: A Handbook of Psychological Research and Clinical Applications. Erlbaum, Mahwah, NJ, pp. 163–178.

Filipp, S.-H., Ferring, D., 2002. Die Transformation des Selbst in der Auseinanderstzung mit kritischen Lebensereignissen [transformation of the self in coping with critical life events]. In: Jüttemann, G., Thomae, H. (Eds.), Persönlichkeit und Entwicklung [Personality and Development]. Beltz, Weinheim. S. 191–228.

Ferri, C.P., Prince, M., Brayne, C., Brodaty, H., Fratiglioni, L., Ganguli, M., Alzheimer's Disease International, 2005. Global prevalence of dementia: a Delphi consensus study. Lancet 366 (9503), 2112–2117.

Ferring, D., 2010. Intergenerational relations in aging societies: emerging topics in Europe. Journal of Intergenerational Relations 8 (1), 101–104.

Ferring, D., Boll, T., 2010. Subjective well-being in old age. In: Bovenberg, L., van Soest, A., Zaid, A. (Eds.), Aging, Health, and Pensions in Europe. Palgrave Macmillan, Houndmills, Basingstoke, Hampshire, pp. 173–205.

Ferring, D., Filipp, S.-H., 2000. Coping as a "reality construction": on the role of attentive, comparative, and interpretative processes in coping with cancer. In: Harvey, J., Miller, E. (Eds.), Loss and Trauma. General and Close Relationship Perspectives. Brunner/Mazel Routledge, Philadelphia, PA, pp. 146–165.

Haass, C., 1996. Presenile because of presenilin: the presenilin genes and early onset Alzheimer's disease. Current Opinion in Neurology 9 (4), 254–259.

Helzner, E.P., Scarmeas, N., Cosentino, S., Tang, M.X., Schupf, N., Stern, Y., 2008. Survival in Alzheimer disease: a multiethnic, population-based study of incident cases. Neurology 71, 1489–1495.

Hicks, J.A., Routledge, C., 2013. The Experience of Meaning in Life. Classical Perspectives, Emerging Themes, and Controversies. Springer, New York.

Horowitz, M.J., 1982. Stress response syndromes and their treatment. In: Goldberger, L., Breznitz, S. (Eds.), Handbook of Stress. Free Press, New York, pp. 711–733.

Jack, C.R., Albert, M.S., Knopman, D.S., McKhann, G.M., Sperling, R.A., Carrillo, M.C., Phelps, C.H., 2011. Introduction to the recommendations from the national institute on aging–Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia 7 (3), 257–262.

Lazarus, R.S., Folkman, S., 1984. Stress, Appraisal, and Coping. Springer, New York.

Maki, Y., Amari, M., Yamaguchi, T., Nakaaki, S., Yamaguchi, H., 2012. Anosognosia: patients' distress and self-awareness of deficits in Alzheimer's disease. American Journal of Alzheimer's Disease and Other Dementias 27 (2), 339–345.

Monin, J.K., Schulz, R., 2009. Interpersonal effects of suffering in older adult caregiving relationships. Psychology and Aging 24 (3), 681–695.

Monin, J.K., Schulz, R., 2010. The effects of suffering in chronically ill older adults on the health and well-being of family members involved in their care: the role of emotion-related processes. GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry 23 (4), 207–213.

Pearlin, L.I., Aneshensel, C.S., 1994. Caregiving: the unexpected career. Social Justice Research 7 (4), 373–390.

Pinquart, M., Sörensen, S., 2003. Differences between caregivers and noncaregivers in psychological health and physical health: a meta-analysis. Psychology of Aging 18 (2), 250–267.

Pinquart, M., Sörensen, S., 2007. Correlates of physical health of informal caregivers: a meta-analysis. Journals of Gerontology Series B: Psychological Sciences and Social Sciences 62B, P126–P137.

Pinquart, M., Sörensen, S., 2011. Spouses, adult children, and children-in-law as caregivers of older adults: a meta-analytic comparison. Psychology and Aging 26 (1), 1–14.

Sabat, S., 2002. Epistemological issues in the study of insight in people with Alzheimer's. Dementia 1 (3), 279–293.

United Nations, Department of Economic and Social Affairs, Population Division, 2013. World Population Ageing 2013. ST/ESA/SER.A/348.

van Vliet, D., de Vugt, M.E., Köhler, S., Aalten, P., Bakker, C., Pijnenburg, Y.A., Vernooij-Dassen, M.J., Koopmans, R.T., Verhey, F.R., 2013. Awareness and its association with affective symptoms in young-onset and late-onset Alzheimer disease: a prospective study. Alzheimer Disease and Associated Disorders 27 (3), 265–271.

Wilkins, V.M., Bruce, M.L., Sirey, J.A., 2009. Caregiving tasks and training interest of family caregivers of medically ill homebound older adults. Journal of Aging and Health 21 (3), 528–542.

Wimo, A., Winblad, B., Jonsson, L., 2010. The worldwide societal costs of dementia: estimates for 2009. Alzheimer's & Dementia 6, 98–103.

Zittoun, T., Valsiner, J., Vedeler, D., Salgado, J., Gonçalves, M.M., Ferring, D., 2013. Human Development in the Life Course. Melodies of Living. Cambridge University Press, Cambridge.

Alzheimer's Disease, Neural Basis of

Heiko Braak and Kelly Del Tredici-Braak, University of Ulm, Ulm, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by H. Braak, pp. 425–430, © 2001, Elsevier Ltd.

Abstract

Evolutionally speaking, the human species appears to have a marked tendency to develop lesions that may ultimately lead to the clinical picture of Alzheimer's disease. Intraneuronal lesions initially consist of variably soluble aggregates of abnormal and hyperphosphorylated tau protein (pretangles) that gradually convert into insoluble fibrillar neuropil threads and neurofibrillary tangles. Extracellular plaque-like deposits of β-amyloid protein develop after the tau pathology and occasionally induce the formation of neuritic plaques. Although the pathological process underlying Alzheimer's disease lasts nearly an entire lifetime, only a small number of individuals ultimately develop the clinical symptoms of Alzheimer's disease.

Alzheimer's disease (AD), first described by the physician Alois Alzheimer in 1906, is an insidious progressive neurodegenerative disorder that only can be detected clinically in its final phase. A definitive diagnosis based on antemortem observations is complicated and often misleading (Montine et al., 2012). A major histopathological criterion of AD, and one that is decisive for its postmortem diagnosis, is the assessment of distinctive alterations within the neuronal cytoskeleton that appear in the form of nonargyrophilic pretangle material (precursor of neurofibrillary lesions that consists of variably soluble aggregates of abnormal tau protein that cannot be degraded by involved nerve cells), argyrophilic neuropil threads (NTs), and neurofibrillary tangles (NFTs) in specific subsets of nerve cells of the human central nervous system (see Section The Intraneuronal Formation of Abnormal Tau Protein). Accompanying pathological alterations that appear later than the aforementioned intraneuronal changes include extracellular deposits of the pathological protein β-amyloid (see Section The Extracellular Deposition of Abnormal β-Amyloid Protein) and the formation of neuritic plaques (see Section The Appearance and Composition of Neuritic Plaques).

The Neurodegenerative Process

The Intraneuronal Formation of Abnormal Tau Protein

Tau is a special protein that is mainly associated with the cytoskeleton of cells. Microtubules are normally stabilized by the tau protein. In a hyperphosphorylated state, abnormal tau loses its binding capacity to microtubules. A turning point in the degenerative process are pathological alterations of the cytoskeleton, a kind of internal moveable scaffolding occurring in every living cell, which result from the formation of an abnormal and hyperphosphorylated tau protein in a few susceptible types of neurons (Goedert et al., 1997; Goedert, 1999; Mandelkow et al., 2007; Alonso et al., 2008; Iqbal et al., 2009). In healthy nerve cells, the protein tau is one of the several specific cellular proteins that is associated with and stabilizes components of the cytoskeleton. The abnormal but variably soluble and nonargyrophilic pretangle material that initially emerges fills the entire nerve cell (Bancher et al., 1989). At first, the cell body and the cellular processes of such neurons hardly deviate from their normal shape. In a series of additional steps, this material aggregates to form argyrophilic and nonbiodegradable filaments: the NTs and NFTs that are the hallmarks of AD. NFTs, which often have a flame or cometlike appearance, gradually fill large portions of the nerve cell body and appear black after staining with special silver techniques (Figure 1(a)) (Braak and Braak, 1991a,b; Trojanowski et al., 1995; Esiri et al., 1997). Nerve cells that contain NFTs can survive for years despite marked cytoskeletal alterations. However, they forfeit many of their functional capacities long before premature cell death occurs. After deterioration and disappearance of the host cell, a cluster of the pathological material remains visible in the surrounding brain tissue as a remnant, or the so-called tombstone tangle where it marks the site of the lost neuron (Figure 1(a)). The fact that tombstone tangles are never observed in the absence of fresh pretangle or neurofibrillary lesions accounts for the absence of spontaneous remission in AD patients. In the course of the illness, all of the involved nerve cells proceed through a pretangle phase before developing the argyrophilic filaments. The potential for reversing the pathological process is most probably at its peak during the pretangle phase. Neurofibrillary changes of the Alzheimer type include stable (insoluble) fibrillary inclusions that mainly consist of abnormal and hyperphosphorylated tau protein and develop in a few susceptible types of nerve cells in the central nervous system of the human brain.

The Extracellular Deposition of Abnormal β-Amyloid Protein

Among the many ingredients of the fluid filling the small space between brain cells (the extracellular space) are soluble proteins of still unknown function. Some of them probably result from processing of a normal component of nerve cell membranes known as the amyloid precursor protein (APP). Under certain conditions and for reasons that are still little understood, abnormal processing of the APP takes place, thereby resulting in the formation of the pathological β-amyloid protein, a hydrophobic self-aggregating peptide (Selkoe, 1994; Beyreuther and Masters, 1997; Esiri et al., 1997). It is unclear whether all of the nerve cell types of the human central nervous system contribute to the production of the abnormal protein to the same degree. Within the extracellular space, the β-amyloid

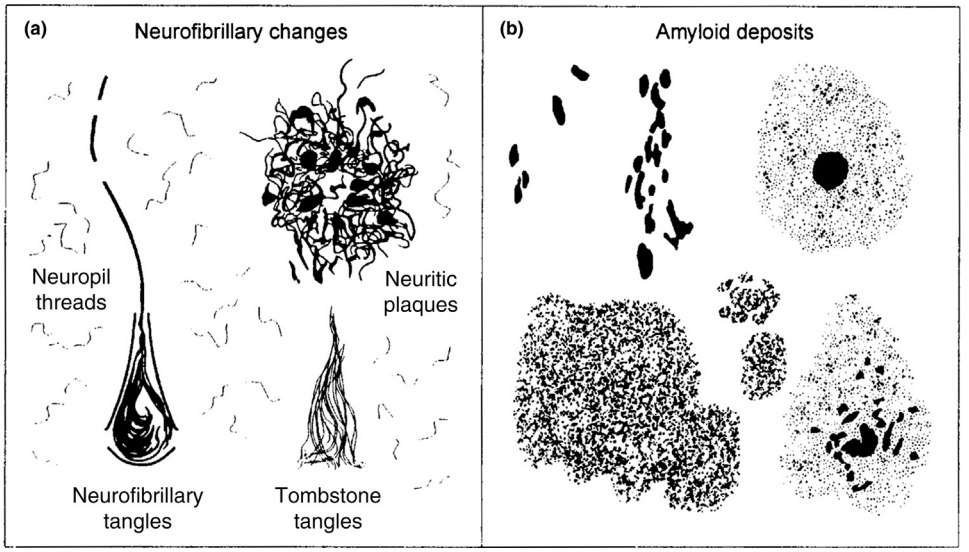

Figure 1 The hallmark of AD is precipitation of abnormal proteins at both intraneuronal and extracellular cerebral locations. (a) The final intraneuronal tau deposits represent the neurofibrillary alterations of the Alzheimer type and include three distinct kinds of lesions: Strands of abnormal tau protein located within involved nerve cell bodies (neurofibrillary tangles) and within their dendritic processes (neuropil threads), as well as abnormal fibrous material which accumulates in swollen cellular processes of neuritic plaques. Extracellular 'tombstone' tangles mark the sites at which the host nerve cells perished. (b) The plaquelike extracellular deposits are composed chiefly, but not exclusively, of β-amyloid protein. Depending on the texture of the neuropil (i.e., brain tissue consisting of nerve cell and glial cell cellular processes; see Section The Extracellular Deposition of Abnormal β-Amyloid Protein), plaques occur in different sizes and shapes. Most cortical β-amyloid deposits evolve as globular structures with or without a condensed core.

protein eventually builds plaquelike deposits of varying sizes and shapes (Figure 1(b)). Globular forms predominate in the cerebral neocortex and occur preferentially in layers III and V. (Neocortex is a part of the cerebral cortex that is typically formed of six layers. The human cerebral cortex consists chiefly of neocortical regions.) Bandlike formations are seen in layer I of the cortex, whereas layers II, IV, and VI are often spared. In the course of the disease, the number of plaques reaches a maximum; yet, even at their greatest extent, large areas of the brain tissue remain free of these deposits. The β-amyloid plaques develop later in the course of the pathological AD process than the intraneuronal abnormal tau material. In addition, they do not correlate with the severity of the clinical symptoms (mild cognitive impairment (MCI), dementia) associated with AD, although intensive research efforts have been devoted to establishing the existence of such a relationship (Schönheit et al., 2004; Korczyn, 2008; Pimplikar, 2009; Braak et al., 2011).

The Appearance and Composition of Neuritic Plaques

Neuritic plaques consist of aggregations of altered glial cells and swollen cellular processes of nerve cells, which, in part, contain argyrophilic and fibrillary masses of the abnormal tau protein (Figure 1(a)). Taken together, the various types of glial cells in the brain outnumber nerve cells by a ratio of approximately 10 to 1: they are the maintenance and repair forces, so to speak, for the neurons. Deposits of β-amyloid fill the extracellular space within the reaches of neuritic plaques. Neuritic plaques are more patchily distributed than simple β-amyloid deposits and are present at much lower densities. Factors inducing the formation of neuritic plaques and the circumstances accompanying their disappearance from the tissue are currently unknown. Some cortical areas remain devoid of neuritic plaques, whereas others display high densities of these lesions quite early in the disease.

Developmental Sequence of the Lesional Distribution Pattern of the Intraneuronal Cytoskeletal Alterations

The intraneuronal tau lesions develop gradually and nearly bilaterally symmetrically at predisposed sites of the central nervous system. Many neuronal types, cortical areas, and subcortical nuclei remain uninvolved, whereas others undergo severe damage. The pathological process begins in select subcortical nuclei, progresses eventually in a topographically systematic manner throughout the cerebral cortex, and requires nearly a lifetime to reach its full extent. The sequence of the spread is predictable and remarkably consistent across cases (Arnold et al., 1991; Braak and Braak, 1991a, 1997; Fewster et al., 1991). By pinpointing the locations of involved neurons, five stages in the development of nonargyrophilic pretangle material and six stages in the evolution of argyrophilic NTs/NFTs can be differentiated (Braak and Braak, 1991a; Braak et al., 2011).

The relationships between age and the AD-associated cytoskeletal pathology can be studied by staging large numbers of nonselected autopsy brains (Braak and Braak, 1997). (The pathological process progresses in a systematic topographical manner through vulnerable brainstem nuclei and cortical regions with relatively little interindividual differences. Staging allows one to distinguish between different neuropathological stages.) The vast majority of

recently studied cases (i.e., $n = 2322$ of 2332 cases) can be assigned to one of these stages, and Figure 2 illustrates a continuum of the cytoskeletal alterations ranging from the very first pretangle material-containing nerve cell (appearing at subcortical stage a) to the extensive destruction seen in fully developed AD (NFT stage VI) (Braak et al., 2011).

Pretangle Stages (a–c, 1a, 1b)

Figure 2 shows the percentage of cases at the recently described pretangle stages (Braak and Del Tredici, 2011a). The most recent study showed that it was impossible to assign an NFT stage to 342 of the total number of 2332 cases (Braak et al., 2011). The brains of only 10 individuals were entirely devoid of abnormal tau pathology (=0.4% of the total 2332 cases) (Figure 2(a)). Fifty eight cases showed only pretangle material-containing nerve cells in a few particularly susceptible nuclei of the brainstem. These cases are referred to as subcortical pretangle stages a–c and occur preferentially in young persons (Figure 2(b)).

Early pretangle material in the brainstem develops only in projection neurons that generate a long and poorly myelinated

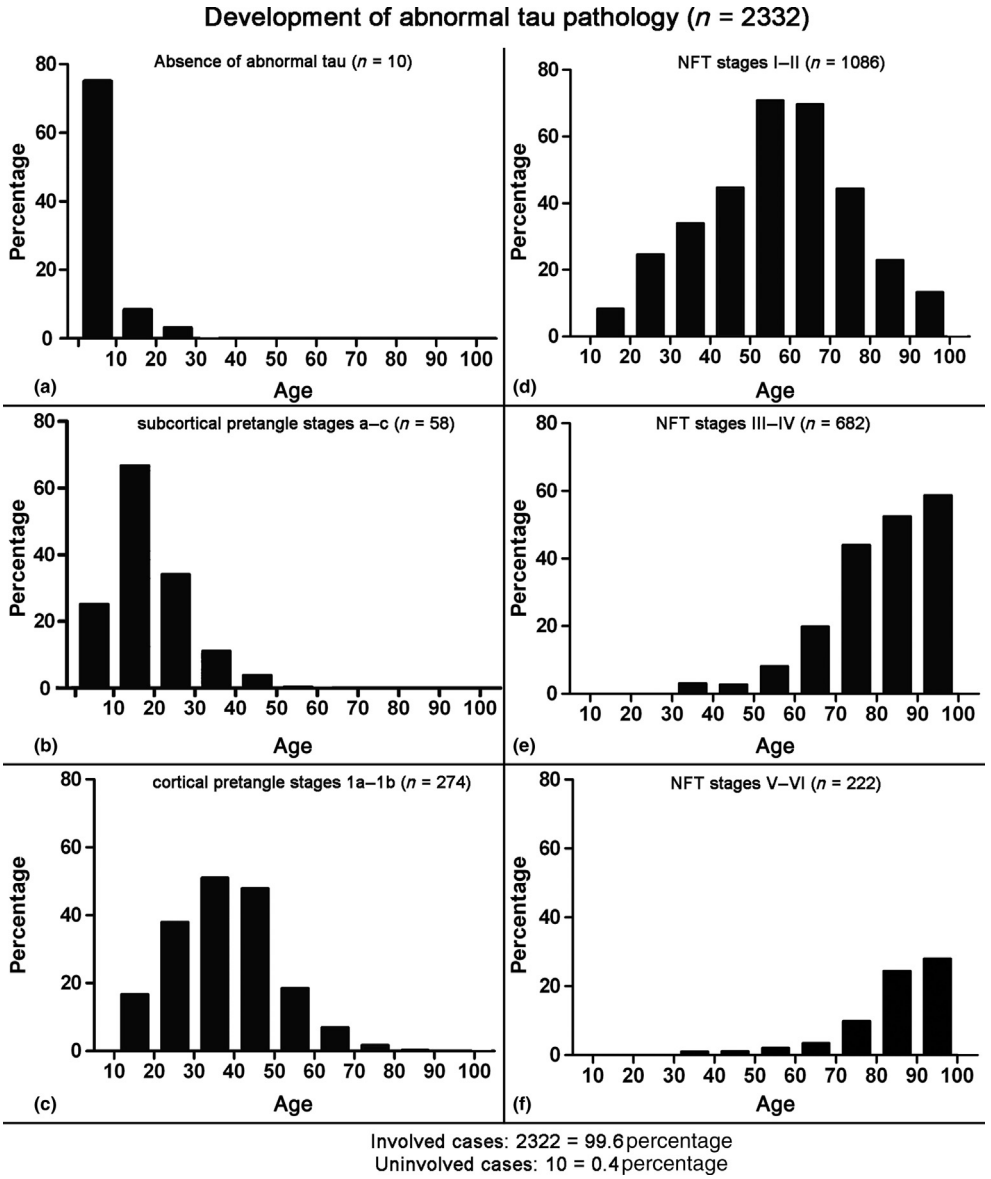

Figure 2 Relative frequency of occurrence of abnormal tau stages in 2332 nonselected autopsy cases. (a) Brains of only a minority of cases show absence of any abnormal tau deposits. (b) Nonargyrophilic and nonfibrillar pretangle material occurs in nerve cells or nerve cell processes of subcortical nuclei that diffusely project to the cerebral cortex (stages a–c pathology). (c) Pretangle material additionally develops either in nerve cell processes (pretangle stage 1a) or in cortical pyramidal cells (pretangle stage 1b) usually within the transentorhinal region. (d–f) Fibrillar and argyrophilic lesions (neuropil threads and neurofibrillary tangles) characterize the cortical neurofibrillary stages I–VI. (d) Stage I–II cases, (e) Stage III–IV cases, (f) Stage V–VI cases.

ascending axon with extensive and diffuse ramifications within the cerebral cortex (e.g., cholinergic basal forebrain nuclei, serotonergic upper raphe nuclei, noradrenergic locus coeruleus) (German et al., 1987; Mesulam et al., 2004; Grinberg et al., 2009). Pretangle stage c cases have pathological cells in more than one of these diffusely projecting nuclei. Abnormal pretangle material is most frequently observed in the locus coeruleus of the pontine tegmentum (Grudzien et al., 2007; Sara, 2009; Braak and Del Tredici, 2011a,b, 2012), and involved cells are confined to this nucleus in pretangle stages a and b (Figure 2(b)). It now appears that the AD process begins in the axon (pretangle stage a) and then goes on to involve the somatodendritic compartment of coeruleus neurons (pretangle stage b). The locus coeruleus is also most probably the site where the AD-associated pathological process in the brain begins (Braak and Del, Tredici 2011a; Braak et al., 2011). Noradrenergic and neuromelanin-containing projection cells comprise the locus coeruleus in the pontine tegmentum of the human brainstem. This nucleus sends diffuse projections to virtually all portions of the central nervous system and modifies functions of the cerebral cortex (cognition, attentional states).

Hence, the earliest cytoskeletal alterations are capable of evolving in brains of otherwise healthy and young individuals (Elobeid et al., 2011). Advanced age is itself not a prerequisite for the development of abnormal tau. Accordingly, the initiation of the pathological process underlying AD is by no means an age-dependent one (Nelson et al., 2011). By incorporating the pretangle stages into the natural history of the tau lesions, the preclinical phase of AD lasts for nearly a lifetime (Figure 2). Thus, the pathological process underlying AD offers a much larger window of opportunity for disease-modifying interventions than previously imagined (Braak and Del Tredici, 2012).

In addition to the subcortical lesions described, 274/2332 cases displayed the presence of pretangle material in the cerebral cortex (Figure 2(c)). Cases with involvement confined to neuronal processes alone in the cortical transentorhinal region are referred to as cortical pretangle stage 1a, whereas in cortical pretangle stage 1b entirely immunoreactive pyramidal cells are seen in the transentorhinal region (Braak et al., 2011), a primate-specific transition zone between the allocortical entorhinal region, on the one hand, and the temporal neocortex, on the other side. This transition zone is particularly well developed in the human brain. It is the first site within the cerebral cortex to become involved in the pathological process underlying AD.

NFT Stages (I–VI)

NFT Stages I and II

In NFT stages I and II (Figure 2(d)), argyrophilic tau lesions are mainly localized within specific projection cells located in small areas of the medial temporal lobe that are important cortical components of the limbic system. Deposition of β-amyloid protein is nearly always absent during development of these initial stages. Bilateral structural preservation of these neurons in the medial temporal lobe is one of the prerequisites for retaining memory and learning capacities. During NFT stage I, vulnerable cortical neurons occupy mainly the transentorhinal region of the temporal lobe (Braak et al., 2011). In some NFT stage I cases, the transentorhinal tau pathology may be confined to no more than a single argyrophilic (Gallyas-positive) pyramidal cell. The negligible to mild neuronal destruction seen in NFT stages I and II still remains below the threshold required for the manifestation of clinical symptoms (see Section NFT Stages (I–VI)). These first two NFT stages are already found in teenagers and increase in frequency steadily up to the sixth decade. Thereafter, they decrease and become replaced by more advanced NFT stages (Figure 2(d)).

Lesions that exist prior to the clinical manifestation of many diseases are generally regarded as prodromes. NFT stages I–II and/or pretangle stages are not detectable by means of currently available diagnostic instruments or tests. Nevertheless, these lesions represent pathological changes because the contention that they are normal requires the problematic definition of a point at which abnormal tau 'normally' converts into 'disease-related' abnormal tau. AD as a clinical entity includes dementia, but the AD-associated pathological process is much longer and includes an extended prodromal (preclinical) phase.

NFT Stages III and IV

Subsequently, the lesions reach the hippocampal formation (NFT stage III) and neocortical areas of the basal temporal lobe that adjoin the transentorhinal region laterally. Thereafter, the process encroaches upon anterior cingulate, insular, and more distant neocortical destinations of the basal temporal lobe (NFT stage IV). NFT stages III–IV begin to occur in the fourth decade and increase thereafter in frequency (Figure 2(e)).

The clinical protocols of many individuals at NFT stages III and IV may make reference to MCI, e.g., difficulties solving simple arithmetical or abstract problems, short-term memory or recall deficits (lapses of the so-called working memory), and the presence of changes in personality ranging from suspiciousness to irascible or aggressive behavior, apathy, withdrawal, and bouts of depression (Albert et al., 2011; Markesbery et al., 2006). Imaging techniques for the clinical detection of pathological tau in the brain are in the earliest phase of development (Fodero-Tavoletti et al., 2011; Jensen et al., 2011), and large autopsy-controlled prospective studies are needed to confirm the reliability of currently used biomarkers, particularly in cases with MCI (Jack et al., 1992, 2010).

NFT Stages V and VI

The tau pathology spreads superolaterally (NFT stage V) and finally reaches the primary neocortical motor and sensory fields (NFT stage VI) (Figure 2(f)) (Braak and Braak, 1991b; Braak et al., 2006). With the widespread devastation of the neocortex, patients present with severe dementia (i.e., severe loss of memory, cognitive faculties, and executive functions). The final NFT stages V and VI correspond to clinically fully developed AD. These persons become incontinent for urine and feces, cannot dress themselves unaided, or recognize persons once familiar to them (spouse, children). With the passage of time, they can no longer walk, sit up, or hold up their head unassisted. Increasing rigidity of the large joints leads to irreversible contractures of the extremities and

immobility. Primitive reflexes normally seen only in infants also reappear in the last clinical phases of the illness (Nelson et al., 2007; Dolan et al., 2010; Reisberg and Franssen, 1999; Reisberg et al., 1999; Sabbagh et al., 2010). The period of time, which elapses between the onset of clinical symptoms and death averages approximately 8 to 9 years (Francis et al., 1999).

The prevalence of AD cases at stages V and VI increases markedly with age (Figure 2(f)) and thereby imposes major socioeconomic burdens on societies with a growing life expectancy. The initial and provisional clinical AD diagnosis is usually made at stage V or VI. At these stages, therapeutic interventions are useless because available antidementia drugs cannot arrest, reverse, or modify the tau pathology (Molnar et al., 2009; Wollen, 2010).

AD and Evolution

The marked proclivity to develop abnormal tau inclusions in predisposed neuronal types appears to be intrinsic to the human brain, inasmuch as 99.6% of the total number of recently examined cases show the presence of AD-associated tau lesions (Figure 2) (Braak et al., 2011). A comparable but remarkably lower propensity for the formation of abnormal tau in similar neuronal types has been reported in a few nonhuman mammalian species (Nelson et al., 1994; Cork et al., 1998; Schultz et al., 2000). However, in these species pretangle material develops only at an advanced age in animals living in captivity and not in the wild.

The proclivity of the human brain to develop abnormal tau probably increased considerably during the last phase of evolution when humans appeared among higher primates. This evolutionary phase was characterized by a remarkable increase in brain volume. The neocortex became more expansive and, at the same time, the allocortical entorhinal region (in macrosmatic mammals chiefly devoted to the processing of olfactory data) underwent reorganization in microsmatic primates. (Allocortex is a part of the cerebral cortex that shows a lamination pattern different from that of the neocortex – that is, fewer or more layers as the six-layered neocortex. Important components of the allocortex are the hippocampal formation, the entorhinal region, and the olfactory bulb.) The entorhinal region lost most of its olfactory functions and established new interconnections to the neocortex via the transentorhinal region, thereby preparing the way for massive indirect neocortical input to the hippocampal formation. (The entorhinal region is a part of the allocortex. In the human brain, the entorhinal region functions as kind of an association cortex for the hippocampal formation. It indirectly receives important information from neocortical regions and transfers these data to the hippocampal formation.) These phylogenetically late events left their mark on late-differentiating neuronal types during ontogenesis (Braak and Del Tredici, 2004). The final step in neuronal differentiation is the production of the myelin sheath around the axon (van der Knaap et al., 1991). A singular feature of neurons that are susceptible to the AD process is that their axons undergo myelination postnatally. Late-maturing oligodendrocytes myelinate large numbers of axons simultaneously and, for this reason, some of these axons may fail to achieve an adequate degree of myelination (Ullén, 2009). Thus, the propensity to produce abnormal tau may be associated with these evolutional brain developments (Bartzokis, 2004; Braak and Braak, 1996; Reisberg et al., 1999). However, they would have had no evolutional significance for the survival and further propagation of the human species because our forebearers, owing to their limited life expectancy, would not have lived long enough to develop cognitive dysfunction or dementia.

See also: Alzheimer's Disease: Behavioral and Social Aspects; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Dementia: Psychiatric Aspects.

Bibliography

Albert, M.S., Dekosky, S.T., Dickson, D., Dubois, B., Feldman, H.H., Fox, N.C., Gamst, A., Holtzman, D.M., Jagust, W.U., Petersen, R.C., Snyder, P.J., Camillo, M.C., Thies, B., Phelps, C.H., 2011. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's and Dement 7, 270–279.

Alonso, A.C., Li, B., Grundke-Iqbal, I., Iqbal, K., 2008. Mechanism of tau-induced neurodegeneration in Alzheimer disease and related tauopathies. Current Alzheimer Research 5, 375–384.

Arnold, S.E., Hyman, B.T., Flory, J., Damasio, A.R., van Hoesen, G.W., 1991. The topographical and neuroanatomical distribution of neurofibrillary tangles and neuritic plaques in the cerebral cortex of patients with Alzheimer's disease. Cerebral Cortex 1, 103–116.

Bancher, C., Brunner, C., Lassmann, H., Budka, H., Jellinger, K., Wiche, G., Seitelberger, F., Grundke-Iqbal, I., Iqbal, K., Wisniewski, H.M., 1989. Accumulation of abnormally phosphorylated tau precedes the formation of neurofibrillary tangles in Alzheimer's disease. Brain Research 477, 90–99.

Bartzokis, G., 2004. Age-related myelin breakdown: a developmental model of cognitive decline and Alzheimer's disease. Neurobiology of Aging 25, 5–18.

Beyreuther, K., Masters, C.L., 1997. Serpents on the road to dementia and death. Nature Medicine 3, 723–725.

Braak, H., Braak, E., 1991a. Demonstration of amyloid deposits and neurofibrillary changes in whole brain sections. Brain Pathology 1, 213–216.

Braak, H., Braak, E., 1991b. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathologica 82, 239–259.

Braak, H., Braak, E., 1996. Development of Alzheimer-related neurofibrillary changes in the neocortex inversely recapitulates cortical myelogenesis. Acta Neuropathologica 92, 97–101.

Braak, H., Braak, E., 1997. Frequency of stages of Alzheimer-related lesions in different age categories. Neurobiology of Aging 18, 351–357.

Braak, H., Del Tredici, K., 2004. Poor and protracted myelination as a contributory factor to neurodegenerative disorders. Neurobiology of Aging 25, 19–23.

Braak, H., Del Tredici, K., 2011a. The pathological process underlying Alzheimer's disease in individuals under thirty. Acta Neuropathologica 121, 171–181.

Braak, H., Del Tredici, K., 2011b. Alzheimer's pathogenesis: is there neuron-to-neuron propagation? Acta Neuropathologica 121, 589–595.

Braak, H., Del Tredici, K., 2012. Alzheimer's disease: pathogenesis and prevention. Alzheimer's and Dement (PM Epub 2012 March 30 ahead of print PMID: 22465174).

Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H., Del Tredici, K., 2006. Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. Acta Neuropathologica 112, 389–404.

Braak, H., Thal, D.R., Ghebremedhin, E., Del Tredici, K., 2011. Stages of the pathologic process in Alzheimer disease: age categories from 1 to 100 years. Journal of Neuropathology and Experimental Neurology 70, 960–969.

Cork, L.C., Powers, R.E., Selkoe, D.J., Davies, P., Geyer, J.J., Price, D.L., 1998. Neurofibrillary tangles and senile plaques in aged bears. Journal of Neuropathology and Experimental Neurology 47, 629–641.

Dolan, D., Troncoso, J., Resnick, S.M., Crain, B.J., Zonderman, A.B., O'Brien, R.J., 2010. Age, Alzheimer's disease and dementia in the Baltimore Longitudinal Study of ageing. Brain 133, 2225–2231.

Elobeid, A., Soininen, H., Alafuzoff, I., 2011. Hyperphosphorylated tau in young and middle-aged subjects. Acta Neuropathologica 123, 97–104.

Esiri, M.M., Hyman, B.T., Beyreuther, K., Masters, C., 1997. Aging and dementia. In: Graham, D.L., Lantos, P.I. (Eds.), Greenfield's Neuropathology. Arnold, London, pp. 153–234.

Fewster, P.H., Griffin-Brooks, S., MacGregor, J., Ojalvo-Rose, E., Ball, M.J., 1991. A topographical pathway by which histopathological lesions disseminate through the brain of patients with Alzheimer's disease. Dementia and Geriatric Cognitive Disorders 2, 121–132.

Fodero-Tavoletti, M.T., Okamura, N., Furumoto, S., Mulligan, R.S., Connor, A.R., McLean, C.A., Cao, D., Rigopoulos, A., Cartwright, G.A., O'Keefe, G., Gong, S., Adlard, P.A., Barnham, K.J., Rowe, C.C., Masters, C.L., Kudo, Y., Cappai, R., Yanai, K., Villemagne, V.L., 2011. 18F-THK523: a novel in vivo tau imaging ligand for Alzheimer's disease. Brain 134, 1089–1100.

Francis, P.T., Palmer, A.M., Snape, M., Wilcock, G.K., 1999. The cholinergic hypothesis of Alzheimer's disease: a review of progress. Journal of Neurology, Neurosurgery and Psychiatry 66, 137–147.

German, D.C., White, C.L., Sparkman, D.R., 1987. Alzheimer's disease: neurofibrillary tangles in nuclei that project to the cerebral cortex. Neuroscience 21, 305–312.

Goedert, M., 1999. Filamentous nerve cell inclusions in neurodegenerative diseases: tauopathies and alpha-synucleinopathies. Philosophical Transactions of the Royal Society London B 354, 1101–1108.

Goedert, M., Trojanowski, J.Q., Lee, V.M.Y., 1997. The neurofibrillary pathology of Alzheimer's disease. In: Rosenberg, R.N. (Ed.), The Molecular and Genetic Basis of Neurological Disease, second ed. Butterworth-Heinemann, Woburn, MA, pp. 613–627.

Grinberg, L.T., Rüb, U., Ferretti, R.E.L., Nitrini, R., Farfel, J.M., Polichiso, L., Gierga, K., Jacob-Filho, W., Heinsen, H., Brazilian Brain Bank Study Group, 2009. The dorsal raphe nucleus shows phospho-tau neurofibrillary changes before the transentorhinal region in Alzheimer's disease. A precocious onset? Neuropathology and Applied Neurobiology 35, 406–416.

Grudzien, A., Shaw, P., Weintraub, S., Bigio, E., Mash, D.C., Mesulam, M.M., 2007. Locus coeruleus neurofibrillary degeneration in aging, mild cognitive impairment and early Alzheimer's disease. Neurobiology of Aging 28, 327–335.

Iqbal, K., Liu, F., Gong, C.X., Alonso, A.C., Grundke-Iqbal, I., 2009. Mechanisms of tau-induced neurodegeneration. Acta Neuropathologica 118, 53–69.

Jack, C.R., Petersen, R.C., O'Brien, P.C., Tangalos, E.G., 1992. MR-based hippocampal volumetry in the diagnosis of Alzheimer's disease. Neurology 42, 183–188.

Jack, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q., 2010. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurology 9, 119–128.

Jensen, J.R., Cisek, K., Funk, K.E., Naphade, S., Schafer, K.N., Kuret, J., 2011. Research towards tau imaging. Journal of Alzheimer's Disease 26 (Suppl. 3), 147–157.

Korczyn, A.D., 2008. The amyloid cascade hypothesis. Alzheimer's and Dementia 4, 176–178.

Mandelkow, E., von Bergen, M., Biernat, J., Mandelkow, E.M., 2007. Structural principles of tau and the paired helical filaments of Alzheimer's disease. Brain Pathology 17, 83–90.

Markesbery, W.R., Schmitt, F.A., Kryscio, R.J., Davis, D.G., Smith, C.D., Wekstein, D.R., 2006. Neuropathologic substrate of mild cognitive impairment. Archives of Neurology 63, 38–46.

Mesulam, M.M., Shaw, P., Mash, D., Weintraub, S., 2004. Cholinergic nucleus basalis tauopathy emerges early in the aging-MCI-AD continuum. Annals of Neurology 55, 815–828.

Molnar, F.J., Man-Son-Hing, M., Fergusson, D., 2009. Systematic review of measures of clinical significance employed in randomized controlled trials of drugs for dementia. Journal of the American Geriatrics Society 57, 536–546.

Montine, T.J., Phelps, C.H., Beach, T.G., Bigio, E.H., Cairns, N.J., Dickson, D.W., Duyckaerts, C., Frosch, M.P., Masliah, E., Mirra, S.S., Nelson, P.T., Schneider, J.A., Thal, D.R., Trojanowski, J.Q., Vinters, H.V., Hyman, B.T., 2012. National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease: a practical approach. Acta Neuropathologica 123, 1–11.

Nelson, P.T., Greenberg, S.G., Saper, C.B., 1994. Neurofibrillary tangles in the cerebral cortex of sheep. Neuroscience Letters 170, 187–190.

Nelson, P.T., Jicha, G.A., Schmitt, F.A., Liu, H., Davis, D.G., Mendiondo, M.S., Abner, E.L., Markesbery, W.R., 2007. Clinicopathologic correlations in a large Alzheimer disease center autopsy cohort: neuritic plaques and neurofibrillary tangles "do count" when staging disease severity. Journal of Neuropathology and Experimental Neurology 66, 1136–1146.

Nelson, P.T., Head, E., Schmitt, F.A., Davis, D.G., Neitner, J.H., Jicha, G.A., Abner, E.L., Smith, C.D., Van Eldik, L.J., Kryscio, R.J., Scheff, S.W., 2011. Alzheimer's disease is not "brain aging": neuropathological, genetic, and epidemiological human studies. Acta Neuropathologica 121, 571–587.

Pimplikar, S.W., 2009. Reassessing the amyloid cascade hypothesis of Alzheimer's disease. The International Journal of Biochemistry and Cell Biology 41, 1261–1268.

Reisberg, B., Franssen, E.H., 1999. Clinical stages of Alzheimer's disease. In: de Leon, M.J. (Ed.), An Atlas of Alzheimer's Disease. The Encyclopedia of Visual Medicine Series. Parthenon, New York and London, pp. 11–20.

Reisberg, B., Franssen, E.H., Hasan, S.M., Monteiro, I., Boksay, I., Souren, L.E., Kenowski, S., Auer, S.R., Elahi, S., Kluger, A., 1999. Retrogenesis: clinical, physiologic, and pathologic mechanisms in brain aging, Alzheimer's and other dementing processes. European Archives of Psychiatry and Clinical Neuroscience 249 (Suppl. 3), 28–36.

Sabbagh, M.N., Cooper, K., DeLange, J., Stoehr, J.D., Thind, K., Lahti, T., Reisberg, B., Sue, L., Vedders, L., Fleming, S.R., Beach, T.G., 2010. Functional, global and cognitive decline correlates to accumulation of Alzheimer's pathology in MCI and AD. Current Alzheimer Research 7, 280–286.

Sara, S.J., 2009. The locus coeruleus and noradrenergic modulation of cognition. Nature Reviews Neuroscience 10, 211–223.

Schönheit, B., Zarski, R., Ohm, T.G., 2004. Spatial and temporal relationships between plaques and tangles in Alzheimer-pathology. Neurobiology of Aging 25, 697–711.

Schultz, C., Dehghani, F., Hubbard, G.B., Thal, D.R., Struckhoff, G., Braak, E., Braak, H., 2000. Filamentous tau pathology in nerve cells, astrocytes, and oligodendrocytes of aged baboons. Journal of Neuropathology and Experimental Neurology 59, 39–52.

Selkoe, D.J., 1994. Alzheimer's disease: a central role for amyloid. Journal of Neuropathology and Experimental Neurology 53, 438–447.

Trojanowski, J.Q., Shin, R.W., Schmidt, M.L., Lee, V.M.Y., 1995. Relationship between plaques, tangles, and dystrophic processes in Alzheimer's disease. Neurobiology of Aging 16, 335–340.

Ullén, F., 2009. Is activity regulation of late myelination a plastic mechanism in the human nervous system? Neuron Glia Biology 5, 29–34.

van der Knaap, M.S., Valk, J., Bakker, C.J., Schooneveld, M., Faber, J.A., Willemse, J., Gooskens, R.H., 1991. Myelination as an expression of the functional maturity of the brain. Developmental Medicine and Child Neurology 33, 849–857.

Wollen, K.A., 2010. Alzheimer's disease: the pros and cons of pharmaceutical, nutritional, botanical, and stimulatory therapies, with a discussion of treatment strategies from the perspective of patients and practitioners. Alternative Medicine Review 15, 223–244.

American and French Revolutions: Impact on the Social Sciences

John C Torpey, Graduate Center, City University of New York, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article surveys the significance of the American and French Revolutions in the modern social sciences. It argues that the French Revolution has historically had a greater impact, except among those political sociologists preoccupied with the question of 'American exceptionalism.' Yet this has changed since the collapse of Soviet Communism, and certain features of the American Revolution and the American experience have come to the fore.

The American Revolution, conventionally dated 1776–83, and the French Revolution that followed it beginning in 1789 have traditionally been regarded as the 'founding' events of political modernity (e.g., Palmer, 1959/1964). The emphasis in these discussions is on the *political* features of modern life – popular sovereignty, constitutionalism, citizenship – as distinct especially from its economic aspects, where the stress would be upon the industrial revolution. The two political revolutions, of which the first very much influenced the second (while French thinkers also had a large impact on their American counterparts), are nonetheless regarded as having rather distinctive characteristics that make them very different both in reality and in perception. While sharing a commitment to what would later become known as universal 'human rights,' the social characteristics of the two 'events' diverged sharply – and this divergence has left a lasting legacy in the social scientific understanding of their significance.

As a result, the American and French Revolutions have had very different 'careers' in social scientific writing. The situation was perhaps best expressed by the political theorist Hannah Arendt, who argued in the early 1960s that, because it failed to address 'the social question' (i.e., the question of social inequality), the American Revolution had been an event of 'little more than local importance' in comparison with its French counterpart, and this despite the fact that the latter had 'ended in disaster' while the former had been 'triumphantly successful' (Arendt, 1963: 56). Arendt argued with some sympathy that revolutions are occasions in human life when the world can be made anew, but this can only happen if they institute the 'constitution of freedom.' As against the legislation of freedom, she insisted, no one can dictate an end to poverty – yet that is precisely what the French Revolution sought to do. Eventually, Arendt argued, Karl Marx would transform 'the Rights of Man into the rights of Sans-Culottes' and revolutions would henceforth be concerned with the social question as much as with the constitution of freedom. All this presaged Arendt's dismissal of the idea of 'the social' – the distinctive arena of study of sociology – which she viewed as a misguided understanding of human arrangements that posed inappropriate constraints on human beings' freedom of action. Drawing always on classical Greek understandings of politics, she insisted that persons should be free to act politically, and that their institutions should be designed to guarantee that freedom; economics, the realm of necessity, could have no part in that realm (Arendt, 1958). The American Revolution was a success, in Arendt's eyes, because it concentrated on the only matter that revolutions can address: the institution of human freedom.

The fact that the American Revolution was perceived as purely a *political* event having no social implications, for better or worse, has had lasting consequences for the way in which the social sciences constructed themselves. For example, the writings of Alexis de Tocqueville ended up largely in political science (*Democracy in America*) and in history (*The Old Regime and the Revolution*), respectively, whereas in fact his work is as deeply sociological in nature as it is possible to imagine. Of course, his views of 'democracy' were not as enthusiastic as many in the post-1960s sociology would have preferred, which helped consign him to other, less 'progressive' disciplines. Meanwhile, Arendt was correct to note that Tocqueville was not really interested in the American Revolution itself; like his follower Weber, he saw the roots of the peculiar American trajectory as lying far back in the European past and as more or less unaffected by the course of the Revolution itself.

Tocqueville had famously been impressed by the 'equality of conditions' (de Tocqueville, 2000) that he claimed to have seen during his visit to the country in the early 1830s. His vision of an egalitarian America has thus long contrasted in social scientific eyes with that of an 'aristocratic' Europe, despite the fact that rates of social mobility have in recent years become greater in social democratic Europe than in the sharply stratified United States of the post-1970s era. Even a century ago, however, the Dutch cultural historian Johan Huizinga scoffed at Tocqueville's image of American egalitarianism, dismissing out of hand his claims about the extent to which economic and racial equality prevailed in American life (Huizinga, 1972: 61, 62). Like other more critical observers, Huizinga found the United States a rather more conservative place than Tocqueville had. This was at least in part because the American Revolution had been – like the Dutch Revolt against the Spanish Habsburgs two centuries earlier – "a conservative revolution – a struggle for the preservation of liberties more than for the conquest of freedom" (Huizinga, 1972: 18). The slaves would presumably have seen it that way as well.

Tocqueville was also struck by the Americans' distinctive 'point of departure,' but this had nothing to do with its revolutionary origins. Instead, the Americans' advantage lay in the fact that they had no feudal regime to combat and to eliminate: "they have arrived at a state of democracy without having to endure a democratic revolution." They had been blessed with

a sort of virgin birth whereby they were 'born equal, instead of becoming so.' This passage, quoted in the frontispiece of Louis Hartz's *Liberal Tradition in America*, defined the terms of the subsequent discussion of 'American exceptionalism' in American social science. "It is not accidental that America[,] which has lacked a feudal tradition[,] has uniquely lacked also a socialist tradition" (Hartz, 1955: 6). Werner Sombart had argued along similar lines in 1905 when he wrote *Why is There No Socialism in America?*, and the influential political sociologist Seymour Martin Lipset would explore the basic problematique throughout his professional career (e.g., 1996).

Tocqueville saw the American Revolution not as a uniquely American event, but as an application of theories largely developed in France: "The Americans seemed merely to apply what our writers had thought of" (1998: 201). In other words, Tocqueville regarded the American Revolution as the fulfillment of French abstractions – ones that in his view would have a very bad outcome in France itself, as we shall see. But the Americans had a way of tempering ideological excesses with the judicious application of political freedom, an institution Tocqueville regarded as woefully underappreciated in his native land. Despite his preoccupation with and durable analysis of the United States, then, it was always really the French Revolution that was more central to his concerns.

Notwithstanding Tocqueville's relative lack of interest in the American Revolution, he was deeply traumatized by the French Revolution. He had good reason for this, as his own father had been thrown in prison by the revolutionaries and, it is said, his father's hair turned white overnight from the shock. Tocqueville regarded the French Revolution as an ongoing catastrophe for France that foretold a future of endless political convulsions that did, in fact, persist at least until the declaration of the Third Republic after the Franco-Prussian War (1870–71). Thereafter, the political upheavals in France tended to come in the context of external events and conflicts, such as the 'strange defeat' by the Nazis in 1940 and the process of decolonization in Algeria. With the notable exception of May 1968 (see Zolberg, 2008), such upheavals have been relatively circumscribed since the end of World War II. Aspects of Tocqueville's analysis of the French Revolution have nonetheless had considerable impact in subsequent social scientific discussion.

Similarly, for Karl Marx, France was the most important actor in the revolutionary drama shaping the modern world; the Germans, by contrast, were daydreamers incapable of acting to bring their political reality up to the standards of the day – which were defined by the French Revolution (Marx, 1978a). As far as Marx was concerned, the Americans were mired in religious superstition – a conclusion he derived chiefly from the writings of Tocqueville and his traveling companion, Gustave de Beaumont, who 'assure us in unison' that "North America is pre-eminently the country of religiosity" (Marx, 1978b: 31).

By contrast, the French Revolution had a profoundly anticlerical thrust. Under the reigning notion of the 'divine right of kings,' the Church had been effectively part of the *ancien régime*; accordingly, anyone who wished to bring down the one was compelled perforce to bring down the other as well. The French Revolution thus left a deep legacy of anticlericalism that has shaped 'progressive' politics ever since. Yet this was a historical coincidence, according to de Tocqueville (1998); the revolutionaries were not antireligious per se. This would become obvious when they sought after toppling the monarchy to create a number of their own religious institutions – the cult of the Supreme Being, a new revolutionary calendar (echoing Islam's recalibration of human time), and the like. Notwithstanding its anticlerical qualities, Tocqueville characterized the French Revolution as 'a religious revolution' because the message of the Revolution was widely understood to be not just for the French alone, but one for all humankind. This universalizing bent of the Revolution was a product of its origins in abstract ideas – the ideas of the Enlightenment, which stressed the corruption of political institutions as well as a post-Christian notion of human perfectibility. A revolution that had to strangle a church to achieve freedom was inspired by quasi-religious notions about a common human dignity.

The situation could not have been more different from that in the United States. There, in Tocqueville's view, religion and freedom marched hand in hand: "Among us [in France], I had seen the spirit of religion and the spirit of freedom almost always move in contrary directions. Here [in the United States] I found them united intimately with one another: they reigned together on the same soil" (de Tocqueville, 2000: 282). Despite the fact that religion "never mixes directly in the government of society," it should be "considered as the first of their political institutions" (2000: 280). This contrast between the American and the Continental experiences would have enduring consequences for the self-conception of the society as well as of the social sciences. In particular, it would in the post-Cold War period animate a reconsideration of the long-dominant theory of 'secularization' that paid more attention to American experience and viewed more skeptically the previously normative status of the French. Henceforth, the theory of secularization would come to have more detractors than defenders, although the rediscovery of religion in social science owed perhaps as much to the terrorist attacks of 11 September 2001 as to a real appreciation of religious motivation in social life (Torpey, 2010).

In contrast to the sober-mindedness of the Americans, Tocqueville argued that the French Revolution was shaped by 'men of letters' – the term 'intellectuals' had to await the Dreyfus Affair of the late nineteenth century – who were utterly innocent of practical politics. Tocqueville's analysis of the reasons for the impractical bent of these thinkers' ideas was straightforwardly sociological: "The very situation of these writers prepared them to like general and abstract theories of government and to trust in them blindly. At the almost infinite distance from practice in which they lived, no experience tempered the ardors of their nature; nothing warned them of the obstacles that existing facts might place before even the most desirable reforms; they did not have any idea of the dangers which always accompany even the most necessary revolutions" (de Tocqueville, 1998: 197). In his critique of the abstract ideas animating the French Revolution, Tocqueville advanced a line of conservative criticism of intellectuals that would resonate through the works of later authors such as Joseph Schumpeter (1942) and the American neoconservatives of the last third of the twentieth century. This perspective took intellectuals to task for the abstractness of their thought and its divorce from practical life. The underlying assumptions are Burkean views about the historical 'rootedness' of ways of life

and their refractoriness vis-à-vis 'social engineering' of a rationalist kind.

The influence of Burke can scarcely be underestimated in later conservative critique, and not only there. In his *Reflections on the Revolution in France*, first published in 1790, Burke (1968) denounced the notion of the 'rights of man' as a will-o'-the-wisp that could not, in fact, be guaranteed other than by a constituted political community. Hannah Arendt (1973) would affirm that view more than 175 years later in the aftermath of World War II, and indeed the problem of enforceability remains a conundrum for the idea of 'human rights' to this day. In his diatribe against the revolution, Burke also articulated a conception of social order that would animate some of the central dichotomies in the modern social sciences: *Gemeinschaft* versus *Gesellschaft*, substantive reason versus instrumental reason, personal loyalty versus interest, etc. Tocqueville, Marx, Weber, and many others shared the view that a society of deeply shared understandings was giving way to one dominated by the 'cold skeleton hands' of rational calculation, and that the French Revolution had helped midwife its birth.

Like Tocqueville himself, his countryman Emile Durkheim had virtually nothing to say about the American Revolution. Yet Durkheim's ideas are unthinkable without the French Revolution, of which he was profoundly an heir. His social thought is shot through with the 'principles of 1789,' and he strongly endorsed both the idea of human rights and a qualified 'individualism' that respects not a self-interested 'egoism' but the shared humanity in every person. Despite the fact that Durkheim would become thought of as the chief proponent of the notion of society as essentially a religious entity, he viewed the new 'religion of humanity' in the following stark terms: "man has become a god for man and … he can no longer create other gods without lying to himself" (Bellah, 1973: 52). Meanwhile, Max Weber said virtually nothing about either the American or the French Revolution, presumably because of his pessimistic view that 'democracy' would be accompanied inevitably by greater bureaucratization – even if the extent of bureaucratization might vary in the European and American contexts due to their divergent historical trajectories (see Weber, 1946). Insofar as Weber discussed either upheaval, it is only the French that gets passing mention. As one might see in his discussion of 'Politics as a Vocation,' Weber tended to regard (modern) revolutions as romantic, sterile excitations that fail to measure up to the 'demands of the day.'

Because of the relative marginality of the American experience in European and radical thought, the French revolutionary trajectory and its Enlightenment presuppositions came to be regarded as normative. Even Talcott Parsons (1977), frequently denounced as an uncritical apologist for the American way of life, largely ignored the American Revolution in favor of the French in his discussion of the revolutions that constituted modernity. For Parsons, the revolutions that made the modern world up to the 1970s when he wrote were the democratic, the industrial, and the educational. Together these major social developments comprised the elements of modern citizenship heralded above all by the French Revolution, which, Parsons argued, enshrined equality as a central principle of modern life. In an economy that generates economic inequality, however, equality could only be achieved in a qualified way, through the institution of citizenship. As dissected by the British sociologist T.H. Marshall shortly after World War II, citizenship comprised three principal elements – civil, political, and social. The 'civil' element corresponds to the idea of equality before the law, a key principle associated with the French Revolution and its abolition of feudalism. Next, the 'political' element was embodied in the right to vote, a right that was, however, only gradually extended to many groups (such as women and blacks) during the course of the nineteenth and twentieth centuries. Finally, the 'social' element was associated with the institutions of education and social welfare provision. Taken together, these aspects of citizenship were intended to insure "a share in the social heritage, which in turn means a claim to be accepted as full members of the society, that is, as citizens" (Marshall, 1965: 76).

In keeping with the broader pattern described here, the American Revolution has played little role in modern theories of revolution. Theda Skocpol (1979) stressed the importance of state breakdown in the 'social revolutions' of the modern world, namely the French, Russian, and Chinese. The American Revolution has not, as a general rule, been deemed to measure up to this standard. Ironically, this lack of interest in the American Revolution on the part of comparative scholars of revolution ignores the fact that at least one of the leading revolutionaries of the twentieth century, Ho Chi Minh, drew inspiration in his "Proclamation of the Independence of the Democratic Republic of Vietnam" from the American Declaration of Independence as well as the French Declaration of the Rights of Man.

In contrast, those interested in the problem of 'American exceptionalism,' i.e., the liberal and nonsocialist bent of American history, the lack of a labor party, etc. have been more interested in the American Revolution. Comparing the postcolonial states of the former British Empire (the United States, Canada, Australia, and South Africa) as well as of Latin America, Louis Hartz (1964) argued that these were 'fragment cultures' with a relatively muted sense of class antagonism born of the lack of a settled aristocracy against which to rebel. Yet Lipset argued in his comparative study of the United States and Canada (1990: 9–13) that the American Revolution had been a true revolution, cutting off the fledgling United States from the aristocratic traditions and antagonisms characteristic of the mother country and sending proportionately far more loyalist emigres to Canada than fled revolutionary France. The result, according to Lipset, was the creation of a distinctive political culture with similarities to but also notable divergences from other former British possessions.

Nonetheless, as Arendt had argued, the American Revolution until recently remained a matter of marginal significance to sociologists and revolutionaries, if not necessarily to political scientists and historians, and was long trumped by the French Revolution and its legacy. Insofar as the modern social sciences have inherited a 'progressive' cast from their nineteenth-century progenitors, they have tended to regard religion as superstition or worse, or to simply ignore it as a source of idolatry and mischief. This misunderstanding was based in part on the neglect of the American version of modernization, misconstrued the place of religion in the modern world, and helped lead to deep blind spots in the social sciences with regard to the trajectory of modern history. The worldwide religious resurgence identified in many quarters since the

Iranian revolution of 1979 was only the most important development that fell off the radar of the secular social sciences as a result of the processes Tocqueville had analyzed in his comparison of the French and American trajectories.

Things have changed in this regard since the late twentieth century, when the prominent French historian François Furet declared the French Revolution to be 'over' (1981: 1). This was a matter of some moment, "for the French revolution is not only the Republic. It is also an unlimited promise of equality and a special form of change. One only has to see in it not a national institution but a matrix of universal history, in order to recapture its dynamic force and its fascinating appeal" (Furet, 1981: 5). As Furet had foreseen, this 'special form of change' was disappearing from the historical stage. Soon thereafter, with the collapse of Soviet-style communism, the American Revolution and its legacy came dramatically to the fore, nullifying Hannah Arendt's earlier judgment of its relative obscurity. Furet (1999) would write the autopsy of the successor revolutions inspired by the French, and German social philosopher Jürgen Habermas declared after the fall of the Berlin Wall that the Eastern Europeans were undergoing 'catching-up revolutions' that drew much of their inspiration from the American constitutional experiment. Rather than seeking to 'reinvent the wheel' when it came to constructing new political institutions, the architects of the new regimes (other than in East Germany) were inclined simply to turn to the West for models that were thought tried and true.

In recent years, the Haitian Revolution (starting 1791) has been added to the 'founding revolutions of the modern age' by scholars who regard the focus on the United States and France as 'Eurocentric' or racist in character, or as simply missing one of the chief developments of the period. Notwithstanding the fact that the Haitian Revolution was the greatest slave rebellion in world history and the only one to result in the creation of a state, the subsequent unhappy experience of Haiti – long the poorest country in the Western hemisphere – has not helped cement its status as part of the 'age of the democratic revolution,' however deserved this might otherwise be (see Dubois, 2012).

Taking the American and French revolutionary experiences as paradigmatic, the British sociologist of religion David Martin (1990) has suggestively framed the current period as a contest between the 'trajectory of 1776' and the 'trajectory of 1789.' By this he means a struggle between voluntary religion and its sociological correlates – antistatism, religious pluralism and tolerance, and a stress on personal initiative – and the successor-states of old regimes with established religions and *their* correlates – statism, anticlericalism, and secularism. There would seem to be much to this diagnosis if one considers the contemporary encounter between top-down, statist versions of rule in, say, China, and the efflorescence of voluntaristic Protestantism around the world, particularly in Latin America, Africa, and Asia (including within China itself).

Whether the American Revolution and its antistatist, voluntarist thrust remains the dominant force in the world is now a matter of much debate. As the 'sole remaining superpower,' the United States came to be widely viewed in the post-Cold War period as an imperial hegemon whose military superiority made it unchallengeable – the new Rome. In the aftermath of the terrorist attacks of 11 September 2001, the Bush administration's military adventures in Iraq and Afghanistan inflamed world public opinion against American policies. Since the economic crisis of 2008, brought on by rampant financial speculation and unfunded military expenditures arising from those wars, the United States appears increasingly constrained by debt and dysfunctional political institutions.

Still, whereas in earlier times, evangelical Protestantism in Latin America was sometimes seen as part of a larger, pro-American plot (Martin, 1990), in the meantime it has made very substantial inroads, such that some 25% of Brazilians now identify themselves as evangelicals in a country once almost uniformly Roman Catholic. Similar trends are afoot elsewhere in the poorer parts of the world (Martin, 2002), suggesting that the American style of religion, at least, is making a stealth invasion of many parts of the world. As compared to the lack of government services so frequently encountered in poorer countries, the self-help mode typically associated with the American experience – also important in countries where Islamists have had a prominent role in public life (e.g., Lebanon, Palestine, Turkey, Egypt, etc.) – has had considerable appeal. While state-led development produced an 'Asian miracle' among the 'four Asian tigers' (South Korea, Taiwan, Singapore, and Hong Kong), it is not clear that this sort of growth can be replicated elsewhere. Instead, the so-called Islamic Calvinists are rejuvenating both the economy and the religion of Turkey.

The changes that have taken place in the world since Furet declared the French Revolution 'over' have gradually led to a transformed view of the two 'founding' revolutions of the modern world, or at least of the characteristic aspects of each. Religion is back on the social scientific agenda, even if this had a great deal to do with the shock of '9/11.' The modern social sciences have been profoundly shaped by the perception of the French Revolution as the dominant political event in the creation of the modern world. Yet the events of the past 25 years have given new luster to features of the American Revolution that often seemed less substantial in comparison to the high drama of the French, and this is increasingly affecting the way social scientists think about the world in which we live.

See also: American Revolution, The; Arendt, Hannah (1906–75); Civil Liberties and Human Rights; French Revolution, The; Human Rights, History of; Marx, Karl (1818–83); Revolution, Theories of; Revolution; Revolutions, History of; Revolutions, Sociology of; Tocqueville, Alexis de (1805–59).

Bibliography

Arendt, Hannah, 1958. The Human Condition. University of Chicago Press, Chicago.
Arendt, Hannah, 1963. On Revolution. Penguin, New York.
Arendt, Hannah, 1973. The Origins of Totalitarianism, new ed. Harcourt Brace, New York.
Bellah, Robert N., 1973. Individualism and the intellectuals. In: Emile Durkheim on Morality and Society. University of Chicago Press, Chicago, pp. 43–57.
Burke, Edmund, 1968[1790]. Reflections on the Revolution in France. Penguin, New York.
Dubois, Laurent, 2012. Haiti: The Aftershocks of History. Metropolitan Books, New York.
Furet, François, 1981[1978]. Interpreting the French Revolution. Cambridge University Press, New York.

Furet, François, 1999[1995]. The Passing of an Illusion: The Idea of Communism in the Twentieth Century (D. Furet, Trans.). University of Chicago Press, Chicago.
Habermas, Jürgen, 1990. Nachholende Revolution und linker Revisionsbedarf: was heisst Sozialismus heute? In: Die Nachholende Revolution: Kleine Politische Schriften VII. Suhrkamp, Frankfurt am Main, pp. 179–204.
Hartz, Louis, 1955. The Liberal Tradition in America. Harcourt Brace Jovanovich, New York.
Hartz, Louis, 1964. The Founding of New Societies: Studies in the History of the United States, Latin America, South Africa, Canada, and Australia. Harcourt Brace Jovanovich, New York.
Huizinga, Johan, 1972[1919/1926]. America: A Dutch Historian's Vision, from Afar and Near (H.H. Rowen, Trans.). Harper & Row, New York.
Lipset, Seymour Martin, 1990. Continental Divide: The Values and Institutions of the United States and Canada. Routledge, New York.
Lipset, Seymour Martin, 1996. American Exceptionalism: A Double-Edged Sword. Norton, New York.
Marshall, T.H., 1965[1949]. Citizenship and social class. In: Marshall, T.H. (Ed.), Class, Citizenship, and Social Development. Anchor Doubleday, Garden City, NY, pp. 71–134.
Martin, David, 1990. Tongues of Fire: The Explosion of Protestantism in Latin America. Blackwell, Cambridge, MA.
Martin, David, 2002. Pentecostalism: The World Their Parish. Blackwell, Malden, MA.
Marx, Karl, 1978a[1844]. Contribution to the critique of Hegel's philosophy of right: introduction. In: Tucker, Robert C. (Ed.), The Marx-Engels Reader, second ed. Norton, New York, pp. 53–65.
Marx, Karl, 1978b[1843]. On the Jewish question. In: Tucker, Robert C. (Ed.), The Marx-Engels Reader, second ed. Norton, New York, pp. 26–52.
Palmer, R.R., 1959–1964. The Age of the Democratic Revolution: A Political History of Europe and America, 1760–1800. Princeton University Press, Princeton, NJ.
Parsons, Talcott, 1977. The Evolution of Societies. Prentice Hall, Englewood Cliffs, NJ.
Schumpeter, Joseph, 1942. Capitalism, Socialism, and Democracy. Harper & Brothers, New York.
Skocpol, Theda, 1979. States and Social Revolutions: A Comparative Analysis of France, Russia, and China. Cambridge University Press, New York.
Sombart, Werner, 1976[1905]. Why Is There No Socialism in the United States? M. E. Sharpe, White Plains, NY.
de Tocqueville, Alexis, 1998[1856]. The Old Regime and the Revolution (A.S. Kahan, Trans.). University of Chicago Press, Chicago.
de Tocqueville, Alexis, 2000[1835/1840]. Democracy in America. Translated and Edited by H.C. Mansfield and D. Winthrop. University of Chicago Press, Chicago.
Torpey, John, Spring 2010. A (post-) secular age? Religion and the two exceptionalisms. Social Research 77 (1), 269–296.
Weber, Max, 1946. Bureaucracy. In: Gerth, H.H., Wright Mills, C. (Eds.), From Max Weber: Essays in Sociology. Oxford University Press, New York, pp. 196–244.
Zolberg, Aristide, 2008[1972]. Moments of madness. In Zolberg, Aristide, How Many Exceptionalisms? Explorations in Comparative Macroanalysis. Temple University Press, Philadelphia, pp. 18–42.

The American Conservative Movement

Ann M Horwitz, University of Maryland, College Park, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Initially an intellectual tradition stemming from the late 1700s, conservatism in America has grown into a veritable social movement, comprising a complex network of activists, elected officials, lobbying groups, financial backers, and media outlets. This article traces American conservatism from its roots in the intellectual realm to its present-day significance as a social movement and political force. It describes the three main strands of conservative thought – traditionalism, libertarianism, and anticommunism – citing the major thinkers within each strand. To illustrate its development into a social movement, the article presents four watershed moments for American conservatism – Barry Goldwater's 1964 presidential campaign, Ronald Reagan's presidency, the 1994 Republican takeover of Congress, and most recently, the advent of the Tea Party in 2009.

Introduction

Conservatism is one of America's great political traditions, whose origin stretches back to the reaction to the Enlightenment (Mannheim, 1986). Some of its fundamental principles comprise an integral part of American governance and inform the country's legal and economic systems, and how young Americans are prepared for citizenship in a democracy. Conservatism – a "tendency to preserve what is established" (Nisbet, 1952: p. 167) – has played, and continues to play, a formative role in American society both on a macrolevel (e.g., the promotion of free market capitalism) and on a more microlevel (e.g., the individual's right to bear arms). The conservative preference for limited government is one of the most enduring political themes in the nation's history (Nash, 1976). Given its vital place in the American fabric, inquiring into the nature and dynamics of American conservatism is an important academic endeavor.

With roots in the intellectual realm (Nash, 1976), starting with the reaction to the Enlightenment, American conservatism has evolved since Barry Goldwater's 1964 presidential campaign into a true social movement. Its adherents have succeeded in establishing a sphere of influence within the formal political structure (i.e., the Republican Party), and have been successful in mobilizing resources to further conservative aims (Bjerre-Poulsen, 2002). The time of this writing, in 2014, is an opportune one for examining American conservatism as the United States is embroiled in one of its most politically polarized eras in memory. The country has still not recovered from the economic crisis of 2008, and the two major political parties, of course, express quite different ideas about how to solve the problem. With power in Congress split – a Republican-controlled House and a Democratic-controlled Senate – compromise has been elusive. Multiple media spectacles over the budget and the debt ceiling have come and gone, and in October 2013, the federal government shut down for 16 days (Dockterman, 2013). In late 2013, President Obama described the current mode of American politics as "lurching from crisis to crisis to crisis" (Taintor, 2013).

Part of this impasse may be due to increased polarization between liberals and conservatives (see Abramowitz and Saunders, 2008; DiMaggio et al., 1996; Evans, 2003; Zschirnt, 2011). Some scholars, however, argue that the more consequential schism is the widening gap between moderate Republicans and more hard-line conservatives (see Barreto et al., 2011; Boykoff and Laschever, 2011; Maxwell and Parent, 2012). For these researchers, the advent of the Tea Party in the wake of Barack Obama's election in 2008 sparked this divide and heralded a resurgence of conservative activism that represents fertile ground for social scientists' inquiries (Perrin et al., 2011; Skocpol and Williamson, 2012; Zernike, 2010).

This article divides the American conservative movement into two spheres: conservatism as an intellectual movement and conservatism as a social movement. It presents the intellectual side of the movement and its three major strands: traditionalism, libertarianism, and anticommunism. It then transitions into an exploration of four watershed moments of the movement in the twentieth century, which gave rise to an activist social conservatism: Barry Goldwater's 1964 presidential campaign, Ronald Reagan's presidency, the Republican takeover of Congress in 1994, and the present-day emergence of the Tea Party.

Conservatism as an Intellectual Movement

An initial question is, 'What exactly are conservatives trying to conserve?' There may not be one single, all-encompassing answer to this question, as the conservative movement is somewhat fragmented (Johnson et al., 1986; Zumbrunnen and Gangl, 2008). This article will explain more about this fragmentation and various factions within American conservatism later, but first, it is useful to establish some of the root principles of conservatism that undergird the movement as a whole.

As Shannon explains, "Conservatism as a distinct political term is [180] years of age, coming into general usage after the Great Reform Act of 1832 in Great Britain. The word meant opposition to reform. In general, conservatives wish to preserve present or past values rather than to create or adopt new ones" (1962: p. 14). Nisbet adds that conservatism is the "disposition and tendency to preserve what is established … tending to maintain existing institutions or views" (1952: p. 167). Within the American context, this orientation translates into a reverence for the Constitution and its enshrinement of private property.

After the American colonies gained independence from Great Britain, "In the absence of a crown and nobility,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64088-6

conservatism has centered around the Constitution and the institution of private property" (Shannon, 1962: p. 13). For conservatives, the view of the Constitution as a 'compact' is central: "The nation [is] a union of States, sovereign States, whose powers were confirmed by the Tenth Amendment" (Nash, 1976: pp. 200–201; see also Adams, 1984). The emphasis on states' rights is important because it showcases one of American conservatism's key tenets, namely, a preference for limited central government. This principle is articulated passionately by Barry Goldwater (discussed below): "The founding fathers had a reason for endorsing the principle of limited government … Throughout history, government has proved to be the chief instrument for thwarting man's liberty. Government represents power in the hands of some men to control and regulate the lives of other men" (Goldwater, 2003: pp. 214–215). Conservatives favor a relatively strict interpretation of the Constitution, as they distrust judges or other agents of government who might extrapolate from the document liberties for the government that the framers did not intend it to have; such abiding distrust of government is central to conservatism (Nock, 2003).

Conservatives do see a role for government. Its purpose is to safeguard individual rights, specifically the right to private property (Buckley, 2003; Smith, 2007; Weaver, 2009). As Smith explains, "Government responsibilities should be limited to defining and protecting property rights, enforcing contracts, upholding law and order, and correcting for monopoly power and other market imperfections" (2007: p. 101). Buckley adds, "It is the job of centralized government (in peace-time) to protect its citizens' lives, liberty, and property" (2003: p. 203). To conservatives, any conception of the role of government that goes beyond these provisions is a misreading of the Constitution and a perversion of the framers' intent.

Returning to Shannon's assertion that conservatism "wish [es] to preserve present or past values rather than to create or adopt new ones," American conservatives' reverence for the Constitution is interesting in light of the fact that that very document was, in many ways, a radical break from the past (1962: p. 14). In response, a number of conservative thinkers argue that the American Revolution and the Constitution were, in fact, a conservative reaction to a British regime that attempted to impose a nontraditional order (see Kirk, 2003). For American conservatives, in any case, there is something exceptional in the Constitution that is worth preserving (it should be noted that the idea of American exceptionalism was introduced by the Frenchman Alexis de Tocqueville; see Lipset, 1988). To them, the document's reining in of government and privileging of individual property rights separates the United States from other democracies.

Although the principles of constitutional fealty and respect for private property rights underlie the conservative intellectual tradition writ large, that tradition is not a unitary one. Three separate strands of thought comprise the conservative intellectual movement in America: traditionalism, libertarianism, and anticommunism (Schneider, 2003).

Traditionalism

An abiding faith in the exceptionalism of the Constitution notwithstanding, the traditionalist strand of the conservative intellectual movement takes a number of its cues from European thinkers, most notably Edmund Burke (Nash, 1976). Burke's famous *Reflections on the Revolution in France* mounted a frontal assault on what he saw as the corrupting influences of rationalism and liberalism embodied in the French Revolution (Mannheim, 1986). The Burkean mantle was assumed by Russell Kirk, perhaps the most influential American traditionalist thinker.

As Kirk explains, traditionalism posits that there is an absolute moral order, primordial, that must be preserved. Traditionalism eschews any type of moral relativism. "The essence of [conservatism] is preservation of the ancient moral traditions of humanity. Conservatives respect the wisdom of their ancestors … A divine intent rules society as well as conscience, forging an eternal chain of right and duty which links great and obscure, living and dead" (Kirk, 2003: p. 110). In defense of this moral order, traditionalism is inherently reactive and counterrevolutionary, 'Clinging to old ways' (Mannheim, 1986: p. 83).

Importantly, traditionalists do not see government as the ideal guarantor of the traditional moral order's survival. Rather, they focus on nongovernmental institutions such as the family, the religious community, or even an aristocratic order (per Burke) as the appropriate guardians of ancient tradition and order (Schneider, 2003). In terms of contemporary conservative activism, these tenets have found voice in the rhetoric of 'family values' religious conservatives (see Peele, 1984).

Although traditionalism is an intellectual philosophy that romanticizes an ancient past, and conceives of present and past generations as bound together in something of a collective, it is unmistakably conservative in its defense of property rights: "Property and freedom are inseparably connected … Separate property from private possession, and liberty is erased" (Kirk, 2003: p. 111). Despite this clear endorsement of individual property rights, traditionalism concerns itself more with moral order than with economic mores. That distinction separates traditionalism with the next intellectual strand of the conservative movement: libertarianism.

Libertarianism

Aside from support for limited government and for private property rights, traditionalism and libertarianism share little in common. In fact, in some ways, libertarians directly repudiate traditionalists' worship of the past. Where traditionalists mistrust the Enlightenment's privileging of reason, libertarians exalt it. "Liberty … is our thesis; reason shall be our method. Only reason can build a valid and consistent ideology, and only reason can strip unsparingly the veil of custom and habit and myth from dominant ideas and institutions and hold them up to the harsh light of truth" (Left and Right, 2003: p. 251). In traditionalism, libertarians see an overreliance on authority (not necessarily only government authority, but perhaps church authority), and a misplaced focus on moral control over individual autonomy: "Unmodified traditionalism, stressing virtue and order in disregard of the ontological and social status of the freedom of the individual person, tend[s] dangerously toward an authoritarianism wrong in itself" (Meyer, 2003a: p. 259). This disagreement with traditionalists

leads some libertarians to reject the conservative label for themselves (see Hayek, 2003).

While much philosophical distance exists between traditionalism and libertarianism vis-à-vis their perspectives and emphases, the latter's preoccupation with limited government and free markets leads most scholars to include libertarianism under the umbrella of the American conservative intellectual movement (Nash, 1976; Schneider, 2003). It is worth noting, however, that whereas traditional conservatism is associated with counterrevolution and counterreformism, a number of libertarians embrace a vision of themselves as 'radicals': "The genuine libertarian … is, in all senses of the word, an 'abolitionist'; he would, if he could, abolish instantaneously all invasions of liberty: whether it be, in the original coining of the term, slavery, or it be the manifold other instances of State oppression" (Rothbard, 2003a: p. 256). Indeed, this has led some conservatives to caution that "untrammeled libertarianism … tends as directly to anarchy and nihilism as unchecked traditionalism tends to authoritarianism" (Meyer, 2003b: p. 259).

In response to that criticism, Rothbard defines libertarianism as a perspective that "emphasiz[es] the importance of the individual, of his freedom of will, of natural rights and private property," that in no way precludes religious beliefs, and that most certainly encompasses moral principles (2003b: p. 266). This debate between Rothbard and Meyer points to the idea of conservative 'fusionism,' which the literature treats not as a true strand of the conservative intellectual movement, but as something of a substrand within it (Adler, 2004; Nash, 1976). Frank Meyer is considered the father of conservative fusionism, an attempt to marry traditionalism and libertarianism in some coherent way. In his mind, the fusionist school of thought could be the unique contribution of American conservative intellectuals to the field: "It is my aim to make this demonstration, to vindicate on theoretical grounds the native belief of American conservatives that freedom as a prime criterion in the political and social sphere is not alien to the conservative view of man's nature and destiny" (2003a: p. 174). Ultimately, "While concerned with maintenance of an objective moral order and the pursuit of virtue in the individual, Meyer argued that the freedom of the person is the central and primary end of political society" (Adler, 2004: p. 51). In the end, the fusionist label has not endured as much as the libertarian label. Libertarianism, indeed, has a robust place within the conservative intellectual movement.

Before ending the discussion of libertarianism, there is a brief mention of free markets, as this topic sharply separates libertarianism from traditionalism. One of the central libertarian tracts, F.A. Hayek's *The Road to Serfdom* (1944), posits the importance of free enterprise and economic competition to ensuring individual freedom and liberty. Another libertarian standard bearer, Milton Friedman, argues in *Capitalism and Freedom* that political liberty cannot be obtained without economic liberty (1982; see also Smith, 2007).

In short, free market ideology is inseparable from libertarianism. Arguments like Hayek's and Friedman's are often cast – implicitly or explicitly – against alternative economic arrangements. For example, both men wrote during the cold war, when free markets vs central planning was a quite salient political issue. This context brings us to the third, and final, strand of the conservative intellectual movement: anticommunism.

Anticommunism

Although fusionism, as discussed above, comprises an uneasy, tenuous marriage of traditionalists and libertarians, it turns out that the threat of communism had a significant impact on merging the two strands, at least temporarily. Because it "destroyed traditional social order – family, religion, and community – and established the State, under the banner of the communist party, as the final arbiter of all decisions affecting individuals," communism presented a mortal threat in the eyes of both traditionalists and libertarians, giving them a common enemy (Schneider, 2003: p. 131).

According to Nash (1976), conservative intellectuals' responses to communism touched both on perceived threats from abroad (e.g., the policy of Soviet containment) and threats from within (e.g., McCarthyism) (see also Anderson et al., 1965). Perhaps the most notable aspect of the anticommunist intellectual strand is that a number of its central figures were themselves "ex-communists [who] proved their disenchantment with 'the god that failed' by dedicating themselves wholeheartedly to expose the tyranny and evil of communist regimes" (Schneider, 2003: p. 132).

At present, although the cold war is over, the intellectual impulses of anticommunism remain alive within the conservative movement. Indeed, Schneider dubs the same excommunist thinkers who were active in the strand the first generation of 'neoconservatives' (Schneider, 2003: p. 132). Their support of strong national defense and interventionist foreign policy is evident today in the neoconservative faction of the movement.

William F. Buckley, Jr.

A number of key intellectual figures in conservatism have been introduced in the preceding pages. It is worthwhile to touch briefly on the work of at least one other, William F. Buckley, Jr., who founded *National Review* in 1955. The publication remains one of the most influential clearinghouses for conservative thought to this day. Buckley employed and published conservative thinkers from across the spectrum, not necessarily achieving intellectual fusion à la Meyer, but certainly promoting a wider readership among conservatives. It is difficult to overstate the magazine's – and, by extension, Buckley's – contribution to American conservatism (Hart, 2005). "If *National Review* (or something like it), had not been founded, there would probably have been no cohesive intellectual force on the Right in the 1960s and 1970s" (Nash, 1976: p. 153). Incidentally, it was during that very same era that conservatism as a veritable social movement emerged.

Conservatism as a Social Movement

The midtwentieth century saw the transition of American conservatism from an intellectual movement to a social

movement. Where before there had been thinkers and writers, now there were demonstrators and activists. Of course, it would be simplistic to say that the two are mutually exclusive, and that there were no conservative activists prior to the 1960s. Even so, the literature shows a clear shift in that decade toward conservative political activism (Himmelstein, 1990; Pierson and Skocpol, 2007). Four watershed moments for 'mobilizing resentment' toward government power (Hardisty, 1999; see also Crawford, 1980) are evident from the literature: the 1964 presidential campaign of Barry Goldwater, Ronald Reagan's presidency, the Republican takeover of Congress in 1994, and the emergence of the Tea Party in 2009.

Goldwater 1964

Barry Goldwater was a senator from Arizona when he won the Republican nomination for president in 1964. At the time, the post–Eisenhower Republican Party establishment was moderate and seemed unconcerned about overreaches of government power (Andrew, 1997; Bjerre-Poulsen, 2002). This establishment was shocked when Goldwater easily beat New York Governor Nelson Rockefeller for the nomination. Goldwater was an unabashed conservative, directly attacking the notion that government power was an effective mechanism for solving social problems (Middendorf, 2006). Although he went on to be roundly defeated by Lyndon Johnson in the general election, Goldwater's candidacy mobilized passionate support among swaths of voters. A "Republican in the new Western mode, he appealed to voters who resented the intrusion of big government, especially the federal government, into their personal and economic lives" (Pierson and Skocpol, 2007: p. 96). Moreover, Goldwater opposed federal desegregation efforts, which won him support among many voters in Southern states that had previously been Democratic strongholds (Pierson and Skocpol, 2007: p. 96).

Goldwater was also effective at mobilizing young voters. While much of the scholarship on young activists in the 1960s focuses on the left-wing Students for a Democratic Society, the conservative Young Americans for Freedom (YAF) was a major movement in its own right. Although YAF predated Goldwater's nomination, his campaign galvanized young conservatives in an unprecedented way. "Through Goldwater thousands of youth entered politics. The campaign transformed the uninitiated into activists and united traditionalists and libertarians in common cause" (Klatch, 1999: p. 84). Although he did not win the general election, scholars credit Goldwater with injecting a strong conservative strain into the Republican Party (Andrew, 1997; Shermer, 2013). This development manifested its significance over time, its political force becoming strikingly evident when another conservative, Ronald Reagan, ran for president, and won.

The Reagan Era

The conservative writer George Will said, "Goldwater won the election of 1964. It just took 16 years to count the votes" (see Middendorf, 2006). This statement reflects the view that Ronald Reagan, who was first elected to the presidency in 1980, was the second coming of Barry Goldwater. Such writers argued that Reagan's candidacy reinvigorated the same voters who had been energized in 1964:

> They were white lower to middle strata voters who switched to the Republican Party on the basis of a profound "social conservatism"—manifested in opposition to abortion and the Equal Rights Amendment and in general concern over the decay of morality, family, and religion. These voters were more religious than the average American and more likely to be alienated from and angry at a federal government that they viewed as the source of secular, liberal values and programs.
>
> Himmelstein and McRae, 1984: p. 592

It is certainly true that Reagan was a conservative; in fact, Schoenwald credits him with embodying a fusion of anti-communism, traditionalism, and libertarianism 'in his own person,' which was not an easy feat for a candidate who began his adult life as a New Deal Democrat (2001: p. 197). His electoral success marked an important moment for the conservative social movement, as it "brought national attention to a new political phenomenon, often labeled 'The New Right': a network of right-wing political organizations and single-issue groups committed to the creation of a conservative 'counterestablishment' that could outdo liberals in grassroots organization, lobbying, fund-raising, and self-promotion" (Bjerre-Poulsen, 1991: pp. 152–153; see also Jacobs, 2006; Lowndes, 2008).

To what extent the presence of a charismatic standard bearer like Reagan at its center contributed to its success is not possible to say, but there are some indicators that the New Right did manage to pull together disparate strands of conservatism. For example, both moderate and ultraconservatives in the business community came together during the Reagan era (Jenkins and Eckert, 2000). Conservative Protestants were brought into the fold and played an integral role in conservative activism (Peele, 1984; Woodberry and Smith, 1998). In terms of attitudes and public opinion, conservative self-identification increased in the 1980s, including among young voters (McBroom and Reed, 1990; Miller, 1992). Additionally, from 1974 to 2012, over the course of many decades as a social movement, conservatives went from having the highest trust in science (relative to liberals and moderates) to having the lowest, indicating the New Right's success in melding fundamentalist religious views into its coalition (Gauchat, 2012).

At the close of Reagan's tenure, though, "Satisfaction with the economic situation in the fall of 1988 was much more important than moral conservatism in [George H.W.] Bush's election win" (Johnson et al., 1991: p. 303). Bill Clinton, the Democrat, would sweep to victory by emphasizing the economy in 1992. A mere 2 years after his election win, though, Clinton would be confronted by a third strong wave of the American conservative movement as Republicans took back control of Congress in the midterm elections of 1994.

The Republican Revolution of 1994

If Bill Clinton's electoral success in 1992 deflated the conservative movement, the effect did not last long. In the lead-up to the 1994 midterm elections, Georgia Congressman Newt

Gingrich introduced the 'Contract with America,' a 10-point platform, which "represented a type of 'conservatism-lite', aimed at electing a conservative Congress, but one which would not depart too far from the vital center of political debate" (Schneider, 2003: p. 397). In the Contract, Gingrich generally shied away from contentious cultural and social issues (welfare and criminal justice system reform being partial exceptions), focusing instead on issues of fiscal policy and national security (Gingrich, 2003). Like Goldwater and Reagan before him, Gingrich appealed to conservatives' distrust of central government and to their notions of economic liberty.

The Contract "not only helped to unify Republican candidates around a sweeping set of policy proposals but also energized some of the party's key constituencies for the fall campaign" (Balz and Brownstein, 1996: p. 37). The Contract had the desired effect. Conservative activists mobilized and voters delivered Republicans a historic victory: "For the first time since 1954, the Republicans captured control of both houses of Congress. Gingrich was speaker of the House and Republicans controlled both houses of Congress the remainder of Clinton's presidency" (Schneider, 2003: p. 397). The next time that the Republican Party would enjoy such a momentous mid-term election would be in 2010, when a new movement faction, the Tea Party, would emerge to galvanize conservative activists.

The Tea Party

In early 2009, Washington, DC, had barely finished removing the tickertape from Barack Obama's inauguration before the Tea Party started making headlines. The opportune moment presented by a new, Democratic president (and one viewed, rightly or wrongly, as particularly liberal) with a majority in both houses of Congress stoked fear and anger among a set of Americans, many of whom had not been politically active before, who wanted to "take [their] country back" (Skocpol and Williamson, 2012). Obama's election not only pointed to social and demographic shifts within the country, but also came at a time of economic trauma.

The circumstances surrounding the change in power in 2009 represented an opening for conservative backlash. The economic collapse of 2008, the bank bailouts that ensued under George W. Bush, and the economic stimulus enacted under Obama, occurring in tandem with major job losses and home foreclosures, roiled the discourse around the country's economic future. Moreover, the electoral success of a Black man signaled, in some Americans' minds, a shift in business as usual in the country's social and political life (Boykoff and Laschever, 2011; Pease, 2010). This unrest provided an opportunity for conservatives to mobilize under the Tea Party banner.

Part of the challenge of studying the Tea Party is that it is amorphous. "By philosophy and design, [the Tea Party] lacks a central leadership structure that coordinates nationwide efforts. Instead, the Tea Party is a far-flung patchwork of organizations, some local and some national, with a related set of issue concerns and positions" (Karpowitz et al., 2011: p. 303). Thus, when discussing the Tea Party, it is necessary to specify whether one is referring to the local activists who attend meetings and protests, the professionalized movement organizations that formulate framing and steward financial resources, the media apparatuses that deliver framed messages and mobilize support, or the conglomeration of all three (Skocpol, 2012).

Despite this definitional ambiguity, scholars have gained an understanding of basic philosophical tenets and political dispositions that unite disparate arms of the movement. Foley identifies three central principles of the movement: limited government, US sovereignty, and constitutional originalism (2012). Ashbee points to evidence of "deep pessimism about the country's economy and its economic future" as a common characteristic of Tea Party activists and sympathizers (2011: p. 157). Additionally, scholars find that those in the Tea Party are joined by "an intense personal and political antipathy to President Obama" (2011: p. 157; see also Jacobson, 2011; Maxwell and Parent, 2012). Incidentally, that antipathy, although it has been a boon to the movement in terms of political opportunity and recruitment, has also saddled it with accusations of inherent racism and xenophobia.

Employing qualitative interviews, content analysis, and quantitative analysis of public opinion data, Barreto et al. conclude that the Tea Party, rather than a popular movement to rein in government overreach and taxation, is a vocal manifestation of the 'paranoid style' of extremist racist, anti-immigrant, homophobic American politics (2011: p. 1). Berlet points to demonization and scapegoating of liberals, people of color, and immigrants, along with conspiracy theories about Obama as a foreign-born Muslim, as hallmarks of Tea Party orthodoxy (2011). Interestingly, while concurring with the portrait of the Tea Party as a racist movement, Enck-Wanzer also faults Obama for not responding with forceful antiracist rhetoric (2011). The arguments put forth by these scholars, though, depict an overt and nefarious racism among Tea Party activists, veiled deliberately behind language of small-government principles. While they find evidence of stereotyped racist beliefs among Tea Party members, Skocpol and Williamson also find evidence of activists within the movement pushing back actively against racists who 'infiltrate' it (2012: p. 69). Courser (2011) and Etzioni (2011) highlight the potentially problematic impulse of scholars to write off the Tea Party as a cadre of racist xenophobes, and Skocpol and Williamson's (2012) work points to the need for further research to tease out the racial rhetoric within the movement and ascertain the degree to which racism in the movement drives its agenda.

Wilson and Burack theorize that the Tea Party is an articulation of the economic and social strands of American conservatism, making it instrumental in uniting conservatives at a moment of political opportunity (2012). Conceived of in this way, the Tea Party may be said to represent a successful conservative fusion (Meyer, 2003a). What is unknown at present is the extent to which the Tea Party will continue to enjoy influence in the political system, a question that may only be answered by observing future election cycles.

Conclusion

As a movement, American conservatism has shown itself to be dynamic and resilient. Even when Republicans have been the minority in Congress and absent in the White House, the principles that define conservatism have held considerable

sway in American politics. In recent memory, free market ideology and deregulatory policies have been the norm even under Democrats (Jimmy Carter, for example, has named deregulation as one of his administration's great accomplishments; see Smith, 2005). Moreover, politicians across the spectrum have invoked a preference for limited government when speaking to the public. No less loyal a Democrat than Bill Clinton proclaimed, in his second inaugural address in 1997, "The era of big government is over," presumably referring to the New Deal regime put in place after the Great Depression. Regardless of the party in power, conservative ideals have coursed strongly through the American political vein.

Even so, some in the present-day political class see great challenges ahead for the conservative movement. With the demographics of the country changing, racial and ethnic minorities make up an increasingly large portion of the electorate. This new reality has led some – both inside and outside of the Republican Party – to believe that conservatives must change their stance on issues like immigration in order to remain politically viable. Additionally, some assert that conservative positions on gay rights and reproductive rights are alienating young voters and women, an untenable trend if Republicans wish to win elections. Compounding the uncertainty over conservatism's future is the current turmoil within the Republican Party. Since the rise of the Tea Party, Republican primary elections at the state and federal level have often become intense battles between hard-line conservatives (often backed by the Tea Party) and more moderate candidates, leaving the victor's image and coffers bruised and sometimes giving the Democratic rival the advantage in the general election.

Despite these obstacles, though, any eulogy for American conservatism is premature. American politics is pendulous, but however it swings, conservative articles of faith are consistently in favor, at least in electoral rhetoric. Examining American conservatism going forward presents the opportunity to observe how the movement navigates a rapidly shifting political and social landscape, and to gauge its success. This endeavor engages the academic concerns of multiple disciplines: political science, economics, sociology, and others. American conservatism, both as an intellectual movement and as a social movement, has a profound impact on the political and social structure of American life. It is a dynamic movement, and it deserves the attention of social scientists.

See also: Capitalism: The History of the Concept; Capitalism; Conservatism: Historical Aspects; Democracy, History of; Political History: History of Politics; Political Parties, History of; Religion and Politics in the United States; Social Movements, History of: General; Social Movements: A Social Psychological Perspective; von Hayek, Friedrich A. (1899–1992).

Bibliography

Abramowitz, A.I., Saunders, K.L., 2008. Is polarization a myth? The Journal of Politics 70 (2), 542–555.

Adams, W.P., 1984. Republicanism. In: Greene, J.P. (Ed.), Encyclopedia of American Political History: Studies of the Principal Movements and Ideas, vol. III. Charles Scribner's Sons, New York, pp. 1131–1146.

Adler, J.H., 2004. Frank Meyer: the fusionist as federalist. Publius: The Journal of Federalism 34 (4), 51–67.

Anderson, B., Zelditch Jr., M., Takagi, P., Whiteside, D., 1965. On conservative attitudes. Acta Sociologica 8 (3), 189–204.

Andrew III, J.A., 1997. The Other Side of the Sixties: Young Americans for Freedom and the Rise of Conservative Politics. Rutgers University Press, New Brunswick.

Ashbee, E., 2011. Bewitched – the Tea Party movement: ideas, interests and institutions. The Political Quarterly 82 (2), 157–164.

Balz, D., Brownstein, R., 1996. Storming the Gates: Protest Politics and the Republican Revival. Little, Brown and Company, Boston.

Barreto, M.A., Cooper, B.L., Gonzalez, B., Parker, C.S., Towler, C., 2011. The Tea Party in the age of Obama: mainstream conservatism or out-group anxiety? Political Power and Social Theory 22, 1–29.

Berlet, C., 2011. Taking Tea Parties seriously: corporate globalization, populism, and resentment. Perspectives on Global Development and Technology 10 (1), 11–29.

Bjerre-Poulsen, N., 1991. The heritage foundation: a second-generation think tank. Journal of Policy History 3 (2), 152–172.

Bjerre-Poulsen, N., 2002. Right Face: Organizing the American Conservative Movement. Museum Tusculanum Press, Copenhagen.

Boykoff, J., Laschever, E., 2011. The Tea Party movement, framing, and the US media. Social Movement Studies 10 (4), 341–366.

Buckley Jr., W.F., 2003. National review: credenda and statement of principles. In: Schneider, G.L. (Ed.), Conservatism in America Since 1930: A Reader. New York University Press, New York, pp. 201–205.

Courser, Z., 2011. The Tea 'Party' as a conservative social movement. Society 49 (1), 43–53.

Crawford, A., 1980. Thunder on the Right: The 'New Right' and the Politics of Resentment. Pantheon Books, New York.

DiMaggio, P., Evans, J., Bryson, B., 1996. Have Americans' social attitudes become more polarized? American Journal of Sociology 102 (3), 690–755.

Dockterman, E., 2013. Here's How Much the Government Shutdown Cost the Economy. http://swampland.time.com/2013/10/17/heres-what-the-government-shutdown-cost-the-economy/ (accessed 05.11.13.).

Enck-Wanzer, D., 2011. Barack Obama, the Tea Party, and the threat of race: on racial neoliberalism and born again racism. Communication, Culture & Critique 4, 23–30.

Etzioni, A., 2011. The Tea Party is half right. Society 48, 197–202.

Evans, J.H., 2003. Have Americans' attitudes become more polarized? – An update. Social Science Quarterly 84 (1), 71–90.

Foley, E.P., 2012. The Tea Party: Three Principles. Cambridge University Press, Cambridge.

Friedman, M., 1982. Capitalism and Freedom. University of Chicago Press, Chicago.

Gauchat, G., 2012. Politicization of science in the public sphere: a study of public trust in the United States, 1974 to 2010. American Sociological Review 77 (2), 167–187.

Gingrich, N., 2003. Contract with America. In: Schneider, G.L. (Ed.), Conservatism in America Since 1930: A Reader. New York University Press, New York, pp. 424–427.

Goldwater, B.M., 2003. The conscience of a conservative. In: Schneider, G.L. (Ed.), Conservatism in America Since 1930: A Reader. New York University Press, New York, pp. 211–225.

Hardisty, J., 1999. Mobilizing Resentment: Conservative Resurgence from the John Birch Society to the Promise Keepers. Beacon Press, Boston.

Hart, J., 2005. The Making of the American Conservative Mind: National Review and Its Times. ISI Books, Wilmington.

Hayek, F.A., 1944. The Road to Serfdom. University of Chicago Press, Chicago.

Hayek, F.A., 2003. Why i am not a conservative. In: Schneider, G.L. (Ed.), Conservatism in America Since 1930: A Reader. New York University Press, New York, pp. 180–194.

Himmelstein, J.L., 1990. To the Right: The Transformation of American Conservatism. University of California Press, Berkeley.

Himmelstein, J.L., McRae Jr., J.A., 1984. Social conservatism, new republicans, and the 1980 election. Public Opinion Quarterly 48 (3), 592–605.

Jacobs, A.K., 2006. The new right, fundamentalism, and nationalism in postmodern America: a marriage of heat and passion. Social Compass 53 (3), 357–366.

Jacobson, G.C., 2011. The Republican resurgence in 2010. Political Science Quarterly 126 (1), 27–52.

Jenkins, J.C., Eckert, C.M., 2000. The right turn in economic policy: business elites and the new conservative economics. Sociological Forum 15 (2), 307–338.

Johnson, S.D., Tamney, J.B., Halebsky, S., 1986. Christianity, social traditionalism and economic conservatism. Sociological Focus 19 (3), 299–314.

Johnson, S.D., Tamney, J.B., Burton, R., 1991. Economic satisfaction vs. moral conservatism in the 1988 presidential election. Sociological Focus 24 (4), 303–314.

Karpowitz, C.F., Monson, J.Q., Patterson, K.D., Pope, J.C., April 2011. Tea time in America? The impact of the Tea Party movement on the 2010 midterm elections. PS: Political Science & Politics, 303–309.
Kirk, R., 2003. The conservative mind. In: Schneider, G.L. (Ed.), Conservatism in America Since 1930: A Reader. New York University Press, New York, pp. 107–121.
Klatch, R.E., 1999. A Generation Divided: The New Left, the New Right, and the 1960s. University of California Press, Berkeley.
Left and Right: A Journal of Libertarian Thought, 2003. The libertarian review: editorial statement. In: Schneider, G.L. (Ed.), Conservatism in America Since 1930: A Reader. New York University Press, New York, pp. 251–252.
Lipset, S.M., 1988. American exceptionalism reaffirmed. International Review of Sociology 2 (3), 25–69.
Lowndes, J.E., 2008. From the New Deal to the New Right: Race and Southern Origins of Modern Conservatism. Yale University Press, New Haven.
Mannheim, K., 1986. Conservatism: A Contribution to the Sociology of Knowledge. Routledge & Kegan Paul, New York.
Maxwell, A., Parent, T.W., 2012. The Obama trigger: presidential approval and Tea Party membership. Social Science Quarterly 93 (5), 1384–1401.
McBroom, W.H., Reed, F.W., 1990. Recent trends in conservatism: evidence of non-unitary patterns. Sociological Focus 23 (4), 355–365.
Meyer, F.S., 2003a. A rebel finds his tradition. In: Schneider, G.L. (Ed.), Conservatism in America Since 1930: A Reader. New York University Press, New York, pp. 171–179.
Meyer, F.S., 2003b. Libertarianism or libertinism? In: Schneider, G.L. (Ed.), Conservatism in America Since 1930: A Reader. New York University Press, New York, pp. 259–261.
Middendorf II, J.W., 2006. A Glorious Disaster: Barry Goldwater's Presidential Campaign and the Origins of the Conservative Movement. Basic Books, New York.
Miller, A.S., 1992. Are self-proclaimed conservatives really conservative? Trends in attitudes and self-identification among the young. Social Forces 71 (1), 195–210.
Nash, G.H., 1976. The Conservative Intellectual Movement in America Since 1945. Basic Books, New York.
Nisbet, R.A., 1952. Conservatism and sociology. American Journal of Sociology 58 (2), 167–175.
Nock, A.J., 2003. Radical individualism: the state as enemy. In: Schneider, G.L. (Ed.), Conservatism in America Since 1930: A Reader. New York University Press, New York, pp. 29–44.
Pease, D.E., 2010. States of fantasy: Barack Obama versus the Tea Party movement. Boundary 37 (2), 89–105.
Peele, G., 1984. Revival and Reaction: The Right in Contemporary America. Oxford University Press, Oxford.
Perrin, A.J., Tepper, S.J., Caren, N., Morris, S., 2011. Cultures of the Tea Party. Contexts 10 (2), 74–75.
Pierson, P., Skocpol, T., 2007. The Transformation of American Politics: Activist Government and the Rise of Conservatism. Princeton University Press, Princeton.
Rothbard, M.N., 2003a. Why be libertarian? In: Schneider, G.L. (Ed.), Conservatism in America Since 1930: A Reader. New York University Press, New York, pp. 253–258.
Rothbard, M.N., 2003b. What is libertarianism? In: Schneider, G.L. (Ed.), Conservatism in America Since 1930: A Reader. New York University Press, New York, pp. 262–273.
Schneider, G.L., 2003. Conservatism in America Since 1930. New York University Press, New York.
Schoenwald, J.M., 2001. A Time for Choosing: The Rise of Modern American Conservatism. Oxford University Press, Oxford.
Shannon, J.B., 1962. Conservatism. Annals of the American Academy of Political and Social Science 344, 13–24.
Shermer, E.T., 2013. Barry Goldwater and the Remaking of the American Political Landscape. The University of Arizona Press, Tucson.
Skocpol, T., 2012. Obama and America's Political Future. Harvard University Press, Cambridge.
Skocpol, T., Williamson, V., 2012. The Tea Party and the Remaking of Republican Conservatism. Oxford University Press, Oxford.
Smith, V.L., 2005. Hayek and experimental economics. The Review of Austrian Economics 18 (2), 135–144.
Smith, M.A., 2007. The Right Talk: How Conservatives Transformed the Great Society into the Economic Society. Princeton University Press, Princeton.
Taintor, D., 2013. Obama: We Need to Stop 'Lurching from Crisis to Crisis to Crisis'. http://talkingpointsmemo.com/livewire/obama-we-need-to-stop-lurching-from-crisis-to-crisis-to-crisis (accessed 05.11.13.).
Weaver, R.M., 2009. Integration is communization. In: Critchlow, D.T., MacLean, N. (Eds.), Debating the American Conservative Movement: 1945 to the Present. Rowman & Littlefield, Lanham, pp. 183–185.
Wilson, A.R., Burack, C., 2012. 'Where liberty reigns and God is supreme': the Christian right and the Tea Party movement. New Political Science 34 (2), 172–190.
Woodberry, R.D., Smith, C.S., 1998. Fundamentalism et al: conservative protestants in America. Annual Review of Sociology 24, 25–56.
Zernike, K., 2010. Boiling Mad: Inside Tea Party America. Henry Holt & Company, New York.
Zschirnt, S., 2011. The origins & meaning of liberal/conservative self-identifications revisited. Political Behavior 33, 685–701.
Zumbrunnen, J., Gangl, A., 2008. Conflict, fusion, or coexistence? The complexity of contemporary American conservatism. Political Behavior 30 (2), 199–221.

American Revolution, The

Gwenda Morgan, University of Newcastle, Newcastle upon Tyne, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous article by W.P. Adams, volume 1, pp. 430–434, © 2001, Elsevier Ltd.

Abstract

The American Revolution was the first of the Atlantic Revolutions. Having attained a status of advanced development, the American colonies were not prepared to tolerate the imposition of taxes and tighter controls which the mother country desired to impose on them. Ill-judged policies met resistance which escalated and resulted in war. Aided by the French, the Spanish and the Dutch, the Americans won a famous victory. Most innovative was the creation of a republican government, constructed over a period of three months in the summer of 1787.

By 1750, Britain's North American mainland colonies had achieved the objective conditions to operate as independent states. They had clearly identifiable elites and viable economies. They were distinguished by their wealth, political office-holding social class, education, housing, manners and dress. They had recognized centers of political power in colonial capitals where legislative bodies, dominated by committees employing parliamentary procedures that met to conduct colonial affairs. The franchise was wider than in Britain and there were no rotten boroughs. At local level, county and urban elites exercised power through membership of county benches and town meetings which met on a more regular basis and were more inclusive than their British counterparts. The continuing relationship with Britain, according to Jack P. Greene, was voluntary (Greene, 1973: pp. 132–80).

Although attempts to tighten British control over the colonies could be discerned before the outbreak of the French and Indian War (known as the Seven Years War in Britain), the war changed the dynamic between colonies and mother country. Fought on a global scale and supported by unprecedented levels of taxation, it resulted in British victories at the expense of the French and Spanish, especially in North America, doubled the size of the national debt and made Britain the world's leading imperial power. Little wonder, then, that the British government struggled in the aftermath of the war. As Linda Colley observed, the British were confounded "by the problem of having acquired too much power too quickly over too many people" (Colley, 1992: p. 102). Less sympathetic was T.H. Breen who in a withering indictment of the final years of British rule described Lord North and his cabinet colleagues as 'mediocre,' blinded by their "intellectual rigidity … to the possibility of reforming a system that no longer worked" (Breen, 2010: p. 59).

American expectations at the end of the war were high. Viewing the prospect before the colonists, the Reverend Thomas Barnard joyously proclaimed in his sermon before the Governor and Council of Massachusetts, 'the snare is broken and we are escaped,' but instead of reaping the benefits of their new security, the colonists found themselves not only hedged in by restrictions on western expansion but also subject to demands for taxes to relieve the burden on British subjects, tighter controls on the implementation of trade regulations, the creation of a new Board of Customs Commissioners and the expansion of the jurisdiction of admiralty courts which met without juries (Brown, 2000: pp. 76–78). As a consequence of these intrusions into colonial affairs, traditional rights and privileges, guaranteed to the colonists under their earliest charters, came under threat and, despite repeated setbacks, Parliament insisted on the recognition of its supremacy. This was the rock on which the empire floundered.

The Stamp Act of 1765 which imposed direct taxes on legal documents and printed paper marked a turning point in colonial relations and, although the Sugar Act of the previous year gave some inkling of what lay ahead, its application was more limited. The Stamp Act applied to all the American colonies including those in the West Indies. Led by the Virginia House of Burgesses, the tax was denounced as a violation of English liberties. Only the burgesses could tax their constituents because they alone knew what level of taxation the people could bear since they themselves would be subject to it. Taxation by the representatives of the people was "the distinguishing characteristic of British freedom, without which the ancient constitution cannot exist" (Brown, 2000: pp. 99–100). If Virginians led the way in defining the constitutional opposition to the tax, colonial seaports were the scene of popular demonstrations on an unprecedented scale. Building on existing traditions of popular protest, familiar in eighteenth-century Britain and replicated in Massachusetts, they made the position of stamp tax collector untenable rendering the act unworkable. Opposition to the Stamp Act also marked the beginning of intercolonial cooperation.

At the instigation of the Massachusetts House of Representatives, nine colonies sent representatives to a congress in New York which not only reiterated the constitutional opposition to the tax, but devised further tactics to oppose it by threatening to boycott British goods thus forcing British merchants to lobby for the Act's repeal. They also dismissed the charade that, like the majority of the people of Britain, they, the colonists, were 'virtually represented' in the British parliament. But it was not the threat to the British economy, the violence on the streets or the voices of American friends in Parliament that led to the Act's repeal, but rather the change of government in Britain. Repeal, however, was accompanied by a Declaratory Act that claimed the undisputed right of Parliament to legislate for the colonies 'in all cases whatsoever' (Brown, 2000: p. 112).

In 1767, Chancellor of the Exchequer Charles Townshend proposed new and different duties on selected items such as paint, glass, lead, paper, and tea under the guise of trade regulations, the revenue from which could be used to release royal officials from dependence on colonial legislators for their salaries. He also set up a new Board of Customs Commissioners with its headquarters in Boston, an action that boded ill for a region long associated with smuggling. He also expanded the jurisdiction of vice-admiralty courts giving them authority over breaches of the latest government measures. Townshend also addressed the continuing failure of the New York legislature to provide suitable accommodation and adequate subsistence for soldiers stationed there. In the face of continuing colonial resistance, Parliament resurrected an Act passed in the reign Henry the Eighth under which those guilty of treason or suspicion of treason should be tried in England (*Journals of the house of lords* XXXII, p. 210; *Journals of the house of commons* XXXII, p. 186). Once again the Virginia Assembly responded, members not only repeating their claim to the exclusive right of taxation but also denouncing Parliamentary measures intended to carry men to England for trial on treason charges. Among other colonies, only Rhode Island, a corporate colony enjoying a greater degree of independence than most, followed Virginia's example and denounced England's attempt to hold trials without local juries. The Virginia action led to the dissolution of the House of Burgesses whose members reconvened in a nearby location defining their meeting as a 'convention.'

In Boston, a seaport whose trade was in decline and whose economy had not recovered from the postwar depression, customs men became the butt of local wrath, but not being local men, they were less easily intimidated than the stamp tax collectors. Nonetheless, vulnerable to attack from a hostile population, the customs men requested naval protection in the execution of their duties. Matters came to a head in the wake of the seizure of John Hancock's ship *Liberty* in June 1768 and the riot which followed (Archer, 2010: p. 85). Persuaded that the provincial government of Massachusetts was 'cowed by a popular faction, lawlessness, and imminent insurrection,' the new Secretary of State, the Earl of Hillsborough ordered two regiments stationed in Ireland to Boston. In seeking to promote colonial unity, the House of Representatives dispatched a circular letter to other colonial assemblies. Hillsborough demanded that the letter be rescinded and failing that, that Governor Francis Bernard dissolve the legislature, which was done. At the instigation of the Boston town meeting, a convention of towns was called to meet on September 22 to address the grievances of the colony. Attended by 96 towns and eight districts, its statement of grievances was surprisingly mild (Archer, 2010: pp. 75, 91–93, 100). The weakness of Bernard's position had become only too obvious at the time of the Boston Stamp Act riot when he found that there was no one he could call on to disperse the rioters or even turn to for advice (Brown, 2000: pp. 100–103). Convinced that the governor was misrepresenting them to the authorities in England, it was impossible for the members of the House and the governor to reach any sort of compromise. Two thousand troops in all were billeted in Boston. Not only did this military occupation create tensions between the soldiers and the civilian population but it also raised longstanding issues about the dangers associated with standing armies (Archer, 2010: p. 91, 101). Since the prospect of an uprising was far-fetched, the troops were withdrawn but not before five Bostonians had been shot to death.

Relative calm prevailed after the repeal of the Townshend duties until the attack and burning of the government revenue cutter the *Gaspee* in June 1772. Engaged in patrolling the waters off Narragansett Bay, Rhode Island, the commander of the *Gaspee* was a hated figure among local merchants and mariners in this small New England colony, well known for strong ties to smuggling. Governor Joseph Wanton moved slowly to investigate the *Gaspee* affair and was outmaneuvered by Hillsborough who invoked the recent Parliamentary Act 'for the better securing and preserving his Majesty's dockyards, magazines, ships, ammunition and stores' which provided the death penalty for those found guilty of burning the king's ships and authorized suspects to be brought to England for trial. Yet when the Privy Council met, it overruled Hillsborough's initiative and in its place, instituted a commission of inquiry with a view to securing those charged with high treason under the statute of Henry VIII. The New England press had a field day. The commission was likened to the Inquisition and the court to Star Chamber. Intercolonial committees of correspondence sprang up. The high-powered commission failed to secure local cooperation. Little information was forthcoming and none of the perpetrators were identified. There would be no trials for treason in England, at least not yet (York, 1992: pp. 3–21).

When British efforts to implement the Tea Act met with resistance and the cargoes sent to Boston ended up at the bottom of the harbor, the British government saw it as an opportunity to isolate Boston and to restructure the government of Massachusetts which had singularly failed to submit to the government's will. Convinced that Massachusetts was at the heart of the problem in Anglo-American relations, and that Boston was in the hands of a small radical faction, reestablishing control over the province, it was assumed, would restore normality. The Coercive Acts, called by the Americans the Intolerable Acts, closed the port of Boston, remodeled the character of the Council making it subject to nomination by the Governor rather than election by the members of the House, removed the courts of law outside the town's limits and deprived the town meeting of its autonomy (Brown, 2000: pp. 143–146). Not only would the port of Boston remain closed until its citizens compensated the East India Company for the loss of its cargoes, its punishment would serve as a salutary lesson to the other colonies. Once again, the British government miscalculated as other colonies rallied to Boston's cause. The Quebec Act was not part of the punitive measures but its content confirmed colonists in their perception of the British government as bent on crushing the liberties of the colonists. Some measures in the Quebec Act were progressive in that they recognized French civil law and the use of local taxes for the support of the Catholic Church – both anathema to New Englanders – but in extending the boundaries of the province of Quebec southwards, the Act blocked the westward expansion of Virginia and Pennsylvania. In addition, there was no provision for a representative assembly on the grounds that the population of the area was too small. Until the passage of the Coercive Acts and the Quebec Act, there was little that the

British government had done that warranted the label of tyranny but now its response was harsh, costly and ill-advised. Thwarted at almost every turn, the British government failed to recognize just how flimsy was its authority.

Representatives of 12 colonies gathered in Philadelphia for the first Continental Congress in the summer of 1774. On the one hand, it prepared a loyal address to the king and on the other, laid out the rights and privileges of the colonies. It denounced the Coercive Acts and the Quebec Act as unconstitutional 'and most dangerous and destructive of American Rights.' Its policy of nonimportation, nonconsumption, and nonexportation of goods to and from Britain was buttressed by the newly created 'Association' which raised local enforcement to a new level (Brown, 2000: pp. 152–154). The Congress adjourned but was recalled following the confrontation between British troops looking for stockpiles of arms outside Boston and local militia at Lexington and Concord. As news of the engagement spread, the British soldiers struggling back to Boston suffered heavy losses from snipers along the route. As the militia laid siege to Boston, the second Continental Congress, meeting in May 1775, appointed a Virginian, George Washington, as commander-in-chief of the newly designated continental army surrounding Boston. Delegates did not initially seek independence, but with pressure building up in local communities, and the publication of Tom Paine's *Common Sense*, events moved inexorably toward independence. Paine poured scorn on the institution of monarchy and the outmoded connection whereby a continent could be governed by an island. 'There was a time when it was proper, and there was a proper time for it to cease' (Paine, 1960, pp. 34–35). That time had come.

Why the image of Britain had become so toxic by the end of the colonial period lay in the perception of history as a never-ending struggle between power and liberty. Eighteenth- century Britain had been lauded by continental writers for having solved the conundrum through its form of government which in combining the traditional forms of government – monarchy, aristocracy, and democracy – had created an equilibrium in the form of the mixed or as it was also known, balanced constitution, but political ideas in the colonies were shaped not by John Locke but by the radical republicans of the late seventeenth and early eighteenth centuries for whom liberty was always prey to men of power. Its survival required the vigilance of a virtuous people; its enemies were ambitious men with political patronage at their disposal, not to mention standing armies and city mobs. From the colonial view point, Britain had lost its way, its people had lost their virtue, its government was corrupt and liberty was crushed. The attack on the rights of the colonists and breaches of longstanding political values – in particular the raising of taxes without representation - was a violation of fundamental English rights (Bailyn, 1967). The reason why the colonists embraced the ideas of writers such as Trenchard and Gordon was because the logic of their arguments fitted the circumstances of the colonists. Americans had enjoyed the exercise of certain rights and privileges but they were not theirs by right. The British government had never recognized colonial assemblies as colonial parliaments with all the rights and privileges accruing to them as enjoyed by the English original. There had been no revolutionary settlement in the colonies as had occurred in Britain following the Glorious Revolution. Colonial assemblies were dependent on governors to call them into existence and they could dissolve them at will. Yet the position of the colonial governor in reality was weak. Owing his appointment to the British government, he was dependent on the legislature for his income. Real power, despite appearances, did lie with the lower houses of assembly but British policy toward the colonies following the ending of the Seven Years War was predicated on the assumption of colonial subordination. The disjuncture between theory and practice was resolved by revolution (Bailyn, 1968).

The War

In 1763 the colonies of British North America had closer relations with Great Britain than they did with each other, yet within 15 years, the Americans took on the might of the British empire and prevailed. The conflict lasted a bruising eight years at the end of which and following the loss of two British armies, one in the north at Saratoga (1777) and another in the south at Yorktown (1781), and after protracted peace negotiations in Paris, Britain recognized the independence of their former mainland colonies. British assumptions about their military superiority were slow to evaporate, if, in fact, they ever did, though the importance of the American colonies to Britain declined when a colonial rebellion was transformed into a global conflict as France, Spain and the Netherlands entered the war.

British officers did not believe that the Americans would or could fight according to the rules of war and law of nations while the Americans accused the British of failing to observe both conventions. Atrocities occurred on both sides and each sought to hold the other to account. When General Burgoyne wrote to General Gates about the failure of local militia to give quarter when requested in the case of German troops who fell into a trap when they attacked the town of Bennington, Gates poured scorn on Burgoyne's effrontery reminding him that his Indian auxiliaries had murdered the fiancée of one of his own officers in a case which gave the Americans one of their greatest propaganda coups of the entire war (Morgan, 2007: pp. 221–222, 225, 258). Native Americans, despite the efforts of both the British and Americans to recruit them, like the Irish, were perceived as lying outside the law of nations. Pushed, they chose the British side and paid the consequences when the war ended. Governor Dunmore's proclamation offering freedom to slaves and indentured servants who joined British forces in Virginia heaped more opprobrium on Britain's reputation. Small numbers of former slaves found freedom in Nova Scotia and Sierra Leone, but for most, bondage was perpetual. In the southern states, especially the Carolinas and Georgia where prewar social and political tensions were most acute, partisan militias with scores to settle could not be contained within the conventions of the rules of war and law of nations. When the British launched their southern campaign in the mistaken belief that pro-British sentiment in the region was strong, civil war raged. As the British army under Cornwallis fought its way to north, it became trapped in Virginia and for the second time, the British lost an entire army. Two armies were enough.

Almost until the end of the conflict, the British defined the Americans as rebels and therefore traitors subject under the Treason Act of 1696 to gruesome punishment. As late as November 1780, Lt Colonel Balfour, commander-in-chief of British forces in occupied Charleston, described his opponents as 'fighting in the cause of treason and rebellion' (Moultrie, 1968 [1802], p. 145). When Ethan Allen and some of his soldiers fell into British hands in Canada in 1775, he and 33 others were sent to Britain to await an uncertain fate. Badly mistreated and taunted with the prospect of a traitor's death, Allen was finally sent back to New York, then in British hands, from where he escaped making his way home. Officers might be exchanged when men of equal rank were taken prisoner but soldiers and sailors were not so lucky. In reality the British treated the Americans as enemy combatants but there was no formal recognition of their status other than as rebels though carrying captured soldiers to Britain was not repeated. The British government would not negotiate with the American congress.

When the Americans defeated General Burgoyne at Saratoga in October 1777, the Americans acquired a large number of prisoners who became known as the 'Convention army.' As a consequence of the defeat, 5000 or so British and German soldiers spent the remainder of the war in custody in Massachusetts, Maryland, New Jersey, and Virginia, some of whom were hired out as laborers. Those captured by the British following the taking of New York and Charleston were confined in churches and on prison ships which led to reports of ill treatment, starvation, and high mortality. The situation of sailors was different again. In 1777 Parliament adopted an Act 'to empower his Majesty to secure and detain persons charged with, or suspected of, the crime of high treason, committed in any of his Majesty's colonies or plantations in America, or on the high seas, or the crime of piracy' (*Statutes at Large*, 1780: pp. 18–19). The Act allowed British vessels to seize American sailors and carry them to England where they remained for the duration of the war unless they succumbed to British inducements that they enlist in the British navy. Some 3000 sailors were imprisoned in the jails at Old Mill (Plymouth) and Forten (Portsmouth). There were some exchanges of seamen organized by Benjamin Franklin when he was commissioner in Paris but no formal arrangement was concluded. The Americans seldom had enough British seamen in their custody to exchange for the larger number of Americans in British hands. It was only when the Earl of Shelburne formed a new government in 1782, following the fall of the North administration that Americans captured by the British, were finally recognized as prisoners of war.

The American Revolution in Perspective

The American Revolution was the first of the so-called Atlantic Revolutions and while subsequent revolutions in the Atlantic world had links of various kinds to the American – it certainly bankrupted the French – what prompted American patriots to refer to their contest as a revolution is not altogether clear. Though the ghost of Oliver Cromwell was not dead in New England as Alfred Young has demonstrated, it was unlikely to be the English civil war while hopes of reconciliation were still alive (Young, 2006: pp. 156–164). The most obvious model was the Glorious Revolution of 1688 which overthrew the last of the Stuarts, James II, and with him the divine right of kings and the prospect of a return to Catholicism, guaranteed the role of parliament in national affairs and issued the Declaration of Right, a national charter of liberty (Davis, 1990: pp. 9–10). While some clauses in the latter can be found in the Declaration of Independence and the Bill of Rights, what was truly innovative about the American Revolution was the creation of a republican form of government and its ratification by the American people. Republican governments were not new but they were not found in extensive territories such as the United States, and they did not last.

Confederation and Constitution

In 1787 state representatives gathered in Philadelphia to revise America's first constitution, the Articles of Confederation, that had been adopted during the course of the war and ratified by the last of the states in 1781. A weak instrument of central authority and with no income of its own, it had, nonetheless, kept the nation together during the war and the early years of the peace. Its shortcomings became more evident after the war ended when Britain failed to fulfill its obligations under the Treaty of Paris, social order broke down in some states, borders were contested, interstate rivalries threatened economic warfare and the exclusive authority of Congress in Indian affairs was ignored. Yet it was not without its achievements in peacetime–the passage of the Northwest Ordnance excluding slavery from the Northwest Territories was one of them. If the creation of the Federal Constitution was a reaction to the Articles of Confederation, it was also a riposte to state legislatures that had become the dominant force at state level after the collapse of royal government.

But the delegates to the Federal Convention decided against revising the Articles and started over. The new Constitution was to be the supreme law of the land, it created a power superior to that of the state legislatures but it lacked the power to veto state laws though Madison had favored a veto. Dividing power between the executive, legislature and judiciary, the delegates to the Philadelphia convention in the summer of 1787 sought to balance power and liberty while at the same time having to accommodate the special interests which reflected the diverse economic and social character of the regions that made up the United States and without which nothing could be achieved. There was little problem over the framework of the government, the respective branches and the length of terms of service since they were founded on solid republican principles. The problems came with the question of representation in the two legislative bodies, noninterference in the slave trade and commercial legislation. Compromise was the key to success but it was hard won. Democracy played little role in debates in the Convention and then only to be dismissed. The sentiment expressed by Jefferson in his *Notes on the State of Virginia* (1785) when he wrote that an elected despotism was not the government they had fought for, was widely shared by convention delegates. Slavery was not mentioned by name in the Constitution, but the document was shot through with measures that protected it and it was further reinforced by the Bill of Rights.

Wiececk described it as the fairy-tale witch who though not invited to the party, came anyhow and put a curse on the proceedings (Morgan, 2007: p. 146). Most difficult of all to resolve was the question of representation. James Madison who wrote the Virginia plan which formed the basis of early discussions on the shape of the Constitution, favored proportional representation in both houses, but fearing the loss of power, the small states would not budge. Holding out for equal representation in both houses, they won equal representation in the senate, a concession which one eminent constitutional scholar (Rakove, 1996: p. 157) described as more costly in the long run than the despised three-fifths clause which counted slaves at three-fifths of their total number so artificially boosting southern representation in the lower house of the Congress. Delegates had employed numerous checks and balances to avoid the failure of past republics and made the amendment process unduly cumbersome. Created to meet the concerns of Americans in the late eighteenth-century, it remains the government of the United States in the twenty-first century though the country itself has changed beyond recognition.

Written in the name of the people, the Constitution was submitted to them in specially elected conventions. An early vote of approval was anticipated – the Federalists had the power, the money, and the organization – but as the opposition began to make up ground, the outcome was no longer assured. The Constitution came under attack from many directions. It contained no rights or safeguards; it created a consolidated rather than a federal structure, it would be too powerful, it threatened the states, the size of the lower house was too small, the numbers of congressmen were too few to represent the interests of their constituents; it was too remote. The first major study of ratification was published only in 2010 (Maier, 2010).

Historiography

There is no single narrative of the American Revolution but two approaches, broadly interpreted, tend to dominate its interpretation. Both trace their origins back to the first generation of professional historians who cast the Revolution in two very different contexts: the imperial and the progressive. (For a more detailed account of this earlier historiography see Morgan, 2007) The former approached the Revolution from the perspective of empire, the latter saw the Revolution as a dual struggle about who should rule over America and who should rule in America.

The concept of 'the Atlantic World' developed in the 1980s. Earlier forays into comparative history laid the ground work but when Palmer published his pioneering study of revolution on both sides of the Atlantic, it met with a less than favorable reception (Palmer, 1959, 1964). More successful was Langley's (1998) inter-American study of revolutions in which he compared the patterns of leadership, political history and social character of British North America, Haiti and Latin America. Wim Klooster (2009) examined four Atlantic Revolutions in his comparative study: the American, French, Haitian, and Spanish American, none of which did he think was inevitable and all of which he linked to the vast expenditure on the warfare which preceded them. Critics of the concept point to the fact that the Atlantic is not an enclosed space and its utility as a concept diminishes in the later eighteenth-century. Eliga Gould and Stephen Conway entered a cautionary note, the former pointing to the significance that France still held for the British while Conway stressed the common values shared by the military of both countries. The work of O'Shaughnessy (2000) who reclaimed the Caribbean as an essential theater of war and linked the defeat at Yorktown to Admiral Rodney's dalliance in the West Indies and the plundering of St Eustatius fitted neatly into an Atlantic framework but Armitage's work on the Declaration of Independence did not. Armitage favored a global perspective arguing that the international reception of the Declaration when examined in a global context was more revealing than an Atlantic one since it showed the emergence of a world of states from an earlier world of empires. Jeremy Adelman, however, questioned whether the world of empires can be written off quite so easily since empires could lose large swathes of territory only to reinvent themselves successfully elsewhere in the world. The importance of this wider perspective and longer time frame was evident in Marshall's work which demonstrated how the British succeeded in achieving in India what they failed to accomplish in America – a structure of local government that ensured 'that British territories were effectively defended and that the wealth of the empire was properly husbanded through a system of trade regulation' (Marshall, 2005: p. 12). A global perspective has also proved relevant to historians working on very different types of subjects such as Maya Jasanoff who described her work in *Liberty's exiles* as the 'first global history of the loyalist diaspora.' As 'British power regrouped, expanded and reshaped itself across the world,' she followed the journeys of a small number of loyalist families and individuals (Jasanoff, 2011: pp. 8–11).

Other historians continue to build on the legacy of the progressives and their successors whose work especially in the 1960s and 1970s opened doors to subjects few historians had ever explored. They seek to penetrate below the surface of colonial society as represented by the political and social elite to recover the numerous other actors variously described as 'the people,' the 'ordinary people,' 'the common folk,' and the 'real people,' mechanics, artisans, seamen, plantation slaves, urban slaves, black sailors, the 'middling sort,' women of all classes, Native Americans and loyalists of all sorts. Clearly, they participated in the Revolution – some would say they were the Revolution – but their roles, although becoming clearer, remain relatively obscure especially when compared with the disproportionate amount of time and effort bestowed on the leaders. They made up the crowds who intimidated the stamp collectors, who faced down the customs men and who dumped the East India Company's tea in Boston harbor. They were the enforcers of the Association, the men who filled the ranks of the Continental Army once the early 'rage militaire' was over and the women who ran the households and farms they left behind.

The focus of much of this work was urban. 'It was urban interests that were most threatened, by the Sugar Act, the Stamp Act, the Townshend taxes and the Tea Act' and without crowd action, wrote Countryman 'there would have been no resistance movement' (Countryman, 1985: p. 87). In his study of

urban society in Boston, New York and Philadelphia, Nash argues that the Revolution in these towns was accompanied by a 'profound social upheaval' shaped by preexisting divisions which erupted like a powerful volcano (Nash, 1979: pp. xv–xvi, 340). Ray Raphael (2001), writing history from the bottom up, focuses on army recruits, African Americans, Native Americans, Pacifists and Loyalists and women. Benjamin Carp (2007) asks how different types of urban landscapes shaped the responses of their respective populations to the Revolution. Richard Archer (2010) writes from both the bottom-up and the top-down approaches in his meticulous study of the events that took place in Boston in the years between the Townshend Duties and the Coercive Acts when a standing army swamped the town (Archer, 2010: p. xviii). T.H. Breen agreed that 'the insurgency began not with a familiar cast of American leaders but with the people' but he questioned the emphasis on towns asserting that 'an urban focus is misleading.' Breen reminds his fellow historians that most free white inhabitants lived in small communities comprising a few hundred and in numerous case studies; he reveals their active participation in the Revolution (Breen, 2010: pp. 27, 45–46). Woody Holton shares Breen's focus on the importance of rural society, because in Virginia there were no towns. Small planters made up the bulk of plantation owners in the rural south and Holton and McDonnell ascribe to them a leading role in pushing forward opposition to British policies (Holton, 1999). Native Americans and African slaves were no longer simply victims of the Revolution but active participants in the conflict as demonstrated by Colin G. Calloway (1995); Sylvia R. Frey (1991) and Gary Nash (2006). Nash brings together these fragmented histories, the legacy of the new social history, in *The Unknown American Revolution* (2005) in which he views the legacy of the Revolution for ordinary people as a half finished project, a view also shared by Alfred Young (2006). Terry Bouton, however, reaches a more negative conclusion. If the role of the people is examined over the longer term "to see the Revolution as a democratic victory for the people, one has to cut most of the people out of the story"' (Bouton, 2007: pp. 4–5).

See also: Confederations, History of; English Revolution, The; Freedom, History of; French Revolution, The; National Liberation Movements.

Bibliography

Anon., 1780. Statutes at Large, vol. 13. Eyre and Strahan, London.
Adelman, Jeremy, April, 2008. An age of imperial revolutions. American Historical Revolution 113 (2), 319–340.
Archer, Richard, 2010. As If an Enemy's Country: The British Occupation of Boston and the Origins of Revolution. Oxford University Press, Oxford and New York.
Bailyn, Bernard, 1967. The Ideological Origins of the American Revolution. Harvard University Press, Cambridge, MA.
Bailyn, Bernard, 1968. The Origins of American Politics. Knopf, New York.
Bouton, Terry, 2007. Taming Democracy: "The People", the Founders, and the Troubled Ending of the American Revolution. Oxford University Press, New York.
Breen, T.H., 2010. American Insurgents, American Patriots: The Revolution of the People. Hill and Wang, New York.
Brown, Richard D. (Ed.), 2000. Major Problems in the Era of the American Revolution, second ed. Houghton Mifflin Company, Boston and New York, pp. 1760–1791.
Calloway, G. Colin, 1995. The American Revolution in Indian Country. Cambridge University Press, Cambridge.
Carp, Benjamin L., 2007. Rebels Rising: Cities and the American Revolution. Oxford University Press, New York.
Colley, Linda, 1992. Britons: Forging the Nation 1707–1837. Yale University Press, New Haven, CT.
Countryman, Edward, 1985. The American Revolution. Hill and Wang, New York.
Davis, David Brion, 1990. Revolutions: Reflections on American Equality and Foreign Relations. Harvard University Press, Cambridge, MA.
Frey, Sylvia R., 1991. Water from the Rock: Black Resistance in a Revolutionary Age. Princeton University Press, Princeton, NJ.
Greene, Jack P., 1973. An uneasy connection. In: Kurtz, Stephen G., Hutson, James H. (Eds.), Essays on the American Revolution. University of North Carolina Press, Chapel Hill, pp. 132–180.
Greene, Jack P., 2000. The American revolution. American Historical Review 105 (1), 93–102.
Holton, Woody, 1999. Forced Founders: Indians, Debtors, Slaves, and the Making of the American Revolution in Virginia. University of North Carolina Press, Chapel Hill.
Jasanoff, Maya, 2011. Liberty's Exiles: The Loss of America and the Remaking of the British Empire. Harper Press, London.
Klooster, Wim, 2009. Revolutions in the Atlantic World: A Comparative History. New York University Press, New York and London.
Langley, Lester D., 1998. The Americas in the Age of Revolution, 1750–1850. Yale University Press, New Haven, CT.
Maier, Pauline, 1997. American Scripture: Making the Declaration of Independence. Knopf, New York.
Maier, Pauline, 2010. Ratification: The People Debate the Constitution, 1787–1788. Simon and Schuster, New York.
Marshall, Peter J., 2005. The Making and Unmaking of Empires: Britain, India and America, 1750–1783. Oxford University Press, Oxford.
Morgan, Gwenda, 2007. The Debate on the American Revolution: Issues in Historiography. Manchester University Press, Manchester.
Moultrie, William, 1968 [1802]. Memoirs of the American Revolution, vols. 2. Arno Press, New York.
Nash, Gary B., 1979. The Urban Crucible: Social Change, Political Consciousness, and the Origins of the American Revolution. Harvard University Press, Cambridge, MA.
Nash, Gary B., 2005. The Unknown American Revolution: The Unruly Birth of Democracy and the Struggle to Create America. Viking, New York.
Nash, Gary B., 2006. The Forgotten Fifth: African Americans in the Age of Revolution. Harvard University Press, Cambridge, MA.
O'Shaughnessy, Andrew J., 2000. An Empire Divided: The American Revolution and the British Caribbean. University of Pennsylvania Press, Philadelphia.
Paine, Thomas, 1960 [1776, 1792]. Common Sense and the Crisis. Doubleday, Garden City, New York.
Palmer, Robert R., 1959, 1964. The Age of Democratic Revolution: A Political History of Europe and America, 1760–1800, vols. 2. Princeton University Press, Princeton, NJ.
Rakove, Jack N., 1996. Original Meanings: Politics and Ideas in the Making of the Constitution. Knopf, New York.
Raphael, Ray, 2001. A People's History of the American Revolution: How Common People Shaped the Fight for Independence. The New Press, New York.
York, Neil, 1992. The uses of law and the Gaspee affair. Rhode Island History 50, 3–21.
Young, Alfred F., 2006. Liberty Tree: Ordinary People and the American Revolution. New York University Press, New York and London.

American Studies, Twentieth Century History of

Wil M Verhoeven, University of Groningen, Groningen, The Netherlands

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 1, pp. 454–461, © 2001, Elsevier Ltd.

Abstract

When the academic discipline of 'American Studies' was developed in the course of the 1930s, it was explicitly set up as an area of studies where scholars and students of, notably, history and literature could meet and join forces in an integrated, interdisciplinary context in order to redress what was felt to be an increasingly acute challenge: how to bridge the widening gap between the literary and historical approaches to the study of American society. However, the emergence of this new, holistic approach to American society immediately created its own dialectic: between those seeing American society as a whole, and those emphasizing its plurality. Until the 1960s and 1970s, roughly, there was a fair degree of consensus among scholars about what the common denominators were in American society. Ever since the 'culture wars' of the 1980s and early 1990s, however, American studies has come to reflect methodologically the 'cultural crisis' and fragmentation that many have observed in American society itself. Although it may never have become the holistic approach originally envisaged, then, few in American studies would deny that trying to understand and describe American society is in effect an attempt to fathom what sustains that society, what forces operate it and direct its purposes. That is what American studies, from a wide variety of disciplinary angles, continues to do.

When the academic discipline of 'American Studies' was developed in the course of the 1930s, it was explicitly set up as an area of studies where scholars and students of, notably, history and literature could meet and join forces in an integrated, interdisciplinary context in order to redress what was felt to be an increasingly acute challenge: how to bridge the widening gap between the literary and historical approaches to the study of American society. However, the emergence of this new, holistic approach to American society immediately created its own dialectic: between those seeing American society as a whole, and those emphasizing its plurality; between those wanting to highlight American society's unifying myths and symbols as well as its common values and aims, and those drawing attention to its centripetal forces, sometimes identifying these as a dynamics of cultural diversity, sometimes as a tendency toward social segmentation and discontinuity. Even at the beginning of the twenty-first century – as we emerge from the 'culture wars' of the late 1980s and 1990s – it remains difficult to establish any kind of consensus among analysts about the common denominators in American society, about its inner core or even its outer boundaries.

Nevertheless, a number of recurrent themes can be identified that appear to dominate past and present debates in American studies about the nature of American society. Following that society through history, distinguishing the lasting from the transient, the center from the periphery, was one of the original drives behind the emergence of American studies, and this strategy continues to inspire and energize the discipline. Although it may never have become the holistic approach originally envisaged, few in American studies today would deny that trying to understand and describe American society is in effect an attempt to fathom what sustains that society, what forces operate it and direct its purposes. Hence, even though, as Wiebe has argued in *The Segmented Society* (1975), surveying the whole of a society always tends to emphasize social patterns rather than social processes and to subordinate 'a history of themes inside American society to a history of the society incorporating them,' one may still look meaningfully at those themes, and at the history of their emergence and development. That is what American studies, from a wide variety of disciplinary angles, continues to do.

Terms and Periods

Apart from qualifying the unanimity among scholars in American studies regarding the object and methodology of their inquiry, a further note of caution should be sounded concerning terminology. Within the discipline of American studies – itself an umbrella term – the term 'society' is usually taken to refer to the aggregate of the politics, history, economics, education, geography, religion, environment, and culture of the United States, as well as to the relations and tensions between these elements. However, many Americanists, especially those approaching their object from the realms of literary analysis, the history of ideas, and 'cultural studies,' would refer to the same grouping of aspects of American society and the interplay between them as American 'culture.' As Potter remarked in *History and American Society* (1973), this terminological confusion is symptomatic of 'a serious general problem in the study of societies,' viz. the lack of fixed points of similarity in different societies, and between different aspects of societies, with which to measure continuity and separateness. For the sake of clarity (and following Wiebe), 'culture' will here be defined in its more narrow sense as 'those values and habits conditioning everyday choices in such areas as family governance, work, religious belief, friendship, and casual interchange,' while 'society' will be seen as combining 'these patterns with broader and more systematic realms of behavior, such as the organization of a community's life, the structure of a profession or business or religious denomination, and the formula for apportioning economic rewards.'

Traditionally, the evolution of American society during the first two hundred years since the Revolution has been divided into three phases, each associated with a distinctive social system: from the Revolution to the 1790s; from 1830 to 1890; and from 1920 to the 1970s. The intervening years – from the 1800s to the 1830s, and from the 1890s to the 1920s – are seen as crucial periods of transition from one social system to the next: from 'eighteenth-' to 'nineteenth-,' to 'twentieth-century society.' Whether the last three decades of the twentieth century constitute a fourth phase in American society (cf. for instance Lipset, 1979 [1980]), or a continuation of the transitional period of the 1960s and 70s, or, indeed, the 'disuniting of America' altogether as signaled by others (cf. Schlesinger, 1991 [1993]), is still a moot point. In the following, aspects of American society that have dominated debates and explorations in American studies will be discussed in terms of their emergence and evolutionary history.

Seventeenth- and Eighteenth-century Society

American society, one often reads in accounts of the settlement and early colonization of North America, began with a dream – a vision of a morally pure, socially just New Eden, itself the reflection of a long tradition of European utopianism. European observers and settlers alike represented seventeenth- and eighteenth-century American society in terms of a transatlantic opposition that is still relevant today: an opposition between Old World corruption and New World innocence, between darkness and light, between stasis and progress. For even if the early American settlers were a motley crowd internally divided by differences of ethnicity, language, ideology, religion, politics, and commercial interest, there was always a remarkable consensus among them about what America was, or was supposed to be – a society that, defining itself against the backdrop of European wars, exploitation, inequality, and persecution, was somehow unique, whole, progressive, and civilized.

More remarkable than that, perhaps, was that, despite their differences, the various stakeholders in the new nation were all prepared, if for very disparate reasons, to rally round a symbol of American society first provided by one radical group of early settlers, the New England Puritans. As early as 1622, Puritan leader John Winthrop labeled England as 'this sinfull land,' plagued by poverty, inequitable taxation, a bureaucratic legal system, and religious intolerance; as the first elected governor of the Massachusetts Bay Company, Winthrop set off for New England in 1630, envisioning a 'city upon a Hill' as the utopian foundation for the new society that he and his fellow Puritans would be building. The New England Puritans may have had fairly specific religious reasons for regarding their community as a model for other colonists and as the first step toward establishing a kingdom of God in the New World that would lead the world into a new millennium, but the image of America as 'a beacon upon a hill' sufficiently reflected the ambitions of other, non-Puritan settlers for it to be adopted widely as one of the most resonant and sustaining symbols of American society, both in qualitative, intrinsic terms, and in its relation to the rest of the world. Thus, long after the waning of the Puritan theocracy, the founding fathers of the 1630s continued to be ritually invoked as heroes in the cause of liberty – not just in New England, but throughout other Eastern States and the South, and as much during the Great Revival as during the French and Indian War and the Revolution. The epic of the Puritan exodus became the common property of all Anglo-American settlers, and the rhetoric and ideology of the Puritan religious errand into that 'unredeemed, unchristianized, lawless region' (as Hawthorne put it) became the rhetoric and ideology of America's cultural errand into the 'wilderness.'

As a symbol reflecting the internal dynamics of the group, the idea of America as a bridgehead of civilization beleaguered by a 'vast and desolate wilderness' (Rowlandson, 1682 [1994]) to an important degree determined the social mechanics of what has been termed 'the ritual of consensus' (Bercovitch, 1993). Faced as they were by the challenges of survival and success, consensus filled the need felt by Puritan and non-Puritan settlers alike for a certain social order. In particular, such a ritual of consensus had to regulate the rights and responsibilities of the individual versus the group. According to Bercovitch, there were three basic tenets of Puritan consensus, which in the course of the seventeenth century transformed from a tribal ritual into a national ritual of cultural and social origins. The first tenet was migration, as a function of divine mission or prophecy – which contributed significantly toward rationalizing the expansionist and acquisitive aspects of settlement (the Puritans being as much interested in material gain as in salvation). The second tenet of consensus was related to discipline: seeing that an individual's success made visible the meaning of the errand of the group, the challenge was how to endorse individualism without promoting anarchy. The third tenet of consensus was concerned with progress: constantly affirming that it was en route from a sacred past to a sacred future, the Puritan community established institutions that were geared more toward sustaining progress and growth than toward maintaining stability. Crucially, the Puritan ritual of consensus was constantly enacted – both in the sense of established in law and in the sense of publicly performed – in a series of interlocking covenants to which individuals were invited to subscribe their names if they wanted to be regarded as members of the community. The first and most famous of such civil covenants, the 'Mayflower Compact' (which was signed in 1620 by the Pilgrim Fathers on board the ship that took them to what was to become Plymouth Plantation), reflects how these covenants were aimed at establishing a 'Civil Body Politic' by preserving individual freedom while demanding submission to 'the general good.' It is this Puritan rhetoric of voluntary compacts and consensus, of community grounded in individuals committing themselves freely to their 'common providence,' that gave American society its unique slant on the universal process of introducing systematic relationships among individuals, groups, and institutions; and it is the same rhetoric that, being both universal yet culturally specific, facilitated the transition from the Puritan to the Yankee, and from errand to manifest destiny and the American Dream. It is the legacy of the Puritan errand that lies at the basis of America's social awareness vis-à-vis the world at large: America as the 'redeemer nation,' which, according to Tuveson (1968), can be reduced to the following key elements: 'chosen race; chosen nation; millennial-utopian destiny; fighting God's war between good (progress) and evil (regression), in which

the United States is to play a starring role as world redeemer.' Thus, when Woodrow Wilson said that 'America had the infinite privilege of fulfilling her destiny and saving the world,' he was not saying anything startlingly new.

In the course of the second half of the eighteenth century, the spiritual and the more secular beginnings of American society (symbolized by Plymouth Plantation and Jamestown, respectively, by the 'beacon on the hill' and the Horn of Plenty) had become sufficiently blended for Crèvecoeur to be able to pose the notorious question in his *Letters from an American Farmer* (1782 [1981]), 'What, then, is the American, this new man?' and to assume – or suggest – that this was actually more than a rhetorical question: that there really was such a social subject as 'an American.' What in hindsight is perhaps even more perplexing than Crèvecoeur's question as such, is his answer to that question: Crèvecoeur's 'American' and the society he is said to have created are still by and large what we would recognize as American national identity and society today. An immigrant newly arrived in America, Crèvecoeur informs us, would find himself 'on a new continent; a *modern society* offers itself to his contemplation, different from what he had hitherto seen.' He would see a society that is egalitarian and classless, in which individuals are 'animated with the spirit of an industry which is unfettered and unrestrained, because each person works for himself.' The world's poor, oppressed and persecuted will find safety and plenty of opportunity in 'this great American asylum.' To underline the radical modernity and uniqueness of this new society, Crèvecoeur assures us that in America 'individuals of all nations are melted into a new race of men, whose labours and posterity will one day cause great changes in the world.' The millennial dreams of the New England Puritans have in Crèvecoeur's *Letters* become America's millennial dreams – now all Americans are 'pilgrims': 'Americans are the Western pilgrims who are carrying along with them that great mass of arts, sciences, vigour and industry which began long since in the East; they will finish the great cycle.'

Nineteenth-century Society

The individual's 'self-interest' and his concomitant voluntary submission to his nation's mission form the nucleus of Crèvecoeur's America – and they still do. The American Revolution, which ended the colonial stage in the evolution of American society, did not therefore constitute that sharp, radical break it has traditionally been made out to be. Since the end of the French and Indian War, self-interest had been synonymous with (Protestant) patriotism, and it was through their appeal to this patriotic self-interest that the Revolutionary leaders managed to steer the colonies victoriously through the conflict with Britain and toward nationhood. Thus, while Crèvecoeur's distressed frontiersman announces in his final letter (burdened by his creator's nostalgia for utopian pastoralism) that in 'these calamitous times' he feels forced to flee his farm and to resettle his family 'under the wigwam' in the western wilderness, the Whig leaders succeeded in bringing the violence and disruption of the Revolution under control by turning it into a redemptive, controlling ritual of national identity. The push for independence, which elsewhere (most notoriously in subsequent years in France) often triggered a complete collapse of the social fabric as the forces of change spun out of control, in America engendered a spirit of intense nationalism and enforced the model of consensus. What the Puritan Fathers had begun – overcoming discord in society by turning anxiety into a vehicle of social control – was now finished by the Founding Fathers.

With the 'family quarrel' – so unfortunate but necessary – finally over, America was ready to enter the final phase of its evolution toward full nationhood; the Revolution having given America independence from the Old World, it was now time to complete the errand into the wilderness by going west, not to hide under a wigwam, but to conquer the continent. Space and diversity, as Madison had argued in *Federalist Number Ten*, would be the best guarantors of liberty, because a dispersed and heterogeneous nation would be able to contain more factiousness than a compact and more homogenous one. And the dangers of faction were increasing all the time in the wake of the Revolution, as the system of graded and interlinking social relationships of eighteenth-century society – marked by paternalism, the moral economy, and virtuous republicanism (or 'civic virtue') – began to give way to an increasing emphasis on personal mobility – social mobility as well as geographical.

Uncurbed social ambition and a vigorous spirit of enterprise led to a rapid crumbling of deferential restraints, and the meaning of American wholeness came under significant pressure during these years of transition. The old household economy of colonial America had gone into a complete nosedive in the 1790s; fuelled by the stirrings of industrialism, the opening up of the west, the Louisiana Purchase of 1803, and the Napoleonic Wars in Europe (which virtually eliminated mercantile competition), liberal capitalism triggered a socio-economic revolution that transformed the United States in the course of two decades into a market economy and market society. The end of the War of 1812 gave a further boost to America's economic revolution, with a rapidly growing immigrant population, and the building of the canals and the early railways permitting the republic to enter an era of unprecedented expansion, especially toward the vast western regions.

American geopolitics in the early decades of the nineteenth-century to a large extent shaped the growing awareness of nationhood and state. This resurgence of nationalism peaked with the so-called 'Monroe Doctrine,' which basically meant that the United States declared they would not accept any future colonization by European powers of 'the American continents.' But in social terms a price had to be paid for America achieving hegemony in the Western Hemisphere, and a controversy erupted over the question of not whether, but how the nation should be allowed to expand. With the election of Jackson as President of the United States in 1829 a powerful democratic movement swept the country, paving the way for mass politics and what was called at the time 'the age of the common man.' The fight against social and economical inequality and privilege dominated the national political agenda. And although there is still considerable disagreement among historians as to how much of this drive toward democratic reform and anti-elitism can actually be attributed to Jackson and his followers, the reaction against laissez-faire capitalism from among the working classes and other exploited groups in the new industrial society was considerable.

Liberty and equality became the cornerstones of the new American society – even though the veneer of egalitarianism did not affect a host of distinctions underneath (notably those of race and gender).

A tourist like the French aristocrat Alexis de Tocqueville, who came to America in the 1830s to discover why the efforts at establishing democracy in France, starting with the French Revolution, had failed while the American Revolution had produced a stable democratic republic, may not have seen far beyond the surface respectability but he did make some seminal comments on American egalitarianism. According to Tocqueville, America was 'the first new nation' in the world. In fact, it was Tocqueville who first referred to the United States as 'exceptional,' that is, as a nation qualitatively different from other nations. In his extraordinarily influential book *Democracy in America* (1835–39 [1945]), Tocqueville identified five elements in the American Creed as it had emerged from the revolutionary ideology – liberty, egalitarianism, individualism, populism, and laissez-faire – but, as Tocqueville pointed out, egalitarianism in American involves an equality of opportunity and respect, not of result or condition. Therefore there is no emphasis on social hierarchy in America – no monarchies or aristocracies (which distinguishes the United States from post-feudal nations as such) – although there is social segmentation. It is this, Tocqueville noticed, and many others since him, that was central to American exceptionalism: 'What held Americans together was their ability to live apart …. From this elementary principle emerged a pattern of beliefs and behavior that was recognizably American' (Wiebe, 1975).

The idea of space as a sine qua non of liberty and equality was popularized at the end of the nineteenth century by the historian Frederick Jackson Turner. At a point in America's history when, he claimed, the vast, unsettled lands were gone, Turner identified the westward expansion of the United States as the crucial formative force behind American individualism, nationalism, and democracy. Because he had to constantly reinvent himself and reconfirm his essential Americanness as he pushed further west, the frontiersman, Turner argued, had defined American civilization. With the waning of the western wilderness, a crucial period in American history had closed. In fact, Turner's concerns turned out to be rather premature as in the decades following his bemoaning of the 'passing of the frontier' the US government gave away more land from the public domain than it had done before – but the 'frontier myth' kept a romantic image of the West alive for long after.

Twentieth-century Society

If, as historian Richard Hofstadter, 1948 has observed, it has been America's 'fate as a nation not to have ideologies, but to be one,' then that ideology was very much in place at the beginning of the twentieth century. But although the ideology did not change materially from that of the late nineteenth century, the face of society did. One of the most conspicuous developments was the emergence of a new social elite – the occupational elite. The end of the nineteenth century saw an explosion of the number of Americans in administrative and professional jobs. Industry and the new technologies needed managers, technicians, and accountants; cities needed professionals in the medical, commercial, legal, and educational services. By the turn of the century, these new professionals had come to constitute a distinct social group – the new middle class.

Although the new professional class enabled individuals to climb the social ladder more rapidly than in the nineteenth century, the new class was by no means any more homogeneous – the nineteenth-century segmentation of society on the basis of region, ethnic background, and access to capital was replaced by a segmentation according to professional occupation – each segment (clerks, technicians, lawyers, teachers, engineers) having its distinct identity and social awareness, and the vitality of each depending on the vitality of the whole.

While diversity remained the common denominator in society, the need for cohesion in the national society never disappeared – the ideology of consensus still being as strong as in the days of the Puritans. If in the nineteenth century the outrage of national conscience had directed itself against sloth, drunkenness, and profligacy, in the twentieth century the social hegemony aimed its arrows at the leveling left, communists, and other 'un-American' elements (descendants of Crèvecoeur's 'off-casts') that threatened to undermine the 'volutary order' (Berthoff, 1971).

But the most powerful cement of the new social system was consumption. Starting hesitantly in the latter decades of the nineteenth century, accelerating in the 1920s, and exploding in the 1950s, consumption became the foundation for the modern America. By the 1960s, consumption was no longer a privilege but a right – to some extent even a civil duty. Consumption became everybody's stake in society – the ultimate equalizer. In the course of the twentieth century, Americans began to regard the meaning of work as the power to spend, and increasingly began to discover their identity and social position in their power to consume. Mass consumption became an inalienable right. Women derived their freedom from being primary consumers; consumption was an effective weapon against communism; and the urban African-Americans who looted shops during race riots thereby underlined the fact that they were still being denied their basic civic right: the right to consume.

Consumption has been regarded as the ideological twentieth-century answer to the challenges of differences of class and civil rights. The nineteenth century had failed to manage the problems of class; the economic elite of the twentieth century quite early on accepted class as a social fact. By quickly and consistently demolishing labor organizations and any manifestations of socialism wherever they appeared, and simultaneously offering occupational segmentation as a new model of social mobility, the threat of class conflict was contained. The importance of regular employment was paramount to egalitarianism for the credit it generated: for it was the credit that made the professional or worker a consumer, and the consumer a citizen. In a similar way, twentieth-century society put paid to the issue of race conflict: the economic elite offered members of minorities as well all other Americans a passport to consumerism and a place somewhere in the great interdependency of the system of segmented wholeness. Race, like class, simply was no longer relevant from a socioeconomic point of view. Racial or other minorities can either aspire to

assimilate to the segmented system, even to belong to the social elite, or they can decide to rely on such benefits as are provided within the context of social security.

The much-touted revolutionary drive of the 1960s' 'counterculture' did not materially change the system of segmented wholeness. For a while it looked as if there might be more emphasis on the common denominators in American society, as if there were an America beyond the segmented system (classless, gender-neutral, race-neutral). Indeed, changes in the overarching social fabric have occasionally been affected by pressure from within one or more of the segments (such as racial desegregation, and a more equitable job market). But, by and large, America has remained essentially a nation of segments. However, if America still subscribes to this concept of nationality, it does so voluntarily. In contrast to the rest of the world, as Gover (1948) has observed, Americans believe that nationality is 'an act of will, rather than of chance of destiny.'

That is what Europeans have never understood about America, according to Baudrillard (1986 [1988]). He argues that the great lesson of the success of America's social formation is that 'freedom and equality, like ease and grace, only exist where they are present from the outset. This is the surprise democracy ha[s] in store for us: equality is at the beginning, not the end. That is the difference between egalitarianism and democracy: democracy presupposes equality at the outset, egalitarianism presupposes it at the end. "Democracy demands that all of its citizens *begin* even. Egalitarianism insists that they all *finish* even."' Whereas Europe has remained stuck in the old rut of social difference and is constantly dragged back into the history of its bourgeois culture, America, by contrast, has achieved a state of radical modernity, which in temporal terms can be described as a perpetual present. Having internalized democracy and technological progress, America ducks the question of originality, descent, and mythical authenticity. Everything is exactly what it appears to be in America: the real and the imaginary have been collapsed into the 'hyperreal,' – a constant state of simulation, of signs having escaped their referentials in the real. The concept of history as 'the transcending of a social and political rationality, as a dialectical, conflictual vision of societies,' is alien to Americans, Baudrillard observes: the United States is, in fact, a utopia achieved – a society which has behaved from the beginning as though it were already achieved.

American Society and American Studies

Baudrillard's incisive observations about American society are not unchallenged within the social sciences, but despite the provocative rhetoric, it cannot be denied that his concepts of the 'hyperreal' and the 'utopia achieved' are valuable tools in trying to account for some of the more paradoxical complexities of American society – its fragmented wholeness; its social stability in the face of an unequal distribution of wealth; its belief in 'a pleasing uniformity of descent' (Crèvecoeur 1782 [1981]) and a 'general happy Mediocrity' (Franklin, 1791–1798 [1986]) when the racial divide appears to be as wide as ever. Baudrillard's agenda – to bring the paradoxical nature of American society into a universal model while retaining the apparent incompatibility of its constituent parts – is certainly unique in the evolution of American studies. From its inception in the 1920s, the discipline of American studies has almost inescapably been concerned with the search for the common denominator in American culture and society: the very success of its mission – to establish and institutionalize a scholarly discipline whose object of study was the society of the United States – depended on the new discipline being able to achieve some sort of consensus about the nature of that object of study.

This mission first took off seriously with the publication of Vernon Parrington's *Main Currents in American Thought* (1927 [1958]), and was carried further with enthusiasm by the 'consensus historians' of the 1950s (including Richard Hofstadter, Louis Hartz, Arthur Schlesinger, Jr., Daniel Boorstin) and, most importantly (in terms of its impact on the field), by the 'myth and symbol' school of the 1950s (which included intellectual historians and literary critics such as Lionel Trilling, Henry Nash Smith, R. W. B. Lewis, and Leo Marx). The 'myth and symbol' scholars had indirectly been influenced by a host of literary critics in the 1930s and 1940s who were known collectively as the 'New Critics.' Approaching the literary text as an autonomous whole, the New Critics (which included such people as John Crowe Ranson, Kenneth Burke, Yvor Winters, and R. P. Blackmur) emphasized the transcendental qualities of the text: its moral content; its symbolic meaning as a reflection of national culture and identity; and its place and contribution to a national literary 'tradition.' F. O. Matthiessen's extraordinarily influential *American Renaissance: Art and Expression in the Age of Emerson and Whitman* (1941) established a 'great tradition' of nineteenth-century American authors that represented the 'quintessential' characteristics, concerns, and values of 'mainstream' American society. Inspired by the humanistic ideal of high culture, Matthiessen's canon of nineteenth-century American literature not only introduced a firm hierarchy of 'high' over 'low' artistic expression but also provided the raw cultural materials for the authors of the myth and symbol school. In this way Smith's thesis of the wilderness as the basic ingredient of American civilization, Lewis's insistence on tragic innocence in the myth of the 'American Adam,' and Marx's symbol of the 'machine in the garden' to account for the 'American' dichotomy of pastoralism and technological progress could become the unifying myths and symbols that dominated the study of American society during the 1950s and 1960s.

When, in the course of the 1960s, mainstream American society and culture increasingly came under attack from historically marginalized groups as part of the struggle for emancipation and recognition (women, African-Americans, Native Americans, Hispanics, gays, and other minorities), the theory and methods of American studies were also subjected to critical analysis and reappraisal. If pluralism replaced universalism as the new norm in society, so it did in American studies. Thus the new social history that emerged in the late 1960s and early 1970s began to demolish American exceptionalism, imperialism, and the consensus society, while it tried to recover or create the history of America's 'forgotten' minorities, as well as America's popular culture, and its labor and regional history. Drawing from the techniques and methods of sociology, anthropology, and other human sciences, gender, class, and race became the new 'holy trinity' of American studies.

Increasingly, also new criticism and the myth and symbol school came under attack for being ahistorical and for their philosophical idealism. With the emergence of the 'New Historicism' (ushered in by Stephen Greenblatt and the journal *Representations*) interdisciplinarity and historical contextualization were put firmly on the methodological agenda of American studies. This paradigm shift is illustrated by a seminal collection of essays edited by Sacvan Bercovitch and Myra Jehlen, *Ideology and Classic American Literature* (1988), which contains reassessments of their work by important pioneers of American Studies theory and method including Henry Nash Smith, Leo Marx, and Alan Trachtenberg, as well as essays by a younger generation of scholars, including Houston Baker, Carolyn Porter, Donald Pease, Michael Gilmore, Jane Tompkins, Jonathan Arac and Myra Jehlen – work that exemplified the scholarly output of the 1980s, blending feminist, neo-Marxist, and poststructuralist theories and methods.

However, whereas among the newly emancipated groups this pluralist approach to American society was experienced as liberating (both intellectually and politically), a growing number of scholars began to interpret the new trend toward plurality as undesirable. Christopher Lasch was one of the first to sound the alarm bell. In *The Culture of Narcissism* (1979 [1991]) he argued that the 'devaluation of the past' was symptomatic of a deep 'cultural crisis,' and he accused the 'demoralization of the humanities' of having reduced American 'individualism' to 'a narcissistic preoccupation with the self.' With the publication of Allan Bloom's *The Closing of the American Mind* (1987 [1988]) the 'battle of the books' erupted into an all-out 'culture war.' According to Bloom, the social/political crisis of twentieth-century America was really an intellectual crisis. Despite the use of such edifying labels as 'individual responsibility, experience, growth, development, self-expression, liberation and concern,' Bloom argued, the 1960s had 'bankrupted' American universities and destroyed 'the grand American liberal traditions of education.' The battle over the canon of American literature was now openly a battle about the canon of American culture and society, about nationality and multiculturalism, about consensus and political correctness. The 1990s saw a whole spate of books calling for a truce and reconciliation, offering a wide array of solutions to the increasingly vicious national debate and the resulting ideological stalemate. In *Beyond the Culture Wars* (1992 [1993]) Gerald Graff showed how 'teaching the conflicts can revitalize American education,' while Henry Louis Gates, Jr. advocated in *Loose Canons: Notes on the Culture Wars* (1992) a new civic culture that would transcend the sharp divisions of nationalism, racism, and sexism. Schlesinger rather desperately proposed in *The Disuniting of America* (1991 [1993]) that Americans should attempt 'to vindicate cherished cultures and traditions without breaking the bonds of cohesion – common ideals, common political institutions, common language, common culture, common fate – that hold the republic together'; this appeal was picked up by David Hollinger, who argued in his *Postethnic America* (1995) that Americans should form 'voluntary' rather than 'prescribed affiliations' to bridge the ethnic and multicultural gaps in society.

There are no clear signs that the latest models of reconciliation have effected any kind of movement on the battleground of nationality and multiculturalism – which perhaps is not surprising since the majority of them propose as the solution to the threat of the radical implosion of the consensus model of American society a return to the very neo-humanist, essentialist ideals that are so vehemently being contested in the multiculturalism/pluralism debate. In the meantime, the discipline of American studies is as divided as the various participants in this debate. However, although there has been some degree of disciplinary backlash (with some academic programs having been reduced or cut from university calendars), most working in the field would concede that the questions about the future of American culture and society have never been as open as they are at the beginning of the twenty-first century, and that this in itself more than legitimizes the continuing study of American society within the interdisciplinary framework of American studies.

See also: Civil Society, Concept and History of; Frontiers in History; Integration and Multiculturalism: Focus on Western Europe; 32098; Multiculturalism; Pluralism; Tocqueville, Alexis de (1805–59).

Bibliography

Baudrillard, J., 1986 [1988]. America. Verso, London.
Bercovitch, S., Jehlen, M., 1988. Ideology and Classic American Literature. Cambridge University Press, Cambridge, UK.
Bercovitch, S., 1993. The Rites of Assent: Transformations in the Symbolic Construction of America. Routledge, New York.
Berthoff, R., 1971. An Unsettled People: Social Order and Disorder in American History. Harper & Row, New York.
Bloom, A., 1987 [1988]. The Closing of the American Mind: How Higher Education Has Failed Democracy and Impoverished the Soul of Today's Students. Simon & Schuster, New York.
Crèvecoeur, J H St J de, 1782 [1981]. Letters from an AmericanFarmer. Penguin, New York.
Franklin, B., 1791–1798 [1986]. Autobiography. Norton, New York.
Gates Jr, H.L., 1992. Loose Canons: Notes on the Culture Wars. Oxford University Press, New York.
Gover, G.E., 1948. The American People: A Study in National Character. Norton, New York.
Graff, G., 1992 [1993]. Beyond the Culture Wars: How Teaching the Conflicts Can Revitalize American Education. Norton, New York.
Hofstadter, R., 1948. The American Political Tradition and the Men Who Made It. Knopf, New York.
Hollinger, D.A., 1995. Postethnic America: Beyond Multiculturalism. Basic Books, New York.
Lasch, C., 1979 [1991]. The Culture of Narcissism: American Life in an Age of Diminishing Expectations. Norton, New York.
Lipset, S.M., 1963. The First New Nation: The United States in Historical and Comparative Perspective. Basic Books, New York.
Lipset, S.M., 1979 [1980]. The Third Century: America as a Postindustrial Society. University of Chicago Press, Chicago.
Lipset, S.M., 1996. American Exceptionalism: A Double-edged Sword. Norton, New York.
Matthiessen, F.O., 1941. American Renaissance: Art and Expression in the Age of Emerson and Whitman. Oxford University Press, New York.
Parrington, V., 1927 [1958]. Main Currents in American Thought. Harcourt & Brace, New York.
Potter, D.M., 1973. In: Fehrenbacher, D.E. (Ed.), History and American Society. Oxford University Press, London.
Rowlandson, Mary, 1682 [1994]. A True History of the Captivityand Restoration of Mrs. Mary Rowlandson. Penguin, New York.
Schlesinger Jr, A.M., 1991 [1993]. The Disuniting of America: Reflections on a Multicultural Society. Norton, New York.
Tocqueville, A de, 1835-9 [1945]. Democracy in America. Vintage, New York.
Tuveson, E.L., 1968. Redeemer Nation: The Idea of America's Millennial Role. University of Chicago Press, Chicago.
Wiebe, R.H., 1975. The Segmented Society: An Introduction to the Meaning of America. Oxford University Press, New York.

Americas, Sociocultural Overviews: Central America

Daniel Reichman, University of Rochester, Rochester, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C.A. Smith, volume 3, pp. 1619–1623, © 2001, Elsevier Ltd.

Abstract

The seven countries in the Central American subcontinent – Belize, Costa Rica, El Salvador, Guatemala, Honduras, Nicaragua, and Panama – have, to a certain degree, followed their own social and historical trajectories in the early years of the twenty-first century, but each nation has been indelibly transformed by globalization. This article explores recent changes in the political, economic, social, and cultural life of the region, with special emphasis on new forms of identity and citizenship that have emerged since the late 1980s.

The seven countries in the Central American subcontinent – Belize, Costa Rica, El Salvador, Guatemala, Honduras, Nicaragua, and Panama – have, to a certain degree, followed their own social and historical trajectories in the early years of the twenty-first century, but one can identify common patterns in their recent histories. Each Central American nation has been indelibly transformed by globalization, the integration of the world's nation-states into a single economic and cultural system that is strongly oriented toward the interests of transnational capital. While Central America has been tightly woven into the world economy since the colonial period, the past two decades have been marked by rapid, widespread changes in economics and politics – produced by globalization – that have reverberated through all aspects of society, profoundly reshaping everyday life throughout the region. As a result, in the past 20 years, social scientists have tended to study Central American sociocultural life within a global frame, describing how particular communities, regions, or states are transformed by macrolevel changes in political economy.

Cold War Legacies

In order to understand the impact of globalization on Central American society, one must first recognize that the present era emerged in the aftermath of revolutionary violence. Caught along the fault lines of the US-Soviet conflict, Central America in the 1980s was defined by war. The conflict began with the Cuban Revolution of 1959, which inspired revolutionary movements in El Salvador, Guatemala, and Nicaragua. In each country, a small group of guerillas sought to overthrow corrupt and oligarchic political regimes by force. While the revolutionary fronts were inspired by different ideologies, including Maoism, Soviet communism, and other variants of postcolonial revolutionary thought, they all sought to build support among the rural poor to overthrow the state to bring about a more equitable social order. In response to the threat of leftist revolution in its own backyard, the United States government supported military action against the revolutionaries, providing weapons, training, and financial support to dictatorial regimes in El Salvador and Guatemala. In Nicaragua, the Sandinista revolutionaries successfully overthrew the Somoza government in 1979 and remained in power until 1990. The United States supported and trained the Contra counterrevolutionaries, who carried out a guerilla war against the Sandinistas, who were supported by the Soviet bloc and Cuba.

Belize, Costa Rica, and Honduras did not directly experience large-scale war in the 1980s, but they were affected by it. Honduras, and, to a lesser extent, Costa Rica, were bases of operations for the Nicaraguan counterrevolutionaries. Foreign military and economic aid poured into the region as a strategy to ward off the threat of revolution, and refugees and displaced migrants flowed into all countries that bordered the war-torn regions. The Central American peace process began to take shape in 1987, led by Costa Rican president Oscar Arias, who was awarded that year's Nobel Peace Prize. Beginning with regional meetings held in Esquipulas, Guatemala, the governments of Guatemala, Nicaragua, and El Salvador created a framework to end the wars and establish democracy. Nicaragua held elections in 1990, leading to the electoral defeat of the Sandinistas. The war in El Salvador ended with a peace agreement in 1992, and Guatemala's war ended with the signing of Peace Accords in 1996.

Since the end of the wars, each country has been forced to come to terms with the legacy of mass violence. In El Salvador, more than 75 000 people lost their lives during the war, and a million people were displaced. In Guatemala, tens of thousands of people – mostly indigenous Maya – were killed in a war that was called genocide by the United Nations. Anthropologists have focused on the complex negotiations of blame and healing that mark postconflict societies. In Guatemala, Diane Nelson (2009) and Victoria Sanford (2004) describe the lingering sense of horror that has emerged as the country comes to terms with mass violence, and where, despite the presence of truth commissions, some of the worst abusers of human rights continue to live with impunity. In El Salvador, which is sometimes described as a model of postconflict stability and democracy, Ellen Moodie (2010) has shown how postwar society remains beset by violence and insecurity that seems difficult to understand and explain than what occurred during the Civil War. While peace has officially come to El Salvador and Guatemala, both countries are still working to heal a traumatized body politic.

Within anthropology, the struggle to come to terms with the violence of the 1980s led to the so-called Rigoberta Menchú Controversy (Arias, 2001). The controversy began in the late 1990s when anthropologist David Stoll published a book

(Stoll, 1999) that used ethnographic and historical research to disprove some of the claims made in *I Rigoberta Menchú*, a world-famous biography of Mayan activist Rigoberta Menchú Tum, written by Elisabeth Burgós-Debray in 1982. *I Rigoberta Menchú* played an invaluable role in raising the world's awareness about the atrocities committed by the Guatemalan military against the Maya. Menchú became a worldwide symbol of indigenous resistance, traveled around the world to speak, and was awarded the Nobel Peace Prize in 1992. Stoll's research called into question some of the details of Menchú's account of her life. At a broader level, his book argued that urban intellectuals in Europe and the Americas romanticized Latin American revolutionary movements, turning a blind eye to some atrocities that the revolutionaries committed against the rural poor that they claimed to help. Although Stoll did not intend to criticize the larger truth of Menchú's story, he argued that scholars had glossed over important complexities of the revolution due to their sympathies for the revolutionary cause. The book made headlines around the world, and was used by critics to discredit Menchú. The entire affair instigated an emotionally charged debate about the politics of anthropological knowledge: Is scholarly objectivity desirable or possible in a situation of violent conflict? And should the value of scholarship be measured by its political effects, or by a disinterested standard of truth? In hindsight, the Menchú controversy played out as a scholarly analog to the truth commissions, in that it generated a difficult – and unresolved – conversation about the role and responsibility of scholars in a complex and violent world.

From Revolution to Remittances

As the revolutionary period subsided, scholars who had focused almost exclusively on the causes and consequences of war in Central America in the 1980s began to focus on new transformative changes in the region, none more significant than the rapid increase in migration to the United States from El Salvador, Guatemala, Honduras, and Nicaragua. It is not a coincidence that the countries that were most impacted by war became the major sources of migration to the United States. Between 1980 and 2010, the number of Central Americans living in the United States increased by at least a factor of 20. Prior to the 1970s, Costa Ricans and Panamanians composed the largest percentages of total Central American migration to the United States, which was still relatively small, approximately 100 000 in 1970. In the late 1970s, displaced people and refugees fleeing El Salvador, Guatemala, and Nicaragua began to enter the United States in large numbers, and in the 1980s, the number of Central Americans entering the United States tripled that of the previous decade. As of 2010, approximately 1.2 million people born in El Salvador were living in the United States, along with 800 000 from Guatemala, 500 000 from Honduras, and 250 000 from Nicaragua.

The so-called Central American diaspora has transformed economy, culture, and society throughout the region. By 2011, migrant remittances accounted for almost 20% of GDP in Honduras, 15% in El Salvador and Nicaragua, and 10% in Guatemala. The social and cultural changes brought about by migration were just as significant as its economic impact. Rural communities that had been historically dependent on agriculture came to rely on remittance payments from young men and women in the United States. Formerly isolated towns and villages became socially and culturally connected to cities and towns in the United States, leading to rising consumerism, changes in family dynamics, and new concepts of status and value. Migration has also transformed the relationship between rural and urban areas around the region, turning impoverished rural communities into sites of economic growth. In El Salvador, Sarah Mahler (1995) has described the chain of migration that developed between suburban Long Island and the rural department of La Unión in the 1980s. David Pedersen (2013) describes the rapid transformation of town of Intipucá, also in La Unión, as a result of large-scale migration to Washington, DC that began in the 1980s. Patricia Foxen (2007) describes a similar process that took shape in Guatemala, where Mayans from the community of Xinxuc began to emigrate to Providence, Rhode Island, in the 1980s, creating a transnational circuit that has endured for decades since. In Honduras, which has become the most remittance-dependent country in the region, the boom in migration to the United States began later than it did in El Salvador, Guatemala, and Nicaragua, caused mainly by economic factors (such as a sharp decline in world coffee prices in the early 2000s) and the after effects of Hurricane Mitch, which devastated the country in 1998 (Reichman, 2011). Despite these differences, migration has transformed Honduran society in ways that are quite similar to what has occurred in El Salvador and Guatemala.

Central American politicians have responded to the newfound influence of migrants. Across the region (particularly in El Salvador and Honduras) hometown associations, groups of organized migrants abroad, have supplied funds for municipal development in their native countries. A host of programs has been developed to lessen the expense of remittance payments for senders and recipients, in order to keep more capital 'in the pockets' of remittance-dependent families. NGOs, governments, and international aid organizations have created pilot programs to supplement the work of hometown associations with public funds in order to support rural development. Noting the direct impact that small amounts of remittances can have on family well-being, El Salvador, Guatemala, Honduras, and Costa Rica have all created conditional cash transfer programs to give direct economic assistance to poor families.

While migration and remittances have provided a much-needed source of income to rural families, they have also created a new form of dependency on the United States. Central American economies have long been economically dependent on the United States, the region's most important trading partner and source of foreign investment, but migration-dependency has added a new wrinkle to this enduring pattern. Changes in US immigration law now directly affect families throughout Central America, and the fortunes of rural communities are tied to the vicissitudes of immigration politics in the United States. In 1986, for example, the United States passed the Immigration Reform and Control Act (IRCA) which legalized more than two million unauthorized immigrants. This law provided a way for the relatives of hundreds of

thousands of Central American migrants to enter the United States legally under family reunification visas. The 1986 law also greatly intensified the enforcement of immigration law, including border security, and this crackdown on unauthorized immigration has continued for the past 25 years, creating a sophisticated, militarized border security complex called the 'deportation regime' (De Genova and Peutz, 2010).

The deportation regime, which reached its peak in the years after the attacks of 11 September 2001, had far-reaching impact on Central America. Elana Zilberg (2011) describes how the deportation of El Salvadoran gang members from Los Angeles led directly to the growth of organized crime in El Salvador. Gangs such as the MS-13 have gained infamy around the world, and their existence is directly tied to the illicit trade in narcotics across the Americas, as well as the smuggling of unauthorized immigrants through Central America and Mexico. At a broader level, the demand for immigrant smugglers to help desperate people across a militarized border has created a complex web of corruption, organized crime, and exploitation that reaches across the Americas.

Rates of violent crime have risen dramatically in El Salvador, Guatemala, and Honduras in the 2000s. Gangs (called maras) have become a specter of disorder and criminality across Central America, and fear of gangs has helped to drive the popularity of rightist governments that claim to restore 'law and order' to a body politic threatened by 'delincuencia.' In Honduras and El Salvador, antigang measures known as mano dura (iron fist) have led to brutal crackdowns against suspected gang members, measures that involve suspension of due process and basic civil rights. While the threat of gang violence is real, it has revived some of the authoritarian measures that were used against suspected political enemies under military dictatorships. The hardline tactics have also led to severely overcrowded prisons, which have been home to several violent prisoner riots, most notably in Honduras, where a suspicious prison fire in the provincial town of Comayagua killed 360 people in 2012.

Neoliberalism and Its Opponents

If the 1980s were the decade of war in Central America, the 1990s were the decade of neoliberal globalization. From about 1950 to the late 1980s, governments sought to gain control over national economies by promoting domestic industry and creating policies that would keep profits from leaving Central America. These policies – sometimes called protectionism or economic nationalism – were viewed as an alternative to the neocolonial exploitation that had shaped the region in the modern era. For much of the twentieth century, in places like Honduras, Guatemala, and Costa Rica, foreign capital was often equated with export industries like banana cultivation, mining, and coffee production, which created severe inequality, corruption, and environmental damage. Neoliberalism marked a sharp turn away from economic nationalism, in that it welcomed foreign corporations as agents of economic growth, rather than exploitation. Neoliberal doctrine promoted privatization and openness to foreign trade. It directed governmental expenditures toward areas that would improve a country's competitiveness in the global economy, prioritizing economic growth over social stability. Under the guidance of the World Bank and the International Monetary Fund, neoliberal policy became the most popular recipe for economic growth in Central America in the 1990s.

The most apt symbols of neoliberalism in Central America are the free trade manufacturing zones that have popped up in every country in the region, designed to create economic growth by attracting foreign direct investment. In some cases, such as Costa Rica and Panama, this strategy has yielded notable benefits. In the former, the high-tech manufacturer Intel has become the largest company in the country. In the latter, the free trade zone along the Panama Canal, under Panamanian control since 1999, is one of the world's most important export manufacturing hubs. In Honduras, Guatemala, Nicaragua, and El Salvador, free trade zones are referred to as *maquilas* or *maquiladoras*. They are specially zoned industrial districts that house foreign-owned manufacturing plants, which often specialize in the manufacture of textiles, housewares, and footwear for export, mainly to the United States. These facilities, disparaged as 'sweatshops' in the United States and Europe, attract foreign capital by offering a stable, productive, and inexpensive labor force, a 'business-friendly' political climate with low or nonexistent corporate taxes, highly subsidized infrastructure (roads, power supply, and ports), and weak labor unions. Maquilas attract a young, predominately female, workforce, many of whom come from rural areas to seek stable employment.

Maquilas have provided an economic alternative to farm labor or unemployment for people throughout Central America, but their low wages and harsh working conditions have made them targets of criticism. For example, the National Labor Committee (NLC), a labor solidarity movement in the United States, brought to light labor abuses (including child labor, intimidation, and nonpayment of wages) in Honduran maquilas. The NLC produced a scandalous documentary called *Zoned for Slavery* in 1994, and in 1996, Charles Kernaghan, the NLC director, testified before Congress about some of the labor violations he had seen in Honduras. His testimony led to the public shaming of television celebrity Kathy Lee Gifford, who owned a clothing line that employed children in the Honduran maquilas that Kernaghan had visited. Following this affair, Central American maquilas became targets of international activist campaigns, including UNITE! and the United Students Against Sweatshops (Pine, 2008). Faced with foreign competition from China and Southeast Asia, wages in maquilas have not risen significantly. Since the 1990s, increased national and international monitoring of working conditions has led to some improvements.

The growth in the maquila sector was part of a broad strategy to diversify Central American economies, cutting down their reliance on export monoculture of traditional crops like coffee, bananas, and timber. In 1989, the International Coffee Agreement (ICA) was dismantled, leading to the deregulation of world coffee prices, which had been controlled by treaty since the 1960s. The end of the ICA introduced a new level of risk into the coffee economy, and countries like Guatemala and Costa Rica, which were highly dependent on coffee exports, sought to develop new export industries. Honduras, which had been a 'sleeping giant' in the world coffee market, became Central America's largest coffee producer in the late 2000s.

Tourism also became a high-growth industry in the 1990s. Costa Rica was the first tourist success story in the region, famous for attracting adventurous, environmentally oriented visitors to its beaches and rainforests. Nicaragua, Panama, and Belize have also become tourist hotbeds, and El Salvador, Guatemala, and Honduras have invested heavily in coastal developments to attract foreigners.

The common thread that connects maquilas, export agriculture, and tourism is that all three industries rely strongly on the US consumer market, and to critics, they recreated the patterns of dependency that marked the twentieth century. Viewed as one model of economic progress, neoliberalism prioritized GDP growth over more collective visions of social justice. Within the political sphere, a similar process of ideological narrowing took place. During the peak decade of neoliberalism (1994–2003), nongovernmental organizations (NGOs) became influential players in the political life of the region, filling the void left by the shrinking role of states in environmental policy, human rights, health, and labor issues, particularly in the countries in the region that were most dependent on foreign economic aid (El Salvador, Nicaragua, Honduras, and Nicaragua). NGOs tended to rely on international donors and foreign leadership, and, although they have led many important reforms around the region, they have been criticized for holding a somewhat narrow and opportunistic political philosophy, focusing their energies on single issues that can attract foreign funding, rather than systematic visions of social justice.

By the mid-2000s, the 'neoliberal experiment' had not created the broad based economic growth that its proponents had promised, and people around Central America began to call for changes that would stimulate the domestic economy and provide a more inclusive vision of social progress. A new form of antineoliberal politics emerged, most notably in Nicaragua and Honduras, where Presidents Daniel Ortega (elected in 2006) and Manuel Zelaya (elected in 2005) espoused a new form of populist nationalism, explicitly defining themselves as opponents of neoliberalism. Ortega and Zelaya both signed agreements to join the Bolivarian Alternative for the Americas (ALBA), an international trade agreement led by Venezuela and Cuba, which sought to provide an alternative to the neoliberal model of free trade, epitomized by the CAFTA-DR, the Dominican Republic – Central America Free Trade Agreement (signed in 2005), and the proposed Free Trade Area of the Americas, a plan to reduce barriers to trade across the Americas. Free trade, the central pillar of neoliberal policy, came to be defined as a new kind of imperialism, and movements like ALBA symbolized a rejection of neoliberal doctrine. In Honduras, Zelaya's entry into ALBA, along with an attempt to significantly raise the Honduran minimum wage, were viewed by the business community as a threat to the global competitiveness of Honduran industry. Zelaya's antineoliberal policies began a political crisis that ultimately led to a coup d'etat that forcibly removed him from power in June 2009.

Multiculturalism, Religious Change, and New Forms of Citizenship

Since the late 1980s, the politics of multiculturalism have significantly changed concepts of national identity and belonging in Central America. Throughout the nineteenth and twentieth centuries, groups that were ethnically, linguistically, racially, and/or culturally distinct from national majority were often stigmatized and subjected to official policies of discrimination. Indigenous people were targets of assimilation programs, exploitation, and violence and were sometimes seen as impediments to modernization and 'civilization.' By the end of the twentieth century, people around the world had come to embrace ethnocultural difference as a mark of democratic openness and tolerance. This change created new political spaces for cultural groups to make demands against the state. The era of 'multiculturalism' was exemplified by Rigoberta Menchú's 1992 Nobel Peace Prize and the United Nations' declaration of the years 1995–2004 as the first International Decade of the World's Indigenous People.

In Guatemala, the Pan-Mayan movement arose in the aftermath of genocidal war. The movement embraced Mayan languages, styles of dress, and religion, which had been suppressed or stigmatized for much of the county's history (Warren, 1998). Mayan activists gained influence in politics, calling for affirmative action programs and targeted development aid, forging strong alliances with international NGOs (Fischer and Benson, 2006). While the Maya are still subject to racial discrimination, indigenous activism has led to more inclusive notions of citizenship. At the same time, the government has used colorful, exotic images of Mayan 'authenticity' to brand itself to foreign tourists, using the recent shift toward cultural tolerance as a marketing tool. The new forms of cultural capital that have emerged within the tourist economy have generated complex questions about what it means to be Maya in multicultural Guatemala, and shifting notions of ethnic identity have become a topic of great interest for anthropologists.

Multiculturalism in Honduras has followed a somewhat different course. The Garifuna people, descendents of residents of the island of St. Vincent who were deported to Central America in the late eighteenth century, have developed a new ethnic consciousness over the past two decades (Anderson, 2009). The Garifuna, formerly called the Black Caribs, define themselves as both 'black' and 'indigenous,' and take part in political movements associated with indigenous rights and the Black diaspora in the Americas. The Garifuna have aligned themselves with other Honduran indigenous political organizations to make demands for land rights, often in opposition to tourist developments near their homes on the Caribbean coast. Like the Maya in Guatemala, the Garifuna have been used as a symbol of Honduras' cultural diversity in the county's attempt to attract foreign tourists. More broadly, the Garifuna have become part of the Honduran national imaginary as it recasts itself from a predominately mestizo society to a multiethnic nation.

Along with ethnicity, the relationship between religion and national identity has also undergone important shifts since the 1980s. Roman Catholicism was by far the dominant religion in Central America from the colonial period until the second half of the twentieth century, when evangelical Protestantism began an explosive period of growth. The rise of evangelical religion in the region began in Guatemala, where more than 60% of the population now identifies as protestant (O'Neill, 2010: p. 11). A similar pattern has taken shape in Costa Rica, Honduras, El

Salvador, and Nicaragua, which are all home to significant evangelical populations that have grown rapidly in the past few decades. Evangelical movements in Central America tend to embrace independence from top-down religious bureaucracies and have a flexible stance toward doctrine. As O'Neill argues, they have a certain affinity with neoliberal democracy, in that they celebrate the power of the individual to transform social life through moral behavior, eschewing more collective visions of social change. The evangelical turn is especially striking when viewed alongside the declining popularity of liberation theology, a form of Catholicism that was influenced by Marxist visions of social justice, which was quite influential in the peasant movements 1960s and 1970s. Whereas liberation theology and related forms of social Catholicism embraced collective political action, the new evangelical movements emphasize the power of the moral individual to bring about a better world. The popularity of evangelical religion has begun to feed back on the Catholic Church itself, where the Catholic Charismatic Renewal (CCR) movement has provided an alternative to the traditional Catholicism that involves spiritual healing, ecstatic prayer, and a direct relationship between the worshipper and divinity.

The politics of gender and sexuality are another dimension of social life that has undergone significant change since the 1990s. Women have taken on high-level political positions throughout Central America, with Nicaragua, Costa Rica, and Panama all electing female heads of state since 1990. In countries where manufacturing and the service-sector work have replaced agriculture, women have gained economic independence, often becoming the primary wage earners in households. The rising economic status of women has challenged cultural notions of masculinity, leading some men to "participate in alternate economies to earn the money and respect that is denied them within the current legal economy" (Pine, 2008: p. 33). Across the region, birth rates have fallen as women find work outside of the home and the demand for agricultural labor declines. Another notable change in gender politics has been the rise of LGBT identity around the region. As Florence Babb (2003) and Cymene Howe (2008) describe in Nicaragua, demands for sexual rights emerged within the broader discourse of multicultural tolerance in the 1990s. While LGBT groups continue to be targets of discrimination (particularly in Belize, where same sex sexual activity remains illegal) public awareness of sexual rights has certainly improved.

While the anthropology of Central America has widened its focus since the end of the 1980s, there are still some areas of social life that remain understudied. The contemporary livelihoods of indigenous groups like the Lenca, Pech, Jicaque, and Guaymí are not well known outside of their home countries. Few anthropologists have produced substantial ethnographies of elites or professionals, particularly in Costa Rica and Panama, the most prosperous countries in the region. In general, studies of the middle classes are far less common than studies of the poor, and there are relatively few studies of Central American education, media, and expressive culture. In general, scholarship on the region has focused on the broad themes of political economy, postconflict justice, ethnicity, religion, and the environment. There is good reason for this: In the United States and Europe, much of the academic interest in the region developed in response to the violence of the 1980s, and has been closely tied to activist movements. As the 1980s fade from view, new social changes have taken root throughout Central America, and we can look forward to new directions in social scientific studies of the region that will incorporate voices and viewpoints from an increasingly global community of scholars.

See also: Christianity in Latin America; Cold War, The; Evangelical, Revivalist, and Pentecostal Religions; Globalization and Religion; Globalization and World Culture; Globalization, Anthropology of; Historical Thought and Historiography: Indigenous Cultures in the Americas; International Migration and Remittances; Multiculturalism, Anthropology of; Neoliberalism.

Bibliography

Anderson, M., 2009. Black and Indigenous: Garifuna Activism and Consumer Culture in Honduras. University of Minnesota Press, Minneapolis.

Arias, A. (Ed.), 2001. The Rigoberta Menchú Controversy. University of Minnesota Press, Minneapolis.

Babb, F., 2003. Out in Nicaragua: local and transnational desires after the revolution. Cultural Anthropology 18, 304–328.

De Genova, N., Peutz, N. (Eds.), 2010. The Deportation Regime. Sovereignty, Space, and the Freedom of Movement. Duke University Press, Durham, NC.

Fischer, E.F., Benson, P., 2006. Broccoli and Desire: Global Connections and Maya Struggles in Postwar Guatemala. Stanford University Press, Palo Alto.

Foxen, P., 2007. In Search of Providence: Transnational Mayan Identities. Vanderbilt University Press, Nashville.

Howe, C., 2008. Spectacles of sexuality: televisionary activism in Nicaragua. Cultural Anthropology 23, 48–84.

Mahler, S.J., 1995. American Dreaming: Immigrant Life on the Margins. Princeton University Press, Princeton.

Moodie, E., 2010. El Salvador in the Aftermath of Peace: Crime, Uncertainty, and the Transition to Democracy. University of Pennsylvania Press, Philadelphia.

Nelson, D.M., 2009. Reckoning: The Ends of War in Guatemala. Duke University Press, Durham, NC.

O'Neill, K.L., 2010. City of God: Christian Citizenship in Postwar Guatemala. University of California Press, Berkeley.

Pedersen, D., 2013. American Value: Migrants, Money, and Meaning in El Salvador and the United States. University of Chicago Press, Chicago.

Pine, A., 2008. Working Hard, Drinking Hard: On Violence and Survival in Honduras. University of California Press, Berkeley.

Reichman, D., 2011. The Broken Village: Coffee, Migration, and Globalization in Honduras. Cornell University Press, Ithaca.

Sanford, V., 2004. Buried Secrets: Truth and Human Rights in Guatemala. Palgrave Macmillan, New York.

Stoll, D., 1999. Rigoberta Menchú and the Story of All Poor Guatemalans. Westview Press, Boulder, Colorado.

Warren, K., 1998. Indigenous Movements and Their Critics: Pan-Mayan Activism in Guatemala. Princeton University Press, Princeton.

Zilberg, E., 2011. Space of Detention: The Making of a Transnational Gang Crisis between Los Angeles and San Salvador. Duke University Press, Durham.

Americas, Sociocultural Overviews: Mexico

Paul Liffman, Centro de Estudios Antropológicos, El Colegio de Michoacán, Michoacán, Mexico

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mexican anthropology, long wedded to a national modernization project that required 'backward' indigenous cultures to assimilate culturally, began distancing itself from the state after 1968. Thanks to renewed international dialogue, Mexicanists – both national and foreign – have critiqued political and economic power relations, revalorized ethnic identity and agency, and deconstructed Mexican modernity by questioning the consistency and locus of national identity within official boundaries. New studies of migration, cities, territorial claims, the state, media, and language embody these paradigm shifts. The older, community-centered tradition of Mesoamerican studies has also reinvigorated its Boasian roots with contemporary cultural and ontological theory.

The Indigenista Legacy

Mesoamerica (comprising most of Mexico, Belize, and Guatemala) was the most heavily populated region of pre-Columbian America. Mexico's cultural heritage – including major archaeological sites and some 60 indigenous languages still spoken to varying degrees – has attracted anthropological attention from the government and foreign researchers since the nineteenth century. Now the country has a globalizing, sometimes violent economy, one of the world's largest cities, and a population of 115 million – 15 million of whom identify as indigenous (once again the most in the western hemisphere). Consequently it is a major site for studying modernity, tradition, sociopolitical change, and *mestizaje* (cultural dislocation and recombination), and it is impossible to talk about the anthropology *of* Mexico without analyzing the impact of anthropology *in* Mexico ever since the dawn of the discipline. Turn-of-the-twentieth century foreign sociocultural anthropologists who had a major impact included the Berlin school of ethnology's Konrad Theodor Preuss and the founder of modern American anthropology, Franz Boas (Columbia University), followed in the 1920s by Robert Redfield (University of Chicago). Yet these men all depended on Mexican interlocutors, some of whom also influenced anthropology's development in Mexico and beyond.

When Boas set up the International School of American Archaeology and Ethnology at Mexico's National Museum of Anthropology during the 1910 Revolution, about 15% of the country spoke an indigenous language (vs about 5% now). As a historicist and cultural relativist, Boas was primarily interested in documenting that diversity and the prehispanic legacy. However, to technocrats eager for Mexico to join the top tier of developed nations, that same diversity entailed social, economic, and political burdens that social anthropology ought to ameliorate. They also wanted archaeology to build the historical legitimacy of a state far more taken with France and the US than the communal agrarian societies of the ancient past and indigenous present.

Boas's student Manuel Gamio initiated postrevolutionary Mexican anthropology in the 1920s, and for 40 years his successor Gonzalo Aguirre Beltrán, *indigenismo's* central intellectual and administrator, directed the mission of assimilating 'backward' indigenous communities into dynamic multiethnic regions from the Instituto Nacional Indigenista (INI), while the Instituto Nacional de Antropología e Historia (INAH) managed the nation's cultural heritage (cf González y González, 1997). However, after the government crushed the 1968 student movement, the legitimacy of official institutions eroded. Guillermo De la Peña (1999) describes the subsequent transformation of nearly all Mexican sociocultural anthropology and compares it to other Latin American models (2005). Revisionism rocked the discipline as disaffected young scholars engaged Marxism and structuralism to consider indigenous historical agency and cultural dynamism. As Manchester social anthropologist John Gledhill put it,

> there is a fundamental difference between an anthropology that is about people and the worlds of meaning that they construct through experience, action and social interaction – always within a framework of evolving power relations – and an anthropology that is about how reified 'cultures' themselves 'live' and 'die'. For much of its history, 'what they call Mexican anthropology', to revisit the critique of the generation of 1968, was an example of the latter kind of anthropology. The strong concern of national anthropologists for practical issues of social welfare is perhaps the side of that old project which it is still worth recuperating (in a less authoritarian and paternalist manner).
>
> (2008: 18)

In addition to this agenda, well mapped by José Luis Escalona, another worthwhile legacy of Aguirre Beltrán is the ongoing vitality of regional studies, as evidenced in the work of De la Peña, Eric Van Young, Daniel Hiernaux, and early Claudio Lomnitz (1992), among others.

Still, the INAH carries on and the INI's twenty-first century neoliberal heir, the Comisión Nacional para el Desarrollo de los Pueblos Indígenas, has at least a shadow of its forebear's role in rural development. The INAH project *Etnografía de las regiones indígenas de México en el nuevo milenio* (e.g., Millán and Valle, 2003) exhibits an ongoing concern with Mesoamerican social structure, community organization, ritual territoriality, identity, interethnic relations, religious norms, and migration, as do the heirs of large community-centered projects run by Harvard, Chicago, Tulane, and other US universities in the

1950s and 1960s (e.g., Evon Vogt's *Zinacantan* and Monaghan, 1995) (see Section Community Studies below).

Interdisciplinary and International Connections

Interdisciplinary relations are key to understanding sociocultural research in Mexico. Because of anthropology's historical importance to the national modernization project, sociology is less institutionally prominent than in the United States or South America, excepting a few major institutions like the Universidad Autónoma Metropolitana-Iztapalapa (UAMI), the Universidad Nacional Autónoma de México (UNAM), and El Colegio de México (Colmex) (plus the mass-mediated commentariat). However, this permitted relatively unfettered sociologists (many of them 1970s Argentinian political refugees) to employ more contemporary theory and contribute original research. Historians have enjoyed the greatest prestige and much social anthropology, influenced by Eric Wolf and others, seeks to take a historical approach to power, even as most university and graduate programs allow almost no interdisciplinary study.

Not only are disciplinary boundaries not the arbiter of anthropological relevance, but shared concerns have jumped the fences between 'subfields' (which raises doubts about using them to organize this article except insofar as institutional boundaries themselves provoke new ideas). Moreover, defining a 'national tradition' is complicated by anthropologists' diverse geographical origins, sites of training, and fieldwork locations. With *indigenismo* fading out after 1968, more Mexican anthropologists won federal support for pursuing PhDs, fieldwork, and research networks outside Mexico or for in-country training in international-level doctoral programs with broad theoretical agendas. Here the Centro de Investigaciones y Estudios Superiores en Antropología Social (CIESAS, www.ciesas.edu.mx), UAMI (www.uam-antropologia.net/index.php?option=com_content&view=article&id=82:publicaciones-recientes&catid=9&Itemid=495), Universidad Iberoamericana (www.ibero.mx/web/site/tpl-Nivel2_pos.php?menu=mgPosgrado&seccion=d_antropologiaSoc), Colegio de México (http://ceaa.colmex.mx/index.php/informacion-academica/218), and Colegio de Michoacán (www.colmich.edu.mx/index.php/docencia-cea/planta-docente) stand out. Still, in part because of funding limits most Mexican anthropological fieldwork remains in-country, with only a handful of programs like the Colmex's Centro de Estudios de Asia y África sending students beyond the Americas. More importantly, anthropology's post-1968 concern with power, political economy, and social justice in Mexico should continue as long as the country confronts the dilemmas of uneven development in both urban and rural contexts.

The History of Anthropology and Nation Building

Taking account of the international dynamics in a period of increasing diversification, the Antropología del Sur initiative, identified with Esteban Krotz in Mexico and the late Robert Kemper in the US, drew on Immanuel Wallerstein to substitute a global North–South over East–West analysis of the discipline's development. It has honed a critical, postmodernization – if at times staunchly antipostmodern – Mexican anthropology (Krotz, 1997; Gledhill, 2011).

Some historians of Mexican anthropology are now revising post-1968 revisionism. Engracia Loyo and more recently Elsie Rockwell (2007) have described the central role of primary education in revolutionary state formation. Reappraising *indigenismo*, they have rediscovered progressive aspects of its rural education programs even as they keep questioning the revolution's prime directive to assimilate ethnic minorities to the mestizo project. Their legacy includes Marco Calderón's work on revolutionary-era educational experiments; Alexander Dawson on imaginaries of the indigenous; Emiko Saldívar's ethnography of waning *indigenista* practice; Stephen Lewis, who has collaborated with Mary Kay Vaughan, the preeminent historian of intellectuals and the state in Mexico; and, inspired by Florencia Mallon's notion of nineteenth century popular liberalism, Ariadna Acevedo on popular constructions of citizenship (see Calderón, 2012, for many of these authors). Parallel to education research is the work of historians of science like Mechthild Rutsch, Laura Cházaro, Miruna Achim, and Frida Gorbach on the role of scientific apparatuses (from measuring instruments to museums) in constructing Mexican national and racial identities.

Political Economy, Ecology, and Brokerage

Central to Mexican anthropology's concerns with regional organization and the state is brokerage, possibly its most consistent theme ever since Eric Wolf's classic formulations (1956). Wolf's influential materialist synthesis emerged from collaborations with Mexican researchers, especially Ángel Palerm, who went on in the 1970s to develop a neo-evolutionist Marxist cultural ecology focused on the relationship between political power and control over natural resources (particularly land and water). Even leading non-Marxists like Norman Long, whose social anthropology has emphasized small group interactions in transnational labor contexts, also rely on Wolf's basic model. Characteristically, in both backward rural areas and urban communities dependent on clientelistic connections to political and economic authorities, access to resources is mediated through the *cacique* or boss (see the work of José Eduardo Zárate) and the ritually inflected forms of land tenure, historical memory, and social organization in which power is materialized (e.g., Jorge Uzeta's research on Atarjea, Guanajuato). Since the 1960s, Paul Friedrich's Michoacán corpus (1986) has set the standard on *caciquismo* and violence, and James Greenberg's ethnographies of indigenous Oaxaca (1989) continue addressing Friedrich's and Wolf's concerns. Fernando Salmerón's review article on the *caciquismo* literature (1984) is still valuable as is Wil Panster's more recent volume (2005).

Political economy and hegemony have remained key concerns for other scholars influenced by Wolf, like William Roseberry, Andrew Roth, Gavin Smith, Juan Pedro Viqueira, and Josiah Heyman. In particular, Roseberry's early work on commodity chains and his later incursion into culture theory (1989) inspired younger Mexicanists like Ricardo Macip Ríos and Casey Walsh. Anthropological historian Paul Eiss (2010) has brought new spatial and temporal perspectives to

political economy, ethnic identity, and historical memory in Yucatán.

More broadly, although the country is now 80% urban and 90% mestizo, the crisis of *ejidos* (collective land grants) and indigenous communities, especially in rural southern and central Mexico, still commands popular and scholarly attention because of the centrality these forms of land tenure have to the revolutionary project and national identity. After the Ejército Zapatista de Liberación Nacional's antiglobalization rebellion of 1994, multiculturalism as a basis for territorial and human rights claims became a major line of research (Sieder et al., 2002). This orientation undergirds activist studies of social movements by figures like Víctor Toledo who emphasize indigenous ecological sensibilities. David Barkin and Hernán Salas use *la nueva ruralidad* to emphasize globalization's role in placemaking (Torres, 2012). Others like Héctor Díaz Polanco combine traditional Marxist class analysis with attention to ethnicity and regional autonomy initiatives.

Recent ethnographies analyze new indigenous regional sovereignty claims (Liffman, 2011), critique the wave of transnational mining activity (Claudio Garibay), and study 'food sovereignty' opposed to genetically modified transnational crops, especially maize (Elizabeth Fitting). The current proliferation of private security forces, community *autodefensas*, and *narco* shadow states proclaiming autonomy and justice have darkened the optimistic pictures that emerged after 1994. Instead, scholars inspired by Das and Poole (2004) highlight uneven geographies of power on the margins of the neoliberal state (e.g., Froylán Enciso http://nuestraaparenterendicion.com/index.php/blogs-ok/weary-bystanders). Legal anthropology has gone beyond examining juridical processes for indigenous populations to question the state apparatus itself (Sierra, 2004; Sieder, 2013).

Within the post-1968 critique of *caciquismo*, state power and neoliberalism, drug trafficking is one of the most gripping subjects. Carlos Flores's project *Historias de polvo y sangre* traces drug smuggling in Tamaulipas back to the 1940s and finds links to the highest echelons of the Mexican government, whereas Salvador Maldonado's historical ethnography traces drug production in Michoacán to postrevolutionary modernization programs that failed because (nonillicit) brokerage between rural communities and the state broke down. Recent antecedents are Luis Astorga's work on *narcotraficantes* (2007) and Lomnitz's broader volume on corruption (2000). Narcoethnography is also traceable to earlier studies of urban violence and gangs (*bandas juveniles*) by Rosanna Reguillo and Néstor García Canclini.

Even though Marxism was highly influential here, Macip describes a 'watermelon' strategy since the rise of neoliberalism: leftist anthropologists discovered that turning 'green on the outside' was a way to survive institutionally, as at the Universidad Autónoma Chapingo. Indeed, a number of them became honeydews through and through. Recent work seeks a synthesis of political ecology, cultural property, and science and technology studies (STS), updating Palerm's cultural ecology tradition. Casey Walsh studies the hydraulic issue in northern Mexico, Andrew Mathews takes an innovative approach to forestry, Molly Doane takes on the contradictions of environmental governance, and Cori Hayden has pioneered bioprospecting research. Akin to STS's interest in risk, disaster studies like the CIESAS project inspired by Kenneth Hewitt and Ulrich Beck and directed by Virginia García Acosta critique state development projects. The approach has overcome the false naturalism and randomness once attributed to disasters and considers local interpretations as well, as in the opposition to mining, hydraulic, and wind energy development (Howe, 2014).

Indigenous Identity and Social Movements

Many of the aforementioned approaches echo the social movements literature developed by Jorge Aceves and Alain Touraine in the 1970s and subsequently advanced by Joe Foweraker and Ann Craig. Instead of fixed cultural identities, it found situated strategic alignments even as images of timeless Mesoamerican cultural continuity and resistance proliferated in the 1990s, thanks to Guillermo Bonfil's seminal *México profundo* (1996[1989]) and new political actors' growing reliance on identity politics.

The interest in identity ranges from conservatives who inscribe folklore traditions into the national cultural heritage to leftist sympathizers of Zapatista-inspired movements that demand national policies to benefit the rural sector and defend the country's cultural and ecological sovereignty. Both approaches are haunted by Gledhill's pungent observation that "almost all the assumptions that anthropologists of varying theoretical persuasions had made about the conditions that would foster reproduction of indigenous ethnicities or 'assimilation' in the form of mestizoisation have eventually been falsified by history" (2008: 5). Constructivist or antiessentialist approaches like Gledhill's and the social movements literature in general point to the fact that "the 'identities' that people use do not necessarily correspond to anthropological constructions of 'culture' and 'ethnicity'" (2008: 11). However, another legacy of *indigenista* authoritarianism is the apparently scant awareness or concern among some academics that such observations can erode the legitimacy of grassroots movements (see Briggs, 1996). Still, Gledhill himself takes a broad view:

> In seeking to revalorise their traditions and culture, even in a manner that distorts the historical record, indigenous activists have opened up national debates and challenged entrenched ideologies more effectively than academic anthropologists. This raises important questions about what the future agenda for academic anthropology should be in countries like Mexico.
> (2008: 15)

Indigeneity studies by authors like Michael Kearney and Federico Besserer build on Kearney's classic article with Carol Nagengast (1990) to examine how international migration foregrounds previously unmarked cultural identities, even as its globalizing nature blurs the previously national focus of Mexican anthropology.

Despite the attention paid to identity politics – and widely read studies on gender and indigeneity by researchers like Aída Hernández and Lynn Stephen (Speed et al., 2006) – at least

within the formal bounds of anthropology, gender and queer studies in Mexico remain limited (but see Prieur, 1998). The exceptions are often in cultural studies and sociology, such as the Colmex's Programa Interdisciplinario de Estudios de la Mujer and the UNAM's Programa Universitario de Estudios de Género. The larger exception within anthropology is the research on labor and violence, where *masculinidades* is a dynamic approach emerging from international dialogues (Gutmann, 1996).

Popular Religiosity

Akin to the social movements literature is the burgeoning work on popular religiosity, especially Latin America's rising Protestant wave. This research can be traced back to projects that emerged from the regional studies tradition in the 1980s. Most prominent were Elio Masferrer's (2000, 2004) Asociación Latinoamericana para el Estudio de la Religión, CIESAS's Religión y Sociedad en el Sureste (under well-known regionalist Andrés Fábregas), and the pan-Latin American Comisión para el Estudio de la Historia de la Iglesia en América Latina y el Caribe (CEHILA) under María Alicia Puente Lutteroth, which included Mexicanists María Eugenia Ugarte and Víctor Gabriel Muro (a liberation theology scholar). Currently the Red de Investigadores del Fenómeno Religioso en México (www.rifrem.mx/?page_id=11) coordinates much research and publication. Work influenced by Masferrer's ex-student Carlos Garma stresses Protestantism's endogenous logic over its foreign antecedents, thereby opposing David Stoll's earlier critique of it as an imperialist import. Without reducing new religious phenomena to North American conspiracies, cross-border connections are key (Odgers and Ruiz, 2009). Other topics include the expulsion of Protestant minorities from indigenous communities (Rivera et al., 2005), popular Catholicism, New Age or *la Nueva Mexicanidad* (Argyriadis et al., 2008), and the intersections of religion, gender, and power in religious movements, studied by Elizabeth Juárez Cerdi. As introductions, the *Atlas de la diversidad religiosa en México* (De la Torre and Gutiérrez, 2007) and its companion volume (Hernández and Rivera, 2009) condense basic demographic information on minority religions.

Community Studies

The concern with power, region, movements, and mutable identities reacted against community studies dating back to Robert Redfield and his research partner Alfonso Villa-Rojas. Nonetheless this literature, typically focused on indigenous forms of communal land tenure and social reproduction via civil-religious cargo systems, is alive and well where it has incorporated new theoretical perspectives. Catherine Good (Broda and Good, 2004) exemplifies long-term fieldwork in indigenous central and southern Mexico, updated with political and feminist concerns. Kinship research by scholars like Hugo Nutini and his ex-student David Robichaux also evince the tradition's fine-grained ethnography. Scholars influenced by Claude Lévi-Strauss, the more recent (and already morphing) ontological turn identified with Eduardo Viveiros de Castro, and models of partible personhood elaborated by Roy Wagner and Marilyn Strathern have done important work on indigenous communities. These include analysts of the myth, ritual, cosmology, art and history of Otomís (Jacques Galinier), central Uto-Aztecan cultures (Jesús Jáuregui), Huichols (Johannes Neurath), Guerrero Nahuas (Aline Hémond), Mixtecs (John Monaghan), and Tzeltal Mayas (Pedro Pitarch, whose 2010 book draws on Wagner to recast interethnic relations from an indigenous ontological perspective). Ben Feinberg has turned the tradition on its head with his experiments in literary representation and demonstration of how Mazatec shamans' use of space confounds closed models of community and region.

Linguistic Anthropology

Despite the diverse fieldsites where linguistic anthropologists (many of them US-trained) have worked since the 1970s, they often share approaches inspired by semiotician CS Peirce and literary theorist Mikhail Bakhtin. Jane Hill has contributed both formal linguistic analyses and broadly anthropological work, originally with Nahuatl speakers in the Sierra Norte de Puebla. John Haviland's work in Chiapas spans the linguistic marking of social status to legal anthropology and most recently the interface between Tzotzil Mayan speakers and sign-language users. John Lucy has updated Benjamin Lee Whorf with his examination of Yucatec Mayan language and thought. William Hanks has a broad ethnographic and ethnohistorical corpus on Yucatec, expanding the purview of deixis to link language, geographical space, cosmology, and social hierarchy. Teresa Carbó's discourse analysis has taken formal structural description (the central concern of Mexican linguistics) to an ontological turn. Among the emerging generation of linguistic anthropologists, Hilary Parsons Dick (2011) has published on Bakhtinian chronotopes of movement in Spanish, with a focus on migration.

Many scholars (e.g., Nancy Hornberger) continue to focus on indigenous language revitalization in Latin America, especially where Amerindian speech and writing still have multiple socioeconomic and political functions. However, Shaylih Muehlmann (2013) shows how Cucupá youths' Rabelaisian mistranslations for nonspeakers subvert the official reliance on native language use as an ethnic marker – for them a vestige of *indigenista* paternalism and national patrimony discourses that neoliberal multiculturalism has only resuscitated.

Visual Anthropology

Visual anthropology highlights the growing attention paid to indigenous forms of representation and agency. Early promoters from other anthropological subfields like Ricardo Pérez Montfort, Scott Robinson, Margarita Dalton, Victoria Novelo, and François Lartigue helped establish CIESAS and the UAMI as centers for ethnographic media in Mexico. Interdisciplinary inspiration came from the prominent historian of

photography, John Mraz, and the historian of Latin American cinema, David Wood. The new generation includes Deborah Dorotinsky (UNAM-Instituto de Investigaciones Estéticas), Antonio Zirión (UAMI), Carlos Y. Flores (Universidad Autónoma de Morelos), and the German scholar Axel Köhler. Köhler, Flores, Carlos Antaramián, Laurel Smith, Byrt Wammack Weber, Erica Wortham, and Gabriela Zamorano not only theorize local agency in media production but collaborate with their subjects in filmmaking, addressing both ethnic and gender issues, often in rural southern Mexico but also in urban and foreign settings (http://nmai.si.edu/explore/film-media/native-media-topics/mapping-mexican-media/). This reflects the influence of the University of Manchester's Granada Center for Visual Anthropology, Deborah Poole's (1997) pioneering visual anthropology work in Peru and her deconstruction of race and ethnicity in Oaxaca, as well as Faye Ginsburg's leadership of New York University's Graduate Program in Culture and Media. Jesse Lerner's 'false documentary' and other experimental approaches have influenced recent Mexican work, including Sandra Rozental's *Piedra Ausente/Absent Stone*, a film (and forthcoming book) addressing the instability of national patrimony as the state expropriates and formerly docile subjects reappropriate it.

Urban Space

Traditional urban anthropology, insofar as it assumed *barrios* were structured communities, had reflected *indigenismo's* obsession with coherent pueblos and the *grandes problemas nacionales* of modernity and modernization. This stemmed from a more fundamental reliance on Émile Durkheim's image of organic solidarity instead of Georg Simmel's urban anonymity. The community-oriented approach yields convincing results where rural groups actually manage to establish solidary urban colonies, as pioneered in the work of Larissa Lomnitz and carried forward in studies of planned neighborhoods (e.g., Claudia Zamorano). But inspired by Michael Warner's notion of shifting counter-publics, there is now a complementary image to the persistent community in urban Mexico (Duhau and Giglia, 2008). Linked to urban research predicated on neoliberal anomie and uncertainty are examinations of violence, another area where sociologists are more prominent than anthropologists. Arturo Alvarado (2012) engages the historical sociology of Diane Davis in his work on policing in the Mexico City metropolitan area.

Previously economists and sociologists (e.g., Brian R. Roberts) had examined demographic and political-economic contexts shaping urban space. Other senior researchers in the urban field include Daniel Hiernaux and Raúl Nieto of the UAMI, whose notions of spatial appropriation contextualize urban conflicts. Also at the UAMI is Néstor García Canclini (2001), whose work dialogues with the global cities scholarship of Saskia Sassen. But central to this article, the UNAM-based scholar and jurist Antonio Azuela's landmark work (1999) brought the social relations behind real estate markets into urban anthropology. His interest in patrimony – the inalienable possession of land – dovetails with contemporary indigenous struggles. It also resonates with Elizabeth Ferry's updating of a Marxist approach to silver mine labor in terms of Annette Weiner's "paradox of keeping-while-giving" such possessions. Azuela emphasizes what Laura Roush (2014) dubs "the protection racket model of urbanization": a basic tactic for the great majority who cannot get housing through conventional market transactions is to exploit the *amparo* (injunction or staying order) and resort to clientelism and other extralegal measures to obtain infrastructure development and legalize squatting.

Cristina Mejorada Sánchez and Elías Huaman have followed Azuela's lead, as have more recent US-trained scholars like Roush and Alejandra Leal. Leal's work on the gentrification of Mexico City's *centro histórico* gives proof that Mexican urban anthropology is becoming a channel for inquiring into the nature of the nation, liberalism, crowds, identities, and Derridean self-constitution via the absent Other. In this case, that Other entails Roush's "sinister broker who sells the *amparo*" and the spectral working class 'pueblo'. The yuppie wave may be displacing them but still they threaten to 'invade' the 'purified' urban core made safe for the rule of law.

Migration and (Post-)Borderology

Migration studies emerged from sociological concerns with labor and demography to examine urban phenomena, transnational indigenous movements (Kearney, 1996), and the questionable coherence, boundedness, and sufficiency of the nation-state as a category of analysis. Key literature includes Roger Rouse's seminal article on the relationship between Mexican home communities and their migrant offshoots in the United States (1991), long-term anthropological projects like Gail Mummert's (1999) and Douglas Massey and Jorge Durand's more strictly sociological corpus. Mummert's recent work, among others, examines how the threat of deportation and ad hoc household arrangements in the diaspora transform kinship, gender, and sexuality (Cantú, 2010).

Related to migration studies (primarily by location) is 'borderology' – originally identified with cultural studies – by such figures as José Manuel Valenzuela of the Colegio de la Frontera Norte. Emergent approaches leave behind the earlier, more optimistic Chicano imaginary, which envisioned an interstitial third country around *el bórder*. A more sober assessment, reaching back to Andrés Reséndez's border history and Pablo Vila's emphasis on prejudice, stereotyping, and other kinds of cultural friction, stresses nation-states' constraints on classes of people who cannot move as freely as transnational capital, especially since 11 September 2011.

Also contesting previous generations' confinement of Mexicanists within national boundaries is the work of Claudio Lomnitz and his students who have undertaken an anthropology of the state and the national space. Lomnitz (2001) manifested this agenda in his rejoinder to Bonfil's *México profundo* and his analysis of the temporality of 'crisis' engendered by the fiscal crunch of 1982. This revision inverts standard center-periphery models because it explores the

edges of countries and marginal spaces within them as sites where national identity is produced rather than diluted, even as officialdom attempts to erase or assimilate such impurities. Emergent work includes Leal's and Roush's urban ethnographies and Rihan Yeh's analysis of Tijuana as a site of state operations that inflect different publics' efforts to articulate national identities. Trevor Stack's binational ethnography of Mexican local intellectuals' conceptions of citizenship parallels such concerns.

Looking at the state itself, Josiah Heyman's corpus has focused on border enforcement apparatuses. Mateo Taussig's work on Immigration and Naturalization Service (INS) detainment facilities represents another approach to critical immigration studies, influenced by Giorgio Agamben. Lomnitz's transnational approach to Mexican national identity is explicit in his study of the revolutionary-era Flores Magón brothers, who inhabited both sides of the border as they forged a new ideology of collective belonging to the land (2014). That underscores this article's overall point: the growing academic recognition of popular agency in language and media, the geographical margins, and urban areas, distanced from *indigenista* anthropology's focus on the rural Mesoamerican core where national mestizo culture would one day domesticate otherness.

Acknowledgments

My thanks to the editors of this volume, my colleagues at the Colegio de Michoacán and Rice University for their generous suggestions. My apologies to those scholars whose important work was overlooked or could not fit into this highly condensed and partial format. A more complete bibliography is available on request.

See also: Christianity in Latin America; Education in Latin America and the Caribbean: Systems and Research; Fertility Transition: Latin America and the Caribbean; Identity Movements; Mexican Revolution, The; Migration to and from Latin America; Modernization, Sociological Theories of; Modernization; Neoliberalism; Religion: Nationalism and Identity; Urban Neoliberalism.

Bibliography

Alvarado, Arturo, 2012. El tamaño del infierno: Un estudio sobre la criminalidad en la Zona Metropolitana de la Ciudad de México. El Colegio de México, México DF.

Argyriadis, Kali, et al. (Eds.), 2008. Raíces en movimiento. Prácticas religiosas tradicionales en contextos translocales. IRD, CEMCA, CIESAS, ITESO, Guadalajara, México DF.

Astorga, Luis, 2007. Seguridad, traficantes y militares: El poder y la sombra. Tusquets, México DF.

Azuela, Antonio, 1999. La ciudad, la propiedad privada y el derecho. El Colegio de México, México DF.

Bonfil, Guillermo, 1996[1989]. México Profundo: Reclaiming a Civilization. University of Texas Press, Austin.

Briggs, Charles L., 1996. The politics of discursive authority in research on the "invention of tradition". Cultural Anthropology 11 (4), 435–469.

Broda, Johanna, Good, Catharine (Eds.), 2004. Historia y vida ceremonial en las comunidades mesoamericanas: Los ritos agrícolas. INAH, UNAM, México DF.

Calderón, Marco, 2012. Educación indígena, ciudadanía y Estado en México. Siglo XX. El Colegio de Michoacán, Zamora.

Cantú, Lionel, 2010. The Sexuality of Migration: Border Crossings and Mexican Immigrant Men. NYU Press, New York.

Das, Veena, Poole, Deborah, 2004. Anthropology in the Margins of the State. School of American Research, Santa Fe.

De la Peña, Guillermo, 1999. Territorio y ciudadanía étnica en la nación globalizada. Desacatos 1, 13–27.

De la Peña, Guillermo, 2005. Social and cultural policies toward indigenous peoples: perspectives from Latin America. Annual Review of Anthropology 34, 717–739.

De la Torre, Renée, Gutiérrez, Cristina (Eds.), 2007. Atlas de la diversidad religiosa en México. El Colegio de Jalisco, Guadalajara.

Dick, Hilary P., 2011. Language and migration to the United States. Annual Review of Anthropology 40, 227–240.

Duhau, Emilio, Giglia, Angela, 2008. Las reglas del desorden: Habitar el metrópoli. Siglo XXI, UAM, México DF.

Eiss, Paul, 2010. In the Name of the Pueblo: Place, Community and the Politics of History in Yucatán. Duke University Press, Durham.

Friedrich, Paul, 1986. The Princes of Naranja: An Essay in Anthrohistorical Method. University of Texas Press, Austin.

García Canclini, Néstor, 2001. Consumers and Citizens: Globalization and Multicultural Conflicts. University of Minnesota Press, Minneapolis.

Gledhill, John, 2008. Autonomy and Alterity: The Dilemmas of Mexican Anthropology. Manchester University. http://jg.socialsciences.manchester.ac.uk/autonomy%20&%20alterity.pdf.

Gledhill, John, 2011. Nuevas epistemologías en Antropología: Antropologías desde el Sur. Boletín Colegio Etnólogos y Antropólogos Sociales, A.C., pp. 11–24.

González y González, Luis, 1997. La ronda de las generaciones. Obras completas, vol. 6. Clío, México DF.

Greenberg, James, 1989. Blood Ties: Life and Violence in Rural Mexico. University of Arizona Press, Tucson.

Gutmann, Matthew, 1996. The Meaning of Macho: Being a Man in Mexico City. University of California Press, Berkeley.

Hernández, Aída, 2001. Histories and Stories from Chiapas: Border Identities in Southern Mexico. University of Texas Press, Austin.

Hernández, Alberto, Rivera, Carolina, 2009. Regiones y religiones en México. Estudios de la transformación sociorreligiosa. Colegio de la Frontera Norte, CIESAS, México DF.

Howe, Cymene, 2014. Anthropocenic ecoauthority: the winds of Oaxaca. Anthropological Quarterly 87 (2) (Special Collection: Energopower and Biopower in Transition).

Kearney, Michael, 1996. Reconceptualizing the Peasantry: Anthropology in Global Perspective. Westview, Boulder.

Krotz, Esteban, 1997. Anthropologies of the South: Their rise, their silencing, their characteristics. Critique of Anthropology 17 (3), 237–251.

Liffman, Paul, 2011. Huichol Territory and the Mexican Nation: Indigenous Ritual, Land Conflict, and Sovereignty Claims. University of Arizona Press, Tucson.

Lomnitz, Claudio, 1992. Exits from the Labyrinth: Culture and Ideology in the Mexican National Space. University of California Press, Berkeley.

Lomnitz, Claudio (Ed.), 2000. Vicios públicos, virtudes privadas: La corrupción en México. CIESAS, Porrúa, México DF.

Lomnitz, Claudio, 2001. Deep Mexico, Silent Mexico: An Anthropology of Nationalism. University of Minnesota Press, Minneapolis.

Lomnitz, Claudio, 2014. The Return of Comrade Ricardo Flores Magón. Zone Books, New York.

Masferrer, Elio (Ed.), 2000. Sectas O Iglesias: Viejas Y Nuevas Religiones. Plaza y Valdés, Bogotá.

Masferrer, Elio, 2004. Es del César o es de Dios? Un modelo antropológico del campo religioso. UNAM, Plaza y Valdés, México DF.

Millán, Saúl, Valle, Julieta, 2003. La comunidad sin límites. Estructura social y organización comunitaria en las regiones indígenas de México, vol. I. INAH, CONACULTA, México DF.

Monaghan, John, 1995. The Covenants with Earth and Rain: Exchange, Sacrifice, and Revelation in Mixtec Sociality. University of Oklahoma Press, Norman.

Muehlmann, Shaylih, 2013. Where the River Ends: Contested Indigeneity in the Mexican Colorado Delta. Duke University Press, Durham.

Nagengast, Carol, Kearney, Michael, 1990. Mixtec ethnicity: social identity, political consciousness, and political activism. Latin American Research Review 25 (2), 61–91.

Odgers, Olga, Ruiz, Juan Carlos (Eds.), 2009. Migración y creencias: Pensar las religiones en tiempos de movilidad. Porrua, México DF.

Pansters, Wil, Knight, Alan (Eds.), 2005. Caciquismo in Twentieth-Century Mexico. University of London, London.

Pitarch, Pedro, 2010. The Jaguar and the Priest: An Ethnography of Tzeltal Souls. University of Texas Press, Austin.

Poole, Deborah, 1997. Vision, Race, and Modernity: A Visual Economy of the Andean Image World. Princeton University Press, Princeton.
Prieur, Annick, 1998. Mema's House, Mexico City: On Transvestites, Queens, and Machos. University of Chicago Press, Chicago.
Rivera, Carolina, et al., 2005. Diversidad religiosa y conflicto en Chiapas: Intereses, utopías y realidades. UNAM, CIESAS, México DF.
Rockwell, Elsie, 2007. Hacer escuela, hacer estado: la educación posrevolucionaria vista desde Tlaxcala. El Colegio de Michoacán, Zamora.
Roseberry, William, 1989. Anthropologies and Histories: Essays in Culture, History, and Political Economy. Rutgers University Press, New Brunswick.
Rouse, Roger, 1991. Mexican migration and the social space of postmodernism. Diaspora 1 (1), 8–23.
Roush, Laura, 2014. The Santa Muerte, protection and desamparo: a view from a Mexico City altar. Latin American Research Review. (Special Issue: Lived Citizenship and Lived Religion in Latin America's Zones of Crisis).
Salmerón, Fernando, 1984. Caciques. Una revisión teórica sobre el control político local. Revista Mexicana de Ciencias Políticas y Sociales 30 (117–118), 107–141.
Sieder, Rachel (Ed.), 2002. Multiculturalism in Latin America: Indigenous Rights, Diversity and Democracy. Palgrave Macmillan, London.
Sieder, Rachel, 2013. De las reformas multiculturales al fin de reconocimiento. Justicia, pueblos indígenas y violencia en México y Guatemala. CIESAS, México DF.
Sierra, María Teresa (Ed.), 2004. Haciendo justicia. Interlegalidad, derecho y género en regiones indígenas. CIESAS, México.
Speed, Shannon, R., et al., 2006. Dissident Women: Gender and Cultural Politics in Chiapas. University of Texas Press, Austin.
Toledo, Víctor, Barrera, Nicolás, 2008. La memoria biocultural. La importancia ecológica de las sabidurías tradicionales. Icaria/Junta de Andalucía, Barcelona.
Torres, Gabriela, 2012. La ruralidad urbanizada en el centro de México. Reflexiones sobre las reconfiguraciones locales del espacio rural en el contexto neoliberal. UNAM, México DF.
Wolf, Eric R., 1956. Aspects of group relations in a complex society: Mexico. American Anthropologist 58 (6), 1065–1078.

Americas, Sociocultural Overviews: North America, Native Americans

Russell Thornton, University of California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

An overview of Native Americans north of Mexico is provided, including characteristics of Native American societies and cultures, aspects of Native American history – especially governmental policies aimed at assimilating Native Americans – and the contemporary situation of Native Americans in the United States and Canada. Particular attention is given to Native American origins, aboriginal population size, population decline following the arrival of Columbus, and recent population recovery and current demographic situation. The study of Native Americans is also considered, as it was important in the development of anthropology in North America and elsewhere.

Origins

Scholars have for some time generally accepted the notion that *Homo sapiens sapiens* – those who became the Native Americans – first arrived in the Western Hemisphere from the Asian continent. In this respect, they acknowledge the biological affinities between Native Americans today and peoples in eastern Asia. There was also general agreement that *H. sapiens sapiens* arrived via the Bering Land Bridge (or its western coast), which has existed at least four times in the past 70 000 years. The issue was not so much how *H. sapiens sapiens* arrived, but when. The conservative view, based on clearly dated human skeletal remains, is that humans arrived only around 10 000 years ago; the liberal view, based on less clearly dated human skeletal remains, as well as other types of evidence, e.g., tool making traditions, language, archeological sites, is that humans have been in North America at least twice as long, if not considerably longer. Scholars have also revised their thinking as to how humans arrived in the Western Hemisphere, initially and/or subsequently. Another migration route gaining increasing acceptance is a maritime one, whereby humans migrated along the coast and thus around the edges of large glaciers that covered much of the northern part of North America at that time period. Some have even suggested an arrival from the Iberian peninsula! It is generally accepted now that there was not just one migration, or even just two: the first, the ancestors of American Indians; and the second, the ancestors of Inuit (Eskimos) and Aleutian Islanders. Most scholars would now accept the notion that there were at least three migrations: one, ancestors of most American Indians; two, ancestors of Athabascan American Indians; and, three, ancestors of Inuit and Aleuts. There is even acceptance of possible other migrations, as evidenced by the controversy and speculations surrounding 'Kennewick man.'

Some of these migrations may have come from areas other than Asia. Confusing the scholarly study of arrivals and their timing is the idea of continued contact between peoples of the Pacific Rim following migrations. Posed to provide clarification of migrations, dates and affinities are developing techniques of analysis, e.g., DNA. Future generations of scholars will provide future answers and, undoubtedly, future controversies. It is clear, however, that what is emerging is a complex and dynamic picture of the peopling of the North American continent.

Early Native Americans

However and whenever, and however many separate times humans arrived in the Western Hemisphere before the late fifteenth century, they eventually developed elaborate and diverse societies and cultures. They, for example, developed a remarkable variety of material cultures, ones obviously shaped by the environments in which they lived and by how they exploited these environments. Some were foragers, exploiting different facets of their environments for food, clothing, shelter, etc., during different seasons of the year. Others were hunters, of deer and buffalo, particularly, but earlier of mammoths, mastodons, and giant sloths; or they were fishers, for example, of salmon in the Pacific Northwest and Alaska. Yet others were maritime hunters of whales, seals and walruses. Many American Indians became agricultural peoples, typically growing maize by 1492 – which diffused northward from present-day Mexico – but also other foods (e.g., squash and beans), along with maize, in the Southwest and East, or camas in the Pacific Northwest.

Cultural life also became rich. Multiple rituals and religions developed and flourished, along with diverse patterns of kinship and marriage, political organization and governance, and practices whereby youths became adults. Europeans considered many of these to be exotic, and compared them to ones they were familiar with in Europe and elsewhere in the world. Varied patterns of political organization and governance developed as well. For example, tribal patterns of organization diverged widely, ranging from small, extended kinship systems to triblets to tribes to confederacies and chiefdoms. Probably, the most spectacular Native American peoples north of present-day Mexico – from the European point of view – were the chiefdoms in Florida and the Southeast to the Mississippi River described by the chroniclers of the explorer Hernando de Soto, whose expedition traveled the area from 1539 to 1543. These were large, dense settlements controlled by rulers, one of whom was the well-known 'Lady of Cofitachequi' (from present-day South Carolina). These chiefdoms soon declined following the arrival of European explorers, probably mainly from epidemic diseases brought by the Europeans. Some, however, such as Cahokia, across the Mississippi River from present-day St. Louis and the largest city in the Western Hemisphere north of Mexico, had reached their zenith prior to European contact.

Perhaps a close second from the Europeans' definition of spectacular Native American peoples were the pueblos of the Southwest, described at about the same time as were the Florida chiefdoms by the explorer Hernando Coronado, and located generally along either side of the Rio Grande River. Other regions of aboriginal North America had Native American civilizations that impressed Europeans to some degree, although Europeans were decidedly unimpressed with some Native American societies and cultures they encountered.

Size and Distribution throughout North America

Native American populations, in the form of hundreds of distinct peoples, were distributed widely throughout North America *c.*1492.

Aboriginal Population Size

Scholars continue to debate the size of the Native American population present in North America north of Mexico *c.*1492, and thus also the magnitude of population decline that began soon after European arrival and continued to about 1900. James Mooney (1928) of the Smithsonian Institution formulated the classic estimate of aboriginal population size at, as he expressed it, "first extensive European contact": he asserted that there were some 1.153 million Native Americans north of the Rio Grande River. Subsequent scholars have offered both slightly lower and far larger population estimates, ranging from Kroeber's (1939) 1.025 million to Dobyns (1983) 18 million (for north of Mesoamerica, an area including northern Mexico as well as the United States, Canada, and Greenland). The estimate I formulated and continue to use is 7-plus million, composed of some 5 million for present-day United States, and some 2 million for Canada and Greenland.

Distribution

Populations were not distributed evenly, however. For example, Kroeber (1939) used the concept of 'cultural areas' to divide Native American groups north of Mexico into areas sharing similar cultural traits as adaptations to similar environments. According to him, the largest populations were found in the Southeast and California; the largest population concentrations were found in the Pueblo and Lower Columbia regions (Kroeber, 1939, Table 8). More recently, Ubelaker (1988) argues that the largest area populations were found in the Southeast and the Northeast.

Impact of European Contact

Contact with Europeans following the beginning of their extensive colonization of the Western Hemisphere after 1492 brought fundamental changes to native populations, and their societies and cultures. During the 1990s, 'holocaust' emerged as the metaphor to view the population collapse of Native Americans accompanying European expansion into this hemisphere (Thornton, 1987). Likewise, this 'holocaust' has emerged as crucial to understanding the full impact of colonialism upon Native Americans, and their subsequent social, cultural, biological, and even psychological changes. Native American societies and cultures, and Native Americans as biological and psychological entities were all impacted by demographic collapse after 1492.

Tribal Decimation and Relocation and/or Concentration

During the first four centuries following initial European contact, virtually all Native American groups experienced population decimation, and relocation and/or concentration, to some degree. Early relocations were sometimes voluntary, as American Indians moved westward to escape encroaching Europeans on America's east coast. This in turn produced conflicts with other American Indians as they were pushed farther west by the immigrant Indians. Subsequently, however, relocation became the official policy of the US government.

Native American population size in the United States and Canada declined for some 400 years, reaching a nadir population around the beginning of the twentieth century of perhaps only 375 000. The causes of the population decline, of course, were epidemics and other diseases introduced from Europe and Africa, e.g., smallpox, the plague, measles, forms of influenza, whooping cough, cholera, typhoid, diphtheria, scarlet fever, yellow fever, malaria, and various sexually transmitted diseases. European, Euro-American, and American colonialism in all its forms also contributed to Native American population reduction through warfare and genocide, relocations, and destruction of economic systems as well as the association of disease with all of these. Important, too, in the equation was the inability of Native American populations to recover from population losses as their societies and cultures were undermined by colonialism.

Accompanying this decimation were further patterns of population relocation and/or concentration, typically onto reservations. The largest pattern of population relocation occurred as official US government policy in the first half of the nineteenth century, whereby large populations of Indians in the Southeast – primarily the Cherokee, Chickasaw, Choctaw, Creek, and Seminole – were removed to lands west of the Mississippi River, lands which were designated Indian Territory and eventually became the eastern half of the state of Oklahoma in 1907. Other tribes were removed into what became the state of Oklahoma (named after Oklahoma Territory, the former western half of the state) as well, particularly ones from other parts of the Southern and Western Plains, the Northern Plains, and the so-called Old Northwest, around the Great Lakes.

US government troops, especially following the United States Civil War which ended in 1865, were used in particular to subjugate Native Americans by forcing them onto reservations in Oklahoma Territory, Indian Territory, and elsewhere in order to contain them and reshape their ways of life. American Indian struggles against such repression included new spiritual movements such as the Ghost Dances of 1870 and 1890, and the Peyote Religion, which arose among many Indian peoples. However, struggles also culminated in infamous massacres,

e.g., at Sand Creek, Colorado; Fort Robinson, Nebraska; and Wounded Knee, South Dakota.

Tribal or Individual Indian

After establishing a reservation system for American Indians, the US government soon began to undermine the integrity of many Indian lands when it saw the reservation system as 'not working' to 'civilize' Indians. From the late nineteenth century into the early twentieth century, policy was introduced to allot parcels of land on some reservations, and in Oklahoma and Indian Territory to individual Indians, thereby eliminating many reservations. A reduced reservation pattern therefore emerged, with most remaining reservations being in the west and representing only small remnants of former lands.

The idea was that Indians were to become individual Indian citizens on privately owned parcels of land. This culminated in the Citizenship Act of 1924, which made all American Indians in the United States citizens for the first time and gave them full voting privileges. 'Leftover' or 'surplus' lands were then 'opened' to claims by non-Indians. Accompanying allotment were larger political and corporate desires for American Indian lands. The decades following the period were characterized by frauds whereby individual allotments ended up in the hands of land barons, speculators, and corporations, as Debo (1940) has shown for the Cherokee, Chickasaw, Choctaw, Creek, and Seminole in Oklahoma; and as Meyer (1994) has shown regarding the White Earth Reservation in Minnesota.

American Indian tribal life was undermined greatly by allotment. Tribes able to retain tribal land in the form of reservations survived as tribes and as Indians better than those that were not able to do so. This included Indians on the Plains and in the west and southwest. In 1934, the Indian Reorganization Act was enacted, acknowledging that an Indian tribe had 'rights to organize for its common welfare.' The Act served to codify tribal governments under the all too watchful eyes of the Bureau of Indian Affairs. And, in turn, it provided the basis for the operations of the more than 330 federally recognized tribes, on reservations or not, and the more than 220 federally recognized Alaskan Native Villages.

The larger American society's vision of American Indians as citizens was also manifest in its system of Indian education. In the early days, America established schools to educate Indians. In fact, the American system of higher education which began in the seventeenth century – the 'Ivy League' schools, in particular – was in part to educate Native Americans, as reflected in the charters of, for example, Harvard University and Dartmouth College. With the development of the reservation system and the development of the system of elementary and secondary education in the United States, a natural alliance was born, and some treaties provided funds for reservation schools. By the beginning of the twentieth century, there were more than 200 government schools on reservations, supplemented by schools funded reluctantly funded by states for Indian youth, and some schools in soon-to-become Oklahoma, primarily under the auspices of the Five Tribes there. By the mid-1920s, however it became apparent that these schools – particularly the boarding ones, but also the day schools – were not effective as avenues to mainstream America: rather, they were a national disgrace. In 1928, the Meriam Report (Meriam, 1928: p. 348) stated that the "Indian problem is essentially an educational one," and called for a redirection of Indian education. Eventually, schools for American Indians became more sympathetic to Indian cultures, and began to incorporate Indian history and culture into the curriculum. Following World War II, however, boarding and day schools for Indians began to decline, and by the 1950s public school education for Indians became the norm.

Urbanization

Coinciding with the increase in the American Indian population beginning with the twentieth century, and with the dispossession of Indians from their lands, were the beginnings of a small but steady increase in the proportion of the population that lived in urban areas. Less than one-half of 1% of the American Indian population of the United States lived in urban areas in 1900; by 1950, more than 13% of the population lived in urban areas; at the time of writing, well over half of the American Indian population lives in urban areas, and the remainder live on reservations or in rural areas. (Only about half of the Native Americans in Canada live in cities, a substantial increase from about one-third in the 1970s, and a mere one-eighth in the early 1960s.) In part, this was a result of American Indian servicemen and women returning to cities rather than to small towns or rural areas following World War II. It was also a result of the mid-twentieth-century Bureau of Indian Affairs relocation program that encouraged and assisted American Indians to relocate to urban areas.

Native Americans Today

The Native American population in the United States and Canada experienced significant population recover during the twentieth century. From the nadir population of about 375 000 at the start of the twentieth century, it increased to more than 5 million self-defined Native Americans and mixed Native Americans in the 2010 US census plus more than 1 million self-defined 'Aboriginals' in Census Canada 2011. (See Shoemaker (1999), Ubelaker (1988), and Thornton (2000) for considerations of Native American population recovery.) This occurred because Native American mortality rates decreased, and fertility rates remained higher for Native Americans than for the total population. It also occurred because of changing census definitions of Native Americans in the United States, whereby individuals formerly classified as non-Native Americans became classified as Native Americans through self-identification in the censuses.

Tribal Membership

Most individuals self-identified in US censuses reported a tribal affiliation. The largest reported tribal affiliations include Cherokee, Navajo, Chippewa (Ojibwa), Sioux, Choctaw, Pueblo, Apache, Iroquois, Lumbee, and Creek. Not all of those in the Native American population are actual members of Native American tribes, however. Only about two-thirds of those self-identified as Native American in the recent US

decennial censuses are actual members of one of the federally recognized tribes or Alaskan Native Villages.

In Canada, the situation is different. There one must be registered under the Indian Act of Canada to be an 'official' Indian. Categories of Indians there include either status (or more or less 'registered') Indians – those recognized under the Act – and nonstatus (or more or less 'nonregistered') Indians – those who never were registered or gave up their registration (and became 'enfranchised'). Status Indians, in turn, are subdivided into treaty and nontreaty Indians. Canada also has the *Metis*, individuals of Indian and white ancestry who are recognized as such but not legally recognized as Indians. (This group is not generally 'land based' although there are several land-based *Metis* communities in Alberta.) Most Canadian Indians are registered, however.

Distribution

Only about one-quarter of Native Americans in the United States live on the more than 300 reservations and trust lands in the United States. About half of these live on the 10 largest reservations (and trust lands): Navajo Reservation (and trust lands); Pine Ridge Reservation (and trust lands); Gila River Reservation; Papago Reservation; Rosebud Reservation (and trust lands); San Carlos Reservation; Zuni Pueblo; Hopi Pueblo (and trust lands); and Blackfeet Reservation. Conversely, about two-thirds of the registered Indians in Canada live on more than 600 reserves there.

According to census data, the states with the largest Native American population include California, Oklahoma, Arizona, New Mexico, Alaska, Washington, North Carolina, Texas, New York, and Michigan. United States cities with the largest Native American populations include New York City, Oklahoma City, Phoenix, Tulsa, Los Angeles, Minneapolis–St. Paul, Anchorage, and Albuquerque. Canadian provinces with the largest number of Native Americans are Ontario, British Columbia, Saskatchewan, and Manitoba. And, there is now in Canada a new Native American-controlled province of Nunavut.

Anthropology and the Study of American Indians

Europeans who colonized North Americans soon developed intellectual interests in the native peoples here. They had to explain where the natives came from, but they were also in awe of native customs and cultures. Early explorers even sent Native Americans back to Europe, so that Europeans could see them for themselves. Eventually, prominent European scholars such as Thomas Malthus and G.W.F. Hegel developed an interest in these people and sought to integrate them into their study of human history, although Malthus considered Native Americans as 'wretched,' and Hegel thought they lacked history and would disappear.

In the late nineteenth century, however, Lewis Henry Morgan wrote of the native peoples of North America, including his two classics, *League of the Ho-de-no-sau-nee*, or *Iroquois* (1851) and *Ancient Society* (1877). He argued that they were disappearing and called for their study – documentation – before this happened. Others, such as Emile Durkheim, Sigmund Freud, Karl Marx, and Frederick Engels, also drew upon Native American societies and cultures in their writings. Although the interest here was decidedly on social evolution – with Native Americans being at the lower stages – intellectual interests changed in the early twentieth century toward how individual elements in societies and cultures were integrated to form a whole. Anthropology in the United States was now separating itself from sociology and coming into its own, with a decided interest in the native peoples there. In California, for example, Kroeber heeded Morgan's earlier call and sought in the first decades of the twentieth century to describe California Indians before they disappeared, or changed their ways of life fundamentally. In this regard, the well-known Ishi – the so-called last wild Indian – became a person of curiosity for Kroeber and others. Anthropology and history soon rejoined forces at mid-century, however, with the decided interest in studying Native Americans using the newly developed idea of 'ethnohistory.'

It was not too long before Native Americans became more involved in setting research agendas about themselves, and reacted against the study of them by anthropologists. Many rallied around Vine Deloria Jr.'s critique of anthropology in his *Custer Died for Your Sins* (1969), and his depiction of anthropologists as 'anthros.' Coinciding with this was the emergence of Native American studies in colleges and universities in the United States, and Native studies in Canadian counterparts, as a reaction toward both anthropology and the neglect of the study of Native Americans in other disciplines (Thornton, 1998a).

At the time of writing, anthropology, Native American studies, and Native Americans have achieved a truce, although an occasionally delicate one, in no small part because anthropologists have gained a new respect for Native Americans. Coinciding with an increased degree of self-determination of Native Americans in American society was the passage of the National Museum of the American Indian Act in 1989, both creating a national museum at the Smithsonian Institution devoted to Native Americans, and mandating the return of Native American human remains and funerary objects held at the Smithsonian, primarily in the National Museum of Natural History. A year later, the Native American Graves Protection and Repatriation Act was passed which mandated the repatriation not only of human remains and funerary objects from all institutions receiving federal funding but also the return of Native American objects of cultural patrimony and sacred objects. The Smithsonian, already required to repatriate human remains and funerary objects, was exempt from this legislation. However, the National Museum of the American Indian Act was amended in 1996 to include the required return of objects of cultural patrimony and sacred objects held at the Smithsonian.

Smithsonian repatriation is monitored by a seven member committee mandated by congress, and it is required that the majority of its members be selected from nominations submitted by Native American tribes. From its inception in 1990 to the early 2000s, it was chaired by Russell Thornton. To date, several 1000 sets of human remains and numerous cultural objects have been returned to Indian tribes from the Smithsonian. Included here are human remains from the Sand Creek and Fort Robinson Massacres, Ghost Dance shirts and other objects from the Wounded Knee Massacre, and Ishi's

brain, and Sitting Bull's braid, cut from their bodies at their autopsies and sent to the Smithsonian.

These national repatriation laws and state laws as well, have in turn forged new relationships between museums, archaeologists, physical anthropologists and ethnographers, and Native Americans (Thornton, 1998b).

The study of Native Americans, by both non-Native and Native scholars, in anthropology and other disciplines, will continue; however, Native Americans will have an increasing role in determining what, how, and by whom it is studied.

See also: Ethnic Identity and Ethnicity in Archaeology; Ethnicity: Anthropological Aspects; Hegemony and Cultural Resistance; Holocaust, The; Indigenous Rights.

Bibliography

Debo, A., 1940. And Still the Waters Run: The Betrayal of the Five Civilized Tribes. Princeton University Press, Princeton, NJ.

Deloria Jr., V., 1969. Custer Died for Your Sins: An Indian Manifesto. Macmillan, New York.

Dobyns, H., 1983. Their Number Become Thinned: Native American Population Dynamics in Eastern North America. University of Tennessee Press, Knoxville, TN.

Kroeber, A.L., 1939. Cultural and Natural Areas of Native North America, vol. 38. University of California Publications in American Archaeology and Ethnology, California, pp. 1–242.

Meriam, L., 1928. The Problem of Indian Administration. Johns Hopkins University Press, Baltimore.

Meyer, M., 1994. The White Earth Tragedy: Ethnicity and Dispossession at a Minnesota Anishinaabe Reservation. University of Nebraska Press, Lincoln, NE.

Mooney, J., 1928. The aboriginal population of America North of Mexico. In: Smithsonian Miscellaneous Collections, vol. 80, pp. 1–40.

Morgan, L.H., 1851. League of the Ho-de-no-sau-nee or Iroquois. Burt Franklin, New York.

Morgan, L.H., 1877. Ancient Society. Holt, New York.

Shoemaker, N., 1999. American Indian Population Recovery in the Twentieth Century. University of New Mexico Press, Albuquerque, NM.

Thornton, R., 1987. American Indian Holocaust and Survival: A Population History since 1492. University of Oklahoma Press, Norman, OK.

Thornton, R., 1998a. Institutional and intellectual histories of Native American studies. In: Thornton, R. (Ed.), Studying Native America: Problems and Prospects. University of Wisconsin Press, Madison, WI.

Thornton, R., 1998b. Who owns our past? Repatriating Native American human remains and cultural objects. In: Thornton, R. (Ed.), Studying Native America: Problems and Prospects. University of Wisconsin Press, Madison, WI.

Thornton, R., 2000. Population history of Native North Americans. In: Haines, M.R., Steckel, R.H. (Eds.), A Population History of North America. Cambridge University Press, Cambridge, UK.

Ubelaker, D., 1988. North American Indian population size, A.D. 1500 to 1985. American Journal of Physical Anthropology 77, 289–294.

Americas, Sociocultural Overviews: Caribbean

Karen Fog Olwig, University of Copenhagen, Copenhagen K, Denmark

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by M.R. Trouillot, volume 3, pp. 1484–1488, © 2001, Elsevier Ltd.

Abstract

The modern Caribbean emerged through the transformations that took place with European colonization, which resulted in the establishment of large-scale plantation societies and the repopulation of the area with Europeans, Africans, and Asians. This resulted in societies that are stratified in terms of race and class, culturally diverse and highly mobile. Key issues in research on sociocultural aspects of the area include African continuities, the cohesion of societies, the coexistence of contradictory value systems linked to colonial and local frameworks of life and their (re)interpretation in everyday life, the role of out-migration, and the development of transnational sociocultural systems.

A Geographical Entity

Viewed as a geographical entity the Caribbean consists of a long chain of islands, enclosing the Caribbean Sea and separating it from the Atlantic Ocean, as well as a smaller number of islands located within the Caribbean Sea. Comprising almost 1000 named islands and islets and extending across 2500 miles, the archipelago runs from Florida in a southeasterly direction toward Trinidad where it then continues in a westward direction along the northern part of the South American coast.

Four groups of islands can be distinguished. The Bahamas, consisting of approximately 700 islands, 30 of which are inhabited, is the northern-most group. Further south are the Greater Antilles, home for the vast majority of the area's about 40 million inhabitants and comprising Cuba, Hispaniola (divided into Haiti and the Dominican Republic), Jamaica, Puerto Rico, as well as a number of smaller islands that are associated with them. From the southeastern end of the Greater Antilles and stretching in a north-south direction to Trinidad is a long chain of islands known as the Lesser Antilles. About 40 of these islands are inhabited, and they are divided into the Leeward Islands (in the north) and the Windward Islands (in the south). Finally, a smaller number of islands located along the South American coast, and sometimes referred to as the Leeward Antilles, comprise the fourth group of islands.

The Caribbean is usually also seen to include Guyana, Suriname, and Cayenne (French Guiana) on the South American coast, Belize on the Central American coast, and the Colombian islands off the Nicaraguan coast, due to their shared colonial history with the Caribbean islands. Bermuda, further north in the Atlantic Ocean, is also sometimes regarded as part of the Caribbean because its population is mainly of Caribbean origin. Furthermore, in recent decades large Caribbean communities have developed in various North American and European metropoles, as substantial segments of the Caribbean population have relocated abroad. It is therefore difficult to make a clear-cut delineation of the Caribbean as a geographical region when viewing it from a historical and sociocultural point of view.

A Historical Perspective

Today the Caribbean is probably best known as a major tourist destination, visited every year by millions of North American and European visitors, who are attracted by the white sandy beaches and tropical climate, often moderated by pleasant trade winds. From the late fifteenth to the late nineteenth centuries, however, the area played a key role in European colonial ventures. It attained prominence in modern colonial history because it was easily accessible to the early seaborne colonization of the 'new world,' and a large number of European states have operated in the area, including Spain, Britain, France, the Netherlands, Denmark, and Sweden. After Columbus discovered the Caribbean islands (and thus the Americas) in 1492, the area became known for its pirates and buccaneers. In the course of the seventeenth century, however, it developed into a center of large-scale sugar production for export using a slave labor force imported from Africa, thus setting in motion the trans-Atlantic triangular trade.

The modern Caribbean, as we know it today, emerged through the complete transformation that took place with this European colonization. The Indian (or Amerindian) people living by agriculture and fishing in the area virtually disappeared within 50 years of Columbus's discovery, as they succumbed to new diseases brought by the colonizers, as well as the colonizers' violent forms of subjugation. They were first replaced by indentured European workers and then enslaved Africans brought to the Caribbean through enforced transportation, as the European colonizers needed a massive supply of laborers to run the plantations producing sugar for the world market. When the trans-Atlantic slave trade was stopped in the nineteenth century, an estimated 3 million Africans had been transported to the Caribbean. With the abolishment of slavery, initiated in the 1830s in the British colonies, many of the freed slaves left the sugar estates and established peasant villages of their own. Others migrated for economic opportunities in urban areas or in other Caribbean societies. This led to the recruitment of approximately a half million contract laborers in Asia, most of them deriving from India.

From the end of the nineteenth century, the European colonies experienced economic decline, as plantation production became less profitable. The European plantocracy, however, maintained economic control, making it difficult for the Caribbean population to develop a viable, alternative economic system. At the same time, the colonial regimes instituted a number of 'civilizing' efforts focusing on missionizing and educational activities intended to create loyal

colonial subjects. Nevertheless, out-migration continued and the post-emancipation period saw the establishment of a strong, still vibrant migration tradition involving population movements to areas of greater social and economic opportunity. During the nineteenth and first half of the twentieth century, migration primarily took place within the Caribbean area. Since the 1950s, however, it has been directed especially toward Europe and North America.

Politically, Haiti was the first Caribbean society to achieve independence – in 1804 after a slave uprising. It was followed by the Dominican Republic in 1821 and by Cuba in 1902. The rest of the Caribbean became decolonized in the post-World War II period, with most of the British possessions gaining independence and most of the Dutch and the American areas attaining partial independence, though Suriname gained full independence. The French colonies, on the other hand, were incorporated as overseas departments into France. Local political systems have ranged from dictatorships to parliamentarian systems based on free elections (Trouillot, 2001).

As can be seen from this history, from a sociocultural perspective the Caribbean may be characterized as an assortment of territories linked to the Caribbean Sea sharing a past of early European colonization that led to the extinction (or displacement) of the indigenous population, the establishment of exploitative European colonies, and the massive importation of an enslaved or contracted labor force from Africa and Asia. Out of this colonial past emerged a range of societies that can be described as highly stratified, culturally heterogeneous, and extremely mobile.

Anthropological Beginnings

The anthropological discipline evolved out of a concern to document and analyze indigenous sociocultural entities, whether through ethnographic fieldwork in local communities in Africa or Oceania, or by doing salvage ethnography among survivors of North American tribal societies on the verge of extinction. Anthropology therefore has not been well equipped to deal with a sociocultural area in which the original population has been largely eradicated and the new populations have been highly complex and mobile. Having virtually no indigenous population, the Caribbean, in fact, generated little interest in anthropology before the middle of the twentieth century. One of the few anthropologists to do substantial research in the area was Melville Herskovits who conducted fieldwork in Suriname, Haiti, and Trinidad during the 1920s and 1930s. His goal was to examine how the culture of the descendants of the African slaves could be seen to be based on retentions and reinterpretations of the culture of their African ancestors who were forcibly transported to the New World as slaves. In this endeavor he sought to counter racist charges that the blacks in the New World had no distinct culture of their own; however, his methodological approach – looking for cultural links through a comparative framework establishing the extent to which Caribbean cultural traits had retained their Africanness – has been criticized as crude and lacking an understanding of historical processes. Nevertheless, Herskovits is still known for his central role in establishing the 'New World Negro' as a legitimate area of study (Scott, 1991). Early anthropological research thus was grounded in an interest in the impact of the African past on not only Caribbean but black culture more generally.

Caribbean culture and society became the subject of sustained ethnographic research only after World War II when social and economic change became a topic of interest in the discipline. This reorientation of anthropology has been linked to two different political developments that generated an awareness of the possibility of broadening the scope of anthropological research. It led, however, to rather different theoretical approaches.

One development pertains to concern about the Cold War. According to David Scott (2004: 195) this led, in the United States, to a "new alliance between the intelligence apparatuses of the US state […], the philanthropic foundations [… and] key universities" that "reimagined the globe as a number of strategic research 'areas,' and enabled, underwrote, and shaped the new area-studies-directed social science of the post-war era." A concrete result of this alliance, and the research it enabled, was a major collaborative research project, directed by Julian Steward, that led to the publication of *The People of Puerto Rico* (Steward, 1956). This edited volume examined the development of Puerto Rican society, its varying subcultures associated with different material conditions of life and the ways in which they were integrated into the larger national community.

One of the participants in the project was Sidney W. Mintz, who did research in the southern part of Puerto Rico dominated by large-scale production of sugar cane. Like his American predecessor Melville Herskovits, Mintz, who became one of the leading figures in the anthropology of the Caribbean, was interested in the historical formation of Caribbean culture, but unlike Herskovits he did not look primarily for African continuities. Rather he investigated the culture building that took place within the Caribbean as people created a new life for themselves there, drawing on the range of material, social, and cultural resources available to them (Mintz, 1974; see also Mintz and Price, 1992). In a number of studies, Mintz documented how African-Caribbean communities emerged during slavery in the margins of the plantation systems representing a mode of 'accommodation' and, at the same time, 'resistance' to slavery and the dominant plantation society with which it was associated. These communities, he showed, became an important foundation for the peasant villages that were established after the abolishment of slavery (Mintz, 1974).

Throughout his studies Mintz emphasized that the Caribbean societies were essentially modern, having emerged within the 'agro-industrial enterprises' of the large-scale plantation system that he described as "landmark experiments in modernity" (1996: 295–297). According to Mintz, this modernity, which entailed the recruitment of labor power from different places, required that people had a 'non-valuative' openness toward cultural difference, partly because there was no 'institutional recourse to a standard common tradition.' While the Caribbean societies might display strong socioeconomic similarities, linked to their common colonial past, they were thus, according to Mintz, culturally heterogeneous and therefore not a 'culture area' in the sense of sharing particular cultural institutions and associated values.

The other political development that contributed to the establishment of the anthropological study of the Caribbean had to do with the need to address the social and economic problems that plagued the late colonial societies. In social science this led to an interest in societal integration both at the micro level of the family and the household as well as sociocultural integration at the wider societal level, as the British colonies moved toward political independence (Slocum and Thomas, 2003: 555). The work on micro-level integration, states Michel-Rolph Trouillot, was spurred by a "concern for policy on the part of officials who saw Afro-Caribbean families as 'deviant' simply because they did not fit the nuclear folk-model of Western consciousness" (1992: 26). Thus, even though some of the anthropological studies, published beginning in the 1950s, produced fine ethnographies of local communities (see, for example, Smith, 1956; Clarke, 1957), their analyses became entangled with colonial, bureaucratic understandings of Caribbean family forms as dysfunctional and as an important cause of the poverty of the African-Caribbean population. As noted by Trouillot (1992: 26), Raymond T. Smith's term matrifocal family, meant to point to the central role of women as mothers, came to signify female-headed (or perhaps rather, father-absent) families, whereas Edith Clarke's monograph *My Mother Who Fathered Me* (1957), which documented important aspects of rural communities in Jamaica, including the key role of family land (a kin-based form of shared land holding), became perceived as a study in social pathology.

Family studies became part of the larger discourse on the nature of Caribbean societies, where R.T. Smith and Michael G. Smith were the major protagonists. R.T. Smith contended that West Indian societies were integrated in terms of sharing cultural values and social norms but, at the same time, stratified due to the existence of color and class hierarchies that influenced the ways in which these values and norms were implemented in practice. He thus argued that all layers of Caribbean societies shared an ideal of the nuclear family, based on marriage, but that this ideal tended to be realized as a family form only in certain phases of the life cycle of individuals and the domestic units to which they belonged (Smith, 1956). M.G. Smith, however, maintained that West Indian societies were characterized by cultural pluralism, a concept he had adapted from John S. Furnivall's plural society theory developed in relation to South-East Asian colonial societies. According to M.G. Smith, Caribbean societies thus consist of groups that adhere to disparate social institutions and cultural norms, so they are only kept together through the power of the dominant segment of the society. He therefore maintained, and sought to demonstrate through statistical analysis as well as detailed ethnography, that West Indians do not share a common family ideal but operate according to different sets of cultural values and social norms (Smith, 1962a, b).

During the 1960s, the anthropology of the Caribbean was dominated by the debate between the cultural integrationalist and the cultural pluralist approaches. This debate, Karla Slocum and Deborah Thomas (2003: 555) state, was nourished by the "then-dominant social science view [which] held that social systems needed to be integrated around a common value system in order to thrive without an overarching (external) system of power and control." It was not before this view was superseded by a more dynamic, practice-oriented notion of sociocultural systems that the anthropology of the Caribbean moved beyond this theoretical and analytical impasse.

Caribbean Culture as Negotiated

The ethnographic studies of Peter Wilson (1969, 1973), based on fieldwork on the island of Providencia, played a key role in the development of new approaches that could move beyond the integrationalist-pluralist divide. Wilson proposed that Caribbean societies be conceptualized as shaped by dual cultural value sets centering on the opposing notions of respectability and reputation. According to Wilson, the value set of respectability emphasizes marriage and the nuclear family, sexual restraint, godliness and education, and was primarily linked to the colonial society, the middle class, and to women whose lives tended to center on the domestic unit and the church. Reputation, on the other hand, celebrates sexuality, the begetting of children, and sociability in public arenas such as the rum shop or the street corner – informal arenas of social life in the local communities where men tend to be more active. While Wilson linked the value sets to different social contexts, he did not see Caribbean societies as divided into disparate and separate modes of life. On the contrary, he emphasized that the two, opposing value sets were continuously played out against each other in social life as individuals drew on various aspects of the two, depending on the particular situation in which they found themselves. By seeing the two value sets as different, yet part of the same social system, Wilson thus combined an integrationalist and a cultural pluralist understanding of Caribbean culture.

Wilson's conceptual scheme has been highly influential in stimulating a more practice-oriented approach that can capture how individuals negotiate different, even conflictual, values and norms, as they seek to position themselves in society. It has also stimulated important critiques and suggestions for further theoretical refinement. Diane Austin-Broos (Austin, 1983), among others, has pointed out that Wilson's scheme needs to incorporate the issue of power – not just in terms of political and economic domination, but also in the sense of ability to "constrain ideas and to intervene in attempts to construct alternative models of society" (1983: 223). She therefore calls for a distinction between culture and ideology: "divergent cultures can persist within larger ideological schemata and do not necessarily signal the end of domination" (1983: 225). She proposes that Caribbean societies represent *a situation of conflict contained by domination* and that this should be a key topic in the anthropology of the Caribbean (1983: 223).

Wilson's scheme has been subject to the most criticism among anthropologists concerned with gender roles in the Caribbean. One of them is Carla Freeman who has argued that there are "competing and co-existing models of femininity that encompass not only those attributes and qualities of 'respectability' [mentioned by Wilson] but also elements that he would have located under the masculine heading of 'reputation' " (Freeman, 2002: 76). She further suggests that 'respectability' is a more dynamic value set than Wilson

acknowledges, because it may be renegotiated and find new forms of expression as Caribbean societies change through time. In her own work she has demonstrated this through an analysis of the ways in which women employed in the informatics industry in Barbados use this institutional structure as a framework within which to express their identity as women using clothing as a way of displaying their respectability (Freeman, 2000).

In yet other studies anthropologists have sought to improve Wilson's scheme by redefining the notions of respectability and reputation. In his work on consumption in Trinidad, Daniel Miller (1994) has thus suggested that Wilson's scheme can usefully be reconceptualized in terms of the notions of transcendence and transcience. Whereas the former, according to Miller, is linked to the internal, domestic sphere and validates the significance of enduring relations and values associated with the family, the latter is associated with the external, public sphere and the shifting relations, practices, and norms associated with modernity. In yet another attempt to further develop Wilson's conceptual framework, Huon Wardle has argued that its somewhat rigid duality can be broken down by rethinking the opposite pair of respectability and reputation in terms of 'commitment' and 'disjuncture,' notions that he sees as "less restrictively laden terms for the polarity of cultural accommodation and cultural confrontation that Wilson emphasizes as crucial" (Wardle, 2007: 569). Furthermore, he proposes adding a third dimension to this dualistic opposition: 'ambiguation' – a 'principle of mediation' that enables individuals to negotiate the social gap created by the radical differences between the disparate notions.

Creolization

Another strand of research, examining the complexity of Caribbean culture and society, has taken its point of departure in creolization processes, borrowing a concept of linguistics. In a study of Guyana, Leo Drummond shows that culture can be understood in terms of a continuum of cultural variation, expressed in a Guyanese context in terms of ethnic categories such as 'English' or 'white,' 'bush' (indigenous Indian), 'coolie' (Asian Indian) that are attributed with varying meaning, as individuals interact (1980: 357). Individuals will learn to master the smaller segment of the continuum where they are positioned, but even in these contexts they will encounter divisiveness and conflictual rules that they must negotiate in social practice (1980: 368). A creole model of culture, according to Drummond, leads to a new concept of culture, because a continuum will have no clear societal boundaries and be under constant transformation, as the meaning of cultural categories and customs are constantly defined and redefined in social life.

From a rather different perspective, Jean Besson has examined creolization through time, inspired by Sidney Mintz's historical approach. She conceptualizes creolization as a complex process involving "subaltern opposition and resistance to hegemonic power reflected in class conflict" as well as "adaptation and negotiation across class boundaries" (e.g., Besson, 2011: 29). She argues that creolization at its core concerns practices of 'localization' or 'indigenization' engaged in by the descendants of those who came from Europe, Africa, and Asia. Furthermore, she sees these practices as involving an ongoing quest for belonging within societies dominated by capitalist class relations linked to a global economy and changing colonial and postcolonial regimes (2011: 28). In several studies of Jamaican communities, she thus shows how kinship, land, and religion may be attributed with novel meaning and find expression in new practices, as people respond to changing conditions of life through time.

Some students of Caribbean culture, however, have been rather skeptical of the notion of creolization. Stephan Palmié, for example, maintains that its focus on the mixing of cultures entails a culturally essentialist understanding of the existence of different cultural entities that can be blended. He therefore suggests that theory of creolization should be "an object of, rather than a tool for, anthropological inquiry" (2006: 448). Such an inquiry can be found in Aisha Khan's (2004) ethnographic analysis of Trinidad. She argues that ideologies of mixing are prevalent in the Caribbean and Latin America and should be viewed as "causal forces in social processes." Thus, they often define certain population groups, such as the Indians or the Chinese, as non-Caribbean because they are not viewed as part of the black-white mix seen to constitute the backbone of Caribbean society.

The analysis of the concept of creolization makes an important, more general, point by highlighting the close connection between social science concepts and societal, normative notions. This issue is of particular salience in research on Caribbean culture and society that has developed within a hierarchical colonial context where it was not possible to establish 'traditional' frameworks of cultural particularistic interpretation associated with indigenous groups. The critique of the metaphor of mixing that has been associated with creolization, however, does not invalidate the importance of investigating Caribbean sociocultural processes through time in relation to changing global structures of inequality, as Besson has done in her work.

Mobility and Transnational Relations

During the last decades, anthropological studies of the Caribbean have increasingly moved out of the Caribbean as a geographical entity in acknowledgment of the massive outmigration from the area, which, as noted, has led to the emergence of large populations of Caribbean origin in North America and Europe. With this research has come the recognition that the Caribbean as a sociocultural area is deeply embedded in extensive fields of relations extending to migrant destinations in different parts of the world. One of the first to point to the significance of such fields was Stuart Philpott (1968), who, inspired by the Manchester School's studies of urban Africa, focused on the social networks connecting Caribbean sending societies and migrants abroad. By following the migrants from their Caribbean island of origin to London (thus demonstrating the value of doing multi-sited ethnographic fieldwork 30 years before it became fashionable in anthropology), he documented the emergence of a translocal community based largely on expectations and

obligations rooted in family and kinship that shaped, and were shaped by, life in the Caribbean as well as in London.

Following Philpott's work, a large number of studies (see Slocum and Thomas, 2003) have documented the key role of migration in Caribbean culture and society. They have shown that most people do not view migration as a means of leaving home with the intent to settle in another place, but rather as a resource that makes it possible for them to explore various opportunities abroad and enjoy social mobility, while remaining part of their Caribbean home community of origin. They therefore generally migrate with the intent to return, and many send considerable remittances to relatives in their community of origin, visit as often as possible, and even invest in homes of their own in the Caribbean.

This research has demonstrated that migration is an integral aspect of Caribbean societies. Thus, the great variability of domestic units to be found in the Caribbean should be seen within the context of the extensive networks of relations that link migrants engaged in wage employment abroad with their parents, siblings, and children in the family home in the local community. While some migrants do not succeed at supporting this family – including their own children left behind – many do and continue to be regarded as key members of the family home even if they live and work in distant destinations. This close relationship between migration and family relations has led Karen Fog Olwig (2007) to suggest that migration – despite the often prolonged physical absence that it involves – is an important way of doing family relations within the context of global systems of social and economic inequality. Thus trans-local family relations constitute not only key resources of economic support and social status for relatives left behind but also vital sources of identification and belonging for many migrants who may experience marginalization as ethnic minorities in the receiving societies. Family land can be seen to proffer a similar symbolic resource of import to migrants. As a form of land holding that allows the descendants of the original owner of the property to share land in common, it thus offers an (in principle) inalienable place of belonging in the Caribbean for migrants and their descendants (Besson, 2011).

This does not mean that it is unproblematic for Caribbean societies to be so strongly oriented toward migration. From a local point of view, the institution of family land, for example, can be viewed as a problematic form of land ownership that makes it difficult to earn a living on the land and therefore leads to underuse of local resources. Furthermore, the maintaining of close family relations over great physical distances may be fraught with tensions, disappointments, and personal sacrifice. This may become particularly apparent when migrants return to claim their place of belonging in the Caribbean (see, for example, Olwig, 2007).

If this migration research has shed interesting new light on Caribbean societies, its greatest impact seems to have been the challenge it has posed to notions of immigration and integration in the receiving societies in the Global North. In the United States, for example, immigration has played a major role in the development of the country, and there has been a strong tradition of investigating population movements to the country and the processes of assimilation into American society immigrants are assumed to undergo. As research on Caribbean migrants in the United States has demonstrated the key significance of continued ties to the place of origin, it has been necessary to reconsider the notion of migration as entailing a one-way move followed by processes of assimilation. In several studies, Glick Schiller and associates (Glick Schiller et al., 1992; Basch et al., 1994; Glick Schiller and Fouron, 2001) have argued, for example, that a key motivation for maintaining ties to the Caribbean is the rejection of the position as racialized others that Caribbean immigrants tend to be given in the bottom of the American ethnic hierarchy. Thus, many migrants not only actively sustain family relations across national borders but become involved in transnational sociocultural systems linked to ethnic organizations that can act as pressure groups in the receiving as well as the sending society. This has led many scholars to reconceptualize migrants as diasporas, population groups that remain oriented toward a distant homeland.

While this strong focus on the migration tradition in the Caribbean and the transnational communities, diasporic groups and ethnic mobilization with which it is associated has generated important new insights, it may be relevant to ask – inspired by the critique of the notion of creolization – whether it may end up defining Caribbean people as belonging to a liminal space outside the normal order of things.

See also: Colonialism, Anthropology of; Creolization: Sociocultural Aspects; Diaspora; Globalization, Anthropology of; Immigration, Citizenship, and Integration: Social Work Connections; Matrifocality; Modernity: Anthropological Aspects; Multisited Field Studies; Race and Racism in the Twenty-First Century; Slavery: Comparative Aspects.

Bibliography

Austin, D.J., 1983. Culture and ideology in the English-speaking Caribbean. A view from Jamaica. American Ethnologist 4 (3), 223–240.

Basch, L., Glick Schiller, N., Szanton Blanc, C., 1994. Nations Unbound. Transnational Projects, Postcolonial Predicaments, and Deterritorialized Nation-States. Gordon and Breach, Langhorne.

Besson, J., 2011. M.G. Smith's plural society theory and the challenge of Caribbean creolization. In: Meeks, B. (Ed.), M.G. Smith and the Emergence of Social Anthropology and Social Theory in the Caribbean and Beyond. University of the West Indies Press, Kingston, pp. 22–42.

Clarke, E., 1957. My Mother Who Fathered Me. A Study of Three Selected Communities in Jamaica. George Allen and Unwin, London.

Drummond, L., 1980. The cultural continuum. A theory of intersystems. Man 15 (2), 352–374.

Freeman, C., 2000. High Tech and High Heels in the Global Economy. Women, Work, and Pink Collar Identities in the Caribbean. Duke University Press, Durham.

Freeman, C., 2002. Mobility, rootedness, and the Caribbean higgler. Production, consumption and transnational livelihoods. In: Sørensen, N.N., Olwig, K.F. (Eds.), Work and Migration. Life and Livelihoods in a Globalizing World. Routledge, London, pp. 61–81.

Glick Schiller, N., Basch, L., Szanton-Blanc, C., 1992. Transnationalism: a new analytic framework for understanding migration. In: Glick Schiller, N., Basch, L., Szanton-Blanc, C. (Eds.), Towards a Transnational Perspective on Migration. Race, Class, Ethnicity, and Nationalism Reconsidered, Annals of the New York Academy of Sciences, vol. 645, pp. 1–24.

Glick Schiller, N., Fouron, G.E., 2001. Georges Woke Up Laughing. Long-distance Nationalism and the Search for Home. Duke University Press, Durham.

Khan, A., 2004. Callaloo Nation. Metaphors of Race and Religious Identity among South Asians in Trinidad. Duke University Press, Durham.

Miller, D., 1994. Modernity. An Ethnographic Approach. Berg, London.
Mintz, S.W., 1974. Caribbean Transformations. Aldine Publishing Company, Chicago.
Mintz, S.W., 1996. Enduring substances, trying theories. The Caribbean region as oikoumene. Journal of the Royal Anthropological Institute 2 (2), 289–311.
Mintz, S.W., Price, R., 1992. The Birth of African-American Culture. An Anthropological Perspective. Beacon Press, Boston.
Olwig, K.F., 2007. Caribbean Journeys. An Ethnography of Migration and Home in Three Family Networks. Duke University Press, Durham.
Palmié, S., 2006. Creolization and its discontents. Annual Reviews in Anthropology 35, 433–456.
Philpott, S., 1968. Remittance obligations, social networks and choice among Montserratian migrants in Britain. Man 3 (3), 465–476.
Scott, D., 1991. That event, this memory. Notes on the anthropology of African diasporas in the New World. Diaspora 1 (3), 261–284.
Scott, D., 2004. Modernity that predates the modern. Sidney Mintz's Caribbean. History Workshop Journal 58, 191–210.
Slocum, K., Thomas, D.A., 2003. Rethinking global and area studies. Insights from Caribbeanist anthropology. American Anthropologist 105 (3), 553–565.
Smith, M.G., 1962a. Kinship and Community in Carriacou. Yale University Press, New Haven.
Smith, M.G., 1962b. West Indian Family Structure. University of Washington Press, Seattle.
Smith, R.T., 1956. The Negro Family in British Guiana. Routledge & Kegan Paul, London.
Steward, J.H. (Ed.), 1956. The People of Puerto Rico. University of Illinois Press, Urbana.
Trouillot, M.-R., 1992. The Caribbean region: an open frontier in anthropological theory. Annual Reviews in Anthropology 21, 19–42.
Trouillot, M.-R., 2001. Caribbean: sociocultural aspects. In: Hannerz, U., Boyer, D. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Elsevier Science, pp. 1484–1488.
Wardle, H., 2007. A groundwork for West Indian cultural openness. Journal of the Royal Anthropological Institute (N.S.) 13, 567–583.
Wilson, P., 1969. Reputation and respectability: a suggestion for Caribbean ethnology. Man 4 (1), 70–84.
Wilson, P., 1973. Crab Antics: The Social Anthropology of English-Speaking Negro Societies of the Caribbean. Yale University Press, New Haven.

Amnesia: General

Kristoffer Romero and Morris Moscovitch, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Moscovitch, volume 1, pp. 1–10, © 2004, Elsevier Ltd.

Abstract

Amnesia refers specifically to the focal disruption of anterograde and retrograde memory, with no other major impairments or clinical features. The study of focal lesion patients has greatly advanced our conceptualization of memory systems and their neuroanatomical basis. Although these distinctions are still very relevant, recent research is challenging such distinctions, and the independence of memory systems that they imply. The authors review the major memory systems with relevant patient findings, and subsequently discuss new insights into how different types of memory are affected in amnesia. The authors finish with a discussion on the implications of recent research for theories of amnesia.

Memory loss is among the most widely studied cognitive disorders that occur following neurological insult. The classical amnesic syndrome is defined by a severe deficit in encoding new information (anterograde amnesia; AA) and by severely impaired recall of experiences that occurred some period of time before the trauma (retrograde amnesia; RA), with other cognitive abilities remaining relatively intact. Thus, following recovery from brain damage, if a patient with amnesia is presented with new information, he or she can keep a limited amount in mind if it is actively rehearsed, but that information will not be retained beyond the immediate moment once the patient is distracted. Moreover, if asked about a past event such as a recent holiday or a particular birthday, the patient may only produce a few general details, which are typically factual in nature and devoid of the rich detail associated with vividly recalling a past episode. Yet, when asked about previously learned knowledge, such as types of animals or historical events, patients with amnesia are not as impaired. Aside from these marked impairments in AA and RA, the amnesic patient has no major deficits in general intelligence, simple attention, language comprehension or production, visual perception, problem solving, planning, or regulating behavior. Thus, the description of the classical amnesic syndrome is one of a selective impairment in memory.

Memory loss is a common sequela following damage to a variety of structures, including the medial temporal lobes (MTLs), subcortical structures, frontal lobes, and even the parietal lobes (for a review of memory loss due to damage in other neural substrates, see Berryhill, 2012; Fama et al., 2012; Moscovitch and Winocur, 2002). This article focuses on the classical amnesic syndrome due to damage in the hippocampus, a deep structure in the MTL. In terms of anatomy, major pathways from the hippocampus project via the fornix to the mammillary bodies, anterior and dorsomedial thalamic nuclei, and eventually the prefrontal cortex. Some projections from the medial thalamic nuclei project back to the hippocampus via the cingulum, thus completing the limbic circuit (Brown and Aggleton, 2001). There are also reciprocal connections between the hippocampus and entorhinal cortex, and between left and right hippocampi.

The hippocampus is particularly susceptible to various types of neurological insults. Viral infections (Cermak and O'Connor, 1983), oxygen deprivation (i.e., hypoxia; Di Paola et al., 2008), epilepsy (Lehéricy et al., 1997), traumatic brain injury (McKee et al., 2009), electroconvulsive shock (Nakamura et al., 2013), and cerebrovascular accidents (i.e., strokes) can affect hippocampal tissue (Lim and Alexander, 2009). Amnesia can also result from damage to any regions in the limbic circuit: For example, diencephalic atrophy is a common consequence of Korsakoff's syndrome, which presents with AA, RA, and concomitant personality and cognitive features (Fama et al., 2012). Moreover, lesions to the fornix or thalamus are also sufficient to produce amnesia (Carlesimo et al., 2011).

In their seminal paper, Scoville and Milner (1957) reported on patient H.M., who required surgical resection of the MTL bilaterally to treat intractable seizures. H.M's profound memory deficits for new information, and subsequent studies of human patients and nonhuman primates, have shown that the hippocampus is necessary for encoding new information in order to be remembered subsequently (Alvarez and Squire, 1994; Gaffan and Murray, 1992). This is due to the anatomical position of the hippocampus, which receives input from various cortical regions via the entorhinal cortex, and thus is ideally situated to bind new information together into a multimodal representation (Squire, 1992).

Anterograde Memory Distinctions

Short-Term Memory and Long-Term Memory

Since the initial study on H.M., several key distinctions have been made in terms of different types of memory. In terms of the anterograde aspects of amnesia, one of the earliest distinctions was between short-term memory (STM) or primary memory for information held in mind shortly after presentation, and long-term memory (LTM), or secondary memory for information no longer held in mind but is retained over longer periods of time. Early studies showed that amnesics were generally comparable to healthy controls in terms of the amount of information they could keep in mind. However, once information was displaced with new stimuli, or after a given period of time had passed in which rehearsal was prevented, amnesics showed a discontinuous drop in memory performance. For example, both H.M. and healthy controls

could accurately recall a string of about six digits immediately after hearing them. However, with sufficient practice, healthy controls could recall an increasingly large number of digits (i.e., up to 20), whereas H.M. could not increase the number of digits he could recall beyond seven, despite dozens of repetitions (Drachman and Arbit, 1966). In addition, the ability to actively maintain and manipulate information in mind (i.e., working memory; WM), was generally found to be intact in amnesia (Baddeley, 2012; Cowan, 2001). For example, when shown a simple pattern of black and white squares and asked to recall it after 0, 4, 8, or 12 s, amnesics performed normally compared to healthy controls over short delays (Owen et al., 1995). Conversely, patients with lesions in the lateral temporal or parietal cortices showed deficits in STM but not in LTM: this demonstration of a double dissociation between lateral temporal and medial temporal regions for STM and LTM, respectively, suggests that these two aspects of memory are indeed separate, and that only the latter type required the hippocampus.

Memory above or below Conscious Awareness: Implicit and Explicit Memory

Another distinction was demonstrated between what is termed 'explicit' or 'declarative' memory and 'implicit' or 'non-declarative' memory. The former refers to any memory we are consciously aware of having, be it knowing the capital of France, or recalling a past summer barbeque. Implicit memory, conversely, refers to stored information that is retained or successfully retrieved without conscious awareness: crucially, although this information is retained subconsciously, it can nonetheless influence behavior. In general, despite their impairments in explicit memory, amnesics were largely intact on measures of implicit memory, such as learning a new skill. For example, amnesic people can learn a novel skill such as reading inverted text at a comparable rate to controls (Cohen and Squire, 1980; Moscovitch et al., 1986). Moreover, both amnesic people and controls show good retention of both the general skill and of particular exemplars, in that they can read them more efficiently when tested at a later time. However, although patients can read a list of inverted words more quickly and accurately, their explicit memory for those words will be poor following a delay, demonstrating their intact implicit memory for the skill of reading inverted text in general and specific exemplars in particular, but impaired explicit memory for what they read.

Amnesic people also respond more accurately or quickly to stimuli they had experienced before, but cannot recognize subsequently. An example is priming, where an individual responds more quickly to a previously presented word or concept, even if the presentation was very brief, and thus not consciously perceived. In the word-stem completion task, participants are shown a series of words e.g., HORSE, then subsequently shown a series of word stems e.g., HO____, and are asked to complete the word with the first one that comes to mind, as quickly as possible. Amnesic people are just as likely as healthy controls to fill in the stem HO____ with HORSE, rather than another word, such as HOUSE, even though they have no memory of seeing a list of words. Overall, these findings suggested that newly encountered information can be retained and recalled not only explicitly, as in recalling a past event memory, but also implicitly, such that one can have an enduring trace of memory that influences behavior subconsciously. The studies of priming and skill acquisition in amnesia showed that only explicit memory was impaired following hippocampal damage, whereas implicit memory, even for new associations, was spared (Goshen-Gottstein et al., 2000).

Retrograde Memory Distinctions

Semantic Facts and Episodic Experiences

Although it became clear that amnesia was characterized by AA for learning new explicit memory, patients also showed memory deficits for information acquired prior to their brain damage (i.e., RA). Here, too, there were perplexing inconsistencies. Specifically, variable evidence existed regarding (1) the types of explicit memories and (2) the age of the memories affected.

With regard to the former, it is important to distinguish between the amnesia for semantic and episodic memory (Tulving, 1972). Semantic memories are those pertaining to general knowledge, and are recalled as noncontextual facts without memory for a corresponding situation. Thus, when shown the word 'apple,' one can recall that an apple is a fruit with several varieties, but cannot typically recall when such facts were learned. In contrast, episodic memories contain information pertaining to specific life experiences, such as seeing the word 'apple' and recollecting the last time you had apple pie. Crucially, because episodic memories pertain to experiences, their recollection has the quality of autonoetic consciousness, or a feeling of reexperiencing the actual episode, whereas semantic memories have only noetic consciousness, or awareness only that the memory exists but no associated feeling of reexperiencing. Thus, amnesic people generally have poor episodic memory for autobiographical events spanning years to decades prior to their neurological insult. However, their memory for general knowledge, such as facts learned in school and well-learned information such family members' names, was less impaired. This impression was confirmed by early double dissociation studies showing that amnesic people were able to recall previously learned semantic facts for most of their life, except the more recent periods before their brain damage, extending for up to 10 years. Conversely, patients with damage to the anterior temporal lobes showed a loss of semantic memory, such as the names of objects in varying categories; however, their memory for past experiences remained relatively preserved (Hodges and Patterson, 2007). Similar patterns have been observed in Alzheimer's dementia patients, who have damage to the hippocampus early in the disease progression, and patients with semantic dementia, who have degradation of the anterior temporal lobes, and relative sparing of the hippocampi early in the disease. Alzheimer's patients show episodic memory loss and relatively intact semantic memory, whereas semantic dementia patients show relatively spared episodic memory and degraded semantic memory (Hodges and Patterson, 2007; Irish et al., 2011). Another example of the episodic-semantic distinction is in patients with

developmental amnesia, who sustain hippocampal damage shortly after birth. Although these patients typically have profound memory loss for their experiences, they can nonetheless acquire new facts, develop language skills, and progress in school at a level far above what would be expected based on the severity of their memory impairment (De Haan et al., 2006). Thus, such patients can go to class and acquire new information sufficient to complete a history test, but will have no memory of attending class or any classroom antics that may have occurred.

Autobiographical Recall of Recent and Remote Memories

Another contentious issue has been the extent to which amnesia will affect past memories. Specifically, there is a debate as to whether hippocampal lesions only affect recent memories (i.e., within the past few years) or both recent and remote memories (i.e., those stretching back to childhood). Early reports of patients with global amnesia caused by focal MTL lesions or Alzheimer's disease suggested that their RA follows a reverse gradient, with impoverished recent memories and largely intact remote memories (i.e., Ribot's law), suggesting that recent and remote memories were different in some way. These and other findings were consistent with a standard consolidation theory (SCT) of memory (Squire and Wixted, 2011), which states that the hippocampus helps initially to encode information into a lasting memory trace and to retrieve that memory. However, the hippocampus' influence will decrease over time as memory is eventually coded in the neocortex, during which time the memory can be retrieved without the hippocampus. Thus, recent memories are disproportionately affected in amnesia because they are still in the process of consolidation, which requires the hippocampus. Conversely, remote memories have been consolidated in the neocortex, and no longer require the hippocampus to be retained and retrieved. This would be the case for any explicit memory, be it episodic or semantic in nature. Evidence for SCT came from early work on RA (Moscovitch, 1982; Squire, 1992), and confirmed by later studies showing a severe loss of recent memories, with a varying extent of time period affected, but seemingly intact remote memories (Squire and Bayley, 2007). In addition, the generally poor ability for amnesics to encode new episodic or semantic information suggested that all memories first undergo a hippocampally dependent consolidation period (Squire and Wixted, 2011).

Recent evidence suggests, however, that although this model may apply reasonably well to semantic memory (but see Sharon et al., 2011), it fares less well with regard to episodic memory. Although people with amnesia may retain some memories for personal episodes, they are fewer and less rich and detailed than those of matched controls (Moscovitch et al., 2005; Moscovitch et al., 2006). The source of the discrepancy on the continuing contribution of the hippocampus to long-term retention, and retrieval of detailed autobiographical memories, is not known. The two leading candidates are differences in the site and extent of the lesions and the manner in which the memory was assessed, but neither has been successful in resolving all the discrepancies.

Updates on Current Memory Models: Exceptions and New Directions

Research on amnesia and the hippocampus has provided very useful distinctions as to the organization of memory, with the resultant models being a fruitful source of research. Although these distinctions generally hold, new theories based on recent findings in cognitive neuroscience have blurred the lines between different memory domains. In particular, earlier distinctions depended more on models based on memory stores, such as STM and LTM, their longevity (recent vs remote), and independent systems, such as implicit vs explicit memory. Once distinctions between implicit and explicit memory were questioned; however, it suggested that perhaps the preserved and impaired memory in amnesia was determined by the nature of memory representations. Similarly, as the distinction between STM and LTM began to be eroded by recent models describing the performance of neurologically intact people, so too did the functional dissociations in amnesia. Thus, several current lines of work suggest that the memory systems previously described may not be as independent of one another as had once been assumed, nor are they always easily dissociable in amnesia.

Is the Hippocampus Working during Working Memory?

Early theories of WM suggested information was held in temporary storage systems (i.e., 'buffers') separate from LTM, and acted upon by top-down control processes. Thus, retention of incoming information in WM existed separately from information in LTM, and would be subject to forgetting if not actively rehearsed or processed (Baddeley, 2012). However, other models eliminated the notion that WM buffers are separate from LTM. Instead, unitary-store models suggested WM is a limited capacity system that devotes attention to information represented in LTM. Thus, information in WM is merely activated LTM, rather than existing in a separate representational system. Crucially, the relative amount of information that may be kept in WM depends on factors such as the familiarity of incoming information, suggesting that more information can be kept in mind if it is based on preexisting LTM representations or knowledge (Cowan, 2001). If this is the case, it should be possible to find deficits in WM in amnesics under certain conditions (for reviews, see Jonides et al., 2008; Ranganath and Blumenfeld, 2005).

Indeed, Olson et al. (2006) had amnesic people and healthy controls remember three objects, three locations, or the three object–location combinations in a 3 × 3 array shown on a screen. Amnesic people showed no deficits when tested 1 s after the stimuli was presented, but could not recall object–location combinations after an 8-s delay. Similarly, Hannula et al. (2006) presented amnesic people with a continuous stream of relational information (i.e., face–scene pairings and location of items in a scene), and were probed for memory for the associations after one, five, or nine trials had elapsed. Although memory for face–scene pairs was intact after one intervening trial, amnesic people showed poor memory after as few as five trials.

These findings suggest that it may be inaccurate to characterize amnesia as an impairment of LTM only, and that under

certain conditions WM performance may also be impaired, particularly if the stimuli are novel or complex, such as a busy scene or a human face (Ezzyat and Olson, 2008; Hannula et al., 2006). Simple WM tests used in early studies of amnesic people often used digits or highly familiar words that are well learned, and thus their retention in WM may have been supported by extrahippocampal cortical structures responsible for their perceptual processing (Ranganath and Blumenfeld, 2005). Another interpretation is that WM tasks with sufficiently complex stimuli are more resource demanding, and thus require LTM as well as WM. Consequently, the amnesic deficit under these conditions reflects impaired LTM ability, rather than impaired WM per se (Jeneson and Squire, 2012; Warrington and Taylor, 1973). Regardless of the correct interpretation, it is increasingly clear that whether newly presented information is being stored temporarily in STM, or has been encoded into LTM is not as easy to discern as previously thought, and can be mediated by the nature of the memoranda (Ezzyat and Olson, 2008; Warrington and Taylor, 1973; see also Waugh and Norman, 1965).

Remembering Secretly: Hippocampal Implications during Implicit Memory

Another issue is whether the hippocampus is only involved in explicit memory, or whether it supports memory regardless of conscious awareness. Although early studies supported the independence of implicit and explicit memory, such findings of spared implicit memory were not always consistent, and depended on specific conditions for different tasks. Moreover, some suggested that priming effects were inflated in patient groups because of their lower baseline priming performance, as priming is always calculated as improved performance after exposure to stimuli, compared to no exposure (Ostergaard, 1994). Furthermore, subsequent analyses of priming showed that explicit memory sometimes contaminated performance on implicit memory tests in the healthy controls. For example, in Gooding et al. (1999), some subjects reported being aware that previously studied words were being tested in word-stem completion, with some even reporting purposely not using the previously studied words. These inconsistencies suggested that these tests did not always yield pure measures of implicit memory, weakening the notion that amnesic people have normal implicit memory.

Recent studies using more nuanced measurements have shown that hippocampal damage can indeed affect implicit memory. For example, Ryan et al. (2000) recorded participants' eye movements while viewing complex nature scenes. Critically, the researchers made very quick, subtle changes to the scenes that could not be consciously perceived. Controls made many new eye movements to the location of the scene change, whereas amnesics did not consistently show such behavior, suggesting they did not perceive the change because they did not have a coherent representation of the previous scene. Similarly, Hannula and Ranganath (2009) measured hippocampal activation and eye movements in healthy controls who previously viewed pairs of faces and scenes. During a later recognition test, subjects were shown the scene with several faces and asked to choose which face was previously shown with the scene. Subjects showed eye movements toward the location of the correctly paired face up to 1000 ms before explicitly recalling the correct face paired with that scene. Surprisingly, even on some trials when subjects incorrectly identified the proper face, they nonetheless made eye movements toward the correct face, and these eye movements were correlated with hippocampal activity during these trials.

Finally, there is evidence to suggest that implicit and explicit memory interact. Sheldon and Moscovitch (2010) had subjects study a set of words, and complete an implicit and explicit memory test with those words. In the implicit test, subjects viewed a series of rapidly presented words or nonwords and decided as quickly as possible whether the word was real or not (i.e., lexical decision). Subsequently, subjects were shown the study words with other new words and asked if they recognized seeing them before; if so, they were further asked whether they recalled an associated episodic detail with that word (i.e., recollection), or just knew they had seen it before but did not remember the experience of seeing the word (i.e., familiarity). Numerous studies using this recollection/familiarity distinction have shown the hippocampus is important for recollection, but not familiarity (Cohn et al., 2009; but see Kirwan et al., 2010). Curiously, words that were recollected were identified more quickly during the lexical decision task than familiar words, suggesting that hippocampally mediated processes may underlie performance on both explicit and implicit memory measures. Overall, these findings question whether conscious awareness is an absolute rule in determining which memory processes require the hippocampus (Dew and Cabeza, 2011; Hannula and Greene, 2012).

Experiences as Facts and Vice Versa: The Interdependence of Episodic Memory and Semantic Memory

Although there is wide agreement regarding the distinction between semantic and episodic information, these systems nonetheless are interdependent, which was evident even in the earliest theories delineating episodic memory and semantic memory (Tulving, 1972). Indeed, patients with hippocampal damage sustained in adulthood typically show very poor encoding of new semantic facts, often requiring many trials in order to retain new knowledge. Moreover, their ability to acquire new knowledge seems to correlate with MTL volume (Greenberg and Verfaellie, 2010), suggesting that at initial encoding both episodic memory and semantic memory are dependent on the hippocampus. However, by using mnemonic strategies such as errorless learning or fast mapping, amnesics can demonstrate better or even normal acquisition of new knowledge, although whether such information is retained over the long term is not always clear (McKenna and Gerhand, 2002; Sharon et al., 2011).

Even studies of healthy subjects report interactions between semantic memory and episodic memory. For example, newly encountered information that is congruent with previous knowledge is better recalled than incongruent information, demonstrating how semantic memory influences episodic memory (Craik and Tulving, 1975). More recently, amnesic people have also demonstrated this effect; when asked to remember grocery items with realistic or unrealistic prices,

amnesic people with damage limited to the MTLs showed the congruency effect, whereas those with lesions extending into lateral temporal cortices did not (Kan et al., 2009).

There is also evidence to suggest that episodic memory influences semantic memory at retrieval. For example, when shown faces and asked to identify whether they are famous or nonfamous, participants are quicker to identify a famous face if they have a specific episodic memory associated with that face, compared to faces they know are famous, but are not tied to any specific personal memory (Westmacott and Moscovitch, 2003). Finally, patient and functional magnetic resonance imaging (fMRI) studies both point to a role for the hippocampus during semantic retrieval tasks such as category fluency, where subjects are given a category and asked to think of as many exemplars from that category as possible (Greenberg et al., 2009; Sheldon and Moscovitch, 2012). Notably, some categories can be thought of as more episodic-like in nature, rather than strictly noncontextual facts. For example, thinking of items found in a kitchen may prompt one also to recall their own kitchen, which engages episodic memory, rather than searching through semantic memory only. Overall, it is clear that although semantic information can be accessed independently of the hippocampus, knowledge is inevitably intertwined with past experiences, and the two may complement one another at encoding or at retrieval.

Recent and Remote Memories: A Lasting Role for the Hippocampus?

Although initially a good description of RA, the SCT model, as noted, failed to account for some findings of extensive RA for episodic memories, which were recently confirmed (Moscovitch et al., 2005; Winocur and Moscovitch, 2011). Consequently, an alternate view is that the hippocampus will always be necessary for retrieving past episodic memories, regardless of the age of the memory. Multiple trace theory suggests that the hippocampus provides retrieval traces or tags to guide memory search for a particular memory, which are always required to recover a memory stored in the neocortex (Moscovitch et al., 2005). Previous studies of RA did not precisely consider that past experiences (i.e., autobiographical memories) inherently consist of both semantic and episodic details. Because previous tests of autobiographical memory did not distinguish between episodic and semantic memories, amnesia appeared to affect recent memories much more than remote memories.

Moreover, if one more carefully looks at semantic memory and episodic memory separately, amnesia tended to spare semantic memory for all life periods except the most recent ones. In contrast, specific details (i.e., sights and emotions) and crucially, the subjective reexperiencing associated with episodic memory, tended to be impoverished following hippocampal damage, regardless of whether the memory was recent or remote (Cipilotti et al., 2001; Viskontas et al., 2000). Parallel patterns are seen in patients with Alzheimer's dementia and semantic dementia: in Alzheimer's dementia, episodic memory for recent and remote events is impaired but semantic memory is relatively preserved, while in the early stages of semantic dementia, remote and recent episodic memories are relatively preserved but semantic memory is impaired (Piolino et al., 2003; Westmacott and Moscovitch, 2003). Finally, fMRI studies of autobiographical retrieval showed that it is qualitative factors related to the subjective reliving of past events (i.e., vividness, perceptual detail) that mediate hippocampal involvement in autobiographical recall (Gilboa et al., 2004; St. Laurent et al., 2009), and not the oldness of the memory per se (Bonnici et al., 2012).

It should be noted that older memories do tend to be less detailed and more schematic. The latter observation led researchers to emphasize that memories are transformed over time and experience, such that as memories become more semantic or schematic, they become independent of the hippocampus and rely instead on extrahippocampal structures (Winocur and Moscovitch, 2011). Insofar as memories retain their perceptual details, they continue to depend on the hippocampus.

Looking Backward to Move Forward: Memory-Related Neural Substrates and Future Thinking

Until recently, human memory research has focused on the study of memory for memory's sake, rather than how memory systems contribute to behavior. Bartlett (1932) noted that the quality of memory is fluid, and susceptible to distortion by cultural factors. In the animal memory literature, it was always assumed that memory conferred great survival value, be it avoiding previous dangerous encounters, or recalling where food was cached (Suddendorf and Corballis, 2007). Similarly, theorists have posited that the ability of humans to reminisce is a special case of the ability to engage in mental time travel, which may go backward, as in autobiographical recall, or forward as in when we try to predict some future scenario. The ability to use past experiences to predict future encounters may confer the benefit of simulating the outcomes of possible actions, without suffering the consequences of actually implementing them if they are undesirable (Boyer, 2008; Buckner and Carroll, 2007; Gilbert and Wilson, 2007; Addis and Schacter, 2012).

Hassabis et al. (2007) found that amnesic people showed deficits in imagining a new experience, such as being at the beach, in that they were not able to describe the event beyond generalities, and had particular deficits in describing spatial details indicative of forming a cohesive representation. Similarly, Addis and colleagues looked at patients with Alzheimer's dementia and found that the hippocampus is needed to imagine future events, just as it is needed to recall past events. Strikingly, the ability to imagine future events seems to mirror autobiographical recall performance in these patients; that is, they can describe the general semantics of events, but cannot describe the details of those events, whether past memories or imagined future scenarios (Addis et al., 2009). Moreover, the ability to construct imagined events also relates to subsequent memory of those events (Romero and Moscovitch, 2012). The propensity to imagine future events that are more semantic (e.g., what do you think the economy will be like in a few years?) is also compromised in amnesic people at the level of details, but not at the level of generalities (Race et al., 2013). Finally, fMRI studies show that the hippocampus is activated during autobiographical recall and future imagining, with specific

subregions of activation for each task (Martin et al., 2011). Other studies have extended these findings to show that the hippocampus aids in simulating possible outcomes with varying levels of value (Benoit et al., 2011), suggesting that the ability to recall past experiences in detail serves purposes other than reminiscence (Pillemer, 2003). One implication of these findings is that the hippocampus may not be best characterized as a memory-specific brain region but rather as a region that is implicated in a host of functions that draw on detailed episodic memories for optimal performance.

Summary

Studying amnesic patients has resulted in several key observations in memory, namely, that memory was not a unitary construct, but had several facets that could operate independently, and rely on separate neural structures. The characterization of amnesia as a disorder only affecting memory was based on the magnitude of the observed effects in LTM, as compared to other cognitive functions. However, recent evidence suggests the characterization of the hippocampus in cognition extends beyond a single memory system (LTM), and may perhaps be better characterized by the nature of the representations it mediates, rather than a discrete memory system (Graham et al., 2010; Henke, 2010). Nevertheless, the notion of distinct memory systems remains useful as a paradigm for characterizing amnesia and other cognitive disorders. As the understanding of brain behavior mappings becomes more precise, the conception of cognitive disorders will also undergo further refinement.

See also: Autobiographical Memory; Confabulation; Dysexecutive Amnesia; Episodic Memory; Future Thinking; Hippocampal Amnesia; Implicit Memory; Recognition; Semantic Memory; Working Memory, Neural Basis of; Working Memory, Psychology of.

Bibliography

Addis, D.R., Sacchetti, D.C., Ally, B.A., Budson, A.E., Schacter, D.L., 2009. Episodic simulation of future events is impaired in mild Alzheimer's disease. Neuropsychologia 47, 2660–2671.

Addis, D.R., Schacter, D.L., 2012. The hippocampus and imagining the future: where do we stand? Frontiers in Human Neuroscience 5, 1–15.

Alvarez, P., Squire, L.R., 1994. Memory consolidation and the medial temporal lobe: a simple network model. Proceedings of the National Academy of Sciences of the United States of America 91 (15), 7041–7045.

Baddeley, A., 2012. Working memory: theories, models, and controversies. Annual Review of Psychology 63, 1–29.

Bartlett, F.C., 1932. Remembering: A Study in Experimental and Social Psychology. Cambridge University Press, Cambridge, England.

Benoit, R.G., Gilbert, S.J., Burgess, P.W., 2011. A neural mechanism mediating the impact of episodic prospection on farsighted decisions. Journal of Neuroscience 31 (18), 6771–6779.

Berryhill, M.E., 2012. Insights from neuropsychology: pinpointing the role of the posterior parietal cortex in episodic and working memory. Frontiers in Integrative Neuroscience 6, 1–12.

Bonnici, H.M., Chadwick, M.J., Lutti, A., Hassabis, D., Weiskopf, N., et al., 2012. Detecting representations of recent and remote autobiographical memories in vmPFC and hippocampus. The Journal of Neuroscience 32 (47), 16982–16991.

Boyer, P., 2008. The evolutionary economics of mental time travel. Trends in Cognitive Sciences 12 (6), 219–224.

Brown, M.W., Aggleton, J.P., 2001. Recognition memory: what are the roles of the perirhinal cortex and hippocampus. Nature Reviews Neuroscience 2, 51–61.

Buckner, R.L., Carroll, D.C., 2007. Self-projection and the brain. Trends in Cognitive Sciences 11, 49–57.

Carlesimo, G.A., Lombardi, M.G., Caltagirone, C., 2011. Vascular thalamic amnesia: a reappraisal. Neuropsychologia 49 (5), 777–789.

Cermak, L.S., O'Connor, M., 1983. The anterograde and retrograde retrieval ability of a patient with amnesia due to encephalitis. Neuropsychologia 19, 213–224.

Cipilotti, L., Shallice, T., Chan, D., Fox, N., Scahill, R., et al., 2001. Long-term retrograde amnesia: the crucial role of the hippocampus. Neuropsychologia 39, 151–172.

Cohen, N., Squire, L.R., 1980. Preserved learning and retention of pattern analyzing skill in amnesia: dissociation of knowing how and knowing that. Science 210, 207–209.

Cohn, M., Moscovitch, M., Lahat, A., Mcandrews, M.P., 2009. Recollection versus strength as the primary determinant of hippocampal engagement at retrieval. Proceedings of the National Academy of Sciences 106 (52), 22451–22455.

Cowan, N., 2001. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. The Behavioral and Brain Sciences 24 (1), 87–114.

Craik, F.I., Tulving, E., 1975. Depth of processing and the retention of words in episodic memory. Journal of Experimental Psychology-General 104, 268–294.

De Haan, M., Mishkin, M., Baldeweg, T., Vargha-Khadem, F., 2006. Human memory development and its dysfunction after early hippocampal injury. Trends in Neurosciences 29 (7), 374–381.

Dew, I.T., Cabeza, R., 2011. The porous boundaries between explicit and implicit memory: behavioral and neural evidence. Annals of the New York Academy of Sciences 1224, 174–190.

Di Paola, M., Caltagirone, C., Fadda, L., Sabatini, U., Serra, L., Carlesimo, G.A., 2008. Hippocampal atrophy is the critical brain change in patients with hypoxic amnesia. Hippocampus 18 (7), 719–728.

Drachman, D.C., Arbit, J., 1966. Memory and the hippocampal complex. Archives of Neurology 15, 52–61.

Ezzyat, Y., Olson, I.R., 2008. The medial temporal lobe and visual working memory: comparisons across tasks, delays, and visual similarity. Cognitive, Affective, & Behavioral Neuroscience 8 (1), 32–40.

Fama, R., Pitel, A., Sullivan, E.V., 2012. Anterograde episodic memory in Korsakoff syndrome. Neuropsychology Review 22, 93–104.

Gaffan, D., Murray, E.A., 1992. Monkeys with rhinal cortex lesions succeed in object discrimination learning despite 24-hour intertrial intervals and fail at matching to sample despite double sample presentation. Behavioral Neuroscience 106, 30–38.

Gilbert, D.T., Wilson, T.D., 2007. Prospection: experiencing the future. Science 317, 1351–1354.

Gilboa, A., Winocur, G., Grady, C.L., Hevenor, S.J., Moscovitch, M., 2004. Remembering our past: functional neuroanatomy of recollection of recent and very remote personal events. Cerebral Cortex 14, 1214–1225.

Gooding, P.A., Mayes, A.R., van Eijk, R., Meudell, P.R., MacDonald, F.L., 1999. Do novel associative word stem completion and cued recall share the same memory retrieval process? Memory 7, 323–343.

Goshen-Gottstein, T., Moscovitch, M., Melo, B., 2000. Intact implicit memory for newly formed verbal associations in amnesic patients following single study trials. Neuropsychology 14 (4), 570–578.

Graham, K.S., Barense, M.D., Lee, A.C.H., 2010. Going beyond LTM in the MTL: a synthesis of neuropsychological and neuroimaging findings on the role of the medial temporal lobe in memory and perception. Neuropsychologia 48 (4), 831–853.

Greenberg, D.L., Verfaellie, M., 2010. Interdependence of episodic and semantic memory: evidence from neuropsychology. Journal of the International Neuropsychological Society 16 (5), 748–753.

Greenberg, D.L., Keane, M.M., Ryan, L., Verfaellie, M., 2009. Impaired category fluency in medial temporal lobe amnesia: the role of episodic memory. Journal of Neuroscience 29 (35), 10900–10908.

Hannula, D.E., Greene, A.J., 2012. The hippocampus reevaluated in unconscious learning and memory: at a tipping point? Frontiers in Human Neuroscience 6, 1–20.

Hannula, D.E., Ranganath, C., 2009. The eyes have it: hippocampal activity predicts expressions of memory in eye movements. Neuron 63, 592–599.

Hannula, D.E., Tranel, D., Cohen, N.J., 2006. The long and the short of it: relational memory impairments in amnesia, even at short lags. The Journal of Neuroscience 26 (32), 8352–8359.

Hassabis, D., Kumuran, D., Vann, D.S., Maguire, E.A., 2007. Patients with hippocampal amnesia cannot imagine new experiences. Proceedings of the National Academy of Sciences USA 104, 1726–1731.

Henke, K., 2010. A model for memory systems based on processing modes rather than consciousness. Nature Reviews Neuroscience 11, 523–532.

Hodges, J.R., Patterson, K., 2007. Semantic dementia: a unique clinicopathological syndrome. Lancet Neurology 6 (11), 1004–1014.

Irish, M., Hornberger, M., Lah, S., Miller, L., Pengas, G., et al., 2011. Profiles of recent autobiographical memory retrieval in semantic dementia, behavioural-variant frontotemporal dementia, and Alzheimer's disease. Neuropsychologia 49 (9), 2694–2702.

Jeneson, A., Squire, L.R., 2012. Working memory, long-term memory, and medial temporal lobe function. Learning & Memory 19, 15–25.

Jonides, J., Lewis, R.L., Nee, D.E., Lustig, C.A., Berman, M.G., Moore, K.S., 2008. The mind and brain of short-term memory. Annual Review of Psychology 59, 193–224.

Kan, I.P., Alexander, M.P., Verfaellie, M., 2009. Contribution of prior semantic knowledge to new episodic learning in amnesia. Journal of Cognitive Neuroscience 21 (5), 938–944.

Kirwan, C.B., Wixted, J.T., Squire, L.R., 2010. A demonstration that the hippocampus supports both recollection and familiarity. Proceedings of the National Academy of Sciences 107 (1), 334–348.

Lehéricy, S., Semah, F., Hasboun, D., Dormont, D., Clémenceau, S., et al., 1997. Temporal lobe epilepsy with varying severity: MRI study of 222 patients. Neuroradiology 39 (11), 788–796.

Lim, C., Alexander, M.P., 2009. Stroke and episodic memory disorders. Neuropsychologia 47 (14), 3045–3058.

Martin, V.C., Schacter, D.L., Corballis, M.C., Addis, D.R., 2011. A role for the hippocampus in encoding simulations of future events. Proceedings of the National Academy of Sciences of the United States of America 108 (33), 13858–13863.

McKee, A.C., Cantu, R.C., Nowinski, C.J., Hedley-Whyte, E.T., Gavett, B.E., Budson, A.E., Santini, V.E., et al., 2009. Chronic traumatic encephalopathy in athletes: progressive tauopathy after repetitive head injury. Journal of Neuropathology and Experimental Neurology 68 (7), 709–735.

McKenna, P., Gerhand, S., 2002. Preserved semantic learning in an amnesic patient. Cortex 38, 37–58.

Moscovitch, M., 1982. Multiple dissociations of function in amnesia. In: Cermak, L.S. (Ed.), Human Memory and Amnesia. Erlbaum Associates, Hillsdale, NJ, pp. 337–370.

Moscovitch, M., Nadel, L., Winocur, G., Gilboa, A., Rosenbaum, S.R., 2006. The cognitive neuroscience of remote episodic, semantic and spatial memory. Current Opinion in Neurobiology 16, 179–190.

Moscovitch, M., Rosenbaum, R.S., Gilboa, A., Addis, D.R., Westmacott, R., et al., 2005. Functional neuroanatomy of remote episodic, semantic, and spatial memory: a unified account based on multiple trace theory. Journal of Anatomy 207, 35–66.

Moscovitch, M., Winocur, G., 2002. The frontal cortex and working with memory. In: Stuss, D.T., Knight, R.T. (Eds.), The Frontal Lobes. Oxford University Press, Oxford, pp. 188–209.

Moscovitch, M., Winocur, G., McLachlan, D., 1986. Memory as assessed by recognition and reading time in normal and memory-impaired people with Alzheimer's disease and other neurological disorders. Journal of Experimental Psychology: General 115, 331–347.

Nakamura, K., Ito, M., Liu, Y., Seki, T., Suzuki, T., et al., 2013. Effects of single and repeated electroconvulsive stimulation on hippocampal cell proliferation and spontaneous behaviors in the rat. Brain Research 23, 88–97.

Olson, I.R., Page, K., Moore, K.S., Chatterjee, A., Verfaellie, M., 2006. Working memory for conjunctions relies on the medial temporal lobe. The Journal of Neuroscience 26 (17), 4596–4601.

Ostergaard, A.L., 1994. Dissociations between word priming effects in normal subjects and patients with memory disorders: multiple memory systems or retrieval? The Quarterly Journal of Experimental Psychology 47 (2), 331–364.

Owen, A.M., Sahakian, B.J., Semple, J., Polkey, C.E., Robbins, T.W., 1995. Visuo-spatial short-term recognition memory and learning after temporal lobe excisions, frontal lobe excisions or amygdalo-hippocampectomy in man. Neuropsychologia 33, 1–24.

Pillemer, D., 2003. Directive functions of autobiographical memory: the guiding power of the specific episode. Memory 11 (2), 193–202.

Piolino, P., Desgranges, B., Belliard, S., Matuszewski, V., Lalevee, C., et al., 2003. Autobiographical memory and autonoetic consciousness: Triple dissociation in neurodegenerative diseases. Brain 126, 2203–2219.

Race, E., Keane, M.M., Verfaellie, M., 2013. Losing sight of the future: impaired semantic prospection following medial temporal lobe lesions. Hippocampus 23, 268–277.

Ranganath, C., Blumenfeld, R.S., 2005. Doubts about double dissociations between short- and long-term memory. Trends in Cognitive Sciences 9 (8), 374–380.

Romero, K., Moscovitch, M., 2012. Event construction in aging and amnesia. Journal of Memory & Language 67 (2), 270–284.

Ryan, J.D., Althoff, R.R., Whitlow, S., Cohen, N.J., 2000. Amnesia is a deficit in relational memory. Psychological Science 11 (6), 454–461.

Scoville, W.B., Milner, B., 1957. Loss of recent memory after bilateral hippocampal lesions. Journal of Neurology, Neurosurgery, & Psychiatry 20, 11–21.

Sharon, T., Moscovitch, M., Gilboa, A., 2011. Rapid neocortical acquisition of long-term arbitrary associations are independent of the hippocampus. Proceedings of the National Academy of Sciences 108 (3), 1146–1151.

Sheldon, S.A., Moscovitch, M., 2010. Recollective performance advantages for implicit memory tasks. Memory 18, 681–697.

Sheldon, S., Moscovitch, M., 2012. The nature and time-course of medial temporal lobe contributions to semantic retrieval: an fMRI study on verbal fluency. Hippocampus 22 (6), 1451–1466.

Squire, L.R., 1992. Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. Psychological Review 99 (2), 195–231.

Squire, L.R., Bayley, P.J., 2007. The neuroscience of remote memory. Current Opinion in Neurobiology 17, 185–196.

Squire, L.R., Wixted, J.T., 2011. The cognitive neuroscience of human memory since H.M. Annual Review of Neuroscience 34, 259–288.

St-Laurent, M., Moscovitch, M., Levine, B., McAndrews, M.P., 2009. Determinants of autobiographical memory in patients with unilateral temporal lobe epilepsy or excisions. Neuropsychologia 47 (11), 2211–2221.

Suddendorf, T., Corballis, M.C., 2007. The evolution of foresight: what is mental time travel and is it unique to humans? Behavioral and Brain Sciences 30, 299–351.

Tulving, E., 1972. Episodic and semantic memory. In: Tulving, E., Donaldson, W. (Eds.), Organization of Memory. Academic Press, New York, pp. 381–403.

Viskontas, I.V., McAndrews, M.P., Moscovitch, M., 2000. Remote episodic memory deficits in patients with unilateral temporal lobe epilepsy and excisions. The Journal of Neuroscience 20 (15), 5853–5857.

Warrington, E.K., Taylor, A.M., 1973. The contribution of the right parietal lobe to object recognition. Cortex 9, 152–164.

Waugh, N.C., Norman, D.A., 1965. Primary memory. Psychological Review 72, 89–104.

Westmacott, R., Leach, L., Freedman, M., Moscovitch, M., 2001. Different patterns of autobiographical memory loss in semantic dementia and medial temporal lobe amnesia: a challenge to consolidation theory. Neurocase 7, 37–55.

Westmacott, R., Moscovitch, M., 2003. The contribution of autobiographical significance to semantic memory. Memory & Cognition 31, 761–774.

Winocur, G., Moscovitch, M., 2011. Memory transformation and systems consolidation. Journal of the International Neuropsychological Society 17, 766–780.

Amnesia: Psychogenic

Angelica Staniloiu, University of Bielefeld, Bielefeld, Germany
Hans J Markowitsch, University of Bielefeld, Bielefeld, Germany; and Hanse Institute for Advanced Study, Delmenhorst, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Amnesic disorders are multifaceted conditions, with respect to etiology, course, and prognosis. Herein, data on epidemiology, clinical and neuropsychological aspects, and neurobiology and neuroimaging of amnesia due to psychological mechanisms (psychogenic, dissociative, or functional amnesia) are reviewed. By emphasizing the commonalities and dissimilarities between 'psychogenic' and 'organic' amnesia, biopsychosocial models of illness and a common intellectual approach to both neurological and psychiatric diseases are endorsed.

Amnesic disorders have as signature characteristic a severe memory impairment (*see* Amnesia: General), which occurs in a responsive, alert individual, in the absence of other significant cognitive impairments (Markowitsch and Staniloiu, 2012a). They are multifaceted conditions, with various causes, presentations, and prognoses, being differentiated into acute and transient, and chronic (lasting for more than a month). The inability to consciously acquire new information for long-term storage was named anterograde amnesia, and the inability to consciously recall previously stored information was designated retrograde amnesia. The memory impairment in amnesic disorders is frequently selective with respect to content-based categorizations of long-term memory (Markowitsch and Staniloiu, 2012a). Often, the memory impairment exclusively or predominantly involves the episodic-autobiographical memory domain (*see* Episodic Memory). According to their preponderant etiological link, amnesic disorders are viewed as being due to a general medical condition, direct effects of substance use (persisting beyond the period of intoxication and withdrawal) or psychological mechanisms (dissociative amnesia), or as having a mixed or iatrogenic etiology.

The Thesaurus of Amnesic Disorders due to Prevailing Psychological Mechanisms

The group of amnesic disorders due to prevailing psychological mechanisms has over the years been captured by several constructs, such as hysterical, functional, psychogenic, dissociative, or medically unexplained amnesia, or mnestic block syndrome. In present international nosologies, dissociative amnesia and fugue replaced previous diagnostic entities of psychogenic amnesia and fugue, respectively (Spiegel et al., 2011). Being listed under the category of dissociative disorders, dissociative amnesia and its variant dissociative fugue are considered to be causally bound to psychological stress or trauma. They are regarded as having as grounds the psychological mechanism of dissociation, where the latter represents a disruption or failure of integration of otherwise integrated aspects of cognition (memory), perception, identity, consciousness, and emotion (Janet, 1907; Spiegel et al., 2011). Some authors submitted to criticism the construct 'dissociative amnesia' due to its peremptory theoretical load, favoring the use of other terms (McKay and Kopelman, 2009). The term psychogenic amnesia has subsequently survived in the literature, designating amnesic disorders due to various psychological mechanisms, including dissociation. The terminology 'functional amnesia' initially indicated the opposite of organic amnesia. It was then applied to amnesic conditions that occurred without evidence of significant brain damage and had unclear etiology (de Renzi et al., 1997). When brain damage was identified in functional amnesia by conventional structural imaging methods, the locus and/or severity of brain damage did not typically match the nature and extent of memory impairment (disproportionate amnesia) (Markowitsch and Staniloiu, 2013).

Despite some debates, several writers have used the terms psychogenic, dissociative, and functional amnesias interchangeably, implicitly acknowledging in this manner that many functional amnesias have a psychological mechanism (Fujiwara et al., 2008; Kritchevsky et al., 2004; Markowitsch and Staniloiu, 2013). The lack of readily identifiable psychological factors in some cases of functional amnesia was conjectured to have arisen from the amnesia itself (in functional retrograde amnesia), an impaired capacity for emotional awareness and clarity or the incubation effect of trauma (Markowitsch and Staniloiu, 2013). As opposed to other terms used for describing amnesic disorders due to prevailing psychological mechanisms, the term functional is regarded as possessing several advantages, which may explain a rejuvenated interest in it (Stone, 2009; Markowitsch and Staniloiu, 2013). These include circumventing the mind-brain debate and emphasizing the functional aspect of the illness for the patient (the flight into illness and potential gain from illness; Markowitsch and Staniloiu, 2013). Furthermore, the word functional may map onto the concept of dynamic lesion, which had been described by Charcot (see Markowitsch and Staniloiu, 2013) and has received support from functional imaging studies (Brand et al., 2009).

The construct 'mnestic block syndrome' was coined by Hans Markowitsch to name a form of retrograde (autobiographical) amnesia, without referring to a clear causation, but highlighting the role of massive or long-lasting stress or trauma in the development of the amnesic condition in these patients (e.g., Markowitsch, 1996, 1997, 1998, 2000, 2003; Markowitsch et al., 1999b, 2000; Brand and Markowitsch, 2009). While avoiding the mind-brain dispute, the construct 'mnestic block syndrome' alludes to mechanisms that are purported to be involved in this condition. The memories are not lost. Only the access to the memories is blocked throughout the period of the illness (Markowitsch, 1996; Spiegel et al., 2011). Similar to the expression 'functional amnesia', the term 'mnestic block syndrome' conveys hope for recovery or improvement (Stone, 2009).

Types of Functional or Psychogenic Amnesia

Functional amnesias are heterogeneous conditions, which could be triggered by psychological factors or a combination of psychological and somatic factors. The physical factors implicated in precipitating functional amnesias are usually of mild intensity (mild traumatic brain injury or mild electrocuting accident). Cases of functional amnesia, however, might also arise after more severe brain insults (Kopelman, 2000; Pommerenke et al., 2012; Fast and Fujiwara, 2001). However, in the presence of significant brain pathology (as detected by conventional brain imaging), the psychological factors contributing to the amnesic condition are more likely to be discounted or neglected and subsequently the condition may solely be attributed to an 'organic' cause (Kopelman, 2000).

Sometimes, functional amnesias occur suddenly after a massive psychological trauma; at other times there may be a period of lag between the psychologically traumatic incident and the onset of amnesia. Not uncommonly, functional amnesia was reported to occur after a seemingly objective minor incident; however, in most cases, a careful anamnesis unearthed a history of recurrent psychologically stressful or traumatic experiences, often with onset in childhood or early adulthood.

Functional amnesias can be classified into acute and chronic. Amnesias persisting over a month period were deemed chronic. This duration cutoff for chronicity may have stemmed from older studies, which showed that a high percentage of psychogenic amnesias recovered within a month from their onset (Abeles and Schilder, 1935). Traditionally, functional amnesias were viewed as being preponderantly acute and transient in many countries outside of the North America (Spiegel et al., 2011). For the last two decades, several authors, however, described a chronic course of memory impairments in a substantial number of patients with functional or psychogenic amnesia (Fujiwara et al., 2008; Staniloiu and Markowitsch, 2012; Hennig-Fast et al., 2008). The conflicting findings with respect to the course of amnesia may be accounted for by several factors, such as the referral bias and the progress made in the conceptualization of episodic-autobiographical memory and the development of neuropsychological methods for detecting memory impairments (Kritchevsky et al., 2004; Markowitsch and Staniloiu, 2013).

Epidemiological and Demographic Data

Epidemiological data about psychogenic or functional amnesia are still limited. The prevalence values for dissociative amnesia were recently stated to range between 1.8 and 7.3% (Spiegel et al., 2011). Johnson et al. (2006) reported for dissociative amnesia a 12-month prevalence of 1.8% in a US community of 658 adults who were evaluated with psychiatric interview. Sar et al. (2007) conducted a study in Turkey where 43 of the 97 consecutive outpatients admitted to the psychiatric emergency department were initially screened for dissociative experiences using the screening tool Dissociative Experiences Scale (DES). The patients who had a DES score higher than 25 (39.5% of 43 assessed) were further evaluated with the Dissociative Disorders Interview Schedule and then the Structured Clinical Interview for Dissociative Disorders (the gold standard diagnostic instrument for dissociative disorders). Seven percent (3 patients) of the 43 patients met diagnostic criteria for dissociative amnesia. The average DES score of patients who met criteria for dissociative disorders was reported to be 43.7.

These above-mentioned data pose a challenge to older views of dissociative amnesia as being extremely rare, and subsequently draw attention to the risk of underdiagnosing or misdiagnosing this condition (Spiegel, 2006). Coons and Milstein (1992) had remarked that psychogenic amnesia may be missed when subtler clinical pictures are present or when other primary psychiatric conditions coexist. Furthermore, the explicit emphasis that the ICD-10 (1992) lays on the requirement to find 'convincing associations in time' between the symptoms of the dissociative amnesia and 'stressful events, problems, or needs' may lead some rushed clinicians or researchers to erroneously drop this condition into other diagnostic categories. As Kopelman (2000) noted, the uncovering of precipitating psychological factors in cases of amnesia is not always simple (Markowitsch et al., 1993). Furthermore, "sometimes patients will not allow you to know what had been happening, or they may disappear before you find out" (Kopelman, 2000: p. 612).

Psychogenic amnesia is most commonly diagnosed between ages 20 and 40 years (Coons and Milestein, 1992; Maldonado and Spiegel, 2008; Reinhold and Markowitsch, 2007), as opposed to transient global amnesia that tends to affect older individuals (*see* Transient Global Amnesia). In the sample of Coons and Milestein (1992), the age of patients with psychogenic amnesia ranged from 17 to 51 years. In the case series of Brand et al. (2009), the age of patients with dissociative amnesia ranged from 23 to 54 years. Several possible explanations for the clustering of cases of dissociative amnesia in the young adult segment were put forth (Staniliou and Markowitsch, 2012). Negative correlations were found between age and dissociation scale scores (Putnam, 1997). New research data suggested that mechanisms that capitalize on memory suppression become more taxing (demanding) with age (Anderson et al., 2011).

Furthermore, massive acute stress may have a differentiated effect on retrieval according to age (Pulopulos et al., 2013).

An underestimation of newly occurring, persisting, or recurring psychogenic amnesias in older age, due to a tendency in medical and psychiatric field to attribute cognitive symptoms in older patients to medical or more common psychiatric conditions cannot, however, be excluded (Kluft, 2007).

The sex ratio in dissociative amnesia was reported to be roughly one (Maldonado and Spiegel, 2008). Some studies, however, described a higher frequency of psychogenic amnesia in women (Coons and Milestein, 1992), and others in men (Kanzer, 1939; Kritchevsky et al., 2004; Brand et al., 2009). Some authors suggested that dissociative amnesia and other dissociative disorders may be underestimated in men (Spitzer and Freyberger, 2008). The mean educational level of patients with psychogenic amnesia reported by Coons and Milestein was 12.7 years; the years of education, however, ranged from 9 to 18 years (Coons and Milestein, 1992). These findings are consistent with results from other studies (Brand et al., 2009; Kritchevsky et al., 2004).

Psychogenic amnesias might recur. Coons and Milstein (1992) found that 40% of their sample had recurrent episodes of psychogenic amnesia. Abeles and Schilder (1935) reported a recurrent rate of 24%. Coons and Milstein (1992) attributed the elevated rate of recurrence in their sample to the features of their assessing setting (a tertiary care unit).

Risk Factors and Comorbidity

A previous history of psychiatric difficulties may be indicative of psychogenic or functional amnesia (Kritchevsky et al., 2004), although it was not unanimously reported (de Renzi et al., 1997). Coons and Milestein (1992) reported a high number of comorbidities with affective disorders, anxiety disorders, substance abuse, somatoform disorders, personality disorders, and other dissociative disorders among the 25 patients with psychogenic amnesia they had investigated in a tertiary care setting. As Spiegel et al. (2011) remarked, this high comorbidity rate might have reflected the severity and complexity of their clinical cases, who required referral to tertiary care clinic. Incidentally, this elevated comorbidity rate was not found in the case series described by Brand and colleagues (2009).

The Neuropsychology of Psychogenic Amnesia

On neuropsychological examination, patients with psychogenic amnesia often present with impairments of episodic-autobiographical memory, which are of retrograde nature (Markowitsch et al., 1997a). The episodic-autobiographical memory impairment can be restricted to a specific period of time or could span the entire life, or can be material specific. In comparison to the symptom of amnesia that may occur in posttraumatic stress disorder, the memory impairment in functional amnesia is usually more extensive, encompassing larger periods (Spiegel et al., 2011). With respect to semantic knowledge, autobiographical-semantic knowledge may initially be impaired (e.g., in psychogenic or dissociative fugue), but then it is relatively quickly relearned. In dissociative fugues, retrograde amnesia and personality loss are accompanied by sudden traveling from the customary environment (Markowitsch et al., 1997b). Albeit dissociative fugue currently is considered a separate diagnostic entity, several clinical, neuropsychological, and neuroimaging studies support its reclassification as a variant of dissociative amnesia (Spiegel et al., 2011; Staniloiu and Markowitsch, 2010; Markowitsch et al., 1997a,b; Hennig-Fast et al., 2008). Old general world knowledge is typically preserved in functional amnesia, but in some cases it may be affected in variable degrees (Brand et al., 2009; Kopelman, 2000; Kritchevsky et al., 2004). Priming is usually spared (Staniloiu and Markowitsch, 2012b). A limited number of studies of patients with functional amnesia provided objective evidence for impairments of previously learned procedural skills or for deficits in acquiring new procedural skills. Given that many patients quickly relearn their skills, their initial complaints might reflect a deficit in the 'procedural discourse' due to impairments in the semantic memory domain (Markowitsch and Staniloiu, 2013). In other cases, the 'loss' of skills may just represent difficulties with initiating or carrying out a task, either due to emotional-motivational factors or executive dysfunctions (Smith et al., 2010). Executive functions have been reported to be affected in a substantial number of patients with functional amnesia (Fujiwara et al., 2008; Brand et al., 2009).

Variable degrees of anterograde amnesia (which are detected by standard anterograde memory tests) can accompany retrograde forms of functional amnesia (Brand et al., 2009; Kopelman, 2000; Kritchevsky et al., 2004; Fujiwara et al., 2008). In patients with dense functional retrograde amnesia, the consciously acquired new memories for long-term storage are qualitatively different, in particular with respect to emotional flavoring and first-person autonoetic connection in comparison to memories of normal, matched participants. These qualitative changes may not be documented as they usually escape capturing on testing with standard anterograde memory tests. Functional anterograde amnesia in the absence of significant retrograde memory impairments or that is out of proportion in comparison to retrograde memory impairments can also occur, albeit it was much more rarely reported (Markowitsch et al., 1999a; Smith et al., 2010; Markowitsch and Staniloiu, 2013).

Several changes in personality dimensions were reported to have occurred after the onset of psychogenic or functional amnesia. These changes involved affectivity, perception (e.g., impaired self-face or face-emotion processing), cognition (e.g., social cognition), and behavior (e.g., changing in eating, smoking, drinking, or working habits) (for a review see Markowitsch and Staniloiu, 2013). Anomalous self-face processing was described in some cases (Pommerenke et al., 2012; Markowitsch and Staniloiu, 2013) and linked to a right-hemisphere dysfunction.

Pathogenetic Models of Psychogenic Amnesia

Two main models (that are not mutually exclusive) were espoused to account for the pathogeny of psychogenic

amnesia. Based on his clinical and neuroimaging findings, Markowitsch developed a model of functional amnesia that postulates that the memory impairment in the episodic-autobiographical domain in functional retrograde amnesia preponderantly reflects a stress hormone-mediated memory blockade ('mnestic block syndrome') (Markowitsch, 1998). This mnestic block syndrome is supposedly underlain by a right-hemispheric synchronization anomaly during retrieval attempts between the frontal lobe regions – important for autonoetic consciousness, search initiation and monitoring, and the temporoamygdalar system – important for emotional processing and appraising the self-relevance of social and biological stimuli (Markowitsch and Staniloiu, 2013). This memory impairment is opined to be triggered by adverse life conditions, usually of a recurrent nature and with onset in childhood or early adulthood; it is modulated by several factors, such as genes; personality characteristics; early parent–child attachment style; medical and psychiatric comorbidities; familial, ecological, and sociocultural environment; and epigenetic factors (Dalenberg et al., 2012). Changes in environment (especially these occurring in early life) were hypothesized to lead to genomic adaptations with consequences for stress hormonal response pattern, brain synaptic plasticity, and processing of incoming information (Markowitsch, 2003). Several key brain structures for autobiographical memory and emotional processing have been identified as being sensitive to the consequences of exposure to stressful experiences (such as amygdala and hippocampal formation, prefrontal cortex, and specific white matter tracts). Massive and prolonged stress might induce changes in regions with a high density of glucocorticoid receptors such as the anterior and medial temporal lobe. In animals, it was shown that stress enhances hippocampal long-term depression and blocks long-term potentiation. The effects of stress on the above-mentioned brain structures depend on their vulnerability, the magnitude of the neurotoxic (glucocorticoid) (*see* Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of) cascade, and the stage of development of respective structures (Lupien et al., 2009). The hypothesis that increased glucocorticoid levels play a significant role in the mnestic blockade has received empirical support from several studies (for a review, see de Kloet and Rinne, 2007). The right-hemispheric synchronization defect during retrieval attempts in functional or dissociative amnesia was suggested by a number of functional neuroimaging studies (Brand et al., 2009; Markowitsch et al., 1997b) (*see* Functional Brain Imaging of Language Processes).

The other model that dominates the understanding of the pathogeny of functional amnesia posits that the inability to retrieve personal events in psychogenic amnesia results from an increase in the activity of inhibitory regions of the prefrontal cortex coupled with a subsequent decrease in the activity of the hippocampus, similar to the one that occurs in motivated forgetting (Kopelman, 2000; Anderson and Green, 2001; Anderson et al., 2011). In line with this model, Fujiwara and Markowitsch (2003) argued that the executive control – or supervisory attentional system – is engaged in holding undesired or distressful memories out of self-awareness or autonoetic consciousness (*see* The Neural Basis of Consciousness). This may result in an overload of the executive system and may reduce the frontal cortex-associated cognitive reserve that is necessary for successful retrieval of other nontraumatic or nonstressful personal memories in functional amnesia. In agreement with this model are findings suggesting that executive functions may covary with successful retrieval of autobiographical memory (Brand et al., 2009). In the study of Brand et al. (2009), executive dysfunctions were found in four patients. These patients showed more extensive retrograde memory impairments than those with normal performance on executive function tests.

The Neuroimaging of Psychogenic or Functional Amnesia

Neuroanatomical, neuropathological, and neuroimaging studies of amnesic disorders had in the past promoted sharp distinctions between amnesias with evidence of brain damage ('organic') and amnesias without (overt) evidence of brain damage('functional'). Case reports and group studies of neurological patients delivered evidence of amnesia arising after (usually bilateral) damage of the regions in the limbic system (*see* Limbic System), partitioned into the medial temporal lobe (*see* Hippocampus and Related Structures; Temporal Lobe), the medial diencephalon (*see* Hypothalamus), and the basal forebrain. Since the advent of functional imaging (*see* Functional Brain Imaging of Language Processes) and wider utilization of structural imaging methods, such as diffusion tensor imaging, several studies have been providing evidence in functional, psychogenic, and dissociative amnesia of metabolic or microstructural changes of brain areas that are presently agreed upon to be crucially implicated in conscious mnemonic processing (*see* Consciousness, Neural Basis of); these findings have subsequently been challenging the perceived gap between 'functional' and 'organic' forms of amnesia (Staniloiu and Markowitsch, 2010).

Markowitsch et al. (1997b) investigated a 37-year-old man with persistent retrograde memory impairments after an episode of dissociative fugue. The patient remained unsure of his relationship to family members and changed a number of personality traits. While having been an avid car driver prior to the fugue, he became quite hesitant to enter a car thereafter, lost his asthma and allergies, gained a substantial amount of weight and changed his profession. Regional cerebral blood flow was estimated with water positron emission tomography (PET) during an autobiographical memory paradigm; it revealed a preponderantly left-hemispheric activation instead of the usual right-hemispheric frontotemporal one found with the same paradigm in normal participants. This result suggested that the patient indeed might have been unable to recollect his own past and instead processed this information as new and unrelated to his person (in a semantic, factual way, without the feeling of warmth and without the first-person autonoetic connection, which accompanies a recollecting experience). Related to this finding, another functional imaging technique (single-photon emission computed tomog

raphy) revealed in another patient with psychogenic amnesia a hypometabolic zone in exactly that frontotemporal junction area, which is necessary for normal retrieval of autobiographical memories (Markowitsch et al., 1997a). A similar hypometabolic zone in the same brain region was detected after organic brain damage and retrograde amnesia (Calabrese et al., 1996). These findings suggested that there may be a block or disconnection of memory-processing neural nets, leading to both 'psychogenic' and 'organic' amnesia (Markowitsch, 1999).

In a subsequent study in which glucose-PET data obtained at rest from 14 patients with dissociative amnesia and severe retrograde episodic-autobiographical memory impairments were analyzed, it was found that the right temporofrontal region was hypometabolic in a significant number of patients, with a significant reduction in the right inferolateral prefrontal cortex (Brand et al., 2009) (**Figure 1**). A reinforcement of these results came in the same year from the work of Tramoni et al. (2009), who in a patient with functional retrograde amnesia performed magnetization transfer ratio measurement and magnetic resonance spectroscopic imaging and found evidence of significant metabolic and subtle structural changes within the white matter of the right prefrontal region. These results lend support to previous proposals that the functional amnesic conditions might be underlain by a disconnection syndrome.

With respect to the neural correlates of the functional anterograde amnesia, neuroimaging data are sparse and inconclusive (Smith et al., 2010). Markowitsch et al. (2000) described a patient who developed both retrograde and anterograde amnesia after witnessing an open fire in his house. The retrograde amnesia spanned the last 6 years of his life. When investigated with glucose-PET, the patient displayed an overall reduction in his regional glucose metabolism in the cerebrum, with an even greater reduction in medial temporal and diencephalic regions. The hippocampal formation (*see* Hippocampus and Related Structures; Temporal Lobe) was found to be bilaterally hypometabolic. A metabolism of three standard deviations below that of normals was found in his right temporobasal, left temporomesial, left and right frontobasal, left insular, and dorsoparietal cortex. A metabolism of two standard deviations below normal was found in the left temporobasal, the right and left frontomesial, the right insular, the left cerebellum, and the left putamen. His reduced glucose metabolism corresponded to that of a patient of similar age and same sex, who after a major heart attack displayed the same type and extent of glucose reduction in his brain (Markowitsch et al., 1997b). The patient's amnesic condition persisted for several months but improved thereafter. The improvement in his conscious mnemonic abilities occurred in the context of combined psychotherapeutic and psychotropic treatment and was paralleled by a normalization of his glucose metabolism in the cerebrum.

The Differential Diagnosis of Psychogenic Amnesia

Similar to the transient global amnesia (*see* Transient Global Amnesia), functional amnesias can be triggered by a mixture of psychological and physical factors (Quinette et al., 2006). There are several important features that distinguish functional or psychogenic amnesia from transient global amnesia (cf **Table 1**).

The main challenge posed by the differential diagnosis of psychogenic (dissociative) amnesia is distinguishing between true and feigned amnesia (Barbarotto et al., 1996). In some cases, the preserved ability to retrieve information via other memory systems than the episodic-autobiographical memory system may lead clinicians to conclude that an individual is feigning amnesia. In other cases, there may be a mixture of true amnesia and feigning/malingering, making for a difficult diagnosis (Spiegel et al., 2011).

Treatment, Course, and Prognosis of Psychogenic Amnesia

Functional amnesias differ greatly with respect to recovery; variables that may independently predict prognosis are, however, largely unknown. There is currently a dearth of knowledge about evidence-based treatments for these conditions (Markowitsch and Staniloiu, 2013). Data on neuropsychological rehabilitation for psychogenic amnesia are sparse (Markowitsch and Staniloiu, 2012b). Furthermore, there is a stringent need for developing a broad theoretical framework for rehabilitation (Wilson, 2008) of psychogenic or functional amnesia and adopting a holistic approach that should aim toward achieving optimal 'physical, psychological, social, and vocational well-being.'

Conclusions

The possibility that both somatic and psychological factors make contributions to the clinical picture of various amnesic disorders had been emphasized by Mai's (1995) statement that "the presence of a neurological abnormality does not necessarily rule out psychogenic amnesia" (p. 105). After the legitimization of the transient global amnesia (Fisher and Adams, 1964) as a diagnostic neurological entity, several cases of psychogenic amnesia were recategorized as representing instantiations of transient global amnesia (de Renzi et al., 1997) (cf **Table 1**). Over half a century later, psychological factors are reacknowledged as being potentially involved of transient global amnesia (TGA) symptomatology (Quinette et al., 2006) (cf **Table 1**). Furthermore, functional neuroimaging data point to neural correlates of functional or psychogenic amnesias. These on the one hand challenge the deeply seated organic–psychogenic dichotomy. On the other hand, they open a pathway of hope for the development of effective treatments and comprehensive models of rehabilitation (Wilson, 2002) for these conditions and call for larger scale studies that use increased methodological rigor, combined neuroimaging methods, and rejuvenated research framework. As Markowitsch stated in 1999, "A common framework for psychiatry and neurosciences emphasized 115 [years] ago by Meynert who attempted to explain all psychic functions by anatomy and physiology […], is nevertheless a central aim for … the next century […]" (p. 335).

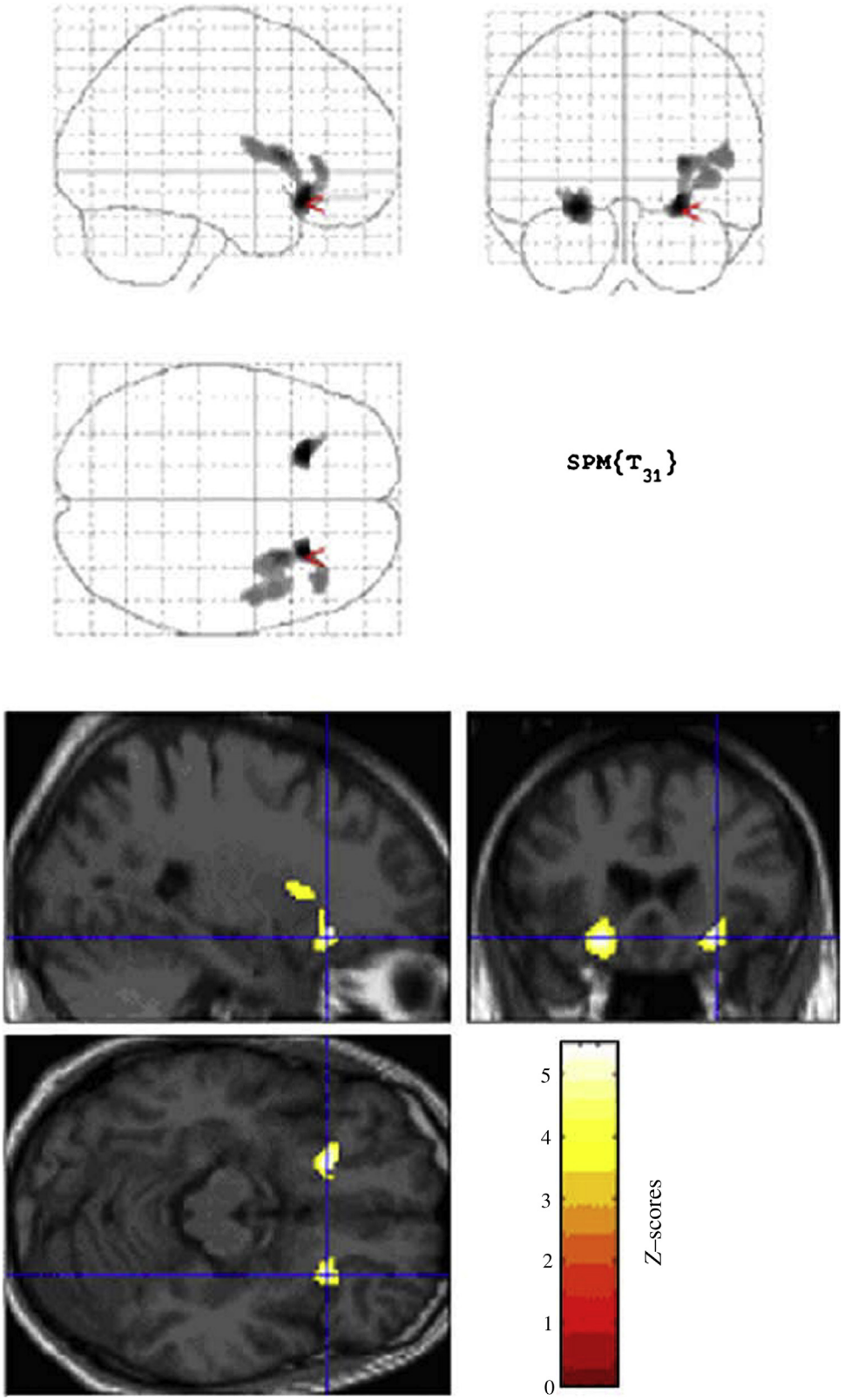

Figure 1 Relative decreases in regional cerebral glucose metabolism in 14 patients with dissociative amnesia relative to 19 control subjects (sagittal, frontal, and horizontal views as 'glass brains' and superimposed on magnetic resonance imaging (MRI) sections, MRI template). The cross indicates the locus of the only significant deactivated spot in the right inferolateral prefrontal cortex ($P_{corrected} < 0.001$, $x =$ "6 mm, $y = 24$ mm, $z = -14$ mm). The homologous hypometabolic region within the left inferolateral prefrontal cortex failed to reach significance ($P_{corrected} < 0.083$, $x = -22$ mm, $y = 24$ mm, $z = -14$ mm). Reproduced from Figure 1 of Brand, M., Eggers, C., Reinhold, N., Fujiwara, E., Kessler, J., Heiss, W.-D., Markowitsch, H.J., 2009. Functional brain imaging in fourteen patients with dissociative amnesia reveals right inferolateral prefrontal hypometabolism. Psychiatry Research: Neuroimaging Section 174, 32–39.

Table 1 General forms of amnesic states and their behavioral characteristics

Type	*Transient global amnesia*	*Functional (psychogenic) amnesia*
Typical age at the time of diagnosis	50–70 Years	20–40 Years
Precipitating factors	Strenuous physical activity or emotional stress or a combination of physical and somatic factors	Psychological stress or a combination of psychological and somatic factors
Occurrence	Suddenly, frequently after significant physical or psychic incidents	Suddenly, either after chronic stress or after a single major stressful incident
Duration	$\leq$24 h	Variable
Affected memory systems	Mainly episodic-autobiographical	Mainly episodic-autobiographical
Loss of personal identity	No	Frequent
Affected time range	Anterograde and variable retrograde	Retrograde and/or (more rarely) anterograde
Patient's behavior	Temporal disorientation, iterative questioning, concern about his/her condition	Often presenting with la belle indifference
Remission/recovery	Usually complete, but mild impairments in episodic memory may persist several months after the remission of the core amnesic syndrome	Sometimes complete, sometimes persistent
Brain changes	CA1 field of hippocampal formation	Prefrontal and limbic regions Fibers projecting to prefrontal cortex

See also: Amnesia: General; Retrieval.

Bibliography

Abeles, M., Schilder, P., 1935. Psychogenic loss of personal identity. Archives of Neurology and Psychiatry 34, 587–604.

Anderson, M.C., Green, C., 2001. Suppressing unwanted memories. Nature 410, 366–369.

Anderson, M.C., Reinholz, J., Kuhl, B.C., Mayr, U., 2011. Intentional suppression of unwanted memories grows more difficult as we age. Psychology and Aging 26 (2), 397–405.

Barbarotto, R., Laiacona, M., Cocchini, G., 1996. A case of simulated, psychogenic or focal pure retrograde amnesia: did an entire life become unconscious? Neuropsychologia 34, 575–585.

Brand, M., Markowitsch, H.J., 2009. Environmental influences on autobiographical memory: the mnestic block syndrome. In: Bäckman, L., Nyberg, L. (Eds.), Festschrift in Honor of Lars-Göran Nilsson. Psychology Press, Hove, UK, pp. 229–264.

Brand, M., Eggers, C., Reinhold, N., Fujiwara, E., Kessler, J., Heiss, W.-D., Markowitsch, H.J., 2009. Functional brain imaging in fourteen patients with dissociative amnesia reveals right inferolateral prefrontal hypometabolism. Psychiatry Research: Neuroimaging Section 174, 32–39.

Calabrese, P., Markowitsch, H.J., Durwen, H.F., Widlitzek, B., Haupts, M., Holinka, B., Gehlen, W., 1996. Right temporofrontal cortex as critical locus for the ecphory of old episodic memories. Journal of Neurology, Neurosurgery, and Psychiatry 61, 304–310.

Coons, P.M., Milstein, V., 1992. Psychogenic amnesia: a clinical investigation of 25 Cases. Dissociation 5, 73–79.

Dalenberg, C.J., Brand, B.L., Gleaves, D.H., Doray, M.J., Loewenstein, R., Cardeña, E., Frewen, P.A., Carlson, E.B., Spiegel, D., 2012. Evaluation of the evidence for the trauma and fantasy models of dissociation. Psychological Bulletin 138 (3), 550–588.

de Kloet, E.R., Rinne, T., 2007. Neuroendocrine markers of early trauma. In: Vermetten, E., Dorahy, M.J., Spiegel, D. (Eds.), Traumatic Dissociation. Neurobiology and Treatment. American Psychiatric Publ., Arlington, VA, pp. 139–156.

de Renzi, E., Lucchelli, F., Muggia, S., Spinnler, H., 1997. Is memory without anatomical damage tantamount to a psychogenic deficit? The case of pure retrograde amnesia. Neuropsychologia 35, 781–794.

Fast, K., Fujiwara, E., 2001. Isolated retrograde amnesia. Neurocase 7, 269–282.

Fisher, C.M., Adams, R.D., 1964. Transient global amnesia. Acta Neurologica Scandinavica 40 (Suppl. 9), 7–83.

Fujiwara, E., Markowitsch, H.J., 2003. Das mnestische Blockadesyndrom: Hirnphysiologische Korrelate von Angst und Stress. In: Schiepek, G. (Ed.), Neurobiologie der Psychotherapie. Schattauer Verlag, Stuttgart, pp. 186–212.

Fujiwara, E., Brand, M., Kracht, L., Kessler, J., Diebel, A., Netz, J., Markowitsch, H.J., 2008. Functional retrograde amnesia: a multiple case study. Cortex 44, 29–45.

Hennig-Fast, K., Meister, F., Frodl, T., Beraldi, A., Padberg, F., Engel, R.R., Reiser, M., Möller, H.J., Meindl, T., 2008. A case of persistent retrograde amnesia following a dissociative fugue: neuropsychological and neurofunctional underpinnings of loss of autobiographical memory and self-awareness. Neuropsychologia 46, 2993–3005.

ICD-10, 1992. Classification of Mental and Behavioral Disorders: Clinical Descriptions and Diagnostic Guidelines. World Health Organization, Geneva, Switzerland.

Janet, P., 1907. The Major Symptoms of Hysteria: Fifteen Lectures Given in the Medical School of Harvard University. Macmillan, New York.

Johnson, J.G., Cohen, P., Kasen, S., Brook, J.S., 2006. Dissociative disorders among adults in the community impaired functioning, and axis I and II comorbidity. Journal of Psychiatric Research 40, 131–140.

Kanzer, M., 1939. Amnesia, a statistical study. American Journal of Psychiatry 96, 711–716.

Kluft, R.P., 2007. The older female patient with a complex chronic dissociative disorder. Journal of Women and Aging 19, 119–137.

Kopelman, M.D., 2000. Focal retrograde amnesia and the attribution of causality: an exceptionally critical review. Cognitive Neuropsychology 17, 585–621.

Kritchevsky, M., Chang, J., Squire, L.R., 2004. Functional amnesia: clinical description and neuropsychological profile of 10 cases. Learning and Memory 11, 213–226.

Lupien, S.J., McEwen, B.S., Gunnar, M.R., Heim, C., 2009. Effects of stress throughout the lifespan on the brain, behaviour and cognition. Nature Reviews Neuroscience 10, 434–445.

Mai, F.M., 1995. Hysteria in clinical neurology. Canadian Journal of the Neurological Sciences 22, 101–110.

Maldonado, J.R., Spiegel, D., 2008. Dissociative disorders. In: Hales, R.E., Yudofsky, S.C., Gabbard, G.O. (Eds.), The American Psychiatric Publishing Textbook of Psychiatry, fifth ed. American Psychiatric Publ., Arlington, VA, pp. 665–710.

Markowitsch, H.J., 1996. Organic and psychogenic retrograde amnesia: two sides of the same coin? Neurocase 2, 357–371.

Markowitsch, H.J., 1997. Varieties of memory: systems, structures, mechanisms of disturbance. Neurology, Psychiatry and Brain Research 5, 37–56.

Markowitsch, H.J., 1998. The mnestic block syndrome: environmentally induced amnesia. Neurology, Psychiatry and Brain Research 6, 73–80.

Markowitsch, H.J., 1999. Neuroimaging and mechanisms of brain function in psychiatric disorders. Current Opinion in Psychiatry 12, 331–337.

Markowitsch, H.J., 2000. Functional amnesia: the mnestic block syndrome. Revue de Neuropsychologie 10, 175–198.

Markowitsch, H.J., 2003. Psychogenic amnesia. NeuroImage 20, S132–S138.

Markowitsch, H.J., Staniloiu, A., 2012a. Amnesic disorders. Lancet 380, 1229–1240.

Markowitsch, H.J., Staniloiu, A., 2012b. The rehabilitation of dissociative amnesia. Brain Impairment 13, 158–159.

Markowitsch, H.J., Staniloiu, A., 2013. The impairment of recollection in functional amnesic states. Cortex 49 (6), 1494–1510.

Markowitsch, H.J., von Cramon, D.Y., Schuri, U., 1993. Mnestic performance profile of a bilateral diencephalic infarct patient with preserved intelligence and severe amnesic disturbances. Journal of Clinical and Experimental Neuropsychology 15, 627–652.

Markowitsch, H.J., Calabrese, P., Fink, G.R., Durwen, H.F., Kessler, J., Härting, C., König, M., Mirzaian, E.B., Heiss, W.-D., Heuser, L., Gehlen, W., 1997a. Impaired episodic memory retrieval in a case of probable psychogenic amnesia. Psychiatry Research: Neuroimaging Section 74, 119–126.

Markowitsch, H.J., Fink, G.R., Thöne, A.I.M., Kessler, J., Heiss, W.-D., 1997b. Persistent psychogenic amnesia with a PET-proven organic basis. Cognition and Neuropsychiatry 2, 135–158.

Markowitsch, H.J., Kessler, J., Kalbe, E., Herholz, K., 1999a. Functional amnesia and memory consolidation. A case of persistent anterograde amnesia with rapid forgetting following whiplash injury. Neurocase 5, 189–200.

Markowitsch, H.J., Kessler, J., Russ, M.O., Frölich, L., Schneider, B., Maurer, K., 1999b. Mnestic block syndrome. Cortex 35, 219–230.

Markowitsch, H.J., Kessler, J., Weber-Luxenburger, G., Van der Ven, C., Heiss, W.-D., 2000. Neuroimaging and behavioral correlates of recovery from 'mnestic block syndrome' and other cognitive deteriorations. Neuropsychiatry, Neuropsychology, and Behavioral Neurology 13, 60–66.

McKay, G.C.M., Kopelman, M.D., 2009. Psychogenic amnesia: when memory complaints are medically unexplained. Advances in Psychiatric Treatment 15, 152–158.

Pommerenke, K., Staniloiu, A., Markowitsch, H.J., Eulitz, H., Dettmers, C., 2012. Ein Fall von retrograder Amnesie nach einer Resektion eines Medulloblastoms – psychogen, organisch, funktionell? [A case of retrograde amnesia after resection of a medulloblastom – psychogenic, organic, functional?]. Neurologie and Rehabilitation 18, 106–116.

Pulopulos, M.M., Almela, M., Hidalgo, V., Villada, C., Puig-Perez, S., Salvador, A., 2013. Acute stress does not impair long-term memory retrieval in older people. Neurobiology of Learning and Memory 104, 16–24.

Putnam, F., 1997. Dissociation in Children and Adolescents. A Developmental Perspective. Guilford Press, New York.

Quinette, P., Guillery-Girard, B., Dayan, J., de la Sayette, V., Marquis, S., Viader, F., Desgranges, B., Eustache, F., 2006. What does transient global amnesia really mean? Review of the literature and thorough study of 142 cases. Brain 129, 1640–1658.

Reinhold, N., Markowitsch, H.J., 2007. Emotion and consciousness in adolescent psychogenic amnesia. Journal of Neuropsychology 1, 53–64.

Sar, V., Koyuncu, A., Ozturk, E., Yargic, L.I., Kundakci, T., Yazici, A., et al., 2007. Dissociative disorders in the psychiatric emergency ward. General Hospital Psychiatry 29 (1), 45–50.

Smith, C.N., Frascino, J.C., Kripke, D.L., McHugh, P.R., Treisman, G.J., Squire, L.R., 2010. Losing memories overnight: a unique form of human amnesia. Neuropsychologia 38, 2833–2840.

Spiegel, D., 2006. Recognizing traumatic dissociation. American Journal of Psychiatry 163, 4.

Spiegel, D., Loewenstein, R.J., Lewis-Fernandet, R., Sar, V., Simeon, D., Vermetten, E., Cardena, E., Dell, P.F., 2011. Dissociative disorders in DSM-5. Depression and Anxiety 28, 824–852.

Spitzer, C., Freyberger, H.J., 2008. Geschlechtsunterschiede bei dissoziativen Störungen. Bundesgesundheitsblatt Gesundheitsforschung, Gesundheitsschutz 51, 46–52.

Staniloiu, A., Markowitsch, H.J., 2010a. Searching for the anatomy of dissociative amnesia. Journal of Psychology 218, 96–108.

Staniloiu, A., Markowitsch, H.J., 2010b. Understanding psychogenic amnesia and psychiatric disorders as causes of dementia. Journal of General Medicine 22, 41–49.

Staniloiu, A., Markowitsch, H.J., 2012. Towards solving the riddle of forgetting in functional amnesia: recent advances and current opinions. Frontiers in Psychology 3, 1–23. Article 403 (Open access).

Staniloiu, A., Markowitsch, H.J., 2012b. The remains of the day in dissociative amnesia. Brain Sciences 2, 101–129 (Open access).

Stone, J., 2009. Functional symptoms in neurology. Neurology in Practice 9, 179–189.

Tramoni, E., Aubert-Khalfa, S., Guye, M., Ranjeva, J.P., Felician, O., Ceccaldi, M., 2009. Hypo-retrieval and hyper-suppression mechanisms in functional amnesia. Neuropsychologia 47, 611–624.

Wilson, B., 2002. Toward a comprehensive model of cognitive rehabilitation. Neuropsychological Rehabilitation 12, 97–110.

Wilson, B.A., 2008. Neuropsychological rehabilitation. Annual Reviews in Clinical Psychology 4, 141–162.

Analysis of Variance and Generalized Linear Models

Ronald Christensen, University of New Mexico, Albuquerque, NM, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp. 473–480, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Analysis of variance (ANOVA) models apply to data that occur in groups. The fundamental ANOVA model is the one-way model that specifies a common mean value for the observations in a group. The ANOVA associated with the one-way model is presented. When the groups in a one-way ANOVA are identified as combinations of two or more factors, models incorporating factor main effects and factor interactions provide a useful device for exploring the underlying structure of the data. The main effects only two-way ANOVA model is discussed along with its interpretation and how it differs from the interaction model. Three-factor models and their interpretations are discussed. Appropriate models can be identified by fitting sequences of successively smaller models and using general testing procedures to identify models within each sequence that fit well. ANOVA models are also useful in generalized linear models. These are discussed with examples involving logistic regression. Again, appropriate models can be identified by fitting sequences of models.

Analysis of variance (ANOVA) models are models that exploit the grouping structure in a set of data and lend themselves to the examination of main effects and interactions. These models are often referred to as being hierarchical in that it makes no sense to test whether main effects can be dropped from a model that includes that factor in an interaction, and it makes no sense to test whether a lower-order interaction can be dropped from a model that includes any higher-order interaction involving all of the same factors. The modeling ideas also extend to generalized linear models (GLIMs).

One-way ANOVA

One-way ANOVA models identify groups of observations using a separate mean for each group. For example, the observations can be the age of suicide categorized by mutually exclusive ethnic groups. Suppose there are a groups with N_i observations in each group. The jth observation in the ith group is written y_{ij} with expected value $\mathrm{E}(y_{ij}) = \mu_i$. The one-way ANOVA model is

$$y_{ij} = \mu_i + \varepsilon_{ij} \qquad [1]$$

$i = 1, \ldots, a$, $j = 1, \ldots, N_i$ where the ε_{ij}s are unobservable random errors with mean 0, often assumed to be independent, normally distributed with variance σ^2, written $\varepsilon_{ij} \sim N(0, \sigma^2)$.

The ANOVA is a procedure for testing whether the groups have the same mean values μ_i. It involves obtaining two statistics. First, the mean squared error (*MSE*) is an estimate of the variance σ^2 of the individual observations. Second, the mean squared groups (*MSGrps*) is an estimate of σ^2 when the means μ_i are all the same and an estimate of σ^2 plus a positive number when the μ_i are different. The ratio of these two statistics, *MSGrps/MSE* should be approximately 1 if the μ_is are all the same and tends to be larger than 1 if the μ_is are not all the same. Under the normality assumptions for the ε_{ij}s, when the μ_is are all the same the exact value of *MSGrps/MSE* is random and follows an F distribution with $a-1$ degrees of freedom in the numerator and $n-a$ degrees of freedom in the denominator. Here $n = N_1 + \ldots + N_a$ is the total number of observations in the data. If the observed value of *MSGrps/MSE* is so much larger than 1 as to be relatively inconsistent with it coming from the F distribution, one concludes that the μ_is must not all be the same.

In the special balanced case where $N_i = N$ for all i, the computations are particularly intuitive as analyzing variances. To find the *MSE*, simply find the sample variance within each group, i.e., compute the sample variance from $y_{i1}, \ldots, y_{iN}$, then average these a numbers to get *MSE*. This provides an estimate of σ^2. To get *MSGrps*, first compute the sample mean for each group, say y_i. Then compute the sample variance of the y_is and multiply by N to get *MSGrps*. *MSGrps* estimates $\sigma^2 + Ns^2\mu$ where s^2_μ is a new parameter consisting of the sample variance of the unknown parameters $\mu_1, \ldots, \mu_a$. The μ_is are all equal if and only if $s^2_\mu = 0$. *MSGrps/MSE* estimates $1 + Ns^2_\mu/\sigma^2$.

Terminology varies considerably for these concepts. The *MSE* is also called the mean squared residual and the mean squared within (groups). *MSGrps* is also called the mean squared treatments (because often the groups are identified as different treatments in an experiment) and the mean squared between (groups).

An alternative but equivalent model used with one-way ANOVA is

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad [2]$$

$i = 1, \ldots, a$, $j = 1, \ldots, N_i$. Here μ is a grand mean and the α_is are differential effects for the groups. The problem with this model is that it is overparameterized, i.e., the μ and α_i parameters are not identifiable. Even if you know the μ_is, there is an infinite number of ways to define the μ and α_i parameters. In fact, you can arbitrarily pick a value for any one of the μ or α_i parameters and still make them agree with any set of μ_is. Without including extraneous side conditions that have nothing to do with the model, it is impossible to estimate any of the μ and α_i parameters. It is, however, possible to estimate some functions of them, like values $\mu + \alpha_i$, or contrasts among the α_is like $\alpha_1 - \alpha_2$. Linear functions of the μ and α_i parameters

for which linear unbiased estimates exist are called estimable functions. *see* Statistical Identification and Estimability.

The F test can also be viewed as testing the full one-way ANOVA model against the reduced model

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

The reduced model can be viewed as either dropping the subscript i from μ_i in model [1] or as dropping the α_is from model [2]. In either case, the reduced model does not allow for separate group effects. The F statistic comes from the error terms of the two models

$$F = \frac{SSE(\text{Red.}) - SSE(\text{Full})}{dfE(\text{Red.}) - dfE(\text{Full})} \Big/ MSE(\text{Full}) = \frac{MSGrps}{MSE}$$

Extensions of ANOVA to more general situations depend crucially on the idea of testing full models against reduced models; *see* Linear Hypothesis.

Two-way ANOVA

Two-way ANOVA can be thought of as a special case of one-way ANOVA in which the groups have a two-factor structure that we want to exploit in the analysis. For example, Everitt (1977) discusses 97 ten-year-old school children who were cross-classified by two factors, first, the risk of their home environment: not at risk (N) or at risk (R), and then the adversity of their school conditions: low, medium, or high. This defines six groups, each a combination of a home environment and a school condition. To illustrate the modeling concepts, suppose the dependent variable y is the score on a test of verbal abilities. In general, we would write a model

$$y_{gk} = \mu_g + \varepsilon_{gk} \qquad [3]$$

where $g = 1, \ldots, G$ indicates the different groups and $k = 1, \ldots, N_g$ indicates the observations within the group. In the specific example, $G = 6$ and $(N_1, N_2, N_3, N_4, N_5, N_6) = (17, 8, 18, 42, 6, 6)$.

To use the two-factor structure, we begin by identifying the factors numerically. Let $i = 1, 2$ indicate home environment: not at risk ($i = 1$), at risk ($i = 2$). Let $j = 1, 2, 3$ identify school adversity: low ($j = 1$), medium ($j = 2$), high ($j = 3$). In general, let $i = 1, \ldots, a$ denote the levels of the first factor and $j = 1, \ldots, b$ denote levels of the second factor. Without changing anything of substance in the model, we can rewrite the ANOVA model [3] as

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \qquad [4]$$

$i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, N_{ij}$. All we have done is replace the single index for the groups g with two numbers ij that are used to identify the groups. This is still a one-way ANOVA model and can be used to generate an F test of whether all the μ_{ij}s are equal. In our example, $a = 2$, $b = 3$, and $(N_{11}, N_{21}, N_{12}, N_{22}, N_{13}, N_{23}) = (17, 8, 18, 42, 6, 6)$. Note that with a two-factor model the number of groups is $G = ab$ and $E(y_{ijk}) = \mu_{ij}$.

The one-way ANOVA model [4] is often rewritten in an overparameterized version that includes numerous unidentifiable parameters. The overparameterized model is called the two-way ANOVA with interaction model, and is written

$$y_{ijk} = \mu + \alpha_i + \eta_j + (\alpha\eta)_{ij} + \varepsilon_{ijk} \qquad [5]$$

Here μ is a grand mean, the α_is are differential effects for the first factor, the η_js are differential effects for the second factor, and the $(\alpha\eta)_{ij}$s are called interaction effects. In reality, all of the μ, α_i, and η_j parameters are extraneous. If we drop all of them, we get the model $y_{ijk} = (\alpha\eta)_{ij} + \varepsilon_{ijk}$, which is just the one-way ANOVA model [4] with the μ_{ij}s relabeled as $(\alpha\eta)_{ij}$s. In model [5], without the introduction of side conditions that have nothing to do with the model, it is impossible to estimate or conduct any test on any function of the parameters that does not involve the $(\alpha\eta)_{ij}$ parameters. Since main effect parameters are extraneous when interactions are in the model, computer programs that purport to give tests for main effects after fitting interactions are really testing some arcane function of the interaction parameters that is determined by a choice of side conditions used by the program.

The interesting aspect of the two-way ANOVA with interaction model is that it suggests looking at the two-way ANOVA without interaction model

$$y_{ijk} = \mu + \alpha_i + \eta_j + \varepsilon_{ijk} \qquad [6]$$

This model is not equivalent to the one-way ANOVA model, but it includes nontrivial group effects. It amounts to imposing a restriction on the μ_{ij}s that

$$\mu_{ij} = \mu + \alpha_i + \eta_j \qquad [7]$$

for some μ, α_is, and η_js. This indicates that the group effects μ_{ij} have a special structure in which the group effect is the sum of a grand mean and differential effects for each factor, hence model [6] is referred to as an additive model. The two-way ANOVA without interaction model is still overparameterized, so none of the individual parameters are estimable, but typically contrasts in the α_is and η_js are estimable as well as values $\mu + \alpha_i + \eta_j$.

To illustrate the modeling concepts, suppose that in our example the group means for the verbal test scores take the values shown in **Table 1**. These μ_{ij}s satisfy the additive model [7] so they do not display interaction. For example, take $\mu = 0$, $\alpha_1 = 110$, $\alpha_2 = 100$, $\eta_1 = 0$, $\eta_2 = -5$, $\eta_3 = -15$. Note that μ, the α_is, and the η_js are not identifiable (estimable) because there is more than one way to define them that is consistent with the μ_{ij}s. For example, we can alternatively take $\mu = 100$, $\alpha_1 = 5$, $\alpha_2 = -5$, $\eta_1 = 5$, $\eta_2 = 0$, $\eta_3 = -10$. However, identifiable functions of the parameters include $\mu + \alpha_i + \eta_j$, $\alpha_1 - \alpha_2$, $\eta_1 - \eta_3$, and $\eta_1 + \eta_2 - 2\eta_3$. Identifiable functions take on the same values for any valid choices of μ, the α_is, and the η_js. For example, the effect of changing from a not at risk home situation to an at risk home situation is a decrease of $\alpha_1 - \alpha_2 = 10$ points in mean verbal test score. Similarly, the effect of changing from a low school adversity situation to a high school

Table 1 Mean verbal test scores, additive model

		j (School)		
	μ_{ij}	*1 (Low)*	*2 (Medium)*	*3 (High)*
i (Home)	1 (*N*)	110	105	95
	2 (*R*)	100	95	85

adversity situation is a decrease of $\eta_1 - \eta_2 = 15$ points in mean verbal test score.

The beauty of the additive model [6] is that there is one number that describes the mean difference between students having not at risk home status and students having at risk home status. The difference is a drop of 10 points *regardless* of the school adversity status. This number can legitimately be described as the effect of going from not at risk to at risk. (Note that unless the students were assigned randomly to their home conditions, this does not imply that changing a student's status from at risk to not at risk will *cause*, on average, a 10 point gain.) Similarly, the effect of school adversity does not change with the home status. For example, going from low to high school adversity induces a 15 point drop regardless of whether the students are at risk or not at risk. Statistically, tests for main effects are tests of whether any of the α_is are different from each other and whether any of the η_js are different from each other, in other words whether the home statuses are actually associated with different mean verbal test scores and similarly for the school adversities.

The existence of interaction is simply any structure to the μ_{ij}s that cannot be written as $\eta_{ij} = \mu + \alpha_i + \eta_j$ for some μ, α_is, and η_js. For example, consider **Table 2**. In this case, the relative effect of having at risk home status for low or medium adversity schools is a drop of 10 points in mean test score, however for highly adverse schools, the effect of an at risk home status is a drop of 15 points. Unlike cases where the additive model [7] holds, the effect of at risk home status depends on the level of school adversity, so there is no one number than can characterize the effect of at risk home status. It makes no sense to consider the effect of home conditions without specifying the school adversity. Similarly, the effects of school adversity change depending on home status. For example, going from low to high school adversity induces a 15 point drop for students whose homes are not at risk, but a 20 point drop for at risk students. Again, there is no one number than can characterize the change from low to high school adversity, so there is no point in considering this change without specifying the home risk status. One moral of this discussion is that when interaction exists, there is no point in looking at tests of main effects, because main effects are essentially meaningless. If there is no one number that describes the effect of changing from not at risk to at risk, what could $\alpha_1 - \alpha_2$ possibly mean?

In practice, analyses such as these are conducted on estimates of the μ_{ij}s, and the estimates are subject to variability.

A primary use of model [6] is to test whether this special group effects structure fits the data. Model [6] is used as a reduced model and is tested against the full model [5] that includes interaction. This test is referred to as a test of interaction. In particular, the interaction test is not a test of whether the $(\alpha\eta)_{ij}$s are all zero, it is a test of whether every possible definition of the $(\alpha\eta)_{ij}$s must be consistent with the two-way without interaction model. The easiest way to think about this correctly is to think about testing whether one can simply drop the $(\alpha\eta)_{ij}$s from model [5].

Table 2 Mean verbal test scores, interaction model

		j (School)		
	μ_{ij}	*1 (Low)*	*2 (Medium)*	*3 (High)*
i (Home)	1 (*N*)	110	105	95
	2 (*R*)	100	95	80

If the two-way without interaction model [6] fits the data, it is interesting to see if any special case (reduced model) also fits the data. Two obvious choices are fitting a model that drops the second factor (school) effects

$$y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk} \qquad [8]$$

and a model that drops the first factor (home) effects

$$y_{ijk} = \mu + \eta_j + \varepsilon_{ijk} \qquad [9]$$

If model [8] fits the data, there is no evidence that the second factor helps explain the data over and above what the first factor explains. Similarly, if [9] fits the data, there is no evidence that the first factor helps explain the data over and above what the second factor explains.

Given model [8] with only the first factor effect, we can evaluate whether the data fit a reduced model without that effect,

$$y_{ijk} = \mu + \varepsilon_{ijk} \qquad [10]$$

This model comparison provides a test of whether the first factor is important in explaining the data when the second factor is ignored. Similarly we can start with model [9] and compare it to model [10]. If model [10] fits, neither factor helps explain the data.

Through fitting a series of models, we can test whether there is evidence for the interaction effects, test whether there is evidence for the second factor effects η_j when the first factor α_i effects are included in the model, test whether there is evidence for the second factor effects η_j when the first factor α_i effects are not included in the model, and perform two similar tests for the importance of the α_i effects. In the special case when the N_{ij}s are all the same, the test for α effects including ηs, that is, the test of model [6] versus model [9], and the test for α effects ignoring ηs, that is, the test of model [8] versus model [10], are identical. Similarly, the two tests for η effects are identical. While this identity greatly simplifies the analysis, it only occurs in special cases and is not generally applicable. Moreover, it does not extend to GLIMs such as log-linear models and logistic regression. The appropriate way to think about these issues is in terms of model comparisons.

Higher-Order ANOVA

The groups in a one-way ANOVA can also result from combining the levels of three or more factors. Consider a three-way cross-classification of Everitt's 97 students by further classifying them into students displaying or not displaying deviant classroom behavior. This determines $G = 12$ groups. The one-way ANOVA model $y_{gm} = \mu_g + \varepsilon_{gm}$, $g = 1, \ldots, G$, $m = 1, \ldots, N_g$ can be rewritten for three factors as $y_{ijkm} = \mu_{ijk} + \varepsilon_{ijkm}$ where the three factors are identified by $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, c$ so that $G = abc$, the group sample sizes are N_{ijk}, and $E(y_{ijkm}) = \mu_{ijk}$. In the example, we continue to use i and j to indicate home and school conditions, respectively, and now use $k = 1, 2$ to indicate classroom behavior with $k = 1$ being nondeviant.

In its most overparameterized form, a three-way ANOVA model is

$$y_{ijkm} = \mu + \alpha_i + \eta_j + \gamma_k + (\alpha\eta)_{ij} + (\alpha\gamma)_{ik} + (\eta\gamma)_{jk} + (\alpha\eta\gamma)_{ijk} + \varepsilon_{ijkm} \quad [11]$$

This includes a grand mean μ, main effects for each factor α_is, η_js, γ_ks, two-factor interactions $(\alpha\eta)_{ij}$s, $(\alpha\gamma)_{ik}$s, $(\eta\gamma)_{jk}$s, and a three-factor interaction $(\alpha\eta\gamma)_{ijk}$s. In model [11], dropping the redundant terms, that is, everything except the three-factor interaction terms, gives a one-way ANOVA model $y_{ijkm} = (\alpha\eta\gamma)_{ijk} + \varepsilon_{ijkm}$. The primary interest in model [11] is that it suggests a wealth of reduced models to consider.

The first order of business is to test whether the three factor interaction is necessary, i.e., test the full model [11] against the reduced model without the three-factor interaction

$$y_{ijkm} = \mu + \alpha_i + \eta_j + \gamma_k + (\alpha\eta)_{ij} + (\alpha\gamma)_{ik} + (\eta\gamma)_{jk} + \varepsilon_{ijkm} \quad [12]$$

Focusing on only the important parameters, model [12] is equivalent to $y_{ijkm} = (\alpha\eta)_{ij} + (\alpha\gamma)_{ik} + (\eta\gamma)_{jk} + \varepsilon_{ijkm}$. In fact, even this version of the model is overparameterized. Model [11] is equivalent to the one-way ANOVA model, so if the three-factor interaction test is significant, there is no simplifying structure to the treatments. Probably the best one can do is to think of the problem as a one-way ANOVA and draw whatever conclusions are possible from the groups.

If model [12] fits the data, we can seek simpler models or try to interpret model [12]. Conditioning on the levels of one factor, simplifies interpretations. For example, if $i = 1$, $y_{1jkm} = \mu + \alpha_1 + \eta_j + \gamma_k + (\alpha\eta)_{1j} + (\alpha\gamma)_{1k} + (\eta\gamma)_{jk} + \varepsilon_{1jkm}$ or $y_{1jkm} = [\mu + \alpha_1] + [\eta_j + (\alpha\eta)_{1j}] + [\gamma_k + (\alpha\gamma)_{1k}] + (\eta\gamma)_{jk} + \varepsilon_{1jkm}$. This is simply a two-factor model with interaction. If we change to $i = 2$, we again get a two-factor interaction model, one that has different main effects, but has the same interaction terms as when $i = 1$. For example, in Table 3 μ_{ijk}s satisfy model [12], that is, $\mu_{jkm} = \mu + \alpha_i + \eta_j + \gamma_k + (\alpha\eta)_{ij} + (\alpha\gamma)_{ik} + (\eta\gamma)_{jk}$ for some definitions of the parameters.

The key point in Table 3 is that, regardless of school adversity, the relative effect of not being at risk is always one point higher for nondeviants than for deviants. Thus, for high adversity the nondeviant home difference is 101–99 which is one point higher than the deviant home difference 99–98.

If model [12] is adequate, the next step is to identify which of the two factor interaction terms are important. For example, we can drop out the $(\eta\gamma)_{jk}$ term to get

$$y_{ijkm} = \mu + \alpha_i + \eta_j + \gamma_k + (\alpha\eta)_{ij} + (\alpha\gamma)_{ik} + \varepsilon_{ijkm} \quad [13]$$

Dropping unimportant parameters, this is equivalent to $y_{ijkm} = (\alpha\eta)_{ij} + (\alpha\gamma)_{ik} + \varepsilon_{ijkm}$. To interpret model [13], condition on the factor that exists in both interaction terms. In our example, this is home conditions. The resulting model for $i = 1$ is a no interaction model $y_{1jkm} = [\mu + \alpha_1] + [\eta_j + (\alpha\eta)_{1j}] + [\gamma_k + (\alpha\gamma)_{1k}] + \varepsilon_{1jkm}$ with a similar no interaction model for $i = 2$ but with different main effects. For example, suppose the m_{ij}s are given in Table 4. Rearranging Table 4 to group together categories with i fixed gives Table 5. For fixed home conditions, there is no interaction. For not at risk students, deviant behavior is associated with a one point drop and high school adversity is associated with a one point drop. For at risk students, deviant behavior is associated with a two point drop and, relative to low school adversity, medium and high adversities are associated with a one point and a three point drop, respectively.

Alternatively, we could eliminate both the $(\alpha\gamma)_{ik}$s and $(\eta\gamma)_{jk}$s from model (12) to get

$$y_{ijkm} = \mu + \alpha_i + \eta_j + \gamma_k + (\alpha\eta)_{ij} + \varepsilon_{ijkm} \quad [14]$$

which is equivalent to $y_{ijkm} = \gamma_k + (\alpha\eta)_{ij} + \varepsilon_{ijkm}$. We can think of this as a model for no interaction in a two-factor analysis in which one factor is indicated by the pair ij and the other factor is indicated by k. In the example, this means there is a main effect for classroom behavior plus an effect for each of the six combinations of home and school conditions. In particular, consider Table 6. Here, deviant behavior is always associated with a one point drop but there is interaction between home and school. For low school adversity, there is no effect of home conditions but for medium or high adversity, being at risk is associated with a one point drop.

From model (14L50), we could then fit a model that involves only the main effects

$$y_{ijkm} = \mu + \alpha_i + \eta_j + \gamma_k + \varepsilon_{ijkm} \quad [15]$$

Model [15] is equivalent to $y_{ijkm} = \alpha_i + \eta_j + \gamma_k + \varepsilon_{ijkm}$. If model [15] is adequate, we could consider models that successively drop out the α_is, the η_js, and the γ_ks.

Fitting this sequence of successively smaller models, [11] then [12] then [13] then [14] then [15], then three models

Table 3 No three-factor interaction

		Adversity of school (j)					
		Low		*Medium*		*High*	
μ_{ijk}	*Home (i)*	*N*	*R*	*N*	*R*	*N*	*R*
Classroom behavior (*k*)	Nondeviant	101	100	101	99	101	99
	Deviant	100	100	100	99	99	98

Table 4 Two two-factor interaction

		Adversity of school (j)					
		Low		*Medium*		*High*	
μ_{ijk}	*Home (i)*	*N*	*R*	*N*	*R*	*N*	*R*
Classroom behavior (*k*)	Nondeviant	101	100	101	99	101	97
	Deviant	100	98	100	97	99	95

Table 5 Rearrangement of **Table 4**

		Home (i)					
		N			*R*		
μ_{ijk}	*School (j)*	*L*	*M*	*H*	*L*	*M*	*H*
Classroom behavior (*k*)	Nondeviant	100	101	100	100	99	97
	Deviant	100	100	99	98	97	95

Table 6 One two-factor interaction

		Adversity of school (j)					
		Low		Medium		High	
μ_{ijk}	Home (i)	N	R	N	R	N	R
Classroom behavior (k)	Nondeviant	101	101	101	100	101	100
	Deviant	100	100	100	99	100	99

with the main effects successively dropped out, provides one tool for analyzing the data. There are 36 different orders possible for sequentially dropping out the two-factor interactions and then the main effects. In the balanced case of $N_{ijk} = N$ for all i, j, k, the order of dropping the effects does not matter, so one can construct an ANOVA table to examine each of the individual sets of effects, that is, main effects, two-factor interactions, three-factor interaction. For unbalanced cases, one would need 36 different ANOVA tables, one for each sequence, so ANOVA tables are almost never examined except in balanced cases. (Two-factor models only generate two distinct ANOVA tables.) For unbalanced cases, instead of looking at the ANOVA tables, it is more convenient to simply report the *SSE* and *dfE* for all of the relevant models using a notation that identifies models by only their important parameters. For four or more factors, the number of potentially interesting models becomes too large to evaluate all of them.

After deciding on one or more appropriate models, the μ_{ijk}s can be estimated subject to the model and the models interpreted subject to the variability of the estimates.

To interpret a three-factor interaction, recall that just as a two-factor interaction exists when the effect of one factor changes depending on the level of the other factor, one can think of a three-factor interaction as a two-factor interaction that changes depending on the level of the third factor. However, it may be more productive to think of three-factor interaction in a three-factor model as simply specifying a one-way ANOVA. As with two-factor models, it makes little sense to test the main effects of a factor when that factor is involved in an important interaction. Similarly, it makes little sense to test that a lower-order interaction can be dropped from a model that includes a higher-order interaction involving all of the same factors.

A commonly used generalization of the ANOVA models is to allow some of the parameters to be unobservable random variables (random effects) rather than fixed unknown parameters. *see* Hierarchical Models: Random and Fixed Effects.

Generalized Linear Models

ANOVA is best suited for analyzing normally distributed data. These are measurement data for which the random observations are symmetrically distributed about their mean values. GLIMs use similar linear structures to analyze other kinds of data, such as count data and time to event data.

We can rethink two-way ANOVA models as independent observations y_{ijk} normally distributed with mean m_{ij} and variance σ^2, write $y_{ijkm} \sim N(m_{ij}, \sigma^2)$. The interaction model [5] has

$$m_{ij} = \mu + \alpha_i + \eta_j + (\alpha\eta)_{ij}$$

The no interaction model [6] has

$$m_{ij} = \mu + \alpha_i + \eta_j$$

Now consider data that are counts y_{ij} of some event. Assume the counts are independent Poisson random variables with means $E(y_{ij}) = m_{ij}$. The use of two subscripts indicates that the data are classified by two factors. For Poisson data the m_{ij}s must be nonnegative, so $\log(m_{ij})$ can be both positive and negative. Linear models naturally allow both positive and negative mean values, so it is natural to model the $\log(m_{ij})$s with linear models. For example, we might use the log-linear model

$$\log(m_{ij}) = \mu + \alpha_i + \eta_j + (\alpha\eta)_{ij}$$

or, without interaction,

$$\log(m_{ij}) = \mu + \alpha_i + \eta_j$$

These log-linear models are also appropriate for multinomial data and independent groups of multinomial data.

Similarly, if the data y_{ij} are independent binomials with N_{ij} trials and probability of success p_{ij}, then we can analyze the proportions $\widehat{p}_{ij} \equiv y_{ij}/N_{ij}$ with $E(\widehat{p}_{ij}) \equiv m_{ij}/P_{ij}$. Probabilities are defined to be between 0 and 1, the odds $p_{ij}/(1 - p_{ij})$ take positive values, and the log-odds can take any positive or negative value. It is natural to write linear models for the log-odds such as

$$\log[p_{ij}/(1 - p_{ij})] = \mu + \alpha_i + \eta_j + (\alpha\eta)_{ij}$$

or, without interaction,

$$\log[p_{ij}/(1 - p_{ij})] = \mu + \alpha_i + \eta_j$$

More generally, in GLIMs we create linear models for $g(m_{ij})$ using any strictly monotone function $g(\cdot)$, called a link function. The linear structures used for one-way ANOVA, two-way ANOVA, and higher-order ANOVA models all apply to GLIMs, and appropriate models can be found by comparing the fit of full and reduced models similar to methods for unbalanced ANOVA. In particular, cross-classified tables of counts often involve many factors.

We now examine in more depth ANOVA-type models for binomial data. Reconsider Everitt's 97 children cross-classified by the risk of their home environment and adversity of their school conditions. Rather than studying their verbal test score performance conditional on their classroom behavior as we did when discussing three-factor ANOVA we now examine models for whether the children display deviant or nondeviant behavior. In particular, we model the probability of a student falling into the deviant behavior category given their membership in the six home–school groups, write the probability of deviant behavior when in the ij group as p_{ij} and model the log-odds with ANOVA type models,

$$\log[p_{ij}/(1 - p_{ij})] = \mu_{ij}$$

The overparameterized model

$$\log[p_{ij}/(1-p_{ij})] = \mu + \alpha_i + \eta_j + (\alpha\eta)_{ij}$$

is equivalent to the original model. Neither of these models really accomplishes anything because they fit a separate parameter to every home–school category, so the models place no restrictions on the observations. The additive model

$$\log[p_{ij}/(1-p_{ij})] = \mu + \alpha_i + \eta_j$$

constitutes a real restriction on the parameters.

Suppose the log-odds, the μ_{ij}s, satisfy Table 7. These have the structure of an additive model. In this case the log-odds for deviant behavior are two larger for at risk homes than for not at risk homes, e.g., $2 = 0 - (-2) = 1 - (-1) = 3 - 1$. This means that the odds are $e^2 = 7.4$ times larger for at risk homes. In particular, the log-odds of deviant behavior for low adversity schools and not at risk homes is -2 so the odds are $O_{11} = e^{-2} = 0.135$ and the probability if $p_{11} = O_{11}/[1 + O_{11}] = 0.12$. Similarly, the odds and probability of deviant behavior for low adversity schools and at risk homes are $O_{21} = e^0 = 1$ and $p_{21} = 0.5$. The change in odds is $0.135 \times 7.4 = 1$. The 7.4 fold increase in odds is the same for all school conditions. Similarly, comparing low to high school adversities, the difference in log-odds is $3 = 1 - (-2) = 3 - 0$, so the odds of deviant behavior are $e^3 = 20$ times greater in highly adverse schools, regardless of the home situation. The odds of deviant behavior in highly adverse schools and not at risk homes is $O_{13} = e^1 = 2.7$ which is 20 times greater than $O_{11} = 0.135$, the odds for low adversity schools and not at risk homes.

An example of a nonadditive model has log-odds as shown in Table 8. For low adversity schools the log-odds only differ by 1, for medium adversity schools they differ by 2, and for high adversity schools they differ by 3. Thus the effect of home conditions on the log-odds depends on the level of school adversity.

Table 7 Log-odds: additive model

		j (School)		
	μ_{ij}	*1 (Low)*	*2 (Medium)*	*3 (High)*
i (Home)	1 (*N*)	−2	−1	1
	2 (*R*)	0	1	3

Table 8 Log-odds: nonadditive model

		j (School)		
	μ_{ij}	*1 (Low)*	*2 (Medium)*	*3 (High)*
i (Home)	1 (*N*)	−2	−1	1
	2 (*R*)	−1	1	4

Conclusions

The fundamental ANOVA model is the one-way model. It specifies different mean values for different groups. When the groups are identified as combinations of two or more factors, models incorporating main effects and interactions become a useful device for examining the underlying structure of the data. Appropriate models can be identified by fitting sequences of successively smaller models and using general testing procedures to identify the models within each sequence that fit well. Having identified a model, it is important to interpret what that model suggests about the underlying data structure. ANOVA models, the sequential model fitting procedures, and the interpretations apply to both balanced and unbalanced data and to GLIMs. The sequential model fitting simplifies in balanced ANOVA allowing for it to be summarized in an ANOVA table.

See also: Multivariate Analysis: Discrete Variables (Overview); Simultaneous Equation Estimates (Exact and Approximate), Distribution of.

Bibliography

Agresti, A., 1990. Categorical Data Analysis. Wiley, New York.
Anderson, Marti J., 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 26 (1), 32–46.
Christensen, R., 1996a. Plane Answers to Complex Questions: The Theory of Linear Models, second ed. Springer-Verlag, New York.
Christensen, R., 1996b. Analysis of Variance, Design, and Regression: Applied Statistical Methods. Chapman and Hall, London.
Christensen, R., 1997. Log-Linear Models and Logistic Regression, second ed. Springer-Verlag, New York.
Everitt, B.S., 1977. The Analysis of Contingency Tables. Chapman and Hall, London.
Fienberg, S.E., 1980. The Analysis of Cross-classified Categorical Data, second ed. MIT Press, Cambridge, MA.
Graybill, F.A., 1976. Theory and Application of the Linear Model. Duxbury Press, North Scituate, MA.
Hocking, Ronald R., 2013. Methods and Applications of Linear Models: Regression and the Analysis of Variance. John Wiley & Sons, Hoboken, NJ.
Hosmer, D.W., Lemeshow, S., 1989. Applied Logistic Regression. Wiley, New York.
Johnson, Richard Arnold, Wichern, Dean W., 2007. Applied Multivariate Statistical Analysis. Prentice Hall, Upper Saddle River, NJ.
Lee, Y., Nelder, J.A., 1996. Hierarchical generalized linear models, with discussion. Journal of the Royal Statistical Society, Series B 58, 619–656.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman and Hall, London.
McCulloch, Charles E., Searle, Shayle R., Neuhaus, John M., 2008. Generalized, Linear, and Mixed Models. Wiley, Hoboken, NJ.
Nelder, J.A., 1977. A reformation of linear models. Journal of the Royal Statistical Society, Series A 140, 48–63.
Scheff, H., 1959. The Analysis of Variance. Wiley, New York.
Searle, S.R., 1971. Linear Models. Wiley, New York.
Seber, G.A.F., 1977. Linear Regression Analysis. Wiley, New York.

Analytical Marxism

Nicholas Vrousalis, University of Cambridge, Cambridge, UK; and KU Leuven, Leuven, Belgium
Philippe Van Parijs, UC Louvain, Louvain-la-Neuve, Belgium

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by P. van Parijs, volume 1, pp. 484–486, © 2001, Elsevier Ltd.

Abstract

This article draws together and briefly summarizes the most important questions raised by and discussed in the broad literature on analytical Marxism since its founding in the early 1980s.

Analytical Marxism is a cross-disciplinary school of thought that attempts to creatively combine a keen interest in some of the central themes of the Marxist tradition and the resolute use of analytical tools more commonly associated with 'bourgeois' social science and philosophy.

Marx is often interpreted as having used a 'dialectical' mode of thought that he took over, with modifications, from Hegel and that can be contrasted with conventional, 'analytical' thinking. Marx himself may have said and thought so. But according to analytical Marxists, one can make sense of his work, or at least of the main or best part of it, while complying with the strictest standards of analytical thought. This need not mean that all of what is usually called dialectics should be abruptly dismissed. For example, the study of the typically dialectical contradictions that drive historical change constitutes a challenging area for the subtle use of analytical thought (see Elster, 1978).

The resolute option for an analytical mode of thought does, however, imply that the research program stemming from Marx should by no means be conceived as the development of an alternative 'logic,' or of a fundamentally different way of thinking about capitalist development, or about social reality in general. It rather consists in practicing the most appropriate forms of standard analytical thought – using conventional conceptual analysis, formal logic and mathematics, econometric methods, and the other tools of statistical and historical research – to tackle the broad range of positive and normative issues broached in Marx's work. Analytical tools may have been developed and extensively used by 'bourgeois' social science and philosophy. This does not make them unfit, analytical Marxists believe, to rigorously rephrase and fruitfully develop some of the central tenets of the Marxian tradition.

Thus, the techniques characteristic of Anglo-American analytical philosophy can be used to clarify the meaning of key Marxian concepts as well as the epistemic status of the central propositions of the Marxian corpus and their logical relations with each other (see, e.g., Cohen, 1978). Formal models resting on the assumption of individually rational behavior, as instantiated by neoclassical economic theory and the theory of strategic games, can be used to understand the economic and political dynamics of capitalist societies (see, e.g., Przeworski, 1985; Carling, 1991). And the careful checking of theoretical conjectures against carefully collected and interpreted historical data can be used to test Marx's grand claims about transitions from one mode of production to another (see, e.g., Brenner et al. 1984). Making use of these tools does not amount to taking on board all the statements they have allegedly helped establish in the hands of conservative philosophers and social scientists. Nor is it meant to serve dogmatic defense of any particular claim Marx may have made. Rather, analytical Marxists view a competent, inventive, and critical use of these tools as an essential ingredient of any effective strategy both for refining and, if necessary, amending Marx's claims and for challenging the political status quo. There is, moreover, substantial disagreement among analytical Marxists as to what the proper methodological tools are (see, e.g., Bertram (2008) and Veneziani (2012) for a map of the differences between analytical Marxism proper and its 'rational choice' variant).

The areas in which analytical Marxists have been active range from medieval history to socialist economics and from philosophical anthropology to empirical class analysis. The issues about which they have been involved in the liveliest controversies include the following:

1. Are the central propositions of historical materialism to be construed as functional explanations, i.e., as explanations of institutions by references to the functions they perform? If so, are such explanations in the social realm as legitimate as in the biological? (See Cohen, 1978; Van Parijs, 1981; Elster, 1983.)
2. Is it possible, indeed is it necessary, for a Marxist to be committed to methodological individualism, i.e., to the view that all social scientific explanations should ultimately be phrased in terms of actions and thoughts by individual human beings? Or are there some admissible structuralist Marxian explanations that are irreducible to an individualistic perspective? (See Elster, 1982, 1985; Roemer, 1985; Wright et al., 1992.)
3. Is there any way of salvaging from the ferocious criticisms to which it has been subjected the so-called theory of the falling rate of profit, i.e., Marx's celebrated claim that capitalist economies are doomed to be crisis-ridden, owing to a systematic tendency for the rate of profit to fall as a result of the very process of profit-driven capital accumulation? If the criticisms are compelling, what are the consequences both for the methodology of Marxian economics and for Marxian crisis theory? (See Roemer, 1981; Elster, 1985; Van Parijs, 1993.)
4. Can one vindicate the labor theory of value, i.e., the claim that the labor time required to produce a commodity is the

ultimate determinant of its price, against the numerous objections raised against it? If not, does this have any serious consequence for positive Marxian theory, bearing in mind that Marx himself considered this theory essential to the explanation of capitalist profits? And does it have any serious consequence for Marxian normative theory (if any), bearing in mind that Marx's concept of exploitation is usually defined in terms of labor value? (See Steedman, 1977; Roemer, 1981; Cohen, 1989.)

5. Does Marx leave any room for ethical statements, i.e., statements about what a good or just or truly free society would be like, as distinct from statements about how a society could be more rationally organized or about what it will turn out to be by virtue of some inexorable laws of history? Or should one instead ascribe to him a consistently amoralist position and, if the latter, can such a position be defended? (See Wood, 1981; Elster, 1985; Geras, 1985; Cohen, 1989.)
6. Can the concept of exploitation – commonly defined as the extraction of surplus labor, or as the unequal exchange of labor value – be made independent of the labor theory of value? Can it be extended to deal with late-capitalist or postcapitalist societies, in which the possession of a scarce skill, or the incumbency of some valued job, or the control over some organizational asset may be at least as consequential as the ownership of material means of production? Can such a more or less generalized concept of exploitation provide the basis for an empirically fruitful concept of social class? By providing a precise characterization of what counts as an injustice, can it supply the core of an ethically sensible conception of justice? (See esp. Roemer, 1982; Wright, 1985; Wright et al., 1989; Wright, 1995)
7. How can the Marxian commitment to equality (if any) be rigorously and defensibly formulated? Could the egalitarian imperative that motivates the demand for the socialization of the means of production also be satisfied by an equal distribution of the latter, or by a neutralization of the impact of their unequal distribution on the distribution of welfare? To what extent is this imperative compatible with every individual owning (in some sense) herself, as taken for granted, it would seem, in the typically Marxian idea that workers are entitled to the full fruits of their labor? (See Roemer, 1994a,b; Cohen, 1995.)
8. After the collapse of East European socialism, is it possible to reshape the socialist project in a way that takes full account of the many theoretical and practical objections that have been raised against it? Can a system be designed in which the social ownership of the means of production is combined with the sort of allocative and dynamic efficiency commonly ascribed to capitalist labor and capital markets? (See Roemer, 1994a.) Or should the radical alternative to capitalism as we know it rather be found in a "capitalist transition toward communism" through the introduction and gradual increase of an unconditional basic income or in a highly egalitarian redistribution of privately owned assets? (See van der Veen et al., 1986; Bowles et al., 1998.)

The boundaries of analytical Marxism are unavoidably fuzzy. Defined by the combination of a firm interest in some of the central themes of the Marxist tradition and the uninhibited use of rigorous analytical tools, it extends far beyond, but is strongly associated with, the so-called September Group founded in 1979 by the Canadian philosopher G.A. Cohen (Oxford) and the Norwegian social scientist Jon Elster (Columbia). The group has crystallized an attitude into a movement, by endowing it with a name, a focus, and a target for criticism (see Roemer, 1985 and 1993; Nielsen and Ware, 1989; Mayer, 1994). Having started with a critical inventory of Marx's heritage, it gradually took a more prospective turn, with a growing emphasis on the explicit elaboration and thorough defense of a radically egalitarian conception of social justice (see Cohen, 1995, 1999; Van Parijs, 1995; Roemer, 1996) and a detailed multidisciplinary discussion of specific reforms (see Van Parijs, 1992; Roemer, 1998; and the volumes in the "Real Utopias Project" directed by Erik O. Wright). This development has arguably brought analytical Marxism considerably closer to left liberal social thought than to the bulk of explicitly Marxist thought. Much contemporary work on analytical Marxism asks whether there are any distinctive Marxist (as opposed to egalitarian liberal) commitments in the realm of social and political theories and, if so, what these are (e.g., Levine, 2003; Veneziani, 2012).

See also: Critical Theory; Historical Explanation, Theories of: Philosophical Aspects; Marx, Karl (1818–83).

Bibliography

Bertram, C., 2008. Analytical Marxism. In: A Critical Companion to Contemporary Marxism. Brill Academic Press, Boston and Leiden.
Bowles, S., Gintis, H., 1998. Recasting Egalitarianism. New Rules for Communities, States and Markets. Verso, London (Part of the Real Utopias Project, 1998).
Brenner, R., et al., 1984. The Brenner Debate: Agrarian Class Structure and Economic Development in Pre-industrial Europe. Cambridge University Press, Cambridge.
Carling, A., 1991. Social Division. Verso, London.
Cohen, G.A., 1978. Karl Marx's Theory of History. A Defence. Oxford University Press, Oxford.
Cohen, G.A., 1989. History, Labour and Freedom. Themes from Marx. Oxford University Press, Oxford.
Cohen, G.A., 1995. Self-Ownership, Freedom and Equality. Cambridge University Press, Cambridge.
Cohen, G.A., 1999. If You are an Egalitarian, How Come You are So Rich? Harvard University Press, Cambridge, MA.
Elster, J., 1978. Logic and Society. Contradictions and Possible Worlds. Wiley & Sons, Chichester & London.
Elster, J., 1982. Marxism, functionalism, and game theory. Theory and Society 11, 453–482.
Elster, J., 1983. Explaining Technical Change. Cambridge University Press, Cambridge.
Elster, J., 1985. Making Sense of Marx. Cambridge University Press, Cambridge.
Geras, N., 1985. The controversy about Marx and justice. New Left Review 150, 47–85.
Levine, A., 2003. A Future for Marxism? Pluto Books, London.
Mayer, T., 1994. Analytical Marxism. Sage, London.
Nielsen, K., Ware, R. (Eds.), 1989. Analyzing Marxism. University of Calgary Press, Calgary.
Przeworski, A., 1985. Capitalism and Social Democracy. Cambridge University Press, Cambridge.
Roemer, J.E., 1981. Analytical Foundations of Marxian Economic Theory. Cambridge University Press, Cambridge.
Roemer, J.E., 1982. A General Theory of Exploitation and Class. Harvard University Press, Cambridge, Mass.
Roemer, J.E. (Ed.), 1985. Analytical Marxism. Cambridge University Press, Cambridge.
Roemer, J.E. (Ed.), 1993. Foundations of Analytical Marxism. Edward Elgar, London.

Roemer, J.E., 1994a. A Future for Socialism. Harvard University Press, Cambridge, MA; Verso, London.
Roemer, J.E., 1994b. Egalitarian Perspectives. Essays in Philosophical Economics. Cambridge University Press, Cambridge.
Roemer, J.E., 1996. Theories of Distributive Justice. Harvard University Press, Cambridge, MA.
Roemer, J.E., 1998. Equality of Opportunity. Harvard University Press, Cambridge, MA.
Steedman, I., 1977. Marx after Sraffa. New Left Books, London.
van der Veen, R.J., Van Parijs, P., 1986. A capitalist road to communism. Theory and Society 15, 635–655.
Van Parijs, P., 1981. Evolutionary Explanation in the Social Sciences. An Emerging Paradigm. Rowman & Littlefield, Totowa, NJ.
Van Parijs, P. (Ed.), 1992. Arguing for basic income. Ethical Foundations for a Radical Reform. Verso, London.
Van Parijs, P., 1993. Marxism Recycled. Cambridge University Press, Cambridge.
Van Parijs, P., 1995. Real Freedom for All. What (If Anything) Can Justify Capitalism? Oxford University Press, Oxford.
Veneziani, R., 2012. Analytical Marxism. Journal of Economic Surveys 26, 649–673.
Wood, A., 1981. Karl Marx. Routledge & Kegan Paul, London.
Wright, E.O., 1985. Classes: Methodological. Theoretical and Empirical Problems of Class Analysis. New Left Books, London.
Wright, E.O., et al., 1989. Debates on Classes. Verso, London & New York.
Wright, E.O., et al., 1992. Reconstructing Marxism: Essays on Explanation and the Theory of History. Verso, London.
Wright, E.O., 1995. Class Counts: Comparative Studies in Class Analysis. Cambridge University Press, Cambridge.

Analytical Sociology

Peter Hedström, Institute for Futures Studies, Stockholm, Sweden
Petri Ylikoski, University of Helsinki, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The core idea of analytical sociology is the importance of mechanism-based understanding of social processes. Sociological theories should provide clear and precise accounts of the social mechanisms by which the intentional activities of social agents bring about social phenomena. Theories about social mechanisms can be characterized as theories of middle range as they provide clear, precise, and simple explanations for specified aspects of range of different phenomena, without pretense of being able to explain all social phenomena. Intentional action plays an important role in social mechanisms, but the analytical sociology perspective suggests that our account of human agency should be based on findings and theories of psychological and cognitive sciences rather than on some preconceived ideas about human motivation or cognitive processing. Much of the development of mechanism-based knowledge consists of developing how-possibly explanation schemes. Agent-based computer simulations can be very useful for this kind of endeavor.

Analytical sociology is a reform movement within sociology and social theory. Its identity is not based on a common object of study, a shared historical tradition in sociological theory, or an empirical research method. Rather, it is founded on the idea that the social sciences should do more than describe and classify social processes. According to its supporters, the primary epistemic aim of the social sciences should be causal explanation of social phenomena. Sociological theory should aim to develop clear and precise accounts of the social mechanisms by which the intentional activities of social agents bring about social phenomena (Hedström, 2005; Manzo, 2010).

The emphasis on causal explanation sets analytical sociology apart from the interpretive – cultural studies-influenced – sociology that either shuns any explanatory ambitions or regards social scientific explanations as noncausal and noncontinuous with the natural sciences. For analytical sociology, the same causal and explanatory ideals that motivate natural sciences should also drive the social sciences. The natural sciences are a source of occasional inspiration and evidential support, rather than an enemy. The emphasis on mechanism-based explanation separates analytical sociology from the quantitatively oriented and variable-based sociology that prefers to stay very close to empirical data and provides explanations mostly on an ad hoc basis. Analytical sociologists do not oppose the use of quantitative data but they argue that social research should be more theory driven and that causal claims should be supported by a theoretical understanding of the causal mechanisms underlying the observed statistical regularities. Finally, the emphasis on precision and clarity and on the importance of tightly linking theory and empirical research contrasts analytical sociology with mainstream social theory. It advocates the idea that sociologists should focus on theories of middle range that can be tested and elaborated via empirical research.

The historical roots of analytical sociology can be traced back to the works of late nineteenth and early twentieth century sociologists such as Max Weber and Alexis de Tocqueville, and to prominent mid-twentieth century sociologists such as the early Talcott Parsons and Robert K. Merton. Among contemporary social scientists, Jon Elster, Raymond Boudon, Thomas Schelling, and James Coleman have profoundly influenced the analytical approach. Although they are rather different types of scholars, they complement each other in important ways, and they all share a commitment to precise, abstract, and action-based explanations.

The current generation of analytical sociologists is building upon the foundations laid by these authors aiming to develop an analytical middle-range approach to sociological theory that avoids the somewhat empiricist and eclectic tendencies of Merton's original middle-range approach. Examples of the recent work in this tradition can be found in anthologies edited by Hedström and Swedberg (1998), Hedström and Bearman (2009), Demeulenaere (2011), and in a monograph by Hedström (2005).

Analytical Realism

The philosophy underlying analytical sociology could be characterized as realistic in a number of respects. First, in contrast to empiricist and instrumentalist views, it regards explanation as the principal epistemic aim of science. Sociological theories are not merely intellectual constructions that can be used for the purposes of prediction and control of social events. The primary goal is to represent causal processes that generate the observable phenomena. For the same reason, sociological theories that merely aim to classify social processes and institutions are regarded as misdirected and lacking proper ambition. The accurate description of social phenomena is a very important part of the social sciences, but it should not be the only one.

Second, in contrast to empiricist and behaviorist views, it does not regard the use of theoretical concepts to be illegitimate or suspicious. As in the natural sciences, social scientists can employ theories that refer to unobservable entities and processes. There is no need to reduce them to observable variables. According to analytical sociologists, it is not possible to represent general causal knowledge in the social sciences as

 http://dx.doi.org/10.1016/B978-0-08-097086-8.44071-7

invariant empirical regularities. Rather, the social scientific knowledge is based on understanding the causal mechanisms underlying observable regularities. In the same spirit, analytical sociologists employ the conceptual scheme of intentional psychology that refers to unobservable psychological states like beliefs and desires. Furthermore, intentional explanations are regarded as causal explanations.

Third, contrary to the instrumentalist attitude common among many rational choice theorists, especially economists, analytical sociologists regard explanation as factive. It is not enough that the theory or model 'saves the phenomena'; it should represent the essential features of the actual causal structure that produces the observed phenomena. Thus, analytical sociology does not accept the 'as if' attitude displayed by many economists; those theoretical assumptions that have a central explanatory role should be both empirically valid and compatible with the results of other scientific fields.

Fourth, analytical sociology embodies the realistic spirit by not letting the considerations of elegance, simplicity, or tractability to override the aim of accurately describing the real causal mechanisms producing the observable phenomena. Contrary to a widespread practice among mathematically oriented economists and rational choice sociologists, analytical sociologists do not accept the instrumentalist attitude according to which assumptions are instruments that can be freely tinkered with until one arrives at simple and elegant models. Parsimonious models with clear analytical solutions deserve praise only if they are not achieved at the cost of implausible theoretical assumptions.

Rather than seeking excessively precise fictions, social scientists should aim for theoretical assumptions known to be at least roughly correct about the phenomenon that they are analyzing. Such a modest and realist strategy characterizes some of the best theoretical work in sociology. Robert K. Merton's work on self-fulfilling prophecies (Merton, 1968) provides a good example. At the core of this elegant and highly influential piece of work is the assumption that the actions of others influence individuals' beliefs and subsequent actions. However, Merton never specified any precise model of the decision calculus for the individuals. Although such model would be possible, it would not add any insights to those found in Merton's analysis. Formalization often is required for explaining social phenomena but, if the model does not properly describe action principles observed in the real world, it is of little explanatory use. It is more advisable to base the analysis on clear and empirically plausible assumptions about the actions and interactions of individuals, as Merton did, and then on this basis develop theoretical models that allow us to get a handle on the social outcomes that the actors are likely to bring about. Such analyses either generate tendency statements about patterns likely to be observed or suggest plausible processes through which the phenomena to be explained could have been brought about. More precision might be desirable, but not excessive precision that simply amounts to precisely stating and assuming to be true what is known to be untrue.

Like most economists, but unlike some sociologists, analytical sociologists regard the method of isolation and abstraction as an indispensable part of theory development: empirical reality is complex and it is futile to try to capture it as a whole. Building a complex model of a complex phenomenon does not provide much explanatory insight. For this reason, the strategy of intentional simplification is preferable: the scientist tries to capture the central features of the phenomenon first. The achievement of understanding is only possible by studying quite simplified but realistic theoretical models and gradually increasing the complexity of these models. The employment of this research strategy does not commit the scientist to the denial of the complexity of social life. To the contrary, the strategy of model building intends to avoid the hubris associated with one sweep attempt to capture complexity. According to analytical sociology, an adequate sociological theory is abstract, realistic, and precise, and it seeks to explain specific social phenomena on the basis of explicitly formulated theories of action and interaction.

Social Mechanisms

The core idea of analytical sociology is that sociological theory explains by specifying causal mechanisms that bring about social phenomena. The specifically social causal mechanisms are called social mechanisms (Hedström and Ylikoski, 2010).

Analytical sociologists are dissatisfied with the traditional covering-law account of explanation that has a large number of philosophical problems and embarrassing counterexamples. It has also provided justification for the use of 'black box' explanations in the social sciences as it does not require that the mechanism linking explanans and explanandum be specified in order for an acceptable explanation to be at hand. This omission has given leeway for sloppy scholarship. In contrast to this fundamentally empiricist account of explanation, analytical sociologists require that explanations should articulate causal mechanisms rather than simply subsume phenomena under empirical generalizations (Hedström and Ylikoski, 2010).

Analytical sociology draws from recent debates about the importance of mechanism-based explanation in the social sciences and philosophy of science (Darden, 2006; Craver, 2007; Hedström and Ylikoski, 2010). Because the entities and processes studied by different sciences are quite heterogeneous, it is difficult to propose a mechanism definition that would both be informative and cover all examples of mechanisms. Some disciplines, such as cell biology (Bechtel, 2006) and the neurosciences (Craver, 2007), study highly integrated systems, whereas others, such as evolutionary biology and the social sciences, study more dispersed phenomena. For this reason, a characterization of a mechanism that applies to one field might not be informative when applied to another.

However, some general ideas are shared by most accepted mechanism definitions (Hedström and Ylikoski, 2010). First, a mechanism is identified by the kind of effect or phenomenon it produces. A mechanism is always a mechanism for something (Darden, 2006). Second, a mechanism is an irreducibly causal notion. It refers to the entities of a causal process that produces the effect of interest. Third, the mechanism has a structure. When a mechanism-based explanation opens the black box, it discloses this structure. It turns the black box into a transparent box and makes visible how the participating entities and their properties, activities, and relations produce the effect of interest. Fourth,

mechanisms form a hierarchy. While a mechanism at one level presupposes or takes for granted the existence of certain entities with characteristic properties and activities, it is expected that there are lower level mechanisms that explain them (Craver, 2007). It is an inherent feature of the mechanism view that the entities and mechanisms of various sciences are ultimately related to each other. Although the explanatory entities and mechanisms employed by one science always bottom out somewhere (Darden, 2006) and are therefore taken as fundamental, their fundamental status is relative because they are mechanistically explainable by other fields of science.

A mechanism-based explanation describes the causal process selectively. It does not aim at an exhaustive account of all details but seeks to capture the crucial elements of the process by abstracting away the irrelevant details. The relevance of entities, their properties, and their interactions is determined by their ability to make a relevant difference to the outcome of interest. If the presence of an entity or of changes in its properties or activities truly does not make any difference to the effect to be explained, it can be ignored. This counterfactual criterion of relevance implies that mechanism-based explanations involve counterfactual reasoning about possible changes and their consequences (Ylikoski, 2011).

Roughly, mechanism-based explanations have two kinds of explananda. First, they might address particular empirical facts. In such cases, the description of the mechanism is often a modified adaptation and combination of more general mechanism schemes. Second, they might address stylized facts. Although the explanation of particular empirical facts is the ultimate goal of mechanism-based theory development, most of the time theorists are addressing highly stylized theoretical explananda that do not necessarily have close resemblance to any particular empirical fact. Consider, for example, the small-world problem of Milgram (1967). Watts and Strogatz (1998) identified some salient and abstract features of the small-world problem – average path length and local clustering – and developed a model that could explain them. The explanation of simplified and heavily idealized facts such as those identified by Watts and Strogatz results in mechanism schemes with wide application domains. By establishing how stylized facts can in principle be explained, theorists contribute to the toolbox of semigeneral mechanisms.

Apart from providing explanatory understanding as to why the dependency holds, the information about the causal mechanisms also provides justification for causal claims. Distinguishing between real causal claims and spurious statistical associations is a major challenge in the social sciences. The real causal dependences are transmitted via causal processes and the search for causal mechanisms directs the attention to these processes.

The idea of causal mechanisms is related to broader ideas about the growth and organization of scientific knowledge. In mechanism-based accounts, scientific knowledge is embedded in mechanism schemes and not in empirical generalizations as in more traditional empiricist accounts. According to this view, scientific knowledge expands by adding items to or improving upon items already present in the toolbox of possible causal mechanisms. Understanding accumulates as the knowledge of mechanisms gets more detailed and the number of known mechanisms increases. This vision of knowledge does not require that mechanisms be ultimately organized into a grand unified theory. It is only required that the accounts of mechanisms provided by different disciplines be mutually compatible and that they form an integrated web in which mechanisms at lower levels of organization explain the mechanisms that higher level disciplines take for granted. For example, psychology explains (and corrects) the assumptions that social scientists make about human cognitive processes.

The mechanism idea is important in a highly specialized and fragmented discipline like sociology. Although empirical data, research methods, and substantial theories differ from one subfield of sociology to another, the general ideas about possible causal mechanisms are something these fields could share and thereby benefit from each other's work. In this vision, sociological theory provides a set of explanatory tools that can be employed and adapted to particular situations and explanatory tasks. The mechanisms are (semi) general in the sense that most of them are not limited to any particular application. For example, the same type of mechanism can be used for (partially) explaining residential segregation (Bruch and Mare, 2006) and success in cultural markets (Salganik and Watts, 2009).

Theories of Middle Range

This mechanism-based vision of knowledge has much in common with Robert K. Merton's idea of sociological theories of the middle range (Merton, 1968). A theory of middle range is a clear, precise, and simple type of theory that can be used for partially explaining a range of different phenomena, but that makes no pretense of being able to explain all social phenomena, and that is not founded upon any form of extreme reductionism in terms of its explanans. Middle-range theories isolate a few explanatory factors that explain important but delimited aspects of the outcomes to be explained (Hedström and Udéhn, 2009).

Merton's account of self-fulfilling prophesies provides a clear example of a social mechanism. The basic idea behind the theory of the self-fulfilling prophecy is what Merton referred to as the Thomas theorem: "If men define situations as real, they are real in their consequences" (Merton, 1968: p. 475). Merton focused on the process through which an initially false belief evokes behavior that eventually makes the false belief come true. His example was a run on a bank. Once a rumor of insolvency gets started, some depositors are likely to withdraw their savings, acting on the principle that it is better to be safe than sorry. These withdrawals strengthen others' beliefs that the bank is in financial difficulties, partly because the withdrawals may actually hurt the financial standing of the bank, but more importantly because the act of withdrawal in itself signals to others that something might be wrong with the bank. This causes even more withdrawals, which further strengthens the belief, and so on. By this mechanism, even an initially sound bank may go bankrupt if enough depositors withdraw their money in the (initially) false belief that the bank is insolvent. The theory of the self-fulfilling prophecy is a good example of a social mechanism as (1) it refers to a dynamic process; (2) the collective outcome is unintended by the individuals who

bring it about; (3) the process is driven by social interactions between individuals; and (4) the process is endogenous and self-reinforcing. The mechanism is in principle quite general and it can be used to explain a range of different types of phenomena (Biggs, 2009).

Social Explananda and Individual Action

The main focus of analytical sociology is on social (as distinct from psychological) explananda that are not definable by reference to any single member of the collectivity. Important examples of such properties include (Hedström, 2005)

- typical actions, beliefs, or desires among the members of a collectivity;
- distributions and aggregate patterns such as spatial distributions and inequalities;
- topologies of networks that describe relationships between members of a collectivity; and
- informal rules or social norms that constrain the actions of the members of a collectivity.

A key insight of analytical sociology is that explanations that directly relate macrovariables to each other are unsatisfactory. They do not specify the causal mechanisms by which the changes in one macrovariable brings about changes in another. Deeper explanatory understanding requires opening up the black box and finding the causal mechanisms that have generated the macrolevel pattern. Rather than analyzing relationships between phenomena exclusively on the macrolevel, one should identify the situational mechanisms by which social structures constrain individuals' action and cultural environments shape their desires and beliefs, describe the action formation mechanisms according to which individuals choose how to act, and specify the transformational mechanisms by which individuals, through their actions and interactions, generate various intended and unintended social outcomes. Only by understanding the whole chain of situational, action-formation, and transformational mechanisms, have we made sense of the observed macrolevel relationship (Coleman, 1990; Hedström and Swedberg, 1998).

This emphasis on detailing mechanisms implies that explanations should refer to individuals, their relationships, and their actions (in addition to whatever other social properties are relevant for explaining the causes and consequences of their actions). It is important to recognize that the basic building blocks of social explanations are not mutually independent actions performed by atomistic individuals, however. Rather, individuals' actions typically are oriented toward others, and their relations to others therefore are central when it comes to explaining why they do and what they do. In addition, social relations are central for explaining why, acting as they do, individuals bring about the social outcomes they do. Consider, as an example, the case of vacancy chains as analyzed by White (1970). A retirement, motivated perhaps by a desire for a more leisurely life, creates an opportunity for others, that is, a job vacancy waiting to be filled by a new occupant. The vacancy generated by the retirement is filled by another individual whose reason for taking the job, perhaps, is to attain more status, a higher salary, or just a change in venue, but this creates another vacancy in this person's old job, and in this way vacancy chains create social interdependencies that are important for explaining mobility. Individuals' desires – for retirement, promotion, status, or change in venue – motivate the system. Without such orientations, people may not move. But explanatory understanding is only achieved by recognizing that actions take place in relational structures that in this case channel mobility opportunities and thereby explain why we observe what we observe (Hedström and Bearman, 2009b).

This line of reasoning does not imply commitment to the doctrine of methodological individualism (Ylikoski, 2012). Most formulations of methodological individualism are much stronger than the mechanism-based perspective requires (Udéhn, 2001). Some form of structural individualism (Coleman, 1990; Hedström, 2005) is sufficient for the purposes of mechanism-based explanations of social phenomena. Structural individualism is a doctrine according to which all social facts, their structure and change, are in principle explicable in terms of individuals, their properties, actions, and relations to one another. Structural individualism differs from most formulations of methodological individualism by emphasizing the explanatory importance of relations and relational structures. It does not require that all explanatory facts are facts about individual agents in the strict sense: facts about topologies of social networks; about distributions of beliefs, resources, or opportunities; and about institutional or informal rules and norms can have a significant role in mechanism-based explanations.

Intentional action plays an important role in social mechanisms, but the idea of a social mechanism in itself does not tell us how to conceptualize human action. Rather than relying on some preconceived ideas about human motivation or cognitive processing, analytical sociology perspective suggests that our account of human agency should be based on findings and theories of psychological and cognitive sciences. Analytical sociology's focus on social, rather than psychological, mechanisms has some methodological implications, however. Understanding of complex phenomena is only possible in a piecemeal way, so sociologists must abstract away from many details of human mental life. Only those aspects of cognition that are relevant for the explanatory task at hand should be included in the explanation, and the explanatory task thus determines how rich the psychological assumptions must be.

So although the mechanism-based approach emphasizes the importance of action in the explanation of social phenomena, it does not subscribe to an axiomatic vision according to which a specific action theory should be used for all purposes. As different theories of action emphasize different aspects of human action, the choice between them should be made on methodological grounds. In some modeling contexts, rational choice theory is a natural choice, while in contexts where the emphasis is on the habitual dimensions of human actions, analytical sociologist might choose to employ some psychological dual-process theory (Wilson, 2002). However, for many social scientific purposes, a relatively simple desire–belief–opportunity model will be sufficient (Hedström, 2005). This simple theory provides a building block for accounts of social mechanisms of interaction through which

the actions of some actors may come to influence the beliefs, desires, opportunities, and actions of others.

Historically, analytical sociology has some of its roots in the tradition of rational choice sociology. However, current analytical sociology does not have any special commitment to the assumptions of rational choice theory. There is nothing in the idea of a mechanism-based explanation that would require the explanation to be articulated in terms of rational choice theory. On the contrary, the requirement that mechanism-based explanations cite actual causes of the phenomenon to be explained often makes rational choice explanations unacceptable, as they are built on implausible psychological and sociological assumptions. Empirically false assumptions about human motivation, cognitive processes, access to information, or social relations cannot bear the explanatory burden in a mechanism-based explanation. In order for unrealistic psychological and social assumptions to be acceptable, they must be simplifying idealizations that help the modeling but do not affect the central explanatory relationships in any crucial manner. This rarely is the case in rational choice theory. For this reason, it is a mistake to take analytical sociology as rebranded rational choice sociology. Rather than relying on some preconceived ideas about human motivation or cognitive processing, the mechanism-based perspective suggests that our account of human agency should be based on findings and theories of psychological and cognitive sciences.

Agent-Based Modeling

A recent development in analytical sociology has been the use of agent-based computer simulations (Hedström, 2005; Macy and Flache, 2009; Squazzoni, 2012). Much of the development of mechanism-based knowledge consists of developing how-possibly explanation schemes. These schemes are not intended to explain any particular empirical facts directly, but to provide a general understanding of how things could work. Given the limitations of experimental methods and the complexity of social phenomena, computer simulations are important for this kind of endeavor. Computer simulations allow systematic exploration of consequences of modeling assumptions and make it possible to model much more complex phenomena than was possible earlier.

The promise of agent-based simulation is based on the fact that the dynamics observed at the social level typically are complex and hard to understand, but often it is possible to describe the basic cogs and wheels of these social processes with rather simple models. Macrolevel outcomes and relationships tell us very little about why we observe the macrolevel outcomes and the relationships we observe. Only by explicitly considering the microlevel actions and relations and how they unfold over time can macrolevel outcomes be explained. This basic insight is at the heart of analytical sociology: to understand collective dynamics, we must study the collectivity as a whole, but we must not study it as a collective entity.

One important feature of agent-based simulations is that they do not impose any a priori constraints on the mechanisms assumed to be operating. Unlike rational choice theory, agent-based modeling is not based on any specific theory of action or interaction. It is a methodology for deriving the social outcomes that groups of interacting actors are likely to generate, whatever the action logics or interaction structures may be. Agent-based simulations should not only be regarded as a tool for theoretical exploration, however. Empirically calibrated agent-based models make it possible to integrate theoretical ideas with the results of empirical research (e.g., Bruch and Mare, 2006).

Conclusion

Mechanism talk is becoming increasingly popular in the social sciences. Many sociologists use the word 'mechanism' in a loose sense without any commitment to the type of mechanism-based explanatory strategy focused upon here. Underlying the mechanism-based approach is a commitment to realism: the explanations should reflect the causal processes responsible for the phenomenon to be explained. This requires stringency in theoretical practice and imagination in the design of empirical research.

One of the basic premises of analytical sociology is that proper understanding of collective processes requires that the researchers pay attention to the entities that mechanisms are made of (the agents, their properties, actions, and relations) rather than treating them as opaque entities. This dogma of analytical sociology means that theoretical models as well as empirical research must focus on actions and relations and how they unfold over time. This basic principle has far-reaching consequences for the type of empirical data that needs to be used as well as for the type of theoretical work that is needed. As far as data and empirical research are concerned, longitudinal data with relational information are essential, given that the basic entities of social mechanisms are actors, their actions, and relations. Similarly, the mechanism approach has considerable consequences for how theoretical work ought to be conducted: rigorous theorizing is needed that explicitly considers the microlevel mechanisms at work and the dynamic processes that they give rise to.

See also: Explanation: Conceptions in the Social Sciences; Mechanism-Based Causal Analysis; Methodological Individualism in Sociology; Rational Choice Theory in Sociology; Social Simulation: Computational Models; Sociological Theory.

Bibliography

Bechtel, W., 2006. Discovering Cell Mechanisms. The Creation of Modern Cell Biology. Cambridge University Press, New York.

Biggs, M., 2009. Self-fulfilling prophecies. In: Hedström, P., Bearman, P. (Eds.), The Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford, pp. 294–314.

Bruch, E.E., Mare, R.D., 2006. Neighborhood choice and neighborhood change. American Journal of Sociology 112, 667–709.

Coleman, J.S., 1990. Foundations of Social Theory. The Belknap Press of the Harvard University Press, Cambridge, MA.

Craver, C., 2007. Explaining the Brain. Mechanisms and the Mosaic Unity of Neuroscience. Oxford University Press, Oxford.

Darden, L., 2006. Reasoning in Biological Discoveries. Essays on Mechanisms, Interfield Relations, and Anomaly Resolution. Cambridge University Press, Cambridge, UK.

Demeulenaere, P., 2011. From Social Mechanisms to Analytical Sociology. Cambridge University Press, Cambridge.
Hedström, P., 2005. Dissecting the Social: On the Principles of Analytical Sociology. Cambridge University Press, Cambridge.
Hedström, P., Bearman, P. (Eds.), 2009. The Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford.
Hedström, P., Swedberg, R. (Eds.), 1998. Social Mechanisms: An Analytical Approach to Social Theory. Cambridge University Press, Cambridge.
Hedström, P., Udéhn, L., 2009. Analytical sociology and theories of the middle range. In: Hedström, P., Bearman, P. (Eds.), The Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford, pp. 25–49.
Hedström, P., Ylikoski, P., 2010. Causal mechanisms in the social sciences. Annual Review of Sociology 36, 49–67.
Macy, M.W., Flache, A., 2009. Population dynamics from the bottom up: agent-based models of social interaction. In: Hedström, P., Bearman, P. (Eds.), The Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford, pp. 245–268.
Manzo, G., 2010. Analytical sociology and its critics. European Journal of Sociology 51, 129–170.
Merton, R.K., 1968. Social Theory and Social Structure. Free Press, New York.
Milgram, S., 1967. The small world problem. Psychology Today 2, 60–67.
Salganik, M.J., Watts, D.J., 2009. Social influence: the puzzling nature of success in cultural markets. In: Hedström, P., Bearman, P. (Eds.), The Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford, pp. 315–341.
Squazzoni, F., 2012. Agent-based Computation Sociology. John Wiley & Sons, Sussex.
Udéhn, L., 2001. Methodological Individualism: Background, History and Meaning. Routledge, London.
Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small world' networks. Nature 393, 440–442.
White, H.C., 1970. Chains of Opportunity: System Models of Mobility in Organizations. Harvard University Press, Cambridge, MA.
Wilson, T.D., 2002. Strangers to Ourselves. Discovering the Adaptive Unconscious. Belknap, Cambridge, MA.
Ylikoski, P., 2011. Social mechanisms and explanatory relevance. In: Demeulenaere, P. (Ed.), From Social Mechanisms to Analytical Sociology. Cambridge University Press, Cambridge, pp. 154–172.
Ylikoski, P., 2012. Micro, macro, and mechanisms. In: Kincaid, H. (Ed.), The Oxford Handbook of Philosophy of the Social Sciences. Oxford University Press, Oxford, pp. 21–45.

Anaphora

Yan Huang, University of Auckland, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Anaphora is definable as a relation between two or more linguistic elements, in which the interpretation of one (called an anaphoric expression) is in some way determined by the interpretation of the other (called an antecedent). This essay defines anaphora and presents typologies of anaphora on the basis of syntactic categories, truth-conditions, and reference-tracking systems. It also provides an authoritative, up-to-date, and critical survey of the latest theoretical treatments of intrasentential anaphora, comparing and contrasting the two main rival approaches: grammatical and pragmatic.

Defining Anaphora

The term 'anaphora' comes from the Greek word *αναφορά*, which means 'carrying back.' In contemporary theoretical linguistics, there are three distinct senses of 'anaphora/anaphor/anaphoric.' ((1) A relation between two or more linguistic elements, in which the interpretation of one (called an anaphoric expression) is in some way determined by the interpretation of the other (called an antecedent). In other words, in an anaphoric relation, an anaphoric expression takes its meaning or reference from its antecedent. Many aspects of anaphora are pragmatic in nature. (2) An anaphoric relationship in which the antecedent comes earlier than the anaphoric expression. Originally, 'anaphora' was a term for a relation 'up' (Greek *aná*). Anaphora (2) is also called variously forward anaphora, backward-looking anaphora, or retrospective anaphora.) In the first place, the term can be used to refer to a relation between two or more linguistic elements, in which the interpretation of one (called an anaphor or anaphoric expression) is in some way determined by the interpretation of the other (called an antecedent). Linguistic elements that can be employed to encode an anaphoric relation in this general sense range from gaps (or empty categories) through reflexives to various discourse reference-tracking systems like switch function (see the section 'Anaphora and Discourse: Reference-Tracking Systems'). (Keeping track of the various entities referred to in an ongoing discourse. Linguistic mechanisms that are deployed for such a purpose are called reference-tracking systems. Cross-linguistically, four main types of reference-tracking systems can be identified:(i) gender or class systems, (ii) switch-reference systems, (iii) switch-function systems, and (iv) inference systems.) Secondly, the term can be deployed in Chomsky's (1981, 1995) generative grammar to refer to a noun phrase (NP) with the features [+anaphoric, −pronominal] 'anaphor' versus an NP with the features [−anaphoric, +pronominal] 'pronominal.' Thus, the reflexive in (1) is treated as an anaphor, and the pronoun in (2) as a pronominal in the Chomskian tradition.

1. Obama praised *himself*.
2. Obama praised *him*.

Thirdly, the term can be used for reference to an anaphoric expression whose antecedent comes earlier, as in (3), as opposed to the term 'cataphora/cataphor/cataphoric,' where the antecedent comes later, as in (4). (An anaphoric (1) relationship in which the antecedent of an anaphoric expression comes later than the anaphoric expression itself. Put the other way round, an expression 'looks forward' for its antecedent. Cataphora is a term for a relation 'down' (Greek *katá*). Also called variously anticipatory anaphora, backward anaphora, forward-looking anaphora, and retrospective anaphora.)

3. After the zoo keeper was wounded, *he* was taken to a hospital.
4. After *he* was wounded, the zoo keeper was taken to a hospital.

Both anaphora and cataphora in this sense can be subsumed under the term 'endophora/endophor/endophoric' (a term that is sometimes used to subsume anaphora and cataphora, viewed as a relation in which an anaphoric expression and its antecedent are within ('*endo-*') the same sentence or discourse), referring to the relation where the anaphoric/cataphoric expression and its antecedent are within the same conversation or text. In contrast, by the term 'exophora/exophor/exphoric' (in opposition to endophora/endophoric, the term refers to a relation in which the referent of an anaphoric expression lies outside ('*exo-*') what is said or written. Also called contextual anaphora and situational anaphora) is meant that the antecedent of an anaphoric, or more accurately a deictic, or indexical, expression lies outside the conversation or text. This is exemplified in (5).

5. (Referents in the physical environment, and with selecting gesture)
   She's not the principal. *She* is.

Anaphora can be intrasentential, in which case the anaphoric expression and its antecedent occur within a single simplex or complex sentence. It can also be discoursal, in which case the anaphoric expression and its antecedent cross sentence boundaries. In addition, there are anaphoric devices that lie in between pure intrasentential anaphora and pure discourse or text anaphora.

Anaphora is at the center of twenty-first-century research on the interface between syntax, semantics, and pragmatics in theoretical linguistics. It is also a subject of key interest in psycho- and computational linguistics, and to work on the philosophy of language, and language in cognitive science. It has aroused this interest for a number of reasons. To begin with, anaphora represents one of the most complex phenomena of natural language, which, in itself, is the source of fascinating problems. Secondly, anaphora has long been regarded as one of the few 'extremely good probes' in furthering our understanding of the nature of the human mind, and thus in facilitating an answer to what Chomsky considers to be the fundamental problem of linguistics, namely, the logical problem of language acquisition – a special case of Plato–Russell's problem. In particular, certain aspects of anaphora have repeatedly been claimed by Chomsky to furnish evidence for the argument that human beings are born equipped with some internal, unconscious knowledge of language, known as the language faculty. Thirdly, anaphora has been shown to interact with syntactic, semantic, and pragmatic factors. Consequently, it has provided a test bed for competing hypotheses concerning the relationship between syntax, semantics, and pragmatics in linguistic theory.

Typologies of Anaphora

Anaphora can be classified on the basis of (a) syntactic categories, (b) truth-conditions, and (c) discourse reference-tracking systems.

Anaphora and Syntactic Categories

In terms of syntactic category, anaphora falls into two main groups: (a) NP- (noun phrase-), including N- (noun-), anaphora, and (b) VP- (verb phrase-) anaphora. In an NP-anaphoric relation, both the anaphoric expression and its antecedent are in general NPs, and both are potentially referring expressions (6). NP-anaphora corresponds roughly to the semantically defined type of 'identity of reference' anaphora. It can be encoded by gaps (or empty categories), pronouns, reflexives, epithets, proper names, and definite descriptions. By contrast, in an N-anaphoric relation, both the anaphoric expression and its antecedent are a noun, rather than an NP, and neither is a potentially referring expression (7). N-anaphora corresponds roughly to the semantically defined type of 'identity of sense' anaphora. Linguistic elements that can be used as an N-anaphor include gaps, pronouns, and nouns.

6. Bill thought that *he* had reached a different conclusion.
7. John's favorite composer is Bach, but Mary's Ø is Handel.

The other main category is VP-anaphora. Under this rubric, five types may be isolated: (a) VP-ellipsis, in which the VP of the second and subsequent clauses is reduced (8); (b) gapping, in which some element (typically a repeated, finite verb) of the second and subsequent conjuncts of a coordinate construction is dropped (9); (c) sluicing, which involves an elliptical construction consisting only of a *wh*-phrase (10); (d) stripping, an elliptical construction in which the ellipsis clause usually contains only one constituent (11); and (e) null complement anaphora, an elliptical construction in which a VP complement of a verb is omitted (12).

8. John adores Goya's paintings, and Steve does, too.
9. Reading maketh a full man; conference a ready man; and writing an exact man.
10. John donated something to Médecins Sans Frontières, but I don't know what.
11. Pavarotti will sing 'Nessun dorma' again, but not in Hyde Park.
12. Mary wanted to swim the English channel, but his father didn't approve.

Anaphora and Truth-Conditions

From a truth-conditional, semantic point of view, anaphora can be divided into five types: (a) referential anaphora, one that refers to some entity in the external world either directly or via its co-reference with its antecedent in the same sentence or discourse (13); (b) bound-variable anaphora, which is interpretable by virtue of its dependency on some quantificational expression in the same sentence or discourse, thus seeming to be the natural language counterpart of a bound variable in first-order logic (14); (c) E- (Evans-) type anaphora, one which, for technical reasons, is neither a pure referential anaphor nor a pure bound-variable anaphor, but which nevertheless constitutes a unified semantic type of its own (15); (d) anaphora or pronoun of 'laziness,' so called because it is neither a referential anaphor nor a bound-variable anaphor, but functions as a shorthand for a repetition of its antecedent (16); and (e) bridging cross-reference anaphora, one that is used to establish a link of association with some preceding expression in the same sentence or discourse via the addition of background assumption (17).

13. The farmer said that he has aerated the soil this morning.
14. Every little girl wishes that she could visit the land of Lilliput.
15. Most people who bought a donkey have treated it well.
16. The man who gave his paycheck to his wife was wiser than the man who gave it to his mistress.
17. John dined in an Italian restaurant. The waiter was very handsome.

Anaphora and Discourse: Reference-Tracking Systems

Reference-tracking systems are mechanisms employed in individual languages to keep track of the various entities referred to in an ongoing discourse. In general, there are four major types of reference-tracking systems in the world's languages: (a) the gender/class system, (b) the switch-reference system, (c) the switch-function system, and (d) the inference system.

In a gender or class system, an NP or (less frequently) a VP is morphologically classified for gender or class according to its inherent features, and is tracked through a discourse via its association with the gender or class assigned. This reference-tracking device is found to be present in a large

variety of languages such as Archi, Swahili, and Yimas. Next, in a switch-reference system, the verb of a dependent clause is morphologically marked to indicate whether or not the subject of that clause is the same as the subject of its linearly adjacent, structurally related independent clause. If both subjects are co-referential, an SS (Same Subject) marker is used; otherwise, a DS (Different Subject) marker is employed. Switch-reference is found in many of the indigenous languages spoken in North America, the non-Austronesian languages spoken in Papua New Guinea, and the aboriginal languages spoken in Australia. It has also been found in languages spoken in North Asia and Africa. Somewhat related to the switch-reference system is the switch-function system. By switch function is meant the mechanism that tracks the reference of an NP across clauses in a discourse by means of verbal morphology indicating the semantic function of that NP in each clause. This system is found in a wide range of languages, including English, Dyirbal, and Jacaltec. Finally, there is the inference system. In this system, reference-tracking in discourse is characterized by (a) the heavy use of zero anaphors, (b) the frequent appeal to sociolinguistic conventions, and (c) the resorting to pragmatic inference. This mechanism is particularly common in East and Southeast Asian languages like Chinese, Javanese, and Tamil. Of the four reference-tracking devices mentioned above, the first is considered to be lexical in nature, the second and third, grammatical in nature, and the fourth, pragmatic in nature (see, e.g., Huang, 2000a and references therein).

Grammatical versus Pragmatic Approaches

With regard to intrasentential anaphora, two main types of analysis can be identified: (a) grammatical including syntactically oriented, semantically oriented, and a mixture of syntactically and semantically oriented, and (b) pragmatic.

Gramatical

Central to the various grammatical analyses is the belief that anaphora is largely a syntactic and semantic phenomenon, and as such reference must be made to conditions and constraints that are essentially syntactic and semantic in nature. This approach is best represented by Chomsky's (1981, 1995) binding theory and Reinhart and Reuland's (1993) theory of reflexivity within the principles-and-parameters theory and its minimalist descendant in the tradition of generative grammar. In addition, a current development in the grammatical analysis of anaphora and binding is to abandon all the conditions that are specific to binding and reduce them to elementary, general, and independent principles of the computational system of language within Chomsky's minimalist program (see, e.g., Safir, 2004; Hicks, 2009; Rooryck and Wyngaerd, 2011; especially Reuland, 2011).

Chomsky distinguishes two types of abstract features for NPs: anaphors and pronominals. An anaphor is a feature representation of an NP that must be referentially dependent and that must be bound within an appropriately defined minimal syntactic domain; a pronominal is a feature representation of an NP that may be referentially dependent but that must be free within such a domain. Interpreting anaphors and pronominals as two independent binary features, Chomsky hypothesizes that one ideally expects to find four types of NP in a language – both overt and nonovert – see Table 1.

Table 1 Chomsky's typology of NPs

	Overt	*Empty*
a. [+anaphor, −pronominal]	reflexive/reciprocal	NP-trace
b. [−anaphor, +pronominal]	pronoun	pro
c. [+anaphor, +pronominal]	–	PRO
d. [−anaphor, −pronominal]	Name	wh-trace

Of the four types of NPs listed in Table 1, anaphors, pronominals, and r- (referential-) expressions are subject to binding conditions A, B, and C, respectively (see Table 2).

Binding is defined on anaphoric expressions in configurational terms, appealing to purely structural concepts like *c-command*, *government*, and *locality*. The binding theory accounts for the distribution of anaphoric expressions in Table 3. It is also applied to empty categories: *NP-trace*, *pro*, and *wh-trace*. It has even been extended to analyzing VP-ellipsis and switch-reference, but that work will not be surveyed here.

In contrast to Chomsky's 'geometric,' syntactic analysis, the semantically based approach maintains that anaphora is subject mainly to semantic constraints. Consequently, it can be explained in semantic terms. Under this approach, binding is frequently defined in terms of argument-structure.

On Reinhart and Reuland's view, reflexivity is not a property of NPs, but a property of predicates. The binding theory is designed not to capture the mirror-image distribution of anaphors and pronominals, but to regulate the domain of reflexivity for a predicate. Putting it more specifically, what the theory predicts is that if a predicate is lexically reflexive, it may not be reflexive-marked by a morphologically complex SELF anaphor (e.g., *zichzelf* in Dutch, *se stesso* in Italian, and *zibun-zisin* in Japanese) in the overt syntax. And if a predicate is not lexically reflexive, it may become reflexive only via the marking of one of its co-arguments by the use of such an anaphor (see Tables 4 and 5).

There are, however, considerable problems with both versions of the binding theory cross-linguistically. Regarding Chomsky's theory, the distribution of reflexives violates binding condition A in both directions: on the one hand, a reflexive can be bound outside its local syntactic domain (so-called long-distance reflexive), as in Chinese, Icelandic, and Tuki; on the other, it may not be bound within its local domain, as in Dutch and Norwegian. Binding condition B is also frustrated, for in many of the world's languages (such as

Table 2 Chomsky's binding conditions

A. An anaphor is bound in a local syntactic domain.
B. A pronominal is free in a local syntactic domain.
C. An r-expression is free.

Table 3 Anaphoric expressions

a. Handel$_1$ admired himself$_1$.
b. Handel$_1$ admired him$_2$.
c. Handel$_1$ admired Handel$_2$.

Table 4 Reinhart and Reuland's binding conditions

A. A reflexive-marked syntactic predicate is reflexive.
B. A reflexive semantic predicate is reflexive-marked.

Table 5 Reinhart and Reuland's typology of overt NPs

	SELF	*SE*	*Pronoun*
Reflexivizing function	+	–	–
Referential independence	–	–	+

Danish, Gumbaynggir, and Piedmontese), a pronominal can be happily bound in its local syntactic domain. Next, given Chomsky's formulation of binding conditions A and B, it is predicted that anaphors and pronominals be in strict complementary distribution, but this predicted complementarity is a generative syntactician's dream world. Even in a 'syntactic' language like English, it is not difficult to find syntactic environments where the complementarity breaks down. Finally, even a cursory inspection of languages like English, Japanese, and Vietnamese indicates that binding condition C cannot be taken as a primitive of grammar.

Next, when we move to Reinhart and Reuland's analysis, we find that in the first place, cross-linguistic evidence has been presented that marking of reflexivity is not limited to the two ways they have identified. In addition to being marked lexically and syntactically, reflexivity can also be indicated morphologically. Secondly, more worrisome is that the central empirical prediction of the reflexivity analysis, namely, only a reflexive predicate can and must be reflexive-marked, is falsified in both directions. On the one hand, a predicate that is both syntactically and semantically reflexive can be nonreflexive-marked; on the other, a nonreflexive predicate can be reflexive-marked.

As mentioned above, a current development in the grammatical analysis of anaphora and binding is to eliminate all the conditions that are put forward specifically for binding such as Chomsky's binding conditions A, B, and C, discussed above, and to reduce these specific conditions to elementary, general, and independent principles of the computational system of language within Chomsky's minimalist program. For example, based mainly on the study of a range of Germanic and Romance languages, Reuland (2011) argues that anaphoric dependencies and their (substantial) cross-linguistic variations (in these languages) can be derived naturally by the interaction of the lexicon, syntax, semantics, and discourse components of the language system, which reflect the cognitive capacities and design of language.

This new development constitutes an important step forward in our understanding of anaphora and binding. Firstly, from a theoretical and methodological point of view, to account for anaphora and binding in terms of some elementary, general, and independent principles of Chomsky's minimalist program is much preferable to explaining them by means of some ad hoc, local conditions that are stipulated specifically for them. Secondly, while their focus is still on the syntax and semantics of anaphora and binding, the latest generative analyses of the type exemplified by Reuland (2011) allow nongrammatical or pragmatic factors to play a role.

On the other hand, however, these accounts are not without problems of their own. Again, take Reuland (2011) as an example. Theoretically and methodologically, Reuland's analysis still suffers from containing too many rules, conditions, and principles. In other words, the theoretical machinery is not parsimonious enough. Next, from an empirical perspective it is unclear to what extent Reuland's account can extend to languages other than the Germanic and Romance ones he has examined. For example, within Reuland's framework, the syntactic binding of a morphologically simplex SE-anaphor (e.g., *zich* in Dutch, *ziji* in Chinese, and *caki* in Korean) needs to enter an *A-chain* (argument-chain) that is created by movement. In order for an SE-anaphor to enter an A-chain, structural Case must be checked, which is achieved via the inflectional system of a language. While this analysis works for a Germanic (such as Dutch) and/or Romance language (such as French), it does not apply to, e.g., Chinese, because Chinese is a language that lacks any inflectional mechanism. Consequently, given that in Reuland's system, no structural Case checking means no A-chain, binding of an SE-anaphor in a language like Chinese has to be ruled out in narrow syntax (but allowed in pragmatics).

Pragmatic

Encouraging progress in the study of anaphora and binding in the last two decades has been the development of pragmatic approaches, the most influential of which is the (revised) neo-Gricean pragmatic theory of anaphora constructed by Levinson (1987, 2000) and Huang (1991, 1994/2007, 2000a,b, 2007, 2010a,b; see also Chiou and Huang, 2010). The central idea underlying this theory is that anaphora including binding is largely pragmatic in nature, though the extent of anaphora being pragmatic varies typologically. Consequently, anaphora can largely be determined by the systematic interaction of some general neo-Gricean pragmatic principles such as Levinson's (2000) Q- (Quantity-) ('Don't say less than is required/What is not said is not the case'), I- (Informativnes-) ('Don't say more than is required/What is generally said is stereotypically and specifically exemplified'), and M- (Manner-) ('Don't use a marked expression without reason/What is said in a marked way conveys a marked message') principles, depending on the language user's knowledge of the range of options available in the grammar, and of the systematic use or avoidance of particular anaphoric expressions or structures on particular occasions. **Table 6** is a revised neo-Gricean pragmatic apparatus for anaphora.

Table 6 Huang's revised neo-Gricean pragmatic apparatus (simplified)

i. The use of an anaphoric expression x I-implicates a local co-referential interpretation, unless (ii) or (iii).
ii. There is an anaphoric Q-scale ⟨x, y⟩, in which case the use of y Q-implicates the complement of the I-implicature associated with the use of x in terms of reference.
iii. There is an anaphoric M-scale {x, y}, in which case the use of y M-implicates the complement of the I-implicature associated with the use of x in terms of either reference or expectedness.

Needless to say, any interpretation generated by the neo-Gricean pragmatic mechanism in **Table 6** is subject to the general consistency constraints applicable to Gricean conversational implicatures. These constraints include world knowledge, contextual information, and semantic entailments.

There is substantial cross-linguistic evidence to show that, empirically, the neo-Gricean pragmatic theory of anaphora is more adequate than both a syntactic and a semantic approach. Consider, for example, Chomsky's binding conditions, as in **Table 2**, and its paradigmatic illustrations, as in **Table 3**. On the neo-Gricean account, Chomsky's binding conditions B and C can be reduced to pragmatics. In somewhat simplified terms, this can be achieved in the following way. If binding condition A is taken to be either grammatically constructed (as in the English-type, syntactic languages) or pragmatically specified (as in the Chinese-type, pragmatic languages), then binding condition B is the direct result of the application of the Q-principle. Given this principle, the use of a semantically weaker pronoun where a semantically stronger reflexive could occur gives rise to a conversational implicature that conveys the negation of the more informative, co-referential interpretation associated with the use of the reflexive, as in the b sentence in **Table 3**. By the same reasoning, binding condition C can also be eliminated. Wherever a reflexive could occur, the use of a semantically weaker proper name Q-implicates the nonapplicability of the more informative, co-referential interpretation associated with the use of the reflexive. This is exactly what has happened in the c sentence in **Table 3**. Furthermore, the revised neo-Gricean pragmatic theory can provide an elegant account of many of the anaphoric patterns that have embarrassed a generative analysis such as the cases where a pronoun is bound in its local syntactic domain. It can also accommodate examples like (3) and (6), which fall outside the scope of Chomsky's binding theory. In the case of long-distance reflexivization, the concept of logophoricity can be invoked to explain why such a marked anaphoric expression is used. (The phenomenon whereby the perspective or point of view of an internal protagonist of a sentence or discourse, as opposed to that of the current, external speaker, is being reported by utilizing some morphological and/or syntactic means. The technical term 'perspective' is intended to include words, thoughts, knowledge, emotion, perception, and space-location. Logophoricity is commonly encoded by a logophor or logophoric expression. It is largely a pragmatic phenomenon.) By logophoricity is meant the phenomenon whereby the 'point of view' of an internal protagonist of a sentence or discourse, as opposed to that of the current, external speaker, is being reported using some morphological and/or syntactic means. The expression 'point of view' is employed here in a technical sense and is intended to encompass words, thoughts, knowledge, emotion, and perception (see, e.g., Huang, 1991, 1994/2007, 2000a, 2002; see also Hagège, 1974; Clements, 1975). This use of long-distance reflexives is accountable in terms of the M-principle. Since the grammar allows the unmarked regular pronoun to be employed to encode co-reference, the speaker will use it if such a reading is intended, as in the Icelandic example (18a) (cited in Huang, 2000a: 227) (INDIC = indicative mood, SBJV = subjunctive mood). On the other hand, if the unmarked regular pronoun is not used, but the marked long-distance reflexive is employed instead, then an M-implicature will be licensed. The implicature is that not only co-reference but logophoricity as well is intended by the speaker. This is the case with the Icelandic example (18b).

18. a.	Jon	veit	að	Maria	elskar	*hann*
	John	knows-INDIC	that	Mary	loves-INDIC	him
	'John$_1$ knows that Mary loves him$_1$.'					
b.	Jon	segir	að	Maria	elski	*sig.*
	John	says-INDIC	that	Mary	loves-SBJV	self
	'John$_1$ says that Mary loves self$_1$.'					

No another correlation here. If relevant, the choice between regular pronouns on the one hand and logophoric long-distance reflexives on the other is correlated with that between indicative and subjunctive mood in the embedded clause. The use of a regular pronoun tends to go with that of indicative mood, as in (18a); the employment of a logophoric long-distance reflexive tends to go with subjunctive mood, as in (18b). This correlation is a reflection of a semantic/pragmatic choice made by the external speaker about the responsibility he or she assumes for the truthfulness of what he or she is reporting. If a regular pronoun and indicative mood are used, it indicates that the speaker asserts that the report is true. On the other hand, if a logophoric, long-distance reflexive and subjunctive mood are deployed, it shows that the speaker does not take responsibility for the truth of the report. More recently, a generalized, unified pragmatic analysis of the switch-reference system in Amele and logophoric verbal suffixation in Gokana has been successfully made in terms of the neo-Gricean pragmatic theory (see, e.g., Huang, 2010c). Finally, the revised neo-Gricean pragmatic theory can also be extended to provide an account of discourse anaphora in naturally occurring conversations (see, e.g., Huang, 1994/2007, 2000a, 2000b).

Next, from a conceptual point of view, the revised neo-Gricean pragmatic theory of anaphora has important implications for current linguistic theorizing. One major implication is that the theory forces a radical rethink of some of the current claims about the nature of grammatical rules and the way in which they interact with pragmatic principles. In the first place, many grammatical rules underlying anaphoric universals such as Chomsky's binding conditions are general, violable tendencies rather than absolute, exceptionless restrictions. Secondly, these rules are not something *sui generis*, but rather have their origins in language use. In other words, they are best seen as, to use Levinson's

(1987) metaphor, 'frozen pragmatics' – the outcome of a gradual, diachronic process from utterance-token-meaning via utterance-type-meaning to sentence-type-meaning. This, of course, does not mean that these rules as they are today are not part of the grammar. On the contrary, they are and as such they should be dealt with in the grammar. But the point is that if they are the result of a historical grammaticalization process, they can no longer be held as evidence for a biologically preprogrammed human linguistic faculty. Thirdly, pragmatics and syntax are interconnected in regulating anaphora and binding, though they are distinct levels and modes of linguistic explanation. The interface between pragmatics and syntax may in general be summarized in a Kantian apothegm: pragmatics without syntax is empty; syntax without pragmatics is blind. Furthermore, the extent to which pragmatics and syntax interact varies typologically. There exists a class of languages (such as Chinese, Japanese, and Korean) where pragmatics plays a central role which in familiar European languages (such as English, French, and German) has hitherto been alleged to be played by the grammar. In these pragmatic languages, many of the constraints on anaphora are primarily due to principles of language use rather than rules of grammatical structure. From a diachronic viewpoint, languages seem to change from being more pragmatic to more syntactic; from a synchronic perspective, different languages may simply be at different stages of this evolution.

A second major ramification the revised neo-Gricean pragmatic theory of anaphora has is concerned with current thinking about important issues such as universals, innateness, and learnability. While a pragmatic approach to anaphora such as ours does not entirely contradict Chomsky's innateness hypothesis, the fact that so many anaphoric processes in such a wide variety of geographically, genetically, and typologically distinct languages can be explained as learnable on the basis of language use does undermine Chomsky's claim to the universality of universal grammar and decreases the plausibility of his innateness hypothesis. This is because anaphora as such can no longer be taken to be mere instantiations of human mental templates. There are aspects of anaphoric universals that clearly are of a grammatical nature; there are also aspects of anaphoric universals that are equally clearly of a pragmatic nature. Put in a slightly different way, anaphoric universals are the product of both nature and nurture. The challenge before the linguist is to work out what are the grammatical rules and what are the pragmatic principles precisely, what is the relationship between them, and how they interact with each other.

It is encouraging that recent developments in generative grammar seem to be moving toward the position of the pragmatics–syntax interface being advocated here. Given that within Chomsky's minimalist program, the syntax/computational system is neither a phonological nor a semantic component, a large amount of current syntactic explanation has to be shifted to the logical form and phonological form interfaces. As a result, syntax has inevitably to interact with semantics and pragmatics in a much more extensive way to link it to the mental world of cognition (see, e.g., Chomsky, 1995). This holds, of course, for the study of anaphora, as we have already seen in Reuland (2011). These new developments will open the way for a more interactive approach between the I- (internalized and intensional-) and E- (externalized and extensional-) models of language study (see, e.g., Chomsky, 1995). It seems unlikely that we can provide a satisfactory answer to what Chomsky has referred to as Humboldt's problem and as (a special case of) Plato–Russell's problem without even trying to tackle Descartes's problem. In other words, the full understanding of the nature and ontogeny of our system of knowledge of language appears to be partially dependent on a better understanding of the (creative) use of that knowledge. This is, no doubt, the case for anaphora and binding too.

See also: Control, Syntactic; Linguistics: Overview; Neural Basis of Language Development; Pragmatics, Linguistic; Semantics; Syntax–Semantics Interface.

Bibliography

Chiou, M., Huang, Y., 2010. NP-anaphora in modern Greek: a partial neo-Gricean pragmatic approach. Journal of Pragmatics 42, 2036–2057.
Chomsky, N., 1981. Lectures on Government and Binding. Foris, Dordrecht, The Netherlands.
Chomsky, N., 1995. The Minimalist Program. MIT Press, Cambridge, MA.
Clements, G.N., 1975. The logophoric pronoun in Ewe: its role in discourse. Journal of West African Languages 2, 141–177.
Hagège, C., 1974. Les pronoms logophoriques. Bulletin de la Société de Linguistique de Paris 69, 287–310.
Hicks, G., 2009. The Derivation of Anaphoric Relations. John Benjamins, Amsterdam.
Huang, Y., 1991. A neo-Gricean pragmatic theory of anaphora. Journal of Linguistics 27, 301–335.
Huang, Y., 1994/2007. The Syntax and Pragmatics of Anaphora. Cambridge University Press, Cambridge, UK.
Huang, Y., 2000a. Anaphora: A Cross-Linguistic Study. Oxford University Press, Oxford, UK.
Huang, Y., 2000b. Discourse anaphora: four theoretical models. Journal of Pragmatics 32, 151–176.
Huang, Y., 2002. Logophoric marking in East Asian languages. In: Güldemann, T., von Roncador, M. (Eds.), Reported Discourse: A Meeting Ground for Different Linguistic Domains. John Benjamins, Amsterdam, pp. 213–224.
Huang, Y., 2007. Pragmatics. Oxford University Press, Oxford, UK.
Huang, Y., 2010a. Neo-Gricean pragmatic theory of conversational implicature. In: Heine, B., Narrog, H. (Eds.), The Oxford Handbook of Linguistic Analysis. Oxford University Press, Oxford, pp. 607–631.
Huang, Y., 2010b. The pragmatics of anaphora. In: Cummings, L. (Ed.), The Pragmatics Encyclopedia. Routledge, London, UK, pp. 9–13.
Huang, Y., 2010c. Switch-reference in Amele and logophoric verbal suffix in Gokana: a generalized neo-Gricean pragmatic analysis. In: Shu, D.F., Turner, K. (Eds.), Contrasting Meaning in Languages of the East and West. Peter Lang, Berlin, pp. 75–101.
Levinson, S.C., 1987. Pragmatics and the grammar of anaphora. Journal of Linguistics 23, 379–434.
Levinson, S.C., 2000. Presumptive Meanings: The Theory of Generalized Conversational Implicature. MIT Press, Cambridge, MA.
Reinhart, T., Reuland, E., 1993. Reflexivity. Linguistic Inquiry 24, 657–720.
Reuland, E., 2011. Anaphora and Language Design. MIT Press, Cambridge, MA.
Rooryck, J., Wyngaerd, V., 2011. Dissolving Binding Theory. Oxford University Press, Oxford, UK.
Safir, K., 2004. The Syntax of Anaphora. Oxford University Press, London, UK.

Anarchism

Maple Razsa, Colby College, Waterville, ME, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Anthropology's relationship with anarchism is complex and multivalent because anarchism is at once (1) a revolutionary tradition that sometimes draws on ethnographic research in developing its critique of contemporary society; (2) a theoretical orientation that has played a significant, if largely unacknowledged, role in the development of anthropology; (3) a form of social organization characteristic of societies without a state, including a great number studied by ethnographers and archaeologists; and (4) more recently, a set of antiauthoritarian principles and practices that have been integral to global social movements that have themselves become objects of ethnographic study.

Introduction

The sustained – if largely unacknowledged – relationship between anthropology and anarchism is not easily characterized. Peter Kropotkin, perhaps the most important anarchist thinker, personifies the complexity and multivalence of this relationship. Born a Russian nobleman, Kropotkin renounced his title and became a committed revolutionary, living much of his life in exile. Writing for the 1910 Encyclopedia Britannica, he provides one of the more enduring definitions of anarchism:

> (from the Greek, ἀν, and ἀρχή, contrary to authority) principle or theory of life and conduct under which society is conceived without government – harmony in such a society being obtained, not by submission to law, or by obedience to any authority, but by free agreements concluded between the various groups, territorial and professional, freely constituted for the sake of production and consumption as also for the satisfaction of the infinite variety of needs and aspirations of a civilized being. (p. 914)

In developing this understanding of anarchism, Kropotkin drew on his own zoological, geographic, and ethnographic observations in the Russian Far East. In challenging the dominant social Darwinism of his day – especially the emphasis on competition (e.g., 'survival of the fittest'), which Kropotkin felt reinforced support for the capitalist system – he stressed the crucial role of mutual aid in survival (1902). Since he understood anarchism not only as a potential future order but also as a long-standing human tendency, he drew on examples of existing and historic stateless societies – especially from the emerging discipline of anthropology – in elaborating his theory of a social order without state authority. To further complicate the relationship between these traditions, during Kropotkin's exile in England, he came to have a strong influence on Alfred Radcliffe-Brown, the founder of modern British anthropology (Perry, 1975). In fact, the influence of anarchist ideas, if rarely appreciated, is a thread woven through the history of anthropology.

Kropotkin thus provides an initial glimpse into anarchism and its varied intersections with anthropology. Anarchism is at once (1) a revolutionary tradition that sometimes draws on ethnographic research in developing its critique of contemporary society; (2) a theoretical orientation that has played a significant, if largely subterranean, role in the development of anthropology; (3) a form of social organization characteristic of societies without a state, including a great number studied by ethnographers and archaeologists; and (4) more recently, a set of antiauthoritarian principles and practices that have been integral to global social movements that themselves have become objects of ethnographic study. To date, there is a complex interaction between anthropology and these different facets of anarchism. Before examining this interaction in more detail, it is necessary to have in mind the history and key principles of anarchism as a political movement.

The Anarchist Revolutionary Tradition

Classical Anarchism

Many readers will associate anarchy – the lack of authority, especially state authority – with disorder, violence, and chaos – that is, either with the breakdown of established order or with an era preceding the emergence of the state, what Thomas Hobbes called the "natural condition of mankind." Second, some will call to mind bomb-wielding terrorists bent on bringing about just such a collapse of ordered society. These apocalyptic images are not surprising; 'anarchist' and 'anarchy' were deployed as epithets – including against the most radical and antiauthoritarian elements of the English Civil War and French Revolution – long before Pierre-Joseph Proudhon affirmed them as a self-description in *What Is Property?* (1840). Although earlier movements embraced similar political goals and practices – including the seventeenth-century English 'Diggers' – a broad-based and self-consciously anarchist movement first developed within the International Workingmen's Association (The First International, 1864–72). During these years, anarchists consolidated as a distinct bloc within the broader workers' movement. Although they largely shared with communists and socialists the goal of socialism, anarchists differed sharply over how to achieve radical political transformation. The storied debates between Marx and Bakunin crystallized these differences – differences that continue to define these distinct radical traditions – and foretold the schism that would end the First International. First, Bakunin objected to Marxists' revolutionary means: seizing state power and using that power to direct radical social change. Kropotkin

summarized anarchists' (prescient) fears about the Marxist approach when he wrote that to hand over

> to the state all the main sources of economical life – the land, the mines, the railways, banking, insurance, and so on – as also the management of all the main branches of industry, in addition to all the functions already accumulated in its hands (education, state-supported religions, defense of territory, etc.), would mean to create a new instrument of tyranny. State capitalism would only increase the powers of bureaucracy and capitalism. (1910: p. 914)

Second, anarchists challenged the preeminence of the proletariat in Marxist analysis, which posited that only the industrial working class – at once utterly dispossessed and absolutely integral to capitalism – was positioned to act as the subject of revolutionary activity. Bakunin, in contrast, asserted that outrage against the contemporary order came from many quarters, and the movement also should embrace the most excluded, including the peasants and the *lumpenproletariat* (the unemployed, common criminals, slum dwellers, scavengers, etc.), the very populations that Marx dismissed as secondary, or worse, counterrevolutionary (Avrich, 1988: pp. 7–13). The perceptiveness of his critiques aside, Bakunin was himself a veteran conspirator, often acting in a manner fundamentally at odds with antiauthoritarian principles. What is more, although these key anarchist figures – and their theoretical differences with Marx – are useful signposts in narrating the history of the anarchist movement, and distinguishing it from its primary revolutionary rival, one should not conflate anarchist movements with their most prominent authors and activists. Nearly all Bakunin's ideas, for example, were drawn from the practices of the anarchist struggles of their day (Marshall, 2010) and these anarchist thinkers were primarily pamphleteers and moralists rather than theoretical philosophers (Graeber, 2008: p. 211).

Although these strategic differences are significant in themselves, they also point to a fundamental difference between Marxist and anarchist approaches to politics. There were (and continue to be) a great variety of anarchist factions – including Mutualists, Anarcho-Syndicalists, Libertarian Communists, Anarcho-Pacifists, Individualists, Anarcho-Naturists, Green Anarchists, Insurrectionists, and Primitivists – but most shared an attitude toward radical change that is better characterized as a commitment to certain principles rather than a specific utopian goal, a commitment to means rather than ends. Anthropologists and historians of anarchism generally have identified these core principles as including autonomy, decentralized organization, direct democracy, antiauthoritarianism, and mutual aid. In other words, rather than being familiar with canonical texts, or acting as acolytes of key figures, as has often been the case for Marxists (e.g., Leninists, Maoists, Trotskyites), anarchists have exhibited "an adherence to a set of principles that guides much of anarchism and provides connections among its strains" (Angelbeck and Grier, 2012: p. 551).

Spain and the Violent Demise of Classical Anarchism

Popular and broad-based anarchist organizing reached its apotheosis in 1930s Spain – and this period is often cited by anarchists for the glimpse it provides of an anarchist order in an industrialized society. Already better organized and more numerous than movements elsewhere in Europe, Spanish anarchism blossomed during the Civil War – in fraught alliance with other social forces resisting General Franco's fascist rebellion against the republic. In the countryside, especially impoverished Andalusia, peasants embraced insurrectionary anarchism, seizing control of their pueblos to establish autonomous communes, destroy the power of the church and redistribute the land. By 1937 the anarchist labor federation, the Confederación Nacional del Trabajo, had more than 2 million members. In the industrialized Northeast, especially around Barcelona, these anarcho-syndicalists seized control of the factories and ran much of the public infrastructure. In May of 1937, however, armed conflict broke out between anarchists and communists, who were backed by Stalin. By January of 1939 the anarchists were crushed and Franco's forces entered Barcelona.

Beyond armed defeat and violent repression – not only in Spain but elsewhere in fascist Europe as well as the United States – two major causes of anarchism's decline as a significant political force stand out. First, the Bolsheviks, emerging victorious from the Russian Revolution, established the highly disciplined, hierarchical, and centralized party as the preeminent model of twentieth-century militant organization. Second, despite the pacifism of many of its key thinkers, including Godwin, Proudhon, Kropotkin, and Tolstoy, anarchism had come to be tainted by the acts of the minority of adherents who championed the 'propaganda of the deed' – and who had successfully assassinated the heads of state of Spain, France, and the United States in an effort to spark a broader social revolution.

Anarchism as Theory

> It is said that the histories of peoples who have a history is the history of class struggle. It might also be said, with at least as much truthfulness, that the history of peoples without history is the history of their struggle against the State. (Clastres, 1987: p. 218)

Stateless Societies and Anarchist Theory

Although the anarchism of the international movements of the nineteenth and early twentieth centuries cannot, of course, be conflated with modes of life of communities living without states – those known as primitive peoples in early anthropology – there has been a dynamic interplay between the antiauthoritarian ideas and practices developed by revolutionary anarchism and the anthropological study of stateless societies. Radcliffe-Brown, for example – a figure anthropologists more likely think of in relation to colonial anthropology – was known as 'Anarchy Brown' during his student days at Cambridge (Stanner, 1968: p. 287). By the time he was established as the founder of political anthropology, however, he had recanted anarchism and espoused a decidedly more moderate politics. As Perry argued, anarchist ideas – Kropotkin's in particular – nevertheless profoundly influenced Radcliffe-Brown and helped to set the tone of British social

anthropology during the first half of the twentieth century (1975: p. 61). Radcliffe-Brown's emphasis on mechanisms for maintaining order and social harmony in stateless communities – especially community centered sanctions, like avoidance or the withdrawal of reciprocity, rather than law – reflected the theoretical influence of anarchist ideas (Goody, 2000: p. 5). In any case, Radcliffe-Brown's anarchist tendency was not simply a personal quirk. His intellectual rival, Evans-Pritchard, more commonly identified with a conservative and Catholic worldview, wrote that "The Nuer constitution is highly individualistic and libertarian. It is an acephalous state, lacking legislative, judicial and executive organs. Nevertheless, it is far from chaotic. It has a persistent and coherent form which might be called 'ordered anarchy'" (1940: p. 296). In part, the anarchist tendency in anthropological theorizing of this period can be seen as part of an agenda – one they shared with Kropotkin – to challenge social Darwinists' emphasis on competition. Malinowski, for example, in a rare acknowledgment of anarchist theory, noted that "Prince Peter Kropotkin was quite right in pointing out that mutual aid between individuals of a cooperative community is the dominant concept, while the struggle between the individuals for survival can not be applied to human societies as a whole" (1944: p. 143).

If anthropologists of this generation, the first to carry out field research, were keen to challenge social Darwinism, and its associated belief in evolutionary hierarchy – the theories dominant among the 'arm-chair' anthropologists who preceded them – they were nonetheless deeply committed to developmentalism. In other words, if for this generation Aborigines, the Nuer, and Trobriand Islanders were equally human, at least in biological terms, their social systems were viewed as more simple – in a word, primitive – than 'complex' modern societies. This developmentalism later would be systematized in Elman Service's classification of sociopolitical stages, which ran in ascending order of increasing population, specialization, and elaboration from band through tribe and chiefdom to the state (1963).

With his avowedly anarchist theorizing, Pierre Clastres offers a fundamental challenge to this implicit – and persistent – developmentalism within anthropology. Drawing on his fieldwork with the Guyaki (Aché) of Paraguay, Clastres (1987 [1974]) questioned the assumption that primitive societies *lack* a state and as such are archaic remnants of an earlier epoch of human development that – given sufficient time – eventually would generate a state. On the contrary, as seen in the epigraph above, Clastres argued that it was only due to the active effort of stateless peoples that hierarchy and inequality were prevented from emerging. To put it another way, if humans are to preserve their autonomy, they must collectively create, cultivate, and maintain a range of practices that actively avert, ward off, and refuse the emergence of authority. In a similar vein, Clastres questioned dominant ideas of 'subsistence,' which implied that hunter-gatherers were incapable of producing a surplus and therefore were living desperate, hand-to-mouth existences (see also Sahlins, 1972). Clastres' interventions have far-reaching implications for political theory, especially for some of the ways theorists have imagined human existence before statehood. By positing active, and at times self-conscious, hunter-gatherer resistance to hierarchy, he countered the Rousseauian notion of man's natural state of innocence (the myth of the noble savage); tribal peoples, far from innocent, are painfully aware of the dangers of, and strive to prevent, the emergence of political authority. Similarly, by arguing that hunter-gatherers lead lives of plenty – yet studiously avoid producing material surplus for fear that the concentration of surplus value would institutionalize inequality – Clastres refuted Hobbes' famous claim that life without a state is one of "continual fear and danger of violent death, and the life of man, solitary, poor, nasty, brutish, and short" (1651: p. XIII.9). As Harold Barclay, another anthropologist with open anarchist sympathies, summarized the matter in *People Without Government*, "I believe many anthropologists, in their own projection of personal and cultural values, have obstinately refused to apply the one truly clarifying term to these numerous societies which are without government and are, therefore, anarchies" (1982: p. 18). Although anarchist theory has been useful in challenging the implied lack or absence imputed to nonstate social entities, this tendency is difficult to avoid; indeed, it persists in the term anarchism itself.

Recently, a number of anthropologists and archaeologists, in the tradition of Clastres and Barclay, have more openly championed the potential and actual contribution of anarchist theory to better understanding existing and historical stateless communities. In *The Art of Not Being Governed: An Anarchist History of Upland Southeast Asia*, for example, James Scott offered a provocative rereading of the ethnographic and historic record of Zomia, a vast highland region stretching from Vietnam to North India (2009). He argued that peoples like the Mien in China and Hmong across Laos, Vietnam, and Cambodia, far from representing the remnants of prestate societies, embody creative responses to the formation of states in the agricultural lowlands. Rather than existing in a not-yet-state condition, they actively fled the consequences of state formation: slavery, conscription, taxes, forced labor, epidemics, and warfare. What is more, Scott argued, they developed a range of creative strategies to maintain their autonomy and resist the rise of centralized authority within their own communities. Similarly, in 'Anarchism and the Archaeology of Anarchic Societies,' Bill Angelbeck and Colin Grier (2012) focus on long-standing questions related to the unusual political structure of the Coast Salish region of the Pacific Northwest. They argued that anarchism makes possible fresh and more revealing interpretations of decentralized social networks, especially those that allow for complex intergroup relations while staving off the establishment of centralized political authority. As they concluded, "Simply put, societies without governments are anarchies. Given that, we propose that the rich intellectual tradition of anarchist theory and practice has something to offer those studying the material record of those anarchic societies" (Angelbeck and Grier, 2012: p. 568). Charles MacDonald, teasing out the epistemological implications of his decades of intermittent fieldwork with the anarchic Palawan of the Southern Philippines, argued that anthropologists must be more discerning in labeling a given community anarchist. He pointed out that some communities that are not characterized by state authority are anything but anarchist in their attitudes toward inequality. He nevertheless echoed other anthropologists of anarchist societies when he concluded that "Equality is constructed, an active process enforced by various means and by constant

effort. Egalitarianists do not just abstain from inequality, like they would from consuming alcohol or wearing T-shirts; they produce and actively maintain equality" (Macdonald, 2012: p. 59).

The relatively minor yet persistent anthropological engagement with anarchist theory, especially in relation to the tradition of studying 'stateless peoples,' can be explained by the unique insights it generates – and the ways it militates against ethnocentrism. Here, we can see a parallel to the way anarchist revolutionary movements have served as an important corrective against the reductionist tendencies – and ends over means orientation – that have plagued Marxist movements. Analogously, anarchist theory serves as an important corrective to the persistent developmentalism of anthropological theorizing.

Anarchist Prospecting in the Ethnographic Record

Before turning to the ethnography of anarchist social movements, it is important to note two other anarchist theoretical developments of interest for anthropologists. First, anarchists have continued to show interest in anthropology, much as Kropotkin did earlier, in part due to the very anthropological tradition of studying and theorizing social order among stateless people outlined previously. There is obviously a chasm separating anthropology as professionalized academic practice from anarchism as a revolutionary one, but the specific nature of how and why anarchists engage anthropology perhaps indexes some underlying affinities between the two traditions. Indeed, it is difficult to think of any engagement with anthropology by nonanthropologists that more closely approximates the core principles of anthropology, especially in its more critical and publicly engaged forms. In their survey of anarchist prospecting in the ethnographic record, Andrew Robinson and Simon Tormey (2012) outlined at least four main modes of engagement. First, anarchists draw on anthropology to generate critique, that is, to develop tools with which to denaturalize the present by revealing alternative forms of social organization. Second, they seek techniques for sustaining stateless relations as found among those peoples anthropologists have studied. In this, they rely precisely on the affirmative reversal of statelessness-as-a-lack evidenced in the above history of anthropological theorizing of nonstate social entities. Third, they draw on anthropology to develop reflexivity, to help recognize the contingency of their own beliefs, which makes possible dialogue with communities that have different ontologies and epistemologies. Finally, activists draw on ethnography to cultivate solidarity, especially with indigenous peoples, whom they see as embodying models of struggle for autonomy (Robinson and Simon, 2012).

Anarchist Critical Theory?

The second anarchist theoretical trend is of anthropological interest because it concerns those social theorists who have been among the most influential in the recent history of the discipline. Although anarchism has not been associated with high theory, especially in comparison with Marxism, there have been recent efforts to develop a body of critical theory informed by anarchism. Often under the heading of post-anarchism – the post implying not so much the overcoming of the anarchist tradition, but its resonance with poststructural and postmodern theorizing – this literature (Rousselle and Evren, 2011) has noted significant affinities between anarchism and the thought of Nietzsche, Lacan, Derrida, Foucault, Deleuze, and Rancière, among others. At times, this work has drawn on these theorists to critique reductionist or essentialist strains within anarchism, such as, for example, the Manichaean tendency within some strains of the anarchist movement to see the state as a mortal enemy rather than a social relationship. Defenders of anarchism are quick to point out, however, that there have always been critics internal to the anarchist tradition who have resisted these tendencies. In any case, postanarchism draws connections between anarchism and those (mainly French) theorists who are highly influential in contemporary anthropology. It is possible to draw such links, Morris argued, because Deleuze and Foucault, for example, have implicitly drawn on the anarchist tradition in developing their ideas (2005: p. 12). To further complicate any clear genealogy, Deleuze and Guattari extensively deployed Clastres' concepts, especially in developing their ideas of a nomadic war machine (Deleuze and Guattari, 1987: pp. 357–360). Finally there is a robust tradition of autonomist Marxist theory – nonparty and anti-authoritarian in orientation, largely popularized in the English-speaking world by Hardt and Negri – that is deeply resonant with anarchism and sometimes is drawn on in analyzing recent movements that are highly influenced by anarchism (Razsa and Andrej, 2012). This recent theoretical production is viewed skeptically in many anarchist quarters, however, confirming Graeber's claim that while "Marxism has tended to be a theoretical and analytic discourse … anarchism is an ethical discourse about revolutionary practice" (2009: p. 211). What is more, it is plausible that the recent 'anarchist turn' in critical theory has as its condition of possibility the dramatic rise of a new wave of anarchist inflected social movements over the past few decades.

The Ethnography of Anarchist Movements (and Actually Existing Anarchy)

Colored perhaps by an inherent antipathy between academic and anarchist practice, scholars often have been dismissive of anarchist-inspired revolutionary movements. Graeber attributed this scholarly hostility to the fact that anarchism does not have the same affinity with the academy that Marxism does. First, anarchism is critical of the notion of a revolutionary vanguard, a role that Marxist scholars often have felt they could play. What is more, anarchism – as primarily an ethical discourse about revolutionary practice – has fewer scientific pretensions than Marxism, which was developed as a theoretical and analytic discourse (Graeber, 2004). Historical research on Republican Spain is symptomatic of scholarly antipathy toward anarchism.

The Anthropology of Spanish Anarchism

Early historians of the Spanish Civil War viewed anarchism as highly moralistic in character, reflecting a millenarian fanaticism that, they argued, owed much to the Catholic

faith it condemned (Brenan, 1950). In *Primitive Rebels*, Eric Hobsbawm elaborated on this position, describing anarchism as "a form of peasant movement almost incapable of effective adaptation to modern conditions" (1965: p. 92). In short, for Hobsbawm as a lifelong communist, anarchism was the embodiment of the 'primitivism' and 'ritualism' typical of social movements lacking modern forms of party organization and discipline. Anthropologists conducting research in Spain, especially in communities deeply affected by anarchist uprisings, were the first to carry out field research on these movements. Evans-Pritchard's student, Pitt-Rivers, for example, worked in a pueblo nationally renowned as the rural 'cradle of anarchism.' On the basis of fieldwork conducted during Franco's reign, however, *People of the Sierra* (1954) is circumspect about the town's anarchist legacy, and when it does refer to the movement, it does little to challenge the dominant portrait of rural anarchism as archaic and millenarian. Jerome Mintz's *The Anarchists of Casas Viejas*, the first anthropological account of anarchists in their own words, humanizes these radicals in a way earlier studies do not (1982). Mintz made clear that scholars' belief in evolutionary hierarchies preemptively discredits anarchism – not only when it is practiced by Fourth World peoples but also when it is adopted by twentieth-century movements in Western Europe. He noted that,

> although Eric Hobsbawm wrote that he visited Casas Viejas in 1956 and spoke with the townspeople, his account is based primarily on a preconceived evolutionary model of political development rather than on data gathered in field research. The model scales labor movements in accord with their progress toward mass parties and central authority … Unfortunately his evolutionary model misled him on virtually every point. (Mintz, 1982: p. 271)

Despite Mintz's important intervention – and even though the revealing testimonies he gathered indicates he developed relationships of great trust with his informants – he also conducted research during the waning years of Franco's rule. One wonders what affect this had, if any, on anarchists' nearly universal recanting of their beliefs in the course of interviews with him. Maddox (1995), who also conducted research in Andalusia, provided a rich contextualization for anarchists' committed anticlericalism, focusing on the stripping and dynamiting of the historic churches during another rural uprising. Despite the subtlety of his analysis, Maddox's Gramscian approach produced some grating theoretical dissonance. Framing anarchist actions in terms of hegemony and counterhegemony is incongruous with movements that probably are better understood as fundamentally antihegemonic. Finally, these anthropologists of Spanish anarchism primarily focus on ideology – as reported in either hostile sources or recollected in different political circumstances.

The Ethnography of the Alterglobalization Movement

Within a few years of these debates about historical anarchism – and in the wake of the collapse of state socialism from 1989 to 1993 – anarchism would once again come to have a powerful grip on the leftist imagination, arguably usurping the place that Marxism held for earlier generations of radicals (Epstein, 2001). Indeed, anarchist-inflected movements soon would burst disruptively onto the international stage – and anthropologists would have the opportunity to study not just anarchist ideas but anarchist daily practices. The streams of antiauthoritarian practice that merged to precipitate this anarchist turn among social movements in the late 1990s were varied – including feminist, antinuclear, peace, Quaker, environmentalist, and autonomist tributaries (Maeckelbergh, 2009: pp. 13–21). In light of the history of the anthropology of anarchism and 'stateless peoples' outlined previously, however, it is telling that most participants of what came to be known as the global justice or alterglobalization movement – and which the media often insisted on calling antiglobalization – traced the origins of their movement to the 1994 indigenous Zapatista uprising in Chiapas, Mexico. The Zapatistas embodied an antiauthoritarian politics that drew, in part, on indigenous practices of self-rule – including local autonomy, *mandar obediciendo* (leading by following), and consensus decision making. As such they marked a departure from the statist orientation of most twentieth-century movements, which, if they were reformist, sought to pressure the state to implement changes on their behalf and, if they were revolutionary, sought to seize the state and use it as a means to implement radical change. What is more, the first global network of movements, People's Global Action (PGA), which facilitated the convergence of various strands of activism, developed from the Zapatista-initiated Second Intercontinental Encounter for Humanity and Against Neoliberalism. The PGA, in turn, organized the Global Action Days, most notably the 1999 Seattle protests against the World Trade Organization that brought the movement to broad public attention.

The wave of major protest mobilizations that became the most recognizable face of the movement – in Washington, Prague, Quebec, Genoa, Quebec, and Miami, among other cities – and the global networking that made them possible provided the context for a generation of anthropologists to conduct fieldwork with movements that seemed to embody the core principles of anarchism. While there were countless links between these mobilizations and movements in the Global South – including important founding roles within the PGA for the Brazilian Landless Farmers' Movement in Brazil and the Indian Karnataka State Farmers' Association, in addition to the Zapatistas – anthropologists studying this wave of direct action movements did so largely in North America and Europe (Juris, 2008; Graeber, 2008; Maeckelbergh, 2009; Razsa and Velez, 2010).

A number of scholarly and political trends made the ethnographic study of these movements especially attractive at this time, including disciplinary interest in globalization, resistance, and activist anthropology. Anthropologists argued, moreover, that the decentralized, practice-centered nature of these movements seemed particularly suited to ethnographic approaches – even as it posed new challenges. As Graeber puts it,

> If anarchism is not an attempt to put a certain sort of theoretical vision into practice, but is instead a constant mutual exchange between inspirational visions, antiauthoritarian attitudes, and egalitarian practices, it's easy to see how ethnography could become such an appropriate tool for its analysis … to tease out the implicit logic in a way of life, along with its related myths and rituals, to grasp the sense of a set of practices. (2008: pp. 221–222)

Or as Maeckelbergh asserted, "taking an ethnographic approach to the movement allows for the exploration of the actual decision-making practices already in place, which make visible the beginnings of an emerging democratic alternative" (2009: p. 5).

While ethnographic methods were well suited to the study of these everyday practices, it proved difficult to delimit one's object of study when the density and frequency of links in these decentralized networks could lead one almost anywhere – after all, the PGA's slogan was "we are everywhere!" As Maeckelbergh asked, echoing broader methodological challenges in studying globalization, "How can we do ethnography of a global network that is not clearly delimited in time and space and has not singular identity, adversary or goal?" (2009: p. 22). Some ethnographers responded to this methodological dilemma by situating themselves within particular activist nodes and following activists from these nodes as they engaged the broader network (Juris, 2008; Razsa, 2008), while others took the process of creating a globally networked movement itself as their object of study (Maeckelbergh, 2009). Beyond methodological dilemmas, ethnographic writing on these movements posed its own challenges, some of them more politicized instances of the representational crisis precipitated by the writing culture debates of the 1980s. As Graeber expressed it, "to employ standard narrative techniques and allow some individuals to typify others would be to employ exactly the logic of representation that the activist decision-making structures I was trying to describe were trying hardest to avoid" (2008: p. 14). Although nearly all anthropologists of alterglobalization practiced some form of engaged, activist, or militant ethnography, this did not mean they shied away from describing the limits, conflicts, and contradictions of this movement's attempts to organize in antiauthoritarian and decentralized ways, reflecting perhaps the vigorous, even ruthless, internal debates and inbuilt practices of critical reflection among participants.

Indeed, a common characteristic of ethnographies of this "movement of movements" has been a focus on what Maeckelbergh calls its "unglamorous side" (2009: p. 3), namely, the organizational and decision-making practices that made possible the dramatic protest actions and mass mobilizations that were the public face of the movement. Juris's *Networking Futures*, for example, details the concrete practices "through which decentered networking logics are produced, reproduced, and transformed within particular social, cultural and political contexts" (2008: p. 11). The forms these took varied, but they included affinity groups, spokes, and spokescouncils, which could be traced to the Spanish Civil War, although refracted through other struggles, especially the antinuclear and peace movements of the 1980s. Activists and ethnographers alike emphasized the ways that directly democratic decision making is prefigurative – that is, it models more egalitarian, participatory, and inclusive modes of self-rule here and now.

These claims to prefigure change – along with other key principles of the movement – including direct action, consensus, horizontality, and diversity – would seem to echo classic anarchist ideas, especially the central method of "building the new world in the shell of the old." Graeber makes the starkest claim for the anarchist character of the movement overall, calling anarchists the "heart of the movement, its soul; the source of most of what's new and hopeful about it" (2002: p. 62). In fact, there is some debate about whether these movements should be considered anarchist. Juris (2008) argued, for example, that even as principles previously associated with anarchism have become preeminent in the movement, many activists do not identify with the label, in part because of the multiple sources of antiauthoritarianism – including, in addition to anarchism, Zapatismo, autonomous Marxism, and ecological movements. What is more, ironically given that Juris was based in the historic anarchist stronghold of Barcelona, many Catalan activists resisted being pigeonholed as anarchists, especially because some anarchist formations were relatively sectarian. Razsa, though he conducted fieldwork primarily with self-declared anarchists in the former Yugoslavia, nonetheless found that it was at those moments when anarchism was expressed most clearly as a political and (sub)cultural identity, rather than a set of principles, that it became an obstacle to the practical implementation of those principles (Razsa, 2008). Finally, Juris highlighted that the emphasis on networking as a *norm* was deeply resonant with anarchist ethics, but this ethic dovetailed with broader cultural and social shifts associated with late capitalism and associated modes of networked communication and production, and themselves were not necessarily radical in nature. This convergence, especially with web-based communications and social media, may indicate why practices that are deeply resonant with anarchist principles seized the radical imagination at this time. They also help to explain why geographically dispersed movements have found nonhierarchical ways to coordinate.

Global Uprisings

Although the scale and intensity of major mobilizations associated with the alterglobalization movement declined during the mid- to late 2000s – especially in the United States following the terrorist attacks of 11 September 2001 – a new cycle of protest exploded through 2011 and 2012. Beginning first in Tunisia before spreading to Egypt and much of the Middle East, the Arab Spring inspired unrest elsewhere. The movement of the *indignados* took up the torch in squares across Spain, kindling US activists' desire to Occupy Wall Street and create protest encampments in hundreds of US cities. Israel, Greece, and Chile also saw major protest movements. If the targets of these protests, what they opposed, varied – corruption, dictatorship, austerity, and wealth inequality – their practices almost universally have involved the embrace of direct democracy. Many analysts therefore have seen these movements as responses to a fundamental crisis of representative politics and as embodying a radical alternative to liberal democracy. Interestingly, those anthropologists who researched the alterglobalization movement were at the forefront of research on these new movements, perhaps because they were well positioned to tease out the continuities and discontinuities between these protests and earlier struggles. They were joined by a new generation of anthropologists who, like the previous one, were swept up in these dynamic movements as participants as much as researchers (Juris and Razsa, 2012). Many of the debates that marked the alterglobalization movement emerged anew around the

dynamics of inclusion and exclusion, the quality of democratic practice, prefiguration versus efficiency, the legitimacy of forming demands or not, and perennial questions of property destruction and violence. If the globalization movement wove together various strands of existing activism, developing innovative ways of coordinating different constituencies in decentralized and nonhierarchical ways, these new protest movements, in part because of the deep financial crisis since 2008, brought in a much wider cross section of people, many with no previous experience with direct democracy and direct action. Although this lead to tensions around 'process' it also offered a new wave of participants a personal experience of collective action and antiauthoritarian decision making.

Actually Existing Anarchism

Although the ethnography of anarchism primarily has been concerned with stateless communities on the one hand, or radical social movements on the other, there is an emerging interest in what might be described as actually existing anarchism. In many ways, this notion returns us, once again, to Kropotkin who insisted that anarchism was not only a model for a better future, or a condition of some Edenic past, but a persistent tendency in all social relations. Colin Ward, the British housing and education activist, pioneered this kind of anarchist research and action, keenly discerning the ways that anarchist principles are already operative in the cracks and fissures of an otherwise hierarchical British society (1973). Graeber even argued that mutuality is the foundation of all sociability, because social relations are always undergirded by the assumption that anyone who is not actually an enemy can be expected to engage others – at least to an extent – on the principle of "from each according to their abilities" (2011: p. 96). In a related movement toward identifying extant forms of autonomous organization, radical migrant activists and theorists have developed an understanding of the 'autonomy of migration,' which looks at the already existing practices that migrant communities use to 'manage the border from below,' asking how these practices might be cultivated and expanded as a basis for expanding migrant power.

Conclusion

The dynamic and multivalent relationship between anthropology and anarchism has enriched both traditions. From the standpoint of anthropology, anarchism in its various guises – as militant tradition, theoretical provocation, prefigurative practice, or mode of living – productively has disrupted the discipline's assumptions, especially regarding teleology, temporality, and the state. Anarchism contributed to the dislodging of first evolutionist and then developmentalist assumptions – and with them certain strains of ethnocentrism. Anarchism – with its slogan 'another world is possible' – also reaffirms the fundamentally anthropological disposition that a great variety of social orders and ways of being human are possible; politics, therefore, always can be radically reimagined. In this spirit, anarchism may continue to catalyze a new political anthropology, even a return to anthropology at its best: The exploration of ways of being human that are at odds with what appears natural and inevitable from the vantage point of the present.

See also: Analytical Marxism; Common Sense, Anthropology of; Cultural Critique, Anthropological; Ethnocentrism; Globalization, Anthropology of; Hegemony and Cultural Resistance; Hegemony: Anthropological Aspects; Ideology, Sociology of; Primitive Society; Social Movements and Political Violence.

Bibliography

Angelbeck, Bill, Grier, Colin, 2012. Anarchism and the archaeology of anarchic societies: resistance to centralization in the Coast Salish region of the Pacific Northwest Coast. Current Anthropology 53 (5), 547–587.
Avrich, Paul, 1988. Anarchist Portraits. Princeton University Press, Princeton, NJ.
Barclay, Harold B., 1982. People without Government. Kahn & Averill, London.
Brenan, Gerald, 1950. The Spanish Labyrinth: An Account of the Social and Political Background of the Civil War. Cambridge University Press, Cambridge, England.
Clastres, Pierre, 1987. Society against the State: Essays in Political Anthropology. Zone Books, New York.
Deleuze, Gilles, Guattari, Félix, 1987. A Thousand Plateaus: Capitalism and Schizophrenia. University of Minnesota Press, Minneapolis, MN.
Epstein, Barbara, 2001. Anarchism and the anti-globalization movement. Monthly Review 53 (4), 1.
Evans-Pritchard, Edward E., 1940. The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People. Clarendon press, Oxford.
Goody, Jack, 2000. Anarchy Brown. Cambridge Anthropology 21 (3), 1–8.
Graeber, David, January–February 2002. The new anarchists. New Left Review 13, 61–73.
Graeber, David, 2004. Fragments of an Anarchist Anthropology, vol. 14. Prickly Paradigm Press, Chicago.
Graeber, David, 2008. Direct Action: An Ethnography. AK Press, Oakland, CA.
Graeber, David, 2011. Debt: The First 5,000 Years. Melville House, Brooklyn, NY.
Hobbes, Thomas, 1651. Leviathan, or, the Matter, Forme, & Power of a Commonwealth Ecclesiasticall and Civill. Printed for Andrew Ckooke i.e. Crooke, at the Green Dragon in St. Pauls Church-yard, London.
Hobsbawm, Eric J., Paul, Avrich, 1965. Primitive Rebels: Studies in Archaic Forms of Social Movement in the 19th and 20th Centuries. W.W. Norton, New York.
Juris, Jeffrey S., 2008. Networking Futures: The Movements against Corporate Globalization. Duke University Press, Durham, NC.
Juris, Jeffrey S., Razsa, Maple, 2012. Introduction: Occupy, Anthropology, and the 2011 Global Uprisings. http://www.culanth.org/?q=node/641.
Kropotkin, Peter A., 1902. Mutual Aid: A Factor of Evolution. McClure, Phillips & co, New York.
Macdonald, Charles, 2012. The anthropology of anarchy. Indian Journal of Human Development 6 (1), 49–66.
Maddox, Richard, 1995. Revolutionary anticlericalism and hegemonic processes in an Andalusian town, August 1936. American Ethnologist 22 (1), 125–143.
Maeckelbergh, Marianne, 2009. The Will of the Many: How the Alterglobalisation Movement Is Changing the Face of Democracy. Pluto Press, London.
Malinowski, Bronislaw, 1944. A Scientific Theory of Culture. The University of North Carolina Press, Chapel Hill, NC.
Marshall, Peter, 2010. Demanding the Impossible: A History of Anarchism. PM Press, Oakland, CA.
Mintz, Jerome R., 1982. The Anarchists of Casas Viejas. University of Chicago Press, Chicago, IL.
Morris, Brian, 2005. Anarchism and anthropology: their elective affinity. Goldsmiths Anthropology Research Papers 11, 1–16.
Perry, Richard J., 1975. Radcliffe-Brown and Kropotkin: the heritage of anarchism in British social anthropology. Kroeber Anthropological Society Papers 51-52, 61–65.
Pitt-Rivers, Julian A., 1954. The People of the Sierra. Weidenfeld and Nicolson, London.
Razsa, Maple, 2008. Mutual aid, anti-authoritarianism, and Dumpster diving: anarchist activism in Croatia since 2000. In: Ramet, Sabrina P., Clewing, Konrad, Lukić, Renéo (Eds.), Croatia since Independence: War, Politics, Society, Foreign Relations. Oldenbourg Verlag, Munich.

Razsa, Maple, Andrej, Kurnik, 2012. The occupy movement in Žižek's hometown: direct democracy and a politics of becoming. American Ethnologist 39 (2), 238–258.
Razsa, Maple, Velez, Pacho, 2010. Bastards of Utopia. Documentary Educational Resources, Watertown, MA.
Robinson, Andrew, Simon, Tormey, 2012. Beyond the state: anthropology and 'actually-existing-anarchism'. Critique of Anthropology 32 (2), 143–157.
Rousselle, Duane, Evren, Süreyya, 2011. Post-anarchism: A Reader. Pluto, London.
Sahlins, Marshall D., 1972. Stone Age Economics. Aldine-Atherton, Chicago, IL.
Scott, James C., 2009. The Art of Not Being Governed: An Anarchist History of Upland Southeast Asia. Yale University Press, New Haven, CT.
Service, Elman, 1963. Profiles in Ethnology. Harper & Row, New York.
Stanner, William E.H., 1968. Radcliffe-Brown, A.R. Encyclopedia of the Social Sciences 13, 285–290.
Ward, Colin, 1973. Anarchy in Action. George Allen & Unwin, London.

Ancestors, Anthropology of

John Middleton, Yale University, New Haven, CT, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 1, pp. 494–496, © 2001, Elsevier Ltd.

Abstract

The main question asked about ancestors by anthropologists is how the living define them, transform the dead into them, and use them to safeguard social and individual moralities. Ancestors are distinct from the mass of the dead: their descendants select and make them by mortuary rites. The selected deceased are linked to their descendant by descent and are usually those who hold a structurally important genealogical position as founders of family segments. Those not selected for ancestorhood join the generality of the dead and are soon forgotten. Ancestors are thought to dwell in a home of the dead, or near their descendants with whom they may communicate; they may punish if offended. The latter communicate with the ancestors by offerings of food or goods. They are remembered in genealogies or memorialized by graves or written monuments. Their importance tends to lessen in 'modern' societies in which kinship may be of little importance, and where Christianity, Islam, or other 'world' religions give more emphasis to deities and less to ancestors.

The study of beliefs and rites involving ancestors has long been a central anthropological concern. Many early anthropologists suggested that ancestral worship was the forerunner of all religions and devoted much time to building conjectural religious histories. Twentieth century anthropologists have devoted attention to the working of ancestral cults as aspects of wider religious systems and to their links to rights in property, succession to office, and cohesion of descent groups. The main questions asked include how the living define ancestors, transform the dead into them, and use them as guardians of morality and agents of social continuity.

The word 'ancestor' is often used as a synonym for 'the dead': but they should be distinguished – not all the dead are remembered and defined as ancestors. Those selected as ancestors provide a history memorializing certain events of the past that give significance to claims of the present. Defined ancestors may either actually have lived at a remembered time, may be given a place in a genealogy irrespective of the date of death, or may be invented so as to provide their believed descendants with history and identity.

The status of ancestor is given by a rite of transition performed by the living descendants of the deceased. Robert Hertz (1960) listed the three components of mortuary rites: the corpse, the mourners, and the soul. The functions of the rites are to dispose of the corpse; to remove the deceased from the immediate domain of the living and to transform him or her into an ancestor; to re-form the descent group of the deceased's living kin; and to ensure the continuity of the essential element of the deceased's human identity, the soul or some similar concept. The forms the rites take reflect the importance of the deceased: those for a king may be highly elaborate and last many months or years, for a respected elder be less so and take a few days, for a child be minimal and take only a few minutes (Huntington and Metcalf, 1979; Bloch and Parry, 1982).

The living decide which of the dead shall be transformed into ancestors. The key criterion is the link of kinship between a living group or person and an ancestor. Without that link the deceased is not an ancestor but merely a dead person and soon forgotten. To possess recognized ancestors gives the living descendants full identity and status; to lack them (as in the case of slaves in many societies) is to lack full status.

Several ancestral categories may be defined. For example, the Lugbara of Uganda (Middleton, 1960) distinguish the *ori*, a descent group's direct and renowned male patrilineal forebears, to whom are made ritual offerings; *oku-ori*, female *ori*, daughters of the patriline who, although they bore sons to other descent groups (by the rule of exogamy), are remembered as powerful women; and *a'bi*, the general mass of other dead members of a given clan. Only relatively few of the dead become *ori* and *oku-ori*, whereas the number of *a'bi* is beyond reckoning. The ancestors of other clans are of no concern and are merely *a'bi* without definition.

Ancestral categories may also depend upon the manner of death. The Akan of Ghana, for example (Fortes, 1959), distinguish three kinds of death: good death, the natural death of man or woman followed by a future in the land of the ancestors; bad death, due to evil behavior during life and a double death, first that in the living world and later within the land of the ancestors, whose inhabitants refuse to accept the deceased who is returned to the land of the living to linger as a bad and dangerous specter; and sudden death, as in war, childbirth, accident, or suicide – such a person is not buried properly in a clan cemetery but in the bush land and is then 'lost' as an ancestor.

Among the Chinese, as a third example, Confucianism makes a general distinction between the ancestors and the non-ancestral dead (Hsu, 1948; Ahern, 1973; Wolf, 1974). The former are men of strictly patrilineal descent groups, senior kinsmen with clearly recognized rights due from their descendants and obligations towards them. Women are not defined as ancestors, although they have been daughters, wives, and mothers to the ancestors of a particular descent group.

These cases show that whether individuals are defined as ancestors depends upon their kinship ties with their descendants, linked by mortuary rites and by inheritance of social position and property. Ancestors possess no independent status in themselves but are given it by their living kin.

Once made into ancestors, where they dwell and the nature of their relationships to their descendants vary from

one society to another. Christian cultures hold that claimed ancestors and other dead dwell either with the Deity in a heaven, a place of peace without further passing of time, and/or a hell, a place of punishment and atonement. The Akan peoples of Ghana hold that ancestors dwell in a land of the dead across a river, which has the same patterns of stratification and authority as that of the living. The Lugbara of Uganda consider that they dwell underground near their descendants' settlements, where they themselves have been buried. In virtually all religions ancestors are considered to be senior kin living apart from their descendants yet still linked by spatial propinquity and kinship interdependence. They are typically believed to see and hear their living kin, who as their juniors are unable to see or hear them except in unusual circumstances, such as their appearance as specters or when people are possessed by them.

Associated with the concept of ancestor is that of soul, linked to individual morality rather than to descent and kinship. The concept of soul was of great concern to early anthropologists such as Frazer (1911-15), Lévy-Bruhl (1927) and Tylor (1871), who coined the term 'animism' for the belief in souls attributed to all living things. Their views were essentially conjecture: more recently ethnographic research, including that on religion, has become based on detailed study of the beliefs and practices of particular cultures. For example, the Lugbara word *ori* is related to the word *orindi* (literally 'the essence of the *ori*') which might be translated as 'soul,' an element of the living that is not exterminated by death but which persists into the afterlife. Only men of full adult status have what might be called full souls; women who are the first-born of a set of siblings may be attributed souls if they acquire respected status before death, and it is they who become *oku-ori*. Other men and women lack souls or, more accurately, do not realize the potentiality of developing them, and it is they who are placed in the loose category of *a'bi*. Different societies order these concepts and elements in differing ways, but the basic factor is that of an individual's moral passage through life, the soul being that element that bestows him or her with morally responsible behavior.

Ancestors have to be remembered, usually by being listed in genealogies. Evans-Pritchard (1940) showed for the Nuer of the Sudan that the generational depth of genealogies reflects the width and composition of descent groups on the ground: the wider the group the more the number of generations needed to include all its members in a single genealogy. The identity of effective ancestors usually depends upon the recognition of a particular mode of descent (patrilineal, matrilineal, or cognatic) as the basis for local social organization. Genealogies are not, or only extremely rarely, historically accurate statements. There is typically a telescoping of generations so as to maintain the same depth of a genealogy over time; and in patrilineal systems, certainly, the exclusion of most women is usual as they do not exercise the same degree of authority as do men. The situation is different in those societies in which ancestors are memorialized in writing or on tablets (e.g., China and Mormon America); here the purpose may be to include all forebears by name irrespective of their immediate and active links to their descendants.

Terms commonly used in discussions of ancestors include 'ancestral spirits,' 'ancestral worship,' and 'ancestral cults,' that refer to the relationships between the living and their own ancestors. The beliefs and practices concerning ancestors nowhere form a distinct religion but rather a religious practice within a wider societal religion. This practice typically includes lineage and family rites, periodic rites on ancestral death days, and often annual rites for collectivities of a group's ancestors. Rites linked to the non-ancestral dead in general should not be included. Ancestors are virtually everywhere held to care for and protect their descendants; they may teach, criticize, and punish them for unfitting behavior. They may send sickness or other affliction to those who disobey their rules and wishes, and they are in general held to be conservative in their views and to represent an idealized stability and moral order. They may act on their own initiative or may be evoked by their descendants. The living may consult oracles or diviners, or interpret omens, to discover the identities and motives of specific ancestors concerned.

The living may communicate with their ancestors both by making verbal requests and by offering them food (often flesh and blood) and drink, commensality being a central act of kinship bonding. Offerings may be regular or made only when the ancestors show that they wish to receive them, at shrines of many kinds, from ancestral graves to special buildings, and may be simple and cursory or elaborate and long-lasting, depending on the ancestors' remembered status.

The term 'ancestral spirit,' used for those aspects of ancestors thought to come into direct contact with the living, is clumsy: ancestors and spirits are in their natures quite distinct. Ancestors have been among the living, who remember the more recent of them as individuals and understand their wishes and demands. Even if their individual personalities are in time forgotten, they represent human experience, skills, and sentiments, and are senior kin. They are used by the living both to maintain conservative authority and moral order and to help bring about orderly settlement of disputes. They provide and maintain the long-lasting stability of otherwise ever-changing and fluid descent groups. On the other hand, spirits are refractions of divinity, have never been living and so have no descendants: their natures and demands are different and may be beyond human understanding. The living can exercise a more immediate control over their ancestors by stressing their link as kin, than they can over spirits. Ancestors may act as unvarying custodians of order, unlike spirits which are by contrast individualistic, capricious, and unpredictable.

An important point is that of the persistence of beliefs in ancestors in the 'modern' industrialized world. The rites of invocation and appeasement of ancestors may lose importance for small-scale descent groups precisely as these lose importance in radically changing societies; ancestors are those who, when they lived, were important in the lives and continuity of their descent groups of the time. By being remembered and commemorated in ritual, they are objects of the group's memory of the past and its identity of the present. Ancestors used as mnemonic figures remain important, as do their material representations such as tombs and shrines. This is especially so in the case of those who in their lives were kings, sacred personages who embodied the identity and continuity of their kingdoms. Such royal ancestors are not only used as signs of memory by their genealogies: they are typically given

royal tombs that materially represent the past in the present and so act as centers of societal identity. In societies that accept Christianity, Islam, or other world religions, greater emphasis may be given to spirits: ancestors' powers are limited to their own descendants whereas those of spirits are typically held to affect all members of a society irrespective of their descent groups.

See also: Death, Anthropology of; Kinship in Anthropology; Religion, Family, and Kinship; Religion: Evolution and Development; Ritual.

Bibliography

Ahern, E., 1973. The Cult of the Dead in a Chinese Village. Stanford University Press, Stanford, CA.
Bloch, M., Parry, J. (Eds.), 1982. Death and the Regeneration of Life. Cambridge University Press, Cambridge, UK.
Evans-Pritchard, E.E., 1940. The Nuer. Clarendon Press, Oxford, UK.
Fortes, M., 1959. Oedipus and Job in West African Religion. Cambridge University Press, Cambridge, UK.
Frazer, J.G., 1911-15. The Golden Bough. Macmillan, London.
Hertz, R., 1960. (1st edn. 1909) Death, and the Right Hand. Cohen and West, London.
Hsu, F.L.K., 1948. Under the Ancestors' Shadow. Norton, New York.
Huntington, R., Metcalf, R., 1979. Celebrations of Death. Cambridge University Press, Cambridge, UK.
Lévy-Bruhl, L., 1927. L'Ame Primitive. Alcan, Paris (English edn., The Soul of the Primitive 1928).
Middleton, J., 1960. Lugbara Religion. Oxford University Press, London. new edition 1999 James Currey, Oxford, UK.
Tylor, E.B., 1871. Primitive Culture. Murray, London.
Wolf, A., 1974. Religion and Ritual in Chinese Society. Stanford University Press, Stanford, CA.

Ancient Egypt, Archaeology of

Marleen D Meyer and Stefanie Vereecken, KU Leuven, Leuven, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The civilization of ancient Egypt flourished between 4000 BC and AD 395 along the banks of the River Nile. During the late nineteenth century the archaeology of this region started to take shape, originally focusing mainly on cemetery sites, but since the 1960s settlement sites have been increasingly excavated. Recent developments in the field of archaeometry are opening up new research perspectives for the study of the material culture.

The archaeology of ancient Egypt is concerned with the remains of prehistoric cultures (*c*.700 000–4000 BC), as well as predynastic and pharaonic civilization (*c*.4000 BC–AD 395) that flourished along the banks of the River Nile and further into the desert in an area that roughly corresponds with modern-day Egypt. Chronologically and geographically several delimitations are possible.

Chronology

The historic sequence of pharaohs is based on kings' lists, annals, and other documents that survived from ancient Egypt (**Table 1**; Redford, 1986). However, the division of ancient Egyptian history in 30 dynasties is based on the writings of the Egyptian priest Manetho, who wrote the *Aegyptiaca* (History of Egypt) in the third century BC. The original manuscript is not preserved, but passages were handed down by later authors. Egyptologists have added two more to Manetho's 30 dynasties: First, Dynasty 0 encompasses a series of rulers that preceded Narmer (now generally seen as the first king of Dynasty 1) and are buried in cemetery U at Abydos. During Dynasty 0, writing was known, but Egypt was not yet unified. The second is Dynasty 31, which refers to a short period between 343 and 332 BC when Egypt was a Persian satrapy for the second time in its history. These dynasties were grouped together by nineteenth-century Egyptologists into several larger periods when Egypt possessed a strong unified state (Old, Middle, and New Kingdoms) and intermediate periods, when central authority was waning for different reasons (First, Second, and Third Intermediate Periods). After the Third Intermediate Period follows the Late Period, which ends with Dynasty 31. Alexander the Great conquered Egypt in 332 BC, and the Macedonian dynasty that occupied the throne in Egypt after his death is referred to as the Ptolemaic Dynasty. In 30 BC Egypt became a Roman province until AD 395. Pharaonic civilization had already died a slow death through the spread of Christianity starting in the third century AD (Coptic Period) when Egypt was conquered by Arabs in AD 641.

Prehistory and the era immediately preceding and leading up to the pharaonic civilization – the Pre- and Early Dynastic Period – have received a greater attention from scholars since the 1980s, so much so that this has almost grown into its own discipline. However, the chronology of this period consists mainly of local site- or region-specific chronologies, and an all-encompassing study still remains to be made (Hendrickx, 2006). The beginning of the historical period of pharaonic Egypt is traditionally defined by the unification of Upper and Lower Egypt under the legendary King Narmer. However, the discovery in 1988–89 of the oldest hieroglyphic writing on jar labels in Tomb U-j at Abydos has pushed back the date of the origin of writing in Egypt about 150 years, thus also pushing back the start of the historical period when this is defined by the presence of writing (Dreyer, 1998).

Table 1 Chronological overview of ancient Egyptian history

Paleolithic Period	c.700 000 BP-7000 BC
Saharan Neolithic Period	*c*.8800–4700 BC
Predynastic Period	*c*.5300–3000 BC
Early Dynastic Period (Dynasties 1–2)	*c*.3000–2686 BC
Old Kingdom (Dynasties 3–8)	2686–2160 BC
First Intermediate Period (Dynasties 9–mid-11)	2160–2055 BC
Middle Kingdom (Dynasties mid-11–14)	2055–1650 BC
Second Intermediate Period (Dynasties 15–17)	1650–1550 BC
New Kingdom (Dynasties 18–20)	1550–1069 BC
Third Intermediate Period (Dynasties 21–25)	1069–664 BC
Late Period (Dynasties 26–31)	664–332 BC
Ptolemaic Period	332–30 BC
Roman Period	30 BC–AD 395
Egypt ruled from Byzantium	AD 395–641
Arab conquest of Egypt	AD 641

Source: After Shaw, Ian., 2000. The Oxford History of Ancient Egypt. University Press, Oxford.

Geography

The Nile Valley is clearly defined by the natural environment: the Mediterranean Sea to the north, the First Cataract in the Nile to the south, and the desert to the east and west form clear natural buffers. In antiquity, the southern border of Egypt was generally located in the area of the First Cataract, near modern-day Aswan. At different periods of Egyptian history, this border moved further south to include large parts of Nubia in present-day Sudan north of Khartum. The eastern border also moved further east, especially during the New Kingdom, to incorporate parts of the Levant.

The Development of Egyptian Archaeology

While interest in the pharaonic culture was already cultivated by the ancient Greeks and Romans, and later reappeared during

the Renaissance, the scientific study of ancient Egypt in the west only took off in the beginning of the nineteenth century. The expedition of Napoleon to Egypt in 1798 and the resulting publication of the *Description de l'Égypte* in 24 volumes (1809–29) are generally considered to be the starting point for this. During the late eighteenth and early nineteenth centuries, several other descriptive works were published that documented the state of monuments throughout Egypt at that time, such as the *Denkmaeler aus Aegypten und Aethiopien* in 12 volumes (1849–59) produced by the Prussian expedition of Karl Richard Lepsius. Based on such publications, the study of ancient Egypt originally focused on philology rather than archaeology, with several scholars attempting to decipher the hieroglyphic script. This reached a breakthrough in 1822 when Jean-François Champollion outlined the basics of the system in his *Lettre à M. Dacier*. Even though around the same time major European museums such as the Louvre and the British Museum were starting to amass collections of ancient Egyptian antiquities, the excavation of these objects was not of a scientific nature. This started to change in 1859 when the Department of Antiquities was founded in Cairo by the French Egyptologist Auguste Mariette to regulate excavations throughout Egypt. While this institution was originally headed by a series of French Egyptologists, the Egyptians have directed it since 1953, and it still exists, albeit in a modified form. Since the revolution of 2011, this institution became its own ministry and was renamed as the Ministry of State for Antiquities (MSA). The MSA regulates excavations in Egypt both by foreign missions as well as Egyptian missions, oversees conservation projects, and is generally in charge of protecting Egypt's heritage.

Archaeology primarily lingered in the realm of treasure hunting until scholars such as William Flinders Petrie (1853–1942, the UK) and George Andrew Reisner (1867–1942, the USA) started to define methodological approaches in excavation during the late nineteenth century. Petrie was not only pioneering within the field of Egyptian archaeology itself, but his methods of seriation applied to Pre- and Early Dynastic Egyptian pottery were revolutionary for the field of archaeology in general (Petrie, 1901). Petrie also introduced the concept of stratigraphy into Egyptian archaeology. Reisner was influential for his methodological record keeping and documentation, describing for instance at length his principles for archaeological photography (Manuelian, 1992), and he was innovative in surveying techniques as well. Around the same time scholarly societies were being formed, mainly in Europe, that funded and supported archaeological fieldwork in Egypt, such as the Egypt Exploration Fund (presently still operating under the name Egypt Exploration Society) in the United Kingdom, which was founded by Amelia Edwards in 1882, and the Institut français d'archéologie orientale, which was founded in 1898 and is also still active to this day.

In the 1960s there was an international effort to salvage sites and monuments that were under threat by the rising water table of Lake Nasser, the artificial lake that formed as a result of the construction of the second High Dam at Aswan. Monuments in the south of Egypt and the Sudan were moved to higher grounds, sites were excavated, and some monuments were donated to museums in participating countries. This also led to Nubiology as a discipline.

In Egyptian archaeology, the focus has historically mainly been on funerary archaeology. Settlement archaeology did not seriously begin until the 1960s, a time when archaeological methodology also started to change under the influence of processual archaeology (a theoretical framework, which found its origins in the late 1950s). The reasons for this focus on funerary archaeology are outlined below. Presently, the study of the material culture is advancing rapidly with the application of techniques from the physical sciences.

Funerary Archaeology

The spectacular discovery of a cache of royal mummies (dating to the New Kingdom and later) in the Theban necropolis by the Abd el-Rasul family in 1870 set in motion the scientific exploration of this area in the late nineteenth century. The Valley of the Kings underwent more systematic exploration since then, which led to the discovery of the tomb of Tutankhamun in 1922 by Howard Carter. The latter provoked a wave of Egyptomania in the decorative arts that had not been seen since the expedition of Napoleon over a century earlier. In the meantime, Reisner had started systematic exploration of the Giza plateau – the necropolis of the kings and high officials of the Fourth Dynasty – and Petrie was working at Abydos in the tombs of the early Dynastic kings. A bit later, right before World War II, Pierre Montet had discovered royal tombs of the Twenty-first and Twenty-second Dynasty at Tanis in the Delta. Thus the tombs of many of the kings of the pharaonic civilization had been under investigation in the first half of the twentieth century.

In addition to royal tombs, necropoli of the general population were also being excavated, although often they too belonged to the elite rather than to commoners. Early twentieth century excavations at cemetery sites tended to cover vast surfaces with very little publication of the work afterward. An example of such an excavation is the work of John Garstang at the Middle Kingdom necropolis of Beni Hasan (Garstang, 1907). This practice has led to a paradox within the archaeological record of cemetery sites in Egypt, namely that while the surface areas that were excavated in the early days are exponentially larger than the surface areas that can be covered by modern-day techniques and methodology, the documentation of these old large-scale excavations does not live up to modern standards. Additionally, many of the original records are still locked away in museum cupboards or institute basements. Digital imaging and online publication of old excavation documentation, however, have recently started to open up many new research perspectives. A pioneering project in this regard is the Giza Archives Project at the Museum of Fine Arts in Boston (www.gizapyramids.org). At the core of the project is the aim to publish online all the photos, excavation diaries, drawings, plans, in short all the documentation that was produced by George Reisner during his excavations at Giza from 1902 to 1942. However, the project goes beyond that by aiming to include other sources of information on the site of Giza as well, and likewise reconstructs the site in 3D virtual reality.

Despite the fact that old records are not up to modern standards, this does not mean that the data from the old excavations should be disregarded all together, but rather that techniques should be applied to distill as much information

out of them as possible. A good example of the type of results that can be obtained by applying modern statistical analyses to old excavation data from cemeteries is Seidlmayer (1990). Methodical reexcavation of sites that were hastily dug through in the past, also proves to be rewarding, such as is clear from the reexcavation of the early Dynastic cemetery of Umm el-Qa'ab at Abydos by the German Archaeological Institute (Dreyer, 1998). Presently, cemetery sites are studied at a much slower pace and in an interdisciplinary way. Physical anthropologists, zooarchaeologists, botanists, geographers, etc., complement the team of Egyptologists and archaeologists, such as for instance at the site of at Hierakonpolis in Upper Egypt (www.hierakonpolis-online.org) or at Dayr al-Barsha in Middle Egypt (www.dayralbarsha.com). The latter also forms part of a recent trend in research of provincial sites, as does the Asyut Project (http://www.aegyptologie-altorientalistik.uni-mainz.de/139.php).

Settlement Archaeology

Settlement archaeology has developed at a much slower pace in the field of Egyptian archaeology than it has for instance in Near Eastern archaeology, where *tell* sites present accumulated layers of human occupation. There are several reasons for this. Firstly, there has been a bias in the research due to the overwhelming presence of well-preserved temples and tombs in Egypt. These monumental and generally elite constructions have drawn away attention from settlements during much of the nineteenth and early twentieth centuries, with some exceptions such as Deir el-Medina. It is only from the 1960s onward that serious research interest in ancient Egyptian settlements began, the pioneering project being the excavations of the Austrian Archaeological Institute of Cairo under the direction of Manfred Bietak at the site of Tell el-Daba' in the eastern Delta (www.auaris.at).

Secondly, settlements in the Nile Valley are generally badly, if at all, preserved due to the specific geographic situation. Settlements, located in the narrow strip of fertile land along the banks of the Nile, were subject to annual flooding, and were frequently built over in subsequent periods of Egyptian history, even until this day. Therefore, even if settlement remains are preserved, they are often very difficult to reach. This is especially true for organically grown settlements in the Nile Valley, the villages and towns that made up the backbone of the Egyptian landscape. A good analogy for the development of urbanization within these organically grown settlements is the game-playing model as presented by Kemp (2006). The best preserved example of an at least partially organically grown settlement in the Nile Valley is the site of Tell el-Amarna, the short-lived capital of Egypt during the reign of Akhenaton, which was built and subsequently abandoned between years 6 and 17 of his reign (www.amarnaproject.com). Although this city is an abnormality in itself, and large portions of it were laid out in a planned manner, the houses in between the official structures tend not to follow a preplanned layout.

There are also a number of settlements, however, that are not located within the Nile Valley itself, but immediately outside of it in the desert fringe, which have managed to escape the same fate. Such settlements are generally not organically grown, but rather (state) planned settlements that were built with a specific purpose in mind. Settlements of this type include workmen's villages to house workers of large building projects (e.g., at Deir el-Medina for the construction of the New Kingdom royal tombs in the Valley of the Kings, or at Heit el-Ghorab for the building of the Old Kingdom pyramids at Giza; www.aeraweb.org); pyramid towns where personnel were housed who were in charge of maintaining the mortuary cult of deceased kings (e.g., the Old Kingdom Khentkawes town at Giza, or the Middle Kingdom town of Lahun next to the mortuary complex of Senwosret II); and fortresses built to defend the borders of Egypt (e.g., the Nubian fortresses of Mirgissa, Shalfak, Uronarti, Buhen, and others which were mainly in use during the Middle Kingdom; or the fortress at Zawiyet Umm el-Rakham about 300 km west of Alexandria, built during the New Kingdom as a defense against the Libyans). Such planned settlements are usually laid out along an orthogonal plan, with modular living spaces of an identical or highly similar structure. While such settlements are better preserved in the archaeological record, they cannot be taken as type examples of an average ancient Egyptian town. The population of such settlements is also often not representative of an average population (concentrations of workmen or soldiers instead of families), and since food was often prepared outside such settlements and then delivered as finished meals to the workers, the material culture does not reflect what can be expected in organically grown towns. In areas further away from the Nile Valley, such as the oases in the Western Desert, and in particular in the Dakhla Oasis and Kharga Oasis, well-preserved settlements have also been found. These locations were of strategic importance for trade routes, and for instance at Balat/'Ayn Asil (Dakhla Oasis) an entire settlement evolved around the palace of local Sixth Dynasty governors, including an industrial area of potters' workshops (Soukiassian et al., 1990). The administration in this remote locality seems to have been remarkably similar to that of the residence in the Nile Valley, albeit on a more modest scale (Pantalacci, 2013).

Remote sensing, satellite imagery, and ground-penetrating radar are tools that are increasingly being used to try and remedy the lack of knowledge about organically grown settlements in the Nile Valley. But even if settlements can be located this way, the archaeological layers on top will often date to the Greco-Roman Period. To gather information about lower and older layers of occupation, traditional archaeological methods will need to supplement the data obtained through remote sensing.

Study of Material Culture

While during the first half of the twentieth century large collections of artifacts had already been excavated in Egypt, the analytical study of the material culture only seriously began with the publication of *Ancient Egyptian materials and industries* (Lucas, 1926). This remained for a long time the standard work of reference for Egyptologists and archaeologists in this field, and several revised editions appeared until 1962 with supplementary material added by John R. Harris. Over the last few decades a more multidisciplinary approach has been applied to different materials in the archaeological record, involving

scientific analytical techniques, which provide new insight into different technological and socioeconomic questions. This has led to the new standard work *Ancient Egyptian Materials and Technology* (Nicholson and Shaw, 2000), which forms a solid base for different aspects of the material culture and its technological aspects.

Pottery

In spite of the fact that pottery occurs in large quantities on every site in Egypt, there was no profound interest in the pottery of ancient Egypt until the mid-twentieth century. Although Petrie's methods of seriation applied to Pre- and Early Dynastic Egyptian pottery were revolutionary for the field of archaeology in general, his example was hardly followed and the most common practice on excavations in those days consisted of collecting only complete vessels or decorated ware without paying further attention to the identification of materials and their source of trade, nor to the physical characteristics of vessels to show their place in technological development.

A turning point came with the formation of the international group for the study of ancient Egyptian pottery, an initiative undertaken by Dorothea Arnold of the German Archaeological Institute in 1975. Here the foundation was laid for a new approach of pottery analysis and from then on, there has been a change in the emphasis of pottery analysis, with more attention being paid to disciplines such as fabric analysis and archaeometry. An important milestone in this evolution was the publication of *An Introduction to Ancient Egyptian Pottery* (Arnold and Bourriau, 1993), a pioneering work for the processing and description of pottery in the field. This publication pointed out the importance of the study of technology and fabric and introduced the Vienna System, a visual classification system for fabrics, which could be applied in the field. This system does not depend on petrological analysis of thin sections or other laboratory-based techniques, which are also necessary, but which are difficult to obtain because only few laboratories in Egypt offer the possibility for such analysis, and taking samples out of the country is forbidden. Despite these problems, archaeometry is also finding its way into pottery analysis as a subsequent step intended to test and refine the groupings of fabrics (e.g., Bourriau et al., 2000 for the petrographic analysis of New Kingdom pottery fabrics). The introduction of the Vienna System has greatly facilitated comparison between the pottery assemblages of different sites all over Egypt. A guideline for the drawing of pottery, which was established by Aston (1998), is also being used on many contemporary excavations.

During the last decade, the analysis of ancient Egyptian pottery has become a subdiscipline of Egyptian archaeology, not only through the addition of scientific techniques, but also through experimental and ethnoarchaeological research (e.g., the work of Nicholson for the Tell el-Amarna Project, www.amarnaproject.com; or Vereecken for the Dayr al-Barsha Project, www.dayralbarsha.com). This has resulted in the publication of large comprehensive works with not only a detailed typological overview of pottery vessels, but also information on the technological and the socioeconomic aspects, shedding new lights on trade patterns and regional differentiation between the pottery assemblages (e.g., Schiestl and Seiler, 2012; Wodzińska, 2009–2010).

Stone

The exploitation of stone in Egypt goes back to at least 40 000 BP, when pits and trenches were dug for surface exploitation. As early as 35 000 BP, the first vertical shafts and galleries appear at the site of Nazlet Khater, but it is from the Old Kingdom onward that large-scale exploitation took place as the building of numerous royal complexes resulted in an unprecedented demand for stone for architecture, sculpture, and stone vessels. For decades, interest was mainly focused on the large architectural constructions (Clarke and Engelbach, 1930 remained for a long time the reference work on ancient Egyptian architecture), and stone sculptures and vessels were mainly analyzed from an art-historical point of view. The study of stone material and its technology in all its aspects has found a renewed interest during the last decades mainly because of new scientific analytical techniques and experimental archaeology (e.g., Stocks, 2003). This has resulted in the development of many subdisciplines of the analysis of the stone material culture, such as the geological analysis of quarries and their different exploitation techniques (e.g., Klemm and Klemm, 2008), or the typology and mineral composition of stone vessels (e.g., Aston, 1994). The study of quarry marks, and nontextual marks in general including for instance also pot marks, has recently been generating a wide interest and yields important information about the organizational and bureaucratic aspects of quarrying (e.g., Depauw, 2009). A pioneering project is the Giza Plateau Mapping Project under the direction of Mark Lehner, which has mapped out the quarries and exploitation of stone during the building of the pyramids at Giza (www.aeraweb.org/gpmp-project/great-pyramid-quarry). The socioeconomic aspects of quarrying are for instance investigated at Dayr al-Barsha, where research has been conducted on the quarry marks and the road system of the quarries that were exploited to build the nearby city of Tell el-Amarna, the short-lived capital of Egypt during the reign of Akhenaton. The recent discovery of a stoneworking site with hundreds of drills and other stoneworking tools at al-Shaykh Sa'id offers insight in the technology of the production of stone vessels during the New Kingdom (Willems et al., 2009).

Egyptian Faience, Blue Frits, and Glass

Faience is a glazed nonclay ceramic material. Its name has been the subject of confusion as it also refers to a type of pottery now known as majolica, a tin-glazed ware made in Faenza in northern Italy from late medieval times onward. The term faience in Egyptian archaeology is used for objects made entirely from the glazed material, as distinct from glazed stone or glazed ware, which are objects with a coating of glaze. In publications it is also possible to come across the terms 'blue frits' or 'Egyptian blue,' but one should be aware that these terms are not exchangeable with (Egyptian) faience, as has often been done by scholars (see Nicholson and Shaw, 2000). Egyptian blue appears to be an Egyptian invention and is,

although it appears much earlier, closely allied to glass (the earliest glass pieces date back to the Twelfth Dynasty, but only from 1500 BC onward does glass start to appear more commonly in Egypt). Faience on the other hand is a material that occurred widely in the ancient world and can be found in Egypt from the Predynastic period on. Egyptian blue frits are easily distinguishable from faience in a broken section, as the latter always shows a core and a separate glaze layer, contrary to the homogenous section of a frit. Experimental archaeology and archaeometrical research on these different materials open up new insights, such as for instance the work of Nicholson for the Tell el-Amarna Project; www.amarnaproject.com.

Metal

The knowledge of ancient sources of metal ores in Egypt has increased since the 1940s through geological studies on metal ore distribution and geochemistry. The main copper-producing areas were located in the Eastern Desert and the Sinai, but copper workshops have been identified throughout Egypt (e.g., a major copper-alloy working center was discovered at Qantir, ancient Piramesse (Pusch, 1990)). The oldest copper objects were already produced during the Predynastic period (Badarian Period). These were probably produced from native copper, i.e., copper found in nature in metallic form, not as ores. The introduction of smelted copper with, in general, the use of iron oxide as a flux occurred in Egypt around 4000 BC, but more efficient production techniques were not introduced until the New Kingdom. Our understanding of ancient Egyptian metals comes from the careful study of surviving examples using, increasingly, technological advances.

Future Directions in Research in Egyptian Archaeology

The field of Egyptian archaeology has undergone significant changes in the past few decades, mainly under the influence of the input of scientific technology in the field. The interdisciplinary character of excavations is still increasing, and will lead to new research questions to be posed and answered. An area where this will certainly pay off is in the establishment of chronology. A large research project at the Research Laboratory for Archaeology and History of Art, University of Oxford, is working toward linking the traditional historical chronology to dating obtained through the physical sciences (Shortland and Ramsey, 2013). This research unit has carbon (C^{14}) dated a large sample of short-lived plant remains that were well attributable to individual reigns of kings. Good results were obtained mainly for the New Kingdom, and can be incorporated into the revision of the historical chronology carried out by the SCIEM2000 Project (the Synchronization of Civilizations in the Eastern Mediterranean in the second millennium BC, Austrian Academy of Sciences; www.oeaw.ac.at/sciem2000).

A challenge that also awaits Egyptian archaeology is to draw up regional typologies and chronologies of material culture. Regional differentiation often remains either unrecognized or unstudied. This is something that is already well underway in ceramics research, but is to a large extent lacking for other archaeological remains. Material culture did not change at the same time or at the same speed throughout the entire country, and patterns need to be recognized in this evolution.

See also: Archaeology, Theory In; Archaeometry; Ceramics in Archaeology; Chronology, Stratigraphy, and Dating Methods in Archaeology; Ritual and Symbolism, Archaeology of; Settlement and Landscape Archaeology.

Bibliography

Arnold, Dorothea, Bourriau, Janine D. (Eds.), 1993. An Introduction to Ancient Egyptian Pottery. Verlag Philipp von Zabern, Mainz am Rhein.

Aston, Barbara, 1994. Studien zur Archäologie und Geschichte Altägyptens 5: Ancient Egyptian Stone Vessels: Materials and Forms. Heidelberger Orientverlag, Heidelberg.

Aston, David, 1998. Die Keramik des Grabungsplatzes Q1. Teil 1: Corpus of fabrics, wares and shapes. Verlag Philipp von Zabern, Mainz am Rhein.

Bourriau, Janine D., Smith, Laurence M.V., Nicholson, Paul T., 2000. Egypt Exploration Society Occasional Publications 14: New Kingdom Pottery Fabrics. Nile Clay and Mixed Nile/Marl Clay Fabrics from Memphis and Amarna. Egypt Exploration Society, London.

Clarke, Somers, Engelbach, Reginald, 1930. Ancient Egyptian Masonry: The Building Craft. Oxford University Press, Oxford.

Depauw, Mark, 2009. Quarry marks in Deir el-Barsha and the logistics of building materials in late period and Graeco-Roman Egypt. In: Haring, Ben, Kaper, Olaf E. (Eds.), Pictograms or Pseudo Script? Non-textual Identity Marks in Practical Use in Ancient Egypt and Elsewhere. Proceedings of a Conference in Leiden, 19–20 December 2006. Nederlands Instituut voor het Nabije Oosten, Leiden, pp. 93–106.

Dreyer, Günter, 1998. Archäologische Veröffentlichungen 86: Umm el-Qaab 1. Das prädynastische Königsgrab U-j und seine frühen Schriftzeugnisse. Verlag Philipp von Zabern, Mainz am Rhein.

Garstang, John, 1907. Burial Customs of Ancient Egypt as Illustrated by Tombs of the Middle Kingdom. A Report of Excavations Made in the Necropolis of Beni Hassan during 1902–3–4. Archibald Constable & Co, London.

Hendrickx, Stan, 2006. Predynastic–early dynastic chronology. In: Hornung, Erik, Krauss, Rolf, Warburton, David E. (Eds.), Handbook of Oriental Studies. Section 1: The Near and Middle East 83: Ancient Egyptian Chronology. Brill, Leiden, pp. 55–93.

Kemp, Barry John, 2006. Ancient Egypt: Anatomy of a Civilization, second ed. Routledge, London.

Klemm, Dietrich, Klemm, Rosemarie, 2008. Stones and Quarries in Ancient Egypt. British Museum Press, London.

Lucas, Alfred, 1926. Ancient Egyptian Materials and Industries, first ed. Kessinger Pub, London.

Manuelian, Peter D., 1992. George Andrew Reisner on archaeological photography. Journal of the American Research Center in Egypt 29, 1–34.

Nicholson, Paul T., Shaw, Ian (Eds.), 2000. Ancient Egyptian Materials and Technology. Cambridge University Press, Cambridge.

Pantalacci, Laure, 2013. Balat, a frontier town and its archive. In: Garcia, Moreno, Carlos, Juan (Eds.), Handbuch der Orientalistik 104: Ancient Egyptian Administration. Brill, Leiden, pp. 197–214.

Petrie, William M.F., 1901. Memoir of the Egypt Exploration Fund 20: Diospolis Parva. The Cemeteries of Abadiyeh and Hu, 1898–9. Egypt Exploration Fund, London.

Pusch, Edgar B., 1990. Metallverarbeitende Werkstätten der frühen Ramessidenzeit in Qantir-Piramesse/Nord. Ägypten und Levante 1, 75–113.

Redford, Donald B., 1986. Society for the Study of Egyptian Antiquities Publications 4: Pharaonic King-lists, Annals and Day-books. A Contribution to the Study of the Egyptian Sense of History. BenBen Publications, Mississauga.

Schiestl, Robert, Seiler, Anne (Eds.), 2012. Handbook of the pottery of the Egyptian Middle Kingdom, 2 vols. Verlag der Österreichischen Akademie der Wissenschaften, Vienna.

Seidlmayer, Stephan J., 1990. Studien zur Archäologie der Ersten Zwischenzeit 1: Gräberfelder aus dem Übergang vom Alten zum Mittleren Reich. Heidelberger Orientverlag, Heidelberg.

Shaw, Ian, 2000. The Oxford History of Ancient Egypt. University Press, Oxford.
Shortland, Andrew J., Ramsey, C. Bronk, 2013. Radiocarbon and the Chronologies of Ancient Egypt. Oxbow Books, Oxford.
Soukiassian, Georges, Wuttmann, Michel, Pantalacci, Laure, 1990. Fouilles de l'Institut français d'archéologie orientale 34: Balat 3. Les ateliers de potiers d'Ayn-Asīl: fin de l'Ancien empire–première période intermédiaire. Institut français d'archéologie orientale du Caire, Cairo.
Stocks, Denys A., 2003. Experiments in Egyptian Archaeology. Stoneworking Technology in Ancient Egypt. Routledge, London and New York.
Willems, Harco, Vereecken, Stefanie, Kuijper, Lucia, et al., 2009. An industrial site at al-Shaykh Sa'id/Wadi Zabayda. Ägypten und Levante 19, 293–331.
Wodzińska, Anna, 2009–2010. AERA Field Manual Series 1: A Manual of Egyptian Pottery, 4 vols. Ancient Egypt research associates, Boston, MA.

Relevant Websites

www.aegyptologie-altorientalistik.uni-mainz.de/139.php – Asyut Project.
www.dayralbarsha.com – Dayr al-Barsha Project.
www.gizapyramids.org – Giza Archives at the Museum of Fine Arts in Boston.
www.aeraweb.org – Giza Plateau Mapping Project.
www.hierakonpolis-online.org – Hierakonpolis Project.
www.oeaw.ac.at/sciem2000 – SCIEM2000 Project.
www.amarnaproject.com – Tell el-Amarna Project.
www.auaris.at – Tell el-Daba'a Project.

Ancient Greece and Rome, History of

Wilfried Nippel, Humboldt-Universität zu Berlin, Philosophische Fakultät I, Unter den Linden, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article covers the development of the Greco-Roman World from Early Greece to Late Antiquity. It focuses on those political and cultural achievements that constituted the legacy of antiquity for European culture: the notion of citizenship and individual rights, the ideas of republicanism and democracy as well as universal emperorship, the continuity of Roman law, and of the Church (including the use of the Latin language).

Antiquity is commonly understood as the history of the Mediterranean world and its neighboring regions before the Middle Ages. The concept is a legacy of the tripartite division of history that has developed since the Renaissance. There was always a tension between mere concentration on Greek and Roman history as representing 'classical' antiquity, and an approach including the high cultures of Egypt and the Near East. On the one hand, the Assyrian and Persian empires were due to the theory of the four world monarchies (adopted from oriental sources in the Book of Daniel), part of a Christian view of history. On the other hand, there was a strong tradition (going back to Aristotle) of contrasting Western freedom with oriental despotism where subjects lived under slavelike conditions. Until the nineteenth century oriental history could only be viewed through classical and biblical sources. Since then, decipherment of hieroglyphs and cuneiform as well as excavations has led to a reconstruction based on monumental sources. Their interpretation, however, required a linguistic and archaeological specialization, which led to a new departmentalization between classical and oriental studies.

Early Greece

Until the late nineteenth century, the beginnings of Greek history were equated with the first literary sources, the Homeric epics (eighth century BC). Since then excavations have enabled detection of the main structures of the Bronze Age civilizations – the 'Minoan' in Crete and the 'Mycenaean' in parts of mainland Greece, where great palaces served as centers of a redistributive economy. Those cultures were finally destroyed during the twelfth century BC by waves of invaders whose identity is far from clear, although they were probably not identical with the Dorian Greeks who afterward settled in parts of mid-Greece, the Peloponnese, and Crete. The structures of the 'Dark Ages' up to the eighth century BC can only be reconstructed conjecturally. There was no need to build up greater units, since there were no attempts by foreign powers to establish control. Social life took place in village communities with some collective organization for warfare and regulation of internal conflicts. The leading part was played by well-to-do landowners, who had to rely on the support of the peasants and permanently competed for a precarious leadership within the community.

The so-called archaic age from *c*.750 BC to *c*.500 BC is the formative period of Greek culture. The adaptation of the Phoenician alphabet was a cultural breakthrough; the epics of Homer and Hesiod (*c*.700 BC) canonized the ensemble of gods and heroes, and Panhellenic festivals such as the Olympic Games presented places of contact for aristocrats from all over the Greek world. Cultural unity was also fostered by the 'colonization', which led to numerous Greek settlements on the coasts of Asia Minor, Sicily, southern Italy, North Africa, and in the Black Sea area; contacts with indigenous populations reinforced the consciousness of an ethnic homogeneity based on language, religion, and customs. Colonization reacted to population growth, aiming at occupation of arable land. Newly created settlements became independent units. Constructing a new community implied a process of rationalization by which the order at home was also understood to be subject to change by deliberative decisions.

The City-State

The city-state (*polis*), first developed in colonization areas, became the typical form of organization in most (yet not all) parts of mainland Greece. It consisted of a fortified center on a hill, a residential town at the foot of a hill, and the agricultural hinterland. Its political unity was symbolized by a meeting place and communal temples. Collective decision making was formalized by the establishment of magistrates with defined competences and terms of office, councils with administrative functions, and assemblies, which took decisions about war and peace, and the administration of justice. For a long time, public functions could only be fulfilled by well-to-do people, but they had to take into account the opinions of the peasants who made up the bulk of the military.

The development toward formalized political structures overlapped with increasing tensions between rising and declining families of the social elite on the one hand, and between the aristocracy and the peasantry on the other. Considerable parts of the peasantry were in danger of indebtedness and finally enslavement. There were demands for cancellation of debts and redistribution of land. In several cities internal tensions led to the seizure of power by a tyrant, an aristocrat who controlled the city through a mixture of military force and patronage. Tyrants improved the condition of the peasantry, developed the infrastructures of the cities, and fostered communal identity by establishing central cults. Only later were tyrants considered to be exercising arbitrary rule

incompatible with the freedom of citizens. In certain cities, 'legislators' were entrusted with the formulation of rules that should stabilize the community.

At the end of the archaic age, a firm political structure was achieved in most cities, with considerable differences as to the degree of participation open to the bulk of the citizenry. Priests were responsible for the administration of cults and temples but did not dispose of political power. There were hundreds of autonomous *poleis*, most of them with a territory of 50–100 km^2 and some hundred or a few thousand (male, adult) citizens. There were, however, some exceptions, in particular Sparta and Athens.

Sparta

Sparta had embarked on an unparalleled expansion. During the second half of the eighth century BC, it conquered Laconia and Messenia, the southern part of the Peloponnese (a territory of *c*.8500 km^2). The subjugated populations were treated as helots, state serfs, or as *perioikoi* ('those who dwell about'), who lived in communities enjoying local autonomy but bound to follow Sparta's military leadership.

Uprisings among the Messenians in the late seventh century BC led to a total reorganization of Spartan society, which was achieved by a series of reforms probably completed about 550. The citizens (Spartiates, *c*.9000) received almost equal plots of land, which were tilled by the helots, whose products enabled the Spartiates to serve as full-time hoplites. Every Spartiate underwent a public education and afterward lived in a warrior community. Society had a uniquely militaristic character, distinguished by an austere lifestyle, which has provoked admiration as well as abhorrence (from antiquity until the present day). The political system was unique again: there was a hereditary kingship in the form of a dual monarchy (of unknown origin), the kings being the leaders of the army; home affairs were administered by five magistrates (*ephors*) appointed every year and a council of men over 60 elected for life (*gerousia*); decisions on foreign policy were taken by the assembly of the Spartiates. In spite of the ostentatious stress on equality within the Spartiates, there was probably an informally ruling elite, although internal politics remain obscure. During the sixth and fifth centuries BC, Sparta enjoyed remarkable stability, and the strength of its army was unequaled in the Greek world.

Athens

Athens was an exceptional case due to the fact that the whole territory of Attica (*c*.2500 km^2) had been formed into one political unit during the Dark Ages. Solon's legislation in 594 BC ensured social peace by establishing that the indigenous population should not be enslaved (which in turn created the demand for chattel slaves from outside), and by codifying large parts of Athenian law. The tyranny of the Peisistratus (561–510 BC) had further unifying effects. In 508/7 BC, Cleisthenes laid the foundations for a political system based on the participation of the citizenry. A reorganization of the subdivisions of the citizenry made newly constituted local communities (demes) the backbone of the system. A council of 500, consisting of members delegated by every deme, took on the preparation of assembly meetings and a number of administrative functions. The council's importance grew, especially when the older council, the Areopagus, consisting of life members, lost political influence in 462/1 BC. During the fifth century BC, appointment by lot was introduced for councilors and the majority of (several hundreds of) magistrates, as well as daily allowances for them and the hundreds of laymen who served as jurors in courts that decided in criminal and civil cases. Every honest citizen (above the age of 30) was considered as being able to undertake public functions, and, due to the great number of positions that had to be filled each year, a great part, if not the majority, of the 30 000–40 000 citizens must have acquired this experience.

All important decisions, especially on foreign affairs, were taken in the assembly, which met regularly. Leadership in the assembly was assumed by generals (strategoi) whose office (a board of 10) constituted an exception, since they were elected (not appointed by lot) and reelection was permitted. But only a few generals acquired a reputation through military achievement and rhetorical ability such that the assembly was likely to follow their lead. The trend toward democracy (a term coined in the late fifth century BC) was not accompanied by disputes of an ideological nature. The central position of the assembly was the result of a dramatic increase in military and diplomatic activities, as tremendous successes in this field fostered the legitimacy of the system.

The Greek World during the Fifth and Fourth Centuries BC

At the Battle of Marathon (490 BC), Athens achieved a sensational victory over Persian troops. In 480 BC, Persia launched an attack to conquer Greece and was beaten off by an alliance of most Greek states, Sparta taking the lead in the land battles, Athens in the maritime operations (with a great victory in the Salamis naval battle). Due to the strength of its fleet, Athens became the leading power in the Aegean Sea and built up a collective military system from which Sparta kept away. To continue fighting Persia an alliance (the Delian League) was established, with a great number of *poleis* financially contributing to the maintenance of the Athenian fleet. Athens enjoyed a period of prosperity and became a cultural center that in the era of Pericles (the leading statesman since the 440s BC) attracted intellectuals, artisans, and craftsmen from all over the Greek world. Athens' political and cultural hegemony materialized in the great temple buildings on the Acropolis and in the great festivals including the theater performances.

Athens' claim for leadership was based on its role in the Persian Wars but operations against Persia ceased after *c*.450 BC. Nevertheless, Athens did not allow any city to drop out of the alliance and showed an increasing tendency to interfere in the affairs of member states. Suspicion of Athenian ambitions to dominate all Greece arose in Sparta and its allies (especially Corinth) with the result that minor conflicts culminated in the 'Peloponnesian War' between Athens and Sparta and their respective allies. Both sides probably considered a final confrontation as inevitable and thought to fight a preventive war. The first period of warfare (431–421 BC) led to a confirmation of the status quo *ante*; the second one opened with

Athens' attack on Sicily (415–413 BC), a disaster for the Athenians, and ended finally in Sparta's total victory (404 BC), which was facilitated by Persian financial aid.

Sparta was not able to keep its hegemonic position for much longer. There were quarrels within the ruling group and, due to military losses and a concentration of landed property, the number of full citizens dwindled continually (down to 2500 Spartiates by 370 BC). Athens partly regained its naval strength, although a new alliance, founded in 377 BC, lasted for only two decades. For a short time, 371–362 BC, Thebes maintained the position of a leading military power. The instability of the power system was increased by Persia paying subsidies to varying allies, which led to a permanent reversal of coalitions. In the midfourth century BC, Macedonia, under King Philip II, entered the scene and tried to acquire a hegemonic position. Athens organized an anti-Macedonian alliance, but this was decisively beaten in 338 BC.

Athenian democracy had survived the defeat in the Peloponnesian War. The terror regime ('Thirty Tyrants') during the final phase of the war had definitively discredited any constitutional alternative. Democracy was restored in 403 BC. Fundamental criticism of political equality as articulated by Plato and Aristotle did not reflect even the opinion of the upper classes. Present scholarship is agreed that there was no decline in political participation. The growing tendency toward a separation of political and military leadership had certain negative effects on military efficiency, but the final defeat by Macedonia was due to the lack of sufficient resources. That Athens (and the other Greeks) should better have accepted a 'national' unity under Macedonian leadership (as was often said in nineteenth-century historiography) is an anachronistic idea, incompatible with the tradition of the autonomous city-state.

The Hellenistic Age

Macedonia did not formally abolish the autonomy of the *poleis*, but they were put under firm control and lost their capacity to pursue an independent foreign policy. Macedonia wanted to legitimize its leadership by embarking on retaliation against Persia. Philip's plan was realized by his successor Alexander ('the Great'). Heading an army of Macedonian and Greek troops, Alexander conquered Asia Minor, Syria, Egypt, Mesopotamia, and Iran with incredible speed (334–330 BC). Having declared the Panhellenic revenge war terminated, he dismissed the Greek contingents and started conquering the eastern parts of the Persian Empire with Macedonian and Iranian troops. In 326 BC, he reached the Indus Valley, but his intention to march to the 'edges of the world' failed as his exhausted troops refused. In his lifetime Alexander won a reputation for invincibility; his death in 323 BC, at the age of 33, made him a figure that was reflected upon by philosophers, historians, and rhetoricians. He became the prototype of a world conqueror later rulers would imitate, and a subject of popular romances, which would receive adaptations in the major languages of the Middle Ages. Modern historians often cannot resist following their own imaginations with respect to Alexander's 'final plans'.

Alexander took over the provincial system of the Persian Empire and considered himself a successor to the Persian Great King and the Pharaoh. His adoption of Persian court ceremonial and attempts to integrate Iranians into his elite units provoked alienation with his Macedonian troops. At the time of Alexander's death there was no successor. Control was taken over by several generals, who acted as governors in various parts of the empire. In a series of wars between them the unity of the empire was broken. Finally, around 280 BC, a new power structure consisting of Macedonia, Egypt under the Ptolemies, and Asia Minor, Syria, Mesopotamia, and parts of Iran under the Seleucids was established. Macedonian generals had succeeded in founding dynasties. In the Ptolemaic and Seleucid kingdoms leading positions in the army and bureaucracy were occupied by Macedonians and Greeks; Alexander's attempts to include native elites were not repeated. Greek-Macedonian and native populations lived under different legal orders. The system was kept together by the king, who was considered absolute ruler and overlord of the territory.

In spite of these limits of integration, there was a tendency toward assimilation of Greco-Macedonian and indigenous cultures, the spread of Greek as an international language, and the spread of the Greek lifestyle to cities and garrisons with Greek populations. Mercenaries, tradesmen, intellectuals, and artists enjoyed mobility throughout the Mediterranean and Near Eastern world. Alexandria in Egypt became a cultural center where the literary heritage of the fifth and fourth centuries BC was cultivated. These achievements of the 'Hellenistic age' (a nineteenth-century coinage) were still persisting when the Romans established their domination over the eastern parts of the Mediterranean world.

Early Rome and the Middle Republic

According to late Republican Roman tradition, the city of Rome was founded in 753 BC. Archaeological evidence suggests that settlements on the hills of the future city were established from the tenth to the eighth centuries BC. Formation of a city-state took place considerably later, and was due to strong Etruscan influences, although Rome was never dependant on an Etruscan city. The primordial kingdom, ever in need of cooperation with a landowner patriciate, was finally abolished around 500 BC. After that Rome was dominated by an aristocracy, which, however, changed its character. There were tensions between the patriciate and rising families, as well as peasant demands for land distributions and debt relief. The plebeians, i.e., all nonpatricians irrespective of their social status, combined in a separate organization with magistracy (tribunes of the people) and assembly to achieve access to political functions hitherto monopolized by the patriciate, and to secure the fulfillment of social demands. The 'Struggle of the Orders' is a somewhat dramatic term for a long process in which phases of conflict and reconciliation alternated. Piecemeal changes from the fifth to the early third centuries BC led to a political system with an equilibrium between magistrates, senate (a council of former magistrates sitting for life), and assemblies, which decided on war and peace, legislation, and the election of magistrates. Politics was determined by a nobility formed of the old patriciate and the plebeian elite who now had access to the magistracy. Office tenure for 1 year, collegiality in all magistracies, a hierarchy of magistracies to be

occupied in successive order, and the senate's overall control ensured a certain equality within the ruling class. Ordinary citizens participated through the assemblies. Voting, however, took place in units structured according to wealth and place of residence, which implied a blatant weight in favor of the propertied classes; initiative for legislation was the monopoly of magistrates. Roman assemblies were far from being democratic in the Greek sense. But their very existence and their function as a forum for communicating with the people ensured consensus within the citizenry. The nobility presented itself as a ruling class not indulging in conspicuous consumption, but absorbed by their duties in public office.

Constitutional development and territorial expansion mutually enforced each other; the peasantry serving in the army put pressure on the nobility, who in turn would fulfill their demands by land distributions in newly acquired territories. During the fourth and early third centuries BC, Rome won control first over the neighboring region of Latium, then over mid- and south Italy. This tremendous success was not only due to the strength of a highly disciplined military but also the result of a prudent policy of integration. Whole communities were admitted to Roman citizenship (with or without the right to vote); others gained a special status by which their members could opt for Roman citizenship. All others were treated as allies enjoying autonomy, their obligation to military support being rewarded by a share in the profits of war. Thus Rome could rely on indirect domination based on the cooperation of local elites.

Roman involvement in Southern Italy led to confrontation with the Carthaginians who occupied parts of Sicily. As a result of the First Punic War (264–241 BC) Rome established domination over Sicily, Sardinia, and Corsica. The second war with Carthage (218–201 BC) opened with catastrophic Roman defeats against Hannibal, who had invaded Italy from Spain. But finally the Romans were able to make Hannibal retreat to North Africa, where he was decisively beaten. In the end, Rome extended its domination over Spain. During the second century BC, Rome established its hegemony in wider parts of the Mediterranean area. The Macedonian state was destroyed in 168 BC, Carthage in 146 BC. The Seleucid position in Asia Minor was undermined in favor of the rise of Pergamum, which in 133 BC would fall to Rome.

Beginning with Sicily, newly acquired overseas territories were made provinces under permanent administration by Roman magistrates and 'promagistrates' (whose authority was prolonged for that purpose after the regular 1-year term). Governors relied on the cooperation of local elites and, as a rule, did not change internal administration structures, especially the systems of tax collection. There were, however, always complaints about exploitation by governors who impudently enriched themselves.

Rome's overseas expansion did not follow a blueprint, and actual decisions were not taken primarily with respect to the acquisition of resources. But a tendency toward 'imperialism' was inherent in a system in which magistrates had just a short period to gain prestige and wealth, and soldiers were accustomed to participate in booty. The more conquest, the more chances to 'defend' Roman interests; Rome's determination to impose its will upon other states was compatible with its self-representation as fighting only 'just wars'.

The Late Republic

The establishment of a large empire had unforeseen repercussions at home. Part of the vast resources available because of the conquests were invested in Italian land. Aristocrats acquired great landed estates, which were worked by slaves for the production of cash crops. Peasants were displaced and no longer met the property qualification for army service. This picture, going back to ancient narrative sources, represents a general trend, although it certainly needs qualification to take account of regional differences. The land reforms by Tiberius and Gaius Gracchus (133 BC and 123/2 BC, respectively) had only limited effects. Later the recruitment system was changed: volunteers without property qualification were enrolled; these soldiers showed special loyalty to their commanders, who in turn felt obliged to secure landed property for veterans. Sulla, who came to power (82 BC) after a series of civil wars (starting with the Italian allies' forcing their access to Roman citizenship) and the war with Mithridates of Pontus (who had attacked the Romans in Asia Minor and Greece), was the first to satisfy the soldiers' demands by large-scale confiscations in Italy. Sulla's attempt as *dictator* (traditionally only a short-time supreme commander) to strengthen the senate's authority was not really successful. It became necessary to entrust the great general Pompey with long-term command, but after his complete reorganization of Roman domination in Asia Minor and Syria (66–62 BC), Pompey could not arrange for the land allotments he had promised his soldiers. He engaged in a coalition with the ambitious Caesar, who used his provincial command in Gallia for a large-scale conquering campaign (58–51 BC). At last Pompey realigned with the senate to crush Caesar. Caesar won the ensuing 'civil war', which was fought all over the Roman Empire (49–45 BC). Being in sole power, Caesar was finally appointed dictator for life and the senate indulged in conferring extreme honors upon him.

Caesar's ostentatious disgust with aristocratic equality caused his murder by a group of senators (44 BC). His supposed 'final plans' have bedeviled generations of historians, but it seems doubtful whether he had had a clear-cut conception of a lasting reorganization of the political system. That a monarchical head of the empire was a historical necessity (as nineteenth-century scholarship put it in Hegelian terms) was an idea apparently not familiar to contemporaries, since military success over centuries had been achieved by an aristocracy. But the traditional republic could not be restored either, and after a further series of civil wars a new system emerged.

The Early Empire

The new order was established by Octavian, Caesar's personal heir, on whom the senate conferred the title 'Augustus' in 27 BC. Through cautious moves he constituted a monarchy in disguise, proclaimed as a 'restoration of the republic'. The emperor's formal authority consisted of a bundle of competencies derived from traditional magistracies, which allowed him to control all affairs. Continuance of magistracies and senate, and employment of senators as military commanders and governors ensured the loyalty of the senatorial elite, although the subsenatorial 'equestrian order' got a share in

administration. The emperor played the part of the patron to the capital's population by grain and cash distributions as well as public games. The problem of veteran settlement was solved by the purchase of land in Italy and overseas colonization. Provincial governors still relied on the cooperation of local elites, who were rewarded with legal privileges and offered chances for ascendance into the equestrian and senatorial ranks.

Such a summary of the 'principate' inaugurated by Augustus should not result in ignoring internal tensions. Keeping a republican facade demanded the self-restraints of rulers, but the lack of constitutional controls and, for example, the tendency of Eastern provinces to confer godlike honors on them made emperors show their true faces as autocratic rulers again and again. Provincial government was improved, but charges of corruption and maladministration against governors did not cease (and led to a great uprising in case of the Jews in AD 66–70). The absence of a law of succession reflected the tension between meritocratic and dynastic elements within the emperorship. Designation of a successor was not necessarily decisive – at the crucial moment, the senate, the imperial elite troops (praetorians), even the capital's population might be influential, or a provincial army might promote a pretender.

All in all, during the first and second centuries AD the empire enjoyed stability and prosperity, which was manifest in the sophistication of urban culture in all parts of the Roman world. Expansionist policy was given up in favor of territorial consolidation (although Britain was subjected to Roman rule throughout the first century AD). The majority of troops were stationed in fortified positions near the frontiers, especially the Rhine-Danube and Euphrates borders. During the third century AD, Rome came under increasing pressure by the Goths (who were invading the Danube area) and the new Sassanid dynasty in Persia. The situation was aggravated by instability on the throne: again and again army units proclaimed officers as emperors; most of them met violent deaths after a short time.

The Later Empire

After half a century of turmoil, Diocletian (AD 284–305) embarked on a comprehensive reform. It was based on devolving the imperial office on two senior and two junior emperors (as prospective successors), who would each have authority in a part of the empire but would govern jointly. The emperorship was symbolically enhanced by attributing a 'sacral' character to it and adopting oriental court ceremonial. However, regulated succession failed; from AD 305 onward various pretenders fought each other, until Constantine and Licinius could consolidate their positions in the West and East, respectively. Finally Constantine became sole ruler (AD 324).

The administrative reforms started by Diocletian were continued by Constantine. They included a reorganization of the army (now divided into field army and border troops) and of the provincial system (now including Italy), a separation of military and civil administration, and a new taxation system to meet the demands of the army and the enlarged governmental apparatus. These led to considerable consolidation although the intention to regulate more and more features of social life was partly defeated by corruption and maladministration, as well as the attempts of various social groups to escape the pressure put on them from above – local grandees sheltered their dependants from the state's demands and exploited them at the same time. Citizenship (since AD 212 extended to all inhabitants of the empire) lost its importance, as can be seen in the harsh penal system with distinctions according to the social status of citizens. Peasants, who had to take the burden of taxation, were reduced to a serflike status; craftsmen were bound to their occupation and place of residence. These points signify tendencies; they should, however, not be understood as covering the multifariousness of social reality. There were considerable differences between the various regions of the empire, and discrepancies between the increasing legislation and its enforcement. The political and societal order of the post-Diocletian era had manifestly changed in comparison with the early empire, but there were also remarkable elements of continuity, for example, with respect to higher education, local administration, and civil law. The traditional label of 'despotism' is not helpful.

A most important change happened with respect to the Christians. From the first century AD onward, Christianity had spread through all parts of the empire, and had taken on firm organization with the episcopal constitution. Christians who did not participate in the ruler cult were considered disloyal citizens. During the first two centuries AD, there were sporadic persecutions of Christians, mostly inspired by local communities. The Christians reacted, on the one hand, by praising martyrdom; on the other, by stressing the coincidence of Christ's coming and the Augustan reign of peace and (answering the charge of being a lower-class religion) by adopting 'pagan' literary culture. Since the midthird century AD, rulers had considered the restoration of state cults necessary for the preservation of the empire. That led to systematic persecutions under the reigns of Decius (AD 250), Valerian (AD 257), and Diocletian (AD 303); indictments of clerics, prohibition of church meetings, confiscation of holy scripts and church property were aimed at destroying the church's very existence. As a whole the measures failed and were stopped in AD 311. After AD 312, Constantine pursued a policy not only of toleration but also of actively promoting the church by donations, granting privileges to clerics, conceding a public role to bishops, promoting Christians in public offices, and finally giving the new capital Constantinople, founded at the site of Byzantium (AD 324), a Christian outlook. Constantine's religiosity and the motives of his policy remain obscure. The strengthening of imperial authority by a theology that declared the emperor God's vicar on earth had to be paid for with involvement in inner-church conflicts, especially the Arian schism, which was not to be solved by the theological compromise found at the Council of Nicaea (AD 325) under Constantine's personal lead. Finally, Theodosius declared Christianity the state religion (AD 380) and pagans, heretics, and Jews became subject to persecution. Bishops played an important role as the heads of cities, and some, especially in the West, in imperial politics as well. The increasing interweaving of state and church, however, fostered ascetic monasticism.

The division of the empire into Western and Eastern parts after the death of Theodosius (AD 395) would last, although the fiction of unity was never given up. For a number of reasons, the Eastern Empire did far better with the pressures from

migrating Germanic tribes, especially the Goths. The weakness of the Western Empire is symbolized by the Visigoths' sack of Rome (AD 410). The end of Western emperorship in AD 476 is traditionally taken as signifying the end of Antiquity. Whereas in the West, Germanic states were established, the East took up the Roman heritage (as symbolized by Justinian's codification of Roman Law, AD 534) and preserved its continuity through the Middle Ages.

The Legacy of Antiquity

In spite of manifold contacts with the oriental world, the political and cultural achievements of Greece and Rome were products of indigenous development. The notions of citizenship and individual rights, the ideas of republicanism and democracy, as well as universal emperorship, the continuity of Roman law and of the church (including the use of the Latin language), Christian preservation of 'pagan' literature and other traditions became formative for European culture through structural continuities and conscious renaissance. In this sense, it is still appropriate to treat Greco-Roman antiquity as a historical epoch of its own.

See also: Christianity, Origins of; Citizenship, Historical Development of; Classical Legacies; Democracy, History of; Democratic Theory; Historical Thought and Historiography: Classical Period (Especially Greece and Rome); Imperialism, History of; Renaissance; Warfare in History.

Bibliography

Brown, P.R.L., 1978. The Making of Late Antiquity. Harvard University Press, Cambridge, MA.

Brown, P.R.L., 1992. Power and Persuasion in Late Antiquity: Towards a Christian Empire. University of Wisconsin Press, Madison, WI.

Cambridge University Press, 1970–2005. The Cambridge Ancient History, 12 vols. Cambridge University Press, Cambridge, UK.

Gruen, E.S., 1984. The Hellenistic World and the Coming of Rome. University of California Press, Berkeley, CA.

Meier, C., 1980. Res publica Amissa. Eine Studie zu Verfassung und Geschichte der spten rmischen Republik. Suhrkamp, Frankfurt am Main, Germany.

Millar, F.G.B., 1977. The Emperor in the Roman World (31 BC–AD 337). Duckworth, London.

Momigliano, A.D., Schiavone, A. (Eds.), 1988–1993. Storia di Roma. Einaudi, Torino.

Nicolet, C., 1977. Rome et la conquête du monde méditerranéen 264–27 avant J.-C. I: Les structures de l'Italie romaine. Presses Universitaires de France, Paris.

Nicolet, C. (Ed.), 1978. Rome et la conquête du monde méditerranéen 264–27 avant J.-C. II: Genèse d'un empire. Presses Universitaires de France, Paris.

Osborne, R., 1996. Greece in the Making 1200–479 BC. Routledge, London.

Will, E., 1972. Le monde grec et l'orient. I: Le Ve siècle. Presses Universitaires de France, Paris.

Will, E., Moss, C., Goukowsky, P., 1975. Le monde grec et l'orient. II: Le IVe siècle et l'époque héllenistique. Presses Universitaires de France, Paris.

Animal Rights in Research and Research Application

Lesley J Rogers, School of Science and Technology, University of New England, Armidale, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of animal rights is discussed and distinguished from animal welfare. A brief history of attitudes to animal sentience and welfare is followed by contemporary activity in protecting animals used in research. Differences between countries in the legal protection of animals used in research in the Western world are covered, as are the movements active in campaigning for animal rights and/or welfare. Finally, debates about which species are covered by legal protection in research are discussed, drawing attention to the current inconsistencies regarding which invertebrates should be included in ethical guidelines for research and which need not be so protected.

The Concept of Animal Rights

Animal Rights has two different meanings. One meaning refers to specific law-based protections of nonhuman animals and hence to animal rights activity in formulating fundamental legal protections of animals, as well as monitoring adherence to these clear legal requirements. The other meaning refers more generally, or some say more vaguely, to the moral rights of animals, and advocates of this position argue for moral protections of nonhuman animals along the same lines as the rights of humans (i.e., the rights to life, liberty, and freedom from torture), but not necessarily treating animals and humans in the same way for all purposes (see Francione, 2000; Armstrong and Botzler, 2003). In essence, extending rights to animals means that animals have an inherent right to live according to their species characteristics/requirements, free from harm, abuse, and exploitation. In other words, nonhuman animals (hereafter referred to simply as animals) have value in their own right completely separate from their usefulness to humans.

Some organizations concerned with animal welfare and rights adopt variations of these polarities, depending on their specific views and the state of play considered possible in improving the treatment of animals. There are a number of ongoing debates on animal rights (see articles in Sunstein and Nussbaum, 2004), many stemming from difficulties arising where biological knowledge meets the law.

Animal Rights versus Animal Welfare

There is a fundamental difference between the animal rights and animal welfare positions. As Francione and Garner point out in their book *The Animal Rights Debate: Abolition or Regulation?* published in 2010, the welfare approach is utilitarian and regulationist. Within this framework, animals are considered to be of lesser moral worth than humans and hence it is acceptable to use them for human purposes as long as they are treated 'humanely.' To those who favor the welfare position, the moral issue is *how* we use animals, not *whether* we use them. Also, advocates of this position argue that death in itself is not harmful to animals. Although regulations must be in place to ensure that death of animals killed for research, to eat, or for other reasons is not painful, death itself is not considered to be a particular issue. This means that, according to this position, animals can be used in experimental research and in testing pharmacological and pharmaceutical substances, provided the benefits of the research outweigh the costs in terms of animal suffering. Here one notes that the benefits are mostly for humans, whereas the suffering is of animals.

On the other hand, the position of animal rights considers death in itself as being 'a harm' for all sentient animals since they must be aware of life versus death. This raises a question about which species are sentient and, therefore, can we draw a line separating those to whom rights are extended from those not worth such privileges, or is all life sentient to varying extents? The animal rights position is, in its purest form, abolitionist and considers that animals should not, under any circumstances, be exploited by humans. This includes animals used in research, issues of farming animals for food, using animals for sport and other forms of human entertainment, keeping animals as companions, and killing animals considered to be pests. Different groups or organizations tend to give emphasis to one or two of these areas. However, to quote Andrew Knight, author of *The Costs and Benefits of Animal Experiments* 2011, "Animal experimentation is arguably the single most contentious issue in the wider debate concerning the rights of humans to use animals."

History

Animal welfare and animal rights ideas developed in the Western world against anthropocentric ideas and ethics. The superiority of humans and their right to use animals to their own advantage was part of the Christian tradition from its inception (humans alone being created in the image of God) (see Scully, 2002; article number 5 in Singer, 2006). There are other cultures and religions that believe in the sanctity and equality of all life (e.g., Hinduism and Buddhism) and hence exercise animal rights to various degrees, the strictest interpretation being that of the Jains in India, who wear a cloth over the mouth to prevent death of small organisms by inhalation.

In the Western world, the ideas of the French philosopher René Descartes (1596–1650) dominated and animals were seen as mechanical automatons unable to experience pain or any other feeling. Descartes and his followers reasoned that the

cries of animals subjected to vivisection, at that time without analgesics or anesthesia, were merely the creaking and groaning of the mechanical system, as in the spring-driven clockwork devices made by humans, and so could be ignored. This perception of animals put an end to any positions that had until then considered animal welfare and it opened the door to all manner of experimentation that we now see as brutality. Descartes' philosophy is said by some to have sparked modern anthropocentrism.

Attitudes to animals changed during the Enlightenment. By 1800, animals had become to represent new standards of truth and morals (Senior, 2008). Humans admired animals and through them sought to regain their own animality. Contemporary understanding of 'rights' stems from this period, be the term applied to human rights, the rights of slaves and prisoners, or the rights of animals.

Then came the theories of Charles Darwin (1809–82) on evolution and the continuity between animals and humans in both structure and mental function (*Origin of the Species*, 1859, and *The Descent of Man*, 1871). Publication of these ideas was met with heated debate since they blurred the line between humans and animals, particularly between humans and primates, and pockets of that debate persist even today: Darwin's theory of evolution is still opposed by the Creationist Movement. Darwin's theory of evolution made a revolutionary change to the study of animals. For the first time, it made clear that research on animals was relevant to understanding humans. At the same time, it suggested that animals share emotions with humans, to varying degrees (see *The Expression of Emotions in Man and Animals*, 1872). This meant that the difference between humans and animals was a matter of degree, not of kind. Hence, animals were considered to experience pain and a range of other emotions including fear, pleasure, and misery, as well as higher cognitive abilities (memory, problem-solving, tool using) and the rudiments of mind.

Contemporary Activity in Protecting Animals in Research

Policy on the use of animals in research has changed substantially in recent years, particularly where nonhuman apes are involved. The Great Ape Project was founded in 1993 as an international organization of primatologists and researchers in related fields to campaign for extension of basic legal rights to the great apes. The book *The Great Ape Project* edited by Cavalieri and Singer argued that great apes share many cognitive abilities with humans (e.g., self-awareness, rationality) and so should be given rights.

In 1999, the New Zealand government passed an Animal Welfare Act banning experimental research on nonhuman hominids. The Austrian Parliament, in 2005, banned experimental procedures on apes unless they were in the interests of the individual ape. Beginning in 2008 in Austria, there has been a law suit fighting to get a chimpanzee called Matthew Hiasl Pan declared 'a person' and so achieve for him the same rights as humans, including a pension (see Knight, 2011). This case was rejected recently, but it continues now as an appeal to the European Court of Human Rights.

In the United States, following a public controversy and a congressional request made in 2010, the U.S. National Institutes of Health (NIH) commissioned formation of an 'Institute of Medicine Committee on the Use of Chimpanzees in Biomedical and Behavioral Research.' This committee reported in 2011 that chimpanzees are largely unnecessary for biomedical research and it recommended guidelines for any research using chimpanzees. In 2013, the NIH adopted most of the recommendations and announced that it would release most of the federally owned or sponsored chimpanzees to sanctuaries in the United States. Some 110 chimpanzees will be retired from research, with a projected 5-year time-frame for transferring them to life in more natural conditions than in the laboratory. Fifty chimpanzees will be retained in the NIH laboratories for potential future use in research.

Protection of animals used in research has increasingly become policy in several countries. In 1997–98, laws of the European Union referred to animals as sentient beings, and in 2002, the German constitution was amended to protect animals at the most fundamental level.

These changes were stimulated by changing public attitudes and by activities of a range of organizations with an interest in curtailing research using animals (see below). They want it either abolished or continued under strictly enforced guidelines. In response to early activity by animal rights advocates and changing public opinions, universities in the UK and Australia, for example, formed 'Animal Ethics' committees that assess all research proposals and issue licenses on individual basis. In the United States, 'Americans for Medical Progress Educational Foundation' was founded in 1991, in order to increase the public's understanding of humane use of animals in research. Scientists are now more open about their research procedures and results and tend to engage more actively with animal welfare and, in some cases, with animal rights groups. Both sides of the debate seem to see importance in education of the public about research in the life sciences and its costs and benefits.

In many countries, attitudes of the public to the use of animals in research have shifted quite markedly in recent years: for example, a 2013 survey in the United States of 1000 people found that 41% considered medical testing on animals as morally wrong, this being an increase of 12% since 2001 (results presented at the American Association for the Advancement of Science, February, 2014, listed below).

Along with these developments, we have seen the inclusion of courses in Animal Law in the United States (Harvard University course commenced in 2000), the UK, Australia, and Spain (e.g., Masters in Animal Law and Society at Universidad Autonoma de Barcelona), and the appearance of new journals and academic societies focused on animal ethics (e.g., the Animals and Society Institute in the United States, which is an independent research and educational institution promoting knowledge in human–animal studies and action to protect animals through adoption of ethical, compassionate public policy). New peer-reviewed journals reporting research and ideas in this field have also been launched (e.g., *Animal Law*, *Sloth*, *Animals and Society* and the *Journal for Critical Animal Studies*). Special issues of other specialist journals are now being devoted to articles discussing the ethics of research on animals (e.g., *Theoretical Medicine and Bioethics*, vol. 35 (2),

published in 2014) and a new focus of enquiry on how we treat animals has emerged in departments of sociology, psychology, and even gender studies (see Taylor and Twine, 2014).

Conferences devoted entirely to animal welfare and rights are now held regularly (e.g., Minding Animals International Inc. has held triannual conferences since 2009).

Guidelines for the Use of Animals in Research

A guiding principle of research using animals is that of the three Rs: meaning to *reduce* the number of animals used in research and still achieve meaningful results, where possible to *replace* use of animals by other techniques (e.g., tissue culture), and to *refine* the methods used in order to minimize pain and suffering. Knight (2011) argues in favor of adding two more Rs: viz., reuse (recycling of research animals) and rehabilitation.

Guidelines for committees formed to assess research in universities and other research institutions vary between countries and even between states in some countries (e.g., Australia).

In most places, there are regular steps taken to upgrade guidelines and legal requirements. For example, in New South Wales, Australia, the 'Animal Research Review Panel,' formed in 1986, is the overarching organization monitoring research on animals and issues accreditation to universities and related organizations. It also acts to review guidelines for housing and conducting research on different species, taking into account relevant knowledge and research on the various species. The National Health and Medical Research Council in Australia issues guidelines that include all nonhuman vertebrates and, as added in 2004, invertebrate cephalopods, such as octopus and squid. In 2010, the European Union passed a directive requiring member states to follow certain guidelines when conducting research and including cephalopods but, contrary to scientific advice, excluding other invertebrates including decapod crustaceans (lobsters, crabs, etc.), spiders, and social insects (*Official Journal of the European Union*, 22 September 2010). The latter species were not included despite sufficient evidence indicating that they can feel pain and experience suffering, distress, and lasting harm. These abilities to feel pain and suffer are considered probable, since it has been demonstrated that these invertebrate species are capable of performing tasks that require complex cognition. This raises the difficult issue of exactly what species are to be included in guidelines and legislation and what are not, as discussed further below.

Animal Rights Movements

Organizations concerned with animal welfare and animal rights currently extend their attention and activities more broadly than to the field of research, although use and abuse of animals in research is also part of their platform. Not all of them can be listed here, but the following are pertinent examples.

Animal Liberation formed in Australia in the late 1970s not only with an emphasis on animals used in factory farming, but also with an interest in animal rights in research. Its activities were an extension of the book *Animal Liberation* by the Australian philosopher Peter Singer, first published in 1975. Then and now, the goal of Animal Liberation is to "end suffering of exploited and confined animals, through legislation, consumer advocacy, action, and humane education." Members of Animal Liberation believe that animals have a right to live free from human intervention, use, and abuse. Incidentally, while Singer's publications, especially his book *Animal Liberation*, have reached a wide audience and led to much activity for animal rights, his arguments have drawn criticism from many of his academic peers on the grounds of rather poor academic quality and, by featuring sentience as the sole criterion of the need for protection of rights, allowing argument for sacrifice or use in research of humans lacking sentience (e.g., anencephalic neonates or comatose patients).

Animal Liberation Front (ALF) is radically abolitionist with a manifesto of total liberation of animals by any means necessary, which can include violence such as arson of animal laboratories and even kidnapping, as it did so in 1999. Then the journalist Graham Hall, who had made an undercover documentary of the violent methods of ALF, was kidnapped.

People for the Ethical Treatment of Animals (PETA), formed in 1980 in the United States, arose out of the Silver Spring monkey case, Maryland, USA. This case began in 1981 with an undercover investigation of monkeys used in research at the Institute of Behavioral Research. PETA's documentation of the cruelty to which these primates were being subjected led to the first criminal conviction of an animal researcher and the first confiscation of animals from a laboratory. It also led to additions to the Animal Welfare Act in the United States. PETA is now is a very large and active organization for animal rights that focuses on four areas in which it considers the largest numbers of animals suffer most intensely for the longest periods of time: in factory farms, in the clothing trade, in laboratories, and in the entertainment industry. PETA keeps to nonviolent methods to achieve its goals although these methods may not always be legal. PETA now has international affiliates in the UK, France, Germany, the Netherlands, India, Australia, and Asia.

Voiceless in Australia focuses its finances and activities on factory farming and the kangaroo industry. This organization funds a prize for research aimed at replacing animals in research.

Animals Australia campaigns against the live export of cattle, factory farming, 'puppy factories,' and, of relevance here, its overall policy is against the use of animals in experimentation, especially in the pharmaceutical industry and in teaching.

Mercy for Animals is an animal rights organization in the United States, founded in 1999, focused on preventing cruelty to farm animals and promoting diet that does not include animals. The use of animals in research is not part of their current activities. They conduct undercover investigations primarily of battery farm practices and have had successes in improving conditions in some cases.

One of the more recent developments is the establishment of a National Museum of Animals and Society in Los Angeles, USA, chiefly exploring the bond between humans and animals and educating the public on the cognitive and emotional abilities of animals (e.g., cognition of domestic chickens). It aims to be a repository of materials and writings from both the welfare- and right-based movements.

What Animals Are Included?

There is no consensus on which species of nonhuman animals should be covered by welfare guidelines and legislation, or should be included in the fight for animal rights (see Wise, 2002). Inclusion of species depends ultimately on which ones can experience pain, but this is not the only criterion. Intelligence, consciousness, and self-awareness are also used or implied as characteristics of species deserving rights and protections. The expansion of interest in methods of assessing cognitive function in animals has thrown these latter qualities into the limelight. Avant-guard research in the perceptual and cognitive abilities of different species has been important in drawing attention to which species should be afforded protection and rights and, as mentioned above, it has brought about changes, such as the inclusion of cephalopods under guidelines or legislation previously limited to vertebrate species. However, since research on these aspects of animal behavior is in its relative infancy, some of the changes made are piecemeal and biased toward the few species that have been tested for their cognitive ability. As explained in some detail by Rogers and Kaplan (2006), a number of behavioral attributes have been assessed in some species, including the abilities of following the direction of another's eye gaze, communicating intentionally, thinking of objects that have disappeared from sight, making and using tools, caching food and remembering from where to retrieve it later, theory of mind, and the concept of self, to list just some of the main indicators of higher cognitive ability (for examples of research on cognition in animals, see Wasserman and Zentall, 2006; Rogers and Kaplan, 2004).

Species that have not yet been tested and shown to have any of these cognitive abilities may be allowed protection if other closely related species have shown these exceptional abilities (e.g., all cephalopods although only one or two species of octopus or squid have been tested). Others may miss out simply because they have not been tested or have 'failed' the tests for reasons to do with experimental design or the particular subjects selected for testing. In other words, the process can be flawed and the outcomes problematic: indeed, inclusion of some species and not others may worsen the treatment of those not included.

Very few would disagree with the fact that a boundary has to be drawn, but there is much debate about where. Some would extend rights only to apes, others to all primates, yet others to all mammals; but, recent research on higher cognition in birds (see Kaplan, 2014) demonstrates that some avian species have cognitive abilities equivalent to those of primates, even apes. Furthermore, recent research on bees shows that they have more complex cognitive abilities than previously considered possible. In summary, research on the cognitive abilities in animals is an active focus of interest and it is leading to revolutionary changes in our understanding of and attitudes to animals.

Currently, however, apart from the recent inclusions of cephalopods discussed above, no ethical guidelines for research on animals include invertebrates and no case is being made to extend rights to other invertebrates. Moreover, vertebrate species that are covered by laws or guidelines concerning research on animals varies from country to country. In the United States, the Animal Welfare Act (as of November 2013: see Animal Care Blue Book and Animal Welfare Information Center Web site) defines 'animal' as any live or dead dog, cat, monkey, guinea pig, hamster, or rabbit but excludes birds, rats of the genus *Rattus*, and mice of the genus *Mus* bred for use in research. In the UK, Germany, and Australia, at least, all of the latter species are covered by the legislation on the use of animals in research.

The Federation of European Laboratory Animal Science Associations represents views of the laboratory animal science community in Europe and issues guidelines that are not mandatory but aimed at keeping up with changes based on knowledge and community views.

There is a general perception that laws and guidelines relating to the use of animals in research need continual revision and modification to encompass new scientific information about the ability of different species not only to perceive pain, but also to suffer stress and distress. The gray area where we should draw the line between those species deserving protection and those not requiring it is broadening and will continue to do so as research provides broader and more reliable evidence of cognition and species differences (see Rogers and Kaplan, 2006). In future, it may well be recognized that not all species should be afforded the same rights but instead that species-relevant legislation should be applied.

See also: Bioethics: Philosophical Aspects; Consequentialism (Including Utilitarianism); Ethical Dilemmas, Research, and Access to Treatment; Ethics and Values; Experiment in Science and Technology Studies; Human Rights, Anthropology of; Objectivity of Research: Ethical Aspects; Research Ethics: Research.

Bibliography

Armstrong, S.J., Botzler (Eds.), 2003. Animal Ethics Reader. Routledge, London.
Cavalieri, P., Singer, P. (Eds.), 1993. The Great Ape Project: Equality Beyond Humanity. Fourth Estate Publishing, London.
Darwin, C., 1890. The Origin of the Species by Means of Natural Selection, sixth ed. John Murray, London.
Darwin, C., 1904. Expression of Emotions in Man and Animals. John Murray, London.
Darwin, C., 1981. The Descent of Man, and Selection in Relation to Sex. Princeton University Press, NJ. Facsimile from 1871 ed. held in the Firestone Library, Princeton University.
Francione, G.L., 2000. Introduction to Animal Rights. Temple University Press, Philadelphia.
Francione, G.L., Garner, R., 2010. The Animal Rights Debate: Abolition or Regulation? Columbia University Press, New York.
Kaplan, G., 2014. Bird Minds. CSIRO Press, Melbourne.
Knight, A., 2011. The Costs and Benefits of Animal Experiments. Palgrave Macmillan, Oxford.
Rogers, L.J., Kaplan, G. (Eds.), 2004. Comparative Vertebrate Cognition: Are Primates Superior to Non-primates?. Kluwer Academic/Plenum Publishers, New York.
Rogers, L.J., Kaplan, G., 2006. Think or be damned: the problematic case of higher cognition in animals and legislation for animal welfare. Animal Law 12, 151–191.
Scully, M.S., 2002. Dominion: The Power of Man, the Suffering of Animals, and the Call to Mercy. St Martin's Griffin, New York.
Senior, M. (Ed.), 2008. A Cultural History of Animals in the Age of Enlightenment. Bloomsbury, London.
Singer, P., 1976. Animal Liberation: A New Ethics for the Treatment of Animals. Jonathan Cape, London.
Singer, P. (Ed.), 2006. In Defense of Animals: The Second Wave. Blackwell, Oxford.

Sunstein, C.R., Nussbaum, M.C. (Eds.), 2004. Animal Rights: Current Debates and New Directions. Oxford University Press, Oxford.
Taylor, N., Twine, R., 2014. The Rise of Critical Animal Studies. Routledge, Oxford, UK.
Wasserman, E.A., Zentall, T.R. (Eds.), 2006. Comparative Cognition: Experimental Explorations of Animal Intelligence. Oxford University Press, Oxford.
Wise, S.M., 2002. Drawing the Line: Science and the Case for Animal Rights. Perseus Books, Cambridge, MA.

Relevant Websites

www.eurekalert.org/pub_releases/2014-02/pfte-nss021314.php – American Association for the Advancement of Science, February 2014.
www.animalsandsociety.org/ – Animals and Society Institute.
http://www.animalsaustralia.org – Animals Australia.
www.animal-lib.org.au – Animal Liberation.
www.animalliberationfront.com – Animal Liberation Front.
www.nal.uasda.gov/awic – Animal Welfare Information Center (USA).
www.felasa.eu – Federation of European Laboratory Animal Science Associations.
www.mercyforanimals.org – Mercy for Animals.
www.mindinganimals.com – Minding Animals International Inc.
www.museumofanimals.org – National Museum of Animals and Society.
www.peta.org – People for the Ethical Treatment of Animals.
www.voiceless.org.au – Voiceless.

Annales School

Eamon O'Flaherty, University College Dublin, Dublin, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Annales group of historians, including the founders of the journal Lucien Febvre and Marc Bloch, were responsible for a historiographical revolution, which challenged and then transformed historical writing in France and, from the 1970s onwards, had an international impact. The germ of the Annales challenge to established historiography in France grew out of a commitment by the founders to a widening of the field of historical research and an engagement with the human sciences. Over three generations between the 1960s and the end of the century, Annales moved through a number of clearly definable phases during which it continued to foster methodological innovation and multidisciplinary engagement, while remaining committed to a rigorous historical approach to evidence.

I

The historiographical revolution that took place in France between the 1930s and the 1960s is indelibly associated with the journal *Annales d'histoire économique et sociale*, founded in Strasbourg in 1929 by Marc Bloch and Lucien Febvre. The journal quickly moved its base to Paris and continued publication until 1939. During the war and the occupation of France, it appeared as *Annales d'histoire sociale* and *Mélanges d'histoire sociale*. It was effectively refounded in 1944 as *Annales: histoires, sociétés, civilizations* and, since 1994 has been titled *Annales: histoire, sciences sociales*. Throughout this period, the journal has been the center of a movement in historical writing, which sought to transform history by opening it up to the concepts and developments in other branches of the human sciences, including geography, sociology, and anthropology. At the core of the wide variety of practices and approaches associated with the Annales school was a view of history, which sought the widest possible perspective on historical change. This necessitated a broad understanding of the complex range of factors, which contribute to historical change including psychological, cultural, economic, and environmental factors. The question of human agency is central to this concern, but in the Annales tradition, human agency could only be understood by studying the collective elements in human consciousness and activity, against which any analysis of individual action needed to be set. A corollary of this was a broadening of the chronological and methodological approach to historical research. To some extent, this involved the expansion of the subject matter of historical enquiry and a new approach to the subject matter of historical research. A historical practice primarily concerned with the study of power and power elites gave way to a new emphasis on hitherto neglected areas of historical human experience, where the newer developments in other areas of the human sciences in the early twentieth century seemed to offer far more interesting and incisive perspectives. Thus, a reformulation of the domain of historical research and a new emphasis on research as collective as well as an individual procedure, focused on clearly identified problems and themes, was a central part of the methodology of the school. But as well as opening the field up to engagement with a variety of new approaches, the *Annales* also brought about an important change in the way in which historians conceived of historical time and periodization. Again, the reevaluation of the role of human agency was central here, particularly given the way in which the successive generations associated with the journal expanded the chronological limits of historical study to evaluate phenomena and processes from the perspective of longer-term changes, which transcended individual consciousness or agency. In seeking to understand deep-seated processes behind historical change, the *Annales* historians were interested in identifying underlying structures, which shaped profound change in human societies in ways that were invisible to individual human agents or to contemporary historical actors. This concern with deep structures was partly shaped by the engagement with economic history and the long-term cycles identifiable from the serial study of economic data such as prices. The result of these latter developments brought quantitative methods to the center of the *Annales* in the middle decades of the twentieth century, identifiable with the leadership of Ernest Labrousse and Fernand Braudel. The emphasis on quantitative methods, serial history, and the study of deep underlying structures of change and immobility over long periods of time represented an evolution from the concerns of Febvre and Bloch in the first generation, and also coincided with the dominance of the school in the French historical establishment. The dominance of economic and statistical approaches, over the psychological and cultural elements of the initial program, revealed an implicit tension between different views of what was primary and secondary in terms of historical causation and social history. The study of mentality, conceived in the widest sense as a history of collective representations and of the evolution of individual consciousness in a social and historical context, had been present from the beginning as a core part of the *Annales* program, however overshadowed by quantitative history. From the 1960s, however the inclusiveness of the totalizing approach, characteristic of the school's founders, and of its ambitions in the postwar era, was clearly still producing a constructive engagement between the study of physical and environmental structures and the analysis of cultural forms and collective mentalities. A variety of regional studies, most famously Le Roy Ladurie's on Languedoc, bore witness to a further shift in the development of the school. To some extent, this reflected the long-standing commitment to the project of a 'total history,' in which even the most austere

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61146-7

structural and quantitative methodologies were approached in the belief that these were ultimately a basis for the understanding of the whole of human experience, including mental and cultural experiences and structures, rather than a form of determinism. Thus, history could be conceived as manifesting itself at a variety of levels, 'from the cellar to the attic,' i.e., from fundamental environmental, and economic constraints to the quintessentially human forms of mental and social organization and, ultimately collective and individual human experience. Of course this extension involved a substantial degree of refinement of the conceptual apparatus for analyzing and describing cultural and mental change. In this respect, the Annales's continuing engagement with theoretical developments across the human sciences in the postwar decades also played a part in the refinement and expansion of the range of study. Thus historical anthropology, the history of mentalities, and cultural history became the most impressive aspects of the contribution of the third generation of historians associated with the school. In all these areas, one can also point to congruent developments in historical approaches to the human sciences in postwar France, which paralleled the work of historians of the *Annales*. Toward the end of the century, even some of the most characteristic elements of the *Annales* approach were called into question by the practice of the third generation. Microhistory and case studies offered a contrast to the previous emphasis on the collective and the anonymous, while the resurgence of narrative history from the 1980s seemed to fly in the face of the reaction against *histoire historisante*, which had united the new history and the other human sciences, especially sociology, in the 1920s.

II

Febvre and Bloch were reacting against the historical positivism associated with Gabriel Monod, Charles Seignebos, and the *Revue Historique*, founded in 1876, but the French historiographical tradition was much wider and, since the Enlightenment, had produced philosophical histories, which sought to probe the underlying cultural and social factors of historical change. Voltaire's *Essai sur les moeurs* sought to place study of human behavior at the center of philosophical history. In the nineteenth century, a number of influential writers, including François Guizot, Alexis de Toqueville, and Jules Michelet produced wide-ranging analyses of the relationship between cultural and political change, which, in the case of the first two, owed much to Edmund Burke's development of a theory of society as the product of an essential historical process of organic institutional and social development. Michelet's breadth of interests ranged from cultural history to the history of the common people, and included the assertion that history must also be about the people and not just about power or institutions. Nevertheless, the influence of the scientistic mode of historical research associated with Leopold Von Ranke and the German school was evident in France after the defeat of 1870. History in the Third Republic was dominated by two main characteristics, one scientific, the other national. The first involved an admiration for the rigor of the German historical school, which Monod praised highly for its 'publication of documents, critique of sources, its patient elucidation of all aspects of history' (*Revue Historique*, I (1878), 315–316) and lay behind much of the academic organization of historical research led by Ernest Lavisse in the last decades of the nineteenth century. In the first issue of the *Revue Historique*, Monod proclaimed that the journal would be 'a collection of positive science and free discussion; moreover, it will not depart from the realm of fact and will be closed to political and philosophical theories' (*Revue Historique*, I (1878), 315–316). Although Monod showed himself open to a wider conception of historical knowledge held by other positivist historians, a more rigorously defined definition of historical science, which seemed to eschew most forms of intellectualism or generalization was evident in the *Introduction aux études historiques* published by Charles-Victor Langlois and Charles Seignobos in 1898. The great monument of this generation of historians was the multivolume history of France conceived by Lavisse in 1890 and published in 17 volumes between 1901 and 1911. The project was highly influential partly because of the academic credentials of the authors, but also because it provided a history that underpinned the political unity of the Third Republic. "Lavisse … created the myth of a nation-state, which came into being between the time of the Gauls and the Merovingians, was consolidated by the administrative changes and military conquests of the Capetian monarchy and culminated in an ideal constitution: the democratic, moderate, and centralizing Republic" (Bourdé and Martin, 1989: 209).

Bloch and Febvre emerged out of a reaction against this approach to history, which was increasingly criticized in the early decades of the twentieth century as limiting and partial. They were not alone in this critical approach to the reigning historical orthodoxy. Emile Durkheim criticized the focus on particular events as opposed to larger processes in the journal *L'Année sociologique* in 1896. But, wider perspectives were also being canvassed by historians around this time. In 1896, Henri Berr founded the *Revue de synthèse* in Paris bringing together a broad spectrum of scholars and intellectuals including psychologists, sociologists, economists, and historians, including Lucien Febvre. In his 1893 thesis *L'Avenir de la philosophie: esquisse d'une synthèse des connaissances fondée sur l'histoire* (The future of philosophy: outline of a synthesis of knowledge founded on history), and afterward, Berr saw historical synthesis as capable of providing a unifying and organizing element in all the sciences. Berr's regular seminars at the *Centre de synthèse* in Paris offered a model for the collective projects launched by Febvre and Bloch in *Annales* in the 1930s, although Berr's project was more theoretical in scope than the *Annales*. Febvre's first work was begun at the École Normale Supérieure in Paris where he encountered the geographer Paul Vidal de la Blache, founder of *Annales de géographie*. Febvre's first work was his thesis on *Philippe II et la Franche Comté* (1911) a regional monograph, which placed strong emphasis on the geographical, religious, and social history of the region. In *La Terre et l'évolution humaine* (1922), Febvre developed his ideas on historical geography and the interaction of humanity and the physical environment, 'seeking to understand the role in human history of a number of specifically geographical factors – distance, space, position.' Here Febvre signaled a concern with the spatial dimension of history, which was central to many of the future concerns of the school. During the 1920s, Febvre moved to the University of Strasbourg where he

was part of a brilliant group of scholars from many disciplines who reshaped the university in the aftermath of the First World War. Strasbourg was the scene of many fruitful exchanges between historians, sociologists, and anthropologists during the 1920s. Here also Febvre began his friendship with Marc Bloch. Bloch had also attended the École Normale Superieure and, like Febvre, had first completed a regional monograph, on the Île de France, published in 1913. Febvre and Bloch produced a number of major works in the 1920s and 1930s and both were appointed to chairs in Paris, Febvre at the Collège de France in 1933 and Bloch at the Sorbonne in 1936. In 1929, the two collaborated in founding *Annales d'histoire économique et sociale*. Febvre seems to have played the leading organizational role in the journal during the 1930s and took over completely when Bloch, who was Jewish, was excluded from academic life during the occupation. But both made ground-breaking contributions, both in the journal and in their books, during the interwar period. Two books in particular stand out as classics which, in separate ways, demonstrated the new approach to the history of mentalities, which was at the heart of their contribution. In *Les Rois thaumaturges* (1924) Bloch studied the phenomenon of royal healing in France over a long time frame, borrowing the concept of collective representations from sociology to analyze the belief system surrounding thaumaturgic kingship in the medieval and early modern periods. The result was an innovative study of the historical sociology of religious belief, which could perhaps be regarded as the first example of the history of mentalities, which was to be so prominent from the 1960s. Lucien Febvre's *Problem of Unbelief in the Sixteenth Century*, while primarily a study of Rabelais, was really addressing a more fundamental problem. Febvre disagreed profoundly with the standard current interpretation, which presented Rabelais as a free thinker and rationalist. Febvre's reinterpretation centered on stripping away anachronistic views of Rabelais's beliefs and reinstating the context of sixteenth-century belief and dissent. In arguing that Rabelais's mental universe could only be reconstructed within the terms and limits of possibility of the language and mental equipment of a sixteenth-century mind, Febvre gave an object lesson in the enlarged possibilities of an approach based on the reconstruction of the mental world of the past, including, in the case of Rabelais, the impossibility of atheism within the mental structures of the sixteenth-century mind.

The history of the journal in its early years reflected the concerns of the directors and especially Febvre, who contributed over 900 articles, reviews, and notices between 1929 and 1948. The battle against *histoire historisante* continued to be waged by Febvre in book reviews, which was also an important medium for pursuing the pluridisciplinarity of the journal. *Annales* also reflected Bloch's belief in the importance of delineating historical problems as the principal object of study, hence transcending the limitations of a fixed periodization. The sense of a relativity of historical time, with different rates of change and different time spans applying to different phenomena was clear, even if the subject matter was more familiarly historical compared to the *semaines de synthèse*. The first collective enquiry looked at the comparative history of land records across Europe. Business archives and price lists were studied in 1930. In 1935, there was a special issue on 'technologies, history, and life' and in 1936 an article 'les noblesses, reconnaissance générale' announced the launch of a collective enquiry on the nobility. During the war, when the journal changed its name, two new *enquêtes* were launched on alimentary habits and on associations in *ancien régime* society. The list alone conveys something of the excitement and the sense of possibility involved in extending the frontiers of research in so many new directions. Indeed Bloch's idea that the historian's craft involved the mastery of a vast range of disciplines made collective enquiry a practical necessity, as well as a reflection of his belief in the essential unity of the human sciences: 'there is only one human science in time, which strives incessantly to unite the study of the dead to that of the living.' The collective enquiries also reflected the journal's engagement with the contemporary world, and the explicit desire to use the new approach to history to cast light on some of the problems of the contemporary world in the aftermath of the Crash of 1929.

Although Febvre's work often concentrated on the study of individuals, albeit with a attention to deep contextualism, which anticipated later developments in intellectual history beyond France, he seems to have been more favorable to the increasing concentration on economic history, which grew in prominence in the 1930s and 1940s. Febvre saw the price surveys undertaken by François Simiand as a model for serial quantitative history, which would transcend the subjectivity of the documentary evidence and grasp patterns and recurring phenomena, which would reveal the deep structures of history. Bloch joined the resistance and was executed in 1944. He was succeeded at the chair of economic history at the Sorbonne by Ernest Labrousse, and this marked a turn away from the focus on the history of mentalities toward socioeconomic serial history. Febvre became more powerful after the war when in 1947 he set up the VIth section of the École Pratique des Hautes Études and director of the Centre des Recherches Historiques. The *Annales* school now had an institutional presence at the heart of French academic life, and its dominance was assured for the next four decades. Although never one of the directors, Ernest Labrousse was extremely influential in the direction taken by the work of the Annales historians from the 1940s to the 1960s. Labrousse's *Outline of the movement of prices and revenues in France* (1933) and his *Crisis of the French economy* (1944) provided a structure and a methodology, which was adopted by a whole generation of French historians. Pierre Chaunu called them 'the breviary of my generation' and saw them as more influential than even the work of Braudel. Labrousse introduced the idea of long-term trends and short-term crises and produced an interpretation of the long-eighteenth century and the French Revolution in terms of a century-long trend of rising prices, followed by a decade-long intercycle of contraction and a bad harvest of 1788 triggering the revolutionary events. Such a deterministic schema, clearly linked to Labrousse's Marxism, was very persuasive in the ideological climate of the 1950s, when many of the leading historians were associated with the left and, combined with the support of Braudel and the obvious affinities of Braudel's work with Labrousse's structural model, account for the leading place of quantitative history in the 1950s and 1960s.

More central than Labrousse, and far more wide-ranging and accessible as a historian, Fernand Braudel was Febvre's successor as leader of the Annales. Born in 1902, Braudel spent most of the period 1923–32 in Algeria, apart from 2 years in

Brazil. On Febvre's advice, he turned a conventional thesis on Mediterranean politics in the reign of Philip II into a study of the Mediterranean at the time of Philip II. The whole project took several decades to complete, the first draft written in a prisoner-of-war camp in Lubeck and the first edition published in 1949. The result was a monumental piece of scholarship, but also an innovative 'geohistory' with a radically new conception of historical time. Braudel wrote of 'les temps de l'histoire,' emphasizing the plurality of historical time. In this, he was following the Annales tradition of matching historical processes and phenomena to their specific durations, but also going much farther in presenting a total history of the Mediterranean as a series of overlapping or layered temporalities. In addition, Braudel wished to construct a geohistory, which would be attentive to space as well as time, recognizing the conditions of human society as intimately related to the landscape and emerging from the landscape in a series of overlapping developments corresponding to three major rates of temporal change. He was critical of the event-oriented history, which Febvre and Bloch had rebelled against, but also of a geography that suffered from analogous limitations when it neglected the environmental and temporal dimensions of human experience. He defined a geohistory as a task, which would 'construct a truly representative human geography; to make the geographers pay more attention to time (which should be fairly easy) and (what will be rather less easy) to make the historians more concerned about space' (Braudel, 1989a: ii, 295). The three times of history correspond to the three sections of *The Mediterranean*. The final section of the book deals with the time frame of conventional history, the short term, the *temps court*, where Braudel describes the political structures, wars, territories, and politics of the period. Braudel is less interested in these 'surface agitations,' 'oscillations,' 'surface disturbances, crests of foam that the tides of history carry on their strong backs.' These and many other comments reinforce Braude's hierarchical view of the relative unimportance of the short term, which is also, he points out, the time frame of the individual, but not of humanity itself. Braudel's distinction of times is also a distinction of rates of change. The short term is the time frame of rapid change, whereas the middle term is slower and rhythmical, 'a structural history, even a social history of groups and movements' what he also calls collective destinies. To this category belong networks of communication, systems of exchange, demographic trends, economic cycles of the kind described by Labrousse, and also the territorial empires of Spain and the Ottomans. Braudel's first level, corresponding to the opening section of the book, is the category of immobile, virtually timeless time. The processes of change at this level take place at a rate incommensurate with individual consciousness and yet they are, for Braudel, the most profound and inescapable forces in history. Understanding them involves an appreciation of human society as a phenomenon existing in and conditioned by the physical landscape and the environment – geology, soil, vegetation, climate, landscape, topography.

Braudel's concept of history as largely shaped by processes that are larger than humanity is fundamentally deterministic. It may be that his pessimistic view on the limitations of human agency and freedom ('a tiny island, a prison') may have been influenced by the fact that he was himself a prisoner when he first wrote the book. Unlike the possibilism of Vidal de la Blache, Braudel does not see the landscape as something that can be endowed with meaning, rather he seeks to see the human actors as shaped by forces emanating from the environment in ways that they cannot determine. Of course it is arguable that this consciousness is itself a means of humanizing the impersonal. Related to this determinism is the fact that culture is relatively neglected in *The Mediterranean*. Yet, despite criticisms, Braudel's massive work had an enormous impact on historical understanding. His assertion of the importance of the *longue durée* as the time frame, which enabled us to understand the relative importance (and unimportance) of human agency in the shorter time frames was nevertheless capable of supporting sophisticated analyses of historical processes in all three. His determinism was by no means reductive, but was rather an affirmation of the complexity of historical process, characterized by the interaction of forces operating at different levels of temporality and causation. It was not the great length of the work but its breadth which accounts for its enduring influence – the revelation of how the aspiration to a total history could provide a more rounded picture of the interaction between human society and the physical environment by expanding the field of historical enquiry in spatial and temporal dimensions. As well as taking over increasingly from Febvre in the direction of the school, through the VIth Section, the journal and the Collège de France, including inspiring and supervising a new generation of historians, Braudel produced two further major books, *Civilization and Capitalism*, a three-part work completed in 1980 and *The Identity of France*, an unfinished history of his own country published after his death. The first part of *Civilization and Capitalism*, *The Structures of Everyday Life*, again saw Braudel make a major contribution to extending the scope of history by focusing on the historical structures and forms of material culture, viewed in the perspectives of the *longue durée* and across continents. The study of material culture is now almost commonplace, indicating just how innovative and influential the book was. *The Wheels of Commerce* and *The Perspective of the World* studied the development of mechanisms of production and exchange, and then proceeded to describe the succession of dominant economic powers in the early modern period from Venice to the rise of Britain in the late-seventeenth century. Here again, Braudel's work was characterized by a combination of meditative insights, daring initiatives of interpretation and analysis based on the extended perspectives of the *longue durée*, and the incorporation of perspectives from across the human sciences, which cast fresh light on the evidence of the sources. Of course, the range of these sources had been dramatically increased by the fundamental reappraisal of the nature and significance of historical evidence associated with Annales. Epistemologically, the relationship between historical sources and historical knowledge was the same for Annales as for the scientistic historians of the previous century. What was fundamentally different was the range of sources discussed, the methodology of their use, the recognition of the need to apply the conceptual sophistication of other disciplines in the human sciences, and, thanks to Braudel and others, the demonstration that this would profoundly change our temporal and spatial understanding of the past. Braudel's example inspired Pierre Chaunu's monumental work on Seville and the Atlantic in the

early modern period, which revealed the power of the battery of methodologies assembled by the Annales.

To concentrate on the leading figures does not do justice to the breadth of the Annales achievement, or to its commanding role in French historical writing in the second half of the twentieth century. If the first and second generations can be conveniently assessed via Febvre, Bloch, and Braudel, the leading figure of the third generation, not least in terms of the international impact and popularity of his work, is Emmanuel Le Roy Ladurie. The 'Labrousse moment' and the dominance of Braudel were reflected in a preponderance of regional monographs based on serial quantitative methodology. Outstanding examples were Pierre Goubert's work on Beauvaisis in the early modern period, which was faithful to the Braudelian/Labroussian example, but also regional studies by Paul Bois and Maurice Agulhon, which restored the interest in cultural forms and mentalities, which Bloch and Febvre had displayed in the interwar period. Le Roy Ladurie's thesis, *The Peasants of Languedoc*, was also a regional monograph, which displayed the strengths of quantitative history in the *longue durée*, especially in the analysis of the interaction of Malthusian demographic structures and the rural economy, but where the standard model described a kind of homeostatic system maintained by short-term crises, Le Roy Ladurie's Languedoc traced a great cycle of recovery, stability, and decline from the mid-fifteenth to the end of the seventeenth century. More significantly, Le Roy Ladurie's analysis, while broadly comparable to the structural model of Goubert's Beauvaisis, did not describe an immobile history dominated by the impersonal interaction of structure and conjuncture, but interwove this with an account of the development of peasant society, which incorporated evidence of popular consciousness, including illuminating case studies of heresy and social unrest of the kind he later expanded into the best-selling *Montaillou*, a historical anthropology of a Cathar village, and *Carnival at Romans*, a case study of the breakdown of social relations in an urban setting, which became a microcosm of early modern urban mentalities and behavior. The range of Le Roy Ladurie's work was in some ways Braudelian, as in his study of the history of climate since the year 1000, comparable to the global perspectives of Braudel's *Capitalism*, but also represented a rediscovery of the specific and the short term, which had been in danger of becoming obscured by an excessively programmatic focus on the quantitative and the impersonal. The case studies reflected an awareness of further waves of affinity in the human sciences in cultural anthropology and cultural theory associated with writers like Pierre Bourdieu, Marcel Mauss, and the slow impact of Norbert Elias. What Michel Vovelle half-jokingly called the bulimia of historians ensured that the danger in all schools of creating a sterile orthodoxy was subverted by the constant evidence of creativity in addressing problems and sources, and a continued willingness to listen to others, including other historians. Was the fertile interaction between Annales and non-Annales historians from the 1960s a reflection of the success of the school in redefining the objects and nature of historical enquiry? Certainly, the thoroughly original contributions of Philippe Ariès and Michel Foucault on death, childhood, madness, prisons, and the epistemic structures of historical systems of thought all posed a challenge to Annales and yet a validation of the notion that history would proceed by identifying and illuminating important areas of research and enquiry, which were the subject of important reflections across the human sciences. For the Annales school, the response was to grapple with the questions of how historians could contribute to this kind of historical anthropology, what kind of sources could be identified, and how should they be studied. Michel Vovelle's study of attitudes to death, the afterlife, and religion, *Baroque Piety and Dechristianisation* (1973) owed much to the work of the Catholic religious sociologist Gabriel Le Bras, but Vovelle's background was as a Marxist and follower of Labrousse's approach to serial history. Vovelle's study of changing religious attitudes in eighteenth century Provence was based on a study of 30 000 wills – a forbidding kind of documentation whose potential he dramatically revealed and in doing so also showed that an austere quantitative approach could be fruitfully applied to the history of mentalities, which once again occupied a central place in the Annales, and could build on its sophisticated tradition of intellectual innovation, which did not sacrifice historical rigor. Collective enquiries were organized into attitudes to death and popular literacy from the 1970s onward and these in turn provoked a further refinement of the problematics of analyzing cultural phenomena through historical sources. Adopting perspectives from cultural theory, Roger Chartier's work on cultural history suggests a further refinement of the development of the study of mentalities by emphasizing the primacy of cultural and symbolic constructions. Peter Burke, who was one of the earliest British historians to appreciate the significance of the Annales, and who has been a constant innovator himself, and André Burguière, an Annales insider who has written an intellectual history of the movement, see the revival of narrative, the reemergence of mentality (as historical anthropology), and the revival of interest in politics as key themes in the development of Annales since the late 1980s. Burke, writing in 1990, saw the divergence of themes and approaches as evidence of continuing vitality but queried whether it was still possible to speak of an Annales school because of its achievement in transforming history internationally as much as because of the divergences of the fourth generation: 'the movement is dissolving, in part as a result of its success' (107). Both Burke and Burguière, writing in 2006, agreed on the centrality of the return to mentalities for the relationship between history and the human sciences, and outlined the challenges faced by historians in responding to some of the demands of cultural theory and anthropology, still facing the problem confronted by the founders of *Annales* between the individual agent and the collective consciousness.

See also: Bloch, Mark Léopold Benjamin (1886–1944); Burke, Peter (1937–); Historical Geography; Historical Thought and Historiography: Current Trends; Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; Social Geography.

Bibliography

Berr, Henri, 1899. L'Avenir de la philosophie: esquisse d'une synthèse des connaissances fondée sur l'histoire. Hachette, Paris.

Bloch, Marc, 1924. Les Rois thaumaturges. Istra, Strasbourg.

Bourdé, Guy, Martin, Hervé, 1989. Les Écoles Historiques. Seuil, Paris.

Braudel, Fernand, 1989a. La Méditerranée et le monde méditerranéenne a l'époque de Philippe II. Librairie Générale, Paris.
Braudel, Fernand, 1981 (Sian Reynolds, Trans.). Civilization and Capitalism, 15th–18th Century, 3 vols. Harper and Row, New York.
Braudel, Fernand, 1989b. The Identity of France (Sian Reynolds, Trans.). Fontana, London.
Burguière, Andre, 2009. The Annales School: An Intellectual History. Cornell UP, Ithaca (Jane Marie Todd, Trans.).
Burke, Peter, 1990. The French Historical Revolution: The Annales School, 1929–89. Polity, Cambridge.
Chartier, Roger, 1989. Cultural History: Between Practices and Representations. Polity, Cambridge.
Clark, Stuart, 1995. The *annales* historians'. In: Quentin Skinner (Ed.), The Return of Grand Theory in the Human Sciences. CUP, Cambridge, pp. 177–198.
De Certeau, Michel, 1975. L'écriture de l'histoire. Gallimard, Paris.
Cullen, Louis, 2005. Labrousse, the *Annales* school and *histoire sans* frontieres. Journal of European Economic History 34, 309–350.
Febvre, Lucien, 1911. Philippe II et la Franche Comté: étude d'histoire politique, réligieuse et sociale. Champion, Paris.
Febvre, Lucien, 1982. The Problem of Unbelief in the Sixteenth Century: The Religion of Rabelais. Harvard UP, Cambridge, USA.
Febvre, Lucien, 1922. La Terre et l'évolution humaine: introduction géographique a l'histoire. La Renaissance du Livre, Paris.
Fink, Carole, 1991. Marc Bloch: A Life in History. UP, Cambridge.
Goody, Jack, 1990. Marc Bloch and historical anthropology. In: Atsma, H., Burgiere, A. (Eds.), Marc Bloch Aujourd'hui: histoire comparée et sciences socialesl. EHESS, Paris.
Hexter, J.H., 1979. Fernand Braudel and the monde Braudellien…. In: Hexter, J.H. (Ed.), On Historians: Reappraisals of Some of the Makers of Modern History. Harvard UP, Cambridge, USA, pp. 61–147.
Labrousse, Ernest, 1944. La Crise de l'économie francaise a la fin de l'ancien régime. PUF, Paris.
Labrousse, Ernest, 1933. Esquisse du mouvement des prix et des revenues en France, 2 vols. Dalloz, Paris.
Vovelle, Michel, 1985. Idéologies et mentalités. La Découverte, Paris.
Vovelle, Michel, 1973. Piété baroque et déchristianisation en Provence au XVIIIe siècle. Plon, Paris.

Anomie

Philipe Besnard[†] Observatoire Sociologique du Changement, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp. 510–513, © 2001, Elsevier Ltd., with revisions made by the Editor, Douglas Massey.

Abstract

The word 'anomie' comes from the ancient Greek *anomia*, meaning absence of rules, norms, or laws, and is still sometimes used in this highly general sense. The term 'anomie' was reinvented by Jean-Marie Guyau, but it was Emile Durkheim who inscribed the term in the vocabulary of sociology. In *The Division of Labor in Society* (1893), Durkheim called 'anomie' an abnormal form of division of labor, defining it as the absence or insufficiency of regulations necessary to ensure cooperation between specialized social functions. In *Suicide* (1897), anomie (whether economic or conjugal) refers to the insufficient social regulation of individual aspirations: the indetermination of the object of the desire leads to frustration. Anomie underwent American naturalization in the 1930s at Harvard University. Merton's 'Social Structure and Anomie' metamorphosed the concept: Durkheim's anomie refers to the lack of restriction on goals; Merton's anomie refers to restriction on means. This semantic revolution attained its full effect between 1955 and 1970, in sample surveys that used attitude scales designed to measure the psychological counterpart of social anomie. At that time 'anomie' came to be considered the sociological concept *par excellence*, despite (or because of) the semantic confusion surrounding it.

'Anomie' comes from the ancient Greek *anomia*, meaning absence of rules, norms, or laws (the adjective form *anomos* was more common). The word is still sometimes used in this highly general sense. In Greek thought it had negative connotations of disorder, inequity, and impiety; and when it reappeared briefly as 'anomy' in English theological literature of the seventeenth century (Orru, 1987), its meaning was somewhat similar to the ancient Greek usage of the word. The term came back at the end of the nineteenth century in the writings of French philosophers and sociologists. The French spelling imposed itself in the language of English sociology despite attempts to anglicize it to 'anomy.' 'Anomie' was particularly popular among American sociologists in the 1960s. It is one of the few lexical inventions in sociology, perhaps the only one, which has been specific to the sociological community. But sociologists used it with differing and even contradictory meanings – sometimes without any meaning at all – and its career eventually ended in total confusion.

Durkheim's Concept

'Anomie' was reinvented by Jean-Marie Guyau, a French philosopher with a sociological bent, in two books: *Esquisse d'une morale sans obligation ni sanction*, published in 1885, and *L'Irréligion de l'avenir*, published in 1887. Guyau opposed *anomie* to *autonomie* (Kantian autonomy). Because of the progressive individualization of the beliefs and criteria for ethical conduct, the morality of the future would be not only autonomous but anomic – determined by no universal law. For Guyau this process was both ineluctable and desirable.

After Guyau, the word appeared occasionally in late nineteenth century, in the writings of French philosophers and sociologists, such as Gabriel Tarde. But it was Emile Durkheim who would inscribe 'anomie' in the vocabulary of sociology, the new discipline he intended to found.

Durkheim could not accept Guyau's anarchizing individualism since for him every moral fact consisted of a sanctioned rule of behavior. He therefore considered anomie a pathological phenomenon. In *De la division du travail social* (*The Division of Labor in Society*), his doctoral thesis defended and published in 1893, Durkheim called anomie an abnormal form of the division of labor, defining it as the absence or insufficiency of the regulation necessary to ensure cooperation between different specialized social functions. As examples of anomie he cited economic crises, the antagonism between capitalists and workers, and science's loss of unity due to its increasing specialization. In all three of these cases, what was missing were continuous contacts between individuals performing different social roles, who did not perceive that they were participating in a common undertaking. At the end of his chapter on anomic form, Durkheim discussed another kind of pathological characteristic of industrial societies: the alienation of the worker performing an overspecialized task. This explains why a number of Durkheim's readers understood job meaninglessness to be an integral part of anomie. In fact, they introduced a 'parasite' connotation into his concept. Such alienation is not only different from anomie, it is in many ways its opposite.

Durkheim further developed the theme of anomie in *Le Suicide* (*Suicide*) (1897), defining anomic suicide as that resulting from insufficient social regulation of individual aspirations. In this connection, he discussed economic anomie, caused by a period of economic expansion; and conjugal or sexual anomie, due to the introduction and spread of divorce. In both these cases, an opening up of the horizon of the possible brings about the indetermination of the object of desire, or unlimited desire (which amounts to the same thing). This leads to frustration. The anomie that follows a passing crisis may be acute, but Durkheim was especially interested in chronic, even institutionalized anomie. This type of anomie, the 'morbid desire for the infinite,' was an unavoidable counterpart of modern industrial society, residing simultaneously in its value system, e.g., the doctrine of constant

[†] Deceased; updated and refreshed by Douglas Massey.

progress; its institutions, e.g., divorce law; and its functioning, e.g., competition in an ever-expanding market. However, along with 'progressive' anomie, Durkheim also evoked a 'regressive' type, which actually pertained to 'fatalism.' This last notion, which Durkheim did not really develop, refers to the impossibility of internalizing rules considered unacceptable because they were unjust or excessively repressive. Here, desires and aspirations run up against new norms, deemed illegitimate; there is a closing down of possibilities. This situation is in fact the opposite of progressive – true – anomie, where the wide, open horizon of the possible leads to unlimited desires. Fatalism, then, is the opposite of anomie, just as altruism is the opposite of egoism (Durkheim's terms for the other types of suicide). In fact, this bipolar theory of social regulation of aspirations was not clearly laid out by Durkheim, and this no doubt explains the concept's peculiar later career. After Durkheim, 'anomie' underwent a complete semantic revolution.

Anomie was not a permanent theme in Durkheim's thought; he was more concerned with the question of social integration. Indeed, both word and concept disappeared from his work after 1901 and they are not to be found anywhere in the works of his disciples or collaborators in the 'French school of sociology' and in *L'Année sociologique*. In *Les Causes du suicide*, a purported continuation of Durkheim's study of suicide published in 1930, Maurice Halbwachs confirmed Durkheim's results on many points; but he either neglected or rejected everything that could possibly pertain to Durkheim's theory of anomie, he made no reference to conjugal anomie, and he refuted Durkheim's hypothesis of a progressive economic anomie.

Yet it was in the sociology of suicide, particularly in American studies of the subject, that Durkheim's notion of anomie would endure for a time, following the 1951 English translation of *Le Suicide*. In this connection, mention should be made of Henry and Short's (1954) study, in which they sought to apply Durkheim's concept of anomie to their analysis of the effect of the economic cycle on suicide frequency. The largely negative results of this and other tests of Durkheim's progressive anomie hypothesis did not encourage sociologists to explore this avenue of research further. Conjugal anomie, meanwhile, was left almost entirely aside in suicide studies, which gave primacy to Durkheim's egoistic suicide, reflecting a preference for his theory of integration over the less developed one of regulation.

The Americanization of Anomie

As with sociological studies of suicide, the future of 'anomie' was linked to American sociology's reception of Durkheim's work, particularly *Suicide*. Interest in Durkheim began to manifest itself only in the 1930s. With the notable exception of Parsons (1937), however, little attention was paid at first to 'anomie.' The term made little headway at the University of Chicago, where a tradition of research on social disorganization had developed in the 1920s, though Louis Wirth (1938) famously argued that the size, scale, and heterogeneity of cities would inevitably produce anomie in urban society. It was at Harvard University that 'anomie' underwent its American naturalization. Used by Elton Mayo (1933), Talcott Parsons (1937), and Robert K. Merton (1938), it appeared as a new label – and possibly a new theoretical framework – for the study of social disorganization and deviant behavior; and enabled the new department of sociology at Harvard to position itself in relation to the old Columbia–Chicago struggle for intellectual supremacy in the discipline. This strategic aspect may be seen in the young Harvard theoreticians' use of the French spelling (sociologists at Columbia, like Durkheim's translators, spelled it 'anomy'). Distinguishing themselves from the American tradition and appropriating the continental European one, that of Durkheim, Weber, and Pareto, the Harvard sociologists spelt the term as 'anomie.'

Merton's successful *Social Structure and Anomie* in its various versions (1938, 1949, 1957, 1964) both made the term familiar and metamorphosed the concept. In fact, Merton never clearly defined what he meant by anomie. We can nonetheless say that Merton's theory ultimately identified anomie as the contradiction between the culturally defined goal of success and the individual's lack of access to legitimate means of achieving this goal. To resolve this contradiction, the individual may engage in deviant behavior, one form of which is using illegitimate means. Merton called this mode of adaptation 'innovation.'

Merton's theory of anomie is fairly equivocal, but it was understood and used as a theory of differential access to the means to success, one which predicted that delinquent behavior was most likely in the least privileged social strata. This theory, which reached its apogee in the mid-1960s, has been considered the core of the 'Strain Theory,' a major sociological approach to deviance. It should be noted that Merton's theory was only rarely tested in empirical research on delinquency and that when it was, the results were generally negative.

The Mertonian concept of anomie – to the extent that its content can be identified and as it has been generally understood – is in many ways not only different from Durkheim's anomie but opposed to it. The opposition was not perceived at the time because Merton never related his own use of the term to Durkheim's. Durkheim's anomie is about the lack of restriction on goals; Merton's anomie refers to restriction on means. In Merton's theory, goals are given, defined, even prescribed by the cultural system, whereas for Durkheim the central feature of anomie is indeterminacy of goals. Durkheim's anomic individual is uncertain of what he should do because the horizon of possibilities is so open; whereas for Merton that individual, with the clearest of ideas of the objective to be reached, finds the possibilities for success closed down.

This semantic revolution attained its full effect in American sociology between 1955 and 1970; There were numerous sample surveys using attitude scales designed to measure the psychological counterpart of social anomie. This research in turn was based on theoretical essays in which anomie had appeared for the first time outside Harvard: at the end of the 1940s sociologists and political scientists such as David Riesman, Robert MacIver, and Sebastian De Grazia had psychologized the concept, using the word 'anomie' to designate individual feelings of anxiety, insecurity, and distrust. These feelings figured in 'anomie scales,' meant to measure the degree of pessimism in world views, feelings of having no control over one's situation, and the renunciation of hope. The most famous of these was developed by Leo Srole in 1956 and called the 'anomia scale.' In Durkheim's thought, these phenomena were

characteristic of fatalism – the opposite of anomie. Anomie had become interchangeable with the notion of alienation, which sociologists were trying to apply in attitude scales; normlessness had been associated with powerlessness. All these empirical studies using attitude scales were based on the (mistaken) postulate of a single unified concept of anomie: the myth of a continuous tradition of research, with Durkheim as precursor and Merton as prophet.

Anomie scales were contemporaneous with the other main usages of the term. As we have seen, it was between the mid-1950s and the late 1960s that 'anomie' was most heavily used, in its Durkheimian meaning, in the sociology of suicide; and with its Mertonian meaning (mainly) in the sociology of deviant behavior. That contradictory meanings of one and the same term should have been successful at the same time shows that the function and use of the word 'anomie' in sociology was more decorative than cognitive.

In the 1960s, anomie came to be considered the sociological concept par excellence, despite (or because of) the semantic confusion surrounding it. In the 1968 edition of the *International Encyclopedia of the Social Sciences*, Talcott Parsons explained that anomie was one of the few concepts truly central to contemporary social science. At that time the term had reached the peak of its glory. Because 'anomie' made possible a kind of bridge between references to classic theoretical sociology and sample surveys using attitudes scales purporting to test hypotheses drawn from those theories, it served as the emblem of the dominant research practice of the time (opposed to the ecological analysis and fieldwork favored by the Chicago School).

Though the word 'anomie' has lost much of its attraction for sociologists, some still use it rather vaguely to mean deregulation or disorder. It would be preferable to restrict usage of the term to Durkheim's meaning of an excessive growth in aspirations or expectations. This concept refers to a paradox present in Tocqueville as well as in the notion of relative deprivation, for which Raymond Boudon (1977), constructed a model: when individual objective conditions improve, collective subjective dissatisfaction may increase.

Recent Departures

Whereas Wirth (1938) drew on Durkheim and other classical theorists to predict rising anomie as society urbanized, Fischer (1975) argued that what Wirth observed in cities was not deviance stemming from a state of anomie, but a high rate of unconventional behavior stemming from the proliferation and intensification of subcultures within large, dense, and diverse urban settings. Subsequent evidence indeed showed that people engaging in unconventional or 'deviant' behaviors were not suffering from an absence of values, but in fact were well integrated into the values of specific subcultures that validated and sustained their behavior – behavior that only appeared 'deviant' relative to mainstream actors but not to members of the subculture (Fischer, 1995). Thus urban society was not disorganized, but very well organized into different, partially overlapping, and partially competing subcultures.

The latest incarnation of Merton's conceptualization of anomie as stemming from a gap between legitimate goals (e.g., earning money) and access to the legitimate means to achieve them (e.g., access to high-paying jobs) is Messner and Rosenfeld's (2006, 2007) theory of institutional anomie, which argues that in contemporary societies the market has been allowed to dominate over other institutional spheres, such as the family and the polity, to create an imbalance in values that leads to a particularly criminogenic form of anomie, a view that finds considerable support in recent work. Baumer and Gustafson (2007) show that crime rates are significantly greater in areas where a strong commitment to monetary success coexists with a weak commitment to legitimate means of getting ahead. Burkatzki (2008) finds that the willingness of people to subordinate the common welfare for individual economic success rises as actors become more integrated into the market sphere; and Bjerregaard and Cochran (2008) demonstrate that homicide is strongly predicted by a three-way interaction between GDP (gross domestic product), economic freedom, and income inequality. In general, anomie rises during periods of rapid social change and is less likely among those of higher socioeconomic status, who are buffered from the negative effects of social, economic, and political change (Zhao and Cao, 2010; Legge et al., 2008).

See also: Alienation, Sociology of; Alienation: Psychosociological Tradition; Control, Social; Control, Social; Delinquency, Sociology of; Durkheim, Emile (1858–1917); Integration, Social; Labor, Division of; Norms; Parsons, Talcott (1902–79); Solidarity, Sociology of; Suicide, Sociology of.

Bibliography

Baumer, E.P., Gustafson, R., 2007. Social organization and instrumental crime: assessing the empirical validity of classic and contemporary anomie theories. Criminology 45 (3), 617–663.

Besnard, P., 1986. The Americanization of anomie at Harvard. Knowledge and Society 6, 41–53.

Besnard, P., 1987. L'anomie, ses usages et ses fonctions dans la discipline sociologique depuis Durkheim. Presses Universitaires de France, Paris.

Besnard, P., 1990. Merton in search of anomie. In: Clark, J., Modgil, C., Modgil, S. (Eds.), Robert K. Merton: Consensus and Controversy. Falmer Press, London.

Besnard, P., 1993. Anomie and fatalism in Durkheim's theory of regulation. In: Turner, S.P. (Ed.), Emile Durkheim: Sociologist and Moralist. Routledge, London.

Bjerregaard, B., Cochran, J.K., 2008. Want amid plenty: developing and testing a cross-national measure of anomie. International Journal of Conflict and Violence 2 (2), 182–193.

Boudon, R., 1977. Effets pervers et ordre social, first ed. Presses Univesitaires de France, Paris.

Boudon, R., 1982. The Unintended Consequences of Social Action. St. Martin's Press, New York.

Burkatzki, E., 2008. Is the market eroding moral norms? A micro-analytical validation of some ideas of anomie theory. International Journal of Conflict and Violence 2 (2), 268–287.

Durkheim, E., 1893. De la division du travail social. Alcan, Paris.

Durkheim, E., 1897. Le suicide. Etude de sociologie. Alcan, Paris.

Durkheim, E., 1933. The Division of Labor in Society. Free Press of Glencoe, New York.

Durkheim, E., 1951. Suicide. A Study in Sociology. Free Press, Glencoe, IL.

Fischer, C.S., 1975. Toward a subcultural theory of urbanism. American Journal of Sociology 80, 1319–1341.

Fischer, C.S., 1995. The subcultural theory of urbanism: a twentieth-year assessment. American Journal of Sociology 101, 543–577.

Guyau, J.M., 1885. Esquisse D'une Morale Sans Obligation Ni Sanction. F. Alcan, Paris.

Henry, A.F., Short Jr., J.F., 1954. Suicide and Homicide: Some Economic, Sociological, and Psychological Aspects of Aggression. Free Press, Glencoe, IL.

Legge, S., Davidov, E., Schmidt, P., 2008. Social structural effects on the level and development of the individual experience of anomie. International Journal of Conflict and Violence 2 (2), 248–267.
Mayo, E., 1933. The Human Problems of an Industrial Civilization. Macmillan, New York.
Merton, R.K., 1938. Social structure and anomie. American Sociological Review 3, 672–682.
Merton, R.K., 1949. Social structure and anomie. In: Merton, R.K. (Ed.), Social Theory and Social Structure. Free Press, Glencoe, IL.
Merton, R.K., 1957. Continuities in the theory of social structure and anomie. In: Merton, R.K. (Ed.), Social Theory and Social Structure. Free Press, Glencoe, IL.
Merton, R.K., 1964. Anomie, anomia and social interaction. In: Clinard, M.B. (Ed.), Anomie and Deviant Behavior. Free Press of Glencoe, New York.
Messner, S.F., Rosenfeld, R., 2006. The present and future of institutional anomie theory. In: Cullen, Francis T. (Ed.), Taking Stock: The Status of Criminological Theory, Advances in Criminological Theory, vol. 15. Transaction Books, Piscataway, NY.
Messner, S.F., Rosenfeld, R., 2007. Crime and the American Dream, fourth ed. Wadsworth, Belmont, CA.
Orru, M., 1987. Anomie: History and Meanings. Allen & Unwin, Boston.
Parsons, T., 1937. The Structure of Social Action: A Study in Social Theory with Special Reference to a Group of Recent European Writers. McGraw-Hill, New York.
Parsons, T., 1968. Durkheim, Emile. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences. MacMillan, New York.
Srole, L., 1956. Social integration and certain corollaries: an explanatory study. American Sociological Review 21 (4), 709–716.
Wirth, L., 1938. Urbanism as a way of life. American Journal of Sociology 44, 1–24.
Zhao, R., Cao, L., 2010. Social change and anomie: a cross-national study. Social Forces 88 (3), 1209–1229.

Anomie: History of the Concept

Mathieu Deflem, University of South Carolina, Columbia, SC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Anomie is a term that, in various forms, originally appeared in writing in Greek antiquity and biblical history. Introduced in modern sociology by means of an appropriation from social and moral philosophy at the end of the nineteenth century, the concept of anomie was first applied in the seminal works of Emile Durkheim. It was subsequently elaborated and examined in the context of the sociology of deviant behavior on the basis of the writings of Robert K. Merton. With varying levels of success, anomie has remained a mainstay in sociological theory and research until today. The meaning of anomie has thereby undergone several transformations of more and less conceptual significance.

The term anomie dates back, in slightly differing linguistic forms, to Greek antiquity, the biblical writings of the Old and New Testaments, and late-nineteenth-century philosophy. As a scientific concept, anomie was introduced in sociology by Emile Durkheim and subsequently popularized through the writings of Robert K. Merton and other scholars working from within the tradition of structural-functionalism. Yet, Durkheim and Merton were not the only scholars to have developed sociological and other theories related to the term. Over the course of the development from classical to modern sociology, moreover, the concept has been used in a number of ways, both more and less distinctly related to the seminal works of Durkheim and Merton. A resurgence of the concept of anomie can be noted since the end of the twentieth century. As a result, in the contemporary sociology of the beginning of the current millennium, multiple understandings of the concept of anomie and related theories coexist.

Anomie Before Sociology

The concept of anomie is not the invention of sociology, but is a term that has historically existed in various incarnations dating back to ancient Greece (Orrù, 1987). Etymologically, the concept is derived from the Greek '*anomia*', which, containing the root term '*nomos*', literally means a lack of law or, more broadly, deregulation and/or normlessness. As first introduced in Greek antiquity, the term not only refers to a factual state of affairs, denoting a lack of order, but also carries with it an ethical judgment of the undesirability of that condition. Thus, typical of a society's culture that is as yet undifferentiated into scientific, ethical, and aesthetical realms, anomie was in the Greek tradition at once a descriptive and normative term, presenting both a rudimentary sociology and a philosophy of society.

Specifically, Greek philosophy contained references to *anomia* during debates in the fifth-century BC, which, foreshadowing modern sociological debates on anomie, was a period in history marked by profound social changes and turmoil. Greek philosophers in the sophist tradition that was oriented at educating political leaders and members of the nobility differentiated at least three aspects of these changes in qualifying a person as 'anomic', referring to a lack of personal character (inhumaneness), a disregard for religious norms (impiety), and defiance toward the unwritten rules of traditional custom (injustice). The Greek term *anomia* was mostly referenced in its adjective form (*anomos*) to relate to an order of rules and norms broader and more fundamental than only formal laws. The social and ethical problems discussed were therefore of both a moral character and political in nature, affecting people's dispositions and conduct as well as the social order in which human relations were organized. *Anomia*, in other words, was held to be a quality existing in people as well as in the order around them.

A next important presociological step in the development of the concept of anomie is its treatment in the biblical writings of Judaism and Christianity (Orrù, 1987). The biblical usage of *anomia* was partly influenced by relevant Greek philosophical writings because it entered the biblical tradition when the Old Testament was translated into Greek. Initially, the word *anomia* was used for about 20 Hebrew words, referring to various aspects of wickedness, ungodliness, injustice, wrongdoing, evil, depravity, transgressions, and sin. In the latter meaning of sin, *anomia* was used interchangeably with the Greek term '*hamartia*' which refers to a fatal flaw. The biblical understanding of *anomia* gradually became more delineated in meaning to refer, not to individual shortcomings, but to a state of society whereby the forces of evil took control over and against the divine order during the last days of the world. This apocalyptic understanding of *anomia* bears some resemblance to the sociological notion of anomie Durkheim would introduce in the late-nineteenth century as a counterpart to Karl Marx's historical-materialist theory of alienation.

A third understanding of anomie before the coming of sociology was represented in certain religiously inspired strands of Western philosophy, specifically in sixteenth- and seventeenth-century England, which developed a conception of natural law based on selected aspects of ancient Greek thought. This renaissance recaptured the Greek notion of *anomia*, which had mostly vanished during the Middle Ages, in legal as well as biblical writings by means of the Anglicized word 'anomy'. A perspective of natural law was thereby proposed in which the lawlessness of anomy was situated in opposition to an order that defended a sense of nature with implications for both the religious order and the social world. This understanding of anomy, however, was not successful in inspiring further

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03067-1

developments of philosophical and social thought. Stripped of its religious connotations, a new theory of anomie would not be developed until the late-nineteenth century.

Anomie in Classical Sociology

The French sociologist Emile Durkheim was the first to introduce the concept of anomie and a related theory of society into the discipline sociology, but he did not independently reinvent the term. Instead, Durkheim had discovered the concept of anomie first in the writings of his philosophical compatriot Jean-Marie Guyau (Besnard, 1987; Orrù, 1987). A prolific writer on moral theory, Guyau had written six books and an additional three manuscripts by the time of his untimely death at 33 years in 1888. Guyau's French-language concept of '*anomie*' appeared in two of his books, dealing with religion and morality, and was directly derived from the Greek '*anomia*', a term Guyau mentions explicitly in his writings. Interestingly, Guyau does not conceive of anomie as a sign of evil, but as the distinguishing quality of an epoch whereby the decline of religious traditions has brought about a continued individualization of morality and rules. Under conditions of freedom, a morally good social order can only mean the absence of collective order and obligation.

When Durkheim reviewed Guyau's work in the 1880s, he favorably commented on its ability to develop elements of a sociology of religion, manifested in Guyau's observations on the decline of religious traditions over the course of history. But Durkheim refuted Guyau's individualistic moral anarchism in favor of developing an understanding of anomie as being in contraction with a society's morality, the foundations of which to Durkheim could only be social, not individual. Thus, in Durkheim's sociology, morality is equated with social order, and any profound disturbance of the social order is captured under the heading of anomie.

Durkheim discussed the concept of anomie in two main instances in his sociological work (Besnard, 1987; Deflem, 1989). In his 1893 book on the social division of labor, Durkheim (1893) defends the perspective, squarely oriented against Marx's theory on the inherent contradictions of capitalism, that the division of labor under normal societal circumstances does not produce any negative outcomes. However, under exceptional circumstances, Durkheim argues, the division of labor that marks society as a whole produces pathological or anomic outcomes because and when, first, there are no regulatory bodies to oversee the process and direction of the division of labor and, second, the actions of such bodies are not prespecified by a particular set of rules. Durkheim mentions three such instances of an anomic division of labor taking place: during sudden periods of economic crisis; when conflicts occur between labor and capital; and with the increasing specialization of the moral and social sciences. In these cases, there is either no sufficient regulation over social relations or existing regulations are not congruent with the degree of societal development. The underlying theoretical premise of this concept of anomie as deregulation is that in so-called organic societies, with which Durkheim refers to modern societies that are marked by a high degree of individualism and functional specialization, integrative mechanisms are needed to accomplish unity and order through an appropriate recognition of diversity. According to Durkheim, the central organ necessary to achieve this integration and avoid anomie is the government as it can formally administer all necessary laws for a national collectivity.

In his seminal study on the social conditions of suicide, Durkheim (1897) also relies on a notion of anomie as part of a general theory of society, which he develops to account for various forms of suicide as a social fact. Distinguishing different types of suicide at the aggregate level, Durkheim first differentiates altruistic from egoistic suicide to refer to suicides influenced by the two extreme conditions of social integration. Egoistic suicide results from an insufficiency or lack of integration, while altruistic suicide is brought about by an extreme excess of integration. The consistently higher suicide rate among Protestants as compared to Catholics is an example for the former type, while the higher suicide rate in the military as compared to civilian life is an example of the latter.

A third form, anomic suicide, Durkheim explains on the basis of an absence or a weakening of social regulation. Under these circumstances of inadequate or ineffective regulation, the suicide rate is abnormally high because of the suddenness of the changes taking place, irrespective of whether those changes are perceived as beneficial or harmful. Examples include moments of economic crisis, divorce, and rapidly acquired political strength through military expansion. Underlying Durkheim's concept of anomie is the theory that social rules are needed in order to limit human desires that otherwise would be insatiable. Therefore, according to Durkheim, society, which exists exterior to the individual, holds a favorable coercive force over the individual's desires. Under exceptional circumstances, including both fortunate and unfortunate but always sudden moments of transition, societal constraints are not effective in regulating human conduct. Acute forms of anomie then ensue, raising the level of suicide to pathological levels. Additionally, Durkheim argues, that in some spheres of social life, specifically in the realm of the economy, an absence of regulation can be of a more chronic nature. In such cases of a materialistically driven lack of any restraint, the anomic suicide rate is consistently high. The type of fatalistic suicide that results from an excess of regulation, as a logical counterpart to anomic suicide, Durkheim considered only a theoretical possibility. Also, anomic conditions could according to Durkheim only be effectively dealt with by means of well-functioning corporations or professional groups that, positioned in between singular individuals and the whole of society, can take on important regulatory functions.

Anomie and Deviant Behavior

Durkheim's concept of anomie did not readily lead to develop a relevant strand of empirical research and theorizing, although the classic's underlying theory of society was, of course, enormously influential in the further evolution of sociology in the modern era. Neither did a corresponding Durkheimian sociology of law and norms develop until more recent times of advanced sociological specialization. However, Durkheim's concept of anomie would be reintroduced in modern sociology by way of the intellectually closely aligned emergence of

American structural-functionalism and its attention toward social integration and regulation. The modern concept of anomie would thereby change in meaning and become associated more closely with the sociology of deviant behavior.

Robert K. Merton performed the most valuable work to renew and popularize an interest in anomie when he developed a novel sociological perspective of deviant behavior. Merton's initial version of this theory was published in 1938 at a time when the scholar was heavily influenced by Durkheimian thought while a student at Harvard (Merton, 1938). That initial paper was a rather exploratory sketch of a theory on the social conditions of deviant behavior, which he subsequently presented in a more systematic fashion (Merton, 1949), and elaborated with an additional chapter that was published, along with the theory's mature version, in the 1957 edition and all subsequent versions of his influential collection of papers, *Social Theory and Social Structure* (Merton, 1957a,b).

Merton's theory of anomie is presented as part of a sociological approach to the study of deviant behavior. Merton developed this perspective with the manifest objective of producing a sociological theory that conceived of deviant behavior as the normal outcome of certain social conditions, rather than an aberration on the basis of a psychology or biology of criminal dispositions. Merton's theory unfolds on the basis of a perspective of society that distinguishes between the cultural goals preferred in a given social context, on the one hand, and the institutionalized norms or legitimate means to pursue those objectives, on the other. This viewpoint is applied to the conditions that exist in the context of the United States, whereby, Merton argues, the predominant goal is individual success, especially in the form of monetary wealth, while the institutionalized norms in the American context value hard work, especially by means of education and employment. Merton additionally observes, on the force of corroborating empirical evidence, that the cultural objectives of American society are typically emphasized much more than the legitimate means to reach them, thus unsettling the equilibrium between goals and means. As a result of this disjunction, Merton argues that anomie or normlessness takes place in the form of a deinstitutionalization of the legitimate means of society.

Merton's related, but otherwise distinct theory of deviant behavior explains the various modes of conduct that can take place under conditions of anomie (Featherstone and Deflem, 2003). Merton maintains that anomic conditions place specific pressures on individuals in terms of their ability to adopt cultural goals and means, which varies according to their respective positions in the social structure. Several modes of adaptation are possible, the most common one being conformity, whereby both the goals and means of society are accepted. Other forms of conduct, however, are considered deviant (although not necessarily dysfunctional) and involve an inability to adopt culturally approved goals and/or institutionalized means. In the case of innovation, for instance, alternative and often illegal means are appropriated to attain the cultural goal of individual success and monetary gain, whereas the adaptation of ritualism implies a holding on to legitimate means even though the success goals are out of reach. Developing a related theory of opportunity structures (Marwah and Deflem, 2006), Merton maintains that the likelihood of various adaptations is influenced by the extent to which individuals do or do not have access to legitimate means in pursuit of widely shared goals. Merton in this respect referred to economic as well as cultural and other relevant conditions of opportunity.

It can readily be observed that Durkheim's and Merton's concepts of anomie are not identical, as the former primarily relates to a society's goals, whereas the latter emphasizes varying opportunities of means. However, unlike Durkheim's original notion, Merton's reintroduction of the anomie concept in modern sociology brought about a veritable avalanche of research and theory, especially in the realm of the sociology of deviant behavior and crime in the post–World War II era, with several hundreds of articles and books devoted to the subject until the 1970s. Of special interest are efforts to measure anomie empirically and related attempts to extend the notion of anomie from describing a state of society to analyzing the conditions of individuals in anomic societies (Deflem, 1989). Most famous among the attempts to operationalize anomie is a scale developed by Leo Srole to measure the subjective experiences of people exposed to the social condition of anomie. Because Srole was aware that this measure did not capture anomie as a social condition, but instead examined the attitudes and dispositions of individuals, he reintroduced the corresponding term anomia in a sociopsychological meaning. Relatedly, other modern sociologists who were influenced by Merton's writings, but who were also conscious of a growing theoretical awareness to distinguish between macro and micro aspects of social life, likewise argued for the relevance of the psychology of anomie to conceptually differentiate the subjective experience of anomie from its objective conditions. With variable levels of success, and causing some unneeded terminological confusion, the term anomie was then substituted by, and/or interchangeably used with, anomia and anomy. In the present era, however, only the concept of anomie has survived.

Besides the microtheoretical readings of anomie, conceptually most interesting during the period of the dominance of American structural-functionalism was the effort by mastertheorist Talcott Parsons to explore anomie within a broadly Durkheimian framework. In his first major work, *The Structure of Social Action*, Parsons (1937) explains the notion of anomie as central to Durkheim's thought in revealing that the coercive force of society over the individual, unlike the physical force of nature, is moral in kind. This insight leads Parsons to develop, and argue for the gradual unfolding in Durkheim's work of, a voluntaristic theory of action, whereby the human will is seen as central to the functioning of society. Anomie, conversely, denotes a breakdown of the hold of norms over individual conduct. In *The Social System*, Parsons's (1951) second *magnum opus*, the concept of anomie is again introduced in a Durkheimian fashion, albeit with explicit reference to a sociological theory of deviant behavior and social control. Critically responding to the writings of his former student Robert Merton, Parsons uses the concept of anomie to refer to a state of disjunction between expectations and reality, involving a breakdown of the social order in the form of an ineffective regulation of human conduct. Anomic forms of suicide and deviant behavior are the result of this state of anomie. Mechanisms of social control are then called

upon to restore the social system. If social control is not successful, anomie may endure to the point where structural changes lead the way to a different collective order.

Contemporary Travels and Adventures of Anomie

With the decline of the dominance of structural-functionalist sociology during the 1970s, the concept of anomie had all but vanished from sociology, especially from the practice of social research where once it had gained a central place. However, by the late 1980s, as the vulgar criticisms against structural-functionalism had been scrutinized and at least partly overcome, the concept of anomie witnessed something of a resurgence that has lasted, and has even been strengthened, until this day. At least two currents can presently be discerned in contemporary sociological perspectives of anomie.

First, the Mertonian concept of anomie has remained of considerable interest in contemporary criminology and sociology (Alder and Laufer 1995; Marwah and Deflem, 2006; Passas and Agnew, 1997). The attention has shifted, however, toward treating Merton's contributions as a classic reference point, rather than a direct source of research, and toward the further exploration of Merton's theory of deviant behavior rather than the theory of society, as comprising cultural goals and institutionalized norms, that Merton developed along with it. Thus, the attention has moved away from anomie theory toward strain theory, whereby the latter dimension is often explored in individualistic terms. An exception, adequately focusing on the structural dimensions of society that are central to Durkheim, Merton, and like-minded sociologists, is the perspective of institutional-anomie theory that focuses on the criminogenic impact of the relative lack of regulation of an unbridled free-market economy by regulatory institutions such as the family.

Second, the concept of anomie has reappeared in contemporary sociological discourses in a meaning that relates more distinctly to the concept Durkheim originally introduced. Some of these efforts are of a theoretical nature, but most are explored through empirical research. On a theoretical level, among the noteworthy reintroductions of anomie is the work of Jürgen Habermas, specifically his sociological theory of system and lifeworld. Foregoing a lengthy exposition of this theory of society, Habermas (1981) introduces a concept of the lifeworld that refers to the dimension of society where actors are meaningfully coordinated in their social relations through communicative action. Habermas posits that the lifeworld contains the three structural components of culture, society, and personality, which, respectively, serve the functional needs of cultural reproduction, social integration, and socialization. Habermas describes the problems that can occur with respect to these three functions as, respectively, a loss of meaning, anomie, and psychopathologies. This Habermasian understanding of anomie is thus closely akin to, and directly derived from, the concept as it was introduced by Durkheim.

More elaborate than the odd theoretical elaboration of anomie in contemporary sociology are the many new lines of empirical research that have used the concept of anomie in recent decades. This new and still growing literature on anomie is oriented at describing and analyzing issues of deregulation, normlessness, and/or norm-ineffectiveness that have come about since some of the most major societal changes that took place during and since the end of the twentieth century, such as the fall of communism, the global spread of capitalism, and the diffusion of, and difficulties involved in implementing with, the ideals of democratic government. Particularly noteworthy are the instances of social research on anomie in the context of post-Communist Eastern Europe and along with the rapid modernization of China. In the European context, important societal changes relevant to the study of anomie are additionally brought about by the encroachment of the European Union and the high levels of immigration into the European continent. In the Chinese context, the coexistence of a still Communist political regime with a rapidly modernizing economy has been observed to have produced strains and discrepancies in the social system that bring about a form of disorder the concept of anomie can highlight. Finally, there are some, albeit as yet few signs that the concept of anomie has gradually also begun to gain some new attention in other sociological studies, especially at the level of organizations. A veritable revival of anomie cannot yet be predicted at the present time.

See also: Alienation, History of; Anomie; Durkheim, Emile (1858–1917); Functionalism, History of; Habermas, Jürgen (1929–); Merton, Robert K (1910–2003); Norms; Strain Theories and Crime; Suicide, Sociology of.

Bibliography

Adler, F., Laufer, W.S. (Eds.), 1995. The Legacy of Anomie Theory. Transaction, New Brunswick, NJ.

Besnard, P., 1987. L'anomie: Ses Usages et Ses Fonctions dans la Discipline Sociologique Depuis Durkheim. Presses Universitaires de France, Paris.

Deflem, M., 1989. From anomie to anomia and anomic depression: a sociological critique on the use of anomie in psychiatric research. Social Science and Medicine 29 (5), 627–634.

Durkheim, E., 1893/1984. The Division of Labor in Society. The Free Press, New York.

Durkheim, E., 1897/1951. Suicide: A Study in Sociology. The Free Press, New York.

Featherstone, R., Deflem, M., 2003. Anomie and strain: context and consequences of Merton's two theories. Sociological Inquiry 73 (4), 471–489.

Habermas, J., 1981/1987. The Theory of Communicative Action. System and Lifeworld: A Critique of Functionalist Reason, vol. 2. Beacon Press, Boston, MA.

Marwah, S., Deflem, M., 2006. Revisiting Merton: continuities in the theory of anomie-and-opportunity-structures. In: Deflem, M. (Ed.), Sociological Theory and Criminological Research: Views from Europe and the United States. Elsevier/JAI Press, Amsterdam, pp. 57–76.

Merton, R.K., 1938. Social structure and anomie. American Sociological Review 3, 672–682.

Merton, R.K., 1949. Social structure and anomie: revisions and extensions. In: Anshen, R.N. (Ed.), The Family: Its Functions and Destiny. Harper, New York, pp. 226–257.

Merton, R.K., 1957a. Social structure and anomie. In: Merton, R.K. (Ed.), Social Theory and Social Structure. The Free Press, New York, pp. 185–214.

Merton, R.K., 1957b. Continuities in the theory of social structure and anomie. In: Merton, R.K. (Ed.), Social Theory and Social Structure. The Free Press, New York, pp. 215–248.

Orrù, M., 1987. Anomie: History and Meanings. Allen & Unwin, Boston.

Parsons, T., 1937. The Structure of Social Action. McGraw Hill, Boston, MA.

Parsons, T., 1951. The Social System. The Free Press, New York.

Passas, N., Agnew, R., 1997. The Future of Anomie Theory. Northeastern University Press, Boston, MA.

Anthropocene Epoch

Anson W Mackay, University College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Anthropocene, as a geological term and disciplinary concept, has gained widespread acceptance in recent years. It represents the period of time when humans have had major, global impacts on many of the Earth's systems (e.g., atmosphere, oceans, and terrestrial biospheres), post AD 1800. The extent of human impact on the Earth is initially discussed, followed by events characterizing the start of the Anthropocene and the Great Acceleration after the end of the Second World War. Increasing greenhouse gases, ocean acidification, and biodiversity loss are considered in detail as case studies.

The Anthropocene: The Beginnings

It has been recognized since at least the middle of the nineteenth century that human activities have had major environmental impacts on the Earth (Marsh, 1864) and its atmosphere (Smith, 1872). At this time, the Western nations were undergoing a revolution in technological innovations, linked primarily to the efficient exploitation of stored energy in fossil fuels: the Industrial Revolution. Fossil fuels provided an almost unlimited supply of energy, which led to widespread synthesis of nitrogen-based fertilizers and the mechanization of large-scale land clearance for agriculture. These, together with many other innovations in, e.g., sanitation and healthcare, led to an exponential increase in human populations from approximately 1000 million in AD 1800 to over 7000 million by the year AD 2013, and a 50-fold increase in economic production (Steffen et al., 2011). This is age of the Anthropocene.

The Anthropocene's first appearance in print was by Crutzen and Stoermer (2000) "… to emphasize the central role of mankind in geology and ecology by proposing to use the term 'anthropocene' for the current geological epoch." Its novelty, and in part its attraction, is that it describes a period during which modern humans have globally become a significant force of nature, rivaling natural environmental, and geological processes. Intrinsic to the definition are the observations that humans have pushed many of the Earth's systems beyond natural variability. For example, carbon dioxide concentrations (CO_2) in the atmosphere are higher than they have been for millions of years, while human activities (such as agriculture) dominate global nitrogen budgets (Galloway et al., 2004).

Perhaps surprisingly, one of the most contentious aspects related to the concept of the Anthropocene is when it started. For example, did it start with the onset of the Industrial Revolution (e.g., Crutzen and Stoermer, 2000; Steffen et al., 2011), or much earlier in human history with the commencement of Neolithic forest clearance approximately 8000 years ago, followed by the expansion of paddy fields for rice cultivation *c*.5000 years ago (e.g., Ruddiman, 2003, 2013). This is known as the early anthropogenic hypothesis. Attempts to define the start of the Anthropocene are not a trivial task and have implications beyond the nature of recognizing a period of intense human impact. For example, there is a push for the Anthropocene to be designated as a new, formal geological epoch (Zalasiewicz et al., 2011). For several years, an Anthropocene committee has been looking into the implications of such a designation, and if a new designation based on stratigraphy is even possible (Gale and Hoare, 2012). If the Anthropocene was formally ratified, then that would mean we have left the era of the Holocene. Key to this debate is the nature of a stratigraphical golden spike; the golden spike is the transition that marks the end of one geological era and the start of another. The transition is usually characterized by well-defined changes in biological fossils and/or sedimentology, discernible in different regions of the world. Considerable debate has ensued as to defining a golden spike between the Holocene and the Anthropocene, and at the time of writing, there is no resolution. The International Commission on Stratigraphy hopes to make a decision by 2016.

In this article I take the view point that the Anthropocene started with the onset of the Industrial Revolution purely on the basis that the temporal designations suggested by Steffen et al. (2007) provide a good structure for the assessment of human impact on the earth, including Pre-Anthropocene impact; Stage 1: utilization of energy from fossil fuels: AD 1850–1945; and Stage 2: exponential increase in energy consumption and population growth from 1945 – to the near future.

Pre-Anthropocene Impact/Early Anthropogenic Hypothesis

Prehistory

Phylogenetic evidence suggests that the early hominin *Homo erectus* used fire to modify their environment and cook food *c*.1.8 Ma years ago, which influenced her evolution and growth of brain size. Modern humans (*Homo sapiens*) evolved in Africa approximately 0.2 Ma, and undertook diachronous migrations into other continents. It is widely recognized that modern humans played a major role in altering biodiversity and landscapes across continents and islands since at least the past 50 000 years. For example, humans have been implicated in the decline and often extinction of hundreds of large mammalian species (called megafauna) across North and South America, Eurasia, Australia, and many oceanic islands (see Barnosky et al., 2004, for a review). Only in Africa and the oceans have megafauna largely escaped wide-scale extinction, although hundreds of species are currently under severe pressure. However, although humans may have contributed to increased rates of megafauna extinctions (e.g., through hunting and alteration of habitats), climate change is also implicated (Wroe et al., 2013), and when considering

megafauna extinctions worldwide, it seems likely that both climate and anthropogenic activity acted in tandem (Prescott et al., 2012).

Agricultural Revolution

The expansion of domestication and agriculture in multiple regions around the planet (e.g., see Fuller et al., 2011) has had major impacts on landscapes, biodiversity, and atmospheric chemistry since the start of the Holocene. The Neolithic revolution from about 8000 years ago led to large areas of forest being felled and burned to improve agricultural land. Ruddiman (2003) put forward the hypothesis that forest clearance led to an overall increase in atmospheric (CO_2) (**Figure 1(b)**, which has contributed to small but sustained elevated global temperatures (**Figure 1(c)**). Some 3000 years later, agricultural expansion in southeast Asia led to widespread cultivation of rice in paddy fields, and possibly concomitant global increases in global methane (CH_4) concentrations (Ruddiman, 2003; Fuller et al., 2011) (**Figure 1(a)**). Whilst there is still debate about the contribution of these land-use practices toward early Holocene greenhouse gas (GHG) concentrations in the atmosphere, early human alteration of the landscape at broad regional scales is increasingly being recognized (Ellis et al., 2013).

Anthropocene Stage 1

A principal reason why population numbers and energy consumption were kept in check for much of human history was because societies were dependent on rather inefficient mechanisms to provide energy, which relied upon either natural forces (e.g., wind, flowing water) or organic fuels especially wood, peat, and in Europe and China at least, coal over the last millennium. The turning point seems to have been when the Scottish inventor James Watt made improvements to the steam engine at the end of the eighteenth century. This allowed for much more efficient energy generation, which contributed to the start of the Industrial Revolution. For the first time, sufficient energy could be used to chemically generate fertilizers from atmospheric nitrogen, literally getting nutrients from air. This allowed much better agricultural yields to be gained, which together with better medical care, led to

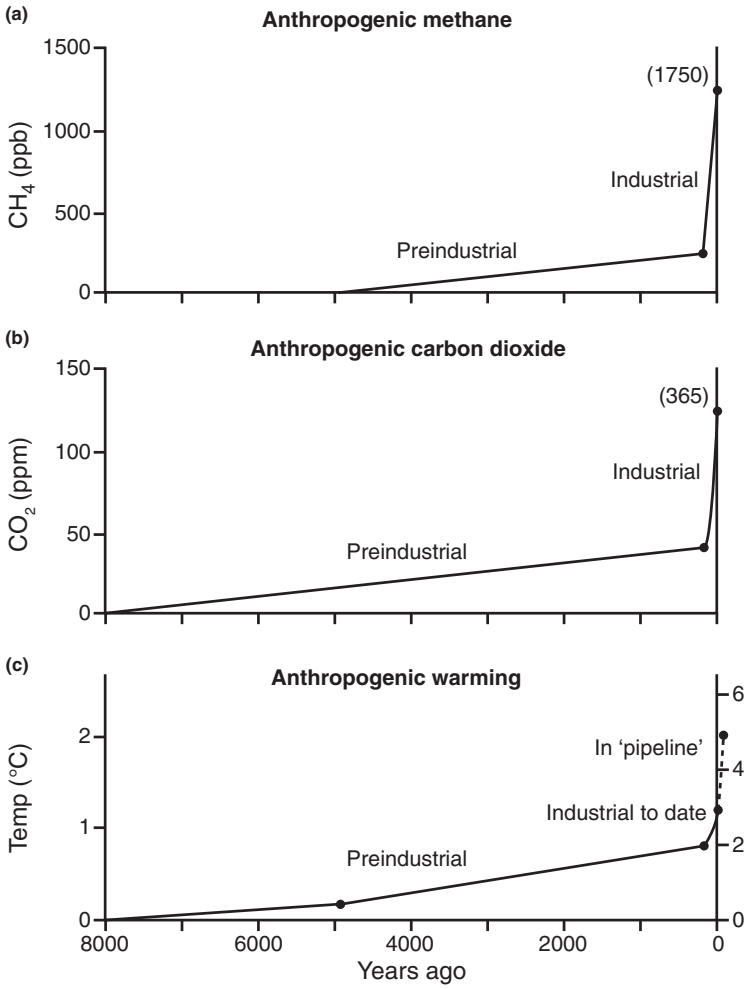

Figure 1 Early anthropogenic hypothesis: (a, b) Preindustrial and industrial-era CH_4 and CO_2 increases from anthropogenic sources. (c) Warming caused by anthropogenic greenhouse gases, with global mean temperature on scale at left and high-latitude temperature on scale at right. From Ruddiman, W.F., 2003. The anthropogenic greenhouse era began thousands of years ago. Climate Change 61, 261–293.

rapidly increasing human populations. Burning of fossil fuels led to increases in atmospheric GHGs, especially CO_2, whilst increasing agricultural practices generally led to increases in CH_4 and nitrous oxide (N_2O) (IPCC, 2007). That these concentration increases are apparent in Antarctic ice cores highlights their global atmospheric distribution.

Increasing fossil fuel combustion and agricultural activities also led to the production of vast quantities of sulfur dioxide (SO_2) and nitrous oxides (NO*x*), which in the atmosphere converts to sulfate (SO_4) and nitrates (NO_3), and leads to acidification of freshwaters and terrestrial ecosystems. Acidification has been especially problematic in regions where catchment geology is base poor, including large parts of northern Europe and North America. Continental scale changes in freshwater diversity were recognized as far back as the early 1980s, and although international legislation was adopted to reduce acidification, biological recovery is being hampered due to climate change (Battarbee et al., 2010).

Anthropocene Stage 2

Post Second World War, global networks of communication and finance were set up to promote greater co-operation between countries. Forty-three countries met at Bretton Woods, New Hampshire, USA, to rebuild the decimated global economy based on neoliberal/capitalist thinking. The conference led to the formation of the International Monetary Fund (and eventually to the World Bank) and open trade agreements. The conference also led to networks of international scientists and engineers leading to major technological advances such as development of nuclear energy and construction of massive structures such as rigs for drilling oil in deep waters. In the early 1960s, subsidies for agriculture were widespread (e.g., the European Union Common Agricultural Policy), which resulted in more intense land use and application of fertilizers. This in turn resulted in widespread nutrient enrichment of freshwater and coastal ecosystems and declines in biodiversity. New oil stores were exploited, concomitant with generation of cheap energy via power stations, which led to very rapid increases in GHG emissions. Factory production of goods were internationalized and made cheap, further integrating global economies.

The rate of societal change, energy consumption, and population growth post Second World War was so dramatic that this period is frequently referred to as the 'Great Acceleration' (Hibbard et al. in Steffen et al., 2007) (**Figure 2**). It is worth noting that energy consumption during this period was driven mainly by countries such the USA, Germany, UK, and France. But in the last decade, new nations are rapidly increasing consumption rates, including Brazil, Russia, India, China, and South Africa, with other countries predicted to become major consumers in the near future (e.g., Nigeria). Impacts on the environment include accelerating increases in GHGs, rapid increases in coastal pollution and exploitation of fisheries, and worrying increases in species extinctions. These impacts stem especially from population increases, energy consumption, and land-use changes.

During this period, scientific endeavors with respect to understanding impacts on the environment have on balance been reactive rather than proactive. For example, realization that acid rain was having transboundary negative impacts in Europe and North America was initially tackled by the UNECE Convention on Long-Range Transboundary Air Pollution (CLRTAP) which was signed in 1979 and entered into force in 1983. The potential danger of artificial greenhouse chlorofluorocarbon gases (CFCs) on ozone in the atmosphere was recognized by Molina and Sherwood in 1974. Observations of the ozone hole across Antarctica in 1985 quickly led to the internationally ratified Montreal Protocol in 1987 which resulted in start of the decline of the manufacturing of CFCs and other ozone-depleting substances. The Montreal Protocol is widely seen as being very successful, to such an extent that the curb on these GHGs may also have acted to protect climate from further warming (Velders et al., 2007).

Case Studies

Atmosphere

Some of the most dramatic human impacts on the Earth linked to the Anthropocene are related to increases in GHGs, especially CO_2, CH_4, and N_2O. Although the latter two have greater greenhouse potencies with respect to CO_2, because CO_2 is much more abundant, it has the highest global radiative forcing (e.g., approximately 1.66 W m^{-2} in 2005; IPCC, 2007). Increased concentrations of atmospheric CO_2 are mainly from combustion of fossil fuels (*c*.74%), although land-use change, such as deforestation, is a small (*c*.22%), but important component. The oceans and terrestrial biosphere both play an important role in regulating atmospheric CO_2 concentrations, and since the 1980s have removed approximately half of anthropogenic emissions (IPCC, 2007). In recent years annual CO_2 emissions into the atmosphere have surpassed 30 billion tonnes, leading to CO_2 levels surpassing 400 ppm at Mauna Loa Laboratory for the first time in 2013. This far exceeds preindustrial values of *c*.280 ppm, and now far exceeds natural variability of atmospheric CO_2 over the past few million years (Monastersky, 2013), long before the arrival of modern humans onto the Earth.

Recent increases in atmospheric concentrations of methane are also ascribed to human activities including exploitation of fossil fuels and agricultural practices. Methane is about 20 times more potent a GHG than CO_2, and current levels in the atmosphere reached over 1800 ppb in 2011, and far surpass natural variability over the past 650 000 years (320–790 ppb; IPCC, 2007). The third important GHG is N_2O, and although approximately two-thirds of this gas derives from natural sources, excess emissions caused by agricultural activities have resulted in 2011 values of 324 ppm in comparison to preindustrial values of *c*.270 ppm (IPCC, 2007).

Why are these increases important? Globally, increased GHGs from anthropogenic activities has resulted in unequivocal warming of the climate system (IPCC, 2007), including both atmosphere and oceans. Impacts on the oceans, cryosphere and terrestrial biosphere are wide ranging with varying degrees of understanding. For example, increased atmospheric temperatures have resulted in nine of the hottest years ever recorded since 2000 (with the exception of 1998). The oceans

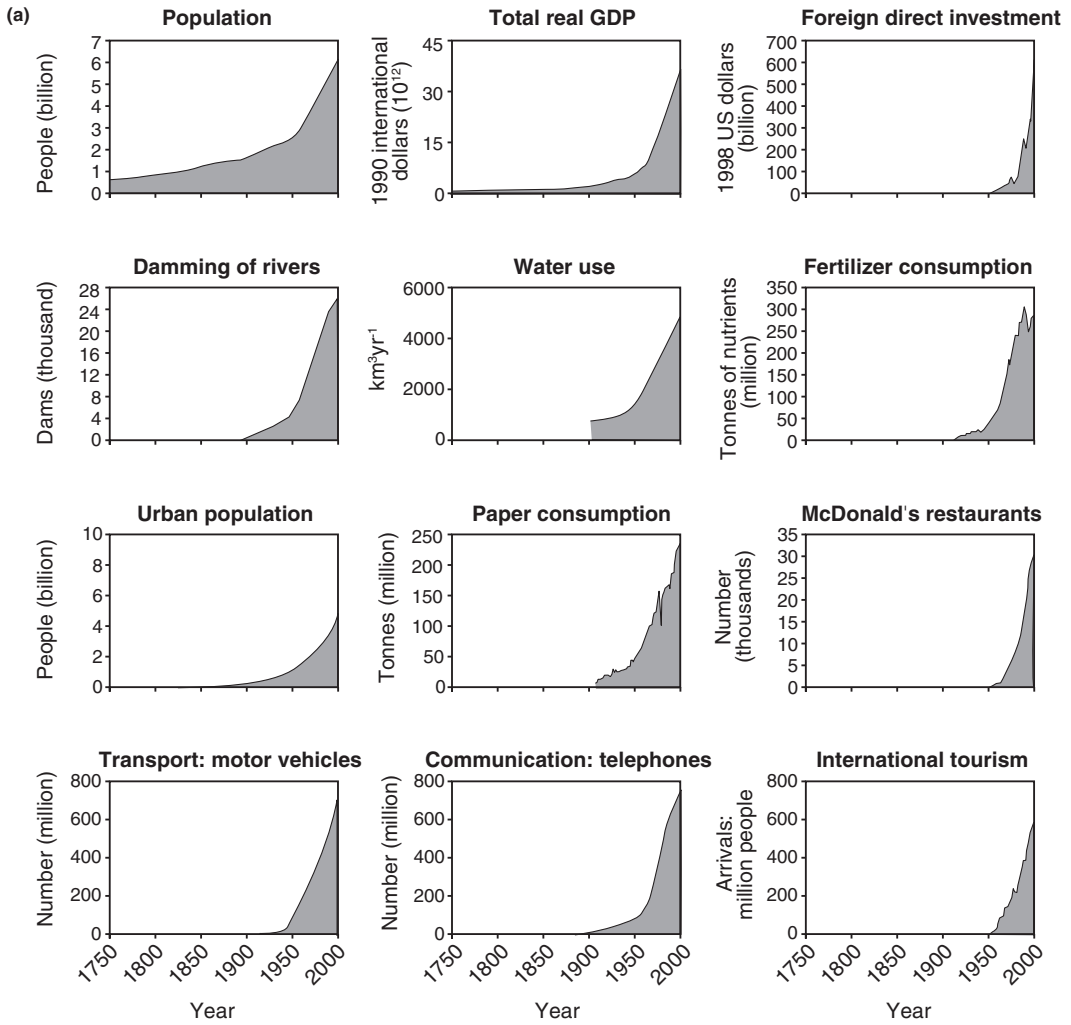

Figure 2 (a) The increasing rates of change in human activity since the beginning of the Industrial Revolution. Significant increases in rates of change occur around the 1950s in each case and illustrate how the past 50 years have been a period of dramatic and unprecedented change in human history. (b) Global scale changes in the Earth System as a result of the dramatic increase in human activity: (i) atmospheric CO_2 concentration; (ii) atmospheric N_2O concentration; (iii) atmospheric CH_4 concentration; (iv) percentage total column ozone loss over Antarctica, using the average annual total column ozone, 330, as a base; (v) Northern Hemisphere average surface temperature anomalies; (vi) natural disasters after 1900 resulting in more than 10 people killed or more than 100 people affected; (vii) percentage of global fisheries either fully exploited, overfished or collapsed; (viii) annual shrimp production as a proxy for coastal zone alteration; (ix) model calculated partitioning of the human-induced nitrogen perturbation fluxes in the global coastal margin for the period since 1850; (x) loss of tropical rainforest and woodland, as estimated for tropical Africa, Latin America, and South and Southeast Asia; (xi) amount of land converted to pasture and cropland; and (xii) mathematically calculated rate of extinction. Reproduced from Steffen, W., Grineald, J., Crutzen, P., McNeill, J., 2011. The anthropocene: conceptual and historical perspectives. Philosophical Transactions of the Royal Society A 369, 842–867.

have also been warming up, and actually absorb over 80% of extra heat produced by higher GHGs (Tyrrell, 2011). Increased warming has also resulted in glacier retreat around the world and decline in Arctic sea-ice cover. Both melting ice and ocean expansion through increased heat have resulted in rapid increases in eustatic sea levels.

Ocean Acidification

The oceans play a crucial role in regulating atmospheric CO_2 because they absorb up to one-third of CO_2 emitted every year through pollution and deforestation. Since the start of the Industrial Revolution, increasing global atmospheric CO_2 concentrations have major impacts on ocean biogeochemistry, which is highly likely to have major consequences for marine biodiversity in the future. When CO_2 dissolves in sea water, it forms carbonic acid which acts to naturally increase the acidity of sea water. Since the start of Anthropocene stage 1, the pH or surface waters of the ocean has dropped by 0.1 pH unit, to pH 8.1. This may not sound a lot, but in fact the rate of acidity increase is over an order of magnitude faster than has occurred for millions of years (Doney et al., 2009), and if CO_2 emissions are not curbed, then ocean acidity may fall by a further 0.3–0.4 units (Schiermeier, 2011).

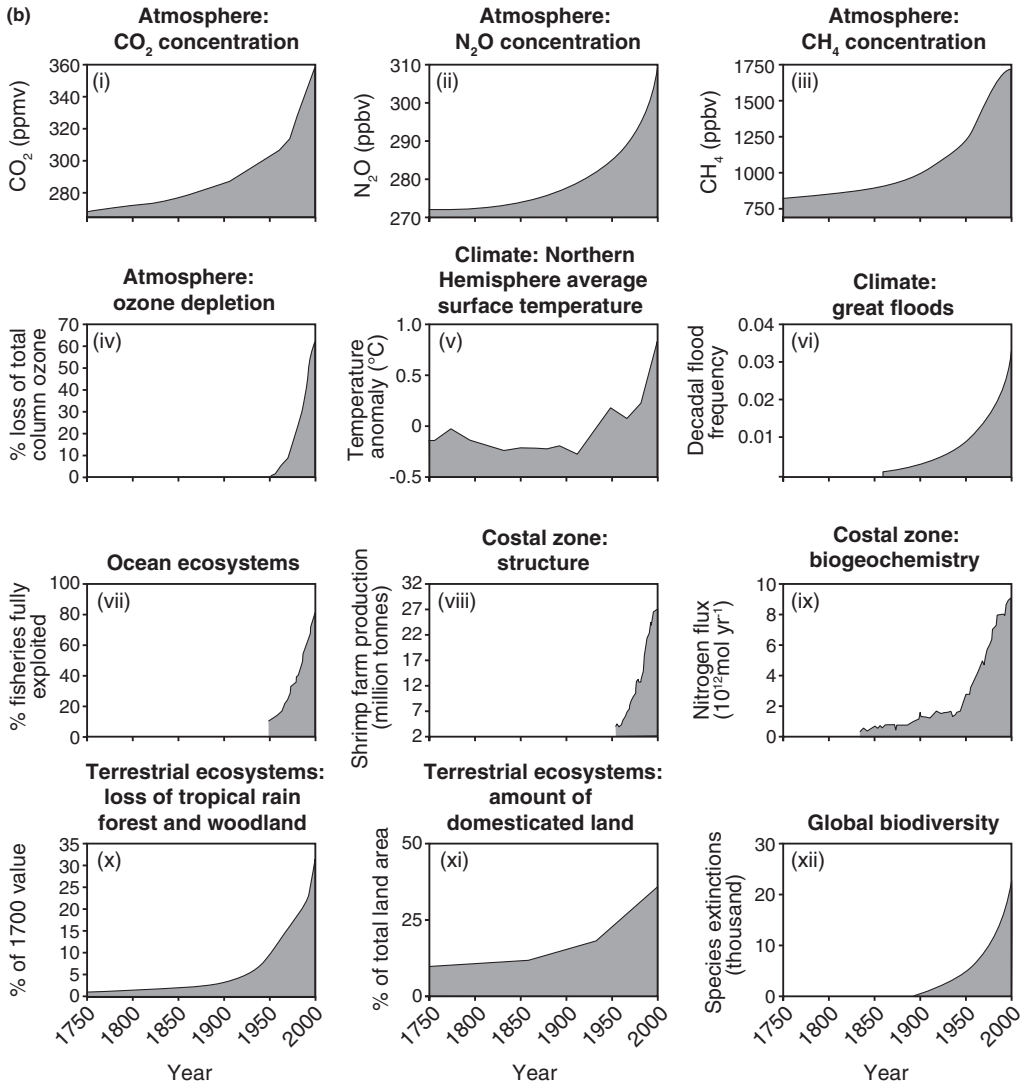

Figure 2 (*continued*).

Why is ocean acidification important? Many organisms that live in the sea, from plankton to oysters to corals, use calcium carbonate ($CaCO_3$) to build their shells and skeletons. However, increased dissolved CO_2 in marine water combines with carbonate ions (to form bicarbonate), which means that $CaCO_3$ can no longer be used. Furthermore, if the concentration of carbonate declines to low enough levels, shells and skeletons weaken and may start to dissolve. For example, coral reefs are seriously under threat from ocean acidification, as highlighted by several laboratory studies which demonstrated a decline in calcification rates related to acidification (Doney et al., 2009). However, with increasing observations in the field, impacts from acidification are not as straightforward as early studies suggested, with apparent greater spatial and temporal heterogeneity (e.g., Pandolfi et al., 2011).

Biodiversity and Land-Use Change

Biological diversity (biodiversity) helps to maintain healthy ecosystems, e.g., water quality, primary production, and decomposition. Biodiversity is therefore an important component of ecosystem services, which contribute to healthy, human livelihoods. Increasing biodiversity results in increased ecosystem stability (Cardinale et al., 2012). Conversely, declines in biodiversity can have major implications for ecosystem functioning and the services that they provide, with complex outcomes. For example, local species losses can have as significant an impact on, e.g., local primary production as effects from global drivers such as climate change (Hooper et al., 2012). However, in a meta-analysis of over 100 studies the introduction of new species resulted in increases in primary production and other ecosystem services, despite declines in richness of native species (Vila et al., 2011).

Anthropogenic activity during the Anthropocene may be driving the planet to a 'modern extinction crisis' because extinction rates are orders of magnitude greater than background rates (Barnosky et al., 2011). Drivers of species loss are varied, and include direct hunting, habitat change and fragmentation, competition for resources, pollution, and anthropogenic global warming, on top of natural factors such as climate change.

However, it is far from clear whether biological change, including declines in biodiversity and land-use change, will lead to a planetary-scale state shift in the Earth's biosphere (Barnosky et al., 2012; Rockstrom et al., 2009). At local and regional scales, species losses are often compensated for by species introductions and invasions (Ellis et al., 2012). This is important as ecosystem functioning at local and regional scales is dependent on species maintaining high levels of biodiversity. Brook et al. (2013) further challenge the concept that the terrestrial biosphere has a planetary tipping point as responses to anthropogenic change are likely to be heterogeneous in time and space, limiting the usefulness of any single aggregate such as global species richness.

Conclusions

The Anthropocene epoch is increasingly recognized as a paradigm shift in global environmental change research. Since the beginning of the Industrial Revolution, human activities have played an ever-increasing role in the functioning of the Earth System. Although global climate has emerged from the end of the Little Ice Age into the early stages of global warming, many of the dramatic environmental and ecological changes that have occurred since the early nineteenth century owe their origins to technological advances leading to unconstrained use of fossil fuels, increased supply of energy and rapid population growth, and rapidly increasing production and consumption of goods and services. We now live in a no-analog world in which a deeper understanding of the principal environmental, socioeconomic, and cultural transformations of the last three centuries is a vital prerequisite for working toward a viable future for human populations. There is little doubt that humankind will remain a major geological force into the foreseeable future. How Earth System functioning responds will be dependent upon key societal and technological decisions made now and in the near future with respect to, e.g., restructuring of national and international institutions through Earth System governance.

See also: Agroecology and Agricultural Change; Capitalism: Global; Climate Change and Global Poverty; Climate Change: Economics; Climate Change; Fire, Human Use, and Consequences; Global Development: Markets and Wellbeing.

Bibliography

Barnosky, A.D., Koch, P.L., Feranec, R.S., Wing, S.L., Shabel, A.B., 2004. Assessing the causes of late Pleistocene extinctions on the continents. Science 306, 70–75.

Barnosky, A.D., Matzke, N., Tomiya, S., Wogan, G.O.U., Swartz, B., Quental, T.B., Marshall, C., McGuire, J.L., Lindsey, E.L., Macguire, K.C., Mersey, B., Ferrer, E.A., 2011. Has the Earth's sixth mass extinction already arrived? Nature 471, 51–57.

Barnosky, A.D., et al., 2012. Approaching a state shift in Earth's biosphere. Nature 486, 52–58.

Battarbee, R.W., Charles, D.F., Bigler, C., Cumming, B.F., Renberg, I., 2010. Diatoms as indicators of surface water acidity. In: Smol, J.P., Stoermer, E. (Eds.), The Diatoms: Applications in the Environmental Sciences. Cambridge University Press, Cambridge, pp. 98–121.

Brook, B.W., Ellis, E.C., Perring, M.P., Mackay, A.W., Blomqvist, L., 2013. Does the terrestrial biosphere have planetary tipping points? Trends in Ecology & Evolution 28, 396–401.

Cardinale, B.J., et al., 2012. Biodiversity loss and its impact on humanity. Nature 486, 59–67.

Crutzen, P.J., Stoermer, E.F., 2000. The "Anthropocene". IGBP Newsletter 41, 17–18.

Doney, S.C., Fabry, V.J., Feely, R.A., Kleypas, J.A., 2009. Ocean acidification: the other CO_2 problem. Annual Review of Marine Science 1, 169–192.

Ellis, E.C., 2011. Anthropogenic transformation of the terrestrial biosphere. Philosophical Transactions of the Royal Society A 369, 1010–1035.

Ellis, E.C., Antill, E.C., Kreft, H., 2012. All is not loss: plant biodiversity in the anthropocene. PLoS ONE 7, e30535, 9 pp.

Ellis, E.C., Kaplan, J.O., Fuller, D.Q., Vavrus, S., Goldewijk, K.K., Verburg, P.H., 2013. Used plant: a global history. Proceedings of the National Academy of Sciences 110, 7978–7985.

Fuller, D.Q., Willcox, G., Allaby, R.G., 2011. Cultivation and domestication had multiple origins: arguments against the core area hypothesis for the origins of agriculture in the near East. World Archaeology 43, 628–652.

Gale, S.J., Hoare, P.G., 2012. The stratigraphic status of the anthropocene. The Holocene 22, 1491–1494.

Galloway, J.N., Dentener, F.J., Capone, D.G., Boyer, E.W., Howarth, R.W., Seitzinger, S.P., Asner, G.P., Cleveland, C.C., Green, P.A., Holland, E.A., Karl, D.M., Michaels, A.F., Porter, J.H., Townsend, A.R., Vörösmarty, C.J., 2004. Nitrogen cycles: past, present and future. Biogeochem 70, 153–226.

Hooper, D.U., Adair, E.C., Cardinale, B.J., Byrnes, J.E.K., Hungate, B.A., Matulich, K.L., Gonzalez, A., Duffy, J.E., Gamfeldt, L., O'Connor, M.I., 2012. A global synthesis reveals biodiversity loss as a major driver of ecosystem change. Nature 486, 105–108.

IPCC, 2007. Climate change 2007: the physical science basis. In: Solomon, S.D., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L. (Eds.), Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK and New York, NY, USA, p. 996.

Marsh, G.P., 1864. Man and Nature: or, Physical Geography as Modified by Human Action. C. Scribner, New York.

Molina, M., Rowland, E., 1974. Stratospheric sink for chlorofluoromethanes: chlorine atom-catalysed destruction of ozone. Nature 249, 810–812.

Monastersky, R., 2013. Global carbon dioxide levels near worrying milestone. Nature 497, 13–14.

Pandolfi, J.M., Connolly, S.R., Marshall, D.J., Cohen, A.L., 2011. Projecting coral reef futures under global warming and ocean acidification. Science 333, 418–422.

Prescott, G.W., Williams, D.R., Balmford, A., Green, R.E., Manica, A., 2012. Quantitative global analysis of the role of climate and people in explaining late quaternary megafaunal extinctions. Proceedings of the National Academy of Sciences 109, 4527–4531.

Rockström, J., et al., 2009. A safe operating space for humanity. Nature 461, 472–475.

Ruddiman, W.F., 2003. The anthropogenic greenhouse era began thousands of years ago. Climate Change 61, 261–293.

Ruddiman, W.F., 2013. The anthropocene. Annual Review of Earth and Planetary Sciences 41, 45–68.

Schiermeier, Q., 2011. Earth's acid Test. Nature 471, 154–156.

Smith, A.R., 1872. Air and Rain. The Beginnings of a Chemical Climatology. Longmans, Green, & Co., London, 600.

Steffen, W., Crutzen, P.J., McNeill, J.R., 2007. The anthropocene: are humans now overwhelming the great forces of nature. Ambio 36, 614–621.

Steffen, W., Grineald, J., Crutzen, P., McNeill, J., 2011. The anthropocene: conceptual and historical perspectives. Philosophical Transactions of the Royal Society A 369, 842–867.

Tyrrell, T., 2011. Anthropogenic modification of the oceans. Philosophical Transactions of the Royal Society A 369, 887–908.

Velders, G.J.M., Andersen, S.O., Daniel, J.S., Fahey, D.W., McFarland, M., 2007. The importance of the montreal protocol in protecting climate. Proceedings of the National Academy of Sciences 104, 4814–4819.

Vila, M., Espinar, J.L., Hejda, M., Hulme, P.E., Jarosik, V., Maron, J.L., Pergl, J., Schaffner, U., Sun, Y., Pysek, P., 2011. Ecological impacts of invasive alien plants: a meta-analysis of their effects on species, communities and ecosystems. Ecology Letters 14, 702–708.

Wroe, S., Field, J.H., Archer, M., Grayson, D.K., Price, G.J., Louys, J., Faith, J.T., Webb, G.E., Davidson, I., Mooney, S.D., 2013. Climate change frames debate over the extinction of megafauna in Sahul (Pleistocene Australia-New Guinea). Proceedings of the National Academy of Sciences 110, 8777–8781.

Zalasiewicz, J., Williams, M., Haywood, A., Ellis, M., 2011. The anthropocene: a new epoch of geological time? Philosophical Transactions of the Royal Society A 369, 835–841.

Anthropological Approaches to the Study of Education: The United States and Beyond

Bradley A Levinson, School of Education, Indiana University, Bloomington, IN, USA
Norma González, School of Education, University of Arizona, Tucson, AZ, USA
Kathryn Anderson-Levitt, Graduate School of Education and Information Sciences, University of California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R. Goodman, volume 9, pp. 4192–4196, © 2001, Elsevier Ltd.

Abstract

This article proceeds by highlighting the key organizing questions that animate the variety of anthropological approaches to education, past and present. With a focus on U.S. contributions to the field of anthropology of education, the article also acknowledges the contributions of other national traditions. With its early roots in studies of cultural transmission and 'culture and personality,' the field has grown to incorporate questions about cultural continuity and change, schools and inequality, and the education of racially minoritized groups. A wide range of interdisciplinary dialogues and advocacy efforts for social justice now characterize anthropological approaches to education.

Anthropology of Education as a Tradition of Inquiry

Most anthropological approaches to education have been developed in the subfield of cultural anthropology known as the anthropology of education. In the United States, this subfield has been institutionalized in one form as the Council on Anthropology of Education, which holds annual meetings in conjunction with the American Anthropological Association and publishes the refereed journal, *Anthropology and Education Quarterly*. However, many other scholars, including those serving as educators, activists, organizers, or consultants outside the academy, may identify themselves as anthropologists of education, if they are principally concerned with the social and cultural dimensions of teaching and learning. We describe the primary questions, concepts, and approaches of the field primarily from the U.S. perspective, while drawing in questions from other vibrant traditions of educational anthropology, such as those of Mexico and Argentina.

Perhaps the best way to articulate anthropological approaches to education is to state the questions that they most urgently pose and attempt to answer. Anthropology as a discipline asks the broad fundamental question, "How and why do human beings behave the way they do?," and it typically pursues the answer to these questions along five analytic dimensions: *historical*, *comparative*, *local*, *holistic*, and *cultural*. Anthropology is first and foremost a historical and comparative discipline. The discipline studies the evolutionary dimension of human existence, the sweeping historical development of human society from early hominid origins to postindustrial globalization. In and through such a historical perspective, anthropology also constructs knowledge about human universals and particulars through systematic comparison of different social groups.

The anthropology of education is firmly rooted in the branches of sociocultural, linguistic, and applied anthropology. Sociocultural anthropology's emphasis on *local* action and local meaning cannot be separated from its methodological preference for long-term participant-observation and ethnographic analysis. Related to this emphasis on local action is an insistence on understanding the context in which such action takes place. Yet even as the local is stressed, anthropology also has long advocated the principle of *holistic* analysis. This means that whatever the particular activity or practice under investigation, it is important to understand how it is interconnected with other dimensions of social life. In other words, nothing can be understood in isolation; each contextualized practice is always part of a larger whole, which ranges from biological and environmental influences to political and economic. Finally, anthropologists typically stress the *cultural* dimension of human society and action. Culture refers to the symbolic meanings that inform any observable social behavior or institution. For anthropologists, all human action is significantly influenced by cultural meanings, practices, and beliefs.

As a subfield of the discipline, then, anthropology of education begins by asking, first and foremost, "How and why do human beings *educate* the way they do?" Within this broad purview, the process of education can be defined as humanity's unique methods of 'teaching and learning' – that is, of *acquiring*, *transmitting*, *producing*, and *transforming* cultural knowledge for interpreting and acting upon the world. Anthropologists of education wish to understand how teaching and learning are organized socially and culturally. Of course, education is much broader than schooling, which is an institution of more recent historical invention. Until the development of agriculture and the rise of city-states some 10 000 years ago, tribal and band societies surely educated their young through complex and deliberate practices, but not in separate institutions like those we call schools. Rather, education was probably a seamless part of everyday life, and took place through the productive and ritual activities characterizing a society's way of life. At least as far back as Margaret Mead's classic 1928 work, *Coming of Age in Samoa*, anthropologists have examined the myriad forms of teaching and learning that comprise everyday social life – ranging from observation and imitation, and guidance and demonstration, to more intentional didactic techniques. Such studies of informal education have led to important findings about the nature of 'situated learning' (or situated cognition) and the 'communities of

practice' through which such learning takes place (Lave and Wenger, 1991).

In contrast to informal education, a school is typically a hierarchical setting, often age-graded, where, as Judith Friedman Hansen (1979) puts it, "learners learn vicariously, in roles and environments defined as distinct from those in which the learning will eventually be applied." Only since the beginnings of the modern period some 200 years ago – a period characterized by the rise of capitalism, large-scale urbanization, the consolidation of the nation-state, and the ubiquity of the printing press – have mass school systems been created, and much of human learning been channeled through schools. And even within schools, it is important to note, teaching and learning is not confined to the hierarchical roles of teacher and student; amongst students themselves, peer-to-peer learning constitutes a good deal of school-based education, too. So the second big question the anthropology of education asks is some variant of, "How exactly does education occur in schools, both within and beyond the classroom, and what relationship does that kind of teaching and learning have to educational processes occurring in many places outside of schools?"

Of course, the more specific questions about schooling and education often have emerged out of the encounter with the specific social and political realities, and the social movements, of the countries where anthropologists have found themselves working. Basic scientific questions have become entwined with political and applied questions. In the United States, for instance, the civil rights movements, racial desegregation, and the growth of ethnic diversity through the 'new immigration' of the 1960s brought forth the need to conceptualize the relationship between ethnic minorities and their schools. Thus, many U.S. anthropologists asked the question, "How do ethnic and racial minority groups organize their education in the community and home, and how do the educational practices of schools contrast or conflict with the educational practices of community and home?" Much work driven by such a question came to be known as the 'cultural difference' or discontinuity approach. By the 1980s, new patterns of immigration in Europe provoked a similar interest in cultural differences and in intercultural communication on that continent. At that time, too, U.S.-based anthropologists of education frequently noted the entrenchment of race, class, and gender divisions in the U.S. society. Though engagement with feminism, neo-Marxist scholarship, Pierre Bourdieu, and critical race theory, anthropologists of education asked questions about power, such as, "How do schools organize teaching and learning in ways that effectively reproduce social structural inequalities, and how do students and their families respond to such practices?" As we shall show further down, anthropologists of education also began to pursue social justice through their work, asking questions that were at least as action-oriented as analytic, such as, "How can teachers and learners be empowered to question and transform their educational circumstances, and how can institutions be designed to facilitate such transformation?"

In Latin America, of course, the questions have varied. Anthropologists of education in Mexico at an early stage drew together and critiqued theory from both the Anglo-American and the French traditions (e.g., 'cultural difference' and 'reproduction'), but soon developed their own concepts and questions. On the one hand, the ongoing denial of school access to millions of impoverished Mexicans prompted the question, "Whom does schooling serve, and why?" On the other hand, the prevalence of dominant assumptions about how schools and teachers worked brought forth the question, "How is the 'everyday life' of the school organized, and in what does 'teachers' knowledge' consist?" This prompted a deeper *historical* anthropological investigation of the roots of local Mexican school structures and cultures, in relation to national reforms and policies. More recently, in the context of contradictory neoliberal education policies and the new discourse of the 'pluralistic' nation, anthropological work has emphasized education in nonformal settings, youth cultures, interculturality, and the educational implications of class and racial–ethnic differences.

In Argentina, meanwhile, questions of diversity and democracy have been foremost. The development of educational anthropology went hand-in-hand with the struggle against military dictatorship and the restoration of democracy in the early 1980s. Argentine anthropologists of education have focused on mechanisms of power and control in schools, as well as class-based discrimination and exclusion. Among the key questions they have asked are, "How does power work in schools, and what is the role of the teacher vis-à-vis the state and the family?" More recently, with the rise in cultural diversity through in-migration from Bolivia and Paraguay, and the growing recognition of Argentina's own indigenous and ethnic minorities, Argentine anthropologists have increasingly asked, "How are cultural differences amongst students negotiated in schools, and how can schools recognize such differences without using them as a basis for exclusion or discrimination?"

In Brazil, anthropologists pursue several questions in concert with sociologists, social psychologists, and other colleagues, including these: "How does racism operate in the lives of teachers and students in the context of a myth of racial democracy? How should we understand youth not simply as victims of poverty and violence but also as strategic actors themselves? How should indigenous peoples' access to mainstream education be supported while respecting indigenous knowledge and pedagogical projects?" (Gomes and Gomes in Anderson-Levitt, 2011).

In Europe, meanwhile, research questions also develop out of local social contexts and disciplinary histories. In Denmark, in the context of the welfare state, many anthropologists and their colleagues in related disciplines focus on the question: "Are children and youth growing up in environments that allow them to grow as moral beings and to gain skills to support themselves and take part in democratic society?" (Anderson, Gulløv and Valentin in Anderson-Levitt, 2011). In Germany, where educational scholars conduct ethnographic studies on a range of topics, one group inspired by philosophical anthropology, by the Annales School of France, and by U.S. cultural anthropology asks: "How do children learn to become social persons through everyday rituals at home and at school?" (Wulf in Anderson-Levitt, 2011). Meanwhile, the new immigration has inspired in Italy, Spain, and other countries studies of children's languages and cultures under the rubric of interculturalism (Gobbo in Anderson-Levitt, 2011), which emphasizes learning across cultures, as opposed to simply

recognizing and problematizing cultural differences, as a multicultural or cultural difference approach is said to do. In Spain, anthropologists of education (Díaz de Rada and Jiménez Sedano, 2011) continue to raise fundamental questions about the ways that bureaucratic school categories propagate harmful assumptions about immigrant students and their families.

Key Concepts

From a broad anthropological perspective, we could say that the educational process in different societies fundamentally oscillates between an emphasis on continuity and an emphasis on change. This is because the challenges of evolution, broadly speaking, are thought to require a social group to adapt to novel circumstances through innovation, and then consolidate and perpetuate this adaptation through repeated inculcation. Education is seen as underwriting every human group's ability to adapt to its environment and thereby reproduce the conditions of its existence. According to Jules Henry (1963), Fred Gearing, George Spindler, and others of their generation, whose roots often lay in village or community-based studies of child socialization, education is, in this sense, a fundamentally conservative social process, only occasionally tinged by innovation. Accordingly, the main research questions are: "What allows a society to adapt to its environment while retaining some historical cohesion and continuity? How does a society conserve essential features of its cultural and technological repertoire?" Early anthropology of education largely sought to address these questions through an analysis of the process of *cultural transmission* – the passing on of basic cultural knowledge and values across the generations.

However, as questions of conflict and power came to the fore in the social sciences, along with the cognitive and linguistic turn, anthropologists of education began to question their functionalist assumptions, and to examine more closely how education contributed to change. "If cultural transmission occurred smoothly, how did societies challenge their own inertia? If education mainly served to mold the young into the cultural patterns of a society, how did innovation ever occur? If children were educated 'traditionally,' how did they respond to sudden and dramatic social change? And in complex, socially stratified societies, how were cultural messages differentially transmitted and received?" In response to such questions, attention turned away from cultural transmission, and the role of 'teachers,' toward *cultural acquisition*, or the role of the 'learner.' "A related emerging question was, how did relatively novice individuals acquire the basic cultural knowledge of a society, and what distinctive personal or categorical interests and traits might they bring to the learning process?" This question spawned a tremendously fruitful collaboration between anthropology and psychology, giving rise to the new field of 'cross-cultural psychology.' The work of the Soviet psychologist L.S. Vygotsky, with its sociohistorical emphasis on *activity* and the role of symbolic 'tools of mediation' in the relation between individual and society, has become central to this field. More recently, as we have mentioned, Jean Lave and Etienne Wenger (1991) have proposed a powerful theory of 'situated learning,' in which society is fundamentally composed of overlapping 'communities of practice' that serve as the vehicles for cultural acquisition.

Hand in hand with the emphasis on cultural acquisition came the increasing recognition of cultural complexity. Rather than a body of 'shared' or common meanings that one could identify with a particular society, culture was reconceived as a pool of symbolic resources, distributed unevenly across societies and social groups, which individuals can negotiate and appropriate with varying degrees of agency (Goodenough, 1976). An important overarching concept that has emerged to encompass the processes of cultural transmission and acquisition is that of *cultural production*. Education may be seen as a constant process of cultural production. Even in those cases where education is oriented primarily to the achievement of continuity, in a relatively closed social system, the theoretical possibility of modification and change always exists. If this is true in small-scale societies, then it must be all the more true in highly stratified, differentiated, and urbanized societies. So new key questions emerge: "In the process of acquiring transmitted cultural knowledge, how can individuals or subcultures modify or extend the knowledge, in effect organizing the knowledge for themselves while producing and adding new knowledge to the common stock? Or how can individuals or subcultures *resist* the transmission of knowledge, and thereby produce their own alternatives?" Indeed, one stream of work analyzes how differing 'models of the educated person,' produced both in and out of school contexts, may clash and conflict with one another (Levinson et al., 1996). For instance, in Doug Foley's (2010[1990]) critical study of a high school in south Texas, *Learning Capitalist Culture*, teachers and administrators seem to collude with mainstream (mostly white) students to produce a model of the smooth-talking, superficially compliant student who knows how to display highly packaged, or 'commodified,' knowledge. Meanwhile, rebellious (mostly Mexican) students, like the *vatos*, come to produce their own alternative model of the truth-talking, worldly-wise educated person, keen to point out the hypocrisy they see around them.

Research Methods and Related Disciplines

In approaching their specific contexts and exploring these questions, anthropologists of education typically employ ethnographic methods. Ethnography relies principally on in-depth participant observation for prolonged periods of time, accompanied by various kinds of interviews (individual, group, focus group, etc.), document analysis, and/or qualitative data prompts (conceptual mapping, open-ended surveying, photo-voice, etc.). Depending on the questions or context, anthropologists of education may also employ in-depth life history interviewing or mixed methods (some quantitative analysis).

Long-term participant observation is an approach also used for nearly a century by some schools of sociology, and also by sociolinguists, social psychologists, and scholars trained specifically in education studies. Thus, a sociology of education using similar research methods and asking similar questions has developed in parallel with educational anthropology. In the United States, sociologists using qualitative methods such as Hugh Mehan, Margaret LeCompte, and Annette Lareau, not

to mention social psychologists like Michelle Fine, or critical education scholars like Gloria Ladson–Billings and Sofia Villenas, have significantly influenced educational anthropology and been influenced by it.

In some countries, such as the United Kingdom and France, sociology of education is the dominant field and ethnographic studies are usually conducted by scholars who identify as sociologists or as education scholars – as in the journal *Ethnography and Education* or in the qualitative studies published in *Éducation et Sociétés*. While their research methods and the analysis of inequality tend to parallel educational anthropology's, the sociologists are more likely to focus on social class and, at least in France, less likely to turn their attention to race or ethnicity. In addition, sociologists of education generally study education 'at home,' whereas for reasons of history and structural dominance, some anthropologists in the United States, as well as in many wealthy Northern countries (e.g., Denmark, the UK, Japan), have conducted their studies 'abroad,' outside the country of their birth or training.

Until recently, anthropologists of education typically spent many months observing and interviewing people at a particular site (a community, a school, even a particular classroom), in order to develop a deep and nuanced understanding of the social and cultural dynamics of the context. More recently, anthropologists of education have extended the ethnographic method to multiple sites and contexts, in an effort to understand how signs and social forces circulate through and across many different contexts (e.g., Nespor, 1997), or how cross-cultural and cross-national contexts can be empirically compared (Anderson-Levitt, 2002; Tobin et al., 2009). Relatedly, a burgeoning anthropology of education policy and reform has yielded methodological models for studying processes and discourses across ethnographic sites and levels of social scale (Vavrus and Bartlett, 2009; Rosen, 2009; Koyama, 2010).

Directions and Trends in the U.S. Anthropology of Education

Concepts from critical social theory and a growing sense of urgency about the damage done by educational inequality have informed the field's growing commitment to social justice. Once anthropologists of education came to conceive of the complexity and full range of education as a dynamic and contested process of cultural production, this allowed for a link to *theories of schooling for social reproduction and for liberation*. As we have seen, in the late 1960s and early 1970s U.S. anthropologists of education tried to address persistent racial inequalities in educational achievement. One of the prevailing popular explanations for such inequality was that ethnic minority students were 'culturally deprived' in relation to their White counterparts, and therefore needed a kind of cultural remediation. This came to be known as the 'deficit' approach to educational achievement. Anthropologists quickly countered that it was not a matter of cultural deprivation, but rather of cultural *difference*. They argued that ethnic minority students tended to do more poorly in school because their cultures of origin used different epistemologies, styles of communication, and participant structures to educate children in the home. Such cultural attributes, they argued, differed sharply from the mainstream, middle-class cultural rules governing school life.

The Nigerian-born immigrant John Ogbu developed a critique of this reigning 'cultural difference approach.' He argued that anthropologists needed to take into account the 'cultural ecology' of a group's response to schooling. Making an important distinction between so-called 'voluntary' or immigrant minorities, and so-called 'involuntary' or castelike minorities (who typically had become part of a society through conquest), Ogbu (2008) showed that the question of school achievement could not be separated from the history of a group's structural position in society, and the repertoire of attitudes and practices that had developed out of that position. Involuntary minorities, such as African-Americans in the United States, had developed alternative and 'oppositional' cultural forms and strategies – what Ogbu called secondary cultural characteristics. Their problems with school could not be reduced to mere differences in primary culture or linguistic style. The problems were more deeply rooted in the history of subordination and its consequences.

Both Ogbu and the cultural difference anthropologists accepted the liberal premises of schooling, and they believed in the positive opportunities for social mobility that increased school achievement might provide. Yet by the middle of the 1980s, newer work in the sociology of education (e.g., Paul Willis, Pierre Bourdieu) and social theories of power helped anthropology focus a more critical lens on schools. This might explain why some anthropologists and educators would criticize John Ogbu for tending to place the burden of school performance on the involuntary minority families and students themselves. Now more than ever, schools were seen to serve dominant interests in the social reproduction of inequality, even as anthropologists also explored the paradoxical potential of schools as contexts to produce knowledge that might liberate individuals and groups from systems of domination.

For example, in a major work that culminates many years of fruitful collaboration, Hervé Varenne and Ray McDermott (1999) show how 'failure' is deeply and inexorably embedded in the U.S. school system through linguistic categories and interactional repertoires. This work also illustrates how the field has been enriched by important advances in the ethnography of communication and the ethnomethodological approach, which originally developed out of phenomenological sociology. Meanwhile, other work has taken up the basic insights of the cultural difference approach and the critique that John Ogbu made. Taking seriously the influence of historical experience and racial categories, new work has nonetheless tried to nuance or challenge Ogbu's formulations by looking more closely at institutional (school-level) effects and nondominant students' varying personal histories (Davidson, 1996; Fordham, 1996; Mehan et al., 1996).

Anthropologists have also worked alongside sociolinguists to challenge rigid, school-based definitions of literacy by documenting the many different ways that people use written texts, from grocery lists to love letters to political petitions seeking redress. Their work particularly underlines the functions of multiple literacies in multilingual societies, arguing against the presumption that what counts as literacy is academic proficiency in a dominant, often colonial language (Collins, 1989; Lee, 2007; Street, 1993). Other scholars of literacy, like Mexican

Elsie Rockwell and the French scholar Anne-Marie Chartier, place in deep historical and cultural context the formal instruction of literacy in schools, showing, among other things, how print materials are always mediated by teachers' verbal practices. Meanwhile, studies of multilingual and multimodal literacy practices within transnational social fields have explored the identities, relationships, and social networks that undergird concepts of transnational literacies. Transnational literacy studies explore language and literacy practices that transmigrants develop across geographical borders and emphasize the always situated, contextual, and ideological nature of reading and writing (Lam and Warriner, 2012).

In the United States, at least, many educational anthropologists teach in Schools of Education. Their work often and necessarily involves teacher training. At the same time, educational anthropologists have steadily undergone a process of self-critique which has led to increases in collaborative research design and the democratization of research relationships. Both of these trends partly explain the emergence and strength of school-based applied and action research, especially oriented to the empowerment of racially minoritized groups. As noted earlier, a common question in the field now is, "How can teachers and learners be empowered to question and transform their educational circumstances, and how can institutions be designed to facilitate such transformation?" Examples of such work across a range of contexts include Hugh Mehan (2012) and colleagues' attempts to study the consequences of a program to provide special support and mentoring to nondominant high school students, González et al. (2005) work on connecting students' 'funds of knowledge' with school curriculum, and Wyman et al. (2013) work on the ways that Indigenous youth often yearn to speak endangered Indigenous languages, regrouping resources from linguistic and cultural heritages while integrating wide-ranging symbolic resources from broader, dynamic worlds to build on local strengths in the face of language shift, endangerment, and inequities in schooling and society. Much of this work has been conducted by a new generation of scholars who are relative 'insiders' to the communities they choose to study, and thus challenge many of anthropology's traditional epistemological assumptions about the value of an 'outsider' perspective (Foley et al., 2001). Such scholars negotiate new 'decolonizing' methodological stances that involve community members throughout the research process with new youth studies approaches that foreground youth as important cultural actors in their own right (e.g., Cammarota, 2008; Lamont Hill, 2009; Urrieta, 2009).

Another trend in the field is the engagement with globalization processes as constituted by global flows of people, capital, and ideologies. These flows are seen to be in constant contact, simultaneously merging and diverging, and anthropological studies of education have embraced notions of continuous change and transformation. 'Multiple' subjects and 'multiple' cultures bear out the effects of globalization, fragmentation, and hybridity. Thus students, especially children of immigration, diaspora, and transnationalism, are increasingly represented as syncretizers, selecting and hybridizing cultural practices from multiple discursive fields. Within globalization processes, anthropologists of education have shown even greater concern with the "interrelationships among power, knowledge and identity" (Hall, 1999: p.123). Thus, along with cultural production, we may now speak of processes of *transculturation*, *interculturality*, and *transnational flows*, in which cultural knowledge becomes complexly combined and reformulated within both individual and collective repertoires.

A Concluding Note

As we have made clear, anthropologists of education have an interdisciplinary interest in education broadly defined, and with the many ways that teaching and learning has happened historically and culturally, and will continue to happen. Thus, they connect intellectually with historians and sociologists of education on the one hand, and with action researchers and reformers who seek to conceive of schooling and education more broadly outside the current historical constraints on the other. Indeed, many anthropologists of education have attempted to blur artificial binaries between scholarship and advocacy. Anthropologists of education have found that the margins between disciplines can be a space-where authorizing paradigms are breached and often displaced. As the field of anthropology as a whole has positioned itself as an 'engaged anthropology,' and anthropologists as social critics, it is worthwhile to note that the subfield of anthropology and education always has had deep historical roots in engagement, policy, and praxis. The confluence of advocacy anthropology and educational policy continues to be a space for the interrogation of inequities both in and out of schools.

See also: Communities of Practice; Cross-Cultural Study of Education; Educational Sociology; Ethnicity and Educational Achievement; Globalization in Education; Situated Learning (Learning In Situ); Social Inequality and Schooling; Socialization and Education: Theoretical Perspectives.

Bibliography

Anderson-Levitt, K.M., 2002. Teaching Cultures: Knowledge for Teaching First Grade in France and the United States. Hampton, Cresskill, NJ.

Anderson-Levitt, K.M. (Ed.), 2011. Anthropologies of Education: A Global Guide to Ethnographic Studies of Learning and Schooling. Berghahn Books, New York.

Bertely, M., 2002. Conociendo Nuestras Escuelas. Paidós, Barcelona.

Cammarota, J., 2008. The cultural organizing of youth ethnographers: formalizing a praxis-based pedagogy. Anthropology & Education Quarterly 39 (1), 45–58.

Collins, J., 1989. Hegemonic practice: literacy and standard language in public education. Journal of Education 17 (1), 9–35.

Davidson, A.L., 1996. Making and Molding Identity in Schools. State University of New York Press, Albany, NY.

Díaz de Rada, A., Jiménez Sedano, L., 2011. Variations on diversity and the risks of bureaucratic complicity. In: Levinson, B., Pollock, M. (Eds.), A Companion to the Anthropology of Education. Wiley-Blackwell, Malden, MA, pp. 408–424.

Foley, D., 2010[1990]. Learning Capitalist Culture: Deep in the Heart of Tejas, second ed. University of Pennsylvania Press, Philadelphia.

Foley, D.E., Levinson, B., Hurtig, J., 2001. Anthropology goes inside: the new educational ethnography of ethnicity and gender. Review of Research in Education 25, 37–98.

Fordham, S., 1996. Blacked Out: Dilemmas of Race, Identity, and Success at Capital High. University of Chicago Press, Chicago.

González, N., Moll, L., Amanti, C., 2005. Funds of Knowledge: Theorizing Practices in Households, Communities, and Classrooms. Lawrence Erlbaum, Mahwah, NJ.

Goodenough, W., 1976. Multiculturalism as the normal human experience. Anthropology and Education Quarterly 7 (4), 4–7.
Hall, K., 1999. Understanding educational processes in an era of globalization: the view from anthropology and cultural studies. In: Lagemann, E.C., Shulman, L.S. (Eds.), Issues in Education Research: Problems and Possibilities. Jossey-Bass, San Francisco, pp. 121–156.
Hansen, J.F., 1979. Sociocultural Perspectives on Human Learning: An Introduction to Educational Anthropology. Prentice-Hall, Englewood Cliffs, NJ.
Heath, S.B., 1983. Ways with Words: Language, Life, and Work in Communities and Classrooms. Cambridge University Press, Cambridge, England.
Henry, J., 1963. Culture against Man. Vintage, New York.
Jacob, E., Jordan, C. (Eds.), 1993. Minority Education: Anthropological Perspectives. Ablex, Norwood, NJ.
Koyama, J.P., 2010. Making Failure Pay: For-Profit Tutoring, High-Stakes Testing, and Public Schools. University of Chicago Press, Chicago.
Lam, E., Warriner, D., 2012. Transnationalism and literacy: investigating the mobility of people, languages, texts and practices in contexts of migration. Reading Research Quarterly 47 (2), 191–215.
Lamont Hill, M., 2009. Beats, Rhymes, and Classroom Life: Hip-Hop Pedagogy and the Politics of Identity. Teachers College Press, New York.
Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, Cambridge, England.
Lee, C.D., 2007. The Role of Culture in Academic Literacies: Conducting Our Blooming in the Midst of the Whirlwind. Teachers College Press, New York.
Levinson, B.A., Foley, D.E., Holland, D.C. (Eds.), 1996. The Cultural Production of the Educated Person: Critical Ethnographies of Schooling and Local Practice. State University of New York Press, Albany, NY.
Levinson, B.A., Pollock, M. (Eds.), 2011. A Companion to the Anthropology of Education. Wiley-Blackwell, New York.
Mead, M., 1928. Coming of Age in Samoa. Norton, New York.
Mehan, H., et al., 1996. Constructing School Success: The Consequences of Untracking Low-Achieving Students. Cambridge University Press, Cambridge.
Mehan, H., 2012. In the Front Door: Creating a College-Going Culture of Learning. Paradigm Publishers, Boulder, CO.
Moll, L., Amanti, C., Gonzalez, N. (Eds.), 2005. Funds of Knowledge. Lawrence Erlbaum, Mahwah, NJ.
Nespor, J., 1997. Tangled up in School: Politics, Space, Bodies, and Signs in the Educational Process. Lawrence Erlbaum, Mahwah, NJ.
Neufeld, M.R., Thisted, J.A. (Eds.), 1999. 'De eso no se habla': Los usos de la diversidad sociocultural en la escuela. Universidad de Buenos Aires, Argentina.
Ogbu, J.U. (Ed.), 2008. Minority Status, Oppositional Culture and Schooling. Routledge, New York.
Rockwell, E., 1995. La Escuela Cotidiana. Fondo de Cultura Económica, Mexico City.
Rosen, L., 2009. Rhetoric and symbolic action in the policy process. In: Sykes, G., Schneider, B., Plank, D. (Eds.), Handbook of Education Policy Research. Routledge, New York, pp. 267–285.
Spindler, G., Spindler, L. (Eds.), 2001. Education and Cultural Process, third ed. Waveland Press, Prospect Heights, IL.
Street, B.V., 1993. Cross-Cultural Approaches to Literacy. Cambridge University Press, Cambridge.
Tobin, J., Hsueh, Y., Karasawa, M., 2009. Preschool in Three Cultures Revisited. University of Chicago Press, Chicago.
Urrieta Jr., L., 2009. Working from within: Chicana and Chicano Activist Educators in Whitestream Schools. University of Arizona Press, Tucson.
Varenne, H., McDermott, R., 1999. Successful Failure: The School America Builds. Westview Press, Boulder, CO.
Vavrus, F., Bartlett, L. (Eds.), 2009. Critical Approaches to Comparative Education: Vertical Case Studies from Africa, Europe, the Middle East, and the Americas. Palgrave Macmillan, New York.
Wax, M., Diamond, S., et al. (Eds.), 1971. Anthropological Perspectives on Education. Basic Books, New York.
Wolcott, H.F., 1973. The Man in the Principal's Office. Holt, Rinehart and Winston, New York.
Wyman, L., McCarty, T., Nicholas, S. (Eds.), 2013. Indigenous Youth and Multilingualism: Language Identity, Ideology, and Practice in Dynamic Cultural Worlds. Routledge, New York.

Relevant Websites

Bibliotheksservice-Zentrum Baden-Württemberg (BSZ). Pädagogische Anthropologie: Allgemeines und Deutschland. From: http://bibscout.bsz-bw.de/bibscout/D/DF/DF4500-DF4509/DF.4500 (retrieved 16.02.14.).
http://www.aaanet.org/sections/cae/ – Council on Anthropology of Education.
http://www.ethnographyandeducation.org/ – The Ethnography Network (Network 19) of the European Educational Research Association.
Jarvis, H.W., 2014, 2010. Worldwide Email Directory of Anthropologists. From: http://wings.buffalo.edu/WEDA/ (retrieved 16.02.14.), search for 'education'.
http://ica.institutos.filo.uba.ar/seanso/pae/boletin/index.html – Program in Anthropology and Education, University of Buenos Aires, Argentina.
http://www.berghahnbooks.com/title_detail.php?rowtag=Anderson-LevittAnthropologies – Anthropologies of Education by Kathryn Anderson-Levitt.

Anthropological Research, Ethics of

Robert Albro, Center for Latin American and Latino Studies, American University, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Anthropology's relationship to ethics in research has been driven by periodic disciplinary controversies, often in close association with the involvement of anthropologists in successive war efforts, primarily in the U.S. Over time disciplinary ethics have increasingly focused on the unique exigencies of ethnography, as a disciplinary-specific method. While professional anthropological organizations developed codes of ethics beginning in the Vietnam era, historically, disciplinary ethical language has been used in two ways: as a basis for internal disciplinary self-policing and as a means to claim public professional standing as a social science. These uses continue to be prevalent today. But anthropology's relationship to its ethics also has been historically dynamic, changing with changes in disciplinary identity, values, and priorities. Less recognized are the ways that disciplinary ethics have been regularly reconstituted in close proximity to the frontiers of the identity of anthropology as a changing project. Anthropology's ethics reflect the discipline's specific history and identity debates, and they are one key index of these changing frontiers. This is most evident with respect to anthropology's ambivalent relationship to: science, the nation-state, and encompassing normative structures.

One sign of the increasing 'primacy of the ethical,' to borrow from Nancy Scheper-Hughes (1995), is the extent to which anthropology is inclined to regularly revisit and to revise its own ethics standards, usually in reaction to perceived ethical crises. If historically these crises have been precipitated by the involvement of anthropologists in different war efforts, over the years they include other cases where the discipline has left the academy. These crises have seemed to come more often of late, and are more regularly revisited, with anthropologists frequently electing to bring the language of ethics to bear when raising questions about the scope and limits of disciplinary practice. As has been often noted, the use of the discourse of ethics accompanies accusations about breaches of professional conduct in the mode of inward- and outward-looking disciplinary self-policing.

'Ethics' increasingly inform or subsume disciplinary debate about researchers' relationships to counterparts. But ethics in the mode of self-policing is only one sign of what has been a broader turn to critical examinations of ethical, moral, and normative structures and agencies. Anthropology's relation to ethics has become more expansive, informing new research on audit cultures, identity, personhood, the state, rights, advocacy, and relationships of the local to the global. This more expansive relationship still includes a concern for the ethics of anthropology, that is, the set of normative principles understood to govern moral, right or acceptable professional conduct. But it is also informed by the notable growth of an emerging anthropology of ethics, which takes 'ethics' as a subject of ethnographic projects. As Faubion (2011) explains it, such research comparatively pursues the 'becoming and maintenance of the ethical subject.' The fruitful traffic between these two forms of ethical engagement currently informs the growing presence of ethics as a part of the disciplinary landscape.

Ethics and Anthropological Science Wars

During the first decades of the twentieth century, when Franz Boas was busy promoting his eventually triumphant historical particularist vision for cultural anthropology at Columbia University, he also provoked the discipline's first ethics controversy in 1919 with his published assertion that during WWI certain American anthropologists 'prostituted science' by using disciplinary activities as a cover for government-sponsored espionage. Much can be said about this event. But of interest here is that the force of Boas's claim of an ethical breach depends, in turn, upon his strong appeal to the integrity of 'science,' as a clearly demarcated professional vocation which – Boas wanted to be clear – should definitely include a then nascent and actively professionalizing *science of anthropology*.

As ongoing debate about the extent and kind of anthropology's involvement in the activities of the national security state continues to make clear, disciplinary invocations of ethics are a particularly public mode of professional self-policing. But Boas's appeal to science also expressed his specific time and place, articulating a specific professional version of anthropology, which has undergone multiple transformations since. The *science of anthropology* Boas defended in his outraged letter to *The Nation* did not begin with 'value judgments.' Nor should scientists, 'in the service of truth,' involve themselves in 'political affairs.' In broad terms, Boas's anthropology was outside of government and in academic departments, objective, empirically driven, nonnormative, but also skeptical of classification and of a science of culture formulated as generalizable laws.

If an example of the self-reflexive use of ethics to claim professional high ground – a usage still prevalent today – Boas's claims for an anthropological science here suppressed several well-known and founding ambiguities of his project for a newly professional anthropology, many of which are on clear display in his 1887 essay on 'The Study of Geography.' To *The Nation* he strategically foregrounded the scientific over the humanistic, his training in physics over his interest in history, and the naturalist over what he called the cosmographer. At this key professionalizing moment, and even as his understanding of science was considerably more nuanced, Boas at least publicly aligned the ethics of anthropology with an empirically rigorous objective evidence-based model of science consistent with the positivist natural sciences of the day.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12190-7

We can compare the ethical crisis provoked by Boas's (1919) letter to a current controversy involving another eminent anthropologist, Marshall Sahlins. Sahlins resigned in 2013 from the National Academy of Sciences to publicly protest the election of Napoleon Chagnon. The latter's widely cited conclusions about human evolution, genetic selection, and the role of violence among the Yanomami of Venezuela and Brazil, have been highly controversial among anthropologists. And according to Sahlins and others, Chagnon's ethics in the field have also been suspect. Sahlins has described Chagnon's methods as a case of 'research and destroy,' a description linking Sahlins's objections to his other concern, the militarization of the social sciences.

The cases of Boas and Sahlins are notably similar. Over their careers, each has influentially represented the discipline as a 'discipline' among peers and beyond. Both voiced uncompromising disagreement with any anthropological involvement with the U.S. national security state. Both have regularly promoted cultural explanations to discredit the biologically rooted pseudosciences of their respective eras (for Boas, eugenics and racialist evolutionary thinking; for Sahlins, sociobiology). When engaged in debate, both variously emphasized the primacy of empirical findings as an answer to bad science or bad theory. And, of course, both took colleagues publicly to task for violating perceived disciplinary ethical standards, leading to their resignations from the National Research Council.

But these similarities conceal equally notable differences between the two, most evidently, the ways they connect the discipline of anthropology to different notions of science (and of history), and as these connections lend force to their ethical claims. Sahlins connects Chagnon's poor ethics, specifically, his apparent failure to obtain 'informed consent' from Yanomami research subjects, to sociobiology's 'theoretical and empirical underpinnings,' succinctly, the selfish-gene Darwinism of Richard Dawkins and ilk with which Chagnon's work is engaged. And sociobiology fits with a hypothesis-testing 'natural science sense of anthropology' preferred by the National Academies, as opposed to Sahlins's empirically grounded historical explanations of cultural determination. Chagnon's defenders, in turn, regularly depict Sahlins's criticisms as 'antiscientific.'

Most tellingly, Sahlins draws a direct connection between Chagnon-type sociobiology and the military-funded research projects of the National Research Council. In his words, both traffic in the same Western ideology of the 'innate character of mankind,' and aspire to 'making the world safe for self-interest.' For Sahlins, Chagnon's work is an egregious illustration of the 'confusion of capitalist greed and human nature.' Here Sahlins reveals a very different set of commitments to those of Boas, indeed, an ideological axe. Sahlins is making a value judgment, while using professional ethics to defend a moral position about scientific practice as it is connected to competing cultural conceptions of personhood and identity, a judgment which he understands to be part and parcel of the ethical practice of anthropology.

Unsurprisingly anthropology has changed dramatically over the course of the intervening years between the ethical crises provoked by Boas and Sahlins, changes which cannot be summarized here. But Sahlins's objections should be taken in a context of the changing expectations about ethnography specifically, and anthropology generally, as a kind of project. If Boas's ethics were focused on purging ideology (specifically, nationalism) from professionalism, the force of Sahlins's objections assumes the greater ethical urgency invested in the politics of ethnography, with a concomitant greater emphasis on 'informed consent' and 'do no harm' than in Boas's time. His ethical accusations express a change in the ways ethics are used now: less about what is defensibly professionally 'scientific' (as with Boas) and more about the circumstances of ethnographic relationships, where the work of ethnography is often also understood to include previously extraethnographic responsibilities to the values and priorities of counterparts in the field.

Despite the fact that the discipline, in its diversity, has equivocated about the ethics of advocacy for some time, AAA leaders carefully parsed this most recent iteration of the Sahlins–Chagnon debate, responding that both advocacy and science are 'at the core of our discipline.' They go on to explain, "Advocacy suggests at minimum an ethical position to try to protect and better the lives of individuals." The mode of science conversant with this is one recognizing science as a practice 'undertaken in a social context.' This is a far cry from Boas's value-neutral empirical *science of anthropology*. Sahlins's ethics appeal more to what Boas identified as 'humanistic,' and as preferable to an anthropological science modeled on the physical sciences. His updated humanistic anthropology is also equated with more politically-engaged disciplinary practice. This shift is, at once, an expression of the discipline's ethical history and has implications for the contemporary meaning of professional ethics.

Boas and Sahlins, both, address alignments of causality and explanation among biology, history, and culture. Both demote biology while elevating culture. Both labor to describe empirically-informed historical explanations for cultural agency. Where they most obviously differ is in the identification of a properly scientific anthropology. The frontier of this difference has been a fault line of anthropological struggle for self-identification since its inception. The ethical turn of each directly troubles this frontier, which Boas originally conceived in the terms of physicist and cosmographer. Almost a century later, this frontier had shifted to the tensions between ethnography and advocacy. Even in cases of self-policing focused on delimiting professional conduct, ethics are also a dynamic catalyst for theory-building, in this case, for right relations among biology, history, and culture. Navigating what Zigon (2007) describes as moments of 'moral breakdown,' successive ethical turns have also productively defined changing positions along the movable frontiers of anthropological science. When it comes to the contentious relationship of anthropology to science, ethics do productive work.

Ethical Genealogies and Normativity

We now have various historical overviews of anthropology's ongoing, changing, relationship to ethics (see Fluehr-Lobban, 2003; Whiteford and Trotter, 2008), which I will not aspire to summarize here. Boas and Sahlins have their parts. But I do want to emphasize implications of the fact that this has

been a dynamically changing relationship. While not in any way an exhaustive account, Murray Wax (1987) charts disciplinary shifts in ethical attention as closely related to successive engagements with, and attitudes about, the U.S. nation-state. He begins with the WWI era, citing Boas's drive for a *science of anthropology* with 'an ethical calling above the petty rivalries of nationalism.' Between the wars, the discipline became more engaged with challenges of relativism and description of the 'varieties of ethical systems.' During WWII, Wax tells us, ethics became more associated with defense of the 'free world.' After WWII they came to align with a refusal to be employed by national agencies. During the Vietnam era, this coalesced into a more focused rejection of any dealings with the military – a still very active disciplinary discussion today.

I am less concerned with whether Wax's account is accurate – it has been debated – and more interested in Wax's observation that anthropology's ambivalent and evolving relationship to the nation-state, including the machinery, agencies, logics, and interests of the state, is one disciplinary ethical frontier. This ambivalence is of course brightly written into the often fraught history of anthropology's relationship to colonialism and other exercises of state power, a long story that cannot be briefly told. But a postcolonial anthropology has been preoccupied with the effects of agencies of state power upon marginal populations, including the regulatory logics of the state. As the ongoing fruitfulness of concepts like Foucault's 'biopower' or 'governmentality' attest, the exercise of state power upon local worlds continues to be a major disciplinary focus even as it remains a source of ethical discontent.

But the fact that, as Wax notes, disciplinary ethical relationships to the state have been anything but static points to the limits of ethics when conceived as a codified body of abstract knowledge and as an extradisciplinary or universalizing language of self-policing. Nevertheless, when the Code of Ethics of the American Anthropological Association (AAA) has in recent years been used as a lexicon for criticism and sanction of wayward practitioners, this is often with the strategic fiction of ethics as if self-evident, transhistorical, legalistic, prescriptive, regulatory, and as effectively bounding the discipline's professional conduct at any given moment and over time. The dynamic genealogical composition of the discipline's relationship to ethics is ignored and, with it, the productivity of ongoing ethical debate.

The categorical evocation of ethics is at odds with where an emerging anthropology of ethics seems to be heading, suggesting that the discipline's self-reflexive understanding of its practice to its theory production requires more attention. A notable part of this picture is anthropology's increasing rapprochement with the normative after decades of deep suspicion about the totalizing or universalizing claims of normative frameworks. This is a story of several converging genealogies of contemporary anthropological concerns, of which I will point to just a few to highlight anthropology's expanding ethnographic engagement with normative structures.

One genealogical thread has been the discipline's response to the evolving federal oversight of scientific work in the U.S. beginning in the 1970s – itself an expression of the discipline's entanglement with the state. During this period many ethnographers shifted attention from stateless societies to an anthropology in and of the state. In this earlier context of shifting priorities, medical anthropologists – with one foot in the practicing world – engineered a version of 'studying up,' as among the first to take up ethics as an explicit subject of disciplinary research in their own right. In this context, anthropology's emergent critique of bioethics was catalyzed by the development and subsequent regulatory influence of the Belmont Report in the U.S. in 1979, which provided the framework to establish federal standards for the ethics of human subjects protection.

The Report characterizes human subjects as autonomous agents who can exercise individual choice, to the ends of seeking to maximize benefits and to minimize risks. In the broadest terms, medical anthropologists took bioethics, as an abstract framework, to task for disregarding the relevance of social contexts for medical ethical decision-making, while uncritically incorporating a rational choice model of decision-making that assumes the primacy of individual choice. In contrast to bioethicists, these medical anthropologists emphasized the need for a culturally informed understanding of choice, personhood, or 'wellness,' as these vary across different local social worlds, and as caught up in and apprehending the biomedical regimes of modern medical practice in the U.S. and elsewhere.

Several anthropologists have now told the story of the emergence of professional scientific institutional review boards (IRBs), with their priorities framed by the Report's conclusions and the methods of biomedical research, and how this was at best an awkward fit for participant-observation ethnography. But one effect of this emergent federal framework was to further focus anthropological ethics upon the particularities of interaction between fieldworker and hosts, or ethnographers and their counterparts. Whiteford and Trotter's (2008) recent discussion of common problems in anthropological ethics registers this ethnographic slant, given its systematic attention to privacy, rights, harm, justice, advocacy, deception, coercion, voluntary participation, confidentiality, and informed consent. Recent discussions by anthropologists of IRBs continue to dwell on the various questions relating to informed consent, all bracketed by the problem of power in research relationships.

Established critiques developed by Arthur Kleinman (1995) and other medical anthropologists of the medical regime of bioethics – as a regulatory and normative expression of state authority – are an early precedent for the discipline's attention to encompassing normative frames and an example of the ways ethics have become a subject of research. One outcome of this engagement between medical anthropologists and bioethics has been the disciplinary rejection of radical or naïve ethical relativism, as untenable, and as corresponding to a now outmoded and value-neutral comparative exercise among cultures treated as if self-contained distinct worlds. In other words, disciplinary ethics – as with the case of its critical response to bioethics – have made more room for critical engagements with encompassing normativities as these often translate the local agency of the state, other trans-local or global agencies.

The questions of informed consent and prohibition to do no harm, which form the center of anthropology's professional ethics, acquired their current meaningful force in large part in the context of the discipline's grappling with bioethics, as it was then coming to be the prevailing regulatory and normative frame for federally-funded scientific research (including social

sciences like anthropology), and as this gained momentum in the 1970s. Practicing medical anthropologists initiated regular critical dialogue between disciplinary ethical conceptions and bioethics, given their status as a cornerstone of the extra-disciplinary landscape of science ethics, and as one expression of the logic of state power. In the process, anthropology's own ethical conversation has become more sharply defined by its particular disciplinary history and the ethnographic encounter, as, in the words of Rena Lederman (2013), about a specific 'culture of research.'

An Anthropology of Values

Anthropology has been called a 'reformer's science.' It would not be inaccurate to describe anthropology's early professional development as the pursuit of an ethnographically informed social science of comparative ethics. But such a pursuit was pervasive and not foregrounded, instead implicit in what was then understood to be a comparative science of religion, government, politics or culture. This has been suggested as a reason for the relative lateness on the part of the discipline's professional organizations to address ethical oversight as a professional responsibility. The AAA, for example, only adopted a formal position on ethics in 1967 and an ethics code in 1971, since modified multiple times (most recently in 2012).

In parallel and recently, attention has been given to developing a comparative anthropology of morality. This impetus is intended to rectify the deficiency that, in Laidlaw's (2002: 311) words, "Our discipline has not developed a body of theoretical reflection on the nature of ethics." In the context of these discussions, different arguments have been advanced about why this has been the case. One of these has highlighted anthropology's debt to Durkheim, identifying a tendency to equate the moral with the social. As Joel Robbins (2012) has explained, "Because Durkheim defined all properly social action as ethical, anthropologists could assume that everything they studied was ethical, and thus that issues of ethics needed no special attention of their own." Durkheim's legacy, in short, confined the consideration of ethics to the moral lockstep of mechanical solidarity.

A second, more specifically American, account identifies a lack of sustained attention to ethics as a Boasian legacy. A brief recapitulation of this argument, as rehearsed by Mark Goodale (2006), goes like this: since U.S. anthropology came to be synonymous with so-called cultural relativism, and since cultural relativism was typically equated with the categorical rejection of all universalizing normative frameworks (e.g., 'universal human rights'), ethics never evolved into a specific topic of consideration. The 1947 American Anthropological Association's Statement on Human Rights is often exhibit A in this account. A stance of relativistic neutrality and insistence on the comparative moral and political equality of human cultures effectively paralyzed the discipline's methodological and theoretical ability to address normative questions.

Whether anthropology historically has in fact been insufficiently cognizant of the moral life of societies is debatable, but current lively discussion of both the Durkheimian and Boasian reasons for the apparent historical neglect of ethics is in fact one sign of the growing attention to ethics in both ways: as a regulatory discourse of disciplinary conduct and as a subject of research programs. Michael Brown's articulation of an updated 'cultural relativism 2.0' illustrates this turn. In Brown's hands anthropology is a discipline constructively situated to engage the traffic between the 'values' of specific societies and the larger scale moral communities in which they are embedded. Revealingly, Brown (2008: 372) defines cultural systems as themselves "morally charged fields of action and meaning." Cultures, in other words, are significant because they organize specific moral commitments.

Brown's updated discussion of relativism notably makes room for an anthropology that is also regularly concerned with normative frameworks, if primarily at the conjuncture of the global with the local. The discipline's engagement with the normative is part of an ethical turn, and relatively new. The AAA's 1999 Statement on Anthropology and Human Rights, for example, represents an about-face from its 1947 relativist position. If resisting conceptions of the normative as a totalizing system of rationalized abstractions distilled into sets of codes or rules, the 1999 Statement nevertheless establishes an 'ethical commitment' on the discipline's part to recognize and enable 'the human capacity for culture,' defined as a human right. The 1999 Statement stakes out a specifically anthropological vernacular with regard to human rights that at least partially recasts the AAA as an international human rights NGO. Goodale has called this the 'normativization of culture through rights discourses as a political strategy.'

In recent decades, ethnographic research, most obviously with indigenous peoples, has also often incorporated an advocacy role, with anthropologists critiquing, in negotiation with, and leveraging national or international normative precedents on behalf of, and in concert with, indigenous groups increasingly inclined to treat 'human rights' as a part of their own legal arsenal in ongoing transactions to extract resources or recognition from recalcitrant states. With bioethics, this composes part of the ethical frontier of anthropology's postcolonial critique of the agency of the nation-state. Advocacy, treated as an extension of ethnography, encourages normative judgments to be made about particular cultural practices or prevailing global norms while also strategically promoting instrumental uses of normative orders to achieve political goals with counterparts. This orientation to ethics is consistent with what Goodale calls an 'ecumenical anthropology,' that is, an ethically committed ethnographically grounded criticality with respect to normative projects. It is also characterized by the return of values talk in disciplinary discourse.

Anthropological values talk has several sources. The effort to animate a comparative moral anthropology is one of them, concerned as it is with 'cultural values' defined as 'cultural conceptions of the good or desirable.' If resisting any view of morality as a totalizing system, in this emerging research the assignation of values situationally hierarchizes priorities in ways helping to define what counts as moral action. David Graeber's (2001) anthropological 'theory of value' offered as a counter to the neoliberal vision of people as calculating maximizers, draws on anthropological discussions of language, exchange, wealth, and desire, to develop an appreciation of value as a model for human 'meaning-making,' comparatively connecting symbolic systems to the work of defining what is important as a moral project.

These several renovations of value talk have departed entirely from the Boasian suspension of 'value judgments,' reinserting politics, hierarchy, and normative structures back into considerations of culture. In the process the attention to values is reinventing an anthropological conception of culture by opening up new frontiers with once competitive uses of the term (e.g., Roy Wagner's argument about the relationship of anthropological cultures to high culture), while moving away from once prominent disciplinary interpretive vocabularies (e.g., Geertzian 'culture as text'). An ethnographic focus on ethics informs this disciplinary increase in values-talk, and supports new theory-building (e.g., new relations of the moral to the cultural). It has also informed how ethnographers engage indigenous legal struggles for identity and autonomy (e.g., the work of advocacy). All of this points to, in Zigon's words, the creative aspect of ethics as 'conscious work on the self.' Anthropology's more expansive turn to ethics, then, includes self-policing but also a growing appreciation for the ways ethics are theoretically productive.

Ethics along Emergent Frontiers

Revealingly, anthropology's ethics arguments often elide its ethical lexicon with the moral language of values, couched within a broader political critique of the behavior of states and of state institutions like military and security agencies. Ethical injunctions (e.g., against 'secret and clandestine' work) are represented as disciplinary 'core values.' And these core values define the possible extent of what anthropology can, or better, should be. As such, much of the back-and-forth of recent years has been over the extent to which the language of ethics effectively circumscribes the discipline as a self-reflexively moral community. As Lederman and others have emphasized (cf Plemmons and Albro, 2012), ethics, used as a linchpin to identify the core values of anthropology as a moral community, subjects disciplinary identity to its own emergent ethical theorizing in such a way that the discipline, itself, is presented as if a Sahlinsesque historically contingent cultural or moral order.

A distinguishing characteristic, then, of the ways that anthropology has taken up ethics of late is the regular interaction of these two modalities – the ethics of anthropology and the anthropology of ethics. And the discipline continues to complicate the historical 'duplexity' (Pels, 1999) of its relationship to ethics. Disciplinary ethical discussions at present combine appeals to extradisciplinary and impartial moral standards with political interests in the mode of advocacy, but now also with value-laden public commitments to anthropology as a historically contingent moral community – as a 'culture of research' shaped by the steady reframing of its ethics in terms of the unique particularities of ethnography and its incumbent relationships. A potential trap of the application of an anthropology of values to disciplinary practice is public disciplinary self-presentation, again borrowing from Zigon, as a 'unique local moral and ethical assemblage.' Increased public perception of the vernacularization of anthropology's ethics could, however, have the undesirable effect of undermining its standing as a social science.

At the same time, treating disciplinary ethics as occasional conflicts, regular debate, and moving frontiers is to imagine ethics as a necessary but fruitful source of vital disciplinary questions, knowledge and theory. The present review of several of anthropology's ethical genealogies, and of the dynamism of disciplinary ethical frontiers with science, diverse agencies of the state, bioethics, changing ethnographic relationships, normativities, and comparative moralities, points to the productive, theoretical, potential of regular ethical debate, a potential anthropological ethical researchers and theorists are now starting to realize.

And this is ongoing. One such ethnographically informed frontier is the more fluid disciplinary ethical conceptions of, and relationships between, privacy, secrecy, transparency, and new proprietary realities. If anthropology's relationship to secrecy has been fraught, this relationship is being creatively stretched by the changing shapes and meanings of 'openness,' now as incorporated in digitally-engaged 'open source' and 'open access' debates, ethnographic projects, collaborations, and new 'circulations of culture' (see Kelty et al., 2008). Current research and debate on openness is exploring often non-self-evident relationships to collective action, intellectual property, system standardization, interoperability, technical networks, free speech, collaborators now conceived as 'source communities,' and other new working relationships. 'Open access,' in other words, is one current and constructive source of ethical theory in anthropology's broadening relationship with its ethics.

See also: Advocacy in Anthropology; Anthropology, History of; Bioethics; Ethical Questions in Social and Behavioral Sciences, History of; Ethics and Values; Ethnography; Militaries and Militarization, Anthropology of; Norms; Reflexivity; Research Ethics: Research; Science, History of; Social Work Ethics; Value: History of the Concept.

Bibliography

Boas, F., 16 October 1919. Scientists as Spies. The Nation.
Brown, M.F., 2008. Cultural relativism 2.0. Current Anthropology 49 (3), 363–383.
Faubion, J., 2011. An Anthropology of Ethics. Cambridge University Press, Cambridge, UK.
Fluehr-Lobban, C., 2003. Ethics and the Profession of Anthropology: Dialogue for Ethically Conscious Practice. Altamira Press, Walnut Creek, CA.
Goodale, M., 2006. Ethical theory and social practice. American Anthropologist 108 (1), 25–37.
Graeber, D., 2001. Toward an Anthropological Theory of Value: The False Coin of Our Own Dreams. Palgrave Macmillan, New York.
Kelty, C.M., et al., 2008. Anthropology of/in circulation: the future of open access and scholarly societies. Cultural Anthropology 23 (3), 559–588.
Kleinman, A., 1995. Writing at the Margin: Discourse between Anthropology and Medicine. University of California Press, Berkeley, CA.
Laidlaw, J., 2002. For an anthropology of ethics and freedom. The Journal of the Royal Anthropological Institute 8 (2), 311–332.
Lederman, R., 2013. Ethics: practices, principles, and comparative perspectives. In: Carrier, J., Gewertz, D. (Eds.), The Handbook of Sociocultural Anthropology. Bloomsbury Press, New York.
Pels, P., 1999. Professions of duplexity: a prehistory of ethical codes in anthropology. Current Anthropology 40 (2), 101–136.
Plemmons, D., Albro, R., 2012. Practicing ethics and ethical practice: the case of anthropologists and military humanitarians. Humanity 3 (2), 179–197.

Robbins, J., 2012. On becoming ethical subjects: freedom, constraint, and the anthropology of morality. Anthropology of the Century 5. http://aotcpress.com/articles/ethical-subjects-freedom-constraint-anthropology-morality/.

Sahlins, M., 2013. The national academy of sciences: goodbye to all that. Anthropology Today 29 (2), 1–2.

Scheper-Hughes, N., 1995. The primacy of the ethical: propositions for a militant anthropology. Current Anthropology 36 (3), 409–440.

Wax, M., 1987. Some issues and sources on ethics in anthropology. In: Cassell, J., Jacobs, S. (Eds.), Handbook on Ethical Issues in Anthropology. American Anthropological Association, Washington, DC, pp. 4–10.

Whiteford, L.M., Trotter II, R.T., 2008. Ethics for Anthropological Research and Practice. Waveland Press, Long Grove, IL.

Zigon, J., 2007. Moral and ethical assemblages. Anthropological Theory 10 (1–2), 3–15.

Anthropological Writing

Alma Gottlieb, University of Illinois at Urbana-Champaign, Urbana, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Until the 1970s, the modal career in anthropology revolved around the regular publication of articles in scholarly journals, and the intermittent publication of book-length 'monographs.' Anthropologists now challenge the supposition that these two writing genres embody the only legitimate means to share knowledge. This article charts the variety of new modes of writing that anthropologists increasingly embrace – from memoir, biography, and fiction to flip books, graphic novels, and invented child care guides – as means to disseminate their findings and unique perspectives to broader readerships.

The Myopia of Monopoly

As Johannes Fabian has observed (2011: p. 16), the very phrase 'ethnographic writing' is redundant given that the Greek roots of the word 'ethnography' mean, literally, 'writing (about) people.' Despite this etymological foundation, the early decades of the discipline of anthropology saw little critical reflection on our writing practices. Instead, until the 1970s, the modal career in anthropology revolved around the regular publication of articles in scholarly journals, and the intermittent publication of a particular form of book termed the 'monograph' (Wulff et al., 2009). Typically, both writing forms featured a predictable structure. A general overview of a single, seemingly complete, and relatively isolated society classically preceded (or underlay) an analytic argument revolving around a basic theme, and one or more 'domains.'

Give or take the occasional effort to stray from the path of writing orthodoxy, both the article and the monograph typically relied on two problematic assumptions revolving around their literary production. First, our writing habitus normally presumed that the anthropologist-as-analyst was best left invisible to the reader, a scholarly Peeping Tom peering through the windows of the other's house. Second, we routinely took for granted that the cumulative emotional effect of this and other features of the work produced a reading experience that should approximate that provided by the natural science report more than that offered by the literary essay. Notwithstanding a few adventurous scholars, until recently the academy has largely left unchallenged, the grip that this model for the properly written scholarly article and monograph have long held on our discipline. To borrow a legal trope, the dominance of these two writing genres constituted a veritable monopoly.

Over a century ago in the United States, the Sherman Antitrust Act of 1890 effectively restricted the reach of monopolies in business; elsewhere, the US law has its analogue in most other modern nations (Joelson, 2006). By contrast, we have not yet passed our own Antitrust Act within the academy. Nevertheless, anthropologists have begun challenging *de facto* the supposition that the scholarly article and book-length monograph, structured in the ways summarized above, embody the only legitimate means for us to share knowledge. Equally important, we now question the notion that our fellow anthropologists constitute the only legitimate audience for our work. The remainder of this article explores the variety of ways that anthropologists are charting new paths in writing. As we have allowed ourselves to reimagine our writing practices, we have implicitly taken steps to embrace three new possible commitments: aiming our work toward new readerships; injecting new liveliness into the conventional writing genres of our discipline; and experimenting with writing in new genres for our discipline.

New Readerships

Central to the new ethnographic writing practices we explore in this article lies an increasingly urgent commitment to share our scholarly expertise with readers beyond our immediate colleagues (cf Gottlieb, 1997). Foundational to this trend lies a broader political commitment toward using our disciplinary knowledge to promote social justice.

Many scholars enter the field of anthropology in a quest to expose unequal opportunities – whether due to race, gender, sexual orientation, class, religion, health status, career choice, neighborhood, educational level, or other structural components of identity. After exposing such inequalities, many further hope that our writings might help expand opportunities for the disadvantaged – whether on a local, national, or global scale. Yet we have discovered that writing jargon-filled work destined for scholarly journals or book series that normally attract dozens or, at best, hundreds of readers limits severely whatever impact we may hope to have on whatever corner of the world (whether 'at home' or 'abroad') where we have chosen to focus our research. Frustrated by 'preaching to the choir,' we have begun expanding our 'bully pulpit' to new readerships (Sabloff, 2011).

The American Anthropological Association (AAA) – the discipline's flagship professional association in the United States, with increasing international membership (AAA, n.d.g), and the largest professional association in the world – now promotes this commitment in myriad ways. In recent years, the AAA has urged anthropologists to share our expertise with a broad public and increase our relevance, creating a host of new initiatives to offer realistic means for achieving this goal, including working with teachers from preschool to adult education classes (AAA, n.d.d), and encouraging anthropologists to read about and participate in public policy debates and other

work outside the academy (AAA n.d.a, n.d.c, n.d.e, n.d.f). Other smaller associations and entities likewise promote reaching out to broader publics in myriad ways, including sponsorship of book series aimed at general readers, and production of 'alternative' rankings of US doctoral programs in anthropology ranged according to the public effectiveness of each program's faculty, rather than the conventional academic criteria based on 'publications, status, and funds obtained' (Center for a Public Anthropology, n.d.). Outside the United States, many European, Latin American, and Asian scholars have engaged earlier and more comfortably with broader publics than have scholars than in the United States (e.g., Afonso, 2006; Barth, 2001; Eriksen, 2005; Hastrup et al., 2011; Hong Kong Anthropological Society, n.d.; Howell, 2010; Ribeiro, 2004, 2012; Rogers, 2001; Uquillas and Larreamendy, 2006).

Behind all these efforts – both collective and individual – lies not just the content of what we write, but the style of how we write it (Geertz, 1989). In effect, the above-mentioned initiatives, and others elsewhere like them, constitute a mandate for creating texts to engage readers beyond doctorate-bearing scholars. Insofar as writing is central to our scholarly practice, reaching new readers requires writing for those readers. How have anthropologists changed their writing style to make their message appealing to an 'educated public' that has an interest in an anthropological approach to understanding social life but lacks a doctorate in anthropology?

New Liveliness

An accessible text easily draws new readers to our research and our insights. And who are better than anthropologists to provide the comparative framework linking the global and the local, to humanize both near and distant others, to explain seemingly inexplicable behavior … in short, to propose creative ways to address the extraordinary opportunities and challenges of the twenty-first century?

Before the 1970s, the clamor for alternative writing remained marginal, with few feeling courageous enough to venture beyond the strict contours of the scholarly article or book; most saved the human stories-from-the-field for professional parties and the occasional classroom lecture. But the major cultural and political upheavals across Western societies during the 1970s saw echoes in major intellectual upheavals in the social sciences in general, and anthropology in particular – including upheavals in writing conventions. With the anti-Vietnam War movement, civil rights movement, women's rights movement, and associated 'sexual revolution,' growing interest in Hinduism and Buddhism, and development of alternative pop culture all combining to question long-accepted practices of both quotidian and political life, is it any wonder that anthropologists – observers of society par excellence – responded by questioning our own practices? The feminist revolution had an especially pronounced effect on our writing practices, although it took some time to be acknowledged (Gordon, 1988).

Unlike earlier generations, many junior scholars today aim to convey the sensory and emotional texture, and range of the human experience via a greater variety of textual styles. No longer are most of us content to write ethnographies that are, as Jeremy MacClancy wrote of an earlier generation of scholarship, "boring, and … virtually unreadable" (1996: p. 237). No longer do we consider it a badge of honor when a friend or relative expresses excitement on hearing about our research, only to confess confusion after dipping into our scholarly publications. Fewer among us feel content to exclusively write scholarly texts that turn off ordinary readers despite the inherent fascination of our topics.

Moreover, fewer academic publishers are willing to publish such works. It should be no surprise that good writing sells, and that means, in one way or another, good stories. Our species has aptly been dubbed homo narrans (Niles, 1999). Thoughtful observers of the human condition ranging from communication scholars (Fisher, 1987), literature professors (Gottschall, 2012), and psychiatrists (Coles, 1989) to urban planners (Eckstein and Throgmorton, 2003) and organizational sociologists (Gabriel, 2000) argue that storytelling is central both to our survival and to our nature. To sell more books, editors now advise authors via a simple homily: "More stories, less theory."

Many junior scholars are heeding the editors' advice. Increasingly, they express a desire either for a nonacademic career, or a career that combines academic research with nonacademic projects and publications. In both cases, their goal is to use their expertise to make a difference in arenas beyond the college classroom. Key to success in achieving this aim is honing the ability to write for readers beyond fellow scholars.

With this in mind, graduate students often seek out opportunities that will train them to hone their writing skills. In dissertation-writing workshops, students respond positively when urged to 'write from the heart' (Aronie, 1998) in turning their research data into theses. A long-ago piece of advice by the respected East Africanist, Ed Winter, may resonate today with doctoral students more forcefully than it did when it was first preferred over three decades ago (Gottlieb, n.d.).

"Don't look at your fieldnotes," Winter urged students just back from their field research. "Just write what you remember…You'll remember the good stuff. The important stuff. You can always check your notes later, make sure you got the details right."

If students were unlikely to take that antipositivist advice three decades ago, that is less the case today. Psychologists tell us that we humans live our lives at least as much in our feelings as we do in our thoughts (e.g., LeDoux, 1998). When neuroscientist Antonio Damasio (2010) studied people who could no longer feel emotions due to brain injuries, he discovered that these individuals had difficulty in making decisions about matters ranging from what to eat to where to live, despite being able to articulate rationally the advantages and disadvantages of their options. A recent theory of politics posits that a combination of emotion and reason accounts for the political stances we all take as citizens (Marcus et al., 2000). And within cultural anthropology, a growing number of scholars suggest that the same reality characterizes the methodological heart of our discipline – our fieldwork practice. That is, we bring our emotional biographies with us to the field, where they meet up with myriad emotional biographies of those in the communities we are studying (e.g., Ahmed, 2004; Hunt, 1989; Kleinman and Kopp, 1993; Kulick and

Willson, 1995; Wulff, 2007). However else we might portray it, the anthropological field encounter can be described as an emotional cauldron.

Anthropologists increasingly acknowledge that it does not make intellectual sense to divorce affective considerations from our analyses when they are a key component of the experiences that form the bedrock of our understanding. And given that narrative inherently offers the capacity to effectively evoke human emotion (Hogan, 2003; Hsu, 2008), anthropologists are increasingly discovering that narrative and related writing genres may actually offer more accurate – hence, more scientific – means for us as scholars to convey the full range of the human experience than do the conventional scholarly article and monograph. Echoing the wisdom of this growing interdisciplinary body of research, the best ethnographies, many now assert, engage our hearts and minds in equal doses.

In some cases, a traumatic incident in the field may so shake fieldworkers that they feel compelled to write about it in some deeply personal way. The usual distanced academic prose simply cannot capture enough of such events to satisfy either author or reader. For example, witnessing (or, in other cases, experiencing) sexual violence has produced riveting narratives by female anthropologists, who have drawn on their traumatic experiences to explore disturbing issues (di Leonardo, 1997; Winkler, 2002). Such emotionally engulfing events can be turned into compelling platforms from which to think frankly about the workings and bodily effects of power inequities, in ways that more conventional scholarly accounts cannot accomplish. Other returned fieldworkers may not experience such a single dramatic moment but nevertheless may have felt every aspect of newness so deeply, and listened so deeply to these feelings, that the cumulative result compels them to work through their emotions on the printed page. In either case, ethnographers who pay attention to the emotional core of their fieldwork experiences are likely to produce rich texts that leave an impact, and much to contemplate, well beyond the last page.

The mainstreaming of such texts signals important changes occurring in our discipline. The Victor Turner Prize in Ethnography (sponsored by the Society for Humanistic Anthropology, a unit of the AAA) has awarded $500 annually since 1990 for the most gripping ethnography published in the previous 2 years. As of 2013, 74 books have won a prize or honorable mention in the competition (Society for Humanistic Anthropology, n.d.). This corpus of engagingly written ethnographies suggests that many among the current generation of scholars seem inclined to write from the heart as much as from the mind.

For example, John Chernoff's two-volume biography of Hawa, a Ghanaian bar girl who is herself a storyteller's storyteller, by turns enchants, delights, and outrages the reader with a riveting account as hard to put down as the most gripping novel. The factual narrative describes an actual life that dignifies the particularities of a woman who has encountered and, in some ways triumphed over, hardships great and small. At the same time, the account evokes the challenges and life strategies of so many other working women whose lives are little short of heroic yet rarely generate written biographies (Chernoff, 2003, 2005). The rich ethnography conveyed by such accounts offers incomparable insights into the texture of human lives as lived in particular times and places.

The mandate to communicate our knowledge to a broad public becomes more realistic if we provide training for scholars in the writing skills that can convey the content of our research in an inviting manner. Thankfully, more scholars now teach such writing skills. Workshops on writing ethnographic fiction, poetry, and other 'alternative' genres are offered annually at the annual AAA conference (sponsored by the Society for Humanistic Anthropology), and they are increasingly oversubscribed. Moreover, semester-length courses on ethnographic writing are increasingly sprouting up in top-rated doctoral programs in anthropology across the United States and, more recently, in Europe. As such courses proliferate, programs increasingly train students to take the craft of writing seriously.

Writing for broad readerships lends itself to engaging with public issues and, in some doctoral programs, thus, goes hand in hand with a new emphasis on public anthropology (in the United States, as of this writing, in programs such as the University of Pennsylvania, the University of Oregon, Wayne State University, and the University of South Florida, among others). Such an emphasis characterizes doctoral programs beyond the United States as well. In Brazil, for example, "a minimum amount of advocacy on behalf of indigenous groups is expected…The lack of a dichotomy between 'getting involved' in the social and political struggles of study populations…and the academic ivory tower means that Brazilian anthropologists are university professors *and* political activists at the same time" (Chelekis, n.d.). Such a focus also characterizes a masters degree program in "applied anthropology and participatory development" in Australia, which is "designed for people working in government, non-government organisations, or as independent consultants … to manage development projects, including undertaking social assessment and advising on their cultural and social impacts" (Australia National University, n.d.). Such programs train students based on the premise that anthropological expertise in understanding the human condition merits promoting beyond the academy (on India, see Mahapatra, 2006).

In the United States, to increase the likelihood that the intellectual contributions of faculty in such programs receive proper recognition, the AAA has developed guidelines for tenure and promotion committees to use in evaluating efforts beyond scholarly writing (AAA, n.d.h, 2011). The AAA has also launched an online 'Writers Circle' that pairs scholars eager to write for a broader audience with seasoned scholars who have successfully done just that, and that makes such published writings available online as models (AAA, n.d.b).

New Genres

Anthropologists are staking out new territory by dipping into other genres, in a dual quest to reinvigorate traditional readerships and attract new ones. In the United States, Margaret Mead set an early precedent (e.g., Mead, 2005). Despite the many detractors who scorned the status she constructed for herself as public intellectual – especially through monthly columns in the popular women's magazine, *Redbook*, among other public venues – Mead established a powerful case for the relevance of an anthropological approach to

understanding the challenges of contemporary Western(ized) lives. Nowadays, after some especially vigorous critiques from colleagues, her intellectual legacy is being restored (di Leonardo, 2001; Lutkehaus, 2008; Shankman, 2009). Inspired by Mead's example, contemporary anthropologists are rediscovering that well-written editorial ('Op-Ed') pieces, letters to the editor, magazine articles, and radio pieces offer further tools to draw readers to our work. In the United States, a new Public Voices Fellowship Program (organized by an entity termed the Op-Ed Project) is providing training in just such genres: "This yearlong experience is designed to help professors…reach beyond the academic audience and share our ideas and experiences with a larger lay audience" (Colón-Ramos, 2013; *cf* The Op-Ed Project, n.d.). The Internet now offers unprecedented access to expanding readerships. Influential anthropologists such as Thomas Eriksen, Paul Stoller, and others now write blogs and guest posts for websites with enormous reach, while junior scholars experiment with new software programs to share findings in creative new formats (Balakian, 2011).

The classic genres of literature also offer appealing avenues for sharing our expertise. Early anthropologists such as Edward Sapir and Ruth Benedict published poetry that gained them readers beyond their scholarly colleagues (Caffrey, 1989; Flores, 1986; Handler, 1986). In recent years, cultural anthropologist Renato Rosaldo has won the widely respected American Book Award for his poetry (Rosaldo, 2003). Likewise, the early ethnographic novels of Adolf Bandalier (1971/1890), Ella Deloria (1988), Zora Neale Hurston (1937), and Laura Bohannan (Bowen, 1954) have many contemporary counterparts in novels by anthropologists Tobias Hecht (2006), Kirin Narayan (1994), Paul Stoller (1999, 2005), and others (cf. Laterz, 2007). Indeed, the Society for Humanistic Anthropology now offers annual awards in ethnographic poetry and short fiction, with winning pieces published in its peer-reviewed journal, *Anthropology and Humanism* (Society for Humanistic Anthropology, n.d.).

Beyond the two classic literary genres of poetry and fiction, anthropologists in recent years have begun exploring the possibility of sharing their expertise through other creative writing genres as well. Some anthropologists have written up their ethnography in the form of plays (e.g., Allen and Garner, 1997; Parada, 2008; Saldaña, 2005), mysteries (Nanda, 2009; Price and Price, 1995), biographies (Behar, 1993; Brown, 1991; Crapanzano, 1980; Shostak, 1981), and memoirs (e.g., Briggs, 1970; Cesara, 1982; Dumont, 1978; Gottlieb and Graham, 1993, 2012; Orlove, 2002; Stoller and Olkes, 1987; Tedlock, 1981).

Beyond these established writing genres, the world of writing offers a plethora of genres that we normally exclude from our scholarly purview, though we may engage them daily outside our academic lives. As participants in what we might term a veritable riot of literacy, scholars are intimately acquainted with a dazzling array of texts of every shape, size, and sort. Creatively engaging with such texts, some anthropologists venture beyond even the 'new' genres discussed above to create their own.

Back in the 1960s, not content with the conventional genres, the ever-maverick Gregory Bateson went so far as to invent a charming genre of writing that he called 'metalogues' (1972) in which he wrote partly fictionalized conversations about the most arcane of topics – conversations that he supposedly held with his partly imagined daughter (whose real-life model grew up to become anthropologist Mary Catherine Bateson). In recent years, a few experimental works by anthropologists and other social scientists have followed in Bateson's footsteps to draw creatively from other genres to convey ethnographic information in lively formats. Examples include a graphic exposition of the discipline (Galman, 2007), an imagined Socratic dialogue between European and African historical figures (Nzegwu, 2006), an extended photo essay (Behar, 2007), a political cookbook (Fair, 2008), imagined local child care guides *à la* Dr. Spock (DeLoache and Gottlieb, 2000), 'factional ethnography' (Sillitoe and Sillitoe, 2009), and an ethnographic, children's book-style 'flip book' (Taylor, 1998).

This growing corpus of 'alternative' texts suggests that increasing numbers of anthropologists are now 'finding their own jig to dance with genre.' As writing practices expand to attract multiple new audiences, the Internet beckons, with its extraordinary opportunities to find new readers. Anthropologists stand ideally poised to augment the discipline's reach in offering the insights into our lives as social creatures that are the hallmark of this unique academic field of inquiry.

Acknowledgments

Short portions of this article appear(ed) in different forms in:

- "The Anthropologist as Storyteller" in *The Anthropologist as Writer: Training, Practice, Genres*, ed. Helena Wulff (under review at the University of Chicago Press).
- "Ways of Writing Anthropology," talk presented at the Stockholm Roundtable on *"The Anthropologist as Writer: Training, Practice, Genres" (October 2009)*.
- "Dancing a Jig with Genre" in *Anthropology News* (April 2005).
- "The Perils of Popularizing Anthropology" in *Anthropology Today* 13 (1) (1997).

See also: Affect and Emotion, Anthropology of; Anthropology: Overview; Ethnography; Feminism in Anthropology; Fieldwork in Social and Cultural Anthropology; Inequality: Comparative Aspects; Internet and Social Media: Anthropological Aspects; Interpretation in Anthropology; Knowledge, Anthropology of; Reflexivity in Anthropology; Research Publication: Ethical Aspects; Senses, Anthropology of the.

Bibliography

Afonso, A.I., 2006. Practicing anthropology in Portugal. NAPA Bulletin 25 (1), 156–175.

Ahmed, S., 2004. The Cultural Politics of Emotion. Routledge, New York.

Allen, C.J., Garner, N.C., 1997. Condor Qatay: Anthropology in Performance. Waveland Press, Prospect Heights, IL.

American Anthropological Association, 2011. Public interest anthropology for academic promotion and tenure. AAA statement produced by the Committee on Practicing, Applied and Public Interest Anthropology. Adopted by the AAA Executive Board May 28, 2011. http://www.aaanet.org/resources/departments/upload/Final-T-P-Document-2011.pdf.

Aronie, N.S., 1998. Writing from the Heart: Tapping the Power of Your Inner Voice. Hyperion, New York.

Balakian, S., November 16–20, 2011. Mapping a lost friend: technology and tracing a transnational life. Paper presented at the 110th Annual Meeting of the American Anthropological Association (Montréal).

Bandelier, A., 1971 [1890]. The Delight Makers. Harcourt Brace Jovanovich, New York.

Barth, F., April 18, 2001. Envisioning a more public anthropology: an interview with Fredrik Barth [with Robert Borofsky]. Center for a Public Anthropology. http://www.publicanthropology.org/interview-with-fredrik-barth/.

Bateson, G., 1972. Metalogues. In: Bateson, G. (Ed.), Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology. University of Chicago Press, Chicago, pp. 3–60.

Behar, R., 1993. Translated Woman: Crossing the Border with Esperanza's Story. Beacon, New York.

Behar, R., 2007. An Island Called Home: Returning to Jewish Cuba. Rutgers University Press, New Brunswick, NJ.

Bowen, E.S., Bohannan, L., 1954. Return to Laughter. Harper & Brothers, New York.

Briggs, J., 1970. Never in Anger: Portrait of an Eskimo Family. Harvard University Press, Cambridge, MA.

Brown, K.M., 1991. Mama Lola: A Vodou Priestess in Brooklyn. University of California Press, Berkeley.

Caffrey, M.M., 1989. Ruth Benedict: Stranger in This Land. University of Texas Press, Austin.

Cesara, M., 1982. Reflections of a Woman Anthropologist: No Hiding Place. Academic Press, New York.

Chelekis, J. Brazilian Anthropology – A National Discipline. Indiana University/Dept. of Anthropology, n.d. http://www.indiana.edu/wwanthro/theory_pages/Brazilian.htm.

Chernoff, J., 2003. Hustling Is Not Stealing: Stories of an African Bar Girl. University of Chicago Press, Chicago.

Chernoff, J., 2005. Exchange Is Not Robbery: More Stories of an African Bar Girl. University of Chicago Press, Chicago.

Coles, R., 1989. The Call of Stories: Teaching and the Moral Imagination. Houghton Mifflin, New York.

Colón-Ramos, D., April 22, 2013. Found in Translation: A Professor Searches for a Public Voice. Chronicle of Higher Education. http://chronicle.com/article/Foundin-Translation-A/138651/?cid=at&utm_source=at&utm_medium=en.

Crapanzano, V., 1980. Tuhami: Portrait of a Moroccan. University of Chicago Press, Chicago.

Damasio, A., 2010. Self Comes to Mind: Constructing the Conscious Brain. Pantheon, New York.

DeLoache, J., Gottlieb, A. (Eds.), 2000. A World of Babies: Imagined Childcare Guides for Seven Societies. Cambridge University Press, New York.

Deloria, E.C., 1988. Water Lily. University of Nebraska Press, Lincoln.

di Leonardo, M., 1997 [1992]. White lies, black myths: rape, race, and the black "Underclass". In: Lancaster, R., di Leonardo, M. (Eds.), The Gender/Sexuality Reader: Culture, History, Political Economy. Routledge, New York, pp. 53–70.

di Leonardo, M., May 21, 2001. Margaret Mead vs. Tony Soprano: feminist anthropology fights for public voice in a new era. The Nation 272 (20), 29–36.

Dumont, J.-P., 1978. The Headman and I: Ambiguity and Ambivalence in the Fieldworking Experience. University of Texas Press, Austin.

Eckstein, B., Throgmorton, J.A. (Eds.), 2003. Story and Sustainability: Planning, Practice, and Possibility for American Cities. M.I.T. Press, Cambridge, MA.

Eriksen, T.H., 2005. Engaging Anthropology: The Case for a Public Presence. Berg, Oxford.

Fabian, J., 2011. Ethnography and memory. In: Melhuus, M., et al. (Eds.), Ethnographic Practice in the Present. Berghahn, New York, pp. 16–27.

Fair, C.C., 2008. Cuisines of the Axis of Evil and Other Irritating States: A Dinner Party Approach to International Relations. Globe Pequot Press, Guilford, CT.

Fisher, W.R., 1987. Human Communication as Narration: Toward a Philosophy of Reason, Value, and Action. University of South Carolina Press, Columbia.

Flores, T., 1986. The poetry of Edward Sapir. Dialectical Anthropology 11, 157–168.

Gabriel, Y., 2000. Storytelling in Organizations: Facts, Fictions, and Fantasies. Oxford University Press, New York.

Galman, S.C., 2007. Shane, the Lone Ethnographer: A Beginner's Guide to Ethnography. AltaMira Press, Lanham, MD.

Geertz, C., 1989. Works and Lives: The Anthropologist as Author. Stanford University Press, Stanford.

Gordon, D., 1988. Writing culture, writing feminism: the poetics and politics of experimental ethnography. Inscriptions 3 (4), 7–24.

Gottlieb, A., 1997. Guest editorial: the Perils of popularizing anthropology. Anthropology Today 13 (1), 1–2.

Gottlieb, A., The anthropologist as storyteller. In: Wulff, H. (Ed.), The Anthropologist as Writer: Training, Practice, Genres. Under Review at the University of Chicago Press, Chicago.

Gottlieb, A., Graham, P., 1994 [1993]. Parallel Worlds: An Anthropologist and a Writer Encounter Africa. University of Chicago Press, Chicago.

Gottlieb, A., Graham, P., 2012. Braided Worlds. University of Chicago Press, Chicago.

Gottschall, J., 2012. The Storytelling Animal: How Stories Make Us Human. Houghton Mifflin Harcourt, New York.

Handler, R., 1986. Vigorous male and aspiring female: poetry, personality, and culture in Edward Sapir and Ruth Benedict. In: Stocking, G. (Ed.), Malinowski, Rivers Benedict and Others: Essays on Culture and Personality. University of Wisconsin Press, Madison, pp. 127–155.

Hastrup, K., et al., 2011. Social and Cultural Anthropological Research in Norway – An Evaluation. Research Council of Norway, Oslo.

Hecht, T., 2006. After Life. Duke University Press, Durham, NC.

Hogan, P., 2003. The Mind and Its Stories: Narrative Universals and Human Emotion. Cambridge University Press, Cambridge.

Howell, S.L., 2010. Norwegian academic anthropologists in public spaces. Current Anthropology 51 (2), S269–S278.

Hsu, J., September 18, 2008. The secrets of storytelling: why we love a good yarn. Scientific American.

Hunt, J.C., 1989. Psychoanalytic Aspects of Fieldwork. Sage, Newbury Park, CA.

Hurston, Z.N., 1937. Their Eyes Were Watching God. J.B. Lippincott, New York.

Joelson, M.R., 2006. An International Antitrust Primer: A Guide to the Operation of United States, European Union, and Other Key Competition Laws in the Global Economy. Kluwer Law International, Alphen aan den Rijn, The Netherlands.

Kleinman, S., Kopp, M., 1993. Emotions and Fieldwork. Sage, Newbury Park, CA.

Kulick, D., Willson, M. (Eds.), 1995. Taboo: Sex, Identity and Erotic Subjectivity in Anthropological Fieldwork. Routledge, London.

Laterz, V., 2007. The ethnographic novel: another literary skeleton in the anthropological closet? Suomen Antropologi: Journal of the Finnish Anthropological Society 32 (2), 124–134.

LeDoux, J., 1998. The Emotional Brain: The Mysterious Underpinnings of Emotional Life. Simon & Schuster, New York.

Lutkehaus, N., 2008. Margaret Mead: The Making of an American Icon. Princeton University Press, Princeton.

MacClancy, J., 1996. Fieldwork styles: Bohannan, Barley, and Gardner. In: MacClancy, J., McDonaugh, C. (Eds.), Popularizing Anthropology. Routledge, London, pp. 225–244.

Mahapatra, L.K., 2006. Anthropology in policy and practice in India. NAPA Bulletin 25 (1), 52–69.

Marcus, G., Neuman, W.R., Mackuen, M., 2000. Affective Intelligence and Political Judgment. University of Chicago Press, Chicago.

Mead, Margaret, 2005. The World Ahead: An Anthropologist Anticipates the Future. Berghahn Books, Oxford.

Nanda, S., 2009. The Gift of a Bride: A Tale of Anthropology, Matrimony and Murder. AltaMira Press, Lanham, MD.

Narayan, K., 1994. Love, Stars and All That. Pocket Books/Simon and Schuster, New York.

Niles, J., 1999. Homo Narrans: The Poetics and Anthropology of Oral Literature. University of Pennsylvania Press, Philadelphia.

Nzegwu, N.U., 2006. The conclave: a dialogic search for equality. In: Nzegwu, N.U. (Ed.), Family Matters: Feminist Concepts in African Philosophy of Culture. SUNY Press, Albany, pp. 157–197.

Orlove, B., 2002. Lines in the Water: Nature and Culture at Lake Titicaca. University of California Press, Berkeley.

Parada, A., 2008. Playing Carmen: A one-act play in three scenes. [Written for ANTH 414–Ethnographic Writing, University of Illinois at Urbana-Champaign.] (Unpublished MS).

Price, R., Price, S., 1995. Enigma Variations. Harvard University Press, Cambridge, MA.

Ribeiro, G.L., 2004. Practicing anthropology in Brazil: a retrospective look at two time periods. Practicing Anthropology 26 (3), 6–10.

Ribeiro, G.L., 2012. From local to global ethnographic scenarios. In: Gottlieb, A. (Ed.), The Restless Anthropologist: New Fieldsites, New Visions. University of Chicago Press, Chicago, pp. 25–34.

Rogers, S.C., 2001. Anthropology in France. Annual Review of Anthropology 3, 481–504.

Rosaldo, R., 2003. Prayer to Spider Woman. Gobierno del Estado de Coahuila, Instituto Coahuilense de Cultura, , San Ángel. Del. Álvaro Obregón, Mexico.

Sabloff, J.A., 2011. Where have you gone, Margaret Mead? Anthropology and public intellectuals. American Anthropologist 113 (3), 408–416.

Saldaña, J. (Ed.), 2005. Ethnodrama: An Anthology of Reality Theatre. AltaMira Press, Lanham, MD.

Shankman, P., 2009. The Trashing of Margaret Mead: Anatomy of an Anthropological Controversy. University of Wisconsin Press, Madison.

Shostak, M., 1981. Nisa: The Life and Words of a !Kung Woman. Harvard University Press, Cambridge, MA.

Sillitoe, P., Sillitoe, J., 2009. Grass-Clearing Man: A Factional Ethnography of Life in the New Guinea Highlands. Waveland Press, Long Grove, IL.
Stoller, P., 1999. Jaguar: A Story of Africans in America. University of Chicago Press, Chicago.
Stoller, P., 2005. Gallery Bundu: A Story about an African Past. University of Chicago Press, Chicago.
Stoller, P., Olkes, C., 1987. In Sorcery's Shadow: A Memoir of Apprenticeship among the Songhay of Niger. University of Chicago Press, Chicago.
Taylor, J., 1998. Paper Tangos. Duke University Press, Durham.
Tedlock, B., 1981. The Beautiful and the Dangerous: Encounters with the Zuni Indians. Viking, New York.
Uquillas, J.E., Larreamendy, P., 2006. Applied anthropology in Ecuador: development, practice, and discourse. NAPA Bulletin 25 (1), 14–34.
Winkler, C., 2002. One Night: Realities of Rape. Altamira Press/Rowman & Littlefield, Blue Ridge Summit, PA.
Wulff, H. (Ed.), 2007. The Emotions: A Cultural Reader. Berg, London.
Wulff, H., Handelman, D., Berghahn, M., Dominguez, V., Moeran, B., 2009. Debate section: what is happening to the anthropological monograph? Live debate at the EASA conference in Ljubljana 2008. Social Anthropology 17 (2), 217–234.

Relevant Websites

http://www.aaanet.org/gvt/index.htm – American Anthropological Association, n.d.a. AAA Government Relations and Public Policy Efforts.
http://www.aaanet.org/resources/AAA-Writers-Circle.cfm – American Anthropological Association, n.d.b. AAA Writers circle.
http://www.aaanet.org/publications/ameranthro.cfm – American Anthropological Association, n.d.c. American Anthropologist Launches "Public Anthropology Reviews" – Request for Submission of Review Materials.
http://www.aaanet.org/committees/commissions/aec/whatis.htm – American Anthropological Association, n.d.d. Anthropology Education Committee (formerly the Anthropology Education Commission).
http://www.aaanet.org/cmtes/ppc/index.cfm – American Anthropological Association, n.d.e. Committee on Public Policy.
http://www.aaanet.org/cmtes/copapia – American Anthropological Association, n.d.f. Committee on Practicing, Applied and Public Interest Anthropology (CoPAPIA).
http://www.aaanet.org/cmtes/cwa/ – American Anthropological Association, n.d.g. Committee on World Anthropologies.
http://www.aaanet.org/resources/departments/upload/Final-T-P-Document-2011.pdf – American Anthropological Association, n.d.h. Guidelines for evaluating scholarship in the realm of practicing, applied, and public interest anthropology for academic promotion and tenure. AAA statement produced by the Committee on Practicing, Applied and Public Interest Anthropology, adopted by the AAA Executive Board, May 28, 2011.
https://studyat.anu.edu.au/programs/7101XMAAPD;overview.html – Australia National University, College of Arts and Social Sciences, n.d.
http://www.publicanthropology.org/ – Center for a Public Anthropology, n.d.
http://blogs.scientificamerican.com/anthropology-in-practice/ – Costa, K. Anthropology in Practice: The Human Experience. Scientific American, n.d.
http://folk.uio.no/geirthe/ – Eriksen, T. H., Engaging with the World: Eriksen's Site.
http://www.wcaanet.org/members/members_descr.shtml – Hong Kong Anthropological Society, n.d.
http://www.theopedproject.org/ – The OpEd Project. Echoing Green, n.d.
http://www.aaanet.org/sections/sha/sha-prize-winners/ – Society for Humanistic Anthropology, n.d.
http://www.huffingtonpost.com/paul-stoller/ – Stoller, P., The Huffington Post.
http://www.wcaanet.org/ – World Council of Anthropological Associations.

Anthropology and History

James D Faubion, Rice University, Houston, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Anthropology engages history not as one but instead as many things: (1) sociocultural change or diachrony; (2) a domain of events and objects that make manifest systems of signification, purpose, and value; (3) a domain of variable modalities of the experience and consciousness of being in time; and (4) a domain of practices, methods, and theories devoted to the recording and the analysis of temporal phenomena. It emerged, and continues to serve, as that branch of 'natural history' which investigates the psychophysical origins and diversification of the human race. As 'ethnohistory,' it investigates the documents of the pasts of native or 'first' peoples, paying special attention to the dynamics and consequences of colonization. Emile Durkheim's *Elementary Forms of the Religious Life* (1995/1912) opens the arena of an 'anthropology of history' with its argument for the social causation of the experience and conceptualization of time, but anthropologists remain divided over what the anthropology of history is or should be. Their disagreements are instructive because they recapitulate a much larger and more enduring controversy over whether anthropological knowledge is a mode of historical or instead a mode of scientific knowledge. The controversy is probably also irresolute, at least until either anthropology or history comes to an end.

For anthropology, history is not one but many things. First, it is the past, especially the past that survives in archives and other written or oral records; 'prehistory' is its more remote counterpart. Second, it is change, diachronic as opposed to synchronic process. Third, history is a domain of events and artifacts that make manifest systems of signification, purpose, and value, the domain of human action. Fourth, it is the domain of all the diverse modes of the human experience and consciousness of being in time. Finally, it is a domain encompassing all those practices, methods, symbologies, and theories that human beings – professional academic historians among them – have applied to the collection, recollection, and comprehension of the past, the present, and the relations between the two.

Anthropology and Natural History

At the end of the sixteenth century, anthropology emerged in Europe not in contrast to history but rather within it. Thereafter, and for some two and a half centuries forward, it would broadly be understood as that branch of 'natural history' which investigated the psychophysical origins and diversification of the human race, or races, as was very often the case. Demurring throughout the period to the theological calculus of the creation of Adam, it confined itself to treating developments presumed to have transpired over only a few millennia. In 1858, miners at England's Brixham cave unearthed tools and other human remains that stratigraphers could prove to be at least 70 000 years old. Theological authority suffered a blow from which it would not recover; anthropological time suddenly acquired much greater archaeological depth. Meanwhile, a growing scholarly coalition was coming together to support the doctrine that 'culture' was that common human possession which made manifest the basic psychic unity of all the putative races of mankind. In 1854, James Pritchard would accordingly launch 'ethnology,' and send the racialists off on their separate ways (Stocking, 1987). 'Cultural' and 'physical' anthropologists would never again keep their earlier company. Yet, even through the turn of the twentieth century, natural history remained the largely uncontested source of the methods and aims of both. Questions of origin and development continued to have pride of place. The 'savage' or the 'primitive' became all the more entrenched as a disciplinary preserve, among other things as the rudimentary pole of any number of ambitious reconstructions of the probable steps or stages that had marked the human passage to 'civilization' or 'modernity.' E. B. Tylor's *Primitive Culture* (1871), Henry Lewis Morgan's *Ancient Society* (1877), and the several volumes of James Frazer's *Golden Bough* (1890) are classic examples of the genre.

Such far-reaching treatises would strike the more meticulous of the subsequent generation of their readers as undisciplined, even whimsical. By the 1920s, and despite all their other differences, Franz Boas, Bronislaw Malinowski, and A. R. Radcliffe-Brown were inaugurating a turn away from 'speculative history' in favor of meticulous observational attention to the 'ethnographic present.' Not even these sober empiricists were, however, opposed in principle to reconstructive or evolutionary analysis. With due regard for rigor, the more adventurous of their colleagues would continue to pursue it; and virtually every subfield of anthropology has contributed to it ever since. Physical anthropology is now a 'genetic' science in both the larger and the stricter sense of the term. Morgan's project survives most explicitly in the United States, from Leslie White to Marvin Harris, as 'cultural materialism' (see Harris, 1968), but also endures implicitly in the technological determinism that informs Jack Goody's argument in *The Domestication of the Savage Mind* (1977). Tylor and Frazer are the precursors not simply of Claude Lévi-Strauss's structuralism (see infra) but also of the burgeoning interdisciplinary and neo-Darwinist vocation known as 'evolutionary psychology.' Boas himself is among the bridges between an older 'comparative philology' and efforts to trace the family tree of all the world's languages (see Kroeber, 1935).

Anthropology and Ethnohistory

A more modest anthropological tradition of diachronic research has borrowed its methods less from natural history

than from empirical historiography. It has a partial foreshadowing in the particularistic study of the drift and dissemination of traits and artifacts from the centers to the peripheries of cultural production with which 'diffusionists' in England, Germany, and the United States were occupied between the 1890s and the 1930s. It has its more definitive commencement in the immediate aftermath of the Second World War. Its most familiar designation is still 'ethnohistory,' however misleading the suggestion of affinities or parallels with 'ethnoscience' or 'ethnomethodology' may be. In any event, the signature task of ethnohistory has always been the investigation and documentation of the pasts of those native or 'first' peoples whom anthropologists had until rather recently proprietarily or conventionally claimed as 'their own.' In the United States, its more concrete initial impetus came with the 1946 ratification of the Indian Claims Act, which soon led to anthropologists serving as expert witnesses – sometimes for the plaintiffs, sometimes for the defense – in the readjudication of the treaties of the pioneer era. In the United States and elsewhere, it would have a catalyst in the granting of public access to the administrative archives of the pioneers, the missionaries, and the bureaucrats of European colonization. Hence, its characteristic focus: the dynamics of contact and conflict between the subaltern and their would-be overlords: pioneering, missionary, colonizing, or enslaving (Spicer, 1975; Cohn, 1968, 1980; Leacock and Lurie, 1971; Rosaldo, 1980; Dirks, 1988, 2001, 2006; Kirch and Sahlins, 1992, 1994; Vansina, 2004, 2010; Hanks, 2009). The most wide-ranging examples remain Eric Wolf's *Europe and the People Without History* (1997) and Sidney Mintz's *Sweetness and Power* (1986).

Narrowly delimited, ethnohistory remains the specialist's craft. Since the 1970s, however, its methods and themes have met with an ever-widening embrace, and if 'historical ethnography' and 'historical anthropology' are not yet synonymous with standard disciplinary practice, they are certainly of a piece with it. The anthropological gaze is now less often 'from afar' than it is longitudinal. Such a vantage has been pivotal in the renovation of political economy. It has also brought fresh and stimulating perspectives to the address of kinship, race, national and personal identity, and gender (e.g., Stoler, 1995). Its deployment and impact may or may not be indications of greater disciplinary enlightenment, but they are by no means indications of passing intellectual fashion alone. The lengthening of the anthropological gaze has rather gone hand in hand with the ascendance of a postcolonial order in which social and cultural boundaries have become increasingly permeable and structures increasingly indistinguishable from processes. It has gone hand in hand as well with the ascendance of the postcolonial demand that anthropology offer a reckoning, not simply of its relation to the colonial past but also of the status of the knowledge that it claims to produce (Hymes, 1972; Asad, 1973; Clifford and Marcus, 1986).

Anthropology and Hermeneutics

In our postcolonial order, anthropology itself needs interpreting; so, too, do many other sociocultural phenomena. Anthropology's current troupe of interpreters espouses, however, an even stronger postulate: that sociocultural phenomena, as historical phenomena, permit only of interpretation; that they permit of contextual understanding, but not of general explanation. 'Interpretive anthropology' thus stands starkly at odds with the loftier versions of the natural history of culture, but further with any empirical historiography that has the inductive abstraction of general types or causal relations as its end. It now comes in many versions of its own. The most venerable of them commences with the Boasians. In 1935, Alfred Kroeber would accordingly remark that the ethnographies that Ruth Benedict and his other colleagues were busy producing were 'historical' in type. True enough, their temporality was synchronic, not diachronic. Their mode, however, was particularistic; their method contextualist; their analyses rarely if ever causal; and their model consequently that of what Wilhelm Dilthey, pursuing a 'critique of historical reason,' had defined as the *Geisteswissenschaften* – 'sciences of spirit,' literally, but better glossed as 'sciences of meaning' or simply 'hermeneutics' (Dilthey, 2002). In fact, Kroeber was quite correct; Benedict, Margaret Mead, and the other cultural anthropologists of Boas's circle were indeed hermeneuticians. They were hence establishing the methodological legacy to which Clifford Geertz is the most celebrated heir (Geertz, 1973). It should be added that the long-standing, if methodologically variable, tradition of life-historical or 'ethnobiographical' research, in which a limited diachronic inquiry serves as the means to a synchronically contextualizing end, is also Boasian in its widest parameters (see, e.g., Mintz, 1986; Crapanzano, 1980; Shostak, 1981; Faubion, 2011).

The more proximate wellspring of contemporary interpretation flows out of the tumult of the late 1960s. Elaborating in 1972 upon the call for a 'critical and reflexive anthropology,' Bob Scholte was among its earliest manifestologists, though he borrowed many of his own philosophical and methodological tenets from Johannes Fabian's slightly earlier exposition of Hans-Georg Gadamer's revision of Dilthey's thought (Scholte, 1972; Fabian, 1971, see also 1983). The channels thus opened have remained critical and reflexive, if perhaps not always so politically committed as Scholte might have hoped. Gadamer's heremeneutics has, moreover, served as only one of many subsequent grounds on which to establish interpretive license (see Rabinow and Sullivan, 1988). Walter Benjamin (see Taussig, 1993) and Giambattista Vico (see Herzfeld, 1987) have won admirers of their own. Geertz's inspirations were diverse: Suzanne Langer; Gilbert Ryle; Ludwig Wittgenstein; above all, Max Weber.

Such a well-populated census might suggest that, as much now as at its beginnings, anthropology belongs to history (as a discipline) no less than in it (as contingent process). Yet, even many of those who label themselves 'historical anthropologists' or 'interpreters' of one or another stripe would object to such a subsumption. No doubt, professional territorialism plays a part in their resistance. Often very palpable differences of professional sensibility play another part. Matters more strictly intellectual, however, play a part of their own, and their stakes are no more evident than within the anthropology of history itself.

The Anthropology of History

Durkheim's *Elementary Forms of the Religious Life* (1995) opens the arena of the anthropology of history with its argument for the social causation of the experience and conceptualization of time. Maurice Halbwachs' *Collective Memory* (1980) expands it and, after many decades of neglect, has come to be among the foundational texts of a recent surge of ethnographic and comparative inquiry into the techniques, the media, and the politics of remembering and forgetting (see, e.g., Boyarin, 1994; Shryock, 1997). Paul Connerton is the most systematic of Halbwachs' followers, and moves beyond Halbwach's own emphasis on the spatial objectification of collective memory to a consideration of the embodiment of memory in and as habit (1989).

The most imposing shadow hanging over the broader anthropology of history is, however, Lévi-Strauss's *The Savage Mind* (1966). A treatise devoted to the analysis of the analogical – and ahistorical – matrices of mythic and totemic thought, *The Savage Mind*, concludes with an extended polemic against the 'historical, structural anthropology' that Jean-Paul Sartre had advocated in the introduction to his *Critique of Dialectical Reason* (2004). Sartre had no time for totemists. His anthropology instead rested with the charting, and the heightening, of 'historical consciousness.' Against it, Lévi-Strauss had two general rejoinders. The first was that 'historical consciousness' was the expression not of dawning wisdom but instead of a collective devotion to 'development,' and its absence was not of the expression of error but instead of a collective devotion to homeostasis. Some societies ran historically 'hot'; others ran 'cold.' All were equally human. The second was that history – as narrative, as diachronic interpretation – was always 'history-for,' always biased. Stripped of its bias, it amounted to nothing more than the methodical application of temporal scales of measure to the flow of human and nonhuman events alike. If not that, then it amounted simply to the preliminary 'cataloging' with which any 'quest for intelligibility' – the anthropological quest included – had to begin. But it could be a beginning only: 'as we say of certain careers, history may lead to anything, provided you get out of it' (Lévi-Strauss, 1966).

Lévi-Strauss came to grant that his division between 'hot' and 'cold' societies was in need of much refinement; that instead of functioning as a strictly binary opposition, the division could better be conceived as defining a continuum. It is good that he did so, since a great deal of ethnographic evidence had already pointed and continues to point to just that. The classic cases are those of millenarian movements and so-called cargo cults, which have been widely documented in Melanesia and Polynesia (Worsley, 1957; Burridge, 1960, 1969), Amazonia (Hill, 1988), Africa (Peires, 1989), Native North America (Mooney, 1896), the Caribbean (Littlewood, 1992) and elsewhere, many but not all of them emerging in the aftermath of one or another colonial encounter and many but not all of them influenced by the New Testament Book of Revelation (see also Lindstrom, 1993; Faubion, 2001). Historically grounded studies of Christianization offer more complex versions of the same theme (Comaroff and Comaroff, 1991, 1997; Robbins, 2004). Michael Lambek has explored a decidedly non-millenarian but far from chilly sense of the relationship among the past, present, and future in Madagascar (2003). Penelope Papailias (2005) has delved into the writings of modern Greek historians to discover senses of the same relationship that are very far from the boiling point. The list could continue at length.

For all this, Lévi-Strauss never retreated from his division between historical and properly anthropological knowledge. On the contrary, from the outset of his career, he consistently counted history among those disciplines limited to the merely statistical representation and analysis of their objects. His anthropology is for its part a model-theoretic discipline, an axiomatic and deductive science. Its object remains what it was for Tylor: the psyche. Its quest, however, culminates not in the hypothetical reconstruction of the path toward enlightened modernity but instead in the formal exegesis of the universal 'grammar,' the structural and structuring properties, of the mind itself. 'Structuralism,' thus construed, is far less influential than it was in the 1960s, but by no means bereft. In 'cognitive anthropology,' it has its most secure contemporary home; and there, history (as intellectual or epistemological paragon) continues to meet with a cool reception.

For Lévi-Strauss as for other philosophical and scientific rationalists, there is an insuperable gap between the 'facts' and their theoretical intelligibility. For positivists and empiricists, the relation between facts and theories is putatively more seamless. Unabashedly positivist anthropologists are a rather rare breed at present, at least in the sociocultural field, though many cultural materialists and evolutionary psychologists might quietly reckon themselves as such. The positivist anthropology of historical consciousness or the 'historical sensibility' is rare indeed, but has a singularly unabashed spokesperson in Donald Brown. In *History, Hierarchy, and Human Nature* (1988), Brown offers a cross-cultural survey of those traits – from divination to record-keeping – most suggestive of a preoccupation with the meaning and significance of events. Among literate peoples, he discerns a relatively stable correlation: between the presence and prominence of such a preoccupation and the absence of caste or other fixed hierarchies. He concludes that history (as sensibility, as worldview, and as mode of inquiry) takes its most regular nourishment from an ideology of social mobility. Whether or not correct, the conclusion is compatible with Lévi-Strauss's own considered judgments. Here, too, the anthropologist takes history as an analytical object.

Marshall Sahlins offers an alternative, which also takes history as its analytical object, but further as its analytical mode. It preserves the structuralist principle that systems of signification are never mere derivatives of their social or material environment. Yet it casts them less as revelations of the grammar of the psyche than as interfaces or differentials between the past and the future of given social practice. Their effect is threefold. First, they determine the internal historical 'temperature' of practice, which is relatively colder when governed by prescriptions, warmer when not. Second, they vary in their capacity to accommodate potentially anomalous or disruptive events. They are more or less historically 'sensitive,' and the greater their sensitivity, the more the continuity of practice is itself at risk. Finally, they influence the symbolic 'weight' or import of actors and their acts. Where some men are

kings, and their authority unchallenged, the historian is right to monitor them with especial care; where democracy holds sway, he or she would do better to monitor status groups or classes or parties. Hence, the best historian should be a good anthropologist, and the best anthropologist always prepared to be a good historian as well (Sahlins, 1985).

Sahlins's standard of goodness is still the standard of objective accuracy. It thus stands in partial contrast to the standard that has at least implicitly guided interpretive anthropology since the Boasian 'golden age.' Though subject to diverse formulations – some more vividly critical than others – the latter standard is practical or pragmatic, a matter of consequences. Perhaps for a majority of interpretive anthropologists, it has been nothing short of 'liberation,' whether from psychosexual repression, as for Benedict and Mead, or from political domination and economic exploitation, as for Scholte. Many of its prominent inflections remain reformist, though of more qualified scope. For the Geertzian, however, the pragmatic proof of an interpretation lies in its facilitating conversation, its translational efficacy. For Geertz and others, it lies in a broadening or enrichment of our imagination of the ways of being human. For a few others still, it lies in therapeutic release – perhaps from prejudice, perhaps from alienated isolation. Interpretation must in every case be factually informed; but at its best, is always also 'informative.' Its analytical mode is historical; but however much Lévi-Strauss or Sahlins might disapprove, it is always also 'history-for' and 'anthropology-for' alike. So, too, the interpretive history of anthropology and the interpretive anthropology of history must occupy the same epistemological plane. Here, historical and anthropological knowledge are of precisely the same kind.

Consistent with such a position is Matti Bunzl's proposal for the development of a 'neo-Boasian' anthropology, a hybrid of the Boasian tradition of particularistic contextualization and Michel Foucault's genealogical 'history of the present' (Bunzl, 2003; Foucault, 1998). In Foucault's first sustained articulation of it, genealogy has as its exemplar Friedrich Nietzsche's *Genealogy of Morals* (1956). Following in Nietzsche's footsteps, it is grounded in the marshaling of evidence of the partiality of the presumptively universal. It exposes the contingency of the presumptively necessary. It is most definitely a history-for. It does not, however, offer a design for historical inquiry in close accord with Dilthey's or other standard historical precedents. It is 'meticulous,' but its aim is not merely nor even primarily to elucidate the contextual nexus that unites a historical ensemble into a meaningful complex. It is not hermeneutical in this sense. Its aim is instead to bring to light that one or another aspect of the present need not be taken as much for granted as at first sight it might seem to need to be. Bunzl rightly notes that the rhetorical force of Foucault's 'genealogy of the present' notwithstanding, it does not as a matter of its own methodological constraints arrive at the present as such. Bunzl's proposal is that a synthesis of Foucauldean genealogy and Boasian hermeneutics could yield a genealogical hermeneutics or hermeneutical genealogy not merely of the present but also in it.

Bunzl points to Paul Rabinow's *French Modern* (1989) as the first anthropological installment of a history of (if not yet quite in) the present. In the work leading to the publication of *Marking Time* (2007), however, and in the work since (Rabinow and Stavrianakis, 2013), Rabinow has been developing his own diagnostics not merely of the present but also of the near past that precedes it and the near future that might follow it. Drawing from Foucault but also from John Dewey and Gilles Deleuze, he offers an 'anthropology of the contemporary' that would track 'ratio(s) of modernity, moving through the recent past and near future in a (non-linear) space' and serving as 'gauges' in light of which one or another aspect of the modern can be registered as 'becoming historical.' Once again, a history-for but also as much an anthropology-for, it borrows some portion of its critical framework from Foucault's late formulation of a 'critical ontology of ourselves,' an exercise that in its successes would extract 'from the contingency that has made us what we are, the possibility of no longer being, doing, or thinking as we are, do, or think' (1997). If the specific temporality of Rabinow's anthropology of the contemporary is of modest scope, it is in accord with the modesty that Ulrich Beck et al. (1994) have implicitly endorsed in their account of the 'risk society' in which we now all live and that Niklas Luhmann (1998) has similarly endorsed in his account of the increasingly global 'ecology of ignorance' with which we now all have to cope. Both accounts underscore that relations between the recent past and the near future are about the only relations in whose understanding we might dare to have some confidence. Contemporaneity is thus our lot. It is the distinctive order of the day. An anthropology of the ratios between the recent past and the near future could accordingly be thought to afford an analytical lens through which we might examine the environment – epistemic, sociocultural, and political – that is in fact all of our own.

In any event, it is clear that history is not simply a thing of many anthropological refractions. It is also a thing of plural and incompatible anthropological estimations. These in turn are among the most telling indices of the most persistent and most persuasive of plural and incompatible visions of the basic enterprise of anthropology itself. One can hence condemn the discipline for its intellectual indecisiveness or incoherence. One can applaud it for its perspectival diversity. One can in any event note that its many byways have a common point of departure in the question of whether human nature itself is transhistorically fixed, or instead historically variable. This is anthropology's first question, and if the past is any indication of the future, it is likely to remain so, at least until either Man or history comes to an end.

Bibliography

Asad, T. (Ed.), 1973. Anthropology and the Colonial Encounter. Humanities Press, Atlantic Highlands, NJ.

Beck, U., Giddens, A., Lash, S., 1994. Reflexive Modernization: Politics, Tradition and Aesthetics in the Modern Social Order. Stanford University Press, Stanford, CA.

Boyarin, J. (Ed.), 1994. Remapping Memory: The Politics of Timespace. University of Minnesota Press, Minneapolis, MN.

Brown, D., 1988. History, Hierarchy, and Human Nature. University of Arizona Press, Tucson, AZ.

Bunzl, M., 2003. Notes toward a neo-Boasian anthropology. American Anthropologist 106 (3), 435–442.

Burridge, K., 1960. Mambu: A Melanesian Millennium. Methuen, London.

Burridge, K., 1969. New Heaven, New Earth: A Study of Millenarian Activities. Basil Blackwell, Oxford.
Clifford, J., Marcus, G. (Eds.), 1986. Writing Culture: On the Poetics and Politics of Ethnography. University of California Press, Berkeley, CA.
Cohn, B., 1968. Ethnohistory. In: Sills, D. (Ed.), International Encyclopedia of the Social Sciences. Macmillan, New York, pp. 440–448.
Cohn, B., 1980. History and anthropology: the state of play. Comparative Studies in Society and History 22 (12), 198–221.
Comaroff, Jean, Comaroff, John, 1991. Of Revelation and Revolution, Volume 1: Christianity, Colonialism, and Consciousness in South Africa. University of Chicago Press, Chicago.
Comaroff, John, Comaroff, Jean, 1997. Of Revelation and Revolution, Volume 2: The Dialectics of Modernity on the South African Frontier. University of Chicago Press, Chicago.
Connerton, P., 1989. How Societies Remember. Cambridge University Press, Cambridge, UK.
Crapanzano, V., 1980. Tuhami: Portrait of a Moroccan. University of Chicago Press, Chicago.
Dilthey, W., 2002/1910. In: Makkreel, R., Rodi, F. (Eds.), The Formation of the Historical World in the Human Sciences. Princeton University Press, Princeton.
Dirks, N., 1988. The Hollow Crown: Ethnohistory of an Indian Kingdom. Cambridge University Press, Cambridge, UK.
Dirks, N., 2001. Castes of Mind: Colonialism and the Making of Modern India. Princeton University Press, Princeton, NJ.
Dirks, N., 2006. The Scandal of Empire: India and the Creation of Imperial Britain. Harvard University Press, Cambridge, MA.
Durkheim, E., 1995/1912. The Elementary Forms of the Religious Life (K.E. Fields, Trans.). Free Press, New York.
Fabian, J., 1971. Language, history, and anthropology. Journal of the Philosophy of the Social Sciences 1 (1), 19–47.
Fabian, J., 1983. Time and the Other: How Anthropology Makes Its Object. Columbia University Press, New York.
Faubion, 2001. The Shadows and Lights of Waco: Millennialism Today. Princeton University Press, Princeton, NJ.
Faubion, J., 2011. An Anthropology of Ethics. Cambridge University Press, Cambridge, UK.
Foucault, M., 1998. Nietzsche, genealogy, history. In: Faubion, J. (Ed.), Essential Works of Michel Foucault. Aesthetics, Method, Epistemology, vol. 2. The New Press, New York, pp. 369–392.
Foucault, M., 1997. What is enlightenment? In: Rabinow, P. (Ed.), Essential Works of Michel Foucault. Ethics: Subjectivity and Truth, vol. 2. The New Press, New York, pp. 309–320.
Frazer, J.G., 1890. The Golden Bough: A Study in Magic and Religion. Macmillan, London.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Goody, J., 1977. The Domestication of the Savage Mind. Cambridge University Press, Cambridge, UK.
Halbwachs, M., 1980/1950. The Collective Memory (F.J. Ditter Jr. and V.Y. Ditter, Trans.). Harper Colophon Books, New York.
Hanks, W., 2009. Converting Words: Maya in the Age of the Cross. University of California Press, Berkeley, CA.
Harris, M., 1968. The Rise of Anthropological Theory: A History of Theories of Culture. Crowell, New York.
Herzfeld, M., 1987. Anthropology through the Looking-Glass: Critical Ethnography in the Margins of Europe. Cambridge University Press, New York.
Hill, J. (Ed.), 1988. Rethinking History and Myth: Indigenous South American Perspectives on the Past. University of Illinois Press, Chicago, IL.
Hymes, D. (Ed.), 1972. Rethinking Anthropology. University of California Press, Berkeley, CA.
Kirch, P., Sahlins, M., 1992. Anahulu: The Historical Anthropology of the Kingdom of Hawaii. In Historical Ethnography, vol. 1. University of Chicago Press, Chicago.
Kirch, P., Sahlins, M., 1994. Anahulu: The Historical Anthropology of the Kingdom of Hawaii. In: Historical Archaeology, vol. 2. The University of Chicago Press, Chicago.
Kroeber, A., 1935. History and science in anthropology. American Anthropologist 37 (4), 539–569.
Lambek, M., 2003. The Weight of the Past: Living with the Past in Mahajanga, Madagascar. Palgrave Macmillan, New York.
Leacock, E., Lurie, N.O. (Eds.), 1971. North American Indians in Historical Perspective. Random House, New York.
Lévi-Strauss, C., 1966/1962. The Savage Mind. (La Pensée sauvage). University of Chicago Press, Chicago.
Lindstrom, L., 1993. Cargo Cult: Strange Stories of Desire from Melanesia and Beyond. University of Hawaii Press, Honolulu.
Littlewood, R., 1992. Pathology and Identity: The Work of Mother Earth in Trinidad. Cambridge University Press, New York.
Luhmann, N., 1998. Observations on Modernity (W. Whobrey, Trans.). Stanford University Press, Stanford, CA.
Mintz, S., 1986. Sweetness and Power: The Place of Sugar in Modern History. Viking Press, New York.
Mooney, J., 1896. The Ghost-dance Religion and the Sioux Outbreak of 1890. University of Chicago Press, Chicago.
Morgan, L.H., 1877. Ancient Society. Holt, New York.
Nietzsche, F., 1956/1887. The Birth of Tragedy and the Genealogy of Morals (F. Francis Golffing, Trans.). Doubleday Anchor Books, Garden City, NY.
Papailias, P., 2005. Genres of Recollection: Archival Poetics and Modern Greece. Palgrave Macmillan, New York.
Peires, J., 1989. The Dead Will Arise: Nongqawuse and the Great Xhosa Cattle Killing. Indiana University Press, Bloomington, IL.
Rabinow, P., 1989. French Modern: Norms and Forms of the Social Environment. MIT Press, Cambridge, MA.
Rabinow, P., 2007. Marking Time: On the Anthropology of the Contemporary. Princeton University Press, Princeton, NJ.
Rabinow, P., Stavrianakis, A., 2013. The Demands of the Day: On the Logic of Anthropological Inquiry. University of Chicago Press, Chicago.
Rabinow, P., Sullivan, W., 1988. Interpretive Social Science. University of California Press, Berkeley, CA.
Robbins, J., 2004. Becoming Sinners: Christianity and Moral Torment in a Papua New Guinea Society. University of California Press, Berkeley, CA.
Rosaldo, R., 1980. Ilongot Headhunting, 1883–1974: A Study in Society and History. Stanford University Press, Stanford, CA.
Sahlins, M., 1985. Islands of History. University of Chicago Press, Chicago.
Sartre, J.-P., 2004/1960. The Critique of Dialectical Reason, Volume 1: Theory of Practical Ensembles (J. Rée and J. Sheridan-Smith, Trans.). Verso, London.
Scholte, B., 1972. Toward a critical and reflexive anthropology. In: Hymes, D. (Ed.), Reinventing Anthropology. Vintage Books, New York, pp. 430–457.
Shostak, M., 1981. Nisa, Portrait of the Life and Words of a !Kung Woman. Harvard University Press, Cambridge, MA.
Shryock, A., 1997. Nationalism and the Genealogical Imagination: Oral History and Textual Authority in Tribal Jordan. University of California Press, Berkeley, CA.
Spicer, E., 1975. Perspectives in American Indian Culture Change. University of Chicago Press, Chicago.
Stocking, G.W., Jr., 1987. Victorian Anthropology. Free Press, New York.
Stoler, A.L., 1995. Race and the Education of Desire: Foucault's History of Sexuality and the Colonial Order of Things. Duke University Press, Durham, NC.
Taussig, M., 1993. Mimesis and Alterity: A Particular History of the Senses. Routledge, New York.
Tylor, E.B., 1871. Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Language, Art, and Customs, vols. 2. John Murray, London.
Vansina, J., 2004. Antecedents to Modern Rwanda: The Nyiginya Kingdom (J. Vansina, Trans.). University of Wisconsin Press, Madison, WI.
Vansina, J., 2010. Being Colonized: The Kuba Experience in Rural Congo, 1880–1960. University of Wisconsin Press, Madison, WI.
Wolf, E., 1997/1982. Europe and the People Without History. University of California Press, Berkeley, CA.
Worsley, P., 1957. The Trumpet Shall Sound: A Study of "Cargo Cults" in Melanesia. Schocken, New York.

Anthropology at Home

Sophie Chevalier, Université de Franche-Comté, Besançon, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Distinction between doing research at home and among the others is built into the social sciences as a division of labor between sociologists and anthropologists. Anthropology at home runs counter this distinction, historically based on the Great Divide between 'us' and 'them.' Yet more and more anthropologists work today in their own society and the gap between self and other is maintained less sharply than in the time of empire and nation-building. The decentred comparative perspective that goes with an ethnographic sensibility can be brought to bear on everyday life anywhere, including on the more familiar institutions of mainstream urban life.

Introduction

At one level, the distinction between doing research at home and among the others is built into the social sciences as a division of labor between sociologists and anthropologists. This is despite their historical unity as 'comparative sociology' in the work of Durkheim and Radcliffe-Brown. Thus for Claude Lévi-Strauss (1958), the sociologist observes his own society or one like, while the anthropologist tries to make sense of other cultures. The latter consists of taking a 'view from afar' on these societies, a perspective on the Great Divide between 'us' and 'them' that was intrinsic to colonial empire. Yet more and more anthropologists work today in their own society and some traditions, such as the French, have done so since the beginning, largely as part of its own nation-building project. As a field with the wider discipline, therefore, 'anthropology at home' can be complicated. Apart from anything else, working in one's own country, especially when it is large and diverse, may not count. The term is more likely to be used in countries that once had foreign colonies to denote a gaze turned inward rather than outward.

Ethnography is a method now used widely in the social sciences. Anthropologists have long claimed that their version of it was distinctive in at least three senses: their 'view from afar' allowed them to juxtapose the familiar and the unfamiliar in interesting ways; they always tried to fit their concrete observations into a wider social or cultural whole; and they brought to their ethnography a range of highly contrasted comparative examples. 'Anthropology at home,' the idea of a native working in their own society, runs counter to most of this, insofar as everything is or ought to be familiar, society as a whole is harder to envisage and the comparative method draws more on books than the fieldworker's own experience of contrast. The kinds of knowledge accessible to 'natives' and 'outside observers' are held to be significantly different; and certainly Lévi-Strauss reinforced this separation in contrasting structural anthropology with sociology. Of course, this is all connected to making the universalizing claim to capture the human experience in general, rather than just describe particular instances of it. As a result, anthropologists working at home have several techniques available to reintroduce a comparative perspective – through studying classes other than their own or adopting a historical approach. This was in fact the case with the first folklorists in Europe when they inserted themselves into rural worlds, other with an evolutionary paradigm in mind. Today, long after the collapse of European empires and of the rural/urban divide, the decentred comparative perspective that goes with an ethnographic sensibility can be brought to bear on everyday life anywhere, however familiar or unfamiliar the circumstances are supposed to be.

The idea of territory – and national territory in particular – is central to anthropology at home. In former imperial countries like France and Britain, metropolitan society was opposed to the colonies abroad which were considered to be the main source for a general social anthropology because they provide simple models of our complex civilization. Later of course, when members of the former colonies migrated in numbers to metropolitan countries, they provided objects of study at home whose cultural difference from the ethnographers introduced further confusion in the inside/outside dialectic of modern anthropology.

At this point, we will have to draw some lines around our topic. There are several very large countries – such as the United States, India, China, Brazil, and Russia – whose anthropologists worked within their own national territory, but mainly with groups considered to be external to mainstream society, perhaps as exotic as Pacific islanders to European colonists. These anthropologists could not be thought of as doing anthropology at home on remote hill tribesmen or Amazonian hunter-gatherers. In societies without external or internal colonies, anthropology at home was linked to the process of nation-building. In recent decades, all these societies have seen a shift to doing anthropology anywhere, including on the more familiar institutions of mainstream urban life. The typology of overseas empire, inland empire, and nation-states is not hard and fast; but in this article, I will focus on anthropology at home in the first and third types, less so in the second, but always with a regard for recent developments which have eroded such distinctions to a large degree. This means that my focus will largely be on Greater Europe (including Russia). We should also note that our subject has appeared under a number of different names: folklore studies, *Volkskunde*, ethnography, *ethnologie*, or anthropology at home.

In the four main schools which formed modern anthropology – Britain, France, Germany, and the United States – a variety of approaches to anthropology at home developed,

always keeping in mind George Stocking's (1982: p. 172) distinction between nationalist ('nation-building') and imperialist ('empire-building') strands that have shaped the discipline, sometimes separately, sometimes combined, but with different emphasis.

Anthropology at Home in France and United Kingdom

The intellectual and institutional context for the emergence of anthropology at home is very different for Britain and for France. There was, for instance, no classical school of British sociology. The founders of the British school of ethnographers sought to differentiate their discipline, as a kind of comparative sociology, from ethnology or folklore; and while British folklorists focused primarily on their own country, social anthropologists worked in faraway overseas. But folklore never evolved into a modern academic discipline under any name.

In a review of British intellectual life, Perry Anderson (1968) contrasted the British situation with sociology in continental Europe and North America, which took root as a conservative discipline celebrating the achievements of the bourgeoisie. Anderson argued that the British ruling class had no interest in social conditions at home, but instead promoted social anthropology as a means of studying the peoples of the Empire. This presented itself as 'the sociology of primitive societies,' basing its claim for universality on the presumption that social order everywhere was revealed most clearly in 'simple' societies (a notion shared by Durkheim and Mauss).

So British social anthropology in its prime did not have to distinguish itself from sociology and folklore was a weak competitor. As late as the 1960s, anthropologists took up posts that might otherwise have been filled by sociologists. Despite the relatively open field, British anthropology at home was slow to take root in the main centers. It was encouraged only by Max Gluckman's department at Manchester, which combined social anthropology with sociology. As a result, the leading figures of British anthropology at home, such as Ronald Frankenberg and Anthony Cohen, started out there.

In France, however, anthropology did not benefit from the expansion of the universities in the 1960s, in striking contrast to sociology. Whereas exotic anthropology in Britain long enjoyed an unchallenged monopoly, ethnologists working at home in France had to compete with both exotic anthropologists and sociologists. Sociology has long been the strongest social science discipline in France and ethnology was always on the defensive there, as it is today. 'The ethnology of France' has long been hidden away in corners of the universities, a marginal field that never had distinguished chairs similar to those held by professors of Volkskunde in Germany. Whereas the latter played a major part, with history, in nation-building, in France, history and geography dominated studies of the national question. Ethnographic descriptions of France and folklore museums were rare at the end of the nineteenth century; and anthropology at home slowly built a reputation as a legitimate discipline and source of collective memory in the interwar period.

In both Britain and France exotic anthropology led the way; the Institute of Ethnology in Paris catered exclusively for ethnography in the colonies and the ethnology of France struggled to establish itself. Both colonial empire and nation-building posed problems of reconciling political imperatives with scientific practice. This was particularly acute when the idea of citizenship and the national territory itself were at stake. In both countries, the scientific and political projects of exotic anthropology found a measure of convergence, but in France this convergence also took place at home.

The study of French history and society was placed in the service of the Republic. National feeling was always modified by attachment to one's 'little homeland' or local identity. Regionalism, starting out as a literary movement, remained loyal to the republic, even if tensions between the regions surfaced here and there. Indeed after the war regional chauvinism was a way of showing pride in being French which has generally managed the tension between centralization (jacobinism) and regionalism. Only during the Second World War did the Vichy regime require social scientists to separate their science from the republican ideal. The period immediately after the war saw an acceleration of the transition from folklore (the term vanished after 1945) to an ethnology of France influenced by social anthropology.

Arnold Van Gennep played a particularly significant role in the development of the anthropology at home in France, acting as a bridge between folklore and ethnology and, with 'rites of passage' was the first to coin a concept for the study of Europe that was taken up by the exotic anthropologists. But he never had a real job and remained marginal to the Durkheimian establishment all his life. In Britain, history and political geography always constrained anthropologists to express themselves with caution when addressing the national question and regional attachments were never part of nation-building.

The ethnography of Britain is a fragmented field, for sure, and its visibility was for a long time very weak. But the disciplines were much less closed there. Individuals and the contexts in which they work are more open to interdisciplinary collaboration in Britain, whereas their French counterparts withdrew into relative isolation from the 1970s. Anthropological dialogue with geography, psychology, and sociology was a hallmark of British studies published during the 1950s. Despite intensified policing of disciplinary boundaries since then, cross-fertilization has continued and it clearly adds to the strength and wider influence of British social anthropology. Today, many anthropologists whose primary interest is Britain, work in departments other than anthropology, where they develop successful collaborative relationships.

British ethnographers also became interested in cities long before their French counterparts. Urban and even industrial research was launched in the 1950s: they were looking at urban life, focusing particularly on kinship, which was seen to provide the framework that made the sprawling metropolis manageable. Engagement with novel field situations, compared with the classical village sites, led British anthropologists to be more innovative in their approach and at an earlier stage. French ethnologists, on the other hand, were penned into their rural worlds for much longer and were slow to visit the city. This is surely one another major contrast between the two countries.

In both, the rise of anthropology at home was linked to the end of colonial empire. For a long time doing fieldwork in one's own society was a way of either getting started before

doing the real thing abroad or of making do at home when exotic fieldwork was no longer possible. Only in the last three decades have significant numbers of anthropologists in both countries carried out fieldwork at home, with the result that these domestic efforts are now beginning to have a theoretical impact on the discipline equal to that of exotic ethnography.

British social anthropology's early emphasis on studying what were presumed to be 'simple' societies was reflected in the fact that much of the ethnography of the British Isles started in the Celtic margins – especially Wales and Ireland – which were considered relatively remote and undeveloped. The first fieldwork-based studies of British societies were undertaken by North Americans. But anthropology at home in Britain took off with the publication in 1957 of Frankenberg's 'Village on the Border.' From the 1960s, some British social anthropologists were influenced by more cultural perspectives: studies of 'community' brought a discourse of 'belonging' into considerations of identity in Britain (Cohen, 1982). Anthropologists discovered that cultural difference easily survives in the face of the homogenizing forces of modernity: this celebration of British cultural heterogeneity also contained more than a hint of nostalgia and salvage ethnography.

French ethnologists too cultivated the social and cultural margins long before they entered the city. If French research at home got going a lot earlier, its institutional framework encouraged collective work and was less reliant on individual initiatives than the British. The National Museum of Popular Arts and Traditions launched a number of large-scale team exercises in the 1930s, such as one in Sologne with Malinowski and Dumont, long-term studies of rural architecture and technology during the war and finally the Aubrac study of 1964–68. Other research centers undertook similar studies, such as the big interdisciplinary project at Plozévet (1960s), exploring a theory of biological anthropologists about a deformed hip common among the women of the region. By contrast to their British counterparts, French ethnologists left reflections on social class to the sociologists, and gender roles were taken up in France only at the end of the 1970s. From the 1980s onward, as a result of government policy, ethnologists were integrated in research programs organized around an interdisciplinary theme such as skills seen in a heritage context. A number of ethnologists today work on heritage and the history of anthropology.

Since the 1970s, there has been a convergence in both countries between the exotic and home branches of anthropology. In Britain, the earlier emphasis on comparative sociology has given way to a cultural anthropology influenced by French structuralism and American symbolic anthropology. Kinship was the focus of French ethnology's self-renewal in the 1970s, under the influence of Lévi-Strauss's social anthropology. The outstanding work of this period was carried out by 'the Ladies of Minot,' members of the Laboratoire d'Anthropologie Sociale, within the framework of the Châtillonais team project, which brought the basic themes of exotic anthropology – kinship, ritual, identity, and socialization – to the ethnology of France. All of these studies were of rural areas. Only from the early 1980s did French ethnologists begin to take an interest in the city, industrial life, ethnic groups, sport, and complex organizations, in short the whole range of social practices or almost so.

In both France and Britain today, anthropology at home is now flourishing, drawing on each other and on their fellow branches in the rest of Europe. One can identify two trends in the dominant anthropological approaches of both countries. One, the majority, is still concerned with trying to understand the properties of social and cultural groups through the articulation of individual actions with collective structures. The other, identified with a branch of French sociology and some anthropologists in Britain, has abandoned entirely the ambition of locating its findings within any notion of social structure, being concerned with specific practices that individuals have in common, sometimes without even locating these in time or space, never mind in society.

'Anthropology at Home' in the Rest of Europe

The majority of schools in the rest of Europe may be summed up as 'nationalist anthropology' in that they were strongly influenced by the contingencies of national political history. The details of each case lie beyond the scope of this essay; but it is possible to identify some common strands by highlighting the formative example of Germany, followed by Russia's influence after the formation of the Soviet bloc.

According to Gingrich (2005), the German Enlightenment tradition of anthropology turned to introspection in the nineteenth century, first by giving priority to German rather than non-German topics and then to spirit and soul rather than the facts of what people really do. This trend combined with the concept of *Kultur* developed by Herder – a key figure in the history of German anthropology – in which he emphasized languages, customs, and mentalities to arrive at an ethnocentric division between German and nearby European *Kulturvölker* and most of the others (*Naturvölker*). Early anthropological writing there had little to do with colonialism (Germany acquired some colonies in Africa only after 1884). The preeminent tendency was to look inward, promoting historicist folklore studies with a nationalist flavor. Eventually, this led to a separation between folklore studies at home (*Volkskunde*) and ethnology aboard (*Völkerkunde*). Ethnography as such was more common in the latter.

As Arendt (1958) shows, the Romantic imaginary of nationalism in the German-speaking countries – and in most of the Central and Northern European countries – was perceived as a return to the nation's alleged rural roots. Rural folklore studies were thus linked to a mass movement whose inspiration was literary as much as anything else, the great fairy tale collections of the Grimm brothers being the prime example. Folklore studies were a form of historicism and even today, *Volkskunde* has remained strongly attached to historical methods and to the academic field of history. At the margins of the German-speaking world, popular and academic folklore studies were less nationalist and more interculturally oriented: Switzerland and Austria stand out in this respect, one as a multicultural democracy and the other as the center of a multinational empire. In Poland, Hungary, and Czechy, ethnography was linked to a movement for national independence. This is why, from a methodological perspective, Gellner (1959) locates the intellectual roots of ethnographic fieldwork in Central Europe at the turn of the century.

Anthropology at home in Central Europe, with its core in folklore studies, was separated intellectually and institutionally from anthropology abroad, but it also had no relations with sociology or with the long debate over the future of the peasantry ('the agrarian question' launched by Lenin, Kautsky, and Chayanov) before the 1970s. In the twentieth century, the two branches of German anthropology were likewise detached from mainstream developments in international sociocultural anthropology until the same period. From the 1930s, anthropology and folklore studies were smoothly integrated into the Nazi project. They became inseparable from physical anthropology, which was now the main scientific arm of this racist ideology. Folklore studies were also integrated into a pangermanic project, including the study of German minorities in Central Europe: ethno-cartography and folkloric atlases were their principal tools. Empirical studies were also carried out in several countries aimed at establishing a 'New European order' with Germans at the top.

The new intellectual environment after the war saw *Volkskunde* persisting separately in Germany until the late 1970s, but now influenced by new concepts and methods in history such as 'history from below.' The leading figure here was Hermann Bausinger in Tübingen who transformed folklore studies from a nationalist approach into a fieldwork-based ethnography of sociocultural process at home. The name changed too from *Volkskunde* to *Europaïsche Ethnologie* which made it a serious partner for sociocultural anthropology until now. The objects and methods of these anthropologists scarcely differ from those of their colleagues who work outside Europe, except perhaps for their emphasis on history.

Scandinavian anthropology went through a similar transformation. The division between anthropology at home and abroad continues there, with local variations, but always with a national, even nationalist dimension. Moreover, there was little or no indigenous tradition of sociology there. The Nordic countries – Sweden, Finland, Denmark, Norway, and Iceland – were influenced at first by German intellectual life and so anthropology found its inspiration there; but after 1945 it developed in different directions in each country. Anthropology at home, however, increasingly took its concepts and methods from the general international discipline from the 1960s. Thus in Sweden before that time, 'studying the life of the people' (*Folkslivsforskning*) was a department attached to a section of the Nordic Museum. Borrowing the concept of 'community' from the Anglo-Saxons facilitated this transition. Although the institutional separation between departments of ethnology and anthropology may persist, working on one's own society is no longer restricted to the ethnologists.

The countries of Central and Eastern Europe came under a double influence: German before the Second World War and Soviet after it. We might speak of a third influence today, the global hegemony of American anthropology. Poland is a case in point. Folklore studies were an instrument of national survival after the country was carved up between foreign powers, being essential to an ethnic nationalism whose object was to raise national consciousness through a revival of peasant traditions. After the First World War, Polish anthropology was renamed as ethnology and it combined the science of peoples with the science of the people (*Völkerkunde* and *Volkskunde*). But after the Second World War, Polish anthropologists, along with other countries of the Soviet bloc, adopted the term ethnography for their discipline. This was intended to highlight the dialectical unity of fact (*-graphy*) and theory (*-logy*) in research. Ethnology made a comeback in the 1970s. Then after 1989, departments were renamed as 'ethnology and cultural anthropology,' indicating an openness to both the European and Anglo-American traditions.

In Hungary, the term *néprajz* (description of the people) denoted an ethnography aimed at documenting the phenomena of Hungarian popular culture that was based on fieldwork, analysis, and comparison of these phenomena. *Egyetemes néprajz* (universal ethnography) included studies undertaken outside Europe. As in Poland, the origins and development of ethnography were linked to the politics of external domination, with the objective of establishing the authentic character of Hungarian culture linked to its origins in the history of the Magyars. Ethnography was devoted to studying Hungarian popular traditions in the countryside, but also took in other peoples who lived in the same political space. Fascination with peasant culture carried over into music through the great twentieth century Hungarian composers, Bela Bartok and Zoltán Kodaly. Peasant studies during the socialist period allowed Hungarians to highlight national identity while observing the restrictions placed on their discipline by the regime; they were even able to study religion under the guise of 'survivals.'

Similarly in Bulgaria, for example, from the 1960s, folklore studies allowed anthropologists to study national roots without exciting the suspicions of the ruling party. These studies included religious icons, material culture, music, and popular literature. The school of descriptive ethnography which flourished in Czechoslovakia still exists, but research there tends to be focused on oral literature. Gradually, methods developed for the study of popular literature were extended to other subjects.

In Russia, then in the Soviet Union, anthropology at home has been through ups and downs, which at times came close to its being banned. Of course we are dealing here less with a national tradition, more with several waves of inland empire. Expansion toward the East led the Russian tsars of the nineteenth century to establish museums and similar institutions to organize ethnographic studies. Peter the Great founded a museum of foreign cultures in the Russian empire in 1714; another, also in St Petersburg, was founded in 1895 to exhibit collections from the various peoples integrated into Russia. A strong institutional distinction, which exists today, developed between those who studied Slavs – anthropology at home – and those who worked on other peoples incorporated into the Soviet Union. The revolution itself did not disturb this work, but things became serious by the end of the 1920s. Researchers could continue their work on peasant traditions as long as they were far from the center. This led to the publication of major works by Zelenin (1927) on the eastern Slavs (a book banned in 1929 and republished in 1991) and above all by Propp (1928) on the morphology of folktales which was very influential. Eventually ethnography came to be seen as a bourgeois science and some anthropologists were forced to give up their work or to go into exile. Those who remained were discouraged from drawing attention to 'retrograde' elements of popular culture. They tended to show that peasants remained

strongly attached to their traditional culture and were largely immune to modernization and to socialism. Obviously, drawing attention to religious 'superstitions' was out. As a survival strategy, ethnographers had to keep their heads down, for example, by working in institutes not ostensibly concerned with ethnography or by taking refuge in provincial universities. Sometimes these strategies succeeded, as in the case of the Tartu School in Estonia whose work on the semiotics of culture had a significant influence in the countries of the Soviet bloc. It was recognized and approved in 1970 and gave rise to a new subject, 'culturology' that links semiotics to a broader definition of culture.

Several new research themes have emerged since 1989: religion, in particular the role of orthodoxy in Russian identity; Soviet 'traditions' and the political manipulation of traditional culture in the socialist period; and the social transformations brought about by recent political and economic change.

American Anthropology

American anthropology was late to arrive on the international academic scene and a lot of foreign-born or trained scholars played a critical role in the discipline's development there. Although the early American anthropologists tended to work within their national territory, their focus was mainly on the Amerindians who certainly did not share the notion of belonging to the same nation (many of them were employed in the Indian service). Some worked on immigrant groups and small towns in a process that might be likened to nation-building; others on neighboring countries, Mexico in particular. Rather than stress anthropology and home and abroad, American anthropologists, with their four-field approach, preferred to make a distinction between simple and complex societies of the kind that sustained ethnography in the British and French colonies. Venturing abroad or studying 'complex' societies was a big step for American anthropologists. The urban sociology of the Chicago ecological school was part of a joint department with anthropology, so that the city's neighborhoods were studied side by side with Mexican peasants.

The main impetus for work inside the United States came from Melville Herskovits' pioneering studies of race relations and communities studies such as Robert Lynd's in Middletown which was itself modeled on Robert Redfield's study of the rural–urban continuum in Mexico during the 1920s. From the early 1930s, urban communities were studied by anthropologists, such as William F. Whyte's (1943) innovative study of an Italian slum in New York city. Many of these small-scale ethnographic studies addressed problems of contemporary American life, such as class, race, and sometimes the power structures and economic organization of these communities. During the 1960s, two trends dominated research carried out in the United States: urban anthropology and studies of ethnicity studies, each stressing the multiplicity of cultural groups coexisting within national society. An exception to this was David Schneider's more traditional study of kinship in Chicago which was intended to be part of a comparative project with Raymond Firth in London.

Whatever the relevance of the distinction between doing anthropology at home and abroad, it has vanished since the 1980s when anthropologists everywhere, but especially in the United States, discovered that we are all living in one world unified by neoliberal globalization. So they studied that.

Specific Institutions

For anthropological practice to take root in a national territory, as opposed to abroad, requires institutions able to support specialized researchers and to communicate their results to a public. This means museums, universities, research centers and journals, as well as umbrella organizations. In regions like the former Soviet bloc, specialized institutions would be part of an overarching Academy of Sciences. There is a section of the French Ministry of Culture called the *Mission du patrimoine ethnologique* (Mission for the ethnological heritage) which was created in the 1980s to promote and fund research on France.

Museums

Museums have played a crucial role in sustaining anthropology at home everywhere, usually focusing on the material culture of the people studied. Museums of this kind had their origin in competition between European nations. In countries which distinguish between anthropology at home and abroad, there are usually two distinct types of museums corresponding to this division.

Museums in France and Britain illustrate the general point and the contrast between the two countries. Both countries developed local and regional museums – the Welsh Museum near Cardiff was an early pioneer in Britain, in contrast to the British Museum and other imperial collections in London, and the Museon Arlaten in Arles was founded in 1899. The National Museum of Popular Arts and Traditions (MNATP) eventually broke away from the Trocadero museum in Paris (later the *Musée de l'Homme* and now the *Musée du Quai Branly*) which specialized in exotic anthropology, but had a 'French hall' since 1884. The MNATP was formed as an independent institution in 1938 to encourage and shape ethnology in France; it finally had its own building in 1969. It was at first modeled after the Scandinavian example and later the Soviet Union and Germany (*Heimatmuseen*). It aimed to provide a national synthesis of 'rural material civilization' and it took charge of all regional and local museums. The museum was closed down in 2005 and its collections moved in 2010 to Marseille, where it is now being transformed into the Museum of European and Mediterranean Civilizations. The idea of a national museum of ethnology never took off in Britain.

Most of the other European countries have some kind of National Museum of popular arts or ethnography, for example, Bucarest's Museum of the Peasantry, the Museum of Ethnography in Kiev and Budapest, and numerous local museums or *Heimatmuseen*. All of these museums share the objective of integrating popular traditions in all their variety into the national culture. Since the nineteenth century, these have been complemented by open air constructions exhibiting the diversity of national architecture in one synthetic site, the leading

example being Skansen in Sweden (1885). A development of this idea later was the 'ecomuseum' which sought to present a particular territory through everything it contained. Museums, more than anything else, entrenched the deep split between 'exotic' and 'popular' material culture that has characterized anthropology in Europe ever since, in the process undermining anthropology's pretension to being a universal discipline.

In a few countries, especially the United States, one can also find community museums, which allow these places to portray their own role in national history.

Associations and Journals

Anthropologists are organized in professional associations at different levels: regional (European Association of Social Anthropologists-EASA); national (American Anthropological Association, Association of Brazilian Anthropologists, etc.); and topical (medical anthropology, anthropology of religion). Inside these big associations, there are sometimes sections concerned with 'anthropology at home' like the 'Europeanist network' of the EASA whose members usually work in their own country. Some associations are explicitly specialized on anthropology at home like the *Société d'Ethnologie Française*. But more and more, these kinds of associations take a wider geographical area. The *American Folklore Society*, for example, defined itself as an association of folklorists whose members study and communicate knowledge about folklore throughout the world, and not only in America. But this association's publications have retained a strong historical approach.

Most associations publish journals which are rarely today restricted to 'anthropology at home,' despite their titles, *Ethnologie Française* being a case in point. This journal now publishes European ethnology and sometimes papers based on research carried out outside Europe; but in general, the majority of issues have a regional focus, such as 'Ethnologia Europaea' or 'Ethnologia Balkanica.' All of this points to the growing difficulty today of maintaining a clear-cut distinction between anthropology at home and abroad. One might indeed ask if there is indeed a future for 'anthropology at home.'

The Future of Anthropology at Home

In the twenty-first century, most ethnographers work in their own country and the gap between self and other is maintained less sharply than in the time of empire and nation-building. The dominance of European and American anthropology is now challenged by the rise of emerging societies, such as Brazil, India, Mexico, and Nigeria, as well as by the residue of British Empire in Canada, Australia, and South Africa. Anthropology might be said to be more dynamic in these places than in the old imperialist centers. In some cases, such as India and Nigeria, anthropologists cultivate a traditional niche, studying remote tribal groups that are still perceived as requiring integration into national society. In China, anthropology is concerned with the 7% of the population that is not Han. The rest is the domain of sociology. Brazil now has the second largest anthropology association in the world. Until recently, the bulk of research there was carried out in Amazonia – and this is still the case for the National Museum in Rio de Janeiro. But Brazilian anthropologists have taken advantage of the rigidity of neighboring social science disciplines, sociology, and political science in particular, to enter the field of contemporary urban society and culture, with impressive results for their national visibility. Of course, such incursions into home society are no longer separated institutionally or theoretically from a world anthropology that studies people everywhere; and in this respect, the anthropology of emerging countries resembles Europe at the present time.

We should never forget that relations between ethnographers, their subjects, and their readers have changed profoundly in recent decades, all of which has altered the status of anthropology at home as opposed to anywhere else. Anthropologists study here and there. The anthropologists themselves come from here and there. Maybe we will soon discover that everyone needs to feel at home in the world.

See also: Community Sociology; Community Studies: Anthropological; Community and Society: History of the Concepts; Community, Social Context of: The US Case; Ethnography in Applied Social Research; Ethnology; Folklore; Home, Anthropology of; Museums; Peasants and Rural Societies in History (Agricultural History); Peasants in Anthropology.

Bibliography

Anderson, P., 1968/1992. Components of the national culture. In: Anderson, P. (Ed.), English Questions. Verso, London, pp. 48–104.
Arendt, H., 1958. The Human Condition. University of Chicago Press, Chicago.
Barth, F., Gingrich, A., Parkin, R., Silverman, S., 2005. One Discipline, Four Ways: British, German, French and American Anthropology. University of Chicago Press, Chicago.
Bausinger, H., 1990. Folk Culture in a World of Technology. Indiana University Press, Bloomington.
Bendix, R., 1997. In Search of Authenticity. The Formation of Folklore Studies. The University of Wisconsin Press, Madison.
Bendix, R., Löfgren, O., 2008. Thematic issue: rethinking Europe as a field for European ethnology. Ethnologia Europaea 38 (1), 5–7.
Chevalier, S., forthcoming. The rise and fall of French anthropology at home. In: Chevalier, S. (Ed), French Anthropology Today. Sean Kingston Publishing, Oxford.
Chevalier, S., Edwards, J., Macdonald, S. (Eds.), 2007. L'anthropologie de la Grande-Bretagne: une discipline en plein essor. Ethnologie Française XXXVII (2), 197–211.
Cohen, A., 1982. Belonging: Identity and Social Organisation in British Rural Cultures. Manchester University Press, Manchester.
DaMatta, R., Hess, D., 1995. The Brazilian Puzzle. Columbia University Press, New York.
Frankenberg, R., 1957. Village on the Border: A Social Study of Religion, Politics and Football in a North Wales Community. Cohen and West, London.
Gellner, E., 1959. Words and Things. Gollancz, London.
Hann, C., Sárkány, M., Skalník, P. (Eds.), 2005. Studying Peoples in the People's Democracies. Socialist Era Anthropology in East-Central Europe. LIT Verlag, Berlin.
Hannerz, U., Gerholm, T., 1983. The shaping of national Anthropologies. Ethnos, 12–35.
Kuper, A., 1996. Anthropology and Anthropologists: The Modern British School. Routledge, London.
Levi-Strauss, C., 1958/1983. Structural Anthropology. University of Chicago Press, Chicago.
Lins Ribeiro, G., Escobar, A. (Eds.), 2006. World Anthropologies. Disciplinary Transformations within Systems of Power. Berg, Oxford.

Ntarangwi, M., Mills, D., Babiker, M. (Eds.), 2006. African Anthropologies: History, Critique and Practice. CODESRIA, Dakar, Senegal.
Pereino, M., 1998. When anthropology is at home. Annual Review of Anthropology 27, 105–128.
Propp, V., 1928/1968. Morphology of the Folk Tale. University of Texas Press, Austin.
Sharp, J., 1980. Two separate developments: anthropology in South Africa. Royal Anthropological Institute News 36, 4–5.
Srinivas, M., 2002. The insider versus the outsider in the study of cultures. In: Srinivas, M. (Ed.), Collected Essays. Oxford University Press, Delhi, pp. 553–560.
Stocking, G., 1982. A view from the center. Ethnos 47 (1–2), 172–186.
Stocking, G., 1991. Colonial Situations: Essays on the Contextualization of Ethnographic Knowledge. University of Wisconsin Press, Madison, WI.
Wallerstein, I. (Ed.), 1996. Open the Social Sciences. Report of the Gulbenkian Commission on the Restructuring of the Social Sciences. Stanford University Press, Stanford.
Whyte, W.F., 1943/1993. Street Corner Society: The Social Structure of an Italian Slum. Chicago University Press, Chicago.
Zelenin, D., 1927. Russische (Ostslavische) Volkskunde. de Gruyter, Berlin and Leipzig.

Anthropology, Genomics, and Human Variation: National Roots

Gisli Palsson, University of Iceland, Oddi, Reykjavik, Iceland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Emphasizing concerns with cultural identity, belonging, ancestry, and health, this article discusses the national roots of the science of human variation and differences, focusing on the significance of genomic studies for national projects and the significance of national concerns for genomics. This is illustrated through a discussion of national biobanks, genomic projects, and genealogical databases, while highlighting purposes, limitations, and future scenarios.

Introduction

Emphasizing concerns with cultural identity, belonging, ancestry, and health, this article discusses the national roots of the science of human variation and differences, in particular anthropology, focusing on the significance of national concerns for genomics and the significance of genomic studies for 'national projects,' for attempts to identify and manage national bodies, populations, and identities – in sum, for nation building. This is illustrated through a discussion of national biobanks, genomic projects, and genealogical databases, while highlighting purposes, limitations, and future scenarios. From the beginning of national projects, with the emergence of nation states and empires, phenotypic markers (skin color, hair texture, skull shapes, etc.) played important roles in identifying and demarcating social groups and for establishing criteria of citizenship (Lindee and Santos, 2012). While national projects were focused on the broad and apparently neutral and culture-free documentation of 'difference' and 'variation,' they usually separated 'us' from 'them,' expressing ethnocentric concerns with race and purity (Marks, 2008). A series of theoretical and technical innovations during the second half of the twentieth century, in particular the discovery of the double helix, allowed geneticists and scholars in related fields to finally explore the contours and structures of genetic material, to 'voyage' into the 'universe within,' in the language of the Human Genome Project in 2000. These developments, sometimes identified as the 'new genetics' (Palsson, 2007), have redefined the conceptual landscape and biometrics of human difference. Genotypic differences seem not to correspond or submit to earlier framings of phenotypic differences and classifications. Many biological anthropologists, in fact, have sought to change popular understandings of race, emphasizing that race has no biological basis. This does not, however, rule out the possibility of somehow reintroducing earlier categorizations of races under the modern labels of genotypes, populations, and ethnic groups.

Background

National collections and databases of the kind discussed here draw upon a series of developments in the history of states and bureaucracies, including practices of writing, accounting, and monitoring represented by 'statistics of sickness.' Thus, in nineteenth-century England the medieval documentation that traditionally had focused on births, deaths, and marriages was extended to statistics and probabilistic methods to describe the health risks and tendencies of the population and its subdivisions: "Disease, madness, and the state of the threatening underworld, *les misérables*, created a morbid and fearful fascination for numbers upon which the bureaucracies fed" (Hacking, 1982: p. 287). Developed in one nation state after another, such statistics represented biopolitical concerns with monitoring and governing the health of the national body and the national population. While this new 'avalanche in numbers,' Hacking suggests, was rarely efficient in managing the population of study, it had subversive effects, namely "to create new categories into which people had to fall, and so to create and to render rigid new conceptualizations of the human being" (1982: 281). Keeping in mind that the term 'nation' is etymologically derived from *nascere* (to be born), health statistics, in a sense, establish the birthmarks of the national body.

At the beginning of this century, a series of biomedical projects normally described as 'biobanks' were developed in different parts of the world. Drawing upon the scientific perspectives of the new genetics and the politics of neoliberal governance, such projects were usually designed and developed as *national* projects, focusing on national populations, health statistics, and lifestyles. These biomedical innovations are not just innovative technical, scientific, and economic enterprises; they are radical experiments in biopolitics with potentially diverse social implications for individuals, families, communities, and the public at large. The broad responses to such experiments, however, vary from one case to another.

The History of 'Human Variation'

For a long time, the study of human bones was an important component of physical anthropological and archaeological research on human variation and history. Skulls in particular were endlessly recovered from graves, measured, classified, analyzed, stored, and displayed in museums and archives. With its set of tools for measuring human bodies, from the 'anthropometron' of the seventeenth century and the 'craniometer' of the nineteenth century onward, physical anthropology carried an aura of objectivity and precision, of proper science. It helped, too, that rules were drawn up for the

standardizing of anthropological measurements, at the International Congress of Anthropology in Monaco in 1906 and in Geneva in 1912. Perhaps this hammered home the point that racial classifications and national typologies were somehow engraved in the biological world, and not imposed by the spectator.

Passionate bone collection and measurements continued in some contexts throughout the twentieth century, both constituting and being constituted by the imagined communities (cf. Anderson, 1983) of nations, cultures, and disciplines and their traditions of remembrance and authority (Connerton, 1989). As Verdery (1999: 27) notes, because bones are material things and 'indisputably there,' they have political lives of their own: "A body's materiality can be critical to its symbolic efficacy." Hooton's (1931) influential work *Up from the Apes*, originally published in 1931, discussed the 'blood stream of the human races,' comparing the history of races with the circulatory system. The blood ties assumed by Hooton clearly reflected his own racist ideas. Nordic peoples were, for instance, according to him, both fairly pure races and near the forefront in human development, while bringing up the rear were 'Arctic Mongoloids.' Commenting on race, Hooton suggested that "confusions of usage are usually confined to the non-anthropological writing public. All anthropologists agree that the criteria of race are physical characteristics" (Hooton, 1946: 447). For Hooton, skeletal material was the main source of information on human variability and national types. The measurement and classification of bones became an obsession. This is underlined by the extensive cross-cultural bone collections of the Peabody Museum at Harvard University, which are stored in hundreds of cardboard boxes in a large hall of the museum and labeled with the names of nations, peoples, and places.

Saturated with listings and descriptive statistics, the massive catalogs often produced by physical anthropologists provide a useful clue to the 'problem orientation' of their time: "amass enough data and problems will sort themselves out" (Scott et al. 2000: p. 347). Often devoid of theory and focusing on preconceived human types, passionate skeletal studies and catalogs continued in some contexts throughout the twentieth century. Sometimes the goal was partly to understand particular medical conditions evident in skeletal structures of particular populations.

Modern genetics was born at the beginning of the twentieth century with the rediscovery of Mendel's laws of inheritance. By the middle of the century, a series of critical developments revealed the molecular mechanisms of inheritance. Twentieth-century advances in cell research also made it possible to extract, store, and manipulate cellular material outside the host organism. With the 'modern evolutionary synthesis' of the 1940s and the 'new genetics' of the 1960s onward, biological anthropology took off. It was not until the 1990s, however, that DNA gradually began to replace skeletal material and other sources (texts, languages, and archaeological remains) as the avenue into human history and variation. A few geneticists and biological anthropologists began to see bones as almost trivial antiques (see Sommer, 2008), somewhat like ancient manuscripts, secondary to DNA sequences and gene frequencies. At the same time, genomics acquired a central place in national projects, as a powerful avenue into health and history. Keeping in mind, however, that formerly discarded sources of data sometimes acquire new meaning as a result of conceptual and methodological advances, skeletal morphology should not be discounted.

Intensive genetic research over recent decades has revealed what are usually seen as the biologically encoded instructions necessary for constructing and maintaining a living example of an organism. What people make of this in different contexts, however, is an important ethnographic question. One form of current gene talk assumes that genetic makeup largely accounts for individual development and characteristics. Along with biotechnological advancements, such a perspective has led to the escalation of research on the genes assumed to be implicated in human diseases. A related development is 'biobanking,' the assembly of a variety of medical and genetic information on specific populations, for the purpose of advancing the industrial development of medical remedies. Anthropologists and other social scientists have studied several aspects of these developments, particularly their implications for biopolitics and the construction of genetic citizenship, notions of belonging informed by genetic theory and practice (Rapp et al., 2001).

Biobanks

Cambon-Thomsen (2004: p. 866) refers to 'biobanks' as 'organized collections of biological samples and data associated with them,' one subset of which is 'population biobanks.' Such a definition is no longer exact as many databases now do not contain actual biological samples, being restricted to numeric representations of molecular properties. Since the biobank concept is rather broad and fuzzy, several others have been suggested for the kind of large-scale biomedical assemblies employed by the new genetics, including 'genetic databases,' 'biomedical collections,' 'population biomedical collections,' 'genebanks,' and 'biothèques' or 'biolibraries.' Tutton and Corrigan (2004: 3) see 'genetic databases' as "involving the collection, storage and use of physical tissue (usually blood, but by no means exclusively), genotype and other biological information derived from that tissue, and a variety of personal data from populations of various sizes." In such a definition, a genetic database typically combines (temporarily at least) two kinds of banking, one corporeal (DNA samples) and another informational (including medical records). A number of national or regional genetic database projects and proposals have been discussed in a variety of contexts since the 1990s – in Iceland, Tonga, Sweden, the United Kingdom, Estonia, Québec, Taiwan, Germany, Israel, Japan, Latvia, Newfoundland, Norway, Sardinia, Singapore, and the United States – with concerns with nationhood, health, monitoring, research, and commercialization, revealing at the same time different scales and designs (Palsson, 2007; Gottweis and Peterson, 2008).

Some database projects involve tiny and rather homogeneous island populations, others are set up for 'ethnic' regions, and still others cover huge and complex populations, potentially a substantial part of humanity. Apart from differences in the scale of the populations involved, database projects differ with respect to the processes and criteria of sampling, property

arrangements and management regimes, the nature and degree of commercial involvement, forms of consent, as well as measures for the protection of personal information. There are important similarities as well. All contexts represent a concern with the governing of bodies, populations, and welfare, monitoring the health of the national or ethnic body. One of the implications of this monitoring is the geneticization of notions of citizenship and belonging. DNA signatures not only specify who belongs to the nation or the ethnic community, who is entitled to the relevant privileges and who is not (the privileged rights to casino profits, for instance, in some indigenous North American contexts), but also help to establish new biosocialities along genetic lines, real or virtual communities based on shared characteristics or conditions (breast cancer, arthritis, etc.). This development is accelerated by digital genealogies, globalization, and network society.

The National and the Transnational

In 1998, the Icelandic Parliament passed a bill authorizing the construction of a Health Sector Database. In doing so, it launched a 'genome biobank boom' (Hinterberger, 2012), a new concept in biomedicine that has had a significant impact on biomedicine in many national contexts. The license to construct the database was open to competition; the licensee would finance it, and the resulting assembly would belong to the National Health Service, with the licensee retaining privileged rights to commercialize it for 12 years. Eventually, a company called deCODE genetics, Inc. was granted the license to construct the database. The text for the license specified that the company should pay the Icelandic state a certain fee for the assembly and use of the records of the medical service. For the spokespersons of the database, the power of genetic and epidemiological analyses would be greatly enhanced. The medical records available since 1915, it was argued, would allow for the exploration of a set of new questions on the interaction among a number of variables apart from genetic makeup and genealogical connections, including variables pertaining to lifestyle, physical and social environments, the use of particular medicine, and degree and kind of hospitalization.

The project seemed plausible and scientifically promising because the Icelandic population has been small, relatively homogeneous, and comparatively isolated. Such populations, it was argued, would simplify and speed up the hunting for mutant genes potentially responsible for common diseases. It also helped that Icelanders have had a passion for keeping genealogical records, facilitating analyses of genetic distance and correlations between genotypes and health. A relatively homogeneous population with good genealogical as well as medical records (for precise phenotypic identification of symptoms), it was argued, was the ideal experimental site for biomedical analysis. According to deCODE genetics and their investors, Icelanders met such conditions better than most other populations. Similar arguments were advanced in some other contexts, although some maintained that larger and more heterogeneous populations were better for such purposes. From the point of view of members of the Icelandic Parliament, ministers and representatives of the state, the Health Sector Database would render the health care system more efficient, making it possible to monitor the impact of drugs, health care, and the role of lifestyle. The results might be useful for pharmaceutical companies and for the medical service, producing information on potential drugs, particular genes or proteins, and possible preventive measures in terms of consumption and lifestyle. From the point of view of deCODE genetics, a database would speed up the search for genes responsible for common diseases and the development and manufacture of 'personalized' drugs. As Winickoff (2006: p. 97) observes, the Icelandic Health Sector Database became an experimental site for genomics and genomic governance in the sense that it "helped produce the technological, political, and normative terrain of all large-scale genomics initiatives today, not just Iceland's."

Another illuminating case is the Canadian project Cart@gène focusing on Québec. The plan was to approach 90 000 people aged between 25 and 74 years, randomly selected from medical records (Hinterberger, 2012). It was hoped that two-thirds of those approached would respond positively and serve as the foundation of the database through their biological samples. The participants would then respond to questionnaires about family history and health profile and donate blood for DNA analysis. The Cart@gène project is a joint venture of the Balsac project of the University of Québec, research institutes at McGill University, and the Faculty of Law at the University of Montréal. The aim is similar to that of the deCODE project – to map the local genomes, to throw light on the 'predispositions' of particular groups for developing certain diseases (e.g., hypertension, diabetes, asthma, and cancer), to explore the relationship between genetic characteristics and environmental variables, and to establish the local economy as an important site in the globally emerging genomic research (Bibeau, 2004).

During the planning process, important changes were made to the original outlines, partly as a result of debates on the relative roles of the state and private companies. For one thing, information entered into the bank would remain anonymous. Eventually, the project was defined as a 'population' health survey, a nonprofit enterprise with no specific commercial involvement. Bibeau (2004: 100) suggests that despite differences with respect to commercial involvement, the contexts of Québec and Iceland have much in common, in terms of public confidence in researchers, the progress of medicine, attitudes toward the pharmaceutical industry, and support to public health care services. Last but not least, Québec and Iceland share the scientific advantages provided by great interest among the public in family histories. About 10 000 French immigrants settled in 'New France' in Québec between 1608 and 1760 and, apparently, many people in Québec can trace their roots back 10 or even 15 generations. The success of modern genealogy companies can be partly explained by the desire of the descendants of the pioneers to trace their roots to a particular region in France and, obviously, the Cart@gène project benefits from this interest in roots. Bibeau (2004: 120) suggests, however, that the Québec genome has never been 'homogeneous' and that the true reason for interest among academics and industry in genome research, irrespective of the rhetoric of roots, has to do with the fact that Québec has demographic files and genealogical information 'without a doubt more complete

than those in Iceland and just as rich in information as the genealogical sources of the Mormons in Utah.'

Emerging in many contexts throughout the world, communities of relatedness based on genetic testing represent a revitalization of biological metaphors to mark belonging beyond established political boundaries, signaling the growth of what Glick Schiller (2005) calls 'long-distance nationalism.' Significantly, the Canadian Cart@gène project has partnered with a genealogical project in Québec (Balsac), a population register in a computerized database established for tracking family histories of French Canadians (Hinterberger, 2012). Such partnership seems geared to nourishing long-distance French nationalism, the imagined community of descendants of the 'Québec founder population' and their European ancestors, across national boundaries. Transnational political projects to build a state or an ethnic community thus increasingly draw upon notions of biological belonging informed by genetic testing. Israel is a case in point (Prainsack and Hashiloni-Dolev, 2009).

Despite the national or regional rhetoric of many database projects, there is much formal and informal collaboration among them. Recently, L. L. Cavalli-Sforza, one of the pioneers of the Human Genome Diversity Project, has suggested that national genetic databases might contribute to the project's collection by donating blood samples, 'representing a minute fraction of their collection, ideally forming a geographically (and if necessary, linguistically and ethnically) stratified sample of the country' (Cavalli-Sforza, 2005: 340). One of the recent territories of biomedicine is the move beyond 'national' genomes to comparative datasets such as that of the International Haplotype Map Project, drawing upon the power of genome-wide scans and related technological and methodological advances. Drawing upon its comparative tool kits, anthropology has often been involved in such projects, sometimes as a key player (Palsson, 2008). Quite possibly in the future, large-scale genetic and genomic databases will be operative at numerous sites, either permanently or from time to time. For gaining larger numbers and samples, which may be important for analyzing many common diseases, different database projects may find it feasible to temporarily share data, hunting and gathering in the human genome on an unprecedented scale under some kind of permanent institutional umbrella.

Genealogies and Ancestry

Iceland's Health Sector Database project collapsed for several reasons, including financial, ethical, and legal ones. While the project failed to materialize, it had several spin-offs, including the development of flexible, interdisciplinary research teams; powerful genomic laboratories; bioinformatic frameworks; the personal genomics of deCODEme (Palsson, 2012); and, last but not least, the Book of Icelanders, a genealogical database that lists the majority of the total number of people ever born in Iceland. The point is not simply to record individuals but rather to be able to connect them to each other. deCODE genetic's chief aim by the construction of the Book of Icelanders was to advance genetic and biomedical research. One of its two versions is available only to deCODE researchers. In this case, no names are included, only numbers or IDs that allow for the combination of different datasets on a limited basis for particular research purposes.

Some of the imagery of the Book of Icelanders and the deCODE project can be taken as emblematic of national genomic projects in general. One of the images (Figure 1), reproduced on deCODE genetics' T-shirts and web pages, has functioned as the informal logo of the company. It shows a pedigree with 11 generations. The current-day Icelanders diagnosed with asthma (filled squares and circles in the figure) are limited to the last generation or two, but their relationships are known through 11 generations and all of them are descendants of a single founder born in the seventeenth century. Apart from its visual appeal, the circular image may serve, consciously or not, to foster a sense of unity and national belonging and a genetic notion of citizenship.

While one version of the Book of Icelanders was produced for deCODE genetics, another version has been available since 2003, free of charge, on the Internet (http://www.islendingabok.is). This version, of course, has no encryption. It enables 'Icelanders' to trace their family histories and to explore their genealogical connection to almost any Icelander living or dead. The responses to the genealogies on the web were overwhelming. Overnight, the Book of Icelanders became a popular pastime. In a few weeks, 100 000 Icelanders, about one-third of the total population, requested a password, exploring their relations with neighbors, colleagues, and friends. Soon the database became a party game. Recently, the Book of Icelanders was linked to Facebook, illustrating some of the connections between genomics, genealogies, and network society.

Increasingly, family histories are explored and celebrated on the web by the curious and enthusiastic public. Thus, in the United Kingdom several websites – including Ancestry.co.uk,

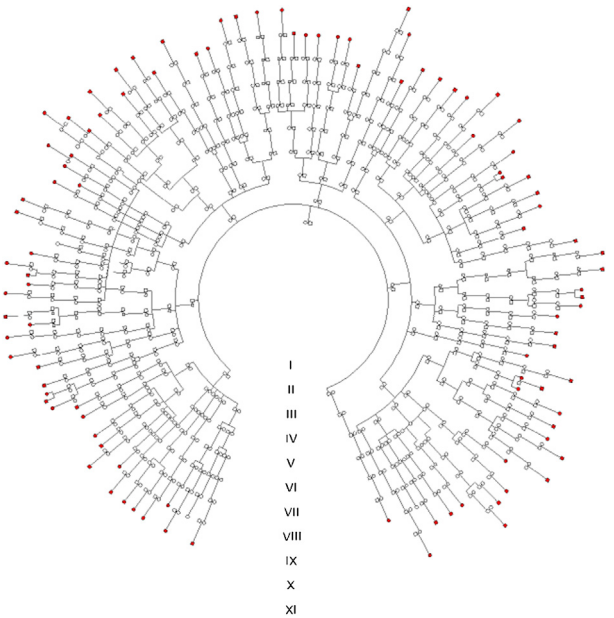

Figure 1 An asthma-pedigree with 11 generations. Source: deCODE genetics.

Genes Reunited, 1837online.com, and Achievements.co.uk – offer a range of genealogical services, either free of charge or for a certain fee. Some allow for the making and storing of family albums with details of family events, photos, and videos; others help users to create family trees; and still others provide access to records of births, baptism, deaths, and marriages. TV programs such as Channel 4's 'Extraordinary Ancestors' and BBC2's 'Blood Ties' testify to an extensive interest among the general public. In the United States too, there are several powerful genealogical websites, among them MyFamily.com and FamilySearch.org. The latter is run by the Church of Jesus Christ of Latter-Day Saints in Utah, drawing upon the extraordinary genealogical sources assembled by the Mormons over the years.

Many of the genealogical projects on the web are a by-product of the new genetics, combining in one fashion or another genetic and genealogical research (Tutton, 2004). The Huntsman Cancer Institute of the University of Utah has assembled life and health histories for 1.6 million Utahans, living or dead, into the Utah Population Database. 'Mormon genes are hot,' as a front-page article in the *New York Times* had it (Johnson, 2004), mainly because of the wealth of family histories but also because of the practice of polygamy in the past that has amplified the genetic imprint of a relatively small number of males. Another Utah-based project is the Sorenson Molecular Genealogy Foundation. Appealing to volunteers from all over the world to participate in its project, it combines Y chromosome, mtDNA, and autosomal genetic markers with established genealogies. Thousands of people from different countries have participated in the project, providing genealogical information and donating a DNA sample taken through mouthwash. The sole purpose of this enterprise is to identify possible genealogical links on the basis of 'molecular genealogy.'

The extent to which genomic evidence informs politics of belonging and exclusion and notions of ancestry largely remains an open empirical question. This is an important avenue for social scientific and humanities scholarship. Prainsack and Hashiloni-Dolev (2009) explore several cases in which genomic experts have tried to use genetics to 'verify' or 'falsify' particular group identities, including the case of genetic distance between Arabs and Jews and the 'Jewish' ancestry of the Lemba of South Africa. It turns out that the picture is more complicated than current discourses on genetic citizenship would lead one to believe and that DNA evidence need not have significant political implications.

The Future

To conclude, how will national projects and the sciences of human variation, in particular genomics and anthropology, be affected by scientific advances, conceptual change, and social developments? In her early historical account of the instruments of physical anthropology, Hoyme (1953: p. 425) had no doubts about the bright future ahead; speculating about the next 50 years, she suggested that it is "fairly certain … that anthropometry, the hallmark of physical anthropology, will be retained among the techniques of the anthropologist, although techniques from other biological sciences will undoubtedly be adopted where needed… There are many problems involving size and proportion not yet solved, and perhaps not yet suggested by genetics, physiology, ecology, and comparative anatomy, for which anthropometry can provide basic data." Yet, modern readers cannot help but wonder why people bothered with all the measurements and indices produced. Hoyme (1953: p. 423) admitted that it "is somewhat … difficult to draw conclusions as to why these measurements and indices were considered worthy of investigation," adding that "it is necessary to examine the studies against the climate of opinion of the period in which they were carried out." One of their main rationales was their role in sustaining national projects, establishing and maintaining identities and boundaries.

Just as the anthropometry of past centuries was moved from the center of national projects and studies of human variation, the focus on human genomics of recent decades is bound to change. While genomic avenues into history, difference, and relatedness have proved quite powerful, it seems likely that they, too, will be redirected. Already, a series of developments can be observed on the horizon, among them the escalation of computing power with advances in bioinformatics and computer technology facilitating complete genome sequencing at low price, the growth of personal genomics and digital networking, the federation of private and national biobanks, the growing evidence of microbiomes and epigenetic processes, and new forms of biopolitics and governance. The last two developments mentioned are probably of particular interest for shaping the future of genomics as national projects.

The discovery of the double helix by the middle of the twentieth century fostered the notion of genes as the 'secret of life.' The theoretical and analytical scope, however, has now been significantly broadened. Not only has the failure to make significant medical predictions on the basis of single genes increasingly drawn attention to multigenic studies assuming complex interaction and articulation, but also the stability of the genome has been seriously questioned. While the genome has largely been seen as 'an ensemble of genes strung along the chromosomes' (Barnes and Dupré, 2008: 76) with identical copies in every cell, it now seems that the dogma that all the cells of an individual are identical in terms of DNA needs rethinking. If it turns out that human genomes are unstable and variable assemblies, it becomes increasingly difficult to sustain claims about authentic or 'real' genomes, whether individual or national.

The fact that what is normally considered the 'human' genome is fundamentally mixed with the microbiomes of other organisms further complicates the picture. The human body, it turns out, is a kind of supraorganism carrying with it trillions of microbes constituting about 99% of its genes. The spokespersons for the so-called Human Microbiome Project (Turnbaugh et al., 2007) anticipate the establishment of microbial observatories worldwide for the purpose of mapping diversity and for linking microbiomes to the planetary environment. The evidence already accumulated necessarily destabilizes the issue of human variation and national genomes. If recent excursions into the human genome have been fixated on the 'wrong' genes, or at least a tiny set, and humans are best seen as supraorganisms, will 'human' genomics and genomic anthropology, in the traditional sense, be relegated to history as

trivial, narrow-minded exercises? What could possibly be 'national' in mapping the genomes of trillions of bugs in our guts and exploring their implications for our health – and who would bother to engage in their study?

In fact, this is a thriving field already. Alternatively one might ask, assuming growing public awareness of the role of microorganisms for health and development, will human genomics 'simply' become more complicated, casting its nets much wider than before and vastly expanding its range? The latter scenario, it seems, would call for enhanced governance concerns, emphasizing contexts of both community and environment. It seems likely, too, that casting the genomic net more widely to incorporate the vast terrain of microbiomes is bound to radically change current notions of human variation. The tentative results of this trend, sometimes called the 'microbiome revolution,' seem to indicate a clustering of human supraorganisms that does not correlate with classical categorizations based on 'race,' color, morphology, and continental background. To what extent and how they will affect national projects and imagined communities remain to be seen.

Moving beyond genes and genomes, organisms are partly regulated through a host of environmental forces that leave their mark on their genomes. The results of a number of recent studies suggest that for humans, early environments condition the possibility for gene expression, potentially affecting health and well-being. Such 'epigenetic' regulation seems to be prevalent in the human genome. Epigenetic 'imprints' are not changes in genetic sequence, but their effects are nevertheless 'memorized' and passed on as inheritance (Landecker, 2011). To account for the growing awareness of nongenetic factors, new terms have been entering biology and related fields, such as epigenome, methylome, and interacteome. While the notion of 'epigenetics' is used in different fashions among biologists, social scientists, and humanities scholars, this should not confuse here. 'Biology,' in any case, is a far more fleeting and complex phenomenon than normally imagined. The notion of 'national genomes' begins to look incredibly unstable and messy, if not meaningless. What bodily signatures, if any, one may ask, will replace those of genomes in the national projects of the future?

One driving force behind the growing governance discourse of the late twentieth century was the advancement in many contexts of the 'new economy' and free-market, neoliberal movements and governments, highlighted by the theorizing of Friedrich Hayek and the politics of Ronald Reagan and Margaret Thatcher. This was the discourse that partly drove the development of national biobanks and related developments, many of which were the private initiatives of multinational start-up companies. Much debate on bioethics and biopolitics focused on how to regulate the activities of biotech companies and the free-market through national machinery. In general, the rapidly expanding discourse on bioethics tended to apply highly individualistic notions of privacy and autonomy, including the notion of informed consent. Much indicates that this discourse is undergoing significant change. In some instances, research institutions utilizing a person's DNA are seen as custodians, stewards, or trustees (Winickoff and Winickoff, 2003) of genetic information, underlying the language of 'benefit sharing' and the 'Common Heritage' of international law. Indeed, governance models alternative to those of the individual, the market, and the state seem to be needed for dealing with the embedded person or body. A recent report on 'Solidarity' (Prainsack and Buyx, 2011) published by the Nuffield Council on Bioethics is one significant step forward in the direction of collective governance, outlining the history of the concept of solidarity and related terms, their growing salience in political theory and public discourse, and their usefulness in the context of concerns with pandemics, lifestyle-related diseases, and national biobanks.

See also: Bioethics: Philosophical Aspects; Community Genetics; Cultural Studies of Science; Direct-to-Consumer Personal Genetic Testing; Epigenetic Inheritance; Eugenics as an International Movement; Evolutionary Medicine; Genetics and Anthropology; Genetics and Forensics; Genetics and Indigenous Communities: Ethical Issues; Genetics and Society; Genetics: The New Genetics; Genotype and Phenotype; Human Genome Diversity Project: History; New Genetics and Race; Race: Genetic Aspects.

Bibliography

Anderson, Benedict, 1983. Imagined Communities. Verso, London.
Barnes, Barry, Dupré, John, 2008. Genomes and What to Make of Them. University of Chicago Press, Chicago.
Bibeau, Gilles, 2004. Le Québec Transgénique: Science, Marché, Humanité. Boréal, Québec.
Cambon-Thomsen, A., 2004. The social and ethical issues of post-genomic human biobanks. Nature 5, 866–873.
Cavalli-Sforza, L. Luca, 2005. The Human Genome Diversity Project: past, present and future. Nature Reviews / Genetics 6 (April), 333–340.
Connerton, Paul, 1989. How Societies Remember. Cambridge University Press, Cambridge.
Gottweis, Herbert, Peterson, Alan (Eds.), 2008. Biobanks: Governance in Comparative Perspective. Routledge, London.
Hacking, Ian, 1982. Biopower and the avalanche of printed numbers. Humanities in Society 5, 279–295.
Hinterberger, Amy, 2012. Publics and populations: the politics of ancestry and exchange in genome science. Science as Culture 21 (4), 528–549.
Hooton, Earnest Albert, 1946, 1931. Up from the Ape. The Macmillan Company, New York.
Hoyme, Lucile E, 1953. Physical anthropology and its instruments: an historical study. Southwestern Journal of Anthropology 9 (4), 408–430.
Johnson, Kirk, July 31 2004. By accident, Utah is proving an ideal genetic laboratory. The New York Times.
Landecker, Hannah, 2011. Food as exposure: nutritional epigenetics and the new metabolism. BioSocieties 6 (2), 167–194.
Lindee, Susan, Santos, Ricardo Ventura, 2012. The biological anthropology of living human populations: world histories, national styles, and international networks. Current Anthropology 53 (S5), 1–14.
Marks, Jonathan, 2008. Race across the physical-cultural divide in American anthropology. In: Kuklick, Henrika (Ed.), A New History of Anthropology. Blackwell, Oxford, UK, pp. 242–258.
Palsson, Gisli, 2007. Anthropology and the New Genetics. Cambridge University Press, Cambridge and New York.
Palsson, Gisli, 2008. Genomic anthropology: coming in from the cold? Current Anthropology 49 (4), 545–568.
Palsson, Gisli, 2012. Decode me! Anthropology and personal genomics. Current Anthropology 53 (S5), 185–195.
Prainsack, Barbara, Buyx, Alena, 2011. Solidarity: Reflections on an Emerging Concept in Bioethics. Nuffield Council on Bioethics, London.
Prainsack, Barbara, Hashiloni-Dolev, Yael, et al., 2009. Religion and nationhood: collective identities and the new genetics. In: Atkinson, Paul, et al. (Eds.), Handbook of Genetics and Society: Mapping the New Genomic Era. Routledge, London, pp. 404–421.

Rapp, Rayna, Heath, Deborah, Taussig, Karen-Sue, 2001. Genealogical dis-ease: where hereditary abnormality, biomedical explanation, and family responsibility meet. In: Franklin, Sarah, McKinnon, Susan (Eds.), Relative Values: Reconfiguring Kinship Studies. Duke University Press, Durham, NC, pp. 384–409.

Schiller, Glick Nina, 2005. Blood and belonging: long-distance nationalism and the world beyond. In: McKinnon, Susan, Silverman., Sydel (Eds.), Complexities: Beyond Nature & Nurture. University of Chicago Press, Chicago.

Scott, G. Richard, Scott, Legge, Lane, Robert W., Steen, Susan L., Street, Steven R., 2000. Physical anthropology of the Arctic. In: Nuttall, Mark, Callaghan, Terry V. (Eds.), The Arctic: Environment, People, Policy. Harwood Academic Publishers, Amsterdam, pp. 339–373.

Sommer, Marianne, 2008. History in the gene: negotiations between molecular and organismal anthropology. Journal of the History of Biology 41, 473–528.

Turnbaugh, Peter J., Ley, Ruth E., Hamady, Micah, Fraser-Liggett, Claire M., Knight, Rob, Gordon, Jeffrey I., 2007. The human microbiome project. Nature 449, 804–810.

Tutton, Richard, 2004. They want to know where they came from: population genetics and family genealogy. New Genetics and Society 23 (1), 105–120.

Tutton, Richard, Corrigan, Oonagh (Eds.), 2004. Genetic Databases: Socio-Ethical Issues in the Collection and Use of DNA. Routledge, London.

Verdery, Catherine, 1999. The Political Lives of Dead Bodies: Reburial and Postsocialist Change. Columbia University Press, New York.

Winickoff, David E, 2006. Genome and nation: Iceland's Health Sector Database and its legacy. Innovations: Technology, Governance, Globalization 1 (2), 80–105.

Winickoff, David E., Winickoff, Richard N., 2003. The charitable trust as a model for genomic biobanks. New England Journal of Medicine 349 (12), 1180–1184.

Anthropology, History of

Thomas H Eriksen, University of Oslo, Oslo, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Anthropology, the study of humanity seen from the perspective of social and cultural diversity, was established as an academic discipline in the mid-nineteenth century. At the time, broad, evolutionist perspectives were predominant, but would be eclipsed in the early twentieth century by the cultural relativism introduced by Franz Boas and the fieldwork revolution championed by Bronislaw Malinowski. During the twentieth century, anthropology became a diverse pursuit ranging from intensive functionalist studies of small-scale societies to ambitious modeling of the human mind, interpretive approaches and ecological accounts. Toward the end of the century, Marxist, feminist, and postcolonial critiques led to a heightened sensitivity toward methodological and ethical issues, and a major preoccupation in the early twenty-first century has consisted in accounting for the impact of globalization on local communities. In this intensely interconnected world, where formerly influential theoretical paradigms have collapsed, anthropology remains committed to comparison and ethnographic fieldwork, applying methods initially developed to study tribal society to research on complex societies.

Early attempts at accounting for cultural variation and human universals, the subject matter of anthropology, include ancient Greek geography and history, Renaissance and early modern travelogues, and philosophical treatises from the European Enlightenment. They fell short of being scientific because of inherent bias, lack of theoretical sophistication, and, frequently, poor empirical material. In the nineteenth century, theorists such as Henry Lewis Morgan, Henry Maine, and Edward Tylor developed ambitious evolutionist theories of culture, implying a hierarchical view of cultures. Modern anthropology emerged in the early decades of the twentieth century, when a new generation of scholars, notably Franz Boas, A.R. Radcliffe-Brown, Bronislaw Malinowski, and Marcel Mauss, rejected evolutionism replacing it with cultural relativism, also refining research methodologies and methods of comparison. In the decades following World War II, new approaches such as structuralism, cultural ecology, and neo-Marxism appeared, and anthropology grew and diversified, branching into various specializations including psychological and medical anthropology, eventually moving increasingly toward the study of modern phenomena such as nationalism and consumption, yet retaining its commitment to the intensive, fieldwork-based study of social life in local settings.

Social or cultural anthropology can be defined, loosely and broadly, as the comparative science of culture and society, and it is the only major discipline in the social sciences that has concentrated most of its attention on non-Western people. Although many of the classic problems investigated by anthropologists are familiar to the European history of ideas, the subject as it is known today emerged only in the early twentieth century, became institutionalized at universities in the Western world in midcentury, and underwent a phenomenal growth and diversification in the latter half of the century.

Foundations and Early Schools

Proto-Anthropology

Interest in cultural variation and human universals can be found as far back in history as the Greek city-state. The historian Herodotos (fifth century BC) wrote accounts of 'barbarian' peoples to the east and north of the peninsula, comparing their customs and beliefs to those of Athens, and the group of philosophers known as the Sophists were perhaps the first philosophical relativists, arguing (as many twentieth-century anthropologists later did) that there can be no absolute truth because, as one would put it today, truth is contextual. Yet their interest in cultural variation fell short of being scientific, chiefly because Herodotos lacked theory while the Sophists lacked empirical material.

Centuries later, scholarly interest in cultural variation and human nature reemerged in Europe because of the new intellectual freedom of the Renaissance and questions arising from European overseas exploits. Michel de Montaigne (sixteenth century), Thomas Hobbes (seventeenth century), and Giambattista Vico (eighteenth century) were among the thinkers of the early modern era who tried to account for cultural variability and global cultural history as well as dealing with the challenge from relativism. Eighteenth-century philosophers such as Locke, Hume, Kant, Montesquieu, and Rousseau developed theories of human nature, moral philosophies, and social theories, taking into account an awareness of cultural differences. The early German romantic Herder challenged Voltaire's universalistic vision by arguing that each people (*Volk*) had a right to retaining its own, unique values and customs – in a manner reminiscent of later cultural relativism romanticism. Indeed, by the end of the eighteenth century, several of the general questions still raised by anthropologists had already been raised: Universalism vs relativism (what is common to humanity; what is culturally specific), ethnocentrism vs cultural relativism (moral judgments vs neutral descriptions of other peoples), and humanity vs the rest of the animal kingdom (culture vs nature). Twentieth-century anthropology has taught that these and other essentially philosophical problems are best investigated through the detailed study of living people in existing societies through ethnographic fieldwork, and by applying carefully devised methods of comparison to the bewildering variety of 'customs and beliefs.' It would take several generations after Montesquieu's comparative musings about Persia and France,

in his *Lettres persanes*, before anthropology achieved this mark of scientific endeavor.

Victorian Anthropology

The first general theories of cultural variation to enjoy a lasting influence were arguably those of two men trained as lawyers: Henry Maine (1822–88) in Britain and Lewis Henry Morgan (1818–82) in the United States. Both presented evolutionist models of variation and change, where Western European societies were seen as the pinnacle of human development. In his *Ancient Law* (1861), Maine distinguished between *status* and *contract* societies, a divide which corresponds roughly to later dichotomies between traditional and modern societies, or, in the late nineteenth-century German sociologist Ferdinand Tönnies' terminology, *Gemeinschaft* (community) and *Gesellschaft* (society); status societies are assumed to operate on the basis of kinship and myth, while individual merit and achievement are decisive in contract societies. Although simple contrasts of this kind have regularly been severely criticized, they continue to exert a certain influence on anthropological thinking.

Morgan's contributions to anthropology were wide ranging and, among other things, he wrote a detailed ethnography of the Iroquois. His evolutionary scheme, presented in *Ancient Society* (1877), which influenced Marx and Engels, distinguished between seven stages (from lower savagery to civilization). His materialist account of cultural change influenced Marx and Engels. His pioneering work on kinship divided kinship systems into a limited number of types, and saw kinship terminology as a key to understanding society. Writing in the same period, the historian of religion Robertson Smith and the lawyer J.J. Bachofen offered, respectively, theories of monotheistic religion and of the (wrongly assumed) historical transition from matriliny to patriliny.

An untypical scholar in the otherwise evolutionist Victorian era, the German ethnologist Adolf Bastian (1826–1905) reacted against simplistic typological schemata. Drawing inspiration from Herderian romanticism and the humanistic tradition in German academia, Bastian wrote prolifically on cultural history, avoiding unwarranted generalizations, yet he held that all humans have the same pattern of thinking, thus anticipating structuralism.

The leading British anthropologist of the late Victorian era was Edward Tylor (1832–1917), whose writings include a famous definition of culture (dating from 1871), seeing it as the sum total of collective human achievements (thus contrasting it to nature). Tylor's student James Frazer (1854–1941) published the massive and very influential *Golden Bough* (1890, rev. ed. 1911–15), an ambitious comparative study of myth and religion.

Intellectual developments outside anthropology in the second half of the nineteenth century also had a powerful impact. Charles Darwin's theory of natural selection, first presented in his *Origin of Species* from 1859, would both be seen as a condition for anthropology (positing that all humans are closely related) and, later, as a threat to it (arguing the primacy of the biological over the cultural). The emergence of classic sociological theory in the works of Comte, Marx, and Tönnies, and later Durkheim, Weber, Pareto, and Simmel, provided anthropologists with general theories of society, although their applicability to non-European societies continues to be disputed (*see* Sociology, History of).

The quality of the data used by the early anthropologists was variable. Most of them relied on written sources, ranging from missionaries' accounts to travelogues of varying accuracy. The need for more reliable data began to make itself felt. Expeditions and systematic surveys now provided researchers around the turn of the twentieth century with improved knowledge of cultural variation, which eventually led to the downfall of the ambitious theories of unilineal evolution characteristic of nineteenth-century anthropology.

An Austro-German specialty proposed both as an alternative and a complement to evolutionist thinking was diffusionism, the doctrine of the historical diffusion of cultural traits. Never a part of the mainstream outside of the German-speaking world, elaborate theories of cultural diffusion continued to thrive, particularly in Berlin and Vienna, until after World War II (Gingrich, 2005). As there were serious problems of verification associated with the theory, it was condemned as speculative by anthropologists committed to fieldwork and, furthermore, research priorities were to shift from general cultural history to intensive studies of particular societies.

In spite of theoretical developments and methodological refinements, the emergence of anthropology, as the discipline is known today, is rightly associated with four scholars working in three countries in the early decades of the twentieth century: Franz Boas in the United States, A.R. Radcliffe-Brown and Bronislaw Malinowski in the United Kingdom, and Marcel Mauss in France.

Boas and Cultural Relativism

Boas (1858–1942), a German migrant to the United States, who had briefly studied anthropology with Bastian, carried out research among Eskimos and Kwakiutl Indians in the 1890s. In his teaching and professional leadership, he strengthened the 'four-field approach' in American anthropology, which still sets it apart from European anthropology, including both cultural and social anthropology, physical anthropology, archaeology, and linguistics. Although cultural relativism had been introduced more than a century before, it was Boas who made it a central premise for anthropological research. Against the evolutionists, he argued that each culture had to be understood in its own terms and that it would be scientifically misleading to rank other cultures according to a Western, ethnocentric typology gauging 'levels of development.' Boas also promoted *historical particularism*, the view that all societies or cultures had a unique history that could not be reduced to a category in some universalistic scheme. On related grounds, Boas argued incessantly against the claims of *racist* views of human diversity.

Perhaps because of his particularism, Boas never systematized his ideas in a theoretical treatise. Several of his students and associates nevertheless did develop general theories of culture, notably Ruth Benedict, Alfred Kroeber, and Robert Lowie. His most famous student was Margaret Mead (1901–78). Although her best-selling books on Pacific societies have been criticized for being superficial, she used material from non-Western societies to raise questions about gender

relations, socialization, and politics in the West, and Mead's work indicates the potential of cultural criticism inherent in the discipline.

One of Boas' most remarkable associates, the linguist Edward Sapir (1884–1939), formulated, with his student Benjamin Lee Whorf, the Sapir–Whorf hypothesis, which posits that language determines cognition. Consistent with a radical cultural relativism, the hypothesis implies that, for example, the Hopi perceive the world in a fundamentally different way from Westerners, due to differences in the structure of their respective languages. A controversial perspective from the beginning, the hypothesis has nonetheless continued to be discussed in linguistics and linguistic anthropology.

The Two British Schools

While modern American anthropology had been shaped by the Boasians and their relativist concerns, as well as the perceived need to record native cultures before their anticipated disappearance, the situation in the major colonial power, Great Britain, was different. The degree of complicity between colonial agencies and anthropologists is debatable, but the very fact of imperialism was an inescapable premise for British anthropology until after World War II.

The man who is often hailed as the founder of modern British social anthropology was a Polish immigrant, Bronislaw Malinowski (1884–1942), whose fieldwork for over 2 years in the Trobriand Islands (1914–18) set a standard for ethnographic data collection that is still largely unchallenged. Malinowski argued the need to learn the local language properly and to engage in everyday life, in order to see the world from the actor's point of view and to understand the interconnections between social institutions and cultural notions. Malinowski placed an unusual emphasis on the acting individual, seeing social structure not as a determinant of but as a framework for action, and he wrote about a wide range of topics, from garden magic, economics, and sex to the puzzling *kula* trade. Although he dealt with issues of general concern, he nearly always took his point of departure in his Trobriand ethnography, demonstrating a method of generalization very different from that of the previous generation with its scant local knowledge.

The other major British anthropologist of the time was A.R. Radcliffe-Brown (1881–1955). An admirer of Durkheim's sociology, Radcliffe-Brown did relatively little fieldwork himself, but aimed at the development of a 'natural science of society' where universal laws of social life could be formulated. His theory, known as structural functionalism, saw the individual as unimportant, emphasizing instead the social institutions (kinship, norms, politics, etc.). Most social and cultural phenomena were seen as functional in the sense that they contributed to the maintenance of the overall social structure (*see* Functionalism in Anthropology).

Despite their differences in emphasis, both British schools had a sociological concern in common (which they did not share with most Americans), and tended to see social institutions as functionally integrative. Both rejected the wide-ranging claims of diffusionism and evolutionism, and yet, the tension between structural explanations and actor-centered accounts remains strong in British anthropology even today.

Malinowski's students included Raymond Firth, Audrey Richards, and Isaac Schapera, while Radcliffe-Brown, in addition to enlisting E.E. Evans-Pritchard and Meyer Fortes – arguably the most powerful British anthropologists in the 1950s – on his side, taught widely, and introduced structural functionalism to several foreign universities. British interwar anthropology was characteristically oriented toward kinship, politics, and economics, with Evans-Pritchard's masterpiece *The Nuer* (1940) demonstrating the intellectual power of a discipline combining detailed ethnography, comparison, and elegant models. Later, his models would be criticized for being too elegant to fit the facts on the ground – a very Malinowskian objection.

Mauss

No fieldwork-based anthropology developed in the German-speaking region, and German anthropology was marginalized after World War II. In France, the situation was different. Already in 1902, Durkheim had published, with his nephew Marcel Mauss (1872–1950), an important treatise on primitive classification; in 1908, Arnold van Gennep published *Les Rites de Passage*, an important analysis of initiation rites, and Lucien Lévy-Bruhl elucidated a theory, later refuted by both Evans-Pritchard, Mauss, and others, on the 'primitive mind,' which he held to be 'prelogical.' New empirical material of high quality was being produced by thorough observers such as Maurice Leenhardt in New Caledonia and Marcel Griaule in West Africa.

Less methodologically purist than the emerging British traditions and more philosophically adventurous than the Americans, interwar French anthropology, under the leadership of Mauss, developed a distinct flavor, witnessed in the influential journal *L'Année Sociologique*, founded by Durkheim and edited by Mauss after Durkheim's death in 1917. Drawing on his vast knowledge of languages, cultural history, and ethnography, Mauss, who never did fieldwork himself, wrote several learned, original, compact essays ranging from gift exchange (*Essai sur le Don*, 1924) to the nation, the body, and the concept of the person. Mauss' theoretical position was complex. He believed in systematic comparison and the existence of recurrent patterns in social life at all times and in all places, yet he often defended relativist views in his reasoning about similarities and differences between societies.

Not a prolific writer, Mauss exerted an enormous influence on later French anthropology through his teaching. Among his students and associates were most of the major French anthropologists at the time, and the three leading postwar scholars in the field – Louis Dumont, Claude Lévi-Strauss, and Georges Balandier – were all deeply indebted to Mauss.

Some General Points

The transition from evolutionist theory and grand syntheses to more specific, detailed, and empirically founded work, in reality amounted to an intellectual revolution. The work of Tylor and Morgan had been relegated to the mists of history by the mid-twentieth century, and the discipline had been taken over by small groups of scholars who saw intensive fieldwork, cultural relativism, the study of single societies, and rigorous

comparison as its essence. Today, the academic institutions, the conferences, and the learned journals all build on the anthropology of Boas, Malinowski, Radcliffe-Brown, and Mauss. This is to a great extent also true of the anthropological traditions of other countries (Vermeulen and Roldán, 1995), including India, Australia, Mexico, Argentina, the Netherlands, Spain, and Scandinavia. Soviet/Russian and East European anthropologies have followed different itineraries, retaining a connection with the German tradition, where *Volkskunde* (ethnology, studies of rural folk culture in one's own society) was distinguished from *Völkerkunde* (literally the study of peoples, that is other cultures). In Russia, there had existed a rich tradition of empirically based anthropological research since the early nineteenth century, and twentieth century Soviet anthropology combined *Völkerkunde* and Marxist influence in original ways (see Eriksen and Nielsen, 2012), but there was little contact with the Western mainstream until the end of Communism (but see Gellner, 1980).

Anthropology in the Second Half of the Twentieth Century

The numbers of anthropologists and institutions devoted to teaching and research in the field grew rapidly after World War II. The discipline diversified. New specializations such as psychological anthropology, political anthropology, and the anthropology of ritual emerged, and the geographical foci of the discipline multiplied: whereas the Pacific had been the most fertile area for theoretical developments in the 1920s, and Africa had played a similar part in the 1930s and 1940s, while the preoccupation with North American Indians had been stable throughout, the 1950s saw a growing interest in the 'hybrid' societies of Latin America and the Caribbean as well as the anthropology of India and South-East Asia, and the New Guinean highlands became similarly important in the 1960s. Such shifts in spatial emphasis had consequences for theoretical developments, as each region posed its own peculiar problems.

From the 1950s, the end of colonialism also affected anthropology, both in a banal sense – it became more difficult to obtain research permits – and more profoundly, as the subject–object relationship between the observer and the observed became problematic as the traditionally 'observed' peoples increasingly had their own intellectuals and spokespersons who frequently objected to Western interpretations of their way of life.

Structuralism

The first major theory to emerge after World War II was Claude Lévi-Strauss' structuralism. Lévi-Strauss (1908–2009) developed an original theory of the human mind, drawing on structural linguistics, Mauss' theory of exchange, and Lévy-Bruhl's theory of the primitive mind (which Lévi-Strauss opposed). His first major work, *Les Structures Elémentaires de la Parenté* (*The Elementary Structures of Kinship*, 1949), introduced a grammatical, formal way of thinking about kinship, with particular reference to systems of marriage (the exchange of women between groups). Lévi-Strauss later expanded his theory to cover totemism, myth, and art. Never uncontroversial, structuralism had an enormous impact on French intellectual life far beyond the confines of anthropology. In the English-speaking world, the reception of structuralism was delayed, as Lévi-Strauss' major works were not translated until the 1960s, but he had his admirers and detractors from the beginning. Structuralism was criticized for being untestable, positing as it did certain unprovable and unfalsifiable properties of the human mind (most famously the propensity to think in terms of contrasts or binary oppositions), but many saw Lévi-Strauss' work, ultimately committed to human universals, as an immense source of inspiration in the study of symbolic systems such as knowledge and myth.

A different, and for a long time less influential, brand of structuralism was developed by Louis Dumont (1911–99), an Indianist and Sanskrit scholar who did fieldwork both in the Aryan north and the Dravidian south. Dumont, closer to Durkheim than Lévi-Strauss, argued in his major work on the Indian caste system, *Homo Hierarchicus* (1969), for a holistic perspective (as opposed to an individualistic one), claiming that Indians (and by extension, many nonmodern peoples) saw themselves not as 'free individuals' but as actors irretrievably enmeshed in a web of commitments and social relations, which in the Indian case was clearly hierarchical.

Most later, major French anthropologists have been associated with Lévi-Strauss, Dumont, or Balandier, the Africanist whose work in political anthropology simultaneously bridged gaps between France and the Anglo-Saxon world and inspired both neo-Marxist research and applied anthropology devoted to development.

Reactions to Structural Functionalism

In Britain and her colonies, the structural functionalism now associated chiefly with Evans-Pritchard and Fortes was under pressure after the war. Indeed, Evans-Pritchard himself repudiated his former views in 1949, arguing that the search for 'natural laws of society' had been shown to be futile and that anthropology should fashion itself as a humanities discipline rather than a natural science. Retrospectively, this statement has often been quoted as marking a shift 'from function to meaning' in the discipline's priorities; Kroeber expressed similar views in the United States. Others found other paths away from what was increasingly seen as a conceptual straitjacket. Edmund R. Leach, whose *Political Systems of Highland Burma* (1954), suggested a departure from functionalist orthodoxies, notably Radcliffe-Brown's dictum that social systems tend to be in equilibrium and Malinowski's view of myths as integrating 'social charters,' would later be a promoter and critic of structuralism in Britain. Leach's contemporary Raymond Firth proposed a distinction between social structure (the statuses in society) and social organization, which he saw as the actual process of social life, whereby choice and individual whims were related to structural constraints. Later in the 1950s and 1960s, several younger social anthropologists, notably F.G. Bailey and Fredrik Barth, followed Firth's lead as well as the theory of games (a recent development in economics) in refining an actor-centered perspective on social life. Max Gluckman, a former student of Radcliffe-Brown and a close associate of Evans-Pritchard, also abandoned the strong

holist program of the structural functionalists, reconceptualizing social structure as a loose set of constraints, while emphasizing the importance of individual actors. Gluckman's colleagues included several important Africanists such as A.L. Epstein, J. Clyde Mitchell, and Elizabeth Colson. Working in southern Africa, this group pioneered urban anthropology and ethnicity studies in the 1950s and 1960s.

Neo-Evolutionism, Cultural Ecology, and Neo-Marxism

The number of anthropologists has always been larger in the United States than anywhere else, and the discipline was always diverse there. Although the influence from the Boasian cultural relativist school remains strong, other groups of scholars have also made their mark. From the late 1940s, a resurgent interest in Morgan's evolutionism led to the formulation of neo-evolutionist and materialist research programs. Julian Steward, a student of Robert Redfield (who had been a student of Radcliffe-Brown), proposed a theory of cultural dynamics distinguishing between 'the cultural core' (basic institutions such as the division of labor) and 'the rest of culture' in a way strongly reminiscent of Marx. Steward led research projects among Latin American peasants and North American Indians, encouraging a focus on the relationship between culture, technology, and the environment. Leslie White held more deterministic materialist views, but also – perhaps oddly – saw symbolic culture as a largely autonomous realm. Among the major scholars influenced by White, Marvin Harris has strengthened his materialist determinism, while Marshall D. Sahlins in the 1960s made the move from neoevolutionism to a symbolic anthropology influenced by structuralism.

Cultural ecology sprang from the teachings of Steward and White, and represented a rare collaboration between anthropology and biology. Especially in the 1960s, many such studies were carried out, including, notably, Roy Rappaport's *Pigs for the Ancestors* (1968), an attempt to account for a recurrent ritual in the New Guinean highlands in ecological terms. The upsurge of Marxist peasant research, especially in Latin America, in the 1970s, was also clearly indebted to Steward.

The appearance of radical student politics in the late 1960s, which had an impact on academia until the early 1980s, had a strong, if passing, influence on anthropology. Of the more lasting contributions, the peasant studies initiated by Steward and furthered by Eric Wolf, Sidney Mintz, and others must be mentioned, along with French attempts, represented in the very sophisticated work of Maurice Godelier and Claude Meillassoux, at synthesizing Lévi-Straussian structuralism, Althusserian Marxism, and anthropological comparison. Although Marxism and structuralism eventually became unfashionable, scholars – particularly those engaged in applied work – continue to draw inspiration from Marxist thought (*see* Field Theory).

Symbolic and Cognitive Anthropology

More true to the Boasian legacy than the materialist approaches, studies of cognition and symbolic systems developed and diversified in the United States. A leading theorist was Clifford Geertz (1926–2006), who wrote a string of influential essays advocating hermeneutics (interpretive method) in the 1960s and 1970s (*see* Hermeneutics, History of). While his originality as a theorist can be questioned, his originality as a writer is obvious, and Geertz ranks as perhaps the finest writer of contemporary anthropology. Marshall Sahlins is, with Geertz, the foremost proponent of cultural relativism today, and has published a number of important books on various subjects (from Mauss' theory of exchange to sociobiology and the death of Captain Cook), consistently stressing the autonomy of the symbolic realm, thus arguing that cultural variation cannot be explained by recourse to ecology, technology, or biology.

In Britain, too, interest in meaning, symbols, and cognition grew after the war, especially from the 1960s (partly due to the belated reception of Lévi-Strauss). British anthropology had hitherto been strongly sociological, and two important scholars who fused the legacy from structural functionalism with the study of symbols and meaning were Mary Douglas (1921–2007) and Victor Turner (1920–83). Taking his cue from van Gennep, Turner, a former associate of Gluckman, developed a complex analysis of rituals among the Ndembu of Zambia, showing their functionally integrating aspects, their meaningful aspects for the participants, and their deeper symbolic significance. Douglas, a student of Evans-Pritchard, famous for her *Purity and Danger* (1966), analyzed the human preoccupation with dirt and impurities as a way of thinking about the boundaries of society and the nature/culture divide. Prolific and original, Douglas is a major defender of a reformed structural functionalism.

Against all these (and other) perspectives regarding how 'cultures' or 'societies' perceive the world, anthropologists emphasizing the actor's point of view have argued that no two individuals see the world in the same way and that it is preposterous to generalize about societies. The impact of feminism has been decisive here. Since the 1970s, feminist anthropologists have identified often profound differences between male and female worldviews, indicating how classic accounts of 'societies' really refer to male perspectives on them as both the anthropologist and the main informants tended to be male. For example, in a restudy of Trobriand society, Annette Weiner (1976) showed that Malinowski's famous work was incomplete and ultimately misleading as he had failed to observe important social processes confined to females. Feminism also influenced anthropology by emphasizing the need for a critical reflexivity addressing the methodological and epistemological issues arising from the ethnographer's identity as, for example, a man or a woman.

Anthropology and the Contemporary World

Since the pillars of modern anthropology were erected around the World War I, the former colonies became independent, 'natives' got their own educated elites (including social scientists), economic and cultural globalization led to the spread of capitalism and consumer culture, and transcontinental migration blurred the boundaries between the traditional 'us' and 'them.' This situation entailed new challenges for anthropology that were met in various ways – revealing continuities as well as breaks with the past.

A late field to be incorporated into anthropology, but one that became the largest single area of interest from the 1970s, was the study of identity politics, notably ethnicity and nationalism. Since the publication of several important texts

around 1970 (by, inter alia, Barth and Abner Cohen in Europe, and George DeVos in the United States), anthropological ethnicity studies investigated the interrelationship between ethnic identity and ethnic politics, and explored how notions of cultural differences contribute to group identification. Since the publication of several important texts on nationalism in the early 1980s (by Ernest Gellner, Benedict Anderson, and others), this also became an important area for anthropologists. Ethnicity and nationalism are partly or wholly modern phenomena associated with the state, and thus denote a departure from the former mainstay of anthropology, the study of nonmodern small-scale societies. While ethnicity and (especially) nationalism could not be studied through participant observation only (other kinds of data are required), it was evident that anthropologists who engaged in this field remained committed to the classic tenets of the discipline; ethnographic fieldwork, comparison, and a systemic view of social reality. Also, the study of identity politics emerged as an interdisciplinary field where anthropologists, sociologists, historians, and political scientists profited from each other's expertise.

Other modern phenomena also received increased attention in anthropology from the 1970s, including consumption, 'subcultures,' wagework, and migration. The boundary between the 'Western self' and the 'non-Western other' became blurred. Anthropological studies of Western societies became common, and Europe was established as an ethnographic region along with West Africa, South Asia, and so on. Even anthropologists working in traditional settings with classic topics increasingly had to see their field as enmeshed, to a greater or lesser extent, in a global system of communication and exchange.

Because of the increased penetration of the formerly tribal world by capitalism and the state, and accompanying processes of cultural change, there was a growing demand for a reconceptualization of culture in the 1980s and 1990s, and scholars such as Ulf Hannerz and Marilyn Strathern developed fluid concepts of culture seeing it as less coherent, less bounded, and less integrated than the Boasian and Malinowskian traditions implied.

Some scholars saw the postcolonial situation as sounding the death knell of anthropology: Since the 'primitive' was gone, and the former informants were now able to identify and describe themselves (they no longer needed anthropologists to do it), the science of cultural variation seemed to have lost its *raison d'être*. Following the lead of Edward Said's *Orientalism* (1978), an influential critique of Western depictions of the 'East,' and often inspired by Michel Foucault, they saw anthropology as a colonial and imperialist enterprise refusing non-Western peoples a voice of their own and magnifying the distance between 'us' and 'them.' Especially in the late 1980s and early 1990s, this view had many followers, some of whom abandoned empirical research, while others tried to incorporate the autocriticism into their work. Yet others saw these pessimistic views as largely irrelevant, since anthropology had always been fraught with similar tensions, to which each new generation found its solutions. In this regard, it must be pointed out that the earlier feminist critique of anthropology, far from repudiating the subject, led to its enriching by adding new implements to its toolbox and new dimensions to its worldview. The same could be said of the reconceptualizations of culture that arguably offered an improved accuracy of description (*see* Postcoloniality).

The Situation at the Turn of the Millennium

Over the course of the twentieth century, anthropology became a varied discipline with a strong academic foothold in all continents, although its centers remained in the English- and French-speaking countries. It was still possible to discern differences between American cultural anthropology, British social anthropology, and French *ethnologie*, but the discipline was more unified than ever before – not in its views, but in its approaches. Hardly a part of the world had not now been studied intensively by scholars engaging in ethnographic fieldwork, but since the world changes, new research is always called for.

Specializations proliferated, ranging from studies of ethnomedicine and the body to urban consumer culture, advertising, and cyberspace. Although the grand theories of the nineteenth and twentieth centuries – from unilinear evolutionism to structuralism – had been abandoned, new theories claiming to provide a unified view of humanity were being proposed; for example, new advances in evolutionary psychology and cognitive science offered ambitious general accounts of social life and the human mind, respectively. The problems confronting earlier generations of anthropologists, regarding, for example, the nature of social organization, of knowledge, of kinship, and of myth and ritual, remained central to the discipline although they were explored in new empirical settings by scholars who were more specialized than their predecessors.

Anthropology has thrived on the tension between the particular and the universal; between the intensive study of local life and the quest for general accounts of the human condition. Is it chiefly a generalizing science or a discipline devoted to the elucidation of the unique? The general answer is that anthropologists ultimately do study society, culture, and humanity, but that in order to do so, they must devote most of their energies to the study of societies, cultures, and humans. As long as their mutual differences and similarities are not fully understood, there will be an intellectual space in the world for anthropology or, at least, a discipline like it (*see* Anthropology).

See also: Anthropology: Overview; Boas, Franz (1858–1942); Cognitive Anthropology; Colonialism, Anthropology of; Community and Society: History of the Concepts; Darwin, Charles Robert (1809–82); Human Behavioral Ecology; Structuralism; Symbolism in Anthropology.

Bibliography

Clifford, J., 1988. The Predicament of Culture: Twentieth-Century Ethnography, Literature and Art. Harvard University Press, Cambridge, MA.

Eriksen, T.H., Nielsen, F.S., 2012. A History of Anthropology, second ed. Pluto, London.

Gellner, E. (Ed.), 1980. Soviet and Western Anthropology. Duckworth, London.

Gingrich, A., 2005. The German-speaking countries. In: Barth, F., Gingrich, A., Parkin, R., Silverman, S. (Eds.), One Discipline, Four Ways: British, German, French, and American Anthropology. University of Chicago Press, Chicago, pp. 61–156.

Kuper, A., 1996. Anthropology and Anthropologists: The Modern British School, third ed. Routledge, London.
Lévi-Strauss, C., 1987 [1950]. Introduction to the Work of Marcel Mauss (F. Baker, Trans.). Routledge, London.
Moore, J.D., 1997. Visions of Culture: An Introduction to Anthropological Theories and Theorists. AltaMira Press, Walnut Creek, CA.
Stocking Jr., G.W., 1987. Victorian Anthropology. Free Press, New York.
Stocking Jr., G.W. (Ed.), 1996. Volksgeist as Method and Ethic: Essays on Boasian Ethnography and the German Anthropological Tradition. University of Wisconsin Press, Madison, WI.
Vermeulen, H.F., Roldán, A.A. (Eds.), 1995. Fieldwork and Footnotes: Studies in the History of European Anthropology. Routledge, London.
Weiler, B., 2005. Die Ordnung des Fortschritts: Zum Aufstieg und Fall der Fortschrittsidee in der 'jungen' Anthropologie. Transcript, Bielefeld.
Weiner, A., 1976. Women of Value, Men of Renown: New Perspectives on Trobriand Exchange. University of Texas Press, Austin, TX.

Anthropology: Overview

Ulf Hannerz, Stockholm University, Stockholm, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Anthropology is a study of human diversity. This overview article discusses global variations in the delimitation of the discipline, and its development from a study closely linked historically to the Western expansion into other parts of the world, to the current situation with anthropologists from varied nations studying social and cultural life everywhere. Key concepts – culture, cultural translation, comparison, and fieldwork – are identified; main dimensions of specializations and practical applications are discussed; and the future of anthropology in an increasingly interconnected world is considered.

Anthropology as a discipline is concerned with human diversity. In its most inclusive conception, this is what brings together the four fields of sociocultural anthropology, archaeology, biological (or physical) anthropology, and linguistics. With its formative period in the historical era, when Europeans and people of European descent were exploring other parts of the world and establishing their dominance over them, and when evolutionary thought was strong, it also came to focus its attention especially on what was, from the Western point of view, distant in time or space – early humans or hominoids, and non-European peoples. In that early period, there was indeed a strong tendency to conflate distances in time with those in space: some living non-Europeans could be taken to be 'contemporary ancestors,' dwelling in 'primitive societies.' Understandings of the discipline have changed over time, however; and they are not now entirely unitary across the world. What is held together under one academic umbrella in one place may be divided among half a dozen disciplines somewhere else. A mapping of the contemporary state of the discipline, consequently, may usefully begin by taking some note of international variations.

Terminologies and Boundaries

It is particularly in North America that academic anthropology has retained what has come to be known as 'the four-field approach.' Generally, it seems to have had its greatest strength in countries where dominant settler populations faced sizable indigenous populations, and may have found it practical to assemble academic knowledge about these in a single discipline. In recent times, the continued viability of this arrangement has come under debate. Among the founders of the discipline, some were perhaps able to work (or at least dabble) in all the main branches, but with time, in American anthropology as well, it has certainly been recognized that most scholars reach specialist skills in only one of them – even as it may be acknowledged that a broad intellectual sweep across humanity has its uses, and at the same time as it is recognized that, here as elsewhere, research in the border zones between established disciplines or subdisciplines often brings its own rewards. On the whole, in any case, scholars in archaeology, biological anthropology, linguistics, and sociocultural anthropology now mostly work quite autonomously of one another, and while terminologies vary, in many parts of the world, they are understood as separate disciplines.

In Europe, varying uses of the terms 'anthropology,' 'ethnology,' and 'ethnography,' between countries and regions as well as over time, often reflect significant historical and current intellectual divides (Vermeulen and Alvarez Roldán, 1995). In parts of the continent, in an earlier period, the term 'anthropology' (in whatever shape it appeared in different languages) tended to be used mostly for physical anthropology, but since the later decades of the twentieth century, it has largely been taken over by what we here term 'sociocultural anthropology' – itself a hybrid designation for what is usually referred to either as social anthropology or as cultural anthropology (in German usage especially, however, 'anthropology' also frequently refers to a branch of philosophy). Physical or biological anthropology, meanwhile, was absorbed in many places by other disciplines concerned with human biology, while archaeology and linguistics maintained their positions as separate disciplines.

In some European countries, now or in the past, the term 'ethnography' has been used, unlike in present-day usage in Anglophone countries, to refer to sociocultural anthropology as a discipline. Matters of discipline boundaries are further complicated, however, by the fact that sociocultural anthropology, especially in Northern, Central, and Eastern Europe, is often itself divided into two separate disciplines, with separate origins in the nineteenth and early twentieth centuries. One, which was often originally designated something like 'folk life studies,' had its historical links with cultural nationalism, and concerned itself with local and national traditions, especially with regard to folklore and material culture. This discipline mostly did not acquire a strong academic foothold in those Western European countries that were most involved in exploration and colonialism outside Europe, particularly Great Britain and France, where, on the other hand, sociocultural anthropology that focused on non-European forms of life was earliest and most securely established. The more Europe-oriented, or nationally inclined, 'folk life studies' discipline has tended, in recent decades, to redesignate itself as 'European ethnology' – or, in some contexts, simply as 'ethnology' – in contrast to a rather more globally oriented 'anthropology.' In another usage, 'ethnology' has been taken to refer to a more historical and museological orientation (in contrast with what was for a time a more presentist social anthropology), while in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12011-2

other contexts again, it is more or less synonymous with 'sociocultural anthropology.' Yet further national variations in terminology continue to make direct transpositions of terms between languages treacherous.

Furthermore, if in its historical beginnings anthropology tended to be a matter of 'the West studying the rest,' with the colonial powers and their settler offshoots taking the greatest interest in establishing the discipline as an academic field, this is no longer the situation. Since the latter half of the twentieth century, the number of practitioners has grown conspicuously, and their global distribution has changed. There are now anthropologists everywhere, locals of just about all countries, based in their home academic institutions or other organizations (Fahim, 1982; Ribeiro and Escobar, 2006; Boskovic, 2008).

Yet this is not always clearly reflected in the terminology of the academic landscape. In India, where anthropology has an extended history, its distinctiveness is frequently taken to involve a particular preoccupation with 'tribal' populations, and perhaps with physical anthropology – while some of the scholars recognized internationally as leading Indian anthropologists, concerned with the mainstream of Indian society, may be seen as sociologists in their own country (Uberoi et al., 2007). In African universities, too, founded in the late colonial or the postcolonial period, there has often been no distinction made, at least organizationally, between anthropology and sociology. If in Africa, anthropology has also for a period had to carry the stigma of being historically associated with the evils of colonialism, it seems now to find its own intellectual base in a collaboration between local and expatriate scholars (Devisch and Nyamnjoh, 2011).

Back in North America and Europe, again, the framework of academic life is not altogether stable over time. 'Cultural studies,' born in Britain but expanding from there, and putting itself on the intellectual map mostly from the 1970s onward, has been most successful as a cross-disciplinary movement in the Anglophone countries but has had an impact elsewhere as well. Its center of gravity may have been mostly in literary and media studies, but insofar as it has engaged with methods of qualitative field research and with issues of cultural diversity, not least in the areas of multiculturalism and diaspora studies, it has sometimes come close to sociocultural anthropology and – in certain European countries – to ethnology, as discussed in this article. While some anthropologists may see this as an undesirable intrusion into their disciplinary domain, others see it as a source of new stimuli (Dominguez, 1996; Nugent and Shore, 1997).

Conceptions of the Core

Again, a concern with human diversity may be seen to define anthropology as a whole (Hannerz, 2010). Even when the discipline at times turns to questions of what unites humanity, and what can be seen to constitute human nature, it does so more than other disciplines against the background of its panoramic understanding of diversity. If that concern may seem somewhat airy, one can also discern its more concrete implications for many of the distinctive working realities of the anthropological practitioner. Yet there has been a somewhat shifting cluster of assumptions here of what is the core of anthropological work and thought. As we inspect it, we will concentrate on sociocultural anthropology – by now the segment of the wider field that remains, internationally, most clearly identified as anthropology.

Obviously in North America and much of Europe, the tradition of studying more exotic forms of life has remained strong, and some might still maintain that this is what a 'real anthropologist' does. Yet few are now likely to find it entirely defensible at intellectual, moral, and political levels to have a separate discipline for the study of the 'non-West' (or 'non-North'). By the latter half of the twentieth century, past vocabulary such as 'primitive societies' had certainly become mostly an embarrassment. In a more egalitarian mood, many anthropologists would now undoubtedly prefer to see the practice of research among distant others in terms of a worldwide intellectual exchange effort: it is quite acceptable for North American and European anthropologists to go to Africa, Asia, or Oceania, as long as anthropologists from these regions are also welcomed to do research in the Occident. In practice, however, such a symmetrical exchange has hardly occurred on any significant scale. The professional anthropologists originating in non-Western, non-Northern parts of the world have seldom had the inclination, or the funding, to do their research abroad. They have been much more likely to conduct their studies in their home countries – where they may thus sometimes share their fields with expatriate anthropologists from Europe or North America.

In the latter part of the twentieth century, moreover, the emphasis on studying exotic others was weakening in American and European sociocultural anthropologies as well. The growing interest in and increasing legitimacy of 'anthropology at home' have been based partly on a sense that humanity is indivisible. The preoccupation with exploring human diversity is to a certain extent retained by emphasizing the internal diversity of forms of life (lifestyles, subcultures, etc.) of contemporary societies. The exotic may be just around the corner. Moreover, the study of one's own society is often motivated by a sense of relevance: one may be better placed to engage actively with whatever are perceived to be the wrongs of the place where one has the rights and the duties of a citizen. In some parts of Europe, too, where the discipline was relatively late in getting established, most anthropologists from the very beginning did their research in their own countries, even as they drew intellectual inspiration from the classic studies of exotics imported from older national anthropologies. Of course, such a repatriation of sociocultural anthropology may again blur certain disciplinary boundaries, perhaps with sociology or perhaps in some contexts with an 'ethnology' of the European type, which could also identify itself as an 'anthropology at home.'

If anthropology is no longer quite so committed to identifying itself in terms of exotic fields, its methodology appears, in some views, to offer another sense of distinctiveness. Anthropologists, since at least the early twentieth century, have typically done 'fieldwork.' They emphasize 'participant observation' or 'qualitative research' – in contrast with the handling of more or less impersonal statistical data – or 'ethnography,' in the double sense of the process of field study and the product of a largely descriptive account of a slice or

a way of life. In this discipline, the pure theorist is an anomaly, if not entirely nonexistent. The direct involvement with another way of being and behaving, whether it is in a village on the other side of the world, in a neighborhood across town, or in another occupation, tends to become more than a methodological choice, an effective way of reaching out to something unknown. It becomes a central personal experience, a surrender with strong moral and aesthetic overtones, and a potential for rich satisfaction and lifelong memories.

Even so, placing fieldwork at the core of the distinctiveness of anthropology is also problematic. In part, the problem is that 'doing ethnography' has become increasingly common in various other disciplines as well, even if anthropologists sometimes complain that it is then done badly or superficially ('ethnography lite'). It is also true that fieldwork of the most orthodox variety does not fit readily into every setting. In all field studies, anthropologists tend to do not only participant observation but other kinds of work as well. They talk to informants, elicit life histories, collect texts, conduct surveys, and engage in some variety of activities to acquire new knowledge. To what extent their observational work can really be actively participatory must depend a great deal on the contexts. But furthermore, in contemporary life, observational work is sometimes not very rewarding. Following the daily round of an agricultural village is one thing; observing an office worker at his desk, staring at his computer screen, is another. On the other hand, fields in the present-day world may involve other kinds of data than those of classic anthropology – more media use, for example. Notions of 'field' and 'fieldwork' are increasingly coming to be debated in anthropology, as the world changes and as anthropologists try to fit their pursuits into it; there are clearly also anthropologists who prefer less distinctively anthropological methodological repertoires (cf Gupta and Ferguson, 1997; Bernard, 1998; Coleman and Collins, 2006).

The idea of gaining knowledge and understanding from fieldwork is certainly not entirely separate from the issues of where, and among whom, anthropologists work either. While anthropologists have been inclined to assume, or argue explicitly, that there are special insights to be gained from combining an outsider view with immersion in another way of life, the rise of 'anthropology at home,' the increasingly frequent copresence of indigenous and expatriate anthropologists in many fields, and the growth elsewhere in academic life of more emphatically insider-oriented fields of study, such as some varieties of ethnic studies, leave assumptions and arguments of this kind more clearly open to debate. Some would question as a matter of principle what kind of validity an outsider's perspective can have. Others could argue that one can never be an anthropologist and an insider at the same time, as these are distinct intellectual stances. And then, surely, such issues are further complicated by the fact that 'insider' and 'outsider' are not necessarily either-or categories, and hardly altogether stable either. An outsider can perhaps become an insider – and, on the other hand, one can probably start as an insider, drift away, and become an outsider. Here is an enduring complex of problems that anthropologists and their colleagues in adjacent disciplines will continue to debate (Merton, 1972; Narayan, 1993).

One may wish to object to the somewhat unreflective inclination to put fieldwork at the center of the discipline that it would tend to make anthropology primarily a methodological specialty – perhaps, because of its qualitative emphasis, the counterpart of statistics. Undoubtedly, many anthropologists would prefer to identify their discipline with particular ideas, theories, and intellectual perspectives, and the history of anthropology is most often written in such terms. It may still turn out to be difficult to point to any single, uncontested, enduring central concept, or structure of concepts.

A key idea in the exploration of human diversity has obviously been that of 'culture,' not least in the plural form of 'cultures.' This is a key term put to many uses in contemporary life (Breidenbach and Nyíri, 2009). Most fundamentally, in its conventional but multifaceted anthropological form, 'culture' stands for whatever human beings learn in social life, as contrasted with whatever is inborn (i.e., genetically given). But with the idea that human beings are learning animals also goes the understanding that they can learn different things, so that what is cultural tends to exhibit variation within humanity. Moreover, the tendency has been to conceptualize culture at a level of human collectivities (societies, nations, and groups), so that members are held to be alike, sharing a culture, while on the other hand there are cultural differences between such collectivities. Culture, that is, has been taken to come in distinct packages – cultures. In the classic view of culture, there has also been a tendency to try to see each such culture 'as a whole' – to describe entire ways of life and thought, and to throw light on the varied interconnections among ideas and practices.

Such a basic notion of culture has allowed anthropologists to proceed with the task of drawing a panoramic map of human diversity. There has certainly also been a widespread preoccupation with showing how much of human thought and behavior is actually cultural (i.e., learned) and thus variable, rather than altogether biologically based and uniform. This emphasis has been visible, for example, in anthropological contributions to gender studies.

Yet culture has also been a contested concept in anthropology, not least in recent times. It is true that it was never equally central to anthropological thought in all varieties of anthropology. American sociocultural anthropology, which has more often been identified as 'cultural anthropology,' has focused rather more on it, for example, than its British counterpart, which has been more inclined over the years to describe itself as 'social anthropology.' By the late twentieth century, however, critics in the United States as well as elsewhere argued that the established style of cultural thought tended to exaggerate differences between human collectivities, to underplay variations within them, to disregard issues of power and inequality and their material bases, and to offer overly static, ahistorical portrayals of human ways of life. While some anthropologists have gone as far as to argue for the abolition of the culture concept, others would be inclined to accept many of the criticisms yet take a more reformist line. Whichever view is taken, it seems that debates over the understanding of culture offer the discipline one of its lively intellectual foci (Hannerz, 1993; Brumann, 1999; Kuper, 1999).

The notion of cultural translation has also had a part in defining anthropological work, and inevitably it is drawn into the arguments about the culture concept generally. Indeed, bridging cultural divides by making the ideas and expressions

of one culture understandable in terms of another has been a major anthropological activity, although this linguistic analogy is certainly too limiting to describe all anthropology (Asad, 1986; Pálsson, 1993). In a somewhat related manner, it has gone with the interest in human diversity to describe anthropology as a discipline centrally involved with comparison (Gingrich and Fox, 2002). In a general way, this is clearly valid. Much of anthropology is at least implicitly comparative, in its inclination to emphasize what is somehow notably different about the ideas, habits, and relationships of some particular population. To the extent that the origins of the discipline have been European and North American, the baseline of such comparisons and contrasts has no doubt been in a general way Western. At the same time, there has been a strong tendency to use the knowledge of diversity to scrutinize Western social arrangements and habits of thought, and thus to destabilize 'common sense.' In such a way, anthropological comparison can serve the purpose of cultural critique, a possibility used in different ways at different times (Marcus and Fischer, 1986). On the other hand, the form of large-scale comparative studies, focusing on correlations between particular sociocultural elements across a large number of social units, which was prominent in the discipline in some earlier periods, has not been a central feature for some time, and the intellectual and methodological assumptions underlying comparison are also a more or less continuous topic of argument (Köbben, 1970; Holy, 1987).

Specialties and Subdisciplines

If culture in the singular and plural forms, fieldwork, translation, and comparison may count as ideas and practices that hold anthropology together, not in consensus but often rather more through engagement in debate, it is often suggested that the discipline also tends to be systematically segmented along various lines – not only in terms of what is understood to make up the discipline, as noted in this article, but also with regard to what draws communities of specialists closer together. One of the three major dimensions here is area knowledge. Anthropologists tend to be Africanists, Europeanists, Melanesianists, or members of other region-based categories. Rather few of them ever commit themselves to acquiring specialized knowledge of more than one major region to the extent of doing field research there. It is true that if fieldwork is often highly localized, it does not necessarily lead to a wider regional knowledge either, but as a matter of convention (perhaps at least to fit into normal job descriptions in the discipline), the tendency has been to achieve specialist status by reaching toward an overview of the accumulated anthropological knowledge of some such unit, and perhaps seeking opportunities to familiarize oneself with more of it through travel. In countries where 'area studies' have been academically institutionalized, such regional anthropologies have had a significant role in the resulting interdisciplinary structures, and, generally, shared regional specialization has often been important in scholarly exchanges across discipline boundaries.

The second major dimension of specialization can be described as topical. Although anthropology has a traditional orientation toward social and cultural 'wholes,' there is yet a tendency among many anthropologists to focus their interest on particular kinds of ideas, practices, or institutions. Frequently, such specializations have tended to follow the dominant dividing lines between other academic disciplines – there has been political anthropology, economic anthropology, psychological anthropology, anthropology of law, anthropology of art, and ecological anthropology, for example. Drawing on knowledge of social and cultural diversity, one aspect of such specialization has been to scrutinize and criticize concepts and assumptions of the counterpart disciplines with respect to their tendencies toward Western-based intellectual ethnocentrism; but obviously there is also a continuous absorption of ideas from these other disciplines. Here, then, are other interdisciplinary connections. It may be added that such subdisciplines often have their own histories of growth in some periods, and stagnation in others (Collier, 1997).

Thirdly, anthropologists sometimes specialize in the study of broad societal types, based on dominant means of livelihood – hunter-gatherers, peasants, pastoralists, fishermen, and so forth. In some ways, urban anthropology may be seen as a similar kind of specialty. In fields like these, too, intensities of collective engagement and intellectual progress have tended to vary in time.

The Practical Uses of Anthropology

It has occasionally been argued that anthropology is less engaged with practical application than many other academic disciplines, and more concerned with achieving a somewhat lofty overview of the human condition, in all its variations. One factor underlying such a tendency may be a sort of basic cultural relativism – it may go with the acceptance, and even celebration, of human diversity to be somewhat skeptical of any attempt to impose particular arrangements of life on other people. This stance may well be supported by the fact that anthropologists have often been outsiders, working in other societies or groups than their own, and feeling that they have no mandate for meddling, or for that matter any actual realistic opportunity.

Nonetheless, there have always been a number of varieties of 'applied anthropology.' In an early period, when the study of non-Western societies was carried out in contexts of Western colonialism, it was held that anthropological knowledge could be useful in colonial administration. While some administrators had anthropological coursework before taking up their overseas assignments, and while some anthropologists did research oriented toward such goals, it seems that such connections remained rather limited, and administrators were often impatient with the anthropologists' preference for a long-term buildup of basic knowledge. Particularly toward the end of the colonial period, anthropologists also often saw such involvements as morally and politically questionable.

In the postcolonial era, anthropological knowledge became engaged in a larger scale, and in many different ways, in development work in what had now (at least for some time) turned into the 'Third World.' Many governments, international agencies, and international nongovernmental organizations have thus drawn on anthropological advice, and have

offered arenas for professional anthropological activity outside academia. At the same time, with 'development' turning into a global key concept, around which an enormous range of activities and organizations revolve, this again has become another focus for critical theoretical scrutiny among anthropologists (Dahl and Rabo, 1992; Escobar, 1995). In recent times, particularly in the United States, there has also been considerable debate over the uses of anthropological expertise in contexts of war and other international conflict (Kelly et al., 2010).

When anthropologists work in their own countries, the range of applications may be wider, and as already noted, practical and political relevance is frequently one reason for doing anthropology 'at home.' Not least as transnational migration has made more societies increasingly ethnically and culturally heterogeneous, anthropologists have been among the social scientists engaged, in one way or other, in the handling of minority affairs and multiculturalism. Educational and medical anthropology are not always involved in practical application, and research in such fields is also carried out in societies other than the researcher's own, but often work in these subdisciplines similarly centers on the encounter between major institutional complexes and culturally heterogeneous populations. Some number of anthropologists in Europe and North America now also make use of their disciplinary perspectives as specialists on organizational culture, and as marketing analysts, in a growing field of business anthropology (Moeran, 2005). A further practical field that emerged in the late twentieth century as its own profession is 'intercultural communication,' which offers training and consultancy in the handling of concrete situations of cultural difference and, putting it somewhat dramatically, 'culture shock.' This field has links to several academic disciplines, but to a considerable degree it draws on a more or less anthropological conception of culture – of a type, it may be added, that many anthropologists might now find rather old-fashioned and clumsy.

The Future of Anthropology

In all its varying shapes, in space and over time, anthropology has tended to straddle conventional academic classifications of disciplines. In its scope of subject matter – for example, family and kinship, politics, and market and exchange, on the one side, and art, music, and dance, on the other – it extends across the social sciences and the humanities. Insofar as it has to take into account the biological givens of human thought and action, and inquires into the interactions of humankind with its natural environment, it reaches into the natural sciences as well. But its multiple affiliations are not only a consequence of its varied subject matter. They are also implied in the variety of intellectual approaches: in field research, in theoretical work, and in styles of presentation. In what ways, or to what extent, anthropology is an art or a science continues to be a matter of lively internal debate.

In many ways the enduring characteristics of anthropology, throughout this range of forms, continue to be expressions of the concern with diversity – with the highly varied manners of being human. To the global public stock of ideas, it brings such notions as taboo, witchcraft, cargo cults, totemism, or the potlatch exchange feasts of Northwest Coast American Indians. Concepts such as 'Big man' (out of Melanesia), 'patron–client relationships' (not least out of the Mediterranean area), or 'caste' (out of India), allowed to travel out of their areas of origin, can enrich our thinking about power, politics, and inequality in many contexts. There is a rich intellectual universe here, to be drawn on within the discipline and from outside it. And anthropology has its classic preoccupations, such as ritual or kinship, concerning which new materials about yet more variations are continuously gathered worldwide, and around which theoretical debates never seem to cease.

At the same time, anthropology goes on reconfiguring itself (cf Wolf, 1964; Hymes, 1972; Fox, 1991; Ingold, 1996; Rabinow et al., 2008). One might perhaps have thought that a discipline revolving around the diversity of human forms of life and thought would find itself in difficulty at a time when increasing global interconnectedness may lead to cultural homogenization, and a loss of local or regional cultural forms. Indeed, it is true that many people in the world no longer stick to the beliefs they used to hold, and are discarding some of their past practices, ranging from spirit possession to headhunting. In part, the responsibility of anthropology here becomes one of preserving a record of the ways of humans, past and present, and keeping that record alive by continuing to scrutinize it, interpret it, and bring it to bear on new developments.

It is also true, however, that diversity is not diminishing as much as some, perhaps fairly superficial, views of globalization might suggest. Traditions may be remarkably resilient, adapting to new influences and absorbing them in various ways; there may be more than one 'modernity.' Moreover, new interconnections may generate their own social and cultural forms, so that there may be cultural gain as well as cultural loss. A growing interest in anthropology with such notions as hybridity and creolization bears witness to this. In addition, however, anthropology has continuously expanded the field of social and cultural variations with which it has actively engaged. As the discipline was practiced in the earlier decades of the twentieth century, one might have discerned a tension between, on the one hand, the ambitious claims of offering a view of humanity, and, on the other hand, the actual concentration on villages of horticulturalists, bands of hunter-gatherers, or other exotic and small-scale sociocultural arrangements. But then by the mid-twentieth century, there was an increasing involvement with peasant societies, and the civilizations of which they were a part, and later yet urban anthropology appeared as another subdiscipline, practiced in field settings in every region. In recent times, the anthropology of science has been another growing specialization, as the practices of scientists are scrutinized as yet another kind of cultural constructs, and as laboratories are added to the range of possible sites of fieldwork (Rabinow, 1996; Downey and Dumit, 1997). It appears, consequently, that the enduring wider claims of the discipline to an overall engagement with human modes of thought and action are increasingly being realized.

Obviously, in many of its fields of interest, anthropologists now mingle with scholars from a variety of other backgrounds, and the division of labor between disciplines here may not seem obvious. Undoubtedly, there are blurred boundaries as well as cross-fertilization, but anthropologists would emphasize

the intellectual potential of a conceptual apparatus trained on, and informed by a knowledge of, a great variety of cultural assumptions and institutional arrangements. There is also the commitment to close-up observation of the relationships between human words and deeds, and the strain toward understanding 'wholes,' which, despite its vagueness, may usefully involve a particular commitment to contextualization as well as skills of synthesis.

Anthropologists now find themselves at work in a global ecumene of increasing, and increasingly polymorphous, interconnectedness. This is a time in which that diversity of social and cultural forms with which they are preoccupied is constantly as well as rapidly changing, and where new social, political, economic, and legal frameworks encompassing and rearranging that diversity are emerging. More people may have access to a larger part of the combined human cultural inventory than ever before; conversely, whether one likes it or not, more of that cultural diversity can also come in one's way. This is also a time when debates over the limits of diversity are coming to new prominence, for different reasons. Evolutionary biologists are setting forth new views of human nature that need to be carefully confronted with understandings of cultural variation. As people sense that they live together in one world, questions also arise over what is, or should be, shared humanity, for example, in the area of human rights (Wilson, 1997; Goodale and Merry, 2007). There would seem to be a place in the public life of this era for a cosmopolitan imagination that both recognizes diversity and seeks the ground rules of a viable and humane world society. For such a cosmopolitan imagination, one would hope, anthropology could continue to offer materials and tools.

Bibliography

Asad, T., 1986. The concept of cultural translation in British social anthropology. In: Clifford, J., Marcus, G.E. (Eds.), Writing Culture. University of California Press, Berkeley, CA.

Bernard, H.R. (Ed.), 1998. Handbook of Methods in Cultural Anthropology. AltaMira Press, Walnut Creek, CA.

Boskovic, A. (Ed.), 2008. Other People's Anthropologies. Berghahn, Oxford.

Breidenbach, J., Nyíri, P., 2009. Seeing Culture Everywhere. University of Washington Press, Seattle, WA.

Brumann, C., 1999. Writing for culture: why a successful concept should not be discarded. Current Anthropology 40, S1–S13.

Coleman, S., Collins, P. (Eds.), 2006. Locating the Field. Berg, Oxford.

Collier, J.F., 1997. The waxing and waning of "subfields" in North American sociocultural anthropology. In: Gupta, A., Ferguson, J. (Eds.), Anthropological Locations. University of California Press, Berkeley, CA.

Dahl, G., Rabo, A. (Eds.), 1992. Kam-ap or Take-Off. Stockholm Studies in Social Anthropology, vol. 29. Almqvist and Wiksell International, Stockholm.

Devisch, René, Nyamnjoh, Francis B. (Eds.), 2011. The Postcolonial Turn: Re-imagining Anthropology and Africa. African Studies Centre, Leiden, the Netherlands, and Langaa Research and Publishing, Bamenda, Cameroon.

Dominguez, V.R., 1996. Disciplining anthropology. In: Nelson, C., Gaonkar, D.P. (Eds.), Disciplinarity and Dissent in Cultural Studies. Routledge, New York.

Downey, G.L., Dumit, J. (Eds.), 1997. Cyborgs and Citadels. School of American Research Press, Santa Fe, NM.

Escobar, A., 1995. Encountering Development. Princeton University Press, Princeton, NJ.

Fahim, H. (Ed.), 1982. Indigenous Anthropology in Non-Western Countries. Carolina Academic Press, Durham, NC.

Fox, R.G., 1991. Recapturing Anthropology. School of American Research Press, Santa Fe, NM.

Gingrich, A., Fox, R.G. (Eds.), 2002. Anthropology, by Comparison. Routledge, London.

Goodale, M., Merry, S.M. (Eds.), 2007. The Practice of Human Rights. Cambridge University Press, New York.

Gupta, A., Ferguson, J. (Eds.), 1997. Anthropological Locations. University of California Press, Berkeley, CA.

Hannerz, U., 1993. When culture is everywhere: reflections on a favorite concept. Ethnos 58, 95–111.

Hannerz, U., 2010. Anthropology's World. Pluto, London.

Holy, L. (Ed.), 1987. Comparative Anthropology. Blackwell, Oxford.

Hymes, D. (Ed.), 1972. Reinventing Anthropology. Pantheon, New York.

Ingold, T. (Ed.), 1996. Key Debates in Anthropology. Routledge, London.

Kelly, J.D., Jauregui, B., Mitchell, S.T., Walton, J. (Eds.), 2010. Anthropology and Global Insurgency. University of Chicago Press, Chicago, IL.

Köbben, A.J.F., 1970. Comparativists and non-comparativists in anthropology. In: Naroll, R., Cohen, R. (Eds.), A Handbook of Method in Cultural Anthropology. Natural History Press, Garden City, NY.

Kuper, A., 1999. Culture. Harvard University Press, Cambridge, MA.

Marcus, G.E., Fischer, M.M.J., 1986. Anthropology as Cultural Critique. University of Chicago Press, Chicago, IL.

Merton, R.K., 1972. Insiders and outsiders: a chapter in the sociology of knowledge. American Journal of Sociology 78, 9–47.

Moeran, B., 2005. The Business of Ethnography. Berg, Oxford.

Narayan, K., 1993. How native is a "native" anthropologist? American Anthropologist 95, 671–686.

Nugent, S., Shore, C. (Eds.), 1997. Anthropology and Cultural Studies. Pluto Press, London.

Pálsson, G. (Ed.), 1993. Beyond Boundaries. Berg, Oxford.

Rabinow, P., 1996. Making PCR. University of Chicago Press, Chicago.

Rabinow, P., Marcus, G.E., Rees, T., 2008. Designs for an Anthropology of the Contemporary. Duke University Press, Durham, NC.

Ribeiro, G.L., Escobar, A. (Eds.), 2006. World Anthropologies. Berg, Oxford.

Uberoi, Patricia, Sundar, Nandini, Deshpande, Satish (Eds.), 2007. Anthropology in the East: Founders of Indian Sociology and Anthropology. Permanent Black.

Vermeulen, H.F., Alvarez Roldán, A. (Eds.), 1995. Fieldwork and Footnotes. Routledge, London.

Wilson, R.A. (Ed.), 1997. Human Rights, Culture and Context. Pluto Press, London.

Wolf, E., 1964. Anthropology. Prentice-Hall, Englewood Cliffs, NJ.

Antidepressant Drugs

Carmine M Pariante, Institute of Psychiatry, King's College London, London, UK
O Benkert, A Szegedi, and MJ Müller, University of Mainz, Mainz, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by O. Benkert, volume 1, pp. 529–535, © 2001, Elsevier Ltd.

Abstract

Antidepressants are psychotropic drugs for the treatment of depressive syndromes of various etiologies. This historical definition nowadays reflects only one facet of the spectrum of their possible therapeutic applications. More than 40 years of clinical experience and scientific research with antidepressant drugs have resulted in a variety of different indications with proven efficacy for these agents, particularly anxiety disorders or obsessive–compulsive disorder. As antidepressants are effective across different psychopathological syndromes of traditional nosological categories, an up-to-date definition for this heterogeneous group of agents with different chemical structures and pharmacological profiles still needs to be established.

Because of historical developments, a stringent classification for all available agents with antidepressant properties has not yet been agreed upon. Initial classificatory attempts focused on the chemical structure of the drugs. Recent classifications emphasize the pharmacological profile of the drug. This latter approach is of greater clinical relevance, because it depicts the primary target sites in the central nervous system (e.g., transporter proteins or various receptor subtypes), allowing a better prediction of the potential clinical application and the side-effect profile of the drug. As a result of both approaches, many clinicians today distinguish the following groups of antidepressants: (1) monoaminooxidase inhibitors (MAOI), (2) tricyclic antidepressants (TCA), (3) selective serotonin reuptake inhibitors (SSRIs), and (4) antidepressants with different pharmacological profiles. In addition, drugs initially developed for other indications have shown some promising effects as antidepressants, such as anti-inflammatory medications.

History

Monoaminooxidase Inhibitors

Iproniazid and its precursor isoniazid had been developed in the laboratories of the company Hoffmann-La Roche primarily as antituberculosis agents. As a consequence of casuistic observations on mood-elevating properties of iproniazid, this effect was first confirmed by Loomer et al. (1957) in a clinical trial. However, significant hepatotoxicity was observed during iproniazid treatment. Subsequent development of drugs inhibiting monoamine oxidase (MAO) systems led to the introduction of isocarboxazid, phenelzine, and tranylcypromine as irreversible inhibitors of MAO. Their use has been limited because of possible hypertensive crises after ingestion of food containing large amounts of tyramine, forcing patients to follow dietary restrictions during pharmacotherapy with these drugs. This led to the development of reversible monoaminooxidase inhibitors (MAOI) such as moclobemide, which no longer required dietary restrictions.

Tricyclic Antidepressants

After the discovery of the antipsychotic effects of chlorpromazine in 1952, the compound G22355, later called imipramine, was developed in the laboratories of the company Geigy. Imipramine was a derivative of chlorpromazine and was first tested in clinical trials in schizophrenia. While no relevant antipsychotic effects were noted, mood-elevating effects in patients with a co-occurring depressive syndrome were observed. At this point, both the company and the clinical investigator Kuhn decided to test imipramine in patients with depression as a potential antidepressant. The results of the clinical trials were first published in 1957, and are still a remarkable document of skillful clinical observation, as the report contains many details about the antidepressant effects of imipramine which still hold true today (Kuhn, 1957).

A variety of compounds that were chemically related to imipramine have been developed. They all shared the basic chemical structure of a tricyclus. Subsequent research revealed that in most instances, while the tricyclic basal structure was preserved, minor alterations of its side chains resulted in markedly different pharmacological profiles of the drugs. The chemically related compound maprotiline, often termed a tetracyclic, was developed as a consequence of systematic variation of the tricyclic structure.

From the 1960s to the 1980s, the tricyclic antidepressants (TCA) were the dominating group of antidepressants in clinical practice throughout the world. As many of the TCA compounds induce substantial side effects, they gradually lost their predominance with the availability of the selective serotonin reuptake inhibitors (SSRIs) during the 1990s.

Selective Serotonin Reuptake Inhibitors

In the 1970s, on the basis of increasing knowledge about the mechanism of action of MAOI and TCA, pathophysiological models of depression were proposed. Consequently, potential antidepressants with more specific neurotransmitter receptor effects were developed. According to the norepinephrine depletion hypothesis of depression, drugs to selectively potentiate central noradrenergic neurotransmission were developed. On the other hand, the various available TCAs were scrutinized for their potency of serotonin reuptake inhibition. Clomipramine was shown to have the strongest effects on serotonin reuptake inhibition; however, the substance does not only selectively inhibit the serotonin transporter but also shows effects on norepinephrine.

Subsequently, in a collaboration between a pharmaceutical company (Astra) and the research team of Carlson in Sweden, the first SSRI zimelidine was developed. The first results from a clinical trial on zimelidine were published in 1977 (Benkert et al., 1977). However, the antidepressant had to be withdrawn from the market because of the emergence of neurological side effects.

In the 1980s, fluoxetine ('Prozac'), an SSRI developed by Eli-Lilly Company, was made available in the USA. The introduction of this drug represents the breakthrough of SSRIs in the treatment of depression. Successful marketing strategies and the low toxicity of fluoxetine may explain the broad spectrum of indications for this drug. The prescription of fluoxetine in patients with personality disorders and, finally, the use of Prozac by healthy people to improve their general wellbeing, led to an ongoing discussion regarding the acceptability of so-called 'lifestyle' drugs.

In the following years, four additional SSRIs were approved: fluvoxamine, paroxetine, sertraline, and citalopram. The latter two substances have favorable pharmacokinetics and a low risk for drug–drug interactions – an issue that became increasingly important in the 1990s.

Antidepressants with Different Pharmacological Profiles

The introduction of selective antidepressants reduced the rate of unwanted side effects. The higher selectivity of these drugs was, however, not accompanied by higher efficacy. Therefore, research and drug development focused on compounds which exerted their effects on different and interrelated neurotransmitter receptors in the brain to enhance their benefit–risk ratio based on theoretical reasoning about the anticipated profile of effects and side effects.

The push to develop highly selective antidepressants targeting a single type of neurotransmitter receptor or transporter had ended. Highly effective and well-tolerated drugs like mirtazapine, venlafaxine, and, more recently, agomelatine, are examples of the new strategy of drug design.

Rather independent from this line of progress, the herb St. John's Wort has achieved great popularity throughout the world, based on the drug's low toxicity, its high acceptance, and a proven efficacy in the treatment of mild to moderate depression. The exact mechanism of action is not completely known.

Mechanism of Action of Antidepressants

There is considerable evidence that many antidepressants facilitate neurotransmission of one or more monoamines – serotonin, norepinephrine, or dopamine. This effect appears to be mediated by different mechanisms that are described in the following sections.

Blockade of One or More Types of Synaptic Monoamine Transporters

Synaptic monoamine transporters are protein structures which rapidly retrieve released neurotransmitter molecules from the synaptic cleft into the originating neuron. By specifically blocking the transporter, the reuptake of the neurotransmitter is hampered, thus increasing its availability in the synapse. Examples of compounds that predominantly or selectively inhibit the reuptake of serotonin are the SSRIs listed above, or the tricyclic compound clomipramine. Drugs predominantly or selectively inhibiting the norepinephrine reuptake are reboxetine, and the tricyclic compounds desipramine and nortriptyline. Many antidepressants block both the serotonin and norepinephrine transporter, more or less with the same potency, for example, imipramine, amitriptyline, doxepine, milnacipran, or venlafaxine. Examples of dopamine reuptake inhibitors are bupropion (also affecting norepinephrine uptake) or amineptine.

Blockade of the Enzymes That Inactivate Monoamines

The main enzyme relevant in this context is the MAO. Its isoform MAO-A degrades mainly serotonin and norepinephrine, while MAO-B predominantly metabolizes dopamine. MAO can be blocked irreversibly and unselectively by compounds such as phenelzine or tranylcypromine. Reversible MAO-A inhibitors such as moclobemide selectively block the metabolism of serotonin and norepinephrine within therapeutic ranges. Selective MAO-B inhibitors have not consistently proved to be antidepressants.

Stimulation and/or Blockade of Receptor Subtypes

In this category, various compounds with different modes of action are mentioned. From the tricyclic compounds, trimipramine blocks dopamine D2, alpha-1-adrenergic H1-histaminergic, and muscarinic receptors without any substantial effect on monoamine transporters. Mirtazapine blocks alpha-1 and 2-adrenergic receptors, as well as 5-HT_2 and 5-HT_3 and H1 histaminergic receptors. Nefazodone mainly blocks 5-HT_2 receptors, with weak inhibition of serotonin and norepinephrine uptake. The most recent development in novel pharmacological strategies is agomelatine, a drug that has antagonist activity at serotonergic (5HT2C) receptors and agonist activity at melatonergic (MT1/MT2) receptors.

The pharmacological profile of most antidepressants has been well characterized, providing an explanation for some of the acute and prolonged effects of these agents. In particular, the side-effect profile of antidepressants can be readily derived from their pharmacological profile. However, the postulated enhancement of monoaminergic function may not be the only or even the crucial factor explaining why antidepressants work. The pharmacological profile of a drug does not sufficiently explain its antidepressant properties. This is borne out by the following findings:

1. The pharmacological target sites within the CNS are affected within minutes after administration of the drug. Many side effects of these agents occur early in the course of treatment, in good time correlation with the administration of the drug. However, virtually all antidepressants need several days to weeks of regular application until their full therapeutic effect becomes clinically visible.
2. Other pharmacological agents acting as monoamine reuptake inhibitors comparable to the above-mentioned

antidepressant drugs have not been found to possess convincing antidepressant properties (e.g., cocaine, amphetamine).

Several hypotheses have been developed to explain the delay of full antidepressant effects. It has been postulated that the effects of antidepressants at their primary target sites (receptors or transporters as first messengers) induce secondary effects, for example, adaptatory changes in intracellular signal transduction at the level of second messengers (such as the adenylyl cyclase system) or of gene expression. This may lead to altered receptor sensitivity (as described for beta-adrenergic receptor down-regulation or decreased sensitivity of serotonergic receptors, such as 5-HT_2 receptors). Changes of the synthesis of brain-derived neurotrophic factor and/or alterations of the hypothalamic–pituitary axis may represent pathways by which antidepressants finally exert their therapeutic effects. Changes in synthesis of protein structures might explain the time lag needed for antidepressants to develop their full therapeutic effect on mood and affect. More recently, clinical trials have demonstrated antidepressant effects by drugs with anti-inflammatory action, indicating that increased inflammation maybe one of the mechanisms by which depression lead to disrupted brain function (Fond et al., 2014).

Indications

By now, many drugs with antidepressant effects have also been evaluated as effective agents in several indications, which are briefly outlined below.

Depressive Disorders

The classical indication for TCA, MAOI, SSRI, and other antidepressants is depressive episodes (e.g., major depression). Their use is the treatment of choice in moderate or severe depression. Antidepressants are effective both in acute treatment and in preventing relapse of unipolar depression during long-term treatment (continuation or maintenance phase, after the acute symptoms have been treated effectively). Several authors have found that depressive episodes with melancholic features are more likely to respond to drug treatment than to psychotherapy. So-called atypical depressive episodes are reportedly more likely to respond to MAOI (and perhaps to SSRI) than to TCA such as imipramine. In the treatment of psychotic depression, the combination of an antidepressant with an antipsychotic is warranted. Contrary to former belief, long-lasting mild to moderate depressive states, currently labeled as dysthymia, also respond well to antidepressant drug treatment in a majority of cases. In bipolar disorder, the use of antidepressants during a depressive episode may trigger a switch into a manic state; therefore their use in this indication is currently a matter of debate.

Panic Disorder

Starting with TCA such as imipramine or clomipramine, a variety of antidepressant drugs have been demonstrated to be an effective treatment in panic disorder with or without agoraphobia. Currently, SSRI are the treatment of choice in this indication, if a drug treatment with an antidepressant is established. SSRI have shown convincing anxiolytic properties in combination with a favorable side-effect profile for a majority of patients.

Obsessive–Compulsive Disorder

This often chronic and disabling disorder, characterized by intruding thoughts or repetitive compulsions that often are time-consuming and may interfere considerably with an individual's functioning, seems to respond selectively to compounds such as clomipramine or SSRI that act as predominant or selective serotonin reuptake inhibitors, because norepinephrine reuptake inhibitors or MAOIs have not consistently shown therapeutic effects in this indication.

Generalized Anxiety Disorder

In the treatment of this chronic anxiety disorder characterized by waxing and waning anxiety symptoms and excessive worrying, several antidepressants such as SSRI or venlafaxine have demonstrated beneficial effects in reducing cognitive anxiety symptoms in particular.

Social Phobia

MAOI and SSRI are effective in the treatments for this disorder, characterized by marked anxiety in social situations, where the individual is exposed to unfamiliar people or to possible scrutiny by others.

Bulimia Nervosa

This disorder, with recurrent episodes of binge eating, loss of control, overeating impulses, and inappropriate compensatory behavior to prevent weight gain, can also be beneficially influenced by antidepressants. Binge frequency and purging behavior in particular can be significantly reduced by regular treatment.

Premenstrual Syndrome

Symptoms of tension, dysphoria, and irritability, accompanied by various somatic complaints and occurring during the late luteal phase of the menstrual cycle, can be effectively treated by continuous or intermittent administration of SSRI or clomipramine. As in the case of OCD, there is evidence that only predominant or selective serotonin reuptake inhibitors are effective from the spectrum of antidepressant drugs.

Further Indications

Antidepressants have also been demonstrated to be of therapeutic value in the treatment of pain syndromes of various etiologies, in the prophylaxis of migraine, in the treatment of posttraumatic stress disorder, in the treatment of children with enuresis or attention-deficit and hyperactivity syndromes, or in patients with narcolepsy (especially clomipramine). SSRI may also be of therapeutic value in treating premature ejaculation,

obesity, or anger attacks associated with mood or personality disorders. In impulsive and affectively unstable personality disorders (e.g., borderline personality disorder) SSRI may be beneficial, whereas MAOI may be helpful in avoidant personality disorder.

Therapeutic Principles

If treatment with antidepressant drugs is established, the prescribing clinician and the treated patient need to consider some general guidelines, which apply to most indications for which these agents have been evaluated:

1. Treatment with antidepressants generally requires continuous treatment over several weeks on a regular basis; intake on an as-needed base generally is not sensible.
2. Therapeutic effects generally develop gradually during continuous treatment with antidepressants. In a majority of cases, 2–3 weeks are needed until more than 50% of symptom reduction is achieved.
3. During the course of treatment, side effects usually appear before the desired therapeutic effect reaches its maximum. This particular issue needs to be clarified with the patient in order to ensure compliance with treatment up to the point of clinical remission.
4. Antidepressants need to be given in a sufficiently high dose to achieve optimal treatment effects. Efficient doses for TCA usually lie between 100 and 300 mg day^{-1}, SSRI such as citalopram, fluoxetine, or paroxetine need 20 mg day^{-1} and sertraline or fluvoxamine, at least 50 mg day^{-1}. Depending on the side-effect profile of a drug, initial treatment requires gradual dose increase until effective doses are achieved (for TCA recommended starting doses are 25–50 mg day^{-1}; SSRI can usually be started with the minimal effective dose).

After remission is reached, it is generally recommended to maintain the effective dose for at least several weeks or months in order to prevent recurrence of the symptoms or relapse. In cases of known high risk of recurrence, continued treatment over years is needed. Combining antidepressant drug treatment with psychotherapy (especially cognitive–behavioral or interpersonal therapy) usually provides additive benefit for the patient; both approaches are not contradictory.

Recent large, naturalistic clinical studies have shown that most antidepressant strategies (including augmentation, combinations, and switching) are equally effective, and that around 25% of patients do not respond to any available antidepressants (Rush, 2011). However, other authors have compared data from large datasets of clinical trials and identified that some antidepressants, such as sertraline, escitalopram, and venlafaxine, are more effective than others (Cipriani et al., 2006).

Antidepressants do not have 'addictive' properties, but a discontinuation syndrome has now been recognized, with symptoms including symptoms such as nausea, vomiting, diarrhea, headaches, sweating, insomnia, tremors, and vertigo, as well as mood disturbances that may be mistaken as depression relapses. This has led to clear guidelines indicating the need for the slow tapering of the dose before stopping any antidepressants.

Disputable Issues: Ethical, Methodological, and Clinical

Despite the widely accepted beneficial effects of antidepressants, there is ongoing debate on how these effects and benefits can be assessed. The scientific gold standard established over a period of about 50 years, and representing the basis of approval of new antidepressant drugs by legal authorities, is to assume *efficacy* of antidepressants if they are clearly superior to placebo in at least two well-conducted randomized controlled trials. Ethical issues are primarily related to the use of placebo in such pivotal trials. A recent study has shown that suicide rates and mortality in patients treated with placebo in randomized controlled trials were not discernible from those of patients treated with an antidepressant verum (Khan et al., 2000).

The demonstration of efficacy of antidepressants in controlled clinical trials is necessary but not sufficient for evaluating the therapy of major depression and related disorders. The focus of interest is shifting from pure efficacy to *effectiveness* studies of antidepressants. This implies asking the question whether a treatment still works when used by the average clinician with the average patient in an average therapeutic setting. Additionally, *efficiency* studies are essential to assess the required level of resources to produce reasonable benefit. Research at the interface of clinical trials and effectiveness studies, including cost–utility and econometric studies, is indispensable to estimate the benefit and the shortcomings of available antidepressant drug therapy (Wells, 1999).

Evidence from double-blind, placebo-controlled trials of antidepressant drugs from all classes has shown them to be effective in alleviating symptoms in approximately 60–75% of patients with major depression. In contrast, the placebo response rate averages 35% after 1 month of treatment. However, the very high variability of verum and placebo responses makes it nearly impossible to extrapolate from one study to another or from the results of a clinical trial to real clinical situations. This variability of study outcomes can be reduced by patients' selection based on DSM-IV criteria, and by clinicians trained on standard depression assessment scales. A symptom reduction of at least 50%, for example, in the Hamilton Depression Rating Scale score, is widely used as an arbitrarily chosen but operational measure of treatment response. However, a 50% improvement may leave many patients with significant symptoms, and, paradoxically, it is more difficult to achieve a 50% response if the baseline severity of depression is high. As a consequence, antidepressant drug efficacy in less severely depressed patients is not easily demonstrable, although most of these patients require effective treatment.

The effects of antidepressive drugs are seen in all of the symptoms of major depression; however, vegetative symptoms, for example, sleep disturbances and loss of appetite, are often the most rapidly improved. When only a single global measure is used to assess depression severity or improvement, differential drug effects and differential response patterns in subgroups of patients are overlooked. Thus, the methodology used for the assessment of symptoms and symptom changes, as well as the distribution of prevailing symptoms in different patients' samples, can influence study results. There is still

a lack of studies considering these issues and other factors which can potentially influence treatment response, for example, patients' gender, age, or personality.

Furthermore, the efficacy of antidepressants is often confounded with the time course of improvement. There is still controversy over whether or not some antidepressants exert a faster onset of action. Survival–analytical approaches (Stassen et al., 1993), exponential recovery models (Priest et al., 1996), and pattern analyses (Quitkin, 1999) are statistical methods to assess differential dynamics of improvement under antidepressants. There are additional important clinical and methodological issues relevant to treatment with antidepressant drugs.

One of these issues is the question whether patients with mild or moderate depression should be pharmacologically treated. This issue is linked to the issue of reliable screening and diagnosis.

There is a lack of convincing trials on the comparative and combined effectiveness of biological and psychological treatments in depression. Despite the long experience in the use of antidepressants and psychotherapy, there is no evaluation standard for these issues.

Another controversy directly related to mental health policy is the recommendation of antidepressants to prevent manifestation of major depressive disorder at an early stage, when only prodromal or single symptoms are present. This issue also includes the questionable prescription of antidepressants for children and adolescents – a generally underinvestigated field.

Secondary prevention, that is, reducing the risk for recurrence and chronicity of depressive disorders, also raises the question of how long antidepressants should be prescribed after remission of the target symptoms. The potentially advantageous prophylactic use of antidepressants is challenged by the risk of triggering switches into manic states, particularly in bipolar disorder. Decisive studies are required to demonstrate which antidepressants have the lowest risk for mood switches and phase acceleration.

Conclusion and Future Prospects

Today, a respectable armamentarium for the treatment of depressive disorders is at hand, and antidepressants are standing in the front line. When used adequately, the available antidepressant drugs are effective in reducing an enormous burden of costs, distress, suffering, and death.

Notwithstanding the tremendous progress in the field of antidepressant drugs the ultimate goals have not yet been reached. The key objectives of antidepressive drug treatment remain as follows:

1. to reduce and ultimately to remove all signs and symptoms of the depressive syndrome in the largest possible proportion of patients in the shortest possible time;
2. to restore occupational and psychosocial function to that of the asymptomatic state; and
3. to attenuate and ultimately to eliminate the risk of relapse, recurrence, and chronicity.

There is still an urgent need to develop new and better treatments for depression, including drugs and other biological treatments, psychotherapeutic strategies, and combinations of some or all of these. The benefit of having a wide array of antidepressant drugs and other treatment options is based on evidence that there may be different forms of the illness, which respond to different mechanisms of action. More research is needed to develop predictors of differential responsiveness in order to tailor antidepressant treatment to the individual patient's requirement (Preskorn, 1994).

Depression is – among other neurological and psychiatric disorders – at present a treatable, but still little understood, disease. From the very beginning, the history of antidepressive drug treatment was accompanied by scientists' expectation that the knowledge about the mechanism of action of antidepressants would lead to an extended insight into the biological nature of the treated illnesses. In fact, the progress in research on antidepressants had an immense impact on biological psychiatry. However, there is still a lack of knowledge concerning the molecular understanding of the etiology and pathophysiology of depression, and the complex mechanisms of action of antidepressant drugs are still enigmatic.

Despite the chemical and pharmacological diversity of the presently available antidepressant drugs, they have some properties in common. The differences between antidepressant drugs lie mainly in side effects, tolerability, and the potential for pharmacokinetic interactions. Regarding the mechanism of action of the diversity of available antidepressants, a final common pathway is likely to be involved. A modulation of a number of different types of neurotransmitter receptors seems to contribute to the long-term effects of antidepressants, in most cases irrespective of the acute effects of the diversity of available drugs on a particular system. On a higher level of action, shared effects of different antidepressants on second-messenger systems and gene expression can be assumed but specific data are not available. For instance, an active cross-talk between serotonergic and noradrenergic receptors which can be modulated by a protein kinase has been demonstrated. Furthermore, a link between central glucocorticoid receptors and neurotransmitter receptor systems is likely.

Promising new candidates of alternative mechanisms of action for novel antidepressive drugs include inflammation, as mentioned above, and neurotrophic factors, especially following the recent evidence that the fast-acting antidepressant, ketamine, activates these mechanisms zunszain (Zunszain et al., 2013). It does not seem overoptimistic to expect that with the rapid progress of genetic and molecular research, the core pathophysiology of depressive disorders could be elucidated in the near future. This attainment alone could provide a rationale for a causal therapy of depression and related disorders and for logically designed antidepressant drugs with a clear therapeutic superiority over those available today.

Superiority should be in future understood not only in terms of higher efficacy, but also with regard to lower toxicity, better subjective tolerability and treatment adherence, improved functional outcome, and quality of life. The increasingly available evidence should serve as the basis for the prescription of a particular drug to a given patient. Long-term mental health issues for the development of better antidepressants should be the cost-effective reduction of patients' disability, morbidity, and mortality caused by depression and related disorders.

See also: Depression; Health Psychology in Childhood; Psychological Treatment, Effectiveness of; Psychotherapy and Pharmacotherapy.

Bibliography

Benkert, O., Laakmann, G., Ott, L., Strauss, A., Zimmer, R., 1977. Effect of Zimelidine (H 102/09) in depressive patients. Arzneimittel-Forschung 27, 2421–2423.

Benkert, O., Hippius, H., 1996. Psychiatrische Pharmakotherapie. Springer, Berlin.

Bloom, F.E., Kupfer, D., 1996. Psychopharmacology: The Fourth Generation of Progress. Raven Press, New York.

Carlson, A., Wong, D.T., 1997. A note on the discovery of selective serotonin reuptake inhibitors. Life Sciences 61, 1203.

Cipriani, A., Barbui, C., Brambilla, P., Furukawa, T.A., Hotopf, M., Geddes, J.R., 2006. Are all antidepressants really the same? The case of fluoxetine: a systematic review. Journal of Clinical Psychiatry 67, 850–864.

Feldman, R.S., Meyer, J.S., Quenzer, L.F., 1997. Principles of Neuropsychopharmacology. Sinauer Associates, Sunderland, MA.

Fond, G., Hamdani, N., Kapczinski, F., Boukouaci, W., Drancourt, N., Dargel, A., Oliveira, J., Le Guen, E., Marlinge, E., Tamouza, R., Leboyer, M., 2014. Effectiveness and tolerance of anti-inflammatory drugs' add-on therapy in major mental disorders: a systematic qualitative review. Acta Psychiatrica Scandinavica 129 (3), 163–179.

Khan, A., Warner, H.A., Brown, W.A., 2000. Symptom reduction and suicide risk in patients treated with placebo in antidepressant clinical trials: an analysis of the Food and Drug Administration database. Archives of General Psychiatry 57 (4), 311–317.

Kuhn, R., 1957. Über die Behandlung depressiver Zustände mit einem Iminodibenzylderivat G 22355. Schweizerische Medizinische Wochenschrift 87, 1135–1140.

Loomer, H.P., Saunders, I.C., Kline, N.S., 1957. A clinical and pharmacodynamic evaluation of iproniazid as a psychic energizer. Psychiatric Research Publications of the American Psychiatric Association 8, 129.

Nathan, P.E., Gorman, J.M., 1998. A Guide to Treatments That Work. Oxford University Press, New York.

Preskorn, S.H., 1994. Antidepressant drug selection: criteria and options. Journal of Clinical Psychiatry 55 (9 Suppl. A), 6–22.

Priest, R.G., Hawley, C.J., Kibel, D., Kurian, T., Montgomery, S.A., Patel, A.G., Smeyatsky, N., Steinert, J., 1996. Recovery from depressive illness does fit an exponential model. Journal of Clinical Psychopharmacology 16, 420–424.

Quitkin, F.M., 1999. Placebos, drug effects, and study design: a clinician's guide. American Journal of Psychiatry 156, 829–836.

Rush, A.J., 2011. Star-D: lessons learned and future implications. Depress Anxiety 28, 521–524.

Schatzberg, A.F., Nemeroff, C.B., 1998. Textbook of Psychopharmacology, second ed. American Psychiatric Press, Washington, DC.

Stassen, H.H., Delini-Stula, A., Angst, J., 1993. Time course of improvement under antidepressant treatment: a survival-analytical approach. European Neuropsychopharmacology 3, 127–135.

Stogner, K.A., Holmes, P.V., 2000. Neuropeptide-Y exerts antidepressant-like effects in the forced swim test in rats. European Journal of Pharmacology 387, R9–R10.

Wells, K.B., 1999. Treatment research at the crossroads: the scientific interface of clinical trials and effectiveness research. American Journal of Psychiatry 156, 5–10.

Zunszain, P.A., Horowitz, M.A., Cattaneo, A., Lupi, M.M., Pariante, C.M., 2013. Ketamine: synaptogenesis, immunomodulation and glycogen synthase kinase-3 as underlying mechanisms of its antidepressant properties. Molecular Psychiatry 18, 1236–1241.

Relevant Websites

http://www.bap.org.uk – British Association for Psychopharmacology.

http://www.ecnp.eu – European College of Neuropsychopharmacology.

http://www.rcpsych.ac.uk/healthadvice/treatmentswellbeing/antidepressants.aspx – Royal College of Psychiatrists Statement of Antidepressants.

Antisemitism

Robert Fine, University of Warwick, Coventry, UK
Glynis Cousin, University of Wolverhampton, Wolverhampton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses social work compliance with antisemitic policies under National Socialism and the strengths and limitations of addressing this record of compliance through ethical frameworks provided by professional bodies. It addresses factors that militate against learning from this history of compliance: textual silences about antisemitis, the radical separation of the present from the past, and the disconnection of antisemitism from other forms of racism. It confronts the difficult relation between antisemitism and the 'Israel question' and how professionals today might think about the challenges antisemitism continues to pose.

Introduction

We explore social work compliance with antisemitic policies under National Socialism and the strengths and limitations of addressing this record of compliance through ethical frameworks provided by professional bodies. We address three impediments to learning from history: textual silences about antisemitism, the divorce of the present from the past, and disconnection of antisemitism from other racisms. We address the relation between antisemitism and criticism of Israel and how professionals today might think about the challenges antisemitism poses.

Compliance: Antisemitism in Social Work under National Socialism

Sünker and Otto (1997) plot a radical shift under National Socialism from a system of welfare whose primary values were universalistic and based on the care of the individual, to one centered on education for the strengthening of the *Volk* community and the elimination of those categorized as 'useless.' Social work practice and policy accommodated this shift largely by donning a defensive professional mask rather than through enthusiastic embrace, although this mask was unavailable to Jewish social workers and other social workers who fell victim to political and antisemitic purges (Schnurr, 1997: 126; Burleigh, 1994: 170–171). The remaining social workers kept their jobs to administrate the demands of the new regime. They participated in National Socialist (NS) pedagogic projects to instil true German self-reliance instead of welfare, worked in educational camps and cultivated NS values about motherhood and community in maternity training programs.

NSs held that certain categories of the population (disabled, mentally defective, racially impure, sexual deviants, political dissenters) were 'useless' to the foundation of a new Germany, and that the ideologically and economically sound way of dealing with these groups was to confine, sterilize, or murder them. At the same time the 'useful' were to be supported to help grow a hygienic nation cleansed of contaminating groups. The sociologist Zygmunt Bauman invokes the metaphor of the gardener to capture this phenomenon: the gardener who selects and uproots weeds to allow other plants to flourish (Bauman, 1989: 91–93). Aly and Roth (cited in Sünker and Otto, 1997: 22) summarize the NS project thus: "the 'final solution to the Jewish issue' was part of a 'final solution' to the social issue." Reducing the population to a desirable *Volk* community had the secondary benefit of cutting the welfare budget so that funds could be diverted to nation building. So, how were social workers enmeshed in this cruel project?

When Stefan Schnurr (1997: 127–128) interviewed former social workers who had served under the NS regime, he found that they put forward apologetics centered on the defense of 'professionality.' They all said they did not like Nazism but were able to exploit its spaces. Here is one testimony:

> Although you did receive training at the *Gauschule*, you just didn't listen to what they were saying, right? The important thing was the commitment to people, to the mothers whom I had got to know in the factories [the speaker's previous field of work], and I knew that they urgently needed something where they would gain recognition, where they could get answers to their questions, and so on. And that is why I went into it positively.

Doing 'good' remained a possibility and a defense for this practitioner. He pleaded ignorance: "it was a long time before I heard about these negative things" (Schnurr, 1997: 128). This plea was in tension with the explicit duties placed on social workers to assist in population selection. Another former social worker presented her participation in the sorting process as a 'neutral administrative act' (Schnurr, 1997: 134),

> And among my clients, I was also a social worker for the health department, I knew a whole lot, a whole lot, well, more than two hands-full, you know, who were suffering from hereditary diseases. And I had to make a note of that. So I wrote down their names for my medical officer. There was no other way at the time. Because he has also had to produce something and was not a fervent National Socialist.
>
> Schnurr, 1997: 132

Such dissociation from the actuality of the task recalls Bauman's discussion of the role of 'distantiation' and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28005-7

'differentiation' in the administration of the Holocaust (Bauman, 1989: 98–102). By representing the list compilation in this fashion, this woman's retrospective account can hardly be judged to be reflexive. However, the predicament of social workers had to do with their specific positionality between state and client in public and private institutions. It was a predicament shared with other professionals – teachers, lawyers, judges, and doctors.

Many social workers played a critical role in the eugenic policies of National Socialism precisely because they were strategically placed in prisons, children's homes, reform schools, maternity wards, and family homes. They were able to support the sorting, selection, and ultimately sterilization and slaughter of those labeled disturbed, disabled, dangerous, or racially impure. Waaldijk (2011) writes that social workers' expertise in classifying individuals and families as worthy or unworthy during this period helped prepare the ground for genocide, even if their collaboration did not require positive adherence to NS ideals. From the testimonies, Schnurr gathered, many social workers were contemptuous of what they saw as "the boundless stupidity of the NS officials who did not understand anything about our work" (Schnurr, 1997: 133). This posture enabled professional distancing from the cruelties that accompanied this perceived stupidity. Schnurr concludes that adherence to National Socialism was not a condition of collusion with its aims. Collusion simply required the demarcation of a professional space – one in which it could be claimed that allegiance to the client was the prime determinant of actions and in which compliance with partial, demarcated tasks was dissociated from their functionality for genocide. Social workers, like everyone else, created masks for themselves and even after some distance from the past, following Schnurr's discussion in the 1990s with a group of social workers who had practiced under National Socialism, they could not lift the masks.

Dealing with the Past

As Sünker and Otto (1997) and Waaldijk (2011) stress, the collusion of social workers should not be treated as a closed episode but rather absorbed into the profession's sense of its own formative history. So too, should stories of social workers who heroically refused to cooperate with the Nazi regime. While most people complied, there were some who did not and relied on their own sense of professional judgment and common humanity. They could not do what they were instructed to do. We might call this, following Theodor Adorno, the 'primal phenomenon of moral behavior' that occurs when "the element of impulse joins forces with the element of consciousness to bring about a spontaneous act" (Adorno, 2006: 240). Johnson and Moorhead (2011) recount some inspiring examples, like that of Alice Salomon, one of the founders of German social work, who advocated peace, disarmament, and women's rights. She was deprived of her doctorate, which was one of the first studies of women at that time, and was ultimately expelled from Germany. The scale of compliance and collaboration of the social work profession with National Socialism has been the cause for reflection on the purpose and ethical foundation of social work practice and policy. It has become a profession now anxious to get it right, having once got it so horribly wrong.

In his survey of social work practice, Waadjik summarizes the profession's relation to the past in an interesting way,

> Social workers operate daily in societies shaped by the disasters and by the triumphs of the past, they work with clients whose lives, whose parents' lives and whose expectations are shaped by this past. In order to provide assistance to these clients, it is crucial that social workers are aware of the way their own professional ideals and practices have been shaped by these pasts as well… Social workers have in different circumstances behaved differently, some resisted, other assisted racist political regimes. Some were cogs in a machine of repression; others combined social work with struggles for survival, liberation and freedom. The remains of these choices continue to play a role in professional ideals – we'd better address them.
>
> Waaldijk, 2011

This self-aware and self-critical account of the profession's relation to its own past may be contrasted with the idealized accounts found in official statements of the profession, such as the *International Definition of Social Work* (presently under consultation) supported by the International Association for the Schools of Social Work (IASSW) and International Federation of Social Workers (IFSW, 2012b). It offers an altogether rosier version of the profession's formative history:

> Social work grew out of humanitarian and democratic ideals, and its values are based on respect for the equality, worth, and dignity of all people. Since its beginnings over a century ago, social work practice has focused on meeting human needs and developing human potential. Human rights and social justice serve as the motivation and justification for social work action. In solidarity with those who are disadvantaged, the profession strives to alleviate poverty and to liberate vulnerable and oppressed people in order to promote social inclusion. Social work values are embodied in the profession's national and international codes of ethics.
>
> IASSW, 2012

This international definition of social work exemplifies what Sünker and Otto refer to as "a mentality of repression" in the profession:

> Social work is today still largely disinterested in what it has inherited from NS … In this sense it moves within the same spectrum of mentalities of repression as other professions and disciplines.
>
> Sünker and Otto, 1997: 144

The choices made by German social workers in this dreadful period are not so much addressed as glossed over. This glossing of the past applies to other episodes of complicity in oppressive practices (Dominelli, 2002).

The Ethical Turn

Rather than address their highly equivocal *past*, the professional bodies of social work turned to ethical guidelines to shape *future* practice. There is general recognition that 'the loyalty of

social workers is often in the middle of conflicting interests' and that 'social workers function as both helpers and controllers.' The tension social workers face between help and control is handled through explicit commitments to human rights and professional responsibilities. For example, there is a statement of "responsibility to challenge negative discrimination on the basis of characteristics such as ability, age, culture, gender or sex, marital status, socio-economic status, political opinions, skin colour, racial or other physical characteristics, sexual orientation, or spiritual beliefs" (IFSW, 2012a). A duty is placed on social workers to "bring to the attention of their employers, policy makers, politicians and the general public situations where resources are inadequate or where distribution of resources, policies, and practices are oppressive, unfair or harmful" (IFSW, 2012a). Absent, is any confrontation with the historical failure of many social workers to live up to these responsibilities and/or perform these duties. The IASSW-IFSW Global Standards document (2012), currently under revision, mentions pressures toward complicity with apartheid policies once imposed on social work educators in South Africa, and states that the profession resisted these pressures. Nonetheless, the orientation of this document, like that of the others, is to the future. This is not in itself a problem but it risks displacing an ugly history of malpractice. Here the deployment of the future tense tends to repress that of the past.

It is noteworthy that there is no mention of antisemitism in these mission statements. The professional organizations align themselves to the *Convention on the elimination of all forms of racial discrimination* (1969), which declares inter alia that "any doctrine of superiority based on racial differentiation is scientifically false, morally condemnable, socially unjust and dangerous, and that there is no justification for racial discrimination, in theory or in practice, anywhere." There is no explicit mention of antisemitism but rather a reliance on the expression 'all forms of racism.' This has the advantage of not having to list categories of vulnerable groups – lists that inevitably exclude some groups and invite argument about which categories should be included. Yet, a question remains as to what is read into 'all forms of racism' and the place of antisemitism within this frame. We put this question because, if the notion of racism in social work as with other public services is associated with racism against groups suffering from social and economic discrimination, an unintended consequence of this approach may be to make invisible forms of antisemitism that do not fit into this frame. Some multicultural practices have included a concern with faith, including Judaism, which makes an antiracist contribution in terms of respect and tolerance for diverse faith communities, but this emphasis does not address other troubling and difficult forms of antisemitism (Levine, 2004; Sicher, 2009).

The invisibility of antisemitism can be discerned in the practices of ethnic monitoring in public institutions, which arrived in the UK from the US during the 1970s. Its purpose was to produce statistical evidence of access to social goods by ethnic group on which to make judgments about where discrimination might be occurring. This purpose stands on its head the kind of 'ethnic monitoring' characteristic of Nazism. A concession to those for whom ethnic monitoring continues to be a reminder of a sinister past, has been the addition of a 'prefer not to say' box – now largely replaced with a box marked 'Other.' Regardless of good intentions, the textual silence regarding antisemitism remains. This textual silence may create awkwardness for Jewish individuals, but more importantly, it may withhold recognition of the weight of the past on the present, thus presenting a tension. On the one hand, there is something troubling about omitting any specific mention of antisemitism from the language of antiracist ethics and ethnic monitoring. On the other hand, 'counting Jews' is part of the repertoire of the antisemite whose refrain is that Jews are overrepresented in places of control and privilege. The motive force behind contemporary ethnic monitoring is about underrepresentation of minorities, but this may be reversed if it becomes associated with suspicion of overrepresentation.

Reports of a curious incident in a Department of Social Work at Toronto University may illustrate the problem (Lightman, 2010). Some 'visible minority' students expressed discomfort at having to participate in a placement at a pioneering nursing home as part of their study and said they felt threatened by 'so many rich Jews.' The Professor to whom these objections were expressed is reported to have been sympathetic to this view and permitted the students to conduct a 'Jew count' of Jewish benefactors to the University as a marker of their undue influence. Complaints were made and as far as we know quietly resolved, but it is troubling that the students' discomfort over learning from 'rich Jews' was not itself more discomforting to some of the department's academics. We should not place too much weight on one contested incident, but it illustrates how antisemitic threads from the past can reemerge.

Social work has an ugly past regarding antisemitism. As Waaldijk (2011) urges, this past has to be addressed to understand how it reverberates in the present. The creation of universalistic and forward-looking ethical frameworks goes some way to addressing the problem but does not name the past. It condemns racial discrimination but does not explore why it happened once and could happen again.

Connecting and Disconnecting Antisemitism

There has been a tendency to present our age as having radically broken from the conditions of life that gave rise to antisemitism in the past, to give the story of European antisemitism a happy ending, and to say that other kinds of racism now existing in Europe – notably against Muslims, immigrants, and Roma – have replaced antisemitism (Beller, 2007; Bunzl, 2007). This tendency is underwritten by shifting representations of Jews in the post-Holocaust world: as part of the European white elite, and as victimizers of Palestinians. The effect of this shift is to render incidents of antisemitic hate speech, abuse, and violence more amenable to other interpretations – from allusions to Jewish oversensitivity to displacing the onus of responsibility onto the actions of Jews themselves (Community Security Trust, 2010).

Antisemitism has not simply vanished in the new, post-Holocaust Europe (Judt, 2007). Many scholars now reject the 'eternal antisemitism' approach, in which antisemitism appears as a permanent and natural feature of relations between Jews and non-Jews (Arendt, 2004: 7). However, the answer to an ahistorical frame of reference that eternizes antisemitism is not to turn it on its head and confine antisemitism to history, but to

explore the ways in which the forces that created antisemitism in the past may persist in the present. Antisemitism in the past provided an epistemic framework to explain social pathologies from defeat in wars to financial crises; this epistemic function persists into the present (Julius, 2010: 54–59, 483–488; Nirenberg, 2013; Jacobs, 2011). As Naomi Klein has written,

> [E]very time I log on to activist news sites like Indymedia.org which practise 'open publishing', I am confronted with a string of Jewish conspiracy theories about September 11 and excerpts from the "*Protocols of the Elders of Zion*".
>
> Klein cited in Jacobs, 2005

Old antisemitic typifications of 'the Jews' – that Jews are indifferent to the suffering of others, treacherous toward their host countries, protective of their own exclusive interests, and incapable of genuine universality – create racist myths parallel to those that claim the incapacity of black people to achieve moral or intellectual maturity or the incapacity of Muslim men to treat women as equals or to understand human rights (Finkielkraut, 2001: 46–53). The content of these typifications is different – Jews tend to be typified as bearers of an intellectually distorted modernity, black people as bearers of premodern savagery, Muslims as bearers of an antagonistic civilization – but in all cases, racism represents the same human bankruptcy.

The disconnection of antisemitism from racism has not been total (notable exceptions are Gilroy, 2001; Lapierre, 2011; Sundquist, 2005), but where it exists it has created significant difficulties of understanding. In the postwar period, professionals have tended to focus on the social disadvantages and discriminations suffered by black immigrants and their descendants. This focus has shown its worth by contributing to the amelioration of social conditions faced by them and the formation of a more inclusive democracy. However, the Jews who survived after the Holocaust have not figured strongly in this framework. Other factors also contributed to the disconnection of racism and antisemitism. These include the imperatives of intellectual specialism, where the largely separate development of Postcolonial and Holocaust Studies has engendered fruitful research in their respective areas but at the cost of constructing disconnected frames of analysis, and the emergence in some urban situations of antagonism between black and Jewish social actors that has served to naturalize their separation (Altglas, 2012; Cousin and Fine, 2012). The difficulties of sustaining the connectivity of racism and antisemitism can be exemplified in relations between two contemporary academic discourses: 'critical race theories' (Delgado and Stefancic, 2011) and 'new antisemitism theories' (for an overview see Judaken, 2008). Each represents a collective endeavor to recognize and resist 'new' forms of racism and antisemitism, which wear the masks of liberalism, radicalism, or even antiracism. While the strength of these theories has been to draw attention to covert and subtle forms of prejudice, the downside is to see racism and antisemitism, respectively as exclusive and ubiquitous (Fine, 2009).

The concern expressed by proponents of 'new antisemitism theory' is that antisemitism has entered into the depiction of Israel as a uniquely illegitimate state, Zionism as a uniquely noxious ideology, supporters of Israel as a uniquely powerful lobby, and memory of the Holocaust as a uniquely self-serving reference to the past (MacShane, 2008; Harrison, 2006; Hirsh, 2007). Behind what passes as 'criticism' of Israel, the fear is that there sometimes lurks a reconstruction of old antisemitic typifications in a new guise: blood libel (the charge that Jews murdered Christian children for their blood), global conspiracy, and indifference to the suffering of others. The fear is that antisemitism is being mediated through forms of political criticism of Israel and Jewish nationalism, whose premise is that 'the Jews should know better' and that a particular burden of moral responsibility rests on 'the Jews.'

Critics of new antisemitism theory (Karpf et al., 2008; Klug, 2004; Rose, 2005) have argued inversely that behind the charge of antisemitism lies an unwillingness to face up the human rights abuses committed by the Israeli state. The fear is that the value of the term 'antisemitism' is debased by its political misuse, and that it is being misused to stigmatize those who express solidarity with Palestinians. For these critics, concern over antisemitism is portrayed as a dishonest way of justifying the unjustifiable: the ethnic character of the Jewish state. In this vein, the Marxist philosopher, Alain Badiou has argued that the signifier 'Jew' is now used as the sign of a victim ideology which places Jews beyond reproach, renders invisible other racisms, declares that Israeli crimes must be tolerated, and accuses those who do not tolerate them (like Badiou himself) of antisemitism (Badiou, 2011). He and his coauthor maintain that, because an ethnically defined state can no longer be defended openly and honestly, Israel's defenders resort to the charge of 'antisemitism' to defend it covertly and dishonestly (Badiou and Hazan, 2011). Badiou treats new antisemitism theory not as an expression of concern over the reemergence of antisemitism, but as the signifier of a reactive consciousness that is on the side not only of Israel against Palestinians but also of all repressive, occupying powers against popular resistance.

Badiou's polemic entrenches the disconnection of antisemitism from other racisms. In order to construct a consistent antiracism, however, it is vital to deconstruct the terms of this debate. To this end we offer three observations. First, it should be no more controversial to say that *some* criticism of Israel is antisemitic than to say that *some* criticism of India or Zimbabwe is racist or that some criticism of Iran or Saudi Arabia is Islamophobic. We learn to distinguish between 'orientalist' and legitimate forms of criticism in all cases (Sen, 2006). In relation to criticism of Israel in particular, the European Union Monitoring Commission (2005) has put forward guidelines designed to inform judgment of when criticism of Israel is or is not antisemitic. It maintains, for example, that when criticism of the Israeli government is like criticism of any other government, it is not to be treated as antisemitic, but that criticism of Israel can, depending on context, become antisemitic through the use of certain prejudicial arguments: (1) the Nazification of Israel (e.g., stating that Jews treat Palestinians like the Nazis treated the Jews), (2) the pathologization of Jews (e.g., claiming that Jews have become indifferent to the suffering of other peoples), (3) the use of old antisemitic tropes (e.g., asserting that Zionists engage in a world conspiracy to protect Israel or that Israeli forces steal Arab body parts), (4) the erasure of the distinction between state and civil society (e.g., saying that all Jews in Israel are responsible for the policies pursued by the

government), or (5) systematic treatment of Israel as culpable by standards not applied equally to other states. As the title 'working definition of antisemitism' suggests, these guidelines are presented as provisional, not final (Klug, 2004).

Second, both within new antisemitism theory and critical race theory there has arisen a temptation to treat the other as an adversary and homogenize and stereotype its views: for example, to declare that new antisemitism theory is racist toward Muslims or that critical race theory is racist toward Jews. Behind these mutual recriminations, however, both bodies of thought address real and related problems: the new forms of racism and antisemitism that have emerged in modern times but do not speak their name.

Third, there is a professional and moral obligation to take any charge of racism or antisemitism seriously. There are some who address racism and antisemitism from an exclusively 'national' point of view, and some who instrumentalize the charge of 'racism' and 'antisemitism' for their own economic or political ends. However, the ethical obligation is not to dismiss the problem of racism and antisemitism for these reasons. It is a paradox that those who criticize new antisemitism theory for its nationalism may nonetheless hold the view that the nationalism of the oppressed is in general an appropriate way to combat racism. To slip from a general critique of exclusivity in antiracist struggles to the critique of new antisemitism theory as the embodiment of exclusivity has specific echoes of an old prejudice concerning the exclusivity of Jews (Hirsh, 2007).

Conclusion

All professionals, including social workers, have the responsibility critically to challenge a conception of compliance that accepts whatever happens to be a policy at a given time and place, including antisemitic policies. To recognize antisemitism requires thought and understanding. This involves questioning the common sense notion that antisemitism is so readily identifiable as contrary to antidiscrimination policies that there is no further need to become reflective about what it is (Sartre, 1995). It is not enough simply to look inward to ask if one's own actions or motives are antisemitic. This brief survey suggests the importance of reflecting on the lessons of a past in which there was compliance with antisemitic policies, and hearing the voices of those who feel themselves to be subjected to discrimination without prejudging the validity of their claims (Gidley, 2011). To this end we have urged increased connectivity between our understandings of racism and antisemitism, which have their own distinctive characteristics but are best understood in relation to one another.

Acknowledgments

The authors wish to thank Gurnam Singh for signposting us to scholarship on social work and National Socialism, and Lena Dominelli for her insightful and careful editorial comments.

See also: Oppression; Social Work Theory; Social Work in Eastern Europe.

Bibliography

Adorno, T.W., 2006. History and Freedom: Lectures 1964–65. Polity Press, Cambridge.
Altglas, V., 2012. Antisemitism in France. European Societies 14 (2), 259–274.
Aly, G., Roth, H.K., 1984. Die Restlose Erfassung: Volkszählen, Identifizieren, Aussondern in Nationalsozialismus. Rotbuch, Berlin.
Arendt, H., 2004. The Origins of Totalitarianism. Schocken Books, New York.
Badiou, A., 2011. The Uses of the Word "Jew". file:///Users/robertfine/Documents/Antisemitism/Articles/Alain%20Badiou%20-%20Uses%20of%20the%20Word%20%22Jew%22.webarchive (accessed 09.01.14.).
Badiou, A., Hazan, E., 2011. L'Antisémitisme Partout. La Fabrique, Paris.
Bauman, Z., 1989. Modernity and the Holocaust. Polity Press, Cambridge.
Beller, S., 2007. Antisemitism: A Very Short Introduction. Oxford University Press, Oxford.
Bunzl, M., 2007. Anti-semitism and Islamophobia: Hatreds Old and New in Europe. Prickly Paradigm Press, Chicago.
Burleigh, M., 1994. Death and Deliverance: 'Euthanasia' in Germany 1900–1945. Cambridge University Press, Cambridge.
Community Security Trust (CST), 2010. Antisemitic Incidents Report 2010. CST, London.
Cousin, G., Fine, R., 2012. A common cause: reconnecting the study of racism and antisemitism. In: Achinger, C., Fine, R. (Eds.), European Societies, Special issue on Racism and antisemitism in contemporary Europe, vol. 14 (2), pp. 166–185.
Delgado, R., Stefancic, J., 2011. Critical Race Theory: An Introduction. NYU Press, New York.
Dominelli, L., 2002. Anti-oppressive Social Work Theory and Practice. Palgrave Macmillan, London.
European Union Monitoring Commission, 2005. Working Definition of Antisemitism. http://fra.europa.eu/sites/default/files/as-workingdefinition-draft.pdf (accessed 09.01.14.).
Fine, R., 2009. Fighting with Phantoms: contribution to the debate on antisemitism in Europe. Patterns of Prejudice 43 (5), 459–479.
Finkielkraut, A., 2001. In the Name of Humanity: Reflections on the Twentieth Century. Pimlico, London.
Gidley, B., May 26, 2011. The Case of Antisemitism in the University and College Union. Engage. http://engageonline.wordpress.com/2011/05/26/ben-gidley/ (accessed 09.01.14.).
Gilroy, P., 2001. Between Camps: Nations, Culture and the Allure of Race. Penguin, Harmondsworth.
Harrison, Bernard, 2006. The Resurgence of Antisemitism: Jews, Israel and Liberal Opinion. Rowman and Littlefield Publishers.
Hirsh, D., 2007. Anti-zionism and Antisemitism: Cosmopolitan Reflections. YIISA Working Paper 1, Yale Initiative for the Study of Antisemitism, New Haven, CT.
International Association for the Schools of Social Work, 2012. Definition of Social Work. http://www.iassw-aiets.org/index.php?option=com_content&task=blogcategory&id=27&Itemid=50 (accessed 09.01.14.).
International Federation of Social Workers, 2012a. Statement of Ethical Principles. http://ifsw.org/policies/statement-of-ethical-principles/ (accessed 09.01.14.).
International Federation of Social Workers, 2012b. Global Standards. http://ifsw.org/policies/global-standards/ (accessed 09.01.14.).
Jacobs, S., 2005. Antisemitism and other forms of racism: continuities, discontinuities (and some conspiracies…). In: CRONEM: The Future of Multicultural Britain. Roehampton University, pp. 14–15.
Jacobs, S., 2011. Globalisation, anti-globalisation and the Jewish question. European Review of History 18 (1), 45–56.
Johnson, S., Moorhead, B., 2011. Social eugenics practices with children in Hitler's Nazi Germany and the role of social work: lessons for current practice. Journal of Social Work Values and Ethics 8 (1). http://www.socialworker.com/jswve/spr11/spr11johnson.pdf (accessed 09.01.14.).
Judaken, J., 2008. So What's new? Rethinking the 'New antisemitism' in a global age. Patterns of Prejudice 42 (4–5), 531–560.
Julius, A., 2010. Trials of the Diaspora: A History of Antisemitism in England. Oxford University Press, Oxford.
Judt, T., 2007. Postwar: A History of Europe since 1945. Pimlico, London.
Karpf, A., King, B., Rose, J., Rosenbaum, B. (Eds.), 2008. A Time to Speak Out. Verso, London.
Klug, B., Feburary 2, 2004. The Myth of the New Antisemitism. The Nation.
Lapierre, N., 2011. Cause Commune: Des Juifs et des Noirs. Stock, Paris.
Levine, B., 2004. Antisemitism and anti-racist social work. In: Lavalette, M., Penketh, L. (Eds.), Race, Racism and Social Work. Policy Press, Bristol, pp. 85–113.

Lightman, E., 2010. Antisemitic at the University of Toronto. Journal for the Study of Antisemitism 2 (2), 363–376.
MacShane, D., 2008. Globalising Hatred: The New Antisemitism. Weidenfeld and Nicolson, London.
Nirenberg, D., 2013. Anti-Judaism: The Western Tradition. W.W. Norton, New York.
Philips, A., 2010. What's wrong with essentialism? Distinktion: Scandinavian Journal of Social Theory 11 (1), 47–60.
Rose, J., 2005. The Question of Zion. Princeton University Press, Princeton.
Sartre, J.-P., 1995. Antisemite and Jew: An Exploration of the Etiology of Hatred. Schocken Press, New York.
Schnurr, S., 1997. Why did social workers accept the new order? In: Sunker, H., Otto, H.-U. (Eds.), Education and Fascism: Political Identity and Social Education in Nazi Germany. Falmer Press, London, pp. 121–143.
Sen, A., 2006. Identity and Violence: The Illusion of Destiny. Penguin, Harmondsworth.
Sicher, E., 2009. Multiculturalism, Globalization and Antisemitism: The British Case. In: Analysis of Current Trends in Antisemitism, vol. 32. Hebrew University, Vidal Sassoon Center for the Study of Antisemitism, Jerusalem.
Sundquist, E., 2005. Strangers in the Land: Blacks, Jews, Post-Holocaust America. Harvard University Press, Cambridge, MA.
Sünker, H., Otto, H.-U. (Eds.), 1997. Education and Fascism: Political Identity and Social Education in Nazi Germany. Falmer Press, London.
United Nations, 1969. International Convention on the Elimination of All Forms of Racial Discrimination. http://www.unhcr.org/refworld/docid/3ae6b3940.html (accessed 09.01.14.).
Waaldijk, B., 2011. Social work between oppression and emancipation: histories of discomfort and inspiration in Europe. Social Work and Society 9 (2). file:///Users/robertfine/Documents/Antisemitism/Articles/Social%20Work%20Between%20Oppression%20and%20Emancipation.%20Histories%20of%20Discomfort%20and%20Inspiration%20in%20Europe.%20.webarchive (accessed 09.01.14.).

Antisocial Behavior

Stephen Crossley, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Issues of 'deviance,' 'delinquency,' 'disorder,' and 'incivilities' have occupied politicians and policy makers for many years. The current widespread policy focus on 'antisocial behavior' draws heavily on theories, which advocate early intervention for low-level nuisance and disorder, to break perceived links with more serious subsequent criminal behavior. This 'widening of the net' of forms of social control associated with these issues has necessitated that social workers curb the behavior of alleged unruly people. This control function potentially places practitioners at odds with widely held social work ethics and value and has implications for social work practice and education.

> Sociology tended to rest heavily upon the idea that deviance leads to social control. I have come to believe that the reverse idea, i.e. social control leads to deviance, is equally tenable and the potentially richer premise for studying deviance in modern society.
> (Lemert, 1972: p. ix)

> What to some are the highest reaches of the welfare state are to others the furthest extension of social control.
> (Lipsky, 1980: p. 11)

Introduction

Issues such as 'deviance,' 'delinquency,' 'disorder,' 'incivilities,' and 'nuisance' have occupied theorists, researchers, politicians, and policy makers for many years. In 1959, Barbara Wootton suggested that:

> No-one can embark on a discussion of antisocial behavior (ASB) without making assumptions as to the criteria by which any specific actions are defined as such; and those assumptions are bound to reflect, not only the norms of a particular culture, but in some degree also, the subjective preferences of the person who makes them. What is understood to be unacceptable behavior varies from place to place and from time to time; and even where standards are much the same, actual manifestations will vary.
> (Wootton, 1959: p. 13)

This article explores some of these concerns in more detail, beginning with what antisocial behavior (ASB) is understood to involve at the current time and where its most recent origins can be found. The focus on ASB spreads across continents and some of the similarities and differences in approaches across cultures are also explored. The differences between them can be partly explained by the influence of wider social policies and welfare systems, which differ between countries and which themselves are borne out of different political and economic visions and circumstances. The withdrawal from a wider social solidarity project to a narrower vision of maintaining order has implications for social work practice and these tensions are explored, making use of an extensive literature that highlights the competing roles that social workers are often asked to carry out. Examples are drawn from a number of different countries and from a wide range of sources to highlight the international and historical saliency of this policy area. However, the robust approach to ASB in the United Kingdom is utilized to examine some points in greater depth.

What Is ASB?

ASB is currently defined under UK law as "acting in a manner that causes or is likely to cause harassment, alarm or distress to one or more persons not of the same household as the person" (Anti-Social Behaviour Act, 2003: p. 26). A new ASB, Crime, and Policing Bill introduced to the UK Parliament in 2013 includes proposals covering a wide range of areas, including dangerous dogs, firearms offences, forced marriage, policing reforms, extradition, and criminal justice, as well as ASB. It is perhaps not surprising, therefore, that this 'version' of 'ASB' has encountered some criticism for the breadth of its coverage. Burney has called it an "elastic concept" (2005: p. 7) and a "hydra-headed monster" (p. 16) while Ashworth has suggested it as "a vague term with a broad definition, which in the last few years has become a rallying call for some onerous and intrusive measures against individuals" (2004: p. 263).

The history of the concept of ASB and related terms has been covered extensively elsewhere (see, e.g., Morris, 1994; Burney, 2005; Welshman, 2006; for histories relevant to the United Kingdom). It is, however, notable that there are some strong similarities between the current focus on 'ASB' and previous historical social concerns about the behavior and/or depravity of certain social groupings. For example, it has been suggested that the contemporary media and political concern surrounding ASB hold a number of similarities, in both application and appeal, to the widespread public anxiety about the so-called dangerous classes in Victorian times (Burney, 2005; Morris, 1994) and it has been argued that these concerns tend to be periodic (Cohen, 2002).

The present focus on low-level nuisance and disorder owes much to the 'Broken Windows' thesis, which originated in the United States (Kelling and Wilson, 1982) and the accompanying policy response of 'zero-tolerance policing,' which was credited with being the primary factor in a significant reduction of crime and nuisance behavior in New York. Kelling and Wilson suggest that there is general agreement between social psychologists and police officers that if a broken window is left unrepaired for a period of time, other windows will get broken as well. These authors argue that 'because some areas are

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28006-9

inhabited by determined window breakers,' people are not only fearful of a sudden, violent attack in public places, they are also fearful of:

> being bothered by disorderly people. Not violent people, nor, necessarily, criminals, but disreputable or obstreperous or unpredictable people: panhandlers, drunks, addicts, rowdy teenagers, prostitutes, loiterers, the mentally disturbed.
> (Kelling and Wilson, 1982: p. 1)

The need to 'maintain order' therefore becomes a policy priority. In this scenario, low-level nuisance cannot be left unchecked for fear of informal community controls breaking down and the precursory warning signals of more serious crime and disorder need 'nipping in the bud.' Despite a number of critics questioning this deterministic view, the concept has traveled well. Aided by wider processes of globalization, the neoliberal and international 'policy market' has helped to ensure that a number of different countries have adopted and adapted different aspects of the zero-tolerance approach to suit their own particular needs and systems.

Burney (2005) draws on a wide range of research to highlight a number of different examples from Europe, including Holland, Sweden, France, and Austria. While there are undoubtedly similarities between some of the policies and approaches used to tackle nuisance behavior, there are also differences. In the original Broken Windows article, Kelling and Wilson note that the majority of the foot patrol officers are white while the 'people on the street' were usually black. These people were then 'sorted' into 'regulars' who the police officers recognized and 'strangers,' who they did not know. The similarities with some countries on mainland Europe are apparent as racial tensions, immigration, and the perceived lack of integration of relatively new, predominantly Muslim communities (especially young men) that have been main drivers of new legislation in that continent. However, the primary target of extended ASB legislation in the United Kingdom has been white working class youths, stereotypically located in areas of industrial decline, deprivation, and unemployment (Burney, 2005: p. 162), although black and minority ethnic groups are overrepresented in penal establishments. Another marked difference between countries is the extent to which state agencies get involved and the way in which they do so. It has been noted, for example, by Burney (2005: p. 157) that Swedish welfare agencies will intervene early to offer support to families and that their system does not include 'purely negative' instruments such as the Antisocial Behavior Order (ASBO) (Anti-Social Behaviour Act, 2003: p. 62) that form part of the UK legislative arsenal. Burney also notes that there appears to be genuine surprise and bemusement among Swedish welfare agencies regarding the politically motivated deployment and 'reach' of ASB legislation in England and Wales.

It has been suggested that, in the United Kingdom, New Labour 'invented' ASB but others have instead proposed a 'rediscovery' of ASB. Drawing on the claim by a former Prime Minister, Tony Blair, Squires and Stephen argue that when he first used the phrase in 1988, "anti-social behaviour did not arise, perfectly formed, in the mind of this politician then. Rather, it reemerged as a result of a complex combination of influences" (Burney, 2005: p. 20) including a focus on social exclusion and victimization and the residualization of social housing. Tracing the development of the concept through the 1980s and 1990s, Squires and Stephen also highlight the 'Broken Windows' influence, and suggest that, by arguing for the need to break the links between juvenile delinquency and persistent adult criminality and 'intervene early,' Labour merely succeeded in strengthening already perceived links in the minds of the electorate. This move placed Labour in the center of the 'law and order' debate initiated earlier by the Conservative government, especially under Michael Howard as Home Secretary, about being 'tough' on law breakers, as it maneuvered for electoral success.

As Squires and Stephen note, the 'rediscovery' of the term 'ASB' fitted with the wider New Labour focus on 'social exclusion,' which was subsequently critiqued as a vague concept, which houses competing discourses (Levitas, 1998). The concept has also been useful in mobilizing communitarian government projects such as New Labour's Together and Respect Agendas, which called on 'active citizens' to utilize and develop their social capital and play a role in supporting state agencies to tackle ASB and instill the old fashioned virtue of respect in an apparently wayward younger generation.

The links between 'ASB' and 'social exclusion' highlight that criminal justice policies are not developed in isolation, away from other social and economic policies. David Garland has highlighted that "the same premises and purposes that transformed criminal justice are evident in the programmes of 'welfare reform' that have been adopted by governments (and opposition parties) on both sides of the Atlantic" (Garland, 2001: p. 196). He argues that the social democratic 'solidarity project' of the middle decades of the last century has given way to a "more reactionary, less ambitious one of reimposing control" (2001: p. 199). This has led to poor people and criminals being again "viewed as members of a culturally distinct and socially threatening 'underclass,' in which all of the pathologies of late modern life are concentrated" (2001: p. 196). Mobilizing the concept of 'othering' (see Lister, 2004 for a discussion on this specifically in relation to poverty), he argues that "so long as offenders and claimants appear as 'other,' and the chief source of their own misfortune, they offer occasions for the dominant classes to impose strict controls without giving up freedoms of their own" (2001: p. 198).

This 'labeling' of others as somehow different and, more usually, 'deviant' (Becker, 1963) has been extended, in the United Kingdom at least, from a focus primarily on ASB to include the other 'pathologies' that Garland alluded to, with a range of perceived familial failings identified as worthy of state intervention. This is most acutely demonstrated in policy programmes designed to 'tackle' 'troubled families' or 'families with multiple disadvantages,' in which ASB features prominently.

Following the riots in a small number of towns in England in 2011, something akin to a moral panic ensued and 120 000 families that had previously been identified as having 'multiple disadvantages' became known as 'troubled families,' with a large centrally organized programme operationalized

to 'turn their lives around.' The UK Prime Minister, David Cameron, stated that his 'mission in politics' was 'fixing the responsibility deficit,' almost perfectly articulating what Garland has called the "responsibilization strategy" (2001: p. 124). In launching the Troubled Families Programme, Cameron said:

> Officialdom might call them 'families with multiple disadvantages'. Some in the press might call them 'neighbours from hell'. Whatever you call them, we've known for years that a relatively small number of families are the source of a large proportion of the problems in society. Drug addiction. Alcohol abuse. Crime. A culture of disruption and irresponsibility that cascades through generations.
> (Cameron, 2011)

Critics of the Troubled Families Programme have noted that the original figure of 120 000 families was identified from a government report exploring families with multiple disadvantages, such as maternal mental health issues, low income, poor quality housing, and low parental skills or qualifications. However, the current group of 'troubled families' are identified as those that have an adult on out-of-work benefits, where children are not attending education and where crime and/or ASB is taking place. This highlights the view, as articulated by Cohen, that "The 'deprived' are not very different from 'the depraved'" (1985: p. 60).

Each 'troubled' family is allocated a 'keyworker' from a local state or voluntary sector welfare agency to help them 'turn their lives round,' with a strong emphasis on the importance of locally developed solutions, highlighted by a lack of statutory guidance supporting the programme. Drawing on the work of Cohen again, the Troubled Families Programme therefore represents a good example of a 'flexible system' where 'Crime and delinquency nets thus not only become blurred in themselves but get tangled up with other welfare, treatment and control nets.' Cohen goes on to suggest, perhaps unfairly, that, "In Britain, social workers are the most powerful human service professionals operating in these waters" (1985: p. 61).

Care or Control?

As I have noted above, the development of social policy does not take place in an intellectual or political vacuum and it therefore follows that these policies cannot be applied without consideration of broader social, economic, and political concerns. Policies need to be implemented – rules need enforcers – and social workers, in a broad sense, are the professionals tasked with implementing a range of social and welfare related policies. Tensions can, however, occur when professionals whose primary role is to ensure the welfare of their clients are tasked with carrying out the function of 'controlling' certain kinds of individuals or families and/or their behaviors.

In his classic work *The Policing of Families*, Jacques Donzelot (1997) argued that social workers were gradually taking over from teachers in civilizing members of society. He suggested that, although social workers are "scattered throughout a multiplicity of inscription sites" they focused on what he called the 'less favored' classes' and, more specifically:

> the pathology of children in its dual form: children in danger – those whose upbringing and education leaves something to be desired, and dangerous children, or delinquent minors.
> (Donzelot, 1997: p. 96)

Originally writing in 1977, Donzelot predicted the current focus on 'prevention' and 'early intervention' via an extensive state apparatus. He argued that so-called experts would look at delinquents' first encounters with state agencies, examine their families' backgrounds and then draw up a 'portrait' of their future life. Following this process of examination, "An infrastructure of prevention will then be erected around him (sic) and an educative machinery will be set into motion, a timely action capable of stopping him short of a criminal violation" (1997: p. 97).

This extension of the state apparatus and the corresponding increased interest in low-level nuisance has led to what Squires has called a "disciplinary welfare system" (1990: p. 1) where some of the tensions between the care and control functions of social policy – working 'with' families or working 'on' them – have become increasingly apparent. Drawing on Foucault's work around discipline and governmentality and Bentham's notion of a "disciplinary continuum" (1990: p. 49), Squires suggests that "a generalised terrain of social problems came to be the accepted target of a regime of state interventions" (Squires, 1990: p. 47) and that these interventions are increasingly being carried out by social workers.

This expansion and dispersal of discipline means that social workers themselves, who are often employed or funded by the state and therefore expected to carry out the functions of the state, have been drawn into potential roles in the disciplinary net (Cohen, 2002). These roles conflict with the original aims and the wider ethics and values of social work and have led to social workers being called "reluctant policemen" (Burney, 2005: p. 115). 'Russian dolls' enveloping an initial judicial state model (Donzelot, 1997: p. 98) and "agents of social control, disguised storm troopers of the state" (Cohen, 1985: p. 130). Bourdieu argues, somewhat more sympathetically that these 'minor civil servants' have work "shot through with the contradictions of the state" (Bourdieu, 1999: p. 184). They become the left hand of the state "whose right hand no longer knows, or worse, no longer wants, what the left hand is doing" (Bourdieu, 1999: p. 183).

In his book *The Deviant Imagination*, Pearson (1975) explores the 'ambiguous politics of social work' in more detail and refers to social workers as 'social policemen,' 'social tranquilizers' (1975: p. 133), and 'professional Robin Hoods' (p. 134) who take part in 'moral hustling' (Pearson, 1975: p. 136). He argues that attempts to 'start where the client is' risk ignoring the influence of the state in locating individuals in certain subject positions for 'the sake of administrative convenience' (Pearson, 1975: p. 128). The nature of a 'dual commitment' to society as well as the individual is explored and this commitment and accountability to a wider 'society' is, according to Pearson, poorly articulated by the social work profession.

Using more robust language, Pearson goes on to argue that:

> Social work perpetually dithers about whether it represents the technical business of deviance-control, or some kind of artistic enterprise built on faith.
> (Pearson, 1975: pp. 129–130)

> These ambiguities mirror social work's ambiguous relationship to social and political structure, and the question of whether social work represents a personal counselling service for active, free willed human agents or a mopping up operation for society's 'dregs'.
> (Pearson, 1975: p. 130)

These views, articulated in the 1970s are still relevant today. If we return to the UK example of the Troubled Families Programme, similar tensions become apparent. Social workers in the United Kingdom have been accused by Ministers and senior civil servants of colluding with families who are allegedly 'fluent in social work,' or of being 'politically correct' and 'naïve' when highlighting poverty as an issue when dealing with behavioral problems. They have been portrayed as lacking authoritative challenge and numerous high profile child protection cases and reports have highlighted 'missed opportunities' for 'intervention' by a range of welfare agencies, usually with a strong focus on children's services (see, e.g., Munro, 2011). Social workers have been made aware of the potential to use 'sanctions' such as having children removed into care, withdrawal of benefits and eviction from homes. In a speech to the Association of Police and Crime Commissioners in 2013, the Secretary of State for Communities, Eric Pickles said:

> And the challenging, authoritative voice of the police is crucial. Some of the most successful family intervention projects are those where the police are heavily involved. Because sometimes it's only when a family is truly confronted with consequences - whether that's the threat of eviction, of having kids taken into care, or criminal proceedings - that they start taking things seriously.
> (Pickles, 2013)

In this political climate, it is understandable that social workers often feel compelled to carry out the wishes of the state without critically challenging or engaging with the policies they are operationalizing. However, the tensions between the need to care for certain groups and the need to control the behavior of some, sometimes overlapping, groups do not just exist between a centralized state and a localized delivery workforce. Tensions and differences in approach can be found within and between various local agencies tasked with participating in multiagency approaches to tackling ASB. Managers within these services are sometimes viewed as being overly bureaucratic and unsympathetic to the complexities of frontline social work. However, the main tensions often arise between services tasked with working together on ASB issues, which often include housing providers, children's services, neighborhood wardens, environmental officers, education professionals, and representatives from probation and criminal justice agencies.

Drawing again on UK examples, recent work has highlighted some of the problems of the 'multiagency' and 'partnership' approaches to tackling ASB. Burney (2005) has indicated the difference that demography, geography, and history can play in localized approaches to ASB, while Matthews and Briggs (2008) have noted the varying levels of significance attached to different forms of ASB by different agencies and organizations. In a study of the approaches used by local Crime and Disorder Reduction Partnerships in three different areas in the United Kingdom, they identify two primary divisions: the first being between the community safety teams and housing departments and the second being between what were loosely defined as enforcement and welfare or caring agencies. They argue that there are "pronounced differences in the attitudes toward ASB among youth services, social services, and education departments on the one hand and the police on the other" (Matthews and Briggs, 2008: p. 93). These different attitudes often led to frustrations with the approach of some of the partner organizations, with police often viewed as being 'heavy handed' and unsympathetic and welfare and social workers being perceived as too soft and potentially obstructive to enforcement action being taken against individuals.

Edwards and Hughes (2008: p. 67) have argued that such differences lead to the brokering of 'satisficing' deals between the various agencies that recognize the 'various statutory, occupational, and populist pressures to which they are subject.' In some cases these deals have not been achieved and Edwards and Hughes note that, as a result of competing statutory duties for different agencies, the issuing of ASBOs has been challenged by some local authority social services departments 'as counter to the welfare of those children subject to these orders and, therefore, in contravention of section 1 of the 1989 Children Act' (Edwards and Hughes, 2008: p. 67).

Similar tensions can also be found in other countries where social workers are caught up in the expectations of local citizens and the policy responses identified by politicians and policy makers. The Netherlands has been highlighted as an example where a strategy of 'responsibilization' (Garland, 2001) has also been implemented with different public bodies expected to play their part in tackling crime and nuisance. The traditional view of Dutch society as tolerant and lenient has been contrasted with a tougher approach increasingly concerned with perceived deviation from Dutch norms and standards and which has argued that "social workers should focus on social cohesion and diversion of juveniles from the street – so participating with police to reduce feelings of insecurity and promote social control" (Burney, 2005: p. 146).

Implications for Practice

The potential for social workers to have 'split personalities' has been extensively documented and there is often a disconnection between the values and ethics associated with social work and the functions they are asked to carry out in everyday practice. The International Association of Schools of Social Work (IASSW) and the International Federation of Social Work (IFSW) state that the profession "promotes social change, problem solving in human relationships and the empowerment and liberation of people to enhance well-being." It goes on to define the ethics of social work as "based on respect for the equality, worth, and dignity of all people," "focused on meeting human needs and developing human potential" and states that "In solidarity with

those who are disadvantaged, the profession strives to alleviate poverty and to liberate vulnerable and oppressed people in order to promote social inclusion" (IFSW, 2013).

The IASSW and the IFSW recognize that social workers act as both helpers and controllers and these values, as we have seen, are often forced to 'compete' or 'collaborate' with the goals and aims of other parts of the state, especially when 'multiagency' approaches are lauded as the most effective way of delivering public services. Far from helping individuals to flourish and reach their full potential, social workers are often expected to help curb behavior that is viewed as unacceptable by the state, often among already disadvantaged and marginalized groups. Drawing on examples from Australian settings, Briskman has highlighted how these competing accountabilities can sometime lead to social workers implementing dominant and repressive policies. She argues that "A benign acceptance of practice norms can take hold, with social workers focusing on 'doing good' without critically reflecting on underlying ideologies that perpetuate inequalities and injustices" (Briskman, 2013: p. 51). Briskman advocates a 'dual loyalty continuum' as an explanatory framework for negotiating the opposing roles of control agent and care agent.

The 'dual loyalty continuum' highlights the shifting and dynamic nature of much of social work. The deep complexities of social work ensure that workers are constantly engaged in negotiating processes with clients and their families, colleagues within departments, and partners from other agencies. Depending on how their work with service users is progressing and the expectation or demands of fellow professionals, social workers will need to adopt different positions on this continuum and the question of care or control is not one which lends itself to an 'either – or' solution.

The concept of 'street-level bureaucrats' (Lipsky, 1980) is potentially useful in exploring the dilemmas faced by frontline workers. Lipsky highlights the room that social workers often have to maneuver and argues that most people 'encounter' government through their interactions with 'street-level bureaucrats' such as social workers, teachers, police officers, nurses, health visitors, and so on as opposed to more formal interactions with politicians and more senior bureaucrats in the traditional sense. These encounters are more frequent for poorer people as more affluent individuals and families can purchase private services in respect of health, education, housing, and other areas and there is consequently less of a perceived need to change or control their behavior as they are not seen as a drain on public resources. Lipsky (1980: p. 8) highlights that "the policy delivered by street-level bureaucrats is most often immediate and personal" and that decisions have implications for how people relate to and are perceived by their neighbors, and how they perceive and think of themselves:

> Clients of street-level bureaucrats respond angrily to real or perceived injustices, develop strategies to ingratiate themselves with works, act grateful and elated or sullen and passive in relation to street-level bureaucrats' decisions.
> (Lipsky, 1980: p. 9)

Despite an increasing interest in surveillance and control, opportunities do still exist that allow the left hand of the state to operate without the full knowledge or approval of the right hand, to use Bourdieu's analogy. Neoliberal discourses around 'freeing up' resources, 'local solutions,' and 'empowerment' provide space for social workers and street-level bureaucrats to develop alternative narratives, often with the support and inclusion of service users. Where marginalized individuals or groups struggle to have their voices heard by those in positions of authority, social workers can play a role in supporting them to do this and, where necessary, advocate on their behalf. This approach has long been recognized in relation to welfare support and advice from social workers but advocacy also takes place across different issues such as disability rights, legal advice and support, discriminatory practice, and so on. As was highlighted earlier, this role has been enacted in respect of ASB by some social service departments challenging the issuing of ASBOs through a legal process as it was viewed as being counter to the welfare of the children involved (Edwards and Hughes, 2008: p. 67).

As well as having considerable scope within the rules that they are expected to work within, Pearson (1975) has argued that bending or breaking the rules is the only way that social workers are often able to achieve the aims of their profession. He suggests that 'industrial deviance' is alive and well in the profession. Bartley (2006) has also carried out research in the United Kingdom that highlighted how welfare professionals who went above and beyond what was necessary for their job were highly valued by clients/service users and made a lasting and positive difference to their lives. However, she notes that:

> These were the exception, however, and when interviewed it was often the case that state welfare workers felt that their approach was despite their employing agency, rather than positively endorsed by the agency.
> (Bartley, 2006: p. 23)

The implications of these tensions extend beyond social work practice, and also resonate in the realms of social work education. Garrett has highlighted how, in the United Kingdom, government documents and reviews of social work practice have helped to focus on social work as a 'practical job' with little requirement for critical reflection and theoretical distractions, with 'learning by doing' prioritized over education time spent outside the workplace. Drawing on a wide range of academic work, he identifies the 'fallacy of theory-less practice' and argues that whether social work practitioners realize it or not, they are constantly drawing on and negotiating theories when carrying out their job (Garrett, 2013: pp. 2–3).

Educating social workers to adopt a critical stance in both their reading and their practice encourages them to see dominant or official discourses in different lights and may help them to understand what it may feel like to be a recipient of some social policies and interventions. This is certainly the case with an issue such as how best to 'deal with' ASB when, as Barbara Wotton (1959: p. 13) suggested over half a century ago, the study of this topic can often reveal as much about the rule makers as it does about the rule breakers.

See also: Criminal Justice and Social Work Practice; Social Work Theory.

Bibliography

Anti-Social Behaviour Act, 2003. Chapter 38. TSO, London. http://www.legislation.gov.uk/ukpga/2003/38/pdfs/ukpga_20030038_en.pdf (accessed 17.12.12.).

Ashworth, A., 2004. Social control and "anti-social behaviour": the subversion of human rights? Law Quarterly Review 120, 263–291.

Bartley, M., 2006. Capability and Resilience: Beating the Odds. UCL, London.

Becker, H., 1963. Outsiders: Studies in the Sociology of Deviance. Macmillan, Toronto.

Bourdieu, P., 1999. The Weight of the World: Social Suffering in Contemporary Society. Stanford University Press, Stanford.

Briskman, L., 2013. Courageous ethnographers or agents of the state. Critical and Radical Social Work 1 (1), 51–66.

Burney, E., 2005. Making People Behave: Anti-social Behaviour, Politics and Policy. Willan, Cullompton.

Cameron, D., 15 December 2011. Troubled Families Speech. https://www.gov.uk/government/speeches/troubled-families-speech (accessed 30.11.13.).

Cohen, S., 1985. Visions of Social Control. Polity Press, Cambridge.

Cohen, S., 2002. Folk Devils and Moral Panics, third ed. Routledge, Abingdon.

Donzelot, J., 1997. The Policing of Families. Johns Hopkins, Baltimore.

Edwards, A., Hughes, G., 2008. Resilient fabians? Anti-social behaviour and community safety work in Wales. In: Squires, P. (Ed.), ASBO Nation: The Criminalisation of Nuisance. The Policy Press, Bristol, pp. 57–72.

Garland, D., 2001. The Culture of Control: Crime and Social Order in Contemporary Society. Oxford University Press, Oxford.

Garrett, P.M., 2013. Social Work and Social Theory. Policy Press, Bristol.

IFSW, 2013. Definition of Social Work. http://ifsw.org/ (accessed 30.11.13.).

Kelling, G.L., Wilson, J.Q., 1982. Broken Windows: The Police and Neighbourhood Safety. http://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/2/ (accessed 25.11.13.).

Lemert, E., 1972. Human Deviance, Social Problems and Social Control, second ed. Prentice-Hall, New Jersey.

Levitas, R., 1998. The Inclusive Society? Social Exclusion and New Labour. Palgrave, Basingstoke.

Lipsky, M., 1980. Street-level Bureaucracy: Dilemmas of the Individual in Public Services. Russell Sage Foundation, New York.

Lister, R., 2004. Poverty. Polity Press, Cambridge.

Matthews, R., Briggs, D., 2008. Lost in translation: interpreting and implementing anti-social behaviour policies. In: Squires, P. (Ed.), ASBO Nation: The Criminalisation of Nuisance. The Policy Press, Bristol, pp. 87–100.

Morris, L., 1994. Dangerous Classes. Routledge, London.

Munro, E., 2011. The Munro Review of Child Protection. TSO, London.

Pearson, G., 1975. The Deviant Imagination: Psychiatry, Social Work and Social Change. Macmillan, London.

Pickles, E., 2013. Speech to the Association of Police and Crime Commissioners. https://www.gov.uk/government/speeches/speech-to-the-association-of-police-and-crime-commissioners (accessed 30.11.13.).

Squires, P., 1990. Anti-social Policy. Harvester Wheatsheaf, Hemel Hempstead.

Squires, P., Stephens, D.E., 2005. Rougher Justice: Anti-social Behaviour and Young People. Willan, Cullompton.

Welshman, J., 2006. Underclass: A History of the Excluded 1880–2000. Hambeldon Continuum, London.

Wootton, B., 1959. Social Science and Social Pathology. Allen & Unwin, London.

Antitrust Policy: Lessons from the US

Daniel L Rubinfeld, University of California, Berkeley, Berkeley, CA, USA; and New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Whether applied through US, European, or other national competition laws, antitrust has come to represent an important competition policy instrument that underlies many countries' public policies toward business. As a set of instruments whose goal is to make markets operate more competitively, however, antitrust often comes into direct conflict with regulatory policies, including forms of price and output controls, antidumping laws, access limitations, and protectionist industrial policies. Courts in the US have widely adopted economic analysis as the theoretical foundation for evaluating antitrust concerns. Yet, antitrust statutes in the European Union also place heavy emphasis on the role of economics. This article provides a view of antitrust primarily from the perspective of the US policy; the review that follows illustrates a theme that has worldwide applicability. This article highlights: (1) the early antibig business period in the US, in which the structure of industry was paramount; (2) the period in which performance as well as structure was given significant weight, and there was a systematic attempt to balance the efficiency gains from concentration with the inefficiencies associated with possible anticompetitive behavior; and (3) the most recent period, which includes the growth of high technology and network industries, in which firm behavior has been given particular emphasis.

The term antitrust, which grew out of the US trust-busting policies of the late nineteenth century, developed over the twentieth century to connote a broad array of policies that affect competition. Whether applied through US, European, or other national competition laws, antitrust has come to represent an important competition policy instrument that underlies many countries' public policies toward business. As a set of instruments whose goal is to make markets operate more competitively, antitrust often comes into direct conflict with regulatory policies, including forms of price and output controls, antidumping laws, access limitations, and protectionist industrial policies.

Because its primary normative goal has been seen by most to be economically efficient, it should not be surprising that antitrust analysis relies heavily on the economics of industrial organization. But, other social sciences also contribute significantly to our understanding of antitrust. Analyses of the development of antitrust policy are in part historical in nature, and positive studies of the evolution of antitrust law (including analyses of lobbying and bureaucracy) often rely heavily on rational choice models of the politics of antitrust enforcement.

The relevance of other disciplines notwithstanding, there is widespread agreement about many of the important antitrust trade offs. Indeed, courts in the US have widely adopted economic analysis as the theoretical foundation for evaluating antitrust concerns. Interestingly, however, antitrust statutes in the European Union also place heavy emphasis on the role of economics. Indeed, a hypothetical conversation with a lawyer or an economist at a US competition authority (the Antitrust Division of the Department of Justice or the Federal Trade Commission (FTC)) or the European Union (The Competition Directorate) would be indistinguishable at first sight.

While this article provides a view of antitrust primarily from the perspective of the US policy, the article that follows illustrates a theme that has worldwide applicability. As our understanding of antitrust economics has grown throughout the past century, antitrust enforcement policies have also improved, albeit sometimes with a significant lag. In this survey the following are highlighted: (1) the early antibig business period in the US, in which the structure of industry was paramount; (2) the period in which performance as well as structure was given significant weight, and there was a systematic attempt to balance the efficiency gains from concentration with the inefficiencies associated with possible anticompetitive behavior; and (3) the most recent period, which includes the growth of high technology and network industries, and in which behavioral theories have been given particular emphasis.

The Antitrust Laws of the US

In the USA, as in most other countries, antitrust policies are codified in law and enforced by the judicial branch. Public cases may be brought under Federal law by the Antitrust Division of the Department of Justice, by the FTC, and/or by each of the 50-state attorneys general. (The state attorneys general may also bring cases under state law.) Further, there is a broad range of possibilities for private enforcement of the antitrust laws, which plays a particularly significant role in the USA.

The Sherman Act

Antitrust first became effective in the US near the end of the nineteenth century. Underlying the antitrust movement was the significant consolidation of industry that followed the Civil War. Following the war, large trusts emerged in industries such as railroads, petroleum, sugar, steel, and cotton. Concerns about the growth and abusive conduct of these combinations generated support for legislation that would restrict their power. The first antitrust law in the US – the Sherman Act – was promulgated in 1890. Section 1 of the Act prohibits: "Every contract, combination in the form of trust of otherwise, or

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71010-5

conspiracy, in restraint of trade or commerce among the several states, or with foreign nations."

Section 2 of the Sherman Act states that it is illegal for any person to "…monopolize, or attempt to monopolize, or combine or conspire with any other person or persons, to monopolize any part of the trade or commerce among the several states, or with foreign nations…." These two sections of the Act contain the two central key principles of modern antitrust policy throughout the world – conduct that restrains trade and conduct that creates or maintains a monopoly – is deemed to be anticompetitive.

The Clayton Act

Early in the twentieth century, it became apparent that the Sherman Act did not adequately address combinations, such as mergers, that were likely to create unacceptably high levels of market power. In 1914, Congress passed the Clayton Act, which identified specific types of conduct that were believed to threaten competition. The Clayton Act also made illegal conduct whose effect "may be to substantially lessen competition or tend to create a monopoly in any line of commerce." Section 7 of the Clayton Act is the principal statute for governing merger activity – in principle Sect. 7 asks whether the increased concentration will harm actual and/or potential competition.

Other sections of the Clayton Act address particular types of conduct. Section 2, which was amended and replaced by Sect. 1 of the Robinson–Patman Act in 1936, prohibits price discrimination between different purchasers of the same type and quality of a commodity, except when such price differences are cost-justified, or if the lower price is necessary to meet competition.

Section 3 of the Clayton Act specifically prohibits certain agreements in which a product is sold only under the condition that the purchaser will not deal in the goods of a competitor. This Section has been used to challenge exclusive dealing arrangements (e.g., a distributor that is obligated to sell only the products of a particular manufacturer) and 'tied' sales (e.g., the sale of one product is conditioned on the buyer's purchase of another product from the same supplier). The Act does not prohibit all such arrangements – only those whose effect would be likely to substantially lessen competition in a particular line of commerce.

The FTC Act

The US is nearly unique among competition law countries in having two enforcement agencies. In part to counter the power granted to the Executive branch under the Sherman Act, Congress created the FTC in 1914. Section 5 of the FTC Act, which enables the FTC to challenge 'unfair' competition, can be applied to consumer protection as well as mergers. In addition, the FTC has the power to enforce the Clayton Act and the Robinson–Patman Act. While the agencies act independently of one another, the FTC and the Department of Justice's (DOJ) enforcement activities have generally been consistent with one another during most of enforcement history.

What then are the differences between the two enforcement agencies? A simple answer is that the FTC is responsible for consumer protection issues, whereas criminal violations of Sect. 1 of the Sherman Act (e.g., price fixing and market division) are the responsibility of the Antitrust Division. It is also important to note that most enforcement activities, when successful, lead to injunctive remedies, where the party that has violated the law is required to cease the harmful activity. Exceptions are the criminal fines that are assessed when Sect. 1 of the Sherman Act is criminally violated, and penalties that are assessed in certain consumer protection cases. Damages are rarely assessed by the federal agencies, otherwise, although have been exceptions in occasion (e.g., when the US represents the class of government employees or when disgorgement is deemed a necessary remedy).

The Goals of Antitrust

Despite considerable economic change in the economy over the twentieth century, the federal antitrust laws have continued to have wide applicability in part because their language was quite vague and flexible. This has led, quite naturally, to extensive historical study of the legislative history and substantial debate about congressional intent, especially where the Sherman Act is concerned. Some scholars have argued that the Sherman Act was directed almost entirely toward the achievement of allocative efficiency (e.g., Bork, 1966). Others have taken the view that Congress and the courts have expressed other values, ranging from a broad concern for fairness, to the protection of specific interest groups, or more simply to the welfare of consumers writ large (e.g., Schwartz, 1979; Stigler, 1985).

The debate concerning the goals of the antitrust laws continues today. The enforcement agencies, for example, evaluate mergers from both consumer welfare and total welfare (consumer plus business) points of view, in part because the courts are not clear as to which is the appropriate standard under the Clayton Act. While it seems clear that the Sherman Act was intended in part to protect consumers against the inefficiencies of monopolies and cartels, it is significant that at the time of the passage of the Act many economists were in opposition because they believed that large business entities would be more efficient. Whatever the goals of the Sherman Act, it is notable that a merger wave (1895–1905) soon followed the passage of the Act. It may be that outlawing various types of coordinated behavior (Sect. 1) may have encouraged legal coordination through merger.

Historical Developments

Because antitrust focuses on the protection of competitive markets, it was natural to suspect other nonstandard organizational forms as potentially anticompetitive. During the early part of the 1900s, most antitrust enforcement was public, and it was directed against cartels and trusts. Early successes included convictions against the Standard Oil Company (*Standard Oil v. United States*, 1911) and the tobacco trust for monopolization in 1911.

It is important to note, however, that the Sherman Act does not prohibit all restraints of trade – only those restraints that

are unreasonable. The distinction between reasonable and unreasonable restraints remains a subject of intense debate today. It is significant, however, that the distinction between per se analysis (in which a practice is deemed illegal on its face) and rule of reason (in which one trades off the procompetitive and anticompetitive aspects of a practice) was made during this early period, and it remains important today.

By its nature, a per se rule creates a rebuttal presumption of illegality once an appropriate set of facts is found. Per se rules have the advantage that they provide clear signals and involve minimal enforcement costs. Yet actual firm behavior in varying market contexts generates exceptions that call for in-depth analyses. It is not surprising, therefore, that the per se rule is the exception to the rule of reason norm.

An example of the application of rule of reason is price discrimination. Viewed as an exercise of monopoly power, the practice was seen as suspect. Indeed, the Robinson–Patman Act presumes that with barriers to entry, price discrimination marks a deviation from the competitive ideal that was presumed to have monopoly purpose and effect. Efficiency justifications for price discrimination and other 'restrictive practices' – promoting investment, better allocating scarce resources, and economizing on transaction costs – were overlooked during the early enforcement period. Confusion also arose between the goal of protecting competition and the practice of protecting competitors. At the beginning of the twenty-first century, however, efficiency arguments are clearly pertinent to the evaluation of Robinson–Patman claims.

There is no doubt that antitrust enforcement is difficult, even with the more sophisticated tools of industrial organization that are available today. The Sherman Act does not offer precise guidance to the courts in identifying illegal conduct. In both the Clayton Act and the Sherman Act, Congress has chosen not to enumerate the particular types of conduct that would violate the antitrust laws. Instead, Congress chose to state general principles, such as attempts to monopolize, and contracts or combinations that restrain trade, without elaborating on what actually qualifies for the illegal behavior. It was left to the courts to ascertain the intent of the antitrust statutes and to distinguish conduct that harms competition from conduct that does not.

In a significant early case, *Board of Trade of the City of Chicago v. USA* (1918), the Supreme Court reiterated the reasonableness standard for evaluating restraints of trade. The Court concluded that "The true test of legality is whether the restraint imposed is such as merely regulates and perhaps thereby promotes competition or whether it is such as may suppress or even destroy competition." In this and in subsequent cases, the court failed to provide a clear statement of when and how a rule of reason analysis should be applied.

Antitrust enforcement agencies and the courts continue to debate the mode of analysis that is most appropriate in particular market contexts and when particular practices are at issue. Both courts and agencies deem it appropriate to undertake some form of 'quick-look analysis,' one which goes beyond the application of a per se rule, but which falls short of a full rule of reason inquiry (see *California Dental Association v. FTC*, 1999, Melamed 1998). Per se rules are applied to certain horizontal restraints, involving two or more firms operating in the same line of business. Quick-look and rule of reason analyses are more prevalent when the restraints at issue are vertical (involving two or more related lines of business, e.g., a manufacturer and a supplier).

Quick-look and more complete rule of reason analyses have relied on a number of basic principles. First and foremost is the market power screen. Courts have appreciated that antitrust injury necessitates the exercise of market power. Antitrust analysis typically begins with the measurement of market share possessed by firms alleged to engage in anticompetitive conduct. As many scholars have noted (e.g., Stigler, 1964), in the absence of collusion, the exercise of market power in unconcentrated markets is unlikely. Market concentration is often seen as a necessary but not sufficient condition for the exercise of market power, since ease of entry, and demand and supply substitutability can limit the ability of firms to raise prices even in highly concentrated markets (e.g., Baumol et al., 1982). But, market concentration – market power screens are not essential; witness the Supreme Court's opinion in *Federal Trade Commission v. Indiana Federal of Dentists* (1986).

Although courts have long recognized the importance of market power to conclusions about antitrust injury, the standards by which they adjudicate antitrust cases, and their willingness to apply sophisticated economic analysis, have varied significantly over time. Antitrust policy once treated any deviation from the competitive ideal as having anticompetitive purpose and effect. Vertical arrangements were particularly suspicious if the parties agreed to restraint that limited reliance on market prices. Over time, however, the critique of such arrangements diminished as economists began to appreciate the importance of efficiencies associated with a range of contractual practices *Leegin v. OSKS* (vertical minimum resale price fixing); *Texaco, Inc. v. Dagher* (2006) (joint ventures).

The Structure-Conduct Interventionist Period

During the 1950s and the 1960s, the normative analysis of antitrust was dominated by the work of Joe Bain on barriers to entry, industry structure, and oligopoly (Bain, 1968). Bain's relatively interventionist philosophy was based on the view that scale economies were not important in many markets, that barriers to entry are often high and can be manipulated by dominant incumbent firms, and that supracompetitive monopolistic pricing is relatively prevalent. This tradition held, not surprisingly, that nonstandard and unfamiliar practices should be approached with skepticism. During this era, antitrust enforcers often concluded that restraints which limited the number of competitors in a market necessarily raised prices, and that the courts could protect anticompetitive conduct by appropriate rule making.

In this early structure-conduct era, industrial organization economists tended to see firms as shaped by their technology. Practices, such as joint ventures, that reshaped the boundaries of the firm (e.g., joint ventures) were often seen as suspect. Significant emphasis was placed on the presence or absence of barriers to entry, which provided the impetus for a very tight market power screen.

Because the government was often seen as benign, it was not surprising that antitrust looked critically at mergers and acquisitions. Moreover, in absence, a broader theory that encompassed a variety of business relationships among firms,

it is not surprising that the Supreme Court often supported government arguments without seriously evaluating the trade offs involved (e.g., *USA v. Von's Grocery Co.*, 1966).

Economies of scale are illustrative. That they can create a barrier to entry was emphasized under Bain's point of view, whereas the clear benefits that scale economies provide were given little recognition. For example, the Supreme Court argued that 'possible economies cannot be used as a defense to illegality' (*Federal Trade Commission v. Procter and Gamble Co.*, 1967).

Government hostility during this period also applied to markets with differentiated products. In *USA v. Arnold, Schwinn & Co.* (1907), for example, the government was critical of franchise restrictions that supported product differentiation because the restrictions were perceived to foster the purchase of inferior products at higher than competitive prices. The possibility that exclusive dealing was procompetitive was not given serious consideration during the 1950s and 1960s.

The failure of antitrust enforcers to appreciate the benefits of discriminatory contractual practices was also evident in the early enforcement of the Robinson–Patman Act. Through the 1960s, the Act was vigorously enforced. Legitimate reasons for discriminatory pricing were very narrow (e.g., economies of scale in production or distribution). Since the mid-1960s, however, there has been a significant shift. Fears that the Act might discourage procompetitive price discrimination have led to less aggressive enforcement by the FTC.

The 1960s and 1970s marked a period of substantial empirical analysis in antitrust, motivated by structural considerations. The literature on the correlation between profit rates and industry structure initially showed a weak positive correlation, suggesting that high concentration was likely to be the source of anticompetitive firm behavior. However, this interpretation was hotly disputed. If behavior that is described as a barrier to entry also served legitimate purposes, what can one conclude even if there is a positive correlation between profitability and concentration? The challenge to this line of empirical work is typified by the debate between Demsetz (1974) and Weiss (1974). Demsetz argued that concentration was a consequence of economies of scale and the growth of more efficient firms – in effect the early empirical work suffered from problems of simultaneity. If concentrated markets led to higher industry profits, these profits were the consequence and not the cause of the superior efficiency of large firms and they consequently are consistent with competitive behavior. Today, these early studies are viewed critically, as having omitted variables that account for research and development, advertising, and economies of scale. It is noteworthy that the observed positive correlation reflected both market power and efficiency, with the balance varying on a case-by-case basis.

The Influence of the Chicago School

All of the assumptions that underlie the tradition of the 1950s and 1960s came under increasing criticism during the 1970s, led in part by the influence of the Chicago School. While the views of its proponents (e.g., Posner, 1979; Easterbrook, 1984) are themselves both rich and varied, they have come to be typified as including the following: (1) a belief that the allocative efficiencies associated with economies of scale and scope are of paramount importance; (2) a belief that most markets are competitive, even if they contain a relatively few number of firms. Accordingly, even if price competition is reduced, other nonprice forms of competition will fill the gap; (3) a view that monopolies will not last forever. Accordingly, the high profits earned by dominant monopolistic firms will attract new entry that in most cases will replace the monopolist or at least erode its position of dominance; (4) a view that most barriers to entry, except perhaps those created by government, are not nearly as significant as once thought; (5) a belief that monopolistic firms have no incentive to facilitate or leverage their monopoly power in vertically related markets (the 'single monopoly rent' theory); (6) a view that most business organizations maximize profits; firms that do not, will not survive over time; and (7) a belief that even when markets generate anticompetitive outcomes, government intervention, which is itself less than perfect, is appropriate only when it improves economic efficiency.

The period of the late 1960s and 1970s not only marked a significant influence by the Chicago School; it also saw an upgrading of the role of economists in the antitrust enforcement agencies. The increased role continued through the end of the twentieth century in the US, as it did in the European Union. For example, two economists play decision-making roles at the Department of Justice – one a Deputy for Economics and the other a Director of Economic Enforcement (the FTC also has several economists in decision-making roles).

Because of the influence of the Chicago School, nonstandard contractual practices that were once denounced as anticompetitive and without a valid purpose were often seen as serving legitimate economic purposes. Not surprisingly, for example, the per se rule limiting exclusive distribution arrangements in *Schwinn* was struck down by the US Supreme Court in *GTE Sylvania, Inc. v. Continental TV, Inc.* (1977). Although the US DOJ/FTC Merger Guidelines does not explicitly spell out a trade-off analysis of efficiencies and competitive effects, accounting for the economic benefits of a merger is reflected in the Guidelines and is now standard practice. Antitrust enforcement is alert to trade-offs, and less ready to condemn conditions for which there is no obvious superior alternative.

The Post-Chicago School Analyses of Strategic Economic Behavior

In the 1970s and continuing through the 1990s new industrial organization tools, especially those using game-theoretic reasoning came into prominence. These tools allow economists to examine the ways in which established firms behave strategically in relation to their actual and potential rivals. The distinction between credible and noncredible threats, which was absent from the early entry barrier literature, has been important to an assessment of the ability of established firms to exclude competitors and to the implications of exclusionary conduct for economic welfare (e.g., Dixit, 1979). These theories also illuminated a broader scope for predatory pricing and predatory behavior. Previously scholars such as Robert Bork (1978) had argued that predatory pricing imposes high costs on the alleged predator and is unlikely to be profitable in all but extraordinary situations. Developments in the analysis of

strategic behavior provide a richer perspective on the scope for such conduct. Thus, nonprice competitive strategies that 'raise rivals' costs' (Krattenmaker and Salop, 1986; Ordover et al., 1990) are now thought to be quite prevalent. Indeed, models of dynamic strategic behavior highlight the ability and opportunity for firms to engage in coordinated actions and for firms to profit from conduct that excludes equally efficient rivals (e.g., Kreps and Wilson, 1982; Milgrom and Roberts, 1982).

The implications of the various models of strategic behavior remain hotly debated. On the more skeptical side is Franklin Fisher, who stresses that while these new developments offer insights, they are insufficiently complete to provide firm conclusions or to allow us to measure what will happen in particular cases (Fisher, 1989). A more optimistic view is given by Shapiro (1989) who believes that we can now analyze a much broader range of business competitive strategies than before. We also know much more about what to look for when we are studying areas such as investment in physical and intangible assets, the strategic control of information, network competition and standardization, and the competitive effects of mergers.

The analysis of strategic behavior has emphasized the potential for exercising market power, often to the detriment of consumers. At the same time, other analyses have stressed that there are gains to consumers from coordinated behavior among firms with market power. Thus, Coase (1937) argued that market forces were only one means to organize market activities, and that nonmarket organizations could provide a viable alternative. Further, Williamson (1985) developed the theory of nonmarket organization to show that contractual restraints can provide improved incentives for investments in human and physical assets that enhance the gains from trade. This transaction cost approach helps to provide a foundation for understanding the efficiency benefits of contractual restraints in vertical relationships.

Coupled with the game-theoretic analyses of firm behavior in imperfect markets has been the application of modern empirical methods for the analysis of firm practices. These newer methods allow for the more precise measurement of market power, and they indicate generally that market power is relatively common, even in markets without dominant firms (Baker and Rubinfeld, 1999).

The Public Choice Approach

Today, a balanced normative approach to antitrust would involve reflections on the broad set of efficiencies associated with various organizational forms and contractual relations, as well as the possibilities for anticompetitive strategic behavior in markets with or without dominant firms. A positive (descriptive) approach focuses on the relationship between market structure and the politics of the interest groups that are affected by that structure.

The theory of public choice has been used as the basis for limiting competitor standing to bring antitrust suits. In support is the argument that competitors' interests deviate from the public interest because competitors view efficiencies associated with alleged exclusionary practices to be harmful, and therefore are not in a position to distinguish efficient from inefficient practices. On the other hand, competitors are often the most knowledgeable advocates, and are likely to know more about the socially harmful effects of exclusionary practices long before consumers do. The debate over standing to sue has surfaced with respect to indirect purchaser suits (e.g., *Illinois Brick Co. v. Illinois*, 1977) and with respect to predatory pricing (*Brooke Group Ltd. v. Brown and Williamson Corp.*, 1993).

High Technology and Dynamic Network Industries

In the latter part of the twentieth century and continuing into the twenty-first, rapid changes in technology altered the nature of competition in many markets in ways that would most likely have surprised the antitrust reformers of the late nineteenth century. The early debates centered on scale – did the benefits of economies of scale in production outweigh the associated increase in market power? In many dynamic high technology industries today, demand, rather than supply is often the source of substantial consumer benefits and significant market power. Economies of scale on the demand side arise in industries such as computers and telecommunications because of the presence of network effects, whereby each individual's demand for a product is positively related to the usage of the product (and complementary products) by other individuals. Network effects apply to communications networks (where consumers value a large network of users with whom to communicate, such as compatible telephone systems and compatible fax machines), and they apply to virtual networks or hardware–software networks (where there is not necessarily any communication between users).

In industries in which network effects are significant, a host of issues challenges our traditional views of antitrust. Because network industries are often characterized by large sunk costs and very low marginal costs, there is a substantial likelihood that a successful firm will come to dominate a market and to persist in that dominance for a significant period of time. Indeed, while there is no assurance that a single standard will arise in network industries, it is nevertheless often the case that users will gravitate toward using compatible products. This combination of economic factors makes it possible for firms to adopt price and nonprice policies that exclude competition and effectively raise prices significantly above what they would be were there more competition in the market (Rubinfeld, 1998).

A host of hotly debated antitrust policy issues are raised by the increasing importance of dynamic network industries (Whinston, 2006). Whatever view one holds, there is little doubt that the antitrust enforcement stakes are raised. On the one hand, because the path of innovation today will significantly affect future product quality and price, the potential benefits of enforcement are huge. This perspective clearly motivated the Department of Justice and 20 states when they chose to sue Microsoft for a variety of antitrust violations in 1998 (*USA v. Microsoft*, 2001). On the other hand, because the path of innovation is highly uncertain and technology is rapidly changing, barriers to entry that seem great today could disappear tomorrow, and the potential costs of enforcement are large as well. The threat of potential entry by innovative firms was a significant part of Microsoft's defense in the Department of Justice case.

A related set of issues has arisen with respect to the overlap between intellectual property and antitrust. The law gives the owners of patents, copyrights, and other intellectual property monopoly rights over their property. But, what is the appropriate scope of those rights? When, if at all, should the owners of those rights be required to give access or otherwise deal with competitors (Carlton, 2001)? In its more expansive period, the Supreme Court had taken a rather broad view of those facilities that were deemed to be essential and required competitor access (*Aspen Skiing Co. v. Aspen Highlands Skiing Corp.*, 1985). But, over the past decade the Court has made it clear that it has little patience for the so-called essential facilities doctrine (*Verizon Communications v. Law Offices of Curtis v. Trinko*, 2004).

Standard Setting

In network industries, competitive problems may arise in the competition to develop market standards. Dominant firms may have an incentive to adopt competitive strategies that support a single standard by preventing the products of rivals from achieving compatibility. Indeed, when the dominant firm's product becomes the standard for the industry, firms that are developing alternative standards may find it difficult to compete effectively (Farrell and Katz, 1998). Alternatively, firms might collude to affect the outcome of the standard-setting (price-setting) process. Such collusion can be difficult to detect because firms often have procompetitive reasons for cooperating in the race to develop new technology. Indeed, firms often possess assets and skills that can make collaboration in developing a market standard an efficient arrangement. An example is the pooling of technology related to video and audio streaming on the Internet. In its business review letter of the MPEG (Moving Picture Experts Group)-II patent pooling agreement in 1997, the Antitrust Division of the Department of Justice spelled out a set of conditions under which the pooling of assets would be deemed to be procompetitive. Since that time, the Division has published a number of additional business review letters involving other patent pools (IEEE (Institute of Electrical and Electronics Engineers), 3G, DVD (Digital Video Disc)). Where and how to draw the line between procompetitive sharing of assets of skills and other anticompetitive activities that will discourage competitors in the battle for the next generation standard remains a highly significant antitrust issue (DOJ/FTC 2005).

As a general rule, one should expect that sufficient competition will exist to develop a new product or market standard whenever more than a few independent entities exist that can compete to develop the product or standard. This rule of thumb is consistent with case law and conclusions in guidelines published by the US antitrust authorities. For example, the DOJ/FTC *Antitrust Guidelines for the Licensing of Intellectual Property* concludes that mergers or other arrangements among actual or potential competitors are unlikely to have an adverse effect on competition in research and development if more than four independent entities have the capability and incentive to engage in similar R&D activity (DOJ/FTC 1995). Similarly, the 1984 DOJ/FTC *Merger Guidelines* (revised in 1992, 1997, and most recently in 2011) concludes that mergers among potential competitors are unlikely to raise antitrust concerns if there are more than a few other potential entrants that are similarly situated.

Leveraging

Leveraging occurs when a firm uses its advantage from operating in one market to gain an advantage in selling into one or more other, generally related markets. Leveraging by dominant firms in network industries may take place for a variety of reasons that can be procompetitive or anticompetitive, depending on the circumstances. The challenge for antitrust policy is to distinguish between the two. On the one hand, leveraging can be seen as a form of vertical integration, in which a firm may improve its distribution system, economize on information, and/or improve the quality of its products. Leveraging, however, can be anticompetitive if it serves as a mechanism by which a dominant firm is able to raise its rivals' costs of competing in the marketplace.

Leveraging can be accomplished by a variety of practices (e.g., tying, bundling, exclusive dealing), each of which may have anticompetitive or procompetitive aspects, or a combination of the two. For example, with tying, a firm conditions the purchase (or license) of one product – the tied product – on the purchase (or license) of another product – the tying product. A firm might choose a tying arrangement for procompetitive reasons, including cost savings and quality control. Suppose, for example, that a dominant firm offers to license its dominant technology only to those firms that agree to also license that firm's complementary product, and suppose that the complementary product builds on the firm's next generation technology. Such a tying arrangement could allow the dominant firm to create a new installed base of users of its next generation technology in a manner that would effectively foreclose the opportunities of competing firms to offer their products in the battle for the next generation technology (Farrell and Saloner, 1986). Alternatively, suppose that a dominant seller offers discounts to buyers that remain loyal to the seller by buying all or a substantial percentage of their product from that seller (*LePage's Inc. v. 3M*). In the extreme this could amount to an exclusive dealing arrangement, but short of that being fully exclusive, the practice could make it difficult for competitors to achieve adequate scale and ultimately be exclusionary (Melamed, 2005; Rubinfeld, 2005).

International Antitrust

The emergence of a global marketplace raises significant antitrust policy issues. On the one hand, free trade and the opening of markets makes it less likely that mergers will significantly decrease competition, and indeed, creates greater opportunities for merging or cooperating firms to generate substantial efficiencies. On the other hand, international price-fixing agreements are more difficult for US or other domestic enforcement agencies to police, since success often requires some degree of cooperation with foreign governments or international organizations. The 1990s we saw a significant increase in the prosecution of international price-fixing conspiracies. Indeed, from the late 1990s through the first decade of the twenty-first

century, the Department of Justice criminal fines received were in the hundreds of millions and occasionally billions of dollars from investigations that included vitamins and dynamic random access memory.

The internationalization of antitrust raises a host of difficult jurisdictional and implementation problems for all antitrust enforcement countries. One can get a sense of the enormous difficulties involved by looking at the problems from the perspective of the US. The Sherman Act clearly applies to all conduct, which has a substantial effect within the US. However, crucial evidence or culpable individuals or firms may be located outside the US. As a result, it is imperative for the US antitrust authorities to coordinate their activities with authorities abroad. This is accomplished in part through mutual assistance agreements, as, for example, in the agreement between the US and Australia under the International Enforcement Assistance Act. It is also accomplished through informal 'positive comity' arrangements, whereby if one country believes that its firms are being excluded from another's markets, it will conduct a preliminary analysis and then refer the matter to the foreign antitrust authority for further investigation, and, if appropriate, prosecution. (Such an agreement was reached with the EU in 1991.) Today, those arrangements continue formally through a range of bilateral agreements and informally through the International Competition Network.

Conclusions

Antitrust policy has undergone incredible change over the twentieth century. As the views about the nature of markets and arrangements among firms held by economists and others have changed, so has antitrust. The early focus was on the possible anticompetitive effects of mergers and other horizontal arrangements among firms. The primary source of market power was the presence of scale economies in production. However, vertical arrangements receive significant critical treatment, and the sources of market power are seen as coming from the demand side as well as the supply side. Further, a wide arrange of contractual practices are now judged as creating substantial efficiencies, albeit with the risk that market power will be used for exclusionary purposes. Finally, significant changes in international competition, resulting from free trade and the communication and information revolution, have reinforced the importance of reshaping antitrust to meet the needs of the twenty-first century. Significant new antitrust challenges lie ahead. Whatever those challenges may be, we can expect that industrial organization economics and antitrust policy will be sufficiently flexible and creative to respond appropriately.

Statutes

Clayton Act, 38 Stat. 730 (1914), as amended, 15 USCA §§12-27 (1977).

Federal Trade Commission Act, 38 Stat. 717 (1914), as amended, 15 USCA §§41-58 (1977).

Robinson-Patman Act, 49 Stat. 1526 (1936), 15 USCA §13 (1977).

Sherman Act, 26 Stat. 209 (1890), as amended, 15 USCA §§1-7 (1977).

Cases

Aspen Skiing Co. v. Aspen Highlands Skiing Corp., 472 US 585 (1985) (see Sect. 4).

Board of Trade of the City of Chicago v. USA, 246 US 231 (1918) (see Sect. 3).

Brooke Group Ltd. v. Brown & Williamson Tobacco Corp., 509 US 209, 113 S. Ct. 2578 (1993) (see Sect 3.4).

California Dental Association v. FTC, 119 S. Ct. 1604 (1999) (see Sect. 3).

Federal Trade Commission v. Indiana Association of Dentists, 476 US 447 (1986) (see Sect. 3).

Federal Trade Commission v. Procter & Gamble Co., Inc., 386 US 568, 574 (1967) (see Sect. 3.1).

GTE Sylvania, Inc. Continental TV, Inc., 433 US 36, 45 (1977) (see Sect. 3.2).

Illinois Brick Co. v. Illinois, 431 US 720, 97 St. Ct. 2061 (1977) (see Sect. 3.4).

LePage's v. 3M, 324 F.3d 141 (3d Cir. 2003) (en banc) (see Sect. 6).

Leegin Creative Leather Products, Inc. v. PSKS, Inc., 551 US 877 (2007) (see Sect. 3).

Standard Oil Co. v. United States, 221 US 1 (1911) (see Sect. 3).

Texaco, Inc. v. Dagher, 547 US 1 (2006) (see Sect. 3).

USA v. Arnold, Schwinn & Co., 388 US 365 (1907) (see Sect. 3.1).

USA v. Microsoft, Civil Action, 253 F.3d 34 (D.C. Cir.), cert. denied 534 US 952 (2001) (see Sect. 4).

USA v. Von's Grocery Co., 384 US 270, 301 (1966) (see Sect. 3.1).

Verizon Communications v. Law Offices of Curtis v. Trinko, 540 U.S. 398 (2004) (see Sect. 4).

See also: Business History; Business Law; Regulation and Administration; Regulation, Economic Theory of; Regulation, Empirical Analysis of; Regulatory Agencies.

Bibliography

Bain, J., 1968. Industrial Organization. Wiley, New York.

Baker, J.B., Rubinfeld, D.L., 1999. Empirical methods in antitrust: review and critique. American Law and Economics Review 1, 386–435.

Baumol, W., Panzar, J., Willig, R., 1982. Contestable Markets and the Theory of Industry Structure. Harcourt Brace, New York.

Bork, R., 1966. Legislative intent and the policy of the Sherman Act. Journal of Law & Economics 7, 7–48.

Bork, R., 1978. The Antitrust Paradox: A Policy at War with Itself. Basic Books, New York.

Carlton, D., 2001. A general analysis of exclusionary conduct and refusal to deal – why Aspen and Kodak are misguided. Antitrust Law Journal 68, 659–683.

Coase, R., 1937. The nature of the firm. Economica 4, 380–405.

Demsetz, H., 1974. Two systems of belief about monopoly. In: Goldschmid, I., Mann, H.M., Weston, J.F. (Eds.), Industrial Concentration: The New Learning. Little Brown, Boston, pp. 164–183.

Dixit, A., 1979. A model of duopoly suggesting a theory of entry barriers. Bell Journal of Economics 10, 20–32.
Easterbrook, F., 1984. The limits of antitrust. Texas Law Review 63 (1), 1–40.
Farrell, J., Katz, M.L., 1998. The effects of antitrust and intellectual property law on compatibility and innovation. The Antitrust Bulletin 43, 609–650.
Farrell, J., Saloner, G., 1986. Installed base and compatibility: innovation, product preannouncements, and predation. American Economic Review 76, 940–955.
Fisher, F.M., 1989. Games economists play: a noncooperative view. RAND Journal of Economics 20, 113–124.
Krattenmaker, T.G., Salop, S.C., 1986. Anticompetitive exclusion: raising rivals' costs to achieve power over price. Yale Law Journal 96, 209–293.
Kreps, D.M., Wilson, R., 1982. Reputation and imperfect information. Journal of Economic Theory 27, 253–279.
Melamed, A.D., 2005. Exclusive dealing arrangements and other exclusionary conduct – are there unifying principles? Antitrust Law Journal 73, 375–412.
Milgrom, P., Roberts, J., 1982. Predation, reputation and entry deterrence. Journal of Economic Theory 27, 280–312.
Ordover, J.A., Saloner, G., Salop, S.C., 1990. Equilibrium vertical foreclosure. American Economic Review 80, 127–142.
Posner, R., 1979. The Chicago School of Antitrust Analysis, vol. 127. University of Pennsylvania Law Review, 925–948.
Rubinfeld, D.L., 1998. Antitrust enforcement in dynamic network industries. The Antitrust Bulletin, Fall-Winter, 859–882.
Rubinfeld, D.L., 2005. 3M's bundled rebates: an economic perspective. Chicago Law Review 72, 1–22.
Schwartz, L.B., 1979. Justice and Other Non-economic Goals of Antitrust, vol. 127. University of Pennsylvania Law Review, 1076–1081.
Shapiro, C., 1989. The theory of business strategy. RAND Journal of Economics 20, 125–137.
Stigler, G.J., 1964. A theory of oligopoly. Journal of Political Economy 72, 44–59.
Stigler, G.J., 1985. The origin of the Sherman Act. Journal of Legal Studies 14, 1–12.
US Department of Justice and Federal Trade Commission, 1984. MeRger guidelines. Reprinted in 4 trade Reg. Rep. (CCH) Paras. 3, 103.
US Department of Justice and Federal Trade Commission, 1992. Horizontal Merger Guidelines. Reprinted in 4 Trade Reg. Rep. (CCH) Paras. 13, 104.
US Department of Justice and Federal Trade Commission 1995 Antitrust Guidelines for the Licensing of Intellectual Property, 6 April 1995. Reprinted in 4 trade Reg. Rep. (CCH) 113, 132.
US Department of Justice and Federal Trade Commission, 2007. Antitrust Enforcement and IP Rights: Promoting Innovation and Competition. Available at: www.ftc.gov/reports/index.shtm.
Weiss, L., 1974. The concentration-profits relationship and antitrust law. In: Goldschmid, H., Mann, H.M., Weston, J.F. (Eds.), Industrial Concentration: The New Learning. Little Brown, Boston, pp. 184–232.
Whinston, M., 2006. Lectures on Antitrust Economics. MIT Press, Cambridge.
Williamson, O., 1985. The Economic Institutions of Capitalism. Free Press, New York.

Anxiety and Anxiety Disorders

Klaus Wiedemann, University Hospital Hamburg Eppendorf, Hamburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Anxiety is one of the important normally and regularly occurring emotions. Some of the actual public health problems like anxiety disorders are based on the pathology of feelings. Anxiety disorders comprise a heterogeneous group of disorders and each of these disturbances has a different etiology and outcome, and different physiological characteristics. Among the most important influences are actually genetic disposition, environmental impact upon one individual, and acute stressors that result in adaptational changes. Considering a large body of findings, several monoamine neurotransmitters and anxiogenic or anxiolytic-like neuropeptides are involved in the regulation of fear and anxiety.

Anxiety is one of the important normally and regularly occurring emotions, which can be observed throughout all human cultures and in several animal species. Some of the actual most prominent medical and public health problems like anxiety disorders or depression are based on the pathology of feelings (Damasio and Carvalho, 2013). While recent data show that the basic facial expression of emotions is also modulated by culture-specific influences, fundamental emotional expressions such as anxiety or fear originate presumably from the beginning of human ontogeny as biologically hardwired and universal signals (Jack et al., 2013).

Anxiety disorders comprise a heterogeneous group of disorders and each of these disturbances has a different etiology and outcome, and different physiological characteristics. Depending on actual classification rules according to the DSM 5 (2013), panic attacks and panic disorder, phobias, social phobia, and generalized anxiety disorder are summarized in anxiety disorders. In contrast to former classifications, post-traumatic stress disorder and obsessive compulsive disorder are no longer included, despite the fact that anxiety is a relevant dimension of these disorders. Moreover, anxiety has to be differentiated from fear and the so-called stress responses.

Anxiety derives from complex origins and among the most important influences are actually genetic disposition; environmental impact upon one individual, especially early adverse events; and acute stressors that result in adaptational changes. Considering a large body of findings, several monoamine neurotransmitters, for example, gamma-aminobutyric acid (GABA), serotonin, and noradrenalin, are involved in the regulation of fear and anxiety. In addition, an increasing variety of either anxiogenic neuropeptides like corticotropin-releasing factor (CRF), cholecystokinin tetrapeptide (CCK-4), vasopressin, and others, or anxiolytic-like neuropeptides like neuropeptide Y (NPY), neuropeptide S, and atrial natriuretic peptide (ANP) are most important in the modulation of anxiety.

Several neurobiologic findings also stem from investigations in humans by eliciting panic attacks. Panic attacks are unique in the spectrum of psychiatric disorders; since their core psychopathology is temporally limited, they can be provoked under laboratory conditions and can in part be conditioned. Due to the experimental character of these investigations, close comparisons with experiments in animals can be drawn. Hence it has been proposed that panic disorder involves the same pathways that support conditioned fear in animals. However, while induction of panic anxiety has been studied intensively, little is known about the intrinsic mechanisms for the termination of panic and anxiety.

Anxiety or fear is, in addition to happiness, sadness, anger, disgust, and desire, one of the important normally and regularly occurring emotions, which can be observed throughout all human cultures and in several animal species (Ekman, 1982), but the facial expression of such basic emotions is also modulated by culture-specific influences. Anxiety per se is a complicated concept since several difficulties arise in defining this emotion and, in addition, it has to be differentiated from fear and stress (see below). Anxiety occurs, besides in anxiety disorders, also comorbid in several other psychiatric conditions. Moreover, anxiety refers to a variety of other emotional experiences, for example, apprehensiveness, tension, and agitation, which occur also in other emotional states. Anxiety is defined by subjective, behavioral, and physiological characteristics. Anxiety involves the experience of dread and apprehensiveness, and the physiological reactions of anxiety usually include trembling, sweating, elevated heart rate and blood pressure, and increases in muscle tone. Anxiety is defined as pathological when occurring inadequately or with much more pronounced severity and debilitating features. An additional defining criterion in standardized diagnostic manuals is the concomitant occurrence of anxiety and avoidance. Representatives for these diagnostic entities, in which anxiety is the leading symptom, are panic disorder and generalized anxiety disorder. Anxiety is experienced in phobic disorders when the subject is confronted with the feared stimulus, which results in its avoidance. One of the reasons for the declaratory confusion of the term *anxiety* is its psychological similarity to fear and its vegetative similarity to stress. Similar to anxiety, *fear* also includes the experience of dread, and fear seems to be largely included into the concept of anxiety. Moreover, anxiety and fear induce similar bodily reactions, which include the so-called *stress* responses. In general, these can be separated in active responses, known as 'fight or flight' reactions or behavioral inhibition, like freezing. The accompanying stress-related sequelae can again be divided into two large entities: the excitatory fight or flight response postulated by Cannon (1929), and the endocrine stress concept raised by Selye

(1956). Complementarily, the results of several studies indicate that stress involves brain structures, which also mediate the perception of anxiety such as amygdala, hippocampus, and other limbic structures (see below).

Differentiation of Anxiety

Both anxiety and fear are regularly experienced within a range of normal emotional responses of everyday life. Specifically, fear is necessary to achieve personal growth and individual freedom during ontogeny.

Anxiety

Anxiety represents a basic emotional state, which is present in men and can be defined by affective (basic emotional feelings), perceptive (realization of bodily or psychomotor sensations), and cognitive components. Besides these subjective components, behavioral and physiological characteristics can be used to define anxiety phenomenologically. In contrast to experiencing anxiety during everyday life, anxiety as psychopathologic disturbance or anxiety disorder includes specific diagnostic criteria, neurobiological dysfunctions, and a specific genetic background and leads to social and occupational disabilities.

Fear

Fear is the normal reaction to threatening stimuli and is very common in everyday life. When fear is greater than warranted by the situation or starts to occur in inappropriate situations, a specific phobia arises, which belongs to the diagnostic entity of anxiety disorders. One distinction between fear in contrast to anxiety is based upon the presence of commonly defined stimuli, a realistic relation between dangerousness and elicited fear, and the potential to cope with or to adapt to the stimulus. Specific phobias are defined as persistent, irrational, exaggerated, and pathological dreads of a stimulus or situation combined with a compelling desire to avoid this feared challenge.

Stress

Stress is regularly experienced by all organisms, and refers generally to physical or psychological stimuli or alterations that are capable of disrupting the homeostasis of one individual or animal. With regard to psychological aspects of stress, predictability, control, and coping skills are important determinants, which, however, are also threatened during anxiety or fear. Hence, anxiety and fear also represent important psychological stressors with their physiological sequelae being similar to stress reactions. The differentiation between stress and fear is difficult in some situations like free speech, since psychological and biological aspects of stress are linked to each other and are mutually interdependent.

Anxiety Disorders

As defined by means of the diagnostic and statistical manual of mental disorders (*DSM-V* American Psychiatric Association, 2013), anxiety disorders comprise a heterogeneous group, which share *anxiety* as a symptom. However, each of these disturbances has a different etiology and outcome, and different physiological characteristics.

Panic Attacks and Panic Disorder

Panic disorder is characterized by recurrent paroxysmal anxiety, which can even surmount the fear of death during an acute myocardial infarction. These attacks are regularly combined with bodily sensations such as tachycardia, suffocation, shaking, trembling, sweating, abdominal distress, and dizziness. They typically have a sudden onset and are either unpredictable or occur before or during specific situations. The duration of panic attacks is short and ranges between some minutes up to about few hours. If these attacks are affiliated to specific situations, they can lead to avoidance of these specific events and agoraphobia develops, which sustains for longer periods.

Phobias

Phobias are usually differentiated into distinct subtypes: (1) agoraphobia, as frequent sequel of panic disorder, (2) social phobias, and (3) simple phobias. Agoraphobia is the fear of being in situations from which escape is not immediately possible. The symptoms regularly include depersonalization, derealization, dizziness, and cardiac symptoms. Agoraphobia may occur without preceding a panic attack, but remain consolidated between attacks. Social phobias are characterized by the fear that someone may be exposed to a situation where this person is inappropriately scrutinized by others or where this person may behave inadequately. Exposure leads to prominent symptoms of anxiety including bodily alterations, and anticipatory anxiety leads to the avoidance of these situations. Simple phobias are characterized by a persistent fear of a defined object or situation such as fear of spiders or fear of height. The anticipatory anxiety is common and these stimuli are largely avoided, which can impair daily life routines for long periods.

Generalized Anxiety Disorders

This disorder is characterized by excessive worry and by an unspecific, unrealistic, and excessive apprehension about a large variety of future events, which are difficult to control for the person. In addition, symptoms of physiological arousal such as restlessness, autonomic hyperreactivity, sleep disturbances, and muscle tension are characteristics. It has been classified as a chronic disorder lasting longer than 6 months.

Sources of Anxiety in Humans

Anxiety has to be derived from complex origins and interplay of genetic, biological, social, and psychological events and influences. Among the most important factors are the genetic or biological disposition, the developmental and environmental impact upon one individual, and acute stressors and experiences that challenge one person and lead to a variety of adaptational changes.

Genetic and Biological Disposition

Hints for a genetic background of anxiety disorders and indicators for their heritability have been considered as long as for mood disorders, despite the change of diagnostic criteria and labels for different anxiety disorders over the years. Among anxiety disorders, the genetics of panic disorder and generalized anxiety disorder have been most studied. Although panic disorder is the most common anxiety disorder, its underlying etiology is still not well understood. However, several studies have consistently shown that genetic factors explain about half of the variance and that most cases have a complex genetic basis. Existing data suggest that the genetic structures underlying panic disorder seem to be heterogeneous and seem also to differ between cases. The pattern of genes involved might differ in familial versus nonfamilial cases, in early- versus late-onset cases, in different comorbid conditions, but also gender effects and potential subphenotypes have to be considered (Schumacher et al., 2011). From a methodological point of view, family studies, twin studies, linkage and association studies have to be differentiated. Regarding panic disorder, it has been shown that relatives of patients have an increased risk of a similar disturbance. Among relatives a risk of up to 30% is reported, which is significantly different from a lifetime prevalence of about 2% in the general population. Also, twin studies support a heritable component since several studies indicate that the concordance rates for panic disorder are higher in mono- than in dizygotic twins. Linkage studies have been attempted several times, but no single gene loci could be identified. Considering the great complexity of this disorder, it is to be expected that single gene loci are unlikely to be responsible for the diagnostic entity *panic disorder*. However, association studies might lead to the detection of genes responsible for an enhanced vulnerability to anxiety disorders. Interestingly, while single polymorphisms like the transporter for the neurotransmitter serotonin seem to be linked to an altered sensitivity for panic attacks, the respective studies are possibly influenced by selection biases of investigated subjects, which illustrate the complexity of genetic correlations.

Support for a genetic basis of anxiety stems also from preclinical studies. By selective breeding, different lines of rats can be established that differ markedly in their innate anxiety behavior. In genetic knockout, strategies in mice for specific receptors was shown that deficiency of receptors, which are considered to be involved in anxiety and stress reactions, is correlated with a lower innate anxiety behavior. These receptors include particularly the CRF receptor (Steckler and Holsboer, 1999).

Regarding neurogenesis, genetic and epigenetic factors may also regulate the proliferation, survival, and integration of cells into the hippocampus, which is involved in information processing of fearful events. Future studies may therefore elucidate the effect of modulating neurogenesis in animal models of anxiety disorders and the development of proneurogenic compounds may have therapeutic potential (Kheirbek et al., 2012).

Social and Environmental Influences

Although the strong impact of untoward events during childhood is evident, it is worth remembering that simple relationships cannot be constructed. Besides family conflict situations, several factors such as parental support, child-rearing style, and personality traits have been linked to incidence of anxiety disorders in adolescence. Childhood separation anxiety, childhood parental loss, common stressful events, and major life events were predictive for individual susceptibility to develop panic disorder. Interestingly, also the liability to experience a panic attack after inhaled carbon dioxide as panicogenic agent (see below) was increased (Ogliari et al., 2010).

Prospectively ascertained child maltreatment was associated with anxiety disorders, indicating that maltreatment in fact carries an association of anxiety disorders as well as depression and alcohol abuse (Scott et al., 2010). Besides these factors, age also contributes to the expression of anxiety. Anxiety and anxiety disorders have a higher incidence in adolescence that cannot be reduced only to the use of different diagnostic tools, but seems to be related to other reasons.

Life Events

The amygdala and the prefrontal cortex show structural plasticity. Acute and chronic stress cause (see above) an imbalance of neural circuitries involved in anxiety and mood that can increase or decrease expression of those behaviors. In the short term, during increased fearful vigilance and anxiety in a threatening environment, these changes may be adaptive. However, when the danger passes and the behavioral state persists along with changes in neural circuitries, such maladaptation may require intervention in the case of anxiety disorders (McEwen et al., 2012). In context with environmental and developmental influences, traumatic events, which are regularly out of the realm of normal human experience, are of special importance. A traumatic event can lead to increased anxiety after the event, but may have also long-term effects that emerge with future traumas.

Neuronal Basis of Anxiety

Both the increasingly differentiated analysis of anatomical structures, and biochemical and neurophysiological pathways has led to a more detailed concept about the neurobiology of anxiety and especially of panic attacks.

Neuroanatomical Structures

While fear is one of the best investigated emotions in terms of brain mechanisms, a direct comparison of animal models of fear is limited with respect to the spectrum of human anxiety disorders. It has been proposed that panic disorder involves the same pathways that support conditioned fear in animals. These findings support the theory that panic attacks arise from loci in the brain stem that control especially serotonergic and noradrenergic neurotransmission and respiratory control. In addition, also glutamate receptors seem to play a role in fear memory. Further, it was postulated that anticipatory anxiety arises from kindling of limbic areas and phobic avoidance from precortical activation.

Sensory inputs for conditioned stimuli are mediated through the connection of the anterior thalamus to the lateral

and then to the central nucleus of the amygdala. The amygdala is a structure that plays a prominent role in both fear and anxiety: the two components of the amygdala are the central nucleus of the amygdala and the bed nucleus of the stria terminalis. These regions exhibit a high degree of connectivity and play a central role in generating negative emotional responses like fear and anxiety to environmental or conditioned stimuli (Johansen et al., 2011). Interestingly, besides its role in fear processing, the amygdala is considered as structure that detects salient and personally relevant stimuli in cooperation with ventral and dorsal medial prefrontal cortex with both engaged in the processing of socially relevant stimuli (Fossati, 2012).

The central nucleus of the amygdala coordinates physiological and behavioral responses related to anxiety. Efferents of this nucleus have several targets, for example, the parabrachial nucleus producing an increase in respiratory rate, the lateral nucleus of the hypothalamus activating the sympathetic system, the locus coeruleus resulting in an increase in noradrenalin release with its sequelae of increased blood pressure and heart rate and behavioral fear responses, and the nucleus paraventricularis of the hypothalamus causing an increase in corticosteroids via release of neuropeptides like CRF. As outlined by LeDoux (1998), the overlap between effects of brain stem activation by the central nucleus of the amygdala in animals with physiological effects in humans during panic attacks is striking. Besides these connections, mutual interactions between the amygdala and the thalamus, the prefrontal and the somatosensory cortex are obvious. An impairment of the cortical processing could lead to a misinterpretation of visceroafferent cognitions, leading to the activation of the above-mentioned systems. Because of these complex interactions with the autonomic and endocrine regulation, panic attacks apparently result in equivocal physiological and behavioral sequelae (see below) (Gorman et al., 2000).

Transmitter Systems

At least four monoamine transmitter systems, i.e., the noradrenergic, serotonergic, glutamatergic, and GABA system, are involved since several pharmacological compounds provide therapeutic benefit via respective receptors. However, additional neurotransmitter systems have been proposed to underlie anxiety and fear, such as the endocannabinoid system which gives new challenges to investigators.

From clinical experience, the involvement of the GABA system in anxiety and anxiety disorders seems highly important, and for short-term treatment, ligands at the GABA-A receptor like benzodiazepines are extremely helpful due to their rapid onset of anxiolytic action. However, all GABAergic compounds have more or less sedative properties and may induce tolerance, abuse liability, and withdrawal symptoms.

In the past, for many years the balance between GABAergic and glutamatergic systems was pursued in anxiety research. Decreasing the activity of glutamatergic neurotransmission may attenuate excitation in the CNS, thus resulting also in anxiolysis. Moreover, glutamate (especially NMDA receptors) may be involved, besides in fear acquisition, also in fear extinction, which is important in treatment: modulating glutamate neurocircuitries may result directly in therapeutic effects of cognitive therapies.

Further, the monoamine transmitters, serotonin and noradrenalin, and neuropeptides like CRF are important in the regulation of anxiety and fear. Moreover, serotonin neurons in the raphe nuclei have an inhibitory effect on noradrenergic neurons at the locus coeruleus. In addition, these neurons act at the periaquaeductal gray modifying the escape responses and are also thought to inhibit the hypothalamic release of CRF. From a clinical point of view, these findings are supported by therapeutic effects of serotonin reuptake inhibitors. These pharmaceuticals inhibit the uptake of serotonin back into the presynaptic neuron and increase the amount of serotonin in the synapse to bind both to pre- and postsynaptic sites with more than 13 subtypes of serotonin receptors (Kent et al., 1998): a long-term increase of serotonergic transmission by these compounds exerts antipanic and anxiolytic effects.

Another important system is the noradrenergic system (Sullivan et al., 1999). Noradrenalin neurons largely originate in the locus coeruleus and some other nuclei in the medulla and pons. Projection sites include the prefrontal cortex, the amygdala, the hippocampus, the hypothalamus, the thalamus, and the nucleus tractus solitarius. Conversely, the locus coeruleus is innervated by the amygdala. Therefore, the locus coeruleus seems to integrate external sensory and visceral afferents influencing a wide range of neuroanatomical structures related to fear and stress. Clinically it has been proven that, for example, noradrenergic alpha-2 receptor antagonists such as yohimbine can be used to provoke panic attacks acting, whereas clonidine, an alpha-2 adrenergic agonist, exerts anxiolytic-like effects in experimentally induced panic attacks (see below).

All in all it may be summarized that besides serotonin and noradrenalin also glutamatergic and GABAergic transmitters are interacting in complex circuitries, which involve also neuropeptides such as CRF (Arborelius et al., 1999; Koob, 1999). Neurons containing CRF and its receptors have been shown to be distributed throughout the brain, especially the amygdala, the hypothalamus, and the locus coeruleus, and CRF has emerged as a neurotransmitter/modulator that plays a central role not only in stress regulation, but also in anxiety and depression. CRF neurons project from the amygdala to the locus coeruleus. Hence, there are strong hints that CRF could act as a modulator of cognitive and physiological symptoms of anxiety. CRF initiates on one hand a humoral cascade, which enhances via the secretion of corticotropin the release of glucocorticoids, which in turn act at central gluco- and mineralocorticoid receptors. On the other hand, CRF seems directly involved in the modulation of anxiety and depression within the central nervous system. Stress results in increased CRF concentrations in the locus coeruleus and CRF increases the firing rate of noradrenergic neurons. In contrast, noradrenalin also potently stimulates the release of CRF. The involvement of CRF is interesting also with respect to respiratory alterations during panic attacks, which have led to the 'suffocation false alarm theory' (Klein, 1993), since CRF seems to be an important modulator of respiratory centers in the brain stem. Several studies support the contention that antagonists and inhibitors of the synthesis of CRF exert anxiolytic-like effects. Antagonists of CRF receptors have also been examined in clinical trials for their anxiolytic and antidepressant potency.

Since serotonin reuptake inhibitors are involved in the inhibitory regulation of noradrenergic neurons of the locus coeruleus and are thought to reduce the hypothalamic release of CRF, these complex interactions suggest that noradrenergic, serotonergic, and CRF-regulated neurotransmission are linked together with other transmitter circuitries mediating the responses to anxiety, fear, and stress.

Besides CRF, several other neuropeptides are involved in the regulation of anxiety: it was consistently demonstrated that NPY has an inhibitory role in fear acquisition and facilitates fear extinction in animals. In line, also in man abnormally low levels of plasma and cerebrospinal fluid levels of NPY have been found in patients with anxiety disorders (Bowers et al., 2012). Another neuropeptide, i.e., neuropeptide S is of special interest since it exerts anxiolytic effects in animals (Dine et al., 2013), however, effects in humans are still pending.

Models of Anxiety

Anxiety is not merely one of the most important naturally occurring emotions throughout phylogeny and ontogeny, but is also provocapable by different means and can then be readily observed under experimental conditions. Both in humans and in animals, a variety of investigations have been conducted, which allow thorough insights into the pathophysiological conditions and the cognitive and neurobiological processes involved in these specific emotional states.

Animal Models

Animal studies in anxiety can be used both for investigating the physiological and anatomical substrates of anxiety and for studying pharmacological strategies for potential anxiolytic or anxiogenic effects (Westenberg et al., 1996). Basically there are three types of animal behavioral models to detect anxiolytic-like effects. They can be classified in two major groups, i.e., conditioned behavior comprising conflict tests and cognitive-based tests, and unconditioned tests, i.e., ethological behavior. Conditioned behavior detects responses controlled by operant conditioning procedures. Unconditioned behavior is mainly based upon naturally occurring behavior and is called an ethological-based model. A different type is separation models, which mainly investigate the behavior of an offspring during separation from its mother and involves the investigation of developmental disturbances.

Conditioned Emotional Responses

The most important conditioned models comprise conflict models where behavior is suppressed by aversive stimulation. The release of the suppressed behavior without altering the levels of punished responding following pharmacological intervention is estimated as the anxiolytic-like effect. Using these models in rodents, for example, benzodiazepines are consistently effective, whereas for other compounds such as serotonin reuptake inhibitors anxiolytic-like effects are difficult to find. Other important models are the fear-potentiated startle response where the startle response of rats is augmented by fear conditioning. During the conditioning phase, another stimulus is presented signaling the presence of, for example, a shock stimulus. During the startle response, presentation of the stimulus enhances the startle amplitude. Also in this paradigm, benzodiazepines exert anxiolytic effects. In addition to these models, a variety of other conditioned responses and active and passive avoidance reactions can be determined.

Ethological Models

In contrast to the conditioned responses, the ethological models are based upon naturally occurring behavior. The most important and frequently used models are the elevated plus-maze, the open-field, and the dark–light-box models. The elevated plus-maze uses the conflict between exploration and aversion to elevated open places. In this test, anxiety is generated by placing the animal on an elevated open arm, where height and openness rather than light are responsible for anxiogenic effects. The device is shaped as a plus sign with two open arms and two arms enclosed by high walls. The time that rodents spend on the open arms and the number of entries, both, are indicative for anxiolytic-like effects. The open-field test investigates the distance traveled by rodents in a locomotor box within a given time interval. Usually, rodents avoid open areas and try to remain at the edge of the locomotor box. The overall distance traveled and the transitions of the central area of the box are related to the anxiolytic potency by a treatment. The dark–light box uses the number of transitions between a light and a dark, closed compartment as measure of anxiety, since rodents prefer the dark compartment. Peptide-receptor ligands such as CRF (Arborelius et al., 1999) and CCK-4 (Bradwejn and Vasar, 1995) show anxiogenic effects in all three paradigms, whereas other substances such as ANP (Wiedemann et al., 2001) and NPY (Heilig and Widerlöv, 1995) indicate anxiolytic-like effects.

Human Models

The interest in human models of anxiety has been catalyzed to a large extent by findings that panic attacks can be stimulated by a variety of different psychological, physiological, and pharmacological paradigms. Attempts to alter basic anxiety levels and especially those via induction of psychological stress have led to equivocal findings. This might indicate that within the above-mentioned neuroanatomical and physiological systems strong interfering factors exist, which modulate the responses to anxiety and stress.

State and Trait Anxiety

When investigating human anxiety, the distinction between state and trait anxiety is most important and still expanding. State anxiety can be defined as a transitory emotional state consisting of feelings of apprehension, nervousness, and physiological sequelae such as an increased heart rate or respiration (Spielberger, 1979; Endler and Kocovski, 2001). While everyone can experience state anxiety occasionally, there are large differences among individuals in the frequency, duration, and severity. State anxiety can be determined by several rating instruments developed in the past. Trait anxiety represents a fairly stable characteristic related to personality. Experiencing more frequently state anxiety combined with a general view of the world as being threatening and dangerous is used as marker of trait anxiety. The initiation and maintenance of trait anxiety have been related to several factors as outlined above.

Challenge Studies

The profound interest in state anxiety and especially panic attacks stems from a large variety of investigations provoking anxiety and panic attacks experimentally (Nutt and Lawson, 1992). Panic attacks are unique in the spectrum of psychiatric disorders since their core psychopathology of paroxysmal anxiety is temporally limited and can be provoked and investigated under laboratory conditions. Information provided by these studies has led to new cognitive and physiological theories about the basis of panic anxiety and anxiety diseases. Moreover, owing to the experimental character of these investigations, closer comparisons with experiments in animals can be drawn in contrast to other psychiatric animal models. Panic attacks can be elicited by various means, which are listed in Table 1. As indicated, the different paradigms can be differentiated into cognitive, metabotropic, and direct receptor-mediated mechanisms. Especially naturally occurring, cognitive, and metabotropic panic attacks share many features. One of the most amazing findings is that, despite the dramatic anxiety felt, a uniform stress response either of the hypothalamo–pituitary–adrenocortical (HPA, Selye, 1956) or the sympathetic system (Cannon, 1929) is largely missing. These findings led to the hypothesis that in addition to a variety of stimulating agents, strong inhibitors also exist, which physiologically antagonize the altered transmitter and modulator systems involved in panic anxiety. Considering the hypothesis that CRF is one important modulator of anxiety in humans and rodents, it is astonishing that no activation of the HPA system occurs in naturally occurring or metabotropic panic attacks. In contrast, compounds interfering with monoamine and peptide receptors stimulate the HPA activity and the noradrenalin system. Of the latter group, one of the most potent panicogens is CCK-4 (Bradwejn and Vasar, 1995), which seems to exert its effect via CRF. CCK-4 leads dose-dependently to an increase of panic anxiety and a rise of stress hormones. Interestingly, during placebo stimulation following CCK-4 stimulations a psychophysiological conditioning effect could be observed with panic anxiety not accompanied by increases in HPA-axis activity (Hinkelmann et al., 2010). Up to now only a few modulators have been identified, which exert anxiolytic-like effects and, in addition, are able to inhibit the exaggerated HPA system activity. One of these inhibitors might be ANP, which is secreted in the atria of the heart and in various brain regions involved in anxiety. Hence it may be speculated that peptides such as ANP might help to explain the so far unknown mechanisms of terminating panic anxiety (Wiedemann et al., 2001). Despite a tremendously increased knowledge about the induction of anxiety, fear, and stress, the mechanisms of coping and terminating these emotional alterations need further investigations.

Table 1 Experimentally induced panic attacks

Panicogen	*Heart rate stimulation*	*Respiratory stimulation*	*HPA stimulation*	*NE stimulation*
Cognitive stimuli	+	+	−	−
Metabotropic agents				
L-Lactate	+	+	−	−
D-Lactate	+	+	−	−
Bicarbonate	−	+	−	−
CO_2	+/−	+	−	−
Receptor ligands				
Yohimbine	+	−	+	+
Fenfluramine	+	−	+	−
β-Carboline	+	−	+	+
Caffeine	+	+	+	−
CCK-4	+	+	+	(+)
CRF	+	+	+	(+)

HPA, hypothalamo–pituitary–adrenocortical system; NE, noradrenergic system; CCK-4, cholecystokinin tetrapeptide; CRF, corticotropin-releasing factor.

See also: Behavior Therapy: Psychiatric Aspects; Cognitive Behavioral Therapy; Hysteria: History and Critiques; Obsessive-Compulsive Disorder; Psychoanalysis: Current Status; Somatoform Disorders.

Bibliography

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Diseases, fifth ed. American Psychiatric Association, Washington, DC.

Arborelius, L., Owens, M.J., Plotsky, P.M., Nemeroff, C.B., 1999. The role of corticotropin-releasing factor in depression and anxiety disorders. Journal of Endocrinology 160, 1–12.

Bradwejn, J., Vasar, E. (Eds.), 1995. Cholecystokinin and Anxiety: From Neuron to Behavior. Springer, New York.

Cannon, W.B., 1929. Bodily Changes in Pain, Hunger, Fear and Rage. Appleton, New York.

Damasio, A., Carvalho, G.B., 2013. The nature of feelings: evolutionary and neurobiological origins. Neuroscience 14, 143–152.

Dine, J., Ionescu, I.A., Stepan, J., Yen, Y., Holsboer, F., Landgraf, R., Eder, M., Schmidt, U., 2013. Identification of a role for the ventral hippocampus in neuropeptide S-elicited anxiolysis. PLOS ONE 8 (3), 1–8.

Ekman, P. (Ed.), 1982. Emotion in the Human Face, second ed. Cambridge University Press, Cambridge, UK, New York. Editions de la Maison des Sciences de l'Homme, Paris.

Endler, N.S., Kocovski, N.L., 2001. State and trait anxiety revised. Anxiety Disorders 15, 231–245.

Fossati, P., 2012. European Neuropsychopharmacology 22, 487–491.

Gorman, J.M., Kent, J.M., Sullivan, G.M., Coplan, J.D., 2000. Neuroanatomical hypothesis of panic disorder, revised. American Journal of Psychiatry 157, 493–505.

Heilig, M., Widerlöv, E., 1995. Neurobiology and clinical aspects of neuropeptide Y. Critical Reviews in Neurobiology 9, 115–136.

Hinkelmann, K., Yassouridis, A., Mass, R., Tenge, H., Kellner, M., Jahn, H., Wiedemann, K., Wolf, K., 2010. CCK-4: psychophysiological conditioning elicits features of spontaneous panic attacks. Journal of Psychiatric Research 44, 1148–1153.

Jack, R.E., Garrod, O.G.B., Yu, H., Caldara, R., Schyns, P.G., 2013. Facial expressions of emotion are not culturally universal. PNAS 109 (19), 7241–7244.

Johansen, J.P., Cain, C.K., Ostroff, L.E., LeDoux, J.E., 2011. Molecular mechanisms of fear learning and memory. Cell 147, 509–524.

Kent, J.M., Gorman, J.M., Coplan, J.D., 1998. Clinical utility of the selective serotonin reuptake inhibitors in the spectrum of anxiety. Biological Psychiatry 44, 812–824.

Kheirbek, M.A., Klemenhagen, K.C., Sahay, A., Hen, R., 2012. Nature Neuroscience 15 (12), 1613–1620.

Klein, D.F., 1993. False suffocation alarms, spontaneous panics, and related conditions. An integrative hypothesis. Archives of General Psychiatry 50, 306–317.

Koob, G.F., 1999. Corticotropin-releasing factor, norepinephrine, and stress. Biological Psychiatry 46, 1167–1180.

LeDoux, J., 1998. Fear and the brain: where have we been, and where are we going? Biological Psychiatry 44, 1229–1238.

McEwen, B.S., Eiland, L., Hunter, R.G., Miller, M.M., 2012. Stress and anxiety: structural plasticity and epigenetic regulation as a consequence of stress. Neuropharmacology 62, 3–12.

Nutt, D., Lawson, C., 1992. Panic attacks: a neurochemical overview of models and mechanisms. British Journal of Psychiatry 160, 165–178.

Schumacher, J., Suhl Kristensen, A., Wendland, J.R., Nöthen, M.M., Mors, O., McMahon, F.J., 2011. The genetics of panic disorder. Journal of Medical Genetics 48, 361–368.

Scott, K.M., Smith, D.R., Ellis, P.M., 2010. Prospectively ascertained child maltreatment and its association with *DSM-IV* mental disorders in young adults. Archives of General Psychiatry 67 (7), 712–719.

Selye, H., 1956. The Stress of Life. McGraw-Hill, New York.

Spielberger, C.D., 1979. Understanding Stress and Anxiety. Harper and Row, New York.

Steckler, T., Holsboer, F., 1999. Corticotropin-releasing hormone receptor subtypes and emotion. Biological Psychiatry 46, 1480–1508.

Sullivan, G.M., Coplan, J.D., Kent, J.M., Gorman, J.M., 1999. The noradrenergic system in pathological anxiety: a focus on panic with relevance to generalized anxiety and phobias. Biological Psychiatry 46, 1205–1218.

Westenberg, H.G.M., DenBoer, J.A., Murphy, D.L. (Eds.), 1996. Advances in the Neurobiology of Anxiety Disorders. Wiley, Chichester, UK.

Wiedemann, K., Jahn, H., Yassouridis, A., Kellner, M., 2001. Anxiolytic-like effects of atrial, natriuretic peptide on cholecystokinin tetrapeptide-induced panic attacks. Archives of General Psychiatry 58, 371–377.

Anxiety and Fear Conditioning, Neural Basis of

Rheall F Roquet and Marie H Monfils, University of Texas at Austin, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Radulovic, volume 1, pp. 567–570, © 2001, Elsevier Ltd.

Abstract

Fear conditioning (FC) is a model that has been extremely well characterized over the years, and which has proven invaluable in improving our understanding of the neural basis of anxiety-related disorders. The neural mechanisms that contribute to anxiety per se are less easily isolated into components; yet, owing to significant overlap between the neural circuitry that is engaged in FC and many forms of anxiety-related disorders, researchers in this field have uncovered a role for a number of brain regions in dysregulations of the fear system, including the amygdala, hippocampus, and medial prefrontal cortex.

What Is Anxiety?

Anxiety is a state of prolonged anticipatory worry about a potential threat characterized by an exaggerated fear state, often without the presence of immediate danger, in which there is a hyperexcitability of the fear circuit that includes hypervigilance with increased response (Rosen and Schulkin, 1998). Because pathological anxiety is a chronic and prolonged disorder, it is believed to be a dysregulation within the fear circuit leading to neural and behavioral sensitization of this network. Anxiety is believed to develop from initially adaptive fear responses to environmental threats. Such fear responses may include behavioral manifestations such as freezing, startle, defensive withdrawal, or escape, as well as physiological responses, including changes in heart rate, and blood pressure. However, unlike an acute response to an immediate threat, anxiety develops into a cognitively modulated chronic fear state, even without the presence of a threat, which can then cause generalization over time and situations.

One of the biggest challenges to the study of anxiety disorders is that they consist of a cognitively subjective sense of worry or fear that often occurs in the absence of true danger. It is difficult for anxiety to be measured explicitly, especially in nonhuman animal models, because we cannot easily determine, if an animal is feeling chronically anxious. However, some measures of anxiety-like behaviors in animal models have been developed based on innate threats that animals face. In such models, rather than presenting a direct threat, we can manipulate situations in which danger may be possible and measure responses to the situation. Still, anxiety-related responses remain difficult to assess. Rather, most researchers in the field rely on fear-learning paradigms (such as fear conditioning (FC)) to make inferences about anxiety and anxiety-related disorders.

What Is FC?

Pavlovian FC is a form of associative learning in which an innately neutral stimulus, conditioned stimulus (CS, e.g., a tone or context), is paired with an emotionally eliciting stimulus, an unconditioned stimulus (US, e.g., a shock) causing an unconditioned response (e.g., freezing) to the US, which then becomes a conditioned response (CR, e.g., freezing). This process creates an association between the CS and US from which the presentation of the CS alone will begin to elicit the CR without the US. FC is one of the most studied forms of learning because it occurs quickly, a single CS–US pairing is sufficient to create a long lasting and stable memory. In addition, the brain mechanisms involved in FC have been extensively studied and mapped over the last few decades, and overlap greatly with those engaged in anxiety-related disorders.

Why Is FC Important and How Can It Be Used to Better Understand Anxiety?

The neurological roots of anxiety lie within the same innate responses found in FC, as such FC provides a more direct way to better understand some of the mechanisms that underlie both fear and anxiety. One of the biggest benefits to the study of FC is that the different stages of conditioning can be isolated and targeted to better understand the specific mechanisms that regulate learning and associated brain responses. The most common way fear is examined is through freezing (that is, cessation of movement in the presence of a threatening stimulus), which can be reliably measured across species.

Brain Regions Implicated in Different Phases of FC

Acquisition

Acquisition is the initial phase of FC in which there is convergence of information about the CS and the US, which leads to associative learning and plasticity.

Two of the most common ways in which FC is studied is through the use of either a CS to a discrete cue (such as a tone) or a context. While both conditioning paradigms work in similar fashions, the sensory information from each stimulus is integrated in different ways. In a condition in which a tone is used as the CS, inputs from the medial geniculate nucleus of the thalamus and auditory association cortex (TE3) converge in the lateral amygdala (LA) where the auditory and somatosensory inputs are then processed (Doron and Ledoux, 1999). Alternatively, in a conditioning paradigm in which a context is used as the CS, projections from the hippocampus to the

amygdala have been identified to be important for integrating the two stimuli (Phillips and LeDoux, 1992; Maren et al., 1998). Additional work has also found that the hippocampus plays a role in trace FC, which determines how time between the introduction of the US and CS effects conditioning (Phillips and LeDoux, 1992; McEchron et al., 1998).

However, through more refined techniques, which allow functional inactivation of subregions through GABA agonists, for brief periods of time, region specificity has been further examined. Wilensky et al. (1999) were able to show that functional inactivation of both the LA and basolateral amygdala (BLA) subnuclei could block fear acquisition. Yet, when those regions were blocked after initial acquisition there was no effect on consolidation of the memory, distinguishing these regions specific roll in the acquisition of a fear memory.

Additionally work has been done to confirm findings from nonhuman animal studies through human imaging, identifying that both the amygdala and hippocampus become active during acquisition (LaBar et al., 1998; Knight et al., 2004). Additional support of fear acquisition in humans has come from trace conditioning, which requires a larger cognitive load. Imaging studies have shown that again both the amygdala and hippocampus become activated during acquisition (Büchel et al., 1999).

In one study, Phelps and colleagues instructed participants that one stimulus would predict shock whereas a second stimulus would not and found that in fact the left amygdala of participants became activated when the stimulus they were told would predict shock was presented (Phelps et al., 2001). This result is consistent with research done with participants who have had left temporal lobotomy who also show deficits in their ability to acquire fear from instruction rather than experience (Funayama et al., 2001). Work by Anderson and Phelps (2002) found that in fact patients that have had lesions to their amygdala can still experience emotion and both positive and negative affective states. This comes with the fact that patients cannot, however, learn new fear associations (LaBar et al., 1995; Weike et al., 2005).

Consolidation

Another important aspect of the formation of a fear memory is that an association between the two stimuli is created and a memory is retained. If during conditioning, the aversive US is introduced with the neutral CS and in that instant a response to the US is made, yet an association is not acquired and retained into a memory no conditioning may have occurred.

Over the last decade, lot of work has been done to determine what regions are specifically involved in the consolidation process because unlike other aspects of conditioning it must be targeted in a time-sensitive way, after initial acquisition but not effect expression. Through methodologies previously discussed functional inactivation immediately posttraining are used to access what brain regions are important for consolidation of fear-related memories.

As expected contextual and cued FC show differences in region specificity of consolidation. Using a neurotoxin to functionally inactivate different regions, the BLA, Entorhinal Cortex (EC), and Perirhinal Cortex (PC) resulted in deficits in consolidation of cued FC, while inactivating the thalamus did not have an effect (Sacchetti et al., 1999). Conversely, similarities of functional inactivation of the BLA, EC, and PC were also seen in contextual FC, however, in addition inactivation of the hippocampus caused deficits in consolidation.

Expression

The BLA was one of the first regions assessed for its effects on expression of fear, as it is the region in which information from the CS and US in both cued and contextual fear. Research has supported, with lesions to the BLA 1 or 15 days after training can eliminate freezing expression to both a context and cued fear (Cousens and Otto, 1998; Angladа-Figueroa and Quirk, 2005). Similarly, functional inactivation of the BLA resulted in attenuated fear expression if induced 24 h after training but before test (Helmstetter and Bellgowan, 1994). However, the action of the BLA is still under debate because, while lesioning or inactivating the BLA shows effects on expression, it is not known if the BLA has a direct behavioral contribution to expression, because it has both direct and indirect connections to the Central Nucleus of the Amygdala (CeA) (LeDoux, 2000). Whereas the CeA is known to have direct connections to the startle circuit, including the periaqueductal gray, lateral hypothalamus, and paraventricular nucleus of the hypothalamus, which modulate the physiological and behavioral responses to FC (LeDoux, 2000; Maren, 2005; Kim and Jung, 2006).

Differently from other aspects of FC, a few new regions of interest become important for fear expression including, the Ventral Tagmental Area (VTA), and Nucleus Accumbens (NAcc) (see review, Pezze and Feldon, 2004). These regions contribute to the motor output of fear expression. In FC, the NAcc receives input both from the VTA as well as the BLA, from which it then can relay information to other regions of the basal ganglia to evoke motor output.

Additionally, within recent years a greater interest has begun to examine the role of the Prefrontal Cortex (PFC) plays in FC. Research in fact has found that the medial prefrontal cortex (mPFC) may be required to express proper fear responses in FC (Lee and Choi, 2012). Similarly to the specialties of the amygdala subregions, the PFC subregions also show distinctions in their roles. While the infralimbic subregion of the mPFC works by inhibiting amygdala output, the prelimbic mPFC works to modulate expression through exciting amygdala output (Vidal-Gonzalez et al., 2006).

Extinction

One of the most important reasons to study FC is not just to better understand fear, but also how to eliminate fear or the aspects that contribute to fear and anxiety-related disorders. Extinction is the most commonly practiced technique to attempt to alter the effects of the original CS–US relationship. Through presentations of the CS over multiple trials without the aversive US, extinction training leads to decreases in the CR to the presentation of the CS. Data support that similar to initial fear learning, the amygdala acts to acquire and store extinction memories, while the hippocampus processes contextual information, and the infralimbic region of the

mPFC engages in consolidation and the retrieval of extinction memory (Myers and Davis, 2006; Quirk and Mueller, 2008).

Supporting the translational relationship between animal work and humans, the amygdala as well as the hippocampus are critical to extinction in humans (Knight et al., 2004). In addition, the mPFC is necessary in extinction, possibly through inhibitory signaling, this corresponding to work that has found that the orbitofrontal and ventromedial prefrontal cortices become activated during extinction learning and recall (Phelps et al., 2004; Gottfried and Dolan, 2004; Milad et al., 2007; Kalisch et al., 2006).

However, one concern with fear extinction is that it is a form of new learning, and has all the components of the original FC, so rather than changing the old memory extinction is creating a new memory that competes with the old association, from which you hope that when retrieving a memory upon presentation of the CS, one retrieves the safety memory (or extinguished memory) (Bouton, 2002). However, when the CS is presented after extinction spontaneous recovery can occur, in which with the passage of time rather than the safety memory being expressed the fear response is elicited (Rescorla, 2004). Alternatively, if the US is presented but not signaled with the original CS reinstatement of the original association can occur, causing an expression of fear in the presence of the CS without reconditioning (Delamater, 1997). Finally, renewal can occur, in which the fear response is elicited in the context it was learned but not in the context in which it was extinguished, this is also an issue with generalizability of extinction (Effting and Kindt, 2007).

Reconsolidation

When new memories are formed consolidation must occur over time, to create long-term memories (LTMs). Similarly, when memories are retrieved from LTM they become progressively unstable (Hong et al., 2013; Kim et al., 2010), at which point they must then be reconsolidated into LTMs once again. While, within a window of accessing a previously established memory and it becoming reconsolidated the memory is considered unstable and in a labile state, and by such can be altered or strengthened before being reconsolidated into LTMs again. Adaptively, this makes sense because if you encounter the same situations over time you want to keep updating existing information rather than creating a new memory or association every time similar circumstances arise. This also facilitates an alternative to extinction for the reduction of fear responses to a stimuli and a way to update rather the overwrite an already established memory. Research has found that this both occurs within the hippocampus for contextual FC as well as the amygdala for both context and cued FC, both of which occur in a temporally graded manner (Debiec et al., 2002; Parsons et al., 2006).

Neural Mechanisms and Modulators Implicated in FC

With the development of better and more precise techniques to assess region specificity for different behaviors and/or disorders, we have begun to identify some of the neural modulators and mechanisms that are critical in FC (for a recent comprehensive review see Johansen et al., 2011). One neural modulator, however, has been found to play an integral role in learning, FC, and as we later discuss anxiety, the ionotropic glutamatergic receptor *N*-methyl-D-aspartate (NMDA) (Tsien, 2000). In FC, NMDA receptors in the LAs have been found to play multiple roles, including roles in both contextual and cued fear, acquisition, expression, and extinction (Rodrigues et al., 2001; Maren et al., 1996; Zimmerman and Maren, 2010). In the expression of fear responses, NMDA receptors within the NAcc modulate Dopamine (D) within the VTA under stressful conditions, potentially working in a feedback manor (Doherty and Gratton, 1997). Consolidation of extinction can be impaired through blockade of NMDA receptors in the mPFC or reduction of Brain-derived Neurotropic Factor signaling within the amygdala (Burgos-Robles et al., 2007; Chhatwal et al., 2006).

Additionally the neural process of plasticity and long-term potentiation (LTP) function though NMDA modulation between synapses (Samson and Paré, 2005), through L-type voltage-gated calcium channels on NMDA signaling between the thalamus to the amygdala (Weisskopf et al., 1999). The process of neural plasticity and LTP is believed to be important in FC, in that after an association has occurred the strength of that association follows the Hebbian theory of plasticity, in which, when a presynaptic neuron fires causing a response in a postsynaptic neuron it increase the strength or relationship between the two neurons (Hebb, 1949). In associative learning, it is interpreted that if you have a strong presynaptic input from one neuron, and another presynaptic neuron also acts on the same postsynaptic neuron it will increase the relationship between the strong acting neuron as well as the weaker acting neuron, through an increased nonspecific association. This is how an association between a CS and US is strengthened to elicit the same response, even though prior to experience with the CS and US, the CS did not elicit the CR.

The effects of plasticity have been shown through changes in synaptic transmission and strength between pre- and postsynaptic neurons in both contextual and cue FC between the hippocampus to amygdala, and thalamus to the amygdala, respectively (Maren and Fanselow, 1995; McKernan and Shinnick-Gallagher, 1997; Quirk et al., 1997). For example, previously unresponsive cells in cued conditioning can become tone-responsive cells with training (Quirk et al., 1995). Additionally, other mechanisms have been found to be important in plasticity, norepinephrine (NE) works through trafficking alpha-amino-3-hydroxy-5-methyl-4-isoxazoleproprionic acid (AMPA)-type glutamate receptors AMPA to the synapse, which in turn decreases the threshold necessary for LTP to occur (Tully and Bolshakov, 2010; Hu et al., 2007).

Other phases of FC show differences in the neural mechanisms that contribute to their processes. In reconsolidation, the blockade of downstream mitogen-activated protein kinase cascade within the amygdala has an effect on protein modulation within a cell and can be targeted to inhibit reconsolidation (Duvarci et al., 2005). This function has been found to be important because of its use in fear memory updating, similar to initial consolidation when protein synthesis is inhibited within the consolidation window it can block the memory and specifically in reconsolidation block spontaneous recovery, reinstatement, and renewal (Schafe and LeDoux,

2000; Duvarci and Nader, 2004). However, non-pharmacological interventions have been shown to modulate this neural process. One such technique employs both reconsolidation, by introduction of a single retrieval session to reactive initial memory trace, then an extinction trial is used to modulate the initial fear memory within the critical reconsolidation window, such that the memory is now updated to the CS no longer predicts the US (Monfils et al., 2009).

Neural and Behavioral Mechanisms of Anxiety

While FC is one of the most commonly and well-respected ways of studying fear, other behavior paradigms can be employed to study fear and anxiety. Similarly, to FC anxiety can be studied in animal models to better understand innate behavioral and emotion responses under a specific circumstance. One challenge to the study of anxiety-like tasks in animal models is that they are not necessarily a direct comparison to human anxiety in the same way that fear responses are between nonhuman animals and humans. As previously mentioned, pathological anxiety in humans is a chronic and relapsing disorder; however, the current models of anxiety-like behavior in animals measure acute interactions of animals when introduced to a biologically significant challenge. Some work, however, has tried to use chronic stress models to emulate the chronic nature of anxiety; however, this will not be discussed here. Anxiety-like measures do offer some benefits and similarities to human anxiety, because behavior can both be assessed under conditioned or unconditioned terms, so one can measure learning and associations as well as an ecologically influenced choice of response.

One paradigm used to assess fear and anxiety through conditioning is fear-potentiated startle (FPS), in which a loud noise (aversive stimulus) is presented with a CS previously paired with a US and differences in startle response to the loud noise alone verse the loud noise plus CS is assessed. Similarly, to basic FC lesions to the amygdala block FPS (Kim and Davis, 1993). More specifically, the blockage of NMDA receptors within the amygdala block both acquisition and extinction of FPS, with some support that it can also block expression (Campeau et al., 1992; Falls et al., 1992; Fendt, 2001). Consistent with FC, FPS employs the mescorticolimbic pathway in startle responses, and both functional inactivation as well as dopaminergic blockade within this pathway can attenuate expression (Munro and Kokkinidis, 1997; de Oliveira et al., 2009).

One of the most commonly used tests to assess anxiety-like behavior is an unconditioned task, the Elevated-Plus Maze (EPM). This task examines the challenge of an animal to explore novel environments (innately desirable), at the cost of two of the four arms, being open to the outside environment, creating vulnerability and potential danger (Montgomery, 1955). This test has been integral in development of new pharmacological tools to understand anxiety disorders. The most well-documented include benzodiazepines, which work through enhancing GABAnergic transmission, as well as buspirone, which functions as a partial agonist of serotonin and dopamine. Both types of pharmacological manipulations increase open arm entries and as such are believed to reduce anxiety (Cole and Rodgers, 1993, 1994, 1995). Additionally, the EPM task and other anxiety-like measures have provided availability to understand what factors influence increases in anxiety, such as social defeat or isolation (Heinrichs et al., 1992; Maisonnette et al., 1993), restraint (Martijena et al., 1997), or previous shock exposure (Korte et al., 1999).

Anxiety-like tasks have been essential in finding neural mechanisms through which anxiety is modulated. Corticotrophin-releasing factor (CRF) and hyperactivity in the hypothalamic-pituitary-adrenal (HPA) axis have been found to influence fear and anxiety (Lelas et al., 2004). Additionally, the glutamatergic system has been found to be involved not only in FC but anxiety, with the use of antagonists for the mGluR5 subunit of glutamate receptor reducing anxiety-like behavior on anxiety tasks (Brodkin et al., 2002).

Neuronal activation upon exposure to the EPM has identified regions involved in anxiety that also overlap with FC, including; the amygdala, hippocampus, thalamus, and periaqueductal gray matter, as well as other regions within the brain stem (Silveira et al., 1993). One area that has been shown to be differentially activated in anxiety from FC is the noradrenergic Locus Coeruleus and may provide a good target for anxiety treatment (Chrousos and Gold, 1992).

One research tool that is being developed is animal models that encompass similar pathologies of human anxiety. Steps for the future of developing these pathological models of anxiety in animals come for the fact that anxiety disorders run in families, supporting the idea that there may be a genetic influence on the development of anxiety disorders (Merikangas and Pine, 2002). Current data from twin studies provide evidence that anxiety disorders are heritable (Hettema et al., 2001; Skre et al., 2000; Goldberg et al., 1990). Studies of children whose parents have been diagnosed with an anxiety disorder have found that the children exhibit greater startle responses, indicative of an increased anxiety response (Grillon et al., 1998). However, it has also been found that genetic disposition is not the only factor that plays into the development of anxiety disorder that environmental influences play an important role as well (Hettema et al., 2005).

Post-traumatic stress disorder (PTSD) is one of the most common anxiety disorders effecting approximately 7% of people over a lifetime (Kessler et al., 2005). Closely following similar acquisition of Pavlovian FC, PTSD is believed to develop when a person is exposed to a traumatic event, which then leads to a chronic anxiety state, as well as potentially direct or generalized associations between stimuli and the trauma. Patients with PTSD have been found to have increased startle and aversive CRs as well as resistance to extinction (Grillon and Morgan III, 1999; Jovanovic et al., 2010). These behavioral metrics may be due to increases in heart rate and skin conductance at baseline, which may be indicative of an overactive sympathetic nervous system and HPA axis. These findings also follow patterns of anxiety in animal models.

Also contributing to the understanding of PTSD in human imaging studies that expand on the fact that in both human and animal models of FC show an importance of the PFC for extinction in healthy individuals. Consistent with those findings, patients with PTSD show deficits in activation in both the mPFC and the anterior cingulate while engaged in emotional tasks (Shin et al., 2005, 2001). Additional alterations include

increased amygdala activation during acquisition, as well as exaggerated responses in the putamen, insula, caudate, and amygdala to the US (Bremner et al., 2005; Linnman et al., 2011; Shin et al., 2005).

Consistent with PTSD patients, patients with panic disorder have also been found to show increased startle responses in a FPS test (Grillon et al., 1994). Patients with generalized anxiety disorder also show analogous deficits in anterior cingulate activation and connectivity with the amygdala within an emotional processing task (Etkin et al., 2010).

Conclusion

In conclusion, there are many overlapping neural mechanisms and brain regions specifically involved in FC and anxiety. This is seen through both assessing the isolated phases of FC, as well as through multiple anxiety-like tasks in nonhuman animals, while demonstrating support for the neural correlates in human imaging and case studies. These include the amygdala, hippocampus, thalamus, periaqueductal gray, and mPFC, which have been shown to be important in both FC and anxiety, see Table 1. Additionally, neural mechanisms in the mesocorticolimbic pathway modulating dopamine are important for fear expression and FPS responses in nonhuman animals. It is clear that studying basic fear models has dramatically improved our understanding of anxiety-related disorders over the years. As we proceed forward, many researchers in the field believe that continued interactions between basic researchers and clinicians will facilitate the development of better tools to help individuals with anxiety-related disorders.

Table 1 Brain regions discussed to be involved in fear conditioning, all bolded regions are those that overlap regions discussed to be involved in anxiety

Phases of fear conditioning	*Brain regions*
Acquisition	**Amygdala**
	Auditory association cortex
	Hippocampus
	Thalamus
Consolidation	Amygdala
	Hippocampus
	Medial temporal lobe
Expression	**Amygdala**
	Hypothalamus
	Periaqueductal gray
	Medial prefrontal cortex
	Nucleus accumbens
	Ventral tagmental area
Extinction	**Amygdala**
	Hippocampus
	Medial prefrontal cortex
Reconsolidation	Amygdala
	Hippocampus

Note. For review of all brain regions involved in fear conditioning and anxiety, see, Kim, J.J., Jung, M.W., 2006. Neural circuits and mechanisms involved in Pavlovian fear conditioning: a critical review. Neuroscience & Biobehavioral Reviews 30 (2), 188–202; LeDoux, J.E., 2000. Emotion circuits in the brain. Annual Review of Neuroscience 23 (1), 155–184; Shin, L.M., Liberzon, I., 2009. The neurocircuitry of fear, stress, and anxiety disorders. Neuropsychopharmocology 35 (1), 169–191.

See also: Emotion, Neural Basis of; Fear Conditioning: Overview; Fear: Psychological and Neural Aspects.

Bibliography

Anderson, A.K., Phelps, E.A., 2002. Is the human amygdala critical for the subjective experience of emotion? Evidence of intact dispositional affect in patients with amygdala lesions. Journal of Cognitive Neuroscience 14 (5), 709–720.

Anglada-Figueroa, D., Quirk, G.J., 2005. Lesions of the basal amygdala block expression of conditioned fear but not extinction. The Journal of Neuroscience 25 (42), 9680–9685.

Bouton, M.E., 2002. Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. Biological Psychiatry 52 (10), 976–986.

Bremner, J.D., Vermetten, E., Schmahl, C., Vaccarino, V., Vythilingam, M., Afzal, N., Grillon, C., Charney, D.S., 2005. Positron emission tomographic imaging of neural correlates of a fear acquisition and extinction paradigm in women with childhood sexual-abuse-related post-traumatic stress disorder. Psychological Medicine 35 (6), 791–806.

Brodkin, J., Busse, C., Sukoff, S.J., Varney, M.A., 2002. Anxiolytic-like activity of the mGluR5 antagonist MPEP: a comparison with diazepam and buspirone. Pharmacology Biochemistry and Behavior 73 (2), 359–366.

Büchel, C., Dolan, R.J., Armony, J.L., Friston, K.J., 1999. Amygdala–hippocampal involvement in human aversive trace conditioning revealed through event-related functional magnetic resonance imaging. The Journal of Neuroscience 19 (24), 10869–10876.

Burgos-Robles, A., Vidal-Gonzalez, I., Santini, E., Quirk, G.J., 2007. Consolidation of fear extinction requires NMDA receptor-dependent bursting in the ventromedial prefrontal cortex. Neuron 53 (6), 871–880.

Campeau, S., Miserendino, M.J., Davis, M., 1992. Intra-amygdala infusion of the N-methyl-D-aspartate receptor antagonist AP5 blocks acquisition but not expression of fear-potentiated startle to an auditory conditioned stimulus. Behavioral Neuroscience 106 (3), 569.

Chhatwal, J.P., Stanek-Rattiner, L., Davis, M., Ressler, K.J., 2006. Amygdala BDNF signaling is required for consolidation but not encoding of extinction. Nature Neuroscience 9 (7), 870–872.

Chrousos, G.P., Gold, P.W., 1992. The concepts of stress and stress system disorders. JAMA: The Journal of the American Medical Association 267 (9), 1244–1252.

Cole, J.C., Rodgers, R.J., 1993. An ethological analysis of the effects of chlordiazepoxide and bretazenil (Ro 16-6028) in the murine elevated plus-maze. Behavioural Pharmacology 4 (6), 573–580.

Cole, J.C., Rodgers, R.J., 1994. Ethological evaluation of the effects of acute and chronic buspirone treatment in the murine elevated plus-maze test: comparison with haloperidol. Psychopharmacology 114 (2), 288–296.

Cole, J.C., Rodgers, R.J., 1995. Ethological comparison of the effects of diazepam and acute/chronic imipramine on the behaviour of mice in the elevated plus-maze. Pharmacology Biochemistry and Behavior 52 (3), 473–478.

Cousens, G., Otto, T., 1998. Both pre-and posttraining excitotoxic lesions of the basolateral amygdala abolish the expression of olfactory and contextual fear conditioning. Behavioral Neuroscience 112 (5), 1092.

Debiec, J., LeDoux, J.E., Nader, K., 2002. Cellular and systems reconsolidation in the hippocampus. Neuron 36 (3), 527–538.

Delamater, A.R., 1997. Selective reinstatement of stimulus-outcome associations. Animal Learning & Behavior 25 (4), 400–412.

Doherty, M.D., Gratton, A., 1997. NMDA receptors in nucleus accumbens modulate stress-induced dopamine release in nucleus accumbens and ventral tegmental area. Synapse 26 (3), 225–234.

Doron, N.N., Ledoux, J.E., 1999. Organization of projections to the lateral amygdala from auditory and visual areas of the thalamus in the rat. The Journal of Comparative Neurology 412 (3), 383–409.

Duvarci, S., Nader, K., 2004. Characterization of fear memory reconsolidation. The Journal of Neuroscience 24 (42), 9269–9275.

Duvarci, S., Nader, K., LeDoux, J.E., 2005. Activation of extracellular signal-regulated kinase–mitogen-activated protein kinase cascade in the amygdala is required for memory reconsolidation of auditory fear conditioning. European Journal of Neuroscience 21 (1), 283–289.

Effting, M., Kindt, M., 2007. Contextual control of human fear associations in a renewal paradigm. Behaviour Research and Therapy 45 (9), 2002–2018.
Etkin, A., Prater, K.E., Hoeft, F., Menon, V., Schatzberg, A.F., 2010. Failure of anterior cingulate activation and connectivity with the amygdala during implicit regulation of emotional processing in generalized anxiety disorder. American Journal of Psychiatry 167 (5), 545–554.
Falls, W.A., Miserendino, M.J., Davis, M., 1992. Extinction of fear-potentiated startle: blockade by infusion of an NMDA antagonist into the amygdala. The Journal of Neuroscience 12 (3), 854–863.
Fendt, M., 2001. Injections of the NMDA receptor antagonist aminophosphonopentanoic acid into the lateral nucleus of the amygdala block the expression of fear-potentiated startle and freezing. The Journal of Neuroscience 21 (11), 4111–4115.
Funayama, E.S., Grillon, C., Davis, M., Phelps, E.A., 2001. A double dissociation in the affective modulation of startle in humans: effects of unilateral temporal lobectomy. Journal of Cognitive Neuroscience 13 (6), 721–729.
Goldberg, J., True, W.R., Eisen, S.A., Henderson, W.G., 1990. A twin study of the effects of the Vietnam War on posttraumatic stress disorder. JAMA: the journal of the American Medical Association 263 (9), 1227–1232.
Gottfried, J.A., Dolan, R.J., 2004. Human orbitofrontal cortex mediates extinction learning while accessing conditioned representations of value. Nature Neuroscience 7 (10), 1144–1152.
Grillon, C., Morgan III, C.A., 1999. Fear-potentiated startle conditioning to explicit and contextual cues in Gulf War veterans with posttraumatic stress disorder. Journal of Abnormal Psychology 108 (1), 134.
Grillon, C., Ameli, R., Goddard, A., Woods, S.W., Davis, M., 1994. Baseline and fear-potentiated startle in panic disorder patients. Biological Psychiatry 35 (7), 431–439.
Grillon, C., Dierker, L., Merikangas, K.R., 1998. Fear-potentiated startle in adolescent offspring of parents with anxiety disorders. Biological Psychiatry 44 (10), 990–997.
Grillon, C., Pine, D.S., Lissek, S., Rabin, S., Bonne, O., Vythilingam, M., 2009. Increased anxiety during anticipation of unpredictable aversive stimuli in post-traumatic stress disorder but not in generalized anxiety disorder. Biological Psychiatry 66 (1), 47–53.
Hebb, D.O., 1949. The Organization of Behavior: A Neuropsychological Approach. John Wiley & Sons.
Heinrichs, S.C., Pich, E.M., Miczek, K.A., Britton, K.T., Koob, G.F., 1992. Corticotropin-releasing factor antagonist reduces emotionality in socially defeated rats via direct neurotropic action. Brain Research 581 (2), 190–197.
Helmstetter, F.J., Bellgowan, P.S., 1994. Effects of muscimol applied to the basolateral amygdala on acquisition and expression of contextual fear conditioning in rats. Behavioral Neuroscience 108 (5), 1005.
Hettema, J.M., Prescott, C.A., Kendler, K.S., 2001. A population-based twin study of generalized anxiety disorder in men and women. The Journal of Nervous and Mental Disease 189 (7), 413–420.
Hettema, J.M., Prescott, C.A., Myers, J.M., Neale, M.C., Kendler, K.S., 2005. The structure of genetic and environmental risk factors for anxiety disorders in men and women. Archives of General Psychiatry 62 (2), 182.
Hong, I., Kim, J., Kim, J., Lee, S., Ko, H.G., Nader, K., Kaang, B.K., Tsien, R.W., Choi, S., 2013. AMPA receptor exchange underlies transient memory destabilization on retrieval. Proceedings of the National Academy of Sciences 110 (20), 8218–8223.
Hu, H., Real, E., Takamiya, K., Kang, M.G., Ledoux, J., Huganir, R.L., Malinow, R., 2007. Emotion enhances learning via norepinephrine regulation of AMPA-receptor trafficking. Cell 131 (1), 160–173.
Johansen, J.P., Cain, C.K., Ostroff, L.E., LeDoux, J.E., 2011. Molecular mechanisms of fear learning and memory. Cell 147 (3), 509–524.
Jovanovic, T., Norrholm, S.D., Blanding, N.Q., Phifer, J.E., Weiss, T., Davis, M., Duncan, E., Bradlet, B., Ressler, K., 2010. Fear potentiation is associated with hypothalamic–pituitary–adrenal axis function in PTSD. Psychoneuroendocrinology 35 (6), 846–857.
Kalisch, R., Korenfeld, E., Stephan, K.E., Weiskopf, N., Seymour, B., Dolan, R.J., 2006. Context-dependent human extinction memory is mediated by a ventromedial prefrontal and hippocampal network. The Journal of Neuroscience 26 (37), 9503–9511.
Kessler, R.C., Berglund, P., Demler, O., Jin, R., Merikangas, K.R., Walters, E.E., 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. Archives of General Psychiatry 62 (6), 593.
Kim, J.J., Jung, M.W., 2006. Neural circuits and mechanisms involved in Pavlovian fear conditioning: a critical review. Neuroscience & Biobehavioral Reviews 30 (2), 188–202.
Kim, J., Song, B., Hong, I., Kim, J., Lee, J., Park, S., Eom, J.Y., Lee, C.J., Lee, S., Choi, S., 2010. Reactivation of fear memory renders consolidated amygdala synapses labile. The Journal of Neuroscience 30 (28), 9631–9640.
Kim, M., Davis, M., 1993. Lack of a temporal gradient of retrograde amnesia in rats with amygdala lesions assessed with the fear-potentiated startle paradigm. Behavioral Neuroscience 107 (6), 1088.
Knight, D.C., Smith, C.N., Cheng, D.T., Stein, E.A., Helmstetter, F.J., 2004. Amygdala and hippocampal activity during acquisition and extinction of human fear conditioning. Cognitive, Affective, & Behavioral Neuroscience 4 (3), 317–325.
Korte, S.M., De Boer, S.F., Bohus, B., 1999. Fear-potentiation in the elevated plus-maze test depends on stressor controllability and fear conditioning. Stress: The International Journal on the Biology of Stress 3 (1), 27–40.
LaBar, K.S., Gatenby, J.C., Gore, J.C., LeDoux, J.E., Phelps, E.A., 1998. Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. Neuron 20 (5), 937–945.
LaBar, K.S., LeDoux, J.E., Spencer, D.D., Phelps, E.A., 1995. Impaired fear conditioning following unilateral temporal lobectomy in humans. The Journal of Neuroscience 15 (10), 6846–6855.
LeDoux, J.E., 2000. Emotion circuits in the brain. Annual Review of Neuroscience 23 (1), 155–184.
Lee, Y.K., Choi, J.S., 2012. Inactivation of the medial prefrontal cortex interferes with the expression but not the acquisition of differential fear conditioning in rats. Experimental Neurobiology 21 (1), 23–29.
Lelas, S., Wong, H., Li, Y.W., Heman, K.L., Ward, K.A., Zeller, K.L., Sieracki, K.K., Polino, J.L., Godonis, H.E., Ren, S.X., Yan, X., Arneric, S.P., Robertson, D.W., Hartig, P.R., Grossman, S., Trainor, G.L., Taub, R.A., Zaczek, R., Gilligan, P.J., McElroy, J.F., 2004. Anxiolytic-like effects of the corticotropin-releasing factor1 (CRF1) antagonist DMP904 [4-(3-pentylamino)-2, 7-dimethyl-8-(2-methyl-4-methoxyphenyl)-pyrazolo-[1, 5-a]-pyrimidine] administered acutely or chronically at doses occupying central CRF1 receptors in rats. Journal of Pharmacology and Experimental Therapeutics 309 (1), 293–302.
Linnman, C., Zeffiro, T.A., Pitman, R.K., Milad, M.R., 2011. An fMRI study of unconditioned responses in post-traumatic stress disorder. Biology of Mood & Anxiety Disorders 1 (8), 1–12.
Maisonnette, S., Morato, S., Brandao, M.L., 1993. Role of resocialization and of 5-HT 1A receptor activation on the anxiogenic effects induced by isolation in the elevated plus-maze test. Physiology & Behavior 54 (4), 753–758.
Maren, S., 2005. Synaptic mechanisms of associative memory in the amygdala. Neuron 47 (6), 783–786.
Maren, S., Fanselow, M.S., 1995. Synaptic plasticity in the basolateral amygdala induced by hippocampal formation stimulation in vivo. The Journal of neuroscience 15 (11), 7548–7564.
Maren, S., Aharonov, G., Stote, D.L., Fanselow, M.S., 1996. N-methyl-D-aspartate receptors in the basolateral amygdala are required for both acquisition and expression of conditional fear in rats. Behavioral Neuroscience 110 (6), 1365.
Maren, S., Anagnostaras, S.G., Fanselow, M.S., 1998. The startled seahorse: is the hippocampus necessary for contextual fear conditioning? Trends in Cognitive Sciences 2 (2), 39–42.
Martijena, I.D., Calvo, N., Volosin, M., Molina, V.A., 1997. Prior exposure to a brief restraint session facilitates the occurrence of fear in response to a conflict situation: behavioral and neurochemical correlates. Brain Research 752 (1), 136–142.
McEchron, M.D., Bouwmeester, H., Tseng, W., Weiss, C., Disterhoft, J.F., 1998. Hippocampectomy disrupts auditory trace fear conditioning and contextual fear conditioning in the rat. Hippocampus 8 (6), 638–646.
McKernan, M.G., Shinnick-Gallagher, P., 1997. Fear conditioning induces a lasting potentiation of synaptic currents in vitro. Nature 390 (6660), 607–611.
Merikangas, K.R., Pine, D., 2002. Genetic and other vulnerability factors for anxiety and stress disorders. Neuropsychopharmacology: the fifth generation of progress. American College of Neuropsychopharmacology 867–882.
Milad, M.R., Wright, C.I., Orr, S.P., Pitman, R.K., Quirk, G.J., Rauch, S.L., 2007. Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. Biological Psychiatry 62 (5), 446–454.
Monfils, M.H., Cowansage, K.K., Klann, E., LeDoux, J.E., 2009. Extinction-reconsolidation boundaries: key to persistent attenuation of fear memories. Science 324 (5929), 951–955.
Montgomery, K.C., 1955. The relation between fear induced by novel stimulation and exploratory drive. Journal of Comparative and Physiological Psychology 48 (4), 254.
Munro, L.J., Kokkinidis, L., 1997. Infusion of quinpirole and muscimol into the ventral tegmental area inhibits fear-potentiated startle: implications for the role of dopamine in fear expression. Brain Research 746 (1), 231–238.
Myers, K.M., Davis, M., 2006. Mechanisms of fear extinction. Molecular Psychiatry 12 (2), 120–150.
de Oliveira, A.R.D., Reimer, A.E., Brandão, M.L., 2009. Role of dopamine receptors in the ventral tegmental area in conditioned fear. Behavioural Brain Research 199 (2), 271–277.

Parsons, R.G., Gafford, G.M., Baruch, D.E., Riedner, B.A., Helmstetter, F.J., 2006. Long-term stability of fear memory depends on the synthesis of protein but not mRNA in the amygdala. European Journal of Neuroscience 23 (7), 1853–1859.

Pezze, M.A., Feldon, J., 2004. Mesolimbic dopaminergic pathways in fear conditioning. Progress in Neurobiology 74 (5), 301–320.

Phelps, E.A., Delgado, M.R., Nearing, K.I., LeDoux, J.E., 2004. Extinction learning in humans: role of the amygdala and vmPFC. Neuron 43 (6), 897–905.

Phelps, E.A., O'Connor, K.J., Gatenby, J.C., Gore, J.C., Grillon, C., Davis, M., 2001. Activation of the left amygdala to a cognitive representation of fear. Nature Neuroscience 4 (4), 437–441.

Phillips, R.G., LeDoux, J.E., 1992. Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning. Behavioral Neuroscience 106 (2), 274.

Quirk, G.J., Mueller, D., 2008. Neural mechanisms of extinction learning and retrieval. Neuropsychopharmacology 33 (1), 56–72.

Quirk, G.J., Armony, J.L., LeDoux, J.E., 1997. Fear conditioning enhances different temporal components of tone-evoked spike trains in auditory cortex and lateral amygdala. Neuron 19 (3), 613–624.

Quirk, G.J., Repa, J.C., LeDoux, J.E., 1995. Fear conditioning enhances short-latency auditory responses of lateral amygdala neurons: parallel recordings in the freely behaving rat. Neuron 15 (5), 1029–1039.

Rescorla, R.A., 2004. Spontaneous recovery. Learning & Memory 11 (5), 501–509.

Rodrigues, S.M., Schafe, G.E., LeDoux, J.E., 2001. Intra-amygdala blockade of the NR2B subunit of the NMDA receptor disrupts the acquisition but not the expression of fear conditioning. The Journal of Neuroscience 21 (17), 6889–6896.

Rosen, J., Schulkin, J., 1998. From normal fear to pathological anxiety. Psychological Review 105 (2), 325.

Sacchetti, B., Lorenzini, C.A., Baldi, E., Tassoni, G., Bucherelli, C., 1999. Auditory thalamus, dorsal hippocampus, basolateral amygdala, and perirhinal cortex role in the consolidation of conditioned freezing to context and to acoustic conditioned stimulus in the rat. The Journal of Neuroscience 19 (21), 9570–9578.

Samson, R.D., Paré, D., 2005. Activity-dependent synaptic plasticity in the central nucleus of the amygdala. The Journal of Neuroscience 25 (7), 1847–1855.

Schafe, G.E., LeDoux, J.E., 2000. Memory consolidation of auditory pavlovian fear conditioning requires protein synthesis and protein kinase A in the amygdala. The Journal of Neuroscience 20 (18), 96RC.

Shin, L.M., Liberzon, I., 2009. The neurocircuitry of fear, stress, and anxiety disorders. Neuropsychopharmocology 35 (1), 169–191.

Shin, L., Whalen, P., Pitman, R., Bush, G., Macklin, M., Lasko, N., Orr, S., Mcinerney, S., Rauch, S., 2001. An fMRI study of anterior cingulate function in posttraumatic stress disorder. Biological Psychiatry 50 (12), 932–942.

Shin, L.M., Wright, C.I., Cannistraro, P.A., Wedig, M.M., McMullin, K., Martis, B., Macklin, M.L., Lasko, N.B., Cavanagh, S.R., Krangel, T.S., Orr, S.P., Pitman, R.K., Whalen, P.J., Rauch, S.L., 2005. A functional magnetic resonance imaging study of amygdala and medial prefrontal cortex responses to overtly presented fearful faces in posttraumatic stress disorder. Archives of General Psychiatry 62 (3), 273.

Silveira, M.C.L., Sandner, G., Graeff, F.G., 1993. Induction of Fos immunoreactivity in the brain by exposure to the elevated plus-maze. Behavioural Brain Research 56 (1), 115–118.

Skre, I., Onstad, S., Torgersen, S., Lygren, S., Kringlen, E., 2000. The heritability of common phobic fear: a twin study of a clinical sample. Journal of Anxiety Disorders 14 (6), 549–562.

Tsien, J.Z., 2000. Building a brainier mouse. Scientific American-American Edition 282 (4), 62–68.

Tully, K., Bolshakov, V.Y., 2010. Emotional enhancement of memory: how norepinephrine enables synaptic plasticity. Molecular Brain 3 (1), 15.

Vidal-Gonzalez, I., Vidal-Gonzalez, B., Rauch, S.L., Quirk, G.J., 2006. Microstimulation reveals opposing influences of prelimbic and infralimbic cortex on the expression of conditioned fear. Learning & Memory 13 (6), 728–733.

Weike, A.I., Hamm, A.O., Schupp, H.T., Runge, U., Schroeder, H.W., Kessler, C., 2005. Fear conditioning following unilateral temporal lobectomy: dissociation of conditioned startle potentiation and autonomic learning. The Journal of Neuroscience 25 (48), 11117–11124.

Weisskopf, M.G., Bauer, E.P., LeDoux, J.E., 1999. L-type voltage-gated calcium channels mediate NMDA-independent associative long-term potentiation at thalamic input synapses to the amygdala. The journal of Neuroscience 19 (23), 10512–10519.

Wilensky, A.E., Schafe, G.E., LeDoux, J.E., 1999. Functional inactivation of the amygdala before but not after auditory fear conditioning prevents memory formation. The Journal of Neuroscience 19 (24), 48RC.

Zimmerman, J.M., Maren, S., 2010. NMDA receptor antagonism in the basolateral but not central amygdala blocks the extinction of Pavlovian fear conditioning in rats. European Journal of Neuroscience 31 (9), 1664–1670.

Aphasia

David Caplan, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Aphasia is a disturbance of the use of language due to disease of the brain. Early approaches to describing aphasia emphasized patients' abilities to speak, understand spoken language, and read and write, and paid little attention to the elements of language that were affected in these tasks or how they were affected. These theories have provided a basis for classifying aphasias and for relating some aspects of language to the brain that is still in clinical use. Modern approaches to aphasia try to delve deeper into the nature of aphasic impairments, by saying what parts of language are affected (e.g., problems with comprehension of sentences could be due to problems affecting word recognition, not knowing what words mean, or not being able to relate words to one another in a sentence), and how they are affected. This article presents an overview of aphasia, beginning with the traditional clinical descriptions and ending with more modern approaches.

Aphasias are disturbances of language following diseases of the brain. Clinical approaches to describing aphasias evolved from nineteenth-century descriptions that emphasized the language task that was affected – speaking, understanding spoken language, reading, and/or writing – and a limited description of the linguistic impairments in patients, such as whether a disturbance of speech affected grammatical forms or content words. This approach was related to a model of where the processing of speaking, understanding spoken language, reading, and writing simple words was located in the brain, and has given rise to a set of aphasic syndromes that are widely recognized by clinicians today. Modern descriptions of aphasic patients have taken major steps toward capturing the details of the linguistic impairments in patients. This research has led to better understanding of how language processing is organized and how it breaks down. However, the goal of understanding the relation between language processing operations and brain structures remains to be achieved.

Definition

The term 'aphasia' refers to disorders of language following diseases of the brain. As is discussed in other articles in this encyclopedia, language can be conceived of as a distinctly human symbol system that relates a number of different types of forms (words, words formed from other words, sentences, discourse, etc.) to various aspects of meaning (objects, properties of objects, actions, events, causes of events, temporal order of events, etc.). The forms of language and their associated meanings are activated in the processes of speaking, understanding speech, reading, and writing. The processes whereby these forms are activated are largely unconscious, obligatory once initiated, fast, and usually quite accurate. Disturbances of the forms of the language code and their connections to their associated meanings, and of the processes that activate these representations in these ordinary tasks of language use, constitute aphasic disturbances. By convention, the term 'aphasia' does not refer to disturbances that affect the functions to which language processing is put. Lying (even transparent, ineffectual lying) is not considered a form of aphasia, nor is the garrulousness of old age or the incoherence of schizophrenia.

Language consists of a complicated system of representations, and its processing is equally complicated, as described in other articles in this encyclopedia. Just the representation of the minimal linguistically relevant elements of sound – phonemes – and the processing involved in recognizing and producing these units constitute a highly complex domain of functioning. When all the levels of language and their interactions are considered, language processing is seen to be enormously complex. Aphasic disturbances would therefore be expected to be equally complex. Researchers are slowly describing the very considerable range of these disorders.

History of the Field: The Classic Aphasic Syndromes, and Alternative Views

The first modern scientific descriptions of aphasia were quite modest with respect to the descriptions of language processing that they contained. These descriptions were made by neurologists in the second half of the nineteenth century. Though modest with respect to the sophistication of the descriptions of language, these studies laid important foundations for the scope of work on aphasia and for the neural basis for language processing, which has always been a closely associated topic. The first of these late nineteenth-century descriptions was that by Broca (1861), who described a patient, Leborgne, with a severe speech output disturbance. Leborgne's speech was limited to the monosyllable 'tan.' Broca described Leborgne's ability to understand spoken language and to express himself through gestures and facial expressions, as well as his understanding of nonverbal communication, as being normal. Broca claimed that Leborgne had lost 'the faculty of articulate speech.' Leborgne's brain contained a lesion whose center was in the posterior portion of the inferior frontal convolution of the left hemisphere, an area of advanced cortex just adjacent to the motor cortex. Broca related the most severe part of the lesion to the expressive language impairment. This area became known as 'Broca's area.' Broca argued that it was the neural site of the mechanism involved in speech production.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54001-X

In a second very influential paper, Wernicke (1874) described a patient with a speech disturbance that was very different from that seen in Leborgne. Wernicke's patient was fluent; however, her speech contained words with sound errors, other errors of word forms, and words that were semantically inappropriate. Also unlike Leborgne, Wernicke's patient did not understand spoken language. Wernicke related the two impairments – the one of speech production and the one of comprehension – by arguing that the patient had sustained damage to 'the storehouse of auditory word forms.' The lesion in Wernicke's case was unknown, but a lesion in a similar case was the area of the brain next to the primary auditory receptive area, which came to be known as Wernicke's area.

These pioneering descriptions of aphasic patients set the tone for much subsequent work. First, they focused the field on impairments of the usual uses of language – speech, later on reading and writing. This seems like an obvious area for aphasiology to be concerned with, but not all researchers of the period agreed with this focus. In another famous paper, the influential British neurologist John Hughlings Jackson (1878) described a patient, a carpenter, who was mute but mustered up the capacity to say 'Master's' in response to his son's question about where his tools were. Jackson's poignant comments convey his emphasis on the conditions that provoke speech, rather than on the form of the speech itself:

> The father had left work; would never return to it; was away from home; his son was on a visit, and the question was directly put to the patient. Anyone who saw the abject poverty the poor man's family lived in would admit that these tools were of immense value to them. Hence we have to consider as regards this and other occasional utterances the strength of the accompanying emotional state.
>
> Jackson, 1878, p. 181

Jackson sought a description of language use as a function of motivational and intellectual states, and tried to describe aphasic disturbances of language in relationship to the factors that drive language production and make for depth of comprehension. Broca, Wernicke, and the researchers who followed focused aphasiology on patients' everyday language use under what were thought to be normal emotional and motivational circumstances.

These and related subsequent papers tended to describe language impairments in terms of the entirety of language-related tasks – speaking, comprehending, etc. – with only passing regard for the details of the language forms that were impaired within a task. A patient's deficit was typically described in terms of whether such an entire function was normal or not, and whether one such function was more impaired than another was. Here, for instance, is the description of Broca's aphasia by two twentieth-century neurologists whose work follows in this tradition:

> The language output of Broca's aphasia can be described as non-fluent … Comprehension of spoken language is much better than speech but varies, being completely normal in some cases and moderately disturbed in others.
>
> Benson and Geschwind, 1971, p. 7

The one level of language that descriptions did tend to concentrate on was the level of words. For instance, in the same passage, Benson and Geschwind point out that patients with Broca's aphasia often omit function words (e.g., determiners such as 'the') and word endings, but produce at least some nouns. Altogether, the clinical approach to aphasia, derived most directly from early work, emphasized the usual tasks of language use and just began to describe impairments in linguistic and psycholinguistic terms. This approach to aphasia led to the recognition of 10 classical aphasic 'syndromes.' These are listed, along with their proposed neural bases, in **Table 1**.

Psycholinguistic Approaches to Aphasia

As noted in the previous section, the classical aphasic syndromes do not give a complete account of the range and specificity of aphasic impairments. More recent descriptions of aphasia add many details to the linguistic and psycholinguistic descriptions of these disorders. It is impossible to review all these impairments in the space of a short article, but a few examples will illustrate these results. For instance, at the first step of speech processing – converting the sound waveform into linguistically relevant units of sound – researchers have described specific disturbances affecting the ability to recognize subsets of phonemes such as vowels, consonants, stop consonants, fricatives, nasals, etc. (Saffran et al., 1976). In the area of word production, patients have been described with selective impairments of the ability to produce the words for items in particular semantic categories, such as fruits and vegetables, but sparing animals and man-made tools (Hart et al., 1985); selective impairments affecting the ability to produce nouns and verbs (Damasio and Tranel, 1993); and other highly restricted deficits. In the area of reading, patients have been found to have impairments of the ability to sound out novel written stimuli using letter–sound correspondences but retained abilities to read familiar words, and vice versa (Marshall et al., 1980; Patterson et al., 1985).

Linguistic theory provides a basis for exploring the nature of aphasic disorders by providing evidence for different types of linguistic representations. Models of the psychological processes involved in activating linguistic representations suggest other possible loci of impairment. Many researchers have worked backward from clinically observed phenomena, developing or modifying theories of language structure and processing on the basis of the disorders seen in aphasic patients. For instance, the finding of patients with disorders of the ability to sound out novel written stimuli but retained abilities to read familiar words and vice versa has led to a model of reading in which there are two ways of pronouncing a written word – sounding it out and recognizing it as a whole. This model is controversial, and some of the strongest evidence for it remains the performance of aphasic patients (Coltheart et al., 2001). Characterizing aphasic disorders is an interactive, interdisciplinary, bootstrapping process that is presently in active evolution.

The psycholinguistic approach to aphasia is based upon a model of language structure and processing. Experts disagree about these models. Some of the greatest disagreements center on the issue of the extent to which linguistic representations are

Table 1 Classical aphasic syndromes

Syndrome	*Clinical manifestations*	*Postulated deficit*	*Classical lesion location*
Broca's aphasia	Major disturbance in speech production with sparse, halting speech, often misarticulated, frequently missing function words, and bound morphemes	Disturbances in the speech planning and production mechanisms	Posterior aspects of the third frontal convolution (Broca's area)
Wernicke's aphasia	Major disturbance in auditory comprehension; fluent speech with disturbances of the sounds and structures of words (phonemic, morphological, and semantic paraphasias)	Disturbances of the permanent representations of the sound structures of words	Posterior half of the first temporal gyrus and possibly adjacent cortex (Wernicke's area)
Pure motor speech disorder	Disturbance of articulation, apraxia of speech, dysarthria, anarthria, aphemia	Disturbance of articulatory mechanisms	Outflow tracts from motor cortex
Pure word deafness	Disturbance of spoken word comprehension	Failure to access spoken words	Input tracts from auditory system to Wernicke's area
Transcortical motor aphasia	Disturbance of spontaneous speech similar to Broca's aphasia with relatively preserved repetition	Disconnection between conceptual representations of words and sentences and the motor speech production system	White matter tracts deep to Broca's area connecting it to parietal lobe
Transcortical sensory aphasia	Disturbance in single word comprehension with relatively intact repetition	Disturbance in activation of word meanings despite normal recognition of auditorily presented words	White matter tracts connecting parietal lobe to temporal lobe or portions of inferior parietal lobe
Conduction aphasia	Disturbance of repetition and spontaneous speech (phonemic paraphasias)	Disconnection between the sound patterns of words and the speech production mechanism	Lesion in the arcuate fasciculus and/or corticocortical connections between Wernicke's and Broca's areas
Anomic aphasia	Disturbance in the production of single words, most marked for common nouns with variable comprehension problems	Disturbances of concepts and/or the sound patterns of words	Inferior parietal lobe or connections between parietal lobe and temporal lobe; can follow many lesions
Global aphasia	Major disturbance in all language functions	Disruption of all language processing components	Large portion of the perisylvian association cortex
Isolation of the language zone	Disturbance of both spontaneous speech (similar to Broca's aphasia) and comprehension, with some preservation of repetition	Disconnection between concepts and both representations of word sounds and the speech production mechanism	Cortex just outside the perisylvian association cortex

highly abstract structures that are produced and computed in comprehension tasks by rules (Chomsky, 1995), as opposed to far less abstract representations that are processed largely by highly developed pattern associations (Rumelhart and McClelland, 1986). If language is seen in the former perspective, many aphasic impairments are considered to be the result of damage to specific representations and/or processing operations. If language is seen in the latter perspective, aphasic disturbances are largely conceptualized as resulting from reductions in the power of the associative system, due to loss of units, increases in noise, etc. Empirical study suggests that both specific impairments and loss of processing power are sources of aphasic disturbances. This can be illustrated in one area – disorders affecting syntactic processing in sentence comprehension.

Disorders of syntactically based comprehension affect the ability to extract the relationships between the meanings of words in a sentence that are determined by the syntactic structure of a sentence. For instance, in the sentence 'The dog that scratched the cat killed the mouse,' there is a sequence of words – the cat killed the mouse – which, in isolation, would mean that the cat killed the mouse. However, this is not what the sentence means, because of its syntactic structure. 'The cat' is the object of the verb 'scratched'; 'the dog' is the subject of the verb 'killed' and is the agent of that verb. Caplan and his colleagues have explored the nature of these disturbances (Caplan et al., 1985, 1996). They found that, in many hundred of aphasic patients, mean group performance deteriorated on sentences that were more syntactically complex and that more impaired groups of patients increasingly performed more poorly on sentences that were harder for the group overall. These patterns suggest that the availability of a processing resource that is used in syntactic comprehension is reduced to varying degrees in different patients. A second finding in their studies has been that individual patients can have selective impairments of syntactic comprehension, just as is the case in the other areas of language processing previously mentioned. Published cases have had difficulty constructing hierarchical syntactic structures, disturbances affecting reflexives or pronouns but not both, and other more subtle impairments of syntactic processing (Caplan and Hildebrandt, 1988).

Overall, these studies suggest that a patient's aphasic impairment can be described in terms of a reduction in the processing resources needed for this function and disruption to specific operations. An unresolved question is whether the entire pattern of performance seen in these disorders can be attributed to just one of these types of impairments, as the two types of models previously outlined maintain. This may be possible, but the challenges in explaining all these aspects of these (and other) aphasic disorders within a model that

either does not incorporate the idea of a processing resource limitation or does not recognize specific operations are considerable.

Functional Consequences of Aphasic Impairments

The focus of this article has thus far been on aphasic disturbances as impairments of the largely unconscious processes that activate the elements of language in the usual tasks of language use. The functional consequences of these disorders deserve a brief comment. Functional communication involving the language code occurs when people use language to accomplish specific goals – to inform others, to ask for information, to get things done, etc. There is no simple, one-to-one relationship between impairments of elements of the language code or of psycholinguistic processors, on the one hand, and between abnormalities in performing language-related tasks and accomplishing the goals of language use, on the other. Patients adapt to their language impairments in many ways, and some of these adaptations are remarkably effective at maintaining at least some aspects of functional communication. Conversely, patients with intact language processing mechanisms may fail to communicate effectively. Nevertheless, most patients who have disturbances of elements of the language code or psycholinguistic processors experience limitations in their functional communicative abilities. In general, as the intentions and motivations of the language user become more complex, functional communication is more and more affected by disturbances of the language code and its processors. Thus, though 'high-level' language-impaired patients may be able to function well in many settings, their language impairments can cause functional limitations that affect family and professional life.

This article should not end on this negative note. Rather, it is important to appreciate that many aphasic patients make excellent recoveries, for a variety of reasons. The natural history of many aphasic impairments is for considerable improvement, especially those due to smaller or subcortical lesions. Though still in their infancy, modern approaches to rehabilitation for aphasia are developing a sounder scientific basis. Technological advances allow for more professionally guided home training using computers, improved augmentative communication devices, and other useful support mechanisms. Support groups for patients and their families and friends are increasing in number; these help patients adjust to the changes in their lives and remain socially active. Though aphasia deprives a person of an important function to a greater or lesser degree, reactions to aphasia are as important as the aphasia itself in determining functional outcome and many aphasic patients function in vital ways after their loss. Lecours et al. (1983) cite a patient described by the Soviet psychologist A. R. Luria, who continued to compose music after a stroke that left him very aphasic; some critics thought his work improved after his illness. Time, rehabilitation, support, and a positive attitude can allow many aphasic patients to be productive and happy.

See also: Speech Production, Neural Basis of; Speech Production, Psychology of.

Bibliography

Benson, D.F., Geschwind, N., 1971. Aphasia and related cortical disturbances. In: Baker, A.B., Baker, L.H. (Eds.), Clinical Neurology. Harper and Row, New York.

Broca, P., 1861. Remarques sur le siège de la facultè de la parole articulèe, suives d'une observation d'aphemie (perte de parole). Bulletin de la Société d'Anatomie 36, 330–357.

Caplan, D., Hildebrandt, N., 1988. Disorders of Syntactic Comprehension. MIT Press (Bradford Books), Cambridge, MA.

Caplan, D., Baker, C., Dehaut, F., 1985. Syntactic determinants of sentence comprehension in aphasia. Cognition 21, 117–175.

Caplan, D., Hildebrandt, N., Makris, N., 1996. Location of lesions in stroke patients with deficits in syntactic processing in sentence comprehension. Brain 119, 933–949.

Chomsky, N., 1995. The Minimalist Program. MIT Press, Cambridge, MA.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., Ziegler, J., 2001. The DRC model: a model of visual word recognition and reading aloud. Psychological Review 108, 204–258.

Damasio, A., Tranel, D., 1993. Nouns and verbs are retrieved with differently distributed neural systems. Proceedings of the National Academy of Sciences 90, 4957–4960.

Hart, J., Berndt, R.S., Caramazza, A., 1985. Category-specific naming deficit following cerebral infarction. Nature 316, 439–440.

Jackson, J.H., 1878. On affections of speech from disease of the brain. In: Taylor, J. (Ed.), Selected Writings of John Hughlings Jackson 1958. Basic Books, New York.

Lecours, A.R., Lhermitte, F., Bryans, B., 1983. Aphasiology (Chapter 19). Bailliere Tindal, Paris.

Marshall, J.C., Patterson, K., Coltheart, M., 1980. Deep Dyslexia. Routledge, London.

Patterson, K., Coltheart, M., Marshall, J.C., 1985. Surface Dyslexia. Lawrence Elrbaum, London.

Rumelhart, D., McClelland, J.C., 1986. Parallel Distributed Processing. MIT Press, Cambridge, MA.

Saffran, E.M., Marin, O., Yeni-Komshian, G., 1976. An analysis of speech perception and word deafness. Brain and Language 3, 209–228.

Wernicke, C., 1874. Der Aphasische Symptomenkomplex. Cohn and Weigart, Breslau, Germany. Reprinted in translation in, Boston Studies in Philosophy of Science 4, 34–97.

Appeals: Legal

Paul D Carrington and Marin K Levy, Duke University School of Law, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P. Carrington, volume 1, pp. 585–589, © 2001, Elsevier Ltd.

Abstract

An appeal is a proceeding in a court of law initiated by a party contending that a judgment of a subordinate court or agency is erroneous. Appeals are not universal in all legal systems, although they tend to be characteristic of many court systems, including the American court system. Contemporary American practice took its present form in the federal courts and in the courts of the states over the course of the nineteenth century. The most important general traits of the American appellate system are six principles of restraint, expressing deference to the judgment of the lower court or agency.

An *appeal* is a proceeding in a court of law initiated by a party contending that a judgment of a subordinate court or agency is erroneous. The appeal as it is known in western legal systems appears to have originated in the early Byzantine Empire. Today, although not universal in all legal systems, appeals exist in many court systems around the world, including the American court system. Contemporary American practice took its present form in the federal courts and in the courts of the states over the course of the nineteenth century – though these practices have shifted as the judiciary has faced a significant increase in workload over the past few decades. Still, the most important general traits of the American appellate system remain. These include six principles of restraint, enumerated below, expressing deference to the judgment of the lower court or agency.

Appeal is to be distinguished from other proceedings initiated in courts for purposes other than the correction of error by a subordinate court or agency. It is also to be distinguished from review proceedings conducted without regard for any previous disposition by the subordinate court or agency, as where a case is subjected to *trial de novo* in the higher court. The latter form of proceeding is, for example, standard practice at the first level of review in Germany (Meador et al., 2006; pp. 907–928). (It is not unknown in the United States; its use is most common when a higher court is reviewing the judgment of a small claims court, in which professional lawyers seldom represent the parties and the judge may be a nonprofessional.) Appeal is also to be distinguished from the *discretionary review* that may be observed in the Supreme Court of the United States and most other highest courts of states or of large nations. A court performing discretionary review may select the issues or rulings that it chooses to review; its primary purpose in making its selection is to explain the correct resolution of issues that are likely to recur in the lower courts or are otherwise important to persons other than the parties. In the United States and in many other countries, most appeals are taken to intermediate courts that are subject to discretionary review by courts of last resort. This article is an account only of the principles governing such appeals, primarily in the intermediate courts of the United States. Some variations in other systems will be noted, but those variations are so numerous that they defy synthesis.

Not Universal

The appeal is not a universal feature of legal systems. Some tribal courts, for example, bring to bear in the first instance most of the wisdom and authority of the community (see, e.g., Gluckman, 1955). There is then no body of higher persons to whom an appeal might appropriately be addressed. Also, there are autocratic systems, generally those having strong religious roots, in which an individual chief, priest, or judge is accorded the power of decision without possibility of review. Islamic law is noted as an example; the *kadi* administers an elaborate code of conduct contained in holy writ that no mere lawyer can presume to interpret, and the *kadi's* decision is not subject to review (Shapiro, 1980).

Although not universal, the appeal is a central feature of many court systems throughout the world today. To name just a few, countries including Canada, the United Kingdom, France, Italy, Germany, Israel, Japan, Brazil, South Africa, and the United States all have legal systems that include at least one (and often more than one) appellate court. The nature of appeals in these different countries varies considerably, depending upon whether appeals may be taken by right or by permission of a court, and whether the appellate courts consider general or specialized subject matters. There are, however, key similarities; typically in those systems using appeals, the review is conducted by a larger bench of judges than the court whose judgment is under review. In the United States, there is with rare exception a single judge presiding over a trial court from whose decisions an appeal may be taken to a three-judge court. Judges sitting on an appellate court are generally designated as persons of higher rank, and are likely to receive marginally higher salaries than judges sitting on the trial courts they review.

On the matter of whether appeals exist as a matter of right, even in the United States, the right of appeal in a civil case is generally a matter of legislative grace. Additionally, until 1888, there was no appeal from a criminal conviction in a federal court, even if the convicted person was subject to capital punishment (Frankfurter and Landis, 1928). However, the constitutions of some states have long guaranteed the right of appeal in criminal cases. Furthermore, it is still the law in all American jurisdictions that a state may not appeal a judgment of acquittal in a criminal case because a second trial would place the defendant in double jeopardy of conviction.

However, some states, such as New York, permit the state to appeal the legal basis for the acquittal in a proceeding that does not affect the present defendant but will serve as precedent for future cases.

Before the right to appeal became widespread, trial judges were seen to accumulate excessive discretionary power over the individuals engaged in disputes brought before them and on occasion to engage in seemingly lawless behavior. The Congress of the United States and state legislatures in the past century were attentive to the importance of constraining that discretionary judicial power and provided for appellate review in civil cases and in criminal cases not resulting in judgments of acquittal.

Principles of Restraint

The development of the appeal in American courts has resulted in the formulation of principles of restraint that are also in use in various forms in other legal systems. The institution described here is analogous to the *writ of error* familiar to ancient practice of English common law courts as an instrument for bringing a local one-judge decision to the attention of the larger court sitting in Westminster, but it departs from that usage in important respects. Contemporary American practice took its present form in the federal courts and in the courts of the states over the course of the nineteenth century. Six principles of restraint emerged.

Reversible Error: The Adversary Tradition

The first of these is the concept of *error*, a principle rooted in the Anglo-American adversary tradition placing primary responsibility for the conduct of litigation on the parties and their counsel. Because the burden is on the parties to present the evidence and inform the court of their claims or defenses, it is not generally error for a court to fail to identify a fact not proven or to fail to enforce a legal principle not invoked. The political function of this principle is to reduce the role of the court: disappointed parties often share responsibility for their own defeats. The administrative purpose of the principle is to encourage thorough preparation and presentation of the case by counsel and to protect the court from being trapped by deceitful or negligent counsel. It follows from this principle of reversible error that an appellant, to be successful on appeal, must generally point to a ruling made by the lower court to which the appellant made timely *objection* (Tigar and Tigar, 2011). In times past, it was often required not only that there be an objection to the erroneous ruling, but also that an appellant have taken *exception* to the adverse ruling, thus putting the trial judge on notice that the ruling may be challenged on appeal. The requirement of an exception has been eliminated from the practice of federal and most state courts.

Even in American courts, the requirement of a timely objection is sometimes disregarded for erroneous rulings so egregious that the judgment under review deeply offended the appellate court's sense of propriety. Such errors are denoted as *plain errors*. Rarely is the plain error doctrine invoked because it allows counsel to proceed in a no-lose situation, knowing that if the error is not corrected by the trial judge there will be a successful appeal affording a nonobjecting party with a fresh start. On the other hand, appellate courts may be reluctant to punish a party having a clearly meritorious claim or defense for no reason other than a lapse on the part of counsel.

Harmless Error

A second and more universal requirement is that no error, however egregious, is an occasion for reversal unless it was consequential. The familiar expression is that a *harmless error* is not grounds for reversal. Thus, an appellant who should and would have lost on other grounds cannot secure relief from an adverse judgment, even if that party can identify blatant errors of fact or of law that were committed by the court below and to which timely objection was made. This requirement conserves judicial resources; cases will only return to the trial court if there is some chance that they will result in a different outcome.

Standing to Appeal

A third requirement is *standing* to appeal. Generally, only a party who has participated in the proceeding is bound by the decision below. Hence, nonparties are not directly harmed by a judgment, even though they may strongly disapprove of the outcome and assert that a grave injustice has been done. The requirement of standing may be extended to bar appeal by nonparties who are only remotely affected by a judgment. Thus, investors or employees lack standing to appeal a judgment rendered against a corporation in which they own shares or are employed. The decision to appeal a judgment rendered against a corporation usually resides exclusively with those who may speak for the corporation itself.

Ripeness for Review

A fourth common requirement is that of ripeness for review. In general, an appeal is premature until a *final decision* has been reached in the court below. This is another principle of economy. An error of the lower court may turn out to be harmless; the higher court cannot know until a final decision has been reached. Moreover, there would be serious inefficiency if every litigant adversely affected by a provisional ruling of a trial court were allowed to take an immediate appeal. Not only would this afford parties a means of imposing needless financial costs on adversaries, but it would also force the trial court, or perhaps even the appellate court, to decide in each instance whether the lower court proceedings should be stayed pending the outcome of the appeal. If a stay is granted, the appellant has been empowered to delay the proceeding; and the delay itself may often result in injustice. If a stay is denied, there is the risk that further proceedings will be set aside as a result of the interlocutory decision of the higher court.

Most jurisdictions have numerous exceptions to the ripeness requirement. For example, in the federal practice, an *interlocutory appeal* may be taken from the grant or denial of a preliminary injunction for the reason that an error in such a ruling can have grave consequences for the party adversely affected (Steinman, 1998). Immediate appeal may also be allowed with respect to rulings involving substantial procedural rights that are very likely to affect the outcome of

a case. With respect to important issues of judicial administration, it may be said that an otherwise unripe appeal should be entertained in order to sustain the supervisory power of the appellate court, that is, to prevent lawless behavior by a trial judge. Appellate courts in the United States also have the power to issue several *extraordinary writs*, such as a writ of *mandamus* (a tool of Roman origins), to forestall an abuse of discretion by a subordinate court. In addition, a lower court may be permitted to circumvent the requirement of ripeness on its own motion by certifying to a higher court a question of law to which the lower court has no answer and that is central to a lengthy trial.

The Record on Appeal

A fifth concept shared by all American and most other jurisdictions is that of the *record on appeal*. The record is an official account of the proceedings below, or at least that part of the official account that the lawyers deem pertinent to the issues raised on appeal and therefore worth the cost of reproduction. The record is generally produced by the clerk of the court below and will contain documents filed with the court and a transcript of any oral proceedings prepared by a professional court reporter. In a civil case, the appellant must generally advance the cost of preparing the record, but a state may be required to bear the cost in a criminal case if the convicted person is indigent.

In general, an appellate court will not consider information that is not contained in the record (Marvell, 1978). There is an exception to this principle generally known as the *judicial notice* doctrine. A court may without proof take notice of common knowledge not contained in the record, or knowledge that is readily available to all, such as the day of the week of a particular date. Further, appellate courts frequently accept submissions from third parties known as *amici curiae*, whose representations of general facts, if uncontroverted, are sometimes relied upon despite being outside the record of the trial court (Gorod, 2011).

Deferential Review of Factual Determinations

Finally, there is a principle of deference to trial courts and agencies. Among the aims of this principle are to dignify the proceedings below and to discourage appeals challenging the guesswork inevitably done in the trial court to resolve issues on which there is conflicting evidence. The principle is expressed in an elusive distinction between *law* and *fact*. It is generally agreed that trial courts are entitled to no deference in their rulings on questions of law. Such rulings may be embodied in the *instructions* on the law given to a jury, or if there is no jury demanded, in the *conclusions of law* stated by the court to explain its disposition. An error in either statement may be a sufficient ground for reversal.

On the other hand, trial courts having dealt directly with the evidence submitted by the adversaries are extended the benefit of some doubt with respect to the determinations of fact. The appellate court hears no witnesses and sees only a transcript of the testimony. If the trial court judgment rests on a *jury verdict*, the factual determination in a civil case can be reversed on appeal only if the appellate court finds that there is no *substantial evidence* in the record to support it. In the absence of a jury, a trial judge's decision on evidence will in a civil case be expressed in the judge's *findings of fact*; such findings may be reversed in American courts only if the reviewing court finds *clear error*. It is generally assumed that the latter is a lower standard and that fact finding by a judge is more closely scrutinized on appeal than is fact finding by a jury.

The distinction between fact and law is subtle and even sometimes circular; a much-simplified summary is that issues of fact are those involving specific past events about which there is doubt, whose resolution has little or no bearing on future cases. Issues of law are typically those involving an interpretation of legal texts; but it is often said that whether there is evidence sufficient to support a jury verdict is itself a question of law. All that this means is that sufficiency of the evidence is a question in the first instance for the trial judge but that the ruling on the issue will be reviewed without deference. It is therefore not wrong to say that an issue of fact is simply one that the courts leave for decision by a trier of fact, while an issue of law is any issue that the appellate court chooses to decide on its own.

This distinction is not made in some other legal systems. For example, it is unknown to Japanese practice, which is modeled on the French. It is generally less useful in systems placing heavy reliance on written submissions of evidence. In such systems, the appellate court has access to the same information as the trial court, and hence there is less reason for deference to the judicial officer who saw and heard the adversary presentation of the parties.

Abuse of Discretion

Notwithstanding these settled principles of restraint, an appellate court is also empowered to correct discretionary actions of a trial judge that it deems to be an *abuse of discretion*. This principle is most frequently invoked to challenge and correct procedural rulings that are seen to be idiosyncratic or manifestly unjust (Friendly, 1982).

Appellate Procedure and Practice

In addition to these principles of appellate jurisdiction, American courts share a traditional procedure that is replicated in many legal systems. Until about 1960, this process was universal in the United States, and included (1) the submission of written *briefs* prepared by counsel on both sides that present their legal arguments and provide citations to pertinent legal texts and authorities; (2) an *oral argument*, at which counsel might engage the appellate judges in dialogue and answer questions they might pose; (3) a conference of the judges responsible for the appellate decision; and (4) a published *opinion of the court* explaining the legal principles underlying the disposition on appeal. Many of these procedural amenities have since been foreshortened in most American appellate courts as their caseloads have grown (Levy, 2011). Specifically, the majority of appeals in the federal courts of appeals do not receive oral argument and do not result in published opinions. It has also become common, especially in federal appellate courts, for

much of the responsibility to be delegated to law clerks and central staff attorneys. As the caseload has grown, it has been argued that the right of appeal should be abolished in federal practice, in recognition of the reality that many appeals are never seriously studied by the judges commissioned as members of the court (e.g., Parker and Chapman, 1997). Others have resisted this trend (Arnold, 1995). It has even been contended that the erosion of appellate procedure has diminished the *raison d'être* of the federal appellate courts (Carrington, 2000). Still, even in their constricted form, the American appellate courts continue to develop the law and correct errors in tens of thousands of cases each year.

See also: Courts and Adjudication; Judicial Review; Supreme Courts.

Bibliography

Arnold, R.S., 1995. The future of the federal courts. Missouri Law Review 60, 540.
Carrington, P.D. (Ed.), 1984. Civil aappellate jurisdiction, part 2. Law and Contemporary Problems 47 (3).
Carrington, P.D., 1998. Virtual civil litigation: a visit to John Bunyan's celestial city. Columbia Law Review 98, 501.
Carrington, P.D., 2000. The obsolescence of the United States courts of appeals. Journal of Law & Policy 20, 266.
Carrington, P.D., Orchard, P., 2010. The federal circuit: a model for reform? George Washington Law Review 78, 575.
Damaska, M.R., 1986. The Faces of Justice and Authority. Yale University Press, New Haven, CT.
Edwards, H.T., Elliott, L.A., 2007. Federal Courts and Standards of Review: Appellate Court Review of District Court Decisions and Agency Actions. Thomson West.
Frankfurter, F., Landis, J., 1928. The Business of the Supreme Court: A Study in the Federal Judicial System. Macmillan.
Friendly, H.J., 1982. Indiscretion about discretion. Emory Law Journal 31, 747.
Gluckman, M., 1955. The Judicial Process Among the Barotse of Northern Rhodesia. Manchester University Press.
Gorod, B.J., 2011. The adversarial myth: appellate court extra-record factfinding. Duke Law Journal 61, 1.
Jolowicz, J.A., 2000. On Civil Procedure. Cambridge University Press, Cambridge, MA.
Levy, M.K., 2011. The mechanics of federal appeals: uniformity and case management in the circuit courts. Duke Law Journal 61, 315.
Marvell, T., 1978. Appellate Courts and Lawyers: Information Gathering in the Adversary System. Greenwood.
Meador, D.J., 2006. Appellate Courts in the United States, second ed. West.
Meador, D.J., Baker, T.E., Steinman, J.E., 2006. Appellate Courts: Structures, Functions, Processes and Personnel. LexisNexis.
Parker, R.M., Chapman, R., 1997. Accepting reality: the time for accepting discretionary review in the courts of appeals has arrived. SMU Law Review 50, 573.
Phillips, J.D., 1984. The appellate review function: scope of review. Law and Contemporary Problems 47, 1.
Pound, R., 1941. Appellate Procedure in Civil Cases. Little Brown.
Shapiro, M.L., 1980. Islam and appeal. California Law Review 68, 350.
Steinman, J., 1998. The scope of appellate jurisdiction. Hastings Law Journal 49, 1337.
Tigar, M.E., Tigar, J.B., 2011. Federal Appeals, Jurisdiction and Practice, third ed. West.

Appeasement: Political

Stephen R Rock, Vassar College, Poughkeepsie, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D.W. Larson, volume 1, pp. 589–592, © 2001, Elsevier Ltd.

Abstract

Appeasement is a policy by which a state hopes to avoid war (or achieve other goals) by making concessions to an adversary, thereby eliminating its grievances. For many years after the failed attempts by Britain and France to conciliate Nazi Germany during the 1930s, appeasement was regarded as being futile, dangerous, and immoral. The so-called 'lessons of Munich' exercised a profound influence on foreign policy, especially in the United States, for more than a generation. More recently, however, scholars and policymakers have come to recognize that appeasement can, in certain circumstances, be an effective and appropriate strategy.

Defining Appeasement

In classical European diplomacy, appeasement meant the "reduction of tension between [two states] by the methodical removal of the principal causes of conflict and disagreement between them" (Craig and George, 1990: 250). The strategy involved making concessions to one's adversary, and some definitions thus emphasize this aspect of the policy (Ripsman and Levy, 2008). Terms frequently associated with appeasement or used as synonyms for it include 'inducements,' 'positive sanctions,' 'conciliation,' and 'accommodation.'

Prior to World War II, the word 'appeasement' had a neutral or even positive connotation. Historians Paul Schroeder (1976) and Paul Kennedy (1983) have identified a tradition of appeasement in British foreign policy dating back to the 1860s, noting that the strategy was employed as a means of settling disputes between Britain and rising powers seeking a greater share of the world's territory and resources. For British statesmen, appeasement was a pragmatic policy of realpolitik aimed at maintaining the balance of power and avoiding war over issues that did not threaten their country's most vital interests. It was also a just and honorable policy that admitted and sought to satisfy an adversary's legitimate grievances.

The meaning of appeasement changed dramatically with the experience of the 1930s. The disastrous attempts of the British government and especially Prime Minister Neville Chamberlain to accommodate the Nazi German government of Adolf Hitler caused the strategy to fall into disrepute. Shortly after World War II, international relations theorist Hans Morgenthau (1951: 137) termed appeasement "politically unwise," adding that appeasement occurs "when a nation surrenders one of its vital interests without obtaining anything worthwhile in return." Others came to regard the policy as being not merely unwise, but immoral.

Munich and the 'Lessons' of the 1930s

Beginning in the mid-1930s, Britain and France pursued an appeasement policy with respect to Nazi Germany, which each conceived as the principal long-term threat to its security. Initially, the strategy was largely passive. Thus, the two Western powers acquiesced to German violations of the Treaty of Versailles, including the remilitarization of the Rhineland in 1936. While this aspect of the policy never disappeared – it is reflected, for example, in the failure to act in the face of the 1938 Anschluss with Austria – the British government subsequently adopted a more active approach, trying to negotiate both economic and territorial arrangements with Germany that would allow for the preservation of peace. British policymakers were particularly hopeful that they might maintain the political and territorial status quo in Europe by redrawing the colonial map of Africa in a manner that was more favorable to German interests.

The famous conference between Neville Chamberlain and Adolf Hitler at Munich in September 1938 is widely regarded as the epitome of appeasement. By agreeing to cede the Sudetenland region of Czechoslovakia to Germany, Chamberlain hoped that he had satisfied Hitler's territorial ambitions and preserved peace. Unfortunately, he had not. In March 1939, Germany seized the remainder of Czechoslovakia, and in September it invaded Poland, despite promises by Britain and France to intervene.

The Munich Conference and subsequent events had a profound effect on both international relations theory and policymaking. After Munich, appeasement was judged to be both futile and dangerous. It was futile because aggressors could not be satisfied; indeed, by feeding them, one only increased their hunger for territory. As the French proverb says, 'L'appetit vient en mangeant' – 'appetite comes (or grows) with eating.' It was dangerous because it undermined subsequent attempts at deterrence. The Germans, having been handed the Sudetenland and allowed to take the rest of Czechoslovakia, did not believe French and British threats to defend Poland, even though Britain had issued a 'guarantee' to do so and had later joined Poland in a formal treaty of mutual defense.

Scholars, particularly in the United States, took what they believed to be the lessons of the 1930s and developed the deterrence model of international politics (Jervis, 1976). This model holds that, when facing an aggressor, credible threats of retaliation are necessary to deter attack and prevent war. U.S. policymakers relied heavily on this model in developing national security policy during the Cold War. It formed the basis of America's strategy of containment vis-à-vis the Soviet

Union, prescribing a massive arms buildup as well as military intervention in places of little strategic or material value to the United States for the purpose of maintaining credibility in other, more vital locations. The wars in Korea and Vietnam were fought not so much because the United States had vital interests in these regions, but because U.S. policymakers feared that Soviet and other communist leaders would not believe their threats to defend Western Europe if they failed to respond to provocations in Asia.

Since the end of the Cold War, far lesser menaces than the Soviet Union and international communism have been viewed by U.S. leaders through the lens of the 1930s. When Iraq invaded and conquered Kuwait in August of 1990, the administration of President George H.W. Bush quickly invoked the so-called Munich analogy, comparing Iraqi leader Saddam Hussein to Hitler in arguing that Iraq could not be allowed to retain Kuwait because doing so would only encourage further Iraqi aggression. Indeed, in the United States, appeasement continues to own such an odious reputation that the label is still employed frequently and almost indiscriminately by politicians and pundits seeking to discredit diplomatic initiatives to which they object.

Challenging the Conventional Wisdom

Although in much of popular and political discourse appeasement remains in disrepute, within the scholarly community the conventional wisdom concerning appeasement has been repudiated to a considerable degree. Some historians have challenged the notion that Hitler regarded the Munich Agreement as a great diplomatic victory, arguing instead that he considered it a crushing defeat, having been thwarted in his desire to take the Sudetenland by force. They contend that Hitler, having accepted Chamberlain's offer and stepped back from the brink of war, began to doubt his own – not his enemies' – fortitude and courage. Hitler, in this view, found Anglo-French threats to defend Poland credible, but decided to invade, anyway, determined, after Munich, not to back down again (Weinberg, 1980). Other scholars suggest that to the degree that Hitler found Anglo-French threats not credible, this may have been due to his own motivated bias – wishful thinking – rather than previous behavior by Britain and France (Rock, 2000).

Students of international politics have cast serious doubt on the belief that appeasement is dangerous because it undermines subsequent efforts at deterrence. This belief rests on the assumption that policymakers base predictions about a state's future behavior on its past actions. But according to Jonathan Mercer (1996: 227), "decision-makers do not consistently use another state's past behavior … to predict that state's behavior." Daryl Press (2005) tests what he calls 'past actions' theory against 'current calculus' theory and concludes that states typically do not assess the credibility of deterrent threats on the basis of a state's past behavior; instead, they rely on an analysis of current factors and conditions, including the perceived military balance.

Finally, researchers have challenged the view that appeasement as a diplomatic strategy is doomed to failure (Rock, 2000). Cases of successful appeasement exist. Of these, perhaps the most notable is British conciliation of the United States during the last decade of the nineteenth century and the first decade of the twentieth century. At this time, Britain's security position was precarious. In Europe, it faced a hostile France and Russia, along with a rapidly rising Germany. In the Western Hemisphere, it confronted the United States, like Germany a late industrializer and growing maritime power, but with far greater potential, given its geographic position, population, and natural resource base. Beginning in late 1895, the United States began to assert itself against British interests. Invoking the Monroe Doctrine, it claimed the right to intervene in a dispute between Britain and Venezuela concerning the boundary between Venezuela and the colony of British Guiana. Although the British government initially rebuffed U.S. overtures, once a crisis erupted and war seemed possible, it quickly reconsidered and agreed to submit the matter to international arbitration.

In subsequent years, additional disagreements were resolved in a similar manner. Under pressure from the U.S. government, British leaders abandoned the Clayton–Bulwer Treaty, renouncing their right to a voice in the construction of a canal through the Central American isthmus. They agreed to submit a dispute concerning the boundary between Alaska and the Canadian Yukon to a panel of six impartial jurists, three from the United States, two from Canada, and one from Britain. When U.S. President Theodore Roosevelt stacked the panel, the British government swallowed hard and accepted his appointments. Then, when it appeared that the panel might be deadlocked, the British representative sided with the Americans against his Canadian colleagues. Between 1904 and 1906, Britain withdrew its naval forces from the Americas, effectively ceding control of the Western Hemisphere to the United States.

The wisdom of Britain's appeasement strategy is evident in hindsight. By placating the United States, British leaders avoided a war that would have been ruinous economically and militarily, one that Britain might well have lost. Having also achieved a rapprochement with France and Russia, British policymakers freed themselves to concentrate on their most dangerous foe: Germany. And, ultimately, they gained an ally, with the United States fighting on Britain's side in World War I. It is nearly impossible to imagine such a favorable outcome had Britain elected to pursue an unyielding, hard-line policy.

When Does Appeasement Work?

Since the historical record of appeasement is mixed, it is important to understand the conditions under which it succeeds and fails. To do so requires specifying the goals that appeasement might be expected to achieve. Traditionally, the policy has been regarded as being aimed at the avoidance of war. But, in the case of Britain and the United States, avoiding war was only the minimum aim for British leaders. Their maximum aim was considerably more ambitious: to make a friend or even an ally of the United States. Generally, the objectives of appeasement may differ according to time horizon (short-term vs long-term) and whether they aim at maintenance or alteration of the international status quo. As a short-term strategy, appeasement may seek to resolve a crisis or result in a limited political trade. As a long-term approach, it

may seek crisis prevention or to produce friendship or alliance. For the most part, these goals are not mutually exclusive. As the Anglo-U.S. case demonstrates, an appeasing state may have both short-term (crisis resolution) and long-term (friendship/alliance) objectives (Rock, 2000).

Appeasement, when it succeeds, may do so via multiple mechanisms. The dominant method, at least according to the conventional wisdom, is satiation: the satisfaction of the (usually territorial) demands of a revisionist state. However, British policymakers during the 1930s had at least three other mechanisms in mind as they sought to conciliate Nazi Germany. One of these was reassurance: an effort to increase an insecure state's confidence in its security position. A second was socialization: an attempt to teach belligerent, bumptious German leaders how to behave in polite, international society. The third was manipulation of Germany's internal political balance. Believing that Germany was not a monolithic state and that there existed a powerful group within the German government that favored peace, the British government sought by means of concessions to strengthen the hand of supposed German moderates in opposing Hitler's bellicose policies (Rock, 2000).

Whether or not appeasement succeeds depends on whether or not these mechanisms are effective in a given case. Perhaps the most critical factor is the nature of the adversary. War-seeking states cannot be appeased because concessions do not address their primary objective: entry into armed conflict. Indeed, as the Munich experience demonstrates, inducements (particularly offers that cannot be refused because they grant what a state claims to be seeking) only serve to frustrate such states and their leaders by delaying the accomplishment of their intended purpose. By contrast, greedy states with limited, tangible ambitions can be appeased. The same is true of states that are insecure. The former can be satiated, while the latter can be reassured. The United States, at the turn of the twentieth century, was a state both greedy and insecure. British concessions satisfied America's limited territorial ambitions (and the country's desire for recognition as the dominant power in the Western Hemisphere) while simultaneously alleviating U.S. concerns about Britain's intent to expand its empire in Latin America. There is no evidence, contrary to the conventional wisdom regarding appeasement, that Britain's inducements caused the United States to enlarge the scope of its aims or to become more aggressive in its pursuit of them (Rock, 2000).

A second factor that helps to determine the success or failure of appeasement is the nature of the inducements offered by the appeasing state. In order to be effective, concessions must be properly directed – that is, they must address the needs or demands of the target state. Early British efforts to conciliate Germany, such as allowing German rearmament in violation of the Treaty of Versailles, did not address Germany's territorial ambitions, although they did, counterproductively, enhance Germany's prospects for achieving these ambitions through the use of force. Concessions must also be sufficient in the sense that they must be adequate to satiate (or reassure) the target state. Britain's willingness to redraw the colonial map of Africa, for example, did not come close to satisfying Hitler's ambitions, which lay mainly in the continent of Europe. By contrast, British concessions to the United States directly and fully addressed U.S. needs and wants. Finally, concessions must be properly perceived by the target state. Perception depends both on the transmission of signals and on their reception. U.S. efforts to conciliate Saddam Hussein's Iraq in the years leading up to the Iraqi invasion of Kuwait were thwarted in part by Congressional attempts to impose economic sanctions in response to Iraqi human rights abuses, as well as anti-Iraqi sentiments expressed by other organs of the U.S. government, including the Department of State. The U.S. government sent mixed signals that were difficult for the Iraqi government to interpret. On the receiving end, in some cases, the offer of inducements is perceived as not being genuine, or even as a trick; in others, the message of conciliation is lost in the flow of diplomatic exchanges (Rock, 2000).

A third factor that helps to explain the success or failure of appeasement is the existence (or lack thereof) of incentives external to the relationship between the appeaser and the target state for the latter to accept and be satisfied by the concessions that are offered. Such factors may be international – e.g., the presence of a foreign adversary against which the cooperation of the appeasing state might be valuable. They may be internal – ideological, political, or economic reasons for the state to decide that abandoning a rivalry would be advantageous to it. One can see such factors at work in the U.S. decision to accept British concessions in the late nineteenth and early twentieth centuries. The United States faced what it perceived to be a serious security threat from Germany, which better relations with Britain would alleviate. It also possessed strong cultural reasons (in the form of Anglo-Saxon nationalism), political/ideological reasons (in the form of shared liberal values and democratic institutions), and economic reasons (due to a high level of commercial and financial interdependence) to respond positively to Britain's overtures (Rock, 2000).

Why Do States Choose Appeasement?

Appeasement is a policy adopted by states in dealing with a threatening or potentially threatening adversary. It is not, however, the only option available. Rather than seeking to redress an opponent's grievances, a state may choose to stand firm, rejecting the demands that are placed on it and attempting to defend its interests by upholding the balance of power, maintaining a posture of deterrence, and, if necessary, going to war. Why do states in some instances choose appeasement over what might be regarded as the preferred option, one that, if successful, enables a state to uphold its interests and maintain its international position?

During World War II and immediately thereafter, the answer to this question typically took the form of the so-called guilty men thesis. Taking its name from the title of a devastating if not entirely accurate polemic published pseudonymously in 1940 by opponents of British appeasement, the essence of this explanation is that the policy was undertaken by leaders who were too weak and pusillanimous to confront an aggressor or too feeble-minded and foolish to understand what was at stake (Cato, 1940). Neville Chamberlain, who was singled out for particular criticism, and many of his colleagues in the British government were vilified and are sometimes still vilified as the embodiment of this sort of leader.

With the benefit of temporal distance from the failures of the 1930s, scholars began to take a more dispassionate approach to the matter. It is now well understood that a balancing strategy may not always be feasible. External balancing, which involves reliance on other states, requires the availability of powerful allies. Internal balancing, which involves the mobilization of domestic resources, demands the production of armaments, an increasingly large military force, and ultimately the commitment of extensive economic and physical assets.

Beyond the matter of feasibility, there are other reasons that states may appease rather than balance or deter. Potential allies may be rejected because they are deemed to be unsuitable by political leaders or public opinion. The citizenry of a country may be opposed to preparations for war, fearing that they indicate the intention of the government to become involved in one. Because internal balancing is expensive, it may not be possible for a government to pursue simultaneously a balance-of-power or deterrent strategy and satisfy its (and its supporters') domestic welfare aspirations. National leaders who cannot provide both guns and butter at the same time must choose between them, and domestic political calculations flowing from the desire to remain in power may push them in the latter direction (Trubowitz and Harris, forthcoming).

Britain's decision to appease Nazi Germany during the 1930s reflected many of these factors. For the British government, the most essential allies were not available. France, having been devastated in World War I, had scant desire to confront Germany. The United States, wary of European entanglements, was profoundly isolationist. Even the British dominions favored appeasement. And the Soviet Union was rejected as a potential ally because the British government (with some justification) deemed it to be untrustworthy.

If effective external balancing seemed impossible to British leaders, internal balancing appeared scarcely less so. Britain was a relatively small country with a limited population and resource base. Its economy had suffered greatly as a result of World War I and the Great Depression. Chamberlain and the Conservative Party more generally had staked their electoral prospects on domestic reforms and economic recovery. To choose guns over butter – a choice could not be avoided, since Britain could not afford both – would have been unpopular and politically problematic. Beyond this, the so-called anti-appeasers were unable to unite in support of an alternative policy. Given the international and domestic environments in which the Chamberlain government was operating, it is hardly surprising that it opted to pursue a strategy of appeasement (Trubowitz and Harris, forthcoming).

In grand strategic terms, appeasement is mainly an alternative to balancing. However, appeasement can also be an essential component of a state's efforts to maintain a favorable balance of power. This is true most obviously when a state decides to appease one adversary in order to focus its attention on another, as when Britain elected, prior to World War I, to accommodate the United States partly in order to confront more effectively the danger from Germany. Appeasement may also serve tactical purposes consistent with balancing. For example, if the intentions of an adversary are unclear, appeasement may enable a state to test and clarify them. This was one motivation for Britain's appeasement policy during the 1930s. More importantly, perhaps, appeasement may buy time so that an unprepared state may ready itself to meet a foreign threat on the field of battle. Ripsman and Levy (2008) argue that British leaders, convinced that they were in a position of relative weakness, sought to appease Germany in order to delay war until Britain had rearmed.

The Ethics of Appeasement

Following the failure of Munich, appeasement rapidly acquired a reputation as a morally suspect, if not altogether bankrupt policy. Chamberlain's decision, made with the support of French President Edouard Daladier, to abandon the Czech government and cede the Sudetenland to Germany in an attempt to pacify Hitler seemed in retrospect unconscionable. Although none other than Winston Churchill wrote after the war that "appeasement in itself may be either good or bad depending on the circumstances," his was a view not widely shared (Record, 2005: 1). In 1956, Webster's offered this definition of appeasement: "To conciliate by political, economic, or other considerations; now usually signifying a sacrifice of moral principle" (Herz, 1964: 298).

The moral status of appeasement, however, merits a more thorough consideration than it has usually been given. Inasmuch as a state's right to individual and collective self-defense is generally acknowledged by scholars and enshrined in international law, there would seem to be few situations in which appeasement would be obligatory. Helen Frowe (2011: 39) suggests that appeasement might be required when the harm that would be produced by it is not serious, but she notes, "it will almost never be the case that appeasement of aggression between states avoids serious harm." Michael Walzer (2006: 51) is more definitive, writing that those confronted with aggression are "always justified in fighting." Pacifists typically accept that aggression may be resisted; they differ from non-pacifists in insisting that resistance must not involve the use of deadly force (Holmes, 1989). Given the apparent consensus that appeasement is rarely, if ever, required; the more interesting question is whether or not it is morally permissible.

One argument against appeasement is that it violates a duty to resist evil. According to Andrew Bacevich, neoconservative critics of appeasement in the United States have drawn two lessons from the 1930s. First, "that evil is real." And, second, "that for evil to prevail requires only one thing: for those confronted by it to flinch from duty" (Bacevich, 2005: 73). To condemn appeasement as a failure to resist evil, however, requires one to assume that an adversary's grievances or demands are never legitimate and that there are never other evils more worth resisting. That these assumptions are valid is not self-evident. Indeed, during the 1930s, many British policymakers explicitly rejected them. They saw as their primary responsibility the avoidance of another bloody, financially ruinous war. Moreover, they did not regard German demands as illegitimate, believing that Germany deserved relief from the onerous burdens of the Treaty of Versailles and that the right of self-determination for the Sudeten Germans and other ethnic Germans living outside of Germany should be upheld. Had their policy succeeded, Chamberlain and his colleagues might today be applauded for their principled stance. Although they

failed, it is hard to see how a duty to confront evil could rule out appeasement in every case.

A second, more plausible argument against appeasement is that it violates a duty to defend. Is it morally permissible for one state to refuse to defend another, especially when the latter is incapable of defending itself? This is the question posed by the British cession of the Sudetenland to Germany in 1938. It seems evident that had Britain been bound to Czechoslovakia by a formal, contractual agreement (i.e., a treaty of mutual defense), then it would have been obligated to defend the Czechs. But no such agreement existed. Any obligation must therefore have come from Britain and Czechoslovakia's shared membership in an international community of states in which the members had reciprocal rights – above all, territorial integrity and sovereignty – and duties, including a duty of collective security. Whether or not an international community existed in the late 1930s, particularly following the collapse of the League of Nations, may be debated. In the twenty-first century, with a more robust United Nations claiming almost universal membership, such a notion, associated with what one scholar calls the 'state-moralist' position, would seem easier to sustain. The rise of a 'cosmopolitan' school of thought, which posits the existence of a global community of human individuals, rather than of states, may make the idea of moral responsibilities beyond one's own borders still more plausible (Nye, 1986).

The issue nonetheless remains complicated by the fact that states have competing moral responsibilities: to the governments and peoples of other states, but also, and perhaps primarily, to their own citizenry. Was it not the first duty of British leaders at the time of Munich to protect the lives and welfare of the British people rather than the sovereignty and territorial integrity of the Czech state? Neville Chamberlain appears to have believed that it was. So, too, did the British public, judging by the adoring crowds that greeted the prime minister on his return from Germany. The British government failed to understand that the interests of Britain's citizens might have been better defended at Munich by fighting for Czechoslovakia rather than by acquiescing to Hitler's demands, but this does not necessarily mean that its priorities were misplaced.

Beyond the question of whether or not a state must defend another, there is the issue of whether or not it is morally permissible for a state to refuse to defend itself, to appease a potential aggressor by sacrificing its own territory, population, or resources – at the extreme, by sacrificing its independence – by declining to engage in armed combat. Utilitarian reasoning suggests that under some circumstances, it may be. When armed resistance is futile, the harms of war and forcible submission are almost certain to exceed the harms associated with surrender (Orend, 2013). In such situations, it would seem better to accede to the adversary's demands than to resist. The eminent just war theorist Michael Walzer (2006) concedes the force of this argument. Nevertheless, he writes, "In most cases … fighting is the morally preferred response." "We want," he says, "to live in an international society where communities of men and women freely shape their separate destinies. But that society is never fully realized; it is never safe; it must always be defended" (51, 72). States must fight, even in the face of almost certain defeat; otherwise, the core values of the international community and the rights of states (and of individuals from which the rights of states are derived) will be diminished or lost.

Although he regards fighting as the morally preferred option in most cases, Walzer does not rule out appeasement in every instance. The obligation to fight existed with respect to Germany during the 1930s because Germany was "committed to the continual use of violence, to a policy of genocide, terrorism and enslavement" (Walzer, 2006: 69). It existed with respect to Stalinist Russia, which, like Germany, posed a fundamental challenge to international norms crucial to the rights of states and those living within them. Walzer suggests, however, that appeasement of a more conventional great power would be morally permissible as long as it did not "involve absolute subjection to the 'rule of violence'" (69). He must believe, therefore, that a state may also decline to defend itself and its interests against an adversary whose ambitions are less extensive and whose intentions are less malign.

Appeasement is inherently neither moral nor immoral; it must be evaluated on a case-by-case basis. In any given case, the moral permissibility of appeasement rests on a number of factors that include, but are not necessarily limited to, the following: (1) the legitimacy of the adversary's grievances and demands, (2) the harm produced by the concessions that are made, (3) the likely harm produced in the event that appeasement fails, (4) the likely harm produced by refusing to appease and adopting an alternative strategy (e.g., balancing, deterrence, war), and (5) the likelihood that appeasement will succeed. At one extreme, appeasement would appear to be justified in cases in which the adversary's grievances are legitimate, the harm resulting from concessions is small, the likely harm resulting from refusing to appease is large, the failure of appeasement is not likely to produce severe harm, and the prospects for successful appeasement are good. At the other extreme, appeasement would appear to be unjustified in cases in which the adversary's demands are not legitimate, the harm resulting from concessions is large, the likely harm resulting from refusal to appease is small, the damage from unsuccessful appeasement would be great, and the prospects for successful appeasement are poor.

By these criteria, Britain's late-nineteenth- and early-twentieth-century appeasement of the United States appears to have been justifiable, while its attempted conciliation of Nazi Germany during the 1930s appears to have been much less so. Of course, evaluating the morality of appeasement is relatively easy in hindsight, when outcomes are known. Political leaders confronted with the decision to appease or not to appease operate in an uncertain environment and must make calculations regarding the consequences that might ensue from whatever policy they choose. The moral status of their strategy cannot be evaluated by its results, which can never be entirely foreseen. It must be evaluated on the basis of whether or not the leaders' assessment of the situation and their calculations of likely outcomes seem reasonable or defensible, given the information available to them at the time.

Engagement: A Modern Form of Appeasement?

Appeasement is now considered by many scholars to be an effective and legitimate tool of statecraft under certain

circumstances. Practitioners of foreign policy largely share this view. They are careful, however, to refrain from using the term 'appeasement' because the word retains its odious reputation in the public mind and represents an easy target for political foes.

Since the 1980s, scholars and policymakers, particularly in the United States, have preferred to speak and write of 'engagement.' Engagement, by one definition, is "the use of non-coercive diplomatic tools and initiatives to elicit cooperative behavior from the target state." This is accomplished through: "(1) accommodation of the target state's interests (regardless of normative consideration); (2) an 'enmeshing' of the state in international institutions upon which its well-being becomes based; or (3) the growth of a variety of transnational linkages that transform the preferences of the state" (Cha, 1999: 36). The Ostpolitik of West German leader Willy Brandt during the 1960s is an example of engagement. More recently, U.S. administrations have pursued engagement with respect to the white-minority government in South Africa and with respect to Vietnam, North Korea, China, Iraq, and Iran, among other states.

Engagement strongly resembles appeasement as it was practiced in classical European diplomacy and by the British government during the 1930s. Johnston and Ross (1999: xii) argue that the two strategies differ in name only, having "identical policy options and objectives." Like appeasement, engagement aims at the reduction of tension. It "depends to a significant degree on positive incentives to achieve its objectives" (Haass and O'Sullivan, 2000: 113). It operates by means of satiation, reassurance, socialization, or alteration of an adversary's domestic political balance – or some combination of these.

There is thus considerable truth to the claim that engagement is "a modern form of appeasement" (Rock, 2000: 22). But the two are not exactly the same. Engagement can be a comprehensive strategy, operating on a variety of fronts. It may seek to maintain open lines of communication, cooperate in matters of mutual interest, and integrate an adversary into the international legal, political, and economic systems in order to gain leverage and constrain its behavior. Appeasement is generally narrower in scope, relying more heavily on inducements to satisfy grievances. It is perhaps most appropriate to think of appeasement as a type of engagement. Indeed, Randall Schweller (1999: 14) has written, "the most common form of engagement is the policy of appeasement."

See also: Conflict Mediation; Deterrence: Sanction Perceptions; Military, War, and Politics; Nation-State and War; National Security Studies and War Potential of Nations; Peace Processes; Peace-Making in History; Second World War, The; Territorial Conflicts; Wars among Nation-States: Patterns and Causes.

Bibliography

Bacevich, A.J., 2005. The New American Militarism: How Americans Are Seduced by War. Oxford University Press, Oxford.
Baldwin, D., 1971. The power of positive sanctions. World Politics 24, 19–38.
Beck, R.J., 1989. Munich's lessons reconsidered. International Security 14, 161–191.
Cato (Owen, F., Foot, M., Howard, P.), 1940. Guilty Men. Gollancz, London.
Cha, V.D., 1999. Engaging China: the view from Korea. In: Johnston, A.I., Ross, R.S. (Eds.), Engaging China: The Management of an Emerging Power. Routledge, London, pp. 32–56.
Cortright, D. (Ed.), 1997. The Price of Peace: Incentives and International Conflict Prevention. Rowman and Littlefield, Lanham, MD.
Craig, G.A., George, A.L., 1990. Force and Statecraft: Diplomatic Problems of Our Time, second ed. Oxford University Press, New York.
Frowe, H., 2011. The Ethics of War and Peace: An Introduction. Routledge, New York.
Haass, R.N., O'Sullivan, M.L., 2000. Introduction. In: Haass, O'Sullivan (Ed.), Honey and Vinegar: Incentives, Sanctions, and Foreign Policy. Brookings Institution Press, Washington, DC, pp. 1–11.
Herz, J., 1964. The relevancy and irrelevancy of appeasement. Social Research 31, 296–320.
Holmes, R.L., 1989. On War and Morality. Princeton University Press, Princeton.
Jervis, R., 1976. Perception and Misperception in International Politics. Princeton University Press, Princeton.
Johnston, A.I., Ross, R.S., 1999. Preface. In: Johnston, Ross (Ed.), Engaging China: The Management of an Emerging Power. Routledge, London, pp. xi–xv.
Kennedy, P.M., 1976. The tradition of appeasement in British foreign policy, 1865–1939. British Journal of International Studies 2, 195–215.
Layne, C., 2008. Security studies and the use of history: Neville Chamberlain's grand strategy revisited. Security Studies 17, 397–437.
Mercer, J., 1996. Reputation and International Politics. Cornell University Press, Ithaca, NY.
Morgenthau, H., 1951. In Defense of the National Interest: A Critical Examination of American Foreign Policy. Alfred A. Knopf, New York.
Nye Jr., J.S., 1986. Nuclear Ethics. Free Press, New York.
Orend, B., 2013. The Morality of War, second ed. Broadview, Toronto.
Press, D.G., 2004/2005. The credibility of power: assessing the credibility of threats during the 'appeasement' crises of the 1930s. International Security 29, 136–169.
Press, D.G., 2005. Calculating Credibility: How Leaders Assess Military Threats. Cornell University Press, Ithaca, NY.
Record, J., 2005. Appeasement Reconsidered: Investigating the Mythology of the 1930s. Diane Publishing, Darby, PA.
Richardson, J.L., 1988. New perspectives on appeasement: some implications for international relations. World Politics 40, 289–316.
Ripsman, N.M., Levy, J.S., 2008. Wishful thinking or buying time? The logic of British appeasement in the 1930s. International Security 33, 148–181.
Rock, S.R., 2000. Appeasement in International Politics. University Press of Kentucky, Lexington.
Schweller, R.L., 1999. Managing the rise of great Powers: history and theory. In: Johnston, A.I., Ross, R.S. (Eds.), Engaging China: The Management of an Emerging Power. Routledge, London, pp. 1–31.
Thornton, B.S., 2011. The Wages of Appeasement: Ancient Athens, Munich, and Obama's America. Encounter Books, New York.
Trubowitz, P., Harris, P. When states appease: British appeasement in the 1930s. Review of International Studies, forthcoming.
Walzer, M., 2006. Just and Unjust Wars, fourth ed. Basic Books, New York.
Weinberg, G., 1980. The Foreign Policy of Hitler's Germany. University of Chicago Press, Chicago.

Applied Criminology and Criminal Justice

Brian Stout and Kerry Clamp, School of Social Sciences and Psychology, University of Western Sydney, Penrith, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Applied criminology may be thought of as being concerned with the application of the discipline of criminology to 'real world' problems of crime and criminal justice. It is both critical and engaged and seeks to find solutions to particular issues of crime and justice, as well as to problematize suggested approaches. The article sets out to review and explore attempts to address crime, to deal with offenders, and to respond to victims.

Introduction: Applied Criminology: What Is It and How Does It Work?

This article is concerned with surveying the field of applied criminology or criminology in its applied form. Applied criminology may be thought of as being concerned with the 'real world' problems of crime and criminal justice and how the research and scholarship within the discipline of criminology can relate to them. It is influenced by both contemporary academic criminology, which seeks to explore and conceptualize crime, its causes, and the responses to it in theoretical terms, and administrative criminology, which has largely become an attempt to "implement pragmatic means of crime opportunity reduction and to manage crime through situational prevention measures" (Muncie, 1998: 4).

Applied criminology seeks to address and compensate for the perceived deficiencies of other criminological approaches. Academic or theoretical criminology is vulnerable to the criticism of being distant from real-world concerns and detached from the public, criminal justice agencies and practitioners. Administrative criminology can sometimes be atheoretical or uncritical, and it suggests that the causes of crime are relatively unimportant or politically impossible to tackle (see Young, 1988). Applied criminology can be viewed as an area of study, which not only seeks to find solutions to particular issues raised by crime and criminal justice, but also to problematize current and suggested approaches.

Applied criminology, as conceptualized by Stout et al. (2008), seeks to draw on both academic criminology and administrative criminology. Canton and Yates (2008: 5) define applied criminology as "a criminology which self-consciously and deliberately explores the insights of criminology for their relevance and application to policy and practice" and which "should have a critical edge, casting a discriminating, analytical gaze over the process of criminalisation, crime enforcement, and the criminal justice system" (2008: 6). Criminological research and theoretical knowledge has significant relevance for addressing crime, offending, and victimization, and the collection makes explicit the relationship between governmental agendas and knowledge production.

Criminology, in its applied form, should engage with real-world concerns but should not be simply pragmatic or functional. It should have a critical edge, casting a discriminating, analytical gaze over the process of criminalization, crime enforcement, and the criminal justice system. It should seek to expose the relationship between governmental agendas and knowledge production (Canton and Yates, 2008: 6). Although the term 'applied criminology' is a familiar one, appearing in the titles of courses, conferences, and research centers, it is not possible to claim that it is, in itself, a distinct area of criminology. It does not claim to offer major premises or unifying theories, which define crime, isolate the causes of offending, and offer a solution to the crime problem (such as critical criminology, peacemaking criminology, and so on). Applied criminology should be thought of as an approach, which not only seeks to find solutions to particular issues raised by crime and criminal justice through established criminological endeavors, but also to problematize current and suggested approaches embraced by governments and criminal justice agencies. It directly takes on the challenge implied in Loader and Sparks's (2010) conceptualization of criminology as a 'successful failure' – a discipline that is successful in attracting the attention of students and academics, but that has a declining influence outside the academy. Applied criminology addresses the practice implications of criminological scholarship.

With this orientation in mind, we use the term to discuss criminology that has a direct connection with the 'real' world and thus highlight the role that criminological research has had on shaping criminal justice and its responses to offenders and victims at a policy level. In their characterization of applied criminology, Canton and Yates (2008: 2) identify three principal questions, which they suggest that applied criminology should address:

- What is to be done about offenders?
- What is to be done about crime?
- What is to be done on behalf of the victims of crime?

This article uses a version of these questions, in a slightly different order, to outline how criminology is being applied to address each of these questions in turn.

What Is to Be Done about Crime?

A starting point for an account of the application of criminology is to take stock of the context of the understanding of crime. Until the early postwar period, criminology had been dominated by the positivist approach, which viewed the causes of crime as a consequence of one's circumstances and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10512-4

neoclassicism, which was based on the belief that certain responses would reduce the likelihood of repeat offending. However, during the late 1960s and 1970s, respectively, the discipline underwent two major crises (Young, 1988). The first related to what Young (1988) referred to as the "aetiological crisis." For the most part, positivism was based on the belief that as social conditions improved, postwar crime would decrease. However, crime did not decline as many had suspected, instead it continued to rise. The second was a crisis of penalty following the publication of a number of police studies in the United States, which questioned the effectiveness of the police and the reformative potential of prisons in dealing with crime. This found expression in the 'nothing works' pessimism, discussed in the next section, and, coupled with increasing crime rates, led to questioning of the ability of the state to control crime (Garland, 2001: 62).

Many advanced liberal democracies responded to this perceived limitation by reforming criminal justice institutions and mobilizing nonstate mechanisms in the "fight against crime" (Crawford, 2008). The 1970s and 1980s were characterized by the revival of the 'just deserts' model, which focused on the punishment of the offense rather than on the needs of the offender; an increased role for both the victim and the 'community' in the administration of justice (rather than professionals only); and the 'politicization' of crime control in which 'public opinion' became a central issue. The rise of "populist punitiveness" (Bottoms, 1995) or the "law and order" ideology (Cavadino and Dignan, 2002) has often involved politicians talking tough and introducing ever more stringent penal policies in order to secure public support (Young and Matthews, 2003). Garland (2001: 172) succinctly describes this development as:

> Criminal justice is now more vulnerable to shifts of public mood and political reaction … The populist current in contemporary crime policy is, to some extent, a political posture or tactic, adopted for short-term electoral advantage.

This, unsurprisingly, has resulted in increased resources being devoted to policing, prosecutions, and prisons (Roach, 2005; McEvoy, 2007) as advanced liberal governments seek to demonstrate to their publics that they are able to control crime. A widespread fear of crime and the ability of politicians and the media to generate "moral panics" (Cohen, 1972) have made this a popular strategy for those wanting to serve vested interests. However, this has also had the adverse effect of dramatically increasing prison populations and creating a perception of increased criminal incidents, despite a relatively stable decline in crime rates in the decade since the mid-1990s (Cesaroni and Doob, 2003; Young and Matthews, 2003).

Many developed nations have therefore increasingly become aware that an exclusive adherence to a punitive strategy has not produced the desired result in terms of "greater satisfaction with criminal justice or an increased sense of security of justice" (Roach, 2005: 2). An increase in caseloads; a perceived lack in the ability of the criminal justice system to deal with crime effectively; and an increase in victim advocacy (Albrecht, 1999; Garland, 2001; Zedner, 2002; McEvoy, 2007) have therefore subsequently resulted in concerted efforts to find new and innovative ways of dealing with offenders and victims.

A highly visible attempt to reformulate the relationship between criminal justice and its public has been located within policing practice. The introduction of community-based policing, problem-oriented policing, and restorative policing during the 1980s and 1990s offered the police an alternative to the reactive, incident-focused style (Paterson and Clamp, 2012). Such initiatives acknowledge the central role of the community in identifying, reporting, and responding to crime, and it is thought that the police are more likely to successfully tackle crime where policing strategies are based within and informed by community members themselves. However, this ideological shift runs contrary to the historic policing mission where independent police professional knowledge directs local developments and it has thus generated a certain level of resistance by police officers. Alongside this, a body of critical literature has developed that questions the extent to which this shift in power and responsibility has, or even can, take place (Hobsbawm, 1995; Bauman, 2001). This literature questions whether placing 'the community' at the center of policing policy makes sense, especially during a historical period in which many communities (in their traditional sense) are understood to have disappeared (see Clamp and Paterson, 2011).

Criminology has also engaged with policing in questioning its very definition of crime and its approach to how offending is measured and recorded thus leading to high-profile and heated debates between police stakeholders and the academic community (Eterno and Silverman, 2012). It has often been the role of criminologists to hold the police to account and to strongly criticize police practice, when required, such as in the two high-profile British examples of the Hillsborough football stadium disaster (Scraton, 2009) and the murder of Stephen Lawrence (Bowling and Phillips, 2002). However, increasingly, across the world, the application of criminology to policing is taking place within police training with university-level preentry police training programs being offered in, among others, the UK, the United States, and Australia. Criminology can then play a dual role of an outsider, researching and commenting on practice, and an engaged insider, contributing to the development of the profession.

What Is to Be Done about Offenders?

Criminological research has informed debates in criminal justice with regard to the difficult question of what should be done with offenders. Criminological research has been applied in determining what approach should be taken in effectively rehabilitating offenders, but this issue has not always been simple or straightforward.

Discussions of the effectiveness of interventions with offenders conventionally start with the notorious meta-analysis of 231 cases (Lipton et al., 1975) and the conclusion from that analysis, most often associated with Robert Martinson, which almost nothing works with offenders. This research was often mischaracterized and was used for political purposes, by both left and right (Miller, 1989), but it provided an impetus for the lengthy and ongoing work of establishing what is effective in

working with offenders. Martinson himself later retracted his 'nothing works' conclusion and accepted that some things work in some circumstances (Martinson, 1979). Applied criminology has been concerned with determining what is effective in working with offenders. However, there are also arguments that simply asking 'what interventions are effective?' is misdirected and reductive; it is also important to take account of offenders' rights and of the narratives that offenders themselves relate about their desistance from offending.

Over the last decades, the question of 'what works' or what is sometimes known as 'effective practice' has dominated the application of criminology to work with offenders. Mackenzie (2000) summarized the extent of knowledge by identifying that some of the approaches that worked included:

- Rehabilitation programs with particular characteristics
- Incapacitating offenders who commit crime at high rates
- Some prison-based therapeutic programs
- Programs based on cognitive behavioral therapy approach
- Nonprison-based sex offender treatment programs
- Vocational education programs in prison or in other residential settings
- Multicomponent correctional industry programs
- Community employment programs

Mackenzie also identified approaches that had been consistently shown not to work in reducing reoffending. These included specific deterrence programs, interventions based solely on control or surveillance, and boot camps and other wilderness programs. Counseling approaches that were vague and unstructured also had no discernible effect on recidivism rates. This vein of research continues with increasingly detailed knowledge being produced regarding what is effective in working with offenders. In the UK, the Probation Studies Unit of the Centre for Criminological Research at the University of Oxford had the stated goal of carrying out research into the effectiveness of various interventions, and they emphasized the need for well-designed studies to evaluate accredited programs (Roberts, 2004). In Australia, the New South Wales Bureau of Crime Statistics and Research (BOCSAR) is the official source of crime statistics for the state but also evaluates initiatives designed to reduce crime and reoffending (BOCSAR, 2013).

The focus of criminological attention on, largely quantitative, studies to determine the effectiveness of particular interventions on changing offender behavior has been subject to challenge and criticism within criminology. Feeley and Simon (1992) identified a new penology with a focus on actuarial and efficient and effective processes that could lower our expectations of the system and become detached from traditional concerns of rehabilitation or punishment. Garland (2001: 19) went further in describing the "all pervasive managerialism" that affected all aspects of criminal justice and led to a social way of reasoning being replaced by an economic way of reasoning. Cohen's (1985) concerns were that an applied criminology could be merely technicist and that measures introduced ostensibly to help and support an offender could simply become another mechanism of control. If applied criminology is to be more than merely technicist or administrative, then it must address these concerns and there are a number of ways that this has been done. Applied criminology is also concerned with the rights of offenders and with their own self-narratives regarding how they desist from offending, and in supporting them to do this.

A simple focus on the effectiveness of programs in addressing offending behavior can lead to crime being considered as a natural disorder, analogous to illness, and can leave crucial questions unexamined. For example, a simple focus on the effectiveness of interventions can ignore the fact that crime, and thus offending behavior, is social constructed. It can also neglect questions of the differential representation of particular groups in the criminal justice system and the link of that representation to questions of social justice. Viewing rehabilitation as a right, regardless of the effectiveness of particular interventions, can help address this problem. Rehabilitative and therapeutic services as well as educational and employment support can be provided without a need to calculate their effectiveness (Rotman, 1986). Carlen (2013) has gone further, arguing that a focus on rehabilitation is a distraction from the injustices in society and in the criminal justice system. She argues that we need an inclusive social justice that reduces inequality and promotes citizenship.

The increasing attention given to what is now known as desistance research and its focus on hearing from offenders regarding how they have moved away from offending has allowed criminological research and insights to become applied to work with offenders. Desistance research has early roots in the United States, particularly in the longitudinal life-course studies analyzed by Sampson and Laub (1993), which investigated how individuals became involved in offending and later came to cease offending behavior. They found that significant life events, such as getting married or finding employment, were correlated with desisting from offending. More recently, primarily in the UK, research and scholarship (see, for example, Maruna, 2001; Farrall and Calverley, 2006; McNeill, 2006) has built on and developed this research to create a desistance paradigm. The desistance paradigm emphasizes the involvement of the offender in his own progress toward desistance and the need to work through partnership, collaboration, negotiation, and an emphasis on strengths. Desistance is an active process, involving the building of social capital and the support of criminal justice practitioners; it is not simply a program that can be imposed on an individual to change their thinking. Crucially, what makes desistance research work in applied criminology is its explicit drawing of links to practice and its desire to provide a framework for both practitioners and policy makers (see McNeill, 2006; Farrall and Calverley, 2006).

A further approach to the application of criminology that avoids the simply technicist use of programmed interventions and that considers wider social justice and relational issues is restorative justice. Restorative justice will be discussed in the next section, within the context of how criminology can be applied to respond to victims of crime.

What Is to Be Done on Behalf of the Victims of Crime?

Criminology, and specifically the subdiscipline of victimology, has had a particular influence in the increased importance given to the needs of victims of crime in the criminal justice system. By the end of the 1970s, "many diverse forces had

converged to draw attention to the neglected role and importance of the victim in the justice system" (Strang, 2001: 72). These forces, from within and outside criminology, included (1) the rise of the feminist movement, which highlighted important issues around particular 'hidden' crimes including rape, sexual harassment, domestic violence, and child abuse; (2) efforts to establish children's rights, which increased acknowledgment that children were particularly vulnerable to victimization; (3) the introduction of victimization surveys, which showed that crime was unreported and grossly underestimated, thus highlighting the difficulties and distrust of the justice system felt by many crime victims; (4) the rise of victim compensation initiated; and finally (5) the introduction of legal and social reforms aimed at protecting and helping crime victims (see Goodey, 2004; Spalek, 2006 for a more detailed overview).

As these developments in the victim movement progressed during the 1980s, the rise of consumerism resulted in agencies being directed to care about the lay people using their "services" (Mawby and Walklate, 1994: 81; Simmonds, 2009). Until then, responding to victims involved providing degrees of compensation for physical injury, including medical services and income maintenance or cash benefits, but ignored those needs that arose as a result of mental or emotional harm (Webster, 1994). Initiatives were subsequently devised to provide a service to victim 'consumers,' victim advocacy groups enjoyed a more significant status, and 'service delivery' targets were used to monitor whether or not the structural and systematic problems encountered by victims were being addressed.

While the 'victim movement' and rise of consumerism have certainly marked an important change within criminal justice, the extent to which this can be said to have improved the plight of victims is still subject to considerable debate. The work of two critical scholars may be drawn on here to illustrate the point further. Nils Christie, a Norwegian criminologist, has substantially shaped how we think about who victims are and how they are treated by the criminal justice system. In 1977, Christie suggested that victims had their 'conflict' stolen by the criminal justice system, and in 1986 he argued that not all individuals are able to acquire legitimate victim status – the less 'ideal' a victim was the less 'justice' they received from the criminal justice process.

Later, Robert Elias (1993) contended that victims are often manipulated by governments to achieve punitive aims. This reality can leave the victim feeling bewildered in which justice is equated with severe sentencing as opposed to tangible outcomes for their suffering or losses. Elias (1990) went so far as to say that the victim movement in the United States could no longer be classified as a social movement at all, so completely had it been corrupted by right-wing political forces. However, this cooption can take other forms as well. Spalek (2006), for example, draws attention to how Victim Support in England and Wales has increasingly been seen as an adjunct to the formal justice system because of its success in securing a place at the center of government policy, casting other victims' groups such as rape crisis centers to the margins.

To further illustrate some of these practices, two important developments will now be reviewed: restorative justice, and the legal provisions, which cover the treatment of victims and victim participation within the criminal justice system.

Restorative Justice

Although the recent origins of restorative justice are widely contested (Clamp, 2014), most agree that the inherent failings of criminal justice created an impetus for alternative models to be devised. The emergence of restorative justice around the same time as the apparent 'demise' of the rehabilitative ideal (or the etiological and penal crises) is perhaps unsurprising as it mirrors objections to state monopoly; questions the role of the professional; and seeks to 'rebalance' responses to crime so that the views of the victim, the offender, and the wider community are taken into consideration. As such, it departs from current trends of 'punitive populism' and 'crime control' outlined in the previous sections in preferring humanitarian notions of restoration and reintegration.

Perhaps, the most significant theorist in this area is John Braithwaite who put forward the idea of reintegrative shaming, whereby the actions of the offender are denounced, but not the offender (Braithwaite, 1989). During the process, the offender is encouraged to acknowledge wrongdoing and make amends for his or her actions by responding to the harm caused to the victim, "thus demonstrating that he or she remains part of the law-abiding community and recognises its norms of acceptable behaviour" (Roche, 2003: 29). Reintegration not only places a requirement on the offender, but also on his/her family and other agencies in supporting the offender to overcome the underlying causes of the offending, which may stem from a variety of practical, emotional, and/or psychological need. This work, from Braithwaite and others, is one of the strongest recent examples of criminological insights being applied in practice.

Experimentation with processes that provide both victims and offenders with an active role in the resolution of their offences can be traced back to victim and offender reconciliation programs that originated in Kitchener, Ontario in the late 1970s (Zehr, 1990; Galaway and Hudson, 1996). Until recently, restorative justice has, for the most part, operated on the margins of the criminal justice system offering no real challenge to the dominant law and order ideology that prevails in many Western societies (Garland, 2001). It was not until the 1980s that the increased status of both the 'victim' and the 'community' contributed to a reconsideration of the role of the state in the resolution of offences (Van Ness, 1996; Albrecht, 1999; Zauberman, 2000; Zedner, 2002). At this point, the restorative justice experiment extended to the reform of youth justice systems on the basis of restorative justice processes and principles such as in New Zealand, Northern Ireland, and South Africa (see Clamp, 2008; O'Mahony et al., 2012), experimentation by the police (see Bazemore and Griffiths, 2003; Moore and Forsythe, 1995; Paterson and Clamp, 2012), and various communities (see Clamp and Paterson, 2011).

Adversarial approaches in criminal justice are often perceived as a zero-sum game between victims and offenders; thus restorative processes are increasingly vulnerable to cooption by the system. Restorative justice is susceptible to degenerating into a shaming ceremony in which wrongdoers are not reintegrated, but stigmatized and humiliated and where victims

are used as pawns to bring about change in the offender (McCold, 2002). Restorative justice can also have a tendency to endorse similar stereotypical notions of victimhood as the traditional criminal justice system (Walklate, 2005; Young, 2002). It is an alternative to formal justice, yet relies on established legal definitions of victims and victimization (Pavlich, 2005).

Developments in Victim Policy

International instruments have had considerable influence on the treatment and position of victims within the criminal justice process and the United Nations Declaration of Basic Principles of Justice for Victims of Crime and Abuse of Power (1985) is the most influential of these. The Declaration is not legally binding but can be used as a 'benchmark' to measure domestic policy progress (Van Dijk, 1989: 12), and its principles have become staples of policy discussion. These include access to justice, provision of information, fair treatment, and assistance in both formal and/or informal procedures. This has provided the impetus for a worldwide trend of designing and implementing victim rights charters, serving as official acknowledgment that victims have been marginalized and adversely treated by criminal justice agencies leading to secondary victimization (Spalek, 2006). The introduction of charters is a positive step for victims, but the use of the phrase 'victims' rights' can unduly raise expectations that victims are being empowered in some significant way (Fenwick, 1995).

Victims' rights legislation can communicate the services that victims can expect to receive from agencies of the criminal justice system (Spalek, 2006; Wolhuter et al., 2009) thus encouraging victims to think of themselves as individual consumers (Bennett, 2007). Such an approach reduces the potential for significant systemic change to occur so when victims are unhappy with the service that they receive, they are only entitled to make complaints – there is no mechanism through which they can enforce their rights. Changes in policy and practice that have been implemented across agencies of the criminal justice system under the guise of development of a 'victim perspective' or 'victim rights' have not necessarily led to significant improvements for victims (Spalek, 2006).

Promoting victims' rights might lead to giving victims a more active role in the criminal justice system such as by participating in prosecution or sentencing decisions. Many countries have now introduced victim impact statements, allowing victims to explain the impact of the crime on them. Proponents of this approach argue that victim involvement can increase prosecution efficiency, improve sentencing effectiveness, and increase victim satisfaction (McLeod, 1986; Kelly, 1987; Erez, 1990). However, critics are wary of the use of victim impact statements to legitimize a punitive stance against offenders (Ashworth, 1998) and have highlighted that the process can create unachievable expectations among crime victims regarding their influence over sentencing decisions (Fattah, 1986).

In a similar way as discussed previously with regard to offenders, applied criminology has been able to claim an influence over changes in approaches to victims in the criminal justice system in the recent decades, while also maintaining a critical stance over the meaning and consequences of some of these changes.

Conclusion

Applied criminology is a term now in widespread use in program titles, conference programs, and names of research centers, and the terminology sits alongside the applied iteration of other disciplines such as applied sociology or applied forensic science. The use of the term 'applied criminology,' however, has neither been accompanied by concerted attempts to define the term, nor are there leading scholars who prominently declare themselves as applied criminologists. It is perhaps due to the nature of the criminological discipline that the boundaries around applied criminology are so unclear; in a sense all criminology is practical and applied, just as all criminology is reflective and theoretical.

This article has outlined how criminology has been applied to the problem of what to do about crime, to work with offenders, and to responding to the needs of victims. Punitive strategies remain popular with politicians and the public, but they have little impact on crime rates. Police practice is held to account both by criminological scholars and researchers and increasingly by contributions from criminology to police training. Research insights inform the nature of the work carried out with offenders and contribute both to the nature of the interventions and, more recently, to the way objectives are set and notions of success are framed. Work with victims is more prominent and better funded than it was a few decades ago. However, victims' emotions have been translated into individual need using the discourses of consumerism and active citizenship so, although they are acknowledged, victims' needs are framed and used according to the bureaucratic and political goals that prevail.

Criminologists are concerned about how their work is received by wider society and the weight given to their research findings and insights. Two recent attempts to conceptualize this relationship between criminology and the public provide assistance in determining the nature of applied criminology. Braithwaite (2011: 137) asks criminology scholars to be 'sparkers,' stimulating debate and providing ideas for the players in policy and practice. Loader and Sparks (2010: 124) suggest that the role of the criminologist in policy formation is that of the democratic underlaborer, promoting a better political and policy understanding of crime and the response to it. McNeill (2011) observed that such a role could also be played by criminology in relation to practice and, that, perhaps, is the place of applied criminology – to spark ideas and inform the practice of those who work within criminal justice, particularly in direct work with offenders and victims.

See also: Applied Sociology; Crime and Criminal Justice: Applied Research from Routine Monitoring to Evidence-Based Practices; Crime, Desistance, and Persistence; Crime, Sociology of; Crime: Knowledge about and Prevalence; Criminal Justice, Sociology of; Deterrence Theory: Crime; Mass Imprisonment and Its Consequences; Police, Sociology of; Policing; Prevention of Crime and Delinquency; Re-entry and Reintegration After Incarceration; Social Justice and Criminal Justice.

Bibliography

Albrecht, H., 1999. Countries in transition: effects of political, social and economic change on crime and criminal Justice – sanctions and their implementation. European Journal of Crime, Criminal Law and Criminal Justice 7 (4), 448–479.

Ashworth, A., 1998. Victims' rights, defendants' rights and criminal procedure. In: Paper Presented to the International Conference on Integrating Victim Perspectives in Criminal Justice, in York, 17–18 July 1998.

Bauman, Z., 2001. Community: Seeking Safety in an Insecure World. Polity Press, Cambridge.

Bazemore, G., Griffiths, C., 2003. Police reform, restorative justice and restorative policing. Police Practice and Research 4 (4), 335–346.

Bennett, C., 2007. Satisfying the needs and interests of victims. In: Johnstone, G., Van Ness, D. (Eds.), Handbook of Restorative Justice. Willan Publishing, Collumpton.

BOCSAR, 2013. Official website: http://www.bocsar.nsw.gov.au/bocsar/bocsar_index.html (accessed 03.12.13.).

Bottoms, A., 1995. The philosophy and politics of punishment and sentencing. In: Clarkson, C., Morgan, R. (Eds.), The Politics of Sentencing Reform. Clarendon Press, Oxford.

Bowling, B., Phillips, C., 2002. Racism, Crime and Justice. Longman, London.

Braithwaite, J., 1989. Crime, Shame and Reintegration. Cambridge University Press, Cambridge.

Braithwaite, J., 2011. Opportunities and dangers of capitalist criminology. In: Parmentier, S., Walgrave, L., Aertsen, I., Maesschalck, J., Paoli, L. (Eds.), The Sparking Discipline of Criminology. Leuven University Press, Leuven.

Canton, R., Yates, J., 2008. Applied criminology. In: Stout, B., Yates, J., Williams, B. (Eds.), Applied Criminology. Sage Publications, London.

Carlen, P., 2013. Against rehabilitation: for reparative justice. Criminal Justice Matters 91 (1), 32–33.

Cavadino, M., Dignan, J., 2002. The Penal System: An Introduction. Sage Publications, London.

Cesaroni, C., Doob, A., 2003. The decline in support for penal welfarism: evidence of support among the elite for punitive segregation. British Journal of Criminology 43 (2), 434–441.

Christie, N., 1986. The ideal victim. In: Fattah, E. (Ed.), From Crime Policy to Victim Policy. Macmillan, Basingstoke.

Christie, N., 1977. Conflicts as property. British Journal of Criminology 17 (1), 1–15.

Clamp, K., 2014. Restorative Justice in Transition. Routledge, London.

Clamp, K., 2008. Assessing alternative forms of localised justice in post-conflict societies – youth justice in Northern Ireland and South Africa. In: Frenkel, D., Gerner-Beuerle, C. (Eds.), Selected Essays on Current Legal Issues. ATINER, Athens.

Clamp, K., Paterson, C., 2011. Rebalancing criminal justice: potentials and pitfalls for community neighbourhood panels. British Journal of Community Justice 9 (2), 21–35.

Cohen, S., 1985. Visions of Social Control: Crime, Punishment and Classification. Polity Press, Oxford.

Cohen, S., 1972. Folk Devils and Moral Panics. MacGibbon, London.

Crawford, A., 2008. Refiguring the community and professional in policing and criminal justice: some questions of legitimacy. In: Shapland, J. (Ed.), Justice, Community and Civil Society: A Contested Terrain. Willan Publishing, Devon.

Elias, R., 1993. Victims Still: The Political Manipulation of Crime Victims. Sage Publications, Newbury Park, California.

Elias, R., 1990. Which victim movement? The politics of victim policy. In: Lurigio, A., Skogan, W., Davis, R. (Eds.), Victims and Criminal Justice. Sage Publications, Newbury Park, California.

Erez, E., 1990. Victim participation in sentencing: rhetoric and reality. Journal of Criminal Justice 18 (1), 19–31.

Eterno, J., Silverman, E., 2012. The Crime Numbers Game: Management by Manipulation. CRC Press Inc, New York.

Farrall, S., Calverley, A., 2006. Understanding Desistance from Crime. Open University Press, Maidenhead.

Fattah, E., 1986. From Crime Policy to Victim Policy. Pagrave Macmillan, London.

Feeley, M., Simon, J., 1992. The new penology: notes on the emerging strategy of corrections and its implications. Criminology 30 (4), 449–474.

Fenwick, H., 1995. Rights of victims in the criminal justice system: rhetoric or reality? Criminal Law Review 843–853.

Galaway, B., Hudson, J., 1996. Restorative Justice: International Perspectives. Criminal Justice Press, Monsey, NY.

Garland, D., 2001. The Culture of Control. Oxford University Press, Oxford.

Goodey, J., 2004. Victims and Victimology: Research, Policy and Practice. Pearson Education, Canada.

Hobsbawm, E., 1995. Age of Extremes: The Short Twentieth Century, 1914–1991. Abacus, London.

Kelly, D., 1987. Victims. Wayne Law Review 34, 69–86.

Lipton, D., Martinson, R., Wilks, J., 1975. The Effectiveness of Correctional Treatment: A Survey of Treatment Valuation Studies. Praeger, New York.

Loader, I., Sparks, R., 2010. Public Criminology. Routledge, London.

McCold, P., 2002. Evaluation of a restorative milieu: CSF Buxmont School/Day Treatment programs 1999–2001 evaluation outcome technical report. In: Paper Presented at the American Society of Criminology Annual Meeting, Chicago, Illinois. http://www.realjustice.org/library/erm.html (accessed 28.11.06.).

McEvoy, K., 2007. Beyond legalism: towards a thicker understanding of transitional Justice. Journal of Law and Society 34 (4), 411–440.

McLeod, M., 1986. Victim participation at sentencing. Criminal Law Bulletin 22, 501–517.

McNeill, F., 2011. Review: public criminology? Ian loader and Richard Sparks. Probation Journal 58 (1), 84–87.

McNeill, F., 2006. A desistance paradigm for offender management. Criminology and Criminal Justice 6, 37–60.

Mackenzie, D.L., 2000. Evidence-based corrections: identifying what works. Crime and Delinquency 46 (4), 457–471.

Martinson, R., 1979. New findings, new views: a note of caution regarding sentence reform. Hostra Law Review 7, 242–258.

Maruna, S., 2001. Making Good: How Ex-Convicts Reform and Re-build Their Lives. American Psychological Society, Washington, DC.

Mawby, R., Walklate, S., 1994. Critical Victimology. Sage, Thousand Oaks, CA.

Miller, J., 1989. Does nothing work? In: Priestley, P., Vanstone, M. (Eds.), Offenders or Citizens?: Readings in Rehabilitation. Willan, Cullompton.

Moore, D., Forsythe, L., 1995. A New Approach to Juvenile Justice: An Evaluation of Family Conferencing in Wagga Wagga: A Report to the Criminology Research Council. Centre for Rural Social Research, Charles Sturt University-Riverina, Wagga Wagga, AUS.

Muncie, J., 1998. Deconstructing criminology. Criminal Justice Matters 34 (1), 4–5.

O'Mahony, D., Doak, J., Clamp, K., 2012. The politics of youth justice reform in post-conflict societies: mainstreaming restorative justice in Northern Ireland and South Africa. Northern Ireland Legal Quarterly 63 (2), 269–290.

Paterson, C., Clamp, K., 2012. Exploring recent developments in restorative policing in England and Wales. Criminology and Criminal Justice 12 (5), 593–611.

Pavlich, G., 2005. The Governing Paradoxes of Restorative Justice. Glasshouse Press, London.

Roach, K., 2005. Enhancing criminal justice reform, including restorative justice. In: Paper Presented at the 11th United Nations Congress on 'Crime Prevention and Criminal Justice Synergies and Responses: Strategic Alliances in Crime Prevention and Criminal Justice', Bangkok, Thailand April 18–25, 2005 [Online]. Available from: www.icclr.law.ubc.ca/Publications/Reports/11_un/KENT%20ROACH%20Speaking%20Notes.pdf (accessed 19.03.09.).

Roberts, C., 2004. Offending behaviour programmes: emerging evidence and implications for practice. In: Burnett, R., Roberts, C. (Eds.), What Works in Probation and Youth Justice. Willan, Cullompton.

Roche, D., 2003. Accountability in Restorative Justice. Oxford University Press, Oxford.

Rotman, E., 1986. Do criminal offenders have a constitutional right to rehabilitation? The Journal of Criminal Law and Criminology 77 (4), 1023–1068.

Sampson, R., Laub, J., 1993. Crime in the Making: Pathways and Turning Points through Life. Harvard University Press, Cambridge, MA.

Scraton, P., 2009. Hillsborough: The Truth. Mainstream Publishing, Edinburgh and London.

Simmonds, L., 2009. What victims want! Victim support, an objective or relative: approach to victims' needs. Social and Public Policy Review 3 (2), 11–29.

Spalek, B., 2006. Crime Victims: Theory, Policy and Practice. Palgrave Macmillan, Basingstoke.

Stout, B., Yates, J., Williams, B., 2008. Applied Criminology. Sage Publications, London.

Strang, H., 2001. The crime victim movement as a force in civil society. In: Strang, H., Braithwaite, J. (Eds.), Restorative Justice and Civil Society. Cambridge University Press, Cambridge.

Van Dijk, J., 1989. Benchmarking legislation on crime victims: the UN victims declaration of 1985. In: Jounsen, M., Shapland, J. (Eds.), Changing Victim Policy: The United Nations Declaration and Recent Developments in Europe. HEUNI, Helsinki.

Van Ness, D., 1996. Restorative justice and international human rights. In: Galaway, B., Hudson, J. (Eds.), Restorative Justice: International Perspectives. Kugler Publications, Amsterdam.

Walklate, S., 2005. Victimhood as a source of oppression. Social Justice 32, 88–99.
Webster, B., 1994. Government strategies to co-ordinate victims' policies and services within a crime prevention framework in New Zealand. In: Paper Presented at 8th International Symposium on Victimology, Adelaide.
Wolhuter, L., Olley, N., Denham, D., 2009. Victimology: Victimisation and Victims Rights. Routledge, London.
Young, R., 2002. Testing the limits of restorative justice: the case of corporate victims. In: Hoyle, C., Young, R. (Eds.), New Visions of Crime Victims. Hart Publishing, Portland, OR.
Young, J., 1988. Radical criminology in Britain: the emergence of a competing paradigm. British Journal of Criminology 28 (2), 159–183.
Young, R., Matthews, J., 2003. The New Politics of Crime and Punishment. Willan Publishing, Collumpton.
Zauberman, R., 2000. Victims as consumers of the criminal justice system? In: Crawford, A., Goodey, J. (Eds.), Integrating a Victim Perspective within Criminal Justice: International Debates. Dartmouth Publishing Company Ltd, Aldershot.
Zedner, L., 2002. Victims. In: Maguire, M., Morgan, R., Reiner, R. (Eds.), The Oxford Handbook of Criminology. Oxford University Press, Oxford.
Zehr, H., 1990. Changing Lenses. Herald Press, Pennsylvania.

Applied Demography

David A Swanson, University of California Riverside, Riverside, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The field of demography can be divided into two general areas, basic or academic demography and applied demography. In this article, I differentiate the latter from the former and go on to discuss characteristics of applied demography in general as well as those found in the practice of applied demography in the private sector and in the public sector. Because demography generally uses population data gathered under the *De Jure* concept, the article also discusses the importance of *De Facto* population information to the private and public sectors. Professional recognition of applied demography and training are discussed. In addition to references for citations in the text, a set of suggested readings is provided.

Introduction

Population estimates and projections are the heart and soul of applied demography and people have been making them for a long time.[1] In 1741, for example, Johann Süssmilch made the first of several estimates of world population (Swanson and Stephan, 2004) and in 1755, Benjamin Franklin was publishing projections of the population of the world (Swanson and Stephan, 2004). It was not until the middle part of the twentieth century, however, that these efforts became institutionalized within the private and public sectors.

Since applied demography is part of the larger field of demography, it is useful to start by defining what demography is so that we can put the discussion of applied demography into the correct conceptual context. Demography can be defined as the study of population, typically focused on five aspects: (1) size, (2) geographic distribution, (3) composition, (4) the components of change (births, deaths, migration), and (5) the determinants and consequences of population change (Swanson and Stephan, 2004, p. 758).

As the preceding definition suggests, demographers share conceptual frameworks, methods, and data in common. For example, a demographer who specializes in studying the geographic distribution of populations typically deals with population size, composition, the components of population change, and the determinants and consequences of population change. As such, this demographer would be using conceptual frameworks, methods, and data that would be familiar to other demographers, regardless of their specialties.

In spite of the shared commonalities, the work done by demographers can be distinguished in several ways. A distinction of interest here is that between basic (academic) demography and applied demography. Basic demography is primarily concerned with solving endogenously defined problems by offering convincing explanations of demographic phenomena, while viewing time and resources as barriers to surmount in order to maximize precision and explanatory power (Swanson et al., 1996).

Applied demography is primarily concerned with solving exogenously defined problems by producing the information necessary to effect practical decision-making, while minimizing the time and resources needed to produce this information (Swanson et al., 1996). Because basic demographers are largely working on problems defined by demographers, they tend to look at both the determinants and consequences of population change. However, applied demographers tend to focus on the consequences of population change because they usually are involved in answering questions posed by nondemographers.

Both basic and applied demography can, themselves, also be subdivided into two subsets. Basic demography can be viewed as being composed of (1) formal demography and (2) population studies. Formal demography can be viewed as mathematics applied to human populations, while population studies can be viewed as the application of sociological concepts and methods to human populations. There is overlap between formal demography and population studies in terms of concepts, methods, and data, but the distinction, nonetheless, is useful.

Applied demography's two distinct subsets are private sector demography and public sector demography. Both use the same methods, some of which are unique to applied demography, but also include some of the methods found in formal demography and many of the methods found in population studies. Private and public sector demography both use a great deal of data generated in the public domain, but because private sector data are too costly or proprietary, little of its data is even known within public sector demography, much less used by it.

As presently practiced around the world, applied demography largely owes its present configuration to two major factors, largely initiated in the United States: (1) the need by government to have data to develop policies and implement them and (2) the need by business to have data to guide its investment, production, and marketing efforts.

Although governments have long used census data, it was the advent of US government policy-making under the administration of President Franklin Roosevelt in the 1930s that set the stage for the need for data beyond the census, which, in turn, kicked off the move by governments at all levels to generate and acquire data for policy and administrative needs, both within and outside the United States (Prewitt, 2013). The utility of low-cost, timely, and useful data being generated by the US government was quickly recognized by the private sector. The private sector demand for such data soon outpaced what could be provided by the government and by 1970, statistics became a commodity independent of government and a statistical services industry developed, an industry in which

'demographics' found itself playing a key role (Martins et al., 2011).

Applied Demography in the Private Sector

There are multiple aspects to the statistical services industry and while there are often less-than-clear-cut distinctions among them, several can be identified, including econometrics, database development and maintenance, survey research, and demographics. Broadly defined, the latter can cover a wide range of activities, including actuarial work, public health, and market demographics. Of interest to us here is market demographics, which makes up the lion's share of the 'demographics' found in the private sector.

There are three major components of private sector demography: (1) the aggregation of personal demographic characteristics by geographic areas such as postal zip codes, blocks, block groups, and client-defined areas; (2) psychographic profiles of individuals aggregated into 'lifestyle' segments; and (3) the aggregation of household characteristics by geographic areas.

The personal demographic characteristics component sprang directly from two factors: (1) the growing interest on the part of major corporations in market segmentation and targeted marketing, which emerged as mass marketing methods began to falter and (2) technology developed within the public sector that enabled the association of demographic characteristics with addresses. Although it came to prominence in the 1970s, following the public release of the first-ever electronic files for census data from the 1970 census, the origin of this component can be traced back to 1950, with the founding of Market Statistics, Inc. (Pentilla, 2005; Morrison and Smith, 2005; Russell, 1984).

The psychographic component also came to prominence in the 1970s, but its origin is found in a 1960 report, 'Consumer Values and Demands,' issued by the Stanford Research Institute (Mitchell, 1983). Ideas from the report ultimately emerged as the 'Values and American Lifestyles' (VALS) project. In 1978, a full-scale VALS project sponsored by 39 corporations (Mitchell, 1983) bore fruit in 1983 in the form of the book by Arnold Mitchell, *The Nine American Lifestyles*.

As the first two components matured, they started feeding into one another. However, even when supplemented by psychographic profiles, 'zip code' demographics could not cover all marketing needs. One of these needs was the ability to segment and target down to the level of the individual household. As was the case for the zip code and psychographic providers, this need led to the third component, 'household characteristics.' Like the other two components, its implementation also required a substantial leap forward in electronic data storage and manipulation.

The origins of the third component, 'household characteristics,' can be traced to precomputer times, but these attempts were very limited. However, these attempts led to the assembly of credit and other data on households held by private companies with the goal of having 100% coverage of all households in the United States. By the early 1970s, these privately held household records approached the same number of households counted in the 1970 US census and have not lost ground since.

To update their records, household providers still rely on credit reporting services, but their efforts have extended to many other sources of data, including purchase data, supermarket savings cards, white pages, telephone usage, surveys, sweepstakes and contest entries, medical records, dental records, insurance policies, financial records, property records, motor vehicle data, automatic number information, credit card transactions, phone records, credit records, product warranty cards, the sale of magazine and catalog subscriptions, customer loyalty cards – and the United States Postal Service (EPIC, n.d.; Tripod, n.d.; US Postal Service, 2011).

While there are only a few companies in the credit bureau business, there are many small companies in the business of selling customized estimates, projections, and other types of analyses. Many companies define their market areas by postal delivery areas and while some national statistical agencies (e.g., the US Census Bureau, Statistics Canada) provide census data for these pieces of geography, they do not routinely produce estimates and projections for them. This is one of the reasons for the existence of the substantial market demographics industry.

While the private sector provides characteristics of interest (e.g., age, race, sex, recent purchases, income), it still relies on data from the public sector. The US Postal Service, for example, is a major source of information on changes of address (EPIC, n.d.; Pitney Bowes, 2008). This is important in the United States, where around 45 million people (about 15%) of the population aged 1 year and over is estimated to have moved in the past year (US Census Bureau, 2009).

The information on changes in address is provided by the National Change of Address Service (NCOA), which is administered by the US Postal Service. Groups that subscribe to the NCOA can obtain updates when a current customer makes a permanent change of address request to the Postal Service. This is the one of the primary methods by which companies obtain individuals' addresses after they move. Businesses can purchase the updates for as little as $5 per 1000 names.

As the preceding suggests, the private sector statistics industry in the United States relies on the support of the public sector. While the US Postal Service provides the private sector with individual-level information in the form of address changes, the aggregated information provided by the US Census Bureau provides benchmarks against which the private sector data can be compared and adjusted if needed. For example, if a private data vendor finds that its estimate of median household income for a given area is less than that of the Census Bureau's estimate, the vendor can adjust its modeling and other processes to match it.

Applied Demography in the Public Sector

Although there is overlap, the origin of applied demography in the public sector precedes its establishment in the private sector. In 1943, the state of Washington established the Washington State Census Board, which in of itself was not new since many states had conducted their own census counts in the late-nineteenth and early twentieth centuries. However, the act (House Bill 72, Chapter 81, Revised Code of Washington) establishing the Board set the stage for the generation of annual

population estimates (something that was, in fact, new) rather than census counts as a way to equitably allocate funds to incorporated cities struggling under the impact of huge population increases resulting from the entrance of the United States into World War II in 1941. The Board immediately started producing estimates (Washington State Census Board, 1944) and the process of developing annual population estimates for incorporated towns was soon on the path to being institutionalized at the University of Washington under the direction of Professor Calvin F. Schmid, who was selected as one of the Board's initial members.

By the 1950s, the Washington State Census Board was not only producing annual population estimates of the state's incorporated towns, but also population and enrollment projections. The utility of these activities was not missed by other states (and Canadian provinces) and California, Oregon, Alberta, British Columbia, among other subnational governments started establishing their own demographic units. This process was greatly assisted by the fact that many of the graduate students who worked with the Washington State Census Board at the University of Washington had moved into positions in these adjoining areas after completing their degrees. The state of Washington ultimately absorbed the Census Board directly into state government and today its legacy lives on with the forecasting unit of the Office of Financial Management, which is the Governor's Budget Office.

By the late-1940s, the US Census Bureau also started producing annual estimates for states and other subnational areas and by 1950s, it was producing population projections on a regular basis. Since both the federal government and several states were producing population estimates for the same areas, it should not be surprising that in the early 1950s, an informal movement took hold that was aimed at coordinating these activities.

In 1966, the National Governor's Conference, in cooperation with the Council of State Governments, initiated and sponsored the 'First National Conference on Comparative Statistics,' held in Washington, DC. This conference gave national recognition to the increasing demand for subnational population estimates. Between 1967 and 1973, a group of Census Bureau and state employees charged with developing annual subnational population estimates, formalized the 'FSCPE,' the Federal-State Cooperative Program for Local Population Estimates (US Census Bureau, n.d.). Jumping off a task force formed within the FSCPE, a little more than 10 years later, a similar organization (The Federal-State Cooperative Program for Population Projections) was established to coordinate population projection efforts.

Although many states reduced or eliminated their demographic activities between the heyday of these activities in the 1970s and the severe economic recession of 2007–12, the original core states of Washington, Oregon, and California still have highly functional and well-respected demographic operations, as do the states of Alaska, New Mexico, Texas, and Florida, and the provinces of Alberta, British Columbia, and Ontario, among others.

Demographic activities in public sector also are found within sub-state governmental units. Typically, these are in the planning departments of large cities and regional organizations such as Metropolitan Planning Organizations (also known as Councils of Government). The regional organizations often run very sophisticated demographic operations. One notable example is the San Diego Association of Governments, which hosts an annual research conference for demographers working in these regional organizations.

De facto Population Estimates

By now, you may have realized that the discussion of both private and public sector activities has been in terms of the definition of population used by the US Census Bureau, namely, the *'de jure'* population. This is the idea that in a census everybody should be counted once, only once, and in the right place (Cork and Voss, 2006). It is based on place of 'usual residence.' However, important to both the private and public sectors is another definition, namely the *'de facto'* population. This is the idea that in a census you count everybody once, only once, but in the place where they are found at the time of the census (Swanson and Tayman, 2011). Examples of *de facto* populations are many. They include vacationers (of interest, for example, to the hotel owners in Cancun, the casino industry in Macao, and the Hawaii Visitors Bureau), migratory workers (of interest, for example, to health care, school, and other social service providers), and the people who work in the central business district of a large city each day, but leave it largely vacant in the evenings (of interest to the San Francisco City Planning Office, for example).

While estimates of *de facto* populations are of great interest, they are difficult to make in countries that use the *de jure* concept of residency for their census counts. In addition, they are also difficult to evaluate in these countries because there is no 'gold standard' against which to measure them in terms of validity and reliability (Swanson and Tayman, 2011).

Professional Recognition[2]

As you have read, by the end of the 1970s, applied demographers in the public sector had a history of interaction and were involved in formal organizations that facilitated the advancement of the field. The federal-state cooperative programs for population estimates and projections, respectively, were instrumental in this regard. Demographers in the private sector were more isolated from one another and their colleagues in the public sector. Partly, this was due to the proprietary nature of their work and partly to the slow pace of hiring demographers on the part of major companies.

One affiliation that was shared between applied demographers in the private sector and their colleagues in the public sector changed that isolation in the late-1970s. The organization was the Population Association of American (PAA), arguably the world's largest single professional association for demographers of all stripes.

Founded in the 1930s, the PAA was almost entirely made up of basic demographers until the 1960s, when applied demographers started showing up at its conferences. By 1975, Harry Rosenberg, who was moving from the University of North Carolina to the US Center for Health Statistics, wrote to the

then-president-elect, Sidney Goldstein, of PAA urging him to include a session in the 1976 conference on state and local demography. While the letter came too late to have a session on state and local demography, it resulted in the first of an annual set of 'breakfast meetings.' By 1977, a formal session on state and local demography was included in the PAA's annual conference, as was the case in 1978, which also saw the first appearance of the official PAA Committee on State and Local Demography.

With annual breakfast and regular sessions, the 'state and local demographers' were soon joined by their colleagues in the private sector and by the early 1980s, a Committee on Business Demography was spun-off by the Committee on State and Local Demography. The two committees then joined forces to launch the 'Applied Demography' newsletter under the auspices of the PAA. By 1993, the two separate committees merged to form the PAA's official Committee on Applied Demography, which continues to sponsor not only the annual breakfast, but also an evening reception and several sessions in each of the annual PAA conferences. In addition to about 125 members who specialize in estimates and projections, the PAA has around a 100 members who specialize in business or state and local demography.

A separate track to professional recognition was provided by the Southern Demographic Association (SDA), which from its founding in 1970 (as the Southern Demographic Interest Group) was focused on applied demography. This organization was initially sponsored by the Oak Ridge National Laboratory (located near Knoxville, Tennessee) and its associated universities. Contact was made between Oak Ridge and Professor Everett Lee at the University of Georgia in regard to developing current population estimates that could be used for emergency planning, especially the evacuation of people from large population centers in the event of nuclear war. Like the PAA, this group held an annual conference at which demographers congregated and it quickly attracted the interest of applied demographers located in or near the southern region of the United States. Over time, this regional organization has grown to include members from outside the southern region of the United States.

The SDA began publishing upon its founding and eventually acquired a journal, *Population Research and Policy Review*, which became a publication outlet for many applied demographers. In addition to its journal, the SDA sponsors the Walter E. Terrie award, which is given to the author(s) of the best paper on applied demography presented at its annual meeting. This is the only award given anywhere in the world specifically for a contribution to applied demography.

Perhaps, the apex of professional activity and recognition is found in the form of the biennial conference on applied demography, which is sponsored and hosted by the University of Texas, San Antonio every other January since 2008. This conference filled a 15-year hiatus that resulted when the applied demography conferences started at Bowling Green State University in 1986 were discontinued in the 1990s. Selected papers from this conference are published in a proceedings book by Springer, which is edited by faculty members in the Department of Demography and staff at the Texas State Demographic Center, both of which are housed within the University.

Training

Training in applied demography has both a formal education component and an on-the-job component. For the most part, applied demographers have a bachelor's degree in one of the social sciences, typically sociology and geography. Many universities offer courses in applied demography, but at the undergraduate level there are neither formal degrees nor concentrations offered.

At the graduate level, however, there are several programs in addition to many individual courses. The University of California Irvine offers a 1-year program in applied demography that leads to a Master's Degree. A similar program is offered by Florida State University, while The University of Texas, San Antonio offers a PhD in applied demography. An entrance requirement to this program is possession of a Master's degree, but it need not be in applied demography. The graduate programs typically include courses on data sources, geographic information systems, and methods for estimating and projecting populations. Information on the courses and other requirements of these programs can be found at the Web site of the universities where the programs are offered.

Summary

The development of applied demography has been driven by the interaction of several elements that have shaped the nature of certain streams of research in demography. The first is the rise in the use of the demographic perspective, data and methods in addressing business opportunities and problems. Encouraged in part by the need for small area data, many demographers more recently have focused their efforts on the improvement of techniques and technology needed to generate, display, and analyze small area data, both for the private and the public sectors. While the vast majority of demographic research focuses on the demography of geographic units, it is worthwhile to note that an interest in organizational demography and other nongeography-based applications has emerged in the last decade or so.

The development of data through estimation and projection (particularly for small areas) has and will continue to be affected by continued improvements in methods and technology. Although the focus here has been on applied demography in the United States, developments in methodology and technology transcend national boundaries. These developments will, in turn, influence both applied demography not only in the United States, but elsewhere.

End Notes

1. A population estimate is the calculation of a current or past population, typically based on symptomatic indicators of population change, while a population projection is the numerical outcome of a particular set of assumptions regarding future population trends (Swanson and Stephan, 2004). It is worthwhile to add here that there is a distinction usually made between the term 'projection' and the term 'forecast,' with the latter being defined as

a population projection deemed most accurate for the purpose of predicting future population (Swanson and Stephan, 2004).

It is also useful to note here that there are two distinct traditions in regard to population estimates: (1) demographic and (2) statistical – that is, the methods used by those who do sample surveys. Demographic methods are used to develop estimates of a total population as well as the ascribed characteristics – age, race, and sex – of a given population. Statistical methods typically employ sample surveys in order to estimate the achieved characteristics of a population – educational attainment, employment status, income, and martial status, for example. Among survey statisticians, the demographer's definition of an estimate is generally termed an 'indirect estimate' because unlike a sample survey, the data used to construct a demographic estimate are symptomatic indicators of population change (e.g., K-12 enrollment data, births, deaths) and do not directly represent the phenomenon of interest. Among demographers, however, the term 'indirect estimate' has a different meaning (Swanson and Stephan, 2004). Most national statistical agencies produce estimates using both traditions, demographic and statistical. With regard to population projections, the distinction between the demographic tradition and the statistical tradition is less pronounced, however.

2. The information about the PAA's Committee on Applied Demography is largely drawn from Voss (2005).

See also: Applied Social Research, History of; Censuses: Current Approaches and Methods; Censuses: History and Methods; Data Bases and Statistical Systems: Applied Social Research; Demography, Early History of; Forecasting; Immigration, Applied Research on.

Bibliography

Bogue, D., June 1950. A technique for making extensive population estimates. Journal of the American Statistical Association 45, 149–163.

Bongaarts, J., Bulatao, R. (Eds.), 2000. Beyond Six Billion: Forecasting the World's Population. National Academy Press, Washington, DC.

Bryan, T., 2004. Population estimates. In: Siegel, J., Swanson, D. (Eds.), The Methods and Materials of Demography, second ed. Elsevier Academic Press, New York City, NY, pp. 523–560.

Cameron, M., Poot, J., 2011. Lessons from stochastic small-area population projections: the case of Waikato subregions in New Zealand. Journal of Population Research 28, 245–265.

Cannan, E., 1895. The probability of a cessation of growth in England and Wales during the next century. The Economic Journal 5 (2), 505–515.

Carroll, G., Hannan, M., 2004. The Demography of Corporations and Industries. Princeton University Press, Princeton, NJ.

Cork, D., Voss, P., 2006. Once, Only Once, and in the Right Place: Residence Rules in the Decennial Census. The National Academies Press, Washington, DC.

De Gans, H., 1999. Population Forecasting 1895–1945: The Transition to Modernity. Kluwer Academic Publishers, Dodrecht, The Netherlands.

EPIC., n.d., Privacy and Consumer Profiling. Available online at: http://www.scribd.com/doc/20289541/EPIC-Information-on-Consumer-Profiling.

Gauthier, A., Chu, C.Y.C., Tuljapurkar, S. (Eds.), 2006. Allocating public and Private Resources across Generations: Riding the Age Waves, vol. 2. Springer, Dordrecht, The Netherlands.

George, M., Smith, S., Swanson, D., Tayman, J., 2004. Population projections. In: Siegel, J., Swanson, D. (Eds.), The Methods and Materials of Demography, second ed. Academic/Elsevier Press, Los Angeles, pp. 561–601.

Hamilton, C.H., Perry, J., 1962. A short method for projecting population by age from one decennial census to another. Social Forces 41, 163–170.

Hardy, G., Wyatt, F., 1911. Report of the actuaries in relation to the scheme of insurance against sickness, disablement, etc. Journal of the Institute of Actuaries XLV, 406–443.

Hollmann, F., Mulder, T., Kallan, J., 13 January 2000. Methodology and assumptions for the population projections of the United States: 1999 to 2100. In: Population Division Working Paper No. 38. U.S. Bureau of the Census, Washington, DC.

Hoque, N., Swanson, D. (Eds.), 2012. Opportunities and Challenges for Applied Demography in the 21st Century. Springer B.V. Press, Dordrecht, Heidelberg, London New York.

Judson, D., Swanson, D., 2011. Estimating characteristics of the foreign-born by legal status: an evaluation of data and methods. In: Springer Briefs in Population Studies, vol. 2. Springer, B.V. Press, Dordrecht, Heidelberg, London New York.

Kintner, H., Swanson, D., 1996. Ties that bind: a case study of the link between employers, families and health benefits. Population Research and Policy Review 15 (5–6), 509–523.

Kintner, H., Merrick, T., Morrison, P., Voss, P., 1994. Demographics: A Casebook for Business and Government. Westview Press, Boulder, CO.

Long, J., 1993. An overview of population estimation methods. In: Bogue, D., Arriaga, E., Anderton, D. (Eds.), Readings in Population Research Methodology, vol. 5. United Nations Population Funds and Social Development Center, Chicago, pp. 20-1–20-5.

Martins, J., Yusuf, F., Swanson, D., 2011. An Introduction to Consumer Demographics and Behaviour: Markets Are People. Springer B.V. Press, Dordrecht, Heidelberg, London New York.

McVey, W., 1974. An Empirical Assessment of a Modified Censal Ratio Estimation Technique. Ph.D. dissertation. Department of Sociology, University of Alberta, Edmonton, Alberta, Canada.

Mitchell, A., 1983. The Nine American Lifestyles. MacMillan, New York, NY.

Moen, E., 1984. Voodoo forecasting: technical, political, and ethical issues regarding the projection of local population growth. Population Research and Policy Review 3, 1–25.

Morrison, P.A., Smith, S., 2005. Small-area and business demography. In: Micklin, M., Poston, D. (Eds.), Handbook of Population. Kluwer Academic/Plenum Publishers, New York, NY, pp. 761–786.

Murdock, S., Ellis, D., 1991. Applied Demography. Westview Press, Boulder, CO.

Murdock, S., Swanson, D. (Eds.), 2008. Applied Demography in the 21st Century. Springer B.V. Press, Dordrecht, The Netherlands.

Murdock, S., Hoque, M.N., Michael, M., White, S., Pecotte, B., 1997. The Texas Challenge: Population Change and the Future of Texas. Texas A&M University Press, College Station, TX.

Myers, D. (Ed.), 1990. Housing Demography: Linking Demographic Structure and Housing Markets. University of Wisconsin Press, Madison, WI.

Pentilla, C., May 2005. A Brief History of Demographics. Entrepreneur Magazine. http://www.entrepreneur.com/article/77152-4#.

Pitney Bowes, 2008. Pitney Bowes Group 1® Software Unveils Aura™, Industry's Most Comprehensive Address Automation Solution. Available online at: http://www.pbinsight.com/about/newsroom/press-releases/detail/2466_pitney-bowes-group-1reg-software-unveils-auratrade/.

Pittenger, D., 1978. The role of judgment, assumptions, techniques, and confidence limits in forecasting population. Socio-Economic Planning Sciences 12, 271–276.

Plane, D., Rogerson, P., 1994. The Geographical Analysis of Population, with Applications to Planning and Business. John Wiley and Sons, New York City, NY.

Pol, L., Thomas, R., 1997. Demography for Business Decision Making. Praeger, New York, NY.

Pol, L., Thomas, R., 2000. The Demography of Health and Health Care, second ed. Springer B.V. Press, Dordrecht, The Netherlands.

Popoff, C., Judson, D., 2004. Some methods of estimation for statistically underdeveloped areas. In: Siegel, J., Swanson, D. (Eds.), The Methods and Materials of Demography, second ed. Amsterdam, pp. 603–641.

Preston, S., 1993. The contours of demography: estimates and projections. Demography 30 (4), 593–606.

Prewitt, K., 2013. The 2012 Morris Hansen lecture: thank you, Morris, et al., for Westat, et al. Journal of Official Statistics 29 (2), 223–231.

Purcell, N., Kish, L., 1980. Post-censal estimates for local areas (or domains). International Statistical Review 48 (1), 3–18.

Rives, N.W., Serow, W., Lee, A., Goldsmith, H., Voss, P. (Eds.), Basic Methods for Preparing Small-area Population Estimates. Applied Population Laboratory, University of Wisconsin, Madison, WI.

Russell, C., 1984. The business of demographics. Population Bulletin 39 (3). Population Reference Bureau, Washington, DC.

Schmitt, R., Crosetti, A., 1954. Accuracy of the ratio-correlation method for estimating post-censal population. Land Economics 30, 279–281.

Siegel, J., 2002. Applied Demography: Applications to Business, Government, Law, and Public Policy. Academic Press, San Diego, CA.

Simpson, S., Diamond, I., Tonkin, P., Tye, R., 1996. Updating small area population estimates in England and Wales. Journal of the Royal Statistical Society, Series A 159 (Part 2), 235–247.

Simpson, S., Middleton, E., Diamond, I., Lunn, D., 1997. Small-area population estimates: a review of methods used in Britain in the 1990s. International Journal of Population Geography 3 (3), 265–280.

Singh, M., Hidiroglou, M., Gambino, J., Kovacević, M., 2001. Estimation methods and related systems at statistics Canada. International Statistical Review 69 (3), 461–485.

Smith, S.K., 1989. Toward a methodology for estimating temporary residents. Journal of the American Statistical Association 84, 430–436.

Smith, S., Tayman, J., Swanson, D., 2013. A Practitioner's Guide to State and Local Population Projections. Springer B.V. Press, Dordrecht, Heidelberg, London New York.

Snow, E.C., 1911. The application of the method of multiple correlation to the estimation of post-censal populations. Journal of the Royal Statistical Society 74 (Part 6), 575–620.

Statistics Canada, 1987. Population Estimation Methods, Canada. Ministry of Supply and Services, Ottawa, Ontario, Canada.

Swanson, D., August 2004. Advancing methodological knowledge within state and local demography: a case study. Population Research and Policy Review 23, 379–398.

Swanson, D., 2008. Population Estimation: An Annotated Bibliography and Glossary. Working Paper 8–1. Center for Sustainable Suburban Development, University of California Riverside. http://cssd.ucr.edu/Papers/PDFs/1-Population%20Estimation.pdf.

Swanson, D., 2009. Hurricane Katrina: a case study of its impacts on medical service providers and their client populations. The Open Demography Journal 2, 8–17.

Swanson, D., 2013. Consumer demographics: welcome to the dark side of statistics. Radical Statistics 108, 38–46.

Swanson, D., Burch, T., Tedrow, L., 1996. What is applied demography? Population Research and Policy Review 15 (5–6), 403–418.

Swanson, D., Forgette, R., McKibben, J., Van Boening, M., Wombold, L., 2009. The socio-demographic and environmental effects of katrina: an impact analysis perspective. The Open Demography Journal 2 (11), 36–46.

Swanson, D., Hough, G., Clemans, C., Rodriguez, J., 1998. Merging methods and judgment for K-12 enrollment forecasting. Educational Research Service Spectrum 16 (Fall), 24–31.

Swanson, D., Hough, G., Lycan, D., Sharkova, I., 1997. Forecasting for school attendance zone changes: merging GIS and demographic methods with local expert judgment. Market and Demographic Analysis 3 (6), 23–31 (in Chinese).

Swanson, D., Hough, G., 2012. An evaluation of persons per household (PPH) data generated by the American community survey: a demographic perspective. Population Research and Policy Review 31, 235–266.

Swanson, D., McKibben, J., 2010. New directions in the development of population estimates in the United States? Population Research and Policy Review 29 (6), 797–818 (with McKibben J.).

Swanson, D., McKibben, J., 1997. Linking substance and practice: a case study of the relationship between socio-economic structure and population estimation. Journal of Economic and Social Measurement 24, 135–147.

Swanson, D., Morrison, P., 2010. Teaching business demography using case studies. Population Research and Policy Review 29 (1), 93–104.

Swanson, D., Pol, L., 2005. Contemporary developments in applied demography within the United States. Journal of Applied Sociology 21 (2), 26–56.

Swanson, D., Stephan, G.E., 2004. Glossary and a demography timeline. In: Seigel, J., Swanson, D. (Eds.), The Methods and Materials of Demography, second ed. Elsevier Academic Press, San Diego, CA, pp. 751–786.

Swanson, D., Tayman, J., 2011. On estimating a De Facto population and its components. Review of Economics and Finance 5, 17–31.

Swanson, D., Tayman, J., 2012. Subnational Population Estimates. Springer B.V. Press, Dordrecht, Heidelberg; London, and New York.

Swanson, D., Walashek, P., 2011. CEMAF as a census method: a proposal for a re-designed census and an independent census bureau. In: Springer Briefs in Population Studies, vol. 1. Springer, B.V. Press, Dordrecht, Heidelberg, London New York.

Swanson, D., Yusuf, F., 2010. Business demography in the 21st century. Population Research and Policy Review 29 (1), 1–3.

Swanson, D., Schlottmann, A., Schmidt, R., 2010. Forecasting the population of census tracts by age and sex: an example of the Hamilton-Perry method in action. Population Research and Policy Review 29 (1), 47–63.

Tripod. n.d. Experian (Formerly TRW) Requesting all Insider Information (ASAP). Available online at: http://nobillsman.tripod.com/experianmain.html.

U.S. Census Bureau. n.d. FSCPE Overview. https://www.census.gov/popest/fscpe/overview.html.

U.S. Census Bureau, 2009. Geographic Mobility by Selected Characteristics in the United States. S070. American Community Survey. Available online at: http://factfinder.census.gov/servlet/STTable?_bm=y&-qr_name=ACS_2009_1YR_G00_S0701&-geo_id=01000US&-ds_name=ACS_2009_1YR_G00_&-_lang=en&-format=&-CONTEXT=st.

U.S. Postal Service (USPS), 2011. NCOALink® Systems. Available online at: http://www.usps.com/ncsc/addressservices/moveupdate/changeaddress.htm.

Voss, P., 2005. History of the Committee on Applied Demography. http://www.populationassociation.org/about/committees/about-cad/.

Washington State Census Board, 1944. Report of the Washington State Census Board for the Years 1943 and 1944. Washington State Printing Plant, Olympia, Washington.

Wilson, T., 2012. Forecast accuracy and uncertainty of Australian bureau of statistics state and territory population projections. International Journal of Population Research. Article ID 419824.

Applied Geography: A US Perspective

John W Frazier, SUNY, Binghamton, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Applied geography is client driven and user oriented, and emphasizes useful information. It can bridge the gap between university and nonuniversity settings. Applied geography has a long history of leadership by both academicians and practicing geographers in the government and private sectors. Some public and private institutions have woven the use of geography into their institutional fabric. American examples include the US Census Bureau, the American Geographical Society, and the Association of American Geographers. At the international level, the International Geographical Union has institutionalized applied geography in its core. Applied geography also has a notable history in the private sector, including in the business community. Geography's main teaching and research themes are expressed topically in applied geography's pursuit of problem resolution in specific human–environmental issues, site selection, and many other subfields of the discipline. Applied geography has a very bright future in the twenty-first century, especially with the increased public awareness of the field and the value of its knowledge base and geotechnologies.

Applied geography provides a user orientation and an effectuation plan that are different from those of curiosity-driven academic research (Frazier, 1982, 2004). Applied geography emphasizes the social relevance of geographic research. It is characterized by the value context placed on knowledge. A dialectic between academic and nonacademic (or applied) geography is reflected in the work of many famous geographers of the twentieth and twenty-first centuries and is institutionalized in public and private agencies. Many current problems require an applied geography approach.

Applied geography emphasizes the social relevance of geographic research that focuses on human–environment interactions, area study, movements, and spatial location problems. In general, the spatial distributions and patterns of physical and human landscapes are examined, as well as the processes that create them. These provide the basis for recommended actions for a client. Applied geography takes place within and outside of university settings and often bridges the gap between academic and nonuniversity perspectives.

Wellar (1998a) has referred to the difference between purely academic research and applied research as curiosity-driven vs client-driven research. Whether university or nonuniversity based, geography becomes applied when the researcher undertakes a problem for a client. While academic geographers' research is driven by their curiosity to understand patterns and processes, applied geographers perform research for a client with a 'real-world' problem. This results in an approach that differs markedly from an academic endeavor. The client can be a retailer in a capitalist economy; a group that receives unequal treatment from a capitalist or some other group and, therefore, would benefit from the empowerment that applied research findings may provide; or any other person or agency seeking or needing relief from a solution to a geographic problem.

The 'applied' approach provides a user orientation and an effectuation plan quite different from those of curiosity-driven academic research. The research results are also presented in different formats and used differently. Applied research results may be input into the design of a new product, may result in a strategy for delivering a new product, or may be a set of recommendations that inform decisions. Academic research pursued for curiosity typically results in the production of a book, monograph, or research article in a journal publication for the purpose of informing colleagues of findings, especially in terms of the contribution to existing theory. Applied geographers in vs out of the university often differ in their views of what applied geography is. Despite these differences, applied geography is united by emphasis on useful knowledge.

University-based applied geography is narrower in scope than is nonuniversity applied geography. The university presumes a theoretical and an empirical basis for applied research. Non-university-based applied geography occurs within business and government cultures, which often results in highly specialized roles for individuals applying geographic knowledge and skills in the solution of problems. The differences between the two workplaces shape research perspectives, define broader nonresearch obligations, and result in different perspectives. For example, in the university context, the researcher controls and participates in all aspects of research, leading most academics to define 'applied' as the research performed by the researcher for a client. Most academics also emphasize research over effectuation planning.

By contrast, in a nonuniversity setting, an applied geographer has a job description that indicates specific tasks and places the employee in a reporting hierarchy (Frazier, 1994). Individual responsibilities are tied to a 'client relationship' that is inherent to a production environment. The employment specializations of nonacademic geographers range from planner–technician to analyst, from a research scientist to director of operations, and from project director to executive (Frazier, 1994). In the nonacademic world, every cog in the wheel is crucial. In short, the very nature of the nonacademic world requires multiple stages in the creation of a product. An applied geographer may be performing research or implementing the research of others, doing none himself or herself, but contributing greatly to the process. Being directly

involved in all parts of the applied research process is not a prerequisite to doing applied geography outside academia.

What holds applied geography together is the value context placed on knowledge. The applied geography approach seeks useful knowledge that helps solve a specific problem for a specific client. Pacione (1999) describes examples of practical recommendations for policy that address the issue of ethical standards in applying science. Specifically, he used British and American examples that illustrate applied geographers taking antiestablishment, unpopular positions to implement research findings that would change existing urban policies.

The goals of applied geographic research differ from those of basic research but the boundary between the two can be fuzzy. The value of 'useful knowledge' helps clarify the differences between the two approaches.

Applied Geography: An Historical Context

Early Leaders by Example

A dialectic between academic and nonacademic geography is reflected in the work of many famous geographers in the early twentieth century. Perhaps the most frequently cited case of geography in action is the work of L. Dudley Stamp in the Land Utilization Survey of Great Britain. Stamp called for geography not only to use its methods and concepts to interpret the world but also to help forward the solution of some of the great world problems (as quoted in Frazier, 2008: p. 3). Stamp's work in the 'Survey' put geographic methods and concepts into action and provided a touchstone for their utility. The 'Survey' methods and findings were replicated by others and informed geographic theory and approach for decades, providing not only planning inputs but also research topics in academic geography.

In the United States, several geographers of national reputation converted their academic interest in human–environmental topics into government employment. Among them were Harlan Barrows, who between 1933 and 1941 worked on a variety of water resource projects, including the design of comprehensive development schemes on a regional basis (Colby and White, 1961). Even earlier, Carl O. Sauer, an architect of academic cultural–historical geography, played a leadership role in the Michigan Land Economic Survey. The frequently employed theme of 'human as agent' was applied in this land assessment and classification effort. Sauer applied well-established academic geographic concepts and methods in an applied research mode. He dealt directly with specific forms of 'destructive exploitation' and their specific impacts on the land. Further, his scheme included a model of potential future uses, including modifications that would realize better results. Sauer's efforts were utilized in land management planning and were also published in the academic literature (Sauer, 1919); they informed future action and future academic research. Sauer followed up on applied research by being involved in the implementation stages as an activist.

The most internationally respected expert on natural hazards is the late Gilbert F. White. He is another example of an academic geographer who served the nonacademic community for half a century. White's early concern for victims of flooding and other natural hazards led to applied work for the administration of Franklin D. Roosevelt and new policy. White established academic and nonacademic followings and his leadership resulted in the creation of The Center for Natural Hazards Studies at the University of Colorado, Boulder. White and his colleagues have used geographic concepts and methods to understand different types of natural hazards and find ways to mitigate their impacts on people. Stressing physical and human variability, White focused on improving the plight of human kind (White, 1974).

While the review of applied geography in this article does not aim to be comprehensive, it would be remiss to exclude the founder of business geography in the nonacademic world. In the 1930s, William Applebaum accepted a position with The Kroger Company. His introduction of basic geographic concepts and methods for analyzing market regions was highly successful and created a career path for others to follow. Initially, applied geographers focused on measurement techniques, market delimitation, and competitor analyses. However, Applebaum maintained a close relationship with a small group of academic geographers and hired new geography graduates. Slowly, applied business geographers expanded their roles to include location analyses, site selection, and the application of models. As a result of this dialectic relationship between Applebaum and, later, others from the business world with academics, a new subspecialty, business/marketing geography, emerged in the United States and new concepts and methods were exchanged between the two environments. In Europe, Ross L. Davies is an example of an academically based applied geographer who spent a career conducting research that guided understanding of retail structure and behavior and that has informed business decisions.

Institutionalizing Applied Geography: Some Examples

Geography has been incorporated into government and business at various levels and in various ways. At one level, geography is inherent in business decisions and has resulted in many firms hiring geographers as location and/or site analysts (Grosvenor Center for Geographic Education, 2012; Solem et al., 2013). Some have risen to managers and executives. At another level, however, agencies formalize geography through their missions and/or internal organizational structure. Three examples suffice.

The American Geographical Society (AGS): Useful Knowledge and Geography in Action

When writing the history of the American Geographical Society (AGS), J.K. Wright noted its earliest public purposes: "The advancement of geographical science and the promotion of business interest of a 'great maritime and commercial city' were the Society's two leading purposes" (Wright, 1952: p. 69). These were most often expressed in the society's early years through published research reports, including those resulting from Society-backed field expeditions. An early example, which also reflects another Society purpose, supporting religious causes was the appointment of the 'Committee on Syrian Exploration,' which was championed by a rabbi and, according to the Committee, would "lead to important discoveries and human development by reclaiming the

desert, creating self-sufficiency and linking the region to world trade" (Wright, 1952: pp. 39–40). This was in keeping with the Society's purposes and exemplifies its efforts to put geography into action, which would promote human welfare.

In attempting to institutionalize applied geography in society, the AGS provided important data useful to government and business decision making. This is obvious in a number of the Society's efforts. In its early years, statistics had been hailed as knowledge capable of eliminating 'the fogs of human ignorance and suffering' (Wright, 1952: p. 47). The Society attempted to influence census taking and even proposed a new federal tax system as its 'most ambitious attempt to influence government policy' (Wright, 1952: p. 50). The intended applied thrust of the AGS is perhaps best stated in the inaugural issue of its major journal, *Geographical Review*: "It is the essence of the modern ideal that knowledge is of value only when transformed into action that tends to realize the aspirations of humanity. It is precisely this view that the Society has always taken" (*Geographical Review*, 1916: pp. 1–20, as quoted by Wright, 1952: p. 195).

Other national and international roles by the Society and its members are well documented. Among the most prominent are Society activities associated with the Congressional Act that established the International Meridian Conference in Washington, DC, and the leadership role of Society Director, Isaiah Bowman, in 'The Inquiry' after World War I. This enterprise was headquartered at the AGS. At the Paris Peace Conference that followed, Bowman served in an executive capacity at the request of President Wilson.

Most recently, the AGS has undertaken work that mirrors its long-term mission of creating useful information for foreign lands. Under the leadership of President Jerome Dobson, the Society created the Bowman Expeditions in the early twenty-first century. One of these expeditions, to Mexico, he labeled as a 'modern incarnation' of that mission that extended the early twentieth century effort termed 'The Millionth Map of Hispanic American,' by which the entirety of Latin America was mapped at the scale of 1:1 000 000 (Dobson, 2009: p. 3). Useful information has remained a cornerstone of the Society.

Applied Geography in Government: The US Bureau of Census Geography Division

Perhaps no federal agency anywhere has been more influenced by the geographic perspective than the US Bureau of the Census. For more than a century, the value of geographic knowledge has been applied in the acquisition and reporting of data for the establishment and monitoring of national and state programs in Commerce, Agriculture, Housing and Urban Development, Health and Human Resources, Indian Affairs, Transportation, Energy, Defense and Central Intelligence Agency, and other federal agencies. Applied geographers have played important roles throughout the federal government in such activities. However, the Geography Division of the US Census has long been charged with very specific duties, including the delineation of meaningful statistical areas as well as the creation and/or implementation of methods for the capture, organization, and reporting of useful geographic knowledge for problem solving. Torrieri and Radcliffe (2006) noted the early work of Census Chief Geographer, Henry Gannett, who by 1890 led the development of area definitions and mapping techniques, and was largely responsible for a variety of population reports, including the 'Report of the Nation' (Torrieri and Radcliffe, 2006).

Over the next century, geography's role expanded and, by 1978, Chief Geographer, Jacob Silver, provided a summation of responsibilities that included the development of national geographic reference files and digital mapping techniques (Silver, 1978). The development of the Geographic Base File-Dual Independent Map Encoding system provided the basis for future census taking and choropleth mapping of the US Census in the 1970s. By the 1980s, its national database development continued with Topologically Integrated Geographic Encoding and Referencing (TIGER), a digital street network for the entire nation, that permitted compliance with the federal mandate that all future census maps be digital. By the close of the twentieth century, the Geography Division's TIGER system became the most widely used digital base file in the world. Geographers also played an important role in the creation and administration of the American Community Survey. The Division is planning new census techniques, area definitions, and portrayal methods for census data in the twenty-first century.

Applied Geography in the Missions of Its Largest Organizations

Two notable examples of Geography's largest organizations incorporating applied geography into their missions are the International Geographical Union (IGU) and the Association of American Geographers (AAG). The IGU has seven broad objectives, including the promotion of applied geography through its 'Applied Geography' Commission. Its members have created multiple means of broadening the understanding of applied contributions, such as the 2004 volume that offered a global perspective from many scholars involved in these pursuits (Bailly and Gibson, 2004).

One of the most striking efforts of the late twentieth and early twenty-first centuries are the efforts of the AAG, which has effectively linked geographic education with the increasing demand for those with geographic skills. Through a series of National Science Foundation grants entitled 'Enhancing Departments and Graduate Education,' the AAG reached out to public agencies and private sector firms to better understand their current use of geographers and their future needs. These efforts have increased the visibility and depth of applied geography at the annual meetings of the AAG, attracting significant representation from the employment sectors hiring geographers. AAG-supported publications have also emerged that benefit students, faculty, and employers, such as *Practicing Geography: Careers for Enhancing Society and Environment* (Solem et al., 2013). Although Geographic Information Science (GIS) has been a big consumer of academic geographic talents, it has become clear that the wider applicability of GIS exists in the employment arena. For example, in the United States, *CareerBuilder.com* and the nationally syndicated *Sunday Jobs* noted that geographers 'study the earth and land (and inhabitants) and examine political and cultural structures as they relate to geography.' It also noted that geographers are in 'the top 10 jobs with above average salaries' (Auerbach, 2012).

Themes in Applied Geography

Most subdisciplines of geography contain examples of applied geography. Examples include a concern for environmental processes and patterns that influence human development and well-being, such as the physical processes of erosion, sedimentation, and desertification. Human–environmental relationships take on many forms but prominent among them are natural hazards behaviors, climate change, and environmental degradation of all types. The applied regional approach, or area study, involves any effort to create or analyze geographic zones for the benefit of human welfare. The leadership role of census geography has already been reported, but others include the creation of planning regions by applied geographers to plan, administer, and monitor urban and environmental areas. Location principles are applied to a wide range of analyses but are probably best known in applications for site selection in business and government and for retail market analyses, including the internationally known Huff model (Huff, 1964).

Torrieri and Radcliffe identified seven general classes into which they believe most applied geography falls: (1) market/location analysis; (2) medical geography; (3) settlement classification and statistical geographic areas; (4) land use, environmental issues, and policy; (5) transportation planning and routing; (6) geography of crime; and (7) developmental tools and techniques for aiding geographic analysis (Torrieri and Radcliffe, 2006). They admit that these classes are not exhaustive. It is unlikely that any nomenclature will satisfy all applied geographers. However, such classifications are useful in understanding the wide range of applied geography that occurs internationally.

One contemporary example illustrates geography at work in the transportation and routing area. Barry Wellar, who worked for the Provincial government of Ontario before becoming a professor at the University of Ottawa, led a research team under contract with that government to develop a 'Walking Security Index.' This index was used to assist Canadian officials and citizens in assessing the safety of signalized intersections for pedestrians (Wellar, 1998b). After developing the necessary indices and examples for the government, Wellar and his colleagues presented their results in public meetings and through the media. The final report contained 17 intersection recommendations based on the research and interactions with the three clients, government officials, staff professionals, and citizens. It also contained implementation issues and strategies.

Future of Applied Geography

Several trends indicate a bright future for applied geography. As noted above, applied geography was institutionalized both inside and outside academia in important ways. One trend is its increasingly formal treatment as part of academic debates (Pacione, 1999). Further, the British journal, *Applied Geography*, created a publication forum for those interested in particular aspects of this approach. Finally, in America, the annual Applied Geography Conference can be considered an institution that has brought thousands of applied geographers together from business, government, and the university for nearly four decades.

The second trend in geography is the recent improvements in geographic education at all levels. Professional European Geography was founded in the nineteenth century at least in large part as a result of educational concerns. At the close of the twentieth century, concerns for geographical ignorance spawned a revolution in geographic education. In America, the generosity of the National Geographic Society and the labor of educators of the National Council for Geographic Education have contributed significantly to improving the quality of geographic teaching through Geographic Alliances, which bring together K-12 teachers and university faculty in all 50 states. If applied geographers can be viewed as the 'outputs' of geographic education, then children are the 'inputs' that are shaped and mentored at all levels of the educational process. This trend of improved geographic education to rid the world of geographical ignorance has contributed positively to a third trend, that of greater public awareness of the importance of geography and its value, not only in better appreciating and understanding our local and global environments but also in contributing to the solution of significant global (e.g., global warming), national (e.g., poverty and inequality), and local (e.g., environmental health) problems. The establishment of the Grosvenor Center for Geographic Education at the Texas State University and a PhD specialization there in geographic education illustrate the increasing importance of this topic. So, too, do the employment trends previously mentioned.

Another trend that has contributed to our ability to recognize and solve geographic problems and to institutionalizing geography in the public and private sectors is the creation of geographic-based automated technology. Geography is now inseparably linked to information systems via GIS. Many employers seek technical and analytical employees who can use this technology in problem solving. Among the other digital technologies often linked to GIS and, therefore, extended to applied geography, are remote sensing and global positioning systems. Both technologies frequently find their academic homes in geography departments and are requirements for undergraduate and graduate degrees. In short, technical skills are now often associated with applied geography and its practitioners because they acquire, portray, and analyze useful geographic knowledge for problem solving.

Finally, and perhaps most important, there is no shortage of pressing global, regional, and local problems that require an applied geography approach. Many have been mentioned here. As long as there are geographers willing to contribute to the solution of problems that are inherently geographic or have geographic dimensions, there will be applied geography. Where geography is put into action, applied geography occurs.

See also: Cultural Geography; Electoral Geography; Geographic Information Systems; Geography: Overview; Global Positioning System; Industrial Geography; Medical Geography; Social Geography; Spatial Optimization Models; Transportation Geography.

Bibliography

Auerbach, D., 2012. 10 Jobs with above Average Salaries. Sunday Jobs. Classified. Press & Sun Bulletin, 1D.

Bailly, A., Gibson, L.J., 2004. Applied Geography. A World Perspective. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Colby, C.C., White, G.F., 1961. Harlan Barrows 1877–1960. Annals of the Association of American Geographers 51, 395–400.
Dobson, J.E., 2009. Let the indigenous people of Oaxaca speak for themselves. Ubique XXIX (1), 1–3.
Frazier, J.W. (Ed.), 1982. Applied Geography Selected Perspectives. Prentice-Hall, Englewood Cliffs, NJ.
Frazier, J.W., 1994. Geography in the workplace: a personal assessment with a look to the future. Journal of Geography 93 (1), 29–35.
Frazier, J.W., 2004. Applied geography in North America. In: Bailley, A., Gibson, L.J. (Eds.), Applied Geography: A World Perspective. Kluwer Academic Publishers, London, pp. 187–210.
Frazier, J.W., 2008. Applied geography as life-long learning, research in geographic education. Geographical Review 9 (1), 15–20, 1916 1: 1–2.
Gilbert, M., 2012. Grosvenor Center for Geographic Education. Why Geography is Important. Grosvenor Center, San Marcos, TX.
Harris, C.D., 1997. Geographers in the U.S. Government in Washington, D.C. during World War II. Professional Geographer 49 (2), 245–256.
Huff, D.L., 1964. Defining and estimating a trade area. Journal of Marketing 28, 36–38.
James, P.E., 1972. All Possible Worlds. A History of Geographical Ideas. Odyssey Press, Indianapolis, IN.
Mayer, H.M., 1982. Geography in city and regional planning. In: Frazier, J.W. (Ed.), Applied Geography Selected Perspectives. Prentice-Hall, Englewood Cliffs, NJ, pp. 25–57.
Pacione, M., 1999. Applied geography: in pursuit of useful knowledge. Applied Geography 19, 1–2.
Sant, M., 1982. Applied Geography: Practice, Problems and Prospects. Longman, London.
Sauer, C.O., 1919. Mapping the utilization of the land. Geographical Review 8, 47–54.
Silver, J., 1978. Bureau of the census—applied geography. Applied Geography Conference SUNY Binghamton 1, 80–92.
Solem, M., Foote, K., Monk, J. (Eds.), 2013. Practicing Geography. Careers for Enhancing Society and the Environment. AAG and Pearson Education, Glenview, IL.
Taylor, P., 1985. The value of a geographic perspective. In: Johnston, R.J. (Ed.), The Future of Geography. Methuen, London, pp. 92–110.
Torrieri, N.K., Radcliffe, M.R., 2006. The history of applied geography. In: Gaile, G.L., Willmott, C.J. (Eds.), Geography in America at the Dawn of the Twenty-first Century. Oxford University Press, Oxford, UK.
Wellar, B., 1998a. Combining client-driven and curiosity-driven research in graduate programs in geography: some lessons learned and suggestions for making connections. Papers and Proceedings of Applied Geography Conferences 21, 213–220.
Wellar, B., 1998b. Walking Security Index. DOT Regional Municipality of Ottawa-Carleton (RMOC), Canada.
White, G.F. (Ed.), 1974. Natural Hazards. Oxford University Press, New York.
Wright, J.K., 1952. Geography in the Making. The American Geographical Society. American Geographical Society, New York, pp. 2851–2951.

Applied Social Research, History of

Sara M Strickhouser and James D Wright, University of Central Florida, Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The history of applied social research is reviewed. The social sciences generally are a product of the Enlightenment and most of the social sciences originated in applied concerns. Thus, in contrast to the conventional view that the basic science is developed first and then applied to solving certain problems, most of the social sciences developed in the opposite direction: in the effort to understand and solve various social problems (education, crime, poverty, pestilence), the pioneers of the social sciences developed the theories, concepts, and methods that came eventually to comprise the basic disciplines. The history of applied social research is, indeed, the history of the social sciences themselves.

At the most general level, our topic is the historical relationship between basic and applied social science. Many discussions of this relationship posit the development of the basic scientific discipline – its theories, concepts, methods, and corpus of findings – as the necessary first step, and then the application of those theories, methods, and findings to problems of direct practical importance as a derivative activity, as an inferior or at least much less glamorous second step. Our thesis is that the historical development of the social sciences tended to follow an exactly opposite pattern. The earliest social science was developed by practical people seeking solutions to the pressing social problems of their day. Most of the social sciences were applied enterprises long before they developed as 'basic' scientific disciplines.

Origins

The idea of 'social science' emerged during and immediately after the Enlightenment, as a way to provide answers to questions about how individuals and institutions would cope with modernization, industrialization, mass political participation, the end of aristocracy, secularization, the spread of capitalism, the growth of the natural sciences, and, more generally, the end of traditional society. The Enlightenment, Edward Shils says, "was antithetical to tradition" (1981: 325) – the very traditions which, then and now, are "conducive to an ordered life." Without substantive traditions, how would modern societies function? What new bases of order would emerge? What, if anything, would prevent 'a widely ramifying disorder?' Would social life even be possible?

Scholars who eventually came to be defined as the founders of the disciplines of sociology, psychology, economics, political science, and even geography, and anthropology were all searching for an explanation of the post-Enlightenment world. Capitalism, science, democracy, and secularization were busily creating a new social order in the late-seventeenth and eighteenth centuries. By the nineteenth century, the Enlightenment ideas of reason, individualism, and freedom had created a nascent social science whose object was both to explain and justify this new order.

Those now considered the founders of the various social science disciplines all lived and worked at approximately the same time in history. Max Weber (sociology, 1864–1920), William Wundt (psychology, 1832–1920), and Franz Boas (anthropology, 1858–1942) were all born within three decades of each other. Adam Smith, arguably the founder of modern economics, was a century earlier (1723–90); and ditto Alexander von Humboldt (1769–1859), arguably the founder of scientific geography. The social sciences as we know them today were 'birthed' in the tumultuous period between 1750 and 1900, a period that includes the American and French Revolutions, the emergence of socialism, the unifications of Germany and Italy, the eclipse of aristocracy and the spread of mass democracy, and the political uprisings of 1848 on the Continent. This period simultaneously saw a shift in the scientific climate, where previously accepted thoughts and observations were challenged and empirical evidence was demanded.

Our point is that the questions the founders first asked were practical questions addressed to what we now call applied concerns. These applied concerns are the foundation of these disciplines as we know them today. As Alfred Marshall says (specifically about economics but more generally about all the social sciences), these disciplines are the "study of mankind in the ordinary business of life" (Marshall, 2011[1920]). They were not created for the purpose of producing theory-laden ideas in an icy laboratory or comfortable ivory tower but rather lively, warm-blooded, practical ideas that would not only explain, but also improve upon the welfare of society and everyone in it (Byrne, 2011).

The Enlightenment and the massive social transformations that ensued were important to the development of the social sciences both historically and conceptually. The rise of democracy and mass political behavior invested individuals and their actions with a new significance – democracy implied that the rulers were no more important than the ruled. Understanding variability in the behavior of individuals and locating those behaviors within specific social and institutional contexts became the *leitmotif* of a new breed of intellectual – not the philosopher, not the natural historian, but rather the social scientist. A century later, disciplinary lines of differentiation emerged, but the earliest of this new breed of scholar had wide-ranging interests in what we would now recognize as psychology, history, economics, sociology – indeed, the entire gamut of contemporary social science disciplines.

If the behavior of individuals was now of interest, then so too were their well-being, political outlooks, their views on

economy and society, all of which became fodder for social–scientific explanations of everyday life. As Jeroen Jansz (2004: 29) explains, "Rather than brute imposition of social order or religiously inspired charity, social management became infused with notions of rational, scientific social planning." Social management, not social theory, was the initial basis for what became social science as we know it. The animating concerns were things like crime and how to reduce it, education and how to increase it, poverty and how it might be eliminated, and on through a very large list of practical, applied concerns. The challenge was a data-based understanding of the world as it was, not the development of some concrete, general, and unwavering social theory that appeared to govern the happenings of society. A general theory of how the social world works was eventually the outcome, but by no means the origins of the applied work and concerns of the real founders of the social science disciplines.

Our intention here is not to debunk the importance of the traditional founders in the development of their respective disciplines, but rather to point out the applied and practical concerns of many of them to highlight applied origins and demonstrate that what is now considered 'applied research' is the precondition to the development of these disciplines as we know them today. Mainly, we aim to encourage modern contributors to these disciplines to pay homage to their applied roots.

Sociology

Johan Goudsblom and Johan Heilbron (2015), this Encyclopedia, argue that "American sociology developed primarily on the basis of empirical research into current social problems. The main topics were the issues which were widely debated by politicians, religious leaders and social reformers: massive immigration, urbanization, industrialization, local communities, social (dis)integration, ethnic and race relations. Empirical studies proliferated, based either on statistical material or on actual fieldwork. By the 1920s the sociology department in Chicago, founded by Albion Small in 1893, had become the predominant center." And while American sociology is not quite the same as sociology in general, this is a useful observation with which to begin.

If one were to ask the typical graduate student in sociology to name the 'founders' of the discipline, most would respond almost instinctively: 'Marx, Weber, and Durkheim.' A few might add Comte or Spencer to the list. And yet a recent paper by one of us (Wright, 2009) was entitled, *The Founding Fathers of Sociology: Francis Galton, Adolphe Quetelet, and Charles Booth; Or: What Do People You Probably Never Heard of Have to Do with the Foundations of Sociology?* (From whence much of the following discussion is adapted.) How, if at all, can these conflicting visions be resolved?

The notion that sociology began as a theoretical discipline originating in the ideas of European scholars such as Max Weber or Emile Durkheim is mainly due to the American Talcott Parsons, whose *The Structure of Social Action* (1937) and *Social System* (1951) pretty well *defined* the history of the discipline as it was taught in graduate departments from the 1950s forward. The 1937 book was particularly influential and was dominated by exegeses on Weber, Marshall, Pareto, Freud, and Durkheim. A like view of the discipline is promulgated in Robert Nisbet's well-known *The Sociological Tradition*, first published in 1966. Both Parsons and later Nisbet described the theoretical discipline as though it were the discipline in its totality and thus downplayed what Anthony Oberschall calls *The Establishment of Empirical Sociology* (1972), which owes a great deal more to, say, Galton, Quetelet, and Booth than to Marx, Weber, and Durkheim.

Even were we to grant that sociology was somehow 'founded' by the Big Three, we would also need to grant that all of them were motivated by strongly applied concerns. Emile Durkheim (1858–1917), for example, is often credited with making sociology a 'science' by insisting that social facts can only be explained by other social facts, a proposition suggesting that the individual acquires characteristics from society and not the other way around, thus appearing to deny the very idea of the individual in a radical way. It is true that Durkheim's many scholarly works and his founding and editing of the first journal of sociology, *L'Année Sociologique*, helped establish sociology as an accepted social science. But as George Ritzer has pointed out, Durkheim was quite open about suggesting social reforms based on his observations of the dismal state of education and the effects of the division of labor on social solidarity (Ritzer, 2011: 103, 106). *Suicide* too was dominated by practical concerns. The concluding passage of the first chapter expressed Durkheim's aspirations for the work, to discover "the means that can be used to counteract [suicide]" (1951[1897]: 52).

Max Weber (1864–1920) was known to his contemporaries primarily as a historian and an economist. Today, we tend to think of Weber as an abstract thinker, a theoretician, but he was intimately engaged in practical affairs and German politics: for example, he was a member of the German delegation during the peace negotiations in Versailles and a member of the committee that drafted the Weimar Constitution that paved the way for Hitler's ascension to power a decade or so later. Paul Honigsheim's *The Unknown Max Weber* (2000) refers to Weber as a 'rural sociologist' and elsewhere as a 'historian of agricultural and rural life.' It is hard to quarrel with the conclusion that "Weber's applied sociology has been given short shrift" (Swatos et al., 1998), due largely to Parsons' myopic insistence that Weber was more important as a theorist than as a practicing hands-on social scientist.

As for the third of the Big Three, Karl Marx, little need be said beyond the inevitable quotation from the *Theses on Feuerbach*: "The philosophers have hitherto only interpreted the world in various ways. The point, however, is to change it" (1969[1845]).

Venerated as these historical figures are, just what aspects of their work survive in contemporary sociology? Marx probably remains the most influential of the batch, although Marxian sociological themes have fallen out of favor in recent years. No one writes about suicide without a tip of the hat to Durkheim, but do the *Rules of Sociological Method*, the *Elementary Forms of the Religious Life*, or even *The Division of Labor in Society* still inform much current sociological research or practice? Weber, of course, lives on in studies of bureaucracy and in studies of 'class, status, and power,' and his *Protestant Ethic and the Spirit of Capitalism* is a hardy perennial in undergraduate social theory classes, but beyond that, what?

Consider, in sharp contrast, Sir Francis Galton (1822–1911), a contemporary of Marx, Weber, and Durkheim. At various points in his career, he would have been described as an anthropologist, sociologist, explorer, geographer, meteorologist, geneticist, psychometrician, psychologist, criminologist, and statistician, among others. There is not a single reference to Galton in Nisbet's 350 ponderous pages and it appears that few sociologists have ever heard of him. And yet his intellectual contributions include the following:

- According to Karl Pearson, Galton invented the concept of statistical correlation (Pearson, 1914).
- Galton was the first to apply statistical methods to the study of human differences and the inheritance of intelligence.
- Galton introduced the use of questionnaires and surveys for collecting data on human communities, which he employed in his genealogical and biographical works and for anthropometric studies. In an important sense, he invented the modern survey questionnaire and other features of the modern survey.
- He coined the phrase 'nature versus nurture' and was the first person to study genetic inheritance versus environmental conditioning and socialization as determinants of human characteristics.
- He is also considered the founder of psychometrics (the science of measuring mental faculties and psychological and social–psychological properties).
- In the late-1860s, Galton conceived the concept of the standard deviation.
- Galton also invented ordinary least squares regression and the use of the regression line; he was also the first to describe and explain the common phenomenon of regression toward the mean.
- In the 1870s and 1880s, he pioneered the use of the normal distribution to fit histograms of actual tabulated data. He also discovered the properties of the bivariate normal distribution and its relationship to regression analysis.

It does not overstate to say that any current issue of the *American Sociological Review* (or, for that matter, the *American Political Science Review*, the *American Economic Review* – indeed, any current journal in any social science discipline published anywhere in the world!) would owe as large an intellectual and historical debt to the Galton *corpus* just summarized as it would owe the Big Three combined. Contemporary social science, as it is actually practiced, is inconceivable without reference to Galton's techniques, concepts, and statistical theories.

What motives drove Galton to this dizzying string of discoveries and breakthroughs? Clearly, the motives were practical, not theoretical. Galton's passion was the statistical study of variations in human ability. The fundamental observation that drove his scientific work was that some people possessed talent in great abundance, and others, barely at all. What accounts for that variation and could anything be done about it? A more practical concern is difficult to conjure! As James Newman (1956) put it in his monumental history of mathematics, Galton was "essentially a social reformer," "imbued with a strong sense of social responsibility and of mission" (p. 450).

Adolphe Quetelet (1796–1894) was another contemporary of the founders and one of the most influential social scientists of the nineteenth century. He expanded a field of investigation that he termed 'moral statistics,' a field originally identified as 'political arithmetic' by English economist and philosopher Sir William Petty in the late-seventeenth century. Like Galton, Quetelet was a pioneer in the use of statistical reasoning to study social phenomena and as such, he profoundly influenced the course of European and eventually American social science. It is of some significance that when Paul Lazarsfeld was given a named chair at Columbia in 1962, the name he chose was the Quetelet Professor of Social Science.

Quetelet is rarely mentioned in modern-day histories of the development of social science, but his efforts to apply probability theory and statistical analysis to data on crime, suicide, and anthropometry were pioneering efforts whose progeny include criminology in its contemporary quantitative form. Nisbet's history of sociology identifies Quetelet as one of the 'political arithmeticians' who "compiled quantities of social statistics, pointing to correlations or 'patterns,' and preening themselves on the quantitative exactitude of their handiwork." From this, one would never know that Quetelet was a venerated figure throughout the nineteenth century, among empirical researchers in Victorian England and on the Continent (Cole, 1972).

What Nisbet (2004) described as 'preening' was in fact a near-obsession with accurate measurements of human characteristics. Quetelet believed that if measurements were sufficiently accurate, then observed sample values could be used to infer properties of entire populations. He did not quite grasp our modern concept of statistical inference, but as can be plainly seen, he came very close. He applied these concepts to statistics on crime, mortality, weight, disease, intelligence, occupations, levels of education, religious affiliation, economics, and other phenomena. His emphasis on accuracy in measurement led him to devise many improvements in census taking and data gathering. In a pamphlet on crime issued in 1831, he intuited what we know today as the law of large numbers and also intuited the effects of increasing sample sizes on what we now call 'the standard error of the estimate.'

Like others, Quetelet's scientific work was motivated by a reformist vision. He was interested in crime because he wanted to end it; in mortality because he wanted to delay it; in education because he wanted to maximize it. To the extent that these scholars had an academic agenda, it was to demonstrate that a true science of society was indeed possible and to show what such a science would look like.

Also worthy of mention in this context are the various statistical societies in Britain and on the Continent that came into being in the mid-to-late nineteenth century (see Elesh, 1972) – the Manchester Statistical Society (1833), the Royal Statistical Society (1834) (founded, incidentally, by Quetelet), the International Statistics Congress (1853), and even the American Statistical Association (1839). As David Elesh notes, all of these societies were sponsoring or in many cases doing themselves applied social research work that today we easily recognize as precursors to empirical social science.

Finally, there is Charles Booth (1840–1916) who, like Galton, goes unmentioned in Nisbet's history. And yet "many people trace the origins of modern survey research to Charles Booth, who produced a landmark study titled *Life and Labor of the People of London*," published in 17 volumes (!) over the

period 1889–1903 (Groves et al., 2009: 4). If we accept the conventional idea that Booth's London surveys were the first 'true' sample surveys of the modern form, then we must again acknowledge the primary role of applied issues in the conception and maturation of the discipline. John Madge's (1962: 536) comments on Booth's work and on the 'social survey movement' that Booth initiated are insightful and worth quoting. 'The social survey movement,' he says, "… was generally concerned with the type of sociography needed to establish the incidence of social evils, primarily with poverty, with the purpose of arousing public opinion into taking practical measures to deal with such evils. One cannot disagree with the verdict … that the social survey is not concerned with evolving any comprehensive sociological theory."

American sociology was also rooted in applied concerns. The focus of the early Chicago School on applied, practical problems was quoted earlier in this article. Many people consider the Chicago School to be the original foundation of American sociology, but this overlooks the important sociological work on race undertaken in the Atlanta Sociological Laboratory under the direction of W.E.B. DuBois at the end of the nineteenth century. DuBois's *The Philadelphia Negro*, although routinely ignored for generations, was not only the first social scientific study of recently freed African-Americans, it was the first seriously scientific sociological study ever done in the United States (see Wright, 2012, for details).

Psychology

Psychology provides another example of how the reality of discipline development is different from textbook caricatures. In most accounts, the first recognizably modern psychologist was the German Wilhelm Wundt (1832–1920). Wundt was a professor, a scholar, when he first proposed that psychology should be seen as an independent discipline, separate from and not reducible to biology, physiology, or philosophy (Jansz, 2004). His version of psychology focused on consciousness, that is, on a person's subjective experiences of self, others, mind, and the world at large. He is notable within psychology for having founded the first experimental psychology laboratory, in Leipzig in 1879.

It is of some significance that the first line of research taken up in Wundt's laboratory concerned human reaction times, and was animated by very practical, applied concerns. The story goes that Wundt was inspired to study this topic when astronomers were having trouble in accurately measuring the passing of a star through the crosshairs of a telescope. The measured time was different for each observer, which was attributed to the incompetence of the assistants doing the measurements. Wundt's insight was that the variable was not incompetence but reaction times that differed for every observer. "The reaction time studies conducted during the first few years of Wundt's laboratory constitute the first historical example of a coherent research program, explicitly directed toward psychological issues and involving a number of interlocking studies" (Danziger, 1980a: 106).

Wundt maintained that experimental psychology could only measure sensory input and motor output and that the most interesting inner workings of the mind would remain the province of the armchair philosophers (Bolles, 1993). Although his achievements were cited often by his contemporaries, few of his theoretical concepts or empirical results survived his lifetime (Jansz, 2004: 32). As Kurt Danziger concludes, Wundt's "crucial contribution seems to fall in the area of methodology rather than in the area of substantive theoretical concepts or specific empirical discoveries" (1980b: 109). Aside from the original reaction time studies, Wundt's psychology could not be said to have been animated by strong practical concerns. His mixture of physiology, psychology, lab procedure, and philosophy of science would seem to mark him as a classical 'academic' psychologist. Yet one of Wundt's earliest students was the Philadelphia-born Lightner Witmer, who earned his PhD in psychology under Wundt in 1893. Witmer "was one of the first to recognize the need for a field that applied psychology to real life, as opposed to it remaining purely experimental" (Pernaski et al., 2013). He is also credited with establishing the first psychology clinic in the United States and is often cited as the founder of clinical psychology.

Witmer's clinic was not based on any particular 'theory' of mind or 'school of thought' in psychology; his practical psychology was based more or less exclusively on "investigating the mental development of schoolchildren using statistical and clinical methods" (Fancher and Rutherford, 2012). A key clinical innovation was his "training school for the treatment of children suffering from retardation or physical defects that interfered with school progress" (Fancher and Rutherford, 2012). In one famous case, Witmer observed that a child having academic difficulties could not see well and thus procured for this student a proper pair of glasses, whereupon the student's academic progress improved. There was no Jamesian epistemology, no theory of consciousness, involved in this 'practical psychology.' The boy could not see!

In addition to Wilhelm Wundt and his students, Sigmund Freud (1856–1939) is often cited as a 'founding father' of academic psychology. Like Wundt, Freud's background was in biological science (neurology) and early in his career as a physician, he conducted research on cerebral palsy, aphasia, and neuroanatomy. But Freud's interests rapidly transcended biological confines. It is well known that his interests in psychotherapy originated in his interactions with Josef Breuer (1842–1925) and Breuer's treatment of one 'Anna O' (real name: Bertha Pappenheim). Anna O suffered various psychosomatic symptoms whose origins were diagnosed as 'hysterical,' the treatment for which was making her recall past events. Whether the reported remission of symptoms was real and whether the remission was due to the recalling intervention remain matters of contention. What is not disputed is that Freud saw in this experience a method to treat a wide variety of psychiatric disorders, an insight from which psychoanalysis was born, and that his animating interest was in developing methods to treat mental disease (not to develop grand theories of the Unconscious, as many commentators would have it).

To the early American psychologists, the application of psychology to a wide variety of pressing social problems and issues was "As American as Apple Pie" (Goodwin, 2012). The range of applications of early American psychology was far reaching and sought to further the discipline while simultaneously serving humanity. Applications included mental testing, the use of psychology in business and industry, and the

use of psychology to address personal psychological problems (what eventually became clinical psychology).

Mental testing began as a way of identifying students who were 'behind' and aiding them in achieving better academic performance. The method, of course, was the modern intelligence quotient test developed by Alfred Binet (1857–1911) and his colleague Theodore Simon (Goodwin, 2012). Binet and Simon believed that if one could identify low performing, academically disabled students reliably, these students could be sent away to receive the attention they needed to catch up (in this case to the Vineland Training Institute in New Jersey, a group home for 'feebleminded' children) (Goodwin, 2012). Lewis Terman expanded on the Binet–Simon test in 1916 with the Stanford-Binet intelligence test which, after many revisions, is still used today. Robert Yerkes (1876–1956) expanded on Terman's work by developing the first group tests given to army soldiers to determine whether soldiers were in the right positions or which soldiers might make good officers. Later in the 1920s, Yerkes and Terman united to create the first group tests in schools, the Scholastic Aptitude Test, which, also after many revisions, is still in use today.

A large group of psychologists who had been initially trained as experimentalists saw that their expertise had potential practical (and profitable) application when utilized to solve recurring problems in the private sector. Hugo Munsterberg (1863–1916) was the first to explore what is today called forensic psychology and later what is now known as industrial and organizational psychology (both with their own sections in this Encyclopedia). Munsterberg saw the critical importance of psychology in selecting juries (Munsterberg, 1908) as well as the benefits of understanding how psychology could affect consumer patterns (Munsterberg, 1913). Walter Dill Scott (1869–1955) also investigated the role of psychology in industry and business. He was commissioned by an advertising agency to research the psychology of advertising; today, psychological research is an integral and significant part of the advertising and marketing industry (Scott, 1903).

In the United States, clinical psychology began as a means to treat students with academic problems. Lightner Witmer, as discussed earlier, set up a section of his lab for the treatment of these children and created the term clinical psychology (this field too has its own section in this Encyclopedia). A key impetus to the growth of clinical psychology was the aftermath of World War II, when nearly 60% of returning veterans appeared to suffer from some type of mental illness. Since then, the application of psychology to solve practical problems has continued to grow within the discipline (Jansz, 2004). A shift from a primarily academic and laboratory-based endeavor to a social enterprise, accepting of its unavoidable existence within the practical context of modern society, has meant psychology is increasingly paying tribute to its roots in applied social science.

Economics

As with all the other social sciences, strands of economic thinking go back to the ancient Greeks; important precursors from the Enlightenment era include the utilitarians and the social contract thinkers (Bentham, Locke, Hume). But most economic historians trace the origin of the modern discipline to Adam Smith and his *Inquiry into the Nature and Causes of the Wealth of Nations* (usually rendered as *The Wealth of Nations*). The question Smith raised was what accounted for the obvious (even then) variation in national wealth. Some nations, that is, were obviously very wealthy; others were not. How could these differences be explained? The mercantilists argued that the basis of national wealth inhered in the accumulation of precious metals (mainly silver and gold); Smith eventually argued himself to the position that the true difference lay in the productivity of labor, but that became practically a footnote in the work as a whole. What Smith produced in the process of puzzling over this question was nothing less than a paradigm for the understanding of the political economy of capitalism. This, surely, is the foundational basis for all of modern economics.

Today's economists are criticized by some as maintaining relatively little interest in the history of their discipline outside obligatory courses in the history of economic thought. In economics, 'history' is but "a time series embedded within a set of data" (Kates, 2013: 17), not events and ideas to go back to, analyze, and learn from. Students of economics are shown the classical theories then "told not to go there" (Kates, 2013: 2). So, as with many of the social and behavioral sciences, the true origins of the discipline of economics have given way to texts and articles 'no more than 20 years old' that contain, at best, a caricature of the discipline's foundation and development (Kates, 2013: 1).

Contemporary economists often study the economic implications of various policy changes in order to frame and concretize policy decisions. This directly implies that a crucial charge of economics today is to solve real-world problems (Kates, 2013). Take, for example, the economic controversies surrounding proposed increases in the minimum wage, changes in immigration policy, educational policy, or welfare policies of all sorts – indeed, about nearly any policy being considered at the national level anywhere in the world. This is reminiscent of David Colander's (1992) lament that "applied economics" is "what most economists do" (p. 197).

The phrase 'applied economics' was first used by the British economist John Neville Keynes, father of John Maynard Keynes, arguably the most influential economist of the twentieth century. Keynes drew a distinction between the 'science of economics,' by which he meant formal economic theory that could be developed deductively, by reasoning from first principles without the pesky interference of economic facts; and the 'art of economics,' using the scientific laws and principles of the discipline to solve practical, applied problems. The "art of economics" was labeled "applied economics" (Colander, 1992).

Today, most of economics inside the academy and all of economics outside of it is "applied economics" in Keynes' sense. Many university departments of economics will have economic theorists on the faculty and most of the Nobel Prizes in economics are for theoretical breakthroughs. But most economists are economists *of* something: labor force economists, demographic economists, agricultural economists, environmental economists, education economists, and on through a long list that is more or less isomorphic with the undergraduate economics curriculum in most universities.

A key tension in modern economics is between the elegance of formal economic theory and the grubbiness of modern economic realities. As David Byrne (2011) explains, many foundational economic principles have failed to predict or explain much of anything in the real world, due largely to unrealistic assumptions required to maintain *ceteris paribus* conditions within models, which can never be the case in reality (Byrne, 2011: 182). Formal theory demands these assumptions; the real world regularly defeats them. A key element in the economic controversy over increasing the minimum wage was simply whether neoclassical price theory could be successfully applied to wages and employment. This was much more important to the economists than to the economy (Leonard, 2000).

Political Science

Founders of political science are often said to include Thomas Hobbes, John Locke, and Niccolo Machiavelli, among others. Later Herbert Baxter Adams (1850–1901), John Burgess (1844–1931), William Dunning (1857–1922), and Albert Bushnell Hart (1854–1943) established the discipline in the United States during the late-nineteenth century and early twentieth centuries. Adams founded the "Johns Hopkins Studies in Historical and Political Science" in 1883, one of the first journals devoted specifically to political science. Burgess found *Political Science Quarterly* at Columbia in 1886. And Dunning and Bushnell both contributed greatly to the Reconstructionist literature after the American Civil War.

Despite the European precursors (Hobbes, Machiavelli), political science is often said to be an American invention. It was the Progressive Era in the United States (1890s–1920s) and the concerns of that era gave rise to a discipline dedicated to the application of its knowledge, concepts, and techniques to address current political and societal issues. Raymond Seidelman and Edward Harpham (1985) propose that the persons most directly involved in taking the discipline in an applied direction included Woodrow Wilson (1856–1924), Charles A. Beard (1874–1948), Charles Merriam (1874–1953), and Harold Lasswell (1902–78). In Seidelman and Harpham's (1985) account, Wilson is rendered as the grandfather of the discipline and Beard, Merriam, and Lasswell as forgotten uncles.

Even Woodrow Wilson's work as a political scientist prior to his Presidential win in 1912 included advocating for the application of the science to aid the government in knowing what they could do better and how they could work more efficiently. Wilson began his piece *The Study of Public Administration* (1887) by pointing out (in the first sentence) that the only sciences that get studied are those sciences that are necessary in a society (i.e., sciences with some *application*). And in the first sentence of his other major work, *Constitutional Government in the United States* (1908), Wilson pointed out that his goal throughout the piece was to explore the US Constitutional system "with an eye to practice, not to theory" (p. 1).

Charles E. Merriam was a builder of the modern discipline of political science and his influence can be seen in essentially every facet of political science as it is practiced today. He is best known academically as the founder of value-free behavioralism, which aimed to understand, free of ideological lenses, the psychological and social factors that shaped mass political behavior. His more pronounced influence on students and professionals of political science came from his ability to persuade a generation of political scientists of the importance of understanding the reasons behind specific political behaviors and how the understanding of those underlying reasons could be the bridge between academic political science and the workings of the federal government. This was the behavioralism revolution that lives on today within the discipline.

Merriam, along with many others, sensed a crisis within America's democracy at the beginning of the twentieth century and dedicated much of his science to finding a solution. He was obsessed by the possibly deleterious effects of rapid technological change on politics and society and wanted to ensure that scientific knowledge was being produced in a way that could be useful to the masses as well as the federal government. Along with many others, he was fascinated by the apparent contradiction between Americans' zealous embrace of ingenuity, innovation, all that was modern, versus an often parochial, even antique political culture. His solution, rendered in modern terms, was to integrate science, including social science, in the political culture *and* in the political process. Science, he felt, would need to become "a non-revolutionary extension of democratic culture in America" (Barrow, 2000: 114).

It was not until Roosevelt's presidency began in 1933 that Merriam saw the uniting of academic political science study and its application. The onset of the Great Depression meant Washington was more interested in the advice of social scientists than ever before. As Seidelman and Harpham describe, applied political science was "vindicated" (Barrow, 2000: 126), as academics secured a place at the public policy table. Roosevelt mirrored Merriam's affinity for practicality and for an experimental approach to democracy.

Charles A. Beard is considered more of an historian than a political scientist by present-day scholars, but he played a major role in the development of the discipline (Barrow, 2000). Beard's work was popular among intellectuals and laymen alike. He felt that one of the most important questions to be asked of the modern (late-nineteenth century) political system was who would conduct the industrial and economic planning for the country. He founded the Rand School of Social Science in 1906 in order to train officials, party organizers, labor journalists, and others in socialist theory and applied social science in order to give them the academic edge he thought they needed to be successful.

Other Disciplines

Both space and expertise limit the attention we are able to give to the applied origins of the other social sciences: of geography, anthropology, and the like, not to mention all the subdisciplines: gerontology, criminology, demography, and all the rest. Anthropology, as all the social sciences, can trace its roots to the ancients, but modern anthropology greatly expanded as a direct result of European colonialism, the contact of the colonizing powers with indigenous populations, and the obvious administrative needs of the colonial and imperial powers to understand these indigenous populations, their languages, customs, institutions, and cultures. Anthropology and related sciences

sprang into existence to satisfy these managerial needs to study scientifically the cultures of colonized people – a source of some embarrassment to modern-day anthropology.

Geography is arguably the oldest of the social sciences, since the need to understand and record where things are is prehistoric. But the evolution of geography into a recognized academic discipline took place, again, in the post-Enlightenment era, during the late-eighteenth and nineteenth centuries. Alexander von Humboldt is usually considered the father of geography as an academic and scientific discipline. Humbodlt's famous survey of the geography of Latin America (1817) laid the foundation for the science of physical geography and to some extent meteorology also. He was among the first to realize that latitude provided a means to compare climatic conditions in different countries and to measure the decline in temperature with altitude, and he intuited the geophysical causes of tropical storms. He is also the progenitor of the idea, now common in geographical circles, that the proper domain of geography is at the intersection of geology, biology, and meteorology, that these intersections determined where plants grew and thus where animals would concentrate; from this, the spatial ecology of the human species could be properly inferred.

Today, the geographers have expanded their rightful domain to include the spatial aspect of all social activity and structure; thus, there is a geography of knowledge, crime, disability, even finance and children, among many, many others, all with corresponding entries elsewhere in this Encyclopedia. The situation in geography is thus reminiscent of that in modern economics: a few core techniques and basic principles as the 'science of geography' and an enormous range of applied work that defines the discipline's 'art.'

Conclusion

The history of 'applied social science' is indeed the history of the social sciences themselves. There can be no finer a conclusion to this article than a passage from an article by the sociologist Alvin W. Gouldner (1920–80) on "Explorations in Applied Social Science" (1989[1956]: 27)

> There seems to be no close correlation, therefore, between the development of generalizations by the pure disciplines and the multiplication of opportunities for, and varieties of, applied sociology. The applied sciences cannot be fruitfully regarded as springing Athena-like from the furrowed brow of the pure disciplines. Any metaphor which conceives of applied social science as the offspring, and of the basic disciplines as parents, is misleading. It obscures the point that the applied sciences often contribute as much to pure science as they receive from it.

See also: Anthropology, History of; Clinical Psychology in Europe, History of; Criminology, History of; Demography, Early History of; Enlightenment: Impact on the Social Sciences; Enlightenment; Macroeconomics, History of up to 1933; Political Economy, History of; Political Science, History of; Psychology, History of (Twentieth Century); Psychotherapy, History of: Psychiatric Aspects; Social Ecology: The Chicago School; Social Scientists as Experts and Public Intellectuals; Sociology, History of.

Bibliography

Barrow, Clyde W., 2000. More than a Historian: The Political and Economic Thought of Charles A. Beard. Transaction Publishers, New Brunswick, NJ.
Bolles, Robert C., 1993. The Story of Psychology: A Thematic History. Brooks/Cole Publishing Company, Pacific Grove, CA.
Byrne, David, 2011. Applying Social Science. The Policy Press, Bristol, United Kingdom.
Colander, David, 1992. Retrospectives: the lost art of economics. Journal of Economic Perspectives 6 (3), 191–198.
Cole, Stephen, 1972. Continuity and institutionalization in science. In: Oberschall, Anthony (Ed.), The Establishment of Empirical Sociology. Harper and Row, New York, pp. 73–129.
Danziger, Kurt, 1980a. In: Rieber, Robert W. (Ed.), Wilhelm Wundt and the Making of a Scientific Psychology. Plenum Publishing Company Limited, New York.
Danziger, Kurt, 1980b. Wundt's psychological experiment in the light of his philosophy of science. Psychological Research 42, 109–122.
DuBois, W.E.B., 1899. The Philadelphia Negro. Lippincott & Co., New York.
Durkheim, Emile, 1997. The Division of Labor in Society. Free Press, New York (Originally Published in 1893).
Durkheim, Emile, 1895. The rules of sociological method. In: Lukes, Steven (Ed.), The Rules of Sociological Method and Selected Texts on Sociology and its Method. Free Press, New York.
Durkheim, Emile, 1915. Elementary Forms of Religious Life. George Allen & Unwin Ltd, London.
Durkheim, Emile, 1951. Suicide. Free Press, New York (Originally Published in 1897).
Elesh, David, 1972. The Manchester statistical society. In: Oberschall, Anthony (Ed.), The Establishment of Empirical Sociology. Harper and Row, New York, pp. 31–72.
Fancher, Raymond, Rutherford, Alexandra, 2012. Pioneers of Psychology: A History, fourth ed. W.W. Norton, New York.
Goodwin, C. James, Baker, David B. (Eds.), 2012. The Oxford Handbook of the History of Psychology. Oxford University Press, Inc., New York.
Goudsblom, Johan, Heilbron, Johan., 2015. Sociology, history of. In: Wright, James D. (Editor-in-Chief), International Encyclopedia of Social & Behavioral Sciences, second ed. Oxford: Elsevier Ltd.
Gouldner, Alvin W., 1989. Explorations in applied social science. Sociological Practice 7 (1), 26–42 (Originally Published in 1956).
Groves, Robert M., Fowler Jr., Floyd J., Couper, Mick P., Lepkowski, James M., Singer, Eleanor, Tourangeau, Roger, 2009. Survey Methodology, second ed. John Wiley & Sons, Inc, Hoboken, NJ.
Honigsheim, Paul, 2000. In: Sica, Alan (Ed.), The Unknown Max Weber. Transaction Publishers, New Brunswick, NJ.
Jansz, Jeroen, 2004. A social history of psychology. In: Jansz, Jeroen, van Drunen, Peter (Eds.). Blackwell Publishing Ltd, Malden, MA.
Kates, Steven, 2013. Defending the History of Economic Thought. Edward Elgar Publishing, Inc., Northampton, MA.
Leonard, Thomas C., 2000. The very idea of applying economics: the modern minimum-wage controversy and its antecedents. History of Political Economy 32 (4), 117–144.
Madge, John, 1962. The Origins of Scientific Sociology. Free Press, New York.
Marshall, Alfred, 2011[1920]. Principles of Economics, eighth ed. Liberty Fund, Inc., Indianapolis, IN. http://files.libertyfund.org/files/1676/Marshall_0197_EBk_v6.0.pdf. (retrieved 13.04.14.).
Marx, Karl, 1969. Theses on Feuerbach. Progress Publishers, Moscow, USSR (Originally Published in 1845).
Munsterberg, Hugo, 1908. On the Witness Stand. York University, Toronto, ON. http://psychclassics.yorku.ca/Munster/Witness/ (retrieved 01.04.14.).
Munsterberg, Hugo, 1913. Psychology and Industrial Efficiency. Houghton Mifflin Company, New York.
Newman, James, 1956. The World of Mathematics, vol. 2. Novello & Co. Ltd, London.
Nisbet, Robert, 2004. The sociological tradition. Transaction Publishers, New Brunswick, NJ (Originally Published in 1966).
Oberschall, Anthony, 1972. The Establishment of Empirical Sociology. Harper & Row, New York.
Parsons, Talcott, 1951. The Social System. Free Press, Glencoe, IL.
Parsons, Talcott, 1968. The Structure of Social Action, vol. 1. The Free Press, USA (Originally Published in 1937).
Pearson, Karl, 1914. The Life, Letters and Labours of Francis Galton. Cambridge University Press, London.
Pernaski, Casey, Brown, Cecilia, DePetro, Emily, Fraire, Nicoletta, Wagner, Samantha, Whitaker, Harry, Smith, Erin, Carey, Larry, Clark, Tyler, Moore, Michelle, 2013. Approaches to a History of Western Psychology, second ed. Available online: contact. whit@nmu.edu (for access).

Ritzer, George, 2011. Sociological Theory. McGraw-Hill, New York.
Scott, Walter Dill, 1903. The Psychology of Advertising in Theory and Practice. Small, Maynard & Company, Boston.
Seidelman, Raymond, Harpham, Edward J., 1985. The Disenchanted Realists: Political Science and the American Crisis, 1884–1984. State University of New York Press, Albany, NY.
Shils, Edward, 1981. Tradition. The University of Chicago Press, Inc., Chicago, IL.
Swatos Jr., William H., Kivisto, Peter, Gustafson, Paul M., 1998. Weber, Max. In: Swatos Jr., William H. (Ed.), Encyclopedia of Religion and Society. AltaMira Press, Walnut Creek, CA.
Smith, Adam, 2005. An Inquiry into the Nature and Causes of the Wealth of Nations. The Electronic Classic Series, Hazleton, PA (Originally published in 1776). http://www2.hn.psu.edu/faculty/jmanis/adam-smith/wealth-nations.pdf (retrieved on 02.04.14.).
Weber, Max, 2009. The protestant ethic and the spirit of capitalism. Information Age Publishing, USA (Originally Published in 1905).
Wilson, Woodrow, 1887. The study of public administration. In: Braman, Sandra (Ed.), Communicating Researchers and Policy-Making. Massachusetts Institute of Technology, Boston, MA, pp. 61–85.
Wilson, Woodrow, 1908. Constitutional Government in the United States. The Columbia University Press, Norwood, MA.
Wright II, Earl, 2012. Why, where and how to infuse the Atlanta Sociological Laboratory into the sociology curriculum. Teaching Sociology 40 (3), 257–270.
Wright, James D., 2009. The founding fathers of sociology: Francis Galton, Adolphe Quetelet, and Charles Booth; or: what do people you probably never heard of have to do with the foundations of sociology? Journal of Applied Social Science 3 (2), 63–72.

Applied Sociology

Jammie Price, Journal of Applied Social Science, Wilmington, NC, USA
Jeff Will, University of North Florida, Jacksonville, FL, USA; and Journal of Applied Social Science, Wilmington, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term 'applied sociology' refers to a diverse group of practitioners all using sociology to 'understand, intervene, or enhance human social life'. Many different approaches to sociological application exist. We use the term 'sociological practice' or 'practitioner' to inclusively refer to applied, clinical, and public sociologists, as well as those who identify more with methods used across the social sciences: community-based researchers, participatory-action researchers, and translational researchers. With their work, all of these sociologists intend to impact groups of people in the present day. They primarily differ in the clients with whom they work and their level of engagement in implementing the action steps of a project.

Introduction

As Jim Wright points out (2009), applied sociology is and has been the foundation of basic sociology since the discipline began some 200 years ago. In the United States, Lester F. Ward called for applied sociology in his 1906 address to the first gathering of the American Sociological Association (then Society): "…sociology, established as a pure science, is now entering upon its applied stage, which is the great practical object for which it exists." The tradition of applied sociology grew with the Chicago School's focus on the social problems of that era and with the work of W.E.B. Dubois and Lewis (1999). (For a more direct discussion of this evolution, see Miller Iutcovich (1997).) Social scientists applied sociology by evaluating the impact of the New Deal programs during the Great Depression and the impact of World War II on soldiers and families (Hooks, 1983). Today, applied sociology is a diverse field, encompassing clinical sociology, public sociology, community-based research, participatory-action research, and translational research. (Some scientists refer to this method as translation, others as translational.)

In light of this variety, we see the topic of 'what is applied sociology' as an ongoing conversation. Members of the applied community routinely discuss this in a variety of venues, including the Association of Applied and Clinical Sociology and the Commission on Accreditation Programs in Applied and Clinical Sociology, as well as the American Sociological Association's section on Sociological Practice and Public Sociology. Members of sister organizations often apply sociology too, including the Association for Humanist Sociology, Sociologists without Borders, Society for the Study of Social Problems, and the International Sociological Association, among others. With all these organizations and all these types of sociologists, there is much we could discuss here. We offer a few key points.

A Working Definition

Applied sociology is using the sociological tools to "understand, intervene, or enhance human social life" (Steele and Price, 2004: p. 4). People apply sociology when they use sociological methods, theories, concepts, or perspectives (tools) to address a social problem or issue (Steele and Price, 2008). Lots of social scientists 'use' sociological tools in their teaching or in their research. But only a subset uses those tools to plan or engage in direct, concurrent social intervention or enhancement. We generally refer to those that do as applied sociologists. Many different approaches to sociological application exist. We use the term 'sociological practice' or 'practitioner' to inclusively refer to applied, clinical, and public sociologists, as well as those who identify more with methodology used across the social sciences: community-based researchers, participatory-action researchers, and translational researchers (Weinstein and Goldman Schuyler, 2008).

Drawing on the works of David Cooper (2011) and Price et al. (2009), we argue that what unites these groups is their commonality in goals, tools, and conversations – all of them focus on 'use;' that is, use-based research and use-based applications. Members of all of these groups use sociological tools to "understand, intervene, or enhance human social life" (Steele and Price, 2004). These groups differ distinctly from academic sociology in that they usually do not work just for the purposes of publishing. With their work, they intend to impact a group of people in the present day. The types of groups they want to impact differ. And, the type of impact they want to make differs.

What Do Applied Sociologists Do?

Sociological practitioners come in many colors, and occupy a wide variety of occupations. Many receive their pay as professors and teachers in institutions of higher education. Federal and state agencies or nongovernment organizations employ a large variety of practitioners. Corporate or business enterprises hire some, with a smaller minority serving as private consultants and owning firms. Some impact groups via concrete suggestions for change inside of survey findings, evaluation reports, needs assessments, legislative summaries, or editorials in newspapers. All of these reach a much wider audience than a publication in a typical academic journal, the end goal for academic sociology. In contrast, some practitioners want to engage in the action steps of change – in not just

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10517-3

evaluating or assessing, but in implementing the improvements along with group members. Hence, a continuum of engagement runs along sociological practitioners.

Clinical sociologists usually identify or diagnose a problem with an organization and then design an intervention to impact the problem. Sometimes they engage in the implementation of their recommendations, but not always. It is called 'clinical' because practitioners apply a medical frame to social groups or situations; assess, diagnose, treat, and assess again. The methodology of clinical sociology mirrors that of translational sociology. While the two are not mutually exclusive, clinical sociology, as a set of ideas and values, does not specify that a researcher participate in the action steps of a social intervention. Translational sociology does. Some translational sociologists, like public sociologists, take the findings of sociological research and try to make them useful to communities or organizations. They 'translate' the findings into goals, objectives, and action steps for groups, thereby making sociological research more useful. Some take their 'use value' a bit further by involving themselves in every conceivable step of improvement on a given issue from grant writing, to community organizing, to researching, to implementing change, evaluating the implementation, and building programs to institutionalize the intervention. In this way, translational research represents a sustained commitment to an issue and a community. Many funding agencies now require translational work.

Participatory-action researchers and community-based researchers similarly apply sociological perspectives and tools to help local organizations and community members understand and improve local issues. They partner with the vested organization or community members, or part of the project involves acquiring community participation. Hence, the terms 'community' and 'participatory.' The members, not the researchers, 'own' the project. It is their community and an issue that directly affects them, likely for years to come. Theoretically, then, the community members will lead the application of the research findings with the researcher(s) help. Unlike translational research, community-based and participatory-action researchers may not remain with a project through the implementation of all the action steps. They may not have the time, as many are college professors with students in the research. Each semester brings new classes and new students. Some schools maintain a focus on a particular neighborhood or issue across semesters by starting a 'center' or 'institute' with organizational structure and resources (i.e., a professor works to obtain grant funding) to support the research. (For a discussion on how to start a local research center, see Steele et al. (1989).)

Public and applied sociologists vary widely in direct purpose and method from all of the above groups. If you look at all of the projects in which a given public or applied sociologist participated, you would likely observe more selection in goals, clients, and methods than the above groups. For example, they may or may not work with a community organization or involve community members. They may or may not use an intervention or participate in the implementation of action steps. We largely socially construct the difference between public and applied sociology. The term 'public' is a relatively new term (largely accredited to Buroway, 2002, 2004), and public science of all kinds is currently en vogue. Whereas, the term 'applied' existed for a 100 years earlier in the United States and longer overseas. So your age and your mentors' ages largely impact whether you identify as an applied or public sociologist. That said, some tensions occasionally flare between members of the groups concerning how the two define 'client' and 'action.'

For a few applied sociologists, a client is a for-profit business from whom they receive monetary payment for services such as marketing products, satisfaction surveys, or utilization studies. This pushes the envelope for acceptable social 'enhancement' for some public sociologists. On the other hand, a few public sociologists remain too distanced from the 'public' and from concurrent, concrete action steps for some applied sociologist's taste. However, members of these groups increasingly put these differences aside and focus on the larger commonality – enhancing social life. As noted earlier, the Sociological Practice section of the American Sociological Association is now the section on Sociological Practice and Public Sociology. And the Journal of Applied Social Science (JASS) invites and regularly publishes public sociology, as well clinical and translational sociology, and community-based and participatory-action research.

Examples

To further illustrate applied sociology, we describe two projects from coauthor Jeffry Will and the Northeast Florida Center for Community Initiatives (CCI). He and his colleagues manage numerous projects annually with local community organizations to help improve services or to impact local social problems such as hunger, homelessness, high-risk or adolescent pregnancies, and substance abuse. (For more information on projects that the Center completed, see https://www.unf.edu/coas/cci/Past_Projects_/_Reports.aspx.) In one project they recently surveyed, interviewed, and observed 2117 clients of local food pantries to compose a demographic description of the client population (Will and Milligan, forthcoming). As part of that project they also documented their own observations and interactions with clients, staff, and settings. Subsequently, they informed the food bank of critical numbers from the data collection such as the percentage of clients below the poverty line and the percentage of food-insecure clients. But they also presented suggestions on organizational change that might allow the food bank to more effectively serve their client population, such as how to network better with other community organizations and how to make better use of their Web site. In another project, for a number of years CCI staff coordinated the local census and survey of the local homeless as part of a 1-day homeless count. Will and his team talked with as many people in shelters as they could find in 24 h. The survey data and the homeless count provided invaluable demographic data to the local social service agencies for planning purposes and to the state legislature for funding purposes.

Conclusion

In conclusion, many different approaches to sociological application exist. We use the term 'sociological practice' or

'practitioner' to inclusively refer to applied, clinical, and public sociologists, as well as to those who identify more with methods used across the social sciences: community-based researchers, participatory-action researchers, and translational researchers. With their work, all of these sociologists intend to impact groups of people in the present day. They primarily differ in the clients with whom they work and their level of engagement in implementing the action steps of a project. A great deal of commonality unites these groups – they all 'use' sociological tools to "understand, intervene, or enhance human social life" (Steele and Price, 2004).

See also: Academic Careers in Comparative Perspective; Academic Research and Employment: Recent Changes in Europe and the United States; Applied Social Research, History of; Homelessness in the United States; Needs Assessment; Participatory Action Research in Social Research; Peer Review: Organized Skepticism; Sample Survey Methodology, History of; Schools in the Social and Behavioral Sciences: Concepts and Historical Relevance; Science and Politics: Value Neutrality; Social Science Learned Societies and Professional Associations; Social Scientists as Experts and Public Intellectuals; Sociology, History of.

Bibliography

American Sociological Association, Section on Sociological Practice and Public Sociology. http://sspps.wordpress.com/.

Association of Applied and Clinical Sociology. www.aacsnet.net.

Association for Humanist Sociology. http://www.ccsu.edu/page.cfm?p=6817.

Buroway, Michael, 2002. Public Sociologies and the Grassroots. Address to Sociologists for Women in Society.

Buroway, Michael, 2004. Public sociologies: contradictions, dilemmas, and possibilities. Social Forces 82 (4), 1603–1618.

Commission on Accreditation Programs in Applied and Clinical Sociology. http://www.sociologycommission.org/.

Cooper, David, 2011. The University in Development: Case Studies of Use-oriented Research. HSRC Press, Cape Town, South Africa.

Dubois, W.E. Burghardt, Lewis, David Levering, 1999 (1935). Black Reconstruction in America, 1860–1889. Free Press.

Hooks, Gregory M, A., 1983. A new deal for farmers and social scientists: the politics of rural sociology in the depression era. Rural Sociology 48 (3), 386–408.

International Sociological Association. http://www.isa-sociology.org/.

Iutcovich, Joyce Miller, 1997. Professionalization of applied sociology: the role of sociological practice organizations. In: Steele, Steven F., Iutcovich, Joyce M. (Eds.), Directions in Applied Sociology 1985-1995, pp. 7–39.

Journal of Applied Social Science. http://www.sagepub.com/journals/Journal202060.

Northeast Florida Center for Community Initiatives. https://www.unf.edu/coas/cci/.

Price, Jammie, Straus, Roger, Breese, Jeff (Eds.), 2009. Doing Sociology: Case Studies in Sociological Practice. Rowman and Littlefield, Lexington.

Society for the Study of Social Problems. http://www.sssp1.org/.

Sociologists without Borders. http://sociologistswithoutborders.org/.

Steele, Stephen F., Lamp, Joseph, Counihan, Harold, Albert, Joan, 1989. Establishing a local research center at a college or university. Sociological Practice 7 (1). Article 28. Available at: http://digitalcommons.wayne.edu/socprac/vol7/iss1/28 (accessed 21.03.14.).

Steele, Stephen F., Price, Jammie, 2004. Applied Sociology: Topics, Terms, Tools and Tasks, first ed. Thomson/Wadsworth.

Steele, Stephen F., Price, Jammie, 2008. Applied Sociology: Topics, Terms, Tools and Tasks, second ed. Thomson/Wadsworth.

Ward, Lester, 1906. The establishment of sociology. In: Presidential Address to the First Annual Meeting of the American Sociological Society. http://www.asanet.org/about/presidents/Lester_Ward/PA1906.cfm (accessed 28.02.14.).

Weinstein, Jay, Schuyler, Kathryn Goldman, 2008. The ASA and sociological practice. ASA Footnotes 36 (7). http://www.asanet.org/footnotes/septoct08/asa_soc.html (accessed 28.02.14.).

Will, Jeff, Milligan, Tracey, 2015. Toward an understanding of food pantry food recipients and the agencies that serve them. Journal of Applied Social Science, vol 1. http://jax.sagepub.com/content/early/2013/11/08/1936724413509249.full.pdf+html (accessed 21.03.14.).

Wright, James, 2009. The founding fathers of sociology: Francis Galton, Adolphe Quetelet, and Charles Booth. Journal of Applied Social Science 3 (2), 63–72.

Apprenticeship and School Learning: Lessons from Germany

Peter Noack and Bärbel Kracke, University of Jena, Jena, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H. Gruber, H. Mandl, volume 1, pp. 601–604, © 2001, Elsevier Ltd.

Abstract

Two aspects in which school prepares students for their transition to work-life are addressed. First, schools are supposed to foster their students' occupational exploration and help them in deciding and planning concerning their future vocational training (or other educational careers after graduation). Second, cognitive apprenticeship is pointed out as a way of school learning in academic domains that promises to result in knowledge relevant to everyday situations and students' future work-life, in particular. As a basis for discussing vocational orientation and apprenticeship learning at school, the apprenticeship system of vocational training is introduced in more detail.

In a narrow sense, apprenticeship is a structured training of future practitioners in the real-life context of the workplace. In this context, apprentices acquire skills and knowledge necessary to work in their field of practice, and at the same time, they are socialized into the community of practitioners. While there is not the one system of apprenticeship, the German 'dual system' that combines training at the workplace and theoretical instruction in vocational schools in a quite successful way is often referred to as a prototype and has, for instance, strongly influenced the School-to-Work Opportunities Act of 1994 in the United States (cf Hamilton, 1990).

The German educationalist Kerschensteiner (1912), a prominent representative of the Reformpädagogik (educational reform) movement, strongly contributed to the development of the German vocational training system but had more in mind than only an adequate preparation of young people for a profession and work-life. He wanted schools to socialize responsible citizens for which he considered active modes of learning in a context of real tasks and challenges superior to instruction that builds on a merely passive reception of knowledge. His didactic approach to teaching chemistry, physics, and even art mirrors crucial aspects of the vocational training of apprentices and can be seen as a first step toward ways of school learning that are referred to as apprenticeship models.

When discussing apprenticeship and school learning, two aspects are highlighted in the following that address schools' mission to guide young people toward their adult role as working citizen. First, schools are expected to support students in developing realistic representations of their competencies and interests as well as in acquiring knowledge concerning, for instance, different domains of work, apprenticeships offered, ways of being trained for jobs of interest, and the labor market. That is, schools prepare for apprenticeship by fostering students' occupational orientation. (It should be noted that basically the same processes are also relevant concerning decisions between an apprenticeship and university education or the entry into unskilled work after school graduation as well as concerning possible disciplines represented at universities or domains of unskilled work.) Second, high-school graduates are expected to leave their schools equipped with knowledge and skills that they can readily apply to their future tasks and challenges. Situated learning (Lave and Wenger, 1991) and apprenticeship models of learning (Collins et al., 1989), in particular, are concepts of instruction in school aimed at an acquisition of knowledge that is not inert (Renkl et al., 1996) but can be successfully applied to real-world situations and problems. Before these two aspects are addressed, the apprenticeship system of vocational training is described in more detail to provide a better basis for understanding to readers who are less familiar with this system.

The Apprenticeship System of Vocational Training

Being systematically trained for skilled work by way of apprenticeship has a long tradition dating back to the Middle Ages, when (then, mostly male) adolescents were introduced to the knowledge and skills necessary in a given domain of crafts or trades by helping, working together with, and being supervised by a skilled practitioner or master in the field. The efforts and costs on the part of the adult expert were balanced by the apprentice's labor and its (real) outcomes. In fact, not only was the recruitment of future practitioners based on this exchange of training and labor but to a certain extent this was also true of the economic basis of many crafts and trades that were controlled by craft guilds, which defined the goals of the training. These goals included knowledge and skills but also aspects such as attitudes and values associated with a given field. These latter aspects, in particular, but also skills and their application to novel situations, cannot easily be conveyed by explicit instruction but were trained by the cooperation of novice and expert in the mastery of the real tasks in the everyday context of their work.

While the essence of apprenticeship training thus has a long tradition, variations depending on the type of work, on the historical time, as well as on the individual master-apprentice constellation are considerable, and it is difficult to pinpoint systematic patterns of development over time. In Germany, for example, the present dual system of vocational education (BMBF, 2012; cf Hamilton and Lempert, 1996), which combines training at the workplace and part-time (1–2 days) theoretical instruction at vocational schools, was not established before the beginning of the twentieth century, and

nationwide legal regulations were established only in 1969. At present, about 550 000 young Germans, roughly half of a birth cohort, start their training in one of a little more than 340 fields of apprenticeship each year (BMBF, 2013).

High school graduates apply for an apprenticeship of their choice with a company or institution. If they get a contract, they receive a small regular income during the typically 3 years of apprenticeship, which, however, is clearly lower than regular wages of skilled practitioners in their field. Upon passing a final exam at the end of the apprenticeship time, the successful apprentices obtain a certificate. Companies working in a given domain strongly rely on these credentials when hiring new employees. Presently, companies change the apprenticeship contract into a regular employment after the exam for about 60% of their apprentices, with the likelihood of apprentices to be kept in the company increasing with company size (Pahnke et al., 2013).

International comparisons (e.g., Gries et al., 2005; Berlin, 2013) show that the apprenticeship system of vocational education is not a unique German institution but is also offered in countries such as Austria and Ireland and can exist with a purely school-based system (e.g., France, Portugal). More widespread, however, vocational education is fully integrated in the school context with varying degrees of integration of experiences at real workplaces, typically by way of internships, as can be found, for instance, in Spain, Poland, or Greece. Finally, in some countries such as the United States, on-the-job training in a learning-by-doing mode without a systematic organization and regulations beyond the expectations and rules of the employing company is the modal form of vocational training. It should be added that college education plays a sizeable role as an institution of vocational training in the US, paralleling to some extent the role played by vocational education at schools in various European countries.

Despite rules and regulations on the national level, state certificates earned upon the successful completion of the training, and controls on the part of the regional crafts and trades chambers, the inclusion of individual institutions and companies in the Dual Vocational Training System results in considerable variations in the experiences of young apprentices depending on, for instance, the domain of work or the size of the companies. Differences already start at admission to apprenticeship, where widely varying criteria apply on which the companies decide. Moreover, large companies have their own training departments and employ special training personnel for their apprentices, where training may resemble practical education in vocational schools. In contrast, apprentices trained in small crafts shops typically work together with their skilled colleagues on construction sites.

Likewise, the cooperation of companies and vocational schools tied together in the system can work out well or poorly. Some small companies might consider the hours that their apprentices spend in school wasted time (and a waste of inexpensive labor). It may also not come as a surprise that small crafts shops as well as state-financed schools can often not offer machinery and other equipment that represents the technical state of the art, which is what apprentices in larger companies typically find.

Still, young people in apprenticeships report having mostly positive experiences (e.g., being involved in real tasks, knowledgeable experts), as shown by an extensive questionnaire study including about 6000 apprentices (Beicht et al., 2009). A major critical aspect that they like is the coordination of instruction in school, on the one hand, and practical learning in the company context, on the other hand, which could be improved by more intensive communication and a better mutual understanding.

A study of advantages and shortcomings of vocational education in six European countries, which included interviews of national experts (Grünewald and Moraal, 2001), reached favorable conclusions concerning the apprenticeship system, particularly with regard to the quality and state-of-art nature of the training, the national acceptance of the training outcomes and certificates, and the motivational advantages for the young people in training. At the same time, the reported findings pointed out that the system is less flexible in matching the changes of the supply and demand sides of the labor market. Related to this point is the observation of portions of young people who find it difficult to commit themselves to an apprenticeship in only one company. These observations suggest that a more flexible solution, based on a more modularized system that allowed trainees to change companies and acquire different partial certificates might be preferable. In general, the experts question the strong orientation of the dual system toward a final certificate because this orientation is unlikely to strongly encourage a positive orientation toward lifelong (vocational) learning. Finally, it has been pointed out that the organizational and financial requirements of the German dual system of vocational training exceed those of alternative systems. Additionally, from an economic perspective, having higher shares of vocational training as in the dual system approach is related to lower unemployment rates but not to income inequalities (Busemeyer, 2011).

Vocational Orientation in School

In addition to teaching students basic skills in languages, mathematics, sciences, and other domains, as well as fostering more general competences related to, for instance, self-regulation to prepare youth for work, schools are expected to provide career guidance facilitating decision-making concerning students' future work-life and the school-to-work transition. While this task is acknowledged as important in many countries, the ways in which it is approached within schools vary widely (Härtel, 2013). Whereas in some countries, such as Finland, Greece, and Spain, specific curricular elements have been established to prepare young people for the important transition to work, these issues are more often integrated into other classes, such as social science education and ethics, in several other countries (e.g., Hungary, Malta). Yet, another path is followed in, for example, France and Poland, where special workshops are offered.

In order to arrive at sound occupational decisions, students should be helped to gain a better general understanding of work and work-life; they need guidance in thinking about such basic questions as why people should work and why school should play an important role in preparing them for their future as working adults. At the same time, students need to be encouraged to think which jobs might be best for themselves

(cf Schneider). This involves self-reflective activities and gathering information on different domains of work and more general perspectives on the labor market. These processes need to be stimulated and guided by teachers, who consider themselves responsible not only for their students' grades at school, their success in final examinations, and their school graduation but also for the occupational future of these adolescents beyond school.

Obviously, schools and teachers cannot accept the responsibility for their students getting an apprenticeship or job after leaving school, but they should support students' efforts toward an active and systematic engagement in the search for a suitable apprenticeship, program of university education, or job. What career guidance school has to promote can be referred to as occupational decision competence (Berufswahlkompetenz; Driesel-Lange et al., 2010). Instruction should include fostering an active exploration (Kracke, 1997; Kracke and Schmitt-Rodermund, 2001) of own abilities and interests as well as the world of work and specific jobs. These processes of exploration meant to promote a realistic self-representation and goal-setting should be complemented by teaching specific practical skills such as writing an application for an apprenticeship or job and by supporting students' attempts to actually apply. To this aim, schools have to be open to the world of work so that they can establish and keep up close contacts with companies in the region in order to be informed about the content and quality of apprenticeships.

Empirical research shows that internships can be a particularly powerful tool among the various strategies schools employ in their career guidance efforts. Internships are systematically integrated into efforts of occupational preparation in, for instance, Denmark, Finland, and Germany (Härtel, 2013). In Germany, for example, one or two 1- to 2-week internships are part of the mandatory school curriculum. During internships, students experience (or at least observe) situations of real work-life in its everyday context. As a side effect, internships guided by schools can contribute to students' belief that school can be a meaningful tool to prepare for real life. During adolescence, in particular, when many students feel somewhat alienated from school, such real-life experiences could be fruitful in stimulating school engagement.

However, beneficial effects can only be expected under certain circumstances. It is crucial that the internship is well designed. That is, before students apply for or start their internships, they should have had a phase of thoroughly assessing their interests and values. After this phase of self-assessment, it would be helpful for students to explore possible jobs on the Internet or by interviewing people as well as by experiencing a real application process and to then set goals for the timing of their internship. Similarly important is a guided reflection phase after completion of the internships when students are encouraged to understand whether and in which ways their occupational self-concept might have changed due to their experiences as well as to think seriously about the possible consequences of their internship experiences for their future school engagement.

The successful design of internships as well as an appropriately fruitful reflection phase requires that teachers know about the development of vocational self-concepts and how to stimulate processes of self-regulation (goal-setting, goal-pursuit, evaluation; cf Zimmerman). These teachers also need to have an idea as to which goals are realistic for which students as these students approach selecting and then enacting an internship. The latter aspects, in particular, call for values and in-class behavior of teachers that parallel the conditions of fostering intrinsic motivation as specified in self-determination theory (Deci and Ryan) such as opportunities for students' experiences of competence and autonomy as well as of positive social relations. While it is feasible to more strongly highlight corresponding elements in teacher education at universities, parallel efforts targeting the supervisors of internships at the respective workplaces are quite a challenge.

School Instruction as Apprenticeship

Fostering an active acquisition of information and competencies as well as self-regulation is a goal of learning in school that is not only relevant in career guidance. Rather, outcomes of instruction in the different academic domains addressed in class have often been criticized because they do not match the requirements of the later real-life situations to which they are supposed to be applied. The acquisition of knowledge and skills at school traditionally takes place in a situation that is isolated in two ways. First, the focus is on individual ability and performance. Second, school is designed as a peculiar learning setting for children and adolescents that is often detached from the rest of their everyday lives. It is this isolated and very limited ecological validity, in a Bronfenbrenner (1979) sense, of school learning that is pointed to as a major barrier to students being able to fruitfully employ the acquired knowledge and competences in the world beyond school.

Against this backdrop, processes of learning inspired by the apprenticeship system of vocational training have been suggested as a strategy to extend and improve instruction at school. While the basic principles had been introduced earlier, Collins et al. (1989), in particular, have conceptually specified them in their model of 'cognitive apprenticeship.' Several advantages of apprenticeship learning have been pointed out. In particular, there is often more to mastering tasks and problems in a given domain than could be easily taught by explicit instruction. In fact, there could even be aspects of which experts in the domain are not readily aware. Such implicit aspects of the knowledge and skills needed for particular vocations when modeled by an expert can be observed and picked up by learners cooperating with the expert. Moreover, apprentices sequentially go through the different steps toward the product of work in a holistic way and thus experience the meaning and usefulness of the different skills and pieces of knowledge required on the way. Likewise, learning in academic domains arranged in similar ways, such as by way of projects run in class, are also more likely to provide students with experiences that help them understand all of the intermediate steps needed as they move toward the result of their activity.

Apprenticeship learning typically follows a sequence that resembles the acquisition of procedural knowledge (Anderson, 1983) such as when learning motor skills in a sports discipline. In the beginning phase, an expert performs the different steps in the accomplishment of a task and tries to make them explicit. At the same time, he or she acts as a model to be

observed by the student. Then, the student becomes more active and performs these steps with the help of the expert. Subsequently, the expert's help steadily decreases. Finally, the learner works autonomously observed by the expert, who can comment or help if necessary. This could happen by way of direct support or by restructuring the tasks to adjust them to the level of a student's competence.

Several strategies, described by Collins (1991), play a crucial role in apprenticeship learning. Three strategies directly refer to the acquisition of knowledge in class: modeling, scaffolding, and coaching. When modeling, which is typical of the initial phase of learning, a teacher performs the tasks so that they can be observed by students. For modeling to be successful, it is important that the teacher explicitly describes and explains the purpose of what he or she is doing. This should include – as far as possible – conveying aspects of procedural knowledge involved in a declarative form. For instance, experts could verbalize considerations or problem-solving strategies that are not typically stated aloud. The method of scaffolding sets the learner into the center of activity. Depending on the progress in learning, the task is done by the learner observed and supported by the expert, who, in turn, gives feedback on the performance, hints about how to improve it, and may take over again parts of the task not yet in reach of the learners' abilities. With a decrease of direct expert activity (sometimes referred to as fading), the method of coaching becomes more important in the process: that is, continued observation of the learner so that the teacher can give helpful feedback to allow the student to accomplish the task and find solutions for upcoming problems as autonomously as possible.

Two further methods described by Collins (1991) are designed to raise the learners' awareness of their activities and improve the strategies they use to accomplish their tasks. Articulation by way of loud thinking or elicited by systematic questions helps students getting a clearer idea of strategies employed and helps learners to better organize their conceptual representations. Reflection, namely comparisons of learners' own performance with experts' performance, is also expected to result in an improved understanding. The method of exploration is meant to take the learners even one step further toward the expert level by letting them try to solve problems autonomously and to master unforeseen challenges.

Throughout the learning process, apprenticeship learning takes place in a social situation and builds on cooperation. What happens in the interaction between expert and learner is a process of coconstruction. It can, but does not necessarily, include cooperation between several learners such as conceptualized as cooperative learning.

While the apprenticeship system of vocational training is mostly a prototypical instance of situated learning (Lave and Wenger, 1991) that is taking place in an everyday work context solving real-life problems, this is obviously not the modal case of school learning in the framework of the cognitive apprenticeship model. At the same time, apprenticeship learning in school can be assumed to benefit from reaching beyond the acquisition of textbook knowledge detached from the real world. For learning to be experienced as being meaningful, a contextualization of the knowledge and skills to be acquired is helpful. To this aim, a selection of tasks and challenges reflecting everyday problems and challenges would be beneficial as would be the inclusion of approximations of real-life situations by project work and role play. By the same token, apprenticeship learning does not imply an actual extension of the learning setting to include real-world contexts, but elements such as field trips included in the curriculum would be likely to support the construction of meaning among the learners as do internships in school-based career guidance.

See also: Academic Engagement; Mastery Learning; Motivation, Learning, and Instruction; Motivation: Life Course and Sociological Perspectives; Personal Projects; School Burnout and Engagement: Lessons from a Longitudinal Study in Finland; Self and Emotional Development in Adulthood and Later Life; Self-Determination Theory; Self-Regulated Learning: Theories, Measures, and Outcomes; Self-Regulation in Adulthood; Vocational Interests, Values, and Preferences, Psychology of.

Bibliography

Anderson, J.R., 1983. The Architecture of Cognition. Harvard University Press, Cambridge.

Beicht, U., Krewerth, A., Eberhard, V., Granato, M., 2009. Viel Licht – aber auch Schatten (Much Light – But Also Dark Spots). BIBB-Report 9/09, 1–16.

BMBF, 2012. Ausbildung und Beruf (Education and Vocation). BMBF, Berlin.

BMBF, 2013. Berufsausbildungsbericht 2013 (Report of Vocational Education 2013). BMBF, Berlin.

Bronfenbrenner, U., 1979. The Ecology of Human Development. Harvard University Press, Cambridge.

Busemeyer, M.R., September 2011. Berufliche Bildung aus Sicht der v-ergleichenden Wohlstandsstaatenforschung (Vocational education under the perspective of comparative research on welfare states). Paper presented at the BIBB Congress, Berlin, Germany.

Berlin, I.H.K., 2013. Berufsbildungssysteme in Europa (Systems of Vocational Education in Europe). http://www.ihk-berlin.de/aus_und_weiterbildung/bildungspolitik/Internationalisierung_der_Berufsausbildung/813740/EU_Laendervergleich.html (retrieved 05.08.13.).

Collins, A., 1991. Cognitive apprenticeship and instructional technology. In: Idol, L., Jones, B.F. (Eds.), Educational Values and Cognitive Instruction: Implications for Reform. Erlbaum, Hillsdale, pp. 121–138.

Collins, A., Brown, J.S., Newman, S.E., 1989. Cognitive apprenticeship: teaching the craft of reading, writing, and mathematics. In: Resnick, L.B. (Ed.), Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser. Erlbaum, Hillsdale, pp. 453–494.

Driesel-Lange, K., Hany, E., Kracke, B., Schindler, N., 2010. Ein Kompetenzentwicklungsmodell ür die schulische Berufsorientierung. In: Sauer-Schiffer, U., Brüggemann, T. (Eds.), Der Übergang Schule – Beruf (The School-to-Work Transition). Waxmann, Münster, pp. 157–176.

Gries, J., Lindenau, M., Maaz, K., Waleschkowski, U., 2005. Bildungssysteme in Europa. Kurzdarstellungen (Educational Systems in Europe. Short Descriptions). ISIS, Berlin.

Grünewald, U., Moraal, D., 2001. Duale Ausbildungssysteme (Dual Vocational Training Systems). Bertelsmann, Bielefeld.

Hamilton, S.F., 1990. Apprenticeship for Adulthood: Preparing Youth for the Future. Free Press, New York.

Hamilton, S.F., Lempert, W., 1996. The impact of apprenticeship on youth: a prospective analysis. Journal of Research on Adolescence 6, 427–455.

Härtel, P., 2013. Professional Orientation in Schools in a European Comparison. http://www.lifelongguidance.net/llg.nsf/All/85E4F1ED614570ABC12572840028792F/$file/3_Professional%20orientation%20in%20schools_format.pdf (retrieved 12.08.13.).

Kerschensteiner, G., 1912. Der Begriff der Arbeitsschule (The Concept of Work School). Teubner, Leipzig.

Kracke, B., 1997. Parental behaviors and adolescents' career exploration. Career Development Quarterly 45, 341–350.

Kracke, B., Schmitt-Rodermund, E., 2001. Adolescents' career exploration in the context of educational and occupational transitions. In: Nurmi, J. (Ed.), Navigating through Adolescence. Routledge, New York, pp. 141–168.

Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, Cambridge.

Pahnke, A., Icks, A., Kay, R., 2013. Übernahme von Auszubildenden – betriebsgrößenspezifische Analysen (Keeping Apprentices – Analyses Depending on Company Size). Institut für Mittelstandsforschung, Bonn.

Renkl, A., Mandl, H., Gruber, G.H., 1996. Inert knowledge: analyses and remedies. Educational Psychologist 31, 115–121.

Apprenticeship, Anthropological Aspects

Roy Dilley, University of St Andrews, St Andrews, Scotland, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Recent studies of apprenticeship have raised the method of skill-based learning in anthropology to new levels of sophistication and applied it to new areas of social life. Three predominant themes arise from this body of work: (1) apprenticeship as an indigenous form of social organization for learning specific skills and crafts, (2) apprenticeship as the way in which anthropologists have themselves undergone specific training as part of a fieldwork research method, and (3) how apprenticeship provides an opportunity to learn about learning in a variety of cultural settings, which has led to anthropological debate about theories of learning and practice.

The last 10 years have witnessed a blossoming of publications in the field of anthropological treatments of apprenticeship. When in 2001, I wrote an entry on apprenticeship for the first edition of this Encyclopedia, I commented that "over the last 10 years or more it [apprenticeship] has received much more sustained attention." (This followed a period of relatively sporadic and piecemeal contributions to the subject in the years prior to that.) That sustained attention of nearly 15 years ago has grown steadily and has now resulted in apprenticeship fast developing into one of the key areas of anthropological method and of anthropological interest about how people learn what they learn. Many new and exciting full-length studies have emerged since 2001, and apprenticeship among all manner of people – such as craft and occupational groups, trade specialists, athletes and artists, and ordinary folk engaged in everyday activity – is an ethnographically rich and theoretically innovative domain of study. Recent studies of apprenticeship have raised the method of skill-based learning in anthropology to new levels of sophistication and have developed the method into new areas of social, cultural, and economic life. The method has become almost canonical in the discipline, and has been employed widely by a range of anthropologists and other field-based researchers.

Three predominant themes in anthropological approaches to apprenticeship are still apparent in much of the contemporary literature: (1) apprenticeship as a form of social organization occurring in a variety of social and cultural contexts, (2) apprenticeship as an anthropological method of field research, and (3) apprenticeship as a domain for the analysis of social, cognitive, and bodily processes of learning. What has emerged quite clearly from a range of studies, however, is that forms of apprenticeship and modes of learning are not unchanging over time or across cultures. They are intimately linked to the economic, political, and social conditions in which learning takes place. The three aspects highlighted here will be dealt with in turn, but first some remarks to indicate how the concept of apprenticeship has been formulated in the West.

The subject of apprenticeship comes with a good deal of conceptual baggage derived from European and North American history. An institution known in ancient Greece and Rome, as well as from the early history of the Middle East, apprenticeship in Europe in the Middle Ages was regulated by craft guilds; specifically in Britain under sixteenth-century Elizabethan statutes, which were later repealed in the early nineteenth century. As a nineteenth-century institution, it received particularly bad press in Britain and the United States, where it was used as a means of poor relief for pauper children who were bound and indentured into trades so as to relieve local communities of the burdens placed on them by the poor. The lives of such children have been described as 'at best a monotonous toil, at worst a hell of human cruelty.'

Anthropological perspectives on apprenticeship and the use of child labor in non-Western cultures are confronted today by a not dissimilar problem: what relationship should be drawn between the condemnation of the use of child labor and the frequently accompanying economic poverty, on the one hand, and the recognition on the other hand that in some cultural contexts children are not excluded practically or ideologically from relations of production (see Nieuwenhuys, 1996). The dilemma over how to approach this problem is informed in part by the history of the institution in the West, and by our knowledge of the abuses to which it has been put in the past. In order for apprenticeship to be viewed as a cross-cultural concept, therefore, it has to be disentangled from our own conceptual baggage. This is not a simple or easy task.

Apprenticeship is a key institution, then, in the West's history of craft production, of capitalist industrialized manufacture, and of the transition between the two. Although Adam Smith regarded it as an archaic institution whose abolition should be welcomed, Karl Marx saw it as a form of organization that protected a skilled workforce from the logic of machine manufacture and from the subsequent deskilling of labor (Marx, 1946). Apprenticeship, viewed in the context of craft production or of industrial manufacture, is a mode of organization for acquiring trade skills and for the supply of able practitioners in a trade. (Contemporary politicians in the United Kingdom and the United States have recently started to proclaim the virtues of apprenticeship in order to fill gaps in the labor market and to maintain continuity of the supply of skilled personnel.) It also involves a relationship between a novice, who is usually a minor, and a master (usually not the novice's parent) who teaches a trade and who is recompensed for the instruction given in the form of the novice's products. Furthermore, the master stands *in loco parentis* to the novice, for he or she is responsible for the

minor's moral development and general conduct, as well as for furnishing his or her board and lodgings.

Apprenticeship as a Form of Social Organization

M. Coy's edited collection on the topic (Coy, 1989) was one of the first works by a group of anthropologists to treat apprenticeship explicitly as a social institution in comparative cross-cultural perspective. His definition of it echoes that given in the preceding paragraph, but he suggests also that it is 'rite of passage' involving 'specialized' and 'implicit' knowledge; also he emphasizes observation and participation as a mode of learning about a craft, about social relations, and about the social self and forms of cultural identity (Coy, 1989: pp. xi–xii). Marchand, however, provides us with a more recent and more nuanced description of what apprenticeship involves:

> In the exchange of 'toil' for 'ethnographic knowledge', fieldworkers are exposed viscerally to the learning environments and livelihoods of fellow workers, craftspeople and athletes … By adopting an explicitly apprentice-style method, or by actively engaging in the practices they study, [anthropologists] have developed their individual understandings about learning and knowing by 'doing' what they study. Cultivating such understanding … demands long immersion, perceptual and kinaesthetic awareness, careful reflection, persistent questioning, and a constant probing of the complex and multiple factors that constitute any field of practice. (Marchand, 2010: S10)

Being apprenticed to a master and directly participating in the practical activities of making and doing, field researchers are able to acquire first-hand experience of, and even some level of expertise in, the practices they describe, analyze, and theorize.

Ethnographic examples in Coy's volume range from North to South America, from Africa to the Far East. One of the major conclusions of the volume is that apprenticeship as a mode of social organization "displays more similarity cross-culturally and historically than any of us realised" (Coy, 1989: p. xv). Goody (1989) goes on to theorize about these commonalities, arguing that apprenticeship is a characteristic of social systems undergoing increasing differentiation in a division of labor that has breached the limits of domestic, kin-based production; systems that are undergoing structural changes by virtue of entry 'into the market.' Although there is a debate about the factors that determine the similarity of forms that apprenticeship takes across the world, some anthropologists have highlighted as well the unique features it takes in particular contexts (see Singleton in Coy, 1989; Kondo, 1990).

Coy's volume set a contemporary benchmark for developments in the study of apprenticeship: first, it provided a set of comparable ethnographic descriptions of apprenticeship systems in a range of cultures; second, it examined in depth the idea of apprenticeship as an anthropological field method. More recent work on the topic has moved debates along in a number of directions by asking such questions as: Is apprenticeship a means of imparting knowledge or of restricting it (cf. Singleton in Coy, 1989, on 'stealing the master's secrets')? What are the relationships of dominance operating within, and indeed outside of, apprenticeship systems (Herzfeld, 1995)? Is there an objective body of knowledge that is passed on to novices or is knowledge negotiated and situational (Lave and Wenger, 1991; Keller and Keller, 1996)?

The first two questions introduce the issue of power into the organization of apprenticeship, viewed as a means of the transmission of knowledge and skills to future generations. Training apprentices produces potential future competition in the form of skilled practitioners who might one day, once they are economically independent, take trade away from their former master. These issues must be seen also within the purview of the wider political economy in which the reproduction of a trade does or does not take place. The third question about the form that knowledge takes relates to whether specialists possess a body of knowledge that exists in the conventional terms in which we often think about it. Even if knowledge does not exist in the form of books, databases and so on, but in the form of oral traditions, memory, practical skill sets, and so on, the question of how the learner achieves 'knowledgeable' performance of an activity is raised in intricate and subtle ways. In many cases of apprenticeship, very little seems to be passed verbally from master to novice; in which case what is being passed on? Indeed, Trevor Marchand was struck during the course of his studies with Yemeni and Malian builders (2001 and 2009) that no templates or drawings were used in the conceptualization and realization of architectural forms. Rather, he argues, conceptualization took place through the direct physical engagement with the act of making, and ideas took shape through the use of materials and through the deployment of tools: mind and bodies and things were related within the learning environment. The development of knowledgeable practice might then be viewed in the framework of interaction and negotiation between various parties – apprentice, master, and environment – and might too be connected with bodily praxis, mimesis, and so on. These are topics dealt with more fully in the section Apprenticeship and Forms of Learning later in the article.

The significance of the master–apprentice relation is another topic of debate, especially whether it should be seen as a central characteristic of apprenticeship. Whether apprenticeship is a formal or an informal arrangement, many studies emphasize the relation between teacher and pupil as crucial to the processes and the experience of learning (e.g., Stoller and Oakes, 1987; Singleton in Coy, 1989). Cooper (in Coy, 1989) describes the constant criticism by his master of his own practice as an apprentice woodcarver, and it seems that the theme of discomfort between master and pupil is one that also recurs in many other studies (e.g., Marchand, 2001). However, following Lave and Wenger's (1991) suggestions about a 'decentred' approach that focuses on 'communities of practice' rather than single dyadic relations, some analyses have now shifted emphasis away from the master–apprentice dyad to focus on whole communities of practitioners (e.g., Palsson, 1994; Lave, 2011). The question raised here concerns how much apprentices actually learn directly from a master and how much is simply absorbed through participation with a group of skilled practitioners. There seems to be a degree of variation in the emphasis placed on master–novice relations depending on the kind of specialism being learned. Lave's studies of Liberian tailors in Monrovia, for instance,

conducted over a period of many years from the time of her first trip to the country in 1973, have highlighted the centrality of practice-based learning in which didactic teaching methods barely play any role. In her long-awaited monograph on apprenticeship, published in 2011, she analyzes the social relations created through learning among communities of fellow craft practitioners, and how she herself was immersed in this 'decentred community of practice.'

Apprenticeship as an Anthropological Field Method

The second aspect of apprenticeship is as a method of anthropological field research. An increasing number of field researchers have apprenticed themselves to practitioners of a variety of crafts, trades, and other specialties in order to acquire particular social and practical skills and to gain a deeper insight to cultural practice and processes; in short, to learn about cultural learning. (See chapters by Cooper, Coy, Dilley, and Dow in Coy, 1989 as examples, or Keller and Keller, 1996; Marchand, 2001, 2009, Stoller and Oakes, 1987; Lave, 2011; among others.) The intense interaction entailed by such participation provides a point of entry into a community and a way of learning through practice. Types of knowledge, secrets, and specific skills are also accessed, and these might otherwise remain hidden to investigations by other methods. Tedlock (1992) described how she and her husband used the method of apprenticeship successfully to learn the art of Mayan divination.

Recent studies of a wide range of crafts and other specialist activities have broadened the focus on the type of communities of practice considered by anthropologists. For instance, Venkatesan's study (2009) among Labbai mat weavers in the town of Pattamadai in South India focuses on low-status female craft workers, who are not driven by a passion to learn weaving but by a need to take it up primarily for economic reasons. Within the context of the anthropology of work, he analyzes the relationship between the processes by which a skill is learned on the one hand, and knowledge about the possibilities and ramifications of the skill on the other hand. Their labors cause the women physical pains and debilitating problems in the long term, and their regular participation in the activity exposes them to wider social values and attitudes, and reinforces gendered identities and social class positions. Girls learn weaving at home in a piecemeal fashion of instruction, observation, imitation, and 'having a go,' a training that is devoid of any formalized elements. Although the mats are woven by poor Muslim women of low-social status, their products are bought by elite Tamil Hindus and other affluent sectors of the population. Paradoxically, the women are lauded as producers of fine traditional craft objects, but are marginalized because of the occupation they follow.

One of the more remarkable studies of apprenticeship of the last decade has been Loïc Wacquant's intriguing, agonizing, and revealing investigation into the world of boxing as practised at a gym in a black neighborhood in Chicago's south side (Wacquant, 2004). For 3 years, he immersed himself in the pugilist's art, never intending at the outset to step into the ring but eventually succumbing to the lure of the gloves. His training regime was strenuous, the discipline severe; the drills of shadow-boxing and sparring against other amateur boxers transformed his body, his mind, and his awareness of himself. Once trained and fit, he entered the Golden Gloves tournament, the city's most prestigious amateur event, as a competitor, and he gives us a compelling account of his participation in it. Nicknamed 'Busy Louie' by the owner of the gym, Wacquant established a sound reputation for himself as a boxer, and he learned not only the art of boxing but also how the gym worked as a social space, and how the controlled violence of the gym stands in contrast to the violence on the streets of Chicago's south side. While his reporting of events, personalities, and social relations of Chicago's world of boxing is evocative and persuasive, Wacquant develops a theoretical line of argument that attempts to capture a 'carnal sociology' within a theory of practice. He does so by developing Bourdieu's concept of habitus, which embraces the idea of embodied knowledge, the type acquired through his training as a boxer. I will return to this theme.

Apprenticeship as a field method is not without its critics, who point out the methodological dangers of compromising 'objectivity' and succumbing to partiality. Few anthropologists, however, rely totally on this method and often use other techniques to investigate wider, macro social processes beyond the workshop or specific context of learning. Cooper discusses the 'fiction' of apprenticeship: the requirement for fieldworker and fellow practitioners to enter into a pretence and to know when the pretence ends. A kind of 'enforced schizophrenia prevails' (Cooper, 1989: p. 138). The 'enforced schizophrenia' inherent in the method allows for a continual methodological movement in and out of roles, and indeed sets up a distance from, as well as an engagement with, the subjective and experiential aspects of being apprenticed. Apprenticeship has of course much in common with traditional participant-observation, in which the fieldworker is as much a kind of 'cultural apprentice' as he or she is a detached observer (Coy, 1989; see also Goody, 1989; Keller and Keller, 1996; Palsson, 1994 on this comparison). Apprentice participation is thus an extension of traditional anthropological methods of participation and observation, and is pertinent to the investigation of particular kinds of activities that may be closed, specialized, or in some way inaccessible. The method also opens up areas of investigation about the nature of social and cultural learning in specific places; and this is a topic that can only be studied by researchers actually having a go at the process in question.

Apprenticeship and Forms of Learning

The third aspect of apprenticeship is the way it has been linked to the anthropology of education (see Pelissier, 1991) and to the analysis of cultural theories of learning. Lave and Wenger (1991) have reworked this area to propose a shift away from the study of apprenticeship per se to the more general idea of 'situated learning' and 'legitimate peripheral participation,' in which the learning process is embedded in 'communities of practice.' The analysis of cognitive activity is,

therefore, firmly placed back in the domain of everyday life. Their work reviews a range of learning contexts from Yucatec midwifery to the US Navy, and it raises questions about the relationship between participation and knowledge.

This relationship can be seen as problematic in that much of the knowledge learnt in such contexts is not verbalized and indeed is nonpropositional in nature. What is learnt is gained through mimesis, practice, and repetition of tasks performed routinely by skilled practitioners. Bodily knowledge, bodily techniques, manual skills, and aesthetics have now become the focus of study. William Hanks made a similar point some time ago in his foreword to Lave and Wenger's *Situated Learning* when he suggested that the apprentices' ability to understand a master's performance depends not on their possessing the same representation of it, or of the object it entails, but rather on they being engaged in the performance in congruent ways (1991: p. 21). Examples of this type of approach are Marchand's studies of master builders and of British woodworkers (2001, 2009, 2010, in preperation), in which he talks of the production of knowledge through the 'shared performance' of everyday mundane occurrences; or Greg Downey's work among skilled practitioners of Brazilian capoeira (2005); or Wacquant's examination of his own apprenticeship as a boxer among prizefighters in Chicago (1998 and 2004). Novices learn through their bodies as well as their minds, and the bodily discipline of a trade also creates moral and aesthetic sensibilities. Moreover, these learning processes are not just about practical skills or producing objects, but are also concerned with the creation of cultural identities and selves. As Kondo points out with regard to Japanese artisans, work involves self-realization, a "polishing of self through hardship," for "a mature artisan is a man who, in crafting fine objects, crafts a finer self" (Kondo, 1990: p. 241). These studies suggest, therefore, that the rigid Cartesian distinction between mind and body, thought and action, needs to be critically reexamined.

Studies by a number of anthropologists into the way in which minds, bodies, thoughts, action, tools, and environment come together in learning new skills and in perfecting the performance of them have led to a debate about how best to theorize practice, process, and structure. Recent work in particular contains a critique of Bourdieu's concept of habitus as a means through which embodied learning occurs. This critique revolves around the central idea that practice-based learning involves imitation and mimicry; a performer imitates a practice and repeats it until the action has been made habitual and perfect. But this is not blind mimicry, an unthinking imitation, a system of rote learning without conscious engagement with the process of skill acquisition. Repetition is not repetitive; rather embodied learning involves awareness by the practitioner of his or her movements and gestures, and an opportunity to reflect on, alter, and finesse the performances of an action as it is reworked and replayed. (It is here that the possibility of innovation and creativity can emerge.) In other words, there is a continual and conscious process of evaluation and adjustment, and this idea challenges the central assumption of practice theory: namely, that mastered practices are executed unreflectively. Anna Portisch's study of felt carpet production among Kazakh women in western Mongolia effectively illustrates the point that carpet-makers keep a number of aspects of the repetitive process of production in their awareness at the same time (2010). The limitations of practice theory, as suggested here, have been pointed out also in the works of Marchand (2010) and Downey (2005).

Greg Downey conducted research among practitioners of capoeira for more than 10 years, taking lessons in Brazil from masters who passed on their expertise in spoken and unspoken ways. In his book entitled *Learning Capoeira: Lessons in Cunning from an Afro-Brazilian Art* (2005), Downey examines the effect that bodily training in this form of dance and martial art has upon the performers' perceptions of self, of others, and of their social interactions within the 'ring,' or *roda*, where the game takes place, as well as outside the ring in their daily lives. Downey takes his analysis beyond Bourdieu's theory of practice by arguing that habitus is not beyond the realm of conscious awareness, and by drawing on recent developments in psychology and neurosciences that deal with the effects of physical education on human perception and self-awareness.

One of the major developments in the study of apprenticeship, therefore, has been that although anthropologists have long acknowledged that people (including themselves) learn through participation, they have now come to shed light on the cognitive processes taking place during learning, and the way in which mental representations are constructed through skilled performance, and how in turn those representations inform performances.

Trevor Marchand has emerged as one of the leading figures in the study and analysis of apprenticeship as a method of anthropological inquiry, and as an avenue into the investigation of the process of learning. He conducted his first apprenticeship among minaret builders in San'a, Yemen, which resulted in a monograph published in 2001 entitled *Minaret Building and Apprenticeship in Yemen*. Later, he worked as a laborer and an apprentice as he conducted research among mud masons in the town of Djenne, Mali, which was published in a monograph entitled *The Masons of Djenne* in 2009. Marchand's apprenticeship allowed him not only to come to an appreciation of how learning occurs through participation among masons, but also to glimpse into the esoteric arts that surround the performance of masonry, arts that are used to protect craftsmen and laborers at work and ensure the safe and proper conduct of the occupation. His third project begun in 2005 is still ongoing and it employs similar apprentice-style field methods to the earlier studies. This research, however, is on carpenters and furniture makers in London, and it engages with skill-based practice learning among trainees at the Building Crafts College in East London. His ethnography details social relations, professional aspirations, and economic challenges faced by his fellow woodworkers as they attempt to gain the skills necessary to set themselves up as professionals. The research findings will appear in a forthcoming book entitled *The Pursuit of Pleasurable Work*. This project (along with his earlier ones) has offered Marchand an opportunity to develop a theory of embodied cognition and communication, which has drawn on ideas taken from current research in neuroscience. His aim is to grasp the relationship between the brain, the human hand, and the tools used in carpentry, and he has proposed new methods of data collection to meet this challenge. Going

beyond the use of participant observation, interviews, and so on, Marchand has employed in particular digital recording technology and motion-capture analysis software to produce a database for examining tool use, human gesture and movement, and the expression of craft skills in the making. His goal is to advance a refined anthropological theory of embodied knowledge that directly challenges the nature of the concept of habitus as conceived in Bourdieu's practice theory.

Apprenticeship has looked specifically at how ways of knowing are constructed and how individuals participate in social relations of learning to acquire new skills, abilities, expertise, and knowledge. What has not received so much attention is how these relationships of learning are predicated on conceptions of ignorance, and how the process of learning and knowing mutually construct correlated domains of ignorance and nonknowledge. While much research over the last 20 years or more has quite rightly and very productively been conducted into the ways in which knowledge is made and how people learn skills and construct their own ways of knowing, there is a flip-side to all of this. And this flip-side concerns not-knowing, nonknowledge, and ignorance. Not-knowing is a precondition for the organization and distribution of relations of learning within specific contexts. It is not only the ground on which learning takes place (knowledge in one sense banishes ignorance), but it also informs the larger field of relations about who knows what and, as importantly, who is ignorant of what. For example, weavers in Senegal, West Africa, learn the skills of weaving in the context of formalized apprenticeships. But members of other social ranks who would not consider it their proper role to practice such a craft – indeed, they would deem it beneath them to do so – desire to remain ignorant of the techniques, skills, and abilities of weaving. This ignorance can be 'real' in the sense that nonweavers are usually excluded by master weavers from entry into an apprenticeship; but this ignorance can also be contrived in the sense that those living cheek-by-jowl with weavers may observe, pick up, and absorb the ways of the weaver by simply being in their presence for long periods. These 'voyeurs' of the weaver's art, as it were, often feel compelled to ignore what they have witnessed; they turn their eyes away from what is there for them to observe. Their ignorance maintains an appropriate social boundary and allows for the pretense of the monopoly on craft skills being contained within the domain of formalized apprenticeship.

The concept of apprenticeship is a matrix connecting a set of keystone issues within the contemporary frame of social and cultural anthropology. When employed in cross-cultural comparison, it demands explanations for its apparent commonalities across a wide range of social settings as well as for its cultural particularities specific to place or time. The practitioners of apprentice field methods have raised questions about the delicate balance between subjective participation and distanced observation that lie at the heart of empirical fieldwork procedures: indeed, apprenticeship suggests a way of bridging this distinction. As a domain of inquiry, apprenticeship highlights cultural conceptions of knowledge and the processes of learning, of embodiment and bodily praxis, of self and identity formation. Also it highlights the need to examine the connections between knowledge and power in dynamic interactive contexts embracing both individual practitioners situated in different social roles and communities of skilled actors whose practices create wider webs of relations. These concerns about knowledge and power extend, moreover, beyond the parochial boundaries of the workshop or place of learning to the broad social world in which apprenticeship takes place.

See also: Body: Anthropological Aspects; Craft Production, Anthropology of; Education: Anthropological Aspects; Fieldwork in Social and Cultural Anthropology; Knowledge, Anthropology of; Practice: An Anthropological Approach.

Bibliography

Cooper, E., 1989. Apprenticeship as field method: lessons from Hong Kong. In: Coy, M.W. (Ed.), Apprenticeship: From Theory to Method and Back Again, pp. 137–148.

Coy, M.W. (Ed.), 1989. Apprenticeship: From Theory to Method and Back Again. State University of New York, Albany.

Downey, G., 2005. Learning Capoeira: Lessons in Cunning from an Afro-Brazilian Art. Oxford University Press, New York.

Goody, E.N., 1989. Learning, apprenticeship and the division of labour. In: Coy, M.W. (Ed.), Apprenticeship: From Theory to Method and Back Again, pp. 233–256.

Herzfeld, M., 1995. It takes one to know one. In: Fardon, R. (Ed.), Counterworks: Managing the Diversity of Knowledge. Routledge, London, etc, pp. 124–142.

Keller, C.M., Keller, J.D., 1996. Cognition and Tool Use: The Blacksmith at Work. Cambridge University Press, Cambridge, etc.

Kondo, D., 1990. Crafting Selves: Power, Gender and Discourses of Identity in a Japanese Workplace. University of Chicago Press, Chicago and London.

Lave, J., 2011. Apprenticeship in Critical Ethnographic Practice. University of Chicago Press, London, Chicago.

Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Practice. Cambridge University Press, Cambridge, etc.

Marchand, T., 2001. Minaret Building and Apprenticeship in Yemen. Curzon, Richmond Surrey.

Marchand, T., 2009. The Masons of Djenné. Indiana University Press, Bloomington.

Marchand, T. (Ed.), 2010. Making knowledge. Special Issue of the Journal of the Royal Anthropological Institute (NS).

Marchand, T., in preperation. The Pursuit of Pleasurable Work: Studies in the Art of Furniture Making.

Marx, K., 1946 [1887]. Capital: A Critical Analysis of Capitalist Production, vol. I. George Allen and Unwin Ltd, London.

Nieuwenhuys, O., 1996. The paradox of child labor and anthropology. Annual Review of Anthropology 25, 237–251.

Palsson, G., 1994. 'Enskilment at sea', man. Journal of the Royal Anthropological Institute 29 (4), 901–927.

Pelissier, C., 1991. The anthropology of teaching and learning. Annual Review of Anthropology 20, 75–95.

Portisch, A.O., 2010. The craft of skilful learning: Kazakh women's everyday craft practices in western Mongolia. In: Marchand, T. (Ed.), Making Knowledge, pp. S62–S79.

Stoller, P., Oakes, C., 1987. In Sorcery's Shadow: A Memoir of Apprenticeship among the Songhay of Niger. The University Press of Chicago, Chicago and London.

Tedlock, B., 1992 [1982]. Time and the Highland Maya. University of New Mexico Press, Albuquerque.

Venkatesan, S., 2009. Craft Matters: Artisans, Development and the Indian Nation. Orient Blackswan, Hyderabad.

Wacquant, L., 1998. The Prizefighter's three bodies. Ethnos 63 (3), 325–352.

Wacquant, L., 2004. Body and Soul: Ethnographic Notebooks of An Apprentice-Boxer. Oxford University Press, New York.

Apprenticeships and School Learning: General Considerations

Hans Gruber, Institute of Educational Science, University of Regensburg, Regensburg, Germany
Heinz Mandl, Department Psychology, Ludwig-Maximilians-Universität, München, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by H. Gruber, volume 8, pp. 601–604, © 2001, Elsevier Ltd.

Abstract

Qualification on the one hand and integration and enculturation on the other hand are among the most important functions of schools for the society as well as for the individuals. School learning thus should teach students facts and skills that are necessary for later life. However, evidence exists that school learning often is far away from application situations out of school. Critics have argued that it frequently leads to knowledge that is inert and thus cannot be used for solving real-world problems. A plausible reform idea is to make school learning resembling learning out of school. This article reviews arguments and evidence for and against this proposition.

Apprenticeship Models as Solutions for Problems with School Learning

Resnick (1987) identified four principles in which learning in and out of school differ: (1) Individual cognition in school versus shared cognition outside. Learning in school focuses on individual performance; that is, students have to work in isolation. In contrast, learning out of school usually focuses on shared knowledge and cooperative problem solving. (2) Pure mentation in school versus tool manipulation outside. School learning focuses on abstract mental activities that have to be done without the use of any external support like one's own notes or Internet research systems. In contrast, learning out of school heavily relies on individuals' ability to use adequate tools in an adequate way. (3) Symbol manipulation in school versus contextualized reasoning outside school. With its focus on symbol-based reasoning, school learning often lacks close connections to events and objects in the daily world that are characteristic for learning out of school. (4) Generalized learning in school versus situation-specific skills outside. School learning aims at the acquisition of general, widely usable principles, whereas learning out of school focuses on solving problems that actually arise at places and contexts the individual is situated in.

The analysis of discrepancies between learning in and outside the school includes criticisms of present school instruction concerning both the qualification function of school and its integration and enculturation function. Similar arguments were brought forward by the German 'Reformpädagogik' a century ago. Alternative instructional models developed in these years are closely related to the proposals nowadays made by situated learning theorists. One common principle is that students participate in more advanced individuals' activities and thus increasingly get part in the community of practice. In such *apprenticeship models of learning*, besides knowledge students have to acquire the ways of thinking in communities of practice, they have to become part of what constitutes the respective epistemic culture (Knorr-Cetina, 1999).

A number of situated learning models have been developed in order to decrease the discrepancy between learning in and out of school and to avoid the emergence of inert knowledge (Renkl et al., 1996). In each, students learn within complex contexts like apprentices by solving authentic problems in a community of practice. The approaches are based on the idea that knowledge is socially shared so that plain teaching of 'objective' knowledge does not suffice because each working situation includes demands for adaptive problem solving (Derry and Lesgold, 1996). The view that knowledge in principle is situated in the environment within which it was acquired is the perspective of *situated learning*.

Theoretical Background: 'Reformpädagogik' and Situated Learning

Many aspects of the mentioned situated learning approaches resemble concepts of the *Reformpädagogik*. When the agents of this German educational movement expressed their ideas a century ago, they were well aware of developments in the Deweyan Chicago school (e.g., the project plan; Kilpatrick, 1922), and vice versa. Reformpädagogik criticized the 'book school' in a similar way the school learning nowadays is criticized. The central counter-idea then was the 'work school,' most prominently proposed by Kerschensteiner (1912). He emphasized the importance of manual work that should be closely tied to mental work. Among his school reforms was the introduction of school kitchen instruction and school garden instruction. Based on authentic activities, students learned chemistry, physics, physiology, and mathematics. Instructional questions usually were not posed by the teacher, but rather by the students when working on problems. Such self-regulated activity reduced the discrepancy between school and everyday life. Work school no longer was an *Ersatzwelt* in which *ersatz activities* had to be carried out. Since the problems posed were of relevance to the students' life outside school, besides skills and knowledge learners acquired relevant social attitudes and a sense of responsibility as members of society.

Kerschensteiner (1912) was extremely concerned about the question whether school prepared for later professional life. He thus initiated the German vocational training system, the *dual system*, which is a form of training tied to the workplace with

supplementary teaching in the compulsory, part-time vocational school. In this system, not only manual work is stressed, but the notion of work can also be ascribed to cognitive learning processes or, in terms of Gaudig (1922), 'free mental work.'

Many of the Reformpädagogik ideas were theoretically reconsidered in the situated learning movement under cognitive psychological and constructivist perspectives. Five major instructional principles can be identified (Gruber et al., 1995).

Learning by Solving Authentic Problems

In problem-oriented learning, an authentic, rather complex problem marks the starting point. Its function is not only to motivate students. By embedding the learning process in meaningful contexts, it also aims at knowledge acquisition through application instead of abstract teaching. Knowledge is conditionalized to application conditions from the beginning.

Multiple Perspectives

Domains are analyzed from multiple perspectives and multiple contexts in order to foster multiple conditionalization. Preventing the emergence of dysfunctional oversimplifications and fostering active abstraction processes are the main instructional functions, thus supporting flexible use of knowledge and skills.

Articulation and Reflection

Externalizing mental processes enables students to compare their own strategies with those of experts and other students. Articulation facilitates reflection and thus fosters the decontextualization of acquired knowledge.

Cooperative Format

Cooperative learning facilitates the comparison with experts and peer learners. In addition, cooperation skills are useful in later professional life.

Learning as Enculturation

A particular function of learning is to become enculturated into communities of practice. To achieve this, students have to get a sense of authentic practice. They have to acquire functional belief systems, ethical norms, and the tricks of the trade.

These principles of situated learning have been realized in several instructional models. The most convincing of these models is cognitive apprenticeship.

Microanalytic Use of Apprenticeship in School Learning: The Model of Cognitive Apprenticeship

The apprenticeship metaphor for designing situated learning arrangements has its origin in craft domains: for example, tailoring or midwifery. Collins et al. (1989) adapted the concept for cognitive domains. In cognitive apprenticeship as in craft apprenticeship, students are introduced into an expert culture by authentic activities and social interactions. Important is the sharing of problem-solving experiences between students and mentors (i.e., experts, teachers, or advanced students), thus negotiating their understandings and actions through dialogue. Since in cognitive domains mental activities prevail, the explication of cognitive processes is extremely important in order to publicly expose knowledge and thinking processes involved in the cooperative problem solving (Simons and Ruijters, 2004).

The core of cognitive apprenticeship is a particular instructional sequence. Learning takes place in sequenced learning environments of increasing complexity and diversity. In early stages of learning, mentors provide overall direction and encouragement. As students improve, mentors gradually fade support, encouraging students to work and think more independently. At all phases of apprenticeship, the mentor is assigned an important role as a *model* and as a *coach* providing *scaffolding*. However, the student has to increasingly take over an active role, as the mentor gradually *fades out*. *Articulation* and *reflection* are promoted by the mentor in order to persist on externalization of cognitive processes.

This instructional sequence supports students to increasingly work on their own (*exploration*) and to take over the role initially held by the mentor. Lave and Wenger (1991) similarly described professional learning in nonexplicit learning contexts as development from legitimate peripheral participation to full participation. Learning is thus not confined to the acquisition of knowledge or skills, but a social process of enculturation by getting a full participation in a community of practice that can cope with the problems typical for the domain in a flexible manner. Considering learners as apprentices ascribes them a novel role. From the beginning, they are active participants in authentic practices instead of passive recipients.

Empirical evidence exists that cognitive apprenticeship learning supports the acquisition of applicable knowledge in a variety of domains (e.g., programming and medicine). However, successful learning is not guaranteed. High demands are made on learners, so that careful instructional support is necessary (Gegenfurtner et al., 2009).

Macroanalytic Use of Apprenticeship in School Learning: Communities of Practice – Implementation of Apprenticeship in School Systems

Cognitive apprenticeship stresses that learning performance, like all authentic cognitive activity, is socially constituted. Learning environments inevitably are part of a complex social system, and the structure and meaning of learning processes are substantially influenced by that system. It is unlikely that an adequate understanding of learning processes can be developed without taking the social context into consideration in which learning processes are situated. Conceiving learning as becoming enculturated into communities of practice thus concerns not only the individual processes during learning but also the implementation of learning environments in large social contexts, frequently denoted as the prevailing 'learning culture.' Taking the apprenticeship idea seriously, such macroanalytic concepts of learning are inseparably connected with microanalytic analyses of learning processes and, at the same

time, learning has to be bound to application situations and has to be integrated into a large system with adequate learning culture.

The principle of the dual system of vocational training in Germany tries to fulfill these requirements. It includes simultaneous qualification at two learning locations, namely vocational school and workplace. Thus, school learning and professional practice are closely tied, so that opportunities for learning by doing are realized.

In medical education, the principles of situated learning are applied in the common teaching format of bedside teaching. The idea is that students learn in an interactive manner to treat a real patient. The authenticity of such a learning environment should provide the opportunity to acquire medical as well as interpersonal skills relevant for the profession (Eraut, 2000). The conditions of the work environment need to foster learning, and therefore the environment should be designed accordingly (Festner and Gruber, 2008; Gierk and Harendza, 2012; Spencer, 2003). In practice this is not always easy to achieve. In a study of guidance at work, Filliettaz (2011) investigated the pedagogical aspects of initial vocational training in the context of the Swiss Vocational Education and Training (VET) system. Apprentices interacted with various colleagues in the workplace, and it was examined how apprentices were guided and supported by experienced workers. The results indicate that the pedagogical quality of guidance in the workplace needs to be considered in order to increase efficiency during the apprenticeship and facilitate gradual transitions into work experiences.

Efforts to implement the dual system in countries other than Germany (Harhoff and Kane, 1997; Heikkinen and Sultana, 1997) (Austria, Switzerland, Denmark, and parts of Italy) showed that many organizational preconditions are required, such as flexibility of enterprises in the adaption of the system, an elaborated widespread school system, adequate teacher education, and so on. The dual system cannot serve as panacea if the learning culture within a society is not readily prepared. Additionally, modern development of workplaces (such as globalization of enterprises and new principles of working organization) yield difficulties for the dual system. New approaches in educational research are being developed as response to emerging imbalances between the learning locations, namely school and workplace. For example, 'discipline and location-crossed training' aims at implementing new ways of cooperation between learning locations.

Outlook: What Remains of Traditional School Learning?

The view that knowledge is essentially situational has major educational implications. Learning is regarded as a process of enculturation and means to take part in a particular community of practice, to get acquainted with the culture of the community, the jargon used, the beliefs held, the problems raised, and the problem-solving methods used. This entails the notion of apprenticeship for learning.

Traditional school learning is accused of producing inert knowledge as students fail to transfer their knowledge to tasks embedded in contexts other than school settings. Lave (1988) argued that traditionally school is treated as if it were a neutral place, with no social and cultural features of its own, in which abilities are acquired that can easily be transferred later to any other situation.

As mentioned above, there is evidence for the phenomenon of inert knowledge. However, this evidence not necessarily requires the importance of traditional school learning to be denied. Even if there exist resources other than teaching through which apprentices acquire their skills, apprenticeship is not the only effective means of learning. The academic context of school learning proves useful for a variety of learning situations. It is very likely that an interaction exists between learning purposes and preferable learning contexts. Educational psychology has to carefully decide which kind of learning context is appropriate for which kind of learning, and should take into consideration research results for the design of school learning as well as for the design of teacher education.

Thus, the notion cannot be maintained that traditional school learning should be entirely abolished. However, it is worth to analyze under which conditions apprenticeship models are an attractive alternative to traditional school learning. For which kinds of skills and knowledge, in terms of their complexity (i.e., elementary, intermediate, or sophisticated) and nature (i.e., declarative, procedural, or strategic), is apprenticeship deemed necessary? Which means help students to benefit most? These questions have to be addressed in future research in order to determine the role apprenticeship can and should play in future so that an adequate balance between traditional school learning and situated learning will be found.

See also: Apprenticeship and School Learning: Lessons from Germany; Apprenticeship, Anthropological Aspects; Communities of Practice; Cooperative Learning in Schools; Education: Skills Training; School Learning for Transfer; Situated Learning (Learning In Situ); Vocational Education.

Bibliography

Collins, A., Brown, J.S., Newman, S.E., 1989. Cognitive apprenticeship: teaching the craft of reading, writing and mathematics. In: Resnick, L.B. (Ed.), Knowing, Learning and Instruction. Erlbaum, Hillsdale, pp. 453–494.

Derry, S., Lesgold, A., 1996. Toward a situated social practice model for instructional design. In: Berliner, D.C., Calfee, R.C. (Eds.), Handbook of Educational Psychology. Macmillan, New York, pp. 787–806.

Eraut, M., 2000. Non-formal learning and tacit knowledge in professional work. British Journal of Educational Psychology 70, 113–136.

Festner, D., Gruber, H., 2008. Conditions of work environments in fostering transfer of training. In: Billett, S., Harteis, C., Eteläpelto, A. (Eds.), Emerging Perspectives of Workplace Learning. Sense, Rotterdam, pp. 215–231.

Filliettaz, L., 2011. Collective guidance at work: a resource for apprentices? Journal of Vocational Education and Training 63, 485–504.

Gaudig, H., 1922. Freie geistige Schularbeit in Theorie und Praxis (Free Mental School Work in Theory and Practice). Hirt, Breslau.

Gegenfurtner, A., Veermans, K., Festner, D., Gruber, H., 2009. Motivation to transfer training: an integrative literature review. Human Resource Development Review 8, 403–423.

Gierk, B., Harendza, S., 2012. Patient selection for bedside teaching: inclusion and exclusion criteria used by teachers. Medical Education 46, 228–233.

Gruber, H., Law, L.-C., Mandl, H., Renkl, A., 1995. Situated learning and transfer. In: Reimann, P., Spada, H. (Eds.), Learning in Humans and Machines: Towards an Interdisciplinary Learning Science. Pergamon, Oxford, pp. 168–188.

Harhoff, D., Kane, T.J., 1997. Is the German apprenticeship system a panacea for the U.S. labor market? Journal of Population Economics 10, 171–196.
Heikkinen, A., Sultana, R.G. (Eds.), 1997. Vocational Education and Apprenticeships in Europe. Challenges for Practice and Research. Tampere, Tampereen Ylioposto.
Kerschensteiner, G., 1912. Begriff der Arbeitsschule (The Idea of the Work School). Teubner, Leipzig.
Kilpatrick, W.H., 1922. The Project Method: The Use of the Purposeful Act in the Educative Process. Teachers College Press, New York.
Knorr-Cetina, K., 1999. Epistemic Cultures – How the Sciences Make Knowledge. Harvard University Press, Cambridge.
Lave, J., 1988. Cognition in Practice: Mind, Mathematics, and Culture in Everyday Life. Cambridge University Press, Cambridge.
Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, Cambridge.
Renkl, A., Mandl, H., Gruber, H., 1996. Inert knowledge: analyses and remedies. Educational Psychologist 31, 115–121.
Resnick, L.B., 1987. Learning in school and out. Educational Researcher 16 (9), 13–20.
Simons, P.R.-J., Ruijters, M.C.P., 2004. Learning professionals: towards an integrated model. In: Boshuizen, H.P.A., Bromme, R., Gruber, H. (Eds.), Professional Learning: Gaps and Transitions on the Way from Novice to Expert. Kluwer, Dordrecht, pp. 207–229.
Spencer, J., 2003. ABC of learning and teaching in medicine: learning and teaching in the clinical environment. British Medical Journal 326, 591–594.

Apraxia

Georg Goldenberg, Bogenhausen Hospital, Munich, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term apraxia refers to 'higher level' disorders of motor control. The core manifestations of apraxia affect the imitation of gestures, the performance of communicative gestures, and the use of tools and objects. All of them occur predominantly after left brain lesions and are frequently, though not invariably, associated with aphasia. In contrast to elemental 'low level' motor disturbances, they affect not only the side of the body opposite to the cerebral lesion but also the ipsilateral side.

The term apraxia refers to 'higher level' disorders of motor control. There is, however, no general agreement as to what counts as a high- or a low level of motor control. Consequently, the history of apraxia has brought forward a wide variety of diverging definitions and classifications of apraxia. The most influential of them was proposed some 100 years ago by the German psychiatrist Hugo Liepmann (Liepmann, 1908; Goldenberg, 2003). Liepmann based a classification of apraxia on a model of voluntary action that distinguished two consecutive phases. The first is the creation of mental images of the intended action and the second its transduction into appropriate motor commands. Their sequence constitutes a path leading from high to low levels of motor control. Liepmann proposed to name disturbance of the first phase 'ideational' and that of the second phase 'ideokinetic' apraxia. He equated this theoretical model with an anatomical schema. He reasoned that mental images of intended actions are predominantly visual and therefore located in posterior brain regions whereas motor commands are sent out from the precentral motor cortex. The transduction of ideas into motor actions thus corresponds to the transmission of signals through white matter tracts leading from the occipital to the frontal lobe. Parietal lesions interrupt this posterior-to-anterior stream of action control and are thus the prime cause for ideokinetic apraxia.

The opposition between mental image and motor execution of actions has not remained the only attempt to conceptualize a dichotomy between high and low levels of motor controls. Other proposals set autonomous action in opposition to environmental dependency, coping with novelty to routine action, abstract symbolic gestures to material interactions with concrete objects, or planning of goal-oriented action sequences to single actions (Pick, 1905; Denny Brown, 1958; De Ajuriaguerra and Tissot, 1969). Arguably, the dichotomy between high and low levels of action control ultimately refers to a mind–body dichotomy that constitutes a driving force behind many attempts of conceptualizing and classifying variants of apraxia (Goldenberg, 2013). The formulation of the dichotomy that fits best with the terminology of cognitive neurosciences today would be the opposition between cognitive and motor aspects of action control.

In view of the diversity of theoretical accounts of apraxia it is not surprising to note a similar diversity of opinions concerning the range of clinical symptoms that should be classified as apraxia. For example, defective imitation of meaningless gestures was central for the understanding of apraxia for some authors (Kimura and Archibald, 1974; Goldenberg and Hagmann, 1997) but considered as irrelevant by others (Geschwind and Damasio, 1985). Likewise, 'constructional apraxia' that had been a central piece of theories of apraxia in the midst of the twentieth century (De Ajuriaguerra et al., 1969) is hardly included into surveys of apraxia any more.

There is, however, a core of symptoms that have already been emphasized by Liepmann and that were recognized as belonging to apraxia by a majority of subsequent authors. They concern three domains of human actions: The imitation of gestures, the performance of communicative gestures, and the use of tools and objects. All of them occur predominantly after left hemisphere brain lesions and are frequently, though not invariably, associated with aphasia. In contrast to elemental 'low level' motor disturbances, they affect not only the side of the body opposite to the cerebral lesion but also the ipsilateral side. All results reported in this article refer to actions of the hand ipsilateral to the lesion.

Figure 1 gives an overview of critical locations for each of the manifestations. A first conclusion from inspecting them is that they vary widely and there is no single region that is involved in all of them. In other words, there is no indication of a 'praxis center' or, respectively, a 'praxis system,' disturbance of which underlies all manifestations of apraxia. We will discuss underlying mechanisms and neural substrate for each of the manifestations separately but then ask again whether there are substantial communalities unifying them.

Imitation

For the clinical examination of apraxia, testing of imitation has the great advantage of being an essentially nonverbal task. The observation that patients with left hemisphere brain damage have nonetheless more difficulties than controls or patients with right hemisphere brain damage contradicts attempt to explain apraxia as a mere consequence of aphasia. The theoretical interest goes beyond the corroboration that apraxia is an independent sequel of brain damage. To Liepmann, disturbed imitation yielded the main support for his contention that patients were unable to transmit a correct image of the intended gesture into appropriate motor commands. The view that defective imitation indicates a problem of motor execution has been upheld by influential modern authors. For example,

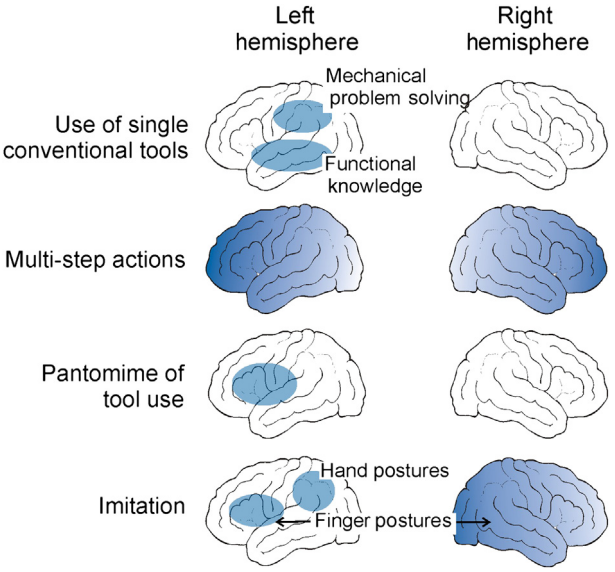

Figure 1 A summary of locations of different manifestations of apraxia: shading of the whole hemisphere does not exclude the possibility of more fine-grained localizations within the hemisphere, but indicates that there are no firm empirical data for delimiting the critical areas. Adapted from Goldenberg, G., 2007. Neuropsychologie – Grundlagen, Klinik, Rehabilitation. 4. Auflage, München, Jena: Elsevier: Urban und Fischer, p. 149, by permission of Elsevier.

De Renzi (1990: p. 246) wrote: "Since the examiner provides the model of the action and the patient must simply copy it, errors can only be due to an executive deficit."

Defective translation of the mental image of the intended gesture into motor command should affect gestures independently of the origin of the mental image. Consequently, patients who have problems with imitation should also fail when motor actions of similar complexity are produced on verbal command as, for example, when patients demonstrate pantomimes of tool use. This prediction is refuted by observations of patients with visuoimitative apraxia. They fail imitation of meaningless gestures but perform pantomimes of object use and other meaningful gestures on command flawlessly (Goldenberg et al., 1997). They may even achieve correct imitation of meaningful gestures by first understanding their meaning and then reproducing them from long-term memory.

A plausible account of these dissociations can be given by the assumption of two parallel routes for imitation. One passes via semantic memory where the meaning of a gesture is recognized and from where it can be reproduced. This route cannot support imitation of meaningless gestures because they are not represented in semantic memory. A second route connects the visual percept of the shape of demonstrated gestures directly with the motor programs for reproducing the same gesture. This route can accommodate both meaningful and meaningless gestures. As meaningful gestures can take the alternative deep route, destruction of the direct route will result in a selective deficit of imitating meaningless gestures, that is, in visuoimitative apraxia (Rothi et al., 1991; Tessari et al., 2007).

The proposal of a direct route connecting perception of motor actions with their motor reproduction has gained popularity by the detection of 'mirror neurons' that fire both when monkeys or human beings see an action and when they perform the same action (Rizzolatti et al., 2002).

The directness of the route from perception to motor execution of gestures is, however, called into doubt by the finding that patients who show apraxia for the imitation of meaningless hand postures are also impaired when they are asked to replicate the hand postures on a mannikin (Goldenberg, 1995). As the motor actions of manipulating a mannikin are very different from those of imitating the target posture on the own body, correspondence between demonstration and replication of the gesture must be established by an intermediate representation that is distinct from the motor action for achieving the target posture.

The plausibility of a direct route from perception to imitation of gestures is further undermined by the observation that defective imitation of meaningless gestures can be body-part specific depending on the location of the causal lesion. **Figure 2** shows examples of hand-, finger-, and foot postures. Whereas defective imitation of hand postures is nearly exclusively bound to left hemisphere lesions, imitation of finger- and foot postures is also susceptible to right hemisphere lesions (Goldenberg and Strauss, 2002). Within the left hemisphere, defective imitation of hand postures is a symptom of parietal lobe damage whereas defective imitation of finger postures is more likely to be due to inferior frontal lesions (Goldenberg and Karnath, 2006). The attempt to reconcile the body-part specificity of causal lesions with interruption of a direct route from perception to motor execution of gestures would lead to the rather unlikely conclusion that there are multiple direct routes distributed across both hemispheres. The improbability of such multiplication prompts the search for other possible mechanisms underlying defective imitation.

I proposed that a crucial intermediate step between perception and motor execution of meaningless gestures is body-part coding. Body-part coding segments the body into distinct body parts that are combined for creation of gestures. It thus reduces the visual manifold of the gesture to simple spatial relationships between a limited set of defined elements. For example, in **Figure 2**, the left upper panel of each kind of gesture can be coded as 'fingers 2, 4, and 5 extended,' 'thumb touches inferior side of nose with back of hand toward face,' and 'heel touches instep.' Body-part coding can accommodate novel and meaningless gestures to combinations of familiar elements and alleviates the load on working memory for holding the gesture from its perception to the accomplishment of its reproduction. Furthermore, translating the gesture's visual appearance into relationships between body parts produces equivalence between demonstration and imitation that is independent of the different modalities and perspectives of perceiving and moving one's own, other persons' or a mannikin's bodies.

According to this hypothesis, body-part coding is bound to integrity of the left inferior parietal lobe. Left frontal and right hemisphere regions come into play when imitation puts additional demands on visuospatial analysis of demonstrated gestures or on selection between perceptually highly confusable postures.

Positions of the hand relative to parts of head and face put high demands on body-part coding because they involve

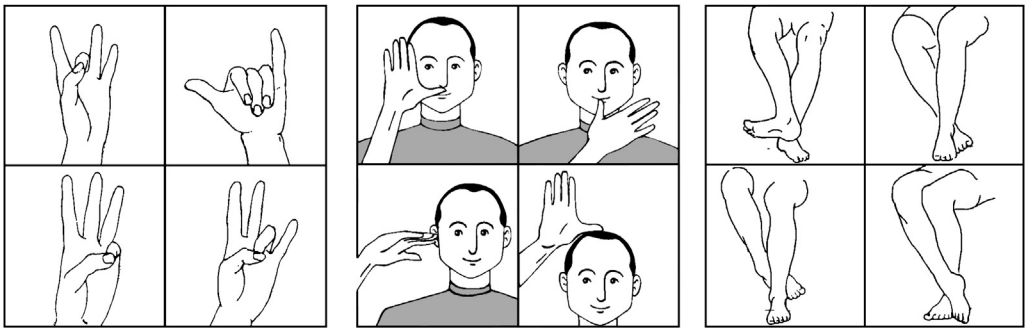

Figure 2 Examples of finger, hand, and foot postures used for imitation of meaningless gestures. Reprinted from Goldenberg, G., 2007. Neuropsychologie – Grundlagen, Klinik, Rehabilitation. 4. Elsevier, Auflage, München, Jena. Urban und Fischer, p. 145, by permission of Elsevier.

a multitude of different body parts like chin, lips, back and flanks of the nose, cheek or ears, palm and back of hands. By contrast, finger postures are, with a possible exception for the thumb, constituted by a set of uniform elements, which differ only in their serial position. Thus, body-part coding of finger postures can be reduced to numbering extended and flexed fingers according to their serial position from index to small fingers. On the other side, the body parts involved in hand postures are perceptually salient and unmistakable, whereas the uniformity of fingers makes their distinction difficult and demands evaluation of the entire hand in order to determine the spatial position of the fingers that are bent or extended (Goldenberg et al., 2009). Finally, foot gestures put high demands on both components of imitation. They involve a considerable number of conceptually different body parts like ankle, heel, calf, big toe, and little toe that are, however, perceptually less salient than the parts of the face determining the hand postures. Their perceptual analysis is further complicated by the three dimensional overlapping and consequent occlusion of sight of one leg by the other (see **Figure 2**).

In sum, modern research does not support Liepmann's contention that defective imitation testifies insufficient transformation of mental images into motor commands nor that it results from interruption of a direct route from perception to motor execution. Indeed motor mechanisms seem to play a subordinate role for this manifestation of apraxia. Although the gestures themselves are meaningless, their imitation taxes the cognitive rather than the motor side of action control.

Communicative Gestures

Hundred and fifty years ago, in the very early days of cognitive neuroscience, Paul Broca noted in his description of the aphasic and hemiplegic patient 'Tantan' that he was unable to manifest his ideas or desires other than by movements of the left hand but frequently made incomprehensible gestures (Broca, 1861). Ten years later, the German psychiatrist Finkelnburg (1870) described an aphasic woman who had been raised as a devout catholic but now failed to make the sign of the cross after the grace. "When asked by her surrounding to make it, she hesitantly reached sometimes behind the ear, sometimes to the neck until it was demonstrated to her. Then she imitated it exactly." From this and other observations of nonverbal communicative impairment of aphasic patients Finkelnburg concluded that these patients suffer not merely from loss of speech but from a more profound inability of communication, which he termed 'asymbolia.'

Liepmann assimilated the deficiency of communicative gestures into his classification of apraxia. He thought that performance of communicative gestures shares with defective imitation the need to perform movements 'wholly from memory' without support by instrumental interaction with external objects, and hence brings forward the inability to transmit a mental image of an intended action into appropriate motor command. Consequently, he subsumed disturbance of imitation and of communicative gesture under the common heading 'ideokinetic apraxia' that was later renamed into 'ideomotor apraxia' (Morlaas, 1928). The existence of 'visuoimitative' apraxia, discussed in the previous section, constitutes a strong argument against the alleged communality. It is complemented by observation of patients who fail the production of communicative gestures but can imitate flawlessly, like Finkelnburg's catholic lady (Barbieri and De Renzi, 1988).

Broca and Finkelnburg had derived their conclusions from observation of spontaneous gestures of aphasic patients. Liepmann introduced two modifications of the assessment procedure that are still followed by the majority of authors today.

The first modification was the replacement of observations of spontaneous gestures by examination of gestures on command. The production of communicative gestures on command creates a conflict between the message of the requested gesture and the concrete communicative situation. As the examiner knows in advance the message that is expressed by the patient's gesture, the actual communicative function of its production is at best confirmation of the patient's willingness to comply with the instruction. Difficulties of aphasic patients to produce gestures that are unrelated to the concrete situation have been central to the argument that they suffer from a general loss of 'abstract attitude' and that apraxia is only one manifestation of this pervasive loss (Goldstein, 1928).

A second modification was the restriction of the examination to gestures that can transmit defined messages without reference to accompanying speech. These may be 'emblems' that have a conventional meaning like thumb up for 'okay' or touching the temple with the extended hand for 'military salute,' or they may be pantomimes of tool use that indicate objects by miming their use. Usually, diagnosis and research on

apraxia concentrate on pantomime of tool use. A practical reason for this preference is that aphasic patients may not understand the verbal label of emblems, whereas comprehension of the name of the tool whose use they should demonstrate can be facilitated by showing the tool or a picture of it. A theoretically more interesting reason for the preference may be that pantomime of tool use is a communicative gesture that is derived from the material interaction with external objects. It thus spans a bridge from instrumental to communicative functions of the hands.

Theoretical accounts of defective pantomime emphasize either its derivation from real tool use or its communalities with the disturbance of verbal communication in aphasia. We will consider both aspects in turn.

Knowledge about how to use a tool is a prerequisite for pantomiming its use, but the conclusion that pantomime is equivalent to running the motor program of real use without holding an implement does not withstand scrutiny. Rather than replicating the motor program of use, pantomime segments the compound image of hand, action, and object into distinctive features of the object and the acting hand that are combined for giving a comprehensible image of the pretended object and its use. For example, when miming to grasp a glass, normal subjects open their hand to the approximate width of the pretended glass at the start of the transport movement and stop the transport at its pretended location without further changing the aperture of their grip (Laimgruber et al., 2005). The width of the aperture demonstrates the width of the glass, and stopping the transport of the hand suffices for indicating the grasp. By contrast, in real grasping, scaling of grip width to the width of the glass is achieved by first opening the hand in proportion to, but wider than, the visually perceived width of the glass and then closing it around the glass. Opening and closing are important for swift and secure grasping but have little value for discrimination of the grasped object. Pantomime may neglect them. Instead, the pantomime must specify features that in actual use are determined by the manipulated object. For example, in real drinking from a glass the distance of the holding hand from the mouth is determined by the contact of the glass with the mouth, but in pantomime a plausible distance must be chosen without external support. Arguably, it is the selection of distinctive features of the tool use action and their combination to a continuous movement of the hand that pose the greatest difficulties for patients with apraxia. Thus, the kinematic features of grasping a glass are the same in patients with apraxia as in normal controls, but their pantomimes omits the opening of the hand that indicates the width of the pretended glass (Laimgruber et al., 2005).

Selection and combination of significant features permit the construction of pantomimes of actions that one has never before executed in reality. For example, most persons are able to pantomime playing a violin or a trumpet, but only a small minority can actually perform the pantomimed action. It is even possible to pantomime actions that are out of the range of human motor competence. For example, rhythmical up and down beat of both arms beside the body will be recognized as a pantomime of flying like a bird. The need to create a gesture rather than to retrieve it ready-made from stored motor programs can also explain the paradoxical finding that aphasic patients have more problems with pantomime of tool use than with the production of emblematic gestures on command, although the verbal comprehension of the instruction is more difficult for emblems than for pantomime (Goodglass and Kaplan, 1963; Buxbaum et al., 2007). In the same vein, we found that in aphasic patients the accuracy of pantomime on command predicts the diversity and comprehensibility of gestures used for compensating the loss of speech when retelling short stories (Hogrefe et al., 2012). The stories did not put any particular emphasis on tool or object use. The most plausible link between success on pantomime and the efficiency of spontaneous gestures is their creation by selection and combination of distinctive features.

The conclusion that apraxic patients' difficulties with pantomime concern its communicative function rather than its descent from actual use leads to the question as to how they relate to accompanying impairments of other communicative channels, notably the impairment of verbal communication in aphasia. Like aphasia, defective pantomime of tool use is, strictly bound to left hemisphere lesions (Goodglass et al., 1963; Goldenberg et al., 2003). About one half to two third of patients with aphasia have problems with pantomime, and the proportion raises to more than 90% when only patients with global aphasia are considered (Goldenberg et al., 2003). Moreover, there is no right-handed patient on record who had apraxia of pantomime without aphasia. The suspicion that aphasia and apraxia are both symptoms of one and the same pervasive destruction of communicative abilities is, however, invalidated by observations of left-handed patients who had apraxia for both imitation and communicative gestures on command but not aphasia following right brain damage (Margolin, 1980; Goldenberg, 2013).

Contrary to widespread belief but in accordance with the sparing of pantomime in visuoimitative apraxia, the crucial lesions for defective pantomime are located in inferior frontal regions and adjacent white matter rather than in parietal regions (Goldenberg et al., 2007).

Tool Use

Defective tool use was the first symptom that received a designation as 'apraxia' (Steinthal, 1871). Liepmann classified defective use of tools and object as 'ideational' apraxia to indicate that the difficulty concerns the planning of the action rather than its motor execution, and later authors mostly adhered to this taxonomy. The uniform designation conceals substantial differences concerning its interpretation. Liepmann owed the description of this variant of apraxia mainly to the Prague psychiatrist Arnold Pick (1905) who described errors of demented patients on everyday multistep actions with multiple objects like dressing or preparing and lighting a pipe. Errors were due to omission or preservation of single steps and to loss of the superordinate goal of action sequences. Liepmann concluded that "ideational apraxia is a mental insufficiency which manifests itself in the domain of action but has its roots in deficits which are not specific for action" (Liepmann, 1929: p. 199). In the modern literature, this variant of apraxia has been rediscovered

independently from the traditional descriptions and has been named 'action disorganization syndrome' (Humphreys and Forde, 1998; Cooper et al., 2005) or 'naturalistic action impairment' (Schwartz et al., 1999). By contrast, the term 'ideational apraxia' is now predominantly used for defective use of single tools that manifests itself also outside of the context of multistep action sequences (Morlaas, 1928; De Renzi and Lucchelli, 1988). As I will outline in the following sections, the two interpretations of 'ideational apraxia' correspond to different variants of deficient tool and object use that have different neural substrates (see **Figure 1**). I will discuss them separately.

Use of Single Tools

Defective use of single familiar tools is the least frequent but most impressive manifestation of apraxia. For example, patients may try to cut bread with the reverse edge of the knife or even with a spoon, may press the head of the hammer upon the nail rather than hitting, or may try to press toothpaste out of a firmly closed tube. It seems that patients have lost their knowledge about the correct use of these familiar tools. Experimental studies reveal that there are at least two sources of such knowledge. One is the storage of 'functional knowledge' in semantic memory. It associates types of tools with their purpose, their recipient, and the action of their use. For example, a screwdriver serves for connecting or disconnecting parts; the recipient of its action is a screw, and the action rotation. Retrieval of such functional knowledge can be probed without requiring the actual use of tools by asking subjects to match pictures of tools with pictures of other tools serving the same purpose or with pictures of the typical recipient of their action.

Functional knowledge can exist only for familiar tools and objects, because acquisition and storage of knowledge about an object makes it familiar. If there are several possible uses of a tool, functional knowledge is likely to assign weightages to them by their relative frequency and familiarity. Consequently, the prototypical use predominates. For example, a screwdriver is registered as being useful for screwing and unscrewing rather than for stabbing. One can, however, use tools in non-prototypical ways and find out possible uses of unfamiliar tools. One can scratch with a screwdriver and screw with a coin. There must be a second source of knowledge about tool use that permits detection of possible uses independently from their storage in semantic memory.

The alternative source of knowledge about tool use can be a direct inference of possible functions from structure. Reasoning based on such inference has been termed 'mechanical problem solving' (Goldenberg and Hagmann, 1998), 'mechanical reasoning' (Hegarty, 2004), or 'technical reasoning' (Osiurak et al., 2009). The elements of mechanical problem solving are not the prototypical functions of entire tools, but the functional compatibilities of their parts (Vaina and Jaulent, 1991). Tools and objects are segmented into functionally significant parts and properties, and combinations of these parts and properties with parts and properties of other tools and objects are used for construction of mechanical chains.

Mechanical problem solving can be probed by confronting patients with novel tools and asking them to find out how to manipulate them or by presenting objects together with an array of tools but without the tool usually employed and asking them to choose the tool that would best complete the task (Heilman et al., 1997). In a variant of this task, Osiurak et al. (2009) presented objects paired with tools that usually are not employed for their use, and asked patients to show how the prototypical action associated with the object could nonetheless be achieved by employing the tool in an unusual way. For example, they presented a pot of yogurt with a fork. Eating of the yogurt could be achieved by turning the fork around and using the handle as a provisional spoon.

In patients with unilateral brain damage, the tests indicate that both, functional association and mechanical problem solving are impaired only after left brain damage (Hartmann et al., 2005), but whereas retrieval of functional knowledge is contingent upon integrity of the middle temporal lobe, mechanical problem solving depends on integrity of parietal regions (Goldenberg and Spatt, 2009). Impressive behavioral dissociations between preservation and disturbance of both components have been demonstrated mainly in patients with bilateral degeneration of either the temporal or the parietal lobes (Sirigu et al., 1991; Hodges et al., 2000; Spatt et al., 2002). In patients with unilateral left-sided lesions, defective use of single familiar tools seems to be bound to large lesions that interrupt both sources of knowledge, but there are indications that insufficient appreciation of functionally significant parts and properties of tools has more impact on the use of familiar tools and objects than disturbed retrieval of functional knowledge (Osiurak et al., 2009; Goldenberg et al., 2009).

Multistep Actions with Multiple Tools and Objects

In daily living, one rarely encounters a situation like in testing for use of single tools where one is handed a tool and asked to perform its prototypical action on an adequately prepared recipient. Typically, the use of the single tool is embedded in a chain of actions involving several tools and objects and aiming at a superordinate goal transgressing and modulating the purposes of each single action step. The need to keep track of completed and outstanding steps of actions, to avoid possible interferences between sequential action steps and to maintain the superordinate goal against digressions create more opportunities for error than the isolated use of single tools. Whereas single tool use by healthy persons is virtually errorless, slips of actions in multistep tasks are a common experience of everyday life, but they rarely attain the severity of errors that are committed by some patients with brain damage.

In contrast to the exclusive association of defective use of single tools with left brain damage, multistep actions are affected also by lesions in the right hemisphere or diffuse brain damage (Schwartz et al., 1998, 1999; Hartmann et al., 2005; Goldenberg et al., 2007). Clinical experience and detailed descriptions of single cases (Poeck and Lehmkuhl, 1980; Rumiati et al., 2001; Forde et al., 2004) suggest more severe aberrations from the normal course of action in patients whose brain damage affects the left hemisphere, but group studies failed to confirm this impression, nor did they show differential effects according to the intrahemispheric localization of lesions (Schwartz et al., 1999; Hartmann et al., 2005; Goldenberg et al., 2007).

The apparently uniform impact of any brain damage on multistep actions reminds of Liepmann's contention that 'ideational apraxia' is a manifestation of 'general mental insufficiency.' In the modern literature, a very similar proposal has been formulated as 'resource limitation.' According to this view, any brain damage limits the resources of attention and effort, and this limitation has a homogenizing effect that obscures any manifestations of underlying specific neuropsychological deficits (Buxbaum et al., 1998). However, some insight into differential mechanisms underlying the superficially uniform impairment of patients with different kinds of lesions has been provided by analyses of correlations between defective multistep actions and accompanying other symptoms of brain damage. In patients with left brain damage and aphasia, severity of impairment on multistep actions correlates with their difficulties with single tool actions and, for some multistep actions also with the severity of aphasia and of semantic memory impairment. Right hemispheric brain-damaged patients have neither aphasia, nor semantic memory impairment, nor problems with single tools. Their errors on multistep actions with multiple objects are correlated with the severity of hemineglect, a tendency to ignore the contralateral half of one's body or of external space (Schwartz et al., 1999; Hartmann et al., 2005). Subsumption of the diversity of accompanying impairments under the umbrella of 'resource limitation' is not very satisfactory. A more convincing assumption would be that there are in fact different mechanisms at work in different patient groups but that the inherent constraints and difficulties of multistep actions with multiple objects restrict the loci and kinds of possible errors and leave little freedom for neuropsychological disorders to manifest themselves in specific and recognizable ways.

Conclusions

Our survey clearly demonstrates that apraxia is not a unitary disorder. It rather appears as a collection of heterogeneous deficits that differ from other motor sequels of hemisphere damage by the bilateralism of their manifestation but have otherwise little in common. However, this bilateralism is not trivial. The deviation from the basic anatomy of motor cortex indicates that problems concern the planning and control rather than the motor execution of actions. It thus points to the cognitive rather than purely motor character of apraxia.

It has been proposed that the intervention of high level cognitive mechanisms comes to the fore most clearly in three core manifestations of apraxia: Imitation of meaningless hand postures, pantomime of tool use, and use of single tools (Goldenberg, 2013). Anatomically, they have in common a strict dependency on left hemisphere lesions. Functionally they have in common that they involve segmentation and combination of elements: For imitation of hand postures, the multiple visual features of the demonstrated gesture are segmented into distinct body parts which are combined to distinct postures. For pantomime of tool use, the compound image of hand, action, and object is segmented into distinctive features of the object and the acting hand, which are combined for giving an understandable image of the pretended object and its use. For use of single tools, the structures of tool and recipient are segmented into functionally significant traits, which are combined to form mechanical chains that transmit manual actions into the effects of the tool upon its recipient.

Segmentation and combination are the essential constituents of generative systems that can create a multitude of entities out of a limited number of elements (Abler, 1989). Their application to motor control enables subjects to understand and reproduce complex actions and to create adaptive actions in response to novel demands. Apraxia interferes with these cognitive contributions to motor control.

See also: Aphasia; Brain Asymmetry; Mental Imagery: Visual Cognition; Mirror Neurons, Theory of; Motor Control Models: Learning and Performance; Motor Cortex; Motor Skills: Development in Infancy and Early Childhood.

Bibliography

Abler, W.L., 1989. On the particulate principle of self-diversifying systems. Journal of Social and Biological Structures 12, 1–13.

Barbieri, C., De Renzi, E., 1988. The executive and ideational components of apraxia. Cortex 24, 535–544.

Broca, P., 1861. Perte de la parole, ramollissement chronique et destruction partielle du lobe antérieur gauche du cerveau. Communication Société d'Anthropologie, Séance du 18 avril 1861.

Buxbaum, L.J., Kyle, K., Grossman, M., Coslett, H.B., 2007. Left inferior parietal representations for skilled hand-object interactions: evidence from stroke and corticobasal degeneration. Cortex 43, 411–423.

Buxbaum, L.J., Schwartz, M.F., Montgomery, M.W., 1998. Ideational apraxia and naturalistic action. Cognitive Neuropsychology 15, 617–644.

Cooper, R.P., Schwartz, M.F., Yule, P., Shallice, T., 2005. The simulation of action disorganisation in complex activities of daily living. Cognitive Neuropsychology 22, 959–1004.

De Ajuriaguerra, J., Tissot, R., 1969. The apraxias. In: Vinken, P.J., Bruyn, G.W. (Eds.), Handbook of Clinical Neurology, vol. 4. North Holland, Amsterdam, pp. 48–66.

De Renzi, E., 1990. Apraxia. In: Boller, F., Grafman, J. (Eds.), Handbook of Clinical Neuropsychology, vol. 2. Elsevier, Amsterdam, New York, Oxford, pp. 245–263.

De Renzi, E., Lucchelli, F., 1988. Ideational apraxia. Brain 111, 1173–1185.

Denny Brown, D., 1958. The nature of apraxia. Journal of Nervous and Mental Disease 126, 9–32.

Finkelnburg, C.M., 1870. Sitzung der Niederrheinischen Gesellschaft in Bonn. Medizinische Section. Berliner Klinische Wochenschrift, 7, 449–450, 460–462.

Forde, E.M.E., Humphreys, G.W., Remoundou, M., 2004. Disordered knowledge of action order in action disorganisation syndrome. Neurocase 10, 19–28.

Geschwind, N., Damasio, A.R., 1985. Apraxia. In: Frederiks, J.A.M. (Ed.), Handbook of Clinical Neurology, Clinical Neuropsychology, vol. 1(49). Elsevier, Amsterdam, New York, pp. 423–432.

Goldenberg, G., 1995. Imitating gestures and manipulating a mannikin – the representation of the human body in ideomotor apraxia. Neuropsychologia 33, 63–72.

Goldenberg, G., 2003. Apraxia and beyond – life and works of Hugo Karl Liepmann. Cortex 39, 509–525.

Goldenberg, G., 2013. Apraxia – the Cognitive Side of Motor Control. Oxford University Press, Oxford, New York.

Goldenberg, G., Hagmann, S., 1997. The meaning of meaningless gestures: a study of visuo-imitative apraxia. Neuropsychologia 35, 333–341.

Goldenberg, G., Hagmann, S., 1998. Tool use and mechanical problem solving in apraxia. Neuropsychologia 36, 581–589.

Goldenberg, G., Hartmann, K., Schlott, I., 2003. Defective pantomime of object use in left brain damage: apraxia or asymbolia? Neuropsychologia 41, 1565–1573.

Goldenberg, G., Hartmann-Schmid, K., Sürer, F., Daumüller, M., Hermsdörfer, J., 2007. The impact of dysexecutive syndrome on use of tools and technical equipment. Cortex 43, 424–435.

Goldenberg, G., Hermsdörfer, J., Glindemann, R., Rorden, C., Karnath, H.O., 2007. Pantomime of tool use depends on integrity of left inferior frontal cortex. Cerebral Cortex 17, 2769–2776.

Goldenberg, G., Karnath, H.O., 2006. The neural basis of imitation is body-part specific. Journal of Neuroscience 26, 6282–6287.

Goldenberg, G., Münsinger, U., Karnath, H.O., 2009. Severity of neglect predicts accuracy of imitation in patients with right hemisphere lesions. Neuropsychologia 47, 2948–2952.

Goldenberg, G., Spatt, J., 2009. The neural basis of tool use. Brain 132, 1645–1655.

Goldenberg, G., Strauss, S., 2002. Hemisphere asymmetries for imitation of novel gestures. Neurology 59, 893–897.

Goldstein, K., 1928. Beobachtungen über die Veränderungen des Gesamtverhaltens bei Gehirnschädigung. Monatschrift für Psychiatrie und Neurologie 68, 217–242.

Goodglass, H., Kaplan, E., 1963. Disturbance of gesture and pantomime in aphasia. Brain 86, 703–720.

Hartmann, K., Goldenberg, G., Daumüller, M., Hermsdörfer, J., 2005. It takes the whole brain to make a cup of coffee: the neuropsychology of naturalistic actions involving technical devices. Neuropsychologia 43, 625–637.

Hegarty, M., 2004. Mechanical reasoning by mental simulation. Trends in Cognitive Sciences 8, 280–285.

Heilman, K.M., Maher, L.M., Greenwald, M.L., Rothi, L.J.G., 1997. Conceptual apraxia from lateralized lesions. Neurology 49, 457–464.

Hodges, J.R., Bozeat, S., Lambon Ralph, M.A., Patterson, K., Spatt, J., 2000. The role of conceptual knowledge in object use – evidence from semantic dementia. Brain 123, 1913–1925.

Hogrefe, K., Ziegler, W., Weidinger, N., Goldenberg, G., 2012. Non-verbal communication in severe aphasia: influence of aphasia, apraxia, or semantic processing? Cortex 48, 952–962.

Humphreys, G.W., Forde, E.M.E., 1998. Disordered action schema and action disorganisation syndrome. Cognitive Neuropsychology 15, 771–812.

Kimura, D., Archibald, Y., 1974. Motor functions of the left hemisphere. Brain 97, 337–350.

Laimgruber, K., Goldenberg, G., Hermsdörfer, J., 2005. Manual and hemispheric asymmetries in the execution of actual and pantomimed prehension. Neuropsychologia 43, 682–692.

Liepmann, H., 1908. Drei Aufsätze aus dem Apraxiegebiet. Karger, Berlin.

Liepmann, H., 1929. Klinische und psychologische Untersuchung und anatomischer Befund bei einem Fall von Dyspraxie und Agraphie [posthumos publication]. Monatschrift für Psychiatrie und Neurologie 71, 169–214.

Margolin, D.I., 1980. Right hemisphere dominance for praxis and left hemisphere dominance for speech in a left-hander. Neuropsychologia 18, 715–719.

Morlaas, J., 1928. Contribution à l'étude de l'apraxie. Amédée Legrand, Paris.

Osiurak, F., Jarry, C., Aubin, G., Allain, P., Etcharry-Bouyx, F., Richard, I., et al., 2009. Unusual use of objects after unilateral brain damage. The technical reasoning model. Cortex 45, 769–783.

Pick, A., 1905. Studien zur motorischen Apraxia und ihr nahe stehenden Erscheinungen; ihre Bedeutung in der Symptomatologie psychopathischer Symptomenkomplexe. Franz Deuticke, Leipzig und Wien.

Poeck, K., Lehmkuhl, G., 1980. Das Syndrom der ideatorischen Apraxie und seine Lokalisation. Nervenarzt 51, 217–225.

Rizzolatti, G., Fadiga, L., Fogassi, L., Gallese, V., 2002. From mirror neurons to imitation: facts and speculations. In: Meltzoff, A.N., Prinz, W. (Eds.), The Imitative Mind. Development, Evolution, and Brain Bases. Cambridge University Press, Cambridge, pp. 247–266.

Rothi, L.J.G., Ochipa, C., Heilman, K.M., 1991. A cognitive neuropsychological model of limb praxis. Cognitive Neuropsychology 8, 443–458.

Rumiati, R.I., Zanini, S., Vorano, L., Shallice, T., 2001. A form of ideational apraxia as a selective deficit of contention scheduling. Cognitive Neuropsychology 18, 617–642.

Schwartz, M.F., Buxbaum, L.J., Montgomery, M.W., Fitzpatrick-DeSalme, E.J., Hart, T., Ferraro, M., et al., 1999. Naturalistic action production following right hemisphere stroke. Neuropsychologia 37, 51–66.

Schwartz, M.F., Lee, S.S., Coslett, H.B., Montgomery, M.W., Buxbaum, L.J., Carew, T.G., et al., 1998. Naturalistic action impairment in closed head injury. Neuropsychology 12, 13–28.

Sirigu, A., Duhamel, J.R., Poncet, M., 1991. The role of sensorimotor experience in object recognition – a case of multimodal agnosia. Brain 114, 2555–2573.

Spatt, J., Bak, T., Bozeat, S., Patterson, K., Hodges, J.R., 2002. Apraxia, mechanical problem solving and semantic knowledge – contributions to object usage in corticobasal degeneration. Journal of Neurology 249, 601–608.

Steinthal, H., 1871. Abriss der Sprachwissenschaft. (vols. Erster Band: Einleitung in die Psychologie und Sprachwissenschaft). Ferd.Dümmlers Verlagsbuchhandlung Harrwitz und Gossmann, Berlin.

Tessari, A., Canessa, N., Ukmar, M., Rumiati, R.I., 2007. Neuropsychological evidence for a strategic control of multiple routes in imitation. Brain 130, 1111–1126.

Vaina, L.M., Jaulent, M.C., 1991. Object structure and action requirements: a compatibility model for functional recognition. International Journal of Intelligent Systems 6, 313–336.

Archaeology and Cultural Memory

Cornelius Holtorf, School of Cultural Sciences, Linnaeus University, Kalmar, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses several theoretical concepts of how collective memories function in society. An overview is given of the different implications of these concepts for both the archaeological study of ancient sites and objects, and archaeology as an academic discipline in the context of a given society or nation of the present.

Individual humans can remember occurrences of their own lifetimes as well as of a collective past, whether near or far removed in time from the moment of remembering. While the former is restricted to events and processes at which the particular person was physically present as a conscious human being, the latter includes accounts of occurrences happening both before and during an individual's life. Although records of personal memories may be an important aspect in some areas of archaeological research as has been demonstrated, e.g., in the case of twentieth-century archaeology (Burström et al., 2009), most archaeologists are dealing with collective memories.

This article discusses several theoretical models of how collective memories function in society. An overview will be given on the different implications for both the archaeological study of ancient sites and objects and archaeology as an academic discipline in the context of a given society or nation of the present.

Archaeology and Theoretical Approaches to Memory

For the longest time in intellectual history, human memory has been seen as a huge archive in every human being from which, at any point in time, specific items can be retrieved in the process of remembering. This was, for example, the view of Saint Augustine (*Confessions*, X. viii (12)). In this perspective, individuals may create mnemonics in order not to forget on particular occasions (Yates, 1966) and they are able, in principle, to remember past events accurately. Collective memories may work accordingly. More recently, however, it has been argued that the issues involved in remembering and forgetting are not that simple (Cubitt, 2007).

Remembrance as a Social Construction

Remembering, and indeed forgetting, are strongly influenced by circumstances of the time in which they take place. Moreover, memory can fail completely or 'make up' stories or events that allegedly took place in the past. Maurice Halbwachs argued that some key factors affecting memory derive from the social arena that people always inhabit when they remember. Halbwachs introduced the concept of *mémoire collective* (collective memory), and emphasized how strongly social processes influence not only people's personal memories of their own lifetimes, but also a community's shared memories of the past. Collective memories are crucial for the identity of groups such as families, believers of a religion, or social classes (Halbwachs, 1992). Although memories vary between individuals and no two persons share identical 'collective' memories, anyone who has ever lived for some time in a foreign country knows what it means not to share the same collective memories as colleagues and friends. Collective memories can be evoked at particular sites (see below), and they have, in turn, the potential to shape such sites. As an example of the latter, Halbwachs discussed the medieval 'creation' of the Holy Land by the superimposition of an imaginary landscape on Palestine (Halbwachs, 1971).

The collective memory of archaeology provides a topical example. Although little research has been conducted on that topic, it is clear that some historical figures such as J. J. Winckelmann, C. J. Thomsen, H. Carter, and H. Schliemann are considered widely to be among the 'fathers' of the discipline. Similarly, sites such as the Valley of the Kings, Troy, Rome, Pompeii, Stonehenge, Teotihuacan, and Machu Picchu as well as museums such as the Louvre, the British Museum or the Metropolitan Museum of Art are remembered collectively as key sites of archaeology and are now major tourist attractions. While the significance of any of these persons, sites, and museums is not disputed here, today they are inseparably linked with how one collectively remembers the history of archaeology including events that happened long after these individuals were dead, at other places than those mentioned, and concerning artifacts displayed elsewhere. During the nineteenth and twentieth centuries, archaeology became a rich cultural field in which literature, film, tourism, advertising, and even other academic disciplines like psychoanalysis together remembered the spectacular discoveries at famous ancient sites by celebrity archaeologists of the past (Frayling, 1992; Zintzen, 1998).

Going beyond Halbwachs's argument, social scientists have long argued that memory of the past is not only influenced but also constituted by social contexts of the present (Middleton and Edwards, 1990; Zerubavel, 2003). Such reasoning questions the separation of past and present in a fundamental way. It becomes pointless to discuss whether or not a particular remembered event or process corresponds to what 'actually' happened; all that matters are the specific conditions under which such memory is constructed. The distinction between personal and collective memory is thus not necessarily a sharp one. Both reflect, first and foremost, the conditions of the present in which they originate. Individual persons learn collective memories

through socialization, but they retain the freedom to offer alternative views of the past, which may themselves later inform the collective memory. It may eventually become impossible to tell which memory is indeed an accurate remembrance of an occurrence at which a person was present, and which is merely a remembrance of an earlier remembrance that was created by a story read in a book or watched on television (Morris-Suzuki, 2005). As a consequence of such reasoning, the focus of memory studies in the social sciences and to some extent in the humanities has shifted from studies seeking to establish the truthfulness or social constructedness of memory to studies of how particular mnemonic and commemorative practices in past and present contributed to the formation of personal and collective human identities and social strategies and conflicts linked to them (Olick and Robbins, 1998). In archaeology, this trend has become particularly visible in a number of studies of remembrance and commemorative practices in historic and modern societies (Williams, 2003, 2006; Walls and Williams, 2010).

Les Lieux des Mémoire

Pierre Nora edited a monumental work of seven volumes about the loci memoriae of France, entitled *Les lieux des mémoire* (Nora, 1984–92). A *lieu de mémoire* (site of memory) is any significant entity, whether material or nonmaterial in nature, which by dint of human will or the work of time has become a symbolic element of the memorial heritage of any community. Nora deals specifically with sites where modern France constructs its national identity by constructing its past. Crucially, these sites of memory not only include places, such as museums, cathedrals, cemeteries, and memorials, but also concepts and practices, such as generations, mottos, commemorative ceremonies, and objects, e.g., inherited property, monuments, classic texts, and symbols. Archaeological sites of memory discussed in individual contributions include the prehistoric art genre known as palaeolithic cave paintings, megalithic monuments as a set of particular objects, and the concept of the Gauls.

According to Nora, all sites of memory are artificial and deliberately fabricated. Their purpose is to stop time and the work of forgetting, and what they all share is 'a will to remember.' Such sites of memory are not common in all cultures, but a phenomenon of our time; they replace a 'real' and 'true' living memory that was with us for millennia but has now ceased to exist. Nora thus argues that a constructed history replaces true memory. Although Nora's idealization of memory at the expense of history and heritage has now been almost forgotten, the variety of empirical phenomena he considered relevant and invited contributors to study in detail has inspired many others and led to similar projects in other countries (e.g., Francois and Schulze, 2001). Many of the sites and notions addressed are directly relevant to the fields of historical and contemporary archaeology as well as landscape archaeology.

Cultural Memory

Jan Assmann employed the term *kulturelles Gedächtnis* (cultural memory) in a study of the past in ancient Egypt (Assmann, 1992). Cultural memory can be defined as the cultural expressions of a collective memory, including written texts, symbols, monuments, memorials, etc. It is distinct from the 'communicative memory' of individuals' lifetimes, which is expressed by means of direct communication. Cultural and collective memories are 'retrospective.' Correspondingly, the hopes and expectations of people for what will be remembered in the future can be termed 'prospective memory.' Such hopes find their expression not only in the design and 'content' of a site of memory, but also in its monumentality and the durability of the materials employed. If prospective memory is the desired memory when building a site of memory, retrospective memory is how this site is in fact later remembered.

According to Assmann, cultural memory embraces both *Erinnerungskultur* (memory culture) and *Vergangenheitsbezug* (reference to the past). Memory culture is the way a society ensures cultural continuity by preserving, with the help of cultural mnemonics, its collective knowledge from one generation to the next, rendering it possible for later generations to reconstruct their cultural identity and maintain their traditions. Memory culture relies on later references to (mnemonics of) the past. Such references reassure the members of a society of a shared past and collective identity, and supply them with an awareness of their unity and singularity in time and space. Cultural mnemonics, such as preserved ancient monuments or artifacts, trigger recollections of past times, although these can be fabrications rather than accurate representations of the past. Arguably, this second aspect of cultural memory has wider implications. The fact that collective memories of the past occur without being able to ensure cultural continuity suggests that the past can be given specific meanings for reasons other than accurate reconstruction. In effect, as interpretations of the past have changed over time, ancient sites and artifacts have been interpreted quite differently during their long 'life histories' (Holtorf, 2008).

For archaeologists, both actual continuities and various archaistic reinventions or other symbolic references to a rediscovered or recreated past are interesting phenomena to study. Because both already occur widely in antiquity, Nora is arguably misguided in limiting the existence of sites of memory to the modern era. There is ample archaeological evidence for deliberate attempts to stop time and make contemporaries remember by 'artificial' means. Indeed, archaeologists have in recent years devoted much attention to studies in many parts of the world of what has been called 'the past in the past' (Assmann, 1992; Jonker, 1995; Van Dyke and Alcock, 2003; Williams, 2006; Holtorf, 2008).

The Relationship between Archaeology and Collective Memories

Archaeology versus Memory

It has often been assumed that the academic study of the past is superior epistemologically to popular notions of the past, as they are reflected in folklore, myths, or in other expressions of collective memory. Maurice Halbwachs contrasted memory and history as two oppositional ways of dealing with the past. Historians aim at writing a single objective and impartial

universal history, whereas collective memories are numerous, limited in their validity to members of a particular community, subject to manifold social influences, and restricted to the very recent past in scope.

Likewise, Pierre Nora argued that memory and history are two very different phenomena, but his preference was the opposite to that of Halbwachs. Nora distinguished true memory, borne by living societies maintaining their traditions, from artificial history, which is always problematic and incomplete, and represents something that no longer exists. For Nora, history holds nothing desirable. In each case, the term 'history' might be replaced with 'archaeology.'

Archaeology as Memory

Other scholars fundamentally challenged the split between history/archaeology and memory, and proposed instead a more fluid transition (e.g., Burke, 1989). In this view, history and archaeology are seen as special cases of social and cultural memories. Doing archaeology simply means (1) to recognize, and treat, certain things as 'evidence' for the past and (2) to describe and analyze it in a way that is valuable according to the standards of archaeologists. These practices are not categorically different from how people engage with the past in their everyday lives. History and archaeology, like all forms of memory, give particular meanings to ancient sites and artifacts. Some of these 'academic' meanings eventually influence collective memories of the past in the present. But many never leave the sphere of academia, and are short lived there too. What is interesting in studying collective memories of the past, academic or not, is, therefore, not only how truthfully they may represent (some) part of a past reality, but also why particular memories are created and adopted (or not) in particular contexts, and what sort of controversies may arise from them. Topical archaeological examples are provided by Paul Shackel's (2003) discussion of competing present-day memories of American Civil War sites and by Gabriel Moshenska's (2010) analysis of contested memories at specific places associated with the Nazi area in Berlin.

Archaeology and Nationalist Agendas

The history of archaeology provides many examples of the great importance of the past for collective identities of modern nations (Kohl and Fawcett, 1995; Diaz-Andreu, 2007). Archaeology originated as an academic discipline in the nineteenth century – the great age of national Romanticism and the European nation-states. This is no coincidence; archaeology has from its beginnings been directly bound up with politics. At times of fundamental social change, people turn to their 'origins' and seek reassurance either in a 'better' past, or in historical traditions and collective identities that legitimate a radical political movement or new social order. This was the situation in early nineteenth-century Europe in the aftermath of the social changes brought about by the French Revolution. By projecting backward contemporary collective identities and establishing (or even inventing) the origins of their respective national people in the distant past, the young European nations reassured their citizens of a shared history, and thus a common identity, legitimizing their existence in the present. As a result, archaeology became institutionalized in the form of national museums, professorships at universities, and legislation protecting the archaeological heritage. A case in point are the defenses of Masada in Palestine, which fell to the Romans after several years of Jewish resistance in AD 73, but not until the last defenders had committed collective suicide. They were not rediscovered as a prime object of Jewish collective memory until the 1920s coinciding with the rise of Zionism (Schwartz et al., 1986).

Since the nineteenth century, a lot of archaeological research has been adopting a 'culture-historical' approach, investing much effort in the ethnic interpretation of material evidence according to which clearly defined types of sites and artifacts regularly recurring together were associated with particular ethnic cultures corresponding to modern cultures. Outside of Europe, colonial and imperialistic archaeologies attempted to establish similar ways of conducting archaeology. The ultimate aim was the establishment of cultural sequences connecting present-day societies with their distant origins in a more or less direct line of historic continuity. In extreme cases, as in Nazi-Germany 1933–45, such reasoning could ultimately lead to aggressive territorial claims when archaeological finds in neighboring countries were seen as connected with archaeological cultures in contemporary Germany. Today, some 'first nations' of indigenous populations in North America also use ancient sites and finds to establish collective identities and legitimize territorial claims.

The connection between archaeology and modern nations is still visible at the beginning of the twenty-first century in many 'national museums' of archaeology, in national academic journals with titles such as *Germania* and *Gallia*, and in the key role of the national pasts in both teaching curricula and popular culture. During the last decade of the twentieth century, and with the emergence of new nation-states in the former Yugoslavia and the former Soviet Union, similar processes were repeated (Kohl and Fawcett, 1995). Although in postwar Germany, remembrance of the Nazi past, e.g., at the sites of former concentration camps, has been rendering national memories into something that is deeply problematic and troubling, in other countries, collective memories may still support nationalist ideologies effectively, while contradicting results gained from archaeological research. Some archaeologists may lend their authority to support such ideologies. At first, this appears to raise a dilemma for those archaeologists who are prepared to accept the equal legitimacy and value of all collective memories of the past, whether created by academic archaeology or other groups; either they have to accept all constructions of the past as legitimate alternatives to their own accounts, or they would contradict themselves by dismissing some as false. A realistic solution is to accept many alternative collective memories of the past as manifestations of particular needs and desires in present-day contexts, but be prepared to fight some of them on political or other grounds, and publicize widely any dangerous consequences or implications. Such issues have recently come to the fore in several regions of the world. They demonstrate that archaeology and the politics of collective/national memory are deeply intertwined and ultimately interdependent. Instead of arguments about theoretical principles, future archaeologists will be challenged increasingly to find

pragmatic guidelines as to how to behave and act in politically highly charged situations.

See also: Archaeology, Politics of; Collective Beliefs, Sociological Explanation of; Collective Identity; Collective Memory, Anthropology of; Collective Memory, Psychology of; Collective Memory, Psychology of; Cultural History; Ethnic Identity and Ethnicity in Archaeology; Nationalism, Sociology of; Nationalism: General; Reconstructive Memory, Psychology of.

Bibliography

Assmann, J., 1992. Das Kulturelle Gedächtnis. Schrift, Erinnerung und Politische Identität in Frühen Hochkulturen. Beck, Munich, Germany (English translation, 2011. Cultural Memory and Early Civilization: Writing, Remembrance and Political Imagination. Cambridge University Press, Cambridge).

Burke, P., 1989. History as social memory. In: Butler, T. (Ed.), Memory: History, Culture and the Mind. Blackwell, Oxford, UK, pp. 97–113.

Burström, M., Diez Acosta, T., González Noriega, E., Gustafsson, A., Hernández, I., Karlsson, H., Pajón, J.M., Robaina Jaramillo, J.R., Westergaard, B., 2009. Memories of a world crisis: the archaeology of a former Soviet nuclear missile site in Cuba. Journal of Social Archaeology 9 (3), 295–313.

Cubitt, G., 2007. History and Memory. Manchester University Press, Manchester and New York.

Diaz-Andreu, M., 2007. A World History of Nineteenth-Century Archaeology. Nationalism, Colonialism, and the Past. Oxford University Press, Oxford.

Francois, E., Schulze, H. (Eds.), 2001. Deutsche Erinnerungsorte, 3 vols. Beck, Munich.

Frayling, C., 1992. The Face of Tutankhamun. Faber and Faber, London and Boston.

Halbwachs, M., 1971 [1941]. La Topographie Légendaire des Évangiles en Terre Sainte. Etude de Mémoire Collective. Presses Universitaires de France, Paris (Conclusion translated in Halbwachs 1992).

Halbwachs, M., 1992. On Collective Memory (L.A. Coser, Ed., Trans., and Intro.). University of Chicago Press, Chicago.

Holtorf, C., 2008. Monumental Past: the Life-Histories of Megalithic Monuments in Mecklenburg-Vorpommern (Germany). CITD Press, Toronto. Electronic monograph: http://hdl.handle.net/1807/245.

Jonker, G., 1995. The Topography of Remembrance. The Dead, Tradition and Collective Memory in Mesopotamia. Brill, Leiden.

Kohl, P., Fawcett, C. (Eds.), 1995. Nationalism, Politics, and the Practice of Archaeology. Cambridge University Press, Cambridge.

Middleton, D., Edwards, D. (Eds.), 1990. Collective Remembering. Sage, London.

Morris-Suzuki, T., 2005. The Past within Us: Memory, Media, History. Verso, London and New York.

Moshenska, G., 2010. Working with memory in the archaeology of modern conflict. Cambridge Archaeological Journal 20 (1), 33–48.

Nora, P. (Ed.), 1984–92. Les Lieux de Mémoire, 7 vols. Edition Gallimard, Paris (English translation, Trouille, M.S., 2001–10. Rethinking France, 4 vols. University of Chicago Press, Chicago).

Olick, J.K., Robbins, J., 1998. Social memory studies: from "collective memory" to the historical Sociology of mnemonic practices. Annual Review of Sociology 24, 105–140.

Schwartz, B., Zerubavel, Y., Barnett, B.M., 1986. The recovery of Masada: a study in collective memory. The Sociological Quarterly 27, 147–164.

Shackel, P.A., 2003. Memory in Black and White: Race, Commemoration, and the Post-Bellum Landscape. Altamira, Lanham.

Van Dyke, R., Alcock, S.E. (Eds.), 2003. Archaeologies of Memory. Malden and Oxford, Blackwell.

Walls, S., Williams, H., 2010. Death and memory on the home front. Second World War commemoration in the South Hams, Devon. Cambridge Archaeological Journal 20 (1), 49–66.

Williams, H. (Ed.), 2003. Archaeologies of Remembrance. Death and Memory in Past Societies. Kluwer/Plenum, New York.

Williams, H., 2006. Death and Memory in Early Medieval Britain. Cambridge University Press, Cambridge.

Yates, F., 1966. The Art of Memory. Routledge & Kegan Paul, London.

Zerubavel, E., 2003. Time Maps. Collective Memory and the Social Shape of the Past. University of Chicago Press, Chicago and London.

Zintzen, C., 1998. Von Pompeji Nach Troja. Archäologie, Literatur und Öffentlichkeit im 19. Jahrhundert. Universitätsverlag, Vienna.

Relevant Websites

http://archive.archaeology.org/0507/etc/conversations.html – Alcock, S., 2005. Conversations: monuments and memory, Archaeology Magazine.

http://www.cric.arch.cam.ac.uk/home.html – Cultural Heritage and the Re-construction of Identities after Conflict (CRIC) Project homepage.

http://eunamus.eu/ – European National Museums: Identity politics, the uses of the past and the European citizen, (EUNAMUS) Project homepage.

http://hdl.handle.net/1807/245 – Holtorf, C., 2008. Monumental past. The life-histories of megalithic monuments in Mecklenburg-Vorpommern (Germany). Electronic monograph.

Archaeology, Politics of

Lynne Goldstein, Michigan State University, East Lansing, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Most people do not think of archaeology as political, but issues focused on the repatriation of human remains and sacred objects have made archaeology a very political discipline. Repatriation of human remains and sacred objects has become a focus of concern in the United States, Canada, Australia, and other countries over the last 20 years. This article examines the definitions and nature of repatriation, the history of repatriation laws and approaches, the politics of the issues, international vs US laws, key issues and foci of current debate, and potential future issues associated with repatriation. There has been a marked culture change among archaeologists and museums concerning these issues, and this shift will have a lasting impact on archaeology, on indigenous peoples, and on the nature of museums. Archaeology can no longer claim to exist outside politics.

Introduction

Archaeology has not traditionally been seen as a political discipline. Most of the complexities associated with 'doing' archaeology focus on whether or not a particular archaeologist has received the appropriate permits, is conducting fieldwork properly, understands and analyzes the cultural materials, and publishes findings. However, there are political issues associated with archaeology, particularly in relation to cultural resources management work, in which archaeologists must mitigate the impact of development and construction on archaeological resources. Archaeology is also not immune to infighting and other types of academic jockeying for positions, but on the whole, the discipline has not been the focus of major political battles. Any apolitical view of archaeology changed in the mid-1980s and early 1990s when the issue of repatriation of human remains and sacred objects became hotly contested by Native Americans. Archaeology was suddenly in the news for topics other than new things found, and archaeologists were not cast as the 'good guys.' Eventually, many states and countries passed legislation concerning human remains and sacred objects, and the two laws that were the leads in this trend were the National Museum of the American Indian Act (NMAIA; 20 United States Code (U.S.C.) 80q et seq.) and the Native American Graves Protection and Repatriation Act (NAGPRA; 25 U.S.C. 3001 et seq.)

A focus on repatriation is appropriate and relevant within an article on the politics of archaeology since there have been no other laws or developments within the last 30 years that have had as major an impact on the discipline.

For most members of the general public in the United States (and some other parts of the world), repatriation or reburial (as applied to Anthropology and museums) appears to be a straightforward notion – tribes and/or other appropriate indigenous groups or relatives ask for the return of one or more objects or human remains that are historically or culturally related to the group, and the museum or university gives it back, or should give it back. Some are aware that the process can be a bit more complicated, but most people have little idea of the nature and process of repatriation, or the problems and complications that are associated. This is not to say that repatriation is problematic and should not be done, but rather that all parties are obligated to follow an explicit legal process. However, the biggest differences between repatriation in the twenty-first century and repatriation in the twentieth century are: (1) repatriation of human remains and associated objects is now law in the United States and some other countries of the world, and (2) repatriation of human remains (in particular) is now widely accepted as something that responsible and ethical institutions do.

Definition

The dictionary definition of 'repatriation' is: "send (someone) back to their own country, or send or bring back to one's own country." Wikipedia defines repatriation similarly:

> Repatriation (from Late Latin *repatriare*) is the process of returning a person back to one's place of origin or citizenship. This includes the process of returning refugees or soldiers to their place of origin following a war. The term may also refer to the process of converting a foreign currency into the currency of one's own country.

The repatriation and reburial of human remains, funerary and sacred objects is a special case of repatriation. Groups representing indigenous peoples around the world have requested that human remains and sacred objects originally taken from their respective historical territories be repatriated from museums and other institutions. Sometimes institutions repatriate human remains and items because they agree that they should rightfully be in the possession of the group that is descended from them, other times the return is based on the requirements of specific laws.

One reason that repatriation – as applied to archaeology and physical anthropology – is not a simple process is because the motivation for these requests is about people in the present and their perception of the past in the present. Repatriation claims are often associated with politics, ethnic identity, and other debates or problems in contemporary society that have or claim to have a historical link to the object or human remains. Such motivations do not make the claim less valid or less appropriate, but the context of the request can be important to understanding the request. Archaeologists, physical

anthropologists, and museum directors do not generally approach repatriation from this perspective; they try to contextualize the materials based on archaeological and historical knowledge and science. It is a discussion that has often led to misunderstandings and differences because the parties are focused on different matters. However, since repatriation laws have come into effect, it is a discussion that is happening in many places, and consequently museums are getting better at responding to requests, and indigenous groups are getting better at making their requests. Repatriation is most active in the United States, primarily because repatriation laws have been in effect for the longest period of time.

Please note that although the term 'archaeologist' is generally used in this article, the reader should understand that this is a shorthand or simplified way of encompassing people who are archaeologists, physical anthropologists, human osteologists, museum studies scholars, and those in related fields who deal with human remains, grave goods, and sacred or religious objects. Furthermore, archaeology has historically taken the lead in these discussions about repatriation, so this shorthand approach is reasonably accurate as well. Archaeologists have most consistently found themselves in the center of the political debate.

Historical Background

During the 1960s, 1970s, and 1980s, some US institutions listened and responded to repatriation requests from tribes, subsequently returning human remains and sacred objects before specific laws were in place. However, there is no question that the passage of two federal laws in the United States – NMAIA (20 United States Code (U.S.C.) 80q et seq.) and the NAGPRA (25 U.S.C. 3001 et seq.) – dramatically changed the nature and process of repatriation. There is no doubt that the effect of this change in process and change in culture has had international implications.

Buikstra (2006), Lambert (2012), Mihesuah (2000), Trope and Echo-Hawk (1992), and Ubelaker and Grant (1989) provide detailed discussions of the history and some of the problems of repatriation in the United States. During the 1980s, the position held by many institutions, as well as by many national professional organizations, was that the ethics of today supersede those of the past, and museums should consider how to change accordingly. At the basic level, repatriation should be done for living descendants upon request and examination of the claim. As Ubelaker and Grant (1989) note, if a group requested remains where there were no living descendants, that group must demonstrate compelling religious or cultural values that outweigh scientific interests. This position also reflected the position of the Society for American Archaeology (SAA) at the time. However, the SAA increasingly assumed the role of representing the scientific community during the 1980s and their view broadened to some degree.

Since the 1980s, the SAA has been remarkably consistent in their position on repatriation (Lovis et al., 2004). The SAA argued for repatriation, but on a case-by-case basis. Indeed, the SAA Executive Committee passed an initial resolution in 1983, followed by a revised version in 1986, and a reaffirmed position in 1999. The key points of the SAA position include: (1) both native and scientific viewpoints have legitimacy; (2) scientific importance must be considered against the strength of affiliation; (3) implementation of repatriation should be done on a case-by-case basis because of the variability in circumstances and histories; and (4) communication and consultation with affected groups is of critical importance. Since 1986, the SAA (Lovis et al., 2004) has recognized the importance and legitimacy of repatriation, and worked actively with Congress to create balanced legislation. Note, however, that not all tribes welcomed the SAA position.

Some native tribes and organizations, archaeologists, and other activists argued that repatriation represented something that was morally and ethically important for the archaeological profession to do, even if archaeology lost access to the remains and artifacts. However, even though several states enacted legislation that in some manner dictated repatriation and/or consultation with native groups, there was no consistent national policy until the 1990s.

During the late 1980s, a number of tribes in the United States argued that repatriation of human remains was a human rights issue – that these were human bodies that had to be treated with respect and respect was not possible if the remains were excavated and placed in a museum. Tribes did not see much difference between archaeologists and looters. One incident that angered many (including many archaeologists) happened in Iowa. A cemetery area was excavated due to road construction. All of the historic white settlers were immediately moved to a new location where they were reburied, and the Native American remains were sent to a museum. The lack of consistency and parity, and the continuing idea that Native American remains belong in a museum (in part suggesting that this may be because they have no clear descendants) represented the perfect example of what was wrong with archaeology and museums. Ironically, it is likely that in the Iowa case, if archaeologists and physical anthropologists had their way, they would have treated all remains the same, but would have studied them all before reburying.

A number of issues and events changed archaeological and other perspectives on repatriation. First, there were several instances, like the aforementioned Iowa case, that led many archaeologists to become increasingly uncomfortable over the idea of study and curation. In the 1980s and early 1990s, the archaeologist who perhaps best personified this position was Larry Zimmerman. In one article, Zimmerman reflects on his experiences and the present and future of archaeology (2000). Zimmerman notes (2000: pp. 294–295) that the reburial issue's trajectory has followed classic syncretism, or a coalescing or reconciliation of differing beliefs. Zimmerman (2000: p. 295) argues that "the problem has been that the belief system of Western science about the past, through archaeology, had imposed itself on belief systems of indigenous peoples about their pasts and those they consider their ancestors." Because archaeologists had little resistance from these indigenous people over a long period of time, they assumed that these people agreed or did not object. The past 25–30 years has demonstrated that archaeologists were wrong. Public discussion – as represented in various polls and forums – demonstrates that the general public supported the Native American position.

As Buikstra (2006: pp. 395–396) points out, another influence on the relationships between archaeologists and indigenous peoples was the development and impact of post-processual theories in archaeology in the 1980s and 1990s. This approach to archaeology moved away from being scientific, and was humanistic and contextual in nature, with practitioners arguing for multivocality in interpreting the past (cf Hodder, 1985; Shanks and Tilley, 1987). Such an approach was certainly more open and welcoming to resolving repatriation disputes, but not everyone agreed with the approach or the benefit for repatriation.

The first US law that directly addressed issues of repatriation was the NMAIA (20 United States Code (U.S.C.) 80q et seq.), which was signed into law on 28 November 1989. The NMAIA was written and passed primarily to establish a new national museum dedicated to the history, music, and art of American Native cultures. The collections of the private Heye Foundation in New York formed the basis of this new museum, to be part of the Smithsonian and built on the national mall in Washington, DC.

The new museum opened in 2004. As a part of the new law, Native Americans had lobbied for a Smithsonian repatriation requirement to be included. In particular, the law affects all collections of American Native human remains and funerary items held by the Smithsonian Institution (although the vast majority of relevant human remains and funerary objects are in the Smithsonian Museum of Natural History, the law affects all of the Smithsonian museums, including the new Museum of the American Indian, American History Museum, etc.). The NMAIA dictates that all designated items are to be inventoried and, where possible, the Smithsonian will determine tribal or descendant origins. When the law was passed, the Smithsonian was estimated to have 19 250 sets of Native American remains (Smithsonian Repatriation Office, 2012). As of 31 December 2010, approximately 29% of this number have been offered for return, and about 20% of the total have actually been repatriated (estimated from Smithsonian Repatriation Office's website: http://anthropology.si.edu/repatriation/faq/index.htm).

Lambert (2012: p. 21) makes the point that the NMAIA marks a real "shift in archaeology law from the protection of archaeological resources to the legal acknowledgement of private ownership (or rights of stewardship) by a sector of the American population of a component of the archaeological record."

Legal representatives of appropriate tribes or descendants can request the return of remains and objects. The Smithsonian notifies tribes if they determine an affiliation, but tribes are also encouraged to submit a request for remains to be returned. If there is a disagreement between the Smithsonian and the requesting tribe, the Smithsonian Repatriation Review Committee can examine the case in detail and submit its recommendations to the head of the Smithsonian. The Review Committee is composed of archaeologists, physical anthropologists, tribal representatives, and tribal elders. The tribal representatives on the Committee have one more representative (and therefore one more vote) than the scientific representatives.

One of the most important results of the Smithsonian act is that systematic study of human remains and objects prior to return is a priority, and the law provides funding for a Repatriation Office. As noted by Killion and Molloy (2000), as well as by Ousley et al. (2005), an enormous amount of contextual knowledge and a systematic catalogue have been gained as a result of this law (Smithsonian Repatriation Office, 2012). Data recovered can be used in the future, and data recording has been done in a consistent manner. It is important to realize, however, that the NMAIA *only* affects collections held by the Smithsonian.

The NAGPRA was passed on 16 November 1990 (NAGPRA; 25 U.S.C. 3001 et seq.). NAGPRA is a far-reaching law that was crafted as a compromise between Congress, Native Americans, and the scientific community. Note that although NAGPRA does not apply to the Smithsonian, the Smithsonian has generally accepted and incorporated the regulations of NAGPRA, and the NMAIA was amended in 1996 to reflect this incorporation of NAGPRA regulations.

NAGPRA has several components:

1. It applies to federally recognized Native American tribes, as well as native Hawaiian organizations and Alaska native corporations.
2. In addition to human remains, NAGPRA includes the repatriation of associated and unassociated funerary objects (objects clearly part of a burial as well as objects that were likely funerary objects but for which details are unknown), sacred objects (objects required for rituals and ceremonies), and objects of cultural patrimony (objects that are owned by the group and cannot be owned or sold by an individual).
3. NAGPRA requires museums and federal agencies to identify and inventory human remains and associated funerary objects, and also produce a summary of unassociated funerary objects, sacred objects, and items of cultural patrimony by 16 November 1995. These inventories and summaries were distributed to the federal government as well as appropriate tribes.
4. NAGPRA established tribal ownership of Native American remains and cultural items recovered from federal and tribal lands after 1990, requiring procedures prior to excavating on such land whether the excavations were inadvertent or intentional.
5. Violation of NAGPRA includes prohibitions against trafficking, sanctions against museums for not meeting deadlines and procedures, and enforcement of the Act by the ability to go to court.

NAGPRA established a Review Committee similar in composition to the Smithsonian's Repatriation Review Committee, with the balance tilted toward Native Americans. NAGPRA also established some grants to museums, Indian tribes, and native Hawaiian organizations (although these funds have been very small relative to the need), and the law promulgated regulations to make the law work.

As Lambert (2012: p. 20) notes:

> These laws reveal significant changes in attitudes concerning the scientific value of archaeological human remains over the course of the 20th century, and illustrate the complexities that can arise when secular and religious ethics collide in the conduct of scientific research.

NAGPRA is important – its passage meant that the Federal government was acknowledging that some things in museums were not obtained in a way consistent with modern American ethics, even though the original acquisition may have been legal at the time. Fortunately, NAGPRA was also a legal impetus for beginning conversations between tribes and the scientific and museum communities. Some tribes and archaeologists had already been talking with each other, but NAGPRA mandated such discussions. As Lambert (2012: p. 22) notes: "In light of the history of disenfranchisement of indigenous people in North America, … NAGPRA's passage was an ethical victory that could be shared by all communities impacted by the law."

Key Issues and Current Debate

Archaeologists and museums operating in the United States today adhere to federal laws and policies concerning repatriation. While some were opposed to these laws initially, they now honor the law. More importantly, there has been a significant culture change among the public and among archaeologists and museum professionals, who accept that repatriation is appropriate federal policy. It is important not to minimize this culture change: repatriation is now considered the *de facto* result of most burial excavation in the United States.

Under NAGPRA and the NMAIA, repatriation does not have a timetable. Except for specific dates by which museums were required to report their holdings, there is no deadline for the time when repatriations must be completed. This was done deliberately since it was impossible to estimate how long this process would take, and no one wanted to put the burden of a timetable on tribes who were already overwhelmed by the numbers of potentially related remains and artifacts in museums.

A key issue in the passage of NAGPRA was the concept of 'cultural affiliation.' Cultural affiliation is the basis for decisions about which remains are open to repatriation claims. NAGRPA's definition is "a relationship of shared group identity which can be reasonably traced historically or prehistorically between a present day Indian tribe or Native Hawaiian organization and an identifiable earlier group" (25 USC 3001, Sec. 2[2]). If there is no clear evidence for such a relationship, NAGPRA specifies establishing cultural affiliation based on a "preponderance of the evidence based upon geographical, kinship, biological, archaeological, anthropological, linguistic, folkloric, oral traditional, historical, or other relevant information or expert opinion" (25 U.S.C. 3005, Sec. 7[4]). It is important to understand that preponderance of evidence is a low legal threshold of simply >50%, and all types of evidence are to be treated equally. The notion of cultural affiliation has been problematic in some cases and is the basis for a number of legal and ethical debates about the meaning and intent of the law.

Many in the public think that the so-called Kennewick Man case legally settled the issue of cultural affiliation, but this is not the case (Thomas, 2000). Kennewick Man (found in Washington state) was a legal battle over a very ancient skeleton. A group of scientists and a group of citizens (Friends of America's Past) were on one side, and five Native American tribes and bands from the Columbia River Basin were on the other side (although technically the case was against the United States). Not only did the skeleton date to a very early time, some physical anthropologists argued that the skeleton had characteristics that were more European than Native American in nature. The details of the case can be found in Thomas (2000) and elsewhere. Unfortunately, instead of deciding the case on the basis of cultural affiliation – a finding that might have helped clarify decisions made under NAGPRA – the US Court of Appeals for the Ninth Circuit argued that the Federal Government had failed to prove that the skeleton was Native American and therefore it was not subject to NAGPRA. It was a decision that angered Native Americans and also angered many archaeologists because the decision was based on an idea about which Native Americans and archaeologists generally do *not* disagree – both sides would say that ancient remains found in the United States can generally be considered Native American.

Subsequent to the Kennewick decision, various attempts were made to amend and/or adapt NAGPRA. Most importantly for the present discussion was the Department of the Interior's 2007 decision to address the issue of 'culturally unidentifiable human remains' (CUHR). This category includes the largest majority of archaeological human remains in the United States. Included are those remains that museums and federal repositories have not been able to affiliate with a modern Native American tribe or descendant. Following the lines of evidence provided in the definition of 'cultural affiliation' (and outlined earlier), these are remains for which institutions have not been able to assign an affiliation. For the most part, these remains are those of greater antiquity, with fewer ties to the present. Members of the scientific and museum communities had worried about this group of remains from the very earliest repatriation discussions because archaeologists and museums were concerned about the impact of repatriation law on the study of prehistory. Some Native American groups were also concerned about the rule from both practical and legal perspectives.

The final rule for CUHR was enacted in May 2010. The original negotiations involved in crafting NAGPRA had carefully maintained a balance between various interests. The new rule ignores that balance and favors native issues over any scientific interests or concerns; it also does not necessarily give tribes much leeway in how they wish to respond to repatriation. It is likely that this rule will be the focus of future legal challenges.

One issue in the final CUHR rule focuses on the term 'disposition,' or the outcome of a repatriation claim. NAGPRA does not mandate the nature or type of disposition for successful repatriation claims, intentionally avoiding use of the term 'reburial.' The reason for this avoidance is that any particular tribe might or might not wish a particular or mandated outcome. While most tribes would likely wish reburial, if the repatriation decision is that the remains belong to the group, the specifics of the final disposition should also be the group's prerogative. The CUHR Final Rule effectively calls for the reburial of all Native American human remains currently in museums and federal repositories, even if evidence for any relationship of shared group identity is lacking.

Although many contentious issues remain, current debate in archaeology has moved from the question of whether or not

repatriation is right or wrong, to more complicated issues. Colwell-Chanthaphonh and Ferguson (2006) present one aspect of this new direction. Instead of principles and rules of ethics, they promote the idea of Virtue Ethics. "Instead of beginning with questions of obligations and oughts, Virtue Ethics begins with questions of character, focusing on relationships and the subjectivities of social interaction" (Colwell-Chanthaphonh and Ferguson, 2006: p. 118). They argue that we need to focus on agents and sustained actions, rather than single acts. Effectively, Colwell-Chanthaphonh and Ferguson (2006) argue for a more complex, layered discussion about archaeological practice. They note that archaeologists have a series of unique relationships with multiple interactions, and these interactions have consequences for ethical standpoints.

International Perspectives

In their broad discussion of ethics of bioarchaeology, Larsen and Walker (2004: p. 114) outline three ethical principles for the study of human remains, whether in the United States or internationally: (1) human remains should be treated with dignity and respect; (2) descendants should have the authority to control the disposition of their relatives' remains; and (3) because of the importance of human remains for the understanding of our common past, human remains need to be preserved when possible so that they are available for scientific research. From the discussions earlier, it is clear that there may not be universal agreement on these principles.

Canada and Australia share with the United States a basic issue associated with repatriation: the majority of the contested collections derive from minority indigenous populations where treatment of the dead is an important part of broader indigenous rights issues. Most repatriation efforts focus on trying to redress historical wrongs associated with Western colonialism (see DeLeiuen's discussion and summary of the Australian situation in this volume).

Buikstra (2006: pp. 408–409) notes that Canada has no federal legislation similar to NAGPRA, but relationships between First Nations and archaeologists are strong. Why the difference between countries? First, the Canadian archaeological profession voiced their sympathy with First Nation concerns relatively early, with physical anthropologists practicing *in situ* analyses and reburials during the 1970s. In other words, Canadian bioarchaeologists were collaborating and consulting with First Nations over a decade before US archaeologists, and they did so without being mandated by legislation, producing impressive work with multivocal perspectives. Interestingly, there are multiple examples of First Nations requesting osteological analyses, and allowing destructive tests. The National Museum of Canada stopped accessioning human remains in the early 1970s. As Buikstra (2006: p. 412) notes, "Canadian museums, physical anthropologists, and archaeologists appear to favor mediation over litigation in addressing repatriation issues."

Lambert (2012: pp. 25–26) notes that while the United States, Canada, and Australia have seen repatriation on a national scale coming from indigenous populations and issues relating to colonialism, some similar historical situations exist in Argentina, Norway, and South Africa. However, to date, repatriation efforts in these countries have been limited, and generally involve the return of named, historically significant individuals. Recent developments indicate that laws and policies similar to those in the United States are likely.

In Israel (Nagar, 2011: p. 8), opposition to the excavation of human remains is associated with sites dating to Classical Antiquity and comes from a group of extremist religious Jews who believe that graves of Jewish dead should always remain undisturbed; their political power has been successful in gaining some severe restraints on research. Nonetheless, the Israeli Antiquity Authority created and has maintained a human osteological database since 1994; all time periods are included (Nagar, 2011).

In Great Britain, considerable dialogue has focused on repatriation and is in part the result of connections to developments in Canada, Australia, and the United States. However, Parker Pearson et al. (2011) provide a recent summary of the issues and concerns. In 2008, "the British government announced that all human remains archaeologically excavated in England and Wales should be reburied after a 2-year period of scientific analysis" (Parker Pearson et al., 2011: p. 1). Not surprisingly, this announcement caused great consternation on the part of museums and archaeologists. As of the moment, it appears that the policy will not be implemented as stated, but Parker Pearson et al. (2011: p. 3) point out that

> The weaknesses and contradictions in the legal frameworks that govern archaeology are shocking in many respects, and cannot remain unresolved; archaeologists are pulled hither and thither by property law and planning law, while human remains exist in a peculiar (but ethically and morally appropriate) legal limbo in that they cannot technically be owned.

In sum, outside of Canada and Australia, laws regarding repatriation and reburial are scattered and sometimes under debate; they resemble the issues raised for the United States when there is a similar historical colonial situation. Whether or not the twenty-first century sees an increase in repatriation and reburial remains to be seen.

Future Directions

Larsen and Walker (2004) discuss the ethics of bioarchaeology and their view of the future. They argue that compromise is the key to reaching an ethical solution to skeletal studies of ancient ancestors. Compromise can be legislated, as it originally was for NAGPRA, or it can come about through working together. However, Larsen and Walker (2004: p. 118) note that this balance achieved through compromise is now endangered with "a perceptible shift in the balance of power toward extremist native perspectives on repatriation." They close their discussion with the comment that "compromise is developed when there is trust between two parties who have invested time and effort in reaching solutions that may not be perfect, but nonetheless attempt to [address] competing interests of different groups" (Larsen and Walker, 2004: p. 118).

An excellent example of a historic cemetery study that was inclusive from the beginning to the end of the project is the Tucson (Arizona) Alameda-Stone cemetery (Heilen, 2012). The planned building of a new court complex in downtown Tucson required the discovery and removal of a very large historic (1860–80) multiethnic cemetery. Although the project requirements were outlined by a court order, the project was successful because of transparency of the work to the affected parties, and the development of relationships between the various parties. The project developed community-sensitive methods for identifying cultural affinities, as well as addressing a variety of community concerns and requirements. The cemetery included Native American, Hispanic, European, and various other groups, all of whom were involved in the excavation, analysis, and reburial of the remains and artifacts.

The notion of trust is an important one, and will continue to be so in the future, both in the United States and internationally. As noted earlier, Colwell-Chanthaphonh and Ferguson (2006) use the concept of Virtue Ethics to look to the future without specific reference to NAGPRA. They also discuss the concept of trust, and the nature of trust in archaeological practice. They draw on the work of Baier (1994) to outline and discuss 'virtue ethics,' citing Baier's contention that goodwill is an essential component of trust. Colwell-Chanthaphonh and Ferguson (2006: pp. 122–124) outline the complex web of categories of trust relationships in archaeological practice, noting that there are not only relationships within categories, but also relationships between them. Trust requires goodwill, but each group has to interpret acts of goodwill as benevolent (Colwell-Chanthaphonh and Ferguson (2006: p. 127). If we wish to develop trust between archaeologists and museums and indigenous groups, all parties must move to a focus on people and actions over time, rather than things and single interactions at one time. Although it takes time to develop and establish trust, the goal has an excellent reward.

See also: Genetics and Indigenous Communities: Ethical Issues; Historical Thought and Historiography: Indigenous Cultures in the Americas; Indigenous Knowledge and Intellectual Property Rights; Indigenous Management Styles.

Bibliography

Baier, A., 1994. Moral Prejudices. Harvard University Press, Cambridge.

Buikstra, J.E., 2006. Repatriation and bioarchaeology: challenges and opportunities. In: Buikstra, J.E., Beck, L.A. (Eds.), Bioarchaeology: The Contextual Analysis of Human Remains. Academic Press, Elsevier, London, pp. 389–415.

Colwell-Chanthaphonh, C., Ferguson, T.J., 2006. Trust and archaeological practice: towards a framework of virtue ethics. In: Scarre, C., Scarre, G. (Eds.), The Ethics of Archaeology: Philosophical Perspectives on Archaeological Practice. Cambridge University Press, Cambridge, pp. 115–130.

Dongoske, K.E., Adenderfer, M., Doehner, K. (Eds.), 2000. Working Together: Native Americans and Archaeologists. Society for American Archaeology, Washington, DC.

Fforde, C., Hubert, J., Turnbull, P. (Eds.), 2002. The Dead and Their Possessions: Repatriation in Principle, Policy and Practice. One World Archaeology Series, No. 43. Routledge, New York.

Heilen, M.P. (Ed.), 2012. Uncovering Identity in Mortuary Analysis: Community-Sensitive Methods for Identifying Group Affiliation in Historical Cemeteries. Left Coast Press and Tucson: SRI Press, Walnut Creek, CA.

Hodder, I., 1985. Postprocessual archaeology. Advances in Archaeological Method and Theory 8, 1–26.

Killion, T.W., Molloy, P., 2000. Repatriation's silver lining. In: Dongoske, K.E., Aldenderfer, M., Doehner, K. (Eds.), Working Together: Native Americans and Archaeologists. Society for American Archaeology, Washington, DC, pp. 111–117.

Lambert, P.M., 2012. Ethics and issues in the use of human skeletal remains in paleopathology. In: Grauer, A.L. (Ed.), A Companion to Paleopathology. Wiley-Blackwell, Malden, MA, pp. 17–33.

Larsen, C.S., Walker, P.L., 2004. The ethics of bioarchaeology. In: Turner, T. (Ed.), From Repatriation to Genetic Identity: Biological Anthropology and Ethics. State University of New York Press, Albany, pp. 111–119.

Lovis, W.A., Kintigh, K.W., Steponaitis, V.P., Goldstein, L.G., 2004. Archaeological perspectives on the NAGPRA: underlying principles, legislative history, and current issues. In: Richman, J.R., Forsyth, M.P. (Eds.), Legal Perspectives on Cultural Resources. AltaMira Press, Walnut Creek, CA, pp. 165–184.

Mihesuah, D. (Ed.), 2000. Repatriation Reader: Who Owns American Indian Remains? University of Nebraska Press, Lincoln.

Nagar, Y., 2011. Human osteological database at the Israel antiquities authority. Bioarchaeology of the Near East 5, 2–18.

National NAGPRA office, 2012. http://www.nps.gov/nagpra/

Ousley, S.D., Billeck, W.T., Hollinger, R.E., 2005. Federal repatriation legislation and the role of physical anthropology in repatriation. Yearbook of Physical Anthropology 48, 2–32.

Parker Pearson, M., Schada-Hall, T., Moshenska, G., 2011. Resolving the Human Remains Crisis in British Archaeology, in Papers from the Institute of Archaeology. University College, London. http://dx.doi.org/10.5334/pia.369.

Scarre, C., Scarre, G. (Eds.), 2006. The Ethics of Archaeology: Philosophical Perspectives on Archaeological Practice. Cambridge University Press, Cambridge.

Shanks, M., Tilley, C.Y., 1987. Social Theory and Archaeology. Cambridge University Press, Cambridge.

Smithsonian Repatriation Office, 2012. http://anthropology.si.edu/repatriation/index.htm

Thomas, D.H., 2000. Skull Wars: Kennewick Man, Archaeology, and the Battle for Native American Identity. Basic Books, New York.

Trope, J.F., Echo-Hawk, W.R., 1992. The native American graves protection and repatriation act: background and legislative history. Arizona State Law Journal 24, 35–77.

Ubelaker, D.H., Grant, L.G., 1989. Human skeletal remains: preservation or reburial? Yearbook of Physical Anthropology 32, 249–287.

Walker, P.L., 2002. Bioarchaeological ethics: a historical perspective on the value of human remains. In: Katzenberg, M.A., Saunders, S.R. (Eds.), Biological Anthropology of the Human Skeleton. Wiley-Liss, New York, pp. 3–39.

Watkins, J.E., 2003. Beyond the margin: American Indians, first nations, and archaeology in North America. American Antiquity 68 (2), 273–285.

Zimmerman, L., 2000. A new and different archaeology? With a postscript on the impact of the Kennewick dispute. In: Mihesuah, D. (Ed.), Repatriation Reader: Who Owns American Indian Remains. University of Nebraska Press, Lincoln, pp. 294–306.

Archaeology: Philosophy and Science

Timothy Webmoor, University of Colorado at Boulder, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A. Wylie, volume 1, pp. 614–617, © 2001, Elsevier Ltd.

Abstract

Archaeology and philosophy share a long history of engagement. This engagement was most pronounced in the 1960s and 1970s with the so-called New Archaeology that drew upon philosophy of science to build an epistemological and methodological framework. The ensuing debates concerning how to adequately do this resulted in archaeologists being largely disaffected with philosophy of science. In the 1980s and 1990s, while philosophy of science shifted toward a more practice-based post-positivist philosophy, increasingly merging with the interdisciplinary field of science studies, archaeology began to engage with a wider suite of philosophies, including continental and pragmatist sources. Currently, rather than epistemic issues, archaeology and science studies share a growing interest in ontology and the role of materiality in human life.

Archaeology is a dirt-driven enterprise. Despite appearances and certain clichés aside, it is also philosophically sophisticated. In part, this is due to the long history of engagement between archaeology and philosophy. This interaction was most pronounced during the ascendancy in the 1960s and 1970s of what was self-described as the New Archaeology. A legacy of the New Archaeology is the continued association of philosophy in archaeology with a particular set of issues. These issues center upon the scientific nature of archaeology and the type of philosophy involved is, appropriately enough, the philosophy of science. While this was an especially influential period of mutual interaction between the fields, the effects of which continue to reverberate through archaeology, this is only a portion of the history. Recent developments, both within philosophy of science and in archaeology's approach to philosophy more broadly, have considerably altered how the fields engage. There is increasing importance of the interdisciplinary field of science studies and a broader range of philosophies that archaeologists find useful.

In this contribution, I sketch archaeology and philosophy's history of engagement, emphasizing the legacy that explicit epistemological questioning continues to have upon archaeologists' attitude toward philosophy. I discuss the proliferation of 'styles of reasoning' in archaeology, contributing to many camps operating somewhat autonomously within research traditions. I relate this trend to philosophy of science's turn to practice-based reasoning and its increased identification with science studies. I note the broader, more eclectic range of philosophies used by archaeologists, including continental and pragmatist sources. I conclude with discussions of ontology, as opposed to epistemology, that are increasingly encountered in science studies and archaeology and the wider emphasis across the social sciences and humanities upon the material dimensions of life.

History of Engagement

Antiquarianism and Early Calls for Science

There has been a long history of explicit engagement between the fields of archaeology and philosophy. As with other sciences to emerge from the antiquarianism of the eighteenth and nineteenth centuries, the discipline's earliest practitioners were Renaissance men like John Aubrey or Walter Scott who were well versed in the philosophical writings of the day. By the late nineteenth century, archaeology increasingly distinguished itself from the supposed dilettantism of this earlier natural philosophy. Eager to develop their field as a scientific enterprise, archaeologists at the turn of the twentieth century drew upon discussions concerning hypothesis testing and the use of the scientific method in other, more mature sciences like geology. From then on, archaeology experienced philosophical crises roughly every 20–30 years (Wylie, 2002). During these introspective periods, engagement with philosophical sources and debate within archaeology over scientific procedures and reasoning become more visible.

Around World War I, Roland Dixon (1913) and Clark Wissler (1917) published works advocating a more systematic and problem-oriented form of artifact collection. As Wylie (in Rathje et al., 2012) notes, it was hoped that this promotion of scientific reasoning would move the discipline from artifact and data collection for its own sake, a practice associated with antiquarianism, toward a 'real, new archaeology,' as Wissler called it (1917). Despite these discussions, data collection remained the *modus operandi* of archaeology during this culture-historical period in archeology. Disparagingly branded as narrow empiricism by later archaeologists, this period of data amassment and systematization could be approvingly characterized as epistemologically cautious, that is, a period when scholars did not draw conclusions that were unsupported by their data. Nonetheless, dissatisfaction with the interpretive timidity of culture-historians reemerged in the 1930s. These critics, such as Clyde Kluckhohn (1939) and Walter Taylor (1948), drew upon a range of philosophical literature to insist upon the need to build up a body of theory in order to provide explanations for the growing corpus of archaeological data. In Kluckhohn's opinion, archaeologists had barely progressed beyond antiquarian collecting. What was needed was a theoretical framework that posed questions and guided the recovery of relevant data.

A wider, trans-Atlantic set of debates developed in the 1940s and 1950s. These debates continued the previous disagreements over the role of theory in archaeology. They were much farther reaching, however, questioning the function and

veracity of typologies and the nature of the relationship of archaeologist to the archaeological record. These have come to be regarded as the typology debates. Essentially, the dilemma was as follows: were typological schemes, which archaeologists developed based upon formal variations of artifacts, directly indicative of inherent past culture traits, or did such schemes reflect the heuristic purposes of archaeologists? Within philosophy this resembles the nominalism versus conceptual realism controversy. The various responses within archaeology to this perennial problem were situated along a spectrum polarized by Ford's (1954) constructivism and Spaulding's (1953) common-sense objectivism. Constructivism, much like philosophical nominalism, embraced what would later be explicitly formulated as the theory-ladenness, or the impossibility of completely separating theory from data, of archaeological explanation. It describes the typological schemes that archaeologists deploy as tools for imposing order on what would otherwise be indistinguishable material. Archaeologists, interested in certain questions, inevitably impose a subjective ordering on the continuum of artifact assemblages. For the constructivists, such orderings could not be tested against the reality of the archaeological record.

Against this subjectivist position, and holding much in common with conceptual realism in philosophy, the objective stance viewed the categorization by archaeologists as a process of making manifest, or discovering, what was inherent in the record. These scholars rejected this version of theory-ladenness, or the supposed subjective ordering on the part of archaeologists, and adhered to a relationship of independence between the observer (archaeologist) and the real world (archaeological record). This latter position found resonance with the influential logical positivist philosophers of science of the time (e.g., Carnap, 1947). While the dilemma was often cast in the form of a polarity, the majority of archaeologists opted for a middle ground between the two positions. Such a middle position seemed to offer a workable solution, allowing archaeologists to get back to business and create the typologies so central to the operation of the discipline without accepting either relativism, the outcome of a subjectivist approach, or naïve empiricism, the outcome of an objectivist approach (*see* Cultural Relativism, Anthropology of). The general pattern of archaeologists bolstering polarizing epistemic positions with philosophical resources and then settling upon a workable middle ground will be encountered again.

The New Archaeology and Philosophy of Science

In the 1960s and 1970s, philosophy of science was drawn upon to support the theoretical and methodological development of a 'new' archaeology. This time Carl Hempel's (1965) logical empiricism, a modified version of positivism, was imported by outspoken archaeologists, most notably Lewis Binford. While many of the themes encountered in previous periods of contact with philosophy were the same, such as the explicit goal of establishing criteria for the scientific formulation and evaluation of research questions, the New Archaeology represented an especially sustained engagement. It also drew in the participation of more archaeologists and, eventually, philosophers of science. At first there was a fairly direct importation of Hempel's ideas concerning covering-law (C-L) explanation. Modeled upon physics, which was considered to be the paragon of scientific method and reasoning, Hempel nonetheless felt that other, 'softer' disciplines could equally generate universal laws applicable across time and cultures.

Archaeologists such as Binford were keenly aware of the discipline's inability to directly observe its subject matter: namely past societies. However, the natural sciences, like physics, routinely produce generalizations concerning indirectly observed phenomena through the use of mediating instruments of measurement and observation. By adopting a Hempelian epistemological framework, New Archaeologists reasoned that the discipline could equally develop a body of law-like generalizations through the mediating aid of middle-range theory (MRT). It was hoped that MRT would serve as a tool permitting archaeologists to observe past dynamic societies in the static archaeological record (*see* Archaeology, Theory in). Explanation of the archaeological record would be according to Hempel's C-L model. Under the C-L model, particular facets of the archaeological record could be sufficiently explained by fitting or subsuming the data to the body of law-like correlations established by MRT. This would move archaeological reasoning from induction, or amassing enough data until patterns emerge to suggest explanations, to deduction. Deductive reasoning inverts this equation by beginning with hypotheses pertaining to proposed laws, and then systematically searching for data relevant to the empirical implications of such proposed laws. Relevant data meant that it could potentially either support or refute the hypotheses. This was the basic workflow to Hempel's hypothetico-deductive (H-D) model of confirmation.

However, to begin confirming hypotheses about the observed record and build explanations about what happened in the past, archaeologists needed the linking universal laws. According to Binford this meant establishing "unambiguous cause-and-effect relationships between the causal dynamics and the derivative statics" (1983: p. 14). The operationalization of the New Archaeology would be a two-step process, then. For this reason, much archaeological attention turned to the processes responsible for transforming past behavior into the observed archaeological record. Two influential programs developed to secure the sociomaterial laws necessary for New Archaeology's fulfillment of the Hempelian ideal of scientific inquiry. Binford himself turned to ethnoarchaeology to study the archaeological record 'in the making' for his MRT (*see* Ethnoarchaeology). Michael Schiffer and his colleagues established a behavioral archaeology (BA) to study the relationships between human behavior and material culture in all times and places through ethnoarchaeology, experimental and modern material culture studies (Reid et al., 1975). There are important differences between these two approaches that bear upon New Archaeology's Hempelian philosophy. Most significant is the scale or applicability of generated laws. For instance, Binford, drawing upon a model of culture as a contained, self-functioning system, limits his generalizations derived from observation to the system-wide level. For Hempel inspired science, established laws were to be akin to those in physics and the other natural sciences, not restricted to spatiotemporal dimensions. Schiffer picks up on this inferred deficiency and attempts to begin at a nonculturally specific level of processes in order to generalize universally and cross-culturally

(Tschauner, 1996). Nonetheless, it was less the specific differences between these two influential currents of research and more the general assumption of uniformitarianism that united them which presaged the stringent criticisms of the New Archeology.

Post-Positivist Philosophy

Uniformitarianism, or the assumption that the principles observed in the present operated identically in the past, is central to both MRT and BA. It is difficult to imagine pursuing archaeology without such a belief. For philosophers of science, uniformitarianism represents the intractable problem of theory-ladenness. It is a more general example of where specific background assumptions influence the course of research. For Hempel, the division of theory from observation was crucial in his H-D model of independent testing. Yet, by the time New Archaeology was producing manuals on how to conduct Hempelian science (e.g., Watson et al., 1971), philosophy of science had largely conceded theory-ladenness. In large part this was due to internal commentators like Thomas Kuhn (1962) who convincingly demonstrated the role of 'extra-scientific' or social, political, and ideological factors that propel scientific progress. For archaeology, the substantive criticism of theory-ladenness and the possibility of the New Archaeology program as modeled upon Hempel's ideals quickly digressed into disciplinary quarrels. Philosophers of science chastised the new archaeologists for not only misunderstanding the details of Hempel's positivistic platform, but for being behind the times in not engaging the recent literature on post-positivist philosophy (Levin, 1973; Morgan, 1974). Archaeologists were being accused of not only doing bad science, but bad scholarship too.

At the same time, there was growing concern within archaeology that, despite drawing upon Hempel's models, the pursuit of meaningful generalizations was proving disappointing. Generalizations about the past were either severely restricted to specific systems or contexts, or too general as to be derided as 'Mickey Mouse laws' (Flannery, 1973). For instance, "objects left on a sloping archeological site wash downhill" (Flannery, 1982: p. 270). The quest for Hempelian covering laws was fairing little better in other sciences. Unsurprisingly, many New Archaeologists became somewhat disillusioned.

A result of these developments was an understandable disaffection with philosophy of science. Kent Flannery (1982) poignantly channeled this frustration with philosophy in archaeology in his parody of the 'Golden Marshalltown.' Philosophically minded archaeologists had lost touch with the dirt-driven details of actually doing archaeology. Robert Dunnell (1982) suggested that concept borrowing from philosophy of science had done a disservice to the discipline and that what was called for was an introspective examination of metaphysical principles. Schiffer (1981) urged the continuation of philosophical analysis in archaeology, but he noted that philosophers needed to better understand the epistemic problems particular to archaeological practice. This stance toward philosophy echoed David Clarke's (1973) earlier call for archaeologists to develop a discipline-specific philosophy.

Several philosophers sympathetic to archaeological pragmatics offered post-Hempelian advice. Drawing upon statistics, Merrilee Salmon (1982) discussed statistical-relevance (S-R) accounts of explanation and Bayesian confirmation and inference. Given the observed data, these models generate a statistical probability for how likely a specific hypothesis or explanation is relative to competing hypotheses. They also account for the updating of belief in a particular explanation with the accumulation of new data. The partnership of archaeologist Jane Kelley and philosopher Marsha Hanen (1988) produced an account drawing, in part, from an altogether different source of philosophy with pragmatism, embracing inference to the best explanation. For them, a satisfactory explanation is one that has more 'cash value,' to use pragmatist wording. Relative to competing explanations, it is the one that empirically coheres with more of the data and so accounts for, or explains, more. At the same time, an adequate explanation also theoretically satisfies or fits established principles. Guy Gibbon (1989), an archaeologist trained in philosophy of science, presented arguments from scientific realism that, contrary to positivist traditions, admitted 'unobservables' into the explanatory process. This is an important concession for archaeology, given the previous discussion of MRT and the admission that the causal processes responsible for the archaeological record are no longer directly observable. Similar to Kelley and Hanen, his scientific realism advocates a general version of inference to the best explanation. Finally, Peter Kosso (1992) unpacked inference to observation and draws upon models of multicomponent analysis used in the natural sciences.

Epistemic Settlements

Alison Wylie (2002), in a work that collects and develops her many essays bridging philosophy of science and archaeology, identifies the 'interpretive dilemma' that the discipline recurrently faces, demonstrating how theory-ladenness represents the most recent formulation of the intractable issue of relativism versus objectivity. Drawing upon scientific realism and feminist standpoint theories, notably Sandra Harding's 'strong objectivity' proposal, Wylie makes a compelling argument for mitigating the subjective bias of the observer (theory-ladenness) through, in part, turning the argument on its head. She argues that there are multiple checks, or 'tackings,' that occur both within and across the source-side (context of interpreter) and subject-side (relevant data) in the process of archaeological interpretation. Rather than attempt to remove subjective bias involved with background knowledge, assumptions, and proposed models, an ideal that post-positivist philosophy discredited, Wylie suggests bringing more interpreters and source-side models and assumptions to bear upon the data. Put simply, multiple sources of bias serve to mutually constrain one another in the interpretive process. For the subject-side of the equation, Wylie similarly explores how different sources and types of data interact to place limits on potential interpretations. Taken as a whole, these empirical and subjective constraints enter interpretation at several points, making it difficult for a coherent account to emerge. Wylie deploys the Peircean metaphor of the cable to indicate the strength of the inferences that result from weaving together these various components or strands of the interpretive process. Rather than a deductive procedure, Wylie's 'ampliative inference' nevertheless results in what she terms a 'mitigated objectivism.'

Wylie's epistemic settlement, much of which was worked out in earlier essays, appeared during this period of circumspect rapprochement with philosophy of science. Other epistemic settlements, in the wake of the New Archaeology, appeared over the same time. Notably Binford (1989), while accepting that observations are paradigm dependent (theory-laden), nonetheless advanced his more moderate 'relative objectivity.' By distinguishing between 'first order observation' and 'pattern recognition work,' Binford argued that archaeological data have a chance in the latter step to talk back, or challenge our initial observations. Bias enters into 'first order observation,' where human choice influences which attributes or qualities of the archaeological record are deemed relevant. However, the next stage of 'pattern recognition work' removes bias from the gathered data because it involves nonhuman, and so objective, measurements by technology. For example, while the archaeologist must identify and sample features or artifacts for radiocarbon analysis, a process that involves a host of decisions, the instruments and technology producing a range of dates for the samples do not involve a human 'subject.' They therefore produce objective data.

Worrying about a discipline strung along the spectrum from idealism to realism to positivism, Bruce Trigger (1998) drew upon Darwinian common sense to suggest that knowledge claims, in the present or past, work to aid our existence in the real world. The general idea is that for knowledge claims to be true they must correspond to reality. No matter how probable they initially appear, if such explanations do not match reality then they will hinder our adaptation to our complex environment. Consequently, false claims will be deselected over time. The hope was that such an evolutionary approach to epistemology, one that resonates with pragmatist accounts of truth confirmation, might offer a unifying, common-sense 'moderate relativism.'

Initially cast as hyper-relativist, Ian Hodder, over the course of the 1990s, drew upon hermeneutics, or the philosophy and method of interpretation of texts, to propose a 'guarded objectivity' as a solution to the evaluation of archaeological claims. With other so-called post-processualists, such as Michael Shanks or Christopher Tilley, Hodder argued for a strong contextualism of the archaeologist. While these archaeologists drew upon an alternate philosophical literary network, this might be understood as an expanded form of theory-ladenness. More strongly than its formulation in the philosophy of science, Hodder and his colleagues argued that the contemporary context of the archaeologist must be considered. It is a contextualism that draws more attention to the wider social, economic, and political factors brought to bear in archaeological activities. Hodder's (1999) framework for how these influences take part in archaeological research was described as the hermeneutic spiral of interpretation (**Figure 1**).

Rather than a vicious circle of interpretation, one leading to relativism, Hodder presents a reflexive analysis of how archaeologists make claims about the past and argues that the process is never ending (or only arbitrarily suspended). No definitive hypothesis testing or protocol is given for determining when a question has been decisively answered. Instead, questions lead to partial answers and, more importantly, the answers provided lead to further questions. These furthering on questions then prefigure, as incorporated into part of the archaeologist's own context, additional queries of the past. There are multiple, ongoing opportunities for the data to resist or disrupt the process, which would generate new questions. The result is a dialectic or, to use Gadamer's (1989[1960]) hermeneutic phrase, a 'fusing of horizons,' between archaeologist and past materials.

While there are sympathies in this process to Wylie's idea of multiple points of contact between the empirics and the archaeologist's background assumptions, Hodder's epistemic settlement differs in granting the archaeological material more agency in the process. It marks an increasing concern with ontology, or questions of what entities exist and how they relate, as opposed to epistemological concerns with how we

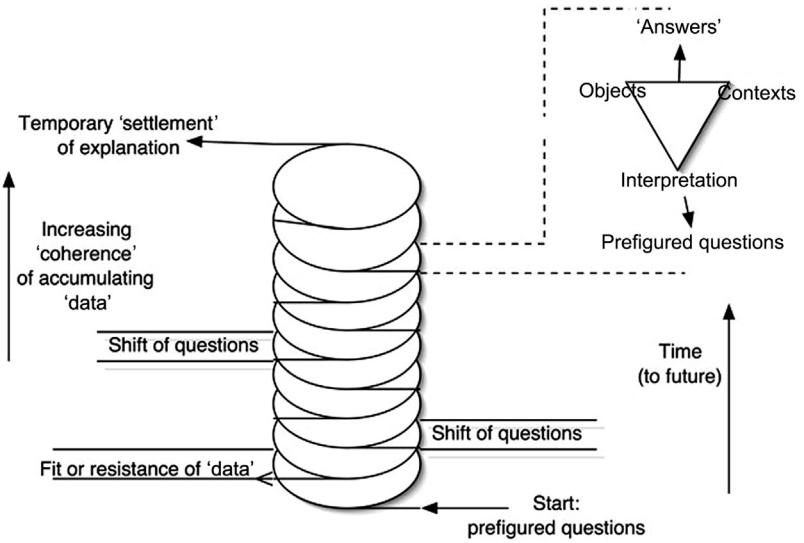

Figure 1 Visual rendering of Hodder's hermeneutic spiral of interpretation After Hodder, I., 1999. The Archaeological Process. Blackwell, Oxford.

know about such entities. While twinned to epistemology to form the pillars of Western philosophy, ontology had been little regarded by analytic philosophy, of which philosophy of science is a part. Furthermore, ontology and broader issues of metaphysics were explicitly regarded by the Vienna School, the progenitors of positivism, as nonsense and superfluous to the purpose of creating an observational language to facilitate making knowledge claims in science.

The Disunity of Science and Science Studies

Put within disciplinary context, these epistemic settlements can be appreciated as attempts for unification at a time when archaeology was seen to be splintering into a "thousand archaeologies" (Schiffer, 1988: p. 479). Some view the plurality of approaches as beneficial, drawing upon metaphors of democratic inclusion and racial diversity (e.g., Hodder, 2003). Others believe it to be worrisome and attempt to accentuate shared operational and reasoning principles among archaeologists despite apparent theoretical divergences (e.g., VanPool and VanPool, 1999). For example, despite the polarization, processual and post-processual practice was argued earlier on to be very similar. Tschauner (1996) suggested that post-processualists such as Hodder rely upon MRT and principles of BA to make arguments about the past in a manner similar to processualists. It has also been pointed out that regional techniques aside, methodologically both camps excavate in a comparable manner (Chadwick, 2003).

While regional specialties and methodologies create differences in practice, the perception of splintering is due in large part to groups of archaeologists orienting their craft according to versions of the epistemic settlements presented above. While these settlements are general heuristics of how archaeological reasoning proceeds, they nevertheless are informed by more specific and divergent background assumptions (such as how culture is modeled), operational procedures (is data acquisition strictly problem-oriented), and evaluative work flow (fitting in coherentism or testing under falsificationism). A result is that some archaeologists practice hypothesis testing or statistical relevance, others operate according to interpretive hermeneutics, and many more mix the various procedures in their overall practice.

Whether detrimental or beneficial for the discipline, the disunity of archaeology into semi-autonomous working camps is related to the discipline's former reliance upon the philosophy of science as well as the larger trend within philosophy of science toward dispersion. Former debates about how to properly pursue explanation of the past were more focal for the discipline, particularly with respect to the New Archaeology. They also drew, importantly, upon a small range of philosophies in their arguments (or in the case of the New Archaeology, two or three prominent philosophers). Discussion could anchor itself to common philosophical ground, as it were. Indeed, the unification of science was the explicit function of positivism according to the Vienna School founders. But whereas philosophers of science or philosophically minded archaeologists formerly could aid in arbitrating debates, there is now a philosophical vacuum left by the dispersion of philosophy of science into science studies.

Like Clarke (1973) or Schiffer (1981), within archaeology, the need to attend to the specific problems, practices, and history and social setting of a discipline was increasingly voiced by philosophers of science themselves. The problems of fit between philosophical models for inquiry and disciplinary practitioners' ability to put them to work were not restricted to archaeology. Other disciplines were experiencing a troubled engagement. Kuhn's (1962) analysis of how science progresses meant, among other things, that scientific practice was not served by idealized heuristics. It was far messier and contingent. Philosophers of science would need to work in the trenches (sometimes literally) to appreciate these contingencies and incorporate them into their models if they were to be of help. This recognition created a sort of crisis of relevance for philosophy of science (Stoll, 2000). Clearly, generating a single, comprehensive, and unified theory for how science works was no longer the goal of philosophy of science. Rather, it was necessary to acknowledge the specific workings of particular scientific disciplines in order to create accurate descriptions and models. In practical terms this meant that there were many sciences. In the face of what was called the disunity of science (Dupre, 1993; Galison and Stump, 1996), philosophers interested in science began to embed themselves in specific disciplines. Historians and philosophers of science, sociologists, and anthropologists, among others, increasingly found themselves under the umbrella of the emerging field of science studies. As Wylie (in Rathje et al., 2012) relates, the changes endemic to philosophy of science meant that if one was interested in pursuing a career, one had to be an 'amphibious philosopher': immersing oneself in a scientific discipline and obtaining practical training and credentials in the relevant fields of study. With philosophers of science attending more to the pragmatic details of particular disciplines, fewer general, trans-disciplinary programmatic statements have been offered. While some archaeologists still hold to concepts and heuristics of philosophy of science of the pre-Kuhnian period, these are largely implicit and infrequently discussed in open forums. Philosophy of science has largely gone underground in archaeology. Other varieties of philosophy have come to the fore of archaeological conversations.

Philosophical Pluralism in Archaeology

Due to the prominence of analytical philosophy in Anglo-American philosophy departments beginning in the early twentieth century, continental and pragmatist traditions within philosophy received little attention from archaeologists. There had been isolated exceptions, such as Raymond Thompson's (1956) use of the American pragmatist John Dewey. Yet with the growth of science studies and the decline in influence of analytic philosophy as handmaiden to science, many more archaeologists began to import ideas from continental sources. Interpretive or post-processual archaeologists drew upon a range of philosophers in their arguments against the positivist leanings of the New Archaeology. As noted above, Hodder cited hermeneutic authors, such as Collingwood, Dilthey, and Heidegger, and later Gadamer and Riceour, in developing his interpretive platform. Other post-processualists, such as Shanks and Tilley (1992), looked to social and political philosophers like Althusser, Bachelard, Derrida, and Foucault. Much less

consolidated around epistemic questions of scientific explanation, continental sources tend to be more interested in subjectivity and the role of the individual in society (as are some pragmatist thinkers). A broad range of social theorists were likewise enlisted to build arguments against some of the fundamental tenets that New Archaeologists had advanced. Among other differences, interpretive archaeologists construed culture as a dynamic between individual agency and societal structures rather than framed as the extrasomatic means of adaptation to the environment. The corollary meant that past cultures were not framed in ecomaterialist (or ecodeterminist) models, and so were less amenable to causal analysis. Hempelian models for scientific inquiry were viewed not only as overly idealistic, as post-positivist philosophers had argued with archaeologists, but also as inapplicable to archaeology's domain of study.

Such interpretive accounts, like Hodder's hermeneutic spiral, greatly contrasted with previous forms of evidential reasoning discussed in archaeology, whether analogical reasoning, simple inference and inference to the best explanation, falsificationism, and verificationism. Given the already circumspect relationship of archaeologists to philosophy of science, the new flurry of less familiar and seemingly eclectic philosophical sources in archaeological research cumulatively effected a largely negative reaction to the post-processualists. Perhaps perceived as adding insult to injury, there had been a fairly quick succession of (at times acrimonious) theoretical debates involving philosophy in archaeology.

The array of philosophical sources introduced to archaeology continues unabated. As one example, the compendium, *Philosophy and Archaeological Practice* (Holtorf and Karlsson, 2000), added philosophers as disparate as Feyerabend, Irigaray, Levinas, Merleau-Ponty, and Wittgenstein. The use of these philosophical sources varies greatly. Many are employed in a manner similar to what philosophers of science would regard as source-side work, that is, the development of new models for investigation in a manner similar to the role that ethnographic analogy served MRT. At other times, certain philosophies are used to build up a more thorough-going archaeological program, extending from problem formulation through to modes of investigation and sometimes explanatory justification. Tilley's (1994) use of Merleau-Ponty and Julian Thomas' (1996) work with Heidegger to develop phenomenological approaches to landscapes are notable examples. Here the archaeologist himself or herself is used as an analog for past individuals' perception of archaeological features. Another example is Robert Preucel's (2010) refinement of Hodder's initial work on hermeneutics and the active role of archaeological materials for the interpretive process with his use of American pragmatist C.S. Peirce's semiotics. Here Peirce's tripartite scheme of icon, index, and symbol is used to demonstrate how cultures and social groups variably ascribed meaning to materials and how such materials had a part within social contexts to determine their own significance.

In addition to these continental- and pragmatist-inspired approaches, there continues to be sporadic engagement with post-positivist philosophy. As a sampling, William Krieger (2006) argues for the mutual benefit that would result from archaeology reembracing the scientific realism of Bunge and Harré. Sandra Wallace (2011) presents Bhaskar's critical realism and suggests that its concern with open systems, in contrast to scientific ideals of closed, laboratory-like systems, could better model past societies. She also contends that critical realism's concern with 'depth ontology' could aid archaeology in avoiding unhelpful dichotomies, such as that between the social and material, or between explanatory tendencies of reduction and conflation. This is because critical realism is a relational approach to ontology that emphasizes shifting internal relations among different levels or strata that make up reality, such as the social, biological, or material. The hope is to avoid conventional conceptions of reality that are argued to be either too atomistic, with reality composed of individual and distinct bits, or overly holistic, with reality being an undifferentiated soup. In addition, and similarly to Gibbon's 'unobservables' in his scientific realism, absence is theorized to be real and have effects in Wallace's relational account.

Current Orientations

Despite this diversity of uses for, and types of, philosophy in recent archaeology, several emerging research orientations have refocused upon science studies as the contemporary inheritor of philosophy of science. Most of these work with ideas developed by ethnographers of science in observing the practicalities of 'science in action' (e.g., Latour, 1986). Sometimes identifying themselves as science and technology studies (STS), these observers have influentially drawn attention to the active role of things (objects, instruments, inscriptions) as they operate as networks with humans to accomplish tasks (Webmoor, 2013). Many of these STS scholars have explicitly rejected epistemology as being unhelpful due to its deep-seated assumption of a subject–object split (often attributed to René Descartes). One resulting mantra has been to avoid making assumptions beforehand regarding what or who the 'actants' are in any given investigation (hence modifying the human-centric term in favor of actant).

Andrew Jones (2002) draws upon science studies in his account of archaeological practice. Demonstrating that archaeological materials have agency in partially determining their interpretation, he offers an integrationist approach that obviates concerns over subjectivity and objectivity and might unite science-based and interpretive archaeologists. Combining insights from philosophy of mind, cognitive science, and STS, Carl Knappet and Lambros Malafouris (2008) develop Colin Renfrew's cognitive archaeology in nonanthropocentric directions. Drawing upon a range of archaeological examples, from experimental studies with pottery making to reflexive analyses of archaeological recording, Knappet and Malafouris focus attention on the manner in which action in the past is never the exclusive domain of social or material factors.

Those advocating relational ontology likewise bring attention to nonanthropocentric ideas of identity and agency in the past, such as the mutual constitution and dependence of humans, nonhumans, and other life-fellows. Christopher Watts (2013) brings together an assortment of archaeologists who draw upon STS authors, indigenous metaphysics, and continental philosophers, such as Deleuze, Guattari and Manuel De Landa, to develop a concept of existence as an

emergent, fluid assemblage of connected components rather than a fixed and stable given. The emphasis upon ontology for most of these scholars alters the role and practice of the archaeologist. It shifts from a traditional concern with accurately representing the reality of the archaeological record, to presenting alternate possible realities of the past.

Olsen et al. (2012) draw together this growing interest in ontology and material agency by articulating a dissatisfaction with the increased prevalence of social constructivism and social theory more generally in a discipline that is, as they remind archaeologists, surely 'the discipline of things.' Working with STS, but also engaging with a group of trans-Atlantic philosophers who identify as speculative realists (e.g., Harman, 2010), the authors present a series of arguments for how taking seriously the active (but nondeterministic) role of materials involved in the practice of archaeology might contribute to disciplinary self-confidence vis-à-vis philosophy or other social sciences. To do this they offer the principle of symmetry as an attitude regarding reality that avoids a priori theorizations about the nature of things, as in relational, atomistic, or holistic accounts. Instead, they advocate the unique expertise of archaeologists to empirically document settings whereby any conception of the realities involved is an outcome, not a starting point. For example, careful recording of an artifact and its context enables postexcavation analysis where it may emerge that objects and their qualities are the locus of explanation, or, alternatively, that the entire assemblage with its relations needs emphasizing, or that a dialectical explanation moving between singular objects and the larger contexts fits best. The point is that the principle of symmetry applied along the entire workflow keeps archaeologists agnostic until the empirics overwhelmingly compel. This ontological outlook avoids privileging relations among things, the relations themselves, or the inherent qualities of things themselves.

In all there has been a discernible shift in archaeology toward an engagement with science studies and a wider suite of philosophies. This has accompanied a greater complacency toward issues explicitly concerned with epistemology and a growing interest in ontology and the role of the material in human life past and present. Given the wider conversation taking place across the humanities and social sciences around the importance of material things, archaeology and philosophy's current engagement stands to make a vital contribution to these wider matters of concern. Despite the changes over their long history of interaction, the partnership of philosophy and archaeology stands to continue in new and interesting directions.

See also: Archaeology, Theory in; Cognitive Archaeology; Ritual and Symbolism, Archaeology of; Science and Technology, Anthropology of.

Bibliography

Binford, L., 1983. Working at Archaeology. Academic Press, New York.

Binford, L., 1989. Debating Archaeology. Academic Press, New York.

Carnap, R., 1947. Meaning and Necessity: A Study in Semantics and Modal Logic. University of Chicago Press, Chicago.

Chadwick, A., 2003. Post-processualism, professionalization, and archaeological methodologies. Towards reflective and radical practice. Archaeological Dialogues 10, 97–117.

Clarke, D., 1973. Archaeology: the loss of innocence. Antiquity 47, 6–18.

Dixon, R., 1913. Some aspects of North American archaeology. American Anthropologist 15, 549–566.

Dunnell, R., 1982. Science, social science, and common sense: the agonizing dilemma of modern archaeology. Journal of Anthropological Research 38, 1–25.

Dupre, J., 1993. The Disorder of Things: Metaphysical Foundations of the Disunity of Science. Harvard University Press, Cambridge, MA.

Flannery, K., 1973. Archaeology with a capital 'S'. In: Redman, C. (Ed.), Research and Theory in Current Archaeology. Wiley, New York, pp. 47–58.

Flannery, K., 1982. The golden Marshalltown: a parable for the archaeology of the 1980's. American Anthropologist 84, 265–278.

Ford, J., 1954. On the concept of types. American Anthropologist 56, 42–57.

Gadamer, H.-G., 1989[1960]. Truth and Method, second ed. Sheed and Ward, London.

Galison, P., Stump, D. (Eds.), 1996. The Disunity of Science: Boundaries, Contexts, and Power. Stanford University Press, Stanford.

Gibbon, G., 1989. Explanation in Archaeology. Basil Blackwell, Oxford.

Harman, G., 2010. Toward Speculative Realism. Essays and Lectures. Zero Books, Melbourne.

Hempel, C., 1965. Aspects of Scientific Explanation and Other Essays in Philosophy of Science. Free Press, New York.

Hodder, I., 1999. The Archaeological Process. Blackwell, Oxford.

Hodder, I., 2003. Archaeology as discontinuous domain. In: VanPool, T., VanPool, C. (Eds.), Essential Tensions in Archaeological Method and Theory. University of Utah Press, Salt Lake City, UT, pp. 5–10.

Holtorf, C., Karlsson, H. (Eds.), 2000. Philosophy and Archaeological Practice, Perspectives for the 21st Century. Bricoleur Press, Gotëborg.

Jones, A., 2002. Archaeological Theory and Scientific Practice. Cambridge University Press, Cambridge.

Kelley, J., Hanen, M., 1988. Archaeology and the Methodology of Science. University of New Mexico Press, Albuquerque, NM.

Kosso, P., 1992. Observation of the past. History and Theory 31, 21–36.

Kluckhohn, C., 1939. The place of theory in anthropological studies. Philosophy of Science 6, 328–344.

Knappet, C., Malafouris, L. (Eds.), 2008. Material Agency: Towards a Non-Anthropocentric Approach. Springer, New York.

Krieger, H., 2006. Can There Be a Philosophy of Archaeology? Processual Archaeology and Philosophy of Science. Lexington Books, Lanham, MD.

Kuhn, T., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.

Latour, B., 1986. Science in Action. Harvard University Press, Cambridge, MA.

Levin, M., 1973. On explanation in archeology: a rebuttal to Fritz and Plog. American Antiquity 38, 387–395.

Morgan, C., 1974. Explanation and scientific archeology. World Archaeology 6, 133–137.

Olsen, B., Shanks, M., Webmoor, T., Witmore, C., 2012. Archaeology: The Discipline of Things. University of California Press, Berkeley.

Preucel, R., 2010. Archaeological Semiotics. Blackwell, Oxford.

Rathje, W., Shanks, M., Webmoor, T., Witmore, C., 2012. Interview with Alison Wylie. In: Rathje, W., Shanks, M., Witmore, C. (Eds.), Archaeology in the Making: Conversations through the Discipline. Routledge, New York, pp. 93–121.

Reid, J.J., Schiffer, M., Rathje, W., 1975. Behavioral archaeology: four strategies. American Anthropologist 77, 864–869.

Salmon, M., 1982. Explanation and Archaeology. Academic Press, New York.

Schiffer, M., 1981. Some issues in the philosophy of archaeology. American Antiquity 46, 153–158.

Schiffer, M., 1988. The structure of archaeological theory. American Antiquity 53, 461–485.

Shanks, M., Tilley, C., 1992. Re-Constructing Archaeology: Theory and Practice, second ed. Routledge, London.

Spaulding, A., 1953. Statistical techniques for the discovery of artifact types. American Antiquity 18, 305–313.

Stoll, A., 2000. Twentieth-Century Analytic Philosophy. Columbia University Press, New York.

Taylor, W., 1948. A Study of Archeology. Southern Illinois University Press, Carbondale, IL.

Thomas, J., 1996. Time, Culture, Identity: An Interpretive Archaeology. Routledge, London.

Thompson, R., 1956. The subjective element in archaeological inference. Southwestern Journal of Archaeology 12, 327–332.

Tilley, C., 1994. A Phenomenology of Landscape. Berg, Oxford.

Trigger, B., 1998. Archaeology and epistemology: dialoguing across the Darwinian chasm. American Journal of Archaeology 102, 1–34.

Tschauner, H., 1996. Middle-range theory, behavioral archaeology, and postempiricist philosophy of science in archaeology. Journal of Archaeological Method and Theory 1 (3), 1–30.
VanPool, T., VanPool, C., 1999. The scientific nature of postprocessualism. American Antiquity 64, 33–54.
Wallace, S., 2011. Contradictions of Archaeological Theory: Engaging Critical Realism and Archaeological Theory. Routledge, London.
Watson, P.J., LeBlanc, S., Redman, C., 1971. Explanation in Archaeology: An Explicitly Scientific Approach. Columbia University Press, New York.
Watts, C. (Ed.), 2013. Relational Archaeologies: Humans, Animals, Things. Routledge, New York.
Webmoor, T., 2013. STS, symmetry, archaeology. In: Graves-Brown, P., Harrison, R., Piccini, A. (Eds.), The Oxford Handbook of the Archaeology of the Contemporary World. Oxford University Press, Oxford, pp. 105–120.
Wissler, C., 1917. The new archaeology. American Museum Journal 17, 100–101.
Wylie, A., 2002. Thinking from Things: Essays in the Philosophy of Archaeology. University of California Press, Berkeley.

Archaeology, Theory in

Staša Babić, Faculty of Philosophy, Belgrade, Serbia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.W. Preucel, volume 23, pp. 15647–15652, © 2001, Elsevier Ltd.

Abstract

From the inception of academic archaeology, the position of theory in the discipline has been marked by influences from other fields of research, both in terms of the concepts applied and the general outlook on the proper modes of investigation. The relationship between theory and method has remained an open issue for archaeologists, not resolved by two massive changes usually marked by the processual shift of the 1960s and the postprocessual critique from the 1980s. The recent developments aim to address this tension, but the academic community remains fragmented and archaeological theory diverges in several disparate directions.

The aim of the discipline of archaeology is to make inferences about the past based upon the material evidence of human activities. Consequently, the role of theory in archaeology is to provide the general framework of propositions and principles according to which the material record may be investigated in order to reach scholarly verified reconstructions of the past. This set of premises has changed throughout the history of the discipline and it has become customary, over the last couple of decades, to regard these developments in archaeological theory according to a pattern of stages, starting with the antiquarian beginnings and then proceeding to the culture-historical, processual, and eventually postprocessual strands of research (cf e.g., Hodder, 1991a; Jones, 2002; Johnson, 1999; Olsen, 2002; Renfrew and Bahn, 2005). This pattern, though a useful heuristic device in its own right, opens up several questions that need to be addressed. Firstly, as is always the case with generalized statements, it leaves aside a plethora of divergent paths and local particularities of archaeological theory and practice around the globe (Babić, 2014; Biehl et al., 2002; Gramsch and Sommer, 2011). Furthermore, this neat sequence may imply that as the discipline matured, so too did its theoretical apparatus, constantly moving toward more sophisticated ways of understanding the past (cf Hodder and Preucel, 1996: 4). And yet, "every once in a while archaeological theory gets a bashing" (Lucas, 2012: 1). This is a somewhat peculiar position for theoretical considerations inside an academic discipline since theory is usually considered to be the very foundation of any particular field of research. It has been suggested that these 'occasional bashings' are the result of the potential schism between archaeological theory and method, apparent at least from the 1960s (Lucas, 2012; cf also Jones, 2002). This enduring tension may be the result of the underlying assumptions and traditions that structure the work of archaeologists and the ways they accept, adapt, and apply theoretical concepts (Johnson, 2006).

Theory and Method

Indeed, from the inception of the discipline to the present day, archaeologists have been influenced by various ideas shaped in other fields of research, but have not always been inclined to overtly acknowledge these sources of inspiration. These idea transfers have been induced by the perceived similarities of the object and aims of research, as well as by the generally accepted premises of scientific work, not only in the field of humanities and social sciences, but also in the domain of hard science. Some of the basic principles of archaeological research have therefore been shaped by the generally prevailing ideas on proper scientific procedures and methods for procuring knowledge, regardless of the type of evidence and goals of a particular field of inquiry. Commenting on this, Julian Thomas (2004: 55) emphasizes archaeology's adherence to a conception of knowledge that privileges method over theory, ascribing this inclination to the discipline's eminently modernist character. He traces the roots of the primacy of method over theory back to early modern philosophy and René Descartes' *Discourse on Method*, which defined a program of rational understanding of the world through rigorous adherence to a defined procedure. Descartes' aim was to overcome the prejudices and uncertainties of the time by means of radical doubt and dismissal of any knowledge that could not be substantiated by hard evidence. In his program, Descartes was responding to his own age, when the European world was changing; fighting to come to terms with these changes, most of them caused by the upheavals of the Reformation; and rapidly expanding scientific discoveries and encounters with non-Western societies through colonizing and mercantile expansion (cf also Gosden, 1999, 2004). Therefore, Descartes asserted that the gathering of knowledge from the world had to be carefully regulated, guided by the use of reason, ultimately leading to universal truth. Methodology thus became a means to order and regulate the process of knowing the world and its absolute and objective truths that was not dependent upon theoretical assumptions. The influence of Descartes was profound and, through the Enlightenment philosophy of the eighteenth century, his ideas pervaded all fields of inquiry, implying that there is one single legitimate linear procedure for the acquisition of knowledge (Thomas, 2004: 55f.). Method prevailed and theory drifted far away from its original meaning.

The word *theory*, along with its many forms in modern languages, is derived from the Ancient Greek *theoria*. The term has a number of meanings, some of them preserved in its modern usage: "a set of statements or principles devised to explain a group of facts or phenomena; the branch of a science or art consisting of its explanatory statements, accepted

principles, and methods of analysis, *as opposed to practice*" (American Heritage Dictionary, 2009, emphasis added). And yet, in its original usage, the Greek *theoria* encompassed a particular practice, that of traveling to foreign lands and observing the customs of other peoples, in order to achieve wisdom (*sophia*) necessary to improve the home state (Euben, 2006: 21). The travelers were appointed from among the most experienced and esteemed citizens. Their voyage was not to exceed a 10-year period and, upon their return to the homeland, the knowledge they acquired was examined, as well as the possible consequences of the experience of travel upon the *theoroi*. If their observations proved to have enhanced their wisdom, they were celebrated and honored in life and death. On the other hand, if a traveler – *theoros* – returned from his voyage 'corrupt,' he was to be isolated from his country-mates and even punished (Plato, *Laws XII*, 951 a–c; see also: Euben, 2006: 21; Redfield, 2002: 25; Leontis, 2001).

So, knowing about aliens was not a light affair among the Ancient Greeks – it was carefully planned, evaluated, and sanctioned. Indeed, it never is perceiving similarities among members and differences with outsiders is the basic principle of forming a group and laying out its ground rules, a process inevitably linked to some form of authority (cf Levinas, 1989; Todorov, 1989). With the advent of the age of modernity, the task of discerning and legitimizing the limits of human groups of all kinds was assigned to academic institutions and scholars in the fields of social sciences and humanities (Thomas, 2004). In this new distribution of authorities, archaeology is positioned as the field of knowledge in charge of the deep past of humankind, reaching back to periods of illiteracy and therefore inaccessible to historians. For lack of written evidence, the explanations of the customs of our ancestors were to be sought in another direction, that of the peoples still living in far off lands, unaffected by the progress of the European civilization. The notes of travelers to Africa or the Americas, describing the strange ways of life of the Others who lived there, shook the world at the times of Descartes and thus became a source for the understanding of our own remote past (Johnson, 1999: 133, 134; Wylie, 2002: 137).

General Law of Development

This initial alliance between distance places and European pasts foreshadowed several points crucial for the future development of theory in archaeology. First of all, archaeology's subject matter has been defined as a culture remote in time, and yet comparable to the ones remote in space. At the same time, this comparison between the Other and the past implied that there is a general framework according to which these comparisons are made, encompassing all possible forms of human organization, ordered in a logical sequence, leading from simpler, less developed, to more complex forms. This general pattern of development was introduced to archaeological thinking from other fields of research: it has a long history in social sciences and humanities, starting again with the Enlightenment of the middle of the eighteenth century and is still present in its refined, but essentially unchanged forms, under the title of 'social evolution' (Pluciennik, 2005). This concept implies that human societies are expected to move along a more or less predictable trajectory of development, with individual groups being 'caught' by observation in one particular 'stage' at a particular time. The remote past and the remote lands have thus been brought into a permanent metaphorical association: the state in which the European researchers have encountered their non-European informants during the ethnographic observations has been removed from modernity, pushed into the past, wherefrom 'we' have long progressed, and this progress is testified by the archaeological record of the European prehistory (Fabian, 1983; Johnson, 1999: 132–137; Gosden, 1999, 2004, 2007).

This starting premise, like most of the other theoretical assumptions of the late nineteenth and early twentieth century archaeology, has remained largely implicit, leading subsequent researchers to emphasize the essentially descriptive character of the culture-historical approach – the set of rules and methods typical of archaeological research in this period (Johnson, 1999: 15, *passim*; Olsen, 2002: 30–33). Nevertheless, the descriptions produced on the grounds of these implied but never clearly stated assumptions have pervaded archaeological interpretations of the past, underlying the cultures uniquely successful on the trajectory of development and emphasizing their value as role models for the present. In the process, the most desirable ancestors were singled out, such as the great civilizations of Greece and Rome (Babić, 2010). Archaeology, along with the other fields of research, was fulfilling its social purpose: the past was uncovered, described, compared to other accessible samples of human behavior, valorized, and offered to the public in the form that enabled the audience to make sense of its present condition. However, this practice left out huge tracts of the past that still needed meaningful links to the present, and arrowheads and sherds still floated over the spaces now inhabited by the modern nations without the proper names for the people who made and used them. Archaeologists were in dire need for the principle to organize these traces of the past into comprehensive units. The solution is best described in the words of one of the most influential figures in the history of the discipline, Vere Gordon Childe:

> We find certain types of remains – pots, implements, ornaments, burial rites, and house forms – constantly recurring together. Such a complex of associated traits we shall term a 'cultural group' or just a 'culture.' We assume that such a complex is the material expression of what today would be called a 'people.'
>
> Childe, 1929: v, vi, emphasis added

It may safely be argued that this statement discloses the fundamental theoretical principle underlying archaeological research over the long period from the beginning of the twentieth century into the present. The logic behind it is the premise that the objects made and used by humans are the direct expressions of cultural norms – rules according to which people behave (Johnson, 1999: 16, 17). These rules are then equated to the modern experience of belonging to a nation state, with its fixed boundaries and authorities unifying social, political, and economic behavior of subjects throughout a contiguous territory (Thomas, 2004: 96–118). Furthermore, the assemblages of material culture identified over certain territories are again sorted into a chronological sequence,

from technologically and therefore culturally less developed ones to the more advanced forms, all ultimately leading to the present (Jones, 1997; Lucy, 2005; Thomas, 2004: 111). The notorious case of Gustaf Kossinna offers a drastic illustration of the ultimate consequences of this line of inference: by the beginning of the twentieth century he established a direct relationship between material culture, the people producing it, and the language they spoke, enabling the researchers to equate sharply delineated archaeological cultures with the areas occupied by clearly recognizable peoples or tribes. This idea lays heavily upon the works of the late nineteenth-century German anthropologists, such as Friedrich Ratzel, Leo Frobenius, and Fritz Graebner, who established the idea of *Kulturkreise*, or homogeneous, organically integrated cultural complexes, as a part of their *Kulturhistorische Methode*, devised to ascertain historical sequences on the basis of the contemporary geographical distribution of culture complexes (Jones, 1998: 46; Lucy, 2005: 87; Olsen, 2002: 117–121). By projecting the areas of modern ethnic or national units onto the past, Kossinna established cultural and therefore ethnic continuity, opening up the path for the justification of the control over territories based on ancestral rights (Jones, 1997: 16–17; Lucy, 2005: 87; Thomas, 2004: 110–111).

The possibility to harness the archaeologically recovered past for the present political purposes has not remained unexplored, and from the days of Kossinna and his overt involvement with the official German claims over parts of Poland (Thomas, 2004: 110), to the much more recent turmoils in many parts of the world, archaeological interpretations of the past have been harnessed for present-day purposes, arguing that cultural groups of the past are in fact embryos of the modern nation states (cf Kohl and Fawcett, 1995). The danger of slipping into the realm of abuses was already apparent to Childe, and he rejected Kossinna's racist assumptions and stated especially in his later writings that an archaeological picture of culture was built up out of many fragments, registered at various times in different places (Lucy, 2005: 88; Olsen, 2002: 37). However, the basic premises according to which archaeological cultures have been discerned remained largely the same and, along with the procedures for organizing material, the theoretical premises of Ratzel and Frobenius upon which these procedures were designed percolated to archaeology too, bringing along current ideas about human groups in general. In this fashion, cultural groups were invested with the characteristics of the modern nation states and the archaeological record interpreted along the lines of the eminently modern experience of group identity, one based upon ethnic affiliations organized into large political entities. Once more archaeologists, Childe and Kossinna alike, partook in the prevailing intellectual climate of the age, received ideas from other fields of research, and adapted them for the purposes of their own quest for the past (cf Thomas, 2004: 94ff.; Olsen, 2002: 33–37). The abuses of archaeological inferences drawn in such a manner were apparent even to Childe, who amended his views and already in the mid-1930s started experimenting with other theoretical premises for archaeological thinking, devoting considerable attention to social and economic aspects of past societies, informed by Marxism (Olsen, 2002: 142–143; Patterson, 2003: 42–62). Nevertheless, the concept of cultural group proved to be extremely durable and has marked almost a century of archaeological research. It survived almost intact the first explicitly theoretical shift in archaeology, unfolding during the 1960s under the name of the 'New Archaeology,' and later renamed 'processual archaeology,' a perspective that expressly aimed at overcoming the fallacies of the culture-historical approach.

Explicitly Scientific Archaeology

The proponents of this new outlook ardently argued that the discipline needed a much more explicit theoretical foundation than the one previously practiced in the culture-historical framework. They purported that its descriptive character did not offer answers beyond classification of material culture of the past (Johnson, 1999: 12f.; Olsen, 2002: 43–56; Wylie, 2002: 23 f.). In order to overcome this petrified state of the discipline Lewis Binford, the main protagonist of the critique of the 'traditional approach,' claimed that archaeology is and indeed must be anthropology, both in terms of disciplinary goals and the wider frame of reference (Binford, 1962). This proposition was strengthened by the institutional divisions of North American universities, where archaeologists have been members of anthropology departments and are basically trained under the same banner. Binford himself was a student of Leslie White, the anthropologist whose tenets soon formed one of the building blocks of New Archaeology (Olsen, 2002: 44, 45, 48, 49). White's functionalist view of culture as an organism with mutually dependent parts or subsystems closely related to the natural environment heavily influenced the newly formulated ideas about the research of the past (Johnson, 1999: 64f.; Sabloff, 2005: 214). Binford and his associates argued that the technological, economic, social, political, and ideological aspects of past societies, inextricably intertwined with the environment they inhabited, can be approached through the material culture and other physical traits that past activities left behind.

This explicit agenda opened up the issue of the practical procedures enabling archaeologists to meet the goals of translating the material traces, now tangible, but static, into the dynamics of past processes or, in other words, to bridge the gap between the archaeological record and past behaviors. This new device came to be known as middle-range theory (Gibbon, 1989; Johnson, 1999: 49–63; Olsen, 2002: 50–51; Sabloff, 2005; Wylie, 2002: 17). The most productive way to link behaviors with their material consequences were found in ethnoarchaeology and experimental archaeology – controlled observations by archaeologists of living populations' activities and their material consequences under particular environmental and cultural situations. The data gathered in this way are then compared to archaeological record, following the premise that the behavior observed by researchers of modern populations may be analogous to ones in the past leaving the same material trace. Binford himself studied the North American Nunamiut population, with the goal of explaining the archaeological record of the prehistoric hunter-gatherers (Binford, 1983). The comparisons are then guided by the mode of explanation predominant in hard

sciences: problem-oriented hypothesis testing, following the hypothetico-deductive model of confirmation (Gibbon, 1989; Wylie, 2002: 2–4). However, this procedure was merely making explicit and methodologically rigorous what archaeologists have been doing from the late nineteenth century – comparing the Other in space to the Other in time.

The insistence on experiment and observation still apparent in the program of processual archaeology testifies to the commitment to a rigorously empirical form of inquiry. This approach echoes the foundations of modern science laid by Descartes and may be traced back to Carl Hempel, philosopher of science who strongly and consistently argued for logical positivism, rooted in the empiricist conviction that legitimate knowledge is properly grounded in testable observables (Gibbon, 1989: 10f.; Wylie, 2002: 2–4). This approach inevitably led to a strong emphasis on various scientific methods applied in archaeological research, such as petrography or palaeobotanics, and accelerated the development of disciplinary branches such as zooarchaeology. The texts produced by processual archaeologists did change substantially in their form and content, with all kinds of diagrams and quantified observations taking center stage. At the same time, equally quantified means came to be used to describe and explain the actions of humans in their environment, producing models of behavior. Site catchment analysis (Bailey, 2005) and blueprints of exchange systems (e.g., Renfrew, 1975) were followed by check-lists that would ensure the identification of certain forms of social organization in the archaeological record (e.g., Peebles and Kus, 1977). However, these forms were sought on a universal ladder of social evolution (e.g., Friedman and Rowlands, 1977).

A focus on universal social forms should not come as a surprise, since the principal aim of New Archaeology has been to formulate lawlike conclusions about the past and bring them in accordance to scientific knowledge in general. As a consequence, the discipline reestablished the concept present in its theoretical toolkit from the very beginning – that human culture progresses through universal and predictable stages of development. Admittedly, the iteration induced by processual archaeology differs from the previous situation in the discipline, since now the concept is clearly articulated and explicitly stated. Nevertheless, the burden of its inception still remains, and substantial efforts invested in its refinement over the last decades have not yet fully resolved the in-built threat of valorizing human groups according to their 'stage of development' (cf Pluciennik, 2005). Even more importantly, the adherence of New Archaeology to the concept of social evolution highlights the fact that at a time when archaeologists explicitly and programmatically turned to social anthropology for theoretical support in the project of making the discipline more explicitly scientific, anthropologists themselves were slowly but steadily abandoning the evolutionary and functionalist approaches proposed by Leslie White. Simultaneously, in other fields of the humanities and social sciences, philosophy as well as history, the positivistic ideal of scientific objectivity was severely challenged (Olsen, 2002: 55). Instead of bringing the discipline closer to other fields of research and endowing it with a sound theoretical foundation, it may be argued that the movements during the 1960s and 1970s were in fact leaving intact some of the crucial points present in archaeology from its inception as an academic discipline. This may be obvious in the fact that the mighty tide of change left aside the concept of cultural group without challenging it and even sporadically used it as a practical principle for organizing material before subjecting it to the scientific analysis.

Troubles with Science

The first critical reactions to the program of New Archaeology were articulated almost simultaneously to its publication. Particularly important, although perhaps too harsh, are the objections judging that the program offered was predominantly a methodological shift, rather than theoretical, leaving the conceptual core of the discipline almost intact (Wylie, 2002: 25–41). The modernist notions about the scientifically valid procedures in line with the writings of Descartes present though not always explicit in the history of the discipline, in the case of the processual program are indeed deliberately emphasized. On the other hand, even the authors acknowledging the theoretical dimension of the processual program point to the limitations of the logical positivistic agenda in general and in particular to its applicability to archeological research (Gibbon, 1989; Wylie, 2002: 17–19). The critique may start with the position that science itself is a social process, in which theory is dependent on observations mediated by methodological rules. These rules, in turn, direct what kind of observations will be considered valid, generating a circular line of inference, where methods based on assumptions about the nature of the object of study produce observations that must confirm the starting assumptions. Here, the general objection to logical positivistic program, that all sciences need not share the same methodological and substantive base, is extended to emphasize that humans possess consciousness, volition, and reflectivity, setting them apart from natural phenomena. The mechanistic conception of human beings and their actions, imminent in positivistic research agenda, are therefore seen as particularly damaging to disciplines such as archaeology (Gibbon, 1989: 35–60). The valorization of the decades' long implementation of the principles of processual archaeology based upon logical positivism seems to prove these claims, since very few formal lawlike statements or theories were ever produced, but a series of hypotheses testing in need of further systematic restatements (Gibbon, 1989: 94).

In retrospect, many claim that "an empiricist account of science must fail to provide laws of social life equivalent in scope, certainty and predictive capacity to those offered by natural science," so "New Archaeologists launched a program that was impossible to carry to completion" (Gibbon, 1989: 117). There can be little doubt, however, that the processual claims for a more methodologically rigorous and problem-oriented archaeology have induced significant changes in the discipline and many archaeologists around the globe still test its interpretive potential. On the other hand, the universal consensus about the theoretical grounds upon which archaeologists should seek knowledge about the cultures different from our own and remote in past – about the Other

in time – has not been reached and whether Lewis Binford put in motion a revolutionary paradigm shift in theoretical terms remains an open question (cf Olsen, 2002: 54–56, *passim*).

The proponents of the next reconsideration of the aims and means of the discipline, emerging by the beginning of 1980s, picked up many of the already mentioned weak points of the processual stratagem. The term to denote these diverging critiques, postprocessual archaeology, announces the dissatisfaction with the processual achievements, but at the same time seeks allegiance with all forms of thinking loosely encompassed with the label postmodern. Following the cue from other humanities, although with a delay of several decades, archaeologists started questioning the suitability of the positivistic imperative for their own research purposes (cf Hodder, 1986; Shanks and Tilley, 1987). Some archaeologists started arguing that the deep embeddedness of research into its contemporary social, cultural, and political circumstances cannot be overcome by pursuing the rigorously scientific procedures, themselves also subject to various pressures and conditionalities of the context of inquiry (Bourdieu, 1988). Rather than working hard at closing out these 'disturbances,' the attempts have been made to contextualize the discipline in its modern setting (Hall, 2000; Thomas, 2004). The issue of authority, apparent already in the Ancient Greek practice of theory, came to the forefront of discussions on the social role of archaeologists (Shanks and Tilley, 1987; Tilley, 1989). The path has been opened to reconsider the conceptual burden of the discipline built into its modern foundations (Thomas, 2004), enabling the researchers to realize that our reconstructions of the past have been heavily influenced by our own preconceptions of ourselves and of the modern Other. These preconceptions range from the ones concerning not only the cultures different from ours (Gosden, 2004, 2007) but also the differences among the individuals in our own society, such as gender (Gero and Conkey, 1991; Gilchrist, 1999; Sørensen, 1999), age, status, and religious or ethnic affiliations (Díaz Andreu et al., 2005; contributions to Insoll, 2007). The change of focus brought into the archaeological research the ideas of possible ways of articulating individual identity other than the ones constructed during the modern age (Fowler, 2004; Thomas, 2004: 119f.). Even the experiences of time and space, previously thought of as universal to all human beings and therefore susceptible to objective observation and plotting onto archaeological charts, have become the subject to investigation (Johnson, 2007; Lucas, 2005).

This turn requires a reconsideration of the material we work with – the archaeological record. While culture-historical and processual approaches equally consider the record to be comparable to a contemporary material context, yet remote in time, the postprocessually inclined authors tend to emphasize its fundamentally historical nature (Lucas, 2012: 8; Patrik, 1985). This basic distinction of the ways in which archaeologists approach their subject matter may be translated into the two models of archaeological record: the physical and the textual, each with a corresponding theoretical background and methodological implications (see also Jones, 2002: 10, 11, 17). As opposed to the concept of (past) material culture as a static physical representation of human actions, susceptible to objective observation, classification, and quantification in a manner similar to natural phenomena, the textual model of archaeological record presupposes that material culture is meaningfully constituted, active, and used by people to convey meanings. The rules according to which these meanings and messages are structured and understood are culturally specific, and the task of archaeologists is therefore to 'read' (Hodder, 1986) or interpret them (Thomas, 2000), rather than explain them in terms similar to hard science.

In order to meet this proposition of Ian Hodder and others, archaeologists had to depart from the previous inspirations in the anthropological literature, such as social evolutionism and functionalism, and seek allegiance with the theoretical reconsiderations transcending disciplinary boundaries between philosophy, sociology, and literary theory among others. The hermeneutics of Paul Ricoeur and the poststructuralism of Michel Foucault brought into archaeological thinking the issues of subjectivity, discourse, and intertextuality (Bapty and Yates, 1990; Tilley, 1990). Furthermore, the concepts developed by Pierre Bourdieu (1977) and Anthony Giddens (1979) were picked up by archaeologists (e.g., Dobres and Robb, 2000; Johnson, 1989; Thomas, 2000), introducing a theory of practice into the discipline. Humans, both present and past, are seen as endowed with *agency*, or the capacity to act individually and to make choices, that may be opposed to structural constraints imposed by *habitus* – a set of values and dispositions acquired under the particular cultural circumstances. Human groups therefore construct collective knowledge with shared meanings ascribed to the world and have in-built potential to change. The change, however, need not be easily predictable according to preconceived rules. Rather than search for universal regularities in social processes, the discipline shows interest for particular experiences of individuals.

Archaeology and Its Contemporary Context

In the course of these reconsiderations of the discipline, its own practices have become the object of research in their own right. It has become apparent that underlying theoretical assumptions influence the methodological aspects of archaeological work down to the very procedures of data collection in the field (Lucas, 2001). At the same time, the valuable legacy of the processual strain of research – the application of hard science analyses – is not neglected, but this remains an uneasy cohabitation, due to the differing logic behind the positivist and constructivist approaches (Jones, 2002). In spite of the efforts made in this direction, the frequent objection that the postprocessual approach represents 'theory without method' (Lucas, 2012: 2) still looms large. It may be argued that the departure from the initial theoretical foundations of the discipline brought about in the 1980s induced changes more profound than the processual shift, but stripped off its methodological certainty. It is therefore not surprising that a significant number of archaeologists around the world still remain largely uninterested in the theoretical debate and adhere to the methodological practice produced as a result of hybridization of culture-historical and processual approaches. Furthermore, various locally specific conditions of the institutionalization of archaeology have brought an array of particular responses to the general pattern of 'the three stages,' similar to the one proposed for the European prehistory by early researchers such as Christian Thomsen (Babić, 2014;

Biehl et al., 2002; Gramsch and Sommer, 2011; Sklenář, 1983), sometimes resulting in diametrically opposing outcomes from the same theoretical starting point. The archaeological involvement with Marxism may serve as a highly illustrative example (cf McGuire, 1992; Patterson, 2003). Matthew Johnson therefore rightfully concludes that "the claims for a distinctive way of thinking about the world in theoretical terms specific to archaeology, to which most or even the largest group of archaeologists would willingly or knowingly subscribe, are over-stated" (Johnson, 2006: 117).

See also: Archaeology and Cultural Memory; Archaeology, Politics of; Classical Archaeology; Cognitive Archaeology; Colonialism, Anthropology of; Ethnic Identity and Ethnicity in Archaeology; Ethnicity: Anthropological Aspects; Ethnoarchaeology; Europe, Archaeology of; Identity in Anthropology; Marxist Archaeology; Modernity: Anthropological Aspects; Postcoloniality; Structuralism.

Bibliography

American Heritage Dictionary, 2009. American Heritage Dictionary of the English Language, fourth ed. Houghton-Mifflin, New York.

Babić, S., 2010. Prošlost kao Drugi – Drugi kao prošlost (The past as the other – the other as the past). Etnoantropološki problemi 5 (2), 259–268.

Babić, S., 2014. Identity, integration and power relations in the study of the iron age: implications from Serbia. In: Popa, C.N., Stoddard, S. (Eds.), Fingerprinting the Iron Age: Approaches to Identity in the European Iron Age: Integrating South-Eastern Europe into the Debate. Oxbow, Oxford, pp. 295–302.

Bailey, G., 2005. Site catchment analysis. In: Renfrew, C., Bahn, P. (Eds.), Archaeology. The Key Concepts. Routledge, London; New York, pp. 230–235.

Bapty, I., Yates, T. (Eds.), 1990. Archaeology after Structuralism. Post-Structuralism and the Practice of Archaeology. Routledge, London.

Biehl, P., Gramsch, A., Marciniak, A. (Eds.), 2002. Archaölogien Europas: Geschihte, Methoden und Theorien/Archaeologies of Europe: History, Methods and Theories. Tubinger Archaologische Taschenbucher, band 3.

Binford, L.R., 1962. Archaeology as anthropology. American Antiquity 28, 217–225.

Binford, L.R., 1983. In Pursuit of the Past. Decoding the Archaeological Record. Thames and Hudson, London.

Bourdieu, P., 1977. Outline of a Theory of Practice. Cambridge University Press, Cambridge.

Bourdieu, P., 1988. Homo Academicus. Polity Press, Cambridge.

Childe, V.G., 1929. The Danube in Prehistory. Oxford University Press, Oxford.

Díaz Andreu, M., Lucy, S.J., Babić, S., Edwards, D.N., 2005. The Archaeology of Identity, Approaches to Gender, Age, Status, Ethnicity and Religion. Routledge, London; New York.

Dobres, M.-A., Robb, J. (Eds.), 2000. Agency in Archaeology. Routledge, London; New York.

Euben, R.L., 2006. Journeys to the Other Shore: Muslim and Western Travelers in Search of Knowledge. Princeton University Press, Princeton.

Fabian, J., 1983. Time and the Other: How Anthropology Makes its Object. Columbia University Press, New York.

Fowler, C., 2004. The Archaeology of Personhood. Routledge, London; New York.

Friedman, J., Rowlands, M.J. (Eds.), 1977. The Evolution of Social Systems. Duckworth, London.

Gero, J., Conkey, M.W. (Eds.), 1991. Engendering Archaeology: Women and Prehistory. Blackwell, Oxford.

Gibbon, G., 1989. Explanation in Archaeology. Blackwell, Oxford.

Giddens, A., 1979. Central Problems in Social Theory: Action, Structure and Contradiction in Social Analysis. University of California Press, Berkeley.

Gilchrist, R., 1999. Gender and Archaeology: Contesting the Past. Routledge, London; New York.

Gosden, C., 1999. Anthropology and Archaeology. Routledge, London; New York.

Gosden, C., 2004. Archaeology and Colonialism. Cultural Contacts from 5000 BC to the Present. Cambridge University Press, Cambridge.

Gosden, C., 2007. The past and foreign countries: colonial and post-colonial archaeology and anthropology. In: Meskell, L., Preucel, R.W. (Eds.), A Companion to Social Archaeology. Blackwell, Oxford, pp. 161–178.

Gramsch, A., Sommer, U. (Eds.), 2011. A History of Central European Achaeology. Theory, Methods, and Politics. Archaeolingua, Budapest.

Hall, M., 2000. Archaeology and the Modern World. Routledge, London; New York.

Hodder, I., 1986. Reading the Past. Cambridge University Press, Cambridge.

Hodder, I., 1991a. Archaeological theory in contemporary European societies: the emergence of competing traditions. In: Hodder, I. (Ed.), Archaeological Theory in Europe – The Last Three Decades. Routledge, London, pp. 1–24.

Hodder, I. (Ed.), 1991b. Archaeological Theory in Europe – The Last Three Decades. Routledge, London.

Hodder, I., Preucel, R. (Eds.), 1996. Contemporary Archaeology in Theory. Blackwell, Oxford.

Insoll, T. (Ed.), 2007. The Archaeology of Identities. A Reader. Routledge, London; New York.

Jones, A., 2002. Archaeological Theory and Scientific Practice. Cambridge University Press.

Jones, S., 1997. The Archaeology of Ethnicity – Constructing Identities in the Past and Present. Routledge, London; New York.

Johnson, M., 1989. Conceptions of agency in archaeological interpretation. Journal of Anthropological Archaeology 8, 189–211.

Johnson, M., 1999. Archaeological Theory. Blackwell, Oxford.

Johnson, M., 2006. On the nature of theoretical archaeology and archaeological theory. Archaeological Dialogues 13 (2), 117–132.

Johnson, M., 2007. Ideas of Landscape. Blacwell, Oxford.

Kohl, P., Fawcett, C. (Eds.), 1995. Nationalism, Politics, and the Practice of Archaeology. Cambridge University Press, Cambridge.

Leontis, A., 2001. Mediterranean theoria: a view from Delphi. Thesis Eleven 67 (1), 101–117.

Levinas, E., 1989. In: Hand, S. (Ed.), A Levinas Reader. Blackwell, Oxford.

Lucas, G., 2001. Critical Approaches to Fieldwork – Contemporary and Historical Archaeological Practice. Routledge, London.

Lucas, G., 2005. The Archaeology of Time. Routledge, London; New York.

Lucas, G., 2012. Understanding the Archaeological Record. Cambridge University Press.

Lucy, S., 2005. Ethnic and cultural identities. In: Díaz Andreu, M., Lucy, S.J., Babić, S., Edwards, D.N. (Eds.), The Archaeology of Identity, Approaches to Gender, Age, Status, Ethnicity and Religion. Routledge, London; New York, pp. 86–109.

McGuire, R., 1992. A Marxist Archaeology. New York University Press.

Olsen, B., 2002. Od predmeta do teksta. Teorijske perspektive arheološkog istraživanja. Geopoetika, Beograd (translation of the Norwegian original: Fra ting til tekst. Teoretiske perspektiv i arkeologisk forksning. Univesitetsforlaget, Oslo, 1997).

Patrik, L.E., 1985. Is there an archaeological record? Advances in Archaeological Method and Theory 8, 27–57.

Patterson, T.C., 2003. Marx's Ghost. Conversations with Archaeologists. Berg, Oxford; New York.

Peebles, C.S., Kus, S.M., 1977. Some archaeological correlates of ranked societies. American Antiquity 42 (3), 421–448.

Pluciennik, M., 2005. Social Evolution. Duckworth, London.

Redfield, J., 2002. Herodotus the tourist. In: Harrison, T. (Ed.), Greeks and Barbarians. Edinburgh University Press, Edinburgh, pp. 24–49.

Renfrew, C., 1975. Trade as action at a distance; questions of integration and communication. In: Sabloff, J.A., Lamberg-Karlovsky, C.C. (Eds.), Ancient Civilization and Trade. University of New Mexico Press, Albuquerque.

Renfrew, C., Bahn, P. (Eds.), 2005. Archaeology. The Key Concepts. Routledge, London; New York.

Sabloff, J., 2005. Processual archaeology. In: Renfrew, C., Bahn, P. (Eds.), Archaeology. The Key Concepts. Routledge, London; New York, pp. 212–219.

Shanks, M., Tilley, C., 1987. Reconstruction Archaeology. Cambridge University Press, Cambridge.

Sklenář, K., 1983. Archaeology in Central Europe: The First 500 Years. Leicester University Press.

Sørensen, M.L.S., 1999. Gender Archaeology. Polity Press, Cambridge.

Thomas, J., 2004. Archaeology and Modernity. Routledge, London.

Thomas, J. (Ed.), 2000. Interpretive Archaeology – A Reader. Leicester University Press, London; New York.

Tilley, C., 1989. Archaeology as socio-political action in the present. In: Pinsky, V., Wylie, A. (Eds.), Critical Traditions in Contemporary Archaeology. Cambridge University Press, Cambridge, pp. 104–116.

Tilley, C. (Ed.), 1990. Reading Material Culture. Structuralism, Hermeneutics and Post-Structuralism. Blackwell, Oxford.

Todorov, Tz, 1989. Nous et les autres. La réflexion française sur la diversité humaine. Editions du Seuil, Paris.

Ucko, P.J. (Ed.), 1995. Theory in Archaeology – A World Perspective. Routledge, London.

Wylie, A., 2002. Thinking from Things. Essays in the Philosophy of Archaeology. University of California Press, Berkeley.

Archaeometry

Alan Watchman, James Cook University, Townsville, QLD, Australia

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 1, pp. 626–629, © 2001, Elsevier Ltd.

Abstract

Archaeometry is the application of the various sciences in archaeology to determine age, composition, structure, origin, production technique, resource exploitation, and environment. This field of research includes radiocarbon, luminescence, and other numerical dating methods, geochemical and mineralogical analyses, metallurgical studies, and stable isotope chemistry. The scope of archaeometry is expanding and diversifying as new physical and chemical analytical techniques are employed to provide archaeologists with answers to questions about natural and human artifacts, site characteristics, organic remains, and past environments.

Archaeometry

'Archaeometry' is a specialized discipline within archaeology in which various scientific methods of chemical and physical analysis are applied to archaeologically derived materials. Archaeometry therefore centers on research whose aim is to explain and test archaeological questions about ancient things or phenomena related to human cultural activities. The research measures or quantifies parameters using analytical techniques borrowed from earth science, chemical, biological, and other scientific disciplines. The field of archaeometry includes, for example, determining the ages of sites and artifacts, sourcing objects to the original localities of raw materials, identifying the components and processes involved in converting earth materials into metals and ceramics, and determining patterns of dietary exploitation. Many analytical techniques have been applied in these investigations and some common ones are briefly described below to illustrate the diversity of archaeometric techniques.

Age Determinations

Probably the first example of the use of archaeometrical methods was the realization that the annual growth rings of trees could be used to determine the age of construction of prehistoric pit houses in the southwest of the USA using cores taken from wooden beams (Douglass, 1936). The tree rings were also shown to indicate variations in climate within the time span of the life of the tree (Judd, 1954). This method of tree-ring dating, or dendrochronology, is still widely used to date habitation sites in the Americas and in Europe.

The dating of carbon-bearing substances associated with archaeological deposits has revolutionized the determination of absolute ages in archaeology. Radiocarbon dating (Libby et al., 1949) has been used to place time markers on important periods of human activities and climatic change, and to date the extinctions of animals, for example the woolly mammoth and sabre-toothed tiger (Ho et al., 1969). Refinements to the method and development of the accelerator mass spectrometer (AMS) have allowed for the dating of microgram quantities of carbon (Nelson et al., 1977). For example, a famous case involves the AMS radiocarbon dating of the Shroud of Turin (Damon et al., 1989). While many people believe the shroud was wrapped around Christ's body, fragments of its linen fibers were dated independently by three radiocarbon laboratories to the late thirteenth century.

Archaeologists have also been able to date individual seed grains, beeswax resins (Nelson et al., 1995), charcoal scrapings from Paleolithic cave paintings (Valladas et al., 1992), and buried Australian rock art (Watchman, 1993). Pigment painted on rock in South Africa was the first application of the AMS radiocarbon technique for dating rock art (Van der Merwe et al., 1987). Innovative methods have since been developed for extracting carbon from paintings. Oxidation of carbon compounds associated with pigments or rock surface mineral deposits using either an oxygen plasma (Russ et al., 1990), laser energy (Watchman et al., 1993), or permanganate chemistry (Gillespie, 1997) can be used to prepare samples prior to AMS radiocarbon dating. Rock art in Texas has been dated to between 4100 to 3200 years ago (Russ et al., 1990). Rock paintings in northern Australia have been dated using carbon in rock surface encrusted 'canvasses' and mineral coatings (Watchman, 1993) and in plant fibers used as paint binders (Watchman and Cole, 1993).

Analytical methods employing light (luminescence) are also being used to date sediments in occupation shelters and in pottery. The basis of the luminescence methods is that natural radiation provides energy to the electronic structure of some crystals, particularly quartz and feldspar minerals (Aitken, 1985). Grains that are heated (thermoluminescence, TL) or illuminated by green light (optical luminescence) emit small amounts of light that reflect the level of radiation and length of time since they were incorporated in a sediment or pot. In a controversial case the floor sediments containing stone artifacts at Jinmium in northern Australia were dated at 116 000 years old using TL and quartz grains (Fullagar et al., 1997). The age of those sediments were disputed by proponents of an optical stimulated luminescence method (OSL) who found that the maximum age of the deposits was only 10 000 years (Roberts et al., 1998). Arguments about the reliability of the age determinations for the sediments are concerned essentially with bulk-samples versus singlegrain analyses, incomplete bleaching by sunlight of the luminescence centers in quartz grains generated by *in situ* disintegration of

rock fragments compared with well-bleached sand grains. These controversial situations and measurements highlight the complex nature of many archaeometric techniques, and indicate the potential problematical results obtained from applications where the experimental conditions are not well known.

Identifying the Sources of Artifacts

Determining the characteristics of raw materials obtained from quarries and recognizing the same attributes in archaeological objects allows an archaeologist to describe trading routes and interactions between groups of people. The simplest archaeological technique is to look at earth materials using a microscope. Such petrological analyses can be used, for example, to identify the rock types selected for production of prehistoric polished edge-ground axes. The texture and assemblage of minerals in stone tools can be compared with similar features in rocks at known quarry sites. This method was used to indicate complex prehistoric trading patterns in southeastern Australia. Hard, fine-grained hornfels from the Mt. William quarry was traded across hundreds of kilometers (McBryde, 1984). Similarly, Neolithic communication networks and boundaries were defined in England based on the petrological analysis of edge-ground hand axes (Cummins, 1980).

Subatomic particles can also be used to provide geochemical information about rocks and minerals to substantiate petrological and stylistic information about artifacts. For example, in the proton induced X-ray emission (PIXE) analysis of artifacts, high-energy protons are used to induce X-ray excitations from a range of elements. The measurement of multiple trace element abundances in artifacts, waste materials, and known sources can differentiate between quarries and can allocate artifacts to individual deposits. An example of this archaeometric method is the characterization of obsidian artifacts and sources for the investigation of the production, trade, and patterns of consumption of that natural resource in Melanesia. The systematic analysis of obsidian deposits and excavated flakes on New Britain have demonstrated that two exposures and possibly a third site supplied most of the obsidian for use as stone tools for more than 11 000 years (Summerhayes et al., 1993). Another analytical technique employing subatomic particles is neutron activation. Energetic neutrons, like the protons in PIXE, allow the measurement of elemental abundances in artifacts. Neutron activation analysis (NAA) has been used to source Late Neolithic and Early Bronze Age obsidian in Macedonia (Kilikoglou et al., 1996). Sources more than 300 km north and south of the archaeological site at Mandalo were shown to have provided raw obsidian for use as artifacts.

Neutron activation analysis has also been used to characterize steatite (soapstone) sources in eastern North America (Truncer et al., 1998). Relatively small quantities of rock can be analyzed using the NAA method to produce gamma rays from many elements, which are measured over four weeks. A wide range of trace and major elements, as well as rare earth elements can be measured after a single long irradiation in a flux of neutrons from a reactor. Statistical analysis of the large numbers of analytical measurements eases the burden of discriminating between quarry materials, and permits the identification or 'fingerprinting' of elements that are characteristic of each quarry.

The chemical and mineralogical characteristics of Australia ochers have been determined to discriminate between various sources and to confirm ethnographic accounts of long-distance trade. PIXE analyses have been used to characterize ocher sources in central Australia, with implications for defining trading networks and, delineating boundaries between Aboriginal populations (Smith et al., 1998). X-ray fluorescence has also provided major and trace element analyses, and Rietveld X-ray diffraction has identified the principal mineral phases for each known ocher source in southern Australia (Jercher et al., 1998). Ochers smeared on bones and objects have been traced to potential sources within geologically defined areas. The high degree of natural variability in ocher compositions makes such sourcing studies extremely challenging, so much so that it may be easier to exclude certain locations rather than identify specific sources.

Chemical analyses of other materials can also be used by the archaeometrist to provide useful information for the archaeologist. For instance, the minor ingredients in glass, especially the presence of lead, can be used to indicate the likely sources of glass production and the existence of trading networks. The ecclesiastical glass found at Koroinen, Finland illustrates how the tools of archaeometry provide insights into medieval manufacturing and trading processes (Kuisma-Kursula and Raisanen, 1999). Using X-rays generated from focused electrons in a scanning microscope and a beam of protons directed onto small glass fragments, the average chemical composition of the Finnish glasses were found to be remarkably similar to northwestern European glasses. The abundances of minor elements, particularly lead, sodium, and calcium, in medieval Finnish glasses are inconsistent with likely Russian sources, but match German and southern European glasses. The indication is that glass-making was not practiced in Finland at that time, but that supplies of colored glass for Finnish monasteries depended on trade links with Western and Central Europe.

Metallurgy

Unlike other geochemical sourcing studies that rely on finding significant archaeological variations in the intrinsic trace element compositions between different sources, the study of metal processing is far more challenging for an archaeometrist. Concentrations of trace elements vary between the ores and the processed metal, and complications arise because of the introduction of fluxes and refractory components. Lead isotope analyses, on the other hand, provide an alternative means for sourcing alloyed or leaded artifacts because the relative proportions of isotopes from ores to artifacts are not measurably affected by chemical and pyrometallurgical processes (Srinivasan, 1999). For example, thermal ionization mass spectrometry produces lead isotope ratios ($Pb^{208/207}$ and $Pb^{207/206}$) which were used to discriminate between known deposits of Indian lead ores and artifacts made from them. Matching artifacts to specific ore deposits using the isotopic ratios resolved contentious chronological issues based on art-historical criteria. Western Indian lead ores rather than local sources were found to have been used for brass in northern

India during the premedieval period, and also in lead and brass in southern India during the medieval period. Recycling of materials was also observed in Indian coins where lead isotope analyses of later bronzes fitted the trends established for earlier groupings.

Ceramics

Geochemical studies of pottery and porcelain have focused mainly on grouping complete objects or sherds into products that were either made locally or imported (*see* Ceramics in Archaeology). This has been done to establish or confirm suspected cultural links and trading associations between groups of people. Basic analyses of pottery sherds usually include determining the mineralogical compositions of the clay and temper to find out what specific materials were used and where a pot was made. This is done using petrographic and scanning electron microscopy. More detailed geochemical analyses may include the use of inductively coupled plasma emission spectroscopy, NAA, PIXE, or gamma-ray emissions. An example of one of these techniques is the confirmation of the extent of local production of fineware between the seventh and second centuries BC at a village on the Calabrian coast, Italy (Mirti et al., 1995). Often, distinctive patterns of elemental abundances are not readily evident in the large amount of geochemical data that is collected. Multivariate statistics are needed to separate disparate groups of pots, as in the case of Roman Samian pottery (Argyropoulos, 1995).

Paleodiets

Adaptation by people to changing environments and the transition in prehistoric societies from hunting and gathering to agricultural subsistence are topical themes in archaeology. These changes can be better understood through the measurements of stable carbon ($*^{13}C$) and nitrogen ($*^{15}N$) isotopes obtained from skeletal materials, and edible plants and animals (DeNiro and Epstein, 1978, 1981). Stable carbon isotopes were first used to study the introduction of maize into northeastern North America (Vogel and Van der Merwe, 1977). Their use is based on the observation that different groups of plants differ markedly in their isotopic compositions, and that therefore animals that live mainly on particular plants will have bones of matching composition. For example, studies of the paleodiets of people occupying southern Ontario, Canada (Katzenberg et al., 1995) and coastal New England, USA (Little and Schoeninger, 1995) have shown reliance on animal proteins rather than on maize or legumes during the Late Woodland period.

These examples of the applications of various analytical techniques reveal how archaeometry plays an important role in archaeology. Archaeometric methods are not used in isolation, but complement a range of other archaeological observations that build a more complete picture of the past. Archaeometry therefore provides a variety of tools that allow archaeologists to understand better human relics of the past.

See also: Bioarchaeology; Ceramics in Archaeology; Environmental Archaeology; Geoarchaeology.

Bibliography

Aitken, M.J., 1985. Thermoluminescence Dating. Academic Press, London.

Argyropoulos, A., 1995. A characterization of the compositional variations of Roman Samian pottery manufactured at the Lezoux production center. Archaeometry 37, 271–285.

Cummins, W.A., 1980. Stone axes as a guide to Neolithic communications and boundaries in England and Wales. Proceedings of the Prehistoric Society 46, 45–60.

Damon, P.E., Donahue, D.J., Gore, B.H., Hatheway, A.L., Jull, A.J.T., Linick, T.W., Sercel, P.J., Toolin, L.J., Bronk, C.R., Hall, E.T., Hedges, R.E.M., Housley, R., Law, I.A., Perry, C., Bonani, G., Trumbore, S., Woelfli, W., Ambers, J.C., Bowman, S.G.E., Leese, M.N., Tite, M.S., 1989. Radiocarbon dating of the Shroud of Turin. Nature 337, 611–615.

DeNiro, M.J., Epstein, S., 1978. Influence of diet on the distribution of carbon isotopes in animals. Geochimica et Cosmochimica Acta 42, 495–506.

DeNiro, M.J., Epstein, S., 1981. Influence of diet on the distribution of nitrogen isotopes in animals. Geochimica et Cosmochimica Acta 45, 341–351.

Douglass, A.E., 1936. The Central Pueblo Chronology. Tree Ring Bulletin 2, 29–34.

Fullagar, R.L.K., Price, D.M., Head, L.M., 1996. Early human occupation of northern Australia: Archaeology and thermoluminescence dating of Jinmium rock-shelter, Northern Territory. Antiquity 70, 751–773.

Gillespie, R., 1997. On human blood, rock art and calcium oxalate: Further studies on organic carbon content and radiocarbon age of materials relating to Australian rock art. Antiquity 71, 430–437.

Ho, J.Y., Marcus, L.F., Berger, B., 1969. Radiocarbon dating of petroleum-impregnated bone from tar pits at Rancho La Brea, California. Science 164, 1051–1052.

Huntley, D.J., Godfrey-Smith, D.I., Thewalt, M.L.W., 1985. Optical dating of sediments. Nature 313, 105–107.

Jercher, M., Pring, A., Jones, P.G., Raven, M.D., 1998. Rietveld X-ray diffraction and X-ray fluorescence analysis of Australian Aboriginal ochres. Archaeometry 40, 383–401.

Judd, N.M., 1954. The Material Culture of Pueblo Bonito. Smithsonian Miscellaneous Collections 124, Washington, DC.

Katzenberg, M.A., Scwarcz, H.P., Knyf, M., Melbye, F.J., 1995. Stable isotope evidence for maize horticulture and paleodiet in southern Ontario, Canada. American Antiquity 60, 335–350.

Kilikoglou, V., Bassiaskos, Y., Grimanis, A.P., Souvatzis, K., Pilali-Papasteriou, A., Papanthimou-Papaefthimou, A., 1996. Carpathian obsidian in Macedonia, Greece. Journal of Archaeological Science 23, 343–349.

Kuisma-Kursula, P., Raisanen, J., 1999. Scanning electron microscopy-energy dispersive spectrometry and proton induced X-ray emission analyses of medieval glass from Koroinen (Finland). Archaeometry 41, 71–79.

Libby, W.F., Anderson, E.C., Arnold, J.R., 1949. Age determination by radiocarbon content: world-wide assay of natural radiocarbon. Science 109, 949–952.

Little, E.A., Schoeninger, M.J., 1995. The Late Woodland diet on Nantucket Island and the problem of maize in coastal New England. American Antiquity 60, 351–368.

McBryde, I.M., 1984. Kulin greenstone quarries: The social contexts of production and distribution for the Mt William site. World Archaeology 16, 267–285.

Mirti, P., Casoli, A., Barra Bagnasco, M., Preacco Ancona, M.C., 1995. Fine ware from Locri Epizephiri: A provenance study by inductively coupled plasma emission spectroscopy. Archaeometry 37, 41–51.

Nelson, D.E., Korteling, R.G., Stott, W.R., 1977. Carbon-14: Direct detection at natural concentration. Science 198, 507–508.

Nelson, D.E., Chaloupka, G., Chippindale, C., Alderson, M.S., Southon, J.R., 1995. Radiocarbon dates for beeswax figures in the prehistoric rock art of northern Australia. Archaeometry 37, 151–156.

Roberts, R., Bird, M., Olley, J., Galbraith, R., Lawson, E., Laslett, G., Yoshida, H., Jones, R., Fullagar, R., Jacobsen, G., Hua, Q., 1998. Optical and radiocarbon dating at Jinmium rock shelter in northern Australia. Nature 393, 358–362.

Russ, J., Hyman, M., Shafer, H.J., Rowe, M.W., 1990. Radiocarbon dating of prehistoric rock paintings by selective oxidation of organic carbon. Nature 348, 710–711.

Smith, M.A., Fankhauser, B., Jercher, M., 1998. The changing provenance of red ochre at Puritjarra rock shelter, central Australia: Late Pleistocene to Present. Proceedings of the Prehistoric Society 64, 275–292.

Srinivasan, S., 1999. Lead isotope and trace element analysis in the study of over a hundred south Indian metal coins. Archaeometry 41, 91–116.

Summerhayes, G.R., Gosden, C., Fullagar, R., Specht, J., Torrence, R., Bird, J.R., Shahgholi, N., Katsaros, A., 1993. West New Britain obsidian: Production and consumption patterns. In: Fankhauser, B.L., Bird, J.R. (Eds.), Archaeometry: Current Australasian Research. Department of Prehistory Research School of Pacific Studies, Australian National University, Canberra, pp. 57–68.

Truncer, J., Glascock, M.D., Neff, H., 1998. Steatite source characterization in eastern North America: New results using instrumental neutron activation analysis. Archaeometry 40, 23–44.

Valladas, H., Cachier, H., Maurice, P., Bernaldo de Quiros, F., Clottes, J., Cabrera Valdes, V., Uzquiano, P., Arnold, M., 1992. Direct radiocarbon dates for prehistoric paintings at the Altamira, El Castillo and Niaux caves. Nature 357, 68–70.

Van der Merwe, N.J., Seely, J., Yates, R., 1987. First accelerator carbon-14 date for pigment from a rock painting. South African Journal of Science 83, 56–57.

Vogel, J.C., Van der Merwe, N.J., 1977. Isotopic evidence for early maize cultivation in New York State. American Antiquity 42, 238–242.

Watchman, A.L., 1993. Evidence of a 25,000-year-old pictograph in Northern Australia. Geoarchaeology 8, 465–473.

Watchman, A., Cole, N., 1993. Accelerator radiocarbon dating of plant-fibre binders in rock paintings from northeastern Australia. Antiquity 67, 355–358.

Watchman, A.L., Lessard, A.R., Jull, A.J.T., Toolin, L.J., Blake Jr., W., 1993. Dating of laser-oxidized organics. Radiocarbon 35, 331–333.

Archival Methods

John Markoff, University of Pittsburgh, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The modern public archive dates from the era of the French Revolution. 'Archival methods' may denote the work of the profession of archivists in acquiring documents, classifying them, preserving them, providing them to users, and developing finding aids to assist those users. It also may refer to the work of historians with archival materials, the use of which has been a distinguishing feature of history as a discipline. Beginning in the late-twentieth century, new electronic information technologies were raising new questions about the future of archival methods in both senses of the term.

"[N]o documents, no history," declared Charles-Victor Langlois and Charles Seignobos (1898: p. 2) at the end of the nineteenth century as they begin their guide to historical research. They continued: "To hunt for and to gather the documents is therefore a part, logically the first part, and one of the most important parts, of the historian's craft" (1898: p. 2). They thus stated a methodological principle as well as a defining practice for a social science discipline. They offered a standard for progress: If more recent historians are in any way superior to those of the past it is because they have "the means to be better informed" (1898: p. 3). These means are the modern public archive, with its large and effectively cataloged collection, and the careful training of fledgling historians in the critical use of documents. In the early twenty-first century, historians still were far more likely than other social scientists to exchange tales of archival discomforts and delights and to guide their graduate students in crafting essays built around properly referenced primary sources. But new information technologies were blurring the distinction between archival and published sources, between primary and secondary sources, and between the methods of history and the methods of other social sciences.

Archives as Research Sites

Archives are the records of an organization's activities. The term is used both for the place those records are housed and for the body of records themselves. These records may have been produced by that organization or gathered by it. Such organizations include businesses, religious bodies, and government agencies, and we may stretch the notion to include the records of individual persons. Since the records were produced for the purposes of that person or organization, they tend to be organized for those purposes and not those of researchers. This means that archives are often difficult and sometimes impossible to use effectively without detailed knowledge of the organization in question. Researchers could hardly make much of the documents in some government archives without understanding the agencies that produced or collected the documents.

Archives, therefore, tend to be organized idiosyncratically. One archive may well differ from another not only in its classificatory categories but also in the extensiveness of inventories, catalogs, and other aids to research. There may be no list of individual documents, but just indications of more or less broad categories. The records of some country's regional or local governments may not be organized in a uniform manner. As a result of such challenges to a user's ingenuity and patience, a researcher may derive considerable pleasure from the sense of mastering an archive. By contrast, a researcher can get quite far in most libraries the world over by expecting to find an alphabetical catalog of authors and titles.

The Modern Public Archive

The linking of the modern practice of historical research and the modern public archive generally is dated from the French Revolution, but archives of some form are considerably older. Because states often have needed records of laws, tax assessments, and interstate treaties, state archives existed in antiquity. *Arkheia* is the Greek for an authority's records. It is probable that the invention of writing was spurred by the state need for documentation. So archives were part of the technology of state power, but they also could play an important role in the power claims of other social actors. In the European middle ages, lords and ecclesiastical institutions maintained records spelling out their privileges (etymologically: private laws). The possession of documents attesting to immunities from royal claims or the antiquity of claims over others (e.g., for seigneurial dues) was an important instrumentality of such privilege.

Early modern European states had their document collections, too, but these often were semiprivate, with individual agencies and even individual officials having their own archives. Steps toward centrally organized collections were part of the history of European statemaking; early centralizers included the Spanish archives of Charles V in 1545, the English State Paper Office in 1578, and the Archivum Secretum Vaticanum in 1611 (Favier, 1958: p. 24). But these were not accessible publicly, secrecy being as significant an attribute of the monarch's documentary collection as of the lord's.

The French Revolution launched the modern public archive. In France's villages, peasant attacks on the documentary collections of lords and monasteries testified to their continuing significance. Although the new revolutionary authorities promoted some archival destruction, the overwhelming thrust of their efforts ran in a very different direction. The separate collections of royal agencies, old regime corporate

institutions, and local powerholders were made public in two senses:

1. They were brought under state management and thereby opened up to systematic classification, not only in the National Archives housed in Paris, but in the uniformly organized departmental archives as well. This development greatly facilitated measures for preservation as well as cataloging. To organize the vast body of materials, some seized from lords or ecclesiastical bodies, the revolutionary state sought those with relevant experience, including those who had advised the lords or former Benedictines. With time, a modern profession of archivists formed around these tasks. France's first school for training this new corps, the Ecole des Chartes, was founded in 1821.
2. They were made accessible to the public. With exceptions for state secrets (a far more limited notion than before the Revolution) and the privacy of individual citizens whose identities might figure in state records, documents were to be made available for the scrutiny, and research, of French citizens. Archival documents were redefined from being primarily instruments of power, often private power, to aspects of the national cultural heritage. National archives took their place alongside national museums and national libraries as repositories of and as emblematic of a national heritage of which a respectable national government would claim stewardship.

In establishing France as the center of a reconstructed Europe, Napoleon had vast archival holdings transferred from Spain, Vienna, and much of Italy to Paris. In the wake of the French defeat, the recovery of other states' documents from France further strengthened the notion that a respectable modern state needed its own respectable modern archives. During the nineteenth century, and beyond, this increasingly meant that, like France's, other archives were going to be open to a far broader public than in the past (Favier, 1958).

Professional History

The historical profession that emerged in the nineteenth century combined a distinctive sense of evidence, a distinctive form of professional apprenticeship, and a distinctively organized intellectual product. This distinction was being developed in many places, but the model for the new professional history was provided in the writings and seminars of German historian Leopold von Ranke, beginning in the late 1820s (Novick, 1988).

1. Historical claims were supposed to be grounded in the best possible primary sources, generally those located in searching relevant archives. Historical knowledge would be disciplined, a notion that acquired important additional baggage as the German *wissenschaft* was transmuted into the English *science* as a description of the kind of learning produced by historical research.
2. Apprenticeship included practice in making something of primary sources in graduate seminars. To the extent that the seminar was seen as a sort of 'laboratory' (Novick, 1988: p. 33), the sense of history as a sort of science was augmented. Mastery would be further demonstrated by the completion of an original work dependent for its primary sources on unpublished, generally archival, materials. The budding historian would acquire not only the cognitive skills but also the character needed for archival labors. Historians needed to be "calm, reserved, circumspect; in the midst of the torrent of contemporary life which swirls about him, he is never in a rush." Those "always in a hurry to get to the end of something may manage to find honorable employment in other careers" (Langlois and Seignobos, 1898: p. 103).
3. Professionally respectable intellectual products consisted of books and articles whose factual claims, in principle, were verifiable from the primary documents to which the text would explicitly refer. The footnote to sources was a vital part of the prose of professional history (Grafton, 1997). Denoting the structure of reference – footnotes plus bibliography – as the scholarly 'apparatus' fostered rhetorical associations with science.

Important and diverse consequences followed from its characteristic methods for such a professionalized history:

1. *National specialization*. Historical specialties tended to be organized thematically around national histories because the archives were organized by states. Beyond linguistic familiarity, practitioners who had invested their energies in mastering one archive or set of archives found it advantageous to continue future work that would reap the benefit of those investments.
2. *Building national identities*. Partly as a consequence of this technology of history, and partly because states were supportive of such work, the writing of history became an important component of the very forging of national identities in the nineteenth and twentieth centuries.
3. *Focus on elites*. Because what was held in these state archives, by definition, were documents that had been of interest to state managers, professionalized history had a tendency to privilege the doings of states and other human practices of interest to administrators while paying lesser attention to other ways of exploring human experience.
4. *Geographically uneven attention*. Because the standards of professional history could best be attained where there were rich archival collections, cared for by skilled practitioners of preservation and cataloging, those parts of the world relatively well endowed in those regards were attractive to researchers who hoped to achieve professional advance. Such conditions meant governments (1) with rich documentary collections, (2) with the resources to support the preservation and cataloging of these collections, and (3) for which the French Revolutionary model of a public national archive as an essential component of a modern national identity had some resonance. Nowhere were these conditions better met than in France. Part of what had made France such an important center of innovation in professional history in the twentieth century was its marvelously organized archives. Countries (1) lacking extensive documentation, (2) having documents but lacking the resources or the interest in preserving them or organizing them coherently, or (3) with strong traditions of archival secrecy were all, in their different ways, more difficult terrain for professional historians.

Bureaucratized institutions other than states maintain archives as well, such as firms, universities, labor unions, social movement organizations, and hospitals, and these institutions have established their own practices of collection and preservation and their own rules of access for curious researchers.

5. *Marginalization*. The inverse of the previous points: social activities, social strata, and entire continents that either have been of little interest to producers of documents or have not developed such extensive and usable documentary collections have tended to be marginal areas for professional history. Historians interested in such subject matters have had to be methodologically innovative and sometimes have faced an enormous challenge for professional recognition.
6. *Eschewing explicit methodological discussion*. A professional socialization oriented to the careful scrutiny of the provenance of particular documents and the opportunities for knowledge and the motivations for falsification on the part of that document's author encouraged historians to delight in the particular. This was reinforced by the particularities of archives, knowledge of whose quirks was not readily transferable to other archives. Historians understood their own methods in large part as the location of useful texts in idiosyncratically organized archives in conjunction with the careful criticism of documentary sources. Such skills were acquired by experience and the patient application of a few readily assimilable critical rules of thumb. The capstone of training involved immersion in some idiosyncratic archive. Thus, as Langlois and Seignobos noted, "[t]o learn to distinguish in this enormous confused literature of printed inventories (to confine ourselves to such), that which merits confidence from that which does not, in a word to be able to make use of them, is a complete apprenticeship" (1898: p. 12). Historians therefore have directed far less of their intellectual efforts at methodological discussion and also have had fewer methods courses as part of graduate training than the other social sciences. In 1971, one study of graduate education commented that "methodology is the orphan of the history curriculum" (Landes and Tilly, 1971: p. 82); there is no reason to think that decades later things had changed significantly.
7. *Eschewing explicit theory construction*. Dependence on the fortunes of document survival, as Murray Murphey observed (1973: p. 148), is a major reason historians have been less prone than other social scientists to pursue a model of inquiry in which abstract hypotheses are formulated and then tested against available evidence. The unpredictabilities of archival research suggest instead the development of hypotheses in tandem with the exploration of data or in modern sociological lingo, the development of *grounded theory*. One often finds wholly unexpected documents sitting on a shelf. Theory building as an explicit and valued agenda therefore is less characteristic of people in departments of history than it is of those in sociology or political science and very much less so than in economics. Nor does a broad knowledge of history's big themes without a deep reserve of pertinent facts earn much esteem from professional peers. By way of contrast, graduate students in top US economics programs do not regard a deep knowledge of the economy as nearly as important to their professional futures as they do skill in mathematics (Klamer and Colander, 1990). One could not imagine graduate students in first-rate departments of history similarly downplaying the significance of a rich and concrete knowledge of some specific time and place in favor of mastering methodological or theoretical tools.

Working with Documents

The notion of 'archival methods' comprehends two rather different sorts of activity. First, there are the methods of collection, preservation, and cataloging on which documentary research in archives depends (Schellenberg, 1956; Brooks, 1969). Second, there is the work of historians (or others) with those documents. Essential to nineteenth-century notions of historical method was the 'criticism of sources.' Classic works of method, like that of Langlois and Seignobos (1898) provided a broad taxonomy of ways in which one document might be superior to another. For example, historians became skilled at considering the following:

1. possible errors of transcription in the reproduction of or quotation from ancient texts;
2. whether a document's author was in a position to reliably know that which was claimed;
3. what motivations that author might have for slanting a story this way and that;
4. how to select from among multiple documents the most credible account.

According to Langlois and Seignobos (1898: p. 131), historians need to cultivate an attitude they call 'methodical distrust': an author's 'every claim' must be suspected of being 'mendacious or mistaken' (p. 132). There follows a taxonomy of possible guises assumed by documentary deceit and error as a guide for avoiding the snares and pitfalls that might lead a researcher too easily to accept some statement as historical truth.

More recent reflection has pondered the forms of distortion that might exist in whole collections of documents, and recognizes along with Murray Murphey (1973: p. 146) that the survival of documents often has depended on fires, floods, and 'the concern of loving daughters.' If there is still much to be said for the formulation of Langlois and Seignobos that there is no history without documents, we may consider how it is that documents come to be available to the historian's scrutiny. We should consider (Shapiro et al., 1987) the following:

1. *Recording*. What social processes bring documents into existence? This includes taking into consideration what sorts of things are or are not of interest to states, economic enterprises, religious bodies, and other recorders of words. An increase in counted murders may indicate a greater interest of states in whether or not people are killing each other, for example. No doubt in many times and places social struggles in and near centers of government generate far more paper than those far away.
2. *Preservation*. What social processes destroy or preserve documents? We must consider not only the destructiveness

of war and fire, which often are random destroyers of texts and the wishes of loving (or hostile) relatives for preservation (or elimination), but also the existence of organized agencies for preserving or eliminating the records of the past. The existence of well-trained and dedicated archival professionals is more probable in some countries than others, more (we may speculate) in large urban centers, and more in relation to some subject matters (but which?) than others.

The study of the professional cultures of archivists would seem an important agenda for a self-critical history. Archivists are not simply neutral agents of an imperative to preserve and arrange. They play important roles in deciding what is to be preserved and how it is to be organized. Like other professionals, archivists develop a sense of collective identity, professional organizations, journals, codes of ethics, conferences, and disagreements. In the United States, for example, the Society of American Archivists organized itself separately from the American Historical Association in 1936 and its journal, *American Archivist*, was launched 2 years later. Archivists study the optimal conditions for the preservation of paper, including the challenges of the different papers in use at different times and places, and consider the potential erosion of moisture, light, and various temperatures. Archival culture tends to strongly favor keeping collections together, if they are to be kept at all, and in the original order, because that order may be an important clue to proper interpretation, but sometimes such principles have been violated in the interests of ease of use. Electronic records are a significant challenge to previously established practice (Schellenberg, 1956; Daniels and Welch, 1984; Bearman, 1991).

When economists, sociologists, or political scientists turn to archival sources to construct statistical series about social processes that extend over time and space, they are likely to encounter several problems. Some of these problems may be amenable to amelioration by statistical manipulation but others demand the kinds of analysis of sources that are characteristic of historians – and still others call for both in conjunction.

1. *Missing data*. Random document loss from moisture, fire, and the gnawing criticism of mice; deliberate destruction of troublesome records by those who might be embarrassed by some datum or data; and errors in filing by clerks all generate considerable missing cases in many potentially valuable data series. Statistical judgments about the appropriateness of completing the series by extrapolation and interpolation from surviving data may be complemented by the institutional knowledge that sometimes permits a researcher to find an alternative source for the same data.
2. *Changing definition*. Much data of interest to states is subject to changing definition and therefore many series of great interest to social scientists cannot be used intelligently without a study of those definitions. Crime, ethnicity, conflict, and poverty, for example, are all subject to considerable redefinition, both formally and informally. To make use of any long data series on crime, one would have to have acquired considerable knowledge of the changing interests of states in accurate data collection for different kinds of offenses. One also would have to know something of the changing mores that redefine actions as criminal or not (and which kind of crime they are). In addition, the boundaries of such administrative subdivisions as counties and municipalities often change, and a study of changing administrative geography often is an essential step on the way to some form of statistical correction.

Challenges to Professional Traditions

The last quarter or so of the twentieth century saw emerging challenges to such notions of research methods from several converging sources; these challenges were continuing into the next century. Because what was easy to examine with prevailing methodologies were those facets of human experience that had left traces in the documentary collection of states, the subject matters of professional history gave great weight to adults, men of weapon-bearing and income-earning age, dominant ethnicities, state politics, and dominant powers on the world scene. Those curious about other realms of human experience, including members of some of the underrepresented groups and places, were led to develop new methodologies.

To some extent, these involved innovative use of state archives, as in the mobilization of records of births and deaths to develop the new historical demography, a vital window on the past lives of ordinary people. Because average height in adults closely tracks childhood nutrition, the extensive records of height maintained by militaries has been used for new understandings of historical variation in wealth and poverty. Because wills are a good place to indicate wishes about how one is to be remembered, they have been used to track long-term changes in religiously connected commitments, and how these varied from place to place, over time, and among social groups. To some extent, innovation involved the use of new kinds of sources, as in the discovery of the value of visual materials for clues to the history of childhood or aging. Additionally, innovation involved extensive interaction with other social science disciplines, as in important interchanges with anthropology in the hope of getting a handle on people whose activities have left less easily pursued traces in the written records of states or in the importation of quantitative methods from economics, sociology, and political science (Rabb and Rotberg, 1982; Revel and Hunt, 1995).

Fueled by these trends, and fueling them, some professional historians became inclined to doubt the clarity and coherence of the entire project of a scientific history. This further eroded the rationale for a methodologically distinctive enterprise, thus opening the way for even further cross-disciplinary fertilization. But it also opened the way for a radical skepticism about any form of historical knowledge or inquiry whatsoever, which some dubbed 'postmodernism.' In the last quarter of the twentieth century, increased sensitivity to the ways in which the subject matter and methods of past historical research had been intertwined with social power was leading some to abandon the notion of any coherent historical truth altogether. But others held that such questions only showed that a historian's job was an engagement with more difficulties than covered in the meticulous account of Langlois and Seignobos and that

the deployment of new sorts of evidence on previously unresearched subjects simply made the past more interesting (Appleby et al., 1994).

Challenge and Opportunity of Electronic Information Technology

The new information technology becoming increasingly available at the end of the twentieth century was reshaping the ways in which archives were used and seemed certain to reshape what an archive was. These technologies were raising challenges to established archival methods, both in the sense of how archivists went about their tasks of developing and maintaining archives and in the sense of how historians made use of them.

1. *Cataloging could now be enhanced radically*. Particular items could be cross-classified in innumerable ways because electronic data files are not subject to the same sorts of space limits as a catalog's page or a card file, making it easier for researchers to find all information that is relevant. User-manipulated databases meant that users could sort through files according to the categories important to them, relieving them, to at least some limited extent, of the challenging burden of mastering the intricacies of the classificatory systems of particular archives. Because electronic information could be transmitted around the world at high speed through the Internet, researchers were becoming less dependent on visits to distant archives to have some notion of their contents.
2. *Reproduction of documents was immeasurably enhanced*. Age-old errors of transcription, inherent in hand copying of archival materials, were alleviated by the development of microfilming technologies and further reduced by photocopying. Both microfilming and photocopying technologies enhanced the labors of researchers from afar. But they also meant that archives and research libraries could duplicate their own scarce holdings and distribute them either to individual scholars or other research institutions. The potential of high-quality electronic scanners as input devices to computers was only at the very beginning of use in archival work at the beginning of the twenty-first century, but no one could doubt that the potential impact on research activities was going to be even greater in the future than it already had proven be early in the twenty-first century. Not only could a visiting researcher store large collections of source materials on a portable computer, but also such materials could be transmitted readily to far-off scholars. Digital files, moreover, often were searchable, which meant that one could, for example, look for how frequently certain terms or combinations of terms occurred in specifiable proximity to other terms or combinations of terms in very large numbers of documents. All this was opening up the prospect of electronic archives searched according to multiple criteria, further freeing researchers from the idiosyncrasies of particular catalogs. This also meant that as researchers became experienced with particular document collections, they could revise and refine their search criteria with ease rather then have to accept their own initially flawed classification scheme because they lacked the time to redo it. So a researcher initially did not have to be as erudite as previously before approaching a document collection, and beginning students could practice interpreting some of the same documents with which their teachers had made their reputations.

One could readily imagine digitized versions of the holdings of national libraries, made freely available to all, scholars and citizens, or even extending access beyond national frontiers. Or one could imagine separate research libraries participating in the common project of making their digitized holdings publicly available. Some European states were taking steps in these directions in the early twenty-first century (Darnton, 2010).

3. *Publishing*. New modes of electronic publishing seemed to some to have similarly important implications for the redefinition of professional scholarly standards. Robert Darnton has noted that a conventionally published book, for example, might omit its apparatus of learned footnotes and bibliography, but make them available to other interested scholars on a disk or an Internet site. One might speculate that this would lead to new forms of scholarly precision and richness (as scholars supplemented a basic text with footnoted references to sources, critiques of learned predecessors, methodological commentaries on sources or predecessors, annotated rather than bare-bones bibliographies, and even scanned copies of vital original documents). At the beginning of the twenty-first century, for example, it was technologically feasible for the learned footnote at the bottom of the printed page to be replaced by the electronic note in some online text in the form of a hyperlink to a scanned copy of the referenced text itself. Such superreferences actually would send the primary sources to the reader's screen or printer, rather than tell the reader in which place the source could be consulted (if that reader had the time and the good fortune to win a fellowship and to have a dean kind enough to grant leave). The capacity of historians to check each other's research against the primary sources would be greatly facilitated. Graduate students with no research funds to travel to document collections nonetheless could hone their critical faculties by studying the sources of prominent works without mastering archival idiosyncrasies. One could imagine future professional historians developing an ethical code requiring making electronic sources publicly available. This would raise again the question of whose responsibility it would be to make those materials useable (Darnton, 1999).

Conversely, the new technologies supported new ways of publishing traditionally structured books as well (including 'publication on demand') and books actually were continuing to increase in number (Darnton, 2010). And since the near-zero cost of transmitting electronic files somewhere multiplied the number of people with an opportunity to print them, archived paper, like archived books, did not seem to be decreasing. Archivists seemed much more ready than historians to define one of their central issues as deciding what to preserve.

The Future of Archival Methods

The new technologies were opening up what could be called *the electronic archive* (Shapiro and Markoff, 1998; Cohen and Rosenzweig, 2006). Databases, possibly accompanied by programs to facilitate use, once created, could be distributed readily, and even augmented by new scholars. But the archiving of electronic materials posed new challenges. First of all what was now technologically feasible was bumping into financial and legal obstacles in the form of copyright issues, the costs of digitization, and the hopes of profit-seeking enterprises to gain control of pieces of these revolutionary processes. Although the technology suggested an erosion of the distinction between a research library and an electronic archive, and scholarly journals in electronic form could nearly instantaneously be read around the world, those who controlled those journals were charging universities enormous fees for access, with that access to be restricted to properly certified members of the universities who paid. On the other hand, early in the twenty-first century, a variety of movements were calling for free access to all, not just academic researchers but beyond, and some electronic archives were organized on that basis. But even then, use was implicitly limited to those with access to appropriate hardware and software (Rosenzweig, 2011).

The proliferation of electronic materials was raising anew important issues of appraisal and accession, maintenance, and access, as storage media, software, and hardware platforms evolved, displacing previous technologies. Only a little way into the twenty-first century, the magnetic tapes and floppy disks of the near past seemed antiques and only some information stored on them had been transferred to newer devices.

Possible consequences for the future culture of historical research might include:

1. Blurring the distinction between research on primary and secondary materials as researchers grew comfortable with electronic archives built of many components that might range from scanned primary sources, quantitative data extracted from such sources, quantified data created from such sources by a researcher, and new variables or indexes added by a community of users.
2. Blurring the distinction between archival and published primary sources as information became available in electronic form, regardless of whether it was extracted from a book or found in an archival dossier.
3. Increasing attention to issues of representativeness of large bodies of data (and diminished concern for the traditional criticism of sources?).
4. Increasing research using primary sources by social scientists with the appropriate technical skill who were not employed by departments of history. User-friendly systems could mean that the level of technical skill needed might not be great. Whether this would mean more collaboration across disciplinary boundaries, or just more work using primary sources by people not having understood the institutional contexts of those sources, or both, was not easy to foresee.
5. Increasing use of archival materials by scholars, whether professional historians or not, who are not primarily specialists in the national history of the country from which the data are derived as it becomes easy for someone on the other side of the planet to use an electronic archive without having to master the particularities of traditional repositories of documents.
6. Increasing use of these resources by non academics, too, perhaps blurring somewhat the boundary between professional historians and history buffs. One could imagine as well the use in complex data-driven projects of a collaboration of amateurs and professionals, on the model of the ways in which professional astronomers were directing large numbers of interested volunteers in participating in sky mapping.
7. Increasing accessibility of materials vital to historical problems that cross national frontiers, as teams of scholars assemble relevant electronic archives.
8. Challenging the notion that an archive was a physical space, housing diverse important collections. An electronic archive need not be in a single place at all. One could imagine a 'distributed' archive, maintained by a consortium of physically distant places, all of which could be accessed from a computer screen by a user who might not even need to know where in the world each piece of information being accessed was physically housed.

One could envisage a possible democratization of the doing of history, as established scholars of more varied backgrounds, students at various levels, and nonacademics, too, joined in scrutinizing readily accessible primary materials. But one could also imagine a constriction, in which only those able to purchase computers, pay for subscriptions for commercially available data sets, and with access to one's own institution's expensive journals, could claim to be in the forefront of producing historical knowledge (Rosenzweig, 2011).

In such ways, the boundaries between history and the other social sciences may become less sharp than they have been since the emergence of the social science disciplines. These boundaries also were raising questions both exciting and troubling about how the pervasive yet evanescent structure of electronic materials might be preserved in useful form for the historians of the future, a challenge raised by what Roy Rosenzweig (2011: p. 7) called 'the simultaneous fragility and promiscuity of digital data.'

See also: Data Bases and Statistical Systems: Archives and Historical Databases; Historical Thought and Historiography: Current Trends; History and the Social Sciences; Internet and Culture.

Bibliography

Appleby, J., Hunt, L., Jacob, M., 1994. Telling the Truth about History. Norton, New York and London.

Bearman, D. (Ed.), 1991. Archival Management of Electronic Records. Archives and Museum Informatics Technical Report, no. 13.

Brooks, P.C., 1969. Research in Archives. The Use of Unpublished Primary Sources. University of Chicago Press, Chicago and London.

Cohen, D.J., Rosenzweig, R., 2006. Digital History. A Guide to Gathering, Preserving and Presenting the Past on the Web. University of Pennsylvania Press, Philadelphia.

Daniels, M.F., Welch, T., 1984. A Modern Archives Reader: Basic Readings on Archival Theory and Practice. National Archives and Records Service, Washington, DC.

Darnton, R., 1999. The new age of the book. New York Review of Books 46 (5).
Darnton, R., 2010. The library: three jeremiads. New York Review of Books 57 (20).
Favier, J., 1958. Les Archives. Presses Universitaires de France, Paris.
Grafton, A., 1997. The Footnote: A Curious History. Harvard University Press, Cambridge, MA.
Klamer, A., Colander, D., 1990. The Making of an Economist. Westview Press, Boulder, CO.
Landes, D.S., Tilly, C. (Eds.), 1971. History as Social Science. Prentice-Hall, Englewood Cliffs, NJ.
Langlois, C.V., Seignobos, C., 1898. Introduction aux études historiques. Hachette, Paris.
Murphey, M.G., 1973. Our Knowledge of the Historical Past. Bobbs-Merrill, Indianapolis, IN and New York.
Novick, P., 1988. That Noble Dream. The 'Objectivity Question' and the American Historical Profession. Cambridge University Press, Cambridge, UK.
Rabb, T.K., Rotberg, R.I. (Eds.), 1982. The New History. The 1980s and Beyond. Princeton University Press, Princeton, NJ.
Revel, J., Hunt, L., 1995. Histories. French Constructions of the Past. New Press, New York.
Rosenzweig, Roy, 2011. Clio Wired. The Future of the Past in the Digital Age. Columbia University Press, New York.
Schellenberg, T.R., 1956. Modern Archives. Principles and Techniques. University of Chicago Press, Chicago.
Shapiro, G., Markoff, J., 1998. Revolutionary Demands. A Content Analysis of the Cahiers de Doléances of 1789. Stanford University Press, Stanford, CA.
Shapiro, G., Markoff, J., Baretta, S., 1987. The selective transmission of historical documents: the case of the parish cahiers of 1789. Histoire et Mesure 2, 115–172.

Archiving: Ethical Aspects

Josefina J Card, Sociometrics Corporation, Los Altos Hills, CA, USA
Tamara J Kuhn, ETR, Scotts Valley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Archiving refers to the process of transferring records from the individual or organization who created the material, to a repository for appraising, cataloging, organizing, preserving, and providing access to others. Social and behavioral science research engenders by-products that are worthy of archiving for a variety of future uses, including the furthering of basic and applied research, policy making, and the development and replication of effective intervention programs. Examples of such archive-worthy by-products are research data, their associated data collection instruments (questionnaires, interview protocols), and, more recently, intervention programs shown to be effective by evaluation research. Ethical issues occasionally arise in the archiving process. These issues typically relate to protecting the integrity of the selection process; protecting respondents' confidentiality; censoring potentially controversial or offensive material; the timing of the release of information to an archive; assignment of due credit to both original producer and archivist; the tension between fidelity and usability in the archiving process; and the ownership of the research and development by-products contained in the archive. Approaches to handling and resolving these ethical challenges are discussed.

Archiving By-Products of Social and Behavioral Research

Archiving refers to the process of transferring records from the individual or organization who created the material to a repository for appraising, cataloging, organizing, preserving, and providing access to others (National Archive of Australia, 2013; Pearce-Moses, 2005).

The social and behavioral sciences produce intellectual by-products at various stages of the research process that, if preserved and organized, could further basic and applied research, aid policy making, and facilitate the development and replication of effective social intervention programs. A variety of institutions preserve such materials. They include government archives, academic data archives and libraries, and specialized organizations in both the public and private sectors. Professional organizations of social science archivists and librarians have been formed to further the field.

The National Archives and Records Administration (NARA) is the US federal agency that preserves and ensures access to those official records which have been determined by the Archivist of the United States to have sufficient historical or other value to warrant their continued preservation by the Federal Government; and which have been accepted by the Archivist for deposit in his custody (44 U.S.C. 2901). Information about NARA's electronic records holdings (most of which are data files) can be obtained from the Internet site http://www.archives.gov/.

CESSDA ERIC, while not itself a data archive, is a large consortium for social science data archives across Europe. Established in the 1970s, the organization provides access through their catalog to information and links to nearly 9000 datasets stored at CESSDA archives throughout Europe. Those member organizations serve more than 30 000 social science researchers yearly, together deliver over 70 000 data collections each year, and continue to add more than 1000 data collections annually (http://www.cessda.org/).

The Inter-university Consortium for Political and Social Research (ICPSR) is the largest academic social science data archive. Founded in 1962 at the University of Michigan, ICPSR is a membership-based organization, which provides access to a large archive of computer-based research and instructional data in political science, sociology, demography, economics, history, education, gerontology, and criminal justice. More information about ICPSR and its holdings is available from http://www.icpsr.umich.edu.

The UK Data Archive, established more than 40 years ago, is the UK's largest collection of digital research data in the social sciences and humanities. The archive contains several thousand datasets acquired from the academic, public, and commercial sectors; and some datasets are available through Nesstar so that frequency counts and questions can be viewed online (http://data-archive.ac.uk/).

Norwegian Social Science Data Services, owned by the Ministry of Education and Research, is one of the largest archives of Social Science research data. Data are organized in four main categories: Individual Level Data, Regional Data, Data About Institutions, and Data About the Political System. Many of their data sets can be accessed, analyzed, and visualized with maps, graphs, and tables using Nesstar. In addition to direct on-line access, data can be downloaded and/or exported to preferred format (http://www.nsd.uib.no/nsd/english/index.html).

An example of private sector archives, Sociometrics Corporation was established in 1983. The company's primary mission is the development and dissemination of social science research-based resources for a variety of audiences, including researchers, students, policymakers, practitioners, and community-based organizations. Sociometrics has pioneered in the establishment and operation of topically focused data, instruments, and (since the mid-90s) program archives (http://www.socio.com): (1) *Data Archives*: collections of original machine-readable data from over 500 exemplary studies, many of them longitudinal, on the American family,

teen sexuality and pregnancy, social gerontology, disability, AIDS and STDs, maternal drug abuse, and geographic indicators; (2) *Instrument Archives*: the questionnaires, interview protocols, and other research instruments that were used to collect the data in the data archives; (3) *Program Archives*: collections of program and evaluation materials from several dozen intervention programs that have proven effective in preventing risky behaviors such as unprotected sex and drug use and in treating psychological and behavioral disorders, such as anxiety, depression, and aggression in children. These topically focused archives synthesize research in the field in one place; facilitate further research with the best existing data and accompanying instruments; promote data-based policy-making; and help service providers and practitioners use the insights gained from research (Card et al., 2007, 2006; Card and Kuhn, 2006).

Two professional organizations of social science data archivists and librarians are the Association of Population Libraries and Information Centers (APLIC) and the International Association for Social Science Information Service & Technology (IASSIST). APLIC's membership, consisting of both individuals and organizations, represents some of the oldest population and family planning agencies and institutions in the United States. IASSIST is an international organization dedicated to the issues and concerns of social science data librarians, data archivists, data producers, and data users. This unique professional association assists members in their support of social science research. The APLIC and IASSIST membership lists provide pointers to the various social science data collections housed all over the world (http://www.aplici.org/; http://www.iassistdata.org/).

Ethical Aspects in Archiving

The development of collections such as those contained in the above archives involves a series of decisions with ethical considerations and implications.

Protecting the Integrity of the Selection Process

Given the limited nature of resources allocable to archiving, how should the contents of collection be selected? Some archives sidestep this challenge by merely cataloging and warehousing archival material (e.g., data sets) donated to them by the field. While this procedure undoubtedly results in the lowest per-capita archiving cost, the quality of the resultant archival collection is uncertain at best. A better procedure is to set objective technical and substantive standards for inclusion in the archival collection and then actively recruit material that meets or surpasses such standards. Setting standards for inclusion can be problematic and contentious. Archivists and curators may often disagree on the future research needs in the organization or in the field. There are some general guidelines that can be used to minimize these challenges, such as making sure that the process used to make data selection decisions is transparent and accountable and guided by policy and legal requirments (Whyte and Wilson, 2010). Decisions about inclusion need to be based on a set of objective criteria for assessing the long-term significance of the research data, which should be developed and agreed upon by larger research communities, institutions, or experts in the field. These guidelines are unlikely to be static. Another challenge is that over time it is likely that standards and thinking in the field will change and inclusion criteria will need to be modified. For example, several years ago criteria for inclusion in an archive of effective behavioral interventions may have required that an evaluation of the intervention be conducted and that it should demonstrate continued behavior change at a 3-month follow up time period. However current standards require a 6-month follow up to demonstrate effectiveness and merit inclusion.

The previously described data and program archives at Sociometrics have worked with Scientist Expert Panels in establishing criteria for inclusion in various collections. For the data archives, the selection criteria are scientific merit, substantive utility, and program and policy relevance of the data sets comprising the collection. For the program archives, the selection criteria are documented effectiveness in preventing the particular social problem or disease (e.g., drug use, teen pregnancy, sexually transmitted disease, HIV/AIDS) or in changing these problems' risky-behavior antecedents (e.g., delaying age at first intercourse, increasing the use of contraception and/or an STD-prophylactic at first and every act of sexual intercourse, and abstaining from or reducing the frequency of drug use). Having established objective inclusion criteria, archive staff then work with their respective Scientist Expert Panels to identify and prioritize available data sets and intervention programs for inclusion in the collections. The end result is an archival collection with integrity and credibility.

Protecting Respondents' Confidentiality

Respondent confidentiality is one of the areas of greatest ethical concern, because once data are released for public use it is impossible to monitor use and ensure that confidentially is respected (ICPSR, 2012). Data archives often contain responses to sensitive questions, some of which, for example, ask respondents to admit to illegal, immoral, or 'private' behavior, such as abortion, premarital or extramarital sexual activity, mental illness, alcohol abuse, and drug use. How can the researcher's need to know (the incidence, prevalence, antecedents, and consequences of these social problems) be balanced against respondent's rights to privacy? This ethical consideration is most often addressed by stripping all archival material of information that could be used to identify individual subjects. Direct identifiers, e.g., name, address, social security number, and exact date of birth (often only month and year of birth are included in a public-use database), are typically easily removed. However indirect identifiers, variables that make unique respondents visible when paired with another, such as a participant with a low-incidence disability attending a rural school with unique characteristics, pose a greater challenge.

A problem arises when data holders want to strip the data set of key variables, such as those measuring the sensitive behaviors listed above, prior to placing a data set in a public use archive. This desire is motivated by the fear that such information could be linked to particular respondents by malicious, hardworking sleuths, even without the help of individual identifiers such as name, address, and so forth. Such censorship restricts the range of uses to which the data set can be put by

future researchers and archives typically make an active attempt to find alternate solutions. There are several solutions to problem. Indirect identifiers can be eliminated or modified by restricting the upper range of the variables, by combining or collapsing variables, or by adding random variation or stochastic error to the variable (ICPSR, 2012; Lawrence, 2010). Data donors typically consult with the archive producers to resolve concerns about identifiers. Other solutions focus on restricting use of the data. It is becoming common to require data users to sign a data use agreement that specifies a set of highly controlled conditions for the use of data prior to being granted access. This may be something as simple as signing a confidentially agreement and pledging to use the data only for research purposes.

However for datasets that have highly confidential material, more requirements are put in place. Many agreements require that the researcher provide an abstract of their research questions and explain why access to the restricted use file is required. They also have to outline a plan for safeguarding the data, and in some cases provide evidence of Institutional Review Board approval. Researchers may be given access to the data only for a limited time period, at the end of which they are expected in good faith to destroy them. There are instances when access to the most confidential data may require even more restrictions and users are typically only able to access those materials through a data enclave. This has meant that a researcher has to access those materials in a secure physical location. These enclaves allow access to the original data in a controlled setting and have physical security measures, such as video monitoring, key card entry, and controlled access; also, the computers housing the data are not connected to the Internet. Use is monitored by archive staff who confirm that materials are not removed and that analyses do not contain any confidential information. Virtual enclaves are replacing physical enclaves in some archives. Virtual enclaves are data portals that allow remote but restricted access to the application system and the restricted data. This access is monitored and the systems restrict users from emailing, copying, or moving files outside of the environment.

The original researcher should collaborate with the archive to determine the appropriate level of protection, aiming to allow the widest access to the data while maintaining confidentiality.

Consent for Data Sharing

While data sharing has long been encouraged and even required by research funders and a data sharing plan is often a required part of a research proposal, the original researcher often has not asked for the consent of the participants for sharing the data beyond the original research project. This can present an ethical dilemma for both the data donor and the archive. While requirements for consent vary by country and funder, this issue can be most easily addressed in the research-planning process and in planning for data sharing when developing study materials. For example, the Australian government has established the position that at minimum, consent forms should not preclude data sharing, e.g., by promising to destroy data; and more specifically, the National Statement defines three levels of consent for the future use of data that must be made clear to the participant: (1) Specific: which is limited only to the project; (2) Extended: which gives permission for the use of the data in future projects in the same general area of research; or (3) Unspecified: which allows for the use of data or tissue for any future research (NHMRC, 2007). If consent was not originally given for data sharing then the archive and the developer have to weigh the benefits of distributing the data against the problem of confidentiality and consent, with one likely outcome – the distribution only of data that is fully anonymous.

Special Considerations for Clinical Trial and Biomedical Data

While most social and behavioral science research projects and archives do not include clinical trial data, there are instances in which clinical trial or biomedical data may be included within a social science archive, such as the recently released Reaching for Excellence in Adolescent Care and Health (REACH) dataset (Adolescent Medicine HIV/AIDS Research Network, 2000; Pardini and Wu, 2011) which combines data on medical visits and laboratory results with access to tissue samples and a series of behavioral questions. REACH data were placed in a social science archive by the US National Institutes of Health specifically to expand the reach and analyses of the data to those beyond the medical sciences. The archiving of clinical trial and biomedical data bring additional ethical considerations. A majority of clinical trials are conducted by biopharmaceutical companies or clinical research groups and the data are typically not made accessible to other researchers or the public, and are rarely placed within a public archive. The lack of sharing within the scientific community has reduced the ability of scientists to validate the findings and examine the methodology of these studies. In many cases, (estimated at between 25 and 50% of trials), the results of these clinical trials are not published or made public (Ross et al., 2009, 2012). For those that have been published, in comparing published articles with trial protocols, 50% of efficacy and 65% of harm outcomes per trial have been shown to be incompletely reported and biased toward the reporting of statistically significant findings (Chan et al., 2004). This lack of transparency and availability of clinical trial data have led to calls for stronger data-sharing requirements. In response, in 2013, EFPIA (which represents the pharmaceutical industry operating in Europe) and PhRMA (the Pharmaceutical Research and Manufacturers of America which represents several US biopharmaceutical research and biotechnology companies) committed to Principles for Responsible Clinical Trial Data Sharing (PhRMA and EFPIA, 2013). Under this agreement, member companies commit to enhance data sharing with qualified researchers, share results with the patients who participate in clinical trials, enhance public access to clinical study information, and reaffirm their commitment to publish clinical trial results. However, rather than committing to archiving the data in a public archive it is the companies that determine who is a qualified to access the data, which again may limit the public availability of such data. One final consideration as clinical trial data becomes available to more researchers is the question of what to do with new results that may be found for participants in the study. In order to make the data available to researchers, the data is deidentified. This makes the conveying

of new results to participants nearly impossible, without the cooperation of the original researchers.

Censoring Potentially Controversial or Offensive Material in Collections

Intervention program archives occasionally encounter an analogous censorship-related challenge. For example, several of the effective programs selected for archiving by the Scientist Expert Panel of PASHA (Card et al., 2007) contain sexually explicit material that could be viewed as offensive and inappropriate by some individuals and communities. However, because these prevention programs are targeted at high-risk, already sexually active youth, the material could also be seen as appropriate, even necessary, to drive home relevant points. In addition, these programs, like other PASHA programs, meet the collection's inclusion criterion of demonstrated effectiveness in changing sexual risk-related behavior in at least one subgroup of teens. The decision was made to include the material without alteration or censorship but to publicize and disseminate the collection as an eclectic one, with different schools and communities being encouraged to replicate programs consistent with their own values, norms, and target populations. An accompanying database of program abstracts was developed so that both the approach and the content of the program packages could be perused, prior to requesting the program from the archive.

Timing of Release of Information to an Archive

Holders of data sets and developers of effective programs often, and understandably, want to reap some payoff from their professional investments by keeping the data or programs to themselves until they have published what they wish to from the data (or have tweaked the intervention program in several implementation cycles to their satisfaction). The ethical issue arises when this 'private' or 'proprietary' period of time stretches to what the field would view as abnormally long. This is especially true when the data were collected, or the intervention program developed, with government funds. Several US federal agencies are trying to forestall the problem by building resource-sharing ground rules into the original funding award document. For example, the National Science Foundation states that researchers with quantitative data should be prepared to place their data, in fully cleaned and documented form, in a data archive or library within a year after the expiration of an award (NSF, 2013). In contrast, the US National Institutes of Health is not as specific and instead expects those receiving large awards to share their data in a timely fashion with 'timely release and sharing,' defined as no later than the acceptance for publication of the main findings from the final data set. Ideally this solution gives the original developer the fair 'head start' his or her efforts have earned, while ensuring that the data collected with government funds will be shared with the field before it gets stale. However two primary issues remain; first, unfortunately the data sharing promises or pledges included in the funded grant proposal are typically not enforced and secondly, in many instances the timely release is based on the discretion of the researcher and many are reluctant to release the data, sometimes for more than a decade, until the data are outdated.

Assignment of Due Credit to Both Original Producer and Archivist

Data sets and program materials are typically received by an archive in a format that data developers and their colleagues find workable but one not yet suitable for public use. The archivist contributes significant additional value in preparing the database for public use. For example, with the approval of the data donor, inconsistencies in the database are eliminated, or at least documented. The documentation is augmented, both at the study level (describing study goals, sampling, and data collection procedures) and at the variable level (assigning names and labels for each variable; documenting algorithms for constructed scale variables). Occasionally, the variable and scale documentation is done using the syntax of a popular statistical analysis package such as SPSS (Statistical Package for the Social Sciences) or SAS (Statistical Analysis System), facilitating future data analysis. Archivists who prepare intervention program packages for public use make analogous contributions. Program materials are edited and 'prettified' for public use. User's Guides, Facilitator's Manuals, checklists, and handouts are created so that the package is implementation-ready in the absence of the original developer. In short, the archiving process is best viewed and executed as collaboration between original developer and archivist. Care must be taken to give due credit for the final product to all individuals, teams, and institutions involved.

Tension between Fidelity and Usability in the Archiving Process

The collaborative model is productive not only for assignment of due credit but also for joint resolution of the fidelity versus usability issues that occasionally arise during the archiving process. Others will be using the materials and have the opportunity to note errors in data or methodology, which leads to several ethical questions. Should obvious errors in the data base be corrected or only documented? Should resulting concerns about methodology be highlighted for those who acquire a dataset or program? Should original program materials that were found effective in the developmental site be altered when replication sites find them unclear or when the curriculum they present is based on out-of-date data? Issues such as these are best resolved on a case-by-case basis by the archivist and original developer, working side by side in collaborative fashion.

Ownership of the By-Products of Research and Development

The purposes and procedures of the archive accepting a donation should be made clear to the donor at the outset. It should be communicated to data donors that the research by-product they are donating toward is being put in a public archive whose main goal is the preservation of the resource.

Some archives also actively publicize and disseminate their holdings. In addition, as seen above, archives vary in the extent to which they work with the donor in 'upgrading' the material for public use. The donor should be informed in advance of what to expect along these lines. Issues of credit and ownership should also be agreed to before archiving work begins. How will professional credit for the collaborative product be allocated? Will the resultant product be sold to the end user (at cost or for profit) or given away? If the product will be sold for profit, will royalties be given to the original developer? If the product will be sold at cost, free, or discounted copies be made available to the original developer?

Conclusion and Looking to the Future

Social science research yields many by-products that, if properly archived, can be used to further future research, aid policymaking, and foster the development and replication of effective prevention and treatment programs. Several challenges, some with ethical implications, arise in the archiving process as it exists today. Many are resolvable with good will, a commitment to upholding the standards established by those in the research and archiving communities, and commitment to the public good on the part of both original developers and archivists. However, the recent discussions around data transparency of research findings and publications based on unreleased clinical data and social science data collected by companies, such as Google and Facebook, are likely to lead to new discussions around the ethics of archiving and the concept of archived data as being publically available. As we move into the age of 'big data' and collections of social networking data owned by corporations, discussions around confidentiality and privacy will probably evolve. The burden will fall to the community of scientists, archivists, and funding agencies to ensure that data sharing and data archiving for public use continues and is conducted ethically.

See also: Data Bases and Statistical Systems: Archives and Historical Databases; Privacy of Individuals in Social Research: Confidentiality; Surveys and Polling: Ethical Aspects.

Bibliography

Adolescent Medicine HIV/AIDS Research Network, 2000. Reaching for Excellence in Adolescent Care and Health (REACH): 1996–2000 [Computer File]. SUNY Health Sciences Center (Producer), Brooklyn, NY. Sociometrics Corporation (Producer & Distributor), Los Altos, CA.

Card, J.J., Lessard, L., Benner, T., 2007. PASHA: facilitating the replication and use of effective adolescent pregnancy and STI/HIV prevention programs. Journal of Adolescent Health 40 (3), 275.e1–e14.

Card, J.J., Kuhn, T., Wells, T., 2006. User-centered design and innovation in the sociometrics social science electronic data library (SSEDL). IASSIST Quarterly 30 (2), 12–17.

Card, J.J., Kuhn, T., 2006. Development of online suites of social science-based resources for health researchers and practitioners. Social Science Computer Review 24 (2), 255–261.

Chan, A.W., Hrobjartsson, A., Haahr, M.T., Gotzsche, P.C., Altman, D.G., 2004. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. Journal of the American Medical Association 291, 2457–2465.

Inter-university Consortium for Political and Social Research (ICPSR), 2012. Guide to Social Science Data Preparation and Archiving: Best Practice throughout the Data Life Cycle, fifth ed. Ann Arbor, MI.

Lawrence, D., 2010. Analysis of Public Use Microdata Files: A Researcher's Perspective. National Statistical Service. http://www.nss.gov.au/nss/home.nsf/NSS/58CC18F435E0800CCA2579E20006BA4B?opendocument – Analysis of public use microdata files.

National Archives of Australia, 2013. Glossary. Retrieved 08.01.13. http://www.naa.gov.au/records-management/publications/glossary.aspx.

NHMRC, 2007. National Statement on Ethical Conduct in Human Research. National Health and Medical Research Council, Australian Research Council, Australian Vice-Chancellors' Committee. Australian Government.

National Institutes of Health, 2012. NIH Grants Policy Statement. http://grants.nih.gov/grants/policy/nihgps_2012/nihgps_ch8.htm.

National Science Foundation, 2013. Directorate for Social, Behavioral and Economic Sciences, "Data Archiving Policy".

Pardini, B., Wu, J., 2011. Reaching for Excellence in Adolescent Care and Health (REACH): 1996–2000: A User's Guide to the Machine-readable Files and Documentation (Dataset 0158). Sociometrics Corporation, Los Altos, CA.

Pearce-Moses, R., 2005. A Glossary of Archival and Records Terminology. Archival Fundamentals Series II. Society of American Archivists, Chicago.

PhRMA & European Federation of Pharmaceutical Industries and Associations, 2013. Principles for Responsible Clinical Trial Data Sharing. Retrieved 08.05.13. http://phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf.

Ross, J.S., Mulvey, G.K., Hines, E.M., Nissen, S.E., Krumholz, H.M., 2009. Trial publication after registration in ClinicalTrials.gov: a cross-sectional analysis. PLoS Medicine 6, e1000144.

Ross, J.S., Tse, T., Zarin, D.A., Xu, H., Zhou, L., Krumholz, H.M., 2012. Publication of NIH funded trials registered in ClinicalTrials.gov: cross sectional analysis. British Medical Journal 344, d7292.

Whyte, A., Wilson, A., 2010. "How to Appraise and Select Research Data for Curation". DCC How-to Guides. Digital Curation Centre, Edinburgh. Available online: http://www.dcc.ac.uk/resources/how-guides.

Arctic Archaeology and Prehistory

Owen K Mason, Institute of Arctic and Alpine Research, University of Colorado, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Arctic region in the northern hemisphere presented humans with dietary and temperature challenges and required a maritime sea mammal hunting or sea-ice adaptation. The colonization of this frigid and florally restricted environment occurred earlier than suspected, during the late Pleistocene *c*.27 000 years ago in northern Eurasia. In geographic terms, only northern Greenland, the coastal margins of Scandinavia, the Taymyr, and Alaska lie beyond the Arctic Circle. By contrast, sustained settlement of the North American High Arctic and Greenland did not occur until late in the Holocene, *c*.5000 years ago. Complex whaling societies developed in Bering Strait by 2000 years ago. For the most part, Arctic archaeology is synonymous with a single Mongoloid population, the Inuit (formerly Eskimo) who still dominate the region from Alaska to Greenland. A brief incursion between AD 900 and 1400 by the European Norse was renewed by Danish colonists in the early eighteenth century and continues to the present.

The Arctic region in the northern hemisphere presented humans with dietary and temperature challenges and required a maritime or sea-ice adaptation. The colonization of this frigid and florally restricted environment occurred earlier than suspected, during the late Pleistocene in northern Eurasia. In geographic terms, only northern Greenland, the coastal margins of Scandinavia, the Taymyr, and Alaska lie beyond the Arctic Circle. However, sustained settlement of the North American High Arctic and Greenland did not occur until late in the Holocene, *c*.5000 years ago. For the most part, Arctic archaeology is synonymous with a single Mongoloid population, the Inuit (formerly Eskimo) who still dominate the region from Alaska to Greenland. A brief incursion between AD 900 and *c*.1400 by the European Norse was renewed by Danish colonists in the early eighteenth century and continues to the present.

Ecological definitions of the Arctic, e.g., the absence of trees, typically extend its limit considerably south of the astronomical Arctic Circle at 66° 33″ N. latitude (Nuttall and Callaghan, 2000). The extent of perennial sea ice and the 10 °C July isotherm are convenient boundaries for the Arctic region. Several marine mammal species, seals, polar bears, walrus, and whales are ice-obligate; distributions that are synonymous with the Arctic. Several diverse geomorphic zones define the North American sector of the Arctic, broadly, the western consists of the mostly low and unglaciated coasts about the Bering Strait and Yukon River delta, northern Chukotka and the northern slope of Alaska and the Yukon, and Northwest territories. The treeless zone of Alaska extends far south to 50° N, which includes the lengthy volcanic arc of the Aleutian Islands. To the east, extensive Pleistocene glaciations sealed, compressed, and carved the ancient cratonic rocks of the Canadian Arctic Archipelago and maritime Greenland. The limits of the Arctic track sea ice distribution into Hudson Bay and as far south as Newfoundland. As sea ice retreats and temperatures rise, the boundaries of the ecological Arctic and its inhabitants shift across time and space; the Arctic sea ice of the Little Ice Age or Neoglacial was far more extensive than that of the anthropogenic twenty-first century.

Archaeological research commenced in the Arctic as a byproduct of eighteenth and nineteenth century colonial or commercial enterprises by Russians, Danes, and Americans (Collins, 1984; Fitzhugh et al., 2002). The sophisticated clothing and hunting technology of the indigenous Inuit impressed explorers and naturalists who subsequently sought its origins by archaeological methods. Archaeological investigations in the Arctic were conducted as early as 1906 as a colonial project by Danes, Americans, and Russians, following several centuries of interaction between Europeans and Inuit during which ethnographic observations were detailed, focusing on material culture, linguistics, and kinship. Anthropology largely matured as a discipline in the Arctic as early cultural anthropologists like Franz Boaz and Diamond Jenness conducted pioneering field work among the central Eskimo of Arctic Canada and Alf Porsild and William Thalbitzer did similar research in Arctic Greenland (Collins, 1984). High regard for the technological achievement of Inuit peoples among anthropologists, explorers, and colonial administrators led to the development of several Danish theories of Inuit origins by Hinrich Rink and Hans Steensby who proposed an interior source, followed by a descent along north flowing rivers through the forest into the treeless Arctic.

Although isolated objects were collected by naturalists and ethnographers such as the Swede Otto Nordenskjold in Siberia and American John Murdoch in Alaska in the late 1870s, the first large-scale effort at archaeological collection, was undertaken by Native laborers, directed by Vilhjálmur Stefánsson at Barrow, Alaska, in 1909, and produced evidence of an elaborate harpoon technology, often decorated. Systematic archaeological investigation started around 1925 in both the eastern and western Arctic. In the east, the Dane Therkel Mathiassen (1927) excavated in the central Canadian Archipelago and described a Thule culture, with an inferred age of 1000 years based on an extrapolated rate of uplift of formerly glaciated landscapes. In the western Arctic, Diamond Jenness inaugurated archaeological research *c*.1925 along the Bering Strait by discovering an 'archaic' Bering Sea culture on the Diomede Islands. Similar elaborated and decorated harpoon heads were uncovered in the late 1920s and early 1930s in large mounds on St. Lawrence Island and northern Alaska by Collins and his colleague James Ford. Research intensified in the Interwar Period, with most of the large coastal mounds investigated by

1945. The discovery of the Ipiutak site at Point Hope in 1939 revealed hundreds of houses and burials and stood for a generation as an unexplainable anomaly (Larsen and Rainey, 1948; Mason, 2006). Superimposed houses and occupation layers at a few sites (e.g., Onion Portage (Dumond, 1987)) allowed the construction of relative cultural chronologies that led to inferences of a rather limited prehistory of several thousand years, constrained by geomorphic factors such as continental glaciation and sea level history in regions transgressed by rising sea levels in the late Holocene. Since 1945, several generations of grant and government funded archaeologists and cultural resource managers have elaborated and refined the basic chronology. Research incorporated modern theoretical and methodological advances after 1970 under the aegis of research institutions such as the Smithsonian, Danish National Museum, Museum of Oriental Art (Moscow), Canadian Museum of Civilisation, the state universities of Wisconsin and Washington, as well as private schools like Brown University and Bryn Mayr. Interdisciplinary research commenced in the Aleutian Islands in the early 1960s (Dumond, 1987).

Modern humans, hunting extinct fauna, reached the 60th parallel in Europe and in Siberia during an interglacial of the Upper Pleistocene, *c*.42 000–35 000 years ago. Hunters in Yakutia were harvesting mammoth, woolly rhinoceros, caribou and horse at the Ikhine site around 34 000 years ago. For eastern Eurasia, humans ventured close to the Arctic Ocean during the Last Glacial Maximum, *c*.33 000–31 000 years ago. Extensive research by Pitulko (2012) at the Yana site shows a sophisticated bone technology and a highly developed linear aesthetic centered on pendants and zoomorphic figures.

A major hiatus in Arctic occupation followed after 25 000 years ago, the peak of the Last Glacial Maximum; apparently intense cold or decreased fauna restricted human population or interest in the North. Continental glaciation capped the Canadian and Greenlandic sectors, while the Bering Strait region was exposed as the subcontinent termed Beringia. The first evidence of a cross-continent Eurasia to North America migration occurs in the late Pleistocene, preserved in the Subarctic of Siberia and Alaska, *c*.14 000 years ago at well-dated archaeological sites in central Kamchatka and in the Nenana and Tanana valleys (Hoffecker and Elias, 2007). Direct evidence in the Bering Strait is assumed to lie on the then exposed, but now flooded, continental shelves.

The deglaciation of the Arctic Archipelagos proceeded over several millennia, after 10 000 BP, the onset of the Holocene. As polar bear and sea mammals recolonized the new lands, humans followed: first into the New Siberian Islands off of the Taymyr Peninsula, *c*.9000 years ago, where a distinctive microlithic technology was employed to dispatch bears and walrus (Pitulko, 2003) Canada and Greenland did not witness a similar early Holocene peopling, although the biota were present, based on ^{14}C dated whale bone on uplifted beaches. In the eastern zone, evidence of a late Pleistocene entry into northern Labrador is inferred from the transport of the distinctive Ramah chert into New England.

In the western Arctic, evidence of maritime hunting from *c*.9000 years ago was preserved on several islands, including Anangula and Unalaska in the eastern Aleutians. The origins of these populations remain uncertain (Knecht et al., 2001). Early researchers argued that the Anangulans relied on an exclusively unifacial technology, whose origins lie to the south. Subsequent researchers noted continuity with other microblade technologies along the Alaska Peninsula and onto Kodiak Island. Absent physical evidence, the use of boats, and the hunting of sea mammals are inferred from geomorphic setting and lithic technology.

A variety of technological innovations were required for sustained and successful Arctic adaptations. These included: a sophisticated repertoire of maritime craft (kayak and umiaq), dog traction, tailored skin clothing, snow goggles, as well as specialized sea mammal hunting equipment. The antiquity of this technology is not yet established, although its use is well authenticated following the development of large mounds in the last millennium BC. Architectural innovations were more limited but include subterranean houses as well as portable tents.

Around 5000 years ago, the coasts of northern and western Alaska witnessed their first intensive occupation, associated with a widespread and distinctive technology renowned for its fine and minute flint knapping. This Arctic Small Tool tradition (ASTt) was first encountered by J. Louis Giddings in 1948 at Cape Denbigh on the eastern shore of Norton Sound. The ASTt technology consisted of very small flaked stone cutting objects that were inset into bone hafts. ASTt peoples were predominantly sea mammal hunters and relying on the atlatl, or spear-thrower, and the toggling harpoon head, as well as on kayaks for transport.

Climate was possibly warmer at this time, which facilitated a rapid expansion of ASTt people from the Bering Strait into the interior and to northern Greenland. ASTt people ascended the rivers and mountains of northern Alaska, especially the Brooks Range, often using lakes to spear disadvantaged swimming caribou. Camp sites at Onion Portage are dated *c*.4200 years ago (Dumond, 1987). The Alaska migrants to the High Arctic must have already developed a complex marine hunting technology, including the bow and arrow as well as the kayak. At the frozen Greenland Qeqertasussuk site, Bjarne Gronnow (in Morrison and Pilon, 1994) obtained well-preserved wood, including kayak, bow and arrows, lances, spear and atlalt parts, barbed dart and leister points from a Saqqaq occupation around 4500 years ago (cf various in Morrison and Pilon, 1994). Subsistence relied on migrating harp seals and walrus, as well as ringed and bearded seals, fish (cod and char), birds, and several types of small whales both hunted and scavenged. The taking of the great auk at more northerly regions indicates a climatic warming. By 4000 years ago, the Saqqaq peoples had settled along western Greenland coasts, with a transition into the succeeding Independence I culture. Both Saqqaq and Independence I ASTt peoples were absent from Greenland after 800 BC, when another group makes its appearance from Arctic Canada. From the Canadian Archipelago to the Hudson Strait, another culture closely related to the ASTt, known as pre-Dorset, arose around 1200 BC. Pre-Dorset culture was also marked by distinctive harpoon heads, subsistence pursuits, and architecture (Maxwell, 1985).

The southern expansion of sea ice during the Neoglacial also allowed the ASTt to extend into the eastern Aleutian Islands (Knecht et al., 2001). Occupied *c*.5000 years ago, Margaret Bay on Unalaska Island is remarkable for its bone

artifact preservation and for a >1 m thick shell midden, with a diverse marine fauna. The bone objects include multiple barbed, tapering, cruciform stem, nontoggling harpoon darts – a very distinctive technology occurring for the rest of the Holocene in southern Alaska. Throwing board pins establish the use of the atlatl. Although harbor seal predominates, northern species were apparent on Unalaska in the mid-sixth millennium BP, most prominently polar bear and walrus that are not present today. Apparently, Margaret Bay was occupied during a cooler episode – archaeologists are still uncertain whether the site was the result of wandering northerners or whether southerners innovated due to the pressure of colder conditions and thrived beyond all measure, thus leading to a northward migration as the cold ceased. Although no boat pieces are known, it is inconceivable that maritime hunting on a large scale could have been entertained in the absence of watercraft.

In northern Alaska, the late Holocene development of sedentary and complex societies closely tracks the distribution of Pacific walrus, as Henry Collins first hypothesized in 1940. As a consequence of its cultural uses (cordage, food, skin for boat coverings, bone, and ivory for tools), walrus enhanced the role of women in skin work and fostered artistic production. Walrus hunting was possibly a precursor to hunting whales. By 5000 years ago on the prominent haul-out on Round Island in northern Bristol Bay, people were using spears tipped with notched bifaces. Several sites from Bristol Bay, Kotzebue Sound, and Wrangel Island parallel the distribution of walrus. This lithic technology was termed Chukchi Archaic by Mason and Gerlach (1995) and bears close resemblances to the technology of the first complex societies along the Bering Strait.

Around 4500–3500 years ago, a settling in of dramatic proportions occurred along the coast from the eastern Aleutians to the southern Alaska Peninsula (Dumond, 1987). In the next millennium, to 2500 years ago, a similar situation prevailed along the Bering Strait, especially on St. Lawrence Island, Cape Dezhneva, and the Diomede Islands (Fitzhugh and Crowell, 1988). This period was colder and stormier with enhanced upwelling, heightened nutrients, and marine productivity that benefited people across the region and led to a population surge. Following 2500 years ago, most cultural mounds added meter upon meter of height at critical and useful locations, occasionally accompanied by human bodies purposefully (or accidentally) interred. In addition, dedicated cemeteries appear in great numbers across the Aleutians and at the Bering Strait, providing evidence of sedentism. In northern Alaska and Chukotka, three cemeteries inferred to date to around 1000–500 BC contain high status graves with evidence of warfare and long distance trade (Mason, 1998; *see* Conflict and War, Archaeology of: Behavior and Social Organization) and in Chukotka, the first evidence of bronze in the far north.

Shell and bone middens, often forming elevated mounds, are ubiquitous on the Aleutian Islands – most with extensive viewsheds to sight prey and enemies. To further avoid attack, precontact sites were rarely at river mouths (McCartney, 1984). Many middens, extending 4 m in depth, contain whale bone houses and debris, with evidence of warfare, such as distinctive multi-barbed arrow points, increasing toward the more recent levels. Like most island dwellers across the Bering Sea and adjacent Gulf of Alaska, the lack of land mammals led to an increasing reliance on marine resources and required watercraft, the kayak, and the umiaq.

The two site types – mounds and village aggregations – reflect larger populations on the southern Alaska Peninsula from Port Moller north, and adjacent islands including the eastern Aleutians, Shumagin, and Sanak islands (McCartney, 1984; Maschner, 1999). The archaeological mounds formed after 4500–2700 years ago in core areas adjacent to hot spots for walrus and seal along Bering Strait and north, or along salmon streams to the south. Aggregations of abandoned houses number in the hundreds and a sophisticated harpoon technology supported fisheries based on cod or salmon, as ecologies shifted with colder or warmer sea surface conditions. The distributions of distinctive artifacts (such as harpoon heads and arrowpoints) track several migrations: (1) *c.*3000 years ago, a south to north trajectory by Choris peoples and (2) the Thule migration from west to east across Arctic North America and Greenland. Common stone building techniques and aesthetic ideas have led some archaeologists to posit direct genetic links between Aleutian islanders and the later, widespread Dorset culture.

The effects of climate on eastern Arctic peoples are difficult to gauge. Gaps in settlement history during cool episodes associated with Neoglacial expansion have fed interpretations of climate-induced retreat, especially in northern and central Greenland. Across the eastern Arctic from Greenland to Arctic Canada, *c.*1000 BC, the ASTt societies developed *in situ* from Saqqaq to pre-Dorset and Dorset. There were profound similarities in cultural practice over time among these archaeological cultures, especially in the use of a distinctive mid-passage house with upright slab hearths and side-notched spear points and side scrapers. A distinctive harpoon head, 'Tyara sliced' has traditionally marked the spread of Dorset culture around 600 BC. Sea mammal hunting sustained most societies, with core regions in Baffin Island in the Foxe Basin and along the Hudson Strait, where early Dorset occupations date to 800 BC. The major sites are Igloolik, the Tanfield site, and Arnapik (Maxwell, 1985).

From 500 BC to AD 1200, the Dorset culture was distributed across northern Canada, Greenland, and south to Labrador and Newfoundland. Early Dorset houses were square with short entry ways (Maxwell, 1985). Late Dorset societies were socially complex, occupying large multi-hearth stone structures that some archaeologists envision as theaters for shamanic performance. A repertoire of wooden masks distributed at many locales suggests performance. Dorset art is complex and figural, and is attributed to shamanistic beliefs, as revealed by its X-ray skeletal motifs and the portrayal of dangerous beings, including humans (Maxwell, 1985). Around 3000–2500 years ago, across much of mainland and southern Alaska, sizable villages of the Norton tradition developed at the mouths of salmon bearing streams, marked by the use of slate, net sinkers, the adoption of Siberian-derived ceramics, and dedicated mortuary precincts (Giddings, 1964; Dumond, 1987). Some researchers link the onset of sedentism with the Neoglacial and increased productivity. Larger villages, likely consisting of dozens of houses, are common east of Nome. Southern, possibly Aleutian influences are witnessed in its initial Choris phase by labrets and barbed harpoon heads. In

lithic technology, commonalities in flaking place both Choris and Norton in a line of descent from ASTt. The fate of the Norton villages is problematic; in the Yukon delta and to the south, a lengthy transformation into modern Yup'ik peoples occurred, with an intrusion and the influence of Inuit peoples of northern Alaska occurring around AD 700–900. In northern Alaska, Norton evolved into Ipiutak, an ally of Bering Strait societies, possibly with some influences from the interior.

Contemporaneous with Norton, the archaeology of the core regions adjacent to the Bering Strait hotspots witnessed a radical conceptual and socioeconomic breakthrough as early as 400–250 BC (Fitzhugh and Crowell, 1988). In Old Bering Sea (OBS) and Ipuitak societies, virtually every artifact was overlain with formalized circle and dot motifs and as figural and facial representations was crafted, both of animals and humans. Three principal artistic styles are defined: Okvik, OBS (I, II, and III), and Ipiutak (Collins, 1973; Dumond, 1987; Fitzhugh et al., 2009). Cemeteries from the earliest cultures contain evidence of a few high status individuals, both men and women, some of whom engaged in long-distance trade in obsidian across the Bering Strait (Mason, 1998). Art appeared earliest along the northern Alaska coast, at Battle Rock, and north of Cape Krusenstern probably dating to the Norton phase (*c*.500 BC) and was spare and linear, apparently decorative in function. The radical shift is that animals and humans were portrayed in Okvik and Old Bering cultures; common subjects include animal faces, mother and child, as well as submissive males (Collins, 1973). Ethnographic parallels can be read so that the designs represent cosmologies although sympathetic hunting magic, and the strong function to ward off evil (cf various in Fitzhugh et al., 2009). The objects reflect that only a few shamans had power and were a symptom of a society in need of power; some archaeologists interpret Ipiutak mortuary art as part of a crisis cult, related to adverse climatic conditions (Mason, 2006, 2009).

The OBS culture occurs on both north and south Chukotka coasts, on the Diomedes, and on St. Lawrence Island. It had only a limited distribution in Alaska, with isolated finds from the Yukon Delta to Pt. Barrow. Prevalent between AD 200 and 900, the distinctive art styles of OBS and Ipiutak delineate several allied societies that traded and engaged in warfare with each other or with an external enemy, possibly the Birnirk societies of the more northerly Chukchi Sea (Mason, 1998). Evidence of warfare is unequivocal and includes distinctive weaponry (daggers and spear heads) as well as human bone impaled with bone points. Mortuary contexts record specialized leaders and shamans. Although the dating of the two art styles is still in development, it appears that the most elaborate art was produced within a few centuries between AD 600 and 900 (Mason, 2006), part of a crisis cult associated with anomalous weather that stressed societies across both the western and eastern Arctic. Iron and obsidian were used for cutting or engraving bone and indicate the distant extent of extraregional exchange systems. The iron is metallurgical, not meteoric, but evidence of iron use by OBS peoples is restricted to two burins from Ekven and one from Uelen, and one large axe from St. Lawrence Island. Iron burins also occur at three sites in the contemporaneous Ipiutak culture, which dominated the Alaska mainland (Fitzhugh and Crowell, 1988). Obsidian was obtained from considerably farther south, in Kamchatka. The role of dogs in transportation and diet remains uncertain although small breeds were kept, possibly for dietary and sacrificial purposes, based on the number of dog bones and the treatment of their skulls. On the other hand, sleds for ice hauling are evident from the numerous ivory runners and shoes recovered.

Between AD 800 and 1000, a drastic shift occurred in nearly all Arctic societies (Dumond, 1987; Mason, 2009; McGhee, 1984), with the crisis cults, like Dorset, Ipiutak, and Kachemak disappearing and new whale-based societies gaining momentum. These changes possibly represented new ethnic amalgams, if not new peoples (Mason, 1998). The Thule culture originated along a wide margin from northern Chukotka to northern Alaska (Dumond, 1987; McGhee, 2001). Innovations in sea mammal hunting (drag-float apparatus) and nautical technology co-occurred with architectural shifts and mortuary patterns that favored subfloor burial over cemeteries. A de-emphasis of decorated ivory co-occurs in Alaska with decline in iron supply and the more extensive use of jade, amber, and slate, possibly representing a revaluing of aesthetic frameworks. In northern Alaska, status differentials in burials are rare, as are large village sites. The long-standing belief among archaeologists is that the pan-Arctic migration of Thule peoples was catalyzed by high populations stressed by the unfamiliar warmth of the Medieval Climate Anomaly. Very little data support this scenario and proxy climate show colder conditions *c*.AD 1000 precede the movement from Alaska to the east. The Thule culture continued to adjust to the cooler climates of the Little Ice Age, with increasing success in whaling within the core regions of central Canadian Arctic, in the Mackenzie delta and in the northwest Alaska (McGhee, 2001; various in Arneborg and GrØnnow, 2006; Maschner et al., 2009). In the latter area, a series of large villages dominated, with populations of several hundred at Point Hope, Point Barrow, and Cape Prince of Wales (Dumond, 1987; Maschner et al., 2009). Seasonal trade fairs were established at several centers, including the Colville River, Barter Island, Wales, Kotzebue, and the Anadyr, possibly as early as AD 1400, with Russian trade goods entering the western Arctic in the seventeenth century, if not earlier. Population increases were associated with increased warfare, intensified by the introduction of East Asian military practices, especially slat armor.

The well-documented migration of Thule people resulted in an occupation throughout Greenland and the Canadian Arctic, including northern Quebec, eastern Hudson Bay, and as far south as Labrador. This migration commenced in the thirteenth century AD and occurred in several stages (GullØv, 1997; Arneborg and GrØnnow, 2006; McGhee, 2001; Maschner et al., 2009). A complex relationship existed between in-migrating Thule and the preceding Dorset peoples, possibly one of intermarriage. The two cultures were contemporaneous between AD 1200 and 1300. Artifact forms, especially harpoon heads, show contact, but no firm evidence of warfare exists and slat armor is not found in Arctic Canada or Greenland (Maschner et al., 2009). Thule settlements were typically composed of several houses, and are well-known for their use of whale bone for structural elements and flat rock slabs for flooring and low rock walls.

Thule peoples thrived during the Little Ice Age and were accomplished whalers, employing harpoon heads similar to those of western Alaska. Thule whale hunters specialized in the capture of 7–11 m animals, according to faunal data interpreted by McCartney (1995; *see* Zooarchaeology). Thule life ways and whaling techniques are illustrated in carvings on whale bone. The Thule culture forms the material substrate of the late prehistoric and contact Inuit who were encountered first by the Norse *c*.AD 1000 and subsequently by the Danes and British, commencing with Martin Frobisher's landing on Baffin Island in 1576. The archaeology of the Thule culture was launched by Therkel Mathiassen (1927) and the Fifth Thule expedition of the 1920s.

The late tenth century AD also witnessed the colonization of southern Greenland by Norse peoples from Iceland and Scandinavia (Lewis-Simpson, 2002). The Greenlandic colonies represent a total break with the marine-centric subsistence economies that prevailed across the Arctic. The Norse did attempt to rely on transhumant pastoralism of sheep, goats, and cattle; nonetheless, seal and other marine animals were a significant part of the diet. Two principal settlements, Western and Eastern, were situated on the southernmost shore of Greenland, within the inner fjords. Houses and churches were constructed of stone and sod, with animals kept within the houses. The Greenlandic Norse maintained close economic and social ties with Iceland and Europe, communicating with Church officials until the early fifteenth century. The Norse successfully transplanted stone architecture and woven woolen clothing in distinction to skin clothing of their Inuit neighbors, with whom iron was transferred intermittently. The abandonment of the Norse settlements remains uncertain; various scholars attribute their demise to adverse climate change, adverse pasture management, plague, or excess taxation. In the early eighteenth century, Danish colonists returned to Greenland and established villages along its southern and western shores; archaeological studies of such remains provide a record of European trade items and distinctive architecture.

European contact with Native societies in the western Arctic started in the late seventeenth century AD as Russian fur traders entered eastern Siberia and supplied tobacco and other goods to local trade fairs along the Chukotkan coast (Vaughan, 1999). A brutal military campaign reduced the Koryak and Chukchi peoples and intensified militarism through the Bering Strait region. The Russians established only two settlements, on the Kuskokwim River and at Saint Michael, north of the Aleutian Islands, preferring to employ Aleuts to harvest fur seals on the Pribilof Islands. Capt. James Cook was the first European to map northwest Alaska, reaching as far as Icy Cape in 1778. The British under Capt. William Beechey were the first Europeans at Barrow by 1826. Only two Russian expeditions ventured along the coast of northwest Alaska, that of Otto von Kotzebue in 1816 and that of I. Kasevarov in 1838. Following its transfer to the United States, American whalers and entrepreneurs established trading and whaling stations at various communities on the north coast of Alaska (Vaughan, 1999). Native populations were decimated in the nineteenth century by a combination of epidemic disease and the effects of overharvesting of whale. Archaeological and ethnohistorical research has established the early to mid-nineteenth century introduction of firearms, metal and ceramic cooking equipment, and other dietary changes such as the use of flour and alcohol. The later nineteenth century also witnessed drastic changes in mortuary practices, a shift in housing from subterranean to above-ground frame structures, and in the adoption of wood and coal-fired stoves. Metal tools, especially for woodworking, cooking, whaling, etc. replaced stone tools after 1867. Due to overharvesting by Euro-American whalers, Native whaling nearly vanished by 1920, but was successfully revitalized after 1940. The introduction of the rifle after 1870 led to Native overhunting of caribou. The U.S. territorial government response to this crisis by Education administrator Sheldon Jackson led to the introduction of reindeer and a European herding culture. Christian missionizing by various sects led to the construction of churches and the establishment of bounded cemetery precincts. In the twentieth century, Euro-American/Canadian concepts of settlement design and housing took precedence with the establishment of schools and stores in central places – a process that led to the abandonment of many smaller camp sites. The absorption of Native communities into the global system continues in the twenty-first century, tempered by the persistence and resilience of communities in the face of this onslaught and the increasingly dire effects of global anthropogenic climate change.

See also: Chiefdoms, Archaeology of; Conflict and War, Archaeology of: Behavior and Social Organization; Ethnic Identity and Ethnicity in Archaeology; Ethnoarchaeology; Historical Archaeology; Hunter-Gatherer Societies, Archaeology of; Migrations, Colonizations, and Diasporas in Archaeology; Ritual and Symbolism, Archaeology of; Zooarchaeology.

Bibliography

Arneborg, J., GrØnnow, B., 2006. Dynamics of Northern Societies. In: Publications of the National Museum of Denmark. Studies in Archaeology and History, vol. 10. Copenhagen.

Collins, H.B., 1984. History of research before 1945. In: Damas, D. (Ed.), Handbook of North American Indians, vol. 5. Arctic, Smithsonian Institution Press, Washington, DC, pp. 8–16.

Collins, H.B., 1973. Eskimo art. In: Collins, H.B., de Laguna, F., Carpenter, E., Stone, P. (Eds.), The Far North: 2000 Years of American Eskimo and Indian Art. National Gallery of Art, Washington, DC, pp. 1–132.

Dumond, D.E., 1987. Eskimos and Aleuts, second ed. Thames and Hudson, London.

Fitzhugh, W.W., Crowell, A. (Eds.), 1988. Crossroads of Continents. Smithsonian Institution Press, Washington, DC.

Fitzhugh, W.W., Hollowell, J., Crowell, A. (Eds.), 2009. Gifts from the Ancestors: Ancient Ivories of Bering Strait. Princeton University Press, Princeton.

Fitzhugh, W.W., Loring, S., Odess, D. (Eds.), 2002. Honoring Our Elders: A History of Arctic Archaeology. Contributions to Circumpolar Anthropology 2. Smithsonian Institution, Washington, DC.

Friesen, M., Mason, O.K. (Eds.), 2015. Oxford Handbook of Arctic Archaeology. Oxford University Press, Oxford.

Giddings, J.L., 1964. The Archeology of Cape Denbigh. Brown University Press, Providence.

GullØv, H.C., 1997. From Middle Ages to Colonial Times: Archaeological and Ethnohistorical Studies of the Thule Culture in Southwest Greenland, 1300–1800 AD. Med. GrØnland, Man & Society 23.

Hoffecker, J.F., 2005. A Prehistory of the North: Human Settlement of the Higher Latitudes. Rutgers University Press, Brunswick, NJ.

Hoffecker, J.F., Elias, S., 2007. Historical Ecology of Beringia. Cambridge University Press, Cambridge.

Knecht, R.A., Davis, R.S., Carver, G.A., 2001. The Margaret Bay site and eastern Aleutian prehistory. In: Dumond, D.E. (Ed.), Archaeology in the Aleut Zone: Some Recent Research. University Oregon, Anthropological Papers 58, pp. 9–34.

Larsen, H., Rainey, F., 1948. Ipiutak and the Arctic Whale Hunting Culture. American Museum Natural History Anthropological Paper 42, New York.

Lewis-Simpson, S. (Ed.), 2002. Vinland Revisited: The Norse World at the Turn of the First Millennium. Historic Sites Foundation of Newfoundland and Labrador, St. Johns.

Mason, O.K., Gerlach, S.C., 1995. The archaeological imagination, zooarchaeological data, the origins of whaling in the western Arctic and "Old Whaling" and the Choris cultures. In: McCartney, A.P. (Ed.), Hunting the Largest Animals: Native Whaling in the western Arctic and Subarctic, Studies in Whaling No. 3, Occasional Paper No. 36. Circumpolar Institute, University of Alberta, Edmonton, pp. 1–32.

Maschner, H.D.G., Mason, O.K., McGhee, R. (Eds.), 2009. The Northern World A.D. 1100–1350: The Dynamics of Climate, Economy and Politics. University Utah Press, Salt Lake City.

Maschner, H.D.G., 1999. Sedentism, settlement and village organization on the lower Alaska Peninsula: a preliminary assessment. In: Billman, B.R., Feiman, G.M. (Eds.), Settlement Pattern Studies in the Americas: Fifty Years Since Viru. Smithsonian Institution Press, Washington, pp. 56–76.

Mason, O.K., 1998. The contest between Ipiutak, Old Bering Sea and Birnirk polities and the origin of whaling during the first millennium A.D. along Bering Strait. Journal of Anthropological Archaeology 17 (3), 240–325.

Mason, O.K., 2006. Ipiutak remains mysterious: a focal place still out of focus. In: Grønnow, B. (Ed.), Dynamics of Northern Societies. Proceedings of a Symposium. Danish National Museum, Danish Polar Center, Copenhagen, pp. 106–120.

Mason, O.K., 2009. Flight from the Bering strait: did Siberian punuk/Thule military cadres conquer northwest Alaska. In: Maschner, H., Mason, O.K., McGhee, R. (Eds.), The Northern World A.D. 1100–1350: The Dynamics of Climate, Economy and Politics. University Utah Press, Salt Lake City, pp. 76–128.

Mathiassen, T., 1927. Archaeology of the Central Eskimos. Report of the Fifth Thule Expedition. vol. IV (Parts i and ii). Glendendalske Boghandel Nordsik Forlag, Copenhagen.

Maxwell, M., 1985. Prehistory of the Eastern Arctic. Academic Press, New York.

Morrison, D., Pilon, J.L. (Eds.), 1994. Threads of Arctic Prehistory: Papers in Honour of William E. Taylor, Jr. Canadian Museum of Civilisation, Hull.

McCartney, A.P., 1984. Prehistory of the Aleutian region. In: Damas, D. (Ed.), Handbook of North American Indians, vol. 5. Arctic, Smithsonian Institution Press, Washington, DC, pp. 119–135,.

McCartney, A.P., 1995. Whale size selection by precontact hunters of the North American western Arctic and Subarctic. In: McCartney, A.P. (Ed.), Hunting the Largest Animals: Native Whaling in the Western Arctic and Subarctic, Studies in Whaling 3, Occas. Paper 36. Circumpolar Institute, University Alberta, Edmonton, pp. 83–108.

McGhee, R., 1984. Thule Prehistory of Canada. In: Damas, D. (Ed.), Handbook of North American Indians vol. 5. Arctic, Smithsonian Institution Press, Washington, DC, pp. 369–376.

McGhee, R., 2001. Ancient People of the Arctic. UBC Press, Vancouver.

Nuttall, M., Callaghan, T. (Eds.), 2000. The Arctic: Environment, People, Policy. Harwood Academic Publishers, Amsterdam.

Pitulko, V., 2003. The bear hunters of Zhokhov Island, eastern Russian arctic. Senri Ethnological Studies 63, 141–152.

Pitulko, V., 2012. The oldest art of the Eurasian Arctic: personal ornaments and symbolic objects from Yana RHS, Arctic Siberia. Antiquity 86, 642–659.

Vaughan, R., 1999. The Arctic: A History. Sutton Publishing, Ltd., Sparkford, UK.

Arctic: Sociocultural Aspects

Peter P Schweitzer, University of Vienna, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Arctic and its peoples have intrigued southeners for millennia. Currently, there is heightened interest in the North due to its resource potential and the pronounced effects of climate change. This article focuses on a discussion of 'traditional' indigenous cultures in the Arctic in the nineteenth to the twentieth centuries. The material will be presented along topical domains: ecology and economy, social and political organization, religion and worldview. The final section will include a discussion of contemporary cultural and social trends among indigenous peoples of the Arctic, as well as of their recent political struggles for self-determination in the face of climate change and globalization.

Contemporary news flashes about the Arctic typically focus on melting sea ice, drowning polar bears, increased Arctic shipping, and northern oil and gas prospects. Long before the current interest in the geopolitics of the 'New Arctic,' the North captivated the European imagination. Western curiosity about the peoples of the North can be traced back to the ancient Greeks. Still, the scholarly study of the Arctic as an area of shared cultural traits and environmental conditions hardly predates the year 1900. The early twentieth century paradigms of diffusionism and environmental determinism were instrumental in creating the simplistic notion of a unified Arctic or circumpolar culture (e.g., Bogoras, 1929; Hatt, 1914). Detailed ethnographic research conducted since that time has demonstrated that variation is an intrinsic characteristic of Arctic sociocultural systems (some of this research is summarized in Berg, 1973; Graburn and Strong, 1973; Irimoto and Yamada, 1994). Nevertheless, this article approaches the subject by focusing on sociocultural similarities, without denying the existence of considerable differences.

Recent years have seen a tremendous increase in northern circumpolar social science research. In 2004, the *Arctic Human Development Report* (AHDR, 2004) appeared; it provided a first pan-Arctic review and assessment of humanities and social science scholarship. Among other things, the report stated that 'fate control,' 'cultural integrity,' and 'contact with nature' are closely associated with human development and well-being in the Arctic. The *Arctic Social Indicators* project (ASI, 2010) – a follow-up report to AHDR 2004 – made use of these insights and added the above-mentioned domains to the more conventional domains of education, health and material well-being, to come up with a small set of social indicators for the Arctic.

Another important report of these years – endorsed by the Arctic Council (AC) like the others – was the *Arctic Climate Impact Assessment* (ACIA, 2005), which was the first comprehensive regional climate change report that paid attention to the impacts on the residents of that region. Taking into consideration other major projects of the time, such as the *Encyclopedia of the Arctic* (Nuttall, 2005), it is fair to say that the first decade of the twenty-first century was characterized by an unprecedented output of comprehensive and synthesizing Arctic social science and humanities publications.

The Arctic and Its Indigenous Inhabitants

Culturally, the Arctic can be subdivided into a North American and a northern Eurasian part. The North American Arctic is primarily inhabited by speakers of Eskimo–Aleut languages. The old collective ethnonym Eskimo is little used today, and commonly replaced by Inuit and Yupik. Aleut, Yupik, and Inuit societies inhabit the coastal areas of northern North America, stretching from southern Alaska to eastern Greenland. Here the geographical boundary between Arctic and Subarctic coincides more or less with the cultural boundary between Inuit/Yupik/Aleut and Athapaskan and Algonquian groups of North American Indians (*see* Americas, Sociocultural Overviews: North America, Native Americans).

In the northern part of Eurasia, the physical boundary between Arctic and Subarctic does not coincide as neatly with cultural boundaries. In Siberia, the cultural realm of the Arctic extends into the Subarctic ecological zone, and reaches its limits only in the steppes of southern Siberia. Samoyedic Tungusic, and Paleoasiatic languages are spoken by the culturally 'most typical' Siberians. Speakers of Turkic languages inhabit large parts of the eastern Siberian tundra and boreal forest, but their historical and cultural background points to Central Asia. In northern Europe, where the vegetational and climatic zone of the Arctic is narrow, conventionally only the Saami (speakers of Finno-Ugric languages) are considered an indigenous Arctic people. Other peoples, who also inhabit the northern margins of Europe but are organized into large-scale agricultural societies, are not considered in this overview (*see* Europe: Sociocultural Overview).

Indigenous and Colonial Histories

Human habitation of the circumpolar North extends over several thousand years. While the first modern humans expanded into areas up to 60° northern latitude between 45 000 and 20 000 years ago, it was only after the end of the Last Glacial period that the Arctic became fully inhabited by people (Hoffecker, 2005: 8). While the direct presence of European colonial powers in the Arctic is a relatively recent phenomenon, findings of iron and other items that were not

produced locally attest to long-standing connections with trade centers to the South. The territory of the Saami has a history of almost 2000 years of outside intervention, as Vikings pushed north along the western coast of contemporary Norway to extract natural resources and to acquire items of Saami production through trade and tribute. Similarly, European expansion into other parts of the North was fueled by the quest for marketable resources. From the seventeenth century onwards, the rich boreal forests of Siberia and Canada became staging areas for the fur trade. The areas north of the tree line were little affected by fur trapping prior to the twentieth century, but the coastal areas of the Arctic close to the Atlantic and Pacific Oceans became important destinations for the Euro-American whaling industry in the eighteenth and nineteenth centuries.

The early days of colonial rule had little direct administrative impact in many areas of the North. Often it amounted to little more than the purely nominal claim to 'owning' the land, and to sporadic resource extraction. In certain areas, however, the state took a more active role in regulating the lives of the indigenous population: For example, in Greenland a particular form of 'enlightened paternalism' was the guiding principle of Danish rule from the eighteenth century onwards. The second half of the nineteenth century was in this respect far more disruptive than earlier periods: The newly emerging ideology of nationalism brought indigenous peoples from northern Scandinavia to Alaska under increasing pressure to adopt the language and culture of the respective dominant societies.

Domains of Arctic Sociocultural Systems

The following section will provide a discussion of various aspects of 'traditional' Arctic cultures. Using the turn of the twentieth century as the ethnographic present, the material is organized along topical lines.

Ecology and Economy

Characteristic of the Arctic fauna is the relative small number of available species, while the number of individual animals at particular times can be high. Similarly, vegetation north of the tree line is sparse, enjoying only a brief but dramatic growing season. Thus, periods of abundance at a particular locale are followed by periods of shortage, and human groups must schedule their movements accordingly. Due to the wide distribution of permafrost, in combination with climatic and vegetational factors, the practice of agriculture was impossible until recently. Thus, all indigenous Arctic peoples practiced foraging forms of subsistence, such as hunting, fishing, and gathering (*see* Hunting and Gathering Societies: Anthropology).

In the tundra zone, collective hunts of wild reindeer (caribou) during spring and fall migrations constituted extremely important subsistence activities. Along the coasts of the Arctic Ocean the pursuit of sea mammals was life sustaining. Seals formed the staple, and areas visited by migrating whales and walrus had the opportunity to acquire large quantities of meat through one successful hunt. The lower reaches of major rivers provided excellent opportunities for seasonal fishing. The inland areas of boreal forests could generally rely less on one or two major resources, but had to combine hunting (wild reindeer, moose, etc.), fishing, and gathering. The changes brought about by the fur trade triggered dramatic shifts in the seasonal rounds of the peoples inhabiting the Subarctic. The hunt for furbearers, which previously had been of little significance, became the most important economic activity.

The major form of subsistence outside of hunting/gathering/fishing is reindeer herding, a form of pastoralism (*see* Pastoralism in Anthropology). Reindeer domestication is found in northern Eurasia, from northern Scandinavia to the Bering Strait, but did not penetrate into northern North America until the late nineteenth century. Among the various local forms of reindeer herding, two major types emerge. Small herds of domesticated reindeer are found throughout the boreal forest regions of Siberia; their primary use is for transportation during hunting. Large-scale reindeer herding of the tundra, on the other hand, is geared toward the maximization of animal numbers, and provides the group with meat and other reindeer products. The only pan-Arctic domesticated animal is the dog, which is used for transportation (and sometimes as a sacrificial animal).

Most economic production was geared toward household and community consumption. Sharing of resources within these limits was a consistent feature of Arctic and Subarctic societies. Exchange relations with other communities were often facilitated through ritual partnerships, which extended the limits of reciprocity. With the advent of colonialism, production for outside markets commenced: furs, meat of domesticated reindeer, and other products were individually appropriated; they fueled the emergence of economic stratification.

In recent years, many parts of the Arctic have experienced dramatic environmental change, be it climate induced or not (e.g., Keskitalo, 2008). Indigenous residents of the North, who have a long track record of monitoring environmental dynamics, provided detailed observations of these changes (see, e.g., Krupnik and Jolly, 2002). The impact of climate change is not limited to the environment. Traditional ecological knowledge and the social prestige it carries are being eroded by abrupt changes, which prevent accurate forecasts. In addition, modernization and globalization have impacted all so-called 'traditional' economic activities in the North, resulting in 'mixed economies' combining subsistence and cash elements. As has been demonstrated with regard to Aleut commercial fishing (Reedy-Maschner, 2010), contemporary indigenous identities require both economic success and traditional values.

Social and Political Organization

It has long been suggested that Arctic societies are characterized by a bilateral type of social organization (Gjessing, 1960). Traditional Saami societies, the Paleoasiatic groups of northeastern Siberia, as well as the majority of Inuit and Yupik societies clearly fit this pattern. The societies inhabiting the western, central, and eastern parts of Arctic Siberia, however, display patrilineal types of organization. Given the fact that

vectors of cultural influence in Siberia run generally from south to north, it is possible to suggest that unilineal social organization among those groups was stimulated by southern (Central Asian) influences. The unilineal tendencies of certain Yupik and Aleut societies are less easy to comprehend. There are indications of patrilineal 'clans,' sometimes with endogamous tendencies (Petrivelli, 2004); however, the basic bilateral framework of Eskimo–Aleut societies (privileging horizontal links over lineal ties) is also present in these cases.

It is noticeable that none of the Arctic or Subarctic societies of Eurasia show any traces of matrilineal organization. In Subarctic North America, on the other hand, the majority of northern Athapaskans (including the linguistically related Eyak and Tlingit) are textbook examples of descent reckoned in the female line. Generally speaking, the Subarctic forests are dominated by unilineal kinship systems. However, it is unclear whether there is a causal relationship between unilineality and Subarctic lifeways and between bilaterality and tundra lifestyles. While bilateral systems are more flexible and seem more appropriate for small-scale groups scattered over large territories than more rigid unilineal systems, population densities in the boreal forests were hardly larger than in the tundra. On the contrary, certain Arctic coastal communities (e.g., near the Bering Strait) sustained quasi-sedentary settlements of several hundred inhabitants; other than along the mouths of salmon-rich rivers, no comparable population concentrations are known from the aboriginal Subarctic.

Parallel to the fate of kinship studies elsewhere, relatedness as a topic of arctic investigations saw its heydays in the 1960s and 1970s (see, e.g., Burch, 1975; Guemple, 1972; Pehrson, 1964). After a 20-year period of little attention to kinship in the north, the topic has returned albeit with a different set of foci. On the one hand, carefully crafted ethnographic accounts of kinship as lived social relations help us understand the meanings of relatedness in particular Arctic contexts and situations (e.g., Bodenhorn, 2000; Nuttall, 2000). On the other hand, the role of social relations and kinship for economic organization, the sharing of resources, and for the constitution of property relations has become more prominent in recent Arctic studies (see, e.g., Magdanz et al., 2002; Petersen, 2003; Ventsel, 2005).

Throughout the circumpolar North, political organization was local-group oriented and did not entail hierarchies based on hereditary status. Leadership continues to be situational: It is directed toward the solution of group problems, and the decision to comply is voluntary. In addition to seniority, individual achievements are most relevant in gaining leadership positions. The major exception is the case of the Sakha (or Yakut as they were known previously), whose political organization has been characterized as a chiefdom (Graburn and Strong, 1973). For the known cases of ranking in the southern parts of Alaska (e.g., Aleut, Alutiiq), influence from the northern northwest coast of North America (especially from the Tlingit) can be conjectured (Schweitzer, 2003).

Men have often been the only political actors visible to outside observers. Although it is true that many Siberian societies with patrilineal descent have pronounced male-centered ideologies, not all Arctic societies were dominated by 'man, the hunter.' For example, Inuit males, who seemingly provided the large majority of food resources, were entirely dependent on women to process the meat and skin of the slaughtered animals, to make them into usable food and clothing. Thus, the roles of men and women were strongly complementary and did not sustain rigid gender hierarchies.

Religion and Worldview

Shamans and shamanism are probably the most evocative symbols of circumpolar religion and worldview (*see* Shamanism). There is no doubt that – until recently – most Arctic communities had religious functionaries who were able to communicate with and to 'master' spirits. These 'shamans' were engaged in healing and other activities aimed at improving communal and individual well-being. In the small-scale societies under consideration here, these functionaries held extremely important social positions, which sometimes led to an abuse of power. However, the notion of 'shamanism' can easily be misconstrued as a unified system of beliefs, which it never was in the Arctic. Instead, in addition to a limited number of common elements, circumpolar shamanisms show profound differences in the belief systems with which they are associated. Especially in northern Eurasia, elements of worldviews associated with highly organized religions (such as Buddhism or Christianity) found their way into localized forms of shamanism long before the direct impacts of colonialism.

Animism – the belief that all natural phenomena, including human beings, animals, and plants, but also rocks, lakes, mountains, weather, and so on, share one vital quality – the soul or spirit that energizes them – is at the core of most Arctic belief systems. This means that humans are not the only ones capable of independent action; an innocuous-looking pond, for example, is just as capable of rising up to kill an unsuspecting person as is a human enemy. Another fundamental principle of Arctic religious life is the concept of humans being endowed with multiple souls. The notion that at least one soul must be 'free' to leave the human body is basic to the shaman's ability to communicate with the spirits.

Since the killing and consumption of animals provides the basic sustenance of circumpolar communities, ritual care taking of animal souls is of utmost importance. Throughout the North, rituals in which animal souls are 'returned' to their spirit masters are widespread, thus ensuring the spiritual cycle of life. While most prey animals receive some form of ritual attention, there is significant variation in the elaboration of these ceremonies. One animal particularly revered throughout the North is the bear (both brown and black), as has been demonstrated by Hallowell (1926) in his classic comparative study of 'bear ceremonialism.' A recent phenomenological study of the spiritual dangers of hunting among the Siberian Yukaghirs was provided by Willerslev (2007).

The ontologies and cosmologies of the circumpolar north are not unique. Many elements of animistic and shamanic worldviews are widespread among hunter-gatherers anywhere. Similarities among the ontologies of American Indian and Siberian peoples might also have historical foundations, since the peopling of the Americas is thought to have happened through northeastern Asia. Recently, a volume devoted to exploring the similarities and differences between notions of nonhuman personhood in Siberia and Amazonia has

conceptually and geographically extended the notion of Arctic animism (Brightman et al., 2012).

By the twentieth century, hardly any Arctic community had not yet felt the impact of Christian missionary activity. However, there is considerable variation as to when these activities commenced: Christianity reached the Arctic areas of Europe almost 1000 years ago, while the indigenous inhabitants of the Chukchi Peninsula (Russia) had little first-hand experience of Christianity before the 1990s. Generally speaking, the eighteenth and nineteenth centuries mark the major periods of religious conversion in the Arctic. Although no other major world religion has significantly impacted the North, the spectrum of Christian denominations represented in the Arctic is considerable. There is also considerable variation in how 'nativized' the individual churches have become. A good example for the latter, the indigenization of Christianity, are the developments in the southern parts of Alaska after the sale of Russian America to the United States in 1867. These regions – inhabited primarily by Aleut, Alutiiq, Eyak, and Tlingit people, as well as by some Denaina and Yupik groups – came under Russian cultural and religious influence in the eighteenth and nineteenth centuries. While this initial spread of Russian Orthodox Christianity is not different from colonial processes elsewhere on the globe, the persistence of the faith after the end of Russia's colonial rule deserves attention and explanation. Such explanations include references to the specifics of the Orthodox religion (e.g., Mousalimas, 1995), to analogies between indigenous and Russian religious concepts (e.g., Csoba DeHass, 2009), as well as to the identity politics of the early US period of Alaska (e.g., Kan, 1999).

The late twentieth and early twenty-first centuries have experienced renewed missionary activities by 'new' churches throughout the Arctic. A particular focus of missionary activity has been the Russian Arctic, which has been viewed by outside (evangelical) churches as a kind of religious *terra nullius* after 70 years of state atheism.

Contemporary Developments

Throughout the circumpolar North, World War II triggered developments that made the once-distant Arctic frontiers into strategically important areas. The resulting infrastructural and demographic developments altered the social fabric of the North and put an end to official policies of isolationism. The first two decades after 1945 were generally characterized by state policies geared toward 'modernization' and assimilation. It is frightening how similar state policies from Siberia to Greenland were: Native people were relocated from ancient villages to faceless new towns, a new emphasis on 'productivity' favored newly introduced economic activities over traditional subsistence pursuits, and educational systems were reoriented toward nonnative knowledge and skills. It may seem ironic that the very same policies provided the educational infrastructure for those native elites who would challenge the political and cultural hegemony of the colonial powers in the years and decades to come.

Between the 1960s and the 1980s, radical political change affected most parts of the circumpolar North. With the exception of the Soviet North, the general tendency was to repeal colonial status and chart a course toward self-determination. In many cases, opposition to large-scale development projects served as a rallying point for newly developing indigenous movements. For example, the conflict around the construction of the Alta hydroelectric dam in Norway in the late 1970s brought international attention to indigenous causes, and marks a sea change in the political history of the Saami. In Alaska, it was the discovery of oil in 1968 that put the long-neglected topic of native title to land on the agenda. The resulting 'Alaska Native Claims Settlement Act' (1971), which awarded title to 12% of Alaska's land, and cash compensation to newly created native corporations, was then considered a spectacular success.

One of the most impressive results of the era was the 1979 passing of the 'Home Rule Act' in Greenland, which provided limited sovereignty within Denmark (control over most public affairs except defense and foreign relations). In 2009, Greenlandic 'self-rule' was established, the next step toward independence from Denmark. Between 1973 and 1993, Finland, Norway, and Sweden installed so-called Saami parliaments, which – although only advisory – provide Europe's northernmost indigenous peoples with powerful symbols of sovereignty. The most recent change on Canada's political map was the creation of Nunavut in 1999, the first and only Canadian territory in which Inuit make up the majority population.

Developments in the Soviet North followed a different pace. The first 'ethnic' administrative units were formed in Siberia in the 1920s, when such events would have been unthinkable in most other countries, but the political conditions of the subsequent decades reduced notions of 'national autonomy' to mere propaganda instruments. Thus, the reforms of the late 1980s, which led eventually to the demise of the Soviet Union, gave hope for advances in the realm of native rights. Although many changes were undoubtedly positive, the social and economic situation of most native communities in the Russian North deteriorated throughout the 1990s. Many Siberian natives have entered the twenty-first century with nostalgic memories of the Soviet period.

Cooperation among Arctic nations improved after the end of the Cold War, which resulted in the establishment of the AC in 1996 (its predecessor – the Arctic Environmental Protection Strategy – had been founded in 1991). The AC has eight Arctic member states – Canada, Denmark (including Greenland and the Faeroes), Finland, Iceland, Norway, Russia, Sweden, and the USA – and includes so-called 'Permanent Participants,' major indigenous organizations from these countries. It is notable that many of these indigenous organizations are transnational (e.g., the Saami Council or the Inuit Circumpolar Council), that is they represent people who ended up in different nation states as a result of colonial border policies.

The fact that most indigenous Arctic communities have come a long way since the days of colonial rule and outright discrimination should not blind us to persistent problems. Indeed, most communities face a plethora of cultural and social ills, which are enhanced by precarious economic and ecological conditions (see Smith and McCarter, 1997; AHDR, 2004). While many of these problems can be attributed directly to the impact of colonization, the option of isolation from global forces has long since disappeared (if it ever existed). Contemporary Arctic indigenous peoples have

understood that the challenge is not to choose between 'Western' modernity and 'unchanging tradition,' but to find a livable combination of the two. Given the political sophistication of local communities, to work in the North has become a tremendously rewarding learning experience for anthropologists and other social scientists. No longer content with being mere objects of study but, at the same time, realizing the potential benefits of social science, Arctic communities are engaged in an ongoing process of defining collaborative and mutually beneficial research.

Climate change and the wealth of untapped nonrenewable resources in the Arctic make the circumpolar North a region of heightened geopolitical interest. Speculations about the shipping consequences of diminished ice in the Arctic Ocean abound (see, e.g., Anderson, 2009; Smith, 2010). Despite significant increases in Arctic shipping in recent years, future marine traffic in the Arctic Ocean is not just a function of sea ice retreat but of economic demand as well (AMSA, 2009). Awareness of the potential and realized impacts of climate differs significantly in various regions of the Arctic. While local (Inuit) communities in Alaska and Canada have been very engaged in climate change discussions (Marino and Schweitzer, 2009), some communities in the Russian only recently became familiar with the notion of global climate change (Crate, 2008). Drastic consequences in some areas, including coastal erosion necessitating community relocation (Marino, 2012), are counteracted by local perceptions that there are more important things to worry about (unemployment, language loss, high suicide rates, etc.) elsewhere. Likewise, the prospect of increased energy production is met with opposition (see, e.g., Chapman, 2013), as well as with support on the local level. A particular example of the latter is Greenland, where oil exploration is frantic and the parliament recently (October 2013) revoked a law prohibiting uranium mining. This seems to be the price of moving toward the "world's first independent Inuit state" (Nuttall, 2008). While media attention remains fixed on energy riches and territorial claims by nation states, sociocultural processes will decide the viability and future of Arctic communities.

See also: Americas, Sociocultural Overviews: North America, Native Americans; Colonialism, Anthropology of; Europe: Sociocultural Overview; Hunting and Gathering Societies: Anthropology; Indigeneity; Shamanism.

Bibliography

ACIA, 2005. Arctic Climate Impact Assessment. Cambridge University Press, Cambridge, MA.

AHDR, 2004. Arctic Human Development Report. Stefansson Arctic Institute, Akureyri.

AMSA, 2009. Arctic Marine Shipping Assessment 2009 Report. Arctic Council.

Anderson, A., 2009. After the Ice: Life, Death, and Geopolitics in the New Arctic. HarperCollins, New York.

Anderson, D.G., Nuttall, M. (Eds.), 2004. Cultivating Arctic Landscapes: Knowing and Managing Animals in the Circumpolar North. Berghahn Books, New York.

ASI, 2010. Arctic Social Indicators. Nordic Council of Ministers, Copenhagen.

Beach, H., 1990. Comparative systems of reindeer herding. In: Galaty, J.G., Johnson, D.L. (Eds.), The World of Pastoralism: Herding Systems in Comparative Perspective. Guildford Press, New York, pp. 255–298.

Berg, G. (Ed.), 1973. Circumpolar Problems: Habitat, Economy, and Social Relations in the Arctic. Pergamon Press, Oxford, UK.

Bodenhorn, B., 2000. 'He used to be my relative': exploring the bases of relatedness among Iñupiat of northern Alaska. In: Carsten, J. (Ed.), Cultures of Relatedness: New Approaches to the Study of Kinship. Cambridge University Press, Cambridge, pp. 128–148.

Bogoras, W.G., 1929. Elements of the culture of the circumpolar zone. American Anthropologist (n.s.) 31, 579–601.

Brightman, M., Grotti, V.E., Ulturgasheva, O. (Eds.), 2012. Animism in Rainforest and Tundra: Personhood, Animals, Plants and Things in Contemporary Amazonia and Siberia. Berghahn Books, New York.

Burch Jr., E.S., 1975. Eskimo Kinsmen: Changing Family Relationships in Northwest Alaska. West Publishing, St. Paul, MN.

Chapman, C., 2013. Multinatural resources: ontologies of energy and the politics of inevitability in Alaska. In: Strauss, S., Russ, S., Loue, T. (Eds.), Cultures of Energy: Power, Practices, Technologies. Left Coast Press, Walnut Creek, CA, pp. 96–109.

Crate, S.A., 2008. Gone the bull of winter? Grappling with the cultural implications of and anthropology's role(s) in global climate change. Current Anthropology 49, 569–595.

Csoba DeHass, M., 2009. Sugpiaq Russian Orthodoxy – conceptual analogy in religious syncretism in Nanwalek, Alaska (Ph.D. dissertation). University of Alaska Fairbanks.

Gjessing, G., 1960. Circumpolar social systems. In: Larsen, H. (Ed.), The Circumpolar Conference in Copenhagen 1958. Ejnar Munksgaard, Copenhagen, Denmark, pp. 75–81.

Graburn, N.H.H., Strong, B.S., 1973. Circumpolar Peoples: An Anthropological Perspective. Goodyear Publishing, Pacific Palisades, CA.

Guemple, L. (Ed.), 1972. Alliance in Eskimo Society. University of Washington Press, Seattle.

Hallowell, A.I., 1926. Bear ceremonialism in the northern hemisphere. American Anthropologist (n.s.) 28, 1–175.

Hatt, G., 1914. Arktiske Skinddragter I Eurasien Og Amerika. En Etnografisk Studie. J H Schultz, Copenhagen, Denmark [Transl., 1969. Arctic skin clothing in Eurasia and America: an ethnographic study. Arctic Anthropology 5, 3–132].

Hoffecker, J.F., 2005. A Prehistory of the North: Human Settlement of the Higher Latitudes. Rutgers University Press, New Brunswick, NJ.

Hoppál, M., Pentikáinen, J. (Eds.), 1992. Northern Religions and Shamanism. Hungary and Helsinki, Budapest (Akadémiai Kiadó and Finnish Literature Society, Finland).

Huntington, H.P., Goodstein, E., Euskirchen, E., 2012. Towards a tipping point in responding to change: rising costs, fewer options for arctic and global societies. AMBIO 41, 66–74.

Irimoto, T., Yamada, T. (Eds.), 1994. Circumpolar Religion and Ecology: An Anthropology of the North. University of Tokyo Press, Tokyo.

Kan, S., 1999. Memory Eternal: Tlingit Culture and Russian Orthodox Christianity through Two Centuries. University of Washington Press, Seattle, WA.

Keskitalo, E.C.H., 2008. Climate Change and Globalization in the Arctic: An Integrated Approach to Vulnerability Assessment. Earthscan, London.

Krupnik, I., Jolly, D. (Eds.), 2002. The Earth Is Faster Now: Indigenous Observations of Arctic Environmental Change. Arctic Research Consortium of the United States, Fairbanks, AK.

Krupnik, I., Aporta, C., Gearheard, S., Laidler, G.J., Holm, L.K. (Eds.), 2010. SIKU: Knowing Our Ice. Documenting Inuit Sea Ice Knowledge and Use. Springer, Dordrecht, The Netherlands.

Magdanz, J., Utermohle, C.J., Wolfe, R.J., 2002. The Production and Distribution of Wild Food in Wales and Deering, Alaska. Alaska Department of Fish and Game, Division of Subsistence and Western Arctic National Parklands, National Park Service, Juneau, AK and Kotzebue, AK.

Marino, E., 2012. The long history of environmental migration: assessing vulnerability construction and obstacles to successful relocation in Shishmaref, Alaska. Global Environmental Change 22, 374–381.

Marino, E., Schweitzer, P., 2009. Talking and not talking about climate change in northwestern Alaska. In: Crate, S.A., Nuttall, M. (Eds.), Anthropology and Climate Change: From Encounters to Actions. Left Coast Press, Walnut Creek, CA, pp. 209–217.

Minority Rights Group (Ed.), 1994. Polar Peoples: Self-determination and Development. Minority Rights Publications, London.

Mousalimas, S.A., 1995. The Transition from Shamanism to Russian Orthodoxy in Alaska. Berghahn Books, Providence, RI.

Nuttall, M., 1998. Protecting the Arctic: Indigenous Peoples and Cultural Survival. Harwood Academic Publishers, Amsterdam.

Nuttall, M., 2000. Choosing kin: sharing and subsistence in a Greenlandic hunting community. In: Schweitzer, P.P. (Ed.), Dividends of Kinship: Meanings and Uses of Social Relatedness. Routledge, London, pp. 33–60.

Nuttall, M., 2008. Self-rule in Greenland: toward the world's first independent Inuit state? Indigenous Affairs 3–4, 64–70.

Nuttall, M., 2012. Tipping points and the human world: living with change and thinking about the future. AMBIO 41, 96–105.

Nuttall, M. (Ed.), 2005. Encyclopedia of the Arctic, 3 vols. Routledge, New York.

Paulson, I., Hultkrantz, Å., Jettmar, K., 1962. Die Religionen Nordeurasiens und der amerikanischen Arktis [The religions of northern Eurasia and the American arctic]. W. Kohlhammer Verlag, Stuttgart, Germany.

Pehrson, R.N., 1964. The Bilateral Network of Social Relations in Könkämä Lapp District. Utgitt av Norsk Folkemuseum, Universitetsforlaget. No place.

Petersen, R., 2003. Settlements, Kinship and Hunting Grounds in Traditional Greenland: A Comparative Study of Local Experiences from Uppernavik and Ammassalik. Danish Polar Center, Copenhagen.

Petrivelli, P.J., 2004. Unangam Aleut social system. Arctic Anthropology 41, 126–139.

Reedy-Maschner, K.L., 2010. Aleut Identities: Tradition and Modernity in an Indigenous Fishery. McGill-Queen's University Press, Montreal.

Schweitzer, P.P., 2003. Levels of inequality in the north pacific rim: cultural logics and regional interaction. In: Habu, J., Savelle, J.M., Koyama, S., Hongo, H. (Eds.), Hunter-gatherers of the North Pacific Rim. National Museum of Ethnology, Osaka, Japan, pp. 83–101.

Shephard, R.J., Rode, A., 1996. The Health Consequences of 'modernization': Evidence from Circumpolar Peoples. Cambridge University Press, Cambridge.

Smith, E.A., McCarter, J. (Eds.), 1997. Contested Arctic: Indigenous Peoples, Industrial States, and the Circumpolar Environment. University of Washington Press, Seattle, WA.

Smith, L.C., 2010. The World in 2050: Four Forces Shaping Civilization's Northern Future. Dutton, New York.

Ventsel, A., 2005. Reindeer, Rodina and Reciprocity: Kinship and Property Relations in a Siberian Village. Münster: Lit Verlag.

Willerslev, R., 2007. Soul Hunters: Hunting, Animism and Personhood Among the Siberian Yukaghirs. University of California Press, Berkeley, CA.

Area and International Studies: Cultural Studies

Benjamin Lee, Rice University, Houston, TX, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 2, pp. 657–662, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

Area studies have been one field in American higher education where an in-depth understanding of languages and cultures of other societies has been nurtured. Its institutional development since World War II was the product of financial pressures in funding from the government and foundations. Its intellectual development was the product of changing views of language, discourse, and culture that came to be known as the 'linguistic turn' in the humanities and social sciences. The linguistic turn provided intellectual resources for transforming area studies as it reacted to feminism, ethnic studies, cultural studies, and globalization. The challenge of globalization has revealed some inadequacies in area and cultural studies' uses of a linguistically based concept of culture that relegated culture to shared local differences in the face of macroprocesses of economic globalization. An alternative is to link culture with processes of circulation. In the social sciences and humanities, both the objects of study and their investigators are increasingly in motion; coordinating this 'double circulation' of communities of research and what they study is a central challenge for contemporary area and cultural studies.

Historical Background of Culture and Area Studies

In the twenty-first century, 'area studies' finds itself at familiar crossroads between funding pressures and new intellectual issues. The development of area studies has always relied upon a mix of foundation and federal support. In the early part of the twentieth century, the establishment of the great philanthropic institutions such as the Carnegie Foundation for the Advancement of Teaching (1905), the Rockefeller Foundation (1913), and then the Ford Foundation (1936) provided the initial financial base for supporting research and teaching about other parts of the world, often as an adjunct to the foundations' own activities abroad. Particularly important was the Rockefeller Foundation, whose experiences in China led to its establishment of a humanities program that focused on modern languages and area studies, including the co-funding of an East Asian studies program with the Carnegie Foundation and the American Council of Learned Societies. These activities paralleled the growing public interest in China and Japan; the first area studies organization was the Far Eastern Association (now the Association for Asian Studies), founded in 1943.

World War II brought government support for area studies for explicitly strategic and intelligence gathering purposes. The Far Eastern Association was created under the specter of World War II, and the prospects of a communist China. The containment of communism required understanding the enemy, and it was in this light that the centers of Russian studies at Columbia and Harvard were established with a mix of government (including 'laundered' CIA funding) and foundation support (Cummings, 1997). The Ford Foundation soon became the major private support for area studies. From 1951 to 1966 its International Training and Research Program granted US$270 million to US universities and other institutions, to build research and training programs in area studies. In addition, from 1950 to 1996, it also provided US $87 million to the Social Science Research Council (SSRC) and the American Council for Learned Societies (ACLS) for joint area studies programs.

The key event that changed the public and government perception of area studies was the Soviet launch of the unmanned satellite Sputnik in 1957. The United States was immediately seen as having fallen behind the Soviet Union at every level of education, and by autumn 1958 Congress was set to pass the National Defense Education Act (NDEA). Under Title VI of this act, funds were set aside 'to teach modern foreign languages if such instruction is not readily available'; in addition to language instruction, the government would provide half of the cost of 'National Resource Centers' devoted to the study of specific languages and areas. From an initial annual appropriation of US$500 000, Title VI would grow to US $14 million in 1966. Despite the failure of Congress in 1966 to pass an International Education Act that would have provided additional support for area studies, Title VI and Fulbright-Hays funding continued to support area research and training, including academic infrastructure and library resources. By 1999, there were 119 Title VI centers divided across 11 areas of the world, forming the backbone for area studies.

By the 1990s, 'area studies' was in crisis. Foundation funding had leveled off, enrollments were declining, and the end of the Cold War eliminated one motivation that played so important a role in federal funding: containing communism. At the same time, the intellectual configurations that nurtured area studies were changing. In the early 1990s, the fall of the Soviet Union and a newly emancipated Eastern Europe revitalized interest in civil society and even international civil society, but the initial optimism soon faded with the resurgence of embittered nationalisms. Further, democratization and economic reforms enhanced the status in policy circles of economics and political scientists who wielded more universal models (sometimes criticized as 'one-size-fits-all'), usually based on quantitative analyses of large data sets rather than institutional analyses grounded in intimate knowledge of local histories, languages, and cultures. This approach, embracing

universal theories over those featuring historical and local particularities, challenged the very legitimacy of area studies in the 1990s. (Problems and even failures of democratization and market building came to haunt these more generalized theories and brought back area studies, but only by the end of the 1990s.) Finally, another term began to dominate discussions of area studies: globalization. The challenge of globalization was soon reflected in the proposed reorganization in 1996 of the joint ACLS and SSRC international program that funded much area research and new SSRC policies reflecting this new stress: "Now free from the bipolar perspective of the Cold War and increasingly aware of the multiple migrations and intersections of people, ideas, institution, technologies and commodities, scholars are confronting the inadequacy of conventional notions of world 'areas' as bounded systems of social relations and cultural categories" (Prewitt, 1996, p. 1).

Almost at the same time, the Ford Foundation initiated its 'Crossing Borders: Revitalizing Area Studies' initiative, in which it described area studies as being "at a significant and somewhat tumultuous turning point in its history as it attempts to respond to and illuminate dramatic changes in the world in recent decades and to understand complex relationships between the 'local' and the 'global'." The program would provide US$25 million dollars over a 6-year period in a major effort to respond to the changing conditions that area studies found itself in as it approached the new millennium. By reaching out beyond established area research centers to include colleges and smaller universities, the pool of over 200 proposals provided a snapshot of what were the leading edges of area and international studies in the United States, as they confronted an age of globalization. The initial wave of project proposals took various approaches to rethinking area studies: for example, examining the geographies of diaspora communities, challenging American exceptionalism, exploring legacies of authoritarianism, studying social movements, and alternative modernities.

The intellectual motivation behind many of the proposals was rethinking the very notion of culture and its ties to explicitly national frameworks, as revealed by such new topics as diaspora, ethnicity, and transnational phenomena. These interests contrasted sharply with those in international relations. There were no proposals that dealt with the two themes of civilizational perspectives on culture and globalization as economic development, which were popular in political science at the time (driven in part by the work of Samuel Huntington). Just as area and cultural studies began to 'deconstruct' notions of shared, national cultures, international relations, and development economics combined Max Weber's Protestant Ethic thesis with a comparative civilizational framework from anthropologists such as Robert Redfield, and made civilization-based cultural differences the causal forces behind development (in the West) and underdevelopment (everywhere else) – reflecting the previous (and later defunct) approach to culture and development (or underdevelopment) of the modernization theory of 1950s and 1960s.

The origins of these different approaches to globalization lie in the historical development of the term 'culture.' One important input into area studies has always been the study of languages, especially those of cultures that had their own written histories and literature. These language-based approaches focused on the historical examination of texts, and required both philological and hermeneutic methodologies taken primarily from the humanities; the result was a close tie both in theory and practice between notions of language and those of culture. These concerns often clashed with social sciences-based interests in contemporary societies that focused on strategic issues and economic development, in which language played basically a utilitarian role and was never the source of theoretical insight. An uneasy truce emerged in which area departments were placed in the humanities, while area centers were interdisciplinary and more social science oriented. History departments often acted as a bridge, since they could be located either in humanities or in social science divisions.

One result is that department-based language and civilization introductions to world cultures have often been part of the undergraduate curriculum in many colleges and universities as a part of 'general education.' With the support of federal (NDEA Title VI) and foundation funding, area programs, such as those at Columbia, Chicago, Berkeley, and Harvard universities, were typically built around a large library collection, and tried to create 'vertical' depth in language instruction, history, and literary studies. The area committees of ACLS and SSRC were staffed by area experts specializing in particular parts of the world (Eastern Europe, Southeast Asia, East Asia, etc.) and funded research on these regions. The emphasis was not on international relations issues, but more area-specific work that required in-depth language, culture, and historical training. Area research, thus, tended to draw theoretical and methodological frameworks from the various disciplines they were historically connected to; theoretical advances made first in the disciplines were then applied to area research. The distinctiveness and autonomy of area studies lay not in its theoretical advances, but in language teaching and the translation and preservation of specialized texts.

On the other hand, international relations programs, especially those affiliated with professional schools such as Woodrow Wilson at Princeton, SAIS at Johns Hopkins, the Kennedy School at Harvard, and Fletcher at Tufts, focus on the economics and politics of nation-states. Professional schools usually exist independently of the intellectual agenda of the university, especially the undergraduate curriculum, and, therefore, have had little influence on the debates about general education. However, their impact on undergraduate curriculum has been accelerating, as many universities have made 'internationalizing' their institutions a top priority for both research and teaching; further, such programs tend to have a more 'practical' aura to them (much as business schools do), which has an increasing appeal to students and their parents as higher education has become more expensive and labor markets have become more competitive. International studies majors are proliferating, and often require some mix of economics and area-specific courses, leading some institutions to see international studies as a way to catalyze multidisciplinary work.

New Approaches to Cultural Studies

The close links between traditional area studies and its place within the humanities explains the influence of what has been

termed the 'linguistic turn,' in which problems of meaning, interpretation, and discourse became central to much of the humanities and social sciences. The linguistic turn was triggered by the development of structuralism after World War II. In 1969, Claude Levi-Strauss published *The Elementary Structures of Kinship*, which applied the structural linguistics of Ferdinand de Saussure and the Prague School to the analysis of kinship and exchange relationships. The result was what might be called a 'paradigm' shift that would have a profound influence on French humanities and social sciences, influencing several generations of thinkers that included Roland Barthes, Michel Foucault, Jacques Lacan, Louis Althusser, Jacques Derrida, and Pierre Bourdieu. It would serve as a lightning rod for the development of feminist, ethnic, and cultural studies in the 1970s and 1980s.

Structuralism arrived in the United States in the mid-1960s, and in 1966, there was a famous colloquium at Johns Hopkins on 'The Languages of Criticism and the Sciences of Man,' in which Derrida, Barthes, Lacan, Gerard Genette, Jean-Pierre Vernant, Lucien Goldmann, Tzvetan Todorov, and Nicholas Ruwet participated. What was impressive was not only the caliber of thought, but its range, including literary criticism, psychoanalysis, history, philosophy, semiotics, and linguistics. Although united underneath a structuralist banner, the conference also introduced poststructuralist thought to US audiences. Poststructuralism would make its initial beach-head at Johns Hopkin's University and would soon conquer comparative literature departments such as Cornell and Yale where the so-called 'gang of four' of Paul DeMan, James Miller, Geoffrey Hartmann, and Harold Bloom would consolidate what came to be known as 'deconstructionism.'

While structuralism had little direct effect on area studies, the focus on language and meaning would have a huge impact across many disciplines. By the late 1960s, the triumvirate of Noam Chomsky, Jean Piaget, and Levi-Strauss promised to uncover the 'deep structures' of the human mind and society. Analytic philosophy focused on the 'ordinary language philosophy' of John Austin, Ludwig Wittgenstein, and the more formalist work of Quine, Kripke, and Davidson. Besides French-based structuralism and poststructuralism, the more German hermeneutic tradition represented by Hans-Georg Gadamer combined with the Weberian tradition of *verstehen* analysis to influence thinkers as diverse as Clifford Geertz and Jurgen Habermas. The result was that at some of the leading 'area studies' research universities, such as the universities of Chicago and Columbia, language and discourse became models for cultural analysis. The rise of symbolic anthropology at the University of Chicago quickly set off an 'interpretive turn' associated with such figures as Marshall Sahlins, Victor Turner, David Schneider, Stanley Tambiah, and Clifford Geertz; all except Sahlins would move elsewhere, thereby transforming the field's core from social to cultural anthropology.

At Columbia, the presence of Edward Said and Gayatri Spivak would pave the way for what would become known as postcolonial studies. At Berkeley, Stephen Greenblatt would develop what became known as the 'new historicism,' which was itself heavily influenced by the mix of history and anthropology developed by Geertz. The interest in meaning and discourse would quickly spread from anthropology and literary studies into area studies as part of what became more generally known as the 'linguistic turn,' a term coined by the philosopher Richard Rorty (1967) and applied first to analytic philosophy. The term was quickly applied to all language-based inquiries, as indicated by Rorty's own work on deconstructionism and Derrida. At the same time, there was a dramatic increase in the number of foreign graduate students, especially from South Asia, East Asia, and Latin America whose interests in contemporary cultural developments in their home countries contributed to the growing interest in popular culture triggered by cultural studies. The relaxation of immigration restrictions in the 1960s also increased their presence among the US faculty.

The introduction of these language and discourse themes in the late 1960s and early 1970s would intersect with the development of new social movements inspired by the civil rights movement and the events of 1968. The rise of feminism and ethnic studies (especially on the West Coast where there were degree-granting programs in ethnic studies) catalyzed research into alternative, nonmainstream histories. Poststructuralist formulations of subjectivity and identity, along with neo-Marxist analyses of culture provided not only new theoretical frameworks, but also new areas of study that challenged traditional area studies. British cultural studies introduced a theoretically sophisticated discussion of the mass media and popular culture, including Althusser's reworking of Marx and Lacanian film analysis. Cultural studies entered the United States through communications programs, most notably at the University of Illinois at Champaign–Urbana. Because of its origins in England (the so-called Birmingham school) and its location in communications studies in the United States, it initially had no impact on area studies or anthropology. However, as it became increasingly apparent (most dramatically by the events of 1989) that popular culture and the mass media were playing a crucial role in the development of societies that were the traditional object of area studies, its influence quickly spread to the point where anthropologists began complaining about how cultural studies were invading their domain and control of the concept of culture.

Even approaches that were not structuralist or poststructuralist contributed to the linguistic turn. The Frankfurt School critiques of mass culture were updated by Jurgen Habermas' work on the public sphere (Habermas, 1989; Calhoun, 1992) and then his explicitly linguistically based theory of communicative action (1984, 1987) and his critiques of poststructuralism. His ideas would introduce a distinctly communications-oriented dimension to the civil society discourses that became popular after the downfall of communism. Benedict Anderson (1983) would draw upon Walter Benjamin's work for his provocative theory of the communicative origins of nationalism, while Charles Taylor's hermeneutic explorations into the origins of Western concepts of the self (1989) invited comparative work.

The development of these trends paralleled the expansion of area, ethnic, and cultural studies in the 1970s and 1980s. Yet, if the events of 1968 could be said to roughly herald in the linguistic turn in the United States, the events of 1989 would prove to be the beginning of its demise. The downfall of communism began the undoing of not only Marxist discourses, but also the relevance of the more politicized domestic versions

of multiculturalism and identity politics for global comparisons. The civil society discourse that seemed to hold such promise at the end of the 1980s would quickly unravel in the 1990s, as the uneven development of Eastern Europe and Russia, the clampdowns in China, and the rise of new nationalisms and fundamentalisms questioned the viability of the civil society concept in an age of globalization.

Whereas in the 1970s and 1980s, theory was very much 'at the cutting edge' of the linguistic turn, by the mid-1990s, 'globalization' had replaced it, not so much as a theoretical category, but rather as an empirical challenge to the standard disciplines and area studies. The transition from the linguistic turn to globalization formed the intellectual backdrop to the Ford Foundation's Crossing Borders Initiative. The over 200 proposals represented the state-of-the-art in area studies, as it attempted to deal with contemporary global transformations. While there was a large variety in theoretical approaches, none emerged as predominant. Instead, most striking was the range of phenomena that the proposals tried to deal with, from opening up ethnic studies and globalizing identity politics, to reconfiguring notions of areas and regions (e.g., thinking of oceans as areas). One lacuna was particularly noticeable and undoubtedly fed into the crisis feeling about area and cultural studies. Despite the calls for interdisciplinarity, there were no proposals that dealt substantively with the relation between economics and culture.

This problem was not unrelated to the linguistic turn in cultural analysis. Despite some innovative works by the literary scholar Mary Poovey (1998) and the philosopher Ian Hacking (1970, 1975), interpretive approaches developed for and applied to literary and historical texts have not been easily transferred to the analysis of mathematical and statistical data. The inability to treat these discourses as cultural phenomena has heightened the split between economic and cultural analyses, just at the moment when historical economics was being eliminated in many economics departments. This gap was exacerbated by the collapse of Marxism, which in the hands of the Frankfurt School and contemporary thinkers such as David Harvey (1982, 1989) and Frederic Jameson (1998), had at least tried to keep the analysis of the relations between culture and capital alive.

Cultural Studies and Globalization

The textual focus also overlooked dynamics of circulatory processes in which texts are embedded and transmitted. Text-oriented approaches often assumed the fixity of the text artifact and such categories as author, reader, audience, and the act of reading itself; even destabilizing moves of deconstructionism presupposed elitist conceptions of reading and re-reading. The linguistic turn's focus on the interpretation of texts, and discourses and the analytic frameworks derived, produced a paradox: the more approaches treated texts or discourses as intrinsic objects of analysis, the more difficult it was to understand the norms of the interpretive communities that made such approaches possible, that is, the context that makes interpretation possible. A focus on texts and discourses per se overlooks how they serve as ways of connecting individuals and groups and how these connections create various kinds of interpretive communities. Yet, if contemporary globalization involves the creation of new forms of connectedness through a globalized market, mass media, and new technologies, then it is not surprising that the linguistic turn has become increasingly anachronistic as it confronts these phenomena.

Globalization presents a unique challenge to area and cultural studies. A focus on language and identity leaves economic processes unanalyzed. If the seeming triumph of capitalism indicates that the leading edge of contemporary changes is the spread of homogenizing economic processes, then analyses that focus on the cultural dimensions of identity formation seem doomed to play catch-up with forces beyond their control. The civilizational approaches developing in international relations invoke cultural values to explain the failure to develop, thereby presupposing that economic success depends on adopting what are basically Western cultural values; a holistic conception of shared culture as civilization-based values provides the diacritics for developmental differences. Given the historical trajectory of area and cultural studies, it seems highly unlikely that they can go back to a conception of culture that they have been abandoning over the last several decades.

Instead, there is a growing realization in area and cultural studies that while globalization has produced many convergences in institutions and practices across societies – market economies, bureaucratic states, parliaments, elections – these differ significantly from society. Unlike the rhetoric of economic development, globalization is not a single process spreading across the world. Instead, it spreads through diffusion, influence, imitation, negative reaction, borrowing, and so on that work across complex circulations of goods, technologies, ideas, images, and forms of collective imagination. At the same time, the velocity, scale, and form of these processes and circulations challenge virtually all existing narratives of culture, place, and identity and the intellectual and academic frameworks used to study them.

The complexity of these global processes suggests that if circulation is to be a useful analytic construct that replaces more traditional notions of shared culture, it must be more than simply the movement of people, ideas, and commodities from one culture to another. Instead, recent work suggests that circulation is a cultural process with its own forms of abstraction, evaluation, and constraints that are created by the interactions between specific types of circulating forms and the interpretive communities built around them; the circulation and interpretation of specific cultural forms creates new ways of connecting individuals and groups that can be the bases for social movements, identity formations, and communities. A much cited example would be Benedict Anderson's claim that novels and newspapers help people imagine forms of connectedness among strangers that are at the heart of the idea of nationalism. However, similar themes run through Habermas' work on the public sphere, the Chartier group on print mediation (Chartier, 1989, 1994; Martin, 1994), and Arjun Appadurai's on global flows and 'scapes' (1996).

The linking of circulation with the construction of different types of imagined and interpretive communities also provides a way of overcoming some of the barriers between cultural and economic approaches to globalization. In much of the

contemporary work on globalization, culture is seen as that which is subjectively shared by a community and is, therefore, local. Economic processes are seen as objective, universal, and global. Often overlooked, however, is the interpretive dimension of economic processes. Global capital flows presuppose the intertranslatability of financial instruments and information technologies while at the same time demanding that 'local' economic activities be translated into cross-culturally comparable statistical categories. The circulation of financial instruments rests upon this work of translation and depends upon an interpretive community of institutions (e.g., exchanges, banks, and clearing houses) that understand and use these forms.

What these phenomena share is an internal connection between circulating forms and the forms of connection and community built around them. Communities are built around the ways they connect individuals and groups. The ways they are connected depend upon the circulation of specific cultural forms that both enable and constrain how they are to be used and interpreted. Whether it is the novels, newspapers, magazines, publishing houses, and coffee houses and salons of Anderson's and Habermas' accounts of nationalism and the public sphere, financial derivatives or junk bonds, or the architecture and codes of the Internet, interpretation and circulation interact to create new forms of collective subjectivity. Instead of simply being the movement of people, ideas, and commodities from one culture to another, circulation is a cultural process with its own forms of abstraction, reification, and constraint. The circulation of a 'community' of forms creates new forms of community with hierarchies of evaluation, contrast, and difference; what might be called a 'culture of circulation.' A culture of circulation would be a 'translocal' form of connectedness that might have the geographical reach of a civilization, but not its fixity or sharedness. It is from these cultures of circulation that new forms of subjectivity and consciousness arise that are the bases for the political cultures of modernity, including the public sphere and nationalism.

Arjun Appadurai (1999) has suggested that research should be thought of as an imagined community that spreads through the global circulations of people and ideas as well as institutions and practices. The peculiarity of area studies is not that it is an interpretive community within the larger cultural circulations that characterize modern research; that would be true of any research community. Instead, its potential as a form of knowledge lies in its self-reflexivity. Its objects of study are cultures of circulation of which it is a part. Both the objects of study and those who study them are increasingly in motion. This 'double circulation' is already creating new communities of research that challenge traditional models of area study by decentering the Euro-American Academy as the main source of innovative research; postcolonial studies may be the welcome harbinger of things to come. These confrontations are not simply producing competing hypotheses that can be tested for their validity, but are rather the indices of new research imaginaries in the making. A self-reflexive area studies, instead of just providing the data to test theories generated in other disciplines, would be at the cutting edge of understanding the conditions for the production of knowledge in an age of globalization.

See also: British Cultural Studies; Critical Theory: Contemporary; Deconstruction; Globalization and World Culture; Language and Society; Linguistic Turn; Postmodernism: Philosophical Aspects; Structuralism.

Bibliography

Anderson, B., 1983. Imagined Communities. Verso Press, London.
Appadurai, A., 1996. Modernity at Large. University of Minnesota Press, Minneapolis, MN.
Appadurai, A., June 1999. Globalization and the research imagination. International Social Science Journal 51 (160), 229–238.
Calhoun, C. (Ed.), 1992. Habermas and the Public Sphere. MIT Press, Cambridge, MA.
Chartier, R., 1989. The Cultural Uses of Print in Early Modern France. Princeton University Press, Princeton, NJ.
Chartier, R., 1994. The Order of Books. Stanford University Press, Stanford, CA.
Cummings, B., 1997. Boundary displacement: area studies and international studies during and after the Cold War. Bulletin of Concerned Asian Scholars 29, 6–26. January–March.
Habermas, J., 1984, 1987. Theorie des kommunikativen Handelns (The Theory of Communicative Action). Polity Press, Cambridge, UK.
Habermas, J., 1989. Strukturwandel der Öffentlichkeit: Untersuchungen zu Einer Kategorie de Bügerlichen Gesellschaft (The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society). MIT Press, Cambridge, MA.
Harvey, D., 1982. Limits to Capital. Blackwell, London.
Harvey, D., 1989. The Condition of Post-modernity. Blackwell, Oxford, UK.
Hacking, I., 1970. The Taming of Chance. Cambridge University Press, Cambridge, UK.
Hacking, I., 1975. The Emergence of Probability. Cambridge University Press, Cambridge, UK.
Harrison, L., Huntington, S. (Eds.), 2000. Culture Matters: How Values Shape Human Progress. Basic Books, New York.
Jameson, F., 1991. Postmodernism, or, the Cultural Logic of Late Capitalism. Verso Press, London.
Jameson, F., 1998. The Cultural Turn. Verso Press, London.
Levi-Strauss, C., 1969. Les Structures Élémentaires de la Parenté (The Elementary Structures of Kinship). Beacon Press, Boston.
Martin, H.-J., 1994. The History and Power of Writing. University of Chicago Press, Chicago.
Poovey, M., 1998. The Invention of the Modern Fact. University of Chicago Press, Chicago.
Prewitt, K., 1996. Presidential items. ITEMS 50 (2–3), 1–9.
Rorty, R., 1967. The Linguistic Turn: Recent Essays in Philosophical Method. University of Chicago Press, Chicago.
Taylor, C., 1989. Sources of the Self: The Making of the Modern Identity. Harvard University Press, Cambridge, MA.

Area and International Studies: Political Economy

Ian Robinson, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 2, pp. 719–723, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Modern 'political economy' explores relationships among economic and political organizations, institutions, policies, and outcomes. There has never been a consensus on the theories and methods that should structure the exploration of these relationships and the debate over how best to understand these matters has evolved over time. The interpretation offered here simplifies both of these dimensions of difference. It distinguishes two periods of postwar history – the Fordist and Neoliberal eras – based on the dominant paradigm of economic organization prevailing in the rich capitalist democracies of the global North. For each period, the most important theoretical approaches to political economy, the regions of the world to which they were applied, and the social science departments in which they were based are identified. It is argued that, in the second or neoliberal era, political economy analyses can be broadly grouped into those that took a critical stance on existing economic paradigms, institutions, and policies, and those which accepted the core assumptions of neoclassical economics and the neoliberal policies and institutions that this approach to economics has legitimated. The relationship between area and international studies and political economy varies, both by period and, to some degree, by the type of political economy in question.

Modern 'political economy' explores relationships among economic and political organizations (e.g., states, corporations, and unions), institutions (e.g., laws and practices regulating trade and competition), policies (e.g., restrictions on international capital mobility), and outcomes (e.g., rates of economic growth, political regime stability). Political economists differ along several important dimensions, as discussed below. This article offers an overview of the evolution of competing versions of political economy, and their relationship to Area and International Studies, since World War II.

The Fordist Moment

The classical political economists of the late eighteenth and nineteenth centuries – Smith, Ricardo, Malthus, Marx, and J.S. Mill – addressed fundamental questions such as the appropriate economic role of the state, the implications of trade liberalization for the economic fortunes of different economic classes, the ecological constraints on continuous economic growth, and the economic and political contradictions of different modes of production, including capitalism.

In the first quarter-century after World War II, some of these questions remained at the center of debates between modernization theorists (Rostow, 1960) and dependency theorists (Cardoso and Faletto, 1979) who argued about the logics of, and possibilities for, economic and political development in what the Cold War framed as the Third World. Marxist versions of political economy became the new orthodoxy in the Second World, which soon encompassed most of Eastern Europe and much of Asia. In the First World, however, most students of market dynamics within economics departments began to abandon political economy approaches. Prior to World War II, when institutional economics remained the dominant tendency in the economics departments of the United States, this break with the core assumptions and research agendas of political economy had gone furthest in the United Kingdom. However, after the war, British neoclassical microeconomics and Keynesian macroeconomics gained ground rapidly in the United States. In the Cold War context, the dominant theoretical orientations of the US hegemon exerted a powerful gravitational pull on social science in the academies of the First World.

The new mainstream economists had a much narrower intellectual agenda. At the micro level, they drew on the pioneering work of Marshall and Pareto in an effort to demonstrate by formal, mathematical means the superiority of competitive markets as efficient allocators of resources. Arguments for trade liberalization and critiques of most forms of state intervention in market allocation processes were developed in this spirit. At the macro level, in First World economies, the new mainstream drew on Keynes in an effort to theorize how best to employ fiscal and monetary policies to reduce the amplitude of business cycle fluctuations. There were significant tensions between these micro- and macroeconomic agendas; however, they only became salient toward the end of this period.

Most of the new mainstream economists – whether micro or macro in focus – sought quasinatural laws governing market dynamics regardless of time and place. They paid little attention to the political and institutional parameters within which markets existed or to the balance of power among social forces that shaped these parameters. There was an irony here. The divorce between politics and economics in First World academic economics was possible because a new kind of political economy – sometimes called 'Fordism' (Lipietz, 1987) – was developed after World War II. The institutional and political details of Fordist regulation varied from country to country but everywhere it represented a fundamental shift away from wage regulation through competitive labor markets maintained by the repression of worker rights, to form democratic unions and engage in collective bargaining. Fordist

institutions linked national real wage growth to the expansion of national labor productivity through some combination of collective bargaining (e.g., pattern bargaining in the United States) and state regulation (e.g., minimum wages) (Piore and Sabel, 1984). Fordist regulation generated higher rates of economic growth and distributed the gains from that growth more broadly among the workforce than any form of economic regulation before or since (Marglin and Schor, 1990). These successes contributed to the depoliticization of economic policy, which could more easily be seen as an administrative matter to which there were technical answers. This facilitated the shift away from political economy's focus on the reciprocal relationship between political power and economic outcomes.

Non-Marxist versions of political economy remained a significant current within the comparative politics subdiscipline of political science in this period. In the Cold War struggle, a great deal of public research funding was made available to those pursuing such studies (Gendzier, 1985). Area studies encompassed the countries of the Second and Third Worlds, where the influence of Marxist political economic analysis was strong among state elites, academics, and organizations such as unions and cooperatives. Hence, First World academic analysts found it necessary to engage questions of class power, organization, and institutions in studying these countries. To this task most brought Weber and Durkheim, as interpreted and synthesized by Talcott Parsons, under the rubric of modernization theory (Leys, 1996).

Area studies thus helped to preserve political economy when it was marginalized in the economics departments of the First World countries in which it had originated. In turn, political economy offered a coherent basis for distinguishing among different areas. Latin America, for example, made sense as a region to be contrasted with others because most of its states shared a particular kind of political economy. In the nineteenth century, the countries in this region were characterized by primary commodity production for export, and republican regimes that secured their independence from Iberian empires. In the crisis of the 1930s, most of these countries embraced a particular kind of economic development strategy – import-substitution industrialization – that gave rise to parallel economic and political dynamics, including the rise of a significant industrial working class and the formation of corporatist political systems.

There was no parallel symbiosis between political economy and the International Relations (IR) subdiscipline as it existed in US political science in this period. IR was dominated by international security debates between those who supported a *realpolitik* firmly rooted in a narrow account of national self-interest and those who asserted that international cooperation rooted in shared liberal values was a surer guide to national security and world peace. IR paradigms, particularly as they were formalized by Waltz and his followers, took it as axiomatic that states were highly autonomous from the domestic societies in which they were embedded, at least as regards the formation of foreign policy (Waltz, 1959). On this view, the international distribution of state power resources (e.g., concentrated in two rival superpowers versus dispersed more evenly among leading countries organized into alliances), and differences in state elite strategies for realizing their power maximizing objectives were the main explanations for variations in state behavior and resulting international dynamics (Krasner, 1976). States might liberalize trade as part of their grand strategies for enhancing their power relative to their rivals, but there was little reciprocal causality in this model. That is, neither international economic dynamics nor classes defined in economic terms had much impact on state goals or strategies.

Beyond IR as practiced in the United States, International Studies in these years was roughly equivalent to diplomatic history. Greater methodological and theoretical eclecticism created more space for recognizing the significance of domestic factors in international relations. A few approached these matters from standpoints that paid greater attention to the kinds of factors highlighted by political economy. Still, most studies of international diplomacy remained in the realm of 'high politics,' and so had only passing contact with the methods and concerns of political economy.

The Neoliberal Moment

As the Fordist economic order began to disintegrate in the late 1960s, many argued that both the causes of the crisis and the remedies for it lay in changes in the balance of economic and political power among nations, between labor and capital within nations, or both (e.g., Gourevitch, 1986). Economics and economic policy were thus repoliticized as an intense political struggle over how to understand and respond to the crisis of Fordism under way. In this context, rival strands of political economy emerged, each associated with advocates of a different response to the crisis. Critical political economists – a diverse group influenced in varying degrees by Marx, Weber, and Polanyi – were concentrated in political science and sociology departments. This strand of political economy was strongest among area specialists who focused on Latin America, Asia, and Western Europe and among students of peasant rebellion and social revolution. Their diagnoses of the crisis tended to support policies that would reinforce and extend the basic principles of Fordist regulation, or a move beyond capitalism to some form of democratic socialism.

Neoclassical political economists began from the premise that neoclassical accounts of economic dynamics were basically sound, as were capitalist economies. The problem as they saw it was how to develop an equally sound science of political dynamics, a science that would explain why state intervention could seldom improve on market outcomes, even when market failures were acknowledged. Some sought to build a new political science on the 'rational choice' premise that all individuals and organizations are instrumentally rational, self-interested actors (Alt and Shepsle, 1990). Others were less programmatic, turning traditional analytic tools to the service of policy goals deriving from neoclassical economics. The first strategy generated theories of 'rentier' states and 'political failure' paralleling the theory of market failure that justified Fordist regulation on efficiency grounds (Kruger, 1974; Bates, 1988). These analyses, together with neoclassical microeconomic doctrines, provided the intellectual rationale for 'neoliberal' policy prescriptions – that is, redefinition as the primary economic role of the state as the creation and maintenance of competitive markets. The second strategy generated

(among other things) assessments of strategies for implementing structural adjustment policies successfully in democracies where popular opposition to such policies was widespread. The neoclassical strand of political economy was concentrated in economics and political science departments (particularly in the United States).

The neoclassical approach to political economy was very much in tune with the neoliberal response to the crisis of Fordism championed by the United States under Ronald Reagan and the United Kingdom under Margaret Thatcher. The Third World debt crisis soon facilitated the export of neoliberal policies, via conditions imposed on debtor nations in return for assistance in restructuring their loans. In this context, neoclassical political economy became the more prominent of the two approaches, particularly among policy elites and in the United States. However, there were important intellectual innovations within both tendencies. The result was a renaissance of political economy analysis in the neoliberal era, with both the form and the implications of that analysis intensely contested.

World systems theory was an important strand of the critical political economy analysis that emerged in this period (Wallerstein, 1976). While influenced by earlier dependency theories, world systems theory was distinctive in two ways: the degree to which it treated the international economy as a system governed by its own systems-level logic, and the degree to which that logic was seen to determine the development possibilities of the nations whose system functions marked them as peripheral or semi-peripheral. World systems theory took hold primarily in sociology departments, where Marx enjoyed more equal status with that of Durkheim and Weber.

A second strand of critical political economy, with institutional roots in sociology and political science, emerged under the banner of 'bringing the state back in' (Evans et al., 1985). This strand explored the significance of national differences in state characteristics and relations between the state and societal actors, factors that were treated as secondary in most world systems analyses. Within the Third World, critical political economists began to explore the social and political consequences of structural adjustment policies in Africa and Latin America in the wake of the debt crisis (e.g., Bierstecker, 1995). There was also increased interest in the role of 'developmental states' in enabling a small number of countries – concentrated in Asia's Newly Industrializing Countries – to escape the travails of the debt crisis and successfully reorient national economies to an export-driven model of industrialization (e.g., Evans, 1995). Finally, a literature emerged on the export processing zones created in many countries subject to neoliberal restructuring, and on the international supply chains that linked them to First World corporate producers and retailers (Gereffi and Korzeniewicz, 1994).

A strand of neoclassical political economy also focused on the world system. Most important here was the 'hegemonic stability' theory advanced by the self-styled neorealists and their neoliberal interlocutors. These analysts, located within international relations sections of political science departments, mainly in the United States, debated the extent and significance of the decline of US economic hegemony evident by the late 1960s. Hegemonic stability theories asserted that, without disproportionate US economic power, the international monetary system created at Bretton Woods and multilateral trade liberalization under the auspices of GATT (General Agreement on Tariffs and Trade) would not have been possible (Gilpin, 1987). They interpreted the collapse of fixed exchange rates and an alleged shift toward protectionism in the form of 'nontariff barriers,' as evidence that the postwar international economic regimes constructed by the United States were indeed unraveling.

Neoliberals such as Keohane drew on game theory to argue that states *qua* rational actors might choose to support and extend trade liberalization and other aspects of international regulation out of an enlightened sense of self-interest, even in the absence of a hegemon, under certain conditions (Keohane, 1984). Other neoclassical political economists took a different tack, identifying a variety of societal, institutional, and ideological factors that might explain why the US state maintained its trade-liberalizing trajectory in the 1970s and 1980s, despite declining US economic hegemony and rising social costs (e.g., Goldstein, 1993). Among neoclassical political economists focusing on the global South, attention was devoted to explaining the wave of democratization that began in the late 1970s, particularly to possible links between economic liberalization and democratization.

In the 1990s, the end of the Cold War and the acceleration of the economic globalization – by which analysts generally meant increased international trade and capital mobility, and sometimes also neoliberal policies such as privatization and deregulation – shifted the focus of research and the terms of debate. The concepts of the Second and Third Worlds were rendered obsolete; many analysts began dividing the world into the 'global North' (i.e., rich capitalist democracies) and the 'global South' (i.e., all others). An important debate developed concerning whether there was anything sufficiently novel about the international economy of the 1990s to warrant the use of the term globalization (Held et al., 1999). There was also great interest in the causes and consequences of economic globalization. Neoclassical political economists tended to take a positive view of economic globalization and often treated that the process as natural and/or inevitable. Critical political economists typically saw the shift as the product of power politics within and among nations and its negative affects as more substantial (Cox, 1994). There was great interest in whether this new economic order was significantly narrowing state policy autonomy, forcing governments toward a more *laissez-faire* model of economic organization regardless of their political stripe and the preferences of voters (e.g., Rodrik, 1997). There was also interest in the implications of economic globalization for the power of organized labor (e.g., Kitschelt et al., 1999). Finally, there was growing interest in the origins and character of organized resistance to the neoliberal model of globalization, a discussion leavened with the insights of social movement theory (e.g., Castells, 1997), as well as more traditional political economy approaches (Arrighi et al., 1989).

These developments had important implications for the evolution of area and international studies. The resurgence of political economy strongly legitimated international studies' supranational and interdisciplinary character and added another important approach to how such work might be organized. As to area studies, political economy may afford new grounds for drawing area boundaries. For example, since

the passage of the North American Free Trade Agreement (NAFTA) in 1993, Canada and Mexico have become much more integrated with the US economy. For many of the questions of interest to political economists, it now makes sense to treat North America as a region to be studied as a unit. This contrasts with the old area study practice of studying the United States in splendid isolation, Mexico as part of Latin America, and largely ignoring Canada. Similarly, as East European countries have joined the European Union (EU), it remains sensible to frame many political economy questions in terms of a new EU region that straddles what were once regions in the First and Second Worlds. This said, this last statement depends on whether the EU will survive the shocks of the 2008 global economic crisis and the subsequent crisis in the euro. While the Greek crisis raised the specter of the collapse of the euro or its reduction to a 'northern euro' (centered on Germany and northern Europe), the stabilization of the European economy (albeit without hoped-for growth, given continuing austerity measures) reduced the threat that the EU would fall apart (although as of this writing it is unclear whether the United Kingdom will retain its ties with the EU). After the shocks of 2008 threatened EU implosion, Vladimir Putin's takeover of Crimea and sponsored civil war in eastern Ukraine might be the stimulus to bring EU countries closer together. However, the years since 2008 have been a period of geopolitical instability, and at this moment it would be foolish to presume either EU expansion or collapse.

A Brief and Uncertain Postscript: Post-2008 Political Economy

This brief discussion of the EU after the global crisis of 2008 highlights a more general issue of this moment in time: that the crisis might provide a potential moment of transition or retrenchment of models of political economy. The extent to which national economies were interconnected was laid bare once the American housing bubble burst and American financial institutions suddenly faced severe challenges to their very survival. At the same time, the degree of local autonomy was also revealed in how various actors went in different directions vis-à-vis responses to recessions: for example, from more state-centered responses in China and Russia, to austerity policies in the EU, to less austerity-oriented but still market-friendly policies in Barack Obama's United States. Further, the degree to which public policies have reflected interests and ideologies deserves study that at this point in time remains to be done. As Paul Krugman argued repeatedly in his *New York Times* blog, austerity policies have been disastrous, and so the only reason they were adopted and persisted was a combination of particular elite interests and ideologies about the 'normal' way states approach political economy and economic crises. As Piketty (2014) revealed in his rigorous study, inequality has been a defining feature of modern economies, and while there are possible responses to improve equality and reduce risks of future shocks, it is unclear whether political and economic elites will accept these. For example, German politicians and public have been averse to allowing inflation to rise: while this might have a negative short-term impact on Germans themselves, it would benefit the EU, and Germany as well, in the medium to long term. Yet, in German political discourse, inflation is framed in terms of the catastrophic post–World War I economic situation, as well as morality play in which German productivity and frugality would appear to be punished for the sins of other European countries and societies that took on too much debt. As well, American elites for the most part have been willing to pursue austerity measures that ultimately have hindered the American recovery; circumstantial evidence suggests this is not only because of interests but also because relatively more Keynesian measures (such as fiscal stimuli and more equitable distribution of the tax burden) would contradict the hegemonic elite ideology in which the state is the source of economic troubles and the business elite the source of its solutions. Finally, the global economy has come under some scrutiny since 2008. While globalization still appears inevitable, actors from various social movements have asked whether that global order need appear as it does. The rise of the BRIC countries (Brazil, Russia, India, and China) presented a potential competitor to the hegemony of the United States and Europe, and while the 2008 crisis reduced BRIC growth, the idea of a new bipolar world – with power balanced between the United States and the EU on one side, and the BRICs on another – is a distinct possibility, especially as Russian leaders Vladimir Putin and Dmitrii Medvedev more than once proposed that BRIC countries plus South Africa form their own form of economic alliance to counterbalance 'the West.'

The possible rearrangement of global relations and positions, and the problematic and contentious (and sometimes seemingly puzzling) nature of public policies in response to the 2008 crisis should force us to revisit what we mean by 'political economy' and how to analyze it. If the story since 2008 is anything to go by, Karl Marx and Max Weber would be happy, as their stresses on the centrality of elites, states, and especially power and contention to any analysis seem borne out by recent events.

See also: Dependency Theory; Development and the State; Nations and Nation-States in History; Political Economy in Anthropology; Political Economy, History of; State Formation, Theory of; World Systems Theory.

Bibliography

Alt, J.E., Shepsle, K.A. (Eds.), 1990. Perspectives in Positive Political Economy. Cambridge University Press, New York.

Arrighi, G., Hopkins, T., Wallerstein, I., 1989. Antisystemic Movements. Verso, London.

Bates, R.H. (Ed.), 1988. Toward a Political Economy of Development: A Rational Choice Perspective. University of California Press, Berkeley, CA.

Bierstecker, T.J., 1995. The "triumph" of liberal economic ideas in the developing world. In: Stallings, B. (Ed.), Global Change, Regional Response: The New International Context of Development. Cambridge University Press, New York.

Cardoso, F.H., Faletto, E., 1979. Dependency and Development in Latin America. University of California Press, Berkeley, CA.

Castells, M., 1997. The Power of Identity. The Information Age: Economy, Society and Culture, vol. 2. Blackwell, Cambridge, MA.

Cox, R.W., 1994. Global restructuring: making sense of the changing international political economy. In: Stubbs, R., Underhill, G.R.D. (Eds.), Political Economy and the Changing Global Order. St. Martins Press, New York, pp. 45–59.

Evans, P., 1995. Embedded Autonomy: States, Firms, and Industrial Transformation. Princeton University Press, Princeton, NJ.

Evans, P., Rueschemeyer, D., Skocpol, T., 1985. Bringing the State Back in. Cambridge University Press, New York.
Gendzier, I., 1985. Managing Political Change: Social Scientists and the Third World. Westview Press, Boulder, CO.
Gereffi, G., Korzeniewicz, M. (Eds.), 1994. Commodity Chains and Global Capitalism. Greenwood Press, Westport, CT.
Gilpin, R., 1987. The Political Economy of International Relations. Princeton University Press, Princeton, NJ.
Goldstein, J., 1993. Ideas, Interests and American Trade Policy. Cornell University Press, Ithaca, NY.
Gourevitch, P., 1986. Politics in Hard Times: Comparative Responses to International Economic Crises. Cornell University Press, Ithaca, NY.
Held, D., McGrew, A.G., Goldblatt, D., Perraton, J., 1999. Global Transformations. Stanford University Press, Stanford, CA.
Keohane, R., 1984. After Hegemony. Princeton University Press, Princeton, NJ.
Kitschelt, H., Lange, P., Marks, G., Stephens, J.D. (Eds.), 1999. Continuity and Change in Contemporary Capitalism. Cambridge University Press, New York.
Krasner, S.D., 1976. State power and the structure of international trade. World Politics 28 (3), 317–347.
Kruger, A., 1974. The political economy of the rent-seeking society. American Economic Review 64, 291–303.
Leys, C., 1996. The Rise and Fall of Development Theory. Indiana University Press, Bloomington, IN.
Lipietz, A., 1987. Mirages and Miracles: The Crises of Global Fordism. Verso, London.
Marglin, S., Schor, J. (Eds.), 1990. The Golden Age of Capitalism: Reinterpreting the Post-War Experience. Clarendon Press, Oxford, UK.
Piketty, Thomas, 2014. Capital in the Twenty-first Century. Harvard University Press, Cambridge.
Piore, M., Sabel, C., 1984. The Second Industrial Divide: Possibilities for Prosperity. Basic Books, New York.
Rodrik, D., 1997. Has Globalization Gone Too Far? Institute for International Economics, Washington, DC.
Rostow, W.W., 1960. Stages of Economic Growth: A Non-communist Manifesto. Cambridge University Press, New York.
Wallerstein, I., 1976. The Modern World-System. Academic Press, New York.
Waltz, K.N., 1959. Man, the State, and War: A Theoretical Analysis. Columbia University Press, New York.

Area and International Studies: Stakeholders

Gilbert W Merkx, Center for International Studies, Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In most nations, the stakeholders in area and international studies (A&IS) are government agencies. However, in the United States, A&IS stakeholders are found in education and the private sector as well as in government, resulting in a high level of organizational development, complexity, and productivity. A&IS represent a crossing point between academic disciplines, research and policy institutions, numerous ethnic foreign-language communities, various international exchange organizations, foreign and domestic corporate interests, national security agencies, and issue-oriented oppositional movements. As a result, the A&IS enterprise is contested by shifting coalitions defined by such factors as professional self-interest, intellectual paradigms, national security threats, and ideological agendas.

The various A&IS stakeholder communities can be divided into three major categories: the producers of knowledge (individuals and institutions, as well as the associations that represent them), the consumers of knowledge (in the form of information or in the form of A&IS-trained personnel), and the investors (in both production and consumption).

The Producers of A&IS

The institutional locus of A&IS varies across nations. This locus is a key determinant of the complexity of the stakeholder community. In many countries, including France, the former Soviet Union, and China, A&IS research and training is located in national academies, or government-funded think tanks (Lambert, 1990: pp. 712–732). These are self-contained nonuniversity institutions whose staff and students are devoted to language training, empirical research, and policy analysis. A variant, most notably in Great Britain, is the free-standing research institute devoted to philology and historical studies of a foreign civilization. Such self-contained institutes are separate, and often distant, from institutions of higher education in their countries. As a consequence, A&IS expertise in these countries is generally not conducted with reference to the standards of the academic disciplines, but represents a nondisciplinary or multidisciplinary tradition that serves as its own point of reference. The stakeholders are basically limited to the area specialists in the national academies and their client government agencies with foreign diplomatic, economic, military, or national security concerns.

In contrast, A&IS in the United States are found primarily in institutions of higher education, particularly in the research universities. The standard organizational model consists of a coordinating center or institute that offers the less commonly taught languages, while the more commonly taught languages and relevant courses in the various academic fields are offered by the disciplinary departments. The A&IS faculty are employed by these departments and judged by them for tenure and promotion. As a result, research and teaching in A&IS are judged primarily by disciplinary standards. Faculty in A&IS programs have primary appointments in disciplinary departments, teach courses in those departments, and conduct research in those disciplines. Thus, while the A&IS programs are interdisciplinary inasmuch as they list faculty and courses from different disciplines, A&IS teaching and research are almost exclusively disciplinary in nature. A&IS graduate degrees, especially PhDs, are therefore normally conferred in a discipline. Department of Education data for all graduate degrees produced by federally funded comprehensive A&IS centers in the 1991–94 period show that 91.5% received disciplinary or professional degrees, while only 8.5% received area studies degrees, and these were largely at the MA level (Schneider, 1995a: p. 9, Table C). Even those programs that award a graduate degree in study of a foreign area do so through a curriculum based on disciplinary courses offered by departments, usually with a disciplinary major.

Because of this integral relationship to higher education, A&IS is far more developed in the United States than in other countries. In the year 2013, federal funding through Title VI of the HEA supported 127 comprehensive (graduate and undergraduate) A&IS centers at colleges and universities, as well as smaller numbers of undergraduate international studies programs, undergraduate international business programs, graduate centers for international business education and research, foreign-language resource centers, and overseas research centers. A&IS programs that do not receive federal funding are far more numerous, at least four times the number receiving aid. The total number of formally organized A&IS programs in the USA is in excess of 500.

The A&IS stakeholders in the U.S. system are far more varied than in countries with the national academy model. The faculty members involved in the academic programs are usually members of disciplinary associations (such as the American Historical Association), and members of A&IS professional associations (such as the Latin American Studies Association and the International Studies Association), which means that these associations also are stakeholders.

The membership of the A&IS professional associations are a guide to the size of the knowledge producer population. Although only about two-thirds of the area studies association's membership are faculty, not all foreign-area studies faculty belong to these associations. Past studies have therefore estimated that total membership in the area studies association provide a good approximation of faculty employment in foreign-area studies (Lambert et al., 1984: p. 13). The total membership of the five major area studies associations, which cover Africa, Asia, Latin America, the Middle East, and the

former Soviet Union and Eastern Europe, was approximately 16 000 in 1990 (NCASA, 1991). This compares with a 1979 membership in all area studies associations of about 18 000. If the 1990 memberships of the smaller area studies associations such as Brazilian, Canadian, European, and Caribbean studies are added to those of the five major associations, the total remains approximately 18 000, about the same as in 1979 (Barber and Ilchman, 1977: p. 15). These numbers have been relatively constant in the two decades since then. Membership in the International Studies Association adds another 3000 persons, for a combined total of 21 000. Another faculty stakeholder group is foreign-language and literature teachers, for whom two sources of information exist, membership in professional organizations, and surveys of the National Center for Education Statistics of the U.S. Department of Education. Using these two sources, a recent study estimated a total 36 000 postsecondary language and literature faculty (Merkx, 2000: pp. 93–100). Added to the previous groups, the estimated total of faculty stakeholders is about 47 000 persons.

Campus units that support A&IS, such as the library and the language departments, have an investment, which means that their professional organizations (such as the Association of Research Libraries or the National Council for the Less Commonly Taught Languages) are also involved. The numerous academic disciplines that contribute to A&IS programs, including foreign languages and literature, have an interest. Universities with strong A&IS programs have a stake, as do the presidential associations that represent their interests in Washington (such as the Association of American Universities (AAU)).

Campus-based components of A&IS face internal rivals on campus, however, and these rivalries may be reflected in division within their respective associations. Within key departments, such as history or political sciences, there may be struggles for resources and faculty lines with non-A&IS factions. Within the library, the A&IS collection effort competes with other collection priorities. The language departments may be unwilling to offer less commonly taught foreign languages that are needed by A&IS programs. The entire A&IS community of a university must compete for funding against student aid, faculty compensation, science and technology programs, and other priorities. Even at the national level, the associations and organizations representing the A&IS campus-based community may face internal conflicts over competing priorities, or find themselves lobbying at cross-purposes. In turn, collaboration or conflict at the national level has a significant influence on levels of investment in A&IS.

Additional A&IS knowledge is produced in the USA by government intelligence agencies, military institutions, government-sponsored think tanks, and risk analysts employed by corporations. This information is generally not accessible to the public, and hence is a relatively minor component of A&IS knowledge. These agencies and institutions are important stakeholders, however, as consumers of A&IS information and training, and to a lesser extent, as funders.

The Consumers of A&IS

A second category of stakeholders is constituted by those institutions that need A&IS information or need A&IS-trained personnel. These can be divided in turn into three broad sectors: government, business, and education.

There is a surprising consistency over time in estimates of U.S. government manpower needs for foreign-language and area-trained personnel. The most cited and thorough study is that of James R. Ruchti of the U.S. Department of State, prepared in 1979 for the Perkins Commission, which surveyed more than 25 agencies and concluded that the federal government employed between 30 000 and 40 000 individuals whose jobs required competence in a foreign language, and that, of these persons, between 14 000 and 19 000 were in positions that required skills in the analysis of foreign countries and international issues (Ruchti, 1979, cited at length in Berryman et al., 1979: pp. 75–114). Although estimates of declining government employment have been assumed to reduce the need for foreign-language skills in the federal government, this was not evident in the mid-1990s. The most recent survey of foreign-language needs at 33 federal agencies, undertaken by Stuart P. Lay in 1995, concludes that these agencies have over 34 000 positions that require foreign-language proficiency, of which an estimated 60% are found in the defense and intelligence community (Lay, 1995). Anecdotal evidence suggests that reductions in force since 1995 may have lowered these figures in the nondefense government sectors, but the preponderance of defense and intelligence employment would reduce the effect of such reductions. If 30 000 positions are taken as a possible lower end estimate to account for reductions, and it is assumed that because of the relatively high turnover of military personnel 20% of the federal positions will require replacement in any given year, a replacement need of 6000 government positions per year can be estimated.

Business demand is hard to estimate. Anecdotal evidence suggests that the U.S. business climate has an increased need for A&IS information and skills. In the 1970s, U.S. business was widely seen as uncompetitive on world markets and lagging in productivity. A survey from this period indicates that less than 1% of jobs at 1266 U.S. firms, which accounted for the great majority of industrial exports, required foreign-language skills. Nevertheless, there were 57 000 jobs at these firms that required or benefited from foreign-language skills (Wilkins and Arnett, 1976). Given the predominance of employment in small firms as opposed to large industrial firms, this was clearly an underestimation of private-sector demand at that time.

Adding an equal-sized small-firm component leads to an estimate of about 100 000 positions, which with a 10% turnover would have required replacement of 10 000 positions per year. Since the 1970s, the share of the U.S. gross national product resulting from international trade has quadrupled. It is therefore reasonable to assume that employment in the private sector of personnel with foreign-language and area skills has increased by several magnitudes. Even if such employment has only doubled since the 1970s, the business sector would need to fill 20 000 positions per year involving foreign-language and area competence.

Educational demand for A&IS has been better studied, but not without controversy. The 1970s recession in higher education employment, combined with the Nixon-inspired drop in Title VI funding, led to gloomy projections about the future academic demand for language, international, and area-

trained personnel (see, for example, the extended discussion in Berryman et al., 1979: pp. 30–74). There was also a concomitant reduction in the production of international and area studies PhDs, compared with the 1960s. However, the optimistic projections of the Barber and Ilchman study of 1977 proved more accurate: they noted that tenured area studies faculty were significantly older than the general tenured faculty population, and predicted a surge of retirements over the following 10 years (Barber and Ilchman, 1977). The academic employment market for language and foreign-area specialists was indeed strong during the 1980s.

Another dimension of change in higher education created additional need for foreign-language, international, and area studies faculty, namely the expansion of public-sector undergraduate-teaching institutions, most notably community colleges and branch campuses. The growth of this sector included a substantial and largely unforeseen growth in international education activities, including the teaching of language, international, and to a lesser extent, foreign-area content courses. By the end of the 1980s a sizable proportion of members of the foreign-area studies association were located at undergraduate-teaching institutions.

During the 1990s the demand for postsecondary faculty strengthened. Unemployment among PhDs dropped by more than one-third in the mid-1990s, and all major professional associations, including those in foreign-language and area studies, reported increased postings of job announcements. In part, this reflected the retirement of faculty hired during the higher education boom of the 1960s. Because A&IS grew even more rapidly than higher education as a whole in the 1960s and 1970s, the job market for these specialists was relatively becoming strong through the first decade of the twenty-first century.

Detailed projections of retirement patterns based on the age cohorts of the area studies association's membership were prepared in 1991 by the National Council of Area Studies Associations, which represents the five major area studies associations. Exit rates of present humanities and social science faculty were based on respondents' plans to retire, estimated at 16.9% for 1997 through 2001, and at 16.8% for 2002 through 2007, for a total of 33.7%, or one-third of current faculty. These estimations do not include projections of exits due to morbidity or mortality based on the age structure of the cohorts, which would, if included, lead to an overall exit rate of approximately 40%. Using the latter figure and assuming that exiting faculty are replaced but there is no growth in academic demand, 40%, or 7000, of the current 18 000 area studies faculty will need to be replaced in the first 10 years of the twenty-first century, or 700 positions per year. An additional 1200 positions will be opening in international studies, or 120 faculty positions a year.

Another faculty stakeholder group consists of foreign-language and literature teachers, for whom two sources of information exist, membership in professional organizations, and surveys by the National Center for Education Statistics of the U.S. Department of Education. Using these two sources, a recent study estimates the combined foreign-language and literature faculty in higher education at the start of the twenty-first century at 36 000 persons. If the exit projection of 40% for area studies faculty over the first decade of the new century is applied to the estimated total of 36 000 postsecondary language and literature faculty, an additional 14 000 positions would need to be filled.

Demand and Supply

The demand in federal government, business, and higher education for personnel who have foreign-language skills, and international and foreign-area knowledge will be substantial in the coming decades. Additional demand from state and local government and from secondary education may also be expected. Overseas employment offers another, as yet unexplored, source of demand. Annual demand for the first decade of the twenty-first century is estimated at 20 000 business jobs, 6000 government jobs, and 10 000 education jobs, for a total of 36 000 foreign-area, international, and language-trained personnel.

The supply side of the equation is far simpler to estimate, as are the implications for Title VI legislation. The number of federal Foreign Language and Area Studies (FLAS) Fellowships awarded annually through Title VI of the HEA, which was 600 in 2000, does not meet even the academic demand, although it is a stimulus for recruiting superior students. The overall annual production of PhDs by Title VI centers was about 1400 per year in the early 1990s. Beginning in 1993 there was a substantial increase in the number of universities receiving National Resource Center for Foreign Language and Area Studies (NRC) or FLAS funding, leading to a jump of PhD production to about 1900 language- and area-trained personnel (Schneider, 1995b). These numbers, which have since been stable, are less than the estimated annual higher education demand for 2100 foreign-language or area studies faulty over the first decade of the twenty-first century.

The production of MA degrees by Title VI centers is higher, averaging about 6000 per year in the 1990s (Schneider, 1995b). That number is far below the annual combined demand for about 14 000 persons coming from the K-12 education sector and the federal government. Production of BAs by Title VI centers approximated 27 000 students annually by the early 1990s, of which an estimated 6000 enter graduate school, and 21 000 of the BA graduates enter the job market. Of the 6000 MA recipients, about 2000 continue graduate study and 4000 enter the market. Thus a combined total of 25 000 BA and MA graduates with foreign-language or area training enter the job market, compared with a demand from business, government, and K-12 education estimated above at 34 000 positions. These positions must be filled by persons trained in other programs or on the job.

As a consequence of the shortfall of trained personnel, the federal government spends considerable sums of money on in-house training programs, such as the Department of Defense's Defense Language Institute (DLI), the National Security Agency's National Cryptologic School (NCS), and the Department of State's Foreign Service Institute (FSI). The DLI and the NCS together train about 4600 students annually. In 1995, for example, the Department of Defense alone spent over $78 million to train linguists to meet its need, considerably more than the cost of all Title VI programs that year (Lay, 1995: p. 1). In-house foreign-language training at the FSI cost an

additional $10 million. These figures do not include the salaries of the personnel who are being trained.

The Investors in A&IS

Investment in A&IS in most of the world is supported almost exclusively by national governments. In the United States, funding for A&IS has come from the federal government, but also from two other sources – private foundations, and from colleges and universities themselves. Title VI of the HEA, formerly the National Defense Education Act (NDEA), administered by the Department of Education, has been the primary instrument through which the federal government supports A&IS (cited in McDonnell et al., 1981: p. 11). Beginning in the mid-1990s, the National Security Education Act provided additional support through the Department of Defense.

The history of A&IS has been of a roller coaster. Federal funding was critical to the development of A&IS in the United States, but has been threatened or reduced a various times. Foundation funding was highly important at one time, but was later reduced to modest levels. Despite this variation of support through time, the commitment of higher education institutions support for A&IS, once established, has proven relatively constant.

The original rationale for NDEA as a whole was narrow and clearly articulated: "To insure trained manpower of sufficient quality and quantity to meet the national defense needs of the United States" (cited in McDonnell et al., 1981: p. v). In the post-Sputnik atmosphere of concern about Soviet achievements in science and technology, the NDEA legislation focused primarily on training in the physical sciences and engineering. However, *prior* to Sputnik the U.S. Office of Education had prepared draft legislation on foreign-language and area training. This legislation, according to one contemporary official, had been drafted because, "By the mid-1950s responsible people in the Government were beginning to realize that university resources in non-Western studies were wholly inadequate to meet present and anticipated national needs. Some measure of Government assistance to language and area studies seemed essential" (Mildenberger, 1966: pp. 26–29).

The Office of Education's foreign-language and area studies draft was incorporated in the NDEA bill as the result of negotiations between Assistant Secretary of Health, Education, and Welfare, Eliot L. Richardson, on behalf of the Eisenhower Administration and the sponsors of the legislation, Senator Lister Hill and Representative Carl Elliott (Clowse, 1981). Like all sections of NDEA, the original Title VI emphasized training, in this case of individuals in modern foreign languages "needed by the Federal Government or by business, industry, or education" and "not readily available in the United States." Such individuals were also to be trained "in other fields needed to provide a full understanding of the areas, regions, and countries in which such language is commonly used," including "fields such as history, political science, linguistics, economics, sociology, geography, and anthropology" (NDEA of 1958, as Amended, reproduced in Bigelow and Legters, 1964). In subsequent legislation, Fulbright-Hays 102 (b)(6) was added to Title VI to provide funding for overseas training components, such as doctoral dissertation research abroad.

The original NDEA Title VI legislation envisioned a 50–50 partnership in which the costs of foreign-area centers would be divided equally between the universities and the federal government (a munificent arrangement in comparison to the 20:1 ratio of university to federal support for Title VI centers at the time of writing) (McDonnell et al., 1981: p. 38). Federal support for foreign-area studies was augmented by sizable investments from the philanthropic community, led by foundations such as Ford, Rockefeller, Mellon, Carnegie, and Tinker. The Ford Foundation alone contributed about $27 million annually between 1960 and 1967 for advanced training and research in international affairs and foreign-area studies, more than federal appropriations for Title VI in the same period.

The establishment of foreign-area programs was a costly venture for universities even with federal and foundation subsidies. Yet universities responded to NDEA Title VI with enthusiasm. This reflected a consensus between academia and government on national needs in the field of international education, forged by World War II and reinforced during the early stages of the Cold War. Many, if not all, foreign-area scholars, university administrators, and foundation executives had served in State Department, intelligence, or military positions during World War II or the Korean War. The boundaries between government and academic institutions were permeable and amicable. The relatively small size of the community of foreign-area specialists inside and outside government meant that people knew one another. Perhaps most important of all, they shared a similar perspective on the U.S. role in world affairs. The alliance between academia and government for foreign-area training and research was a natural outgrowth of these affinities.

The Title VI experiment proved highly successful. Within a decade the United States had established a network of centers covering most foreign areas, generating unprecedented quantities of research, and producing substantial numbers of new foreign-language and area specialists. Ancillary institutions such as foreign-area studies associations and research journals multiplied quickly. The prestige and the funding conveyed by Title VI designation continued to be rewarded by university administrations (Merkx, 1990: pp. 1, 18–23).

Following the promising early start of Title VI came the Vietnam War, which by the late 1960s led to funding freezes and then to declines. Moreover, the Vietnam War itself produced a significant deterioration in the relationship between government and academia. Academic dissent from U.S. foreign policy in Southeast Asia began to grow during the Johnson Administration and reached high levels during the Nixon Administration. The events of Watergate did little to increase academic confidence in government. Resentment by government officials of criticisms from academia, including from some of the very foreign-area specialists trained under Title VI, grew as well. The gradual retirement of the World War II generation of leaders further contributed to the growing gulf between academic and government cultures.

The Carter Administration made some effort to improve relations with the academic community. Carter appointed a Presidential Commission on Foreign Language and International Studies (known as the Perkins Commission), which in 1979 issued a strong call for increasing federal support of international education to more than three times existing levels

(US Department of Health, Education, and Welfare, 1979). Among its recommendations were the establishment of undergraduate and regional foreign-area centers as complements to the Title VI national resource centers, and the provision of annual federal grants of $50 000 to each of the national resources centers for library costs. However, the timing of the Perkins Commission report was not propitious. Double-digit inflation and high interest rates were leading to reductions, not increases, in federal expenditures.

A more successful initiative by the Carter Administration was to repeal NDEA and incorporate its Title VI functions in a new HEA. This step ratified the separation of university-based foreign-area studies from the national defense needs that had been the original justification for federal funding. While pleasing many campus-based area specialists, the change of name made Title VI a more vulnerable target for those in the next administration who were to argue that federal subsidies were no longer in the national interest.

The years of the Reagan Administration represent the nadir of the partnership between academia and government in foreign-area studies. Due to the high inflation of the late l970s and early 1980s, and essentially level funding in current dollars, Title VI appropriations in real terms were at their lowest level since the inception of the program, falling to well below half their 1967 high point. The Administration was formally opposed to the program. Higher education in turn was suffering from the consequences of inflation, the growth of university enrollments had taken a sharp downturn for demographic reasons, and foundation support for international education had evaporated.

Nonetheless, the partnership between government and academia survived. Support for Title VI came from three directions. A few well-placed government officials recognized the value of Title VI research and training, and were moved to intervene on its behalf. A second source of support came from the directors of Title VI-funded area centers, who were galvanized by the realization that the survival of Title VI programs could no longer be taken for granted. The second-generation center directors of the 1980s lacked the Washington contacts and political knowledge of the founding generation. They were forced, however, to increased activism by the disastrous implications of a total cutoff of federal support for their programs. After the failure of their efforts to reverse the Reagan Administration zero-funding of Title VI in the early 1980s, the center directors focused on the Congress. These early attempts at congressional relations were for the most part uncoordinated individual initiatives, but were not without effect. Congress proved more responsive to the universities than the administration had been, and Title VI funding survived.

The third source of support came from individuals outside of government and academia that were affiliated with institutions involved directly or indirectly with international education, such as the major foundations, higher education associations, and professional organizations.

The growing sense of shared concern about the chronic underfunding of international education in general, and the threat to Title VI in particular, coalesced in an initiative that began with a dinner at the Smithsonian Institution in late 1984 "to discuss what might be done to stabilize long-term federal support for international studies and foreign language training" (Prewitt, 1986: p. v). Several foundations funded, under the aegis of the AAU, a study which resulted in the monograph *Points of Leverage: An Agenda for a National Foundation for International Studies* (Prewitt, 1986: p. v), accompanied by the recommendation of an AAU advisory committee chaired by Kenneth Prewitt of the Rockefeller Foundation (AAU, 1986). The Prewitt committee offered draft legislation for the establishment of a National Foundation for Foreign Languages and International Studies.

The AAU Legislative proposal was intended to serve as a rallying point for the various international education constituencies. It had the opposite effect. The proposal was viewed with suspicion as being the product of the elite institutions associated with the AAU. It was widely misinterpreted as constituting a threat to existing programs. The process by which the proposal was developed was criticized for having too narrow a base and not including a sufficiently broad spectrum of international education interests.

In a constructive response to these criticisms, the AAU obtained additional foundation support to broaden the dialogue about international education needs, establishing in 1987 a 2-year effort known as the Coalition for the Advancement of Foreign Languages and International Studies (CAFLIS). An open invitation to participate was sent to virtually every type of group that might be interested. Ultimately, some 150 organizations participated in one or more stages of CAFLIS discussion, including most higher education associations, area studies associations, language groups, organizations engaging in foreign exchanges, peak organizations in the social sciences, and professional associations.

Despite, or perhaps because of, the inclusive nature of the process, CAFLIS ultimately ended in failure. The various constituencies represented in the CAFLIS process had different agendas and found difficulty coming to an overall agreement. A fundamental cleavage existed between those groups that emphasized the need for increased investment to meet the original Title VI goals of advanced training and research in foreign-language and area studies (a focused and less costly agenda), and those who were advocating federal subsidies for language training in primary and secondary schools, for internationalization of undergraduate education, or for adding international dimensions to professional training in fields such as business, medicine, and engineering (a diffuse and more costly agenda).

The disappointing outcomes of both the AAU legislative proposal and the CAFLIS process led the core A&IS groups to refocus attention on Title VI. The HEA was to expire in 1990. The AAU, the National Association of State Universities and Land Grant Colleges, and the other higher education presidential associations organized a series of meetings on Title VI reauthorization that culminated in the appointment of an Inter-Association Task Force on HEA–Title VI Reauthorization. The directors of foreign-area centers receiving federal funding formed a Council of Title VI National Resource Center Directors (CNRC). The foreign-area studies associations organized a coordinating group of executive directors and presidents known as Council of Area Studies Associations.

This resurgence led to a strategy of building support for Title VI programs in particular, and A&IS in general, by mobilizing

the campus-based core constituencies, their professional associations, and national organizations such as the AAU. At the invitation of CNRC a meeting of all these groups took place at the Library of Congress in 1991, for the purpose of defining a common agenda with respect to Title VI reauthorization. When the final law reauthorizing HEA was approved by the Congress in 1992, Title VI incorporated, virtually intact, almost all of the recommendations of the Inter-Association Task Force which had been supported by the core constituencies.

This success of the 1991–92 reauthorization effort resulted in the establishment of the Coalition for International Education, which includes 26 national organizations representing the various A&IS stakeholder communities. This membership includes six higher education associations representing college and university presidents, two associations of international education administrators, four associations of Title VI-funded program directors (representing area centers, business centers, language centers, and schools of international administration), two library associations, two international exchange associations, one council of overseas research centers, one area studies association, a social science association, a humanities alliance, an overseas university, and an association of graduate schools.

Federal funding for A&IS reached a low point in constant dollars during the mid-l980s. Following the establishment of the Coalition for International Education, appropriations grew steadily. Adjusted for inflation, funding for Title VI grew 6.5% between the 1994 and 2000 federal budgets. The original Title VI programs, such as the area centers, are funded at about one-half their constant dollar figures of the late l960s. The overall International Education and Foreign Language Studies account in the year 2000 federal budget was approximately $70 million. In the first decade of the twenty-first century, Title VI appropriations grew more rapidly, reaching a high point in FY 2010 of $125.7 million. However, the FY 2011 appropriation was reduced by $50 million to $75.7 million. It declined to $74.2 million in FY 2012, and was restored to $75.7 million for FY 2013. The political stalemate in Washington at the start of the second term of the Obama Administration does not bode well for restoring the cuts in the Area and Foreign Language (A&FL) programs of the Department of Education.

See also: Cold War, The; Policy Knowledge: Universities.

Bibliography

AAU, 3 October 1986. To Strengthen the Nation's Investment in Foreign Languages and International Studies: A Legislative Proposal to Create a National Foundation for Foreign Languages and International Studies. Association of American Universities, Washington, DC.

Barber, E.G., Ilchman, W., 1977. International Studies Review. The Ford Foundation, New York.

Berryman, S.E., et al., September 1979. In: Foreign Language and International Studies Specialists: The Marketplace and National Policy. The Rand Corporation, Santa Monica, CA, pp. 30–74.

Bigelow, D.N., Legters, L.H., 1964. NDEA Language and Area Centers: A Report on the First Five Years. US Department of Health Education, and Welfare, Office of Education. US Government Printing Office, Washington, DC.

CAFLIS, December 1989. The Federal Government: Leader and Partner. Report on the recommendations and findings of the Coalition for the Advancement of Foreign Languages and International Studies: Working Group on Federal Support for International Competence. CAFLIS, Washington, DC.

Clowse, B.B., 1981. Brainpower for the Cold War: The Sputnik Crisis and the National Defense Education Act of 1958. Greenwood Press, Westport, CT.

Lambert, R.D., July/August 1990. Blurring the disciplinary boundaries: area studies in the United States. American Behavioral Scientist 33 (6), 712–732.

Lambert, R.R., et al., 1984. In: Beyond Growth: The Next Stage in Language and Area Studies. Association of American Universities, Washington DC, p. 13.

Lay, S.P., 1995. Foreign Language and the Federal Government: Interagency Coordination and Policy. MA thesis. University of Maryland.

McDonnell, L.M., et al., 1981. In: Federal Support for International Studies: The Role of the NDEA Title VI. The Rand Corporation, Santa Monica, CA, p. 11.

Merkx, G.W., Summer 1990. Title VI accomplishments, problems, and new directions. LASA Forum 21 (2), 1, 18–23.

Merkx, G.W., 2000. Foreign language and area studies through Title VI: assessing supply and demand. In: Lambert, R.D., Shoharny, E. (Eds.), Language Policy and Pedagogy. Essays in Honor of A. Ronald Walton. John Benjamins, Amsterdam, pp. 93–110.

Mildenberger, K.W., 1966. The federal government and the universities. In: International Education: Past Present, Problems and Prospects. Task Force on International Education, John Brademas, Chairman, Committee on Education and Labor, House of Representatives. US Government Printing Office, Washington, DC, pp. 26–29.

NCASA, 1991. Prospects for Faculty in Area Studies. NCASA, Stanford, CA. Report of the National Council of Area Studies Associations.

NCFLIS, 1981. Report of the Task Force on National Manpower Targets for Advanced Research on Foreign Areas. Mimeo. National Council on Foreign Languages and International Studies, New York.

Prewitt, K., 1986. Preface. In: Lambert, R.D. (Ed.), Points of Leverage: An Agenda for a National Foundation for International Studies. Social Science Research Council, New York, p. v.

Ruchti, J.R., 12 July 1979. The U.S. Government Employment of Foreign Area and International Specialists (Paper prepared for the President's Commission on Foreign Language and International Studies).

Schneider, A.I., 15 September 1995a. Title VI FLAS Fellowship Awards, 1991–1994. Memorandum to Directors of Title VI Centers and Fellowships Programs, Center for International Education, US Department of Education, p. 9 (Table C).

Schneider, A.I., 26 September 1995b. 1991–1994 Center graduates: Their Disciplines and Career Choices. Memorandum to Directors of Title VI Centers and Fellowships Programs. Center for International Education, US Department of Education.

US Department of Health, Education, and Welfare, November 1979. Wisdom: A Critique of U.S. Capability. Report to the President from the President's Commission on Foreign Language and International Studies. US Government Printing Office, Washington, DC.

Wilkins, E.J., Arnett, M.R., June 1976. Languages for the World of Work. Olympus Research Corporation, Salt Lake City.

Area Studies, History of

Matthias Duller, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Area studies are the systematic efforts of knowledge production about particular world-regions through academic institutions often used to inform foreign policy strategies of national governments. Their history is a classic example to show the interrelatedness between the political history of the Cold War and the intellectual history of the social sciences in the twentieth century. This article outlines the emergence of area studies in the Soviet Union and the United States of America before turning to more recent trends and criticisms that characterize the field until today.

Introduction

The history of area studies is a particularly illustrative example for the interrelatedness of the intellectual history of the social sciences and the political history of the twentieth century. Area studies are the systematic efforts of knowledge production about particular world-regions through academic institutions often used to inform foreign policy strategies of national governments. Within the realm of foreign policy analysis, area studies were supposed to provide familiarity with various local contexts in the non-Western world through linguistic, historical, cultural, sociological, economic, and political research that usually relied on extensive field experience in the respective countries. In that, they explicitly challenged disciplinary boundaries in the social sciences while emphasizing an integrative interdisciplinary approach to a geographically limited space. The definition of area boundaries and allocation of resources always mirrored the geopolitical interests of the powers behind these endeavors. African studies, Russian and Soviet studies, North American studies, East European studies, Near or Middle Eastern studies, South Asian studies, Southeast Asian studies, East Asian studies, and Latin American studies denote some of the most typical research fields of area studies institutions.

Area studies have typically been described as a US-American Cold War enterprise (*see* Cold War: Impact on Social and Behavioral Sciences). Having evolved from the extensive use of scientific research groups in military, government, and intelligence agencies during World War II, area studies centers and departments were one of the institutional offsprings of interdisciplinary policy-oriented social science research organizations exemplifying the close relationship between the sciences and politics in Cold War America. While it might be true that no other nation has yielded area studies in comparable scope and breadth as the USA did during the Cold War years (Szanton, 2004: 8), the history of area studies can neither geographically nor temporally be reduced to Cold War America. The case of the Soviet Union shows how Russia's colonial tradition in Oriental studies has been reorganized toward politically more useable knowledge production about its peripheral and neighboring cultures already during the 1920s in a way that is remarkably similar to U.S. area studies 20 years later. Furthermore, area studies have not lost their significance since the Cold War has ended.

Area studies' uncomfortable position between the diverging normative demands from political and scientific institutions was at the heart of many controversies concerning the ethics of applied social science. At best, area scholars drew theoretical and methodological innovations from their high degrees of entanglement with actual political processes and institutions that fed back into the social science disciplines in fruitful ways. At worst, area scholars were part of highly questionable intelligence operations without any scientific value.

This article begins by outlining the development of area studies in the Soviet Union since the 1920s, followed by a discussion of the situation in the USA. In both cases, emphasis is laid on its developmental phase and in particular on the interplay between the political context and the intellectual history of area studies. Finally, the field's diversification including its central conflicts and criticisms will be reviewed and current trends be traced (for detailed accounts *see* Area and International Studies in the United States: Stakeholders; Ethical Questions in Social and Behavioral Sciences, history of; Funding of Social Science, History of; Modernity: History of the Concept; Cold War: Impact on Social and Behavioral Sciences; Counter insurgencies and the Social Sciences; Western and Non-Western Views of the Social and Behavioral Sciences: De-Westernizing the Social and Behavioral Sciences). Due to the field's tremendous diversity, a limited number of examples will serve to illustrate the general problem of area studies as much as some specific problems that resulted from this influential and often controversial field of scholarship.

Area Studies in the Soviet Union

Oriental Studies in the Interwar Period

Soviet foreign area scholarship had its roots in the tsarist Russian empire where studies of the languages and cultures of 'the Orient' have a long tradition that reaches back into the eighteenth century. By the turn to the twentieth century, the Institute of Oriental Studies of the Russian Academy of Sciences in St. Petersburg, founded in 1818, was Europe's leading research institution in the field (Schimmelpenninck van der Oye, 2011) housing linguistic and historical specialists on a vast geographical area from Northern Africa, the Middle East, and the entire rest of Asia, including Russia's own territories in Central and Eastern Asia. From the beginning, this institution had advisory functions to the tsarist regime. After the Russian Revolution in 1917, however, Oriental studies were fundamentally reorganized in order to meet the new political requirements of the Bolshevik regime.

In its historical mission to spread the Communist revolution across the world, the Russian Communist Party redirected its foreign policy priorities to the Middle East, after revolutionary attempts had failed in Europe in the aftermath of World War I. As early as in 1919, Leon Trotsky advocated a strategy to destabilize the bourgeois regimes in Europe by instigating national revolutions in the Middle East against the European colonial powers. This emancipatory revolutionary agenda for the Oriental peoples constituted the political basis for Soviet area studies in North Africa and Asia during the interwar period and beyond.

The St. Petersburg (then called Petrograd) Institute of Oriental Studies was skeptically viewed by the Bolshevik leadership due to its historical legacy as a bourgeois research institution and its former ties to the tsarist regime. Therefore an additional Institute of Oriental Studies was established in Moscow in 1921 that became the primary institution to train political, administrative, diplomatic, and educational personnel for the Soviet state in questions regarding the Orient for several decades (Kemper, 2011: 3). Besides doubts in the political loyalty of the Petrograd Institute, the Moscow Institute was to replace Petrograd's emphasis on classical languages and history by contemporary social sciences that were able to provide politically immediately useable skills and knowledge.

Key event for the new Soviet Oriental studies was the Congress of the Peoples of the East held in September 1920 in Baku/Azerbaijan where leading politicians and intellectuals from Oriental countries including Azerbaijan, Daghestan, Central Asia, Iran, and Turkey convened to corroborate the Soviet Union's alliance with anticolonial revolutionary movements in the Middle East (Kemper, 2010: 448). Oriental studies, in this situation, had the double function of systematically generating knowledge about the local conditions in the Oriental countries for the Soviet elite and creating a national political consciousness among, what was called, the uneducated masses of the Orient, which would finally lead to revolution. The second rationale is especially visible in the plan to establish a University of Social Sciences for Workers of the Orient in Baku, comprising training in history and culture, geography and ethnography, economy, legal sciences, socialism, and revolutionary movements in the East. Additionally to this social science education, the designated curriculum allowed for specializations in a specific country (Kemper, 2010: 454). Though never realized in Baku, two Moscow institutions, the Communist University of the Toilers of the East (1921) and the already mentioned Institute of Oriental Studies (1921) were built in accordance to these plans, educating both Oriental and Russian students in Oriental languages and social sciences.

As the intellectual key figure of early Soviet Oriental studies emerged Mikhail P. Pavlovich (1871–1927), a former Menshevik intellectual who gained a reputation as a theorist of revolutionary movements in Africa and Asia (Kemper, 2010). Having held advisory positions in the young Soviet state before, Pavlovich became chairman to the newly funded All-Russian Scientific Association of Oriental Studies (Vserossiiskaia Nauchnaia Assotsiatsiia Vostokovedeniia, VNAV) in Joseph Stalin's People's Commissariat for Nationalities in December 1921 and director of the Moscow Institute for Oriental Studies in 1923. Different from the Baku initiative with its direct revolutionary purpose, the VNAV was designed to coordinate scholarly activities in Oriental studies throughout the Soviet state. Its purpose was to inform policy makers as much as the broader public. Quickly, it published a journal called Novyi Vostok (New Orient, 1922–30) and opened a number of new branches during the 1920s from the Ukraine to Siberia (Kemper, 2010: 455f.).

Intellectually, early Soviet Oriental studies were mostly concerned with determining the historical stage within the Marxian development model on which the Oriental peoples were to be located and what role Islam played in that process. In accordance with the Comintern's interwar strategy to support national identity policies (*korenizatsiia*) in order to gain popularity within the peasant and workers' masses, Islam was initially seen far less critically than Christian orthodoxy in Russia. By the 1930s, however, with the political repressions escalating toward Stalin's Great Purge, moderate views on religion became gradually impossible to express. Many Orientalists suffered severe repression, imprisonment, forced labor, and execution (Kemper, 2011: 6–8). With the German attack on the Soviet Union in 1941 and the subsequent devastating war, the situation further aggravated. Political repression and censorship continued after the war and culminated once again in Stalin's xenophobic campaign against 'cosmopolitanism' during the early 1950s. Only after Stalin's death in 1953 did Soviet area studies enter into a phase of institutional and intellectual consolidation and diversification.

Diversification of Area Studies in the Soviet Union during the Cold War

With the expansion of the Soviet Union's global activities during the Cold War, its area studies programs expanded in geographical scope too. North American studies, lacking prewar traditions in Russian scholarship, emerged in the Soviet Union during the 1950s and 1960s as a distinct field. The first important center for the study of North America was located at the Institute for World History at the Russian Academy of Sciences. Besides its interest in American foreign policy, especially with respect to US–Russian relations as well as U.S. policies in Latin America and the Far East, Soviet Americanists paid much attention to the emergence of capitalism in the USA, the history of slavery, and in particular the study of African Americans in past and present (Bolkhovitinov, 1969). The second was the Institute for US and Canadian Studies at the Russian Academy of Sciences (ISKRAN), whose founding director from 1967 to 1995 was the political scientist Georgy Arbatov (1923–2010). Different from its counterpart at the Institute for World History whose members' often maintained a critical distance to the Soviet regime (Zhuk, 2009), ISKRAN's work has been very close to foreign policy decision-making with Arbatov personally advising all General Secretaries of the Communist Party of the Soviet Union since Leonid Brezhnev (1964–82).

Institutes for African Studies (1960) and Latin American Studies (1961) (Westad, 2007: 68) completed the geographical scope of area-specific research organizations in the Soviet Union. In all cases, these institutions prepared briefs for Party and intelligence agencies and their research agendas displayed the political and ideological interests of the Soviet regime. Special attention has been given to national liberation movements and decolonization (Desai, 1968: 252). The ubiquitous

criticism of American foreign and domestic policies reveals that area studies also fulfilled an important propagandistic function. Nevertheless, Soviet area studies centers produced valuable scholarship that received appreciation also among Western colleagues (Bartley, 1970). Though rigorously surveilled by the Komitet gosudarstvennoy bezopasnosti (KGB, Committee for State Security), Soviet area scholars increasingly engaged in fieldwork in all world areas, including the USA.

Area Studies in the United States of America

World War II and the Origins of Area Studies in the USA

In US-American scholarship knowledge about the non-Western world was hardly available in any organized fashion prior to World War II (cf Chomsky, 1997), possibly with the exception of the rather small community of cultural anthropologists. With the United States entering the World War II in December 1941, however, background knowledge on historical, cultural, and political issues in a variety of Western and non-Western countries suddenly became urgently needed. Various military and government agencies began seeking support from scientists. Most notably, the Office of Strategic Services (OSS) and the Office of War Information (OWI), the prime wartime institutions for intelligence and propaganda, created large and well-organized research units that recruited scholars from various social science disciplines and the humanities from the top universities in the US and abroad (Katz, 1989). Additionally, the Army Specialized Training Program was launched at 227 U.S. universities by the end of 1942 to train military personnel in various specializations, among them specific area and country education at 55 campuses (Lambert, 2001: 687).

Besides language training, a certain familiarity with cultural specificities was considered necessary which gave cultural anthropologists an opportunity for war-related activities. John M. Cooper estimated that 'probably a good half' of American professional anthropologists, engaged in 'full-time war activities' during World War II (Cooper, 1947: 2). The two most outstanding figures in this respect were Margaret Mead, working, among other things, on the improvement of war propaganda and Ruth Benedict, who contributed the famous study on Japanese culture *The Chrysanthemum and the Sword: Patterns of Japanese Culture*, first published in 1946 (Mandler, 2012). Commissioned by the OWI in June 1944 in light of a planned U.S. occupation of Japan, Benedict produced a comprehensive account of Japanese culture in order to predict Japanese reactions to military intervention and a possible occupation scenario after the war. While influencing American views on Japanese culture for decades to come, the study indicates how social science knowledge could be used in foreign policy decision-making. Characteristic features of wartime social science research in government organizations were interdisciplinary research teams, a keen focus on analysis and solution of concrete policy problems and close attention to the local particularities. In that sense, McGeorge Bundy later called the OSS the 'the first great center of area studies' in the USA (Engerman, 2009: 3). The importance of this kind of research for intelligence agencies was mirrored in the Central Intelligence Agency's (CIA) continuation of OSS's Research and Analysis Branch after the war (Müller, 2010).

Postwar: Institutional Takeoff

By the end of World War II there was a broad consensus in US-American science and policy circles that general knowledge about the non-Western world was in a very poor state and that the new international role the United States had achieved in the course of World War II required massive investment in producing such knowledge. The Soviet Union solidifying its strength in Europe and Asia, the rise of China as another Asian giant, and early signs of decolonization in Asia and Africa were perceived as the threats that necessitated this investment. In this situation, international and area studies had the function to provide 'culturally and historically contextualized knowledge' about world-regions of political interest (Szanton, 2004: 10).

This governmental impetus notwithstanding, postwar area studies were primarily organized by nongovernmental organizations, above all the Social Science Research Council (SSRC) and the American Council of Learned Societies. Funding has been predominantly provided by the universities themselves though only through contributions from a number of private foundations could the additional financial needs for language training and overseas research be met. Among them, the Ford Foundation stood out as the single most important organization, investing more than $ 270 million only between 1951 and 1966 and funding 2.050 doctoral students between 1951 and 1972 (Szanton, 2004: 11). Other organizations included the Rockefeller and Mellon Foundations or the Carnegie Corporation. Government funding became an additionally important factor with the 1958 National Defense Education Act (NDEA) following the Sputnik shock the year before. Title VI of NDEA and later the Higher Education Act established language and area programs as recipients of government support. Together, these institutions were able to trigger a massive expansion of area studies organizations across the United States. Almost inexistent before World War II, area-specific curricula have been offered at nearly every second university in the United States by the 1990s (Lambert, 2001: 687; *see* Area and International Studies in the United States: Stakeholders; Funding of Social Science, History of).

Two institutional forms have emerged in area studies training: area studies centers and area studies departments. The latter were modeled after wartime organizations and offered interdisciplinary training including languages, literature, history, religion, and occasionally politics on undergraduate and graduate level. Area studies centers (or institutes and programs respectively), on the other hand, were designed for students who have been trained in one of the social science disciplines before, among them anthropology, art history, geography, history, language and literature, music, sociology, political science, and others. This type gained importance during the 1960s and dominated the field since the 1970s typically resulting in double identities of area scholars with regards to their home disciplines and the region they specialized in (Szanton, 2004: 7).

Early Russian or Soviet Studies

World War II ended in the bipolar world order of the emerging Cold War between a Western and an Eastern bloc with the USA and the Soviet Union as their respective leaders. Not surprisingly, Russian or Soviet studies immediately emerged as the

first full-fledged field of area studies in the USA. Three centers stand out particularly for the early phase of US Soviet studies. The first was the Russian Institute (RI) at the Columbia University, opening as early as in 1946 with the support of the Rockefeller Foundation and working closely with the U.S. army and navy. The second was the Russian Research Center (RRC) at Harvard, established in 1948, getting most of its support from the CIA and the air force (Engerman, 2009: 7). The third major center of early Soviet studies in the USA was the Center for International Studies (CENIS) at the Massachusetts Institute of Technology (MIT), founded in 1951, receiving significant support from the CIA too. Together, these centers display the field of early Russian and Soviet studies, which shows strong personal and institutional continuities with wartime research experiences and in which entanglements with government and intelligence agencies were exceptionally close.

Columbia's RI, with eminent historian Geroid T. Robinson the only center having an expert on Russia as its director, was the politically most diverse center of the three, including open admirers of socialism and the Soviet Union. Unusual for this early phase of American Soviet studies, the RI was able to send researchers to the Soviet Union for conducting fieldwork. During the rampantly anti-Communist McCarthy era, this fueled suspicion about the faculty's loyalty to the anti-Communist stance of U.S. foreign policy. While being largely insignificant in publishing research on the Soviet Union, the RI's credentials lay in its important efforts to build the academic infrastructure for US-American Soviet Studies. It was by far the most productive institute to train experts on the Soviet Union for government service and, primarily through the Joint Committee on Slavic Studies, to secure library facilities, access to Soviet newspapers and periodicals, and to collect the existing academic literature on Russian affairs in the USA. Therefore, Columbia's RI has "rightfully claimed to be the founder of modern Russian Studies in the United States" (Engerman, 2009: 13–42, here: 36).

Upon its inauguration, Harvard's RRC could already benefit from the research infrastructure set up by the RI. In contrast to Columbia, however, the RRC assembled social scientists with strong disciplinary backgrounds rather than prior research experience in or about the Soviet Union and displayed a general leaning toward the emerging paradigm of behavioral science. Its first director was Clyde Kluckhohn, an anthropologist who established his reputation in the OWI's Foreign Morale Analysis Division during World War II where also Ruth Benedict's study on Japanese culture originated. Though variously involved in government-sponsored projects, the RRC was successful in maintaining some degree of scholarly distance to politics and established itself as the leading research center on Soviet affairs in the USA.

Probably the most conspicuous example of intelligence sponsored academic research related to area studies was the CENIS at the MIT. The MIT had been heavily engaged in wartime research for government and intelligence agencies and has continued to work on classified and secret research projects during the years that followed World War II on a regular basis. One offspring of this political entanglement was Project Troy, a 3-month top-secret research project funded by the State Department in autumn 1950 to explore US-American possibilities for propaganda actions behind the Iron Curtain including technical and social science perspectives. The 21 distinguished scholars in Project Troy found their work a stimulating experience and suggested to establish an organization dedicated to Cold War social research. Their recommendations led to the establishment of CENIS in 1951 (Needell, 1998). Its intellectual core group was composed of economists around Max Millikan and Walt Whitman Rostow who, not being experts on the Soviet Union, quickly developed research interests in international economics and foreign policy more broadly (Blackmer, 2002).

Diversification and Liberalization of Area Studies

Russian or Soviet studies' intimate entanglements with American Cold War foreign policy, including covert intelligence operations, illustrate the aspect of political application orientation in the service of governments that is inherent in the area studies approach. From the beginning of area studies during World War II, however, its proponents also pursued another objective: the deprovincialization of US-American higher education (*see* Western and Non-Western Views of the Social and Behavioral Sciences: De-Westernizing the Social and Behavioral Sciences). Different from the political instrumentality of social science research, this argument emphasized the epistemological benefits of accounting for the plurality of worldviews in a globalizing world. With the higher education sector drastically expanding and the parallel politicization and liberalization of American universities during the 1960s, these two objectives increasingly seemed incompatible with each other.

Area studies' principle to institutionalize in-depth knowledge of non-Western cultures essentially favored interpretative social science disciplines and could only meaningfully be realized through extensive fieldwork. The Ford Foundation's prestigious Foreign Area Fellowship Program, established in 1950, was the most important institution to enable doctoral students to gain the necessary familiarity with the countries they specialized in. Although area scholars have continuously worked with government and intelligence agencies, these activities naturally remained limited in number. With area studies centers proliferating across the USA, most of these institutions had no direct involvement with the government and emphasized academic research instead.

With African studies, Asian studies, and Latin-American studies, area studies quickly diversified its geographical scope. Instead of the 'Know your enemy'-approach prevalent in Russian or Soviet studies (Engerman, 2009), these fields usually encompassed more sympathetic attitudes toward the peoples and cultures they studied, engaged in intensive exchanges with local scholars and often turned expressly critical toward US-American foreign policies in their respective regions. Two events stand out as particularly important that changed the public and professional understandings of the roles social scientists took in foreign policy-related research.

In 1965, the disclosure of Project Camelot brought the topic of government-sponsored social science research to widespread public debates. Planned within the Special Operations Research Office in the American University and the U.S. Department of Defense, the purpose of the project was to "measure, predict and control internal conflict within peripheral countries" (Navarro, 2011: 807) through a series of

case-studies across the world and was to be implemented in Chile. After failing attempts to recruit Chilean social scientists for it, the project had been first publicly discussed in June 1965, which caused massive reactions in Chile, the US and beyond that saw an unacceptable political agenda behind it. Eventually, what was to become "the largest social science project in US history" (Solovey, 2001: 171) was canceled before it actually got started. More generally, the disclosure of Project Camelot fueled already existing discussions about the relation of social science research and espionage and the impact of government funding on the autonomy of the social sciences (*see* Counter Insurgencies and the Social Sciences; Ethical Questions in Social and Behavioral Sciences, History of).

The increasingly critical attitudes of area scholars toward the involvement of their profession in U.S. foreign policy culminated during the escalation of the Vietnam War. In 1968, a group of young Southeast Asian scholars from various area studies centers in the USA formed the Committee of Concerned Asian Scholars (CCAS) in order to express organized protest from within their academic field against the U.S. warfare in Vietnam and the role area scholars had played in it. In particular, the CCAS founding statement criticized the "complicity or silence of [their] profession with regard to that policy" in Asia that served expansionist interests of the USA. Instead, they demanded from Asian scholars to take 'responsibility for the political consequences of their research' and stressed the need to "develop a humane and knowledgeable understanding of Asian societies" instead (Committee of Concerned Asian Scholars, 1969).

Intellectual Trends

With respect to the academic disciplines, area studies stood for strengthening the social sciences as politically engaged applied science – be it in the service of governments or of marginalized groups – and openly challenged traditional disciplinary differentiations in a way that transformed the relations of the disciplines with each other in significant ways (Wallerstein, 1996). Though practically all social science disciplines were represented in area studies in one way or another, their degree of involvement differed.

Psychology, for example, being of central importance to many interdisciplinary endeavors in the social sciences during the early postwar period (Backhouse and Fontaine, 2010), played only a minor role in area studies. While being relevant to certain streams in Russian or Soviet studies, the general tendency of most area studies centers to emphasize the study of cultural particularities stood in opposition to psychology's attempts of establishing a universalist science of human behavior.

Economics, another discipline often cited to have championed the universalist tradition in social science, has produced area specialists especially within its branch of development economics. Modernization theory is maybe the most distinct outcome of economists, and to a lesser degree, political scientists and sociologists, engaging with area studies (*see* Modernity: History of the Concept). Dependency theory and other critiques of development theories are good examples of the dynamics that the engagement of Western social science with non-Western regions and their local scholarly communities evoked (Drake and Hilbink, 2004). The dominance of neoclassical theory, however, made economics play no important role in area studies since (Szanton, 2004: 7).

The disciplines that could establish themselves most firmly in area studies were anthropology, geography, history, literature, sociology, and political science. In all cases, the most innovative impulses were drawn from the confrontation of established concepts from the social science disciplines with in-depth studies in non-Western regions. Increasingly, contributions from scholars from the regions studied gained importance, which sometimes resulted in explicit critiques of the area studies approach itself. The best example for this might be the emergence of Subaltern studies which radicalized the idea of prioritizing the epistemological position of marginal rural populations as opposed to the elite-centered view prevalent in conventional South Asian studies (Dirks, 2004; Guha and Spivak, 1988; Chakrabarty and Bhabha, 2002). Such developments together with rising numbers of students with cultural and ethnic backgrounds from the respective regions inside Western universities contributed to area studies' promise to deliver culturally sensitive knowledge about the non-Western world.

Criticisms and Newer Trends

Besides criticisms referring to the political instrumentalization of area studies for an imperialistic U.S. foreign policy (cf Simpson, 1998), area studies were also criticized on academic grounds. A common argument has been the accusation of being passive consumers of the methodological work done in the core disciplines resulting in a generally inferior quality in scientific standards. This accusation was often coupled with the more far reaching blame for adhering to a merely 'idiographic' research style that confines itself to collecting curiosities from remote world areas without developing generalizable theories. Naturally, this criticism was especially audible from scholars and disciplines that championed universalist approaches to social science, especially dominant in psychology, economics, and political science. Set aside the deeper philosophical discussion about universalism versus particularism in the social sciences, this argument cannot be upheld as an accusation to area studies in general. First, universalist approaches were not absent in area studies centers as has been shown above. Second, the increasing particularist leaning in area studies did not mean absence of theoretical generalizations. To the contrary, area studies most innovative moments emerged from reappraising established theories in the light of new empirical work outside the context where the theories were initially developed.

Another line of criticism has been launched from the epistemologically opposite camp. Poststructuralist and postcolonial theorists have accused area studies of reproducing an imperial perspective on the areas they have been studying by adhering to an established Western conceptual apparatus. Instead, they insisted that cultures could only be studied in a nonhegemonic fashion if they are analyzed in their own terms. Area studies self-understanding of 'translating' foreign cultures to a Western audience (Mirsepassi et al., 2003), even when detached from direct political or developmentalist projects, inevitably contained a sharply asymmetric distribution of power between those studying and those studied. This criticism, most famously related to Edward Said's critique of Orientalism (1978), has been largely endorsed by many area scholars and has become extremely influential since the 1980s.

The relation between power and epistemology inherent to the program of studying non-Western cultures within Western academic institutions has since been excessively debated and constitutes a dominant theme in today's literature on area studies in general.

With the Cold War coming to a sudden end by the end of the 1980s, a third thread of criticism set out to change area studies' intellectual shape. After Communism had disappeared from global politics by 1989/1991, theories of global convergence received widespread support, making area studies' concern with regional particularisms appear as "an outdated concern for a world that no longer exists, or at best, that is rapidly fading away" (Szanton, 2004: 24). While the hypothesis of the "End of History" (Fukuyama, 1992) proved wrong, the rising importance of transnationalism did lead to reorientations in the internationally oriented social sciences in another way. Under the conceptual pair global/local, a new paradigm gained ground in which area studies redefined their role within increasingly comparative approaches (Schäfer, 2010). Early examples of this trend have already emerged during the 1980s, when the SSRC established a Committee on the Comparative Study of Muslim Societies in 1986 and launched a new Dissertation Fellowship Program in 1994 favoring cross-area comparative approaches (Szanton, 2004: 26). Recently, Comparative Area Studies has become a widely accepted approach that integrates in-depth case research in a given area in a systematic comparative research design with the aim of generating generalizable theory (Ahram, 2011).

Besides these conceptual considerations, transnationalism has affected area studies in yet another way. The grossly increasing mobility of well-trained scholars from ever more regions of the world has led to far more intense forms of cooperation between Western area scholars and native scholars from the region under study. Additionally, the role of diasporas has been centrally important to area studies. The intimacy of migrant students with their cultures of origin demanded more critical, reflexive, and better grounded teaching than has previously been the offered adding to the quality and a more careful interpretations of foreign cultures (Szanton, 2004: 27).

See also: Area and International Studies: Stakeholders; Cold War: Impact on Social and Behavioral Sciences; Counter-Insurgencies and the Social Sciences; Ethical Questions in Social and Behavioral Sciences, History of; Funding of Social Science, History of; Modernity: History of the Concept; Western and Non-Western Views of the Social and Behavioral Sciences: De-Westernizing the Social and Behavioral Sciences.

Bibliography

Ahram, A.I., 2011. The theory and method of comparative area studies. Qualitative Research 11, 69–90.

Backhouse, R.E., Fontaine, P., 2010. Toward a history of the social sciences. In: Backhouse, R.E., Fontaine, P. (Eds.), The History of the Social Sciences since 1945. Cambridge University Press, Cambridge; New York, pp. 184–234.

Bartley, R.H., 1970. On scholarly dialogue: the case of US and soviet Latin Americanists. Latin American Research Review 5, 59–62.

Benedict, R., 1946. The Chrysanthemum and the Sword: Patterns of Japanese Culture. Houghton Mifflin, Boston.

Blackmer, D.L.M., 2002. The MIT Center for International Studies: The Founding Years 1951–1969. MIT Center for International Studies, Cambridge.

Bolkhovitinov, N.N., 1969. The study of United States history in the Soviet Union. American Historical Review 74, 1221–1242.

Chakrabarty, D., Bhabha, H.K., 2002. Habitations of Modernity: Essays in the Wake of Subaltern Studies. University of Chicago Press, Chicago.

Chomsky, N. (Ed.), 1997. The Cold War & the University: Toward an Intellectual History of the Postwar Years. The New Press, New York.

Committee of Concerned Asian Scholars, 1969. BCAS Founding Statement. http://criticalasianstudies.org/about-us/bcas-founding-statement.html (accessed 13.08.14.).

Cooper, J.M., 1947. Anthropology in the United States during 1939–1945. Journal de la Société des Américanistes 36, 1–14.

Desai, R., 1968. The explosion of African studies in the Soviet Union. African Studies Bulletin 11, 248–258.

Dirks, N.B., 2004. South Asian studies: futures past. In: Szanton, D.L. (Ed.), The Politics of Knowledge: Area Studies and the Disciplines. University of California Press, Berkeley, pp. 341–385.

Drake, P.W., Hilbink, L., 2004. Latin American studies. Theory and practice. In: Szanton, D.L. (Ed.), The Politics of Knowledge: Area Studies and the Disciplines. University of California Press, Berkeley, pp. 34–73.

Engerman, D.C., 2009. Know Your Enemy: The Rise and Fall of America's Soviet Experts. Oxford University Press, Oxford.

Fukuyama, F., 1992. The End of History and the Last Man. Free Press, New York.

Guha, R., Spivak, G.C., 1988. Selected Subaltern Studies. Oxford University Press, New York.

Katz, B.M., 1989. Foreign Intelligence: Research and Analysis in the Office of Strategic Services, 1942–1945. Harvard University Press, Cambridge.

Kemper, M., 2010. Red Orientalism: Mikhail Pavlovich and Marxist oriental studies in early Soviet Russia. Welt Islams 50, 435–476.

Kemper, M., 2011. Introduction: integrating soviet oriental studies. In: Kemper, M., Conermann, S. (Eds.), The Heritage of Soviet Oriental Studies. Routledge, Abingdon; New York, pp. 1–25.

Lambert, R.D., 2001. Area and international studies in the United States: institutional arrangements. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Elsevier, Amsterdam, pp. 686–692.

Mandler, P., 2012. Deconstructing "cold war anthropology". In: Isaac, J., Bell, D. (Eds.), Uncertain Empire: American History and the Idea of the Cold War. Oxford University Press, Oxford; New York, pp. 245–266.

Mirsepassi, A., Basu, A., Weaver, F., 2003. Introduction: Rethinking Local-Global Connections, in: Localizing Knowledge in a Globalizing World: Recasting the Area Studies Debate. Syracuse University Press, Syracuse.

Müller, T.B., 2010. Krieger und Gelehrte. Herbert Marcuse und die Denksysteme im Kalten Krieg. Hamburger Edition, Hamburg.

Navarro, J.J., 2011. Cold war in Latin America: the Camelot project (1964–1965) and the political and academic reactions of the Chilean Left. Comparative Sociology 10, 807–825.

Needell, A.A., 1998. Project Troy and the cold war annexation of the social sciences. In: Simpson, C. (Ed.), Universities and Empire: Money and Politics in the Social Sciences during the Cold War. The New Press, New York, pp. 3–38.

Said, E.W., 1978. Orientalism. Vintage Books, New York.

Schäfer, W., 2010. Reconfiguring area studies for the Global Age. Globality Studies Journal 22.

Schimmelpenninck van der Oye, D., 2011. The Imperial roots of soviet orientology. In: Kemper, M., Conermann, S. (Eds.), The Heritage of Soviet Oriental Studies. Routledge, Abingdon; New York, pp. 29–46.

Simpson, C. (Ed.), 1998. Universities and Empire: Money and Politics in the Social Sciences during the Cold War. The New Press, New York.

Solovey, M., 2001. Project Camelot and the 1960s epistemological revolution: rethinking the politics-patronage-social science nexus. Social Studies of Science 31, 171–206.

Szanton, D.L., 2004. Introduction: the origin, nature, and challenges of area studies in the United States. In: Szanton, D.L. (Ed.), The Politics of Knowledge: Area Studies and the Disciplines. University of California Press, Berkeley, pp. 1–33.

Wallerstein, I., 1996. Open the Social Sciences: Report of the Gulbenkian Commission on the Restructuring of the Social Sciences. Stanford University Press, Stanford.

Westad, O.A., 2007. The Global Cold War: Third World Interventions and the Making of Our Times. Cambridge University Press, Cambridge; New York.

Zhuk, S.I., 2009. Nikolai Nikolaevich Bolkhovitinov, 1930–2008. Slavic Review 68, 225–226.

Relevant Website

http://criticalasianstudies.org/about-us/bcas-founding-statement.html – Bulletin of Concerned Asian Scholars (BCAS) Founding Statement.

Areal Linguistics

Lyle Campbell, University of Hawai'i Mānoa, Honolulu, HI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Areal linguistics is defined. A few of the better known linguistic areas of the world are surveyed for the features that characterize them. Issues concerning how areally diffused features are identified, how linguistic areas are established, and what impact areal linguistics has on other aspects of historical linguistics are addressed. Conceptual problems are considered.

Introduction: Defining the Concept

Areal linguistics is about the diffusion of structural features across the languages of a geographical area. The term *linguistic area* refers to a geographical area in which, due to borrowing and language contact, languages of a region have come to share structural features – not just loanwords, but also shared phonological, morphological, syntactic, and other traits. Alternative names sometimes used to refer to a linguistic area are *Sprachbund*, *diffusion area*, *convergence area*, *adstratum*, and *linguistic league*. The key characteristic of a linguistic area is the existence of structural similarities shared among languages of a geographical area, where usually some of the languages are genetically unrelated or at least not all are closely related to one another. It is assumed that the reason the languages of the area share these traits is because they are borrowed.

There have been two approaches to establishing linguistic areas. The more common *circumstantialist* approach lists similarities found in the languages of a geographical area, allowing the list of traits to suggest diffusion, but typically not seeking the specific historical linguistic evidence that could demonstrate that the traits are indeed diffused. The circumstantialist method has been criticized, since it does not eliminate chance, universals, and possibly undetected genetic relationships as alternative possible explanations for the traits shared among the languages. The *historicist* approach seeks concrete evidence showing that the shared traits are diffused. The historicist strategy is preferred because it is more rigorous and reliable, although the lack of clear historical evidence in many cases makes reliance on the circumstantialist approach necessary in many situations.

Linguistic areas are often defined, surprisingly, by a rather small number of shared linguistic traits.

Examples of Linguistic Areas

In what follows, some of the best known linguistic areas are inspected briefly together with the more important of the generally accepted defining traits shared by the languages of each area. A look at these linguistic areas gives a good foundation for understanding some of the issues involving the criteria employed for defining linguistic areas, taken up later below.

The Balkans

The Balkans is the best known linguistic area. The languages of the Balkans area are: Greek, Albanian, Serbo-Croatian, Bulgarian, Macedonian, and Romanian (to which some scholars also add Romani and Turkish). Some salient traits of the Balkans linguistic area are

1. a central vowel/ɨ/(or/ə/) (not present in Greek or Macedonian);
2. syncretism of dative and genitive cases (dative and genitive merged in form and function), illustrated by Romanian *fetei* 'to the girl' or 'girl's,' as in *am data o carte fetei* 'I gave the letter *to the girl*,' and *frate fetei* 'the *girl's* brother';
3. postposed articles (not in Greek), as for example, in Bulgarian *mə^{j}ət* 'the man,' compare *mə^{j}* 'man';
4. periphrastic future (with an auxiliary verb corresponding to 'want' or 'have,' not in Bulgarian or Macedonian), as in Romanian *voi fuma* 'I will smoke' (literally 'I want (to) smoke') and *am a cínta* 'I will sing' (literally 'I have sing');
5. periphrastic perfect (with an auxiliary verb corresponding to 'have');
6. absence of infinitives (rather with constructions equivalent to, for example, 'I want that I go' for 'I want to go');
7. double marking of animate objects by use of a pronoun copy, as in Romanian *i-am scris lui Ion* 'I wrote to John,' literally 'to.him-I wrote him John,' and Greek *ton vlépo ton jáni* 'I see John,' literally 'him.Acc I see him.Acc John.' (Joseph, 1992; Sandfeld, 1930; Schaller, 1975)

South Asia (The Indian Subcontinent)

The South Asian linguistic area is composed of languages belonging to the Indo-Aryan (branch of Indo-European), Dravidian, Munda, and Tibeto-Burman (branch of the Sino-Tibetan) language families. Some traits shared among different languages of the area are

1. retroflex consonants, particularly retroflex stops;
2. absence of prefixes (accept in Munda);
3. presence of a dative-subject construction (that is, dative-experiencer, as in Hindi *mujhe maaluum thaa* 'I knew it' (*mujhe* 'to me' + know + Past));
4. Subject-Object-Verb (SOV) basic word order, including postpositions;
5. absence of a verb 'to have';

6. 'conjunctive or absolutive participles' (tendency for subordinate clauses to have nonfinite verbs, that is, participles, and to be preposed; for example, relative clauses precede the nouns they modify);
7. morphological causatives (causatives signaled by verbal suffixes);
8. so-called 'explicator compound verbs' (where a special auxiliary from a limited set is said to complete the sense of the immediately preceding main verb, and the two verbs together refer to a single event, as for example Hindi *le jaanaa* 'to take (away)' ('take' + 'go'));
9. sound symbolic forms based on reduplication, often with *k* suffixed (for example in Kota, a Dravidian language: *kad-kadk* '(heart) beats fast with guilt or worry'; *a:nk-a:nk* 'to be very strong (of man, bullock), very beautiful (of woman)').

Some of these areal features are not limited to the Indian subcontinent, but can be found also in neighboring languages (for example, SOV basic word order is found throughout much of Eurasia and Northern Africa) and in languages of many other parts of the world. Some of the traits are not necessarily independent of one another, for example, languages with SOV basic word order tend also to have nonfinite (participial) subordinate clauses, especially for relative clauses, and not to have prefixes (Emeneau, 1956, 1980, 2000; Masica, 1976).

Mesoamerica

The language families (including language isolates) which make up the Mesoamerican Linguistic Area are Nahua (branch of Uto-Aztecan), Totonacan, Otomanguean, Mixe-Zoquean, Mayan, Xinkan, Tarascan, Cuitlatec, Tequistlatecan, and Huave. Five areal traits are shared by nearly all Mesoamerican languages, but not by neighboring languages beyond this area, and these are considered particularly diagnostic of the linguistic area:

1. nominal possession of the type *his-dog the man* 'the man's dog,' as in Pipil (Uto-Aztecan): *i-pe:lu ne ta:kat,* literally 'his-dog the man';
2. relational nouns (locative expressions composed of noun roots and possessive pronominal affixes), of the form, for example, *my-head* for 'on me,' as in Tz'utujil (Mayan): (*č-r-i:x* 'behind it, in back of it,' composed of *č-* 'at, in,' *r* 'his/her/its,' and *i:x* 'back,' contrasted with *č-w-i:x* 'behind me,' literally 'at-my-back');
3. vigesimal numeral systems based on combinations of 20, such as that of Chol (Mayan): *hun-k'al* '20' (1 × 20), *ča?-k'al* '40' (2 × 20), *uš-k'al* '60' (3 × 20), *ho?-k'al* '100' (5 × 20), *hun-bahk'* '400' (1-*bahk'*), *ča?-bahk'* '800' (2 × 400), etc.;
4. nonverb-final basic word order (that is, generally no SOV languages) – although Mesoamerica is surrounded by languages both to the north and south which have SOV basic word order, languages within the linguistic area have VOS, VSO, or SVO basic order, not SOV;
5. a number of calques (loan translation compounds) shared by the Mesoamerican languages; these include, for example, 'boa' = 'deer-snake,' 'egg' = 'bird-stone/bone,' 'lime' = 'stone(-ash),' 'knee' = 'leg-head,' and 'wrist' = 'hand-neck.'

Since these five traits are shared almost unanimously throughout Mesoamerican languages but are generally not found in the languages just beyond the borders of Mesoamerica, these traits are considered strong evidence supporting the validity of Mesoamerica as a linguistic area. Four of these five traits have essentially the same distribution, clustering at the borders of Mesoamerica. Such bundling is uncommon in linguistic areas.

A large number of other features are shared among several Mesoamerican languages, but some are not found in all the languages of the area, while others are found also in languages beyond the borders of the area (for details see Campbell et al., 1986).

The Northwest Coast of North America

The Northwest Coast, the best known of linguistic areas in the Americas, includes Tlingit, Eyak, the Athapaskan languages of the region, Haida, Tsimshian, Wakashan, Chimakuan, Salishan, Alsea, Coosan, Kalapuyan, Takelma, and Lower Chinook. These languages are characterized by elaborate systems of consonants, which include series of glottalized stops and affricates, labiovelars, multiple laterals, and uvular stops. There are typically few labial consonant (labials are completely lacking in Tlingit and Tillamook and are quite limited in Eyak and most Athabaskan languages); in contrast, there are several uvular consonants in most of these languages. There are typically few vowels, only three (*i, a, o,* or *i, a, u*) in several of the languages, four in others. Several of the languages have pharyngeals (ħ, ʕ), and most have glottalized resonants and continuants. Shared morphological traits include extensive use of suffixes and the near absence of prefixes; reduplication (of several sorts, signaling iteration, continuative, progressive, plural, collective, distribution, repetition, diminutive, etc.); numeral classifiers; alienable/inalienable oppositions in nouns; evidential markers in verbs; and verbal locative-directional markers; masculine/feminine gender (shown in demonstratives and articles); visibility/invisibility opposition in demonstratives. Aspect is more important than tense. All except Tlingit have passivelike constructions. The negative appears as the first element in a clause regardless of the usual word order. Overt marking of nominal plurals is absent or limited, as in many American Indian languages. Northwest Coast languages also have lexically paired suppletive singular and plural verb stems (that is, a lexical root may be required with a plural subject which is entirely different from the root used with a singular subject).

Some other traits are shared by a smaller number of Northwest Coast languages (see Campbell, 1997: 333–334; cf Sherzer, 1976).

The Baltic

The Baltic linguistic area includes at its core (Balto-)Finnic languages, especially Estonian and Livonian (members of the Finno-Ugric branch of the Uralic family); Baltic languages (Indo-European); and Baltic German (also Indo-European). However, all of the following have been included in different treatments of the Baltic linguistic area: Old Prussian (extinct), Lithuanian, Lavian (Baltic languages, a branch of Indo-European); the 10 Saami (Lapp) languages, Finnish, Estonian, Livonian, Votian, Vepsian, Karelian, and others

(of the (Balto-)Finnic subgroup, a member of the Finno-Ugric branch of the Uralic family); High German, Low German, Baltic German, Yiddish (West Germanic); Danish, Swedish, Norwegian (North Germanic); Russian, Belorussian, Ukranian, Polish, Kashubian (Slavic); Romani (Indo-Aryan); and Karaim (Turkic).

Shared features of the Baltic area include

1. first-syllable stress;
2. palatalization of consonants;
3. tonal contrasts;
4. partitive case/partitive constructions (to signal partially affected objects, equivalent to, for example, 'I ate [some] apple,' in Finnic, Lithuanian, Latvian, Russian, Polish, etc., as for example, in Finnish *söin omena-a* (I.ate apple-PARTITIVE) 'I ate [some] apple');
5. direct objects in the nominative case in a number of constructions which lack overt subjects (Finnic, Baltic, North Russian), as in Finnish *on hyvä syödä omena-Ø* (is good to.eat apple-NOMINATIVE), where it is not possible to substitute *omena-n* (apple-ACCUSATIVE) in this utterance;
6. evidential mood ('John works hard [it is said/reported/inferred]'), in Estonian, Livonian, Latvian, and Lithuanian, as for example in Estonian *ta tege-vat töö-d* (he.NOMINATIVE do-EVIDENTIAL work-PARTITIVE) 'he reportedly is working'/'they say he is working';
7. prepositional verbs (as German *aus-gehen* (out-to.go) 'to go out'), in German, Livonian, Estonian, Baltic, and others;
8. Subject-Verb-Object (SVO) basic word order;
9. agreement of adjectives in number with the nouns which they modify (all languages of the area except Saami languages and Karaim); they also agree in case in all except the Scandinavian languages (which have lost case distinctions for adjectives); they also agree in gender in Baltic, Slavic, Scandinavian, and German, Yiddish, and some others.

For a more complete list of traits attributed to the Baltic linguistic area, see Zeps (1962), Koptjevskaja-Tamm (2002), and especially Koptjevskaja-Tamm and Bernhard Wälchli (2001); cf also Jakobson (1931).

Ethiopia

Languages of the Ethiopian linguistic area include Cushitic (Beja, Awngi, Afar, Sidamo, Somali, etc.), Ethiopian Semitic (Ge'ez, Tigre, Tigrinya, Amharic, etc.), Omotic (Wellamo (Wolaytta), Kefa, Janjero (Yemsa), etc.), Anyuak, Gumuz, and others. Among the traits they share are

1. SOV basic word order, including postpositions;
2. subordinate clause preceding main clause;
3. gerund (nonfinite verb in subordinate clauses, often inflected for person and gender);
4. a 'quoting' construction (a direct quotation followed by some form of 'to say');
5. compound verbs (consisting of a nounlike 'preverb' and a semantically empty auxiliary verb);
6. negative copula;
7. plurals of nouns are not used after numbers;
8. gender distinction in second and third person pronouns;
9. reduplicated intensives;
10. different present tense marker for main and subordinate clauses;
11. the form equivalent to the feminine singular is used for plural concord (feminine singular adjective, verb, or pronoun is used to agree with a plural noun);
12. a singulative construction (the basic nonaffixed noun may be a collective or plural and it requires an affix to make a singular);
13. shared phonological traits such as *f* but no *p*, palatalization, glottalized consonants, gemination, presence of pharyngeal fricatives (ħ and ʕ) (Ferguson, 1976; Thomason, 2001; cf Tosco, 2000).

Several other Linguistic Areas have been proposed, some widely recognized, others with less accepted. These include Amazonia, Ancient Anatolia, Andean Area, Australia, Caucasus, the Cape (Africa), Chaco and the Southern Cone (South America), Central Eurasia, Clear Lake Area (California), Colombian–Central American Area, East Nusantara and Bird's Head (Papua New Guinea), Ecuadoran–Colombian subarea, Ethiopian Highlands, Europe (Standard Average European Area), Great Basin (US), the Guaporé-Mamoré area (South America), the Kalahari Basin, Macro-Sudan Belt, Mainland Southeast Asia, Northeast Area (North America), Northern Asia, Northern California, Northern Northwest Coast (North America), Orinoco-Amazon Area, Plains Area (North America), Plateau Area (North America), Pueblo Area, Qinghai-Ganzu area (China), Sepik River Area (Papua New Guinea), Siberia, South Coast Range (California), Southeast Area (North America), Southern Africa, Southern California and Western Arizona Area, the Tanzanian Rift Valley area, Upper Xingú incipient area (Brazil), the Vaupés (Içana-Vaupés) area (South America), Venezuelan-Antillean Area, among others.

How Are Linguistic Areas Defined?

How linguistic areas are defined has been subject to considerable debate. The following criteria have at times been considered relevant for establishing a linguistic area: (1) the number of traits shared by languages in a geographical area, (2) bundling of the traits in some significant way (for example, clustering at roughly the same geographical boundaries), and (3) the weight or complexity of different areal traits (some accorded more significance for determining areal affiliation on the assumption that it is more difficult for them to be borrowed than others).

The general idea has been that for establishing a linguistic area, the more shared features, the better. Linguistic areas in which many diffused traits are shared among the languages are considered better established. Nevertheless, some scholars believe that even one shared trait may be enough to define a weak linguistic area (Campbell, 1985, 2006). In any event, it is clear that some areas have stronger support than others because their languages share a larger number of traits. In the linguistic areas described above, the number and kind of shared traits vary considerably. Urban (2007) would add 'mutual influence' among the languages to the criteria, which would presumably keep out cases of linguistic areas based on a single diffused trait.

The idea that greater weight should be attributed to some traits than to others for defining a linguistic area can be illustrated with the borrowed word order patterns in the Ethiopian linguistic area. Ethiopian Semitic languages exhibit a number of interconnected word-order patterns that are borrowed from neighboring Cushitic languages. Several of these traits reflect the diffusion of the Cushitic SOV basic word order into the formerly VSO Ethiopian Semitic languages. Typologically the orders Noun-Postposition, Verb-Auxiliary, and Relative Clause-Head Noun are all correlated and tend to co-occur with SOV order cross-linguistically. Their presence in Ethiopian Semitic languages (some of the languages having all of these, others fewer) might seem to reflect several different diffused traits (SOV counted as one, Noun-Postposition as another, and so on), and might be taken as several independent pieces of evidence supporting the linguistic area. However, from the perspective of expected word-order co-occurrences, these word-order arrangements are not independent traits, but reflect the diffusion of a single complex feature, the overall SOV word-order type with its tendency for the various expected coordinated orderings in typologically interrelated constructions to co-occur. However, if borrowed SOV word order is counted as a single diffused areal trait, it may be argued that it should rank high in significance for defining the linguistic area, since it is much more difficult for a language to change so much of its basic structure under areal influence than it is to acquire less complex traits.

Some scholars had thought that the bundling of areal traits, clustering at the boundaries of a linguistic area, might be necessary for defining a linguistic area, though this has proven a poor criterion. Linguistic areas are similar to traditional dialects, where often one trait spreads across more territory than another trait, so that their boundaries (or territories) do not coincide, that is, do not 'bundle.' Typically the geographical extent of individual traits may vary considerably. However, in the rare situation where the areal traits do coincide at a clear boundary, the definition of a linguistic area matching their boundary is relatively secure. As mentioned, several of the traits in the Mesoamerican linguistic area do have the same boundary, though many other traits do not, and typically in other linguistic areas the various areal traits do not share the same geographical boundaries, offering no clearly identifiable outer border of the linguistic areas in question.

Implications of Areal Linguistics for Linguistic Reconstruction and Subgrouping

Areal diffusion can have important implications for comparative reconstruction and for subgrouping within known language families. Nootkan provides a good example which illustrates this. The sound correspondences upon which Nootkan subgrouping is based are given in Table 1.

Table 1 Nootkan sound correspondences

	Makah	Nitinat	Nootka	Proto-Nootkan
1.	b	b	m	*m
2.	b'	b'	m'	m'
3.	d	d	n	*n
4.	d'	d'	n'	*n'
5.	q'	ʕ	ʕ	*q'
6.	Q'^{w}	ʕ	ʕ	*q'^{w}
7.	χ^{w}	χ^{w}	ħ	*χ^{w}
8.	χ	χ	ħ	*χ

Reproduced from Haas, Mary R., 1969. The Prehistory of Languages. Mouton, The Hague.

Nitinat and Makah appear to share the innovation which changed nasals to corresponding voiced stops (in (1)–(4)), while Nitinat and Nootka (Nuu-chah-nulth) appear to share the change of the glottalized uvulars to pharyngeals (in (5) and (6)). Makah and Nitinat also share the retention of uvular fricatives, which Nootka has changed to a pharyngeal (in (7) and (8)); however, shared retentions are not valid evidence for subgrouping. Here, one innovation (denasalization) suggests a subgrouping of Makah-Nitinat together, with Nootka as more distantly related, while the other innovation (pharyngealization) suggests Nitinat-Nootka together, with Makah less closely related. This seeming impasse is solved when we take into account the fact that the absence of nasals is an areal feature shared by several other languages of the area; it diffused into both Makah and Nitinat under areal influence and is thus not real evidence of a shared common development before the languages separated, rather it reached these two languages independently from elsewhere in the linguistic area. The innovation shared by Nitinat and Nootka of glottalized uvulars changing to pharyngeals (in (5) and (6)) is real evidence of subgrouping – a true (nondiffused) shared innovation. So, Nitinat and Nootka together constitute one branch of the family, Makah the other branch. Moreover, with respect to areal implications for reconstruction, if we did not know about the areal diffusion in this case, we might be tempted to reconstruct the voiced stops in Proto-Nootkan, since they occur in more languages than the nasals do, and to postulate a change of these to nasals in Nootka (for (1)–(4)), getting it wrong in this case. Thus, areal linguistic traits can have important implications for classification (subgrouping) and for reconstruction.

Areal Linguistics and Proposals of Distant Genetic Relationship

Similarities among languages that are due to areal diffusion are sometimes mistakenly taken to be evidence of a possible distant genealogical relationship among languages whose classification is in question. The Mosan hypothesis, which proposed a genetic relationship between the Salishan, Wakashan, and Chimakuan language families, illustrates this problem many long-range comparisons have. Several scholars noted structural similarities among languages of these three Northwest Coast language families, but the Mosan hypothesis was not found convincing because much of the evidence turned out to rely on areal traits that are widely borrowed in the Northwest Coast Linguistic Area. Morris Swadesh (1953) presented 16 shared structural similarities in support of Mosan, but most of these are Northwest Coast areal features (some of the traits are also typologically commonplace,

found independently in languages throughout the world), for example:

1. "Extensive use of suffixes."
2. "Nearly complete absence of functioning prefixes in Chimakuan and Wakashan, minor role in comparison to the suffixes in Salish." (Typologically it is not unusual for suffixing languages to lack prefixes.)
3. "Extensive use of stem reduplication, including initial reduplication … and … full stem reduplication."
4. "Aspect, including at least the dichotomy of momentaneous and durative."
5. "Tense is an optional category."
6. "Plural is an optional category."
7. "Dichotomy of non-feminine versus feminine gender shown in demonstratives and articles."
8. "Numeral classifier notions, shown by suffixes."
9. "Two alternate stems for number" (lexically paired distinct singular and plural verb stems).
10. "Lexical suffixes … referring to body parts and other space references."
11. "Predicative use of nouns."
12. "Demonstrative distinctions such as the present versus absent, or visible versus invisible."

As is clear, the traits which Swadesh listed as evidence for the Mosan hypothesis are better explained as the results of diffusion within the Northwest Coast Linguistic Area (see Campbell, 1997 for details).

From this case, it is easy to see why the identification of areal traits is so important in historical linguistics. In this instance, failure to recognize the areal borrowings led to an erroneous proposal of genetic relationship among neighboring language families.

Areal Linguistics and Other Linguistic Concerns

Some implications of areal linguistics for several other areas of linguistics have been addressed in the literature, sometimes only raising new questions, and sometimes helping to provide greater perspective on certain issues. Some of these include linguistic typology, language universals, multilingualism, linguistic prehistory, language ecology, lexical diffusion, dialectology, origins and evolution of language, language classification, diversification and dispersal of the languages, how language changes spread, models of horizontal versus vertical transmission of change, attempted correlations of human genetics and linguistics, explanation and motivation of language change, social interaction (and network theory), discourse, ethnography of communication, etc.

Kinds of 'Linguistic Areas'

It is generally recognized that things that have been called linguistic areas include entities that diverge widely in character and historical backgrounds, depending on marriage patterns, trade, invasion and conquest, prestige, and many sort of social, cultural, political, geographical, and other factors that correlate with diffusion of linguistic features in different regions (Dahl, 2001: 1458; Kuteva, 1998: 308–309; Thomason, 2001: 104). There is discussion in the literature of such different sorts of linguistic areas as incipient areas, only beginning to form and with as yet few shared traits; moribund and decaying areas, where due to many changes after the area was actively formed, fewer traits are currently recognizable among the languages; overlapping areas, where different areas formed on top of or partially overlapping one another at different times for different reasons; multilateral areas (with areal traits spreading from various languages of the region) and unilateral areas (with most traits shared throughout the languages of an area stemming predominantly from one language); areas due to rapid conquest, to population spread and migration (traits moving with movement of speakers), others through home-grown, stay-in-place contact (movement of traits but not of peoples); and disrupted areas with "latecomers, earlier drop-outs, and temporary passers-by" (Stolz, 2002: 265). "In short, the notion 'linguistic area' does not refer a uniform phenomenon, either socially or linguistically" (Thomason, 2001: 115). This array of different kinds of linguistic areas raises questions about whether the concept of 'linguistic area' is justified, whether all these different 'objects' legitimately qualify as 'linguistic areas,' given their very different natures and composition and given the very different circumstances of their birth (and decay). The notion 'linguistic area' offers little upon which these different sorts of linguistic areas can be united other than the fact that they all involve borrowing in some way, but borrowings of different sorts, for different reasons, in different settings, and at different times.

Linguistic Areas versus Borrowing Generally

It is generally acknowledged that linguistic areas are 'notoriously messy,' 'notoriously fuzzy' things (Thomason and Kaufman, 1988: 95; Heine and Kuteva, 2001: 396; Tosco, 2000: 332), and that "what we understand about linguistic areas is depressingly meager" (Thomason, 2001: 99). A common perception is that the term 'linguistic area' is difficult to define (cf Heine and Kuteva, 2001: 409). As Thomason (2001: 99) observes, "linguistics has struggled to define the concept [linguistic area] ever since [Trubetzkoy, 1928], mainly because it isn't always easy to decide whether a particular region constitutes a linguistic area or not." Stolz's (2002: 259) believes "the search for clear-cut definitions [of '*Sprachbund*, linguistics area, and areal type'] has been largely futile and will probably never come to a really satisfying conclusion." In spite of prolonged efforts to define 'linguistic area,' there is no general agreement on its definition, and even for the most widely accepted linguistic areas, such as the Balkans, scholars do not agree wholly on which languages belong to the area, what linguistic traits characterize the area, and what its precise geographical extent is. Campbell (2006) relates this difficulty to the lack of any clear distinction between areal phenomena and borrowing generally. Dahl (2001: 1458) asks

> In the end, we are led to the following more far-going question about the notion of area: to what extent do areas … have a reality of their own and to what extent are they just convenient ways of

summarizing certain phenomena? At the most basic level, linguistic contact relationships are binary: one language influences another. An area is then simply the sum of many such binary relationships.

Campbell (2006) argues that the various definitions of 'linguistic area' offered in the literature confirm that linguistic areas amount to no more than the study of linguistic borrowing in particular languages and its history. Every 'linguistic area,' to the extent that the notion has any meaning at all, arises from an accumulation of individual cases of 'localized diffusion'; it is the investigation of these specific instances of diffusion, and not the pursuit of defining properties for linguistic areas, that increases our understanding and explains the historical facts. With focus rather on specific instances of borrowing, and not on particular matters of geography, many of the unresolved issues and indeterminacies which have dogged areal linguistics from the outset cease to be relevant questions. It is the linguistic changes due to diffusion that count and not the attempt to seek meaning in the geography that secondarily is involved. A linguistic area, to the extent that it may have a legitimate existence at all, is merely the sum of borrowings in individual languages in contact situations. If we focus rather on understanding borrowings, on those historical events, the difficulty of determining what qualifies as a legitimate linguistic area is no longer a problem.

See also: Linguistic Typology; Linguistics: Overview.

Bibliography

Aikhenvald, Alexandra Y., Dixon, R.M.W. (Eds.), 2001. Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics. Oxford University Press, Oxford.

Campbell, Lyle, 1985. Areal linguistics and its implications for historical linguistic theory. In: Fisiak, Jacek (Ed.), Proceedings of the Sixth International Conference of Historical Linguistics. John Benjamins, Amsterdam, pp. 25–56.

Campbell, Lyle, 1997. American Indian Languages: The Historical Linguistics of Native America. Oxford University Press, Oxford.

Campbell, Lyle, 2004. Historical Linguistics: An Introduction, second ed. Edinburgh University Press, Edinburgh. MIT Press, Cambridge, MA.

Campbell, Lyle, 2006. Areal linguistics: a closer scrutiny. In: Matras, Yaron, McMahon, April, Vincent, Nigel (Eds.), Linguistic Areas: Convergence in Historical and Typological Perspective. Palgrave Macmillan, Houndmills, Basingstoke, Hampshire, pp. 1–31.

Campbell, Lyle, Kaufman, Terrence, Smith-Stark, Thomas, 1986. Mesoamerica as a linguistic area. Language 62, 530–570.

Dahl, Östern, 2001. Principles of areal typology. In: Haspelmath, Martin, König, Ekkehard, Oesterreicher, Wulf, Raible, Wolfgang (Eds.), Language Typology and Language Universals: An International Handbook, vol. 2. Mouton de Gruyter, Berlin, pp. 1456–1470.

Emeneau, Murray B., 1956. India as a linguistic area. Language 32, 3–16.

Emeneau, Murray B., 1980. Language and Linguistic Area, Essays by Murray B. Emeneau (Selected and Introduced by Anwar S. Dil). Stanford University Press, Stanford.

Emeneau, Murray B., 2000. Linguistics in India: past and future. In: Lockwood, David G., Fries, Peter H., Copeland, James E. (Eds.), Functional Approaches to Language, Culture, and Cognition: Papers in Honor of Sydney B. Lamb. John Benjamins, Amsterdam, pp. 545–554.

Ferguson, Charles, 1976. The Ethiopian language area. In: Bender, M.L., Bowen, J.D., Cooper, Robert, Ferguson, Charles, A. (Eds.), Language in Ethiopia. Oxford University Press, Oxford, pp. 63–76.

Haas, Mary R., 1969. The Prehistory of Languages. Mouton, The Hague.

Heine, Bernd, Kuteva, Tania, 2001. Convergence and divergence in the development of African languages. In: Aikhenvald, Alexandra Y., Dixon, R.M.W. (Eds.), Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics. Oxford University Press, Oxford, pp. 393–411.

Jakobson, Roman, 1931. Über die phonologischen Sprachbünde. Travaux du Cercle Linguistique de Prague 4, 234–240.

Joseph, Brian, 1992. Balkan languages. In: Bright, William (Ed.), International Encyclopedia of Linguistics, vol. 1. Oxford University Press, Oxford, pp. 153–155.

Koptjevskaja-Tamm, Maria, 2002. The Circum-Baltic languages – a contact-superposition zone in the European periphery. In: Ramat, Paolo, Stolz, Thomas (Eds.), Mediterranean Languages: Papers from the MEDTYP Workshop, Tirrenia, June 2000. Brockmeyer, Bochum, pp. 209–222.

Koptjevskaja-Tamm, Maria, Wälchi, Bernhard, 2001. The Circum-Baltic languages: an areal-typological approach. In: Dahl, Östen, Koptjevskaja-Tamm, Maria (Eds.), Circum-Baltic Languages. John Benjamins, Amsterdam, pp. 615–750.

Kuteva, Tania, 1998. Large linguistic areas in grammaticalization: auxiliation in Europe. Language Sciences 20, 289–311.

Masica, Colin P., 1976. Defining a Linguistic Area: South Asia. University of Chicago Press, Chicago.

Masica, Colin P., 1992. Areal linguistics. In: Bright, William (Ed.), International Encyclopedia of Linguistics. Oxford University Press, Oxford, pp. 108–112.

Sandfeld, Kristian, 1930. Linguistique balkanique: Problemes et resultants. Champion, Paris.

Schaller, Helmut W., 1975. Die Balkansprachen: Eine Einführung in die Balkanphilologie. Winter, Heidelberg.

Sherzer, Joel, 1976. An Areal-Typological Study of American Indian Languages North of Mexico. North Holland Publishing Co., Amsterdam.

Stolz, Thomas, 2002. No sprachbund beyond this line! on the age-old discussion of how to define a linguistic area. In: Ramat, Paolo, Stolz, Thomas (Eds.), Mediterranean Languages: Papers from the MEDTYP Workshop, Tirrenia, June 2000. Brockmeyer, Bochum, pp. 259–281.

Swadesh, Morris, 1953. Mosan I: a problem of remote common origin. International Journal of American Linguistics 19, 26–44.

Thomason, Sarah G., 2001. Language Contact: An Introduction. Edinburgh University Press, Edinburgh.

Thomason, Sarah G., Kaufman, Terrence, 1988. Language Contact, Creolization, and Genetic Linguistics. University of California Press, Berkeley.

Tosco, Mauro, 2000. Is there an Ethiopian language area? Anthropological Linguistics 42, 329–365.

Trubetzkoy, Nikolai S., 1928. Proposition 16. Acts of the First International Congress of Linguists, 17–18. Leiden.

Urban, Mateusz, 2007. Defining the linguistic area/league: an invitation to discussion. Studia Linguistica Universitatis Iagellonicae Cracoviensis 124, 137–159.

Zeps, Valdis, 1962. Latvian and Finnic Linguistic Convergences. Indiana University Press, Bloomington.

Arendt, Hannah (1906–75)

Antonia Grunenberg, Berlin, Germany

Published by Elsevier Ltd.
This article is reproduced from the previous edition, volume 2, pp. 733–737, © 2001, Elsevier Ltd.

Abstract

Hannah Arendt, American political thinker of German origin, has recently received increasing attention from the academic community as well as from a broader public. This can be attributed to various developments. First, the end of communism, the collapse of the Berlin Wall, and the fading authority of Marxism after 1989 have generated the need for a different understanding of the world. Second, the ongoing debates in Central European countries about the legacy of totalitarianism have rekindled the interest in Arendt's approach to the analysis of totalitarianism. Third, the growing difficulty of Western political systems to integrate various social groups and the upcoming discussion on renewing democracy have given rise to basic questions about how to revive the political sphere. And last but not least, Arendt's concepts of 'freedom' and 'the political' have again given fresh impetus to political theory and to the debate in the political public. Thus, Arendt's way of reflecting the political dimension of democracy fills a void in Western theory and political thinking, which did not arise until after the end of the Cold War. At the same time, her views are increasingly considered in the ongoing debate on establishing democracy in Central and Eastern Europe. Arendt stressed the importance of political thinking and political theory in modern times. She was deeply influenced by the experience of the destructive power in both German National Socialism and Stalinist Communism, but her work cannot be categorized easily. Her notions can be classed within the field of philosophy as well as with historiography and political theory. But she does not object to being described as a political theorist in the sense of public political thinking. Throughout her life she was engaged in political thinking. Feeling the need to meet the intellectual challenge which came up with the self-destruction of European democracies after World War I that led to the strengthening of Stalinism and National Socialism, she published her first famous book *Origins of Totalitarianism* in 1951. Her last work, *The Life of the Mind*, was left unfinished, and was not published until two years after her death in 1975. The objective she pursued was to stimulate a way of human thinking, judging, and acting which orients toward renewing and protecting the political community of the citizen.

Born on 14 October 1906 in Hannover, Arendt grew up in a liberal Jewish family in Koenigsberg. Later, she studied philosophy, protestant theology, and Greek philology in Marburg, Heidelberg, and Freiburg. During this time, she got to know and was greatly influenced by two philosophers: Martin Heidegger and Karl Jaspers. In 1928 she received her Ph.D. in Heidelberg for her thesis on 'The concept of love in the work of St. Augustin.' Understanding the implications of Hitler's accession to power, she left Germany in 1933. She temporarily settled in Paris where she volunteered for a Jewish refugee organization. In 1941 she emigrated to the USA, and became an American citizen 10 years later. She worked as a lecturer at various American universities and colleges, and was a well-known freelance writer who contributed to the American intellectual culture as well as to the postwar European political discourse. Arendt died on 4 December 1975 in New York.

Arendt's Methodological Approach to Theory and Thinking

One cannot comprehend Arendt's thought without taking into account her personal experience of National Socialism and the genocide of the European Jewry. Throughout her life, Arendt always pointed to the existence of National Socialist and Stalinist terror and to concentration camps, which she considered the utmost challenges to political thinking.

It appears that she interlinked all her concepts and categories as well as the history of thought on which she based her notions with the experience of totalitarianism. She considered totalitarianism not only to be a breach in modern civilization but was even convinced that totalitarian rule challenges modern political thinking. This view gave rise to her seemingly unscientific methodological approach to thinking: How to create a world to which mankind feels attached and in which the individual does not lose their ability to judge? How to protect the 'body politic' from self-destruction? In this context, the concept of 'understanding the world' seems to become the overarching hermeneutic idea which interrelates all other categories: judging, acting, being in the public, etc.

Arendt focuses on reconsidering the instruments applied in social sciences. She believes that these were reduced to an absurdity by totalitarianism (see also Lefort, 1988, p. 48). Her considerations encompass a wide range of historical, philosophical, and political references. Her concept of 'the world,' and her radicalization of hermeneutics by questioning the self as a subject have political and philosophical connotations.

There are some specific methodological aspects that need to be considered to understand Arendt's way of thinking:

1. Historically Arendt focuses on the disintegration of the public sphere in the European societies (e.g., France and Germany but also Russia) at the end of the nineteenth century. Her point of reference is the emergence of a new form of total domination in the 1930s, which basically differs from former types of tyranny or dictatorship but refers to the concept of natural law and/or of dialectical law. For Arendt, the need to understand totalitarianism epistemologically arises from the fact that totalitarianism cannot be sufficiently explained by historical and social sciences

(Arendt, 1994, p. 317). She argues that totalitarian rule has had a major impact on the categories applied in social sciences (Arendt, 1994, p. 318). She criticizes social sciences for assuming that political action is basically a rational process (in the sense of Max Weber, *see* Weber, Max (1864–1920)). But totalitarian terror is, in Arendt's view, not at all rational but meaningless and contingent. Therefore, it is not possible to grasp the absolute meaninglessness of destruction, evil, and mass murder by adopting social-science approaches. This is the reason why Arendt's hermeneutic concept is closely interlinked with her criticism of social sciences.

2. Arendt aims to 're-open' the political dimension of modern political thinking by applying her concept of understanding. In this context, 'reopening' means that a dimension, which has been lost, has to be rediscovered: the public sphere as a sphere of human interaction. It is the public sphere that keeps the body politic (the political system) alive. In Arendt's view, the removal of institutions and legal systems as well as the abolition of political freedom under totalitarian rule are direct consequences of the dissolution of the public sphere in modern times.
3. In terms of philosophical categories, understanding is taken to be different from cognition. Understanding is not simply an intellectual concept of the world but is a hermeneutic process, and thus also a philosophical and emotional approach to the world. Understanding is interlinked with experience and judgement.
4. Understanding is also supposed to be different from having 'correct information and knowledge,' although the two are interrelated. Understanding also differs from scientific knowledge. In Arendt's critical view knowledge is meaningless without preliminary understanding. Here, she breaks with the tradition of natural and social sciences according to which the knowledge about facts and details of a phenomenon and putting these pieces together in systematical order means to recognize these facts and details or to realize their truth.

Furthermore, 'understanding' does not mean 'causal and historicist explanation': 'The necessity which all causal historiography consciously or unconsciously presupposes does not exist in history. What really exists is the irrevocability of the events themselves, whose poignant effectiveness in the field of political action does not mean that certain elements of the past have received their final, definite form, but that something inescapably new was born' (Arendt, 1994, p. 326).

In this way, Arendt again draws a line between political thinking and the world of scientific analysis. Her approach denies all teleological and deterministic as well as deductive thinking. However, it includes a fundamental criticism of the concept of progress and rationality in modern history.

Methodologically, her work is challenging insofar as she opposes the traditional (usual) meaning of concepts. In all of her greater works she critically reviews concepts and conceptions as facets of the history of thinking; she eventually tries to re-establish an authentic relation between a given concept, reality and the acting individual; furthermore she seeks to reveal what is new about the object of recognition. Thus, her work is also a major contribution to hermeneutics in social sciences.

Basic Concepts in Arendt's Thinking

Arendt did not evolve a systematic political theory nor did she elaborate a philosophical system. However, her approach to political thinking remains a challenge for both political theory and philosophy. It is based on a different understanding of the major concepts of political theory. Her critical view of the techniques applied in social sciences induced Arendt not only to do research on etymological changes in social science concepts (e.g., freedom, morality) but also to relate her own concept to the overarching idea how to reveal the political dimension in the community of citizens.

The Concept of Society

Society is considered a never-ending process in which citizens act and judge towards an open future within a public sphere; it is not perceived as a set of institutions and ways of living. Arendt opposes traditional views of theory on society by denying that society is to be seen as a (national) entity.

In her view institutions have a protective function. However, they can also pose a threat to the 'humanness of the world.' In this respect, Arendt follows Max Weber in his criticism of bureaucracy. She agrees that the public realm cannot do without institutions, but here again her concept of institution is different, for example, the act of setting up a 'political body' during a revolution is taken to be an institution by itself.

The 'humanness' is supposed to be maintained only by the citizens. Thus, it may not be surprising that Arendt relates her basic concepts to certain abilities of citizens: thinking, judging, and acting in the public realm.

The Political

Arendt's concept of 'the political,' which is basically different from 'politics,' is based on her distinction between the public sphere and the needs and desires dominating the private sphere.

The political is not a substance of humankind but arises in the midst of citizens in the public realm. This is another example of the way in which Arendt relates her concepts to the sphere of interaction. In her view, the political gives rise to a new beginning in the sphere of interaction.

Arendt's way of thinking can be closely related to the tradition of political thinking which emerged in Greek and Roman antiquity. She acknowledged the Greek 'polis' and the Roman republic as historical forms of political sphere, to which die American founding fathers referred to perceiving the republic as a permanent founding of itself (self-instituting) within a system of checks and balances.

Political Freedom

The term 'freedom' (like 'society') refers to a process of interaction in the political sphere and not primarily to institutions. Freedom is not restricted to the individual or to the private sphere. Moreover, 'political freedom is distinct from philosophic freedom in being clearly a quality of the I-can and not of the I-will. Since it is possessed by the citizen rather than by man in general, it can manifest itself only in

communities, where the many who live together have their intercourse both in word and in deed regulated by a great number of *rapports* – laws, customs, habits, and the like. In other words, political freedom is possible only in the sphere of human plurality, and on the premise that this sphere is not simply an extension of the dual I-and-myself to a plural We' (Arendt, 1981a, p. 200).

Basically, political freedom appears to have a double meaning: on the one hand there is the act of setting up a 'political body' (e.g., by means of a constitution) and on the other hand there are the *rapports* of citizens in favor of their political body.

Power

As with the concept of freedom to which it is linked, power is interrelated to the public realm. Arendt does not interlink power with domination, and violence. She disagrees with the definition of power given by major social scientists and especially by Max Weber (Weber, 1978, p. 53).

In Arendt's thinking power, that is, political power, cannot be 'possessed' by individuals nor – as this was assumed during the French Revolution – by 'the people.' On this point Arendt differs from the liberal as well as from the revolutionary concept of power. 'Power corresponds to the human ability, not just to act in concert. Power is never the property of an individual; it belongs to a group and remains in existence only so long as the group keeps together.'

This means that power belongs to the public sphere. It can be incorporated by means of a revolution, a political uprising or an act of founding. However, if power is institutionalized it tends to become bureaucratic. In this respect Arendt agrees with the skepticism of Max Weber about modern bureaucracy.

Thus, power comes up and disappears again depending on the contingency of history.

'Alienation from the World' and 'Being in the World'

Arendt experienced emigration, exile and the deplorable status of a stateless person over many years. It is no accident that the issue of alienation plays a major role in her works. But again, alienation is not related to the private realm but to the public sphere. Alienation refers to the estrangement of individuals from their political community. According to Arendt's theory of humanness, individuals have to 'appear' in the world, join their fellow men and commit themselves to the preservation of the world they live in through their actions. This is what 'citizen' means.

In this context, 'world' (and opposed to 'being without a world'/worldlessness) refers to the fundamental fact of being born into an already existing world.

However, Arendt's concept of alienation differs from the one of Marx or other philosophical authors insofar as it focuses not on the individual but on the relationship between the citizens and the world which they have created (*see* Marx, Karl (1818–83)).

This is the starting point from which Arendt pursues two objectives: to reflect upon how to make the world worldly and to humanize thinking. This is a constant theme throughout her work.

Totalitarianism

All of Arendt's works on political thinking refer to the political events of the time. Her research on totalitarianism during the 1940s focuses on various aspects.

1. She identifies totalitarianism as a new type of total domination. It bases on violent antidemocratic movements which later become part of a political system which engages in the control of institutions and the personal lives of the people, sets up a specific ideology, and establishes systematic terror.
2. She focuses on the 'The origins of totalitarianism,' that is, imperialism and anti-Semitism as the major sources of destroying the political body of the European democracies since the nineteenth century.
3. She considers totalitarianism as representing a breach in modern civilization which began a long time ago. Here again she argues that the categories of thinking are deeply affected by this process.

Arendt considers totalitarian rule as a possible outcome of the weaknesses of modern age. It is bewildering that modern democracies themselves may generate totalitarian elements. Hence, in Arendt's view totalitarian rule cannot be dismissed as an 'accident in history'; there is always the possibility that this form of rule emerges.

This is why it is problematic to discuss moral behavior in a traditional sense after the emergence of totalitarian rule has occurred. Wherever human nature was reduced to the level of 'material' it seems inappropriate to simply re-establish rules that were violated or manipulated. For Arendt, it is important to reflect on the dimensions of a world which gave rise to such a development and which continues to exist after the breakdown of totalitarianism.

Radical and Banal Evil

One word within Arendt's concept of totalitarianism generated major criticism: 'evil.' Arendt distinguishes between radical and banal evil. 'Radical evil' makes human beings superfluous. Arendt uses this concept to find an explanation for Nazi death camps. 'Radical evil' stands for both an absolute and a contingent negation of humanness. The other side of evil is its banality. Arendt takes the Nazi functionary Adolf Eichmann as an example. Her characterization of Eichmann as 'banal' has been widely misinterpreted as a diminution of his personal responsibility (Arendt, 1963). But what Arendt really meant has more to do with another dimension of evil: the absence of thinking (that means, of self-reflection and of conscience). Thus, Arendt presents Eichmann as a specific type of modern person who has lost their relationship to the world as a 'common good.'

Arendt as a Political Thinker of Morality

At first sight it appears strange to describe Arendt's way of thinking as moral (ethical) because Arendt does not evolve a theory of 'acting correctly.' Neither does she construct a system of values. She does not apply moral standards to judge political thinking and acting. Moreover, she deconstructs the

popular meaning of morality by uncovering its etymological roots. It turns out that 'morality' has more to do with being committed to the world than with internalizing values.

Seen from her perspective, acting within the political sphere means to establish civil (civic) manners; this can only be achieved if political action is related to 'home' which may also be called 'the world people share.'

Arendt's principal work 'The origins of totalitarianism' concludes by referring to Heidegger's philosophy. There is a quasi-existential fixed point which is inaccessible to totalitarian rule: the fact of natality. The simple fact that people are born cannot be removed; the birth of a child will continue to represent a new beginning for the world. 'Beginning, before it becomes a historical event, is the supreme capacity of man; politically it is identical with man's freedom… This beginning is guaranteed by each new birth; it is indeed every man' (Arendt, 1981b, p. 479).

Hence the basic prerequisite for human action is there; it is one of the fundamentals of human existence. The question that arises is: How can this existential potentiality of political action be given shape?

Acting politically is only possible in a public sphere. This sphere provides the common world people share in which thinking in public and acting take place. Citizens have to see to it that this public sphere can be renewed. At this point Arendt differs from Kant's theory of morally acting. Unlike Kant, Arendt does not perceive 'polical community' as a limited place, that is, a city (polis) but as a never-ending process of taking action in a sphere that can only be protected by the citizens themselves. The citizens' concerns arise from a joint interest in the world they live in, a world, which keeps on giving new beginning to them.

Arendt considers a renewal of the public sphere to be the only possibility to counteract the (self-)destructive potential of modern times. The conclusion that can be drawn after experiencing totalitarian rule is that evil in the world can only be faced if people commit themselves to making the world they share fit to live in for their fellow citizens and if they permanently renew this 'habitability.'

The difficulty which has arisen is that the distinction between public and private sphere becomes increasingly blurred: it was intentionally destroyed by totalitarian rule, and similarly, it is jeopardized by the increasing importance of social interests and needs in modern society. Nevertheless Arendt argues, a world that is shared by people can only emerge by creating a public sphere. However, this can only be achieved to the extent to which citizens succeed in agreeing on making freedom the overall objective of their actions.

See also: Antisemitism; Balance of Power, History of; Citizenship: Political; Civil Liberties and Human Rights; Civil Society and the Public Sphere: History of the Concept; Civil Society, Concept and History of; Democracy, History of; Democratic Theory; Dictatorship; Freedom: Political; Genocide, Historical Aspects of; Marx, Karl (1818–83); National Socialism and Fascism; Political Theory, History of; Power in Society; Power; Public Interest; Public Sphere: Eighteenth-Century History; Social Work Ethics; Totalitarianism; Weber, Max (1864–1920).

Bibliography

Arendt, H., 1958. The Human Condition. University of Chicago Press, Chicago (1960 Vita Activa oder Vom tätigen Leben. Kohlhanno, Stuttgart, Germany).

Arendt, H., 1963. Eichmann in Jerusalem. A Report on the Banality of Evil. Viking, New York (1964 Eichmann in Jerusalem. Ein Report von der Banalität des Bösen. Piper, München, Germany).

Arendt, H., 1973. The Origins of Totalitarianism. Harcourt Brace Jovanovic, New York (1980 Elemente und Ursprünge totaler Herrschaft. Piper, München, Germany).

Arendt, H., 1981a. On violence. In: The Crisis of the Republic. Harcourt Brace, San Diego, CA (1990 Macht und Gewalt. Piper, München, Germany).

Arendt, H., 1981b. The Life of the Mind. Harcourt Brace, New York (1979 Vom Leben des Geistes. Piper, München, Germany).

Arendt, H., 1994. In: Kohn, J. (Ed.), Essays in Understanding. Harcourt Brace, New York.

Bernstein, R., 1996. Hannah Arendt and the Jewish Question. MIT Press, Cambridge, MA.

Canovan, M., 1992. Hannah Arendt. A Reinterpretation of Her Political Thought. Cambridge University Press, New York.

Forti, S., 1994. Vita della mente e tempo della polis. Hannah Arendt tra filosofia e politica. Angeli, Milan.

Lefort, C., 1988. Democracy and Political Theory (D. Macey, Trans.). University of Minnesota Press, Minneapolis, MN.

Villa, D.R., 1996. Arendt and Heidegger. The Fate of the Political. Princeton University Press, Princeton, NJ.

Weber, M., 1978. In: Roth, G., Wittich, C. (Eds.), Economy and Society. University of California Press, Berkeley, CA (1980 Wirtschaft und Gesellschaft. Mohr, Tübingen, Germany).

Young-Bruehl, E., 1982. Hannah Arendt. For Love of the World. Yale Unversity Press, New Haven, CT (1986 Hannah Arendt. Leben und Werk. Fisher Frankfurt-an-Main, Germany).

Aristocracy, Nobility, and Gentry, European History of

Antoni Mączak, University of Warsaw, Poland

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 2, pp. 737–742, © 2001, Elsevier Ltd.

Abstract

Aristocracy as an identifiable element of a social structure has characterized human societies from ancient times, and was more often than not directly associated with land ownership. Different countries developed different forms of aristocracy, sometimes encompassing a small noble class and sometimes an extensive group of farmers on a small or large scale, often labeled the gentry. As a form of government aristocracy was associated with the rule of the wise, and debates about the extent to which this was actually the case, and could be extended and given a social and political role, characterized intellectual life in every society. It was only toward the end of the eighteenth century that aristocracy began to be challenged as a social form, and visions put forward of societies without rank or hierarchy. Although this vision was not brought to reality after the French Revolution (especially after Bonaparte ennobled many of the believers in equality who had survived the revolutionary turmoil), the idea of a society without an aristocratic class continues to attract adherents, combated in turn by those for whom class and social hierarchy are natural social forms.

These three terms embrace the elite groups originally connected with land ownership. They have been used to describe different social types, chiefly in Europe. However, along with 'court,' 'king,' 'prince,' etc., they are also being used to describe diverse tribal (African or Asian) systems of social stratification and power. This article will concentrate on Europe and the Anglophone world with explicative forays to other parts of the world. Much depends upon the language in use. In the British usage, 'nobility' equals 'aristocracy'; the term 'gentry' being used for the landowners without hereditary titles. This does not necessarily fit the continent of Europe, where in early modern times 'the noble' corresponds rather to the English 'squire' than to the 'aristocrat' (Cooper, 1970). All these terms (and their numerous synonyms, like *Ritter*, *Ritterschaft*) have been value-charged, often strongly, but not necessarily and not always positively, such as 'workers aristocracy.' The terms in question change with time: some can be said to be losing their value (including once common terms such as 'Sir' or 'Esq.'). There follow sections on: seclusion of the aristocracy (and/or nobility); interplay of 'immemorial origins,' 'old riches,' and 'royal grace' in the reproduction process of the traditional elites; 'patricians' as an urban aristocracy; *dérogeance* and various patterns of behavior of landed and urban elites; traditional values; inheritance and preservation mechanisms of landed elites; their relations to other estates and classes; the patronage (political and cultural); the royal household; orders of chivalry; growth of the modern state, traditional elites, and state service; and the twilight of the aristocracy in the industrial society, its adaptability to new conditions, aristocratic preserves.

Troublesome Terminology

This article concentrates on Europe. No room can be spared for scrutiny of other parts of the world but one should mention the striking similarities between medieval nobility in Europe and the Japanese *samurai*, as well as in modern times the export of European aristocratic structures to other parts of the world. Even within the European core, diverse terminologies reflected different social structures of particular countries. In the British usage, as noted above, 'nobility' equals 'aristocracy' and was often associated with a peerage; it does not include the wealthy or poor 'gentry' without hereditary title; in France or Poland and across much of Europe, by contrast, the ranks of the gentry would be considered members of a noble class. In this article the term 'nobility' will be used in this comprehensive meaning, corresponding with the German *Adel*, Italian *nobiltà*, or French *noblesse*, all terms originally connected with land ownership. In particular languages there were parallel terms that laid stress on the origins of the nobility. In German *Ritter*, as well as in Swedish *riddare* and French '*chevalier*' or Spanish '*caballero*,' originally signified 'horseman.' In Poland the word *szlachta* stressed its hereditary character (from German *Geschlecht* over Czech *šlechta*), while in Sweden medieval *frälse* underlined their freedom from taxes. All these terms have been widely employed in political argument. They gave birth to various modern descriptive terms like 'workers' 'aristocracy,' or 'trade-union aristocracy.' The derivatives of 'gentry' have in English a particular value, being associated with the improvement of an area (as in 'to be *gentrified*'). In other words, gentrification in English was not synonymous with a process of aristocratization, whereas elsewhere in Europe this might well have been the case.

Ancient-World Origins

The word 'aristocracy' is of ancient Greek origin and signifies the 'rule of the best.' In Homeric times 'the best' signified chiefs of the noble families who pretended to share with the king descent from the gods, and were also prominent because of their wealth and personal prowess. They formed a class of 'horsemen' or 'knights' (*hippeis*) connected by blood and by various community institutions. They governed the state by means of the council of the *gerontes* (the elders). In the eighth and early seventh centuries BC, the social position of the

aristocrats was derived from their land ownership but also could be acquired by commerce, robbery, and piracy. The aristocrats dominated the communities (*poleis*) and organized colonization. Many factors contributed to destruction of aristocratic rule, like change of military tactics (riders in single combat were replaced by *phalanx* of heavily armed foot soldiers) and the ascent of nonagrarian social groups striving for power, and collected in the courts of emperors or in large urban enclaves.

The fate of the aristocracy in ancient Greece followed a pattern that was replicated by many other ruling groups through history: from undisputed moral and political domination to the challenge of rival groups and on to a near-complete loss of power. Prestige related to ancient roots (real or fictitious) survived, however, and would become a constituent element of all aristocracies. Aristocracy as the 'rule of the best' was a moral ideal; if birth was replaced by wealth as the decisive qualification, as tended to happen, the result was oligarchy.

In republican Rome several groups consecutively established an oligarchy of prestige and power. The earliest hereditary estate was that of the patricians (the *patriciate*) who in the late fifth century BC came to dominate almost all of the offices of state. In the later fourth century, their competitors, the plebeians (*plebs*), also began to gain access to power. The outcome was a sort of convergence of the top strata of both estates through matches and family alliances. In the late fourth and the third centuries a new aristocracy began to emerge, the *nobilitas*. Its foundation was the great landed estates run by the slaves and peasants, who were more and more dependent as clients. About 30 houses had access to power both civil and military. The electoral system and honorary unpaid offices secured their domination. Territorial expansion offered them benefits (the fruits of power in the provinces) but at the same time created new dangers. 'The mighty few' (*pauci potentes*, as Cicero called them) were being pressed by the *equites*, originally a moneyed group that in stormy times of civil wars strove for power and was virtually assimilated to the *nobiles*.

Augustus and his imperial successors changed the role of *nobilitas* by the very introduction of the Imperial Court. Later on, along with spatial expansion of the Roman Empire, there emerged provincial aristocracies whose role increased when the center was losing its grip on more distant provinces invaded by the 'Barbarians.' Roman traditions influenced medieval and modern vocabularies of elites.

The Early Middle Ages and La Féodalité

The dissolution of the Roman Empire in the West (fifth century AD) brought about new nobilities from among the retinues of Germanic chieftains and local Roman governors. Their structures were stabilized and let to the emergence of new states. The milieu of princely (royal) households offered opportunities for advancement of diverse social types like the prince's companions, allied tribal chiefs, valiant warriors, and also bondsmen or slaves who had a chance to serve the prince's person directly. A parallel but closely interwoven network was created by the Church. The bishops exercised secular control over large territories, and the Church offered a convenient upward path to able and ambitious persons of modest origin. This is true for later periods as well. When, after the twelfth century, nunneries multiplied, they constituted an opportunities for the daughters of the prominent, who might easily become abbesses or prioresses of great prestige and considerable power. Individual success signified family advancement as well. Close relation to the Church has been for many centuries an important factor in aristocratic modes of life. In Germany even Protestant noble families claimed a right to particular benefices granted by the Catholic Church, which they had traditionally enjoyed before the Reformation.

Medieval and modern aristocracies with their hereditary titles of honor had their roots in the royal household and in the administration of the early state. Office holders strove to obtain permanent hereditary positions. Instability within the early medieval states helped them. In Carolingian times a 'count' (Latin: *comes*) was a judge who accompanied the Emperor. In the twelfth century there emerged other *comites*, provincial governors, and 'viscounts,' who were their representatives. Border provinces of the Empire were governed by the 'marquesses' (German *Markgraf*), 'Duke,' or (German) *Herzog*, which meant 'leader of hosts.' 'Baron' (German *Freiherr*) was either a great landowner or just a substantial 'free man,' i.e., primarily the prince's companion, not his simple subject; it was also an equivalent of 'nobleman.' Accordingly the recipients of the *Magna Carta* (1215) were *barones*.

More rigid rules of interpersonal relationship were conceived originally in France in the twelfth century, and developed into sophisticated structures in subsequent centuries. In that system, now often called 'feudal,' whoever counted was somebody's vassal and had his own vassals as well. Everybody knew his specific rung on a long ladder on top of which was the ruler, who was deemed to enjoy a special relationship with God. This formed a power system based on personal relationships, a skeleton of the medieval 'state,' called by the German constitutional historians *Personenverbandstaat*. Diverse alterations of these rules cannot be analyzed here, although they determined the nature of particular aristocracies. It is enough to mention the notion of 'pairs' or 'peers' (the king's direct vassals equal to each other and also his companions like the legendary 'Knights of the Round Table') making the council of prince's advisors with juridical privileges, and a pressure group conscious of their common interests. In Norman-dominated countries (the most important being England and the Kingdom of Naples) the ruler was recognized as direct *seigneur* of all vassals; in other countries the dominant principle asserted that 'the vassal of my vassal is not my vassal.' The Norman principle became a legal tool whenever the king wished to curb his vassals. In the thirteenth century it helped to create in Naples a bureaucratic state. The other system tended to become a complicated mixture of conflicting loyalties, leading to particular turmoil in France and Germany.

From the late Middle Ages the nobilities in particular countries were shaped by factors as diverse as the intensity of the money economy and customs or laws of inheritance. Money had a destructive influence upon feudal relationships: debts and mortgages haunted landed estates. A system, later called 'bastard feudalism,' replaced the fiefs with monetary payments. In the Hundred Years War the feudal hosts of France proved helpless time and again against English foot militia

employing the longbow. Eventually mercenary troops raised by entrepreneur-commanders replaced them. The most successful commanders aspired to give themselves aristocratic titles. In Italy mercenary commanders (*condottieri*) were nobles, even sovereign princes, and they often recruited their troops among lesser nobles from Romagna, Friuli, or Marche. The nobles were looking for a role within a constantly growing state machine.

Early Modern State: Crisis and Adjustment

Not everybody was seeking to become a nobleman, however. The fifteenth and early sixteenth centuries witnessed multiform crises: rents were falling, and robber knights emerged in Bohemia (they were especially active among the Hussites) and in East Germany. Revolts of lesser nobles broke out in many parts of the Empire: princes and bishops eventually won and razed many knights' castles. In the Renaissance, assemblies of estates, all over Latin Europe, became a source of power parallel to that of the monarch. The nobles participated in them chiefly through a process of election (along with the clergy and the burghers) but the pairs and titled nobles usually had their personal seats secured. The Assemblies became a forum for open competition between the nobility and the burghers. The nobles of a given province usually aimed at a monopoly of posts and offices, against the burghers and any outsiders.

Research in the second half of the twentieth century was particularly concentrated on the 'crises' of the nobility. In most countries the nobles (in the broadest sense) rather successfully adapted themselves to new economic and political conditions but this created major shifts within the estate; hence the 'crisis.' The ideal of an unspoiled and bucolic noble life could be contrasted with that of public service. A precondition for state service was legal training, and the gentry largely disdained formal education. Therefore some rulers encouraged foreign travel or founded special lay colleges (like the Knights' Academy in Siegen or nobles' colleges in Denmark and Sweden). In Elizabethan and Stuart England, young gentlemen crowded into the Inns of Court, and in Scotland the legal training of young noblemen was encouraged in law. In the seventeenth century German nobles embraced a legal training as they needed it in the service of their princes. Not so for the French *noblesse*.

Internal stratification of the estate was of utmost importance to the future of noble ranks. It differed according to size. Here are some estimates: in Denmark, around 1600, the nobles made up 0.25% of population; in Bavaria, Saxony, and Holland slightly more of the population were nobles; in Bohemia and in Prussia, in 1700, 1%; in Poland 7%; and in Castile about 10% (see Labatut, 1978; Meyer, 1973). In some regions (Vizcaya in Spain, Masovia in Poland) membership of the noble estate was far greater.

As states became more centralized rulers took the opportunity to organize their noble subjects into clienteles. 'Liveries' and noble retinues were forbidden in Tudor England and virtually suppressed in Spain in the mid-sixteenth century. In France, a century later, they failed as a weapon of the prince during the *Fronde*. The seventeenth century may be regarded as a century of opposition between the 'Court' and the 'Country.' The former became a principal source of prestige and privilege, a center of culture and power. It was never as important in England, at least after the Tudors, as it was in France, Spain, and Austria.

In the East, where demesne farming prevailed, the state was 'coercion-intensive' (as Charles Tilly, 1990 has termed it). In Prussia and Livonia, the landed gentry (*Junkers*) identified with the state and its army. In Russia, they were the Tsar's dependent servants and did not become an estate before the mid-eighteenth century. In Poland a sequence of fifteenth-century statutes secured independence from the royal administration. The structure of politics in Poland–Lithuania was shaped by the relationship between the gentry and the landed magnates. Neither of these groups was interested in a strong executive, after an elective monarchy was introduced, and the political patronage of the magnates became the dominant factor in the politics of the state.

In Sweden, nobility, having virtually no medieval traditions, developed into an estate of civil and military officers through the ennoblement of commoners. Absolutist reforms in 1680 then challenged the aristocracy and diminished its authority. The system of ranks (known also in Denmark and later introduced into Russia) served the interests of the monarch. In France, the *noblesse d'epée* felt endangered by the robins (*noblesse de robe*), or newcomers, who were buying land and titles of nobility on a large scale. Strict *dérogeance*, a ban on improper and disqualifying sources of income, such as trade, created economic problems. The high aristocracy had enjoyed military commands and governorships of provinces as their *chasse gardée*. But their influence was rather restricted to the Court and its intrigues, masterfully described in the memoirs of Duke de Saint-Simon (1699–1752).

Pride and Prejudice

One may argue that noble culture has been based on these two vices. The whole idea of nobility drew from tradition, and much imagination was employed to deepen the roots of each family tree. The Renaissance discovery of Roman literature enriched noble imagination with hosts of ancient heroes. Most seminal were the escapees from burning Troy. When in 1624 a scion of the distinguished Lithuanian family Pac visited Florence, both he and his hosts, the Pazzi, were happy to get 'proof' of their common heroic, Homeric origins. The nobilities favored imagined glorious ancestry: the Germanic Francs (as opposed to the commoners descending from the Celtic Gauls) in France, and in England the Norman knights of William the Conqueror. The Lithuanian gentry descended from a Palemon (from Troy) and the Polish one from Japhet son of Noah (as contrasted with Cham for the peasantry and Sem for the Jews). In Italy, France, and Spain noblemen and scholars discussed the relative importance of long and impeccable pedigree, which were seen to be the basis of a noble lifestyle and accompanying traits such as personal valor.

Of more practical importance to the maintenance of noble classes were property rights and the laws governing inheritance. These determined the economic continuity of families. Chiefly in the south (Naples, the Islands, Spain) feudal property was dominant, which in cases of conflict, became a tool of the monarch. In the England of the early Stuarts, the Court of

Wards was an instrument of fiscal oppression of the nobles and the Long Parliament abolished it in 1640. In Naples and Sicily the nobles continuously pressed their Spanish sovereigns for more liberal rules concerning feudal inheritance.

The preferred rule of inheritance was the entail, which treated landed estates as family property and neutralized the tendency to division. However, it limited opportunities to mortgage the real estate and left open the fates of younger sons and daughters. In the English case, entail served landed property well: it kept property intact and the younger sons easily found jobs in business, the professions, royal service, and eventually in the colonies. On the Continent, ambitious, less substantial nobles as 'noblemen-errant' searched for service abroad. Count Achatius zu Dohna (from Prussia) and Prince Eugene (of Savoy) were among the best known. In Poland, shared inheritance contributed to the impoverishment of the gentry and the growth of wealthy magnates. The latter secured entail, for themselves alone, as a legal exemption.

Urban Aristocracies

The cities had aristocracies of their own. North of the Alps, they were usually called the 'patriciate' and in Italy simply *nobiltà*. In the High Middle Ages, before town guilds took over the government, most cities were run by royal officers of noble extraction. The most impressive relics of that early stage and its lifestyle are 13 towers that dominate the cityscape of San Gimignano and the incredible constructions still towering over Bologna. In most cities, however (Florence and Basle for example) victorious town guilds destroyed these bastions of noble power, only to create new elites of their own. The large, wealthy town offered great benefits to the people of influence, and led to strong oligarchic trends in urban 'communes.' The top families were often more aristocratic than the aristocrats *sensu strictiori*. In Italy and later also in south Germany, ruling elites introduced the *serrata*, or closed lists of families with exclusive right to city offices in a given city. The most famous was the Golden Book of Venice (1297) but it was by no means unique. The patricians cultivated their special ways and observed *dérogeance* taboos. In Germany, the symbol of urban independence was 'Roland,' the legendary companion of Charlemagne and personification of knightly valor. King Arthur too became the eponym of exclusive clubs (*Artushöfe*). The urban patricians built country residences and acquired landed property chiefly as a status symbol. It was often apparent that they treated this as a collective enterprise. In the fifteenth and sixteenth centuries, in the Swiss city of Luzern, a small cluster of patrician families raised urban troops and leased them to neighboring powers. The patricians commanded them and collected most of the money paid for their services. Profits from power and politics were closely interwoven.

The *connubium* or marriage market strictly limited to elite families created problems, and between the sixteenth and eighteenth centuries many urban aristocracies from Venice to Lübeck experienced a shortage of skilled candidates for urban offices. The medieval register of noble families, centuries later, did not secure wealth compatible with their social standing. Therefore the Venetian *nobiltà* created, at the city's expense, a system of social welfare for the noble poor.

In general, the urban patriciate in many respects was similar to the aristocracy. They were closely related too, naturally enough, because landed aristocrats did not scorn urban money. In Germany, patricians from many cities of the Empire (*Reichsstädte*) enjoyed the status of the *Reichsadel*. The Habsburgs granted aristocratic titles to wealthy burghers who supported them with loans and banking services. Names like Fugger (Augsburg), Doria, and Spinola (Genoa) are only the most famous among many. In some cases the patrician was automatically regarded as a noble: in the tiny imperial city (*Reichsstadt*) of Wels, owners of important local salt (brine) wells were, by imperial patent, regarded as nobles (incidentally, this status was enjoyed by Franz von Papen, who preceded Hitler as German Chancellor). In France some city posts were ennobled almost automatically. Even the most exclusive Maltese Order excepted candidates from Florence, Genoa, Lucca, and Pisa from its very strict *dérogeance* rules.

Members of the urban elite merged with the nobility *tout court* in many ways. They met at Royal Councils, acquired landed estates and prestigious titles, and profited from the inflation of honors in the sixteenth and seventeenth centuries. They defended their exclusive status against the bureaucratic encroachments of enlightened autocrats. The elites of some larger cities made a particularly closed, self-conscious caste. The twilight of the patrician civilization was best described by Thomas Mann in *Buddenbrooks* (1901).

Triumph and Twilight of Aristocracy

The French Revolution closed a chapter, but not the whole story, of European aristocracy. Although aristocrats were given preference onto the scaffold (even an heiress of Tchernobyl, who visited Paris at the wrong time, lost her head), the First Consul and later Emperor Napoleon Bonaparte created his own Court and constituted a new stratum of *notables* based on loyalty to the ruler. Political changes largely destroyed the prestige of French élites whose membership reflected, in the nineteenth century, the vicissitudes of the nation. But still, in 1839, the Marquess Astolphe de Custine argued that the true nobility cannot be bestowed or purchased. The ruler, he continued, can appoint dukes; education, opportunity, and genius or virtue can create heroes; but nobody and nothing can make a noble. Yet everywhere commoners were being ennobled: bankers, industrial entrepreneurs, statesmen and politicians, army and navy officers, across the states of Europe. Monarchies, even restricted by the constitutions of the second half of the nineteenth century, needed traditional elites as political support and as an ornament. In Berlin, the last emperor Wilhelm II recreated a very formal and lavish court ceremonial. In Britain, in 1876 Queen Victoria became Empress of India and made Benjamin Disraeli, a statesman of Jewish extraction, Earl of Beaconsfield. Immemorial traditions were being cheerfully reinvented. In Britain, the 'first industrial state,' the high nobility had a greater share of land than in other states and continued to enjoy substantial political influence.

By contrast, in the East, after the dissolution of the Polish Republic (1795) the three partitioning powers inherited numerous lesser nobles but almost no titled ones. This unusual gap had to be filled and eventually anybody who was able to

prove membership of the Senate among his ancestors, and who pursued a noble lifestyle, could become baron or count. 'Austrian Count' became, in Poland, an ironic comment. In general, however, in the monarchies, ennoblement and aristocratic titles were becoming a mere ornament and a final prize for distinguished service and/or business performance.

Constitutional and political changes in Europe in 1917–18 and after were destructive of the nobility, and not only those in Europe. Neither republican Spain nor Franco's Spain were supportive of traditional Spanish aristocracies, either at home or in the colonies of Latin America.

After 1918 radical projects for agrarian reforms threatened noble power in many countries but it was entirely destroyed only in the Soviet Russia. The former Russian aristocrat who became a taxi driver in Paris became a literary *topos*. Another totalitarian system, Nazi Germany, was supported by the Prussian Junkers. In the beginning, Adolf Hitler's party seemed to them too radical, too plebeian, but the preceding Weimar Republic had also been disappointing for them. Nevertheless the Junkers turned to Hitler because they shared nationalistic commitments and a fear of the left. At the same time, as the war went badly, many aristocrats were prominent within the internal opposition.

Perhaps the *Credo* of the late aristocrats eventually challenging the Nazi dictator could have been outlined that way: the aristocracy never collaborates with the tyrant in defense of order; its tradition is on the one hand the defense of the people against despotism, and on the other the defense of civilization against revolution, that most dreadful of tyrants (or so Astolphe de Custine had had it). The failed coup against Hitler on 20 July 1944 brought the lives of many aristocratic Prussians to a bloody end, and the postwar loss to Poland of the territories beyond Oder-Neisse frontier cut off their economic base.

The conspiracy against Hitler was perhaps the last conscious political action of an aristocracy. The political influence of that class seemed lost for good, notwithstanding the distinguished intellectual, scientific, or political performance of their individual members. What remains is the value of their collective tradition and the sentiment or flavor so well presented to the modern world by a *Principe* Giuseppe Tomaso di Lampedusa or a Jean Marquis d'Ormesson. Having said this, it may well be the case that neoliberal economic reforms are paving the way for the establishment of a new global aristocracy of bankers and hedge-fund managers, owners of land and owners of scarce natural resources, and owners of new technology companies. A corporate elite, equally, is emerging, many members of which have strong personal and private links to the politicians whom they serve through the outsourcing of public services, including defense. Whether these trends will lead to new forms of aristocratic living is an open question.

See also: Bourgeoisie and Middle Classes, History of; Class: Social; Elites: Sociological Aspects; Family and Kinship, History of; Feudalism; French Revolution, The; Inequality: Comparative Aspects; Middle Ages, The; Peasants and Rural Societies in History (Agricultural History); Revolutions, History of; Social History; Social Inequality in History (Stratification and Classes).

Bibliography

Adonis, A., 1993. Making Aristocracy Work. The Peerage and the Political System in Britain 1884–1914. Oxford University Press, Oxford, UK.
Asch, R.G., Birke, A.M., 1991. Princes, Patronage and the Nobility: The Court at the Beginning of Modern Age c. 1450–1650. Oxford University Press, Oxford, UK.
Bush, M.L., 1983. Noble Privilege. Manchester University Press, Manchester, UK.
Bush, M.L., 1988. Rich Noble, Poor Noble. Manchester University Press, Manchester, UK.
Cannadine, D., 1990. The Decline and Fall of the British Aristocracy. Longman, London.
Cooper, J.P. (Ed.), 1970. The New Cambridge Modern History, vol. 4. Cambridge University Press, Cambridge, UK.
Dewald, J., 1996a. The European Nobility, 1400–1800. Cambridge University Press, New York.
Dewald, J., 1996b. The European Nobility, 1400–1800. Cambridge University Press, Cambridge, UK.
Donati, C., 1995. L'Idea di nobiltà in Italia. Secoli XIV–XVIII. Laterza, Roma-Bari, Italy.
Duggan, A., 2000. Nobles and Nobility in Medieval Europe: Concepts, Origins, Transformations. Boydell Press, Woodbridge, Suffolk.
Duggan, A.J., 2000. Nobles and Nobility in Medieval Europe: Concepts, Origins, Transformations. Boydell & Brewer Ltd.
Gemmill, E., 2013. The Nobility and Ecclesiastical Patronage in Thirteenth-century England.
Given-Wilson, C., 2003. The English Nobility in the Late Middle Ages: The Fourteenth-century Political Community. Taylor & Francis e-Library, London.
Goody, J., Thirsk, J., Thompson, E.P. (Eds.), 1976. Family and Inheritance. Rural Society in Western Europe. Cambridge University Press, Cambridge, UK, pp. 1200–1800.
Kujala, A., 2003. The Crown, the Nobility, and the Peasants 1630–1713: Tax, Rent, and Relations of Power. SKS/Finnish Literature Society, Helsinki.
Labatut, J.-P., 1978. Les Noblesses européennes de la fin du XVe sicle à la fin du XVIIIe sicle, first ed. Presses Universitaires de France, Paris.
Leonhard, J., Wieland, C., 2011. What Makes the Nobility Noble?: Comparative Perspectives from the Sixteenth to the Twentieth Century. Vandenhoeck & Ruprecht, Göttingen.
Lukowski, J., 2003. The European Nobility in the Eighteenth Century. Palgrave Macmillan, Houndsmills, Basingstoke, Hampshire.
Mączak, A., 1989. Der Staat als Unternehmen. Adel und Amtsträger in Polen und Europa in der Frühen Neuzeit. Oldenbourg, Munich, Germany.
Meyer, J., 1973. Noblesses et pouvoirs dans l'Europe d'Ancien Régime. Hachette, Paris.
Mingay, G.E., 1977. The Gentry. The Rise and Fall of a Ruling Class, London.
Nierop, H.F.K., 2009. The Nobility of Holland: From Knights to Regents, 1500–1650. Cambridge University Press, Cambridge.
Powis, J., 1984. Aristocracy. Oxford University Press, Oxford, UK.
Romaniello, M.P., Lipp, C., 2011. Contested Spaces of Nobility in Early Modern Europe. Ashgate, Farnham, Surrey, England.
Scott, H.M. (Ed.), 1995. The European Nobilities in the Seventeenth and Eighteenth Centuries. Longman, London.
Smith, J.M., 2006. The French Nobility in the Eighteenth Century: Reassessments and New Approaches. Pennsylvania State University Press, University Park, PA.
Stone, L., Fawtier Stone, J., 1964. An Open Elite? England 1540–1880. Clarendon Press, New York.
Stringer, K.J., 2004. Essays on the Nobility of Medieval Scotland. John Donald, Edinburgh.
Tilly, C., 1990. Coercion, Capital and European States, AD 990–1990. Blackwell, Oxford, UK.
Wehler, H.-U. (Ed.), 1990. Europäischer Adel, 1750–1950. Vanderhoek and Ruprecht, Göttingen, Germany.

Armed Conflict and Social Work: Significance and Implications

Shulamit Ramon, Anglia Ruskin University, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

An underresearched area, this article is aimed at understanding the complexity of the impact of armed conflict on individuals and communities in the context of social work. For this purpose, key definitions, the scale of those affected by the phenomenon, and its impact are considered. Issues related to researching armed conflict are highlighted. The relevance of social work support during different facets of armed conflict follows, including the dilemmas that this type of intervention raises for social workers. Throughout the text, the international character of armed conflict is emphasized, and examples are given from different parts of the world.

Introduction

Making sense of the significance and implications of armed conflict for social work requires clarity about what constitutes armed conflict. Armed conflict is a political conflict in which arms are used to attack the other side, defined as 'the enemy,' and to defend one's own side. Although this is the definition of war, the concept of war is usually kept for an armed conflict between two or more countries, while the term 'civil war' is used to describe a war within one nation or country.

Armed conflict is not a useful term for describing hostilities between two groups in which arms are not applied, but which are nevertheless powerful, and may be either the outcome of an armed conflict, or lead to one in the future. Most conflicts between ethnic minorities and majorities, coexisting religions, and travelers and nontravelers do not result in armed conflict most of the time, but may plant the seed for such a conflict in the future. The war in the then Yugoslavia in the 1990s provides a good example of a political conflict that simmered for 50 years before it erupted into outright war. This article considers the relevance of social work support during different facets of armed conflict, including the dilemmas that intervening poses for social workers. It emphasizes the international character of armed conflict, giving examples from different parts of the world.

Contextualized Background Information

There are 15.3 million refugees of war from 21 countries, according to the United Nations (UN) Higher Commissioner for Refugees (UNHCR, 2012). Refugees are people who have left their own country due to a violent conflict and 10.5 million of them are looked after by the UNHCR, while 4.8 million of these are looked after by the UN Refugees Welfare Agency in the Middle East refugee camps based in Jordan, Lebanon, and the Palestinian Authority (PA) where they have been since 1949.

Twenty-seven million people are recognized as internally displaced people (IDP). They live mainly in their country of origin, but away from where they lived when the conflict began, which they left due to the impact of an ongoing internal war (civil war) or occupation by another country. This brings the total number of refugees and IDP to 42 million, likely to be an underestimate, because many of those leaving in a hurry in the chaotic situation of an armed conflict, are often not counted and do not declare themselves as either IDP or refugees.

Asylum seekers are people who declare themselves as refugees, live outside their country, and whose status has not been confirmed as refugees by the authorities in the country in which they reside. The total number of asylum seekers is not known, as not every country releases this information, and also because some of them enter a country illegally, or without the knowledge of the authorities. Recent information from the UNHCR (2012) states that in 2012, there were 108 912 asylum seekers in industrialized countries, a rise of 25% from the 2009 to 2010 figures.

Their treatment in the host countries and country of origin should follow the Geneva Convention (established in 1951). It should also follow the UN Convention and Protocol (1967) on the status of refugees, endorsed by the UN, although not signed by all countries (see Ramon and Magljlic, 2012). In reality, this is not always the case, especially for IDP or asylum seekers, where each country decides how they should be treated.

Experiencing Armed Conflict

Some of the most negative outcomes of armed conflict appear daily on television screens. Losses of many types characterize the impact of armed conflict, such as loss of life, limbs, family members and friends, animals, property, or items of sentimental value. The loss of the sense of security and of future prospects for many of the survivors of an armed conflict may take a very long time to heal, and in many instances remain unhealed and potentially lead to further conflict in future.

Economic devastation is often a direct outcome of an armed conflict, in which the means of production have been destroyed, leading to supply shortages of food, water, and means of construction. The lack of food and shelter leads to an increased rate of illness and death among the more vulnerable population groups, such as children, women, and old people (Watchlist, 2003 (Web site); ECHO, 2013 (Web site); Doucet and Denov, 2012; Krill, 2000).

The destruction of medical facilities and the disappearance of health personnel lead to a high rate of disability and increased mortality among those wounded. The long-term

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28007-0

physical, psychological, and social impact on those affected has been documented (e.g., Doucet and Denov, 2012). These include the long-term impact on women, Beah's (2007) personal account of being forced to be a child soldier, and Nuttman-Shwartz et al.'s (2011) account of the long-term impact of posttraumatic stress disorder (PTSD).

Long-Term Impact

Most of the current ongoing armed conflicts, across all continents, have their roots in the past history of the communities engaged in them. For instance, the current Syrian conflict has its roots in the history of the relationships between the Alawite minority groups and the Sunni majority; the aftermath of the war in Iraq is marked by the renewed conflict between the Sunni and the Shia communities, which has also been at the background of the war between Iraq and Iran in the 1980s.

It is the interpretation of the past, and what is remembered of it by those who witnessed the conflict and subsequent generations who reinterpret the past in the light of the present, that counts and makes the differences between looking for another round of armed conflict or its prevention. Memories are always constructed, although often individual memories are suppressed in favor of memories perceived by those in power as suitable for collective consumption (Halbwachs, 1992).

Social work practitioners need to be aware of these issues in their work with individuals and communities, to enable individuals to resolve their memories that conflict with the collective injunction. For example, in Bosnia, Moslem women experiencing rape in the 1990s were not allowed to mention this experience due to the wish to repress the assumed guilt of their menfolk in being unable to prevent the rape (Zavirsek, 2008: p. 162).

Considerable efforts have gone into collectively remembering some conflicts, such as the Second World War, the Holocaust of Jews and Gypsies in Europe, and beyond these groups at that time, or in subsequent events of a similar nature. Moreover, some other conflicts, no less horrendous or far reaching, are not equally collectively commemorated, e.g., the Armenian genocide and the Biafra war.

The wish to remember, and for such events to be etched into the collective memory, is understandable in the light of the magnitude of what has happened (Zavirsek, 2008). Furthermore, where the underlying reasons for the conflict have not been resolved, and where people continue to lead an uprooted life, it is likely that the conflict will resurface, as is the case in the Middle East, where a third generation of those who left Palestine in 1948 live in refugee camps, do not have full citizen rights in the countries where the camps are located, live on UNWRA financial support, have poor future prospects, and are highly unlikely to return to the land of Palestine–Israel, a country they have never been to, but which has acquired a mythical significance for them, as it did for Jews for over 2000 years.

At the individual level, there is a range beginning with people who wish not to forget any element of the conflict, ending in those who wish to forget all of it, with people who have nightmares connected to it to those for whom it does not matter in the middle. PTSD is a recognized psychiatric diagnosis, attributed to people suffering from reliving selected situations of the conflict and finding it difficult to continue with their ordinary life (Nuttman-Shwartz et al., 2011).

Enabling to share significant memories both collectively and individually is an important part of community social work/development work in rebuilding a community impacted by armed conflict, helping to cement community cohesiveness. Yet, carrying out this work without rekindling the wish for revenge and reliving traumatic events presents a challenge to social workers.

Positive Outcomes of Armed Conflict

Side by side with the high level of mindless brutality that war unleashes, a kinder human approach during armed conflicts and their aftermath can be observed. Positive outcomes include in-group and outer-group solidarity; instances of human generosity; readiness to risk one's own life for the sake of the lives of others; tireless work toward conflict resolution and peace building; health and social reforms that would not have taken place without being stimulated by the wish that something positive will come out of such negative events; and creative art (see L'Art en Guerre, 2013).

Specific examples include Anna Freud and Dorothy Burlingham (who was a social worker) revising their psychoanalytically oriented work with children following the lessons learned from their involvement in the experience of mass evacuation of British children during the Second World War (Freud and Burlingham, 1943). This has led to more caution in considering moving children from home when a parent becomes ill and unable to look after them, to much greater focus on the family as a unit during family consultation services (Felsman, 2001). The recognition that many of the behaviors displayed by the evacuated children were normal reactions to abnormal situations helped to reduce the detection of pathology, and diagnosis of these children as suffering from forms of mental illness.

When Sarajevo's psychiatric hospital was bombed out of existence during the Bosnian war, the option of not to rebuild it was accepted, enabling the more radical change of the mental health system there to proceed and move it from being a heavily institutionalized and segregating system to one with a number of community services (Kucukalic et al., 2005).

The value of an individual acting courageously to raise awareness and encourage international organizations to take the necessary steps has been exemplified by Malala Yousafzia recently. Aged 14 years, she was shot in the SWAT Valley of Pakistan in October 2012 by the Taliban because she protested against their policy of excluding girls from schooling in Pakistan. The shooting was not only condemned internationally, but also resulted in expressed mass solidarity with her in Pakistan and elsewhere (Amnesty International Newsletter, November 2012: pp. 10–17). She has survived the shooting, but had to pay the price of emigrating from her country to the United Kingdom. Malala has continued her activities for girls' education, and has received during 2013 the International Peace Prize for Youth.

Researching Armed Conflict

Research into armed conflict is needed to achieve better conceptual understandings, give meaning to people's experiences of it, and develop strategies for preventing the next round of armed conflict and the harm to the population from the ensuing outcomes. Yet, carrying out methodologically sound research is not an easy task. There is potential physical risk to researchers and those participating in the research during an acute conflict stage. Risk is also heightened because such research touches on highly sensitive and emotionally charged issues.

Researchers of human behavior are likely to use qualitative methods that encompass their own values and biases (Lincoln and Guba, 1985). This may introduce a lack of 'objectivity' that becomes multiplied in armed conflicts, when sympathy with one side provides the credentials that enable the researcher who is asking for permission to carry out research to do so. Furthermore, the sensitivity of the themes requiring research is difficult to encompass through simple methodology and methods, while researching in harsh conditions calls for simplicity.

Although impartiality is unattainable, commitment to uphold universal ethical rules and reflect on one's own biases go some way to ensure that this research is carried out ethically (Kelly, 2004). Two types of research can be identified: research aimed at better understanding the context and stakeholders' positions and evaluation of policies and interventions, including social work interventions.

Despite these difficulties, examples of sound research in relation to values and methodology exist, some of which are outlined below. Miljenovic and Zganec (2012) researched the views of local residents in a region of Croatia that showed signs of community disintegration after the 1990s war there. They developed a framework of the factors of community life that mattered to the residents, and interviewed key people representing these elements. Deteriorating economic development topped the list, followed by dysfunctional social relations between the ethnic groups that were leading a separate coexistence, difficulties in the way public institutions were functioning, and crisis of local government. Meanwhile, international nongovernmental organizations (NGOs) were active, but internal NGOs were operating in an uncoordinated fashion that lacked continuity. On the basis of this research, the authors proposed how community building could proceed.

A survey of 10 countries across Africa, Asia, Europe, and South America, undertaken recently by the World Health Organization (Fisher et al., 2011), highlights the difficulties that adolescents face in war situations. In total, 4868 adolescents from armed conflict areas participated in these studies, and although they come from different contexts, the similarities in the findings are striking. The findings include high exposure to traumatic events, high to severe levels of PTSD symptoms, varied therapeutic responses, and the impact of the context on the severity of difficulties experienced and adequacy of responses. For example, those who became child soldiers had not only experienced traumatic events, but also scored much higher than the clinical cutoff point on depression and PTSD scales. Much needed evaluation of policy implementation is provided in a recent study of UNCHR community empowerment projects in Sierra Leone (Skran, 2012).

Only a small number of studies looked at the impact of armed conflict on social workers (e.g., Campbell and McCrystal, 2005; Ramon et al., 2006; Baum and Ramon, 2010), although this group is highly likely to be impacted by armed conflicts. Al Mikhamresh et al. (2012), researching social work activities with Iraqi refugees in Jordan, highlight the need for politically competent social work practice, in the sense of having to understand well the 'political map' in the situation and use this understanding to benefit clients. Guru (2012) refers to the dilemma of the researcher attempting to work with families of young Moslem men accused of engaging in terrorist activities. This research sympathizes with the families and wishes to understand better their context, but is not sure how ethical it is to research such families at the crisis stage in which their lives are embedded, while the researchers simultaneously experience difficulties in getting ethical permission to carry out this research on the grounds of potential risk to them.

The Relevance of Social Work to the Humanitarian Activities Undertaken in Armed Conflict Contexts

Local social workers in an armed conflict context are usually part of the workforce providing different types of support, working alongside doctors and administrators, and local and international organizations. At the time of an acute conflict, they will provide any help they can – food, shelter, accompanying people to medical services, taking children left without parents to a safer place, and sharing in their grief.

Support at the postacute stage may differ. At times, this focuses on economic support, including rebuilding local enterprises, emotional support in the form of formal and informal counseling, bereavement work, referring survivors to disability services, reconnecting people, supporting people to go back to their original geographic or ancestral location, and enabling memory work to take place (Zavirsek, 2008). The social work response needs to be tailored to the variations in the reactions to the traumas experienced during an armed conflict in a given population. Focus on community and group work is often more effective than individual work in increasing solidarity and mutual support. Good practice examples range from Northern Ireland to South Africa (O'Brien, 2003) to the development of advocacy practice centers in Israel, Jordan, and Palestine (Moshe-Gorodofsky, 2012), and memory work with three generations in Chile (Zavirsek, 2008). Doucet and Denov (2012) describe the work of social workers in Sierra Leone with women returnees from the civil war. In this, the clients were engaged in agricultural work alongside their social workers, in a situation in which these women were excommunicated by their own community and the social workers too had to find additional source of subsistence to their meager salary (Doucet and Denov, 2012).

Lindsay (2008) describes the work carried out by Palestinian school social workers who work with school children, their parents, and teachers when children are arrested, or who have to spend hours at checkpoints just to get to their workplace. Israeli social workers outlined what it meant to provide group mediation and emotional support to settlers evacuated from the Gaza strip, despite their vocal objection to this move (Nuttman-Shwartz, 2008).

Being local is an advantage in terms of knowing what is available and how to behave, an essential element of offering emotional and social support. Yet, being local can be a disadvantage when a civil war is raging, as it will be assumed that the social worker would side with the position taken by her or his national/tribal group, and hence defined as either part of 'us' or of 'the enemy,' and treated as such. Wairire (2008) describes such a situation in Kenya, where social workers try to mediate between tribes engaged in war on territory and cattle, and may risk their lives as a result. Working for international humanitarian agencies often entails carrying out social work activities, such as rapid assessment, responding to the need for immediate relief to a traumatic war event, and distributing food, medicine, and shelter. At the postconflict stage, the complexity to be confronted is even more demanding, including displaced people with little to do, buy, or sell, and worried about the people left behind, sick, or disabled.

Maglijlic-Holicek and Rasidagic (2008) have looked at the role of international NGOs in Bosnia-Herzogovina during the armed conflict there in the 1990s and its aftermath. Being a major tool in the application of the Dayton Peace Agreement, alongside the UN Peace Corps, they brought a new language, one of Western policy terms, and favored people who spoke English regardless of their other qualities or lack of them, over local people with the qualities and experience required for welfare work. The usual pitfalls in understanding local conditions, power games, and the diversified needs of the local population, played an additional role in making these overseas workers ineffective in their activities.

The Impact of Armed Conflict on Social Workers

Social workers participate in armed conflicts as both citizens and professionals. Like any other citizen, they too, experience increased physical and psychological vulnerability during the conflict. Yet, being citizens makes them more sensitive to the plight of specific groups, and alert to dangers that their loved ones may be facing while they are working as professionals. They may be committed to one of the two sides in the conflict.

As professionals, their mandate is similar to that of doctors – namely, to support their clients, and those people who become their clients during the armed conflict, regardless of the political affiliation of the latter. Working during a conflict may be difficult, in terms of physical danger, in being exposed to more acute losses and suffering than is normally the case (e.g., being the bearer of bad news to families), lacking necessary resources, or having to provide services to people whose views they object to. They may at times feel that they have to provide an example to others who are more vulnerable than they are (e.g., Palestinian school counselors, as described by Lindsay, 2008) or rise to the challenges that the conflict has provided and with that an opportunity to acquire a higher level of skills while highlighting to the general public the centrality of the social work contribution (Ramon, 2004; Baum and Ramon, 2010). For some social workers, the solutions adopted included deflecting the political nature of their work (e.g., mental health social workers in Belfast, as described by Campbell and McCrystal (2005)).

This context inevitably contains ethical dilemmas, which social workers may or may not be aware of. Cemlyn (2008) and Cemlyn and Nye (2012), looking at working with asylum seekers and travelers in England, indicate a range of such dilemmas resulting from the misfit of English legislation and policy and social work values. A minority of social workers are also engaged politically in campaigning for the values they believe in. Doing so from a platform of social work values is challenging, as highlighted in Moshe-Gorodofsky and Yudelevich's (2012) recent Israeli example.

Training and Educational Initiatives

Training and education offer the possibility of preparing service providers to offer targeted service for vulnerable groups caught in armed conflict. The trainers must be responsive to local conditions and sensitivities. For example, Moshe-Gorodofsky (2012) reports on a program aimed at preparing community advocates in Israel, Jordan, and Palestine as volunteers who respond to citizens experiencing the negative outcomes of armed conflict. The program enables the advocates to meet with their counterparts from the three countries in a neutral location, a once in a lifetime opportunity to meet the enemy as human beings who share values and ways of working with each other. Butler (2005) has led a fieldwork placement on a social work degree in Plymouth, in which students worked as social work trainees with refugees. Using a community development and welfare rights approach, this enabled students to have a much better understanding of the issues faced by refugees, develop empathy and demonstrate it, and try to resolve individual issues within the limitations imposed in the UK system on what refugees can and cannot do. This placement enabled students to test their adherence to the values taught on the degree course, and acquire the necessary knowledge and skills for this type of work.

Lindsay (2008) developed a training program for school counselors in the PA in Ramallah with local educators. This project enabled a 2-year training program for people from a variety of different disciplinary backgrounds to acquire knowledge and skills of providing a social work service to children and their teachers in an area where daily confrontation took place with Israeli soldiers and settlers. As the children were also engaged in protest activities against the occupation, and the workers were experiencing frequent difficulties in getting to the workplace through the roadblocks manned by the Israeli army, the newly qualified workers had to grapple with the issues that working under these complex conditions raised.

Examples of encountering obstacles related to the armed conflict and attempting to overcome them include a period of not being paid salaries due to the temporary withdrawal of EU funding because of an escalation of the political conflict in 2007. This affected directly the participants in the training who were in many cases the sole bread winners of their families. They considered taking strike action, but decided against it because of the importance of the training to them (Lindsay, 2008: pp. 233–234). Many of the participants were often stopped at border crossing by the Israeli army and it took them a long time to get to work, but put up with this obstacle as an example to the children they were working with

to strive toward a normal, nonviolent life as much as possible (Lindsay, 2008).

These examples raise the question of what input is needed within the basic qualification for social workers and in post-qualifying programs on armed conflicts. This question is particularly pertinent given the universal lack of attention to this issue in most well-established social work programs, despite attempts to encourage engaging with it (Duffy et al., 2013). The main reason usually given for this blatant lack is the already burdened curriculum, which is a polite way of saying that this issue is perceived as being of lesser importance than those already in the curriculum, and constitutes a denial of the significance of armed and political conflicts to an ever increasing number of countries across the globe.

The few existing initiatives are located in countries that are in a postacute armed conflict stage or have become more acutely aware of the impact of a long simmering conflict. An example of an initiative developed in the peace-building phase is provided by the school of social work at Queens University Belfast, Northern Ireland (Duffy, 2006; Campbell and Duffy, 2008). Service users who have experienced the impact of this conflict (locally known as 'The Troubles') were invited to address social work students in their introduction to the social work course. The invitation was to a local voluntary sector organization that works with families affected by 'The Troubles.' A well worked-out protocol was developed relating to the aims of this initiative, how to approach the students over the several seminars that were focused on this issue, what to tell or not tell about oneself, preparing the students who come increasingly from the two sides of the conflict, and how to evaluate the outcomes. This has been a successful initiative to judge by students' and presenters' comments (Duffy, 2006). However, social workers who have been affected in this conflict both as citizens and as professionals have not been invited to address the students thus far.

Conclusion

This article has highlighted the complexity of the issues entailed in social workers' understanding of and responses to the needs of populations experiencing armed conflict and included the impact of such conflicts on social workers themselves. The text covered research evidence, and the need for further research was also identified.

The insufficient attention paid to the impact of armed conflict in social work, and in social work education, and good practice examples were identified. Given that armed conflicts are part of the reality of the twenty-first century, it is hoped that the social work profession, its researchers, and educators will move from paying lip service to this crucial issue to give it the centrality it deserves. Likewise, there is a place for multidisciplinary work between social workers and other relevant disciplines to occur soon. In the context of armed conflict, social work knowledge takes key concepts concerning trauma, displacement, rebuilding communities, and research methods from psychology and sociology in particular. Practitioners work closely with varied health practitioners and educators and planners.

Epidemiologists, economists, and system managers also play an important role in different stages of an armed conflict. In meeting other disciplines on this issue, the role of social workers is both to inform and be informed, to decide which of the contributions of the other disciplines is of particular relevance to social work, as well as whether and how it fits social work values. An interesting example is provided by Dan Bar-On in his seminal article on the silence of psychologists as a discipline when faced with armed conflict (Bar-On, 2001). He argues that in this case, the allegiance to science, perceived as objective and beyond biases, prevents psychologists from taking a moral stand as well as a knowledge-building one. This article helps to understand similar positions taken by social workers, social work researchers, and social work educators, and understand how such a reaction does not fit the values that social work stands for, because these include social workers' defense of human rights. Compared to other social sciences disciplines, social workers' activities in armed conflict situations, and the research led by social work researchers on this issue, are far less known outside the discipline, calling for more effort to be put into wider dissemination of existing understandings of the issues entailed and evidence accumulated within social work.

See also: Asylum Seekers; Child Soldiers as Victims and Actors: Lessons from Ethiopia; Child Soldiers; Collective Memory, Anthropology of; Collective Memory, Psychology of; Collective Memory, Social Psychology of; Commemoration; Conflict and Socioemotional Development; Demographic Impact of Disasters; Disasters in Social, Cultural and Political Context; Gendered Care and Social Work Practice; Hazards and Disaster Research in Contemporary Anthropology; Human Rights and Social Work; Human Rights, History of; Mass Trauma: Psychopathological Effects across the Life Span; Natural Disasters: Health-Related Aspects; Social Work Theory; Social Work and Disasters; War, Political Violence, and Effects on Children.

Bibliography

Al Mikhamresh, S., Saneus, S., Neocleous, G., 2012. The need for political competence social work practice: lessons learned from a collaborative project on Iraqi refugees – the case of Jordan. British Journal of Social Work 42 (2), 1074–1092.

Amnesty International Newsletter, November 2012, pp.10–17.

Bar-On, D., 2001. The silence of psychologists or why there is no "post-Zionist" psychology? Political Psychology 22 (2), 331–345.

Baum, N., Ramon, S., 2010. Professional growth in turbulent times. Journal of Social Work 13, 163–182.

Beah, I., 2007. A Long Way Gone. Harper Perennial, London.

Butler, A., 2005. A strengths approach to building futures: UK students and refugees together. Community Development Journal 40 (2), 147–157.

Campbell, J., Duffy, J., 2008. Social work, political violence and citizenship in Northern Ireland. In: Ramon, S. (Ed.), Social Work in the Context of Political Conflict. Venture Press and IASSW, Birmingham, pp. 57–76.

Campbell, J., McCrystal, P., 2005. Mental health social work and the troubles in Northern Ireland: a study of practitioner experiences. Journal of Social Work 5 (2), 173–190.

Cemlyn, S.J., 2008. Social work education responses to political conflict issue in the UK. In: Ramon, S. (Ed.), Social Work in the Context of Political Conflict. Venture Press and IASSW, Birmingham, pp. 189–213.

Cemlyn, S.J., Nye, M., 2012. Asylum seekers young people: social work value conflicts in negotiating age assessment in the UK. International Social Work, Special Issue on Armed Conflict 55 (5), 675–688.

Doucet, D., Denov, M., 2012. The power of sweet words: local forms of intervention with war-affected women in rural Sierra Leone. International Social Work, Special Issue on Armed Conflict 55 (5), 612–628.

Duffy, J., 2006. Participating and Learning: Citizen Involvement in Social Work Education in the Northern Ireland Context. Social Care Institute for Excellence/ Social Work Northern Ireland Social Care Council, London.

Duffy, J., Ramon, S., Guru, S., Cemlyn, S., Lindsay, J., Shwartz-Nutman, O. Opportunities and challenges in developing a social work curriculum on political conflict. European Journal of Social Work 16, forthcoming.

Felsman, K., 2001. Young children in complex emergencies. Reading for Child and Youth Care Workers. Issue 34 online.

Fisher, J.R.W., Cabral de Mello, M., Izutsu, T., Vijayakumar, L., Belfer, M., Omigbodun, O., 2011. Adolescent mental health in resources-constrained settings: a review of the evidence in the nature, prevalence and determinants of common mental health problems and their management in primary health care. International Journal of Psychiatry 57 (Sl), 57–78.

Freud, A., Burlingham, D., 1943. War and Children. Medical War Books, New York.

Guru, S., 2012. Reflections on research: families affected by counter-terrorism in the UK. International Social Work, Special Issue on Armed Conflict 55 (5), 689–703.

Halbwachs, M., 1992. On Collective Memory. Chicago University Press, Chicago.

Kelly, M., 2004. Qualitative evaluation research. In: Seale, C., Gobo, G., Gubrium, J., Silverman, D. (Eds.), Qualitative Research Practice. Sage Publications, London, pp. 463–477.

Krill, F., 2000. The Elderly in Situations of Armed Conflict. International Committee of the Red Cross, Geneva.

Kucukalic, A., Dzubar-Kulemovic, A., Ceric, I., Jacobsson, L., Bravo-Mehmedbasic, A., Preibe, S., 2005. Regional collaboration in reconstruction of mental health services in Bosnia. Psychiatric Services 56 (11), 1445–1447.

Lincoln, Y.S., Guba, E.G., 1985. Naturalistic Inquiry. Sage Publications, London.

Lindsay, J., 2008. Dilemmas, opportunities, obstacles and achievements in training Palestinian social care workers and social workers. In: Ramon, S. (Ed.), Social Work in the Context of Political Conflict. Venture Press and IASSW, Birmingham, pp. 217–244.

Maglijlic-Holicek, R.A., Rasidagic, E.K., 2008. The role of NGOs in societies facing war: experiences from Bosnia and Herzogovina. In: Ramon, S. (Ed.), Social Work in the Context of Political Conflict. Venture Press and IASSW, Birmingham, pp. 123–146.

Miljenovic, A., Zganec, N., 2012. Disintegration and possibilities for rebuilding of war-affected communities: the Vojnic municipality case. International Social Work, Special Issue on Armed Conflict 55 (5), 645–661.

Moshe-Gorodofsky, M., 2012. Community-based human rights advocacy practice and peace education. International Social Work, Special Issue on Armed Conflict 55 (5), 740–753.

Moshe-Gorodofsky, M., Yudelevich, D., 2012. Organising across-identity group division in the context of acute political conflict – operation cast lead. British Journal of Social Work 42, 1060–1073.

Musee d'Art modern de la Ville de Paris-Musee, the Guggenheim Museum, Bilbao, 2013. L'Art en Guerre: France 1930–1947: From Picasso to Dubuffet. Guggenheim Museum, Bilbao visited 21st August 2013.

Nuttman-Shwartz, O., 2008. Working with "others" in a context of political conflict: is it possible to support clients whose views you disagree with? In: Ramon, S. (Ed.), Social Work in the Context of Political Conflict. Venture Press and IASSW, Birmingham, pp. 36–56.

Nuttman-Shwartz, O., Dekel, R., Tuval-Mashiach, T., 2011. Post-traumatic stress and growth following forced relocation. British Journal of Social Work 41 (3), 486–501.

O'Brien, C.R., 2003. Community Development and Post Conflict Resolution: An Examination of the Potential for Complementary Strategies in Post-Settlement Contexts, with Special References to Northern Ireland and South Africa. Ph.D. thesis. University of Bradford, Bradford.

Ramon, S., 2004. The Impact of the 2nd Intifada on Israeli Arab and Jewish social workers. European Journal of Social Work 7 (3), 285–303.

Ramon, S. (Ed.), 2008. Social Work in the Context of Political Conflict. Venture Press and IASSW, Birmingham.

Ramon, S., Magljlic, R.A., 2012. Social work, political conflict and displacement. In: Lyons, K., et al. (Eds.), The Sage International Handbook of Social Work. Sage Publications, London, pp. 311–325.

Ramon, S., Zavirsek, D. (Eds.), 2012. Armed conflict. International Social Work 55, 2 (special issue).

Ramon, S., Campbell, J., Lindsay, J., McCrystal, P., Baidun, N., 2006. The impact of political conflict on social work: experiences in Northern Ireland, Israel and Palestine. British Journal of Social Work 36 (3), 435–450.

Skran, C.M., 2012. Research on UNCHR Community Empowerment Projects in Sierra Leone. UNCHR Policy development evaluation services. UNCHR, Geneva.

Wairire, G., 2008. The challenge of social work in the Kenyan context of political conflict. In: Ramon, S. (Ed.), Social Work in the Context of Political Conflict. Venture Press and IASSW, Birmingham, pp. 101–122.

Zavirsek, S., 2008. Social work as memory work in times of political conflict. In: Ramon, S. (Ed.), Social Work in the Context of Political Conflict. Venture Press and IASSW, Birmingham, pp. 147–168.

Relevant Websites

ECHO, 2013. Factsheets Children in Conflict. http://ec.europa.eu/echo/files//aid/countries/factsheets/…/children_conflict (accessed 29.08.13.).

UNHCR The United Nationals Refugees Agency. www.unhcr.org (website accessed 30.12.12.). Statistics and operational data: statistical catalogue. Global trends 2011 18.6.12. Asylum Levels in industrialised countries 2011 15.10.12. Asylum levels in industrialised countries first half of 2012 27.3.12.

Watchlist, 2003. The impact of armed conflict on children in the Democratic Republic of Congo. www.watchlist.org (accessed 28.08.13.).

Arms Control

Coit D Blacker, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 2, pp. 748–751, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The article begins by defining the term and distinguishing between 'arms control' and 'disarmament' in both theory and practice. The analysis continues with a review of the negotiating history, with particular reference to US–Soviet efforts to delimit the bilateral nuclear competition from the mid-1960s to the early 1990s. The third section focuses on the disruption to and the broadening of the arms control agenda occasioned by the collapse of the Soviet Union in 1991 and the spread of nuclear weapons and ballistic-missile technologies (i.e., the problem of 'horizontal proliferation') to second- and third-tier states. The article concludes with a detailed look at the enduring tension in arms control and nuclear strategy between efforts to prevent the outbreak of war through the codification of nuclear deterrence and the renewed interest in national (and theater) missile defense as the preferred approach to the maintenance of national and international security.

Arms control, a term popularized in the early 1960s, may be defined as the effort, between and among countries, to negotiate binding limitations on the number and types of armaments or armed forces, on their deployment and disposition, or on the use of particular types of armaments. It also includes measures designed to reduce the danger of accidental nuclear war and to alleviate concerns about surprise attack. Although the two terms are often used interchangeably, it is distinct from disarmament, which has the more ambitious objective of seeking to eliminate, also by international agreement, the means by which countries wage war (Blacker and Duffy, 1984). The goal of eliminating war extends far back in history, but in modern times disarmament came into focus with the Hague Peace Conferences in 1899 and 1907. These and subsequent efforts met with what can only be termed limited success.

Schelling and Halperin, in their seminal work, *Strategy and Arms Control*, first published in 1961, listed three objectives for arms control: to reduce the likelihood of war; to limit the extent of damage should war occur; and to reduce expenditures on military forces. As US and Soviet weapons arsenals mushroomed during the Cold War and each country spent liberally to keep pace with the other militarily, later analysts tended to offer less sweeping goals. These included reducing the number of nuclear weapons and redirecting the arms race into areas less likely to threaten the stability of the international system.

The end of the Cold War and the collapse of the Soviet Union gave rise to expectations, both at the expert level and among publics more broadly, that the sudden cessation of the superpower arms race would create opportunities for radical reductions in the nuclear and conventional weapons arsenals of the major powers.

Although the leading industrialized countries have taken some steps in that direction – US armed forces, measured in terms of total numbers, have declined by one-third since 1990 – progress in eliminating the nuclear-weapons stockpiles of the US and Russia has proven to be an elusive goal. To compound the problem, the number of nuclear-armed states has actually grown in recent years as India and Pakistan, in a series of highly publicized weapons tests in 1998, announced their arrival as full-fledged nuclear powers. The means to deliver these weapons, as well as chemical and biological agents, across hundreds (and even thousands) of miles has also spread as countries such as North Korea, Iraq, and Iran continue to invest heavily in programs to develop and deploy long-range ballistic missiles.

Arms control, particularly between the US and the Soviet Union, aroused controversy from the outset. The Limited Test Ban Treaty, signed by representatives of the US, the United Kingdom, and the Soviet Union in 1963, provoked sharp debate when US president John F. Kennedy submitted this arms control 'first' for Senate approval. Some critics saw the treaty, which prohibits the testing of nuclear weapons above ground, under water, and in space, as unenforceable. These and other concerns notwithstanding, the treaty eventually was ratified and fears of Soviet noncompliance proved to be unfounded.

More favorably received in the US was the United Nations-sponsored Nuclear Non-Proliferation Treaty (NPT), which sought to restrict the size of the 'nuclear club' by inducing non-nuclear weapons states to renounce the acquisition of such weapons in exchange for a commitment (among other pledges) on the part of the nuclear-weapons countries to reduce their own arsenals. The NPT, which entered into force in 1970, was renegotiated in 1995, at which time the signatory states agreed to extend the treaty's provisions for an unlimited period of time. Conspicuous by their absence as states-parties to the NPT are the two newest declared nuclear powers, India and Pakistan, as well as Israel, which is believed to possess a small but sophisticated arsenal of nuclear weapons.

The NPT was critically important in facilitating the start of bilateral US–Soviet negotiations in 1969 to limit central strategic forces. Known by the acronym SALT, for Strategic Arms Limitation Talks, the negotiations resulted in two arms-control agreements in 1972. The first and more important was the Antiballistic Missile (ABM) Treaty, by which the two countries agreed to limit the number of ABM sites and thus not deploy nationwide defensive systems to protect their homelands against nuclear-missile attack. The second accord was a 5-year freeze on the construction of long-range land- and sea-based ballistic missile 'launchers' (underground silos and submarine missile tubes, respectively). The so-called Interim Agreement on Offensive Weapons was a temporary measure to slow

the competition in offensive weaponry, pending negotiation of a more permanent and restrictive treaty (Newhouse, 1973).

The second phase of the negotiations, lasting from 1972 to 1979, led to the signing of several agreements, including an accord further limiting US and Soviet ABM deployments and the 1974 Threshold Test Ban Treaty, which restricts the yields of underground nuclear weapons tests. The most important agreement concluded during this period was the 1979 SALT II treaty, a lengthy and complex document that attempted to extend and refine many of the provisions of the 1972 Interim Agreement (Talbott, 1979). The US Senate never ratified the treaty, however, largely because of the dramatic deterioration in superpower relations that began during the abbreviated administration of President Gerald Ford and accelerated during the term of his White House successor, Jimmy Carter. Despite the ambiguous legal status of the treaty following the failure to obtain its ratification, both the US and the Soviet Union abided by most of its provisions well into the 1980s.

Negotiations on US and Soviet strategic nuclear forces, renamed the Strategic Arms Reduction Talks (START) by President Ronald Reagan, resumed in the summer of 1982. Through the remainder of Reagan's presidency, the two sides reached consensus on many key points, including the desirability of 50% reductions in long-range nuclear forces and the need for intrusive, on-site inspections to prevent cheating.

Among the issues not resolved was how to construct the preferred relationship between strategic offensive and defensive forces. In a dramatic reversal of its earlier negotiating position, the US now favored the rapid development and deployment of nationwide defensive systems, as shown by its sponsorship of the Strategic Defense Initiative (SDI), first outlined by Reagan in March 1983 in a nationally televised address (Fitzgerald, 2000; Nolan, 1989). The Soviet Union was sharply critical of SDI and resisted the conclusion of any new agreement to reduce strategic offensive weapons, pending a commitment by the US to abide by the terms of the ABM treaty, narrowly interpreted, through at least the end of the century.

The rise to power of the Soviet leader Mikhail Gorbachev in 1985 signaled a fundamentally new phase in arms control. One of Gorbachev's most important foreign policy objectives was to curtail sharply the political rivalry between the USSR and the West by, among other steps, delimiting their military competition. The first major result of this policy was the Intermediate-Range Nuclear Forces (INF) Treaty, concluded in 1987, which eliminated all US and Soviet land-based nuclear missiles with ranges between 500 and 5500 kilometers. This was followed in 1990 by a treaty to reduce conventional forces in Europe (CFE), signed by 22 member-states of the North Atlantic Treaty Organization (NATO) and the now-defunct Warsaw Pact. At about the same time, US and Soviet negotiators completed work on the long-awaited START treaty, and in July 1991, President George Bush traveled to Moscow to join Gorbachev in signing the accord. Less than 6 months later, Gorbachev – the architect of the most ambitious reform program in the 74 year history of the Soviet Union – resigned as president and the USSR ceased to exist.

At a hastily called summit meeting in June 1992, Bush and Russian Federation president Boris Yeltsin agreed to press for early ratification of the START treaty. They also pledged to conclude a second and more ambitious agreement to reduce US and Russian strategic nuclear arsenals by up to two-thirds within a decade and to eliminate all multiple-warhead land-based missiles. Some 2 weeks before the inauguration of Bill Clinton as US president in January 1993, the two sides made good on their promise and concluded the START II treaty. As relations between the US and Russia deteriorated during the second half of the 1990s, the treaty itself languished; although finally ratified by both sides, nearly a decade after its signing most of the agreement's provisions still had not been implemented.

The US–Russian arms-control agenda has also been complicated by the confused, and confusing, nuclear legacy of the Soviet Union. In the wake of the Soviet collapse, four of the country's now independent republics, including Russia, found themselves in possession of thousands of nuclear weapons and hundreds of long-range delivery systems. Under pressure from Russia and the West, in May 1992 the governments of Belarus, Kazakhstan, and Ukraine promised to abide by the terms of the 1991 START I agreement and to join the NPT as non-nuclear weapons states. With financial and technical assistance from the US, all nuclear warheads and their associated missile systems deployed on Belarusian, Kazakh, and Ukrainian territory eventually were disarmed and dismantled. The denuclearization of the three former Soviet republics constitutes the most important – and perhaps the only unambiguous – arms control success story of the 1990s.

The abrupt end of Soviet rule gave rise to a second kind of security problem for which policymakers everywhere were ill prepared. The frequency and intensity of ethnically and religiously inspired conflicts in such heretofore remote parts of the former USSR as Tajikistan and the southern Caucasus increased dramatically following the collapse of central authority in Moscow. The breakdown of political control in these and proximate regions, coupled with the extreme poverty that afflicted many of the people caught in the fighting, served both to prolong these struggles and to frustrate diplomatic efforts to contain them.

In addition, the unraveling of the Soviet Union's alliance relationships left a number of countries politically orphaned and therefore less secure. This too served to increase regional instability. It also encouraged some states – North Korea being a case in point – to seek to develop nuclear weapons of their own despite the determined opposition of the major powers, especially the US.

At the start of the twenty-first century the most important challenge to arms control, however, is the prospective change in the relationship between long-range, offensive nuclear forces and weapons designed to defend against them. For the last 40 years, what Philip Green characterized in the mid-1960s as the deadly logic of nuclear deterrence helped preserve the uneasy truce between Washington and Moscow. In its simplest form, deterrence held that no rational leadership in possession of nuclear weapons would ever intentionally authorize their use against a nuclear-armed adversary because of the near-certain knowledge that the victim of such an attack would retaliate in kind. With effectively no ability to ward off or deflect such a retaliatory strike, the would-be aggressor would thus be deterred from initiating a nuclear exchange in the first place (Brodie, 1965; Schelling, 1960).

As the capacity to build nuclear weapons spreads to countries other than the so-called great powers, interest in acquiring the means to defend against their possible use has grown, particularly in the US. Even a comparatively modest system of active defenses – one designed to defeat a handful of incoming ballistic missile warheads launched from whatever quarter – arouses concern, however, because of the latent ability of such a system to expand and improve over time. The larger and more robust a system of national missile defense becomes, the better it will be at defending against more complex threats. The more able it becomes, in other words, the more threatening it will seem to the more established nuclear powers, including Russia and China, for whom the unchallenged ability to retaliate with overwhelming force constitutes the bedrock of their security (Wilkening, 2000). According to the classic tenets of arms control, the large-scale deployment of strategic defensive systems could therefore erode stability and increase the likelihood of war.

It is probably not beyond human ingenuity to design and construct a 'mixed' strategic environment that allows for a modicum of defense while preserving nuclear deterrence in its essentials. Given the understandable urge to escape the persistent threat of nuclear annihilation, it seems safe to assume that governments will persist in their efforts to square this strategic circle. Policymakers everywhere would do well to remember, however, that nuclear deterrence, for all its flaws, has kept the peace for half a century, and that any attempt to replace it with something else is likely to entail serious risks and potentially enormous costs.

Editor's Note

Dr Blacker's recounting of arms control leaves off near the end of the 1990s and only hints at the arms control issues that would plague the world in the ensuing years, specifically, how to keep weapons of mass destruction out of the hands of non-state actors, terrorist organizations, splinter 'states' and other potentially dangerous groups. For a comprehensive update covering all this and more, see Williams and Viotti, 2012.

See also: Conflict and Consensus; Conflict and War, Archaeology of: Weapons and Artifacts; Geopolitics; Military Geography; Military History; War, Social Causes and Consequences.

Bibliography

Blacker, C.D., Duffy, G. (Eds.), 1984. International Arms Control: Issues and Agreements, second ed. Stanford University Press, Stanford, CA.

Brodie, B., 1965. Strategy in the Missile Age. Princeton University Press, Princeton, NJ.

Fitzgerald, F., 2000. Way Out There in the Blue: Reagan, Star Wars, and the End of the Cold War. Simon & Schuster, New York.

Green, P., 1966. Deadly Logic: The Theory of Nuclear Deterrence. Ohio State University Press, Columbus, OH.

Newhouse, J., 1973. Cold Dawn: The Story of SALT, first ed. Harper & Row, New York.

Nolan, J.E., 1989. Guardians of the Arsenal: The Politics of Nuclear Strategy. Basic Books, New York.

Schelling, T.C., Halperin, M.H., 1961. Strategy and Arms Control. Twentieth Century Fund, New York.

Schelling, T.C., 1960. The Strategy of Conflict. Harvard University Press, Cambridge, MA.

Talbott, S., 1979. Endgame: The Inside Story of SALT II, first ed. Harper & Row, New York.

Wilkening, D., 2000. Ballistic Missile Defense and Strategic Stability. International Institute for Strategic Studies, London. Adelphi Paper 334.

Williams, Robert E., Viotti, Paul R. (Eds.), 2012. Arms Control: History, Theory and Policy, vols. 1 and 2. ABC-CLIO, Santa Barbara, CA.

Aron, Raymond (1905–83)

Panagiotis Christias, University of Cyprus, Nicosia, Cyprus

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Raymond Aron was too respectful of each person's freedom to found a school of thought. But his mark in French political history and thinking is just as profound due to the fact that, far from relying on a network of disciples and institutions, it is anchored in an intellectual heritage that transcends partisan differences or academic studies. Aron succeeded in continuing and revitalizing, in the midst of the century of ideologies, the tradition of French political liberalism. More than that, he was the academic teacher and the intellectual who defended the *primacy of politics* in an era dominated by social sciences and economy. The rebirth of the political thinking in France in the 1980s is his legacy. In France, his attitude and his writings – particularly the publication of *L'Opium des Intellectuels* (1955) – contributed decisively in progressively detaching the intellectuals from antiliberal regimes and, from there, to the resistance of French society to antiliberal regimes. At the same time, Aron's action and thinking were of tremendous importance in the conversion of a large part of governmental and administrative elite to the Atlantic Alliance and the building of the European Community. Abroad, he acquired a huge moral notoriety and had a privileged relationship with numerous scholars, philosophers, and statesmen, such as Henry Kissinger with whom he maintained a frequent dialogue. Aron thus is a unique example of a twentieth-century French intellectual who was both patriot and European, republican and liberal, antitotalitarian and cosmopolitan. In the face of collective passions and demagogy, which overtake regularly democracies in general and French political life in particular, Aron played the major part, according to Claude Lévi-Strauss of a 'teacher of intellectual hygiene.'

Biographical Elements

Raymond Aron was born on 14 March 1905 in Paris. His family originated from Lorraine, a territory partly attached to the German Empire at the moment. His father was a professor of law. In 1924, he enters the *École Normale Supérieur* (ENS), the most prestigious public institution of higher education, along with Paul Nizan and Jean-Paul Sartre. All three were intellectuals, philosophers, and editorialists with great influence on public opinion on major political issues. He succeeded in his philosophy aggregation concourse in 1928, and in 1930, he taught as a lecturer at the University of Köln, and right after at the Academic House of Berlin from 1931 to 1933. These years in Germany and Berlin would shape the intellectual choices of the young philosopher, and would orientate him toward German Sociology, especially Max Weber, giving him a fleur of the tumultuous twentieth-century politics that was about to transform Europe. Back to France, he taught at the College of Le Havre (1933–34), and then occupied the post of secretary of the *Social Documentation Centre* in the ENS till 1939. In 1938, he publicly defended his major thesis, *An Introduction to the Philosophy of History* and his minor one, *The Critical Philosophy of History*. This second work along with *Contemporary German Sociology* (Aron, 1935) was the first presentation of great German philosophers and sociologists, like Dilthey, Tönnies, Weber, Simmel, and Tröelsch in France.

In 1939, he was drafted and fought the Nazi Germany in what was known as the *drôle de guerre* that led to the occupation of France. By the end of June 1939, he embarked for London, where he joined General De Gaulle and the Free French Forces, whose emblem was Larraine's Cross. In London, he became chief of the editorial board of the journal *La France Libre*. After the liberation, he continued his public interventions. In 1945, he founded avec Jean-Paul Sartre la revue cosmopolite *Les Temps modernes*, and collaborated with Albert Camus in the *Combat* (1945–46). He worked as an editorialist at the *Figaro* (1947–77), and the *Express* (1977–83). Thus, he participated actively in the political and ideological debates of his times. He would defend political and economical liberalism and liberal values in a public space obsessed by ideas about socialism and communist revolution. His opposition to Sartre and the communist camp is legendary and incarnated in the intellectual mode of the 1960s: "better be wrong with Sartre than right with Aron." His opposition would culminate in his major work on *The Opium of the Intellectuals* (1955), where he denounces a self-blinded left intelligentsia bringing to light its erroneous ideological framework. For more than 40 years, he analyzed international and interior politics, strategies, diplomacies, and conjectures, and established himself as one of the best analysts of the Cold War era.

From 1945 to 1978, he occupied a number of prestigious positions in the French academic system: professor at the *Institute of Political Studies* in Paris (1945–54) and the *National School of Administration* (1945–47), professor holding the chair of sociology in the Faculty of Letters of Paris-Sorbonne University (1955–67), Director of the 6th section of the *École Pratique des Hautes Études* (1960–78), which would become the *École des Hautes Études en Sciences Sociales*, Professor at the *Collège de France* (chair of sociology of modern civilization, 1970–78). During these years, he contributed to the creation of a number of political philosophical reviews like *Commentaire*, *Preuves*, *Contre-point*, which formed the ideas of liberal political thinking and philosophy in France.

One of the great voices of his time, he had a particular pathos for dialogue and open mindedness, that made him an eminent interlocutor of philosophers with conflicting ideologies, like Cornelius Castoriadis and Leo Strauss. He has formed a great number of philosophers and political thinkers like

Pierre Manent and Jean-Claude Casanova, who he then sent to Chicago to conclude their studies under the direction of Leo Strauss, Jean Bæchler, Alain Besançon, and influenced the works of thinkers like Kostas Papaïonnou, Cornelius Castoriadis, Julien Freund, André Glucksmann, and Pierre Hassner. Raymond Aron died in Paris on 17 October 1983, a little after the publication of his *Mémoires*, which was an immediate success. French politician and major European political actor, Jacques Delors described him as "one of the greatest intellectuals of our time," and the anthropologist Claude Lévi-Strauss as "our last sage."

Works

Aron has left behind him a polyvalent and multi-axis work. He was an acute observer of social reality and contemporary history. He undertook studies on philosophy of history, on sociology and sociological theory, on political theory and ideology, on war and international relations. He wrote major works on liberty, ideology and totalitarianism, economy, class and industrial society, Cold War politics, war theory and diplomacy in an open dialogue with Marx, Weber, Pareto, Tocqueville, and Clausewitz, with a constant orientation to the actual and present time conjunctures and interests.

On Philosophy of History

Influenced by the German school of sociology, Aron's work on philosophy of history focuses on the historical condition of the social researcher and theoretical knowledge of our societies. In his *Introduction to the Philosophy of History*: *An Essay on the Limits of Historical Objectivity* (Aron, 1938), Aron enters in a dialogue between Hegelian heritage and Weber's critique of Hegel's philosophy. He distinguishes between 'objective spirit' and 'objectivated spirit.' If 'objective spirit' is a quasi-metaphysical entity, 'objectivated spirit' is a part of everyday life and understanding. Aron takes the example of a bell in the chambers of a bourgeois apartment. He argues that in order to understand the use and function of this object, we have to understand the logic of class society, the distinction between masters and servants: the master calls using the ring, and the servant answers the call obeying to the wishes of his master. This interaction implies of course the logic of salary work and moneyed interest. The argument is based on the fact that historical interpretation must be based in a reconstruction of a theatrical scene that will ultimately reveal the nature of the object in question. Historical knowledge is thus contextualized, and the limits of historical interpretation are the limits of our imagination to figure out contextualized situations. In this sense, there is no unique interpretation of historical events, as the process of historical understanding implies an open dialogue between the present state of the researcher and his ability to project to past time experiences. The very meaning of historical experience and experience of the past is constantly varying regarding to the everyday experience in and of the present. In other words, both Hegel's theory of the possibility of an objective knowledge and Weber's intuition of the distinction between values and facts are put here in difficulty. Facts are value permeable, and values can be understood as facts, that is, they can be 'objectivated,' in the process of historical understanding. Furthermore, the very idea of an 'objective' knowledge of our present experiences can only be understood as a sum of social interactions in which the social researcher is both social actor and historical judge. This does not, however, invalidate any historical knowledge: it just emphasizes on the fact that there is no ultimate interpretation and understanding of history, but only a perpetual dialogue and even conflict between social scientists.

But can such a knowledge be pertinent for practical action? In *Politics and History* (Aron, 1978), Aron gives an example, using Thucydides' historical narrative as to how ancient history, or our understanding of it, can be useful for political actors in the twentieth century. He argues that despite of the social and economical complexity of modern day societies, Thucydides' narrative sheds some light in basic political interactions, as war and war-based diplomacy. Similarities between fifth century BC Athens and twentieth century AD Cold War experience can only be understood if we consider the basic human structures of war and politics. Facts such as alliances and power diplomacy, dilemmas such as justice and the use of force remain intact as an anthropological basis to the understanding of human societies how complex may they become. In other words, and as far as power confrontation is concerned, Thucydides' experience is as good as ours. History, as Aron understands it, that is as a "skeptical intellectual shaped by the experience of fascism and World War II" understands it, leaves little doubt as to the inevitable preeminence of politics and war in human societies. Historical synthesis of human progress by explaining events to come in the light of human creativity and accomplishments only tells one side of the story. Eventually, we will come to Aristotle's conclusion that human events are of probabilistic and not of deterministic nature, and such is the nature of their narratives and understandings of them.

Aron believed in the utility of historical research combined to philosophical dialogue with the great thinkers of the past. Difficult and fundamental concepts, such as power and liberty, shaped in the past can help us forge our future: political decisions can hardly be taken without the experience and skillful interferences with such concepts and philosophical understandings. As history interferes with politics, it becomes clear to Aron that present and past historical events can only be understood in terms of a complex of variables to be determined and examined alongside the comparison of political experiences and philosophies in different historical and social contexts. This is actually the meaning and the sense of a comparison between Machiavelli and Marx, or between Tocqueville and Marx on liberal society: to define the fundamental concepts of power and liberty that influence our political action. Without this historical knowledge, modern era's computer-based war games are blind.

On Sociology and Sociological Theory

Raymond Aron is mostly known in the academic world as a sociologist. He held the chair of Sociology at the Paris-Sorbonne University (1955–67) and later a chair of sociology in the *College of France* (1970–78), a most prestigious advanced studies research institution in Paris. In 1967, he published *Les Étapes de la pensée sociologique* (Aron, 1967, 1968c), a book that

would determine the mainstream sociological thinkers in France for the years to come. Aron counts between the fathers of sociological thinking Comte, Marx, Pareto, Durkheim, and Weber, and also Montesquieu and Tocqueville. Regrettably, he did not include Georg Simmel, although he was the first to present him to the French public as an historian. An "intellectual gallery of portraits," this book is a consecration of the founding fathers of sociology. In the first part of the book, Aron distinguishes three schools: (1) The French school of political sociology, following Montesquieu and Tocqueville, a non-dogmatic school mostly interested in politics but who did not ignore social infrastructure. Aron includes himself in this school. (2) The school of Auguste Comte that leads through Durkheim to contemporary French sociology. This school minimizes the role of politics and economy emphasizing in the social aspect of unity through consensus. (3) The Marxist school. The Marxist theory, explains Aron, exists in two versions: a doctrine of the State leading the believers in a classless society fantasy, and a subtle one, destined to be understood only by intelligent people. According to this theory, the socioeconomical structures and the historical process can explain society, as a whole. The latter leads necessarily to the final triumph of socialism. The second part of this book presents Pareto, Durkheim, and Weber. All three, comments Aron, belong to the generation *tournant de siècle*, who lived in a relative social stability, but shared a pessimistic sentiment of Europe in crisis. All three, notes Aron, felt that the withdrawal of religion caused by the rise of modern science and technology left societies without moral guidance. They believed that social science could, in a certain extent, contribute to the restoration of social morality, but despite this belief, they understood well that societal values were beyond the rational grasp of science.

Aron often described himself as a *Marxian* but not as a Marxist, rejecting Marxism's poetical enthusiasm and embracing Marx's scientific and analytical prose. In 1962, Aron pronounced his courses on *Marx's Marxism* in the Sorbonne, and repeats the same seminars in the *Collège de France* in the 1970s. Jean-Claude Casanova, publisher of *Le Marxisme de Marx* (Aron, 2002), notes the very words of Aron in his inaugural course: "Today, it has been 30 years that I have been studying Marx." Although he confesses his inability to become a Marxist, Aron acknowledges that Marx was his initiator on social philosophy. Aron hopes to establish that Marx's critical power resides not in the concept of alienation, as the young Marxists of his time believed, but in his critique of religion, law, and economy. Aron refuses to accept that the Marx of the *Manuscripts of 1844* should shadow his mature work. Alienation was a concept of the young Marx that did not last long. Marx's genius consists in questioning not only social reality but also our own conscience of reality. Beginning by a critique of religion, he continues with the critique of Law, and arrives naturally in the critique of political economy. The critique of religion is also the critique of the illusions of religion and in the same time the critique of the society that was produced by these illusions. This kind of reasoning can be transposed in economy. Economists think that profit comes from the earth or the capital, but this is an illusion: it comes from the exploitation of workers. Economical actors live in a false world, and cannot realize it because they refuse to recognize themselves for what they are. This blindness of all classes, proletarians and bourgeois, even of economists is what allows the system to work. For Aron, the *Capital* is an 'economico-philosophical' book that combines German philosophy and English political economy. Aron himself went through the same exercise: write a philosophical book in an economical language.

In the same year 1962 when he commenced his seminars on Marx, Aron published *Eighteen Courses on Industrial Society* (Aron, 1962a), courses taught in the Sorbonne during 1955 and 1956. In this work he notes that industrial societies are based on the idea of progress, and that productivity means above all progress in science, technology, and evolution of means. A strong growth needs perpetual transformations of industry and production, which leads to social and political transformations. He then examines the growth of Soviet industry and understands that quantitative industrial growth is not accompanied by positive qualitative change of citizen's living but rather to forced population displacement from the countryside to the industrial areas with a much lower standard of living. This is according to Aron the sign of the Soviet failure to provide a better way of life. Finally, he argues that the capitalist system leads to democracy – the market, syndicates, and consumers are counterpowers – better standards of living, and deceleration of growth. His previsions of 1955 both came true less than 50 years later, when Soviet economy collapsed and capitalist economy is decelerating.

On Political Theory and Ideology

Aron belongs to the French School of political sociology and follows the works and methods of Montesquieu and Tocqueville. He analyses together liberal democracy and its enemies, totalitarian regimes (Aron, 1965). Modern societies, he argues, tend by their very nature to equality, a lesson taught by both Montesquieu and Alexis de Tocqueville. In this longing for even more equality, two perspectives are open for modern societies: either relative equality through democratic reforms, which would lead to a pluralist constitutional regime, or absolute equality through revolution, and the imposition of a monopolistic party regime. Aron insists on the conflictual and pluralist character of liberal democracies, as he opposes two types of regime: a constitutional-pluralist one and another of a monopolistic party. This opposition actually recovers a series of oppositions: political and economical concurrence or monopole, constitution or revolution, social pluralism or absolutism, party State or partisan State (Audier, 2004). Respect of legalism and the sense for compromise (Aron, 1965) is the principle of the constitutional-pluralist regime, argues Aron. It is what is denied in the monopolistic party regime. Such a regime would come to extinction if it were corrupted by the democratic sense of compromise. Instead, absolutism is its strength, and its principle is the sentiment of fear and faith. Militants' faith and non-militant's fear is the force and capacity for action for totalitarian regimes as dialogue and open-minded conflict is for liberal pluralist ones. Of course, Aron argues, there are many variations of one or multiple party regimes and even mixed forms, like monopolistic regimes introducing a Constitution and respecting individual rights or pluralistic regimes presenting an oligarchic character. And then again, liberal regimes are also open to corruption.

It was the special historical circumstances of the academic year 1957–58, when Aron offered the seminars on *Democracy and Totalitarianism* (Aron, 1965) that pushed his reflection on the corruption of constitutional-pluralist regimes. As Aron states in his introduction, "a legal savior inherited a corrupted Republic." The passage from the 4th to the 5th French Republic, and the end of the Algerian War, which ended with the independence of Algeria, gave Aron a critical insight into the liberal pluralistic regimes. There are three major types of corruption that can endanger liberal constitutional regimes claims Aron. The first appears when the party system does not any more correspond to the various groups of interests or when no stable authority can come out of the rivalry between parties. The second is the corruption of the public spirit, or the corruption of the *principle*, as Montesquieu would have called it. Either the partisan spirit of the parties' militants is too strong and erases any notion of common good, or the spirit of compromise is too strong and there can be no decision making. Last, corruption could come of the failure of the social infrastructure, if class conflicts are too intense and the political system of the parties is incapable of withholding them within limits. Excessive oligarchy or excessive demagogy can lead to social dissolution and political decay. Aron notes that corruption of liberal regimes is either due to not much time in power, which means that the political system is fresh and does not have yet deep roots into society, or too much time in power, and a long period of decomposition. A corrupted regime, concludes Aron, can continue existing for a long period of time, as revolution is not necessarily what follows. But it can also, under specific conditions, like the Republic of Weimar in Germany in 1933, crumble into totalitarianism.

Aron defines the totalitarian regimes as a mixture of ideological frenzy and police terror. In addition, totalitarian regimes center on a controlled economy based on State planning. In the case of the Soviet regime, Aron makes some interesting remarks. First, that there have been free elections after the takeover in 1917. This first Parliament was of course dissolved, as it did not serve the Bolsheviks. Second, that there have been many Constitutions over the years. And third, that there was a separation between the State and the party, even if only for diplomatic reasons. As the party proclaims liberation by revolution, and the State has to negotiate and develop diplomatic relations with Western capitalistic forces, it would be difficult for the State to bear openly the party's ideology. This shows according to Aron that if Hitler and the fascist regimes show open hostility to democratic principles, the communist regime embraces them, even if it does not apply them. One of the forces of the communist totalitarianism is that it proclaims his faith in the principles it fights against. This is a 'passionate problem,' notes Aron. The answer to this apparent contradiction lies in the distance between the ideological 'truth' and the social and economical reality. The very organization of the party and the State are by ideological definition transient, suspended on the destruction of class society and the arrival of a class free society. Democratic principles can be restored only in this latter phase of the dictatorship of the proletariat. Just like Stalin said that the State must be reinforced until it could be abandoned, democratic principles must be suspended before they can be reinstated in a truly democratic society free of class conflicts and exploitation. This is the ideological structure that was used by left wing intellectuals during the Cultural Revolution in the West.

In 1955, Aron publishes one of his most famous books: *The Opium of the Intellectuals* (Aron, 1955). Adopting a critical perspective, Aron denounces the myths on which intellectuals base their argumentation. "Aron's 'skeptical' assault on the myths of the left, the proletariat and the revolution, and his philosophical dissection of the 'idolatry of history,' are at the service of restoring political judgment to its rightful place as the guardian of the human world," writes Dan Mahoney (Aron, 2001). The book is an attack on French left wing intellectuals like Sartre, and their tentative to fusion existentialism with Nietzsche's philosophy in the service of a Marxist interpretation of political action, explains Harvey Mansfield (Aron, 2001, 1969b). Aron interprets the intellectual mobilization in search of an earthly paradise as a new religion. Intellectuals were in need of a new religion, one that would promise salvation and the certitude of future restoration of a just and equalitarian society. Intellectuals were therefore the modern society preachers appealing to the faith in historical necessity that would transform pessimism regarding social reality into optimism of a future liberation and emancipation of the exploited.

Aron's understanding of the conflictual nature of human societies made it impossible for him to believe in the quest of the "best regime." Oppositions such as liberty and efficacy, consensus and political direction point out to the value of civil and political diversity (Mahoney, 1992).

On War and International Relations

Between 1972 and 1980, Aron elaborated his masterpiece: *Penser la guerre: Clausewitz* (Aron, 1976: vol. 1, and 1984: vol. 2). In the same time he wrote a number of articles, conference papers, and occasional texts on the Prussian "strategist and patriot." As Pierre Hassner points out, Aron's central points are the Trinitarian definition of war (the reasoning of the political authority, the free activity of the chief general of war, and people's passions), the theoretical status of the absolute war, and the subordination of the primacy of the war of annihilation to the primacy of politics, that is achieving political goals through war (Aron, 1987). But Aron saw more in Clausewitz than a work to be understood and interpreted. He used Clausewitz to understand the complex situations of world diplomacy and political arena in the twentieth century. As political relations in Europe are an inheritance from the nineteenth century, the themes of equilibrium of powers between rival National States and alliances of States are preeminent in the Cold War era. Two new developments joined this heritage only to render Clausewitz's thought of great actuality. On one hand, in the era of thermonuclear dissuasion, the possibility of Clausewitz's absolute war became from pure concept a real possibility and his theory of limited conflicts a common practice. On the other hand, the fact that Lenin was a reader of Clausewitz's textbook on popular war, and the implication of civilians in guerilla actions in the chessboard of the East–West conflict for global geopolitical positioning, rendered the theoretical framework of the Prussian General indispensible. In the aftermath of

World War II, the weakening of Western colonial powers, the chain reaction of wars of independence all over the globe, and the Soviet military dogma that called for a planetary armed opposition to the expansion and domination of Western capitalism, render the unfinished chapter of Clausewitz's work on civilian armament and popular resistance particularly challenging. Furthermore, as Clausewitz suggested, National State's rationality and interest do not act alone. Sentiments and passions like national pride, popular hostility, and animosities are always at steak in major geopolitical events as army or diplomatic confrontations.

But the most important point in Clausewitz's analysis of the phenomenon of war is his call on individual judgment. Following Clausewitz, Aron accords a special interest in the analysis of personalities of individual actors, politicians, or military chiefs in command. It is them that are called to take action in an uncertain landscape. Aron sees in the paradigm of war the very expression of conflictual situations implicating theoretical knowledge, trained action, and decision making. As Hassner (Aron, 1987) points out, the special relation between Aron and Clausewitz lies in the status they accord to the theoretical analyses of world history and historical events: they are not there to give us a doctrinal view of world's history or a system of scientific propositions that can be applied in every circumstance, functioning as prophecies by making right predictions; they only educate and train the glance and the judgment and thus help us make the right decisions or call the best historical and sociological judgments in the battlefield of world history, not only as chiefs of war or social scientists but also as plain observers and citizens. And this is Aron's greatest legacy.

Dealing with the international diplomacy and the indetermination of strategic relations, Aron interrogates the future of the European idea in the context of the planetary era. He analyses the conditions of possibility for a European cooperation and union rather than the 'dream' of an autonomous and united European power. He argues that, first, Europe is only a branch of a larger defense coalition, and, second, that one cannot create fatherlands (*patries*) on command. One should not count on the hate of Nations and the faith in international entities to promote the European ideal, but instead work on creating more dense cooperation and collaboration channels between the European Nations (Launay, 1995).

See also: Ideology, Sociology of; Liberalism: Historical Aspects; Sociological Theory; Sociology, History of.

Bibliography

Aron, R., 1935. La Sociologie Allemande Contemporaine. Alcan, Paris [1981, new ed. PUF, Paris].

Aron, R., 1938. Introduction to the Philosophy of History. An Essay on the Limits of Historical Objectivity. Gallimard, Paris [1961. Weidenfeld and Nicolson, London].

Aron, R., 1951. The Century of Total War. Gallimard, Paris [1954. Verschoyle, London].

Aron, R., 1955. L'Opium des Intellectuels. Calmann-Lévy, Paris [1991, new ed. Hachette, Paris].

Aron, R., 1958. War and Industrial Society. Oxford University Press, London.

Aron, R., 1962a. Dix-huit Leçons sur la Société Industrielle. Gallimard, Paris [1988, new ed. Gallimard, Paris].

Aron, R., 1962b. Paix et Guerre Entre les Nations. Calmann-Lévy, Paris [1992, new ed. Calmann-Lévy, Paris].

Aron, R., 1964. La Lutte de Classes. Nouvelles Leçons sur les Sociétés Industrielles. Gallimard, Paris [1981, new ed. Gallimard, Paris].

Aron, R., 1965. Démocratie et Totalitarisme. Gallimard, Paris [1992, new ed. Gallimard, Paris].

Aron, R., 1967. Les Étapes de la pensée sociologique. Gallimard, Paris.

Aron, R., 1968a. De Gaulle, Israel and the Jews. Plon, Paris [1969. Deutsch, London/Praeger, New York].

Aron, R., 1968b. La Révolution Introuvable, Réflexions sur la Revolution de Mai. Fayard, Paris.

Aron, R., 1968c. Main Currents of Sociological Thought. Weidenfeld and Nicolson, London.

Aron, R., 1969a. Les Désillusions du Progrès. Essai sur la Dialectique de la Modernité. Calmann-Lévy, Paris [1987, new ed. Julliard, Paris].

Aron, R., 1969b. Marxism and the Existentialists. Gallimard, Paris [1970. Harper and Row, New York].

Aron, R., 1973a. History and the Dialectic of Violence: An Analysis of Sartre's 'Critique de la Raison Dialectique. Gallimard, Paris [1975. Blackwell, Oxford, UK].

Aron, R., 1973b. The Imperial Republic. The United States and the World 1945–1973. Calmann-Lévy, Paris [1973. Prentice Hall, New York].

Aron, R., 1976. Penser la Guerre. Clausewitz. Vol. I, L'Age Européen, Vol. II, L'Age Planétaire. Gallimard, Paris [new edn. 1989 (vol. I) and 1984 (vol. II), Gallimard, Paris].

Aron, R., 1977. In Defense of Decadent Europe. R. Laffont, Paris.

Aron, R., 1978. Politics and History. Free Press, New York.

Aron, R., 1983. Mémoires. Cinquante Ans de Réflexion Politique. Julliard, Paris [1990, new ed. Presses Pocket, Paris].

Aron, R., 1985. History, Truth, Liberty: Selected Writings of Raymond Aron. University of Chicago Press, Chicago.

Aron, R., 1987. Sur Clausewitz. Complexe, Bruxelles.

Aron, R., 2001. The Opium of the Intellectuals. Transaction Publishers, New Brunswick, NJ.

Aron, R., 2002. Le Marxisme de Marx. Fallois, Paris.

Audier, S., 2004. Raymond Aron. La démocratie conflictuelle. Michalon, Paris.

Baverez, N., 1993. Raymond Aron, un moraliste au temps des ideologies. Flammarion, Paris.

Colquhoun, R., 1986. Raymond Aron: The Philosopher in History 1905–1955, The Sociologist in Society 1955–1983, 2 vols. Sage, London.

De Ligio, G., 2007. La tristezza del pensatore politico. Raymond Aron e il primato del politico. Bologna University Press, Bologna, Italy.

Launay, S., 1995. La Pensée politique de Raymond Aron. Presses Universitaires de France, Paris.

Mahoney, D.J., 1992. The Liberal Political Science of Raymond Aron. Rowman and Littlefield, Lanham, UK.

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 2

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser—Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 2

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 2 *xlv*

VOLUME 2

Art and Culture, Economics of 1
Victor A Ginsburgh

Art and Socialisation 7
Ineke Nagel and Harry BG Ganzeboom

Art, Anthropological Aspects of 15
Nelson Graburn

Art History 21
Horst Bredekamp

Artificial and Natural Computation 27
Achim G Hoffmann

Artificial Intelligence: Combinatorial Search 32
Richard E Korf

Artificial Intelligence: Connectionist and Symbolic Approaches 35
Ron Sun

Artificial Intelligence: Genetic Programming 41
Wolfgang Banzhaf

Artisans and Guilds, History of 46
Josef Ehmer

Artists, Competition and Markets 52
Pierre-Michel Menger

Asceticism 60
Christopher Adair-Toteff

Asia, Sociocultural Aspects: China 65
Frank N Pieke

Asia, Sociocultural Overviews: Korea 71
Robert Oppenheim

Asia, Sociocultural Overviews: South Asia 76
Andre Beteille

Assimilation of Immigrants 81
Rubén G Rumbaut

Associations, Scholarly 88
Laurent Jeanpierre and Thibaud Boncourt

Associative Learning in the Cerebellum 92
Peter Mariën, Frank Van Overwalle, Marie Vandekerckhove, and Mario Manto

Assortative Mating in the Marriage Market 100
Aloysius Siow

Asylum Seekers 108
Linda Briskman

Asymmetry of Body and Brain 115
Frank R Boutsen

Atomism and Holism: Philosophical Aspects 131
Michael Esfeld

Attachment During Infancy and Early Childhood: Cultural Variations, Development of 136
Hiltrud Otto

Attempted Utopias and Intentional Communities 144
Adrienne Redd and Tsvi Bisk

Attention-Deficit and Hyperactivity Disorders, Neural Basis of 155
Irene M Loe and Heidi M Feldman

Attention, Neural Basis of 163
Manuela Sellitto and Giuseppe di Pellegrino

Attitude Formation and Change 169
William D Crano and Antonis Gardikiotis

Attitude: History of Concept 175
Christian Fleck

Attitude Measurement 178
Norbert Schwarz

Attitudes and Behavior 183
Joshua J Guyer and Leandre R Fabrigar

Attitudes, Political and Civic Culture 190
GA Almond

Attitudes, Political and Public Opinion 193
Larry M Bartels

Attitudes, Political: Authoritarianism and Tolerance 197
Stanley Feldman

Attributional Processes: Psychological 201
BF Malle

Auctions 205
S Müller

Audience Measurement 211
James G Webster

Audiences, Media 216
Andrea L Press, Fan Mai, Francesca Tripodi, and Michael L Wayne

Auditory Scene Analysis: Computational Models 223
Martin Cooke

Australia and New Guinea, Archaeology of 229
Jim Allen and Ian Lilley

Australian Aboriginal Society and Culture: An Overview 234
Robert Tonkinson

Australia, Sociocultural Overviews: Australian Settler Society 239
Adrian Peace

Authentic Leadership 245
William L Gardner and James D Carlson

Authenticity, Anthropology of 251
Richard Handler

Authoritarian Personality 255
John Duckitt

Authoritarianism 262
Markus Kemmelmeier

Authority, Social Theories of 269
Howard G Schneiderman

Autism Spectrum Disorders 275
Giacomo Vivanti and Erica Salomone

Autobiographical Memory 282
Daniel L Greenberg, Lauren L Deasy, and Amelia L Zasadski

Autobiographical Memory during Early Childhood and Cultural Variations 289
Qi Wang and Robert C Blenis

Automation: Organizational Studies 296
Federico Butera

Automobility 303
Tim Schwanen

Autonomous Agents 309
Tom Bösser

Autonomy, Philosophy of 314
Natalie Stoljar

Avant-Garde Art and Artists 320
Diana Crane

Avoidance and Approach Motivation: A Brief History 326
Rachel M Korn and Andrew J Elliot

Awards, Public Recognition, and Evaluation 332
Rafael Yann Schögler

Axiomatic Theories 338
Patrick Suppes

Balance of Power, History of 345
G van Benthem van den Bergh

Baltes, Paul B (1939–2006) 349
Ulman Lindenberger, Julia AM Delius, and Ursula M Staudinger

Bankruptcy 353
Rafael I Pardo

Barthes, Roland (1915–80) 357
Dustin B Garlitz

Bayesian Graphical Models and Networks 363
David Heckerman

Bayesian Models in Neuroscience 368
Will Penny

Bayesian Statistics 373
Hal S Stern

Bayesian Theory, History, Applications, and Contemporary Directions 378
Scott M Lynch

Beauty and the Labor Market 383
Eva M Sierminska and Xing (Michelle) Liu

Beauvoir, Simone de (1908–86) 392
Elizabeth Fallaize

Beck, Ulrich (1944–2015) 396
Adam J Pritchard

Becker, Howard (1928–) 401
Suzie Guth

Behavior Analysis, Applied 407
Ruben Ardila

Behavior Therapy: Background, Basic Principles, and Early History 412
Kyle E Ferguson and William O'Donohue

Behavior Therapy: Evidence-Based Practice, Clinical Applications, Unresolved Issues 418
Kyle E Ferguson and William O'Donohue

Behavior Therapy: Psychiatric Aspects 426
Lynne M Drummond and Isaac Marks

Behavior Therapy: The Second and Third Waves 431
Kyle E Ferguson and William O'Donohue

Behavioral Economics 437
Sendhil Mullainathan and Richard H Thaler

Behavioral Economics, History of 443
Michiru Nagatsu

Behavioral Genetics: Psychological Perspectives 450
Valerie S Knopik

Behavioral Inhibition and Social Withdrawal across Cultures 456
Jonathan D Lane and Ju-Hyun Song

Behavioral Medicine 463
Claus Vögele

Behavioral Problems, Effects of Parenting and Family Structure on 470
Helena R Slobodskaya

Behavioral Problems, Effects of Socio-Economic Status on 477
Patrick J McGrath and Frank J Elgar

Behavioral Theories of Organization 481
Henrich R Greve and Linda Argote

Belief in a Just World 487
Claudia Dalbert and Matthias Donat

Belief, Anthropology of 493
Byron Good and Andrew McDowell

Bell, Daniel (1919–2011) 498
Howard Brick

Bellah, Robert N (1927–2013) 506
Matteo Bortolini

Bendix, Reinhard (1916–91) 512
John Bendix

Benedict, Ruth (1887–1948) 515
Judith [Modell] Schachter

Benhabib, Seyla (1950–) 519
Leandro Aramburu

Bentham, Jeremy (1748–1832) 525
William Twining

Bereavement 531
Margaret Stroebe, Henk Schut, and Kathrin Boerner

Bereavement and Complicated Grief across the Lifespan 537
Hans Jörg Znoj

Berlin, Isaiah (1909–97) 542
Hans-Peter Müller

Best Worst Scaling: Theory and Practice 548
AAJ Marley and Terry N Flynn

Bielefeld School of History 553
Lutz Raphael

Big Five Factor Model, Theory and Structure 559
Boele de Raad and Boris Mlačić

Big Man, Anthropology of 567
Rena Lederman

Bilingual Aphasia 574
Onésimo Juncos-Rabadán

Bilingual Education: International Perspectives 580
Rosemarie Tracy

Binary Response Models and Logistic Regression 587
Garrett M Fitzmaurice and Nan M Laird

Binet, Alfred (1857–1911) 596
Robert J Sternberg and Linda Jarvin

Bioarchaeology 600
Clark S Larsen

Biobanking: Ethical Issues 604
Klaus Hoeyer

Biocitizenship 610
Carlos Novas

Bioethics 613
Felix Thiele

Bioethics: Genetics and Genomics 619
Ruth Chadwick

Bioethics in the Post-genomic Era 625
Alena Buyx and Barbara Prainsack

Bioethics: Philosophical Aspects 631
Richard E Ashcroft

Biography and Society 637
Roswitha Breckner

Biography: Historical 644
Birgitte Possing

Biological Preconditions for Language Development 650
Jutta L Mueller, Claudia Männel, and Angela D Friederici

Biomarkers in Demographic Research 656
Thomas McDade

Biomedical Sciences, History and Sociology of 663
Robin W Scheffler and Bruno J Strasser

Biosocial Approaches: Crime 670
Michelle A Coyne and John P Wright

Biotechnology 676
Stephen Hilgartner

Bipolar (Manic-Depressive) Disorder 683
Leonardo Tondo and Ross J Baldessarini

Birdsong and Vocal Learning During Development 691
Sarah W Bottjer

Birth, Anthropology of 697
Lucile Newman

Bleuler, Eugen (1857–1939) 700
Paul Hoff

Bloch, Mark Léopold Benjamin (1886–1944) 703
Peter Schöttler

Bloomfield, Leonard (1887–1949) 705
Manfred Bierwisch

Boas, Franz (1858–1942) 710
Georg Elwert

Bobbio, Norberto (1909–2004) 714
Richard Bellamy

Body: Anthropological Aspects 719
Thomas J Csordas

Body, Evolution of 723
Christopher B Ruff

Body, History of 728
Jakob Tanner

Body Image 735
Gemma L Tatangelo, Marita P McCabe, and Lina A Ricciardelli

Bologna Process 741
Annamaria Campanini

Boltanski, Luc (1940–) 747
Simon Susen

Borderline (Patient) Personality 755
Otto F Kernberg

Borders, Anthropology of 760
Hastings Donnan

Boserup, Ester (1910–99) 765
Gale Summerfield

Boundaries and New Organization Forms 768
Walter W Powell and Birthe Soppe

Bounded Rationality 773
John Bendor

Bourdieu, Pierre (1930–2002) 777
Gisèle Sapiro

Bourgeoisie and Middle Classes, History of 784
Hannes Siegrist

Bowlby, John (1907–90) 790
Joan Stevenson-Hinde and Robert A Hinde

Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences 795
William J Hoyer

Brain and Behavioral Lateralization in Animals 799
Lesley J Rogers

Brain Asymmetry 806
Sebastian Ocklenburg

Brain Damage: Neuropsychological Rehabilitation 813
Bryan D Fantie

Brain Development: Ontogenetic Neurobiology of 821
Robbin Gibb

Brain, Evolution of 827
Ralph L Holloway

Brain: Response to Enrichment 835
Janice M Juraska and Leslie M Wise

Brain Stimulation Reward 841
Peter Shizgal

Breast-feeding: Effects on Cognitive and Neural Development 847
Fahmida Tofail and Jena Hamadani

British Cultural Studies 852
Graeme Turner

Broadbent, Donald Eric (1926–93) 856
Patrick Rabbitt

Broadcasting: Regulation 859
Eli Skogerbø

Broaden-and-Build Theory of Positive Emotions 864
Suzannah F Isgett and Barbara L Fredrickson

Brokers and Brokerage, Anthropology of 870
Johan Lindquist

Bruner, Jerome Seymour (1915–) 875
Keiichi Takaya

Bruner's Theory of Cognitive Development 880
Keiichi Takaya

Buddhism 886
David N Gellner and Richard Gombrich

Budgeting and Anticipatory Management 894
Michel J Lebas

Budgeting: Political Science Aspects 901
James D Savage

Burckhardt, Jacob (1818–97) 906
Wolfgang Hardtwig

Bureaucracy and Bureaucratization 908
Marshall Meyer

Bureaucracy, Sociology of 913
Amanda K Damarin

Bureaucratization and Bureaucracy, History of 919
Rolf Torstendahl

Burnout, Psychology of 929
Christina Maslach

This page intentionally left blank

CONTRIBUTORS TO VOLUME 2

Christopher Adair-Toteff
University of South Florida, Tampa, FL, USA; and University of Kent, Canterbury, UK

Jim Allen
Archaeology, La Trobe University, Melbourne, VIC, Australia

G.A. Almond
Stanford University, CA, USA

Leandro Aramburu
Instituto de Investigaciones Gino Germani, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina

Ruben Ardila
National University of Colombia, Bogotá, Colombia

Linda Argote
Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA

Richard E. Ashcroft
School of Law, Queen Mary University of London, London, UK

Ross J. Baldessarini
Harvard Medical School, Boston, MA, USA; and Mailman Research Center, McLean Hospital, Belmont, MA, USA

Wolfgang Banzhaf
Memorial University of Newfoundland, St. John's, NL, Canada

Larry M. Bartels
Princeton University, Princeton, NJ, USA

Richard Bellamy
School of Public Policy, University College London, London, UK; and European University Institute, Florence, Italy

John Bendix
Universität Zürich, Zürich, Switzerland

John Bendor
Stanford University and Center for Advanced Study in the Behavioral Sciences, Stanford, CA, USA

Andre Beteille
Delhi School of Economics, University of Delhi, Delhi, India

Manfred Bierwisch
Berlin, Germany

Tsvi Bisk
Center for Strategic Futurist Thinking, Ziv Group, Ltd., Kfar Saba, Israel

Robert C. Blenis
Cornell University, Ithaca, NY, USA

Kathrin Boerner
Jewish Home Lifecare, New York, NY, USA

Thibaud Boncourt
Centre Européen de Sociologie et de Science Politique (CESSP), Paris, France

Matteo Bortolini
Università di Padova, Padova, Italy

Tom Bösser
University of Münster, Münster, Germany

Sarah W. Bottjer
University of Southern California, Los Angeles, CA, USA

Frank R. Boutsen
College of Allied Health, Health Sciences Center, University of Oklahoma, Oklahoma City, OK, USA

Roswitha Breckner
University of Vienna, Institute of Sociology, Vienna, Austria

Horst Bredekamp
Humboldt University, Berlin, Germany

Howard Brick
University of Michigan, Ann Arbor, MI, USA

Linda Briskman
Swinburne University of Technology, Melbourne, VIC, Australia

Federico Butera
University of Milan-Bicocca, Milano, Italy

Alena Buyx
University College London, London, UK; and University of Münster, Münster, Germany

Annamaria Campanini
Università di Milano Bicocca, Milano, Italy

James D. Carlson
Texas Tech University, Lubbock, TX, USA

Ruth Chadwick
Centre for Social Ethics and Policy, The University of Manchester, Manchester, UK

Martin Cooke
University of the Basque Country, Vitoria, Spain

Michelle A. Coyne
University of Cincinnati, Cincinnati, OH, USA

Diana Crane
University of Pennsylvania, Philadelphia, PA, USA

William D. Crano
Claremont Graduate University, Claremont, CA, USA

Thomas J. Csordas
University of California, San Diego, CA, USA

Claudia Dalbert
Martin Luther University of Halle-Wittenberg, Halle/Saale, Germany

Amanda K. Damarin
Georgia Perimeter College, Dunwoody, GA, USA

Lauren L. Deasy
College of Charleston, Charleston, SC, USA

Julia A.M. Delius
Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany

Boele de Raad
University of Groningen, Groningen, The Netherlands

Giuseppe di Pellegrino
Centre for Studies and Research in Cognitive Neuroscience, University of Bologna, Cesena, Forlì-Cesena, Italy

Matthias Donat
Martin Luther University of Halle-Wittenberg, Halle/Saale, Germany

Hastings Donnan
Queen's University, Belfast, UK

Lynne M. Drummond
South West London and St George's Mental Health NHS Trust, London, UK; and St George's, University of London, London, UK

John Duckitt
The University of Auckland, Auckland, New Zealand

Josef Ehmer
University of Salzburg, Salzburg, Austria

Frank J. Elgar
Institute for Health and Social Policy, McGill University, Montreal, QC, Canada

Andrew J. Elliot
University of Rochester, Rochester, NY, USA

Georg Elwert[†]
Free University of Berlin, Berlin, Germany

Michael Esfeld
University of Lausanne, Lausanne, Switzerland

Leandre R. Fabrigar
Queen's University, Kingston, ON, Canada

Elizabeth Fallaize[†]
Oxford University, Oxford, UK

Bryan D. Fantie
American University, Washington, DC, USA

Heidi M. Feldman
Stanford University School of Medicine, Stanford, CA, USA

Stanley Feldman
State University of New York at Stony Brook, Stony Brook, NY, USA

Kyle E. Ferguson
Providence St. Peter Family Medicine Residency Program, Renton, WA, USA; and University of Washington, Seattle, WA, USA

Garrett M. Fitzmaurice
Harvard School of Public Health, Harvard University, Boston, MA, USA

Christian Fleck
University of Graz, Graz, Austria

Terry N. Flynn
University of Technology Sydney, Sydney, NSW, Australia

[†] Deceased.

Barbara L. Fredrickson
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Angela D. Friederici
Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Harry B.G. Ganzeboom
Free University Amsterdam, Amsterdam, The Netherlands

Antonis Gardikiotis
Aristotle University of Thessaloniki, Thessaloniki, Greece

William L. Gardner
Texas Tech University, Lubbock, TX, USA

Dustin B. Garlitz
University of South Florida, Tampa, FL, USA

David N. Gellner
University of Oxford, Oxford, UK

Robbin Gibb
University of Lethbridge, Lethbridge, AB, Canada

Victor A. Ginsburgh
ECARES, Université Libre de Bruxelles, Brussels, Belgium; and CORE, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

Richard Gombrich
University of Oxford, Oxford, UK

Byron Good
Faculty of Arts and Sciences, Harvard Medical School, Boston, MA, USA

Nelson Graburn
University of California, Berkeley, CA, USA

Daniel L. Greenberg
College of Charleston, Charleston, SC, USA

Henrich R. Greve
INSEAD, Singapore

Suzie Guth
Strasbourg University, Strasbourg, France

Joshua J. Guyer
Queen's University, Kingston, ON, Canada

Jena Hamadani
International Centre for Diarrhoeal Disease Research, Bangladesh, Dhaka, Bangladesh

Richard Handler
University of Virginia, Charlottesville, VA, USA

Wolfgang Hardtwig
Humboldt University, Berlin, Germany

David Heckerman
Microsoft Corporation, Redmond, WA, USA

Stephen Hilgartner
Cornell University, Ithaca, NY, USA

Robert A. Hinde
St. John's College, Cambridge, UK

Klaus Hoeyer
University of Copenhagen, Copenhagen, Denmark

Paul Hoff
Psychiatric University Hospital, Zurich, Switzerland

Achim G. Hoffmann
The University of New South Wales, Sydney, NSW, Australia

Ralph L. Holloway
Columbia University, New York, NY, USA

William J. Hoyer
Syracuse University, Syracuse, NY, USA

Suzannah F. Isgett
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Linda Jarvin
Paris College of Art, Paris, France

Laurent Jeanpierre
Universite Paris 8, Paris, France

Onésimo Juncos-Rabadán
University of Santiago de Compostela, Santiago de Compostela, Spain

Janice M. Juraska
University of Illinois, Champaign, IL, USA

Markus Kemmelmeier
University of Nevada, Reno, NV, USA

Otto F. Kernberg
Personality Disorders Institute, New York Presbyterian Hospital, New York, NY, USA; Joan and Sanford I. Weill Medical College of Cornell University, New York, NY, USA; and Columbia University Center for Psychoanalytic Training and Research, New York, NY, USA

Valerie S. Knopik
Rhode Island Hospital, Providence, RI, USA; Brown University, Providence, RI, USA; and Bradley Hasbro Children's Research Center, Providence, RI, USA

Richard E. Korf
University of California, Los Angeles, CA, USA

Rachel M. Korn
University of Rochester, Rochester, NY, USA

Nan M. Laird
Harvard School of Public Health, Harvard University, Boston, MA, USA

Jonathan D. Lane
Harvard University, Cambridge, MA, USA

Clark S. Larsen
The Ohio State University, Columbus, OH, USA

Michel J. Lebas
HEC School of Management, Paris, France

Rena Lederman
Princeton University, Princeton, NJ, USA

Ian Lilley
University of Queensland, Brisbane, QLD, Australia

Ulman Lindenberger
Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany

Johan Lindquist
Stockholm University, Stockholm, Sweden

Xing (Michelle) Liu
The University of Arizona, Tucson, AZ, USA

Irene M. Loe
Stanford University School of Medicine, Stanford, CA, USA

Scott M. Lynch
Princeton University, Princeton, NJ, USA

Fan Mai
The University of Virginia, Charlottesville, VA, USA

B.F. Malle
University of Oregon, Eugene, OR, USA

Claudia Männel
Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Mario Manto
Fonds de la Recherche Scientifique (FNRS), ULB Erasme, Bruxelles, Belgium

Peter Mariën
ZNA Middelheim Hospital, Antwerp, Belgium; and Vrije Universiteit Brussel, Brussels, Belgium

Isaac Marks
King's College, University of London, London, UK

A.A.J. Marley
University of Victoria, Victoria, BC, Canada; and University of Technology Sydney, Sydney, NSW, Australia

Christina Maslach
University of California, Berkeley, Berkeley, CA, USA

Marita P. McCabe
School of Psychology, Deakin University, Burwood, VIC, Australia

Thomas McDade
Northwestern University, Evanston, IL, USA

Andrew McDowell
McGill International TB Centre, McGill Anthropology, Montreal Chest Institute, Montreal, QC, Canada

Patrick J. McGrath
IWK Health Centre and Capital District Health Authority, Dalhousie University, Halifax, NS, Canada

Pierre-Michel Menger
Collège de France, Paris, France; and Ecole des Hautes Etudes en Sciences Sociales, CESPRA, Paris, France

Stephen Mennell
School of Sociology, University College Dublin, Dublin, Ireland

Marshall Meyer
University of Pennsylvania, Philadelphia, PA, USA

Boris Mlačić
Institute of Social Sciences Ivo Pilar, Zagreb, Croatia

Jutta L. Mueller
Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany; and University of Osnabrück, Institute for Cognitive Science, Osnabrück, Germany

Sendhil Mullainathan
Massachusetts Institute of Technology, Boston, MA, USA

Hans-Peter Müller
Humboldt-University of Berlin, Berlin, Germany

S. Müller
Humboldt-Universität zu Berlin, Berlin, Germany

Michiru Nagatsu
University of Helsinki, Helsinki, Finland

Ineke Nagel
Free University Amsterdam, Amsterdam, The Netherlands

Lucile Newman
Brown University, Providence, RI, USA

Carlos Novas
Carleton University, Ottawa, ON, Canada

William O'Donohue
University of Nevada, Reno, NV, USA

Sebastian Ocklenburg
Ruhr-Universität Bochum, Bochum, Germany

Robert Oppenheim
University of Texas at Austin, Austin, TX, USA

Hiltrud Otto
The Hebrew University of Jerusalem, Mount Scopus, Jerusalem, Israel

Rafael I. Pardo
Emory University School of Law, Atlanta, GA, USA

Adrian Peace
School of Social Science, University of Queensland, Ipswich, QLD, Australia

Will Penny
Wellcome Trust Centre for Neuroimaging, University College, London, UK

Frank N. Pieke
Leiden University, RA Leiden, The Netherlands

Birgitte Possing
The Danish National Archives, Copenhagen, Denmark

Walter W. Powell
Stanford University, Stanford, CA, USA

Barbara Prainsack
King's College London, London, UK

Andrea L. Press
The University of Virginia, Charlottesville, VA, USA

Adam J. Pritchard
University of Central Florida, Orlando, FL, USA

Patrick Rabbitt
University of Oxford Department of Experimental Psychology, University of Oxford, Oxford, UK

Lutz Raphael
Universität Trier, Trier, Germany

Adrienne Redd
Schools for Sustainability, Philadelphia, PA, USA

Lina A. Ricciardelli
School of Psychology, Deakin University, Burwood, VIC, Australia

Lesley J. Rogers
School of Science and Technology, University of New England, Armidale, NSW, Australia

Christopher B. Ruff
Johns Hopkins University, Baltimore, MD, USA

Rubén G. Rumbaut
University of California, Irvine, CA, USA

Erica Salomone
King's College London, Institute of Psychiatry, London, UK

Gisèle Sapiro
École des hautes études en sciences sociales, Paris, France; and CNRS (Centre européen de sociologie et de science politique), Paris, France

James D. Savage
University of Virginia, Charlottesville, VA, USA

Judith [Modell] Schachter
Carnegie Mellon University, Pittsburgh, PA, USA

Robin W. Scheffler
Program in Science, Technology, and Society, Massachusetts Institute of Technology, MA, USA

Howard G. Schneiderman
Lafayette College, Easton, PA, USA

Rafael Yann Schögler
University of Graz, Graz, Austria

Peter Schöttler
Centre National de la Recherche Scientifique, Institut d'histoire du temps présent, Paris, France

Henk Schut
Utrecht University, Utrecht, The Netherlands

Tim Schwanen
University of Oxford, Oxford, UK

Norbert Schwarz
USC Dornsife Mind & Society Center, University of Southern California, Los Angeles, CA, USA

Manuela Sellitto
Centre for Studies and Research in Cognitive Neuroscience, University of Bologna, Cesena, Forlì-Cesena, Italy

Peter Shizgal
Concordia University, Montreal, QC, Canada

Hannes Siegrist
University of Leipzig, Leipzig, Germany

Eva M. Sierminska
CEPS/INSTEAD Research Institute, Esch-sur-Alzette, Luxembourg

Aloysius Siow
University of Toronto, Toronto, ON, Canada

Eli Skogerbø
University of Oslo, Oslo, Norway

Helena R. Slobodskaya
Novosibirsk State University, Novosibirsk, Russia

Ju-Hyun Song
University of Michigan, Ann Arbor, MI, USA

Birthe Soppe
Stanford University, Stanford, CA, USA

Ursula M. Staudinger
Robert N. Butler Columbia Aging Center, Columbia University, New York, NY, USA

Hal S. Stern
University of California Irvine, Irvine, CA, USA

Robert J. Sternberg
Cornell University, Ithaca, NY, USA

Joan Stevenson-Hinde
University of Cambridge, Cambridge, UK

Natalie Stoljar
McGill University, Montreal, QC, Canada

Bruno J. Strasser
University of Geneva, Geneva, Switzerland

Margaret Stroebe
Utrecht University, Utrecht, The Netherlands

Gale Summerfield
University of Illinois, Champaign, IL, USA

Ron Sun
Rensselaer Polytechnic Institute, Troy, NY, USA

Patrick Suppes
Stanford University, Stanford, CA, USA

Simon Susen
City University London, London, UK

Keiichi Takaya
Faculty of Letters, Kokugakuin University, Tokyo, Japan

Keiichi Takaya
Kokugakuin University, Shibuya-ku, Tokyo, Japan

Jakob Tanner
Forschungsstelle für Sozial- und Wirtschaftsgeschichte, Universität Zürich, Zurich, Switzerland

Gemma L. Tatangelo
School of Psychology, Deakin University, Burwood, VIC, Australia

Richard H. Thaler
University of Chicago, Chicago, IL, USA

Felix Thiele
Europaeische Akademie, Bad Neuenahr-Ahrweiler, Germany; and Institute for Philosophy, University of Duisburg-Essen, Essen, Germany

Fahmida Tofail
International Centre for Diarrhoeal Disease Research, Bangladesh, Dhaka, Bangladesh

Leonardo Tondo
Harvard Medical School, Boston, MA, USA; McLean Hospital, Belmont, MA, USA; and Lucio Bini Mood Disorders Centers, Cagliari and Rome, Italy

Robert Tonkinson
The University of Western Australia, Crawley, WA, Australia

Rolf Torstendahl
Uppsala University, Uppsala, Sweden

Rosemarie Tracy
University of Mannheim, Mannheim, Germany

Francesca Tripodi
The University of Virginia, Charlottesville, VA, USA

Graeme Turner
University of Queensland, Brisbane, QLD, Australia

William Twining
University College London, London, UK

G. van Benthem van den Bergh
The Hague, The Netherlands

Marie Vandekerckhove
Vrije Universiteit Brussel, Brussels, Belgium

Frank Van Overwalle
Vrije Universiteit Brussel, Brussels, Belgium

Giacomo Vivanti
School of Psychological Science, Olga Tennison Autism Research Centre, La Trobe University, Melbourne, VIC, Australia

Claus Vögele
Institute for Health and Behaviour, Research Unit INSIDE, University of Luxembourg, Walferdange, Luxembourg

Qi Wang
Cornell University, Ithaca, NY, USA

Michael L. Wayne
The University of Virginia, Charlottesville, VA, USA

James G. Webster
Northwestern University, Evanston, IL, USA

Leslie M. Wise
University of Illinois, Champaign, IL, USA

John P. Wright
University of Cincinnati, Cincinnati, OH, USA

Amelia L. Zasadski
College of Charleston, Charleston, SC, USA

Hans Jörg Znoj
Institut für Psychologie, Abteilung Klinische Psychologie und Psychotherapie, Universität Bern, Switzerland

This page intentionally left blank

Art and Culture, Economics of

Victor A Ginsburgh, ECARES, Université Libre de Bruxelles, Brussels, Belgium; and CORE, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article provides an overview since the starting point of the field in the mid-1960s, its main developments, strengths, and weaknesses and suggests new research avenues.

Introduction

The two landmarks of art history and economics, Johann Winckelman's *Geschichte der Kunst des Altertums* and Adam Smith's *An Inquiry into the Nature and Causes of the Wealth of Nations*, are almost contemporaneous. This should have paved the way to a good understanding between the two emerging disciplines. Unfortunately, they did not, though most forefathers of the science that came to be called 'economics' made forays into 'art.' Most of them were concerned with art as a 'luxury' or as the object of 'conspicuous consumption'; prices; talent needed to create; 'progress' in the arts; why people buy art; why competition is needed; exorbitant rewards; copyright; public support of the arts (or its regressive nature, since anyway, art is for the riches); moral utility of the arts, and so on, but aesthetics and art history were far from being important in the discussion (Goodwin, 2006).

Research on the economics of art really started with the essay by Baumol and Bowen (1966) on why, unless they receive financial support, the performing arts may disappear. In a nutshell, today as in 1750, a quartet needs four musicians, and while there may be productivity gains in the rest of the economy, and therefore, wage increases, they will also have to be passed through to musicians. Otherwise, nobody will be attracted to become a professional musician, and quartets will disappear. If, on the other hand, these wage increases are passed through to musicians, and since no productivity gains can be made, quartets will become too expensive to run and will also disappear. Therefore, intervention (public or other) is needed. Many arguments have been suggested to counter this theory that came to be called 'the cost disease, but it still has a lot of appeal and generated a large literature.

The field is now reasonably well established, has an association, the Association of Cultural Economics International, publishes a quarterly journal, the *Journal of Cultural Economics*, organizes, every other year, a conference that is attended by 200–300 scholars, and has been able to generate teaching positions in a few economics departments in Europe and in the United States. Though the discipline is now the subject of several handbooks and collections of articles (Ginsburgh and Throsby (2006, 2013), Towse (2003, 2008a). See also Frey and Pommerehne (1989), which contains a very complete and useful list of references written before 1989.), it is still in its infancy and, as will be discussed, is looking for some firm ground.

Indeed, the subject is hard to define, since it is located at the crossroads of several disciplines: art history, art philosophy, sociology, law, management, and economics. It also started to make forays into topics that used to be discussed in other fields, such as anthropology, development economics, cultural conflicts, psychology, religion and customs, and languages. These extensions will not be discussed here. (The interested reader is encouraged to visit Hutter and Throsby (2008) and Ginsburgh and Throsby (2013).)

The issue of whether economists interested in cultural economics should simply apply their usual tools to questions related to and data coming from the arts, or whether they should take culture as an opportunity to add new issues to the existing economic literature is still open. Should economists take as granted that prices and income are the main determinants of the demand for theater plays or should they also understand what quality means in such heterogeneous markets, define it, measure, and enter it as a variable in their empirical models? Blaug (2001) concludes in a pessimistic way: "We have all been hoping that the field might actually suggest and promote developments that would spill over with benefit to economics outside its own domain. Has this happened? Not really." Not only did nothing spill over from art or cultural economics to economics in general. But there is worse, since many tools are or have been 'invented' and imported from economists who certainly have some feeling for art and culture, but whose main field of research was or is different. We shall give some examples later, but here are some articles that are worth singling out: Stigler and Becker (1977) who emphasized rationality in cultural consumption; Montias (1982, 1996, 2004) and his choice of looking at inventories and other microdata to analyze painters and collectors in the Dutch seventeenth-century Golden Age; Rosen's (1981) superstar model showing that small differences in talent make for large differences in income; Baumol (1986) who suggested that the difference between the returns on art and financial markets could be interpreted as accounting for the aesthetic pleasure provided by looking at artworks; Ashenfelter (1989) who made auctions known outside of the difficult theoretical field in which they were entrenched and made economists who were interested in the arts aware that hammer prices could be used to estimate price indices; and Peacock for his many contributions, but especially for his *Paying the Piper* (1993).

We start our discussions with two topics that are of concern to all types of art: Should the arts benefit from public or private support and from regulation (Section Public Support of the

Arts) and How well labor markets for artists work (Section Labor Markets and Artists' Careers). Next we go to specific problems, distinguishing the visual arts (paintings, sculpture, art objects) in Section The Visual Arts, the performing arts (music, theater, opera, dance) in Section The Performing Arts, and cultural heritage (museums, historical buildings, monuments, and sites) in Section Cultural Heritage, Where Do We Go, though there are some unavoidable intersections (museums accumulate paintings, some artists in the visual arts produce unique performances in galleries). There is a tendency, nowadays, to add what came to be called the cultural (or creative) industries, which are interested in books, movies, popular music, records, the media, the internet, and copyright, and are very briefly discussed in Section The Cultural and Creative Industries. In each case, we try to point out what cultural economists have found interesting, discuss the tools that they have imported from economics to analyze art and culture, and some issues that are still open. Section Concluding Remarks concludes.

Public Support of the Arts

Many arguments have been invoked to justify public support and correct for assumed market failures. Some arguments hold for all forms of artistic activity (including the cultural industries), some are more specific. Here is a nonexhaustive list of the most important economic rationalizations for such support.

1. The oldest and most often invoked argument is that art, whatever its form, is a public good. It benefits not only those who see it and who pay for it but also all other consumers, who do not necessarily wish to contribute voluntarily to its production (performing or visual arts) or to its preservation (museums) and free ride or who cannot contribute, since they are not yet born (heritage). If the arts are left to the market, they will not be priced correctly and will thus be underproduced or not saved for future generations. Artistic activities are also said to produce externalities that cannot be sold on the marketplace, such as civilizing effects, national pride, prestige, and identity (justifying, for example, the French position in its fight against free trade of movies and TV programs), social cohesion, and so on, which benefit all consumers. Similar arguments had already been suggested by Stanley Jevons, Adam Smith, and Arthur Pigou.
2. But the arts are also said to yield economic externalities. Old castles, well-known opera houses or orchestras, and art festivals attract visitors and tourists. So do museums with good collections, while newly constructed museums are claimed to contribute to city renewal (an argument used, for instance, to attract public support for the Guggenheim museum in Bilbao). This is supposed to have spillover effects on hotels, nearby restaurants, and shops, generate new activities and came to be called the 'arts multiplier' (see Grampp, 1989: p. 247).
3. Art is a 'merit' good. It "is a means of educating the public's taste and the public would benefit from a more educated taste" (Scitovsky's, 1972). Since consumers are not fully informed, they are unable to evaluate all its benefits without public intervention. Moreover, even if there is addiction to the arts, it develops only slowly, and consumers have to be exposed as much as possible.
4. For equity reasons, art should be made available also to low-income consumers who cannot afford to pay. Poor artists should be supported alike. Schemes for doing this will be discussed below.
5. Culture is transmitted by education but also from parents to children. Since parents can hardly be considered as purely altruistic, an additional externality is generated, which needs support for efficiency (and equity) reasons.
6. The last argument on the difficulty or impossibility to achieve productivity gains is more specific to the performing arts: the Baumol cost disease, which we discussed earlier as being the starting point of cultural economics.

These issues have of course their supporters and their detractors, the most enjoyable to read being Grampp (1989). Detractors base their case on the public choice theoretic claim that some people wish to use the state's resources for their own benefit and on the regressive nature of supporting high-income groups who are more likely than others to attend and who can afford to do so. In 1997, in the United States, for example, 35% of the high-income population (more than $75 000 per year) attend classical music live performances, while this percentage drops to 6.8% for the lowest income group. See Heilbrun and Gray (2001: p. 47). See also Benhamou (2000) for Europe.

The cost disease has led to a host of quantitative studies, which show that the curse is not as severe as what Baumol and Bowen had originally suggested. Empirical studies on the performing arts point to price elasticities which are smaller than one so that there is still some room for price increases; bundling – bookstores at opera houses, museums, drinks during intermissions – can also be used more systematically; technical progress (sound and lighting, for instance) can make for larger audiences; broadcasts and records could be used to cross-subsidize live performances, and make these less expensive; opera music is performed under the form of concerts, and both are broadcasted in real time; contemporary theater plays save on the number of characters, and so on. See Towse (1997b) for a collection of papers on the disease.

In most developed countries, the arts are supported, though the volume of public intervention can vary substantially. It is very low in the United States ($3–4 per head), reaches $16 in the United Kingdom, $19 in Australia, and $30–40 in Continental Europe where private support is negligible. (See Throsby, 1994: p. 21 as well as Netzer, 1978 for more detailed data.) No convincing argument has yet been put forward to explain these large differences in behavior between the Anglo-Saxon and other cultures.

Government intervention takes several forms. It can be direct: Some art institutions or companies are state-owned (museums in most European countries, but also opera companies, theaters, symphony orchestras) and are, therefore, supported, at least in part, by public funds. Their efficiency has been questioned on many occasions. Arts education programs are also often supported by the state. In the 1960s, Peacock (1993: p. 122) suggested to distribute to poor consumers art vouchers exchangeable for seats or visits to museums.

In most cases, public funds, including lottery funds, are channeled to companies, or directly to artists, through subsidies or tax relief (reduced VAT rates, no or lower rates of taxation on profits or on income). In the United States, most of the effort is concentrated on tax deductions or exemptions to donors (individuals or corporations) for charitable contributions; gifts or bequests of artworks are also encouraged by tax concessions, and more and more often, inheritance taxes can be paid for by artworks. This indirect subsidization system is much less used in European countries, though similar incentive schemes are present. See also O'Hagan (1998) for more information.

The less direct (and, sometimes, most insidious) form of intervention is regulation, which is mainly concerned with (1) ensuring moral rights, that is, protecting artists against being plagiarized, or against alterations or destruction of their works; (2) ensuring financial rewards to living artists and to their heirs during as much as 70 years after their death under the form of copyright. Copyright has led to a very large literature, at the confines of law and economics. See Landes and Posner (2003) and Towse (2008b) for a collection of recent articles. Its economic justification is based on the idea that it provides incentives to create and that the absence of protection would lead to underproduction. (Waldfogel (2011a,b) shows that this is wrong in the music industry.) The same idea is used to justify protection of industrial patents.

Copyright applies mainly to plays, literary, and musical works. Authors usually do not 'sell' their creation, but get compensated through authors' rights for each copy (book or record) sold or for each performance of a play or of a score. However, the idea has been extended to all other forms of art, including the performance itself (the staging of a play can be copyrighted and so do live performances by visual artists), and the visual arts. Reproduction of a work pays rights to the artist even after he sold it. One can wonder whether copyright laws still provide incentives for art creation, or whether they generate incentives to use Mozart's music in movies and Schiele or Klimt paintings (no longer covered by copyright) to illustrate book covers. Obviously, regulation has led to overregulation, much more driven by the rent-seeking behavior of the music industry and author societies than by artists.

The international trade of cultural goods has also attracted the attention of regulators. Some countries fully prohibit the export of artifacts of some age – although illegal exports from Africa, Mexico, Italy, or Greece keep going. Some countries (France and the United Kingdom) are more lenient and require export licenses for 'national treasures'; this procedure is often used to delay the export and give time to the potential exporting country to raise funds and buy the object at the price at which it could be sold abroad. France insists on maintaining import quotas on American TV programs to 'protect' local culture and national identity. This is the so-called cultural exception argument, which is very reminiscent of its forefather, the infant industry argument.

Labor Markets and Artists' Careers

A folk theorem states that most artists are poor – or have lower income than the average consumer, die poor, while salerooms, their dealers, and editors become rich.

Using the 1980 US Census data, Filer (1986) concludes that, on average, artists do not earn any less than they would in a nonartistic employment. It has later been shown, however, that this is due to multiple job holdings, including in nonartistic activities (Wassal and Alper, 1992). If these extra jobs are not accounted for, artists' average incomes are smaller than those of workers with comparable education. It is also said that artists devote less time to their work on art than do others to their main occupation, that they may have to go to lower paid jobs to supplement their income, and that their earnings are more variable and less directly influenced by education than those of other workers. To explain why artists accept this income differential, one has to appeal to differences in motivation; this is summarized by Throsby (1994: p. 17) for whom "the primary desire to create art as a principal occupation must be recognized as the essential driving force behind an artist's labor supply decisions."

But the income distribution of artists is also markedly skewed and there exist superstars, whose earnings are impressive. Rosen (1981) shows that talent only does not explain these differences. He models income as a convex function of talent so that small differences in this endowment may result in large differences in income. This is compounded by the fact that there is little if any substitution between various degrees of talent or between quality and quantity. As Rosen points out, "hearing a succession of mediocre singers does not add up to a single outstanding performance."

In a series of articles and a book, Galenson (2001) studies the careers and life cycles of impressionist and modern French and American painters (as well as writers) and consistently finds that some artists bloom early, while others produce their best works later during their life cycle. The two types of careers are related to innovative behavior. *Conceptual* artists, who bloom early "think of [their work] as primarily making a preconceived image, and [the execution is] often simply a process of transferring an image they have already created from one surface to another." *Experimental* artists, on the other hand, bloom late: "Their goals are imprecise, so their procedure is tentative and incremental" (Galenson, 2003). The idea that some artists bloom early, while others are late, and that artists may follow two paths in their creative work seems very reasonable. What is less so is the association between age and type of creativity and the dichotomous behavior of being either conceptual or experimental; most artists are both at different times in their careers and even at the same moment in different works. See Ginsburgh and Weyers (2006) and Simonton (1997).

The Visual Arts

Art works are, with some exceptions, heterogeneous. Each creation is unique and markets can hardly be analyzed by the usual supply–demand mechanism. The very high prices reached by some works has led economists to study (1) what can be said about the characteristics that determine values, (2) whether the returns obtained on artworks outperform those of financial markets, and (3) why tastes change. None of these issues has been given a definitive answer.

Values are obviously mainly determined by the name (or the reputation) of the artist, but this explains little, and one should understand why reputations differ. Decomposing 'quality' in terms of (hopefully) objective characteristics or properties should be possible. After all, this is what the French art historian de Piles (1708) had in mind and led him to grade on a scale between 0 and 20, characteristics such as composition, drawing, color, and expression, for a group of painters, and rank them accordingly. Analytic art philosophers have developed the idea of properties, which is quite close to economists who work on horizontal and vertical differentiation, so that there is some common ground.

Financial returns on works of art have been analyzed very often (see Frey and Eichenberger, 1995 and Ashenfelter and Graddy, 2006 for surveys), but Baumol's (1986) article was probably the most influential one. Baumol finds that between 1650 and 1960, the average real return rate is equal to 0.55% per year, some 2% lower than the return on bonds. He suggests that the difference is attributable to the return provided by aesthetic pleasure.

Tastes change, and prices commanded by works of art often follow these movements. There are many examples of painters who have been 'rediscovered,' including Botticelli, Vermeer, Goya, or the Pre-Raphaelites. See Haskell (1980) for an account of these changes in France and in the United Kingdom between the late eighteenth century and the early 1920s. It is unclear whether such changes are due to fads or fashion or to more fundamental reasons. Grampp (1989: pp. 66–68) suggests that the rapid succession of styles in contemporary art is due to the rapid increase in consumers' income, "who seek novelty more often than they would if their income was lower." Montias, an economist turned to famous art historian, respected for his work on Vermeer and on Dutch Golden Age painting. See among others Montias (1982). Instead of analyzing leading figures and facts, his work centers on ordinary people. In his 1996 analysis on seventeenth-century Dutch art collectors in Amsterdam, he illustrates how quantitative methods can be used in art history. He also uses archives to examine the intricate relations between artists, art dealers, and their patrons and shows that dealers were only modestly successful. They relied on friends, business as well as religious acquaintances, and their family for their working capital, of which they seemed to be in constant need. Their role in dealing works by the major artists of their time seems also to be quite limited (Montias, 2004). Much work in similar directions is described in the articles jointly written by economist de Marchi and art historian Van Miegroet. Their articles on copies (1996), on early auction markets (2004), and more generally on economic art history (2006) should be singled out. So should the book by Nelson and Zeckhauser (2008), who use information economics to understand art markets from the Italian Renaissance.

The status and the prices of copies (to be distinguished from fakes) have also varied over time. While copies were close substitutes for originals during the seventeenth century, and priced at about half the price of originals (De Marchi and Van Miegroet, 1996), they are hardly considered to have any value nowadays. Again, little is known about the reasons and the period at which the break occurred. (See Benhamou and Ginsburgh, 2006.)

Many issues discussed above have often been analyzed on the basis of works sold and prices obtained at auctions. This has led to several observations on the possible inefficiency of auction mechanisms, which theory considers to be efficient. Among these observations, let us single out the declining price anomaly (prices of homogeneous lots sold in sequence decrease over time), but this is also the case for the ratio between hammer prices and presale estimates for heterogeneous works of art (Ashenfelter, 1989; Beggs and Graddy, 1997), the existence of systematic price differences across salerooms and/or countries (Pesando and Shum, 1996), the biasedness of presale estimates by salerooms and the fact that presale estimates do not reflect all the information that is available to them, though theory shows that it is efficient for the auctioneer to reveal the truth (Bauwens and Ginsburgh, 2000).

The Performing Arts

The performing arts were the first to be studied after Baumol and Bowen (1966) introduced their cost disease alluded to earlier. Blaug (1976) contains a collection of papers written in the late 1960s and early 1970s that are mainly devoted to the performing arts. See also Throsby and Withers (1979). It is also the sector that can be thought to conform best to the classic supply–demand framework. Demand is modeled as depending on own price, prices of substitutes, income, quality, and the 'price of leisure,' since the performing arts are time consuming.

However, both quality and the price of leisure are usually difficult to assess (Throsby, 1983). Supply is multidimensional and can hardly be represented by the number of seats or the number of performances offered. Indeed, what is offered is also variety and width of repertoire, categories of seats sold at different prices, as well as intertemporal price discrimination and peak load pricing during certain days, which all affect demand. (See Courty and Pagliero, 2013 for a highly interesting survey and application of price discrimination by rock singers.)

The performing arts are prone to the superstar phenomenon, which is difficult to study empirically, since talent can hardly be measured and has to be circumvented by other measures.

Cultural Heritage, Where Do We Go

Forty-five thousand monuments are listed in France, and this number is increasing every year, without any legal possibility to drop a monument once it is listed. In the United Kingdom, there are 40 000 scheduled monuments and sites and 500 000 listed buildings. Maintaining this heritage obviously implies costs, and these costs are increasing. The number of specialists who are able to reconstruct and restore these sites is dwindling, the materials needed are becoming rare and often expensive to extract, and monuments deteriorate at an increasing rate, since they are subject to pollution and visited more often (Benhamou, 2000: p. 58). The decision to list heritage sites less often will have to be made, and perhaps also the decision to drop monuments and allow their demolition. Property rights

may have to be redefined to induce private owners to care for their properties, and tools, such as contingent valuation, polls, or even referenda may be needed to elicit preferences and help decision makers.

The number of museums has also increased dramatically during the past 20 years, and so have collections, generating serious constraints on exhibition space: 80–90% of the collections are not shown (97% in the case of the Art Institute in Chicago). Though the International Council of Museums (ICOM) recommends that each museum should have a statement describing its collecting policy, many have no such written documentation. Why can't existing museums stop increasing their collections, and why is it that so many new museums are being born? Why is it that if curators are empowered with the possibility to buy works of art, it is made difficult (in the United States) if not impossible or unlawful (in Europe) for them to sell works, including the thousands of broken Greek vases that almost each museum holds? This would not only ease their financial situation but also reduce the incentive for illegal excavations and sales.

Should museums charge entrance fees? Curiously enough, the French, whose successive governments claim to have the largest budgets to support the arts, insist that museums should charge visitors and they do so, while free admission at the six British national museums, some of which currently charge, was introduced in April 2001. Some experiments are being run in French museums on the effect of charging and not charging entrance fees once every 4 weeks, on Sundays, but the evidence that this increases the number of visitors is scant. More serious econometric work is obviously needed in this area.

How would the management of museums be affected if the value of their collections were considered as fixed assets? How should museums (as well as many other cultural institutions) be evaluated? Their output can obviously not be measured by profits, or by the number of visitors, as has often been suggested and done. According to the ICOM definition, a museum "acquires, conserves, researches and communicates, and exhibits for the purpose of study, education and enjoyment, material evidence of people and their environment." This makes for an activity with multiple outputs and calls for using appropriate efficiency frontier methods to measure performance. (Detailed accounts faced by heritage policies are discussed by Towse (1997a), Schuster (1997), and Peacock and Rizzo (2008).)

The Cultural and Creative Industries

Cultural industries mainly produce 'multiples,' such as books (Though some books can be viewed as works of art. This is obviously the case for old manuscripts, but also for limited editions which become collectors' items.), records, movies, newspapers, and TV programs, and the value of the work is not conferred by its uniqueness or its scarcity. Like in other sectors, these industries are faced with globalization and follow the move to more concentration, both in terms of production and of distribution, which may open the road to more uniformity.

At issue are also (1) the public support to (and even the ownership by the state of) television and radio stations and, in some countries, newspapers, and the complex problems of (dis) information this may generate, (2) import quotas on cultural programs in a world which moves to free trade in all other industries, (3) the complex economic and legal copyright problems (Landes and Posner 2003; Towse 2008b) as well as those generated by the internet and what came to be called 'piracy,' though, as shown by Waldfogel (2011a,b), this has not affected the creativity of singers and composers. Protecting very concentrated industries such as the recording industry by tougher copyright laws does not necessarily increase consumers' welfare.

The terms 'cultural industries' has recently morphed into 'creative industries,' the title of Caves' (2000) important volume in which contract theory is used to illuminate the workings of the industry.

Concluding Remarks

We have covered the main issues that look more like a puzzle the many pieces of which have still to be assembled, but the subject is still young, and economics itself, though it is much older, is also an unassembled puzzle. Additionally, we are left with many unresolved questions. The economics of certification or authentication ('authentic' works versus copies, or fakes) is one of these. While many art philosophers maintain that this does not change the aesthetic qualities of a work, it definitely changes its price. The formation of values also raises the interesting issue of who are the agents endowed with the power to discriminate between good and bad art. What is the expertise of those who grant prizes and awards? Who would today award the 1959 Oscar for best movie to *Ben Hur* and forget about *Some Like it Hot* produced during the same year? See English (2005) and Ginsburgh and Weyers (2008) on prizes and competitions.

Why did blockbuster exhibitions or music, opera, and theater festivals become so numerous? Don't such events simply displace the public from more traditional venues? Much has been said on the rent-seeking behavior of those who organize such events, but little is known on the long-term effects on the behavior of those who attend. See Frey and Busenhart (1996) and Gergaud and Ginsburgh (2014).

The last question, which is also one of the most important ones, is how does culture affect the preferences of consumers. This is the subject of the important article by Stigler and Becker (1977), who were among the first to discuss changing tastes including addiction to culture in terms of a household production function and the stock of past consumption (of culture) and not consumption itself, appearing as argument in the utility function. Addiction (if positive) should enhance the role of education to culture. More of it would certainly lead to more participation in cultural activities and, therefore, to less need for public support. It would also increase efficiency, since consumers would then have enough information to choose in a sovereign way, and we would have to accept the consequences to which this could lead.

I am not fully happy with the field since it is going in the direction of much more empirics than theory and more management and cultural policies than economics. There is little in terms developments that spilled over outside its own domain, as observed 13 years ago by Blaug (2001). He adds "that [this

was] perhaps asking too much of a subject no older than 30 years?" We are unfortunately still in the same situation 13 years later.

See also: Art History; Artists, Competition and Markets; Cultural Heritage; Cultural Policies in East Asia; Cultural Policy Regimes in Western Europe; Culture Policy Regimes in China; Culture and Economy; Culture, Production of: Prospects for the Twenty-First Century; Markets: Artistic and Cultural; Social Inequality in Cultural Consumption Patterns.

Bibliography

Ashenfelter, Orley, 1989. How auctions work for wine and art. Journal of Economic Perspectives 3, 23–36.

Ashenfelter, Orley, Graddy, Kathryn, 2006. Art Auctions. In: Ginsburgh, Throsby (Eds.), Handbook on the Economics of Art and Culture. Elsevier Science, Amsterdam, pp. 909–945.

Baumol, William, 1986. Unnatural value: or art investment as floating crap game. American Economic Review 76, 10–14.

Baumol, William, Bowen, William, 1966. Performing Arts. The Economic Dilemma. Twentieth Century Fund, New York.

Bauwens, Luc, Ginsburgh, Victor, 2000. Art experts and auctions: are pre-sale estimates unbiased and fully informative. Louvain Economic Review 66, 131–144.

Beggs, Alan, Graddy, Kathryn, 1997. Declining value and the afternoon effect: evidence from art auctions. Rand Journal of Economics 28, 544–565.

Benhamou, Françoise, 2000. L'économie de la culture. La Découverte, Paris.

Benhamou, Françoise, Ginsburgh, Victor, 2006. Copies of artworks: the case of paintings and prints. In: Ginsburgh, Throsby (Eds.), The Handbook of the Economics of Arts and Culture. vol. 1. Elsevier, Amsterdam.

Blaug, Mark (Ed.), 1976. The Economics of the Arts. Martin Robinson, London.

Blaug, Mark, 2001. Where are we now on cultural economics? Journal of Economic Surveys 15, 123–144.

Caves, Richard, E., 2000. Creative Industries. Harvard University Press, Cambridge, MA and London.

Courty, Pascal, Pagliero, Marco, 2013. In: Ginsburgh, Throsby (Eds.), The Pricing of Art and the Art of Pricing: Pricing Styles in the Concert Industry, 2006.

De Marchi, Neil, Van Miegroet, Hans, 1996. In: Ginsburgh, Menger (Eds.), Pricing Invention: Originals, Copies and Their Relative Value in Seventeenth Century Art Markets.

De Marchi, Neil, 2004. Auctioning paintings in late 17th century London. In: Ginsburgh, Victor (Ed.), Economics of Art and Culture. Elsevier, Amsterdam.

De Marchi, Neil, Van Miegroet, Hans, 2006. The History of Art Markets. In: Ginsburgh, Throsby (Eds.), Handbook on the Economics of Art and Culture. Elsevier Science, Amsterdam, London, and Tokyo, pp. 69–122.

English, James, 2005. The Economy of Prestige. Harvard University Press, Cambridge.

Filer, Randall K., 1986. The starving artist – myth or reality? earning of artists in the United States. Journal of Political Economy 94, 56–75.

Frey, Bruno, Busenhart, Isabelle, 1996. In: Ginsburgh, Menger (Eds.), Special Exhibitions and Festivals.

Frey, Bruno, Eichenberger, Reiner, 1995. On the rate of return in the art market: survey and evaluation. European Economic Review 39, 529–537.

Frey, Bruno, Pommerehne, Werner, 1989. Muses and Markets: Explorations in the Economics of the Arts. Basil Blackwell, Oxford.

Galenson, David, 2001. Painting Outside the Lines: Patterns of Creativity in Modern Art. Harvard University Press, Cambridge.

Galenson, David, 2003. The Life Cycles of Modern Artists. Theory, Measurement, and Implications, manuscript.

Gergaud, Olivier, Victor, Ginsburgh, 2014. Evaluating the economic effects of cultural events. In: Ateca-Amestoy, Victoria, Fernandez-Blanco, Victor, Ginsburgh, Victor, O'Hagan, John, Mazza, Isidoro (Eds.), Assessing Tools to Enhance Cultural Participation. Springer.

Ginsburgh, Victor, Menger, Pierre-Michel (Eds.), 1996. Economics of the Arts. Selected Essays. North-Holland, Amsterdam.

Ginsburgh, Victor, Throsby, David (Eds.), 2006, 2013. The Handbook of the Economics of Arts and Culture, vol. 1 and 2. Elsevier, Amsterdam.

Ginsburgh, Victor, Weyers, Sheila, 2006. Creation and life cycles of artists. Journal of Cultural Economics 30, 91–107.

Ginsburgh, Victor, Weyers, Sheila, 2008. Quantitative approaches to valuation in the arts, with an application to movies. In: Hutter, Michael, Throsby, David (Eds.), Beyond Price. Value in Culture, Economics, and the Arts. Cambridge University Press, New York.

Goodwin, Craufurd, 2006. In: Ginsburgh, Throsby (Eds.), Art and Culture in the History of Economic Thought.

Grampp, William, 1989. Pricing the Priceless. Art, Artists and Economics. Basic Books, New York.

Haskell, Francis, 1980. Rediscoveries in Art. Some Aspects of Taste, Fashion and Collecting in England and in France. Phaidon Press, London.

Heilbrun, James, Gray, Charles, 2001. The Economics of Art and Culture. Cambridge University Press, Cambridge.

Hutter, Michael, Throsby, David (Eds.), 2008. Beyond Price. Value in Culture, Economics and the Arts. Cambridge University Press, Cambridge, UK.

Landes, William, Richard, Posner, 2003. The Economic Structure of Intellectual Property Law. Harvard University Press, Cambridge, MA.

Montias, John Michael, 1982. Artists and Artisans in Delft. A Socio-Economic Study of the Seventeenth Century. Princeton University Press, Princeton, NJ.

Montias, John Michael, 1996. Quantitative methods in the analysis of 17th century Dutch inventories. In: Ginsburgh, Menger (Eds.), Economics of the Arts. North Holland, Amsterdam.

Montias, John Michael, 2004. Art dealers in Holland. In: Ginsburgh, Victor (Ed.), Economics of Art and Culture. Elsevier, Amsterdam.

Nelson, Jonathan, Zeckhauser, Richard, 2008. The Patron's Payoff. Conspicuous Commissions in Italian Renaissance Art. Princeton University Press, Princeton and Oxford.

Netzer, Dick, 1978. The Subsidized Muse: Public Support for the Arts in the United States. Cambridge University Press, Cambridge.

O'Hagan, John, 1998. The State and the Arts. An Analysis of Key Economic Policy Issues in Europe and the United States. Edward Elgar, Cheltenham.

Peacock, Alan, 1993. Paying the Piper. Culture, Music and Money. Edinburgh University Press, Edinburgh.

Peacock, Alan, Rizzo, Ilde, 2008. The Heritage Game. Economics, Policy and Practice. Oxford University Press, Oxford.

Pesando, James, Shum, Pauline, 1996. Price anomalies at auction: evidence from the market for modern prints. In: Ginsburgh, Menger (Eds.), Economics of the Arts. Elsevier, Amsterdam, pp. 113–134.

de Piles, Roger, 1708[1992]. Cours de peinture par principes. Gallimard, Paris.

Rosen, Sherwin, 1981. The economics of superstars. American Economic Review 71, 845–858.

Schuster, Mark (Ed.), 1997. Preserving the Built Heritage. Tools for Implementation. University Press of New England, Hanover and London.

Scitovsky, Tibor, 1972. What's wrong with the arts is what's wrong with society. American Economic Review 62, 62–69.

Simonton, Dean, 1997. Genius and Creativity: Selected Papers. Ablex Publishing, Greenwich, CT.

Stigler, George, Becker, Gary, 1977. De gustibus non est disputandum. American Economic Review 67, 76–90.

Throsby, David, 1983. Perception of quality in demand for the theater. In: Hendon, William, Shanahan, James (Eds.), Economics of Cultural Decisions. Abt Books, Cambridge, MA.

Throsby, David, 1994. The production and consumption of the arts: a view of cultural economics. Journal of Economic Literature XXXII, 1–29.

Throsby, David, Withers, Glenn, 1979. The Economics of the Performing Arts. St. Martin's Press, New York.

Towse, Ruth (Ed.), 1997a. Cultural Economics: The Arts, the Heritage and the Media Industries, 2 vols. Edward Elgar, Cheltenham.

Towse, Ruth (Ed.), 1997b. Baumol's Cost Disease: The Art and Other Victims. Edward Elgar, Cheltenham.

Towse, Ruth (Ed.), 2003, 2008a. A Handbook of Cultural Economics, first and second ed. Edward Elgar, Cheltenham.

Towse, Ruth (Ed.), 2008b. Recent Trends in the Economics of Copyright. Edward Elgar, Cheltenham.

Waldfogel, Joel, 2011a. Bye, Bye Miss American Pie? the Supply of New Recorded Music since Napster. National Bureau of Economic Research Working Paper 16882.

Waldfogel, Joel, 2011b. Copyright Protection, Technological Change, and the Quality of New Products: Evidence from Recorded Music since Napster. Working Paper, the Carlson School and Department of Economics, University of Minnesota, Minneapolis.

Wassal, Gregory, Alper, Neil, 1992. Toward a unified theory of the determinants of the earnings of artists. In: Towse, Ruth, Kakhee, Abdul (Eds.), Cultural Economics. Springer Verlag, Heidelberg, pp. 187–200.

Art and Socialisation

Ineke Nagel and Harry BG Ganzeboom, Free University Amsterdam, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the impact of arts socialization on participation in arts. Arts participation is characterized by large social inequalities. We focus on early arts socialization by parents and later socialization at school and discuss two models that explain the link between social inequalities and arts socialization: the cultural reproduction model and the cultural mobility model. The models differ in the importance they attach to early socialization by the parents and later socialization in schools. In addition, we discuss the status theory and the information theory, which explain effects of arts socialization, respectively, by social context and cognitive capacities. Finally, we review some relevant empirical studies, noting their substantive findings and/or methodological shortcomings.

Introduction

Visiting museums, going to the theater, attending a concert, or reading a novel are all expressions of personal taste. However, they are also reflective of a person's position in society, a connection that is much more salient in the arts than in any other lifestyle domain (Bourdieu, 1984[1979]). Those who enjoy the arts are often better educated and were exposed to the arts at home. The strong influences of both family and education on arts participation suggest that early socialization is decisive: cultural preferences start early, with parents and schools acting as socializing agents. Research shows that these effects endure over the life cycle and may even become stronger in early adulthood. This article examines the relative impact of family and education as socializing agents on participation in the arts in early and later life.

Socialization is the process by which skills, knowledge, values, and habits prevalent in a particular social group are conveyed to newcomers who become part of that group, either as the next generation or as new members. Parents are responsible for the socialization processes of early childhood, whereas school and peer groups gain importance as the child grows older. In adulthood, socialization processes continue, e.g., in friendship networks, in partner relationships, and at the workplace.

Studies on arts socialization distinguish between primary socialization, i.e., early socialization by parents, and secondary socialization, outside the family of origin, starting somewhat later in life, particularly in education (De Jager, 1967). Socialization processes are partly intentional, to the extent that parents or teachers overtly encourage and correct children's attitudes and behavior, and partly unintentional, to the extent that children observe (or model) others (Bandura, 1969).

Research on social stratification and social mobility identifies participation in the arts as a primary component of 'cultural capital' (Bourdieu and Passeron, 1990[1977]) that pays off in both educational and occupational attainment. The literature proposes two models explaining how people acquire such cultural capital (DiMaggio, 1982): the cultural reproduction model and the cultural mobility model (see **Figure 1**). The cultural reproduction model holds that family arts socialization induces arts participation and is used by high-status parents to ensure a similar social position for their offspring. The cultural mobility model implies that school and peers can compensate for a lack of cultural resources at home, and arts participation can become a means of upward mobility.

On the effective ingredients of arts socialization, two theories can be identified. Status theory considers arts participation to be a normatively driven activity and stresses the social context as the operative part of socialization. By way of

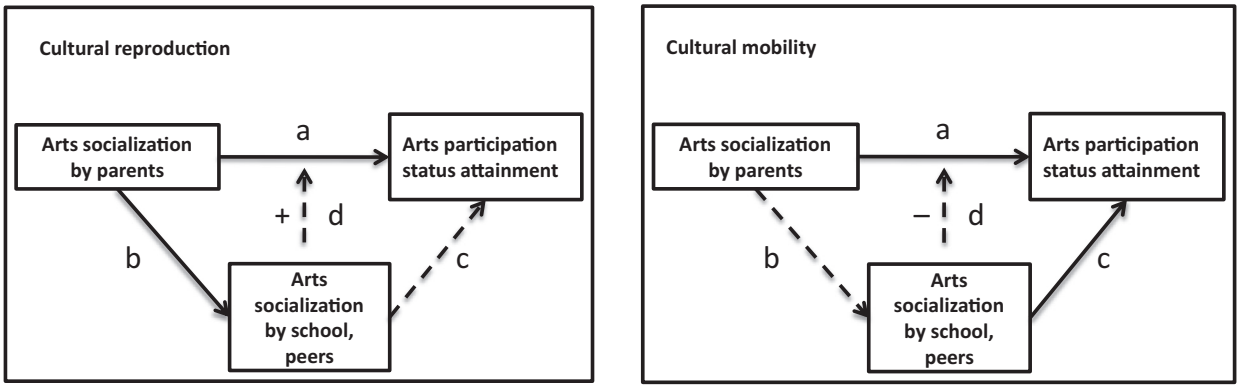

Figure 1 Cultural reproduction model and cultural mobility model.

contrast, information theory views arts participation as a cognitive activity and emphasizes instruction and the transmission of knowledge on art as crucial to socialization.

In what follows, we briefly discuss the cultural reproduction model and the cultural mobility model, before turning to basic theories of arts socialization. We then summarize the empirical evidence.

Cultural Reproduction and Cultural Mobility

The cultural reproduction model and the cultural mobility model both stress the social advantages that accrue from cultural capital, but they disagree on how it is acquired: the former says it comes from parents (Bourdieu and Passeron, 1990[1977]), and the latter points to influences outside the parental home (DiMaggio, 1982). In other words, they offer different views on the relative importance of primary socialization by the parents and secondary socialization outside the family of origin, especially in education.

Cultural Reproduction Model (Bourdieu)

The cultural reproduction model states that arts socialization is a key strategy by which parents pass their social position onto their children. In this view, arts socialization is exclusively provided by parents and is only effective when acquired at home. Bourdieu et al. (1991[1969]: 64–70) argue that the capabilities and attitudes needed to appreciate the fine distinctions that characterize art and art forms can only be acquired at home. In their view, appreciating art requires a process of gradual familiarization and regular exposure over a long period, and only parents can provide this. Children of culturally active parents, without being explicitly taught about art, regularly come into contact with various art forms and develop the attitude and knowledge required for appreciation. Bourdieu et al. (1991[1969]) argue that schools generally cannot provide the extensive and uninterrupted exposure necessary to develop a lasting appreciation of art. Schools may take their students to artistic events (plays, exhibits, etc.) and help them to appreciate certain works of art (books, paintings, etc.), but this can never be as extensive as the exposure provided at home. Dutch sociologist De Jager (1967) offers similar arguments, claiming that the impact of family socialization is much stronger; the influence of other socializing agents is only temporary and not comparable to the persistent and overriding parental influence.

According to cultural reproduction theory (Bourdieu and Passeron, 1990[1977]), schools reproduce existing social and cultural inequalities. A school diploma represents a particular status culture that corresponds to the graduating students' social positions; the prestige of a school varies according to the students it attracts and the value of the diplomas it grants (Bourdieu, 1984[1979]: 25–26). In addition, teachers evaluate students not only by their cognitive abilities, but also by their fitness for future positions. Students who are familiar with culture and art are assessed as suitable for higher status destinations and are favored by teachers who belong to these status groups themselves. As a result, students from culturally active families do better in school. The process is reinforced by the students who, in a process of anticipatory socialization (Roe, 1992), adapt their aspirations to accord with the social positions they are expected to assume.

According to the cultural reproduction model depicted in **Figure 1**, arts socialization in the parental family strongly affects arts participation in adult life (a) In addition, the model assumes that children who have been raised with art go to more prestigious schools where arts are valued and mix with peers from similar family backgrounds (b) The school environment may offer independent arts socialization, but its effects are minor (c) if arts socialization in school is at all helpful, it will help those who have already been familiarized with art at home, simply strengthening the effects of parental arts socialization (d) Arts socialization at school does not fill a gap for those children who have not acquired cultural capital in the parental family.

Cultural Mobility Model (DiMaggio)

A contrasting idea on the formation of cultural capital is offered by the cultural mobility model (DiMaggio, 1982). When DiMaggio (1982) empirically tested Bourdieu's cultural reproduction model in US high schools by examining whether students' cultural capital affected their grades, he found that over and above students' cognitive abilities, teachers gave higher grades to those with more cultural capital. Interestingly, DiMaggio also found that students' cultural capital was not or hardly related to the social position of their parents. He termed this latter finding cultural mobility.

While the cultural mobility model shown in **Figure 1** acknowledges the role of cultural preferences in status attainment processes and does not deny the importance of the family in the formation of cultural capital (a) it assumes there are ample opportunities to acquire cultural capital outside the family, and these are open to lower status groups (b) It further assumes that cultural capital gained this way will be as effective as cultural capital gained at home (c) In other words, students who are not raised with art in their family of origin may find other opportunities to acquire cultural capital, particularly at school (DiMaggio, 1982; DiMaggio and Useem, 1978). Opportunities include the school curriculum, extracurricular activities, and peer groups. As cultural capital is valued irrespective of how it is acquired, it can be used as a strategy to achieve upward mobility. Ambitious students from lower status families will be able to bring their cultural preferences and arts participation to a higher level to accord with their desired social positions. In such cases, schools compensate for a lack of parental influence (d).

Cultural Reproduction vs Cultural Mobility

Both models indicate the importance of participation in the arts for status achievement. Sensitivity to the arts is a key issue, especially for those lacking a high-status background. Particularly interesting in this respect, not just for academics and educators, but also for policy makers, is the role of the school and the extent to which the school offers students opportunities to learn about the arts, thereby increasing their chances of entering higher status groups.

Two Mechanisms of Arts Socialization: Status Motives and Cultural Competence

While the cultural reproduction model and the cultural mobility model problematize the relative influence of family vs school (and other social contexts), they do not focus on the underlying mechanisms of socialization. Socialization processes inherent to the formation of cultural capital include the transmission of social norms and cultural competence. On the question which of the two is more effective in furthering arts participation, two competing theories have been proposed (Ganzeboom, 1982): status theory and information theory.

Status theory assumes that arts participation is normatively driven; therefore, a student's various social contexts (family, school, peer group), and their respective norms are the operative parts of the socialization process. According to information theory, however, arts participation is a cognitive activity and the transmission of knowledge (i.e., information transfer) is the crucial part of socialization. It also points to generic cognitive capabilities that are present before parental and/or school socialization starts. The literature on the effects of cultural capital on educational attainment proposes a similar distinction in mechanisms of arts socialization by differentiating between reading, assumed to be driven entirely by cognitive capacities, and other, more socially driven disciplines (Crook, 1997; De Graaf, 1986; De Graaf et al., 2000; De Graaf and De Graaf, 2002; Evans et al., 2010; Kingston, 2001).

Status Theory

Cultural capital is defined by Lamont and Lareau (1988: 156) as consisting of "institutionalized, i.e., widely shared, high status cultural signals (attitudes, preferences, formal knowledge, behaviors, goods and credentials) used for social and cultural exclusion." Status theory (Ganzeboom, 1982) emphasizes that arts participation, one aspect of cultural capital, provides a visible demarcation in the social hierarchy, and is driven by a desire to maintain these boundaries (Bourdieu, 1984[1979]). It expresses membership in higher social status groups and can be used to assess the suitability of newcomers to enter these groups. Thus, it maintains the boundaries between social status groups.

According to status theory, arts socialization includes learning definitions of correct taste or what is and what is not appropriate, in short, the social norms of arts consumption. Status theory assumes that social contexts in which the arts are valued and where arts participation is the norm will stimulate participation among their members. New members will gradually develop the tastes and behaviors that match the social group. Normative arts socialization starts in the parental family, but in a later stage, other social contexts – school, peers – add relevant socialization contexts. Status theory works for both the cultural reproduction and cultural mobility models: for those already socialized into the arts by the family, socialization at school accords with the family background; for others, introduction to the arts will start at school or in the peer group. The status mechanism can be reconciled with both cultural reproduction and cultural mobility. The two models differ in the relative importance they attach to early and later social contexts.

An important implication of status theory bears on participation in various art disciplines. To the extent that arts participation represents a single status culture, participation will occur in several art disciplines at the same time (DiMaggio, 1982: 191–192). In other words, normative arts socialization and its effects will be diffuse and involve multiple disciplines, as long as they procure prestige in the social status group.

Information Theory

Information theory (Ganzeboom, 1982, 1984) emphasizes that the arts constitute complex sources of information and their enjoyment requires a considerable amount of cognitive capacity. Those who lack these capacities will experience art as difficult, making them likely to refrain from arts participation. An enhanced knowledge of art that leads to greater participation results from both intended and unintended instructions, and both acquired knowledge and generic cognitive abilities play a role. Information theory does not make predictions on the effects of the timing of arts instruction (i.e., in early or later childhood, or in adolescence), but as noted above, the cultural reproduction and cultural mobility models offer contrasting accounts of the value of arts instruction in childhood (by parents) and adolescence (by school and peers).

Although knowledge of art may be transferable from one art discipline to another (Bourdieu et al., 1991[1969]), to the extent that art socialization concerns the transfer of specialized knowledge, it will particularly affect the art discipline in which the instruction takes place. For the most part, discipline-specific arts lessons will lead to separate art publics. Information theory also predicts that the effects of arts socialization will be strongest for art practices for which more artistic competence is a prerequisite, hence, the more complex disciplines.

Finally, information theory asserts that generic cognitive abilities enhance the capacity to process the complexity of art works. Assuming these generic cognitive abilities are present early in life, this implies that differences in arts participation are due to differences in already existing cognitive capacities. This has important implications for assessing the empirical evidence. Because measures for generic cognitive abilities are scarce in research (see below), these abilities are generally represented by educational level. An effect of educational level, therefore, cannot directly be interpreted as an effect of arts socialization during schooling.

To the extent that generic cognitive abilities affect arts participation, there will be overlap between art forms; however, assuming that reading is a more cognitive activity, the effects of generic cognitive abilities will be stronger for reading.

Methodological Challenges

Empirical research on the total and relative impact of family and school as socializing agents in arts participation faces many challenges. Ideally, determining the causal effects of arts socialization requires (quasi-)experimental designs, but their use is limited to assessing the effects of specific arts instruction programs. In practice, the effects of arts socialization can only be studied through observational survey designs in which respondents are asked to report on the issues of interest.

Information is collected on respondents' current arts participation and their education, but not always on the arts socialization they have experienced in different social contexts. For instance, the number of studies that include explicit measures of parents' arts socialization is limited (Sullivan, 2011: 197).

A first serious pitfall of arts socialization research is that education effects are attributed to normative socialization and instruction in schools, but they may simply reflect the filtering effect of education. In modern and meritocratic societies, one role of education is to sort students by their cognitive capacities. The differences found in arts participation between higher and lower educated persons in studies that only include education as a variable cannot be attributed to arts socialization by school. At best, education may reflect arts socialization if a measure of generic cognitive capacities is controlled, and measures of generic capacities, in turn, capture only part of the educational filtering effect.

Second, studies that include information on school socialization practices may suffer from selection bias: arts socialization provided by school in optional courses may be dependent on family background or prior interest in the arts. Therefore, to determine the effects of arts socialization in school, parental socialization practices should be controlled, or at least parents' level of education; otherwise, the effects of the school's arts socialization could be due to the parents' socialization. The problem of selection bias does not apply to the effect of parental arts socialization, because it can be safely assumed that parental participation represents independent family socialization, once education and parental education are controlled. Causality in observational research can be identified by applying longitudinal designs comparing respondents over time, before and after exposure to socialization contexts (Nagel et al., 2010). Such panel designs allow for controlling prior arts participation and assessing the additional influence of secondary socialization.

Third, an approach that uses only measures of arts socialization to assess the relative influence of primary and secondary socialization contexts remains necessarily incomplete. Ideally, to estimate the total and relative impact of family and education, all influences that children have been exposed to by the family and by school should be taken into account. Such total effects of family and school contexts can be established in multilevel models by comparing the degree of resemblance in arts participation between students and their siblings (family context) and their schoolmates (education context). However, it is virtually impossible to combine comparisons between siblings and schoolmates simultaneously in a single design. The family context is more or less stable, whereas children may experience several school and class contexts.

Fourth, most measurement of early socialization occurs retrospectively and by proxy and may be vulnerable to both random and systematic measurement error; as such, reports may contain too much noise or be biased toward current practices (De Vries and De Graaf, 2008). Researchers have occasionally been successful in obtaining nonproxy measures, i.e., by obtaining reports of the socializing agents themselves.

Fifth, the effects of parents' arts participation controlled for education are often interpreted as the normative dimension of arts socialization, while the effects of education controlled for parents' socialization are considered to represent the effect of cognitive capacities. But part of the arts socialization by parents could be the transmission of cultural competence, and part of the effect of education could be due to social norms on arts participation. One way to learn more about the active ingredients in family and school socialization would be to include explicit measures of social norms or cultural competence. However, we do not know of any studies with such measures that can safely be assumed to be causally prior to actual arts participation. Another way to identify the active ingredients of arts socialization is to differentiate among various art forms in the degree to which they signal high status and cognitive capacity, and to see if the effects are stronger for one or the other.

Empirical Evidence

Much of the existing empirical research is from the Netherlands, where a strong research tradition has developed, beginning with the work of Wippler (1968) and De Jager (1967) in the 1960s. This tradition continued with Ganzeboom (1984), De Graaf (1987), Kraaykamp (1993), and Van Eijck (1996), along with a number of other dissertation projects. As a consequence, in the Netherlands, many data sets on cultural participation are available, including arts socialization by parents and school. (Many of these are accessible via the Dutch data archive www.dans.knaw.nl.)

Family Socialization

Without exception, studies that include parents' arts participation report moderately strong effects. Typically, parents' arts consumption is measured by adult respondents' retrospective reports on their own parents' arts participation while they were growing up (usually between 12 and 16 years of age). These studies find that when education is controlled, parents' cultural participation is a strong determinant of the children's arts participation, with standardized effects around .30 years (e.g., Kraaykamp and Nieuwbeerta, 2000; Kraaykamp and Van Eijck, 2010; Yaish and Katz-Gerro, 2012). Slightly stronger effects are reported in studies that control random measurement error (De Graaf and De Graaf, 1988; Ganzeboom, 1982) and systematic measurement error (De Vries and De Graaf, 2008).

Strong resemblances between parents' and children's contemporaneous arts participation are found in many studies (Crook, 1997; Damen et al., 2010; Jæger, 2009; Sullivan, 2001; Van Wel et al., 2006; Willekens and Lievens, 2014), albeit to a lesser extent when parents report on their arts participation independently (Nagel, 2010). Interestingly, Van Wel et al. (2006) and Willekens and Lievens (2014) show that a mother's influence is generally stronger than a father's influence. Van Wel et al. (2006) also find that her influence is stronger on her daughters than on her sons. Although arts participation in adolescence is part of the socialization process itself, the contemporaneous effects of family imply that cultural careers start and take decisive turns quite early.

Similar strong associations are found between parents' literary reading during the socialization period and later in life (Kraaykamp, 2003; Notten, 2011). Research using independent measures of parents' socialization also finds that parents'

reading levels affect the reading of prestigious books in adulthood (Verboord and Van Rees, 2003). Nagel and Verboord (2012) find that reading frequency among adolescents and young adults is strongly determined by parental reading and secondary education. In fact, the effects of parents' reading frequency increase a little in this period.

In addition to the effects of parental example (unintentional learning), Kraaykamp (2003) and Notten (2011) find effects for deliberate parental strategies to stimulate children's literary reading, such as reading to them, giving them books as presents, and showing interest in their reading. However, Notten (2011: 102) concludes that direct imitation is the main process by which parents transmit their preferences in reading and television to their children. Verboord and Van Rees (2003: 295) do not find additional effects of parents' reading socialization activities on the reading level of their children. Lack of effect of explicit instruction is confirmed by Kraaykamp and Dijkstra (1999) who, in a sample of book readers, find no effects of parents' socialization activities on reading level when the educational level is taken into account. By contrast, Verboord (2005) finds effects of parental socialization practices and attitudes on the frequency of reading, and Tepper (2000) finds positive effects of parental encouragement on reading fiction in the United States.

The above studies assess the impact of parental arts socialization by explicit measures of parents' arts participation or socialization practices, but parents may affect their children's interest in arts in other ways. Total effects of family of origin can be estimated by studying sibling resemblance in arts participation. The few studies that study arts participation by using a sibling design generally report stronger influences of families (Ganzeboom and De Graaf, 1991; Van Eijck, 1997; Nagel and Ganzeboom, 2002; Willekens and Lievens, 2014) than those using measured parental characteristics. Van Eijck (1997: 215) concludes that shared family characteristics account for 37% of the variance in the cultural consumption of adult siblings. Van Eijck (1997: 221) infers that parents' cultural resources 'are the strongest determinant of siblings' cultural participation,' more important than education.

Nagel and Ganzeboom (2002) find 38% explained sibling variance in a sample of former secondary school students aged 25–35 years; when the secondary school context is controlled, this figure drops to 28%, still far above the findings when explicit socialization measures are used. They also find that the family effects are rather stable in the period from adolescence to adulthood. Although these total effects include other shared family characteristics besides arts socialization practices (i.e., shared general cognitive ability), parents' arts participation explains a large part of these total effects; the stability in arts participation is mainly caused by family of origin. While the effects of parents' arts socialization are stable between adolescence and early adulthood, however, the effects of education increase in that period. By way of contrast, in a Belgian study, Vander Stichele and Laermans (2007) find no effects of cultural participation at age 12 or 14 years on later arts participation.

To summarize, many studies report moderately to strong effects of parental socialization on a person's participation in the arts, both in adolescence and in adulthood. Studies of reading suggest that the example set by parents is more important than their intentional socialization practices. The available evidence also suggests that the family influence starts early in life and remains stable afterward. The strong effects of parental socialization fit both the cultural reproduction and the cultural mobility model; both predict that arts socialization by parents has an early and lasting effect on children's arts participation. However, the stability of the effects of parents' socialization (strong in adolescence and in adulthood) favors the cultural reproduction model.

School Socialization

The school as a socialization context encompasses the level of education (lower/higher primary, secondary, and tertiary), the curriculum (including school-based arts instruction), the school climate (extracurricular activities), and the peer group. Of these, only the effects of arts education and the level of education have been researched systematically.

Research has found that parents' arts participation, next to their education, affects their children's type of schooling, more so in the early school career (Aschaffenburg and Maas, 1997). Moreover, enrollment in school-based arts education turns out to be highly affected by prior interests in the arts and by parents' arts participation (Nagel et al., 1997). Finally, peers in school classes select each other partly because of similarity in their parents' arts participation (Nagel et al., 2011). Thus, parental and school contexts are often at least partly aligned, and to assess the effects of socialization at school, it is imperative to control the effects of parental arts socialization.

Arts Education

Using the US Survey on Public Participation in the Arts (SPPA) 1992, Kracman (1996) finds that school-based arts education affects museum visiting and performing arts attendance later in life. The effects arise over and above those of the mother's and father's education and the educational level of the respondent, but the research is entirely based on retrospective reports. For the 2008 SPPA survey, Christin (2012) also reports a positive effect of art lessons in childhood on a broad spectrum of highbrow culture. However, as both analyses lack parents' arts socialization practices as a control, the effect of arts education may have been overestimated.

Nagel et al. (1997) use a prospective design in which enrollment in a specific arts examination program was measured by secondary school archival records. Although enrollment in this program was found to be selective according to parental socialization, prior arts participation, and humanities in the curriculum, it turned out that 10–20 years later in the life course, former secondary school students who had taken an exam in visual arts and music more often participated in visual arts and music (attendance at events and participation at home) than their schoolmates who graduated from the same schools but did not take these arts exams. Overall, the effects of arts education are smaller than the combined effects of prior and parental arts participation, but larger than the effect of the level of secondary education. The effects occur mainly within disciplines: visual art education enhances museum attendance; music affects concert attendance. Two other studies on the same data

reveal that the effects are smaller if arts participation is restricted to arts attendance (Nagel, 2004) or when a composite measure of different art disciplines is used (Nagel and Ganzeboom, 2002). Nagel (2004) shows that the effects of arts education on arts attendance are highest in young adulthood and then fade away.

Nagel et al. (2010) evaluate an arts course in the Netherlands that compels secondary school students (age 15–16 years) to visit 6–10 arts events in several art disciplines. Two to six years after completion of the course there were no differences in arts participation between students who had been enrolled in the course and a previous cohort who had not. Damen (2010:31) characterizes the course as 'too little, too late,' pointing at the multidisciplinary character of the arts course and the timing (ages 15–16 years) in the curriculum.

With respect to reading literary works, positive effects of school socialization practices have been reported. Kraaykamp (2003) finds that literary reading is stronger among those who took more humanities courses in the secondary school curriculum and who reported a stronger interest in culture when they were in secondary school. In a sample of book readers, Kraaykamp and Dijkstra (1999) find (equally) positive effects of school stimulation on the complexity and literary prestige of reading preferences. Although measures of parental socialization practices are taken into account in these studies, the measure of school socialization will probably include some self-selection effects. Verboord and Van Rees (2003) find that the time spent at secondary school on literature enhances the reading of prestigious books in adulthood, over and above parental socialization. In this study, the direct effects of school socialization are comparable in size to the direct effects of parental socialization and secondary education (but as the model comprises intervening variables, the relative sizes of the total effects remain unclear). Verboord (2005) finds that reading frequency is not affected by the amount of literary education in secondary school, but by the didactic approach of the teacher (student-centered teaching methods enhance reading; teacher-centered methods have the opposite effect).

To summarize, art lessons in secondary school may enhance arts participation later in life, but the effects are weak and seem somewhat more convincing for reading than for arts attendance. The effects of art lessons appear to be confined to the same arts discipline in which instruction took place, suggesting that the active ingredient of art socialization by school is the specific cultural competence.

General Education

Without exception, studies of educational effects on arts participation find strong associations with the level of education (DiMaggio and Useem, 1978; Kracman, 1996; Christin, 2012), but it is important to control for parents' arts socialization and educational filtering.

Studies including parents' arts participation report effects of education that are roughly equivalent to the effects of parents' socialization: while some report slightly larger effects of parents' arts participation (Kraaykamp and Nieuwbeerta, 2000), others find somewhat stronger effects of education (De Graaf and De Graaf, 1988; De Vries and De Graaf, 2008; Kraaykamp and Van Eijck, 2010; Yaish and Katz-Gerro, 2012). Still, the level of schooling may largely represent cognitive abilities; thus, the effect of educational level cannot be interpreted as evidence of the effect of arts socialization at school. One way to overcome this problem is to include a measure of cognitive capacities. Unfortunately, survey data sets rarely contain such a measure.

Another approach to overcome selection problems is to use dynamic designs that compare arts participation before and after enrollment in education. For example, Nagel and Ganzeboom (2002) find that the effects of final attained education (over and above secondary schooling) are not yet present in adolescence, but do appear around the age of 30 years. Apparently, the effect of final education rises during the life course and can be interpreted as the consequence of schooling. Other studies of adolescents corroborate the finding that the short-term effects of educational level are relatively small compared to the effects of education in adulthood (Van Wel et al., 2006). Nagel (2010) finds that in the period from adolescence to adulthood, the effects of education partly exist at the beginning of adolescence and partly increase as the educational career progresses.

Nagel and Verboord (2012) assert that this does not hold for reading. Reading frequency varies more strongly with secondary education than with parental arts socialization, but the effects of secondary education are transitory and probably have to be attributed to actual school socialization practices. They say there are no additional school effects on reading frequency over and above those already existing early in adolescence (at 14 years of age).

Daenekindt and Roose (2013: 321) differentiate between private and public cultural participation. They find that cultural participation is determined more by attained education than by the parents' education, more so in public participation (arts attendance) than private participation (media use). Although the larger effect of attained education could be due to educational filtering, the stronger effect on public than on private arts participation points to 'social motives.'

To summarize, although the association of education with arts participation is comparable in size to that of parents' arts socialization, only part of it can be attributed to the school as a socialization context. Educational differences in arts participation in adolescence, prior to completion of the educational career, are most likely caused by generic cognitive abilities but may also result from former education or be signs of anticipatory socialization (Nagel, 2010).

Conclusions

When we look at the relative importance of cultural reproduction and cultural mobility processes, we derive the following conclusions from the empirical evidence.

First, there is a strong degree of intergenerational transmission of arts participation: parents' arts participation is a strong determinant of their offspring's arts participation. Their example (unintentional socialization) is especially important, not, or to a lesser extent, their intentional socialization practices. Differences in arts participation emerge early

in life: these are already present in adolescence and remain stable in adulthood.

The large effects of parental socialization fit both the cultural reproduction model and the cultural mobility model. The stability of the effects of parental socialization (strong in adolescence and adulthood) favors the cultural reproduction model, but these effects do not negate the cultural mobility model's contention that arts socialization outside the family can be equally effective. The differences between the models depend on the effects of education as a socialization context, as well as how the two contexts combine.

Although arts participation is strongly associated with education, drawing conclusions on the causal effects of education as a socialization context is difficult, as there are multiple selection problems in the available empirical evidence. To the extent that we can attribute the school socialization effects to school, there is limited evidence of cultural mobility. Arts instruction in school may enhance arts participation later in life, but the evidence is weak, albeit a bit more convincing for reading than arts attendance. The few studies that try to unravel the additional influence of schooling by using a dynamic design find additional influences of educational level on arts participation, suggesting that the education effects are not solely a matter of cognitive filtering.

While the results generally favor the cultural reproduction model over the cultural mobility model, future research should aim for more decisive conclusions on the tenability of either model. First, researchers should decompose the net effect of the level of education into its underlying mechanisms. This could be done by including (independent) measures of generic cognitive abilities (in childhood) and measures of the mechanisms of schooling, not only of arts instruction, but also of the social context, i.e., the social composition of the school and the peer groups in the classroom. Second, to differentiate between the status and information mechanisms, more detailed information should be collected on arts that differ in their requirement of cognitive abilities and their status-enhancing qualities. Finally, instrumental variables (i.e., effects of socialization that affect arts participation exclusively via education) should be used to explicitly model intervening variables and examine to what extent education enhances cultural competence or status motives. Such econometric designs have not yet been applied in the research literature on art socialization.

See also: Cultural Capital and Education; Cultural Participation, Trends In; Culture, Cognition and Embodiment; Embodiment and Culture; Leisure and Cultural Consumption: The European Perspective; Leisure and Cultural Consumption: US Perspective; Social Inequality in Cultural Consumption Patterns; Symbolic Boundaries.

Bibliography

Aschaffenburg, K., Maas, I., 1997. Cultural and educational careers: the dynamics of social reproduction. American Sociological Review 62 (4), 573–587.

Bandura, A., 1969. Social learning theory of identificatory processes. In: Goslin, D.A. (Ed.), Handbook of Socialization Theory and Research. Rand McNally, Chicago, pp. 213–262.

Bourdieu, P., 1984[1979]. Distinction. A Social Critique of the Judgement of Taste. Routledge & Kegan Paul, London.

Bourdieu, P., Passeron, J.-C., 1990[1977]. Reproduction in Education, Society and Culture. Sage, New Delhi, Thousand Oaks, London.

Bourdieu, P., Darbel, A., Schnapper, D., 1991[1969]. The Love of Art. European Art Museums and Their Public. Polity Press, Cambridge.

Christin, A., 2012. Gender and highbrow cultural participation in the United States. Poetics 40, 423–443.

Crook, C.J., 1997. Cultural Practices and Socioeconomic Attainment. The Australian Experience. Greenwood Press, Westport, CT.

Daenekindt, S., Roose, H., 2013. Cultural chameleons: social mobility and cultural practices in the private and the public sphere. Acta Sociologica 56, 309–324.

Damen, M.-L., 2010. Cultuurdeelname en CKV. Studies naar effecten van kunsteducatie op de cultuurdeelname van leerlingen tijdens en na het voortgezet onderwijs [Cultural participation and CKV. Studies on Effects of Arts Education on Students' Cultural Participation during and after Secondary Education]. Dissertation Utrecht University, Utrecht.

Damen, M.-L., Nagel, I., Haanstra, F., 2010. Short-term effects of compulsory multidisciplinary secondary school arts education on cultural participation in the Netherlands. Studies in Art Education: A Journal of Issues and Research 51 (2), 147–161.

De Graaf, P.M., 1986. The impact of financial and cultural resources on educational attainment in the Netherlands. Sociology of Education 59, 237–246.

De Graaf, P.M., 1987. De invloed van financiële en culturele hulpbronnen in onderwijsloopbanen [The Impact of Financial and Cultural Resources on Educational Careers]. Instituut voor Toegepaste Sociale Wetenschappen, Nijmegen. Dissertation Utrecht University, Utrecht, Utrecht.

De Graaf, N.D., De Graaf, P.M., 1988. Family background, postmaterialism and lifestyle. The Netherlands Journal of Social Sciences 24, 50–64.

De Graaf, N.D., De Graaf, P.M., 2002. Formal and popular dimensions of cultural capital: effects on children's educational attainment. The Netherlands Journal of Social Sciences 38 (2), 167–186.

De Graaf, N.D., De Graaf, P.M., Kraaykamp, G., 2000. Parental cultural capital and educational attainment in the Netherlands: a refinement of the cultural capital perspective. Sociology of Education 73 (2), 92–111.

De Jager, H., 1967. Cultuuroverdracht en concertbezoek [Cultural Transmission and Concert Attendance]. Stenfert Kroese, Leiden [Dissertation Utrecht University], Utrecht.

De Vries, J., De Graaf, P.M., 2008. Is the intergenerational transmission of high cultural activities biased by the retrospective measurement of parental high cultural activities? Social Indicators Research 85, 311–327.

DiMaggio, P., 1982. Cultural capital and school success: the impact of status culture participation on the grades of U.S. high school students. American Sociological Review 47 (2), 189–201.

DiMaggio, P., Useem, M., 1978. Social class and arts consumption: the origins and consequences of class differences in exposure to the arts in America. Theory and Society 5 (2), 141–161.

Evans, M.D.R., Kelley, J., Sikora, J., Treiman, D.J., 2010. Family scholarly culture and educational success: books and schooling in 27 nations. Research in Social Stratification and Mobility 28, 171–197.

Ganzeboom, H.B.G., 1982. Explaining differential participation in high-cultural activities: a confrontation of information-processing and status-seeking theories. In: Raub, W. (Ed.), Theoretical Models and Empirical Analyses. ES-Publications, Utrecht, pp. 186–205.

Ganzeboom, H.B.G., 1984. Cultuur en informatieverwerking [Culture and Information Processing]. Dissertation Utrecht University, Utrecht.

Ganzeboom, H.B.G., De Graaf, P.M., 1991. Sociale herkomst, culturele socialisatie en cultuurparticipatie: een sibling-analyse [Social background, cultural socialization and cultural participation: a sibling analysis]. Sociale Wetenschappen 34 (4), 272–287.

Jæger, M.M., 2009. Equal access but unequal outcomes: cultural capital and educational choice in a meritocratic society. Social Forces 87 (4), 1943–1971.

Kingston, P.W., 2001. The unfulfilled promise of cultural capital theory. Sociology of Education 74, 88–99. Extra Issue: Current of Thought: Sociology of Education at the Dawn of the 21st Century.

Kraaykamp, G., 1993. Over lezen gesproken. een studie naar sociale differentiatie in leesgedrag. [Talking about Reading. A Study on Social Differentiation in Reading behavior.] Thela Thesis, Amsterdam [dissertation Utrecht University].

Kraaykamp, G., 2003. Literary socialization and reading preferences. Effects of parents, the library, and the school. Poetics 31, 235–257.

Kraaykamp, G., Dijkstra, K., 1999. Preferences in leisure time book reading: a study on the social differentiation in book reading for the Netherlands. Poetics 26, 203–234.

Kraaykamp, G., Nieuwbeerta, P., 2000. Parental background and lifestyle differentiation in Eastern Europe: social, political, and cultural intergenerational transmission in five former socialist societies. Social Science Research 29, 92–122.

Kraaykamp, G., Van Eijck, K., 2010. The intergenerational reproduction of cultural capital: a threefold perspective. Social Forces 89 (1), 209–231.

Kracman, K., 1996. The effect of school-based arts instruction on attendance at museums and the performing arts. Poetics 24 (2–4), 203–218.

Lamont, M., Lareau, A., 1988. Cultural capital: allusions, gaps and glissandos in recent theoretical developments. Sociological Theory 6 (2), 153–168.

Nagel, I., 2004. Cultuurdeelname in de levensloop [Cultural Participation in the Life Course]. Dissertation Utrecht University, Utrecht.

Nagel, I., 2010. Cultural participation between the ages of 14 and 24: Intergenerational transmission or cultural mobility? European Sociological Review 26 (5), 541–556.

Nagel, I., Damen, M.-L., Haanstra, F., 2010. The arts course CKV1 and cultural participation in the Netherlands. Poetics 38, 365–385.

Nagel, I., Ganzeboom, H.B.G., 2002. Participation in legitimate culture: family and school effects from adolescence to adulthood. The Netherlands Journal of Social Sciences 38 (2), 102–120.

Nagel, I., Ganzeboom, H.B.G., Haanstra, F., Oud, W., 1997. Effects of art education in secondary schools on cultural participation later in life. Journal of Art & Design Education 16 (3), 325–331.

Nagel, I., Ganzeboom, H.B.G., Kalmijn, M., 2011. Bourdieu in the network: the influence of high and popular culture on network formation in secondary school. In: Rössel, J., Otte, G. (Eds.), 'Lebensstilforschung', Kölner Zeitschrift für Soziologie und Sozialpsychologie (Sonderheft 51), pp. 424–446.

Nagel, I., Verboord, M., 2012. Reading behaviour from adolescence to early adulthood: a panel study of the impact of family and education on reading fiction books. Acta Sociologica 55 (4), 351–365.

Notten, N.J.W.R., 2011. Parents and the Media. Causes and Consequences of Parental Media Socialization. Dissertation Radboud University (Nijmegen).

Roe, K., 1992. Different destinies – different melodies: school achievement, anticipated status and adolescents' tastes in music. European Journal of Communication 7, 335–357.

Sullivan, A., 2001. Cultural capital and educational attainment. Sociology 35, 893–912.

Sullivan, A., 2011. The intergenerational transmission of lifestyles. In: Rössel, J., Otte, G. (Eds.), 'Lebensstilforschung', Kölner Zeitschrift für Soziologie und Sozialpsychologie (Sonderheft 51), pp. 196–222.

Tepper, S.J., 2000. Fiction reading in America: explaining the gender gap. Poetics 27, 255–275.

Van Eijck, K., 1996. Family and Opportunity. A Sibling Analysis of the Impact of Family Background on Education, Occupation and Consumption. Tilburg University Press, Tilburg [Dissertation Tilburg University].

Van Eijck, K., 1997. The impact of family background and educational attainment on cultural consumption: a sibling analysis. Poetics 25, 195–224.

Van Wel, F., Couwenbergh-Soeterboek, N., Couwenbergh, C., Ter Bogt, T., Raaijmakers, Q., 2006. Ethnicity, youth cultural participation, and cultural reproduction in the Netherlands. Poetics 34, 65–82.

Vander Stichele, A., Laermans, R., 2007. Wie doet wat en waarom? De invloed van culturele socialisatie op het cultuurparticipatiegedrag in Vlaanderen [Who does what and why? the influence of cultural socialization on the cultural behavior in Flanders]. Mens & Maatschappij 82 (3), 247–271.

Verboord, M., 2005. Long-term effects of literary education on book-reading frequency: an analysis of Dutch student cohorts 1975–1998. Poetics 33, 320–342.

Verboord, M., Van Rees, K., 2003. Do changes in socialization lead to decline in reading level? How parents, literary education, and popular culture affect the level of books read. Poetics 31, 283–300.

Willekens, M., Lievens, J., 2014. Family (and) culture: the effect of cultural capital within the family on the cultural participation of adolescents. Poetics 42, 98–113.

Wippler, R., 1968. Sociale determinanten van het vrijetijdsgedrag [Social Determinants of Leisure Behaviour]. Assen: Van Gorcum [Dissertation University of Groningen].

Yaish, M., Katz-Gerro, T., 2012. Disentangling 'Cultural Capital': the consequences of cultural and economic resources for taste and participation. European Sociological Review 28, 169–185.

Art, Anthropological Aspects of

Nelson Graburn, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Anthropologists assume art is a human universal, but have had to struggle with the cross-cultural identification of art, either by finding common forms and practices, e.g., painting, sculpture, and dance, or by exploring the aesthetic locus in each cultural system. All humans enjoy aesthetic appreciation, but their visual, aural, tactile, olfactory, and sensory preferences vary from place to place and over time. Anthropologists have examined the meaning of arts, in their psychological or psychoanalytic aspects and emotional power, as semiotic media of communication like language and signaling systems, or by examining their sociopolitical agency in guiding human relationships. Arts like other local practices have experienced cross-cultural exchange, colonization, and globalization. They have evolved both as symbols of specific ethnic identities, and as commercial and tourist arts. Crossing boundaries subjects arts to different audiences who make different interpretations, and meanings may be manipulated for power or commerce by mediators such as traders, anthropologists, and critics. As trade and travel erase barriers, artists of formerly separate peoples encounter metropolitan practices such as art school education, art galleries and museums, and the thriving and competitive mainstream art world. Many creatively enter that world while endeavoring to maintain ethnic distinctiveness, so that for instance, British, Inuit, Japanese, Spanish, and Maori arts are circulated and exhibited by the same or similar institutions. The contemporary world of dissolving boundaries, often called postmodernity, has allowed many practices formerly segregated as crafts, often made by women, to enter these same circuits and institutions; metropolitan artists have appropriated features of non-Western arts as their own, as non-Western artists take up metropolitan practices. There is a convergence as artists now engage in ethnographic research, and anthropologists emphasize the poetic (rather than the analytical) power of their works. Art historians emphasize the personal and contextual in art production while anthropologists extend their gaze at the metropolitan and global art world.

The term 'art' has several meanings, all having to do with the performance and products of human skills, associated with aesthetic (rather than functional, sacred, or scientific) qualities, appealing to the human senses. Art can appeal to a single sense, e.g., easel painting, or to multiple senses, e.g., ritual dances (vision, kinesthetic, sound). This article focuses on the visual arts rather than song, music, or dance.

The Term 'Art'

The meaning of 'art,' from the Latin *ars*, *artis* (skill), has changed over time, reflecting technological and social contexts. In English, 'art' is often twinned with 'craft' from the Germanic *kraft* ('skill, cunning' in English, since AD 850). After the Norman Conquest and the Renaissance, 'craft' became associated with manual skills and lower class occupations, and 'art,' from the French *art*, with the conquerors and the Church, hence with learned practices such as mathematics, rhetoric, logic, and grammar (as in Master of Arts).

In Enlightenment Europe, the meaning of 'art' changed again, becoming a synonym for *'beaux arts*,' a set of primarily expressive skills entertaining to the senses but with few practical aims. These arts were the enjoyment of the wealthy, educated classes whose tastes brought into being nonecclesiastical forms such as opera, symphonic music, figurative painting, and choreographed dance (and later ballet). Once detached from 'bread-and-butter' issues, art was justified as a superior expression of the human soul, whose practitioners embodied inner genius and were owed freedom of expression.

Many aesthetic practices of the unschooled members of European populations: folk song, dance, and handicrafts, and of women: quilting, lullabies, and sewing, were not considered arts. Others having both aesthetic and practical ends, architecture (versus building), landscape architecture (versus gardening), or cuisine (versus cooking), bestowed 'artist' status only on trained practitioners, usually men. These distinctions are fading with the postmodernist revaluations of many folk and vernacular practices as 'arts.'

The Challenge for Anthropologists

All these arts are embedded in the historical, social, and cultural structures of the Western world, and anthropologists find it difficult to identify 'art' in other societies. They have tackled the problem in two ways.

First, identify practices and products resembling Western arts, in form (appeal to certain senses) and in relation to context, e.g., religion, hierarchy, ritual. This was easy in the West's encounter with other complex civilizations (often archaeological), but difficult in more egalitarian nonliterate societies studied by early anthropology. In his *Primitive Art* (1955), Boas emphasized the universality of arts as skill-based productions of aesthetically valued forms

> When the technical treatment has attained a certain standard of excellence…such that certain typical forms [later identified as 'styles'] are produced, we call that process art…perfection is essentially an aesthetic judgement…[but] we cannot tell just where the aesthetic attitude sets in.
>
> 1955: 10

He added, "the twofold source of artistic effect, the one based on form alone, the other on ideas associated with form"

(1955: 13). Presumably the former would be aesthetic universals and the latter culturally specific, but Boas admits that other authors claim that all arts iconically represent something.

For Maquet (1979), all societies have aesthetic foci, where excellence is judged by the senses. But 'art' comprises only those productions that circulate within the metropolitan 'art world' (Becker, 1982), and 'primitive art' only exists in the galleries and homes of Western societies. Maquet, following Malraux (1967), posited that Arts are created two ways (1) by destination, made by self-conscious 'artists' for the art world and (2) by metamorphosis, produced either in the past or elsewhere, outside the art world. For instance, eighteenth-century quilts were not considered art until recently, and contemporary Dogon granary doors made for use in West Africa, only become 'art' when imported to the West. The current globalization of commodities makes Maquet's suggestion cogent for almost everything that would be 'art' by Boas' definition.

Meaning, the Content of Art

After the demise of nineteenth-century museum-based anthropology, with its evolutionary and diffusionary overtones, anthropological approaches to art since the 1960s have been concerned with the mechanisms and nature of the messages carried by art. They have drawn upon either psychology or linguistics, i.e., semiotics, reflecting functionalist and structuralist modes of anthropology.

Psychological Approaches

Psychoanalytic approaches treated art, like dreams and myths, as expressions of the subconscious, translated into forms acceptable to the conscious Ego. Most assume the universality of the Oedipus complex and the sublimation of aggression. Jungian approaches posit a universal set of archetypes, whose meanings are personally or locally adapted to solving problems.

Devereaux (1961) said that art expresses taboos in palatable, aestheticized forms. Taboos may be universal, e.g., incest, culturally specific, e.g., sex in puritanical society, or idiosyncratic (neurotic). Arts are admired locally when the taboos are cultural and are more widely when they speak to universal human problems. This 'aesthetics as alibi' approach also underlies the works of Margaret Mead, Bateson, Scheiffelin, and to some extent Kardiner.

Fischer (1961), following Wallace and Barry, proposed that artists sense how their patrons receive satisfactions to needs generated by local sociopsychological contexts: the latent (nonrepresentational) content of art, the 'style,' presents aestheticized versions of social fantasies giving security or pleasure. For instance, symmetrical designs with repetitive simple elements characterize the arts of egalitarian societies, whereas the arts of complex hierarchical societies present enclosed asymmetrical designs integrating disparate elements. He also found that preferences for the curvilinear reflected male security and fantasizing about females. Tielhet (1978) suggested that the equivocal role of women artists in nonliterate societies resulted from men's jealousy of women's reproductive powers, compensating by control of artistic creativity and the hoarding of valuable materials.

Linguistic Approaches

Art, like language, is a communicative form, using a medium that carries a coded message understood by both creator and audience. Forge claimed that for the Sepik peoples, visual art could be read directly, but not reduced to words. In contrast, Munn (1973) showed Australian Walbiri painting consists of a number of iconic but polyvalent elements, combined in sentencelike structures, to make symbolic statements about dream-time ancestors, often suggesting speech.

Levi-Strauss (1966) posits that "art lies halfway between scientific knowledge and mythical and magical thought … midway between design [the scientist builds structures from first principles] and anecdote [the *bricoleur* makes things from the 'debris of past events'] … The aesthetic emotion is the result of this union between the structural order and the order of events." Other structuralist analyses, such as Adams (1969) on Sumbanese textiles, reveal underlying systems of meanings interpreted locally.

For Bloch (1974), relations between aesthetics and emotion may be summarized "You cannot argue with a song!" He makes the analogy: phatic/illocutionary speech is to ritual/art as discursive/propositional speech is to secular behavior and mundane expression. Formal mechanisms, e.g., repetition, symmetries, limited range of elements, fixed sequences, archaic images, give the authorities who control ritual, rhetoric, and the arts power over those enculturated to these patterns.

Sociopolitical Approaches

Social anthropologists have stressed the functions of art in maintaining the social. Turner (1967) showed how Ndembu political rites use normative and naturalistic symbols and colors to structure the event and foster *communitas* among the participants, and Biebuyck (1973) demonstrated how human and animal figurines recall proverbs, functioning to train initiates in the Lega *bwami* cult. Gell (1996) asserts that anthropologists wrongly focused on aesthetics and on relationships between formal style and other cultural characteristics, because (1) 'art' is not always about aesthetics and (2) the expression of a culturally specific aesthetic system is not the function of art objects. Art is created to function in social relations to do something, i.e., it has agency, and art objects can serve as substitutes for social persons. In the West, art serves class and power interests, in gender exclusion (Nochlin, 1971) and the maintenance of social hierarchies (Bourdieu, 1984). Graburn (2005) went beyond Gell's original scope by extending his analytical framework to commercial ethnic arts.

Aesthetics, Ethnoaesthetics, and Universals

Little is understood yet about aesthetic universals. A number of anthropologists have explicated the local, i.e., ethnoaesthetic systems of, e.g., the Yoruba (Thompson, 1973), the Kilenge (Dark, 1974), and the Inuit (Graburn, 1973, 2005). Some have tried inconclusively to test the cross-cultural nature of aesthetic taste by, for instance, showing photographs of art objects both

to their non-Western makers and to Western students of art (Child and Siroto, 1965). However, Graburn (1978, 2001) drawing on the work of Salvador (1976) among the Kuna, was among the first to use a test exhibit of the actual art objects in comparing Native (Inuit) and Euro-American preferences. Other approaches drawing on Wittgenstein's philosophical work and other works on Paleolithic and ethnographic arts (Conkey, 1995; Marshack, 1972) show promise; there are possibilities that underlying aesthetic universals are related to neuro-ocular phosphenes and the structure of the brain itself (Onians, 2008).

Art in the Contemporary World

Ethnic and Tourist Arts

Early anthropological analyses focused on embedded 'authentic, traditional' arts, known as 'primitive' or 'non-Western' arts (Gerbrands, 1957). In *Ethnic and Tourist Arts* (1976), Graburn examined their entry into the 'World Art System,' (Becker, 1982) looking at the primary audience (locals or the external market) and the sources of aesthetic forms (traditional, external models, or novelties). Others have focused on such arts, in Africa and the Americas where traditional 'primitive' arts were found, and in literate civilizations, which have fallen under Western influences (Phillips, 1992). Questioning the historically contingent distinctions hybrid/purebred, Morphy (1992) revealed identical structural and aesthetic patterns underlying functional (locally directed) and commercial (market) Yirrkala bark paintings. Neich (1993) examined a colonial era (1870–1910) form of Maori figurative painting stylistically influenced by *pakeha* (white) culture but expressing Maori cosmology.

Cole (1985), Appadurai (1986), Clifford (1988), and Thomas (1991) have shown the historical depth and socio-cultural complexity of art production in colonial or post-colonial, often touristic, contexts. Dealers, curators, and tour directors may present historical artifacts as 'purebred' symbols of rich heritages, but practically everything in our ethnographic collections was acquired in colonial situations. They may be trade replicas or, like the carved arts of the Northwest Coast Indians of North America, they may have enlarged and proliferated with new tools and economic stimuli. Often, colonists both consumed and imitated contact arts, attempting to assume a new 'ethnic' identity by appropriating the culture of the conquered.

Recent works have focused less on the objects and more on the agency of the personnel of the art world. Ethnohistorical analyses: Cohodas (1997), Batkin (1987) and Phillips (1998), and ethnographic research (Steiner, 1995), reveal the complex activities of artists, intermediaries, and consumers. Artists may consciously work in a variety of styles, aiming at a number of consumer markets (Jules-Rosette, 1984), or they may themselves play the role of intermediaries.

When artists and consumers are culturally or geographically separated, mediating agents assume greater importance (Steiner, 1994). They not only transmit the physical art object from the producer to the consumer but also control the flow of information about its origin, age, meaning, and status as a commodity or a treasure (Appadurai, 1986). They also transmit back the demands and the ideology of the collectors, in terms of price, repeat orders, or information about form, content, color, and materials.

Artists dependent on the cross-cultural market are rarely socialized into its symbolic and aesthetic system, knowing little about the art world, or even 'art' itself. They live in a minefield of rules that they overstep at their peril, unless they assimilate into the 'art world' (Graburn, 1993). The Australian Aborigine watercolorist Namatjira painted successfully in the style of his white mentor, engendering both praise and jealous racism among the Australian public, leading to his downfall (Batty, 1963). Today's Third and Fourth World peoples have increasing exposure to mainstream arts and art materials. Many have attended art schools and assumed the burdens of (Western) art history and the ideological values of originality, artistic freedom, and individual creativity.

The separation of the art circuits for Native arts and mainstream (Euro-American) arts is eroding. Many Native artists work in genres of the metropolitan world art. The contents of their art may not be recognizable as 'ethnic,' possibly because white artists have appropriated so many ethnic motifs. When competing in the world market, Native artists signal their ethnicity through presentation of traditional content, inclusion of stereotypical motifs and color schemes; choice of title or use of a Native name (or 'tribal' designation with a nonethnic name); or the inclusion of written statements expressing a Native point of view (Sturtevant, 1986; Smith, 2009). Inclusions of non-Native contents, motifs, names, and titles may draw criticism or rejection by the market.

Ethnic artists are everywhere pressured to create arts conforming to the image desired by the mainstream market. Even the 'Magiciens de la Terre' (Martin, 1989), which exhibited 'world artists' in Paris, revealed by juxtaposition how the 'ethnic' artists were expected to create something stereotypically traditional, but the metropolitan (white) artists drew their motifs or materials from natural or cultural alterity (otherness) without limitation.

The Metropolitan Art World

Very few anthropologists have tackled the core institutions and activities of the metropolitan art world itself. Berlo (1992), Gerbrands (1957), Graburn (1976), Krech (1999), and Price (1991) have all examined how this art world has classified and examined non-Western arts. Myers, in his work on contemporary Australian aborigine dual purpose (sacred and commercial) arts, has best analyzed the Western art discourse about such arts (Marcus and Myers, 1995).

A number of excellent detailed ethnographies of contextualized contemporary arts in the globalizing world have followed Myers (2002a,b) who examined the complete set of events and transaction from the making of (Central Australian) Pintupi art to its exhibition of the highest realms of the metropolitan world. He showed that in order for the authentic meaning of the art to be transmitted to its final audience, a task undertaken by art historians and art critics for 'art by destination' (Maquet, 1979) produced in and for the metropolitan world, anthropologists have to serve as the mediators for they,

as exemplified by his book, have a thorough knowledge of the artists and the social and material circumstances of production. This is the antithesis of the (1) attitudes of the Cubists and others of a century before who only took inspiration from the visual forms of 'primitive' arts while disdaining the beliefs that lay behind them as well as (2) the transmission of misinformation in the sales pitches of contemporary African arts by 'runners' (middlemen, often Muslims) to Europe and North America (Steiner, 1991, 1994). Morphy confirmed the situation of 'Anthropology as a Mediating Discipline' (Morphy, 2007: 8) in his analysis of the comparable 'arts by metamorphosis' of the North Australian Yolngu. Adams working in Toraja, Indonesia, shows how art serves as 'an active ingredient in identity politics' (2006: 209) within the multicultural globalizing nation of Indonesia, and within the grasp – photographically for the many metropolitan tourists who witness but cannot purchase the complex and colorful rituals around the *tongkonan* funeral houses – and literally for the purchased or stolen – or faked – funeral figures that end up as 'arts of destination' in the galleries and homes of those tourists. Fair's comprehensive posthumous compendium on Alaska Native Arts (2006) is based on more than 20 years of intimate knowledge of the artists in their villages and in urban areas, first as a commercial collector and later a study of anthropology and art history. While nostalgic for a less threatening time of culture contact, the work contains no hesitation or excuses about the often commercial motivations of the Native artists nor purist longing for traditional or 'precontact' materials and techniques, treating them as artists with aesthetic and technical problems to solve trying to make a livelihood, much as did Inupiaq artist Ron Senungetuk, founder (in 1965) of the Alaska Native Arts Centers at the University of Alaska, Fairbanks. Jonaitis and Glass (2010), an anthropological art historian and an art historical anthropologist, have together written a complex account of the transformations of the Northwest Coast totem pole for ethnic, national, commercial, souvenir, and advertising purposes, illustrating the multiple paths to which such iconic forms are subjected over time and across cultural boundaries. These paths are not always linear nor are they always debasing in their evolved commercial forms. Bunten (2006) shows how such transformed arts may, in turn, be consumed by the very people who made the originals in the ongoing evolution of their ethnic identity.

By the 1990s, anthropologists were focusing on artistic production, distribution, consumption, and meaning from sources hardly 'outside' (in Maquet's sense) the metropolis itself. In Bright and Bakewell (1995), the contributing anthropologists were literally "Looking High and Low" (the title) in the greater Los Angeles region, covering such cases as Chicano graffiti and murals (Sanchez-Tranqillo 55–88), Low Riders (Bright 89–123), and the aesthetics and politics of representation of Zuni war gods and Hopi *kachinas* (Tedlock 151–172), showing how hegemonic cultural categories of identity produced classifications of products as crafts versus arts and high and low within the latter. In the last chapter, Marcus (173–198) focuses on the John Paul Getty Museum, and dissents from Bourdieu (1984) by asserting that 'high arts' in the United States are supported by the wealthy 'middlebrow' classes, whereas in France, there still was a 'highbrow' elite life style. Continuing the anthropological examination of control of symbolically laden categories, Plattner (1997) has showed how hierarchies of prestige, stemming from New York, control economic rewards that dictate the life choices of a range of professional artists in St. Louis. The inclusion of Western and World arts into the mainstream of anthropology is signaled by Ong's (2012) analysis of a 2008 exhibition at the New York Guggenheim Museum by the Chinese mainstream artist Cai Guo-Qiang.

The anthropology of art overlaps with visual anthropology, which has come to mean a focus on 'modern' forms of representation such as the media, including photography, film, video, television, and computers in themselves or as media through which to analyze more traditional anthropological subjects such as art, ritual, clothing, and performance (Banks and Morphy, 1997; Pink et al., 2004). The study of art also overlaps with the resurgent study of material culture, with a journal of that name founded in 1996. Emphases may be on vernacular or folk materials (Glassie, 1999) but the major thrust of this new Marxist-oriented purview is on ordinary mass-produced objects and their consumption in metropolitan and increasing postcolonial societies (Miller, 1997, 1998) and particularly on personal possessions and the home (Miller, 2010). Miller and his colleagues emphasize how centrally important 'things' are in everyday life, both as expressions of personal and family identity and as items giving meaning to everyday life.

Anthropology as Art as Anthropology

Norwegian anthropologist Arnd Schneider organized a meeting at the Tate Modern Gallery, London, in 2003 to explore the overlaps and possible convergence of *Contemporary Art and Anthropology* (Schneider and Wright, 2006). This volume and many other articles and collections (Schneider and Wright, 2010) examine the methodological and epistemological convergences of contemporary anthropology and art practice. On the one hand, many artists claim that they are doing ethnography, conducting research and revealing the results of their explorations of the world – usually in visual forms – a topic that Taylor (1994) claimed anthropology had long shunned. On the other, since the 'crisis of representation' in the 1980s, many anthropologists claim that their exploration and presentations are as or more convincing for their aesthetic forms than their scientific rigor. Common are works on "the poetics of…," while others focus on creativity in anthropology. While the contributors are about equally divided between anthropologists of many stripes, and artists broadly defined, key authors include anthropologist George Marcus and cultural studies maverick Lucy Lippard who admits to being an archaeologist manquée.

Contemporary Anthropology and Disciplinarity

The anthropology of art and the discipline of art history have long been growing closer and overlapping, especially the 'new art history' (Baxandall, 1972), which emphasized the personal and material relationships and social networks within which art was created; this focus on particular historical societies and cultures rather than formal aspects resembles anthropological emphases on contextualization. The anthropology of art is in

fruitful dialogue with art history in covering the archaeological, non-Western ethnic and cosmopolitan arts of the world. Morphy and Perkins (2006) retrace the history of the subdiscipline, while Venbrux et al. presents papers from the 2003 American Anthropological Association meeting *Exploring World Art* (2006). Svasek (2007) presents a masterful summary of *Art, Anthropology and Cultural Production* (2007) emphasizing the role of emotions. Art historian Leuthold draws upon anthropological works in framing his collection *Cross-Cultural Issues in Art* (2011), whereas in Westermann's interesting collection *Anthropologies of Art* (2005), only Morphy and Errington represent anthropology among the 14 authors who are mainly art historians.

There is a growing field in which the disciplinary boundaries between anthropology, art history, and cultural studies dissolve in synthetic collections in books and in journals such as *Res, Art in America,* or *Visual Anthropology Review,* reflecting the emerging breakdown of the boundaries between art, craft, and material culture and perhaps of the singularity of 'art' itself.

See also: Art History; Art and Culture, Economics of; Art and Socialisation; Authenticity, Anthropology of; Colonialism, Anthropology of; Cultural Heritage; Diffusion: Anthropological Aspects; Hybridity; Markets: Anthropological Aspects; Markets: Artistic and Cultural; Materiality and Culture; Performance: Anthropological Aspects; Tourism.

Bibliography

Adams, K., 2006. Art as Politics: Re-crafting Identities, Tourism, and Power in Tana Toraja, Indonesia. University of Hawaii Press, Honolulu.

Adams, M., 1969. System and Meaning in East Sumba Textile Design: A Study in Traditional Indonesian Art, vol. 16. Yale University Press Studies in Southeast Asian Art.

Appadurai, A., 1986. Social Life of Things: Commodities in Cross-cultural Perspective. Cambridge University Press.

Banks, M., Morphy, H. (Eds.), 1997. Rethinking Visual Anthropology. Yale University Press, New Haven.

Batkin, J., 1987. Pottery of the Pueblos of New Mexico, 1700–1940. Taylor Museum, Colorado Springs.

Batty, J.D., 1963. Namatjira, Wanderer between Two Worlds. Hodder and Stoughten, Melbourne.

Baxandall, M., 1972. Painting and Experience in Fifteenth Century Italy. Oxford University Press.

Becker, H., 1982. Art Worlds. University of California Press, Berkeley.

Berlo, J.C. (Ed.), 1992. The Early Years of Native American Art History: The Politics of Scholarship and Collecting. University of Washington Press, Seattle.

Biebuyck, Daniel, 1973. Lega Culture; Art, Initiation, and Moral Philosophy Among a Central African People. University of California Press, Berkeley.

Bloch, M., 1974. Symbols, song, dance and features of articulation: is religion an extreme form of traditional authority. European Journal of Sociology 15, 55–98.

Boas, F., 1955. Primitive Art. Dover, New York [originally 1927].

Bourdieu, P., 1984. Distinction: A Social Critique of the Judgment of Taste (Richard Nice, Trans.). Harvard University Press [originally 1979].

Bright, B.J., Bakewell, L., 1995. Looking High and Low: Art and Cultural Identity. U of Arizona Press, Tucson.

Bunten, A., 2006. Commodities of authenticity: when native peoples consume their own "tourist art". In: Venbrux, E., Rose, P.S., Welsch, R.L. (Eds.), Exploring World Art. Waveland Press, Long Grove, IL, pp. 317–336.

Child, I., Siroto, L., 1965. BaKwele and American aesthetic evaluations compared. Ethnology 4, 349–360.

Clifford, J., 1988. The Predicament of Culture: Twentieth Century Ethnography, Literature and Art. Harvard University Press, Cambridge, MA.

Cohodas, M., 1997. Basket Weavers for the California Curio Trade Elizabeth and Louise Hickox. University of Arizona Press, Tucson.

Cole, D., 1985. Captured Heritage: The Scramble for Northwest Coast Artifacts. University of Washington Press, Seattle.

Conkey, M., 1995. Making things meaningful: approaches to the interpretation of ice age imagery of Europe. In: Lavin, I. (Ed.), Meaning in the Visual Arts: Views from the Outside. Institute for Advanced Study, Princeton, NJ, pp. 49–64.

Dark, P., 1974. Kilenge Life and Art. Academy Editions, London.

Devereaux, G., 1961. Art and mythology: a general theory. In: Kaplan, B. (Ed.), Studying Personality Cross-culturally. Harper and Row, New York, pp. 361–403.

Fair, S.W., 2006. In: Blodgett, J. (Ed.), Alaska Native Art: Tradition, Innovation, Continuity. University of Alaska Press, Fairbanks, AK.

Fischer, J., 1961. Art styles as cultural cognitive maps. American Anthropologist 63, 79–93.

Gell, A., 1996. Art and Agency: An Anthropological Theory. Clarendon, Oxford.

Gerbrands, A.A., 1957. Art as an Element of Culture, Especially in Negro Africa. Brill, Leiden.

Glassie, H., 1999. Material Culture. University of Indiana Press, Bloomington.

Graburn, N., 1973. A preliminary analysis of symbolism in Eskimo art and culture. Proceedings of the XL International Congress of Americanists, Rome 2, 165–170.

Graburn, N. (Ed.), 1976. Ethnic and Tourist Arts: Cultural Expressions from the Fourth World. University of California Press, Berkeley.

Graburn, N., 1978. 'I like things to look more different than that stuff did': an experiment in cross-cultural art appreciation. In: Greenhalgh, M., Megaw, J. (Eds.), Art in Society: Studies in Style, Culture and Aesthetics. Duckworth, London, pp. 51–70.

Graburn, N., 1993. Arts of the fourth world: the view from Canada. In: Whitten, D., Whitten, N. (Eds.), Imagery and Creativity; Ethnoaesthetics and Arts Worlds in the Americas. University of Arizona Press, Tucson, pp. 171–204.

Graburn, N., 2001. White evaluation of the quality of inuit sculpture. Inuit Art Quarterly 16 (3), 30–39.

Graburn, N., 2005. From aesthetics to prosthetics and back: materials, performance and consumers in Canadian inuit sculptural arts; or, Alfred Gell in the Canadian Arctic. In: Coquet, M., Derlon, B., Jeudy-Ballini, M. (Eds.), Les Cultures à l'oeuvre – Rencontres en Art [Cultures at Work – Encounters in Art]. Biro Editeur, Paris, pp. 47–62.

Jonaitis, A., Glass, A., 2010. The Totem Pole: An Intercultural History. University of Washington Press, Seattle.

Jules-Rosette, B., 1984. The Messages of Tourist Art: An African Semiotic System in Comparative Perspective. Plenum Press, New York.

Kopytoff, I., 1986. The cultural biography of things: commoditization as a process. In: Appadurai, A. (Ed.), The Social Life of Things: Commodities in Cultural Perspective. Cambridge University Press, Cambridge, UK, pp. 64–91.

Krech, S., 1999. Collecting Native America. Smithsonian Press, Washington, DC.

Leuthold, S., 2011. Cross-Cultural Issues in Art: Frames for Understanding. Routledge, New York.

Levi-Strauss, C., 1966. The science of the concrete. In: The Savage Mind. University of Chicago Press.

Malraux, A., 1967. Museum without Walls; the Voices of Silence. Secker and Warburg, London.

Maquet, J., 1979. Introduction to Aesthetic Anthropology, second ed. Undena, Malibu, CA.

Marcus, G., Myers, F. (Eds.), 1995. The Traffic in Culture: Refiguring Art and Anthropology. University of California Press, Berkeley.

Marshack, A., 1972. The Roots of Civilization: The Cognitive Beginning of Man's First Art, Symbol and Notation. McGraw-Hill, New York.

Martin, J.-M. (Ed.), 1989. Magiciens de la Terre. Editions du Centre Pompidou, Paris.

Miller, D., 2010. Stuff. Polity Press, Cambridge.

Miller, D. (Ed.), 1997. Material Culture and Mass Consumption. Wiley, New York.

Miller, D. (Ed.), 1998. Material Culture: Why Some Things Matter. University College London Press.

Morphy, H., 1992. From dull to brilliant the aesthetics of spiritual power among the Yolngu. In: Coote, J., Shelton, A. (Eds.), Anthropology, Art and Aesthetics. Clarendon Press, Oxford, pp. 181–208.

Morphy, H., 2007. Becoming Art: Exploring Cross-Cultural Categories. Berg, Oxford.

Morphy, H., Perkins, M. (Eds.), 2006. The Anthropology of Art: A Reader. Blackwell, Oxford.

Munn, N., 1973. Walbiri Iconography: Graphic Representation and Cultural Symbolism in a Central Australian Society. Cornell University Press, Ithaca, NY.

Myers, F., 2002. Painting Culture: The Making of an Aboriginal High Art (Objects/Histories). Duke University Press.

Myers, F. (Ed.), 2002. The Empire of Things: Regimes of Value and Material Culture. School of American Research, Santa Fe, NM.

Neich, R., 1993. Painted Histories: Early Maori Figurative Painting. Auckland University Press.

Nochlin, Linda, January 1971. Why have there been no great women artists? ARTnews, 22–39, 67–71.
Ong, A., 2012. "What Marco Polo Forgot": contemporary chinese art reconfigures the global. Current Anthropology 53 (4), 471–494.
Onians, J., 2008. Neuroarthistory: From Aristotle and Pliny to Baxandall and Zeki. Yale University Press.
Phillips, H., 1992. The Integrative Art of Modern Thailand. University of Washington Press, Seattle.
Phillips, R., 1998. Trading Identities: The Souvenir in Native North American Art from the Northeast, 1700–1900. University of Washington Press, Seattle.
Phillips, R., Steiner, C. (Eds.), 1996. Unpacking Culture: Art and Commodity in Colonial and Post-Colonial Worlds. University of California Press, Berkeley.
Pink, S., Kurti, L., Afonso, A.I. (Eds.), 2004. Working Images: Visual Research and Representation in Ethnography. Routledge, London.
Plattner, S., 1997. High Art Down Home: An Economic Ethnography of a Local Art Market. University of Chicago Press.
Price, S., 1991. Primitive Art in Civilized Places. University of Chicago Press.
Salvador, M.-L., 1976. The clothing arts of the Cuna of San Blas. In: Graburn, N.H.H. (Ed.), Ethnic and Tourist Arts. University of California Press, Berkeley, pp. 165–182.
Schneider, A., Wright, C. (Eds.), 2006. Contemporary Art and Anthropology. Berg, Oxford.
Schneider, A., Wright, C. (Eds.), 2010. Between Art and Anthropology: Contemporary Ethnographic Practice. Berg, Oxford.
Smith, P.C., 2009. Everything You Know about Indians Is Wrong. University of Minnesota Press.
Steiner, C., 1991. In and Out of Africa (Video 60 mins). University of California Extension, Berkeley [co-produced with Barbash, I., Gabai, B., Taylor, L.].
Steiner, C., 1994. African Art in Transit. Cambridge University Press.
Steiner, C., 1995. The art of the trade: on the creation of value and authenticity in the african art market. In: Marcus, G.E., Myers, F.R. (Eds.), The Traffic in Culture: Re Configuring Art and Anthropology. University of California Press, Berkeley, pp. 151–165.
Sturtevant, W., 1986. The meanings of native American art. In: Wade, E.L. (Ed.), The Arts of the North American Indians: Natives Traditions in Evolution. Udson Hills, New York.
Svasek, M., 2007. Anthropology, Art and Cultural Production. Pluto, London.
Taylor, L. (Ed.), 1994. Visualizing Theory. Routledge, New York.
Tielhet, J., 1978. The equivocal role of women artists in non-literate societies. Heresies 1, 96–102.
Thomas, N., 1991. Entangled Objects: Exchange, Material Culture and Colonialism in the Pacific. Harvard University Press, Cambridge.
Thompson, R.F., 1973. Yoruba artistic criticism. In: D'Azevedo, W. (Ed.), The Traditional Artist in African Societies. Indiana University Press, Bloomington, pp. 19–61.
Turner, V., 1967. The Forest of Symbols; Aspects of Ndembu Ritual. Cornell University Press, Ithaca, NY.
Venbrux, E., Rose, P.S., Welsch, R.L. (Eds.), 2006. Exploring World Art. Waveland, Long Grove, IL.
Wade, E. (Ed.), 1986. Native American Arts in Transition. Hudson Hills, New York.
Westermann, M. (Ed.), 2005. Anthropologies of Art. Yale U Press, Williamstown, MA. New Haven, for S and F Clark Art Institute.

Art History

Horst Bredekamp, Humboldt University, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The main task of art history is to collect, preserve, classify, appreciate, and mediate historic artworks of all periods. It deals with materialized forms of historical interventions that shape the concepts of time and thereby continuously change the character of the present. Art history is continuously in conflict with modernism and defines it in the same moment. 'Globalization' can count on a prehistory that goes back before the age of imperialism, but at the same time the traditional concept of art history as 'Image history' (Bildgeschichte) is in danger, caused by the separation between art history and visual studies.

The Character of Time

In the year 1983, the artist Bogomir Ecker had the idea to create an artwork that would not only last for the centuries to come, but would be realized over the inconceivable time span of 500 years; he hence began to search for a place that promised to survive so far into the future. His search was by no means trivial: his artwork, the so-called *Tropfsteinmaschine* (*Dripstone Machine*), required an incessant supply of dripping rainwater for its apparatus to create a sculpture that, through mineral sedimentation, would grow 1 cm every 100 years (Ecker, 1999). Not only the place but also the maintenance of the apparatus had to be warranted for the next 500 years. The first building that came to the artist's mind was the house of parliament in Bonn. However, in 1987, he decided by intuition that even the future of this center of the West German state was too uncertain for it to serve as the site for his installation. History proved him correct, as the Reunification put an end to the 'Bonn Republik'. In his search for a lasting establishment, the artist ultimately decided against a political building and instead chose a museum, the safekeeper of art: With his choice of the Hamburger Kunsthalle, he incorporated his generation-spanning production space into the institution that to him held the highest promise of surviving all future changes, even over the next centuries (**Figure 1**).

This example puts in a nutshell the main tasks of art history, shared with archaeology, namely, to collect, preserve, classify, interpret, and mediate historic artworks of all periods. In doing so, art historians instantiate a concept of time that counteracts the passing of the present. Already the decision to allocate or withdraw means for this expansive task touches a community's identity at the heart. Also oral, textual, and musical tradition provides a space for reflection; but the sphere of visual art confronts the viewers and their projections, which stem from the present, in the unique form of a counterreality that can be scrutinized in its material presence. Therefore, the concept of time developed and maintained by art historians carries a profound social dimension.

Museums at the same time reveal the paradox endeavor of art history. Since the discipline was systematically included in an academic curriculum for the first time in 1810 at the University of Berlin (Bredekamp and Labuda, 2010), art history, through its museums, has left an imprint on modern culture only comparable to the spread of hospitals through the progress of medicine and the erection of temples of transportation, such as train stations and airports, through engineering science. Since the opening of the Altes Museum in Berlin in 1830 (**Figure 2**) – the first autonomous building of modernity constructed for this specific purpose – the permeation of the history of art by art history has exerted an eminent influence on metropolitan life. Through museums, art history contributes to the urbanity even of small- and medium-sized towns, but also of metropoles such as Berlin, London, Madrid, Paris, Mexico, New York, and Beijing. Whenever newly developed centers wish to document their supraregional relevance, these claims find expression in the erection of museums: recent examples are Abu Dhabi (**Figure 3**) and Hong Kong. As a hard factor contributing to the attractiveness of human communities, museums in their paradoxical construction have remained unchallenged over almost 200 years; and even

Figure 1 Bogomir Ecker, Dripstone Machine, Installation, 1996, Kunsthalle, Hamburg.

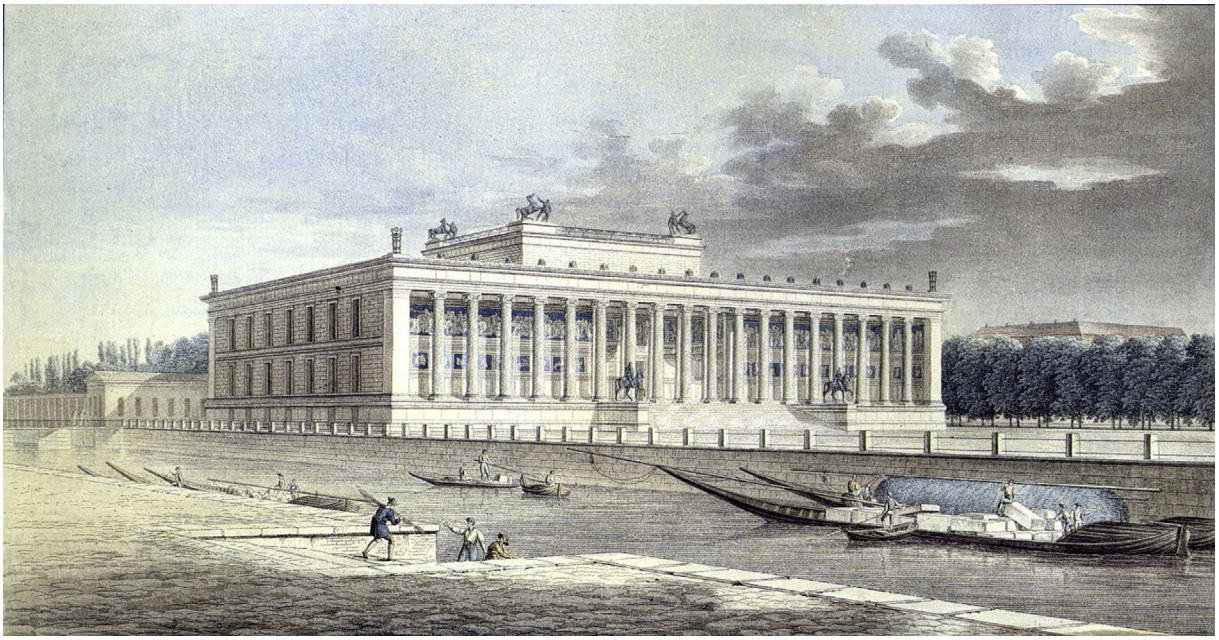

Figure 2 Friedrich Alexander Thiele, View of Karl Friedrich Schinkel's museum of Berlin, colored etching, 1830.

Figure 3 Frank Gehry, Model of the museums in Abu Dhabi, 2007.

presently, the continuing increase of visitor streams conveys no trace of fading appeal. The 100 most visited art museum range between the Palace Museum in Beijing (*c.* 12 million visitors per annum) and the Istanbul Modern (*c.* 666 000 visitors). The 100 most frequented exhibitions of the year 2012 range between *c.* 10 500 (Tokyo, Metropolitan Art Museum) and *c.* 160 000 (Stockholm, Moderna Museet) per day (Source: Visitor Figures, 2012). The numbers for Europe can be checked regularly through the *European Group on Museum Statistics* (www.egmus.eu/en/statistics/).

As a warrantor of the diachronic development of art, however, art history stands in conflict with a modernity that it brought forth itself. Modernity calls for acceleration, presentism, and creative destruction, which in the Futurist Manifesto of 1909 comprised the destruction of all museums as a prerequisite for further development. The fact that art history, by having integrated Futurism into the historical conditions of the Modern Canon, bears this conflict within itself, is evidence for the paradoxical disposition of the discipline in the modern world.

Space: Globalization

This conflict affects all aspects of art history's presence in our days. Because art history has a century-old history, it has to reinvent itself by providing new answers in conflicting fields, of which the problems of globalization are perhaps the most pressing.

The criteria for form analysis, comparison, and the history of form development and the search for iconological semantics, which has been differentiated into a variety of within themselves highly complex stylistic periods, have been developed and refined since the age of the so-called 'father of art history', Giorgio Vasari, mostly with regard to European art; however, over the centuries, the criteria have been tested and improved in most diverse ways.

From this, two possibilities arise for the future. According to the first, these methods hold a similarly objective status as Euclid's geometrical laws, able not only to transcend time but also to operate transculturally. This precludes their continuous discussion and repeated adaptation to new subject matters. According to the second possibility, these methods must be defined as historically determined skills that will be transformed in an undetermined future within a new framework, in which they will unfold their meaning in a different way. The future beneficiary of such a process would then have to compensate possible losses of the historical methods.

This fundamental question is related to the European origin of the discipline. The Western industrial countries have brought forth by far the most art historians. However, from the beginning, it was considered a sign of quality not to confine art-historical studies to the art of one's own culture but instead to explore the congruent as well as antagonistic intertwinings between different cultures. Regarding the artifacts of the so-called 'Third World', the colonial aspect of this phenomenon is not far to seek; but the interest in foreign cultures was, albeit less frequently, also motivated by curiosity and high regard for other domains. For this reason, those museums that stem from the tradition of the *Kunstkammer* had to deal with the internal conflict that their collections were predominantly the products of colonial overthrows, but also of a non- or anticolonial interest and admiration (example of a scholar, Nicolas-Claude Fabri de Peiresc: Miller, 2012; feathers as an example of shaped materials: Russo, 2002; Penacho, 2012; in general: Götterbilder, 2012). In this lies the second open question that the discipline is confronted with today.

An early, liberal answer to the problem is Franz Kugler's magnum opus of liberal globality, the *Handbuch der Kunstgeschichte* (*Handbook of Art History*), published in 1842 (Kugler, 1842; Franz Theodor Kugler, 2010). Written in the climate of Johann Gottfried Herder's definition of the *Weltbürger* (world citizen), the text tries to unite all artworks from the Stone Age to the present in a non-hierarchical narrative. Through the deepening of nationalisms in the second half of the nineteenth and above all in the twentieth century, this global orientation was lost for the most part. A fortiori, the non-colonial interest in foreign art today has become a driving impulse in most of the contemporary debates worldwide (Crossing Cultures, 2009).

It was a programmatic decision of the founders to locate the *Getty Center* as one of the most ambitious research institutes for art history and material culture worldwide, not on the East coast of the United States, but instead in Los Angeles, to make strong its pacific perspective (**Figure 4**). The process of transcending borders regarding materials and cultures is spurred by the umbrella organization *Comité International d'Histoire de l'Art* (CIHA). The CIHA was founded after the Second World War to counteract the mental and material destructions of fascism and National Socialism. Until recently, all world congresses, which are organized by the CIHA every 4 years since the first Vienna congress in 1973, had taken place in Europe and the United States. In 2008, the first congress to take place in the Pacific Rim was held in Melbourne, as a statement regarding the spread of art history throughout the world. While 50 nations convened there at the time, the CIHA-congress in Nuremberg 2012 brought together art historians from unprecedented 82 nations. The congress of the year 2016 will take place in Beijing. Further, no less telling symbols of this global orientation are two encyclopedias: the *Dictionary of Art* (1996) and the German-language *Allgemeines Künstlerlexikon* (General Artist Lexicon, 1991ss.), dedicated to the artists of all times and peoples (the online-edition already contains over 1 million entries (http://www.degruyter.com/view/db/akl)).

In view of this comprehensive documentation of artists and artworks of all times, the methodological globalization of art history lies not merely in adding new fields or ordering existing ones but in developing perspectives that emphasize tensions and differences and thereby produce criteria for assessing the respective strengths and weaknesses. The study of perspective in different cultures can function as an example for this (Is Art History Global? 2007; Kritische Berichte, 2012) as well as the reformulation of crossing cultures in larger areas like the Mediterranean (Wolf, 2009).

An attempt was made to increase this problem awareness also institutionally by internationalizing the directorship of major art-historical institutions. Hence, in the year 2013, an Australian art historian is the director of the Norwegian museum in Trondheim, while a Norwegian art historian directs the National Museum in Mombasa; a German ethnologist and art historian is the head of the Victoria and Albert Museum, while – the perhaps most historic example – a curator educated in Herder's sense, who was born in Nigeria, worked in Africa and the USA, and curated the documenta 11 in Kassel in 2002, since 2011 heads the *Haus der Kunst* in Munich – a museum that was inaugurated by Adolf Hitler in 1936. This tendency was pushed to an almost Dadaistic extreme at the Venice Biennial in 2013, where France and Germany exchanged their pavilions and Germany featured foreign artists, among them the Chinese Ai Weiwei (La Biennale di Venezia, 2013). Taking all of these answers to the question of

Figure 4 Richard Meyer, Getty Center, Los Angeles, 1991–1997.

globalization, it appears that the universal self-definition of the discipline has been sealed also for the future.

Image: Art History as Bildgeschichte (Image History)

One of the conditions of this process is the reactivation of a wide definition of 'art' as developed mainly in the German-speaking realm of Austria, Switzerland, Germany, and several eastern European countries throughout the nineteenth and twentieth century until the fatal year of 1933. 'Art' was not understood as an autonomous, self-obliging form of creativity, but in the sense of the Greek *techné* and the Roman *ars*. This meant that not only the products of high art but also all forms of applied arts lay in the responsibility of art historians. The museums of art and design have lived up to this standard worldwide.

Important representatives of this direction were Franz Kugler, Gottfried Semper, Alois Riegl, and Aby Warburg. Namely Riegl (1985) and Warburg (Warburg, 1998, 1999) set a still relevant example for the ability of art historians to overcome the inner paradox of the discipline, i.e., to stand in the present and simultaneously secure the timeless quality of the artworks, to be devoted to every kind of form (*Gestaltung*) by emphasizing the unique character of the visual arts while also realizing the finesse of the applied arts.

Within this framework, art historians appropriated new media such as photography and film not only as parts of a new corpus of visuality but also as methodological enrichments. The early theory of film by Rudolf Arnheim (1931) became a landmark in film studies that were founded through the continuous activity of Erwin Panofsky (Bredekamp, 2006). What is more, all forms of video (the earliest attempt was: Film als Film, 1977) and digital media – such as web art – were quickly analyzed and adopted. The medias of video and digital art are included in many museums, and the *Zentrum für Kunst und Medientechnologie* in Karlsruhe (1996) has for the first time devoted an entire complex to these new media, consisting of both an art museum and a laboratory. As the response to a media revolution encompassing the world but remaining ephemeral in its production, museums as the primary seats of art history remain the fixed points of a locally accessible but materially permanent realm of experience, one that also incorporates the products of the art-technological revolution.

Against common intuition, art history was one of the first humanities to support the use of computers as an auxiliary means for archiving and searching (Kohle, 2013). One of the oldest and most successful digital programs within the humanities may be the *Census of Antique Works of Art and Architecture Known to the Renaissance*, whose presentation in Pisa in 1983 and in London in 1992 set the course for the discipline as a whole (Nesselrath, 1993). The search engines of the image archive *Foto Marburg* (www.fotomarburg.de), the *Prometheus Image Archive* (www.prometheus-bildarchiv.de), *Digitale Diathek* (www.digitale-diathek.net), and the Andrew F. Mellon Foundation's ARTstor (www.artstor.org) have achieved in the field of images what the Getty Research Portal of the Getty Research Institute has developed for digital research in art-historical literature (www.getty.edu/research/tools/portal/index.html) and what other institutions have created for the documentation of national art (e.g., in the Netherlands: www.rkd.nl) or national collections (e.g., France: http://www.culture.gouv.fr/documentation/joconde/fr/pres.htm). Nearly every larger museum has in recent years made at least parts of its collection accessible online as digital reproductions, and the Department of Prints and Drawings of the British Museum set an example regarding the free use of images in the sense of open access (www.britishmuseum.org/research/collection_online/search).

Furthermore, the products of everyday culture have been and continue to be a systematic subject of this type of art history. Already at the beginning of the twentieth century, Warburg suggested the use of the term *Bildwissenschaft* (image science) to better describe the subject matter of the discipline.

After National Socialism severed this tradition, it has been successfully reestablished since the 1960s (landmarks: Freedberg, 1989; Belting, 2001; Didi-Huberman, 2002). If today the image worlds of the natural sciences and medicine (Das Technische Bild, 2008; The Technical Image, 2013) and most notably political iconography (Handbuch der Politischen Ikonographie, 2011) belong to the natural areas of art-historical analysis, then this is an effect of the renaissance of a once dominant tradition.

A part of this revival is also the close relationship between art history, philosophy, and linguistics as it was developed at the cultural historical library *Kulturwissenschaftliche Bibliothek Warburg*, the work of whose members gave some of the most important impulses that ever came from art history. Attempts have been made to resuscitate this tradition as well, which was mostly destroyed after the library's emigration to London in 1933. If the present period has been placed under the terms *pictorial turn* (Mitchell, 1992) and *iconic turn* (Boehm, 1994), this by no means implies playing off images against language. Instead, what is at stake is to radically question the interpretive predominance of language championed by the *linguistic turn*. Both *turns* are related to the conviction that in a highly technologically advanced civilization in which communication takes place for the most part visually it becomes all the more necessary to develop a philosophy of the image that is equal to the present task.

The collaborative effort between *Embodiment Theory* (Krois, 2011) and *Picture Act Theory* (Theorie des Bildakts, Bredekamp, 2010) has tried to answer these demands. Also in this endeavor, artifacts have been examined in their entire breadth, from automatons to diagrams.

Material, Language, Visual Studies

However, at the same time, these achievements are threatened by the breaking away of visual studies from art history in the English-speaking realm (Elkins, 2003; Bredekamp, 2003; Art History and Visual Studies in Europe, 2012). This has had the effect that Anglo-Saxon art historians tend to feel relieved of the pressure to face the entire breadth of the discipline as it was conceived by Riegl and Warburg. This contradicts their own tradition that tried to embrace the whole frame, as developed from most different perspectives by scholars like George Kubler (1962), David Freedberg (1989), Barbara Stafford (1991), James Elkins (1999), Whitney Davis (2011), and Neil MacGregor (2011). Although representatives of visual studies have succeeded in opening up everyday visual culture to scholarly research in a remarkable way, they also have to account for dire losses. By distancing themselves from art-historical methods, the sensitivity for the value of each individual form becomes less developed.

This process creates the third profound problem, strengthened by developments in language policy. While up to the year 1933 the dominant language of art history was German, the CIHA committee chose English, French, Italian, and German as the official congress languages. In 2004 Spanish was added. So far, each language, independent of nationality, has brought forth specific modes of thought that defy immediate translation. Conjointly, the five internationally defined languages, to which, depending on context, must be added Arabic, Chinese, Japanese, Netherlandish, Russian, Portuguese, and all other languages in which the respective artworks have been reflected, have generated methodological diversity and a fruitful wealth of perspectives. The globalization of science in general, however, has also caused art history to increasingly sacrifice its plurality of languages to the predominant use of English. Despite all the advantages of facilitated mutual understanding, a danger is involved that threatens art history at its core; in this lies the major problem and potential of art history today.

With the global transition toward English as the universal language of art history, the danger of a methodological reduction of the discipline lights up and could cause its marginalization. Art history will continue to be responsible for the high arts inalienably and for indeterminable time to come. With regard to the deeper-lying battlefields of coming to terms with a visually dominated contemporary culture, however, art history could be pushed into a minor role, or even worse, cease to play any role at all. Its future disciplinary as well as social reputation depends on the solution of this conflict.

Bibliography

Anderson, Jaynie (Ed.), 2009. Crossing Cultures: Conflict, Migration and Convergence. The Proceedings of the 32nd International Congress in the History of Art. Miegunyah Press, Carlton.

Arnheim, Rudolf, 1931. Film als Kunst. Rowohlt, Berlin.

Art History and Visual Studies in Europe, 2012. In: Rampley, Matthew, Lenain u. a., Thierry (Eds.), Transnational Discourses and National Frameworks. Brill, Leiden and Boston.

Allgemeines Künstlerlexikon. Die Bildenden Künstler aller Zeiten und Völker, vol. 1ss. K. G. Saur / De Gruyter, Leipzig 1991ss.

Belting, Hans, 2001. Bild-Anthropologie. Entwürfe für eine Bildwissenschaft. Fink Verlag, München (Engl. transl.: An Anthropology of Images: Picture, Medium, Body, Princeton University Press, Princeton 2011.).

Boehm, Gottfried, 1994. Die Wiederkehr der Bilder. In: Was ist ein Bild? Fink Verlag, Munich, pp. 11–38.

Bredekamp, Horst, 2003. A neglected tradition? Art history as bildwissenschaft. Critical Inquiry 29 (3), 418–428.

Bredekamp, Horst, 2006. "On movies". Erwin Panofsky zwischen Rudolf Arnheim und Walter Benjamin. In: Koebner, Thomas, Meder, Thomas (Eds.), Bildtheorie und Film. Verbindung mit Fabienne Liptay, München, pp. 239–252. Edition Text & Kritik.

Bredekamp, Horst, 2010. Theorie des Bildakts. Frankfurter Adorno-Vorlesungen 2007. Suhrkamp Verlag, Berlin.

Bredekamp, Horst, Labuda, Adam, 2010. Kunstgeschichte, Universität, Museum und die Mitte Berlins 1810–1873. In: Bredekamp, Horst, Labuda, Adam S. (Eds.), In der Mitte Berlins. 200 Jahre Kunstgeschichte an der Humboldt-Universität. Gebr. Mann Verlag, Berlin, pp. 25–54.

Bredekamp, Horst, Schneider, Birgit, Dünkel, Vera (Eds.), 2013. The Technical Image. A History of Styles in Scientific Imagery. Bard College, New York.

Das Technische Bild, 2008. In: Bredekamp, Horst, Schneider, Birgit, Dünkel, Vera (Eds.), Kompendium zu einer Stilgeschichte wissenschaftlicher Bilder. Akademie Verlag, Berlin.

Davis, Whitney, 2011. A General Theory of Visual Culture. Princeton University Press, Princeton, NJ.

Didi-Huberman, Georges, 2002. L'Image survivante. Histoire de l'Art et Temps des Fantômes selon Aby Warburg, Paris.

Ecker, Bogomir, 1999. Die Tropfsteinmaschine. 1996–2496, Hamburger Kunsthalle. Hatje Cantz, Hamburg.

Elkins, James. (Ed.). Is Art History Global? Routledge, New York and Milton Park.
Elkins, James, 1999. The Domain of Images. Cornell University Press, Ithaca.
Elkins, James, 2003. Visual Studies. A Skeptical Introduction. Routledge, New York and London.
Freedberg, David, 1989. The Power of Images. Studies in the History and Theory of Response. University of Chicago Press, Chicago and London.
Fleckner, Uwe, Warnke, Martin, Ziegler, Hendrik (Eds.), 2011. Handbuch der Politischen Ikonographie, vols. I and II. Beck Verlag, München.
Götzenbilder und Götzendiener in der Frühen Neuzeit – Europas Blick auf fremde Religionen, Exhibition catalogue. In: Effinger, Maria (Ed.), Cornelia Logeann and Ulrich Pfisterer. Universitätsverlag Winter, Heidelberg.
Hein, Birgit, Herzogenrath, Wulf (Eds.), 1977. Film als Film, 1910 bis heute. Hatje Cantz, Stuttgart.
Kohle, Hubertus, 2013. Kunstgeschichte Digital. Reimer Verlag, Berlin.
Kritische Berichte, Jg 40 (2), 2012. Universalität der Kunstgeschichte?
Krois, John M., 2011. Bildkörper und Körperschema. Akademie Verlag, Berlin.
Kubler, Georges, 1962. The Shape of Time: Remarks on the History of Things. Yale University Press, New Haven and London.
Kugler, Franz T., 1842. Handbuch der Kunstgeschichte, 2 vols. Ebner & Seubert, Stuttgart.
Kugler, Franz T., 2010. In: Espagne, Michel, Savoy, Bénédicte, Trautmann-Walter, Céline (Eds.), Deutscher Kunsthistoriker und Berliner Dichter. Akademie Verlag, Berlin.
MacGregor, Neil, 2011. A History of the World in 100 Objects. Penguin, New York.
Miller, Peter N., 2012. Peiresc's Orient. Antiquarianism as cultural history in the seventeenth century. Farnham. Ashgate Variorum.
Mitchell, William J.T., March 1992. The pictorial turn. In: Artforum, pp. 89–94.
Nesselrath, Arnold, 1993. The census of antique works of art and architecture known to the renaissance. In: Boardman (Hg.), John (Ed.), Data and Image Processing in Classical Archaeology, Archeologia e Calcolatori, vol. IV, pp. 23–27, pp. 237–241.
Penacho, 2012. Pracht & Passion. Der altmexikanische Federkopfschmuck in Wien, Exhibition catalogue. Museum für Völkerkunde, Vienna.
Rampley, Matthew, 2012. Bildwissenschaft: theories of the image in German-language scholarship. In: Rampley, Matthew, Lenain u. a., Thierry (Eds.), Art History and Visual Studies in Europe. Transnational Discourses and National Frameworks. Brill, Leiden and Boston, pp. 119–134.
Riegl, Alois, 1985. Spätrömische Kunstindustrie, Wien 1901 (Late Roman Art Industry (Rolf Winkes, Trans.)). Bretschneider, Rome.
Russo, Alessandra, 2002. Plumes of sacrifice. Transformations in sixteenth-century Mexican feather art. Res. Anthropology and Aesthetics, vol. 42, pp. 226–250.
Stafford, Barbara, 1991. Body Criticism. Imaging the Unseen in Enlightenment Art and Medicine. MIT Press, Cambridge/MA and London.
Turner, Jane (Ed.), 1996. The Dictionary of Art, 36 vols. Macmillan Publishers, London since 2003: Oxford University Press.
Visitor Figures, 2012. The Art Newspaper (245), 15–29. April 2013.
di Venezia, La Biennale, 2013. In: Pavillon, Deutscher, Ai Weiwei, Begleitpublikation, Karmakar, Romuald, Mofokeng, Santu, Singh, Dayanita (Eds.), Susanne Gaensheimer. Gestalten, Berlin.
Warburg, Aby M., 1998. In: Bredekamp, Horst, Diers, Michael (Eds.), Die Erneuerung der heidnischen Antike. Kulturwissenschaftliche Beiträge zur Geschichte der europäischen Renaissance (Gesammelte Schriften. Studienausgabe, Erste Abteilung, vols. I, 1 and I, 2. Akademie Verlag, Berlin.
Warburg, Aby, 1999. The Renewal of Pagan Antiquity. Getty Research Institute for the History of Art and Humanities, Los Angeles.
Wolf, Gerhard, 2009. In: Fluid Borders, Hybrid Objects. Mediterranean Art Histories 500–1500, Questions of Method and Terminology. Crossing cultures, pp. 134–137 and the following section 138–178.

Artificial and Natural Computation

Achim G Hoffmann, The University of New South Wales, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the notion of computation in natural as well as in artificial systems. Particular emphasis lies on artificial computation models that are derived from natural systems that perform computation. The article sketches the main direction of research in the past decades on modeling computational processes in natural systems. This concerns on the one hand computation in nervous systems, such as the human brain and their artificial models. On the other hand there is the process of biological evolution, of which artificial computer models proved very useful as a general approach to solving difficult design and optimization problems.

In the 1930s, a precise mathematical notion of computation had been developed, which captures everything that can be computed in our modern computers. What exactly natural computation is and how it is conducted in natural systems, such as the human brain, is less well understood. However, the mathematical notion of computation proved a convenient and very useful form for describing information processing in natural systems.

Wiener (1948) introduced the discipline of cybernetics, which was concerned with control systems, including natural systems, such as the human body, in which feedback loops control the secretion of glands.

In the following decades, computational models of biological systems were developed and used to address a variety of difficult problems for which conventional solutions were not readily available. The research on models of natural computation has focused mainly on models of the human brain and its neurons on the one hand and on biological evolutionary processes on the other hand.

Another type of computation in 'natural systems' was shown by Adleman (1994), when he demonstrated the use of DNA molecules as a means of solving computational problems. The work showed the potential for building computers on the basis of molecular biology.

This article focuses on the main research efforts in the past decades and discusses artificial models of natural computation in neural systems. Furthermore, the use of models of biological evolution for solving computational problems will be sketched.

The article is organized as follows: first, the mathematical notion of computation is discussed. This is followed by a description of computation in those natural systems, for which most of the artificial models of natural computation have been developed: the human brain and evolutionary processes in biology. The relationship between the natural computing processes and the artificial models are discussed thereafter.

The Mathematical Notion of Computation

In the 1930s, it was an important problem in mathematics to delineate those function definitions for which the function value can be computed for all possible arguments from those functions for which the function value may not be effectively computable. (For functions that contain themselves either directly or indirectly in their definition, it is not straightforward to decide whether a value can effectively be determined for all possible arguments.)

Turing (1937) developed a model for what a mathematician can possibly do, when computing the value of a function according to clearly defined rules, and he argued that no human mathematician could possibly follow other clearly defined rules that are not covered by his model. This model became known as the Turing machine. Other mathematicians developed alternative models for how to effectively compute mathematical functions. However, it was proved that with all these different models, exactly the same class of functions was effectively computable.

The Turing Machine

A Turing machine has a finite number of internal states. A tape on which the Turing machine can read and write symbols from a finite alphabet via its read/write head is unlimited on both sides. The Turing machine starts its operation in a particular internal state and terminates its operation in another particular state. The read/write head will focus on the first square on the tape, where the input string (in consecutive squares) for computation is provided. The Turing machine will execute a sequence of operations. Each of these operations depends on the current symbol under the read/write head and its internal state. The operation may erase the old symbol on the tape in the square under the read/write head and replace it by another symbol. Furthermore, the Turing machine may change its internal state and move (optionally) its read/write head one square to the left or to the right. The entire computation terminates when the Turing machine reaches the special internal termination state. The result of the computation is found at a prespecified location on the tape relative to the position of the read/write head. Which operation is executed depends on the internal state and the current symbol under the read/write head and is specified in the corresponding Turing table. An example is shown in Figure 1.

An important concept is the universal Turing machine, which is a special Turing machine that is capable of simulating any other Turing machine, if a description of that machine is provided as input on the tape. The existence of the universal Turing machine demonstrates the relativity of the actual device on which the computation is performed. Generally speaking,

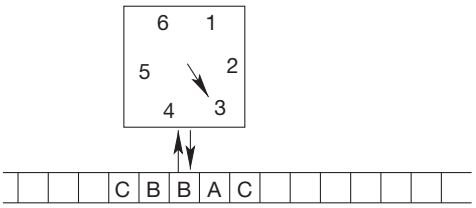

State	Symbol		State	Symbol	Head
1	A	→	3	B	R
1	B	→	2	C	L
1	C	→	6	C	L
...	...		...	...	...
...	...		...	...	...
6	D	→	3	A	R

Figure 1 The schema of a Turing machine (left) and the corresponding Turing table defining the operations of the Turing machine (right).

a computing device can behave in quite different ways depending on what data are provided to it. In the case of a computationally universal computing device, all other computing devices can be simulated, if appropriate data, i.e., an appropriate program, is provided.

Today's computers are generally said to be computationally universal. Hence, natural computation can, in principle, always be simulated on an artificial computation device. In practice, however, enough is not known about what 'programs' are actually 'executed' in natural computation processes.

Artificial and Natural Computation in Neural Systems

Artificial computation in artificial neural networks is a metaphorical model of natural computation in biological neural systems, such as the human brain.

Usually, large sets of computing units, connected in a network-like manner, are used to simulate the way natural neural networks in man and animal perform information processing. Often, the involved computing units are all of the same type and compute some sort of linear threshold function, i.e., they sum up the weighted input signals and compare the sum against a threshold value. This simple model of a neuron's functionality is due to McCulloch and Pitts (1943).

The unit computes internally the function

$$f(i_1, \ldots, i_n) = \sum_{j=1}^{n} w_j i_j$$

where i_j is the jth input and w_j is the weight associated to the jth input. This internal function is used to calculate the output value by applying, e.g., a sigmoid function to it, such as

$$\text{out} = \frac{1}{1 + e^{-af(i_1,\ldots,i_n)}}$$

with a scaling parameter a. This maps the internal function value in the range of $[-\infty, \infty]$ to a value in the range [0,1] strictly monotonically as shown in Figure 2 (right). The weights for each individual unit usually vary and are often determined by an automatic adjustment process, also called training or learning. The first well-known learning algorithm for such neuronal models had been developed by Rosenblatt (1962), who used a step function as opposed to the mentioned sigmoid output function.

Artificial neural networks are usually composed of a large number of these units; e.g., in engineering applications, the number of units typically ranges between ten and several thousands. The units are normally arranged in a layered fashion, similar to the structure in Figure 2 (left), where the outputs of the units on one layer are fed as inputs to the units on the next layer. Such networks are also known as multilayer perceptrons (usually abbreviated as MLP since the hallmark publication of Rumelhart et al. (1986)). See also Figure 2 (left). In combination with learning techniques that find appropriate weights for each unit, these networks have been successfully applied to a large number of engineering problems; for general classification problems; for recognition of visual, acoustic, and other patterns; for adaptive control of nonlinear systems; and many other types of problems.

It must be noted that the computational models of cognitive systems are not at all accurate replications of what is known about their biological functionality. To contrast the artificial neural networks, the following presents some of the current knowledge about the physiological basis of computation in the human brain.

Natural Computation in the Human Brain

The human brain is made up of a vast network of neurons coupled with receptors and effectors. The input to the network is provided by millions of receptors while the output from the brain is transformed into effects on our body and the external world via effectors. The receptors continually monitor changes in the internal and external environment. Hundreds of thousands of cells control the effectors, such as muscles and the secretion of glands.

The human brain is made up of thousands of millions of neurons. A typical estimate is about 10^{10}–10^{11}, i.e., 10–100×10^9 neurons. Many neurons have between 1000 and 10,000 connections to other neurons in the brain. There are thousands of different types of neurons in the brain. This includes neurons whose purpose has been identified as being specialized to the processing of sensory information. Those specialized neurons that deal directly with the signals coming from sensors, which are receptors for light, sound, touch, etc., are called sensor neurons. Other neurons are responsible to control our muscles and are, therefore, called motor neurons.

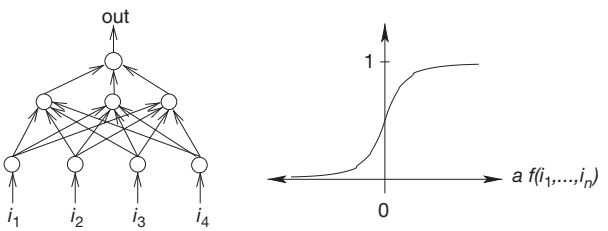

Figure 2 Left: A feed-forward neural network (multilayer perceptron). Right: A sigmoid output function.

The brain can hardly be considered as a simple stimulus–response chain from receptors to effectors. The vast network of billions and billions of neurons is rather interconnected in extremely complex and diverse ways, including simple loops as well as complex feedback cycles. Thus, signals coming from the receptors are mixed with signals already traveling in this vast network resulting in a change of the signals traveling inside the network as well as to generate signals for controlling the effectors.

In that way, the brain, or the central nervous system, manages to generate actions that depend on both the current stimulation as well as on some residue of past experiences. Past experiences may be reflected in current activities of the network, i.e., certain signals traveling inside the network as well as the current pattern of connections.

There are more than 1000 different types of neurons estimated to occur in the human brain. As a consequence of the large variety of neurons in the brain, it is impossible to speak of the typical neuron. However, a model of a neuron is sketched in Figure 3, which captures some of the basic features and functionality found in many different types of neurons.

More than 40 properties of biological neurons are known to influence a neuron's information processing. One distinguishes three different parts in biological neurons: soma, axons, and dendrites (Figure 3). Biological neurons are connected with each other. Their connections are called synapses. The axon may be considered as the output channel of the neuron, which conveys its activation level to the synaptic connections with other neurons. The ends of the axons are called nerve terminals or endbulbs. The dendrites on the other hand act as the neuron's input receptors for signals coming from other neurons. They convey the (electrical) input potential to the neuron's soma, i.e., to the neuron's body. The input potential in that case is also called postsynaptic potential. The neuron's soma, in particular where it connects to the axon, i.e., in the axon hillock, acts as an accumulator of input potentials and/or as an amplifier of these signals.

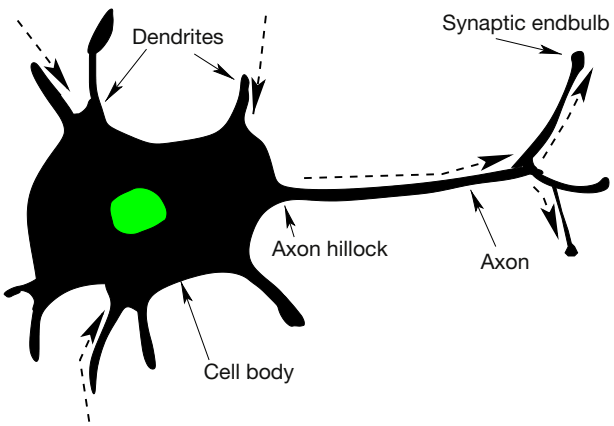

Figure 3 A schematic view of a neuron. The activity coming from receptors or from other neurons changes the membrane potentials on the dendrites and cell body (soma). The effects of these stimulations accumulate at the axon hillock where, for appropriate spatiotemporal patterns of stimulation, a threshold is reached, which triggers a pulse of membrane potential to be propagated along the axon. This pulse is carried via the branching of the axon to the synaptic endbulbs, which in turn stimulate the membrane potential of the neurons or muscle fibers to which the synaptic endbulbs are connected. The dashed arrows indicate the direction of activity propagation through the axons.

The tips of the branches of the axon impinge upon other neurons or upon other effectors. The locus of interaction between an endbulb and the cell upon which it impinges is called a synapse, and one says that the cell with the endbulb synapses upon the cell with which the connection is made. In fact, the axonal branches of some neurons can have many 'endbulbs' that allow them to make synapses along their lengths as well as their ends.

Dynamics of Biological Neurons

Neural activations are mostly circularly stimulated. A neuron is activated by other neurons to which it is connected. In turn, its own activation stimulates other connected neurons to activation.

If an impulse is started at any one place on the axon, it propagates in both directions. However, if the pulse is started at one end of the axon (normally the axon hillock), it can travel only away from that end, since once a section has been triggered it becomes refractory until well after the impulse has passed out of range. An impulse traveling along the axon triggers new impulses in each of its branches, which in turn trigger impulses in their even finer branches. Axons come in two kinds: myelinated and unmyelinated. A myelinated fiber is wrapped in a sheath of myelin, a sequence of Schwann cells wrapped tightly around an axon with small gaps, called nodes of Ranvier, between adjacent segments. Instead of the somewhat slow active propagation down an unmyelinated fiber, the nerve impulse in a myelinated fiber jumps from node to node, thus speeding passage and reducing energy requirements. When an impulse arrives at one of the endbulbs, after a slight delay it yields a change in potential difference across the membrane of the cell upon which it impinges. The membrane on the endbulb is called the presynaptic membrane, and the membrane of the surface upon which the endbulb impinges is called the postsynaptic membrane. At most synapses the direct cause of the change in (electrical) potential of the postsynaptic membrane is not electrical but chemical. There are some exceptional cell appositions that are so large or have so tight coupling that the impulse affects the polarization of the postsynaptic membrane without chemical mediation. However, the normal process is that the electrical pulse that reaches the endbulb causes the release of transmitter molecules from little packets called vesicles through the presynaptic membrane. The transmitter then diffuses across the very small synaptic cleft to the other side. When the transmitter reaches the postsynaptic membrane, it causes a change in the polarization of this membrane. The effect of the transmitter may be of two basic kinds: either excitatory, tending to move the potential difference across the postsynaptic membrane in the direction of the threshold (depolarizing the membrane), or inhibitory, tending to move the polarization away from the threshold (hyperpolarizing the membrane).

The many small changes in potential differences across the membrane of a neuron, due to the activity of many or all the synapses that impinge upon it, propagate – in most

cases – passively through the membrane of its dendrites and its cell body. They decay as they propagate. At the axon hillock, many different and often small potential changes converge, so that the total contribution by which excitation exceeds inhibition may be quite large. In other words, the transmitter released when the impulse of a single synapse arrives at the endbulb, and generally causes a change in the postsynaptic membrane, is usually not large enough to trigger the neuron. Nonetheless, the cooperative effect of many such subthreshold changes may yield a potential change at the axon hillock that exceeds threshold, and if this occurs at a time when the axon has passed the refractory period of its previous firing, then a new impulse will be fired down the axon. See Arbib (2003) and Bower (1997), or the *Journals Biological Cybernetics*, *Journal of Neurobiology*, and others.

From the description it is clear that the current artificial neural networks are only very rough approximations of the natural neural systems. Furthermore, it has been proved that certain types of artificial neural networks (cellular automata) are even computationally universal. This, however, implies also that nothing has been gained from the natural role model with respect to the function the cellular automata will compute. The function depends completely on the 'activity levels of the individual cells.' That is, with suitable initial activity patterns the system can compute any function one chooses.

Evolutionary Computation

Evolutionary computation is another field, which is strongly inspired by nature. This field was pioneered independently in the 1960s by Rechenberg (1973) and Holland (1975); both published their work in books only in the 1970s. Rechenberg used evolutionary strategies to develop highly optimized devices, such as irregularly shaped reduction pieces for pipes, e.g., for air conditioning, which proved to have lower air flow resistance than ordinary reduction pieces.

In evolutionary computation, the process of natural evolution is used as a role model for a strategy for finding optimal or near-optimal solutions for a given problem. In genetic algorithms, an important class of evolutionary computing techniques, candidates for a solution are encoded in a string, often a binary string containing 0s and 1s only. Evolution takes place by modifying the genetic code of a candidate.

Evolutionary techniques are generally applied to optimization problems. Many of those problems are combinatorial optimization problems, which are computationally hard (NP-hard). This means roughly that programs are expected to require a computing time that grows exponentially with the size of the problem. One such problem is the traveling salesperson problem, where a number of cities are given along with a distance matrix providing the traveling distance or cost for each pair of cities. The problem asks for the shortest possible round trip, i.e., the traveling salesperson departs from a given city and should visit all other cities before returning to the first city. For n cities, apart from the city of origin, there are $n!$ different round trips. A substantial proportion of them has to be evaluated before one can be sure that the optimal round trip has been found. Hence, the number of computing steps grows exponentially with the number of cities to be visited. That is, the required computing time to find the optimal solution becomes impractical even for moderate numbers of cities on today's fast computers and those computers of the near future. As a consequence, one has to be content with a good solution as opposed to an optimal solution.

The traveling salesperson problem can be handled by a genetic algorithm according to the following sketch: a population of potential solutions (more or less randomly generated orderings of the cities) is created initially as the first generation in the evolutionary process. The fitness of each individual for a solution to the posed problem is assessed, i.e., the traveling distance is calculated. The better suited individuals in the population, i.e., those with smaller traveling distances, will be retained and the less fit individuals are discarded when the next generation is created. In addition to the retained individuals are new individuals created as 'offsprings' from the current population. Offspring individuals are usually created by random mutations of the 'parents.' This may also include the combination of the genetic code of multiple parents, such as 'crossover' operators on the genetic code of two parent individuals.

From a more abstract point of view, such evolutionary approaches fall into the category of probabilistic algorithms. Other probabilistic algorithms, such as simulated annealing (which constitute another computer model of a natural, physical process), share the following common feature with the sketched genetic algorithms: a solution to the given optimization problem is developed by iteratively modifying one or more candidate solutions until a satisfactory solution is found. The iterative modification is based on some random decisions. Usually, a fitness measure is used to decide whether a new candidate solution is deemed any closer to an acceptable solution than previously generated candidates. Consequently, if a new candidate appears closer to an acceptable solution, the new candidate will normally be used to derive another modification. This is also called hill climbing in a search space. However, since no exact measure for the closeness of a candidate to an acceptable solution can generally be expected, a candidate that appears less close to an acceptable solution may still be selected for modification, but with less probability.

While the evolutionary strategies appear to be universally applicable, experience showed that in most cases problem-specific adaptations of the evolutionary scheme need to be made in order to allow the evolutionary process to generate a good solution in reasonable computing time. Such adaptations may consist of finding a suitable genetic encoding of a solution candidate and determining an effective size of the population, suitable probabilities for generating mutations, etc. Another application of the sketched genetic techniques is the development of programs, as it is done in the field of genetic programming. See Koza et al. (1999) for more details. For more details on evolutionary computation, see, e.g., Rozenberg et al. (2012).

Other Approaches

Various other ideas have been put forward about how processes from nature can be used to build computational systems. These include the following:

- Swarm intelligence: these systems utilize observations from animal swarms such as ants, fish, or even bacteria, where

a collective information gathering process can be seen that is utilized in order to optimize the behavior of the group. For example, ants have been found to use pheromones on their way back to the nest, when they have found a food source. As a consequence, other ants can perceive the relevant information about where they can find food. If multiple food locations were found by multiple 'explorer' ants, the strengths of the pheromone trails would indicate in what region perhaps even more food sources may be found. See, e.g., Engelbrecht (2006) for further reading.

- Membrane systems: these systems are based on the idea of using compartments that are separated by membranes. Inside the compartments a number of objects can reside and can be manipulated in parallel according to specific rules for each compartment (corresponding to a biological cell). Certain rules apply when an object can migrate from one compartment to another (i.e., moving through a membrane). Such a system can again be applied to optimization problems or be used to simulate the actual processes occurring in biological cells. See, e.g., Cardelli (2005).

Summary and Prospects

The presented artificial computing models of natural processes were the most studied models in recent years. They were models of biological neural systems and models of evolutionary processes in biology.

While nature inspired substantial research on these models, the models are rather metaphorical models of the natural processes. Many known details of the natural processes are missing in these models. The models are not accurate, and they are not meant to be accurate. Once simple models were developed, they were studied for their use in particular applications, such as solving certain combinatorial optimization problems or learning to classify images of characters. Hence, from an engineering perspective it is less important whether the models represent the natural processes authentically, as long as the models are useful for solving practical problems.

In regard to artificial neural networks, it is certainly desirable to enhance their capabilities, so that they come closer to the performance of the human brain. Given the substantial gap between the artificial models and what is known about the brain's physiology, an accurate computational model of the brain's physiology is not in sight. While it is unknown to what extent the brain's physiology need to be modeled in order to produce a behavior, which would be comparable to a human, it seems likely that not all aspects of the physiology need to be modeled. Which aspects are necessary and sufficient is unknown at this stage.

The artificial evolutionary computation models are no accurate models of the biological processes either. However, their application in engineering and elsewhere are clearly directed toward artificial domains, i.e., engineering problems. Hence, one would not expect that the evolutionary computation models need to be accurate models of biological evolution.

It remains an open question, however, how far will the study of such metaphorical models of nature lead? Perhaps, there exist more suitable approaches that have no resemblance with nature. After all, there are plenty of examples, where excellent engineering solutions are very different from what is found in nature, for example, the invention of the wheel or the steam engine. On the other hand, there are also examples where a partial, metaphorical model of nature's solution to a problem can be supplemented with an artificial engineering approach, for example, the development of airplanes, which do not move their wings, but have rather 'artificial' propulsion mechanisms. See also Hoffmann (1998) for a discussion of the requirements of a computing model for being used as an engineering approach to the design of complex systems.

Further reading on the topic can be found in, e.g., Rozenberg et al. (2012) or in the journals *International Journal of Natural Computing Research (IGI)* or *Natural Computing (Springer)*.

See also: Artificial Intelligence: Combinatorial Search; Artificial Intelligence: Connectionist and Symbolic Approaches; Artificial Intelligence: Genetic Programming; Cognitive Modeling: Connectionist Approaches; Information Processing Architectures: Fundamental Issues; Self-Organizing Dynamical Systems.

Bibliography

Adleman, L., 1994. Molecular computation of solutions to combinatorial problems. Science 266, 1021–1024.

Arbib, M.A. (Ed.), 2003. The Handbook of Brain Theory and Neural Networks. MIT Press, Cambridge, MA.

Bower, J.M. (Ed.), 1997. Computational Neuroscience: Trends in Research. Plenum Publishing Corporation, New York.

Cardelli, L., 2005. Brane calculi: interactions of biological membranes. In: LNCS 3082. Springer, Berlin, pp. 257–280.

Engelbrecht, A.P., 2006. Fundamentals of Computational Swarm Intelligence. Wiley.

Hoffmann, A., 1998. Paradigms of Artificial Intelligence – A Methodological and Computational Analysis. Springer-Verlag, Secaucus, New Jersey, ISBN 981-3083-97-2.

Holland, J.H., 1975. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, MI.

Koza, J.R., Bennett III, F.H., Andre, D., Keane, M.A. (Eds.), 1999. Genetic Programming III: Darwinian Invention and Problem Solving. Morgan Kaufmann Publishers, San Francisco, CA.

McCulloch, W.S., Pitts, W., 1943. A logical calculus for the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5, 115–143.

Rechenberg, I., 1973. Evolutionsstrategie: Optimierung Technischer Systeme und Prinzipien der Biologischen Evolution. Frommann-Holzboog, Stuttgart, Germany (in English: Strategy of Evolution: Optimization of Technical Systems and Principles of Biological Evolution).

Rosenblatt, F., 1962. Principles of Neurodynamics. Spartan, Chicago, IL.

Rozenberg, G., Bäck, T., Kok, J.N. (Eds.), 2012. Handbook of Natural Computation. Springer, Berlin, ISBN 978-3-540-92909-3.

Rumelhart, D.E., McClelland, J.L., The PDP Research Group, 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. I & II MIT Press, Cambridge, MA.

Turing, A.M., 1937. On computable numbers, with an application to the Entscheidungs problem. Proceedings of the London Mathematical Society 2 (42), 230–265 and (43) 544–546.

Wiener, N., 1948. Cybernetics. Wiley, New York.

Artificial Intelligence: Combinatorial Search

Richard E Korf, University of California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Combinatorial search is the study of computer algorithms designed to solve problems by systematic trial-and-error exploration of possible solutions. Problem-solving tasks include single-agent path-finding problems, such as the Fifteen Puzzle, two-player games, such as chess, and constraint-satisfaction problems, such as the Eight-Queens problem.

Single-Agent Path-Finding Problems

A classic example of a single-agent path-finding algorithm is the *Fifteen Puzzle* (Figure 1). It consists of a 4×4 frame containing 15 numbered tiles and an empty or blank position. Any tile horizontally or vertically adjacent to the blank can be slid into the blank position. The *states* of the problem are the different permutations of the tiles, and the *operators* are the legal moves. The task is to rearrange the tiles, via a sequence of legal moves, from some random initial state into a particular goal state, such as that shown in the figure, ideally using the fewest moves possible.

A search algorithm may be systematic or nonsystematic. A *systematic search algorithm* is guaranteed to find a solution if one exists and may also guarantee a lowest-cost or optimal solution. If no solution exists, a systematic algorithm may or may not detect this fact. Nonsystematic algorithms, such as *simulated annealing*, *genetic algorithms*, *tabu search*, and other *stochastic local search* algorithms, are not guaranteed to find a solution even if one exists and are not considered here.

The simplest systematic search algorithms are *brute-force search* algorithms, which do not use any knowledge about the problem other than the states, operators, initial state, and a description of the goal state. For example, *breadth-first search* starts with the initial state, generates all states that can be reached from the initial state with one operator, then all states two operators away, etc., until a goal state is reached. *Uniform-cost search*, or *Dijkstra's* single-source shortest-path algorithm (Dijkstra, 1959), allows operators to have different costs and visits states in increasing order of their total cost from the start, where the cost of a state is the sum of the operator costs on the path from the initial state to the given state. A drawback of both these algorithms is that they require enough memory to store all the states generated, which is prohibitive in large problems.

	1	2	3
4	5	6	7
8	9	10	11
12	13	14	15

Figure 1 Fifteen Puzzle.

Alternatively, *depth-first search* applies a sequence of operators from the initial state until some depth limit is reached. It then backtracks to the last state that has not been completely explored and applies a new sequence of operators from there. The primary advantage of depth-first search is that it only requires enough memory to store the current operator sequence from the initial state.

A drawback of all brute-force search algorithms is that they do not consider the location of the goal until it is reached. In contrast, a *heuristic search*, such as the *A* algorithm* (Hart et al., 1968), makes use of a *heuristic evaluation function*, which estimates the cost to reach a goal from any given state. For example, *Manhattan distance* is a common heuristic estimate for the Fifteen Puzzle. It is computed by counting the number of grid units each individual tile is away from its goal position and summing these values over all tiles. For every state n visited by A*, it adds the number of moves made from the initial state to reach state n, $g(n)$, to the heuristic estimate of the number of remaining moves to a goal, $h(n)$, to estimate the total number of moves on a path from the initial state to a goal that passes through state n. A* starts with the initial state and generates all the states one move away from it. It then evaluates those states using the cost function $f(n) = g(n) + h(n)$. At each step, it generates and evaluates the neighbors of an unexplored state n with the lowest total cost estimate, $f(n)$. It stops when it chooses a goal state to expand next.

A* is guaranteed to find a solution if one exists. Furthermore, if the heuristic function $h(n)$ never overestimates actual cost, A* is guaranteed to find a shortest solution. In the Fifteen Puzzle, for example, since each tile must move at least the Manhattan distance to its goal position, and each move only moves one tile one grid unit, Manhattan distance is a lower bound on the actual number of moves required to solve any given state. Thus, A* using the Manhattan distance heuristic function is guaranteed to find a shortest solution to the Fifteen Puzzle, if one exists, given sufficient memory.

The drawback of all combinatorial search algorithms is *combinatorial explosion*. For example, the Fifteen Puzzle contains $16!/2 \approx 10^{13}$ different reachable states. Even at a 100 million states per second, a brute-force search examining all these possibilities would take over 29 hours. The 5×5 *Twenty-Four Puzzle* contains $25!/2 \approx 10^{25}$ states. At a 100 million states per second, brute-force search would take over 2.4×10^9 years. With heuristic search techniques, optimal solutions have been found for both the Fifteen and Twenty-Four Puzzles, but not for the larger versions.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.43007-2

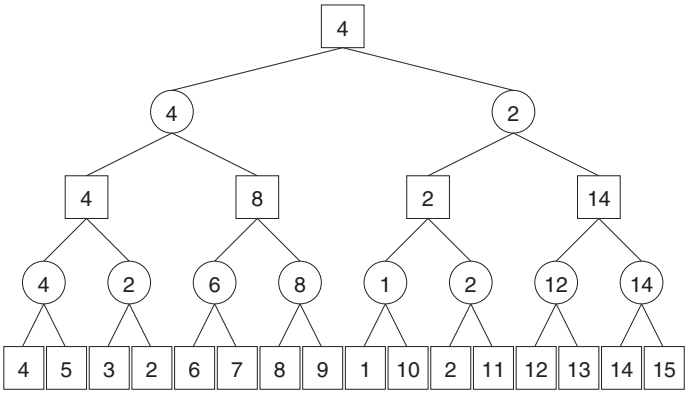

Figure 2 Minimax search tree example.

Two-Player Games

The classic *two-player game* is *chess* and has been an object of research in artificial intelligence for over 60 years. In 1997, an IBM machine named Deep Blue defeated Garry Kasparov, the world champion at the time, in a six-game match. The current best programs are comparable to the best human players. The standard algorithm for these problems is *minimax search* with static evaluation, described below, and *alpha–beta pruning*, a technique that makes minimax search much more efficient. The states of the problem are the legal board positions, and the operators are the legal moves of the game. From the current board position, minimax search considers all the legal moves, then all the possible responses by the opponent, then all the possible countermoves to those, etc. In small games, like *tic-tac-toe*, there is sufficient time for the machine to explore the entire *game tree*, or all possible plays by both players, allowing the machine to play perfectly.

In more complex games, brute-force search and hence perfect play are impossible. For example, the chess game graph is estimated to contain over 10^{40} different board positions. In that case, the machine searches ahead as many moves as it can, given the time available to make each move. The limit of how far it can look ahead is called the *search horizon*. Each of the game states at the search horizon is then evaluated by a heuristic function that returns a number, which estimates the relative merit of that board position for each player. A large positive value is advantageous for one player called MAX, and a negative value of large magnitude is beneficial to the player called MIN. For example, the simplest evaluation function for chess would sum all the pieces on the board belonging to MAX, weighted by their relative values, and subtract the weighted sum of all of MIN's pieces on the board, to get the relative material advantage of MAX over MIN. A weighting function for chess might count the queen as worth nine points, the rooks as five points each, the bishops and knights as three points, and one point for each pawn, for example.

Given the numeric evaluation of the board positions at the search horizon, the move to make is computed by the minimax algorithm. The value of each position in the game tree is computed recursively from the value of the positions below it that results from each different legal move. If the MAX player is to move in some position, the value of that position is the maximum of the values of the positions that result from each legal move. Similarly, if the MIN player is to move in a given position, the value of that position is the minimum of the values of the positions that result from each legal move. By backing these values up the tree, one can compute the value of the current game state and make one move toward the position that gave rise to this value.

Figure 2 shows an idealized four-level game tree, where the square nodes represent the game states where MAX is to move, and the circle nodes represent states where MIN is to move. The numbers at the bottom level of the tree come from the heuristic evaluation function, and the numbers at the interior nodes are computed from the nodes immediately below them by the minimax rule. The move recommended by this procedure is to the left child of the root node. A much more efficient version of this algorithm, known as alpha–beta pruning (Knuth and Moore, 1975), is used in practice.

Constraint-Satisfaction Problems

A classic *constraint-satisfaction problem* is the Eight-Queens problem. The task is to place eight queens on a standard 8×8 chessboard so that no two queens attack each other on the same row, column, or diagonal. Figure 3 shows one solution to the Eight-Queens problem. In a constraint-satisfaction problem, the task is not to find a sequence of moves, but to identify a state that satisfies a set of constraints. The standard

Q							
				Q			
							Q
					Q		
		Q					
						Q	
	Q						
			Q				

Figure 3 A solution to the Eight-Queens problem.

algorithm for these problems is known as *backtracking*. We systematically place queens on the board one at a time, one per row, so that no two queens on the board are in the same column or diagonal. If we get to a state in which no additional queens can be placed without attacking an existing queen, and all the queens have not yet been placed, we backtrack by moving the last queen that was placed to the next column where it does not attack any of the previously placed queens. The algorithm continues until one solution is found, all solutions are found, or it is determined that no solutions exist, which is the case with the two and three queens problems.

Further Reading

The classic reference in this field is by Pearl (1984). A recent book on the subject is by Edelkamp and Schroedl (2012). A recent survey article is by Korf (2010).

Bibliography

Dijkstra, Edsger W, 1959. A note on two problems in connexion with graphs. Numerische Mathematik 1, 269–271.

Edelkamp, Stefan, Schroedl, Stefan, 2012. Heuristic Search: Theory and Applications. Morgan Kaufmann, Waltham, MA.

Hart, Peter E., Nilsson, Nils J., Raphael, Bertram, 1968. A formal basis for the heuristic determination of minimum cost paths. IEEE Transactions on Systems Science and Cybernetics 4 (2), 100–107.

Knuth, Donald E., Moore, Ronald W, 1975. An analysis of alpha-beta pruning. Artificial Intelligence 6 (4), 293–326.

Korf, Richard E, 2010. Artificial intelligence search algorithms. In: Atallah, Mikhail J., Blanton, Marina (Eds.), CRC Handbook of Algorithms and Theory of Computation, second ed. CRC Press, Boca Raton, FL, pp. 22-1–22-23.

Pearl, Judea, 1984. Heuristics. Addison-Wesley, Reading, MA.

Artificial Intelligence: Connectionist and Symbolic Approaches

Ron Sun, Rensselaer Polytechnic Institute, Troy, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes the two competing paradigms of artificial intelligence: connectionist and symbolic approaches. Each paradigm has its strengths and weaknesses. No single existing paradigm currently can fully address all major issues concerning intelligence and cognition. This condition indicates the need to integrate these existing paradigms.

Perhaps the most significant feature of current computational research on intelligence and cognition is the coexistence of a number of vastly different and apparently conflicting paradigms that compete for the attention of the research communities. Artificial intelligence (AI) aims to reproduce human-level intelligence (and even beyond) in computational forms. Various computational paradigms have been developed or utilized to understand intelligence and cognition based on different theoretical and practical considerations.

This article describes two major competing paradigms of AI: the connectionist and the symbolic approaches. Brief analysis of each paradigm is provided. As a result of the analysis of their respective strengths and shortcomings, possible integration of the two is also discussed.

The Two Paradigms

The two main competing paradigms in AI can be summarized as follows:

1. The traditional symbolic paradigm (Newell and Simon, 1976). The field of AI, since its inception, has been conceived mainly as the development of computational models that capture intelligent behaviors on the basis of symbol manipulation. Computation in such models is performed based on explicit representations with symbols organized in some specific ways. Aggregate information is explicitly represented with aggregate structures that are constructed from syntactic combinations of constituent symbols.
2. The more recent connectionist paradigm (Rumelhart and McClelland, 1986; Grossberg, 1982). The emergence of the connectionist paradigm largely resulted from dissatisfactions with computational models centered on symbol manipulation, especially with their inability to deal with flexible and robust information processing in a computationally efficient manner, which is needed to capture human-level intelligence. The connectionist paradigm leads to massively parallel models that consist of a large number of simple and relatively uniform processing elements interconnected with extensive links – that is, artificial neural networks and their various generalizations. In many connectionist models, representations are distributed throughout a large number of processing elements (in correspondence with the structure of such models). Sometimes, the constituent symbolic structures of aggregate information are embedded in a network and are hard to identify. Because of their massively parallel nature, such models often are good at flexible and robust information processing and have shown promise at dealing with some issues of human intelligence and cognition that have been difficult for the symbolic paradigm.

Thus far, no single existing paradigm can fully address all the issues and problems in the computational study of intelligence and cognition. Each paradigm has its strengths and weaknesses, and each excels at certain tasks while falling short at some others. This situation indicates the need to integrate these existing paradigms (see Section Hybrid models).

Symbolic AI

The *physical symbol system hypothesis* introduced by Newell and Simon (1976) articulated the tenets of symbolic AI. They defined a *physical symbol system* as follows:

> A physical symbol system consists of a set of entities, called symbols, which are physical patterns that can occur as components of another type of entity called an expression (symbol structure). Thus a symbol structure is composed of a number of instances (or tokens) of symbols related in some physical way (such as one token being next to another) (p.116).

They further claimed that symbols can designate arbitrarily: 'a symbol may be used to designate any expression whatsoever;' 'it is not prescribed a priori what expressions it can designate.' 'There exist processes for creating any expression and for modifying any expression in arbitrary ways.' Based on that, they concluded: 'A physical symbol system has the necessary and sufficient means for general intelligent action,' which is the well-known *physical symbol system hypothesis*.

The physical symbol system hypothesis has spawned (and was used to justify) enormous research effort in traditional AI (and in cognitive science). This approach (classical symbolism) typically uses discrete symbols as primitives and performs symbol manipulation in a sequential and deliberative manner.

Representation and Search

Two fundamental ideas that have played central roles in symbolic AI are search and representation, which originated in the earliest days of AI.

Let us first discuss the idea of search space. In any problem to be tackled with AI, there is supposed to be a space of *states*, each of which describes a step in problem solving (inference). *Operators* can be applied to reach a new state from a current state. Search techniques adopted from the early days of AI include depth-first search and breadth-first search. In depth-first search, from the current state, the system examines one alternative (a path) at a time by applying one of the operators to the current state, which leads to a new state. From the new state, the same process then is repeated, until it reaches a goal state or hits a dead-end at which point no operator can be applied. In the latter case, the system backs up to a previous state and tries a different alternative. In breadth-first search, however, from the current state, the system examines all alternatives (paths) by applying all the applicable operators. Such exhaustive search techniques are inefficient. To speed up search, many heuristic search algorithms have been proposed and applied to many domains. The idea of search space has been used in all areas of AI, including in problem solving, natural language processing, robotics, vision, knowledge representation/reasoning, and machine learning (see Russell and Norvig, 2010, for more details regarding search).

Another important idea in symbolic AI is representation. Representation is based on the belief that knowledge should be expressed in an internal form that facilitates its use, corresponding to the requirement of the task to be handled and mirroring the external world in some way. A variety of symbolic representational forms have been developed over the years, and most of them are used in conjunction with search algorithms. One of the earliest representational forms is rule-based reasoning, in which discrete rules are used to direct search (inference). Rules are composed of conditions, which specify the applicability of a rule, and conclusions, which specify actions or outcomes. Rules are modular: ideally, each rule can be added to or deleted from a system, without affecting the other parts of the system (however, modularity may negatively affect dynamic interaction).

A popular form of rule-based reasoning is the production system, which evolved from psychological theories that emerged in the 1960s and 1970s. A production system consists of (1) a production rule base (for storing all the rules), (2) a working memory (for storing initial, intermediate, and final results), and (3) a control structure (for coordinating the running of the system). The system can perform either forward chaining (from conditions to actions) or backward chaining (from actions to conditions).

Formal logics constitute an alternative to production systems for rule-based reasoning, as advocated by John McCarthy (McCarthy, 1968). Logics are relatively simple, formally defined languages capable of expressing rules in a rigorous way. Logical inference is performed in formally defined ways that guarantee the completeness and soundness of the conclusions, which can be carried out using a variety of algorithms beyond forward and backward chaining. Formal logics (and most production systems) are restrictive: one needs to have all conditions precisely specified to perform one step of inference. Thus, they often are unable to deal with partial, incomplete, or approximate information. There is also no intrinsic way for handling similarity-based reasoning (Sun, 1994).

Another type of representation aims to capture the aggregate *structures* of knowledge (instead of dispersing such structures as in production systems). Knowledge can be organized in structured chunks, each of which is centered on a particular entity or concept, and each can contain, or be contained in, other chunks. Each chunk contains all the pieces of information regarding a certain entity as well as their interrelations. For instance, a *frame* (as proposed by Marvin Minsky) represents a concept in terms of its various attributes (slots), each of which has a name (label) and a value. By organizing knowledge in frames, all the relevant pieces of information can be accessed in a structured way. A *semantic network* (as proposed by Ross Quillian) consists of a set of nodes, each of which represents a particular concept, and labeled links, each of which represents a particular relation between nodes. Semantic networks also allow efficient access of knowledge, albeit in a different way, by following links that go from one concept to all the others that have a particular relation to the original one, by, for example, 'spreading activation.' *Scripts* (proposed by Roger Schank and associates) are used to represent prototypical event sequences in stereotypical situations, for example, eating in a restaurant. In such a situation, there is an almost fixed sequence. Scripts help with the recognition and handling of these sequences. The shortcomings of these types of representations include (1) structures often need to be determined a priori and hand coded, (2) they usually are fixed and cannot be changed dynamically, and (3) they may become too costly and unwieldy to capture the full extent of complex real-world situations.

Some developments in the symbolic paradigm aim to remedy some of the problems with traditional symbolic representation, especially those with logic-based approaches as discussed earlier. These extensions include, for example, default logic, among many others. *Default logic* is proposed to capture default reasoning: it deals with beliefs based on incomplete information, which might be modified or rejected later based on subsequent observations. Such logics have shortcomings, including the lack of capabilities for dealing with (1) approximate information, (2) inconsistency, and (3) reasoning as a complex, interacting process (Sun, 1994). Thus, they are somewhat deficient from the standpoint of capturing human reasoning. For an overview of these models, see Davis (1990).

In a different vein, Zadeh proposed *fuzzy logic* (see Zadeh, 1988), primarily to capture vagueness or approximate information in linguistic expressions, such as 'tall' or 'warm,' which have no clear-cut boundaries. The basic idea is as follows: for each concept, a set of objects satisfying that concept to various degrees form a subset, namely, a fuzzy subset. This fuzzy (sub) set contains as its elements pairs consisting of an object and its grade of membership, which represents the degree to which it satisfies the concept associated with the set. A fuzzy logic therefore can be constructed with which one can reason about the fuzzy truth values of concepts.

The *probabilistic approach* (especially the Bayesian network approach) treats beliefs as probabilistic events and utilizes probabilistic laws for belief combinations (Pearl, 1988). Human reasoning, however, may not always conform to the assumptions and the laws of probability theory, in part because of the complexity of the formal theory. In addition, it is not always possible to obtain accurate probability measures in practice.

Symbolic Learning

Learning is a major aspect of AI, as well as being important for cognitive science and other related fields. Although sophisticated representations have been developed for symbolic AI, learning has been difficult. This is in part due to the fact that, from its inception, symbolic AI has been centered on representation, not learning.

Nevertheless, there have been some significant developments of symbolic learning, especially since the late 1980s. The majority of work in symbolic machine learning focuses on 'batch learning.' Typically, the learner is given all the instances, positive or negative, before learning starts. Most of these algorithms handle the learning of concepts with simple rules or decision trees (Quinlan, 1986). They typically involve (1) a division of all the instances into mutually exclusive or overlapping classes (i.e., clustering), or (2) the induction of a set of classification rules describing the makeup of a concept (Michalski, 1983; Quinlan, 1986).

Some symbolic learning algorithms have been extended to deal with noisy or inconsistent data and to deal with incremental learning. In addition, inductive logic programming tries to induce more powerful first-order rules from data. Some reinforcement learning algorithms also have handled dynamic sequences, which necessarily involve temporal credit assignment (i.e., attributing a success or failure to some preceding steps; Sutton and Barto, 1998). Probabilistic learning algorithms also have been developed. Many of these more recent learning methods, however, are more numerical than symbolic.

Criticisms

Although the symbolic paradigm dominated AI and cognitive science for a long time, it has received criticisms from various sources (e.g., from Rodney Brooks, John Searle, and Hubert Dreyfus; see Dreyfus and Dreyfus, 1986). The criticisms focused largely on the disembodied abstractness of this approach, especially in relation to interacting with the world. Some, such as John Searle, emphasized the importance of biological substrates of intelligence. Others have more specific criticisms some of which were identified earlier.

Connectionist AI

In the 1980s, the publication of the parallel distributed processing book (Rumelhart et al., 1986) started the so-called connectionist revolution in AI and cognitive science. The basic idea of using a large network of extremely simple units to tackle complex computation seemed completely antithetical to the tenets of symbolic AI and has met both enthusiastic support (from those disenchanted by symbolic AI) and acrimonious attacks (from those who believed in the symbolic AI agenda). Even in the twenty-first century, we can still feel, to some extent, the divide between connectionist AI and symbolic AI, although hybrids of the two paradigms and other alternatives have flourished. Much of the controversy, however, was the result of misunderstanding, overstatement, and terminological differences.

Connectionist models are believed to be a step in the direction toward capturing the intrinsic properties of the biological substrates of intelligence in that they have been inspired by biological neural networks and seem to be closer to neurobiological processes (Grossberg, 1982). They often are capable of dealing with incomplete, approximate, and inconsistent information as well as generalization.

Connectionist Learning

Connectionist models excel at learning: Unlike the formulation of symbolic AI, which focused on representation, connectionist models have very much been about learning. Learning in connectionist models generally involves the adjustment of weights in a large network of units, so that complex computations can be accomplished through activation propagation through these weights (although there have been other types of learning algorithms, such as constructive learning and weightless learning). The weight tuning usually is based on gradient descent or its approximations for gradually minimizing an error function. The best known of such learning algorithms is the backpropagation algorithm (Rumelhart et al., 1986).

Connectionist learning algorithms have been devised for (1) supervised learning, similar in scope to aforementioned symbolic learning algorithms for classification rules but resulting in a trained network instead of a set of rules; (2) unsupervised learning, similar in scope to symbolic clustering algorithms, but without the explicit use of rules; and (3) reinforcement learning, either implementing symbolic reinforcement learning methods or adopting uniquely connectionist ones.

Connectionist learning has been applied to learning some forms of symbolic knowledge. For example, Pollack (1990) used the standard backpropagation algorithm to learn tree structures, through repeated applications of backpropagation at different branching points of a tree, in an auto-associative manner (known as recurrent auto-associative memory). Because trees are a common symbolic form, this approach is applicable to learning many symbolic structures. Similarly, Giles and Gori (1998) used backpropagation to learn finite-state automata, another common symbolic structure. Connectionist learning algorithms combine the advantages of their symbolic counterparts with the connectionist characteristics of being noise and fault tolerant and being capable of generalization. For an overview of both symbolic and connectionist learning, see Mitchell (1997).

Connectionist Knowledge Representation

Although it is relatively difficult to develop complex, structured representations in connectionist models (compared with symbolic models), there have been some significant developments of connectionist knowledge representation. Many so-called 'high-level' connectionist models have been proposed that employ representations that are comparable to, and sometimes even surpass, symbolic representations, and these models remedy certain problems of traditional representation methods. For an overview of connectionist knowledge representation, see Sun and Bookman (1995).

Let us look into some of these developments in more detail. First, logics and rules can be implemented in connectionist models in a variety of ways. For example, in one type of

connectionist system, inference is carried out by constraint satisfaction by minimizing an error function. The process, however, is slow. Another type of system, as proposed by a number of researchers in the early 1990s, uses more direct means by representing rules with links that directly connect nodes representing conditions and conclusions, respectively, and inference in these models amounts to activation propagation (e.g., Shastri and Ajjanagadde, 1993). They thus are more efficient. They also deal with the so-called variable binding problem in connectionist networks. Advanced logics that go beyond classical logic also can be incorporated into connectionist models (see, e.g., Sun, 1994).

Aggregate information can be represented in connectionist models. A system developed by Miikkulainen and Dyer (1991) encodes scripts by dividing input units of a backpropagation network into segments, each of which encodes an aspect of a script in a distributed fashion. The system is capable of dealing with incomplete (missing) information, inconsistent information, and uncertainty. In localist alternatives, a separate unit is allocated to encode an aspect of a frame (Sun, 1994).

Search, the means of utilizing knowledge in representation, may be employed or embedded in connectionist models. Either a search can be conducted through a settling or error minimization process (as discussed earlier), or a search can be conducted through a massively parallel activation propagation. Symbolic search requires global data retrieval and thus is costly in terms of time. Global error minimization (as in some connectionist models) is also time consuming. Local computation in connectionist models is a viable alternative. Knowledge may be stored in a network connected by links that capture search steps (inferences) directly. Search amounts to activation propagation by following links (similar to semantic networks in a way), without global control.

The advantage of connectionist knowledge representation is that such representation can not only handle symbolic structures in some ways but also goes beyond them by dealing with incompleteness, inconsistency, uncertainty, approximate information, and partial match (similarity) and by treating reasoning as a complex dynamic process. Developing representation in highly distributed media, such as connectionist networks, is inherently difficult.

Hybrid Models

Given the different strengths of connectionist and symbolic systems, it seems possible that combining them would be a promising avenue for developing more robust, more versatile, and more powerful systems. Research on such systems has been slowly but steadily growing. A great deal of developments have led to hybrid connectionist-symbolic systems.

The relative advantages of connectionist and symbolic models have been amply argued for – see, for example, Waltz and Feldman (1988) and Sun and Bookman (1995). The computational advantage of the combination is thus relatively easy to justify. Naturally, we want to take advantages of both types of models and their synergy (for additional justification, see Dreyfus and Dreyfus, 1986; Sun, 1994).

Many issues need to be addressed in developing hybrid connectionist-symbolic systems. These issues concern architectures, learning, and various other aspects of hybrid models. First, hybrid models likely involve a variety of different types of processes and representations. Multiple heterogeneous mechanisms interact in likely complex ways. We need to consider ways of structuring these different components; in other words, we need to consider *architectures*, which thus occupy a prominent place in this line of research. Second, although purely connectionist models, which constitute part of any hybrid system, are known to excel in their learning abilities, hybridization makes it more difficult to perform learning. In a way, hybrid systems inherit the difficulty with learning from the symbolic side and mitigate to some extent the advantage that (purely) connectionist models have in terms of learning.

Architectures and Representations

We may divide hybrid systems into two broad categories: *single-module* and *multimodule* architectures. Among single-module models, along the representation dimension, there can be the following types (Sun and Bookman, 1995): symbolic, localist (with one distinct node for representing each concept; see, e.g., Shastri and Ajjanagadde, 1993), and distributed (with a set of nonexclusive, overlapping nodes for representing each concept; see, e.g., Pollack, 1990). Among multimodule systems, we can distinguish between *homogeneous* and *heterogeneous* systems. Homogeneous systems are similar to single-module systems, except they contains several copies of the same structure, each of which can be used for processing different inputs, or the same set of inputs, to provide redundancy for various reasons.

Heterogeneous multimodule systems are more complex and interesting. This category constitutes the true hybrid systems. As an example, CONSYDERR (Sun, 1994) belongs to this category. It consists of two 'levels' (modules): the top level is a network with localist (symbolic) representation, and the bottom level is a network with distributed representation; concepts and rules are represented diffusely in the bottom level by sets of feature units. This is a similarity-based representation in which concepts are 'defined' in terms of their similarity to other concepts in these representations. The localist network is linked with the distributed network by connecting each node in the top level representing one concept to all the feature nodes in the bottom level representing the same concept. Through a three-phase interaction between the two levels, the model is capable of both rule-based and similarity-based reasoning, dealing with incomplete, inconsistent, and approximate information, and accounts for a large variety of seemingly disparate patterns in human reasoning.

A variety of distinctions can be made here. For instance, a distinction can be made in terms of representations of constituent modules. In heterogeneous multimodule systems, there can be different combinations of different types of constituent modules. For example, a system can be a combination of localist modules and distributed modules (as in CONSYDERR), or it can be a combination of symbolic modules and connectionist modules (either localist or distributed).

Another distinction can be made among heterogeneous multimodule systems in terms of the *coupling* of modules: a set of modules can be either loosely coupled or tightly coupled. In loosely coupled situations, modules communicate with each

other, primarily through message passing, shared memory locations, or shared files. Such loose coupling enables some loose forms of cooperation among modules. One form of cooperation is in terms of pre- and postprocessing versus main processing: while one or more modules take care of pre- and postprocessing, such as transforming input data or rectifying output data, a main module focuses on the main part of the task. Another form of cooperation is through a master-salve relationship: while one module maintains control of the task at hand, it can signal other modules to handle some specific aspects of the task. For example, a symbolic expert system, as part of a rule, may invoke a neural network to perform a specific classification. Yet another form of cooperation is the equal partnership of multiple modules. In this form, the modules (the equal partners) can consist of (1) functionally complementary processes, (2) multiple functionally equivalent but representationally different processes, or (3) multiple differentially specialized and heterogeneously represented experts each of which constitutes an equal partner in accomplishing a task.

In tightly coupled systems, on the other hand, the constituent modules interact through multiple channels or may even have node-to-node connections across two modules (such as CONSYDERR in which each node in one module is connected to a corresponding node in the other module; Sun, 1994). A variety of forms of cooperation among modules exist, similar to loosely coupled systems.

Learning

Learning, which can include (1) learning the content (knowledge) in a hybrid model or (2) learning and developing the structure of the model itself, is a fundamental issue that is clearly difficult. Learning is indispensable, however. Over the years, some progress on learning has been made. Although some researchers have tried to extend connectionist learning algorithms to learn complex symbolic representations, others have incorporated symbolic learning methods. For example, Sun (2002) developed CLARION, in which symbolic knowledge is extracted online from a reinforcement learning connectionist network and is used, in turn, to speed up connectionist learning and to facilitate transfer. The work showed not only the synergy between connectionist and symbolic learning but also that symbolic knowledge can be learned autonomously online, from subsymbolic knowledge, which is useful in developing autonomous agents.

Learning methods that may be applied to hybrid systems include gradient descent and its many variations (extending typical connectionist learning algorithms), expectation-maximization and its many instantiations (including hidden Markov model algorithms), search algorithms, probabilistic algorithms, evolutionary algorithms, and heuristic methods (such as decision trees or rule induction; see Mitchell, 1997). Some of these methods may be combined with others (as in Sun, 2002), which likely results in enhanced learning. A variety of other learning approaches also are being proposed, including many rule extraction or insertion algorithms.

There is a sense that future advance in this area depends on the development of new learning methods for hybrid systems and the integration of learning and complex representations. As mentioned (see Sun, 2002), symbolic representation and reasoning may well emerge from subsymbolic processes through learning, and thus an intimate and synergistic combination of symbolic and subsymbolic learning processes should be pursued.

Discussion

Despite the diversity of the connectionist, symbolic, hybrid, and other approaches discussed thus far, there is an underlying common theme: bringing together symbolic and connectionist models to achieve a synthesis of the two seemingly different paradigms.

Without doubt, we need to invest substantial efforts in continuing analyzing and exploring various paradigms, especially the symbolic paradigm and the connectionist paradigm, to reveal their respective strengths and limitations, as well as underlying assumptions, which can lead to better understanding and more rapid development of the AI field. Although significant advances have been achieved so far, this process is still on-going, and it appears that it will not be concluded until a better understanding is achieved of the nature of intelligence in computational terms.

In the meantime, it is an appealing idea that representation and learning techniques from both symbolic models and connectionist network models shall be brought together to tackle problems that neither type of model alone can apparently handle very well.

In particular, one such problem is the modeling, simulation, and explanation of human cognition, which requires dealing with a variety of psychological capacities. Several researchers (e.g., Dreyfus and Dreyfus, 1986; Sun, 1994, 2002) have argued consistently that cognition is multifaceted and better captured with a combination of symbolic and connectionist methods. Many approaches and frameworks reviewed also share the belief that connectionist and symbolic methods can be integrated usefully, and such integration may lead to significant advances in our understanding of cognition. Some probabilistic models may be seen in this light. Hybridization indeed may be a theoretically sound approach, in addition to being a practically expedient approach.

See also: Artificial and Natural Computation; Cognitive Modeling: Connectionist Approaches; Connectionist Models of Concept Learning; Intelligence: History of the Concept; Knowledge Representation; Scientific Discovery, Computational Models of.

Bibliography

Davis, E., 1990. Representations of Commonsense Knowledge. Morgan Kaufmann, San Mateo, CA.

Dreyfus, H.L., Dreyfus, S.E., 1986. Mind Over Machine. Free Press, New York.

Giles, C.L., Gori, M., 1998. Adaptive Processing of Sequences and Data Structures. Springer, New York.

Grossberg, S., 1982. Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control. Kluwer Academic Publishers, Norwell, MA.

McCarthy, J.L., 1968. Programs with Common Sense. In: Minsky, M. (Ed.), Semantic Information Processing. MIT Press, Cambridge, MA.

Michalski, R.S., 1983. A theory and methodology of inductive learning. Artificial Intelligence 20, 111–161.
Miikkulainen, R., Dyer, M., 1991. Natural language processing with modular PDP networks and distributed lexicons. Cognitive Science 15 (3), 343–399.
Mitchell, T., 1997. Machine Learning. McGraw-Hill.
Newell, A., Simon, H., 1976. Computer science as empirical inquiry: symbols and search. Communication of ACM 19, 113–126.
Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo, CA.
Pollack, J.B., 1990. Recursive distributed representations. Artificial Intelligence 46 (1–2), 77–106.
Quinlan, R., 1986. Inductive learning of decision trees. Machine Learning 1, 81–106.
Rumelhart, D.E., McClelland, J.L., PDP Research Group, 1986. Parallel Distributed Processing: Explorations in the Microstructures of Cognition. MIT Press, Cambridge, MA.
Russell, S.J., Norvig, P., 2010. Artificial Intelligence: A Modern Approach, third ed. Prentice-Hall, Englewood Cliffs, NJ.
Shastri, L., Ajjanagadde, V., 1993. From simple associations to systematic reasoning: a connectionist representation of rules, variables and dynamic bindings using temporal synchrony. Behavioral and Brain Sciences 16 (3), 417–451.
Sun, R., 1994. Integrating Rules and Connectionism for Robust Commonsense Reasoning. Wiley, New York.
Sun, R., 2002. Duality of the Mind. Lawrence Erlbaum Associates, Mahwah, NJ.
Sun, R., Bookman, L.A. (Eds.), 1995. Computational Architectures Integrating Neural and Symbolic Processes. Kluwer, Boston.
Sutton, R., Barto, A., 1998. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.
Waltz, D., Feldman, J.A. (Eds.), 1988. Connectionist Models and Their Implications. Ablex, Norwood, NJ.
Zadeh, L.A., 1988. Fuzzy logic. Computer 21 (4), 83–93.

Artificial Intelligence: Genetic Programming

Wolfgang Banzhaf, Memorial University of Newfoundland, St. John's, NL, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the method of genetic programming by first setting it into the context of artificial intelligence and the theory of evolution, and then by describing its mechanisms and the progress of the field in recent years. The second part of this article explains applications of the method. The article starts with a general discussion, then points out some of the striking results of the method, and finally turns to the area close to social and behavioral sciences. The article concludes with a discussion of methodological issues and perspectives.

Introduction

Genetic programming (GP) describes a research area within the field of artificial intelligence (AI) that deals with the *evolution* of computer code. The term *evolution* refers to an artificial process analogous to the natural evolution of living organisms according to Darwin's theory of evolution by natural selection (Darwin, 1859). The process has been abstracted and stripped of most of its intricate details. The resultant algorithms then yield approximate solutions to problems in machine learning or induce precise solutions in the form of grammatically correct (language) structures for the automatic programming of computers.

GP is part of a growing collection of evolutionary algorithms that apply heuristic search principles inspired by natural evolution to a variety of different problem domains, notably, parameter optimization. Genetic algorithms, evolutionary programming, evolutionary strategies, and differential evolution are other branches of the area of evolutionary algorithms that mostly find applications as optimization techniques. The major distinction between GP and these other areas of evolutionary algorithms is that GP handles active components like symbolic expressions, computer instructions, or other tools to control behavior, as opposed to simple parameters. Additionally, GP is able to develop its own representation of a problem by allowing variable complexity of its individuals.

All evolutionary algorithms follow Darwin's principle of differential natural selection in a variation - selection loop Figure 1. This principle states that the following preconditions must be fulfilled for evolution to occur via (natural) selection:

- Entities called individuals form a population. These entities can reproduce or can be reproduced.
- There is heredity in reproduction, that is to say that individuals produce similar offspring.
- In the course of reproduction, variety affects the likelihood of survival and therefore of reproducibility of individuals.
- Finite resources cause individuals to compete. Because of overreproduction, not all individuals can survive the struggle for existence. Differential natural selection exerts a continuous pressure toward improved individuals.

Mechanisms behind GP

GP works with a population of computer programs that are executed or interpreted to judge their behavior. Usually, a scoring operation called fitness measurement is applied to the outcome of the behavior of a program individual. For instance, the deviation between the quantitative output of a program and its target value (defined through an error function) could be used to judge the behavior of the program. This is a straightforward procedure if the function of the target program can be clearly defined. Results may be defined as side-effects of a program, such as consequences of the physical behavior of a robot controlled by a genetically developed program. Sometimes, an explicit fitness measure is missing, for instance, in a game situation, and the results of the game (winning or loosing) are taken to be sufficient scoring for the program's strategy. The general approach is to apply a variety of programs to the same problem and to compare their performance relative to each other.

The outcomes of fitness measurement are used to select programs. There are a number of different methods for selection, both deterministic and stochastic. Selection determines (1) which programs are allowed to survive (overproduction selection), and (2) which programs are allowed to reproduce (mating selection). Once a set of programs has been selected for further reproduction, the following operators are applied: reproduction, mutation, and crossover.

Reproduction simply copies an individual to an offspring population (the next generation), *mutation* varies the structure of an individual under control of a random number generator, and *crossover* mixes the structures of two (or more) programs to generate one or more new programs for the offspring population. Additional variation operators are applied in different applications. Most of these contain problem knowledge in the form of heuristic search recipes that are adapted to the problem domain.

In this way, fitness advantages of individual programs are exploited in a population to lead to better solutions. In GP, as well as in other evolutionary algorithms, differential selection can be realized in a number of different ways.

The simplest approach is a generational selection scheme called fitness proportional or roulette wheel selection. In roulette wheel selection, the fitness of all individuals in the population is summed up. The fitness f_i of each individual i is then normalized by the sum of all fitness values found, and this normalized value determines its probability of being selected for reproduction, mutation, and crossover. Thus, $p_i = f_i/\sum_j f_j$. Based on the stochasticity of random events, fitness proportional selection sometimes allows weaker individuals to succeed in reproduction.

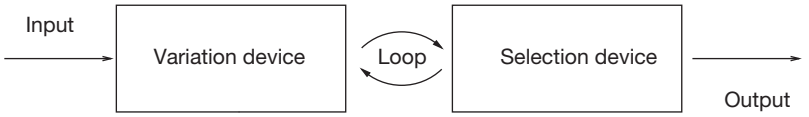

Figure 1 The variation selection loop of GP.

Another selection method is called tournament selection. In tournament selection, a subset of individuals in the population is drawn randomly, and their fitness is compared with one another. The individuals with higher fitness are allowed to replace (directly as in reproduction, or with a variation as in mutation and crossover) the individuals with lower fitness. Normally, tournament selection is done with a minimum amount of $k = 4$ individuals, which carries the lowest selection pressure. Larger tournaments, however, have been used in the literature, up to the extreme of holding tournaments among the whole population, at which point the selection scheme is called truncation selection.

Ranking selection is another selection scheme introduced to address some weaknesses of fitness proportional selection. Each individual is assigned a rank in the population, and selection proceeds either by a linearly or by an exponentially associated probability, based on the rank of an individual. A detailed comparison of different selection schemes is given in Blickle and Thiele (1996).

A key effort in GP is the definition of the fitness measure. Sometimes the fitness measure has to be iteratively improved for the evolved solutions to actually perform the function for which they were intended. The entire process can be seen in close analogy to breeding animals. The breeder has to select those individuals from the population that carry the targeted traits to a higher degree than others.

Turing (1950), Friedberg (1958), and Samuel (1963) suggested ideas similar to GP in the early days of AI, but their results were quite limited at the time. It was only after other techniques in evolutionary algorithms had been developed successfully that GP emerged. Earlier work concerned *genetic algorithms* by Holland (1975), *evolution strategies* by Rechenberg (1974) and Schwefel (1981), and *evolutionary programming* by Fogel et al. (1966). These methods have been applied successfully to a wide spectrum of problem domains, especially in optimization. It was unclear for a long time, however, whether the principles of evolution could be applied to computer code, with all its dependencies and structural brittleness. Negative results from early experiments seemed to indicate that evolution of computer code was not possible. Successes were all realized in the area of constraint optimization (Michalewicz, 1996), where methods were made available to deal with structural brittleness. These methods found their way into programming subsequently and gave rise to the new field of GP (Koza, 1992).

Progress and State-of-the-Art

In his seminal work of 1992, Koza established the field of GP by arguing convincingly that manipulation of structures of symbolic expressions is possible with evolutionary algorithms and that the resulting technique would have a wide variety of applications. In subsequent years, the field experienced both broadening and deepening (Banzhaf et al., 1998). Many different representations for GP were studied, among them other generic data structures, such as sequences of instructions (Figure 2) or directed graphs, as well as more exotic data structures, such as memory stacks or neural networks. In the twenty-first century, different approaches are considered to be GP, from the evolution of expression trees to the evolution of arbitrary structures. The overarching principle is to subject structures with variable complexity to forces of evolution by applying mutation, crossover, and fitness-based selection. The results must not necessarily be programs, but instead they could be descriptions for designs (like structures of bridges) or other manipulatable elements.

An ever-present difficulty with GP is that the evolution of structures of variable complexity often leads to large structures with considerable redundancy. Notably, it was found that variable complexity often leads to inefficient and space consuming code. It was subsequently recognized that the evolutionary forces exerted a pressure toward more complex solutions, parts of which could be removed after evolution without doing any harm to the behavior of the evolved solution. By drawing an analogy from biological evolution of genomes, this phenomenon originally was called 'code bloat,' 'intron growth,' or 'growth of ineffective code.' It was found that code growth is not the only unintended result of evolutionary processes, but it has been the most examined emergent phenomenon to date (Nordin et al., 1997). At least three different influences are at work promoting the growth of complexity during evolution. The most important influence has to do with the protection effect of redundant code if it is subjected to the action of crossover or mutation. Redundant code is more resistant to

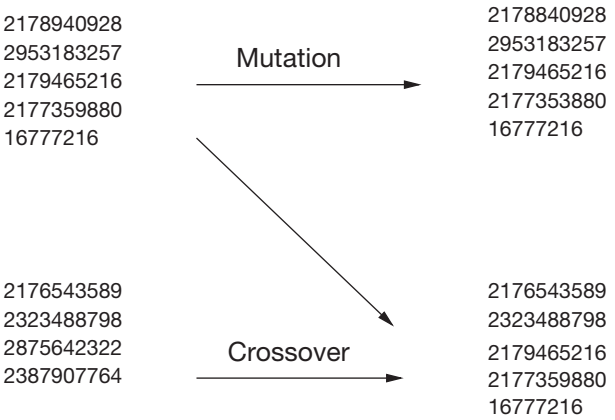

Figure 2 The primary operations of GP, mutation and crossover, are applied here to programs represented by sequences of instructions. The instructions are coded as integer numbers that allow for easy manipulation by access to these numbers.

crossover and mutation and allows its carrier solution to survive better than other individuals that do not possess this redundancy (Nordin and Banzhaf, 1995). Removal bias in crossover operations (Soule and Foster, 1998) describes the fact that code can grow to infinity from any size; another explanation is that it only can be reduced to zero for a particular size. Finally, a genetic drift toward larger solutions (Langdon, 1997) has been named as an important influence.

Over the past decade, the relation between robustness of organisms and their evolvability has been under intense study in biology (Wagner, 2005). A seeming paradox between these two features frequently found in nature has been resolved. A key aspect of robustness was found to be neutral evolution – that is, the capability of evolution to change genotypes without changes to phenotypes of individuals. This capability has been known for decades and previously had been discovered to be an important process in evolution, called neutral evolution (Kimura, 1983). The understanding of neutrality led to new mathematical models like the idea of neutral networks (Forst et al., 1995). A transfer of these ideas into GP is currently under way (Hu et al., 2011).

Although the theory of genetic algorithms has been well established, theoretical progress in GP has been more difficult to achieve because GP works with variable complexity and multiple fitness cases for fitness scoring. Many researchers are working to produce results for GP by gleaning information from the research on genetic algorithms. The schema theory of genetic algorithms (Goldberg, 1989; Vose, 1999) has been a primary target of knowledge transfer. In the meantime, a number of different schema theorems have been formulated for GP, and theory has progressed substantially (Poli et al., 2010).

When analyzing search spaces of programs, it was realized that their size is many orders of magnitude larger than search spaces of combinatorial optimization problems. A typical size for a program search space might be $10^{100,000}$, as opposed to a typical search space for a combinatorial optimization problem of the order of 10^{100}. Although this might be a discouraging interpretation for search mechanisms, it also was realized that the solution density in program spaces is, above a certain threshold of complexity, constant with changing complexity (Langdon, 1999). In other words, there are proportionally many more valid solutions in program spaces than in the spaces of combinatorial optimization problems.

Applications

The main application areas of GP are as follows (from narrow to wide) (Banzhaf et al., 1998):

- Computer science
- Science
- Engineering
- Arts and entertainment

In computer science, the development of algorithms has been a focus of attention. By being able to manipulate symbolic structures, GP is one of the few heuristic search methods for algorithms. Sorting algorithms, caching algorithms, compression algorithms (Kattan and Poli, 2011), random number generators, and algorithms for automatic parallelization of code (Ryan, 2000), to name a few, have been studied. The spectrum of applications in computer science spans from the generation of proofs for predicate calculus to the evolution of machine code for accelerating function evaluation. The general tendency is to automate the design process for algorithms of different kinds. Recently, the process of debugging code, that is, the correction of errors, has been added to the list of applications (Forrest et al., 2009).

Typical applications in science are modeling and pattern recognition. Modeling certain processes in physics and chemistry with the unconventional help of evolutionary creativity supports research and understanding of the systems under study. Pattern recognition is a key ability in molecular biology and other branches of biology, as well as in science in general. In this context, GP has delivered first results that are competitive if not better than human-generated results.

In engineering, GP is used in competition or cooperation with other heuristic methods such as neural networks or fuzzy systems. The general goal is again to model processes, such as production plants, or to classify results of production. In recent years, design in engineering has gained some prominence (Lohn et al., 2004). Control of manmade apparatus is another area in which GP has been used successfully, with process control and robot control being the primary applications.

In arts and entertainment, GP is used to evolve realistic animation scenes and appealing visual graphics. It has been used to extract structural information from musical composition to model the process so that automatic composition of music pieces becomes possible.

Many of these problems require a huge amount of computational power on the part of the GP systems. Parallel evolution has been a key engineering aspect of developments in GP. As a paradigm, GP is well suited for a natural way of parallelization. With the advent of inexpensive parallel hardware, in recent years – in particular through graphics processing units (Banzhaf et al., 2009) – a considerable proliferation of results is expected from GP systems.

Human-Competitive Results of GP

In the past decade, a substantial number of results have been published in various fields that claim to have produced human-competitive results by applying GP as a problem-solving method.

These claims are based on a comparison between the presently best-known human solutions to a problem and their respective counterparts produced by GP. Applications are from areas like quantum computing algorithms, analog electrical circuit design and other mechanical and electrical designs, game-playing applications, finite algebras and other mathematical systems, bioinformatics and other scientific pattern recognition problems, reverse engineering of systems, and empirical model discovery. Some of the similarities of these successes were summarized by Koza (2010) as follows:

- Usually, a large amount of computational power has to be invested to gain human-competitive results from GP runs.
- Most times, a dedicated representation for the solution, known to be efficient by the specialist, has been applied to

allow the full power of expression of solutions to be born on the problem.
- The GP system has been equipped with dedicated growth or development operators such that the adaptation of complexity of a description can be achieved smoothly.

Because of the ability of the human mind to quickly grasp the recipes of problem solution that an artificial system has applied, the question remains open whether solutions found by a GP system will remain qualitatively better than solutions discovered by human.

Evolution, Development, and Learning

The relation between evolution and development has been studied for decades in biology (West-Eberhard, 2003). It was found that the time-dependent process of gene expression and gene regulation through both internal and external cues is the mechanism by which both processes can be unified (Ben-Tabou de Leon and Davidson, 2007). Some progress has been made in GP to couple evolution and development. The developmental approach in GP takes the form of a recipe that, upon its execution, generates a structure that is subjected to fitness tests (Spector and Stoffel, 1996). Thus, it is not the GP program itself that is tested, but the result of its execution.

Similar to the coupling between evolution and development, the coupling between development and learning was considered an important link for understanding the mechanisms of development and learning processes. Cognitive neuroscience has presented evidence for this coupling by finding that there are critical periods in development in which certain learning tasks are facilitated (and sometimes only possible) to take place. If the critical period is missed, learning success in a task is substantially reduced (Daw et al., 1978).

The coupling between development (or evolution) and learning has only recently been explored in GP. The problem is to clearly separate adaptations or fitness gains resulting from development or evolution versus those from learning. Its reason is the less stringent separation of time scales between evolution, development, and learning in GP systems. Although biological evolution can happen over many thousands or millions of years, development over the lifetime of an organism, and learning over phases of that lifetime, all three mechanisms are on similar time scales in GP and usually are tied into single runs of a GP system. In addition, in most GP systems, there is no notion of species (and their evolution). Rather, the entire population is essentially mixed and therefore belongs to one single species. Finally, the goal under the influence of evolution is behavior, the same entity usually associated with learning. First attempts have been made to examine learning as a separate task for which evolution and development have to provide the means (Harding et al., 2009), yet this area requires much more investigation.

Methodological Issues and Future Directions

Although GP has made great strides over the past two decades, a number of issues are still open and require continued investigation. The theory of GP has (Poli et al., 2010)

- succeeded in finding appropriate schema theorems that allow for an understanding of how the search space and the population representation interact,
- started to analyze Markov chain models of the search process dynamics, and
- found ways to characterize search spaces (difficulty, complexity) and relate them to the performance of GP systems.

In coming years, this theory is expected to make progress on the treatment of dynamical problems, proofs of convergence of the search algorithms, and in classifying problem spaces. On the practical side, research in GP will target the following (O'Neill et al., 2010):

- identifying appropriate representations for GP in particular problems,
- designing open-ended evolutionary systems with GP,
- addressing the problem of generalization in GP,
- establishing benchmarks for measuring and comparing GP performance, and
- pursuing modularity and scalability in GP.

Implementation of GP will continue to benefit in coming years from new approaches, including results from developmental biology and epigenetics. Application of GP will continue to broaden. Many applications focus on controlling behavior of real or virtual agents. In this role, GP may contribute considerably to the growing field of social and behavioral simulations. Genetic algorithms already have been found to be beneficial in optimizing strategies of social agents (Gilbert and Troitzsch, 1999). With its ability to adjust the complexity of a strategy to the environment and to allow competition between agents, GP is well positioned to play an important role in the study and simulation of societies and their evolution (Edmonds et al., 2007).

See also: Artificial Intelligence: Combinatorial Search; Artificial and Natural Computation; Autonomous Agents; Cognitive Modeling: Connectionist Approaches.

Bibliography

Banzhaf, W., Harding, S., Langdon, W., Wilson, G., 2009. Accelerating genetic programming through graphics processing units. In: Riolo, R., Soule, T., Worzel, B. (Eds.), Genetic Programming Theory and Practice VI. Springer, New York, pp. 1–23.

Banzhaf, W., Nordin, P., Keller, R., Francone, F., 1998. Genetic Programming – An Introduction. Morgan Kaufmann.

Ben-Tabou de Leon, S., Davidson, E., 2007. Gene regulation: gene control network in development. Annual Reviews in Biophysics and Biomolecular Structure 36, 191–212.

Blickle, T., Thiele, L., 1996. A comparison of selection schemes used in evolutionary algorithms. Evolutionary Computation 4, 361–394.

Darwin, C., 1859. On the Origin of Species. John Murray, London.

Daw, N., Berman, N., Ariel, M., 1978. Interaction of critical periods in the visual cortex of kittens. Science 199, 565–567.

Edmonds, B., Hernandez, C., Troitzsch, K. (Eds.), 2007. Social Simulation – Technologies, Advances and New Discoveries. IGI Global.

Fogel, L., Owens, A., Walsh, M., 1966. Artificial Intelligence through Simulated Evolution. Wiley, New York.

Forrest, S., Nguyen, T., Weimer, W., Le Goues, C., 2009. A genetic programming approach to automated software repair. In: Rothlauf, F. (Ed.), Proceedings of the 11th International Conference on Genetic and Evolutionary Computation. ACM Press, New York.

Forst, C., Reidys, C., Weber, J., 1995. Evolutionary dynamics and optimization. In: Moran, F., Moreno, A., Merelo, J., Chacon, P. (Eds.), Advances in Artificial Life. Springer, Berlin.
Friedberg, R., 1958. A learning machine: part I. IBM Journal of Research and Development 2, 2–13.
Gilbert, N., Troitzsch, K., 1999. Simulation for the Social Scientist. Open University Press, Buckingham, UK.
Goldberg, D., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley, Reading, MA.
Harding, S., Miller, J., Banzhaf, W., 2009. Evolution, development and learning using self-modifying cartesian genetic programming. In: Rothlauf, F. (Ed.), Proceedings of the 11th International Conference on Genetic and Evolutionary Computation. ACM Press, New York.
Holland, J., 1975. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, MI.
Hu, T., Payne, J., Moore, J., Banzhaf, W., 2011. Robustness, evolvability, and accessibility in linear genetic programming. In: Silva, S., Foster, J.A., Nicolau, M., Machado, P., Giacobini, M. (Eds.), Proceedings EuroGP 2011. Springer, Berlin.
Kattan, A., Poli, R., 2011. Evolution of human-competitive lossless compression algorithms with gp-zip2. Genetic Programming and Evolvable Machines 12, 1–30.
Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK.
Koza, J., 1992. Genetic Programming. MIT Press, Cambridge, MA.
Koza, J., 2010. Human-competitive results produced by genetic programming. Genetic Programming and Evolvable Machines 11, 251–284.
Langdon, W., 1997. Fitness Causes Bloat, Technical Report, Technical Report CSRP-97-22. University of Birmingham.
Langdon, W., 1999. Boolean function fitness spaces. In: Poli, R., Nordin, P., Langdon, W., Fogarty, T. (Eds.), Proceedings EuroGP'99. Springer, Berlin.
Lohn, J., Hornby, G., Linden, D., 2004. An evolved antenna for deployment on NASA's space technology 5 mission. In: O'Reilly, U., Riolo, R., Yu, G., Worzel, W. (Eds.), Genetic Programming Theory and Practice II. Kluwer Academic, Boston, MA.
Michalewicz, Z., 1996. Genetic Algorithms + Data Structures = Evolution Programs. Springer, Berlin.
Nordin, P., Banzhaf, W., 1995. Complexity compression and evolution. In: Eshelman, L. (Ed.), Genetic Algorithms: Proceedings of the Sixth International Conference (ICGA95). Morgan Kaufmann, San Francisco, CA, USA, pp. 310–317.
Nordin, P., Banzhaf, W., Francone, F., 1997. Introns in nature and in simulated structure evolution. In: Lundh, D., Olsson, B., Narayanan, A. (Eds.), Proceedings of Biocomputing and Emergent Computation (BCEC97), Skovde, Sweden, September 1–2, 1997. World Scientific, Singapore, pp. 22–35.
O'Neill, M., Vanneschi, L., Gustafson, S., Banzhaf, W., 2010. Open issues in genetic programming. Genetic Programming and Evolvable Machines 11, 339–363.
Poli, R., Vanneschi, L., Langdon, W., Freitag, N., 2010. Theoretical results in genetic programming: the next ten years. Genetic Programming and Evolvable Machines 11, 285–320.
Rechenberg, I., 1974. Evolution Strategy. Holzmann Froboog.
Ryan, C., 2000. Automatic Re-engineering of Software Using Genetic Programming. Kluwer Academic, Boston, MA.
Samuel, A., 1963. Some studies in machine learning using the game of checkers. In: Feigenbaum, E., Feldman, J. (Eds.), Computers and Thought. McGraw-Hill, New York.
Schwefel, H., 1981. Numerical Optimization of Computer Models. Wiley, Chichester, UK.
Soule, T., Foster, J., 1998. Removal bias: a new cause of code growth in tree-based evolutionary programming. In: IEEE International Conference on Evolutionary Computation. IEEE Press.
Spector, L., Stoffel, K., 1996. Ontogenetic programming. In: Koza, J., Goldberg, D., Fogel, D., Riolo, R. (Eds.), Proc. Genetic Programming 1996. MIT Press, Cambridge, MA.
Turing, A., 1950. Computing machinery and intelligence. Mind 59, 433–460.
Vose, M., 1999. The Simple Genetic Algorithm: Foundations and Theory. MIT Press, Cambridge, MA.
Wagner, A., 2005. Robustness and Evolvability in Living Systems. Princeton University Press, Princeton, NJ.
West-Eberhard, M., 2003. Developmental Plasticity and Evolution. Oxford University Press, New York.

Artisans and Guilds, History of

Josef Ehmer, University of Salzburg, Salzburg, Austria

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 24, pp. 816–821, © 2001, Elsevier Ltd.

Abstract

Artisanal labor can be traced to the Neolithic Revolution, and can be found as a significant social group in every subsequent society, including contemporary societies of whatever economic level across the globe. For part of European history especially artisans gather in guilds, or associations of artisans, were especially important, and their significance is explained and underlined. In doing so the role of artisans in the preindustrial craft economy is compared with their role in industrializing and modern societies, underlining the very different ideas about skill entailed in each instance.

This article starts with a definition of artisanal labor and artisans. It presents the origins of artisans in the Neolithic Revolution, their social positions in ancient and medieval times as well as in early modern European corporate societies, and finally it discusses the redefinition of the artisan in eighteenth and nineteenth century class societies. In the second part, the article describes guilds as a particular type of artisanal association which existed almost everywhere in late medieval and early modern urban Europe, and it analyzes the role of artisanal and guild traditions in modern middle class and labor movements. A discussion of recent revisionism in guild and artisan historiography completes the article. This discussion stresses the dynamic and flexible character of the preindustrial craft economy, the perception of guilds as dynamic fields of social relations, the persistence of artisanal small-scale production in industrializing societies the gender dimension of the concept of the artisan, and, finally, the ambiguities of the concept of skill.

Definitions

Artisanal labor is a particular form of labor that has existed in most – if not all – human societies. Historically, it first appeared as nonagricultural production, and during the last 200 years primarily as small-scale or small-commodity production, in contrast to mass production and mechanized factory industry. The focus of the definition is on production, thus also distinguishing artisanal labor from economic activities such as commerce. Indeed, one cannot draw strict boundaries here. The English term 'trade,' for instance, has been in use for artisanal production as well as for commerce and exchange. Furthermore, artisanal labor is not necessarily limited to production, but also includes some activities that in modern economic thought are assigned to the tertiary or service sector. Artisan is a narrower concept than artisanal labor. While artisanal labor may be performed by men and women as one activity among many others, artisans can be defined as persons who specialize in artisanal labor and make their living mainly, if not exclusively, from this kind of work. There are some almost universal elements in the concept of the artisan: manual labor and the use of tools; specialization of labor, which connects the concept of the artisan with the concept of profession or occupation; skill and some kind of training as a precondition of skill; and work not for subsistence but for consumers or markets. Besides these universal elements, there are also some historically variable ones that fluctuate in accordance with broader social and economic change.

Artisans: Concepts and History

The origins of artisans are connected with social changes at the end of the Neolithic Revolution (see Work, History of) in approximately the fourth millennium BC. Foremost among these was the development of technology, above all in the fields of metallurgy, ceramics, and woodworking. This promoted specialization in the production of metals as well as in the fashioning of tools, weapons, and a growing number of practical implements. Second, the emergence of social stratification brought about increased demand for luxury goods by the upper classes, the production of which called for special training and particular skills. Third, there was the formation of urban cultures, especially in the Near East and the Mediterranean. Artisans were concentrated in cities, and it was here that the division of labor within artisanal production accelerated and specialized crafts and trades came into being. During the first millennium BC, dozens of different artisanal occupations emerged in the urban cultures of India, the Near East, and the Mediterranean.

The value assigned to labor in the crafts and trades and the social position of the artisan has been characterized by a high degree of ambivalence. In all societies in which work was held in low esteem, this applied to artisanal labor too. In European history, though, there is also a long tradition of valuing the work of artisans more highly than other forms of manual labor. Whereas in some non-European cultures, such as ancient China, artisans were assigned a lower social status than peasants, in the European tradition they enjoyed a relatively high level of prestige. In early ancient Greece, this was expressed in the myth of Prometheus the god of fire, pottery, and metallurgy. "In the eighth century BC the crafts of blacksmith, pottery, and weaving were considered in the same light as that of priests, bards and doctors (…). All these activities were called *technai*, that is, activities implying the use of secret processes, specialized knowledge based on long apprenticeship and an initiation ritual." (Godelier, 1980)

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62012-3

Some crafts appeared as sacred in Gallic and Germanic mythology as well, and in the European Middle Ages, for instance, in the scale of social values in early Germanic law, artisans clearly had a higher rank than persons engaged in agricultural activities. "The craftsmen who forged the weapons for the aristocratic warriors (…); and the goldsmiths who decorated these weapons and created jewels for the women of these warriors (…) were important personages who upheld the prestige of technical skill" (Le Goff, 1980).

Artisans in Corporate Society

The urbanization process in twelfth- to fourteenth- century Europe strongly influenced the concept and the social standing of the artisan. Most importantly – and not at all usual for people working with their hands – artisans, as urban citizens, were freemen. In many English towns, for instance, completion of an apprenticeship entitled an individual to the freemanship of the respective town (which meant permission to take up residence, claim to social relief if the need arose, and often some form of political participation). The increasing division of labor in towns created a wide range of artisanal occupations, and the expanding guild system (cf guilds, see Section Guilds: Concepts and History) served as an institutional framework for training and work in these occupations. Both of these developments had consequences, which became basic to the concept of the artisan in the corporate societies of Europe from the late Middle Ages until the eighteenth century.

First, artisanal work – even if it was manual labor – was not simply regarded as labor. The negative connotations of labor (travail, hard work) as toil, pain, burden, or penitence were replaced by an understanding of artisanal labor as art. *Arts et métiers* (arts and trades) was the common French term used to group together all artisans (a generalizing term, which appeared in the sixteenth century). In eighteenth-century France, artisanal crafts were regarded as honorable because they were based not only on the 'effort of the body' but also on the "subtlety of the mind" (Sewell, 1980: p. 23). Second, by organizing in guilds or other associations, artisans had a recognized standing in society and the state. Whether they enjoyed royal privileges, municipal statutes, or the permission of the church to form a confraternity, all these defined them as a corps or as an estate (French: *état*; German: *Stand*). In both respects, inherent in the concept of the artisan was a sharp distinction vis-à-vis unskilled laborers.

The Ambiguous Journeyman

Traditionally, the concept of the artisan referred primarily to the master artisan – an independent craftsman, who ran his own workshop and casually or regularly employed a few hands. As head of the household, the master was part of the corporate structure and exercised domestic authority over his family and his workforce. On the other hand, the conception of the artisan also included different categories of employees. First, it referred to apprenticeship, which was understood as a period (and a social position) of training – generally, although not necessarily, connected with adolescence. In the late Middle Ages, there appeared a further group: artisans who had finished their apprenticeship but had not – or not yet – become masters. In English, the term journeyman (which stems not from the journey of the tramping artisan, but from the French word *jour* meaning day, and designating one who works and is paid by the day) became most widespread to describe this group. While the concepts of apprentice and master are relatively clear, the concept of the journeymen is fluid and ambiguous. On one hand, it refers to a life-cycle stage, as the completion of training, mainly by tramping a couple of years. From this perspective, journeymen were young, single men subject to the domestic authority of the master (when at work and not on the road), until they themselves became masters and householders. On the other hand, the concept of journeymen refers to wage laborers – often on a lifelong basis – having their own households and families. In practice, it is almost impossible to make a clear distinction between these two concepts and modes of life. Very roughly, the first type prevailed in small-scale production (bakers, tailors, and the like), the second in textiles (particularly where the craft was transferred into outwork or domestic industry) and in the building trades. Furthermore, the first seems to be more common in Central Europe, whereas the second is more familiar in the West, particularly in England. Many historians also believe that over the course of the late medieval and early modern history, the first type was gradually replaced by the second, but this is probably a far too general and too linear model (cf Reith, 1999). Generally, the relations between masters and journeymen included solidarity and moral community as well as conflict.

Redefinition of the Artisan in Class Society

In the emerging class societies of eighteenth- and nineteenth-century Europe, the concept of the artisan became redefined both in *petit bourgeois* and in proletarian terms. On one hand, the concept still focused on the independent master craftsman working in his own workshop with his own tools. Apprentices and journeymen were included in this concept as future or anticipated masters. On the other hand, the concept of the artisan increasingly referred to lifelong employed workers working in a wide range of economic settings – even in factories – but being distinguished from ordinary laborers by their common 'property of skill.' Artisans were "those who through apprenticeship or its equivalent had come to possess a skill in a particular craft and the right to exercise it" (Rule, 1987: p. 102).

In social theory, the first model became predominant. In Enlightenment philosophy as well as in political economy, the artisan began to be regarded as a countermodel to capitalism. Karl Marx conceptualized the artisan as a personification of a non- (or pre-) capitalist mode of production, which he called small-commodity production (*kleine Warenproduktion*). At the turn of the century, the German social scientist Werner Sombart regarded the artisan as the 'ideal type' of the precapitalist economic system, which, in his view, was characterized by empiricism and traditionalism (as opposed to rationality) and a lack of interest in accumulation and profit. This perspective became extremely powerful and shaped more or less completely the historical interpretation of the artisan.

In practice, however, in the economically advanced regions of Western Europe, most of the workmen who identified

themselves as artisans were in wage employment. Adam Smith in his 'Wealth of Nations' (1776) estimated that in England there were 20 men working for wages for everyone who was his own master (Rule, 1987: p. 103). Here, the artisan as a skilled workman became central to the self-image, identity, and rhetoric of the emerging working class. On the continent, and particularly in Germany, even in the mid-nineteenth century, the number of masters almost equaled the number of journeymen in the crafts and trades. Here, the concept of the artisan became synonymous with independent master craftsmen and their respective *petit bourgeois* political movements.

Guilds: Concepts and History

In Western history, both the concept and the practice of artisans are closely connected with guilds. Associations of artisans existed in many historic societies and cultures. For example, there is evidence of their existence in classical India, ancient Rome, Tokugawa Japan, and Colonial America. Nevertheless, the term guild or craft guild refers to a particular type of artisanal association, which had its place in the corporate societies of Europe from the late Middle Ages to the nineteenth century. Despite enormous variations from city to city, region to region, and one historical period to another, the concept of the guild refers to a combination of several characteristics. First, a major organizational principle was occupation. Usually, although not necessarily, guilds organized artisans of the same occupation or of related/similar occupational groups, and occupation formed the major line of differentiation separating one guild from another. Changes in the division of labor and in occupational structures are therefore the major dynamic forces in the development of the guild system. Second, guilds were strongly concerned with economic matters in their respective sector of the economy. Market regulation – both in respect to commodity markets and labor markets – was an essential part of their economic activities. Third, guilds were primarily organizations of master craftsmen. Occasionally they also included apprentices and journeymen in minor positions; on other occasions, they definitely excluded these groups, thus providing the impetus for journeymen to find their own clandestine or legal associations. Fourth, guilds were formally constituted and officially recognized bodies. Statutes or privileges granted by municipal or state authorities made them a solid part of the corporative structures of society. In this context, guilds also performed political functions, casually participating in the political system, occasionally even dominating urban politics. Lastly, guilds were not only economic and political bodies, but also multifunctional institutions engaged in religious, cultural, and social activities.

The first guilds – in the sense of this definition – were established in France, England, Germany, and Italy in the twelfth century. In the thirteenth and fourteenth centuries, guilds existed almost everywhere in urban Europe with a wide variety of titles and designations. Latin sources use terms such as *officium*, *ministerium*, *magisterium*, *artificium*, *collegium*, and many others. The French language speaks of *métiers*, *corps des métiers*, or *corporations*; in English they are called guilds, crafts, or mysteries; in German *Zunft*, *Zeche*, *Innung*, or *Amt*; in Italian *arti*, *gremi*, and so on.

Guilds in Corporate Society

Despite these obviously medieval roots, European guilds became a mass phenomenon and a nearly universal institution in the Early Modern Era. In the Netherlands and France, Spain and the various Italian states, German-speaking Central Europe, and the Nordic countries, most guilds were founded in the sixteenth, seventeenth, and eighteenth centuries. Even in England with its highly developed market economy, many guilds were reestablished in the Restoration, and in cities like Bristol and Bath, there were even new guilds being founded as late as the mid-eighteenth century. A guild system also developed in the Ottoman Empire, particularly in the Balkans, the Near East, and Northern Africa. In the Russian Empire, on the other hand, guilds were unknown. In Colonial America during the early eighteenth century, many artisans took an interest in setting up craft guilds. Even though some associations were actually founded (such as the Carpenter's Company of the City and County of Philadelphia in 1724), no general guild system came into being.

Even when guilds were at the pinnacle of their importance in Europe, not all artisans were incorporated. In Western Europe and particularly in France and England, guilds were limited to cities, and did not even exist in all cities. In many Central European states – for instance, Austria – guilds transcended municipal boundaries and became regional or nationwide institutions organizing urban artisans as well as craftsmen from villages and the countryside. In those places where guilds did exist, there were few instances in which all artisans belonged to them. A large proportion of artisanal labor always took place outside of the guild system.

Mercantilist theorists and political leaders of the seventeenth and eighteenth centuries regarded guilds as instruments of unified state regulation of trades, labor, and taxes. During this period, many European states indeed tried to make guilds universal and to establish the guild system as a framework for all economic activities; one such example is the effort of French Prime Minister Colbert in his trade laws of 1673.

The End of the Guild System

A hundred years later, the opinion of guilds held by political and intellectual elites had changed completely. During the second half of the eighteenth century, an antagonistic discourse emerged. Mainstream economic and political thought – expressed in the fields of political economy, physiocracy, or in Enlightenment thought in general – came to regard guilds as obsolete relics of the past that sought to retard or prevent economic development. Proponents of this approach called for complete freedom of trades and the abolition of guilds. Conclusions diametrically opposed to these were drawn by conservative thinkers who saw guilds as the embodiment of a corporative ideal and as a means of combating the assumed decay of authority and order.

When the leading physiocrat and Enlightenment thinker Turgot served briefly (1774–76) as French prime minister, he issued a decree in 1776 completely abolishing all guild privileges as well as expressly forbidding "all masters, journeymen, workers, and apprentices (…) to form any association or assembly among themselves (…)" (Sewell, 1980: p. 73).

Despite the fact these measures were repealed only a few months after Turgot's fall from power, they nevertheless initiated the process of the elimination of guilds from Europe over the course of the next 100 years. Revolutionary France took a radical step as early as 1791 and banned guilds along with all other labor organizations. England's Combination Acts of 1799/1800 also mandated the 'prohibition of all combinations' – those of masters as well as of journeymen (Chase, 2000: p. 84). Most Central European states oscillated between anti-guild and proguild principles. They indeed forbade the organization of artisanal workers, but permitted the guilds – as organizations of masters – to remain in existence. It was not until the 1860s that the guilds were dissolved once and for all in the German states and in the Habsburg Monarchy. This process reflected the guilds' loss of their social function. In most countries, dissolving the guilds met no opposition, but it was precisely in Central Europe that guilds remained alive as the object of the projections engendered by a corporative social ideal and conservative nostalgia.

Guild Traditions in Petit Bourgeois and Labor Movements

The end of the guild system did not also mean the end of artisanal organizations. These developed over the course of the nineteenth century in two separate directions. On one hand, there was the organizational continuity of master artisans, independent craftsmen, and employers. This line of tradition was significant above all in Central Europe. In the German and the Habsburg Empires, in Denmark, and in some other countries, cooperative organizations in the crafts and trades continued to exist even after the dissolution of the guilds. From the 1880s, these organizations were granted wide-ranging authority in their respective fields, such as control over the training of apprentices. Politically, these trade associations were closely associated with the 'middle-class movements,' the political movements of the *petits bourgeois*. They demanded the restoration of guilds and developed corporative ideals in the framework of conservative and right-wing politics (cf Haupt, 1996).

The tradition of guild crafts and trades, however, also played an important role in the emerging labor movement. Although journeymen's associations had been banned in many European countries at around 1800, confraternities, *compagnonnages*, houses of call, and similar institutions continued to exist on an informal or illegal basis. Although there was no direct institutional continuity between guilds and trade unions, there was nevertheless a close connection on the conceptual, experiential, and personnel levels. Artisanal workmen who came out of guild traditions were used to organizing in such associations, which they considered their political right. They were likewise accustomed to settling labor disputes collectively, be that in the form of negotiations or job actions. Furthermore, they also carried on the tradition accentuating and defending their social status, which was based upon the 'property of skill.' All these traditions were important incentives for the founding of the first trade unions in the first half of the nineteenth century.

In a broad sense, artisans were also the driving force behind the early political labor movement. Of course, the concept of class contradicted the tradition of craft consciousness, since it stressed the common experiences of all workingmen, skilled or unskilled, in whichever craft or trade they engaged. Nevertheless, there is almost universal agreement "that skilled artisans, not workers in the new factory industries, dominated labor movements during the first decades of industrialization. Whether in France, England, Germany, or the US, whether in strikes, political movements, or incidents of collective violence, one finds over and over the same familiar trades: carpenters, tailors, bakers, cabinetmakers, shoemakers, stonemasons, printers, locksmiths, joiners, and the like. The nineteenth-century labor movement was born in the craft workshop, not in the dark, satanic mill" (Sewell, 1980: p. 1).

Revisionism in Guild and Artisan Historiography

The negative assessment of guilds in eighteenth- and nineteenth-century thought also characterized historiography. Until recently, most historians had assumed that guilds were technologically regressive, that they restricted economic growth and strait-jacketed the development of free market capitalism.

All of these assumptions have been subjected to critique since around 1980. A growing number of historians is now engaged in a 'revisionist version of guild history' (cf Epstein et al., 1998; Ehmer, 1998). With respect to early modern Europe, artisans and guilds have become reintegrated into the mainstream of the development toward capitalism. The Dutch case shows most clearly what can be seen in many other European regions as well: 'guilds were part and parcel of commercial capitalism,' showing extensive growth particularly in the 'Golden Age' (Lucassen, 1995). The increasing dependency of Europeans on markets and the growing variety of available goods and of artisans who made them are now seen as evidence of a "dynamic, flexible, and creative craft economy" (Farr, 1997: p. 25). The guilds' emphasis on skill, their system of formalized training, and the technological transfer through tramping artisans, are now said to explain, 'the undisputed technological leadership of guild-based production', particularly in comparison to the rural putting-out system (proto-industry) and early centralized factories (Epstein, 1998: p. 705).

Furthermore, guilds are no longer regarded as stable institutions of equals, but as highly stratified and dynamic fields of social relations. Merchant capitalists used some guilds to control and discipline economically dependent masters and journeymen, whereas other guilds were instruments of resistance to merchant capital. There are also new perspectives on artisanal cultures, conceptualizing them as historically variable and adaptable sets of practices. In this new light, customs that were regarded as 'ancient' by the artisans themselves as well as by the historians who came after them are now often being shown to have a rather brief history. For instance, French *compagnonnages*, journeymen's associations with highly elaborate rituals and an aura of mystery, experienced their 'golden age' from the mid-eighteenth century to the 1830s or 1840s – the very period of redefinition of the position of artisans and guilds in society (cf Sonenscher, 1987: pp. 31–63).

Another area in which revision is taking place is the role of the artisans in the Industrial Revolution. In eighteenth- and nineteenth-century Britain, small-scale and labor-intensive production, skill and 'hand technology' drawn from a huge reservoir of urban and rural artisans played a key role, not as

a mere remnant of the past, but as a central and dynamic component of industrial capitalism (cf Berg, 1985). Generally, artisanal small-scale production displayed remarkable persistence in industrializing societies, adapting to the needs of industry, and benefiting from the growing demand of consumer society. The generally successful process of adaptation was not harmonious and free of conflict. Some artisanal branches survived, some new ones were born, but others – including all textile production and most metal manufacturing – declined and vanished. Artisans in these trades experienced the devaluation of their skills. The same happened to the journeymen of the metropolitan garment industries, which left the artisans' workshops and became reorganized as an urban domestic industry based on the cheap labor of women and children instead of that of skilled workmen.

The role of women in the world of the artisans is a major point of recent discussions (Hafter, 1995; Hanawalt, 1986). It is generally agreed that women in most human societies regularly performed artisanal labor, but rather seldom became artisans in the sense of professional, organized, and honorable workers. In European history, the concepts of artisan and, to an even greater extent, guild were male concepts. Artisans who practiced crafts and trades in European cities during the late Middle Ages produced commodities that had previously been produced in households by women as well as by men. The difference was that the professional artisans then produced for the market and their labor became regulated by a set of formal and informal rules. Even if 'the early guild ordinances (…) did not regard skilled guild labor as a male preserve,' in the long run the switch from household production to market production, from unregulated artisanal labor to a 'craft,' excluded women (Wiesner, 1989).

Recent research clearly shows that women participated in artisanal labor in many respects, as masters' wives and daughters, as domestic servants in artisans' households, as widows, and, last but not least, as autonomous independent producers. Nevertheless, in all these roles they were not officially recognized as skilled workers. This shows that the very concept of skill, which is central to the conception of the artisan, is highly ambiguous. On one hand, skill indeed requires a certain body of knowledge, experience, and training; on the other hand, skill is a historically variable social construct. Not all labor activities that require some expertise are regarded as skilled work. In European history, skill was defined by male artisans and guildsmen not only with respect to actual work practices, but also as property, status, and a means of social distinction.

See also: Capitalism; Industrialization, Typologies and History of; Industrialization; Labor History; Labor Unions; Labor, Division of; Middle Ages, The; Social Stratification; Work, History of; Work, Labor: History of the Concept.

Bibliography

Arbeit, W., 2012. Links to the Past: The Work of Early Hawaiian Artisans. University of Hawai'i Press, Honolulu.

Barbieri-Low, A.J., 2007. Artisans in Early Imperial China. University of Washington Press, Seattle.

Béal, J.-C., Goyon, J.-C., Institut d'archéologie et d'histoire de l'Antiquité (Lyon), 2002. Les artisans dans la ville antique. Université Lumière-Lyon 2, Lyon.

Berg, M., 1985. The Age of Manufactures Industry, Innovation and Work in Britian 1700–1820. Fontana, London.

Chase, M., 2000. Early Trade Unionism. Fraternity, Skill and the Politics of Labour. Ashgate, Aldershot, Brookfield, Singapore, Sidney.

Cooper, L.H., 2011. Artisans and Narrative Craft in Late-medieval England. Cambridge University Press, Cambridge.

Coquery, N., 2004. Artisans, industrie: Nouvelles révolutions du Moyen Age à nos jours. Société française d'histoire des sciences et des techniques, Paris.

Crossick, G., Haupt, H.-G., 1995. The Petite Bourgeoisie in Europe 1780–1914. Routledge, London and New York.

Ehmer, J., 1998. In: Lenger (Ed.), Traditionelles Denken und neue Fragestellungen zur Geschichte von Handwerk und Zunft, pp. 19–77.

Epstein, S.A., 1991. Wage Labor and Guilds in Medieval Europe. The University of North Carolina Press, Chapel Hill and London.

Epstein, S.R., 1998. Craft guilds, apprenticeship, and technological change in preindustrial Europe. The Journal of Economy History 58, 684–713.

Epstein, S.R., Haupt, H.G., Poni, C., Soly, H. (Eds.), 1998. Guilds, Economy and Society. Publicacioines de la Universidad de Sevilla, Seville, Spain.

Faroqhi, S., 2009. Artisans of Empire: Crafts and Craftspeople under the Ottomans. I.B. Tauris, London.

Farr, J.R., 1997. On the shop floor: guilds, artisans, and the European market economy, 1350–1750. Journal of European Modern History 1, 24–54.

Farr, J.R., 2000. Artisans in Europe, 1300–1914. Cambridge University Press, Cambridge, UK.

García-Bryce, I.L., 2004. Crafting the Republic: Lima's Artisans and Nation Building in Peru, 1821–1879. University of New Mexico Press, Albuquerque.

Gillespie, M., 2000. Free Labor in an Unfree World: White Artisans in Slaveholding Georgia, 1789–1860. University of Georgia Press, Athens.

Godelier, M., 1980. Work and its representations: a research proposal. History Workship Journal 10, 164–174.

Guenzi, A., Massa, P., Piola Caselli, F. (Eds.), 1998. Guilds, Markets and Work Regulations in Italy, 16th–19th Centuries. Ashgate, Aldershot, Brookfield, Singapore, Sidney.

Hafter, D.M. (Ed.), 1995. European Women and Preindustrial Craft. Indiana University Press, Bloomington and Indianapolis, IN.

Hanawalt, B. (Ed.), 1986. Women and Work in Preindustrial Europe. Indiana University Press, Bloomington and Indianapolis, IN.

Haupt, H.-G., 1996. Zum Fortbestand des Ancien Régime im Europa des 19. Jahrhunderts: Zünfte und Zunftideale. In: Hettling, M., Nolte, P. (Eds.), Nation and Gessellschaft in Deutschland. Historische Essays. Verlag C. H. Beck, Munich, Germany.

Haupt, H.-G., 2000. Das Ende des Ancien Régime. Zünfte in Europa um 1800. Vandenhoek und Ruprecht, Göttingen, Germany.

Hugger, P. (Ed.), 1991. Handwerk zwischen Idealbild and Wirklichkeit. Kultur- und sozialgeschichtliche Beiträge. Verlag Paul Haupt, Bern and Stuttgart, Germany.

Icher, F., 2007. Die Weisheit des Handwerks: Traditionelles Wissen neu entdeckt. Knesebeck, München.

Jan, S., 2004. Forgotten Identities: Women Artisans of Kashmir. Decent Books, New Delhi.

Johns, P.S., Rizzo, J., Bommel, J., 2001. Traditionelle Spezialitäten aus Italien: Ausgewählte Familienbetriebe und überlieferte Rezepte. Christian-Verl, München.

Kristensen, T.M., Poulsen, B., Birk, S., Archaeological Institute of America, 2012. Ateliers and Artisans in Roman Art and Archaeology. Journal of Roman Archaeology, Portsmouth, RI.

Le Goff, J., 1980. Time, Work and Culture in the Middle Ages. The University of Chicago Press, Chicago.

Lenger, F., 1998a. Sozialgeschichte der deutschen Handwerker seit 1800. Frankfurt am Main, Suhrkamp, Germany.

Lenger, F. (Ed.), 1998b. Handwerk, Hausindustrie und die historische Schule der Nationalökonomie. Wissenschafts- und gewerbegeschichtliche Perspektiven. Verlag für Regionalgeschichte, Bielefeld, Germany.

Lis, C., Soly, H. (Eds.), 1997. Werelden van Verschil. Ambachtsgildern in de Lage Landen. VVB Press, Brussels.

Lovell, M.M., 2005. Art in a Season of Revolution: Painters, Artisans, and Patrons in Early America. University of Pennsylvania Press, Philadelphia.

Lucassen, J., 1995. Labour and early modern economic development. In: Davids, K., Lucassen, J. (Eds.), Miracle Mirrored. The Dutch Republic in European Perspective. Cambridge University Press, Cambridge, UK, pp. 367–409.

Reith, R., 1999. Lohn und Leistung. Lohnformen im Gewerbe 1450–1900. Franz Steiner Verlag, Stuttgart, Germany.

Rohr, D.A., 2001. The Careers of British Musicians, 1750–1850: A Profession of Artisans. Cambridge University Press, Cambridge.

Rule, J., 1987. The property of skill in the period of manufacture. In: Joyce, P. (Ed.), The Historical Meanings of Work. Cambridge University Press, Cambridge, UK, pp. 99–118.

Schulz, K. (Ed.), 1999. Handwerk in Europa. Vom Spätmittelalter bis zur Frühen Neuzeit. R. Oldenbourg Verlag, Munich, Germany.

Sewell, W., 1980. Work and Revolution in France. The Language of Labor from the Old Regime to 1848. Cambridge University Press, Cambridge, UK.

Sonenscher, M., 1987. Mythical work: workship production and the compagnonnages of eighteenth-century France. In: Joyce, P. (Ed.), The Historical Meaning of Work. Cambridge University Press, Cambridge, UK, pp. 31–63.

Wiesner, M.E., 1989. Guilds, male bonding and women, work in early modern Germany. Gender and History 2, 125–137.

Artists, Competition and Markets

Pierre-Michel Menger, Collège de France, Paris, France; and Ecole des Hautes Etudes en Sciences Sociales, CESPRA, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Income distribution in art occupations is highly skewed. Abilities and talent seem to be better candidates than initial training to explain skewness. Yet quality differences among artists and works play a puzzling role: They can be detected only through comparisons, but the standards of comparative evaluation must remain indeterminate enough to allow for substantial differentiation and competition through originality and innovation. Careers are patterned by comparative tournaments: Even tiny differences at the outset suffice to trigger a mechanism of cumulative advantage, with attention and reward being concentrated on those favorably compared, out of proportion to the underlying distributional gap of abilities and skills. Assortative matchings lower the cost of creative invention. Thus initially boundless interindividual competition, with its distinctive excess supply mark, gradually results in stratification by reputation and status.

Income distribution in art occupations generally follows the Pareto curve: One-tenth of professionals in the given field earn half of all annually distributed income; one-fifth earn 80%. In art there are more individuals earning nothing – or less than nothing after art-related expenses – than in any other higher occupation. At the other extreme we have the elongated tip of the distribution, signaling the presence of artists with astronomically high incomes – a level that brings to mind lottery payoff matrices. To what mechanisms of the artistic labor market should this discrepancy and the resulting extreme inequalities be attributed? Earnings distribution is structured entirely differently from the skills and qualifications distribution associated with wage equations. Whereas the distribution of human capital factors of the sort included in wage equations typically forms a bell curve wherein individuals of the given population are fairly symmetrically distributed around mean values and the majority of individuals are at the center of the distribution, in creative jobs we have an extremely asymmetrical curve. This is where market organization of artistic production and consumption come under suspicion. The critical view of inequality-generating mechanisms suggests the Paretian distribution of rewards demonstrates the true pathologies of a competition-driven market organization. The primary aim of this article is to explain such a skewness and to refer careers in creative undertakings to mechanisms of quality detection and evaluation that are both indeterminate enough to let innovation lead the game and efficient enough to translate uncertain occupational prospects into wide gaps in reputation, status, and rewards.

Job Stratification

Acccording to the theory of job stratification (Stinchcombe, 1986; Jacobs, 1981; Baron and Kreps, 1999), each occupation can be assigned certain characteristics and capacity requirements whose social and economic values are assessed according to their degree of scarcity and the nature of the collaboration between those working together. The degree of scarcity reflects the vertical ranking of abilities, skills, and performance: An extraordinary talent will be admired and, where a market value exists, exploited at considerable profit, provided that it is sought after by a sufficient number of people willing to pay for it. Moreover, a talented worker may make an unusually high contribution to the success of his organization, more than proportionate to the differential between his qualities and those of his teammates. It is in these professions that competition to attract and reward individuals deemed exceptionally talented is fiercest and it is here that earnings concentration creates winner-take-all or winner-take-the-most situations (Frank and Cook, 1995). In such professions, the probability of obtaining an exceptionally fine result is low and most performances produce average results. The cost for the company of hiring an average professional is low compared to the profits it will reap if it finds someone exceptional, and this leads to an employment policy or contract relationship that brings in a great number of different individuals, the aim being to find the 'real gem.'

Let's call these jobs 'star jobs,' following Baron and Kreps' reworking of the stratification model of Stinchcombe, with its distinction between 'star jobs,' 'guardian jobs,' and 'foot-soldier jobs.' To look where to find 'star jobs,' one may think of scientific research, sports, the entertainment industry (cinema, radio and television, concerts, shows, and performances for a wide audience), and the arts. Yet not all of the artistic occupations entail this 'star job' profile. The creative artist (novelist, painter, composer, and solo performer) obviously holds a 'star job,' as characterized above.

Artistic jobs like that of teacher or orchestra musician primarily falls into a second category, that of 'guardian jobs': In these, even spectacularly excellent individual contributions cannot bring the organization or team any additional reputation or profit. In these jobs, the skills required for performing the activity are an 'additive' production factor and they are more homogeneously distributed among the individuals concerned. Lastly, extra-artistic activities are usually the equivalent

This article draws on models and analyses expounded in my forthcoming book *The Economics of Creativity: Art and Achievement under Uncertainty*, Cambridge: Harvard University Press, 2014.

of 'foot-soldier jobs': Within them, variation in individual performance has limited impact and the range of individual differences is of very limited value. Here the success of the organization depends on the aggregation of all individual performances. Employees are hired on the basis of a simple wage negotiation: Anyone who accepts the proposed wage gets hired.

Star Jobs and Paretian Distribution of Earnings

Clearly, star jobs (i.e., primary creative activities) are those that earn their successful incumbents the highest rewards (monetary and nonmonetary, like esteem and social recognition) and those for which cultural enterprises such as publishers, gallery owners, and recording and film companies are very clearly looking for the rare gem.

In these jobs, high value is laid on expertise and talent, and perceived differences in quality are of decisive importance in orienting the buyer's preferences. Thus given that demand is highly sensitive to differences in quality, at a given price for a good or service, a buyer's utility will be greater if she chooses a professional who is considered more talented than another. This seems quite obvious when objective quality superiority is at stake. A surgeon with a 10% greater ability to save lives than others will be in greater demand. His fees and aggregate earnings will exceed those of his colleagues by much more than 10%. Thus, if the question is why differences in reward can be extremely disproportionate to differences in ability, the first answer lies in the mechanism that concentrates earnings on the basis of perceived quality differences and nonlinear increase in expected utility from higher quality (Rosen, 1981; Goode, 1979).

The situation in the arts looks somewhat different at first sight, as it leans toward the subjective side. In contrast to the surgeon, artistic talent and art commodity quality differentials point to subjective utility. Yet a quality difference that yields greater subjective utility is an inherent feature of the service demanded by the audience: It is precisely what the consumer is looking for. Without the assumption that quality differences play a fundamental role in orienting consumer preferences, we could not understand why there is competition among artists. And this assumption was precisely ruled out by the Marxian attack against the market organization of production as functionally and economically separated from consumption. For, as in the case of the superior surgeon who saves more lives than another (but with much less dramatic consequences), an artist deemed superior is much more desirable than an artist of inferior quality, and this should hold without consumers being subjected to any kind of external influence. Two concerts, exhibitions or films of moderate quality will not give me as much satisfaction as one high-quality concert, exhibition or film. The quality perceived as superior is powerful enough to trigger demand concentration and therefore celebrity and wealth for those artists reputed to have the greatest talent.

Providers deemed to have superior talent have a larger market share that may extend immensely due to joint consumption technology: Digitization, trading networks, and duplication of commodities enable artists to serve instantly very large markets. Audiences are potentially enormous relative to the scale most of us operate on. And even if a provider is unable to clone himself and the commodities or services he supplies cannot be reproduced (surgery, consultancy, painting, and concert performances), his career can develop on a worldwide scale like that of a small number of famous practitioners in great demand in each trade.

How to Cope with Quality Uncertainty? Talent, Tournaments, and Excess Supply

Up to this point, we stand on the demand side. The argument relies on two key assumptions: (1) quality differences in production and performance can be perceived; (2) quality differences should be perceived without bias, otherwise quality evaluation is subject to manipulation and as it is well known, bad quality eventually crowds out the good one.

How to detect and rate quality differences? It would be easy to evaluate artists and their work and perceive qualitative differences if everything could be assessed in absolute terms, on the basis of a univocal scale and a stable set of perfectly unambiguous criteria. The selection process in the course of initial artistic training would perfectly filter the many candidates through some simple tests and contests. Things take another course once ability is viewed as a multidimensional and multifactorial component. Indeed, the fundamental properties of a creative activity are unlimited differentiation of its products (Caves, 2000) and originality-driven competition. Therefore, in stark contrast to a timed athletic performance or problem resolution, esthetic originality and artistic value can only be measured in relative terms. And reasoning in terms of success factors would become spurious: It is, of course, the *combination* of various types of qualities and skills that counts, but there is no detectable ideal formula for an optimal combination or optimal proportions of those qualities and skills. We suspect that the skewed distribution of those qualities and their indecipherable combinations may create sharp inequalities in chances of success, but it is impossible to estimate that distribution a priori. This is why people engage in relative comparison.

Competition in a given market determines the value of artists' accomplishments. It does so through the intensity of immediate preferential demand, and through a flow in demand which is linked to the durability of the artwork and to the interdependencies between the artworks that are successively produced over the course of an artist's career (since the success of one artwork can trigger enthusiasm for the artist's previous works, and bring heightened attention to those that follow). The quality of each artistic good supplied is uncertain. It is impossible to directly assess the abilities of artists through skills measurements and standardized tests, no less than it is to assess the value of artworks using an unequivocal scale of measurement and a stable set of unambiguous criteria, as the fundamental property of artistic activity is driven by competition for originality. Thus, in contrast to a timed sports performance or the solving of a problem, esthetic originality and artistic value can only be measured in relative terms. The rankings, reward schemes, and career advancement profiles must take the form of tournaments (competitive tournaments in music, recruitment through auditions, awarding of prizes,

hit parade rankings, critics' evaluations and scorings, etc.) in which evaluations are based on incessant comparison. Artists work to differentiate themselves from one another on multiple levels so as to make their mark in the competition. However, critics, art world professionals, market intermediaries (producers, employers, organizers, and agents), and consumers never cease their rankings. Initially dispersed, evaluations gradually align, due to affiliation processes (Van Rees, 1987). The cultural knowledge required to appreciate and assess artworks can be defined as the sum of significant comparisons an individual is capable of making, explicitly or implicitly, for the purpose of attributing meaning and value to an artwork. Thus, artworks initially juxtaposed through the law of originality become hierarchically ordered by audiences and art world professionals, through a series of competitive and comparative tests, according to their preferences and investments. What is called "talent" can be defined as the quality gradient attributed to the individual artist through these relative comparisons. The difficulty in defining talent rests in that it is not an arbitrary value, but rather a purely differential quality (Menger, 2014).

These characteristics are reflected in the twofold, operational strategy of cultural entrepreneurs: the exploitation of uncertainty, which is a condition for entrepreneurial profit, and the reduction of uncertainty. Very little is known about the ingredients for success. Uncertainty about the market potential of each artwork and innovation therefore pushes each firm to hedge its bets across a broad range of artists; this drives cultural industry entrepreneurs as a whole to feed, through a composition effect, a structural excess supply characterized by seasonal peaks and short term fluctuations (Hirsch, 1972).

Yet, as soon as cultural entrepreneurs manage to identify an artist with 'high potential,' they set about overexposing him and activating the mechanisms of contagious imitation in the general public. They do this by exploiting the self-reinforcing dynamic that transforms an artist's success into both an effect and a cause of the quality attributed to him by consumers. They may then seek to 'develop' the artist who has enjoyed his first successes, just as is done in research and development with scientific inventions or technical innovations. Thus, after having taken advantage of the uncertainty about who will emerge as winners by exploiting competition through differentiation, cultural entrepreneurs endeavor to reduce uncertainty about a promising artist's chances of future success by seeking to transform his instant value into a lasting one – a sure asset in which it is possible to invest.

Should the problem of artistic overproduction simply be regarded as the modern phenomenon of the industrialization of culture? In fact, nearly every sociologist, economist, and historian who has ever studied artistic labor markets has underscored the excess supply of artists (Ehrlich, 1985; Graña, 1964; Lenman, 1989; Montias, 1996; White and White, 1965). If overproduction is as frequently invoked as a disease caused by the industrialization of culture, it is because cultural industries are constantly organizing and reorganizing themselves so as to turn overproduction into an advanced technique for managing uncertainty and into a mechanism for increasing profitability. The wide range of influences exerted by cultural entrepreneurs over consumer choice and over the evaluation of competing artists is illustrated here with the greatest clarity.

In effect, the entire architecture of the cultural industries (music, book, cinema, and audiovisual production) has been built on organized relations between the production, distribution, promotion, and consumption of artistic goods and services – thereby generating competition through tournaments (hit parades) to identify profitable talents. The basic economic principles at work here are well known (Curien and Moreau, 2006; Benhamou, 2008; Connolly and Krueger, 2006). The form of competition in this sector is that of oligopolistic concentration. In the music industry, for example, a few big firms realize three-quarters of the world-wide turnover for the sector, and a nebula of so-called independent producers maintain relations of 'coopetition' with these major labels. Concentration is explained by the cost structure: In the music and book industries, the cost of producing a good (the acquisition and payment of the raw material – creative labor) and the cost of material production are proportionately low and variable, whereas the costs of distribution and promotion are high and fixed (accounting for three-quarters of total costs). Industrial concentration is, consequently, motivated by economies of scale linked to the distribution and commercial exploitation of a raw material – artists' creativity – that can be obtained and controlled at low cost, even though its market value is difficult to predict.

The structure of competition reflects this way of proceeding. Large firms find it more profitable to let independent producers act as adventurous explorers – as risk-takers and fine connoisseurs of market niches and emerging trends – and to engage in cooperative competition with them, via the distribution of their products and financial participation in their capital. This is the classic figure of 'oligopoly with a competitive fringe.' Small firms devote most of their resources to scouting talents and to financing their own productions. The major firms, meanwhile, extract rent via the distribution of independent productions, buy up the contracts of independent producers with successful artists, develop the most promising careers, ally themselves with stars, and work on triggering and reinforcing the dynamics of success amplification through investments in advertising and promotion. The picture should, however, not be painted in such binary terms: The major firms themselves are in reality entire galaxies in which labels behave like autonomous centers of production and profit while also acting as talent scouts. Nonetheless, the distinction according to size continues to hold; it imparts its characteristics on the demographics of firm population, resulting in a high mortality rate for small firms, growth for more skillful or fortunate ones, buyouts, mergers, and firm concentration. Furthermore, the pattern of concentration has changed: In continuity with past practice, vast multimedia groups have been formed that produce and distribute their own products along with those of other producers, and that control publishing companies (with copyrights protected for ever longer periods of time). However, these groups currently have much greater control than before over radio and television stations and over commercial operation networks (e.g., retail sales, on-line sales, and payable downloads), enabling them to directly activate the levers of success amplification.

Under conditions of flexible specialization, the number of independent producers grows, even as the rate of sector concentration increases. What is, in fact, also on the rise is the

density of independent producers' interdependencies with both other firms and dominant companies (Hesmondhalgh, 2002). Thus, the 'openness' of the system of competitive production explains why, despite the fact that the rate of oligopolistic concentration in the recording industry has increased since the early 1970s, rates of innovation and diversity in musical production (as calculated on the basis of hit parades) either remained stable or increased throughout the 1980s (depending on the indicator chosen (Peterson and Berger, 1975; Lopes, 1992)).

Assortative Matchings

In the models I have discussed thus far, spectacular inequalities in success primarily concern the specific situation in which artists or professionals with valued expertise are competing to attract demand – each acting individually, through direct interaction with the market, and without any apparent partners. However, in order to work and to produce and diffuse their products, professionals usually join a permanent or temporary organization (e.g., orchestra, theater company, film production crew, etc.); alternatively, they enter into a contract with an organization that acts as an intermediary (e.g., publishing house, recording company, or art gallery) in order to create material reproductions of a given good, place an artwork into circulation, and gain access to the market. It is here that another trigger of inequality comes into play: assortative matching.

Assortative matching characterizes the multiplicative nature of the production function in artistic labor. Like a scientist in Merton's cumulative advantage model (Merton, 1968), an artist benefits from associating with professionals reputed to be of equal or superior quality in their respective careers. Indeed, in order for an artist to secure the best chances of developing his talent, it is important for him to associate with professionals of comparable value working in the other professions involved in the production and circulation of artworks. For example, a reputed director will seek to enlist top professionals in the key positions of a film production, and the head of a publishing house will entrust his most seasoned literary director with the task of managing labor relations with the company's most talented and promising writers or with the latter's agents (Thompson, 2010). Furthermore, in the very early stages of artistic and scientific careers, formal learning trajectories followed by on-the-job learning (and even by apprenticeship, a persistent form of artisanal organization in certain artistic and technical-artistic trades) are heavily determined by an individual's association with experienced partners. These are the people who provide an artist on the path toward professionalization with better opportunities for developing skills in demanding projects, through contact with fellow artists who have themselves been selected based on their potential.

Art worlds combine labile organizational architectures (e.g., network, project, and vertical disintegration) with teams structured through the association of professionals of equivalent quality or reputation – that is, through assortative matching. Labor markets for the most highly qualified jobs are, in this way, hierarchically ordered by professional pairings. A successful career dynamic constitutes a movement of upward mobility in a world stratified by networks of acquaintances and recurring collaborations. In my above presentation of the existing analyses of job stratification by functional importance, I highlighted that, in the practice of an artistic profession, talent is a complementary rather than additive factor of production. Assembling talents of approximately equal level – each in their respective function (interpretation, organizational mediation, editing, fundraising, etc.) – has a multiplicative effect on a given project's chances for success and on the project collaborators' chances of accumulating reputation (Rosen, 1983; Kremer, 1993). The interdependence of performances renders the benefits of this relationship complementary: A publisher who attracts talented authors will increase his own chances of expanding his experience and renown in the editorial field, just as a talented author will benefit from collaborating with a publisher reputed for his professional qualities.

Faulkner (1983), in his study of Hollywood cinema, demonstrated the power of assortative matching in the cultural industry. Accominotti (2008) has since shown that the mobility of painters in the gallery networks he studied follows a similar mechanism. Numerous studies dedicated to scientific careers, networks of collaboration and copublication among researchers, and upward mobility in the American academic labor market, have likewise revealed how: (1) Stratification of the professor–researcher population by reputation level (in terms of productivity, visibility, and readership) regulates collaborative pairings; and (2) The recruitment and career mobility market is governed by the rule whereby the candidate's relative value must be matched to the institution's position in the hierarchy of excellence.

We can easily connect the assortative matching argument to the dynamic analysis of careers based on the mechanisms of competitive tournament and cumulative advantage. In the course of artists' early formative experiences, capacities manifest themselves differently and unequally according to each individual. What still remains undetermined, however, is the nature of the differences in talent that exist between the creators who, over the relatively long term, will achieve success – either lastingly or not – and those who will fare less well. Expressed in terms of probabilities of success, the advantage gained from demonstrating a potential talent early in one's career may indeed be weak; yet any (great or small) difference perceived in each test of competitive comparison suffices to polarize the investments and wagers of system actors (i.e., artists themselves, trainers, professionals, patrons, entrepreneurs, critics, and consumers). The learning content of different work situations relies on the same mechanism: There exists an optimal profile for expanding one's skills, which is a function of the number and variety of work experiences and of the quality of the collaborative networks mobilized by the artist as he moves from one project to the next.

The relative comparison of artworks and artistic performances conducted in competitive tests, together with the lasting indeterminacy of the course of creative activity, imbue artistic labor with a continual tension. It is on this basis that analyses of the gaps in reputation and success put forth the causal role of networks of interdependence and cooperation in artistic labor. In order to create and diffuse their works, creators and artists enter into contractual relations with

organizations, such as artist agencies, publishing houses, recording companies, painting galleries, production companies, etc.

Selective tests, as well as information on the value of an artist's commitment to creative labor provided step by step to himself, his peers, his colleagues, and his supporters, allow us to understand the career dynamics of artists foremost as a learning trajectory within a segmented system of activity. Whether we focus on the professionals who sponsor the artist's debut, on his partners, or on the various other categories of art world actors with whom he establishes work relationships, it is through the development of collaborative networks that an artist organizes his activities. This is especially so given that he must mobilize diversified resources in order to develop his work capabilities.

Two remarkable consequences of assortative matching must be highlighted. First, the multiple and incessant evaluative judgments, on the basis of which reputation hierarchies are constructed, act as structuring forces that segment a professional milieu whose activities do not fit in a stable organizational mold. Second, gaps in talent that are initially small, or at the very least of uncertain importance with regard to future success, are rapidly increased by the game of assortative matching – and this, due to the multiplicative effect this game has on the expression of the individual qualities of collaborators, and to the authority it bestows on those who co-opt each other for their creative projects.

We are now in a better position to understand how a hierarchy develops. A hierarchy is founded on reputation, which has the functional property of an efficient vector of information and investment in a universe of sharply differentiated activities and goods. Neither art world professionals nor consumers can estimate through direct experience the value of every artist and artwork; nor can they continuously reestimate an artist's value in the shifting context of incessantly renewed competition. As Stinchcombe (1968) explains, the vocabulary of talent, genius, brilliance, and creativity is a common means of attributing extraordinary qualities to individuals in sectors of activity wherein success is highly uncertain (e.g., the arts, research, and advertising). The primary semantic function of these terms is to draw a sharp distinction between these extraordinary qualities and all the other typically measurable qualities of an individual. Furthermore, these terms are essentially assigned a posteriori, as a means of domesticating and categorizing that which is unmeasurable – that which cannot be predicted but only 'retrodicted.' This vocabulary acts to convert an uncertainty dispelled a posteriori (in light of experience) into an expectation of success. Insofar, the attribution of talent acts as a mechanism for reducing social complexity.

The increase in artist reputation reveals the existence of a mechanism of cumulative advantage. The artistic career presents itself to a young professional as a succession of comparative tests in which each performance is judged individually. For an artist who has completed a series of such tests, reputation constitutes a form of capital that can be managed in various ways to protect him from the variability of instantaneous evaluations and to more rapidly increase the benefits derived from his fame. Ascribing talent leads to granting a more stable value to the favored artist and his works; insofar as it is equated with the substantial qualities evoked by the artist's name, this value becomes incorporated into the artist's identity and lowers the costs of attracting critical attention (Van Rees and Vermunt, 1996). Rather than defending only artworks, publishers and art dealers support artists themselves. They thus contribute to building artists' careers, and urge consumers to focus their interest and cultural investments on these specifically (White, 1993).

An artist's reputation rating allows for his integration into a system of stratified relations: It is profitable for an artist to be associated with other artists and professionals of at least equivalent standing. The game of relations of exchange and collaboration defines a hierarchy of artist status. In a competitive system of this type, an artist's position in the hierarchy is associated with the reputation he has accumulated. This position becomes a 'status': a hierarchical indicator of quality utilized in a system of relations and exchanges that is more stable than the sum of information contained in the value attributed to an individual's different accomplishments (Podolny, 2005).

Thus, reputation reduces uncertainty, and status reduces uncertainty about the current informative value of reputational signals linked to past 'enactments.' The complete trajectory of reputation accumulation no longer matters when we assume that the value ascribed to an artist – his status, as Podolny puts it – sums up and guarantees all of the information produced and exchanged in the art world concerning that artist's relative quality.

Reputation, of course, can always be contested so long as it continues to be subjected to tests of interindividual competition. Nonetheless, through the mechanisms of self-reinforcement and assortative matching, reputation offers deeply unequal opportunities for the actualization of creative talent, no matter how small and indeterminate this talent's initial distinctive value may be.

Careers and Markets

If we now consider careers and markets together, we see that there are two causes for the excess supply of artistic goods and candidates for an artistic career. First, the number of artists and the variety of artistic production increase faster than demand, because overproduction is a rational response of firms to an uncertain environment. Second, the organization of artistic production on a project-by-project basis, which helps minimize the fixed costs incurred by the schema of rational overproduction, relies heavily on temporary contractual relations with the diverse categories of professionals involved in key operations (from the creation to the distribution of artistic goods). One characteristic of this mode of organization is that it generates an excess supply of labor, available for any project that might solicit it.

As careers of artists unfold outside of the stable organizational structure provided by a firm, they generally follow a trajectory of projects realized through a string of contractual transactions, and for which there exist none of the guarantees associated with ordinary wage labor. As a result, making a career implies getting through elimination tournament stages and that getting through those stages means beating chance.

The forward progress of a career works to reveal underlying the individual's qualities and strengths (those that enabled him or her to get through the stages), qualities and strengths that are unequally distributed among individuals. An individual who succeeds in developing her career in the project-by-project context can enjoy the benefits of a well-established reputation, padding her relational network with contacts that will convey information and work offers through which to increase her skills. This dynamic is particularly influential in occupations where on-the-job learning plays an important role and where the reputation signal is a highly functional means of passing on information in professional network exchanges when it comes to organizing projects.

The importance of 'on-the-job' learning is explained by the heavy exposure of an individual's work to the uncertainty of an extremely turbulent environment; that is, organization on a project basis and the variable degree of control that the individual has over the result of team work. A successful career can be likened to a gradual increase in artist's control over the relatively variable dimensions of her activity and over relations with her environment, in a world where stratification by reputation – unlike in the sciences – is disconnected from stable work organization. It is the very system of artistic labor that creates conditions allowing chance to intervene. Art careers are constructed from one project to the next, and not all projects are equally likely to be successful. Moreover, the individual's work is usually immersed in a collective undertaking whose chances of success are imperfectly correlated with the quality of each team member. The skill or talent of an actress evaluated in terms of personal performance do not fundamentally differ, of course, by whether the film she plays in is a success or failure, but her visibility and the likelihood that she will be involved in more or less promising projects later on depend in large part on the film's success. Organizing work on a project basis introduces strong variability into professional activity and multiplies possible bifurcation points; for example, being called in at a moment's notice to replace the star opera singer, who has caught cold; discovering just the right information on a future project or employment opportunity; landing a role in which, against all expectations, one can reveal one's aptitudes without having ever been cast in that job category before. Project complexity increases the role of chance and in some cases the sequence in which good or bad luck strikes. Moreover, there are few occupations whose practitioners make such frequent recourse to superstitious practices and rituals, the counterpart to that other essential behavior mechanism characteristic of the art world: excessive self-valuing.

In this context, artists' careers can be analyzed as a stochastic process (MacDonald, 1988; Fox and Kochanowski, 2007). Young artists are uncertain about the quality of their work, and their professional engagements constitute a series of evaluative tests (exhibitions, publications, performances, or concerts). If initial evaluations by peers, critics, and members of their reference group are favorable, they will choose to pursue the profession. Artists who achieve lesser or very little success in this first career phase are exposed to a mechanism of cumulative disadvantage. Whether artists remain in the profession with hopes of overcoming the negative effects of a mediocre debut depends on the resources available to them for managing career risks (e.g., multiple jobholding, unemployment benefits, entrepreneurial initiatives, public subsidies, and diversification of activities in order to acquire visibility), as well as on the value artists attribute to the nonmonetary rewards they obtain from their activity, as compared with alternative activities that may offer greater chances for success.

A cohort of artists who have simultaneously entered the market therefore includes a majority of individuals who earn low income and obtain only modest success or rapidly encounter failure, and a minority of professionals who emerge successfully from the competition in the first career phase. Income inequalities reflect the composition effects of the artistic population, the growth of which is driven by two mechanisms. First, the number of entrants hoping to make a career in the arts increases faster than the proportion of those who are relegated and eliminated by the competition. Second, artists in a given cohort who choose to pursue their careers are made to compete with artists from prior cohorts; their position in the competition does not depend on their employment status, as in an organization with seniority and hierarchical positions, but instead on the estimated value of what they produce and on their chances of maintaining their position or raising their status. Thus, careers distribute artistic professionals according to a reputation hierarchy which is based on their past accomplishments and expectations to deliver work and performance of the same quality.

When viewed in instantaneous cross-section, the hierarchy of artist reputations appears to express substantial differences in quality, as revealed by a series of comparisons and competitions. However, as Rosenbaum (1989) has underlined, comparative rankings do not merely reveal unequally distributed qualities, and then select individuals on that basis. These competitions cause divergences in contenders' career paths, despite the fact that their abilities may have been similar. The concentration of earnings and reputations on a very small number of individuals could therefore correspond to gaps in success that are far greater than the gaps in abilities known as talents. The signal emitted by winning a competition helps trigger the process of reputation accumulation.

We readily see the connection between the assortative matching argument and the analysis holding that careers in art advance by way of tournament competition and cumulative advantage. In the course of their early, formative experiences, would-be artists manifest capacities in ways and degrees that vary by individual. The nature of what kind of difference in talent exists between creators who will succeed (more or less lastingly) and others who will not come out as well remains undetermined. Expressed in terms of probability of succeeding, the benefit that hoped-for talent provides early in an artist's career may be weak, but it will be enough to create a small, or not-so-small and in any case perceptible difference with each competitive comparison test, and this in turn will polarize the investments and 'wagers' of system actors (artists themselves, trainers, professionals, patrons, entrepreneurs, critics, and consumers). The intrinsic learning content of work is of a similar origin. There is an optimal profile for increasing one's skills: It is a function of the number and variety of work experiences an artist has and the quality of the collaboration networks she can mobilize as she moves from project to project.

Conclusion

Given the skewed distribution of artists' income, how is it that artists are not deterred from entering such an occupation in growing numbers, nor withdrawing from artistic careers as much as would be expected? In order to face the constraints of a rationed labor market, artists have to learn to manage risky careers. Pioneering empirical research by Baumol and Bowen (1966) found that artists may improve their economic situation in three main ways which are not incompatible and may be combined: Artists can be supported by private sources (working spouse, family, or friends) or by public sources (subsidies, grants and commissions from the state, sponsorship from foundations or corporations, and other transfer income from social and unemployment insurance); they can work in cooperative-like associations by pooling and sharing their income and by designing a sort of mutual insurance scheme; and finally they can hold multiple jobs.

Another feature is essential to understand how artists behave under occupational uncertainty. Rewards in artistic jobs are of two sorts: monetary and nonmonetary, the latter being 'psychic income' flows which have in fact been regarded for a long time as an essential dimension of work. Analytically speaking, every job can be regarded as a bundle of characteristics; wage differentials compensate for more or less attractive work and equalize among workers the total monetary and nonmonetary advantages or disadvantages (Rosen, 1986). The benefits derived from nonmonetary income are not of a uniform magnitude. Individual differences in hiring probabilities are, by contrast, not subject to compensation. And an analysis in terms of equalizing differences requires that we adjust the total amount of these benefits according to the job, the level of professional achievement, and the portfolio management of multiple jobholding (Menger, 1989; Throsby, 1996).

Ironically enough, the nonmonetary value of creative work is the genuine fuel of market competition. The intrinsic motivation of each artist, which renders the absence of instrumental calculation rational in the context of creative activity, must compromise with that of all the others. Yet competition through the unlimited differentiation of artistic projects and artworks forbids the artist from seeing the other simply as a competitor against whom he must measure himself, by comparing easily calibrated qualities that would help determine who possesses them in greater quantity. In order to render such competition acceptable, the artist must invent a peculiar conception of his relationship to the audience for which he designates his work and of the action of market intermediaries: He must relativize or negate the sanctions produced by the orientation of audience preferences, and crudely divide the art markets between sectors of production for a vast audience, on the one hand, and sectors geared toward limited consumption, on the other. Such partitioning of the art markets largely contributes to inscribing the mechanism of competition between artists of the same sector into the more striking vision of a mechanism of hypercompetition that radicalizes the opposition between the ways of doing art: one oriented toward profit, and the other toward free and authentic invention. The deformation of the statistical reality of success and failure, as well as the recourse to salient counter-examples, are among the cognitive mechanisms by which artists invent the illusions necessary to provide long-term motivation for their professional commitment.

See also: Cultural Mediators and Gatekeepers; Cultural Policies in East Asia; Cultural Policy Regimes in Western Europe; Cultural Production in Networks; Culture and Economy; Culture and Networks; Culture, Production of: Prospects for the Twenty-First Century; Markets: Artistic and Cultural; Organizations and Culture.

Bibliography

Accominotti, F., 2008. Marché et hiérarchie: la structure sociale des décisions de production dans un marché culturel. Histoire & Mesure 23, 177–218.
Baron, J., Kreps, D., 1999. Strategic Human Resources. Wiley, New York.
Baumol, W., Bowen, W., 1966. Performing Arts. The Economic Dilemma. The Twentieth Century Fund, New York.
Benhamou, F., 2008. L'économie de la culture. La Découverte, Paris.
Caves, R., 2000. Creative Industries. Contracts between Art and Commerce. Harvard University Press, Cambridge, MA.
Connolly, M., Krueger, A., 2006. Rockonomics: the economics of popular music. In: Ginsburgh, V., Throsby, D. (Eds.), Handbook of the Economics of Art and Culture. Elsevier, Amsterdam, pp. 667–719.
Curien, N., Moreau, F., 2006. L'industrie du disque. La Découverte, Paris.
Ehrlich, C., 1985. The Music Profession in Britain since the Eighteenth Century. Clarendon, Oxford.
Faulkner, R., 1983. Music on Demand. Composers and Careers in the Hollywood Film Industry. Transaction Books, New Brunswick.
Fox, M., Kochanowski, P., 2007. Multi-stage markets in the recording industry. Popular Music and Society 30, 173–195.
Frank, R., Cook, Ph, 1995. The Winner-Take-All Society. The Free Press, New York.
Goode, W., 1979. The Celebration of Heroes. Prestige as a Social Control System. University of California Press, Berkeley.
Graña, C., 1964. Bohemian versus Bourgeois. Basic Books, New York.
Hesmondhalgh, D., 2002. The Cultural Industries. Sage, London.
Hirsch, P., 1972. Processing fads and fashions. American Journal of Sociology 77, 639–659.
Jacobs, D., 1981. Toward a theory of mobility and behavior in organizations: an inquiry into the consequences of some relationships between individual performance and organizational success. American Journal of Sociology 87, 684–707.
Kremer, M., 1993. The O-ring theory of economic development. Quarterly Journal of Economics 108, 551–575.
Lenman, R., 1989. Patronage and the art market in Germany 1850–1914. Past and Present 123, 109–140.
Lopes, P., 1992. Innovation and diversity in the popular music industry, 1969 to 1990. American Sociological Review 57, 56–71.
MacDonald, G., 1988. The economics of rising stars. American Economic Review 78, 155–166.
Menger, P.-M., 1989. Rationalité et incertitude de la vie d'artiste. L'Année Sociologique 39, 111–151.
Menger, P.-M., 2014. The Economics of Creativity: Art and Achievement under Uncertainty. Harvard University Press, Cambridge, MA.
Merton, R., 1968. The Matthew effect in science. Science 159, 56–63.
Montias, J.M., 1996. Le marché de l'art aux Pays-Bas, XVè-XVIIè siècles. Flammarion, Paris.
Peterson, R., Berger, D., 1975. Cycles in symbol production: the case of popular music. American Sociological Review 40, 158–173.
Podolny, J., 2005. Status Signals: A Sociological Study of Market Competition. Princeton University Press, Princeton.
Rosen, S., 1981. The economics of superstars. American Economic Review 71, 845–858.
Rosen, S., 1983. Authority, control and the distribution of earnings. Bell Journal of Economics 13, 311–323.
Rosen, S., 1986. The theory of equalizing differences. In: Ashenfelter, O., Layard, R. (Eds.), Handbook of Labor Economics. North-Holland, Amsterdam, pp. 641–692.
Rosenbaum, J., 1989. Organization, career systems and employee misperceptions. In: Arthur, M., Hall, D., Lawrence, B. (Eds.), Handbook of Career Theory. Cambridge University Press, Cambridge, UK, pp. 337–338.

Stinchcombe, A., 1968. Constructing Social Theories. The University of Chicago Press, Chicago.
Stinchcombe, A., 1986. Stratification and Organization. Cambridge University Press, Cambridge, UK.
Thompson, J., 2010. Merchants of Culture: The Publishing Business in the Twenty-first Century. Polity, Cambridge.
Throsby, D., 1996. Disaggregated earnings functions for artists. In: Ginsburgh, V., Menger, P.-M. (Eds.), Economics of the Arts. North-Holland, Amsterdam, pp. 331–346.
Van Rees, K., 1987. How reviewers reach consensus on the value of literary works. Poetics 16, 275–293.
Van Rees, K., Vermunt, J., 1996. Event history analysis of authors' reputation: effects of critics' attention on debutants' careers. Poetics 23, 317–333.
White, H., 1993. Careers and Creativity. Westview Press, Boulder.
White, H., White, C., 1965. Canvases and Careers: Institutional Change in the French Painting World. Wiley, New York.

Asceticism

Christopher Adair-Toteff, University of South Florida, Tampa, FL, USA; and University of Kent, Canterbury, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S. Abbruzzese, volume 2, pp. 826–829, © 2001, Elsevier Ltd.

Abstract

Asceticism is the practice of self-discipline, self-denial, or self-restraint used in order to reach a particular goal. Its origins are found in the activities of Greek athletes and warriors and its practice is found in almost every religion in the world. It was most prominently found in Early and Medieval Christianity and is closely connected to monks. Max Weber developed the sociological study of asceticism and introduced a new type of asceticism – the 'inner-worldly' asceticism of the early Protestants.

Asceticism is the practice of self-discipline, self-denial, or self-restraint in order to achieve a certain goal. This goal can be a physical one, but more often it is a spiritual one. The term 'asceticism' was used originally to refer to ancient Greek athletes who practiced it in order to increase their physical strength and maintain their mental discipline. The term 'ascetic' can be applied to athletes, warriors, artists, and politicians, but it is most commonly applied to religious figures. Ascetics can be found in almost every religion (Pfister, 1927: p. 571). There have been Jewish, Hindu, Buddhist, and Islamic ascetics, but the term is most frequently used in conjunction with early and medieval Christian ascetics and specifically with monks. Different religions emphasize different ascetic practices, but there are five universal traits: fasting, sexual continence, poverty, seclusion, and self-inflicted pain (Heussi, 1936: p. 13; Kaelber, 2005: p. 527). The body is regarded as being evil; thus, there is a fundamental and critical opposition between body and spirit. For Christians, asceticism is the denial of the body; it is a means to an end. The sole goal is to see God. In order to do this, one believes it necessary to 'flee the world,' to leave it as humanly possible without dying (Harnack, 1904: pp. 84, 95, 101–102, 106). The focus here is on early and medieval Christian asceticism and on the modern sociological study of it.

Origins of Asceticism and Early 'Desert Asceticism'

The term 'asceticism' comes from the ancient Greek term 'askesis' ('ἄσκησις') and was used to mean 'exercise,' 'practice,' or 'training.' Karl Heussi, one of the greatest authorities on asceticism wrote that it originally referred to the athletes' practice of exercise or training for sports competition (Heussi, 1908: p. 727; Derrett, 1995: p. 88, Pinsent, 1995: pp. 211–212). The term 'asceticism' was later used to mean a 'mode of life,' 'philosophical life,' or especially 'religious practice' (Liddel and Scott, 1978: p. 257). Asceticism as a type of practice of discipline is found in Aristotle and this notion was carried over into nineteenth century German scholarship. Rheinhold Seeberg's article 'Askese' states that the philosophical use of the term was based upon the athletic term for 'practice' and Otto Zöckler translated 'askesis' as 'Uebung' meaning 'practice' (Seeberg, 1897: p. 134; Zoeckler, 1863: pp. 1–3; Pfister, 1927: p. 570). However, Zöckler emphasized the practice of morality, so he connects its usage to Plato. While both Plato and Aristotle were concerned with ethical conduct, Aristotle believed in happiness and the good life, whereas Plato extolled the notion of rigorous self-control and emphasized the denial of the importance of the world. Thus, the notion of asceticism as possessing a negative attitude toward the world and the body is found primarily in Plato (Dillon, 1995: pp. 80–83). It is found in *The Republic* with need for discipline and ordering of the soul and the city, and it is found especially in the *Phaedo* with its belief that the body tends to corrupt the soul. The world is something to be devalued and denied, and life is not to be enjoyed but should be viewed as a training for death – both important aspects of asceticism for Christian asceticism.

There is some question whether Jesus should be regarded as an ascetic; however, there is no doubt that Paul is considered to have been one (Heussi, 1936: pp. 16–19). For Paul, there was an 'antithesis between *the spirit* and *the flesh*' and his comments on self-control, on sexuality, and on general conduct, all point toward an ascetic doctrine (Brown, 1988: pp. 47–48; Lohse, 1969: pp. 124–126). However, the earliest Christian ascetics were uninterested in doctrine; they were interested in one thing: seeing God and they withdrew from family, friends, and society in order to accomplish this (Gribomont, 1979: p. 217). It was said that those who were visited by people could not be visited by angels (Adam, 1960: p. 1073). These ascetics were the first 'monks'; meaning those who live alone in the wilderness. While monks have long been associated with early medieval monasteries, ascetic monks predate monasticism by several centuries (Heussi, 1912: p. 427; Rubenson, 1995: p. 49). These ascetic monks who withdrew to the desert to live alone were also referred to as hermits or anchorites (Bertholet, 1930: p. 130; Heussi, 1936: p. 54; Adam, 1960: pp. 1072–1074; Derrett, 1995: p. 97). There is no clear evidence regarding who the first of these 'desert hermits' were, but it is relatively clear that they began appearing in the late third and early fourth centuries (Heussi, 1936: pp. 70–76, 109–113). They did not belong to any church and they followed no doctrines or rules; they mistrusted the increasingly powerful Church and they rejected its dictates (Völter, 1900: p. 9; Harnack, 1904: pp. 94, 102–103). They considered the world as a place of temptation, and they regarded the desires of the body as evil. They rejected everything worldly – possessions, marriage, and even honor – in order to live their lives in the service of God (Grützmacher, 1903: p. 217; Harnack, 1906: p. 97). These

monks withdrew from the human world in order to create a new and special world devoted solely to God (Heussi, 1912: p. 427, 1936: pp. 40, 54, 55).

The world was dualistic; it had been created by God but had become filled with devils (Harnack, 1906: p. 113; Heussi, 1936: p. 28). There was the possibility to live a life devoted to God, but the reality was that this life was filled with temptations. These temptations most often came in the forms of devils and demons, and they were everywhere (Harnack, 1906: p. 113; Rubenson, 1995: pp. 51, 53; Ware, 1995: p. 7). While the belief in the devil and demons is found in the New Testament, the belief in them grew considerably during the second, third, and fourth centuries (Harnack, 1906: pp. 108–111). The monks were convinced that these demons sought to inhabit humans and to tyrannize and plague them; thus, it was absolutely imperative that the monks struggled against them (Weiß, 1897: p. 411). Although some ascetics saw devils and demons, most were aware that they were invisible. Satan would tempt them in their own thoughts and their own words (Heussi, 1936: pp. 257, 262–265).

The dualism of God and the world is replicated in the dualism of the soul and body: the soul is spiritual and good, whereas the body is sensible and bad. Asceticism was the struggle against one's own body and the devil (Seeberg, 1897: p. 135; Pfister, 1927: p. 571). These ascetic monks fought against the devil and all of them did so with the power of prayer. But, many of them employed various ascetic practices to drive the demons out and away (Seeberg, 1897: p. 137; Harnack, 1906: p. 110). Some tried to do this by determining how they lived. Some lived in caves, in mountains, and in deserts; the point was to live in isolation. Some ascetics went barefoot and some wore rags. Some wore various kinds of clothing intended to inflict pain if not torture. Some wore nothing at all. Some ascetics ate only roots, others vegetables, others ate insects, but all of them fasted. Some slept on the ground, some standing upright, while others tried to do without any sleep. Some beat themselves, some whipped themselves, and some carried crosses and chains to remind themselves of the temptations of the body and to imitate the suffering of Jesus (Harnack, 1906: p. 115; Zoeckler, 1863: pp. 17–80; Heussi, 1936: pp. 221–229).

It was Tertullian who emphasized the importance of the ascetic life for God and it was he who first wrote about it (Heussi, 1908: p. 727; Preuschen, 1913: pp. 1145–1146). And, it was Antonius who provided us with the earliest form and goals of the ascetic life (Heussi, 1927: p. 399). The ascetic life was solitary, but it is a mistake to think that there were only a few people who left family and friends to settle in the desert. By the beginning of the fifth century, there were approximately 5000 hermits living in one rather small area in Egypt (Brown, 1988: p. 215). It is also a mistake to believe that monasteries developed only in the Middle Ages; the ascetic monks 'fled the world' and its cities only to create a new 'city.' Monasteries were already being formed by the third century (Brown, 1988: p. 217).

Monastic Asceticism

During the third century, a number of changes began to affect the Church. Greek philosophical concepts began to be incorporated into Christian doctrine and the Church began to organize itself (Gribomont, 1979: p. 221). In addition, the notion of asceticism began to play an increasingly important role in the development of Christian ways of life, specifically the monastic life (Harnack, 1906: p. 304; Heussi, 1908: p. 727). Instead of living alone ('kata monas') as anchorites or in small groups, the ascetic monks began to organize and live in communal life ('koinos bios') (Heussi, 1936: p. 115). By the fourth century, monks were living together within walled communities (Heussi, 1930: p. 135). The point of the monasteries was to 'wall off' the monks from the outside world (Arens, 1959: p. 1671). But, it would be a mistake to regard this as a triumph for the Church, because the monks were distrustful of its increasing authority and its expanding hierarchy. The monks were also wary of the Church's perceived increasing accommodation to the world. These newly formed cloisters were intended to further the monks' service to God and to introduce structure to their lives. Their lives continued to be devoted to prayer and sacrifice, but by the beginning of the sixth century, study and work were not only encouraged but expected (Heussi, 1936: pp. 214–220). The earlier anchorites did not beg; they worked. They needed to work to provide for their own food, but they did this individually and with some degree of regularity (Heussi, 1936: pp. 214–217). In this respect, the importance of Benedict of Nursia cannot be underestimated, for he radically transformed the monks' ascetic lives. It was Benedict who was largely responsible for replacing the older, harsh ascetic practices with a highly structured and extremely disciplined life. Now, monks not only lived the life of contemplation ('vita contemplativa'), but also the life of activity ('vita activa'). And, it was Benedict who gave them the new way of monastic life by following his 73 'Rules' for conduct (Arens, 1959: p. 1672; von Lilienfeld, 1993: p. 169; Köpf, 1998: p. 836). Benedict was by no means the first to try to instill order and regularity within the monastic community, but his 'Rules' were probably the most definitive and likely the most influential of any (Prinz, 1980: pp. 34–35). Within two centuries, Benedictine monasteries were being established through many parts of Western Europe – all with the intention of following Benedict's rigorous rules (Prinz, 1980: p. 38). The scope of Benedict's influence is spelled out clearly with the claim that the entire development of Western monasticism from the early Middle Ages cannot be understood without consideration of his writings (Lohse, 1969: p. 226). Bernard McGuinn maintained that Benedict's 'Rules' is "the single most important document in the history of western monasticism" (McGuinn, 1999: p. 27). And, the central focus of the 'Rules' is that the monk is to devote his entire being to God (Lohse, 1969: p. 229).

In Chapter One of his 'Rules,' Benedict discusses the four types of monks; but his interest is primarily in the solitary anchorite, who with God's help, learned to fight the devil, and the Coinobites, who lived together. Benedict evidently believed that establishing a monastery and determining rules was the best way to serve God. While prayer was extremely crucial, Benedict also emphasized the importance of the active life, which included exercises intended to help the monk in his 'ascetic combat' (Fry, 1980: p. 95). In the second and third chapters, Benedict discusses the calling of the Abbot and his function in the monastery. In Rules 4–73, Benedict details the orders that regulate everything about the ascetic life in the monastery, from

the layout of the buildings to the monks' diet to the times for prayer, and especially regarding the importance of work. In Rule 48, Benedict explains that 'idleness is the enemy of the soul' and he instructs the monks to engage in manual labor (Fry, 1980: p. 349). Everyone is to work; even the sick and infirm should be given suitable work (Fry, 1980: p. 253).

The earlier Greeks and Romans had viewed work as lacking intrinsic value, and this certainly applied to physical labor. But, Augustine did not view physical work as a punishment for sin; rather, he believed that it was part of God's creation (Prinz, 1980: p. 68). But, it was Benedict who not only wrote of the necessity of work, but also emphasized its importance (Fry, 1980: p. 70). And, it is not just the importance of work, but the increasing sense of rational order that leads to Max Weber and his social theories of asceticism.

Social Theories of Asceticism

Max Weber was not a religious person and he had no theological interest in the conception of asceticism. He was, however, an astute investigator of human behavior and he sought to understand it within the full range of economics, politics, and sociology. While he examined a wide range of issues, he concentrated on the one defining characteristic of the Western world – its increasing rationality. In his most famous work, *The Protestant Ethic and the Spirit of Capitalism*, Weber attempted to draw a correlation between asceticism and capitalism. He did not try to prove this and he acknowledged that what he offered was only a sketch. However, the *Protestant Ethic* does contain a fascinating study of one type of asceticism – 'inner-worldly asceticism' (Köpf, 1998: p. 837; Kaelber, 2005: p. 529). This type of asceticism differs in fundamental ways from traditional Catholic asceticism: It is Protestant, it is 'world rejecting,' and it is 'active.'

Asceticism had been closely associated with Catholicism in general and with monks in particular. Weber's analysis of Protestant asceticism showed that there were clear differences between the types of asceticism. First, while monks were not enthusiastic about Church Doctrine and they occasionally even rebelled against it, they were devoted to the Abbot's authority and they were comfortable in the monastic environment. In contrast, the early Protestant ascetics rejected any type of doctrinal authority and believed only in the authority of the Bible. Second, while the early Protestants lived in communities and participated in a Church, they continued to maintain that religious life was not communal but was intensely personal. This is not to say that Weber believed that all early Christians adhered to the same beliefs. On the contrary, for him, one of the most distinguishing differences between Luther and Calvin was the differing conceptions of God. Luther's God is closer to the Catholic God in that Luther's notion is of a loving father, whereas Calvin's deity resembles the angry and vengeful God of the Old Testament (Weber, 1996: pp. 58–61). Moreover, Luther still tended to believe in the Catholic tradition of sin and redemption, but Calvin rejected such notions. For him, God ordained that a select few were chosen, whereas the majority was damned. Calvin's Doctrine of Double Election held that God had preordained everyone; as a result no one could be certain whether he or she was among the Elect. This uncertainty caused the Calvinists extreme anxiety and pushed them into extreme inner loneliness. Unlike Catholics, the Calvinists could not seek comfort in a sacrament or a church (Weber, 1996: pp. 62–66). What it did prompt them to do is to work, with the conviction that work was solely for the greater glory for God (Weber, 1996: pp. 66–71, 73–74, 78).

The second major difference is that the Catholic monks sought to 'flee the world,' whereas the early Protestants 'rejected the world.' This notion is not only important for Weber's *Protestant Ethic* but also crucial for Weber's entire sociology of religion. Thus, Stefan Breuer is correct to call it world-rejecting ('Weltablehnung') a "key concept in Weber's sociology of religion" (Breuer, 2001: p. 227; Mulert, 1913: p. 1914; Weber, 2001: p. 324). While Catholics and early Protestants regarded the world as a place of temptation, only Catholic ascetics sought to flee from it. In contrast, ascetic Protestants sought only to deny certain aspects of it and in certain respects even had a certain 'world-enjoyment.' Weber did not mean that they relished life, but he did believe that they 'valued life as a task' (Weber, 1996: p. 48). They did not work for pleasure; rather, they viewed their work in the world as their service for God. They did not create wealth for themselves; they did it solely for the benefit of God. Weber notes this paradox applies to the rational ascetics, but he also points to another paradox that pertains to the monks: as individuals they live lives without possessions, but the monasteries and churches that they inhabit and serve are filled with them (Weber, 1989: pp. 489–490).

Weber acknowledges certain similarities between the ascetic monks and the ascetic Protestants and he allows that Benedict and his successors were responsible for introducing regulations into monastic life (Weber, 1996: p. 78). However, the monks did not lead a rational life and they withdrew behind the walls of the monastery in order to 'flee the world.' In contrast, the early Protestants were 'inner-worldly' ascetics; that is, they chose to remain in and participate within the world, but as ascetics. Weber emphasized that these ascetics practiced rigorous and systematic self-control and he stressed that this self-control was rational and methodical (Weber, 1996: pp. 77–80).

The third major difference is Weber's distinction between asceticism and mysticism. Weber never provided a complete theory of asceticism and he chose to contrast asceticism with mysticism (Adair-Toteff, 2010: p. 109). This should not be taken to mean that Weber did not understand or appreciate mysticism, because he did (Adair-Toteff, 2002: pp. 339, 352–353). In the *Protestant Ethic*, Weber noted that the Lutherans were like the Catholic mystics in that the ultimate goal for both was the 'unio mystical' – the 'mystical union' with God. And, he stressed that both engaged in passive contemplation in order to achieve this union (Weber, 1996: pp. 71–72). In his later 'Zwischenbetrachtung' Weber sets out his 'theory of the steps and directions of religious world-rejection' and it is there that he spells out in more detail the opposition between active asceticism and contemplative mysticism (Weber, 1989: p. 481). In fact, he regards the contrast between the ascetic and the mystic as being a 'polar opposition.' He maintains that the ascetic regards himself as a 'tool' or 'instrument' of God and it is his duty to be active. In contrast, the mystic is a 'vessel' and it is imperative that he remain passive and quiet in order to listen to God (Weber, 1989: p. 482). Weber acknowledges that the distinction between ascetic and mystic can collapse but he insists that the

opposition persists because the former is inherently rational while the latter is predominately irrational (Weber, 1989: pp. 482–483). The rational ascetic seeks to control the world by imposing form; the irrational mystic seeks to flee the world into a totally 'formless experience' (Weber, 1989: pp. 501, 507).

The account of asceticism provided by Weber seems compelling, but it can be challenged on several fronts. First, his 'polar opposition' between ascetics and mystics is questionable, both in a historical and theoretical way. Throughout history, there have been individuals who have been regarded as being 'ascetic mystics' beginning with Paul (Romans 8: pp. 7–8; Derrett, 1995: p. 93; Lohse, 1969: p. 172). Weber was quite familiar with the medieval German mystic tradition, which was attracting scholarly attention. Heinrich Seuse, a follower of Meister Eckhart, was a mystic who wrote about the 'unio mystica' and who adopted many of the practices of the 'desert ascetics,' including fasting, wearing rough and irritating clothing, and sleeping on freezing cold surfaces (Wentzlaff-Eggebert, 1947: pp. 6, 18, 28).

Asceticism and mysticism are also closely connected theoretically. Evelyn Underwood, whose *Mysticism* first appeared in 1911, wrote about the connection between ascetic self-discipline and the striving for mystical visions (Underwood, 1990: pp. 224–277). Karl Heussi, the leading authority on asceticism, maintained that asceticism is the preparation for the reception of God's grace and the unification with him (Heussi, 1936: pp. 45, 266–269). And, Bernard McGuinn, the author of the multivolume work on mysticism frequently connects mysticism with asceticism (McGuinn, 1995: pp. 109, 110, 137, 144–148; McGuinn, 1999: pp. 26–29).

The preceding criticism focused on Weber emphasizing the differences between asceticism and mysticism, whereas there is another one that addresses Weber minimizing the differences between monastic life and industrial capitalism. This complaint is found in Hubert Treiber and Heinz Steinert's *Die Fabrikation des zuverlässigen Menschen* (Treiber and Steinert, 1980) but it receives its fullest discussion in Lutz Kaelber's *Schools of Asceticism*. Kaelber's account is important because he correctly sets out Weber's thesis about rational asceticism by emphasizing the two fundamental traits: (1) 'restraint and self-control' and (2) 'the systematic control of this plan' (Kaelber, 1998: p. 3). But, it is even more important because he examines two medieval ascetic movements that basically embody these traits: the Waldesians and the Catharists. Kaelber does not condemn Weber for ignoring these; instead, he compliments Weber for at least acknowledging that there are certain similarities between monastic life and the rational ascetic life and he provides an account of western Christianity that Weber was unable to do because of his premature death. Weber may not have fully understood asceticism, but he clearly recognized many of its critical features. Moreover, he was the first to understand and appreciate the revolutionary power of 'inner-worldly' rational asceticism.

Conclusion

The practice of asceticism is found in various forms throughout history, but its two most studied forms have occurred in Christianity. It was found first in the 'world-fleeing' form in the early desert ascetics and then in medieval monks, and then in the 'inner-worldly' form of some of the early Reformers. While most of the adherents to these forms believe in a dualism of soul and body, the 'world-fleeing' ascetic denigrates the world while the 'inner-worldly' ascetic accepts the world and works in it in order to praise God. All reject the notion of the material enjoyment of wealth, but the Puritan ascetic believed that wealth gained through hard work was a sign that one was a member of the Elect.

The greatest impetus for someone to become an ascetic monk was the realization of one's own sense of sin. This is not the recognition of doing one or two sinful deeds; rather, it was the steady and vividly self-conscious sense of continual sin (Heussi, 1936: p. 190). Without that sense of sin, there is little reason to engage in asceticism. It was Adolf von Harnack, one of the greatest of the Protestant Church historians, who indicated the reasons for the decline in asceticism. He suggested that when the belief in sin and guilt is replaced by the Paulinian-Augustinian trinity of 'faith, love, and hope,' asceticism no longer has such a compelling power over people (Harnack, 1916: pp. 156–158). But, for those who despise the body and renounce the world, asceticism will always be a way of life.

See also: Christianity, Origins of; Monasticism: West and East; Mysticism; Ritual; Sacrifice; Self-Regulated Learning; Spirituality.

Bibliography

Adair-Toteff, Christopher, 2002. Max Weber's mysticism. Archives Européenes de Sociologie XLIII (3), 339–353.

Adair-Toteff, Christopher, 2010. Max Weber's notion of asceticism. Journal of Classical Sociology 10 (2), 109–122.

Adam, A., 1960. Mönchtum II. Christliches Mönchtum. In: Die Religion in Geschichte und Gegenwart, Dritte, völlig neubearbeitet Auflage, Vierter Band. J.C.B. Mohr (Paul Siebeck), Tübingen, pp. 1072–1081.

Arens, F., 1959. Kloster I. Religionsgeschichtlich/II. Im Christentum. In: Die Religion in Geschichte und Gegenwart, Dritte, völlig neubearbeitet Auflage, Dritter Band. J.C.B. Mohr (Paul Siebeck), Tübingen, pp. 1671–1676.

Bertholet, Alfred, 1930. Mönchtum. 1. Religiongeschichtlich. In: Die Religion in Geschichte und Gegenwart. Zweite, völlige neubearbeite Auflage. J.C.B. Mohr (Paul Siebeck), Tübingen, Band IV, pp. 130–134.

Breuer, Stefan, 2001. In: Kippenberg, Hans G., Riesebrodt, Martin (Eds.), Weltablehnung, pp. 227–240.

Brown, Peter, 1988. The Body and Society. Men, Women, and Sexual Renunciation in Early Christianity. Columbia University Press, New York.

Derrett Jr., M.J. Duncan, 1995. Primitive Christianity as an ascetic movement. In: Wimbush, Valantasis (Eds.), Asceticism. Oxford University Press, New York, pp. 108–118.

Dillon, John M., 1995. Rejecting the body, refining the body: some remarks on the development of platonist asceticism. In: Wimbush, Vincent L., Valantasis, Richard (Eds.), Asceticism. Oxford University Press, New York, pp. 80–87.

Fry, Timothy (Ed.), 1980. The Rule of St. Benedict. in Latin and English with Notes. The Liturgical Press, Collegeville, Minnesota.

Gribomont, Jean, 1979. Askese IV. Neues Testament und Alte Kirche. In: Theologische Realenzyklopädie, Band 4. Walter de Gruyter, Berlin, New York, pp. 204–225.

Grützmacher, G.K., 1903. Mönchtum. In: Realencyklopädie für protestantische Theologie und Kirche, dritte Auflage, Band 13. J.C. Hinrichs'sche Buchhandlung, Leipzig, pp. 214–235.

Harnack, Adolf, 1904. Das Mönchtum, seine Ideale und seine Geschichte. In: Reden und Aufsätze, Erste Auflage. J. Ricker' sche Verlagsbuchhandlung (Alfred Töpelmann), Gieszen, pp. 81–139.

Harnack, Adolf, 1906. Mission und Ausbreitung des Christentums in den ersten drei Jahrhunderten, Erster Band. J.C. Hinrichs'sche Buchhandlung, Leipzig.

Harnack, Adolf, 1916. Askese. In: Aus der Friedens- und Kriegsarbeit. Reden und Aufsätze, Neue Folge, Dritter Band. Verlag von Alfred Töpelmann, Giessen, pp. 141–161.

Heussi, Karl, 1908. Askese. II. Kirchengeschichtlich. In: Die Religion in Geschichte und Gegenwart, Band 1. Verlag von J.C.B. Mohr (Paul Siebeck), Tübingen, pp. 727–729.

Heussi, Karl, 1912. Mönchtum. In: Die Religion in Geschichte und Gegenwart, Band 4. Verlag von J.C.B. Mohr (Paul Siebeck), Tübingen, pp. 427–446.

Heussi, Karl, 1927. Antonius, 1.der Große. In: Die Religion in Geschichte und Gegenwart, Zweite, völlig neubearbeitete Auflage, Erster Band. Verlag von J.C.B. Mohr (Paul Siebeck), Tübingen, p. 399.

Heussi, Karl, 1930. Mönchtum II. Christliches Mönchtum. In: Die Religion in Geschichte und Gegenwart, Zweite, völlig neubearbeitete Auflage, Vierter Band. Verlag von J.C.B. Mohr (Paul Siebeck), Tübingen, pp. 134–145.

Heussi, Karl, 1936. Der Ursprung des Mönchtums. J.C.B. Mohr (Paul Siebeck), Tübingen.

Kaelber, Lutz, 1998. Schools of Asceticism. Ideology and Organization in Medieval Religious Communities. University of Pennsylvania Press, University Park.

Kaelber, Walter O., 2005. Asceticism. In: The Encyclopedia of Religion, second ed., vol. 1. Thomas Gale, Detroit, pp. 526–530.

Kippenberg, Hans G., Riesebrodt, Martin, 2001. In: Kippenberg, Hans G., Riesebrodt, Martin (Eds.), Max Webers 'Religionssystematik'. J.C.B. Mohr (Paul Siebeck), Tübingen.

Köpf, Ulrich, 1998. Askese IV. Kirchengeschichtlich. In: Religion in Geschichte und Gegenwart, Vierte, völlig neubearbeitete Auflage, Band I. J.C.B. Mohr (Paul Siebeck), Tübingen, pp. 834–837.

Liddel, Henry George, Scott, Robert, 1978. A Greek-English Lexicon. Revised and augmented throughout by Sir Henry Stuart Jones. Clarendon Press, Oxford.

von Lilienfeld, Fairy, 1993. Mönchtum II. Christlich. In: Theologische Realenzyklopädie, Band 23. Walter de Guyter, Berlin, New York, pp. 150–193.

Lohse, Bernhard, 1969. Askese und Mönchtum in der Antike und der alten Kirche. R. Oldenbourg, München, Wien.

McGuinn, Bernard, 1995. Asceticism and mysticism in late antiquity and the early middle ages. In: Wimbush, Vincent L., Valantasis, Richard (Eds.), Asceticism. Oxford University Press, New York, pp. 58–74.

McGuinn, Bernard, 1999. The presence of God: a history of Western mysticism. In: The Growth of Mysticism. Gregory the Great through the 12th Century, vol. II. Crossroads Publishing Co., New York.

Mulert, H., 1913. Welt usw.: V. Weltbejahung und Weltverneinung. In: Religion in Geschichte und Gegenwart, Band V. Verlag von J.C.B. Mohr (Paul Siebeck), Tübingen, pp. 1912–1920.

Pfister, Fr, 1927. Askese: I. Religionsgeschichtlich. In: Die Religion in Geschichte und Gegenwart, Zweite, völlig neubearbeitete Auflage, Band 1. Verlag von J.C.B. Mohr (Paul Siebeck), Tübingen, pp. 570–574.

Pinsent, John, 1995. Ascetic moods in Greek and Latin Literature. In: Wimbush, Vincent L., Valantasis, Richard (Eds.), Asceticism. Oxford University Press, New York, pp. 211–219.

Preuschen, E., 1913. Tertullian. In: Die Religion in Geschichte und Gegenwart, Band IX. Verlag von J.C.B. Mohr (Paul Siebeck), Tübingen, pp. 1145–1147.

Prinz, Friedrich, 1980. Askese und Kultur. Vor- und frühbenediktinisches Mönchtum an der Wiege Europas. C.H. Beck, München.

Romans, The Holy Bible. Rubenson, Samuel, 1995. Christian Asceticism and the Emergence of the Monastic Tradition. In: Wimbush, Vincent L., Valantasis, Richard (Eds.), Asceticism. Oxford University Press, New York, pp. 49–57.

Seeberg, Reinhold, 1897. Askese. In: Realencyklopädie für Protestantische Theologie und Kirche, Dritte Auflage, Band 2. J.C. Hinrich'sche Buchhandlung, Leipzig, pp. 134–142.

Treiber, Hubert, Steinert, Heinz, 1980. Die Fabrikation des zuverlässigen Menschen. Über die "Wahlverwandschaft" von Kloster- und Fabrikdisziplin. Heinz Moos Verlag, München.

Underwood, Evelyn, 1990. Mysticism. The Preeminent Study in the Nature and Development of Spiritual Consciousness. Image Books/Doubleday, New York.

Völter, Daniel, 1900. Der Ursprung des Mönchtums. Ein Vortrag. Verlag von J.C.B. Mohr (Paul Siebeck), Tübingen.

Ware, Kallistos, 1995. The way of the ascetics: Negative or Affirmative?. In: Wimbush, Vincent L., Valantasis, Richard (Eds.), Asceticism. Oxford University Press, New York, pp. 3–15.

Weber, Max, 1989. In: Schmidt-Glintzer, Helwig, Kolonko, Petra (Eds.), Die Wirtschaftsethik der Weltreligionen. Konfuzianismus und Taoismus. Schriften 1915-1920. J.C.B. Mohr (Paul Siebeck), Tübingen. Max Weber Gesamtausgabe. I/19.

Weber, Max, 1996. In: Lichtblau, Klaus, Weiß, Johannes (Eds.), Die protestantische Ethik und der 'Geist' des Kapitalimus, Zweite Auflage. Beltz Athenäum. Neue Wissenschaftliche Bibliothek, Weinheim.

Weber, Max, 2001. In: Kippenberg, Hans G., Schilm, Petra, Niemeier, Jutta (Eds.), Wirtschaft und Gesellschaft. Die Wirtschaft und gesellschaftlichen Ordnungen und Mächte. Nachlaß, Religiöse Gemeinschaften, Teilband 2. J.C.B. Mohr (Paul Siebeck), Tübingen. Max Weber Gesamtausgabe. I/22–2.

Weiß, Johannes, 1897. Dämonische. In: Real encyklopädie für protestantische Theologie und Kirche, Dritte Auflage, Band 4. J.C. Hinrichs'sche Buchhandlung, Leipzig, pp. 410–419.

Wentzlaff-Eggebert, F.W., 1947. Heinrisch Seuse. Sein Leben und seine Mystik. Jan Thorbecke Verlag, Lindau.

Wimbush, Vincent L., Valantasis, Richard (Eds.), 1995. Asceticism. Oxford University Press, New York and Oxford. With the assistance of Gay L. Myron and William S. Love.

Zoeckler, Otto, 1863. Kritische Geschichte des Askese. Ein Beitrag zur Geschichte Christlichen Sitte und Cultur. Verlag von Heyden und Zimmer, Frankfurt am Main und Erlangen.

Asia, Sociocultural Aspects: China

Frank N Pieke, Leiden University, RA Leiden, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by H-D. Evers, volume 22, pp. 14661–14666, © 2001, Elsevier Ltd.

Abstract

China's modernization brings to the fore the challenges of nationalism, state building, and social integration. This article traces the development of Chinese society and culture since the dissolution of the Qing empire in 1911 through the collective period after 1949 and the reforms since 1978. The article focuses in particular on the changes in the social, cultural, political, and ethnic diversity of Chinese society; the impact of market capitalism on life, work, and social organization; the impact of China on the world and the world's impact on China; and the challenges currently facing Chinese society.

Introduction

Modern China is a country that is heir to the vast empire conquered by a non-Chinese group called the Manchus, whose origins were in what is now China's northeast, or Manchuria. The Qing dynasty that they established lasted from 1644 to 1911. The Manchus incorporated in their empire large non-Chinese territories, such as Manchuria, Mongolia, Xinjiang, Tibet, and Taiwan. All that these territories had in common was some form of allegiance to the emperor, regardless of their culture, ethnic orientation, religion, or even political system.

The heritage of empire makes contemporary China a political entity that struggles to find a political form. This task is made even more difficult by the legacy of Western and Japanese imperialism and the Cold War during the last two centuries. Not only Taiwan (asserting itself as an independent republic), Hong Kong, and Macao (with the status of special administrative region) but also Tibet, Xinjiang, Inner Mongolia, Ningxia, and Guangxi (all autonomous regions) claim or have been given various degrees of autonomy or independence from the government of the People's Republic of China. As a modern nation-state, China therefore is anything but a finished product, and is probably best described as a work in progress.

China is also a civilization, connected to but by no means coterminous with any contemporary political unit that calls itself China. Viewed from this angle, China and being Chinese are not just a matter of belonging, but also of becoming. This has at least two aspects. Chinese culture, goods, influence, and people have for centuries fanned out far beyond the reach of the Chinese state, mixing and mingling with other cultures and civilizations in a variety of settings all across the world.

This article asks three broad questions. What unites and what divides China as a country, society, and culture? What are the distinguishing features of China's developing modernities? What is the impact of globalization on China, and what impact on the world do Chinese society and culture have? Below, each question will be discussed in a separate section. The article will draw mainly on recent anthropological research, supplemented with work by sociologists, social geographers, historians, and political scientists.

Anthropologists of China have long been drawn to China, fascinated by its scale and complexity. Yet for most of the twentieth century, anthropologists mainly tended to view China through village studies, communities that were seen as largely self-contained and bound together by the bonds of family (major ethnographies include Ahern, 1973; Cohen, 1976; Lin, 1998; Wolf, 1968). Connections between these communities existed, of course, but were seen as largely limited to traditional patterns of trade (Skinner, 1964), intervillage marriage and supralocal kinship organization (Freedman, 1958, 1966; Fried, 1953), and religious networks (Sangren, 1987). Anthropologists for the most part ignored the impact of the traditional or early modern state, with the exception of some work on Chinese towns and local elites (Fei, 1953) and the interactions between Chinese popular, official, and organized religion (Feuchtwang, 1977; Yang, 1961). This tendency was strengthened by the fact that, until the early 1980s, the People's Republic did not allow foreign researchers in. China anthropologists had to be content with proxy fieldwork in Taiwan or Hong Kong or among overseas Chinese, where they sought to reconstruct prerevolutionary Chinese society from artificial slices of culture carefully dissected from complex and rapidly modernizing societies outside China itself. (An excellent overview and discussion of the anthropology of China of this period is given by Steven Sangren in his article on China in the first edition of this encyclopedia (Sangren, 2001).) It has only been in the last 30 years that China anthropologists have squarely dealt with contemporary society and culture in the People's Republic.

Unities and Diversities

China is a continent-sized country. Since the first unification of China in 221 BCE, ruling dynasties expanded and unified their empires either directly through warfare, tribute relationships, and the recognition of indigenous rulers or indirectly because of the effect of migration, intermarriage, trade, road building, land reclamation, and cultural assimilation. As a result, over the centuries China incorporated a rich variety of peoples and cultures. These peoples adopted aspects of Chinese culture, such as ancestor worship, cosmology, religion, or gender relations, gradually translating differences between Chinese and non-Chinese into part of the endless diversity of Chinese culture. Other groups, however, stayed beyond imperial rule, either by choice or exclusion. These non-Chinese 'barbarians'

living within or immediately adjacent to the empire owed neither tax nor corvée labor, but could also not count on imperial protection and the opportunities for upward mobility that Confucian education offered their elites. These processes have been best documented for parts of China that were firmly incorporated during the last dynasty of the Qing, such as Taiwan or Yunnan (Brown, 2004; Giersch, 2006; see also the articles in Crossley et al., 2006).

Early Chinese nationalists around the turn of the twentieth century imagined the new Chinese nation as the successor of the empire, thus opting for a 'maximal' definition of the nation and its territory that required from the onset recognition and accommodation of non-Chinese groups and territories in the new nation (Crossley, 2005). The Chinese nation built after the fall of the Qing in 1911 therefore incorporates two very different, at times even conflicting, narratives. The Chinese nation was to be the home of the Chinese, a safe haven and instrument to undo the injustices of Manchu alien rule and 'national humiliation' by Western and Japanese imperialists after the Opium Wars. The nation would unify all Chinese and restore China's rightful place in the world. But the Chinese nation was also imagined as the reincarnation of an empire consisting of groups no longer glued together by dynastic rule, but by the cultural, economic, and political dominance of the Han majority.

After coming to power in 1949, the Chinese Communist Party (CCP) continued to embrace these two conceptualizations of the nation. The CCP views China as a composite 'multination country' (*duo minzu guojia*), consisting of 55 'minority nationalities' (*shaoshu minzu*) and one majority nationality, the Han. The identification of these nationalities took place mostly in the 1950s; since then they have remained largely fixed with individual membership determined by birth. Minorities were given certain rights and privileges (including the right to practice their own culture and religion) and also a degree of self-administration (autonomy) of the provinces, prefectures, or counties where they lived (Heberer, 1989).

In contradistinction to the many different minorities, China's majority population is universally called Han. Yet their culture and language display a rich diversity, predominantly rooted in regional variety (Cantonese, Fujianese, Pekingese, Shanghainese, etc.). In addition, Han Chinese are also divided along ethnic (cf the Hakka of Southeastern China, see Constable, 1996) or religious (cf Muslim Chinese, see Gladney, 2004) lines that uncomfortably straddle the divide between the Han majority and the non-Han minorities. China's minority policies thus not only emphasized the minority–majority distinction, but also put an almost total ban on the political expression of regional or cultural differences among the Han.

The recognition of minority nationalities is part of an elaborate administrative system that was set up after 1949 to regulate and manage forms of diversity: religious, cultural, and ethnic. This system is supervised by the CCP through its United Front Department. After the turmoil of the Cultural Revolution, the system was reinstated and even strengthened and it continues to play an important role until the present day, turning forms of diversity into bureaucratic statuses that come with highly regulated rights, duties, privileges, and restrictions. In addition to 56 nationalities, China thus has practitioners of five recognized religions (Catholicism, Protestantism, Islam, Buddhism, and Taoism; Yang, 2012); compatriots from Hong Kong, Macao, and Taiwan; overseas Chinese; returned overseas Chinese and their dependents (Barabantseva, 2011); and even foreign friends and guests (Brady, 2003).

During the collective period (1949–78) society was forced into the straightjacket of a socialist planned economy. Industry, agriculture, trade, and services were nationalized or collectivized. The establishment of the household registration system, unified job allocation, and rationing of foodstuffs and many other consumption goods tied people to one place: the rural population in their villages as members of their collective team, brigade, and commune (Heins Potter and Potter, 1990) and the urban population in their town or city of residence as members of their 'work unit' (*gongzuo danwei*; Bray, 2005). Except for times of great upheaval, such as the famines of the Great Leap Forward between 1958 and 1962 or the political mayhem of the Cultural Revolution between 1966 and 1969, mobility was subject to bureaucratic allocation, either as part of the economic plan or through more informal arrangements made by individual work units or collectives.

After 1949, diversity became much less prominent and politically relevant in Chinese society than it had been before. Minorities, religions, overseas Chinese, and foreigners were contained in their own separate places and spaces. Social life mainly took place in the main building blocks of society, urban neighborhoods and work units and rural villages and collectives. Due to the restrictions put on mobility, Chinese from different places and backgrounds mixed and met much less frequently. Furthermore, with the establishment of a unified administration under the leadership of the Communist Party, native place (with its associated linguistic, cultural, and religious aspects) ceased to be a principle of organization and vehicle of political participation among city populations (about pre-1949 China on this point, see Rowe, 1989).

This picture of strict regulation and containment of social life in general and diversity and mobility more in particular changed after the start of reform in 1978. This happened slowly at first, but quickly gained speed, particularly during the second phase of reform that started in 1992–93. In the countryside, the reforms unleashed the enormous entrepreneurial potential of China's rural population that had remained bottled up for 30 years. In the early 1980s, land was decollectivized, giving farmers not only much better incentives to step up agricultural production, but also the freedom to decide whether to stay on the farm, start another business, work for a local enterprise, or move elsewhere in search of work or business opportunities. In the 1980s, millions of farmers found their way to China's cities, working in construction, manufacturing, petty trade, and other jobs that the established city population, entitled as they were to state employment, did not have or did not want to do. In the 1990s, reforms targeted the urban sector and a further opening up to the world economy. The result was an urban investment boom and the rapid growth of the export-processing sector. At the same time, the rural population suffered from declining agricultural prices, land expropriation for development projects, deteriorating services, and increasingly rapacious government (Bernstein and Lü, 2003). As a result and despite an initial surge in urban unemployment, the 'floating population' (*liudong renkou*) of rural–urban migrants grew well to over 100 million, reaching around 200 million, or about one-sixth of China's total population of 1.3 billion, in the 2000s.

Members of the floating population are second-class citizens. The barriers to internal migration thrown up between cities and the countryside by the household registration system are in many ways comparable to the impact of national borders on international migration. Despite that fact that by now many rural migrants have lived in a city for many years, they continue not to be denied the entitlements of permanent city dwellers like a house rented from or purchased from the work unit, health care and health insurance, and education. Rural migrants therefore continue to be tied to the soil of their village of origin: their key entitlements and obligations continue to be had not where they happen to live but where they are from and where one or more members of their family (parents, spouse, or children) quite often continue to reside and work the land (Chan, 1994).

Rural–urban migration brings people from many different places together in China's larger and smaller cities, (re) producing forms of diversity that could not develop during the collective period. However, prerevolutionary forms of organization and representation (native place organizations and guilds) have not reappeared, and migrants are largely powerless in the face of the established population and local authorities. The best known case is 'Zhejiang village' (*Zhejiang cun*) in Beijing in the 1990s, a community of often very successful traders and entrepreneurs from the area around the city of Wenzhou in Zhejiang province. The Beijing authorities deemed their presence a nuisance and even a threat. Throughout the 1990s, the police repeatedly raided the area, demolishing houses, the wholesale market, and workshops that made up the 'village' (Xiang, 2005).

Migration and mobility are by no means limited to rural–urban migration to China's industrial centers in the eastern coastal areas. Politically very important is the long-standing migration from China's interior to peripheral, originally non-Chinese territories, such as Tibet, Xinjiang, and Inner Mongolia. Much of this is state-planned or state-induced, an element in the strategic incorporation of these areas into the Chinese nation-state (Rossabi, 2004). The reforms period has added increasing flows of nonstate planned migration of entrepreneurs, skilled workers and professionals, farmers, traders, and tourists (Halskov Hansen, 2005; Schein, 2000). Non-Han Chinese are now becoming minorities even in many of their own autonomous areas, a politically combustible development especially in Xinjiang and Tibet, where the Chinese presence is the most controversial and where tensions have traditionally been high.

During the reform period, Chinese not only started to migrate to destinations inside the country itself, but also abroad. In the nineteenth century after the defeat by Western imperial powers during the Opium Wars of 1842 and 1860s, China had been forced to open up the outside world. This starting some of the largest emigration flows of the century, with Chinese fanning out to destinations all across the world (McKeown, 1999). In 1949, China quickly shut itself off from further emigration, but emigration quickly rebounded in the 1980s, reaching levels in the early twenty-first century unprecedented even at the peak of emigration 150 years earlier. China now is one of the most important sending countries. Apart from sheer numbers, the most important fact about Chinese international migration today is the diversity of migratory flows. Migrants now hail from all over China, and include business and government expatriates, investors and entrepreneurs, students, professionals, contract workers, unskilled job seekers, and family migrants. Added to this must be the vastly larger number of Chinese who travel abroad for shorter periods as tourists or visitors on business, on exchange programs, or as members of delegations. The world is indeed becoming a Chinese space, with long-term emigration being a vital, but by no means the only component (Chu, 2010; Fong, 2011; Nyiri, 2010).

More recently, international migration to rather than from China has been added to the mix. Since the mid-1990s, a rising number of foreigners live in China, many of whom do so long term or even permanently. These immigrants are attracted by the opportunities of investment, employment, trade, or business. Others come to China for the promise of a better life or a politically more stable environment. International migrants still are only a minute fraction of China's huge population, but according to the 2010 census their absolute number has already surpassed the 1 million mark. Moreover, foreigners are highly concentrated in specific parts of the country, such as Guangzhou, Shanghai, or Beijing. In these places, foreigners have become a long-term and important aspect of the urban landscape, contributing to the increasing diversity of cultural, religious, linguistic, and even political life. In the last 10 years, China's large coastal cities have become global cities, cosmopolitan hubs that attract people from all over East and Southeast Asia and beyond (Pieke, 2012).

Despite the persistence of the household registration system and minority policies that continue to be predicated on the assumption that it is the norm that people stay where they belong and have specific ascribed and unalterable statuses, Chinese society has changed almost beyond recognition over the past 20 years. Indeed, the paradigms of migration and diversity no longer seem adequate to describe many of these developments, especially in the last 10–15 years. Social and spatial mobility have become increasingly normal, an enduring aspect of life in China. A new generation of rural migrants – often the children of parents who were migrants themselves in the 1980s or 1990s – no longer seems content with temporary employment in the expectation of an eventual return to the village. Indeed, it is increasingly difficult to distinguish these highly rural migrants and city high school and college graduates who travel across China in search for work and a career, often staying in one place for no more than a few weeks or months (Siu, 2007). China is moving rapidly in the direction of a nationally integrated job market, where regional and rural–urban differences no longer define and limit one's chances, aspirations, and destinations.

Unfolding Modernities

The gradual dissolution of the work units and village collectives has broken down many of the old barriers and certainties of socialist society. In their stead, a society is developing where the population consists of families, households, and individuals who are free to achieve their life goals in ways and places of their own choosing. We should not mistake this for a gradual demise of the socialist system under the impact of market

capitalism: in China, capitalism and socialist governance have proved to be symbiotic rather than antagonistic (Pieke, 2009). Yet despite the strengthening of the regime of the CCP and the absence of a multiparty democracy and human rights, substantial gains in personal, economic, and even political freedom have been real and sustained.

China has become a highly aspirational society with rapidly growing and increasingly wealthy and self-conscious middle and upper classes (Goodman, 2008). For many such middle class Chinese, especially in the eastern coastal regions, modernity is both an achieved status and a promise of a better future. A belief in the opportunities to move up in the world by dint of hard work, thrift and ability is widely shared among the population, even among the most disadvantaged groups and in the face of rising inequality and the absence of a level playing field (Whyte, 2010). The state promotes a discourse on the greater or lesser 'quality' (*suzhi*) of the population, individuals, and groups. The concept of 'quality' quite deliberately conflates many different vectors and explanations of success (education, wealth, genetic makeup, rural or urban residence, to name just a few). 'Quality' discourse puts the blame for backwardness on those who are deemed backward. It points to a strategy to achieve national strength and modernity that raises the 'quality' of the population, in effect telling the losers of the reform process to catch up or else be permanently left behind (Anagnost, 2004).

Perhaps the most visible example is education. Rising standards of living and the marketization of education have impelled hundreds of millions of middle class families to invest all their resources in the education of their one child. The aim is to score high enough in the nation-wide college entrance examination to get into a good university, and beyond that perhaps the opportunity to study abroad. For many families, education has become a bottomless pit that they will continue to throw money in, often regardless of the expected return on the education received, a will to elite status with strong echoes of a Confucian past (Fong, 2004; Kipnis, 2011).

Housing is another aspect of the fundamental changes that have taken place. Rural Chinese always owned their own homes, but in the second half of the 1990s urban work units ceased to rent highly subsidized housing to their members. China became a nation of home owners. Residence now reflects choice, wealth, power, and status, leading to a spatial sorting of the population in different types of neighborhoods that looks more and more like capitalist societies (Zhang, 2010). Quite often, this is anything but a smooth process. Demonstrations, petitions, and law suits against the demolition of old neighborhoods or the appropriation of rural land for development are rife; indeed, land and housing demolition are the most important issue of political contestation in China today (Hsing, 2010; O'Brien and Li, 2006). In conjunction with this shift, the government has responded to the weakening of work units to building up residential 'communities' (*shequ*) as a key interface between local government and the population. In these new communities, local police, government agencies, property management agencies, and home owners associations are jointly responsible for local governance (Read, 2012).

In rural China, the hand of government weighs less heavily as in urban areas. Villages are no longer collective units of agricultural production, but continue to be central as units of residence, land rights, and other entitlements. In the 1980s, villages played a crucial role in fostering enterprises, either privately or collectively owned; some of the more successful among them evolved into genuine business conglomerates with a diverse portfolio of holdings. Although the age of such 'super villages' seems to have passed, villages located close to (or even enveloped by) cities control valuable land that they can turn into a substantial source of income for the village and its members (Hsing, 2010; Siu, 2007). After the reforms, many villages have reinvented prerevolutionary ways to recreate and celebrate their sense of community. Fueled by the increased wealth brought by the reforms, patrilineal kin groups, Christian churches, popular religious temples, festivals, and celebrations have become crucial aspects of village life and governance, with formal village heads or party secretaries frequently doubling up as lineage leaders, temple elders, or organizes of fairs, festival, or ritual celebrations (Ashiwa and Wank, 2009; Chau, 2006; Jing, 1996; Tsai, 2007).

In 1949, China had only just started on the road to nation building. By 1978, after 30 years of unification and political integration, China had progressed immeasurably farther. A unified spoken and written national language was now taught in most high schools and universities across the country. An elaborate network of national and local media (newspapers, radio, and television (TV)) reached into virtually all areas and households. A sense of Chinese identity was shared across the country regardless of the many local, ethnic, and religious identities that existed. Yet it was only the generations after the start of the reforms that fully became nationally Chinese. For the first time, large numbers of children grew up speaking standard Chinese (Mandarin) as their first language rather than a local dialect. Travel, tourism, and migration created spaces and experiences that encompassed China as a whole (on tourism, see Nyíri, 2006). The commercialization of the mass media meant that national TV began to cater to popular tastes, creating audiences of hundreds of millions of households for shows, soaps, and movies. More recently, the Internet and social media have united Chinese across the country into a plethora of virtual communities around any issue, interest, or hobby imaginable. The hand of government censorship usually does not weigh as heavily as is often thought and largely is limited to attempts at organizing collective action or explicitly political activities (Yang, 2009; Zhou, 2006). Against this background, it is not all that surprising that the 1990s and 2000s have been the age of the growth of popular nationalism, which sometimes is virulently jingoistic or even racist. Some of this is no doubt due to careful nurturing and manipulation by the regime, but a more important explanation seems to be that for many people only in this era China began to exist and be experienced as a lived-in reality (Gries, 2004).

The erosion of work units and collectives, social and spatial mobility, and the emergence of a national space have been complemented by the individualization of Chinese society, or, as Chinese sociologists sometimes put it, Chinese have changed from being "people of work units" (*danwei ren*) to "people of society" (*shehui ren*). Individualization in China is sometimes linked to a deterioration of social norms and a lack of social responsibility and at odds with traditional Chinese culture that emphasizes group membership (especially the family) and the relatedness of the social person (cf Yan, 2010).

Norms of reciprocity and social obligations, already under pressure during the collective and early reform period when they often deteriorated into the instrumental use of "personal connections" (*guanxi*, see Yang, 1994), are now thought to be further in decline. Market exchanges and contractual obligations backed up by the rule of law leave less space for favoritism and the manipulation of personal relations (Gold et al., 2002). In the private sphere, the traditional Chinese emphasis in the Chinese family on the obligations between parents and children has given way to personal choice in marriage and the primacy of conjugal relations between husband and wife (Yan, 2003).

In terms of gender relations, the reforms have proved to be a mixed blessing. While the Maoist period was certainly not a feminist utopia, increasing labor participation and education of women nevertheless meant a huge improvement. In certain respects, the reform period further improved the position of women, especially in the cities. With the one-child-policy (Greenhalgh and Winckler, 2005), many couples are raising a single daughter and cannot and do not want to treat her any different than they would if their single child had been a son. However, in many other places, things were not quite so simple. Rural couples in need of a son to continue the family line and as security for old age often feel forced to abort, abandon, or not register a baby daughter. China's sex ratio is now dangerously skewed within some places 150 or so men for every 100 women. Tens of millions of children and young adults, many more of them women than men, were born illegally and have none or only very few rights.

The forces of the market and social inequality also impinge on gender relations. Rural-to-urban migration has provided a way for many young rural women to become socially and financially independent from their parents, husbands, or parents-in-law. However, many of these women find themselves into potentially very exploitative situations as live-in domestic help, in sweatshops, or in the sex industry (Yan, 2008; Zheng, 2009). The wealth, social inequality, and scarcity of women generated by the reforms and the one-child-policy have led to many forms of commercialization and commodification of women and sexuality. These include a steep hike in bride-price for marriage, concubinage, prostitution, and smuggling or trafficking of women. This is rapidly developing a transnational dimension. Women from, for instance, North Korea or Southeast Asia come to China to marry Chinese men, while others end up in the sex industry. Conversely, married Taiwanese or Hong Kong men are known to maintain a mistress in Shanghai or the Pearl River Delta (Grillot, 2012; Shen, 2005).

Modernization, wealth, individualization, and social and spatial mobilities create a whole new diversity of life styles. Like in other developed societies, these life styles are techniques of the self that are driven personal choice and inclination rather than ascription and are informed by fashion, mass media, peer groups, and popular culture. Some of these life styles define communities of belonging within the anonymous environment of China's megacities; others are a matter of individual consumer choice and identity construction (Gerth, 2010). Sexual orientation and preference (Jeffreys, 2006, modern art or pop music De Kloet, 2010), and conversionary religions (Buddhism, Christianity, syncretic sects Goosaert and Palmere, 2011) are only some examples. The vast majority of such communities have little if any interest in political participation or activism. Nevertheless, the example of the Falun Gong demonstrations in 1999 shows that their impact can be very sudden and substantial, enough to trigger immediate suppression from the authorities.

China's modernization of the last 30 years has been an unprecedented story of sustained success. Chinese have become immeasurably wealthier than they have ever been. Yet some real problems have emerged that in the long term threaten to undo many if not most of the quality of life gains of reform. Most elusive are the twin issues of environmental degradation and long-term demographic imbalances. Air, water, and soil in many areas are now so polluted that they have become virtually uninhabitable. In many places, the environment has become a major cause of local political contention; added to this should be the rising discontent about food safety, particularly after scandals with tampered baby formula in 2006 and 2008. The Chinese authorities are now increasingly aware that a "pollute now, pay later" attitude no longer works and an impressive regulatory framework has been put into place. Yet genuine and sustainable solutions of the many problems have to be weighed against the competitive advantage of major sectors of the economy. In other places, the damage already is irreparable or, perversely, pollution has become so much part of life that the local population either are ignorant of its harmful effects or has become so dependent on payoffs from polluting factories that they actually resist change (Lora-Wainwright, 2009; Weller, 2006; Yan, 2012).

Demographically, China lives on a time bomb. Until 2013, the economy benefited from a demographic dividend (relatively few old and young people and thus a large working-age population). This dividend has now run out. Henceforth, there will be a population deficit with relatively small cohorts of young people entering the labor force and relatively large cohorts of pensioners leaving it (Park et al., 2010). This problem is only in part a normal aspect of the demographic transition that all developing countries go through. It is made much worse by 30 years of strict, state-mandated family planning that still has not been abandoned for reasons that with each passing year become increasingly obscure (Greenhalgh, 2008).

Whether China is able to overcome the challenges that lie ahead is not just a Chinese problem. China's global impact is now such that what will happen there will be felt by us all. A stable, prosperous, and confident China is in everybody's best interest.

See also: Adoption, Demography of; Afrocentric Approaches to Social Work; Chinese Revolutions: Twentieth Century; Communism, History of; Culture Policy Regimes in China; Fertility Transition: China; Sexual Minorities in China; Sexuality, Theories of; Socialism: Historical Aspects; socialism; Socialist Societies: Anthropological Aspects; Special Economic Zones and Cluster Dynamics: China; Urban Planning in China; Urbanization in China.

Bibliography

Ahern, E.M., 1973. The Cult of the Dead in a Chinese Village. Stanford University Press, Stanford.

Anagnost, A., 2004. The corporeal politics of quality (*Suzhi*). Public Culture 16, 189–208.
Ashiwa, Y., Wank, D.L. (Eds.), 2009. Making Religion, Making the State: The Politics of Religion in Modern China. Stanford University Press, Stanford.
Barabantseva, E., 2011. Overseas Chinese, Ethnic Minorities, and Nationalism: De-centering China. Routledge, London.
Bernstein, T.P., Lü, X., 2003. Taxation without Representation in Contemporary Rural China. Cambridge University Press, Cambridge.
Brady, A.-M., 2003. Making the Foreign Serve China: Managing Foreigners in the People's Republic. Rowman and Littlefield, Lanham.
Bray, D., 2005. Social Space and Governance in China: The Danwei System from Origins to Reform. Stanford University Press, Stanford.
Brown, M.J., 2004. Is Taiwan Chinese? The Impact of Culture, Power, and Migration on Changing Identities. University of California Press, Berkeley.
Chan, K.W., 1994. Cities with Invisible Walls: Reinterpreting Urbanization in Post-1949 China. Oxford University Press, Hong Kong.
Chau, A.Y., 2006. Miraculous Response: Doing Popular Religion in Contemporary China. Stanford University Press, Stanford.
Chu, J.Y., 2010. Cosmologies of Credit: Transnational Mobility and the Politics of Destination in China. Duke University Press, Durham.
Cohen, M.L., 1976. House United, House Divided: The Chinese Family in Taiwan. Columbia University Press, New York.
Constable, N. (Ed.), 1996. Guest People: Hakka Identity in China and Abroad. University of Washington Press, Seattle.
Crossley, P.K., 2005. Nationality and difference in China: the post-imperial dilemma. In: Fogel, J. (Ed.), The Teleology of the Modern Nation-State: Japan and China. University of Pennsylvania Press, Philadelphia, pp. 138–155.
Crossley, P.K., Siu, H.F., Sutton, D.S. (Eds.), 2006. Empire at the Margins: Culture, Ethnicity, and Frontier in Early Modern China. University of California Press, Berkeley.
De Kloet, J., 2010. China with a Cut: Globalisation, Urban Youth and Popular Music. Amsterdam University Press, Amsterdam.
Fei, H.-T., 1953. China's Gentry: Essays in Rural-Urban Relations. University of Chicago Press, Chicago.
Feuchtwang, S., 1977. School-temple and city-god. In: Skinner, G.W. (Ed.), The City in Late Imperial China. Stanford University Press, Stanford, pp. 581–608.
Fong, V.L., 2004. Only Hope: Coming of Age under China's One-Child Policy. Stanford University Press, Stanford.
Fong, V.L., 2011. Paradise Redefined: Transnational Chinese Students and the Quest for Flexible Citizenship in the Developed World. Stanford University Press, Stanford.
Freedman, M., 1958. Lineage Organization in Southeastern China. The Athlone Press and Humanities Press Inc., London and New York.
Freedman, M., 1966. Chinese Lineage and Society: Fukien and Kwangtung. The Athlone Press, London.
Fried, M.H., 1953. Fabric of Chinese Society: A Study of the Social Life of a Chinese County Seat. Praeger, New York.
Gerth, K., 2010. As China Goes, So Goes the World: How Chinese Consumers Are Transforming Everything. Hill & Wang, New York.
Giersch, C.P., 2006. Asian Borderlands: The Transformation of Qing China's Yunnan Frontier. Harvard University Press, Cambridge, Mass.
Gladney, D., 2004. Dislocating China: Reflections on Muslims, Minorities, and Other Subaltern Subjects. Hurst & Co., London.
Gold, T., Guthrie, D., Wank, D.L. (Eds.), 2002. Social Connections in China: Institutions, Culture, and the Changing Nature of Guanxi. Cambridge University Press, Cambridge.
Goodman, D.S.G. (Ed.), 2008. The New Rich in China: Future Rulers, Present Lives. Routledge, London.
Goosaert, V., Palmere, D.A., 2011. The Religious Question in Modern China. University of Chicago Press, Chicago.
Greenhalgh, S., 2008. Just One Child: Science and Policy in Deng's China. University of California Press, Berkeley.
Greenhalgh, S., Winckler, E.A., 2005. Governing China's Population: From Leninist to Neoliberal Biopolitics. Stanford University Press, Stanford.
Gries, P.H., 2004. China's New Nationalism: Pride, Politics, and Diplomacy. University of California Press, Berkeley.
Grillot, C., 2012. Cross-border marriages between Vietnamese women and Chinese men: their integration of otherness and the impact of popular representations. In: Haines, D., Yamanaka, K., Yamashita, S. (Eds.), Wind over Water: Migration Inian Context. Berghahn, New York, pp. 125–137.
Halskov Hansen, M., 2005. Frontier People: Han Settlers in Minority Areas of China. Hurst, London.
Heberer, T., 1989. China and Its National Minorities: Autonomy or Assimilation? M.E. Sharpe, Armonk, NY.
Hsing, Y.-T., 2010. The Great Urban Transformation: Politics of Land and Property in China. Oxford University Press, Oxford.
Jeffreys, E. (Ed.), 2006. Sex and Sexuality in China. Routledge, Abingdon.
Jing, J., 1996. The Temple of Memories: History, Power, and Morality in a Chinese Village. Stanford University Press, Stanford.
Kipnis, A.B., 2011. Governing Educational Desire: Culture, Politics, and Schooling in China. University of Chicago Press, Chicago.
Lin, Y.-H., 1998. The Golden Wing: A Sociological Study of Chinese Familism. Routledge, London.
Lora-Wainwright, A., 2009. Of farming chemicals and cancer deaths: the politics of health in contemporary rural China. Social Anthropology 17, 56–73.
McKeown, A., 1999. Conceptualizing Chinese diasporas, 1842 to 1949. Journal of Asian Studies 58, 306–337.
Nyiri, P., 2010. Mobility and Cultural Authority in Contemporary China. University of Washington Press, Seattle.
Nyíri, P., 2006. Scenic Spots: Chinese Tourism, the State, and Cultural Authority. University of Washington Press, Seattle.
O'Brien, K.J., Li, L., 2006. Rightful Resistance in Rural China. Cambridge University Press, Cambridge.
Park, A., Cai, F., Du, Y., 2010. Can China meet her employment challenges? In: Oi, J., Rozelle, S., Zhou, X. (Eds.), Growing Pains: Tensions and Opportunities in China's Transformation. Stanford Asia-Pacific Research Center, Stanford, pp. 27–55.
Pieke, F.N., 2009. The Good Communist: Elite Training and State Building in Today's China. Cambridge University Press, Cambridge.
Pieke, F.N., 2012. Immigrant China. Modern China 38, 40–77.
Potter, S.H., Potter, J.M., 1990. China's Peasants: The Anthropology of a Revolution. Cambridge University Press, Cambridge.
Read, B.L., 2012. Roots of the State: Neighborhood Organization and Social Networks in Beijing and Taipei. Stanford University Press, Stanford.
Rossabi, M. (Ed.), 2004. Governing China's Multiethnic Frontiers. University of Washington Press, Washington.
Rowe, W.T., 1989. Hankow: Conflict and Community in a Chinese City, 1796–1895. Stanford University Press, Stanford.
Sangren, P.S., 1987. History and Magical Power in a Chinese Community. Stanford University Press, Stanford.
Sangren, P.S., 2001. China: sociocultural aspects. In: International Encyclopedia of the Social & Behavioral Sciences, pp. 1733–1738.
Schein, L., 2000. Minority Rules: The Miao and the Feminine in China's Cultural Politics. Duke University Press, Durham, NC.
Shen, H.-H., 2005. 'The first Taiwanese wives' and 'the Chinese mistresses': the international division of labour in familial and intimate relations across the Taiwan Strait. Global Networks 5, 419–437.
Siu, H.F., 2007. Grounding displacement: uncivil urban spaces in postreform South China. American Ethnologist 34, 329–350.
Skinner, G.W., 1964. Marketing and social structure in rural China, Part I. Journal of Asian Studies 24, 5–43.
Tsai, L.L., 2007. Accountability without Democracy: Solidary Groups and Public Goods Provision in Rural China. Cambridge University Press, Cambridge.
Weller, R.P., 2006. Discovering Nature: Globalization and Environmental Culture in China and Taiwan. Cambridge University Press, Cambridge.
Whyte, M.K., 2010. Myth of the Social Volcano: Perceptions of Inequality and Distributive Injustice in Contemporary China. Stanford University Press, Stanford.
Wolf, M., 1968. The House of Lim: A Study of a Chinese Farm Family. Prentice Hall, Englewood Cliffs.
Xiang, B., 2005. Transcending Boundaries. Zhejiangcun: The Story of a Migrant Village in Beijing. Brill, Leiden.
Yan, H., 2008. New Masters, New Servants: Migration, Development, and Women Workers in China. Duke University Press, Durham.
Yan, Y., 2003. Private Life under Socialism: Love, Intimacy, and Family Change in a Chinese Village, 1949–1999. Stanford University Press, Stanford xvi, 289 pp.
Yan, Y., 2010. The Chinese path to individualization. The British Journal of Sociology 61, 489–512.
Yan, Y., 2012. Food safety and social risk in contemporary China. Journal of Asian Studies 71, 705–729.
Yang, C.K., 1961. Religion in Chinese Society. University of California Press, Berkeley.
Yang, F., 2012. Religion in China: Survival & Revival under Communist Rule. Oxford University Press, Oxford.
Yang, G., 2009. The Power of the Internet in China: Citizen Activism Online. Columbia University Press, New York.
Yang, MM-H., 1994. Gifts, Favors & Banquets: The Art of Social Relationships in China. Cornell University Press, Ithaca.
Zhang, L., 2010. In Search of Paradise: Middle Class Living in a Chinese Metropolis. Cornell University Press, Ithaca.
Zheng, T., 2009. Red Lights: The Lives of Sex Workers in Postsocialist China. University of Minnesota Press, Minneapolis.
Zhou, Y., 2006. Historicizing Online Politics: Telegraphy, the Internet, and Political Participation in China. Stanford University Press, Stanford.

Asia, Sociocultural Overviews: Korea

Robert Oppenheim, University of Texas at Austin, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by N. Abelmann, volume 12, pp. 8159–8162, © 2001, Elsevier Ltd.

Abstract

This article examines the historical development of the anthropology of Korea and the Koreas from the nineteenth century to the present, highlighting the shifting multinational character of this anthropology over time. It examines the contexts of colonialism and the predominance of social anthropological village studies in the years after 1945, while devoting its greatest attention to trends in sociocultural research on both Koreas over the past two decades.

Late Nineteenth Century and Colonial Origins of the Anthropology of Korea

With respect to US anthropology, Korea had a greater role in late-nineteenth-century evolutionary ethnology, institutionally based in museums, than it would in the immediate wake of the early twentieth century disciplinary turn to long-term fieldwork as a defining research method. The 1882 treaty that established diplomatic relations between the United States and the kingdom of Chosŏn (i.e., Korea) coincided with a general expansion of collecting activity at the United States National Museum of the Smithsonian Institution, the dominant anthropological institution of the era. More specifically, it coincided also with a joint program between the Smithsonian and the US Navy to train young naval officers in scientific techniques and employ them for museum purposes. The first major anthropological collection of Korean materials to arrive in the United States came courtesy of a product of this program, John B. Bernadou. Drawing on this and on other collections of the late 1880s, National Museum Curator of Ethnology Otis Mason (1886: 115) offered an initial gloss of the significance of Korean objects as "a human exemplification of the survival of whole genera of industries and customs, while in surrounding regions these have been swept away or transformed." Mason's location of Korea's importance in relation to E.B. Tylor's notion of evolutionary survivals, but also simultaneously in universal ('human') and regional terms, bespoke a larger ambivalence that attended the evolutionary anthropological understanding of East Asian societies in this period. One resolution would come almost a decade later in the form of the first book-length study of Korea in the US anthropological tradition, Stewart Culin's *Korean Games*. Culin's book stemmed from a joint project with his fellow anthropologist and close friend Frank Hamilton Cushing, whose research focus was on the US Southwest. Together, Culin and Cushing sought to demonstrate parallel sequences of the evolutionary development of games of chance from practices of divination in East Asia and North America, and thus to make a strong argument for 'independent invention' – the like working of a (universalized) human mind in similar circumstances – over diffusion as a primary explanation for the occurrence of similar customs or artifacts in widely separated areas. Even as Korea was thus being enlisted into metropolitan anthropological theory, a generation of missionaries, travelers, and diplomats resident in the country itself were producing texts that, while often quite judgmental, were also more descriptive and ethnographic in feel (Chun, 1999: 41–42). Ironically, it is these works, more than those of the nineteenth-century US anthropological establishment, that have tended to be appropriated as useable precursors for more ethnographic anthropologies that followed the eclipse of the evolutionary paradigm.

Yet from this early moment, Korea has also offered a contrast to situations in which a dominant national tradition of anthropological analysis has grown up within relatively settled colonial relations, as, for example, in the case of British social anthropology in Africa or studies of Native Americans in the United States. The relative lateness of colonialism in Korea contributed to a situation in which the anthropology of Korea was from the first deeply multinational in its authorship. French, Russian, British, and German anthropologists also produced a volume of scholarship. The foundation of the Tokyo Anthropological Society under Tsuboi Shōgorō in 1884 was followed around 1890 by the first appearance of articles on Korea in its journal, *Jinruigaku Zasshi*, setting the stage for subsequent research during the colonial era (Chun, 1999: 42–43). At the same time, Korean intellectuals of this period, in their attempt to imagine and configure national modernity under the shadow of imperialist encroachment, drew both on extant local argumentative traditions and on imported frameworks such as evolutionism and Social Darwinism, producing many writings with an auto-anthropological character (for the example of Ko Ŭijun, see Chun, 1999: 35–41). At moments, the dialogue of Korean indigenous scholarship and proto-institutional international anthropology was quite literal. The intellectual Yu Kilchun toured the United States in 1883 as part of an official study mission and then stayed on to attend US schools, living for a period with the natural historian, archaeologist, and ethnologist Edward S. Morse near Salem, Massachusetts. Yu went on to publish *Sŏyu kyŏnmun* (*Observations on a Journey to the West*), one of the most important late-nineteenth-century Korean accounts of world geography, history, and Western governmental and social institutions; Morse wrote 'Korean Interviews,' an anthropological overview of Korean social customs, which, for the most part, was anonymously based on his conversations with Yu.

With Japan's victories in the Sino-Japanese War (1894–95) and the Russo-Japanese War (1904–05), the establishment of a Japanese protectorate in Korea in 1905, and the Japanese annexation of Korea in 1910, the balance of these intellectual formations shifted. One US anthropologist, Frederick Starr,

continued to do research in Korea during the Japanese colonial period, making six research trips from 1911 to 1930. But the combination of Starr's sparse publication on Korea and his marginality in the discipline ensured that this work was poorly known, then and since. Instead, under colonial conditions, Japanese anthropology and related disciplines such as archaeology and folklore became dominant in the institutionalized study of Korean society and culture; prominent Japanese scholars of Korea included Torii Ryūzō, Sekino Tadashi, Imamura Tomo, Murayama Chijun, Zensho Eisuke, and Akiba Takashi. Some of this research took place under the auspices of the colonial Government General of Chōsen or other allied organs and was at least nominally oriented to the problems of colonial governance. Much also played into thematic poles of colonial ideology, emphasizing either the difference of Koreans from Japanese in more or less essentialized terms, or the ostensible common ethnoracial origins of Koreans and Japanese and thus the supposed preservation in Korea of traces of a more primordial Japan, with the latter relatively pronounced given the assimilationist design of much Japanese colonial policy in the peninsula. For these reasons, post-1945 critiques of colonial knowledge in the Koreas, the earliest of which significantly preceded the rise of postcolonial and post-Orientalist critique in Anglo-American anthropology, have sometimes offered blanket repudiations of this Japanese scholarship. Even as this has been the case, however, some research foci, conceptual frameworks, and findings of colonial anthropology have carried forward into subsequent social and cultural analysis of the Koreas. Akiba, for instance, gave impetus to later work on Korean shamanism, which he portrayed as located within a distinct female cultural sphere opposed to the Confucian sphere of men. The topic of shamanism, the problematic of gendered spheres, and the revision of this problem remained important to much village-based ethnography from the 1940s to the 1980s (Pai, 2000; Atkins, 2010: 52–101).

As in the late nineteenth century, Japanese and other foreign anthropological research during the colonial era coexisted with Korean scholarship on national folklore and cultural history. Much of this Korean scholarship in turn bore connections to the broader intellectual movement of cultural nationalism, which gained strength in the colonial milieu from the beginning of the 1920s. Of the four folklorists whom Roger L. Janelli (1986) regards as central, Yi Nŭnghwa and Ch'oe Namsŏn undertook mostly philological research. Both treated shamanism as deeply imbedded in Korea's history and culture and as a locus of its indigenous worldview, and Ch'oe in particular sought linkages for this and other elements in the Asian mainland, thus displacing colonial claims of a primordial Japanese–Korean racial and cultural unity. Son Chint'ae and Song Sŏkha, meanwhile, undertook fieldwork to examine aspects of Korea's folk culture; the former understood this in Tylorian fashion as suffused with survivals from the national past while the latter, more pragmatically and romantically, sought its preservation against encroaching change.

Divided Korea and Village Ethnographies

The end of Japanese colonialism in 1945 was followed quickly by separate occupations of the south and north of Korea by the United States and Soviet Union, respectively. Great power sponsorship of ideologically compliant regimes hardened these separate zones of occupation into competing states by 1948, while the accelerant it poured onto Korean social divisions set the stage for the outbreak of full-scale fighting in the Korean War of 1950–53. The political economy of sociocultural research in both halves of the divided nation was interwoven with these changes. In the south, Song Sŏkha and Son Chint'ae continued as central figures, and the predecessor organization to the Korean Anthropological Association, a museum of Korean folklore and anthropology, and the first academic department of anthropology were all established by 1946 (Chun, 1999: 125–160). Meanwhile, US anthropological researchers Cornelius Osgood and Eugene Knez conducted or began their village-based ethnography under the auspices of the US occupation government. During the Korean War itself, military-sponsored team anthropological research on the effects of temporary North Korean occupation of southern villages, along with refugee research, was among the techniques used to attempt to understand the 'Sovietization' of North Korean society itself. Notwithstanding the framing of this wartime research by conceptually blinkered Cold War tropes, and the apparent assumption of a monolithic 'culture of communism,' it actually found complex interactions between existing village social processes and real or ostensible political commitments, and in many respects was of a piece with the peacetime political and social anthropology of villages that would follow. For decades after the occupation, the United States would remain the major training site for perhaps most South Korean researchers, and the traditions have had strong ties as a result, though Japanese research on South Korea has also continued. Simultaneously, in North Korea, anthropology has had a role as a contributor to the intellectual project of socialist modernity, but the precise nature of that role, as well as the character of the connections of this research practice to other socialist anthropologies, remains a topic of ongoing investigation.

At the risk of overgeneralization, it can be said that the classic anthropology of South Korean village life that was predominant from the 1950s through approximately the mid-1980s had a sociological emphasis that distinguished it from other adjacent anthropologies. Village organization, its relation with structures and practices of kinship, and religion were among its major topics. Even in the last case, however, the tendency to abstract out religion or culture as an autonomous system was little in evidence, certainly in comparison to the postwar anthropology of Japan, for which the influence of the culture and personality school and its apotheosis in the work of Ruth Benedict's *The Chrysanthemum and the Sword* was important. The first English-language ethnography of Korea published in this period, Osgood's *The Koreans and Their Culture* of 1951, was also internally heterogeneous; the author appended to his village study a historical account of Korea from ancient times, out of a conviction that such information was lacking and thus necessary background for his readership.

Some specific topics of this era, such as the prevalent clan/lineage structures present within villages and their relation to village sociopolitical authority, were more pronounced in Korean-language work (e.g., Kim, 1979). But there were also several crosscutting common issues. In the last edition of this encyclopedia, Nancy Abelmann wrote of a set of

dichotomies that tended to structure Korean village studies, whether sustained or contested in various ways. These were between elite and non-elite culture and practice, between Confucianism and shamanism, and between male and female. One of the strongest conceptual developments of these dichotomies (and others) occurred in the work of Vincent S.R. Brandt (1971). Confronting an unexpected diversity of villagers' orientations toward kinship and community, and following perhaps most closely the model of models in Edmund Leach's *Political Systems of Highland Burma*, Brandt postulated the existence of dual, hierarchical and egalitarian cognitive ethical structures differently expressed in different situations. With the general anthropological turn toward practice and performance, and a specific interest in ritual, ethnographies of South Korea called into question the cleanliness of these dichotomies as either cognitive map or social actuality. In the early writings of Roger L. Janelli and Dawnhee Yim Janelli (1982) and Laurel Kendall (1985), it was in a sense deceased ancestors who served as the vehicles for this complexification, insofar as they are the common focus of much Confucian and shamanic ritual practice. Kendall, especially, was specifically concerned to question the notion of a separate 'women's sphere' of Korean religious and ritual activity.

There were, of course, works that fell outside this set of issues. In a more materialist direction, some research examined the transformations of objects of everyday life that was an aspect of the 'modernization' of villages. With a more explicitly critical attitude toward 'modernization' as a concept, Clark Sorensen (1988) related village changes to the larger political economy of industrialization and urbanization and the energetics of peasant households. Meanwhile, tracking more toward French structuralism, though also deeply rooted in classical Korean sources, Kang Shin-pyo explored the possibility of a distinctive Korean cultural grammar; he also offered an early, critical, cultural examination of the rhetoric and experience of high-speed development in South Korea (Kang, 1986).

Recent Trends

Between roughly the mid-1980s and the early 1990s, the anthropology of South Korea underwent a shifting of its sites away from a village focus in dialogue with changes in the discipline and in Korean society. The rise of discourse-based approaches within anthropology encouraged attentiveness to the imagination of tradition, culture, and nation. Simultaneously, within South Korea, movements of students, laborers, farmers, urban squatters, and others opposed to the ruling authoritarian regime and the political economy it supported grew in strength, rooting themselves ideologically in an acutely critical historical and cultural consciousness. Some ethnography was directly concerned with these movement formations. Seung-Kyung Kim's (1997) examination of female workers and labor mobilization in an export-processing zone was focused also on the dynamic tension between gender ideologies and new worker identities. Nancy Abelmann's (1996) study of an agrarian and student movement was permeated with activists' own understanding of their historicity, much of which revolved around the praxiological category of the *minjung*, very approximately the 'folk' or 'common people.' Even in anthropological writing on other topics, however, self-conscious attention to the construction and politics of culture suffusing anthropology and Korean public discussion was evident. Roger Janelli's ethnography with Dawnhee Yim Janelli (1993) of a South Korean corporation was cast against deterministic understandings of the relations between culture and capitalism, even as it tracked also the deployments of notions of familial and village culture by the firm's managers and employees. Laurel Kendall's (1996) account of marriage and wedding practices treated them as an arena in which conceptions of tradition and modernity are negotiated. Further work on shamanism in this period, by Kendall, Kim Kwang-Ok, and others, examined its increasing use as a social and political symbol, as well as the social relations of spokespersonship and marginalization entailed in that appropriation.

Since the 1990s, the international anthropology of the Koreas has continued its diversification of topics. Over roughly the same period, anthropology as practiced in South Korea itself has become thoroughly international in its research scope. No longer is Korean society its almost sole focus, though of course domestic research remains significant. These factors, combined with the large number of new scholars entering the field, make a complete overview impossible. What follows should thus be read as simply an identification of some important recent and contemporary research directions.

Much ethnography of the past two decades has tracked the shifting fortunes of South Korea's economy and the policies and practices of social governance with which it is intertwined. From the early 1990s, Korean anthropologists devoted attention to the apartment-based lifestyles, status consciousness and stratagems, educational focus, and transformed family and gender relations of the new urban middle class, and from around the turn of the millennium anthropologists based abroad have followed suit. Laurel Kendall's (2009) most recent work on shamans includes examination of the intersection of the world of spirits with the animating forces of consumptive capitalist desire. In a good deal of this research, the whiplash of the recalibration of narratives of self and nation in concert with the vertiginous ups and downs of the economy, particularly with the Asian financial crisis of the late 1990s, is front and center. In Laura Nelson's (2000) ethnography, middle class consumption and the gendered moral panic about (over)consumption are considered together. The supposed excessive spending habits of young female consumers, enamored with foreign goods, becomes the anchor for a discourse about the loss of national common purpose. Nancy Abelmann (2003) treats the melodramatic turns and keywords that stretch between South Korean media and women's life stories of rapid transformation. The consequences of the financial crisis included a state embrace of neoliberalism, manifest in new discourses of the self-cultivating individual and shifts in a host of labor and social welfare policies. These are familiar aspects of neoliberalism as a global governmental technology, but Jesook Song (2009), focusing centrally on initiatives concerned with homelessness, highlights the retrenchment of conservative familialist ideology with which they were connected in the South Korean case.

Other research has examined South Korea's increasingly multifarious variety of transnational articulations over the past several decades. Examinations of multiple diasporic communities with links to both Koreas have been a forerunner of this trend; in Korean studies, the tendency to highlight such historical and ongoing connections has been pronounced and long-standing. Recent work has turned to specific, mediated forms of Korean outmigration. Eleana J. Kim (2010), for instance, has focused on the transnational adoption complex, which facilitated the movement overseas of large numbers of South Korean children over a course of decades. Kim's work weaves the social, ideological, and bureaucratic production of this form of adoption together with its political location and intimate effects among South Koreans and adoptees abroad. Other studies have begun to examine the strategic practice and affective work of families that split, often for long periods, in order to secure their children access to education in other countries while one spouse remains employed in South Korea. Meanwhile, in South Korea itself, socioeconomic shifts since around 1990 have led to an influx of two major classes of migrants: labor sojourners, increasingly employed in South Korean manufacturing, construction, and other industries, and foreign brides, recruited in response to a perceived national crisis of familial reproduction, particularly in rural regions. Both are the subject of a good deal of emerging research. At the interstices of these movements, Caren Freeman (2011) examines the case of ethnically Korean Chinese migrants who strategically deploy and manipulate kinship in efforts to stabilize their residence in South Korea. Other anthropologists and sociologists have considered the discourse and actuality of 'multiculturalism' (*tamunhwa*), simultaneously highly structured by gender, that has arisen in the wake of these phenomena within South Korean society, long ideologically constructed as monoethnic and monocultural. Still another focus of recent work is on the transnational imaginaries engendered by various instances of Korean participation in global events or by the circulation of Korean products in foreign settings. Transnational sport is one important topic (e.g., Joo, 2012), while anthropological and sociological considerations of the 'Korean Wave' (*Hallyu*) – the new popularity of South Korean media abroad – have examined its reception and localization in a variety of international settings as well as the contested understanding of its significance in the country of its origin.

The past in South Korea is rarely really past, and from the 1990s onward the traumatic pasts of the twentieth century have increasingly been opened to public discussion with a loosening of authoritarian-era restrictions on speech and publication, a process that reached a momentary apex in the 2000s with the troubled efforts of a national Truth and Reconciliation Commission. Anthropologists have turned an eye to the historical, memorial, and cultural politics of these processes. One of the first to do so was Seong Nae Kim (1989), who considered the return of the repressed April Third Cheju Uprising and massacres of 1948 in shamans' rituals and other locations. Other studies have examined the legacies of violence and the local politics of division in villages torn apart during the Korean War, or the fraught and contested memories of other traumatic historical events, such as the 1980 Kwangju Uprising. More recently, C. Sarah Soh's (2008) work on the 'comfort women' – the institution of wartime sexual slavery in the Japanese empire – contextualizes the phenomenon both in terms of its appropriations by various postcolonial states and activist groups and in relation to longer histories of mobilized public sexuality. Other approaches to the presence of history in contemporary processes draw inspiration rather from considerations of the rendering of history and culture as national heritage, and the selectivity of narratives of the national past fostered by Korean states and a variety of other actors. Moreover, with the return of local self-government at the beginning of the 1990s and new imperatives for regions and municipalities to compete for tourism and development projects, local heritage in South Korea began to draw increased attention. Scholars have examined the intersection of local heritage with national and local electoral politics, its commodification for tourism, and its entanglements with technical disputes (e.g., Kim, 1996; Oppenheim, 2008). Tying together these trends are new attempts to write deep histories of regional communities, historical consciousness, and place-making across long spans of time, exemplified by a collective examination of the Sŏsan region by a team of Korean anthropologists and by a variety of interdisciplinary projects on local areas often rooted in regional South Korean universities.

Recent research on Korean division, and on North Korea, has built on past trends while also embarking in new directions. Work of the late 1990s emphasized the imagination of the North in the South, including various idealizations of North Korean life and authenticity fostered in the absence of much actual contact. Since that time, against the backdrop of a variety of political shifts, such contact between North and South Koreans has increased greatly. Temporary reunions of families originally divided by the Korean War accelerated for a period after 2000 and have drawn additional attention since then. Other studies have examined North Korean refugees, following both their movement through China and other areas and the resettlement of some in the South. Insofar as North Koreans newly resident in South Korea both claim and are imputed with Korean identity while simultaneously experiencing displacement and discrimination, they also pry open the question of the heterogeneity of Koreanness. This research thus bears articulations with the other examinations of migration and multiculturalism noted above. At the same time, however, some scholars have attempted to shine new anthropological (if not ethnographically based) light on the ideological organization of North Korean leadership, state, and society. Sonia Ryang (2012), with an explicit goal of humanizing North Koreans against a tide of portrayals that do the opposite, uses North Korean literature to consider the transactional relationship between individual selfhood and the sovereignty of the Great Leader. The abstract infinitude of the Great Leader's love in these writings demands continual acts of love and sacrifice on the part of his subjects. Viewed as political theology, the idolization of Kim Il Sung (and in differing ways, of his successors) ceases to be *sui generis*, sharing, for example, structural similarities to some Christian understandings of the love of God. Heonik Kwon and Byung-Ho Chung (2012) pursue similar deexoticizing goals along a different pathway. For them, what is basic to the North Korean political system and the rule of the Kims is Weberian

charismatic authority; what needs explanation is its persistence and routinization. In analyzing this, the authors draw on Clifford Geertz's classic examination of the 'theater state' as a useful basis for understanding contemporary North Korea.

In only the past decade, anthropological work on the Koreas has proliferated greatly in quantity and variety. It shows every sign of continuing to do so. Some emerging scholarship takes old issues in new directions, examining shamanism through the lens of the commodification of fortune telling, for example, or social movement practice through the question of new millennium movements' commitment to pragmatism, to the ideology of no ideology. Other topics are relatively new for the field. Anthropologies of media use and reception follow on the intense interest in film within South Korea, and the growing significance of internet communities and digital media in its social life and politics. Research is being done on advertising as a cultural process, and on 'national branding' by organs of a state very explicitly and self-consciously concerned with building its 'soft power.' The anthropology of science, and science and technology studies more generally, has an expanding place in Korean studies – with examinations of biotechnology perhaps especially prominent in the wake of the Hwang Woo-suk cloning scandal. Linguistic anthropology, long a somewhat peripheral subfield, is increasingly in dialogue with sociocultural research. If there remains, in some circles, the perception that anthropological work on the Koreas stands apart from the discipline as a whole, immersed in its own distinctive concerns, that perception has become somewhat distant from reality.

See also: Anthropology at Home; Asia, Sociocultural Aspects: China; Globalization, Anthropology of; Japan: Sociocultural Aspects; Liberalism, Anthropology of; Modernity: Anthropological Aspects; Nationalism: General; Orientalism; Postcoloniality; Refugees in Anthropology; Ritual; Shamanism; Tradition, Anthropology of; War: Anthropological Aspects, Historical Development of.

Bibliography

Abelmann, N., 1996. Echoes of the Past, Epics of Dissent: A South Korean Social Movement. University of California Press, Berkeley.

Abelmann, N., 2003. The Melodrama of Mobility: Women, Talk, and Class in Contemporary South Korea. University of Hawaii Press, Honolulu.

Atkins, E.T., 2010. Primitive Selves: Koreana in the Japanese Colonial Gaze, 1910–1945. University of California Press, Berkeley.

Brandt, V.S.R., 1971. A Korean Village: Between Farm and Sea. Harvard University Press, Cambridge.

Chun, K.-S., 1999. Han'guk illyuhak paengnyŏn. Iljisa, Seoul.

Freeman, C., 2011. Making and Faking Kinship: Marriage and Labor Migration between China and South Korea. Cornell University Press, Ithaca.

Janelli, R.L., 1986. The origins of Korean folklore scholarship. The Journal of American Folklore 99 (391), 24–49.

Janelli, R.L., Janelli, D.Y., 1982. Ancestor Worship and Korean Society. Stanford University Press, Stanford.

Janelli, R.L., Yim, D., 1993. Making Capitalism: The Social and Cultural Construction of a South Korean Conglomerate. Stanford University Press, Stanford.

Joo, R.M., 2012. Transnational Sport: Gender, Media, and Global Korea. Duke University Press, Durham.

Kang, S.-P., 1986. Kŭndaehwa wa chŏnt'ong munhwa. In: Han'guk sahoe kwahak hyŏbŭihoe (Ed.), Han'guk sahoe ŭi pyŏnhwa wa munje. Hyŏnmunsa, Seoul, pp. 362–383.

Kendall, L., 1985. Shamans, Housewives, and Other Restless Spirits: Women in Korean Ritual Life. University of Hawaii Press, Honolulu.

Kendall, L., 1996. Getting Married in Korea: Of Gender, Morality, and Modernity. University of California Press, Berkeley.

Kendall, L., 2009. Shamans, Nostalgias and the IMF: South Korean Popular Religion in Motion. University of Hawaii Press, Honolulu.

Kim, E.J., 2010. Adopted Territory: Transnational Korean Adoptees and the Politics of Belonging. Duke University Press, Durham.

Kim, K.-O., 1996. 'Chibang' ŭi saengsan kwa chŏngch'ijŏk iyong. Han'guk Munhwa Illyuhak 29 (1), 3–35.

Kim, S.N., 1989. Lamentations of the dead: the historical imagery of violence on Cheju Island, South Korea. Journal of Ritual Studies 3 (2), 251–285.

Kim, S.-K., 1997. Class Struggle or Family Struggle?: The Lives of Women Factory Workers in South Korea. Cambridge University Press, Cambridge.

Kim, T., 1979. Ssijok purak ŭi kujo yŏn'gu. Ilchogak, Seoul.

Kwon, H., Chung, B.-H., 2012. North Korea: Beyond Charismatic Politics. Rowman and Littlefield, Lanham.

Mason, O.T., 1886. Corea by native artists. Science (o.s.) 8 (183), 115–118.

Nelson, L.C., 2000. Measured Excess: Status, Gender, and Consumer Nationalism in South Korea. Columbia University Press, New York.

Oppenheim, R., 2008. Kyŏngju Things: Assembling Place. University of Michigan Press, Ann Arbor.

Pai, H.I., 2000. Constructing "Korean" Origins: A Critical Review of Archaeology, Historiography, and Racial Myth in Korean State-Formation Theories. Harvard University Asia Center, Cambridge.

Ryang, S., 2012. Reading North Korea: An Ethnological Inquiry. Harvard University Asia Center, Cambridge.

Soh, C.S., 2008. The Comfort Women: Sexual Violence and Postcolonial Memory in Korea and Japan. University of Chicago Press, Chicago.

Song, J., 2009. South Koreans in the Debt Crisis: The Creation of a Neoliberal Welfare Society. Duke University Press, Durham.

Sorensen, C., 1988. Over the Mountains Are Mountains: Korean Peasant Households and Their Adaptations to Rapid Industrialization. University of Washington Press, Seattle.

Asia, Sociocultural Overviews: South Asia

Andre Beteille, Delhi School of Economics, University of Delhi, Delhi, India

Published by Elsevier Ltd.

Abstract

'South Asia' stands for seven countries: India, Pakistan, Bangladesh, Nepal, Bhutan, Sri Lanka, and the Maldives. The article discusses the overall features of Indian civilization, religion, and language; sociocultural diversity including collective identities of caste and tribe, village organization, family, and kinship; and the important recent changes in Indian society resulting from globalization and related factors.

Area and Population

The term 'South Asia' has come to stand for a group of seven countries: India, Pakistan, Bangladesh, Nepal, Bhutan, Sri Lanka, and the Maldives; and it is in that sense that it will be used in this article. These countries differ greatly in area and population, and in geographical aspect and location. India, with over a billion inhabitants, has the second largest population in the world; the Maldives have a population of only a few hundred thousand people. Nepal and Bhutan are entirely landlocked; Sri Lanka is an island in the Indian Ocean; and the Maldives comprise a chain of more than 1000 islands, most of them uninhabited.

The people of these countries have a shared past by which many of their present social and cultural characteristics have been shaped. Demographically, economically, and politically, India, Pakistan, and Bangladesh are the three most important countries of the region. Until 1947, they were one country. It was partitioned in the same year, when India and Pakistan became separate and independent nations. Pakistan was in turn partitioned in 1971, leading to the creation of Bangladesh as a separate country. The first partition was occasioned by differences of religion, and the second by differences of language. In each case the partition was attended by considerable movements of population.

The Indian subcontinent, comprising India, Pakistan, and Bangladesh, was under British rule for a long time, as was also Sri Lanka. The two landlocked kingdoms of Nepal and Bhutan never came directly under British rule and remained relatively isolated from the economic and political forces released by it. Although many social and cultural changes have taken place in the region since the end of colonial rule in the middle of the twentieth century, large parts of it still remain economically backward.

The subcontinent is the home of a very ancient and distinct civilization that may be described as the 'pan-Indian' or the 'Indic' civilization. This civilization had its origin in the Indo-Gangetic plain in the second millennium BC, and it antedates the birth of both Christianity and Islam. Its characteristic bearers today are the Hindus. It has undergone many changes, and its influence has not only permeated throughout India and its neighboring countries, but also extended to lands beyond the seas such as Indonesia, Cambodia, and Thailand. India is not only the home of Hinduism, it also gave birth to Buddhism, which is the official religion of Bhutan and the predominant religion of Sri Lanka.

Religion and Language

Although Hinduism has had the longest and most pervasive influence on society and culture in the subcontinent, Islam too has a strong presence in it, including present-day India. Pakistan is an Islamic country; Bangladesh has a predominantly Muslim population; and although the Muslims are a minority in India, they are a very large minority. The medieval phase of Indian history was dominated by Muslim sultanates and empires; and the Muslim population of Bangladesh, Pakistan, and India taken together make up a considerable part of the Muslim population of the world as a whole.

In South Asia, the divisions of religion cut across those of state and nation, and the region as a whole may be viewed as a mosaic of religious communities, sects, and denominations. India provides an extreme example of coexistence of diverse religious faiths and practices. Although the Hindus predominate in the country as a whole, there are districts in which Sikhs, Muslims, or Christians predominate. Again, Nepal is a predominantly Hindu country, but there are areas in which Buddhists predominate; in Sri Lanka it is the other way around.

South Asia is the home not only of many religions but also of innumerable languages, and, whether on the plane of cultural values or of social identity, language is as important as religion. Like the divisions of religion, those of language also cut across the boundaries of state and nation. The Punjabi-speaking community is divided between India and Pakistan; the Bengali-speaking community between India and Bangladesh; and the Tamil-speaking community between India and Sri Lanka. Sometimes, as in Sri Lanka, where the Sinhala-speakers are predominantly Buddhist and the Tamil-speakers predominantly Hindu, the divisions of language and religion follow the same line. But this is not always the case. It is common in the subcontinent, if not characteristic, for the divisions of language and religion to cut across each other; there are Muslims, Hindus, and Sikhs among the Punjabi-speakers in India and Pakistan, and Hindus and Muslims among the Bengali-speakers in India and Bangladesh.

Because of historical and demographic reasons India has the largest variety of languages among South Asian countries: it is indeed the most polyglot country in the world. There are more

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12019-7

than a dozen literary languages, divided between the Indo-Aryan and the Dravidian families of language. In addition, there are numerous tribal languages and dialects belonging to the Austro-Asiatic, the Tibeto-Burman, and other families. In India, as in the subcontinent generally, the divisions of language correspond more closely to geographical divisions than do the divisions of religion; but here too, population movements through the centuries have led to a certain dispersal of language communities.

If the civilization of South Asia had a design, it was most clearly visible in the Indian subcontinent until its first partition in 1947. Visitors to the area from far and near have been for centuries struck by the distinctness and continuity of that design, and have commented on its influence on society and culture in the different countries of the region. Despite the changes and upheavals of the twentieth century, its distinctive features are still visible, not only in India, but elsewhere as well.

Accommodation of Diversity

A brief sketch of the design of the traditional civilization of the subcontinent will be provided as an aid to the understanding of the multitude of beliefs, practices, habits, customs, and manners prevalent in the region to this day. The first striking feature of this design is the accommodation of diversity. The rich, ethnographic record available from the middle of the nineteenth century onward shows an endless, almost inexhaustible, variety in material culture, social arrangements, and religious beliefs and practices. Some of the variation in material traits, in matters such as food, dress, and habitation, may be explained by variations in geographical environment and natural resources. But the cultural variation is far in excess of what would be required by variations of natural environment. For instance, even where its basic ingredients are the same, food might be prepared and served differently among the different castes and communities in the same village.

Then there is the great diversity of social arrangements in matters relating to family, marriage, descent, inheritance, residence, and so on. What needs to be stressed is not simply the diversity of practices, but also the diversity of rules; they were the despair of the colonial administrators of the nineteenth century who sought to codify the laws and customs of the subcontinent. There is finally the profusion of religious beliefs, doctrines, rituals, customs, and usages, which have been mentioned.

The presence of diversity has been more than a matter of mere existence. In Indian society in particular and South Asian societies in general, it has been sustained by law, custom, and popular religion. Accommodation without assimilation has been a core value of Indian civilization until modern times. It has led to the coexistence of a large multiplicity of beliefs and practices as well as of social groups. Adding new components has not meant discarding old ones, and new and old components of the most heterogeneous kinds have existed cheek-by-jowl to a greater extent than in other civilizations. The Indian subcontinent has been and still largely remains a vast mosaic of tribes, clans, castes, sects, and communities, each with its own identity.

Collective Identities

The second distinctive feature of South Asian societies is the predominance of collective identities. In the past, what counted was not so much the individual but the group of which he was a member by birth, and this continues to a large extent even now. It has often been said that traditional Indian society was a society not of individuals but of castes and communities. As has been noted, there was a great multiplicity of groups, and this multiplicity was a feature even of small local communities. The members of a single village might observe different religious practices, speak more than one language, and belong to a dozen or more different castes and subcastes.

Collective identities were maintained by fidelity to the customs of one's ancestors and by rules for the regulation of marriage. In course of time, subtle differences developed in craft practices, dress, and ornamentation, marriage rules and ritual observances between groups having the same origin. Each group and subgroup maintained its distinctive style of life and transmitted it from generation to generation. The group preserved a sense of its own identity despite changes in its social position. The high value placed on the accommodation of diversity allowed and indeed encouraged each group to retain a strong sense of its own identity.

The persistence of collective identities may be illustrated with two examples from the subcontinent. The first relates to what has been called 'the Hindu method of tribal absorption.' In the past, tribal groups, which were relatively isolated from the wider society sometimes established relations with that society, discarding some old practices and acquiring some new ones. In this way, a tribe was often transformed into a caste but, despite a change of name and to some extent also of social position, it preserved its old collective identity. The second example relates to the survival of caste identities despite changes of religion. The census reports of British India show that in the Punjab, on both sides of the present international frontier, the same castes, such as Ahir, Jat, and Gujjar were present among Hindus, Muslims, and Sikhs. When a section of a Hindu caste embraced Islam or Sikhism, it adopted a new religion but did not abandon its old and distinct collective identity.

Hierarchical Organization

The accommodation of diversity did not mean that the tribes, clans, castes, sects, and communities that were allowed or even encouraged to maintain their distinctive identities were all placed on the same plane of equality. Diversity was organized on the basis of hierarchy and not of equality, and that is the third distinctive feature of the traditional civilization of South Asia. It is true that all communities were allowed or even encouraged to maintain their distinctive styles of life; but they were not all equally esteemed. While all premodern civilizations were hierarchical to a greater or lesser extent, the principle and practice of hierarchy were carried furthest on the Indian subcontinent, particularly among the Hindus.

Hierarchy is not simply a matter of inequality in the distribution of material resources or of life chances; it is above all a system of values according to which the division of the

world into social groups deemed to be of unequal worth is a part of the natural scheme of things. The hierarchical conception of society found its fullest expression in the classical Hindu texts known as the *Dharmashastra*, which provided the framework for Hindu law. Although the law has changed radically in the direction of equality, the bias of custom is still often in favor of hierarchy.

The Hindu caste system has been represented as the prototype of a hierarchical social order. But hierarchy extends beyond caste and permeates social relations and social institutions of the most diverse kinds among Hindus, Muslims, Buddhists, and others throughout the region. Despite the doctrine of equality characteristic of Islam, agrarian relations and gender relations are cast in a deeply hierarchical mold in Pakistan and to some extent also in Bangladesh.

The Village

Colonial administrators throughout the nineteenth century saw the subcontinent as essentially a land of villages. They have left behind memorable accounts of the unity, the self-sufficiency, and the continuity of the Indian village. A large number of village studies have been conducted in India by professional anthropologists and others since the country became independent in 1947. Similar studies have been conducted in Sri Lanka, Pakistan, Bangladesh, and Nepal. In the paragraphs below there follows a brief account of the broad social characteristics of the Indian village, many of which may also be found in villages in the other countries just mentioned.

In the past, the Indian village had a distinct physical identity, which became particularly prominent during floods, famines, and other natural calamities. Most villages were nucleated, although there were also dispersed villages, depending in part on the terrain. The life of the village was based on an economy of land and grain, and its rhythm was, and to a large extent still is, governed by the annual cycle of agricultural activities. The village also has an annual cycle of religious ceremonies and festivals, and the two cycles are related closely. With the development of transport and communication in the twentieth century, and particularly since the 1950s, the isolation of the village is becoming progressively reduced, although not at the same pace in all parts of the country or the region. The less-isolated villages are now becoming urbanized, though in a highly uneven manner.

The social morphology of the village shows considerable variation in terms of size and internal differentiation. Irrigated villages in river valleys and close to the ancient and medieval centers of civilization tend to have an elaborate structure; 'dry' villages in the remote areas tend to be smaller and simpler in their structure. Only the latter may be described as communities of peasants. The former are often highly differentiated and stratified. They include noncultivating landowners, owner–cultivators, cultivating tenants, sharecroppers, and agricultural laborers, in addition to groups of families associated with various specialized crafts and services, usually on a hereditary or quasi-hereditary basis. Agrarian reform and the expansion of the market are changing some of this, although many of the older features remain.

Its relative isolation and self-sufficiency and its well-knit division of labor between agriculture, crafts, and services gave the Indian village a certain unity. In the nineteenth century – and even later – it was often described as a 'little republic.' But it was not a republic of equals. In the larger irrigated villages the social boundaries between the superior, intermediate, and inferior sections of the community – landowners, tenants, agricultural laborers, priests, potters, and scavengers – were clearly maintained. If there was a certain unity in such villages, as indeed there was, it is best described as a 'vertical unity.' The unity of the village has been weakened in recent decades, but that does not mean that the inequalities have all disappeared.

While it is true that South Asia is a land of villages, one cannot ignore the presence of towns and cities in it. The subcontinent has cities going back to ancient and medieval times, and India had urban centers before most of Europe did. In the past, towns and cities grew around places of worship, the courts of kings and princes, and commercial and trading centers. Colonial rule led to the creation of a new kind of city, mainly along the coast, such as Calcutta, Madras, Colombo, Bombay, and Karachi. In the years since the end of colonial rule, industrialization has added to the population of existing cities and led to the creation of new ones. The urban middle class and the industrial working class are important factors in the economic and political life of contemporary India.

Returning briefly to the Indian village, a prominent place was occupied in it by caste. Indeed, caste may be viewed as the social basis of the division of labor in it, and the differentiation and stratification in the Indian village cannot be understood without taking caste into account.

Caste

Caste is an institution of great antiquity, and it has shown resilience in the face of the economic and political changes of the twentieth century. In its origin and fullest development, it is an institution of the Hindus, but near analogs of the Hindu caste system may be found among Sikhs, Muslims, and Christians on the subcontinent and among Buddhists in Sri Lanka; and it is of course very important in the predominantly Hindu population of Nepal. On the doctrinal plane, Hinduism, Buddhism, and Islam are different from each other; on the plane of everyday life at the local level, social divisions among Hindus, Muslims, and Buddhists act in broadly similar ways in most parts of South Asia.

The caste, the subcaste, and even the subsubcaste, to each of which the term *jati* may be applied, maintained its identity and continuity through the regulation of marriage. The most general rule was the rule of endogamy, and among Hindus this was supplemented by the rule of hypergamy. In the past, Hindu law prohibited intercaste marriage except under very, specific conditions; the law has now changed, but the bias of custom is still in favor of endogamy among the Hindus, and to some extent also among the others.

Every caste or *jati* was in principle associated with a specific occupation, although agriculture was open to most if not all castes. The association between caste and occupation was stronger among the Hindus than among the others, but from

the end of the nineteenth century, it began to weaken even among the former. A very important factor behind the weakening of the association has been the emergence, since the middle of the nineteenth century, of a large and increasing number of occupations that are 'caste-free' in the sense that they have grown outside the framework of the old division of labor based on caste.

The different castes are socially ranked; indeed, it is in the caste system that the principle of social ranking finds its most extreme and elaborate expression. In South Asia, castes are ranked among all religious groups, but the ranking is not as rigid or as exhaustive among the non-Hindus, as it is among the Hindus. Caste ranking among the Hindus was expressed in the idiom of purity and pollution, but that idiom is now less pervasive than before, and new criteria of social ranking, having little to do with caste in the traditional sense, are becoming increasingly salient, particularly in the urban areas.

Family, Kinship, and Marriage

Given the great significance of the rules regulating marriage and in particular the rule of endogamy, it is not surprising that there is a close association between caste and kinship. Through repeated intermarriage from generation to generation, all the members of the local section of a caste or subcaste become related to each other by real or putative ties of kinship. This is generally true of all the major religious and linguistic divisions of South Asian society.

Much has been written about the prevalence of the joint family in the subcontinent and the changes it is undergoing as a result of urbanization, industrialization, and modernization. The Hindu undivided family had certain legal characteristics peculiar to itself, and some of these are still recognized by the law in India. One important feature of the family in the subcontinent, whether among Hindus, Sikhs, Muslims, or Christians, is that marriages are still to a large extent arranged by parents and older relatives, and the new couple generally lives in the parental home immediately after marriage, even though it may set up a separate household after one or more children are born.

Reference has already been made to wide variations in the rules of descent, inheritance, and marriage across the many and diverse communities that inhabit the region. But through all these variations, kinship has a significant place in social life everywhere and among all communities. Relatives are numerous and various, and the obligations of kinship are cherished and cannot be denied easily. The rules of endogamy act against the formation of real ties of kinship between members of different castes; but they do not prevent the formation of fictive ties of kinship between them. The idiom of kinship is used extensively in interpersonal relations across castes and communities in villages throughout the region.

Changes are undoubtedly taking place in the size and composition of the household, particularly in the rapidly growing cities. But even where nuclear households outnumber extended ones, the matrix of kinship within which they operate gives them a different character from their counterparts in the West. The boundaries between households are porous, and no household is secure against intrusion by cousins, uncles, aunts, nephews, nieces, and grandparents.

In discussing caste, kinship, and family it is impossible to ignore the subject of gender, although only a passing reference to it is possible here. Despite variations according to language, religion, and caste, and between patrilineal and matrilineal communities, women have since time immemorial occupied a subordinate position in South Asian societies. Polygamy was sanctioned by law among both Hindus and Muslims. In India, it is now prohibited by law among Hindus but not among Muslims; it is permitted by law in Pakistan, and attempts to regulate polygamy by law have had little success among Muslims. The position of women is changing in all countries and among all communities in South Asia, although the change is highly uneven. Here, more important than the variations of language, religion and caste are those that arise from education, occupation, and income.

New Institutions and Classes

The first impediment against attempting a concise account of society and culture in South Asia arises from the problem of variation; the second impediment arises from the problem of change. Two important developments in the subcontinent since the middle of the nineteenth century have to be noted. These are, first, the growth of open and secular institutions such as schools, colleges, universities, libraries, laboratories, hospitals, law courts, and newspapers; and, second, the emergence of a new middle class based on a new educational system and a new occupational system. These two related developments have taken place everywhere, but most conspicuously in India where they are now a significant feature of the social and cultural landscape.

Open and secular institutions are sustained by norms and values that are very different from the norms and values by which the traditional institutions of South Asian societies have been sustained. The new institutions are often at odds with the old ones, which exert various kinds of pressures on them and sometimes threaten their very survival. At the same time, and particularly in a country like India, they have grown to an extent that makes it difficult to dispense with them or to devise alternatives to them that are fully in harmony with the institutions of the past. As a result, the conflict of norms and values has become endemic in these societies.

The new middle class is both the creation and the creator of the open and secular institutions just referred to. The origins of the Indian middle class go back to the first half of the nineteenth century. Its first beginnings took place in the presidency capitals of Calcutta, Bombay, and Madras in each of which a new university was established in 1857. The graduates of these universities and of the others which came up in course of time became administrators, managers, lawyers, doctors, scientists, teachers, and clerks. The Indian middle class is now very large, at well over 100 million persons. Although it is still a relatively small proportion of the population of the country, it is among the largest and the most differentiated of the middle classes in the world.

The emergence of a new and increasingly dynamic middle class has not led to the effacement of the old identities and

distinctions based on language, religion, and caste. Those still remain, although they are now coming to be organized in a different way. The middle class in turn is differentiated not only in terms of occupation, income, and education, but also in terms of language, religion, and caste. Caste and class reinforce each other in some respects and cut across each other in others. While there is some individual mobility, it does not lead to the equal distribution of all castes and communities among the different economic strata. The social mosaic is, thus, becoming in some sense even more complex than in the past.

Media, Movements, and Modernity

Since the turn of the millennium, South Asia has rapidly appeared as a prominent region in a globalizing world. The Indian economy, during the decades after independence largely restricted by state control, has been liberalized and this has strongly influenced not only trade and industry, but also social and cultural aspects of society. A wide range of 'foreign' consumer goods are now available for those that can afford them, and members of an urban middle class are attracted by job opportunities in multinational companies or new domestic enterprises producing for the global market. Thus, the social and cultural environment of the middle class is not confined to postcolonial institutions in the South Asian social landscapes. The horizon of expectations, not least regarding matters of education and occupation, extends far outside the region.

This is of course a reality more accessible for wealthier sections of urban India – and particularly for people in cities which have emerged as hubs of the global Information Technology industry, such as Bangalore and Hyderabad – than for that vast majority of the population which is still outside this world of new possibilities. But also for the less fortunate, modern consumer goods, proper education, and successful occupational careers have become visions to struggle for. One reason for this is the enormous growth of mass media and other communication technologies through which images and knowledge of emergent life styles are disseminated.

Only a few decades ago, newspapers and television were media relevant mostly for a privileged, often English-speaking, middle class; while domestic film industries produced escapist entertainment for poor mass audiences. Today this cleavage has decreased. Daily newspapers and regional television channels are available in vernacular languages and popularized forms; simultaneously, the film industry in India – popularly known as Bollywood – is increasingly producing movies in order to attract the growing middle class and even a global audience, beside village India. The mobile phone has rapidly become a technical commodity in mass use.

While, in some sense, the cultural rift (between urban and rural, or modernity and tradition) within South Asia has diminished, social and economic differences have not. Much due to the influence of mass media, inequalities have rather become more visible and not least contested. Later decades has seen a rise of social movements and new political formations, which have changed the complexion of politics. In India, Hindu nationalism, low caste (*Dalit*) mobilization, environmentalism, anticorruption campaigns, and various regional movements have been on a steady rise, and at times agitations on specific issues have escalated in violence. Also this development could be interpreted as a sign of a how South Asia is transforming toward universal forms of norms and discourses. However, it could also be argued that certain traditional social and cultural aspects still prevail, for example the importance of family relations, religion, and the privileges of caste – even if in modified forms. Thus South Asia accommodates its own forms of hybrid modernities.

See also: Area and International Studies: Cultural Studies; Caste; Colonization and Colonialism, History of; Communalism.

Bibliography

Alavi, H., Harriss, J. (Eds.), 1989. South Asia. Macmillan, Basing-Stoke, UK.
Aneesh, A., 2006. Virtual Migration: The Programming of Globalization. Duke University Press, Durham and London.
Baviskar, A., 2004. In the Belly of the River: Tribal Conflicts over Development in the Narmada Valley. Oxford University Press, New Delhi.
Baviskar, A., Ray, R., 2011. Elite and Everyman: The Cultural Politics of the Indian Middle Class. Routledge, London and New Delhi.
Béteille, A., 1991. Society and Politics in India. Athlone Press, London.
Biao, X., 2007. Global "Body Shopping": An Indian Labor System in the Information Technology Industry. Princeton University Press, Princeton, NJ.
Breckenridge, C., 1996. Consuming Modernity: Public Culture in Contemporary India. Oxford University Press, New Delhi.
Breman, J., Kloos, P., Saith, A. (Eds.), 1997. The Village in Asia Revisited. Oxford University Press, New Delhi.
Dumont, L., 1980. Homo Hierarchicus. University of Chicago Press, Chicago.
Fuller, C.J. (Ed.), 1996. Caste Today. Oxford University Press, New Delhi.
Ganti, T., 2012. Producing Bollywood: Inside the Contemporary Hindi Film Industry. Duke University Press, Durham and London.
Hardtmann, E.-M., 2009. The Dalit Movement in India: Local Practices, Global Connections. Oxford University Press, New Delhi.
Jeffrey, R., 2000. India's Newspaper Revolution: Capitalism Politics and the Indian Language Press 1077–1999. Oxford University Press, New Delhi.
Madan, T.N. (Ed.), 1976. Muslim Communities of South Asia. Vikas, New Delhi, India.
Mandelbaum, D.G., 1970. Society in India, 2 vols. University of California Press, Berkeley, CA.
Mazzarella, W., 2003. Shoveling Smoke: Advertising and Globalization in Contemporary India. Duke University Press, Durham.
Rajagopal, Arvind, 2001. Politics after Television. Hindu Nationalism and the Reshaping of the Public in India. Cambridge University Press, Cambridge.
Srirupa, R., 2007. Beyond Belief: India and the Politics of Postcolonial Nationalism. Duke University Press, Durham and London.
Shah, A.M., 1998. The Family in India. Orient Longman. India, New Delhi.
Srinivas, M.N., 1976. The Remembered Village. Oxford University Press, New Delhi.

Assimilation of Immigrants

Rubén G Rumbaut, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the evolution of the concept of assimilation in American social science. It distinguishes *assimilation* from *accommodation* as modal adaptation outcomes of different immigrant generations, as well as various aspects that are commonly conflated by the concept (cultural adaptations, economic mobility, social acceptance into a native mainstream); discusses interrelated cultural (subtractive and additive acculturation), structural (primary and secondary integration), and psychological (identification) dimensions of the concept; and describes the process of 'segmented assimilation' – how it is that different groups, in varying contexts of reception and incorporation, adapt to and are absorbed into different sectors of the society.

"What happens 'when peoples meet'?" was the question with which Milton Gordon (1964: p. 60) opened his seminal chapter on "The Nature of Assimilation," referring to the "processes and results … of ethnic meetings in the modern world," which take place in a wide range of contexts from colonial conquest and military occupation to the displacement of indigenous peoples and to large-scale voluntary immigration. He identified seven 'variables' (or 'stages or subprocesses') of a complex 'assimilation process' to describe, in the American context, what happens to the 'sense of peoplehood' of ethnic groups, that is, of "groups defined by race, religion, and national origin, or some combination of these categories" (1964: pp. 27, 69–83).

International migrations produce profound social changes in both sending and receiving societies, in intergroup relations within receiving societies, and among the immigrants and their descendants. In varying contexts of reception, immigration is followed predictably not only by acculturative processes on the part of the immigrants, but also by varying degrees of acceptance, intolerance, or xenophobia about the alien newcomers on the part of the natives, which in turn shape the immigrants' own modes of adaptive response and sense of belonging. And immigration engenders ethnicity – collectivities who perceive themselves and are perceived by others to differ in language, religion, 'race,' national origin or ancestral homeland, cultural heritage, and memories of a shared historical past. Their modes of incorporation across generations may take a variety of forms – some leading to greater homogenization and solidarity within the society (or within segments of the society), others to greater ethnic differentiation and heterogeneity (Yinger, 1981, 1994; Massey and Sánchez, 2010). Those outcomes are shaped by the complex interaction of the (welcoming, neutral, or hostile) contexts of incorporation into which the immigrants are received by the society, and the resources and adaptations of the immigrants and their descendants (Portes and Rumbaut, 2014).

Assimilation is one such possible outcome. Sociologically, it has been defined as a multidimensional process of boundary reduction and brokering, which blurs or dissolves an ethnic distinction and the social and cultural differences and identities associated with it (Alba and Nee, 2003). At its hypothesized end point, formerly distinguishable ethnocultural groups become effectively blended into one. At the group level, assimilation may involve the absorption of one or more minority groups into the mainstream, or the merging of minority groups – e.g., second-generation West Indians 'becoming black Americans.' At the individual level, assimilation denotes the cumulative changes that make individuals of one ethnic group more *acculturated, integrated,* and *identified* with the members of another.

Ideologically, the term has been used in the past to justify selective state-imposed policies aimed at the eradication of minority cultures and the 'benevolent' conquest of other peoples. Two notorious examples are the campaigns, encouraged by the Dawes Act of 1887, to Americanize, Christianize, and 'civilize' American Indian children by removing them from their families and immediate environments and moving them into boarding schools like the Carlisle Indian School in Pennsylvania; and the 1898 'Benevolent Assimilation' policy of the United States to colonize and pacify the Philippines and quash its struggle for independence, pursuing an imperial interest under the guise of idealized purpose and beneficent intent (Miller, 1982). But the ideal of assimilation, more often tied to the metaphor of the 'melting pot,' has sought to dramatize, legitimize, and celebrate the consensual integration of immigrants and their descendants into a common national life – while ignoring the realities, inequalities, and potentialities of conflictual intergroup relations marked by enduring segregation, discrimination, marginalization, and exclusion.

In *Beyond the Melting Pot* (1963), their classic study of "the Negroes, Puerto Ricans, Jews, Italians, and Irish of New York City," Nathan Glazer and Daniel P. Moynihan concluded that the concept of assimilation could not elucidate the complex realities of the nation's largest city:

> To what does one assimilate in modern America? The 'American' in the abstract does not exist … The point about the melting pot is that it did not happen … The notion that the intense and unprecedented mixture of ethnic and religious groups in American life was soon to blend into a homogeneous end product has outlived its usefulness, and also its credibility. The persistent facts of ethnicity demand attention, understanding, and accommodation. (xcvii, 20, 290)

Evolution of the Concept

In academic and colloquial usage, in social science, public policy, and popular culture, the idea and the ideal of 'assimilation' have had a checkered history. A protean concept, over time the term has conflated various empirical descriptions and normative prescriptions to make sense of the incorporation of 'ethnic' difference in American life. The concept is used to refer at once to cultural adaptations, economic mobility, patterns of settlement and intermarriage, and social acceptance into a changing native mainstream. After more than a century of use and misuse, the term itself remains elusive, confusing, and contentious. For a 'canonical' concept, there remains surprising ambiguity as to its meaning, measurement, and applicability.

The earliest uses of 'immigrant assimilation' confirm the sociologist Robert Park's observation that it was a metaphor derived from physiology to describe, as in a process of digestion and absorption, how "alien peoples come to be incorporated … with a community or state" (1914: p. 611). As originally used, assimilation was a *transitive* verb, entailing the 'swallowing and digesting' (by the incorporating community or state, the subject of the action) of alien peoples (the object). Thus, a 19 May 1852 article in the newly founded *New York Times* observed that "The population of the United States is supplied by the world … Assimilation in America, thanks to a healthy and useful digestion, is equal to the largest supplies of aliment … Politically, the influx of life is only pernicious where the assimilative functions of labor and compensation are quiescent or disordered." A generation later, at the start of a new immigration from Southern and Eastern Europe, the *Times* editorialized on 15 May 1880, that "There is a limit to our powers of assimilation and when it is exceeded the country suffers from something very like indigestion … We know how stubbornly conservative of his dirt and ignorance is the average immigrant who settles in New York … these wretched beings change their abode, but not their habits in coming to New York." By 9 August 1903, another *Times* editorial noted that "Little more than one half of the people of the country at the opening of the twentieth century are of American parentage … Obviously the task of assimilation imposed on the American people is considerable." It was the 'American people' who did the assimilating; it was up to the 'Anglo-Saxon race' to absorb the foreigners.

In the twentieth century, 'assimilation' came increasingly, if almost imperceptibly, to be used as an *intransitive* verb. It was now the aliens who were the subject of the action of adapting and changing themselves *to* American cultural standards – 'acculturating,' as the process would later be called (Redfield et al., 1936) – and thereby embracing (or expected to embrace) a common national loyalty and identity. Already in Chicago in 1914, with immigration unabated and the large majority of its residents consisting of immigrants and their children, Robert Park (who saw the modern construction of national identities as entailing both the incorporative and acculturative modes of intergroup change) could write that:

> In America it has become proverbial that a Pole, Lithuanian, or Norwegian cannot be distinguished, in the second generation, from an American born of native parents … As a matter of fact, the ease and rapidity with which aliens, under existing conditions in the United States, have been able to assimilate themselves to the customs and manners of American life have enabled this country to swallow and digest every sort of normal human difference, except the purely external ones, like color of the skin.
>
> 1914: p. 608

'Americanization' became the synonym of assimilation – all the more during the national mobilizations for '100 percent Americanism' spurred by World War I, the Bolshevik Revolution, and their aftermath (Higham, 1955). Popular usage reflected the dominant currents of thought of the times. The sociological concept of 'assimilation' was honed during an era of rapid industrialization, urbanization, and a mass 'new immigration' from Southern and Eastern Europe in the late nineteenth and early twentieth centuries. So was American sociology, especially in Chicago, which became a natural laboratory for the study of the immigrant and the city. Grand narratives, such as Emile Durkheim's depiction of the shift from 'mechanical' to 'organic solidarity' in *The Division of Labor in Society* (1893), sought to grasp the transition from premodern folk to modern industrial society; they undergird the subsequent development of functionalist theories of modernization. Progressive reforms, rooted in a belief in the human ability to improve social conditions, especially with the aid of experts and rational efficiency, sought to grapple with the attendant problems of large-scale social integration. The teleological notion of an endlessly improving future and positivist assumptions of linear progress were made credible by the rapid expansion of science, technology, and economic innovation. The idea of progress came to dominate the worldview of the entire culture. The ideal of assimilation reflects that optimism in progress as a universal law.

Assimilation in American Sociology: Three Classic Formulations

The concept of assimilation seeks to grasp a contextual and not solely an individual reality that is complex, relational, and multidimensional. In what became arguably the most influential text ever published in the history of American sociology, Robert Park and Ernest Burgess gave the concept of assimilation its first classic definition: "a process of interpenetration and fusion in which persons and groups acquire the memories, sentiments, and attitudes of other persons and groups, and, by sharing their experience and history, are incorporated with them in a common cultural life" ([1921]1924: p. 735). They distinguished systematically between 'four great types of interaction' – competition, conflict, accommodation, and assimilation – which they related respectively to economic, political, social, and cultural institutions. *Competition* and *conflict* sharpen ethnic boundaries and the consciousness of intergroup difference. An *accommodation* (of a conflict or to a new situation) may take place quickly, and the person or group is typically a highly conscious protagonist of the process of accommodating those circumstances. In *assimilation*, by contrast, the changes are more subtle and the

process is typically unconscious, so that the person is incorporated into the common life of the group largely unaware of how it happened. Assimilation is thus very unlikely to occur among immigrants who arrive as adults. Instead, *accommodation* most closely reflects the modal adaptation of first-generation adult immigrants, while *assimilation* can become a modal outcome ultimately only for the malleable young and for the second and third generations, and then only if and when permitted by structural conditions of inclusion at the primary group level.

Assimilation thus defined takes place most rapidly and completely in primary (intimate and intense) social contacts, including intermarriage. *Accommodation* may be facilitated through secondary contacts, but they are too distant and remote to promote assimilation. Since the nature (especially the interpersonal intimacy, 'the great moral solvent') of the social contacts is what is decisive, it follows that 'a common language is indispensable for the most intimate associations of the members of the group,' and its absence is 'an insurmountable barrier to assimilation,' since it is through communication that gradual and unconscious changes of the attitudes and sentiments of the members of the group are produced.

But – crucially – language and acculturation alone cannot ensure assimilation if a group is categorically segregated, racially classified, and 'regarded as in some sense a stranger, a representative of an alien race.' That is why, Park emphasized in a later encyclopedia article (1930: p. 282), the English-speaking Protestant "Negro, during his three hundred years in this country, has not been assimilated … not because he has preserved in America a foreign culture and an alien tradition, for with the exception of the Indian … no man in America is so entirely native to the soil." Race and place (i.e., racial discrimination and residential segregation) become critical structural determinants of the degree of assimilation or dissimilation precisely insofar as they delimit possible forms of primary social contact and heighten social contrasts and conflict. To be considered assimilated, it is not enough to acquire "the language and social ritual of the native community" – i.e., to be acculturated – but also to be able to "participate, without encountering prejudice, in the common life, economic and political" (1930: p. 281). A century later, much the same could be said of millions of highly acculturated undocumented immigrants from Mexico, Central America, and elsewhere who are nonetheless excluded from the legal labor market and subject to systematic state persecution and unprecedented levels of deportations (Kanstroom, 2007; Ngai, 2005).

The 'melting pot' metaphor had been dismissed by Park and Burgess as a "'magic crucible' notion of assimilation" where "the ideal of assimilation was conceived to be that of feeling, thinking, and acting alike" (1924: p. 735). The end result of assimilation is not 'like-mindedness,' but rather "a unity of experience and orientation, out of which may develop a community of purpose and action … The extent and importance of the kind of homogeneity and 'like-mindedness' that individuals of the same nationality exhibit has been greatly exaggerated. Like-mindedness … contributes little or nothing to national solidarity" (Park, 1914; reproduced in Park and Burgess, 1924: p. 759). Instead, Park and Burgess argued,

> Not by the suppression of old memories, but by their incorporation in his new life is assimilation achieved … Assimilation cannot be promoted directly, but only indirectly, that is, by supplying the conditions that make for participation. There is no process but life itself that can effectually wipe out the immigrant's memory of his past. The inclusion of the immigrant in our common life may perhaps best be reached, therefore, in cooperation that looks not so much to the past as to the future. The second generation of the immigrant may share fully in our memories, but practically all that we can ask of the foreign-born is participation in our ideals, our wishes, and our common enterprises.
>
> 1924: pp. 739–740

Ironically, despite his wide-ranging writings on many subjects, Park may be best known for the formulaic notion that assimilation is the final stage of a natural, progressive, inevitable, and irreversible 'race relations cycle.' But in a prolific career, Park only wrote about a 'race relations cycle' twice: first in a single sentence near the end of a 1926 article, "Our Racial Frontier in the Pacific"; then a decade later in a brief introduction to a book on interracial marriage in Hawaii written by one of his former students. In the first instance, he was arguing against the likelihood that a 'racial barrier' – which the passage of exclusionary laws sought to establish by barring Asian migration to the United States – could be much of a match against global economic, political, and cultural forces that have brought about "an existing interpenetration of peoples … so vast and irresistible that the resulting changes assume the character of a cosmic process" (Park, 1926: pp. 141, 149). And in his 1937 introduction, he explicitly rebutted any notion of a unilinear assimilative outcome to race conflict and change ("what are popularly referred to as race relations"), arguing instead that when stabilization is finally achieved, race relations would assume one of three configurations: "They will take the form of a caste system, as in India; they will terminate in complete assimilation, as in China; or the unassimilated race will constitute a permanent racial minority within the limits of a national state, as in the case of the Jews in Europe … All three types of change are involved … in what we may describe as the 'race relations cycle'" (Park, 1937: xiii).

What has come to be called 'straight-line theory' (see Gans, 1992a,b) has little basis in the theories of Park or Gordon, as is often if mistakenly asserted, but gained traction from W. Lloyd Warner and Leo Srole's *The Social Systems of American Ethnic Groups*, a definitive statement on the subject published in 1945. (The data, collected in the early 1930s, formed the basis of Srole's 1940 doctoral dissertation, "Ethnic Groups and American Society: A Study in the Dynamics of Social Assimilation."). They described the progressive advance of eight European-origin groups in the major status hierarchies of 'Yankee City' (Newburyport, Massachusetts): the Irish, French Canadians, Jews, Italians, Armenians, Greeks, Poles, and Russians. In the immigrant generation, all except the Jews had come from rural backgrounds. From the first chapter ("The Melting Pot," consisting of seven illustrative personal stories) to the last ("The American Ethnic Group"), the book systematically analyzed their spatial distribution, economic life, class system, family, church, language and the school, and associations, and spelled out a set of innovative methods for the study of ethnic groups.

Warner and Srole developed six-point linear indices to measure residential, occupational, and class status; and 'assimilation' and 'subordination' scales based on specific criteria to estimate the time for an entire group to disappear (the final result of assimilation), the proportionate number of people who drop out of a group in each generation, and the amount and kind of participation permitted members of the group by the 'host society.'

They explicitly linked upward social mobility to assimilation, which they saw as determined largely by the degree of ethnocultural (religion and language) and, above all, racial difference from the dominant group. While 'racial groups' were subordinated and excluded through caste restrictions on residential, occupational, associational, and marital choice, the clash of 'ethnic groups' with the dominant institutions of the host society was not much of a contest, particularly among the young. The industrial economy, the polity, the public school, popular culture, and the American family system all undercut and absorbed ethnicity in various ways, so that even when "the ethnic parent tries to orient the child to an ethnic past … the child often insists on being more American than Americans" (1945: p. 284). For the upwardly mobile, with socioeconomic success came intermarriage and the further dilution of ethnicity. They concluded that "it is the degree of racial difference from the white American norms which counts most heavily in the placement of the group and in the determination of its assimilation"; absent such discrimination and structural inequalities, however, "the future of American ethnic groups seems to be limited; it is likely that they will be quickly absorbed" (pp. 294–295).

That general if qualified view of assimilation as linear progress, with sociocultural similarity and socioeconomic success marching in lock step, was reformulated by Milton Gordon in *Assimilation in American Life* (1964), published ironically on the eve of the beginning of the latest era of mass immigration to the United States – and of the denouement of the concept itself in the wake of the 1960s. From the opening sentence of the book, Gordon focused on the relational and contextual character of the process: "This book is concerned, ultimately, *with problems of prejudice and discrimination arising out of differences in race, religion, and national background* among the various groups, which make up the American people" (p. 3; emphasis added).

Although he meticulously reviewed a wide variety of definitions of 'assimilation' in the scholarly literature, he did not explicitly provide his own. Instead, he broke down the assimilation sequence into seven steps, of which 'identificational assimilation' – a self-image as an unhyphenated American – was the end point of a *hypothetical* process that began with cultural assimilation, proceeded through structural assimilation and intermarriage, *and* was accompanied by an absence of prejudice, discrimination, and value conflict in the 'core society.' Once structural assimilation occurred (i.e., extensive primary-level interaction with members of the 'core group'), either in tandem with or subsequent to acculturation, "the remaining types of assimilation have all taken place like a row of tenpins bowled over in rapid succession by a well-placed strike" (1964: p. 81). For the children of white European immigrants, at least, the acculturation process was so 'overwhelmingly triumphant' that the greater risk consisted in alienation from family ties and in role reversals of the generations that could subvert parent–child relationships.

Nonetheless, in reviewing the actual evidence for the assimilation sequence in American life, Gordon reached very different conclusions than those that are habitually ascribed to him. He coined the term 'ethclass' to refer to the stratified segment of social space created by the intersection of ethnicity and social class, which he saw as "fast becoming the essential form of the subsociety in America" (p. 51), and proposed a series of hypotheses about contextual variations in cultural behavior, social participation, and group identity. He found that "The most salient fact … is the maintenance of the structurally separate subsocieties of the three major religions and the racial and quasi-racial groups, and even vestiges of the nationality groupings, along with a massive trend toward acculturation of all groups – particularly their native-born – to American culture patterns." In short, "the sense of ethnic belonging has survived … American society has come to be composed of a number of 'pots' … The entire picture may be called a 'multiple melting pot'" (24–25, 67, 130–131). Anticipating what 'segmented assimilation' would assert in the 1990s, he concluded that "Structural pluralism is the major key to the understanding of the ethnic makeup of American society, while cultural pluralism is the minor one" (p. 159).

Gordon was aware of the ways in which the real and the rhetorical, the ideal and the ideological, get wrapped up in the idea of assimilation. 'Cultural pluralism' as an ideology did not match empirical realities, and the theory of the 'Melting Pot' exhibited "a considerable degree of sociological naiveté" (p. 129). And while 'Anglo-conformity' was the most prevalent ideology of assimilation in American history, he also noted that "Structural assimilation turned out to be the rock on which the ship of Anglo-conformity foundered. And if structural assimilation, to a large degree, did not take place, then in similar measure amalgamation and identificational assimilation could not" (p. 114).

Historians have seen the apogee of the concept of assimilation in the 1950s and early 1960s as reflecting the need generated by World War II for national unity, and the postwar tendency to see American history as a narrative of consensus rather than conflict; and the political and social upheavals of the 1960s (nationally and internationally) as shattering the 'consensus school' and the rationale for studying assimilation, bringing back instead a focus on the ethnic group and ethnic resilience, and more inclusive conceptions of American society. As the notion of an Anglo-American core was delegitimized amid the conflicts and ethnic reassertions of the 1960s, assimilation lost its allure (Kazal, 1995). But by the 1990s, once more well into a new era of mass immigration, a systematic reevaluation of the concept of assimilation emerged, with applications in contemporary scholarship seeking to contrast differences and similarities between the old and the new immigration.

Incorporation Processes: Acculturation, Integration, Identification

Combining the various emphases currently given to the term, 'assimilation' involves a series of interrelated but analytically

distinct cultural (acculturation), structural (integration), and psychological (identification) dimensions. Gordon referred to them as "the three crucial variables of group identity, social participation, and cultural behavior as they pertain to the subsociety of the ethclass" (1964: p. 51). These processes do not unfold in a social vacuum; contextual and group factors shape each of these dimensions either by promoting or precluding assimilative outcomes (Yinger, 1981, 1994; Massey and Sánchez, 2010). In the end, there can be no 'assimilation' without acceptance and inclusion by the dominant society.

Acculturation, which comes closest to the common sense notion of 'melting,' involves complex processes of cultural diffusion and changes producing greater linguistic and cultural similarity between two or more groups. Its homogenizing influences are generally more extensive among members of smaller and weaker groups, and particularly (voluntary) immigrant groups – and more rapidly achieved with respect to what Gordon called 'extrinsic culture traits.' Nonetheless, acculturation is never exclusively one-sided; dominant groups too are culturally influenced by their contacts with other ethnocultural groups in the society (Alba and Nee, 2003). In the American experience, language shifts have been overwhelmingly one-sided, with the switch to monolingual English typically being accomplished by the third generation. Acculturation proceeds more rapidly among children than adults, and linguistic and other acculturative gaps commonly develop in immigrant households between parents and children; Portes and Rumbaut (2001) have identified three such acculturative patterns in parent–child relationships, labeled *dissonant*, *consonant*, and *selective acculturation*.

At the individual level, a key distinction is between *subtractive* (or *substitutive*) *acculturation* and *additive acculturation*. The first is essentially a zero-sum game that involves giving up some elements of a cultural repertoire (such as language and memory itself), while replacing them from another; the second does not involve losing so much as gaining to form and sustain a more complex repertoire (bilingualism and biculturalism). Available research has yet to examine systematically the multiplicity of conditions and contexts yielding subtractive versus additive acculturative outcomes, although in the United States at least it has proved exceedingly difficult to sustain fluent bilingualism beyond the second generation (Portes and Rumbaut, 2014).

The degree of acculturation, as noted earlier, is by itself not a sufficient condition for 'assimilation' (the two terms are not synonymous but are often used equivalently; see Gans, 1997). *Structural integration* was the crux of the matter for Gordon, although what he had in mind was the entrance of the minority group into "the social cliques, clubs, and institutions of the core society at the primary group level" (not 'parity' with the majority group on such indices as income or education, or 'regression to the mean' – notions of standardization implicit in uses of 'assimilation' as an intransitive verb). Given the many different institutions involved – and the fact that integration into the economy, the polity, and the community at the secondary group level is ignored by that formulation – a conceptual distinction can be made between *primary* and *secondary* dimensions. The latter refers to a wide range of integrative processes within secondary groups, including socioeconomic and spatial (residential) integration, and the acquisition of legal citizenship as a full-fledged member of the polity. The former – extensive interaction within personal networks and primary relationships, including intermarriage – is unlikely to take place under conditions of status inequality. While all of these dimensions (acculturation, integration, intermarriage) are interdependent to varying degrees, the linkages between them are historically contingent and will vary depending on a number of factors, notably social class, racial framing, and the context of reception within which different groups are incorporated.

Conventional accounts of shifts in ethnic *identification* among the descendants of European immigrants, conceived as part of a linear assimilative process, have pointed to the 'thinning' of their ethnic self-identities in the United States. For their descendants, at least, one outcome of widespread acculturation, social mobility, and intermarriage with the native population was that identity became an optional form of 'symbolic' ethnicity (Gans, 1979; Alba, 1990; Waters, 1990). As the boundaries of those identities become blurred and less salient, less relevant to everyday social life, the sense of belonging and connection to an ancestral past faded – into what Richard Alba called "the twilight of ethnicity." This mode of ethnic identity formation, however, was not a simple linear function of socioeconomic status and the degree of acculturation – i.e., of linguistic and other cultural similarities with the dominant group – but hinged also on the context of reception and the degree of discrimination and racialization experienced by the subordinate group.

Identity shifts, like acculturative changes, tend to be from lower to higher status groups. But where social mobility is blocked or hindered by prejudice and discrimination, members of lower status groups may react by reaffirming their shared identity. This process of forging a *reactive ethnicity* in the face of perceived threats, persecution, discrimination, and exclusion is not uncommon (Portes and Rumbaut, 2001). On the contrary, it is another mode of ethnic identity formation, accounting for the 'thickening' rather than the dilution of ethnicity. At the extreme, as reflected in the African-American experience, the result can be the sense of 'double consciousness' of which W.E.B. DuBois wrote eloquently.

Compared to language loyalty and language shift, generational shifts in ethnic self-identification are far more conflictual and complex. Paradoxically, despite the rapid acculturation of European immigrants in the United States, as reflected in the abandonment of the parental language and other ethnic patterns of behavior, the second generation remained *more* conscious of their ethnic identity than were their immigrant parents (Nahirny and Fishman, 1965). The parents' ethnic identity was so much taken for granted that they were scarcely explicitly aware of it, but the marginality of their children made them acutely self-conscious and sensitive to their ethnicity, especially when passing through adolescence. Moreover, as parents and children acculturated at different rates, a generational gap grew so that by the time the children reached adolescence 'the immigrant family had become transformed into two linguistic subgroups segregated along generational lines.' Finally, by the third generation 'the grandsons became literally outsiders to their ancestral heritage,' and their ethnic past an object of symbolic curiosity.

Contemporary Realities: Prospects and Paradoxes

By the late-twentieth century, a new era of mass immigration, now overwhelmingly non-European in composition, had again raised familiar questions about the 'assimilability' of the newcomers and their children, and concerns that many of them might become consigned to vast multiethnic formations on the other side of new color lines. Gans (1992a) questioned the American myth of nearly automatic immigrant success and delineated six theoretical 'scenarios' for the incorporation of the new second generation as they were beginning to enter the workforce, hinging on economic and other conditions. Three were positive scenarios, positing upward mobility driven by educational attainment, ethnic succession, or niche improvement. Three posited negative futures, projecting the reverse of the previous three (educational failure, the stalling of ethnic succession in the legal economy, niche shrinkage) – a 'second-generation decline' potentially exacerbated by a combination of economic downturns or nonlabor-intensive economic growth, the second generation's refusal or inability to accept the jobs their parents held, and competition from successive new waves of immigrants. Rather than experiencing upward mobility, the second generation (or segments of it, especially the children of undocumented immigrants) would join the ranks of urban poor.

The new realities also raised questions about the applicability of explanatory models developed in connection with the experience of European ethnics, despite the fact that contemporary immigrants were being incorporated in a post-Civil Rights context – if also officially categorized by new and pervasive pan-ethnic labels – characterized more by ethnic revivals and identity politics than forced Americanization campaigns. While assimilation (as indexed by acculturation, socioeconomic mobility, residential integration, naturalized citizenship) may still represent a 'master trend' for many of today's immigrants, as Alba and Nee (2003) have argued, it is subject to too many contingencies and affected by too many variables to render the notion of a relatively uniform and straightforward path convincing (aside from the swift switch to English among immigrants' children, which cuts across all classes and nationalities).

Instead, as Portes and Zhou (1993) framed it a year after Gans' six scenarios, the present second generation of children of immigrants can be seen as undergoing a process of 'segmented assimilation' where outcomes vary across immigrant minorities, and where rapid integration and acceptance into the American mainstream represent just one possible alternative. Why this is so hinges on a number of factors: internal characteristics, including the immigrants' level of human capital and the structure and cohesiveness of their families, interact in complex but patterned ways with external contexts of reception – government policies and programs, the economy in areas of settlement, employer preferences in local labor markets, the extent of racial discrimination and nativist hostility, the strength of existing ethnic communities – to form the conditions within which immigrants and their children adapt to different sectors of American society.

'Segmented assimilation' processes – i.e., adaptations that take place within varying opportunity structures and are shaped through differential associations, reference groups, experiences, and attachments, especially in primary social relationships stratified by race, religion, region, and class – are not new in the American experience (Rumbaut, 1997). Caste restrictions based on race (extending to all aspects of social life, including citizenship), the 'triple melting pot' of religion-bounded intermarriages, the structural pluralism of 'ethclasses,' the persistence of ethnic groups as political interest groups, are all indicative of such adaptations. Alba and Nee (2003), thinking about the divergent outcomes that will likely obtain in the first decades of the twenty-first century, concede the point and put it well:

> The contemporary immigration scene displays complex, contradictory patterns, from rapid assimilation apparent among some professionals and their children to the new way of sojourning apparent in some transnational circuits, and to the potential among other immigrant groups for incorporation as racialized minorities … Clearly, assimilation will not apply to all immigrant minorities to the same extent … many in the second generation are likely to experience upward mobility into the American socioeconomic mainstream … [others] may experience lateral or, at best, short-distance mobility … Children of low-wage labor migration are likelier to experience downward mobility into the urban minority underclass than children of human capital migration from the same ethnic group … There is no reason to believe that assimilation is inevitable or that it will be the master trend for all these diverse groups (pp. 50, 273–275).

Immigration and American Pluralism

Despite the grand narratives of modernization, which undergird the concept of assimilation, neither race nor religion nor ethnicity has vanished in American life. Protestants (never a homogeneous category, composed of dozens of disparate denominations and of fundamentalist, evangelical, and apostolic varieties) are actually a 'vanishing majority,' having fallen below 50% for the first time in 2005 – a decline that is sharper still among younger people, in more recent years, and among the first and second generations. The U.S. Supreme Court is now made up of six Catholics and three Jews – not a Protestant among them. But religious pluralism remains alive and well. Catholics have remained a fourth of the population for decades now, a secular decline over time having been more than compensated by new influxes of immigrant Catholics from Latin America, the Philippines, and Vietnam. And there has been an increase in non-Christian religions, notably Buddhists, Hindus, and Muslims; aside from the long-present Jewish population, all other non-Christian religions had increased to about 5.5% of the population by 2000. The United States remains today a profoundly religious country – not least because the American state allowed immigrant groups to develop their own social and cultural institutions, including parishes, schools, hospitals, temples, and synagogues – a *laissez faire* stance now ironically attributed to the American ability to 'assimilate' them (Portes and Rumbaut, 2014).

The fate of immigrant languages other than English is another matter: linguistic 'Anglicization' and other forms of acculturation do proceed rapidly, especially among immigrant children and the second generation, and that may be truer now

than ever before, despite the unprecedented diversity of class, culture, and color in the present era of mass immigration. The United States has been aptly described as a 'graveyard' for languages because of its historical ability to extinguish the mother tongues of immigrants within two or three generations (Portes and Rumbaut, 2014). As documented consistently and compellingly by a variety of sources, including census and survey data, not even Spanish appears to challenge this dubious reputation.

But alongside undeniable upward social mobility from the first to the second generation for most groups, especially the children of the poorest and least educated – though the gains appear to peak in the second generation and decline or plateau thereafter (Telles and Ortiz, 2008) – there is compelling evidence of widening 'ethclass' and legal inequalities, of new conflicts and political mobilizations around ethnic and racial issues, and of downward mobility and marginalization for vulnerable segments of these populations. An undocumented status has become a castelike master status blocking access to the opportunity structure and paths to social mobility for millions of immigrants. A fraught concept like 'assimilation,' weighted by the normative baggage of its past and by its insistent if inclusive expectation of progress and homogenized national cohesion, seems ill-suited to grasp these complex dynamics and to focus critical attention on enduring structural inequalities and persistent ethnic and pan-ethnic formations in a 'permanently unfinished' society.

See also: Generations, Sociology of; Inequality, Social; Integration, Social; Minorities; Minorities; Prejudice in Society: Sociological Perspectives; Race Identity; Racial Relations; Social Identity in Sociology.

Bibliography

Alba, R.D., 1990. Ethnic Identity: The Transformation of White America. Yale University Press, New Haven.

Alba, R.D., Nee, V., 2003. Remaking the American Mainstream: Assimilation and Contemporary Immigration. Harvard University Press, Cambridge, MA.

Gans, H.J., 1979. Symbolic ethnicity: the future of ethnic groups and cultures in america. Ethnic and racial Studies 2 (1), 1–20.

Gans, H.J., 1992a. Second generation decline: scenarios for the economic and ethnic futures of Post-1965 american immigrants. Ethnic and Racial Studies 15 (2), 173–192.

Gans, H.J., 1992b. Ethnic invention and acculturation: a bumpy-line approach. Journal of American Ethnic History 12, 45–52.

Gans, H.J., 1997. Toward a reconciliation of 'assimilation' and 'pluralism:' the interplay of acculturation and ethnic retention. International Migration Review 31 (4), 875–892.

Glazer, N., Moynihan, D.P., 1963. Beyond the Melting Pot: The Negroes, Puerto Ricans, Jews, Italians, and Irish of New York City. The MIT Press, Cambridge, MA.

Gordon, M.M., 1964. Assimilation in American Life: The Role of Race, Religion, and National Origins. Oxford University Press, New York.

Higham, J., 2002[1955]. Strangers in the Land: Patterns of American Nativism, 1860–1925. Rutgers University Press, Brunswick, NJ.

Kanstroom, D., 2007. Deportation Nation: Outsiders in American History. Harvard University Press, Cambridge, MA.

Kazal, R.A., 1995. Revisiting assimilation: the rise, fall, and reappraisal of a concept in american ethnic history. American Historical Review 100 (2), 437–471.

Kennedy, R.J.R., 1944. Single or triple melting-pot? Intermarriage trends in New Haven, 1870–1940. American Journal of Sociology 49, 331–339.

Massey, D.S., Sánchez, M.R., 2010. Brokered Boundaries: Creating Immigrant Identity in Anti-immigrant Times. Russell Sage Foundation, New York.

Miller, S.C., 1982. Benevolent Assimilation: The American Conquest of the Philippines, 1899–1903. Yale University Press, New Haven, CT.

Nahirny, V.C., Fishman, J.A., 1965. American immigrant groups: ethnic identification and the problem of generations. Sociological review. NS 13, 311–326.

Ngai, M.M., 2005. Impossible Subjects: Illegal Aliens and the Making of Modern America. Princeton University Press, Princeton.

Park, R.E., 1914. Racial assimilation in secondary groups with particular reference to the negro. American Journal of Sociology 19 (5), 606–623.

Park, R.E., May 1926. Our racial frontier in the Pacific. Survey Graphic IX, 192–196 [Reprinted in Park's collected papers, 1950. Race and Culture, vol. 1. The Free Press, Glencoe, pp. 138–151.].

Park, R.E., 1930. Assimilation, social. In: Seligman, E.R.A., Johnson, A. (Eds.), Encyclopedia of the Social Sciences, vol. 2. Macmillan, New York, pp. 281–283.

Park, R.E., 1937. Introduction. In: Adams, R. (Ed.), Interracial Marriage in Hawaii. Macmillan, New York.

Park, R.E., Burgess, E.W., 1924[1921]. Introduction to the Science of Sociology. University of Chicago Press, Chicago.

Portes, A., Rumbaut, R.G., 2001. Legacies: The Story of the Immigrant Second Generation. University of California Press and Russell Sage Foundation, Berkeley and New York.

Portes, A., Rumbaut, R.G., 2014. Immigrant America: A Portrait, fourth ed. University of California Press, Berkeley.

Portes, A., Zhou, M., 1993. The new second generation: segmented assimilation and its variants among Post-1965 immigrant youth. The Annals, AAPSS 530, 74–98.

Redfield, R., Linton, R., Herskovits, M.J., 1936. Memorandum for the study of acculturation. American Anthropologist 38 (1), 149–152.

Rumbaut, R.G., 1997. Assimilation and its discontents: between rhetoric and reality. International Migration Review 31 (4), 923–960.

Telles, E.E., Ortiz, V., 2008. Generations of Exclusion: Mexican Americans, Assimilation, and Race. Russell Sage Foundation, New York.

Warner, W.L., Srole, L., 1945. The Social Systems of American Ethnic Groups. Yale University Press, New Haven, CT.

Waters, M.C., 1990. Ethnic Options: Choosing Identities in America. University of California Press, Berkeley.

Yinger, J.M., 1981. Toward a theory of assimilation and dissimilation. Ethnic and Racial Studies 4 (3), 249–264.

Yinger, J.M., 1994. Ethnicity: Source of Strength? Source of Conflict? State University of New York Press, Albany.

Associations, Scholarly

Laurent Jeanpierre, Universite Paris 8, Paris, France
Thibaud Boncourt, Centre Européen de Sociologie et de Science Politique (CESSP), Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, using some of the tools provided by the history and the sociology of science, we assess how scientific societies and scholarly associations emerged and played a role in the development of social and behavioral sciences and in their specialization. They institutionalized in the nineteenth century and became both professionalized and internationalized in the twentieth century. Their number has rapidly grown in most countries. However, even as they multiplied, scientific societies in the social sciences became less important for the scientific enterprise. We analyze this process of internationalization and the role some professional scholarly associations are likely to play in the future as civic actors at national and international levels.

Scholarly associations are one among the various types of social organizations for scientific activity alongside universities, institutes, research laboratories, or learned societies. They are private or government sponsored organizations whose individual or institutional members pursue only one particular or a few related branches of learning. They originally played a key role in the emergence of science and have represented an important factor for the birth and growth of social and behavioral sciences (SBS) since the beginning of the nineteenth century. In particular, scholarly associations have been central to the specialization and discipline wise organization of scientific activities. Yet for more than a century now, they have become more often places of communication and exchanges than sites for producing knowledge. Despite their acknowledged role in the development of the SBS, little research has been done on scholarly associations in the SBS. Studies of science have focused mostly on universities, academies, and learned societies. Scholarly associations have been dealt with sometimes for their role in the history of the natural sciences. Studies dealing with scholarly associations in the SBS are very rare. A majority of the few existing ones focus on their internationalization.

According to the *World Guide to Scientific Associations and Learned Societies* (2006) there are at least 17 500 such associations. Even as this figure includes some research laboratories, it probably underestimates the number of scholarly associations since it depends on self-declarations. In contrast to most professional associations, academies are exclusive organizations, the membership of which is generally restricted to elected fellows. A majority of contemporary scientific associations in any field of science are local, national, and regional and are usually devoted to a specific discipline (Kiger, 1993). There are also a lot of international disparities in the distribution of scholarly associations: the western countries have a lot more of them than countries of the developing world which sometimes have fewer than five science associations (Union of International Associations, (1989), quoted in Crawford, 2003). In the peripheral countries, national scientific infrastructure may be very weak. In that case international scholarly associations are likely to be the main link of scientists to the activities of their field. All the disciplines of the SBS and some of their subdisciplines and cross-disciplines developed their professional associations at both national and international levels.

Early History

While specialized scholarly associations started to develop from the nineteenth century onwards, associations of scholars existed much earlier. One could trace their existence to Confucian China or to Ancient Greece. But this type of intellectual groups started to choose scientific cooperation as a norm much later in history (Zilsel, 2000[1945]). The ideal of cooperation and progress of knowledge through cooperation did not exist in western and nonwestern philosophical traditions before the beginning of the sixteenth century. Conflict was more important than cooperation both in ancient philosophical schools and in medieval universities. According to Edgar Zilsel, "the Middle Ages produced one co-operative scholarly organization. Soon after his accession to the throne in 1252, King Alphonso X of Castile and Leon appointed a commission of Arabic, Jewish and Christian scholars to translate astronomical and astrological works from Arabic to Spanish." They produced the *Libros del Saber* (Zilsel, 2000 [1945]).

In humanist academies, which started to develop in Italy after 1500 (around 700 of them arose in the sixteenth century alone), individualistic ideals were still dominant among members. Moreover, most of these numerous academies were not oriented toward research: they were debating clubs and literary societies for nonscholars, for aristocrats with a high level of education and a variety of interests. Collective work started at least in two places at that period: in the Academia Aldina of Venice founded in 1495 by the printer Aldo Manunzio and in the Academia della Crusca founded in 1582 in Florence. In the former, members were not aristocrats but classical scholars. In the latter, members worked on the first dictionary of Italian language published in 1612, which is thus considered by Zilsel as "the first scientific publication composed by several authors in collaboration … after the astronomical encyclopaedia of the Alphonsine Commission." But the idea of progress through collective cooperation in

research was still absent from this academy. On the other hand, the Accademia del Cimento created by pupils of Galileo in Florence in 1657 was presumably the first association where scientific experimentations were the main goals (Maylender, 1926–1930).

A few years later the Royal Society of London and the Paris Académie des Sciences were created (Ornstein-Bronfenbrenner, 1928). There the Baconian idea of scientific progress was integrated. In London, the small group of original founders expanded to 200 fellows in 1666 to raise more funds. In Paris, the early Academy consisted of two dozens of paid scientists who met privately twice a week. Networks of correspondents sharing the same interests in natural philosophy and science also developed all over Europe in the second half of the seventeenth century. All these new institutions were created outside universities, where in most cases, scholarly activities involved neither any experiments nor any renewal of knowledge: the higher faculties of the medieval and early modern universities only delivered degrees for the liberal professions (theology, law, and medicine). In the case of the Royal Society and the Paris Academy, their relative autonomy toward the patron and monarch was higher than in Renaissance Italian academies. With these associations of aristocrats and scholars, the scientific revolution had its deepest roots in an organizational revolution of intellectual activities. In addition to the Royal Society and the Paris Academy, many other scientific societies were founded in the second half of the seventeenth century, many of which were modeled after one of their two predecessors. In 1669/1670 Leibniz wrote for example a plan for a German Academy that was created in Berlin in 1700. Other academies were later established in Göttingen or in Bavaria.

Modern History

National and local scientific academies spread over Europe as well as in Russia and in the United States during the eighteenth century (McClellan III, 1985). After the French Revolution, the Directoire established in 1795 the Institut National des Sciences et Arts in order to replace the Académie of Sciences. It was divided into three classes of knowledge, one of which would be related today to the SBS and was being called at the time 'Sciences Morales et Politiques.' This class itself was subdivided into six sections: Analysis of Sentiments and Ideas, Ethics, Social Science and Legislation, Political Economy, History, and Geography (Machlup, 1982: p. 92). The French Academy was one of the first scientific institutions where the SBS were recognized as a class of knowledge and where scholars would gather to debate on social scientific topics. A scientific society devoted to anthropology, the *Société des observateurs de l'homme* was also created in Paris a few years later but lasted only for 5 years. In 1803, Napoléon abolished the class for Moral and Political Sciences in the Institute. Almost 30 years later, a national academy again devoted to this class of knowledge was created under the new regime of Louis Philippe of Orléans. It had five sections: Philosophy; Ethics; Legislation, Public Law and Jurisprudence; Economics, Statistics and Finance; and History and Geography (Machlup, 1982: p. 93). There was no comparable move in England until 1902 when the British Academy for the Promotion of Historical, Philosophical and Philological Studies was created because the Royal Society took no interest in the new human and social sciences. In Germany, the scientific academies had devoted some of their sections to philosophy and history since the mid-eighteenth century. In Sweden, the Royal Academy was founded in 1741 and in 1799, one of its 12 classes was devoted to 'economic, statistical, and social sciences.' In most cases, the various disciplines of the SBS except law, history, and geography were fully recognized in public academies at the beginning of the twentieth century, starting with psychology and anthropology.

Scientific societies and scholarly associations outside academies had proliferated in the nineteenth century. Academies remained multidisciplinary and developed in provincial cities. On the contrary, scientific societies became more and more specialized, like the Geological Society of London founded in 1807. They progressively defined norms for realization, presentation, belonging, and qualification among scholars sharing the same interests. Science became, through scholarly associations, a professional activity and a career. The conception of a professionally qualified person replaced the idea that science could only be pursued by a minority with a vocation and money to sustain it. When the British scholar William Whewell coined the word 'scientist' in 1833, he captured the transformation in the definition of the 'man of science' that was taking place at the time. Scientific societies became more elitist both in the physical and in natural sciences and in the SBS. As a result, learned societies, which were more open and public, also developed in the nineteenth century. According to Fox, there were, for example, 665 of them in France in 1886 and a French learned society had a membership of 200 000 on the eve of World War I. Archeology and local history were among the most common interests of these societies (Fox, 1980).

Scientific societies and scholarly associations of the nineteenth century became a central element in the building of a new social organization of higher learning and of science where teaching, research, and the public involvement of scientists were integrated in disciplines. According to Heilbron, "French developments in the natural sciences formed the centre during the first phase of the development of disciplines. If the pioneering role of the French setting is generally overlooked, this is mainly because historical analyses commonly focus on universities while ignoring the role of institutions like academies, learned societies, and professional schools that did not have a university status" (Heilbron, 2004). This was a long-term process, which also involved the founding of chairs in universities, the creation of disciplinary journals, and subsequently creation of autonomous degrees (Daston, 1998; Heilbron, 2004; Kelley, 1997). Nevertheless, from the end of the nineteenth century onwards, scientific societies tended to be less and less implied in learning and research and more in the diffusion of knowledge among scientists and in the public. They were eclipsed by universities and corporate and government laboratories. While scientific societies and academies were devoted to research and universities to teaching and diffusion of science, the roles of these two types of institutions gradually interchanged during the second half of the nineteenth century.

Various interpretations have been made of the emergence of scientific societies and scholarly associations and of their multiplication, specialization, and professionalization. Zilsel has emphasized the technological and social origins of the ideas of cooperation and progress in the first associations of scholars during the sixteenth and seventeenth centuries (Zilsel, 2000 [1945]). He showed that craftsmen on the one hand and cartography, navigation, and applied mathematics on the other hand were decisive for the development of some of the earliest scientific academies and for the invention of the scientific role. The growth in the number of academic societies and in their interest for science has also been related to technology by other scholars (Ornstein-Bronfenbrenner, 1928). Analyzing the data gathered by Maylender on hundreds of European academies created after 1500, Ben David has shown that in the sixteenth and seventeenth centuries the recognition of an autonomous scientific enterprise was more likely in Northern Europe than in Italy (Ben-David, 1971). He thus relates the early developments of scientific societies with Protestantism. Stichweh tried to explain the organization of the modern scientific system and the development of science and scholarly associations in the eighteenth and nineteenth centuries by studying German physical sciences between 1740 and 1890. According to him encyclopedic and humanistic ideals have been replaced by scientific disciplines over this period. The multiplication of professional scientific associations and learned societies in the beginning of the nineteenth century is the first sign of a transformation that can be interpreted as an internal differentiation of knowledge following the social rules of a growing division of labor (Stichweh, 1984). Yet, for other scholars like Heilbron "the world of higher learning already had a fairly differentiated structure well before the second half of the eighteenth century…Science and scholarship in Europe was institutionally more heterogeneous before 1800 than after" (Heilbron, 2004). Even though there is no consensus on the factors that provoked the emergence, the multiplication, and the growing specialization of science associations in the western modern world, most of the scholars agree that they were central in the building of a 'regime of disciplines' both in physical and natural sciences and in SBS (Heilbron et al., 1998).

Internationalization and Current Developments

Development of scholarly associations did not stop at national boundaries. First, trans-boundary science existed before the nation-state: from as early as the late seventeenth century, 'international' forms of scientific gatherings and networks emerged. More modern forms of these gatherings took the shape of a few punctual congresses revolving mostly around the discipline of physics. In the nineteenth century, international congresses developed in the SBS including statistics (1853), prehistory (1866), geography (1871), psychology (1889), and history (1900). Universal exhibitions also played a role in the internationalization of scientific traditions and national associations. After 1850, the number of such congresses started to grow at an exponential pace as these gatherings became more diversified at the disciplinary level, with both natural (e.g., physicists and biologists) and social scientists (e.g., demographers and geographers) organizing such events. Several disciplines, such as statistics, set out to hold congresses on a regular basis, thereby paving the way for the creation of proper international scholarly associations.

Several international scientific organizations were thus created between the mid-nineteenth and the late twentieth century so that, in 1994, 'there were approximately 350 active international science associations' (Schofer, 1999). Two phases may be identified in their development. The late nineteenth century and early twentieth century was a period of slow growth as it saw the creation of only about 30% of these associations. In the second half of the twentieth century, international science associations grew in number at an accelerated pace as over 70% of them were founded during that period (Schofer, 1999). The revival and renewed expansion of international social science associations occurred after the Second World War, when the United Nations Education, Science and Culture Organization (UNESCO) initiated and funded international disciplinary associations: the International Union of Anthropological and Ethnographical Sciences (1948), the International Sociological Association (1949), the International Political Science Association (1949), the International Economic Association (1949), the International Association of Legal Science (1950), and the International Union of Psychological Science (1951). These international associations vary in their geographical scope and purpose. At the geographical level, they may seek to gather scientists from all over the world or from a specific continent or region. In terms of their purpose, they may be either professional organizations (that seek to promote specific professional interests and scientific knowledge) or socially oriented associations (that seek to support science to "address social problems such as economic development, environmental degradation, war, nuclear weapons, ethics, and so on"). Socially oriented organizations were relatively rare before 1945 (Schofer, 1999).

Several explanations have been put forward to explain the exponential rise in the number of international scholarly associations in the twentieth century and particularly after 1945. The first of these is related to academic development. In this perspective, the growth in the number of organizations is seen as a consequence of the rise in the number of scholars that came with the massification of higher education and the creation of new universities. Correlatively, it is also linked to the increasing fragmentation of scientific knowledge and the diversification of disciplines. The creation of specific scholarly associations is thus a step toward the increasing autonomization and institutionalization of sciences (Clark, 1972).

While this is a key explanation, other factors have also been put forward by the literature. Several authors have highlighted the influence of political actors and contexts on the development of scholarly associations. As eighteenth and nineteenth centuries saw the "emergence of the nation as the structuring unit and the principal arena of scientific activity" (Crawford et al., 1993), state building processes were seen as powerful incentives for the national development of sciences. Sciences contributed to creating a shared national culture and assumed a regulatory function by getting involved, for example, in the standardization of weights and measures. The common interests of scholars and governments officials often resulted in the foundation of new scholarly organizations, as in the case of Germany where "in 1887, the joint interests of physicists,

government officials and industrial leaders in precision measurement and technology led to the creation of the Imperial Institute of Physics and Technology" (Crawford, 1992).

Correlatively, the 'institutionalization of the liberal world polity' is identified as a key factor for the rise in the number of international scholarly associations, notably after World War II (Schofer, 1999). Several studies have, for example, documented the role played by UNESCO in the creation and funding of such organizations, particularly in the social sciences (e.g., Platt, 2002). More recently, the European Union also funded the foundation of regional scientific organizations with a view to strengthening the project of European Research Area (ERA). Some nongovernmental organizations also played a key role in sponsoring the creation of scholarly associations. During the interwar years and the Cold War, philanthropic organizations such as the Ford and Rockefeller foundations indeed provided sizable grants for the founding of several scientific endeavors in all kinds of disciplines ranging from molecular biology (Kay, 1993) to public administration.

On the other hand, other political factors may hamper the creation of scholarly associations. Periods of wars and economic depression tend to decrease the founding rate of professional science organizations (Schofer, 1999). The nature of political regimes may also be more, or less favorable to the development of scientific disciplines and the creation of scholarly associations. Authoritarian regimes have thus been shown to deter the foundation of such bodies (e.g., Blum and Mespoulet, 2003), contrary to liberal democracies.

Becoming Civic Actors?

Some scholars relate the growing internationalization of scientific associations with a change in their social role. Crawford writes, "(we are currently living a) major shift from science associations organized as an inward-looking profession to science associations organized around social problems and societal concerns, often sustaining advisory positions (or other authoritative roles) in relation to the United Nations or particular national governments," as in the case of the International Organization for Chemical Sciences in Development or the International Network of Engineers and Scientists for Global Responsibility (Crawford, 2003: pp. 81–82). These socially oriented scholarly associations provide relevant information on scientific activities to both policy makers and public. They promote science or SBS as means to solve some social issues. They also develop appropriate ethical frameworks for the practice and utility of science. According to Crawford, the "rise of professional science INGOs increases the founding rate of socially oriented science INGOs" (Crawford, 2003: p. 92). After a long phase of specialization, of professionalization, and of internationalization of scientific societies and scholarly associations in the natural and social sciences we may now witness a political turn in their role.

See also: Social Science Infrastructure: European Union (Social Science Academies and Related Organizations); Social Science Learned Societies and Professional Associations; Universities and Science and Technology: Europe; Voluntary Associations, Sociology of.

Bibliography

Ben-David, J., 1971. The Scientist's Role in Society. Prentice-Hall, Englewood Cliffs, NJ.

Blum, A., Mespoulet, M., 2003. L'anarchie bureaucratique: pouvoir et statistique sous Staline. La Découverte, Paris.

Clark, T.N., 1972. Les étapes de l'institutionnalisation scientifique. Revue internationale des sciences sociales XXIV (4), 699–714.

Crawford, E., 1992. Nationalism and Internationalism in Science, 1880–1939. Cambridge University Press, Cambridge.

Crawford, E., Shinn, T., Sörlin, S. (Eds.), 1993. Denationalizing Science. The Contexts of International Scientific Practice. Kluwer Academic Publishers, Dordrecht/Boston.

Crawford, E., 2003. International science associations, 1870–1990. In: Drori, Gili S., Meyer, John W., Ramirez, Francisco O., Schofer, Evan (Eds.), Science in the Modern World Polity. Institutionalizaton and Globalization. Stanford University Press, Stanford, pp. 81–99.

Daston, L., 1998. The academies and the unity of knowledge: the disciplining of the disciplines. Differences: A Journal of Feminist Cultural Studies 10 (2), 67–86.

Fox, R., 1980. The Savant confronts his peers: scientific societies in France, 1815–1914. In: The Organization of Science and Technology in France, pp. 241–282. Cambridge and Paris.

Heilbron, J., Magnusson, L., Wittrock, B. (Eds.), 1998. The Rise of the Social Sciences and the Formation of Modernity: Conceptual Change in Context, 1750–1850. Kluwer Academic Publishers, Dordrecht/Boston.

Heilbron, J., 2004. A regime of disciplines: toward a historical sociology of disciplinary knowledge. In: Camic, Charles, Joas, Hans (Eds.), The Dialogical Turn. New Roles for Sociology in the Postdisciplinary Age. Rowman & Littlefield, Lanham, pp. 23–42.

Kay, L.E., 1993. The Molecular Vision of Life: Caltech, the Rockefeller Foundation, and the Rise of the New Biology. Oxford University Press, New York/Oxford.

Kelley, Donald R. (Ed.), 1997. History and the Disciplines: The Reclassification of Knowledge in Early Modern Europe. Rochester University Press, Rochester.

Kiger, J.C. (Ed.), 1993. International Encyclopaedia of Learned Societies and Academies. Greenwood Press, Westport (CT)/London.

Machlup, F., 1982. The Branches of Learning, Volume 2 of Knowledge: Its Creation, Distribution, and Economic Significance. Princeton University Press, Princeton.

Maylender, M., 1926–1930. Storia delle Accademie d'Italia. L. Capelli, Bologna.

McClellan III, J.E., 1985. Science Reorganized: Scientific Societies in the Eighteenth Century. Columbia Unibersity Press, New York.

Ornstein-Bronfenbrenner, M., 1928. The Role of Scientific Societies in the Seventeenth Century. Chicago University Press, Chicago.

Platt, J., 1998. A Brief History of the International Sociological Association 1948–1997. International Sociological Association, Montreal.

Platt, J., 2002. Fifty Years of the International Social Science Council. International Social Science Council, Paris.

Schofer, E., 1999. Science associations in the international sphere 1875–1990. The rationalization of science and the scientization of society. In: Boli, J., Thomas, G. (Eds.), Constructing World Culture: International Non-Governmental Organization since 1875. Stanford University Press, Stanford, pp. 249–266.

Stichweh, R., 1984. Zur Entstehung des modernen Systems wissenschaftlicher Disziplinen: Physik in Deutschland 1740–1890. Suhrkamp, Frankfurt am Main.

Union of International Associations, 1989. Yearbook of International Organizations. K.G. Saur, München.

World Guide to Scientific Associations and Learned Societies, tenth ed., 2006. K.G. Saur, München.

Zilsel, E., 2000[1945]. The genesis of the concept of scientific progress and scientific cooperation. In: The Social Origins of Modern Science. Kluwer Academic Publishers, Dordrecht/Boston, pp. 128–170.

Associative Learning in the Cerebellum

Peter Mariën, ZNA Middelheim Hospital, Antwerp, Belgium; and Vrije Universiteit Brussel, Brussels, Belgium
Frank Van Overwalle and Marie Vandekerckhove, Vrije Universiteit Brussel, Brussels, Belgium
Mario Manto, Fonds de la Recherche Scientifique (FNRS), ULB Erasme, Bruxelles, Belgium

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.E. Steinmetz, volume 3, pp. 1627–1631, © 2001, Elsevier Ltd.

Abstract

Associative learning denotes a specific kind of learning process, in which a change in the behavior of an organism results from adaptation to a particular relation between events occurring in an environment. Current insights in the neurobiological mechanisms subserving associative learning are primarily based on experimental and clinical research with the classical (eyeblink) conditioning paradigms. In the mammalian species, including humans, the cerebellum serves as the cardinal neural substrate for different forms of associate learning including motor, emotional, and cognitive associative learning. This article aims to present a concise overview of the critical involvement of the human cerebellum in associative learning processes in the motor, cognitive, and affective domain.

Introduction

During the past decades, the classical tenet, which posits that the cerebellum, 'the little brain,' solely functions as a coordinator of motor skills has been substantially modified. A large number of neuroanatomical, neurophysiological, neuroimaging, and clinical studies in humans and animals have amassed compelling evidence that fundamentally extends the motor role of the cerebellum to the modulation of a variety of cognitive, behavioral, and affective processes (Baillieux et al., 2008). Although from time to time clinical case descriptions and experimental evidence from animal studies dating back to the early part of the nineteenth century already suggested an association between cerebellar pathology and a variety of nonmotor cognitive and affective dysfunctions, a causal relationship remained unexplored and was dismissed for several decades. However, as exemplified by the work of Snider, Dow, Heath, Cooper, and others (see Schmahmann, 2010 for a review), investigators started in the mid-1900s to examine a possible causal link between the cerebellum and cognitive and affective processes. This line of research laid a robust foundation for the rediscovery of the concept by Leiner et al. (e.g., 1986), who hypothesized that more recently evolved parts of the cerebellum contribute to learning, cognition, and language, and by Schmahmann et al. (see Schmahmann, 2010 for a review), who introduced the influential dysmetria of thought hypothesis providing an historical, clinical, neuroanatomical, and theoretical framework within which a cerebellar role in higher cognitive and affective processes could be considered. In this theory, the cerebellum is viewed as a modulator of behavior function, maintaining it around a homeostatic baseline appropriate to the context. In the same way, the cerebellum regulates motor function (rate, force, rhythm, and accuracy of movements), the speed, capacity, consistency, and appropriateness of affective and cognitive processes. Dysmetria of movement, inherently bound to ataxia of movement, is then matched by dysmetria of thought, an unpredictability and illogic to social and societal interaction. In the late 1990s, Schmahmann and Sherman (1998) introduced in a much cited paper the concept of cerebellar cognitive affective syndrome to identify a range of cognitive and affective disturbances in patients with isolated cerebellar lesions. Based upon bedside screening and formal neuropsychological testing, Schmahmann and Sherman (1998) identified a range of cognitive, linguistic, and affective symptoms following focal cerebellar damage. The core features of this syndrome consist of: (1) executive dysfunctions such as disturbances in planning, set-shifting, abstract reasoning, and working memory (WM), (2) visuospatial deficits, such as impaired visuospatial organization and memory, (3) mild language symptoms including agrammatism and anomia, and (4) behavioral-affective disturbances, consisting of blunting of affect or disinhibited and inappropriate behavior.

Associative learning in its different forms is a good example of the conceptual change of thinking about the role of the cerebellum in motor, cognitive, and affective functioning. Associative learning represents a basic type of learning skill, in which a change in the motor, cognitive, or affective behavior of an organism results from adaptation to a particular relation between events occurring in a particular environment. The paradigm with the longest tradition that attracted ample attention of the scientific community in the research on associative learning is the classical conditioning of the eyeblink response. It was first introduced in a prototypical form in humans by Ernest Hilgard in the 1930s and later popularized during the 1960s in animals as the nictitating membrane/eyelid response by Isadore Gormezano. In this Pavlovian, associative motor learning paradigm an aversive, unconditioned stimulus (US) such as a periorbital shock or corneal air puff is given after the presentation of a behaviorally neutral conditioned stimulus (CS) such as a tone or light. At the onset, the US elicits a reflexive, unconditioned vigorous eyeblink response (UR) while the neutral CS results in no overt motor reaction. In association with repeated, paired stimulus presentations (50–100 pairings), a gradual development of a learned, conditioned eyeblink response (CR) to the CS takes

place consisting of a precisely timed closure of the eye before the time the aversive CS is given.

The most extensive body of evidence on classical eyeblink conditioning has been derived from experimental animal studies (rabbits) but it appears from many cerebellar lesion and functional brain imaging studies that all mammals, including humans, learn this motor association task at similar rates by means of a same neural circuitry, in which the cerebellum plays a cardinal role. Many animal studies have been conducted on the basis of a variety of experimental techniques such as electrophysiological recording of single and multiple units, electrolytic and chemically induced lesions, chemical and physical reversible lesions, neural stimulation, genetic mutations, and pharmacological manipulation. Notwithstanding the fact that an extensive number of neural regions are involved in various aspects of eyelid responses, the primary focus of interest of the majority of research studies and reviews on eyeblink conditioning has been directed to the cerebellum as the primary structure of interest in the acquisition and storage of this type of associative learning. Thompson et al. demonstrated in an important series of neurobiological studies that the dorsolateral interpositus nucleus of the cerebellum ipsilateral to the conditioned eye is the cardinal locus in animals for acquisition and retention of newly acquired eyelid responses, while involvement of the cerebellar cortex during normal acquisition does not seem to be essential (for a review see Woodruff-Pak and Steinmetz, 2000; Thompson, 2005; Lindquist et al., 2013). Studies applying more complicated forms of eyeblink conditioning, such as trace conditioning, which makes use of a stimulus-free time interval between the CS and US, show that in addition to the deep nuclei of the cerebellum-conditioned learning recruits involvement of the hippocampal system and the prefrontal lobe (Delgado-Garcia and Gruart, 2006).

Motor Associative Learning

Eyeblink Conditioning in Animals

There exists a wealth of excellent overviews and surveys, to which the reader can be referred, in which the large amount of animal study data relating to eyeblink conditioning is dealt with in much detail (e.g., Thompson, 2005; Thompson and Steinmetz, 2009). In this survey, a summary of some of the main findings is presented.

First evidence for the involvement of the cerebellum in classical eyeblink conditioning was derived from early experimental lesion studies in animals. Studies showed that decorticated and decerebrated animals, which were trained in the delay eyeblink CR (the offset of the CS overlaps with the US, or is immediately followed by US) before incurring the lesion, retain the eyeblink CR as long as the cerebellum is not structurally affected. Richard Thompson et al. in the 1980s showed that permanent disruption of CRs in trained animals results from damage to the interpositus nucleus of the cerebellum. In addition, permanent abolishment of the skill to acquire and retain the eyeblink CR was also found in association with lesions involving the lateral pontine nuclei, the inferior olive, the red nucleus, and the inferior, middle, and superior cerebellar pedunculi. By contrast, experimental animal studies with lesions of the cerebellar cortex did not produce a similar consistent pattern of eyeblink CR impairment. Ablations limited to lobule VI (located directly above the anterior interpositus nucleus) massively impaired but did not abolish the delay eyeblink CR (Figure 1). Extensive lesions of the intermediate and lateral zones of the anterior and posterior lobe of the cerebellum were associated with a permanent loss of the skill to acquire and retain the delay eyeblink CR. These lesion studies indicate that structural integrity of the olivo-cortico-nuclear loop is a prerequisite for normal acquisition of CR and that the cerebellum may be critically involved in neural plasticity subserving classical eyelid conditioning.

In addition to structural lesion studies, reversible inactivation of brain function together with electrophysiological recordings of neural activity have been frequently applied as efficient research methods to localize memory traces in the brain. Reversible inactivation by means of pharmacological agents (e.g., lidocaine, tetrodotoxin) or cooling has often been used to produce temporary lesions to the major structures and pathways considered to constitute the cerebellar-brainstem circuitry implicated in classical conditioning of discrete responses during acquisition and performance of CR. The development and execution of the eyeblink CR have been considered to result from learning-related increases in neuronal responding within the interpositus nucleus. Electrophysiological

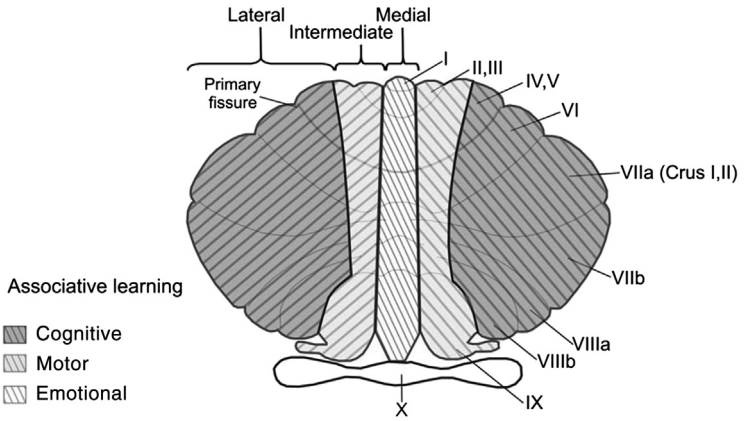

Figure 1 Cerebellar areas hypothesized to be involved in emotional, motor, and cognitive associative learning. I–X = cerebellar lobules according to Schmahmann et al. (2000); anterior lobe = lobules I–V, posterior lobe = lobules VI–X; medial, intermediate, lateral = cerebellar sagittal zones.

recordings in the cerebellar regions, which receive converging CS and US input showed that the Purkinje cells, which represent the sole output of the cerebellar cortex, discharge when the stimuli (CS or US) are presented, resulting in the inhibition of neural firing in the interpositus nucleus via GABA release, preventing a motor response (CR). In addition, it was also demonstrated that the discharge rate of Purkinje cells from the cerebellar cortex (firing rate) either decreased or increased according to a pattern that seemed time-locked to the execution of the behavioral CR. As a result, the excitatory/inhibitory firing rate pattern of the Purkinje cells likely results in the modulation of excitability of the cellular activity of the interpositus nucleus neurons and the subsequent formation of the CR.

In an impressive series of experimental studies using a spectrum of various techniques, Thompson et al. further refined and extended the classic theories of the computational neurobiologists Marr (1969) and Albus (1971), which considered the cerebellum as a favored structure for modeling a neuronal learning system because of its unique modular architecture and its huge number of neurons. Marr (1969) hypothesized that the mossy fibers serve as the environmental input while the climbing fibers of the cerebellum serve as the teaching input in the learning system. Thompson et al. further extended the neurobiological basis of the generation and control of classically conditioned eyelid responses and introduced a basic brainstem-cerebellar circuitry model, in which the interpositus nucleus and lobule VI as well as the anterior lobe of the cerebellar cortex are crucial parts (**Figure 2**). According to this model, sensory information from the CS (typically a tone CS) enters the pontine nuclei via the peripheral auditory system while the sensory information relating to the US (typically consisting of an air puff) is projected via the trigeminal nucleus to the dorsal accessory inferior olive. Simultaneously, the input entering the trigeminal nucleus is sent via the reticular formation to the cranial motor nuclei comprising the nucleus complex of the facial nerve, resulting in the production of the reflexive UR. The CS sensory information reaches the Purkinje cells of the cerebellar cortex and the interpositus nucleus via axons, which originate from the pontine nuclei (mossy fibers), while the US sensory information enters the Purkinje cells of the cerebellar cortex and the interpositus nucleus via granule cell axons originating from the inferior olive (climbing fibers). According to this model, development and execution of the eyeblink CR result from learning-related increases in neuronal responding within the interpositus nucleus. Purkinje cell output from the cerebellar cortex inhibits interpositus nucleus unit excitation and as a result CR performance. The CR pathway projects from the interpositus nucleus to the magnocellular red nucleus and cranial premotor and motor nuclei, including facial (VII), and accessory nerve (VI) nuclei.

However, although many experimental studies clearly indicate involvement of the brainstem and cerebellum in classical eyeblink conditioning, no consensus exists regarding the exact putative role of the cerebellum in this basic form of associative learning. There is ongoing discussion and even disagreement regarding the contribution of the cerebellum to associative learning as the functional neuroanatomical substrate where associative learning processes related to conditioned eyelid responses take place (acquisition and storage) or, alternatively

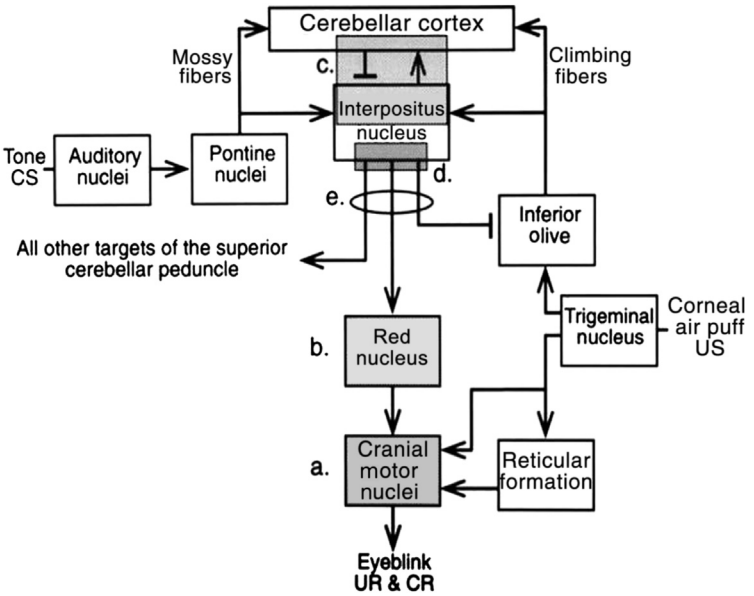

Figure 2 A simplified cerebellar circuit model for classical eyeblink conditioning from Thompson, R.F., Steinmetz, J.E., 2009. The role of the cerebellum in classical conditioning of discrete behavioral responses. Neuroscience 162, 732–755; with permission. The CS (tone) enters the circuit through the pontine nuclei. The US (an air puff or shock) is projected via the trigeminal nucleus to the inferior olive, and at the same time to the facial nucleus, forming a brainstem pathway responsible for production of the UR. Mossy fibers originating from the pontine nuclei carry CS information, while climbing fibers originating from the inferior olive carry US information into the cerebellum. Nucleus interpositus modulates the neuronal discharges of inferior olivary neurons via the nucleo-olivary pathway. Excitability changes occur in cerebellar neurons that receive input from the CS mossy fibers and US climbing fibers both in the cerebellar cortex and nuclei. The net effect of plasticity in these two areas is to increase interpositus nucleus activity, which may be capable of driving activity in brainstem neurons that control eyeblinking. The visual cortex facilitates the eyeblink conditioning and interacts with the prefrontal cortex and the hippocampus.

as a structure essential in the proper performance of eyelid responses independently of their reflex or acquired nature. In line with the performance hypothesis, Delgado-García et al. contested the basic brainstem-cerebellar circuit model on the basis of a number of inconsistencies with anatomical, kinematic, electrophysiological, and pharmacological evidence (Manto et al., 2012). In a number of studies, Delgado-García et al. do not define the role of the cerebellum in classical eyeblink conditioning as the locus of acquisition and storage of the critical memory trace but as a facilitator of the memory processes involved in the generation of CRs (Sánchez-Campusano et al., 2009).

Eyeblink Conditioning in Patients

Following the rapid progress of fundamental insights in the neural mechanisms subserving associative learning in animals during the second half of the twentieth century, human lesion studies involved the patients with structural cerebral damage and cerebellar dysfunction became of increased interest to further elucidate the neural basis of associative learning. The results obtained in a variety of etiologically heterogeneous patient populations are in accordance with the findings from animal studies, indicating that the cerebellum is crucially implicated in the acquisition and timing of the eyeblink CRs (Gerwig et al., 2007).

Weiskrantz and Warrington (1979) and Serdaru and L'Hermitte (1981) were the first authors who used an eyelid CR paradigm in a clinical population of brain-damaged patients. In these patients with medial temporal and diencephalic lesions, study results disclosed that the rate of acquisition of the eyelid CR was not affected. On the basis of these findings, Weiskrantz and Warrington (1979) concluded that the patients who could not recall the conditioning episodes (e.g., the patients with Korsakoff's syndrome) but were unimpaired with regard to the acquisition of the eyelid CR represent examples of long-term retention in amnesia syndromes. Given the emerging role of the cerebellum as the primary site of plasticity for the eyeblink CR, Salomon et al. (1989) for the first time explored the effects of damage to the cerebellum and related structures in humans on acquisition and retention of the CR. They investigated the acquisition of an eyeblink CR in a 54-year-old woman with a vascular lesion in the pontine region that affected the cerebellar afferents. In agreement with the findings obtained in animal studies and in marked contrast to an age-matched control group of five healthy subjects, significant disruption in acquiring an eyeblink CR to a tone CS and an air puff US was found in this patient, indicating that the cerebellum acts as an essential structure of plasticity for classically conditioned somatic responses. From the fact that their patient could not condition but could recall and describe the conditioning episode including the relation between the tone and the air puff, Salomon et al. (1989) concluded that their patient was an example of disrupted procedural memory with preserved declarative memory.

Subsequent group studies of patients with etiologically heterogeneous lesions of the cerebellum confirmed the findings of disrupted eyeblink conditioning reported in the early case reports. Indeed, during the past two decades eyeblink conditioning has been extensively studied across various clinical populations including patients with defined focal vascular damage of the cerebellum due to stroke (Gerwig et al., 2010) or posterior fossa tumors, primary neurodegenerative disorders of the cerebellum (Timmann et al., 2005), neuropsychiatric disorders including Alzheimer's disease (Woodruff-Pak, 2001); Huntington's disease (Woodruff-Pak and Papka, 1996); schizophrenia (Bolbecker et al., 2009a); bipolar disorder (Bolbecker et al., 2009b); abstinent chronic alcoholism and amnestic Korsakoff's syndrome (McGlinchey-Berroth et al., 1995); obsessive–compulsive disorder (Tracy et al., 1999); patients in a vegetative condition or with minimal consciousness (Bekinschtein et al., 2009), or developmental neurocognitive disorders such as attention-deficit and hyperkinetic disorder (Frings et al., 2010), autism (Oristaglio et al., 2013), and dyslexia (Nicolson et al., 2002). In a 1-year follow-up study, Timmann et al. (2005), for instance, examined delay eyeblink conditioning in 24 patients with genetically defined heredoataxias (spinocerebellar ataxia (SCA) type 6, type 3, and Friedreich's ataxia) and 10 healthy age-matched control subjects. Compared to the control group, the group of cerebellar patients was significantly impaired in various aspects of eyeblink conditioning. In contrast to the controls, no indications of retention and learning across the training sessions were found in the cerebellar patient group. In addition, the patient group showed disturbed timing of CRs. In the cerebellar patient group, the onset and peak of the CRs occurred significantly earlier than in the control group. As a result, the study clearly showed that impaired learning of CRs is a stable finding across multiple learning sessions in patients with degenerative disorders that primarily affect the cerebellar cortex. From an anatomoclinical perspective, patients with circumscribed vascular lesions of the cerebellum constitute the preferred study group to localize the conditioning-dependent learning sites within the human cerebellum. However, due to a lack of stroke patients with isolated lesions restricted to the cerebellar nuclei and circumscribed regions of the cerebellar cortex, human lesion studies are more circumspect with regard to the role of the cerebellar cortex and the deep nuclei in eyeblink conditioning. The first group study comprising a larger number of patients with focal cerebellar lesions was conducted by Gerwig et al. in 2003. In order to test the hypothesis that conditioning of the eyeblink response is distorted in patients with lesions involving the superior cerebellar artery – known to supply the deep nuclei, anterior lobe, and lobule VI of the cerebellum – and preserved in patients with lesions involving the posterior inferior cerebellar artery (PICA), Gerwig et al. (2003) compared eyeblink conditioning in 24 cerebellar stroke patients with primarily unilateral lesions (20/24) including the vascular territory of the superior cerebellar artery ($n = 11/24$) and the vascular territory of the PICA ($n = 13/24$). High-resolution three-dimensional magnetic resonance imaging (MRI) was used to identify the extent of the cortical cerebellar damage (i.e., which lobuli were affected) and the possible involvement of the cerebellar nuclei. In the group of cerebellar patients, the ability to acquire classically CRs was significantly reduced. As expected, conditioning deficits were present ipsilaterally in the patients with unilateral cerebellar lesions. Eyeblink conditioning was significantly reduced on the affected side compared with the unaffected side in patients with

superior cerebellar artery lesions including hemisphere lobules VI and Crus I. In patients with lesions restricted to the PICA territory (i.e., Crus II and below), no significant difference between the affected and unaffected sides was present. If PICA lesions extended cranially into Crus I, eyelid conditioning was significantly impaired on the ipsilateral side. Conditioning deficits were neither significantly different in superior cerebellar-artery patients with pure cortical lesions compared with the patients with additional nuclear impairment nor in superior cerebellar artery patients with unilateral lesions compared with the patients with bilateral lesions. In agreement with the findings obtained in animal studies that demonstrated the acquisition of eyeblink conditioning depends on the functional integrity of the cerebellar cortex, Gerwig et al. (2005) showed that lesions of the lobule VI and Crus I of the cerebellar cortex are sufficient to impair eyeblink conditioning in humans. With regard to the relative contribution of the deep nuclei to eyeblink conditioning in humans, no sound conclusions could be made since no lesions were restricted to the deep nuclei. In a subsequent study, Gerwig et al. (2005) investigated timing of CRs in an etiologically heterogeneous study group of 43 patients and 45 age- and sex-matched healthy control subjects. The study group consisted of 16 patients with pure cortical cerebellar degeneration (SCA type 6 and idiopathic cerebellar ataxia), 14 patients with vascular lesions located within the territory of the superior cerebellar artery, and 13 patients with infarctions within the territory of the PICA. In comparison with the control subjects, CR acquisition rate was significantly reduced and CRs occurred significantly earlier in patients with cortical cerebellar degeneration and lesions of the superior cerebellum. Incidence and timing of CRs was not impaired in patients with lesions confined to the posterior and inferior cerebellum. Voxel-based lesion-symptom analysis disclosed that cortical areas within the anterior lobe (lobule V) were most significantly related to timing deficits, whereas reduced CR incidences were related to more caudal parts (lobule VI) of the superior cerebellar cortex. The study data reveal that different cortical parts of the anterior cerebellar lobe may be implicated in the formation of the stimulus association and appropriate timing of CRs in humans. Several authors suggested that the disrupted process of eyeblink conditioning in etiologically heterogeneous cohorts of patients favors a role for the cerebellum in the pathogenesis of a variety of neuropsychiatric disorders.

Involvement of the cerebellum in eyeblink conditioning is also supported by a number of neuroimaging studies using functional MRI (fMRI) and positron emission tomography (PET) in healthy subjects. An event-related fMRI study conducted by Ramnani et al. (2000) in five healthy subjects provides clear evidence regarding the lateralization and localization of eyeblink conditioning related areas in the cerebellum. Differential delay conditioning was investigated with pseudorandomly presented trials containing auditory CS+ with a corneal air puff US (paired), or without an US (unpaired). In agreement with the findings in animal and human lesion studies, hemodynamic learning-related changes were found in the ipsilateral cerebellar cortex overlapping the medial portion of cerebellar lobule VI and the adjacent area of Crus I.

Summing up, integrity of the cerebellar circuitry seems to be a prerequisite in different forms of eyeblink conditioning in animals and humans. Since defective CR timing and UR amplitude changes cannot account for the acquisition impairments of patients with cerebellar damage, the cerebellum may be considered to be intrinsically involved in associative learning mechanisms.

Nonconditioned Learning and the Cerebellum

The role of the animal and human cerebellum in motor learning and automatization of motor sequences has attracted much attention. Animal studies in the 1970s based on cerebellar ablation, for instance, demonstrated impaired learning of motor sequences (e.g., Bleek and Peters, 1974). In addition, studies of patients with cerebellar lesions focusing on vestibulo-ocular reflex adaptation (Yagi et al., 1981) and learning of hand/arm movements (e.g., Ito, 1986) also consistently revealed cerebellar involvement in motor learning. Subsequent research disclosed that the cerebellum in addition to its role in motor learning is also crucially implicated in various forms of nonmotor learning as well. Fiez et al. (1992) described a patient who had impaired learning in implicit nonmotor learning tasks, consisting of practice-related learning tasks. Molinari et al. (1997) demonstrated that patients with cerebellar lesions also consistently failed to accomplish learning on serial reaction time tasks, in which the participants' responses are based on the spatial position of series of stimuli that either follow a sequence or occur randomly. Quintero-Gallego et al. (2006) studied the performance of children and adolescents with posterior fossa tumor resections on declarative and procedural learning tasks and showed significantly impaired procedural learning, while declarative learning was preserved. Neau et al. (2000) identified impairment in the learning condition of Rey's word list in five patients with isolated cerebellar infarcts in the PICA territory. Gottwald et al. (2004) investigated memory functions in a group of 21 patients with acquired cerebellar lesions by means of the Wechsler Memory Scale-Revised (WMS-R) and found difficulties in 'free-recall' memory subtests ('logical memory' and 'visual reproduction' subtests of the WMS-R).

A number of neuroimaging studies in the early 1990s explored the neuronal substrate of WM. Paulesu et al. (1993) reported bilateral cerebellar activation and simultaneous activation in associative and sensorimotor cortical regions in healthy volunteers during verbal WM tasks. To date, several neuroimaging studies have confirmed cerebellar involvement in verbal WM tasks, which is in accordance with clinical findings. Ravizza et al. (2006), for instance, investigated the role of the cerebellum in both visual and verbal WM tasks in 15 patients with isolated cerebellar lesions of vascular or neoplastic origin and showed that in comparison to the control subjects performance on forward and backward verbal spans was significantly depressed in cerebellar patients. In addition, the cerebellar patients tended to be more impaired on the verbal than on the spatial span tasks, which implies a cerebellar contribution to the phonological loop. Results regarding cerebellar involvement in visuospatial WM tasks are less consistent. Gottwald et al. (2004) found normal results on forward and backward visual memory spans in patients with isolated cerebellar lesions. In addition,

a meta-analysis on functional neuroimaging studies on WM (Wager and Smith, 2003) disclosed that cerebellar activity during spatial WM tasks is very rare, while consistent cerebellar activations are found during verbal WM tasks.

Cognitive Associative Learning

Although there is general agreement about involvement of the cerebellum in motor (associative) learning, the available evidence for a role of the cerebellum in cognitive associative learning is less strong. In one of the pioneer studies that reported cognitive associative learning deficits in cerebellar disorders, six patients with etiologically heterogeneous cerebellar disorders performed a reaction time paradigm, in which they had to learn the predictive value of an original neutral visual stimulus indicating a particular movement to be performed (Canavan et al., 1994). By contrast, Tucker et al. (1996) did not observe impaired visuomotor associative learning in all of the seven patients with cerebellar ataxia. However, various criticisms on methodological issues were raised against the early findings indicating impaired cognitive associative learning in patients with cerebellar lesions. Drepper et al. (1999), for instance, studied the performance on a cognitive visuomotor associative learning task of nine patients with idiopathic cerebellar ataxia in comparison to a matched control group. Subjects were asked to learn an association between pairs of colors and numerals by trial and error. Regardless of the degree of motor deficits, results clearly indicated that, in comparison to the control subjects, the ability to learn the correct cognitive associations was significantly depressed in the patient group, suggesting a contribution of the cerebellum to cognitive associative learning. Although possible confounding factors such as age, education, IQ, and visual memory had been carefully controlled for, the authors argued that the motor execution component of the task (pressing a button) might as well have captured increased attentional resources in the patient group than in the controls. In a follow-up study of nine patients with degenerative cerebellar ataxia and nine healthy controls (Timmann et al., 2010), the authors investigated by means of systematically varying the motor and cognitive load of the learning task whether the deficits in cognitive associative learning were due to increased attentional demands. Since associative learning proved to be unrelated to the motor demands of the task, the authors concluded that the cerebellum contributes to motor-independent processes involved in associative learning. In addition, a few studies (Richter et al., 2004; Thoma et al., 2008) have indicated that cerebellar dysfunction mainly affects more complex types of associative learning. Thoma et al. (2008), for instance, showed in a study with reward-based associative learning tasks with 8 cerebellar stroke patients (6 unilateral lesions = 3 right, 3 left hemisphere, 2 bilateral hemisphere lesions, 2 with deep nuclei lesions) and 24 carefully matched control subjects that impaired verbal paired associative learning was not affected in the patient group at the level of the easy items (semantically related words) and only distorted with regard to the hard items (semantically unrelated items). In a visual reward-based association learning task, in which new associations between abstract symbols and colors had to be formed on the basis of monetary feedback, the learning rates of the patients were similar to those of the control subjects in the (easy) acquisition phase of task but significantly impaired when task complexity increased (reward-based reversal learning). Since the role of the cerebellum in nonmotor associative learning has not been systematically addressed in functional neuroimaging studies or anatomoclinical research based on focal cerebellar lesions, there is currently no convergent evidence in favor of a critical location in the cerebellum that subserves nonmotor associative learning (Bellebaum and Daum, 2011).

Emotional Associative Learning

Because of its widespread reciprocal connections with distant parts of the central nervous system that regulate autonomic, emotional, and affective function, it has long been suggested that the cerebellum forms an intrinsic part of the distributed neural network subserving autonomic, emotional, and affective processing. Fear conditioning, commonly assessed by changes in autonomic and behavioral responses, represents a form of emotional associative learning, in which the vermis and fastigial nucleus of the cerebellum (limbic cerebellum) are considered to play a cardinal role as a part of an integrated network comprising the limbic system (including the amygdala and hippocampus) and the prefrontal cortex. Early electrophysiological recordings in animals and humans revealed that fastigial nucleus and vermal stimulation induce anger and fear-related responses including autonomic (blood pressure, heart rate, and gastric motility), endocrine (pituitary adrenal hormone), and behavioral changes (immobility, freezing, gnawing, and grooming) while vermal lesions decrease these responses (Heath et al., 1974).

Only a scarce number of *in vivo* studies have addressed the role of the cerebellum in emotional associative learning in humans. In these studies, the focus of interest was directed to fear conditioning. Maschke et al. (2002), for instance, studied via a classical conditioning paradigm (with a tone as the CS and an electrical shock to the right index finger as the US) fear-conditioned heart rate and skin conductance responses as measures for autonomic fear responses in five patients with medial cerebellar lesions due to posterior fossa tumor surgery. Consistent with the findings of the lesion studies in animals, the cerebellar patients, in contrast to the five nonlesioned control subjects, did not show a significant decrease of heart rate during fear conditioning. The findings obtained in the few patient studies are supported by a number of functional brain imaging studies in healthy subjects. By means of an fMRI thermal stimulus paradigm, in which subjects learned during the imaging session which of the presented color signaled pain and which signaled warmth, Ploghaus et al. (1999) investigated in 12 healthy subjects the neural activation patterns associated with acute pain stimuli and its anticipation. Pain experience activated the caudal anterior cingulate cortex, the mid-insula, and the anterior cerebellum around the vermis ('pain regions'), whereas learned anticipation to pain activated the anterior medial frontal cortex, the anterior insula, and the posterior cerebellar vermis ('anticipation

regions'). These findings indicate that nearby but albeit distinct cerebellar areas are engaged during fear experience and associative learning. However, recent functional neuroimaging studies have shown that cerebellar involvement in fear-conditioned learning is not restricted to the vermis per se but may also comprise the posterolateral cerebellar hemisphere suggesting that the vermis relates to the autonomic and somatic aspects of fear and the posterolateral region of the cerebellum to the emotional content.

Conclusion

Classical conditioning in animals and humans has undoubtedly contributed to a substantial increase of insights in the neural network subserving various forms of associative learning in the motor, cognitive, and emotional domain. Although the exact neurobiological mechanisms of cerebellar involvement in motor, cognitive, and emotional associative learning remain to be elucidated, enormous progress has been achieved in delineating the neural circuitry involved in eyeblink conditioning as the prototype of motor associative learning. In addition, a variety of experimental, clinical, and neuroimaging studies based on paradigms that closely relate to classical conditioning revealed that cerebellar disorders have a detrimental effect on cognitive and emotional associative learning as well. The contribution of the cerebellum to motor, cognitive, and emotional associative learning may relate to the role of the cerebellum in motor control processing, i.e., correct prediction of upcoming events associated with particular stimuli and/or actions. Overall, cerebellar circuits are growingly recognized as critical associative centers for sensory information seen in a broad sense. Associative learning is another example on the historical progression from a pure motor point of view to a more global sensorimotor contribution for cerebellar modules. Further investigations are still needed in larger groups of patients with circumscribed lesions to identify which parts of the cerebellum are crucially implicated in the various forms of associative learning.

See also: Aging and Memory; Amnesia: General; Amnesia: Psychogenic; Autobiographical Memory; Confabulation; Dysexecutive Amnesia; Episodic Memory; False Memories; Hippocampal Amnesia; Hippocampus and Related Structures; Learning and Memory, Neural Basis of; Non Declarative (Procedural) Memory; Post Traumatic Amnesia; Prospective Memory; Semantic Memory; Short-Term Memory: Psychological and Neural Aspects; Synaptic Mechanisms and Protein Synthesis in Memory; Transient Epileptic Amnesia; Transient Global Amnesia; Working Memory, Neural Basis of.

Bibliography

Albus, J.S., 1971. A theory of cerebellar function. Mathematical Biosciences 10, 25–61.

Baillieux, H., De Smet, H.J., Paquier, P.F., De Deyn, P.P., Mariën, P., 2008. Cerebellar cognition: insights into the bottom of the brain. Clinical Neurology and Neurosurgery 110, 763–773.

Bekinschtein, T.A., Shalom, D.E., Forcato, C., Herrera, M., Coleman, M.R., Manes, F.F., Sigman, M., 2009. Classical conditioning in the vegetative and minimally conscious state. Nature Neuroscience 12, 1343–1349.

Bellebaum, C., Daum, I., 2011. Mechanisms of cerebellar involvement in associative learning. Cortex 47, 128–136.

Bleek, C., Peters, M., 1974. Learning performance of rats with cerebellar lesions. Acta Biologica Et Medica Germanica 32, 517–525.

Bolbecker, A.R., Mehta, C.S., Edwards, C.R., Steinmetz, J.E., O'Donnell, B.F., Hetrick, W.P., 2009a. Eyeblink conditioning deficits indicate temporal processing abnormalities in schizophrenia. Schizophrenia Bulletin 111, 182–191.

Bolbecker, A.R., Mehta, C.S., Johannesen, J.K., Edwards, C.R., O'Donnell, B.F., Shekhar, A., Nurnberger, J.I., Steinmetz, J.E., Hetrick, W.P., 2009b. Eyeblink conditioning anomalies in bipolar disorder suggest cerebellar dysfunction. Bipolar Disorders 11, 19–32.

Canavan, A.G., Sprengelmeyer, R., Diener, H.C., Hömberg, V., 1994. Conditional associative learning is impaired in cerebelar disease in humans. Behavioral Neuroscience 108, 475–485.

Delgado-García, J.M., Gruart, A., 2006. Building new motor responses: eyelid conditioning revisited. Trends in Neuroscience 29, 330–338.

Drepper, J., Timmann, D., Kolb, F.P., Diener, H.C., 1999. Non-motor associative learning in patients with degenerative cerebellar disease. Brain 122, 87–97.

Fiez, J.A., Peterson, S.E., Cheney, M.K., Raichle, M.E., 1992. Impaired non-motor learning and error detection associated with cerebellar damage: a single case study. Brain 115, 155–178.

Frings, M., Gaertner, K., Buderath, P., Gerwig, M., Christiansen, H., Schoch, B., Gizewski, E.R., Hebebrand, J., Timmann, D., 2010. Timing of conditioned eyeblink responses is impaired in children with attention-deficit/hyperactivity disorder. Experimental Brain Research 201, 167–176.

Gerwig, M., Dimitrova, A., Kolb, F.P., Maschke, M., Brol, B., Kunnel, A., Böring, D., Thilmann, A.F., Forsting, M., Diener, H.C., Timmann, D., 2003. Comparison of eyeblink conditioning in patients with superior and posterior inferior cerebellar lesions. Brain 126, 71–94.

Gerwig, M., Guberina, H., Esser, A.C., Siebler, M., Schoch, B., Frings, M., Kolb, F.P., Aurich, V., Beck, A., Forsting, M., Timmann, D., 2010. Evaluation of multiple-session delay eyeblink conditioning comparing patients with focal cerebellar lesions and cerebellar degeneration. Behavioral Brain Research 212, 143–151.

Gerwig, M., Haerter, K., Hajjar, K., Dimitrova, A., Maschke, M., Kolb, F.P., Thilmann, A.F., Gizewski, E.R., Timmann, D., 2005. Timing of conditioned eyeblink responses is impaired in cerebellar patients. Journal of the Neurosciences 25, 3919–3931.

Gerwig, M., Kolb, F.P., Timmann, D., 2007. The involvement of the human cerebellum in eyeblink conditioning. The Cerebellum 6, 38–57.

Gottwald, B., Wilde, B., Mihajlovic, Z., Mehdorn, H.M., 2004. Evidence for distinct cognitive deficits after focal cerebellar lesions. Journal of Neurology Neurosurgery and Psychiatry 75, 1524–1531.

Heath, R.G., Cox, A.W., Lustick, L.S., 1974. Brain activity during emotional states. American Journal of Psychiatry 131, 858–862.

Ito, M., 1986. Long-term depression as a memory process in the cerebellum. Neurosciences Research 3, 531–539.

Leiner, H.C., Leiner, A.L., Dow, R.S., 1986. Does the cerebellum contribute to mental skills? Behavioral Neuroscience 100, 443–454.

Leiner, H.C., Leiner, A.L., Dow, R.S., 1989. Reappraising the cerebellum: what does the hindbrain contribute to the forebrain? Behavioral Neuroscience 103, 998–1008.

Lindquist, D.H., Steinmetz, J.E., Thompson, R.F., 2013. Cerebellum and eyeblink conditioning. In: Manto, M., Gruol, D.L., Schmahmann, J.D., Koibushi, N., Rossi, F. (Eds.), Handbook of the Cerebellum and Cerebellar Disorders. Springer, Amsterdam, pp. 1753–1769.

Manto, M., Bower, J.M., Bastos Conforto, A., Delgado-García, J.M., Nascimento Farias da Guarda, S., Gerwig, M., Habas, Chr, Hagura, N., Ivry, R.B., Mariën, P., Molinari, M., Naito, E., Nowak, D.A., Oulad Ben Taib, N., Pelisson, D., Tesche, C.D., Tilikete, C., Timmann, D., 2012. Roles of the cerebellum in motor control – the diversity of ideas on cerebellar involvement in movement. The Cerebellum 11, 457–487.

Marr, D., 1969. A theory of cerebellar cortex. Journal of Physiology 201, 237–270.

Maschke, M., Schugens, M., Kindsvater, K., Drepper, J., Kolb, F.P., Diener, H.C., Daum, I., Timmann, D., 2002. Fear conditioned changes of heart rate in patients with medial cerebellar lesions. Journal of Neurology, Neurosurgery and Psychiatry 72, 116–118.

McGlinchey-Berroth, R., Cemak, L.S., Carillo, M.C., Armfield, S., Gabrieli, J.D., Disterhoft, J.F., 1995. Impaired delayed eyeblinking conditioning in amnesic Korsakoff's patients and recovered alcoholics. Alcoholism: Clinical and Experimental Research 19, 1127–1132.

Molinari, M., Leggio, M., Solida, A., Ciorra, R., Misciagna, Silveri, M.C., Petrosini, L., 1997. Cerebellum and procedural learning: evidence from focal cerebellar lesion. Brain 120, 1753–1763.

Neau, J.P., Arroyo-Anllo, E., Bonnaud, V., Ingrand, P., Gil, R., 2000. Neuropsychological disturbances in cerebellar infarcts. Acta Neurologica Scandinavica 102, 363–370.

Nicolson, R.I., Daum, I., Schugens, M.M., Fawcett, A.J., Schult, A., 2002. Eyeblink conditioning indicates cerebellar abnormality in dyslexia. Experimental Brain Research 143, 42–50.

Oristaglio, J., West, S.H., Ghaffari, M., Lech, M.S., Verma, B., Harvey, J.A., Welsh, J.P., Malone, R.P., 2013. Children with autism spectrum disorders show abnormal conditioned response timing on delay, but not trace, eyeblink conditioning. Neuroscience. S0306-4522(13)00504-6.10.1016/.

Paulesu, E., Frith, C.D., Frackowiak, R.S.J., 1993. The neural correlates of the verbal component of working memory. Nature 362, 342–345.

Ploghaus, A., Trcey, I., Gati, J.S., Clare, S., Menon, R.S., Matthews, P.M., Rawlins, J.N., 1999. Dissociating pain from its anticipation in the human brain. Science 18, 1979–1981.

Quintero-Gallego, E.A., Gomez, C.M., Vaquero Casares, E., Marquez, J., Pérez-Santamaria, F.J., 2006. Declarative and procedural learning in children and adolescents with posterior fossa tumors. Behavioral and Brain Functions 2, 1–9.

Ramnani, N., Toni, I., Josephs, O., Ashburner, J., Passingham, R.E., 2000. Learning- and expectation-related changes in the human brain during motor learning. Journal of Neurophysiology 84, 3026–3035.

Ravizza, S.M., McGormick, C.A., Schlerf, J.E., Justus, T., Ivry, R.B., Fiez, J.A., 2006. Cerebellar damage produces selective deficits in verbal working memory. Brain 129, 306–320.

Richter, S., Matthies, K., Ohde, T., Dimitrova, A., Gizewski, E., Beck, A., Aurich, V., Timmann, D., 2004. Stimulus-response versus stimulus-stimulus-response learning in cerebellar patients. Experimental Brain Research 158, 438–449.

Sánchez-Campusano, R., Gruart, A., Delgado-García, J.M., 2009. Dynamic associations in the cerebellar–motoneuron network during motor learning. Journal of Neuroscience 29, 10750–10763.

Schmahmann, J.D., 2010. The role of the cerebellum in cognition and emotion: personal reflections since 1982 on the dysmetria of thought hypothesis, and its historical evolution from theory to therapy. Neuropsychology Review 20, 236–260.

Schmahmann, J.D., Sherman, J.C., 1998. The cerebellar cognitive affective syndrome. Brain 121, 561–579.

Serdaru, M., L'Hermitte, F., 1981. Etude des syndromes amnesiques par duchotically du conditionnement palpebruk et de la reaction de sursaut et d'orientation. Revue Neurologique 137, 39–47.

Solomon, P.R., Stowe, G.T., Pendlbeury, W.W., 1989. Disrupted eyelid conditioning in a patient with damage to cerebellar afferents. Behavioral Neuroscience 103, 898–902.

Thoma, P., Bellebaum, C., Koch, B., Schwarz, M., Daum, I., 2008. The cerebellum is involved in reward-based reversal learning. Cerebellum 7, 433–443.

Thompson, R.F., 2005. In search of memory traces. Annual Review of Psychology 56, 1–23.

Thompson, R.F., Steinmetz, J.E., 2009. The role of the cerebellum in classical conditioning of discrete behavioral responses. Neuroscience 162, 732–755.

Timmann, D., Drepper, J., Frings, M., Maschke, M., Richter, S., Gerwig, M., Kolb, F.P., 2010. The human cerebellum contributes to motor emotional and cognitive associative learning: a review. Cortex 36, 845–857.

Timmann, D., Gerwig, M., Frings, M., Maschke, M., Kolb, F.P., 2005. Eyeblink conditioning in patients with hereditary ataxia: a one year follow-up study. Experimental Brain Research 162, 332–345.

Tracy, J.A., Ghose, S.S., Stetcher, T., McFall, R.M., Steinmetz, J.E., 1999. Classical conditioning in a nonclinical obsessive-compulsive population. Psychological Science 10, 9–13.

Tucker, J., Harding, A.E., Jahanshahi, M., Nixon, P.D., Rushworth, M., Quinn, N.P., Thompson, P.D., Passingham, R.E., 1996. Associative learning in patients with cerebellar ataxia. Behavioral Neuroscience 110, 1229–1234.

Wager, T.D., Smith, E.E., 2003. Neuroimaging studies of working memory: a meta-analysis. Cognitive, Affective, & Behavioral Neuroscience 3, 255–274.

Weiskrantz, L., Warrington, E.K., 1979. Conditioning in amnesic patients. Neuropsychologica 17, 187–194.

Woodruff-Pak, D.S., 2001. Eyeblink classical conditioning differentiates normal aging form Alzheimer's disease. Integrative Physiological and Behavioral Science 36, 87–108.

Woodruff-Pak, D.S., Papka, M., 1996. Huntington's disease and eyeblink classical conditioning. Normal learning but abnormal timing. Journal of the International Neuropsychological Society 2, 323–334.

Woodruff-Pak, D.S., Steinmetz, J.E. (Eds.), 2000. Eyeblink Classical Conditioning: Animal Models. Kluwer, Boston.

Yagi, T., Shimizu, M., Sekine, S., 1981. A new neurological test for detecting cerebellar dysfunction. Annals of the New York Academy of Sciences 374, 526–531.

Assortative Mating in the Marriage Market

Aloysius Siow, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Using data from the 2000 U.S. census, the first part of this article shows how local log odds are used to measure three forms of marriage matching patterns: ethnic endogamy, positive assortative matching by educational attainment, and social exchange between educational attainment and ethnicity. Using the Choo–Siow marriage matching model, the second part of the article provides a behavioral interpretation of using local log odds to measure marriage matching patterns.

In most marriage markets, there are different types of participants. They differ by age, ethnicity, educational attainment, religion, citizen status, and so on. In spite of this large diversity of participants, individuals tend to marry (match with) spouses who have similar characteristics to their own. This commonly observed tendency to marry someone like themselves is known as positive assortative matching (PAM) in marriage.

The first part of this article is to present a simple way to quantify different forms of PAM. The most common way to quantify PAM on one ordered characteristic, such as years of education, is to compute the correlation in the characteristics between husbands and wives (e.g., Fernandez et al., 2005). While useful, such a correlation cannot be used to quantify matching when characteristics are unordered such as ethnicity or religion. It is also less useful when there are multiple characteristics and we want to consider trade-offs between different characteristics.

Sociologists and demographers also commonly use log-linear models (which is a parametrization of contingency tables) to quantify PAM (e.g., Goodman, 1972; Kalmijn, 1991; Mare and Schwartz, 2005 for U.S. data; Blossfeld, 2009; Smits and Park, 2009 have international comparisons). There is confusion about how to define PAM using log-linear models and how to use them to quantify trade-offs between multidimensional characteristics (Gullickson and Fu, 2010). Log-linear models is a parametric class of local log odds models. Instead of using log-linear models, it will be more transparent to work with local log odds models directly (e.g., Gullickson and Fu, 2010; Siow, in press).

We will use local odds models first to quantify endogamy (propensity to marry someone with the same characteristic) for an unordered characteristic. Then we will quantify them to PAM for one ordered characteristic. Finally, we will use them to study trade-offs between two characteristics when individuals are characterized by more than one characteristic. For illustration, we will estimate the various local log odds using marriage matching data from the U.S. 2000 census.

Recently, researchers have investigated whether the degree of PAM by spousal educational attainment in the US has increased in the last 30–40 years (e.g., Mare and Schwartz, 2005; Siow, in press). Within that time period, women have caught up and passed men in average college attainment. Also marriage rates have fallen. These researchers ignore the issue of changes in population supplies and marriage rates when they study changes in PAM over time. They also often study the evolution of educational PAM holding other characteristics of the spouses, such as age and ethnicity, constant. While empirically convenient, is there a behavioral justification for these two empirical shortcuts?

The second part of this article applies Siow which uses the Choo and Siow (2006; hereafter CS) marriage matching model to provide a unified behavioral interpretation of the matching behavior characterized by these local log odds. Conveniently, the CS model shows that one can study marriage matching (1) independent of variation in population supplies and (2) holding other characteristics of spouses constant.

Measuring Endogamy

Individuals differ along different characteristics. Some of these characteristics are ordered. One can have more or less of an ordered characteristic such as age, income, and years of education. Some characteristics cannot be ordered such as ethnicity and religion. It will be convenient first to discuss matching by unordered characteristics.

When a characteristic is unordered, all participants may not rank the desirability of different potential spouses along those characteristics in the same way. Consider the case of matching by religion. An atheist may regard a potential Jewish spouse as similar to a potential Buddhist spouse. On the other hand, a Muslim may rank a potential Jewish spouse differently from a potential Buddhist spouse. With unordered characteristics, we can distinguish between a tendency of each type of spouse to marry their own type (endogamy) versus intermarrying.

Consider the number of spousal matches by ethnicity: White (W), Black (B), and Hispanic (H), in the U.S. 2000 5% census between 31 and 35 years old females and 32–36 years old males in **Table 1**. (For this article, an individual is Hispanic if the person is of Hispanic origin independent of any other racial designation.)

Table 1 Ethnic marital matching in the U.S. 2000 census

		Female		
		White	*Hispanic*	*Black*
Male	White	112 732	3054	315
	Hispanic	2482	15 646	111
	Black	791	265	9608

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31100-X

A striking feature of the table is that preponderance of marriages along the main diagonal of the table which are marriages where both spouses have the same ethnicity (ethnic endogamy or homogamy). One way to quantify endogamy is to measure the number of marriages where both spouses have the same ethnicity relative to ones where one spouse is another ethnicity. The problem with this measure is that although all Blacks in the above table can potentially marry White spouses, not all White spouses can potentially marry Black or Hispanic spouses. Thus, one has to take into account availability of potential spouses in quantifying endogamy or PAM.

Let T the number of different types individuals of each gender in the marriage market. In Table 1, $T = 3$. Let i (j) be the different types of husbands (wives) by ethnicity where i (j) = W, B, H. Let μ_{ij}, the value of the (i, j) cell of Table 1, be the number of type i husbands married to type j wives.

In this article, we will use local log odds ratio to measure local endogamy or PAM. Consider W husbands and wives who are either married to each other or intermarry B spouses. There are $m_{W(W,B)} = \mu_{WW} + \mu_{WB}$ and $f_{(W,B)W} = \mu_{WW} + \mu_{BW}$ of these husbands and wives, respectively. Also consider B husbands and wives who are either married to each other or intermarry W spouses. There are $m_{B(W,B)} = \mu_{BW} + \mu_{BB}$ and $f_{(W,B)B} = \mu_{WB} + \mu_{BB}$ of these husbands and wives, respectively. The sum of the number of husbands and wives must be equal to each other, $m_{W(W,B)} + m_{B(W,B)} = f_{(W,B)W} + f_{(W,B)B}$.

The marriage pattern between Blacks and Whites is described by the local log odds ratio for these marriages, l (W, B), as

$$l(W,B) = \ln\left[\frac{\mu_{WW}\mu_{BB}}{\mu_{WB}\mu_{BW}}\right] \qquad [1]$$

The numerator consists of the product of the numbers of endogamous marriages. The denominator consists of the product of the numbers of intermarriages. Given the number of marriages of one type, and the numbers of husbands and wives of each type, $m_{W(W,B)}$, $m_{B(W,B)}$, $f_{(W,B)W}$, and $f_{(W,B)B}$, the numbers of the other three types of marriages are determined. The local log odds is also determined. Put another way, given the local log odds and the numbers of husbands and wives of each type, the numbers of marriages of each type is determined. Thus the local log odds summarizes all marriage matching patterns which are feasible given the numbers of husbands and wives of each type.

Considering these four types of marriages, random matching of spouses or independence in marriage matching for males of type t and females of type t' implies

$$\mu_{tt'} = m_{t(t,t')}\left(\frac{f_{(t,t')t'}}{f_{(t,t')t} + f_{(t,t')t'}}\right) \qquad [2]$$

Using the above formula for the different types of marriages in eqn [1], the local log odds for random matching of spouses of two different ethnicities, t and t', is

$$l(t,t') = 0 \qquad [3]$$

That random matching implies $l(t, t') = 0$ is important because it says that the distributions of husbands and wives do not constrain the local log odds in a particular way.

Type W and B individuals practice local endogamy if $l(W, B) > 0$ and local exogamy $l(W, B) < 0$. The degree of local endogamy is the value of $l(W, B)$. From Table 1, $l(W, B) = 8.37$ which is far from zero, the criteria for random matching by race.

The value of $l(W, B)$ is independent of the values of $l(W, H) = 5.45$ or $l(B, H) = 8.54$, i.e., any other degrees of local endogamy. Thus for each racial pair, they all practice endogamy.

The total number of local log odds, or local endogamy measures, is $T(T-1)/2$. Since the number of local endogamy measures depends on T, it may be difficult to compare the degree of endogamy across societies or even in one marriage market across time.

Let $\Phi(x)$ be an indicator function such that

$$\Phi(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \le 0 \end{cases}$$

In order to compute an overall measure of endogamy for a marriage market with T types of an unordered characteristic, consider two overall indexes of endogamy:

$$E_T = \sum_{t=1}^{T}\sum_{k<t}\frac{\Phi(l(t,k))}{T(T-1)/2} \qquad [4]$$

$$E_T^W = \sum_{t=1}^{T}\sum_{k<t}\left[\frac{\ln\left(f_{(t,k)t} + f_{(t,k)k}\right)}{\sum_{h=1}^{T}\sum_{g<h}\ln f_{(h,g)h} + f_{(h,g)g}}\right]\Phi(l(t,k)) \qquad [5]$$

E_t is the fraction of local log odds which are strictly positive. E_T^W is a weighted average of the fraction of local log odds which are strictly positive. The log population weighting scheme gives more weight, at a diminishing rate, to types with larger population shares. If we use simple population shares for the weights, E_T^W will be dominated by the local log odds involving Whites. Under random matching, $E_T^F = E_T^{WF} = 0$. When the values of the indexes are larger than zero, we say the society is endogamous.

Let x^y the value of x in year y. In Table 1 with the 2000 census, $E_3^{F2000} = E_3^{WF2000} = 1$. For comparison, in 1980, $E_3^{F1980} = E_3^{WF1980} = 1$. To compute indices which capture whether endogamy has increased, consider

$$\Delta E_T = \sum_{t=1}^{T}\sum_{k<t}\frac{\Phi\left(l(t,k)^{2000} - l(t,k)^{1980}\right)}{T(T-1)/2}$$

$$\Delta E_T^W = \sum_{t=1}^{T}\sum_{k<t}\frac{\sqrt{\omega^{2000}(t,k)\omega^{1980}(t,k)}}{\sum_{h=1}^{T}\sum_{g<t}\sqrt{\omega^{2000}(h,g)\omega^{1980}(h,g)}}\Phi\left[l(t,k)^{2000} - l(t,k)^{1980}\right]$$

$$\text{where } \omega^y(t,k) = \frac{\ln\left(f^y_{(t,k)t} + f^y_{(t,k)k}\right)}{\sum_{h=1}^{T}\sum_{g<h}\ln\left(f^y_{(h,g)h} + f^y_{(h,g)g}\right)}$$

$\Phi(l(t,k)^{2000} - l(t,k)^{1980}) = 1$ when $l(t,k)^{2000} > l(t,k)^{1980}$ and there was an increase in endogamy in the racial pair (t, k) between 1980 and 2000. It is zero otherwise. So the index, ΔE_T, is the fraction of racial pairs where there were increases in endogamy between the two time periods. If endogamy

increased for all racial pairs, $\Delta E_T = 1$. Conversely, if endogamy decreased for all racial pairs, $\Delta E_T = 0$. ΔE_T^W is a weighted fraction of racial pairs where there were increases in endogamy between the two time periods. The weighting scheme gives more weight, at a diminishing rate, to racial pairs with larger population shares.

With the 2000 and 1980 data, $\Delta E_3^F = 0.0$ and $\Delta E_3^W = 0.0$. For all three racial pairs, endogamy decreased between 1980 and 2000. There is no difference between the unweighted and weighted indices. So racial endogamy unambiguously declined in the United States between 1980 and 2000.

Measuring PAM

When a characteristic is ordered such as the age or schooling attainment of an individual, we can measure more than just whether spouses practice endogamy or not. We will consider a more general notion of spousal matching, PAM.

Let each male in the marriage market be characterized by an ordered characteristic i, $i = 1, \ldots, I$, where $i + 1 > i$. Let each female in the marriage market be characterized by an ordered characteristic j, $j = 1, \ldots, J$, where $j + 1 > j$. Let μ_{ij} be the number of type i males married to type j females.

Table 2 shows the number of spousal matches by four categories of schooling attainment, less than high school (*LHS*), high school graduate (*HS*), some college, BA or more (*BA*), in the U.S. 2000 5% census between 31 and 35 years old females and 32–36 less than high school years old males. This is the same population as in Table 1. The value in cell (i, j), μ_{ij}, is the number of type i males married to type j females.

Table 2 shows there is also endogamy by schooling attainment. Individuals are more likely to marry spouses with the same level of education. But Table 2 also shows that individuals are more likely to marry someone closer to themselves in education than further away. So *LHS* individuals are more likely to have *HS* spouses than *LBA* spouses and even less likely to have *BA* spouses. Again we have to worry about feasibility. For example, not every *HS* male can marry an *LHS* female even if he wanted to.

Here, we extend the local log odds concept above to measure PAM. Consider types i and $i+1$ husbands who are married to types j and $j+1$ wives. There are $m_{i(j,j+1)} = \mu_{ij} + \mu_{i,j+1}$ and $f_{(i,i+1)j} = \mu_{ij} + \mu_{i+1,j}$ of type i husbands and type j wives, respectively. There are also $m_{i+1,(j,j+1)} = \mu_{i+1,j} + \mu_{i+1,j+1}$ and $f_{(i,i+1),j+1} = \mu_{i,j+1} + \mu_{i+1,j+1}$ of type $i+1$ husbands and type $j+1$ wives, respectively. The sum of the number of husbands and wives must be equal to each other, $m_{i(j,j+1)} + m_{i+1,(j,j+1)} = f_{(i,i+1)j} + f_{(i,Wi+1),j+1}$.

Table 2 Educational marital matching in the U.S. 2000 census

		Female			
		LHS	*HS*	*LBA*	*BAM*
Male	LHS	7905	5171	3209	713
	HS	3704	19 920	13 070	4249
	LBA	1850	9183	21 945	10 359
	BAM	430	2864	9830	31 133

LHS, less than high school; HS, high school graduate; LBA, some college.

To characterize the marriage matching patterns among these adjacent types, the local log odds ratio for these marriages, $l(i, j)$, is defined as

$$l(i,j) = \ln\left[\frac{\mu_{ij}\mu_{i+1,j+1}}{\mu_{i+1,j}\mu_{i,j+1}}\right] \qquad [6]$$

The local log odds is defined for any (i, j) pair as long as $i < I$ and $j < J$. There are $(I-1) \times (J-1)$ local log odds in total. Note also that I need not equal J. Table 3 shows the all the local log odds computed from Table 2.

As earlier, given the numbers of husbands and wives of each type, $m_{i,(j,j+1)}$, $m_{i+1,(j,j+1)}$, $f_{(i,i+1)j}$, and $f_{(i,Wi+1),j+1}$, and the local log odds ratio, $l(i, j)$, the numbers of marriages of each type is determined. The local log odds summarizes all marriage matching patterns which are feasible given the numbers of husbands and wives of each type. Each local log odds is independent of the values of the other local log odds. Put another way, the marriage matching distribution $\mu_{ij} > 0$, for all i and j, is completely summarized by the $(I-1) \times (J-1)$ local log odds and the numbers of married individuals of each type.

As earlier, when there is random matching between types i and $i+1$ husbands with types j and $j+1$ wives, $l(i, j)$ will be equal to zero. Again, that random matching implies $l\ (i, j) = 0$ is important because it says that the distributions of husbands and wives do not constrain the local log odds in a particular way. If $l(i, j)$ is positive, there is PAM between these adjacent types which says that individuals prefer to marry spouses who are closer to them in characteristic than further away. If $l(i, j)$ is negative, there is negative assortative matching between these adjacent types which says that individuals prefer to marry spouses who are further away from them in characteristic than closer by.

Statisticians say that spouses in a marriage market practice PAM if all $(I-1) \times (J-1)$ local log odds are larger than zero or are totally positive of order 2 (TP2). This is a strong requirement. Following the statistics literature, Siow provides an estimator for all the local log odds which imposes TP2.

Consider the following two overall indexes of PAM for a society.

$$P_{IJ} = \sum_{i=1}^{I-1}\sum_{j=1}^{J-1}\frac{\Phi(l(i,j))}{(I-1)\times(J-1)} \qquad [7]$$

$$P_{IJ}^W = \sum_{i=1}^{I-1}\sum_{j=1}^{J-1}\left[\frac{\ln\left(m_{i(j,j+1)} + m_{i+1,(j,j+1)}\right)}{\sum_{h=1}^{I-1}\sum_{g=1}^{J-1}\ln\left(m_{h(g,g+1)} + m_{h+1,(g,g+1)}\right)}\right]\Phi(l(i,j))$$

Table 3 Local log odds in education matching in the U.S. 2000 census

		Female		
		LHS/HS	*HS/LBA*	*LBA/BA*
Male	LHS/HS	2.11	0.06	0.38
	HS/LBA	−0.08	1.29	0.37
	LBA/BA	0.29	0.36	1.9

LHS, less than high school; HS, high school graduate; LBA, some college; BA, BA or more.

P_{IJ} is the fraction of local log odds which are positive. P_{IJ}^{W} is the weighted average of the fraction of local log odds which are positive. Again we use log population weights to give more weight, at a diminishing rate, to types with larger population shares. When there is random matching in marriage by all individuals, $P_{IJ} = P_{IJ}^{W} = 0$. Using Table 3, $P_{44}^{2000} = 0.89$, $P_{44}^{W2000} = 0.89$. As a point of comparison, $P_{44}^{1980} = 1$ and $P_{44}^{W1980} = 1$. The equivalent indexes for increases in PAM are $\Delta P_{44} = 0.44$, $\Delta P_{44}^{W} = 0.44$. In other words, slightly more than half of the local log odds fell between 1980 and 2000. Thus there is little evidence in favor of an increase in PAM by educational attainment between 1980 and 2000.

Considering changes in educational endogamy, $\Delta E_4 = 0.33$ and $\Delta E_4^{W} = 0.33$. Again here, slightly more than half of the local log odds for educational endogamy fell between 1980 and 2000. Thus there is also little evidence in favor of an increase in endogamy by educational attainment between 1980 and 2000.

Summarizing the empirical evidence, there is evidence in favor of increases in educational PAM and endogamy between the two periods. The weighted indexes do not give quantitatively different results than the unweighted indexes.

Trade-Offs in Multidimensional Matching

There is no standard definition of PAM when individuals are characterized by multidimensional attributes (See Dupuy and Galichon, 2012 for an overall approach). Our goal here is modest and application based. We follow Gullickson and Fu to use local log odds to study status exchange in marriage between Black and Whites (Merton, 1941; Davis, 1941), an important application of multidimensional PAM.

When individuals are characterized by multidimensional attributes, a husband may have a higher value of an attribute X than his wife and a lower value of another attribute Y than her. Sociologists argue that the husband and wife are engaging in social exchange in marriage. Applications of this theory include: A Black husband with a White wife who has lower educational attainment than him. Indian women may offer larger dowries to their husbands who are from a higher caste. A wealthy man may marry a younger beautiful but less wealthy or educated woman as a trophy wife.

Following Gullickson and Fu, consider the case where individuals are characterized by their race ($r = W, B$) and $E_m = E_f = 4$ levels of educational attainment ($e = LHS, HS, LBA, BA$). For notational convenience, let $e + 1$ denote the educational level immediately higher than e. Table 4 shows the numbers of marriages of each match type for the 2000 U.S. census data. An entry in Table 4, μ_{r_m,e_m,r_f,e_f}, is the number of marriages between husbands of race r_m and schooling e_m, and wives of race r_f and schooling e_f.

A striking feature of the data is that significantly more Black men have White wives than Black women have White husbands. Under the social exchange hypothesis, demographers and sociologists hypothesize that some White women intermarry racially for a more educated husband. Equivalently, some educated Black males are willing to marry a less-educated White woman to intermarry White females.

To quantify the above intermarriage phenomena, consider marriages between Black men with education e_m or $e_m + 1$ and White and Black women with education e_f. Both Black and White women have the same level of education. We want to study their relative propensity to marry up.

Let the local log odds $l_{B,e_m,(B,W),e_f}$ denote

$$l_{B,e_m,(B,W),e_f} = \ln\left[\frac{\mu_{B,e_m}+1,W,e_f}{\mu_{B,e_m,W,e_f}}\right]\left[\frac{\mu_{B,e_m}+1,B,e_f}{\mu_{B,e_m,B,e_f}}\right]^{-1} = \ln\frac{D}{C} \quad [8]$$

D is the ratio of the number of White women who married $e_m + 1$ Black men relative to the number marrying e_m Black men. C is the ratio of the number of Black women who married $e_m + 1$ Black men relative to the number marrying e_m Black men. If $D > C$, then White women with education e_f have a larger log odds of marrying $e_m + 1$ Black men to e_m Black men than Black women with the same education. The local log odd $l_{B,e_m,(B,W),e_f} > 0$ consistent with social exchange in marriage between Black men and White women. If $l_{B,e_m,(B,W),e_f} = 0$, there is random matching between these two types of husbands and wives. With four educational level for each gender, there are $(E_m - 1) \times E_f = 3 \times 4 = 12$ of these local log odds ratios.

Table 5 shows the values of these 12 local log odds ratios.

Let $m_{B,(e_m e_{m+1}),(BW),e_f}$ be the number of Black men with education e_m or e_{m+1} who are married to Black or White women

Table 4 Ethnic and education marital matching in the U.S. 2000 census

		Female							
		White				Black			
Male		LHS	HS	LBA	BA	LHS	HS	LBA	BA
White	LHS	2624	3411	1971	441	5	6	9	2
	HS	2240	15 910	10 530	3574	5	24	34	8
	LBA	1024	7288	17 053	8432	3	18	52	23
	BA	229	2336	8192	27 314	4	9	33	79
Black	LHS	15	22	19	3	403	380	316	74
	HS	25	99	51	16	291	1533	1134	282
	LBA	18	79	167	77	161	623	1880	760
	BA	4	11	59	126	30	141	491	1063

LHS, less than high school; HS, high school graduate; LBA, some college; BA, BA or more.

Table 5 Local log odds for B men with e_m or e_m+1 married to W and B women with e_f

	Female			
Black male	*W_LHS\B_LHS*	*W_HS\B_HS*	*W_LBA\B_LBA*	*W_BA\B_BA*
LHS\HS	0.84	0.11	−0.29	0.34
HS\LBA	0.26	0.67	0.68	0.58
LBA\BA	0.18	−0.49	0.30	0.16

LHS, less than high school; HS, high school graduate; LBA, some college; BA, BA or more.

of education e_f. We compute both unweighted and population-weighted fraction of strictly positive local log odds to get indices for the propensity of White women to intermarry racially in order to marry a more educated spouse:

$$X_{B,e_m,(B,W),e_f} = \sum_{e_m=1}^{E_m-1}\sum_{e_f=1}^{E_f}\frac{\Phi\left(l_{B,e_m,(B,W),e_f}\right)}{(E_m-1)\times E_f} \quad [9]$$

$$X^W_{B,e_m,(B,W),e_f} = \sum_{e_m=1}^{E_m-1}\sum_{e_f=1}^{E_f}\left[\frac{\ln m_{B,(e_m,e_{m+1}),(BW),e_f}}{\sum_{h=1}^{E_m-1}\sum_{g=1}^{E_f}\ln m_{B,(h,h+1),(BW),g}}\right]\Phi\left(l_{B,e_m,(B,W),e_f}\right)$$

If the indices are larger than 0.5, then the society exhibits social exchange in marriage in the sense defined above. The indices are bounded above by 1 and bounded below by zero. Based on data in Tables 4 and 5, $X_{B,e_m,(B,W),e_f} = 0.83$ and $X^W_{B,e_m,(B,W),e_f} = 0.83$. So there is empirical support for social exchange in marriage between White women and Black men.

What about social exchange in marriage between Black women and White men? Consider Black women with education e_f or e_f+1 and White and Black men with education e_m. Both Black and White men have the same level of education. We want to study their relative propensity to marry up.

Interchanging genders in the definition of the local log odds in eqn [8], Table 6 shows the values of the 12 local log odds ratios $l_{B,e_f,(B,W),e_m}$.

Similarly, based on data in Tables 4 and 6, $X^F_{B,e_f,(B,W),e_m} = 0.58$ and $X^{W\,F}_{B,e_f,(B,W),e_m} = 0.59$. As discussed earlier, if there is no social exchange, we expect the value of the index to be 0.5. Thus $X^F_{B,e_f,(B,W),e_m} = 0.58$ and $X^{W\,F}_{B,e_f,(B,W),e_m} = 0.59$ show that there is minimal social exchange in marriage between White men and Black women. Our findings, (1) that there is social exchange in marriage between White women and Black men, and (2) there is minimal social exchange in marriage between White men and Black women replicate the conclusions in Gullickson and Fu (2010).

Table 6 Local log odds for B women with e_f or e_f +1 married to B or W men with e_m

	Female		
Male	*LHS\HS*	*HS\LBA*	*LBA\BA*
W_LHS\B_LHS	0.24	0.59	−0.05
W_HS\B_HS	−0.09	0.65	−0.06
W_LBA\B_LBA	0.44	−0.44	0.09
W_BA\B_BA	−0.74	0.05	0.10

LHS, less than high school; HS, high school graduate; LBA, some college; BA, BA or more.

Finally, in all the three studies of ethnic endogamy, PAM, and social exchange, the values of the unweighted versus population-weighted indices measuring a particular marriage matching pattern were quantitatively similar. So in these cases, there is little gain to compilation weighted indices. The simpler unweighted indices sufficed.

A Behavioral Interpretation of Local Log Odds in Marriage Matching

When there are more marriages of (i, j) couples than marriages of (h, k) couples, analysts often infer that (i, j) couples must be more attracted to each other than (h, k) couples. There are three related concerns with such an inference. First, there may be more h and k individuals in the population than i and j individuals. In other words, we need to control for differences in population supplies. Second, is 'more attracted to' a statement about preferences for different types of spouses or is it a combination of spousal preferences plus resource allocations within those marriages. For example, young women who marry older rich men may prefer to marry younger men but the older men can provide more resources to their young wives. Third, can we learn anything about male versus female preferences from observing marital matching patterns?

This section applies Siow which uses the CS model to provide a behavioral interpretation of local log odds in marriage matching which also deals with the three concerns raised above (Siow and CS have been further characterized and extended in several directions (e.g., Decker et al., 2013; Galichon and Salanie, 2012; Graham, 2011)). The CS model belongs to the class of frictionless transferable utility marriage models which was introduced by Becker (1973; reprinted 1991) to bring demand and supply analysis to marriage markets.

Consider a marriage market where there are I types of men, $i = 1, \ldots, I$, and J types of women, $j = 1, \ldots, J$. M is a population vector where an element m_i is the number of eligible (single) men of type i. F is a population vector where an element f_j is the number of eligible (single) women of type j. μ_{ij} is the number of marriages between type i men and type j women. μ_{i0} and μ_{0j} are the number of type i and type j women, respectively, who remains unmarried.

Each marital match between two different types of individuals constitutes a distinct submarriage market. With I types of men and J types of women, there are $I \times J$ submarriage markets.

Let the systematic marital output of a type i man married to the type j woman be π_{ij} which is divided between the two spouses.

Let $\tilde{\tau}_{ij}$ be the share of the systematic marital output that is obtained by a type j wife. Each wife also gets an idiosyncratic payoff from marriage which depends on her specific identity, the type of spouse that she marries and not his specific identity. Her idiosyncratic payoff also does not depend on $\tilde{\tau}_{ij}$.

In an (i, j) marriage, $\pi_{ij} - \tilde{\tau}_{ij}$ is the share of systematic marital output that is obtained by a type i husband. Each

husband also gets an idiosyncratic payoff that is specific to him, the type of spouse that he marries and not her specific identity. His idiosyncratic payoff also does not depend on $\pi_{ij} - \tilde{\tau}_{ij}$.

The above assumptions imply that every type i male regards every type j female as perfect spousal substitutes and vice versa.

Each individual also gets a systematic payoff from remaining unmarried which depends on their type, π_{i0} and π_{0j} for the type i man and type j woman, respectively, as well as an idiosyncratic payoff which depends on their specific identity.

Given their payoffs, both systematic and idiosyncratic, from every potential spousal choice including remaining unmarried, each individual will choose the spousal choice which maximizes their utility.

The roles of systematic versus idiosyncratic payoffs are as follows. Consider two types of men, i and i', that a type j woman may marry. If the systematic share $\tilde{\tau}_{ij}$ exceeds $\tilde{\tau}_{i'j}$, then all other things equal, type j women will prefer to marry type i men rather than type i' men. With the addition of idiosyncratic payoffs, a particular type j woman may get a very large idiosyncratic payoff from marrying a type i' man compared with her idiosyncratic payoff from marrying a type i man and so choose to marry the type i' man in spite of the unfavorable difference in systematic payoffs. So the difference in systematic payoffs captures the observation that type j women are more likely to marry type i men than type i' men. The idiosyncratic payoffs allow some (i', j) marriages.

We aggregate these individual decisions into demand and supply functions for spouses in every (i, j) submarriage market. For a given transfer, $\tilde{\tau}_{ij}$, the demand by type i men for type j wives may not equal the supply of type j women for type i husbands. If the demand for wives exceeds the supply, $\tilde{\tau}_{ij}$ will rise and vice versa.

The marriage market clears with equilibrium transfers $\tilde{\tau}_{ij} = \tau_{ij}$, when the number of type i men who want to marry type j women will be exactly equal to the number of type j women who want to marry type i men for all (i, j) submarriage markets. In marriage market equilibrium, every individual chooses the type of spouse they want to marry or chooses to remain unmarried.

Using a particular formulation for systematic and idiosyncratic payoffs, McFadden's (1974) additive random utility model, for modeling spousal choice, CS showed that for every (i, j):

$$\ln \frac{\mu_{ij}}{\sqrt{\mu_{i0}\mu_{0j}}} = \frac{\mu_{ij}}{\sqrt{\left(m_i - \sum_k \mu_{ik}\right)\left(f_j - \sum_h \mu_{hj}\right)}} \qquad [10]$$

$$= \frac{\pi_{ij} - \left(\pi_{i0} + \pi_{0j}\right)}{2} \qquad [11]$$

The left-hand side of eqn [10] is observed by the analyst. It is the log of the number of (i, j) marriages divided by the geometric mean of the number of unmarried type i men and unmarried type j women. CS calls the left-hand side of eqn [10] the total gains to marriage. The right-hand side of [11] provides a behavioral interpretation of the total gains to marriage. It is the per spouse systematic payoff to type i man marrying a type j woman minus the systematic payoffs to them remaining unmarried. See CS for an application to changes in the U.S. marriage market induced by the legalization of abortion; Brandt et al. (2008) for estimating the marital outcome of children born during the 1960 famine in rural China; Bruze et al. (in press) on marriage, divorce, and remarriage in Denmark; Chiappori et al. on the evolution of marriage matching by educational attainment in the US (2011).

There are three important points about eqn [10]. First, equilibrium transfers or the equilibrium division of resources between spouses, τ_{ij}, do not appear in [10]. Thus the total gains to marriage do not provide any information on within household allocations between spouses. Second, variations in populations of different type of individuals affect equilibrium transfers but do not affect the total gains to marriage. Third, the total gains to marriage do not provide any information on who, the husband or wife, contributes more to the marriage.

Apply eqn [10] to marriages between men of types i and i' and women of types j and j';

$$\ln \frac{\mu_{ij}\mu_{i'j'}}{\mu_{ij'}\mu_{i'j}} = \ln \frac{\mu_{ij}}{\sqrt{\mu_{i0}\mu_{0j}}} + \ln \frac{\mu_{i'j'}}{\sqrt{\mu_{i'0}\mu_{0j'}}} - \ln \frac{\mu_{ij'}}{\sqrt{\mu_{i0}\mu_{0j'}}} - \ln \frac{\mu_{i'j}}{\sqrt{\mu_{i'0}\mu_{0j}}} \qquad [12]$$

$$= \frac{1}{2}\left(\pi_{ij} + \pi_{i'j'} - \pi_{ij'} - \pi_{i'j}\right) \qquad [13]$$

The left-hand side of eqn [12] is a local log odds of this marriage market. It is observed by the analyst. Equation [13] provides an interpretation of the local log odds in terms of the marital output function, π_{ij}. We will use it to interpret the various local log odds which were discussed earlier.

Rationalizing Endogamy

Let $i = j = W$ and $i' = j' = B$, then eqn [12] becomes

$$l(W, B) = \ln\left[\frac{\mu_{WW}\mu_{BB}}{\mu_{WB}\mu_{BW}}\right] = \frac{1}{2}\left(\pi_{WW} + \pi_{BB} - \pi_{WB} - \pi_{BW}\right) \qquad [14]$$

The local log odd in eqn [14] is the same as that in eqn [1], the local log odd for Black and White marriages. The right-hand side of [14] provides a behavioral interpretation of the local log odd. The right-hand side is positive when the sum of marital outputs from own race marriages exceeds the sum of marital outputs from mixed race marriages. The right-hand side is negative when the sum of marital outputs from own race marriages is less than the sum of marital outputs from mixed race marriages. The right-hand side is equal to zero when the sum of marital outputs from own race marriages is equal to the sum of marital outputs from mixed race marriages.

When $l(W, B) > 0$, CS rationalizes endogamy by saying that the sum of marital outputs from own race marriages is less than the sum of marital outputs from mixed race marriages. Sociologists and demographers interpret local log odds and related statistics as quantifying 'forces of attraction' in marriage. Some economists have worried about the 'force of attraction' interpretation of local log odds because supply considerations are not present, i.e., the number of Blacks and Whites do not appear in $l(B, W)$. CS provides a formal rationalization for 'forces of attraction' interpretation. Moreover, CS derives the interpretation of the local log odds from a model of marriage

market equilibrium which explicitly takes supply considerations into account.

Rationalizing PAM

Now let men and women be differentiated along one dimension of ability which is educational attainment. So with I types of men, let the education of a type i man be i. So the educational attainment of a type i man is lower than that of a type $i+1$ man. And similarly with J types of women, let the educational attainment of a type j woman be j.

Let $i' = i+1$ and $j' = j+1$. Then eqn [12] becomes

$$l(i,j) = \ln\left[\frac{\mu_{ij}\mu_{i+1,j+1}}{\mu_{i,j+1}\mu_{i+1,j}}\right] = \frac{1}{2}\left(\pi_{ij} + \pi_{i+1,j+1} - \pi_{i,j+1} - \pi_{i+1,j}\right) \quad [15]$$

When the marital output function is π_{ij} supermodular, i.e., the right-hand side of eqn [15] is strictly positive, then the local log odds $l(i, j)$ is positive. Put another way, $l(i, j) > 0$ implies that the sum of marital outputs by PAM is larger than the sum of marital outputs by negative assortative matching. Siow was the first to discover this result. Graham provides the most general version of this result to date. These results are the stochastic generalization of Becker's famous deterministic PAM matching result where he showed that if the marital output function is supermodular, there will be strict PAM in marriage. The stochastic generalizations are empirically useful because strict PAM is never observed.

Note again that population supplies do not directly affect eqn [12]. Thus analysts who want to study PAM using log-linear models and local log odds models do not have to worry about population supplies.

Rationalizing Trade-Offs in Multidimensional Matching

Now let men and women be differentiated by race and educational attainment. Let a type i man be a Black man with $e_m + 1$ amount of education. Let a type i' man be a Black man with e_m amount of education. Let a type j woman be a White woman with e_f amount of education. Let a type j' woman be a Black woman with e_f amount of education. Then eqn [12] becomes

$$l_{B,e_m,(B,W),e_f} = \ln\left[\frac{\mu_{B,e_m+1,W,e_f}}{\mu_{B,e_m,W,e_f}}\right]\left[\frac{\mu_{B,e_m+1,B,e_f}}{\mu_{B,e_m,B,e_f}}\right]^{-1} \quad [16]$$

$$= \pi_{B,e_m+1,W,e_f} - \pi_{B,e_m+1,B,e_f} - \left(\pi_{B,e_m,W,e_f} - \pi_{B,e_m,B,e_f}\right) \quad [17]$$

$l_{B,em,(B,W),ef}$ in eqn [16] is the same as that in eqn [8]. So there will be social exchange in marriage between Black men and White women when the difference in marital outputs of between White and Black wives with Black husbands with $e_m + 1$ schooling is larger than the difference in marital outputs between White and Black wives with Black husbands with e_m schooling.

Conditional PAM

Most empirical research on spousal matching by educational attainment condition on other spousal characteristics. For example, this article conditioned on age and ethnicity. Marital outputs are affected by these other characteristics. This subsection will show that the above results apply conditional on other spousal characteristics.

Together with i, let there be K, $k = 1, \ldots, K$ other characteristics which defines a type of male in the marriage market. k may be a White male between ages 32 and 36. Thus a type of male is now indexed by (i, k). There are $I \times K$ types of men on the society. Together with j, let there be L, $l = 1, \ldots, L$ other characteristics which define a type of female in the marriage market. l may be a White female between ages 31 and 35. Thus a type of female is now indexed by (j, l). There are $J \times L$ types of women on the society.

Consider the CS model where the systematic marital output of an $(i, k) \times (j, l)$ marriage is denoted as $\pi_{(i,k),(j,l)}$. The additive idiosyncratic payoff an individual to any spousal choice is as before. Marital transfers are used to clear the $I \times K \times J \times L$ marriage markets simultaneously. Then Siow shows

$$\ln\frac{\mu_{(i,k),(j,l)}}{\sqrt{\mu_{(i,k),0}\mu_{0,(j,l)}}} = \frac{\pi_{(i,k),(j,l)} - \pi_{(i,k),0} - \pi_{0,(j,l)}}{2} \forall (i,k),(j,l)$$

Conditioning on k and l, we obtain the conditional local log odds:

$$\ln\frac{\mu_{(i+1,k),(j+1,l)}\mu_{(i,k),(j,l)}}{\mu_{(i,k),(j+1,l)}\mu_{(i+1,k),(j,l)}}$$

$$= \pi_{(i+1,k),(j+1,l)} + \pi_{(i,k),(j,l)} - \left[\pi_{(i,k),(j+1,l)} + \pi_{(i+1,k),(j,l)}\right]$$

All the previous results in the above subsections apply conditional on k and l. For example, conditional on k and l, if the marital output function $\pi_{(i,k),(j,i)}$ is supermodular in i and j, the conditional spousal educational matching distribution is $TP2$. Thus CS rationalizes the standard practice of studying spousal educational matching conditional on other spousal characteristics as is also done here.

Acknowledgment

I thank Jin Hu for excellent research assistance.

See also: Dowries and Other Marital Transfers; Ethnic Intermarriage; Family Theory: Economics of Marriage and Divorce; Population Dynamics: Two-Sex Demographic Models.

Bibliography

Becker, Gary S., 1973. A theory of marriage: part I. Journal of Political Economy 81, 813–846.

Becker, Gary S., 1991. A Treatise on the Family. Harvard University Press, Cambridge, MA.

Blossfeld, Hans-Peter, 2009. Educational assortative marriage in comparative perspective. Annual Review of Sociology 35, 513–530.

Brandt, Loren, Siow, Aloysius, Vogel, Carl, 2008. How Flexible Is the Marriage Market? Evidence from Famine Born Cohorts in China. University of Toronto Working Paper.

Bruze, Gustaf, Michael, Svarer, Yoram, Weiss. The dynamics of marriage and divorce. Journal of Labor Economics, in press.

Chiappori, Pierre A., Salanié, Bernard, Weiss, Yoram, 2011. Partner Choice and the Marital College Premium.

Choo, Eugene, Siow, Aloysius, 2006. Who marries whom and why. Journal of Political Economy 114, 175–201.

Davis, Kingsley, 1941. Intermarriage in caste societies. American Anthropologist 43, 376–395.

Decker, Colin, Lieb, Elliott H., McCann, Robert J., Stephens, Benjamin K., 2013. Unique equilibria and substitution effects in a stochastic model of the marriage market. Journal of Economic Theory, Elsevier 148 (2), 778–792.

Dupuy, Arnaud, Galichon, Alfred, 2012. Personality Traits and the Marriage Market. manuscript.

Fernandez, Raquel, Guner, Nezih, Knowles, John, 2005. Love and money: a theoretical and empirical analysis of household sorting and inequality. The Quarterly Journal of Economics 120 (1), 273–344.

Fu, Vincent K., 2001. Racial intermarriage pairings. Demography 38 (2), 147–159.

Goodman, Leo A., 1972. Some multiplicative models for the analysis of cross classified data. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability. Theory of Statistics, Vol. I. Univ. California, Berkeley, CA, 1970/1971, pp. 649–696.

Galichon, Alfred, Salanie, Bernard, 2012. Cupid's Invisible Hand: Social Surplus and Identification in Matching Models. manuscript.

Graham, Bryan, 2011. Econometric methods for the analysis of assignment problems in the presence of complementarity and social spillovers. In: Benhabib, Jess, Bisin, Alberto, Jackson, Matthew (Eds.), Handbook of Social Economics, 1B. North-Holland, Amsterdam, pp. 965–1052.

Gullickson, Aaron, Fu, Vincent Kang, 2010. Comment: an endorsement of exchange theory in mate selection. American Journal of Sociology January 115 (4), 1243–1251.

Kalmijn, Matthijs, 1991. Status homogamy in the United States. American Journal of Sociology 97, 496–523.

Mare, Robert D., Schwartz, Christine R., 2005. Trends in educational assortative marriage from 1940 to 2003. Demography 42 (4), 621–646.

McFadden, Daniel, 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka, Paul (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.

Merton, Robert, 1941. Intermarriage and the social structure: fact and theory. Psychiatry: A Journal of the Biology and the Pathology of Interpersonal Relations 4, 361–374.

Siow, Aloysius. Testing Becker's theory of positive assortative matching. Journal of Labor Economics, in press.

Smits, Jeroen, Park, Hyunjoon, 2009. Five decades of educational assortative mating in 10 East Asian societies. Social Forces 88 (1), 227–255.

Asylum Seekers

Linda Briskman, Swinburne University of Technology, Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The treatment of asylum seekers in the modern nation-state violates human right norms. Exclusion from nation, an emphasis on border protection, and the use of immigration detention facilities are among the approaches aimed at deterring people from seeking safe haven. Many of these practices are antithetical to social work ethics and values, and social work practitioners can use their practice wisdom and advocacy skills to promote alternatives to current policies and practices.

Introduction

The rights of asylum seekers are one of the most publicized and politicized public policy issues. This article outlines the global context of asylum seeking and border control and highlights the distinguishing features of the modern nation-state response, including immigration detention. The question of racialized policies is in the forefront. Attention is also given to the role of social work, and suggestions are made for ways forward for more humane approaches that cohere with human rights norms.

Background

The terms asylum seeker and refugee are often confused. An asylum seeker is a person who says that he or she is a refugee, but whose claim has not yet been assessed. A refugee has been assessed by a government or by the United Nations High Commissioner for Refugees (UNHCR) as set out in the 1951 Convention Relating to the Status of Refugees, commonly known as the Refugee Convention (AHRC, 2012). Countries that have ratified the Convention are obliged to assess asylum seekers' claims for protection.

According to the Refugee Convention, a refugee is a person who is outside his or her own country and is unable to return due to a well-founded fear of persecution on the following grounds: race, religion, nationality, membership of a particular social group or political opinion (UNHCR, 1951). More than 140 countries have acceded to the 1951 Convention or its 1967 Protocol, and many of these countries have incorporated the refugee definition of the Convention within national legislation (UNHCR, 2007).

In this article, the terms asylum seeker and refugee are occasionally used interchangeably to denote shared experiences and reflect the literature.

There are approximately 15.2 million refugees in the world (RCOA, n.d.). The 20 nations with the highest number of refugees are all developing countries, including 12 least developed countries, with more than 42% of the world's refugees residing in nations where gross domestic product per capita is below US$3000 (UNHCR, 2010).

The top three countries of origin of refugees under UNHCR's mandate are Afghanistan, Iraq, and Somalia (RCOA, 2010). The top refugee host countries include Pakistan and Iran, both non-Western countries. Pakistan hosts 2 million refugees, the United Kingdom around 238 000, and Australia approximately 22 000 (RCOA, 2010). In terms of resettlement of refugees, the top three countries are the United States, Canada, and Australia (RCOA, 2010). Resettlement opportunities are extremely limited; in 2009, a mere 1% of the world's refugees were resettled and estimates are that it would take 188 years for the world's refugees to be resettled (RCOA, n.d.). Formal protection spaces are shrinking and many people are in protracted refugee situations. A protracted refugee situation is defined as one in which refugees have been in exile for more than 5 years and for whom solutions are not evident in the foreseeable future (Loescher and Milner, 2009).

As signatories to the Refugee Convention harden their entry stance, restrictions imposed have the effect of encouraging refugees to engage people smugglers to assist them in crossing borders, often in a dangerous manner. Between 1993 and 2010, nearly 14 000 people are known to have died trying to enter Europe, while in detention, or during forcible deportations (Weber and Pickering, 2011). The causes of deaths vary according to context, but include drowning during boat journeys, suffocation in lorries or containers, vehicle accidents, suicides, and environmental exposure in deserts (Weber and Pickering, 2011).

Border Security vs Human Security

A preoccupation exists with the sanctity of nation-state borders, which is privileged over the security of people. As border security measures increase, the security of people in many countries is threatened by what are known as push factors, which propel movement to countries perceived to be safe. Lester (2010) argues that the breakdown in the human security of people has a ripple effect that transcends borders. Factors such as conflict, inequality, human rights abuses, terrorism, environmental degradation, and transnational organized crime are not isolated events. Yet, asylum seekers are cast as agents of human insecurity, which is evident in the xenophobic responses of host states (Lester, 2010), and can have devastating consequences for the fleeing asylum seeker.

Many asylum seekers have experienced torture and trauma and the provision of casework services to this group is often the point of entry for social workers. Survivors of torture may experience chronic symptoms that continue for many years

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28008-2

including hyperarousal, sleep problems, irritability, poor memory and concentration, reexperiencing traumatic events, and social withdrawal (Bowles, 2005). The experience of torture and trauma regrettably does not activate a welcoming response for asylum seekers for, as Edwards and Ferstman (2010) argue, rather than being viewed as people exercising their rights to seek asylum, as enshrined in international law, they are denoted as invaders. This means that as noncitizens they are dealt with by discrimination, xenophobia, and criminalization. Grewcock (2009) says that criminalizing people for a legal activity results in people being treated like criminals through naval interception, maximum security prisons, the employment of guards, and the use of physical constraints and weaponry. This compounds the trauma already experienced.

Attempts to eliminate the people smuggling trade are the cornerstone of border protection. Alison Mountz (2010) notes that human smuggling is in many ways routine and a historical phenomenon that has existed for centuries. Among accounts of the need for human smugglers are those who used such services to flee the Nazi Holocaust. Those who assisted them, such as Oskar Schindler, are now portrayed as heroic. In popular discourses, however, smugglers are portrayed as enemies of the state.

Racism

It is not possible to critique the manner in which asylum seekers are treated without examining questions of racism, a construct that is familiar to social workers through antiracist theorizing. Institutional racism takes forms that are ingrained in practices that privilege dominant paradigms. Since the attacks in the United States on 11 September 2001, Muslims, including Muslim asylum seekers, have been depicted as terrorists who are seen as intent on changing societal ways of life and unappreciative of Western values. The so-called war on terror since 2001 has ushered in new state racism through laws, administrative edicts, and judicial systems that shape popular racism (Sivandan, 2007). By conflating asylum seeking with terrorism, nation-states have been able to introduce harsh measures that are supported by their constituents. When fear is injected into the populace about those depicted as 'the threatening other', there is acceptance of inhumanity to vulnerable people who are seeking help.

Asylum policies can disproportionately penalize people who are not legal citizens of a nation. Asylum seekers can be punished merely because they cross borders without documentation to seek asylum as enshrined in international norms. The Canadian group 'No-One is Illegal' argues that granting citizenship to a privileged few represents racist immigration and border policies aimed at exploiting and marginalizing migrants (NOI, n.d.).

Racism and xenophobia are not caused by isolated acts of individuals. Rather, racism expressed at the individual level is representative of "an orchestrated effort by segments of the dominant society to wage a war on the poor and on people who, by virtue of their race, ethnicity, language, and class are reduced at best to half-citizens and at worst to a national enemy responsible for all the ills afflicting our society" (Macedo and Bartolome, 2001: p. 6). Institutionally, the way race thinking materializes is evident in the field of migration, where nations are increasingly putting up their shutters to keep out people who are seen not only as a potential terrorist threat but also as undermining the nation's demographic order and the value base of society. This results in insidious assimilation approaches that require people to relinquish their identities to fit into the dominant society. This is antithetical to social work principles and constructs of cultural diversity and antiracism. In the new world order, assimilation ideologies fuse with security discourses. These security discourses provide the rationale for racialized immigration policies (Burke, 2001). In addition, asylum seekers and refugees are rarely accorded a voice of their own and they are:

> Typically a two-dimensional character, a blank screen upon which other's agendas and narratives can be projected. The actions of refugees are narrated for the general public of the Global North by charities, refugee advocates, health professionals, politicians and others.
>
> (Fiske, 2012: p. 17)

Moreover, the perceptions of asylum seekers and refugees are often mediated by sections of the media, which defy their own codes of ethics by using language that demonizes asylum seekers and colludes with governments, directly and indirectly, to invoke fear.

The Use of Immigration Detention

Immigration detention is one of the starkest illustrations of the tough manner in which asylum seekers are treated. Detention deprives people of liberty and such deprivation is one of the most serious sanctions a state can issue against a person and in other circumstances is normally proscribed except when authorized by a court (Briskman et al., 2008). The Geneva-based Global Detention Project (GDP) (n.d.) explains the use of detention as the practice of detaining on administrative (as opposed to criminal) grounds. The GDP states that because many national legal systems do not have clear rules for such detention, those detained face legal uncertainties, including the absence of limitations on the duration of their detention.

The detention of people deemed unauthorized has grown enormously in scope and scale in the past 30 years and many developed nations hold unauthorized or rejected foreigners in prisonlike facilities (Wilsher, 2012). This is nowhere clearer than in the case of asylum seekers who can often be held in an arbitrary manner and for extended periods, outside the mainstream of the law, in inhumane conditions (Wilsher, 2012). Although the UNHCR guidelines on detention (UNHCR, 2012) specifically prohibit detention for the purposes of deterrence as such policies are generally unlawful under international law, this practice is increasing. The privatization of detention is also a common feature whereby more than a dozen countries have employed private contractors in some form within their immigration detention institutions (Flynn and Cannon, 2009).

Australia has one of the harshest detention regimes in the world and is a country illustration of how stridently this

provision is implemented especially as, on a global scale, it receives a relatively low number of asylum seeker arrivals. Since 1992, legislation has been in place to detain mandatorily people arriving without authorization. Mandatory detention is mostly applied to asylum seeker boat arrivals and frequently takes place in remote and island locations, including in the nations of Nauru and Papua New Guinea.

Detention in Australia applies to men, women, and children. Two Australian social work academics describe the practice of detaining children as 'organized and ritualized abuse' by the Australian government (Goddard and Briskman, 2004: p. 17). As of February 2014 approximately 1000 children were held in closed immigration detention (AHRC, 2014), prompting the announcement of an official inquiry by the Australian Human Rights Commission into child detention. At the same date, around 116 children are believed to be held in Nauru detention facilities (Chilout, 2014).

The detention of children occurs in other nations. A survey of 21 countries found that there is an increasing trend to detain migrant, asylum seeker, and refugee children (IDC, 2012a). The IDC (2012b) explains that such detention has negative physical, psychological, and emotional impacts.

In the United Kingdom although fewer children are being detained than under the previous government, the number of children held in what is known as predeparture facilities is increasing (End Child Detention Now, 2012). Furthermore, a report by the Refugee Council in the United Kingdom accuses the immigration service of detaining child asylum seekers by wrongly classifying them as adults (McVeigh, 2012). Human Rights Watch (2012) in criticizing proposed legislation in Canada speaks of the violation of human rights law by allowing for a year of detention without review for certain groups of people, including 16- and 17-year old children (Human Rights Watch, 2012).

There are alternatives to detention in place in some contexts. According to the International Detention Coalition (IDC) (2012b), countries such as Belgium have alternatives to immigration detention for children and countries like Japan, Holland, and the United Kingdom have recently taken steps to release children or to prevent totally their detention. In Australia, under the Labor government, which was voted out of office in September 2013, there was increasing use of community options during claims processing. However, 1765 children are in situations where their parents have no work rights and limited access to government support (ChilOut, 2014). Social workers are among those who have been employed in programs contracted to nongovernment organizations that provide support to asylum seekers in the community. The emphasis on the change to a Coalition government is now on a policy named Operation Sovereign Borders that gives priority to turning back boats to Indonesia and reducing the number of arrivals to Australia. Clearly, there is no place for social workers in militiarized responses.

Islands as Places of Banishment

A disturbing trend is the use of islands for immigration detention purposes. Holding people on islands makes them more difficult to access and more detached from advocates, lawyers, and other services. Mountz and Briskman (2012) note:

> At the unaccountable border, nation-states can exercise their power with impunity and brutality. In these sites of exclusion, liberty is denied and normative state rights minimised. Identity is discounted and the demonised, criminalised asylum seeker is uncritically perceived by those far away from the island in the nation's heartland as a threat to the nation, creating panic and fear.
>
> (Mountz and Briskman, 2012: p. 23)

The islands of Guam and Saipan, Lampedusa, Christmas Island, and Guantanamo Bay are cases in point. These islands are linguistically, culturally, economically, geographically, and jurisdictionally different from one another. In each region, however, these islands feature in struggles over migration and each island has a history of crisis surrounding marine arrivals (Coddington et al., 2012).

As an archipelago of small Pacific Islands under US jurisdiction, Guam and the Commonwealth of Northern Mariana Islands are geographically remote from, and geopolitically significant to, the United States. Guam and Saipan are apparent in US endeavors to manage and restrict asylum-based immigration (Coddington et al., 2012). Immigration detention facilities are located at the US naval station on Guantanamo Bay; the use of Guantanamo for this purpose is not generally known (Dastyari and Effeney, 2012). The Italian island of Lampedusa has been a significant point of arrival for migrants and refugees from North Africa where they are detained in such substandard conditions that it has been reported that the European Commission may initiate legal action (CBC News, 2013).

Christmas Island, an Indian Ocean Territory of Australia, is a site of detention containment. Until the change to a Coalition government in 2013 and the subsequent Operation Sovereign Borders policy, boat arrivals usually reached Christmas Island from the transit destination of Indonesia. Nearly all were detained on Christmas Island for initial processing before being transferred to the mainland, with adult men usually held in a maximum security facility. Now asylum seekers who have made their way to Christmas Island are held there while awaiting transfer to Nauru or Manus Island (Papua New Guinea).

The Role of Social Workers

Although there are internationally agreed definitions and standards about social work, as Banks (2012) notes how social work is practiced in different countries is linked to context including national and regional welfare regimes, social welfare laws and policies, the roles of the state, as well as market and not-for-profit organizations. There is also an imperative to think beyond the local to international forms of social work (Ife, 2001) and as discussed below, social workers need to traverse the global and the local in considering their roles and responses to situations.

The involvement of social workers with asylum seekers has not been surveyed on an international basis. Collett's (2004: p. 84) observation in the United Kingdom that "asylum seekers

are still not regarded as significant enough to be a focus of social work training" is mirrored in other countries.

Collett (2004) refers to the gatekeeping role of social workers in the United Kingdom. Rather than viewing asylum seekers as clients, they can be drawn into informing on behalf of immigration services. This characteristic has been subject to robust debates within social work of how social control ideologies have characterized appraisals of professionalism, particularly where practitioners may be expected to provide the contradictory roles of care agent and control agent (Briskman et al., 2012). Social work has been criticized for the attention given to the management of social problems on behalf of governments. Although conceptions of abuse have widened to include institutional abuse (Cemlyn and Briskman, 2003), social workers can inadvertently become partners in implementing repressive approaches that are masked as providing caring services. The activities of the profession in the United Kingdom with regard to asylum seekers tend to be derived from values concerned with investigation, rationing, gatekeeping, and surveillance. These are separated from ethics, sensitivity, accountability, and empathy (Humphries, 2004). Although not often expressed in the social work literature, dual loyalty is an explanatory framework.

Dual loyalty issues arise where the loyalty of a health or welfare professional to patient or client is compromised by obligation to the employing body, which may represent the interests of the state. Historically, the ethical obligations of health and welfare professionals have privileged the need for loyalty to patients and clients (Briskman et al., 2012), but these boundaries may be crossed with asylum seekers.

An examination of research conducted with nurses who worked in immigration detention in Australia sheds light on the conflict that besets those who conduct human services work in closed environments. Deborah Zion et al. (2009) interviewed a sample of nurses who had worked on the frontline of health care in immigration detention services. Although nurses were ascribed responsibility and power, it was limited to enforcing the rules set out by the private detention providers and immigration department. Against the odds, many focused on the ethics of nursing to enact a duty of care and display empathy, but this could not compensate for the inhuman conditions in which asylum seekers found themselves. This would have required political action (Zion et al., 2009: p. 550).

Social workers and other human services professionals can be compromised in their work, particularly if they are silenced through contracts of confidentiality and limits to speaking out. Gready's (2007) analysis of medical complicity in human rights abuses in apartheid South Africa resonates. He argues (p. 417) that it is difficult to draw the line between active and passive collaboration, referring to 'moral disengagement'.

Humphries (2004) provides some directives for social workers, arguing that social work is not inevitably an arm of oppression and the profession has a choice to make a new moral effort to engage its knowledge about sources of inequality with a sense of urgency. Asylum seekers, says Humphries, are a dramatic example of people regarded as undeserving, worthy of derision, abysmal treatment, and expulsion.

An International Human Rights Approach

Human rights social work academic Jim Ife (2001) laments the way in which social problems are often framed as national problems, with little acknowledgment of shared global problems, global causes, and global solutions. Khan and Dominelli (2000) refer to how globalization contributes to the internationalization of social problems, with some social workers facing issues that transcend national boundaries including asylum migration. Ife concurs, noting that with refugees and asylum seekers, an issue so obviously global and internationally linked, the response by activists is often confined to national perspectives seen as 'Canada's refugee problem', 'Australia's treatment of illegal immigrants', or 'the need for Kosovo to be made safe again' (Ife, 2001: p. 74). He argues for all of social work to be concerned with the global and all social work in this sense is international social work. Similarly, Phillippe Legrain (2006) says that focusing on individual nation-states provides a distorted and partial view, for the increase in international migration is not occurring in a vacuum but is part of a globalization process.

The Universal Declaration of Human Rights (United Nations, 1948) is a starting point for fair treatment. It states that all people have the right to seek and to enjoy in other countries asylum from persecution. It also states that all people have the right to work, that no one should be subjected to arbitrary detention; cruel, inhuman, or degrading treatment; or interference in their family unit. All these provisions are breached by many nation-states. It should serve as a reminder to violating nation-states and their supporters that at the end of the Second World War the Declaration was designed to guarantee the rights of every individual and to uphold dignity and justice for all. More than 50 states participated in its drafting. Apart from the Refugee Convention of 1951, other relevant instruments that signed-up nation-states need to be cognizant about in their treatment of asylum seekers are the Convention against Torture and the Convention on the Rights of the Child.

To develop clarity about their positions, a human rights approach provides a basis for assertive practice in confronting violations (Briskman and Cemlyn, 2005). A human rights social work approach ought be a political endeavor. Furthermore:

> Social workers with a commitment to anti-oppressive and anti-racist philosophies should inform themselves of the local context, in order to challenge stereotypes of racism. They have a role in counteracting the dehumanisation of asylum seekers and educating the wider community.
>
> (Briskman and Cemlyn, 2005: p. 721)

Advocacy and Social Work: Prospects and Limits

While harsh policies are meted out to asylum seekers, there have been concerted endeavors to protest against and reverse dominant discourses and inhumane practices. These are evident in social movements in a number of countries whereby organizations and individuals form collectives to proffer alternatives to cruel regimes and to expose the injustices. In their contest with the state, social movement actors may be discounted in

dominant narratives of nationhood and nation building that are portrayed by the state (Wills, 2009). Nonetheless, those involved in social movements provide a humanizing ethic of how those discarded by the nation-state can be treated in a responsible manner (Stratton and McCann, 2002). It is unknown how many social workers have been involved in such movements either professionally or individually; however, there are some documented examples including the social work-inspired People's Inquiry into Detention described below.

Among the groups that have coalesced for the rights of asylum seekers, refugees, and undocumented migrants are those on the US-Mexico border, and sanctuary movements based around church organizations to provide protection for people facing deportation (Gosden, 2012). Australia has a long history of asylum seeker activism with the common aim of ending mandatory detention. In the United Kingdom, a large antidetention lobby exists that ranges from exposing detention practices, particularly in the medical sphere, to organizing campaigns for the release of children. In Canada, the 'No-One is Illegal' group fights for the rights of all migrants to live with dignity and respect.

Global organizations also exist including the IDC that has 250 members working in 50 countries to undertake research, advocate, and provide direct services to and on behalf of refugees, asylum seekers, and migrants (IDC, n.d.). The GDP is an interdisciplinary research endeavor that investigates the role detention plays in states' responses to global migration (Global Detention Project, n.d.).

Social workers can form alliances at local, regional, and international levels to promote policies and practices that address fundamental injustices. Doing so requires a review of the knowledge and skills needed and the approaches that are relevant (Lyons et al., 2006). Social workers are among those who have joined movements to bring about change and provide support to asylum seekers. They are among the membership of refugee action collectives, are in touch with legal representatives, join in demonstrations, and visit detention centers (Cemlyn and Briskman, 2003).

Social work ethics guide action. Wherever in the world social workers practice, they deal with ethical challenges about how to treat and respond to people respectfully and whether and how to resist, ameliorate, or tolerate the social injustices they see on a daily basis (Banks, 2012). There are directives and principles enshrined in social work ethics, national and international, which call upon social workers to affirm human rights and to challenge unjust principles. The International Statement of Ethical Principles proclaims that the principles of human rights and social justice are fundamental to social work (IASSW and IFSW, 2004).

Apart from codified ethics, the practice wisdom of social workers can facilitate action. Social workers are witness to the impact of subjugation, oppression, racism, and structural disadvantage. By applying theories of anti-oppressive practice, and a human rights framework to analyzing policies on the ground, an emancipatory outlook can be adopted that challenges and questions the harms done to vulnerable people. Social work educators have an important role.

One Australian example of social workers enacting their moral responsibility was an initiative undertaken under the auspices of the Australian Council of Heads of Schools of Social Work (ACHSSW). From 2005 to 2008, the ACHSSW facilitated a People's Inquiry into Detention to expose and influence the harmful policies that breached core human rights tenets. The ACHSSW had no authority to enact such a project except the authority of 'the people' and a commitment to give voice to the voiceless who were incarcerated in Australia's detention centers (Briskman et al., 2008).

There are risks attached to advocacy endeavors including when initiated by social work educators. One only has to contemplate what the National Association of Scholars (NAS) had to say in a 2007 article titled 'The scandal of social work education'. The article accused social work educators of politicization that undercuts the intellectual foundations of the modern university, which the NAS described as the search for truth. They criticized what they saw as a trend in universities to doctrinaire ideological commitments, polemics in place of teaching, and political activism. Social work education, the NAS said, is an advanced case of politicization, and coerced intellectual conformity. It even went so far as to criticize the American Council on Social Work Education's protocol about affirming social justice ideals, stating that this has meant students have had their consciences coerced.

This criticism is disturbing as social workers are mandated, professionally and morally, to use their expertise, academic freedom, and professional networks to enact projects that scrutinize injustices and the flouting of human rights obligations. The question arises of moral responsibility and collective action when silence occurs. What is social workers' responsibility when people are effectively silenced through the political system, as has been the case with detained asylum seekers? The role of social work academics is crucial. As asserted by Noble et al., (2009: p. 415):

> To work in academic settings is a privilege that comes with both the opportunity and the responsibility to speak on behalf of those whose voices remain suppressed, or too weak to be heard.

Although tensions exist for social work educators, who have a degree of academic freedom, they are magnified for practitioners. Procedural control by organizations can result in limited scope for questioning of policies and practices decided at the corporate center (Hough and Briskman, 2003). Nonetheless, social work educators can assist students in finding a pathway through the limitations by helping them find a way to privilege social work principles rather than organizational tenets. Through instilling the courage to act, students can become moral agents (Ife, 2001). Role modeling by social work educators can inspire social work students to become advocates for asylum seekers. Students need to be cognizant of the risks of acting and not acting, and educators can take on a mentoring role.

Conclusion

Social work is a profession that is practiced within contradictory paradigms. The question remains as to how to position social workers as actors in the political realm of asylum seekers.

In a globalized world, social workers need to emerge from their own context to engage with the way in which dominant discourses are discordant with the value base of social work. Social workers can recognize their practice wisdom and reposition themselves as practice ethnographers. This entails drawing upon the knowledge derived through practice so that it does not merely remain within the annals of case notes and peer discussion, but can also contribute to social change. The experiential knowledge that resides within social work professionals themselves (Lyons et al., 2006) can be at the forefront of social work practice and advocacy.

See also: Armed Conflict and Social Work: Significance and Implications; Critical Social Work Practice; Human Rights and Social Work; Migration, Ethnicity, Aging and Social Work Practice; Refugees in Anthropology.

Bibliography

Australian Human Rights Commission (AHRC), 2012. Face the Facts: Some Questions and Answers about Indigenous Peoples, Migrants and Refugees and Asylum Seekers. AHRC, Sydney.

Australian Human Rights Commission (AHRC), 2014. National Inquiry into Children in Immigration Detention 2014. Discussion Paper. AHRC, Sydney.

Babacan, A., Singh, S., 2010. Migration, Belonging and the Nation State. Cambridge Scholars Publishing, Newcastle, Australia.

Banks, S., 2012. Global ethics for social work. In: Banks, S., Nohr, K. (Eds.), Practising Social Work Ethics Around the World: Cases and Commentaries. Routledge, London, pp. 1–31.

Benhabib, S., 2004. The Rights of Others: Aliens, Residents and Citizens. Cambridge University Press, Cambridge.

Bowles, R., 2005. Social work with refugee survivors of torture and trauma. In: Alston, M., McKinnon, J. (Eds.), Social Work: Fields of Practice, second ed. Oxford University Press, South Melbourne, pp. 249–267.

Brennan, F., 2003. Tampering with Asylum. University of Queensland Press, St. Lucia, Australia.

Briskman, L., Cemlyn, S., 2005. Reclaiming humanity for asylum seekers: a social work response. International Social Work 48 (6), 714–724.

Briskman, L., Latham, S., Goddard, C., 2008. Human Rights Overboard: Seeking Asylum in Australia. Scribe, Melbourne.

Briskman, L., Zion, D., Loff, B., 2010. Challenge and collusion: health professionals and immigration detention. International Journal of Human Rights 14 (7), 1092–1106.

Briskman, L., Zion, D., Loff, B., 2012. Care or collusion in asylum seeker detention. Ethics and Social Welfare 6 (1), 37–55.

Briskman, L., 2012. A clash of paradigms for asylum seekers: border security and human security. In: Bee Chen, G., Offord, B., Garbutt, R. (Eds.), Activating Human Rights and Peace: Theories, Practices and Contexts. Ashgate, Aldershot, pp. 270–286.

Burke, A., 2001. In Fear of Security: Australia's Invasion Anxiety. Pluto Press, Sydney.

Cemlyn, S., Briskman, L., 2003. Asylum, children's rights and social work. Child and Family Social Work 8 (3), 163–178.

Children Out of Detention (Chilout), 2014. Latest News. http://www.chilout.org/statsreports (accessed 04.02.14.).

Coddington, K., Catania, R.T., Lloyd, J., Mithcell-Eaton, E., Mountz, A., 2012. Embodied possibilities, sovereign geographies and island detention. Shima 6 (2), 27–48.

Collett, J., 2004. Immigration is a social work issue. In: Hayes, D., Humphries, B. (Eds.), Social Work, Immigration and Asylum: Debates, Dilemmas and Ethical Issues for Social Work and Social Care Practice. Jessica Kingsley Publishers, London.

Connolly, M., Harms, L. (Eds.), 2009. Social Work: Contexts and Practice, second ed. Oxford University Press, Melbourne.

Dastyari, A., Effeney, L., 2012. Immigration detention in guantanamo bay (not going anywhere anytime soon). Shima 6 (2), 49–65.

Edwards, A., Ferstman, C., 2010. Humanising non-citizens: convergence of human rights and human security. In: Edwards, A., Ferstman, C. (Eds.), Human Security and Non-citizens: Law, Policy and International Affairs. Cambridge University Press, Cambridge, pp. 3–46.

End Child Detention Now (ECDN), 21 October 2012. ECDN Addresses the Medsin Annual Conference at Warwick University. http://ecdn.org/ (accessed 29.11.12.).

Fine, M.D., 2007. A Caring Society? Care and the Dilemmas of Human Service in the 21st Century. Palgrave Macmillan, New York.

Fiske, L., 2012. Insider resistance: understanding refugee protest against immigration detention in Australia 1999–2005. Unpublished Ph.D. Thesis. Curtin University, Perth.

Flynn, M., Cannon, C., September 2009. The Privatization of Immigration Detention: Towards a Global View. A global detention working paper. http://www.globaldetentionproject.org/fileadmin/docs/GDP_PrivatizationPaper_Final5.pdf (accessed 29.12.12.).

Global Detention Project. About the Global Detention Project. http://idcoalition.org/aboutus/ (accessed 29.12.12.), n.d.

Goddard, C., Briskman, L., 19 February 2004. By Any Measure It's Official Child Abuse. The Herald Sun, p. 17.

Gosden, D., 2012. I Had to Do Something. I Couldn't Do Nothing!: Citizen Action in Support of Asylum Seekers in Australia 2001–2006. Unpublished Ph.D. thesis. University of New South Wales, Sydney.

Gready, P., 2007. Medical complicity in human rights abuses: a case study of district surgeons in apartheid South Africa. Journal of Human Rights 6 (4), 415–417.

Grewcock, M., 2009. Border Crimes: Australia's War on Illicit Immigration. Institute of Criminology Press, Sydney.

Hough, G., Briskman, L., 2003. Responding to the changing socio-political context of practice. In: Allan, J., Pease, B., Briskman, L. (Eds.), Critical Social Work: An Introduction to Theories and Practices. Allen and Unwin, Sydney, pp. 202–213.

Human Rights Watch, 2012. Canada: Vote No on Migrant Detention Bill. http://www.hrw.org/news/2012/03/16/canada-vote-no-migrant-detention-bill (accessed 21.11.12.).

Humphries, B., 2004. The construction and reconstruction of social work. In: Hayes, D., Humphries, B. (Eds.), Social Work, Immigration and Asylum: Debates, Dilemmas and Ethical Issues for Social Work and Social Care Practice. Jessica Kingsley Publishers, London, pp. 29–41.

IASSW and IFSW, 2004. Statement of Ethical Principles. http://ifsw.org/policies/statement-of-ethical-principles/ (accessed 19.12.12.).

International Detention Coalition, 7 December 2012a. A call on states to end the detention of children at the occasion of human rights day. Email Communication.

International Detention Coalition (IDC), 2012b. Global Campaign to End Immigration Detention of Children. http://idcoalition.org/children/ (accessed 07.02.14.).

Ife, J., 2001. Human Rights and Social Work: Towards Rights-based Practice. Cambridge University Press, Cambridge.

International Detention Coalition (IDC). About us: http://idcoalition.org/aboutus/ (accessed 29.12.12.), n.d.

Khan, P., Dominelli, L., 2000. The impact of globalization on social work in the UK. European Journal of Social Work 3 (2), 95–108.

Kundnani, A., 2007. The End of Tolerance: Racism in 21st Century Britain. Pluto Press, London.

Legrain, P., 2006. Immigrants: Your Country Needs Them. Little, Brown, London.

Lester, E., 2010. Socio-economic rights, human security and survival migrants: whose rights? Whose security? In: Edwards, A., Ferstman, C. (Eds.), Human Security and Non-citizens: Law, Policy and International Affairs. Cambridge University Press, Cambridge, pp. 314–356.

Loescher, G., Milner, J., 2009. Understanding the challenge. Forced Migration Review. Special Issue: Protracted Displacement 33, 9–11.

Lymberry, M., Butler, S., 2004. Social Work Ideals and Practice Realities. Palgrave Macmillan, Basingstoke.

Lyons, K., Manion, K., Carlsen, M., 2006. International Perspectives on Social Work: Global Conditions and Local Practice. Palgrave Macmillan, Basingstoke.

Macedo, D., Bartolome, L.I., 2001. Dancing with Bigotry: Beyond the Politics of Tolerance. Palgrave, New York.

McVeigh, T., 20 May 2012. Child Asylum Seekers Still Being Imprisoned by Immigration Service. The Guardian. http://www.guardian.co.uk/uk/2012/may/20/child-asylum-seekers-still-being-locked-up (accessed 29.12.12.).

Mountz, A., 2010. Seeking Asylum: Human Smuggling and Bureaucracy at the Border. University of Minnesota Press, Minneapolis.

Mountz, A., Briskman, L., 2012. Introducing island detentions. Shima 6 (2), 21–26.

Mullaly, B., 2002. Challenging Oppression: A Critical Approach to Social Work Practice. Oxford University Press, Ontario.

National Association of Scholars, 2007. The Scandal of Social Work Education. www.nas.org/articles/The_Scandal_of_Social_Work_Education (accessed 20.12.12.).

News, C.B.C., 18 December 2013. Treatment of Migrants on Lampedusa May See European Union Action against Italy. http://www.cbc.ca/news/world/treatment-of-

migrants-on-lampedusa-may-see-eu-action-against-italy-1.2468759 (accessed 07.02.14.).
No One is Illegal (NOI) – Toronto. About No One is Illegal – Toronto. http://toronto.nooneisillegal.org/about (accessed 29.11.12.), n.d.
Noble, C., Henrickson, M., Han, I.Y., 2009. Many voices, many truths, social justice for all. In: Noble, C., Henrickson, M., Han, I.Y. (Eds.), Social Work Education: Voices from the Asia Pacific. The Vulgar Press, Melbourne, pp. 413–415.
Refugee Council of Australia (RCOA). Myths about Refugees and Asylum Seekers. http://www.refugeecouncil.org.au/f/myth-long.php (accessed 11.12.12.), n.d.
Refugee Council of Australia (RCOA), 2010. Global Refugee Statistics. http://www.refugeecouncil.org.au/r/stat-int.php (accessed 25.12.12.).
Sivandan, A., 2007. Foreword. In: Kundnani, A. (Ed.), The End of Tolerance: Racism in 21st Century Britain. Pluto Press, London.
Stratton, J., McCann, S., 2002. Staring into the abyss: confronting the absence of decency in Australian refugee law and policy development. Mots Pluriels 21. http://motspluriels.arts.uwa.edu.au/MP2102jssmc.html (accessed 29.12.12.).
United Nations, 1948. Universal Declaration of Human Rights. http://www.un.org/en/documents/udhr/index.shtml (accessed 29.12.12.).
United Nations High Commissioner for Refugees (UNHCR), 1951. Convention and Protocol Relating to the Status of Refugees. http://unhcr.org.au/unhcr/images/convention%20and%20protocol.pdf (accessed 25.11.12.).
United Nations High Commissioner for Refugees (UNHCR), 2007. UNHCR Statistical Online Population Database: Sources, Methods and Data Consideration. http://www.unhcr.org/statistics/STATISTICS/45c06c662.html (accessed 25.11.12.).
United Nations High Commissioner for Refugees (UNHCR), 2010. UNHCR Global Trends 2010. http://www.unhcr.org/4dfa11499.html (accessed 18.12.12.).
United Nations High Commissioner for Refugees (UNHCR), 2012. Detention Guidelines: Guidelines on the Applicable Criteria and Standards Relating to the Detention of Asylum-seekers and Alternatives to Detention. http://www.unhcr.org/505b10ee9.html (accessed 27.12.12.).
Weber, L., Pickering, S., 2011. Globalization and Borders: Deaths at the Global Frontier. Palgrave Macmillan, Basingstoke.
Wills, S., 2009. Between the hostel and the detention centre: possible trajectories of migrant pain and shame in Australia. In: Logan, W., Reeves, K. (Eds.), Places of Shame: Dealing with Difficult Heritage. Routledge, Oxon.
Wilsher, D., 2012. Immigration Detention: Law, History and Politics. Cambridge University Press, Cambridge.
Winterdyk, J.A., Sundberg, K.W. (Eds.), 2009. Border Security in the Al-Quaeda Era. CRC Press, Boca Raton.
Zion, D., Briskman, L., Loff, B., 2009. Nursing in asylum seeker detention in Australia: care, rights and witnessing. Journal of Medical Ethics 35 (9), 547–551.

Relevant Websites

http://www.globaldetentionproject.org/ – Global Detention Project.
http://idcoalition.org/ – International Detention Coalition.
http://www.refugeecouncil.org.au/f/who-conv.php – Refugee Convention.
http://unhcr.org.au/unhcr/ – United Nations High Commissioner for Refugees (UNHCR).

Asymmetry of Body and Brain

Frank R Boutsen, College of Allied Health, Health Sciences Center, University of Oklahoma, Oklahoma City, OK, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Advances in computational analysis, imaging, and genetics along with increasingly sophisticated research paradigms have brought about an across-the-life-span, if not also an evolutionary, perspective in body and brain laterality research. In addition, a simple or naive structural localizationist approach to asymmetry in brain and body has given way to a developmental and physiological perspective. Evidence shows that asymmetries of body and brain are established early and share basic molecular pathways that are evolutionarily well conserved. That said, some genes show altered (asymmetric) expression levels in human brains and endow the brain unlike the viscera with lifelong adaptability.

Introduction

Although the basic body plan of vertebrates is essentially bilateral, all vertebrates lateralize unpaired organs of the chest and abdomen during embryogenesis. In humans this process leads to the characteristic positions of the heart apex, stomach, and spleen to the left of the midline but also to organ asymmetries as in the lung and heart with the primitive gut coiling itself in a counterclockwise direction (Kosaki et al., 1999; Cohen, 2001). Likewise, the brain while essentially symmetric also has structural asymmetries. Additionally, certain motor and cognitive functions appear to be hosted by one hemisphere more so than the other, promoting the idea of hemispheric specialization. Most basically limb motor control emanates from the contralateral hemisphere and thus is asymmetric from a central nervous system (CNS) axis point of view. This arrangement and other asymmetries in the nervous system are certainly not uniquely human (Corballis, 2014). However, as Geschwind and Miller (2001) and more recently Hervé et al. (2013) point out, the colateralization of hand preference and language dominance is quite unique and also an indicator that genetic underpinnings of body and brain asymmetry are different.

In humans, a link between cilia and L/R determination has been suspected since the discovery of Kartagener syndrome. Kartagener syndrome is an autosomal recessive genetic ciliopathy, also referred to as primary ciliary dyskinisia. In Kartagener patients, dynein arms (motor molecules) are absent from the respiratory cilia and sperm flagella. Leigh et al. (2009) reported that the prevalence of situs inversus totalis tends to be over 60% in the pediatric population, as opposed to 50% in the adult population. In addition, patients with Kartagener syndrome have a greater incidence of congenital cardiovascular defects, while a minority present with heterotaxy (heterotaxy according to Casey (1998) refers to any right/left abnormality including situs ambiguous and situs inversus). Bataille et al. (2011) described three patients with autosomal dominant cystic kidney disease, one presenting with dextro cardia and situs inversis totalis in the other two. Each harbored mutation in the PKD2 gene (vide infra).

More recently, Kosaki et al. (1999) described two Lefty gene mutations in patients with left-right anomalies that are homologues of Lefty-1 and Lefty-2 in the mouse model (vide infra). Gradient-wise expression of human Lefty may well be established by retinoic acid (RA) to which it binds. The patients' phenotype included cardiac abnormalities (e.g., left ventricle hypoplasia) and left pulmonary isomerism. Overall, these findings of heterotaxia point to a possible role of molecular signaling in heterotaxia.

Studies of conjoint twins have been most revealing as to the stage and possible embryonic mechanisms that are critical for normal body asymmetry. This is because conjoined twins are thought to originate from two adjacent embryonic streaks and tend to show situs abnormalities (Levin, 2001). This correlation, thus, invokes events surrounding the formation of the primitive streak as playing a critical role in the establishment of body symmetry.

Interestingly, according to Levin (2004, 2005) there is converging evidence that visceral and brain laterality are not correlated, even though they are statistically associated. He refers to studies by Kennedy et al. (1999) and McManus et al. (2004) showing that in right-handed individuals with situs inversus totalis left language lateralization occurs with the same frequency (95%) as it does in normal situs right-handed individuals while the incidence of left-handedness is also comparable to that in the rest of the population. Levin (2004) further notes that occurrence of unilateral anomalies on opposite sides of the head ("bookending" phenomena such as cleft lip, unilateral eye and ear defects) of otherwise healthy twins is not observed in the body.

Twin studies also point to difficulties of a purely genetic interpretation of laterality of brain function, including handedness (Annett, 2002; Klar, 2003). Klar (2003) suggests that lateralization of structure and function in the brain while certainly influenced by genetics is best interpreted within a developmental if not also a functional and even cultural perspective. On the other hand, familial occurrences of mirror movements (Peng and Charron, 2013), apraxia of speech, (Lai et al., 2001; Turner et al., 2013) and dyslexia (Galaburda et al., 2006; Graham and Fisher, 2013) indicate a molecular basis for lateralized limb motor control and left-lateralized speech and language if not also for hemispheric specialization in general.

MM are unwanted ipsilateral limb movements that co-occur with intended contralateral movements. While affecting mostly movements in the upper limbs, they can also interfere with or alter asynchronous (alternating) movements as in walking (Peng and Charron, 2013; Depienne et al., 2012). They result

from abnormal (re-)crossing of the corticospinal pathway at the level of the corpus callosum (CC), the medulla, or the spinal cord (Peng and Charron, 2013). A genetic basis for MM is evidenced in X-linked Kallmann syndrome. First described in 1944, this syndrome is characterized by hypogonadism and anosmia/hyposmia, presenting with MM (Peng and Charron, 2013). Dyslexia also has long been suspected to have a genetic underpinning as documentation of familial inheritance of this disorder goes back to the end of the nineteenth century (Galaburda et al., 2006). Not until recently, however, have MRI studies unveiled the incidence of periventricular nodular heterotopia and peri-Sylvian polymicrogyria in two groups of children with dyslexia suggesting a molecular disorder during embryogenesis (Galaburda et al., 2006). Perhaps the greatest advance in pinpointing the molecular basis for hemispheric specialization has been the discovery of the FOXP2 gene. Mutations of this gene were first identified in the KE family (Lai et al., 2001) and later in individuals with a small mutation having been discovered in an Australian family (Turner et al., 2013). In these families children present with childhood apraxia of speech, language, and literacy difficulties.

Intriguingly, Geschwind and Miller (2001) draw attention to the notion that understanding of the pathways involved in axis patterning and L-R establishment in particular may well advance understanding of degenerative brain disorders in the adult population, especially frontotemporal dementia (FTD). This is because FTD has been documented to involve asymmetric neurodegeneration that is familial, either presenting with primary progressive aphasia if left-sided or behavioral symptoms if right-sided (Geschwind and Miller, 2001).

Advances made in the above genetic lines of inquiry have been paralleled by improved imaging (e.g., Diffusion Tensor Imaging (DTI); Infrared Spectroscopy) and computational analysis techniques (e.g., Dynamic Component Modeling, Tractography) within the last decade (Hervé et al., 2013). As a result, an across-the-life-span perspective, starting *in utero*, is emerging showing an evolving relation between asymmetry in brain function and structure. Since much of the discovery regarding the asymmetry of body and brain pertains to human embryology that readers may not be familiar with early development is summarized based upon recent in-depth reviews provided by Sadler (2010) and Tan et al. (2013). The developmental time line (in days week^{-1}) is adopted from Sadler (2010).

Human Embryology: A Brief Review

The first 4 weeks of development, from fertilization until the end of gastrulation, is also known as blastogenesis. Blastogenesis establishes the first developmental field which comprises the entire embryo. The embryo proper initially is comprised of a cup-shaped single layer of primitive ectoderm, the epiblast, which is subsequently converted into the three definitive germ layers during gastrulation which include definitive ecto-, meso-, and endoderm. The primary field is then subdivided into progenitor fields, which give rise to primordia of all morphological structures. Organogenesis begins at the end of blastogenesis and is characterized by morphogenesis (formation of organs and other body parts) and histogenesis (differentiation of cells and tissues). It can be seen from this that the prospective adult body axes are fully laid out by the end of the blastogenesis while packaging (rotation of organs) continues during organogenesis. Brain development and laterality are established during differentiation; however, its development is also dependent on function.

Axis Formation and Positioning of Gut, Heart, and Central Nervous System

The axial blueprint for the body and brain is laid out very early during development. Upon sperm entry the fertilized zygote is corticalized. Mitotic division and subsequent cleavage produce by day 3 a spherical mass (early morula) that looks like a blackberry with isomorphic cells (blastomeres) that with compaction, flatten at the periphery (late morula, day 4) to become the first embryonic epithelium, the trophectoderm (later chorion and placenta).

The cells inside the compacted morula become the inner cell mass (ICM), and are totipotent. The dorsoventral axis is defined at the time of the blastocyst (**Figure 1**).

The future dorsal part of the embryo, contained in the ICM, is located at the contact of the blastocoele. In the next few days the blastocyst is implanted and the ICM evolves to become embryoblast (**Figure 1**). In the second week the embryoblast, through cell polarizaton, differentiates into a hypoblast layer (cuboid) and epiblast (columnar). As the amnion and primitive yolk develop dorsally and ventrally, respectively, extraembryonic mesoderm on the inner aspect of the trophoblast surrounds both. By day 12 the embryo reemerges on the endometrial surface. Isolated coelomic spaces appear in extraembryonic mesoderm and consequently fuse to form the extraembryonic coelom (chorion) that is limited by the chorionic plate (chorion is the embryonic component of placenta). This fluid-filled cavity surrounds the amnion and umbilical vesicle, except where they are attached to the chorionic plate by the connecting stalk. As the extraembryonic coelom forms, a second wave of hypoblast migration moves the primary umbilical vesicle away from the embryoblast to be pinched off for the most part. In its wake it is replaced by a secondary umbilical vesicle (the definitive yolk sac).

At the lateral aspects of the disc, extraembryonic coelom splits the extraembryonic mesoderm into extraembryonic somatic and splanchnic mesoderm that will cover the amnion (dorsal) and secondary umbilical vesicle (ventral), respectively. Somato mesoderm gives rise to somatopleure which is precursor to dermis and bone; splanchnic endoderm develops into splanchnopleure that will give rise to heart and blood vessels (vide infra). In the third week, during a process called gastrulation (**Figure 1**), ecto-, meso-, and endoderm are derived from the epiblast so that a trilaminar disc develops. The 14-day embryo still has the form of a flat bilaminar embryonic disc but the epiblastic cells in a localized area are now columnar and form a thickened circular area, the prechordal plate, which is an important organizer of the head region. Around day 16 the primitive streak forms as a groove in the midsagittal plane in the caudal half of the bilaminar disc (**Figure 1**). As the streak elongates, its cranial end thickens to

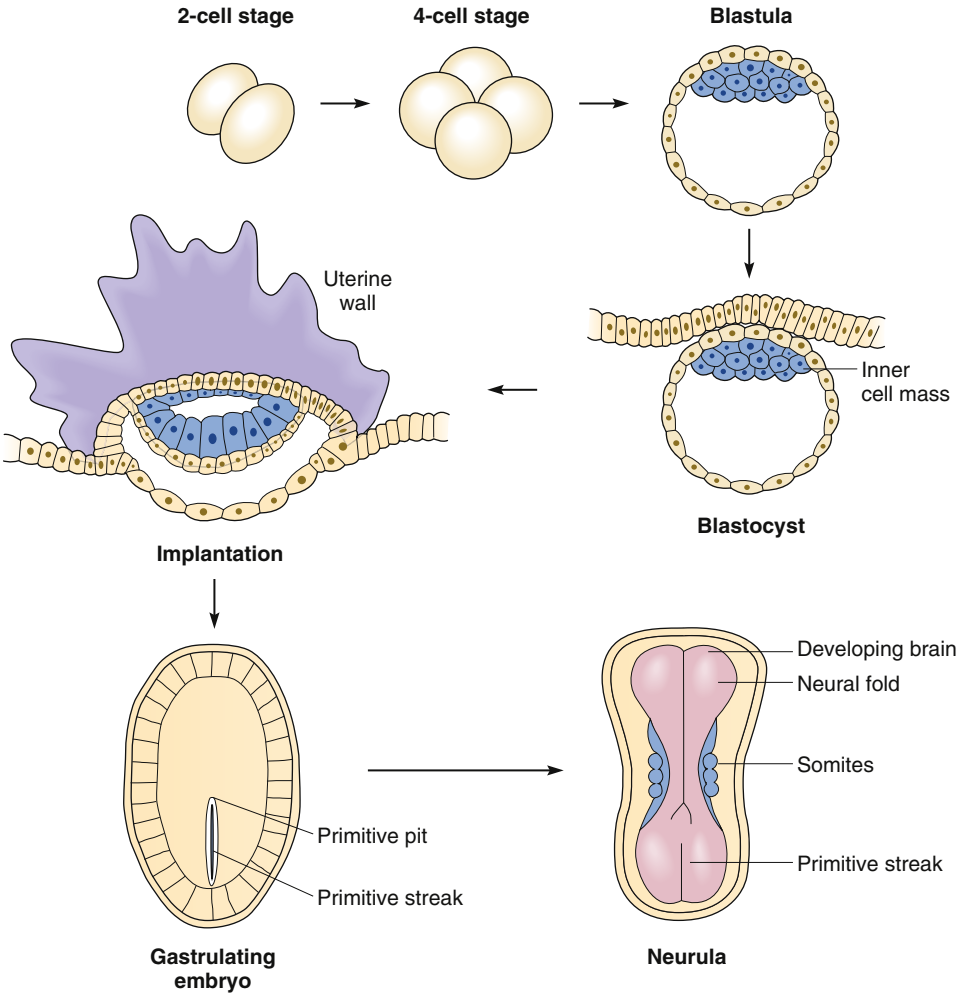

Figure 1 Early division, blastocyst, implantation, gastrulation, and neurulation stages of the developing embryo.

form a primitive knot or node that surrounds a depression termed the primitive pit (Figure 1). Near the primitive streak, epiblast cells delaminate as they undergo epithelial to mesenchym transformation (EMT) and ingress to displace hypoblast to form definitive endoderm. Definitive endoderm is a precursor to the gut tube and later fore-, mid-, and hindgut. In addition, epiblast cells migrate between endoderm and epiblast layers to form intraembryonic mesoderm except in two regions where the oropharyngeal and cloacal membrane form. Cells migrating from the cranial end of the node become the prechordal plate and notochord. Prechordal plate is a progenitor to cardiac mesoderm that forms in a horseshoe manner cranially to the oropharyngeal membrane. While over the course of this time the primitive streak has been shrinking to ultimately disappear, cephalocaudally the notochord makes its appearance as a tunnel under the dorsal surface of the trilaminar embryo. The notochord is located in the rostral half of the presomite embryo and underlies the midline of the neural plate (head plate) that at day 19 is clearly visible. On day 20 neural folds form cranially and a neural groove is visible centrally that is flanked by the first pair of somites in the occipital region of the embryo (Figure 1). Somitization proceeds cephalocaudally and segments paraxial mesoderm along its anterior-posterior (AP) axis. The first somitomeres develop in the head region in association with segmentation of the neural plate into neuritomeres. Cephalad expansion of the developing neural tube along with the endoderm below it results in folding at the cranial end of the embryo's AP axis so that both initially extend over the future pericardial cavity. Ventrally the endoderm is continuous with the yolk sac that is suspended from it in a teardrop shape. Folding of endoderm (superior to the yolk sac) forms the gut tube, which is a progenitor of the gut. Meanwhile, the neural fold expands anteriorly and curves underneath the anterior aspect of arched endoderm leaving space for the developing heart tube and liver that are situated anterior to the superior convergence (neck) of the yolk sac. Thus, at this point aspects of the adult AP axis (in supine position) are established as the brain primordia are in front of those the heart which are in turn in front of the midgut. In addition, aspects of the adult dorsoventral axis are also established where the neuraxis is dorsal to the heart (tube) and gut (tube). As the neural folds continue to elevate they approximate at the

midline dorsally and finally fuse around day 22 in the center of the AP axis leaving unfused areas (neuropores) cranially and caudally. At this point segmentation of the paraxial mesoderm is well underway as seven somites are visible alongside the developing neural tube. In addition, at day 23 the heart tube proliferates causing a bend halfway along its craniocaudal axis and also rightward twisting to form the atrium cranially and the ventricle caudally. Final closure of the cranial neuropore occurs around day 25 at the 18–20 somite stage. At this point neural crest cells (Figure 2) differentiate from neuroectoderm above the neural tube and induce the first and second pharyngeal arches ventrally in head mesenchyme. They are positioned anteriorly to the heart tube that completes rightward cardiac looping around day 28. Around day 28 the caudal neuropore closes as well at which time 25 somites have formed. At this point blastogenesis is complete and the embryo is formed along major body axes. The neural tube is formed between right and left chains of somites (Figure 2). It has been derived from neuroectoderm. Pharyngeal (branchial) arches are visible. The embryo is now slightly curved because of the head and tail folds. The heart produces a large ventral prominence and pumps blood. The forebrain produces a prominent elevation of the head, and folding of the embryo has given the embryo its characteristic C-shaped curvature. A long, curved tail is present. Upper limb buds become recognizable by day 26 or 27 as small swellings on the ventrolateral body walls.

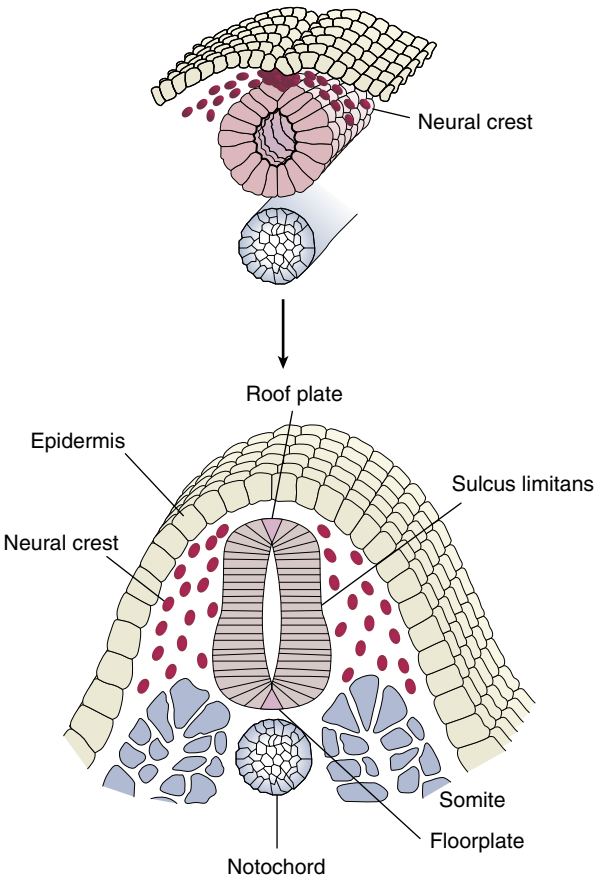

Figure 2 Neural crest cell develop upon closure of the neural tube. The neural tube expresses a roof plate, floor plate, and the sulcus limitans along its length.

Central Nervous System Development: Anterior-Posterior Partitioning

In the 5th week the neural tube expands to form 3 vesicles that pattern its AP axis into fore-, mid-, and hindbrain and future spinal cord. The dorsoventral axis of the lumen of the neural tube is expressed by the appearance of the sulcus limitans (Figure 2). The human cerebral hemispheres originate at about 6 weeks in the embryo, growing from the midline at the rostral end of the neural plate in the morphologically symmetric prosencephalic vesicle. At 7 weeks the prosencephalon divides into diencephalon and telencephalon while the hindbrain divides in metencephalon and myelencephalon. At 8 weeks the two hemispheres become clearly delineated by an interhemispheric fissure, which divides the two cerebral hemispheres.

Cortical Specification

The cortex is formed at the rostral-most portion of the neural tube and is quickly subdivided in the left and right telencephalic vesicles. Cortical neurons are generated from dividing progenitor cells in the ventricular zone (VZ) layer (Figure 3; Buckner and Krienen, 2013; Monuki and Walsh, 2001).

Guided by neuroglia (Figure 3) most postmitotic neurons migrate radially and enter the cortex which will develop from the inside out as is shown in Figure 4: layer I will develop before layer II and so on (Cooper, 2008).

Limited expansion medially between the telencephalic vesicle creates the dorsal midline roof plate while medially in the ventral region of the vesicles continued expansion produces the ganglionic eminence (the future basal ganglia; see Figure 5) that will be an important source of GABA neurons.

Meanwhile, the cortex becomes exposed to several potential sources of secreted signaling molecules or organizers (Buckner and Krienen, 2013; Monuki and Walsh, 2001). They include in order of their development, the anterior neural ridge (ANR), the roof plate, the cortical hem, and other sources of signaling molecules, such as the surface ectoderm (the future skin) and mesenchymal elements between the skin and brain, including the future meningeal cells that surround the adult brain (vide infra). After its initial induction by extrinsic signals, the telencephalon becomes specified into dorsal (pallial) and ventral (subpallial) regions. Following dorsal telencephalic specification, the dorsal telencephalon must be further subdivided into dorsal midline epithelial fates (choroid plexus epithelium and cortical hem) and the cortical VZ. Once the cortical VZ fate is selected, regional specification of the cortical VZ occurs (Buckner and Krienen, 2013; Monuki and Walsh, 2001).

The CC develops relatively late, i.e., around the 12th week (Sadler, 2010). Choe et al. (2012) note that CC development awaits midline glial specializations and guidance cues from

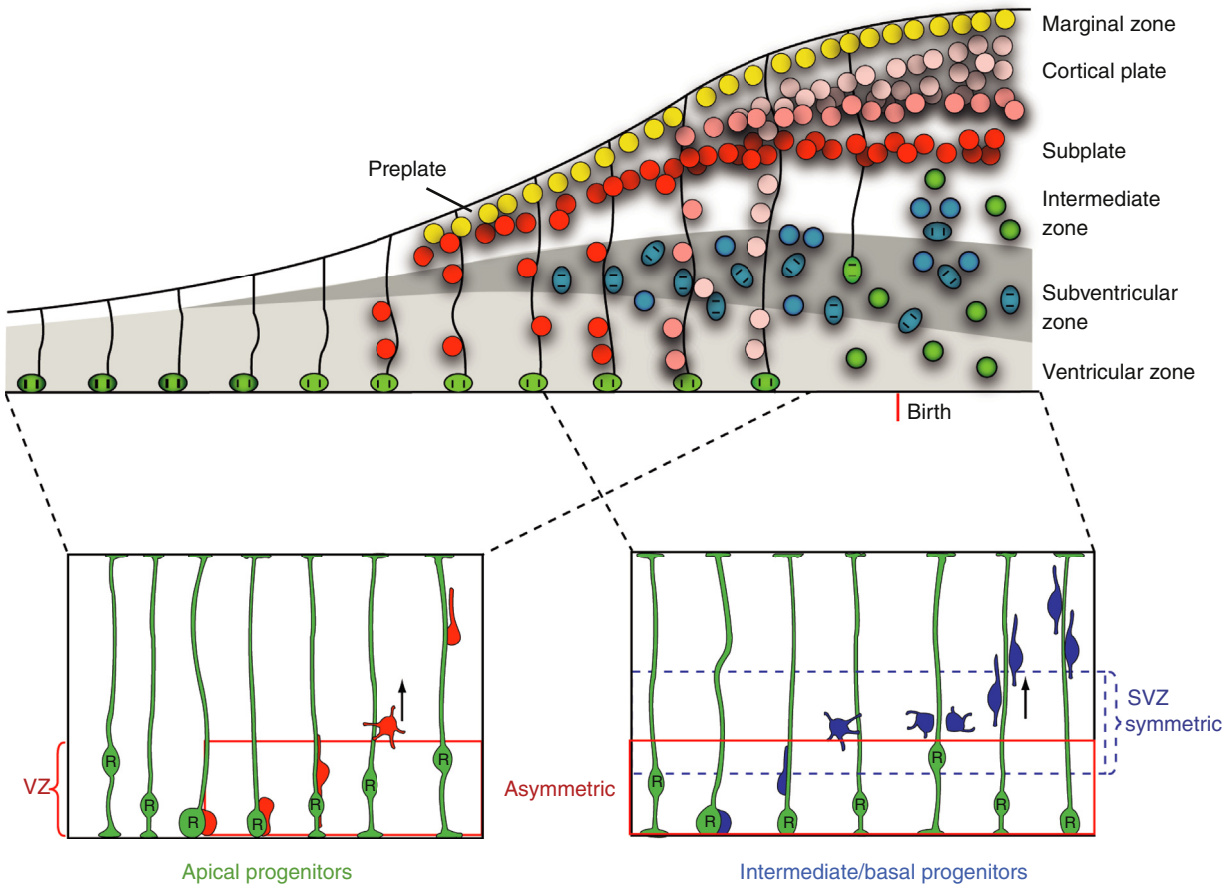

Figure 3 Histogenesis in the cerebral cortex proceeds through three stages. In the first stage of histogenesis, the wall of the cerebral cortex is made up of the progenitor cells, which occupy the VZ. In the next stage of development, the first neurons exit the cell cycle (red) and accumulate in the preplate, adjacent to the pial surface. The neurons of the preplate can be divided into the more superficial Cajal-Retzius cells and the subplate cells. In the next stage of cortical histogenesis, newly generated neurons (red) migrate along radial glial fibers to form a layer between the Cajal-Retzius cells and the subplate. This layer is called the cortical plate, and the majority of the neurons in the cerebral cortex accumulate in this layer. Modified from Noctor, S.C., Martinez-Cerdeno, V., Ivic, L., Kriegstein, A.R., 2004. Cortical neurons arise in symmetric and asymmetric division zones and migrate through specific phases. Nature Neuroscience 7, 136–144; Noctor, S.C., Martlnez-Cerdefio, V., Kriegstein, A.R., 2008. Distinct behaviors of neural stem and progenitor cells underlie cortical neurogenesis. Journal of Comparative Neurology 508, 28–44.

the septum underlying the callosum that do not normally develop until just before the callosal pathfinding axons project across the midline. These cues regulate the

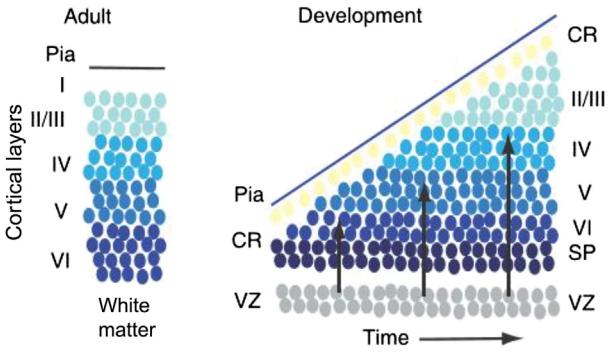

Figure 4 In normal cortical development the early generated neurons bypass the previously generated neurons to produce the inside-out developmental ordering to the lamination.

connectivity of the two hemispheres, as well as the formation of other ascending and descending tracts. They further mention that later-born neurons (layers 2–4) may yet be another source of delay for CC formation as they, along with layer V neurons, are primary sources of callosal axons. Interestingly, Buckner and Krienen (2013) cite evidence from the Allen Brain Institute that genes with species-specific expression patterns in mouse and humans are expressed asymmetrically in cortex. That is, according to these data, mouse brain showed a preponderance of markers in layer V (projection neurons), whereas half these markers showed either no or diminished expression in humans. On the other hand, in humans more of these genes were expressed in layer III (cortico-cortical projection neurons). According to Buckner and Krienen (2013), this finding supports the notion that these layer-III pyramidal neurons with long-range intracortical projections could well be important for the patterning of association cortex (see also role of FOXP2 below). Finally, as it pertains to the development of cortical folds, Geschwind and Miller (2001)

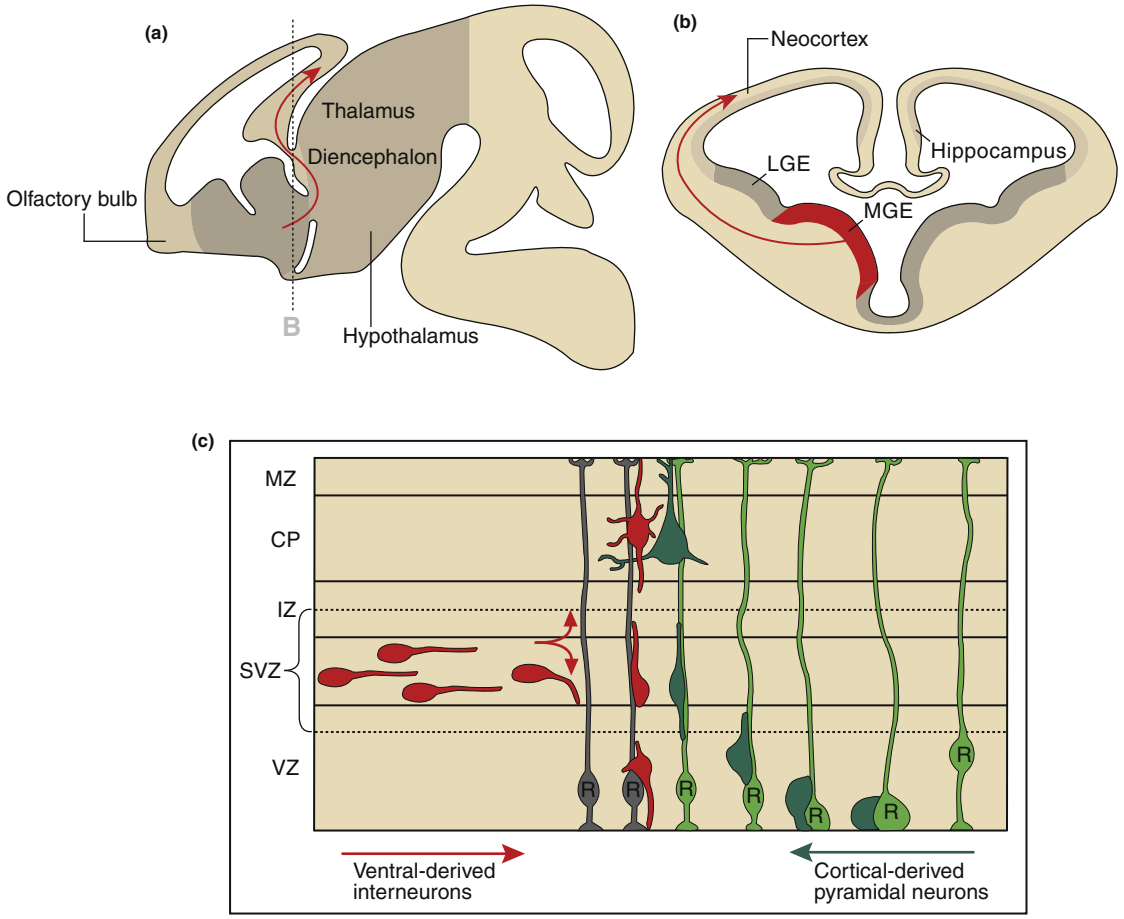

Figure 5 The neurons of the cerebral cortex derive from both intrinsic and extrinsic sources. Most of the neurons in the cortex are derived from the VZ cells immediately below their adult location. (a–c) This figure shows the paths taken by the neurons from the ganglionic eminence in red at low (a,b) and high magnification (c) and the neurons generated within the cortex in green. Modified from Kriegstein, A.R., Noctor, S.C., July 2004. Patterns of neuronal migration in the embryonic cortex. Trends in Neurosciences 27 (7), 392–399.

report that at 26 weeks of gestation, the superior temporal sulcus (STS) is easily distinguished, followed by the angular and supramarginal gyri of the parietal lobe at 28 weeks of gestation. The transverse temporal gyrus appears at 31 week (Geschwind and Miller, 2001).

Asymmetry

Body

The tissue movements initiated during gastrulation eventually establish the distinct left-right orientation of the developing viscera. According to Constam and Robertson (2000) L-R asymmetry first becomes visible after the formation of 8–10 somite pairs when the primitive heart tube loops to the right side. Shortly thereafter the entire embryo turns rightward 180° while concurrently ventral folding of the gut endoderm and embryonic body wall fold results in ventral closure. Constam and Robertson (2000) note that the coordinated timing of axial rotation and ventral closure suggests that they may be regulated by the same molecular signals, i.e., the asymmetric expression of several genes, including the TGFβ family members Lefty-1, Lefty-2, and Nodal (see below) that precedes both processes. More recently, Yamaguchi (2001) concluded that Wnt and Bmp (Wingless first identified in Drosophilia and bone morphogen proteins) inhibitory signals from the midline anterior neural tube and notochord ensure that the heart-forming region remains lateral by inhibiting cardiogenesis in the adjacent paraxial mesoderm (vide infra).

Brain Asymmetries

The earliest sign of anatomical asymmetry so far identified in human fetuses, shows the choroid plexus to be larger on the left than on the right in approximately two-thirds of fetuses in the 11th to 13th week of gestation and this asymmetry may well underlie asymmetry of the planum temporale (Corballis, 2013).

Geschwind and Miller (2001) report evidence that planum temporale and peri-Sylvian language asymmetry areas are first visible around 30 weeks gestation in most human fetuses. These peri-Sylvian structures are visible earlier on the right than on the left by about 1 week.

Likewise, studies by Gilles (1983) and Jammes and Gilles (1983) report that the transverse temporal gyrus, which

appears at 31 week, is observed 1 ± 2 weeks earlier on the right in most fetuses.

Hervé et al. (2013) in a review of imaging evidence and hemispheric specialization report that multiple imaging studies *in utero* or in preterm newborns have reported a larger planum temporale and deeper right STS. Additionally, these studies reveal that sulcal pits (deepest and also earliest parts of cortical sulcal fundi) are more frequent and less variable in the left than the right STS extremities (Hervé et al., 2013). Consistent with these findings are data from Li et al. (2013), who investigated cortical hemispheric asymmetry from birth to 2 years of age. They found that the cortical asymmetries around the peri-Sylvian region and STS evolve modestly from birth to 2 years of age. At birth, males showed larger asymmetries than females and this difference persisted from birth to 2 years of age. Specific asymmetries shown included the left supramarginal gyrus (SMG) being significantly posterior to the right SMG (10.2 mm for males and 6.9 mm for females at birth growing to 12.0 and 8.4 mm by 2 years of age). In addition, they found that the right STS and parieto-occipital sulcus are significantly larger and deeper than their counterparts in the left hemisphere while the left planum temporale is significantly larger and deeper than that in the right hemisphere.

As for white matter, Dubois et al. (2009) conducted a DTI study in 1- to 4-month-old newborns. For each infant the arcuate fasciculus and corticospinal tract were reconstructed using tractography. Whole brain voxel-based and region-specific fractional anisotropy (FA) analyses revealed greater FA in the left compared to the right arcuate fasciculus as well as in the lower aspect of the corticospinal tract, i.e., the segment between the insula and the cerebral peduncle. These results are consistent with DTI findings by Buchel et al. (2004) in adults who found left greater than right FA white matter asymmetries in the arcuate fasciculus along with asymmetries in tracts contralateral to the dominant hand (i.e., higher on the left in right-handers) innervating the precentral gyrus.

In terms of a functional or hemispheric dominance perspective, Corballis (2013) cites evidence from ultrasound recordings that fetuses as early as 11 week of gestation show more movement of the right than of the left arm, and from 15 weeks some 90% of fetuses suck the right rather than the left thumb, an asymmetry that according to Hepper et al. (2005) correlates with handedness at 10–12 years of age. Lin et al. (2013) employed frequency-domain near-infrared spectroscopy to measure hemodynamic and metabolic parameters in newborns and found that they were greater in the temporal and parietal regions than in the frontal region and that these differences increased with age. In addition, most parameters were significantly greater in the right hemisphere than in the left. Compared to age-matched females, males showed greater left than right asymmetry (Lin et al., 2013). On the other hand, Hervé et al. (2011) only found modest T2 signal asymmetry in CST associated with handedness on DTI with 7-T (Tesla is a unit of measurement quantifying the strength of a magnetic field, 7 being ultra-high-field) and were unable to corroborate their earlier finding (Hervé et al., 2009) of greater T1 asymmetry in CST.

Overall, as Li et al. (2013) point out early brain development data indicate sequential development of asymmetry giving primacy to right over left development which is in agreement with the "Right-hemisphere conservatism" theory proposed by Geschwind and Galaburda (1985). Along this line they speculate that early (*mine*) right hemisphere development is protective, since it shields this hemisphere from external influences while the delay in maturation of the left hemisphere may allow for a higher plasticity with functional stimulation, such as language exposure and motor activity (perhaps at the cost of exposure). The delayed expression in the left hemisphere may also be responsible for the left lateralization of sensorimotor control or praxis (Lin et al., 2013). As this takes place there may be subcortical white matter "spurts" that may be greater in the left than in the right giving rise to greater connectivity (vide infra).

Molecular Basis: Signaling Pathways and Gradients

The molecular basis for transformation and patterning of existing into new tissue involves distant and gradient signaling to accomplish spatiotemporal timing of processes often lasting as short as a few hours (Field et al., 2011; Yang, 2012). While there appears to be a redundancy of pathways and genes that can participate in this space-time clock, much of its integration if not coordination seems to be at the level of SMAD receptors that are in much more limited supply (SMAD is a portmanteau for MAD (mothers against decapentaplegic, in Drosophilia (Heldin et al., 1997) and SMA for small body size in *Caenorhabditis elegans* (Savage et al., 1996)) introducing an element of competition as it were among genes/pathways (Inui et al., 2012; De Robertis and Kuroda, 2004). Most basically, molecular signaling affects transformation in existing tissue by undoing and altering the existing cell arrangements (delamination, EMT) in a directional (polarization by Planar Cell Polarity (PCP)) and hierarchical and patterned manner. At the cellular level this involves developmental processes including establishment of polarity and changes in cytoskeleton (microtubules) and (symmetric and asymmetric) division affecting expansion while intercellular processes such as gene transcription or repression instruct processes including cell adhesion (as in compacting or loosening) and guidance (chemo-attraction/repellents, affinity to glial guidance). At the same time molecular signaling informs the timing of cell fates ranging from toti/pluripotent (as in morula or later in VZ) to committed. Regional (patterned) development depends not only on distant intercellular signaling for the dispersion of genes but also on the establishment of temporary organizers. Examples of the latter include the primitive streak, the primitive node (also termed Hensen's node and Spemann-Mangold organizer, see Niehrs (2004) for review) in mesoderm, the head organizer, the notochord, the neural plate, the ANR, cortical hem (and antihem), and glial cells.

Intercellular signaling results in the activation of intracellular proteins and factors through a well-defined cascade that eventually initiates or represses the transcription of specific genes. Based on their shared molecular structures four

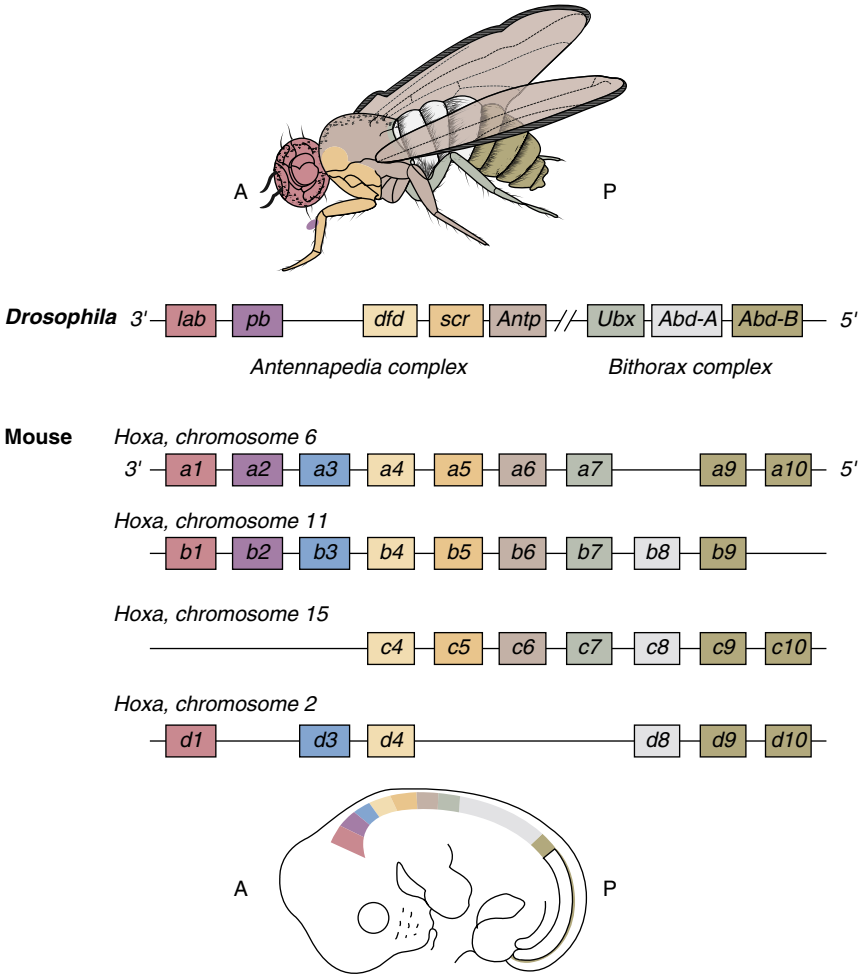

Figure 6 Hox genes are important in the anterior-posterior patterning of the embryo.

signaling families are distinguished (Tan et al., 2013). They include the Hedgehog (Sonic Hedgehog or Shh), Wnt (canonical and noncanical), Fibroblast Growth Factor (FGF), and transforming growth factor-β families (TGFβ). Important targets of this signaling include Homeobox or HOX genes (**Figure 6**) and BMPs. As can be seen in **Figure 6** Hox genes are important in antero-posterior patterning of the embryo.

Processes Laying the Ground Work for Asymmetry

Planar cell polarization is evidenced by uniform asymmetrical localization of key polarity protein and uniform asymmetrical cellular behaviors such as the localization and orientation of cell extensions like cilia. Examples of PCP-regulated processes are uniform orientations of drosophila hair, mouse fur, inner ear sensory hair cells, elongating antero-posterior body axis and closing neural tubes (Yang, 2012). In a context-dependent manner, the directional information provided by PCP is essential for cells to decide in which direction to divide, to move, or to interact with each other (Yang, 2012).

Like the PCP pathway, Nodal, a member of the TGFβ family, partakes in very early signaling in mouse, Nodal is expressed throughout the epiblast and induces formation and anterior movement of the anterior visceral endoderm (AVE) 1 day before gastrulation (Yamamoto et al., 2004; Tan et al., 2013). In doing so, it plays a crucial role in establishing the proximal distal axis (which is a primordial AP axis). Yamamoto et al. (2004) cite evidence that the AVE secretes signaling molecules, including the Nodal antagonists Lefty-1 and Cerl both of which will inhibit Nodal signaling in the underlying epiblast and restrict Nodal activity to the posterior side of the embryo. Such asymmetry in Nodal signaling specifies that the anterior epiblast will form the prospective brain and generates the primitive streak on the posterior side (Yamamoto et al., 2004). It can be seen from this that reductions of Nodal pathway activity albeit indirectly can result in phenotypes resembling human holoprosencephaly.

The Wnt pathway, originally described as the wingless pathway in Drosophila, is involved in axis formation and regionalization of the vertebrate body and nervous system (**Figure 7**) (Geschwind and Miller, 2001). As a morphogen it can form a concentration gradient across a field of hundreds of cells which leads to the generation of different cell types in a distinct spatial order (Yamaguchi, 2001; Yang, 2012). While

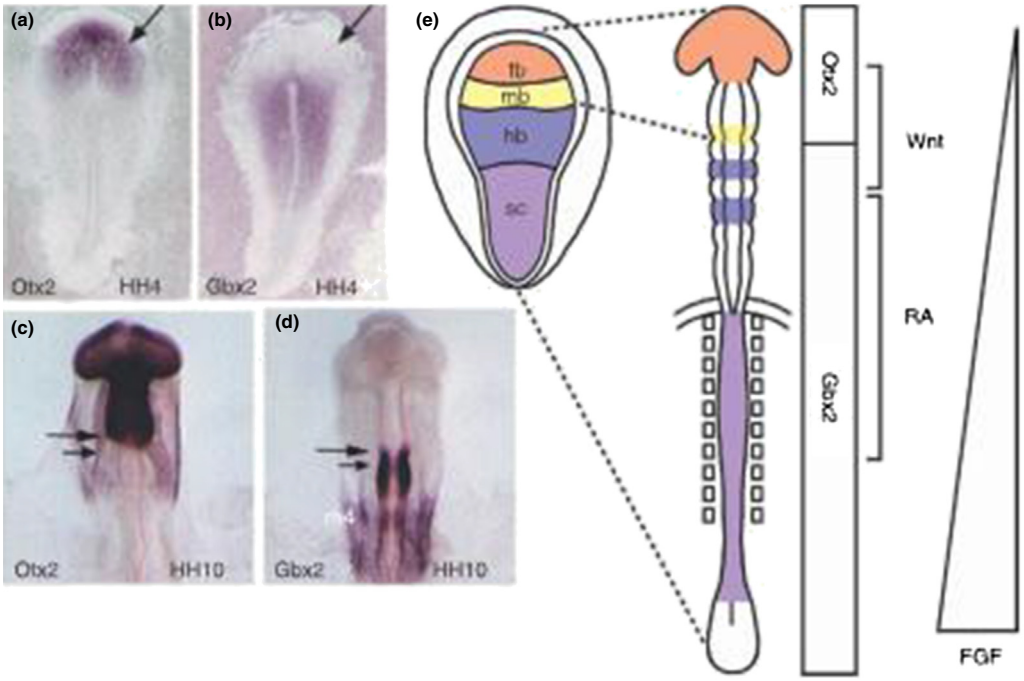

Figure 7 Gradient expression of Wnt, retinoic acid (RA) and FGF is involved in anterior posterior patterning of the embryo. It regulates expression of otx2 which reflects the basic division between the rostral brain and the hindbrain and spinal cord. Otx2 expression is depicted at various stages of embryonic development in the chick brain. Otx2 is expressed in the anterior neural plate (a) and remains expressed in most of the brain throughout development (b–e). The arrowhead points to the midbrain-hindbrain boundary. (e) The relationship between the three transformer signals and the expression of otx2 and gbx2. Modified from Millet, S., Bloch-gallego, E., Simeone, A., Alvarado-Mallart, R.M., 1996. The caudal limit of Otx2 gene expression as a marker of the midbrain/hindbrain boundary: a study using in situ hybridisation and chick/quail homotopic grafts. Development 122 (12), 3785–3797.

Yamaguchi has detailed the importance of Wnt role in vertebrate axis formation and heart development in particular, Yang (2012) recently has described how Wnts are involved in regulating PCP and in doing so by way of their gradient expression provide quantitative information to generate a distinct pattern as in proximal distal limb bud patterning. Interestingly, Wnt along with Shh upstream from it also involved in cleft lip pathogenesis (Kurusaka et al., 2014)

FGF signaling is essential to delimit the expression domain of BMPs as early as the blastula stage (Arnold and Robertson, 2009). Initially BMPs are expressed throughout the bilaminar disc. According to Thisse and Thisse (2005), FGF signaling represses BMP expression dorsally, thus delimiting BMP expression to the ventral embryo. FGF activity thus is an essential factor in the early establishment of dorsoventral patterning (Fürthauer et al., 2004). As reviewed above "dorsalization" in the embryo (see **Figure 8**) involves the development of the primitive streak and its replacement by growing notochord. Movement patterns of anterior streak cells can be explained by an FGF8-mediated chemorepulsion of cells away from the streak followed by chemoattraction toward an FGF4 signal produced by the forming notochord (Yang et al., 2002). Toyoda et al. (2010) recently has shown that in mouse FGF8 establishes a gradient along the AP axis of the neocortical primordium. They cite evidence that the anterior source of FGF8 and FGF17 is an integral patterning center for the entire neocortex (Toyoda et al., 2010).

The HOX gene families are members of the Homeobox superfamily of genes (Sadler, 2010; Tan et al., 2013). The HOX genes are partly responsible for determining the AP axis in the developing embryo (see **Figure 6**) and regulate the formation of musculoskeletal structures in each body segment (Tan et al., 2013). Timed transcription of different Hox genes lined up in AP axis involves a retinoic acid (RA) gradient (**Figure 7**) that is low anteriorly and high posteriorly (Tan et al., 2013). Together they make up an important part of the segmentation clock. Pitx2 is a Homeobox gene that is important in L-R axis asymmetry and organogenesis (Field et al., 2011).

Left-Right Asymmetry Body (Organs)

Understanding the initiation of left-right axis formation has been quite a challenge. Unlike the anterior-posterior and dorsoventral axes with which it is coordinated there are no exogenous cues (e.g., gravity or point of sperm entry) that serve as reference points for its determination (Levin, 2001; Arnold and Robertson, 2009). The "two cilia hypothesis" which has gained traction in recent years offers real promise toward solving this conundrum (Nakaya et al., 2005; Tan et al., 2013). During gastrulation, cilia emanating from the ventral surface of the primitive node play a crucial role in the "breaking" of bilateral symmetry (McGrath and Brueckner, 2003). Recently, Yang (2012) has pointed out that PCP is involved in the polarization of the node cells and in the uniform curved shape of the cilia. This is because the net

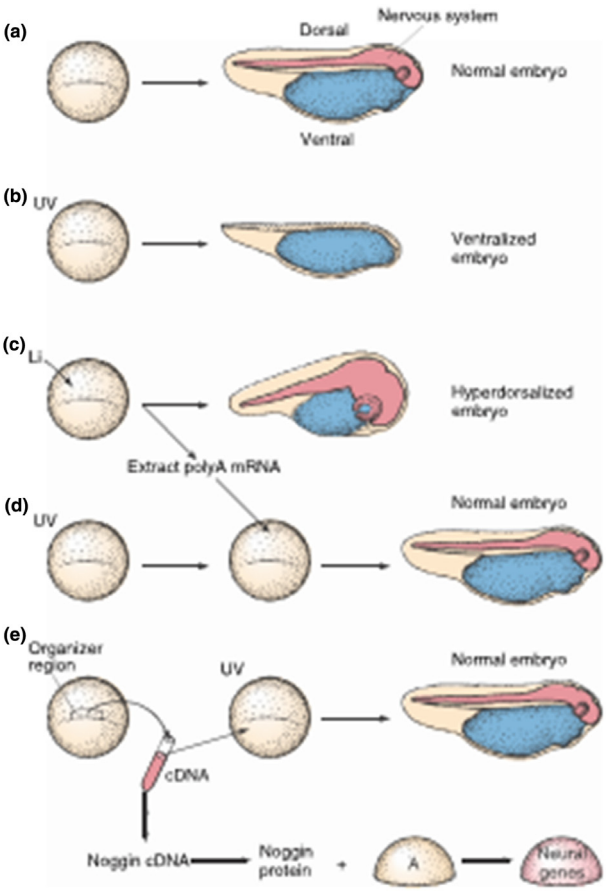

Figure 8 Dorsalization and ventralization of the embryo comprise balanced development of nervous system (red) and heart-gut formation (blue). (a) Normal development of a Xenopus embryo. (b) uV light treatment of the early embryo inhibits the development of dorsal structures by disrupting the cytoskeleton rearrangements that pattern the dorsal inducing molecules prior to gastrulation (ventralized). (c) Lithium treatment of the early embryo has the opposite effect; the embryo develops more than normal dorsal tissue (i.e., hyperdorsalized). (d) If messenger RNA is extracted from the hyperdorsalized embryos and injected into a uV-treated embryo, the messages encoded in the mRNA can "rescue" the uV-treated embryo and it develops relatively normally. (e) Similarly, cDNA from the organizer region of a normal embryo can rescue a uV-treated embryo.

leftward flow of extraembryonic fluid at the node is thought to be generated by motile curved cilia that are positioned in the posterior aspect of the node cells. In addition, nonmotile mechanosensory cilia, sense Nodal flow on the left side of the node and convert it into an asymmetric Ca2+-dependent signal transduction event that upregulates Nodal on the lateral plate mesoderm (LPM) (Yang, 2012; Tan et al., 2013). The two cilia hypothesis is consistent with studies by Pennekamp et al. (2002) and McGrath et al. (2003), who found that mutations in the polycystic kidney disease 2 (Pkd2) gene that partakes in the generation of asymmetric Ca2+ flux cause abnormal LR development.

It now has been well-documented across vertebrates that LPM Nodal partakes in a cascade of events (Hamada et al., 2001; Yamaguchi, 2001). That is, upstream, Nodal activity is upregulated by Notch pathway activity (Shh) and downstream by Nodal itself in a positive feedback loop (Field et al., 2011). Autoactivation of Nodal results in rapid spread throughout the left LPM along with the induction of Lefty-1 and Pitx2 expression. Nodal is then quickly downregulated by Lefty-2 via a negative feedback loop. Lefty-1 most likely acts as a molecular barrier preventing left-sided Nodal signals to cross to the right side. More recently, Nakaya et al. (2005) reported evidence that Wnt, expressed in the primitive streak and dorsal posterior node, acts as a long-range signaling molecule, directly regulating target gene expression throughout the node and presomitic mesoderm.

Nodal expression in the left LPM occurs around the 2–4 somite stage and is turned off by the 8 somite stage (Nakaya et al., 2005). Pitx2 continues to be expressed well past this point and partakes in the lateralized morphogenesis of the heart (Field et al., 2011). Even so, the details of the coupling of somitogenesis and L-R axis are still not well understood (Nakaya et al., 2005). One possible mechanism proposed by Yost (1992) involves (polarized L-R) fibrils that are attached to the ectoderm. These fibrils supposedly change their orientation from parallel to perpendicular to the AP axis during neurulation at which time, coincident with movement of the cardiac and gut primordia across the extracellular matrix, the ectoderm is stretched by elongation of the embryo in the anterior-posterior axis.

Brain: Molecular Pathways Setting the Stage for Organizers and Cortical Patterning

Monuki and Walsh, 2001 summarized the hierarchical arrangement of molecular pathways and establishment of organizers in a recent review of the molecular pathways involved in telencephalic development as follows. Initially, Fgf8-Foxg1 is involved in the induction of the ANR whereas HPE genes, including Shh promote the dorsal midline roof plate. From here on, in their account genetic/molecular pathways diverge from those described above in blastogenesis (although, see also Geschwind and Miller (2001 and Choe et al., 2012) below) for the continued role of Wnt. That is, dorsalization of telencephalic tissue is specified (Neurogenins and Gli3) and a cortical VZ fate is selected from other dorsal fates by Lhx2. The cortical VZ then becomes regionally specified and signaling molecules, including those from the dorsal midline region (roof plate and cortical hem), then act on the cortical VZ to regulate its expansion. Relative sizes of cortical VZ fields and ultimately cortical areas are being determined by gradients of intrinsic transcriptional regulators Emx2 and Pax6 (Muzio and Mallamaci, 2003 see **Figure 9**).

Interestingly, Buckner and Krienen (2013) cite evidence by Fukuchi-Shimogori and Grove (2001), who demonstrated that varying FGF8, in its normally occurring anterior location of the telencephalon, could expand and contract the somatosensory map of the adult mouse cerebral cortex (see **Figure 10**).

Additionally, they showed that by unnaturally expressing FGF8 in a second posterior location a duplicate somatosensory map of mouse barrel cortex was formed (Buckner and Krienen, 2013).

As stated above CC development with axons actually crossing the midline is relatively late. Choe et al., 2012 found

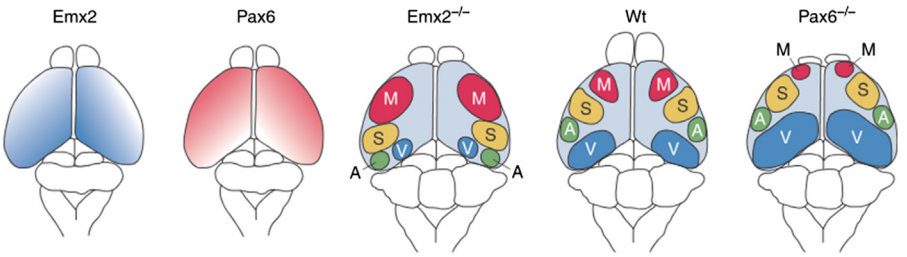

Figure 9 Two transcription factors critical for the specification of regional identities in the cortex are Pax6 and Emx2. Emx2 is expressed primarily in the posterior cerebral cortex and then gradually diminishes in expression toward the rostral cortical pole; Pax6 has the complementary pattern of expression. Loss of either the Pax6 gene or the Emx2 gene affects the cerebral cortical pattern of development. In the wild-type (wt) animal, the motor cortex (m) is primarily located in the rostral cortex, and the other sensory areas for somatosensation (S), auditory sensation (A), and visual perception (v) are located in the middle and posterior cortex, respectively. In the Emx2-deficient mice, the pattern is shifted caudally, and a greater area is occupied by the motor cortex; by contrast, in the Pax6-mice the visual cortex is expanded and the motor cortex is severely reduced. Modified from Muzio, L., Mallamaci, A., 2003. Emx1, emx2 and pax6 in specification, regionalization and arealization of the cerebral cortex. Cerebral Cortex 13 (6), 641–647.

that this late development is effected by BMP7 that serves as a barrier to early projection of axons across the midline. They cite evidence that early CC development involves cingulate (pioneer) neurons adjacent to the Cajal-Retzius cells and meningeal tissues. Wnt3 expression by these cells and as well as indirect Wnt control by the meninges allows them to overcome the negative effects of BMP7 (Choe et al., 2012).

Central Nervous System Asymmetry

Several genes including ROBO (roundabout) and Ephrin signaling pathways (see **Figure 11**) have been shown to be involved in axon crossings in the CC as well as in the medulla and spinal cord. (Kullander et al., 2001; Paixão et al., 2013). That said, consistent with the observation that congenital MM (abnormal midline crossing(s) of CST) primarily affect the upper limbs, mutations in them do not always show interference with alternating movements in gait (Depienne et al., 2012; Paixão et al., 2013; Peng and Charron, 2013). Peng and Charron (2013) report that the first mutation specifically linked to congenital MM without additional neurological defects in humans (i.e., "essential" MM) was in deleted in colon cancer (DCC). More recently a mutation in the RAD51 gene has been shown to produce MM as well (Depienne et al., 2012). DCC is a receptor involved in axon guidance and midline crossing in particular by binding to Netrin to promote a chemoattractive response (Depienne et al., 2012) Rad51 is a gene involved in DNA repair but may also specifically be involved in axon crossing at the pyramids in mouse (Depienne et al., 2012). There is also evidence that Shh made by floor plate cells collaborates with netrin-1 in midline attraction (Ly et al., 2008). Peng and Charron 2013 note that in the DCC mutant family males are more affected than females and that this could perhaps indicative of sex differences in CST development.

Seminal studies by Sun et al. (2005), and Sun and Walsh, (2006) reveal significant asymmetries of gene expression in embryos as early as 12 weeks gestational age. In particular, the transcription factor LIM-only protein 4 (Lmo4) gene is consistently more highly expressed in the right peri-Sylvian cortex than in the left. Lmo4 relates to the regulation of normal sensorimotor control (Huang et al., 2009). Specifically, studies by Cederquist et al. (2013) and Huang et al. (2009) reveal that Lmo4 establishes molecular anatomical diversity of motor control projection neurons within the motor cortex and this diversity of neuronal connectivity centrally includes the development of dual projection neurons that integrate sensory and motor information within the same hemisphere and across hemispheres.

Monuki and Walsh (2001) recently concluded that FOXP2 may be a candidate gene in peri-Sylvian polymicrogyria, which results in difficulties with swallowing, articulation, and language acquisition as well as frequent and severe seizures. They further state that because of its regional nature this disorder may well involve a gene mutation that regulates later events in cortical patterning and thus causes subtle rather than gross malformations of the cortex along with mild cognitive and behavioral defects. In addition to cortical patterning, several lines of studies suggest that FOXP2 in mice and humans acts to regulate dopamine and corticostriatal function related to motor learning (Enard et al., 2009; Murugan et al., in this issue). Galaburda et al. (2006) cite two reports, one describing 8 dyslexic cases with the neuronal migration anomaly and periventricular nodular heterotopia (Chang et al., 2005), and 13 children with peri-Sylvian polymicrogyria and developmental language deficits (de Vasconcelos Hage et al., 2006). The most consistent findings in the original cases are nests of neurons, termed ectopias, in cortical layer 1 and occasional focal microgyria affecting language areas (Galaburda et al., 2006). Additional defects in the thalamus and cerebellum, together with the cortical changes, are proposed to explain the dyslexic phonological deficits, as well as the auditory discrimination and motor deficits that occur in some cases (Galaburda et al., 2006; Giraud and Ramus, 2013). DYX1C1, KIAA0319, DCDC2, and ROBO1 are reported to be dyslexia candidate susceptibility genes (Galaburda et al., 2006; Giraud and Ramus, 2013). The proteins encoded by these genes may be functionally linked in pathways involved in neuronal migration and axon growth. (Galaburda et al., 2006; Graham and Fisher, 2013). In this regard, it is of note that subcortical white matter abnormalities have been demonstrated in recent studies using DTI. Interestingly, Rimrodt et al. (2010) used

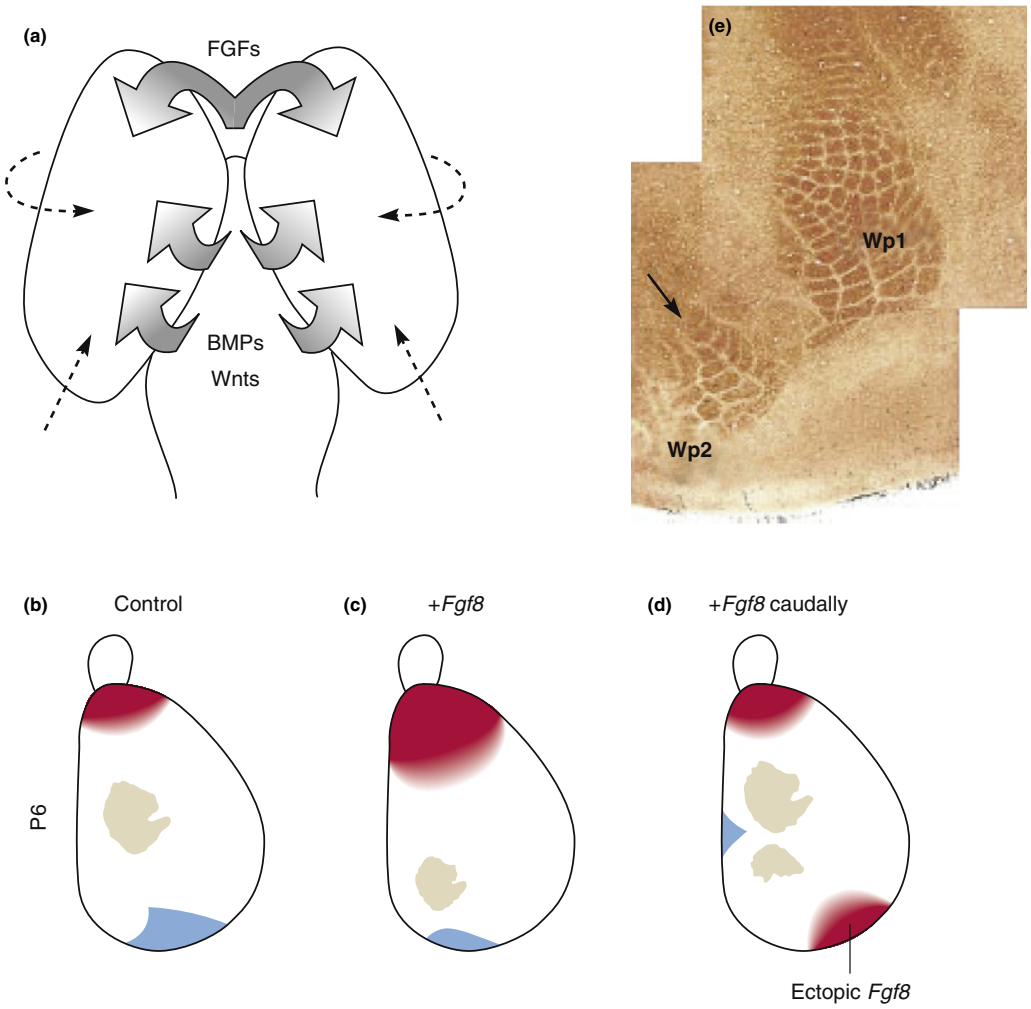

Figure 10 Fgf8 patterns the cerebral cortex. (a) Fgf8 is expressed at the anterior pole of the developing telencephalon, while BmPs and wnt genes are expressed in the posterior pole. Grove and her colleagues have misexpressed Fgf8 in different positions within the developing cortex. (b) In the normal mouse, the barrel fields of the somatosensory map (yellow) are located near the middle of the cerebral cortex, while Fgf8 (red) is expressed anteriorly and BmP (blue) is expressed posteriorly. (c) Increasing the amount of Fgf8 (red) in the anterior pole causes a caudal shift in the cortical regions, including the somatosensory map. (d) Placing a bead of Fgf8 in the caudal cortex causes the formation of a duplicated, mirror image of the somatosensory map. (e) micrograph of duplicated somatosensory maps after the addition of an ectopic Fgf8 bead. Wp1 is the original map and Wp2 is the new map. Modified from Grove, E.A., Fukuchi-Shimogori, T., 2003. Generating the cerebral cortical area map. Annual Review of Neuroscience 26, 355–380.

DTI and tractography and concluded that atypical white matter connectivity (and orientation) especially in the left inferior frontal gyrus in dyslexic participants may well be the result of cortical patterning abnormalities.

As for asymmetric degenerative disease, Geschwind and Miller (2001) entertain the idea that molecular pathways involved in L-R asymmetry in embryonic development may be at play in asymmetric degeneration. Specifically, they implicate the Wnt pathway since it regulates Tau which has been shown to be deficient in FTD.

Substrates Underlying Cerebral Dominance

Over the years cerebral dominance has been defined as the hemisphere where language and hand preference are uniquely colocalized in the majority of people. This uniquely human cerebral arrangement has led to biased views of cerebral asymmetry where the left hemisphere is more special than the right and that uniquely human genetics make it so (see Corballis (2014) for a recent review on facts and fantasies surrounding left brain-right brain). That is, left-right structural asymmetries are often seen as caused by genetics that are "special." While some substitutions on FOXP2 gene indeed appear quite recent (Graham and Fisher, 2013) and the data by Sun et al. (2005), and Sun and Walsh (2006) point to a set of uniquely human genes that have recently evolved, a recent study by Jahanshad et al. (2010) point out that environmental factors codetermine cerebral asymmetry along with genetics. These investigators used DTI and a twin/sibling design to map the 3D pattern of asymmetries and to search for regions where these

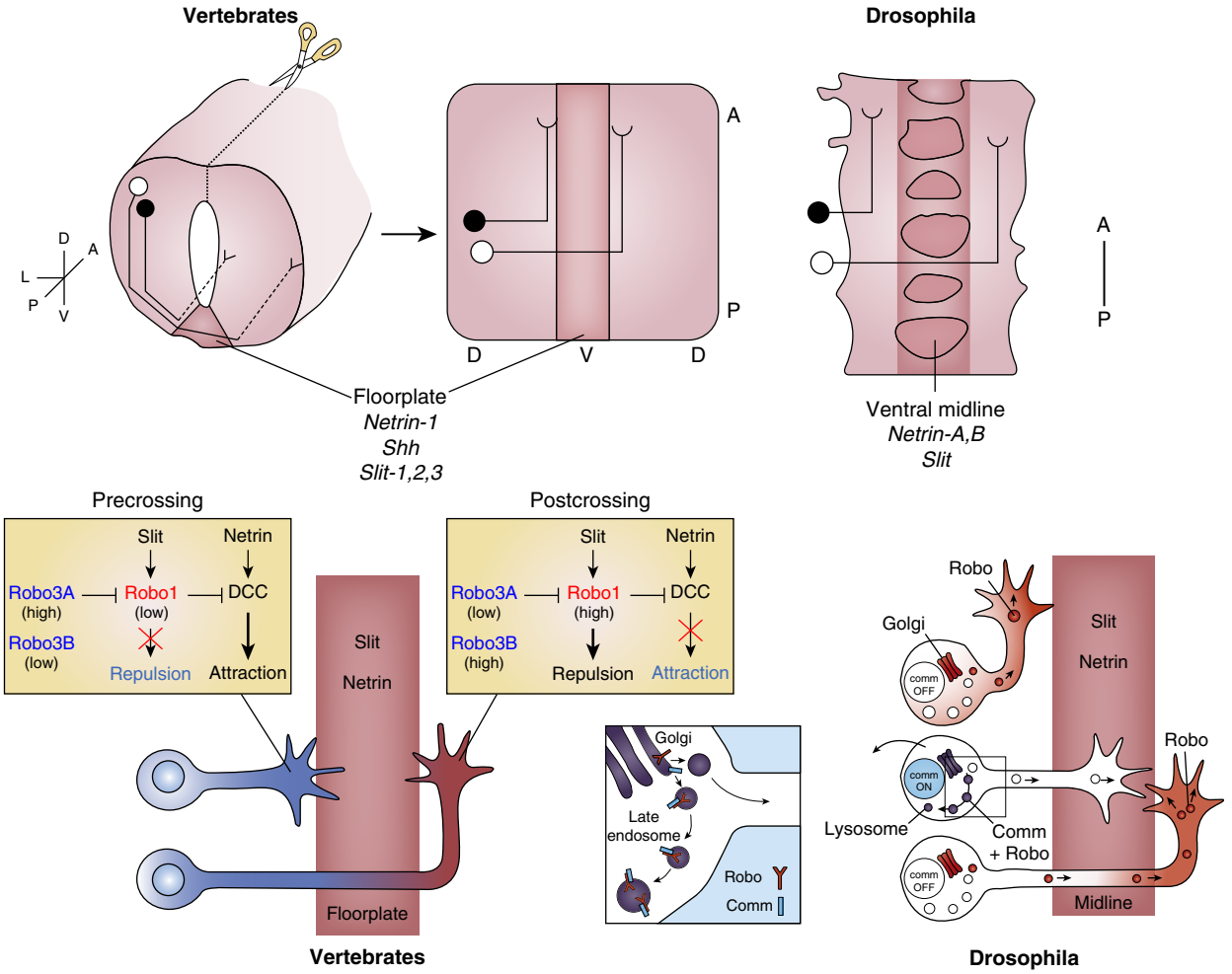

Figure 11 Comparison of conserved mechanisms of midline crossing in vertebrate and fly. Top: filleted vertebrate preparation, obtained by dissecting the embryonic spinal cord of the chick and the embryonic Drosophila ventral nerve cord. Homologous guidance molecules (particularly Netrins and Slits) are expressed at the ventral midline. Bottom right. In flies, Comm in precrossing axons binds to newly made Robo coming from the Golgi apparatus and sorts it to the late endosome for degradation. Bottom left. In vertebrates, Robo3 antagonizes Robo1 to allow crossing in mice. In precrossing commissural axons, Robo3 levels are high, and Robo1 levels low. Robo3A inhibits Robo1-mediated repulsion in these axons so that they are instead attracted to the floorplate by Netrin-1 and Shh. After crossing, Robo3A is replaced by Robo3B, and Robo1 activity is high. Axons are now repelled by signaling of Slit through Robo1. In addition, attraction to Netrin is downregulated, owing to a slit-dependent interaction between Robo1 and DCC. From Dickson, B.J., Gilestro, G.F., 2006. Regulation of commissural axon pathfinding by slit and its Robo receptors. Annual Review of Cell and Developmental Biology 22, 651–675.

asymmetries are highly heritable and thus may be accounted for by genetics. Using this statistical method they report significant asymmetries in FA with frontal R > L and temporal L > R regions. Interestingly, regional analysis of FA asymmetry shows that genetic factors account for 33% of the variance in the inferior fronto-occipital fasciculus, 37% of the variance in the anterior thalamic radiation, and 20% of the variance in the forceps major and the uncinate fasciculus. In addition, shared environmental factors account for about 15% of the variance in the corticospinal tract and about 10% of the variance in the forceps minor. Jahanshad et al., 2010 further report higher similarities in the intraclass correlation maps among monozygotic compared to dizygotic twins, thus providing additional support for the notion that fiber asymmetries are genetically influenced. The overall finding of genetic and (random) environmental factors playing a role in cerebral asymmetry is in line with the proposal that there might be a "right-shift" gene or perhaps more likely a set of "right-shift" genes along with random environmental factors influencing the degree of cerebral dominance and lateralized behavior such as handedness (Annett, 2002. Suzuki and Ando, 2014).

As to the role of the CC in functional asymmetry, Josse et al. (2008) combined DTI assessment of the CC and functional MRI and found revealed that CC size was positively correlated with left lateralization for single word naming and semantic decision tasks. Replicating an earlier measurement by Witelson (1989) it was also shown that in the Hofer II subdivision of the CC, right-handers have a smaller midsaggital surface area than non-right-handers. Faced with this apparent disparity, Josse et al., 2008 concluded that language lateralization is stronger when CC size is larger and that the effects of CC on handedness

and language lateralization are independent. Interestingly, analysis of CC size and language lateralization data in terms of increased/decreased left-right hemisphere activation patterns shows that the effect of CC size on posterior temporal activation is partly driven by right hemisphere inhibition, whereas the effect of CC size on inferior frontal activation is explained by increased information transfer to the left hemisphere from right hemisphere homolog or heterolog regions (Josse et al., 2008).

Aside from the CC, the corticospinal tract has been considered a possible (localized) substrate for hand dominance. As stated above a recent investigation by Hervé et al. (2011) using a Tesla 7 MRI scanner found a leftward asymmetry in the internal capsule in T2 measurements in non-right-handed persons but were unable to confirm T1 asymmetry shown in an earlier DTI study in this group (Hervé et al., 2009). Hervé et al. (2011) caution that while this finding could suggest a greater myelination of the internal capsule in their "motor dominant" hemisphere, the effect was not specific to motor fibres of the CST and was also observed in a small group. Perhaps more revealing are two recent DTI studies one conducted in children and one in elderly analyzing hemispheric connectivity and quality of performance using novel methodologies. Barber et al. (2012) used resting state functional connectivity to examine brain-behavior relationships between motor network connectivity and motor control abilities in typically developing (TD) right-handed children. They found consistent and robust relationships between children's motor abilities and the degree to which connectivity was stronger within the LH motor circuit than it was within the RH circuit. Greater left-lateralized motor circuit connectivity predicted overall better motor ability as reflected in a test of motor skill (PANESS). These results contrast with the findings in a study with young and elderly persons by Davis et al. (2012). This study combined eye-tracking, FMRI, and DTI as well as motor testing to evaluate whether recruitment in the contralateral hemisphere is similar and equally effective in younger and older persons. While in the scanner, participants performed a word matching task where the word pairs were shown unilaterally (i.e., in one visual field) or bilaterally. Results revealed that compared to the younger adults older adults showed greater functional connectivity between left and right PFC, particularly in the bilateral condition. In addition, both behavioral and neural (functional connectivity) measures of cross-hemispheric communication were significantly correlated with DTI measures of callosal integrity in older – but not younger – adults. Overall, these findings indicate that contralateral PFC over-recruitment in older adults is associated with increased left-right PFC connectivity, and furthermore, that this connectivity while dependent on white matter integrity in older adults is beneficial for performance.

Conclusion

The last decade or so has witnessed major methodological advances in imaging and genetics but also in computational analysis. In addition, research paradigms have become increasingly sophisticated lining up multiple approaches as in combining FMRI and DTI or using resting state as the baseline when in addition behavioral indexes are quantified and correlated with indexes of connectivity.

It is fair to say that as a result of this shift in research design a simple or naive structural localizationist approach to addressing questions of asymmetry in brain and body has given way to a developmental and physiological perspective. It appears that asymmetries of body and brain are established quite early and share basic molecular pathways that are evolutionarily well conserved. That said, there are some genes that show altered (asymmetric) expression levels in human brains and play a role in patterning and connectivity providing the brain unlike the viscera lifelong adaptability.

Bibliography

Annett, M., 2002. Handedness and Brain Asymmetry: The Right Shift Theory. Taylor & Francis, New York.

Arnold, S.J., Robertson, E.J., 2009. Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. Nature Reviews Molecular Cell Biology 10, 91–103.

Barber, A.D., Srinivasan, P., Joel, S.E., Caffo, B.S., Pekar, J.J., Mostofsky, S.H., January 2012. Motor "Dexterity"?: evidence that left hemisphere lateralization of motor circuit connectivity is associated with better motor performance in children. Cerebral Cortex 22, 51–59.

Bataille, S., Demoulin, N., Devuyst, O., Audrézet, M.P., Dahan, K., Godin, M., Fontès, M., Pirson, Y., Burtey, S., 2011. Association of PKD2 (polycystin 2) mutations with left-right laterality defects. American Journal of Kidney Diseases 58, 456–460.

Buchel, C., Raedler, T., Sommer, M., Sach, M., Weiller, C., Koch, M.A., 2004. White matter asymmetry in the human brain: a diffusion tensor MRI study. Cerebral Cortex 14, 945–951.

Buckner, R.L., Krienen, F.M., 2013. The evolution of distributed association networks in the human brain. Trends in Cognitive Sciences 17, 648–665.

Casey, B., 1998. Two rights make a wrong: human left-right malformations. Human Molecular Genetics 7, 1565–1571.

Cederquist, G.Y., Azim, E., Shnider, S.J., Padmanabhan, H., Macklis, J.D., 2013. Lmo4 establishes rostral motor cortex projection neuron subtype diversity. The Journal of Neuroscience 33, 6321–6332.

Chang, B.S., et al., 2005. Reading impairment in the neuronal migration disorder of periventricular nodular heterotopia. Neurology 64, 799–803.

Choe, Y., Siegenthaler, J.A., Pleasure, S.J., 2012. A cascade of morphogenic signaling initiated by the meninges controls corpus callosum formation. Neuron 73, 698–712.

Cohen, M.M., 2001. Asymmetry: molecular, biologic, embryopathic, and clinical perspectives. American Journal of Medical Genetics 101, 292–314.

Cooper, J.A., 2008. A mechanism for inside-out lamination in the neocortex. A mechanism for inside-out lamination in the neocortex. Trends in Neurosciences 31, 113–119.

Constam, D.B., Robertson, E.J., 2000. Tissue-specific requirements for the proprotein convertase Furin/SPC1 during embryonic turning and heart looping. Development 127, 245–254.

Corballis, M.C., November 2013. Early signs of brain asymmetry. Trends in Cognitive Sciences 17, 554–555.

Corballis, M.C., 2014. Left brain, right brain: facts and fantasies. PLoS Biology 12 (1), e1001767.

Davis, S.W., Kragel, J.E., Madden, D.J., Cabeza, R., 2012. The architecture of cross-hemispheric communication in the aging brain: linking behavior to functional and structural connectivity. Cerebral Cortex 22, 232–242.

Depienne, C., Bouteiller, D., Meneret, A., et al., 2012. RAD51 haploinsuffciency causes congenital mirror movements in humans. American Journal of Human Genetics 90, 301–307.

De Robertis, E., Kuroda, H., 2004. Dorsal-ventral patterning and neural induction in xenopus embryos. Annual Review of Cell and Developmental Biology 20, 285–308.

Dubois, J., Hertz-Pannier, L., Cachia, A., Mangin, J.F., Le Bihan, D., Dehaene-Lambertz, G., 2009. Structural asymmetries in the infant language and sensori-motor networks. Cerebral Cortex 19, 414–423.

Enard, W., Gehre, S., Hammerschmidt, K., et al., 2009. A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. Cell 137, 961–971.

Dickson, B.J., Gilestro, G.F., 2006. Regulation of commissural axon pathfinding by slit and its Robo receptors. Annual Review of Cell and Developmental Biology 22, 651–675.

Field, S., Riley, K.L., Grimes, D.T., et al., 2011. Pkd1l1 establishes left-right asymmetry and physically interacts with Pkd2. Development 138, 1131–1142.

Fukuchi-Shimogori, T., Grove, E.A., 2001. Neocortex patterning by the secreted signaling molecule FGF8. Science 294, 1071–1074.

Fürthauer, M., et al., 2004. FGF signalling controls the dorsoventral patterning of the zebrafish embryo. Development 131, 2853–2860.

Galaburda, A.M., LoTurco, J., Ramus, F., Fitch, R.H., Rosen, G.D., 2006. From genes to behavior in developmental dyslexia. Nature Neuroscience 9, 1213–1217.

Geschwind, N., Galaburda, A.M., 1985. Cerebral lateralization. Biological mechanisms, associations, and pathology: I. A hypothesis and a program for research. Archives of Neurology 428–459.

Geschwind, D.H., Miller, D.H., 2001. Molecular approaches to cerebral laterality: development and neurodegeneration. American Journal of Medical Genetics 101, 370–381.

Gilles, F., 1983. Telencephalon medium and the olfacto-cerebral outpouching. In: Gilles, F., Leviton, A., Dooling, E.C. (Eds.), The Developing Human Brain: Growth and Epidemiologic Neuropathology. John Wright, Boston, pp. 59–86.

Giraud, A.L., Ramus, F., 2013. Neurogenetics and auditory processing in developmental dyslexia. Current Opinion in Neurobiology 23, 37–42.

Graham, S.A., Fisher, S.E., 2013. Decoding the genetics of speech and language. Current Opinion in Neurobiology 23, 43–51.

Grove, E.A., Fukuchi-Shimogori, T., 2003. Generating the cerebral cortical area map. Annual Review of Neuroscience 26, 355–380.

Hamada, H., Meno, C., Watanabe, D., Saijoh, Y., 2001. Establishment of vertebrate left–right asymmetry. Nature Reviews Genetics 3, 103–113.

Heldin, C.H., Miyazono, K., ten Dijke, P., 1997. TGF-beta signaling from cell membrane to nucleus through SMAD proteins. Nature 390, 465–471.

Hepper, P.G., Wells, D.L., Lynch, C., 2005. Prenatal thumb sucking is related to postnatal handedness. Neuropsychologia 43, 313–315.

Hervé, P.Y., Leonard, G., Perron, M., Pike, G.B., Pitiot, A., Richer, L., Veillette, S., Pausova, Z., Paus, T., 2009. Handedness, motor skills and maturation of the corticospinal tract in the adolescent brain. Human Brain Mapping 30, 3151–3162.

Hervé, P.Y., Cox, E.F., Lotfipour, A.K., Mougin, O.E., Bowtell, R.W., Gowland, Paus, T., 2011. Structural properties of the corticospinal tract in the human brain: a magnetic resonance imaging study at 7 Tesla. Brain Structure and Function 216, 255–262.

Hervé, P.Y., et al., 2009. Handedness, motor skills and maturation of the corticospinal tract in the adolescent brain. Human Brain Mapping 30, 3151–3162.

Hervé, P.Y., Zago, L., Petit, L., Mazoyer, B., Tzourio-Mazoyer, N., 2013. Revisiting human hemispheric specialization with neuroimaging. Trends in Cognitive Sciences 17, 69–80.

Huang, Z., Kawase-Koga, Y., Zhang, S., Visvader, J., Toth, M., Walsh, C.A., Sun, T., 2009. Transcription factor Lmo4 defines the shape of functional areas in developing cortices and regulates sensorimotor control. Developmental Biology 327, 132–142.

Inui, M., Montagnera, M., Ben-Zvib, D., Martelloa, G., Soligoa, S., Manfrina, A., Aragonaa, M., Enzoa, E., Zacchignaa, L., Zanconatoa, F., Azzolina, L., Duponta, S., Cordenonsia, M., Piccoloa, S., 2012. Self-regulation of the head-inducing properties of the Spemann organizer. Proceedings of the National Academy of Sciences 109 (38), 15354–15359.

Jahanshad, N., Lee, A.D., Barysheva, M., McMahon, K.L., de Zubicaray, G.I., Martin, N.G., Wright, M.J., Toga, A.W., Thompson, P.M., 2010. Genetic influences on brain asymmetry: a DTI study of 374 twins and siblings. NeuroImage 52, 455–469.

Jammes, J., Gilles, F., 1983. Telencephalic development: matrix volume and isocortex and allocortex surface areas. In: Gilles, F., Leviton, A., Dooling, E.C. (Eds.), The Developing Human Brain: Growth and Epidemiologic Neuropathology. John Wright, Boston, pp. 87–93.

Josse, G., Seghier, M.L., Kherif, F., Price, C.J., 2008. Explaining function with anatomy: language lateralization and corpus callosum size. The Journal of Neuroscience 28, 14132–14139.

Kosaki, K., Bassi, M.T., Kosaki, R., Lewin, M., Belmont, J., Schauer, G., Casey, B., 1999. Characterization and mutation analysis of human LEFTY A and LEFTY B, homologues of murine genes implicated in left-right axis development. American Journal of Human Genetics 64, 712–721.

Kennedy, D., O'Craven, K., Ticho, B., Goldstein, A., Makris, N., Henson, J., 1999. Structural and functional brain asymmetries human situs inversus totalis. Neurology 53, 1260–1265.

Kurusaka, H., Lulianella, A., Williams, T., Trainor, P.A., 2014. Disrupting hedgehog and WNT signaling interactions promotes cleft lip pathogenesis. Journal of Clinical Investigation 124, 1660–1671.

Klar, A.J., 2003. Human handedness and scalp hair-whorl direction develop from a common genetic mechanism. Genetics 165, 269–276.

Kriegstein, A.R., Noctor, S.C., July 2004. Patterns of neuronal migration in the embryonic cortex. Trends in Neurosciences 27 (7), 392–399.

Kullander, K., Croll, S.D., Zimmer, M., Pan, L., McClain, J., Hughes, V., Zabski, S., DeChiara, T.M., Klein, R., Yancopoulos, G.D., Gale, N.W., 2001. Ephrin-B3 is the midline barrier that prevents corticospinal tract axons from recrossing, allowing for unilateral motor control. Genes and Development 15, 877–888.

Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., Monaco, A.P., 2001. A forkhead-domain gene is mutated in a severe speech and language disorder. Nature 413, 519–523.

Leigh, M.W., Pittman, J.E., Carson, J.L., Ferkol, T.W., Dell, S.D., Davis, S.D., Knowles, M.R., Zariwala, M.A., 2009. Clinical and genetic aspects of primary ciliary dyskinesia/Kartagener syndrome. Genetics in Medicine 11, 473–487.

Levin, M., 2001. Asymmetry of body and brain: embryological and twin studies. In: Smelser, N., Baltes, P. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Elsevier, Oxford, UK, pp. 853–859.

Levin, M., 2004. The embryonic origins of left-right asymmetry. Critical Reviews in Oral Biology and Medicine 15, 197–206.

Levin, M., 2005. Left-right asymmetry in embryonic development: a comprehensive review. Mechanisms of Development 122, 3–25.

Lin, P.Y., Roche-Labarbe, N., Dehaes, M., Fenoglio, A., Grant, P.E., Franceschini, M.A., 2013. Regional and hemispheric asymmetries of cerebral hemodynamic and oxygen metabolism in newborns. Cerebral Cortex 23, 339–348.

Li, G., Nie, J., Wang, L., Shi, F., Lyall, A.E., Lin, W., Gilmore, J.H., Shen, D., 2013. Mapping longitudinal hemispheric structural asymmetries of the human cerebral cortex from birth to 2 years of age. Cerebral Cortex 1–12. http://dx.doi.org/10.1093/cercor/bhs413.

Ly, A., Nikolaev, A., Suresh, G., Zheng, Y., Tessier-Lavigne, M., Stein, E., 2008. DSCAM is a netrin receptor that collaborates with DCC in mediating turning responses to netrin-1. Cell 133, 1241–1254.

McGrath, J., Somlo, S., Makova, S., Tian, X., Brueckner, M., 2003. Two populations of node monocilia initiate left-right asymmetry in the mouse. Cell 114, 61–73.

McManus, I.C., Martin, N., Stubbings, G., Chung, E.M.K., Mitchison, H.M., 2004. Handedness and situs inversus in primary ciliary dyskinesia. Proceedings of the Royal Society of London. Series B. Biological sciences 22, 2579–2582.

Muzio, L., Mallamaci, A., 2003. Emx1, emx2 and pax6 in specification, regionalization and arealization of the cerebral cortex. Cerebral Cortex 13 (6), 641–647.

Monuki, E.S., Walsh, C.A., 2001. Mechanisms of cerebral cortical patterning in mice and humans. Nature Neuroscience Supplement 4, 1199–1206.

Murugan, M., Harward, S., Scharff, C., Mooney, R., 2013. Diminished FoxP2 levels affect dopaminergic modulation of corticostriatal signaling important to song variability. Neuron 80, 1464–1476.

Nakaya, M., Biris, K., Tsukiyama, T., Jaime, S., Rawls, A., Yamaguchi, T.P., 2005. Wnt3a links left-right determination with segmentation and anterior-posterior Axis elongation. Development 132, 5425–5436.

Noctor, S.C., Martinez-Cerdeno, V., Ivic, L., Kriegstein, A.R., 2004. Cortical neurons arise in symmetric and asymmetric division zones and migrate through specific phases. Nature Neuroscience 7, 136–144.

Noctor, S.C., Martlnez-Cerdefio, V., Kriegstein, A.R., 2008. Distinct behaviors of neural stem and progenitor cells underlie cortical neurogenesis. Journal of Comparative Neurology 508, 28–44.

Niehrs, C., 2004. Regionally specific induction by the *Spemann*-Mangold organizer. Nature Reviews Genetics 5, 425–434.

Paixão, S., Balijepalli, A., Serradj, N., Niu, J., Luo, W., Martin, J.H., Klein, R., 2013. EphrinB3/EphA4-Mediated guidance of ascending and descending spinal tracts. Neuron 80, 1407–1420.

Pennekamp, P., Karcher, C., Fischer, A., Schweickert, A., Skryabin, B., Horst, J., Blum, M., Dworniczak, B., 2002. The ion channel polycystin-2 is required for left-right axis determination in mice. Current Biology 12, 938–943.

Peng, J., Charron, F., 2013. Lateralization of motor control in the human nervous system: genetics of mirror movements. Current Opinion in Neurobiology 23, 109–118.

Rimrodt, S.L., Peterson, D.J., Denckla, M.B., Kaufmann, W.E., Cutting, L.E., 2010. White matter microstructural differences linked to left perisylvian language network in children with dyslexia. Cortex 46, 739–749.

Savage, C., Das, P., Finelli, A.L., Townsend, S.R., Sun, C.Y., Baird, S.E., Padgett, R.W., 1996. *Caenorhabditis elegans* genes sma-2, sma-3, and sma-4 define a conserved family of transforming growth factor beta pathway components. Proceedings of National Academy of Sciences 93, 790–794.

Sadler, T.W., 2010. Langman's Medical Embryology, eleventh ed. Lippincott Williams & Wilkins, Philadelphia.

Sun, T., Patoine, C., Abu-Khalil, A., Visvader, J., Sum, E., Cherry, T.J., Orkin, S.H., Geschwind, D.H., Walsh, C.A., 2005. Early asymmetry of gene transcription in embryonic human left and right cerebral cortex. Science 308, 1794–1798.

Sun, T., Walsh, C.A., 2006. Molecular approaches to brain asymmetry and handedness. Nature Reviews Neuroscience 7, 655–662.

Suzuki, K., Ando, J., 2014. Genetic and environmental structure of individual differences in hand, foot, and ear preferences: a twin study. Laterality 19, 113–128.

Thisse, B., Thisse, C., 2005. Functions and regulations of fibroblast growth factor signaling during embryonic development. Developmental Biology 287, 390–402.

Tan, W., Gilmore, E.C., Baris, H.N., 2013. Human developmental genetics. In: Rimoin, D.L., Pyeritz, R.E., Korf, B. (Eds.), Emery and Rimoin's Principles and Practice of Medical Genetics. Elsevier, pp. 1–63.

Toyoda, R., Assimacopoulos, T.S., Wilcoxon, J., Taylor, A., Feldman, P., Suzuki-Hirano, A., Shimogori, T., Grove, E.A., 2010. FGF8 acts as a classic diffusible morphogen to pattern the neocortex. Development 137, 3439–3448.

Turner, S.J., Hildebrand, M.S., Block, S., Damiano, J., Fahey, M., Reilly, S., Bahlo, M., Scheffer, I.E., Morgan, A.T., 2013. Small intragenic deletion in FOXP2 associated with childhood apraxia of speech and dysarthria. American Journal of Medical Genetics Part A 161, 2321–2326.

de Vasconcelos Hage, S.R., et al., 2006. Specific language impairment: linguistic and neurobiological aspects. Arquivos de Neuro-Psiquiatria 64, 173–180.

Witelson, S.F., 1989. Hand and sex differences in the isthmus and genu of the human corpus callosum. A post mortem morphological study. Brain 112, 799–835.

Yang, X., et al., 2002. Cell movement patterns during gastrulation in the chick are controlled by positive and negative chemotaxis mediated by FGF4 and FGF8. Developmental Cell 3 (3), 425–437.

Yang, X., 2012. Wnt signaling in development and disease. Cell & Bioscience 2, 14.

Yamaguchi, 2001. Heads or tails: Wnts and anterior-posterior patterning. Current Biology 11, R713–R724.

Yamamoto, M., Saijoh, Y., Perea-Gomez, A., Shawlot, W., Behringer, R.R., Ang, Siew-Lan, Hamada, H., Meno, Yamamoto, C., 2004. Nodal antagonists regulate formation of the anteroposterior axis of the mouse embryo. Nature 428, 387–392.

Atomism and Holism: Philosophical Aspects

Michael Esfeld, University of Lausanne, Lausanne, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the philosophy of the social sciences, atomism is the view that human beings can be thinking, rational beings independently of social relations. Holism, by contrast, is the view that social relations are essential to human beings insofar as they are thinking, rational beings. This article first provides an overview of different sorts of atomism and holism (see Section Types of Atomism and Holism). It then briefly sketches the historical background of these notions in modern philosophy (Section The Historical Background of Atomism and Holism). The main part is a systematic characterization of atomism and holism (see Section A Characterization of Atomism and Holism) and a summary of the most important arguments for both these positions (see Section Arguments for Atomism and Holism).

Types of Atomism and Holism

If one speaks of atomism or holism, one has to specify with respect to what atomism or holism is claimed. Atomism in one domain can go with holism in another domain, and vice versa. In the social sciences, the discussion on atomism vs. holism focuses on the question of whether social relations with other humans are essential to human nature. The question thus is whether society in the sense of a social community of human, thinking, and acting beings is an atomistic or a holistic system. The question is neither whether a human being can exist in the biological sense without social relations nor whether a human being is dependent on social relations for maintaining a fully satisfactory human life. The point at issue is whether a human being is dependent on social relations to other humans insofar as he or she is a thinking, rational being. Since, on most accounts, thought in the full sense of the term as exhibited by adult human beings is tied to rationality constraints and since rational agency, including moral agency, presupposes thought, the point at issue can be formulated thus: Are social relations necessary for a human being to have thoughts with a determinate conceptual content? Social atomism is the thesis that an individual considered in isolation can have thoughts with a determinate conceptual content. Social holism, by contrast, is the thesis that social relations are essential for a human being in order to have thoughts with a determinate conceptual content. Social holism thereby implies that having thoughts is tied to speaking a public language; social atomism does not have such an implication.

The discussion on atomism vs. holism concerns not only the conditions under which humans can have thoughts; it extends also to aspects of thoughts. The most important issue in this respect is the question of whether meaning is atomistic or holistic. Semantic atomism is the thesis that each thought or sentence has a meaning independently of other thoughts or sentences. Semantic holism, in reverse, is the thesis that the meaning of a thought or a sentence consists in its inferential relations to other thoughts or sentences in a system of thoughts, a theory, etc. There is a middle position, namely semantic molecularism. This is the thesis that only some particular inferential relations to other thoughts or sentences, most prominently the analytic ones, constitute the meaning of a thought or a sentence.

Atomism and holism are put forward with respect to further aspects of thoughts. Confirmation atomism is the thesis that thoughts or sentences can be empirically confirmed or disconfirmed one by one. Confirmation holism, by contrast, is the thesis that only a whole system of thoughts or a whole theory can be confirmed or disconfirmed by experience. Consequently, if a conflict between theory and experience occurs, there are several possibilities for adapting the theory to experience. Justification atomism is the thesis that there are some thoughts or sentences that can be justified without referring to other thoughts or sentences. These thoughts or sentences are the foundation of knowledge. Justification atomism is, therefore, known as foundationalism. Justification holism, by contrast, is the thesis that the justification of any thought or sentence consists in relations to other thoughts or sentences. Justification holism is, therefore, known as coherentism: the justification of a thought or a sentence is given by its being integrated into a coherent system of thoughts or sentences. Although there are arguments that link all these sorts of atomism or holism together, endorsing any one sort of atomism or holism does not logically commit one to endorse any other sort of atomism or holism as well.

The Historical Background of Atomism and Holism

The issue of atomism vs. holism in the social sciences is linked closely with philosophical reactions to what is perceived as the methodology of modern natural science. Thus, social atomism starts when Hobbes applies the method of physical science to social science. Hobbes – in contrast to, for instance, Gassendi – is not a physical atomist: he does not commit himself to the thesis that there are smallest, indivisible bodies that have a few basic properties such as position and motion and of which every complex body is composed (*De corpore/Of Body*, Chap. 15, Sect. 2, Chap. 27, Sect. 1). However, he transfers to the social sciences the method of (1) dividing a complex system up into its constituent parts, (2) considering the parts and their properties in isolation from each other, and (3) understanding the nature and function of a complex system on the basis of the properties of its constituent parts considered in isolation (*De cive/Of the Citizen*, Preface). Hobbes is famous for applying this method to

political science. He thereby commits himself to social atomism: in his political philosophy, he presupposes individuals each of whom is a thinking, rational being independently of the other ones (*De cive*/*Of the Citizen*, Chap. 8, Sect. 1). He reconstructs social and political relations on that basis.

Despite the widespread acceptance of the sketched methodology of the natural sciences as a paradigm for philosophy, holism is proposed as an ontology in early modern philosophy. Most prominently, Spinoza maintains in his *Ethics* that there is only one substance and that everything is a way of being (a mode) of that substance. However, it is only when the supposition that the methodology of physical science is a paradigm for philosophy is challenged in romanticism and its predecessors that social holism emerges as an option. This development is linked with a focus on language and the thesis that language is essential to thought. Thus, for instance, Herder claims (1) that thought is tied to language and (2) that language is tied to social relations. Consequently, he comes to the conclusion that a human can be a thinking being only in virtue of social relations to other humans (see, in particular, *Abhandlung über den Ursprung der Sprache*; partial English translation in Herder, 1993 and see Spencer, 1996). Hegel systematizes this conception and integrates it into his holistic ontology according to which the world is the dialectical development of a spiritual substance. He goes as far as considering individual thinking beings as accidents of a collective substance (for instance, *Grundlinien der Philosophie des Rechts*, Sect. 156; English translation Hegel, 1991). He can be seen as thereby transforming social holism into collectivism.

Social holism in the romanticists is part of their reaction against conceiving philosophy on the model of the methodology of the physical sciences. But the romanticists do not go as far as developing a proper methodology for the social sciences. It is only from the beginning of the twentieth century on that social holism is proposed as a proper methodology for the social sciences. Most prominent in that respect are the works of Durkheim (1927) and Mead (1934). This development again links up with a development in the natural sciences, namely the emergence of systems theory. Systems theory seeks to understand a system on the basis of its constituent parts – but not on the basis of properties that these parts have independently of each other; on the contrary, systems theory sets out to understand a complex system on the basis of the relations into which its constituent parts enter (von Bertalanffy, 1968). Applied to social systems, systems theory can therefore lead to social holism. Moreover, as regards physics, it is claimed that atomism is replaced with a comprehensive holism in our description of the microphysical level consequent upon quantum theory, in the sense that certain relations rather than intrinsic properties are characteristic of fundamental physical systems. This view is worked out in the position known as ontic structural realism in current philosophy of science (Ladyman and Ross, 2007). However, this holism has – as yet? – nearly no repercussions for the social sciences (but see Wagner, 2012).

A Characterization of Atomism and Holism

Both atomism and holism, taken in a general sense, are claims about complex systems in which constituent parts can be distinguished. Having constituent parts is not sufficient for something to be an atomistic system. In reverse, there is no reason to recognize something that does not admit of parts as a holistic system; such a thing would rather be an atom in the literal sense. Both atomism and holism therefore presuppose the analysis of a complex system in terms of parts. For a characterization of atomism in opposition to holism, it is crucial to consider (1) the sort of independence or dependence that holds among the parts and (2) the properties of the parts with respect to which that sort of independence or dependence holds.

Relations of causal dependence among the parts of a complex system are not sufficient to settle the issue of atomism vs. holism, even if they are complicated relations of mutual or circular causal dependence. No one denies that the development of thought and rationality in a human being causally depends on the social environment and that these social relations can be complex ones of mutual and circular causal dependence. But this commonplace does not amount to a refutation of social atomism and an argument for social holism, on pain of trivializing social holism and turning social atomism into an obviously wrong-headed position.

It is, therefore, appropriate to focus on whether or not there are relations of ontological dependence among the parts of a complex system. Ontological dependence is about something depending for its very existence on the existence of some other thing. Two sorts can be distinguished. (1) If an individual *a* cannot exist unless there is a particular individual *b*, then we have a case of rigid ontological dependence. (2) If there can be no individual of a certain kind unless there is some other individual of a certain kind, then we have a case of generic ontological dependence (Simons, 1987, Chap. 8.3; Correia, 2008). The issue of atomism vs. holism is about generic ontological dependence among the parts of a complex system. The holist claims that some social environment is necessary for thought, but not that it has to be, say, one's family, or one's tribe.

According to the standard account of modal sentences in today's philosophy, the necessity that ontological dependence implies is to be analyzed in terms of possible worlds. One can claim that the necessity is metaphysical and thus concerns all possible worlds. But one does not have to go as far as that: one can restrict the scope of ontological dependence so that it applies, for instance, only to all physically possible worlds, i.e., all those worlds in which the same physical laws hold as in our world. Consequently, one can take the following position: Social atomism or social holism applies only to human thinking beings, i.e., to those who share a human nature and who are subject to the physical and biological laws of our world. One thereby refuses to commit oneself as to whether or not in the imagined case of Martians, who are thinking beings, but who do not have a human nature, social relations are necessary for thought. Hence, one can conceive social atomism or social holism in such a way that these positions formulate a conceptual necessity about all possible thinking beings; or one can restrict these positions in such a way that they apply only to us humans.

When it comes to atomism vs. holism, we can further specify generic ontological dependence: (1) We are interested in generic ontological dependence as a symmetric relation, for we are inquiring into the sort of dependence that captures the way in which the parts of a holistic system depend on each other. (2) In

cases such as social atomism vs. social holism and atomism vs. holism with respect to meaning, confirmation, or justification of thoughts, generic ontological dependence concerns the dependence of things of one sort on there being other things of the same sort (Fodor and Lepore, 1992, pp. 1–2). In social atomism vs. social holism, the question is whether there have to be other human thinking beings if there is one human thinking being. This point does, however, not apply to all cases of holism: if an organism is a holistic system, then something that is a heart is not ontologically dependent on there being other hearts, but on there being blood, kidneys, etc.

Most importantly, (3) the issue of atomism vs. holism can be formulated in a more fine-grained way than being about generic ontological dependence that concerns the existence as such of the parts of the kind of systems in which we are interested. The issue is about generic ontological dependence among the parts of a complex system with respect to certain of their properties. The claim of social holism is not that a human being cannot exist unless there are other humans. The claim is that a human being is ontologically dependent on social relations to other humans with respect to the property of thought and rationality, i.e., insofar as it is a thinking, rational being. Otherwise, social holism would be an absurd position: it would have to claim either that a baby already has thoughts in the full sense of the term, or that a baby and an adult cannot be the same individual. Thus, for a characterization of atomism vs. holism, it is necessary to filter out properties for which it is a substantial issue whether or not the parts of a complex system are ontologically dependent on each other with respect to these properties.

For every system of a qualitative kind, there is a family of nondisjunctive, qualitative properties which make something a system of the kind in question. Such a family of properties can include both nonrelational and relational properties. Something is a system of the kind *S* if and only if it has all – or by far most of all – the properties that make something an *S*. For instance, if something is a grain of sand, a certain molecular structure as well as a shape, size, and mass within a certain margin count among the family of properties that make something a grain of sand. Applied to the parts of an *S*, this notion is to say: There is a family of qualitative, nondisjunctive properties that make something a part of an *S* – in case the thing is arranged with other things in a suitable way. Having all or nearly all the properties that belong to such a family of properties is a necessary condition for something to be a part of an *S*; it is a sufficient condition in conjunction with the condition of a suitable arrangement with other things.

The condition of a suitable arrangement has to be imposed in order to exclude those properties for which it is trivial that something can have them only by being part of a complex system. For instance, a human being can exercise a social role – such as being a judge, or being a salesman – only by being part of a social system. Social roles are arrangement properties. Pointing out social roles is not sufficient to make a case for social holism, again on pain of trivializing social holism and turning social atomism into an obviously wrong-headed position. The question is whether properties that are the prerequisite for exercising a social role – such as the property of thought and rationality in the sense of having thoughts with a determinate conceptual content – require social relations. By excluding from the mentioned family of properties those relational properties in which the arrangement with other things consists, we pick out the properties that underlie the arrangement. There is a substantial case of holism if and only if these properties are relational as well – in the sense that one thing can have these properties only if there are other things together with which this thing is arranged in such a way that there is an *S*. Consequently, there is a substantial case of atomism if and only if something can have these properties independently of whether or not there are other things with which this thing is arranged in such a way that there is an *S*.

The issue of atomism vs. holism can now be characterized as follows: consider a system of the kind *S* and its constituent parts. For every constituent part of an *S*, there is a family of qualitative, nondisjunctive properties that make something a constituent part of an *S* provided that there is a suitable arrangement. The issue whether or not an *S* is holistic depends on whether or not the following condition is satisfied by all the things which are its constituent parts: With respect to some of the properties that belong to such a family of properties, a thing is ontologically dependent in a generic way on there being other things together with which it is arranged in such a way that there is an *S* (Esfeld, 1998). If there are some properties that satisfy this condition, the system is holistic with respect to these properties. With respect to all those properties within the mentioned family of properties that do not satisfy this condition, the system is atomistic.

This is a fine-grained characterization of atomism and holism: One and the same system can be atomistic in regard to some properties, but holistic in regard to other properties. Apart from the above-mentioned variable of the scope of ontological dependence, there are two variables with respect to the properties that atomism or holism concerns: (1) the number of these properties: for instance, in the case of a system of thoughts, one can maintain that holism not only concerns both confirmation and meaning, but also that it concerns only confirmation; one can hence claim that a system of thoughts is a holistic system with respect to confirmation, but an atomistic system with respect to meaning (Fodor and Lepore, 1992, Chap. 2); (2) the importance of these properties: all the properties that atomism or holism concerns are important in that they belong to the family of properties which make something a constituent part of a system *S* provided that there is a suitable arrangement. Nonetheless, the properties that are touched by atomism or holism can be more or less important as far as the identity of the system and its constituents is concerned. For instance, in the case of a thinking human being who is a constituent part of a social community, the property of having thoughts is not necessary for the identity of the human being concerned. If a human being develops from a baby into a normal adult and then goes insane, she nevertheless remains the same human being throughout this development. By contrast, consider the case of atomism or holism of thought with respect to meaning: If a constituent of a system of thoughts changes its meaning, it is no longer the same thought.

The outlined characterization makes clear that the issue of atomism vs. holism is distinct from the issue of individualism vs. collectivism. The former issue is about the way in which the constituent parts of a complex system are interrelated with each other. It is noncommittal on the issue of whether or not there are forces or regularities by means of which a whole influences

the behavior of its parts. Collectivism is the claim that there are such social forces; this claim is not implied by social holism. Social holism can go with individualism: It can be conceived as being about the requirements under which a human being can develop into an individual personality (Pettit, 1993, Chaps. 3, 4); *see* Individualism versus Collectivism: Philosophical Aspects.

Arguments for Atomism and Holism

The main argument for social holism and the main challenge that social atomism has to meet stem from the work of the later Wittgenstein, in particular the *Philosophical Investigations* and the interpretation of that work by Kripke (1982). Kripke sets out what is known as the problem of rule-following, i.e., the problem of how the thoughts of finite thinking beings can have a determinate conceptual content. He raises the following issue: Any finite sequence of examples or of actions including linguistic actions of persons can in principle be described in terms of infinitely many different rules. Consequently, (1) there are infinitely many possible ways of continuing any such sequence in any new situation, and (2) for any new situation, it is not determined what is the correct way to go on. The problem of rule-following then is this one: What can select for us one of infinitely many rules so that for us one way of going on is singled out as the correct one? *See* Meaning and Rule Following.

According to Kripke's interpretation of Wittgenstein, to the extent that there is rule-following at all, social interactions are necessary in order to (1) select for a person one of infinitely many meanings in a finite sequence of examples or of actions and (2) enable a person to have a distinction between correct and incorrect rule-following at their disposal. This is the main argument for social holism. This argument speaks against the attempt to reduce the description of thoughts to a description in terms of the natural sciences.

There are two main ways in which social holism can be spelled out. One possibility is to conceive social holism in terms of an asymmetric I–we relation: An individual becomes a thinking being by being integrated into a social community. The members of the community assess the actions of the individual according to what they take to be fit and proper. In this case, norms are reduced to social facts: The fact of community approval or disapproval is identified with correctness or incorrectness. The actions on which the members of a community agree are not themselves liable to an assessment as being correct or incorrect. This position consequently is a social relativism. Its most extreme version is a radical meaning finitism that considers the extension of a concept only to be determined insofar as factual communal agreement reaches at a time (Bloor, 1997, Chap. 2).

The other possibility is to conceive social holism in terms of symmetric I–thou relations between individuals assessing each other's actions as correct or incorrect. This conception does not have to identify norms with social facts. Instead, it can regard these assessments as an open-ended process. Each normative attitude of taking something to be correct or incorrect, including those attitudes that are shared by all the members of a community at a time, can be challenged as not being correct. This position makes it possible to conceive the normative attitudes of taking something to be correct or incorrect as being responsive and responsible to the way the world is. It thus combines social holism with a realism (Pettit, 1993, Chap. 4; Brandom, 1994, Chaps. 1, 8; Hattiangadi, 2007).

If one intends to defend social atomism, one has to meet the challenge that Kripke poses. Two strategies can be distinguished. One strategy is to make a case for meanings as entities beyond the physical, including persons and their interactions (Katz, 1990, Chaps. 2–4). Having thoughts with a determinate conceptual content then does not depend on social relations, but on standing in an appropriate relation to such entities, i.e., grasping meanings. The other, most widespread strategy is to argue for some sort of a naturalization of semantics, i.e., an account of meaning in the vocabulary of the natural sciences. One then has to show how an account in terms of dispositions to behavior can solve both the infinity and the normativity problem that Kripke raises. Prominent proposals are an account in terms of computation (Miscevic, 1996), a teleofunctional account (Millikan, 2005), and an account within an ontology of causal, dispositional properties in general (Heil, 2003, Chap. 18). According to these accounts, certain dispositions that each human being has considered in isolation from other human beings are sufficient for a human being to have thoughts with a determinate conceptual content – although, of course, social relations influence the way in which the thoughts of a person develop.

Social holism is a social externalism: The thoughts of a person are externally individuated, namely by the person's relations to her social environment. Social atomism, by contrast, is an internalism: The thoughts of a person are individuated by factors that are internal to that person, such as her dispositions to behavior. Independently of Kripke's argument, Burge (1979) makes a case for social externalism by claiming that the meaning of the thoughts of a person depends on the way in which the terms in which that person expresses her thoughts are used in her social community. Burge's method is to develop thought experiments in which he keeps all the factors that are internal to a person fixed and varies the social environment. He thereby intends to show that the social environment enters into fixing the meaning of the thoughts of a person. Burge's argument fits into the method that dominates the discussion on externalism in general. A common reply to the externalist arguments is to distinguish between narrow and wide content of thoughts: The externalist argument is granted as far as wide content is concerned. But it is claimed that there is a narrow content of thoughts which is internally individuated and which is all that is needed to explain behavior. Thus, social atomism is vindicated as far as narrow content is concerned.

Social holism can be combined with semantic holism if one maintains that social relations determine a content for a thought of one type only by determining a content for other types of thoughts as well that constitute an inferential context for the thought in question (Brandom, 1994; Weiss et al., 2010). However, one does not have to combine semantic holism in form of an inferential semantics with social holism. The most widespread version of semantic holism is an inferential semantics applied to narrow content (Block, 1986). Hence, a semantic holism can go with a social atomism. Moreover, subsequent to the influential criticism of semantic holism by Fodor and Lepore (1992), arguments for semantic

atomism as well as semantic molecularism have gathered new momentum (Devitt, 1995).

By contrast, holism with respect to the confirmation of thoughts or sentences is widely accepted owing to Quine (1951). Confirmation holism is one of the backgrounds of the view of science of Kuhn (1962), which is also influential in the social sciences. It is in dispute to what extent confirmation holism and Kuhn's view of science lend support to social relativism. Because of the presumed relativistic consequences, Quine (1969) rejects Kuhn's view of science and excludes observation sentences from both confirmation and semantic holism, defending atomism with respect to these sentences.

See also: Autonomy, Philosophy of; Collective Beliefs, Sociological Explanation of; Collective Identity; Collectivism and Individualism: Cultural and Psychological Concerns; Colonialism, Anthropology of; Hegel, Georg Wilhelm Friedrich (1770–1831); Individualism versus Collectivism: Philosophical Aspects.

Bibliography

Block, N., 1986. Advertisement for a semantics for psychology. In: French, P.A., Uehling, Jr., H.K., Wettstein, T.E. (Eds.), Studies in the Philosophy of Mind. Midwest Studies in Philosophy, vol. 10. University of Minnesota Press, Minneapolis, MN, pp. 615–678.

Bloor, D., 1997. Wittgenstein, Rules and Institutions. Routledge, London.

Brandom, R.B., 1994. Making It Explicit. Reasoning, Representing, and Discursive Commitment. Harvard University Press, Cambridge, MA.

Burge, T., 1979. Individualism and the mental. In: French, P.A., Uehling, Jr., H.K., Wettstein, T.E. (Eds.), Studies in Metaphysics. Midwest Studies in Philosophy, vol. 4. University of Minnesota Press, Minneapolis, MN, pp. 73–121.

Correia, F., 2008. Ontological dependence. Philosophy Compass 3, 1013–1032.

Devitt, M., 1995. Coming to Our Senses. A Naturalistic Program for Semantic Localism. Cambridge University Press, Cambridge, UK.

Durkheim, E., 1927. Les règles de la méthode sociologique. University of Chicago Press, Chicago, IL. Alcan, Paris 8 [1938 The Rules of Sociological Method. Trans. Solovay S A, Mueller J H].

Esfeld, M., 1998. Holism and analytic philosophy. Mind 107, 365–380.

Fodor, J.A., Lepore, E., 1992. Holism. A Shopper's Guide. Blackwell, Oxford, UK.

Hattiangadi, A., 2007. Oughts and Thoughts. Rule-Following and the Normativity of Content. Oxford University Press, Oxford, UK.

Hegel, G.W.F., 1991. Elements of the Philosophy of Right. Cambridge University Press, Cambridge, UK (Trans. Nisbet H B).

Heil, J., 2003. From an Ontological Point of View. Oxford University Press, Oxford, UK.

Herder, J.G., 1993. Against Pure Reason. Writings on Religion, Language, and History. Fortress Press, Minneapolis, MN (Trans. Bunge M).

Katz, J.J., 1990. The Metaphysics of Meaning. MIT Press, Cambridge, MA.

Kripke, S.A., 1982. Wittgenstein on Rules and Private Language: An Elementary Exposition. Blackwell, Oxford, UK.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago, IL.

Ladyman, J., Ross, D., 2007. Every Thing Must Go. Metaphysics Naturalised. Oxford University Press, Oxford, UK.

Mead, G.H., 1934. Mind, Self, and Society. From the Standpoint of a Social Behaviorist. University of Chicago Press, Chicago, IL.

Millikan, R.G., 2005. Language: A Biological Model. Oxford University Press, Oxford, UK.

Miscevic, N., 1996. Computationalism and rule following. Proceedings of the Aristotelian Society 96, 215–229.

Pettit, P., 1993. The Common Mind. An Essay on Psychology, Society, and Politics. Oxford University Press, Oxford, UK.

Quine, W.V.O., 1951. Two dogmas of empiricism. Philosophical Review 60, 20–43.

Quine, W.V.O., 1969. Epistemology naturalized. In: Quine, W.V.O. (Ed.), Ontological Relativity and Other Essays. Columbia University Press, New York, NY, pp. 69–90.

Simons, P.M., 1987. Parts. A Study in Ontology. Clarendon Press, Oxford, UK.

Spencer, V., 1996. Towards an ontology of holistic individualism: Herder's theory of identity, culture and community. History of European Ideas 22, 245–260.

von Bertalanffy, L., 1968. General System Theory: Foundations, Development, Applications. Braziller, New York.

Wagner, G., 2012. Function and causality. Revue internationale de philosophie 66, 35–55.

Weiss, B., Wanderer, J. (Eds.), 2010. Reading Brandom. On "Making It Explicit". Routledge, London.

Attachment During Infancy and Early Childhood: Cultural Variations, Development of

Hiltrud Otto, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents an overview of ethological attachment theory as conceptualized by John Bowlby, characterizing the development of attachment relationships during infancy as biologically preprogrammed capacity. It depicts various empirical approaches to measure differences in attachment relationships and presents the ongoing scientific discussion on the psychobiological processes underlying different attachment organizations. The end of the article summarizes major challenges to the original formulation of attachment theory posed by cultural psychology, anthropology, and sociobiology, which led to innovations of the theory, generating a context-informed approach to attachment.

Introduction: Attachment Theory

The psychoanalyst and pediatrician John Bowlby (1990[†]) is considered the founder of attachment theory. After World War II, he practiced at the children's ward at Tavistock Clinic in England, when he was asked to provide a report on behalf of the World Health Organization on the situation of homeless and orphaned children in this period of time. In this report, he described the effects of childhood separation and deprivation he observed on a daily basis and postulated that caregiving environments influence the mental health of children. He argued that psychology was in need of a theory that actually accounts for the developmental consequences of early socioemotional experiences. In the following years, his ideas turned into attachment theory, which became one of the most influential theories of developmental psychology to this day (Bowlby, 1969, 1973, 1980).

Situating Attachment Theory: Sociohistorical Circumstances

Until the 1960s, best practice for child rearing stipulated providing adequate food and hygiene for babies, while keeping physical contact at a minimum to diminish the likelihood of the transmission of diseases, and to not spoil a child (Pauen, 2006). Preverbal infants were considered as creatures without feelings, and emotions between adults and infants did not play a role. Parental guidebooks from the first half of the twentieth century recommended to "place the child in a separate room in which it remains alone" (Haarer, quoted by Chamberlain, 2000: p. 26). The devastating consequences of this child care concept became obvious particularly in orphanages or hospitals where children often showed massive socioemotional disturbances. The observation of these developmental consequences made Bowlby challenge the prevailing principles of child care.

The Roots of Attachment Theory

Drawing on evolutionary theory as outlined by Charles Darwin who stated that natural selection favored behavioral systems, which proved adaptive (Darwin, 1860), Bowlby assumed that the attachment system also had *evolutionary roots*. According to Bowlby, attachment increased the likelihood of infant survival in the environment in which our ancestors lived ('environment of evolutionary adaptedness,' Bowlby, 1969). Children displaying attachment behaviors such as crying or following were more likely to stay in contact with their caregivers, and thereby heightened their chances of survival compared to infants not showing attachment behaviors. As attachment increased individuals' survival and reproduction rates, it became consolidated in the human genome.

Bowlby also incorporated findings from *ethological research*: The phenomenon of imprinting observable in animals reminded Bowlby of attachment behaviors in humans. Imprinting is known in ethology as an irreversible form of learning (Lorenz, 1943): During a relatively short, genetically determined time period ('critical period'), stimulus-response patterns become permanently incorporated in the behavioral repertoire of animals. Lorenz demonstrated that incubator-hatched geese would imprint on the first stimulus they saw within the first hours after hatching, henceforward following this stimulus unconditionally. In natural environments, this stimulus is typically the mother, but in experiments, Lorenz could show that the goslings would also imprint on him. Bowlby assumed that there is a similar mechanism at work for attachment formation in humans: there is a critical period for humans to develop attachments (i.e., the first year of life) when children rely on innate reflexive behaviors (e.g., smiling, crying, following) to form attachments. However, Bowlby assumed that in humans, fixed reflexive behaviors turned into '*open genetic programs*' in the course of phylogeny, and are thereby – within certain biological limitations – way more flexible. In terms of attachment, this means that the shape of attachment relationships does not only depend on genetic programs, but also on various environmental factors. Thereby the attachment system is flexible enough to adapt to different conditions.

In the 1950s, Harry Harlow, a *primate researcher*, explored the basics of mother–child love in primates in the United States and could disprove the assumption of learning theorists that attachment is the result of classical conditioning (Harlow, 1958/1997). He raised young rhesus monkeys with two 'surrogate mothers': one dummy mother was made of wire and provided milk for the young monkeys, another dummy

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23166-8

mother was covered with cloth, but provided no milk. Harlow could show in his experiments that the monkeys stayed with the lactating wire dummy mother only for the time of feeding, but otherwise cuddled with the fabric-covered dummy mother. Put under stress, the young monkeys always fled to the fabric-covered dummy mother. Harlow thereby proved that a cuddly dummy mother is emotionally more satisfying for a young rhesus monkey than a wiry food source.

Due to his training in *psychoanalysis*, Bowlby was well aware of the importance of the first years of life for later development. However, unlike Freud who believed that drives determine our psyche from early childhood on, Bowlby thought that early socioemotional experiences influence the future development of children. Children do not simply mature to become a person: they are shaped by their early social environment, especially by interactions with their main caregivers.

As attachments develop in interactions, Bowlby used *systems theory* to describe the dyadic and dynamic processes of relationship formation (Bischof, 1975). He conceptualized attachment relationships as feedback loops with the goal of establishing physical and/or psychological safety of the child. If stressed, the child displays attachment behaviors, which trigger caregiving behaviors in parents; these caregiving behaviors help the child to regain physical and/or psychological safety. Safety in the presence of a protective caregiver allows a child to freely explore the environment. The attachment system thus faces a complementary behavioral system: the '*exploratory system*.' The two systems function antagonistically: if one is active, the other is disabled and vice versa.

Bowlby's Concept of Attachment

Bowlby defined attachment as an emotional bond – '*a child's tie to his mother*' – that develops during the first year of life (Bowlby, 1969). He assumed that each infant has an innate attachment motivation, which is equal to other basic human needs, e.g., hunger or thirst. This natural attachment motivation evokes attachment behaviors that help to establish proximity to caregivers. The actual behaviors vary with age and motor abilities of the child: smiling, crying, clinging, crawling, and following are some of them. Caregivers usually react intuitively to these attachment behaviors with nurturing or caregiving behaviors. According to Bowlby (1969), attachment develops in four phases:

- During the first 6 weeks of life ('the preattachment period'), infants make little distinctions between persons, who can be exchanged almost arbitrarily.
- In the following 6 months ('attachment in the making'), infants direct signals to particular persons and start showing preferences for these persons.
- Once children are able to move, they actively seek closeness to caregiver(s), showing distinct attachment behaviors ('clear-cut attachment'). With 12 months, individual attachment relationships originated in the context of daily interactions, and children developed expectations about the accessibility and reliability of their caregiver(s) in case of insecurity and danger.
- From about the age of three, children form a more complex partnership ('goal-corrected partnership') with their caregiver(s) on the basis that they begin to see others as independent persons with own goals and feelings.

Bowlby postulated a hypothetical construct, the so-called internal working model (IWM), in which early attachment experiences are represented as a system of thoughts, memories, beliefs, and expectations about the self and others. IWMs can be defined as: "a set of conscious and/or unconscious rules for the organization of information relevant to attachment and for obtaining or limiting access to that information, that is, to information regarding attachment-related experiences, feelings, and ideations" (Main et al., 1985: pp. 66–67). Accordingly, experiences with a reliable mother should result in a positive IWM of self and caregiver, experiences with an unpredictable, unreliable attachment figure in a negative IWM. Though IWMs are formed during the first year of life, attachment theory assumes that they fundamentally influence our future socioemotional development and remain influential over the life span.

Empirical Approaches to Attachment

Mary Ainsworth and her colleagues made attachment theory accessible to empirical research (Ainsworth, 1967; Ainsworth et al., 1978). During a stay in Uganda, Africa, Ainsworth observed individual differences in mother–infant pairs: Separated from their mothers and in the presence of the strange researcher, the majority of children showed distinct attachment behaviors, a second group of children displayed exaggerated or inconsistent attachment behaviors and a third group of children did not show any attachment behaviors at all. She called these three groups of children 'securely attached,' 'insecurely attached,' and 'nonattached,' respectively.

Back in the United States, Ainsworth wanted to replicate her findings with families from Baltimore, but realized that 1-year-old US children were not alarmed when separated from their mothers; obviously, the US children were accustomed to be temporarily separated from their mothers and also to the presence of unfamiliar people in their homes. Thus, Ainsworth invited mothers and their children to a 'foreign' laboratory at the university, creating the paradigm of the '*strange situation procedure*' (SSP), a mini drama with fixed sequences to elicit attachment and exploratory behaviors in children (Ainsworth et al., 1978). During the SSP, the child is exposed to seven 3-min episodes that become increasingly stressful: In the first episode, mother and child are alone in the room; in the second episode, a stranger appears and starts to interact with the child; in the third episode, the child is left alone with the stranger; in the fourth episode, the mother returns and the stranger leaves; in the fifth episode, the mother exits for the second time, leaving the child completely alone; in the sixth episode, the stranger returns; and in the seventh episode, the mother finally returns. On the basis of their separation and reunion behaviors with the mother and their interaction with the stranger, children are classified into the four categories:

- *Securely attached children* (*B*) show their emotions openly: They articulate their fear and sadness in the separation situation, show their joy and elation in the reunion with

their mothers, and calm down quickly. They are able to use the presence of the mother as a 'secure base,' from which they explore the environment and interact with the stranger.

- *Insecure–avoidant children* (*A*) do not show their emotions openly but avoid expressions of attachment. They seem unimpressed by the separation from the mother, and ignore their mother in the reunion situation. Avoidant children prefer to interact with the stranger and the interactions are predominated by exploratory behaviors.
- *Insecure–ambivalent children* (*C*) show attachment behaviors constantly with anger being the dominant emotion. They seem overwhelmed by the pain of separation, do not calm down in the reunion with the mother and hardly interact with the stranger, even in the presence of the mother.
- *Disorganized children* (*D*) represent a fourth category that was added by Main and Solomon in the 1980s (Main and Solomon, 1986). Children in this group show bizarre behaviors and repetitive movements. Disorganized attachment is conceptualized as the lack of a coherent behavioral strategy for dealing with stress.

Using the SSP, Ainsworth and colleagues classified 106 US children, and the distributional pattern of 66% secure, 22% avoidant, and 12% ambivalent children is referred to as the 'American Standard distribution' (Ainsworth et al., 1978). Moreover, Ainsworth found correlations between the attachment security of children and the quality of their mothers' interactive behaviors: Children with mothers who responded sensitively to their children's needs were more likely to be securely attached, whereas children with less sensitive mothers were more likely to be insecurely attached. The quality of maternal interactive behavior was defined by Ainsworth et al. (1978) as *maternal sensitivity* and captured the mothers' ability to

- perceive the child's signals,
- interpret the signals correctly,
- respond adequately, and
- respond promptly.

The correlations between maternal sensitivity and infant attachment organization demonstrate the plasticity of the attachment system: secure children have learned that their signals are perceived and interpreted correctly, so they rely on help from their caregiver and communicate their feelings. Avoidant children have learned to suppress their emotions, as the mothers of these children are stressed by negative signals. Insecure–ambivalent children have mothers who are not consistent in their behaviors; this unpredictability puts children in the situation to have a notoriously activated attachment system. Disorganized children lack a coherent strategy altogether, which is often the product of abuse through the caregiver. In these cases, the attachment figure, who is supposed to provide protection, becomes a threat, which puts the child in a 'double bind situation' with no way out.

The *core assumptions of attachment theory* as established by Bowlby and Ainsworth, can be summarized as follows:

1. *Universality*: All healthy newborns develop attachment relationships with their caregiver(s) during their first year of life.
2. *Normativity*: The majority of children develop secure attachment relationships; insecure attachment patterns are less frequent.
3. *Sensititvity*: Maternal sensitivity is an important antecedent for attachment security.
4. *Competence*: Securely attached children are more competent in the solution of further development tasks.

Psychobiology of Attachment

The attachment system aims at maintaining or restoring homeostasis in the human organism (Bowlby, 1969). On the behavioral level, being reunited with the caregiver serves this goal. On the biological level, physiological systems also aim at the maintenance of homeostasis; here regulation takes place continuously, e.g., by increases/decreases in blood pressure or heart rates. The attachment system and the physiological systems are primarily independent, but can be activated together, e.g., during a stress response. From the 1980s on, the interplay of these two systems has been investigated by various disciplines, such as neuroscience, biology, and neurophysiology.

In humans, the activation of two physiological systems during a stress response can be observed noninvasively: The cardiovascular system responds to changes in heart rate, which can be measured easily (Spangler and Schieche, 1998); the pituitary adrenocortical system releases the stress hormone cortisol, which is detectable after a brief time delay in saliva (Chryssanthopoulou et al., 2005). Cortisol became especially prominent in attachment research because cortisol measurements can be used before and after stressful situations, thereby indicating the degree of a stress experience (e.g., Gunnar, 2005). To date, empirical findings concerning the interplay of attachment behavioral strategies and adrenocortical activation produced no clear picture yet. Instead, two different theoretical models are discussed in the literature: the 'coping model' and the 'arousal model.'

The coping model assumes that different attachment qualities equip children with different coping strategies for regulating their emotions. This model is supported by studies (e.g., Spangler and Grossmann, 1993) reporting that infants with insecure attachments – especially those with disorganized attachments – showed higher cortisol levels following the SSP, which is interpreted as proving the suitability of the secure attachment strategy to deal with stress. In contrast, the arousal model assumes that only infants exhibiting behavioral stress in terms of crying or negative emotions have a heightened adrenocortical activity. This model is supported by studies that reported a fundamental increase of adrenocortical activation in ambivalent and disorganized infants (e.g., Hertsgaard et al., 1995). However, many studies report no differences in the adrenocortical response between the insecure attachment groups (e.g., Gunnar et al., 1996), or found cortisol increases only in insecure infants who were additionally characterized by behavioral inhibition (e.g., Nachmias and Gunnar, 1996). These studies suggest that both temperamental characteristics and attachment quality may influence the function of the adrenocortical system, whereby attachment security may function as a moderator for the activity of the hypothalamic-pituitary-adrenal (HPA) system.

Although the biobehavioral organization of attachment is not completely understood yet, it is obvious that the interplay of the behavioral and the adrenocortical system differs between the attachment groups. Research has revealed the immense complexity of the psychobiology of attachment and future research might provide definite answers to the question concerning the mechanisms of the biobehavioral organization of attachment.

Intergenerational Continuity of Attachment

Another unsolved question is if and how attachment styles are passed over generations. Van IJzendoorn (1995) proposes in the so-called *transmission model* that parental behaviors have their roots in childhood experiences. Thus, securely attached children should become better parents (Main, 1990). However, while children's IWMs can be assessed in the Strange Situation on a behavioral level, the assessment of adults' IWMs requires a move to the level of representation, because a direct correspondence between behavior and mental structures is no longer a given in adulthood. Therefore, the '*adult attachment interview*' (AAI; Main et al., 1985; Main and Goldwyn, 1998), a semistructured interview, was devised by with the idea to assess adults' state of mind with respect to attachment as expressed in discourse about early relationships. Interestingly, the content of these autobiographical memories did not become the main indicator for an adult's attachment classification; instead, communication styles proved more illuminative, especially the concept of coherence. Coherence is based on the four the so-called Gricean maxims: be truthful and have evidence for what you say (*quality*); be succinct, yet complete (*quantity*); be relevant (*relevance*); and be clear, brief, and orderly (*manner*). During the AAI, an interviewee is faced with the task of producing and reflecting upon memories related to attachment while maintaining a coherent discourse with the interviewer. AAI classifications refer to differences in discourse style, four discourse patterns have been identified:

- *Secure–autonomous* adults (*F*) provide a discourse that is open and coherent, presenting even difficult attachment experiences in clear and vivid ways. The discourse shows no contradictions, the focus is on the conversation and interviewees show a rich use of language.
- *Dismissing* adults (*D*) provide extremely brief interviews that are characterized by striking contradictions in the interviewee's discourse about early attachments, with generalized representations of history being unsupported. A strong idealization of caretakers is common, along with contradictory memories of events.
- *Preoccupied* adults (*E*) talk excessively long and include information that is irrelevant to the discourse. Interviewees are not able to describe their attachment biography coherently or to move beyond an excessive preoccupation with attachment relationships. Their discourse is full of passive speech, unfinished sentences, and vague expressions.
- *Unresolved* adults (*U*) show lapses in reasoning, especially when discussing traumatic events involving loss or abuse. For example, they may indicate the belief that a dead person is still alive, lapse into prolonged silences, or enter a trancelike state of mind.

Theoretically, the AAI was supposed to shed light on the intergenerational transmission of attachment, e.g., adults with a secure–autonomous state of mind should be the better, i.e., more sensitive parents; however, meta-analyses could not prove these assumptions empirically (Van IJzendoorn, 1995; De Wolff and Van IJzendoorn, 1997). Instead, it became obvious that other aspects besides sensitivity play an important role in the development of attachment security (Madigan et al., 2006). Measures capturing dyadic dimensions of caregiver–child interactions, e.g., the 'scale of reciprocity' (De Wolff and Van IJzendoorn, 1997), and the inclusion of contextual factors, e.g., the family's socioeconomic status or the child's temperament (Belsky and Fearon, 2008) proved informative. Moreover, mentalization concepts were considered a possibility to close to transmission gap. Meins (1997) could show that mothers who were able to verbalize the inner psychological states of their children were likely to have securely attached children. She called this ability of mothers to put themselves into the mental state of the child 'mind-mindedness.' Alike measures, e.g., the 'reflective functioning scale' (Fonagy et al., 1998) capturing a caregiver's ability to reflect the mental state of the child, revealed correlations with the child's attachment security. But though research on attachment increased tremendously, the transmission of attachment cannot be explained entirely. The ongoing discussion in the literature is referred to under the term 'transmission gap' (Van IJzendoorn, 1995).

Methods and Problems of Attachment Research

The buzz around attachment research brought a variety of new tools and methods with it. Besides *behavioral observations* of infants, e.g., using the SSP or the Attachment Q-sort (Waters and Deane, 1985) and *interview procedures* such as the AAI or the Child Attachment Interview (Target et al., 2003), there are mainly two more classes of methods applied in attachment research: Projective methods and questionnaires.

Projective methods use ambivalent material, such as pictures or stories, to elicit unconscious projections. Projections are supposedly influenced by attitudes, motives, and wishes of a person, they can be interpreted by trained evaluators and thereby obtain diagnostic value. Frequently used projection methods are the 'separation anxiety test' (Klagsbrun and Bowlby, 1976; Kaplan, 1987) and the so-called attachment story completion task (Bretherton et al., 2001).

In contrast, *questionnaires* aim at the self-assessment of the subjects. The first questionnaire to assess attachment styles was formulated by Hazan and Shaver (1987). In their 'attachment self-report,' they designed distinctive vignettes for each attachment style and asked their participants to self-assess their attachment style. A variety of different questionnaires was developed over time, e.g., the 'attachment style measure' (Simpson, 1990), the 'adult attachment scale' (Collins and Read, 1990), or the 'adolescent attachment questionnaire' (West et al., 1998).

As attachment research offered more convenient tools of measurement, attachment research became highly attractive for

many researchers, and till today thousands of papers on attachment have been published. However, the empirical study of attachment comprises a serious method problem as different conceptualizations of attachment become apparent in the different methods: Empirically, correlations between attachment patterns identified with different methods are disappointingly low or trivial (e.g., Roisman et al., 2007). This inconsistency of classifications renders it highly unlikely that the different methods capture indeed the same construct. Instead, it indicates a fundamental lack of clarity in the conceptualization of attachment.

Cultural Variations

In Bowlby's original formulation of attachment theory (Bowlby, 1969), an important feature of the attachment system was its flexibility to adapt to various environmental conditions. For children during their first year of life, social experiences with caregivers constitute a major environmental influence. Moreover, the fact that different experiences result in different attachment patterns documents the plasticity of the attachment system nicely. However, cross-cultural studies demonstrated that early social interactions do not only vary on an individual level, but are also strongly influenced by the ecocultural context within which they take place (Keller, 2007): Parents raise children differently in different cultural environments in order to equip them with the necessary competencies to become successful in a specific environment. This also impacts the development of attachment formation.

Early Pitfalls of (Cross-)Cultural Attachment Research

Unfortunately, the majority of cross-cultural studies on attachment did not take culture seriously for many years and instead relied on Ainsworth's paradigm of the SSP, arguing that the procedure itself represents a culture-sensitive method as it was developed by Ainsworth after research in Africa (Van IJzendoorn, 1990; Van IJzendoorn et al., 2006). This assumption proves incorrect as a close examination of Ainsworth's publications reveals that the invention of the SSP resulted from the very fact that one method, namely natural observations during house visits, did not allow the observation of attachment behaviors in both African and American children. The invention of the SSP resulted as a necessary adaptation of observational methods within the African and US contexts. Albeit cultural realities caused the invention of the SSP, it spread worldwide without accounting for cultural differences other than distributional differences in attachment patterns, which were only post hoc explained as cultural differences. The secure attachment pattern was regarded as the optimal attachment strategy everywhere using the circular argument that secure behavior represents the norm because the majority of infants are characterized as securely attached in the SSP worldwide (Van IJzendoorn and Kroonenberg, 1988). The two insecure patterns were both regarded as less optimal coping strategies, and the disorganized pattern as an extremely disadvantageous strategy (DeOliveira et al., 2004; Main and Solomon, 1990).

Mother–Infant Relationships from a Historical and Anthropological Perspective

Today, the intrinsic role of culture in developmental processes is increasingly acknowledged: Culture represents the primary mode of adaptation to the demands of a particular ecocultural context or niche, thereby creating shared cultural practices and shared meaning systems (Greenfield and Keller, 2004; Super and Harkness, 2002). Regarding development, culture can be considered as the necessary software that gives adaptive direction to the neurophysiological hardware, the inherent capacities all newborns possess (LeVine et al., 1994). Consequently, cultural environments and the biological nature of human development are strongly interrelated (Greenfield, 2002). This view also applies to the study of attachment, which brings reinterpretations and innovations.

Data of Ainsworth's original SSP were collected in a distinct Western population, so infants' reactions should be interpreted as reflecting variations of norms of social distance and interpersonal responsiveness of that historical time (LeVine and Norman, 2001); and not as universal norms per se. This sociohistorical perspective clearly challenges the validity of the American Standard distribution as setting a universal benchmark. Similarly, invoking a historical perspective reveals that in former times the phase of infancy did not resemble the ideas of attachment theory about infancy: In the past, mothers were unlikely to pay a lot of attention to their babies, instead, they often relied on a 'wait-and-see' strategy toward the newborn. This strategy was caused by high infant mortality rates that made infant survival unpredictable and early investment in babies uncommon. Moreover, childbirth and breast-feeding physically weakened a mother's health to the point that it threatened maternal survival. Yet a mother was far more important to the family than a baby, maternal health had always primacy to a baby's well-being, which gained infants a marginal status at the very beginning of their lives (Lancy, in press).

There is also vast anthropological evidence that attachment theory's original formulation was more a cultural ideology, not a universally applicable theory (for a summary see LeVine, in press): In non-Western societies, mothers often lack the resources for exclusive caretaking due to a heavy workload and a high number of children, so caregiving is socially distributed and also not an adult responsibility per se (Weisner and Gallimore, 2008). For example, Aka infants and young children in the Central African Republic can be observed to interact with just over 20 caregivers each day, and half of these caregivers invest in intensive forms of caregiving (Meehan, 2009). Similar observations are reported for Beng infants from Cote d'Ivoire (Gottlieb, 2004), for Nso infants from Cameroon (Keller and Otto, 2009; Yovsi, 2003), and for Gusii children from Kenya (LeVine et al., 1994). These insights clearly question the monotropic understanding of relationships that are prevalent in attachment theory's assumptions, but adhere only to Western nuclear families. Alternatively, sociobiologists (e.g., Hrdy, 2005) put forward the so-called cooperative breeding model as the reproductive strategy in ancient human history where alloparents allowed mothers to redistribute the cost of raising offspring (Hrdy, 1999). This view is supported from an evolutionary

viewpoint: living in social units offered advantages such as a lesser probability to be attacked, a better defense against aggressors, and more efficient food acquisition (Voland, 2000). And while nuclear families could survive and reproduce, they were likely to fall short compared to families that can rely on a more extensive network of support (Mace and Sear, 1997, 2006). If our ancestors lived in larger social units, attachment relationships would have developed not only with mothers, but with many social others that served similar functions (Belsky, 1997; Porges, 2003; Trivers, 1996). Multiple diverse attachment relationships may indeed have constituted the reality for most of our ancestors' children.

Culture-Informed Attachment Research

A side effect of using the SSP across different cultural contexts was that researchers sidestepped the big question: *Are attachment processes essentially the same or significantly different across cultures?* The answer to this question seems to be that attachment indeed varies across cultures. Sagi wrote: "the repertoire of attachment behaviors is similar across countries, but the selection of these behaviors is culturally specific" (Sagi, 1990: p. 19). Differences concerning the development of attachment relationships have been found on three different levels: (1) *parental socialization goals*, which vary in accordance with sociocultural environments, (2) inform actual *parenting behaviors* in interactions with infants, and (3) lead to the formation of culture-specific, ideal *attachment strategies in children*.

1. A study focusing on maternal *socialization goals* and perceptions of attachment compared Anglo and Puerto Rican working- and middle-class mothers in the United States (Harwood et al., 1995). Harwood et al. used vignettes of Strange Situation behaviors, revealing that maternal perceptions of desirable and undesirable attachment behaviors differed from the labels given to them by attachment theory. Mothers could prefer behaviors coded as insecure, and disapprove behaviors coded as secure, based on the SSP ratings. Similar results were found with regard to attachment relationships in the Japanese culture where the indigenous Japanese concept of relatedness, *amae*, involves insecure–ambivalent behaviors (Rothbaum and Kakinuma, 2004). Also, Cameroonian Nso mothers were found to dislike the idea of developing an exclusive relationship with their infants (Otto, 2008). Nso children crying openly upon being separated from the mother were regarded as naughty; in contrast, mothers valued expressionless infants. Nso mothers trained their children to be used to others because "*one mother alone cannot take care of an infant*."
2. Actual *maternal behaviors* toward infants vary across cultures: Whereas sensitivity may be regarded as optimal parenting and represent a precursor of attachment security in the Western world, maternal behavioral strategies look drastically different in non-Western cultures: For example, Cameroonian Nso mothers prefer a controlling parenting style that would be regarded as intrusive parenting following attachment theory (Yovsi et al., 2009). Nso mothers consider themselves as experts in raising their children, so they actively structure interactions, which often includes interfering with the babies' activities. To consider babies as persons with their own will and their own mind is unheard of in many non-Western context, e.g., in the Cameroonian Nso, but only then does it make sense to address them in a mind-minded way as do Western parents. Societies in the Pacific represent an extreme as they claim a fundamental unknowability of other minds in general, which is referred to as the 'opacity doctrine' by anthropologists (e.g., Duranti, 2008; Schieffelin, 2008). In this case, talking to babies to verbalize the baby's state of mind or emotions does not make any sense. Furthermore, research has shown that mothers are attuned differentially to positive or negative signals across cultures (Keller et al., 2008), so that today many researchers arrived at the conclusion that maternal sensitivity is expressed in cultural specific ways, depending on a variety of ecological and cultural constraints and incentives (Thompson, 1997; Thompson et al., 2005).
3. The investigation of *children's attachment strategies* across cultures needs an emic approach, but only few studies to date did not rely on the SSP to assess children's actual attachment patterns. For the Cameroonian Nso, the ideal strategy is characterized by the absence of emotional expressions and an adaptation to the caregiver structuring the interaction: 1-year-old Nso children were expressionless even when approached and picked up by a stranger (Otto, 2008), and they readily followed the stranger's initiatives (Otto, in press). Calm and inexpressive infants were found highly valued in other cultures as well, e.g., in a rural Iranian community (Friedl, 1997) or in the Yucatec Mayan community (Howrigan, 1988). These cultures are characterized by a heavy workload of mothers combined with high fertility rates, which forces mothers to rely on multiple caregiving arrangements. As expressionless infants can easily be taken care of by a variety of caretakers, this pattern proved adaptive for these contexts.

Principally, the variance of attachment strategies suggests that attachment patterns evolved due to the necessity of children to adapt to various circumstances. Therefore, different attachment strategies must be regarded as adaptations themselves, not coercively maladaptations (Belsky, 1999, 2006). Sociobiologists provide support for this view and argue that the steady frequency of the so-called insecure patterns found in studies using the SSP speaks against a genetically preprogrammed biological normativity of the secure pattern (1998). Consequently, attachment patterns must be evaluated against the context within which they develop. A possible outcome of this approach may be that strategies that do not resemble the secure attachment pattern still represent an 'optimal' strategy – when situated in their context.

Outlook

Attachment theory is one of the most influential theories of developmental psychology to date. However, the phenomenon of attachment requires a complex analysis devoid of Eurocentric assumptions. In theory, the attachment system is designed as a system that adapts to environmental conditions and empirically, this has been demonstrated on an individual

level in many cases. Taken to its logical end, attachment theory turns into a cultural theory that integrates cultural values and norms and is therefore 'context informed.' Empirically, 'serious' cultural attachment research would significantly increase our understanding of the cultural determinants and variations of attachment and would allow an interpretation of both cultural and individual differences.

From another angle, practitioners, e.g., therapists or social workers, also call for a context-informed attachment theory. A huge number of intervention and prevention programs are based on attachment theory's assumptions of normal and healthy development, but practitioners working with children and families from non-Western societies oftentimes realize a gap between the needs of their clients and the treatments they can offer. Theory and practice rely on different languages, practices, and ideas. Whereas theoretical knowledge is mainly derived from the so-called weird – *w*hite, *e*ducated, *i*ndustrialized, *r*ich, and *d*emocratic – populations (Henrich et al., 2010), research in the non-Western 'majority world' is still widely lacking; with the consequence that programs often do not fit for their target populations.

Researchers become more and more aware of these problems and work on closing the gap: Recently, scientists from various disciplines such as cultural and cross-cultural psychology, cultural anthropolgy, ethnology, linguistics, and evolutionary science started to bring together their multicultural expertise to show that attachment looks differently and has different developmental trajectories across contexts, highlighting the need and possibility to incorporate contextual variation systematically into attachment theory. Two edited volumes are in press today, one titled 'Different faces of attachment: Cultural Variations on a Universal Human Need' (Cambridge University Press) by Hiltrud Otto and Heidi Keller, the other one titled: 'Attachment Reconsidered: Cultural Perspectives on a Western Theory' (Palgrave Macmillan) by Naomi Quinn and Jeanette Mageo. It seems that the future will bring new conceptual developments to the seminal groundwork that John Bowbly and Mary Ainsworth already laid with their formulation of attachment theory.

See also: Childhood: Anthropological Aspects; Early Emotional Development and Cultural Variability; Emotional Development, Effects of Parenting and Family Structure On; Social and Emotional Development in the Context of the Family; Stress in Childhood: Effects on Development.

Bibliography

Ainsworth, M.D.S., 1967. Infancy in Uganda: Infant Care and the Growth of Love. Johns Hopkins Press, Oxford, England.

Ainsworth, M.D.S., Blehar, M.C., Waters, E., Wall, S., 1978. Patterns of Attachment: A Psychological Study of the Strange Situation. Lawrence Erlbaum, Oxford, England.

Belsky, J., 1997. Theory testing, effect-size evaluation, and differential susceptibility to rearing influence: the case of mothering and attachment. Child Development 68 (4), 598–600.

Belsky, J., 1999. Interactional and Contextual Determinants of Attachment Security. Guilford Press, New York.

Belsky, J., 2006. Determinants and Consequences of Infant-Parent Attachment. Psychology Press, New York.

Belsky, J., Fearon, R.M.P., 2008. Precursors of attachment security. In: Cassidy, J., Shaver, P.R. (Eds.), Handbook of Attachment. Theory, Research, and Clinical Application, second ed. Guilford Press, New York, pp. 295–316.

Bischof, N., 1975. A systems approach toward the functional connections of attachment and fear. Child Development 46 (4), 801–817.

Bowlby, J., 1969. Attachment. Basic Books, New York.

Bowlby, J., 1973. Separation: Anxiety and Anger. Basic Books, New York.

Bowlby, J., 1980. Loss, Sadness and Depression. Basic Books, New York.

Bretherton, I., Sueess, G.J., Golby, B., Oppenheim, D., 2001. "Attachment Story Completion Task" (ASCT) – Methode zur Erfassung der Bindungsqualität im Kindergartenalter durch Geschichtenergänzungen. In: Suess, G.J., Scheurer-Englisch, H., Pfeifer, W.-K.P. (Eds.), Bindungstheorie und Familiendynamik - Anwendung der Bindungstheorie in Beratung und Therapie. Psychosozial-Verlag, Gießen, pp. 83–124.

Chamberlain, S., 2000. Adolf Hitler, die deutsche Mutter und ihr erstes Kind. Über zwei NS-Erziehungsbücher. Psychosozial-Verlag, Gießen.

Chryssanthopoulou, C.C., Turner-Cobb, J.M., Lucas, A., Jessop, D., 2005. Childcare as a stabilizing influence on HPA axis functioning: a reevaluation of maternal occupational patterns and familial relations. Developmental Psychobiology 47 (4), 354–368.

Collins, N.L., Read, S.J., 1990. Adult attachment, working models, and relationship quality in dating couples. Journal of Personality and Social Psychology 58 (4), 644–663.

Darwin, C.R., 1860. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. John Murray, London.

DeOliveira, C.A., Bailey, H.N., Moran, G., Pederson, D.R., 2004. Emotion socialization as a framework for understanding the development of disorganized attachment. Social Development 13 (3), 437–467.

De Wolff, M.S., Van IJzendoorn, M.H., 1997. Sensitivity and attachment: a meta-analysis on parental antecedents of infant attachment. Child Development 68, 571–591.

Duranti, A., 2008. Further reflections on reading other minds. Anthropological Quarterly 18 (2), 483–494.

Fonagy, P., Target, M., Steele, H., 1998. Reflective-functioning Manual, Version 5.0, for Application to Adult Attachment Interviews. University College London, London.

Friedl, E., 1997. Children of Deh Koh: Young Life in an Iranian Village. Syracuse University Press, Syracuse, NY.

Gottlieb, A., 2004. The Afterlife Is Where We Come from: The Culture of Infancy in West Africa. University of Chicago Press, Chicago.

Greenfield, P.M., 2002. The Mutual Definition of Culture and Biology in Development. Cambridge University Press, New York.

Greenfield, P.M., Keller, H., 2004. Cultural psychology. In: Spielberger, C.D. (Ed.), Encyclopedia of Applied Psychology. Elsevier, Oxford, UK, pp. 545–553.

Gunnar, M.R., 2005. Attachment and Stress in Early Development: Does Attachment Add to the Potency of Social Regulators of Infant Stress? MIT Press, Cambridge, MA.

Gunnar, M.R., Brodersen, L., Nachmias, M., Buss, K., Rigatuso, J., 1996. Stress reactivity and attachment security. Developmental Psychobiology 29 (3), 191–204.

Hazan, C., Shaver, P.R., 1987. Romantic love conceptualized as an attachment process. Journal of Personality and Social Psychology 52, 511–524.

Harlow, H.F., 1997. The Nature of Love (1958). American Psychological Association, Washington, DC.

Harwood, R.L., Miller, J.G., Irizarry, N.L., 1995. Culture and Attachment: Perceptions of the Child in Context. Guilford Press, New York.

Henrich, J., Heine, S.J., Norenzayan, A., 2010. The weirdest people in the world? Behavioral and Brain Sciences 33 (2–3), 61–83.

Hertsgaard, L., Gunnar, M., Erickson, M.F., Nachmias, M., 1995. Adrenocortical responses to the strange situation in infants with disorganized/disoriented attachment relationships. Child Development 66 (4), 1100–1106.

Howrigan, G.A., 1988. Fertility, infant feeding and change in Yucatan. In: LeVine, R.A., Miller, P.M., West, M.M. (Eds.), Parental Behavior in Diverse Societies, New Directions for Child Development, 40, pp. 37–50.

Hrdy, S.B., 1999. Mother Nature: Maternal Instincts and How They Shape the Human Species. Pantheon, New York.

Hrdy, S.B., 2005. Evolutionary Context of Human Development: The Cooperative Breeding Model. MIT Press, Cambridge, MA.

Kaplan, N., 1987. The Sixth-Year Separation Anxiety Test Classification System (Doctoral dissertation). University of California, Berkeley, CA.
Keller, H., 2007. Cultures of Infancy. Lawrence Erlbaum Assoc., Mahwah, NJ, US.
Keller, H., Otto, H., 2009. The cultural socialization of emotion regulation during infancy. Journal of Cross-cultural Psychology 40 (6), 996–1011.
Keller, H., Otto, H., Lamm, B., Yovsi, R.D., Kaertner, J., 2008. The timing of verbal/vocal communications between mothers and their infants: a longitudinal cross-cultural comparison. Infant Behavior & Development 31 (2), 217–226.
Klagsbrun, M., Bowlby, J., 1976. Responses to separation from parents: clinical test for young children. British Journal of Projective Psychology 21, 7–21.
Lancy, D. "Babies aren't persons": a survey of delayed personhood. In: Otto, H., Keller, H., (Eds.), Different Faces of Attachment: Cultural Variation on a Universal Human Need. Cambridge University Press, Cambridge, UK, in press.
LeVine, R.A. Attachment theory as cultural ideology. In: Otto, H., Keller, H., (Eds.), Different Faces of Attachment: Cultural Variation on a Universal Human Need. Cambridge University Press, Cambridge, UK, in press.
LeVine, R.A., Dixon, S., LeVine, S., Richman, A., Leiderman, H.P., Keefer, C.H., Brazelton, B.T., 1994. Child Care and Culture: Lessons from Africa. Cambridge University Press, Cambridge, UK.
LeVine, R.A., Norman, K., 2001. The infant's acquisition of culture: early attachment reexamined in anthropological perspective. In: Moore, C.C., Mathews, H.F. (Eds.), The Psychology of Cultural Experience. Cambridge University Press, New York.
Lorenz, K., 1943. Die angeborenen Formen moeglicher Erfahrung. Zeitschrift fuer Tierpsychologie 5 (2), 235–409.
Mace, R., Sear, R., 1997. Birth interval and the sex of children in a traditional African population: an evolutionary analysis. Journal of Biosocial Science 29 (4), 499–507.
Mace, R., Sear, R., 2006. Are humans cooperative breeders? In: Voland, E., Chasiotis, A., Schiefenhoevel, W. (Eds.), Grandmotherhood - the Evolutionary Significance of the Second Half of Female Life. Rutgers University Press, Piscataway, NJ, pp. 143–159.
Madigan, S., Bakermans-Kranenburg, M.J., Van IJzendoorn, M.H., Moran, G., Pederson, D.R., Benoit, D., 2006. Unresolved states of mind, anomalous parental behavior, and disorganized attachment: a review and meta-analysis of a transmission gap. Attachment & Human Development 8 (2), 89–111.
Main, M., 1990. Cross-cultural studies of attachment organization: recent studies, changing methodologies, and the concept of conditional strategies. Human Development 33 (1), 48–61.
Main, M., Goldwyn, R., 1998. Adult attachment scoring and classification system (Unpublished manual). University of California, Berkeley, CA.
Main, M., Kaplan, N., Cassidy, J., 1985. Security in infancy, childhood, and adulthood: a move to the level of representation. Monographs of the Society for Research in Child Development 50 (1/2), 66–104.
Main, M., Solomon, J., 1986. Discovery of an insecure-disorganised/disoriented attachment pattern: procedures, findings and implications for the classification of behaviour. In: Brazelton, T.B., Yogman, M.W. (Eds.), Affective Development in Infancy. Ablex, Norwood, NJ.
Main, M., Solomon, J., 1990. Procedures for Identifying Infants as Disorganized/Disoriented during the Ainsworth Strange Situation. University of Chicago Press, Chicago, IL.
Meehan, C.L., 2009. Maternal time allocation in two cooperative childrearing societies. Human Nature 20 (4), 375–393.
Meins, E., 1997. Security of Attachment and the Social Development of Cognition. Psychology Press, Hove.
Nachmias, M., Gunnar, M., 1996. Behavioral inhibition and stress reactivity: the moderating role of attachment security. Child Development 67 (2), 508–522.
Otto, H., 2008. Culture-specific attachment strategies in the Cameroonian Nso: cultural solutions to a universal developmental task (Doctoral dissertation). University of Osnabrueck, Osnabrueck, Germany.
Otto, H. "Don't show your emotions!" Attachment in the Cameroonian Nso. In: Otto, H., Keller, H. (Eds.), Different Faces of Attachment: Cultural Variation on a Universal Human Need. Cambridge University Press, Cambridge, UK, in press.
Otto, H., Potinius, I., Keller, H., 2014. Cultural differences in stranger-child interactions: A comparison between German middle-class and cameroonian Nso dyads. Journal of Cross-Cultural Psychology 45 (2), 322–334.
Pauen, S., 2006. Was Babys denken: Eine Geschichte des ersten Lebensjahres. Beck Verlag, München.
Porges, S.W., 2003. Social Engagement and Attachment: A Phylogenetic Perspective. New York Academy of Sciences, New York.
Roisman, G.I., Holland, A., Fortuna, K., Fraley, R.C., Clausell, E., Clarke, A., 2007. The adult attachment interview and self-reports of attachment style: an empirical rapprochement. Journal of Personality and Social Psychology 92, 678–697.
Rothbaum, F., Kakinuma, M., 2004. Amae and attachment: security in cultural context. Human Development 47 (1), 34–39.
Sagi, A., 1990. Attachment theory and research from a cross-cultural perspective. Human Development 33 (1), 10–22.
Schieffelin, B., 2008. Speaking only your own mind: reflections on confession, gossip, and intentionality in Bosavi (PNG). Anthropological Quarterly 81 (2), 431–441.
Simpson, J.A., 1990. Influence of attachment styles on romantic relationships. Journal of Personality and Social Psychology 59, 971–980.
Spangler, G., Grossmann, K.E., 1993. Biobehavioral organization in securely and insecurely attached infants. Child Development 64 (5), 1439–1450.
Spangler, G., Schieche, M., 1998. Emotional and adrenocortical responses of infants to the strange situation: the differential function of emotional expression. International Journal of Behavioral Development 22 (4), 681–706.
Super, C.M., Harkness, S., 2002. Culture structures the environment for development. Human Development 45 (4), 270–274.
Target, M., Fonagy, P., Shmueli-Goetz, Y., 2003. Attachment representations in school-age children: the development of the child attachment interview (CAI). Journal of Child Psychotherapy 29, 171–186.
Thompson, R.A., 1997. Sensitivity and security: new questions to ponder. Child Development 68 (4), 595.
Thompson, R.A., Braun, K., Grossmann, K.E., Gunnar, M.R., Heinrichs, M., Keller, H., Wang, S., 2005. Group report: early social attachment and its consequences: the dynamics of a developing relationship. In: Carter, C.S., Ahnert, L., Grossmann, K.E., Hrdy, S.B., Lamb, M.E. (Eds.), Attachment and Bonding: A New Synthesis. MIT Press, Cambridge, MA, pp. 349–384.
Trivers, R.L., 1996. Parental Investment and Sexual Selection. University of Chicago Press, Chicago, IL.
Van IJzendoorn, M.H., 1990. Developments in cross-cultural research on attachment: some methodological notes. Human Development 33 (1), 3–9.
Van IJzendoorn, M.H., 1995. Adult attachment representations, parental responsiveness, and infant attachment: a meta-analysis on the predictive validity of the adult attachment interview. Psychological Bulletin 117 (3), 387–403.
Van IJzendoorn, M.H., Kroonenberg, P.M., 1988. Cross-cultural patterns of attachment: a meta-analysis of the strange situation. Child Development 59 (1), 147–156.
Van IJzendoorn, M.H., Bakermans-Kranenburg, M.J., Sagi-Schwartz, A., 2006. Attachment across Diverse Sociocultural Contexts: The Limits of Universality. Psychology Press, New York.
Voland, E., 1998. Evolutionary ecology of human reproduction. Annual Review of Anthropology 27, 347–374.
Voland, E., 2000. Nature or nurture?–the debate of the century, a category error, and the illuminating impact of evolutionary psychology. European Psychologist 5 (3), 196–199.
Waters, E., Deane, K.E., 1985. Defining and assessing individual differences in attachment relationships: Q-methodology and the organization of behavior in infancy and early childhood. In: Bretherton, I., Waters, E. (Eds.), Growing Points of Attachment Theory and Research, Monographs of the Society for Research in Child Development, vol. 50 (1–2), pp. 41–65.
Weisner, T.S., Gallimore, R., 2008. Child and sibling caregiving. In: New, R.S. (Ed.), Anthropology and Child Development: A Cross-cultural Reader. Blackwell Publishing, Malden, MA, pp. 264–269.
West, M., Rose, S., Spreng, S., Sheldon-Keller, A., Adam, K., 1998. Adolescent attachment questionnaire: a brief assessment of attachment in adolescence. Journal of Youth and Adolescence 27, 661–673.
Yovsi, R.D., 2003. Ethnotheories about Breastfeeding and Mother–Infant Interaction. The Case of Sedentary Nso Farmers and Nomadic Fulani Pastorals with Their Infants 3–6 Months of Age in Mbven Sub Division of the Northwest Province of Cameroon. LIT Verlag, Muenster, TX.
Yovsi, R.D., Kaertner, J., Keller, H., Lohaus, A., 2009. Maternal interactional quality in two cultural environments: German middle class and Cameroonian rural mothers. Journal of Cross-cultural Psychology 40 (4), 701–707.

Attempted Utopias and Intentional Communities

Adrienne Redd, Schools for Sustainability, Philadelphia, PA, USA
Tsvi Bisk, Center for Strategic Futurist Thinking, Ziv Group, Ltd., Kfar Saba, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The authors define 'attempted utopias' as historical communities founded on systematic principles intended to promote optimal cooperation and fulfillment for their inhabitants. Real-life utopian communities have been predominantly Protestant-based reform movements in the United States from the eighteenth century through the present, though there are notable exceptions, such as ecological communities, and wholesale social reorganization as in the Chinese and Russian Revolutions. Other important themes in utopian experiments entail rethinking the family, sharing of work, and management of environmental resources, such as soil, water, energy, and organic residues.

Attempted Utopias and Intentional Communities

Utopia (pronounced/uːˈtoʊpiə/) can be defined as a proposal for harmonious society governed by a reinvented politicolegal system. No known society has achieved this ideal state, but scores of innovative communities have been organized around central principles meant to systematically address members' needs and ameliorate social problems. Because of the inherent impossibility of actualizing a perfect society, 'intentional community' rather than 'utopia' is a more precise term for such projects undertaken in the real world.

An intentional community is a residential group that defines and consents to a philosophy of how to live and work with a high degree of mutual benefit derived from cooperation, camaraderie, and shared values and responsibilities. An intentional community can be as small as an extended family or as large as a nation-state. The word 'intentional' reflects the attempt to consciously base the community on a set of principles consensually shared by participants who enter into membership voluntarily. Some scholars, such as the author of *California's Utopian Colonies*, add the criterion that a utopian community withdraws from mainstream society (Hine, 1953). However, for the purpose of this historical survey, we will also consider revolutionary movements whose intent was to be transformative for society overall. Examples include the Russian and Iranian revolutions. In his definition of 'utopia' Hicks adds that "[U]topian[s]…seek, by their example, to convince the world of the superiority of the new social institutions they have built" (Hicks, 1987).

Christian (and Islamic) Roots of Utopian Thinking

Although there are ancient, medieval, Renaissance, Islamic, and Eastern works of imagination that describe ideal societies, utopian attempts of the eighteenth and nineteenth century predominantly have ideological roots in the Protestant Reformation. A companion article in this reference work traces literary and philosophical utopian thought from Classical writing through the late twentieth century and examines some non-Christian and non-Western works.

Utopianism as it was attempted historically is predominantly based in the apocalypticism of early medieval Christianity, the Protestant Reformation, and the American Enlightenment. "The perfectionist goals of eventually overcoming sin and achieving human perfection on earth had roots in the earliest development of Christianity, with its millennial expectations of Christ's imminent second coming. As Christianity developed into an established religion, such millennial concerns were largely relegated to an indefinite future. Similar concerns resurfaced, however, at the time of the Protestant Reformation as well as in nineteenth-century America among groups that rejected Calvinist beliefs that human beings were totally depraved and could be saved only by the intervention of a mysterious, utterly transcendent God. Methodists, for example, argued that individuals could play an active role in achieving their salvation and saw salvation as an ongoing process, not absolute…" writes Pitzer (1997), a historian of Communal Protestant utopian settlements.

Monastic Christian practice began in the third century AD and gained momentum after the fall of Rome in AD 476 (Gospel of Judas, 2006). Saint Benedict of Nursia (*c.*AD 480–*c.*543) founded the Benedictine order which established ascetic, communal living arrangements (perhaps in emulation of the endurance shown by Jesus during his 40 days in the desert). During the Middle Ages several lay religious groups such as the Beghards and Brothers and Sisters of the Free Spirit also engaged in communal life. In allowing gender integration, these communities differed from the earlier Catholic and Orthodox monasteries who demanded complete separation of the genders (and total celibacy) as a demonstration of the unearthliness of religious life.

At the end of the tenth century in Europe there were plagues, fires, and disasters. Devout, literal-minded Christians believed the events of the Book of Revelations were imminent and that the second coming would take place in the year 1000. Historians, led by Jules Michelet, have drawn a dramatic picture of the expectation of mass apocalyptic events climaxing at the end of the first millennium AD. Michelet's assessment was that eschatological fervor aroused hope in the oppressed – and terror in the oppressors and catalyzed political transformations of eleventh-century France (Landes, 2000).

About a century earlier, Abū Nasr Muhammad bin al-Farakh al-Fārābi (known to European intellectuals as Alfarabius or Avennasar) wrote *On the Perfect State*, which resonated through both the Islamic world and Medieval Europe. The work shares

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32195-X

with Augustine's *City of God* the assertion that only from the sound moral life of the individual can a harmonious community flow. On such individual lives can the perfect city state be built. This reside (*sic*) within a perfect world state and all under the precepts of religious doctrine (Christian for Augustine and Islamic for al-Fārābi) (Al-Farabi, 1985). However, Islamic ideology never separated the sacred from the mundane as did the accountants of the earliest Catholic universities. This Islamic expectation of the unity of the individual in the community, within the polity, within a world system under God may explain why Islamic culture did not nurture the settlements and experiments that the Christian world did.

The Reformation and the American Enlightenment

Though there are earlier Anabaptist breaks from the Catholic Church, most scholars place the origins of the Protestant Reformation in the teachings of Martin Luther (1483–1546) (and John Calvin (1509–64)). The ideas of these two leaders, one German, one Swiss, fundamentally changed western European societal attitudes about relationship between religion and work. Luther's beliefs rejected the medieval conception of labor within a hierarchy of professions. Rather, Luther asserted that all forms of work offered equal spiritual dignity. This influence can be seen in the way that almost New World Protestant utopian attempts are organized around how work is honored and evenly distributed among the hands that might plow fields, mend clothes, and prepare food, etc. This dismantling of the hierarchical nature of work in the Medieval world was continued in twentieth-century intentional communities in which previously gendered work, such as caring for children, became more evenly distributed.

Access to books (particularly the Bible) in Europe after the inception of the movable type printing press in the mid-fifteenth century made it possible for people to interpret sacred texts for themselves. Starting in the late fifteenth and early sixteenth centuries, Pietist and Protestant movements condemned the corruption of the Catholic Church, which had strayed away from practice stipulated in those sacred texts. Max Weber interpreted Protestant urgency for social reform as follows – that the Protestant elimination of monasteries turned the entire world into one immense monastery in which, as Sebastian Franck put it, "every Christian had to be a monk all of his life" (Weber, 1958).

One of the divergences from Catholic practice salient to the utopian impulse was the rejection of the practice of infant baptism, through which members into the religious community were inculcated within the first few days of being born. Christians who held that prospective members of faith should *choose* to join the religious community were called Anabaptists; they did not initiate members until young adulthood. Even twenty-first-century children reared in Amish communities are not technically 'Amish' until they join voluntarily and are baptized between the ages of 16 and 25.

The geographical location and timing of utopias and utopian communities suggest that they spring from the struggle between Catholicism and Protestant Christianity. This struggle dates to the break from the Catholic Church by Henry VIII.

In 1516, the first version of Thomas More's *Utopia* was published, criticized through fantasy the organization of wealth and the exploitation of the poor in late Medieval Europe. The year after *Utopia* was printed, Martin Luther set in motion events which led to the Protestant Reformation – in which Europe divided permanently on religion. Luther broke away from the Catholic Church altogether.

The English King Henry VIII was a personal friend of More, and made him lord chancellor of England (the highest legal office in the country). As chancellor, More was responsible for burning Protestants. More, like other humanists, had criticized many aspects of the church, but he was unwilling to split from it; he thought that if religious unity amongst Christians was undermined, the result would be disaster for Europe and civilization. So More acted ferociously on his belief that heretics had to be eradicated.

When Henry VIII broke with the Catholic Church, More resigned his job, and in 1535 was beheaded for his opposition to the king's divorce from Catherine of Aragon and marriage to Anne Boleyn. More died for his Catholic faith (and was made a saint centuries later); he himself was intolerant of non-Catholics. Yet in More's seminal book the citizens practice religious toleration, suicide, euthanasia, and divorce, all condemned by the Catholic Church. Not even More's life and writing yield a definitive answer to how a utopia should be organized, but they do point to the issues that nearly every utopian attempt must confront.

The New World and Pietist Intentional Communities

As humanistic philosophy spread throughout Europe during the Renaissance, another factor was added to the utopian impulse: the European encounter with the Americas. The 'New World' offered the potential for the actual establishment of communities that were previously only imagined, such as the travelogue of Thomas More's *Utopia*. British and other colonies in North America offered affordable land and religious freedom, as in Pennsylvania, founded on William Penn's promise of toleration of all faiths. In fictional and philosophical proposals meant to improve upon existing social arrangements it has been necessary to make a fresh start – to leave behind or wipe clean the slate of previous civilization. This is why so many literary and attempted utopias have been set on islands or in remote rural areas.

Granted a huge tract of land by King Charles II (in payment for a debt owed to Penn's father), William Penn became a real estate promoter, city planner, and governor when he founded his 'Holy Experiment' and sanctuary for Quakers in the prerevolutionary province of Pennsylvania in 1682. It became a haven of toleration for pietist, communitarian, and free-thinking groups who were leaving an intolerant Old World for the open religious climate of the British colony.

After freedom of religion was encoded into the U.S. Constitution, the Enlightenment idea of progress continued to blend with reformism to create an American incubator of social experiments from Maine to Kentucky and New York to California. These factors generated an energy and moral certainty that has characterized the American public sphere ever since.

The notion that individuals and nations can, by their own actions, make a better life for themselves rather than wait passively for the latter day constituted the very spirit of the Enlightenment and modernity's emphasis on humanism and rationality. Attracted by the promise of social inclusion within Penn's experiment (Pennsylvania) and arable land in the Mid-Atlantic, religious utopian societies from Europe streamed to the United States from the eighteenth century on; these included the *Society of the Woman in the Wilderness* (led by Johannes Kelpius), the 'Solitaries' who established the Ephrata Cloister, and the Harmony Society, among many others.

After freedom of religion was encoded in the U.S. Constitution, the Enlightenment idea of progress continued to blend with reformism to create an American incubator of social experiments from Maine to Kentucky and New York to New Mexico. These factors generated an energy and moral certainty that has characterized the American public sphere ever since. The Progressive movement, with both high ideals and petty prejudices, was the definitive example of this social trajectory.

The notion that individuals and nations can, by their own actions, make a better life for themselves rather than wait passively for the world to come was the very spirit of the Enlightenment and modernity's emphasis on humanism and rationality (Bury, 1920/1932/1955). Even before dissatisfied Lutherans in Germany turned to pietism, the movement, founded by Jan Hus, started what is now the Czech Republic in the 1450s. Hus organized his pietist groups as the Moravian Brethren. The term 'Moravian' comes from the original exiles who came to Saxony in 1722 from the Eastern European region of Moravia to escape persecution, but the Moravian Church actually began in 1457 in Kunvald, Bohemia, an autonomous kingdom within the Holy Roman Empire. Over the centuries they passed on their own practices and also absorbed influences from Anabaptists, then Baptists and Methodists. The Moravian Church (or Unitas Fratrum, meaning 'Unity of the Brethren') is one of the early Christian groups that settled in Pennsylvania. Moravians emphasize Christian unity, personal piety, missions, and music as intrinsic to worship. Their settlements in Bethlehem, Pennsylvania (in 1741) and elsewhere in the New World were communes, in which men and women lived communally and separately until marriage, and property and work were shared. They were among the first industrialists in what would become the United States, and harnessed water power to run their tanneries, grist mills, and oil mills. Over time, the collectivist nature of Moravian settlements changed; however, the church still exerts social and political influence in Moravian communities from Bethlehem, Pennsylvania to Bethel, Alaska.

Johannes Kelpius (1667–1708) was a German pietist, mystic, musician, and writer, interested in the occult, botany, and astronomy, who came to believe with his followers in the pantheist 'Society of the Woman in the Wilderness' that the end of the world would occur in 1694. Kelpius felt that the seventeenth-century Province of Pennsylvania, given its reputation for religious toleration at the edge of a barely settled wilderness, was the best place to be. Philadelphia had been founded in 1682, but the city and the Province of Pennsylvania had quickly become a tolerant haven and refuge for many pietist, communitarian, or free-thinking groups who were leaving an intolerant Old World for the congenial religious climate of the British colony. Kelpius and his followers crossed the Atlantic and lived in the valley of the Wissahickon Creek in Philadelphia from 1694 until his death. It is reported that they lived communally, though they also spent time in solitary meditation in caves and small cells scattered about their common living area. Though no sign or revelation occurred, the faithful, known as the Hermits or Mystics of the Wissahickon, continued to live in celibacy, searching the stars and hoping for the end of earthly time.

The Harmony Society was a Christian theosophy and pietist group founded in Germany by George Rapp in 1785. Its adherents were known as Harmonists, Harmonites, or Rappites. Due to religious persecution by the Lutheran Church and the government in Württemberg, the society moved to the United States in 1803, settled in Pennsylvania, and on 15 February 1805, about 400 followers formally organized the Harmony Society, placing all their goods in common. The society established three communes: Harmony, PA, Harmony, Indiana (the predecessor to Robert Owen's New Harmony), and one at Economy, PA, which was a pioneer in the American silk industry, and became the American silk center in the 1830s and 1840s. The group lasted until 1905, making it one of the longest-running financially successful communes in American history.

Religious Revivalism on American Soil

Historian Sydney Ahlstrom describes a 'great international Protestant upheaval' of the 1730s and 1740s that emphasized personal experience rather than authority and led people to be more passionately involved in their religion (Ahlstrom, 1972). Thomas Kidd goes on to make the extended argument that the First Great Awakening of the early eighteenth century intermingled the passions for salvations with yearnings for political freedom as well, leading the American colonies into a revolution and complete severance from England (Kidd, 2007).

A major figure in the Second Great Awakening in the U.S., Charles Grandison Finney preached in 1835 On Revivals of Religion (1835) that individuals *and* societies could be 'saved' by conscious decisions either by accepting Jesus, in the case of individuals, or by living according to the laws of God and the teachings of Jesus, in the case of societies. Reform-minded, Finney also advocated equality for women and African-Americans and became president of Oberlin College, the first American college to admit female and black students. Since it is hard for the individual to be virtuous in a sinful society, overarching social reform was deemed necessary and this generated a wide and creative variety of projects.

The Shakers and Other Communal Utopians

The longest lived (1747–present) and among the best-known Protestant intentional communities were the Shakers (self-identified as *The United Society of Believers in Christ's Second Appearing*), who originated in England in the eighteenth century but emigrated to America to seek a more complete manifestation of their vision. The members looked to women

for leadership. Jane Wardley, an articulate preacher, urged her followers to repent because the return of the true church was imminent. Ann Lee joined the group in 1758 and soon assumed leadership. The death of her four children in infancy traumatized 'Mother Ann,' as her followers later called her. She claimed numerous revelations regarding the fall of Adam and Eve and directed her followers to abstain from all sexual intercourse. Shakers practiced celibacy and communal ownership of goods, along with a strict separation, but total equality of the genders in work and leadership. They also adopted and educated children left in their care.

Manufacturing was a primary reason for the early success of the Shakers, but automated production and new laws concerning the adoption of orphans contributed to their decline. Membership dwindled in the early twentieth century. Many Shaker settlements have now been converted into museums; as of this writing three Shakers abide on the last working farm in Sabbathday, Maine.

The Zoarites were another communal society that came to Philadelphia from Germany in 1817. Their religious views were inspired by Jakob Böhme, a mystic who rejected salvation by faith alone and believed that the travails of man enabled God to achieve self-awareness – i.e. that human behavior during corporeal life affects the very being of God. This is a proactive belief system amenable to the essence of social reform. Sympathetic Quakers funded their migration to Ohio, where they established a communistic settlement. The death of their charismatic leader led to a decline in the cohesion of the village. By 1898, the village voted to disband the communal society and the property was divided among the remaining residents.

Christian intentional communities and attempted utopias were once so prevalent in the nineteenth century that the western and central region of upstate New York was referred to as the 'Burned-Over District,' meaning that the wildfire of radical religious experimentation and reform had so thoroughly consumed the region that there was no fuel (potential new members for the various movements) left to be ignited. Religious movements in the Burned-Over District included the Mormons, the Shakers, and the Oneida Society. Also emerging from the reformist fervor of New York in the late nineteenth century was Elizabeth Cady Stanton, who organized the Seneca Falls Convention and established the women's suffrage movement in the United States.

Extant Anabaptist Utopian Communities

The extant Amish, Hutterites, and Mennonites are direct descendants of the Anabaptist movement. The Hutterites are a communal branch of Anabaptists who, like the Amish and Mennonites, trace their roots to the Radical Reformation of the sixteenth century. Since the death of their namesake Jakob Hutter in 1536, the collectivism, pacifism, and piety have resulted in hundreds of years of odyssey through many countries. Nearly extinct by the eighteenth century, the Hutterites found a new home in North America. Their population grew from 400 to around 50 000. Like the Amish in the farmlands of Pennsylvania, Ohio, and Texas, Hutterites live in rural regions of Western Canada and the upper Great Plains of the U.S. They live on community farms or *bruderhöfe*, each one of which is home to about 15 families.

The Bruderhöfe Communities ('place of brothers' in German) are Christian religious communities with branches in New York, Florida, West Virginia, and Pennsylvania in the U.S., the United Kingdom, Germany, Paraguay, and Australia. They were previously called The Society of Brothers and were loosely affiliated with the Hutterian Brethren. The group is also known as Church Communities International. Most contemporary communities have a nursery, kindergarten, school, communal kitchen, laundry, various workshops, and offices. Bruderhöfe life is built around the family, though there are also single community members. Children are an important part of each community and participate in most communal gatherings. Disabled and elderly members are cared for within the community and participate in daily life and work as much as they are able. Like the Hutterites, the Bruderhöfe members do not hold private property individually, but rather share everything in common. No Bruderhöfe member receives a salary or has a bank account. Income from all businesses is pooled and used for the care for all members, and for various communal outreach efforts.

The Hopedale Community was a religious commune founded in Worcester County, Massachusetts in 1842. Their principles were temperance, abolitionism, women's rights, spiritualism, and education. When the commune went bankrupt in 1856 it became a textile factory, a resonance of the Oneidans who continued making utensils long after their social experiment dissipated.

Utopians in the American West

Subsequent to the northeastern states welcoming intentional communities, utopian settlers moved west and south, in search of independence and open space in which to conduct their social experiments. Fountain Grove was a utopian religious cooperative in California founded in 1875 by Thomas Harris who established a movement called the Brotherhood of the New Life. He had previously founded two similar communities in New York State. The movement's most prominent member was Laurence Oliphant who was sent to Palestine by Harris to explore the possibility of setting up a similar utopian community. Oliphant tried to obtain a lease for the Galilee region from the Turkish Sultan in order to resettle the Jews and bring about the end of days. This took place a full decade before Theodore Herzl wrote *The Jewish State* (1896).

The Church of the Latter Day Saints, also known as the Mormon Church, was founded in 1830 by Joseph Smith, who claimed that God had led him to a new set of scriptures called the Book of Mormon about the presence of Jesus in North America. It became another collectivist project which in retrospect is arguably utopian in its principles. Like other social innovators who have advocated for a different expression of human sexuality, Smith espoused polygamy as part of his utopian society. Smith and his followers were persecuted in Ohio and the Midwest. In 1844, a mob murdered Smith and his brother Hyrum in Illinois. Smith's follower, Brigham Young, led the followers of Mormonism to what is now Utah.

Orderville, Utah was established at the direction of Latter-day Saint (Mormon) leader Brigham Young in 1875 and followed a voluntary form of communalism defined by Joseph Smith. The town was settled primarily by destitute refugees from failed settlements on the Muddy River in Nevada. The extreme poverty of these settlers may have contributed to their devotion to the principles of the United Order.

Meaning 'House of God' in Hebrew, Bethel or Beth el or Beth-El, Missouri was a Bible-based communist utopian colony established in 1844 by William Keil. He believed that the Book of Acts required that Christians hold all property and means of production in common. He founded another religious commune in 1856 in Aurora, Oregon. He was encouraged to form his communal movement by former members of the Harmony Society. Bishop Hill Colony in Illinois was a utopian communal religious community. It was founded in 1846 by Swedish pietist Eric Janson who believed the commune would become the 'New Jerusalem' and spread reform around the world.

Jewish Intentional Communities

In 1825 Mordechai Noah, an American Jewish playwright, attempted to establish a Jewish refuge on Grand Island in the Niagara River. It was to be called *Ararat* after the resting place of Noah's Ark. Noah purchased land on the island and laid a cornerstone which read 'Ararat, a City of Refuge for the Jews,' but Jews did not settle in the City of Refuge, so the project began and ended with the ceremonial laying of that cornerstone.

In 1891, 3 years before the Dreyfus Affair and 6 years before the First Zionist Congress, the German Jewish philanthropist Baron Maurice de Hirsch created the Jewish Colonization Association (JCA). The goal of the JCA was to encourage mass Jewish emigration from Russia and Eastern Europe in order to settle them in agricultural colonies on lands purchased by the Association. The peak of JCA activity in Argentina was in 1930 with over 20 000 Jews on approximately 1 500 000 acres of land (including Jewish *Gauchos*). But town life eventually outshone the romance of the soil and by 2012 only a few dozen elderly Jews were left.

The JCA attempted something similar in Brazil in 1904 with Jews from Bessarabia (modern-day Moldova). This colony failed and was liquidated in 1928. Another colony was founded in 1909 but was abandoned even before the First World War (WWI). The JCA resettled it but this experiment finally ended in 1965. Another failed attempt was in the 1930s with German Jewish refugees.

The JCA was active in the United States and Canada and in cooperation with other groups had established 78 Jewish farms by WWI in New York, New Jersey, Pennsylvania, and Connecticut – the most famous being Woodbine, New Jersey. By the time of the Second World War (WWII), JCA activity in the United States had all but ceased. In Canada 40 Jewish farms had been established by 1930, though nothing remains of the project.

In Israel itself, the Kibbutz is perhaps the most well-known and successful twentieth-century Jewish utopian experiment. Over 270 Kibbutzim in four major ideological categories survived, grew, and flourished for over 70 years (until their ultimate crises) and were the backbone of the establishment of modern Israel. Other Israeli utopian experiments include the Moshav (small agricultural cooperative) and the Moshav Shitufi. Unlike the Kibbutz, land is not worked in common but 'held' by separate families who live in family units (private homes) and retain their own profits. The Moshav Shitufi is a hybrid of Moshav and Kibbutz. Families live as they do in the Moshav with small family-managed plots of land while certain sectors of agriculture are operated collectively as in the Kibbutz. Services and marketing are dealt with cooperatively in both these social structures; by the twenty-first century, economic downturns have all but destroyed the cooperative aspects of these experimental settlements.

Organizing Principles of Intentional Communities

Although most intentional communities have been religious (and most of those Anabaptist Protestant Christian projects), monotheistic faith is not the only organizing principle upon which utopian experiments have been built. Ownership or property, vegetarianism, ecology, and human sexuality (or celibacy) have been central principles in utopian attempts.

While some groups, such as the solitaries of Ephrata and the Shakers, chose celibacy and asceticism as a path to fulfillment, the Oneidans chose what they called 'complex marriage.' Founded by John Humphrey Noyes, the settlers in Oneida in upstate New York rejected monogamy. The reasoning behind this was that sexual possessiveness leads to overall lack of concern for the larger community. Members were encouraged to engage in free (heterosexual) copulation throughout the community; postmenopausal women were encouraged to teach young men about the joys of sex (particularly because these couplings were less likely to result in pregnancy) and the devout were encouraged to consort with the nondevout in hopes of encouraging greater religiosity.

Decisions about procreation were handled by committee, though unsanctioned children were provided for just as if they had been planned in accordance with the rules. Birth mothers were only responsible for the care of their offspring for the first few years of life; the community as a whole cared for and disciplined older offspring. Now known for its manufacture of Oneida silverware, the Oneida community lasted from 1848 to 1881, making it among the longest-running communes in American history.

The *Amana* Colonies in Iowa constitute another example of a long-running commune. Started by radical German pietists, they lasted from 1855 to 1932. Like the Moravians and Shakers, they embraced manufacturing. The Amana Corporation, manufacturer of refrigerators and household appliances, was originally started by the group.

Communal Societies

Two dominant themes in intentional communities are collectivism and agrarian life. Considering the roots of utopianism in the Protestant Reformation of the sixteenth and seventeenth century, the humanism and empowerment of

the Renaissance, and the rationality of the Enlightenment, it is necessary to consider the impact of these ideas on social upheaval of the Industrial Revolution. Starting since the 1830s in England and continuing for a century in Europe and the U.S., workers flocked to cities to work in factories and struggled visibly as capitalism and the profit motive became organizing principles of both societies. The reform movements of the nineteenth century came to emphasize distributing resources such as housing, medical care, food, and education to struggling workers and the urban underclass. However, some utopians took it upon themselves to start new communities, usually away from the squalor and turmoil of the cities, and to organize those societies around equitable distribution of goods and work.

Connected to the urge to distribute resources fairly is the question of inherent human selfishness and how reshape in the culture of the settlement so that the needs of the group precede individual desires and urges. During the nineteenth century and through the twentieth century, economic utopias sprang up around the United States in reaction to various political conservative movements. They were communal and egalitarian, sharing work and resources.

American secular utopian communities took inspiration from ideas and philosophies originating with Protestant Christian groups, but evolved their philosophy to moral-spiritual, rather than religious doctrine. Transcendentalism began as a term developed by the German philosopher Immanuel Kant (1724–1804) embodying those aspects of man's nature transcending, or independent of, experience. Taking root in America, Transcendentalism created a cultural renaissance in New England during 1830–45 and manifested its chief American expression in Ralph Waldo Emerson's doctrine of self-reliance.

Some Transcendentalists decided to put their theories about 'plain living' into practice. One experiment in communal living was established at West Roxbury, Massachusetts, on some 200 acres of land from 1841 to 1847. The Brook Farm Institute of Agriculture and Education became better known than many other communal experiments due to its distinguished founder and inhabitants. The community wasn't unique for its time; more than 80 utopian communities were launched in the 1840s alone – but it was notable as the first purely secular one. The Brook Farm Institute was organized and directed by George Ripley, a former Unitarian minister and later literary critic for the *New York Tribune*. Others connected with the project were Charles A. Dana and Nathaniel Hawthorne (both shareholders), Ralph Waldo Emerson, Margaret Fuller, William Henry Channing, John S. Dwight, and Sophia Dana Ripley, a woman of wide culture and academic experience. Brook Farm attracted not only intellectuals, but also carpenters, farmers, shoemakers, and printers. The community provided to all members, their children and family dependents, housing, fuel, wages, clothing, and food. There was daycare, a primary school, and a college, preparatory course covering 6 years.

A disastrous fire in 1846 demolished the newly financed Phalanstrey building. In combination with further financial troubles, including Hawthorne's suit against Ripley and Dana to recover his investment in the project, brought about the end of the Brook Farm community the following year. Nathaniel Hawthorne later satirized in his novel, *The Blithedale Romance* (1852).

In 1842 the Northampton (Massachusetts) Association of Education and Industry established another secular commune based on racial, gender, religious, and class equality. They wanted to replace competition with cooperation and were especially dedicated to the abolition of slavery. Sojourner Truth was a member and Frederick Douglass was a visiting lecturer.

The *Skaneateles Community* was established in 1843 by the *Society for Universal Inquiry and Reform* in upstate New York. Several similar communities were established by the society before the Civil War. It survived for 3 years. Its leaders were abolitionists. Members shared living quarters and worked in agriculture and light industry. It was economically successful but broke up because of internal dissent.

The *Icarians* were a utopian movement founded by French philosopher Étienne Cabet. He advocated workers' cooperatives instead of capitalist factories. This appealed to the artisan classes whose living was being endangered by large factories. He had been influenced by Robert Owen and, in turn, influenced Marx and Engels. His social philosophy stemmed from primitive Christianity and was detailed in his book *The Real Christianity according to Jesus Christ*. At its peak the movement had 400 000 adherents in France but still made no impact on actually reforming French society. Consequently he moved with some of his followers to the United States in 1848 where he established several egalitarian communes in Texas, Illinois, and Iowa.

Fruitlands was a utopian agricultural commune founded in 1843 by Bronson Alcott in Concord Massachusetts based on transcendentalist principles. Members were vegan vegetarians, drank only water, bathed in cold water, and used no artificial light. Property was held communally, and no animal labor was used. Celibacy was celebrated as an ideal. Although it was egalitarian in principle, Alcott's wife and the other women did most of the work and the stress of deprivation because of its extreme principles contributed to its collapse after only 7 months.

The anarchist village *Modern Times* was a community of 150 founded on Long Island in 1851. It was based on barter and existed until the Civil War when patriotism trumped social idealism and many members enlisted in the Union army. In 1864 it transformed into the prosperous suburb of Brentwood.

Some utopian writings originated in Europe, but were attempted in North America. Another influence behind Brook Farm was the writing of François Marie Charles Fourier. Fourier believed that society should be organized into small cooperative groups, and that people should largely be assigned productive work that they enjoyed, reorganization that would greatly increase society's productivity (Fourier, 1967). He proposed to eradicate poverty by providing adequate wages and guaranteeing a minimum standard of living for those not able to work. In this he anticipated the modern market-based welfare state. As many as 29 communities were founded in the United States based on Fourier's views; they included Utopia, Ohio; La Reunion, Texas; and the North American Phalanx in Red Bank, New Jersey.

Fourier envisioned cooperative communities he called 'phalanxes' living in structures called Phalanstères or 'grand hotels' (not unlike the Unité d'Habitation architectural

experiment of Le Corbusier). These buildings were four-level apartment complexes in which the richest had the uppermost apartments and the poorest resided on the ground-floor residence. Wealth was determined by one's job; jobs were assigned based on the interests and desires of the individual. There were incentives: jobs people might not enjoy doing would receive higher pay. Fourier considered trade, which he associated with Jews, to be the 'source of all evil' and advocated that Jews be forced to perform farm work in the Phalanstères.

Fourier's vision of cooperative communal subunits and a greater social safety net for workers also influenced the Paris Commune of 1871. While it only lasted 40 days and was fraught with instability, it stemmed from the communal utopian impulse and is regarded as the first socialist worker uprising. The workers of Paris, joined by mutinous National Guardsmen, seized the city and set about reorganizing society in their own interests based on workers' councils. They could not hold out, however, when more troops retook the city and massacred 30 000 workers.

A handful of communal, agrarian settlements of the nineteenth century were founded on European soil. Matti Kurikka was a Finnish utopian socialist and theosophist who helped establish *Sointula*, a utopian island colony of Finnish immigrants on Malcolm Island, British Columbia Canada based on cooperative principles in 1901. The utopian colony was dissolved in 1905.

New Lanark is a village on the River Clyde, approximately a mile and a half from Lanark, in South Lanarkshire, and some 25 miles southeast of Glasgow, Scotland. It was founded in 1786 by David Dale, who built cotton mills and housing for the mill workers. Dale built the mills there in a brief partnership with the English inventor and entrepreneur Richard Arkwright to take advantage of the water power provided by the only waterfalls on the River Clyde. Under the ownership of a partnership that included Dale's son-in-law, Robert Owen, a Welsh philanthropist and social reformer, New Lanark became a successful business and an epitome of utopian socialism as well as an early example of a planned settlement and so an important milestone in the historical development of urban planning.

Robert Owen was a social reformer and wealthy industrialist who made his fortune from textile mills in New Lanark, Scotland. Owen and his 22-year-old son, William, sailed to the United States in 1824 to purchase a site to implement his vision for 'a New Moral World' of happiness, enlightenment, and prosperity through education, science, technology, and communal living. Owen's utopian community would create a 'superior social, intellectual and physical environment' based on his ideals of social reform. Owen and his followers and successors were known as Owenites. Owenism aimed for radical reform of society and is considered a forerunner of the cooperative.

New Harmony became known as a center for advances in education and scientific research. New Harmony's residents established the first free library, a civic drama club, and a public school system open to men and women. Its prominent citizens included Owen's sons, Indiana congressman and social reformer Robert Dale Owen, who sponsored legislation to create the Smithsonian Institution; David Dale Owen, a noted state and federal geologist; William Owen; and Richard Owen, state geologist, Indiana University professor, and first president of Purdue University. The town served as the second headquarters of the U.S. Geological Survey and numerous scientists and educators contributed to New Harmony's intellectual community, including William Maclure, Marie Louise Duclos Fretageot, Thomas Say, Charles-Alexandre Lesueur, and Joseph Neef, Frances Wright.

Literary Influences on Utopian Attempts

One of the most influential fictional visions of a collective society was Edward Bellamy's *Looking Backward: 2000–1887* (1888), which was translated into dozens of languages and influenced popular readers, intellectuals, and reform movements. In the novel, the Eastern United States has peacefully transitioned to a society where there is no money and each citizen is productive to the best of his or her ability and takes what he or she needs from communal dining halls and well-stocked shops. *Looking Backward* had a profound impact on the minds of reformers who were well aware of the upheaval, deprivation, and widening of the gap between rich and poor caused by the Industrial Revolution.

Topolobampo was the site of a utopian colony from 1884 to 1894 in Mexico. It was the socialist utopian vision of Albert K. Owen and was settled by farmers from New England. The members believed in 'Collective ownership and management for public utilities and conveniences and the community being responsible for the health, usefulness, individuality and security of each.' The colony influenced the garden city concepts of Ebenezer Howard who had himself been influenced by Bellamy's utopian novel of 1888, as well as Henry George's economic article, *Progress and Poverty*, an example of the indirect influence of utopian literature on society.

Revolution as Utopianism

While European and American workers struggled with the upheavals of the Industrial Revolution, Russia reeled with the aftermath of the liberation of the serfs and problems with distribution of land. The Russian Revolution of 1905 was a wave of worker strikes, peasant unrest, and military mutinies that spread through the Czarist Russia. It catalyzed the establishment of a limited constitutional monarchy, the *Gosudarstvennaya Duma* (State Assembly), a multiparty system, and the Russian Constitution of 1906. Although he saw violent revolution as an illegitimate source of reform, Max Weber considered this a great opportunity for renewal of Russia and an opportunity for sweeping solutions to the major economic and social problems of turn-of-the-century Russia, namely more equitable distribution of land to the peasants and sharing of political power (Weber, 1995). He specifically refers to the potential for revolution to be utopian; interestingly, the German word 'Utopische' has the same connotation as it does in English – that of a proposal being sweeping, idealistic, and overambitious.

Mao Zedong and Mohandas Gandhi were born in the East but their thinking was based in the forward-thinking, scientifically based paradigm of modernity. The thrust toward

modernity differs in Asia's two leviathans but their modernity is unquestionable (Friedman and Gilley, 2005). Gandhi and Mao expressed their views in Indian and Chinese cultural trappings to popularize them, yet Gandhi's attitudes toward eliminating caste and Mao's desire for women's equality were more western than eastern.

It was not the idealist dimension of the peasant revolutions that caused them to sour, however. It was the willingness to crush the individual in pursuit of the total reorganization of society, and the way that Soviet totalitarianism played out in the public consciousness. The view of peasant revolutions as utopian was quickly quashed as citizens of such states revealed their brutal dimension in dystopian fiction that barely exaggerated the oppression of the Soviet Union, such as Yevgeni Zamyatin's *We* (1924), Ayn Rand's *Anthem* (1938), and scathing satires of Soviet Russia, such as Alexander Zinoviev's *Yawning Heights* (1976). As dissidents fled persecution, even children's fiction incorporated glimpses of what a society becomes when it sacrifices individuality for stability and conformity. The scene in the city of Camazotz of a boy bouncing a ball in perfect rhythm in a reeducation chamber in Madeleine L'Engle's *A Wrinkle in Time* (1962) chillingly captures the horror of Soviet Gulags and reeducation camps, as well as American anxieties in the U.S. at the time the book was written (Halberstam, 1994).

Technology and Futurism

Although the Amish and old order Mennonites reject technology beyond an arbitrary point as distractions from their spiritual pursuits, some groups, notably the Shakers, Oneidans, and Amana Colonies, embraced and even innovated new technologies.

Eventually founding two dozen communities in New York, Connecticut, Massachusetts, New Hampshire, Maine, Kentucky, Ohio, Indiana, Florida and attracting over 20 000 converts, the Shakers enjoyed their heyday in the nineteenth century, creating a number of technological advances including commercially packaged seeds and opium, the circular saw, the flat broom, the pencil-end eraser, as well as highly functional and starkly elegant furniture and architecture.

While the majority of utopian communities have been faith-based, there have also been rational and systematic social experiments. In a scientific approach to finding utopia, the Global Scenario Group, an international group of scientists founded by Paul Raskin, used scenario analysis and backcasting to map out a path to an environmentally sustainable and socially equitable future. Its findings suggest that a combination of global grassroots momentum and political will is necessary to steer political, economic, and corporate entities toward this new sustainability paradigm. Fictional visions of the misuse of technology, as in the surveillance screens in Orwell's *1984*, and industrialized human reproduction in Huxley's *Brave New World* have also contributed to dystopian fears.

Architectural Experiments and Ecovillages

Epiphanies about environmental degradation and the finiteness of the planet were of the cultural forces that stimulated some utopian impulses during the 1960s and the 1970s. Arcosanti, an experimental town on 25 acres in central Arizona, is an example of a secular, environmentally oriented intentional society that has integrated the use of specific technologies in order to make life in the desert workable. Construction began on the project in 1970 and most construction was complete by 1989. Though neither as prosperous nor as populous as Matrimandir (discussed below), Arcosanti has about 150 permanent residents. Founded by Italian-American architect and artisan Paolo Soleri (who retired as president and CEO at the age of 92 in 2011).

Matrimandir of auroville

The purpose of the Arcosanti was to demonstrate that architecture and ecology could be combined into 'archeology' and lessen the destructive impact of human habitation on the planet. The principal economic activity of the community (in addition to growing food) is the casting of bronze sculptural bells. As in a kibbutz, residents live in family apartments, but shared cooking and other work, including the casting of bells, which was the principal economic activity of the complex. One of the challenges of the community has been how members can make a living through a combination of mandatory work at Arcosanti and still have sufficient time to be employed outside the settlement.

One of the most enduring and globally regarded ecovillages is Auroville, a city of 2300 people on 60 acres of denuded land (which had only a solitary banyan tree when the project started) on the east coast of India in Tamil Nadu. The community grew out of the work of an advocate for Indian independence, Sri Aurobindo, who founded an ashram near

Pondicherry on the principle of what he called 'integral yoga.' Aurobindo cofounder, a Frenchwoman named Mirra Alfassa, who came to be known as 'the Mother' outlined a charter for the community, emphasizing cooperation and the idea that material manifestations of harmony with nature (such as food and energy production and water and waste management) are crucial to healthy spirituality.

Endorsed by UNESCO and comprising people from over 40 countries and built around a huge golden dome, called Matrimandir, the city has capacity for 50 000 people. It has continued to grow in population each year and asserts the goal of embodying a global paradigm shift toward sustainability.

Another ecologically oriented architectural community is the Findhorn Ecovillage located near Moray, Scotland. Founded in 1962, the purpose of the project was to demonstrate a human settlement that was sustainable in environmental, social, and economic terms. Approximately 450 members of the 'ecovillage' live and work there, engaging in ecohouse construction; a 75 kW Vestas wind turbine and an ecological waste water treatment system were all undertaken by the Findhorn Foundation itself; other independent practitioners have been attracted to the area and have diversified the ecological and demonstration projects.

The back-to-the-land movement in the United States and elsewhere called for people living on any sizable piece of land from a city lot to ample acreage to grow food from the land on a small-scale basis for themselves or for others. The concept was popularized in the United States at the beginning of the twentieth century by activist Bolton Hall, who set up vacant-lot farming in New York City and wrote on the subject.

Communal Experiments of the 1960s and 1970s

Distributism was an attempt during the 1920s and 1930s to find a third way between capitalism and socialism. It was later used to refer to a North American social phenomenon of the 1960s and 1970s. This latter back-to-the-land movement was a migration from cities to rural areas that took place in the United States, its greatest vigor being in the late 1960s and early 1970s.

During the nineteenth century, California was home to diverse utopian political experiments. These 'old utopian communities' were built on themes ranging from communitarian, socialist and Fourierist, that being the proposal that society be organized into small self-sustaining settlements. Life in these communities entailed hard work and simple living conditions. They were frugal and isolated from the conveniences of late-nineteenth-century life. In sociological terms, most of these communities were 'total institutions' (Goffman, 1961) because they encompassed most of the life-space of their members. Few of the communities lasted more than 10 years. While most of them disbanded by 1900, there were continuing experiments with cooperative utopian communities through the twentieth century in California and other states. In the 1960s, the West Coast was again the cradle of developing small cooperative communities.

Many back-to-the-land and communal arrangements conceived in the heyday of Haight-Ashbury disintegrated; Drop City was an artists' community that formed in southern Colorado in 1965. It became known as the first rural 'hippie commune.' Inspired by the architectural ideas of Buckminster Fuller and Steve Baer, residents constructed domes and zonohedra to house themselves, using geometric panels made from the metal of automobile roofs and other inexpensive materials. In 1967 a group consisting of 10 core people won Buckminster Fuller's Dymaxion award for their constructions. The project was abandoned by the early 1970s.

Zonohedron constructed at the burning man festival

Several intentional communities from that era have endured. One is The Farm, a 1750-acre commune in Tennessee founded by Stephen Gaskin, a former English professor at San Francisco State University. Traveling from California in a caravan of 60 buses, vans, and trucks in 1971, its residents pooled their money to buy land in Lewis County, near the town of Summertown. They installed their own water system and had their own electrical crew, composting crew, farming crew, construction and demolition crew, clinic, firewood crew, alternative energy crew, motor pool, laundromat, tofu plant, bakery, school, and ambulance service. It established The Book Publishing Company, which published the works of the Gaskins and other Farm members. The Farm's midwifery school and Ina May Gaskin's seminal book *Spiritual Midwifery* are well known throughout the world for their emphasis on maternal and newborn compassion, safety and success rates.

The Twin Oaks community in Louisa County, Virginia is an ecovillage and intentional community of about 100 people living on a 450-acre farm in Louisa County, Virginia. Founded in 1967, it is one of the longest-enduring and largest secular intentional communities in North America and was Initially inspired by B.F. Skinner's novel *Walden Two*, which describes a fictional behaviorist utopia. However, behaviorist principles gave way to egalitarian principles and the community has successfully functioned according to those for more than 40 years. The community struggled during its early years because of the problems stereotypically associated with communes, such as freeloading and lack of structure, but developed a structured, but flexible, labor system that has been successful. A member of the Federation of Egalitarian Communities, its basic values are cooperation, egalitarianism, nonviolence, and sustainability.

The Failure of Utopianism

Although there are some extant intentional communities, such as the last few Shakers, Findhorn, Arcosanti, a few remaining kibbutzim, The Farm, Twin Oaks, and Auroville, many utopian initiatives eventually dissipated with the exception of the Hutterites and the Amish in North America and the remnants of the Kibbutz movement in Israel. The failure of so many well-intentioned initiatives pursued with such early energy and idealism has made the term 'utopian' a pejorative, synonymous with naive, unrealistic aspirations, and even oppressive control of the residents.

Disillusionment with Israel's Kibbutz movement played a part in discrediting utopianism. This had been the one twentieth-century utopian experiment given sanction by serious social science. But by the first decade of the twenty-first century only 20 out of 270 Kibbutzim were still flourishing in their original communal form. Others were barely surviving. The Kibbutz movement only survived into the second decade of the twenty-first century by abandoning unconditional equality. Differential income standards were adopted – people doing jobs that required greater skills and responsibility received greater compensation. Previously services had been available on demand according to the request of each member. Gradually, members were required to pay for their electricity, water, and food out of their own income. The consumption of these necessities plummeted (to the delight of environmentalists) enabling many previously failing Kibbutzim to regain financial stability. One of the sad realities of the Kibbutzim was that egalitarianism encouraged waste. People left their heat on high so they could come home to a warm apartment, fed food meant for humans to companion animals, and did not bother to repair dripping faucets (Unpublished interview).

Second- and third-generation kibbutzniks had already radically changed the internal social structure of the Kibbutz before the financial crisis and in doing so contributed to it. Those raised in communal children's houses (an experiment celebrated in scores of social science books as an alternative to the traditional family) had painful memories of their experience and refused to raise their own children in the same way. Anecdotal articles as well as substantial research and lawsuits undermined the halcyon myth of the 'children's house', revealing physical, sexual, and psychological abuse.

The subsequent necessity of expanding family dwellings resulted in great expense, which became a major contributor to subsequent financial problems. Many people sympathetic to the kibbutz ideal concluded that idealistic wishful thinking can produce adverse social structures which result in economic difficulties, social disintegration, and the loss of individual self-esteem for members of the community. This appears to have been the greatest failure of utopianism; the inability to understand the connection between economics and ideals: that no matter how lofty the ideal, if it does not rest upon a foundation of practical, efficient, rational economic behavior it is doomed to fail. The very short life spans of the intentional communities described in this article appear to support this observation. The history of Soviet and Chinese collectivization also supports this assessment.

The examples of the Soviet Union and China, the self-indulgent communes of the 1960s and dropouts of the 1970s, and the bizarre religious communes such as Jonestown and Branch Davidians have done little to improve the reputation of utopian speculations and experiments. The dismal record of China's Cultural Revolution, the Gulags of the Soviet Union, and the killing fields of the Communist Party of Cambodia during the twentieth century have made utopianism less promising. It has been estimated that governments having utopian pretensions as their foundational societal axiom killed over 50 million people in the twentieth century. Stalinism, The Great Leap Forward, The Cultural Revolution, Pol Pot, and other 'utopian' phenomena have piled up bodies in numbers inconceivable to even Genghis Khan, Attila the Hun, and Ivan the Terrible.

Totalitarian Dilemma within Utopianism

"The main reason that utopian thought went out of fashion was the link drawn between utopianism and the totalitarian regimes that blighted the twentieth century," writes Craig Smith (2005). Sargent (2010) also notes the trajectory from utopian thought to totalitarianism and Goodwin and Taylor (1892/2009) wrote, "[F]actors conspired, in the 1920s and 1930s, to alter the prevailing mood from optimism to despair: widespread economic depression, the constant threat of war, the instability of liberal democracy, the rise of Fascism. Not surprisingly dystopian rather than utopian speculation texts figured prominently in the literature of this period, the most notable examples being Zamyatin's *We* (1920), Huxley's *Brave New World* (1932) and Orwell's *Nineteen Eighty-Four* (1949). The realities of Fascism and Stalinism were taken as warnings of what a modern scientific-technological systems, established in the name of utopianism, might be like." The focus on one main social problem makes almost all utopian writing and almost all intentional societies one-dimensional and therefore potentially totalitarian. Since literary (and attempted) utopias are based on a single main ideology this intense focus can lead utopian aspiration over the line into the temptation to force all member of the society to adhere to the precepts and procedures of the supposedly ideal society. That is, if the founders of the utopian experiment have great certainty that they know best, they may be willing to exert brutal means to bring citizens in line with their vision. Total control over

communications with the outside world has been characteristic of every twentieth-century totalitarian society with utopian ambitions. Twentieth-century 'utopians' like Stalin, Mao, Pol Pot, and Kim Il-sung all made their societies social, cultural, and 'economic islands,' and in doing so, also made them dystopias.

Utopianism Promoted Reform

A utopia is a society that attempts to eliminate all social problems. By that definition, because human beings are imperfect, no human society can ever completely avoid all suffering, deprivation, friction, or selfishness among its members. All but a few attempted utopias of the nineteenth and early twentieth century failed in terms of their own standards, but they did lead the rest of society toward meaningful reforms in distribution of resources, the reconceptualization of work, greater equality for women and nonwhites, and even greater productivity through technological innovation.

One example of failure that nevertheless produced social progress can be seen in the efforts of Robert Owen. The Owenite movement established several communities according to communitarian and cooperative principles. One of the best known of these efforts, which were largely unsuccessful, was the project at New Harmony, Indiana, which started in 1825 and was abandoned by 1829. In spite of this short-lived project, Owenism is also closely associated with the development of the British trade union movement, and with the spread of the Mechanics movement, and therefore had a lasting impact on society.

Conclusion

Just as the title of Sir Thomas More's fictional travelogue combining the 'eu' the Greek word root for good (as in eulogy or euthanasia) and 'u' meaning 'not' plus 'topia' meaning place, the words 'utopia' and 'utopian' have many and contradictory connotations: ideal, visionary, oriented toward reform, quixotic, unrealistic, or even totalitarian, as in the title of the biography of one of the ideological architects of the Iranian Revolution, *An Islamic Utopian*.

The intellectual intolerance of Plato and moral intolerance of the Bible produced one particular type of utopian thinking that dominated fictional visions as well as social experiments – a categorical certainty that resulted in burning religious reformers at the stake, guillotines and gulags to purify society from dissent and difference. Until and unless utopian thinking can learn to entertain several mutually contradictory ideas at the same time it will lead to social oppression in practice and remain suspect in theory. If utopian literature is to continue to have relevance and impact on the thinking of social planners, it must resolve this quandary.

Bibliography

Ahlstrom, Sydney E., 1972. A Religious History of the American People. Yale University Press, New Haven, London, p. 263.

Al-Farabi, Abu Nasr. 1985. On the Perfect State. In: Richard Walzer, Seyyed Hossein Nasr (Trans. and Eds.). Clarendon Press, Oxford.

Bury, John Bagnell, 1920/1932/1955. The Idea of Progress: An Inquiry into its Origin and Growth. Macmillan, London. New York: Macmillan reprint; New York: Dover. Chapter VI: The General Progress of Man: Abbe de Saint-Pierre. http://www.gutenberg.org/files/4557/4557-h/4557-h.htm.

Bellamy, Edward, 1888. Looking Backward 2000–1887. Ticknor and Company, Boston.

Friedman, Edward, Gilley, Bruce, 2005. Asia's Giants: Comparing China and India. Palgrave Macmillan, New York.

Fourier, Charles, 1967[written 1816–1818]. Le Nouveau Monde Amoureux. Éditions Anthropos, Paris, pp. 389, 391, 429, 458, 459, 462, and 463.

Goodwin, Barbara, Taylor, Keith, 1982/2009. The Politics of Utopia: A Study in Theory and Practice. Peter Lang AF, International Academic Publishers, Bern, p. 42.

Goffman, Erving, 1961. Asylums: Essays on the Social Situation of Mental Patients and Other Inmates. Anchor Books. Doubleday, New York.

2006. Gospel of Judas. (Kasser, Rodolphe, Marvin Meyer, Gregor Wurst, in collaboration with François Gaudard, Trans.). National Geographic Society, Washington, DC. http://www.nationalgeographic.com/lostgospel/pdf/GospelofJudas.pdf and http://www.nationalgeographic.com/lostgospel/timeline_24.html.

Hine, Robert V., 1953. California's Utopian Colonies. University of California Press, Berkeley, p. 5.

Halberstam, David, 1994. The Fifties. Ballantine Books, New York.

Hicks, George, 1987. Experimental Americans: Celo and Utopian Community in the Twentieth Century. University of Illinois Press, Urbana and Chicago, p. 31.

Huxley, Aldous, 1932. Brave New World: A Novel. Doubleday, Doran & Co, Garden City, N.Y.

Kidd, Thomas, 2007. The Great Awakening: A Brief History with Documents. Bedford/St. Martin's, Boston and New York.

Landes, Richard, January 2000. The fear of an apocalyptic year 1000: augustinian historiography, medieval and modern. Speculum 75 (01), 97–145. A more popular treatment of the subject. http://www.mille.org/scholarship/1000/AHR9.html.

L'Engle, Madeleine, 1962. A Wrinkle in Time. Ariel Books Farrar Strauss & Cudahy, New York.

More, Thomas, 1518/1895. The Utopia of sir Thomas more. Latin from the Edition of March 1518, and in English from the 1st Ed. of Ralph Robynson's Translation in 1551. Clarendon Press, Oxford. Available as an: http://books.google.com/books?id=6REuAAAAMAAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false.

On Revivals of Religion, 1835. Biblical Repertory and Theological Review, vol. 7. No. 4. pp. 626–674.

Orwell, George, 1949. Nineteen Eighty-four. Buccanneer Books, New York. Available: https://archive.org/details/ost-english-1984-george-orwell-1937-dystopia.

Pitzer, Donald. E., 1997. America's Communal Utopias. University of North Carolina Press, Chapel Hill, p. 272.

Rahnema, Ali, 1998. An Islamic Utopian. A Political Biography of Ali Shariati. I.B. Tauris.Oxford: Clarendon Press, London. Portions available: http://www.sacred-texts.com/isl/palf/palf10.htm.

Smith, Craig, Summer 2005. Great reformation in the manners of mankind. Utopian thought in the Scottish Reformation and Enlightenment in Utopian Studies 16 (2), 221–245.

Sargent, Lyman Tower, 2010. Utopianism: A Very Short Introduction. Oxford University Press, Oxford.

Skocpol, Theda, 1979. States and Social Revolutions: A Comparative Analysis of France, Russia, and China. Cambridge University Press, New York.

Unpublished interview: Davidowitz, Amos. Interview by Adrienne Redd. Personal conversation and tour of Gezer Kibbutz. Gezer Kibbutz.

Weber, Max, 1958[1902]. The Protestant Ethic and the Spirit of Capitalism. Scribner, New York, p. 121.

Weber, Max, 1995[1905]. The Russian Revolutions. In: Gordon C. Wells, Peter Baehr (Trans. and Eds.). Cornell University Press, Ithaca. pp. 67, 77, 119, 251.

Attention-Deficit and Hyperactivity Disorders, Neural Basis of

Irene M Loe and Heidi M Feldman, Stanford University School of Medicine, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Understanding that attention-deficit/hyperactivity disorder (ADHD) had a neural basis was first based on association of neurological conditions with behavioral manifestations of the disorder. Since those early observations, ADHD diagnosis continues to be made on the basis of clinical symptoms, although the psychological and biological mechanisms underlying the disorder have been better elucidated. Recent advances in neuroscience coupled with genetic, environmental, and psychopharmacology studies shed light on the neural basis of the disorder. Structural, functional, and connectivity imaging studies of the brain show distinct differences in ADHD. Neurotransmitter systems are also affected by the disorder and targeted in treatment.

Historical Background

Early references to children with attention-deficit/hyperactivity disorder (ADHD) symptoms made no mention of possible neural origins to the condition. A German physician, Heinrich Hoffman, wrote a children's book in 1848 in which he described two characters, 'Fidgety Phil,' who demonstrated hyperactive/impulsive symptoms and was unable to sit still, and 'Johnny Head-In-Air,' who demonstrated inattentive symptoms and, while distracted, fell into a river (Hoffman, 1848). ADHD was first described behaviorally in the medical literature as a 'defect of moral control' in children with ADHD symptoms by George Still over 100 years ago (Still, 1902).

The possibility of a neurobiological basis for the disorder came decades later in the 1920s, when children recovering from influenza encephalitis caused by a worldwide epidemic in 1917 exhibited symptoms of restlessness, inattention, impulsivity, easy arousability, and hyperactivity (Hohman, 1922; Ebaugh, 1923). In 1937, the serendipitous discovery that racemic amphetamine, Benzedrine, improved ADHD symptoms in children provided additional support for a neurobiological basis for the disorder since stimulants affect brain neurotransmitters (Bradley, 1937). Various types of brain injury have been described in association with ADHD symptoms; however, the presentation of symptoms in children without obvious brain damage led to the name 'minimal brain disorder' or 'minimal brain dysfunction.' Stimulants such as methylphenidate and amphetamines were not routinely used for ADHD until the late 1950 to 1960s. Increasing use of stimulants coincided with a switch in nomenclature in the late 1960s that focused on behavioral symptoms rather than etiology, evolving from hyperkinetic syndrome to attention-deficit disorder, and most recently, to ADHD. At that time, little was known about brain mechanisms underlying pathogenesis and treatment. Clinical description of the disorder, including diagnosis and treatment, is included in the preceding article on ADHD. ADHD diagnosis continues to be made on the basis of clinical symptoms, although the psychological and biological mechanisms underlying the disorder have been better elucidated. Recent advances in neuroscience, coupled with genetic, environmental, and psychopharmacology studies, shed light on the neural basis of the disorder through structural, functional, and connectivity imaging studies of the brain, studies of neurotransmitter systems affected by the disorder and targeted in treatment, and interactions with genes and environment.

Neuropsychological Impairments

ADHD has been characterized by a wide variety of cognitive and neuropsychological impairments that accompany the behavioral symptoms of inattention, hyperactivity, and impulsivity used in diagnosis (Nigg, 2005). Various lab-based paradigms have been developed to tap different neuropsychological skills categorized under the umbrella term of executive function or cognitive control. Executive function skills are often broken down into individual constructs such as response inhibition, working memory, cognitive flexibility, organization, and planning. In addition, attention is often subdivided into different skills such as vigilance, orienting, and selective, divided, and sustained attention. Other systems involved in motor planning, motivation and reward, and timing or temporal expectations have also been implicated in ADHD. However, the presence of such impairments varies widely among individuals with ADHD and may or may not be captured by lab-based measures or parent-, teacher-, or self-ratings of such skills (Nigg, 2005). Nonetheless, these neuropsychological skills have been linked to various brain regions, originally through associations with brain lesion or injury studies, and now through anatomic and functional imaging.

Neurobiological Basis

Structural Imaging Studies

Structural imaging studies of the brain in ADHD often focus on brain volume, surface area, or both. Investigators' early focus included frontostriatal regions based on findings from neuropsychological studies in patients with brain lesions. However, findings from several studies indicate that more widespread regions or networks are involved in ADHD. Results vary from study to study. Findings of smaller volumes have been reported in the dorsolateral prefrontal cortex, caudate, pallidum, corpus

callosum, and cerebellum in children with ADHD compared to controls (Seidman et al., 2005). Findings in the caudate have varied, with results ranging from equal, larger, or smaller caudate measurements in ADHD subjects compared to controls (Valera et al., 2007). Differences in findings across studies may be due, at least in part, to low power as many structural studies have had small sample sizes (i.e., often <20 subjects/group). In addition, the influence of medications, comorbid conditions, and gender as well as other sources of heterogeneity in the populations such as ADHD subtype, family history of ADHD, and perinatal complications often were not considered (Seidman et al., 2005). Although sex-specific behavioral differences in ADHD are well described, the current neuroimaging literature, which overall has fewer female than male participants, shows no consensus on sex-specific brain anomalies or differences (Mahone and Wodka, 2008).

A meta-analysis was performed to help resolve inconsistencies across structural imaging studies regarding which brain regions are most affected or have the largest volumetric reductions in children with ADHD (Valera et al., 2007). Pooling results from 21 high-quality studies, global reductions in multiple regions characterized children with ADHD compared to controls. The regions that were most frequently assessed and showed the largest differences included cerebellar regions, the splenium of the corpus callosum, total and right cerebral volume, and the caudate. Significant differences between several frontal regions and the cerebellum were also found, although it was advised that these results be interpreted with caution due to concerns about low power (Valera et al., 2007).

High-resolution structural MRI scans have been used to quantify cortical thickness in children with ADHD. They show significant global thinning of the cortex, which was most prominent in prefrontal and lateral temporal cortices (Shaw et al., 2006). In addition, the ADHD group also had a slower rate of prefrontal cortical thinning during late childhood and adolescence (Shaw et al., 2006). Applying the same longitudinal methods to typically developing children with varying levels of hyperactivity and impulsivity but no diagnosis of ADHD, children with higher symptom levels also demonstrated a slower rate of cortical thinning in several regions: predominantly in prefrontal cortical regions, bilaterally in the middle frontal/premotor gyri, extending down the medial prefrontal wall to the anterior cingulate; the orbitofrontal cortex; and the right inferior frontal gyrus (Shaw et al., 2011). The rate of cortical thinning was linked to the severity of symptoms, mirroring the slower rate of cortical thinning in adolescents with ADHD, who demonstrated the slowest rate. Similar to the findings from epidemiologic studies (Lubke et al., 2009), the results lend neurobiological support for viewing ADHD diagnoses dimensionally, lying at the extreme end of a continuum of symptoms, rather than as discrete categories, since typically developing children with symptoms but no diagnosis demonstrated similar neurodevelopmental changes in the brain (Shaw et al., 2011).

Despite the findings of group differences, structural imaging remains ineffective in its use as a clinical diagnostic tool at the individual patient level. However, early structural imaging findings were also found to relate to functional impairments in children with ADHD (Casey et al., 1997). Efforts to explore structure–function relationships have been greatly expanded through functional imaging studies.

Functional Imaging Studies

Early functional imaging modalities include single photon emission computed tomography (SPECT), first used in the 1950s, and positron emission tomography (PET), which originated in the 1970s. SPECT is an imaging method that uses injected or inhaled radiopharmaceuticals, such as xenon-133 or iodine-123, that distribute throughout the body, including the brain, and emit gamma rays as they decal; higher blood flow, and therefore higher amount of radioactive tracer, occurs in more highly active brain areas which is quantified by SPECT imaging. PET is an imaging method that uses injected or inhaled radiopharmaceuticals that emit positrons during decay which are detected by PET imaging; methods may be flow dependent although others measure cerebral metabolism rates. The majority of SPECT studies in ADHD have had methodological concerns, such as lack of or an inadequate control group. Among the studies of better design, findings include decreased striatal perfusion and increased regional cerebral blood flow after methylphenidate administration in dorsolateral prefrontal cortex, caudate, and thalamus in previously treatment-naïve children and adolescents with ADHD (Bush et al., 2005). A PET study of adults with ADHD reported significantly decreased global cerebral glucose metabolism of 8.1% (Zametkin et al., 1990). Although there were concerns about gender differences affecting the results, separate gender analyses and another study suggest that global metabolism is likely reduced in ADHD (Bush et al., 2005).

The use of PET and SPECT remains limited due to lack of ethical justification for radiopharmaceutical exposure in healthy children and/or volunteers. PET studies have been limited mainly to adults and a few studies of adolescents. SPECT also has poor spatial and temporal resolution. PET and SPECT have largely been replaced by functional MRI (fMRI, imaging technique that uses the blood-oxygen-level-dependent contrast to image the change in blood flow (hemodynamic response) related to energy use by brain cells; allows for inference of the regions of brain activation involved during a cognitive task), but continue to be used for drug and neurotransmitter transporter studies (Bush et al., 2005). Other techniques used to study ADHD include magnetic resonance spectroscopy, a noninvasive MRI-based method to quantify neurochemicals that might be abnormal in ADHD, and electrophysiologic methods such as quantitative electroencephalogram and event-related potentials. These methods are beyond the scope of this chapter and are not described in detail here due to their lack of spatial resolution.

fMRI has been paired with both cognitive tasks and the resting state to identify brain regions that are normally involved in attention, cognition, and various executive functions, and potentially abnormal in ADHD. We first review studies on the roles of cortical areas and then review neurocircuitry models between prefrontal cortex and subcortical regions in the subsequent section.

Cortical Dysfunction

Cingulate, frontal, and parietal cortical regions have been implicated in the pathophysiology of ADHD (Bush, 2011). These regions have been described as part of a cognitive-attention network in ADHD; in addition, striatum, premotor areas, thalamus, and cerebellum serve as nodes within parallel networks of attention and cognition. Cingulate subdivisions of dorsal anterior midcingulate cortex (daMCC, also previously referred to as the anterior cingulate cortex or ACC), dorsolateral prefrontal cortex, and ventrolateral prefrontal cortex have been studied for their individual contributions to pathogenesis of ADHD.

daMCC serves critical roles in a variety of cognitive processes, including attention, cognitive processing, target detection, novelty detection, response selection, response inhibition, error detection, and motivation. Given the wide variety of functions subserved by daMCC, it is relevant to reward/motivation and cognitive theories of ADHD as it may help modulate adjustments in behavior by influencing feedback-based decision making. Hypofunction of the daMCC in children with ADHD compared to healthy controls has been found in a number of fMRI studies using a variety of response inhibition and timing tasks (Bush, 2011). Findings of hypoactivation of daMCC were confirmed in a meta-analysis of neuroimaging studies, which also identified hypofunction of a wider network of regions, including dorso- and ventrolateral prefrontal cortex, superior parietal cortex, caudate, and thalamus (Dickstein et al., 2006). Other imaging modalities such as PET and event-related potential studies have also confirmed daMCC hypofunction in ADHD (Bush, 2011).

Studies of lateral prefrontal cortex have focused on dorsolateral and ventrolateral prefrontal cortex, as these regions are implicated in functions affected in ADHD such as vigilance, selective and divided attention, attention shifting, planning, executive control, and working memory (Bush, 2011). Smaller volumes in these regions in ADHD have been found in structural imaging studies. fMRI studies have also found dysfunction in dorso- and ventrolateral prefrontal cortex using tasks that tap response inhibition (Bush et al., 2005). Meta-analysis that focused only on response inhibition tasks identified ventrolateral prefrontal cortex, daMCC, parietal cortex, caudate, and precentral gyrus, but not dorsolateral prefrontal cortex, suggesting that abnormalities in dorso- and ventrolateral prefrontal cortex may contribute to ADHD in different ways (Dickstein et al., 2006).

The parietal cortex plays important roles in attention and spatial processing, specifically in attention allocation. Although the parietal cortex has been less well studied compared to other cortical areas, a number of functional imaging studies have identified parietal hypoactivation in children with ADHD during a variety of tasks, including a spatial working memory mental rotation task, task switching, mental rotation/spatial processing, and sequential finger tapping (Bush, 2011). Although hypofunction may represent primary abnormality in the parietal cortex, it may also reflect abnormal input from regions connected to it (Bush, 2011).

It has been proposed that the multiple functions subserved by the cortical regions reviewed may be subdivided into the following primary roles: dorsolateral prefrontal cortex, overall planning and goal-setting; ventrolateral prefrontal cortex and daMCC, inhibition of excessive or inappropriate motor behavior; parietal cortex, target detection and attention shifts; and daMCC, integration of information from these inputs and execution of plans by modifying behavior moment-to-moment. Hence, any dysfunction within these regions could lead to the behavioral manifestations of ADHD – inattention due to inability to detect targets or shift attention appropriately; hyperactivity due to failure to inhibit motor activity or use reward-and-error feedback to modify behavior; and impulsivity due to inability to encode motivational goals and/or pursue long-term goals over short-term goals (Bush, 2011). These cortical regions interact with other brain regions, including subcortical regions. We next review links among prefrontal cortex, striatum, and cerebellum.

Cortical and Subcortical Neurocircuitry Dysfunction

The prefrontal cortex has long been thought to exert control over behavior, and is especially important in cognitive control, or the ability to flexibly adjust behavior to changing circumstances. Inputs to and from prefrontal cortex are also vitally important to various functions; hence studies have focused on circuits to and from prefrontal regions. The main prefrontal cortex circuits implicated in ADHD include (1) dorsal frontostriatal, linked to cognitive control; (2) orbitofrontostriatal, linked to reward and motivation; and (3) frontocerebellar, linked to timing and building temporal expectations (Durston et al., 2011). These circuits interact with each other through spiraling loops in the striatum and connections from the cerebellum to the prefrontal cortex and striatum, but also have many more connections within circuits than between circuits.

The dorsal frontostriatal circuit comprises reciprocal connections among the striatum, thalamus, and prefrontal areas, and is critical for cognitive control. Functional imaging studies have shown differences in frontostriatal activity during cognitive control tasks (Dickstein et al., 2006; Durston et al., 2011). Additional support for the importance of connections between prefrontal cortex and striatum has been found in typical development, in which greater microstructural organization of white matter tracts predicted developmental improvements in cognitive control (Liston et al., 2006).

The orbiotofrontostriatal circuit comprises loops between ventral striatum and orbitofrontal cortex and is implicated in reward processing. It has been proposed that differences in sensitivity to reinforcement lead to changes in motivation, ultimately resulting in ADHD symptoms. Decreased ventral striatal activation in adolescents and adults with ADHD in anticipation of reward has been reported in three fMRI studies (Durston et al., 2011). It has been proposed that hypoactivation in response to reward might lead to increased impulsive reward-seeking behavior in ADHD (Durston et al., 2011).

The frontocerebellar circuit includes projections from the cerebellum via the thalamus to many areas of the prefrontal cortex, although there are also likely more direct connections between basal ganglia and cerebellum as well. The cerebellum has long been recognized for its roles in motor control and coordination, but now has also been found to have a role in cognitive functions and learning (Durston et al., 2011). fMRI studies have found attenuated cerebellar activation in ADHD during cognitive tasks tapping cognitive control, working

memory, and timing as well as resting-state fMRI (Durston et al., 2011).

Summary

Dysfunction in several brain regions and neural circuits has been identified in ADHD. It has been proposed that many pathways may lead to the clinical phenotype of ADHD and that perhaps identification of specific prefrontal–subcortical circuits could be used to subtype ADHD (Durston et al., 2011).

Connectivity Studies

The study of connections between prefrontal and subcortical regions, among others, in ADHD has been enhanced by relatively new imaging techniques including diffusion tensor imaging (DTI) and resting-state fMRI. These techniques allow for study of both structural and functional connectivity.

Structural Connectivity

DTI is an MRI technique that generates quantitative measures of white matter microstructure in the brain. It quantifies anisotropic diffusion, the degree to which diffusion is greater in one direction than another, to approximate microstructural integrity of white matter (Feldman et al., 2010). Since myelinated axons restrict water diffusion such that water diffuses more readily in parallel rather than perpendicular to a white matter tract, anisotropy is thought to be proportionate to the degree of myelination, orientation, regularity, and density of axons in a tract. Hence, higher fractional anisotropy (FA, a primary outcome measure generated by DTI which is a ratio ranging from 0 to 1; represents the degree to which diffusion is anisotropic. High FA values indicate that diffusion is greater in one direction than others, whereas low FA values indicate that diffusion is nearly equal in every direction) is thought to reflect more mature or healthy white matter.

DTI studies in children and adolescents with ADHD yield mixed results with findings of reduced (Nagel et al., 2011; Ashtari et al., 2005), increased, or mixed FA (Peterson et al., 2011; Davenport et al., 2010) in multiple regions and/or tracts compared to controls. A variety of tracts have been identified as abnormal in ADHD, including frontostriatal projections, superior longitudinal fasciculus, and tracts involving the cerebellum, corpus callosum, basal ganglia, anterior cingulate, and corticospinal tract (Liston et al., 2011; Konrad and Eickhoff, 2010). Out of 8 DTI studies, 7 studies found structural connectivity deficits in frontostriatal projections (or anterior corona radiata) in ADHD compared to control subjects, whereas 6 of 9 DTI studies that examined the superior longitudinal fasciculus found structural connectivity deficits (Liston et al., 2011). Two studies of adult ADHD report reduced FA in the superior longitudinal fasciculus.

DTI studies have been performed in children born preterm as this population is at high risk for both overt and subtle, diffuse white matter injury. Children born preterm are also at elevated risk for ADHD, and studies often report high levels of attention difficulties even if children do not meet full criteria for the disorder (Johnson and Marlow, 2011). Hence, children born preterm serve as a neurobiologic risk model for ADHD and associated problems. DTI studies of children and adolescents born preterm have found significant associations between white matter properties and attention within preterm samples; regions of association include the corpus callosum and bilateral internal capsules (Skranes et al., 2007; Nagy et al., 2003), left external capsule, bilateral inferior fasciculi, left middle fasciculus, and right superior fasciculus (Skranes et al., 2007). Another study found that both attention and internalizing problems, such as anxious and withdrawn symptoms, were associated with white matter FA in several distributed and overlapping regions, suggesting common underlying neurobiology for behavioral symptoms in this high-risk population (Loe et al., 2013).

An innovative approach to study structure and function combines DTI with fMRI by applying fiber tractography to reconstruct white mater tracts and assess the role of connectivity within anatomically specified projections. A study of children with ADHD and their parents used fMRI and DTI tractography during a cognitive control task. Results showed that tract-specific connectivity in white matter projections from the striatum to the prefrontal cortex correlated positively with activity in the caudate nucleus and inferior frontal gyrus during the cognitive control task (Casey et al., 2007). In addition, individual differences in connectivity as measured by FA predicted performance for children with ADHD and their parents but not control subjects (Casey et al., 2007). Other regions identified as having abnormalities in ADHD using fiber tractography and other DTI measures of connectivity include frontostriatal projections, superior longitudinal fasciculus, cingulum bundle, and cerebellar peduncles (Liston et al., 2011).

The DTI studies described above are heterogeneous in terms of different analytic approaches, such as whole brain, region of interest, and tract-based statistics. Alterations in superior longitudinal fasciculus and frontostriatal white matter appear to be the consistent findings at this time (Liston et al., 2011).

Functional Connectivity

Functional connectivity estimates the strength of long-range connections by inferring the aggregate strength of functional connections between two areas of cortex. It is often defined as the temporal correlation or coherence of spatially remote neurophysiological events (Konrad and Eickhoff, 2010). Although correlated activity might result from a third area that drives activity between two areas or reflect multisynaptic relays across multiple intermediate areas of cortex, it is generally thought that estimates of structural connectivity correlate well with functional connectivity (Liston et al., 2011).

Two general approaches for assessing functional connectivity are (1) the task-based fMRI approach and (2) resting-state fMRI. The task-based approach measures correlated activity in different cortical areas that are coactivated when a specific function is performed during a task (i.e., response inhibition during a go-/no-go task). Five task-based studies in children with ADHD that have examined group differences in correlation strength between regions found connectivity deficits (Liston et al., 2011). One study found reduced coupling in frontocerebellar and frontoparietal circuits during a combined interference and time discrimination task. Two studies found that methylphenidate normalized or ameliorated connectivity deficits and improved behavioral performance during a task

requiring selective attention and suppression of competing information and a rewarded continuous performance task (Liston et al., 2011). In resting-state fMRI, the subject is imaged 'at rest' with no specific task to perform. A specific default mode network (DMN) has been identified from resting-state fMRI, which includes a large and replicable network of brain regions associated with task-irrelevant mental processes and mind wandering. The DMN comprises the precuneus/posterior cingulate cortex, the medial prefrontal cortex, and the medial, lateral, and inferior parietal cortex (Konrad and Eickhoff, 2010). During rest, the DMN shows higher activity and stronger functional connectivity than during externally driven tasks. DMN activity is attenuated, but not extinguished during transition from rest to task states. Stronger deactivation of the DMN has been associated with increased task difficulty. DMN activity is negatively correlated with lateral prefrontal cortex activity during attentional and executive function tasks, and it has been proposed that failure to suppress DMN activity sufficiently interferes with task performance (Liston et al., 2011; Konrad and Eickhoff, 2010). In contrast with the DMN, a second network consisting of dorsolateral prefrontal cortex, intraparietal sulcus, and the supplementary motor area has been identified as active during tasks that require active allocation of attentional resources. These regions have been associated with increased alertness, response preparation, and selective attention in a manner that is largely independent of the specific task. This network is referred to as the 'task-positive' frontoparietal network. It has been found to be anticorrelated (negative correlation in which one value increases as the other decreases; used to describe resting state fMRI data and generally observed between competing neural systems) with the DMN, such that the DMN is attenuated when there is task-specific activation of the task-positive network and vice versa (Konrad and Eickhoff, 2010). Studies using resting-state fMRI in ADHD in general have shown that for both ADHD and control groups, DMN activity is anticorrelated with activity in regions that have important roles in cognitive control such as ACC (Castellanos et al., 2008; Liston et al., 2011). However, the ADHD group often shows a weaker relationship between the DMN and regions implicated in cognitive control, consistent with failure to suppress the DMN activity during tasks requiring cognitive control.

Summary

Relatively new methods of analysis to assess functional connectivity suggest that there are altered interactions between networks implicated in ADHD or changes in DMN network activity. Varied tasks, methods, and analytic strategies do not allow meta-analytic review of connectivity studies at this time. Although correlations do not equal causality, early work showing improvements in both behavioral performance and connectivity with stimulants suggests that these connectivity deficits may play a role in ADHD symptomatology (Liston et al., 2011).

Neurotransmitter Systems

Dysregulation of dopamine and noradrenaline has been implicated in the pathophysiology of ADHD, and these catecholamines are affected by first-line pharmacotherapy with psychostimulants such as methylphenidate and dextroamphetamine (*see* Neurotransmitters). The neurotransmitters modulate activity in fronto-striato-cerebellar circuits, affecting executive functions that are often deficient in ADHD. PET and SPECT studies have been used to assess neurotransmitters directly *in vivo*, at baseline or in response to medication. However, the precise neurobiological mechanisms underlying pathogenesis and treatment remain poorly understood.

Dopaminergic System

The dopamine transporter (DAT) is a protein that spans the cell membrane and serves to regulate dopamine by pulling it from the synaptic space into the presynaptic neuron. There are at least five dopamine receptors on the postsynaptic neuron, each of which is coded by a distinct gene (D1–D5). Receptor D_4 has been of particular interest in ADHD because of polymorphisms in a region of tandem repeats.

Imaging studies of the DAT have yielded mixed results, with early studies showing increased density in ADHD and more recent studies in medication-naïve adults with ADHD showing reduced DAT and D_2/D_3 receptor availability in subcortical regions (Del Campo et al., 2011). These differences have been attributed to differences in medication history as well as genetic differences, including polymorphism of the DAT1 gene (Del Campo et al., 2011). Limitations of imaging studies of dopamine receptors include the low affinity of radiotracers, which restricts the focus to brain areas high in D_2/D_3 density, such as the striatum, rather than prefrontal cortex where low D_2/D_3 receptor density and inadequate neurotransmission may contribute to ADHD. Studies to examine other dopamine markers also show disparate results due to many methodological and individual patient factors that influence expression of dopaminergic markers (Del Campo et al., 2011).

PET studies have confirmed methylphenidate blockade of DAT, resulting in increased extracellular dopamine in the striatum. Levels were not correlated with blockade, indicating that other factors contribute to dopamine release. Animal studies using microdialysis also show increases in dopamine in extrastriatal areas such as frontal cortex with stimulant administration. In the frontal cortex, increased dopamine levels after stimulant administration are attributed to the noradrenaline transporter (NAT), which has high density in the frontal cortex and low density in the striatum, converse to DAT density in those areas. Dopaminergic systems in frontostriatal and nigrostriatal networks (midbrain to striatum) and interactions between prefrontal cortex and subcortical systems are also the focus of investigations to understand the therapeutic effects of psychostimulants (Del Campo et al., 2011).

Noradrenergic System

Noradrenaline has long been implicated in pathophysiology of ADHD but is less well studied than dopamine. Noradrenaline is thought to exert effects primarily in prefrontal cortex as there is sparse noradrenergic innervation in the striatum. Noradrenaline projections to prefrontal cortex originate from the locus

coeruleus. Multiple pharmacologic agents for ADHD have important effects on noradrenaline transmission, including methylphenidate, dextroamphetamine, buproprion (an indirect dopamine and noradrenaline agonist), tricyclic antidepressants, guanfacine (an alpha-2 receptor antagonist), and atomoxetine (selective noradrenaline reuptake inhibitor) (Del Campo et al., 2011). Molecular imaging of the noradrenaline system has been limited due to lack of suitable tracers, although recent work using PET shows that clinical doses of methylphenidate significantly reduce NAT availability in a dose-dependent manner in dense receptor regions, including locus coeruleus, raphe, hypothalamus, and thalamus (Del Campo et al., 2011).

Psychostimulant medications are the first-line medical treatments for ADHD. Methylphenidate blocks reuptake of dopamine and noradrenaline, thereby increasing extrasynaptic transmitter levels. Dextroamphetamine also inhibits reuptake of dopamine and noradrenaline, but also increases extracellular levels by promoting efflux of the neurotransmitters into the extraneuronal space and inhibiting catabolic activity of monoamine oxidase (Del Campo et al., 2011). Initially, the 'calming effect' of psychostimulants on behavior was thought to be a 'paradoxical' reaction unique to individuals with ADHD; however, the behavioral-calming and cognition-enhancing effects of low and clinically relevant doses of psychostimulants were observed in humans without ADHD as well as animals (Berridge and Devilbiss, 2011). The effects of psychostimulants on prefrontal cortex-dependent behavior and cognitive processes are observed acutely and without tolerance that occurs with many medications after extended treatment (Berridge and Devilbiss, 2011). Research on the mechanisms underlying cognition enhancing and therapeutic effects of low-dose stimulants indicates that low-dose stimulants preferentially increase extracellular levels of catecholamines in prefrontal cortex and preserve phasic signaling of catecholamine neurons (Berridge and Devilbiss, 2011). There are regionally selective neurochemical actions of low-dose stimulants and enhanced neuronal responding to afferent signals within prefrontal cortex. This regionally specific effect on neuronal signal processing in prefrontal cortex contrasts sharply with the widespread action of higher and behaviorally activating doses of stimulants, which also results in profound suppression of prefrontal cortex neuronal signaling. In addition, rather than acting directly on dopamine as a primary target, pharmacologic treatments for ADHD appear to involve both dopamine and noradrenaline acting at postsynaptic alpha-2 and D_1 receptors in the prefrontal cortex (Berridge and Devilbiss, 2011).

The evidence that stimulants improve core symptoms of ADHD and exert effects at the neurobiological level is well-established. However, individual variation in treatment response has prompted pharmacogenetic studies to identify genetic predictors of treatment effects. The underlying premise of pharmacogenetics is that by studying individual candidate genes, failure of or variation in response can be linked to specific gene variants that affect receptors, transporters, or drug-metabolizing enzymes (Froehlich et al., 2010). Pharmacogenetic studies in ADHD have focused on catecholamine-related polymorphisms that affect the DAT gene and dopamine receptor D_4 (DRD4). For example, polymorphisms in a variable number tandem repeat in the 3′-untranslated region of the DAT gene have been studied for their possible role in predicting treatment response since DAT inhibition is a primary mechanism of methylphenidate. Other studies have focused on polymorphisms in DRD4, which regulates dopamine synthesis, release, and neuron firing rate. However, results from DAT and DRD4 studies have been mixed in terms of links between various alleles and methylphenidate response (Froehlich et al., 2011b). Other genes of interest include the adrenergic alpha-2A receptor, catechol-O-methyltransferase, which catabolizes dopamine and norepinephrine, D_5 receptor, NAT protein 1, and synaptosomal-associated protein.

Results have been inconsistent across studies with disparate results attributed to small samples; variation in study design (open label vs randomized controlled trials); and differences in samples, outcome measurements, and dosing regimens (Froehlich et al., 2010). A recent, well-designed study of stimulant-naïve children with ADHD examined the role of four catecholamine-related candidate genes in moderating methylphenidate dose response (Froehlich et al., 2011b). This randomized, double-blind, crossover trial of long-acting methylphenidate included parent and teacher evaluations of response to placebo and three methylphenidate dosage levels. They found significant gene-by-dose interactions for hyperactive-impulsive symptoms for DAT and DRD4 polymorphisms. The DAT 10-repeat allele and DRD4 7-repeat allele have each been associated with ADHD risk in other studies, whereas the more common DRD4 4-repeat allele is considered low risk. In this randomized trial, children who lacked the DAT 10-repeat allele showed greater improvement in symptoms with increasing dose compared with 10-repeat carriers. Children who lacked the DRD4 4-repeat allele showed less improvement across doses compared with 4-repeat carriers (Froehlich et al., 2011b). The findings, in which children lacking the DAT 10-repeat allele and those with the DRD4 4-repeat allele had more robust response to methylphenidate compared to other genotypes, are consistent with an improved response for the ADHD-susceptibility low-risk alleles (Froehlich et al., 2011b). The results suggest that DAT and DRD4 polymorphisms may be associated with individual variability in dose response to methylphenidate. With replication, exciting findings such as these could be used to guide tailored prescribing in an era of personalized medicine in the future.

Other Interactions with Neurobiology

Risk factors associated with ADHD include environmental toxins such as lead, pesticides, and industrial chemicals; nutritional factors such as low child zinc and omega-3 fatty acid levels; and lifestyle and psychosocial factors, such as maternal stress, early trauma, and institutional care (Froehlich et al., 2011a). The mechanisms by which these factors exert ADHD-related neurobehavioral effects go through changes to the neurobiological systems described above. For example, lead exposure leads to disruption of synapse formation and

derangements to the brain dopaminergic system (Froehlich et al., 2011a). Twin and family studies of ADHD indicate high heritability, on the order of 0.6–0.8; these figures also reflect that environmental influences also play an important role in disorder susceptibility (Froehlich et al., 2011a). New studies of gene–environment interactions in ADHD may elucidate how both genetics and environmental influences contribute to ADHD. By including gene and/or environmental factors in neuroimaging studies, we may also be able to determine the multiple pathways contributing to neurobiological differences and manifestations of ADHD.

Conclusions

Our understanding of ADHD has greatly evolved since it was first described in the medical literature as a 'defect of moral control.' Recent advances in neuroscience, coupled with genetic, environmental, and psychopharmacology studies, shed light on the neural basis of the disorder and may provide guidance for prevention and treatment. Structural, functional, and connectivity imaging studies of the brain have greatly contributed to our understanding of the networks involved in pathogenesis of the disorder. However, these neuroimaging methods are ineffective at the level of individual diagnosis and treatment and remain research tools at this time. Genetic evaluations may ultimately be able to identify which individuals with ADHD will be responsive to specific medications to improve function.

See also: Attention, Neural Basis of; Brain Development: Ontogenetic Neurobiology of; Dysexecutive Syndromes; Language Learning Impairment.

Bibliography

Ashtari, M., Kumra, S., Bhaskar, S.L., et al., 2005. Attention-deficit/hyperactivity disorder: a preliminary diffusion tensor imaging study. Biological Psychiatry 57, 448–455.

Berridge, C.W., Devilbiss, D.M., 2011. Psychostimulants as cognitive enhancers: the prefrontal cortex, catecholamines, and attention-deficit/hyperactivity disorder (Review). Biological Psychiatry 69, 101–111.

Bradley, C., 1937. The behavior of children receiving benzedrine. American Journal of Pyschiatry 94, 577–585.

Bush, G., 2011. Cingulate, frontal, and parietal cortical dysfunction in attention-deficit/hyperactivity disorder (Review). Biological Psychiatry 69, 1160–1167.

Bush, G., Valera, E.M., Seidman, L.J., 2005. Functional neuroimaging of attention-deficit/hyperactivity disorder: a review and suggested future directions (Review, 101 refs). Biological Psychiatry 57, 1273–1284.

Casey, B.J., Castellanos, F.X., Giedd, J.N., et al., 1997. Implication of right frontostriatal circuitry in response inhibition and attention-deficit/hyperactivity disorder. Journal of the American Academy of Child & Adolescent Psychiatry 36, 374–383.

Casey, B.J., Epstein, J.N., Buhle, J., et al., 2007. Frontostriatal connectivity and its role in cognitive control in parent-child dyads with ADHD. American Journal of Psychiatry 164, 1729–1736.

Castellanos, F.X., Margulies, D.S., Kelly, C., et al., 2008. Cingulate–precuneus interactions: a new locus of dysfunction in adult attention-deficit/hyperactivity disorder. Biological Psychiatry 63, 332–337.

Del Campo, N., Chamberlain, S.R., Sahakian, B.J., Robbins, T.W., 2011. The roles of dopamine and noradrenaline in the pathophysiology and treatment of attention-deficit/hyperactivity disorder (Review). Biological Psychiatry 69, 145–157.

Davenport, N.D., Karatekin, C., White, T., Lim, K.O., 2010. Differential fractional anisotropy abnormalities in adolescents with ADHD or schizophrenia. Psychiatry Research 181, 193–198.

Dickstein, S.G., Bannon, K., Castellanos, F.X., Milham, M.P., 2006. The neural correlates of attention deficit hyperactivity disorder: an ALE meta-analysis. Journal of Child Psychology and Psychiatry and Allied Disciplines 47, 1051–1062.

Durston, S., Van Belle, J., De Zeeuw, P., 2011. Differentiating frontostriatal and frontocerebellar circuits in attention-deficit/hyperactivity disorder (Review). Biological Psychiatry 69, 1178–1184.

Ebaugh, F.G., 1923. Neuropsychiatric sequelae of acute epidemic encephalitis in children. American Journal of Diseases of Children 25, 89–97.

Feldman, H.M., Yeatman, J.D., Lee, E.S., Barde, L.H., Gaman-Bean, S., 2010. Diffusion tensor imaging: a review for pediatric researchers and clinicians (Review, 74 refs). Journal of Developmental and Behavioral Pediatrics 31, 346–356.

Froehlich, T.E., Anixt, J.S., Loe, I.M., Chirdkiatgumchai, V., Kuan, L., Gilman, R.C., 2011a. Update on environmental risk factors for attention-deficit/hyperactivity disorder (Review). Current Psychiatry Reports 13, 333–344.

Froehlich, T.E., Epstein, J.N., Nick, T.G., et al., 2011b. Pharmacogenetic predictors of methylphenidate dose-response in attention-deficit/hyperactivity disorder. Journal of the American Academy of Child & Adolescent Psychiatry 50, 1129–1139.

Froehlich, T.E., Mcgough, J.J., Stein, M.A., 2010. Progress and promise of attention-deficit hyperactivity disorder pharmacogenetics (Review, 107 refs). CNS Drugs 24, 99–117.

Hoffman, H., 1848. Der Struwwelpeter. Imsel Verlag, Leipzig.

Hohman, L.B., 1922. Post-encephalitic behavior disorder in children. Bulletin of the Johns Hopkins Hospital 33, 372–375.

Johnson, S., Marlow, N., 2011. Preterm birth and childhood psychiatric disorders. Pediatric Research 69, 11R–18R.

Konrad, K., Eickhoff, S.B., 2010. Is the ADHD brain wired differently? A review on structural and functional connectivity in attention deficit hyperactivity disorder. Human Brain Mapping 31, 904–916.

Liston, C., Cohen, M.M., Teslovich, T., Levenson, D., Casey, B.J., 2011. Atypical prefrontal connectivity in attention-deficit/hyperactivity disorder: pathway to disease or pathological end point? (Review). Biological Psychiatry 69, 1168–1177.

Liston, C., Watts, R., Tottenham, N., Davidson, M.C., Niogi, S., Ulug, A.M., Casey, B.J., 2006. Frontostriatal microstructure modulates efficient recruitment of cognitive control. Cerebral Cortex 16, 553–560.

Loe, I.M., Lee, E.S., Feldman, H.M., 2013. Attention and internalizing behaviors in relation to white matter in children born preterm. Journal of Developmental and Behavioral Pediatrics 34, 156–164.

Lubke, G., Hudziak, J., Derks, E., Van Bijsterveldt, T., Boomsma, D., 2009. Maternal ratings of attention problems in ADHD: evidence for the existence of a continuum. Journal of the American Academy of Child & Adolescent Psychiatry 48, 1085–1093.

Mahone, E.M., Wodka, E.L., 2008. The neurobiological profile of girls with ADHD. Developmental Disabilities Research Reviews 14, 276–284.

Nagel, B.J., Bathula, D., Herting, M., Schmitt, C., Kroenke, C.D., Fair, D., Nigg, J.T., 2011. Altered white matter microstructure in children with attention-deficit/hyperactivity disorder. Journal of the American Academy of Child & Adolescent Psychiatry 50, 283–292.

Nagy, Z., Westerberg, H., Skare, S., et al., 2003. Preterm children have disturbances of white matter at 11 years of age as shown by diffusion tensor imaging. Pediatric Research 54, 672–679.

Nigg, J.T., 2005. Neuropsychologic theory and findings in attention-deficit/hyperactivity disorder: the state of the field and salient challenges for the coming decade (Review, 122 refs). Biological Psychiatry 57, 1424–1435.

Peterson, D.J., Ryan, M., Rimrodt, S.L., Cutting, L.E., Denckla, M.B., Kaufmann, W.E., Mahone, E.M., 2011. Increased regional fractional anisotropy in highly screened attention-deficit hyperactivity disorder (ADHD). Journal of Child Neurology 26, 1296–1302.

Seidman, L.J., Valera, E.M., Makris, N., 2005. Structural brain imaging of attention-deficit/hyperactivity disorder (Review, 94 refs). Biological Psychiatry 57, 1263–1272.

Shaw, P., Gilliam, M., Liverpool, M., et al., 2011. Cortical development in typically developing children with symptoms of hyperactivity and impulsivity: support for a dimensional view of attention deficit hyperactivity disorder. American Journal of Psychiatry 168, 143–151.

Shaw, P., Lerch, J., Greenstein, D., Sharp, W., Clasen, L., Evans, A., Giedd, J., Castellanos, F.X., Rapoport, J., 2006. Longitudinal mapping of cortical thickness and clinical outcome in children and adolescents with attention-deficit/hyperactivity disorder. Archives of General Psychiatry 63, 540–549.

Skranes, J., Vangberg, T.R., Kulseng, S., et al., 2007. Clinical findings and white matter abnormalities seen on diffusion tensor imaging in adolescents with very low birth weight. Brain 130, 654–666.

Still, G.F., 1902. The Coulstonian lectures on some abnormal physical conditions in children. Lancet 1, 1008–1012.

Valera, E.M., Faraone, S.V., Murray, K.E., Seidman, L.J., 2007. Meta-analysis of structural imaging findings in attention-deficit/hyperactivity disorder. Biological Psychiatry 61, 1361–1369.

Zametkin, A.J., Nordahl, T.E., Gross, M., et al., 1990. Cerebral glucose metabolism in adults with hyperactivity of childhood onset. New England Journal of Medicine 323, 1361–1366.

Attention, Neural Basis of

Manuela Sellitto and Giuseppe di Pellegrino, Centre for Studies and Research in Cognitive Neuroscience, University of Bologna, Cesena, Forlì-Cesena, Italy

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. di Pellegrino, volume 2, pp. 888–894, © 2001, Elsevier Ltd.

Abstract

Over the last three decades, cognitive neuroscience has contributed outstandingly to research on attention. Selective attention is a key mechanism the brain uses to face the enormous flow of stimuli coming from the environment, with the ultimate purpose of focusing on the relevant object, thus achieving goal-directed behavior. Classical studies in monkeys and recent research using novel techniques in humans and other animals have been integrated, thereby contributing to the development of neural and functional models of explanation of several mechanisms through which visual selective attention is deemed to operate.

Selective attention is a key mechanism the brain uses in order to make efficient object recognition and to focus on the relevant stimulus, with the ultimate purpose of achieving coordinated, goal-directed behavior. Indeed, the flow of inputs that is constantly available from the environment to our sense organs is far greater than the brain can handle and cope with, and cognitive psychology research has clearly established that our perceptual system is severely limited in its capacity to fully process and identify more than one object at a time. Moreover, our response system is limited inherently in its ability to perform concurrent actions on a large number of objects: we can only gaze at one object at a time to align it with the fovea of the retina, and reach at most to two objects. Selective attention allows the brain to select the most important aspects of the sensory scene and filter out those that are irrelevant or potentially interfering with the goal. Perceptual experience of the sensory world critically depends on attention. Selected objects can enter perceptual awareness, guide our choices and actions, and undergo memory storage, whereas ignored and irrelevant objects are often left at the margins of our consciousness.

For several decades, selective attention has been studied at the psychological level. More recently, behavioral research has been strengthened by a variety of neuroscience approaches as diverse as single-neuron recordings in monkeys, functional neuroimaging and event-related potentials in humans, lesion studies in neurological patients, and, lately neural-network models. This interdisciplinary and convergent effort is beginning to reveal how the brain activity is regulated through attention. Due to its essential role in building and gating object representations, this article will focus mainly on selective visual attention and review some recent experimental advances in understanding the neural underpinnings and mechanisms of this cognitive function.

Multiple Visual Pathways

The neural representations of visual objects have been demonstrated by contemporary neuroscience to be highly distributed. Visual inputs activate almost simultaneously a variety of cortical and subcortical regions within the brain. Among visually responsive structures are 30 or more separate cortical areas, which appear to be organized within two major parallel pathways, or streams (Ungerleider and Mishkin, 1982; Reynolds and Chelazzi, 2004; Connor et al., 2007). Each visual stream begins with the primary or striate visual cortex (area V1) and then involves multiple areas of the secondary and extrastriate visual cortices beyond V1. One of the pathways, the ventral stream, leads toward the inferior temporal (IT) cortex, and is crucial for object perception and recognition (the 'what' pathway). The other pathway, the dorsal stream, is directed into the posterior parietal cortex and is mostly concerned with spatial localization and action (the 'where' or 'how' pathway). Physiologically, the processing of visual information appears largely hierarchical within each cortical pathway, whereby object representations are synthesized with level of complexity. Specifically, the latency of visual responses, the size of the receptive field (RF), and the complexity of neuronal tuning properties increase steadily as one proceeds from one area to the next in the processing hierarchy. Anatomically, two types of connections between areas can be identified: (1) feed-forward connections that relay information from lower to higher visual areas in each pathway and (2) feedback connections that provide recurrent signals from higher to lower levels of processing. As described below, feedback, re-entrant connections from parietal and frontal areas, may form the anatomical basis of top-down attentional signals that modulate activity elsewhere in the visual system.

The Anatomical Locus of Selective Attention

Historically, the parietal lobe has been viewed as central in visual selection. Damage to this region commonly produces unilateral neglect, a deficit in perceiving and interacting with events on the contralesional side of space, which has often been explained as a difficulty in directing attention to that side. Patients with left neglect, for example, fail to complete the left half drawings, or to read the left side of words or passages of text. The attentional nature of neglect is well shown by the observation that the deficit is greatly exaggerated by competing

inputs on the intact, ipsilesional side ('extinction'). In fact, recent evidence based on large series studies indicates that neglect and extinction may follow a wide variety of unilateral lesions, both cortical and subcortical. This indicates a multiplicity of competitive systems modulated by attention and behavioral relevance, rather than a single anatomical level at which attention affects processing.

Data from lesion studies in monkeys are consistent with this conclusion. Indeed, different methodologies, such as single-unit recording, event-related potential (ERPs), functional imaging, and lesion studies provide converging evidence that attention can affect neuronal activity in most of the extrastriate visual areas where such modulation has been investigated. Paradoxically, many electrophysiological reports suggested that the processing of attended and unattended items does not differ in primary visual cortex. Monkey physiology in the past yielded conflicting results in this regard. In some studies, approximately 30% of V1 neurons revealed some form of attentional modulation. However, in a pioneering study (Roelfsema et al., 1998), monkeys were trained on an ingenious task in which animals had to trace one of two curves – the attended curve – while maintaining fixation. The firing rate of V1 neurons was found to be significantly higher when the receptive field was on the curve that needed to be traced, than on the distracting curve. The long latency of the enhancement (about 200 ms) suggested that this effect might be mediated by feedback processing from higher-order areas in the visual hierarchy.

Functional magnetic resonance imaging (fMRI) studies in humans have also shown that spatial attention is indeed capable of modulating neuronal activity in V1. One study compared the magnetic resonance signals in central and peripheral parts of V1 while subjects were either attending to a rapidly presented series of letters in the center of the visual field or performing a direction-of-rotation discrimination task on a peripheral rotating disk (Somers et al., 1999). In striking contrast with previous imaging studies, substantial increases in activity were found in V1 regions corresponding to the attended stimulus and decreases in regions representing the unattended stimulus. Together, these studies demonstrated that, depending on the visual task, attentional modulation of responses does indeed occur in the primary visual cortex. For the region of V1 representing the attended part of the visual field, neuronal responsiveness increases; conversely, when attention is directed away from a particular location there appears to be a consequent decrease in responsiveness in the retinotopically corresponding V1 region.

How can these recent findings be reconciled with previous failure to find attentional modulation in V1? One explanation is that, in the earlier studies, the tasks were not sufficiently demanding. In the new studies, stimulus parameters were chosen so that tasks were difficult and only 70–90% correct responses were achieved. Alternatively, it is possible that some techniques, like ERP recordings, may reflect only initial, feed-forward processing in V1. Apparently, the early part of the visual response of V1 neurons is unaffected by attention, as single-unit studies have shown. In contrast, the relatively poor time resolution of fMRI measurements would capture both early feed-forward processing and later top-down attentional influences from higher visual areas.

How Does Visual Selective Attention Work?

According to a traditional view, visual attention has been viewed as a mental spotlight or zoom lens that illuminates stimuli in selected regions of space, thus easing their access to higher processing centers (Posner, 1980). Visual attention would be inherently spatial in nature, so that objects are found by rapidly shifting the attentional spotlight from one location to another. The canonical view also poses that attention is mainly controlled by a unified, topographical 'master' map located in the parietal cortex and associated spatial processing structures. However, several recent findings suggest that any effect of attention is best explained in terms of competitive interactions among the neural representations of objects present in the visual field, as outlined by the Biased Competition Model of visual attention (Desimone and Duncan, 1995). According to this model, (1) multiple inputs falling within the same RF compete for controlling the neuron's firing rate at all levels of the visual system (e.g., Zoccolan et al., 2005), and (2) different mechanisms can tip the competitive balance in favor of one stimulus' representation over the others, thus causing the neuron's firing to be determined by the selected object (e.g., Reynolds and Desimone, 2003).

At the behavioral level, competition is reflected in the loss of accuracy or slowing in reaction time that result when subjects are required to identify multiple stimuli at once, rather than a single object. At the physiological level, single-unit recording studies in monkeys have shown that responses to two stimuli that fall inside a single RF (or its surrounding) of visual cells are smaller than responses to either stimulus in isolation (Reynolds et al., 1999). This indicates that multiple stimuli in the visual field, effective or ineffective for that cell, are not processed separately from each other but rather they interact in an inhibitory, suppressive fashion. In agreement with this result, an fMRI study in humans (Kastner et al., 1998) has shown that neural responses from visual areas are lower when stimuli are presented simultaneously than when identical stimuli are presented one after the other in the same locations. The reduced activation in the simultaneous condition may reflect sensory suppression among multiple concurrent visual stimuli ('repetition suppression'; Kanwisher and Wojciulik, 2000; Kastner and Ungerleider, 2000; Vuilleumier et al., 2005).

Competitive interference between stimuli has been especially observed in ventral stream areas. However, there is also evidence for sensory competition in structures associated with the dorsal stream, for example, in the frontal eye field (FEF) and in areas MT and MST. Here, cells respond best to stimuli motion, that is, to particular 'preferred' direction across the RF, but respond weakly or not at all to stimuli moving in the opposite or 'null' direction. Treue and Maunsell (1996) found that the response to an otherwise optimal stimulus (i.e., moving in the cell's preferred direction), placed within the RF of an MT neuron, was critically reduced by the addition of a second, nonpreferred stimulus at a different location in the same RF. Recently, computational neuronal models have been developed to predict different effect of neuronal firing. One of these, the Normalization Model of Attention (Reynolds and Heeger, 2009), posits that the attention multiplies the excitatory drive for the neuron in a selective manner, prior to

normalization (i.e., its division by the sum total drive across the neuronal population).

Bottom-Up and Top-Down Biases

Competitive interactions between neuronal populations encoding different visual objects can be modulated and solved by virtue of bottom-up and top-down biases, such that neurons encoding one object win out over the competition and that object receives further processing in the brain. Bottom-up factors refer to largely automatic and unconscious processes that operate early in vision and presumably contribute to perceptual 'pop-out' phenomenon – the quick and effortless detection of a unique item (for instance, based on luminance contrast) on a homogeneous background, at no increasing cost as a function of the number of distractors (Nothdurft, 2000). A stimulus ineffective, due to its features, in driving a visual response from a given neuron, may nonetheless be effective in determining the response, due to its salience. Bottom-up biases are thought to be mediated by hard-wired neural mechanisms, such as the center-surround structure of RFs found in many areas of the visual system. However, competition among visual neurons can also be biased by top-down cognitive influences, in which objects that meet the requirements of the current task are voluntarily selected for focal attention. For instance, top-down biases allow you on one occasion to search efficiently for your mug on a desk cluttered with many objects, while on another occasion you might search the same complex visual scene for your pencil. Top-down biases, as the name implies, are imposed on low-level sensory processes by higher-level control mechanisms, and are probably mediated by extensive feedback connections from higher areas (parietal and frontal) to low-level sensory areas. It is to these processes and their contribution to resolve competition between stimuli that the article will now turn.

What Does Attention Select?

In contrast to the spotlight model of visual attention, a central tenet of the Biased Competition Model is that the feedback bias from higher areas is not purely spatial. Depending on the context, different visual features – spatial location, color, shape, or motion – can be used to direct attention and assign limited processing resources to a particular object. Some psychologists have arrived at the notion of attentional template, a flexible, short-term description, formed to suit the task at hand, that can be used to give a competitive advantage to visual inputs matching that description. The attentional template can specify any property of the relevant object. Spatial location is, therefore, only one among many features that can contribute to the attentional template. For example, consider a subject that is asked to search for and report numbers colored in red and to disregard numbers in other colors. Following this instruction, neurons selective for red inputs are primed or preactivated in visual areas coding colors, such that a stimulus that possesses the behaviorally relevant color is processed preferentially over others. Information specifying the target of selection, be it spatial- or feature-based, is represented online in frontal networks while making the task ('working-memory'), and affects the activity of V4 neurons in a salience-based manner, even if with distinguishable time courses between spatial and visual processing (Hayden and Gallant, 2005).

Space-Based Selection

In the past, much research on the control of visual attention has focused on selection in space. In most behavioral studies, subjects covertly (i.e., without executing eye movement) direct their attention onto stimuli shown at a specific spatial location in the visual field, whereas distracting information is present at the other locations. Sensitivity has been shown to increase at attended versus unattended locations in the visual field, enhancing performance in terms of reduced reaction times and greater accuracy, and reduced at the distractor locations, suggesting that competition for focal capacity is biased toward objects at the attended region of space. A host of single-cell recording works in monkeys have reported that attention can increase the neural response to a visual stimulus when it appears at an attended location compared with when the animal attends to a different stimulus location. This selective enhancement effect, which increases with task difficulty and with the amount of visual clutter outside of the RF, has been demonstrated in numerous visual structures. These include primary visual cortex, areas V2 and V4, but especially dorsal stream areas implicated in visuomotor control, such as the lateral intraparietal area, the FEF, and the dorsolateral prefrontal cortex (Colby and Goldberg, 1998). Moreover, spatially directed attention also affects neuronal firing at subcortical level. For instance, enhanced visual responses were found in the lateral geniculate nucleus of the thalamus (McAlonan et al., 2008). In humans, brain imaging and ERP studies have reported similar findings. In these experiments, identical stimuli were presented on either side of fixation, while normal subjects focused attention to the left or right visual field. Results showed enhanced activity in extrastriate visual areas, generally lateralized to the hemisphere contralateral to the attended stimulus, in agreement with the fact that many visual areas have a predominantly crossed representation of the visual field (Kastner and Ungerleider, 2000). Overall, these findings suggest that top-down biases operate by increasing the input strength for stimuli at the attended locations. Most notably, however, these studies generally failed to find any effect of spatial attention on activity in human V1 (see above), and the above described effect of response enhancement is relatively large with stimuli of low contrast, whereas this tends to decrease with increasing contrast of the stimulus.

An alternative, not exclusive, possibility, however, is that spatial attention filters out distracting information when multiple stimuli appear simultaneously within the individual neuron's RF. Studies in extrastriate ventral areas V2 and V4, but also in dorsal areas MT and MST, have shown that, when monkeys attend to a stimulus at one location within the RF, competitive interference from stimuli at ignored locations within the RF is greatly attenuated or suppressed (Moran and Desimone, 1985). Apparently, the cell's RF 'shrinks' around the attended location, with unattended stimuli being filtered out. In addition, the effects of attention are much weaker when either the target or the distractor is placed outside the RF. Indeed, in such a configuration stimuli are no longer competing for the cell's response, thus reducing the need of

attentional biases. The suppressive mechanism of attention may explain why we have little awareness of unattended stimuli, and is highly consistent with the idea that spatial attention biases the sensory interactions among neurons, causing them to respond primarily to the attended stimulus.

In accordance with monkey neurophysiology, one fMRI study in humans showed that directing attention to one object in a cluttered visual scene reduces the sensory interference exerted by distracting stimuli at nearby locations (Kastner et al., 1998). Although attention can suppress irrelevant stimuli, it does not imply that it always does. According to one theory of attention, the extent to which spatial attention inhibits irrelevant information depends on the processing load imposed by the primary task, with stronger inhibition occurring when processing load is high. Consistent with this, an fMRI study by Rees et al. (1997) showed that neural responses to irrelevant moving stimuli, as registered in the motion-processing visual areas MT and MST, are much weaker when the primary task is difficult than when it is easy.

Nonspatial Feature-Based Selection

Although spatially directed attention has proven to be an excellent model system for investigating attentional mechanism, objects can be also selected by attention based on other, nonspatial, properties. For example, color can be used efficiently to control attention, particularly when there is a large hue difference between target and distractors. Neurophysiological studies in monkeys have considerably extended our understanding of feature-based processes for directing visual attention. In a single-neuron study in IT cortex, monkeys were instructed to perform a simple visual search task (Chelazzi et al., 1993, 1998, 2001). On each trial, a central cue was presented, followed by a blank delay period. At the end of the delay, an array of stimuli was presented and the monkey was rewarded to make a saccadic eye movement toward the stimulus matching the cue (target). The location of the target was not known in advance and thus selection could not be based on spatial information. Just like attention to location, attention to an object (nonspatial) feature was found to inhibit neuronal responses to irrelevant stimuli. Neurons in IT cortex showed an initial response if the choice array contained their preferred stimulus, without regard of the target identity. When the preferred stimulus was the eye movement target, the initial response remained high. However, when the preferred stimulus was an irrelevant, distracting stimulus in the array, the initial response was rapidly inhibited, long in advance of the onset of the saccade. Interestingly, neurons selective for the cue showed higher sustained activity during the delay. It has been argued that this tonic activity during the interval represents the neural correlate of the attentional template, which can be used to bias competition in favor of cells representing the target. Other neurophysiological studies also reported attentional modulation based on object features. Interestingly, studies that have used a single stimulus within the cell receptive field found facilitation of attended stimuli as well as inhibition of unattended stimuli.

Functional neuroimaging studies provided additional insight into feature selection. In a classic positron emission tomography study (Corbetta et al., 1991), subjects compared two successive visual displays, with reference to the color, shape, or speed of motion. Depending on the relevant visual feature, different extrastriate areas were active: attention to color led to enhanced visual response in area MT, whereas attention to shape and motion induced response enhancement in area V4.

Similar results have been reported using ERPs and fMRI. Together, these studies argue that selective attention to a particular visual feature enhances neural activity in areas selectively specialized for that attribute. Attention selection can also occur from whole object representations. In a recent fMRI study (O'Craven et al., 1999), two stimuli, a face and a house, were superimposed transparently in the same location. In each trial, either the face or the house oscillated back and forth. Subjects' attention was directed in different conditions to the face, the house, the direction of motion, or the spatial location of the stationary stimulus. As expected, different brain regions were activated depending on whether the relevant feature was the face (fusiform face area), the house (parahippocampal place area), or the direction of motion (areas MT and MST). Importantly, however, results showed that attention to one feature of an object enhanced the representation of other features of the same object, even when these features were completely irrelevant for the task. This 'cross-feature attention' phenomenon strongly suggests that attention selects entire objects, rather than selecting simple visual attributes and it evidences how it is easier to divide attention between different features of a single object than between features belonging to different objects (Duncan, 1984). In the case of a single object, for instance during a Stroop test (Stroop, 1935), choices must be guided by a selection of the relevant feature and the minimization of irrelevant ones. In the case of otherwise identical stimuli, responses of V4 neurons to them are modulated depending on the component feature of the stimulus being currently attended (after being cued), which could have been a specific color or orientation (Mirabella et al., 2007). The area V4 may be important in the process of converting selected feature information into a signal available to guide behavioral responses, suggesting that neurons in this area encode the behavioral response associated to the selected feature in a perceptually categorical fashion, instead of being a purely sensory area (Mirabella et al., 2007).

The Neural Mechanisms of Selection

Do attentional signals affect visual processing by facilitation of attended stimuli or by inhibition of irrelevant distractors, or both? One view suggests that attention exerts its influence by inhibiting unwanted stimuli early in visual processing, and by enhancing relevant information later on. The assumption underlying this view is that neurons in early visual areas respond at close to maximal rates to visual stimuli even during passive viewing, and therefore there is little room for response enhancement in early stages. In contrast, in higher visual areas, cells become less stimulus driven and more sensitive to behavioral relevance. Overall, recent single-unit and neuroimaging data do not support such a suggestion, and argue instead for a close interplay between suppressive and enhancing mechanisms at all levels of the visual system.

For example, a recent fMRI study reported that attention to a particular spatial location, in addition to an enhanced response to the attended visual stimulus, results in a widespread suppression of activity at all other locations (Smith et al., 2000). However, ERP studies have shown that the 'costs' of spatial cueing are associated with suppression of unattended inputs at an early stage, whereas the 'benefits' of cueing are associated with enhancement of attended stimuli at a later stage (Hillyard and Anllo-Vento, 1998). Thus, there is at least some suggestive evidence that suppression and enhancement may entail qualitatively different mechanisms, which operate at different levels of the visual pathways (Motter, 1993; Carrasco et al., 2004; Carrasco, 2006).

Top-down attentional signals may also affect visual processing by increasing the spontaneous (baseline) level of neural activity in the absence of any visual stimulation. Single-cell recording studies have clearly shown a substantial elevation of baseline activity for neurons in area V2 and V4, when the monkey was instructed to attend covertly to a location within the receptive field, even before the stimulus was presented there (Luck et al., 1997). The shift in baseline activity during spatial attention closely resembles the sustained, tonic activation found in IT neurons during visual search experiments. This increase in spontaneous firing rate is consistent with the idea of top-down feedback signals priming neurons coding the attended location of space. In a similar vein, several fMRI studies in humans have shown that during the expectation period preceding the arrival of a stimulus at an attended location, baseline activity increased in numerous extrastriate visual areas. The effect is spatially selective, as it occurs only in subregions of visual areas with a representation of the attended location; thus, it cannot be due to nonspecific arousal. Remarkably, an increase in baseline activity has been found even in the primary visual cortex during a demanding threshold detection task, and its magnitude was highly predictive of the subject's behavioral performance (Ress et al., 2000).

The Control of Visual Attention

A central issue in the current debate on selective visual attention concerns the brain sources of the top-down attentional signals to extrastriate and striate visual areas. Studies of brain-damaged patients, single-neuron recordings in primates and more recently blood-flow neuroimaging methods have identified a network of interconnected cortical and subcortical areas that play a key role in the control of selective visual attention. These areas include the dorsolateral prefrontal and posterior parietal cortex, the anterior cingulate gyrus, the pulvinar nucleus of the thalamus, and the superior colliculus (see Petersen and Posner, 2012; for a recent review). It has been proposed that in this circuitry, the prefrontal cortex would initiate and maintain the selective sensory bias in working memory, whereas the parietal cortex would direct and shift attention to specific objects and locations in space. In turn, this control network would enhance or suppress stimulus representations in the extrastriate visual areas according to current concern. Parietal and prefrontal cortex have anatomical and functional properties ideal for providing attentional templates. Both regions are extensively and reciprocally interconnected with virtually all visual cortical areas, and are therefore in a position to influence visual processing at multiple levels of the visual system. Neurons in the prefrontal and parietal cortex are capable of encoding the behavioral relevance or saliency of a visual stimulus. Furthermore, they display a pattern of activity that is well suited to form and maintain online a representation of an object or of a location for as long as that representation is relevant to behavior. In the past years, several functional neuroimaging studies in humans have shown frontal and parietal activations during visual attention tasks (see Kastner and Ungerleider, 2000; for a recent review). However, only recently top-down control signals for allocating attention have been distinguished from the modulatory effects that these signals have on the processing of incoming visual stimuli. For example, Hopfinger et al. (2000) asked which brain regions were activated in response to cues instructing shift or attention, and which were activated by the subsequent visual target presentation. A number of different visual areas were active; the surprising finding was that there was relatively little overlap between the two sets of visual responses. Importantly, posterior parietal and prefrontal areas were selectively engaged by attention-directing cues, thus suggesting that these two structures play a fundamental role in top-down attentional control. It should be noted, however, that the frontoparietal network is not exclusively related to visual attention, but may coincide or overlap with cortical regions involved in oculomotor processing (the FEF). This overlap supports the hypothesis that the neuronal signals used to direct attention correspond to the premotor signals that drive the eyes. One way to disentangle attentional and oculomotor processes would be an attentional task that does not require eye movements. In one study, for instance, subjects performed rapid attention shifting between the shape and color of stimuli presented at fixation, so that the need for eye movements eliminated (Le et al., 1998). Nevertheless, significant activations were observed in superior parietal lobule and other occipital visual areas, as well as the lateral cerebellum. Thus, it seems that at least some activity in the parietal and frontal cortex reflects attentional signals that are independent of eye movement. The frontoparietal circuit may therefore play a very general role in visual attention, governing the access of competing visual representations to higher brain centers and ultimately to perceptual consciousness.

See also: Attention-Deficit and Hyperactivity Disorders, Neural Basis of; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Cognitive Neuroscience; Consciousness, Neural Basis of; Eye Movement, Control of: Oculomotor Control; Hemispatial Neglect, Neural Basis of; Motivation and Actions, Psychology of; Occipital Lobe (Visual Cortex): Functional Aspects; Visual Streams: Dorsal and Ventral.

Bibliography

Carrasco, M., 2006. Covert attention increases contrast sensitivity: psychophysical, neurophysiological and neuroimaging studies. Progress in Brain Research 154, 33–70.

Carrasco, M., Ling, S., Read, S., 2004. Attention alters appearance. Nature Neuroscience 7, 308–313.

Chelazzi, L., Duncan, J., Miller, E.K., Desimone, R., 1998. Responses of neurons in inferior temporal cortex during memory-guided visual search. Journal of Neurophysiology 80, 2918–2940.

Chelazzi, L., Miller, E.K., Duncan, J., Desimone, R., 1993. A neural basis for visual search in inferior temporal cortex. Nature 363, 345–347.

Chelazzi, L., Miller, E.K., Duncan, J., Desimone, R., 2001. Responses of neurons in macaque area V4 during memory-guided visual search. Cerebral Cortex 11, 761–772.

Colby, C.L., Goldberg, M.E., 1998. Space and attention in parietal cortex. Annual Review of Neuroscience 22, 319–349.

Connor, C.E., Brincat, S.L., Pasupathy, A., 2007. Transformation of shape information in the ventral pathway. Current Opinion in Neurobiology 17, 140–147.

Corbetta, M., Miezin, F.M., Dobmeyer, S., Shulman, G.L., Petersen, S.E., 1991. Attentional modulation of neural processing of shape, color and velocity in humans. Science 248, 1556–1559.

Desimone, R., Duncan, J., 1995. Neural mechanisms of selective visual attention. Annual Review of Neuroscience 18, 193–222.

Duncan, J., 1984. Selective attention and the organization of visual information. Journal of Experimental Psychology – General 113, 501–517.

Hayden, B.Y., Gallant, J.L., 2005. Time course of attention reveals different mechanisms for spatial and feature-based attention in area V4. Neuron 47, 637–643.

Hillyard, S.A., Anllo-Vento, L., 1998. Event-related brain potentials in the study of visual selective attention. Proceedings of the National Academy of Science USA 95, 781–787.

Hopfinger, J.B., Buonocore, M.H., Mangun, G.R., 2000. The neural mechanism of top-down attentional control. Nature Neuroscience 3, 284–291.

Kanwisher, N., Wojciulik, E., 2000. Visual attention: insights from brain imaging. Nature Reviews Neuroscience 1, 91–100.

Kastner, S., De Weerd, P., Desimone, R., Ungerleider, L.G., 1998. Mechanisms of directed attention in human extrastriate cortex as revealed by functional MRI. Science 282, 108–111.

Kastner, S., Ungerleider, L.G., 2000. Mechanisms of visual attention in human cortex. Annual Review of Neuroscience 23, 315–341.

Le, T.H., Pardo, J.V., Hu, X., 1998. 4 T-fMRI study of nonspatial shifting of selective attention: cerebellar and parietal contributions. Journal of Neurophysiology 79, 24–42.

Luck, S.J., Chelazzi, L., Hillyard, S.A., Desimone, R., 1997. Neural mechanism of spatial selective attention in areas V1, V2 and V4 of macaque visual cortex. Journal of Neurophysiology 77, 24–42.

McAlonan, K., Cavanaugh, J., Wurtz, R.H., 2008. Guarding the gateway to cortex with attention in visual thalamus. Nature 456, 391–394.

Mirabella, G., Bertini, G., Samengo, I., Kilavik, B.E., Frilli, D., Della Libera, C., Chelazzi, L., 2007. Neurons in area V4 of the macaque translate attended visual features into behaviorally relevant categories. Neuron 54, 303–318.

Moran, J., Desimone, R., 1985. Selective attention gates visual processing in extrastriate cortex. Science 229, 782–784.

Motter, B.C., 1993. Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. Journal of Neurophysiology 70, 909–919.

Nothdurft, H., 2000. Salience from feature contrast: additivity across dimensions. Vision Research 40, 1183–1201.

O'Craven, K.M., Downing, P., Kanwisher, N., 1999. fMRI evidence for objects as the units of attentional selection. Nature 401, 584–587.

Petersen, S.E., Posner, M.I., 2012. The attention system of the human brain: 20 years after. Annual Review of Neuroscience 35, 73–89.

Posner, M., 1980. Orienting of attention. Quarterly Journal of Experimental Psychology 32, 3–25.

Rees, G., Frith, C.D., Lavie, N., 1997. Modulating irrelevant motion perception by varying attentional load in an unrelated task. Science 278, 1616–1619.

Ress, D., Benjamin, T., Backus, B.T., Heeger, D.J., 2000. Activity in primary visual cortex predicts performance in a visual detection task. Nature Neuroscience 3, 940–945.

Reynolds, J.H., Chelazzi, L., 2004. Attentional modulation of visual processing. Annual Review of Neuroscience 27, 611–647.

Reynolds, J.H., Chelazzi, L., Desimone, R., 1999. Competitive mechanisms subserve attention in macaque areas V2 and V4. The Journal of Neuroscience 19, 1736–1753.

Reynolds, J.H., Desimone, R., 2003. Interacting roles of attention and visual salience in V4. Neuron 37, 853–863.

Reynolds, J.H., Heeger, D.J., 2009. The normalization model of attention. Neuron 61, 168–185.

Roelfsema, P.R., Lamme, V.A., Spekreijse, H., 1998. Object-based attention in the primary visual cortex of the macaque monkey. Nature 395, 376–381.

Smith, A.T., Singh, K.D., Greenlee, M.W., 2000. Attentional suppression of activity in the human visual cortex. NeuroReport 7, 271–277.

Somers, D.C., Dale, A.M., Seifert, A.E., Tootell, R.B.H., 1999. Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. Proceedings of the National Academy of Science USA 96, 1663–1668.

Stroop, J.R., 1935. Studies of interference in serial verbal reactions. Journal of Experimental Psychology 18, 643–662.

Treue, S., Maunsell, J.H.R., 1996. Attentional modulation of visual motion processing gain in macaque visual cortex. Nature 399, 575–579.

Ungerleider, L.G., Mishkin, M., 1982. Two cortical visual systems. In: Ingle, D.J., Goodale, M.A., Mansfield, R.J.W. (Eds.), Analysis of Visual Behavior. MIT Press, Cambridge, MA, pp. S49–S86.

Vuilleumier, P., Schwartz, S., Dolan, R.J., Driver, J., 2005. Selective attention modulates neural substrates of repetition priming and 'implicit' visual memory: suppressions and enhancements revealed by fMRI. Journal of Cognitive Neuroscience 17, 1245–1260.

Zoccolan, D., Cox, D.D., DiCarlo, J.J., 2005. Multiple object response normalization in monkey inferotemporal cortex. The Journal of Neuroscience 25, 8150–8164.

Attitude Formation and Change

William D Crano, Claremont Graduate University, Claremont, CA, USA
Antonis Gardikiotis, Aristotle University of Thessaloniki, Thessaloniki, Greece

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by S. Chaiken, volume 2, pp. 899–905, © 2001, Elsevier Ltd.

Abstract

This article presents a summary of developments in the study of attitude formation and attitude change, two defining features of social psychology from its inception as an empirical science. Research on conceptualizations of change is discussed, as are variables implicated as attitude change agents. The interconnection of implicitly and explicitly measured attitudes is discussed, along with models devised to account for their mutual interplay.

Study of attitude formation and change is a central preoccupation of social psychology, and has been so from the beginning days of the discipline (Prislin and Crano, 2012). Research on attitudes provides an aperture into the conceptual system, supplies insights on complexities of human social behavior, and suggests how actions are affected by beliefs, and how beliefs affect actions. Advances in measurement, necessary for the scientific study of any phenomenon, were driven in the early days of attitude research by Louis Thurstone (1931), whose pioneering efforts conveyed the concept of attitude from speculation to operation. His insights were advanced and refined by Likert (1932) and Osgood et al. (1957), among others. Thurstone's view of attitudes as fundamentally evaluative was carried forward by his intellectual successors, and the field has largely followed this lead, though current views also include cognition and affect in definitions of the construct (e.g., "An attitude represents an evaluative integration of cognitions and affects experienced in relation to an object" (Crano and Prislin, 2006: p. 347).

Beyond measurement, programmatic experimentation on attitudes and attitude change grew largely from the efforts of Carl Hovland and his research group at Yale, where much of the foundational research on persuasion was undertaken (Hovland et al., 1953). Hovland's work affirmed Allport's (1935: p. 395) earlier appraisal that identified attitudes as social psychology's 'most distinctive and indispensable concept.' Although other actors have made their entrance on social psychology's stage since that time, attitude formation and change maintains the lead role.

Two fundamentally divergent views of the construct currently vie for adherents: the traditional position follows from and extends Thurstone's original evaluative emphasis, in which attitudes are conceptualized as enduring and stable structures that are stored in memory and vary in strength, which determines the attitude's stability and likelihood of affecting behavior (Fazio, 2007; Petty and Brinõl, 2009). The alternative (constructionist) position also hews to the evaluative nature of attitudes, but holds that attitudes are not stored in memory, but rather are constructed judgments influenced by the context in which they are assessed (Schwarz, 2007). Intermediate positions (e.g., Crano and Prislin, 2006) can accommodate either the traditional or constructivist views, and may provide a means of integrating both camps.

Attitude Formation and Change

Partly in response to Johnson and Eagly's (1990) stout critique of the field's failure to differentiate attitude formation from change, investigators have attended more closely to the distinction between these two cognitive processes. The best research on attitude change focuses on beliefs that are relevant to the individuals under study, whereas investigations of attitude formation are concerned with how attitudes develop in the first place.

Attitude Formation

Early research on attitude formation, initiated when learning-based models were in vogue, tested the possibility that attitudes could be established via classical or operant conditioning procedures. Both approaches were consistent with Thurstone's assumption that attitudes were fundamentally affective. Using classical conditioning processes, Staats and Staats (1958) found that consistent association of positively valenced words (e.g., gift, sacred, happy) with a national name (e.g., Dutch, Swedish) resulted in significantly more positive ratings of the national name than when the same name was associated with negatively valenced words (e.g., bitter, ugly, failure). This result held after subjects who claimed awareness of the study's purpose were removed from the analysis, though this procedure's utility has been challenged (Kattner et al., 2012; Stahl et al., 2009). Operant conditioning, too, can affect expressions of attitude. Insko (1965) verbally reinforced (good) university students for agreeing with positive or negative statements about a campus issue. This operant procedure clearly influenced responses. Furthermore, the effects persisted when students' attitudes were assessed 1 week later in a different context, by different researchers with the critical attitude items buried in a longer 'Local issues questionnaire.' Awareness of reinforcement contingencies is not a strong rival explanation of the persistent effects Insko reported.

More recent research on *evaluative conditioning* (EC) mirrors the findings of the classical conditioning school, but its proponents point to different underlying processes to explain conditioning effects. In the usual EC study, a picture of a neutral, unknown human face is presented repeatedly with a liked or disliked face, resulting in the originally neutral face

taking on the evaluative tone of the valenced one. The mere association of evaluative cues is sufficient for attitude formation. This position has many adherents (e.g., Forgas, 2008; Walther and Langer, 2010), but it is not without critics who maintain that cognitions (beliefs) and their associated evaluations, not mere affect, are the foundation of attitudes (Fishbein and Middlestadt, 1995). The weight of evidence favors the 'mere affective association' position, but whether EC can affect strong and highly vested attitudes remains at issue.

Contrary to results in classical conditioning, some have found that EC effects are strongest when respondents are *unaware* of the contingency between conditioned and unconditioned stimuli (De Houwer et al., 2001). However, controversy surrounding this position remains (Gawronski and Walther, 2012; Kattner et al., 2012). This issue is important as it bears on the effects of implicit attitudes on actions and explicit beliefs (Greenwald et al., 1998), the focus of considerable interest today. What appear noncontroversial are the distinctions between classical and evaluative conditioning. The latter, for example, apparently is not dependent on temporal contiguity between the unconditioned and conditioned stimuli. Furthermore, linkages formed by evaluative conditioning seem considerably more resistant to extinction than classically conditioned associations (Walther and Langer, 2008). These differences suggest the more widespread use of evaluative conditioning in field settings, especially in applications involving attitude formation, that is, on issues on which no firmly established belief exists.

Attitude Change

Study of attitude change logically presupposes preexisting attitudes, which may vary in terms of strength and centrality. An essential task of the attitude change researcher is to understand the factors that render an individual more or less susceptible to persuasion. A standard assumption in almost all of the foundational theories is that people strive to attain and maintain valid beliefs. Consistent with this assumption is the expectation that the more self-relevant the attitude, the more difficult it will be to change. These assumptions have formed the basis of considerable research in traditional studies of attitude change.

Hovland's Base Model

The formative model of attitude change was developed by Hovland et al. (1953), who made a critical but often overlooked distinction between attitudes and opinions. Attitudes were thought to involve nonverbalizable, often unconscious implicit approach–avoidance tendencies; opinions, on the other hand, were the verbal answers that one produced overtly or covertly to questions regarding the evaluation of a person or object. The technology of Hovland's time precluded study of implicit cognitive–behavioral responses (attitudes), and so by his definition, 'attitude change' research actually focused on opinions (explicit evaluative judgments), expressed and assessed on standard attitude measures. (To maintain consistency with contemporary usage, we will use the terms interchangeably.) Hovland's stress on the implicit nature of attitudes foreshadowed current considerations of implicit and explicit attitudes, though his theory of communication and persuasion implies strong interrelatedness between them; that is, that explicit opinions and implicit attitudes were strongly associated, and that large discrepancies, resulting in unpleasant cognitive disequilibrium or imbalance, would create the conditions for change.

Message Design, Elaboration, and Counterargumentation

Hovland et al. (1953) assumed that if a persuasive message were noticed and relevant to the individual, it would be considered carefully (i.e., elaborated). This elaboration process, which involved careful consideration of the persuader's arguments, also was accompanied by a defensive response called counterargumentation, which protected established beliefs. Thus, developing methods to overcome counterargumentation was the central goal of persuasion in Hovland's scheme. To realize this objective, it was necessary that a persuasive message raises a question in the receiver's mind about the legitimacy or validity of an established attitude. Furthermore, to succeed, the communication had to provide an answer to the question that suggested the appropriate response. Hovland's model was developed when social psychology was strongly influenced by learning theory. Consistent with this then-popular approach, it was assumed that opinions were learned and that new learning was needed to overlay old knowledge in producing change. Attitude change was viewed as a function of message learning, and so factors thought to reinforce learning and acceptance of the persuader's position, thus resulting in 'attitude change,' were the fundamental variables of study. Many variables have been shown to influence the attitude change process in ways predicted by Hovland and associates, even though the underlying link between message learning and attitude change often was not evident, even in research that succeeded in changing attitudes. An abbreviated sampling of some of the most established persuasion variables is presented on the following pages.

Source

Features of the source of a persuasive communication have proven important to the success or failure of an attempt at influence. In Hovland's learning-based model, being in agreement (and identified) with a respected message source was reinforcing, thus prompting acceptance of the position advanced. Source credibility, as this variable is commonly termed, was defined as a function of the source's expertise – its capacity to provide valid information – and the source's trustworthiness – its willingness to provide honest information independent of self-interest. Sources of high levels of expertise and trustworthiness induce considerably more change than those that do not possess these features. It should be noted, however, that source credibility effects often prove ephemeral, an aspect of the persuasion equation noted even in the earliest studies of this variable (Hovland and Weiss, 1951).

Minority versus Majority Sources of Influence

An exciting extension of research on source credibility developed from Moscovici's (1980) seminal work on the influence of minority groups on the majority. After Moscovici,

considerable research was conducted to study the different effects of majority and minority message sources on persuasion. The literature indicates that both the minority and the majority can induce influence, but the cognitive processes underlying change in response to one the other of these groups are different and produce differences in the longevity of the altered beliefs (Crano, 2012; Crano and Alvaro, 1998). Attitude change usually occurs rapidly in response to a majority-sourced communication, but such change typically does not persist because it is based, theoretically, on threat and social pressure, and not on careful message elaboration.

Conversely, minorities rarely evoke immediate change; rather, if the minority is a part of the majority (i.e., if it is an in-group minority), and it is active – that is, willing to take the risk involved in confronting majority group members – then the majority may change attitudes on issues related, but not identical to the focus of the minority's communication. Furthermore, if this indirect change of majority members' attitudes is sufficiently large, a delayed change may occur on the attitude that was the original focus of the minority's persuasive communication. Crano and Alvaro (1998) and Martin and Hewstone (2008) have suggested theoretical models that account for this unusual but often observed pattern of results.

Fear Arousal

If positive reinforcement of a behavior results in a higher likelihood of the behavior's being repeated, then the reinforcement experienced by avoiding a negative behavioral outcome should induce avoidance of the behavior. This logic is the basis of considerable research on emotional or fear-arousing communications, which have an enormous appeal in applied research contexts (e.g., alcohol, tobacco, or drug prevention, lifestyle changes). Unfortunately, many such applications have been mounted by change agents who appear to misunderstand the principles involved in their use. It is critical to understand the way reinforcement operates when using fear arousing communications to promote attitude change. Thus, to threaten an individual with dire consequences for, say, using methamphetamines would not be expected to succeed if the target of the message is already concerned about the practice (e.g., is a user of the substance). Avoiding a negative outcome is reinforcing, but if the targeted behavior is ongoing, the fear-arousing message, not the proscribed behavior, is the likely candidate for avoidance (Witte, 1992). This type of communication can be effective if it is not overdone with threats the target finds unrealistic (Skenderian et al., 2008) or punishing (i.e., threatens negative outcomes for ongoing behavior), and the message is presented by a credible source who offers a way of effectively avoiding the threatened consequences of the behavior. Without satisfying these requirements, fear-arousing messages are likely to produce no change or, in the worst case scenarios, iatrogenic effects as a result of message rejection, possibly brought about by source derogation or psychological reactance (Brehm, 2000).

Distraction

Counterarguing draws upon scarce cognitive resources. Thus, it is not likely to be engaged in contexts involving issues that are unimportant to the receiver. This is one of many reasons why it is easier to persuade receivers to accept a message on topics on which they are not vested. Contexts that occupy cognitive resources also are expected to disrupt counterargumentation, and thus, increase the likelihood of successful persuasion. Petty et al. (1976) showed that distraction inhibited the dominant cognitive response to a persuasive communication. Thus, distraction enhanced attitude change when the dominant response was negative, as often is found with illogical and poorly argued communications. However, distraction interfered with the positive cognitive responses that typically follow from elaboration of logical and strongly argued communications. In this case, distraction inhibited attitude change. This research laid the groundwork for Petty and Cacioppo's (1986) elaboration likelihood model of attitude change (*see* Persuasion Theories).

Overload

Research on 'cognitive busyness' suggests that individuals working under cognitive load are less able to counterargue, and thus, are more susceptible to persuasive communications (Gilbert et al., 1988). This hypothesis can be augmented by considering message quality. As in research on distraction and persuasion, it is likely that cognitive overload enhances the relative persuasive power of weak messages by attenuating negative cognitive responses, but weakens the persuasiveness of strong and logical messages by moderating positive responses, which lead to message acceptance. If nothing else, research on distraction and cognitive load counsel against simple univariate explanations of complex human behavior, a theme that is replicated throughout the literature on persuasion.

Mood

Early research on the effects of mood on susceptibility to attitude change suggested that positive moods rendered receivers unlikely to elaborate persuasive communications carefully. Those in a positive mood state seemed susceptible to attitude change regardless of the quality of the persuasive communication. Message quality was thought to be a more important factor for those in neutral or negative mood states (Mackie and Worth, 1989). Findings based on Wegener et al. (1995) hedonic contingency hypothesis, however, suggests a somewhat different interpretation. In Wegener and associates' view, a happy mood state enhances message processing if the communication does not threaten to ruin the happy mood, or if the message is uplifting. However, if the message is depressing, happy people are less likely to process it. Happy people, it seems, are intent on remaining so, and thus are likely to seek out information that supports this goal. Unhappy people, on the other hand, seem motivated to alleviate their negative state, and thus will process information that might result in a happy ending. This possibility was supported in Hullett's (2005: p. 423) *meta*-analysis of research on the mood–persuasion association, which found that "participants' processing of messages seemed to be motivated toward attaining or maintaining positive moods."

This emphasis on the hedonic consequences of a persuasive communication anticipated research that indicated that

positively framed communications were more persuasive for participants in positive than negative mood states, especially when the communication was strong and logical. For those in negative mood states, the hedonistic consequences of message framing and quality were less important, as most messages would be likely to improve mood, and hence were not elaborated carefully (Wegener et al., 1995).

Misdirection

One would not expect an individual to counterargue a persuasive communication if it was addressed to another person. This expectation was confirmed by Walster and Festinger (1962), who found that an 'overheard' communication induced considerably more attitude change than an identical message targeted to the receiver. This misdirection effect, as it has come to be called, is based on the logic of counterargumentation Hovland et al. (1953) advanced, and represents an interesting technique used to circumvent resistance. It has been used successfully in many persuasion contexts. For instance, apparently misdirected messages were shown to be strongly persuasive in antidrug research involving young adolescents, a group that usually is resistant to such communications. The results of an experiment by Crano et al. (2007), for example, indicated that youth who had received a communication apparently directed toward their parents were significantly more affected by it than were adolescents to whom the communication was addressed directly. The communications themselves differed by only a few words in the introduction, which specified the apparent audience of the message, parents or their children.

Ambivalence

To this point, attitudes have been discussed as unidimensional like–dislike, evaluative constructs. Research on attitude ambivalence suggests that this conceptualization may be incomplete. Attitudes may possess positive and negative evaluations simultaneously, thus generating the experience of ambivalence (Newby-Clark et al., 2002). Distinguishing ambivalent from neutral attitudes has attracted considerable research attention (Conner and Armitage, 2008), suggesting the importance of the distinction. High levels of ambivalence are associated with weaker attitude–behavior relations, and less resistance to counterattitudinal communications (Conner and Armitage, 2008). The development of ambivalent attitudes, as distinct from neutral attitudes, is viewed by some as the result of contradictory values (Feather, 2004), which suggests why contradictory evaluations of the same object may be expressed by the same individual at different times, depending on the social context in which the evaluation is elicited. This is not to suggest that attitudinal ambivalence is a positive or sought-after state. It should be understood that the motivation to maintain consistency among cognitions is strong (Gawronski and Strack, 2012), and research has highlighted the ways in which explicit attitudes affect implicit ones, and vice versa, to establish the preferred and probably normative state of cognitive balance (Peters and Gawronski, 2011). As will be discussed, both poles of an ambivalent attitude may affect the other; the general dynamic of the cognitive system, however, appears to press for consistency among beliefs.

Implicit and Explicit Attitude Measurement and Change

Feather's (2004) description of ambivalent attitudes as reflecting the simultaneous possession of positive and negative attitudes toward the same object of judgment provides the basis for a deeper understanding of implicitly and explicitly measured attitudes. Implicit, often nonverbalizable attitudes are thought to develop slowly, the result of a consistent pairing of an unconditioned with a conditioned stimulus. They are slower to change than explicit attitudes, which are formed more rapidly as a function of "evaluative judgments that are based on syllogistic inferences" (Gawronski and Bodenhausen, 2006: p. 694). Research by Rydell et al. (2006) anticipated this view by showing that implicitly and explicitly measured attitudes were formed in different ways, and, importantly, were susceptible to different persuasive processes.

The standard models of attitude change (e.g., Petty and Cacioppo, 1986; Chaiken, 1980) were developed before the technology to assess attitudes implicitly was commonly available, and hence were focused on change of explicit attitudes. In the meantime, with technological advances, new models of attitude structure and change were developed to account for the interaction of implicit and explicit beliefs. Gawronski and Bodenhausen's (2006: p. 693) associative–propositional evaluation (APE) model, for example, "specifies causes of implicit and explicit attitude change and how changes in one kind of evaluation may or may not be associated with changes in the other." This is not to suggest that the 'standard' models cannot also account for the newly available – if not newly theorized – features of attitudes; the possible conflict of implicit and explicit attitudes can be seen as compatible with attitude models developed before implicit attitudes even could be assessed. Petty and Briñol (2009), for example, described one such integration in their metacognitive model (MCM) of attitude structure. In the MCM, evaluative associations can be formed in many ways, and to the extent that they are implicit or below awareness, are treated as valid. If the holder of an implicit attitude is unaware of its existence, it is not subject to dispute or direct falsification. Incongruity between attitudes measured with implicit or explicit measures can induce evaluative change on either the implicit or the explicit belief. These changes are predictable, and in part a function of the circumstances under which the incongruity was induced. A strong persuasive attack on an explicit attitude, for example, would be expected to affect the explicit attitude; if successful (i.e., if the explicit measure indicated attitude change in the advocated direction), movement on the implicit attitude might result as well, likely as a function of the preference for cognitive consistency noted earlier. An intriguing prediction from the MCM is that the more extreme the discrepancy between explicit and implicit attitudes, the greater will be the accommodation of one or the other when pressure for congruity motivates attitude change. Supportive evidence for this expectation is found in the literature on minority-induced indirect attitude change, where minorities often induce the majority to change on issues related, but not identical to the thrust of the persuasive attack (Moscovici, 1980). Research has shown that

extremity of indirect change is associated with likelihood of delayed direct change, as hypothesized in Crano and Alvaro's (1998) Context/Comparison Model of minority influence, Petty and Brinõl's (2009) MCM, and confirmed by Crano and Chen's (1998) empirical research.

Some Concluding Thoughts

Research on attitudes has witnessed enormous progress over the years, even though many of the issue and constructs of current interest have been contemplated and contested since the earliest days of research on persuasion. The march of progress in measurement, along with the accumulation of empirical substantive knowledge resulting from 60 years of intense research interest, has created a literature that is as rich as it is generative. The theoretical debates that have occurred over the years often ended in a draw, in which both sides of the issue owned part, or all of the truth, which they approached in somewhat different ways and from somewhat different points of view. A similar integration of current models also will occur, hopefully before too much more time has passed. The established models of persuasion have focused primarily on the formation and change of explicit attitudes, and often presuppose stored representations of person or object evaluations. These representations can have powerful effects on judgments and behavior. However, as shown, they do not preclude the possibility that implicit attitudes also can affect explicit beliefs and behaviors. The same can be said of the more recent models, which have emphasized the importance of implicitly identified attitudes. From the start, the more recent models of attitude formation and change had to come to grips with implicit associations, which often seemed to operate below the conscious awareness of their holders. These models, however, also had to create mechanisms allowing for the interaction of explicitly and implicitly measured attitudes. Models of persuasion that accommodate implicit and explicit attitudes, and that account for their mutual interplay in light of our clear preference for consistency among beliefs, represent the next peak in the topography of attitude research (Prislin and Crano, 2008).

See also: Attitude Measurement; Attitudes and Behavior; Cognitive Dissonance; Heuristics in Social Cognition; Implicit Social Cognition; Persuasion Theories; Social Cognition.

Bibliography

Allport, G.W., 1935. Attitudes. In: Murchison, C. (Ed.), Handbook of Social Psychology. Clark University Press, Worcester, MA, US, pp. 798–884.

Brehm, J.W., 2000. Reactance. In: Kazdin, A.E. (Ed.), Encyclopedia of Psychology, vol. 7. American Psychological Association, Washington, DC, pp. 10–12.

Chaiken, S., 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. Journal of Personality and Social Psychology 39, 752–766.

Conner, M., Armitage, C.J., 2008. Attitudinal ambivalence. In: Crano, W.D., Prislin, R. (Eds.), Attitudes and Attitude Change. Psychology Press, New York, pp. 261–286.

Crano, W.D., 2012. The Rules of Influence. St. Martin's Press, New York.

Crano, W.D., Alvaro, E.M., 1998. The context/comparison model of social influence: mechanisms, structure, and linkages that underlie indirect attitude change. In: Stroebe, W., Hewstone, M. (Eds.), European Review of Social Psychology, vol. 8. John Wiley and Sons, Hoboken, NJ, US, pp. 175–202.

Crano, W.D., Prislin, R., 2006. Attitudes and persuasion. Annual Review of Psychology 57, 345–374.

Crano, W.D., Siegel, J.T., Alvaro, E.M., Patel, N.M., 2007. Overcoming adolescents' resistance to anti-inhalant appeals. Psychology of Addictive Behaviors 21, 516–524.

Crano, W.D., Chen, X., 1998. The leniency contract and persistence of majority and minority influence. Journal of Personality and Social Psychology 74, 1437–1450.

De Houwer, J., Thomas, S., Baeyens, F., 2001. Association learning of likes and dislikes: a review of 25 years of research on human evaluative conditioning. Psychological Bulletin 127, 853–869.

Fazio, R.H., 2007. Attitudes as object-evaluation associations of varying strength. Social Cognition 25, 603–637.

Feather, N.T., 2004. Value correlates of ambivalent attitudes toward gender relations. Personality and Social Psychology Bulletin 30, 3–12.

Fishbein, M., Middlestadt, S., 1995. Noncognitive effects on attitude formation and change: fact or artifact? Journal of Consumer Psychology 4, 181–202.

Forgas, J.P., 2008. The role of affect in attitudes and attitude change. In: Crano, W.D., Prislin, R. (Eds.), Attitudes and Attitude Change. Psychology Press, New York, pp. 131–158.

Gawronski, B., Bodenhausen, G.V., 2006. Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. Psychological Bulletin 132, 692–731.

Gawronski, B., Bodenhausen, G.V., 2007. Unraveling the processes underlying evaluation: attitudes from the perspective of the APE model. Social Cognition 25, 687–717.

Gawronski, B., Strack, F., 2012. Cognitive Consistency: A Fundamental Principle in Social Cognition. Guilford Press, New York.

Gawronski, B., Walther, E., 2012. What do memory data tell us about the role of contingency awareness in evaluative conditioning? Journal of Experimental Social Psychology 48, 617–623.

Gilbert, D.T., Pelham, B.W., Krull, D.S., 1988. On cognitive busyness: when person perceivers meet persons perceived. Journal of Personality and Social Psychology 54, 733–740.

Greenwald, A.G., McGhee, D.E., Schwartz, J.L.K., 1998. Measuring individual differences in implicit cognition: the implicit association test. Journal of Personality and Social Psychology 74, 1464–1480.

Hovland, C.I., Janis, I.L., Kelley, H.H., 1953. Communication and Persuasion: Psychological Studies of Opinion Change. Yale University Press, New Haven, CT.

Hovland, C.I., Weiss, W., 1951. The influence of source credibility on communication effectiveness. Public Opinion Quarterly 15, 635–650.

Hullett, C.R., 2005. The impact of mood on persuasion: a meta-analysis. Communication Research 32, 423–442.

Insko, C.A., 1965. Verbal reinforcement of attitude. Journal of Personality and Social Psychology 2, 621–623.

Johnson, B.T., Eagly, A.H., 1990. Involvement and persuasion: types, traditions, and the evidence. Psychological Bulletin 107, 375–384.

Kattner, F., Ellermeier, W., Tavakoli, P., 2012. Both trace and delay conditioning of evaluative responses depend on contingency awareness. Learning and Motivation 43, 35–47.

Likert, R., 1932. A technique for the measurement of attitudes. Archives of Psychology 22, 1–55.

Mackie, D.M., Worth, L.T., 1989. Processing deficits and the mediation of positive affect in persuasion. Journal of Personality and Social Psychology 57, 27–40.

Martin, R., Hewstone, M., 2008. In: Zanna, M.P. (Ed.), Majority versus Minority Influence, Message Processing and Attitude Change: The Source-Context-Elaboration Model, vol. 49. Academic Press, New York, pp. 238–326.

Moscovici, S., 1980. Toward a theory of conversion behavior. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 13. Academic Press, New York, pp. 209–242.

Newby-Clark, I.R., McGregor, I., Zanna, M.P., 2002. Thinking and caring about cognitive inconsistency: when and for whom does attitudinal ambivalence feel uncomfortable? Journal of Personality and Social Psychology 82, 157–166.

Osgood, C.E., Suci, G., Tannenbaum, P., 1957. The Measurement of Meaning. University of Illinois Press, Urbana.

Peters, K.R., Gawronski, B., 2011. Mutual influences between the implicit and explicit self-concepts: the role of memory activation and motivated reasoning. Journal of Experimental Social Psychology 47, 436–442.

Petty, R.E., Briñol, P., 2009. Implicit ambivalence: a meta-cognitive approach. In: Petty, R.E., Fazio, R.H., Briñol, P. (Eds.), Attitudes: Insights from the New Implicit Measures. Psychology Press, New York, pp. 119–163.

Petty, R.E., Cacioppo, J.T., 1986. Communication and Persuasion: Central and Peripheral Routes to Attitude Change. Springer-Verlag, New York.

Petty, R.E., Krosnick, J.A. (Eds.), 1995. Attitude Strength: Antecedents and Consequences. Erlbaum, Hillsdale, NJ.

Petty, R.E., Wells, G.L., Brock, T.C., 1976. Distraction can enhance or reduce yielding to propaganda: thought disruption versus effort justification. Journal of Personality and Social Psychology 34, 874–884.

Prislin, R., Crano, W.D., 2008. Attitudes and attitude change: the fourth peak. In: Crano, W.D., Prislin, R. (Eds.), Attitudes and Attitude Change. Psychology Press, New York, pp. 3–15.

Prislin, R., Crano, W.D., 2012. History of social influence research. In: Kruglanski, A., Stroebe, W. (Eds.), Handbook of the History of Social Psychology. Psychology Press, New York, pp. 321–339.

Rydell, R.J., McConnell, A.R., Mackie, D.M., Strain, L.M., 2006. Of two minds: forming and changing valence-inconsistent implicit and explicit attitudes. Psychological Science 17, 954–958.

Schwarz, N., 2007. Attitude construction: evaluation in context. Social Cognition 25, 638–656.

Skenderian, J.J., Siegel, J.T., Crano, W.D., Alvaro, E.E., Lac, A., 2008. Expectancy change and adolescents' intentions to use marijuana. Psychology of Addictive Behaviors 22, 563–569.

Staats, A.W., Staats, C.K., 1958. Attitudes established by classical conditioning. Journal of Abnormal and Social Psychology 57, 37–40.

Stahl, C., Unkelbach, C., Corneille, O., 2009. On the respective contributions of awareness of unconditioned stimulus valence and unconditioned stimulus identity in attitude formation through evaluative conditioning. Journal of Personality and Social Psychology 97, 404–420.

Thurstone, L.L., 1931. The measurement of social attitudes. Journal of Abnormal and Social Psychology 26, 249–269.

Walster, E., Festinger, L., 1962. The effectiveness of 'overheard' persuasive communications. Journal of Abnormal and Social Psychology 65, 395–402.

Walther, E., Langer, T., 2008. Attitude formation and change through association: an evaluative conditioning account. In: Crano, W.D., Prislin, R. (Eds.), Attitudes and Attitude Change. Psychology Press, New York, pp. 87–109.

Walther, E., Langer, T., 2010. For whom Pavlov's bell tolls: processes underlying evaluative conditioning. In: Forgas, J.P., Cooper, J., Crano, W.D. (Eds.), The Psychology of Attitudes and Attitude Change. Psychology Press, New York, pp. 59–74.

Wegener, D.T., Petty, R.E., Smith, S.M., 1995. Positive mood can increase or decrease message scrutiny: the hedonic contingency view of mood and message processing. Journal of Personality and Social Psychology 69, 5–15.

Witte, K., 1992. Putting the fear back into fear appeals: the extended parallel process model. Communication Monographs 59, 329–349.

Attitude: History of Concept

Christian Fleck, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept 'attitude' became used widely in several social science disciplines to indicate a relatively stable evaluative stance of a person toward any object. Originally from the language of painting and sculpture, the term became used to designate an inner preparedness of some stability. Rivaling with competitors in the public space around the turn of the twentieth century such as instinct, imitation, habit, etc., attitude outperformed others primarily because Thurstone suggested a measurement technique which became within a very short period of time the standard technique to gauge opinions. A set of items which could be scaled between opposing extremes is then an attitude variable and the single items are attitudes.

'Attitude' entered the vocabulary of the social and behavioral sciences early in the twentieth century after the term had traveled for a while in other corners of language uses. A look at the prehistory of the concept reveals some of the advantages and problems it later experienced. Unlike many other social science concepts, the word 'attitude' existed in more than one Indo-European language in the same form before the social sciences captured it for their own use: *Attitudine* in Italian, *attitude* first in French and then in English. The word goes back to the medieval Latin *aptitude* and the classical *aptus* (*Oxford English Dictionary*; Fleming, 1967: 292). Before becoming a social science concept, 'attitude' was used in painting and in particular in sculpture to describe a bodily pose of an individual who presents him/herself: as in present day usage 'to strike an attitude.' Here it is less clear whether the term is directed more to the person's preparedness and execution or the position reached finally. By referring to something located inside a human and not easily observable, 'attitude' encompassed something which became attractive for social scientists.

According to Fleming (1967: 293–297) Charles Darwin made use of attitude for the first time conceptually. In *Expression of the Emotions in Man and Animals* (1872) he discusses the, for him surprising, fact that particular strong emotions are accompanied by gestures, bodily postures expressing the emotion where a vis-à-vis can understand the emotion of the other by reading the body gestures. Attitudes are, for Darwin in this book, strong motor expressions of a particular emotion. Since Darwin was read widely one can presume that his repeated use of the term attitude functioned as the way in which the term entered the language of the social sciences.

Fleming (1967) discusses a second way attitude could have become the concept it was from the 1920s onward. Back then psychologists worldwide looked at Germany as the most mature representation of the discipline, and since the majority of professional psychologists and other social and behavioral scientists of that time not only read German but often studied in Germany, it could have been the case that the translation of the German *Einstellung* as attitude was the way attitude became the concept we know today. In this case, it would have been the Würzburg School of *Denkpsychologie* – Oswald Külpe, Narziß Ach, and their pupils – who developed in their experiments their own conceptual language. Someone has to solve a task (*Aufgabe*) and the experimental psychologists interviewed the participants afterward about their inner reasoning, their thinking while executing the task, inviting the selected participants to execute introspection, then regarded as the bridge into the brain, so to speak. German *Denkpsychologen* used several composite words to describe what happened in the person before executing an action. One single word expression for this phenomenon inside an actor was *Einstellung* and when E.B. Titchener worked on a translation of works from the Würzburg School he had chosen attitude for it.

Furthermore we need to recognize that attitude had to fight against the prominent rivals occupying the intellectual space of the late nineteenth- and early twentieth-century discourses. Imitation following Gabriel Tarde and instinct with no similar prominent coiner populated the textbooks of social psychology and sociology of the early years of the twentieth century together with habit as more or less synonymous expressions. But also belief, opinion, wish, idea, mental and emotional 'states,' etc., populated the universe from which social scientists took out what pleased them to speak about unobservable moves in people's inner life. Even if one used 'attitude,' he was free to define it as it pleased the author. For instance, Franklin H. Giddings from Columbia University in his textbook *The Principles of Sociology: An Analysis of the Phenomena of Association and of Social Organization* (1896) describes the phenomenon of communication as a form of association: "The expression of conscious states by means of attitude, muscular movement, and utterance is a language common to animals and men" (Giddings, [1896] 1916: 108).

More broadly and with higher impact is what William I. Thomas from the University of Chicago contributed to the establishment of attitude as a social science concept. He not only studied in Germany but read widely according to his biographers-interpreters and he was teaching a course on 'social attitudes' from 1900 onward (Fleming, 1967: 322). Fleming identified casual usage of attitude in some of the early writings of Thomas but there is widespread consensus in the small literature on the conceptual history of attitude that *The Polish Peasant in America* (1918) was the birthplace of the sociological and social psychological concept of attitudes, initially mostly used with the prefix social. Both Thomas's teaching and the book functioned as media to make attitude, the concept it remains since then.

Thomas positioned attitude explanatorily in a prominent environment (and his coauthor Florian Znaniecki; but there is wide agreement that the famous Methodological Note – of 86 pages – was written by Thomas alone). He claims that there are two "fundamental practical problems … which have constituted the center of attention of reflective social practice in all times" (Thomas and Znaniecki, [1918]1927: 20). On the one hand the 'dependence of the individual upon social organization and culture' and on the other hand the 'dependence of social organization and culture on the individual' (Thomas and Znaniecki, [1918]1927). According to Thomas, the first problem could be reformulated as the question "How should we produce the desirable mental and moral characteristics in the individuals constituting the social group?" whereas the second problem "How should we produce the desirable type of social organization and culture?" In a footnote, he even adds that the two practical turns might strengthen each other in succinct steps of doing. For these two sides – the objective and the subjective – Thomas then proposed two concepts: value and attitude.

> By a social value we understand any datum having an empirical content accessible to the members of some social group and a meaning with regard to which it is or may be an object of activity.
>
> Thomas and Znaniecki, [1918]1927: 21

And he adds after this well-done definition a list of illustrative cases of values: "Thus, a foodstuff, an instrument, a coin, a piece of poetry, a university, a myth, a scientific theory, are social values" because individuals could act according to them transforming any natural object into a meaningful subject of activities.

Moving to the other concept Thomas defines attitude as "a process of individual consciousness which determines real or possible activity of the individual in the social world" (Thomas and Znaniecki, [1918]1927: 22). Less well developed than the other definition, the list of illustrations makes clear what Thomas had in mind: "Thus, hunger that compels the consumption of foodstuff; the workman's decision to use the tool," etc. The list of conceptual correspondences is revealing: Compel, decision, tendency, feeling, expression, sympathy and admiration, needs, fear, devotion, creating, understanding, and the ways of thinking – 'all these are attitudes' (Thomas and Znaniecki, [1918]1927).

Thomas goes even a step further and assigns social psychology to take care of attitudes because they are the subjective side of the phenomenon, entrusts sociology to concentrate on the objective axis, and investigates values. The Methodological Note did not remain the *locus classicus* of the early understanding of attitude because Thomas reworked his conceptual scheme continuously. He puts, figuratively speaking, below attitudes and values wishes – most probably his adoption of Freud's *Trieb* (alternatively translated into English as drive or instinct) – and proposed that the spectrum of wishes could be classified into his famous 'four whishes': desires for new experience, security, response, and recognition. These forcing powers then mold both the selection of values and the formation of attitudes in a particular 'situation' which is accompanied by the 'definition of the situation' either by an individual or offered by social groups or culture as such.

The major drive behind Thomas and others concerned with the concept attitude should be seen in the light of the scholarly discourses in the first decades of the twentieth century. The strong performance of behaviorism prompted authors to propose means to describe observable processes of behavior by abstaining from any introspection as a valid tool for research on humans' actions. What followed was the recession of the tradition of *Denkpsychologie* and the encapsulation of Freud's and others' depth psychology from academic discourses. The tradition to which Thomas contributed switches to a 'behavioralistic' language in presenting their insights.

Robert E. Park and Ernest W. Burgess's *Introduction to the Science of Sociology* ([1921]1942), the 'green bible' of the early Chicago School, contains under the chapter heading Social Forces several concepts aside from attitude (interests, social pressure, idea-forces, sentiments, wishes). The chapter's introduction, the selected exemplary text, and the recommendations for further readings show strong affinities toward psychoanalysis and committed to Thomas's four wishes but the explanation of what an attitude is remains highly elusive. Instead of defining attitude, the authors offer what attitudes are not:

> They are not instincts, nor appetites, nor habits, for these refer to specific tendencies to act that condition attitudes but do not define them. Attitudes are not the same as emotions or sentiments although attitudes always are emotionally toned and frequently supported by sentiments. Opinions are not attitudes. An opinion is rather a statement made to justify and make intelligible an existing attitude or bias. A wish is an inherited tendency or instinct which has been fixed by attention directed to objects, persons, or patterns of behavior, all of which then assume the character of values. An attitude is the tendency of the person to react positively or negatively to the total situation. Accordingly, attitudes may be defined as the mobilization of the will of the person.
>
> Park and Burgess, 1942: 438

There is some family resemblance to what later analytic philosophers did with ambiguous expressions by focusing on the use of them by different speakers, but Park and Burgess do not go explicitly into this direction, very much to their and attitude's disadvantage. Just as one illustration of what Park and Burgess did have in mind one could see from the list of topics suggested for written examinations: "Typical Attitudes: Familism, Individualism, 'Oppressed Nationality Psychosis,' Race Prejudice" or from the list of questions for discussion: "How far would you say that the attitude may be described as an organization of the wishes?" (Park and Burgess, 1942: 502–503).

Attitude entered center stage only when the statistical versatile psychologist L.L. Thurstone applied his measurement competencies to attitudes too. In 1928 he published a programmatic statement 'Attitudes Can Be Measured' which paved the way for the, until then, very vague concept. Similar to the concept intelligence, which became usable via measuring it, also attitude made its success along the road of operationalism. Thurstone borrowed from psychophysical experiments the idea of the 'just noticeable difference' to propose a continuum with opposite extremes and a set of items in between. He illustrated

his proposal by referring to an attitude variable militarism-pacifism, with a neutral zone in the middle. The 10 items running from one extreme to the other have been selected in a multilevel process. From a first long list of brief statements from anywhere (questionnaire, press, publications, letter, etc.) those items have been removed which did not allow to accept or reject them. The reduced list was then given to a group of several hundred 'judges or readers' asking them to order the items "in eleven piles ranging from opinions most strongly affirmative to those most strongly negative" (Thurstone, 1928: 545). Further refinements which became standard afterward generated a much shorter list of scaled attitude items where every single item stands for an attitude related to the topic with its opposite extremes.

Thurstone did not solve a single conceptual ambiguity but suggested a solution in the way to transform an undecidable conceptual problem into an empirical measurable analysis. The next decades demonstrated the attractiveness of this proposal (an early affirmative review of the field of attitude research is Gordon W. Allport (1935)). In particular, in psychology and social psychology measuring attitudes, forming and changing attitudes, and investigating the correlation between attitudes and behavior are still around. The techniques of scaling have been applied to other measurable dispositions and routinization sidestepped any debate about explanatory power of the concept attitude.

See also: Allport, Gordon W (1897–1967); Applied Social Research, History of; Attitude Formation and Change; Attitude Measurement; Attitudes and Behavior; Darwin, Charles Robert (1809–82); Definition of the situation: History of the Concept; Environmental Attitudes and Behavior: Measurement; Mentally Ill, Public Attitudes towards; Psychology, History of (Twentieth Century); Public Opinion: Social Attitudes; Semantic Differential; Sociology, History of; Thorndike, Edward Lee (1874–1949).

Bibliography

Allport, G.W., 1935. Attitudes. In: Murchison, C.A., Allee, W.C. (Eds.), A Handbook of Social Psychology. Clark University Press, Worcester, Mass, pp. 798–844.

Fleming, D., 1967. Attitude: the history of a concept. In: Perspectives in American History, vol. 1, pp. 287–365.

Giddings, F.H., [1896]1916. The Principles of Sociology: An Analysis of the Phenomena of Association and of Social Organization. Macmillan, New York.

Park, R.E., Burgess, E.W., [1921]1942. Introduction to the Science of Sociology. The University of Chicago Press, Chicago.

Thomas, W.I., Znaniecki, F., [1918]1927. The Polish Peasant in Europe and America. A.A. Knopf, New York.

Thurstone, L.L., 1928. Attitudes can be measured. American Journal of Sociology 33 (4), 529–554.

Attitude Measurement

Norbert Schwarz, USC Dornsife Mind & Society Center, University of Southern California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term *attitude* refers to a predisposition to evaluate some object in a favorable or unfavorable manner. This predisposition cannot be directly observed and is inferred from individuals' responses to the attitude object. Different measurement strategies draw on different types of responses. They include answers to direct questions; the observation of overt behaviors; and responses that may be outside of the individual's awareness, such as minute facial expressions and response time differences in the range of milliseconds. The advantages and disadvantages of different measurement strategies and their conceptual implications are discussed.

The term *attitude* refers to a hypothetical construct, namely a predisposition to evaluate some object in a favorable or unfavorable manner. This predisposition cannot be directly observed and is inferred from individuals' responses to the attitude object, which can run from overt behavior (such as approaching or avoiding the object) and explicit verbal statements to covert responses, which may be outside of the individual's awareness (such as minute facial expressions or small differences in the speed of a response). This article reviews diverse measurement procedures that draw on these different manifestations of attitudes (for a history, see Oskamp, 1991).

Explicit Self-Reports

Most attitude researchers rely on respondents' answers to questions, such as "Do you approve or disapprove of how President Obama is handling his job?" Direct questions are the most feasible procedure for assessing the attitudes of the population at large, e.g., in representative sample surveys. However, self-reports of attitudes are highly context dependent and minor changes in question wording, question format, and question order can profoundly alter the obtained results. This section addresses the underlying processes and their implications for the measurement and conceptualization of attitudes.

The Question Answering Process

To answer an attitude question, respondents need to (1) understand the meaning of the question; (2) retrieve relevant information from memory to; and (3) form an attitude judgment. In most cases, they also need to (4) map their judgment onto the response alternatives provided by the researcher; and they (5) may want to edit their response before they report it, due to reasons of social desirability and self-presentation. Respondents' performance at each of these steps is highly context dependent (for comprehensive treatments see Sudman et al., 1996; Tourangeau et al., 2000). The cognitive and communicative processes involved in these tasks are reviewed in *Questionnaires: Cognitive Approaches*, and this material applies directly to attitude questions. The following discussion is restricted to specific aspects of the judgment stage and the use of different response formats.

Evaluating Attitude Objects

Once respondents determined what the question refers to, they need to recall relevant information from memory. In some cases, they may have direct access to a previously formed relevant judgment that they can offer as an answer. In most cases, however, they will not find an appropriate answer readily stored in memory and will need to compute a judgment on the spot. To do so, they need to form a mental representation of the attitude object and of a standard, against which the object is evaluated. The resulting judgment depends on which information happens to come to mind at that point in time and on how this information is used.

As a large body of research demonstrated, individuals rarely retrieve all information that may bear on an attitude object but truncate the search process as soon as enough information has been retrieved to form a judgment with sufficient subjective certainty. Hence, the judgment depends on the first few pieces of information that come to mind. Whereas some information may always come to mind when the person thinks of this object (and is therefore called *chronically accessible*), other information may be only *temporarily accessible*, for example, because it has been brought to mind by preceding questions. Changes in what is temporarily accessible are at the heart of many context effects in attitude measurement, whereas chronically accessible information contributes some stability to attitude judgments.

How accessible information influences the attitude judgment depends on how it is used (Bless and Schwarz, 2010). Information that is included in the mental representation formed of the attitude object results in *assimilation effects*, i.e., the judgment becomes more positive (negative) when information with positive (negative) implications comes to mind. In contrast, information that is used in forming a mental representation of the standard against which the attitude object is evaluated results in *contrast effects*. In this case, the judgment becomes more negative (positive) when information with positive (negative) implications comes to mind and is used as a positive (negative) standard of comparison. Suppose, for example, that a preceding question brings a politician to mind who was involved in a scandal (say, Richard Nixon). When a subsequent question pertains to the trustworthiness of American politicians in general (a superordinate category), Richard Nixon is likely to be included in the representation formed of 'politicians' as a group, resulting

 http://dx.doi.org/10.1016/B978-0-08-097086-8.24006-3

in judgments of decreased trustworthiness (an *assimilation effect*). Suppose, however, that the question pertains to the trustworthiness of Barack Obama (a lateral category), rather than to politicians in general. In this case, Richard Nixon cannot be included in the representation formed of the attitude object 'Obama'; instead, he is likely to serve as a (rather negative) standard of comparison, relative to which Barack Obama will be evaluated as more trustworthy than would otherwise be the case (a *contrast effect*). As this example illustrates, the same context question can have opposite effects on superficially similar subsequent questions, resulting in many apparent inconsistencies. In the present case, a researcher might conclude that political scandals decrease trust in politicians when judgments of the group are used as the dependent variable, but increase trust in politicians when judgments of individual politicians are used as the dependent variable.

Whether a given piece of information is used in constructing a representation of the attitude object (resulting in assimilation effects), or of a standard of comparison (resulting in contrast effects), depends on a host of different variables. Bless and Schwarz (2010) review these variables and present a theoretical model that predicts the direction, size, and generalization of context effects in attitude measurement.

Some Attitude Scale Formats

Textbooks consistently advise researchers to use multiitem scales. However, the classic examples of multiitem attitude scales – namely, *Thurstone scales, Guttman scales*, and *Likert scales* (see Eagly and Chaiken 1993: Chapter 2) – are rarely used in practice, primarily because they require extensive, topic-specific development work. An exception is Osgood and colleagues' *Semantic Differential Scale* (see Eagly and Chaiken, 1993: Chapter 2), which is a ready-to-use scale that can be applied to any topic, making it considerably more popular. Respondents are asked to rate the attitude object (e.g., 'welfare') on a set of seven-point bipolar adjective scales. The adjectives used as endpoint labels reflect three general factors, namely evaluation (e.g., good–bad; pleasant–unpleasant), potency (e.g., strong–weak; small–large), and activity (e.g., active–passive; fast–slow). Of these factors, evaluation is considered the primary indicator of respondents' attitude toward the object, as reflected in the objects' (relatively global) connotative meaning.

In most studies respondents' attitudes are assessed by only one or two questions, presented either with multiple-choice response alternatives or a rating scale. When the question offers several distinct opinions in a *multiple-choice format*, it is important to ensure that the set of response alternatives offered covers the whole range of plausible positions. Any option omitted from the set is unlikely to be reported at all, even when respondents are offered a general 'other' response option, which they rarely endorse. Similarly, few respondents report not having an opinion on an issue when this option is not explicitly provided – yet, they may be happy to report so when 'do not know' is offered as an alternative (see Sudman et al., 1996: Chapter 3).

When *rating scales* are used, respondents are presented with a numerical scale with verbally labeled endpoints (e.g., -3 = 'strongly disagree,' $+3$ = 'strongly agree') and asked to check the number that best represents their opinion. Alternatively, each point of the rating scale may be labeled, a format that is common in telephone interviews. In general, the retest reliability of fully labeled scales is somewhat higher than that of partially labeled scales and retest reliability decreases as the number of scale points increases beyond seven, reflecting the difficulty of making many fine-grained distinctions. What is often overlooked is that the numeric values of the scale may influence respondents' substantive interpretation of the question. For example, when asked to report how highly they think of a politician along a scale from 'not so highly' to 'very highly,' respondents need to determine if 'not so highly' pertains to the absence of positive thoughts or to the presence of negative thoughts. When the numeric values run from 0 to 10, respondents interpret 0 = 'not so highly' as reflecting the absence of positive thoughts; but when the scale runs from -5 to $+5$, they interpret -5 = 'not so highly' as pertaining to the presence of negative thoughts. In general, scales that range from negative to positive numbers suggest that the researcher has a bipolar dimension in mind, implying that one endpoint label refers to the *opposite* of the other; in contrast, scales that present only positive numbers suggest a unipolar dimension, implying that different scale values pertain to different degrees of the *same* attribute. Such shifts in question meaning result in markedly different responses (see Schwarz, 1994 for what respondents learn from scales and other formal characteristics of questionnaires).

Implications

From a measurement perspective, the reviewed influences highlight the need to pay close attention to how respondents make sense of the question asked. Suitable methods can be employed at the questionnaire development stage; they are discussed in *Questionnaires: Cognitive Approaches*. From a theoretical perspective, the emergence of context effects poses important challenges for the traditional conceptualization of attitudes as enduring dispositions. One response to this challenge holds that context effects may only emerge when respondents have not thought much about an issue, but not when they have formed strong, or crystallized, attitudes. Empirically, this *attitude strength* hypothesis has received little support; although strong attitudes are more resistant to explicit persuasion attempts, they are not more resistant to contextual influences in the question answering process (for a review, see Krosnick and Abelson, 1992). A second response emphasized that explicit self-reports of attitudes are subject to self-presentation concerns – respondents may know their true and enduring attitude but may be unwilling to report it, tailoring the answer instead to what seems reasonable in the context of the interview. This concern motivated attempts to reduce socially desirable responding.

Reducing Socially Desirable Responding

Some procedures aim to reduce socially desirable responding by making it easier to report undesirable positions (for a review, see Tourangeau et al., 2000). They range from simple assurances of anonymity and confidentiality to complex randomized response techniques. In the latter case, respondents are presented with two different questions, an innocuous one and a socially sensitive one, and a draw a card that determines which one they are to answer. Given properly worded

response alternatives, the interviewer remains unaware to which question the answer pertains, thus ensuring the highest possible level of confidentiality. Other procedures create conditions that increase the perceived cost of socially desirable responding. For example, a 'bogus pipeline' technique involves convincing participants that the researcher can discern their true attitude independent of what they say, thus making lying an embarrassment. Empirically, these various techniques have been found to increase the frequency of socially undesirable answers. However, they do not affect the emergence of most context effects, suggesting that the most context effects are driven by processes of evaluative judgment rather than by concerns about one's public appearance.

Note, however, that all attempts to reduce socially desirable answers to explicit attitude questions rest on the assumption that respondents know their attitude and merely hesitate to report it to others. Wondering whether deception and self-presentation may also be directed toward the self, some researchers began to develop *implicit attitude measures* that attempt to assess attitudes of which people may not even be aware or may not want to admit to themselves.

Implicit Attitude Measures

Implicit measures are based on the assumption that attitudes exert a systematic influence on people's performance on a variety of tasks and that the size of this influence can serve as an index of the underlying attitude (for reviews of different measures and their theoretical rationales see the contributions in Wittenbrink and Schwarz, 2007). This measurement logic requires clear bridging rules that specify the theoretical and empirical relationship between the attitude and the measurement task. Not surprisingly, these bridging rules have varied widely over the history of attitude research. From the early use of projective tests to the current use of response latency measures, the history of indirect measures mirrors historical shifts in the conceptualization of attitudes and their underlying processes (see Vargas et al., 2007).

Response Time Measures

Response time measures of implicit attitudes take advantage of one of two regularities. First, preceding exposure to a stimulus facilitates subsequent responses to related stimuli. For example, after seeing the word 'doctor,' we are faster at recognizing or pronouncing the word 'nurse.' This *response facilitation effect* can be used to gauge which concepts are, or are not, closely associated in a person's mind. Second, a stimulus is responded to more slowly when it contains multiple features that give rise to competing responses. For example, it takes longer to name the color in which a word is printed when the meaning of the word and the print color diverge, e.g., when the word 'green' is printed in blue ink. This *response competition effect* can be used to gauge which responses are compatible or incompatible in a person's mind.

Response Facilitation Measures

One strategy involves the presentation of target words with evaluative meaning (e.g., lazy, smart) and measures how fast respondents can identify the word. Speed of identification can be assessed by asking respondents to pronounce the word as fast as possible or to decide quickly whether a letter string is a word or a nonword. Of interest is whether a preceding stimulus (called a 'prime') that represents the attitude object of interests affects the speed with which different target words can be identified. For example, a faster response to the word 'lazy' after exposure to the word 'black' than after exposure to the word 'white' is indicative of negative racial stereotypes about Blacks, as first observed by Gaertner and colleagues in the early 1980s. The response time patterns generated by such *concept priming procedures* can be brought to bear on three questions: First, does exposure to the attitude object activate associated beliefs? Any concept that is processed faster after exposure to the attitude object is associated with it, even though the person may not be aware of it. Second, is the activation of beliefs evaluatively biased? For example, are negative attributes identified more quickly than positive ones? Third, does exposure to the attitude object activate general evaluative associations (e.g., good, bad)?

A related strategy focuses solely on evaluative associations. In a typical experiment, participants are exposed to a prime (e.g., a Black or White face) and decide whether a subsequent target word (e.g., 'pleasant,' 'awful') means something positive (press the 'good' key) or something negative (press the 'bad' key). If the attitude prime is strongly associated with a positive evaluation, it speeds up the identification of positive words as 'good' and slows down the identification of negative words as 'bad.' Hence, this *evaluative priming procedure*, first used by Fazio and colleagues in the mid-1980s, gives insight into the extent to which the attitude object is associated with positive or negative evaluations. Wittenbrink (2007) reviews the development and application of these and related procedures and summarizes representative findings.

Response Competition Measures

Other measures take advantage of the interference that results from competing response tendencies. The best known of these procedures is the implicit associations test (IAT) developed by Greenwald and colleagues in the 1990s. It presents two discrimination tasks that are combined in specific ways across a sequence of five steps. To assess attitudes toward African-Americans and European Americans, for example, the first discrimination task may present names that are typical for the respective group and participants are asked to categorize each name as 'white' or 'black.' They do so by pressing a response key assigned to 'white' with their left hand or a response key assigned to 'black' with their right hand. Next, a second discrimination task presents words with pleasant (e.g., love) or unpleasant (e.g., poison) connotations, which participants classify as positive or negative by pressing the left or right response key. At the third step, these two tasks are superimposed and participants press the left key when either a white name or a pleasant word is shown, but the right key when either a black name or an unpleasant word is shown. As in the facilitation paradigms, this task is easier when evaluatively associated categories share the same response key, e.g., when white participants press the left key to categorize white names and pleasant words. Going beyond this assessment of response facilitation, the IAT involves two more steps.

At the fourth step, the assignment of keys to white and black names is reversed, so that participants who first used the left key for white names now use the left key for black names. Finally, the two discrimination tasks are again superimposed, resulting in an assignment of 'black' and 'pleasant' to the left response key and 'white' and 'unpleasant' to the right response key. Of key interest is the speed with which participants can perform the two superimposed discrimination tasks at step 3 and step 5. Do participants respond faster when a given response key pertains either to the pairing of white names + pleasant words or black names + unpleasant words (step 3) than when this pairing is reversed (step 5)? In the present example, a faster response at step 3 than at step 5 is thought to indicate that white names and positive evaluations and black names and negative evaluations are more strongly associated than the reverse pairings. Lane et al. (2007) review the development and application of the IAT, along with representative findings.

Low Tech Measures of Implicit Attitudes

Whereas response time procedures require a high degree of instrumentation and sophistication, other implicit measures are less demanding (for a review, see Vargas et al., 2007). Some take advantage of the observation that attitudes and expectations influence individuals' information processing in systematic ways. For example, people are more likely to spontaneously explain events that disconfirm rather than confirm their expectations; hence, the amount of explanatory activity can serve as an indirect measure of a person's expectations. For example, the number of explanations generated in response to stereotype-consistent versus stereotype-inconsistent behaviors provides an indirect measure of stereotyping. Similarly, people describe expected or stereotype-consistent behaviors in more abstract terms than unexpected or stereotype-inconsistent behaviors. The size of this bias can again be used as an indirect measure to gauge the underlying expectations.

Context Effects on Implicit Measures

To many researchers' surprise, the initial hope that implicit attitude measures may be relatively immune to contextual influences has not been supported (for a review, see Ferguson and Bargh, 2007). Instead, implicit and explicit measures of attitudes typically show the same pattern of context effects. For example, incidental exposure to pictures of liked or disliked members of racial groups shifts racial attitudes on both types of measures, as does exposure to preceding questions that influence which aspects of an attitude object come to mind. Note that these contextual influences cannot be traced to respondents' self-presentation concerns. Respondents lack insight into the rationale of implicit attitude measures and remain unaware of what they disclose about themselves, which renders these measures relatively immune to strategic responding. Instead, context effects on implicit measures reinforce what many researchers inferred from context effects on explicit measures: how people evaluate an attitude object depends on what is on their mind at the time of judgment.

Psychophysiological Measures

A different measurement strategy focuses on the physiological correlates of evaluative responses. Because physiological responses are assumed to be involuntary and hard to control nature, they have long been of interest to attitude researchers who doubted respondents' explicit self-reports. Early uses of physiological measures drew on the observation that strong affective reactions to an attitude object are associated with increased activation of the sympathetic nervous system. Increased sympathetic activation results in increased sweat glands activity, which can be measured by assessing the resistance of the skin to low-level electric currents, a procedure known as *electrodermal measurement*. However, electrodermal responses do not reflect the direction (favorable or unfavorable) of the evaluative response, which limits their usefulness.

More promising are attempts to assess changes in individuals' facial expression in response to an attitude object. Overt facial expressions (like smiling or frowning) may often be observed in response to attitude objects that elicit strong reactions. But these expressions may be intentionally concealed and many evaluative reactions may be too subtle to evoke overt expressive behaviors. Even subtle evaluative reactions are associated, however, with low-level activation of facial muscles that can be detected by *electromyography*. These muscle reactions reflect the direction (favorable vs unfavorable) as well as the intensity of the evaluative response. However, the obtained measures can be distorted by facial movements that are unrelated to the evaluative reaction.

Another approach involves the measurement of brain activity through *electroencephalography*, the assessment of small electric signals recorded from the scalp. This procedure, however, does not lend itself to a direct assessment of positive or negative responses. Instead, it capitalizes on the observation that unexpected stimuli evoke brain wave activity that differs from the activity evoked by expected stimuli. Hence, one may detect if a target object is evaluated positively or negatively by embedding its presentation in a long series of other objects with a known evaluation. The brain activity evoked by the target object will then indicate if its evaluation is consistent or inconsistent with the evaluation of the context objects.

Ito and Cacioppo (2007) review these and other measures, including recent developments in brain imaging techniques, which provide a promising avenue for future work in this area. Throughout, the implementation of psychophysiological measures requires sophisticated technology and high expertise and their analysis poses complex issues of data reduction.

Behavioral Observation

In principle, a person's attitude toward some object may be inferred from his or her behavior toward it. However, people's behaviors are influenced by many variables other than their attitudes. Hence, the attitude–behavior relationship is typically weak, unless other variables are taken into account. Hence, the mere observation of a behavior is a poor indicator of the person's attitude per se and behavioral observation is rarely used as a measurement strategy.

Implications and Future Directions

As this article of attitude measurement issues indicates, responses to attitude measures are highly context sensitive. While this observation is undisputed, its implications are controversial. Some researchers assume that context effects merely reflect some 'noise' that does not call the existence of enduring attitudes into question. From this perspective, attitude measures assess the person's true and enduring attitude plus some temporary contextual influence that gives rise to variation in the expression of an otherwise stable underlying disposition (see Eagly and Chaiken, 2005). Other researchers question the usefulness of the classic attitude concept. They assume that evaluation stands in the service of action. To guide action in an adaptive way, evaluation needs to be informed by past experience but highly sensitive to the specifics of the here and now, turning deplorable context dependency into laudable context sensitivity (see Schwarz, 2007). Independent of these different metatheoretical perspectives, it is increasingly apparent that different attitude measures can give rise to widely diverging conclusions, even when assessed at the same time. For example, subliminal exposure to 'cod liver oil' may facilitate your identification of 'disgusting' as a 'bad' word and may elicit a negative physiological response, yet you report that you think cod liver oil is good for you and you take it regularly. Which of these diverging responses should we consider an unbiased indicator of your 'true' attitude? Each measure reflects a different aspect of a complex phenomenon, rendering the designation of any single one as the only 'true' window on individuals' attitudes counterproductive. Understanding the complex interplay between cognitive, affective, physiological, and behavioral responses in context presents the most challenging issue for future research.

See also: Attitude Formation and Change; Cognitive Dissonance; Heuristics in Social Cognition; Implicit Social Cognition; Persuasion Theories; Social Cognition; Social Psychology; Values, Psychology of; Wisdom, Psychology of.

Bibliography

Bless, Herbert, Schwarz, Norbert, 2010. Mental construal and the emergence of assimilation and contrast effects: the inclusion/exclusion model. Advances in Experimental Social Psychology 42, 319–374.

Eagly, Alice H., Chaiken, Shelly, 1993. The Psychology of Attitudes. Harcourt, Brace, Jovanovich, Fort Worth.

Eagly, Alice H., Chaiken, Shelly, 2005. Attitude research in the 21st century: the current state of knowledge. In: Albarracin, Dolores, Johnson, B.T., Zanna, Mark P. (Eds.), The Handbook of Attitudes. Erlbaum, Mahwah, NJ, pp. 743–768.

Ferguson, Melissa J., Bargh, John A., 2007. Beyond the attitude object: automatic attitudes spring from object-centered-contexts. In: Wittenbrink, Bernd, Schwarz, Norbert (Eds.), Implicit Measures of Attitudes. Guilford, New York, pp. 216–246.

Ito, Tiffany A., Cacioppo, John T., 2007. Attitudes as a mental and neural state of readiness: using physiological measures to study implicit attitudes. In: Wittenbrink, Bernd, Schwarz, Norbert (Eds.), Implicit Measures of Attitudes. Guilford, New York, pp. 125–158.

Krosnick, Jon A., Abelson, Robert P., 1992. The case for measuring attitude strength. In: Tanur, Judith M. (Ed.), Questions about Questions. Russell-Sage, New York, pp. 177–203.

Lane, Kristin A., Banaji, Mahzarin R., Nosek, Brian A., Greenwald, Anthony G., 2007. Understanding and using the implicit association test: IV. What we know (so far) about the method. In: Wittenbrink, Bernd, Schwarz, Norbert (Eds.), Implicit Measures of Attitudes. Guilford, New York, pp. 59–102.

Oskamp, Stuart, 1991. Attitudes and Opinions. Prentice-Hall, Englewood Cliffs, NJ.

Schwarz, Norbert, 1994. Judgment in a social context: biases, shortcomings, and the logic of conversation. Advances in Experimental Social Psychology 26, 123–162.

Schwarz, Norbert, 2007. Attitude construction: evaluation in context. Social Cognition 25, 638–656.

Sudman, Sudman, Bradburn, Norman M., Schwarz, Norbert, 1996. Thinking about Answers: The Application of Cognitive Processes to Survey Methodology. Jossey-Bass, San Francisco, CA.

Tourangeau, Roger, Rips, Lance J., Rasinski, Kenneth, 2000. The Psychology of Survey Response. Cambridge University Press, Cambridge.

Vargas, Patrick T., Sekaquaptewa, Denise, von Hippel, William, 2007. Armed only with paper and pencil: "Low-tech" measures of implicit attitudes. In: Wittenbrink, Bernd, Schwarz, Norbert (Eds.), Implicit Measures of Attitudes. Guilford, New York, pp. 103–124.

Wittenbrink, Bernd, 2007. Measuring attitudes through priming. In: Wittenbrink, Bernd, Schwarz, Norbert (Eds.), Implicit Measures of Attitudes. Guilford, New York, pp. 17–57.

Wittenbrink, Bernd, Schwarz, Norbert (Eds.), 2007. Implicit Measures of Attitudes. Guilford, New York.

Attitudes and Behavior

Joshua J Guyer and Leandre R Fabrigar, Queen's University, Kingston, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A.S.R. Manstead, volume 2, pp. 908–913, © 2001, Elsevier Ltd.

Abstract

Attitudes have generated interest in the social sciences because they have been presumed to exert a strong influence on behaviors. This article reviews the evolution of perspectives on the attitude–behavior link. It begins with early perspectives that assumed a strong attitude–behavior association and then discusses research and perspectives challenging this assumption. Next, various explanations for why attitudes sometimes fail to predict behaviors are reviewed. The article concludes with a discussion of contemporary themes in attitude–behavior consistency research such as the proposition that the processes by which attitudes influence behavior differs as a function of whether behaviors are deliberative versus spontaneous.

Attitude–Behavior Consistency in Historical Context

For decades, the attitude construct has played a central role in many areas of the social sciences. Although definitions of attitude have varied over time and across disciplines, most contemporary researchers have come to view an attitude as a relatively general and enduring evaluation of an object, person, or concept along a positive to negative dimension (e.g., Eagly and Chaiken, 1993; Fabrigar and Wegener, 2010). Attitudes have generated great interest in the social sciences because they have long been presumed to exert a strong influence on behaviors, decisions, and judgments. Thus, in many applied settings, attitudes are measured in an attempt to predict behavior and are the target of persuasive appeals in an effort to shape behavior.

In the early phases of attitude research, the close association between attitude and behavior was largely an article of faith. Indeed, many early definitions of attitude were based on the assumption that attitudes predispose an individual to behave in a certain manner (e.g., Allport, 1935). Likewise, early theories of attitude structure explicitly assumed that behavior was closely intertwined with attitude. For instance, the tripartite theory of attitudes proposed behavior as one of the three fundamental components of attitudes (e.g., Katz and Stotland, 1959). Similarly, early indirect measures of attitudes such as the lost-letter technique relied on directly observable behaviors that were presumed to be manifestations of attitudes (Milgram et al., 1965). Thus, these measures were based on the implicit assumption that attitudes are strongly associated with behaviors.

Despite being the prevalent belief for the first few decades of attitude research, the assumption that individuals typically act in accordance with their attitudes remained largely untested. However, this belief was sharply challenged following the publication of LaPiere's (1934) widely known study on racial prejudice, which appeared to reveal virtually no correspondence between overt behavior toward a Chinese couple and a subsequent measure of attitudes. Subsequent decades produced a number of other studies that appeared to challenge the proposition that attitudes were predictive of behaviors. Indeed, by the 1960s, researchers began to reevaluate the attitude construct and its relationship to behavior. This debate led some researchers to conclude that little, if any, relationship existed between verbal measures of attitude and overt behavior, and some went so far as to recommend abandoning the attitude construct entirely.

Of these various challenges to the utility of the attitude construct, none was more influential than Wicker's (1969) review of 47 studies that appeared to suggest a rather trivial relationship existed between measured attitudes and overt behavior. More specifically, Wicker (1969) asserted that correlations between measures of attitudes and subsequent behaviors rarely exceeded 0.30 and were often considerably closer to zero. Finding an average correlation of 0.15 across studies, Wicker (1969: p. 65) declared that "it is considerably more likely that attitudes will be unrelated or only slightly related to overt behaviors than that attitudes will be strongly related to actions." He further concluded that "Only rarely can as much as 10% of the variance in overt behavioral measures be accounted for by attitudinal data (p. 65)." Despite presenting a rather gloomy outlook on the status of the attitude–behavior link, Wicker acknowledged the possibility that a large number of factors could have led to a reduction in attitude–behavior correspondence and called for a closer examination of their role in this relationship.

The skepticism expressed by Wicker and others regarding the utility of attitudes as predictors of behaviors did not go unchallenged. For example, both Kelman (1974) and Schuman and Johnson (1976) noted that Wicker (1969) based his review on an extremely limited range of domains of social behavior. Indeed, of the 47 articles included in his review, 20 examined attitudes in relation to workplace performance and absenteeism, 16 dealt with racial prejudice, and a small handful (11) assessed miscellaneous topics – several of which contained multiple studies on the same topic. Thus, based predominantly on two domains of social behavior, Wicker reached sweeping conclusions regarding the nature of the attitude–behavior link. These critics noted that when other domains of social behavior were examined, much larger attitude–behavior correlations were often observed. These criticisms notwithstanding, Wicker's review and that of other skeptics of the attitude–behavior link had an undeniable impact on the field and gave rise to what came to be known as the 'attitude–behavior

problem.' Understanding the implications of the apparent weak association between attitude and behavior became one of the central themes of attitude research throughout the 1970s and 1980s.

Explanations for Low Attitude–Behavior Associations

Since the early 1970s, the guiding questions in attitude–behavior consistency research have been when and why measures of attitudes sometimes fail to be associated with behavior. Responses to these questions have taken a myriad of forms. However, at the general conceptual level, most responses to these questions can be broadly classified as falling into one of three broad themes.

Measurement Explanations

One potential explanation for why an attitude measure might be a poor predictor of a behavior is that the measurement procedure might in some way fail to accurately capture the attitude and/or the behavior. Indeed, a number of attitude researchers (e.g., see Kelman, 1974; Schuman and Johnson, 1976) were quick to point out that numerous studies in Wicker's review utilized poor measurement procedures. For example, some investigations actually failed to include measures of attitudes, but rather made assumptions about participants' attitudes. Other investigations obtained measures of attitude that were likely not taken from the same person who performed the behavior. Still other studies examined topics that were socially sensitive and thus participants might not have honestly reported their true attitudes.

Of these various measurement-related explanations for weak attitude–behavior associations, probably the one that has received the most attention in the literature is the attitude–behavior specificity principle. According to Ajzen and Fishbein (1977) there are four elements (i.e., the action, target, context, and time) that contribute to the generality/specificity of an attitudinal or behavioral measurement. An example of these elements might be drinking coffee (i.e., action) with a classmate (i.e., target) in a local coffee shop (i.e., context) at 3p.m. today (i.e., time). According to the specificity matching principle, if the goal is to predict the specific behavior illustrated above, then the respondent should be asked to provide his or her attitude about drinking coffee with a classmate at a local coffee shop at 3 p.m. today. Of course, such a highly specific measure of attitude would not be expected to be a very accurate predictor of an individual's general coffee drinking behavior. Instead, general coffee drinking behavior should be most accurately predicted by a correspondingly general measure of attitude toward drinking coffee. Taken together, to the degree that each of these four elements is of equivalent specificity in a measure of attitude as well as a measure of overt behavior, one can expect to obtain strong attitude–behavior correspondence.

In one of the first demonstrations of this principle, Weigel and Newman (1976) administered a general attitude survey on environmental issues to residents of a New England town. Three months after completing the survey, participants were contacted on various occasions over the next 5 months and provided with opportunities to engage in specific environment-related activities. Consistent with the principle of specificity, their results indicated general attitudes toward environmental issues were relatively poor predictors of any specific environmental behavior (with most correlations below 0.35), but were significantly better predictors of a comprehensive index that aggregated all the individual environmental behaviors ($r = 0.62$). Reinforcing this point, reviews of the attitude–behavior consistency literature have indicated that instances of strong attitude–behavior correspondence were reliably found when measures of attitude and behavior matched in specificity, but weak and inconsistent attitude–behavior correspondence was frequently revealed when mismatches of specificity occurred (Ajzen and Fishbein, 1977).

Clarifying the Role of Attitude in Behavior

Although flaws in measurement can account for many cases in which attitudes have been found to be poor predictors of behaviors, such methodological explanations cannot fully explain variations in attitude–behavior associations. Thus, a number of attitude researchers have sought more conceptual explanations. One general approach has been to more precisely specify how attitudes might influence behaviors and the potential role of other constructs in this process.

Perhaps the best exemplar of this approach, and certainly the most influential, is the Theory of Reasoned Action (TORA; Ajzen and Fishbein, 1980; Fishbein and Ajzen, 1975). This theory begins with the premise that the effects of attitude on behavior are not direct. Instead, the theory postulates that the most proximal determinant of a behavior is a behavioral intention. A behavioral intention represents the amount of effort an individual is willing to exert in order to perform a given behavior. Because intentions are assumed to capture the motivational factors that influence behavior, as intentions increase in strength, so too does the likelihood that a given behavior will be performed. Thus, from the perspective of the TORA, behaviors can largely be conceptualized as intentional actions that typically fall under the volitional control of the individual. Hence, determining an individual's behavioral intention is the key to yielding accurate predictions of behavior. However, it is important to recognize that there are many factors that can intervene and reduce the predictive power of measures of intentions on behaviors. For example, the extent to which a measure of intention and a measure of behavior are of corresponding levels of specificity, the time elapsed between measurement of an individual's intention and a measure of their behavior, and the introduction of new information which may alter the individual's intention can all influence the intention–behavior link.

The TORA further proposes that one of two major determinants of an individual's behavioral intentions is reflected in their attitude toward the behavior. Attitude toward the behavior can be broken down into behavioral beliefs, which reflect an individual's beliefs regarding the consequences of performing a given behavior, and an evaluation of those consequences. The second major determinant of behavioral intentions is subjective norms, which denote an individual's belief that other people important to the individual concerned expect that a particular behavior should or should not be

performed. Subjective norms, in turn, are comprised of an individual's normative beliefs, which represent the belief that other people important to the individual concerned expect the individual to behave in a specific manner, and their motivation to comply with the expectations of others.

The TORA provides a useful framework for understanding why attitudes are sometimes only moderately correlated with behaviors. First, because the effect of an attitude on behavior is indirect, this necessarily sets an upper boundary on the strength of association that can be expected. For example, even if one assumes a relatively strong association between attitude and intention (e.g., $r = 0.60$) and between intention and behavior (e.g., $r = 0.60$), the expected correlation between attitude and behavior will be comparatively modest ($r = 0.36$). Second, in some contexts, one might expect much more modest associations between attitude and intention. For example, when normative pressures for a behavior are strong (a socially sensitive behavior) and subjective norms contradict the attitude, one might quite reasonably expect an attitude to only weakly influence intentions, and thus be only weakly correlated with behavior.

Over the years, a sizeable body of evidence has accumulated supporting the proposed relationships among attitudes, subjective norms, intentions, and behaviors across a wide range of different domains of social behavior (e.g., see Ajzen and Fishbein, 1980; Fishbein and Ajzen, 1975). Furthermore, the efficacy of predicting behavior using intentions rather than attitude was confirmed in a *meta-analysis* of 87 studies by Sheppard et al. (1988), who found an average correlation between intention and behavior of 0.53. Despite the relative success of the TORA, a major theme of subsequent research has been whether attitudes and subjective norms are sufficient to fully explain variations in behavioral intentions or if the prediction of intentions could be improved by the inclusion of additional factors. Among the factors that have been proposed are habit (e.g., Ouellette and Wood, 1998), moral obligations (Gorsuch and Ortberg, 1983), anticipated regret (Richard et al., 1998), and perceived behavioral control (Ajzen, 1985).

Of these additional factors, perhaps the construct that has been most widely accepted is perceived behavioral control. First introduced within the framework of the Theory of Planned Behavior (TPB; Ajzen, 1985), this construct reflects a person's beliefs regarding the difficulty of performing a given behavior. Enacting some behaviors is unhindered by internal or external factors, and thus are instances of high perceived control. In other cases, the performance of behaviors can be impeded by a combination of these factors and thus represent instances of low perceived control. Examples of internal factors include an individual's skills, abilities, knowledge, and adequate planning; whereas examples of external factors might include time, opportunity, and reliance on the cooperation of others. The TPB identifies perceived behavioral control as a separate variable that can influence behavior directly as well as indirectly, mediated by behavioral intentions. The theory predicts a direct effect of perceived control on behavior when an individual's behavior is unlikely to be entirely under volitional control and perceptions of behavioral control are accurate. The indirect effects of perceived behavioral control via intention reflect the assumption that inherent in perceived behavioral control are motivational properties that impact an individual's intentions to perform a behavior. If people believe that a given behavior is outside of their control, behavioral intentions may be reduced even though favorable attitudes and/or subjective norms regarding performance of the behavior exist. This suggests another context in which attitude–behavior correspondence may be weaker than anticipated. If perceived control over performance of the behavior is relatively low, a positive attitude might not be sufficient to produce a strong intention to perform the behavior.

Research investigating perceived behavioral control suggests the addition of this construct has led to significant improvement in the prediction of behavioral intentions as well as actual behaviors (Madden et al., 1992). In line with this, McEachan et al. (2011) recently conducted a *meta-analysis* examining the contributions of perceived behavioral control to the theory of planned behavior. Their data revealed correlations between behavioral intentions and attitudes, subjective norms, and perceived behavioral control ranged from 0.40 to 0.57, and generated a multiple correlation of 0.67.

A second construct acknowledged by many researchers to have an important role in the attitude–behavior relationship is habit (typically operationalized using measures of past behavior; e.g., see Ouellette and Wood, 1998). Most contemporary theorists define habits as tendencies to repeat certain behavioral responses that generally require minimal thought and effort and that occur within the confines of specific supporting contexts. Indeed, research has shown that repetition of a behavior within a specific context frequently leads to the development of cognitive processes that become automatic, thus resulting in minimal effort and attention to perform a behavior. Based on this logic, proponents of the habit construct argued that within contexts that support the development of habits or where specific behavior patterns are already well learned, habits will be strong predictors of future behavior.

Advocates of the habit construct have argued that because the Theory of Reasoned Action assumes that behavior is largely under volitional control of the individual, the theory is better suited for predicting behaviors involving a strong deliberative component rather than those behaviors characterized by habitual or spontaneous responding. Several empirical tests investigating the addition of habit (i.e., past behavior) as a predictor have indicated improved prediction of intentions as well as actual behavior. Most notably, Ouellette and Wood (1998) conducted a *meta-analysis* of 64 studies. Their findings suggested the magnitude of the effect of habit on behavior ($r = 0.39$) was generally equivalent with that of alternative predictors. Although the effect was somewhat weaker than intentions ($r = 0.54$), the data revealed a slight improvement over the effect of attitudes ($r = 0.33$), and a moderate increase compared with the effects of perceived behavioral control ($r = -0.21$) and subjective norms ($r = 0.23$). This suggests that at least some of the low and inconsistent relationships between attitude and behavior found in prior research may be explained when considering the role of habit in behavior.

Attitude Strength

A second conceptual response to understanding variations in the strength of the attitude–behavior link was to focus on more

precisely understanding variations in the underlying cognitive and motivational foundations of attitudes. The central premise of this approach is that although two people might report similar attitudes, the underlying strength of these evaluations might well differ. Thus, when assessing the attitudes of groups of people, it might be inappropriate to assume that the entire sample is reporting attitudes that reflect stable and meaningful evaluations of a particular object. Instead, within any sample of individuals there may only be a subset of people who have relatively well-formed, meaningful attitudes. For this later group, the link between attitudes and behavior might be quite strong, but this strong link might well be obscured by people in the sample who lack strong attitudes on the topic. Indeed, if the group lacking strong attitudes is larger than the group who holds strong attitudes, the overall attitude–behavior correlation might be quite weak. Of course, an important question that naturally arises from this perspective is how one determines whether an evaluation is meaningful and relatively well-formed? To this end, a great deal of research throughout the 1970s and 1980s was focused on identifying the specific properties of attitudes that determine their strength (for reviews, see Fabrigar and Wegener, 2010; Fabrigar et al., 2005; Fazio and Roskos-Ewoldsen, 2005; Petty and Krosnick, 1995; Visser et al., 2006). This research has documented numerous properties of attitudes that have been found to moderate the strength of the attitude–behavior link.

Some of these properties reflected features of the underlying cognitive structure of attitudes. Of these, one of the most widely investigated properties has been the accessibility of an attitude in memory. Attitude accessibility is defined as the strength of the object-evaluation link in memory and is represented by the ease or speed with which the attitude is brought to mind. Numerous studies in both laboratory and field settings have found that greater attitude accessibility is associated with stronger attitude–behavior associations, presumably because of the greater likelihood that an attitude will be activated from memory at the time of the behavior. A second property of cognitive structure that has been studied extensively in the context of attitude–behavior consistency is attitude ambivalence. This construct is defined by the simultaneous possession of positive and negative evaluative responses toward an object. Highly ambivalent attitudes are characterized by both positive and negative evaluations of an object whereas attitudes low in ambivalence reflect predominantly positive or negative evaluations of an object. The majority of studies indicate attitude–behavior associations become weaker as ambivalence increases. Other properties of cognitive structure that have been examined in the context of attitude–behavior consistency include working knowledge, affective-cognitive bases of attitudes, and attitude complexity.

Other determinants of strength explored in the context of attitude–behavior consistency have involved subjective beliefs regarding the attitude or the attitude object. Of these subjective beliefs, the two most widely researched have been attitude importance and attitude certainty. Attitude importance can be conceptualized as the subjective sense of psychological significance an individual attaches to an evaluation or an object (Eaton and Visser, 2008). Increased importance has been found to be associated with higher attitude–behavior consistency across a range of different social behaviors. Attitude certainty reflects the extent to which an individual is confident in their attitude toward a particular object (Tormala and Rucker, 2007). Recently, attitude theorists have proposed that certainty can be decomposed into two conceptually distinct components: attitude clarity (i.e., the subjective sense that one has a clear notion of their evaluation of a given object) and attitude correctness (i.e., the subjective sense that one's evaluation of the object is valid). Research in a variety of domains has found that higher attitude certainty generally translates into stronger attitude–behavior correspondence.

Other determinants of attitude strength have reflected variations in the processes by which attitudes are formed or changed. The first determinant of this type explored in the literature was direct versus indirect experience with the attitude object. Research has indicated that attitudes formed via direct experience are more predictive of behaviors than are attitudes formed via indirect experience. More widely studied has been the extent to which an attitude is formed or changed via extensive cognitive elaboration of attitude-relevant information. Influential theories of attitude formation and change such as the Elaboration Likelihood Model (ELM; Petty and Wegener, 1999) and the Heuristic-Systematic Model (HSM; Chen and Chaiken, 1999) have postulated that attitudes can be formed and changed via highly thoughtful versus nonthoughtful processes. Highly thoughtful attitude formation and change occurs when an individual is highly motivated and able to carefully scrutinize the central merits of information regarding the attitude object (sometimes referred to as the *central route* to attitudes). In contrast, when a person lacks motivation and/or ability to engage in extensive cognitive elaboration of attitude-relevant information, the person will look for simple cues in the formation/persuasion context (e.g., the credibility of the source of the message) to provide a nonthoughtful basis for arriving at an attitude (sometimes called the *peripheral route* to attitudes). The ELM postulates, and subsequent empirical research has confirmed, that attitudes arrived at via careful thought are more enduring and consequential and therefore more powerful determinants of behavior than attitudes arrived at via relatively nonthoughtful processes.

Contemporary Themes in Attitude–Behavior Consistency Research

Since Wicker's influential challenge, the attitude–behavior consistency literature has gradually evolved well beyond its central focus on whether attitudes predict behavior. For the most part, attitude researchers now no longer support or even seriously debate the extremely pessimistic conclusions advanced in the late 1960s regarding the utility of attitudes as predictors of behaviors (e.g., see Eagly and Chaiken, 1993; Fabrigar et al., 2010; Fazio and Roskos-Ewoldsen, 2005). In fact, attitude researchers (e.g., Eagly and Chaiken, 1993) have sometimes expressed surprise that given the vast body of empirical evidence that has accumulated in contradiction to the pessimistic conclusions of the late 1960s, some social psychologists continue to believe that attitudes are only weakly related to behavior. Moreover, it is now recognized that general characterizations of the association between attitudes and behaviors as weak or strong is misleading because conditions

exist in which associations between attitudes and behaviors can be either extremely strong or nonexistent. Thus, attitude–behavior consistency has moved beyond the question of whether attitudes predict behaviors, and even to some extent beyond the question of when attitudes do or do not predict behaviors. In recent years, the focus in attitude–behavior consistency has shifted in an effort to more precisely understand the processes by which attitudes influence behaviors and to more fully understand the processes underlying the effects of various moderators of attitude–behavior consistency.

Distinguishing between Deliberative and Spontaneous Behaviors

Until as recently as the late 1980s, much of the research on attitude–behavior consistency assumed that a single process could be used to describe the manner in which attitudes influence behaviors. Indeed, two of the more prominent theories of the attitude–behavior relationship, the Theory of Reasoned Action (TORA) and Theory of Planned Behavior (TPB), were advanced as theoretical frameworks that could generally be applied under all conditions. However, this notion has been increasingly challenged in contemporary attitude–behavior consistency research.

Of the theories to challenge this assumption, the most prominent has been the *M*otivation and *O*pportunity as *DE*terminants (MODE) model of attitude–behavior consistency (Fazio, 1990; Fazio and Towles-Schwen, 1999). According to the MODE, attitudes serve as guides when making decisions and thus have the potential to influence behavior through affecting an individual's appraisal of their decision alternatives. The MODE model distinguishes between two different types of attitude-to-behavior processes. This distinction revolves around the extent to which choosing a particular behavioral response entails effortfully deliberating between alternative courses of action or a relatively effortless and spontaneous reaction to one's immediate circumstances. When people are motivated and have the opportunity to carefully consider their actions, people will engage in a relatively effortful deliberation of the costs and benefits of performing a specific behavior and thus form an attitude toward the specific behavior. In such a situation, the process by which an attitude influences a behavior will generally resemble the processes as outlined by theories such as the TORA and TPB.

In other situations, people may lack motivation to carefully consider their actions (e.g., if the decision has little importance) or the opportunity to do so (e.g., if distractions are present in the decision environment or the decision must be made very quickly). In spontaneous behavioral contexts of this sort, the MODE proposes that people will be unlikely to carefully consider the positive and negative consequences of a given action and form a specific attitude toward the behavior in question and subsequently a behavioral intention. Thus, theories such as the TORA and TPB may provide a poor description of the processes by which attitudes influence behavior for spontaneous behaviors. Rather, the MODE postulates that more general attitudes toward the target (or targets) of the behavior will be the more consequential attitude (attitudes). Under spontaneous conditions, when an individual encounters the object that is the target of the behavior, the attitude toward that object will be activated in memory (presuming it is sufficiently accessible). This attitude then biases the perception of the object in an attitude-congruent fashion and perceptions of the object will in turn influence how the behavioral situation is construed. Construal of the situation will in turn determine the behavioral response that is undertaken.

To illustrate this process, consider a case in which you are approached by a man who is a member of a particular ethnic group who explains he has lost his wallet and would like you to provide him with $2 to purchase bus fare to get home. If your attitude toward that ethnic group is sufficiently accessible, it will be activated and influence your perceptions of the requestor. A positive attitude might result in you attending to the positive features of the person (e.g., the friendly facial expression and polite wording of the request). A negative attitude might direct your attention to negative features of the person (e.g., shabby clothing). These perceptions in turn could strongly influence if you construe the situation as a sincere request from someone in need (in which case you might well assist him) versus a dishonest ploy to get money (in which case you might decline the request).

Research on the MODE has supported a number of its key premises regarding differences in deliberative versus spontaneous behavior (see Fazio, 1990; Fazio and Towles-Schwen, 1999). For example, in one early test of the MODE, Sanbonmatsu and Fazio (1990) conducted an experiment in which they presented individuals with information regarding two department stores: Brown's Department Store and Smith's Department Store. The information provided about the stores was such that Brown's was described in a positive manner for most of its departments, but its camera department was described as comparatively poor. In contrast, information regarding Smith's was negative for most departments, but information about its camera department was positive. Participants were then asked to indicate which store they would choose if they needed to buy a camera. Motivation at the time of the decision was manipulated by requiring respondents to either justify their responses or not, and opportunity by limiting the amount of time respondents had to make their decision or not.

Following the logic of theories such as the TORA, people should consider the specific positive and negative implications of purchasing a camera from both stores and then form a specific attitude toward purchasing a camera from one of the stores versus the other. This attitude should in turn lead to an intention to purchase a camera from one of the stores. Such a process would suggest that the great majority of participants should select Smith's department store because it has the superior camera department. This outcome in fact did occur when participants had both motivation and opportunity to deliberate about their decisions. However, when motivation and/or opportunity were low, most participants actually selected Brown's despite the fact its camera department was inferior. This finding suggests that people did not consider the specific implications of their behavioral actions, but instead were influenced by their more global attitudes toward the stores (even though those global attitudes implied a non-optimal decision). This later finding is consistent with the predictions of the MODE and not readily explainable using theories such as TORA.

As it relates to attitude–behavior correspondence, findings from tests of the MODE suggest that established theories such as the TORA and TPB remain useful accounts of attitude–behavior consistency when people have sufficient motivation and sufficient opportunity to deliberate about their actions. In such cases an attitude toward performing a specific behavior is likely to be the best predictor of performance of that specific behavior. However, when sufficient motivation and/or opportunity are lacking, more global attitudes toward the target or targets of the behavior might well be more useful predictors of even a specific behavior. In summary, when attempting to predict behavioral responses, the extent to which that behavior is likely to be performed in a deliberative versus spontaneous manner needs to be considered.

Implicit and Explicit Attitude Measures

Another topic that has received significant attention in contemporary attitude–behavior consistency literature is the utility of explicit versus implicit measures of attitudes as predictors of behaviors (for reviews, see Petty et al., 2008). Explicit attitude measures directly ask people to report their attitudes and are represented by various traditional self-report measures that have been commonly used in attitudes research for decades. In contrast, implicit measures do not ask people to directly report their attitudes. Instead, attitudes are indirectly measured, usually using some form of priming paradigm or dual judgment task. Among the implicit attitude measures that have enjoyed substantial popularity in the literature are affective or evaluative priming, the Implicit Association Test (IAT), the Personalized Implicit Association Test (PIAT), the Affect Misattribution Procedure (AMP), and the Go/No Go procedure.

One interesting finding to emerge in this literature has been that implicit measures tend to be better predictors than explicit measures of automatic or spontaneous behaviors and decisions whereas explicit measures tend to be better predictors than implicit measures of more controlled behaviors and decisions (e.g., Dovidio et al., 1997). This basic finding has been demonstrated in a number of studies, but its interpretation remains a matter of much debate. Some have argued that implicit measures and explicit measures capture fundamentally different types of attitudes that are represented in distinctly different memory systems (i.e., implicit/unconscious attitudes versus explicit/conscious attitudes). Thus, these researchers have argued that spontaneous or automatic behaviors are more influenced by our unconscious evaluations whereas deliberative behaviors are more influenced by our conscious evaluations. Others have argued that such measures do not represent fundamentally different attitudes in memory, but instead reflect a single stored evaluation that is manifested in both automatic evaluative judgments (which are comparatively free of motivational concerns and higher level cognitions regarding the evaluation) and more thoughtful evaluative judgments. These more thoughtful evaluative judgments are influenced by the evaluation in memory as well as motivational concerns (e.g., self-presentation concerns and desires to avoid bias) and *metacognitions* regarding the evaluation (e.g., beliefs regarding the validity of the attitude). Thus, these researchers have argued that because spontaneous behaviors should also be less influenced by motivational concerns and *metacognitions*, implicit measures should logically be better predictors of these behaviors. Conversely, because deliberative behaviors are more likely to be influenced by motivational concerns and *metacognitions*, they should be better predicted by explicit measures which are also influenced by these factors.

Processes Underlying Moderators of Attitude–Behavior Consistency

A final theme that has recently emerged in the literature is a greater focus on understanding why specific moderators of attitude–behavior consistency have their effects. For example, as was noted earlier, a number of properties of attitudes have been found to moderate the magnitude of the attitude–behavior link. However, attitude theorists (e.g., Fabrigar et al., 2005, 2010; Fazio and Roskos-Ewoldsen, 2005) have noted that although the list of moderators of attitude–behavior consistency is quite lengthy, the psychological mechanisms responsible for many of these moderator effects are unclear. As such, it is difficult to know to what extent various moderators reflect similar versus distinct underlying processes. Thus, researchers have returned to the study of well-established moderators of attitude–behavior consistency in an effort to better understand the mechanisms responsible for these effects. Indeed, some researchers have postulated that the wide range of moderators in the literature might well be explained by a relatively small set of underlying mechanisms. One recent exemplar of this approach is the Mechanisms Responsible for Prediction and Influence (MRPI) model (Fabrigar et al., 2010; see also Fabrigar and Wegener, 2010). The MRPI proposes that there are three mechanisms that may account for variations in the impact of attitudes on behavior. Building on the MODE, the MRPI postulates that the emergence of these processes depends on whether a behavior is performed in a deliberative versus spontaneous manner.

When people are spontaneous, the theory holds that attitudes influence behavior by serving as simple nonthoughtful cues to appropriate action. Indeed, as discussed earlier, under spontaneous conditions attitudes can provide a simple basis for action even when it might not be logical to rely upon them. However, to serve as a simple cue, an attitude must be activated from memory at the time of the behavior. Thus, under low-deliberation conditions, the theory argues that potential moderators of attitude–behavior consistency will have an effect to the extent that they are related to the likelihood of attitude activation (i.e., attitude accessibility).

When people are motivated and able to be thoughtful, the MRPI postulates that attitude–behavior consistency can be thought of as a type of inference process in which people consider (if the attitude is activated) whether the attitude provides an informative guide for the behavior. That is, people consider whether the bases of their attitude are relevant to the goals of the behavior. If the attitude is seen as a relevant guide, it will influence behavior. If it is judged as uninformative for the behavior, it will have little influence. Consideration of the relevance of attitudes should most likely occur when behaviors are highly deliberative because the metacognitive process of judging the applicability of an attitude and disregarding it or correcting for its influence

requires considerable cognitive resources. Thus, under high deliberation, potential moderators of attitudinal influence will exert effects to the extent that they have an impact on attitude accessibility and/or the extent to which an attitude is perceived as a valid guide to behavior.

To date, testing of the MRPI is still in its early stages. Evidence does exist for each of the processes proposed by the MRPI in the context of some moderators of attitude–behavior associations. However, no single moderator has been examined in light of all the processes and some moderators have yet to be examined in light of any of these processes. Thus, much research remains to be conducted before the utility of the MRPI can be fully determined. However, regardless of the ultimate outcome of these tests, the general approach of more fully exploring the processes responsible for the effects of moderators of attitude–behavior consistency is likely to continue to be a guiding theme in the literature.

Conclusions

The link between attitude and behavior has been of interest to attitude researchers since attitudes first became a topic of systematic empirical inquiry. Although perspectives on this link have varied dramatically over the years, there is little doubt that important advances have been made. The question of whether attitudes can predict behavior has largely been put to rest and influential theories designed to explain the processes by which attitudes exert their influence on behavior have been proposed. Moreover, researchers have documented an impressive array of factors that can account for when attitudes will and will not be useful predictors of behavior, and have now begun to grapple with the mechanisms underlying the effects of these factors.

See also: Attitude Formation and Change; Attitude Measurement; Cognitive Dissonance; Heuristics in Social Cognition; Implicit Association Test; Implicit Social Cognition; Persuasion Theories; Social Cognition; Social Psychology; Values, Psychology of.

Bibliography

Ajzen, I., 1985. From intentions to actions: a theory of planned behavior. In: Kuhland, J., Beckman, J. (Eds.), Action-control: From Cognitions to Behavior. Springer, Heidelberg, pp. 11–39.

Ajzen, I., Fishbein, M., 1977. Attitude–behavior relations: a theoretical analysis and review of empirical research. Psychological Bulletin 84, 888–918.

Ajzen, I., Fishbein, M., 1980. Understanding Attitudes and Predicting Social Behavior. Prentice-Hall, Englewood Cliffs, NJ.

Azjen, I., Madden, T.J., 1986. Prediction of goal-directed behavior: attitudes, intentions, and perceived behavioral control. Journal of Experimental Social Psychology 22, 453–474.

Allport, G.W., 1935. Attitudes. In: Murchison, C. (Ed.), Handbook of Social Psychology. Clark University Press, Worcester, MA, pp. 798–844.

Chen, S., Chaiken, S., 1999. The Heuristic-systematic model in its broader context. In: Chaiken, S., Trope, Y. (Eds.), Dual-process Theories in Social Psychology. Guilford Press, New York, pp. 73–96.

Dovidio, J.F., Kawakami, K., Johnson, C., Johnson, B., Howard, A., 1997. On the nature of prejudice: automatic and controlled processes. Journal of Experimental Social Psychology 33, 510–540.

Eagly, A.H., Chaiken, S., 1993. The Psychology of Attitudes. Harcourt Brace Jovanovich, Fort Worth, TX.

Eaton, A.A., Visser, P.S., 2008. Attitude importance: understanding the causes and consequences of passionately held views. Social and Personality Psychology Compass 2, 1719–1736.

Fabrigar, L.R., MacDonald, T.K., Wegener, D.T., 2005. The structure of attitudes. In: Albarracin, D., Johnson, B.T., Zanna, M.P. (Eds.), Handbook of Attitudes and Attitude Change. Erlbaum, Mahwah, NJ, pp. 79–124.

Fabrigar, L.R., Wegener, D.T., 2010. Attitude structure. In: Baumeister, R.F., Finkel, E.J. (Eds.), Advanced Social Psychology: The State of the Science. Oxford University Press, New York, US, pp. 177–216.

Fabrigar, L.R., Wegener, D.T., MacDonald, T.K., 2010. Distinguishing between prediction and influence: multiple processes underlying attitude–behavior consistency. In: Agnew, C.R., Carlston, D.E., Graziano, W.G., Kelly, J.R. (Eds.), Then a Miracle Occurs: Focusing on Behavior in Social Psychological Theory and Research. Oxford University Press, New York, pp. 162–185.

Fazio, R.H., 1990. Multiple processes by which attitudes guide behavior: the MODE model as an integrative framework. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 23. Academic Press, San Diego, CA, pp. 75–109.

Fazio, R.H., Roskos-Ewoldsen, D.R., 2005. Acting as we feel: when and how attitudes guide behavior. In: Brock, T.C., Green, M.C. (Eds.), Persuasion: Psychological Insights and Perspectives. Sage Publications, Thousand Oaks, CA, pp. 41–62.

Fazio, R.H., Towles-Schwen, T., 1999. The MODE model of attitude–behavior processes. In: Chaiken, S., Trope, Y. (Eds.), Dual-process Theories in Social Psychology. Guilford Press, New York, pp. 97–116.

Fishbein, M., Ajzen, I., 1975. Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research. Addison-Wesley, Reading, MA.

Gorsuch, R.L., Ortberg, J., 1983. Moral obligation and attitudes: their relation to behavioral intentions. Journal of Personality and Social Psychology 44 (5), 1025–1028.

Katz, D., Stotland, E., 1959. A preliminary statement to a theory of attitude structure and change. In: Koch, S. (Ed.), Psychology: A Study of a Science: Vol. 3 Formulations of the Person and the Social ContextMcGraw-Hill, New York, pp. 423–475.

Kelman, H.C., 1974. Attitudes are alive and well and gainfully employed in the sphere of action. American Psychologist 29, 310–324.

LaPiere, R.T., 1934. Attitudes vs. actions. Social Forces 13, 230–237.

Madden, T.J., Ellen, P.S., Ajzen, I., 1992. A comparison of the theory of planned behavior and the theory of reasoned action. Personality and Social Psychology Bulletin 18, 1–8.

McEachan, R.R.C., Conner, M., Taylor, N., Lawton, R.J., 2011. Prospective prediction of health-related behaviors with the theory of planned behavior: a meta-analysis. Health Psychology Review 5, 97–144.

Milgram, S., Mann, L., Harter, S., 1965. The lost-letter technique: a tool for social research. Public Opinion Quarterly 29, 437–438.

Ouellette, J.,A., Wood, W., 1998. Habit and intention in everyday life: the multiple processes by which past behavior predicts future behavior. Psychological Bulletin 124, 54–74.

Petty, R.E., Fazio, R.H., Brinol, P. (Eds.), 2008. Attitudes: Insights from the New Implicit Measures. Psychology Press, New York.

Petty, R.E., Krosnick, J.A. (Eds.), 1995. Attitude Strength: Antecedents and Consequences. Ohio State University Series on Attitudes and Persuasion, vol. 4. Lawrence Erlbaum Associates, Mahwah, NJ.

Petty, R.E., Wegener, D.T., 1999. The elaboration likelihood model: current status and controversies. In: Chaiken, S., Trope, Y. (Eds.), Dual-process Theories in Social Psychology. Guilford Press, New York, pp. 41–72.

Richard, R., de Vries, N.K., van der Pligt, J., 1998. Anticipated regret and precautionary sexual behavior. Journal of Applied Social Psychology 28, 1411–1428.

Sanbonmatsu, D.M., Fazio, R.H., 1990. The role of attitudes in memory-based decision making. Journal of Personality and Social Psychology 59, 614–622.

Schuman, H., Johnson, M.P., 1976. Attitudes and behavior. Annual Review of Sociology 2, 161–207.

Sheppard, B.H., Hartwick, J., Warshaw, P.R., 1988. The theory of reasoned action: a meta-analysis of past research with recommendations for modifications and future research. Journal of Consumer Research 15, 325–343.

Tormala, Z.L., Rucker, D.D., 2007. Attitude certainty: a review of past findings and emerging perspectives. Social and Personality Psychology Compass 1, 469–492.

Visser, P.S., Bizer, G.Y., Krosnick, J.A., 2006. Exploring the latent structure of strength-related attitude attributes. Advances in Experimental Social Psychology 38, 1–67.

Weigel, R.H., Newman, L.S., 1976. Increasing attitude–behavior correspondence by broadening the scope of the behavioral measure. Journal of Personality and Social Psychology 33, 793–802.

Wicker, A.W., 1969. Attitudes versus actions: the relationship of verbal and overt behavioral responses to attitude objects. Journal of Social Issues 25, 41–78.

Attitudes, Political and Civic Culture

GA Almond, Stanford University, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G.A. Almond, volume 3, pp. 1862–1865, © 2001, Elsevier Ltd., with revisions made by the Editor.

Civic Culture theory asserts that democracy is stable or consolidated when specifically democratic attitudes and practices combine and function in equilibrium with certain non democratic ones. It is a literature which reconciles role theory with citizenship theory, recognizing that civic participation and activity is one of the many roles performed by most individuals and that effective and good citizenship requires a certain harmonization of these roles. 'Strong Democracy,' 'Participatory Politics,' various kinds of 'populism,' tend to deny respect to, and to invade, subject, parental, matrimonial, professional, recreational, and other roles. The concept of Civic Culture as an equilibrium is a dynamic one; that is to say, changes in demography, economy, literacy and education, media exposure and impact, culture, and lifestyle, etc. may transform this equilibrium.

Civic Culture theory asserts that democracy is stable or consolidated when specifically democratic attitudes and practices combine and function in equilibrium with certain nondemocratic ones. It was formulated and tested in empirical research in the late 1950s and early 1960s, years still reverberating with the memories of the rise of Communism and Fascism, the collapse of democracy and the catastrophes of World War II. It drew from a contemporary social science literature similarly influenced by the interwar history of the stalemated Third French Republic, the deeply flawed Weimar Germany, the Austrian and Spanish civil wars, and the breakdown of the Fourth French Republic.

It could point to the long tradition in political theory of 'mixed government,' from Plato and Aristotle through to Montesquieu, which supported this anti-populism and prudentialism. Among the post World War II social science influences were works of Joseph Schumpeter, Paul Lazarsfeld and Bernard Berelson, Edward Shils, Robert A. Dahl, and Harry Eckstein, among others.

Schumpeter rejected the 'classic democratic' assumption of the necessity of an informed, activist, rational public for a genuine democracy, and proposed in its place, an 'elites competing for votes' theory (Schumpeter, 1947, Chapters 21 and 22). This minimalist theory could be reconciled with more realistic assumptions of a relatively ignorant and indifferent demos.

Paul Lazarsfeld and Bernard Berelson, theorizing from their 'panel' voting studies of the 1940s similarly saw democracy as associated with a set of cultural and social conditions having the effect of limiting the intensity of social conflict. These included relative economic and social stability, a pluralistic social organization, a basic value consensus and what we would now call a 'civil society.' They described a democratic equilibrium as involving mixes of involvement and indifference, stability and flexibility, consensus and cleavage (Berelson et al., 1954).

Edward Shils, on the democratic prospects of the new nations, emphasized the importance of a 'widely dispersed civility.' By this he meant a moderate sense of nationality, a degree of interest in public affairs, a consensus on values, institutions, and practices, and a recognition of individual rights and obligations. He wrote " … These qualities should not be intense, and they need not be either equally or universally shared" (Shils, 1960: 387ff).

Robert Dahl's theory of polyarchy, elaborated in 1956 in a full-length contrast with populistic and Madisonian democracy, belongs among these early social science influences on the civic culture. His early characterization of the American political system as providing "… a high probability that any active and legitimate group will make itself heard effectively at some stage in the process of decision … " (Dahl, 1956: 145; see also Dahl, 1970, 1989) reflected the minimalist mood that Dahl shared with the generation that emerged out of the great depression and World War II. At the time that *The Civic Culture* was being written, Dahl's concept of polyarchy already had introduced an empirically grounded and quantitative set of concepts into democratic theorizing. Democracy was not an essence. In its full sense it did not exist, and probably could not exist. Hence the concept *polyarchy* was developed to refer to real political entities which attained measured performance levels on specified empirical dimensions.

Harry Eckstein was the first to emphasize the 'mixed' or paradoxical side of democracy, recognizing the necessity for a democracy not only to represent and formulate the will of the public, but to govern it authoritatively. In his *Theory of Stable Democracy* he describes some of the qualities which enable democracies to reconcile responsible authority and democratic responsiveness. Such reconciliation is facilitated by balances among contrasting qualities; participant behavior is balanced by deference to authority, dogmatism by pragmatism, and the like. Institutionally he attributes democratic stability to the degree to which social authority patterns coincide or are 'congruent' with political ones.

Civic culture theory had the advantage of being formulated in the context of a major empirical investigation informed by this historical experience, and benefiting from this prior research and scholarship. Its results were reported in book form in 1963, in paperback form in 1965; and were reprinted in 1989 (Almond and Verba, 1963, 1965, 1989) and remains in print as of this publication. It was widely reviewed in social science periodicals and in 1980 a retrospective volume was published with critiques of the theory and the findings (Almond and Verba, 1980).

The data were made available in the Interuniversity Consortium for Political and Social Research at the University of Michigan, and have been utilized in many secondary studies.

Four of the five countries that it investigated were chosen because they exemplified democratic stability and instability in the first half of the twentieth century – the United Kingdom and the United States exemplifying stability on the one

hand – Germany and Italy exemplifying democratic instability and breakdown on the other. The fifth, Mexico, was a target of opportunity, selected with the thought that it might provide some insight into problems of democratization outside the North American-European area. The method used in the study, of combining structured and open-ended questions administered to probability samples of national populations, provided a richer set of data specifically responsive to questions arising out of this historical experience and body of speculative theory.

The conception of stable democratic political culture as a 'mixed' political culture received a fuller elaboration than it had been given in the earlier work of Berelson and Lazarsfeld and Eckstein. The mix of democratic political culture was based on political role theory. People in stable democracies were both citizens and subjects, and they needed to accommodate their non political, private, and parochial roles. A thriving, stable democracy consists not only of voters, demonstrators, petition signers, and politician button-holers; but of taxpayers, jurors, and military conscripts; as well as parents, mates, work-persons, voluntary association-members, vacationers, and private, self-involved individuals.

It is this mix of roles – participant as well as subject, non political as well as political – that democratic citizens of a stable democracy must balance and accommodate; and which their institutions must choreograph in a process of converting demands and supports into outputs and outcomes. To shift metaphors, civic culture theory was an equilibrium theory in which political buyers and sellers reach prices at which the political market is cleared. Civic culture theory specified what conditions had to be present in order to clear these markets.

These mixes and balances were located empirically in the British and American cases in the late 1950s and early 1960s – this combination of influentialism and deferentialism, involvement and indifference, conflictual and consensual attitudes, principled and instrumental ones. The relative absence of these balances in the German and Italian cases was noticeable. There was more deferentialism and less participationism in the British case than in the American. A 'reserve of influence' was also noted in the American and British cases, based on the finding that Americans and the British acknowledged the obligation to participate far more frequently than they reported actually participating, This discrepancy between performance and obligation could be viewed as a kind of 'default' mode, a reserve supply of participatory energy available for crises. The civic culture would run cool normally and at a moderate speed, but it had a reserve of influence to draw upon in the twists and turns of democratic politics, as the concerns and interests of different groups of voters were engaged.

By the time the retrospective volume, *The Civic Culture Revisited* was published in 1980 it was evident that British and American civic culture were in trouble (Almond and Verba, 1980, Chapters 5 and 6).

The balance of consensus and conflict had moved toward conflict; pride in nation and confidence in government were down. Participationism had declined. In contrast Germany showed dramatic gains in social trustfulness, confidence in government, and civic competence (Almond and Verba, 1980, Chapter 7). In Italy political alienation and extreme partisan antagonism continued largely unchanged. In Mexico the political culture of ambivalent belief in the legitimacy of the democratic revolution, and the corruption of politicians and office-holders, also still survived (Almond and Verba, 1980, Chapter 9).

That patterns of political culture would change in response to changes in economy, demography, politics and public policy, communications technology, and popular education should not have occasioned surprise. That the exemplars of the civic culture of the 1950s – Britain and the United States – should be showing signs of wear and tear in the 1970s and 1980s; and that the problem child of democracy – Germany – was showing strong signs of an emerging civic culture, were not causes for rejecting civic culture theory. The question was whether the changes observed in the two decades after the civic culture study, were in a direction which sustained or disproved the theory. The evidence was moderately supportive of the theory. Thus, for example, the withdrawal of Johnson from the 1968 presidential race, despite the fact that political tradition would have legitimized another term of office, and the resignation of Nixon from the presidency in the 1970s were clear evidences of American instability; and the political disorders of the 1960s and 1970s were clear evidences of cultural disequilibria of one kind or another – conflict had undermined consensus, the legitimacy of government had declined, the modes of participation had radicalized. In contrast Germany had had several decades of experience of effective political leadership and remarkable economic growth appearing to produce a moderating *koalition-fahig* partisanship, growing popular trust in government, civic obligation, and the like.

The place of civic culture theory in the contemporary theory of democracy is in some doubt. In the continuing theoretic polemic about the nature of democracy anything settling for less than perfection partakes of sin. It was precisely to avoid this commingling of the sacred and the secular that led Robert Dahl to invent the concept of polyarchy and to place the concept of *democracy* somewhat, but not completely, off limits. However Dahl's Polyarchy III (which is the closest to ultimate democracy that he gets) is achieved through increasing the depth and extent of attentive publics within the larger mass public, corresponding to the significant issues confronting the polity and policy elites. Modern information technology makes it possible that this gap between policy elites and the mass public might be significantly reduced (Dahl, 1989: 338ff).

The state of the polemic regarding the competence, rationality, and potential effectiveness of mass publics in contemporary polyarchies is well argued in a symposium held during the 1990s. The merits of the various options – elitism, inventive utilization of information technology, or reducing the scope of politics – are among the issues debated (Friedman, 1999).

Civic Culture theory might have enriched this discussion somewhat by affirming the legitimacy of other than political claims upon humankind. If one views the full range of role demands on the time and resources of humans, how do we choose among them? How do these choices interact? What are

the tradeoffs, synergies, and opportunity costs? How do we weight the claims of the civic world against the demands of profession, family, edification, and pleasure?

Editors' Note

Gabriel Almond, along with co-author Sidney Verba, originated the theory and concept of 'civic culture' in their 1963 book of that title. In addition to its rich conceptual and theoretical insights, the book was also an early exemplar of the need for and value of comparative political research. Almond died on Christmas Day, 2002, and his essay from the first edition is being reprinted here as a legacy entry.

Since the publication of Almond's essay above, *The Civic Culture* (1963) has continued to receive both complements and criticisms of its model, its methodology, its typology, and its conclusions (for useful reviews, see Dalton, 2014: Chapter 12; Verba, 2011). In the event, its value remains its account of the cultural requirements of a democratic polity in particular, and of the relationships between political culture and political regime more generally. And, recent citizen actions in currently democratic, democratizing, and non democratic countries, including those in the Middle East or the successor states of the former Soviet Union, such as Ukraine, have subjected this account to additional important research.

This important research focuses on, for example:

1. intra-cultural change, such as from 'allegiance' to 'assertiveness' among citizens and, relatedly, in citizens' (dis) satisfaction with government and political institutions and officials (see. e.g., Dalton and Welzel, 2014 and Norris, 2011; see also 'Attitudes, Political: Trust,' 'Attitudes, Political and Social Capital,' 'Politics and Participation,' and 'Politics, Protest and Civil Disobedience' in this Section);
2. inter-cultural change and difference, involving consensus but especially conflict between, for example, European- and non-European-based, such as Muslim, cultures (see, e.g., Sniderman and Hagendoorn, 2007);
3. the roles of culture and of cultural change in the sequence of 'human development' involving economic development, culture, behavior, and democratic institutions (Inglehart and Welzel, 2005; see also 'Modernization' in this Section);
4. the theorizing and modeling of relationships involving the causal ordering of beliefs-cultures-preferences and of institutions-regimes (e.g., Inglehart and Welzel, 2005; see also 'Democratic Transitions' and 'State Formation' in this Section) that are capable of addressing, for example, "the fundamental human problem" of "how can a society construct, or preserve, institutions that make self-interested individual action not inconsistent with the welfare of others?" (Munger, 2000: 25);
5. the design, collection, and methodological analysis of multiple country, respondent and time-point datasets capable of implementing the logic and the conduct of comparative inquiry that can help, in part, the research above (see, e.g., the European Social Surveys; the World Values Surveys; Davidov, Schmidt and Billiet, 2011).

See also: Democracy: Normative Theory; Democratic Theory; Lazarsfeld, Paul Felix (1901–76); Politics and Participation; Schumpeter, Joseph A. (1883–1950); Trust, Sociology of.

Bibliography

Almond, G.A., Verba, S., 1963. The Civic Culture: Political Attitudes and Democracy in Five Nations. Princeton University Press, Princeton, NJ. Reprinted in 1965 by Boston, Little Brown & Co., and in 1989 by Sage Publications Newbury Park, CA.

Almond, G.A., Verba, S. (Eds.), 1980. The Civic Culture Revisited. Little Brown & Co, Boston. Reprinted in 1989 Sage Publications Newbury Park, CA.

Berelson, B., Lazarsfeld, P., McPhee, W., 1954. Voting: A Study of Opinion Formation in a Political Campaign. University of Chicago Press, Chicago. Chap.14.

Dahl, R.A., 1956. In: A Preface to Democratic Theory. University of Chicago Press, Chicago, p. 145.

Dahl, R.A., 1970. Polyarchy: Participation and Opposition. Yale University Press, New Haven, CT.

Dahl, R.A., 1989. In: Democracy and Its Critics. Yale University Press, New Haven, CT, p. 338 ff.

Dalton, Russell J., 2014. Citizen Politics: Public Opinion and Political Parties in Advanced Industrial Democracies, sixth ed. Sage Publications/CQ Press, Los Angeles.

Dalton, Russell, Welzel, Christian (Eds.), 2014. The Civic Culture Transformed: From Allegiant to Assertive Citizens. Cambridge University Press, Cambridge.

Davidov, Eldad, Schmidt, Peter, Billiet, Jaak (Eds.), 2011. Cross-Cultural Analysis: Methods and Applications. Routledge/Taylor & Francis Group, New York.

Friedman, J. (Ed.), 1999. Special Issue: Public ignorance and democratic theory. Critical Review; An Interdisciplinary Journal of Polilitics and Society 12 (4).

Inglehart, Ronald, Welzel, Christian, 2005. Modernization, Cultural Change, and Democracy: The Human Development Sequence. Cambridge University Press, New York.

Munger, Michael C., 2000. Political Science and Fundamental Research. PS (March): 25–30.

Norris, Pippa, 2011. Democratic Deficit: Critical Citizens Revisited. Cambridge University Press, Cambridge.

Schumpeter, J.A., 1947. Capitalism, Socialism, and Democracy. Harper and Brothers Chap 21 and 22, New York.

Shils, E., 1960. In: Political Development in The New States. Mouton, The Hague, The Netherlands, p. 387 ff.

Sniderman, Paul M., Hagendoorn, Louk, 2007. When Ways of Life Collide: Multiculturalism and Its Discontents in the Netherlands. Princeton University Press, Princeton.

Verba, Sidney, 2011. A Life in Political Science. Annual Review of Political Science 14, i–xv.

Attitudes, Political and Public Opinion

Larry M Bartels, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12560–12563, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Scholars of public opinion attempt to describe and account for the politically relevant preferences and beliefs of ordinary citizens, and to assess the political impact of those preferences and beliefs. The development and proliferation of systematic opinion surveys transformed the study of public opinion in the last half of the twentieth century. Analysts have used survey data to examine the social bases, structure, and sophistication of political opinions in elaborate detail, often with an eye toward evaluating the democratic competence of citizens. They have generally been more successful in characterizing and accounting for public opinion than in specifying its role in the broader political process. Efforts to specify the political causes and consequences of public opinion will no doubt be facilitated by the continuing accumulation of survey data over time and space, but will also require scholars of public opinion to make more ingenious and eclectic use of other relevant evidence in combination with survey data.

Scope and Methods

Public opinion has been usefully defined by V.O. Key Jr. as "those opinions held by private persons which governments find it prudent to heed" (Key, 1961). Scholars of public opinion attempt to describe and account for the politically relevant preferences and beliefs of ordinary citizens, and to assess the political impact of those preferences and beliefs. They investigate a wide range of subject matter, including broad ideologies, specific policy preferences, evaluations of political parties and candidates, assessments of self-interest and collective interest, attributions of political responsibility, and attitudes toward social groups and political institutions (Kinder, 1998).

The study of public opinion draws upon and overlaps the study of political psychology, but tends to focus on collective preferences and beliefs in their broad political context rather than on individual mental processes (*see* Opinion Formation, Theory of). For example, the most notable work in the field begins with a psychological account of "how citizens acquire information and convert it into public opinion," but "pays vastly more attention to the social sources of mass attitudes – in particular, the availability of information in elite discourse – than to the largely autonomous operation of people's minds and psyches" (Zaller, 1992: p. 3).

Obviously, which opinions held by private persons governments will find it prudent to heed depends significantly upon the political setting. It should not be surprising that systematic attention to the workings of public opinion has been most common in modern liberal democracies, where the preferences and beliefs of ordinary citizens are supposed to be broadly and routinely consequential, whether through elections of public officials, voting in referenda, interest group activities, or other mechanisms. By the same logic, the proliferation of new democracies in Eastern Europe, Latin America, and Asia since the 1980s may be expected to stimulate unprecedented scholarly attention to public opinion in those parts of the world.

Contemporary studies of public opinion are profoundly shaped by insights and concerns stemming from the rise of mass politics in the nineteenth century and from the intertwined evolution of modern democratic states and societies. For example, Zaller's (1992) influential model of the dynamics of opinion change builds upon Bryce's (1888) distinction between the 'active class' of political entrepreneurs and the 'passive class' of ordinary citizens who have 'been told what to think, and why to think it,' and upon Lippmann's (1922) analysis of the role of the mass media and political elites (including government bureaucrats) in shaping 'the pictures in our head' that govern our reactions to complex and distant public affairs.

While long-standing insights and concerns are readily recognizable in modern scholarship on public opinion, empirical research in the field has been transformed since the 1940s by the development and proliferation of increasingly ambitious and sophisticated opinion surveys. Semistructured in-depth interviews, focus groups, and laboratory experiments continue to generate valuable insights, but their relatively small scale and problems of external validity make them better suited to the study of individual political psychology than of collective public sentiment. By comparison, an opinion survey with several hundred randomly selected respondents can provide a reasonably accurate representation of the preferences and beliefs of millions of ordinary citizens – public opinion on a scale which governments may well find it prudent to heed.

Pioneering Analysis

The pioneering analyses of public opinion based upon survey data were conducted in connection with US presidential elections of the 1940s (Berelson et al., 1954) and 1950s (Campbell et al., 1960). Berelson et al. concluded that election campaigns mostly activate pre-existing political preferences rooted in ethnic, sectional, class, and family traditions and reinforced by social interactions with like-minded

acquaintances (*see* Campaigning: Political; Voting and Direct Democracy). Campbell et al. emphasized the central importance of long-standing psychological attachments to the political parties (*see* Party Identification and Nonpartisanship), while also acknowledging the significance of such short-term forces as Dwight Eisenhower's personal popularity, which overcame but did not supplant the Democratic advantage in party identification dating from the New Deal era. Both of these studies portrayed public opinion as surprisingly thin and unsystematic, arguing, for example, that "many people know the existence of few if any of the major issues of policy" (Campbell et al., 1960: p. 170) and that political preferences "are characterized more by faith than by conviction and by wishful expectation rather than careful prediction of consequences" (Berelson et al., 1954: p. 311).

These uncomplimentary characterizations of public opinion were refined and elaborated in what remains the single most influential (and controversial) work in the field, Philip Converse's essay on "The nature of belief systems in mass publics" (Converse, 1964). Converse's analysis was based on three primary sorts of evidence. First, he showed that survey respondents answering open-ended questions only rarely made spontaneous use of systematic ideological concepts or language. Second, he showed that respondents' political views were only weakly constrained by conventional notions of 'what goes with what': liberals on one issue often took conservative positions on other issues, even within the same policy domain. Third, repeated interviews with the same respondents demonstrated a great deal of instability in their opinions, and the pattern of that instability suggested that it was caused more by random fluctuation than by systematic responses to political events. On the basis of this evidence, Converse (1964: p. 245) concluded that "large portions of an electorate do not have meaningful beliefs, even on issues that have formed the basis for intense political controversy among elites for substantial periods of time."

Civic Competence

The civic competence of ordinary citizens has remained a central issue of scholarly debate in the decades since Converse's devastating analysis. One line of revisionist argument accepted the accuracy of Converse's characterization of public opinion in the USA of the 1950s, but denied its relevance to other times and places. Most notably, Nie et al. (1976) and others argued that the political upheavals of the 1960s had produced a 'new American voter' more sophisticated, more ideological, and more attuned to political issues than the one portrayed by Campbell et al. (1960) and Converse (1964). While this thesis has been subjected to vigorous methodological critiques – and has come with the passage of time to seem at least as time-bound in its own right as the portrayal it aimed to supplant – it focused attention on the potential responsiveness of public opinion to such powerful contextual factors as social change and the nature of political rhetoric.

A newer wave of revisionist scholarship has acknowledged that most citizens in most circumstances are relatively inattentive to politics and uninformed about public affairs, but has interpreted these facts as evidence of "rational ignorance" in Anthony Downs's sense of avoiding costly information gathering in a realm in which individual investment can have only negligible collective consequences. This interpretation is nicely conveyed by the titles of three notable works of the early 1990s: *The Reasoning Voter* (Popkin, 1991), *Reasoning and Choice* (Sniderman et al., 1991), and *The Rational Public* (Page and Shapiro, 1992). These and other works argued that ordinary citizens make sense of the political world despite their lack of detailed information about ideologies, policies, and candidates. In some cases, the emphasis is on how citizens use 'information shortcuts' and 'gut reasoning' (Popkin, 1991) to deal efficiently with political complexity. In others, the emphasis is on the tendency of statistical aggregation to iron out individual errors and idiosyncrasies in mass publics, building upon an argument first set forth by Condorcet in the eighteenth century. This 'extenuationist' literature (to use a label coined by Robert Luskin) has in turn inspired attempts to measure the extent to which information shortcuts and aggregation actually do mitigate the political effects of 'rational ignorance.' For example, Bartels (1996) estimated that individual vote choices in a series of US presidential elections fell about half way between random behavior and hypothetical 'fully informed' choices (imputed to each voter by observing the choices of better-informed voters with similar demographic and social characteristics). Deviations from 'fully informed' behavior at the individual level were mitigated but not eliminated at the aggregate level: on average, Democrats did almost 2% points better and incumbent Presidents did almost 5% points better than they would have if all voters were, in fact, 'fully informed.'

Normative concerns about the quality of public opinion have also led to the development by James Fishkin and others of 'deliberative opinion polls' designed to combine the efficiency and representativeness of conventional opinion surveys with the informative and deliberative aspects of a seminar or town meeting. Rather than simply reporting their existing opinions, respondents are brought together with policy experts, public officials, moderators, and each other to consider political issues in depth. The aim is to find out – and then to publicize – "what the public *would* think, if it had a more adequate chance to think about the questions at issue" (Fishkin, 1991: p. 1).

Political Consequences of Public Opinion

Political scientists have been more successful in characterizing and accounting for public opinion than in specifying its role in the broader political process. The path-breaking study by Miller and Stokes (1963) of representation in the US Congress provided a model for much subsequent research examining the relationship between public opinion and elite attitudes and behavior. However, the basic research design employed in these studies – cross-sectional analysis of the impact of constituents' opinions on the opinions or behavior of individual legislators – is arguably ill-suited to answering questions about the political impact of public opinion in the aggregate, especially in legislative systems marked by strong party discipline or proportional representation, or when crucial policy decisions are made by executives or bureaucrats rather than by legislatures.

An alternative research design focuses on covariation between aggregate public opinion and policy outcomes across political units or in a single political unit over time. The analysis of 'dynamic representation' by Stimson et al. (1995) is an ambitious example of this approach, relating variations in the liberalism of America's 'public mood' over a 35-year period to subsequent variations in the behavior of Presidents, Congresses, and courts. A more specialized literature treats fluctuations in presidential popularity not only as a product of political, economic, and international events (as in the pioneering work of Mueller (1973)), but also as a factor in accounting for the varying ability of Presidents to get things done.

Efforts to specify the broad political causes and consequences of public opinion will no doubt be facilitated by the continuing accumulation of survey data over time and space. For example, the cumulative data file of the American National Election Studies project now includes more than 40 000 individual survey respondents in more than two dozen national elections spanning half a century, while the Eurobarometer series includes biannual surveys since 1973 in each of the member states of the European Union. By merging these enormous survey data sets with aggregate-level data reflecting variations in national economic conditions, electoral laws, party platforms and campaign appeals, policy initiatives, media coverage of public affairs, and many other contextual factors, analysts should be able to illuminate important connections between mass opinion and elite behavior, as well as the role of social and political institutions in shaping those connections.

On the other hand, the availability of richly detailed survey data has probably distracted scholarly attention from crucial aspects of public opinion that are not readily operationalized in surveys. For example, Converse's (1964: pp. 245–246) 'issue publics' – the various minorities of the public whose 'activated' opinions on particular issues of interest are 'expressed in the writing of letters to the editor, the changing of votes, and the like' – have seldom been subjected to detailed scrutiny. Even more notably, the elusiveness of what Key (1961) referred to as 'latent opinion' – the preferences and beliefs politicians *expect* to confront in future elections, after today's policies have had their effects and tomorrow's political opponents have had their say – has discouraged political scientists from giving it attention commensurate with its undoubted political significance. In these cases, among others, scholars will have to be more ingenious and eclectic in collecting and analyzing relevant data if they are to place the study of public opinion more squarely in the mainstream of political science.

Editors' Note

Since the publication of Bartels' essay above, much important research on public opinion has been conducted and published (see Clawson and Oxley, 2013; Dalton, 2014; and many articles in *Public Opinion Quarterly* and in the *International Journal of Public Opinion Research*). Some of this research involves:

1. the definition (nonopinions/opinions), measurement, survey designs (including question construction, questionnaire design, and survey-error reduction), and content analyses of online communications involving public opinion (see, e.g., Berinsky, 2012, forthcoming; Saris and Sniderman, 2004; Weisberg, 2005);
2. the correlates, distributions, and sources of public opinion:
   a. ranging from 'domestic,' such as economic and social (e.g., ethnicity-race and gender) issues through foreign event-policy issues (Berinsky, 2009, 2012; Burns and Gallagher, 2010),
   b. and involving sociodemographic groups; other attitudes, beliefs, and values (such as ideology or party identification); full information or cue use as well as knowledge and sophistication; personality traits; and various media uses (see, e.g., Berinsky, 2012; Saris and Sniderman, 2004; see also Mass Media and Politics);
3. the connections between public opinion and public policy, notably the 'thermostatic' model (Soroka and Wlezien, 2010; see also Berinsky, 2012; Political Representation);
4. dataset constructions and methodological developments capable of analyzing individual- and aggregate-level change and stability of public opinions and of policy connections (see, e.g., Hagenaars and McCutcheon, 2002; Saris and Sniderman, 2004; as well as Political Science and Polling; Pooled Cross-Sectional and Time Series Analyses, in Political Science.

See also: Attitude Measurement; Attitudes and Behavior; Issue Evolution; Mass Media and Politics; Opinion Formation, Theory of; Party Identification and Nonpartisanship; Political Advertising; Political Communication; Political Representation; Political Science and Polling; Voting and Direct Democracy.

Bibliography

Bartels, L.M., 1996. Uninformed votes: information effects in presidential elections. American Journal of Political Science 40, 194–230.
Berelson, B.R., Lazarsfeld, P.F., McPhee, W.N., 1954. Voting: A Study of Opinion Formation in a Presidential Campaign. University of Chicago Press, Chicago.
Berinsky, Adam J. Conceptualizing public opinion. Annual Review of Political Science 18, forthcoming.
Berinsky, Adam J., 2009. In: Time of War: Understanding American Public Opinion from World War II to Iraq. University of Chicago Press, Chicago.
Berinsky, Adam J. (Ed.), 2012. New Directions in Public Opinion. Routledge/Taylor & Francis Group, New York.
Bryce, J., 1888. The American Commonwealth. Macmillan, New York.
Burns, Nancy, Gallagher, Katherine, 2010. Public opinion on gender issues: the politics of equity and roles. Annual Review of Political Science 13, 425–443.
Campbell, A., Converse, P.E., Miller, W.E., Stokes, D.E., 1960. The American Voter. Wiley, New York.
Clawson, Rosalee A., Oxley, Zoe M., 2013. Public Opinion: Democratic Ideals, Democratic Practice, second ed. Sage Publications/CQ Press, Los Angeles.
Converse, P.E., 1964. The nature of belief systems in mass publics. In: Apter, D.E. (Ed.), Ideology and Discontent. Free Press of Glencoe, Glencoe, IL.
Dalton, Russell J., 2014. Citizen Politics: Public Opinion and Political Parties in Advanced Industrial Democracies, sixth ed. Sage Publications/CQ Press, Los Angeles.
Fishkin, J.S., 1991. Democracy and Deliberation: New Directions for Democratic Reform. Yale University Press, New Haven, CT.
Hagenaars, Jacques A., McCutcheon, Allan L. (Eds.), 2002. Applied Latent Class Analysis. Cambridge University Press, Cambridge.
Key, V.O., 1961. Public Opinion and American Democracy, first ed. Knopf, New York.

Kinder, D.R., 1998. Opinion and action in the realm of politics. In: Gilbert, D., Fiske, S., Lindsey, G. (Eds.), Handbook of Social Psychology, fourth ed. McGraw-Hill, Boston.
Lippmann, W., 1922. Public Opinion. Macmillan, New York.
Miller, W.E., Stokes, D.E., 1963. Constituency influence in congress. American Political Science Review 57, 45–56.
Mueller, J.E., 1973. War, Presidents and Public Opinion. Wiley, New York.
Nie, N.H., Verba, S., Petrocik, J., 1976. The New American Voter. Harvard University Press, Cambridge, MA.
Page, B.I., Shapiro, R.Y., 1992. The Rational Public: Fifty Years of Trends in Americans' Policy Preferences. University of Chicago Press, Chicago.
Popkin, S.L., 1991. The Reasoning Voter: Communication and Persuasion in Presidential Campaigns. University of Chicago Press, Chicago.
Saris, Willem E., Sniderman, Paul M. (Eds.), 2004. Studies in Public Opinion: Attitudes, Nonattitudes, Measurement Error, and Change. Princeton University Press, Princeton.
Sniderman, P.M., Brody, R.A., Tetlock, P.E., 1991. Reasoning and Choice: Explorations in Political Psychology. Cambridge University Press, New York.
Soroka, Stuart N., Wlezien, Christopher, 2010. Degrees of Democracy: Politics, Public Opinion and Policy. Cambridge University Press, Cambridge.
Stimson, J.A., MacKuen, M.B., Erikson, R.S., 1995. Dynamic representation. American Political Science Review 89, 543–565.
Weisberg, Herbert F., 2005. The Total Survey Error Approach: A Guide to the New Science of Survey Research. University of Chicago Press, Chicago.
Zaller, J.R., 1992. The Nature and Origins of Mass Opinion. Cambridge University Press, New York.

Attitudes, Political: Authoritarianism and Tolerance

Stanley Feldman, State University of New York at Stony Brook, Stony Brook, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article S. Feldman, volume 2, pp. 970–972, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The construct of authoritarianism is one of the major attempts in political psychology to understand individual differences in intolerance and prejudice and susceptibility to fascist appeals. The publication of *The Authoritarian Personality* by Adorno, Frankel-Brunswick, Levinson, and Sanford in 1950 initiated a large body of empirical research on the causes and consequences of authoritarianism. The construct has been severely criticized since its inception on both theoretical and empirical grounds. Recent work in this area has made significant progress on some of the empirical issues but substantial theoretical problems remain.

The construct of authoritarianism grew out of attempts by psychologists to understand the mass appeal of fascism in the first half of the twentieth century. As a result of a large body of empirical research, authoritarianism has become one of the major psychological explanations for the persisting finding of generalized prejudice and intolerance: the tendency for negative attitudes toward out-groups to correlate across different targets. It remains one of the most important approaches in the political psychology literature.

Origins

Research on authoritarianism was influenced by the attempts of European psychologists in the psychoanalytic tradition to understand mass support for fascism (Fromm, 1941; Reich, 1946). The major defining work was the publication in 1950 of *The Authoritarian Personality* by Adorno et al. (1950). This book of almost 1000 pages combined psychoanalytic and personality theory with empirical research from in-depth interviews and questionnaires. While the project began as a study of antisemitism, the researchers quickly discovered that antisemitism appeared to be a part of a broader ethnocentrism. They located the origins of this ethnocentrism in basic personality processes as seen from a Freudian perspective.

Adorno et al. argued that punitive child-rearing practices and inconsistent parental affection produce a severe conflict within the young child. The repression of hostility toward the parents is, in adulthood, displaced onto groups perceived to be weak or inferior while authority figures (a generalization from the parents) are deferred to. Another consequence of this psychological conflict is a weak ego, which in turn leads to characteristics such as an unusual adherence to conventional social values, intolerance of ambiguity, and attraction to superstitions. From this perspective, prejudice has little to do with the actions or attributes of the target group but is a consequence of the psychological needs of those high in authoritarianism to direct repressed hostility toward subordinate groups.

Along with their explanation, Adorno et al. developed a paper and pencil measure of authoritarianism. It was labeled the F-scale because it was considered to be an indicator of susceptibility to fascist appeals. Introduced just as quantitative social science research was gaining momentum, an enormous literature soon developed that examined the correlates of authoritarianism, as measured by the F-scale or similar measure. As of 1989 there were well over 2000 publications on authoritarianism and related constructs (Meloen, 1993). Studies cataloged the relationship between the F-scale and social background characteristics, various measures of personality, and a wide range of social and political attitudes (see Kirscht and Dillehay, 1967 for a typical review of this research).

Criticisms

The Authoritarian Personality was subjected to close scrutiny and criticism almost immediately after it was published (see Christie and Jahoda, 1954). Methodologically, problems with the F-scale measure of authoritarianism cast doubt on much of the empirical research. Since all of the questions in this scale were worded so that an agree response indicated authoritarian characteristics, a tendency to agree with questions like this regardless of content (agreement response set) could produce high authoritarianism scores independent of any individual differences in personality. And while the questions in the F-scale were developed to tap the various syndromes hypothesized to make up authoritarianism, it is not clear that the measure reflects the underlying conceptualization. Factor analyses of the F-scale always show the measure to be multidimensional but the factors are always difficult to interpret and never cleanly correspond to the hypothesized components (see Altemeyer, 1981).

Critics also argued that the authoritarianism measure, and possibly the construct itself, was ideologically biased. Although Adorno et al. were clearly responding to the history of fascism in Europe, the growing threat of communism in the 1950s led to concerns about an authoritarianism of the left (Shills, 1954). However, despite observations that communists appeared to possess traits characteristic of authoritarians, there is little empirical evidence to support these claims (Stone, 1980).

Perhaps most importantly, the Freudian-based theory underlying *The Authoritarian Personality* proved far less useful

than the authors thought and as a result research become increasingly divorced from the theory. In addition to the general loss of status of Freudian explanations in empirical social science research, Adorno et al.'s specific explanation of authoritarianism has not fared well. Although quantitative research established relationships between the F-scale, social background characteristics, and social attitudes, evidence on the dynamics and origins of authoritarianism consistent with the Adorno et al. theory proved elusive (see Altemeyer, 1981; 1988; Duckitt, 1989). The bulk of research based on *The Authoritarian Personality* thus consists of short empirical studies with only limited guidance from the original theory. Most of these studies report correlations between measures of authoritarianism and a small number of other individual level characteristics: social background, personality, values, and attitudes.

More Recent Approaches

Several attempts have been made since the publication of *The Authoritarian Personality* to develop alternative explanations for individual differences in prejudice and intolerance. Perhaps the best known of these is Rokeach's research on dogmatism reported in *The Open and Closed Mind* (1960). Rokeach moved from the Freudian perspective of Adorno et al. to a more cognitive explanation based on the structure and function of belief systems. He argued that prejudice and intolerance are a result of closed belief systems that ward off threats by limiting the scope of accepted beliefs and rejecting all challenges. Like Adorno et al., Rokeach developed a measure of dogmatism based on the same methodology of agree–disagree attitude statements. His dogmatism measure was also constructed of a diverse range of statements designed to reflect the various elements of open and closed belief systems.

While research has utilized Rokeach's dogmatism scale, most of it, like research using the F-scale, consists of narrow empirical studies that examine the correlates of the measure. There has been little reported work that has examined Rokeach's conceptualization of open and closed belief systems despite his own detailed and novel explorations reported in the book. In addition, the measurement properties of the dogmatism measure have proved to be no improvement on the F-scale. It too has items that are worded in only the agree direction and the interitem correlations and dimensionality of the measure may be even worse than the F-scale (Altemeyer, 1981).

The most significant new contribution to the authoritarianism literature has been a recent body of research reported by Altemeyer (1981; 1988; 1996). From years of studies using undergraduate students, Altemeyer has both developed a new and more reliable measure of the construct and advanced a new conceptualization based on social learning theory. Rejecting the Adorno et al.'s Freudian explanation, Altemeyer (1988) prefers a simpler conceptualization that sees authoritarianism as a social attitude (or cluster of attitudes) that is learned through interactions with parents, peers, schools, the media, and through experiences with people who hold conventional and unconventional beliefs and lifestyles. His measure of right-wing authoritarianism (RWA) is more reliable and unidimensional than previous scales and has items balanced for agreement response set.

There have been several other attempts to reconceptualize the authoritarianism phenomenon. Kelman and Barkley (1963) argued that high authoritarianism reflects a limited breadth of perspective. Lipset (1960), Selznick and Steinberg (1969), and others have suggested that authoritarianism is a function of social learning associated with lower social class and less educated environments. More recently, Duckitt (1989) has developed a model in which authoritarianism reflects the desire to maintain strong in-group identifications.

Limitations and Problems

In addition to the measurement and theoretical problems discussed above, there are other significant shortcomings of the authoritarianism approach.

How do factors in the social environment exacerbate or moderate predispositions toward prejudice and intolerance? Some researchers have hypothesized that social threats or anxiety will arouse prejudice and punitiveness and encourage support for extreme right-wing parties and even mass violence (Staub, 1989). There is also some evidence from archival and individual level studies that levels of authoritarianism, or the strength of the connection between authoritarianism, prejudice, and intolerance, increase during periods of threat (Sales, 1973; Doty et al., 1991; Feldman and Stenner, 1997). Although there appears to be strong impressionistic evidence that prejudice, intolerance, and support for right-wing groups wax and wane over time, there is nothing obvious in the present conceptualizations of authoritarianism that would help to explain this phenomenon.

More generally, authoritarianism is generally treated as a purely dispositional construct; researchers almost always examine simple additive relationships between measures of authoritarianism and indicators of prejudice and intolerance. This type of approach provides no understanding of how the reality of social and political conflicts – or any other environmental factors – interact with authoritarian predispositions. This leaves authoritarianism a purely 'personality' explanation, with only weak connections to other explanations of prejudice and intolerance.

It is also necessary to make progress on the relationship between authoritarianism and ideology, particularly since the authoritarianism construct is beginning to be used in studies of countries with a history of communist governments. Studies in the USA and Canada consistently find negative relationships between measures of authoritarianism and attitudes toward communism. In Russia, however, researchers have shown that 'right-wing authoritarianism' has been associated with support for communism (McFarland et al., 1992). This relationship can be understood if support for communism in countries like this represents a commitment to the established, authoritarian order. Even so, a positive correlation between 'right-wing' values and communism strongly suggests that the construct of authoritarianism must be defined and measured in a way that allows it to be used in cross-national research.

Conclusions

Despite all of the problems that have been raised with the authoritarianism construct since the 1950s, it continues to be used in research on prejudice and intolerance. Its longevity is, in part, a function of the repeated findings of consistency in expressions of prejudice and intolerance across a wide range of targets. Individual differences account for a significant proportion of the variance in these attitudes and authoritarianism remains one of the major explanations of those individual differences.

Editors' Note

"Authoritarianism" has been defined in at least two major ways in political and social science research. According to the "institutional" definition, authoritarianism is, in particular, a type of government or regime whose few institutions and their incumbents can make and enforce rules that others must obey, and it is, more generally, a context in which power is concentrated and only slightly or completely unrestrained by other actors or organizations (see also Dictatorship). According to the "attitudinal-value-behavioral" definition, authoritarianism is a set of attitudes, beliefs, values and behaviors that are associated with personality characteristics and other factors, and whose origins, measurement, explanations, and problems in research are well discussed by Feldman above.

In recent years, this latter research has developed in important ways (for reviews, see, e.g., Clawson and Oxley, 2013; Gibson, 2008). Five ways involve:

1. the sources, notably socialization, other experiential, and perceived threat bases, distributions, and consequences of authoritarian attitudes, beliefs, values and behaviors (see, e.g., Stenner, 2005; Feldman et al., 2010);
2. the trends in relationships between authoritarianism and (in)tolerance, and in (in)tolerance itself, which may be defined as an individual's recognition that other people have lifestyles and views that differ from her/his own and that they have the right to public expression of them. Several of these studies have questioned whether tolerance of people in different ethnic-racial, including Muslim, and other social groups, and of people with different political or social views has increased, rather than decreased, over time in various countries (see, e.g., Gibson, 2008; Sniderman et al., 2000; Sniderman and Hagendoorn, 2007);
3. the theoretical, conceptual and empirical connections among authoritarianism/(in)tolerance, ideology (see, e.g., Feldman, 2015; see also Ideological Constraint: History and Current Status of the Concept), personality traits, and/or value constructs (see e.g. Cohrs et al., 2005). These personality traits include the "Big Five" of agreeableness, conscientiousness, extroversion, neuroticism, and openness (see, e.g., Mondak, 2010; Gerber et al., 2011). The value constructs refer to seven basic ones of hedonism, power/achievement, security, self-direction, stimulation, tradition/conformity, and universalism/benevolence (see Davidov, Schmidt and Schwartz, 2008; Davidov, Schmidt and Billiet, 2011);
4. the relationships between "institutional" authoritarianism and "attitudinal-value-behavioral" authoritarianism, that is, whether and, if so, then how and why the former precedes the latter and/or vice-versa (see, e.g., Attitudes, Political and Civic Culture; Democratic Transitions and State Formation, Theory of).
5. the conceptualization of authoritarianism within political psychology (see Kessler/Cohrs, 2008; Stellmacher/Petzel, 2005) and its relation to the dual process model in social psychology (see e.g. Duckitt et al., 2002).

See also: Authoritarian Personality; Authority, Social Theories of; Communism, History of; National Socialism and Fascism; Prejudice in Society: Psychological Perspectives; Prejudice in Society: Sociological Perspectives; Totalitarianism.

Bibliography

Adorno, T.W., Frankel-Brunswick, E., Levinson, D.J., Sanford, R.N., 1950. The Authoritarian Personality. Harper, New York.
Altemeyer, B., 1981. Right-Wing Authoritarianism. University of Manitoba Press, Winnipeg, MB.
Altemeyer, B., 1988. Enemies of Freedom. Jossey-Bass, San Francisco.
Altemeyer, B., 1996. The Authoritarian Specter. Harvard University Press, Cambridge, MA.
Christie, R., Jahoda, M. (Eds.), 1954. Studies in the Scope and Method of 'The Authoritarian Personality. Free Press, Glencoe, IL.
Clawson, Rosalee A., Oxley, Zoe M., 2013. Public Opinion: Democratic Ideals, Democratic Practice, second ed. Sage Publications/CQ Press, Los Angeles.
Cohrs, J.C., Moschner, B., Maes, J., Kielmann, S., 2005. The motivational bases of right-wing authoritarianism and social dominance orientation: relations to values and attitudes in the aftermath of September 11, 2001. Personality and Social Psychology Bulletin 31 (10), 1425–1434.
Davidov, Eldad, Schmidt, Peter, Billiet, Jaak (Eds.), 2011. Cross-Cultural Analysis: Methods and Applications. Routledge/Taylor & Francis Group, Hove, UK and New York, USA.
Davidov, Eldad, Schmidt, Peter, Schwartz, Shalom H., 2008. Bringing values back. In: The Adequacy of the European Social Survey to Measure Values in 20 Countries. Public Opinion Quarterly 72, 420–445.
Doty, R.M., Peterson, B.E., Winter, D.G., 1991. Threat and authoritarianism in the United States. Journal of Personality and Social Psychology 61, 629–640.
Duckitt, J., 1989. Authoritarianism and group identification. Political Psychology 10, 63–84.
Duckitt, J., Wagner, C., du Plessis, I., Birum, I., 2002. The psychological bases of ideology and prejudice: testing a dual process model. Journal of Personality and Social Psychology 83, 75–93.
Feldman, Stanley. 2015 forthcoming. 'Authoritarianism and Ideology,' In: Friedrich Funke, Thomas, Petzel, Christopher, Cohrs and John, Duckitt (Eds.), Perspectives on Authoritarianism. Berlin: Springer-Verlag.
Feldman, Stanley, Lavine, Howard, Lodge, Milton, Verhulst, Brad, 2010. Seeing Negative: Authoritarianism and Automatic Vigilance for Threatening Stimuli. In: paper presented at the Annual Meeting of the International Society of Political Psychology, San Franciso, CA.
Feldman, S., Stenner, K., 1997. Perceived threat and authoritarianism. Political Psychology 18, 741–770.
Fromm, E., 1941. Escape from Freedom. Rinehart, New York.
Gerber, Alan S., Huber, Gregory A., Doherty, David, Dowling, Conor M., 2011. The big five personality traits in the political arena. Annual Review of Political Science 14, 265–287.
Gibson, James L., 2008. Intolerance and political repression in the United states: a half century after McCarthyism. American Journal of Political Science 52, 96–108.
Kelman, H.C., Barclay, J., 1963. The F scale as a measure of breadth of perspective. Journal of Abnormal and Social Psychology 6, 608–615.

Kessler, T., Cohrs, J.C., 2008. The evolution of authoritarian processes: Fostering cooperation in large-scale groups. Group Dynamics: Theory, Research, and Practice 12, 73–84. http://dx.doi.org/10.1037/1089-2699.12.1.73.

Kirscht, J.P., Dillehay, R.C., 1967. Dimensions of Authoritarianism. University of Kentucky Press, Lexington, KY.

Lipset, S.M., 1960. Political Man. Doubleday, New York.

McFarland, S.G., Ageyev, V.S., Abalakina-Pap, M.A., 1992. Authoritarianism in the former Soviet Union. Journal of Personality and Social Psychology 63, 1004–1010.

Meloen, J.D., 1993. The F scale as a predictor of fascism. In: Stone, W.F., Lederer, G., Christie, R. (Eds.), Strength and Weakness: The Authoritarian Personality Today. Springer-Verlag, New York, pp. 47–69.

Mondak, Jeffery, 2010. Personality and the Foundations of Political Behavior. Cambridge University Press, Cambridge.

Reich, W., 1946. The Mass Psychology of Fascism. Orgone Press, New York.

Rokeach, M., 1960. The Open and Closed Mind. Basic Books, New York.

Sales, S.M., 1973. Threat as a factor in authoritarianism. Journal of Personality and Social Psychology 28, 44–57.

Selznick, G.J., Steinberg, S., 1969. The Tenacity of Prejudice. Harper & Row, New York.

Sniderman, Paul M., Hagendoorn, Louk, 2007. When Ways of Life Collide: Multiculturalism and Its Discontents in the Netherlands. Princeton University Press, Princeton.

Sniderman, Paul M., Peri, Pierangelo, de Figueiredo Jr., Rui J.P., Piazza, Thomas, 2000. The Outsider: Prejudice and Politics in Italy. Princeton University Press, Princeton.

Shills, E.A., 1954. Authoritarianism: right and left. In: Christie, R., Jahoda, M. (Eds.), Studies in the Scope and Method of 'The Authoritarian Personality. Free Press, Glencoe, IL, pp. 24–49.

Staub, E., 1989. The Roots of Evil. Cambridge University Press, Cambridge, UK.

Stellmacher, J., Petzel, T., 2005. Authoritarianism as a group phenomenon. Political Psychology 26, 245–274.

Stenner, Karen, 2005. The Authoritarian Dynamic. Cambridge University Press, New York.

Stone, W.F., 1980. The myth of left-wing authoritarianism. Political Psychology 2, 3–19.

Attributional Processes: Psychological

BF Malle, University of Oregon, Eugene, OR, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 2, pp. 913–917, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Attribution is primarily the process of ascribing a cause to an event or explaining the event. Secondarily, the term 'attribution' refers to certain inferences or judgments, such as inferences of personality traits from behavior and judgments of responsibility and blame. Psychological research on attribution was inspired by Heider (1958), who argued that the social perceiver makes sense of the world by attributing behaviors and events to their underlying causes. The rise of empirical attribution research in the 1960s and 1970s was sparked by a focusing and narrowing of Heider's broad model of social perception, generating two main strands of research: attribution as trait inference and attribution as causal explanation. The first strand began with Jones and Davis (1965). They took Heider's idea that people infer intentions from behavior a step further by developing a model of 'correspondent inferences' from a given intentional behavior ('he acted friendly') not just to a corresponding intention but to a corresponding trait ('he is friendly'). The second strand began with Kelley (1967) who developed Heider's insight that causal judgments are pivotal in social perception. Kelley proposed that such judgments are based on a simple information-processing rule: people infer those causes that covary with the event in question. Kelley's covariation theory profoundly influenced attribution research by assuming that (1) people break down causes into internal attributions (something about the agent) and external attributions (something about the situation) and (2) that the internal–external dichotomy generally applies to all behaviors and events alike. Recent theoretical developments emphasize, by contrast, that ordinary people sharply distinguish between (1) attributing *intentional actions* to the actor's reasons and motives and (2) attributing *unintentional behaviors* (e.g., failure, depression) to causal factors internal or external to the agent. This approach considers behavior explanations to be an integral part of people's theory of mind, which views people as intentional agents who act in light of their beliefs and desires. One advantage of this approach lies in its compatibility with work on responsibility attribution, which also takes into account the intentionality of the behavior in question. Another advantage lies in the integration of social-psychological work on explanations with developmental research on children's emerging theory of mind as well as with philosophical analysis of the properties and evolution of folk psychology. A further development suggests that attributions are not just cognitive processes but rather communicative acts. Thus, behavior explanations are altered for impression-management purposes, are responsive to an interlocutor's interest and concerns, and have an immediate impact on others' perceptions and evaluations of the explainer and the agent whose behavior is being explained.

Background

Psychological research on attribution was inspired by Heider (1958), who developed models of attribution for object perception as well as person perception. Heider assumed, following the philosophical tradition of phenomenology, that a perceiver faced with observational data tries to attribute these data to their underlying cause – to an object in the case of sensory data, and to characteristics of the agent or the environment in the case of human behavioral data.

Heider recognized, however, that person perception is more complex than object perception – due to the manifold observational data available and the manifold causes (e.g., beliefs, desires, emotions, traits) to which these data can be attributed. Heider believed that people master this complexity by means of a sophisticated conceptual framework that he labeled 'commonsense psychology' – nowadays also called 'folk psychology' or 'theory of mind.' Using this conceptual framework, perceivers bring order to the massive stream of behavioral data by classifying them and attributing them to mental categories – including motives, obligations, sentiments, and personality traits. For example, certain behaviors are classified as 'intentional actions,' and the perceiver looks for the specific intention underlying each action (what the agent was trying to do).

Heider's research method was philosophical. He relied on conceptual analysis complemented by linguistic examples and thought experiments, but he provided no empirical experiments to test his model. The impact of Heider's work, then, was largely mediated by other researchers, who turned some of his ideas into empirical research programs. The evolution of attribution research can therefore be seen as a selective exegesis of Heider – with different scholars picking up different ideas and developing them in more detail.

This survey begins with a brief sketch of Heider's charismatic work, then turns to the two main strands of post-Heiderian attribution research: trait attribution work and causal explanation work. Finally, recent theoretical developments are addressed that return to some of Heider's original concerns and link attribution research to developmental research on children's theory of mind (*see* Mind, Theories of) and philosophical work on folk psychology.

Heider

Despite his early focus on object perception, Heider's deepest concern was always with the topics of person perception and social interaction. Just like Lewin and Asch before him, Heider recognized that a psychology of social interaction must chart out people's *subjective* conceptions of the social world. At first Heider tried to develop a general theoretical framework to describe these conceptions, but then he realized that such a powerful framework already existed – it was commonsense

psychology, the system of concepts that ordinary people use to describe and understand human behavior. In his 1958 book, Heider thoroughly explored this framework, analyzing such commonsense concepts as *can*, *want*, *try*, and *ought*. Heider's blend of empirical observation with conceptual and linguistic analysis was visionary and unlike anything psychology had seen before – resembling more the tradition of ordinary language philosophy or contemporary cognitive linguistics.

Two major insights stand out from Heider's analysis of social perception. The first is that social perception is a process of reconstructing invariance out of variance. According to Heider, the social perceiver is presented with streams of behavior, and the cognitive system conceptualizes these streams in terms of a network of fundamental concepts, trying to reconstruct the invariances underlying behavior. Heider's notion of invariances encompassed perceptions, intentions, motives, abilities, and sentiments, which are all relatively invariant against the stream of ongoing behavior. Among these invariances, the agent's motives occupied a special role: "The underlying causes of events, especially the motives of other persons, are the invariances of the environment that are relevant to [the perceiver]; they give meaning to what he experiences" (Heider, 1958, p. 81). Heider's second major insight was that the reconstructive process of identifying invariances is a form of causal (i.e., attributional) analysis, rendering judgments of causality one of the primary elements in social perception. According to Heider, the commonsense conception of causality breaks down into two separate models. For one, there is *impersonal causality*, applied to unintentional human behaviors (such as sneezing or feeling pain) as well as physical events (such as stones rolling or leaves falling). For another, there is *personal causality*, applied only to human agents who perform an intentional action (such as cleaning the kitchen or inviting someone to dinner): "Personal causality refers to instances in which *p* causes *x* intentionally. That is to say, the action is purposive" (Heider, 1958, p. 100). For Heider, this distinction between unintentional and intentional behavior was central to commonsense psychology because it guides how people predict, evaluate, and influence social behavior.

In the literature, Heider's distinction between personal (intentional) and impersonal (unintentional) causality is typically portrayed as a distinction between internal (or person) causes and external (or situation) causes of behavior. But Heider did not subsume all internal attributions under the label 'personal causality'; only the attribution of purposive behavior to intentions and motives was characterized that way. The remaining internal causes fall under the 'impersonal' label, because they bring about behavior in a mechanical, nonintentional way. Thus, Heider drew a line between intentional and unintentional events, whereby intentional events are based on a special class of person causes (intentions and motives) and unintentional events can be based on person or situation causes.

Attribution Research after Heider

The rise of empirical attribution research in the 1960s and 1970s was sparked by a focusing and narrowing of Heider's broad model of social perception. Two main strands of research can be identified, each of which used Heider's ideas as a starting point and then developed them in more detail.

Attribution as Trait Inference

Heider theorized about the conditions under which people infer intentions from behavior. In their classic paper, Jones and Davis (1965) took this idea a step further by developing a model of 'correspondent inferences' from a given intentional behavior ('he acted friendly') not just to a corresponding intention ('he tried to be friendly') but to a corresponding trait ('he is friendly'). The model identified conditions that increase the likelihood of such correspondent trait inferences. For example, the social perceiver will be more likely to infer a trait the more the action in question deviates from what is normally seen as desirable, and the more unique (positive) effects it has compared to alternative courses of action. If none of these conditions hold, the perceiver will not infer a correspondent trait but instead attribute the agent's behavior to forces of the situation.

The initial model was limited to intentional behaviors and predicted that people do not infer traits from unintentional behaviors or from behaviors performed under strong constraints (e.g., upon request by an authority). However, subsequent research found that people infer traits even from unintentional behaviors such as mistakes and accidents (e.g., 'he's clumsy') and even when the agent acted under strong constraints. This tendency to make correspondent trait inferences, even from single behaviors, and the parallel reluctance to attribute behavior to situational forces, was later dubbed the 'correspondence bias' or 'fundamental attribution error' (Ross, 1977).

Researchers became increasingly interested in the question as to why people are so ready to infer traits from behavior and what conditions would moderate this bias. The currently dominant models of trait inference postulate a multistage process in which a behavior is identified as a certain type of action, a corresponding trait is then inferred by default, and only subsequent corrections take situational influences into account (Gilbert and Malone, 1995). However, this default trait inference does not occur, for example, when the situational influence is salient or when the perceiver suspects that the agent acted out of ulterior motives. A potential limitation of trait attribution work is the scarcity of studies that examine the frequency and conditions of trait inferences in naturalistic social contexts.

Attribution as Causal Judgment

The second major strand of attribution research was inspired by Kelley's (1967) model of causal attribution. Focusing on Heider's insight that causal judgments are pivotal in social perception, Kelley proposed that such judgments are based on a simple information-processing rule: people infer those causes that covary with the event in question. Specifically, when an agent *A* behaves toward object *O*, the cause of the behavior is perceived to lie in *A* (internal attribution) if few other people behave as *A* does (low consensus), if *A* behaves the same way toward *O* over time (high consistency), and if *A* behaves the same way toward other objects (low distinctiveness). By

contrast, the cause of the behavior is perceived to lie in *O* (external attribution) if most other people behave as *A* does (high consensus), if *A* behaves the same way toward *O* over time (high consistency), and if *A* behaves differently toward other objects (high distinctiveness). For Kelley, the principles of covariation applied equally to behaviors (whether intentional or unintentional) and physical events. His model was therefore embraced by causal reasoning researchers in both social and cognitive psychology.

Empirical tests of the covariation model tended to support it, although this support was limited to experimental settings in which explicitly presented covariation information had an effect on judgments. At the turn of the twenty-first century, no studies had demonstrated that people spontaneously, in everyday situations, seek out covariation information before answering a why question. Moreover, the covariation model makes no predictions about situations in which people lack covariation information (e.g., single observations) or in which they are not motivated to make use of the information (e.g., under time pressure). For those instances, Kelley later proposed additional rules of causal reasoning ('causal schemata'). One of them is the discounting principle, which states that, under certain circumstances, a second cause weakens the plausibility of a first cause. For example, if a student aces a difficult exam, we might explain it by assuming that he studied hard; but upon hearing that he cheated we may no longer believe that he studied hard, thereby discounting the previous explanation.

The problem that Kelley's causal attribution models tried to address was the 'causal selection problem' – how perceivers select particular causes for explaining a given behavior or event. Even though Kelley's models themselves did not completely solve this problem (alternative models were developed later), Kelley's work profoundly influenced attribution research by assuming that (1) people break down causes into internal ones (something about the agent) and external ones (something about the situation) and (2) the internal–external dichotomy generally applies to all behaviors and events alike. This dichotomy proved to be a compellingly simple dependent variable that allowed researchers to explore a variety of interesting phenomena. One of these phenomena is the self-serving bias in explanations (see Miller and Ross, 1975) – the tendency for people to explain their own positive and negative outcomes so as to maintain favorable self-perceptions or public impressions. For example, students would be expected to explain a good grade by citing internal causes (e.g., ability or hard work) and a bad grade by citing external causes (e.g., bad luck or an unreasonable teacher). Another important phenomenon is the actor–observer asymmetry, which is the tendency for people to explain their own behaviors and other people's behaviors in systematically different ways. Specifically, Jones and Nisbett (1972) argued, and later studies confirmed, that people tend to explain their own behaviors by reference to external factors (e.g., 'I chose psychology as my major because *it's interesting*') but explain other people's behavior by reference to internal attributes (e.g., 'He chose psychology as his major because *he wants to help people*').

Some researchers proposed alternatives to the internal–external dichotomy of causes. For example, Weiner introduced two additional distinctions – one between stable and unstable causes and one between controllable and uncontrollable causes – and thereby improved predictions for people's emotions and motivations in the wake of explaining achievement outcomes, health outcomes, and deviant social conduct (see Weiner, 1995). Abramson et al. (1978) introduced another distinction, between global and specific causes, to describe a hopeless explanatory style in depression (*see* Depression). However, most attribution work at the beginning of the twenty-first century still uses a single internal–external (or sometimes trait–situation) distinction.

The focus on internal versus external causal attributions to all behaviors and events alike can be contrasted with Heider's claim that ordinary people sharply distinguish between (1) attributing *intentional actions* to the actor's motivation and (2) attributing *unintentional behaviors* (e.g., failure, depression) to causal factors internal or external to the agent. Heider felt, as an interview in 1976 reveals, that people's attributions of events were adequately depicted in Weiner's model, but that attributions of actions to motives were inadequately treated by contemporary research (Ickes, 1976). A second limitation of contemporary research is that it developed primarily cognitive models of the attribution process, whereas Heider demanded careful attention to the communicative functions of attributions and the language of causality. Some recent theoretical developments have tried to address both of these potential limitations.

Recent Theoretical Developments

Many refinements have been proposed since these original conceptualizations and empirical studies. These have been reviewed by Gordon and Graham (2006) and Malle (2004).

Attribution as Folk Theory of Mind

Over the decades, attribution research has offered increasingly sophisticated information-processing models for the basic internal–external distinction, but it has not examined the distinctions that actually guide people's explanations within folk psychology – an omission that Kelley and Michela (1980) called the central irony of attribution research. The study of people's own concepts and distinctions has recently gained force among researchers who endorse what may be called a folk–theoretical approach to explanations. This approach considers behavior explanations an integral part of people's folk theory of mind and assigns the concept of *intentionality* a central role in these explanations, because people reliably distinguish between intentional and unintentional human behaviors and use distinct modes of explanation for each type of behavior (Buss, 1978; Malle, 1999).

In particular, unintentional behavior is explained by *causes*, perceived as simply bringing about the behavior in a mechanical way (e.g., 'She cried because *she was sad*'; 'He fell because *the floor was slippery*'). By contrast, intentional behavior is primarily explained by *reasons*, conceptualized as the beliefs and desires an agent considers when forming an intention to act (e.g., 'She quit her job because *she felt the pay was too low*'; 'They worked extra hours because *they needed to finish the project*'). Thus, when people offer reason explanations, they presume the agent's awareness of the reason

content and a more or less rational process of forming an intention to act in light of those reasons. According to the folk-theoretical approach, the traditional internal–external dichotomy applies well to people's cause explanations of unintentional behavior but not to reason explanations of intentional behavior. For example, the above reason explanation, 'she felt the pay was too low,' refers to the agent's own belief ('she felt'), but the belief itself has a content that refers to the situation ('the pay was too low'), so it is unclear how this reason should be classified within an internal–external scheme. To analyze adequately people's reason explanations of intentional behavior, new theoretical categories have been developed that take into account the complex conceptual and linguistic nature of reasons (Malle, 1999).

One advantage of the folk–theoretical approach lies in its compatibility with work on responsibility attribution. For example, Shaver (1985) argued that assignments of responsibility and blame are critically dependent on the intentionality of the behavior in question, and so the choice of explanations (e.g., causes vs reasons) often manages assignments of blame. Another advantage of this approach lies in the integration of social–psychological work on explanations with developmental research on children's emerging theory of mind and philosophical analysis of the properties and evolution of folk psychology (Malle et al., 2001). These disciplines have highlighted people's unique conception and explanation of intentional action, and the next generation of attribution research will have to flesh out in detail the conceptual, linguistic, and social nature of these unique explanations.

Attribution as Communication

Heider examined the concepts of commonsense psychology in their everyday conversational context because he was convinced that tools of social perception help people to accomplish their goals in social communication and interaction. Along these lines, innovative work by Hilton (1990) and Antaki (1994) suggests that attributions are not just cognitive processes but rather communicative acts that obey the rules of conversation; that is, behavior explanations are altered for impression-management purposes (e.g., to appear rational or fend off blame). Explanations are also responsive to an interlocutor's concerns and questions (e.g., 'How was this possible?' elicits different modes of explanation than 'For what reason?'). Finally, explanations have an immediate impact on others' perceptions and evaluations of the explainer and the agent whose behavior is being explained.

Future Directions

Promising new directions of attribution research may arise from a merging of the folk–theoretical and communicative approaches. By recognizing people's fine-grained theory of mind and their distinctions between various modes of explanation, researchers can more effectively trace the cognitive as well as communicative functions that explanations serve in social interaction. Classic phenomena such as the self-serving bias, the fundamental attribution error, and the actor–observer asymmetry may then be fruitfully reexamined, not just under laboratory constraints, but within their natural contexts of occurrence. Greater attention to the social context and functional roles of attribution processes may perhaps be the most important desideratum for future research. As a result, attribution theory might lead to powerful applications in the domain of social relations (e.g., conflict resolution, relationship dynamics) and would contribute to the scientific knowledge of both social perception and social interaction, reverberating Heider's (1958) timeless goal of providing a psychology of interpersonal relations.

See also: Control Behavior: Psychological Perspectives; Control Beliefs: Health Perspectives; Explanatory Style and Health; Mind, Theories of.

Bibliography

Abramson, L.Y., Seligman, M.E.P., Teasdale, J.D., 1978. Learned helplessness in humans: critique and reformulation. Journal of Abnormal Psychology 87, 49–74.
Antaki, C., 1994. Explaining and Arguing: The Social Organization of Accounts. Sage, Thousand Oaks, CA.
Buss, A.R., 1978. Causes and reasons in attribution theory: a conceptual critique. Journal of Personality and Social Psychology 36, 1311–1321.
Gilbert, D.T., Malone, P.S., 1995. The correspondence bias. Psychological Bulletin 117, 21–38.
Gordon, L.M., Graham, S., 2006. Attribution theory. In: The Encyclopedia of Human Development. Sage Publications, Thousand Oaks, pp. 142–144.
Heider, F., 1958. The Psychology of Interpersonal Relations. Wiley, New York.
Hilton, D.J., 1990. Conversational processes and causal explanation. Psychological Bulletin 107, 65–81.
Ickes, W., 1976. A conversation with Fritz Heider. In: Harvey, J.H., Ickes, W.J., Kidd, R.F. (Eds.), New Directions in Attribution Research, vol. 1. L. Erlbaum Associates, Hillsdale, NJ, pp. 3–18.
Jones, E.E., Davis, K.E., 1965. From acts to dispositions: the attribution process in person perception. Advances in Experimental Social Psychology 2, 219–266.
Jones, E.E., Nisbett, R.E., 1972. The actor and observer: perceptions of the causes of behavior. In: Jones, E.E., Kanouse, D., Kelley, H.H., Nisbett, R.E., Valins, S., Weiner, B. (Eds.), Attribution: Perceiving the Causes of Behavior. General Learning Press, Morristown, NJ, pp. 79–94.
Kelley, H.H., 1967. Attribution theory in social psychology. In: Levine, D. (Ed.), In Nebraska Symposium on Motivation, vol. 15. University of Nebraska Press, Lincoln, NE, pp. 192–240.
Kelley, H.H., Michela, J.L., 1980. Attribution theory and research. Annual Review of Psychology 31, 457–501.
Malle, B.F., 1999. How people explain behavior: a new theoretical framework. Personality and Social Psychology Review 3, 21–43.
Malle, Bertram F., 2004. How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction. MIT Press, Cambridge.
Malle, B.F., Moses, L.J., Baldwin, D.A. (Eds.), 2001. Intentions and Intentionality: Foundations of Social Cognition. MIT Press, Cambridge, MA.
Miller, D.T., Ross, M., 1975. Self-serving biases in the attribution of causality: fact or fiction? Psychological Bulletin 82, 213–225.
Ross, L., 1977. The intuitive psychologist and his shortcomings: distortions in the attribution process. In: Berkowitz, L. (Ed.), In Advances in Experimental Social Psychology, vol. 10. Academic Press, New York, pp. 174–221.
Shaver, K.G., 1985. The Attribution of Blame: Causality, Responsibility, and Blameworthiness. Springer-Verlag, New York.
Weiner, B., 1995. Judgments of Responsibility. Guilford, New York.

Auctions

S Müller, Humboldt-Universität zu Berlin, Berlin, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 2, pp. 917–923, © 2001, Elsevier Ltd.

Abstract

This article surveys the literature on auctions. To a large part it is concerned with revenue and efficiency considerations. It addresses the problems of finding the auction that yields the greatest expected revenues to the auctioneer and that allocates the goods to the bidders which value them most highly. Models with private value, common value, and general value are considered. Open (English, Dutch) and sealed-bid (discriminatory, uniform-price, first-price, Vickrey) auctions are analyzed. The analysis focuses on the strategic aspect of the bidding process, that is, the game-theoretic auction literature will be heavily relied on. The employed equilibrium concept is Bayes–Nash. The theoretical analysis is supplemented by reviewing empirical work resulting from real-life auctions or controlled laboratory experiments. The importance of auctions in designing trading institutions is documented for financial markets.

Introduction

The use of auctions to allocate goods is quite common and has been known for centuries. Numerous and quite different goods are sold by auctions, that is, antiques, fish, flowers, real estate, gold, securities, and oil exploration rights. Cassady (1967) gives a complete list of goods sold by auction. The literature on auctions is enormous, and it is increasing rapidly. It is concerned mainly with revenue and efficiency considerations, namely to find the auction that yields the greatest expected revenues to the auctioneer and that allocates the goods to the bidders which value them most highly. Further issues, like the robustness of auctions, transactions costs, bid preparation costs, and the vulnerability to cheating are addressed rarely.

Definitions

According to Cassady (1967: p. 8), auctioning is

> a unique system of allocating … property based on price making by competition of buyers for the right to purchase.

More formally, Myerson (1981) defines an auction mechanism by (1) a pair of outcome functions, which gives the allocation of the goods and the bidders' fees in dependence of the bidders' strategies, and (2) a description of bidders' strategic plans. According to those definitions, there is a plethora of different auctions. To fix ideas, assume that there are N units of a homogenous good for sale and that each bidder is restricted to one unit of the goods. First of all, open and sealed-bid auctions can be distinguished. An auction is open, if the bids become known to all participants of the auction when submitted. In contrast to this, an auction is sealed-bid, if bidders submit their bids privately. Thus, bidders have no knowledge of the bids of their competitors.

The best known open auctions are the English and the Dutch auction (cf Milgrom and Weber, 1982). In the English auction, the auctioneer raises the price up to the point, where N bidders remain, with the understanding that the N units of the good are sold to the remaining bidders at the last called-out price. In the Dutch auction the auctioneer lowers the price interrupted only by bidders accepting the last called-out price for one unit of the good. The auction ends when all units are sold or a lower price limit is reached.

The best known sealed-bid auctions are the discriminatory and the uniform-price (competitive) auction (cf Smith, 1966; Harris and Raviv, 1981). In the discriminatory auction the N units are allocated to the bidders with the highest price bids in descending order. Successful bidders are charged an amount equal to their bids. Thus, for more than one unit for sale, different bidders pay different unit prices in a discriminatory auction. If there is only one unit for sale, the discriminatory auction is called first-price auction. The rules of the uniform-price auction are the same as those of the discriminatory auction, except for the prices paid by successful bidders. In the uniform-price auction all successful bidders are charged the same price, which equals the highest unsuccessful bid. If there is only one unit for sale, the uniform-price auction is called second-price auction or Vickrey auction. Vickrey (1961) introduced this type of auction.

Allowing for more than one seller, a double auction results. In a double auction sellers submit bids to sell and buyers submit bids to buy. The sellers' (buyers') bids are ranked in ascending (descending) order. The price in a nondiscriminatory double auction is determined by the intersection of the obtained supply and demand schedule. This auction as well as other types may be performed as a discrete-time auction at predetermined times during a trading period, or as a continuous-time auction at any moment during a trading period (cf Friedman, 1991).

The goods sold by auction can be classified with respect to the monetary value that bidders attach to them. Milgrom and Weber (1982) distinguish between goods with private value, goods with common value, and goods with general value. Goods are of the private value type (or of preference uncertainty (cf Myerson, 1981)), if its value to a bidder is a purely personal matter. A painting bought for purely personal reasons provides an example. Goods are of the common value type

(or of quality uncertainty (cf Myerson, 1981)), if the monetary value bidders attach to it is the same for all bidders. This value, however, generally will be unknown at the time the auction takes place. Typically, securities have a future uncertain monetary value. Thus, they are of common value. A good is called to be of the general value type, if its value to a bidder is a combination of private and common values. A painting bought for both personal reasons and later resale could be considered being of the general value type.

Analyzing Auctions

The auction literature can be divided into three groups: *decision-theoretic auction models*, *game-theoretic auction models*, and *empirical studies*. Decision-theoretic auction models deal with the bidder's decision problem in a way which ignores her strategic response to the bidding behavior of her competite. The game-theoretic approach, on the contrary, is based on this strategic aspect of the bidding process. Empirical studies make use of either the decision-theoretic or the game-theoretic approach when examining data sets resulting from real-life auctions or controlled laboratory experiments.

The earliest auction models are of the decision-theoretic type. They are motivated by real-world auction design problems concerning the sale of oil exploration rights (Hughart, 1975) and US Treasury bills (Smith, 1966; Scott and Wolf, 1979). Formal in nature, they assume that bidders are expected utility maximizers, where the expectation is with respect to a subjective probability density function for the lowest accepted bid (cp, e.g., Smith, 1966). In particular, it is assumed that a bidder's action does not affect the realization of the lowest successful bid (cp, e.g., Scott and Wolf, 1979). Despite obvious modeling deficiencies, those early models already succeed in identifying certain effects, which later are proved rigorously within the game-theoretic approach to auction theory.

The earliest game-theoretic auction studies primarily are concerned with the determination of Bayes–Nash equilibrium bidding strategies. The models are quite simple in that they assume symmetric bidders with independent value estimates. Those results are generalized in different directions in order to account for bidders with correlated value estimates, for goods of the general value type (Milgrom and Weber, 1982), and for goods with a resale market (Bikhchandani and Huang, 1989).

Empirical studies of auctions are concerned with the analysis of real-life auction data sets or controlled laboratory experiments. Examples are auctions of Federal leases on the Outer Continental Shelf (Hendricks and Porter, 1988), US Treasury bills (Cammack, 1991), wine (Ashenfelter, 1989), and collectible trading cards over the Internet (Lucking-Reiley, 1999).

A Game-Theoretic Auction Model

For ease of exposition there is only one unit of goods to be sold in an auction. The *auctioneer*, who for simplicity is assumed to be the owner of the goods, sets the auction rules and makes them known to all participants. The auction rules are binding even though it might be in the auctioneer's interest to change them after the bids have been submitted. The auctioneer is risk-neutral, that is, interested in maximizing the expected revenues resulting from the sale of the goods. There are n risk-neutral bidders competing with each other in the auction. When preparing their bids, they do not cooperate. Every bidder i has some private information x_i, a realization of her information variable. Bidders are distinguished solely by their private information; otherwise they are identical. It is assumed that the number of bidders and the joint distribution of all information variables is common knowledge.

In an auction a rational bidder takes the bidding behavior of the competitors into account and behaves strategically. However, a bidder has only incomplete information about the relevant features of the bidding situation. Thus, an auction is to be considered a non-cooperative game with incomplete information among the bidders. To analyze this game the Bayes–Nash equilibrium concept (Harsanyi, 1967/1968) is employed. For that purpose, a bidder's strategy is defined as a real-valued function mapping possible realizations of her information variable into bids. A Bayes–Nash equilibrium (B_1, …, B_n) of bidders' strategies is described by the following properties. B_i maximizes bidder i's expected gains assuming bidding strategies B_1, …, B_{i-1}, B_{i+1}, …, B_n of her competitors and the assumed bidding behavior is correct.

Theoretical Results

Fundamental to the theory of auctions is the bidding behavior under varying assumptions. This naturally includes alternative auction rules. In what follows the Dutch, English, first-price, and second-price auctions will be considered and compared.

The simplest set-up is the one where the good to be auctioned is of private value and the values that different bidders attach to it are independent. This assumption is well suited for goods without resale, but it excludes cases like the painting of an artist whose work has some prestige among collectors.

The Private Value Model

Independent Private Values

Although the rules of the standard open and sealed-bid auctions are quite different, Vickrey (1961) found an astonishing similarity. When comparing the Dutch and the first-price auction, he noted that a bidder would follow the same strategy in both auctions. In either auction, a bidder has to determine a price at which she is willing to accept the good. In the first-price auction this would correspond to her bid; in the Dutch auction this would be the lowest called-out price at which she would claim the good. Thus, the Dutch and the first-price auction are strategically equivalent. A similar argument applies to the English and the second-price auction. In either auction, a bidder would bid her private value. This ensures that she will accept all prices below and no price above her private value of the good.

To be more specific in the first-price auction, consider the decision problem of bidder i. Her expected gains are given by

$$x_i \times \text{Probability of winning} - \text{Expected payments} \qquad [1]$$

where x_i denotes her private value. Note that lowering the bid has two opposing effects on expected gains. The probability of

winning decreases, and the gains from the bid increase. Intuitively, the optimal bid will be such that it balances those two effects, thus maximizing eqn [1].

Since bidders differ only by their private information, bidding behavior can be described by a common bidding strategy B. The resulting bidding strategies in the standard sealed-bid auctions are as follows, where F is the distribution function of X_i:

1. Bidding Strategies (Vickrey, 1961; Harris and Raviv, 1981)
   a. In the second-price auction the symmetric bidding strategy is given by
   $$B(x) = x$$
   b. In the first-price auction the symmetric bidding strategy is given by
   $$B(x) = x - \frac{\int_0^x [F(y)]^{n-1}\,\mathrm{d}y}{[F(x)]^{n-1}} \qquad [2]$$

From this a number of conclusions follow that are subject to empirical testing. Both in the first- and second-price auction, bidding strategies are increasing in bidders' private values. In a first-price auction, bidding becomes more aggressive with an increasing number of bidders. A comparison of first- and second-price auctions yields that a bidder submits a lower bid in the first than in the later auction. This difference in bids is represented by the second term of eqn [2].

Uniqueness of symmetric equilibria can be established (cp Milgrom and Weber, 1982). Furthermore, efficiency is guaranteed in both types of auction, that is the bidder with the highest private value is the one who wins the good.

From the described bidding behavior, bidders' expected gains and the auctioneer's expected revenues can be derived. Both coincide in the case of first-price and second-price auction. This is a special case of the revenue equivalence theorem (Vickrey, 1961; Harris and Raviv, 1981; Myerson, 1981; Riley and Samuelson, 1981).

2. Revenue Equivalence. The auctioneer's expected revenues are the same in the single-unit English, single-unit Dutch, first-price, and second-price auction.

Similarly, a bidder's expected gains coincide in the above mentioned auction types.

Strictly Affiliated Private Values

Suppose that bidders' private values are strictly affiliated, that is, large values of one variable are more likely, when the other variables are large as well. Milgrom and Weber (1982) introduced this generalization of nonnegative correlation. This concept allows for goods of private value, where bidders have some common liking.

Although Dutch and first-price auction are still strategically equivalent, this is no longer true for English and second-price auction. In particular, Milgrom and Weber (1982) established the following ranking of auctions.

1. Auction Ranking. The auctioneer's expected revenues in the first-price auction are never higher than those in the second-price auction. The auctioneer's expected revenues in the second-price auction are never higher than those in the single-unit English auction. The single-unit Dutch auction yields the same expected revenues as the first-price auction.

Similarly, in the second-price auction a bidder's expected gains are smaller or equal to those in the first-price auction.

The Common Value Model

In the common value model the unknown value of the good and bidders' information (about the unknown value) are affiliated. Auctions of financial securities like US Treasury bills and of Federal leases on the Outer Continental Shelf can be analyzed within this framework. The informational effect, however, makes bidding quite complicated and gives rise to the following phenomenon.

Winner's curse

Common value models suffer from an effect called the winner's curse.

> learning that others have bid less than [his own bid] … is bad news about the value of the item being acquired.
> Milgrom, 1987: p. 18

Unsophisticated bidding leads to the effect that the successful bidder tends to be the one who overvalued the goods. Substantial losses in common value auctions are quite often attributed to this effect. Thus rational bidders take account of the winner's curse effect, when preparing their bids. The 'winner's curse' was the subject of a lot of controversy (cf, e.g., Kagel and Levin, 1986). Milgrom (1987) formalizes the winner's curse effect for a sealed-bid common value auction for one unit of the good, where the highest bid wins the auction. A bidder's expected unit value of the goods, when she has won with her bid, is smaller than her expected unit value. Thus, the winner's curse effect can be measured by this difference.

Any information on the unknown common value will weaken the winner's curse effect. In particular, linking the price in an auction to affiliated information of the common value reduces the winner's curse. This is the essence of the linkage principle (Milgrom and Weber, 1982). Reducing the winner's curse effect leads to more aggressive bidding and higher auction prices. Also competition matters, that is, the winner's curse effect gets stronger with an increasing number of bidders.

Independent Information Variables

If the bidders' information variables are independent, the results of the common value model resemble those of the independent private value model. In particular, the auctioneer's expected revenues in the first-price and second-price auction coincide. Similarly, the same is true for a bidder's expected gains in both auctions.

Strictly Affiliated Information Variables

The bidding behavior in the common value set up can be described in the same way as in the private value model. Again, Dutch and first-price auction are strategically equivalent. This, however, is not the case for the English and second-price auction. Fundamental to the analysis is a bidder's expected value of the good given her information and the fact of winning the auction with her bid. This way she takes care of the winner's curse effect. Taking this into account, existence and uniqueness

of symmetric equilibria can be established (cf Milgrom and Weber, 1982).

The central results of the common value auction literature concern the ranking of auctions, the auctioneer's information policy and the information aggregation property of auctions. All those results rely on the comparison of symmetric equilibria of the respective auctions. The basic results are as follows:

1. Auction Ranking (Milgrom and Weber, 1982). The auctioneer's expected revenues in the first-price auction are never higher than those in the second-price auction. The auctioneer's expected revenues in the second-price auction are never higher than those in the single-unit English auction.

If the auctioneer has private information about the unknown common value, she may use this fact to influence the outcome of the auction. She may commit to supplying information publicly in order to increase revenues. There is a whole range of possible information policies. Milgrom and Weber (1982) mention complete revelation (fully reporting of all information), censoring (reporting of favorable information only), randomizing (reporting after having added noise to the information), and concealment (no reporting of private information).

2. Information Revelation (Milgrom and Weber, 1982). The auctioneer's policy of revealing any affiliated information publicly and completely raises her expected revenues in the first-price auction, the second-price auction, and the single-unit English auction.

There has always been the question of whether the price of a good of unknown value reflects (aggregates) the dispersed information of market participants. Under certain assumptions this could be answered in the affirmative in the case of auctions.

3. Information Aggregation (Wilson, 1977). The winning bid in a first-price auction converges almost surely to the common value V as the number of bidders n goes to infinity.

The General Value Model

The general value model gives results in much the same way as the common value model. In particular, the auction ranking and the information revelation results hold true. However, the information aggregation property relies on the simple structure of the pure common value model and thus does not generalize.

Empirical Results

There is an increasing number of studies which provide tests of auctionmarket behavior by means of experimental methods or real-life auction data sets (field data). Whereas the experimental approach allows for controlling a specific auction model including informational assumptions, the field approach provides a test of auction theory in situations where most theoretical assumptions are unobservable.

Quite often, experimental studies are concerned with the independent private values model, in which predictions are being tested. The results obtained are mixed (cf Smith, 1991; Kagel, 1995, and their citations for the following). Bidding behavior is characterized by deviations from the strategic equivalence of first-price and Dutch auctions and second-price and English auctions, respectively. In particular, in first-price (second-price) auctions bidders submit higher bids than in the corresponding Dutch (English) auctions. Consequently, revenue equivalence does not hold. However, as predicted by theory, individual bids are strictly increasing in bidders' private values. As concerns comparative statics, increasing the number of bidders leads to more aggressive bidding on the bidders' part in first-price auctions. When introducing affiliated private values in experiments, the resulting bids could be better explained by Bayes–Nash bidding behavior than by alternative ad hoc bidding models.

Fundamental to experimental work in the common value set-up is the study of Smith (1967), who examines individual bidding behavior in discriminatory and uniform-price auctions. His results are consistent with the predictions of auction theory. He establishes that bids in the discriminatory auction are lower than those in the uniform-price auction and that the same is true for the auctioneer's revenues. Other results are less supportive of auction theory (cf Kagel and Levin, 1986; Kagel, 1995, and his citations for the following). As predicted by theory, individual bids are strictly increasing in bidders' information. However, comparative statics might fail. This is the case in second-price auctions, where increasing the number of bidders leaves bids unaffected.

In general, Bayes–Nash bidding behavior in sealed-bid auctions is questioned by strong winner's curse effects especially for inexperienced bidders. Furthermore, the common value auction rankings could not be confirmed in experiments. Due to the winner's curse effect, sealed-bid auctions yield higher revenues than English auctions. Even information revelation can be affected by the winner's curse effect. Kagel and Levin (1986) find that information revelation reduces average revenues, when the number of bidders is high. Only for smaller numbers of bidders, the theoretical predictions of information revelation could be confirmed. This effect applies to sealed-bid as well as English auctions. Thus, in first- and second-price auctions theoretical predictions are valid only when the winner's curse effect is weak.

Early results of field data support auction theory. For example, Tenorio (1993) establishes that in the Zambian foreign exchange uniform-price auction average revenues are higher than in the corresponding discriminatory auction. Lucking-Reiley (1999), on the contrary, reports a failure of revenue equivalence for auctions of collectible trading cards over the Internet. He finds that revenues in the Dutch auction are higher than those in the first-price auction. With respect to English and second-price auctions, however, the data support revenue equivalence. This contrasts sharply with the results of independent private values experiments.

Double Auctions

Since the nineteenth century, trade at organized exchanges to a large extent was conducted by means of double auctions. Nowadays, double auctions are important tools for computerized trade. They also serve well for the organization of markets for natural gas as well as for electric power networks. The theoretical analysis concerning double auctions is much more involved than that of one-sided auctions. Determination

of Bayes–Nash equilibrium is complicated, if at all feasible. Often the analysis gives rise to a multiplicity problem. For that reason, the theoretical analysis centers around the independent private values model for a single indivisible unit. Satterthwaite and Williams (1993) give results for particular sealed-bid nondiscriminatory double auctions. For m risk-neutral sellers and m risk-neutral buyers, symmetric Bayes–Nash equilibria are determined. Convergence of equilibria to ex post efficiency as both the number of sellers and buyers increase is shown to hold.

However, due to the above mentioned technical problems, the analysis of double auctions is done almost always by means of empirical or experimental studies (cf, e.g., Friedman and Rust, 1993; Section Analyzing Auctions).

Application to Financial Markets

Interest in the market institutions that govern the trade of financial securities grew rapidly during the last decades of the twentieth century. Along with the successful introduction of new exchanges, the trading rules at existing exchanges were examined more closely. Common to all those activities is an interest in the price formation process under alternative trading rules. An important example of this approach is the discussion of the design of the US Treasury bill auction, one of the most important auctions of financial securities. Every Monday the US Treasury auctions a previously announced quantity of Treasury bills maturing in 91 and 182 days, respectively.

Since the late 1950s there has been quite a controversy concerning the design of this auction. This controversy started with a proposal from Milton Friedman to change the Treasury bill auction from discriminatory to uniform-pricing. The reasons given in favor of this proposal were quite different in nature. They ranged from revelation of bidders' true demand curves and an increase in the number of bidders to discouragement of collusion under uniform-pricing. They did not, however, include an increase in the Treasury's average revenues from a sale under uniform-pricing. In fact, the early opponents to this proposal argued by a reduction in revenues through a change from discriminatory to uniform-pricing. Several years later, models of the decision- and the game-theoretic type lent additional support to the Friedman proposal.

In that respect, the results of Smith (1966) and Milgrom and Weber (1982) on the superiority of uniform over discriminatory pricing in terms of auctioneer's expected revenues were especially successful. Despite the fact that a completely satisfactory theoretical model of the Treasury bill auction is not available yet, those results made the Treasury reconsider the Friedman proposal. Those considerations led to experimentation with a uniform-price auction in the 1970s. However, the Treasury retained the discriminatory rules. Only after the Salomon Brothers Inc. admission of violating the US Treasury auction rules, the experiments were taken up again in September 1992 for the 2-year and 5-year note auctions (cf Bikhchandani and Huang, 1993). Experimentation ended in a decision effective with the 11 February 1998 auction to auctioning all bills using the uniform-pricing rules.

Conclusion

The basic game-theoretic auction model gives rise to a number of strong results which have been subject to empirical and/or experimental tests. Naturally, the question arises as to whether those results are robust with respect to a relaxation of assumptions. Unfortunately, there are several assumptions which, when relaxed, give rise to different results. There are a number of factors which have to be dealt with. First, risk neutrality matters. When allowing for risk aversion on the bidders' part, the expected revenue comparison will yield different results. For example, in models with both risk-aversion and strictly affiliated information variables, the established ranking of first- and second-price auctions with respect to expected revenues fails to hold. Only in special cases (e.g., independently and identically distributed (i.i.d) information variables) general results can be established (cf Milgrom and Weber, 1982).

Second, multiunit demand by bidders changes results dramatically. Endogenizing quantity, for example, as in the case of a share auction, might result in a multiplicity of equilibria and a substantial lower expected sale price than in comparable unit auctions (cp Wilson, 1979). Third, in auctions of financial securities bidders usually meet more than once. The introduction of this phenomenon produces all effects that apply to repeated noncooperative games. Finally, a number of features cannot be solved in the outlined theoretical auction model due to mathematical complexities. It is in those cases that one has to rely on controlled laboratory experiments in order to draw conclusions.

See also: Game Theory.

Bibliography

Ashenfelter, O., 1989. How auctions work for wine and art. The Journal of Economic Perspectives 3, 23–36.

Bikhchandani, S., Huang, C., 1989. Auctions with resale markets: an exploratory model of treasury bills markets. The Review of Financial Studies 2, 311–339.

Bikhchandani, S., Huang, C., 1993. The economics of treasury securities markets. Journal of Economic Perspectives 7, 117–134.

Cammack, E.B., 1991. Evidence on bidding strategies and the information contained in Treasury bill auctions. Journal of Political Economy 99, 100–130.

Cassady Jr., R., 1967. Auctions and Auctioneering. University of California Press, Berkeley, CA.

Friedman, D., 1991. The double auction market institution: a survey. In: Friedman, D., Rust, J. (Eds.), The Double Auction Market. Institutions, Theories and Evidence. Addison-Wesley, Reading, MA, pp. 3–25.

Friedman, D., Rust, J. (Eds.), 1993. The Double Auction Market. Institutions, Theories and Evidence. Addison-Wesley, Reading, MA.

Harris, M., Raviv, A., 1981. Allocation mechanism and the design of auctions. Econometrica 49, 1477–1499.

Harsanyi, J.C., 1967/1968. Games with incomplete information played by Bayesian players, Part I. Management Science 14, 159–182. Part II. Management Science 14, 320–334; Part III. Management Science 4, 486–502.

Hendricks, K., Porter, R.H., 1988. An empirical study of an auction with asymmetric information. American Economic Review 78, 865–883.

Hughart, D., 1975. Informational asymmetry, bidding strategies, and the marketing of offshore petroleum leases. Journal of Political Economy 83, 969–985.

Kagel, J.H., 1995. A survey of experimental research. In: Kagel, J.H., Roth, A.E. (Eds.), The Handbook of Experimental Economics. Princeton University Press, Princeton, NJ, pp. 501–585.

Kagel, J.H., Levin, D., 1986. The winner's curse and public information in common value auctions. American Economic Review 76, 894–920.

Lucking-Reiley, D., 1999. Using field experiments to test equivalence between auction formats: magic on the Internet. American Economic Review 89, 1063–1080.
Milgrom, P.R., 1987. Auction theory. In: Bewley, T. (Ed.), Advances in Economic Theory. Cambridge University Press, Cambridge, UK, pp. 1–32.
Milgrom, P.R., Weber, R.J., 1982. A theory of auctions and competitive bidding. Econometrica 50, 1089–1122.
Myerson, R.B., 1981. Optimal auction design. Mathematics of Operations Research 6, 58–73.
Riley, J.G., Samuelson, W.F., 1981. Optimal auctions. American Economic Review 71, 381–392.
Satterthwaite, M.A., Williams, S.R., 1993. The Bayesian theory of the k-double auction. In: Friedman, D., Rust, J. (Eds.), The Double Auction Market. Institutions, Theories and Evidence. Addison-Wesley, Reading, MA, pp. 99–123.
Scott, J.H., Wolf, C., 1979. The efficient diversification of bids in treasury bill auctions. Review of Economics and Statistics 61, 280–287.
Smith, V.L., 1966. Bidding theory and the Treasury bill auction: does price discrimination increase bill prices? Review of Economics and Statistics 48, 141–146.
Smith, V.L., 1967. Experimental studies of discrimination versus competition in sealed-bid auction markets. Journal of Business 40, 56–84.
Smith, V.L., 1991. Auctions. In: Eatwell, J., Milgate, M., Newman, P. (Eds.), The New Palgrave: A Dictionary of Economics. Macmillan, London, pp. 138–144.
Tenorio, R., 1993. Revenue equivalence and bidding behavior in a multi-unit auction market: an empirical analysis. Review of Economics and Statistics 75, 302–314.
Vickrey, W., 1961. Counterspeculation, auctions, and competitive sealed tenders. Journal of Finance 16, 8–37.
Wilson, R.B., 1977. A bidding model of perfect competition. Review of Economic Studies 44, 511–518.
Wilson, R.B., 1979. Auctions of shares. Quarterly Journal of Economics 93, 675–689.

Audience Measurement

James G Webster, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Audience measurement entails regular assessments of the size and composition of media audiences. These are presented as syndicated reports used by media industries. Measurement services use questionnaires, diaries, meters, and/or servers to quantify people's exposure to mass media. All methods of data collection are affected by error, including sampling, nonresponse, and response error, as well as error in the production of reports. Audience measurement is a specialized kind of market research that is deeply enmeshed in the media systems of most industrialized nations. These measures, which include broadcast audience ratings, play an important role in shaping popular culture.

Audience measurement, as the term is commonly used, refers to regular assessments of people's media use. These data often result in estimates of the size and composition of audiences, like television audience ratings, that enable advertiser supported media. Audience measurement can be thought of as a specialized form of market research that is deeply enmeshed in the media systems of most industrialized nations (Anand and Peterson, 2000). It quantifies popularity, guides the allocation of billions of dollars in advertising worldwide and, in doing so, shapes the media environment and popular culture.

Uses of Audience Measurement

The major impetus for audience measurement is advertising. By the late nineteenth century, most newspapers and mass circulation magazines had begun to sell space to advertisers. The value of space was determined largely by the number of readers who would see each publication, and by extension each ad. Paid circulation served as a reasonable surrogate for the size of the readership. Many publishers, however, made inflated claims about their circulation, undermining the confidence of advertisers. To make print media credible, in 1914 the industry created the Audit Bureau of Circulations, an independent organization that verified circulation. To this day, such audits continue in the United States and many other countries around the world.

The growth of radio in the 1920s presented new challenges. Unlike print, there were a few traces of the radio audience, save for fan mail or the sheer number of receivers being sold. For radio to be a viable advertiser-supported medium, the industry needed to authenticate the size and composition of audiences. The problem was especially pressing in countries like the United States and Australia that relied on advertising to fund broadcasting (Balnaves et al., 2011). To provide reliable estimates of the number of people listening to specific programs, the industry sponsored audience surveys. By the 1930s independent research firms had begun to produce what were called 'ratings' (i.e., percentages of the population tuned to a particular program or station). These made the audience visible and allowed advertisers to buy spot announcements in radio with a degree of certainty. When television and other forms of electronic media emerged later in the century, similar mechanisms for measuring the size and composition of audiences were quickly put in place (Beville, 1988). By the 1980s, as advertiser support became more widespread, most European countries put similar systems in place (Gunter, 2000).

Today, advertising is a multibillion dollar business. Audiences are bought and sold like commodities. In many countries, these transactions are made possible by a dominant firm that supplies measurements for a particular medium. Advertising expenditures are typically guided by the cost of reaching various audience segments. In a business world increasingly interested in target marketing, research firms have been called upon to produce ever finer demographic distinctions, as well as data on lifestyles, levels of engagement, and product purchases.

Curiously, audience ratings describe something that has already happened, while most advertising expenditures are made in anticipation of audiences that have yet to materialize. This odd system works because, in the aggregate, media audiences are quite predictable. For example, advertisers and networks know the total number of people who are likely to watch television at any time during the year, and negotiate ad prices on that basis. The stability of mass behavior has also allowed researchers to identify various 'law-like' features of audience formation (e.g., Goodhardt et al., 1987; Webster and Phalen, 1997). Web use appears to follow similar law-like patterns (e.g., Easley and Kleinberg, 2010), although advertisers often pay after the fact as a function of audience impressions or 'click-throughs' (Turow, 2011).

The fact that advertising is a major source of revenue for most forms of media (e.g., print, radio, television, the Internet, etc.) has embedded audience measurement in the operation of these industries. Obviously the system places a premium on audiences that will be attractive to advertisers, either by virtue of their sheer size or other desirable characteristics. Content, therefore, is evaluated with an eye toward its audience-making potential. Moreover, broadcasters pay careful attention to how audiences 'flow' across an entire schedule of programming, relying on structural factors to deliver people to their program offerings. Web publishers exploit hyperlinks and recommendation systems in a similar way. In all of these endeavors, audience measurement is needed to manage and sell the audience. In fact, a good case can be made that audience measurement is a necessary condition for the emergence of any form of advertiser supported media (Ettema and Whitney, 1994).

Even in social systems not fully committed to advertising, audience measurement can become a fixture. Public service broadcasting around the world must often justify its existence by demonstrating that it has an audience (Ang, 1991; Gunter, 2000). Here the techniques of audience measurement are similar to those in fully commercial systems, although they may involve the use of more 'qualitative' measures that assess the appeal or usefulness of media materials (Gunter, 2000). Similarly, government officials concerned with good or bad media effects often relate these to people's exposure to good or bad content (Webster and Phalen, 1997).

Measurement in the Digital Media Environment

Since the 1990s, there has been explosive growth in digital media. This has affected older media, like television, by greatly increasing the number of channels and 'on-demand' options available to viewers. It has also created new ways for users to find, view, download, and share media over the digital networks. These developments pose a challenge to traditional methods of measurement by fragmenting audiences and enabling anywhere anytime media consumption.

Prior to the arrival of digital media, virtually all forms of audience measurement depended on drawing a sample of users to estimate the population. This is a 'user-centric' approach to measurement. As competition for audiences has grown, each new channel or Web site has claimed a piece of public attention. As a result, audiences, which used to concentrate on a few outlets, have been fragmented into smaller and smaller groups. This, coupled with the advertiser's desire to reach more narrowly defined audiences, has taxed the ability of samples to keep pace and increased the likelihood of sampling error – a topic addressed below.

Digital media have also enabled the growth of nonlinear technologies, like digital video recorders (DVRs), video-on-demand (VOD), and Web sites that allow users to stream or download media at will. Further, many of these are available on multiple platforms like computers, tablets, and smart phones, allowing users to fetch media anywhere at any time. As a result, it is no longer possible to rely exclusively on scheduling information to determine what media each person is exposed to.

These challenges are mitigated, to some extent, by harnessing the servers that power digital media. They include the computers that (1) provide content to users, by delivering web pages, videos, or other media; (2) enable access through the Internet service providers; and (3) place display ads across multiple Web sites to deliver impressions. In the process, servers can record each user action, which forms the basis for a new way to measure audiences. This is a 'server-centric' approach to measurement.

Each approach, the user-centric method and the newer server-centric strategy, has strengths and weaknesses. Generally speaking, user-centric methods do a better job of assessing the traits of individual users (e.g., age, gender, etc.) and allowing measurement firms to track media consumption across platforms (e.g., print, broadcast, Web, mobile, etc.). However, since they use samples it is difficult to accurately estimate the audience for smaller outlets without undue sampling error. Server-centric methods avoid the problem of too small samples because they see the actions of everyone who uses the server. In theory they provide a census of media use, although in practice all members of the relevant population may not be visible to the server. More importantly, they cannot see who is visiting the server, unless visitors have provided that information. While that is sometimes the case, for example with 'social media' sites, it is often difficult to accurately describe the characteristics of the users. Moreover, servers cannot see beyond their associated digital networks, making it difficult to assess cross-platform media use.

Measurement companies have devised solutions to many of these problems, though none is without its own drawbacks. Often, data are 'fused' to create a hybrid of user- and server-centric information in an attempt to capture the best of both worlds. Unfortunately, true 'single source' systems that measure all the desirable variables (i.e., user traits, product purchases, moment-to-moment multi-platform media use) across a very large sample or population are costly, can impose undue burdens on respondents, and are therefore rare.

Methods of Audience Measurement

Methods of audience measurement are closely tied to the use of user-centric or server-centric approaches. The latter, more traditional approach, offers greater flexibility about who is studied and how they are measured. As a type of survey research, user-centric measurement confronts issues common in sampling and statistical inference. Its distinctiveness comes from the methods it uses to measure the behavior of audience members. One of the three techniques is typically employed: questionnaires, diaries and meters. Server-centric methods limit who can be studied and confine measurement to actions that can be tracked or inferred from server usage.

The first regular surveys of radio audiences were conducted using telephone interviews. Respondents were called at home and asked either what they were currently listening to (i.e., telephone coincidental method) or what they had listened to in the recent past (i.e., telephone recall method). Both techniques are used today. In fact, telephone coincidentals were historically regarded as the most accurate way to measure broadcast audiences and served as the standard against which other methods were judged (Beville, 1988).

The advent of radio audience measurement forced print media to go beyond circulation as the sole indication of readership. Because a single copy of a magazine or a newspaper might well be read by more than one person, circulation alone seemed to understate the audience for print. The industry eventually adopted survey research to better estimate readership. Personal interviews were conducted in which respondents were shown stripped down copies of various magazines and asked if they recalled reading them. Alternatively, to economize on time and increase the number of publications that could be assessed, respondents were asked to sort cards containing magazine logos and indicate which of these were read.

Today, more and more survey work is being done over the Internet. Respondents, who are either recruited or volunteer, fill out to questionnaires online. This method provides some flexibility in questionnaire design, allowing researchers to

introduce visual materials or scripted clarifications. Using the Internet to administer questionnaires can offer fast turnaround times and be relatively inexpensive. However, since these techniques are limited to those who have Internet access and are willing to cooperate, making inferences to larger populations must be done with caution.

Diaries are small booklets in which respondents keep a log of their radio or television usage throughout the day. Typically, diaries are kept for a week, though time periods of different durations are possible. Diaries are a relatively inexpensive data collection technique and, properly filled out, contain a wealth of information. However, many people are reluctant to accept a diary or are unwilling or are unable to make accurate entries. In a world of push-button radios, multi-channel television, and remote control devices, completing a diary can be a burdensome task. Diaries tend to overrepresent the use of popular outlets or programs that come easily to mind and underrepresent the number of sources that a diary-keeper actually uses. Nonetheless, diaries are a common measurement technique for producing local radio and television audience reports (Webster et al., 2014).

Many of the problems associated with diaries can be solved by the use of meters. These devices attach to a receiver and automatically record the source to which the set is tuned. By the early 1940s, Arthur Nielsen had developed such a device and had placed it in a panel of American households to produce national radio network ratings. Household meters were eventually adapted to the new medium of television. While household meters provide a relatively accurate measure of set usage, they are expensive and offer no information about who is actually using the set. With advertisers increasingly concerned about demographics, the latter is a serious shortcoming.

To provide information about viewers, meters were adapted so that people could identify themselves by pressing a button. These 'peoplemeters' were put into service in the United States and most European countries in the 1980s (Webster et al., 2014). Newer peoplemeters are capable of detecting what content is being viewed, eliminating reliance on program schedule information. They do this by detecting an electronic code embedded in the signal, sometimes called a 'watermark,' or by actively comparing the digital signature of the signal to a library of digital media. Today, peoplemeters are the preferred method of television audience measurement in much of the world.

Conventional peoplemeters are attached to television sets, and are therefore unable to measure out-of-home media use (e.g., radio) or cross-platform use. To remedy that problem, measurement companies also use portable peoplemeters (PPMs). These are either built into pager-like devices or wristwatches that respondents wear throughout the day. They 'listen' for an inaudible code (i.e., a watermark) in the audio portion of a broadcast signal, and credit listenership accordingly. PPMs are then docked at night to be recharged and cleared of the information they contain. This technology is used to measure radio audiences in larger markets. It can also capture other forms of media use (e.g., TV viewing) and so holds promise for studying cross-platform behaviors.

With more and more media use migrating to the Web, measurement firms also meter personal computers. Respondents are asked to load measurement software on their own computers, which records the various pages that the user visits and reports that to the research company. These data are aggregated to describe the number and type of people who visit the Web sites. They also support analyses of how audiences flow from one site to another. Like other exercises in metering, this is a user-centric approach that depends upon the cooperation of sample members.

As noted above, a powerful new way to measure Internet use is to monitor servers. These computers see and record the actions of all those using the server. Learning something about the identity of the visitor is the biggest challenge in server-centric measurement. A variety of techniques are used. Servers will note the IP addresses of the visitors. That may reveal something about their location. Identities can sometimes be inferred by the kinds of materials a user accesses. Most importantly, servers routinely place 'cookies' on the browsers of visitors. These are small computer files that contain a record of what a user did on the site, and serve to identify them as repeat customers. Cookies can also be used to track users across Web sites and target advertising. These practices raise questions about user privacy and are the subject of regulation that varies from country to country.

Servers also manage the traffic on most broadband distribution systems. Some firms aggregate data from Internet service providers to estimate Web site traffic. Cable and satellite systems often stream digital video signals into subscribing households, which are managed by 'set-top-boxes' (STBs) attached to TV sets. When properly programmed, STBs can record channel usage. They become the functional equivalent of household meters, except that they are present in millions of homes. The census-like scope of the sample eases the problem of measuring fragmented audiences. When STBs are used to estimate television audiences the data must be carefully weighted to account for the fact that many homes and television sets are not attached to these digital networks.

Sources of Error in Audience Measurement

There are four sources of error in audience measurement: sampling, nonresponse, response, and production. The first three are, again, problems commonly associated with survey research. The last includes a variety of issues that have to do with bringing a saleable research product to market.

Most syndicated audience measurement is based on some form of probability sampling. Abundant digital media and the ensuing audience fragmentation, make sampling error a persistent problem. In world where some media subsist with a fraction of 1% of the total audience, even a sample with tens of thousands of respondents, quickly becomes too small to accurately estimate the audience for smaller outlets or unpopular products. The error may be random, but it is unacceptably high relative to the estimate. Unfortunately, the cost of some measurement devices (e.g., meters) and the diminishing marginal returns associated with increases in sample size have forced difficult compromises on sample size and allowable levels of sampling error. Often, unpopular offerings are simply 'below minimum standards' for reporting. Servers that monitor millions of users can help provide the needed granularity, but with the limitations noted above.

There are also potential problems of response. Written forms of feedback (e.g., self-administered questionnaires,

diaries, etc.) often suffer from low rates of response. Further, as people become increasingly leery of telemarketers, telephone survey completion rates are threatened. All these lead to concerns about nonresponse (i.e., the people providing data are systematically different from those who don't). Even among those who agree to cooperate, there may be problems with the quality of their responses (e.g., response error). Diaries are particularly prone to response errors, as the flood of modern media simply overwhelms the reporting capabilities of even the most conscientious respondent. These errors can be nonrandom. For example, because diaries systematically favor popular media, they can produce reliable results of questionable validity. The accuracy of self-reported media use has become a contentious issue in some academic circles (e.g., Dillipane et al., 2013; Prior, 2013).

The information collected by interviews, diaries, meters, and servers can be thought of as raw material that must undergo a production process to become a saleable product. Data must be coded and checked for obvious errors. Missing values are occasionally imputed using a process called ascription. New information must be added. For example, the characteristics of server users might be inferred by fusing it with other data sets. Responses from underrepresented or overrepresented groups are often assigned a mathematical weight in a process called sample balancing. Finally, data must be aggregated and projected to create population estimates. Such production processes can introduce errors. Sometimes, it is human error. Sometimes it is applying a less than perfect mathematical correction to the data (Webster et al., 2014).

Institutional Aspects of Audience Measurement

In most developed countries, audience measurement is a business that sells research reports or data to multiple subscribers. Sometimes, as is the case in the United States, firms compete openly for clients. In many countries, however, the consumers of audience measurement create a 'joint industry committee' (JIC) that specifies what measurement services are required and awards a multiyear contract to a sole supplier. Under either system, it is common to see one company become the dominant supplier of the 'currency' within a medium. As such, their reports take on the air of quasi-public, official documents. Though various errors in the data mean the numbers are not nearly as 'hard' as they appear, the organizations that use them have little alternative but to treat them as such. And an enormous amount rides on those numbers. The process of allocating billions of dollars in advertising and programming resources is informed, some would say tyrannized, by audience measurement. This affects the kinds of media that are available and, more generally, how entire media systems operate (Ang, 1991; Napoli, 2011; Turow, 2011).

The institutional significance of audience measurement has not gone unnoticed. In the 1960s the United States Congress held hearings into the quality and consequences of the American broadcast ratings industry. Public scrutiny induced the industry to create an ongoing council to audit and accredit ratings research methods, as well as sponsor industry-wide methodological research (Beville, 1988). JICs are also a forum for reviewing research methods. For these reasons, and because audience metrics are used by powerful institutions with competing interests, it is probably the case that no supplier of audience measurements can drift too far from truthful reporting without sowing the seeds of its own destruction.

Even so, much about audience measurement is negotiable and left to industry consensus. The very definition of exposure to media has changed over time and from method to method. For example, in the US the currency for valuing television advertising used to be program ratings. Advertisers, however, were more interested in how many people saw their ads. So in 2007, the currency became commercial ratings plus 3 days of replays, called 'C3s.' Today servers capture more information, which can be parsed in so many ways, that industry consensus about the definition and utility of metrics is a source of ongoing discussion. Social media, for example, allow research firms to track what people say about programs and products. These 'social listening' data can form the basis of metrics on 'engagement,' though just how they will be used remains to be seen. New metrics, once they are adopted, can alter the way media industries operate (Napoli, 2011).

By its very nature, audience measurement is reductionist. Critics have rightly charged that it fails to capture the idiosyncratic, lived experiences of media users. Conventional forms of audience measurement tell us little about what sense people make of the media use and how that use affects them. Some also charge that measurement turns people into commodities by lumping them into categories where one person is the functional equivalent of all others in the group. These groups are then sold to advertisers at some 'cost-per-thousand.' At best, these practices are dehumanizing. At worst, it is argued that the entire enterprise manipulates and ultimately victimizes audiences (Ang, 1991; Turow, 2011).

Alternatively, audience measurement can be seen as exercising a democratizing influence on popular culture. Material that is successful in the marketplace proliferates, while content that is unpopular fades away. While this notion of cultural democracy is oversimplified, it has been argued that measurement empowers audiences by casting them in a form to which institutions can and do respond (Ettema and Whitney, 1994; Webster and Phalen, 1997). Newer social media, which allow people to express their likes and dislikes, share links, and recommend content, extend the impact of measurement. These activities are supported and rendered visible by 'user information regimes,' such as search engines and recommender systems, that aggregate user data. The metrics they produce affect both the production and consumption of media (Webster, 2014). One way or another, it can certainly be said that audience measurement helps shape popular culture.

See also: Advertising: General; Audiences, Media; Mass Media, Political Economy of; Media Effects.

Bibliography

Anand, N., Peterson, Richard A., 2000. When market information constitutes fields: sensemaking of markets in the commercial music industry. Organization Science 11 (3), 270–284.

Ang, Ien, 1991. Desperately Seeking the Audience. Routledge, London.

Balnaves, Mark, O'Regan, Tom, Goldsmith, Ben, 2011. Rating the Audience: The Business of Media. Bloomsbury, London.
Beville, Hugh M., 1988. Audience Ratings: Radio, Television, Cable, rev. ed. Erlbaum, Hillsdale, NJ.
Dilliplane, Susanna, Goldman, Seth K., Mutz, Diana C., 2013. Televised exposure to politics: new measures for a fragmented media environment. American Journal of Political Science 57 (1), 236–248.
Easley, David, Kleinberg, Jon, 2010. Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press, Cambridge, UK.
Ettema, James S., Whitney, D. Charles, 1994. Audiencemaking: How the Media Create the Audience. Sage, Thousand Oaks, CA.
Goodhardt, Gerald, J., Ehrenberg, Andrew S.C., Collins, Martin A., 1987. The Television Audience: Patterns of Viewing, second ed. Gower, Westmead, UK.
Gunter, Barrie, 2000. Media research Methods: Measuring Audiences, Reactions and Impact. Sage, London.
Napoli, Philip M., 2011. Audience Evolution: New Technologies and the Transformation of Media Audiences. Columbia University Press, New York.
Prior, Markus, 2013. The challenge of measuring media exposure: reply to Dilliplane, Goldman, and Mutz. Political Communication 30 (4), 620–634.
Turow, Joseph, 2011. The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth. Yale University Press, New Haven, CT.
Webster, James G., Phalen, Patricia F., 1997. The Mass Audience: Rediscovering the Dominant Model. Erlbaum, Mahwah, NJ.
Webster, James G., Phalen, Patricia F., Lichty, Lawrence W., 2014. Ratings Analysis: Audience Measurement and Analytics, fourth ed. Routledge, New York.
Webster, James G., 2014. The Marketplace of Attention: How Audiences Take Shape in a Digital Age. MIT Press, Cambridge, MA.

Relevant Websites

http://www.gesis.org/en/services/data-collection/gesis-panel/.
http://www.gfk.com/.
http://www.kantarmedia.com/.
http://www.lissdata.nl/lissdata/.
http://mediaratingcouncil.org/.
http://nielsen.com/us/en.html.
http://www.nielsen.com/content/corporate/global/en.html.
http://www.telecontrol.ch/typo3/.

Audiences, Media

Andrea L Press, Fan Mai, Francesca Tripodi, and Michael L Wayne, The University of Virginia, Charlottesville, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, the authors provide a historical overview of audience research spanning theoretical perspectives from the hypodermic needle model, the transition to British Cultural Studies, and on to postmodern conceptualizations. Applying an interdisciplinary lens, the article provides summaries of pertinent studies in anthropology, cultural studies, and sociology that explore how audiences understand, interact, and use media as part of their daily lives. In addition to the historical background, the article also explores specific subfields in new media, feminist, and global audience research. It concludes with insights into how the field is assessing new developments in social class and reality television.

The Beginning

The notion of audience has been largely derived from the image that media producers have of the actual, or the intended, people or groups of people that they imagine as the main recipients of their products. Media products – newspapers, television shows, films, radio broadcasts – are for the most part manufactured with the aim of capturing the attention of audience members. Often this interest is commercial – when media can capture the attention of particular audiences, their producers can sell this attention to various advertisers who may profit from this exposure. Scholars, therefore, are often rancorous in their disapproval of the commercial nature of the very concept of media audience.

Current audience research comes out of this tradition. At first, a 'hypodermic needle' model of media influence was prevalent among media scholars. This asserted that audience members were powerfully influenced directly by media exposure ('injected' with media 'messages,' so the metaphor goes). This model posited an essentially passive audience, and a media full of clearly understandable content with only one meaning.

This theory of powerful influence and simple media content was supplanted by the work of Paul Lazarsfeld in the 1930s. Lazarsfeld headed the Bureau of Applied Social Research, affiliated loosely with Columbia University, in the 1940s (he later became a professor of sociology at the University). Along with Elihu Katz, Lazarsfeld elaborated the minimal effects model in *Personal Influence* (1955). This work was a more general study of the influence of media on a variety of areas of peoples' everyday lives. It became famous for elaborating the 'two-step' flow model of the influence of media on the audience. Instead of directly and powerfully influencing people, Katz and the study confirmed the idea that media effects, rather than being powerful and evil, were actually for the most part rather minor and noninvasive, mediated through the more important effect of opinion leaders.

The limited effects model was influential in audience research throughout the 1960s. However, by the 1970s, this model was frequently criticized. For example, Todd Gitlin set forth a direct, vitriolic critique of limited effects theory generally, which has been widely read and cited (1978), and which had an important influence on audience research since. In this article, he accuses Lazarsfeld, Katz, Klapper, and other limited effects theorists of misconceiving the problem of media effects. By conceptualizing an 'effect' too narrowly and concretely, Gitlin claimed that researchers missed many less-measurable ways in which the mass media influences its audience. Much of Gitlin's critique of the limited effects model has been picked up today by researchers studying the audience from a culturally contextualized perspective, influenced by anthropological and sociological methods. Nonetheless, the two-step flow model remains relevant in a variety of fields including advertising (e.g., Carr and Hayes, 2014), journalism (e.g., Farnsworth et al., 2010), and information sciences (e.g., Case et al., 2004). As Neuman and Guggenheim (2011) observe in their *meta*-analysis, in recent years, scholars addressing new media are drawing attention to the ways in which the dramatically expanded set of choices facing media audiences necessitate a reconsideration of the central theoretical premises of the media effects paradigm.

British Cultural Studies

The contemporary tradition of critical media audience analysis can largely be traced back to scholars associated with the Centre for Contemporary Cultural Studies in Birmingham. Stuart Hall, David Morley, Angela McRobbie, Paul Willis, Dick Hebdige, and Roger Silverstone have all done work that looks at media audiences not as isolated phenomena, but as individuals and groups of individuals who must be studied in the context of the rest of their lives, and whose nature as a part of the media audience is only one segment of an overall set of cultural practices that characterize their identities. Also emphasized in cultural audience work has been the multilayered nature of texts as well as the complexity of how they are received.

Perhaps the first work to bring together textual and audience analysis from these new, critical perspectives was Stuart Hall's essay about television reception, entitled 'Encoding/Decoding in the Media Discourse' (1973). In this work, Hall theorizes both the complex nature of the meanings 'encoded' primarily in the television text, and the necessarily separate, but equally complex nature of the process by which viewers decode these messages. This seminal article paved the way for the more empirical audience studies which were produced in its wake by cultural audience researchers. Audience research carried out in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.95010-4

this tradition emphasizes overall the ways in which the media are part of culture, and concomitantly, the way in which audience reception of media is only one facet of the way an entire cultural system influences those who live within it.

David Morley's work has been particularly influential. He has authored two famous studies, *The Nationwide Audience* (1980) and *Family Television* (1986). In *The Nationwide Audience*, he looked at the way people of different occupations, social class statuses, and ethnicities interpreted a television news program differently. Even a supposedly 'objective' news show, he found, was open to different interpretations made by different types of audiences. In *Family Television*, Morley took the family locale of the television audience seriously. He and his research team went into the homes of families and both interviewed them about television and observed them watching it. Through this research, Morley was able to convey about the gender dynamics of the family audience for television. One of the phenomena he commented on was the way husbands rather than wives often commandeered the remote control, enabling them to make many of the family viewing decisions.

Morley's work, and the work of other cultural studies researchers, helped to shift scholarly thinking about the media audience from a scientific paradigm, which attempted to measure audience exposure and effects, to a more holistic one which looks at audiences in the context of their everyday lives. As a result of this paradigm shift, the impact of media on audiences became viewed as a broad phenomenon. Rather than searching for more narrow, measurable influences of media, cultural researchers began examining the narratives embedded within media, and looking at how these are interpreted and indeed adopted by audience members in the course of many activities in their lives. Audience research became less focused solely on 'audiences' per se, and has expanded to include the many uses of culture and media we all have in postmodern society. For example, in *The Export of Meaning*, Liebes and Katz looked cross-culturally at the ways in which audiences in different countries, and members of different ethnic and religious groups within each country, interpreted the same episodes of the prime-time soap opera Dallas, which had become a global phenomenon, being broadcast worldwide and achieving an avid following in many different national and cultural contexts. They found striking ethnic and religious group differences in how these audiences read the very same episodes of this popular television show. For example, Israeli Arabs and Russian immigrants were defensive about the American way of life pictured in the show, and attempted to shield their children from it, while others in cultures closer to that pictured read it more as nonthreatening, simple entertainment.

Third Generation Audience Research

The intellectual turn toward 'third generation' research with its focus on 'postmodern' audiences began in cultural studies during the late 1980s and, according to Pertti Alasuutari (1999), emphasizes that "… there isn't really such a thing as the 'audience' out there; one must bear in mind that the audience is, most of all, a discursive construct produced by a particular analytic gaze" (p. 6). However, the definitive statement regarding the place of the postmodern audience in media studies did not emerge until the mid-1990s when Ien Ang (1995) began promoting a critical ethnographic methodology in response to the theoretical needs created by the changing nature of media audiences. Echoing the conception of postmodernity recognizing reality to be an unstable proposition, Ang notes that the state of knowledge regarding television audiences is always in flux. She writes, "What matters is not the certainty of knowledge about audiences, but an ongoing critical and intellectual engagement with the multifarious ways in which we constitute ourselves through media consumption" (Ang, 1995: p. 52). Furthermore, she adds, "Acknowledging the inevitably partial (in the sense of unfinished and incomplete) nature of our theorizing and research would arguably be a more enabling position from which to come to grips with the dynamic complexity and complex dynamics of media consumption practices" (Ang, 1995: p. 67). This postmodern position slowly became the centerpiece of current conceptions of the audience.

Despite the diversity of empirical analyses associated with such work, three basic strains of reception research can be considered as third generation (Grindstaff and Turow, 2006). The first type of reception study attempts to understand the 'decoding' processes associated with television programming; Jhally and Lewis' (1992) *Enlightened Racism* examining audience reception of *The Cosby Show* is an example. They find that reception of the 'Cosby Show' varies with race: African-American viewers are largely ambivalent about the show because it challenges racial stereotypes while simultaneously displaying a standard of living out of sync with the material realities of the black audience, while white viewers are largely positive about the show because it reinforces the American myth of a meritocratic society and thus downplays the continued existence of structural barriers based on race. In *African American Viewers and the Black Situation Comedy: Situating Racial Humor*, Robin Means Coleman (1996) builds on the work of Jhally and Lewis by exploring the diversity associated with a qualitative examination of African-American reception of racially charged situation comedies. The data indicate that some African-Americans think mainstream society believes the representations on television create negative consequences for African-Americans, others think the representations of African-Americans have a positive impact, and others do not think there is a relationship between the representations of African-Americans and the perceptions of mainstream society. The second type of reception work is primary focused on the behavior and interpretations of television fans like Henry Jenkins' (1992) *Textual Poachers*. Borrowing Michel de Certeau's concept of 'poaching,' Jenkins vehemently rejects the dominant stereotypes of fans and shows that fans use material borrowed from popular media to establish subcultures. Jenkins continues this work and adapts his thinking to the new media environment in his more recent book, *Convergence Culture* (2006), which is discussed below.

The last type of reception research examines the consumption of media in the context of a larger social milieu. For example, Andrea Press and Elizabeth Cole (1999) use focus groups to examine the relationship between class subjectivity and the reception of television content. The study produces three important findings: (1) the discourse of pro-choice

middle-class women mirrors the discourse of prime-time television in which abortion is presented as appropriate for only the lower classes; (2) attitudes among pro-choice working-class women vary by class identification; and (3) class divisions are less important than shared beliefs among pro-life women. These findings indicate that the role of categorical identity in audience reception is more complex than previously believed. Similarly, in *New Media Audiences* (1999), Ellen Seiter uses a variety of qualitative methods to paint a more complex picture of the relationship between television and various segments of the American middle-class audience. In particular, Seiter (1999) asserts that middle-class individuals with lower levels of socioeconomic status and higher levels of educational attainment "cannot afford the luxury of indulging a love of popular culture, for fear that they will appear uneducated" (Seiter, 1999: p. 131). As the title of Seiter's (1999) book indicates, however, by the end of the twentieth century scholars were already beginning to confront 'new media audiences.'

New Media Audiences

Initially, theoretical work responded to traditional media's 'diversifying in form and contents' by interpreting new media audiences through a postmodern framework asserting digital technology only exacerbated existing trends making audiences 'less predictable, more fragmented, and more variable' (Livingstone, 1999: p. 63). Yet, rather quickly, many scholars began claiming that something was different about new media audiences. For example, in the influential *Convergence Culture: Where Old and New Media Collide*, Henry Jenkins (2006) argues that contemporary audiences are now defined by their willingness to 'make connections among dispersed media content,' and, as a result, 'consumption has become a collective process' (pp. 3–4). Although Jenkins' theoretical work is widely regarded as significant, much of the earliest empirical research addressing new media audiences were largely guided by the belief that digital age inequality could be addressed through exclusively structural means like expanding fiber-optic networks to ensure the Internet access (for an excellent review of this research, see Livingstone, 2005).

More recently, however, the focus has shifted from issues of access to issues of usage and cultural engagement. Cross-national research examining the relationship between school-aged children and the Internet usage confirms that behavior tends to vary with socioeconomic location. For example, Sonia Livingstone's (2009) mixed-method examination of school children in the United Kingdom finds that for some "the internet is an increasingly rich, diverse, engaging, and stimulating resource of growing importance in their lives; for others, it remains a narrow and relatively unengaging if occasionally useful resource" (Livingstone, 2009: p. 57–58). Yet, access only explains some of the inequality created by new media. As Seiter (2005) notes in an ethnographic study of children and the Internet, closing the technology gap is an 'easier task to address than the wider and deeper deprivations' in society at large (p. 101). Even if access is unconstrained by economic concerns, audience behavior would remain stratified because the availability of the cultural competencies needed to interact with new media is related to socioeconomic status.

Research examining the relationship between young adults and new media finds similarly class-stratified engagement. In her research, Eszter Hargittai and her coauthors find that the tendency to accumulate cultural capital online is related to socioeconomic status using a variety of quantitative data. In a study examining the online behavior of individuals belonging to 18–26 years of age, for example, Hargittai and Hinnant (2008) observe that individuals with higher levels of education and those from materially privileged backgrounds use the Internet for 'capital-enhancing activities,' defined as the frequency with which individuals visit Web sites related to high-status Internet consumption such as political news, economic news, health news, stock prices, or travel information. Recent research addressing the use of social networking Web sites produces similar findings. For example, Danah Boyd (2011) examined the changing cultural preferences associated with the massive shift from MySpace to Facebook; she concludes by noting, "In some senses, the division in the perception and use of MySpace and Facebook seems obvious given that we know that online environments are a reflection of everyday life. Yet, the fact that such statements are controversial highlights a widespread techno-utopian belief that the internet will once and for all eradicate inequality and social divisions" (2011: p. 37).

Like Boyd, many scholars examining new media audiences reject the conceptual division separating mediated behavior from everyday practice. For example, in *Personal Connections in the Digital Age* (2010), Nancy Baym argues that, rather than conceptualizing social cues as something lost when one moves from 'real' to mediated social life, they should be treated as variable; media providing more cues are considered 'rich,' whereas those providing fewer clues are considered 'lean' (2010: p. 9). This work is rather characteristic of empirical analyses of new media audiences in that a social constructivist approach is married with a soft technological determinism to acknowledge the innovative uses of new media and the tendency for existing social relations to be recreated online.

Feminist Audience Research

Feminist audience research now constitutes a sizable body of work with its own history, subfields, and criticisms. Feminist audience study began at the theoretical level with Laura Mulvey's seminal piece about film spectatorship, *Visual Pleasure and Narrative Cinema* (1975), in which Mulvey challenged the notion that audiences unilaterally watch classical Hollywood film. Using the concepts of psychoanalytic theory, Mulvey develops the notion that a 'male gaze' dominates the spectator position in classical Hollywood cinema. Framing gendered audiences as distinct and worthwhile of separate attention pioneered decades-long discussion on the way women view and interact with media.

Later feminist audience analysis broadened Mulvey's discussion of female audiences by including both traditional media as well as specialized 'female' genres. One of the most influential authors to take 'women's media' seriously was Janice Radway. In *Reading the Romance: Women, Patriarchy, and Popular Literature*, she provided a powerful understanding of the way Midwestern women used romance novels, which like

much women's media had formerly been considered too 'trashy' for scholarly attention, as a form of escape from the daily demands of their lives. Using interviewing and ethnography of a group of romance readers, Radway illustrated how women fans found great value in romance novels, admiring the portrayal of strong and independent heroines.

Radway's work was innovative for a myriad of reasons, including her qualitative social scientific study of the *audience* in addition to the text, and her focus on 'women's' genre previously deemed trivial for scholastic attention. By talking with women on their own terms and using their own language and space, Radway was able to understand how women used romance novels to evade household duties and claim time to indulge their own needs and fantasies.

Press (1991) continued in Radway's tradition with her influential book, *Women Watching Television: Gender, Class, and Generation in the American Television Experience*, exploring the ways in which women watch and connect with television. Rather than assuming 'one' female audience, Press sought to understand the way class and generation varied the viewing experiences by interviewing women of different ages in both the working class and the middle class. In doing so, Press documents the differing role that television plays in reinforcing our culture's hegemonic values, finding that "[P]opular television images represent certain social groups, issues, and institutions systematically and repetitively in a manner that often reflects the position of these groups within our society's hierarchical power structure" (1991: p. 27). These findings were a critical juncture for feminist audience research as emerging researchers began combining the contributions of both Press and Radway. While Radway's work validated that female genres warrant scholarly attention, Press brought to light that women's audiences are not unidimensional and that other variables, such as socioeconomic status and generation, must be considered.

Other scholars have explored how 'postfeminist' audiences contextualize and interact with various media. Drawing on Judith Butler's (1990) theory that sexual, economic, and racial facets complicate the category of 'woman,' postfeminism emphasizes a multidimensional female audience. Postfeminist theory suggests that feminist audience researchers must develop a complex theory of the relationship between culture, politics, agency, and women's consumption (Tasker and Negra, 2007).

A leader in researching and theorizing postfeminist audiences, Angela McRobbie questioned how women in today's media environment interact with content that assumes feminism as a 'spent force' that must be both 'taken into account' but framed as though it has already 'passed away' (McRobbie, 2004: p. 256). Her recent works (2007, 2009) focus on the way female audiences negotiate education, earnings, sexual power, and fashion to maintain an aura of femininity while challenging hegemonic masculinity. In particular, her contribution to audience research conceptualizes the absence of critique from today's female audiences who may not have the power to react when their voice is "called upon to be silent … as a condition to her freedom" (McRobbie, 2007: p. 34).

Other scholars have also looked at the way in which postfeminist audiences interact with media. Akass and McCabe (2004) document fans' reaction to the show *Sex in the City* through the lens of feminist critique. New media scholars (e.g., White, 2006, 2009, 2010) have explored 'postfeminist' expression in the public sphere through online forums. Tasker and Negra (2007) allege that postfeminist culture is "exemplified by the figure of a white, middle-class, heterosexual woman" and does not take into account the varied experiences of women. International scholars (Lemish, 2010; Lemish and Reznik, 2008; Parameswaran, 2002; Hegde, 2011) have attempted to bridge this gap by broadening the idea of female audience to look not only at the U.S. and Western audiences, but to look more globally and to understand issues of transnationalism and imperialism as they affect female audiences worldwide.

Global Audiences

The study of global audiences starts largely in response to the 'cultural imperialism' thesis that analyzes media as an instrument of cultural domination, reflecting uneven flows and exchanges between the West and the rest of the world. Canonical studies of cross-cultural audiences (Ang, 1995; Liebes and Katz, 1990; Gripsrud, 1995) challenge the simplicity of this rhetoric based on the misconception of passivity on the part of the local audiences (for a more detailed critique of the cultural imperialism thesis, see Morley, 2006). Far from producing a homogeneous global culture, audience studies have shown that media products moving across national boundaries are often reinterpreted according to local cultural grids. By insisting that international audiences were neither powerless nor uncritical to American television shows, this pioneering research further supports active audience theory on a global scale. However, despite the fact that global audiences have agency to reinterpret or reject what is offered, they have rather limited power to decide which programs are imported in the first place. The United States is still the major media exporter to the global market, even though its relative dominance is declining. The postcolonial approach reminds international media scholars that they must balance a romanticized vision of audience empowerment with a reconstructed politic of media manipulation.

McMillin (2007) notes that postcolonial and critical audience research needs to move beyond the 'cultural imperialism' frame to ask a wider range of questions about international audiences. Based on a mobile and multisited ethnography, Abu-Lughod (1999) studies Egyptian TV drama as a key institution that stitches audiences together for the production of national culture. She watches television and discusses its meanings with two different groups: (1) women in the villages of a rural, underdeveloped region of Upper Egypt and (2) women who work as domestic servants in Cairo. These television dramas provide common reference points for their audiences who contribute to national affinity. Drawing on recent scholarship on the politics and pleasures of reality television, Punathambekar (2010) studies a series of events surrounding the third season of *Indian Idol* to explore how reality television, situated to the changing landscape of Indian television, has enabled new modes of cultural and political expression. In their consideration of the *Super Girls' Voice*'s audience

(a Chinese counterpart of *The American Idol*), Cui and Lee (2010) focus on the issue of whether the reality show has altered the power relationship between the media and the audience. Most of their focus group discussants enjoyed watching the 'super girls' for their ordinariness, but still derived satisfaction from seeing how these ordinary girls became extraordinary through the reestablishment of the media/ordinary boundary.

One of the major trends of the last decade has been the 'regionalization' of media productions and reception within the multicountry markets linked by geography, language, and culture. For instance, Japanese pop cultures and Korean TV dramas have been startlingly successful in the East Asian region since the late 1990s. The regional media flows open up new possibilities to study non-Western audiences in terms of their perceptions of colonial–postcolonial relationships, gender subjectivity constructions, and nationalist responses to transnational media culture exchange. Lin and Tong (2008) conducted a comparative study of the viewing practices of female fans in Hong Kong and Singapore. Based on in-depth interviews with the fan groups, they found that these female audiences are adept in using Korean dramas to construct what they see as their distinctive 'Asian' modern femininities. While both groups appreciated the hybridization of traditional values and modern images in Korean dramas, the female viewers in Singapore seemed to be more inclined toward traditional values and Confucian ideals of femininities than the Hong Kong viewers. Through interviews with middle-aged female fans in Japan, Mori (2008) found that the penetration of Korean TV dramas into actual Japanese family life has political potential for understanding and reconstructing Korea–Japan relationship in new ways. These studies demonstrate the emergence of regional media culture and offer new conceptual insights into global audience studies, which cannot be ascertained through previous US-centric analyses.

Besides the global flow of media production and consumption, audience studies can also be used to examine how migrants, multinational citizens, and other cosmopolitans use the media to negotiate their identities and transnational connections. Gillespie (2006) started a collaborative ethnographic study of the news-viewing practices among multilingual households in the United Kingdom. Interviewees watched the attacks of September 11 on multiple news channels and in a variety of languages, then discussed with researchers their everyday media practices. Given the perceived bias in Western reporting, these transnational viewers actively sought alternative news sources through the use of satellite television and the Internet. This study contributes to the scholarship on the emergence of transnational audiences, publics, and identities.

There is a difference in emphasis, however, between the transnational studies of traditional media (e.g., TV programs and videos) and the research centered on the use of new media (e.g., mobile phones and the Internet). As Lewis and Hirano (2001) found in their study of Thai-Australia families, watching rented Thai videos serves as a site of escapism for Thai brides who are culturally and socially isolated. Although their use of ethnic video drama helps Thai brides to cope with their nostalgic sentiments, it does not change the marginalized social status that they often enjoy in their host country. In contrast, a number of scholars (e.g., Hiller and Franz, 2004; Chan, 2010) who studied how migrants use new communication technologies, especially the use of Internet, suggest that migrant groups benefit concretely from their online activities in terms of cultivating social ties in both home and host countries, and forming cyber-communities that facilitate their social adaptation to the new environment while maintaining strong attachments to the homeland. These studies have illustrated the value of studying audiences or the users in the new media context as a key way to understand the transnational identities and lived realities of diasporas.

As noted in previous examples, cross-national research proves to be one of the most fruitful research approaches in transnational media studies. Based on in-depth interviews conducted in 2001 and 2004, Elias (2008) studied how Russian-speaking returnees from the former Soviet Union use electronic and print media in Israel and Germany. A parallel comparison of returnees in Israel and Germany afforded an interesting case to examine the immigrants' media consumptions in different structural and cultural contexts. The findings showed that these immigrant groups used the media in very different ways. Characterized by assimilative media consumption, returning immigrants in Germany limited themselves nearly exclusively to consumption of German-language media. However, German-speaking media did not guarantee 'a free ticket into German society' (p. 138). These immigrants felt alienated and isolated from German culture even when speaking German. In contrast, Russian-speaking Jews seek out various forms of Russian-speaking media in Israel. Both newcomers (1–3 years in residence) and old-timers (5 years in residence), who get information through Russian-speaking media, however, do feel at home in Israel. This study points out the importance of political, ideological, and cultural factors, which condition both media consumption of the immigrants and their integration into the host society. While traditional ethnography focuses on particular ethnic groups in a specific geographic place, a revised ethnographic approach targeting international audiences needs to draw out the wide connections between local experience and the global context in which their cultural identity is located.

Directions for Future Research

Regarding television audiences specifically, the emergence of cable technology at the end of the network era fundamentally altered the ways in which content becomes socially important. In describing the shift from the multichannel transition period, which began in the 1980s, to the postnetwork era in the early 2000s, Lotz (2007) notes that the emergence of so many new networks and niche channels fundamentally changed the economics of the television industry. Also, rather than delivering large audiences to advertisers within this new environment, producers, like advertiser-supported cable networks, can afford to create content that will only be watched by 1% of the available audience (Lotz, 2007: p. 37). As a consequence of these economic and cultural realities, postnetwork television

differs from the programming of the network era in that the cultural relevance of any content ceases to be a direct function of audience size, and producers are no longer limited to creating content that appeals to 'a multiplicity of social types at once' (Gitlin, 1982: p. 248).

With such changes, the medium has become increasingly legitimate as a cultural form with the production of targeted content appealing to high-income, educated niche audiences. A recent Harvard course centered on HBO's The Wire (Chaddha and Wilson, 2010) confirms Lotz's assertion that only in the context of television's postnetwork era can a particular content be affirmed by "hierarchies of artistic taste and social importance" and become "imbued with what Pierre Bourdieu terms 'cultural capital'" (Bourdieu, 1984; quoted in Lotz, 2007: p. 40). According to Newman and Levine (2012), however, the cultural legitimation of quality television content relies on the same imbalanced binaries that are associated with the medium's historical degradation. As a result, television is now bifurcated. Content reminiscent of network era programming becomes 'bad' television in the postnetwork era because of its association with the passive, 'feminized' viewing experiences of mass audiences; while television shows like *The Sopranos* and *The Wire* become 'good' television because of their association with the active, 'masculine' viewing experiences of elite niche audiences. In contrast with previous research examining the relationship between social class and television audiences (Press, 1991; Morley, 1980, 1988; Seiter, 1999), however, the relationship between social class and postnetwork audiences remains largely unexplored.

Furthermore, as audience researchers we must also acknowledge that the television landscape is changing, blurring the lines between producer and consumer (Meyrowitz, 2010). One of the fastest growing sectors of participatory television is the genre of 'reality television,' a programmatic landscape that purports unscripted, real-life glances at the lives of 'ordinary' people (Holmes and Jermyn, 2004). To date, most scholarship emphasizes the way reality television reinforces class differences and reveals the role television continues to play in constructing hegemonic discourse (Lizardo, 2010; Grindstaff, 2002; Couldry, 2011; Andrejevic, 2011). More attention must be paid to *biopic* reality television, formats that mimic documentaries. Since biopic narratives are situated in identifiable communities, they provide the opportunity for understanding how *indirect participants* negotiate their cultural and class position in relation to reality television. These 'integrated audiences' have friends, relatives, and members of their community featured on national television, but still act as an audience, *consuming* the final product (Tripodi forthcoming).

Conclusion

With the changing media environment and changing media genres and formats, the shape, structure, and even the very definition of the media 'audience' keeps changing in ways that acknowledge the more active and interactive roles that audiences play in the new media environment. As Barker and Mathijs (2012) find in a multimethod research project spanning a dozen countries and generating 25 000 responses, "cultural product such as films are not message-vehicles, to be assessed for their lesser or greater 'effect,' but complexly organized bodies of meaning, which draw on and react back onto their constitutive culture" (676). They also note the increasing importance of ancillary texts as "news, reviews, gossip, leaks, publicity, posters, merchandising, etc., work in complicated ways, and with different groups of people, to steer, influence, or create emergent frames for receiving and affiliating with as a film such as *Lord of the Rings*" (679). Nonetheless, the term 'audience' itself is progressively less used as scholars confront the issue of how to study media influence as media become more fragmented and more interactive.

The 'mass' sense in which we thought of the media audience in the first decades of our field has now evolved into a sense that most audiences are 'niche' audiences of various sorts, and that all audiences are active to some degree, though some are more active, and interactive, than others. This necessarily changes, and is continually changing, the sorts of questions it's important for audience researchers to ask; the design of audience research studies; the way we think about the relationships between production, reception, activity, and passivity in the context of media audience study, and the way we think about the nature of media influence itself, and how we should conceptualize this. The study of media audiences, however, continues to be a central and important part of media research more generally, and will continue to be so even as the media environment, media genres, and the relationship between media production and reception continue to evolve.

See also: Audience Measurement; British Cultural Studies; Mass Media, Representations in; Public Sphere and the Media; Television: General.

Bibliography

Abu-Lughod, Lila, 1999. The interpretation of culture(s) after television. In: Ortner, Sherry (Ed.), The Fate of "Culture": Geertz and eyond. University of California Press, Berkeley, pp. 110–135.

Akass, Kim, McCabe, Janet, 2004. Reading Sex and the City. I.B.Tauris & Co Ltd, London.

Alasuutari, Pertti, 1999. Rethinking the Media Audience: The New Agenda. Sage Publishing, London.

Andrejevic, Mark, 2011. Managing the borders: classed mobility on security-themed reality TV. In: Skeggs, Beverly, Wood, Helen (Eds.), Reality Television and Class.

Ang, Ien, 1995. Living Room Wars: Rethinking Media Audiences for a Postmodern World. Routledge, New York.

Barker, Martin, Mathijs, Ernest, 2012. Researching world audiences: the experience of a complex methodology. Participations 9 (2), 664–689.

Baym, N.K., 2010. Personal Connections in the Digital Age. Polity Press, Malden, MA.

Bourdieu, Pierre, 1984. Distinction: A Social Critique of the Judgement of Taste. Harvard University Press, Cambridge, MA.

Boyd, Danah, 2011. White flight in networked publics? How race and class shaped American teen engagement with MySpace and facebook. In: Nakamura, Lisa, Chow-White, Peter (Eds.), Digital Race Anthology. Routledge, London and New York, pp. 203–222.

Butler, Judith, 1990. Gender Trouble: Feminism and the Subversion of Identity. Routledge, New York and London.

Carr, Caleb T., Hayes, Rebecca A., 2014. The effect of disclosure of third-party influence on an opinion leader's credibility and electronic word-of-mouth in two-step flow. Journal of Interactive Advertising no. just-accepted: 00–00.

Case, Donald O., David Johnson, J., Andrews, James E., Allard, Suzanne L., Kelly, Kimberly M., 2004. From two-step flow to the internet: the changing array of sources for genetics information seeking. Journal of the American Society for Information Science and Technology 55 (8), 660–669.

Chaddha, Anmol, Wilson, William Julius, 12 September 2010. Why We're Teaching the Wire at Harvard. The Washington Post.
Chan, Brenda, 2010. The internet and new chinese migrants. In: Alonso, Andoni, Orarzabal, PedroJ. (Eds.), Diasporas in New Media Age: Identity, Politics, and Community. University of Nevada Press, Reno & Las Vegas, pp. 242–264.
Couldry, Nick, 2011. Class and contemporary forms of 'Reality' production or, hidden injuries of class 2. In: Skeggs, Beverly, Wood, Helen (Eds.), Reality Television and Class. Palgrave Macmillan, Basingstoke, pp. 33–44.
Cui, Li, Lee, Francis, 2010. Becoming extra-ordinary: negotiation of media power in the case of super girls' voice in China. Popular Communication 8, 256–272.
Elias, Nelly, 2008. Coming Home: Media and Returning Diaspora in Israel and Germany. State University of New York Press.
Farnsworth, Stephen, Soroka, Stuart, Young, Lori, 2010. The international two-step flow in foreign news: Canadian and US television news coverage of US affairs. The International Journal of Press/Politics 15 (4), 401–419.
Gillespie, Marie, 2006. Transnational television audiences after september 11. Journal of Ethnic and Migration Studies 32 (6), 903–921.
Gitlin, Todd, 1978. Media sociology: the dominant paradigm. Theory and Society 6 (1), 205–254.
Gitlin, Todd, 1982. Television's screens: hegemony in transition. In: Apple, Michael W. (Ed.), Cultural and Economic Reproduction in Education: Essays on Class, Ideology, and the State. Routledge & Kegan Paul, Boston, MA.
Grindstaff, Laura, 2002. The Money Shot. Trash, Class, and the Making of TV Talk Shows. The University of Chicago Press, Chicago.
Grindstaff, Laura, Turow, Joseph, 2006. Video cultures: television sociology in the 'new TV' age. Annual Review of Sociology 32, 103–125.
Gripsrud, Jostein, 1995. The Dynasty Years: Hollywood Television and Critical Media Studies. Taylor and Francis, London.
Hall, Stuart, 1973. Encoding and decoding in the television discourse. In: Durham, Kellner (Eds.), Media and Cultural Studies (KeyWorks in Cultural Studies), pp. 163–173.
Hargittai, Eszter, Hinnant, Amanda, 2008. Digital inequality: differences in young adults' use of the internet. Communication Research 35 (5), 602–621.
Hegde, Radha (Ed.), 2011. Circuits of Visibility: Gender and Transnational Media Culture. New York University Press, New York.
Hiller, Harry H., Franz, Rara M., 2004. New ties, old ties, and lost ties: the use of the internet in diaspora. New Media & Society 6, 731–752.
Holmes, Su, Jermyn, Deborah, 2004. Understanding Reality Television. Routledge, New York.
Jenkins, Henry, 1992. Textual Poachers: Television Fans & Participatory Culture. Routledge.
Jenkins, Henry, 2006. Convergence Culture: Where Old and New Media Collide. New York University Press, New York.
Jhally, Sut, Lewis, Justin, 1992. Enlightened Racism: The Cosby Show, Audiences, and the Myth of the American Dream. Westview Press, New York.
Juluri, Vamsee, 2003. Becoming a Global Audience: Longing and Belonging in Indian Music Television. Peter Lang, New York.
Katz, Elihu, Lazarsfeld, Paul, 1955. Personal Influence: The Part Played by People in the Flow of Mass Communication. The Free Press, Glencoe, IL.
Lemish, Dafna, 2010. Screening Gender on Children's Television. Routledge, Abington and New York.
Lemish, D., Reznik, S., 2008. To laugh or not to laugh? Performing girlhood through humor. Girlhood Studies 1 (1), 114–137.
Lewis, G., Hirano, C., 2001. Mi arai mai mai mai? Thai-Australian video ways. In: Cunningham, S., Sinclair, J. (Eds.), Floating Lives: The Media and Asian Diasporas. Rowman & Littlefield Publishers, Inc, New York, pp. 185–216.
Liebes, Tamar, Katz, Elihu, 1990. The Export of Meaning: Cross-cultural Readings of DALLAS. Oxford University Press, New York.
Lin, Angel, Tong, Avin, 2008. Re-imagining a cosmopolitan 'Asian us': Korean media flows and imaginaries of asian modern femininities. In: Beng Huat, Chua, Iwabuchi, Koichi (Eds.), East Asian Pop Culture: Analysing the Korean Wave. Hong Kong University Press, pp. 91–125.
Livingstone, Sonia, 1999. New Media, new audiences? New Media & Society 1 (1), 59–66.
Livingstone, Sonia, 2005. Critical Debates in Internet Studies: Reflections on an Emerging Field. LSE Research, London. Online. http://eprints.lse.ac.uk/1011/1/CriticaldebatesTODAY.pdf.
Livingstone, Sonia, 2009. Children and the Internet. Polity, London.
Lizardo, Omar, 2010. Culture and stratification. In: Hall, John R., Grindstaff, Laura, Lo, Ming-Cheng (Eds.), Handbook of Cultural Sociology. Routledge, New York.
Lotz, Amanda, 2007. The Television Will Be Revolutionized. New York University Press.
McMillin, Divya, 2007. International Media Studies. Wiley-Blackwell, Malden, MA.
Meyrowitz, Joshua, 2010. Media evolution and cultural change. In: Hall, John R., Grindstaff, Laura, Lo, Ming-Cheng (Eds.), Handbook of Cultural Sociology. Routledge, New York.
Means Coleman, Robin, 1996. African American Viewers and the Black Situation Comedy: Situating Racial Humor. Routledge, New York.
McRobbie, Angela, 2004. Post-feminism and popular culture. Feminist Media Studies 4, 255–264.
McRobbie, Angela, 2007. Top girls? young women and the post-feminist sexual Contract. Cultural Studies 21 (4), 718–737.
McRobbie, Angela, 2009. The Aftermath of Feminism: Gender, Culture and Social Change. SAGE, London and Thousand Oaks.
Morley, David, 1980. The 'Nationwide' Audience. British Film Institute, London, UK.
Morley, David, 1986. Family Television: Cultural Power and Domestic Leisure. Comedia Pub. Group, London.
Morley, David, 1988. Family Television: Cultural Power and Domestic Leisure. Routledge, New York.
Morley, David, 2006. Globalisation and cultural imperialism Reconsidered: old questions in new guises. In: Curran, James, Morley, David (Eds.), Media and Cultural Theory. Routledge.
Mori, Yoshitaka, 2008. Winter Sonata and cultural practices of active fans in Japan: considering middle-aged women as cultural agents. In: Beng Huat, Chua, Iwabuchi, Koichi (Eds.), East Asian Pop Culture: Analysing the Korean Wave. Hong Kong University Press, pp. 127–142.
Mulvey, Laura, 1975. Visual pleasure and narrative cinema. Screen 16 (3), 6–18.
Newman, M., Levine, E., 2012. Legitimating Television: Media Convergence and Cultural Status. Routledge, New York.
Neuman, W. Russell, Guggenheim, Lauren, 2011. The evolution of media effects theory: a six-stage model of cumulative research. Communication Theory 21, 169–196.
Parameswaran, Radhika, 2002. Reading fictions of Romance: gender, sexuality and nationality in postcolonial India. Journal of Communication 52 (4), 832–852.
Punathambekar, Aswin, 2010. Reality TV and participatory culture in India. Popular Communication 8, 241–255.
Press, Andrea, 1991. Women Watching Television: Gender, Class, and Generation in the American Television Experience. University of Pennsylvania Press, Pennsylvania.
Press, Andrea L., Cole, Elizabeth R., 1999. Speaking of Abortion: Television and Authority in the Lives of Women. University of Chicago Press, Chicago.
Press, Andrea L., Cole, Elizabeth R., 2001. Speaking of Abortion: Television and Authority in the Lives of Women. University of Chicago Press, Chicago, IL.
Press, Andrea L., Williams, Bruce A., 2010. The New Media Environment: An Introduction. Basil Blackwell, Malden, MA and Oxford.
Radway, Janice, 1991. Reading the Romance. University of North Carolina Press, North Carolina.
Seiter, Ellen, 1999. Television and New Media Audiences. Oxford University Press, New York.
Seiter, Ellen, 2005. The Internet Playground: Children's Access, Entertainment, and Mis-education. Peter Lang Publishing, New York.
Tasker, Yvonne, Negra, Diane, 2007. Interrogating Post-Feminism. Gender and the Politics of Popular Culture. Duke University Press, North Carolina.
White, Michele, 2006. The Body and the Screen: Theories of Internet Spectatorship. MIT Press, Massachusetts.
White, Michele, 2009. Networked bodies and extended corporealities: theorizing the relationship between the body, embodiment, and contemporary new Media. Feminist Studies 35 (3).
White, Michele, 2010. What a mess: eBay's narratives about personalization, heterosexuality, and disordered homes. Journal of Consumer Culture 10 (1).

Auditory Scene Analysis: Computational Models

Martin Cooke, University of the Basque Country, Vitoria, Spain

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by G.J. Brown, volume 2, pp. 942–946, © 2001, Elsevier Ltd.

Abstract

Listeners have to make sense of a complex acoustic world containing overlapping sound sources that must be organized into individual auditory objects. Computational auditory scene analysis concerns the use of algorithms inspired by human sound perception whose aim is to extract properties of constituent sound sources in a complex mixture. Starting with representations based on models of how sound is processed in the peripheral auditory system, typical computational auditory scene analysis techniques function by decomposing the mixture into components followed by selective recomposition into groups of components that appear to emanate from a single source. Grouping processes can be informed by information in the signal itself or by the use of prior statistical models of sound sources. This article outlines some of the principal signal decompositions used in models of auditory grouping and goes on to describe a decoder that combines both signal- and model-driven grouping processes.

Introduction

We live in an acoustically cluttered world and are often forced to pick out the sound source of interest from the mixture reaching our ears. Computational auditory scene analysis (CASA) is the branch of machine perception that attempts to extract information about the location, identity, and content of individual acoustic sources in a mixture (Lyon, 1983; Brown and Cooke, 1994; Ellis, 1999; Wang and Brown, 2006). CASA algorithms differ from other engineering approaches to source separation by being both inspired and constrained by human perceptual organization of sound. In practice, this means that CASA algorithms operate with signals from one or, at most, two microphones; employ a first stage of signal analysis using a computational model of the peripheral auditory system; and embody auditory grouping processes informed by auditory scene analysis.

CASA grew in part from investigations of the perceptual organization of sound in listeners (Bregman, 1990; Darwin, 1981), although early research in the nascent field of CASA (e.g., Cooke, 1993) was also influenced by computer vision, especially the work of Marr (1982). Computational techniques in hearing and vision distinguish between bottom-up 'signal-driven' processes that extract information from sensor data, and top-down 'model-driven' processes that use prior knowledge of objects to interpret features derived from the signal. However, there are important differences between the two modalities that are reflected in computational approaches in each field. While vision systems routinely handle large numbers of spatially separated objects, the number of individual acoustic sources of interest is usually far smaller, but the separation problem is more challenging since information about each source is overlapped in both time and frequency. Likewise, while attentional-controlled searches for salient objects in the spatial dimension comprise a key aspect of computer vision systems, temporal processing plays a more important role in accessing information in computational audition (e.g., following a conversation in the presence of other voices or tracking a single instrument in a polyphonic composition).

As described in the sections of this article, CASA algorithms operate by first decomposing the incoming sound mixture in order to transform the signal into domains where the constituent sources are more easily separable, then selectively recombining signal elements, often based on auditory grouping principles, where there is evidence that the elements originated from the same acoustic source.

Signal Decomposition

The decomposition stage converts one or two one-dimensional microphone signals into one or more two- or three-dimensional representations of sound, whose purpose is to more clearly reveal features of the individual acoustic sources such as periodicity or spatial location, thereby facilitating sound separation. Typically, decomposition proceeds by converting each microphone signal into an 'auditory spectrogram,' a representation of auditory nerve activity as a function of frequency and time. Example auditory spectrograms are shown in **Figure 1** for a female talker, a male talker, and a mixture of the two. Each image results from passing the signal through a bank of 100 bandpass filters whose shapes and center frequencies are first approximations to human auditory filter shapes derived from psychophysical studies (Glasberg and Moore, 1990). The temporal envelope in each frequency band has been extracted, smoothed, logarithmically compressed, and converted to a grayscale, in which darker areas correspond to more energetic sound components.

While the auditory spectrogram is primal in documenting the relative energy balance of the constituent sources in a mixture at each time and frequency point, fine timing information is lost in its construction, and other transformations are typically applied to the signal at the high temporal resolution output of each modeled auditory nerve fiber. The two most common transformations are based on auto- and cross-correlation, inspired by early physiological models of Licklider (1951) and Jeffress (1948), respectively. The 'correlogram' representation, shown in **Figure 2(a)**, is the

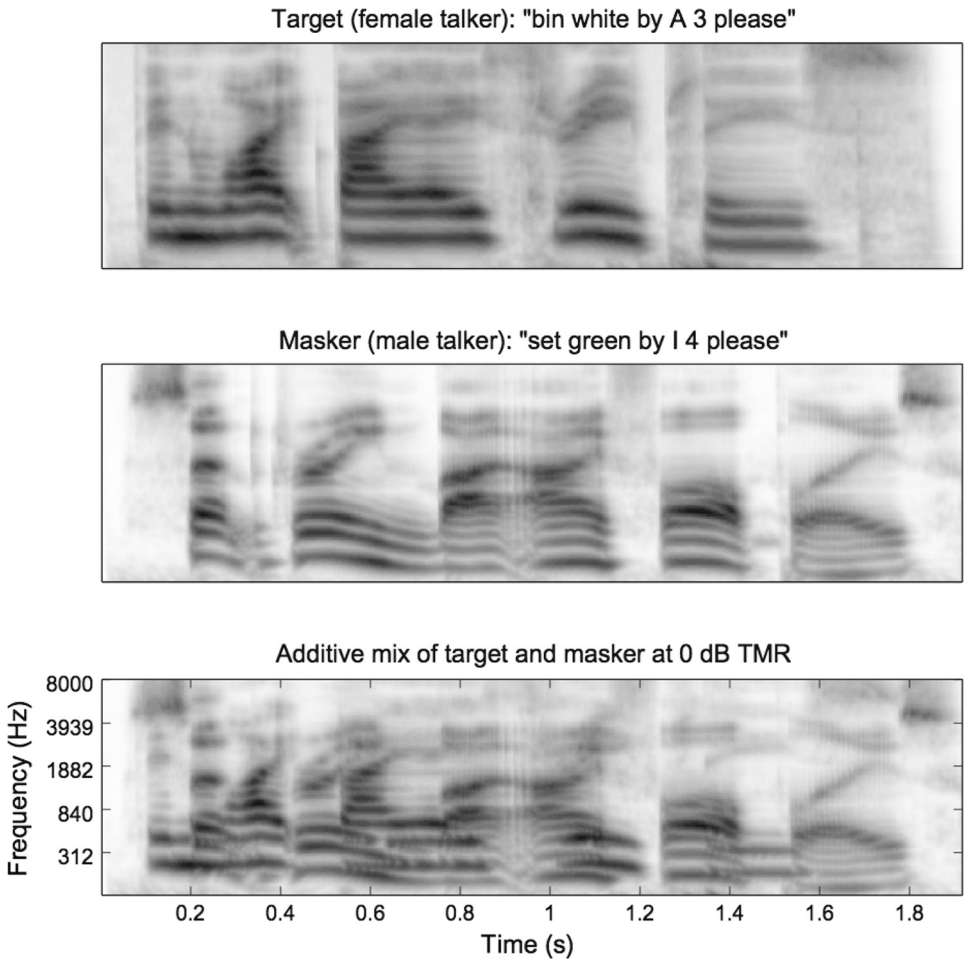

Figure 1 Auditory spectrograms for utterances by a female (top) and male (middle) talker and for their equal-energy mix (bottom).

collection of autocorrelation functions of auditory filter outputs, applied independently in each frequency band. In principle, the correlogram is a three-dimensional representation of sound, with dimensions of cochleotopic frequency, autocorrelation delay ('lag'), and time, but in practice it is computed on short overlapping segments of the signal. Autocorrelation enhances periodicities in each frequency region, and, as explained in this article, it is the across-frequency grouping of common periodicities that can be used to separate sound sources at those time instants where one or both are periodic, as is the case for speech signals at moments where the vocal cords are active, or for plucked instruments.

In a similar manner, pairs of microphone signals can be cross-correlated to highlight the delays at which the two signals are most similar, based on interaural time difference cues to the source azimuth. More recently, for sensors mounted on head simulators, CASA models (e.g., May et al., 2011) have attempted to reinforce time difference cues in the cross-correlogram with level difference estimates motivated by the observation that energy received at the sensor on the opposite side of the head to the source is substantially attenuated relative to that of the closest microphone. Other low-level signal transformations applied in CASA systems include the extraction of temporal envelope modulations, which can encode the fundamental frequency of periodic sources and other modulation components (e.g., Hu and Wang, 2004), as well as source onset and offset enhancement (e.g., Brown and Cooke, 1994).

Grouping: Recovering Individual Sound Sources by Selective Recombination

Having transformed the acoustic mixture into representations in which evidence for each individual source is more salient, the remaining task for CASA algorithms is to determine which elements belong together in order to reconstruct one or more sources. Early algorithms focused exclusively on the application of auditory grouping principles based on information in the signal itself. More recent approaches have incorporated prior knowledge of sound sources, often embodied in probabilistic models. We consider each in turn.

Signal-Driven Grouping

Figure 1 exposes a multiplicity of cues that might be used to recombine signal elements into source descriptions without prior knowledge of the individual sources. Considering the lower panel which depicts the mixture of male and female

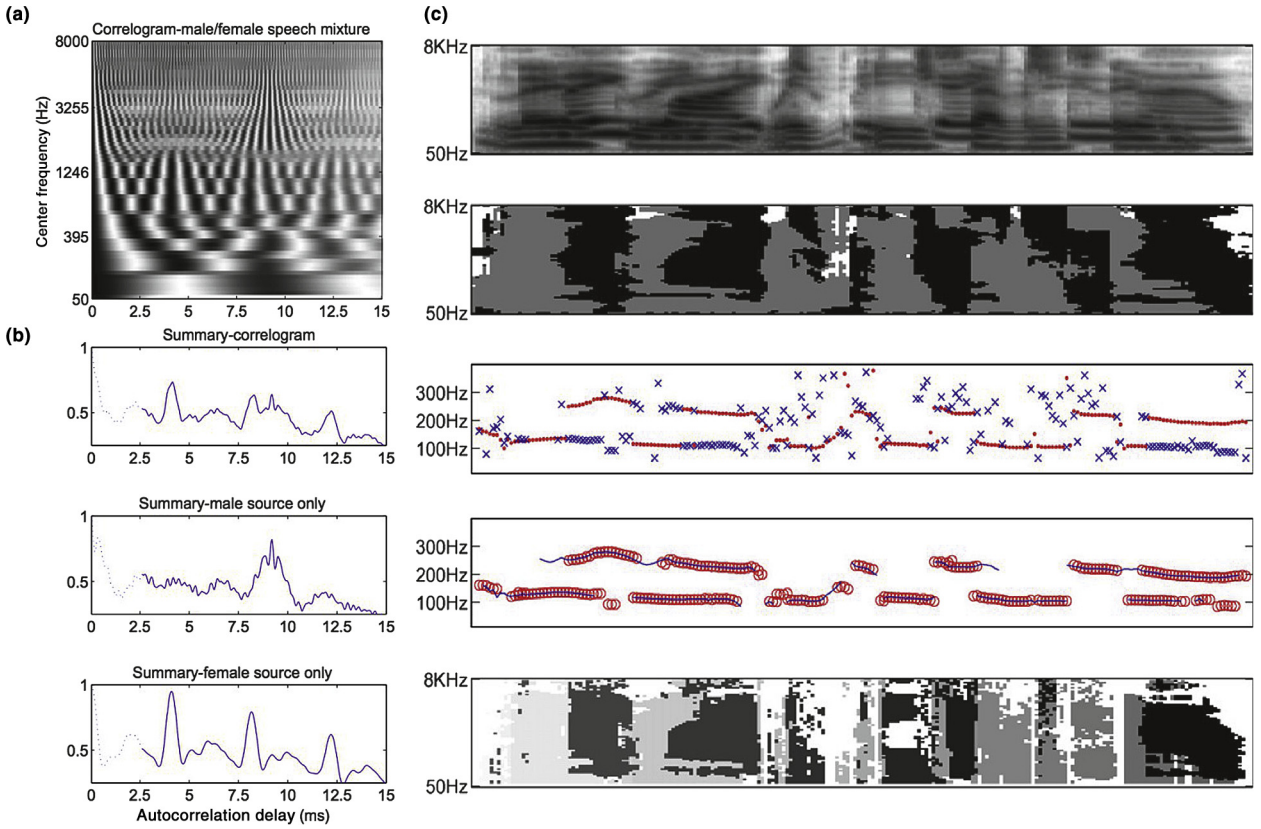

Figure 2 Using pitch cues to separate a two-talker mixture. (a) Correlogram; (b) summary correlograms for the mixture (top) and formed from spectral components dominated by the male (middle) and female (bottom) voice; and (c) sequential grouping of multiple pitch tracks. From top to bottom: auditory spectrogram of the mixture; ground truth for regions dominated by the female (black) and male (gray) voice; pitch estimates derived from the correlogram, with red dots indicating the stronger source and blue crosses the weaker source; pitch estimates (red) produced by the multipitch tracking algorithm, with ground truth in blue; and regions corresponding to estimated pitches. Adapted from Figures 6 and 10 of Ma, N., Green, P., Barker, J., Coy, A., 2007. Exploiting correlogram structure for robust speech recognition with multiple speech sources. Speech Communication 49, 874–891.

voices, it is evident that each voice can be associated with a number of time-varying partials that are harmonics (integer multiples) of a common fundamental frequency. These components, apparent in the low- and midfrequency regions, share a common onset time, maintain harmonicity as they change smoothly over time, and to a large extent possess simultaneous offsets. In the mid- and high-frequency regions, we see evidence of frequency components enhanced by the resonant cavities of the vocal tract. Again, these share common onsets and offsets with the harmonic partials and change smoothly over time. A closer look reveals periodic vertical structures, which are most evident around 0.9 s for the male voice. These correspond to amplitude modulations caused by the interaction of more than one harmonic component in a single auditory filter, and for midfrequencies the modulation period is equivalent to the harmonic separation (i.e., the fundamental frequency). In principle, then, periodically excited portions of a sound source such as speech can be recombined on the basis of a common fundamental frequency along with continuity constraints and onset–offset synchrony.

A representative signal-driven grouping algorithm based on fundamental frequency (Ma et al., 2007) is depicted in Figure 2. Panel A shows the correlogram for a single frame of a mixed male and female talker signal. The correlogram represents the periodic structure of the signal in each frequency region in a small time interval. The main value of the underlying autocorrelation function comes from its property of preserving periodicities in the input signal while removing phase information. In response to a periodic signal, each frequency region of the correlogram contains peaks at delays corresponding to multiples of the frequency of dominant components in that frequency region. So, for a harmonic series, low-frequency channels will produce a peak at the frequency of the lowest component, while progressively higher channels will possess peaks at the frequency of the nearest component and at integer multiples of that frequency. For example, in response to a voice with a fundamental frequency of f Hz, peaks will be present at delays of $1/f$ s, $2/f$ s, and so on in the low-frequency region; at delays of $1/2f$ s, $2/2f$ s, and so on in a slightly higher frequency region; and at lags of $1/3f$ s, $2/3f$ s, $3/3f$ s, and so on for the next frequency region. Critically, note that each frequency region contains a correlogram peak at a delay of $1/f$ s. It is this coincidence that makes the correlogram a powerful representation for discovering fundamental frequencies of constituents of a mixture. The coincident delays can be seen as vertical structures in Figure 2(a). In order to extract dominant fundamentals, it is usual to sum

the correlogram across frequency channels, resulting in a summary correlogram as depicted at the top of Figure 2(b). The largest peak corresponds to the female voice at a lag of around 4 ms (250 Hz), while other peaks can be seen at around 8, 9, and 12 ms. In principle, it is possible to identify which frequency regions contributed to each peak, and thereby group together those that appear to come from the same source. Unfortunately, due to the presence of large numbers of autocorrelation peaks in the midfrequencies, it is not feasible in practice to use the correlogram directly in this way to perform source separation. One solution, adopted by Ma et al. (2007), is to sharpen the correlogram using Gabor cosine operators in what is essentially a matched filter operation taking advantage of the local 'pyramidal' structure of the correlogram. This results in a less ambiguous summary correlogram, but peaks in the summary are still not guaranteed to correspond to the fundamental frequency of one of the constituent sources, as shown by the estimates derived in this way in the third panel of Figure 2(c). Consequently, a multipitch tracking process is normally used to exploit temporal continuity constraints in order to resolve ambiguous local pitch estimates.

Grouping using intrinsic signal properties such as common fundamentals can go some way to reconstructing an individual sound source. However, there are limits as to what the signal-driven approach can accomplish. One specific issue concerns those parts of the signal that are not periodically excited, such as the voiceless consonants /s/, /th/, and /f/ in 'set', 'three,' and 'four' in Figure 1. Grouping of voiced and voiceless elements has received relatively little attention in CASA (but see Hu and Wang, 2011), but it can in principle make use of the onset–offset synchronies that result from an alternation of voiced and voiceless elements, or interaural time and level differences that signal a common source azimuth (Harding et al., 2006).

Pure Model-Driven Grouping

Individual acoustic sources are often highly structured, containing regularities that can be captured by discrete or probabilistic models. For instance, in the case of speech, constraints exist at every level, from articulatory limitations (e.g., vocal-tract size or vocal fold vibration rate) through to linguistic determinants of lexical, grammatical, and semantic possibilities. Similar considerations apply to musical instruments and animal vocalizations. Even less-structured sounds, including those originating from the combination of multiple sources (e.g., babble, traffic, leaves rustling, and waves crashing), exhibit patterning in their spectral profile or temporal modulations. In all these cases, source models can be learned from data, and more recent CASA systems have shown how this kind of prior knowledge can be used to separate a mixture into its individual components. The key idea is to combine statistical models for each source believed to be present at each instant in the mixture in an attempt to explain the acoustic observations. In this way, sources are separated essentially as a by-product of recognition. The original scheme, proposed by Varga and Moore (1990), has since been extended to cope with the computationally complex process of evaluating all possible combinations of a large number of source models. Source model combination can be extremely effective. Indeed, algorithms have been demonstrated that outperform human listeners in limited-domain tasks such as the recognition of speech from one talker in the presence of a competing talker (Rennie et al., 2010).

Identifying a Target Source in an Unknown Background

The fundamental problem faced by listeners is to extract information from a single target sound source in the face of other sounds, whose composition (e.g., type and number) may vary over time. Pure model-driven grouping fares less well in situations where the number of active sources has to be estimated, and also requires statistical models at some level of structural detail for all sources present. A source separation system aimed at tracking a single target in an arbitrary background was proposed by Barker et al. (2005). The system makes use of both signal-driven and source model-driven grouping, and is based on the notion of 'glimpses,' motivated by energetic masking considerations in the auditory periphery. As a consequence of loglike amplitude compression in the cochlea, auditory nerve activity in each frequency region is dominated by the most energetic source active at that instant to such an extent that weaker sound sources make almost no impact on activity unless they are within a few decibels of the stronger source. Because many sounds exhibit sparse concentrations of energy across time and frequency, the likelihood of any pair of sources contributing similar energies in any spectro-temporal location is quite low. In other words, while sources are acoustically transparent, they are better characterized as opaque after encoding in the auditory periphery. The mixture shown in Figure 1 illustrates the opacity approximation. Locally, each part of the auditory spectrogram is nearly identical to the corresponding region of the auditory spectrogram for one of the constituent talkers. Some interactions between the sources are visible (e.g., in the low-frequency region at around 0.6 s), but most of the spectro-temporal plane provides a relatively undistorted view of one or other sources.

The glimpsing idea is further illustrated in the upper panel of Figure 3, which uses a form of color coding to denote spectro-temporal regions in a mixture of two talkers that is dominated by one or the other talker. Each compact region of a single color is a potential glimpse of one source. Of course, glimpses do not arrive at the auditory system with a convenient color code, and the computational task is akin to taking two or more jigsaws, throwing away some large fraction of the pieces of each, and attempting to identify the original pictures. Recast in this manner, the focus of glimpse-based CASA is on the twin problems of glimpse detection and integration. Detection is well suited to signal-driven grouping. Potential glimpses of sufficient spectro-temporal extent typically contain cues based on features such as periodicity or onset synchrony which allows them to be constructed using the kinds of signal-driven grouping algorithms outlined in this article. Glimpse detection avoids the need for extensive grouping of elements, especially across time, and thereby greatly simplifies the CASA problem.

The glimpse integration stage requires probabilistic models for the target sound source. For speech, these would typically be hidden Markov models for phonemes, words, or some other unit. The goal, shown in the panel in Figure 3, is to find both

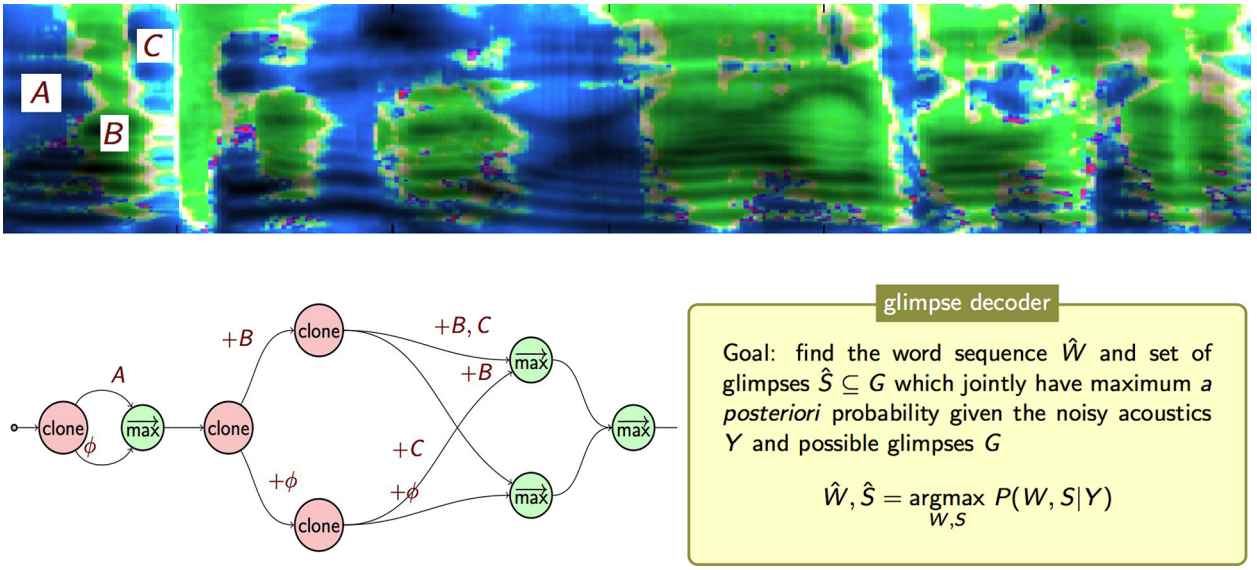

Figure 3 Upper: An auditory spectrogram of a two-talker mix represented by a hue saturation value encoding. The value corresponds to the log energy of the mix, as for conventional auditory spectrograms, while the hue and saturation values are derived from the log of the local target-to-masker energy ratio. Positive values map on to green, while negative map to blue. The magnitude of the log energy ratio maps on to the saturation term. In this way, regions where one source is clearly dominant show up as a single well-saturated hue, while regions where the energy ratio of the two sources is similar possess an intermediate unsaturated hue. Lower left: An illustration of the decoding algorithm applied to the first three glimpses. Lower right: Definition of the glimpse decoder for word models.

the most likely word sequence and the set of glimpses that explain the acoustic observations. Decoders of this type rely on missing data techniques (Cooke et al., 2001; see also Seltzer et al., 2004) to permit models to be evaluated against the partial information contained in glimpses.

A brute force approach to glimpse integration would require the evaluation of all possible subsets of glimpses against all permissible model sequences, something that is clearly infeasible. However, a computationally tractable algorithm for glimpse integration is available, as illustrated in Figure 3. Consider first the way that a conventional decoder for sources such as speech operates. At each moment in time, the decoder maintains a very large number of alternative partial interpretations of the data (e.g., incomplete phrases) up to that point, each of which has an associated likelihood. At the next time instant, a new acoustic observation becomes available and is used to update all existing interpretations. Finally, on reaching the end of the utterance, the most likely interpretation is chosen as the recognized result. The glimpse decoder operates in a similar fashion up to the point at which a new glimpse is encountered. At this point, each ongoing hypothesis is cloned. In subsequent time instants, one hypothesis is evaluated by including the glimpse in its interpretation, while its cloned counterpart excludes it. When the glimpse ends, all pairs of cloned hypotheses are compared and the most likely path of the pair is retained, under the assumption that later-arriving glimpses will not affect the earlier interpretation. Compared to a conventional single-source decoder, the complexity of the glimpse decoder is increased by a factor that in the worst case is 2 to the power of the largest number of temporally overlapping glimpses encountered. This number depends in turn on the power of signal-driven grouping in detecting glimpses. In this way, the decoder trades off signal and model-driven grouping: the more successful the signal-driven grouping stage, the fewer alternatives are presented to the model-driven stage.

Applications, Evaluations, and Future Challenges

The ability to describe the content of complex acoustic environments will find widespread application in many fields, including human–computer interfaces, health, security, entertainment, and robotics. Key beneficiaries of general-purpose CASA will be hearing-impaired individuals, for whom selective enhancement of sources of interest will be possible. Other applications include robust automatic speech recognition for everyday (i.e., potential noisy) conditions, remote audio surveillance, automated music analysis, transcription, and repair, as well as providing audio descriptions and audio awareness for robots.

Much effort in CASA to date has been dedicated to tackling the 'cocktail party problem,' which involves separating or recognizing a voice of interest in a background of one or more distracting voices. This formed the challenge problem for the first global evaluation of speech separation and recognition systems, which took place in 2006 (SSC website). While speech separation in the cocktail party task is an important goal, and one that is particularly relevant for hearing-impaired listeners, it is a largely artificial problem. Realistic sound environments have additional complications such as the presence of sound sources whose type and number vary over time. This scenario forms the basis for current evaluations (CHIME website).

Future systems can be expected to handle moving sources. To counteract these and other dynamic elements of auditory

scenes, CASA systems may well utilize active sensors to simplify the tracking of moving targets while attenuating the contribution of competitors. More sophisticated scenarios demand models of auditory attention capable of focusing on sources of interest while responding to relevant novel events. Richer characterizations of the auditory scene will be possible by using algorithms that construct more complete spatial descriptions than is currently the case, going beyond azimuth estimates to include elevation and distance. An additional challenge is to develop more robust scene analysis algorithms that are capable of handling and indeed exploiting reverberant energy and other convolutional distortions. Finally, it is likely that auditory and visual scene analyses will be more closely integrated in future in order to exploit the advantages of each modality.

See also: Speech Perception; Speech Recognition and Production by Machines; Statistical Pattern Recognition.

Bibliography

Barker, J., Cooke, M., Ellis, D.P.W., 2005. Decoding speech in the presence of other sources. Speech Communication 45, 5–25.

Bregman, A.S., 1990. Auditory Scene Analysis. MIT Press, Cambridge, MA.

Brown, G.J., Cooke, M., 1994. Computational auditory scene analysis. Computer Speech and Language 8, 297–336.

Cooke, M., 1993. Modelling Auditory Processing and Organisation. Cambridge University Press, Cambridge, UK.

Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and uncertain acoustic data. Speech Communication 34, 141–177.

Darwin, C.J., 1981. Perceptual grouping of speech components differing in fundamental frequency and onset-time. Quarterly Journal of Experimental Psychology 3A, 185–207.

Ellis, D.P.W., 1999. Using knowledge to organize sound: the prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures. Speech Communication 27, 281–298.

Glasberg, B.R., Moore, B.C.J., 1990. Derivation of auditory filter shapes from notched-noise data. Hearing Research 47, 103–138.

Harding, S., Barker, J., Brown, G.J., 2006. Mask estimation for missing data speech recognition based on statistics of binaural interaction. IEEE Transactions on Speech and Audio Processing 14, 58–67.

Hu, G., Wang, D., 2004. Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Transactions on Neural Networks 15, 1135–1150.

Hu, K., Wang, D., 2011. Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction. IEEE Transactions on Audio, Speech and Language Processing 19, 1600–1609.

Jeffress, L.A., 1948. A place theory of sound localization. Journal of Comparative and Physiological Psychology 41, 35–39.

Licklider, J.C.R., 1951. A duplex theory of pitch perception. In: Schubert, D. (Ed.), 1951. Experientia, 7. Dowden, Hutchinson and Ross, pp. 128–133.

Lyon, R.F., 1983. A computational model of binaural localization and separation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1148–1151.

Ma, N., Green, P., Barker, J., Coy, A., 2007. Exploiting correlogram structure for robust speech recognition with multiple speech sources. Speech Communication 49, 874–891.

Marr, D., 1982. Vision. W. H. Freeman, San Francisco, CA.

May, T., van de Par, S., Kohlrausch, A., 2011. A probabilistic model for robust localization based on a binaural auditory front-end. IEEE Transactions on Audio, Speech and Language Processing 19, 1–13.

Rennie, S.J., Hershey, J.R., Olsen, P.A., November 2010. Single-channel multitalker speech recognition: graphical modelling approaches. IEEE Signal Processing Magazine, 66–80.

Seltzer, M.L., Raj, B., Stern, R.M., 2004. A Bayesian framework for spectrographic mask estimation for missing feature speech recognition. Speech Communication 43, 379–393.

Varga, A.P., Moore, R.K., 1990. Hidden Markov model decomposition of speech and noise. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 845–848.

Wang, D., Brown, G.J. (Eds.), 2006. Computational Auditory Scene Analysis: Principles, Algorithms and Applications. John Wiley and Sons, Hoboken, NJ.

Relevant Websites

http://spandh.dcs.shef.ac.uk/chime_challenge/ – The CHiME Challenge (CHiME).

http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm – The Speech Separation Challenge (SSC).

Australia and New Guinea, Archaeology of

Jim Allen, Archaeology, La Trobe University, Melbourne, VIC, Australia
Ian Lilley, University of Queensland, Brisbane, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J. Allen, volume 2, pp. 952–956, © 2001, Elsevier Ltd.

Abstract

Various chronologies for the arrival of modern humans in Australia and New Guinea are discussed, together with a consideration of possible routes of entry. The archaeological evidence from the New Guinea Highlands and the islands of Melanesia is reviewed. After a general discussion of human colonization of Pleistocene Australia, some of the key sites in two contrasting environments – the arid zone and Tasmania – are compared.

Dating the Arrival of Humans

Although today Australia, Tasmania, and New Guinea comprise three separate landmasses, during times of lowered sea levels in the late Pleistocene they formed a single continent called *Sahul*. There is general consensus that Sahul's first human colonizers were both anatomically and behaviorally modern humans; indeed the necessary crossing of a minimum water barrier of *c.* 90 km to reach New Guinea or Australia in the late Pleistocene constitutes one measure of that modernity (Davidson and Noble, 1992).

Beyond that, there is little consensus for the date of earliest arrival (O'Connell and Allen, 2004). The oldest claims, in the 100–200 ka (thousands of years) range, all depend upon ecological evidence – usually charcoal counts in pollen cores interpreted as anthropogenic rather than natural fires – and are claimed in both the north and south of the Pleistocene continent, as well as Eastern Indonesia. The second group of claims consists of analyses of mtDNA diversity which put humans in Sahul *c.* 70–100 ka. A third group of claims associate direct archaeological evidence with uranium series, luminescence, and electron spin resonance dates in the 50–70 ka range. The fourth category includes those sites which form the beginning of a continuous sequence of occupation in the Pleistocene continent which may extend to 45–47 ka sidereal years ago. Certainly by *c.* 40 ka all the major ecological zones of Sahul and nearby islands such as New Britain and New Ireland – and perhaps even the main Solomon Islands chain – had been colonized by humans.

Claims in the first very ancient group have little credibility because they lack direct artefactual evidence and because it is difficult to accept on the basis of current global evidence that humans in Sahul before 100 ka could have been anatomically and behaviorally modern. Distinctly premodern *Homo floresiensis* probably reached the island of Flores during this period or even earlier but apparently moved no further east, probably because its original movement across the water barrier known as Wallace's Line was accidental (Morwood and Jungers, 2009). Wallace's Line marks the distributional divide between Old World placental mammals and Australian/Sahulian marsupials, but a few now-extinct Old World mammals such as Stegadon elephants also made the short crossing to Flores then advanced no further toward Sahul. Further afield, pre-sapiens humans also reached the Mediterranean island of Crete at least 110 ka, but only had to cross about 30 km of water, not the 90 km needed to reach Sahul (Strasser et al., 2011).

Three specifically archaeological claims put humans in Australia before or at 50 ka. The western New South Wales burial known as Mungo 3 has been redated to *c.* 56–68 ka using uranium series and electron spin resonance techniques on the skeletal elements and OSL to date the surrounding deposits (Bowler et al., 2003). This is particularly dramatic because the Mungo 3 human was ritually buried and covered in red ocher. Such burial is seen as a marker of modern behavior, and if as old as this, would be the oldest unambiguous ritual burial of *Homo sapiens* presently known anywhere. The second claim is for two rockshelter sites in the Northern Territory, Nauwalabila and Malakunanja, where the deepest stone artifacts are associated with TL and OSL dates having central tendencies between 50 ka and 60 ka (Roberts et al., 1994). Here critics question the *in situ* nature of the deepest artifacts being dated rather than the dating techniques themselves (*see* Chronology, Stratigraphy, and Dating Methods in Archaeology). Similar criticisms have been leveled at the third example, the recent AMS dating of the Devil's Lair cave in southwest Western Australia to *c.* 48–50 ka (Turney et al., 2001), where the lowest levels of the site which produced this date are water-scoured and contain only a few stone flakes, some not clearly of human manufacture.

Defenders of the longer chronologies criticize the shorter chronology on the grounds that the great majority of dates are based on older radiocarbon dating which was unreliable for ages much over 40 ka, so that this age for earliest arrival might easily be a product of the technique (O'Connor and Chappell, 2003). However, modern radiocarbon techniques now produce accurate dates up to 45 ka and have done so regularly in Australia for nonarchaeological samples, but only on relatively few occasions for archaeological sites.

The resolution of this debate is needed to understand the rate at which Sahul was colonized, and the adaptive strategies used to occupy a wide range of new environments. Equally important are the implications of this debate for understanding the spread of modern humans elsewhere in the world, since behaviorally modern humans are unknown before *c.* 50 ka anywhere in Eurasia (Barker et al., 2007; Dennell and Petraglia, 2012; Kaifu and Fujita, 2012).

Routes of Entry

The Sunda shelf of Southeast Asia is separated from the Sahul shelf of Australia, Tasmania, and New Guinea, and this break forms a significant biogeographical boundary first recognized by Alfred Russell Wallace. As noted above, Wallace's Line separates an Asian fauna which includes primates, carnivores, elephants, and ungulates from the marsupial fauna of Australia, Tasmania, and New Guinea. However, the widespread elements of Asian flora in Indonesia, New Guinea, and into northern Australia indicate that this boundary still permitted the natural transfer of this flora.

Two routes of entry from Southeast Asia seem possible, the first and more northern route through the Indonesian Moluccan archipelago to New Guinea, and a second more southern route along the Flores chain to Timor and then to northern Australia (Allen and O'Connell, 2008; Birdsell, 1977). Any of the subroutes which might form the northern route had two sailing advantages which the southern route lacked – intervisibility between departure point and target landfall on all legs and the ability to return to the departure point on existing winds and currents. Sea craft could journey blind from Timor to Australia on the summer monsoon, but not return. Interestingly, altered heights of islands because of sea level fluctuations had no appreciable effect on either intervisibility or distance between landfalls.

The earliest coastal site in New Guinea is at its extreme eastern end, emphasizing the absence of systematic archaeological exploration of the western half of the island. This site, an open site on uplifted coralline terraces on the Huon Peninsula, has yielded dates of *c.* 40 ka or a little older for large tanged and waisted flaked axes which must have been hafted.

New Guinea Highlands

New Guinea is dominated by a steep central cordillera rising to 4000 m asl in places. Because of rainfall patterns and cloud cover, Pleistocene colonists likely settled the series of intermontane valleys ranging between 1200 and 2400 m. Currently, the oldest dates for humans in the highlands are *c.* 45 ka, the lack of any time lag from the coast indicating how adaptable early colonists must have been to settle this remote region so quickly. This date is for the recently re-excavated site of Kosipe, located at 2000 m in the Ivane Valley, where stone tools, hearths, and plant and animal remains were found (Summerhayes et al., 2010). The animal bone cannot be identified but the floral remains indicate the exploitation of local pandanus as well as yams from lower altitudes. The excavators think the site was only visited during this period and not permanently inhabited. The subsistence patterns reflected in other excavated sites like the Nombe rockshelter reflect a strategy of hunting and collecting across a range of high-altitude zones. While forests and grasslands contained mainly small game, Pleistocene New Guinea was also home to a small number of large mammals, all now extinct. At Nombe, four species – all forest browsers – occur between *c.* 25 ka and *c.* 14 ka, but it is uncertain whether they were human prey. The general paucity of game animals in New Guinea is, however, offset by the huge array of plants, of which more than 200 are used for food. While little direct evidence exists for the development of plant processing and plant storage, it is likely that these processes developed from the beginning. The Huon-type axes also occur at Nombe, Kosipe, and other early sites, where they have been identified as tools used to thin forest patches by trimming and ring-barking so as to promote the growth of useful plants. The burning of forests at this time is reflected in the pollen records and supports this interpretation.

If accurate, these data suggest a long evolution of plant management which culminates in a distinctive highlands agriculture at 9–6 ka at Kuk in the Wahgi Valley (Denham, 2006). Here evolutionary stages of garden technology have been identified, including hydraulic practices involving both drainage and irrigation. The most recent phases reflect the familiar root crop/arboriculture/pig-husbandry systems present in the highlands today. Pigs, not indigenous to New Guinea, appear in the archaeological record 5–6 ka with some claims back to 10 ka (Denham, 2006).

Island Melanesia Before Lapita

Moving into the nearer islands of Melanesia required adapting to the simplified terrestrial ecology of an oceanic world. The initial colonists adopted strategies which compensated for this reduced range of edible plants and terrestrial animals through high mobility across large territories. This, in turn, depended on safe, if simple, watercraft. Initially it appears such craft took people to scattered resources, but over time, as populations and specialized resource knowledge increased, moving resources to people increasingly replaced the first strategy.

Because of tectonic uplift and deep submarine contours along the north coast of New Ireland, Pleistocene sites now normally flooded by later rising seas have been preserved (O'Connell and Allen, 2007; Leavesley, 2006). The oldest known of these, Buang Merebak, recently redated to 40–41 ka, and Matenkupkum (*c.* 35 ka) are about 140 km apart, and both contain marine shell middens which are very similar. Both contain marine fish bones indicating some form of fishing at this age, and both reflect strategies of shellfish collecting and processing the shells for food and tools. People also moved inland, quarrying chert outcrops near Yombon in New Britain at *c.* 36 ka. At Kilu Cave in the northern Solomons, stone tools dated between 29 and 20 ka carry residues of *Colocasia* and *Alocasia* taro, in quantities sufficient to suggest that successful selection had already taken place for starch-rich tubers.

Getting to Kilu required a sea journey about twice that of any prior necessary crossing between mainland Southeast Asia and the Bismarck Archipelago, a voyage without the aid of two-way intervisibility. But by 21 ka people had reached the Admiralties Group, a voyage of over 200 km which required that 60–90 km be traversed totally out of the sight of any land; this is the earliest known example of true seafaring anywhere.

By 20 ka, moving resources to people is reflected in New Ireland in the appearance of obsidian from a New Britain source 300–350 km distant and the introduction of *Phalanger orientalis*, a marsupial which became an important food source. It is the first of a series of animal translocations in the region and flags the possibility that useful plants and nut-bearing trees were also moved about well before the end of the Pleistocene.

A reduction in cave use in the region during the early Holocene raises the question of whether these islanders were more dependent on cultivated plants than previously. While

the evidence is piecemeal, gradual economic intensification is reflected in aspects of the archaeology such as intensive and extensive trade in obsidian, plant residues on tools, domesticated and transported nut remains in sites, elaborate stemmed and hafted obsidian and chert tools, stone and shell ground axes, open-site settlement patterns, and expansion onto smaller offshore islands, all of which seemingly demand more substantial food production systems than those provided solely by hunting and gathering.

Maritime trading systems developed in part on Pleistocene foundations elaborated during the later Holocene, as horticulture underwrote village development both in the near islands and around the New Guinea coast (Pavlides, 2006). The growth of these systems, particularly after the spread of Lapita-introduced pottery technology, has been a major research concern of archaeologists in the region (*see* Pacific Islands, Archaeology of).

Australia

Regardless of where first landfall occurred, the move south through Sahul was into environments more foreign in their floras and faunas and in increasing contrast to presumed tropical humid homelands. Although more than 160 Pleistocene age sites are known in the region, they occupy a diverse range of environments and are frequently great distances apart, such that no neat synthesis of past human behaviors is yet possible. A few short generalizations will provide some background (Hiscock, 2008; O'Connell and Allen, 2007). (1) With some glaring exceptions, such as edge–ground axes in Arnhem Land at *c.* 20 ka, Pleistocene flaked stone tools across Australia are mostly amorphous, showing little of the elaboration which characterizes the Upper Palaeolithic in Europe or the Later Stone Age in Africa. (2) A host of larger marsupials, collectively dubbed megafauna, appear to have become extinct around the time of human arrival, although the dating of the extinction event(s) is poor. While it is difficult not to implicate humans in this extinction, the processes are unclear. Direct extinction by hunting appears unlikely, given the range of species, the size of the continent and its varieties of environments, the presumed low numbers of humans, and the conspicuous absence of kill sites. Many commentators appeal to less direct effects, particularly habitat modification by the human use of fire, but not even this effect is universally endorsed. Arguments of climate change are still advanced to counter human involvement, but these are also widely criticized. (3) Various sorts of small blade-tools and points which first appeared in low numbers in different parts of the continent from *c.* 7 ka dramatically increased in popularity from *c.* 4 ka throughout Australia, but not in Tasmania, which had been cut off from Sahul by rising postglacial sea levels some 13 ka (Hiscock, 2006). (4) Despite the example of New Guinea to the north, which remained attached to Australia until cut off by rising postglacial seas *c.* 7 ka, Aboriginal people did not adopt or develop agriculture, remaining a continent of hunter–gatherers until European arrival. Even so, the successful colonization of even extremely marginal areas of the continent reflects the adaptability of these people, while the widely documented development of art and engraving on rocks and cave walls, now demonstrated from earliest arrival, reflects an intellectuality and social awareness among aborigines previously only documented ethnographically.

The Arid Zone

The arid zone occupies an area of *c.* 5 million square km – about 70% of Australia's land surface – and during the Pleistocene this area was even larger owing to glacial aridity. In this region, occupation and movement has always been constrained by the distribution of food and water (Hiscock, 2008; O'Connell and Allen, 2007). Areas of relatively high relief, the ranges of the central desert and the Kimberly and the Pilbara in particular, concentrate run-off in stream channels and waterholes; similarly, riverine floodplains contain perennial and seasonal streams which contain surface water in some seasons and store accessible subsurface reserves. The riverine floodplains, particularly those along the Darling system in New South Wales, offer a range of aquatic resources, including fish, shellfish, hydrophytic plants, and water birds.

Before 25 ka, conditions were perhaps wetter, though cool to cold, but between 25 and 12 ka conditions were very much colder and drier, especially at the last glacial maximum. Much of this zone would have been inimical to human settlement at this time. Although people were on its fringes by 40 ka and occupying the 'dead heart' before 30 ka, many sites were abandoned at the height of the ice age.

People were at Mandu Mandu Creek rockshelter on the arid Western Australian coast *c.* 34 ka and left behind shell beads apparently used for personal adornment, as well as ocher carried there from the Pilbara area. As the sea dropped, offshore islands were incorporated into coastal territories as people followed the receding coastline. Today the bare rocks of the Montebello islands contain small rockshelter sites which reveal human presence at *c.* 27 ka, with both terrestrial and coastal resources represented in the deposits, and again between 7 and 10 ka, when the prograding sea was again near. These sites were not used between times and were abandoned when they again became offshore islands.

Thus models of changing land-use patterns put forward by arid region scholars all emphasize mobility and adaptation to changing environmental conditions and suggest small populations thinly spread across the arid zone. Diets may not have been greatly different from those recorded ethnographically in these regions. They are generally similar and broadly based and in addition to marsupials included reptiles, birds, larvae, tubers, fruits, seeds, and aquatic resources where available. Recent claims that grass and tree seeds might have been gathered and processed in the Pleistocene and particularly before the last glacial maximum are notable because although seeds are nutritious, their gathering and processing is laborious. For this reason, seed processing elsewhere in the world is seen as a key indicator of changes in human behavior connected to the appearance of new technologies, larger and more settled populations, and in places the establishment of agriculture. Its presence in an Australian form in preglacial arid Australia has significant theoretical relevance. Traces of seeds and other plant food remains from Pleistocene levels in Carpenters Gap in the northwest is complemented by pieces of grinding stones dating to *c.* 30 ka at Cuddie Springs some thousands of kilometers to the east. These carry usewear and residue traces suggesting the

processing of starchy and siliceous plants, evidence coincident with a local change to grasslands seen in the pollen record at this time.

Data from the Mungo region of western New South Wales indicate how favorable these lake systems were when lacustral conditions ensued. A range of resources – fish, shellfish, crustaceans, and birds – were taken. Fish may have been taken using nets and spears and possibly traps. A range of mainly small terrestrial game was also captured, and again in this region unspecialized grinding tools suggest at least generalized processing of plant foods and seeds during the Pleistocene.

Tasmania

Today, and for the last 12 ka, humans have occupied the north and east of Tasmania. In the southwest, temperate rainforest continues to exclude settlement. During the late Pleistocene, however, this situation was reversed (Hiscock, 2008; O'Connell and Allen, 2007). At times glacial deserts made the east inhospitable, while colder and drier conditions forced the rainforests in the southwest into refugia at lower altitudes, opening up the montane valleys to open forest with patches of grassland on limestone substrates in the valley bottoms. Since 1980 *c.* 30 sites, predominantly limestone caves have been located in the rainforest and *c.* 12 tested by excavation. The oldest, Warreen, was occupied *c.* 35 ka and all were abandoned by 12 ka, when warmer and wetter conditions encouraged the return of the forests, which drove out game animals and their human hunters. At the height of the last glacial maximum, the mountains had glaciers extending as low as perhaps 700 m above sea level, only a few hundred meters above some of the occupied caves. With temperature reductions of *c.* 7 °C and winds reducing temperatures sometimes tens of degrees, it was indeed an extreme climatic place, some 40° of latitude south of their ancestors' equatorial point of departure.

The cluster of known sites provides a rare glimpse at a regional pattern of Pleistocene behavior extremely different to that depicted for arid Australia. Here the subsistence strategy was the opposite of broad spectrum. There are no obvious vegetable staples which might have supported even small transient groups and the plentiful faunal remains repeat the same pattern from site to site. Up to 80% of bones from humanly consumed animals were from the single species *Macropus rufogriseus* – the Bennetts wallaby, supplemented in small amounts by wombat and platypus. In more than 630 000 animal bones processed from eight sites, there is not a single extinct megafaunal species represented, although some are present in nonarchaeological deposits in caves in the same region, suggesting that those species known from Tasmania had become locally extinct before human arrival, or that they were not hunted or scavenged.

Holocene Continuities and Changes

As the forests drove Tasmanians from the southwest, rising sea levels at the end of the Pleistocene cut off Tasmanian access to the Australian continent and for 8 ka the Tasmanians were deprived of human contact beyond their island. When reunited with the world at the end of the eighteenth century, they no longer made bone points, caught fish, or made fire, and were obliged to keep fire burning or rekindle flame from neighboring groups. Their technology comprised no composite or hafted tools, and consisted only of single-piece wooden spears, clubs, digging sticks, kangaroo skin capes, skin and kelp water containers, string, and small baskets. A simple stone-tool technology was compared by Europeans to Mousterian industries. While the loss of fish in the diet *c.* 4 ka may reflect some established island-wide taboo, this dietary loss was offset by the continued capture of seals and terrestrial mammals and an expanded range of shellfish species. In the recent Holocene, more use seems to have been made of the less-hospitable south and west coasts and people also used nearby offshore islands. Bundles of reeds tied in the shape of small canoes are known ethnographically where they were used as flotation devices.

In contrast, the end point of Holocene changes on the mainland differentiated Tasmanians and mainland Australians so distinctly that early explorers and anthropologists considered them different races. But while the Tasmanian devil and marsupial tiger or thylacine – both extant in Tasmania when Europeans arrived – had disappeared from the mainland when the dingo was introduced there from Southeast Asia or New Guinea *c.* 4 ka, few other Holocene changes can clearly be attributed to external influences. In particular, the temporal and distributional evidence for the appearances of small points, geometric microliths, and other tools during the mid- and late Holocene suggest different and regional origins within the country. While some of these tool types disappeared perhaps 1–2 ka in some regions, in the north larger bifacial points and other macroblade tools appeared more recently (Hiscock, 2006).

This proliferation of regional tool-types, increases in site numbers, site types, and the range of organic and inorganic artifacts to be found in them, together with the appearance of cemeteries, evidence for exchange networks, and labor-intensive specialized food procurement activities such as constructing artificial channels to capture migrating eels, has been taken as a package by some Australian archaeologists to promote a model of socioeconomic intensification in the late Holocene (David et al., 2006; Lourandos, 1997). Even though stationary sea level and the relatively young ages of these sites clearly promote better preservation and higher visibility, this package is contrasted with the Pleistocene and early Holocene, where population densities were supposedly lower, territories larger, people were less sedentary, there was less social ranking, and less complex food procurement strategies. Critics of this model question its testability in the archaeology, pointing to evidence that aspects of Pleistocene life may have been just as complex as that in later periods, and some prefer to continue seeking environmental explanations for the changes seen in the archaeological record (Hiscock, 2008).

The unsurpassed heritage of Australian rock art has been recorded, classified, and interpreted for more than a century. The ability to now date many of the varied styles directly is a recent revolution which is enabling the integration of art with other aspects of the archaeological record. Many new and ingenious techniques, such as luminescence dating of sand grains in mudwasp nests overlying painted surfaces, or dates that use carbon in oxalate salts to indicate the time when such

encrustations covered the art, are now demonstrating a long Pleistocene history for many paintings, providing a further window on the nature of hunter–gatherer intellectuality.

See also: Archaeometry; Chronology, Stratigraphy, and Dating Methods in Archaeology; Hunter-Gatherer Societies, Archaeology of; Migrations, Colonizations, and Diasporas in Archaeology; Pacific Islands, Archaeology of.

Bibliography

Allen, J., O'Connell, J., 2008. Getting from Sunda to Sahul. In: Clark, G., Leach, F., O'Connor, S. (Eds.), Islands of Inquiry: Colonisation, Seafaring and the Archaeology of Maritime Landscapes. ANU E-Press, Canberra, pp. 31–46.

Barker, G., Barton, H., Bird, M., Daly, P., Datan, I., Dykes, A., Farr, L., Gilbertson, D., Harrisson, B., Hunt, C., Higham, T., Kealhofer, L., Krigbaum, J., Lewis, H., McLaren, S., Paz, V., Pike, A., Piper, P., Pyatt, B., Rabett, R., Reynolds, T., Rose, J., Rushworth, G., Stephens, M., Stringer, C., Thompson, J., Turney, C., 2007. The 'human revolution' in lowland tropical Southeast Asia: the antiquity and behavior of anatomically modern humans at Niah Cave (Sarawak, Borneo). Journal of Human Evolution 52, 243–261.

Birdsell, J., 1977. The recalibration of a paradigm for the first peopling of greater Australia. In: Allen, J., Golson, J., Jones, R. (Eds.), Sunda and Sahul: Prehistoric Studies in Southeast Asia, Melanesia and Australia. Academic Press, New York, pp. 113–167.

Bowler, J., Johnston, H., Olley, J., Prescott, J., Roberts, R., Shawcross, W., Spooner, N., 2003. New ages for human occupation and climatic change at Lake Mungo, Australia. Nature 421, 837–840.

David, B., Barker, B., McNiven, I., 2006. The Social Archaeology of Australian Indigenous Societies. Aboriginal Studies Press, Canberra.

Davidson, I., Noble, W., 1992. Why the first colonisation of the Australian region is the earliest evidence of modern human behaviour. Archaeology in Oceania 27, 135–142.

Denham, T., 2006. Envisaging early agriculture in the highlands of New Guinea. In: Lilley, I. (Ed.), Archaeology of Oceania: Australia and the Pacific Islands. Blackwell, Malden, MA, pp. 160–188.

Dennell, R., Petraglia, M., 2012. The dispersal of *Homo sapiens* across southern Asia: how early, how often, how complex? Quaternary Science Reviews 47, 15–22.

Hiscock, P., 2008. Archaeology of Ancient Australia. Routledge, London.

Hiscock, P., 2006. Blunt and to the point: changing technological strategies in Holocene Australia. In: Lilley, I. (Ed.), Archaeology of Oceania: Australia and the Pacific Islands. Blackwell, Malden, MA, pp. 69–95.

Kaifu, Y., Fujita, M., 2012. Fossil record of early modern humans in East Asia. Quaternary International 248, 2–11.

Leavesley, M., 2006. Late Pleistocene complexities in the Bismarck Archipelago. In: Lilley, I. (Ed.), Archaeology of Oceania: Australia and the Pacific Islands. Blackwell, Malden, MA, pp. 189–204.

Lourandos, H., 1997. Continent of Hunter-Gatherers. New Perspectives on Australian Prehistory. Cambridge University Press, Melbourne.

Morwood, M., Jungers, W., 2009. Paleoanthropological research at Liang Bua, Flores, Indonesia. Conclusions: implications of the Liang Bua excavations for Hominin evolution and biogeography. Journal of Human Evolution 57 (5), 640–648.

O'Connell, J., Allen, J., 2004. Dating the colonisation of Sahul (Pleistocene Australia-New Guinea): a review of recent research. Journal of Archaeological Science 31, 835–853.

O'Connell, J., Allen, J., 2007. Pre-LGM Sahul (Australia-New Guinea) and the archaeology of early modern humans. In: Mellars, P., Boyle, K., Bar-Yosef, O., Stringer, C. (Eds.), Rethinking the Human Revolution: New Behavioural and Biological Perspectives on the Origin and Dispersal of Modern Humans. McDonald Institute for Archaeological Research, Cambridge, pp. 395–410.

O'Connor, S., Chappell, J., 2003. Colonisation and coastal subsistence in Australia and Papua New Guinea: different timing, different modes. In: Sand, C. (Ed.), Pacific Archaeology: Assessments and Prospects. Service des musées et du patrimoine, Département Archéologie, Nouméa, pp. 15–32.

Pavlides, C., 2006. Life before Lapita: new developments in Melanesia's long-term history. In: Lilley, I. (Ed.), Archaeology of Oceania: Australia and the Pacific Islands. Blackwell, Malden, MA, pp. 205–227.

Roberts, R., Jones, R., Spooner, N., Head, M., Murray, A., Smith, M., 1994. The human colonisation of Australia: optical dates of 53,000 and 60,000 years bracket human arrival at Deaf Adder Gorge, Northern Territory. Quaternary Geochronology 13, 575–583.

Strasser, T., Runnels, C., Wegmann, K., Panagopoulou, E., Mccoy, F., Digregorio, C., Karkanas, P., Thompson, N., 2011. Dating Palaeolithic sites in southwestern Crete, Greece. Journal of Quaternary Science 26 (5), 553–560.

Summerhayes, G., Leavesley, M., Fairbairn, A., Mandui, H., Field, J., Ford, A., Fullagar, R., 2010. Human adaptation and plant use in highland New Guinea 49,000 to 44,000 years ago. Science 330, 78–81.

Turney, C., Bird, M., Fifield, L., Roberts, R., Smith, M., Dortch, C., Grun, R., Lawson, E., Ayliffe, L., Miller, G., Dortch, J., Cresswell, R., 2001. Early human occupation at Devil's Lair, southwestern Australia, 50,000 years ago. Quaternary Research 55, 3–13.

Australian Aboriginal Society and Culture: An Overview

Robert Tonkinson, The University of Western Australia, Crawley, WA, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Cowlishaw, volume 2, pp. 956–960, © 2001, Elsevier Ltd.

Abstract

This article provides a brief overview of Aboriginal Australian society and culture, past and present. After 1788, those hunter-gatherers who survived introduced diseases resisted the British colonists along an uncontrolled, often violent frontier. Processes of alienation, displacement, institutionalization, discrimination, and racism debilitated them. The First Australians became a small, powerless, oppressed minority, overrun by an expanding settler frontier. There is, however, considerable cause for optimism.

Origins

The ancestors of the Aborigines came out of Asia and crossed the Australian continent during a period of lower sea levels about 50 000 years ago, then went on to eventually populate the entire continent. Mitochondrial DNA evidence shows that Aborigines are unrelated to all their neighbors, such as the Maori, Papuans, and Indonesians. Almost 200 years prior to colonization, Indonesians became the first outsiders to visit in search of highly valued *bêche-de-mer* (sea-cucumber) found along the northwest and north coasts. However, their mark on the land was negligible, in great contrast to that of those who followed. Estimates of the continent's population at the time of British settlement vary between 300 000 and one million (Mulvaney and White, 1987; Edwards, 1988).

Despite inevitable fluctuations in population densities, subsistence activities, and regional politics, there was a high degree of cultural continuity throughout the continent. Much of this sameness derives from their hunter-gatherer adaptation, which limits possibilities of diversification. Australia had few areas (such as the lower River Murray valley) capable of supporting high population densities. Mobility and large ranges of movement typify their adaptation to much of the continent's interior, where travel in small dispersed groups was the norm. Multiple bonds linked individuals to homeland sites and territories; that more than 200 languages were spoken supports the fact of strong anchorage to homeland territories.

Social Organization

An estimated 500 or so 'tribes' existed at the time of the British invasion (see Berndt and Berndt, 1988; Burridge, 1973 for comprehensive accounts of traditional Aboriginal society and culture). The visible, face-to-face local group was the band, averaging between 15 and 25 people, which varied in size, and comprised a core of closely related people. Gender and age were the two most important status differentiating criteria. However, mature males predominated, especially in politics and religion, and they claimed primary responsibility for ensuring the society's cultural reproduction. Strong bonds of kinship, friendship, marriage, dialect, exchange relationships, and shared religious responsibilities also turned people's attention outward to regional cultural connections. The 'range' or normal circuit of a band typically overlapped to a certain extent with those of neighboring groups. This larger society could never be a permanent fixture, but was temporarily realized in the form of periodic gatherings two or three times every year, at a location possessing sufficient water and food resources to support a large group (Tonkinson, 1991). An integral element of Aboriginal society, these 'big meetings' are still very important in the lives of many remote-dwelling Aboriginal people. They include an early and obligatory settlement of all outstanding disputes, initiation of males, performance of rituals, exchange of sacred and mundane knowledge and information, items of material culture and other lore, arrangement of betrothals, and planning for future meetings. They are exciting occasions, filled with ritual performances, gossip, and romance.

Australia's Other Indigenous Minority: The Torres Strait Islanders

Many Australians are unaware that Australia has two Indigenous minorities. Traditionally, most of the 60 000 or so culturally Melanesian people of the Torres Strait Islands and northeastern coast of Queensland were tropical horticulturalists, living in settled villages and subsisting from their gardens, fishing activities, and some foraging for wild foods. Today, most Torres Strait Islanders live on the mainland and many have successfully integrated in the wider society. Socioeconomically, they have generally fared better than the Aborigines in their fortunes since the European invasion. Space precludes further discussion, but the interested reader should refer to Beckett (1990); and AIATSIS (2012).

European Impacts and Aftermath

European colonization of Australia began near Sydney in 1788, when the British established a convict settlement and used the legal fiction of *terra nullius*, 'empty land,' to claim the entire continent. Once settlers arrived in large numbers, an unregulated frontier was soon being pushed westward, provoking

Aboriginal resistance (Reynolds, 2006). The settlers, understanding nothing of the Aboriginal people's complex social organization and religion, increasingly depicted them as 'treacherous savages' who failed to make productive use of the land. Resistance was inevitable, leading to protracted hostilities, including massacres, mostly of Aborigines; eventually, superior technology prevailed. Introduced diseases also killed countless Aboriginal people, prompting characterizations of them as a 'weaker race' and the enactment of ameliorative legislation, which was sometimes labeled 'soothing the dying pillow' (Broome, 2010). In many parts of the continent, colonial authorities sought to contain the survivors (many of whom had been reduced to destitution after it became impossible to maintain traditional subsistence activities). They were institutionalized, given blankets and subject to many restrictions aimed at isolating them from settler society (Rowley, 1970; Rowse, 1998). Some remained on pastoral and cattle properties, where they labored. Others were sheltered by Christian missions, which had scant initial success with the people who, unsurprisingly, for the most part, evinced little interest in this new religion – though Christianity was later embraced, to varying degrees, in many places.

Following federation of the six Australian states in 1901, state authorities passed increasingly restrictive legislation that made Aborigines captives in their own land. Meantime, the Australian frontier had moved west and north into the arid interior. For a long time, 'the outback' or 'bush' was highly masculine, being considered no place for a white woman. Inevitably, then, frontiersmen would take Aboriginal women as sexual partners (sometimes disguising them as males). A major consequence was the rise of a mixed-race population, which eventually came to greatly outnumber Aboriginal people of full descent.

The period of unprecedented national prosperity that followed World War II was not shared by many Aboriginal people. This situation of poverty amid plenty aroused concern in some quarters of the wider populace, and various organizations devoted to Aboriginal welfare sprang up. So did Aboriginal protest and resistance movements. Led mainly by urban activists, these groups decried racism and neglect, and agitated for reforms and urged pan-Aboriginal unity. In 1967, a national referendum was held, the outcome of which was an overwhelming 91% vote to include Aborigines in the Census; it also empowered the Federal Government to make laws pertaining to Indigenous Australians.

On Australia Day, 1972, Indigenous activists set up a 'Tent Embassy' (representing a displaced nation), opposite the federal parliament in Canberra, demanding land rights, compensation, and protection of heritage. The 'Embassy' remains to this day as a constant reminder of unfinished business and a focal point for Indigenous protests. A change of Federal government in the early 1970s brought new policies aimed at increasing the level of Aboriginal self-management; it promoted Indigenous community councils and land councils and created service organizations. There were many setbacks and well-publicized scandals about alleged financial malfeasance, but the following decades saw the rise of powerful Aboriginal-controlled regional land councils and ancillary service organizations.

By 2000, the conservative Coalition federal government in power was taking a tougher line on Indigenous affairs. It later disbanded the Aboriginal and Torres Strait Islander Council, the national Indigenous elected representative body, raising fears of a new paternalism. In 2007, it intervened in the Northern Territory, using Defense personnel, ostensibly to protect Aboriginal children against sexual abuse. The Northern Territory Emergency Response (usually labeled 'the Intervention') imposed strict controls on Aboriginal people living in remote reserves; these included management of welfare benefits, especially where children are deemed to be neglected; restriction of access to alcohol and compulsory acquisition of some land areas. The Intervention continues to divide opinion. When the federal Labor Party won government in the 2008 election, it retained, modified, and extended the Intervention. Labor's stated major policy goal in Indigenous affairs is to close the immense socio-economic gap that persists between Indigenous Australians and the rest of the population. In life expectancy, for example, it is 12 years less for Indigenous men and 10 years for women.

In remote regions of central and northern Australia today, Aboriginal people retain a significant tradition orientation despite strong Westernizing influence on lifestyle and diet. They continue to enjoy hunting and gathering, more as leisure activities than subsistence necessity, but with vehicles, guns, and iron tools. Continuities with the past have remained important in the values and behaviors surrounding kinship and social relations, and in recent decades there has been a strong emphasis on cultural revival. There is much evidence of renewed and determined opposition by Aboriginal people to cultural absorption, consequent on material and cultural impoverishment, a legacy of the colonial past.

Economics and Kinship

Broadly speaking, the subsistence economy was classically hunter-gatherer: men hunted larger animals, such as kangaroos, crocodiles, and emus, while women and children gathered vegetable foods and hunted small animals (Keen, 2004). The nature of the seasonal round depended very much on climate, particularly rainfall, since most of the continent is arid. Everywhere, though, people possessed an encyclopedic knowledge of their environment and, for example, practiced strategic burning to facilitate the yields of both floral and faunal food sources. Like other hunter-gatherer societies, those of the Aborigines were small in scale, 'familistic,' kinship-based, and broadly egalitarian. A key integrating institution was the system of classificatory kinship, in which every kin-term encodes expected behaviors and responsibilities. Wider groupings also featured, and Aboriginal societies are unique in the complexity of their social category systems (moieties, sections, and subsections) that complement and overlie kinship systems. Activated for certain purposes, most often during rituals and major rites of passage, these social categories are divorced from economic activities. The absence of formal institutions of social control meant heavy reliance on individual self-regulation. Aboriginal people use kin terms in addressing and referring to others, indicating that, ideally, the relationship entails mutual obligation and responsibility. In rural and remote Australia, webs of kinship tend to be larger and more complex than in urban areas. Historical factors such as dislocation from country, institutionalization, and removal from parents, separated and dispersed many families.

Religion and Worldview

The Aboriginal cosmic order comprises human society, the plant and animal world, the physical environment, and the spiritual realm. Across most of the continent, worldviews were grounded in two closely interrelated concepts: 'the Dreaming' (also called the Dreamtime) and 'the Law' (Berndt, 1974; Stanner, 1979, 1989). Broadly, the Dreaming explains how things came to be as they are, and embraces past, present, and future. The Law specifies how life is to be lived according to jural rules instituted by ancestral beings in the creative epoch of 'the Dreaming.' It specifies what from the founding era was to be emulated and what was tabooed, and remains part of lived reality for most remote-dwelling Aborigines (Maddock, 1982; Keen, 1996).

Accounts of the Dreaming are at base broadly similar across the continent. Here is a brief example from the Western Desert: long ago, large and beautifully decorated humanoid-animal beings arrived in a featureless, empty land. Possessing super-human magical powers, they hunted, gathered, and exchanged religious objects, rituals, and knowledge. The extensive tracks of the major beings crisscross the land and are commemorated in mythology, song, and ritual. They left behind human societies and a blueprint for living. There were also localized creative beings, whose exploits are commemorated among smaller groups. Prominent or distinctive shapes and features in the landscape inspired much mythology; for example, a claypan created when a weary creator lay down to sleep, or a canyon forged by the passage of a huge rainbow serpent (Meggitt, 1962; Munn, 1973).

The ancestral beings eventually 'died,' worn out by their creative work, and turned into landscape features before, as spirits, they eventually returned to the spiritual realm. The 'life essence' they had exuded to animate the flora and fauna remained as the source of all life. These departed creators are believed to retain a strong interest in human affairs, and their spirit-being associates play a vital role as intermediaries between human and spiritual realms. Though unreachable in the course of everyday mundane activities, they or their spirit-being messengers are amenable to contact during altered states of consciousness, notably dreams, reverie, and rituals.

People's strong sense of security was grounded in their encyclopedic environmental knowledge. Their responsibility was to nurture country and resources, obey the Law, and perform the rituals necessary to maintain the fertility of flora and fauna. Multiple totemic associations link individuals to the Dreaming powers, and manifest a powerful urge to maintain intimate relationships between humans and nature. In the Western Desert region, for example, conception totemism provides every person with a unique account of her/his transition from spirit-child to human, linking everyone to place, spirits, and other humans and, most significantly, into the realm of the Dreaming itself.

Maintaining Ethnic Identity and Pride

About 400 000 Australians (*c.* 2.3% of the total population) today identify themselves as Aboriginal, and many more people with Aboriginal ancestry have melded into the population at large. About 80% of Aborigines are urban dwellers, often somewhat distanced from 'traditional' roots but proud of the cultural achievements of their ancestors. In remote regions, the Dreaming and the Law remain part of most Aboriginal people's lived reality and, together with kin and country, are essential components of their ethnic distinctiveness. Like the population at large, about 80% of Aboriginal people live in urban areas. Just over half of them live in just two states, New South Wales and Queensland, but are small minorities in both, whereas in the Northern Territory they constitute a significant (if still minority) proportion of the population. Rates of intermarriage between Aboriginal and non-Aboriginal people are increasing. Aboriginal identity is officially recognized via two criteria: self-identification and the recognition of this claim by an unspecified 'Aboriginal community.'

The spectrum of Aboriginal people's social situation in Australia today is vast, covering a small minority of well-educated bureaucrats and professionals occupying senior positions, and at the other extreme those at the bottom of Australia's socioeconomic pyramid, including a minority living in remote regions and speaking English as a second language. Although the Aboriginal population is disadvantaged according to virtually all social indicators, it is united by pride in their identity and the ongoing struggle for full equality (Sutton, 2009; ABS, 2011). Under current Australian law, all citizens ideally share the same rights and obligations, but Aboriginal people are demonstrably also subject to inequality and racism, both structural and personal.

In regions where alienation from land and traditional resources has been longest and strongest, Aborigines are of mixed descent and hugely outnumbered. Their sense of unity derives less from any common territorial roots than in a shared history of marginalization, alienation from land, oppression, and racism. For them, identity can be a major concern, since they may face 'hints of inauthenticity', especially from sceptics among the non-Indigenous majority. Some objectify their Aboriginality and associated traditions and have been active in cultural borrowing from other areas. The slogan 'cultural revival is survival' is linked to issues of identity, pride, land rights, local solidarity, and belonging. Many Indigenous people, across the continent, are increasingly aware, partly as a result of the land rights movement and recognition of native title, of the potential held by tradition to legitimate their claims to a share of the country's resources. Legislation, particularly regarding Aboriginal heritage protection and native title, is one of the few tools available to them for the effective exercise of some measure of power against the hegemony of the 'whitefella.'

Language

Although English is the first language of most Aboriginal people, statistics from 2008 report that one fifth of Indigenous people aged 15 years and over and 13% of Indigenous children (aged 3–14 years) speak an Aboriginal or Torres Strait Islander language. In parts of northern Australia, particularly, creoles are spoken. Many language centers operate in rural and remote areas, and there has been considerable investment in language

research and recording, and in supporting language education. There are some bilingual schools, particularly in central Australia and the Northern Territory, but in recent years bilingual education has become controversial, with opponents claiming that it retards the effective acquisition of English and mainstream education. About 70% of Indigenous children (aged 3–14 years) and 63% of Indigenous people aged 15 years or over are involved in cultural events, ceremonies, or organizations. Some two thirds of Indigenous Australians identify with a clan, tribal, or language group. In Southern Australia, many languages have been revived, and many people believe they are crucial to cultural health, a keystone of identity, Law, and land claims.

Art Production and Sport

Petroglyphs (rock engravings) and rock paintings have been a part of Aboriginal culture for at least 40 000 years, and everywhere ochres, and other sources of color, were used. In connection with rituals, body decoration with these pigments was common; also, in the north, decoration of burial poles and in the desert interior, ground drawings and sand paintings, were common. Pigments held symbolic significance, for example, red ochre was highly valued as the metamorphosed blood of the ancestral creative beings. The adoption of synthetic paints and other portable enduring materials has resulted in new media and forms of visual art expression being widely adopted. Basketry, pottery, sculpture, and glasswork are all part of the prodigious Aboriginal art production, and the art market now provides a significant source of income for many Indigenous artists. Outputs on canvas and bark have risen dramatically and state and national awards and major exhibitions encourage prominent artists, whose works also sell well in overseas markets.

The most widely recognized Aboriginal people in Australia's media today are sports stars, who are significantly overrepresented in two of the most popular codes: Australian Rules football (10%) and rugby (11%). Many women have achieved sporting fame, notably tennis star Evonne Goolagong and Olympics gold medal winner Cathy Freeman. Many Aboriginal people take part in a variety of amateur sports such as netball and basketball.

Sovereignty Issues and Relations with National and Local Governments

Organized Aboriginal political activity and calls for reform and equality date back at least to the 1930s, but intensified in the 1970s when articulate leaders drew inspiration for strategizing from Black Power movements in the United States. Resistance accelerated in the 1970s with the institution of an elected advisory body, the National Aboriginal Consultative Committee. The NACC's most prominent leaders and spokespersons were Aboriginal people of mixed descent, whose ethnic identity often came under fire, both from White Australian opponents and some Aboriginal people in remote areas.

Other challenges for Aboriginal nation-builders have been to overcome the problems of distance and diversity of lifestyle and outlook and to present a unified front and positive self-image in the face of media coverage that is often negatively focused on crime and 'dysfunction.' Any claim to represent Aboriginal people is potentially conflict with strongly maintained traditional values surrounding local autonomy, which limit the ability of individuals to 'speak for' others. Following colonization, and especially after Federation, Aboriginal people were not included as citizens with the full array of rights entailed in this status, and relations with governments at all levels have been fraught. Over many decades prior to the 1980s, a largely hidden policy of removing children by government fiat from their natural parents was in force, and for many its consequences were dire. In the 1990s, there was a well-publicized national inquiry into 'the stolen generations.' A 1994 Australian Bureau of Statistics survey revealed that more than 10% of indigenous Australians over the age of 25 reported having been separated and raised in isolation from their natural families. Many non-Aboriginal Australians were deeply moved to learn how severe, widespread, and recent were these traumas and gross violations of human rights. Several organizations are currently working toward the goal of reconciliation between the majority population and the indigenous minorities. In 2008, Labor Prime Minister Kevin Rudd offered a formal apology to members of the stolen generations, the symbolic importance of which was massive. Broadcast nationwide, it triggered emotional reactions across the Australian population.

Land Rights

For the last 50 years, land rights have loomed large in Indigenous political struggle. Notable breakthroughs include both federal and state legislation specifically recognizing Indigenous Australians' claims of ownership (e.g., the Aboriginal Land Rights (Northern Territory) 1976). The most momentous legislation to date is the Native Title Act 1993. In 1992, in *Eddie Mabo and others v the State of Queensland*, Australia's High Court accepted that, under common law, the native title of Australia's indigenous inhabitants could be recognized, invalidating *terra nullius*. However, the Court also found that grants of all freehold and some leasehold titles by the Crown extinguish native title. This left most of southern and eastern Australia unavailable for claim; and only 10–30% of Aboriginal people stand potentially to gain some security of tenure. The early promise of Mabo has been dulled by successful Federal Government moves to weaken the Native Title Act and by protracted legal challenges to the claimants, but there has recently been some progress in speeding the claims process.

Outlook for the Future

Today, generalizing about Aboriginal Australia is difficult because of its enormous regional and social diversity. Certainly, there will be an increasing melding of the two 'nations' in Australia, though the contrasting population pyramids (showing a much younger, but shorter lived, Indigenous population that is increasing more rapidly than the national rate) will not soon blend into one. While there is much discourse, by

politicians, media commentators, and many within and outside the Aboriginal population, there is no consensus as to the causes of enduring inequality or the methods by which it can be ameliorated. Many non-Aboriginal Australians continue to blame 'dysfunction,' welfare state provisions and so on, and discount the debilitating impacts of a long history of dislocation, marginalization, oppression, and racism.

Of considerable concern to Indigenous leadership and organizations, governments and NGOs have been the glaring disadvantage of Aborigines and Torres Strait Islanders on every measure of well-being, especially health, education, and employment (2010). For example, median incomes for Aboriginal people are 35% of those of non-Indigenous Australians. Raising educational attainment and reducing welfare dependency are also often-stated goals, and there is a continuous debate about how best to achieve those goals. Health problems remain severe, and Western lifestyle diseases, notably diabetes, cardiovascular diseases, and renal failure now claim a disproportionately high number of victims. Alcohol abuse and associated violence and death are widespread. Indigenous people are also twice as likely as non-Indigenous people to be daily smokers and to suffer higher levels of psychological distress and shorter life spans than the population at large. In many areas, Aboriginal community life is also severely marred by massive and chronic unemployment and domestic and other forms of violence. The failure of 40 years of policies promising self-management has led many prominent Aboriginal leaders to push for a complete rethinking of government policies. Some urge the empowerment of communities to tackle head-on the social ills plaguing them and to devise their own strategies to escape from 'welfare dependency' and engage productively with the mainstream economy.

Influential Indigenous people are now working in the professions, academia, media, politics, and business. There is a Indigenous-controlled television network. University graduation rates are increasing, and leadership programs have successfully launched hundreds of young Aborigines into well-paid careers. Educational attainment for Aboriginal and Torres Strait Islander Australians has increased appreciably since the mid-1990s. In 2008, 37% of Aboriginal and Torres Strait Islander people aged 18 years and over (adults) had attained a minimum of Year 12 or a skilled vocational qualification, more than double the rate in 1994 (16%). Over the same time period, those completing a minimum of Year 10 or basic vocational qualifications increased from 48 to 71%. While relatively few Aboriginal and Torres Strait Islander adults continued on to complete a Bachelor degree or above, the rate increased to 5% in 2008.

Aboriginal people have, until relatively recently, been excluded from control over any significant resources. However, big changes have been occurring and more are imminent, particularly in Western Australia and Queensland, the mining states currently accounting for much of Australia's strong economic situation. After decades during which local Indigenous people were almost entirely excluded, the industry now provides training programs and employment in mining and associated service industries. Also, with the recognition of land rights and native title, mining royalties and other income from the industry are now flowing to affected Indigenous communities and land councils in many remote regions. Major organizational challenges nonetheless remain: principally, the creation and regulation of systems that ensure an equitable redistribution among Aboriginal land owners, with mechanisms to guarantee long-term benefits to affected populations.

Indigenous Australians are now an increasingly visible physical and cultural presence in most major cities and a significant demographic component of local populations in the sparsely populated northern and central regions. The situation is complex, but the cessation of restrictive discriminatory laws and the success of Aboriginal people in a range of activities have all contributed to notable improvements. (A large surge in people identifying as Aboriginal occurred in the 1990s and could be related to this more amenable social milieu.) Notable Aboriginal achievements in art, music, literature, sport, film, and dance have received much media attention; there are nationally televised award nights, and positive images of indigenous people appear to be seeping into the national consciousness. The rhetoric of survival and resistance has become more muted, displaced by a more culture-centred emphasis on Aboriginal commonalities and shared aspirations to mainstream participation. The gap separating them from the mainstream is still considerable, but is narrowing and there are grounds for optimism.

Bibliography

ABS (Australian Bureau of Statistics), 2011. Overcoming Indigenous Disadvantage: Key Indicators 2011. Productivitiy Commission, Melbourne.

AIATSIS (Australian Institute of Aboriginal and Torres Strait Islander Studies), 2009. The Little Red Yellow Black Book: An Introduction to Indigenous Australia. Aboriginal Studies Press, Canberra.

Beckett, Jeremy, 1990. Torres Strait Islanders: Custom and Colonialism. Cambridge University Press, Cambridge.

Berndt, Ronald M., 1974. Australian Aboriginal Religion. (4 Fascicles). Brill, Leiden.

Berndt, Ronald M., Berndt, Catherine H., 1988. The World of the First Australians, revised ed. Aboriginal Studies Press, Canberra.

Broome, Richard, 2010. Aboriginal Australians: A History Since 1788. Allen & Unwin, Sydney.

Burridge, K.O.L., 1973. Encountering Aborigines. A Case Study: Anthropology and the Australian Aboriginal. Pergamon, New York.

Edwards, William H. (Ed.), 1988. Traditional Aboriginal Society, second ed. Macmillan, South Yarra.

Keen, Ian, 1996. Knowledge and Secrecy in an Aboriginal Religion. Oxford University Press, Oxford.

Keen, Ian, 2004. Aboriginal Economy and Society: Aborigines on the Threshold of Colonisation. Oxford University Press, Melbourne.

Maddock, Kenneth, 1982. The Australian Aborigines: A Portrait of Their Society, second ed. Penguin, Melbourne.

Meggitt, Mervyn J., 1962. Desert People: A Study of the Walbiri Aborigines of Central Australia. Angus & Robertson, Sydney.

Mulvaney, D. John, White, J. Peter (Eds.), 1987. Australia to 1788. Syme and Weldon Associates, Sydney.

Munn, Nancy D., 1973. Walbiri Iconography: Graphic Representation and Cultural Symbolism in a Central Australian Society. Cornell University Press, Ithaca.

Reynolds, Henry, 2006. The Other Side of the Frontier: Aboriginal Resistance to the European Invasion of Australia. UNSW Press, Sydney.

Rowley, Charles D., 1970. The Destruction of Aboriginal Society. Australian National University Press, Canberra.

Rowse, Tim, 1998. White Flour, White Power. Cambridge University Press, Cambridge.

Stanner, William E.H., 1979. White Man Got No Dreaming: Essays 1938–1973. Australian National University Press, Canberra.

Stanner, William E.H., 1989. On Aboriginal Religion. Oceania Monograph 36. Sydney. (Reprinted with an appreciation and introduction.)

Sutton, Peter, 2009. The Politics of Suffering. Melbourne University Press, Melbourne.

Tonkinson, Robert, 1991. The Mardu Aborigines: Living the Dream in Australia's Desert, second ed. Wadsworth/Thomson Learning, Belmont, CA.

Australia, Sociocultural Overviews: Australian Settler Society

Adrian Peace, School of Social Science, University of Queensland, Ipswich, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

When global surveys detail the countries to which people would most like to migrate, Australia consistently emerges at the top of the league table, or very near to it. The grounds on which these global judgments are arrived at are never made clear. But it is assuredly the case that Australia's international image is that of a prosperous, stable, and successfully multicultural nation. It is seen as forward- rather than backward-looking, notably free from the historical enmities and enduring conflicts which afflict other, late capitalist societies. Broadly speaking, this assessment is accurate enough, but Australia is an increasingly complex society in which less progressive cultural features are discretely hidden from view.

Prosperity and Paradox

Australia is assuredly a prosperous society, ranking about sixth in the world today, which means that its affluent middle classes are structurally and culturally all-important. Australians enjoy one of the highest rates of private property ownership in the Western world: household possession of a second or third vehicle, including expensive 4WDs, is commonplace; an extensive range of costly consumer goods is the norm; and even the ownership of a second dwelling for vacation use is considered nothing untoward. There is a strong commitment also to costly hedonistic activities, including multiple forms of gambling which have become a significant source of revenue for the country's two largest states. According to reliable sources, Australian adults lose on average AUD$1300 per year on gambling, the highest rate in the world. The country comes to a complete halt when the Melbourne Cup, its premier annual horse racing event, takes place; the turnover in bets on this race alone totaled AUD$93.6 million in 2011.

Australia is also a society distinguished by political stability, particularly by comparison with other nation states which grew rapidly in the twentieth century through migration. Unlike South Africa or Israel, for example, the political process has throughout been strikingly free from outright conflict. Constitutional crises have been rare, elections are routine affairs, and changes in government at federal and state levels are uncomplicated. This is by no means to suggest that the political arena is free from significant lines of cleavage, or that public confrontations are unknown. The divide between Aboriginal and non-Aboriginal interests is substantial, and public expression of it is recurrent in a diversity of ways. As we shall see, racism and xenophobia sometimes seem only just below the surface of everyday life, and how the country should deal with the arrival of refugees by boat from the north is a source of recurrent dispute. But the absorption of major social conflicts is undoubtedly among the more pronounced qualities of the political system in its broadest sense (see Megalogenis, 2012).

Yet despite this broadly based affluence and noteworthy political order, it is also the case that a distinct sense of unease and uncertainty over Australia's national and cultural identity is often remarked upon, particularly by authors, playwrights, and poets. As the end of the twentieth century drew near, David Malouf, who is one of the best-known authors and respected commentators in the country, suggested that Australians attended excessively to the opinions of others, including Americans and Asians, and should pay more attention instead to "our own better and freer and more adventurous selves" (Malouf, 1998). Now that we are well into the twenty-first century, it is clear that this is easier said than done, especially as a result of momentous changes of late in the political economy of East Asia, for this cultural unease and societal uncertainty continue apace. At one level, this relates to how Australians are different from one another, more specifically what is entailed by multiculturalism, and how it is defined and ordered in particular ways, for there are, in the view of some, cracks in the façade which are not easily papered over. At another level, it concerns the position of Australia in the modern global system, above all in light of the growth in material prosperity being increasingly driven by complex and uneven trading relationships with China.

Rural Change

Australia's small and peripheral economy – the current population is just over 22 million and ranks 50th in the world today – has long been integrated into the global marketplace, mostly to distinct economic advantage. Throughout the first half of the twentieth century, the export of agricultural produce and minerals to Britain and mainland Europe generated a prosperity from which only particular minorities were wholly excluded.

This affluence was sustained after World War II as new international markets were added to traditional ones. By the early 1970s, however, this export-dependent economy proved distinctly vulnerable to changed global conditions (Bell, 1997), and in rural areas the major impact fell on the small to middle-size farms. Declining commodity prices, increased interest rates, price hikes in key inputs (such as fertilizer), and rising transport costs impacted most severely on the bottom 25% of the rural hierarchy; and their problems were compounded by the loss of wage-earning opportunities as well as cutbacks to infrastructure and welfare services. By the end of the twentieth century, the combined impact of these changes had resulted in the loss of large numbers of young people to the cities, the closure of somewhere between 2000 and 3000 production

units each year, and increased dependence on welfare payments among the already marginal, along with a widening of the material divide between the prosperous and those who struggled to survive, the 'battlers in the bush.'

Lawrence (1987) made it clear that these changes to the rural class structure were inseparable from the increased hegemony of a small number of transnational corporations over the agricultural system. These corporations exercised tight control over the supply of critical inputs to the rural sector (mechanical equipment, building materials, seeds, and so on) as well as its major outputs (processing, canning, and transportation). The larger units of rural production that were owned by, or were contractually integrated with, these transnational organizations realized higher rates of profit than previously, but smaller farms found it increasingly difficult to survive and were often swallowed up by larger ones.

These processes have been compounded into the current century to the point at which just two transnational corporations exercise effective dominance over the country's domestic food sector while using their profits to expand into new spheres of commerce and service industry. But the rural sector has also been greatly impacted by a severe decade-long drought which has, somewhat inevitably, taken its greatest toll on moderate to small-size agricultural units. To be sure, there is nothing new about drought conditions in Australia, the driest continent on the planet. But the impact of the most recent one not only highlighted, but was also compounded by, environmental problems generated by long-term economic exploitation and political mismanagement. The most significant region in this regard has been the vast Murray-Darling river system which covers major areas of Queensland, New South Wales, and Victoria, before entering the Southern Ocean in South Australia. Although pivotal to the entire country's material well-being as a major source of foodstuffs for domestic and overseas markets, decades of agricultural pollution, uneven water usage, and widespread wastage of water due to a lack of infrastructural maintenance have been detailed in a series of scientific reports generated over at least two decades. But rivalries between the states with different commercial realities to protect, and the Federal Government's inability to moderate or override these political divisions, have resulted in delay and deadlock.

Community protest has become a recurrent feature across this rural landscape, and not only inside the Murray-Darling river system. As mining corporations have penetrated rich agricultural areas in the eastern states to explore the prospect of shale gas production, families and communities have raised what are by now customary concerns elsewhere about the impact of intensive 'fracking' on valuable farmland and irreplaceable water supplies. What is common to these developments is that ordinary folk, with no track record of political protest, openly vent their substantial concerns about community survival and an equally profound disenchantment with government incompetence and political deception. It is frequently the case that, in the course of grassroots mobilization, these external intrusions bring to the fore and even exacerbate local differences and regional distinctions; but the root sense of belonging to specific locales, and the authentic ways of rural life inscribed over these landscapes, is by no means compromised as a result.

A further environmental development which has impacted greatly on rural communities in recent times has been the explosive spread of feral creatures, including such non-native animals as wild dogs, camels, buffalo, pigs, and most notorious of all, the cane toad. Not only do some of these creatures pose a threat to the survival of small, indigenous species, which might otherwise be a significant attraction in rural tourism, along with the kangaroo population which is widely held to be in 'plague' proportions, their combined impact has the capacity to drive even moderate-size farms to the wall.

In these complex circumstances, the independence and the fortitude of farm family producers, so integral in the past to the material prosperity of Australia, become increasingly illusory. The cultural significance of these changes is as wide-ranging as their economic ones. The bush has always provided the more powerful symbols from which cultural specialists forged narratives about the distinctiveness of the Australian character and its relation to the environment. Self-reliance, egalitarianism, and mateship are all morally weighted idealizations drawn from outback experience, and all have been integral to the process of distinguishing Australian culture and the nation from their European origins. When the realities of rural experience are characterized by increasing inequalities of class, threats to community life, corporate domination and environmental dispute, the bush and its authentic way of life inevitably lose their iconic value in the political construction of Australia as a distinct cultural space.

Migration, Urbanization, and Multiculturalism

The majority of Australians reside in a handful of capital cities and their suburban satellites at different points of a vast coastline. These are the metropolitan centers which expanded exponentially with the arrival of immigrants post World War II and therefore became the crucibles of Australia's experience with multiculturalism. Hugo, a demographer, has noted that Australia underwent an unparalleled transformation throughout the twentieth century from "an overwhelmingly Anglo-Celtic, homogenous population in which 95.2% had been born in Australia, the United Kingdom or Ireland, to one of the world's most multicultural societies" (Hugo, 2002), Richards, a historian, has argued that the relative absence of major disharmony and dispute throughout this social transformation stood as testament to the success of Australia's commitment to multiculturalism(Richards, 2008).

The main warrant for these positive accounts has been the way in which the country absorbed the huge postwar intake of immigrants, and the experience of the second and third generations (see contributions to Jupp, 1988). Between 1947 when the population stood at 7.5 million and 1987, some four million immigrants entered the country. Since the United Kingdom and Ireland were no longer able to meet Australia's demand for labor, by the 1960s substantial numbers were recruited from southern Europe and the Levantine. Regardless of their places of origin, these new immigrants were offered citizenship to encourage permanent residence.

The result was an exceptional increase in Australia's culturally hybrid character. Inner city suburbs were rapidly transformed from predominantly Anglo-Celtic working class neighborhoods

into multiethnic lower-class suburbs. Alternatively, these new Australians gravitated toward the burgeoning outer suburbs of cities like Sydney and Melbourne where opportunities for wage employment in modern factories and relatively cheap housing were to be found. The crucial development was that overall (and there were some exceptions) immigrants from countries like Italy, Greece, and Yugoslavia, tended to loosely cluster in a range of residential areas rather than becoming concentrated in inner city ghettoes. This pattern of clustering, the building blocks of which were informal familial and friendship networks, allowed new arrivals to gain a foothold residence- and occupation-wise with the assistance of fellow ethnics who had arrived before them. Subsequently, it was relatively straightforward for themselves and their children to move upward socially on the occupational ladder and outward spatially to better serviced, more prestigious suburbs. Most important of all, upcoming generations married extensively across ethnic boundaries and expanded their friendship circles similarly.

All this was made possible in the second half of the twentieth century by a steady growth in societal prosperity in which second and third generations were able to share. In addition to acquiring the material trappings of the Australian middle class, they were able to retain select aspects of ethnic identification which, for the most part, were discretely rather than aggressively expressed, as was considered appropriate to 'the Australian way of life.' Speaking in one's native language around the home was coupled with the ordering of domestic space, the choice of foodstuffs, the organization of entertainment, especially participation in sport, in ways considered consistent with being Italian-Australian, Greek-Australian, and so on. On a more public plane, food and dance carnivals as well as ethnic clubs signaled and sometimes satisfied a sense of nostalgia for a cultural inheritance left behind several decades previously. Particularly among immigrants of West European origin, visits to the village or town of birth took on the character of a secular pilgrimage, but the upshot was generally to underscore the all-round benefits of long-term residence in Australia and the formal acquisition of citizenship.

These generalized processes could be readily glossed by the country's political elite as solid, even incontrovertible, evidence of the assimilative properties of a tolerant, open-handed, and relaxed late-twentieth-century Australia. Government agencies highlighted the achievements of multiculturalism, which had been adopted as state policy in the early 1970s. In the past decade or so, however, this progressive narrative has come under close scrutiny for several reasons.

The first is that a more complex picture has emerged from the social analysis of a greater diversity of ethnic immigrants, the organization of their everyday lives, and their connectedness with their places of origin. Thomas (1999) described in detail how migrants from Hanoi construct their suburban homes and gardens in intricate fashion to sustain their cultural traditions; but behind these practices is to be found a deep-rooted sense of fear and resentment, hardly surprising considering the mammoth displacements to which they have been subject. Baldassar (2001) argued that Italian-Australians return to their home villages with all the trappings of material success, but those who have stayed behind are often unimpressed, so that the returnees get bound up in a "circular quest for an increasingly elusive identity." Hyndman-Rizik (2008) argued that Maronite-Christians, who have been migrating to Australia since the 1950s and often been economically successful, can rise socially only so far before hitting a glass ceiling. They return home with high expectations of being respected there, but the result instead is a sense of being out of place in both Lebanon and Australia. These, and other studies, suggest that, behind the immediate benefits of material prosperity, there are many issues concerning the choice and expression of social identity which are hidden from view.

The second reason is that there has been a growing sense of resentment and hostility among some immigrant populations, especially the second generation of adults now assessing the life chances of their own offspring, that in order to become fully recognized as Australian, it has been obligatory to meet the expectations and beliefs of the Anglo-Celtic core which retains, or at least believes it retains, control over the key standards or mores which define membership of Australian society. This reaction from below has become pronounced in recent times because this hegemonic positioning was repeatedly articulated by Prime Minister John Howard who led a highly conservative Federal government between 1996 and 2007. Hage (1998, 2003), a sociologist, has addressed this issue analytically. While not specifically targeting multiculturalism as such, he argues that powerful agents in and around the modern state articulate a nationalist discourse which lays out, if not determines, the terms on which cultural others might be included within mainstream citizenry. This nationalist discourse, Hage states, is "not necessarily about excluding/destroying otherness but about regulating the modality of inclusion" (Hage, 1998: 174), and it is the fantasy of a 'White Nation' shared by groups which would not readily consider themselves ideologically allied. Hage's analysis does not engage much with the commonplace discourses of ethnic populations, nor the tensions and conflicts which disrupt their everyday worlds, but supporting evidence can be gleaned from other sources.

This brings us to the third, closely related factor, which is the proposition that there is a façade-like quality to multiculturalism: below a veneer of tolerance is to be found a sediment of xenophobia which from time to time explodes into the public domain. The impact is considerable because these occasions are so thoroughly turned over by the media. The most dramatic episode of this type comprised 'the Cronulla riots' of 2005 (see Cowlishaw, 2007) named after the beachfront suburb of Sydney in which they occurred, and involved several thousand predominantly young men aiming to target 'Lebanese' people. The trigger was an attack on two lifesavers by men of Middle Eastern appearance, while the rapid mobilization of this assembly was facilitated by text messaging and media coverage. The crucial point to recognize is that the label 'Lebanese' serves as a shorthand in some popular discourse for the Arab Other, an imagined population of outsiders who, it was assumed by the rioters, do not, and never will, integrate into the 'White Nation' of which those involved considered themselves members. The beachfront at Cronulla, itself an icon of Australian paradise and its relaxed culture, had to be reclaimed from 'uncivilized Arabs' with no right to be there. Sociological and pseudosociological accounts of the riots proliferated in the aftermath, but the idea that an ugly thread of racism is barely hidden by a veneer of multicultural harmony was commonly encountered.

By far the most important development has been the long-running crisis focused on the arrival on Australia's northern shores of asylum seekers from the Third World, notably Afghanistan, Iraq, and Sri Lanka. The number of refugees involved are insignificant by contrast with the scale of the problem facing prosperous European countries; Australia currently ranks 68th in the world on a refugee-per-capita basis. But, as has frequently been pointed out, the threat of 'invasion from the north' has long been an element of conservative nationalist discourse, so that the arrival of unheralded boats in unspecified number from Indonesia is rapidly incorporated into overblown scenarios about serious breaches to national sovereignty. The incident which received most international attention involved the *Tampa*, a Norwegian freighter which in 2001 responded to a request by Australian authorities to rescue over 400 refugees from a sinking vessel, only to find that the government had no intention of allowing them to enter Australian waters. There is no doubt that the Howard government's 'hard line on illegals' found extensive support among the Australian public, including its 'Pacific solution' involving the incarceration of refugees in processing centers overseas. The handling of asylum seekers has proved intractable and it continues to generate tensions with Australia's closest neighbors: but it has also exposed latent xenophobic sentiments at the top and the base of the social hierarchy.

The Middle Classes and the Changing Political Economy

The distinctiveness of social class in the Australian context has been regularly addressed in the sociological literature, including the widespread notion that, relatively speaking, this is a classless society. This belief is, or perhaps more accurately was, less about the presence of classes as the prevalence of egalitarianism – the idea of everyone being equal, notwithstanding significant lines of social differentiation – so that patterns of deferential behavior considered pervasive in especially British society were absent by comparison. In historical terms, the absence of an aristocracy or landed gentry was considered critically important (see Connell and Irving, 1992).

Presently, sociological analyses are particularly concerned to unravel the social intricacies of Australia's heterogeneous middle classes, which contrast markedly with those whose influence has been on the wane for some time. These interpretations are informed by the work of Bourdieu, and focus in particular on the cultural distinctions between a relatively young, prosperous, and fully urbanized middle-class stratum, and an older, less well-off, and predominantly regional one. Bennett et al. (2001) argue that the major source of their contrasting cultural tastes lies in the realm of formal education, so that those with tertiary-level qualifications are able to position themselves in Australian society in distinctive ways. It is, for instance, critically important to their self-identification that they take great care of the body, a significant site of pleasure and satisfaction. Survey materials reveal that professionals and managers participate extensively in sporting activities (tennis and golf) rather than being mere spectators. By contrast, supervisors and manual workers are inclined toward passive engagement, so that as a general rule the less well-off you are, the less inclined you are to be physically active.

Then again, significant differences are to be found in the distribution of literary tastes. Professional people are most strongly oriented to a high culture of classical authors, contemporary novels, and biographies, while manual workers, if they read at all, opt for books on gardening, sport, and leisure. Employers and the self-employed are different again, occupying something of a middle ground, but some of these variations are markedly influenced by the gender composition of these classes. Third, there are clear patterns in the relation between class standing and musical preferences, so that those of higher rank opt for 'high brow' genres (classical music, opera, and the like) with manual workers opting for country and Western music. The distribution of musical tastes across the board, however, is more complicated than a straightforward elite-mass distinction. Bennett et al. (2001) found that all classes nowadays prefer the middle-brow form of 'easy listening,' and pop and rock music, which leads the sociologists to propose a novel distinction between omnivores and univores. The former have musical preferences which cut across these boundaries of taste, the latter are much more restricted. At the end of the day, the explanation of these and other cultural tastes consistently turns on the level of education attained by those of different rank, for it is the extent of knowledge about the variety of available genres which opens up a broad field of choices.

Sociological findings such as these are useful to exploring the fine distinctions which constitute and compound the heterogeneity of the Australian middle classes; they raise significant questions about the relation between the knowledge classes, government funding, and the reproduction of economic and social inequalities into the next generation. At the same time, apart from a marked degree of predictability in their findings, a common failure is to leave out of the equation the most important of all middle-class resources, namely the private domestic properties around which revolve many of the securities and insecurities of middle-class existence in twenty-first-century Australia. Domestic property prices in the major metropolitan areas are now among the highest in the world: mortgages and rental payments consume an increasingly substantial percentage of two-income households, which is the norm. It is hardly surprising, then, that commercial television programs centered on home improvement strategies are astonishingly popular: indeed, their close reading might tell us considerably more about the fetishisms focused on the family home than any sociological survey of cultural choices.

The other weakness of analyses based on sociological surveys is that their findings are rarely integrated with broader transformations in the political economy of Australia, above all its relations with its global trading partners. This is, however, an increasingly pressing requirement where the interests and the investments of the more cosmopolitan knowledge classes are concerned.

The crucial development by far has been the emergence of China as Australia's major commercial partner, for this has repercussions across the Australian landscape, including its class structure and the distribution of power. The main factor in this equation has been China's near-insatiable demand for the raw materials of industrial production on the one hand and the ability of Australia's resource sector to meet much of it on

the other. Coal remains the country's main mineral export: Queensland alone provides one-third of the world's coking coal, which is needed to produce steel. Oil production is currently worth about AUD$30 billion each year, although, as this declines, its contribution to the economy is being overtaken by gas from Western Australia. The major source of iron ore is now the Pilbara region of Western Australia: the ore's high quality ensures that it is important to China's vast building program and industrial expansion. As is often remarked, the tyranny of distance in the past between Australia and the major marketplaces in the northern hemisphere has been turned on its head nowadays by the Antipodean advantage of proximity to Asia. The vast movement of commodities has been accompanied by the extensive mobility of people, including tertiary-level students and well-heeled middle-class tourists from China (in 2012, 70 000 and 550 000, respectively).

Inside Australia, the effects of these transformations in political economy have been wide-ranging. The first has been changes to regional economies of significance, so that while, for example, the renowned Iron Triangle of South Australia, once one of the country's major areas of industrial production, inexorably slides into decline, the Pilbara to the country's northwest is fast consolidating its position as the foremost region of capital investment and profit accumulation. Elsewhere, in New South Wales and Queensland, rural areas well-known for high-quality agricultural production and winegrowing are being bought out by transnational corporations for the extraction of coal and coal seam gas destined for export. The scale of some of these extractive enterprises is exceptional. The strikingly named China First Coal Project in Queensland, for instance, requires a single mining lease extending over some 21 561 km^2, which means that it is larger than the Northern Territory's internationally renowned Kakadu National Park, the largest national park in the country (Cleary, 2012: 17).

The second major impact involves changes to the upper reaches of the class structure. At one level, the nature of this is conspicuous enough in the emergence of a new plutocracy of resource entrepreneurs whose prosperity firmly locates them in the upper ranks of wealthy people at the global level. Ownership of the key means of twenty-first-century industrial production is often matched with an interest in the control of select media, which ensures that they become the focus of the public gaze. Far less easy to specify is the extensive stratum of senior managers, finance advisors, lawyers, accountants, major shareholders, and the like whose authority and influence is heavily reliant on the success of this new plutocracy, or more precisely the mobility of capital which is so integral to their power across the nation state.

Mobility is also key to the emergence of a new labor aristocracy of foremen, supervisors, and highly skilled workers, whose incomes and life styles on the one hand set them apart from the bulk of Australia's unionized workforce, but on the other hand lay down benchmarks for remuneration and conditions of employment to which many others can only aspire. As with the labor aristocracy brought into being by the first industrial revolution in Europe, spatial and social mobility is integral to the significance of this growing labor elite, as indexed by the novel FIFO label which has entered the national vocabulary. This refers to 'fly in–fly out' workers who reside in metropolitan cities or regional towns, and are flown to mining locations or oil fields for intensive periods of work while living in basic accommodation. Mining and oil companies see this as a flexible employment arrangement that obviates the need to build company towns, which was the pattern in the 1980s and resulted in 'ghost towns' once the resource was fully extracted. The costs to family and community life which this change involves are little understood, but companies are assuredly prepared to pay exceptional wages in order to exploit and to retain the high-level skills of these employees.

The Middle Classes and Australia's Environmental Crisis

Undoubtedly the most important impact of the 'resources boom,' however, has been the shielding of the mainstream middle classes from the global financial crisis (GFC) of recent times. It is certainly the case that the rate of economic growth in Australia has declined and that the value of some property in major cities and towns has depreciated. By contrast, however, with comparable classes in the northern hemisphere, Australia's middle classes have been able to sustain the high levels of conspicuous consumption to which they were well accustomed by the late twentieth century. This serves to underscore the paradox of the regional political economy in which the country is now inexorably locked: its material prosperity is inseparable from the economic underdevelopment, especially the availability of cheap productive labor, which characterizes its east and southeast Asian trading partners.

In other words, there has been no compunction for these affluent classes to reevaluate the patterns of overconsumption which are now widely acknowledged elsewhere as integral to environmental crises at regional and global levels. A number of surveys have indicated that people are apprehensive, even anxious, about the country's dependence on China as by far the most dominant force in the regional political economy, and this public concern spikes when especially iconic companies or sectors of production are acquired by the Chinese government, or its representatives. Thus, at the time of writing (September 2012), a major debate has broken out as to whether a Chinese textile consortium, Shandong RuYi, should be allowed to purchase a vast irrigation system in Queensland which, as part of the ailing Murray-Darling river system, has the capacity to produce over 10% of the country's entire cotton output. In such instances, the rhetorical terminology of 'the national interest' is parlayed from all possible angles, albeit without much by way of clear resolution. By contrast with such debates over productive resources, however, the excessive consumption patterns endemic to the middle classes receive little critical attention. Overconsumption is normalized and regularized to an exceptional degree in Australia, and the shielding of the middle classes from the most important negative impacts of the GFC has made it all the more so.

This omission is the more notable because the subject of environmental crisis is among the more significant threads running through the national political discourse. A number of environmental issues have been debated at the national level for many years, for example, the logging of old growth forests for woodchips, while others have emerged much more

recently, for instance, the spread of 'fracking' by exploratory companies referred to earlier. To a varying extent, such issues have provoked the spread of recurrent, even widespread, protest activity, and significant environmental gains have been recorded. But these localized activities only involve a small percentage of the population directly, and this has always been the case. For most middle-class Australians, their involvement has ranged from being vicarious to disinterested. The focus has rather been on expressing their environmental concern by spending more rather than less, of purchasing additional 'green' commodities – from organic foodstuffs through hemp-derived clothing to electric-powered vehicles – on top of those customarily consumed in the typical, prosperous, two-income household.

In short, the idea of seriously reducing expenditure and consumption remains quite alien: a fairly typical attitude seems to be that, if using low-energy light bulbs or sorting out the household waste into different containers is all that is necessary, high consumer middle-class existence can proceed as normal. Furthermore, if societal-scale environmental initiatives undertaken by government might increase the cost of living, the middle classes expect to be fully compensated rather than shoulder even a modest burden. There has been no better illustration of this belief, in combination with the new dominant class interests described above, than the public response to an initiative in 2012 by the Labor government to impose a carbon tax on the country's worst 500 polluting corporations. Since Australia is at the head of the global league of developed nations in terms of atmospheric pollution per capita, this was a progressive move, and when it was first mooted public opinion was generally supportive. When it became clear, however, that the carbon tax might result in even very modest increases to the average household budget, a sustained campaign underwritten by the major mining corporations that now dominate the industrial sector had little trouble in turning public opinion against the initiative, which then had to be significantly watered down.

Conclusion

The specifics of social and cultural relationships inside Australia have always been inseparable from changes to its broader political economy. Whether it is hybrid ethnic relations or involved class relations that are under the microscope, all are to some extent or other the product of Australia's high level of integration into its regional context and the global marketplace beyond. While materially prosperous and politically stable, as an ex-colony this nation state has throughout wrestled with the cultural uncertainties generated and sustained by economic dependency. There is no doubt that their specific character, and the politics which they generate, will be markedly different to those of earlier times.

See also: Australian Aboriginal Society and Culture: An Overview; Multiculturalism, Anthropology of.

Bibliography

Baldassar, L., 2001. Visits Home: Migration Experiences between Italy and Australia. Melbourne University Press, Melbourne.

Bell, S., 1997. Ungoverning the Economy: The Political Economy of Australian Economic Policy. Oxford University Press, Melbourne.

Bennett, T., Emmison, M., Frow, J., 2001. Accounting for Tastes: Australian Everyday Cultures. Cambridge University Press, Cambridge.

Cleary, P., 2012. Mine-Field: The Dark Side of Australia's Resources Rush. Black Inc., Collingwood.

Connell, R., Irving, T., 1992. Class Structure in Australian History: Poverty and Progress. Longman Cheshire, Melbourne.

Cowlishaw, G., 2007. Introduction: ethnography and the interpretation of 'Cronulla'. The Australian Journal of Anthropology 18 (3), 295–299.

Hage, G., 1998. White Nation: Fantasies of White Supremacy in a Multicultural Society. Pluto Press, Annandale, NSW.

Hage, G., 2003. Against Paranoid Nationalism. Pluto Press, Annandale.

Hugo, G., 2002. A Century of Population Change in Australia. ABS Year Book Australia 2002. Australian Bureau of Statistics.

Hyndman-Rizik, N., 2008. 'Shrinking worlds': Cronulla, anti-Lebanese racism and return visits in the Sydney Hadchiti Lebanese community. Anthropological Forum 18 (1), 37–55.

Jupp, J. (Ed.), 1988. The Australian People: An Encyclopedia of the Nation, Its People and Their Origins. Angus and Robertson, Sydney.

Lawrence, G., 1987. Capitalism and the Countryside: The Rural Crisis in Australia. Pluto Press, Sydney.

Malouf, D., 1998. A Spirit of Play: The Making of Australian Consciousness. ABC Books, Sydney.

Megalogenis, G., 2012. The Australian Moment: How We were Made for These Times. Hamish Hamilton, Victoria.

Richards, E., 2008. Destination Australia: Australian Immigration from 1901. UNSW Press, Sydney.

Thomas, M., 1999. Dreams in the Shadows: Vietnamese-Australian Lives in Transition. Allen and Unwin, Sydney.

Authentic Leadership

William L Gardner and James D Carlson, Texas Tech University, Lubbock, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Authentic leadership is a pattern of leader behavior that includes self-awareness, balanced processing of information, relational transparency with followers, and an internalized moral perspective. Authentic leader–follower relationships arise when both leaders and their followers remain true to their inner thoughts, beliefs, and experiences. Empirical research has established positive relationships between authentic leadership and positive individual (e.g., engagement, job satisfaction, job performance), group (e.g., teamwork), and organizational (e.g., financial performance) outcomes.

People who lead others while enacting their true selves, rather than a fake or distorted self, are said to engage in authentic leadership. The construct of authentic leadership has been formally defined as a pattern of leader behavior that includes self-awareness, balanced processing of information, relational transparency with followers, and an internalized moral perspective. Within the context of an authentic leader–follower relationship, both leaders and their followers remain true to their own inner thoughts, beliefs, and experiences. Empirical research on authentic leadership has established positive relationships between authentic leadership and many positive individual (e.g., engagement, job satisfaction, job performance), group (e.g., teamwork), and organizational (e.g., financial performance) outcomes.

Historical Overview

Scholarly and practitioner interest in authentic leadership has grown rapidly over the past decade, spurred on by a mixture of economic and societal forces. Particularly influential have been a series of shocks to the world economy, including the global financial crisis of the late 2000s. In attempting to discern the cause of the crisis and develop an understanding of how to prevent such events from occurring in the future, many have come to see leaders in various organizations as being partially culpable. For example, leaders of key US financial institutions authorized ethically questionable practices, such as securitizing subprime 'toxic' mortgages. Other leaders permitted investment or lending practices that were unacceptably risky, hampering responsible recovery when financial conditions worsened. While no one leader's actions alone caused this financial crisis, the cumulative ethical lapses of multiple leaders has been identified as a key factor that contributed to the financial meltdown. This is significant, considering the far-reaching impact of the crisis. Once the US economy fell into recession near the end of 2007, tightly interlocked global financial markets allowed for a chain of events that spread financial hardships throughout the world economy. Some have termed the resulting decrease in World GDP in 2009 as 'the Great Recession'. Today, many nations are still dealing with the effects of the financial crisis (e.g., the Eurozone).

For leadership scholars, the global financial crisis is just one example in a stream of modern events that have proven to be challenging for conventional leadership theories. That is, historically, leadership theories have not addressed what it means to be an authentic leader, or what the outcomes that accrue from such leadership might be. Research on authentic leadership has since helped to fill this gap in leadership theory and knowledge.

Early Conceptions

As with most new theories, initial attempts to explain authentic leadership took varied approaches. Given that current models of authentic leadership build on past research, understanding the history of the construct will help to place current authentic leadership theory and research in context.

An unusual aspect of authentic leadership theory is the fact that its growth and development coincides with the emergence of popular practitioner writings on the same subject. That is, popular works devoted to authentic leadership coevolved with the academic literature in the early 2000s, as practitioner interest in the topic contributed to scholarly interest, and vice versa. Key contributions to the practitioner literature include two highly popular books written by Bill George, the former CEO of Medtronic, titled *Authentic Leadership* (George, 2003) and *True North*: *Finding your Authentic Leadership* (George and Sims, 2007). Lauded by many for being a particularly authentic leader, George interviewed more than a 100 leaders in preparation for writing *True North*. He concluded that authentic leaders possess five important characteristics: (1) an understanding of their purpose; (2) clear and salient values with respect to doing the right thing; (3) a focus on trust in their relationships with others; (4) self-discipline in acting on their espoused values; and (5) a passion for their mission, enabling them to act from the heart. While evidence-based and well thought out, George's model is not intended to serve as a scientific theory.

Scholarly attention devoted to the study of authentic leadership began to rise shortly after the emergence of these practitioner approaches. The earliest theoretical model of authentic leadership development was proposed by Luthans and Avolio (2003). A key feature of this model was the inclusion of positive organizational behavior, which grew out of a movement toward positive psychology (Seligman and Csikszentmihalyi, 2000). Positive psychology emerged as a reaction to a persistent tendency in the field of psychology to focus on deficiencies and weaknesses, rather than on vitality and strengths. Research within the field of positive psychology therefore focuses on

understanding what promotes experiences of joy, happiness, and flourishing.

Luthans and Avolio (2003) suggested that authentic leaders influenced others by invigorating them through the activation of 'states' of positive organizational behavior. In contrast to the term 'trait', which refers to characteristics that are stable across contexts and resistant to change, 'states' can be developed and managed to achieve improved performance – an important implication for the development of authentic leadership. Subsequently, these positive organizational behavior states were subsumed under a more general, multidimensional construct called psychological capital, or PsyCap (Luthans et al., 2007). This construct incorporates the states of self-efficacy, optimism, hope, and resilience.

In their original model of authentic leadership, Luthans and Avolio (2003) suggested that the theory arose from the nexus of positive organizational behavior, transformational/full-range leadership development, and work on ethical and moral perspective-taking capacity. Authentic leadership was defined as "as a process that draws from both positive psychological capacities and a highly developed organizational context, which results in both greater self-awareness and self-regulated positive behaviors on the part of leaders and associates, fostering positive self-development" (p. 243). Moreover, they posited that such leadership worked when "[an individual is] true to him/herself and the exhibited behavior positively transforms or develops associates into leaders themselves" (p. 243). Thus, the mechanism of influence involved role modeling and arose from the leader's alignment of his or her behavior with core personal values and beliefs. As followers observed such behavior, they were often positively impacted by the leader and influenced to likewise align their values and actions. Such positive role modeling provides a stark contrast with other forms of influence, such as coercion, rational persuasion, or inspirational appeals that are rooted in personal leader charisma. Despite these differences, authentic leadership theory was thought to share some overlap with other forms of leadership. For example, Luthans and Avolio (2003) suggested that authentic leaders, like transformational leaders, tend to promote the leadership development of followers enabling them to become leaders themselves.

After the initial conceptions of authentic leadership were introduced, several refinements to the definition and theory emerged. A particularly important event in the history of research on authentic leadership occurred in 2004 with the inaugural Gallup Leadership Summit, which resulted in two key sets of publications focusing on authentic leadership development: (1) a special issue of *The Leadership Quarterly* (Volume 16, Issue 3) the premier journal for leadership research; and (2) an edited volume of the *Monographs for Management and Leadership* (Gardner et al., 2005b). Following these publications, a coalescence of theoretical considerations occurred. It is now possible to identify seminal work by identifying the theoretical foundations underlying research into authentic leadership (Gardner et al., 2011).

Authentic Leadership Theory

In recent years, theoretical attention devoted to authentic leadership has coalesced around a perspective introduced by Bruce Avolio, William Gardner, Fred Walumbwa and colleagues (Gardner et al., 2005a; Walumbwa et al., 2008). In this section, the concepts and literature that are foundational to this perspective are examined to facilitate understanding of the theory's underlying assumptions. It is important to initially explicate the meanings given to terms used in authentic leadership research, which often differ from everyday colloquial usage. The components that lie at the core of authentic leadership from this perspective are then discussed.

In contrast to earlier work, authentic leadership is now thought to lie at the nexus of leadership, ethics, and positive organizational behavior (Walumbwa et al., 2008). The important aspects of these fields, as they relate to authentic leadership research, are reviewed below.

One fundamental assumption of authentic leadership theory is that authentic leadership constitutes the 'root construct' that underlies positive forms of leadership (Avolio and Gardner, 2005). This creates a special relationship between authentic leadership and other positive perspectives of leadership. Whereas other positive leadership theories (e.g., transformational leadership theory, ethical leadership theory, spiritual leadership theory, servant leadership theory) attempt to explicate a range of positive behaviors under a single theory, authentic leadership theory suggests that authenticity provides a foundation for and thereby enhances these forms of leadership. In other words, considering authentic leadership as a root construct means that being an authentic leader does not preclude one from engaging in another type of leadership. To the contrary, authenticity is seen as enhancing and complimenting transformational, ethical, spiritual, or servant leadership because it ensures such leadership is genuine. The resultant alignment between the leader's values, principles, abilities, personality, motives, and behavior helps the leader capitalize on his or her strengths and achieve internal consistency, and thereby fosters follower trust, leader well-being, and leader effectiveness (Gardner et al., 2005a). However, this only holds true for positive forms of leadership that are aligned with authentic leadership, rather than negative forms of leadership. That is, if the form of leadership is dysfunctional (e.g., abusive supervision, narcissistic leadership, personalized charismatic leadership), it is not commensurable with authentic leadership.

Authenticity

The construct of authenticity lies at the center of authentic leadership theory (May et al., 2003). Authenticity has been defined in a variety of ways. Some suggest that authenticity involves owning one's personal experiences, be they thoughts, emotions, needs, preferences, or beliefs (Harter, 2002). Authenticity can also be described as the degree to which one fulfills personal expectations and commitments promised to oneself. When one is authentic, one's true or core self is unobstructed in one's daily enterprise (Kernis, 2003). The notion of authenticity is also captured by the injunction to 'know thyself'. Within the authentic leadership literature there is an implicit assumption that to know one's self includes *acting* in accordance with that known self.

Literature on authenticity assumes that people make commitments to their self-identity (Kernis, 2003). The emotions experienced when these commitments are violated

contribute to feelings of inauthenticity, or the negative feelings associated with 'being fake'. In contrast, there are also positive emotional and cognitive consequences that accrue when one acts in accordance with one's true self. People are generally not considered to be completely authentic or inauthentic; rather, authenticity operates on a continuum, and it is therefore more accurate to refer to people as being more or less authentic (Avolio and Gardner, 2005). Relatively authentic individuals pay more attention and are more devoted to achieving congruence between their internal self-experience and how they interact with the environment. For example, authentic individuals are generally highly aware of their values, and are more likely to act in accordance with them.

Highly authentic individuals have been shown to experience higher levels of psychological well-being, enhanced feelings of engagement, and elevated performance (Harter, 2002). What happens when such a person takes on a leadership role is the basic focus of authentic leadership theory. Authenticity in leadership can be seen as a paradox of sorts. An authentic individual in a leadership capacity is nonetheless self-referential with regard to being true to him or herself. However, leaders, by definition, must have follower(s), and it is a key part of leadership to be aware of follower needs and circumstances. Part of the goal of authentic leadership theory is to explain why this paradox works, answering the question "how does being true to one's self elicit positive relational consequences and outcomes?"

Core Components of Authentic Leadership

Though authentic leadership is often oversimplified as involving an approach to leadership where one can 'be true to one's self', current theory suggests that authentic leadership goes beyond this foundational notion (Walumbwa et al., 2008). Indeed, authentic leadership involves more than being an authentic person since leadership is required. As the definition presented above suggests, authentic leadership involves a pattern of leader behavior. The four components of this pattern suggested by Walumbwa et al. (2008) are reviewed below.

Self-awareness

Leader self-awareness refers to a leader's understanding of how he or she derives meaning from the world, and how this meaning-making process impacts his or her view of the self and the world over time. This understanding includes the recognition of personal strengths and weaknesses, in addition to the many different facets of the self (some perhaps more likable, accepted, or desirable than others). Through interactions with others, leaders with high self-awareness gain valuable personal insights including knowledge about how they impact others in general, and especially those that they lead.

Internalized Moral Perspective

An internalized moral perspective refers to a leader holding and adhering to core ethical values, standards, and principles as an act of personal choice. That is, authentic leaders are driven by internal moral standards regardless of whether they match or deviate from prevailing norms of a group, organization, or social context. Indeed, an internalized moral perspective guides the leader toward decisions and behaviors that are congruent with their values, even in the face of environmental pressures, including rewards and sanctions, to do otherwise. In contrast, the moral conduct of extrinsically motivated leaders is easily swayed by external forces. The internalized moral perspective component of authentic leadership is really a form of self-regulation (Deci and Ryan, 2000). Because authenticity involves the unobstructed operation of one's true or core self, following one's inner sense of right and wrong is a fundamental element of authentic leadership. Hence, authentic leaders are not enticed by opportunities to get ahead that require violations of their core principles and commitments.

Other theories of leadership also specify an ethical component (e.g., transformational leadership; Bass and Avolio, 1994) or provide almost an exclusive ethical focus (e.g., ethical leadership; Brown and Trevino, 2006). Some view authentic leadership as diverging from these perspectives because they believe authentic leaders need not necessarily be ethical. However, Avolio and Gardner (2005) reject this notion. To the contrary, they assert that highly authentic leaders operate at a high level of moral cognitive development. That is, authentic leaders, like transformational leaders, pursue end values that prohibit knowingly engaging in unethical conduct.

Balanced Processing of Information

Leaders often operate at the nexus of various sources of information; as such, one of their key responsibilities is to process this information to make decisions. Balanced processing refers to the evaluation of ego-relevant information – be it favorable or unfavorable – in a relatively unbiased fashion. In addition, authentic leaders seek out objective information even if it challenges their preexisting notions. This is borne out of a sense of security that accrues for authentic persons. Indeed, when one understands and owns the self, including both strengths and weaknesses, one is more receptive to negative feedback that may be ego-threatening and hence subject to distortion for less authentic and secure persons. Balanced processing helps authentic leaders make the right decision for their group, organization, or society, even if doing so may not be personally beneficial. Thus, authentic leaders are open to feedback, in contrast to leaders who only want to hear good news.

Relational Transparency

Relational transparency involves the presentation of one's true self to close others (Gardner et al., 2005a). Organizations and other settings often provide opportunities and even incentives to manage impressions and manipulate relationships. Authentic leaders do not obfuscate their purposes or desires, and do not misrepresent themselves to achieve desired goals. Specifically, to the extent possible, authentic leaders disclose work-related information in a transparent and open fashion, as long as doing so does not violate the privacy or safety of others. In cases where full disclosure would be inappropriate, such leaders release as much information as possible, while communicating the reasons for not withholding certain facts. Relational transparency also refers to sharing one's true thoughts and feelings openly, while maintaining dignity appropriate to the workplace and context. Thus, followers understand the reasoning behind a leader's decision, which decreases fear of the unknown and fosters feelings of trust.

Theoretical and Empirical Antecedents and Outcomes

Consistent with their theoretical perspective, Walumbwa et al. (2008) found empirical support for a multidimensional, higher-order construct of authentic leadership comprised of the four basic components previously discussed. The emergence of a validated measure of authentic leadership, the Authentic Leadership Questionnaire (ALQ; Walumbwa et al., 2008), has stimulated research on the multicomponent model. More recently, an alternative and psychometrically sound measure based on the same four-component model, the Authentic Leadership Inventory, has been advanced (Neider and Schriesheim, 2011).

Authentic leadership has been connected to a variety of antecedents and outcomes, both theoretically and empirically (Gardner et al., 2011). While sensational examples of malfeasance provided the original impetus for understanding authentic leadership, it is now evident that the costs of inauthentic leadership can be great for any organization or team. Conversely, the benefits of authentic leadership have been found to be wide ranging and compelling. In this section, the major empirical findings relative to the theoretical antecedents and outcomes of authentic leadership are discussed.

Antecedents

As noted in Section Early Conceptions, early models defined authentic leadership "as a process that draws from both positive psychological capacities and a highly developed organizational context" (Luthans and Avolio, 2003: p. 243). More recently, however, models of authentic leadership see positive psychological capacities and positive ethical climate as peripheral antecedents and consequences, rather than core components of the construct. To date, research on the antecedents of authentic leadership has been very limited. While some support for PsyCap leading to authentic leadership has been found (Jensen and Luthans, 2006), this research was conducted before Walumbwa et al. (2008) operationalized the multicomponent model of authentic leadership through the introduction of the ALQ.

Outcomes

The majority of empirical work on authentic leadership has sought to explicate its outcomes, rather than its antecedents (Gardner et al., 2011). Many theoretical propositions regarding these outcomes have been supported. Indeed, this may be part of the reason why authentic leadership has sustained the attention of the scientific community. In this section, we briefly review some of the predicted outcomes and relevant empirical findings. Some outcomes are posited explicitly in models of authentic leadership, whereas others have been derived from those models.

Overall, the Gardner et al. (2011) review suggests that authentic leadership is associated with positive leader (psychological well-being), follower (empowerment, job performance, job satisfaction, organizational citizenship behaviors, organizational commitment, trust in leadership, work engagement, and work happiness), and organizational outcomes (openness of organizational climate and firm financial performance). Authentic leadership was also shown to be negatively associated with undesirable outcomes (e.g., follower burnout and contingent leader self-esteem). Together, these results provide evidence of predictive validity of the authentic leadership construct. Subsequent research provides additional empirical evidence of predictive validity, as authentic leadership has been shown to be positively related to behavioral integrity (Leroy et al., 2012b), perceived leader effectiveness (Moriano et al., 2011), follower extra effort (Moriano et al., 2011), satisfaction with the leader (Moriano et al., 2011; Neider and Schriesheim, 2011), follower basic need satisfaction (Leroy et al., 2012a), follower general satisfaction (Neider and Schriesheim, 2011), follower organizational commitment (Neider and Schriesheim, 2011), follower affective commitment (Leroy et al., 2012b), follower social exchange (Chiaburu et al., 2011), follower creativity (Chiaburu et al., 2011), follower work role performance (Leroy et al., 2012a,b), positive (Woolley et al., 2011) and collective psychological capital (Woolley et al., 2011), organizational citizenship behavior (Walumbwa et al., 2011), positive work environments (Walumbwa et al., 2011), group level performance (Woolley et al., 2011), firm level performance (mediated through top management team positive affective tone; Hmieleski et al., 2012), team trust (Walumbwa et al., 2011), team authenticity (Hannah et al., 2011), and teamwork behavior (Hannah et al., 2011).

Gardner et al. (2011) note that while the research to date is informative, much work remains to advance our understanding of authentic leadership. To do so, they recommend a five-pronged agenda for future research: (1) stronger theory building, (2) expansion of the nomological network, (3) more rigorous and diverse methods, (4) attention to authentic followership, and (5) attention to authentic leadership development. In a separate review of the leadership literature, Avolio et al. (2009) came to a similar conclusion that the measurement of constructs relating to authentic leadership is at an early stage of development. They identify several relevant constructs that have not yet been examined, such as moral perspective, self-concept clarity, psychological well-being, spirituality, and judgment. They also suggest that additional research is needed to examine authentic leadership across different cultural and situational contexts.

One way that the theory building can be enhanced is by more clearly articulating the boundary conditions. Though the intent of research behind authentic leadership is to understand a wide range of phenomena, it is likely that some contexts are not fully supportive of authenticity and authentic leadership (e.g., toxic organizational cultures). While some work on boundary conditions of authentic leadership has been done (Eagly, 2005), more theory and empirical research is needed to explicate these boundaries.

Another area that requires greater research is the interplay between authentic leadership and authentic followership. In their initial conceptualization of authentic leadership, Gardner and associates (Gardner et al., 2005a) posited a parallel process to authentic leadership whereby followers come to function authentically, in part by modeling the authenticity of the leader. In their review of the literature, Gardner et al. (2011) pointed out that the role of authentic followership within authentic leader–follower relationships had not been examined, despite it being a crucial component of authentic

leadership theory. In an initial step toward filling this void, Leroy et al. (2012a) demonstrated that work units that were characterized by relatively high levels of authentic leadership provided a supportive context for followers' satisfaction of their basic needs for autonomy, competence, and relatedness, and thereby fostered authentic followership and elevated levels of follower work role performance. Moving forward, additional research is required to explore the role that authentic leadership plays in facilitating follower authenticity at work, as well as the extent to which authentic followership in turn enables leaders to achieve authenticity.

Authentic Leadership Development

As authentic leadership theory continues to evolve, some have recommended that the field pay greater attention to authentic leadership development (Avolio et al., 2010). As discussed earlier, authentic leadership development was the primary concern of some of the earliest models of authentic leadership. This focus may have arisen from the initial impetus for learning about authentic leadership (e.g., corporate malfeasance) and a desire to change what was perceived as a societal trend toward the development of inauthentic leaders.

Some scholars have cautioned that we face many challenges in developing authentic leaders (Cooper et al., 2005), while others have advocated in favor of such development (Avolio and Gardner, 2005). A recent review of authentic leadership development (Avolio, 2010) traces the personal experiences of Bruce Avolio, who has been on the front lines of authentic leadership development research. Investigations into authentic leadership development ask whether leaders are born or made, and whether leaders can be changed by experienced planned interventions. In addition, models of authentic leadership development include trigger events, self-reflection, and developmental readiness, as well as core components of authentic leadership theory. Avolio et al. (2010) note that the future of authentic leadership development will likely incorporate more consideration of contextual influences, as much of the focus on training and development of authentic leaders has taken place outside of the context where they work. One way to better explicate the work context would involve embedding leadership training into organizations internally (rather than addressing the need by using an external source, such as a leadership consultant).

Conclusion

Despite being a relatively new construct, authentic leadership is a popular approach to understanding leadership for both academics and practitioners. It emerged as interest in genuine, veritable leadership increased. The foundation for authentic leadership stems from the attainment of individual authenticity on the part of the leader. For authentic *leadership* to be achieved, however, a pattern of leader influence that consists of high authenticity relative to followers is required. In other words, authentic leadership occurs when leaders achieve high self-awareness, balanced processing, relational transparency, and an internalized moral perspective in their relationships with followers. Authentic leadership has been theoretically and empirically related to many desirable outcomes. While many additional questions remain to fully develop and assess the theory, it is evident that authentic leadership has captured the attention of researchers and practitioners due to its potential utility in fostering positive and sustainable individual, group, and organizational outcomes.

See also: Business Ethics; Identity and Identification, Social Psychology of; Identity and Identification; Leader–Member Exchange Theory; Leadership in Organizations, Sociology of; Managerial and Organizational Cognition; Person and Self: Philosophical Aspects; Personal Identity: Philosophical Aspects; Personality and Values at Work; Power, Politics, and Influence in Organizations; Self-Determination Theory; Self-Regulated Learning; Self: Philosophical Aspects; Social Work Ethics; Subjective Wellbeing, Psychology of; Teamwork and Team Performance Measurement; Values, Psychology of; Values, Social Psychology of; Work Motivation.

Bibliography

Avolio, B.J., 2010. Pursuing authentic leadership development. In: Nohria, N., Khurana, R. (Eds.), The Handbook of Leadership Theory and Practice: A Harvard Business School Centennial Colloquium on Advancing Leadership. Harvard Business School Press, Boston, MA, pp. 739–768.

Avolio, B.J., Gardner, W.L., 2005. Authentic leadership development: getting to the root of positive forms of leadership. The Leadership Quarterly 16 (3), 315–338.

Avolio, B.J., Griffith, J., Wernsing, T.S., Walumbwa, F.O., 2010. What is authentic leadership development? via psyh (EBSCOhost) In: Linley, P.A., Harrington, S., Garcea, N. (Eds.), Oxford handbook of Positive Psychology and Work. Oxford University Press, New York, pp. 39–51. http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2009-21869-004&site=ehost-live.

Avolio, B.J., Walumbwa, F.O., Weber, T.J., 2009. Leadership: current theories, research, and future directions. Annual Review of Psychology 60 (1), 421–449.

Bass, B.M., Avolio, B.J., 1994. Improving Organizational Effectiveness through Transformational Leadership. Sage, Thousand Oaks, CA.

Brown, M.E., Trevino, L.K., 2006. Ethical leadership: a review and future directions. The Leadership Quarterly 17 (6), 595–616.

Chiaburu, D.S., Diaz, I., Pitts, V.E., 2011. Social and economic exchanges with the organization: do leader behaviors matter? Leadership & Organizational Development Journal 32 (5), 442–461.

Cooper, C.D., Scandura, T.A., Schriesheim, C.A., 2005. Looking forward but learning from our past: potential challenges to developing authentic leadership theory and authentic leaders. The Leadership Quarterly 16 (3), 475–493.

Deci, E.L., Ryan, R.M., 2000. "What" and "why" of goal pursuits: human needs and the self-determination of behavior. Psychological Inquiry 11, 227–268.

Eagly, A.H., 2005. Achieving relational authenticity in leadership: does gender matter? The Leadership Quarterly 16 (3), 459–474.

Gardner, W.L., Avolio, B.J., Luthans, F., May, D.R., Walumbwa, F.O., 2005a. "Can you see the real me?" A self-based model of authentic leader and follower development. The Leadership Quarterly 16 (3), 343–372.

Gardner, W.L., Avolio, B.J., Walumbwa, F.O., 2005b. Authentic leadership development: emergent trends and future directions. In: Gardner, W.L., Avolio, B.J., Walumbwa, F.O. (Eds.), Authentic Leadership Theory and Practice: Origins, Effects and Development. Elsevier Science, Oxford, UK.

Gardner, W.L., Cogliser, C.C., Davis, K.M., Dickens, M., 2011. Authentic leadership: a review of the literature and research agenda. The Leadership Quarterly 22 (6), 1120–1145.

George, W., 2003. Authentic Leadership: Rediscovering the Secrets to Creating Lasting Value. Jossey-Bass, San Francisco, CA.

George, W., Sims, P., 2007. True North: Discover Your Authentic Leadership. Jossey-Bass, San Francisco, CA.

Hannah, S.T., Walumbwa, F.O., Fry, L.W., 2011. 'Leadership in action teams: team leader and members' authenticity, authenticity strength, and team outcomes. Personnel Psychology 64 (3), 771–802.

Harter, S., 2002. Authenticity. In: Snyder, C.S., Lopez, S.J. (Eds.), Handbook of Positive Psychology. Oxford University Press, Oxford, pp. 382–394.

Hmieleski, K.M., Cole, M.S., Baron, R.A., 2012. Shared authentic leadership and new venture performance. Journal of Management 38, 1476–1499.

Jensen, S.M., Luthans, F., 2006. Relationship between entrepreneurs' psychological capital and their authentic leadership. Journal of Managerial Issues 18 (2), 254–273.

Kernis, M.H., 2003. Toward a conceptualization of optimal self-esteem. Psychological Inquiry 14, 1–26.

Leroy, H., Anseel, F., Gardner, W., Sels, L., 2012a. Authentic leadership, authentic followership, basic need satisfaction, and work role performance: a cross-level study. Journal of Management. http://jom.sagepub.com/content/early/2012/08/24/0149206312457822 (accessed 27.08.12.).

Leroy, H., Palanski, M.E., Simons, T., 2012b. Authentic leadership and behavioral integrity as drivers of follower commitment and performance. Journal of Business Ethics 107 (3), 255–264.

Luthans, F., Avolio, B.J., 2003. Authentic leadership: a positive developmental approach. In: Cameron, K.S., Dutton, J.E., Quinn, R.E. (Eds.), Positive Organizational Scholarship. Barrett-Koehler, San Francisco, pp. 241–261.

Luthans, F., Avolio, B.J., Avey, J.B., Norman, S.M., 2007. Positive psychological capital: measurement and relationship with performance and satisfaction. Personnel Psychology 60, 541–572.

May, D.R., Chan, A.Y.L., Hodges, T.D., Avolio, B.J., 2003. Developing the moral component of authentic leadership. Organizational Dynamics 32, 247–260.

Moriano, J.A., Molero, F., Levy Manguin, J.-P., 2011. Authentic leadership: concept and validation of the authentic leadership questionnaire in Spain. Psicothema 23 (2), 336–341.

Neider, L.L., Schriesheim, C.A., 2011. The authentic leadership inventory (ALI): development and empirical tests. The Leadership Quarterly 22 (6), 1146–1164.

Seligman, M.E.P., Csikszentmihalyi, M., 2000. Positive psychology. American Psychologist 55, 5–14.

Walumbwa, F.O., Avolio, B.J., Gardner, W.L., Wernsing, T.S., Peterson, S.J., 2008. Authentic leadership: development and validation of a theory-based measure. Journal of Management 34 (1), 89–126.

Walumbwa, F.O., Luthans, F., Avey, J.B., Oke, A., 2011. Authentically leading groups: the mediating role of collective psychological capital and trust. Journal of Organizational Behavior 32, 4–24.

Woolley, L., Caza, A., Levy, L., 2011. Authentic leadership and follower development: psychological capital, positive work climate, and gender. Journal of Leadership & Organizational Studies 18 (4), 438–448.

Authenticity, Anthropology of

Richard Handler, University of Virginia, Charlottesville, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The anthropology of authenticity concerns the relationship of cultural objects and processes to posited social or cultural identities. From an objectivist perspective, it should be possible to validate such assignments of identity; from a constructionist perspective, identity-making is an ongoing social and political process that is often contested, yet also central to the cultural aspirations of many of the people anthropologists study.

Within anthropology, there are two principal approaches to the question of authenticity. The first presupposes that authenticity is an attribute or property inherent in cultural objects, and that the role of anthropology is to determine, using scientifically appropriate evidence and methods, whether particular objects are or are not authentic. The second locates authenticity not in objects but in discourses about them. This approach holds that the authenticity of phenomena is never a matter of objectively determinable fact, but is always socially constructed and, as such, subject to change.

Presently, the second approach is in the ascendance. It would, however, be short-sighted to view this trend as the culmination of a unilineal historical development in which the second approach has eclipsed the first. Rather, these opposed approaches endure because they replicate a fundamental epistemological opposition within modern societies, that between objectivism, realism, or positivism, on the one hand, and constructionism or hermeneutics, on the other. Despite the recent persuasiveness of constructionism in social science, objectivist notions of authenticity remain hegemonic in many late-capitalist institutions, such as the art market, museums, and courts of law.

Authenticity and Reality

Questions about authenticity should be distinguished from questions about reality and its representation. To ask whether an object is real, or an event occurred, is to ask about its existence relative to the empirical world. To declare an object or event 'unreal' or 'illusory' is to claim that it never existed or occurred. Note that such 'unreal' phenomena can exist in imagination. At least since Durkheim, anthropologists have argued that representations of objects are real-world phenomena with real-world effects, whether or not the objects of such representations can be shown (according to some sets of criteria) to exist in the world beyond those representations.

In contrast, to ask whether an object is authentic is to ask not about its existence but about its relationship to a posited social or cultural identity. The opposite of 'authentic' is not 'unreal' but 'inauthentic.' To deem an object inauthentic is to assert that it is not, despite claims to the contrary, an example of an identified class of objects, or not the creation of an identified person or group.

It is easy to confuse questions of reality and those of authenticity due to the commonsense use of the word 'real' to mean 'authentic,' as in the phrase 'this is a real Native American medicine bundle.' Such claims are not intended as predications of existence as such, but as evaluations of what is understood to be the essential nature, or the identity, of the phenomenon in question. Thus, assertions of authenticity always have embedded within them assertions of identity.

Throughout the history of their discipline, anthropologists have been asked to authenticate artifacts, customs, legends, texts, and other cultural objects, as well as the racial, ethnic, or cultural affiliation of persons. Depending on subdisciplinary and areal expertise, anthropologists might offer authoritative answers to such questions as: 'Are these stone weapons Clovis points, from Western North America and 12 000 years old, or are they modern reproductions?' 'Can this narrative be attributed to precolonial Native Americans, is it the result of European influence on Native American traditions, or is it a recent creation, and if the latter, what is the cultural identity of its author?' And most bluntly: 'Are these people Indians with unbroken ties to precolonial America, or have generations of racial mixing and cultural borrowing erased that identity?' Such assessments of the authenticity of particular objects and persons are similar to those required in fields like art history and museology, where it is often important to be able to identify objects by attributing them to named artists, historical periods, or cultural groups.

Authenticity and Modern Individualism

The commonsense notion that authenticity is an expression of the essential identity of an object stems from modern (in the sense of post-fifteenth century) individualism. In this worldview, ultimate reality is understood to be lodged, not in a hierarchically interrelated divine cosmos, but in individuals, from the smallest particles to persons to delineated groups of persons (cultures). To quote the philosopher Cassirer on the modern ontology of nature: "Nature…implies the individuality, the independence and particularity of objects. And from this characteristic force, which radiates from every object as a special center of activity, is derived also the inalienable worth which belongs to it in the totality of being… The part not only exists within the whole but asserts itself against it, constituting a specific element of individuality and necessity" (1965, p. 41).

From such an outlook the notion of authenticity derives: each object, each individuality is itself a center of reality with a unique essence or identity that distinguishes it from all other objects.

Authenticity and Sincerity

In *Sincerity and Authenticity*, Trilling (1972) drew out the implications of this worldview for modern conceptions of sociality and personhood. With the emergence of modern individualism, persons were no longer defined by their position in a social hierarchy (as in medieval European realms). Instead, social positions came to be understood as 'roles' to be filled by persons whose essence, or individuality, precedes and transcends those roles. In this perspective, the inner realm of individual character or personality is more real than the outer world of social relations.

The importance of sincerity in the early modern period, Trilling argued, derived from concerns about whether a person's social actions matched or masked inner attitudes. Indeed, it was insincerity that was at issue. As Rousseau puts it, humans in the 'state of nature,' prelingual and presocial 'savages,' had no means to dissemble; their actions could be nothing other than the direct expression of their inner or essential being. But with the emergence of human language and society, "it…became the interest of men to appear what they really were not" (1973, p. 86). Inner authenticity diverged from outer sociality, and the 'rise' of civilization, as Rousseau saw it, entailed a process of moral corruption.

In Trilling's account, concern for sincerity, with its questioning of the relationship between society and the individual, began to cede place in the Romantic period to concern for authenticity, which privileges inner essences without relating them to an outer world of other people. The Romantics' quest for authenticity was a quest for 'essential identity,' a longing for knowledge of true inner being irrespective of any accidents of social history or circumstance. Artists, in particular, were understood to be individuals possessed of powerfully unique inner realities, uncorrupted and undiminished by social conventions. At the same time, the growing middle classes of the new mass democracies were coming to see their lives or identities as increasingly 'unreal' or 'inauthentic.'

Authenticity and Collectivities

In social philosophy, and later the social sciences, these concerns for individual realities were applied to human collectivities. The leap from the individual to the collective is not as great as it seems, for, as Dumont (1970) has insisted, modern ideology understands groups as 'collective individuals.' From the eighteenth century onward, folklorists and anthropologists assumed that groups of people could be characterized by unique cultural, historical, religious, linguistic, or otherwise traditional features that set them off from all other groups. These models of the group were often naturalistic, using analogies to biological organisms or species: it was assumed that a group was a bounded natural unit or object, that members of the group shared its defining features, and that it was possible objectively to evaluate the cultural identities of individual persons, based on the cultural attributes they 'possessed.'

Authenticity Endangered

A noteworthy accompaniment of this conception of cultural authenticity is a fear of its demise. Authenticity is seen to be perpetually threatened. The authenticity of nonmodern ('savage,' 'primitive,' 'native,' 'indigenous,' 'folk,' 'peasant,' or 'traditional') cultures is imagined to depend on their pristine existence beyond the reach of advancing 'civilization.' Throughout the epoch of European colonial expansion, down to the present day, metropolitan authorities – scholars, travelers, missionaries, colonial officials, collectors – took it upon themselves to delineate unique groups in terms of their 'possession' of specific racial, linguistic, and cultural attributes. The impingement of European culture on such groups, or even the influence of non-European but purportedly more 'complex' cultures, was thought to alter their unique identities and thus to destroy their cultural authenticity. Such conceptions were applied at the level of the individual as well, such that in colonial social orders, the 'half-breed' (to use a racial term) or the culturally bastardized colonial subject was often scorned as the embodiment of the worst of both the 'primitive' and 'civilized' worlds.

In sum, the European 'discovery' of 'Others' followed by ongoing 'contact' made them, almost by definition, inauthentic. Authenticity as a quality of cultures, artifacts, and persons was projected by metropolitan authorities into the past, a past they might be able to document, but only just before its destruction.

Constructionist Accounts of Authenticity

Despite the pervasiveness of such assumptions about groups as bounded objects, there have been recurring social-scientific trends (historical diffusionism, for example) that privileged intergroup connections, contacts, borrowings, and syncretisms. In Anglo-American anthropology since the 1970s, there has been growing rejection of the notion that cultures, societies, and groups in general are naturally bounded and internally homogeneous. Indeed, it is now commonplace to assert the opposite, and to claim that groups are historically and situationally contingent; that the relative rigidity (or porousness) of group boundaries is a function of semiotic, not natural, factors; that collective identities salient in contemporary institutions, politics, and scholarship ought not to be projected into the past; that individual persons have multiple identities that can shift according to social context, although not without regard to institutional and cultural constraints; and that the products of human creativity cannot unproblematically be assigned cultural, racial, or ethnic identities. In addition, recent historical scholarship has argued that the work of colonial cultural classification was carried out in areas of the world where European notions of objective social boundaries encompassing homogeneous, authentic cultural identities made little sense in relation to local people's understanding of sociality.

Negotiating Authenticity

Constructionist approaches to authenticity question whether it is useful to specify the 'essential' identity of groups, artifacts,

and persons. Since human beings are historically and socially situated, they are always connected to other people. Culture exists not simply *in* them but *between* them. Any understanding of a particular person's or group's authentic identity will be a function of specific social encounters, and particular evaluations of authenticity are open to negotiation and dispute. An 'expert' (colonial administrator, art collector, anthropologist) may claim that a person is not a 'real' or 'pure' member of a posited culture, or that an artifact is not an authentic product of it, but others may reject such evaluations based on differing definitions of cultural boundaries and essences.

This does not mean it is impossible or unhelpful to describe and classify cultures. Rather, it means that the terms used to do so must be understood in relation to the discursive project of the classifiers. At any historical moment, some such terms will be relatively unproblematic. For example, given our present technology and understanding of time, we can expect the carbon 14 dating of archaeological artifacts to yield valid information. We need not question the empirical validity of the description to recognize that such data become useful, or meaningful, only in terms of more general hypotheses concerning culture history, hypotheses that can never be reduced merely to facts derived from carbon 14 dating. Moreover, terms of cultural classification that are unproblematic in one context might become contested in another, for example, present-day descendants of precolonial Native Americans might dispute carbon 14 results, claiming a greater antiquity for artifacts than anthropological methods can demonstrate.

The Politics of Authenticity

Assertions of cultural authenticity can have political consequences. Among the most serious are governmental recognition of 'tribal' or 'aboriginal' status, a status that often entitles group members to government subsidies and services, and to government (mis)management. In settler societies such as the United States, Canada, and Australia, some 'indigenous' groups have governmental recognition based on treaties, some of which date to the late eighteenth century; other groups claiming to be authentically indigenous may turn to the courts to gain such recognition. In the courts, authenticity is defined in terms of temporal continuity, cultural homogeneity and distinctiveness, and geographical stability. It may or may not be possible for people who have lived for generations in a situation of 'contact' to demonstrate possession of cultural authenticity defined in such terms. Indeed, there is a cruel irony inherent in the relationship of 'native' authenticity to nation-state citizenship. On the one hand, 'indigenous' peoples are asked to assimilate, through education and economic mobility, into modern society; on the other hand, to do so is to be seen as having become inauthentic and to be scorned on that basis. In practice, both racist barriers and modern fantasies of authenticity fence people in, making it difficult for them to enjoy full citizenship.

Claims concerning the authenticity of individuals and objects can also have political (and economic) consequences. Government criteria to evaluate the affiliation of individuals to recognized tribal groups are often based on quantifications of 'blood'; that is, proportion of 'native' descent. Such criteria usually do not match those that defined social inclusion before the imposition of modern definitions of authentic personhood. Individuals' access to government benefits nonetheless depends upon satisfying the imposed criteria.

The authenticity of cultural objects is negotiated in the markets for art, museum artifacts, and touristic souvenirs. Objects deemed culturally authentic command higher prices and can claim space in more prestigious venues, such as museums and the homes of wealthy collectors, than those considered culturally 'mixed' or 'commercial.' Over time, however, objects can gain authenticity as scholars and collectors objectify and historicize what were once seen as debased cultural forms. 'Indigenous' artists and craftworkers, too, can influence standards of authenticity as they position themselves vis-à-vis the art market and museum world.

Indigenous Control of Authenticity

At the beginning of the twenty-first century, political trends and technological developments have combined to give indigenous peoples around the world renewed control over what they considered to be their own cultural property. The global expansion of intellectual-property laws and United Nations' declarations supporting indigenous cultural-property rights have led many national governments to revise their property regimes, in dialogue with indigenous peoples living within their territories. At the same time, anthropologists, museum directors and curators, art and antiquities dealers, and private collectors have been brought, willingly or otherwise, into dialogue with newly empowered indigenous groups seeking either the return of cultural property (from museums, for example, to tribal institutions) or increased access to it. In practice, this has meant that indigenous peoples, anthropology's traditional object of study, have gained new leverage to participate in debates about authenticity. As objects begin to circulate in new ways – for example, nineteenth-century artifacts might be returned to the descendants of the people who created them, to be used in contemporary religious or cultural activities – their cultural authenticity is renewed and redefined.

At the same time, digital technologies are creating new ways to access stationary cultural properties. Museums, for example, now digitize their collections, allowing visitors to view aspects or images of objects not available in standard displays. Local or indigenous institutions can choose to have items of cultural property exhibited and/or protected in metropolitan institutions while making use of digitized images and digital databases for use in their communities.

Whether new legal regimes and new technologies will lead to changes in concepts of authenticity remains to be seen, but it is already clear that they have opened the debate to new participants. Anthropologists will continue to play an important role as scholars and even arbiters of authenticity, but going forward, we may as often be employed by indigenous communities seeking to define authenticity in relationship to their own ideas and goals as by metropolitan universities and museums seeking to protect theirs.

See also: Advocacy in Anthropology; Center–Periphery Relationships; Craft Production, Anthropology of; Cultural Heritage; Ethnicity: Anthropological Aspects; Human Rights and Social

Work; Identity in Anthropology; Indigeneity; Knowledge, Anthropology of; Materiality and Culture; Modernity; Museums; National Character; Property: Anthropological Aspects; Tourism.

Bibliography

Adorno, T.W., 1964. Jargon der Eigentlichkeit: Zur deutschen Ideologie. Suhrkamp Verlag, Frankfurt am Main, Germany (1973. The Jargon of Authenticity. Northwestern University Press, Evanston, IL.).

Bendix, R., 1997. In Search of Authenticity: The Formation of Folklore Studies. University of Wisconsin Press, Madison, WI.

Bruner, E.M., 2005. Culture on Tour: Ethnographies of Travel. University of Chicago Press, Chicago.

Cassirer, E., 1965. The Philosophy of the Enlightenment. Beacon Press, Boston.

Clifford, J., 1988. The Predicament of Culture: Twentieth-Century Ethnography, Literature, and Art. Harvard University Press, Cambridge, MA.

Dumont, L., 1970. Religion, politics, and society in the individualistic universe. Proceedings of the Royal Anthropological Institute, 31–45.

Errington, S., 1998. The Death of Authentic Primitive Art and Other Tales of Progress. University of California Press, Berkeley, CA.

Gable, E., Handler, R., 1996. After authenticity at an American heritage site. American Anthropologist 98, 568–578.

Geismar, H., 2012. Treasured Possessions: Culture, Property and Indigenous Rights in the Pacific. Duke University Press, Durham, NC.

Orvell, M., 1989. The Real Thing: Imitation and Authenticity in American Culture, 1880–1940. University of North Carolina Press, Chapel Hill, NC.

Rousseau, J.-J., 1973. A discourse on the origin of inequality. In: Cole, G.D.H. (Ed.), The Social Contract and Discourses. Dent, London.

Segal, D.A., 1999. Can you tell a Jew when you see one? Or thoughts on meeting Barbra/Barbie at the museum—teaching Jewish studies. Judaism 48, 234–241.

Trilling, L., 1972. Sincerity and Authenticity. Harvard University Press, Cambridge, MA.

Authoritarian Personality

John Duckitt, The University of Auckland, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The theory of an authoritarian personality was an influential though controversial mid-twentieth-century theory to explain the mass appeal of fascism and ethnocentrism. Methodological and conceptual criticisms of the original theory, however, lead to alternative theories and culminated in research suggesting two distinct dimensions of ideological attitudes, Right-Wing Authoritarianism (RWA) or Social Conservatism and Social Dominance Orientation (SDO) or Anti-Egalitarianism. RWA and SDO were initially thought to be direct expressions of two different authoritarian personalities, but have more recently been seen as describing social or ideological attitude dimensions with multiple social and personal determinants.

The idea of an authoritarian personality emerged in the 1930s as an attempt to explain the popular appeal of fascist movements, such as the Nazis, in many European societies. At that time, European philosophy and social science was strongly influenced by Marxism, and a major prediction from Marxism was that the crisis engendered by the Great Depression should radicalize the working classes and lead to socialist revolutions. This expectation, however, was confounded in much of Europe where the crisis did not produce socialist revolutions but resulted in mass support for fascist movements advocating authoritarian leadership and virulent anti-Semitism. The explanation proposed was that a particular kind of authoritarian personality, which had become socially widespread, would under threatening conditions make people emotionally and cognitively vulnerable to the appeal of fascist ideologies and to ethnocentric hostility to vulnerable minorities.

Reich (1975), for example, argued that capitalist social structures were characterized by punitive families, which used child-rearing practices involving extensive sexual repression to create authoritarian personalities, who would be unlikely to rebel against exploitative social conditions and yearn for powerful leadership. This authoritarian character structure was described as conservative, afraid of freedom, submissive to authority, obedient, yet with 'natural aggression distorted into brutal sadism' (p. 66). Maslow (1943) and Fromm (1941) produced similar descriptions of an authoritarian personality to explain the appeal of fascist movements and ideologies.

These early theories were speculative and did not develop empirically based measures of their constructs. As a result, they produced little research. However, their core idea that social and ideological beliefs are a direct expression of basic needs in the personality did later prove highly influential and dominated much of social scientific thought and theory about the psychological bases of ideology and prejudice for much of the rest of the twentieth century.

Original Theory of the Authoritarian Personality and the F Scale

It was in 1950 with the publication of a classic volume titled *The Authoritarian Personality* that these ideas finally became prominent in the social sciences. In this volume the authors (Adorno et al., 1950) theoretically elaborated a theory of the authoritarian personality and reported the results of a multi-pronged decade-long program of research testing the theory. Their investigation had begun with the objective of explaining the psychological bases of anti-Semitism. This was shown to be part of a much broader ethnocentric pattern involving a generalized dislike of out-groups and minorities, as well as an excessive and uncritical Nationalism. Anti-Semitism and ethnocentrism were also strongly related to political and economic conservatism. These attitudes and beliefs appeared to cluster together to form a coherent pattern, and this patterning seemed best explained as an expression of basic needs within the personality.

Evidence from a number of sources and particularly their own research comparing persons high and low in ethnocentrism on indices and ratings scored blind from interview data, and projective test protocols suggested that a constellation of nine tightly covarying traits characterized this authoritarian personality syndrome. Moreover, these traits seemed to be directly expressed in particular 'implicitly antidemocratic,' or authoritarian, attitudes and beliefs. This meant that it would be possible to identify authoritarian personalities by the degree to which people would agree with these 'implicitly antidemocratic' attitudes and beliefs. On this basis, Adorno et al. (1950) developed their famous F scale consisting of items expressing attitudes which were believed to be direct expressions of each of the nine 'traits' of the authoritarian personality syndrome. These nine 'traits' are listed below with their gist definitions in parentheses followed by an example in quotation marks:

- *Conventionalism* (rigid adherence to conventional middle-class values): "A person who has bad manners, habits, and breeding can hardly expect to get along with decent people."
- *Authoritarian submission* (a submissive, uncritical attitude toward authorities): "Obedience and respect for authority are the most important virtues children should learn."
- *Authoritarian aggression* (tendency to condemn, reject, and punish people who violate conventional values): "Homosexuals are hardly better than criminals, and ought to be severely punished."
- *Anti-intraception* (opposition to the subjective, imaginative, and tender-minded): "Nowadays more and more people are

prying into matters that should remain personal and private."

- *Superstition and stereotypy* (belief in mystical determinants of the individual's fate, disposition to think in rigid categories): "Some day it will probably be shown that astrology can explain a lot of things."
- *Power and toughness* (preoccupation with the dominance–submission, strong–weak, leader–follower dimension; identification with power, strength, toughness): "People can be divided into two distinct classes, the weak and the strong."
- *Destructiveness and cynicism* (generalized hostility, vilification of the human): "Human nature being what it is, there will always be war and conflict."
- *Projectivity* (disposition to believe that wild and dangerous things go on in the world; the projection outward of unconscious emotional impulses): "Most people do not realize how much our lives are controlled by plots hatched in secret places."
- *Sex* (an exaggerated concern with sexual 'goings-on'): "The wild sex life of the old Greeks and Romans was tame compared to some of the 'goings-on' in these regions, even in places where people might least expect it."

Adorno et al.'s (1950) theoretical explanation for the origin of this authoritarian personality drew heavily on psychodynamic theory. It suggested that overstrict, harsh, and punitive parental socialization sets up an enduring conflict within the individual. In this conflict, parental punitiveness engenders resentment and hostility toward parental authority and by extension all authority, but cannot be expressed because of fear of and dependence on the all-powerful parents. The anger and hostility are therefore repressed and replaced by an uncritical idealization of the parents and conventional authority and submission to them. The repressed anger and hostility toward authority does not disappear but is displaced and directed toward substitute targets, notably those seen as being sanctioned by conventional authority, such as vulnerable and culturally deviant out-groups and minorities. These inner impulses and conflicts are then directly expressed in the nine surface trait components of the authoritarian personality and those implicit antidemocratic beliefs sampled by the F scale.

Initially, this theory inspired a great deal of enthusiasm. It seemed to effectively tie together concepts over an extremely broad range – from individual psychodynamics to sociological phenomena of immense significance for human society and history. In the two decades following the publication of their book, the F scale was used in hundreds of studies as a measure of an authoritarian personality dimension. These studies confirmed that scores on the F scale were strongly correlated to right-wing attitudes, political conservatism, nationalism, and generalized prejudice against out-groups and minorities (see, e.g., the review by Brown, 1965).

Critical Reaction to the Original Authoritarian Personality Theory

Adorno et al.'s (1950) theory and research, however, also generated a great deal of criticism (see, e.g., Christie and Jahoda, 1954). First, subsequent research could not find clear support for the psychodynamic propositions of their theory. Second, much of their research validating the F scale, which compared high- and low-prejudice participants, seemed seriously compromised because of their failure to use blind ratings and control for sociodemographic and other group differences. Third, their theory and concepts seemed to explain only authoritarianism of the political right and ignored authoritarianism of the left. And, fourth, the F scale became a particular focus of controversy, because all its items were positively worded, so that agreement always indicated high authoritarianism, and disagreement low authoritarianism. One consequence of this was that acquiescence (or the tendency of people to agree rather than disagree) might influence scores on the F scale. A further and more serious consequence in psychometric terms was that acquiescence might have inflated or even caused the apparently high level of covariation between the items of the F scale. If so, the apparent reliability of the F scale might be spurious and it might not be measuring a single personality trait dimension after all.

Most critics did feel, however, that despite its flaws, the two most basic tenets of the theory were valid. These were the assumptions that social attitudes and beliefs were meaningfully organized along a single broad ideological dimension and that these attitudes were direct expressions of personality. A number of subsequent studies therefore focused on trying to remedy the weaknesses of Adorno et al.'s (1950) original research. One important goal of this research was to try and 'balance' the F scale, that is, to produce an F scale that would contain equal numbers of pro- and anti-authoritarian items so that the total scale score would not be influenced by acquiescence. These attempts generally failed with the resulting balanced F scales having very low internal consistencies, that is, the items simply did not correlate sufficiently strongly to indicate that they were measuring a single unitary syndrome and therefore could not be expressive of a particular personality dimension (see, e.g., Altemeyer, 1981). Most researchers on the topic were reluctant to abandon such a core proposition of the theory and preferred to conclude that failure to develop adequately balanced versions of the F scale was because the nature of its items made it difficult to psychologically reverse their meaning. The alternative possibility that the F scale might simply be covering a range of item content that was not unidimensional was not seriously investigated until very much later.

The failure to balance the F scale and the many other criticisms of the original theory did also result in attempts to develop a better theory and measure of the authoritarian personality. These all shared the basic assumption of the original theory, which was that social attitudes in general were ideologically organized along a single dimension that was a direct expression of personality.

Alternative Theories of the Authoritarian Personality

In his book, *The Nature of Prejudice*, Gordon Allport (1954) described the kind of personality that would be generally prejudiced against out-groups and minorities, which closely followed Adorno et al.'s (1950) nine surface traits of their authoritarian personality. Allport, however, largely discarded the idea that these traits were the result of inner psychodynamic

conflict and instead proposed that they stemmed from personal insecurity and fearfulness, or 'ego weakness.' This personal insecurity and fearfulness, he suggested, would cause authoritarian personalities to need structure, order, and control in their social environments and to react with punitive hostility to social change, deviance from convention, and novelty. Allport did not develop a measure of his concept of the authoritarian personality so that his ideas never acquired prominence of those of Adorno et al. (1950), but his simplified conceptualization was broadly adopted by later theorists, such as Wilson (1973) and Altemeyer (1981) who did develop influential new measures.

A second reconceptualization was proposed by Rokeach (1954), who suggested that the core characteristic of the authoritarian personality was a rigid cognitive style, characterized by "a relatively closed cognitive organization of beliefs … organized around a central set of beliefs about absolute authority" (p. 195), which he therefore named Dogmatism. Persons high in dogmatism would be attracted to authoritarian movements and beliefs of either the political right or the left, and cause them to dislike and reject persons and out-groups with dissimilar beliefs and values to their own.

Rokeach (1960) developed a D scale to measure dogmatism, but it shared many of the flaws of the F scale. The items of the D scale did not seem to directly assess a cognitive style, but tended to be broad statements of opinion or social attitude that were often very similar to those of the F scale. Not surprisingly, therefore, research found that the D scale correlated so highly with the F scale that there seemed to be little difference in what they were measuring. In addition, the items of the D scale were also all formulated so that agreement always indicated high dogmatism so that it shared with the F scale the problem of being potentially contaminated by acquiescent responding. As with the F scale, attempts to balance the items of the D scale also resulted in the collapse of internal consistency of the scale. This suggested that the items of the D scale, like the F scale, could not be assessing a single personality or cognitive style dimension.

A third alternative conceptualization of the authoritarian personality was proposed by Wilson (1973), which he named Conservatism. Closely following Allport (1954), Wilson suggested that the basic personality characteristic involved in conservatism was a generalized susceptibility to experience threat or anxiety when confronted by uncertainty (p. 259). As a result, such personalities would adopt authoritarian or conservative social attitudes, which he developed a conservatism or C scale to measure. The items of the C scale consisted of attitudinal statements expressing dislike for and resistance to change, novelty, and diversity, and a preference for order, structure, tradition, and convention.

The C scale did improve on its predecessors by using items for which high Conservatism required both agreement and disagreement and so controlling acquiescent responding. Unfortunately, it also shared its predecessors' weakness, when balanced, of having extremely low internal consistencies with a typical mean correlation between its items being only about 0.05 (Altemeyer, 1981). This indicated that, like the F and D scales, also the C scale could not be measuring a unidimensional construct. Factor analytic studies of the C scale also did not reveal a meaningful factor structure that would enable the extraction of a unidimensional core of items. Finally, the items of the C scale were also similar in content to those of the F and D scales and it also correlated highly with the F and D scales suggesting little or no differentiation between what these scales were measuring.

In general, therefore, none of these alternatives to the original F scale managed to resolve the primary problem that their items, when balanced, did not covary sufficiently to indicate that they were measuring a single unitary construct. This meant that despite the appeal of the theories of an authoritarian (or dogmatic, or conservative) personality, none was able to produce a valid measure of such a personality dimension. As a result, by the end of the 1970s social scientists had largely lost interest in the idea that an authoritarian personality might be the underlying cause of people's ideological convictions. In the 1980s, however, an important new research revived the idea of an authoritarian personality.

Altemeyer's Authoritarian Personality

In 1981 an important book by Bob Altemeyer finally reported the successful development of a fully balanced and psychometrically sound Right-Wing Authoritarianism (RWA) scale to measure the authoritarian personality. The point of departure for Altemeyer's decade-long program of research had been his view that Adorno et al.'s (1950) nine traits and their attitudinal expressions in the F scale were far too broad to delineate a single unitary dimension. His research instead suggested that only three of the original nine traits and their corresponding attitudinal expressions were correlated together strongly enough to describe a unitary dimension. The three traits were 'conventionalism,' 'authoritarian submission,' and 'authoritarian aggression' and the items of the RWA scale were therefore limited only to attitudinal expressions of these three components. Altemeyer (1981) therefore defined RWA, which he saw as direct measure of the authoritarian personality syndrome, as consisting of the covariation of these three highly correlated traits (Altemeyer, 1981: pp. 147–148). A series of validation studies revealed that the RWA scale had excellent psychometric properties. Despite its items being balanced against acquiescence, they showed a high level of internal consistency and seemed to be assessing a single unitary dimension. Moreover, it substantially outperformed a number of earlier measures of the authoritarian personality in predicting presumed effects of RWA, such as holding right-wing political attitudes, generalized prejudice against out-groups and minorities, and support for punitive and unjust actions by established authorities against persons or groups regarded as deviant.

A series of subsequent studies by Altemeyer (1981, 1988, 1996) explored the psychological origins of RWA. Consistent with prior findings using the F scale, there was little support for Adorno et al.'s (1950) original psychoanalytic theory. Thus, correlations between parental punitiveness and strictness in child rearing and the RWA scores of their offspring were weak or not significant. Moreover, signs of repressed aggression in daydreams and fantasies were no more common among high than among low RWA scorers. Altemeyer's (1996) findings also suggested that RWA did not develop in early childhood but was largely formed through social learning and personal

experiences and crystallized during late adolescence. Moreover, despite RWA scale scores being generally stable over time, his research also showed that they could be substantially changed by experiences throughout the life cycle. Thus, RWA decreased substantially with liberal higher education, increased as a result of becoming a parent, and increased in response to threatening social crises. His research also showed that high scorers on the RWA scale seemed to have been socialized to view the world as a dangerous and threatening place, and that this motivated authoritarian behavior, such as aggression and punitiveness, toward persons targeted by the authorities.

A Second Authoritarian Personality?

Another important development in authoritarianism research came in the 1990s with the concept of Social Dominance Orientation (SDO) (Pratto et al., 1994; Sidanius and Pratto, 1999). SDO was conceptualized as a "general attitudinal orientation toward intergroup relations, reflecting whether one generally prefers such relations to be equal, versus hierarchical" (Pratto et al., 1994: p. 742), and measured by a highly reliable and unidimensional SDO scale with items such as 'Inferior groups should stay in their place' (agreement indicating high SDO) and 'No one group should dominate in society' (agreement indicating low SDO).

Although the item content of the SDO scale was therefore quite different from that of the RWA scale, and scores on the two scales were generally not correlated or only very weakly correlated, some striking similarities between the two measures soon became apparent. Both were strong predictors of right-wing political attitudes, as well as nationalism, militarism, and support for tough, undemocratic, authoritarian government. Both also strongly predicted generalized prejudice against outgroups and minorities. Finally, the social attitudes comprising SDO were, at least initially, also seen by Pratto et al. (1994) as direct expressions of a basic personality dimension. Altemeyer (1998) noted that the RWA and SDO scales seem to relate to different sets of the original nine 'trait' clusters listed by Adorno et al. (1950), and therefore concluded that these scales measure two different kinds of authoritarian personality dimensions (the 'submissive' and the 'dominant').

The idea that there are two 'authoritarian' dimensions did help to explain the checkered history of the authoritarian personality and the difficulties of the early theorists. In retrospect, it was apparent that Adorno et al.'s (1950) original conceptualization of the authoritarian personality and their F scale combined these two dimensions and syndromes, resulting in its lack of unidimensionality. Allport (1954), Rokeach (1960), and Wilson (1973) had attempted to simplify the conceptualization of this personality tending to focus on characteristics of the 'submissive' authoritarian, but they failed to narrow their measures correspondingly, which remained multidimensional. The success of Altemeyer's (1981) RWA scale thus seems largely because of him having stripped off those items tapping the factorially different 'authoritarian dominance' syndrome in his item development studies, which had now reemerged in the SDO scale.

Further research by Altemeyer (1998) and others (Duckitt, 2001; Duriez and Van Hiel, 2002; Van Hiel and Mervelde, 2002) confirmed that RWA and SDO, despite their common ideological outcomes, were two independent, individual, different dimensions correlating quite differently with a range of other variables. RWA was powerfully associated with religiosity and valuing order, structure, conformity, and tradition, seeing the social world as a dangerous place influenced by social threat, while SDO was not. SDO, on the other hand was strongly associated with valuing power, achievement, and hedonism, being male, and with a Social Darwinist view of the world as a ruthlessly competitive jungle in which the strong win and the weak lose, while RWA was not.

By the turn of the century, therefore, Altemeyer's (1998) conclusion that two quite different personality dimensions underlay people's ideological beliefs and a wide range of related social attitudes and behaviors had come to be widely shared by researchers. During the next decade, however, the assumption that these were personality dimensions came to be seriously questioned leading to the development of new theoretical models of the psychological bases of ideology.

The New Critique of the Idea of Authoritarian Personalities

A major criticism of the theory of an authoritarian personality, or personalities, which gathered force during the next decade, was that neither the RWA or SDO scale, nor indeed any of their predecessors, such as the F, D, and C scales, actually measured personality (Duckitt, 1989, 2001; Feldman, 2003; Van Hiel and Mervelde, 2002). The items of all these psychometric measures consisted entirely of statements of belief and attitude of a broadly ideological nature and did not describe behavioral dispositions or tendencies as the items of personality inventories typically do. Indeed, after their initial article presenting the SDO scale, Sidanius and Pratto (1999) have generally described it as a measure of enduring beliefs rather than of personality. The idea that the social attitude and belief items comprising the RWA and SDO scales measure personality was therefore merely an assumption that had never been empirically demonstrated. This meant that these scales were better construed as measuring exactly what their item content specified, that is, social and ideological attitudes or value dimensions.

There was also empirical support for the contention that the RWA and SDO scales were measures of attitude or value dimensions rather than personality. For example, Altemeyer's (1996) own research had shown that RWA did not seem to be established in childhood as one might expect for a personality trait, but only in late adolescence through a process of social learning. Moreover, although individuals' level of RWA once established seemed to remain reasonably stable during adulthood, it could be markedly altered by changes in their social situations or life experience, such as higher education, becoming a parent, or confronting threatening sociopolitical crises, which again seemed more consistent with it being an attitudinal construct rather than personality. In addition, studies investigating the structure of sociopolitical attitudes and sociocultural values had typically revealed two roughly orthogonal dimensions, with one corresponding closely to and correlating strongly with RWA, and the other to SDO

(Duckitt and Sibley, 2009: Table 1; see also; Duriez and Van Hiel, 2002; Duriez et al., 2005; Saucier, 2000). Again, therefore, these findings suggested that RWA and SDO were measuring social attitudes or values rather than personality.

Finally, research has shown that both RWA and SDO were powerfully influenced by social situational factors. Thus, social threats (such as social, economic, political crises, and instability) assessed at both the societal and the individual level significantly increased societal or individual levels of authoritarianism, whether measured by the F, C, or RWA scale or by independent ratings (e.g., Doty et al., 1991; Duckitt and Fisher, 2003; McCann, 1999; Sales, 1973). The role of threat as a possible primary explanation for RWA, had been, however, largely ignored by theorists, presumably because it did not fit well with the prevailing assumption that authoritarian attitudes were a direct expression of personality. In addition, research has shown that SDO was markedly influenced by membership of dominant social groups, and will shift accordingly when in-group power and status changes (Huang and Liu, 2005).

Beyond the Authoritarian Personality

This critique of the idea of an authoritarian personality (or personalities) has led to the emergence of new theories conceptualizing authoritarianism not as personality but as social attitude and value dimensions (Duckitt, 1989, 2001; Duriez and Van Hiel, 2002; Feldman, 2003; Kreindler, 2005). This reconceptualization has opened up issues and questions that had largely been ignored because of the personality assumption. First, an issue that had been completely neglected in the case of RWA (though less so for SDO), was that of identifying the central, core value or set of values underlying the seemingly diverse attitudinal content of the RWA scale's items that resulted in them cohering together to form a single unitary ideological dimension. The second has been the issue of what personal or social factors causally influence people's adoption of the ideological attitudes assessed by the RWA and SDO scales? And third, how do these social and ideological attitudes influence social outcomes such as generalized prejudice against out-groups and minorities?

An early alternative to the personality concept of authoritarianism was the group cohesion model of RWA (Duckitt, 1989). Although this model did not generate much research itself, its core ideas were adopted by more influential later models. The group cohesion model saw RWA as a social attitude dimension of a broadly ideological nature, and not as a personality dimension. It proposed that the three components of RWA identified by Altemeyer (1981) were not covarying personality traits, but attitudinal clusters, which covaried tightly together because all the three expressed the motivational goal or value of group cohesion, articulated in attitudes favoring the subordination of individual autonomy and self-expression to group cohesion and authority. Thus, Altemeyer's (1981) 'conventionalism' component was expressed in attitudes favoring the maintenance of the traditional norms of the group, his 'authoritarian submission' in attitudes favoring obedience and respect to group leaders and authorities, and his 'authoritarian aggression' in attitudes favoring punitiveness toward those who violated traditional group norms and conventions. This model suggested that the degree to which people would value group cohesion and therefore hold authoritarian attitudes would be jointly determined by the degree to which they identified with their social group and perceived threats to it. Out-group dislike or prejudice would therefore be caused by perceiving out-groups as threatening in-group cohesion or security in some way.

An interesting implication of this theory was that authoritarian attitudes could be held in respect of any social group. The items of the RWA scale (and of its predecessors) were formulated at the societal or national level so that the RWA scale measured authoritarian attitudes only in relation to the individual's society or nation. This group cohesion model was later elaborated by Stellmacher and Patzel (2005), who developed a General Authoritarianism scale with items that could be applied to any social group and so measure authoritarian attitudes in relation to any group (e.g., such as political parties).

A second theory influenced by the group cohesion model was proposed by Feldman (2003). He also viewed RWA, or at least its core authoritarian submission items, as an attitudinal expression of the value of group conformity or cohesion as opposed to individual autonomy. Feldman's theory did not attempt to explain what caused individuals to adopt the authoritarian values of group conformity or cohesion, but did suggest that once they held these values their perception of threats to cohesion and conformity in their groups would cause them to adopt and support authoritarian policies and actions (this included Altemeyer's 'authoritarian aggression' component) and direct prejudice at those seen as threatening in-group cohesion and conformity.

A third model influenced by the group cohesion Model was Kreindler's (2005) Dual Group Processes Model (DGPM), which also saw authoritarianism as a group phenomenon, and therefore potentially applicable to any social group. Like the group cohesion model the DGPM saw the primary causal determinant of RWA as identification with a group which was experiencing threats to its core norms, which caused hostility to those seen as responsible for these normative threats. The DGPM extended this group identification hypothesis to explain SDO as well, which it suggested was caused by identification with high status or dominant groups that resulted in high-group identifiers valuing inequality, hierarchy, and group dominance, and derogating members of lower status groups.

The most comprehensive and best researched theory deriving from the original group cohesion model has been the dual process model (DPM) of ideology and prejudice (Duckitt, 2001; Duckitt and Sibley, 2009). The DPM proposes that RWA and SDO are social attitudinal expressions of two different basic value or motivational goal dimensions that are made salient for individuals by their personalities and socialized worldview beliefs. Following the group cohesion approach, the DPM proposes that RWA is an expression of the motivational goals or values of collective security and cohesion, which is causally influenced by personality (the Big Five trait dimensions of low openness and high conscientiousness) and the belief acquired through socialization that the social world is a dangerous and threatening (as opposed to safe and secure) place. SDO is seen as an attitudinal expression of the motivational goal or value of power, dominance, and superiority over others, which is causally influenced by the Big Five personality

trait dimension of low agreeableness and the belief that the social world is a ruthlessly competitive jungle in which the strong win and weak lose (as opposed to a cooperative place where people share with and care for each other). Because people high in RWA value collective security and cohesion, they will direct prejudice and hostility toward persons and groups seen as threatening collective security and cohesion. Persons high in SDO, on the other hand, value power, dominance, and superiority over others and will therefore derogate and dislike out-groups low in power or status (in order to justify their relatively superiority) and out-groups competing with their own group over relative power and status.

Finally, a motivated cognition theory of ideology has been proposed by Jost and his colleagues (Jost et al., 2003), which has similarities to the DPM and group cohesion theories, but also important differences. This theory sees RWA (and related constructs, such as the F, C, or D scales) as expressing attitudinal resistance to change, and SDO as expressing attitudinal support for inequality, as two components of a single higher order political conservatism dimension. Conservatism, and therefore both RWA and SDO, express motives to manage and control threat and uncertainty, which arise from social situational factors likely to activate threat and uncertainty, and dispositional factors indexing personal needs to avoid uncertainty and threat. Jost et al.'s (2003) motivated cognition theory is therefore similar to the DPM and cohesion theories in seeing conservatism not in terms of personality but as a social attitudinal or ideological dimension that is an expression of basic motivational goals involving needs to manage threat and uncertainty and which is influenced by both personality and situational factors. It differs, however, from the DPM in seeing political conservatism as a single higher order dimension comprising both RWA and SDO, which are therefore seen as having the same motivational, personality, and situational bases. It also tends to differ from the cohesion influenced theories by seeing RWA as a more limited construct, involving only attitudes expressing resistance to change instead of a broader range of social attitudes favoring the subordination of individuals to group authority and cohesion.

These new models of authoritarianism therefore differ from the traditional authoritarian personality theories in seeing RWA and SDO as social attitude dimensions that express basic motivational goals or values, rather than as personality dimensions. While these new models see authoritarian attitudes as influenced by personality, they also strongly emphasize the causal influence of social factors as equally or indeed possibly even more important. And finally, these new models help to explain why authoritarian attitudes might cause prejudice, an issue that the traditional authoritarian personality approach largely glossed over. During the past decade, both the DPM and Jost's motivated cognition approach have generated a good deal of supportive research (for reviews, see Duckitt and Sibley, 2009; Jost, 2009).

Conclusions

The concept of an authoritarian personality to explain patterns of relatively stable individual differences in a broad range of social, political, and intergroup attitudes and reactions emerged early in the twentieth century. Since then it has inspired a great deal of research and controversy, which have lead to two important changes in conceptualization and measurement. One change was away from trying to measure the entire range of attitudes and beliefs originally seen as comprising the authoritarian syndrome on a single psychometric dimension. These early measures, such as the F, D, and C scales, invariably failed, and this led to the discovery that this broad social attitudinal domain comprised two distinct dimensions, which seem to comprehensively organize individuals' social, political, and intergroup attitudes and their many manifestations.

The second change has come more recently. This has involved challenging the conception of authoritarianism as personality, be it one dimension or two. New theories have emerged which see the two dimensions of authoritarianism, RWA and SDO, not as personality dimensions but rather as two distinct social attitudinal dimensions expressing two sets of motivationally based social values. The main focus of these newer theories has been to clarify the values that lie at the core of ideological attitude dimensions such as RWA and SDO, how situational, experiential, and personality factors might influence these dimensions, and how and why RWA and SDO in turn affect social and intergroup behavior.

See also: Aggression, Social Psychology of; Intergroup Relations; Levels of Analysis in Social Psychology; Racism: Social Psychological Perspectives; Social Dominance Orientation; Social Psychological Theory, History of; Social Psychology.

Bibliography

Adorno, T., Frenkel-Brunswick, E., Levinson, D., Sanford, N., 1950. The Authoritarian Personality. Harper, New York.
Allport, G., 1954. The Nature of Prejudice. Reading. Addison-Wesley, Massachusetts.
Altemeyer, B., 1981. Right-Wing Authoritarianism. University of Manitoba Press, Winnipeg, Canada.
Altemeyer, B., 1988. Enemies of Freedom: Understanding Right-Wing Authoritarianism. Jossey-Bass.
Altemeyer, B., 1996. The Authoritarian Specter. Harvard University Press, Cambridge, MA.
Altemeyer, B., 1998. The other "authoritarian personality". In: Zanna, M. (Ed.), Advances in Experimental Social Psychology, vol. 30. Academic Press, San Diego, pp. 47–92.
Brown, R., 1965. Social Psychology. Free Press, New York.
Christie, R., Jahoda, M. (Eds.), 1954. Studies in the Scope and Method of "the Authoritarian Personality". Free Press, Glencoe, Illinois.
Doty, R., Peterson, B., Winter, D., 1991. Threat and authoritarianism in the United States, 1978–1987. Journal of Personality and Social Psychology 61, 629–640.
Duckitt, J., 1989. Authoritarianism and group identification: a new view of an old construct. Political Psychology 10, 63–84.
Duckitt, J., 1992. The Social Psychology of Prejudice. Praeger, New York.
Duckitt, J., 2001. A dual process cognitive-motivational theory of ideology and prejudice. In: Zanna, M.P. (Ed.), Advances in Experimental Social Psychology, vol. 33. Academic Press, San Diego, pp. 41–113.
Duckitt, J., Fisher, K., 2003. The impact of social threat on worldview and ideological attitudes. Political Psychology 24, 199–222.
Duckitt, J., Sibley, C.G., 2009. A dual process motivational model of ideology, politics, and prejudice. Psychological Inquiry 20, 98–109.
Duriez, B., Van Hiel, A., 2002. The march of modern fascism: a comparison of social dominance orientation and authoritarianism. Personality and Individual Differences 32, 1199–1213.

Duriez, B., Van Hiel, A., Kossowska, M., 2005. Authoritarianism and social dominance in Western and Eastern Europe: the importance of the socio-political context and of political interest and involvement. Political Psychology 26, 299–320.
Feldman, S., 2003. Enforcing social conformity: a theory of authoritarianism. Political Psychology 24, 41–74.
Fromm, E., 1941. Escape from Freedom. Rinehart, New York.
Huang, L., Liu, J., 2005. Personality and social structural implications of the situational priming of social dominance orientation. Personality and Individual Differences 38, 267–276.
Jost, J., 2009. "Elective affinities": on the psychological bases of left-right ideological differences. Psychological Inquiry 20, 129–141.
Jost, J., Glaser, J., Kruglanski, A., Sulloway, F., 2003. Political conservatism as motivated social cognition. Psychological Bulletin 129, 339–375.
Kreindler, S., 2005. A dual group processes model of individual differences in prejudice. Personality and Social Psychology Review 9, 90–107.
Maslow, A., 1943. The authoritarian character structure. Journal of Social Psychology 18, 401–411.
McCann, S., 1999. Threatening times and fluctuations in American church memberships. Personality and Social Psychology Bulletin 25, 325–336.
Pratto, F., Sidanius, J., Stallworth, L., Malle, B., 1994. Social dominance orientation: a personality variable predicting social and political attitudes. Journal of Personality and Social Psychology 67, 741–763.
Reich, W., 1975. The Mass Psychology of Fascism. Penguin, Harmondsworth, England.
Rokeach, M., 1954. The nature and meaning of dogmatism. Psychological Review 61, 194–204.
Rokeach, M., 1960. The Open and the Closed Mind. Basic Books, New York.
Sales, S., 1973. Threat as a factor in authoritarianism: an analysis of archival data. Journal of Personality and Social Psychology 28, 44–57.
Saucier, G., 2000. Isms and the structure of social attitudes. Journal of Personality and Social Psychology 78, 366–385.
Sidanius, J., Pratto, F., 1999. Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression. Cambridge University Press, Cambridge, England.
Stellmacher, J., Patzel, T., 2005. Authoritarianism as a group phenomenon. Political Psychology 26, 245–274.
Van Hiel, A., Mervelde, I., 2002. Explaining conservative beliefs and political preference: a comparison of social dominance orientation and authoritarianism. Journal of Applied Social Psychology 32, 965–976.
Wilson, G. (Ed.), 1973. The Psychology of Conservatism. Academic, London.

Authoritarianism

Markus Kemmelmeier, University of Nevada, Reno, NV, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S. Feldman, volume 2, pp. 970–972, © 2001, Elsevier Ltd.

Abstract

Authoritarianism refers to a construct originally created to understand individual differences associated with intolerance, prejudice, and support of fascist ideology. Almost since its inception, the construct has been widely criticized, with shortcomings being identified in theory and measurement. A consensus, dominant in the literature of the last 30 years, distinguished authoritarians' submission to authority, adherence to in-group norms, and aggression toward deviants. However, with new approaches challenging this consensus, recent research focused on the dynamic aspects of authoritarianism, and the contextual circumstances under which it is aroused, and the process by which it shapes social and political attitudes.

The concept of authoritarianism was first introduced by Adorno et al. (1950) with the goal of understanding personality differences associated with intolerance, prejudice, and support of fascist ideology. Yet, almost since its inception, the construct has been widely criticized, with shortcoming being identified in theory and measurement. Challenging the psychodynamic origins of the concept, Altemeyer (1981) reconceived authoritarianism as reflecting a willingness to submit to authority, to adhere to and enforce in-group conventions, and a tendency to aggress against those who challenge established authorities or in-group conventions. Whereas these themes continue to persist in contemporary research, the last decade or so has sparked lively research activity and a continuous refinement of the construct. Authoritarianism is one of the most heavily researched individual-difference variables in the social and behavioral sciences, with researchers across various disciplines, including psychology, sociology, and political science, exploring its origins, nature, and consequences.

Origins

Research on authoritarianism is inherently linked to the emergence of fascist regimes in Europe in the 1930s and the subsequent violent oppression and persecution of social, religious, ethnic, and sexual minorities. Rooted in the psychoanalytic tradition, European psychologists sought to answer as to how and why an otherwise cultured and civilized society like Germany would support ethnocentrism making the Holocaust possible (e.g., Fromm, 1941). Theodore Adorno, a German Marxist theorist, and Else Frenkel-Brunswick, a German-Austrian psychoanalyst, both emigrants from Europe, joined forces with Berkeley psychologist Neville Sanford and his doctoral student Daniel Levinson, who were already working on the origins of anti-Semitism in personality structure. The research culminated in the volume *The Authoritarian Personality* by Adorno et al. (1950). Relying on in-depth interviews and survey research, the book presented in almost 1000 pages an ambitious attempt to integrate personality theory and psychoanalytical theory. The central characteristic of Adorno et al. approach was to locate the origins of the authoritarian personality in early childhood experiences. Consistent with the psychodynamic thinking of the day, Adorno et al. argued that authoritarianism was the result of harsh and punitive child-rearing practices where the child could never be certain of parental affection. This resulted in a deep-seated psychological conflict in which the desire for acceptance combined with rage toward the parents. Repressed hostility was eventually transferred onto minority groups, while individuals' desire for acceptance made them susceptible to the appeal for dominant leaders. The authors further argued that authoritarians had a 'weak ego,' which led them to adhere to conventional norms and values and made them intolerant of ambiguity. Thus, authoritarian tendencies, including prejudice and the willingness to aggress against members of other groups, were primarily a reflection of individual psychological needs, rather than the nature of intergroup relations or a particular political structure.

Along with their theoretical account, Adorno et al. (1950) provided a survey measure of authoritarianism, labeled the 'California F-scale.' This F-scale was used in thousands of publications (e.g., Meloen, 1993), which documented its association with an impressive array of social and political attitudes and characteristics.

Early Criticism

Publication of *The Authoritarian Personality* was followed by an extensive debate on the merits of the work, with many of the themes of this debate persisting in authoritarianism research to the present day. One concern was that the construct of authoritarianism was ideologically, rather than scientifically motivated, as it focused exclusively on the political right, but ignored authoritarianism on the left (Eysenck, 1954; Shils, 1954). This critique was justified in that much research on authoritarianism was indeed sparked by the experiences of right-wing extremism in Germany and Austria, with two of the authors having fled Nazi tyranny by migrating to the United States.

Another concern was the fact that the F-scale only included items where a higher level of agreement indicated a higher level

of authoritarianism (Shils, 1954). Even when the original authors made various attempts to control for this type of bias (Adorno et al., 1950), there was the possibility that individuals with a greater tendency to acquiesce might have been falsely identified as authoritarian. Commentators pointed out that the item content was often ambiguous, with double-barreled questions making it hard to interpret the meaning of responses. And, although the F-scale claimed to measure one coherent construct, factor analyses routinely demonstrated that it was in fact multidimensional with factors being difficult to interpret. As Altemeyer (1981) noted, factors did not map onto components of authoritarianism as hypothesized by Adorno et al. (1950).

Altemeyer's Reconceptualization

Within 30 years of the publication of *The Authoritarian Personality*, empirical support for the psychodynamic origins of authoritarianism remained largely elusive. But the F-scale remained empirically useful, and there were various attempts to conceptualize the construct based on learning experiences and social class. The reconceptualization that was by far most successful was by Bob Altemeyer (1981, 1996). Based on a sweeping critique of existing research, Altemeyer proposed a new approach to authoritarianism based on social learning theory. This approach was widely adopted during the 1980s and 1990s, and set the stage for much of the contemporary discussion of authoritarianism construct.

Altemeyer viewed authoritarianism as a cluster of learned attitudes that individuals acquire through their upbringing, and their interactions with peers, institutions, media, and the society at large. In line with the origins of the concept, he focused exclusively on authoritarianism on the political right where he distinguished three components: first, a willingness to submit to authority; second, a tendency to adhere to and enforce in-group norms; and third, a willingness tendency to aggress against those who challenge established authorities or in-group norms. This tripartite structure guided the construction of his Right-Wing Authoritarianism (RWA) scale and was subsequently confirmed through empirical research, even when Altemeyer continued to view RWA as a single coherent construct. Altemeyer continually updated his RWA scale after submitting it to repeated empirical tests, which produced a variety of versions of the same scale. Other than Adorno et al. (1950) F-scale, the RWA scale avoids potential response biases by using reverse-coded items; its psychometric properties are sound with Cronbach's alpha reliabilities routinely approaching, if not exceeding 0.90. Moreover, the availability of short versions of the instrument allowed for the inclusion of RWA measures in a wide range of studies.

Altemeyer's RWA conception is arguably dominant in the literature, and the RWA scale, with its different versions, is the most widely used authoritarianism scale used to date. It is a reliable predictor of a broad range of social and political attitudes, including prejudice, intolerance, religious fundamentalism, and sexual inhibition (e.g., Altemeyer, 1998; Sibley et al., 2006). Still, Altemeyer's approach has met with criticism, which has sparked continuous refinement of the concept of authoritarianism and its empirical assessment (e.g., Duckitt, 2001; Stenner, 2005).

Measuring RWA

In spite of the solid psychometrics of the RWA scale, research has challenged the claim that it represents one coherent construct. Factor analytic studies documented two or three independent factors (e.g., Feldman, 2003; Funke, 2005; Van Hiel et al., 2007). This prompted Duckitt et al. (2010) to measure each of Altemeyer's (1981) three components using separate scales. The authors noted that the components reflect the distinct constructs of Authoritarianism (authoritarian aggression), Conservatism (authoritarian submission), and Traditionalism (conventionalism). The resulting ACT model allows for the testing of differential predictions pertaining to different components of the model.

Another line of criticism refers to a confound between the content of the RWA scale and the content of attitudes that the RWA scale is often used to predict. For instance, some RWA items ask respondents as to what to do about societal minority groups. It is hardly surprising if RWA is then found to correlate with hostile attitudes toward specific minority groups. More blatantly, a correlation between most versions of the RWA scale and attitudes toward gays and lesbians may not be very informative because the RWA scale does include items referencing homosexuality. In other words, the item content of the RWA scale makes it impossible to distinguish between an authoritarian disposition and the social consequences of this disposition, thus producing potentially trivial findings. Feldman and Stenner (1997) and Stenner (2005) proposed that this is the unavoidable consequence when authoritarianism is conceived as a cluster of attitudes. These authors define authoritarianism as a generalized motive aimed at maintaining and creating order and uniformity, and measure it in a manner that is distinct from any social or political preferences (see also Feldman, 2003). Prominently, Stenner (2005) assessed authoritarianism via individuals' preferences of child-rearing practices, with authoritarian tendencies being reflected in a preference for well-behaved and obedient children, rather than independent and curious children. Notably, items focusing on children and child-rearing had always been part of Altemeyer's (1981) RWA scale.

But even though scholars differ in how authoritarianism was conceptualized and measured, Van Hiel et al. (2007) documented a remarkable level of convergence of different scales.

Authoritarianism and Conservatism

Some scholars have treated authoritarianism measures as a *de facto* measure of conservatism (e.g., Jost et al., 2003a). Whereas measures of conservatism and authoritarianism are often substantially correlated, Wilson (1973) argued that authoritarianism is but one specific manifestation of conservatism, and only becomes indistinguishable in the presence of threat (see below). The link between authoritarianism and social conservatism is well-documented, but authoritarianism is less-strongly related to the acceptance of inequality, and similarly

weakly, if not negatively, related to economic conservatism and (e.g., Van Hiel et al., 2007). Jost et al.'s (2003a) recent definitions of conservatism as implying the maintenance of the status quo and acceptance of societal inequality allows for a clearer conceptual distinction between conservatism and authoritarianism, even when this definition is not universally accepted (e.g., Duckitt et al., 2010).

Stenner (2005) offers to date one of the clearest definitional separations (but see Duckitt et al., 2010). According to Stenner (2005), authoritarianism is a motive aimed at reducing differences across space, whereas conservatism is a motive aimed at minimizing differences over time. Whereas authoritarian is focused on reducing diversity and increasing uniformity within today's society, conservatism is focused on maintaining stability and resisting change. Defined in this manner, authoritarians may support policies aimed at social change if they expect that those policies will produce greater uniformity. Conversely, conservatives might be supportive of diversity, civil liberty and individual expression if those are viewed as traditions or an established institutional reality. According to Stenner (2005) conservatives only gravitate toward authoritarianism when public opinion is fragmented, and they view established institutions as dysfunctional. But whereas authoritarians can be expected to limit individual freedom, conservatives are likely to defend freedom and democracy within an established institutional context (see also Feldman and Stenner, 1997).

Authoritarianism on the Left

A recurring theme in the literature on authoritarianism is the notion that left-leaning researchers are oblivious, if not willfully ignorant, of authoritarian tendencies on the left side of the political spectrum. This concern emerged in part because Adorno et al. (1950) had painted right-wingers in a decidedly unfavorable light, with authoritarians, preoccupied with their own childhood conflicts, being ready to engage in blind obedience and mass murder as long as it satisfied their psychological needs. Arguably, Adorno et al. (1950) had defined norm conformity and in-group solidarity as grounded in pathology and irrationality, and ignored left-wing extremism (Eysenck, 1954; Shils, 1954; Rokeach, 1960). A similar sentiment was echoed by Durrheim (1997), who argued that attempts of researchers to associate certain ideological beliefs with irrationality can be interpreted as an attempt to devalue said beliefs. Indeed, some may charge that authoritarianism research was little more than a thinly veiled attempt to denigrate conservative viewpoints. Even if this may be a minority position, the debate surrounding Jost et al.'s (2003a) comprehensive view of cognitive style and political orientation included Greenberg and Jonas (2003) pointing out that the rigidity and dogmatism of extreme left-wingers remained essentially unexamined.

There had been efforts to document that extremists on the left and right were largely similar in terms of their tough-mindedness (Eysenck, 1954) and their rigidity (Rokeach, 1960). However, convincing evidence for a left-wing authoritarianism remained largely elusive during the first 30 years of research on authoritarianism (Stone, 1980). Yet, when the end of the Cold War enabled broader access to eastern European samples, scales that were originally conceived as assessing RWA were found to be related to support for Communism (e.g., McFarland et al., 1992). This underscores that individual differences in authoritarianism might be best understood as disposing individuals to support an established social system that is compatible with authoritarians' preference for order and uniformity.

Additional evidence for left-wing authoritarianism was found in Belgium, even when the factor structure was somewhat different from RWA (Van Hiel et al., 2006). However, at present it is still unclear to what extent this concept and its measurement generalize to other contexts.

RWA versus Social Dominance Orientation

Over the last 20 years research on authoritarianism has been cross-fertilized by the emergence of social dominance theory and the widespread use of the Social Dominance Orientation (SDO) scale, which measures individual differences in the acceptance or support of hierarchy and inequality between social groups (Pratto et al., 1994). Though originally suspected to be redundant with RWA, research quickly documented only a modest correlation between the two (e.g., Altemeyer, 1998). Much of current research treats RWA and SDO as complementary, with RWA referring to authoritarian submission and SDO referring to support for power-seeking leaders. This broader conception of authoritarianism as a two-sided phenomenon is apparent in Duckitt's (2001) dual-process model of ideology and prejudice, which relies equally on RWA and SDO as constituent variables. Although the present review focuses on authoritarianism more narrowly, note that much of contemporary research assesses RWA and SDO in tandem (e.g., Duriez et al., 2008). Results typically reveal RWA and SDO as independent and additive antecedents of intergroup attitudes (e.g., Sibley, et al., 2006).

Developmental Origins of Authoritarianism

Advances in research on human genetics have long suggested that dispositions toward certain sets of attitudes and ideologies might be predicted by one's genetic heritage. Such evidence is also available for authoritarianism, with twin studies indicating that up to 50% of the interindividual variability might be transmitted genetically (e.g., McCourt et al., 1999). Yet, this finding also suggests that a tremendous amount of individual variance occurs due to environmental variability, even when heritability estimates might have been biased downward due to assortative mating (Kandler et al., 2012; see below). Moreover, with genetic influences being not necessarily deterministic, it is critical to understand the environmental circumstances under which certain genes might be expressed.

Even when they differed categorically with regard to the nature of the processes involved, both Adorno et al. (1950) and Altemeyer (1981) located the origins of authoritarian tendencies in individuals' upbringing, specifically the interaction between parent and child. Consistent with this idea, a number of studies documented that authoritarianism seems to be running in families, with there being a substantial correlation between

levels of authoritarianism in parents and their children (e.g., Duriez et al., 2008). However, it is unclear whether the intergenerational transmission of authoritarianism occurred due to socialization experience or the genetic similarity between parents and offspring. Notably, Duriez et al. (2008) documented that parents high and low in authoritarianism systematically differed in the kinds of goals that they fostered in their children, and that parental goals mediated effect of parental authoritarianism on their adolescent children's authoritarianism. This finding demonstrates the importance of socialization practices for the intergenerational transmission of authoritarianism.

Authoritarianism and Information Processing

Early on, research on authoritarianism made strong predictions about the cognitive characteristics of authoritarians. Authoritarians were expected to be less tolerant of ambiguity, more rigid and cognitively inflexible. In part because the former are characteristics of high cognitive functioning, authoritarians were also predicted to be lower in intelligence than nonauthoritarians (Adorno et al., 1950). This *rigidity-of-the-right* perspective was soon challenged by Rokeach (1960) who claimed that extremists, regardless of whether they are authoritarians on the right or committed communists on the left, should be equal in their level of rigidity. This might be dubbed the *rigidity-of-the-extreme* perspective.

Both perspectives received at least some support. Modest support only emerged for the *rigidity-of-the-extreme* perspective, and often authoritarianism was not directly assessed (Jost et al., 2003b). However, there is ample support for the *rigidity-of-the-right* perspective. Consistent with early research, measures of authoritarianism are often found to be associated with self-report measures of cognitive rigidity, intolerance of ambiguity, need for structure and low attributional complexity (Altemeyer, 1998; Kemmelmeier, 2010b; see Jost et al., 2003a for a comprehensive review). Conversely, they are negatively related to need for cognition, a construct measuring an individual's tendency to engage in and enjoy thinking (e.g., Kemmelmeier, 2010a).

However, as noted by Van Hiel et al. (2010), the meaning of a correlation between a self-report measure of authoritarianism and a self-report measure of cognitive style is inherently problematic because the presence of a method-related artifact cannot be conclusively excluded. Moreover, researchers have suspected allegedly value-neutral measures to be ideologically biased (e.g., Durrheim, 1997; Van Hiel et al., 2010). Thus, there has been a call for the assessment of objective performance measures, which allow more reliable inferences about individuals' cognitive processing. In one such attempt, Durrheim (1998) was unable to find a consistent relationship between authoritarianism and attitudinal ambivalence, that is, individuals' tendency to hold both positive and negative feelings toward an attitude object. Though based on self-reports of attitudes, measures of attitudinal ambivalence can be objectively computed and served as measure of tolerance of ambiguity. In contrast, Kemmelmeier (2010a) demonstrated that authoritarians were more likely to trust their intuition and engage in heuristic processing when solving reasoning tasks, even when doing so produced normatively incorrect solutions. This finding is in line with the observation that measures of cognitive ability or intelligence are generally, though not consistently, negatively correlated with measure of authoritarianism and right-wing attitudes (see Van Hiel et al., 2010; McCourt et al., 1999 for reviews).

Authoritarianism and Personality

A recent meta-analysis of the relationship between the Big-5 personality dimensions and RWA revealed a substantial negative correlation with openness to experience (Sibley and Duckitt, 2008). Whereas there was no association with agreeableness, extraversion and neuroticism, RWA also showed a weak, positive association with conscientiousness. These findings are broadly consistent with the notion that authoritarians prefer order and uniformity. But while authoritarianism may be derived from individuals' (lack of) openness to experiences and conscientiousness, Sibley and Duckitt (2008) argue that authoritarianism should not be itself regarded as a basic personality variable – contrary to Adorno et al. (1950). Rather, as suggested by Duckitt (2001), it might function like a value orientation. Consistent with the idea that personality traits motivate value concerns, which are then expressed through attitudes and behavior, Sibley and Duckitt (2008) demonstrate that RWA mediates the relationship between openness to experience and prejudice.

Authoritarianism and Well-Being

The authors of *The Authoritarian Personality* clearly viewed authoritarianism as psychologically maladaptive. Wilson (1973) likened the authoritarian personality to "a kind of *pathological syndrome* like psychopathology or schizophrenia" with various authors often taking it for granted that authoritarians are psychologically conflicted, highly anxious, and of low self-esteem. Yet, evidence of maladaptive implications of authoritarianism is mixed with most studies failing to document a link with psychopathology (see Van Hiel and De Clercq, 2009; Duriez et al., 2012 for reviews). Van Hiel and De Clercq (2009) demonstrated that authoritarianism might even shield individuals from the psychologically maladaptive consequences of adverse life events. This result was not corroborated by Duriez et al. (2012) who found authoritarianism to be linked to an increase in depressive symptoms over time. This is consistent with Bonanno and Jost's (2006) observation that increased levels of depression and posttraumatic symptoms persisted longest among those high-exposure survivors of the 9/11 attacks who also reported higher levels of authoritarianism. However, neither trait anxiety (Onraet et al., 2013) nor self-esteem (Altemeyer, 1998) appear to be associated with authoritarianism.

Authoritarianism and Gender

Authoritarianism has ample implications for gender and gender relations (see Peterson and Zurbriggen, 2010 for a comprehensive review). Women higher in RWA are more likely to pursue traditional career and family goals than women low in RWA,

with nontraditional role combinations being experienced as stressful. Similarly, high RWA women were more attracted to men with traditionally masculine characteristics than low RWA woman. On the other hand, high RWA men rated their own femininity to be low compared to low RWA men even when there was no difference with regard to self-described masculinity. In both men and women, high RWA was related to sexual conservatism, rape myth acceptance (i.e., the tendency to blame a rape victim for being raped), the rejection of casual sex, as well as a rejection of gay men and lesbian women. Overall, high RWA individuals seem to be committed to traditional conceptions of masculinity and femininity, and a traditional understanding of gender roles, with high levels of agreement between men and women. Arguably, this sets the stage for assortative mating, whereby men and women with similar social attitudes and political leanings are disproportionately likely to have children together (Kandler et al., 2012).

Authoritarianism and Threat

Critics of *The Authoritarian Personality* have long charged that Adorno et al.'s (1950) focus on within-individual dynamics failed to take changing social and political circumstances into consideration. Others questioned Altemeyer's (1981) social learning explanation as such processes cannot easily accommodate fluctuations in authoritarianism over time. Indeed, recent research has pointed to the paramount role that external threat plays in authoritarianism (Feldman, 2003; Onraet et al., 2013; Stenner, 2005).

Focusing primarily on archival data (Doty et al., 1991; Feldman and Stenner, 1997), the presence of threat has been found to be associated with an increase in authoritarianism as measured via increases in punitiveness, prejudice, intolerance, and aggression. Though various causal accounts were suggested for this type of main effect, authors typically explained increases in authoritarianism as the result of in-group solidarity and protecting one's society's integrity (e.g., Altemeyer, 1996; Duckitt, 1989; see below).

More recent approaches have focused on the interactive effect of threats and preexisting dispositions, such that threat activates authoritarian tendencies among individuals who are thus disposed, but not necessarily among those who do not harbor such motives (Duckitt, 2001; Feldman and Stenner, 1997). Stenner (2005) described this two-variable interaction as "intolerance = authoritarianism × threat" and labeled it the *authoritarian dynamic*. McCann (2008) illustrates this dynamic in an archival study of death penalties and executions in different US states. In conservative US states, higher levels of threat were associated with more death penalties and executions, consistent with previous findings that threat increases authoritarian responses. However in liberal states, higher levels of threat were linked to fewer death penalties and executions, suggesting that the presence of external threat only intensified the response based on a nonauthoritarian disposition. This finding is difficult to reconcile with the main effects model described above. It is also somewhat at odds with the observation of a rightward shift in ideology observed among high-exposure survivors of the 9/11 attacks in New York observed (Bonanno and Jost, 2006). Witnessing a horrific act of terrorism induced liberals, centrists, and conservatives alike to shift their political views toward the right.

Aside from threat activating any latent authoritarian dispositions, a pervasive finding in the literature is that, regardless of levels of objective threat, authoritarians are more inclined to view the world as a dangerous place (Altemeyer, 1998; Duckitt, 2001). Theorists generally assume that this heightened perception of threat is responsible for authoritarians' increased willingness to engage in aggression against deviants (e.g., McFarland, 2005). A complex simulation game reported by Altemeyer (1996) compared how a group of high-RWA versus a group of low-RWA individuals would run the world for 40 years. The high-RWA simulation eventually ended in a nuclear holocaust and the death of the world's population. By contrast, the low-RWA simulation eventually resulted in a more or less stable world order in which pressing problems were by and large cooperatively addressed. Altemeyer (1996) attributed the failure of the high RWA group primarily to their distrust of others, which steered them toward aggressive action early on.

Authoritarianism as a Group Phenomenon

The question of the importance of threat for authoritarianism is embedded into the broader question as to the origins and purpose of authoritarianism as a social phenomenon. Adopting an evolutionary approach, there is evidence that differences in the prevalence of authoritarian forms of government as well as the prevalence of individual-level authoritarianism are related to the historical prevalence of pathogen prevalence and the stress that pathogens imposed on ancestral human societies (Murray et al., 2013). The explanation of this relationship focuses on the notion that in premodern societies, authoritarianism facilitated a collective response focused on isolating infected individuals. At a time when disease-causing processes were poorly understood and effective treatments were not available, disrupting pathways of infection by avoiding contact with pathogen carriers was the only means of protecting a collective. This may have fostered behavioral norms that focused on conformity because individuals with symptoms of disease were necessarily deviant; thus, rejecting deviant individuals meant avoiding contact with the carrier of a pathogen. Conversely, nonconformity might have either exposed individuals to pathogens or rendered them suspect in the eyes of fellow in-group members, who would then treat them as deviant. In this manner, fear of infection may have also promoted forms of governance that enforced order and uniformity. According to this account, authoritarianism is ultimately an evolutionary adaptation in response to environmental threat (Murray et al., 2013).

Even though this evolutionary account has not yet been widely embraced, it nevertheless conceives of authoritarianism as a response that is inherently embedded in a group process – an idea that has gained increasing prominence in research on authoritarianism. Duckitt (1989) first proposed that authoritarianism is best conceived as the normative expectation of individuals' conformity to and identification with the group for the purposed of enhancing group cohesion. The notion that authoritarianism is inherently a phenomenon

that is aimed at fostering in-group uniformity and cohesion, implying a rejection of deviants and nonconformists, is now a central element in theories of authoritarianism and prejudice (e.g., Duckitt, 2001; Feldman, 2003; Stenner, 2005). Although individuals vary in the extent to which they possess a disposition toward hostile responses toward deviants, such responses should be primarily expected in the presence of perceived threat to one's group, not necessarily in response to threats to one's well-being. Consistent with this idea, Onraet et al. (2013) show that only external threats (e.g., perceived threat to the integrity of one's society) were strongly and consistently related to authoritarian views, whereas the relationship for internal threats (e.g., anxiety, psychological distress) was weak and inconsistent.

Authoritarianism in International Perspective

Much of the research on authoritarianism originated in the United States and Canada, but research on the construct was never limited to North America. The popular RWA has been translated in various different languages or adapted to different culture contexts, and it yields results that are largely consistent across cultures, while also revealing cross-cultural differences (e.g., Duriez et al., 2005). Evidence of a possible link between cultural values stems from a study involving seven societies, in which RWA was found to be correlated with vertical collectivism, a value construct that measures individuals' respect for and willingness to submit to culturally sanctioned authority (Kemmelmeier et al., 2003). At the same time, RWA was not associated with horizontal individualism, a value orientation focused on individual self-determination and personal uniqueness. The only exceptions were postcommunist societies where the link between RWA and horizontal individualism was negative, presumably because totalitarian had rendered it salient that authoritarian tendencies were incompatible with the pursuit of personal freedom.

Conclusion

Since its inception, the concept of authoritarianism has demonstrated a remarkable degree of longevity. Yet, research on authoritarianism has been successful because it was never beholden to its theoretical beginnings, but continuously evolved. Rather than understanding authoritarianism exclusively as a personality variable, progress in theorizing over the last 10–15 years has clarified the nature of authoritarian dispositions and processes, and examined their activation and emergence within threatening social contexts. Of particular importance were new approaches, which allowed for a sharper definition of the authoritarianism concept and enabled its integration in more encompassing theoretical frameworks. And whereas much research has expanded our knowledge of the implications of authoritarianism, often yielding surprising results, contributions of behavioral genetics and developmental research have opened new pathways of understanding how authoritarian dispositions might be passed on across generations. Still, significant challenges remain, including a better understanding of the nature of threats, which arouse authoritarian responses, the variability of such responses over time and across situations, as well as the role of socialization practices and societal institutions in enhancing or mitigating authoritarianism. However, the last two decades have shown that authoritarianism is not only a reliable individual-level predictor of prejudice and intolerance, but that it enables substantial insights into the origin and dynamics of social attitudes.

See also: Aggression, Social Psychology of; Authoritarian Personality; Dark Triad: The "Dark Side" of Human Personality; Epistemological Beliefs and Motivation; Five Factor Model of Personality, Assessment of; Five Factor Model of Personality, Facets of; Group Processes, Social Psychology of; Intelligence: Central Conceptions and Psychometric Models; Levels of Analysis in Social Psychology; Mass Killings and Genocide, Psychology of; Normality and Mental Health; Obedience: Social Psychological Perspectives; Peace Promotion, Psychology of; Personality Development: Systems Theories; Personality Differences and Development: Genetic and Environmental Contributions; Personality Theory and Psychopathology; Political Psychology; Prejudice and Discrimination; Social Beliefs and Attitudes: Personality and the Isms; Social Dominance Orientation; Social and Economic Change: Psychological Challenges for Individuals; Tyranny; Values, Psychology of.

Bibliography

Adorno, T.W., Frenkel-Brunswik, E., Levinson, D.J., Sanford, R.N., 1950. The Authoritarian Personality. Harper, New York.

Altemeyer, B., 1996. The Authoritarian Specter. Harvard University Press, Cambridge, MA.

Altemeyer, B., 1998. The "other" authoritarian. In: Zanna, M.P. (Ed.), Advances in Experimental Social Psychology, vol. 30. Academic Press, San Diego, pp. 47–92.

Altemeyer, R.A., 1981. Right-Wing Authoritarianism. University of Manitoba Press, Winnipeg, Canada.

Bonanno, G.A., Jost, J.T., 2006. Conservative shift among high-exposure survivors of the September 11th terrorist attacks. Basic and Applied Social Psychology 28, 311–323.

Doty, R.M., Peterson, B.E., winter, D.G., 1991. Threat and authoritarianism in the United States, 1978–1987. Journal of Personality and Social Psychology 64, 629–640.

Duckitt, J., 1989. Authoritarianism and group identification: a new view of an old construct. Political Psychology 109, 63–84.

Duckitt, J., 2001. A dual-process cognitive-motivational theory of ideology and prejudice. In: Zanna, M.P. (Ed.), Advances in Experimental Social Psychology, vol. 33. Academic Press, San Diego, CA, pp. 41–113.

Duckitt, J., Bizumic, B., Krauss, S., Heled, E., 2010. A tripartite approach to right-wing authoritarianism: the authoritarianism-conservatism-traditionalism model. Political Psychology 31, 685–715.

Duriez, B., Klimstra, T.A., Luyckx, K., Beyers, W., Soenens, B., 2012. Right-wing authoritarianism: protective factor against or risk factor for depression? European Journal of Personality 26, 536–549.

Duriez, B., Soenens, B., Vansteenkiste, M., 2008. The intergenerational transmission of authoritarianism: the mediating role of parental goal promotion. Journal of Research in Personality 42, 622–642.

Duriez, B., Van Hiel, A., Kossowska, M., 2005. Authoritarianism and social dominance in Western and Eastern Europe: the importance of the sociopolitical context and of political interest and involvement. Political Psychology 26, 299–320.

Durrheim, K., 1997. Theoretical conundrum: the politics and science of theorizing authoritarian cognition. Political Psychology 18, 625–656.

Durrheim, K., 1998. The relationship between tolerance of ambiguity and attitudinal conservatism: A multidimensional analysis. European Journal of Social Psychology 28 (5), 731–753.

Eysenck, H.J., 1954. The psychology of politics. Routledge and Keagan Paul, London.

Feldman, S., 2003. Enforcing social conformity: a theory of authoritarianism. Political Psychology 24, 41–74.
Feldman, S., Stenner, K., 1997. Perceived threat and authoritarianism. Political Psychology 4, 741–770.
Fromm, E., 1941. Escape from Freedom. Rinehart, New York.
Funke, F., 2005. The dimensionality of right-wing authoritarianism: lessons from the dilemma between theory and measurement. Political Psychology 26, 195–218.
Greenberg, J., Jonas, E., 2003. Psychological and political orientation – the left, the right, and the rigid: comment on Jost et al. (2003). Psychological Bulletin 129, 376–382.
Jost, J.T., Glaser, J., Kruglanski, A.W., Sulloway, F.J., 2003a. Political conservatism as motivated social cognition. Psychological Bulletin 129, 339–375.
Jost, J.T., Glaser, J., Kruglanski, A.W., Sulloway, F., 2003b. Exceptions that prove the rule – using a theory of motivated social cognition to account for ideological incongruities and political anomalies: reply to Greenberg and Jonas (2003). Psychological Bulletin 129, 383–393.
Kandler, C., Bleidorn, W., Riemann, R., 2012. Left or right? Sources of political orientation: the roles of genetic factors, cultural transmission, assortative mating, and personality. Journal of Personality and Social Psychology 102, 633–645.
Kemmelmeier, M., Burnstein, E., Krumov, K., Genkova, P., Kanagawa, C., Hirshberg, M., Erb, H.-P., Wieczorkowska, G., Noels, K., 2003. Individualism, collectivism and authoritarianism in seven societies. Journal of Cross-Cultural Psychology 34, 304–322.
Kemmelmeier, M., 2010b. Gender moderates the impact of need for structure on social beliefs: Implications for ethnocentrism and authoritarianism. International Journal of Psychology 45 (3), 202–211.
Kemmelmeier, M., 2010a. Authoritarianism and its relationship with intuitive-experiential and rational-analytical cognitive styles. Personality and Individual Differences 48, 44–48.
McCann, S.J., 2008. Societal threat, authoritarianism, conservatism, and US state death penalty sentencing (1977–2004). Journal of Personality and Social Psychology 94, 913–923.
McCourt, K., Bouchard Jr., T.J., Lykken, D.T., Tellegen, A., Keyes, M., 1999. Authoritarianism revisited: genetic and environmental influences examined in twins reared apart and together. Personality and Individual Differences 27, 985–1014.
McFarland, S.G., 2005. On the eve of war: authoritarianism, social dominance, and American students' attitudes toward attacking Iraq. Personality and Social Psychology Bulletin 31, 360–367.
McFarland, S., Ageyev, V., Abalakina-Paap, M., 1992. Authoritarianism in the former Soviet Union. Journal of Personality and Social Psychology 63, 1004–1010.
Meloen, J.D., 1993. The F scale as a predictor of fascism: an overview of 40 years of authoritarianism research. In: Stone, W.F., Lederer, G., Christie, R. (Eds.), Strengths and Weaknesses: The Authoritarian Personality Today. Springer, New York, pp. 47–69.
Murray, D.R., Schaller, M., Suedfeld, P., 2013. Pathogens and politics: further evidence that parasite prevalence predicts authoritarianism. PloS One 8 (5), e62275.
Onraet, E., Van Hiel, A., Dhont, K., Pattyn, S., 2013. Internal and external threat in relationship with right-wing attitudes. Journal of Personality 81, 233–248.
Peterson, B.E., Zurbriggen, E.L., 2010. Gender, sexuality, and the authoritarian personality. Journal of Personality 78, 1801–1826.
Pratto, F., Sidanius, J., Stallworth, L.M., Malle, B.F., 1994. Social dominance orientation: a personality variable predicting social and political attitudes. Journal of Personality and Social Psychology 67, 741–763.
Rokeach, M., 1960. The Open and the Closed Mind. Basic Books, New York.
Shils, E.A., 1954. Authoritarianism: right and left. In: Christie, R., Jahoda, M. (Eds.), Studies in the Scope and Method of 'The Authoritarian Personality'. Free Press, Glencoe, IL, pp. 24–49.
Sibley, C.G., Duckitt, J., 2008. Personality and prejudice: a meta-analysis and theoretical review. Personality and Social Psychology Review 12, 248–279.
Sibley, C.G., Robertson, A., Wilson, M.S., 2006. Social dominance orientation and right-wing authoritarianism: additive and interactive effects. Political Psychology 27, 755–768.
Stenner, K., 2005. The Authoritarian Dynamic. Cambridge University Press, Cambridge, UK.
Stone, W.F., 1980. The myth of left-wing authoritarianism. Political Psychology 2, 3–19.
Van Hiel, A., De Clercq, B., 2009. Authoritarianism is good for you: right-wing authoritarianism as a buffering factor for mental distress. European Journal of Personality 23, 33–50.
Van Hiel, A., Cornelis, I., Roets, A., De Clercq, B., 2007. A comparison of various authoritarianism scales in Belgium Flanders. European Journal of Personality 21, 149–168.
Van Hiel, A., Duriez, B., Kossowska, M., 2006. The presence of left-wing authoritarianism in Western Europe and its relationship with conservative ideology. Political Psychology 27, 769–793.
Van Hiel, A., Onraet, E., De Pauw, S., 2010. The relationship between social-cultural attitudes and behavioral measures of cognitive style: a meta-analytic integration of studies. Journal of Personality 78, 1765–1800.
Wilson, G.D. (Ed.), 1973. The Psychology of Conservatism. Academic Press, London.

Authority, Social Theories of

Howard G Schneiderman, Lafayette College, Easton, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Authority is a concept denoting a relationship of unequal power, where those who have authority demand obedience from others who are expected to accept the legitimacy of those demands and comply with them. Authority is an elementary form of social life, found in all societies, and in all social institutions.

Theoretical Sources

Theories and Ideologies

For centuries philosophers have written about authority, but they have been less interested in analyzing authority objectively than in using the term ideologically to further their own intellectual agendas regarding justice, power, domination, and government. After all, as one very famous philosopher, Bertrand Russell, has said, "most philosophy has been a reaction against skepticism; it has arisen in ages when authority no longer sufficed to produce the socially necessary minimum of belief, so that nominally rational arguments had to be invented to secure the same result" (Russell, 1960: p. 86). With Russell's pronouncement in mind, this discussion will concentrate on value-neutral social science theories of authority, rather than surveying the vast philosophical literature on the subject.

German Theorists

Most ideas about authority have emanated from the work of Max Weber, who gave us the most basic and up to now most pervasive ideas about how power becomes authority through what he considered its three basic legitimations: traditional authority, rational-legal authority, and charisma. Although Weber is justly famous for his discussions of authority, one should acknowledge that his personal contact with historians, sociologists, and theological thinkers such as Rudolph Sohm, Adolf Harnack, Ernst Troeltsch, and Georg Simmel, all of whom were writing about these three forms of authority before, during, and after he did, played a huge part in Weber's thinking. Indeed, in his well-known work on church history, *Kirchenrecht*, Sohm introduced the term charismatic organization (Sohm, 1892: p. 26), which Weber directly adopted for his theory of charismatic authority. Adolf Harnack's monumental three-volume *Lehrbuch der Dogmengeschichte*, or *History of Dogma*, was published right after Sohm's book, between 1894 and 1898, and also had a major impact on Weber's thinking about authority in the early Christian church, which he then applied to modern societies. Ernst Troeltsch, a close personal friend, and professional colleague of Weber, wrote extensively about premodern and modern forms of institutional authority. Troeltsch's *Protestantism and Progress* (2013) and his *Social Teachings of the Christian Churches* (1912) discussed organizational and personal forms of authority in both church and state, and his work influenced the formation of Weber's famous tripartite theory of authority (Schneiderman, 2013).

Georg Simmel, a German sociologist even more influential than Weber in his day, also wrote about authority, and his excursus on 'super ordination and subordination' remains one of the most inventive ways of thinking about the relationship of authority to power. In this essay, Simmel introduced the concept of the 'prestige leader,' who seems to wield exactly what Weber would later call charismatic authority (Simmel, 1950: p. 185). Robert Michels, Weber's most successful student, published a book, *Political Parties*, which introduced a new term to describe authority, 'the iron law of oligarchy,' which says that the authority of an oligarchic elite in any democratic institution is inevitable. Michels reminds us of the Weberian endorsed idea of a value-free sociology regarding the study of authority: "leadership is a necessary phenomenon in every form of social life. Consequently it is not the task of science to inquire whether this phenomenon is good or evil, or predominantly one or the other" (Michels, 1915: p. 364).With the unification of Germany into a major modern state in the late nineteenth century, these social science thinkers brought a new analytic interest in power and authority, and this is clearly reflected in their work.

French Theorists

French sociology was also quick to become interested in the nationalistic forces boiling up in Europe, and in the new movements to consolidate power and to create the basis of legitimate authority in France and neighboring states. Gabriel Tarde was the first French sociologist to write extensively about authority and its relationship to power in his 1899 book, *Les Transformations du Pouvoir*, where he defined authority as the right to be obeyed. Tarde's well-known three-pronged theory of social order, first introduced in his *Laws of Imitation*, namely, repetition, opposition, and adaptation, are applied to power and authority in modern societies. Tarde sees the origins of political and social authority of the modern state in the history of the family as a social institution, an idea that was later underscored by the American sociologist, Robert Nisbet, in 'Kinship and Power in First Century Rome' (Nisbet, 1999). Another notable French sociologist, Célestin Bouglé, wrote about authority and social class in his famous treatise, *Essais Sur le Régime des Castes*, reiterating a line of thinking originating with Alexis de Tocqueville's famous analysis of the origins of the French Revolution, *The Old Regime and the Revolution*, and leading up to Louis Dumont's work on caste and authority in the 1960s. The most important French sociologist was, of

course, Emile Durkheim who also wrote about authority in his work on law and education. For Durkheim, social authority was not only about roles and statuses in the social order, but also about collective representations and symbols, which exacted obedience to authority through its moral interface with the society. These French sociologists played a part in more contemporary thinking about the cultural and symbolic nature of obedience to authority discussed above.

Italian Theorists

Interest in authority, power, and politics in Italy was set in motion by the three wars of Italian independence fought in 1848, 1859, and 1866 and was further fueled by the unification of the various Italian states as the Kingdom of Italy in 1861. Thus, in Italy, as in Germany, the modern bureaucratic state came into existence late in the nineteenth century and created an interest in power and authority among early social scientists. Among the most important Italian social scientists writing on authority were Gaetano Mosca, Vilfredo Pareto, and Guglielmo Ferrero. As Italy underwent its transformation from an array of tiny principalities, to a loose confederation of small states, and then to a centralized modern state, social scientists became interested in the process of how centralizing power became centralized authority.

Gaetano Mosca, an Italian political scientist whose *Elementi di scienza politica*, first published in 1895, and later in an English translation, *The Ruling Class*, set forth a theory of elites that added a new dimension to theories of authority. Mosca's basic assumption is that all societies are composed of small elite ruling classes that demand obedience from large unorganized majority groups. The pivotal question for Mosca is how the leaders who wield authority do so. As he says, in any form of political organization, authority is either transmitted from above, downward in the political or social scale (he calls this the autocratic principle), or from below, upward (the liberal principle). Mosca also discusses how leaders are chosen to wield authority, or at least from where in the social order they are chosen. The two types of leadership recruitment are aristocratic and democratic. Aristocratic recruitment consolidates power and authority in the hands of a relatively few families over multiple generations, while democratic recruitment emphasizes rejuvenation of elites by bringing up new talented individuals from the lower and middle classes.

Mosca's types of recruitment to authoritative elite positions anticipated Vilfredo Pareto's even more well-known 'circulation of the elite.' Pareto's most famous work, *Mind and Society* (*Trattato Di Sociologia Generale*), published in 1916, analyzed regime change, where authority is transferred from one elite to another because of revolutions from below. Perhaps thinking of the situation in Italy during the decades of the second half of the nineteenth century, Pareto was able to generate a theory of how social and political authority changes hands due to a constant pressure to change leaders.

Another Italian social scientist, Guglielmo Ferrero, wrote extensively about power and authority in his historical works about the Roman Republic and the Empire. He also visited America at the invitation of President Theodore Roosevelt, and wrote a book called *Ancient Rome and Modern America*, in 1914. In this book, Ferrero said of America that "Every authority is losing power." According to Ferrero, everywhere he turned he saw a nation that, judged by European standards, defied tradition and authority. "The people discuss the government and the laws, children take the first opportunity of escaping from the authority of their parents, the younger generation is convinced that it knows more than the older, and values the latter's experience at zero" (Ferrero, 1914: p. 204). These observations led Ferrero to distinguish between authority in what he called defiant democracies and authority in those societies that he called deference democracies. Deference democracies favored legitimate authority, according to Ferrero, while defiant ones bring power to the fore since they question legitimacy incessantly. These theories of how authority circulates throughout the social order later influenced the thinking of American sociologists from Talcott Parsons to E. Digby Baltzell.

British Theorists

Interest in authority among British social scientists is exemplified not only by James Bryce, but also by Walter Bagehot and Harold Laski. As Perry Anderson astutely observed, there really was no effective effort to build a British sociology until after World War II (WWII). Even absent a full-blown sociology, British social scientists showed a lively interest in authority. The English statesman, political scientist, and sociologist James Bryce's study of democracy in the *American Commonwealth* and later in *Modern Democracies* exposed the give and take between leaders and citizens in the modern state, and his essay, 'Obedience,' is a remarkable dissection of why an authority effectively commands obedience on the basis of indolence, deference, sympathy, fear, and reason (Bryce, 1901). Indolence is again emphasized as a reason for obedience in Bryce's 1909 Yale Lectures, *The Hindrances to Good Citizenship*. Bryce's contemporary Walter Bagehot also took up the study of deference and obedience to authority in his well-known *English Constitution*, and comes very close to what Weber later called traditional authority. In his discussion of tradition and authority, Bagehot says, "custom is the first check on tyranny; that fixed routine of social life at which modern innovations chafe, and by which modern improvement is impeded, is the primitive check on base power" (Bagehot, 1867: p. 26). In other words, tradition embedded in custom acts to legitimate power, and it becomes authority. In *Authority in the Modern State*, another English social scientist, Harold Laski, wrote that the basis for obedience to authority is consent. In this he added his voice to many social scientists linking authority to freedom in the contemporary state. If one freely consents to obey authority, its legitimacy becomes more pervasive than if one were forced to obey by dint of raw power.

This brief overview about who among the early European social scientists wrote about authority shows that social science responded to the sociopolitical events dominating the continent. It is interesting that American social science took very little notable interest in authority until the post-WWII period. This American neglect of a major social science concept is neither an accident nor a quirk. It gives us an opportunity to see that social theory is at least in part a mirror for understanding the social and historical conditions of the times and places in which they are produced.

American Theorists

During the second half of the nineteenth century, European nations underwent births and rebirths of nationalism, bureaucratization, and centralization, and while these forces were also at work in America they did not materialize as focal points because politics was subsumed by civil society and treated like any other associational influence. American social scientists generally saw politics as just another part of the associational life. Authority was viewed as just part of the fabric of association in America. Except for E.A. Ross' 1901 book, *Social Control*, hardly any of the great founding generation of American social scientists wrote about authority, and it was not until well after World War I that authority showed up on the American social science horizon. Beginning in the 1930s, Robert MacIver began writing about authority, as did Harold Lasswell, Talcott Parsons, Pitirim Sorokin, and Carl Friedrich, among others who were acquainted with the European literature and began applying it to America. It is notable that these social scientists imported and modified German, French, and Italian theories of authority rather than creating genuinely American theories.

Weberian Concepts

As noted, the most influential discussion of authority is found in the work of Max Weber, one of the founders of modern sociology. It is rare for any social science discussion of authority to omit Weber's analysis of power, authority, and legitimacy. Weber was interested in forms of power, *Herrschaftstypen*, or types of domination. The term *Herrschaft*, which means rule or domination, has come to be popularly translated as authority, and sometimes as legitimate power. Indeed, Weber's studies showed that all societies established and maintained authority on the basis of one or another type of legitimacy. Although Weber understood that authority rarely if ever existed in its pure form in any social organization, he categorized these types of authority as traditional, rational-legal, or charismatic.

Traditional Authority

Traditional authority, or what Weber called *die Autorität des ewig Gestrigen*, the authority of the eternal yesterday (Weber, 1948: p. 78), rests on legitimacy conferred by custom. He says that traditional legitimacy justifies the use of power by sanctifying the mores by appeals to 'ancient recognition,' and to accepted practices. Some of the examples Weber uses to illustrate traditional authority are hereditary rulers, religious priesthoods, parents, and elders in extended families. Later American sociologists, such as Edward Shils and Robert Nisbet, further explored the bases for traditional authority extending the term to explain the overarching authority of the royal family in Great Britain, and the traditional authority of fathers, *patres familias*, regarding citizenship, and life and death authority over their extended families, in ancient Roman society (Shils and Young, 1952; Nisbet, 1999). The sociological theory of traditional authority always takes tradition, custom, hereditary or ascribed status, and social norms burnished over long periods of time into account as essential to this form of legitimacy in relatively small and homogeneous communities, with simple divisions of labor.

Rational-Legal Authority

With the transition from medieval to modern societies and states in Europe, there emerged a powerful alternative to traditional authority: rational-legal authority, or power legitimated by a rational-legal and interdependent division of labor, characterized by economic and occupational specialization, and complex rational-legal systems, where highly specialized and differentiated bureaucracies and associations are commonplace. Weber called this rational justification for obedience to authority, "domination by virtue of legality, by virtue of the belief in the validity of legal statute and functional competence, based on rationally created rules" (Weber, 1948: p. 79). According to Weber, this is the authority "exercised by the modern servant of the state." Aside from bureaucracies, rational-legal authority is situated in constitutions and courts that interest these, achieved offices, such as those held by presidents, governors, and senators, rather than hereditary and ascribed offices held by monarchs and aristocrats. Structural complexity, complex organizations, and constant pressures to change are endemic to the world of rational-legal authority.

Weber's German sociological colleague, Ferdinand Tönnies, distinguished between two types of societies: communally, and rurally oriented *gemeinschaft-like*, rooted in families or neighborhoods, and ascribed statuses, on the one hand, and *gesellschaft-like*, exemplified by associational, rather than communal orientations, and achieved statuses, on the other. Sociologists have clearly understood that *gemeinschaft* societies are marked by traditional authority structures, and *gesellschaft* societies by rational-legal authority.

A younger colleague to both Weber and Tönnies, Herman Schmalenbach, wrote about a third form of social organization, *des bundes*, communions, that transcended the *gemeinschaft-gesellschaft* continuum and its relationship to traditional and rational-legal legitimations of authority. He showed clearly that these ephemeral forms of association are based on what Weber called charismatic authority, the third leg of his famous tripartite theory of authority.

Charisma

Charisma, according to Weber, "is the authority of the extraordinary and personal gift of grace." Weber used this ecclesiastical term to denote "the absolutely personal devotion and the personal confidence in revelation, heroism, or other qualities of individual leadership" (Weber, 1948: p. 79). This form of authority is useful for analyzing small, very cohesive inner circles of political, communal, and religious organizations. It is the authority of the revolutionary religious prophet, the political zealot, radically innovative entrepreneurs, and vanguard artists. Unlike traditional or bureaucratic authorities, these are individuals that demand obedience from their followers not on the basis of either ascribed or achieved status, but on their direct and personal connection to mystical powers. Although many social theorists have used the concept of charismatic authority, none have really transcended what Weber said about this form of legitimacy as "a certain quality of

an individual personality by virtue of which he is set apart from ordinary men and treated as endowed with supernatural, superhuman, or at least specifically exceptional powers or qualities. These are … regarded as of divine origin or as exemplary, and on the basis of them the individual concerned is treated as a leader" (Schneiderman, 2003). Religious leaders from Moses, Abraham, Jesus, or the prophets are seen as charismatic authorities, but so were more recent leaders on the edge of sanity and the good and evil divide, such as John Brown, Jim Jones, or David Koresh. Edward Shils extended the Weberian analysis of charismatic authority to events, places, and things, and Weber's colleague, Joseph Schumpeter, went even farther in applying charismatic authority theories to economic entrepreneurs.

While Herman Schmalenbach, Edward Shils, and Joseph Schumpeter extended Weber's theory of charismatic authority beyond Tönnies' *Gemeinschft und Gesellschaft* continuum, a later sociologist, Robert Bierstedt, raised the question as to whether or not charisma is actually a type of authority at all. Arguing that "a leader can only request, and authority can require," Bierstedt made the case that charisma is merely a type of leadership, not a type of authority. By setting Weber's traditional and rational-legal legitimations of authority along a continuum of organizational statuses terms, Bierstedt demonstrated that authority exists because statuses that carry authority exist. But because charismatic legitimacy inheres in the person, rather than in roles and statuses, "charismatic authority is not authority at all, but leadership." While this interesting line of reasoning runs counter to Weber's theory of authority, and also to Edward Shils' extension of that theory beyond leadership to events and things, it also offers a possible obstacle to an ever expanding list of studies of charisma. In the end, Bierstedt's argument that the entire collective symbolism of the society stands behind the traditional as well as the rational-legal forms of authority, offering no alternatives to obedience, seems in keeping not only with Weber's theory, but also with the French sociologist Emile Durkheim's theory of collective representations. If charismatic leaders cannot bring collective sanctions to bear on disobedience, do they really have authority, as Weber understood that term? This question is still worth pondering, 50 years after Bierstedt first asked it.

Obedience and Authority

One important question raised by Bierstedt's discussion of the relationship between authority and leadership is why authority is obeyed. Whether it is the authority conferred by status in the society, or by followers in smaller bundlike groups, authority always demands obedience, and its success or failure depends upon followers obeying authoritative demands, so why is authority obeyed?

Legitimacy

Before trying to answer the question of obedience, a baseline assumption is usually made, namely, when authoritative demands are considered legitimate by those who thereby willingly obey them, one has authority. When the demands are obeyed mostly out of fear of force being used, one has raw power. Power is the ability to make people do what they do not want to do, but authority is the ability to make them want to do what they start off not wanting to do. This assumption is problematic because it valorizes one important part of Weber's theory of power at the expense of an equally important but contradictory element of it. The problem is centered on Weber's often ambivalent use of terms, such as *Herrschaft*, or domination, *Autorität*, or authority, and *Macht*, or power. Thus, in Weber's most comprehensive chapter on authority in his magnum opus, *Wirtschaft und Gesellschaft*, or Economy and Society (Weber, 1922, 1978), Weber says that "Herrschaft ist ein sonderfall von macht" (domination is a special case of power), and a few paragraphs later he discusses "domination by virtue of a constellation of interests" (die Herrschaft kraft Interessenkonstellation), as opposed to "domination by virtue of authority" (die Herrschaft kraft Autorität).

These ambiguities in Weber's language of domination and authority have been noted and discussed by a number of social theorists including Steven Lukes, Guenther Roth, and Wolfgang Mommsen. This inexactness of terminology is yet to be resolved, and until it is there is a certain cloudiness that hangs over the social science theories of authority resting on Weberian foundations. Weber himself seems to have understood that legitimate domination, or authority, as one has come to understand the term, is different than power insofar as it is accepted by those over whom obedience is demanded. It is this acceptance to which one must turn to understand obedience to authority.

This Weberian perspective – problematic as it may be – has been emphasized by Hannah Arendt, in a famous article, 'What Was Authority?' in which she also demonstrates that there is nothing to be gained by discussing some vague 'authority in general' (Arendt, 1968). As she notes, authority is always alloyed to the dialectic of obedience and disobedience in a particular and specific social structure, and it makes no sense to discuss authority in a free-floating and overly general way. For Arendt, as for Weber in his somewhat convoluted but comprehensive treatment of domination, authority must be understood "in contradistinction to both coercion by force and persuasion through arguments" (Arendt, 1968: p. 93). It is in this theoretical context that one must try to understand why one obeys authority.

Power

The ability of authority to transform resistance to obedience into acceptance of obedience is usually attributed to what Weber called legitimation or justification of power. This transformation is what the author calls a social and cultural 'transduction.' In physics, a transducer is a device that transforms one form of energy into another, as when an electromagnet transforms electrical energy into magnetic energy, or an incandescent bulb transforms electrical energy into light. Weber's forms of legitimacy are sociocultural devices that transform power into traditional authority, or rational-legal authority, or charismatic authority.

What Weber lacked was an understanding about how these transducing justifications tuned power into authority, or, put differently, how a reluctant obedience to power is turned into a willing obedience to authority. Weber was well versed in the

work of James Bryce, and was definitely influenced by him. As noted earlier, Bryce's now much neglected essay, 'Obedience,' had a demonstrable effect on Weber's theory of authority. In 'Obedience,' Bryce was one of the first social scientists to use the term 'legitimate authority' as an analytical tool, and he narrowed down the transformational and transducive qualities that turn power into authority to indolence, deference, sympathy, fear, and reason, all of which would later become important to Weber's understanding of legitimacy.

So why do people obey authority? Traditional values, religious values, rational values, and the like actually become part of who one is, part of his or her identities, and if they justify obedience to authority, reluctantly obeyed power becomes transformed into happily obeyed authority. To get to deeper and more modern answers to the question about sociocultural transduction of power into authority, one must turn to later theorists.

Culture

Philip Rieff, a sociologist, understood cultures to be engines of moral demand, or what he called moral demand systems. Rieff was not the only theorist to see culture as a system of moral demands. In this he was joined by the anthropologist Clifford Geertz who saw culture as a set of symbolic control mechanisms that govern behavior. Both Rieff's and Geertz's theories of culture benefit from Sigmund Freud's psychoanalytical idea that symbols become introjected into the self, and made part of an individual's sense of identity.

These cultural theories were anticipated decades ago by social scientists influenced by pragmatist philosophy, such as George Herbert Mead, whose ideas about 'significant symbols,' lays the groundwork for understanding the transducive quality of cultural values. To this list of theorists one should add the philosopher Susanne Langer, whose *Philosophy in a New Key* has chapters on 'life symbols,' as the root of myth and of sacrament that provide key insights into how cultural symbols transform reluctant obedience to power into willing obedience to authority. As she says, "the human brain is constantly carrying on a process of symbolic transformation of the experiential data that comes to it that causes it to become a veritable fountain of spontaneous ideas" (Langer, 1942: p. 34). It is the seeming spontaneity and freedom of our ideas about the need to obey authority that gives it legitimacy.

A very different point of view about the reasons people obey authority was expressed by the social psychologist, Stanley Milgram, whose experiments about obedience were both brilliant and controversial. Milgram found that a disturbingly high rate of experimental subjects were willing to obey an authority figure wearing a lab coat, and professing to be in charge of a learning experiment. The subjects were instructed to give what they believed were electrical shocks in increasing intensities to a 'learner' whenever he got an answer wrong. After finding that a majority of subjects obeyed the authority figure, often in spite of professing to have moral qualms about inflicting pain on the 'learner,' Milgram tried to explain their obedience, first in part because of the cultural context of their upbringings, and then because of the institutional setting of the experiments. But, what is really interesting about Milgram's explanations, in terms of obedience to authority, is that the subjects responded to the perceived legitimacy of the experimentor. Very much in line with Weber's theory, and that of the culture theorists such as Rieff and Geertz, Milgram says that the first and foremost reason for obedience is "the perception of a legitimate authority" (Milgram, 1974: p. 138). This perception of legitimacy is a direct result of "the internalization of the social order – that is internalizing the set of axioms by which social life is conducted." Thus, from Bryce to Rieff, Geertz, and Milgram, one sees that one must turn to cultural symbolics of obedience to understand the transducive qualities in the Weberian justifications of authority.

The most basic paradigm for conceptualizing authority was the one set into motion by Max Weber, and he and others addressed how obedience to authority comes to replace obedience to power. The intellectual landscape of authority reveals when, where, and how authority was conceived as a social science concept.

While Western philosophers from Plato and Aristotle through Hobbes, Locke, Rousseau, Machiavelli, Kant, Hegel, Kierkegaard, and Nietzsche, among others, have thought about power and authority, their discussions generally pivot on either justifying or condemning hierarchical civic or political relations. Social science theories of authority, on the other hand, tend to ask and answer the problem-solving questions what, why, when, how, where, and who. In other words, social science theorists of authority, such as Weber and the other social scientists discussed above, are less interested in justifying or condemning authority, than in understanding it.

See also: Bureaucracy, Sociology of; Charisma, Social Aspects of; Elites: Sociological Aspects; Legitimacy, Sociology of; Organizations, Sociology of; Social Influence, Sociology of; Work, Sociology of.

Bibliography

Anderson, P., July–August 1968. Components of the national culture. New Left Review. Number 50, 1–51.
Arendt, H., 1968. Between Past and Present. Viking, New York.
Arendt, H., 1970. On Violence. Harcourt, New York.
Bagehot, W., 1867. The English Constitution. Oxford, New York.
Baltzell, E.D., 1987. The Protestant Establishment. Yale, New Haven.
Baltzell, E.D., 1994. Judgment and Sensibility: Religion and Stratification. Transaction, New Brunswick.
Baltzell, E.D., 1999. The Protestant Establishment Revisited. Transaction, New Brunswick.
Bierstedt, R., 1974. Power and Progress. McGraw Hill, New York.
Bouglé, C., 1908. Essais Sur le Régime des Castes. Felix Alcan, Paris.
Bryce, J., 1888. The American Commonwealth. Macmillan, New York.
Bryce, J., 1901. Studies in History and Jurisprudence. Oxford, New York.
Bryce, J., 1921. Modern Democracies. Macmillan, New York.
Bryce, J., 1993. Hindrances to Good Citizenship. Transaction, New Brunswick.
Dumont, L., 1970. Homo Hierarchicus: The Caste System. University of Chicago, Chicago.
Ferrero, G., 1914. Ancient Rome and Modern America. Putnam, New York.
Ferrero, G., 1942. The Principles of Power. Putnam, New York.
Friedrich, C., 1972. Tradition and Authority. Oxford, New York.
Geertz, C., 1973. The Interpretation of Cultures. Basic, New York.
Harnack, A., 1961. History of Dogma. Dover, New York.
Horowitz, I., 1999. Behemoth. Transaction, New Brunswick.
Langer, S., 1942. Philosophy in a New Key. Harvard, Cambridge.

Laski, H., 1919. Authority in the Modern State. Yale, New Haven.
Lasswell, H., 1930. Psychopathology and Politics. University of Chicago, Chicago.
Lasswell, H., 1936. Politics: Who Gets What, When, and How? McGraw Hill, New York.
Lukes, S., 1978. Power and authority. In: Bottomore, T., Nisbet, R. (Eds.), A History of Sociological Analysis. Basic, New York.
MacIver, R., 1947. The Web of Government. MacMillan, New York.
McCraw, T., 2007. Prophet of Innovation: Joseph Schumpeter and Creative Destruction. Harvard, Cambridge.
Michels, R., 1915. Political Parties. Free Press, New York.
Milgram, S., 1974. Obedience to Authority. Harper, New York.
Mommsen, W., 1989. The Political and Social Theory of Max Weber. University of Chicago, Chicago.
Mosca, G., 1939. The Ruling Class. McGraw Hill, New York.
Nisbet, R., 1993. The Sociological Tradition. Transaction, New Brunswick.
Nisbet, R., 1999. Tradition and Revolt. Transaction, New Brunswick.
Pareto, V., 1935. Mind and Society. Harcourt, New York.
Parsons, T., 1937. The Structure of Social Action. McGraw Hill, New York.
Rieff, P., 1959. Freud: The Mind of the Moralist. Viking, New York.
Rieff, P., 1966. The Triumph of the Therapeutic. Harper, New York.
Ross, E.A., 2009. Social Control. Transaction, New Brunswick.
Roth, G., 1971. Scholarship and Partisanship. University of California, Berkeley.
Russell, B., 1960. Authority and the Individual. Beacon, New York.
Schneiderman, H., 2003. Charisma. Encyclopedia of Community, vol. 1. Sage, Thousand Oaks, CA, 126–128.
Schneiderman, H., 2009. The strange career of political sociology in America. Culture and Civilization 1 (1), 96–123.
Schneiderman, H., 2013. Protestantism and progress redux. In: Troeltsch, E. (Ed.), Protestantism and Progress.
Schumpter, J., 1989. Essays: On Entrepreneurs, and Innovations. Transaction, New Brunswick.
Schmalenbach, H., 1977. On Society and Experience. University of Chicago, Chicago.
Shils, E., 1980. The Calling of Sociology. University of Chicago, Chicago.
Shils, E., 1981. Tradition. University of Chicago, Chicago.
Shils, E., Young, M., 1952. The Meaning of the Coronation. Bobbs Merrill, Indianapolis.
Simmel, G., 1950. Sociology of Georg Simmel. Free Press, New York.
Sohm, R., 1892. Kirchenrecht. Duncker & Humblot, Leipzig.
Sohm, R., 1895. Outlines of Church History. MacMillan, London.
Sorokin, P., 1925. The Sociology of Revolution. Lippencott, Philadelphia.
Tarde, G., 1899. Les Transformations du Pouvoir. Felix Alcan, Paris.
Tarde, G., 1903. Laws of Imitation. Henry Holt, New York.
Tocqueville, A., 1998. The Old Regime and the Revolution. University of Chicago, Chicago.
Tönnies, F., 1963. Community and Society. Harper, New York.
Troeltsch, E., 1931. Social Teachings of the Christian Churches. Allen & Unwin, London.
Troeltsch, E., 2013. Protestantism and Progress. Transaction, New Brunswick.
Weber, M., 1922. Wirtschaft und Gesellschaft. C.B. Mohr, Tubingen.
Weber, M., 1948. From Max Weber: Essays in Sociology. Oxford, New York.
Weber, M., 1978. Economy and Society. University of California, Berkeley.

Autism Spectrum Disorders

Giacomo Vivanti, School of Psychological Science, Olga Tennison Autism Research Centre, La Trobe University, Melbourne, VIC, Australia
Erica Salomone, King's College London, Institute of Psychiatry, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Autism spectrum disorders are a family of conditions characterized by early emerging disruptions in social communication and restricted/repetitive behaviors. We will review recent research and current understanding on the evolution of the concept of autism, its clinical characteristics, its causes, as well as issues in clinical management, policy, and intervention. Given the relevance of autism for a scientific and cultural understanding of normative and atypical development of social behavior, research in autism has the potential to crucially inform knowledge in medical, biological, social, and behavioral sciences.

Evolution of the Concept of Autism (from Kanner to the DSM 5)

Autism is the diagnostic concept originally formulated by Leo Kanner (1943) to define a clinical syndrome characterized by early emerging social communication abnormalities and behavioral rigidity. Kanner's concept has proven influential and enduring, and interest in the nature and implications of this condition has grown exponentially over time. In the past decade, autism and related conditions have attracted increasing attention from a variety of different disciplines, including molecular biology, neuroscience, cognitive science, developmental psychology, pedagogy, psychiatry, philosophy of mind, linguistics, artificial intelligence, and anthropology.

Kanner's original description of 11 children affected by 'autistic disturbances of affective contact' emphasized a reduced or completely lacking interest in other people and a striking 'insistence on sameness' as the hallmarks of this condition. The validity of autism as a specific diagnostic concept has been supported in the following decades by various lines of research. In particular, studies documenting the early emergence of social disturbances, the presence of deficits affecting both language and nonverbal communication, and the discrepancy between social and nonsocial abilities in affected children pointed to a distinction between autism and other diagnostic concepts such as schizophrenia, language disorders, and intellectual disability (ID). Over the decades following Kanner's original report, the different definitions used to classify and diagnose autism have continued to be based on these two notions: impairments in social reciprocity and communication and insistence on sameness/resistance to change. The robustness of these diagnostic indicators is reflected in a report of the National Institutes of Health autism working group, written five decades after Kanner's description, which states that "the clinical diagnosis of autism is one of the most reliable diagnoses in psychiatric or developmental research" (Bristol et al., 1996). Despite the continuity between Kanner's original description and current definitions, the concept of autism has evolved over time as a consequence of cultural changes and research advances.

The most recent edition of the *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*, published by the American Psychiatric Association (2013), defines autism as a 'spectrum disorder,' rather than a single entity, with diagnostic criteria grouped in two categories: 'social communication deficits' and 'fixated interests and repetitive behaviors.' The notion of 'autistic spectrum' reflects the homogeneity in the core impairments, as well as the continuum of variability in the clinical presentation of these deficits. The current definition of autism as a spectrum represents a significant departure from previous approaches, which conceptualized autism and related conditions as a discrete number of mutually exclusive diagnostic categories. The International Classification of Diseases (ICD 10), published by the World Health Organization, is likely to incorporate these changes in the next edition. The concept of 'Asperger's disorder,' previously in use to define a 'subtype of autism' (DSM-IV), has become part of the mainstream vocabulary and is still culturally associated to autism, but is no longer relevant to diagnostic purposes. The terms autism and autism spectrum disorder (ASD) are now considered interchangeable.

Epidemiology

While the first studies in the 1960s indicated a prevalence of 4–5 in 10 000 for autistic disorder (as conceptualized in DSM-IV-TR), a recent review of epidemiological surveys conducted worldwide found a median prevalence of 19 in 10 000 for autistic disorder, which becomes 62 in 10 000 considering related conditions (Elsabbagh et al., 2012) with close correspondence in the estimates derived from North American and Northern European studies. Although the progressive rise in the prevalence in the last decades is irrefutable, the changes in the diagnostic criteria and the different methods used for case ascertainment influence the interpretation of the studies and do not allow to conclude that there is a correspondent increase in the incidence (or number of new cases per year) of the disorder. The factors presumably responsible for the increased prevalence include the progressive broadening of the diagnostic criteria (shown, for example, by a reexamination of the clinical records of a survey conducted in the 1980s using the case definition of DSM-IV-TR, Miller et al., 2013), the improved efficacy in the diagnostic procedures, and the increased awareness in health professionals as well as the changes in policies for access to services. In the United States, there is

evidence of ethnic and racial disparities in the identification of ASD (Mandell et al., 2009). The sex ratio (male:female) is 4.3:1, but there is suggestion that females are more likely to remain undiagnosed (Giarelli et al., 2010). Research on gender differences in the clinical phenotypes has mixed evidence, partly due to the fact that few female participants have been involved in research studies.

Clinical Features

Despite its biological roots, research has not identified yet a biological marker distinguishing ASD from other conditions. As a consequence, ASD can be diagnosed solely on the basis of behavioral manifestations. Moreover, autism is a developmental disorder; therefore, the presentations of symptoms are affected by and affect the developmental level of the individual (verbal ability and mental age), and the chronological age of the subjects influences the way in which these symptoms manifest themselves.

Social Communication and Social Interaction

Impairments in social communication and social interaction are a defining feature of autism (Carter et al., 2005); nonetheless, the presentation of symptoms in this area varies considerably across the life span. In infancy and early childhood, eye contact is often poor and the direct gaze may not indicate social contact with the carer. The ability to follow an adult's gaze or pointing is impaired and, while the awareness of other environmental sounds is intact, the response to human voices (for example, when the child's name is called) is delayed or lacking. Behaviors aimed at directing or maintaining the carer's attention to what the child is interested in – such as vocalizing, pointing to request or express interest, bringing objects to show them – occur with lower than expected frequency or are totally absent. When present, nonverbal communication (eye contact, gestures, facial expressions, body language) is poorly integrated with verbal or vocal communication. Social smiling may be lacking or occur only in response to prompts, and the behavioral cues that indicate the toddler's desire to show his/her enjoyment to the adult during joint activities may be very limited or absent. Turn-taking in the communicative exchanges with the carer in early infancy and the typical back-and-forth conversation after the emergence of language are impaired. Social play is very limited with failure to initiate or respond to social routines with an adult (such as peekaboo, patty-cake). Pretend and imaginative play, when present, may lack spontaneity or flexibility and are not shared with adults or peers. Relationships with peers may range from fewer attempts to approach other children to no interest at all and limited response or avoidance of their approaches. Response to others' emotions may be very limited or occur only in the presence of exaggerated expressions (e.g., adult pretending to cry). Delay or lack of development of spoken language may occur but is not a discriminating feature; however, it is highly indicative of ASD if not accompanied by attempts to compensate through the use of gestures. Abnormal speech patterns, odd words or phrases, echolalia, and unusual tone or pitch in vocalizations or words are other distinctive characteristics of early language in ASD and may persist throughout the life span. The use of communicative speech increases from childhood to adolescence and adulthood, but many aspects of communication, particularly the social components (such as pragmatics or prosody), are likely to remain impaired. Deficits in developing, maintaining, and understanding relationships (such as friendship) become more apparent in middle childhood, as the social demands increase. In adolescence and adulthood, for example, social interactions may increase in frequency but retain a qualitative impairment (stilted, exaggerated, or 'socially awkward' behavior).

Restricted, Repetitive Patterns of Behavior, Interests, or Activities

Repetitive and restricted behaviors (RRBs) are a heterogeneous group of symptoms that can emerge in infancy and persist throughout the life span; however, their manifestation may change with age. These behaviors are not a distinctive feature of ASD as similar phenomena are often observed in IDs, psychiatric disorders, neurological conditions and other clinical conditions with sensory deficits as well as in typically developing toddlers; however, ASD is usually characterized by a pattern of co-occurring behaviors. These symptoms include behaviors characterized by stereotyped or repetitive movements (such as motor stereotypies, repetitive use of objects) or repetitive use of speech (echolalia, idiosyncratic phrases) and restricted, fixated interests unusual in their intensity or focus (such as strong attachments to unusual objects, circumscribed or perseverative interests). The insistence on the sameness (such as the need to eat the same food everyday) and adherence to routines or rituals is another characteristic of ASD, often associated to high distress for the affected individual and impairment in everyday functioning. Seeking of specific sensations and unusual sensory interests (such as peering or smelling of objects) have been included in the DSM-5, together with hyper- or hyporeactivity to sensory inputs as important clinical descriptors of ASD. Within the autistic population, there is huge variability in the prevalence and severity of each of these types of behaviors. Behaviors like repetitive use of objects (spinning the wheels of a toy car) or motor stereotypies such as finger flicking or hand flapping are more likely to occur in children with ASD and ID than in children with ASD only, whereas circumscribed interests are more prevalent in higher-functioning individuals (Bishop et al., 2006). There is growing evidence that all kinds of RRBs decrease in frequency and severity with age, independently of co-occurring ID, with the exception of stereotyped movements that tend to be more persistent in individuals with ID (Esbensen et al., 2009).

Co-occurring Conditions

Once thought to be occurring in most cases of ASD, ID is now ascertained in approximately 50% of children diagnosed with ASD (CDC, 2012). In subjects with ASD, comorbidity with several psychiatric conditions (including anxiety, depression, hyperactivity and attention difficulties, oppositional defiant disorder, tics, and epilepsy) is increased compared to rates in the general population (Simonoff et al., 2013). The presence of mental health problems needs to be carefully assessed to

ensure appropriate support; however, the nature of the association between psychiatric conditions and ASD is still unknown.

Causes of Autism

There is incontrovertible evidence that genetic factors are the predominant cause of ASDs. Twin studies and family studies have documented that siblings of children with ASD have a 20- to 50-fold increased risk of having ASD compared to the general population, and the concordance rate in identical twins (who have 100% of genes in common) is up to 90% (Rutter, 2000). There is also evidence for a remarkable genetic heterogeneity in the etiology of ASD. A number of potential risk genes have been identified; however, none of them accounts for more than 1% ASD cases, and the combination of known genetic risk factors accounts for approximately 10% of ASD (Geschwind, 2006); in the remaining 90% of the cases, ASD etiology is unknown. It is currently estimated that many hundreds of genes might contribute to the risk of developing ASD. While genetic findings indicate that ASD is one of the most heritable neuropsychiatric disorders, the lack of complete concordance in identical twins indicates that nongenetic factors contribute to ASD as well. Several environmental factors have been found to be associated with increased risk of developing ASD, including prenatal exposure to rubella, thalidomide (a chemical compound that used to be prescribed to treat nausea and vomiting), and valproic acid. Recent research pointed to a possible association between ASD and exposure to chemicals in the environment (e.g., pesticides and air pollutants), but evidence is still inconclusive (Hertz-Picciotto et al., 2010). Advanced paternal age is also associated with an increased risk of ASD. Current evidence does not support the popular notion of a link between vaccines and risk for autism or the historical notion that poor parenting can determine autism. While it is clear that autism and related conditions result from a complex interplay of multiple genetic and environmental factors, the exact mechanisms that underlie both the homogeneity and the heterogeneity observed in ASD are largely unknown.

Several theoretical models of the neurobehavioral basis of ASD have been proposed. One line of research focuses on the possibility that a single neurocognitive process might explain the phenotypic presentation of autism. The many cognitive/neurocognitive explanations of ASD proposed within this framework share the view that different pathological pathways can ultimately result in one specific primary abnormality at the neurocognitive level, which, in turn, can determine the constellation of features (including strengths and weaknesses) that are observed in ASD. The primary deficits that have been hypothesized to underlie autistic features include difficulties in the ability to attribute mental states to others (i.e., theory of mind), a weak central coherence (i.e., a detail-focused processing style, whereby information is processed in terms of constituent parts, rather than global meanings), a deficit in executive functioning (i.e., difficulties in the flexible organization of behavior), and a deficit in affective connectedness/intersubjectivity (see Frith, 1996). More recently, the 'broken mirrors theory' (Ramachandran and Oberman, 2006) pointed to a primary deficit in the 'embodied' social understanding processes implemented by the mirror neuron system (a neural network that activates in response to both self-generated and observed actions and emotions). While each of these theories has some explanatory power and has generated important research, none of the proposed 'primary mechanisms' provides a satisfactory account for all the features observed in autism.

The current debate is focused on identifying the processes underlying specific features, such as social difficulties, rather than attempting to find a single explanation for all ASD symptoms. The complex information processing model (Minshew and Goldstein, 1998) posits that difficulties in handling information that is complex in nature affects the ability to process social stimuli (as they tend to be inherently more complex than nonsocial information). The social deficit models propose a specific difficulty in processing social versus nonsocial information, regardless of complexity level. Recent developmental models of social impairment suggest that a biologically based decrease in the salience or in the reward value attributed to social stimuli during early developmental stages virtually 'isolate' children with ASD from the social world, with downstream effects on social cognitive development (Chevallier et al., 2012).

Onset and Developmental Course

Historically, the onset of autism has been reported as occurring in one of two patterns. The 'early onset pattern' involves the early emergence of abnormalities affecting development from birth. In the 'regressive autism' pattern, on the contrary, children appear to be developing normally during early childhood, but then undergo a regression and lose previously acquired communicative and social skills. Recent research, however, has questioned the notion of a sharp distinction of 'early onset' versus 'regressive' autism (Ozonoff et al., 2008). In particular, a number of studies based on the prospective investigation of high risk samples indicate that at birth or shortly after birth children with autism are undistinguishable from those without autism (Yirmiya and Ozonoff, 2007). Abnormalities appear to emerge over the first 18 months of life, involving both a decline/regression in certain areas (e.g., social attention) and a failure to progress in other areas (e.g., language). Such delays and diminishments in typical behaviors occur at different points in development in different children, indicating a continuum, rather than a dichotomy, in the age of onset in autism.

Autistic symptoms are relatively stable, with most individuals who meet ASD diagnostic criteria in childhood continuing to meet criteria across the life span (for example, Howlin and Moss, 2012). However, the expression and the severity of the impairments seen in ASD can change over time. Studies on individual developmental trajectories point to a remarkable heterogeneity across functioning domains, which tend to increase with age. For example, a large cohort study (Anderson et al., 2007) documented that a subgroup of children with ASD made significant gains in communication, ultimately reaching 'normative' language level at 9 years of age, while another subgroup made little progress. Similar heterogeneity in developmental trajectories has been documented in the social domain, as well as in cognitive and adaptive functioning. There are also well-documented reports of children who have largely

or completely lost their symptoms of autism (Fein et al., 2013). Despite this remarkable variability in individual trajectories and outcomes, existing evidence suggests that the majority of adults in the autism spectrum, including those who had an average cognitive functioning at the age of diagnosis, experience very limited social inclusion, do not have a paid employment, and are not able to live independently.

The processes underlying these heterogeneous patterns of onset and developmental course of autistic symptoms are largely unknown. Existing research indicates that a complex interplay of biological and environmental risk and protective factors, rather than a single mechanism, might be at the origin of the diverse expression of clinical features in the ASD population.

Clinical Management

Diagnosis and Assessment

In most countries, diagnosis is based on the criteria specified in the DSM-5 or the ICD 10. Ascertainment of behavioral manifestations of ASD is achieved through a combination of direct observation and parent questionnaires focused on the child's developmental history. A number of standardized protocols such as the Autism Diagnostic Observation Schedule (Lord et al., 2006, 2012) and the Autism Diagnostic Interview-Revised (Le Couteur et al., 2003) have been developed to facilitate differential diagnosis between ASD and conditions that share superficial features (e.g., ID, attention deficit hyperactivity disorder, reactive attachment disorder, specific language impairments, and other communication disorders). To lower the age of diagnosis (often exceeding 4 years of age, CDC) and consequently facilitate early access to intervention programs and services, screening and developmental surveillance are recommended (Johnson and Myers, 2007). Further research is needed to establish whether a broad developmental screener is preferable to ASD-specific tools. Beyond the 'core' symptoms, ASD affects the individual's functioning such as cognitive, language, and adaptive skills in a variety of domains. A comprehensive developmental assessment delivered by a multidisciplinary team that considers these as well as other aspects (e.g., the presence of challenging behaviors) is therefore essential to identify strengths and weaknesses of each individual in order to build and monitor intervention programs tailored on the child's specific needs and resources. As the presence of co-occurring ID can affect the symptomatology, it is important to take into account the developmental level of the child in the clinical assessment in order to tease out the differential contribution of ASD and ID to the presentation of symptoms. Given the variability in the behavior of individuals with ASD across settings (e.g., the same child might show few social initiations in an unfamiliar context, but initiate often in a familiar context), the involvement of parents in the assessment as a source of information is critical.

Treatment

Currently there is no cure for autism, and the issue of treatment for such a heterogeneous condition whose etiology is uncertain is a highly debated topic. While the unethical and unsupported treatment approaches that were used until the 1970s (such as the institutionalization of children to remove them from their parents and the use of electric shocks to suppress abnormal behaviors) are no longer included in the current treatment guidelines, there is no general consensus on this subject. A number of self-advocacy organizations question the need of treating autism, on the basis of the notion that autism is a way of being in the world, or a culture, that should be supported and appreciated rather than treated. Others, while advocating for a shift from a medical model (which aims at eradicating/curing deviations from normality) to a social model (which focuses on environmental and social barriers to civil rights and inclusion), acknowledge the importance of interventions aimed at increasing opportunities for self-determination in individuals on the autism spectrum. Within the scientific community, there is a growing effort to promote rigorous research on treatment methods and outcomes within a social model that takes into account preferences and rights of individuals with ASD. Current approaches to treatment are mostly based on special education techniques. Different educational intervention programs vary in terms of the specific domains that are targeted (e.g., social skills versus self-management), the instructional techniques (e.g., directive teaching versus play-based approaches), and the overarching treatment goals (e.g., focus on ameliorating ASD symptoms versus focus on personal self-sufficiency). Some approaches use specific strategies to target a single domain or symptom (e.g., teaching how to communicate using alternative communication systems to compensate for language difficulties, or practicing social skills through adult-mediated social skills training groups). Other approaches involve comprehensive intervention programs, which target all areas of development. These programs usually involve the intensive engagement of the child in a planned educational intervention based on behavioral strategies, with operationalized instructional procedures, and a data collection system to facilitate progress and outcome measurement. Within this framework, different programs vary according to the specific curriculum, strategies, and underlying theoretical framework. For example, the Discreet Trial Training model (Lovaas, 1987) emphasizes a highly structured approach where predefined objectives are broken down in small, measurable units of behavior that are taught through systematic reinforcement, often using extrinsic rewards. In the Early Start Denver Model (Rogers and Dawson, 2010), instead, the focus is on the application of behavioral principles in joint play routines built on the child's spontaneous choices and motivation, with intervention goals organized around a behavioral framework (e.g., joint attention skills are taught before complex sentences, as joint attention precedes phrase speech in typical development). While positive effects have been documented for different focused and comprehensive programs, there is no evidence to indicate whether one treatment approach is more effective than others. Given the variability in learning profiles in ASD, it is likely that different children will benefit more from specific approaches according to their specific pretreatment characteristics; however, current knowledge on 'what works for whom' in autism is limited. Moreover, the degree of change that is attainable for individuals with ASD with autism, especially for those who are more severely affected, is not known. Initial claims of 'recovery' in response to behavioral treatment in half of the treated children have not

been substantiated by research. However, significant gains in cognitive, language, and adaptive skills appear to be possible for many individuals with ASD. Despite the differences in treatment approaches, a number of elements that are shared across most currently recommended programs appear to be particularly important for positive outcomes. These involve the enrollment of children in the intervention program as early as possible, the intensity of the program (at least 15–20 h per week, according to most guidelines), the use of systematic assessment procedures to define individualized sets of goals and monitor progress, and the involvement of parents.

Pharmacological treatment is sometimes used to target specific issues that can be associated with ASD, such as irritability and aggression; however, no medication appears to be effective in improving the core symptoms of autism.

A number of complementary/alternative medical treatments (i.e., treatments that are meant to be used alongside or as a substitute for standard treatments recommended by health professionals such as dietary interventions or supplements) are widely used, despite the lack of evidence for their efficacy.

Cultural Issues

The prevalence of autism is thought to be consistent across cultural groups, races, or countries and variations across sites disappear when accurate case identification methods are followed (Kim et al., 2011). Indirectly, the presence of autism parent organizations and advocacy groups across all continents suggests the universality of the constellation of behaviors originally described as autism by Leo Kanner. Nonetheless, the diagnostic classification manuals of choice, the availability of translated and validated diagnostic tools, as well as the beliefs hold by professionals, and the stereotypes attached to autism, which might influence symptom reporting by parents, vary across countries and play a major role in the diagnostic process. Moreover, the context where symptoms become apparent may be different in different cultural groups: for example, the ascertainment of qualitative impairments in communication should not overlook variations in language exposure, societal expectations, and communicative styles in parents (Grinker et al., 2011).

However, while it has been established that autism is a global public health concern, the vast majority of the research on the clinical presentation and neuropsychological profile of individuals with ASD has been carried out in Western countries. Therefore, it is not clear whether ASD features are affected by cultural differences, or whether ASD represents a culturally universal phenotype. This is particularly relevant with regard to neuropsychological and behavioral factors that are subject to cultural variations, such as, for example, the ability to detect items embedded within a more complex figure (known to be superior in Western versus Eastern cultures; Nisbett and Miyamoto, 2005) or the tendency to solve problems using experience-based versus rule-based strategies (known to be more pronounced in Eastern cultures; Norenzayan et al., 2002). A recent cross-cultural study (Koh and Milne, 2012) documented that, consistent with previous literature, British children with ASD outperformed their typically developing counterparts in tasks involving detection of shapes embedded in a complex figure – however, this was not the case in participants with ASD from Singapore, suggesting an impact of cultural variations in perceptual style of children with ASD. However, more cross-cultural research is needed to determine the degree to which cultural differences impact on symptom expression and neuropsychological functioning in ASD.

Rights/Policies/Advocacy Issues

Alongside the first advocacy campaigns of parents of children with disabilities to ensure free and appropriate public education (such as the Individuals with Disability Education Act in the United States), advocacy in the field of autism has historically focused on education and access to healthcare for children. In the United States, organizations such as the Autism Society of America and Autism Speaks have battled to gain autism insurance legislation to ensure access to services, and such legislation is now adopted or being considered by the majority of the states. In 2003, the Council of Europe has admitted the complaint presented by the umbrella organization Autism-Europe and recognized that the situation of the education provision in France constituted a violation of the right of persons with disabilities to independence, social integration and participation in the life of the community, as well as the right of children and young persons to social, legal, and economic protection (Articles 15 and 17 of the Social Charter). As there is a growing population of adults with autism (who would have went undiagnosed or misdiagnosed in the past), the main advocacy organizations as well as self-advocates are increasingly including in their agenda issues such as transition into adulthood, employment and housing, services for adults, and so on. Funding to services and research in the causes and treatment of autism has seen an unprecedented rise in the past decades, largely thanks to the efforts of these organizations. It is important to note though, that huge difference exists across nations: a recent report (Pellicano et al., 2013) has shown that the average research spend per autistic individual is 18 times higher in the United States than in the United Kingdom, and although no comparison data are available for the rest of the world, this gap is likely to be much wider.

Current Issues and Research Directions in Autism Research

Despite the tremendous increase in research in the last 20 years, autism remains an enigma. The complex phenomenology of autism raises a multitude of questions on the psychological and biological nature of social and emotional connectedness, language, the understanding of self and others' minds, as well as issues on appreciation of diversity, human rights and self-determination, stimulating empirical research and debate across a variety of disciplines. As it has become clear that this disorder does not result from a single dysfunction or pathological pathway, current research is focusing on understanding the interplay of biological, neuropsychological, and environmental factors. One critical issue that is still poorly understood is the interrelation between the seemingly unrelated symptoms of autism and their relation to the other features that are

present in some but not all individuals on the spectrum (e.g., epilepsy, special talents). Research on factors underlying clinical heterogeneity should also inform research on the variability in treatment response in this population, with the potential to identify profiles of responders and nonresponders to available treatment programs. The nature of ID in ASD is another important issue: why do many individuals with ASD (50%) have a cognitive delay, while some have an above average IQ? What is the link between symptoms of autism and cognitive development? Current research is addressing these issues using a variety of approaches, but answers to these questions remain elusive.

Advances in technology in the last decade (e.g., neuroimaging techniques) have offered unprecedented opportunities to investigate brain functioning in ASD and brought to promising research leads, including evidence for atypical functioning of the 'social brain network,' abnormalities in the mirror neuron system activity, as well as evidence for disturbances in neural connectivity (e.g., the synchronization between different brain regions). However, findings in this area are often inconsistent, and research has thus far failed to identify neural correlates of ASD that are universal, specific, and unique to this condition.

A full understanding of the processes underlying ASD symptoms is made difficult by the fact that the processes underlying typical development of social–emotional reciprocity and their role in cognitive, language, and personality development are poorly understood. Therefore, research advances in the biology and neuropsychology of typical social and emotional development have the potential to facilitate the understanding of autism. Conversely, research in autism has the potential to crucially inform knowledge in medical, biological, social, and behavioral sciences.

See also: Diagnostic Processes in Clinical Psychology and Psychotherapy; Disability: Social and Psychological Aspects; Goal-Oriented Behaviour Among Human and Non-Human Primates; Intellectual Functioning, Assessment of; Mirror Neurons, Theory of; Neural Correlates of Social Development; Theory of Mind During Infancy and Early Childhood Across Cultures, Development of.

Bibliography

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Publishing, Arlington, VA.

Anderson, D.K., Lord, C., Risi, S., Shulman, C., Welch, K., DiLavore, P.S., Pickles, A., 2007. Patterns of growth in verbal abilities among children with autism spectrum disorder. Journal of Consulting and Clinical Psychology 75 (4), 594–604. http://dx.doi.org/10.1037/0022-006x.75.4.594.

Bishop, S.L., Richler, J., Lord, C., 2006. Association between restricted and repetitive behaviors and nonverbal IQ in children with autism spectrum disorders. Child Neuropsychology 12 (4–5), 247–267. http://dx.doi.org/10.1080/09297040600630288.

Bristol, M.M., Cohen, D.J., Costello, E.J., Denckla, M., Eckberg, T.J., Kallen, R., Spence, M.A., 1996. State of the science in autism: report to the National Institutes of Health. Journal of Autism and Developmental Disorders 26 (2), 121–154. http://dx.doi.org/10.1007/Bf02172002.

Carter, A., Ornstein Davis, N., Klin, A., Volkmar, F., 2005. Social development in autism. In: Volkmar, F., Rhea, P., Klin, A., Cohen, D. (Eds.), Handbook of Autism and Pervasive Developmental Disorders, third ed., vol. 1. John Wiley & Sons, Hoboken, NJ, pp. 312–334.

CDC, 2012. Prevalence of Autism Spectrum Disorders – Autism and Developmental Disabilities Monitoring Network. 14 Sites Vol. 61, 1–19 (SS03). MMWR, United States.

Chevallier, C., Kohls, G., Troiani, V., Brodkin, E.S., Schultz, R.T., 2012. The social motivation theory of autism. Trends in Cognitive Sciences 16 (4), 231–239. http://dx.doi.org/10.1016/j.tics.2012.02.007.

Dawson, G., 2010. Recent advances in research on early detection, causes, biology, and treatment of autism spectrum disorders. Current Opinion in Neurology 23 (2), 95–96. http://dx.doi.org/10.1097/Wco.0b013e3283377644.

Elsabbagh, M., Divan, G., Koh, Y.-J., Kim, Y.S., Kauchali, S., Marcín, C., Fombonne, E., 2012. Global prevalence of autism and other pervasive developmental disorders. Autism Research 5 (3), 160–179. http://dx.doi.org/10.1002/aur.239.

Esbensen, A., Seltzer, M., Lam, K.L., Bodfish, J., 2009. Age-related differences in restricted repetitive behaviors in autism spectrum disorders. Journal of Autism and Developmental Disorders 39 (1), 57–66. http://dx.doi.org/10.1007/s10803-008-0599-x.

Evans, D.W., Leckman, J.F., Carter, A., Reznick, J.S., Henshaw, D., King, R.A., Pauls, D., 1997. Ritual, habit, and perfectionism: the prevalence and development of compulsive-like behavior in normal young children. Child Development 68 (1), 58–68. http://dx.doi.org/10.1111/j.1467-8624.1997.tb01925.x.

Fein, D., Barton, M., Eigsti, I.M., Kelley, E., Naigles, L., Schultz, R.T., Tyson, K., 2013. Optimal outcome in individuals with a history of autism. Journal of Child Psychology and Psychiatry 54 (2), 195–205. http://dx.doi.org/10.1111/Jcpp.12037.

Frith, U., 1996. Cognitive explanations of autism. Acta Paediatrica. Supplement 416, 63–68.

Gerber, J.S., Offit, P.A., 2009. Vaccines and autism: a tale of shifting hypotheses. Clinical Infectious Diseases 48 (4), 456–461. http://dx.doi.org/10.1086/596476.

Geschwind, D., 2006. Autism genetics: from gene to brain to behavior. American Journal of Medical Genetics Part B-neuropsychiatric Genetics 141B (7), 687.

Giarelli, E., Wiggins, L.D., Rice, C.E., Levy, S.E., Kirby, R.S., Pinto-Martin, J., Mandell, D., 2010. Sex differences in the evaluation and diagnosis of autism spectrum disorders among children. Disability Health Journal 3 (2), 107–116. http://dx.doi.org/10.1016/j.dhjo.2009.07.001.

Gotham, K., Bishop, S.L., Lord, C., 2011. Diagnosis of autism spectrum disorders. In: Amaral, D.G., Dawson, G., Geshwind, D. (Eds.), Autism Spectrum Disorders. Oxford University Press, Oxford, pp. 30–43.

Grinker, R.R., Yeargin-Allsopp, M., Boyle, C., 2011. Culture and autism spectrum disorders: the impact on prevalence and recognition. In: Amaral, D., Dawson, G., Geschwind, D. (Eds.), Autism Spectrum Disorders. Oxford University Press, Oxford, pp. 112–136.

Hertz-Picciotto, I., Delwiche, L., Hansen, R., Pessah, I.N., 2010. Household pesticide use in relation to autism. American Journal of Epidemiology 171, S100.

Howlin, P., Moss, P., 2012. Adults with autism spectrum disorders. Canadian Journal of Psychiatry-revue Canadienne De Psychiatrie 57 (5), 275–283.

Johnson, C., Myers, S., 2007. Identification and evaluation of children with autism spectrum disorders. Pediatrics 120 (5), 1183–1216.

Kanner, L., 1943. Autistic disturbances of affective contact. Nervous Child 2, 217–253.

Kim, Y.S., Leventhal, B.L., Koh, Y.-J., Fombonne, E., Laska, E., Lim, E.-C., Cheon, K.-A., Kim, S.-J., Lee, H.K., Son, D.-H., Grinker, R.R., 2011. Prevalence of autism spectrum disorders in a total population sample. The American Journal of Psychiatry 168, 904–912. http://dx.doi.org/10.1176/appi.ajp.2011.10101532.

Koh, H.C., Milne, E., 2012. Evidence for a cultural influence on field-independence in autism spectrum disorder. Journal of Autism and Developmental Disorders 42, 181–190.

Le Couteur, A., Lord, C., Rutter, M., 2003. The Autism Diagnostic Interview -Revised (ADI-R). Western Psychological Services, Los Angeles, CA.

Lewis, M.H., Bodfish, J.W., 1998. Repetitive behavior disorders in autism. Mental Retardation and Developmental Disabilities Research Reviews 4 (2), 80–89. http://dx.doi.org/10.1002/(sici)1098-2779(1998)4:2<80::aid-mrdd4>3.0.co;2–0.

Lord, C., Risi, S., DiLavore, P.S., Shulman, C., Thurm, A., Pickles, A., 2006. Autism from 2 to 9 years of age. Archives of General Psychiatry 63 (6), 694–701. http://dx.doi.org/10.1001/archpsyc.63.6.694.

Lord, C., Rutter, M., DiLavore, P.C., Risi, S., Gotham, K., Bishop, S., 2012. Autism Diagnostic Observation Schedule, Second Edition (ADOS-2). Western Psychological Services, Torrance, CA.

Lovaas, O.I., 1987. Behavioral treatment and Normal educational and intellectual-functioning in young autistic-children. Journal of Consulting and Clinical Psychology 55 (1), 3–9. http://dx.doi.org/10.1037/0022-006x.55.1.3.

Mandell, D.S., Wiggins, L.D., Carpenter, L.A., Daniels, J., DiGuiseppi, C., Durkin, M.S., Giarelli, E., Morrier, M.J., Nicholas, J.S., Pinto-Martin, J.A., Shattuck, P.T., Thomas, K.C., Yeargin-Allsopp, M., Kirby, R.S., 2009. Racial/ethnic disparities in the identification of children with autism spectrum disorders. American Journal of Public Health 99 (3), 493–498. http://dx.doi.org/10.2105/ajph.2007.131243.

Miller, J., Bilder, D., Farley, M., Coon, H., Pinborough-Zimmerman, J., Jenson, W., Rice, C.E., Fombonne, E., Pingree, C.B., Ritvo, E., Ritvo, R.A., McMahon, W., 2013. Autism spectrum disorder reclassified: a second look at the 1980s Utah/UCLA autism epidemiologic study. Journal of Autism and Developmental Disorders 43 (1), 200–210. http://dx.doi.org/10.1007/s10803-012-1566-0.

Minshew, N.J., Goldstein, G., 1998. Autism as a disorder of complex information processing. Mental Retardation and Developmental Disabilities Research Reviews 4 (2), 129–136. http://dx.doi.org/10.1002/(sici)1098-2779(1998)4:2<129::aid-mrdd10>3.0.co;2-x.

Nisbett, R.E., Miyamoto, Y., 2005. The influence of culture: holistic versus analytic perception. Trends in Cognitive Sciences 9, 467–473.

Norenzayan, A., Smith, E.E., Kim, B.J., Nisbett, R.E., 2002. Cultural preferences for formal versus intuitive reasoning. Cognitive Science 26, 653–684.

Ozonoff, S., Heung, K., Byrd, R., Hansen, R., Hertz-Picciotto, I., 2008. The onset of autism: patterns of symptom emergence in the first years of life. Autism Research 1 (6), 320–328. http://dx.doi.org/10.1002/Aur.53.

Pellicano, E., Dinsmore, A., Charman, T., 2013. A Future Made Together: Shaping Autism Research in the UK. Institute of Education, London.

Ramachandran, V.S., Oberman, L.M., 2006. Broken mirrors: a theory of autism. Scientific American 295 (5), 62–69.

Rogers, S.J., Dawson, G., 2010. Early Start Denver Model for Young Children with Autism: Promoting Language, Learning, and Engagement. Guilford Press, New York.

Rutter, M., 2000. Genetic studies of autism: from the 1970s into the millennium. Journal of Abnormal Child Psychology 28 (1), 3–14. http://dx.doi.org/10.1023/A:1005113900068.

Schreibman, L.E., 2005. The Science and Fiction of Autism. Harvard University Press, Cambridge, MA.

Simonoff, E., Jones, C.R.G., Baird, G., Pickles, A., Happé, F., Charman, T., 2013. The persistence and stability of psychiatric problems in adolescents with autism spectrum disorders. Journal of Child Psychology and Psychiatry 54 (2), 186–194. http://dx.doi.org/10.1111/j.1469-7610.2012.02606.x.

Snow, A., Lecavalier, L., 2011. Comparing autism, PDD-NOS, and other developmental disabilities on parent-reported behavior problems: little evidence for ASD subtype validity. Journal of Autism and Developmental Disorders 41 (3), 302–310. http://dx.doi.org/10.1007/s10803-010-1054-3.

Totsika, V., Hastings, R.P., Emerson, E., Lancaster, G.A., Berridge, D.M., 2011. A population-based investigation of behavioural and emotional problems and maternal mental health: associations with autism spectrum disorder and intellectual disability. Journal of Child Psychology and Psychiatry 52 (1), 91–99. http://dx.doi.org/10.1111/j.1469-7610.2010.02295.x.

Williams, D.L., Goldstein, G., Minshew, N.J., 2006. Neuropsychologic functioning in children with autism: further evidence for disordered complex information-processing. Child Neuropsychology 12 (4–5), 279–298. http://dx.doi.org/10.1080/09297040600681190.

Yirmiya, N., Ozonoff, S., 2007. The very early autism phenotype. Journal of Autism and Developmental Disorders 37 (1), 1–11. http://dx.doi.org/10.1007/s10803-006-0329-1.

Autobiographical Memory

Daniel L Greenberg, Lauren L Deasy, and Amelia L Zasadski, College of Charleston, Charleston, SC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.A. Conway, volume 14, pp. 9563–9567, © 2001, Elsevier Ltd.

Abstract

When we use the word 'memory' in everyday conversation, we are usually referring to a memory of a specific event, whether it be yesterday's lunch meeting, last year's graduation ceremony, or the snowstorm that blanketed our town when we were young. At times, we might recall exceptionally powerful memories, such as the 11 September attacks or a robbery that happened right in front of our eyes. We might reminisce on happy events or ruminate on sad ones, and we might find ourselves lost in memories of our youth. All these are examples of autobiographical memory – in other words, memories of the events of one's life. More than a century of research has led to fascinating discoveries about the distribution of memories across the life span and the factors that affect the vividness and intensity of recall.

What Is Autobiographical Memory?

The study of autobiographical memory is broad in scope, and researchers in this field often investigate topics such as flashbulb memory, eyewitness testimony, the development of autobiographical memory in children, and the loss of autobiographical memory as a result of neurological disorder. This encyclopedia already contains articles on all of these topics, so we will not revisit them here. Instead, we will concentrate on basic research in autobiographical memory, with brief detours to touch on otherwise-uncovered topics such as autobiographical recollection in people with mood disorders.

First, however, we should define what is meant by 'autobiographical memory.' As is common in psychological research, different researchers define the term in different ways. According to many researchers, a defining feature of autobiographical memory is *recollection* – that is, the sense that one is reliving or reexperiencing the original event (see Brewer, 1996, for discussion). This might make it sound as though autobiographical memory is a kind of episodic memory, and some researchers do define the term that way (Rubin, 1998, 2006).

Other researchers draw a distinction between *autobiographical event memory* and *autobiographical facts* or *personal semantic memories*, such as one's childhood address (Kopelman et al., 1989; Levine et al., 2002). Memories like these can be retrieved without having a specific event in mind, and thus they do not seem to qualify as episodic. They seem more like a form of semantic memory, but they differ in one key way. Semantic memory is fairly consistent across people (at least within a language and a culture), but personal semantic memories are not: most of the people reading this article will have a similar idea of what a house is, but your childhood address is almost certainly different from mine. Thus, personal semantic facts have an autobiographical aspect to them even though they are not episodic.

Conway and colleagues (e.g., Conway, 2001; Conway and Pleydell-Pearce, 2000) have developed a hierarchical model that incorporates several different kinds of autobiographical information. *Lifetime periods* are distinct periods with beginnings and endings – for instance, "when I was in school," "when I worked for *x*," or "when I was dating *y*." *General events* are more specific than lifetime periods; they include experiences that span a relatively long period (a family vacation in 1997) as well as repeated events (for instance, Tuesday lunches with my best friend; Neisser's (1981) term "repisodic memory" covers a similar idea). *Event-specific knowledge* consists of vivid and detailed sensory information, often visual, that is drawn from individual events and is thus episodic in nature. In Conway's view, all these levels are autobiographical, and retrieval can proceed up and down this hierarchy as appropriate to generate a rich autobiographical memory that is grounded in a person's self-knowledge.

Nonepisodic autobiographical memories obviously have everyday importance – we all need to be able to remember our home address, for instance. They are also relevant in clinical situations: people with semantic dementia, for instance, are often unable to retrieve event-specific knowledge, even though they can retrieve general events (e.g., Westmacott et al., 2001). For the most part, however, research has focused on episodic autobiographical memories – memories of events – and we will focus on them for the remainder of the article.

A Classic Example

The archetypical autobiographical memory comes from Marcel Proust's (1922/1934) *A la recherche du temps perdu*, variously translated as *Remembrance of Things Past* or *In Search of Lost Time*. When the narrator takes a bite of *petite madeleine* soaked in tea, he is struck with a sense of inexplicable joy, and then, a few bites later, he is transported into the past:

> [S]uddenly the memory returns. The taste was that of the little crumb of madeleine which on Sunday mornings at Combray…my aunt Léonie used to give me, dipping it first in her own cup of real or of lime-flower tea…[I]mmediately the old grey house upon the street, where her room was, rose up like the scenery of a theatre to attach itself to the little pavilion, opening on to the garden, which had been built out behind it for my parents…and with the house the

> town, from morning to night and in all weathers, the Square where I was sent before luncheon, the streets along which I used to run errands, the country roads we took when it was fine….[S]o in that moment all the flowers in our garden and in M. Swann's park, and the water-lilies on the Vivonne and the good folk of the village and their little dwellings and the parish church and the whole of Combray and of its surroundings, taking their proper shapes and growing solid, sprang into being, town and gardens alike, all from my cup of tea. (Proust, 1922/1934: p. 36)

Psychologists like to cite this memory because it illustrates many of the key features of autobiographical recall. The first few steps come involuntarily, triggered by the taste of the petite madeleine. The full memory does not spring to mind at once; it slowly and deliberately unfolds into a story. The narrator feels transported back in time, he sees a series of vivid visual images, he remembers the setting, and then he feels the emotions he felt then. In short, he is *recollecting* his days at Combray – mentally reliving them as though he were experiencing them once more.

Measuring Autobiographical Memory

Before discussing these features in more detail, we should describe how researchers measure autobiographical memory. This is not as simple as it might seem, at least if one wants to measure it in a rigorous and reliable way.

Ever since Ebbinghaus, most memory experiments have followed the same general pattern. Researchers compile a set of stimuli, such as a list of nouns or a series of images. These stimuli are carefully selected so that they have certain desirable properties. In verbal learning experiments, for instance, the researchers might select words with a particular degree of frequency, imageability, and length. They can then administer each word to participants for a precise amount of time, repeat it as often as they like, and measure retention at set intervals. Much of the experiment can be timed to the millisecond, if needed, allowing for a great degree of experimental control.

For the most part, those of us who study autobiographical memory are unable to do any of these. We generally have no control over the original experience, and, in most cases, we will be asking about different experiences from each participant. We cannot be certain how often the participant rehearsed the experiment (whether by thinking about it, talking about it with others, or writing it down in a diary), and we have only the participants' best estimate of when the remembered event occurred. In fact, for the most part, we have no way to know whether the remembered event actually took place at all; for all we know, the participant could be misremembering, confabulating, or just plain making something up.

There are a few exceptions. In diary studies (e.g., Barclay and Wellman, 1986) and studies of flashbulb memory (e.g., Talarico and Rubin, 2003), we have more insight into some aspects of the original events. In some cases, participants are placed in contrived situations (St. Jacques and Schacter, 2013), as is common in studies of eyewitness testimony (Loftus and Palmer, 1974). In clinical situations, investigators sometimes try to have memories verified by a spouse, but this is problematic as well: Spouses do not always agree even in the best of circumstances, and when two people disagree, one can never quite be sure which (if either) is actually correct.

These challenges have led some authors to suggest that carefully designed laboratory experiments are a better way to investigate memory. Others feel that the study of real-world phenomena is critical, no matter how messy it might be. This debate reflects the long-standing tension between naturalistic, ecologically valid psychology and traditional laboratory-based psychology (see Neisser, 1978; Banaji and Crowder, 1989, for discussion.) Wherever one stands on this issue (and it seems to be largely a matter of taste), it is clear that autobiographical memory research must use research methods that are different from the usual verbal learning paradigms.

Thus most measures of autobiographical memory do not focus on accuracy, but rather on aspects such as phenomenological properties, the efficacy of certain cues, or (in the case of amnesia) whether the individual can remember anything at all. In the Galton–Crovitz paradigm, for instance (Galton, 1879; Crovitz and Schiffman, 1974), participants are presented with a series of cue words such as *tree* or *candy* and asked to come up with a memory associated with each. The Autobiographical Memory Test (Williams and Broadbent, 1986) takes this approach a step further by using emotionally valenced words to elicit memories. Researchers can use these tasks to assess how many memories people recall from a particular period and how detailed those memories seem to be; they can also assess how different kinds of people respond to positive and negative cues (see the section on depressive disorders and memory for more details). They can add on a questionnaire that asks people to rate the phenomenological properties of the memory (Rubin et al., 2003; Johnson et al., 1988), allowing them to explore the relations between these properties.

The Galton–Crovitz test is straightforward and easy to use, but as several authors have noted, it is somewhat artificial (Kopelman et al., 1989). Furthermore, it provides very little in the way of retrieval support. The link between cue word and memory is not always obvious, so participants must rely on executive functioning to search for and locate an appropriate memory. Consequently, poor performance on the task might actually reflect a deficit of executive function rather than poor autobiographical memory.

The Autobiographical Memory Interview (Kopelman et al., 1989) tries to address this problem by taking a more structured approach. It divides the participant's life span into three periods – childhood, early adulthood, and recent – and asks people to come up with autobiographical events and personal semantic memories from each. The test manual provides a detailed scoring method, and the results can be compared to established norms. In addition, the manual specifies the precise memories the participant is to give (for instance, a wedding). In this way, it provides a greater degree of retrieval support, presumably reducing the demand on executive function. At the same time, these constraints are not always desirable: If one wants to investigate which of a participant's memories are most vivid, for instance, this test would not be as appropriate.

The Autobiographical Interview (Levine et al., 2002) attempts to balance these two approaches by varying the

amount of retrieval support. Participants are asked to generate a memory from each of five time periods. For each memory, participants are first asked to speak about each event freely, that is, they are provided with very little retrieval support. Next, the experimenter provides a *general probe* (for instance, by asking whether the participant can remember anything else). Finally, the participants are given a semistructured interview that asks about the specifics of the event, the time and place at which it occurred, and various other details. The resulting memories can be scored both for episodic and semantic components, and one can gain some insight into a particular individual's need for retrieval support. As always, there is a price to be paid for this approach: only five memories are obtained, and even so the test can be quite long.

Each of these tests has been used to discover interesting and important results, and each has its advantages and disadvantages. As with much of psychology, 'it depends on how you measure it': the results you get can vary with the measure you select. Ultimately, different tests are appropriate for different circumstances, and the choice of test will depend on what exactly a researcher is trying to measure.

Aspects of Autobiographical Memory

Distribution of Memories across the Life Span

Proust's vivid memory illustrates another aspect of autobiographical memory, namely, the persistence and vividness of memories from youth. For most forms of memory, the retention curve takes the form of a power function: a substantial amount of information is lost upfront, but a core of information lasts much longer (Rubin and Wenzel, 1996). As **Figure 1** shows, however, the retention function for autobiographical memory takes a different form. The data were generated using the memories of 70-year-olds. The *x*-axis shows the decade in which the event occurred, and the *y*-axis shows the percentage of memories from that decade, excluding the last year.

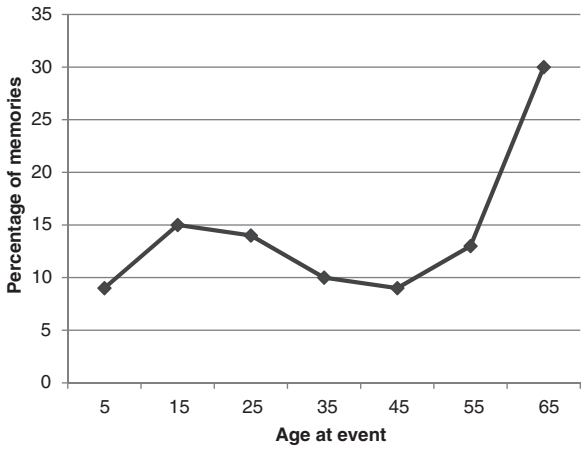

Figure 1 Typical distribution of memories across the life span. Reproduced from Rubin, D.C., Schulkind, M.D., 1997. The distribution of autobiographical memories across the lifespan. Memory and Cognition 25, 859–886.

The right side of the curve shows the normal retention function, and on the left side, there are few, if any, memories from before the age of 3 (this is called childhood amnesia and is covered in more detail in Childhood Memories). Something unusual is happening in the middle, however: people recall a disproportionate recall of events from ages 15 to 25.

This phenomenon is called the reminiscence bump, and it is one of the most reliable phenomena in autobiographical memory research (Rubin et al., 1986). The precise location of the peak varies a bit from study to study, but is always somewhere in this period. It occurs when participants are asked about vivid memories (Fitzgerald, 1988), important memories (Fitzgerald, 1996), olfactory-cued memories (Rubin et al., 1984), involuntary memories (Berntsen and Rubin, 2002), and happy memories (Berntsen and Rubin, 2002). It occurs when people are asked for their life stories (Fromholt and Larsen, 1991) or which memories should go into a book about their life (Fitzgerald, 1996). It does not seem to occur for negative memories, however (Berntsen and Rubin, 2002; Rubin and Berntsen, 2003), a point to which we will return.

Why does the reminiscence bump occur? One straightforward explanation involves neurocognition: neural and cognitive function is thought to be at the peak during this period, so the bump might reflect that encoding is optimal at this point in development. Another explanation involves novelty: this period of life involves many firsts (first romantic relationship, first real full-time job, first time living independently, etc.), and novel information tends to be more memorable (Rubin et al., 1998). Other explanations focus on social or developmental factors. Fitzgerald (1996), for instance, has attributed the reminiscence bump to the development of identity in adolescence and young adulthood. Since events during this period are particularly important for identity function, it makes sense that they would be disproportionately included in the story of one's life. Similarly, Conway and colleagues (Conway and Haque, 1999; Rathbone et al., 2008) attribute the bump to the formation of highly self-relevant memories. In support of several of these views, a few studies have reported a second bump for momentous or transformative events, such as immigration to a new country (Schrauf and Rubin, 1998, 2001) or a struggle for national independence (Conway and Haque, 1999).

While each of these factors might contribute to the reminiscence bump, they cannot entirely explain the phenomenon. Most of these hypotheses (except perhaps the novelty hypothesis) would predict a bump for negative memories, but no such bump is found. Why not?

Berntsen and Rubin (2002, 2004) have suggested that *life scripts* might explain the reminiscence bump. They define a life script as a sequence of events that typically happens to an individual living in a particular culture. They further describe it as an *idealized* life story, meaning that it will exclude common but unpleasant events such as a divorce, the loss of a job, or the early death of a parent. Several lines of evidence suggest that life scripts help structure recall. When adults think about the likely life story of a typical adult in their culture, they tend to list a disproportionate number of events from the 'bump' period (e.g., Berntsen and Rubin, 2004). Similarly, Bohn and Berntsen (2011) asked children to describe their future lives and found

that a disproportionate number of events come from early adulthood. These findings are consistent with the idea that the life script helps structure autobiographical recall by putting greater emphasis on positive events in early adulthood.

Phenomenological Properties of Autobiographical Memory

As we noted above, several autobiographical memory tests can be used to measure the phenomenological properties of autobiographical memory – for instance, its narrative coherence, the vividness of its imagery, the intensity and valence of the emotions it evokes, and so on. Based on data collected with these measures, Rubin and colleagues (Rubin, 1998; Greenberg and Rubin, 2003) have developed a basic-systems model of autobiographical memory that suggests that autobiographical memory is a high-level process that draws on information from a number of other processes, including sensory imagery, emotion, language, narrative coherence, and executive function.

Here we focus on visual imagery, which has been identified as one of the most important of these processes (Brewer, 1996; Brewer and Pani, 1996). In previous work, we used the Autobiographical Memory Questionnaire to investigate the relationship between visual imagery and reliving (Rubin et al., 2003). We found that the intensity of visual imagery during recollection strongly predicted the extent to which participants felt that they were reliving an event. In fact, visual imagery seemed almost necessary for this feeling of reliving: Fewer than 2% of memories were recalled with both a strong sense of reliving and weak visual imagery. D'Argembeau and van der Linden (2006) examined individual differences in visual imagery and showed that the vividness of participants' visual imagery predicted the amount of sensory detail and temporal information in the memory but did not predict the intensity of the memories or their personal importance. Finally, impairments of visual imagery have been associated with memory impairments (e.g., Greenberg and Rubin, 2003), further supporting the idea that visual imagery has a fundamental role to play in autobiographical recall.

Olfactory Cuing of Autobiographical Memory

Proust's description of the memory evoked by the petite madeleine has led a number of researchers to investigate the connection between odor and autobiographical memory. (As a side note, it is not at all clear how this happened. As the quotation above shows, Proust's narrator's memory was triggered by the *flavor* of the madeleine, including its texture and taste, rather than the smell. Taste and smell are certainly linked, in part through the phenomenon of retronasal smell, but most of the studies following Proust have *only* used olfactory cues.) Odor cues can clearly evoke autobiographical memories (Rubin et al., 1984; Larsson and Willander, 2009), and they seem to be particularly good at eliciting vivid and intense memories: odor-cued memories are more emotional (Herz and Cupchik, 1995; Rubin et al., 1984) and more detailed (Chu and Downes, 2002). Furthermore, odor-cued memories tend to be older (Willander and Larsson, 2006) and have not been rehearsed as often (Rubin et al., 1984; Willander and Larsson, 2006). In short, olfactory stimuli seem perfectly tailored to trigger the recollection of a Proustian sort of memory.

Overgeneral Memory and Depression

Most of these studies were trying to find the factors that led to rich and vivid autobiographical memories. Another group of researchers has taken a different approach and has investigated memories that are overly general and devoid of detail. This might seem like a strange topic to study, but it turns out to be important for a very simple reason: the inability to retrieve detailed autobiographical memories is very common in people with various forms of depression and other psychiatric conditions (see Williams et al., 2007, for a review).

Williams et al. (2007) have characterized overgeneral memory as a tendency to summarize past events in categories rather than giving a specific memory, even when asked to be as specific as possible. For example, when asked to recall a specific event in response to the cue word *lonely*, a person without depression might say, "I really felt lonely last Friday when nobody turned up at the party," whereas a person exhibiting overgeneral memory might say, "I feel lonely on the evenings that my wife goes out working" (Hermans et al., 2008: p. 669).

According to several studies, depressed individuals who tend to retrieve overgeneral memories have a poorer prognosis than those who do not (e.g., Brittlebank et al., 1993; Williams et al., 2007). They continue to generate overgeneral memories even when they are not in the midst of a depressive episode (Williams and Dritschel, 1988; Mackinger et al., 2000). Similarly, the production of overgeneral memories generally predicts later depression (Gibbs and Rude, 2004).

Researchers have proposed several theoretical explanations of overgeneral memory (Conway and Pleydell-Pearce, 2000; Williams et al., 2007). First, people might essentially train themselves to retrieve overgeneral memories because it helps them avoid remembering detailed and emotionally painful memories (Conway and Pleydell-Pearce, 2000). Second, people with depression can easily fall into rumination, a state in which they constantly reflect on their depressive symptoms. The connection to overgeneral memory is straightforward: When all mental roads lead to rumination, specific memories are very difficult to retrieve. Third, some evidence suggests that people with depression might have poorer executive function (or might just exert less effort at any cognitive task), which would make it more difficult for them to search for and assemble a detailed memory (Conway and Pleydell-Pearce, 2000; Williams et al., 2007).

Any or all of these factors could induce an individual to retrieve overgeneral memories. Furthermore, Williams et al. (2007) have proposed that these tendencies may grow more severe over time because of mnemonic interlock. That is, as people retrieve overgeneral memories, those memories are rehearsed more and more, making them easier to retrieve than more detailed memories. Recent studies have shown that there are ways to alleviate the problem. Mindfulness-based cognitive therapy can mitigate overgeneral memory as well as depression (Williams et al., 2000; Neshat-Doost et al., 2013), perhaps because they

encourage people to focus on specific aspects of their thoughts and experiences.

Cultural Differences in Autobiographical Memory

An intriguing and relatively recent line of research has revealed cultural differences in autobiographical recall. According to Ross and Wang (2010), people from Western and Eastern cultures tend to describe their memories in different ways. People from Western cultures tend to emphasize individuality and the development of a unique identity, whereas people from East Asian cultures tend to place more emphasis on their role within a larger society. In accordance with this view, people of European descent tended to rate their memories as more personally important than Chinese people did (Wang and Conway, 2004). An examination of the content of autobiographical memories yields similar findings. People of Asian descent tend to report more memories of events involving other people (see Ross and Wang, 2010, for review), and they are more likely to visualize memories from a third-person perspective than people of European descent are (Cohen and Gunz, 2002).

There are other differences as well. Westerners tend to retrieve more autobiographical memories in a variety of tasks (Wang, 2009), and their oldest memories tend to be a bit older (e.g., MacDonald et al., 2000). Interestingly enough, this difference can be eliminated by priming the European participants to focus on group affiliation (or, for that matter, priming the Asian participants to focus on themselves as unique individuals), suggesting that the culture to which they belong creates a default 'mindset' that guides recall (Wang and Ross, 2005). All these findings come with an important caveat: As Ross and Wang (2010) note, the vast majority of studies have compared Asian and European cultures, meaning that differences between other cultures remain relatively unexplored.

Involuntary Memory

So far, most of the studies we have discussed involve voluntary recall – participants were asked to make a deliberate effort to retrieve a specific autobiographical memory. Yet, much of the time – perhaps most of the time – our memory does not work in this way. As with Proust's narrator, our memories often come to mind unbidden, without any apparent cue or any conscious attempt to recall. These are known as *involuntary memories* (Berntsen, 2009), and in the past two decades, they have received an increasing amount of attention in the cognitive psychology literature.

The study of involuntary memories is not new. Berntsen (2009) has noted that the idea goes back at least to Ebbinghaus (1885/1964), who distinguished between non-conscious memories, voluntary memories, and involuntary memories. For the most part, however, psychologists focused on purposeful retrieval and the construction of memory, with some authors suggesting that involuntary memory was not amenable to scientific study at all (Miller, 1962). The problem is straightforward: it is hard to ask participants to conduct a deliberate search for involuntary memories, because then they would not be involuntary at all.

Nevertheless, recent studies have led to new discoveries about this type of memory. Many of these studies used a diary method, in which participants are given a diary and asked to answer a series of questions whenever they realize that an involuntary memory has sprung to mind. These studies show that involuntary memories are quite common. Typically, they are evoked by some kind of identifiable cue, whether from the outside world or the participant's own thoughts (e.g., Berntsen and Hall, 2004; Mace, 2004; Schlagman et al., 2007). Their distribution tends to follow the shape of the forgetting curve (Berntsen and Hall, 2004); the reminiscence bump is still present when older adults are tested, and it is still limited to positive memories (Berntsen and Rubin, 2002; Schlagman et al., 2007). Moreover, the event tends to be positive (Hall and Berntsen, 2008; Walker et al., 2003), as with voluntary autobiographical memory.

There are also intriguing differences between involuntary and voluntary memories. Most studies show that involuntary memories are more likely to refer to specific and distinct episodes (e.g., Berntsen and Hall, 2004); moreover, they do not show the usual decline in specificity with age (Schlagman et al., 2007). They also tend to be more emotional (Berntsen and Hall, 2004). Nevertheless, Rubin and colleagues (2008) found that these memories were less fundamental to a person's life story than voluntary memories were. Berntsen (2009) has suggested that these findings can be reconciled by looking at some of the differences in which these memories are cued. For instance, voluntary memory requires a deliberate search, whereas involuntary memory is elicited by a cue encountered by happenstance. Almost by definition, we are more likely to conduct voluntary searches for personally important memories, but involuntary memories can be about anything. Similarly, a specific and unique episode might be hard to search for voluntarily, but when the right cue is encountered, it will spring to mind. Thus, involuntary memories complement voluntary memories in many ways, forming a fundamental but still-understudied aspect of autobiographical recall.

Neuroimaging of Autobiographical Memory

Cognitive neuroscience has built on this research by using functional neuroimaging to uncover the neural correlates of autobiographical memory (for a review, see St. Jacques, 2012). These studies have generally shown that autobiographical retrieval involves a sequence of activation across the prefrontal cortex, medial temporal lobe, and posterior cortices. Prefrontal activation occurs throughout the medial and lateral prefrontal cortex. Medial activation is generally taken to reflect the processing of self-relevant information, whereas lateral activation is thought to reflect the need to search for an autobiographical memory and maintain it in mind. Within the medial temporal lobe, the hippocampus is thought to mediate vivid recollection by coordinating activity in posterior cortices; amygdala activation reflects the emotional content of the memory. Activation in occipital and retrosplenial regions presumably reflects the vivid visual imagery that is so often associated with autobiographical recall.

Conclusion

All these studies have provided important insights about the ways in which we remember our lives. Dozens of studies have shed light on the durability of memories from early adulthood, the effects of different kinds of cues, and the ways that memories differ across cultures. Other studies have explored the effects of different forms of mental illness, and some have even managed to explore involuntary memory; in the past decade or so, researchers have begun to examine the neural systems that underlie this form of recall. Autobiographical memory will no doubt remain a challenging phenomenon to study, but our efforts so far have clearly paid off.

See also: Amnesia: General; Amnesia: Psychogenic; Autobiographical Memory During Early Childhood and Cultural Variations; Childhood Memories; Confabulation; Declarative Memory, Neural Basis of; Episodic Memory; Eyewitness Testimony; False Memories; Flashbulb Memories; Hippocampal Amnesia; Hippocampus and Related Structures; Reconstructive Memory, Psychology of; Visual Imagery.

Bibliography

Banaji, M., Crowder, L., 1989. The bankruptcy of everyday memory. American Psychologist 44, 1185–1193.

Barclay, C.R., Wellman, H.M., 1986. Accuracies and inaccuracies in autobiographical memories. Journal of Memory and Language 25, 93–103.

Berntsen, D., 2009. Involuntary Autobiographical Memories. Cambridge University Press, New York.

Berntsen, D., Hall, N.M., 2004. The episodic nature of involuntary autobiographical memories. Memory and Cognition 32, 789–803.

Berntsen, D., Rubin, D.C., 2002. Emotionally charged memories across the life span: the recall of happy, sad, traumatic, and involuntary memories. Psychology of Aging 17, 636–652.

Berntsen, D., Rubin, D.C., 2004. Cultural life scripts structure recall from autobiographical memory. Memory and Cognition 32, 427–442.

Bohn, A., Berntsen, D., 2011. The reminiscence bump reconsidered: children's prospective life stories show a bump in young adulthood. Psychological Science 22, 197–202.

Brewer, W.F., 1996. What is recollective memory? In: Rubin, D.C. (Ed.), Remembering Our Past: Studies in Autobiographical Memory. Cambridge University Press, Cambridge, pp. 19–66.

Brewer, W.F., Pani, J.R., 1996. Reports of mental imagery in retrieval from long-term memory. Consciousness and Cognition 5, 265–287.

Brittlebank, A.D., Scott, J., Williams, J.M.G., Ferrier, I.N., 1993. Autobiographical memory in depression: state or trait marker. British Journal of Psychiatry 162, 118.

Chu, S., Downes, J.J., 2002. Proust nose best: odors are better cues of autobiographical memory. Memory and Cognition 30, 511–518.

Cohen, D., Gunz, A., 2002. As seen by the other…perspectives on the self in the memories and emotional perceptions of Easterners and Westerners. Psychological Science 13, 55–59.

Conway, M.A., 2001. Sensory-perceptual episodic memory and its context: autobiographical memory. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 356, 1375–1384.

Conway, M.A., Haque, S., 1999. Overshadowing the reminiscence bump: memories of a struggle for independence. Journal of Adult Development 6, 35–44.

Conway, M.A., Pleydell-Pearce, C.W., 2000. The construction of autobiographical memories in the self-memory system. Psychological Review 107, 261–268.

Crovitz, H.F., Schiffman, H., 1974. Frequency of episodic memories as a function of their age. Bulletin of the Psychonomic Society 4, 517–518.

D'Argembeau, A., van der Linden, M., 2006. Individual differences in the phenomenology of mental time travel: the effect of vivid visual imagery and emotion regulation strategies. Consciousness and Cognition 15, 342–350.

Ebbinghaus, H., 1885/1964. Memory: A Contribution to Experimental Psychology. Dover, New York.

Fitzgerald, J., 1988. Vivid memories and the reminiscence phenomenon: the role of a self narrative. Human Development 31, 261–273.

Fitzgerald, J., 1996. Intersecting meanings of reminiscence in adult development and aging. In: Rubin, D.C. (Ed.), Remembering Our Past: Studies in Autobiographical Memory. Cambridge University Press, Cambridge, pp. 360–383.

Fromholt, P., Larsen, S.F., 1991. Autobiographical memory in normal aging and primary degenerative dementia (dementia of Alzheimer type). Journals of Gerontology: Psychological Sciences 46, P85–P91.

Galton, F., 1879. Psychometric experiments. Brain 2, 149–162.

Gibbs, B.R., Rude, S.S., 2004. Overgeneral autobiographical memory as depression vulnerability. Cognitive Therapy and Research 28, 511–526.

Greenberg, D.L., Rubin, D.C., 2003. The neuropsychology of autobiographical memory. Cortex 39, 687–728.

Hall, N.M., Berntsen, D., 2008. The effect of emotional stress on involuntary and voluntary conscious memories. Memory 16, 48–57.

Hermans, D., Vandromme, H., Debeer, E., Raes, F., Demyttenaere, K., Brunfaut, E., Williams, J.M.G., 2008. Overgeneral autobiographical memory predicts diagnostic status in depression. Behaviour Research and Therapy 46, 668–677.

Herz, R.S., Cupchik, G.C., 1995. The emotional distinctiveness of odor-evoked memories. Chemical Senses 2, 517–528.

Johnson, M.K., Foley, M.A., Suengas, A.G., Raye, C.L., 1988. Phenomenal characteristics of memories for perceived and imagined autobiographical events. Journal of Experimental Psychology: General 117, 371–376.

Kopelman, M.D., Wilson, B.A., Baddeley, A.D., 1989. The autobiographical memory interview: a new assessment of autobiographical and personal semantic memory in amnesic patients. Journal of Clinical and Experimental Neuropsychology 11, 724–744.

Larsson, M., Willander, J., 2009. Autobiographical odor memory. Annals of the New York Academy of Sciences 1170, 318–323.

Levine, B., Svoboda, E., Hay, J.F., Winocur, G., Moscovitch, M., 2002. Aging and autobiographical memory: dissociating episodic from semantic memory. Psychology and Aging 17, 677–689.

Loftus, E.F., Palmer, J.C., 1974. Reconstruction of automobile destruction: an example of the interaction between language and memory. Journal of Verbal Learning and Verbal Behavior 13, 585–589.

MacDonald, S., Uesiliana, K., Hayne, H., 2000. Cross-cultural and gender differences in childhood amnesia. Memory 8, 365–376.

Mace, J.H., 2004. Involuntary autobiographical memories are highly dependent on abstract cuing. The Proustian view is incorrect. Applied Cognitive Psychology 18, 893–899.

Mackinger, H.F., Pachinger, M.M., Leibetseder, M.M., Fartacek, R.R., 2000. Autobiographical memories in women remitted from major depression. Journal of Abnormal Psychology 109, 331–334.

Miller, G.A., 1962. Psychology: The Science of Mental Life. Harper and Row, New York.

Neisser, U., 1978. Memory: what are the important questions? In: Gruneberg, M.M., Morris, P.E., Sykes, R.N. (Eds.), Practical Aspects of Memory. Academic Press, London.

Neisser, U., 1981. John Dean's memory: a case study. Cognition 9, 1–22.

Neshat-Doost, H.T., Dalgleish, T., Yule, W., Kalantari, M., Ahmadi, S.J., Dyregov, A., Jobson, L., 2013. Enhancing autobiographical memory specificity through cognitive training: an intervention for depression translated from basic science. Clinical Psychological Science 1, 84–92.

Proust, M., 1922/1934. Remembrance of Things Past (C.K.S. Moncrieff, Trans.). Henry Holt, New York.

Rathbone, C.J., Moulin, C.J.A., Conway, M.A., 2008. Self-centered memories: the reminiscence bump and the self. Memory and Cognition 36, 1403–1414.

Ross, M., Wang, Q., 2010. Why we remember and what we remember: culture and autobiographical memory. Perspectives on Psychological Science 5, 401–409.

Rubin, D.C., 1998. Beginnings of a theory of autobiographical remembering. In: Thompson, C.P., Herrmann, D.J., Bruce, D., Reed, J.D., Payne, D.G., Toglia, M.P. (Eds.), Autobiographical Memory: Theoretical and Applied Perspectives. Erlbaum, Mahwah, NJ, pp. 47–67.

Rubin, D.C., 2006. The basic-systems model of episodic memory. Perspectives on Psychological Science 1, 277–311.

Rubin, D.C., Berntsen, D., 2003. Life scripts help to maintain autobiographical memories of highly positive, but not highly negative, events. Memory and Cognition 31, 1–14.

Rubin, D.C., Boals, A., Berntsen, D., 2008. Memory in posttraumatic stress disorder: properties of voluntary and involuntary, traumatic and non-traumatic autobiographical memories in people with and without PTSD symptoms. Journal of Experimental Psychology: General 137, 591–614.

Rubin, D.C., Groth, E., Goldsmith, D.J., 1984. Olfactory cuing in autobiographical memory. American Journal of Psychology 97, 493–507.
Rubin, D.C., Rahhal, T., Poon, L., 1998. Things learned in early adulthood are remembered best. Memory and Cognition 26, 3–19.
Rubin, D.C., Schrauf, R.W., Greenberg, D.L., 2003. Belief and recollection of autobiographical memories. Memory and Cognition 31, 887–901.
Rubin, D.C., Schulkind, M.D., 1997. The distribution of autobiographical memories across the lifespan. Memory and Cognition 25, 859–886.
Rubin, D.C., Wenzel, A., 1996. One hundred years of forgetting: a quantitative description of retention. Psychological Review 103, 734–760.
Rubin, D.C., Wetzler, S.E., Nebes, R.D., 1986. Autobiographical memory across the adult lifespan. In: Rubin, D.C. (Ed.), Autobiographical Memory. Cambridge University Press, New York, pp. 202–221.
Schlagman, S., Kvavilashvili, L., Schulz, J., 2007. Effects of age on involuntary autobiographical memories. In: Mace, J.H. (Ed.), Involuntary Memory. Blackwell, Malden, MA, pp. 87–112.
Schrauf, R.W., Rubin, D.C., 1998. Bilingual autobiographical memory in older adult immigrants: a test of cognitive explanations of the reminiscence bump and the linguistic encoding of memories. Journal of Memory and Language 39, 437–457.
Schrauf, R.W., Rubin, D.C., 2001. Effects of voluntary immigration on the distribution of autobiographical memory over the lifespan. Applied Cognitive Psychology 15, S75–S88.
St. Jacques, P.L., 2012. Functional neuroimaging of autobiographical memory. In: Bernsten, D., Rubin, D.C. (Eds.), Understanding Autobiographical Memory: Theories and Approaches. Cambridge University Press, Cambridge, England, pp. 114–138.
St. Jacques, P.L., Schacter, D.L., 2013. Modifying memory: selectively enhancing and updating personal memories for a museum tour by reactivating them. Psychological Science 24, 537–543.
Talarico, J.M., Rubin, D.C., 2003. Confidence, not consistency, characterizes flashbulb memories. Psychological Science 14, 455–461.
Walker, W.R., Skowronski, J.J., Thompson, C.P., 2003. Life is pleasant—and memory helps to keep it that way! Review of General Psychology 7, 203–210.
Wang, Q., 2009. Are Asians forgetful? Perception, retention, and recall in episodic remembering. Cognition 111, 123–131.
Wang, Q., Conway, M.A., 2004. The stories we keep: autobiographical memory in American and Chinese middle-aged adults. Journal of Personality 72, 911–938.
Wang, Q., Ross, M., 2005. Why we remember and what we remember: culture and autobiographical memory. Perspectives on Psychological Science 5, 401–409.
Westmacott, R., Leach, L., Freedman, M., Moscovitch, M., 2001. Different patterns of autobiographical memory loss in semantic dementia and medial temporal lobe amnesia: a challenge to consolidation theory. Neurocase 7, 37–55.
Willander, J., Larsson, M., 2006. Smell your way back to childhood: autobiographical odor memory. Psychonomic Bulletin and Review 13, 240–244.
Williams, J.M.G., Barnhofer, T., Crane, C., Hermans, D., Raes, F., Watkins, E., Dalgleish, T., 2007. Autobiographical memory specificity and emotional disorder. Psychological Bulletin 133, 122–148.
Williams, J.M.G., Broadbent, K., 1986. Autobiographical memory in suicide attempters. Journal of Abnormal Psychology 95, 144–149.
Williams, J.M.G., Dritschel, B.H., 1988. Emotional disturbance and the specificity of autobiographical memory. Cognition Emotion 2, 221–234.
Williams, J.M.G., Teasdale, J.D., Segal, Z.V., Soulsby, J., 2000. Mindfulness-based cognitive therapy reduces overgeneral autobiographical memory in formerly depressed patients. Journal of Abnormal Psychology 109, 150–155.

Autobiographical Memory during Early Childhood and Cultural Variations

Qi Wang and Robert C Blenis, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The development of autobiographical memory is foreshadowed by many early prerequisite skills. It is not an all-or-none phenomenon but involves a gradual incremental process. Young children's memories are initially fragmentary and incoherent. Over the preschool years, their memories show increasing endurance, temporal coherence, complexity, and elaboration as a result of multiple contributing factors. These factors include neurological maturation, language acquisition, development of a sense of self, and participating in adult-guided memory conversations. Evidence emerging from sociocultural approaches further shows that cultural beliefs and the associated narrative practices shape the quantity and quality of autobiographical memory throughout early childhood.

Introduction

Young children often exhibit impressive abilities to remember things that occur in their lives. Memory for distinct personal experiences, namely, autobiographical memory, encompasses knowledge of the self and is critical for personal identity and psychological well-being (Conway and Pleydell-Pearce, 2000; Pillemer, 2009). The development of autobiographical memory is foreshadowed by many early prerequisite skills and takes place rapidly across the preschool years (Bauer, 2007; Nelson and Fivush, 2004). Once toddlers develop general memory skills, rudimentary language, and an emerging cognitive sense of self, they become able to recall fragmentary information about their past experiences with help from adults. With further neurological maturation, language and cognitive growth, and particularly through participation in memory conversations with adults, preschoolers remember personal event information with increasing endurance, temporal coherence, complexity, and elaboration. Furthermore, children's memories of personal experiences are intricately shaped by cultural beliefs and mnemonic practices of the community (Wang and Ross, 2007). During the myriad of daily exchanges with their parents and others, children come to construct personal life stories that reflect both their cultural knowledge and idiosyncratic characteristics.

Autobiographical Memory Development in Early Childhood

Development of Long-Term Memory

The development of autobiographical memory in early childhood is a continuous multifaceted process. First, infants must be able to retain event information over an extended period of time. Arguably, the capacity for long-term memory may be present in very early infancy, at 2 months of age, as reflected in operant conditioning paradigms, such as the mobile reinforcement task, the high-amplitude sucking procedure, or the train task (Reese, 2009). However, a general consensus among researchers is that the ability to form long-term memories emerges at about 6 months of age, as measured in deferred imitation paradigms (Bauer, 2007). In these paradigms, an infant observes an experimenter interacting with novel objects. After a delay, infants are presented with the same objects they saw the experimenter interact with earlier. If the infant imitates the experimenter's earlier interactions with the objects, they are believed to have developed long-term memory. At 6 months, infants can reproduce specific actions after a delay of 24 h, and by 9 months, infants can recall actions of multistep sequences over delays as long as 5 weeks. Still, long-term memory ability is fragile and unreliable in the first year. Infants often require multiple experiences of the to-be-remembered actions in order to recall these actions later, and there exists a wide range of individual differences. The ability to recall sequences of actions improves dramatically over the course of the second year, where older infants can remember complex event information over delays of several months and even a year (Bauer, 2007).

The Self and Autobiographical Memory

Early long-term memories are not '*auto*biographical,' however, until a cognitive sense of self emerges to connect the past experiences to 'me.' The cognitive sense of self – a rudimentary form of self-concept – emerges at about 18–24 months, as indicated by passing a mirror self-recognition test. It is considered a critical precursor for the emergence and development of autobiographical memory (Howe, 2003). With the self in place, personal event information can now be represented and organized in a structured fashion and, consequently, becomes more accessible for later retrieval. Harley and Reese (1999) found that children's self-recognition skills at 19 months uniquely predicted their ability to remember autobiographical events at 32 months, and children who recognized themselves in the mirror at earlier ages developed autobiographical memory at a faster rate than did children who recognized themselves at later ages.

The self continues to develop and enrich over the preschool years, providing children with a sense of psychological continuity in time and enabling them to construct personally meaningful, causal connections among events (Povinelli, 1995; Welch-Ross, 2001). As a result, event representation is gradually transformed from a static collection of isolated episodes to a temporal–causal organization of life stories and, thus, an enduring autobiographical memory system. Notably, the self is only a necessary but insufficient factor for autobiographical

memory development. Among the many other factors that accompany the development of autobiographical memory is children's burgeoning language proficiency.

Language and Autobiographical Memory

Language provides tools for both memory representation and expression (Nelson and Fivush, 2004). Simcock and Hayne (2002) found that children cannot translate prelinguistic experience into language. In their study, young children participated in a unique event and their memories were assessed 6 months or 1 year later. There was not a single instance in which children verbally retrieved information about the event that was not part of their productive vocabulary at the time of encoding. In other words, if the children could not express something verbally the time when the event took place, they did not express it verbally in recall, even though their language skills had improved dramatically.

Although memory may not depend on language, linguistic representation and reinstatement of event information may make the information more enduring (Nelson and Fivush, 2004). Language can be considered to compress the multimodal sensory inputs that form the basis of autobiographical memory into associations between a word or phrase, its meaning, and remembered experience. Language may further serve as retrieval cues to facilitate recollection. As children's language proficiency increases, they can use a wider range of linguistic cues to aid recall and process information more efficiently. Consequently, the expansion of children's verbal repertoire is followed by an increase in the coherence and endurance of their autobiographical memory (Hayne, 2004).

Memory Sharing and Parental Conversational Styles

In addition to serving as a medium for memory representation and expression, the developing language skills allow children to participate in memory conversations with others, particularly parents, which is considered a critical factor for the development of autobiographical memory (Nelson and Fivush, 2004; Reese, 2009). Through memory sharing, parents model to children how to organize and communicate past experiences in narrative forms, which facilitates long-term retention of the event information. Parents also provide evaluations (e.g., 'That's a fun experience') that help children derive personal meanings from past experiences, making the experiences more memorable. Furthermore, memory sharing highlights the social significance of autobiographical memory, thus showing children why the past is worth remembering.

Children begin to participate in joint conversations of the past before the end of the second year, at as early as 15–17 months of age. While their contributions are initially fragmentary and difficult to understand, their recollections are gradually shaped into more vivid, understandable narratives over the preschool years as their parents provide evaluations and feedback and situate the past in personally meaningful contexts (Nelson and Fivush, 2004; Reese, 2009).

Furthermore, the way that parents structure memory conversations is an important variable that affects preschool children's developing autobiographical memory (Fivush et al., 2006). High-elaborative parents often ask open-ended questions, embellish the event under discussion with rich details, and provide evaluative feedback to scaffold and encourage children's participation in the conversation. Low-elaborative parents, in contrast, often attempt to elicit specific answers from children by simply repeating pointed questions and rarely comment on children's responses. These stylistic differences in parent–child memory sharing have both concurrent and long-term effects on children's autobiographical memory: children of high-elaborative mothers typically remember more details and discuss the past with greater elaboration than children of low-elaborative mothers by the end of the preschool years (Fivush et al., 2006; Nelson and Fivush, 2004; Reese, 2009). The effect of early maternal conversational styles on autobiographical memory recall persists through adolescence (Jack et al., 2009).

Other Contributing Factors

There are many other factors that may play important roles in the development of autobiographical memory in early childhood. For example, neurological maturation, particularly developments in medial temporal and cortical structures, enables more efficient consolidation of event information following the initial encoding, thus making the information more enduring (Bauer, 2007). Knowledge about emotion situations helps children understand the personal significance of past events, experience appropriate emotions during the events, and organize the event information in a structured fashion, thereby facilitating the retention of and access to the event memories over the long term (Wang, 2008). An evaluative self-awareness, as reflected in children's understanding of norm violations (e.g., broken toys), use of evaluative words (e.g., 'dirty'), and references to competencies (e.g., 'I can do it'), may serve as a self-referential frame that children use to represent not only personal events as experienced but also their subjective attitudes toward the events (Welch-Ross, 2001). In addition, theory of mind, metamemory, narrative skills, and knowledge of time may all contribute to the development of autobiographical memory in early childhood (Nelson and Fivush, 2004). Given all the contributing factors, as children grow older, they are able to remember more details from their personal experiences, retain the details for a longer period of time, require less support to retrieve the information at recall, focus more on the distinctive aspects of past events, and ascribe more readily personal meaning to the events (Bauer, 2007).

Childhood Amnesia

While children as young as age 2 or 3 years are able to remember autobiographical events for a considerable period of time, most adults do not remember any events that occurred before age 3 years and exhibit a continued scarcity of memories for events occurring between ages 3 and 5 years (Bauer, 2007; Pillemer and White, 1989). This phenomenon is referred to as childhood amnesia or infantile amnesia. It could be argued that childhood amnesia is because events from early childhood occurred so long ago that they become forgotten or inaccessible. However, if the events recalled from different

periods of life are plotted by the time of occurrence, fewer events are recalled from early childhood than would be expected if forgetting occurred linearly as a function of time (Bauer, 2007; Nelson and Fivush, 2004).

Typical studies of childhood amnesia examine adults' recollections of events that occurred in childhood. Recent approaches that examine children's recollections of early childhood have shown that before age 8 or 9 years, children can still recall a great deal of information about salient events that occurred at ages 2 and 3 years (for a review, see Peterson, 2012). The ability to access very early memories decreases with age, such that younger children can recall memories from an earlier age than older children. It appears that early memories vanish or become inaccessible at some point as children grow older. On the other hand, memories for events that occurred after the period of childhood amnesia (after age 5 years) become increasingly well retained, consolidated, and long lasting, as a result of the interactive contribution of a variety of neurological, cognitive, linguistic, and sociocultural factors that were discussed earlier. The emergence of autobiographical memory and the offset of childhood amnesia represent two sides of the same coin.

Cultural Influences on Early Autobiographical Memory

Cultural Views of the Self

Cultural ideologies and practices have pervasive effects on the development of autobiographical memory. In particular, the conception of selfhood differs markedly across cultures (Markus and Kitayama, 1991). In Western, particularly North American, cultures that embrace individuality and autonomy, the self is defined primarily by individuals' inner attributes and qualities independent of social others and contexts – an autonomous self. In contrast, in East Asian cultures such as China and Japan that give prominence to interpersonal harmony and collectivity, the self is largely defined by one's social stratification and responsibility – a relational self.

Given the executive role of the self in modulating autobiographical remembering (Conway and Pleydell-Pearce, 2000), different cultural views of the self may have profound effects on the development of autobiographical memory (Wang and Ross, 2007). An autonomous self, valued in Western cultures, may motivate individuals to attend to, encode, and retain information of events that happened to 'me' and further drive the early development of an organized, articulated, and durable autobiographical memory system. Also, in this cultural context, autobiographical memories may focus on unique personal experiences, especially specific, one-moment-in-time episodes that are unique to the individual and focus on the individual's own roles, feelings, and perspectives (e.g., 'the time I won the spelling bee competition'). In contrast, a relational self, valued in many Asian cultures, may prioritize instead the retention of knowledge – not necessarily in the form of memory – that is critical for social harmony and collective solidarity. The development of a structured memory system for one's own experiences may not be accentuated. When remembering the past, individuals may focus on information about group activities and social interactions. Individuals may also attend to generic routine events (e.g., 'going to parties'), which generally serve to direct one's behavior in particular, oftentimes social, situations (Wang, 2013).

In line with this theoretical analysis, studies have shown that compared with Asian children, European and Euro-American children as young as age 3 or 4 years are able to access more distant and detailed autobiographical memories, retrieve more frequently unique, one-time episodes (as opposed to generic events), and focus more on their own roles and predilections (for reviews, see Wang, 2013; Wang and Ross, 2007). For instance, Wang (2004) examined autobiographical memory as well as self-concept in 180 European American and Chinese preschoolers, kindergartners, and second graders. Children were asked to recount personal events such as a recent time when they did something special and fun. They also answered open-ended questions that elicited self-descriptions. Across all age groups, Euro-American children provided lengthier, more detailed, and more emotionally elaborate memories than did Chinese children. American children also focused more on their own roles, preferences, and feelings (e.g., 'I liked the birthday present,' and 'My mom didn't let me go out but I did anyway') and recalled more specific episodes than Chinese children. The Chinese children, in contrast, often recalled past experiences that centered on social interactions and daily routines. Furthermore, consistent with cultural values and beliefs about the self (Markus and Kitayama, 1991), Euro-American children more often described themselves in terms of personal attributes and predilections (e.g., 'I am smart' and 'I like hockey') than Chinese children, who talked more about social categories and relationships ('I am a child' and 'All my friends like me'). The cultural differences in autobiographical memory and self-concept increased in magnitude with age, presumably reflecting the effect of socialization and enculturation. Importantly, across the entire sample, children who exhibited a greater autonomous sense of self also provided lengthier and more self-focused memory accounts, independent of culture and age.

There is additional evidence that cultural influences on autobiographical memory are mediated by cultural self-views (Wang, 2013; Wang and Ross, 2007). For instance, Wang (2006) examined the question of whether cultural differences in memory accessibility are driven by differing self-concepts that focus on autonomy versus relatedness. Three-year-old children from Chinese families in China, first-generation Chinese immigrant families in the United States, and European American families were asked to recall recent personal events and provide self-descriptions. Euro-American children recalled more specific memory details and described themselves in more personal, as opposed to social, terms than their Chinese and Chinese immigrant peers. Children's self-views mediated the effect of culture on memory details, whereby the self accounted for 50% of variance in the detailedness of memory predicted by culture. Thus, cultural differences in remembering the details of personal experiences are due, at least in part, to cultural views of the self.

Cultural Practices of Memory Sharing

Culture further influences memory development through mnemonic practices. Parent–child memory sharing teaches

children how to represent and reinstate event information in linguistic forms and provides children with the necessary organizational framework around which to structure their autobiographical memories (Nelson and Fivush, 2004; Pillemer and White, 1989). Located in the cultural context, parent–child memory sharing further conveys to children cultural norms and beliefs regarding the purpose and importance of remembering the past and models to children culturally promoted ways of remembering (Wang, 2013; Wang and Ross, 2007).

When conversing with their preschoolers about shared past experiences, Euro-American mothers often use a high-elaborative conversational style in which they dwell on specific episodes, supplement children's responses with rich and embellished information, and provide immediate feedback to scaffold children's participation. In contrast, Chinese mothers often employ a low-elaborate conversational style, where they direct the conversations by frequently posing and repeating factual questions, provide little detail or embellishment to assist the child's participation, and often try to elicit correct answers in a way that emulate a memory test (Wang, 2013; Wang and Ross, 2007). While the high-elaborative conversations between Euro-American mothers and children serve well the purpose of helping children remember details of past events about themselves, the low-elaborative conversations between Chinese mothers and children seem to focus less on memory but reinforce the position of the mother as an authority figure. The use of a high-elaborative style is positively associated with mothers' endorsement of an autonomous self-view and negatively associated with their endorsement of a relational self-view (Wang, 2007).

The thematic content of memory conversations also exhibits marked cultural differences. Euro-American mothers often take a child-centered approach, casting the child as the central character of the story and frequently referring to the child's interests, preferences, opinions, and personal attributes. In contrast, Asian mothers often take a mother-centered, hierarchically organized approach, where they set the direction of the conversation, emphasize social interactions, and frequently refer to moral rules and behavioral expectations with their children (Wang, 2013; Wang and Ross, 2007). The child-centered reminiscing in Euro-American families may encourage remembering self-focused information that highlights one's individuality and uniqueness, and the socially oriented memory sharing in Asian families may facilitate remembering event information related to important others and the community. The following two conversational excerpts illustrate the stylistic and content variations in Euro-American and Chinese mother–child memory conversations (Wang and Fivush, 2005).

American mother–child dyad:

M: … Do you remember doing some crying?
C: Why did I cry?
M: I'm not quite sure why you cried. But do you remember where you were?
C: I cried because I had any, no any balloon.
M: They had no balloon. But then, you were also crying because, did you not want to go home?
C: Yeah.
M: Where were you?
C: At Stewart Park!
M: (Laughs). You did cry a lot at Stewart Park, but, um, this was in Joe's parking lot. Do you remember Joe's Restaurant parking lot? Do you remember standing by the door and crying?
C: Yeah.
M: You do?
C: Yeah.
M: What were you crying about?
C: 'Cause I didn't wanted to leave yet; it was because I wanted to eat.
M: Oh you wanted to eat some more (laughs); is that why?
C: Yeah.
M: Hmm. I remember Mommy tried to pick you up and you put up a little bit of a fight. You were crying real hard. Maybe it was 'cause the balloon and maybe it was 'cause you were hungry. But we knew that you could get another balloon, right?
C: Yep

Chinese mother–child dyad:

M: … Do you remember why Dad spanked you last time?
C: Chess!
M: Why chess? What did you do with chess?
C: Not obedient!
M: How were you not being obedient?
C: (I) threw the pieces on the floor.
M: All over the floor, right? And did you do it on purpose?
C: Umm. I'll be careful next time!
M: Right! That's why Dad spanked your bottom, right? …Did you cry then?
C: (I) cried.
M: Did it hurt?
C: It hurt.
M: It hurt? It doesn't hurt anymore, right?
C: Right. I'll be careful next time.
M: Umm, be careful.

Cultural variations in maternal style of memory conversation have been observed even among Western families. In a study on memory sharing between mothers and their 4-year-old children from Germany, Sweden, and Estonia, Tõugu and colleagues (Tõugu et al., 2011) found that Estonian mothers, who presumably placed less value on autonomy over relatedness, were less elaborative than Swedish and German mothers. Other cultural dimensions may further influence parental conversational styles. For example, Reese and colleagues (Reese et al., 2008) examined mother–child memory conversations in Māori and New Zealand European families. Māori culture traditionally places a strong emphasis on the past and the oral histories of individuals and families. Accordingly, when sharing memories with their preschool and grade-school children, Māori mothers elaborated more in significant past events than did European mothers.

Interestingly, children as young as age 3 years have internalized their mothers' style of conversation when sharing memories about the personal past (Wang, 2007, 2013; Wang and Fivush, 2005). During memory conversations, Euro-American children often contribute more information about the events under discussion and refer more frequently to their own roles, feelings, and thoughts than their Asian peers, who provide less memory information and more often comment on social rules and the roles of others. Furthermore, when

children independently recount their experiences with an interviewer, the volume, style, and content of their memory narratives echo the cultural variations in mother–child memory sharing. Euro-American children tend to provide lengthy, elaborative, self-focused memory narratives and place a great emphasis on personal predilection and autonomy. In comparison, Asian children's memory narratives are often brief, skeletal, relation-oriented, and exhibit a great concern with authority and moral rectitude (Wang, 2004, 2013).

Furthermore, differing parent–child memory sharing has direct consequences for autobiographical memory development. Studies have shown that Euro-American mothers' use of a high-elaborative conversational style early on, compared with Chinese mothers' low-elaborative style, directly facilitates American children's memory recall concurrently and over the long term (Wang, 2007). Maternal style further serves as a potent mediator in explaining cultural differences in children's autobiographical memory (Wang, 2006). Importantly, given the low-elaborative memory sharing early in development within Asian families, Asian children may not develop elaborated narrative schemas for representing and recalling the personal past, which may, in turn, result in less detailed and sparser early childhood memories in adulthood. Indeed, Asian adults exhibit more pervasive and prolonged childhood amnesia than Western adults (Wang, 2003a, 2013).

Culture and Emotion Knowledge

Another important cultural variable that has been investigated in relation to children's developing autobiographical memory concerns cultural beliefs pertaining to emotion. In Western, particularly European American, cultures that emphasize individuality and autonomy, emotion is regarded as a direct expression of the self and an affirmation of the uniqueness of the individual. Explicit expression and communication of emotion is encouraged and considered psychologically beneficial. In contrast, in East Asian cultures that put a premium on social harmony and group interests, such as Korean, Chinese, and Japanese cultures, emotion is often viewed as destructive or even dangerous to ongoing relationships and therefore needs to be strictly controlled. Also, it is believed that emotion should be inferred or implied and explicit communication is discouraged (Markus and Kitayama, 1991; Wang, 2013).

These different cultural beliefs about emotion and emotion communication are reflected in early family socialization. Western middle-class parents are often eager to help children understand and express emotions and feelings in order to raise an 'emotionally intelligent' child. Conversely, Asian parents are not preoccupied with facilitating emotional understanding in their children and emphasize, instead, psychological discipline and behavioral standards (Chao, 1995; Wang, 2013). Moreover, when Euro-American mothers converse with their 3-year-old children about shared emotional experiences, they often engage in extensive discussion of emotion, talking about the causes and consequences of emotions and providing elaborate explanations as to why and how the child was experiencing certain feeling states in the past event. In contrast, Chinese mothers often initiate little discussion about emotion and when they do so, they criticize or comment on the incorrectness of their children's (negative) emotional experience or behavior (Wang, 2013; Wang and Fivush, 2005).

The great emphasis on the personal importance of emotion and emotional understanding and the associated family narrative practices in Euro-American culture directly facilitate children's developing emotion knowledge (Doan and Wang, 2010). When asked to judge the emotional nature of situations in stories or to describe situations likely to provoke various emotions, Euro-American preschoolers outperform their Chinese peers regardless of age (Wang, 2013). For example, in a study with Euro-American and Chinese 3- to 6-year-olds (Wang, 2003b), children were presented with 20 short stories with a protagonist of their age, gender, and ethnicity, and were asked to identify the feeling states of the protagonist by choosing among faces showing happy, sad, scared, or angry emotions. Children's mothers and a second group of adults read the same stories and judged the protagonist's emotions in the same fashion. Based on the concordant judgments between children and adults in each culture, findings showed that American children had a better grasp of emotion knowledge and made more rapid progress in such knowledge than their Chinese peers.

Conceivably, the greater emotion knowledge in Euro-American children helps them better remember autobiographical events for long-term retention. Indeed, compared with their Chinese counterparts, Euro-American children tend to provide more elaborate and detailed memory accounts (Wang, 2004, 2013). The relation of emotion knowledge to autobiographical memory development is directly tested in a longitudinal investigation with European American, Chinese American, and Chinese preschoolers (Wang, 2008). Across the preschool years, Euro-American children showed overall greater understanding of emotion situations and recalled more detailed autobiographical memories than Chinese and Chinese American children. Regardless of culture, children who showed greater understanding of emotion recalled memories with more specific details both concurrently and longitudinally than children who had less emotion knowledge. Emotion knowledge is thus related to autobiographical memory at both the cultural and the individual levels. Furthermore, emotion knowledge functioned as a potent mediator that partially accounted for cultural differences in memory recall.

There are also other cultural factors that may impact on the early autobiographical memory development, such as cultural beliefs concerning the functions or purposes of remembering and sharing personal experiences (Kulkofsky et al., 2009; Wang and Ross, 2007), perceptual styles (Masuda and Nisbett, 2001), language (Wang et al., 2010), prominent lifestyles of a community (Wang, 2003b), and so on. For a comprehensive discussion, see Wang (2013).

Concluding Remarks

The development of autobiographical memory during early childhood exhibits both universality and cultural diversity. It is underlined by general neurological maturation and cognitive and linguistic advances as well as culture-specific beliefs and practices. The end result reflects the interaction between the individual child and the developmental niche embedded in a cultural context.

A functional approach is critical for understanding different modes of autobiographical memory development as the consequence of adaptation to specific cultural settings. We have seen that children's autobiographical memories mirror the conversational styles of their parents and the values and beliefs of their culture. For example, parents who explain the origins and consequences of emotion during memory sharing often have children who have a better understanding of emotion and who recollect emotionally rich memories. Parents who tell personal stories with rich details and elaboration often have children who do the same. Cultures that place value on autonomy and individuality often cultivate children who remember detailed and unique information about their life experiences. Autobiographical memory is evolved as a functional adaptation that supports an individual's ability to navigate in his or her environment and make predictions about the future (Suddendorf and Corballis, 2007). Cultural differences in social and physical environments are bound to produce cultural differences in memory.

Thus, cultural differences in autobiographical memory do not imply any deficit in general cognitive development in one group or another. Instead, the development of autobiographical memory, and social–cognitive skills in general, may reflect specific developmental responses to varying cultural expectations and environmental demands. The general and other-focused autobiographical memory in Asian children, although likely being viewed as unconventional or even deficient in Western cultural contexts, reaffirms a relational self that is privileged in Asian cultures. It further facilitates the development of other social–cognitive skills such as theory of mind (Lu et al., 2008). Detailed remembering of one's own experiences, as much as is taken for granted as a positive personal quality in Western cultures that value independence and autonomy, may signal excessive focus on the psychological self, which is incongruent with the Asian cultural norm that emphasizes relationships.

Ultimately, the development of autobiographical memory is a culturally constructive process. As Bartlett proclaimed 80 years ago (1932, p. 296), "Social organization gives a persistent framework into which all detailed recall must fit, and it very powerfully influences both the manner and the matter of recall."

Bibliography

Bauer, P., 2007. Remembering the Times of Our Lives: Memory in Infancy and Beyond. Lawrence Erlbaum Associates.

Chao, R.K., 1995. Chinese and European American cultural models of the self reflected in mothers' childrearing beliefs. Ethos 23 (3), 328–354. http://dx.doi.org/10.1525/eth.1995.23.3.02a00030.

Conway, M.A., Pleydell-Pearce, C.W., 2000. The construction of autobiographical memories in the self-memory system. Psychological Review 107 (2), 261–288. http://dx.doi.org/10.1037/0033-295X.107.2.261.

Doan, S.N., Wang, Q., 2010. Maternal discussions of mental states and behaviors: relations to emotion situation knowledge in European American and immigrant Chinese children. Child Development 81 (5), 1490–1503. http://dx.doi.org/10.1111/j.1467-8624.2010.01487.x.

Fivush, R., Haden, C.A., Reese, E., 2006. Elaborating on elaborations: role of maternal reminiscing style in cognitive and socioemotional development. Child Development 77 (6), 1568–1588. http://dx.doi.org/10.1111/j.1467-8624.2006.00960.x.

Harley, K., Reese, E., 1999. Origins of autobiographical memory. Developmental Psychology 35 (5), 1338–1348. http://dx.doi.org/10.1037/0012-1649.35.5.1338.

Hayne, H., 2004. Infant memory development: implications for childhood amnesia. Developmental Review 24 (1), 33–73. http://dx.doi.org/10.1016/j.dr.2003.09.007.

Howe, M., 2003. When autobiographical memory begins. Developmental Review 23 (4), 471–494. http://dx.doi.org/10.1016/j.dr.2003.09.001.

Jack, F., MacDonald, S., Reese, E., Hayne, H., 2009. Maternal reminiscing style during early childhood predicts the age of adolescents' earliest memories. Child Development 80 (2), 496–505. http://dx.doi.org/10.1111/j.1467-8624.2009.01274.x.

Kulkofsky, S., Wang, Q., Koh, J.B.K., 2009. Functions of memory sharing and mother–child reminiscing behaviors: individual and cultural variations. Journal of Cognition and Development 10 (1–2), 92–114. http://dx.doi.org/10.1080/15248370903041231.

Lu, H., Su, Y., Wang, Q., 2008. Talking about others facilitates theory of mind in Chinese preschoolers. Developmental Psychology 44 (6), 1726–1736. http://dx.doi.org/10.1037/a0013074.

Markus, H.R., Kitayama, S., 1991. Culture and the self: implications for cognition, emotion, and motivation. Psychological Review 98 (2), 224–253. http://dx.doi.org/10.1037/0033-295X.98.2.224.

Masuda, T., Nisbett, R.E., 2001. Attending holistically versus analytically: comparing the context sensitivity of Japanese and Americans. Journal of Personality and Social Psychology 81 (5), 922–934. http://dx.doi.org/10.1037//0022-35i4.81.5.922.

Nelson, K., Fivush, R., 2004. The emergence of autobiographical memory: a social cultural developmental theory. Psychological Review 111 (2), 486–511. http://dx.doi.org/10.1037/0033-295X.111.2.486.

Peterson, C., 2012. Children's autobiographical memories across the years: Forensic implications of childhood amnesia and eyewitness memory for stressful events. Developmental Review 32, 287–306. http://dx.doi.org/10.1016/j.dr.2012.06.002.

Pillemer, D.B., 2009. Twenty years after Baddeley (1988): Is the study of autobiographical memory fully functional? Applied Cognitive Psychology 23 (8), 1193–1208. http://dx.doi.org/10.1002/acp.1619.

Pillemer, D.B., White, S.H., 1989. Childhood events recalled by children and adults. In: Reese, H.W. (Ed.), Advances in Child Development and Behavior, Vol. 21. Academic Press, New York, pp. 297–340.

Povinelli, D.J., 1995. The unduplicated self. In: Rochat, P. (Ed.), The Self in Infancy: Theory and Research. Advances in Psychology. North-Holland/Elsevier Science Publishers, Amsterdam, Netherlands, 112 (pp. 161–192).

Reese, E., 2009. The development of autobiographical memory: origins and consequences I. The development of autobiographical memory: origins and consequences II. The growth of autobiographical memory from infancy to adolescence. Advances in Child Development and Behavior 37, 145–200. http://dx.doi.org/10.1016/S0065-2407(09)03704-5. Elsevier.

Reese, E., Hayne, H., MacDonald, S., 2008. Looking back to the future: Māori and Pakeha mother–child birth stories. Child Development 79 (1), 114–125. http://dx.doi.org/10.1111/j.1467-8624.2007.01114.x.

Simcock, G., Hayne, H., 2002. Breaking the barrier? Children fail to translate their preverbal memories into language. Psychological Science 13 (3), 225–231. http://dx.doi.org/10.1111/1467-9280.00442.

Suddendorf, T., Corballis, M.C., 2007. The evolution of foresight: What is mental time travel, and is it unique to humans? The Behavioral and Brain Sciences 30 (3), 299–313. http://dx.doi.org/10.1017/S0140525X07001975. discussion 313–51.

Tõugu, P., Tulviste, T., Schröder, L., Keller, H., De Geer, B., 2011. Socialization of past event talk: cultural differences in maternal elaborative reminiscing. Cognitive Development 26 (2), 142–154. http://dx.doi.org/10.1016/j.cogdev.2010.12.004.

Wang, Q., 2003a. Infantile amnesia reconsidered: a cross-cultural analysis. Memory 11 (1), 65–80. http://dx.doi.org/10.1080/741938173.

Wang, Q., 2003b. Emotion situation knowledge in American and Chinese preschool children and adults. Cognition & Emotion 17 (5), 725–746. http://dx.doi.org/10.1080/02699930302285.

Wang, Q., 2004. The emergence of cultural self-constructs: autobiographical memory and self-description in European American and Chinese children. Developmental Psychology 40 (1), 3–15. http://dx.doi.org/10.1037/0012-1649.40.1.3.

Wang, Q., 2006. Earliest recollections of self and others in European American and Taiwanese young adults. Psychological Science 17 (8), 708–714. http://dx.doi.org/10.1111/j.1467-9280.2006.01770.x.

Wang, Q., 2007. "Remember when you got the big, big bulldozer?" Mother–child reminiscing over time and across cultures. Social Cognition 25 (4), 455–471. http://dx.doi.org/10.1521/soco.2007.25.4.455.

Wang, Q., 2008. Emotion knowledge and autobiographical memory across the preschool years: a cross-cultural longitudinal investigation. Cognition 108 (1), 117–135. http://dx.doi.org/10.1016/j.cognition.2008.02.002.

Wang, Q., 2013. The Autobiographical Self in Time and Culture. Oxford University Press, New York, NY.

Wang, Q., Fivush, R., 2005. Mother–child conversations of emotionally salient events: exploring the functions of emotional reminiscing in European-American and Chinese families. Social Development 14 (3), 473–495. http://dx.doi.org/10.1111/j.1467-9507.2005.00312.x.

Wang, Q., Ross, M., 2007. In: Kitayama, S., Cohen, D. (Eds.), Culture and Memory, first ed. Guilford Publications, New York, NY, pp. 645–667.

Wang, Q., Shao, Y., Li, Y.J., 2010. "My way or mom's way?" the bilingual and bicultural self in Hong Kong Chinese children and adolescents. Child Development 81 (2), 555–567. http://dx.doi.org/10.1111/j.1467-8624.2009.01415.x.

Welch-Ross, M.K., 2001. Personalizing the temporally extended self: evaluative self-awareness and the development of autobiographical memory. In: Moore, C., Lemmon, K. (Eds.), The Self in Time: Developmental Perspective. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 97–120.This page intentionally left blank

Automation: Organizational Studies

Federico Butera, University of Milan-Bicocca, Milano, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Applications and social consequences of automation are explored in factories, offices, travel, banking, retail, safety, and others. Automation is not sheer displacement of tasks from men to machines but a stage in the process toward integrated systems of processes, technology, organization, roles, and values, where technology performs a large variety of existing and new tasks and where interaction is designed among human beings and technical systems.

Technology upsets work but only design resets work. Social and technical methodologies take the opportunity of the always existing 'room of maneuver' for planning, design, and experimentation of systems. Social dimensions in technological development are not effects but instead matter of design.

The Concept of Automation

The word automation, as a contraction of automatic production, was first used in 1952 both by John Diebold – author of the book *Automation: the advent of automatic factory* – and by D.S. Harder, vice-president of manufacturing at Ford Motor Company.

According to Gallino (2006), automation, as a popular term, "indicates different situations where the human physical and intellectual work is replaced by machines and by mechanical, hydraulic, pneumatic, electrical, electronic servomechanisms which may automatically perform sequences of operations."

As a result of 50 years of discussions about a changing phenomena, I define *automation as a stage in the process toward integrated systems of processes, technology, organization, roles and values, where technology performs a large variety of existing and new tasks, and where interaction is designed among human beings and technical systems with the goal to get optimal products and services*.

Three levels of interaction are indicated by most authors: (1) *control systems* based upon closed-loop control mechanism of feedback (on standards) and feed forward (on goals or week signals), performed by any kind of technological device; (2) *integration* of different devices, processes into a unitarian architecture at the level of a factory, firm, network, achieving continuity of processes, and management control; and (3) *system adaptation and innovation*, through rapid detection both of the internal state of the system and of the environment (technical, economic, commercial, etc.). In all cases now those components include tight collaboration among information and communication technologies, advanced managerial systems, and operative work.

Applications and literature indicate that automation is not a device to be purchased, but a dynamic process. In fact (1) different levels of automation depend upon the changing quantity of human tasks embedded in the technical system; (2) different patterns of automation evolve from forms and degrees of interaction between human beings and technology and the way they should co-operate; and (3) systems are not developed once in a time, but along subsequent stages and through change management processes.

Essentials of automation are (1) some necessary human tasks (steering, maintenance, logistics, etc.), because without it there would be only automatic machines; (2) the need for collaboration among human beings and machines, thanks to interrelated activities and interaction through interfaces; and (3) integration of technological, organizational, social system adopted.

The essence of automation is *automatic control*. This concept has been developed as an academic discipline and as a set of technologies applied in a large variety of intelligent human artifacts, that is machines that can read complex data, signals, symbols, have feed-back loop and feed-forward mechanisms, interpret external and internal variables, adjust its conduct accordingly, and that, sometimes, can find new solutions, act without human intervention, and do something which resembles thinking. We live in an environment where most of the artifacts display these features at different degrees: an elevator, a fully automated rolling mill, a remote-controlled automatic train, an auto-piloted aircraft, a personal computer, a robot, and many others. Introduction of automated technologies in various domains can decrease operator workload and increase operator awareness, but researchers have recognized that these technologies must not be developed apart from human operators. As a matter of fact, human-centered automation is an approach to create an environment in which humans and machines work together cooperatively (Billings, 1997; Sheridan and Hennessy, 1984). Principles of human-centered automation constitute, among others, the incorporation of the human operator in the execution of the automated tasks, and appropriate information distribution as well as automated functions that are easy to learn and to operate.

In the following discussion, we will concentrate first upon *industrial automation* and also on an increasing number of applications that go beyond factories.

Applications

Concept, research, and debate on automation have evolved since 1952, according to the available technologies and areas of applications. *Factory automation* came first when mechanical engineering together with production engineering mechanized

 http://dx.doi.org/10.1016/B978-0-08-097086-8.73036-4

both the machine tools and the conveyors in the automobile industry, giving rise to that primitive but seminal type of automation called *Detroit Automation*, that is *transfer lines* where products, e.g., engines, are manufactured and moved without human effort.

In 1913 Henry Ford conducted the first notable material-flow automation exercise by automating the assembly line for automobile engines. Electromechanical engineering developed ingenious devices for automatic control of continuous process industries – as oil, chemical products, cement – and of discontinuous flow process – as steel mills, casting. They were applied also to traffic control of material flows – as in railway's control rooms – and of energy flows – as in power plants. Automation introduced major disruptions in assembly, packaging, and bottling in various manufacturing industries, actually turning from discontinuous processes into continuous ones. In the 1960s, the developments of electronics and computer sciences made those innovations much deeper and wider in all industries. Computer control in the 1970s also extended automatic control into individual operating machines of numerically controlled machines (CNC), advanced robots, integrated CNC, and robots in flexible manufacturing systems (FMS). CNC and FMS were a breakthrough: manufacturing operations were automated and at the same time it was possible to manufacture small batches at costs of economy of scale, since the optimization of planning processes was embedded in the system. In recent cases of automation, processes of R&D, manufacturing, logistics, and distribution were powerful means for integrating among internal and external functions or external units, achieving interfunctional and intercompany seamless processes in the hierarchic-functional classic industrial organizations. Architecture of telecommunications and information systems, and organization network were overboosting those processes and organizational integration.

Human task displacement and production integration pursued two distinct models. The first one was the *unmanned factory* aiming at cutting costs and controversies of the blue-collar workers through advanced mechanization and use of robots, with the aim of getting rid of people, as was in Italy (the Fiat Cassino experiment of the 1980s). The second model was the *integrated factory* which deliberately keeps a meaningful portion of human tasks out of machines, as in the late Toyota developments and in Fiat Melfi. Cost savings and productivity gains are reported in Japanese firms making considerable steps back from full automation. In the *integrated factory*, an optimal distribution of skilled work and automatic control takes place where advantages are not related only to the reduction of direct manpower but also to the integration of various functions and units so as to achieve continuity and flexibility in business processes.

Thanks to industrial robots, new technologies, laser cutting and injection molding, and additive manufacturing machines, production lines can be automated to the point that they can run unsupervised for weeks without any human intervention. But this does not mean that people are not necessary: all these automated machines need someone who services them and tells them what to do. Machine operators become machine minders, with new skills and detailed knowledge about the workplace in general and the production process in particular (Karsvall et al., 2003). This is clear above all in cases where human operators must ensure that the automated system is performing properly and must detect variances, that is undesired status of process or product (Dalmine case in Butera and Thurman, 1984), or situations that diverge from the expected performance (*The Economist*, 2012). Nowadays, clever software, novel materials, more dexterous robots, new processes, three-dimensional printing, and a whole range of web-based services have contributed to the diffusion of a new idea of factory automation beyond production lines: stand-alone machines are linked together into intercompany integrated processes.

A different type of automation is *office automation*. Because of the development of large mainframes, electronic data processing in the 1980s automated large amounts of operations of data collection, retrieval, and processing, both in public and private administrations: the work and the organization in accounting, planning, and control, and commercial departments were deeply affected. Companies introduced a wide range of stand-alone applications, such as manufacturing planning, drafting, and design (computer-aided design – CAD). All external-facing communications and relations at that time were carried out through fax and phone lines. The next stage in information-flow automation, then, were the enterprise resource planning applications, centered on sharing information among multiple individuals and functional areas of a single company. Electronic data exchange (EDI) supported the communication among different departments and different firms.

Video terminals first, then stand-alone personal computers, and finally networked PCs in the 1990s deeply affected the typical clerical and professional operations such as calculating, typing, editing, copying, printing, retrieving, and mailing, introducing a revolution in the office work.

Automation for professionals was intended to substitute and accelerate routine tasks of scientists and professionals. Computer graphics and multimedia at first, and then Internet did not automate tasks, but offered opportunities to *inform* and *communicate* which did not exist before (Zuboff, 1988). In the twenty-first century, some technologies such as VoIP, e-mail, tagging, phone/video-conferencing, and instant messaging built a platform on which users can communicate but can also exchange information, express thoughts, discuss, and reconfigure existing explicit knowledge. This platform is a good example of combining and remixing knowledge in order to produce new knowledge (Haigh, 2006).

Task automation, information technology, and new roles gave rise to the knowledge management approach very popular from the 1990s (Nonaka and Takeuchi, 1995; Brown and Duguid, 1991).

Automation in services is a relatively recent appearance. Automated processes are transforming service delivery in many industries (Karmarkar, 2004). Services are not totally automated; in most cases they are *hybrid services*. They are not simply services in which some of the information are processed digitally. They combine human and machine-based capabilities, depending on human talents to deliver, customize, and personalize (Zysman et al., 2010).

In the *banking sector* it is possible to identify two different automation paradigms. The first paradigm (back-office automation, 1951–80) mainly aimed at enhancing internal

activities. In this context, the database management system can be considered as the first technical device developed for banking purposes (Fincham et al., 1994; Batiz-Lazo et al., 2002). Due to the competitive environment of the 1970s, banks were required to further exploit the benefits of automation and adopted the clearing system (Morris, 1986) and the retail money transfer mechanism (Thompson, 1967; Mandell, 1990). Barclays – the bank that issued the UK's first credit card in 1966 and, a few years later, the first automated teller machines (ATM) – is a typical pioneer of automation (Ackrill and Hannah, 2001; Batiz-Lazo et al., 2002). This kind of automation brings with it certain benefits such as greater storage and faster processing capacity to accommodate the increasing demand for banking transactions.

The second paradigm focused on customer service automation (front-office automation, integrated networking, 1981–to date) and the proliferation of the service delivery channels building on the experience of the ATM. Online customer portals represent an opportunity to involve the front office and the customers more actively in the banking processes, in popular language it is the home banking. Automation improves customer service, by enabling the front office to focus on providing highly personalized services to its customers (Consoli, 2005).

In the *travel industry* the computer reservations system and the airline reservations systems enabled the automation of flight and seat control within an airline. For example, automation allows travel agents not to call the airline to make reservations: therefore, they can save more time to help travelers. Although travel agencies operate most of transactions digitally, there are complex cases and particularly high-value customers that absolutely require human agents. So the travel industry is not completely automated: it is a hybrid service where human beings and technology work together.

In the *retail industry*, bar code devices and RFId allow the sharp reduction of shelf replacement, ordering, cashier work, etc. But, in addition to this, the act of purchasing at a supermarket or at a retailer is no longer just a simple monetary transaction, but a data-generating activity in all respect. At the beginning of the application of ICT to retail, inventories were monitored. Then, information have been run in order to capture consumer preferences and consumption patterns, as well as to manage inventories and supply chains and, sometimes, sold to third parties. As the range of products and services expands and customer demand increases, support systems and technological tools become the integrating factor in business operations. The infrastructure moves the data both to the front office operator and to the marketing operator in charge of customer profiling and segmentation. Web interfaces, multimedia, and tracking tools can manage various interactions and follow up work items connecting front office and service management operations.

In *safety critical domains and sector* such as aviation, air traffic control, and healthcare, automated systems are being used to perform tasks. Automated systems were introduced primarily in order to reduce human error, that until then was thought to be the main cause for most of accidents in safety critical domains (Coyne, 1994; Shappell and Wiegmann, 2000). Automated systems offer several benefits. For example, reliable automation has shown its usefulness in improving conflict detection performance of air traffic controllers (Metzger and Parasuraman, 2005; Sethumadhavan, 2011), in providing faster and accurate decision making in military tasks (Rovira et al., 2007), and in improving pilot performance during icing conditions (Sarter and Schroeder, 2001). In the healthcare industry, automation can help to exchange data and files in different formats between many different departments and partners and to reduce errors. So, automation enables healthcare organizations to streamline operations across departmental boundaries while eliminating human error and delays.

Networking is becoming the normal way of doing: networks of enterprises and networks of people for managing operations, information, or social relations.

A huge amount of functions are run automatically by the internet technology. In addition to the instant and worldwide communication of texts, images, and sound, Internet is going to include functions and structures which in the past were work and organizational systems. The analysis of different service sector suggests that the key source of change is not a complete automation, but a combination between technology and people, since automation is the creation of distributed knowledge among human beings and machines (Bagnara et al., 2010). After all, the role of human operators has become more important in comparison with times in which automation was not part of our lives yet.

Social Effects of Automation

The literature on social effects of automation has had a discontinuous progress. In the late 1950s, developments of automation in various industries gave rise to a wide program of researches on social effects of automation financed by the governments of USA and other industrialized countries. Under unions pressure, these researches came to the worries of an imminent wide diffusion of automation. Advantages commonly attributed to automation included higher production rates and increased productivity, more efficient use of materials, better product quality, improved safety, shorter workweeks for labor, and reduced factory lead times. But the fears of potential consequences in terms of unemployment, deskilling, poor working conditions, work organization, and union/management relations were strong. Automation could also increase the mental workload of an operator if he/she, for example, does not understand how the automated system works or what is being performed. Automation might also decrease an operator's awareness of the situation by alienating him/her from the tasks carried out, leaving the operator less attentive.

Those outstanding researches delineated most of what we know in this field, but in the middle of 1960s they came to a sudden stop when the American 'National Commission for Technology, Automation and Economic Progress' came to the conclusion that employment would not have been affected, due to difficult implementation and limited diffusion of automation in most industries. In the 1970s automation, on the contrary, had a continuous incremental diffusion, without a dramatic drop of employment, but a sharp change in the professional composition of work force, that is less blue collars and more knowledge workers. This phenomenon did not

receive great attention by social scientists, unless in case it appeared relevant for the ongoing studies and criticisms of taylor-fordism. In the 1980s, automation grew at a faster pace and worries returned about employment and working conditions. In the 1990s, discussion and research focused more on information and communication technologies and their impact on the service sector. At the beginning of 2000, a great interest was arising about Internet: in the 2010, automation fell into the larger category of innovation creating economic growth and requiring new skills for new jobs (Cedefop, 2010).

Discussion about the nature and scope of automation may be grouped around two basic divides.

The first great divide was around the issue of *continuity/discontinuity* of technical development. Bright (1958) offers a classification of types of mechanization based upon degree of task allocation among men and machines, including automation. For economists automation is no more than a particular case of technical development and no different social effects should be expected: "the term automation is used in a loose way and there is not any advantage in distinguishing it from advanced forms of technical change" (Buckingham).

Crossman (1960) stated that automation – he was among the first and major scholars – is the second major stage in the displacement of human work substitution of information processing to machines, since in automation "to steer, control, operate a machine, calculate, solve problems, take decision, discussion may be performed by machines as by the human brain." Artificial intelligence made computers more able to cope with complex problems. Development of advanced robotics gave machines sensor-motor skills (Sheridan and Hennessy, 1984).

A breakdown on the contrary was envisaged first by Diebold (1952), who saw automation as a new concept of integrated systems allowing continuous flow in production process. When computers were not yet diffused, Pollock defined automation as "the integration of continuous and discontinuous processes under coordination of electronic systems." In Buckingham's opinion (1961), automation is a new pattern of manufacturing.

According to sociologists as Touraine, Woodward, Blauner, Meissner, and Braverman, automation is a breakdown because massive displacement of skilled and unskilled work into machines overcomes both the craft-based organizations and the mass production model. New tasks, roles, and organizations are created while the amount of work force required for product unit sharply diminishes: most of them agree that human beings maintain mainly control tasks. This happened widely in cases of the last 20 years with *supervisory control* of systems fully controlled by a computer and human beings controlling the computer, ranging from a minimum of switching on/off, to controlling how the computer controls the computer.

Automation, in this prevailing view of social scientists, results to be a breakdown, because it is not only the substitution of any kind of human work, but the generation of new functions not affordable by the human being. Automation, in this view, is the creation of distributed knowledge among human beings and machines that is human–machine cooperation (Bagnara, in Hoc, 2000). *Automation is integration of complex technological and human system by means of automatic devices*. Human work on the other hand does not disappear, but changes its interactional composition. Knowledge workers increase and concentrate on the design, development, and supervision of 'intelligent machines' (Butera et al., 2008).

The second great divide was around the *pessimistic/optimistic view* of automation. The pessimistic view holds that a new type of taylorism is under way, the technological taylorism. The 'intelligence of the system' as a whole would shift upward toward functions which will become more specialized. The results will be a new wave of deskilling among operatives and employees who will become 'transitional ancillary workers' and 'data entry workers.' Consequences of this may be the 'polarization of the labor force,' in the growing divergence between the unskilled masses and a super-skilled elite (Braverman, 1974; Rifkin, 1995).

The opposite view – the optimistic view – sees technology as a way of freeing workers from monotonous or dangerous tasks and from physical fatigue. Machines will do the work and human beings will supervise them (Blauner, 1964). Although the new technologies require less manpower, they should boost the cycle of expansion, giving rise to new firms and new professions (Freeman, 1998).

Researches supported one or another of the mentioned positions and we will shortly refer some of them. It was clear since the beginning that different levels of allocation of tasks to machines affect differently the amount and quality of operative and managerial work (Bright, 1958), with potential *unemployment*.

Authors as Jaffe and Froomkin assumed that the continuity of technical innovation should not generate dramatic changes in the content and amount of *employment*. Rifkin, Forester, and De Masi argued that technological development carries large unemployment. Economists note that job losses are due not only to technological development but in larger proportion to international trade, to business process reengineering, and to reorganization. Evidences are reported that technical innovation creates new jobs and new enterprises with a potential gain in jobs. Sylos Labini (1998) maintains that a national GNP steady growth beyond 3.5% will reduce unemployment also during intense technological innovation.

The impact of automation on the *composition of the workforce* has been studied. Mallet (1975) found that the nature of the productive process, under conditions of semi-automation or automation, is the key for the potential development of technicians as a new working class. Kern and Schumann (1984) corrected the dequalification thesis, but reported about potential polarization of work force. This thesis returns with Rifkin (1995) who charged modern technology with the responsibility of changing human work, by encouraging the coming out of selected group of privileged individuals. Burris (1998) indicates that computerized work generates overall upskilling, some deskilling, and skill bifurcation; Barley and Gideon (2004) and Butera et al. (2008) found that the automation of work gave the starting signal to a real revolution, and from that time on a structural change has occurred in the workforce that can be identified with the development of the knowledge economy, where the actors are knowledge workers. Managers, professionals, and technicians are knowledge workers and their presence rapidly increased in Europe countries from 2000 to 2008. Basically, in countries as Belgium, Netherlands,

Germany, Sweden, Denmark, Finland, Slovenia, Norway, Czech Republic, Austria, United Kingdom, and France today there is not a huge difference in the number of knowledge workers and the rest of the employed population: knowledge workers are 45% of the working force. In other countries as Ireland, Italy, Hungary, Croatia, Greece, and Slovakia, knowledge workers are few, but they increased from 30% (2000) to 40% (2008).

Early researches on *skills* focused first on the change of the nature of tasks in automation. Empirical researches found that manipulation, traditional craft skills, dexterity, and operational ability (Bright, 1958) were fading away. On the contrary, tasks of supervision (Friedman, 1971), 'control skills' (Crossman, 1960), responsibility and relationship with the entire production system (Touraine, 1955), mental or visual jobs (Blauner, 1964), and sophisticated skills to respond to stochastic events (Davis and Taylor, 1976) were increasing.

Undesirable effects and *risks* of automation in dangerous processes have been studied as normal accidents (Perrow, 1984), human and organization errors (Reason, 1990), misinterpretation, and sense making of technological events (Weick, 1990). The Three Mile Island accident in 1979 is a typical example of the way automation can reduce situation awareness, therefore human operator has difficulty in having a proper understanding of the situation so that suitable oversight and interventions can occur.

Most early researches in automated plants found that *job design* criteria moved away from the taylor-fordist tradition, and that new criteria were emerging: interdependence of role and fewer levels of supervision, group behavior, multiskilling and shared responsibility, decline of strict job specifications and close supervision, communication, cooperation and influence, and group autonomy to cope with uncertainties. Recent developments of advanced computers enhance the face-to-face and remote group behavior, the human computer cooperation (Hoc, 2000), and, lately, the human–agent interaction (Lewis, 1998).

Automation changes the entire *role of the human operator* and requires proper knowledge and skills about the workplace in general and the production process in particular (Karsvall et al., 2003). Changes in technology (specifically the introduction of computers and ICT in general) are the main cause of increased demand for skilled workers. Problem solving, reflective skills and metacognitions, communication skills, self-regulations of motivations and affections, interpersonal skills and desire to interact by knowledge workers – but also the ability to use the social capital to fulfill personal desires and organization's needs – are new skills of knowledge workers (Kessels, 2001; Cedefop, 2010).

Conclusions – Effects or Design?

Technology in itself has direct influence on the physical universe of work (light, noise, heat, vibrations, radiation) and the operational universe of work (psychosocial conditions). Design of equipment has a major effect on (1) the amount of work per product unit (work displaced by technology); (2) the operational 'tasks' (man/machine task allocation); and (3) the coordination and control functions left to work system. Technology has only a composite impact with other factors on components of work systems such as roles, organizational structures, communications, work allocation systems and policies, work management systems (such as skills, pay, working hours, regulations), and in particular the cultural and value systems which govern cooperation and conflict (Butera, 1987).

Automation implies more than sheer displacement of operational human tasks because it includes the creation of new tasks of design, maintenance, and coordination, and generates new forms of organization of production and restructuring of the labor market. Automation is at the same time a process of social and technical change, a different distribution of knowledge and operational properties among human beings and technical equipment, and the development of advanced forms of sociotechnical systems in manufacturing and services.

Automation is in essence a sociotechnical system as a combination of business processes, technology, organization, and work according to Trist. Different patterns of supervisory control (Sheridan and Hennessy, 1984), human–computer interaction (Norman and Draper, 1986), and human–computer collaboration (Hoc, 2000) may make a difference in the same level of automation. Participation of stakeholders in the design of the system and of new business along with the process of introduction of the new technology may change the final pattern of sociotechnical systems (Davis and Taylor, 1976).

If technology upsets work, only design resets work. Social and technical methodologies take the opportunity of always existing 'room of maneuver' for planning, design, and experimentation of systems (such as territories, firms, factories, offices, professional units), which jointly may meet economic, social, and technical goals (Butera, 1987). In brief, social dimensions in technological development are not effects but instead matter of design.

Designing integrated architectures of sociotechnical systems, patterns of organization and work, and interaction among human beings and computers is a sociotechnical process and artifact and in principle should be matter of cooperation among different disciplines, as economic technology, organization science, sociology, and psychology. Since automation is matter of design, different practitioners should be involved as engineers, computer scientists, consultants, physicians, and many others. Automation involves different social actors as entrepreneurs, managers, legislators, government, and unions. In a word, development of automation should be a participative event.

Nothing is more matter of joint design of technical, economic, and social systems than the automation as we have represented it. But surprisingly, in the amazing history of automation, the sociotechnical approaches and methodologies have been in most cases excluded. The critical issue is that automation in different applications is mostly viewed as a sheer technological phenomenon. As a consequence, insufficient attention is paid to organizational and human dimensions, whose areas deserve an insignificant proportion of research and investments compared to research and investment on technology. The technological drive is so powerful and dominant that scientists and practitioners in technology are supposed to

be experts also in organization, social systems, roles and professions, and usability and are entitled to design sociotechnical systems, not calling in the research and design of other experts and practitioners.

As a result, a countless series of disasters and social disruptions came along the development of technical development: suboptimizations as the illusion of the unmanned factory, disasters as the Three Mile Island event, social disruptions as the case of Manhattan Highway designed by architect Moses (the omnipotent designer of New York) whose uncontrolled technical innovation transformed the Bronx from a middle-class nice neighborhood in a ghetto as described by Marshall Berman (1982). Berman compares the case of overpowered technologists to the case of the engineer Faust described by Goethe, betting his soul in the belief to be able to design alone a new huge technological settlement: but the resistance to leave their house from an old couple required the help of the devil and the consequent loss of Faust's soul.

On the contrary, many successful positive cases of intentional joint design performed with the cooperation of experts of different technical and social disciplines gave rise to extremely successful results in design of complex systems like civil aviation or healthcare, of products and platforms like Apple or Google, and of composite services like Ikea or Zara.

Acknowledgment

A first and different version of this paper is in Smelser N.J., Baltes P.B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences, Pergamon, Amsterdam–New York, 2001.

Professor of Sociology of Organization at University of Milan 'Bicocca'; President of Fondazione Irso-Institute for Action research on Organizations and Systems; Former Chairman of *Social Effects of Automation* Committee of *IFAC (International Federation of Automatic Control)*; Director of *'Studi Organizzativi.'* federico.butera@irso.it

A substantial contribution for the update of this article to non factory automation was given by Federica Persico, researcher and consultant at Fondazione Irso and PhD student at University of Milan 'Bicocca.'

See also: Science and Technology, Social Study of: Computers and Information Technology; Technology and Organization; Technology and Social Control; Technology, Evolution of.

Bibliography

Ackrill, M., Hannah, L., 2001. Barclays: The Business of Banking 1690–1996. Cambridge University Press.

Bagnara, S., Parlangeli, O., Tartaglia, R., 2010. Are hospitals becoming high reliability organizations? Applied Ergonomics 41, 713–718.

Barley, S.R., Gideon, K., 2004. "Gurus, Hired Guns and Warm Bodies: Itinerant Experts", *Knowledge Economy*. Princeton University Press, Princeton, NJ.

Batiz-Lazo, B., Wonglinpiyarat, J., Wood, D., 2002. Corporate Strategy in the UK Credit Card Market: The Case of Barclaycard. Manchester Business School Working Paper.

Berman, M., 1982. All that Is Solid Melts into Air. The Experience of Modernity. Simon & Schuster, New York (trad. it: *L'esperienza della modernità*, Il Mulino, Bologna, 1985).

Billings, C.E., 1997. Aviation Automation: The Search for a Human-centered Approach. Lawrence Erlbaum Associates Publishers, Abingdon.

Blauner, R., 1964. Alienation and Freedom: The Factory Worker and His Industry. University of Chicago Press, Chicago.

Braverman, H., 1974. Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century. Monthly Review Press, New York-London.

Bright, J.R., 1958. Automation and Management. Harvard University Press, Cambridge, MA.

Buckingham, W., 1961. Automation: Its Impact on Business and People. Harper & Brothers, New York.

Butera, F., 1987. Options for the future of work. In: Butera, F., De Martino, V., Koheler, E. (Eds.), Options for the Future. Coogan, London.

Butera, F., Thurman, J.E., 1984. An approach to automation and work design. In: Butera, F., Thurman, J.E. (Eds.), Automation and Work Design. North-Holland, Amsterdam.

Butera, F., Bagnara, S., Cesaria, R., Di Guardo, S. (Eds.), 2008. Knowledge Working. Lavoro, Lavoratori, Società Della Conoscenza. Mondadori Università, Milano.

Brown, J Seely, Duguid, P., 1991. Organizational learning and communities-of-practice: toward a unified view of working, learning, and innovation. Organization Science 2, 40–57.

Cedefop, 2010. New Skills for New Jobs: Action Now. Report by the Expert Group on New Skills for New Jobs, Prepared for the European Commission.

Consoli, D., 2005. The dynamics of technological change in UK retail banking services: an evolutionary perspective. Research Policy 34 (4).

Coyne, P., 1994. Roadcraft: The Police Driver's Handbook. HMSO, London.

Crossman, E.R.F.W., 1960. Automation and Skill. HMSO, London.

Danziger, J.N., 1985. Social science and the impact of computer technology. Social Science Quarterly 66, 3–21.

Davis, L.E., Taylor, J.C., 1976. Technology, organization and job structure. In: Dubin, R. (Ed.), Handbook of Work, Organization and Society. Rand McNally, Chicago.

Diebold, J., 1952. Automation: The Advent of Automated Factor. Van Nostrand, New York.

Endsley, M.R., Garland, D.J. (Eds.), 2000. Situation Awareness Analysis and Measurement. Lawrence Erlbaum Associates, Mahwah, NJ.

Fincham, R., Fleck, J., Procter, R., Scarbrough, H., Tierney, M., Williams, R., 1994. Expertise and Innovation: Information Technology Strategies in the Financial Services Sector. Clarendon Press, Oxford.

Freeman, C., 1998. Unemployment and Long Waves of Economic Development. Pinter, London.

Friedman, 1971. A non cooperative equilibrium for supergames. Review of Economic Studies 28, 1–12.

Gallino, L., 2006. Dizionario di Sociologia. Utet, Torino.

Haigh, T., 2006. Remembering the office of the future: the origins of word processing and office automation. IEEE Annals of the History of Computing.

Helander, M., 1988. Handbook of Human–Computer Interaction. North Holland, Amsterdam.

Hoc, J.M., 2000. Human–Machine Interaction to Human–Machine Cooperation. Taylor & Francis Ltd., London.

Karmarkar, U., 2004. Will you survive the services revolution? Harvard Business Review 82.

Karsvall, A., Frohm, J., Hassnert, M., et al., 2003. On decision-making: making sense of system performance. In: Proceedings of EAM, the 22th European Conference on Human Decision Making and Manual Control, pp. 1–8.

Kern, H., Schumann, M., 1984. Das Ende der Arbeitsteilung. Rationalisierung in der industriellen Produktion. C.H. Beck Verlag, Munchen.

Kessels, J.W.M., 2001. Learning in organizations: a corporate curriculum for the knowledge economy. Futures 33, 497–506.

Kusterer, K.C., 1978. Know-how on the Job: The Important Working Knowledge of "Unskilled" Workers. Westview Press, Boulder, CO.

Lewis, M., 1998. Designing for human–agent interaction. AI Magazine 19 (2).

Mallet, S., 1975. The New Working Class. Spokesman, Nottingham.

Mandell, L., 1990. The Credit Card Industry: A History. Twayne Publishers, Boston.

Meissner, M., 1969. Technology and the Worker. Technical Demands and Social Processes in Industry. Chandler Pub. Co, S. Francisco.

Metzger, U., Parasuraman, R., 2005. Automation in future air traffic management: effects of decision aid reliability on controller performance and mental workload. Human Factors 47, 35–49.

Morris, T., 1986. Innovations in Banking. Croom Helm, London.

Nonaka, I., Takeuchi, H., 1995. The Knowledge Creating Company. Oxford University Press, New York.

Norman, D.A., Draper, S.W. (Eds.), 1986. User Centered System Design: New Perspectives on Human Computer Interaction. Lawrence Erlbaum Associates, Hillsdale, N.J.

Perrow, C., 1984. Normal Accidents: Living with High-Risk Technologies. Basic Books, New York.

Reason, J., 1990. Human Error. Cambridge University Press, Cambridge, UK.

Rifkin, J., 1995. The End of Work: The Decline of the Global Labour Force and the Dawn of Postmarket Era. Putnam, New York.

Rovira, E., McGarry, K., Parasuraman, R., 2007. Effects of imperfect automation on decision making in a simulated command and control task. Human Factors 49, 76–87.

Sarter, N.B., Schroeder, B., 2001. Supporting decision making and action selection under time pressure and uncertainty: the case of in-flight icing! Human Factors 43, 573–583.

Shappell, S.A., Wiegmann, D.A., 2000. The Human Factors Analysis and Classification System – HFACS (DOT/FAA/AM-00/7). Office of Aerospace Medicine, Federal Aviation Administration, Washington, DC.

Sethumadhavan, A., 2011. Effects of Level of Automation on Air Traffic Controller Situation Awareness of Performance. Ph.D. dissertation. Texas Tech University.

Sheridan, T.B., Hennessy, R.T. (Eds.), 1984. Research and Modeling of Supervisory Control Behavior. National Academy Press, Washington, D.C.

Sylos Labini, P., 1998. "Nuove Tecnologie e occupazione", Sviluppo Tecnologico e disoccupazione. Convegno Accademia dei Lincei.

Special Annual Report "Manufacturing and Innovation" The Third Industrial Revolution, April 21, 2012. The Economist.

Thompson, J.D., 1967. Organizations in Action. McGraw-Hill, New York.

Touraine, A., 1955. In: L'évolution du travail ouvrier aux usines Renault. CNRS, Paris.

Trist, E., Murray, H., 1993. The Social Engagement of Social Science, vol. II. University of Pennsylvania Press, Philadelphia.

Weick, K.E., 1990. Technology as equivoque: sensemaking in new technologies. In: Goodman, P., Sproull, L.S. (Eds.), Technology and Organizations. Jossey Bass, San Francisco.

Woodward, J., 1965. Industrial Organization: Theory and Practice. Oxford University Press, London.

Zuboff, S., 1988. In the Age of the Smart Machine: The Future of Work and Power. Basic Books, New York.

Zysman, J., Feldman, S., Murray, J., Nielsen, N.C., Kushidai, K., 2010. Services with Everything: The Digital Transformation of Services from Economic Sinkhole to Productivity Driver. BRIE Working Paper 187a: 1–53.

Automobility

Tim Schwanen, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term 'automobility' is often used to denote the practices, materialities, institutions, and landscapes centered on the private car. This article introduces recent thinking on automobility in sociology, human geography, political science, and other fields. It describes different conceptualizations of automobility as a system and a regime, reviews studies of how cars are used and experienced by their users, and discusses the future of automobility. It cautions against too much optimism about the imminent demise of automobility but also argues that the possibility of radical change over the next three decades should not be disregarded.

Introduction

The term 'automobility' appears fairly straightforward at first sight, presumably referring to automobiles and movement. Indeed, automobility is often used as shorthand for the practices, materialities, institutions, and landscapes centered on the private car. The term is nonetheless imprecise, and multiple ambiguities can be identified. There is the question whether the prefix 'auto' refer to 'self' as in ancient Greece, or to 'autonomy' as is often argued (Urry, 2004; Paterson, 2007). Likewise, does mobility refer to an artifact that moves, to the human(s) who move(s), or both? The conjunction of 'auto' and 'mobility' multiply ambiguities. If the emphasis is on the artifact and 'auto' refers to self, attention will almost certainly be directed toward self-propelled vehicles like the private car. The combination of artifact and autonomy implies a wider focus – any artifact that affords autonomy in and through movement, including the bicycle, wheelchair, or airplane, is brought into focus. If the emphasis is on the moving human(s), automobility can also refer to walking as both a self-propelled bodily practice and as providing and expressing freedom. Finally, if the emphasis is on both human and artifact, automobility refers to the hybrid ensemble of elements – human bodies/minds, machines, surfaces, signs, values, and so forth – through which movement, and especially autonomous movement, is enacted (Sheller and Urry, 2000).

This last understanding of automobility usefully highlights the complex networks of relations, heterogeneous elements, and practices through which different forms of mobility are brought into being: autonomous mobility is not a simple function of free will – as naïve versions of (neo)liberalism would have it – but a thoroughly relational outcome. This way of thinking also dispels the myth that the machine of the private car is unique in affording autonomy to people. The bicycle, boat, and airplane could, at least in theory, enable similar levels of autonomy. What is unique to the private car is the scale and complexity of the ensembles that make car transport possible, as well as the extent to which (liberal) values like freedom, control, and independence have become inscribed into this artifact. Culture and politics are thus inextricably linked to the car.

Although the concept of automobility is not a priori limited to car-facilitated movement, social scientists have privileged the automobile over other modes of transport in theoretical and empirical researches. This preoccupation reflects not simply that the car is the most widely used form of transport across the global North. It is also because the rise of the car and the ensembles that have made its proliferation possible are as defining for modernity as is urbanization (Sheller and Urry, 2000). Three interrelated concerns in the social science literature about automobility pertain to how hybrid ensembles organized around the private car are best conceptualized, how car-facilitated movement is lived, and how automobile-dominated mobility systems will develop in the future.

System or Regime?

Arguably the most well-known theorization of automobility has been offered by John Urry. Inspired by complexity theory he conceives of automobility as "a self-organizing autopoietic, non-linear system that spreads worldwide" (2004, page 27), comprising a wide range of elements. The car, according to Sheller and Urry (2000), is:

- An object manufactured by one of the most powerful economic sectors within the capitalist system of the twentieth century and from which key concepts for understanding the development of capitalism (e.g., Fordism) have emerged.
- The second largest item of consumption (after housing), which provides status to its owner/user through associated values such as speed, safety, flexibility, freedom, family, and masculinity, and that is a major object of desire, affection, and crime.
- A powerful complex linked to other sectors, such as the oil sector, road construction, hotel sector, car sales/repair, suburban housing, shopping malls, advertising, and urban planning.
- A form of mobility that subordinates other transport modes and determines how, when, and where people engage in activities.
- The dominant culture that feeds into major discourses about well-being and citizenship and provides a rich source of inspiration for literature, film, and other arts.
- A key user of nonrenewable environmental resources and major cause of emissions and pollution.

For Urry the system of automobility is autopoietic because it produces the preconditions of its own expansion. Crucial to how it generates the need for more cars is the ways in which it restructures space–time as part of its development. Consider, for instance, the critical role of the private car in enabling the suburbanization of households, services, and employment, which in many places across the global North has marginalized public transport, walking, and cycling. Consider also how cars have, in Urry's terms, coerced individuals into flexibility: the latter are expected to lead a 'fast,' fragmented, and individualized life in which employment, family life, community participation, leisure, and mobility are seamlessly combined into a do-it-yourself biography.

The system of automobility is nonlinear because small changes can have big effects. One or multiple coincidences in particular places at specific times can set in motion a train of events with profound consequences. In the context of automobility, key coincidences were the activities of early car enthusiasts around 1900, who perfected the then recently invented car by adding lights, electric starters, trunks, and so forth; the victory of a petrol-fueled car over steam and electric alternatives in a horseless carriage competition in Chicago in 1896; and the delivery of the first Ford Model T in 1908 (Urry, 2007). Over time and through autopoiesis, many locales – first in the United States, then in Western Europe and the rest of the global North, and now in the global South – have become locked into automobility: particular technologies (e.g., the internal combustion engine (ICE)), infrastructures (e.g., motorways, gas stations), practices (e.g., driving to work, touring among tourist attractions on holidays), meanings (e.g., driving as freedom), and institutions (e.g., automobile associations, fuel taxes) have become deeply embedded. Some locales – especially in North America – have come to experience 'hyperautomobility,' a situation whereby the distances that individuals travel by car continue to grow when car ownership has reached saturation levels (Freund and Martin, 2009).

Insights from innovation studies align with Urry's ideas. Geels and colleagues (2012) conceptualize automobility as a sociotechnical system – a set of interdependent technologies, infrastructures, markets, user practices, cultural meanings, politics, and technoscientific knowledge. Central to this system is the sociotechnical regime, or the set of rules that orient and coordinate the activities of the social groups that reproduce the sociotechnical system of automobility. Examples of such rules are cognitive routines and shared beliefs, competences, lifestyles, institutions, regulations, and contracts. Innovation scientists like Geels seek to understand the dynamics in automobility and other systems. They hold that sociotechnical systems are relatively stable and due to multiple lock-in mechanisms tend to change only incrementally. Incumbent car manufacturers, for instance, are often hesitant to invest in radical inventions because they may render core competencies and existing production facilities superfluous, demand large sunk investments in new facilities, and may not appeal to potential consumers. This is an important reason why the attempts of EU regulation to stimulate radical innovations in car engines since 1990 have been (mis)used by car manufacturers to stimulate incremental change in ICEs (Dijk and Yarime, 2010). This risk-aversive attitude is reinforced by the practices of banks and other capital investors, and also by governments. While sustainability, and more recently low-carbon mobility, have become important elements in policy discourses across the Western world since 1990, policymakers and politicians are not external to the system or even at its helm: they are often effectively lobbied by the automotive industry; they are reliant on the revenues from fuel, car, and road use taxation; they work with policy instruments that favor incremental over radical change and are dependent on research and expertise that reflects the built-in logic of the existing system; and they are 'acculturated' into status quo forms of knowledge and thinking through training and education.

Radical change in a sociotechnical system like automobility is nonetheless possible. This can occur if and when pressures in the 'sociotechnical landscape' destabilize the regime and 'niche innovations' have built up sufficient momentum to displace or reconfigure the existing sociotechnical regime (Geels et al., 2012). The sociotechnical landscape is the external environment in which a regime is situated and over which regime actors have limited influence; it comprises macrolevel developments in demographics (e.g., population aging across most of the global North), economics (e.g., the rise of neoliberal capitalism), and such processes as anthropogenic climate change. Niches are protected spaces – often market niches or demonstration projects – where inventions are tested and developed and where users have specific demands. In the context of automobility, a range of niches have emerged over the past decades in many cities across the Western world (Geels et al., 2012; see also Nykvist and Whitmarsh, 2008):

- Intermodal travel: initiatives to integrate rail, bus, taxi, car, and walking/cycling – park-and-ride schemes are a well-known example.
- Cultural and sociospatial niches that challenge the normality of automobility through land use (transit-oriented development, compact city) or collective ownership, as with car/bike-sharing programs.
- Mobility management policies by (local) governments, public transport providers, and employers who seek to encourage alternative mobility practices through personalized travel planning, information provision, awareness campaigns, and so forth.
- Public transport innovations through new infrastructure, better information provision, and modernized vehicles – French and Swiss cities (e.g., Strasbourg, Zurich) provide oft-cited examples.
- Information and communication technologies that increase transport's efficiency (as with intelligent transport systems) and substitute for physical movement (as e-commerce does).
- Green propulsion technologies, such as hybrid and battery electric vehicles (BEVs).

The enthusiasm of many activists and scholars notwithstanding, these niches have built up only limited internal momentum to date. They have so far been unable to overhaul the system of automobility built around the ICE, even in countries like the Netherlands or Denmark and cities like London, Zurich, or San Francisco. For Geels and colleagues the regime dominating the system of automobility has become somewhat weaker in countries like the Netherlands and the UK since 1990 – particularly in (the central parts of) cities and among young adults – but is also remarkably resilient. So far

the landscape pressures of climate change, oil scarcity, and the rise of the Internet have been counteracted by such trends as the cultural preference for individual ownership; the cultural valorization of speed and time efficiency; individualization; and the growing markets for conventional cars in emerging countries like China, India, Brazil, Mexico, Nigeria, and Ghana.

The conceptualizations of automobility by Urry and to a lesser extent Geels and colleagues can be, and have been, criticized. They are universalist insofar as they downplay the variations in automobility or rather automobilities. These vary significantly in geographical terms (Edensor, 2004; Merriman, 2009): not only does the power of the automobile lobby differ significantly across Western countries, but landscapes of automobility – road signs, the architecture of petrol stations and roadside buildings, type of cars driven, as well as traffic rules and driving styles – differ markedly across nation-states to such an extent that they contribute to a state's identity. Driving a car along US Route 66 is not quite the same as driving on a German Autobahn or in one of India's many distinctive roadscapes. Another criticism is that system thinking "tends to underplay collective human agency in the production of automobility and to avoid political questions about the shaping of the automobile 'system' " (Böhm et al., 2006, page 5). Rather than being a self-organizing system in a state of lock-in, automobility should be thought of as continually rebuilt through human actions and as profoundly political and suffused with power relations (Goodwin, 2010).

To emphasize both the systemic and political dimensions of automobility, Paterson and colleagues prefer the term 'regime' over 'system' (Böhm et al., 2006; Paterson, 2007), but they use this term in a radically different sense from Geels and colleagues. For Paterson (2007), automobility refers to the cultural, political, and economic institutions, norms, and practices that sustain and promote car-dependent transport and its centrality to global capitalism, environmental degradation, socioeconomic injustice, and the formation of subjectivities – particular ways of being and becoming for human subjects. Böhm et al. (2006) draw on Foucault's philosophy to understand automobility as a regime of truth, power, and subjectivity.

It is a regime of truth because the automobile industry and lobby, car drivers, politicians and policymakers, and scholars tend to position cars as:

- Efficient – their use tends to involve less effort compared to alternative modes of transport.
- Convenient – they enable smooth transfer from A to B whenever and wherever people like.
- Cheap – the costs are, or seem to be, lower than for alternatives.
- Stylish and progressive – they enables users to express and construct certain identities and associate them with modernity.
- Democratic and liberating – they level differences among people and offer freedom and autonomy.

Automobility is also a regime of power as it implies a particular governmentality, or style of governing practices and individuals (Merriman, 2007). A wide range of regulations and institutions have emerged to monitor and discipline drivers and other road users, and to mitigate the negative externalities of car use, such as its environmental degradation, congestion, and traffic accidents. In effect, the rise of automobility has expanded the state's power. Automobility also has geopolitical dimensions. Securing reliable and reasonably cheap access to the oil that keeps cars moving was a significant factor in the decision making of the United States and other Western states that culminated in the Gulf Wars (Goodwin, 2010) and is one reason why the Dutch and UK governments more or less condone Shell's detrimental impact on the environment in the Niger delta and Ogoni livelihoods and quality of life in southeastern Nigeria (Mugveni and Engler, 2011). Yet, the governing of practices and individuals is also undertaken by nonstate actors, and car advertising is key here (Paterson, 2007). In the United States such advertising connected particular car types and brands, such as sports utility vehicles and Hummers, with conquering nature and particular forms of masculinity up to at least the late 2000s. Since then other discourses, which were already prominent in Western Europe, have become more important in US car advertising. These discourses emphasize concern for the environment rather than the conspicuous consumption of resources and have been drawn upon to market smaller city cars and hybrid vehicles, such as the Toyota Prius (Goodwin, 2010).

Because these changes in advertising imply that drivers come to be understood in different ways, they also amount to a shift in automobility as a regime of subjectivity: they assemble possible ways of being for car drivers in new ways. The extent of change should not be exaggerated, however. On the whole automobility continues to position drivers as free and autonomous individuals making their own informed choices. For Böhm et al. (2006), this positioning not only couples automobility to individualism and neoliberalism but also normalizes driving as something that normal people do. In this way, traveling by other means of transport is constructed in some places as a form of deviance. This marginalization of non-car transport is exacerbated by the ways in which automobility's regimes of truth, power, and subjectivity reinforce each other. Through myriad actions of governments, car manufacturers, drivers, and many other agents, the dominance of automobility is continually maintained and reconfigured in ways that are only partially captured by complexity theory with its roots in the natural and exact sciences.

Living Automobility

Thinking about automobility at the macrolevel of systems and regimes is important and useful but also only one side of the coin. It needs to be complemented by, and tied to, micro level analyses of what people actually do with cars – how automobiles are used and experienced and what effects are generated. Studies of car advertisements showing which values and meanings are inscribed into cars are very insightful, but drivers are no passive receptacles slavishly absorbing the messages by car manufacturers, governments, or other stakeholders. Hagman (2010) has shown that whereas understandings of driving pleasure as articulated in Swedish ads and magazines emphasize driving pleasure as inhering in the car and as

a function of its power, speed, and drivability, drivers themselves consider the context of driving – the weather, behavior of other drivers, the destination, and so forth – critically important. This difference seems to reflect that the drivers refer to their everyday experience where the car is a utility and its use tightly regulated, whereas the ads and magazines seek to compensate for triviality of driving and keep alive the ideology of driving as freedom.

The factors that motivate individuals to drive have been studied extensively by transport scholars, psychologists, economists, geographers, and sociologists who have drawn on a wide range of theoretical frameworks, including utility theory, attitude theories, time geography, and theories of social practices. A full review is beyond the scope of this article but available in Lucas et al. (2011). Suffice it to say that, as a result of engineering and economics dominating academic thinking on the topic, research has long privileged instrumental factors in the explanation of why, how, when, and where individuals use cars. Compared to alternative means of transport, cars are preferred, the argument goes, because they waste less time on movement and because they are cheaper, more convenient and comfortable, always and (almost) everywhere available, and safer. More recently, psychologists, geographers, and sociologists have shown that cars have also symbolic and affective values. A number of studies have shown how many suburban mothers use cars to create the (middle-class) identity of a 'good' mother – someone who continually strives to do what is best for her children – for themselves by engaging in particular mothering and mobility practices (e.g., Dowling, 2000). Cars enable those mothers to drive their children to schools, nurseries, and extracurricular activities that need not be the nearest available but instead are the 'best match' with parental aspirations and perceptions about their children's needs. Unlike other transport modes, cars are also believed to be unique means to protect children from traffic injury and 'stranger danger' and to spend quality time with one's children. For many mothers, the car trip is not simply about getting from A to B; it is replete with emotion and meaning and a space–time of care and sociability.

Another social group for whom driving is often more than moving from origin to destination is older adults. Studies from across the Western world – though especially from regions and countries where the marginalization of non-car transport is particularly profound such as the United States and Australia – suggest that older people often experience driving cessation as a loss and a rite of passage into deep old age (Schwanen et al., 2012). For many older adults in the global North, being able to drive in later life is a sign and expression of self-reliance and autonomy, enabling them to avoid the stigmas of dependency and passivity and to enact the neoliberal, ageist ideal of the ageless self.

In short, cars are often more than means for getting around. Apart from being implicated in identity construction, they are sites where drivers undertake activities as diverse as paid work, educating children, and building up social relationships (Laurier et al., 2008), and where distinctive affective bonds are forged between man and machine (Sheller, 2004; Thrift, 2004). Sheller (2004) speaks of "automotive emotions" – visceral and other feelings associated with car use and the embodied dispositions of car users. The idea is that below the threshold of consciousness the interactions between driver, car, and the wider computerized networks in which the latter are increasingly enrolled produce a deep-seated attunement to cars and specific ways of sensing and inhabiting the world. A specific type of cyborg (cybernetic organism) – the driver-car (Dant, 2004) – seems to have come into being, and this human–machine hybrid may be less responsive to policy and other stimuli (pricing, taxation, information campaigns, and so forth) attempting to limit car use than mainstream academic and policy thinking assumes. If this is true, then these affective processes will trap drivers even further in automobility, deepen the marginalization of other forms of transport, and lower the odds that sociotechnical niches other than that of green auto-propulsion technologies will durably reconfigure the system of automobility.

Cars' symbolic importance and automotive emotions may also weaken users' sensitivity to many negative externalities of automobility and the extent to which it puts others at risk. Those others include humans who for some reason (e.g., lack of financial resources, disability, age) are excluded from driving; pedestrians, cyclists, children playing in the street, and others who may be harmed in traffic accidents involving cars; persons with respiratory diseases; people in parts of the world that are of geopolitical importance to the supply of oil; future generations across the world; and many of the forms of nonhuman life inhabiting Earth. There is a degree of multiple deprivation here insofar as households excluded from driving are more likely to live in unhealthier locations where access to non-car modes of transport is poor and traffic safety is low (Martin, 2007). Car use and automotive emotions also put users themselves at risk, by contributing, for example, to sedentary lifestyles.

The Future of Automobility

Given automobility's many 'costs,' it is tempting to predict that automobility must somehow at some point dissolve. The inherent fragility of automobility (Böhm et al., 2006) reinforces this temptation, as a range of contingencies reproduce automobility, including its ability to secure access to oil; to keep congestion, emissions and traffic injury within bounds of acceptability; to perpetuate the idea that cars provide freedom and autonomy to individuals in ways that no other forms of transport can; and so forth. Noteworthy is also that growing numbers of local governments in Europe and the United States have committed themselves to the aforementioned sociotechnical niches and have begun to govern automobility in different ways: plans to construct roads for cars are increasingly shelved and roadspace previously devoted to cars is reallocated to public transport, cycling, and walking.

The temptation becomes even greater when one appreciates that the trend toward 'hyperautomobility' is being eclipsed by the recent phenomena of 'peak travel' (Millard-Ball and Schipper, 2011) and 'peak car use' (Newman and Kenworthy, 2012) – a modest decline, stabilization, and/or slowed-down growth in the demand for car travel at the population level since 1990 or thereabouts in North America, Western Europe, Australia, and Japan. The car continues to be the dominant

mode of transport for daily activities (car use for tourism tends to be excluded in peak travel research!), but among young adults – and especially young men – holding a driver's license, car ownership and the per capita distance by car have been decreasing (Kuhnimhof et al., 2012). Newman and Kenworthy (2012) argue that the reversal of the long-term trend toward growing car use reflects six processes:

- In many countries people now spend so much time in cars that large shares of the population have reached some sort of biologically determined, maximum tolerance level regarding mobility.
- A halting of urban sprawl – the outward growth of urban areas in the form of auto-oriented, low-density, and mono-functional developments.
- A revival of public transport in cities and urban regions.
- Population aging, which means that the share of car commuters in the overall population may be stabilizing.
- Emerging cultures of urbanism, which means that more younger adults remain in the city than previously and perhaps also that more retired households exchange their suburban house for city living.
- The increase in fuel prices, particularly since 2000.

Newman and Kenworthy's list can be challenged. Individual factors may be less or more important in particular localities, and others might be added (e.g., the post-2008 economic downturn, the waning influence of the car industry). There are also broader concerns with peak car thinking. One is the possibility that analysts read too much into 'peak travel' and that growth will pick up in the longer run; another is the narrow focus on the global North. The scale of automobility's expansion is unprecedented in newly industrializing countries (especially China and India) and more recently also in parts of Sub-Saharan Africa. Automobility is now truly globalizing, and it is at the global scale that its system and regimes must be examined and conceptualized. On the other hand, peak car use is a significant phenomenon showing that growth through autopoiesis in the mature systems of automobility of the global North can no longer be taken for granted.

Further to its rapid expansion beyond the Western world, there are other reasons to temper expectations about the imminent dissolution of automobility. Returning to the innovation studies perspective on automobility as a sociotechnical system, we can say that pressures on the regime are increasing but regime actors also adjust to these. Governments and industries have implemented, and keep adopting, a range of measures to mitigate resistance and the negative externalities of widespread car use, but most of these do not profoundly alter the system's basic architecture. Following Unruh (2002), we can say that end-of-pipe solutions and continuity measures tend to be favored. The former treat externalities and are exemplified by catalytic converters and more recently initiatives to increase the fuel economy of conventional vehicles. The latter are incremental innovations and include hybrid and electric vehicles as well as congestion pricing schemes as in London. Such schemes are often positioned as radical innovations – or discontinuity measures (Unruh, 2002) – by transport scholars, but actually curb some of automobility's excesses instead of bringing about its complete dissolution. Potentially discontinuous innovations are sometimes promoted by governments but difficult to implement effectively and often unlikely to result in the creation of true (localized) alternatives to automobility. A case in point is the set of integrated transport and compact city policies that the Dutch National government sought to implement in the late 1980s and 1990s (Schwanen et al., 2004). The growth in car use in the Netherlands may have been tempered to some extent, but other elements of the policy package – most notably a national road pricing scheme – failed to materialize.

Another reason for tempered expectations is that there is no realistic alternative to the ICE-powered car. For Urry (2004, 2011), the lock-in effects of automobility are irreversible, and a successor system will have to offer at least the same level of flexibility in the scheduling and conduct of everyday activities and the same sense autonomy as the private car. Viable regimes of truth, power, and subjectivity may emerge around public transport, cycling, and walking but only in specific pockets within cities. Those regimes cannot be generalized across space as New Urbanists in the United States and compact city adherents in Europe sometimes do. Automobility's spatial legacy in the form of sprawling low-density developments and counterurbanization will necessitate some form of motorized individualized mobility in most locales. At this time the BEV is considered the most likely successor to the conventional car. Yet, even under the most favorable assumptions "BEVs may only provide a niche market over the next 20 years" (Tran et al., 2012, page 332) and major uncertainties about their diffusion remain. Biofuels and other advanced fuels may turn out to be stiff competitors, and large-scale adoption of BEVs may require changes to electricity generation and distribution networks, which may dampen further diffusion. Additionally, "battery storage will need to increase by a factor of four while costs need to drop 70% over the next ten years" (Tran et al., 2012, page 331) for mass diffusion to become a reality. Cultural change is also required: drivers' expectations about how far they can drive without charging need to be lowered and 'range anxiety' – the fear of becoming stranded before the destination is reached – needs to be overcome.

Yet, a glass half full is also a glass half empty. Urry (2011) argues that at some point in the future and in particular places electric vehicles may become fashionable and then 'go viral.' For him, automobility might be profoundly reconfigured if mass diffusion of electric vehicles is entwined with a normalization of carbon budgeting and digital surveillance technologies that allow practically all vehicles and individuals to be traced, trends toward urban compaction and deprivatization, and a weakening of discourses that position mobility as freedom. Perhaps newly industrializing countries like China provide the geographical context where these processes can unfold first given that they are less encumbered by the historical legacy of the ICE than is the global North (Tyfield and Urry, 2012).

It is difficult to predict the future of automobility. Its trajectory may become increasingly geographically differentiated: large differences between the global North and South and within each of these mega-regions can be anticipated. However the system, regimes, and lived experience of automobility may change, there will be winners and losers (Cohen, 2012): the structure of urbanization patterns, the fortunes of firms and lobby organizations, the nature of governmental institutions,

the meanings and emotions associated with mobility, and culturally inflected aspirations of what counts as a good life will change. All of these may induce forms of conflict that merit attention from social scientists.

See also: Cultural Geography; Eco-cities and Sustainable Urbanism; Innovation; Mobility and Travel Activity Patterns; Organizational Path Dependence; Transportation Geography; Transportation Planning; Urban Sprawl.

Bibliography

Böhm, S., Jones, C., Land, C., Paterson, M., 2006. Introduction: impossibilities of automobility. The Sociological Review 54 (S1), 1–16.

Cohen, M.J., 2012. The future of automobility society: a socio-technical transitions perspective. Technology Analysis & Strategic Management 24 (4), 377–390.

Dant, T., 2004. The driver-car. Theory, Culture & Society 21 (4/5), 61–79.

Dijk, M., Yarime, M., 2010. The emergence of hybrid-electric cars: innovation path creation through co-evolution of supply and demand. Technological Forecasting & Social Change 77 (8), 1371–1390.

Dowling, R., 2000. Cultures of mothering and car use in suburban Sydney: a preliminary investigation. Geoforum 31 (3), 345–353.

Edensor, T., 2004. Automobility and national identity: representation, geography and driving practice. Theory, Culture & Society 21 (4–5), 101–120.

Freund, P., Martin, G., 2009. The social and material culture of hyperautomobility: "hyperauto". Bulletin of Science, Technology & Society 29 (6), 476–482.

Geels, F.W., Kemp, R., Dudley, G., Lyons, G. (Eds.), 2012. Automobility in Transition? A Socio-Technical Analysis of Sustainable Transport. Routledge, New York.

Goodwin, K.J., 2010. Reconstructing automobility: the making and breaking of modern transportation. Global Environmental Politics 10 (4), 60–78.

Hagman, O., 2010. Driving pleasure: a key concept in Swedish car culture. Mobilities 5 (1), 25–39.

Kuhnimhof, T., Armoogum, J., Buehler, R., et al., 2012. Men shape a downward trend in car use among young adults – evidence from six industrialized countries. Transport Reviews 32 (6), 761–779.

Laurier, E., Lorimer, H., Brown, B., et al., 2008. Driving and 'passengering': notes on the ordinary organization of car travel. Mobilities 3 (1), 1–23.

Lucas, K., Blumenberg, E., Weinberger, R. (Eds.), 2011. Auto Motives: Understanding Car Use Behaviours. Emerald, Bingley.

Martin, G., 2007. Global motorization, social ecology and China. Area 39 (1), 66–73.

Merriman, P., 2007. Driving Spaces: A Cultural-Historical Geography of England's M1 Motorway. Blackwell, Oxford.

Merriman, P., 2009. Automobility and the geographies of the car. Geography Compass 3 (2), 586–599.

Millard-Ball, A., Schipper, L., 2011. Are we reaching peak travel? Trends in passenger transport in eight industrialized countries. Transport Reviews 31 (3), 357–378.

Mugveni, B., Engler, I., 2011. Stop Signs: Cars and Capitalism, on the Road to Economic, Social, and Ecological Decay. RED Publishing, Vancouver.

Newman, P., Kenworthy, J., 2012. 'Peak car use': understanding the demise of automobility dependence. Available at: http://www.parliament.tas.gov.au/ctee/Council/Submissions/Professor%20Newman%20submission%2014%20Aug%2012.pdf.

Nykvist, B., Whitmarsh, L., 2008. A multi-level analysis of sustainable mobility transitions: niche development in the UK and Sweden. Technological Forecasting & Social Change 75 (9), 1373–1387.

Paterson, M., 2007. Automobile Politics: Ecology and Cultural Political Economy. Cambridge University Press, Cambridge.

Schwanen, T., Banister, D., Bowling, A., 2012. Independence and mobility in later life. Geoforum 43 (6), 1313–1322.

Schwanen, T., Dijst, M., Dieleman, F.M., 2004. Policies for urban form and their impact on travel: the Netherlands experience. Urban Studies 41 (3), 579–603.

Sheller, M., 2004. Automotive emotions. Theory, Culture & Society 21 (4/5), 221–242.

Sheller, M., Urry, J., 2000. The city and the car. International Journal of Urban and Regional Research 24 (4), 737–757.

Thrift, N., 2004. Driving in the city. Theory, Culture & Society 21 (4/5), 41–59.

Tran, M., Banister, D., Bishop, J.D..K., McCullogh, M.D., 2012. Realizing the electric-vehicle revolution. Nature Climate Change 2, 328–333.

Tyfield, D., Urry, J., 2012. Greening China's 'Cars': Could the Last Be the First? CeMoRe Working Paper. Lancaster University. Available at: http://www.lancs.ac.uk/staff/tyfield/GreeningChinaCars_CeMoReWorkingPaper.pdf.

Unruh, G.C., 2002. Escaping carbon lock-in. Energy Policy 30 (4), 317–325.

Urry, J., 2004. The system of automobility. Theory, Culture & Society 21 (4/5), 25–39.

Urry, J., 2007. Mobilities. Polity, Cambridge.

Urry, J., 2011. Climate Change & Society. Polity, Cambridge.

Relevant Websites

http://www.cosmobilities.net/ – Cosmobilities Network.

http://www.transitionsnetwork.org/ – Sustainability Transitions Research Network.

Autonomous Agents

Tom Bösser, University of Münster, Münster, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Autonomous agents are software programs that respond to states and events in their environment independent from direct instruction by the user or owner of the agent, but acting on behalf and in the interest of the owner. The term agent is not defined precisely, and agent software can range from simple programs composed of a small number of rules to large and complex systems, notably multiagent systems, which comprise many interacting agents. Agent technology was developed in artificial intelligence (AI) research, and can include AI techniques such as learning or reasoning. Agent applications automate aspects of human performance and support the human user by continuous data analysis, monitoring of data streams and large databases, and where routine reactions to events are required. Often these are related to user interaction, technical control systems and electronic trade. Multiagent systems are also used as simulation models for systems with many actors, for example financial markets or traffic. A weak position sees agents as tools that relieve humans from routine tasks suited to computational solutions, while a strong position asserts that agents can be constructed to mimic or even surpass cognitive functions of the human.

What Are Autonomous Agents?

Agents are software programs that:

1. sense events and analyze data streams in their environment,
2. react to certain events and states with defined responses,
3. exhibit goal-oriented behavior representing the intentions of the owner or user of the agent,
4. act proactively on their own initiative, and
5. function in a temporally continuous manner.

Agents may include cognitive functions such as (1) learning, most often adaptation, the simplest form of learning, based on statistical data analysis and observation of process parameters (potentially more complex forms of learning can be included), or (2) communication with other agents or humans, including the exchange of data, parameters, and goal functions.

The autonomy of agents is limited by the interests of the owner. The degree of autonomy ranges from task-specific agents with a restricted domain of application to virus-like entities that proliferate widely and are not under control by their owner, once activated.

Distinction between Robots and Agents

Autonomous agents and robots are closely related: agents are software programs only, whereas robots comprise functions implemented in hardware. Robot behavior is optimized by tuning sensors, effectors, and the control program; agents are composed of software functions for data analysis and communication, and a control program. Robot behavior must also be optimized with regard to energy and physical performance parameters, which are less important for agents. There is a tendency to use the terms software robot and agent synonymously.

Agents are distinct from software programs in general by their continuous operation, their reactive behavior, and their (partial) autonomy. It should be noted that feedback control systems also have some of these properties but are generally associated with the monitoring of continuous variables and the control of set points or simple forcing functions (e.g., the thermostat).

Constructing Agents

Agents are specific to their tasks and their environment. The design of a specific agent is optimized for its application to maximize performance and reliability, and to minimize error probability, cost, and other relevant properties. For a well-understood task in a predictable context, agent software can be specified as a discrete control system or a state machine. Autonomous agents are expected to be 'intelligent.' Intelligence refers to the extent to which the agent responds in an optimal manner to the events that it monitors, in terms of correctness, speed, and reliability of responses. The goal of an agent is to optimize the benefit for the owner of the agent. It is thus important to note that it is the behavior only that is judged to be intelligent, not whether the underlying computational processes are implemented as a fixed control algorithm, or include planning, learning, and reasoning (McFarland and Bösser, 1993).

Design Principles for Agents

Design principles that can be applied to agents are:

1. Development of a reactive system based on profound analysis and understanding of the tasks and context in which the agent is operating. An agent of this type may be implemented as a limited number of rules, or a state machine. The intelligent behavior is not a function of the real-time computations of the agent, but of the optimal selection of a limited number of rules or behaviors that overall generate successful behavior.
2. Artificial intelligence (AI) principles, including planning and learning, can be used to make an agent adaptive and react successfully in many different states that need not be anticipated in the design phase.

3. Agents often are designed according to inspirations and partial structural analogy to biological and social systems, with features defined in anthropomorphic terms such as mood and emotions. Generally the definitions in anthropomorphic terms are aspirations to realize a certain functionality rather than the claim to reproduce biological or even human capabilities.

Implementation of Agents

Agents may be implemented in widely varying forms, ranging from simple event-response formalisms or state machines (where the specification translates directly into implementation) to complex AI architectures with the aspiration to model human behavior such as SOAR (Tambe et al., 1995). Brooks (1991) has argued that robots do not necessarily have to use cognitive functions, reasoning, or planning to behave efficiently and to appear intelligent, but that the selection of optimal features for sensors and effectors and simple rules governing the response and an appropriate architecture permits the implementation of robust and effective robots. The bottom-up composition of robots in a hierarchical fashion from separate components with their own independent behavior is called subsumption architecture, using, for example, legs comprising mechanical structures, sensors, and motors plus basic wiring that allows them to move independently, or a vision system that includes preprocessing and active control of the sensors as building blocks. The advantage of the subsumption architecture is that there is no need for a centralized representation of all sensor data, and complete planning of all behavior in detail. The central control system may be much smaller, and the behavior of the entire robot more robust. As components are combined, the resulting behavior becomes more complex and assumes behavior not designed into the components, which is often called emergent behavior. Agents, by analogy, would use existing data capture and analysis functions and initiate actions by calling other programs, and thus can be restricted to the decision-making algorithms that are the essence of the agent's behavior.

The combination of a form of learning with autonomy is highly attractive because the ability of the agent to adapt to changing environmental conditions further decreases the need for supervision and instruction by the owner, and increases the robustness of its behavior. In most instances simple forms of learning (adaptation of parameters) are used. More complex forms of learning (concept formation, elaboration of new response patterns based on planning) are hard to implement, require massive amounts of information, and are much less reliable and secure than preprogrammed responses. The ability to learn can have advantages, but also carries high cost during design and run-time.

Multiagent Systems and Cooperating Agents

Autonomous agents are robust due to their task-specific and encapsulated nature, but this also limits them to the particular tasks and contexts for which they are designed. As the functionality of an individual agent increases, the optimization effort and maintenance effort is likely to increase unproportionally. An alternative is to compose systems with complex behavior from a number of loosely coupled agents. As a result the robustness and reliability increase, local decision making using local data demands fewer central resources, and the need for communication between components decreases. The composition of large systems as a collection of agents is attractive because of the open architecture, where new agents can be added as required without the need to modify the existing system.

The behavior of a multiagent system is not specified in detail, and testing has to be done experimentally. Exhaustive tests of a full range of conditions are generally not possible; the behavior of multiagent systems is not verified for unforeseen states. This may be very relevant for systems in safety-critical applications, and systems with fluctuating load, such as communication, emergency, or trading systems.

Swarms are collections of autonomous communicating agents, such as swarms of birds or ant colonies. Characteristically swarms exhibit regular and sophisticated behavior without centralized command and control. The apparently intelligent behavior of the swarm is generated by individuals with limited rules and communication with close neighbors. Aspired applications are underwater exploration and monitoring, where the capabilities of a swarm are an alternative to large hierarchically controlled systems. The swarm is robust, can suffer some fatalities, and can adjust to restricted communication abilities as in deep sea.

Anthropomorphic and Mentalistic Concepts

Agents can be designed to mimic natural language, speech, and animations as output to human users (Babski and Thalmann, 2000). This can help users, but may also be irritating, and create the erroneous impression that human-like reasoning is included in the agent. Mentalistic concepts may also be used in the design of agent features, such as motivation, emotions, consciousness, or moods, to describe the internal state of the agent (Sloman, 1997). At this stage it may be best to consider these concepts as metaphors for some aspects of technical communication, but these functions do not model human communication or even human mental processes.

Applications for Autonomous Agents

Interesting application domains for agents are those where the relative strengths of agent technology can be used to augment or partly replace direct interaction of the human user with a software application. Where continuous monitoring of data streams, repeated quick responses to events, statistical analysis and pattern detection in large amounts of data, and fast access to databases are required, agents can respond faster and more reliably than humans. Web services are interesting domains for the use of agents because they provide access to large amounts of information, create the need to search and select data, and make it attractive to enable automatic responses, relieving the human user from routine tasks, and also to improve performance.

Agents in User Interfaces

Early implementations of user interface agents (Lieberman, 1997; Maes, 1994) observed user behavior, interpreted user requests, and explicitly suggested actions and the automation of routine functions, but it was found that users prefer less obtrusive functions. Today agent-like functions are included in many applications as hints, preselection of options, or corrections and auto-completion of user input. Rather than the agent control core, innovative user interfaces use sophisticated sensors, data and text analysis to produce user support. These functions are called assistants, for example in cars, where they produce warnings and information display based on information not easily accessible to the driver.

Help functions in user interfaces can include agents that analyze previous user input and provide specific help for the user in his or her application context. More specific performance support uses a database of procedures specific to, for example, a particular company, a complex technical device, or an application domain. Performance support consists of guiding users to select and carry out prescribed procedures, relieving the user from the need to memorize and practice these procedures previously, and also to standardize procedures and to reuse knowledge in an organization (Chin, 1991). Studies where users were asked to assess their subjective impression of agents indicate that users are pleased to receive active support in situations where they feel helpless, but they are easily irritated by verbose and redundant information when they have relatively mature knowledge of a domain. Agents should include strategies and thresholds for presenting information in adequate format.

Operators for safety-critical tasks such as power station control or piloting are trained extensively in simulators. Where the nature of the task includes cooperation with other individuals, these are implemented as agents, representing a wide array of human behavior (Tambe et al., 1995).

Information Filtering and Access

Agents to support user information search behavior employ a personalized profile that is continuously constructed and adapted based on this users search behavior, and on known search strategies and preferences of other users. Access to large amounts of information, for example, catalogs, the CD and book market or scientific publications, but also TV programs, can be supported by agents that use instructions from the user in the form of keywords, reference to relevant similar items, and user profiles elaborated from previous user requests to actively search and display information selectively.

Agents in Electronic Trade

In electronic trading systems agents replace human traders in the monitoring of market data and the execution of orders for purchases and sales. Agents provided with strategies based on mathematical models exceed the capabilities of human traders in a number of aspects: agents can supervise more information streams with high precision, and execute orders extremely fast. Software agents do not use approximations, common sense, rules of thumb, or gut feelings, and are not subject to sentiments. The programs that the agents execute are rational in the sense that they do not deviate from the programed strategy to optimize profit. Research on economic decisions has shown that human traders deviate systematically from this form of rational behavior. Experiments have shown that software agents with optimal strategies dominate human robot traders (DeLuca and Cliff, 2011).

As in other multiagent systems the impact of diverse strategies executed by a large number of independent agents for the behavior of the system are not easy to predict, especially the occurrence of extreme states and anomalies of the market, leading to crashes. The benefit expected from electronic markets is efficiency and transparency, but the risk of instability is a concern. The appropriate design of markets taking these aspects into account is a specialized area of research (Sandholm, 2000).

In an open e-commerce environment individual participants provide trading agents with extensive amounts of information about their behavior. Sales agents can use this information to optimize their sales tactic by directing offers toward the single customer that are adapted to the profile and inferred preference of the potential buyer.

Evaluation of Agent Performance

Agent performance is judged by the owner and user according to the advantages obtained from the agent. Most agents incur only minimal costs for storage and computing power but may have considerable costs for communication with other agents, their environment, and the user. The main advantage should be reduced communication effort for the user in instructing his or her systems to perform his or her tasks.

The bandwidth of user–agent communication should be much lower than the bandwidth required for direct user–application interaction. The quality of performance must also be considered. Quality criteria for the information obtained and for the decisions made by the user include the number of hits and errors compared to other ways of accessing the same information, the erroneous decisions made, and opportunities missed. Precise and detailed analyses of the benefit of agent applications are still scarce. In a number of regular competitions, applications based on agent technology for specific tasks compete and compare their relative performance (TAC; DARPA grand challenge; Robocup).

Ethical Considerations

Obviously in systems where humans and agents interact, there is a potential for agents to have an impact on humans and important affairs. These may cause disruptive economic events, breakdown and failure of vehicles in traffic, or threats emanating from security applications. The creation of market crashes by the dynamics of extremely fast automated trading is a subject of research and practical concern. The emergence of risky behavior, or the interlocking of security mechanisms presents obvious risks that may be caused by emergent

behavior of agent systems. The use of agents for decision making and execution of procedures in situations with life-threatening consequences for individuals is subject to the same considerations as in respect of the use of military robots (Lin et al., 2012).

Ownership of Personal Data and Privacy

Agents often observe and subsequently attempt to predict the behavior of humans. The ability to predict behavior is a precondition for influencing, and potentially a means for manipulating, user behavior. Robots, since their conception in the twentieth century, are postulated to have a built-in rule not to harm humans. Agents, however, are mediators of human-to-human communication. An agent represents the interests of an owner, and in this role may pursue goals that are in conflict with the interest of other humans. Because ethical behavior is hard to enforce universally by rules, users may want to be cautious against giving information to agent-based systems that are outside their control. This is a particularly relevant aspect for e-commerce where the integration of data-processing systems opens the possibility to test and model the status of partners and competitors, predict their behavior, and adapt their own strategy. When, for example, it can be observed that the production facilities of a supplier are not fully used, it is likely that lower prices can be negotiated. The ownership and control over personal data is thus a valuable asset and subject of legal considerations and economic competition.

Multiagent Systems as Models of the Behavior of Social Systems

In biological and social systems large assemblies of individuals often exhibit regular behavior without centralized command and control, such as swarms of birds. Agent models are used as models of such behavior, for example in traffic systems or markets. The challenge is to identify the design principles embedded in the relatively simple individual agents that generate highly complex, adaptive, and seemingly intelligent behavior of the entire swarm. The behavior in traffic (pedestrians or cars) has been modeled as agent systems, and rules that generate safer and more efficient traffic flow have been suggested (Kretz and Schreckenberg, 2006; Yu and Helbing, 2010). Of particular interest is the case in which autonomous vehicles are included in the stream of traffic composed of human drivers.

Grand Visions and the Ontological Incongruence between Agent Implementations and Biological Entities

The term autonomous agent was originally a grand vision associated with high ambitions, concentrating on interacting agent technology combined with sensors and effectors. The strong point of view of agent technology (e.g., Moravec, 1988) assumed that the expected further growth of computational power will help to automate most of human work within the lifetime of some of us, and even surpass human cognitive abilities. These visions should be viewed with caution: the ontological incongruence of the theories of natural and biological processes on the one hand, and the engineering principles for constructing autonomous and intelligent systems on the other hand may create a larger distance than anticipated. When the core concepts in the two domains are examined, an essential difference between theories of cognition and the design specifications for artificial intelligent systems becomes apparent: formal and mathematical theories of natural processes are understood to be abstractions of a qualitatively different reality; they are generally not assumed to be able to represent the entire essence of the processes modeled, not even when extended dramatically in size. Specifications of computational artifacts must be assumed to be self-contained and complete – they could not work otherwise and would be useless for applications. The terms used to describe mental processes, such as *emotions* or *consciousness*, are often only names to describe entities that we recognize when we see them, but do not yet understand sufficiently in a reductionist sense. These theories may mislead some to believe that implementing these theories as computing machines will produce the full behavior of the entities in question. If not hubris, the ambition to build humanlike, or even animal-like, intelligence still represents a goal that transcends today's realistic possibilities. Computational models prove to be valuable research tools, but not easy means to generate automata that are capable of complex behavior observed in reality.

Further Developments

The concept of agents is maturing and realized in the implementation of many applications that support humans in everyday tasks. It appears that the further development of agent technology and practical applications profits mainly from extended abilities to sense and analyze large amounts of diverse information, and to execute complex procedures with high speed and precision, rather than the implementation of more capable cognition-like capabilities for decision making.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Human–Computer Interfaces; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Network Models of Tasks; Workplace Deviance.

Bibliography

Babski, C., Thalmann, D., 2000. 3D on the WEB and virtual humans. Software Focus 1, 6–14.

Brooks, R.A., 1991. Intelligence without representation. AI 47, 139–159.

Chin, D.N., 1991. Intelligent interfaces as agents. In: Sullivan, J.W., Tyler, S.W. (Eds.), Intelligent User Interfaces. ACM Press, New York, pp. 177–206.

DARPA, 2007. Automotive Grand Challenge – Next Challenge. www.theroboticschallenge.org.

DeLuca, M., Cliff, D., 2011. Human-agent auction interactions: adaptive-aggressive agents dominate. IJCAI, 178–185.

Kretz, T., Schreckenberg, M., 2006. F.A.S.T. – floor field- and agent-based simulation tool. In: Chung, E., Barcel'o, J., Dumont, A.-G. (Eds.), International Symposium of Transport Simulation (ISTS06).

Lieberman, H., 1997. Autonomous interface agents. In: Pemberton, S. (Ed.), Human Factors in Computing Systems. ACM Press, New York, pp. 67–74.
Lin, P., Abney, K., Bekey, G.A. (Eds.), 2012. Robot Ethics: The Ethical and Social Implications of Robotics. MIT Press, Cambridge, MA.
Maes, P., 1994. Agents that reduce work and information overload. CACM 7, 31–40.
McFarland, D., Bösser, T., 1993. Intelligent Behavior in Animals and Robots. MIT Press, Cambridge, MA.
Moravec, H., 1988. Mind Children. The Future of Robot and Human Intelligence. Harvard University Press, Cambridge.
Robocup. http://www.robocup.org/.
Sandholm, T., 2000. Agents in electronic commerce: component technologies for automated negotiation and coalition formation. Autonomous Agents and Multi-Agent Systems 3, 73–96.
Sloman, A., 1997. What sort of control system is able to have a personality? (Lecture Notes in AI). In: Trappl, R., Petta, P. (Eds.), Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents. Springer, Berlin.
TAC, Trading agents competition. http://aa.tradingagents.org/.
Tambe, M., Johnson, W.L., Jones, R.M., Koss, F., Laird, J.E., Rosenbloom, P.S., Schwamb, K., 1995. Intelligent agents for interactive simulation environments. AI Magazine 1, 15–39.
Yu, W., Helbing, D., 2010. Game theoretical interactions of moving agents. Understanding Complex Systems 0, 219–239.

Autonomy, Philosophy of

Natalie Stoljar, McGill University, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Individual autonomy is the property or capacity of agents of being self-governing or self-determining. An autonomous state is one in which a person's will is genuinely his or her own. This article outlines the main philosophical notions of autonomy under five headings: first, the traditional Kantian account; second, procedural accounts that claim that autonomy requires a kind of self-reflective authenticity; third, neo-Kantian analyses that introduce normative competence, or an ability to identify and apply moral or intersubjective norms; fourth, theories claiming that there are psychological conditions of autonomy, such as self-esteem, or the agent's sense of answerability for his or her own choices and preferences; and fifth, theories that introduce external moral or social conditions of autonomy.

Introduction

Individual autonomy is usually taken to be the property or capacity of agents of being self-governing or self-determining. Within the philosophical literature, 'autonomy' has been construed as a fundamental right, as a basic political good of liberal political theory, as a character ideal, and as a necessary condition of full moral agency. Although there are many different conceptions of autonomy, there is a common theme. An autonomous state is usually taken to mean one in which a person's will is genuinely his or her own.

Individual or agent autonomy (including moral autonomy) is to be distinguished from political autonomy. Political autonomy is a property of nations or states, and typically requires the absence of conditions that would undermine a nation's capacity for self-determination. A nation is self-governing or self-determining when it is not subject to interference or coercion by others. However, for an individual to be self-governing, certain positive conditions are required in addition to the absence of interference. Consider two typical failures of personal autonomy: first, agents who habitually and uncritically adopts others' social preferences and values as their own; and second, agents who have so internalized the outlooks and mores of the family in which they grew up that these are treated as authoritative without critical reflection (Feinberg, 1986: p. 32). Although in neither case is there any overt interference or coercion, agents in both cases lack autonomy because they fail to exhibit a characteristic necessary for autonomy – for example, a capacity for "subjecting these governing ideals to rational criticism and then modifying them when necessary" (Feinberg, 1986: p. 32). The development of such positive capacities can be undermined by social conditions including oppression, and therefore the articulation of a conception of autonomy is an important tool for feminists and other social critics (e.g., Meyers, 1989).

There are different classifications of theories of autonomy in the current philosophical literature. One approach is to divide theories of autonomy into procedural theories and substantive theories (Mackenzie and Stoljar, 2000b). Procedural theories claim that critical reflection is sufficient for autonomy: agents are autonomous when they reflect on their motivations, beliefs, and values in the right way. Procedural theories seek to define autonomy using morally neutral conditions such as reflective endorsement (Frankfurt, 1988; Dworkin, 1988; Friedman, 2003) or nonalienation (Christman, 2009). Substantive theories, however, build in moral or other normative conditions, for instance, moral attitudes to oneself or background moral conditions (Benson, 2005; Raz, 1986). Alternatively, theories of autonomy may be classified as either internalist or externalist. Procedural theories are internalist when they employ only internal psychological conditions to spell out autonomy; they are externalist when they require additional conditions that are external to agents' present internal psychological states, for instance, historical conditions (Mele, 1995). Other externalist theories employ the notion of morally adequate options (Raz, 1986) or 'sociorelational' conditions (Oshana, 2006). Since the latter externalist positions may also require a background moral theory to spell out precisely the external conditions that are incompatible with autonomy, these externalist theories are often substantive as well. Moreover, theories of autonomy distinguish between local and global conditions of autonomy. Many theorists focus on the conditions required for local autonomy, namely, what is required for particular choices, preferences, or desires at particular times to count as autonomous (e.g., Christman, 2009). Others focus on the conditions required for agents to achieve autonomy over an extended period and treat autonomy as a global condition of agents rather than a local feature of preferences and desires (e.g., Oshana, 2006).

This article describes the main philosophical notions of autonomy. The first section outlines the traditional Kantian account, which conceives of autonomy as an ability of rational agents to formulate universal moral principles. These moral principles are said to originate in the will of the agent and to be, in this sense, the agent's 'own.' The second section elaborates different procedural responses to the traditional Kantian account. These claim that what is required for autonomy is a kind of procedural authenticity, but each proposes a different conception of authenticity. The section briefly surveys three conceptions of authenticity: authenticity as endorsement or 'identification,' authenticity as nonalienation, and authenticity as self-discovery. The third section describes what could be called 'neo-Kantian' analyses of autonomy. These propose that autonomy requires 'normative competence,' namely, an ability

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63004-0

to identify and apply moral norms, or norms that are otherwise 'correct' for an agent to endorse. The fourth section identifies theories claiming that there are psychological conditions of autonomy in addition to processes of critical reflection. These conditions include either self-regarding attitudes such as those of self-trust and self-esteem or the agent's sense of answerability for his or her own choices and conduct. The final section briefly outlines theories that introduce external moral or social conditions on autonomy.

The Kantian Conception of Autonomy

A prominent strand within autonomy theory is derived from the philosophy of Immanuel Kant. For Kant, autonomy has a special meaning: "Autonomy of the will is the property the will has of being a law unto itself (independently of every property belonging to the objects of volition)" (1785/1948: p. 108). Rational beings make the moral law for themselves, and so can regard themselves as authors of the law. Thus, individual autonomy is manifested when rational agents 'will' the moral law. An important feature of the moral law is that it is a categorical, not a hypothetical, imperative. When agents formulate a categorical imperative, they formulate a maxim of the form 'I ought to do this,' not of the form 'I ought to do this if I desire that.' For example, agents formulate the moral law when they will 'I ought not to lie,' but formulate a hypothetical imperative when they will 'I ought not to lie if I desire not to be punished for lying.' The act of formulating the moral law is therefore an act of a pure autonomous will: it is untainted by the influence of the desires and interests that an agent may have relative to a particular situation. The pure autonomous will amounts to a metaphysical conception of the self, a self unconstrained by empirical desires, interests, and social forces. The traditional Kantian conception therefore presupposes a version of incompatibilism, namely, the thesis that free will cannot be reconciled with socialization or physical determinism.

A contemporary parallel of Kant's conception of autonomy is John Rawls's notion of free and rational agents formulating principles of justice in the 'original position' (Rawls, 1971). Rawls argues that rational agents formulate principles of justice from behind a 'veil of ignorance,' namely, from a position in which they are making decisions about how a society will function before they know who they will turn out to be in this society. For example, in the original position, agents do not know their social status, natural abilities, or conceptions of the good (Rawls, 1971: p. 12). Such agents, like Kantian agents, are not influenced by the particular desires and preferences that are contingent on being embedded in an actual situation. Because of this, their formulation of the principles of justice is taken to be the product of a 'pure' self, and hence genuinely self-originating.

The Kantian conception of autonomy has two aspects. First, autonomy is a function of a conception of morality that is encapsulated in the categorical imperative, or, in Rawls' account, in the principles of justice. The categorical imperative provides a standard according to which we can judge whether a particular example of reasoning is autonomous. If an agent's reasoning issues in the categorical imperative, then it is autonomous. Otherwise, it is not. (Thus, the Kantian account could be classified as both substantive and externalist.) Second, since autonomy manifests itself when the Kantian 'true' self is expressed through the formulation of the moral law, the Kantian account presupposes a metaphysical or 'noumenal' self. Both aspects of the Kantian view are contentious. Critics challenge the notion of a universally valid moral law as well as the concept of a metaphysically pure autonomous will. In addition, Rawls' account of the agent operating in the original position has been criticized as being unfaithful to empirical reality. Feminists and communitarians have argued that this agent is 'atomistic,' that is, existing in a vacuum divorced from the social relations in which actual agents are embedded (e.g., Sandel, 1982). No such agent exists, because agents are always the product of family influences, socialization, and so on. Moreover, an agent stripped of desires, interests, and preferences cannot engage in the imaginative moral reasoning required to formulate principles of justice. A further difficulty is that the Kantian conception is said to lead to an impoverished account of the value of autonomy. It is associated with the claim that autonomous agents are, and ought to be, substantively independent or self-sufficient, which in turn is associated with the character ideal of the 'self-made man.' Critics – especially feminists – have challenged this character ideal, and questioned whether substantive independence is really a value that a theory of autonomy, or normative theories in general, should promote. They have argued that a conception of autonomy should not exclude the possibility of agents having significant relationships of interdependency, such as those obtaining with family members (e.g., Friedman, 1997).

Procedural Theories of Autonomy

Theorists of autonomy have responded to the problems inherent in the traditional Kantian conception by offering compatibilist accounts, namely attempts to explicate autonomy within the network of physical and social forces in which agents operate. The notion of the self that is employed in procedural theories is empirical not metaphysical; procedural theories do not require that agents achieve substantive independence from the empirical conditions in which they are embedded. Rather, it is sufficient for autonomy that agents achieve procedural independence, that is, that their reasoning processes satisfy certain formal and procedural conditions. Procedural theories are both content-neutral and morally neutral. They are content-neutral because they do not require that the reasoning of autonomous agents delivers preferences or desires with particular moral contents. It is not required, for example, that the outcomes of autonomous reasoning correspond to what would be required by the universal moral law. Procedural theories are also morally neutral: they do not invoke background moral standards external to the agent as conditions of autonomy. It has been proposed that both content neutrality and moral neutrality are crucial aspects of a notion of autonomy that is useful in liberal political theory. It has been argued that only procedural conceptions of autonomy are compatible with liberalism (Christman, 2009).

The key task of procedural conceptions therefore is to distinguish when socialization undermines autonomy from when it does not. Procedural theorists argue that socialization may undermine autonomy by blocking either competency conditions or authenticity conditions. For example, John Christman defines agents as competent just in case they are capable of critical reflection and of forming 'effective' intentions (Christman, 2009: p. 154). If socialization undermines minimal capacities for critical reflection, or an agent's ability to form effective intentions, it also undermines autonomy. However, it is the authenticity condition that is taken by procedural theorists to be the 'core' of the notion of autonomy (Christman and Anderson, 2005: p. 3). In this section, I describe three conceptions of authenticity that have been offered by procedural theorists.

Endorsement or 'Identification'

The most influential procedural approaches are those of Gerald Dworkin and Harry Frankfurt (Dworkin, 1970; Dworkin, 1988; Frankfurt, 1971; Frankfurt, 1988). Both argue that what distinguishes autonomous from nonautonomous critical reflection is endorsement – agents' sense of engagement or 'identification' with their preferences or desires. In Frankfurt's terms, an autonomous will is the will the agent wants. Consider his distinction between a willing and an unwilling drug addict. The free will of both is hampered because addiction has undermined their 'freedom to do otherwise' (Frankfurt, 1971). However, the willing addict identifies with and endorses the desire for the drug, and hence has an important kind of freedom that the unwilling addict lacks. The unwilling addict is a mere 'passive bystander' with respect to the physiological addiction, whereas the willing addict has made the desire for the drug his own by identifying with it.

On Frankfurt's account, freedom of the will is spelled out using a hierarchy of desires. A first-order desire is autonomous just in case it is endorsed by a second-order volition, namely, a second-order desire that the first-order desire be effective (Frankfurt, 1971). His account therefore analyzes free will using conditions that an agent's present motivational state must satisfy. Alternative approaches introduce conditions that allow historical reflection on the processes of development of the motivational state. For example, Gerald Dworkin's analysis is hierarchical yet historical: autonomy is "a second-order capacity of persons to reflect critically on their first-order preferences, desires, wishes and so forth and the capacity to accept or attempt to change these in the light of higher-order preferences and values" (Dworkin, 1988: p. 20). A capacity allows reflection on the processes by means of which an agent acquires desires and preferences. It also distinguishes between an agent's autonomy on the one hand, and a particular exercise of autonomy on the other.

Important objections have been raised to endorsement accounts. The first focuses on their hierarchical nature. Even if first-order desires are endorsed by second-order volitions, this is not sufficient to ensure autonomy unless the second-order volitions themselves are endorsed. It has been argued that the problem results in a regress unless we introduce a qualitative distinction among levels, for example, between higher order values and first-order desires (Watson, 1975). However, the regress objection can be generalized to many hierarchical accounts, because whatever the nature of the higher order level, it can always be questioned whether it truly represents free agency. The higher order level itself will always be the outcome of the social forces influencing an agent. A second objection is that the endorsement condition may be too stringent, because it seems that certain unendorsed desires should count as autonomous. Suppose that, out of habit, an agent goes on vacation to the same destination every year. The fact that the agent fails to attend critically to the desire and does not endorse it each time should not rule out its being autonomous. Third, it has been suggested that the endorsement condition is too weak. Frankfurt's version of the endorsement account relies on the idea that, although an unwilling addict seems to intentionally inject himself with heroin, he prefers not to, and hence he does not endorse the desire to inject himself. It has been argued that (all things considered) intending to do something and preferring to do something are indistinguishable, and thus if agents intend to do something, they also prefer to do it, and to endorse it. Therefore, insofar as unwilling addicts intend to inject themselves with the drug, they also endorse it. The objection claims that endorsement merely distinguishes intentional from nonintentional agency, and not autonomous from nonautonomous agency. The apparent lack of autonomy of the unwilling addict is therefore not due to a lack of endorsement (Buss, 1994).

Authenticity as Historical Nonalienation

In early work, John Christman proposed a counterfactual and historical criterion of autonomy: agents are autonomous with respect to a certain preference or desire only if they do not resist its development when attending to the process of its development, or would not have resisted had they attended to the process (Christman, 1990, 1991). In response, it was pointed out that agents might repudiate the way in which a preference or desire was formed; an agent might, for example, repudiate an extremely doctrinaire religious upbringing, yet nevertheless be autonomous with respect to the religious values that were formed as a result of this process (Mele, 1995: pp. 138–139). In recent work, Christman formulates a modified 'hypothetical reflection condition' to spell out the notion of authenticity required for autonomy. A preference or desire is authentic and hence autonomous, if, were the agent to reflect critically on the historical processes leading to the formation of the desire, he or she would not be alienated from the desire (Christman, 2009: pp. 155–156). Alienation is a "combination of judgment and affective reaction. To be alienated is to feel negative affect, to feel repudiation and resistance…to feel a need to repudiate that desire or trait, to reject it and alter it as much as possible and to resist its effects" (Christman, 2009: p. 144). If one is (or would be) alienated from a desire but does not succeed in repudiating it, one is inauthentic and heteronomous with respect to the desire. Hence, on Christman's account, autonomy is compatible with preferences being habitually or even unconsciously formed, or being formed through processes that the agent on reflection would resist.

However, even this historical criterion is difficult to maintain in the face of apparent counterexamples. Consider Thomas Hill's example of the Deferential Wife: "She buys the clothes he

prefers, invites the guests he wants to entertain, and makes love whenever he is in the mood. She willingly moves to a new city in order for him to have a more attractive job… she tends not to form her own interests, values, and ideals, and when she does, she counts them as less important than her husband's" (Hill, 1991: p. 5). One would think that such excessive deference would be incompatible with autonomy, but the historical criterion cannot adequately explain why. The very fact that gender norms that treat women's submissiveness as desirable are so effectively internalized makes it unlikely that the agent would experience negative affect or alienation when reflecting on the genesis of the norms that she has endorsed (cf Benson, 1991). Christman's historical condition therefore does not account for cases that are often thought to be paradigms of inauthentic and nonautonomous preference formation.

Self-Discovery

Diana Tietjens Meyers describes acting autonomously as "the difference between doing what one wants and doing what one really wants. The autonomous self is not identical with the apparent self; it is an authentic or 'true' self" (Meyers, 1987: p. 619). For Meyers, the authentic self emerges when a person exercises the 'agentic skills' that characterize autonomous people. She writes that "autonomous people exercise a repertoire of skills to engage in self-discovery, self-definition and self-direction, … the authentic self is the evolving collocation [sic] of attributes that emerges in this ongoing process of reflection, deliberation and action" (Meyers, 2005: p. 49). These include skills of introspection, imagination, communication, and memory (Meyers, 2002: p. 20). Hence, Meyers does not propose that competency conditions and authenticity conditions of autonomy should be distinguished, but rather that the ongoing exercise of what she terms 'autonomy-competency' delivers a true or authentic self. Marilyn Friedman argues that being autonomous requires 'self-reflective affirmation' with respect to one's preferences and values. Through the process of critical affirmation (and reaffirmation), preferences and values become a person's own: "When someone reflectively reaffirms wants and values that are important to her…they become part of the perspective that defines her as the particular person she is. They embody the 'nomos' of her self…" (Friedman, 2003: p. 6). Friedman adds that the process of self-reflective affirmation "must be relatively unimpeded by conditions, such as coercion, deception and manipulation" (Friedman, 2003: p. 14).

One important focus in the work of both Meyers and Friedman is the question of whether autonomy is compatible with conditions of oppression including situations in which women themselves appear to be committed to oppressive gender norms. Meyers argues that "patriarchal noise can overwhelm women's voices and choices," but that autonomy skills "endow women with the capacity to fashion self-portraits and self-narratives in their own voices and to lead their own lives" (2002: p. 32). She points out that the development of the specific skills required for autonomy competency, such as introspection and imagination, can be impeded by oppressive practices and cultural stereotypes. It is important therefore that adequate education be available, especially to young girls, to allow children to develop autonomy skills. Friedman considers the impact of familial and community relationships on the skills of critical reflection required for autonomy (Friedman, 2003: p. 97). She notes that restrictive or oppressive social relationships may hamper an agent's ability for autonomy, or they may provide role models, self-trust, self-confidence, and so forth, which enhance the capacity (Friedman, 2003: p. 97).

For Meyers and Friedman, and on procedural accounts in general, there is no reason in principle why choosing subservience, or adopting norms that are oppressive, could not be autonomous. Critics have responded that preferences for one's own subordination are irrational and impair autonomy because they are inconsistent with the agent having an appropriate sense of his or her own moral worth (Superson, 2005). Others have suggested that severely constraining external circumstances, including the lack of a sufficient number of adequate options, compromise agents' freedom and autonomy, and hence procedural accounts that employ only psychological conditions in their analysis of autonomy are unsatisfactory (e.g., Raz, 1986; Brison, 2000; Oshana, 2006).

Neo-Kantian Theories

Procedural accounts usually assume that the conditions they identify, which are a function of actual or historical critical reflection, are both necessary and sufficient for autonomy. What could be called 'neo-Kantian' conceptions introduce additional external, moral conditions: autonomy requires that agents have a capacity or competency to identify and apply relevant norms. Since the norms are external moral or interpersonal norms, this competency is in part defined using external moral or interpersonal conditions. Hence, although certain procedural criteria such as a minimal capacity for critical reflection may be necessary for autonomy, they are not sufficient.

Susan Wolf adopts a normative competence view in which the capacity that is essential for autonomy is the capacity to track objective moral reasons. Wolf considers agents who have experienced morally impoverished or distorting socialization. For instance, suppose that JoJo is the son of an evil and sadistic tyrant who has been raised to respect his father's values and emulate his desires, so that he thoroughly embraces his father's worldview. Suppose that on procedural theories, JoJo counts as autonomous; he holds these values and desires as a result of appropriate self-reflection and therefore is governed by what he really wants. Wolf proposes that he is neither free nor morally responsible because his upbringing has undermined his capacity to distinguish right from wrong: "[i]t is unclear whether anyone with a childhood such as his could have developed into anything but the twisted and perverse sort of person that he has become" (Wolf, 1987: p. 54). On Wolf's account, the failure of autonomy is a failure of a capacity to track an objective aspect of the world, namely, 'the moral' or 'the right.' Since for Wolf the demands of morality are equivalent to the demands of objective 'reason,' in order to be autonomous, agents must be capable of discerning the requirements of reason.

Benson employs a parallel notion of normative competence, though it does not require a capacity to track objective

morality, but rather an ability to identify and deploy norms that are appropriate to a particular domain (Benson, 1987: p. 486). He writes that: "[F]ree agency requires normative competence, an array of abilities to be aware of applicable normative standards, to appreciate those standards, and to bring them competently to bear in one's evaluations of open courses of action…" (Benson, 1990: p. 54). Benson points out that normative competence does not require that the content of agents' preferences correspond to the content of the applicable norms because the exercise of a capacity in particular cases could fall short. In particular cases "normatively competent persons can choose what is unreasonable or wrong or value what is bad" (Benson, 2005: pp. 133–134). Hence, normative competence views do not conflate autonomy or self-rule with 'right-rule.'

There are nevertheless controversial issues raised by normative competence accounts. The first is that of the status of the moral and other norms that are employed to explicate normative competence. Wolf's account is controversial because it relies on objective reason. Benson's normative standpoints are intersubjective standpoints. Nevertheless, it is questionable whether competently deploying such standards is a plausible mark of autonomy. Intersubjective norms can be erroneously applied to agents, in particular in circumstances of oppression. Agents who recognize such norms and treat them as reasons for choice and action may be less autonomous than those who reject outright the applicability of intersubjective and oppressive norms that are applied to them. Finally, normative competence accounts may seem to collapse the distinction between responsibility and autonomy (Benson, 1994: p. 665). In the case of JoJo, for example, although it is plausible that the comprehensiveness of his socialization absolves him from full moral responsibility for his acts, it may be too quick to conclude that his acts are not the product of his own (autonomous) agency.

Self-Regarding Attitudes and Dialogical Approaches

Approaches that employ self-regarding attitudes propose that although critical reflection may be necessary for autonomy, it is not sufficient (Govier, 1993; McLeod, 2002; Benson, 1994). They claim that certain emotional states and attitudes to oneself, such as self-trust, self-esteem, and self-respect, are additional necessary conditions. Paul Benson describes a case of a woman with a certain personality type who is treated as pathological and crazy by her husband and the medical establishment (Benson, 1994: pp. 555–557). The woman's response is helplessness and disorientation leading to a loss of self-worth. Moreover, the woman does not resist the process through which she loses her self-worth because "she arrives at her sense of incompetence and estrangement…on the basis of reasons that are valued by a scientific establishment which is socially validated and which she trusts" (Benson, 1994: p. 657). Benson argues that, although the woman's critical reasoning faculties are intact, she nevertheless lacks autonomy because of a diminished sense of self-worth.

Theorists employing self-regarding attitudes notice that these attitudes can be undermined or enhanced by interpersonal conditions. Thus, self-regarding attitudes are often dialogical. On Benson's approach, self-worth is in part a requirement of "regarding oneself as being competent to answer [to others] for one's conduct" (Benson, 1994: p. 660). He develops this idea in later work to argue that autonomous agents are those who treat themselves as answerable for their conduct, who claim the authority to speak for themselves (Benson, 2005: p. 111 ff.). Others have also proposed that answerability is a key condition of autonomy (Anderson and Honneth, 2005; Mackenzie, 2008; Westlund, 2009). For example, Andrea Westlund argues that autonomy is neither a structural nor a historical capacity of critical reflection, but rather a disposition to treat oneself as answerable to the critical demands of others (Westlund, 2009: p. 35).

Both types of account potentially suffer from the objection that the conditions they identify are not sufficient for autonomous agency. Suppose an agent acquiesces in and endorses a set of circumstances or an ideology that is oppressive to her. In one sense, she may lack a sense of self-worth, but she may also regard herself as competent to answer for her actions in terms of those norms. Indeed, many agents who seem to be paradigms of nonautonomy, such as the Deferential Wife described in 'Authenticity as Historical Nonalienation,' will have the self-trust and self-esteem required to articulate their commitment to the norms they embrace. Indeed, the more deeply committed the agent is to such norms, the more likely it is that he or she will treat himself or herself as answerable for them. Thus, it seems that answerability could be present in many cases in which autonomy is not.

External Moral and Social Conditions

The theories of autonomy surveyed so far have treated autonomy as a psychological feature of agents, although in some cases we saw that the defining conditions of the psychological features include historical or other external conditions. Procedural theories require psychological processes of critical reflection. Normative competence theories rely on the agent's psychological capacity to discern substantive norms that are intersubjective or objective and hence are external to the agent. Dialogical approaches invoke agents' psychological states, such as sense of one's own answerability, although they analyze these states as related to external interpersonal conditions. In contrast to these accounts, which define autonomy primarily in psychological terms, the set of theories described in this section claim that agents with fully functioning, healthy psychologies could nevertheless have their autonomy undermined by external conditions.

Joseph Raz proposes that autonomy requires having morally adequate options in addition to mental ability, and independence from coercion and manipulation by other agents (Raz, 1986: p. 369). Consider a woman on a desert island who is hounded by a wild animal and has to spend all her time and energy planning for survival. The woman has a variety of options in addition to the mental abilities necessary for planning but the options she has are inadequate. For one thing, they are "dominated by her one overpowering need and desire to escape being devoured by the beast" (Raz, 1986: pp. 374–376). For another, having 'adequate options' is understood by Raz as

a moral requirement: "[a]utonomy requires that many morally acceptable options be available to a person" (Raz, 1986: p. 378).

Marina Oshana also claims that autonomy can be undermined by conditions in the world that do not necessarily affect the agent's psychology. She argues that agents who are reflective and psychologically competent can have their autonomy undermined by finding themselves in – or deliberately adopting – a situation in which their 'practical control' is removed: "We correctly attribute autonomy to a person when the person has *de facto* power and authority to direct affairs of elemental importance to her life within a framework of rules (or values, principles, beliefs, pro-attitudes) that she has set for herself" (Oshana, 2007: p. 411). For Oshana, agents living under conditions of severe social constraint, such as those obtaining under slavery, have limited autonomy. Such 'sociorelational' external conditions interfere with autonomy conceived as a temporally extended, 'global' condition of agents; they remove the *de facto* power that is required for autonomy on Oshana's account.

It has been suggested that sociorelational views are objectionable because they permit paternalism or are implicitly committed to an egalitarian ideal and hence a kind of perfectionism (Holroyd, 2009; Christman, 2004, 2009). It has also been alleged that these accounts are too quick to claim that severely constraining or morally inadequate external conditions always undermine autonomy. Rare individuals, such as Rosa Parks and Martin Luther King, despite being subject to constraining and morally objectionable conditions, resist oppression and in so doing seem to manifest autonomy (Noggle, 2011). If this is right, external sociorelational conditions that seem to impede agents' practical control do not *ipso facto* undermine autonomy.

See also: Free Will and Action; Freedom, History of; Freedom: Political; Self: Philosophical Aspects.

Bibliography

Anderson, J., Honneth, A., 2005. Autonomy, vulnerability, recognition, and justice. In: Christman, Anderson (Eds.), Autonomy and the Challenges to Liberalism, pp. 127–149.

Benson, P., 2005. Taking ownership: authority and voice in autonomous agency. In: Christman, Anderson (Ed.), Autonomy and the Challenges to Liberalism. Cambridge University, Cambridge, pp. 101–126.

Benson, P., 1994. Free agency and self-worth. Journal of Philosophy 91, 650–668.

Benson, P., 1991. "Autonomy and oppressive socialization". Social Theory and Practice 17, 385–408.

Brison, S.J., 2000. Relational autonomy and freedom of expression. In: Mackenzie, Stoljar (Ed.), Relational Autonomy Feminist Perspectives on Autonomy, Agency, and the Social Self. Oxford University Press, Oxford and New York, pp. 280–300.

Buss, S., 1994. Autonomy reconsidered. In: French, P.A., Uehling, T.A., Wettstein, H.K. (Eds.), Midwest Studies in Philosophy, vol. XIX. University of Minnesota Press, Minneapolis, pp. 95–121.

Christman, J., 2009. The Politics of Persons. Individual Autonomy and Socio-historical Selves. Cambridge University Press, Cambridge.

Christman, J., 2004. Relational autonomy, liberal individualism, and the social constitution of selves. Philosophical Studies 117, 143–164.

Christman, J., 1991. Liberalism and individual positive freedom. Ethics 101, 343–359.

Christman, J., 1990. Autonomy and personal history. Canadian Journal of Philosophy 20, 1–24.

Christman, J., Anderson, J. (Eds.), 2005. Autonomy and the Challenges to Liberalism: New Essays. Cambridge University Press, New York.

Dworkin, G., 1970. Acting freely. Nous 9, 367–383.

Dworkin, G., 1988. The Theory and Practice of Autonomy. Cambridge University Press, Cambridge.

Frankfurt, H., 1971. Freedom of the will and the concept of a person. Journal of Philosophy 68, 5–20.

Frankfurt, H., 1988. The Importance of What We Care about. Cambridge University Press, New York.

Friedman, M., 2003. Autonomy, Gender, Politics. Oxford University Press, New York.

Friedman, M., 1997. Autonomy and social relationships: rethinking the feminist critique. In: Meyers, D.T. (Ed.), Feminists Rethink the Self. Westview, Boulder, CO, pp. 40–61.

Govier, T., 1993. Self-trust, autonomy, and self-esteem. Hypatia 8, 99–120.

Hill, T., 1991. Autonomy and Self-respect. Cambridge University Press, Cambridge.

Holroyd, J., 2009. Relational autonomy and paternalistic interventions. Res Publica 15, 321–336.

Kant, I., 1785/1948. Groundwork of the Metaphysic of Morals (H.J. Paton, Trans. and analyzed). Harper & Row, New York.

Mackenzie, C., 2008. Relational autonomy, normative authority and perfectionism. Journal of Social Philosophy 39, 512–533.

Mackenzie, C., Stoljar, N. (Eds.), 2000a. Relational Autonomy Feminist Perspectives on Autonomy, Agency and the Social Self. Oxford University Press, New York.

Mackenzie, C., Stoljar, N., 2000b. Introduction. Refiguring autonomy. In: Mackenzie, Stoljar (Ed.), Relational Autonomy. Oxford University Press, New York, pp. 3–34.

McLeod, C., 2002. Self-Trust and Reproductive Autonomy. MIT Press, Cambridge, MA.

Mele, A., 1995. Autonomous Agents: From Self-control to Autonomy. Oxford University Press, New York.

Meyers, D.T., 2002. Gender in the Mirror: Cultural Imagery and Women's Agency. Oxford University Press, New York.

Meyers, D.T., 1989. Self, Society and Personal Choice. Columbia University Press, New York.

Meyers, D.T., 1987. Personal autonomy and the paradox of feminine socialization. Journal of Philosophy 84, 619–628.

Noggle, R., 2011. Review of Marina Oshana, personal autonomy in society. Journal of Value Inquiry 45, 233–238.

Oshana, M., 2006. Personal Autonomy in Society. Ashgate Publishing, Aldershot.

Rawls, J., 1971. A Theory of Justice, 1999, revised ed. Harvard University Press, Cambridge.

Raz, J., 1986. The Morality of Freedom. Oxford University Press, Oxford.

Superson, A., 2005. Deformed desires and informed desire tests. Hypatia 20, 109–126.

Taylor, J.S. (Ed.), 2005. Personal Autonomy: New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy. Cambridge University Press, New York.

Westlund, A., 2009. Rethinking relational autonomy. Hypatia 24, 26–49.

Westlund, A., 2003. Selflessness and responsibility for self. Is deference compatible with autonomy? Philosophical Review 112, 37–77.

Wolf, S., 1990. Freedom Within Reason. Oxford University Press, New York.

Wolf, S., 1987. Sanity and the metaphysics of responsibility. In: Schoeman, F. (Ed.), Responsibility, Character and the Emotions. Cambridge University Press, New York.

Avant-Garde Art and Artists

Diana Crane, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of the avant-garde has been applied to three types of changes in the arts: in the aesthetic content of art, in the social content of art, and in the norms surrounding the production and distribution of artworks. First used in France in the 1830s, the term acquired new meanings and connotations in the twentieth century, ranging from aesthetic innovation to alienation. The best-known literary and philosophical theories of the avant-garde have concentrated on the content of avant-garde works, while sociological theories and research have been concerned with the production and reception of these works. In the context of postmodernism at the end of the twentieth century, alternatives to the aesthetic and social values of art establishments are less likely to come from avant-gardes than from artists outside the system. The visual strategies and tactics of artistic avant-gardes of the prewar period have been appropriated by the electronic media and other forms of popular culture such as advertising, fashion design, and music videos. Works created for popular consumption that use avant-garde techniques are unlikely to contain critical content. As high culture has become less distinct from popular culture, the concept of the avant-garde has lost many of its former connotations. The avant-garde's mission as iconoclast and social rebel has been taken over by cultural avant-gardes in the form of a wide variety of micropolitical movements.

Since the concept of the avant-garde appeared at the beginning of the nineteenth century, it has been applied to a wide variety of aesthetic and social practices. Analysis of different uses of the term shows that it has been applied to three types of changes: in the aesthetic content of art, in the social content of art, and in the norms surrounding the production and distribution of artworks (Crane, 1987: pp. 14–15). For example, the term avant-garde is applied to the aesthetic content of art objects: (1) when those works represent a redefinition of conventions for creating art, often in such a way that they are perceived as violating taboos and as shocking or offensive; (2) when they involve the use of new tools and techniques, or concern the nature or use of techniques per se; and (3) when they redefine what can be considered an artwork. The term is applied to the social content of artworks when they express social or political values that are critical of or different from the dominant culture, when they attack art institutions, and when they attempt to redefine the boundaries between high and popular culture. Finally, the term is applied to creators of artworks when they attempt to alter the social context for the production of art (e.g., appropriate role models, critics, and publics for artists), the organizational context in which art is displayed and distributed, and the social role of the artist in terms of his or her participation in other social institutions, such as education, religion, and politics. This article examines the evolution of artistic avant-gardes since the nineteenth century, theories of the avant-garde, and sociological research on the social organization and reception of avant-gardes.

Origins and Evolution of the Term

Originally referring to a section of an army that marched ahead of the troops, the term avant-garde was first used in France in the 1830s to refer to artists as leaders and creators of a new social order (Nochlin, 1968: p. 5). Until the middle of the nineteenth century, the term generally meant that an artist was both politically and aesthetically progressive. The artist was expected to lead society in new directions and to be in conflict with tradition and the establishment. Subsequently, the term began to refer to artists whose works expressed alienation from bourgeois society in the form of ironic or destructive commentaries on social and artistic values rather than commitment to specific programs for social or aesthetic change. These different elements associated with avant-gardes – political rebellion, aesthetic rebellion, and alienation – continued to be important in the twentieth century.

At the beginning of the twentieth century, the idea of the avant-garde as alienation was given a new dimension by Marcel Duchamp who created artworks or 'ready-mades' by making minor changes in existing artworks or by arbitrarily declaring that commonplace objects, such as a mass-produced urinal, were artworks. This conception of the avant-garde became the basis for the Dadaist movement consisting of groups of artists, located in several European cities and in New York, whose members attempted to destroy artistic conventions and to shock the public. In this period, the Dadaists shared with the Italian Futurists (whose aesthetic innovations were very different) "a belief that they were ahead of the society in which they lived and that they were breaking social bonds of restraint" (Taylor, 1968: p. 97). This attitude led the Futurists to engage in flag burnings and to stage assaults on theaters and opera houses, although they were not actually interested in changing the social order but in aesthetic change and in "continuous revolution as a way of art" (p. 100).

During the twentieth century, the term was applied to groups of artists who developed new aesthetic programs, many of them concerned with problems of visual perception and the interpretation of visual experience. This type of art dealt with aesthetic and technical issues and was detached from the concerns of everyday life. Although the term avant-garde appeared before the development of modernism in the arts, there was a pronounced affinity between avant-gardes and modernism, which was reflected in the modernist's

commitment to the development of a particular style. At the end of the 1930s, Clement Greenberg, an American art critic, provided a theoretical justification for the type of avant-garde that has proven to be most characteristic of the second half of the twentieth century, one which engages in art for art's sake and which is concerned with aesthetic values rather than with social conflict or political struggle (Orton and Pollock, 1981: p. 322). Another American art critic, Harold Rosenberg (1968), argues that the suppression of avant-garde art by the Communists and later by the Nazis marked the beginning of the separation of political and artistic avant-gardes. As an example of this rupture during the postwar period, he cites the 1968 Venice Biennale where militant students denounced the artworks as 'art for dealers and the rich.'

In the late twentieth century, the success of Pop Art, Conceptual art, and neo-Expressionism signaled both the triumph of the ironic, alienated conception of the avant-garde and a decline in the priority attached to aesthetic innovation, reflecting the increasing influence of postmodernism in all forms of culture. Unlike the modernist, the postmodernist is not interested in style conceived as a consistent, integrated set of aesthetic elements or in using a style to express opposition to the dominant culture. Instead, the postmodernist engages in *pastiche*, bringing together disparate elements from many previous texts, regardless of whether they produce a coherent unity. According to Goldman (1992: p. 214), the meaning of postmodernist texts cannot generally be discerned from analyzing the texts themselves but only by asking where the elements in the text come from. Postmodernist works tend to expose or reinterpret rather than criticize or attack the dominant culture.

The relevance of the concept of the avant-garde was confirmed in the 1990s by the intensity of controversies surrounding government funding for artworks that challenged conventional attitudes and values. The allocation of funds by the United States government agency, the National Endowment for the Arts, to support exhibitions of artists such as Robert Mapplethorpe and André Serrano, whose works were perceived as having transgressed moral and religious boundaries, created enormous opposition and even pressure to dissolve the agency itself. However, it is significant that these artists were not members of art movements and that their activities did not engender such movements. Some authors claim that the avant-garde is dead and that contemporary artists do not correspond to this concept. Dunn (1991: p. 124) argues that, in the highly fragmented and ambiguous context of postmodern culture, artistic avant-gardes have been replaced by cultural avant-gardes to which the term avant-garde is seldom applied. These movements engage in collective protest and challenge tradition and hegemonic structures on a wide range of social and cultural issues such as class, gender, sexuality, race, peace, and ecology.

Theories of the Avant-Garde

Theorists who have written about the avant-garde have generally based their analyses on certain types of avant-gardes, particularly those that have rebelled against the dominant culture and excluded others from consideration. A major strand of theory concerning the avant-garde derives from the critical theorists, known as the Frankfurt School. Adorno despised popular culture produced by culture industries on the grounds that it provided shallow entertainment that distracted and manipulated the unsophisticated public. He interpreted the role of high culture and its avant-gardes as that of challenging dominant institutions and conventional ways of thought. For Benjamin (1968), the significance of Dadaism as an avant-garde was that the Dadaists deliberately attempted to produce artworks that lacked an aura of uniqueness and authenticity, thus undermining the status of the arts as sacred.

Two leading theorists of the avant-garde are Bürger (1984) and Poggioli (1968). A contemporary critical theorist, Bürger examines the activities of avant-gardes and the social content of their work, arguing that what is most characteristic of avant-gardes is their challenge to art institutions. Instead of creating works of art that were autonomous, complete, and separate from social life, the avant-garde attempted to create works that challenged the public to make sense of them and to relate them to their daily lives. On this basis, he restricts the title of avant-garde to a small number of movements, including Dadaism, early Surrealism, and some aspects of Futurism.

Poggioli (1968) argues that the dominant characteristic of avant-garde art is alienation in all its forms – psychological, social, economic, historical, aesthetic, and stylistic. Bourgeois, capitalist, and technological society provides avant-garde artists with freedom to create while the tensions underlying this type of society provide them with 'a reason for existing.' In Poggioli's view, the avant-garde artist is motivated by nostalgia for an earlier period when the artist was treated as a creator rather than as a parasite and a producer.

Recent debates about the contribution of artistic avant-gardes to politics in the postwar period also suggest that avant-garde artists in the 1960s perceived themselves as having broader social and political roles than those that are ascribed to artists today. Rasmussen (2009) analyzes three art movements of the 1960s – the Situationist International, the Artist Placement Group, and the Art Workers' Coalition – all of which attempted to use avant-garde art as a political tool to address social issues and bring about social change.

Huston (1992) is critical of these theories on the grounds that they are unable to explain and often do not even consider factors affecting the production and social reception of avant-garde arts. He argues that most styles of avant-garde art ultimately succeed in winning the support of the art establishment, which initially rejected them but that theories of the avant-garde are not helpful in explaining these developments. In his view, these theories tend to legitimate avant-garde art as *true* art and "a certain vision of the artist as a great artist" (p. 78). In other words, he sees the concept of the avant-garde as an ideological category that accepts uncritically artists' conceptions of their situation and tends to be used in such a way as to legitimate certain types of art rather than others (p. 79).

These theories of the avant-garde can also be faulted on the grounds that they treat the distinction between high culture and popular culture as unproblematic and view high culture and its avant-gardes as being the most influential and prestigious forms of culture. In fact, the enormous proliferation of different forms of popular culture (film, television, and popular music) has made it difficult to ignore the aesthetic

influence of these cultures on traditional arts and in everyday life, and has, at the same time, decreased the influence of the traditional arts such as painting, theater, dance, poetry, and experimental music. The latter tend to be appreciated by distinct and relatively small subgroups within the middle and upper classes, generally located in a few major cities.

Bauman (1997: p. 95) argues that avant-gardes have disappeared in this postmodern context because there is no longer any clear-cut distinction between styles that can be considered more or less advanced or more or less progressive than others. Instead, while there is a great deal of change in contemporary arts, these changes are 'random, dispersed, and devoid of clear-cut direction.' The importance of an artwork is determined less by aesthetic elements than by publicity and the notoriety that results from it. Several factors, such as the enormous increase in the prices of works by successful artists as well as in demand for these works, have contributed to a situation where successful artists have more to gain from creating artistic brands, making their work instantly recognizable to art collectors and dealers, than from creating works that critique the status quo or that are difficult to understand (Lury, 2005).

Sociological Studies of Avant-Gardes

While theorists of the avant-garde such as Bürger and Poggioli emphasize the aesthetic, political, and ideological programs of avant-gardes, sociological research has focused on (1) the relationships among artists and (2) the reception of artworks by art institutions and by the public. The dominant view underlying recent studies in the sociology of art is that the significance of all types of art is socially constructed. The question becomes one of identifying the social processes through which this takes place. A major concept is that of the art world, in which art is produced through the efforts of artists in collaboration with members of many other occupations (Becker, 1982).

White and White (1965) document changes in social and institutional structures that supported the activities of artists in nineteenth-century Paris. These changes facilitated the emergence of avant-garde styles. Specifically, the transition from academic art to Impressionism, a style that transformed the aesthetic conventions of the period, necessitated the development of a new occupation for buying and selling paintings, that of the art dealer, and the appearance of a new type of collector, drawn from the expanding middle class.

Becker (1982: pp. 304–305) does not use the term avant-garde, preferring to use the concept of 'artistic revolution.' The latter is set in motion by ideological and organizational activities among art world participants. Ideological changes are revealed by the appearance of 'manifestos, critical essays, aesthetic and philosophical reformulations, and revisionist histories,' while organizational changes lead to shifts in control over 'sources of support, audiences, and distribution facilities.' Like White and White, Becker argues that the most important factor affecting the success of revolutions is the extent to which their proponents are able to take over the art world's organizational structure and facilities. He says (p. 310): "Ideas and visions are important, but their success and permanence rest on organization, not on their intrinsic worth."

Crane (1987: pp. 137–143) traces changes in the roles of artistic avant-gardes during the twentieth century and particularly in the decades after World War II. She identifies three artistic roles that existed in the prewar period, that of the aesthetic innovator who analyzed visual reality in terms of its constituents such as color and form, the iconoclast, who attacked bourgeois conventions and art institutions, and the social rebel whose target was political and economic institutions. In the postwar period, these three roles were gradually transformed. The role of the aesthetic innovator emerged as the dominant role but in response to the expansion of public, private, and corporate expenditures and the creation of new cultural institutions, these artists are increasingly identified with the middle class in terms of lifestyle and the content of their works. Iconoclasts thrived during the 1960s but after that decade concentrated upon exploring the limits of art itself or eliminating the boundaries between the arts and the media. Artists who moved in the latter direction came to see themselves as entertainers rather than members of an alienated avant-garde. There was little support for the role of social rebel. Political art and a humanistically oriented representational art remained on the periphery of the art world. The role of social rebel was replaced by the democratic artist who communicated social and aesthetic ideas to audiences in poor urban neighborhoods and who had virtually no contact with the art world (for an examination of art avant-gardes in Israel see Greenfeld, 1988).

At the end of the twentieth century, in the absence of strong avant-garde movements within the art world, alternatives to the aesthetic and social values of the art establishment often came from outside the system, specifically from the works of untrained and often uneducated artists, including folk and ethnic artists, the homeless, prison inmates, elderly people in nursing homes, and hospice patients (Zolberg and Cherbo, 1997: p. 1). The so-called outsider artists who worked alone and who had formerly been completely ignored by the art establishment were now supported by a network of dealers, curators, and collectors.

Drawing on poststructuralist theories, sociologists view the meaning of an artwork as being constructed in the process of reception as much as in the process of creation. The response of the audience constitutes part of the meaning of the artwork. Consequently, the meaning of an avant-garde movement changes in different social contexts as different social groups with particular institutional interests engage in 'a struggle over the text.' Using this perspective, Halley (1985) examined the reception of the Dada movement and showed that, in its own time, its importance was not recognized. Audiences were very small and the movement was overshadowed by Surrealism. Subsequently, different social groups such as art historians, curators, critics, and artists themselves developed competing and conflicting interpretations of Dada and often misunderstood or ignored the original intentions of Dadaist works. From the beginning, there was more interest in Dada in the United States than in Europe, which helps to explain its influence on American avant-garde groups in music, dance, and art in the postwar period. These groups incorporated Dadaist ideas into their own artworks, contributing to wider recognition of the movement.

Bourdieu's (1984) theory argues that support for art movements is associated with social class but that a specific

aesthetic statement will appeal to different social strata, depending on its significance in the definition of taste in different time periods. The social positions of supporters of an emerging avant-garde movement differ from the social positions of those that admire the movement after it has received recognition. For Bourdieu, the important aspect of the avant-garde is not its aesthetic characteristics but its structural position in the art world, which is subject to change over time.

Attempts to show the relationship between social class and support for artworks that have been considered avant-garde in the past, such as abstract art, have yielded ambiguous results (Halle, 1992). However, appreciation of artworks that are still considered avant-garde is confined to relatively small groups even within the middle and upper classes, as is suggested by the intensity of the conflicts that arise when such works are displayed in public places (Heinich, 1997).

At the present time, the concept of the avant-garde is most meaningful in countries outside Europe and the United States, such as Japan, and particularly in developing countries, such as China and Latin America. Recent studies include (1) an ethnographic study, which attempts to interrogate the notion of the avant-garde in Japan through an analysis of avant-garde music and its audience in Japan (Plourde, 2010); (2) case studies of artists and art worlds in Beijing, which reveal an evolution, not unlike what has taken place in the West, away from avant-garde underground movements toward the development of a profitable, entrepreneurial art market in which market forces define the content of art (Tang, 2009); and (3) an analysis of changes in the types of discourse adopted in the literary works by a group of Mexican avant-garde writers (Parodi, 2006).

Avant-Gardes and Popular Culture

In the past two decades, the electronic media have appropriated most of the visual strategies and tactics of artistic avant-gardes from the prewar period. Caldwell (1995: p. 8) claims that "every framework of the avant-garde … had become highly visible in some form in the corporate world of the new television." Kaplan (1987: p. 55), in an attempt to classify music videos in the mid-1980s, found that all the categories she identified used avant-garde strategies. The originators of punk music used many of the tactics identified with early twentieth-century avant-gardes, such as the blurring of boundaries between art and everyday life, intentional provocation of the audience, and disorganization of accepted styles and procedures of performance (Henry, 1984).

Recent studies have examined avant-gardes in popular music (Lena and Peterson, 2008; Lena, 2012), jazz (Currie, 2010; Krnic, 2010), fashion design (Mears, 2008; Kawamura, 2004), culinary innovation (Hollows and Jones, 2010; Cunningham, 2008), and video games (Kirkland, 2010). These studies show the importance of the concept in understanding the emergence and development of these forms of culture. Use of the term increases the legitimacy of these types of culture.

In their analysis of types and trajectories of music genres (which they define as "systems of orientations, expectations, and conventions that bind together an industry, performers, critics and fans in making what they identify as a distinctive sort of music"), Lena and Peterson (2008: p. 698) identify four types of genres, one of which they labeled 'avant-garde.' Avant-gardes in popular music "engage in experimental practices, including playing standard instruments in unconventional ways, creating new musical instruments, and modifying objects that have not previously been used in the production of music" (p. 703). Two-thirds of new forms of music that appeared in the United States during the twentieth century originated as avant-garde genres.

The term, avant-garde, performs an important role in the long-term controversy over whether jazz is a form of high or popular culture (Krnic, 2010). Currie (2010) treats jazz improvisation as an avant-garde and attempts to understand the formation of subcultural communities surrounding this type of music.

Avant-garde practices are often used by emerging fashion designers to attract attention to their collections. Several, very influential Japanese designers have frequently been labeled avant-garde (Kawamura, 2004; Mears, 2008). Their work has been recognized in many museum exhibitions. The Italian fashion house, Prada, has benefited from its patronage of avant-garde artists and architects, which included gallery ownership, contemporary art collecting, and commissioning architects to design its stores. The company's connection with these artists and architects enhanced its symbolic capital and the distinctiveness of the brand (Ryan, 2007).

Recently, the term, avant-garde, has begun to be applied to culinary innovation. Hollows and Jones (2010) show that an emphasis on avant-garde innovation and experimentation is being used by chefs in high-end restaurants to increase their profits. In one case, a series of television shows in which a British chef demonstrated his intellectual and aesthetic practices had the effect of securing for him the label of 'culinary alchemist.'

Finally, an analysis of a survival horror video game series, Silent Hill, argues that the series qualifies as an art form on the basis that its texts exhibit formal devices used by art cinema, such as 'realism, ambiguity, psychological complexity, and self-reflexivity,' that its advertising emphasizes these qualities rather than gaming, and that its production has been the subject of a documentary (Kirkland, 2010).

However, assimilation by the mass media of oppositional stylistic devices associated with avant-gardes raises the question of the extent to which oppositional themes are being disseminated. High-culture institutions frame oppositional messages in such a way as to highlight their effect: an entire evening at the theater devoted to a particular playwright, an entire gallery or museum wing devoted to the work of a particular artist for several weeks. By contrast, the sheer volume of messages of all kinds being transmitted by the mass media tends to obliterate the effect of oppositional messages when they do appear. Kaplan (1987: p. 65) points out that oppositional messages are likely to be 'overridden by the plethora of surrounding texts.'

The commitment of popular culture creators to the tenets of the avant-garde may be questionable. Bourdieu (1993) differentiates between high culture and popular culture using the concept of the cultural field. Esoteric works are created in cultural fields in which autonomous creators have substantial control over standards for content and economic criteria for success are less important than aesthetic criteria. Popular works

are created in cultural fields in which creators are subject to the demands of managers and entrepreneurs who control the distribution of these forms of culture. Economic interests predominate.

When fashion designers incorporate avant-garde devices in luxury clothing for wealthy clienteles, does this activity represent a genuine avant-garde or is it simply a marketing strategy to attract attention, either to the clothes or to products licensed under their names? (Crane, 2000). Using Bourdieu's theory of cultural fields, the luxury fashion market could be classified as belonging to what he calls the field of 'restricted production,' where commercial interests are subordinate to aesthetic considerations (Bourdieu, 1993; Bourdieu and Delsaut, 1975). This interpretation is plausible in relation to relatively small companies operating in a single location. It becomes more difficult to defend in relation to large fashion firms with clienteles over the world.

Benetton, an Italian clothing manufacturer, has commissioned advertisements in which photographs, often reprinted from the press, graphically portray social and political crises and calamities in order to increase public awareness of social, political, and environmental issues (Guerrin, 1998). However, critics have not accepted this advertising strategy at face value but have claimed that the company is exploiting tragic situations as a means of publicizing its products.

The ambiguity that emerges when an avant-garde creator associates with a commercial enterprise is seen in the outcome of the collaboration between Rem Koolhaas and Prada (Ryan, 2007). The avant-garde architect was selected to design several avant-garde stores for the company because his status of being on the 'cutting edge' was considered desirable for the company's image. The stores he created have dual functions, as art centers and as shops, but may be incomprehensible to many customers who lack the appropriate cultural capital. Koolhaas himself denounced the Prada project in print as an example of 'junkspace' in which all public spaces have been turned into shopping centers. He was in turn accused by critics of appropriating high culture for commercial ends.

Curiously, as some of the properties of high culture are being assimilated or co-opted by certain forms of popular culture, the old hierarchy of high and low is reappearing in the latter. For example, in popular music, certain genres are deemed superior to others and some artists' works are considered 'classics' in comparison with all the rest (Regev, 1994).

The Contemporary Relevance of Avant-Gardes

An analysis of the evolution of avant-gardes in art history, as well as a review of theories of the avant-garde and of relevant sociological research, suggests that the concept of the avant-garde applies to the behavior and activities of artists during a period of about 150 years. At the present time the term is being less frequently used because it is more difficult to identify avant-gardes in an era when diverse and fragmented post-modernist styles predominate. Ironically, the term now tends to be applied to works created for popular consumption that incorporate techniques but not ideological content from avant-garde movements that existed earlier in the century. Meanwhile, the avant-garde artist's mission as iconoclast and social rebel has been assumed by a broad range of 'micropolitical' movements outside the arts.

See also: Art History; Art and Culture, Economics of; Artists, Competition and Markets; Bourdieu, Pierre (1930–2002); Cultural Mediators and Gatekeepers; Culture, Production of: Prospects for the Twenty-First Century; Frankfurt School: Institute for Social Research; Globalization and World Culture; Markets: Artistic and Cultural; Popular Culture; Postmodernism in Sociology.

Bibliography

Agee, William, 1968. New York Dada 1910–1930. In: Hess, Thomas B., Ashbery, John (Eds.), Avant-Garde Art. Collier Books, Collier–Macmillan, London, pp. 125–154.
Bauman, Zygmunt, 1997. Postmodernity and Its Discontents. New York University Press, New York.
Becker, Howard S., 1982. Art Worlds. University of California Press, Berkeley, CA.
Benjamin, Walter, 1968. Illuminations. Schocken Books, New York.
Bourdieu, Pierre, 1984. Distinction: A Social Critique of the Judgment of Taste. Harvard University Press, Cambridge, MA.
Bourdieu, Pierre, 1993. The Field of Cultural Production. Polity Press, Cambridge.
Bourdieu, Pierre, Delsaut, Yvette, 1975. Le couturier et sa griffe: contribution à une théorie de la magie. Actes de la Recherche en Sciences Sociales 1, 7–36.
Bürger, Peter, 1984. Theory of the Avant-Garde (M. Shaw, Trans.). University of Minnesota Press, Minneapolis, MN.
Caldwell, J.C., 1995. Televisuality: Style, Crisis and Authority in American Television. Rutgers University Press, New Brunswick, NJ.
Crane, Diana, 1987. The Transformation of the Avant-Garde: The New York Art World, 1940–1985. University of Chicago Press, Chicago.
Crane, Diana, 2000. Fashion and Its Social Agendas: Class, Gender, and Identity in Clothing. University of Chicago Press, Chicago.
Cunningham, T. LaRae, 2008. Eating soil and air: the culinary avant-garde at the turn of the 21st century. Dissertation Abstracts International, A: The Humanities and Social Sciences 68 (11), 4703.
Currie, A. Scott, 2010. Sound visions: an ethnographic study of avant-garde jazz in New York city. Dissertation Abstracts International, A: The Humanities and Social Sciences 70 (12), 4513.
Dunn, R., 1991. Postmodernism: populism, mass culture, and avant-garde. Theory, Culture and Society 8, 111–135.
Goldman, Robert, 1992. Reading Ads Socially. Routledge, New York.
Greenfeld, Leah, 1988. Professional ideologies and patterns of 'gatekeeping': evaluation and judgment within two art worlds. Social Forces 66, 903–925.
Guerrin, M., 1998. Oliviero Toscani, l'âme damnée de Benetton. Le Monde July (5–6), 8.
Halle, David, 1992. The audience for abstract art: class, culture and power. In: Lamont, Michèle, Fournier, Marcel (Eds.), Cultivating Differences: Symbolic Boundaries and the Making of Inequality. University of Chicago Press, Chicago, pp. 131–151.
Halley, J.A., 1985. The sociology of reception: the alienation and recovery of Dada. In: Herek, L., Rupel, D. (Eds.), Alienation and Participation in Culture. Research Institute of the Faculty of Sociology, Political Science and Journalism. University of Edvard Kardelj, Ljubljana.
Heinich, Natalie, 1997. Outsider art and insider artists: gauging public reactions to contemporary public art. In: Zolberg, Vera L., Cherbo, Joni M. (Eds.), Outsider Art: Contesting Boundaries in Contemporary Culture. Cambridge University Press, Cambridge, UK, pp. 118–127.
Henry, Tricia, 1984. Punk and avant-garde art. Journal of Popular Culture 17, 30–36.
Hollows, Joanne, Jones, Steve, 2010. Please don't try this at home: Heston Blumenthal, cookery TV and the culinary field. Food, Culture and Society 13, 521–537.
Huston, L., 1992. The theory of the avant-garde: an historical critique. Canadian Review of Sociology & Anthropology 29, 72–86.
Kaplan, E.A., 1987. Rocking around the Clock. Methuen, New York.
Kawamura, Yuniya, 2004. The Japanese Revolution in Paris Fashion. Berg, Oxford.
Kirkland, Ewan, 2010. Discursively constructing the art of Silent Hill. Games and Culture 5, 314–328.

Krnic, Raseljka, 2010. Jazz between popular and high art. Drustvena Istrazivanja 19, 1115–1138.
Lena, Jennifer C., 2012. Banding Together: How Communities Create Genres in Popular Music. Princeton University Press, Princeton.
Lena, Jennifer C., Peterson, Richard A., 2008. Classification as culture: types and trajectories of music genres. American Sociological Review 73, 697–718.
Lury, Celia, 2005. 'Contemplating a self-portrait as a pharmacist': a trade mark style of doing art and science. Theory, Culture and Society 22, 93–110.
Mears, Patricia, 2008. Exhibiting Asia: the global impact of Japanese fashion in museums and galleries. Fashion Theory 12, 95–120.
Nochlin, Linda, 1968. The invention of the avant-garde: France. In: Hess, Thomas B., Ashbery, John (Eds.), Avant-Garde Art. Collier Books, Collier-Macmillan Ltd, London, pp. 1–24.
Orton, F., Pollock, G., 1981. Avant-gardes and partisans reviewed. Art History 4, 305–327.
Parodi, Claudia, 2006. Linguistic fractures: the estridentistas. Mexican Studies 22, 311–329.
Plourde, Lorraine, 2010. Difficult music: an ethnography of listening for the avant-garde in Tokyo. Dissertation Abstracts International, A: The Humanities and Social Sciences 70 (8), 3065.
Poggioli, Renato, 1968. The Theory of the Avant-Garde (Gerald Fitzgerald, Trans.). Harvard University Press, Cambridge, MA.
Rasmussen, Mikkel B., 2009. The politics of interventionist art: the Situationist International, Artist Placement Group, and Art Workers' Coalition. Rethinking Marxism 21, 24–39.
Regev, Motti, 1994. Producing artistic value: the case of rock music. The Sociological Quarterly 35, 85–102.
Rosenberg, Harold, 1968. Collective, ideological, combative. In: Hess, Thomas B., Ashbery, John (Eds.), Avant-Garde Art. Collier Books, Collier-Macmillan, London, pp. 81–92.
Ryan, Nicky, 2007. Prada and the art of patronage. Fashion Theory 11, 7–23.
Tang, Ling-Yun, 2009. Commodifying culture in the global city: spatial practices and symbolic boundaries of contemporary art in Beijing. Dissertation Abstracts International, A: The Humanities and Social Sciences 70 (1), 376.
Taylor, Joshua, 1968. Futurism: the avant-garde as a way of life. In: Hess, Thomas B., Ashbery, John (Eds.), Avant-Garde Art. Collier Books, Collier-Macmillan, London, pp. 93–106.
White, Harrison C., White, Cynthia, 1965. Canvases and Careers. Wiley, New York.
Zolberg, Vera L., Cherbo, Joni M. (Eds.), 1997. Outsider Art: Contesting Boundaries in Contemporary Culture. Cambridge University Press, Cambridge, UK.

Avoidance and Approach Motivation: A Brief History

Rachel M Korn and Andrew J Elliot, University of Rochester, Rochester, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of motivation has been of interest to researchers in all fields who are interested in examining the underlying causes of behavior. Approach and avoidance motivation, in particular, has been a topic of much discussion since the inception of the field of psychology. This article begins with a brief definition of the approach–avoidance distinction and covers the history of the construct. It continues with a conceptual discussion of the approach–avoidance distinction followed by a comparison of the varied terminology researchers have used to discuss this issue.

Introduction

The study of motivation has been a topic of interest for scientific researchers since the inception of the field of psychology. Since it can be applied to a variety of types of research in different fields, it is important to have a clearly defined and articulated conceptualization of the fundamental approach–avoidance distinction. This article begins by explaining relevant terms, then covers the history of the approach–avoidance distinction followed by a conceptual overview of approach and avoidance motivation and a discussion of terminological considerations (for additional details, see Elliot, 1999, 2008; Elliot and Covington, 2001, from which the present article draws heavily).

Definitions

Before delving into the history and conceptualization of approach and avoidance motivation, it is important to provide clear definitions of the terminology to be used in this article. 'Motivation' refers to the energization and direction of behavior (Elliot and Covington, 2001). 'Approach motivation' is used to describe the energization by and direction of behavior toward a positive stimulus. 'Avoidance motivation' is used to describe the energization by and direction of a behavior away from a negative stimulus (Elliot, 2006).

The definition of motivation includes two distinct parts: the 'energization' and 'direction' of behavior. Energization is used to describe the preliminary triggering of a response that instigates a behavior (Elliot, 1997). This is not meant to imply that an organism is in a resting state before the introduction of a stimulus; quite the opposite, there is a constant flow of orienting behaviors in response to ever-changing stimuli (Atkinson and Birch, 1970). 'Direction' refers to the way behavior is conducted in response to a stimulus and can be associated with the satisfaction of both biological and psychological needs (Elliot and Thrash, 2001).

The idea of movement is implicit in the definition of approach and avoidance motivation. Organisms move toward positively valenced stimuli and away from negatively valenced stimuli (Arnold, 1960; Corwin, 1921). However, it is important to note that an organism need not make a physical movement toward or away from the stimuli. Behavioral predispositions can also be considered approach and avoidance oriented, even though the overt behaviors may follow a different path according to the goals of the organism (Elliot and Church, 1997). In addition, it does not always have to be physical movement toward or away from a stimulus. For example, if an organism is already close to a positive stimulus, then staying near the stimulus can also be considered approach behavior. Similarly, if an organism is far from a negative stimulus, staying away can be considered avoidance behavior. Accordingly, moving toward or staying near a positive stimulus is conceptualized as approach behavior, while moving away from or staying away from a negative stimulus is conceptualized as avoidance behavior (Elliot, 2006).

The terms 'positive' and 'negative' are used to describe the evaluation of stimuli by an organism. Stimuli that are evaluated positively invoke approach behaviors in organisms, while stimuli that are evaluated negatively evoke avoidance behaviors in organisms. 'Positive' and 'negative' are comprehensive terms that are functionally equivalent to other dimensions such as pleasant/unpleasant and desirable/undesirable (Elliot, 2006).

The term 'stimuli' is used throughout this article to refer to any number of objects, opportunities, or events, whether observable in the physical world or internally generated by the organism (Elliot, 2006).

History

The approach–avoidance distinction has been utilized in a variety of different ways by scholars for over 2000 years, but its roots are in the field of philosophy. The Greek philosopher Democritus (460–370 BCE) proposed that human beings should follow 'ethical hedonism' and be guided by the pursuit of pleasure and the avoidance of pain (Elliot and Covington, 2001). Centuries later, Jeremy Bentham (1779/1879), a British philosopher, offered a similar 'psychological hedonism' in which humans actually are guided by the pursuit of pleasure and avoidance of pain. Although these concepts are similar, Bentham took the idea of ethical hedonism further by providing an explanation for behavior instead of a mere suggestion for how individuals should strive to behave.

Since the inception of the field of scientific psychology, the distinction between approach and avoidance motivation has

 http://dx.doi.org/10.1016/B978-0-08-097086-8.26071-6

been of great interest to researchers (see Elliot, 1999; Elliot and Covington, 2001). One of the founding figures of psychology, Wundt (1887) posited that pleasure and pain were psychological experiences brought into conscious awareness through the processes of cognition and sensation. James (1890) theorized that pain and pleasure were 'springs of action' that cause the energization of behavior and that pleasure served as reinforcement while pain served to inhibit actions. Freud (1915) suggested in his psychodynamic theory that organisms tend to seek pleasure and avoid pain. In addition, he described these tendencies as motivations that guide behavior and are regulated by the superego (Freud, 1923). These early theorists helped to move the focus from the distinction between pain and pleasure to the resultant approach and avoidance behaviors (Elliot, 2008).

These initial forays into the study of approach and avoidance motivation paved the way for other researchers to adapt this conceptualization for their own use. Thorndike (1911) proposed in his 'law of effect' that responses which resulted in 'satisfaction' were more likely to be repeated, while responses that resulted in 'discomfort' were less likely to be tried again. Jung (1921) suggested that extroverts were more likely to move toward social objects whereas introverts were more likely to move away from social objects. Pavlov (1927) distinguished between two distinct responses in his research on classical conditioning: an 'orienting' response (toward a stimulus) and a 'defensive' response (away from a stimulus). Lewin (1935) theorized that goal objects have positive or negative valences that attract or repulse organisms. Skinner (1938, 1953) differentiated between reinforcement, which led to strengthened responses, and punishment, which led to weakened responses.

Murray (1938) distinguished between 'adient,' positive needs which lead organisms toward objects, and 'abient,' negative needs which lead organisms away from objects. Hull (1943) described the approach–avoidance distinction as conditioned appetitive and aversive drives formed through positive and negative associations with stimuli, which lead organisms toward or away from stimuli. Miller (1944) theorized that conflicts result from incompatible valences and that organisms can be both drawn to and repulsed from the same goal object. Hebb (1949) suggested that differing levels of stimulation can result in different behavioral responses: stimulation below a certain level leads to approach behaviors, while stimulation above the level leads to avoidance behaviors. Rotter (1954) posited that individuals form expectancies from past experiences with rewards and punishments.

Maslow (1955) theorized that human beings have two basic sets of needs: growth needs that lead an individual to reach for more positive stimulation, and deficit needs that lead an individual to seek tension reduction in order to avoid negative stimulation. Cattell (1957) identified two innate motives: exploration (an appetitive motive) and escape to security (a nonappetitive motive). Heider (1958) conceptualized the difference between the words 'can' and 'may' in a similar way, claiming that 'can' implies that the person will succeed if they try, while 'may' implies that they will not be punished if they try. Rogers (1961) suggested that personal goals represent either a desire to move toward a positive stimulus or a desire to move away from a negative stimulus. Erikson (1963) differentiated between basic trust and mistrust in his Stages of Psychosocial Development. Eysenck (1967) theorized that introverts had high baseline levels of arousal and were therefore likely to be 'stimulus shy,' while extroverts had low baseline levels of arousal and were thus more likely to be 'stimulus hungry.' Bowlby (1969) differentiated between two types of attachment: secure attachment, which leads individuals to explore and seek challenges, and insecure attachment, which leads individuals to be cautious and concerned with safety.

This brief overview of psychological theorists is intended to establish the fact that the approach–avoidance distinction has been utilized throughout the history of psychology. In addition, this conceptualization has been present in every major theoretical orientation since the inception of psychology as a scientific field, from psychodynamic to behaviorism, humanistic, cognitive, biological, etc. (Elliot and Covington, 2001). It is also clear that there has been a historical lack of a widespread and useful conceptualization of the approach–avoidance distinction, which has led to a tremendous amount of variance in the way researchers describe the same concept. Researchers from different fields and backgrounds have applied the idea of a split between approach and avoidance motivation to their own theories in different ways, sometimes using different terminology, without explicitly defining the concept (Elliot, 2008). In addition, much historical work has focused on a variety of different specific concepts within the field of motivation. For example, research has centered on hedonism, drives, underlying motives of behavior, and valenced stimuli, but it is important to consider how the overarching approach and avoidance distinction can encompass all of these different types of research (Elliot, 2008).

Conceptualization of Approach and Avoidance Motivation

Approach and avoidance motivation have been observed across a tremendous variety of situations and in all different levels of species. Approach motivation has generally been linked to thriving, while avoidance motivation tends to facilitate survival. However, both are integral adaptive mechanisms that are necessary for the physical and psychological well-being of organisms (Elliot, 2006).

The approach–avoidance distinction can be seen in life forms ranging from the single celled to as complex as human beings. At the most basic end of the spectrum, the amoeba has exhibited approach and avoidance behaviors. For example, the introduction of a weak light stimulus results in a movement of protoplasm toward the light. In contrast, an intense light stimulus results in the opposite effect, a movement away from the light (Schneirla, 1959). These basic adaptive mechanisms are present at all evolutionary levels and serve to facilitate need satisfaction in organisms in everything from finding food and shelter to avoiding harm. Although the complexity of these behaviors varies substantially across different types of life forms, the underlying mechanisms of approach and avoidance motivation are present across species (Schneirla, 1959) and

have been documented in a number of different species (see Elliot and Covington, 2001). For example, Budaev (1997) implemented factor analysis to explore the patterns underlying behavior in guppies. Data showed that there were two distinct factors, approach and fear avoidance, and that they were stable over time. Further independent factor analyses conducted with different species have also supported the approach–avoidance distinction (Elliot and Covington, 2001). Research suggests that these approach and avoidance behaviors that lead organisms toward positive stimuli and away from negative stimuli have adaptive benefits (Schneirla, 1959).

Another argument for the approach–avoidance motivational distinction is the instantaneous nature of many approach and avoidance behavioral responses (Elliot and Covington, 2001). Even low-level organisms display immediate reflexive responses to stimuli. For example, the amoeba instantly displays the withdrawal response to bright light. It can be argued that the speed of the response, not just the ability to respond to stimuli, is part of the adaptive significance of such behavior (see Berntson et al., 1993; Orians and Heerwagen, 1992). Humans also experience a number of immediate approach and avoidance reflexive behaviors in response to external stimuli. Some of the most basic survival mechanisms can be categorized as either orienting (linked to approach) or defensive (linked to avoidance) (Dickinson and Dearing, 1979; Sokolov, 1963). For instance, the blink reflex is an automatic physiological response that occurs almost immediately after a startling stimulus such as a puff of air or bright light (Bradley and Vrana, 1993). This is an adaptive mechanism that can protect the eye from injury and prepare the body to cope with environmental threats (Lang, 1995; Öhman, 1997). The speed and strength of this response can vary in accordance with the motivational state of the individual; the reflex will occur more quickly and in greater strength when individuals are in a negative (avoidance) motivational state, in contrast with neutral and positive (approach) motivational states (Lang et al., 1990).

Humans tend to make positive or negative evaluative judgments of all stimuli they encounter (Osgood et al., 1957), both instantaneously and without conscious awareness (Bargh, 1997; Zajonc, 1998). Bargh et al. (1996) conducted a study on the 'automatic evaluation effect' where participants were subliminally primed positively or negatively with words and asked to pronounce other valenced words. The results showed that the subliminal priming of positively valenced words led to faster pronunciation of other positively valenced words. Researchers concluded that since participants were unaware of the primes, they had automatically processed the valence of the primed words despite the lack of instruction for them to do so. Although some researchers have suggested that these evaluations are a form of automatic cognitive processing (Fiske, 1982), others argue that they are an automatic reaction to stimuli, independent of cognitive processes (Kuhl, 1986; LeDoux, 1987; Zajonc, 1980). This automatic evaluation process is understood to immediately invoke approach and avoidance behavioral tendencies (Elliot and Covington, 2001). Researchers from a wide variety of fields (see Elliot and Covington, 2001) have explored the idea that these automatic evaluations can evoke instantaneous predispositions toward or away from stimuli. For example, Chen and Bargh (1999) asked participants to enact an approach behavior (pulling a lever toward themselves) or an avoidance behavior (pushing a lever away) when they saw a word. When the word valence matched the response valence (positive word for pulling the lever), participants reacted faster than when there was a discrepancy between the valence of the stimulus and the response, even when they were not asked to evaluate the words. Bargh and Chartrand (1999) posited that the automatic evaluation process leads to almost immediate approach and avoidance behavioral predispositions.

The behavioral predisposition created by automatic evaluation is not necessarily an explicit behavioral response (Elliot and Covington, 2001). Stimuli may produce a valenced automatic evaluation and preparedness for approach or avoidance behaviors (Arnold, 1960). However, the enacted behavior is not always linked with this evaluative response (Lang et al., 1997). In lower-level organisms, automatic evaluations may always result in predictable behavioral responses. On the other hand, humans may have goals that lead to behaviors not tied to the initial approach or avoidance response (Elliot and Covington, 2001). For example, if individuals walk by a coffee shop, they may experience an automatic approach reaction and take a step forward to enter. However, if they are trying to cut back on their caffeine intake, they may exhibit an overt avoidance behavioral response by turning around and walking away. As such, in order to predict the behavior of complex organisms such as humans, it is important to keep in mind that approach and avoidance motivation may be affected by differing goals on multiple levels (Cacioppo and Berntson, 1994), represented hierarchically (Elliot and Church, 1997) or on the same level (Miller, 1944).

A large portion of the research exploring the approach–avoidance distinction has focused on neuropsychological data. Davidson and colleagues (see Davidson, 1995) conducted a number of studies examining the localization of approach and avoidance motivation in the cerebral hemispheres. They found that greater resting activity in the left prefrontal lobe was associated with approach motivation and positive affect, while greater resting activity in the right prefrontal lobe was associated with avoidance motivation and negative affect. It should be noted that hemispheric activity was monitored using electroencephalograph (EEG) recordings and collapsed into a single scale of differences between the left and right hemispheres. This type of measurement is not ideal if we are to accept the independence of approach and avoidance as motivational constructs because it does not allow us to measure hemispheric activity when there are high levels of both approach and avoidance motivation (Ito and Cacioppo, 1999). However, further research has supported these findings by linking approach motivation with activation of the left hemisphere and avoidance motivation with activation of the right hemisphere using PET and fMRI studies (Murphy et al., 2003).

More recent studies have taken a different approach to studying approach and avoidance motivation at the neocortical level. Friedman and Forster (2005) used a behavioral measure to determine relative hemispheric activation after invoking

approach and avoidance anticipatory states by having participants complete an enacted behavior task. They found that anticipatory approach states led to greater right hemispheric activation (relative to left hemispheric activation), while anticipatory avoidance states led to greater left hemispheric activation (relative to right). However, when they examined approach- and avoidance-related arousal (rather than anticipation), the results were reversed. The results obtained when approach- and avoidance-related arousal was induced matched with earlier EEG studies on the localization of approach and avoidance motivation in different cerebral hemispheres. Continuing the behavioral measure research, Roskes et al. (2011) examined hemispheric activation by studying directional biases. After inducing approach motivational states, they found that participants displayed right-oriented behavioral biases, which implies left hemispheric activation. However, these results were only present when participants were under time constraints. More research is needed to elucidate the hemispheric activation differences associated with approach and avoidance motivation.

In addition to the aforementioned work on hemispheric differences in approach and avoidance motivation, research has also explored the idea that different areas of the brain may be associated with each construct. Historically, theorists have found that electrical stimulation of the brain can lead to the activation of 'pleasure centers' or feelings of pain (Delgado et al., 1954; Olds and Milner, 1954). These early data linked the septal area of the limbic system with pleasure, and the hypothalamic area of the limbic system with pain. More recent data have gone on to link the medial prefrontal cortex and nucleus accumbens to approach motivation and the amygdala and anterior cingulate cortex with avoidance motivation (Wager et al., 2003).

While both approach and avoidance motivation are integral mechanisms of our physical and psychological processing, they often have very different consequences. Approach motivation tends to enable thriving and frequently leads to positive outcomes (Elliot, 2006). Avoidance motivation, on the other hand, is linked to survival needs and a focus on negative stimuli, which can lead to stress and decrements in well-being (Elliot and Sheldon, 1997, 1998). In addition, while successful approach motivation can lead to positive outcomes, successful avoidance motivation can only lead to the absence of negative outcomes. Therefore, the presence of avoidance motivation can result in missed chances for positive growth and development (Elliot, 2006).

In sum, the conceptualization of approach and avoidance motivation is an integral part of research on explanation of behavior in all living organisms. It is an intuitive, instantaneous mechanism that is present in both reflexive and conscious behaviors and has proven biological underpinnings. For these reasons, this construct has been widely studied throughout history by scholars seeking to understand the causes of behavior.

Terminology

Throughout history, a variety of different terms have been used to describe the approach–avoidance distinction. Each deserves a brief overview of its unique conceptualization and usage (see Elliot, 2008; Elliot and Covington, 2001).

Early work in philosophy labeled the approach–avoidance distinction as 'hedonism,' distinguishing between two poles of pleasurable and painful experiences. This definition, utilized by the British Empiricists, has historically been quite narrow, only including physical sensations (Boring, 1950; Cofer and Appley, 1964); however, recently, in the field of psychology, 'hedonism' has been broadened to include other types of sensations, such as the experiences of accomplishment and failure, in addition to bodily feelings and affect (Kahneman et al., 1999; Rozin, 1999). The traditionally narrow definition of hedonism represents only a portion of the construct of approach and avoidance motivation, but the more recent expansion of the range of this definition is quite similar to the way approach and avoidance motivation is currently conceptualized (Elliot and Covington, 2001).

Some theorists have used the 'approach–withdrawal' distinction to examine visible behaviors in order to understand their motivational underpinnings. Schneirla (1959) argued for the use of observable behaviors as a tool for measuring motivation in order to achieve consistency when comparing lower-level and higher-level organisms. This conceptualization holds that organisms are acting on approach motivation when they move toward a stimulus and withdrawal motivation when they move away from a stimulus. More recently, Davidson and colleagues (Davidson, 1992; Sutton and Davidson, 1997; Tomarken et al., 1992) have broadened this definition to include behavioral predispositions as well as observable actions. For instance, an action tendency to approach a stimulus would be considered approach motivation even if there was no resulting movement toward the stimulus. These approach and withdrawal behavioral predispositions have been linked to activity in different hemispheres of the brain: approach tendencies are associated with activation of the left prefrontal cortex, while withdrawal tendencies are associated with activation of the right prefrontal cortex (Davidson, 1995; Roskes et al., 2011). The conceptualization of the approach–withdrawal distinction is similar to the approach–avoidance distinction. However, the approach–withdrawal construct focuses on the energization of behavior and only includes behavioral tendencies that are observable at a biological level. In contrast, the approach–avoidance distinction covers the energization and direction of action tendencies that are both psychologically and biologically based (Elliot, 2008).

Researchers in psychology have frequently used the terminology 'appetitive–aversive' when discussing a construct similar to the approach–avoidance distinction. Both of these conceptualizations focus on motivational valence: the underlying energization and direction of behavior toward positive stimuli and away from negative stimuli (Elliot and Covington, 2001). Craig (1918) suggested that appetites and aversions were experienced as internal states of readiness to approach a positive stimulus or to stay away from a negative stimulus. The use of the terms 'appetitive' and 'aversive' has been adopted by other theorists in the field of psychology (Tolman, 1932), and this operationalization has continued to be used in the areas of emotion and reflexive behaviors (Lang, 1995; Lang et al., 1997). Although the appetitive–aversive

distinction has gained prominence in certain areas of psychology, the conceptualization of approach and avoidance motivation is much more inclusive and can be used to examine a wider range of phenomena (Elliot, 2008).

The terminology 'approach–avoidance' was first used by Lewin (1935) in his work on Field Theory. He suggested that all stimuli have attractive or repulsive properties, which lead organisms to approach or avoid them. These approach and avoidance responses stem from the internal states of the organism and are often associated with the satisfaction of survival needs. Miller (1944) also adopted this conceptualization, although his research focused on the physical movement toward or away from objects. McClelland and colleagues (McClelland, 1951; McClelland et al., 1953) continued using this operationalization of the approach–avoidance distinction but focused on the dispositional preferences of the organism for states or experiences, such as the need for achievement. These theorists all considered the approach–avoidance distinction applicable to the underlying mechanisms of behavior and cognitive processes (Elliot and Covington, 2001).

Although each terminologically distinct conceptualization has its own explanatory strengths, it is best to adopt the most comprehensive definition for the approach–avoidance distinction. Historical work on 'hedonism' has focused too narrowly on only sensory and affective phenomena, while 'approach-withdrawal' has overemphasized observable processes and physical movement. The 'appetitive–avoidance' distinction is similar to 'approach–avoidance,' but the latter has been used more widely in the literature and is easier to understand for those unfamiliar with the field. In addition, the current usage of the 'approach–avoidance' distinction has evolved to become a broad explanatory tool for different types of phenomena ranging from reflexes to motivational states, which makes it the most comprehensive terminological conceptualization (Elliot, 2008).

Concluding Statement

As illustrated, the approach–avoidance distinction provides a conceptual foundation for the scientific study of motivation. Approach and avoidance motivation are fundamental to the study of behavior in humans and lower organisms and have been a topic of much discussion for over 2000 years in a variety of different fields. In order to fully understand the determinants of behavior, it is important to have a widely used conceptualization of approach and avoidance motivation.

See also: Academic Achievement Motivation, Development of; Gender and Academic Motivation; Intrinsic Motivation: A Cultural Perspective; Motivation and Actions, Psychology of; Motivation: History of the Concept; Race and Academic Motivation; School Achievement: Motivational Determinants and Processes; Schooling: Impact on Cognitive and Motivational Development; Self-Determination Theory.

Bibliography

Arnold, M., 1960. Emotion and Personality. Columbia University Press, New York.
Atkinson, J.W., Birch, D., 1970. The Dynamics of Action. Wiley, New York.
Bargh, J.A., 1997. The automaticity of everyday life. Advances in Social Cognition, 20–27.
Bargh, J., Chaiken, S., Raymond, P., Hymes, C., 1996. The automatic evaluation effect: unconditional automatic attitude activation with a pronunciation task. Journal of Experimental Social Psychology 32, 104–128.
Bargh, J., Chartrand, T., 1999. The unbearable automaticity of being. American Psychologist 54, 462–479.
Bentham, J., 1779[1879]. Introduction to the Principles of Morals and Legislation. Clarendon Press, Oxford.
Berntson, G., Boysen, S., Cacioppo, J., 1993. Neurobehavioral organization and the cardinal principle of evaluative bivalence. Annals of the New York Academy of Sciences 702, 75–102.
Boring, E.G., 1950. A Historical of Experimental Psychology. Appleton-Century-Crofts, New York.
Bowlby, J., 1969. Attachment. Basic Books, New York.
Bradley, M., Vrana, S., 1993. The startle probe in the study of emotion and emotional disorders. In: Birbaumer, N., Ohman, A. (Eds.), The Structure of Emotion. Hogrefe & Huber, Seattle, pp. 270–287.
Budaev, S., 1997. "Personality" in the guppy (*Poecilia reticulata*). Journal of Comparative Psychology 111, 399–411.
Cacioppo, J., Berntson, G., 1994. Relationship between attitudes and evaluative space: a critical review, with emphasis on the separability of positive and negative substrates. Psychological Bulletin 115, 401–422.
Cattell, R., 1957. Personality and Motivation: Structure and Measurement. World Book, Yonkers, New York.
Chen, M., Bargh, J., 1999. Consequences of automatic evaluation: immediate behavioral predispositions to approach or avoid the stimulus. Personality and Social Psychology Bulletin 25, 215–223.
Cofer, C.N., Appley, M.H., 1964. Motivation: Theory and Research. Wiley, New York.
Corwin, G., 1921. Minor studies from the psychological laboratory of Cornell University. American Journal of Psychology 32, 563–570.
Craig, W., 1918. Appetites and aversions as constituents of instincts. Biological Bulletin 34, 91–107.
Davidson, R., 1992. Prolegomenon to the structure of emotion: gleanings from neuropsychology. Cognition and Emotion 6, 245–268.
Davidson, R., 1995. Cerebral asymmetry, emotion, and affective style. In: Davidson, R., Hugdahl, K. (Eds.), Brain Asymmetry. MIT Press, Cambridge, MA, pp. 362–387.
Delgado, J.M.R., Roberts, W.W., Miller, N.E., 1954. Learning motivated by electrical stimulation of the brain. American Journal of Physiology 179, 587–593.
Dickinson, A., Dearing, M., 1979. Appetitive–aversive interactions and inhibitory processes. In: Dickinson, A., Boakes, R. (Eds.), Mechanisms of Learning and Motivation. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 203–231.
Elliot, A.J., 1997. Integrating the "classic" and "contemporary" approaches to achievement motivation: a hierarchical model of approach and avoidance achievement motivation. In: Pintrich, P., Maehr, M. (Eds.), Advances in Motivation and Achievement, vol. 10. JAI Press, Greenwich, CT, pp. 143–179.
Elliot, A.J., 1999. Approach and avoidance motivation and achievement goals. Educational Psychologist 34, 149–169.
Elliot, A.J., 2006. The hierarchical model of approach and avoidance motivation. Motivation and Emotion 30, 111–116.
Elliot, A.J., 2008. Approach and avoidance motivation. In: Elliot, A.J. (Ed.), Handbook of Approach and Avoidance Motivation. Psychology Press, New York, pp. 3–14.
Elliot, A.J., Church, M.A., 1997. A hierarchical model of approach and avoidance achievement motivation. Journal of Personality and Social Psychology 72, 218–232.
Elliot, A.J., Covington, M.V., 2001. Approach and avoidance motivation. Educational Psychology Review 13, 73–92.
Elliot, A.J., Sheldon, K.M., 1997. Avoidance achievement motivation: a personal goals analysis. Journal of Personality and Social Psychology 73, 171–185.
Elliot, A.J., Sheldon, K.M., 1998. Avoidance personal goals and the personality-illness relationship. Journal of Personality and Social Psychology 75, 1282–1299.
Elliot, A.J., Thrash, T.M., 2001. Achievement goals and the hierarchical model of achievement motivation. Educational Psychology Review 12, 139–156.
Erikson, E., 1963. Childhood and Society, second ed. Norton, , New York.
Eysenck, H., 1967. The Biological Basis of Personality. Charles Thomas, Springfield, Illinois.

Fiske, S., 1982. Schema-triggered affect. In: Clark, M., Fiske, S. (Eds.), Affect and Cognition. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 55–78.
Freud, S., 1915[1957]. Repression. In: The Standard Edition of Complete Psychological Works of Sigmund Freud, vol. XIV. Hogarth, London.
Freud, S., 1923[1947]. The ego and the id. In: The Standard Edition of Complete Psychological Works of Sigmund Freud, vol. XIX. Hogarth, London.
Friedman, R.S., Forster, J., 2005. Effects of motivational cues on perceptual asymmetry: implications for creativity and analytical problem solving. Journal of Personality and Social Psychology 88, 263–275.
Hebb, D.O., 1949. The Organization of Behavior. Wiley, New York.
Heider, F., 1958. The Psychology of Interpersonal Relations. John Wiley & Sons, New York.
Hull, C., 1943. Principles of Behavior. Appleton-Century-Crofts, New York.
Ito, T., Cacioppo, J., 1999. The psychophysiology of utility appraisals. In: Kahneman, D., Diener, E., Schwarz, N. (Eds.), Wellbeing: The Foundations of Hedonic Psychology. Russel Sage Foundation, New York, pp. 470–488.
James, W., 1890. The Principles of Psychology, vol. II. Henry Holt & Co., New York.
Jung, C., 1921. Psychological types. In: Collected Works of C.G. Jung, vol. 6. Princeton University Press, Princeton, NJ.
Kahneman, D., Diener, E., Schwarz, N., 1999. Preface. In: Kahneman, D., Diener, E., Schwarz, N. (Eds.), Wellbeing: The Foundations of Hedonic Psychology. Russell Sage Foundation, New York, pp. ix–xii.
Kuhl, J., 1986. Motivation and information processing. In: Sorrentino, R., Higgins, E. (Eds.), Handbook of Motivation and Cognition, vol. 1. Guilford, New York, pp. 404–434.
Lang, P., 1995. Studies of motivation and attention. American Psychologist 50, 372–385.
Lang, P., Bradley, M., Cuthbert, B., 1990. Emotion, attention, and the startle reflex. Psychological Review 97, 377–395.
Lang, P., Bradley, M., Cuthbert, B., 1997. Motivated attention: affect, activation, and action. In: Lang, P., Simmons, R., Balaban, M. (Eds.), Attention and Orienting: Sensory and Motivational Processes. Lawrence Erlbaum Associates, Florida, pp. 87–135.
LeDoux, J., 1987. Emotion. In: Plum, F. (Ed.), Handbook of Physiology, vol. 5. American Physiological Society, Bethesda, MD, pp. 419–454.
Lewin, K., 1935. A Dynamic Theory of Personality. McGraw-Hill, New York.
Maslow, A., 1955. Deficiency motivation and growth motivation. In: Jones, M. (Ed.), Nebraska Symposium on Motivation. University of Nebraska Press, Lincoln, pp. 1–30.
McClelland, D., 1951. Personality. The Dryden Press, New York.
McClelland, D., Atkinson, J., Clark, R., Lowell, E., 1953. The Achievement Motive. Irvington Publishers, New York.
Miller, N., 1944. Experimental studies of conflict. In: Hunt, J.McV. (Ed.), Personality and the Behavioral Disorders, vol. 1. Ronald Press, New York, pp. 431–465.
Murphy, F.C., Nimmo-Smith, I., Lawrence, A.D., 2003. Functional neuroanatomy of emotions: a meta-analysis. Cognitive, Affective & Behavioral Neuroscience 3, 207–233.
Murray, H., 1938. Explorations in Personality. Oxford University Press, New York.
Öhman, A., 1997. As fast as the blink of an eye: evolutionary preparedness for pre-attentive processing of threat. In: Lang, P., Simmons, R., Balaban, M. (Eds.), Attention and Orienting: Sensory and Motivational Processes. Lawrence Erlbaum Associates, Hilldale, NJ, pp. 87–135.
Olds, J., Milner, P., 1954. Positive reinforcement produced by electrical stimulation of septal area and other regions. Journal of Comparative and Physiological Psychology 47, 419–427.
Orians, G., Heerwagen, J., 1992. Evolved responses to landscapes. In: Barkow, J., Cosmides, L., Tooby, J. (Eds.), The Adapted Mind. Oxford University Press, New York, pp. 555–579.
Osgood, C.E., Suci, G., Tannenbaum, P., 1957. The Measurement of Meaning. University of Illinois Press, Urbana, Illinois.
Pavlov, I., 1927. Conditioned Reflexes: An Investigation into the Physiological Activity of the Cortex (G. Anrep, Trans.). Dover, New York.
Rogers, C., 1961. On Becoming a Person: A Therapist's View of Psychotherapy. Houghton Mifflin, Boston.
Roskes, M., Sligte, D., Shalvi, S., De Dreu, C.K.W., 2011. The right side? Under time pressure, approach motivation leads to right-oriented bias. Psychological Science 22, 1403–1407.
Rotter, J., 1954. Social Learning and Clinical Psychology. Prentice-Hall, Englewood Cliffs, New Jersey.
Rozin, P., 1999. Preadaptations and the puzzles and properties of pleasure. In: Kahneman, D., Diener, E., Schwarz, N. (Eds.), Well-being: The Foundations of Hedonic Psychology. Russell Sage Foundation, New York, pp. 109–133.
Schneirla, T., 1959. An evolutionary and developmental theory of biphasic processes underlying approach and withdrawal. In: Nebraska Symposium on Motivation. University of Nebraska Press, Lincoln, pp. 1–42.
Skinner, B.F., 1938. The Behavior of Organisms: An Experimental Analysis. Prentice-Hall, Englewood Cliffs, NJ.
Skinner, B.F., 1953. Science and Human Behavior. Macmillan, New York.
Sokolov, E., 1963. Perception and the Conditioned Reflex. Pergamon, Oxford.
Sutton, S.K., Davidson, R.J., 1997. Prefrontal brain asymmetry: a biological substrate of the behavioral approach and inhibition systems. Psychological Science 8, 204–210.
Thorndike, E., 1911. Animal Intelligence. Macmillan, New York.
Tomarken, A.J., Davidson, R.J., Wheeler, R.E., Doss, R.C., 1992. Individual differences in anterior brain asymmetry and fundamental dimensions of emotion. Journal of Personality and Social Psychology 62, 676–687.
Tolman, E., 1932. Purposive Behavior in Animals and Men. The Century, New York.
Wager, T.D., Phan, K.L., Liberzon, I., Taylor, S.F., 2003. Valence, gender, and lateralization of functional brain anatomy in emotion: a meta-analysis of findings from neuroimaging. NeuroImage 19, 513–531.
Wundt, W., 1887. Grundzüge der physiologishen psychologie, third ed. Engelmann, Leipzig.
Zajonc, R.B., 1980. Feeling and thinking: preferences need no inferences. American Psychologist 35, 151–175.
Zajonc, R.B., 1998. Emotion. In: Gilbert, D., Fiske, S., Lindzey, G. (Eds.), The Handbook of Social Psychology, fourth ed. McGraw-Hill, New York, pp. 591–632.

Awards, Public Recognition, and Evaluation

Rafael Yann Schögler, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Scientific authority or the power and legitimation to contribute in scientific debates is granted by the recognition of peers and the public. To be granted recognition and esteemed by colleagues the visibility of discoveries, arguments, or theories are a prerequisite. One form of distinction manifests by being read, cited, or referred to in speeches, i.e., being part of the academic discourse. Public evaluation and distinction manifests in practices of issuing awards, grants, or prizes to researchers and their institutions; issuing honorary degrees to academics and other public figures, but also by evaluating and ranking research, researchers, and their institutions.

Introduction

Scientific authority or the power and legitimation to contribute in scientific debates is granted by the recognition of peers and the public. To be granted recognition and esteemed by colleagues the visibility of discoveries, arguments, or theories are a prerequisite. In this article, the most visible element of such recognition – the awarding of prizes – will be discussed. Public distinctions such as awards, honorary degrees, and other public appraisals help to establish academic recognition and at the same time organize the academic field. Awards make peer recognition, hierarchies, and hegemonic scientific discourses visible. Systematic studies dealing with the contemporary or historical role of awards in the social and behavioral sciences are rare. Rather, existing literature either deals with histories of single prizes, or more broadly with the phenomenon of reputation and recognition (H.S. Becker, P. Bourdieu, R. K. Merton). Taking this lack of empirical examination into account the following lines will shortly elaborate on functions of recognition and public distinction in the academic field before providing some historical background to the awarding of prizes and offering a tentative typology of awards in the academic field. A selection of international prizes will be depicted before focusing on atypical exponents such as honorary degrees and forms of collective distinction such as higher-education rankings.

Recognition and Reputation

Scientific authority or the power and legitimation to contribute in scientific debates is granted by the recognition of peers and the public (Bourdieu, 1976). Prizes, awards, and other forms of public distinction are compelling signs of public and peer recognition. Not only are these an institutionalized sign for the acceptance of the work of a researcher, but they also contribute to making research and researchers visible in the public sphere. To be granted recognition and esteemed by colleagues the visibility of discoveries, arguments, or theories are a prerequisite. In Howard S. Becker's (1982) description of the art world the acceptance of an oeuvre depends on whether one is working in an accepted style or not. Similarly, the use of specific methods, theoretical standpoints, and arguments in the social and behavioral sciences tend to influence the recognition bestowed upon a colleague inside a narrow scientific community. The desire for recognition – manifested by being read and cited – is a main motivation for academics (Hull, 1988).

A latent but nonetheless powerful form of distinction consists of a researcher being read, quoted in print, or referred to in oral communication, i.e., being part of the academic discourse. Histories of disciplines inevitably know something such as classics or key figures that have obtained 'reputational immortality.' This status becomes manifest through legacy-forming mechanisms such as naming an award, foundation, building, or lecture hall, but also a theory, school of thought, or controversy by the name of a researcher. This kind of scientific reputation is also shown by including researchers in textbooks, specialized encyclopedias, or histories of a discipline. On a less prominent level there is the 'Festschrift' and obituaries honoring the work of an academic. Of similar nature are commemorative events taking place in the memory of a particular person. Like in the field of arts (Lang and Lang, 1988) postmortem reputation is only established if the deceased reached a critical mass of publications that was conserved and distributed. Also, the study of 'classics' and other dead scholars could make use of records such as mail correspondence with prominent colleagues, notes, or unfinished manuscripts bequeathed upon death that can contribute to the memory of a particular person.

Patterns influencing the reward system in science deviating from meritocratic ideals show that public recognition of science does not only reward talent, productivity, or the contribution to the advancement of a scientific field but that institutionalized cultural capital plays a major role in the allocation of awards. Merton (1968) described one of these systematic deviations based on Harriet Zuckerman's work on Nobel laureates, which showed that in science "eminent scientists get disproportionately great credit for their contributions to science while relatively unknown scientists tend to get disproportionately little credit for comparable contributions" (Merton, 1968: 2). The first part – the overevaluation of recognized researchers – has been remembered as the Matthew effect. Rossiter (1993) described the (female) equivalent as the Matilda effect, where she specifically highlighted the systematic underappreciation of the work of women in scholarship.

These systematic unbalances in recognizing the work of scholars that ought to be meritocratic have resulted in the practice of creating prizes for those who have not received the

recognition they were supposed to get, as an illustration one could mention the so-called alternative Nobel Prize distinguishing improbable research (IGNobel, 2014).

Awards, Medals, and Prizes

What Bourdieu calls objectified symbolic capital are mechanisms of public evaluation and distinction that fulfill the role of distinguishing an individual or a piece of research from others explicitly. The awarding of such a distinction is enacted in a public ritual and, in the social and behavioral sciences, it manifests in practices of issuing awards, grants, or prizes to researchers and their institutions; issuing honorary degrees to academics and other public figures; but also – more recently – by evaluating and ranking research, researchers, and their institutions. Awards and prizes can relate to past research, teaching, or an entire oeuvre. Prizes can, however, also be awarded for prospective or ongoing research taking the form of research grants or (prestigious) scholarships. A prominent example for the latter are research grants issued by the European Research Council, established in 2007, which awards resources of up to € 2.5 million and is often referred to as an award. Honorary degrees have the function to honor at the same time the recipient of the honorary degree and the institution granting the honors by claiming a relationship between the institution and the person having received the degree.

In the arts and humanities, the practice of awarding prizes, medals, and other forms of distinction goes back at least to "Greek drama and arts competitions in the sixth century B.C." (English, 2002: 1–2). A broader array of prizes was established in the early Renaissance (English, 2002: 2) and in the eighteenth century another impetus came from royal and national academies, professional associations as well as learned societies (English, 2002: 2). In the eighteenth century, a series of contests awarding solutions to societal/scientific questions were advertised. These were organized to promote research in certain fields, a practice which has been taken up again recently by, for example, the Longitude Prize (Nesta, 2014). This latter prize takes reference to the historical Longitude Act of 1714 that offered a prize for a solution of the problem of longitude.

In contrast to that the distinctions offered by academies of science, which were established since the beginning of the seventeenth century, is mainly of symbolic nature and manifests in granting seats in the academies. Usually, ordinary members of these institutions vote to select a new member. The number of seats in academies of science being restricted, only a fixed number of individuals can hold a seat at the same time. This leads to the so-called phenomenon of the 41st chair (Merton, 1968). The 41st chair refers to the limitation of 40 seats in the Académie française and the resulting situation that a great number of distinguished personalities were not admitted into the Académie, such as Descartes, Flaubert, Molière, Pascal, Proust, Rousseau, Stendhal, or Zola. "In all [academies] there are occupants of the 41st chair, men outside the Academy having at least the same order of talent as those inside it" (Merton, 1968: 1).

With the establishment of the Nobel Prize in the early twentieth century, academic distinctions received an international dimension. The internationalization of scientific prizes in the late nineteenth century coincided with the internationalization of academia and the creation of international scientific associations. Ever since, the number of distinctions, awarded by an increasingly large variety of funders, grows continually from one year to the next.

A Typology of Awards

Although one could try to differentiate prizes from awards, for example, along the lines of whether a researcher needs to enter a competition to win, or whether the selection undergoes a nomination process and whether a monetary reward is included or not, empirical evidence suggests that there is no clear differentiation between these terms. However, a typology of awards, medals, and prizes can be suggested that also helps to differentiate between varying functions of these distinctions for the recognition of researchers, science, and associated institutions. Prizes and awards can thus be distinguished by funder; award category; disciplinary, geographical, and institutional scope; or selection procedure. These dimensions cannot be perceived as comprehensive as they overlap and, in many instances, complement each other.

Funder	*Award category*	*Boundaries of eligibility*	*Selection procedure*
Academies of science	Life work	Locality (national–international)	Peers
Foundations	Past research	Membership	Nonresearchers
Associations	Ongoing research	Disciplinary boundary	Competition
Companies	Future research	Nonacademic	Nomination
	Output	Academic	Transparency

Funders can broadly be differentiated between academic and nonacademic institutions, or on a slightly different level, whether a prize is given by members of the scientific community to honor a member thereof, or whether it is an external organization issuing an award to scholars (and other persons). Whereas prizes by sections of academic associations, learned societies, and the like are often quite insignificant in their monetary value their reputational value within a specific scientific community might be high. Thus the main function of these prizes are to honor the recipient of the award and bestow him/her with scientific capital and the (peer) legitimation to take a leading role in ongoing debates of the awarding epistemic community. Other relevant academic funders are international scientific associations, universities, and scientific journals. In contrast to that, there are a number of nonacademic institutions such as cities, regions, or ministries as public bodies on the one hand, and, on the other, private institutions like foundations, NGOs, or corporations that award prizes to scholars. Private institutions might carry out research and/or fund research, however, they are usually only operating on the boundaries of the academic field. The awards given by private funders often reflect aims of these organizations or, in some cases, have the function to influence the public perception of a company.

Prizes can be awarded in a multitude of categories. One (functional) distinction is between past (reaching from single studies to life work), ongoing, or future research. This distinction can be applied to prizes for young scholars, for lifework, or research grants for promising future research. Another category for awards – which in some cases might overlap with the previous differentiation – is along the lines of different types of outputs. The differentiation there can run along the lines of, for example, best book or best paper award, prizes for innovative teaching but also distinctions for museum installations would form such a category.

The 'boundaries of eligibility' describe the phenomenon that only members of certain groups are eligible to receive a certain award thus restricting competition. The geographical scope of an award is characterized by the dichotomy between locality and internationality. Local prizes restrict participation or nomination for example, to members of a university or researchers of a region, city, etc. and often serve to promote and fund local research. By defining thematic priorities, such local prizes can help to establish a research profile for a region, university, etc. Eligibility can also be restricted by requiring membership to specific organizations, associations, etc. In contrast to that are prizes that allow for any person to be eligible for nomination. In this latter category the number of potential nominees is, however, restricted by traditions connected to prizes, dominant ideas concerning the kind of research, and researcher to be distinguished and, the restricted scope of potential nominees known to those involved in the selection procedure. Third, disciplinary or thematic boundaries are defined in call texts, the general aims of an institution, or distinction, e.g., a prize can be designed for cognitive psychology, or open to all disciplines. Especially for international distinctions the boundaries of a prize are even broader, as some awards are of more general nature aiming at distinguishing personalities for their contribution to the 'advancement of humanity' at large. Fur these types of awards members of other social groups than scholars – such as political activists, journalists, writers, and public intellectuals at large – are equally eligible.

Two dimensions concerning the selection of award winners shall be discussed: First, the technical procedures and second, patterns influencing the reward system in science. Broadly speaking two technical procedures exist; the selection through application and the selection through nomination. In both scenarios a more or less complex set of actors, boards, referees from within the academic field or not, but also regulations and traditions associated to the prize need to be taken into account. The social procedure leading to the selection can, for many prizes and awards, be compared to that of peer review procedures leading to fellowships as studied for example by Lamont (2009). The nomination and selection by a board consisting of scholars and nonscholars is a pattern found prototypically for the selection of awardees of international distinctions. The combination of competition and peer review, in contrast, is characteristic for research grants and young researchers' awards. Another promising distinction for the study of awards is the degree of transparency chosen in the selection procedure. Transparency concerning the selection of a winner is not a necessary prerequisite for a prize to be esteemed of high quality. Some institutions try to legitimize their awards by offering explicitly transparent procedures.

International Prizes and Awards

In the social and behavioral sciences, a series of international prizes were established since the end of World War II. To become an asset to a scholar being distinguished with an award it needs continuous funding, visibility inside the relevant peer-group, and a recognition of its own.

International scope and the public recognition of a prize are not only serving the renown of researchers, but are also a resource for local prestige. It is well known, for example, that the Nobel Prize is Swedish. To serve the renown of the awarding institution – and the researcher – awards need to establish a reputation of their own. Awards obtain their reputation through input- and output-legitimizing procedures. High monetary endowment is an element that can help to obtain reputation, as it may increase the public visibility of a prize. However, the correlation between the monetary value and (public) recognition of a prize is not positive. Prizes with low endowment can become more reputable than awards offering a more significant reward, e.g., the Fields Medal in mathematics has a comparably low monetary endowment, is, however, regularly referred to as the Nobel for mathematics.

Transparent procedures leading to a distinction are another criterion, which may increase input legitimacy of a prize. This may include making nominations public and offering insight to the process leading to the selection of an awardee. However, there are examples where transparency is not given and the reputation of a prize mainly depends on the successful output – i.e., the continuous selection of broadly accepted awardees. Most prominent is the Nobel Prize, where "nominations, evaluations, deliberations, and final decisions concerning the prizes were to be kept secret" (Crawford, 2002: 3). Further, at the time of nomination it is unclear who was invited to propose nominations besides those with a permanent entitlement to nominate. In the case of the Nobels, only historical cases older than 50 years can be accessed and scrutinized. Ultimately, it is the output – i.e., selection of individuals and institutions that stand the moral, professional, and public scrutiny that will have the strongest influence on the reputation of the award and award-giving institution. Institutions awarding prizes thus have a high interest in selecting laureates that already have a certain reputation and will not deceive the expectations of the award givers. In the case of the Nobels, this may explain the increasing average age of awardees over time. Although the Nobels are awarded for the best research in a year, by choosing individuals that have already accomplished a great amount of their lifework the probability of unexpected future turns – that might then be associated with the Nobels – are low. In contrast to offering a high monetary endowment to increase public visibility, building a reputation of an award based on the achievements, and recognition of the researchers selected takes time. The character of a prize can also be established through other forms of public distinction: for example, in literature the Booker prize gained its reputation by "delivering annual scandals" (English, 2002: 114).

The Nobel Prize, one of the first international awards, managed to gain reputation in a short period of time. In the

social and behavioral sciences the 'Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel' was only introduced in 1969. Since, 74 laureates, thereof one woman, were distinguished with this prize. Although the endowment of $ 1.2 million in 2013 is higher than most other research prizes the awardees often only receive parts thereof as the prize can be divided among three scholars. The high reputation and public awareness for the Nobel Prize has transformed it into a reference for other prizes especially in domains where no Nobel is given. In mathematics, as mentioned, the Canadian Fields Medals compete with the – much higher endowed – Norwegian Abel prize for status of Nobel Prize in mathematics. A prize deliberately aiming at rewarding fields that are not covered by the Nobel Prizes is the John W. Kluge Prize hosted by the Kluge Center in the Library of Congress and first awarded in 2003 to Leszek Kolakowski. It is endowed with $ 1 million and covers disciplines such as history, philosophy, politics, anthropology, sociology, religion, criticism in the arts and humanities, and linguistics. Next to these Nobel substitutes there are also Nobel Prizes in categories, such as literature and the peace prize that were awarded to individuals related to the social sciences. Among others, the latter was awarded to three women with a more or less close relationship to social sciences: Jane Addams (1931), Emily Greene Balch (1946), and Alva Myrdal (1982) (for details on these three see Misztal, 2009). Also, Theodore Mommsen (1902) and Bertrand Russell (1950), a historian and a philosopher, received the Nobel Prize in literature.

In the natural sciences these international prizes awarded by academies and foundations have most recently been complemented by a series of highly endowed awards funded mainly by US-American, Russian, and Asian Internet entrepreneurs. These latter prizes define as their goal to make the sciences more visible and to raise the awareness of excellent science in certain fields. Among these prizes we find the 'Breakthrough Prize in Life Sciences' (2013) and the 'Fundamental Physics Prize' (2012) both endowed with $ 3 million or the 'Tang Prize' (2013) endowed with over $ 1.5 million (Merali, 2013). The exceptionally high monetary value of these prizes is complemented by a presentation of the laureates reminding events staged for performing arts. The individuals or institutions are put into the center of attention with great media impact. Such presentations provide the opportunity to present the prizewinner as well as the company/brand/institution hosting the prize.

The functions of awarding prizes are thus at least threefold: honoring the recipient, promoting the awarding organization and improving its public profile, and informing a larger public on scientific issues or the aims of an organization.

International Social Science and Humanities Prizes

Historically the first renowned international disciplinary prize was the already mentioned Nobel Prize in economics established on an annual basis in 1969 and first awarded to Ragnar Frisch and Jan Tinbergen for applied dynamic models for the analysis of economic processes (see Nobel Media, 2014). A prize with an even longer tradition is the Italian Antonio Feltrinelli Prize given by the Accademia Nazionale dei Lincei, which names itself the oldest academy founded in 1603. The Feltrinelli Prize is split in distinctions limited to Italian citizens and an international prize. In contrast to the Nobel Prize, the disciplines covered rotate since 1950 on a yearly basis and include the arts, music, literature, history, philosophy, medicine, and physical and mathematical sciences.

Other examples for awards of multidisciplinary nature that remain restricted to the social sciences and humanities, are for example, the European Amalfi Prize for sociology and social sciences (Premio Europeo Amalfi per la Sociologia e le Scienze Sociali) established by the Italian Association of Sociology and first awarded in 1987 to Norbert Elias. The selection was made by a jury that included distinguished social scientists. Another highly renowned prize is the Prince of Asturia Prize that is among others awarded in the categories of social sciences as well as communication and humanities. In the social sciences, it was first awarded in 1981 to Román Perpiñá. More recently we saw the creation of the Holberg Prize by the Norwegian government to distinguish research in the arts and humanities, social science, law, or theology, which is awarded 4.5 million Norwegian Krone. The first awardee in 2004 was Julia Kristeva.

Prizes of even broader scope that are among others relevant for social and behavioral sciences are those that commemorate humanistic goals. These prizes usually do not have disciplinary boundaries and are also awarded to personalities coming from nonacademic fields. Examples for early prizes of this category would be the Heine Prize, which is a biannual prize established by the city of Düsseldorf to commemorate the 175th birthday of Heinrich Heine and currently endowed with € 50 000 or the Peace Prize of the German Book Trade first awarded in 1950 with the aim to "reintroduce humanist thought into society" (Book Prize, 2014) in Germany. In 1958, the Erasmus Prize was installed "for a person who has made an exceptional contribution to culture, of scholarship, in Europe and beyond" (Erasmus Prize, 2014). It was first awarded to the people of Austria and was later awarded to distinguished scholars such as Claude Lévi-Strauss or Jürgen Habermas.

Honorary Degrees – Honoris Causa

Honoris causa or degrees awarded 'for the sake of honor' are distinctions honoring public figures for general societal achievements, support for academia, and/or for academic merits. Inside the scientific community, the practice of awarding honorary degrees is controversial. By borrowing names of academic merits (Bachelor, Masters, Doctor) honorary degrees suggest a honorification of academic work. The practice of awarding such degrees to individuals that have not gathered merits in academia, however, is inconsistent with idealized conceptions of science. By granting such degrees to (potential) donors, politicians, and other public figures to gain their (monetary) support, it is argued that the ideals of free and uninterested research are broken.

The tradition of granting honorary degrees is traced back to fifteenth century England, where academic degrees were a form of currency for being granted recognition by the Church and the State (Buxton and Gibson, 1935: 84). The earliest traces of an honorary degree lead to Lionel Woodville, Dean of Exeter, the brother-in-law of Edward IV and a future bishop who was awarded the first honorific degree in Oxford in 1478 or 1479.

These early honorary degrees furthermore granted considerable privileges in the university (Buxton and Gibson, 1935) whereas in modern times privileges are more restricted and are comparable to those granted to alumni (e.g., donating into alumni funds, using certain sporting facilities, faculty clubs, etc.).

The USA has a strong tradition of granting honorary degrees. Next to demonstrating proximity between leading public figures and academia, such honors also had the function to reward wider societal engagements. The first honorary doctorate is traced back to the sixth president of Harvard College in 1692, Increase Mather (Lady, 1967: 198). The title allowed Harvard College to become a University and grant doctoral degrees of its own. After independence in the late eighteenth century colleges conferred degrees to many military and political leaders. George Washington for example, received five awards (Lady, 1967: 199). In the 1870s, more honorary than earned PhDs were awarded in the USA (see Lady, 1967: 200), which even led to universities adopting regulations limiting the awarding of such degrees at the end of the century. Other institutions like the Massachusetts Institute of Technology, Rice, Cornell, Stanford and military academies have never or only in very rare circumstances awarded honorary degrees (see MIT News, 2001; Freiherr, 1979: 23).

Ranking Research

Public distinctions on the individual level come in the form of awards, prizes, medals, and the like. On the collective level public evaluation of research comes in the form of rankings of institutions of higher education, academic journals, etc. It is a phenomenon of the twentieth and twenty-first century. The rankings contribute in establishing a (seemingly) rationalization of the hierarchy inside and between institutions, researchers, and disciplines. The evaluation and, henceforth, ranking of research based on a limited number of indicators is often attributed to broader developments in the organization of knowledge-systems that took place since the end of World War II. The introduction of the so-called entrepreneurial university, public management, and a more general lack of legitimacy to spend public funds contributed to establishing a comprehensive system of public evaluation of research. Also, the massive expansion of the tertiary educational sector and the consequential growth of the academic field since the 1950s brought an unprecedented increase in competition among educational and research institutions on a global scale.

In the last decade, the Times Higher Education World University Rankings, the Academic Ranking of World Universities, also known as the Shanghai Ranking, and the QS World University Rankings are among the most influential rankings of universities on a global scale (for a comparison see StatSilk, 2014). The rankings use different sets of indicators to compare institutions and combine these with peer review procedures and surveys among researchers and students. On the European scale the Centre for Higher Education Development, founded by the German Rector's Conference and the Bertelsmann Foundation, evaluates German Universities since 1998 and selected universities on a European scale since 2010. In contrast to these global rankings, national mechanisms of collective evaluation that compare the performance (or what is understood as being an academic performance) are used by policy-makers to take decisions concerning the distribution of funds, the establishment of programs of excellence or the penalization of underperforming institutions.

Public evaluation of research does not only take place on the organizational level but is also practiced on the individual level. The evaluation of performance based on predefined indicators, e.g., the publishing of a certain amount of journal articles in a specified timeframe, is used to rationalize personnel decisions such as awarding tenure. A more open approach to evaluation is taken by social media platforms where – similarly to the rating of holiday destinations – regional, national, and international Internet-based services have developed portals for the evaluation of courses, lecturers, and professors.

See also: Becker, Howard (1928–); Bourdieu, Pierre (1930–2002); Habermas, Jürgen (1929–); Levi-Strauss, Claude (1908–2009); Merton, Robert K (1910–2003); Myrdal, Alva (1902–86); Recognition, Politics of: Perspectives from Social Work and Sociology; Reputation.

Bibliography

Becker, H.S., 1982. Art Worlds. University of California Press, Berkeley.
Book Prize, 2014. The History of the Peace Prize. http://www.friedenspreis-des-deutschen-buchhandels.de/445945/ (accessed 08.15.14.).
Bourdieu, P., 1976. Les modes de domination. Actes de la Recherche en Sciences Sociales 122–132.
Buxton, D., Gibson, S., 1935. Oxford University Ceremonies. Clarendon Press, Oxford.
Buxton, D.L., Strickland, B., 1935. Oxford University Ceremonies. Clarendon Press, Oxford.
Crawford, E.T., 2002. The Nobel Population 1901–1950: A Census of the Nominators and Nominees for the Prizes in Physics and Chemistry. Universal Academy Press, Tokyo, Japan.
English, J.F., 2002. Winning the culture game: prizes, awards, and the rules of art. Literary History 33, 109–135.
Erasmus Prize, 2014. Prize and Adornments. http://www.erasmusprijs.org/ (accessed 08.15.14.).
Freiherr, G.T., 1979. Is there an honorary doctor in the house? Change: The Magazine of Higher Learning 11, 22–25. http://dx.doi.org/10.1080/00091383.1979.10569646.
Haefner, A.E., 1958. Honoris causa. The Journal of Higher Education 29, 321–351. http://dx.doi.org/10.2307/1978862.
Helsvig, K.G., 2014. The Abel prize: the missing Nobel in mathematics?: the Abel prize. Centaurus 56, 1–30. http://dx.doi.org/10.1111/1600-0498.12038.
Hull, D.L., 1988. Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science, Science and its Conceptual Foundations. University of Chicago Press, Chicago.
IGNobel, 2014. About the Ig Nobel Prizes. http://www.improbable.com/ig/ (accessed 08.15.14.).
Lady, R.A., 1967. Honoris causa: an examination of the doctor of philosophy degree. The Journal of Higher Education 38, 197–205. http://dx.doi.org/10.2307/1979043.
Lamont, M., 2009. How Professors Think: Inside the Curious World of Academic Judgment. Harvard University Press, Cambridge, MA.
Lang, G.E., Lang, K., 1988. Recognition and renown: the survival of artistic reputation. American Journal of Sociology 94 (1), 79–109. Available at: http://www.jstor.org/stable/2781023.
Merali, Z., 2013. Science prizes: the new Nobels. Nature 498, 152–154. http://dx.doi.org/10.1038/498152a.

Merton, R.K., 1968. The Matthew Effect in science: the reward and communication systems of science are considered. Science (New York) 159, 56–63.

Misztal, B., 2009. A Nobel Trinity: Jane Addams, Emily Greene Balch and Alva Myrdal. The American Sociologist 40. http://dx.doi.org/10.1007/s12108-009-9081-2.

MIT News on Campus and around the World, 2001. No Honorary Degrees Is an MIT Tradition Going Back to … Thomas Jefferson. http://newsoffice.mit.edu/2001/commdegrees (accessed 08.15.14.).

Nesta, 2014. Longitude Prize 2014. http://www.longitudeprize.org (accessed 08.15.14.).

Nobel Media, 2014. All Prizes in Economic Sciences. http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/index.html (accessed 08.15.14.).

Rosenberg, R.P., 1961. The first American doctor of philosophy degree: a centennial salute to Yale, 1861–1961. The Journal of Higher Education 32, 387–394. http://dx.doi.org/10.2307/1978076.

Rossiter, M.W., 1993. The Matthew Matilda Effect in science. Social Studies of Science 23, 325–341. http://dx.doi.org/10.1177/030631293023002004.

StatSilk, 2014. Where Are the World's Top Universities? Interactive Maps Comparing Three Rankings: ARWU, THE and QS. http://www.statsilk.com/maps/where-are-worlds-top-universities-interactive-maps-comparing-three-rankings-arwu-the-qs (accessed 08.15.14.).

Zuckerman, H., 1992. The proliferation of prizes: Nobel complements and Nobel surrogates in the reward system of science. Theoretical Medicine and Bioethics 13.

Axiomatic Theories

Patrick Suppes, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

After considering the historical background of axiomatic methods since the beginning in ancient Greece, this article covers the main ingredients of methods for axiomatizing theories in mathematics or an empirical science within a standard set-theoretical framework. This approach has a natural division into the following steps. First, identify the primitive concepts of the theory being axiomatized; second, state explicitly the prior mathematical basis assumed; third, state the axioms; fourth, characterize models of the theory and give a definition of isomorphism of two such models. Examples and applications are given for various social–science theories, with an emphasis on measurement problems for such concepts as preference, subjective probability, utility, and stimulus intensity. Formal game-theory models for economics and quantum models for finance and decision making are briefly discussed.

After considering the historical background of axiomatic methods since the beginning in ancient Greece, this article covers the main ingredients of methods for axiomatizing theories in mathematics or an empirical science within a standard set-theoretical framework. This approach has a natural division into the following steps. First, identify the primitive concepts of the theory being axiomatized; second, state explicitly the prior mathematical basis assumed; third, state the axioms; fourth, characterize models of the theory and give a definition of isomorphism of two such models. Examples and applications are given for various social–science theories, with an emphasis on measurement problems for such concepts as preference, subjective probability, utility, and stimulus intensity.

Axiomatic development of theories is common practice in pure mathematics and is also now widely used in many sciences. The main ingredients of the methods for axiomatizing theories are the following: statement of the primitive concepts of the theory, statement of the prior mathematics basis assumed, statement of the axioms, characterization of models of the theory, and a definition of two such models having the same structure. Theories formulated in this way can easily satisfy the standard set-theoretical approach to axiomatization. The further step of formalizing the language of the theory, especially in the case of elementary theories, can lead to specific positive and negative results about the axiomatizability of theories in restricted languages.

Historical Background

Of all the remarkable intellectual achievements of ancient Greek civilization, none has had greater subsequent impact than the development of the axiomatic method of analysis. No serious traces are to be found in the earlier civilizations of Babylon, China, Egypt, or India. The exact history of the beginnings is not known, but elements that can now be identified emerged in the fifth century BC. A good reference is Knorr (1975). What can be said, and is important for the subsequent discussion here, is that, already in the next century, the fourth century BC, the detailed and elaborate theory of proportion of Eudoxus emerged in quite a clear and definite axiomatic form, most of which is preserved in Book V of Euclid's *Elements* (Euclid, 1925). What is important about Eudoxus's work, and even the commentaries of the work of this time, for example, by Aristotle, in the *Posterior Analytics* (1994 74a–17), is that the theorems were proved, not for single geometric objects, but for magnitudes in general, when applicable. The recognition of the correct abstraction in the general concept of magnitude, its technical and thorough implementation by Eudoxus and the philosophical commentary by Aristotle, represent a genuinely new intellectual development. The language of Eudoxus's famous Definition V in Book V of Euclid's *Elements* matches in its abstractness and difficulty the standards of modern axiomatic theories in mathematics and the sciences.

Definition 5 *Magnitudes are said to be in the same ratio, the first to the second and the third to the fourth, when, if any equimultiples whatever be taken of the first and third, and any equimultiples whatever of the second and fourth, the former equimultiples alike exceed, are alike equal to, or alike fall short of, the latter equimultiples respectively taken in corresponding order*.

The codification of the Greek axiomatic approach in Euclid's *Elements* was a great success and remained almost unchallenged until difficulties in the details were found in the eighteenth and nineteenth centuries, as discussed later.

Various scientific examples of axiomatic theories existed already in ancient Greek times and, of course, from the Greek standpoint, they were regarded as essentially homogeneous with the axiomatic theory of geometry, no sharp distinction being made between geometry and mechanics, for instance. A good example is Archimedes' set of partial qualitative axioms for measuring weights on balances. This is undoubtedly the first partial qualitative axiomatization of conjoint measurement, a form of measurement that has received much axiomatic attention and manifold applications in modern theories of measurement in the social sciences. (For detailed treatment of conjoint measurement, see Krantz et al., 1971. *See* Measurement Theory: Conjoint.)

Other early examples of axiomatic theories aimed at empirical matters can be found in the large medieval literature on qualitative axioms for the measurement of weight (Moody

and Clagett, 1952). Even more impressive examples are to be found in the medieval literature on physics. Some of the most subtle and interesting work is that of Nicole Oresme (1968) in the fourteenth century. (See especially his treatise, Tractatus de Configurationibus Qualitatum et Motuum.) What is surprising is Oresme's unexpectedly subtle approach in a geometrical framework to the phenomena of intensive qualities and motions. An example for Oresme would be that of a surface being more or less white.

Finally, in the geometric axiomatic tradition of analyzing phenomena in natural science, the two great late examples of deeply original work were the seventeenth century treatises of Huygens' *The Pendulum Clock* (Huygens, 1673/1986) and, of course, as the second example, Newton's *Principia* (Newton, 1687/1946). Both Huygens and Newton formulated their axioms in the qualitative geometric style of Euclid and other Greek geometers 2000 years earlier.

Here is the English translation from the Latin of Newton's purely verbal formulation, in the language of proportion, of his first three justly famous axioms or laws of motion:

> LAW I
> Every body continues in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed upon it.

> LAW II
> The change of motion is proportional to the motive force impressed; and is made in the direction of the right line in which that force is impressed.

> LAW III
> To every action there is always opposed an equal reaction: or, the mutual actions of two bodies upon each other are always equal, and directed to contrary parts. Newton (1687/1964), p. 13

Axiomatic Geometry in the Nineteenth Century

Without question, development in axiomatic methods in the nineteenth century was the perfection and formalization of the informal Greek methods that had dominated for so many centuries. The initial driving force behind this effort was certainly the discovery and development of non-Euclidean geometries at the beginning of the century by Bolyai, Lobachevski, and Gauss. An important development later in the century was the discovery of Pasch's (1882) axiom as a necessary addition to the Euclidean formulation. Pasch found a gap in Euclid's work, which required a new axiom, namely, the assertion that if a line intersects one side of a triangle, it must also intersect a second side. This was just the beginning for Pasch. He created the modern conception of formal axiomatic methods, which has been the central aspect of the model for axiomatic workup until present times. The axiomatic example in geometry that had the most widespread influence was Hilbert's *Grundlagen der Geometrie*, first published in 1897 and still being circulated in later editions (Hilbert, 1956).

Ingredients of Standard Axiomatic Practice

Building, especially, on the work in the later part of the nineteenth century in axiomatizing geometry, early in the twentieth century there was widespread axiomatization in mathematics and, to a lesser extent, in the empirical sciences. The main ingredients of the methods for axiomatizing theories were the following: statement of the primitive concepts of the theory, statement of the prior mathematics basis assumed, statement of the axioms, characterization of models of the theory and a definition for isomorphism of two such models, proof of a representation theorem when possible, and, finally, some analysis of invariance of the models of the theory.

Before turning to an explicit discussion of these ingredients, it is worth noting that the emphasis should really be on models of the theory, which is what axiomatizing a theory makes clear. It is the models of the theory, i.e., structures, that satisfy the theory, as explained in the section Independence of Axioms, and that exhibit the nature of the theory, whether it be in geometry, economics, psychology, or some other science. (For an elementary, but detailed, account of the concepts discussed in this section, see (Suppes 1957/1999, Chapters 8 and 12).)

Primitive Concepts of a Theory

The first point to recognize in axiomatizing a theory is that some concepts are assumed as primitive. Their properties are to be stated in the axioms, and, therefore, it is important to know how many such concepts there are and what their general formal character is, such as relations, functions, etc. In Hilbert's axioms for elementary plane geometry, the five primitive concepts are those of point, line, plane, betweenness, and congruence. In contrast, in psychological theories of measurement, an ordering of the stimuli or other phenomena is almost always assumed – in this case, a weak ordering, that is, a binary relation that is transitive and connected, rather than the geometric ordering of betweenness.

In addition to a primitive concept of ordering, there are other relations, for example, a primitive concept of the comparison of stimulus differences or an operation of combination for extensive measurement, as in the case of subjective probability. These measurement examples also apply to a large literature on utility and subjective probability in economics. In psychological theories of learning, even of the simplest nature, one would need such concepts as that of stimulus, response, and a conditioning relation between stimulus and response. Theories without much more elaboration would require the important concept of similarity or resemblance, in order to develop a theory of the fundamental psychological phenomena of generalization, discrimination, and transfer.

There are also important theories that use much simpler primitive concepts. A good example would be the theory of zero-sum, two-person games in normal form. The primitive concepts are just that of two nonempty sets X and Y that are arbitrary spaces of strategies for the two persons and a numerical function M, defined on the product space $X \times Y$. The intuitive interpretation of M is that it is the payoff or the utility function for the player whose space is X. The negative of that is the payoff for the player whose space is Y. Later, a representation theorem is given for finite games of this sort.

Axioms as Defining Theories

As is widely recognized, axioms are what intuitively characterize a theory, whether it be of geometry, game theory, or a psychological theory of measurement. From a formal standpoint, something more can be said. The essence of the matter is that, to axiomatize a theory is to define a certain set-theoretical predicate. This is just as valid in the empirical sciences as in pure mathematics. The axioms, of course, are the most important part of such a definition.

Because the theory of weak orderings is used widely in both economics and psychology, as part of the theory of choice, it will be useful to give a formal definition and, therefore, the axioms for a weak ordering.

Definition 1 *Let A be a nonempty set and R a binary relation on A, i.e., let R be a subset of the product space* $A \times A$. *A structure (A, R) is a weak ordering if and only if the following two axioms are satisfied, for every a, b and c in A:*

Axiom 1 *R is transitive on A, i.e., if aRb and bRc then aRc.*

Axiom 2 *R is connected on A, i.e., aRb or bRa.*

There are various general methodological and folklore recommendations about the way in which axioms should be written – clarity, lack of ambiguity, and simplicity are familiar. More substantive results about the form of axioms are discussed in the section Theories with Standard Formalization or first-order formalization. There are many definite mathematical results about the form of axioms, some of which are also discussed later.

Independence of Axioms

An early and important substantive recommendation is that the axioms be independent, i.e., none can be derived from the others. The question then arises from a methodological standpoint: how is independence to be established? The answer is in terms of models of the axioms. To make the matter still more explicit, consider possible realizations of the theory. These are set-theoretical structures that have the right form. For example, in the case of weak orders, a possible realization is an ordered pair consisting of a nonempty set and a binary relation on that set. Such an arbitrary pair is called a possible realization because as a possible realization it is not a model of the theory unless the axioms of the theory are also satisfied. An obvious example of a model of the theory of weak ordering is the pair consisting of the set of positive integers and the relation of weak inequality $\geq$.

The independence of a given axiom of a theory is established by finding a possible realization of the theory in which all the axioms, except the particular one in question, are satisfied. The deductive argument to show that this yields a proof of independence is intuitively obvious and will not be made in a more formal manner here. However, the essence is that if the concept were not independent, but definable, then it would necessarily hold in the model, just like the remaining axioms, and so a contradiction of its both holding and not holding in the given model would be obtained. Therefore, by reductio ad absurdum it must be independent. It will be useful to consider an example or two.

To show that the axiom of connectedness for weak orders is independent of the axiom of transitivity, it is necessary to take a possible realization that is also a transitive relation. Let $A = \{1, 2, 3\}$ and $R = \{1, 2\}$. In other words, R is just the strict numerical $<$ relation on the subset $\{1, 2\}$ of A. Then it is obvious that this relation is transitive but not connected for the number 3, which is in the set A, and does not stand in the relation R to any other element in A. To show that transitivity is independent of connectedness, that is, Axiom 1 for weak orders is independent of Axiom 2, it is sufficient to take the same set A as before, but now, the relation R is the set of the ordered pairs $\{(1, 2), (2, 3), (3, 1), (1, 1), (2, 2), (3, 3)\}$. Then it is clear that any two elements in the set A are connected by the relation R, but the relation R is not transitive, for, in order to be transitive, it must also have the pairs $(1, 3)$, $(2, 1)$, and $(3, 2)$.

The examples of independence given are trivial, but it is to be emphasized that it can often be a difficult task to establish whether or not an axiom is independent of the remaining axioms; that is, whether or not it can be derived from the remaining axioms. A very familiar mathematical example, with a long history in the twentieth century, is the proof that the axiom of choice is independent of the other standard axioms of set theory.

Padoa's Principle for Proving Independence of Primitive Concepts

Less familiar than proving the independence of axioms is the method of using models of a theory to prove independence of the primitive concepts of the theory. To prove that a particular primitive concept of a theory, for example, the notion of congruence in Euclidean geometry, is independent of the other primitive concepts, it is sufficient to find two models of the theory, such that the domain of both models is the same, the two models are the same for all the other primitive concepts of the theory, but the two models differ only in their realization of the concept in question. Thus, to prove congruence independence, what is required are two different models of the axiom in question, which are the same for all the other concepts, such as point, line, etc., but which have two distinct notions of congruence.

In the case of weak orders, Padoa's principle can be used in an obvious way to show that the concept of the binary relation is independent of the given set A. It suffices, for example, to use the two different orderings $\leq$, $\geq$ on the set of positive integers. On the other hand, since R is connected, the set A is definable in terms of the relation R for the special case of weak orderings by taking the union of the domain and range of the relation. But for different orderings, for example, partial orderings, which are reflexive, antisymmetric, and transitive on the set A, it is easy to show that the set A is an independent concept. Just let A in one model be a proper subset in the other, and elements in the relation come only from the first set.

Isomorphism of Models of a Theory

The separation of the purely set-theoretical characterization of the possible realizations of a theory from the axioms proper, which characterize models of a theory, is significant in defining certain important concepts. For example, the notion of two

models of a theory being isomorphic is often said to be axiom-free, since the definition of isomorphism of the models of a theory really depends on just the set-theoretical characterization of the possible realizations. A satisfactory general definition of isomorphism for the structures that are possible realizations of any sort of theory is difficult, if not impossible, to formulate. The usual practice is to formulate a special definition for the possible realizations of some particular theory. This is what will be done here, as already illustrated in the case of binary relations. A possible realization is a set-theoretical structure that is a nonempty set and a binary relation whose domain and range are included in that set. In this case the definition of isomorphism is as follows.

Definition 2 *A binary relation structure (A, R) is isomorphic to a binary relation structure (A', R') if, and only if, there is a function f such that*

1. *$D(f) = A$ and $R(f) = A'$*
2. *f is a one–one function*
3. *For a, b in A, aRb if and only if $f(a)R'f(b)$*

The definition of isomorphism for possible realizations of a theory is used to formulate a representation theorem, which has the following meaning. A certain class of structures or models of the axiomatized theory is distinguished for some intuitive or systematic reason and is shown to exemplify within isomorphism every other model of the theory. This means that, given any model of the theory, there exists in this distinguished class an isomorphic model. A good example of this can be found in the axiomatic theory of extensive measurement. Given any empirical structure that represents the data about subjective judgments of probability and satisfies the axioms of the theory, there is a numerical structure satisfying the axioms that is isomorphic to the empirical structure. Note, of course, that there is not any one single such isomorphism. Different individuals can have different empirical structures realizing their subjective probabilities, but there will be for each of them a particular numerical model of the axioms that is isomorphic to each given empirical structure.

Here is a simple and obvious, but useful, example of a representation result for two-person, zero-sum games that are finite. A game (X, Y, M), as introduced earlier, is finite just in case the sets X and Y are finite. Therefore, with the definition of isomorphism obvious from the result now to be stated, any finite game $G = (X, Y, M)$ with $X = (x_1, \ldots, x_m)$ and $Y = (y_1, \ldots, y_n)$ is isomorphic to the game $G' = (I_m, I_n, M')$, and $M'(i, j) = M(x_i, y_j)$, where the notation I_k denotes the set of positive integers $1, \ldots, k$. Therefore, any finite two-person zero-sum game may be represented by a game where X and Y are initial segments of the positive integers, even if vivid substantive descriptions of the individual strategies x_i and y_j have been given.

The use of the formal structure of games has been one of the most striking innovations in economic theory in the second half of the twentieth century, inspired above all by the publication in von Neumann and Morgenstern's (1944), *Theory of Games and Economic Behavior*. The formal isomorphism between many kinds of economic structures and games has been a surprisingly robust theoretical development. A recent elementary introduction to these ideas is Fudenberg and Levine (1988), and a more advanced text is Fundenberg and Tirole (1991).

Quantum Decision Theory

More recently than game theory, and much less well developed, is the introduction into psychology, as well as economics, of models with formal features isomorphic to those of quantum mechanics. It can be argued that the strange nonlocality to be found in quantum entanglement of distant particles is not relevant, but the contextuality, exhibited especially in quantum interference, is the important common feature (de Barros and Suppes, 2009). Applications, and further references, to the psychology of decision making are found in Busemeyer et al. (2006). Quite different applications are found in recent physics-inspired theories of financial markets, e.g., Baaquie (2004) and Haven (2002).

Invariance Theorems for Axiomatized Theories

In addition to having representation theorems for the models of a given axiomatized theory, it is also significant and useful to have explicit invariance theorems. The intuitive idea of invariance is most naturally explained in terms of either geometric theories or measurement theories.

Given a class of models, for example, analytical models in geometry and numerical models in measurement, with respect to which any other model of the theory can be represented, the invariance theorem is an expression of how unique the representation in terms of such an analytic model is. For example, in the case of the representation theorem for axiomatic formulations of Euclidean geometry, the invariance theorem states that any analytic model is unique only up to the group of Euclidean motions. This means that for any analytic representation of Euclidean geometry, any reflections, rotations, or translations of the analytic model will produce a new analytic model, also isomorphic to the set-theoretical model satisfying the qualitative synthetic axioms of Euclidean geometry. In fact, the most general theorem widens the scope to what is often called the group of generalized Euclidean transformations, namely, those that also permit a change in scale, so that no unit of measurement is fixed.

The situation is very similar in the case of theories of measurement. Given an empirical structure and a representing numerical model isomorphic to that structure, then, in the case of intensive quantities such as cardinal utility, for example, the numerical model is unique only up to the group of affine transformations, that is, transformations that multiply the unit of measurement by a positive real number and add a constant as well. This means that in fundamental measurement of a cardinal utility there is nothing but convention in the choice of a unit of measurement or a zero value.

Classical examples of important invariant theorems in physics are that structures of classical physics are invariant up to Galilean transformations, and in special relativity, invariant up to Lorenz transformations – in both these cases, further generalizations arise by also permitting changes in the units of measurement.

Theories with Standard Formalization

The set-theoretical framework for axiomatizing theories just discussed is used implicitly throughout the mathematical social sciences. Most of the axiomatic work that takes place in the social

sciences can be put in a straightforward way within the framework just described. On the other hand, there is a tighter framework for discussing the axiomatization of theories that leads to clarification of problems that arise in the conceptual or qualitative thinking about theories in the social sciences. The purpose of this section is to describe, in an informal way, this more narrowly defined framework and to give a sense of the kind of results that can be obtained. Most of the results are of a negative sort, important because they show what cannot be done by apparently very natural qualitative formulations of theories of preference, of subjective probability or related kinds of qualitative measurement, or scaling problems, especially in economics and psychology, and in anthropology, political science, and sociology.

A language with standard formalization is a language that is given a precise formulation within first-order logic. Such a logical framework can be characterized easily in an informal way. This is the logic that assumes

1. one kind of variable;
2. logical constants, mainly the sentential connectives such as *and*;
3. a notation for the universal and existential quantifiers; and
4. the identity symbol $=$.

A language formulated within such a logical framework is often called an elementary language. Ordinarily, three kinds of nonlogical constants occur in axiomatizing a theory in such a language – the relation symbols, also called predicates; the operation symbols; and the individual constants.

The grammatical expressions of the language are divided into terms and formulas, and recursive definitions of each are given. The simplest terms are variables or individual constants. New terms are built up by combining simpler terms with operation symbols in a recursive manner. Atomic formulas consist of a single predicate and the appropriate number of terms. Compound formulas are built up from atomic formulas by means of sentential connectives and quantifiers.

Possible realizations of elementary theories, i.e., theories formulated in an elementary language, assume an especially simple form. First, there is a nonempty domain for the structure. Second, corresponding to any relation symbol of the theory, there is a corresponding relation, including sets representing predicates as one-place relations. Corresponding to any operation symbols in the theory are operations defined on the domain of the structure, and, finally, individual objects of the domain correspond to the individual nonlogical constants of the theory. It is worth noting in passing that the definition of isomorphism for such elementary structures is straightforward and simple, which is not always the case for theories formulated in more complicated set-theoretical languages.

Some Positive Results about Axiomatizability

The first positive result uses two concepts. First, the set of all the finite models of a theory is called a finitary class of elementary structures. Second, a theory is recursively axiomatizable when there is an algorithm for deciding whether or not any formula of the language is an axiom of the theory in question.

Theorem 1 *Any finitary class of models of an elementary theory is axiomatizable, but not necessarily recursively axiomatizable.*

The importance of this result shows that the expressive power of elementary languages is adequate for finitary classes but not necessarily for the stating of a set of recursive axioms.

A more special positive result about finitary classes of models can also be given:

Theorem 2 *Let K be the finitary class of measurement structures with respect to an elementary language L and with respect to a numerical model R of L such that K includes all finite models of L homomorphically embeddable in R. If the domain, relations, functions, and constants of R are definable in elementary form in terms of* $(Re, \leq, +, \cdot, 0, 1)$ *then the set of sentences of L that are satisfied in every model of K is recursively axiomatizable.*

A universal sentence is one that has only universal quantifiers at the beginning with the scope of a quantifier the remainder of the sentence. In practice, such sentences are written as quantifier-free statements to simplify the notation. The axioms of weak ordering given earlier are of this kind. To be explicit, the conjunction of the two given axioms is a universal sentence, and this single universal sentence is, in this form, the single axiom for a weak ordering.

Theorem 3 (*Vaught's Criterion, 1954*). *Let L be an elementary language without function symbols. A finitary class K of measurement structures (with respect to L) is axiomatizable by a universal sentence iff K is closed under submodels and there is an integer n such that if any finite model M of L has the property that every submodel of M with no more than n elements is in K, then M is in K.*

The intuitive idea of Vaught's criterion for finitary classes of models of a theory is easy to explain. Consider again weak orderings. Because the axioms involve just three distinct variables, it is sufficient to check triples of objects in the domain of an empirical structure to determine if it is a model of the theory. Generally speaking, the number of distinct variables determines the size of the submodels that must be checked to see if universal axioms are satisfied. To have a universal axiom for a theory, or what is equivalent, a finite set of universal axioms, it is necessary that the number of distinct variables be some definite number, say n. By examining submodels involving no more than n objects, it is then sufficient to determine satisfaction of the axiom or axioms, and this is the intuitive idea of Vaught's criterion.

Some Negative Results About Axiomatizability

To begin with, it is useful to state as a theorem what is a corollary of Vaught's criterion. It is simply the negation of it for determining when a finitary class of models of a theory is not axiomatizable in an elementary language by a universal sentence.

Theorem 4 *Let L be an elementary language without function symbols and let K be a finitary class of measurement structures (with respect to L) closed under submodels. Then K is not axiomatizable by a universal sentence of L iff for every integer n there is a finite model M of L that has the property that every submodel with no more than n elements is in K, but M is not in K.*

It is natural to interpret the negative result stated in the general form here as showing that the complexity of relationships in finitary classes satisfying the hypotheses of the theorem

is unbounded, at least unbounded when the theory must be expressed by quantifier-free formulas of elementary logic.

The first application is something rather surprising. A semiorder is a structure $(A, \succ)$ satisfying the following three axioms.

Axiom 1 It is not the case that $(a \succ a)$.
Axiom 2 If $(a \succ b \,\&\, a' \succ b')$ then $(a \succ b' \vee a' \succ b)$.
Axiom 3 If $(a \succ b \,\&\, b \succ c)$ then $(a \succ d \vee d \succ c)$.

When the set A is finite, the following numerical representation holds:

$$\text{for every } a \text{ and } b \text{ in } A,\ a \succ b \text{ if and only if } f(a) > f(b) + 1.$$

What is now surprising is a result about the indistinguishability relation for semiorders, that is, the relation $\sim$ that is the negation of the characterization of the semiorder, namely, for a and b in A.

$$|f(a) - f(b)| \leq 1 \text{ iff } a \sim b.$$

The following theorem is from Roberts (1969).

Theorem 5 *Let L be the elementary language whose only nonlogical symbol is the binary relational symbol* $\sim$. *Then the finitary class J of measurement structures for the indistinguishability relation* $\sim$ *is not axiomatizable in L by a universal sentence.*

The next case of a negative result applying the negation of Vaught's criterion (Theorem 3) is the proof that the qualitative theory of utility differences or the qualitative theory, more generally, of various psychometric sensations is not axiomatizable by a universal sentence, contrary, of course, to the simple theory of order. This result is from Scott and Suppes (1958).

Consider the elementary language whose only symbol is the quaternary relation symbol 'D' with the intended numerical interpretation:

$$ab\ D\ cd \text{ iff } f(a) - f(b) \geq f(c) - f(d)$$

We then define the finitary class of measurement structures for algebraic difference as consisting of all models (A, D) such that

1. A is a nonempty finite set;
2. D is a quaternary relation on A; and
3. (A, D) is isomorphic to (A', Δ), where A' is a finite set of numbers and Δ is the quaternary numerical relation such that for real numbers x, y, u, and v, $xy\,\Delta\,uv$ iff $x - y \geq u - v$.

Theorem 6 *Let L be the elementary language whose only nonlogical symbol is the quaternary relation symbol D. The finitary class D of measurement structures for (algebraic) difference is not axiomatizable in L by a universal sentence.*

The intuitive idea of the proof can be seen from the construction of a 10-element structure, all of whose substructures have a numerical representation, but which does not itself have such a representation. The idea of this construction can then be generalized for arbitrary n in order to apply Theorem 4.

Using the same ideas, Titiev (1972) extended the results of Theorem 6 to additive conjoint measurement and to multidimensional scaling with the Euclidean metric. Titiev (1980) also gives a negative proof for the city block metric for $n \leq 3$ dimensions. Using more sophisticated logical results, it is possible to extend most of the results just stated to not being finitely axiomatizable. This means that existential quantifiers can be introduced, but the number of axioms must be finite in character. The main results here are those reported by Per Lindstrom, which together with the other results mentioned in this section are presented in detail in Luce et al. (1990).

Theorem 7 *The finitary class of measurement structures for algebraic difference is not finitely axiomatizable in the elementary language whose only nonlogical symbol is the quaternary relation symbol D.*

In a way, a still stronger negative result about axiomatizability can be proved for Archimedean axioms. Here is one standard formulation of such an axiom. If $a \geqslant b$ then for some n, $nb \geqslant a$, where nb is the combination of n copies of b. Notice that it is necessary to introduce a quantifier for integers not just for the empirical objects of the domain. Because the exact formulation of the negative result is rather complicated, the following informal theorem is given.

Theorem 8 *For any standard elementary language used to formulate a theory of measurement, there is no set of elementary formulas of the language equivalent to an Archimedean axiom for the theory.*

The best way to think about the Archimedean axiom in this context is that it is a second-order axiom and, therefore, cannot be formulated by an equivalent set of formulas, in first-order logic.

Still another way of looking at this result is that to characterize the real numbers we need some sort of second-order axiom such as Dedekind completeness, the Cauchy sequences, or the least-upper bound axiom, but none of these axioms, including the Archimedean axiom, can be formulated in a first-order language whose variables take real numbers as values. Narens (1974) provides a general account of Archimedean axioms in various forms.

For deeper and more general results on axiomatizability there is more recent work that can be cited, especially concerning the definability of linear orders for classes of finite models and the problem of the complexity of the class (Stolboushkin, 1992; Gurevich and Shelah, 1996; Hella et al., 1997).

See also: Mathematical Models in Philosophy of Science; Measurement Theory: Conjoint; Measurement Theory: History and Philosophy; Measurement, Representational Theory of; Ordered Relational Structures.

Bibliography

Aristotle, 1994. Posterior Analytics. Clarendon Press, New York.
Baaquie, B., 2004. Quantum Finance: Path Integrals and Hamiltonians for Options and Interest Rates. Cambridge University Press, pp. 332.
Busemeyer, J.R., Wang, Z., Townsend, J.T., 2006. Quantum dynamics of human decision-making. Journal of Mathematical Psychology 50, 220–241.
de Barros, A., Suppes, P., 2009. Quantum mechanics, interference, and the brain. Journal of Mathematical Psychology 53, 306–313.
Euclid, 1925. Euclid's Elements Book V [Trans. Heath, T.L.] Cambridge, UK [1956 Dover Publications, New York]. Cambridge University Press.
Fudenberg, D., Levine, D.K., 1988. The Theory of Learning in Games. MIT Press, Cambridge, Massachusetts.

Fudenberg, D., Tirole, J., 1991. Game Theory. MIT Press, Cambridge, Massachusetts.
Gurevich, Y., Shelah, S., 1996. On finite rigid structures. Journal of Symbolic Logic 61, 549–562.
Haven, E., 2002. A discussion on embedding the Black-Scholes option pricing model in a quantum physics setting. Physica A. http://dx.doi.org/10.1016/S0378-4371(01) 00568-4.
Hella, L., Kolaitis, P.G., Lousto, K., 1997. How to define a linear order on finite models. Annals of Pure and Applied Logic 87, 241–267.
Hilbert, D., 1956. Grundlangen der Geometrie, eighth ed. Teubner, Stuttgart, Germany.
Huygens, C., 1673/1986. The Pendulum Clock, or Geometrical Demonstration of Pendula as Applied to Clocks. Trans. Rockwell, R.J. Iowa State University Press, Ames, IA.
Knorr, W.R., 1975. The Evolution of the Euclidean Elements. Reidel, Dordrecht, The Netherlands.
Krantz, D.H., Luce, R.D., Suppes, P., Tversksy, A., 1971. Foundations of Measurement, vol. I. Academic Press, New York.
Luce, R.D., Krantz, D.H., Suppes, P., Tversky, A., 1990. Foundations of Measurement vol. III, Representation, Axiomatization and Invariance. Academic Press, New York, CA.
Moody, E.A., Clagett, M., 1952. The Medieval Science of Weights (Scientia de Ponderibus). University of Wisconsin Press, Madison, WI.
Narens, L., 1974. Measurement without Archimedean axioms. Philosophy of Science 41, 374–393.
Newton, I., 1687/1946. Principia. Trans. Cajori, F. University of California Press, Berkeley, CA.
Oresme, N., 1968. Tractatus de Configurationibus Qualitatum et Motuum. Trans. Clagett, M. University of Wisconsin Press, Madison, WI.
Pasch, M., 1882. Vorlesungen über Neuere Geometrie. Verlang von Julius Springer, Leipzig, Germany.
Roberts, F.S., 1969. Indifference graphs. In: Harary, F. (Ed.), Proof Techniques in Graph Theory. Academic Press, New York.
Scott, D., Suppes, P., 1958. Foundational aspects of theories of measurement. Journal of Symbolic Logic 23, 113–128.
Stolboushkin, A.P., 1992. Axiomatizable classes of finite models and definability of linear order. In: Proceedings of the 7th Annual IEEE Symposium on Logic in Computer Science. IEEE Computer Society Press, pp. 64–70.
Suppes, P., 1957/1999. Introduction to Logic. Van Nostrand, New York.
Titiev, R.J., 1972. Measurement structures in classes that are not universally axiomatizable. Journal of Mathematical Psychology 9, 200–205.
Titiev, R.J., 1980. Computer-assisted results about universal axiomatizability and the three-dimensional city-clock metric. Journal of Mathematical Psychology 22, 209–217.
Vaught, R., 1954. Remarks on universal classes of relational systems. Indagationes Mathematicae 16, 589–591.
von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press, Princeton, New Jersey.

B

Balance of Power, History of

G van Benthem van den Bergh, The Hague, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 2, pp. 1036–1039, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

'Balance of power' in the relations between states can refer first to a situation in which the Great Powers – regional or global – possess roughly equal power resources. But the concept acquired a more specific meaning in the eighteenth century as a policy of Great Powers to prevent one of them from becoming the hegemonic power. This required that alliances between them should be geared to that purpose and thus be flexible. Balance of power policy was quite successful between reasonably satisfied dynastic states whose power resources were reasonably even. But the policy failed when that power balance suddenly became more uneven, as in the cases of Jacobin and Napoleonic France or the unification of Germany by Prussia. Balance of power policy developed further after the Congress of Vienna into the idea and practice of the 'Congress of Europe,' based on the premise of the common interest of the Great Powers in maintaining peace and stability. Because the Great Powers were rivals as well as allies, the 'Concert' idea could not prevent World War I. But it forms the basis of all later attempts to preserve or restore peace between states through collective security and international organization, such as the Concert of Europe, the League of Nations, and the United Nations.

The concept 'balance of power' can be confusing. It has two different, though interrelated, meanings. It can refer first to a situation in which the Great Powers of a particular figuration of states – whether regional or global – possess roughly equal power resources. A power equilibrium is unlikely; the differences must just be small enough to make the risk of going to war too high. To achieve hegemony or empire a highly uneven power balance is required.

A balance of power can be bilateral as well as multilateral. It was considered crucial in the hegemonial rivalry between the Soviet Union and the United States. The rivals saw arms competition as the primary indicator of the balance between them. However, by balance they in fact meant superiority. The unintended result was stalemate. The nuclear danger, at the time called a 'balance of terror,' chained them together and made 'superiority' meaningless.

Balance of power in the first sense thus refers to the state of the power balance between interdependent, rivaling states or their predecessors. The first theory about the power balances in competing kingdoms was probably formulated by Kautilya, around 450 BC in India. When states were first formed in Europe, their rivalry was guided by the maxims of 'reason of state.' But gradually their perception of common interests became stronger, as reflected in the development of diplomacy and international law and the creation of international organizations.

Balance of power in its second meaning was part of the process in which cooperation between Great Powers increased without removing their rivalry. It came to mean a policy to prevent a bid for hegemony of one of the Great Powers, by the threat of the other Great Powers allying against it. If deterrence of a hegemonist failed the alliance could become a war coalition to restore the balance. But the purpose of balance of power policy itself was to maintain the status quo. To succeed, the power differentials between the Great Powers should remain small enough to foreclose hegemonic aspirations. The first meaning of balance of power was therefore a condition for the second.

Background

How and why did balance of power policy emerge? The European peace conference at Westphalia in 1648 to end the many conflicts comprising the Thirty- and Eighty-Years (Dutch independence) Wars was not yet a joint attempt to develop a more stable balance of power in Europe. But the conference brought the representatives of the hitherto separate northern and western power balances together. Though the participating states in Europe recognized one another as such by treaty, repudiating the prior claims to supreme authority of empire and church, dynastic rivalry remained. As Wedgwood remarked about the long delay of actual negotiations in Munster: "The ruling powers … asked for peace always in a general sense: when it came to practical action they were always prepared to fight for a little longer in order to gain their own particular end" (Wedgwood, 1963: p. 474). The peace conference became possible after expansionist ends could no longer be reached. Popular sentiment against war, very strong in the devastated German lands, did not make much difference, however. War in the dynastic period was characterized by Kant as a 'pleasure

party' for princes, who themselves did not suffer from them as the people did. For durable peace, dynastic states should be replaced by republics, by their nature responsive to the people, and willing to form a commonwealth together. Indeed, when the power distribution between dynastic states was still fluid and difficult to assess as in the sixteenth and seventeenth centuries, wars were frequent.

The peace of Westphalia simultaneously ended a plurality of conflicts. It made France more powerful and reduced Habsburg power. It also contributed to the rise of new Great Powers like Russia and Prussia and to the fall of Spain, Sweden, and later the Netherlands. One European power balance came into being. The monopolization processes that had led to the forming of dynastic states out of feudal fragmentation continued in the rise of Great Powers.

In the larger dynastic states, the increase of the king's power and authority had to be at the expense of other centers of power, such as the warrior nobility and autonomous towns. The domestic consolidation of dynastic regimes made stability in interstate relations more desirable. Balance of power policy came in useful. But it was precarious. Power resources changed all the time, so attempts at acquiring hegemony and war between Great Powers could not be excluded.

The Classical Balance of Power in Europe

Balance of power policy was practiced before it became an explicit principle. The power of Louis XIV's France grew to such an extent after the Peace of Westphalia that the other European powers feared it might acquire hegemony. A French attempt to use the Spanish succession to increase its power aroused Britain into forming in 1701 in The Hague an alliance of the other Great Powers against France. This coalition won the war and the Treaty of Utrecht of 1713 considerably reduced the power of France. That treaty also readjusted power relations all over Europe, and France lost some colonies to Britain. But it followed the prescription of balance of power policy that one's adversaries should not be destroyed, as that would upset the power balance between the allies.

In the eighteenth century five Great Powers emerged that would remain until 1914. They included one global and colonial sea power (Britain), two continental powers (Prussia and Austria), one power with both continental and maritime colonial ambitions (France), and an imperial power, with its territory stretching beyond Europe (Russia). Though the Ottoman Empire was in practice part of the European power balance, it was not a member of the quintet. Though the power resources of the five Great Powers were quite different, they were not too widely divergent most of the time to make balance of power policy impossible.

What exactly was classical balance of power policy? As the pamphlet, *Europe's Catechism* (1741) defines it:

> The Balance of Power is such an equal Distribution of Power among the Princes of Europe, as makes it impracticable for the one to disturb the Repose of the other … It may be destroyed by Force or Fraud, by the Pusillanimity of some, and the Corruption of all … When any Potentate hath arrived to an exorbitant share of power, the Rest (ought) to league together and reduce him to his due Share of it. Otherwise there is but one Potentate, and the others are only a kind of Vassals.

The *Catechism* shows that balance of power policy was well understood at the time. It clearly describes its functions: to prevent hegemony includes maintaining the 'Repose' of the participating powers, implying that they were no longer interested in expansionist adventures and had become status quo powers.

Balance of power policy requires that the Great Powers are able and willing to shift their alliances to deter hegemonic aspirations. That goal can be supported by a balancer. Britain, as a global sea power, could fulfill this role by always aligning itself against the leading power on the continent.

In the long run, however, balance of power policy can only be successful if the power balance between the Great Powers does not change too drastically. It still allows for cooperative expansion, as the three divisions of Poland among Russia, Austria, and Prussia in the late eighteenth century make clear. Conquest of Poland by one of them would have disturbed the balance of power.

The Congress of Vienna and the Concert of Europe

Balance of power policy did not always deter all Great Powers. Revolutionary France no longer respected the dynastic order in Europe. First, the Jacobin regime and then Napoleon set out on conquest, supported by the new power of popular conscript and volunteer armies, motivated by nationalism. Then balance of power policy had to change into a war coalition to undo the conquests of the hegemonist and restore the balance. As France occupied most of the continent, this took time. The decisive turn in the war came when Napoleon overextended his power and failed to subdue Russia.

The main members of the war coalition (Russia, Britain, Prussia, and Austria) expressed their war aims as the restoration of a just balance of power and a redivision of their respective forces suitable to assure this equilibrium. The Congress of Vienna followed the rules of balance of power policy: France was readmitted as one of the Great Powers and as member of the conference. As part of the postwar power balance the newly created Kingdom of the Netherlands – being in between Britain, France, and Prussia – was strengthened by the inclusion of the Southern Netherlands in the Kingdom.

With 'A World Restored' (Kissinger, 1957) in Vienna, a relatively stable period began again. No great war occurred until 1914. Still, in the nineteenth century maintaining the balance went together with more intense competition, whether for military and industrial power resources or for colonial possessions. There were few wars between a limited number of Great Powers, such as the Crimean War. Yet, rulers recognized a common interest in preserving a European balance of power.

The Concert of Europe and Its Global Successors

Balance of power policy was extended at the Conference of Vienna by the balancer Britain. In 1815 its Foreign Secretary Castlereagh drafted Article VI of the Quadruple Alliance Treaty

(to which France later acceded), which provided for periodic meetings between the Great Powers: "for the consideration of the measures which shall be considered the most salutary for the repose and prosperity of nations and the peace of Europe" (Kissinger, 1957: p. 186).

This 'conference system' broke down in 1822 when Castlereagh's successor Canning refused to support France's proposal for joint armed intervention in Spain. But the idea of a Concert of Europe to prevent and solve conflicts did not disappear. In 1855, Austria called a conference to prevent a war between Russia and Turkey. But it failed and the Crimean War broke out. That was ended by the Conference of Paris (1856), at which the British Foreign Secretary Lord Clarendon tried to strengthen the role of conference diplomacy by proposing the rule that states would seek 'good offices' before going to war. Two other important conferences were held in Berlin (1878 and 1884–85), both at the initiative of Bismarck. In the first, he acted with success as 'honest broker' between Russia, Britain, Austria–Hungary, and the Ottoman Empire to settle conflicts over the Balkan and in the Near East, resulting from the decline of the Ottoman Empire. The most important result of the second was that the Great Powers agreed to give control over the Congo to King Leopold of Belgium. As with the Netherlands in 1815, balance of power policy required that the Congo would not be controlled by one of the Great Powers, no matter what that would mean for the Congo itself.

The idea of a Concert of Europe went further than balance of power policy. It could imply collective Great Power domination. But it could not prevent a second complete breakdown of the balance of power. The rivalry for colonial possessions, symbolized by the scramble for Africa, became closed toward the end of the century. The world had been divided. Latecomers, such as Germany and Italy could not be satisfied. At the same time, the power balance in Europe became more uneven. After three wars, (with successively Denmark, Austria, and France) Prussia made itself the nucleus of a powerful new German Empire. Its first Chancellor Bismarck still thought in terms the balance of power. But after he left in 1890, Emperor William II began to harbor hegemonic ambitions. The other Great Powers did not form an antihegemonic alliance against Germany's aspirations and growing capability. On the contrary, Austria–Hungary sided with Germany, whereas Britain, France, and Russia formed a Triple Alliance. Two inflexible alliances stood against each other. Czar Nicholas' attempt to revive the conference idea in the form of general peace conferences, held in The Hague in 1899 and 1907, was not successful. The outbreak of World War I in July 1914 demonstrated again that balance of power policy was precarious.

Still, the winning alliance of World War I depended on the balancing effect of the entry of the United States. But restoring a balance of power was not a stated purpose of the postwar conference of Versailles. President Wilson believed that balance of power policy was responsible for the continual scourge of war in Europe. He looked for a lasting basis for peaceful conduct of states in collective security and national self-determination. Collective security was based on all members of the League of Nations uniting against an aggressor. But it proved impotent even against Italy, itself not a Great Power. Wilson's critical attitude toward balance of power policy was representative of the idealist thinking about international politics that was influential after the horrors of World War I. After the even worse horrors of World War II its counterpart realism, stressing the importance of power and military preparedness, became dominant again.

In the Security Council of the UN, the idea of a Concert of Great Powers became institutionalized. American President Roosevelt, who had again balanced the hegemony of Germany and its allies, saw the role of Great Powers in the Security Council as 'international sheriffs.' But a permanent Conference embodied in an international organization could not solve the problem of Great Power competition either. Hegemonic rivalry started again after a few years between the United States and the Soviet Union, thereby paralyzing the Security Council. Their rivalry did not lead to another great war primarily because of the consequences of the shared danger of a nuclear war. After the peaceful end of the hegemonic rivalry and of the Soviet Union, consensus for a time reigned in the Security Council, making a mandate for the use of force against Iraq possible. But China abstained. Rivalry again sets the tone in the Security Council.

Development beyond balance of power policy has thus not come very far. It has not been replaced by international organization. The dynamic of rivalry between states and especially Great Powers has not fundamentally changed, even though their domestic development now requires more cooperation.

Balance of Power in Other Parts of the World

Power balances have been part of the development of state systems everywhere, so have wars as part of hegemonic struggles. Thinkers about interstate relations, such as Kautilya in India or the Legalists in the China of the Warring States were, like Machiavelli, advisors to princes, primarily interested in increasing the power of their own state rather than in a balance of power. That was also the case with dynastic states in Europe. But in Europe balance of power policy replaced reason of state as the dominant policy orientation. Why? The explanation may be found:

1. in the relatively even Great Power balance that developed at the end of the seventeenth century;
2. in the importance of 'Repose' for dynastic regimes;
3. in colonial expansion (and Russia's frontier in Asia) as a substitute for territorial expansion within Europe; and
4. in the increasing repugnance for war in established states, though in the nineteenth century it could be replaced by nationalism.

The European balance of power became global in the twentieth century. The bipolar balance between the two remaining Great Powers after 1945 was perceived by both as covering the world as a whole. The smallest power advantage obtained by the one had to be compensated by the other.

In the post-Cold War world regional balances became more important. In Southeast Asia, for example, the members of the Association of Southeast Asian Nations (ASEAN) had transformed their economic cooperation in a security alliance directed to what they saw as the hegemonic ambitions of Vietnam, demonstrated by Vietnam's intervention and occupation of Cambodia. Recently Vietnam joined the ASEAN

because of an unspoken common interest in balancing China's possible bid for hegemony in the region. American foreign policy has been constantly concerned with regional balances of power, in Europe, but also in East Asia and the Middle East.

Conclusion: A Permanent Feature of International Politics?

Since the Congress of Vienna all attempts to create an international order were based on an extension of balance of power policy. The exception was the League of Nations. But its impotence can be explained by the neglect of balance of power considerations and overestimation of the effectiveness of the all against one requirement of punishing an aggressor, especially if that were a Great Power.

After 1991, the United States was the only global Great Power left. It is not likely that it will remain so indefinitely. Potential challengers, such as China or the European Union or the reformed Russia, are already there, even though still relatively weak. Will balance of power policy become relevant again in the future? As long as a monopoly of violence at the international level does not come into being, balance of power policy may well remain relevant.

See also: First World War, The; International Law and Treaties; Military History; Second World War, The; Warfare in History.

Bibliography

van Benthem van den Bergh, G., 1992. The Nuclear Revolution and the End of the Cold War: Forced Restraint. Macmillan, Houndmills, Basingstoke, UK.

Carr, E.H., 1946. The Twenty Years' Crisis 1919–1939. Macmillan, London.

Chandra, C., 2008. War, Peace and Hegemony in a Globalized World: The Changing Balance of Power in the Twenty-First Century. Routledge, London.

Cooper, A.S., 2011. The Oil Kings: How the U.S., Iran, and Saudi Arabia Changed the Balance of Power in the Middle East. Simon & Schuster, New York.

Elias, N., 1982. The Civilising Process. In: State Formation and Civilization, vol. II. Basil Blackwell, Oxford.

Gulick, E.V., 1955. Europe's Classical Balance of Power: A Case History of One of the Great Concepts of European Statecraft. Cornell University Press, Ithaca, NY.

Ikenberry, G.J., 2002. America Unrivaled: The Future of the Balance of Power. Cornell University Press, Ithaca.

Kant, I., 1957. Perpetual Peace. Bobbs Merril, New York.

Kaufman, S.J., Little, R., Wohlforth, W.C., 2007. The Balance of Power in World History. Palgrave Macmillan, Basingstoke, England.

Kautilya, L.N., 1992. In: Rangarajan, L.N. (Ed.), The Arthasastra. Penguin Books, New Delhi, India.

Kissinger, H.A., 1957. A World Restored: Metternich, Castlereagh and the Problems of Peace. Houghton Mifflin, Boston, MA.

Little, R., 2007. The Balance of Power in International Relations: Metaphors, Myths, and Models. Cambridge University Press, New York.

Meinecke, F., 1929. Die Idee der Staatsraison in der neueren Geschichte. Oldenbourg, Munich, Germany.

Nicholson, H., 1947. The Congress of Vienna. Phoenix Publishing, Bern, Switzerland.

Odgaard, L., 2007. The Balance of Power in Asia-Pacific Security: US-China Policies on Regional Order. Routledge, New York, NY.

Schweller, R.L., 2008. Unanswered Threats: Political Constraints on the Balance of Power. Princeton University Press, Princeton, NJ.

Vincent, R.J., Wright, M. (Eds.). Special issue on the balance of power. Review of International Studies 15 (2).

Vinci, A., 2009. Armed Groups and the Balance of Power: The International Relations of Terrorists, Warlords and Insurgents. Routledge, London.

Watson, A., 1992. The Evolution of International Society: A Comparative Historical Analysis. Routledge, London.

Wedgwood, C.V., 1963. The Thirty Years War. Routledge, London.

Baltes, Paul B (1939–2006)

Ulman Lindenberger and Julia AM Delius, Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany
Ursula M Staudinger, Robert N. Butler Columbia Aging Center, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Developmental psychologist Paul B. Baltes (1939–2006) was born in Saarlouis, Germany. After studying in Germany he moved to the United States, where he continued the conceptual and methodological quest of his doctoral dissertation to transform developmental psychology into a science that seeks to describe, explain, and modify individual differences in behavioral development across the entire life span. His work was pivotal in establishing life span psychology as a comprehensive and distinct theoretical orientation within contemporary developmental psychology. Paul Baltes' career was characterized by internationality, interdisciplinarity, generativity, societal concern, and institutional prudence and sagacity. One of his central research interests was to explore age-graded changes in the range of plasticity in individual development, with the goal to identify and promote personal and societal conditions that favor successful aging. At the same time, he was not negligent of the negative sides of very old age, and spoke of old age and aging as 'hope with a mourning band.'

Life

Paul B. Baltes was born in 1939 in Saarlouis, Germany. Between 1959 and 1963, he studied psychology at the University of Saarland in Saarbrücken with Ernst E. Boesch, who was a student of Swiss psychologist Jean Piaget, and Günther Reinert, who would remain his academic mentor for many years. As Baltes himself noted, "Piagetian developmental cognitive structuralism and Boeschean cultural-psychological theory thus were two of the cornerstones in my study of psychology. The third, provided by the guidance of Günther Reinert, was psychometrics and the developmental study of intelligence" (Baltes, 2000: 11).

During that time he met Margret M. Labouvie who later became his wife and close collaborator. After obtaining his master's degree in psychology on the differentiation hypothesis of intelligence in childhood (Reinert et al., 1965), he and his newlywed wife Margret spent a year at the University of Nebraska. As a young graduate exchange student Paul Baltes collaborated with the eminent developmental psychologist and gerontologist K. Warner Schaie. Back in Saarbrücken, he continued his work on the basic methodological problem of disentangling age from cohort and period effects when studying development, and completed his dissertation on this topic in 1967 (Baltes, 1968; now a 'citation classic'). He then accepted K. Warner Schaie's offer of an assistant professorship at the University of West Virginia.

The first West Virginia conference on life span development (Goulet and Baltes, 1970) marks the emergence of life span psychology as a meta-theoretical framework within modern psychology (cf. Tetens, 1777; Thomae, 1959). It formed the basis of what later became the classic series, "Lifespan Development and Behavior," edited by Paul Baltes in collaboration with social psychologist Orville Gilbert Brim Jr., and later with sociologist David Featherman and developmental psychologist Richard M. Lerner. In this series, a continually growing group of scholars radically criticized the dominant interpretation of developmental psychology as a science primarily oriented toward childhood and biology-based growth processes. According to this critique, development includes the whole life span from conception into old age (Baltes and Goulet, 1970) and comprises losses as well as gains, maturation as well as senescence, and biological as well as cultural influences. Paul Baltes was one of the pioneers of this theoretical orientation in modern developmental psychology, as first described in the article on the life span perspective on human development that was published in 1980 in the *Annual Review of Psychology* (Baltes et al., 1980).

At the age of 33 years Paul Baltes was appointed head of the Division of Individual and Family Studies at the College of Human Development of Pennsylvania State University, where he set up an interdisciplinary graduate program on research questions of lifelong development – together with colleagues such as John R. Nesselroade and Richard M. Lerner. In 1980, the Max Planck Society recruited Paul Baltes back to Germany as Director at the Max Planck Institute for Human Development in Berlin. A quarter of a century later, when he reached emeritus status, life span thinking had become integrated into developmental psychology in both North America and Germany. In the late 1990s he took up the challenge of coediting the first edition of the *International Encyclopedia of the Social and Behavioral Sciences* together with Neil J. Smelser. Its forerunner was the *International Encyclopedia of Social Sciences* edited by David Sills in 1968. As coeditor-in-chief, Paul Baltes ensured that fields such as behavioral and cognitive neuroscience, genetics, evolutionary sciences, law, psychiatry, and, of course, developmental psychology were represented in the 26-volume work that was published in 2001 (Smelser and Baltes, 2001).

In 1999, Margret Baltes, his longtime collaborator, companion, and mother of their two children, died suddenly. Christine Windbichler, Professor of Law at Humboldt-Universität zu Berlin, later became his second wife.

Taking on the title of Director Emeritus in 2004 was only a transition to new challenges for Paul Baltes. He was appointed Distinguished Professor at the University of Virginia and Director of the Max Planck International Research Network on Aging (MaxnetAging), the first virtual Max Planck Institute that was geared toward the multidisciplinary investigation of

human aging. Furthermore, until the very end of his life, he was the Chair of the International Max Planck Research School on the Life Course (LIFE) and a board member of the working group 'Opportunities and Problems of an Aging Society: The World of Working and Lifelong Learning' launched by the German National Academy of Sciences Leopoldina and the German National Academy of Science and Engineering (acatech). Sadly, his own life was cut short by cancer and he died in Berlin aged 67 in 2006.

Honors

Paul Baltes left an oeuvre of more than 300 publications including 27 books, and the 26 volumes of the *International Encyclopedia of the Social and Behavioral Sciences*. His scientific recognition is also reflected in the honorary degrees that he received as well as countless memberships in national and international academies. For his scientific work, he received 12 prestigious prizes and awards from national and international professional societies. He was a foreign member of the American Academy of Arts and Sciences and the Royal Swedish Academy of Sciences. Honorary doctorates were bestowed on Paul Baltes by the Universities of Jvyäskylä, Stockholm, Geneva, and Humboldt-Universität zu Berlin. His international awards included the American Psychological Association's Award for Distinguished Contributions to the International Advancement of Psychology, the Robert W. Kleemeier Award in Recognition of Outstanding Research from the Gerontological Society of America, and the Novartis Prize for Gerontological Research awarded by the International Association of Gerontology. Paul Baltes also received highest recognition beyond the scientific community. He was the only psychologist to be awarded membership in the German Order 'Pour le Mérite' for the Arts and Sciences, and received the Federal Republic of Germany's Great Cross of Merit with Star.

Profile as Scientist and Scholar

Beyond scientific productivity and professional as well as public honors, Paul Baltes as a person and his role in science and society may be best characterized by five qualities: internationality, interdisciplinarity, generativity, societal concern, and institutional prudence and sagacity (see also Staudinger, 2007).

Internationality

Throughout his career, Paul Baltes built bridges and developed international networks. He firmly believed that international exchange increases the quality of scientific work and promotes scientific insights. Only very few German scientists have reached equally high recognition in the USA. Good examples of such networks are the network of life span researchers as captured in the Lifespan series, MaxnetAging, and the international doctoral program LIFE, which continues to flourish to the present day.

Interdisciplinarity

As spiritus rector of life span psychology, Paul Baltes was committed to the interdisciplinary study of human development. During his time as department head at Penn State, he shepherded an interdisciplinary faculty composed of psychologists and sociologists. His work for the Social Science Research Council and the European Research Foundation also provided rich experiences in interdisciplinary exchange. He initiated the Berlin Aging Study (BASE), connecting biomedical and psychological with sociological and economic perspectives on aging. BASE also became a success because Paul Baltes managed to lead the various disciplinary directions to an equal dialogue (Baltes and Mayer, 1999). During his later years, he took great pleasure and pride in developing discourses that transcended disciplinary horizons in the context of the academies of which he was a member, particularly in the Order 'Pour le Mérite' for the Arts and Sciences. He linked his striving for the Renaissance ideal of a comprehensively informed scholar with a wakeful eye for the necessities and opportunities of modern societies.

Generativity

Early on and throughout his career, Paul Baltes's work was characterized by generativity. According to Erik Erikson, caring for the following generations, passing on insights and experiences, is the central developmental challenge of the middle and later years. Paul Baltes took on this challenge much ahead of time. He mentored his doctoral students and postdoctoral fellows with a great deal of personal investment in many areas of life. In the 1980s, soon after his return to Germany, Paul Baltes and other developmental psychologists launched the first graduate fellowship program in Germany funded by the Volkswagen Foundation, which later became a model for graduate programs in Germany funded by the German Research Foundation (DFG). Among the cherished projects of his later years were the foundation of the International Max Planck Research School LIFE (see http://www.imprs-life.mpg.de) and the establishment of an academy of sciences for young academics. The latter project, the 'Junge Akademie,' was implemented as a collaboration project between the Berlin–Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina (of which Paul Baltes was vice-president). The 'Junge Akademie' has since become a role model and example for similar initiatives in many other countries.

Societal Interest

Paul Baltes displayed concern for societal issues. He firmly believed that psychology has something to say and even more strongly, that it can contribute to the solution of challenges that modern societies are facing. Going public as a scientist is an art that requires great care and investment. Paul Baltes excelled in this regard as well. He advanced and influenced societal discussions about lifelong learning, demographic aging, the chances and hardships of old and very old age, and the question of self-determined death.

Institutional Initiative and Prudence

Science takes place within institutions and is influenced by institutional characteristics and their logic. Paul Baltes thought and acted in the interest of the institution and thereby contributed to the flourishing of the Max Planck Institute for Human Development in Berlin. Over the years,

he also took on many responsibilities to promote the standing of the behavioral and social sciences within the Max Planck Society, and took a leading role in numerous networks promoting the study of adult development and aging from a life span perspective.

Contributions to the Behavioral and Social Sciences

Paul Baltes' doctoral dissertation on cohort effects marked the starting point for a comprehensive methodological and conceptual oeuvre (e.g., Baltes, 1968; see also Schaie, 1965). Paul Baltes formulated the scientific tenets and rationale of life span psychology (Baltes et al., 1977; Baltes, 1987; Baltes et al., 2006a). His conceptual and methodological writings have become milestones in the scientific study of behavioral development.

Quite early, Paul Baltes developed interest in the malleability of cognitive functioning in adulthood and old age. His work on cognitive training demonstrated the continued existence of learning potential in adulthood, and foreshadowed much of the recent interest in this topic. Together with Reinhold Kliegl, he introduced the testing-the-limits paradigm to delineate age differences in memory performance after instruction and training in mnemonic skill (Baltes and Kliegl, 1992). Later his interest in cognitive plasticity expanded to include personality functioning and to link the notion of plasticity with traditional conceptions of resilience across the life span (Staudinger et al., 1995). Together with Ursula Staudinger and Jacqui Smith, he succeeded in establishing a measurement model of wisdom, defined as high levels of expertise about the fundamental pragmatics of life, and identified the constellation of personal attributes and social opportunities that are germane to its emergence (Baltes and Staudinger, 2000). Together with Ellen Skinner, he refined the study of control beliefs in childhood by introducing the tripartite conceptualization of control, agency, and means–ends beliefs (Skinner et al., 1988), in line with his emphasis on goal-directed action as a central concept in developmental theory.

In the late 1980s, Paul Baltes initiated and launched the Berlin Aging Study (Baltes and Mayer, 1999), a multidisciplinary, longitudinal study on old age and aging. The coprincipal investigators were Hanfried Helmchen, a psychiatrist, Karl Ulrich Mayer, a life-course sociologist and codirector at the MPI for Human Development, and Elisabeth Steinhagen-Thiessen, a geriatrician. Among the key features of the Berlin Aging Study are a truly multidisciplinary approach, expressing Paul Baltes's conviction that progress in science requires a dialogue among equals in which no discipline dominates the other; a focus on very old age (Gerstorf et al., 2013); and a strong interest in methodological issues such as sample selectivity (Lindenberger et al., 2002).

Besides the continued potential for growth, Paul Baltes was also interested in the reasons for decline in adulthood and old age. He showed that hearing, seeing, and walking require increasing cognitive resources with advancing adult age (Lindenberger and Baltes, 1994). Later, he advocated linking research on behavioral development to research on neural mechanisms. In this context, he promoted the notion of 'bio-cultural co-constructivism' and warned against regarding behavior as the dependent variable of the brain (Baltes et al., 2006b).

Together with his first wife Margret, Paul Baltes was interested in the orchestration of psychological mechanisms promoting successful human development. Together they introduced the selection, optimization, compensation (SOC) model of developmental regulation, which posits the three mechanisms as regulatory mechanisms of human development (Baltes and Baltes, 1990; Freund and Baltes, 1998). The SOC model has been integrated into systems theory (Boker, 2013), and successfully applied to the study of specific developmental tasks, such as the resolution of work–family conflict (e.g., B. Baltes and Heydens-Gahir, 2003).

Paul Baltes has left a rich legacy. His work continues to inspire research in developmental psychology and gerontology. After his first wife's death in 1999, Paul Baltes established the Margret M. Baltes Foundation to promote their common goal of advancing life span developmental and gerontological research. Upon his own death in 2006, it was renamed the Margret M. and Paul B. Baltes Foundation. According to its website, "The Margret M. and Paul B. Baltes Foundation aims to advance research in developmental psychology and gerontology and to underline its relevance in science and society. It encourages researchers to dedicate projects and careers to these fields and to widen the empirical and theoretical understanding of humans and society in adulthood and old age. It aims to highlight societal initiatives that point to the relevance of such research."

Paul Baltes was a master at forging links – in the academic, science management, societal, and personal world. Therefore, his activities, his knowledge, and worldwide recognition were not restricted to his home discipline, developmental psychology; they drew wide circles into the humanities and natural sciences. He put questions into the open and made them the starting point of a new research idea, a new institution, or a new collaboration. Paul Baltes was able to establish developmental psychology as an equal partner among disciplines. In accordance with his theory of the permanent incompleteness of human nature (Baltes, 1997), his scientific work was not finished and will never be finished.

See also: Adulthood: Dependency and Autonomy; Aging, Theories of; Cognitive Aging; Cognitive Psychology: History; Erikson, Erik Homburger (1902–94); Lifespan Development, Theory of; Piaget, Jean (1896–1980).

Bibliography

Baltes, B.B., Heydens-Gahir, H.A., 2003. Reduction of work–family conflict through the use of selection, optimization, and compensation behaviors. Journal of Applied Psychology 88, 1005–1018.

Baltes, P.B., 1968. Longitudinal and cross-sectional sequences in the study of age and generation effects. Human Development 11, 145–171.

Baltes, P.B., 1987. Theoretical propositions of life-span developmental psychology: on the dynamics between growth and decline. Developmental Psychology 23, 611–626.

Baltes, P.B., 1997. On the incomplete architecture of human ontogeny: selection, optimization, and compensation as foundation of developmental theory. American Psychologist 52, 366–380.

Baltes, P.B., 2000. Autobiographical reflections: from developmental methodology and lifespan psychology to gerontology. In: Birren, J.E., Schroots, J.J.F. (Eds.), A History of Geropsychology in Autobiography. APA Books, Washington, DC, pp. 7–26.

Baltes, P.B., Baltes, M.M., 1990. Psychological perspectives on successful aging: the model of selective optimization with compensation. In: Baltes, P.B., Baltes, M.M. (Eds.), Successful Aging: Perspectives from the Behavioral Sciences. Cambridge University Press, Cambridge, pp. 1–34.

Baltes, P.B., Goulet, L.R., 1970. Status and issues of a life-span developmental psychology. In: Goulet, L.R., Baltes, P.B. (Eds.), Life-Span Developmental Psychology: Research and Theory. Academic Press, New York, pp. 4–21.

Baltes, P.B., Kliegl, R., 1992. Further testing of limits of cognitive plasticity: negative age differences in a mnemonic skill are robust. Developmental Psychology 28, 121–125.

Baltes, P.B., Lindenberger, U., Staudinger, U.M., 2006a. Life span theory in developmental psychology. In: Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology, Theoretical Models of Human Development, sixth ed., vol. 1. Wiley, New York, pp. 569–664.

Baltes, P.B., Mayer, K.U. (Eds.), 1999. The Berlin Aging Study: Aging from 70 to 100. Cambridge University Press, New York.

Baltes, P.B., Reese, H.W., Nesselroade, J.R., 1977. Life-Span Developmental Psychology: Introduction to Research Methods. Brooks/Cole, Monterey, CA.

Baltes, P.B., Reese, H.W., Lipsitt, L.P., 1980. Life-span developmental psychology. Annual Review of Psychology 31, 65–110.

Baltes, P.B., Reuter-Lorenz, P.A., Rösler, F. (Eds.), 2006b. Lifespan Development and the Brain: The Perspective of Biocultural Co-constructivism. Cambridge University Press, New York.

Baltes, P.B., Staudinger, U.M., 2000. Wisdom: a metaheuristic (pragmatic) to orchestrate mind and virtue toward excellence. American Psychologist 55, 122–136.

Boker, S.M., 2013. Selection, optimization, compensation, and equilibrium dynamics. GeroPsych 26, 61–73.

Freund, A.M., Baltes, P.B., 1998. Selection, optimization, and compensation as strategies of life management: correlations with subjective indicators of successful aging. Psychology and Aging 13, 531–543.

Gerstorf, D., Ram, N., Lindenberger, U., Smith, J., 2013. Age and time-to-death trajectories of change in indicators of cognitive, sensory, physical, health, social, and self-related functions. Developmental Psychology 49, 1805–1821.

Goulet, L.R., Baltes, P.B. (Eds.), 1970. Life-Span Developmental Psychology: Research and Theory. Academic Press, New York.

Lindenberger, U., Baltes, P.B., 1994. Sensory functioning and intelligence in old age: a strong connection. Psychology and Aging 9, 339–355.

Lindenberger, U., Singer, T., Baltes, P.B., 2002. Longitudinal selectivity in aging populations: separating mortality-associated versus experimental components in the Berlin Aging Study (BASE). Journals of Gerontology Series B, Psychological 57B, P474–P482.

Reinert, G., Baltes, P.B., Schmidt, L.R., 1965. Faktorenanalytische Untersuchungen zur Differenzierungshypothese der Intelligenz. Psychologische Forschung 28, 246–300 [Factor analytic investigation of the differentiation hypothesis of intelligence].

Schaie, K.W., 1965. A general model for the study of developmental problems. Psychological Bulletin 64, 92–107.

Skinner, E.A., Chapman, M., Baltes, P.B., 1988. Children's belief about control, means-ends, and agency: developmental differences during middle childhood. International Journal of Behavioral Development 11, 369–388.

Smelser, N.J., Baltes, P.B. (Eds.-in-Chief), 2001. International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier Science, Oxford.

Staudinger, U.M., 2007. Paul B. Baltes – in memoriam. European Psychologist 12, 72–74.

Staudinger, U.M., Marsiske, M., Baltes, P.B., 1995. Resilience and reserve capacity in later adulthood: potentials and limits of development across the life span. In: Cicchetti, D., Cohen, D. (Eds.), Developmental Psychopathology, Risk, Disorder, and Adaptation, vol. 2. Wiley, New York, pp. 801–847.

Tetens, J.N., 1777. Philosophische Versuche über die menschliche Natur und ihre Entwicklung (Philosophical investigations about the human soul and its development). Weismanns Erben und Reich, Leipzig. Reprinted from: Kantgesellschaft, vol. 1. Reuther & Reichard, Berlin (1913).

Thomae, H., 1959. Entwicklungspsychologie (Developmental psychology). Hogrefe, Göttingen.

Bankruptcy

Rafael I Pardo, Emory University School of Law, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by W.C. Whitford, volume 2, pp. 1042–1044, © 2001, Elsevier Ltd.

Abstract

Bankruptcy is a legal process that generally seeks to effectuate two goals: (1) an orderly distribution of a debtor's pre-bankruptcy assets and/or future income for the benefit of creditors and (2) financial rehabilitation of the debtor. The first goal can be viewed as the law's response to the common-pool problem that arises when a debtor has insufficient resources to repay creditor claims. The second goal can be viewed as the law's response to alleviating the financial distress that renders a debtor economically unproductive.

Bankruptcy is a legal process that generally seeks to effectuate two goals: (1) an orderly distribution of a debtor's pre-bankruptcy assets and/or future income for the benefit of creditors and (2) financial rehabilitation of the debtor. The first goal can be viewed as the law's response to the common-pool problem that arises when a debtor has insufficient resources to repay creditor claims. The second goal can be viewed as the law's response to alleviating the financial distress that renders a debtor economically unproductive – that is, when the debt burden becomes so great that it interferes with the debtor's ability to make future financial commitments. These two goals provide a metric against which the content of bankruptcy law can be evaluated for its soundness.

Bankruptcy's Equality Principle

When a debtor ceases to repay his or her debts voluntarily, creditors must seek through other means to collect the debts owed to them by the debtor. Some creditors might have had an opportunity prior to the debtor's default to contractually create an interest in property of the debtor to secure repayment of the debt (i.e., a consensual lien). Other creditors might not have had such an opportunity. Instead, they would have to rely on the creation of an involuntary lien – whether arising solely by operation of law under certain circumstances or upon the occurrence of certain conditions (i.e., a statutory lien), or whether obtained through judicial process (i.e., a judicial lien) – on the debtor's property to secure repayment of the debt. The property to which the lien, whether consensual or nonconsensual, has attached will serve as a source of repayment. The creditor will apply the proceeds obtained through liquidation of such property to reduce the outstanding amount of the obligation owed to it by the debtor.

When a debtor has insufficient resources (i.e., assets and income) to repay his or her creditors in full, a common-pool problem arises. This problem is exacerbated by the distributive principle that generally governs nonbankruptcy collection remedies – that is, "first in time, first in right". Put another way, the creditor who is first to take the necessary steps to establish legal entitlement to satisfaction of the debt from a particular asset of the debtor will have priority over subsequent creditors who make a claim to the same asset. This legal rule prompts a race of diligence among creditors to dismantle the failing debtor. Each creditor will seek to establish legal entitlement to the debtor's assets in order to assure itself of repayment to the fullest extent possible. Because there are insufficient resources to pay all creditors in full, there will be winners (e.g., those who are fully paid) and losers (e.g., those who do not receive any payment). Moreover, the competition for resources may result in a disorderly, inefficient allocation that does not maximize their value, thereby harming the group as a whole. In other words, the individualistic nature of competition for a debtor's resources will likely prompt creditors to pursue their non-bankruptcy remedies against the debtor in a way that promotes the creditor's self-interest to the detriment of the creditor group.

Bankruptcy seeks to overcome the common-pool problem by effectuating a shift to a compulsory collective proceeding pursuant to which the debtor's assets and/or future income will be administered for the benefit of all creditors (Jackson, 1986: pp. 7–19). The initiation of a bankruptcy proceeding generally creates an estate that includes the debtor's prebankruptcy property and simultaneously enjoins the pursuit of individual remedies by creditors against the debtor and its property. The grant of jurisdiction over the bankruptcy case and the debtor's property enables the court or agency that administers the process to conduct a uniform proceeding in a single forum, sorting out creditors' claims against the debtor and its property.

Instead of providing the same treatment for all creditors, bankruptcy focuses on organizing creditors into various classes based on their nonbankruptcy entitlements. Once those classes have been established, bankruptcy law concerns itself with providing equal treatment *within*, rather than *across*, creditor classes. Because of this, the law allows for certain creditor classes (e.g., secured creditors) to receive more favorable treatment than others (e.g., unsecured creditors). Thus, the process of marshaling and distributing a debtor's resources in bankruptcy is oriented pursuant to the principle of equal treatment of *similarly situated* creditors (the "equality principle").

Bankruptcy's Fresh-Start Principle

To ameliorate a debtor's financial distress, bankruptcy law's primary substantive relief is the discharge, which generally

releases the debtor from personal liability on prebankruptcy debts. Such relief provides the debtor with a fresh start and represents societal consensus to forgive a debtor's financial obligations. Two animating issues underlie such substantive relief: (1) the type of debtor that should be eligible for that relief and (2) the scope of such relief. The former issue turns on normative judgments about when a debtor may seek respite from financial failure under the protective cover of bankruptcy law, whereas the latter issue addresses the extent of relief to be conferred upon debtors who are deemed eligible for it.

Eligibility rules for bankruptcy may focus on a variety of characteristics, such as the organizational structure of the entity seeking bankruptcy relief (e.g., in the United States, domestic insurance companies are not eligible to file for bankruptcy) or the debtor's motivation for seeking bankruptcy relief (e.g., eligibility may be denied based on the debtor's bad faith in seeking relief). On the other hand, defining the scope of bankruptcy relief necessarily entails judgments about the debtor's fresh start. Limits on such relief are generally imposed in one of two forms: (1) the price of discharge and (2) the amount of debt discharged. With respect to the former, the law generally requires a debtor to relinquish either all of its prebankruptcy assets and/or a portion of its future income. With respect to the latter, the discharge may not extend to cover all prebankruptcy debts (e.g., certain tax debts). Accordingly, one might conceive of the benefit of discharge as the extent to which the amount of debt forgiven exceeds the price paid for the discharge.

The Content of Bankruptcy Law

In analyzing bankruptcy law, one can generally point to the equality and fresh-start principles as influencing its content. Put another way, bankruptcy law might be viewed as an operative framework for effectuating these principles. While some of the legal rules may apply across all classes of bankruptcy cases – that is, cases involving both natural persons and legal entities, such as corporations – others may apply only within certain classes.

Implementing the Equality Principle

As previously mentioned, the fresh-start principle generally requires the debtor to pay a price for the discharge of debt – namely, relinquishment of all prebankruptcy assets and/or a portion of the debtor's future income. The price exacted from the debtor will dictate the amount repaid to creditors. Because debtors generally face a choice of seeking relief either through liquidation (i.e., converting their prebankruptcy assets into proceeds for distribution to creditors) or reorganization/rehabilitation (i.e., relying on future income to repay creditors, but nonetheless preserving the option of using prebankruptcy assets for repayment), the law must devise a baseline for the repayment of creditor claims.

The repayment baseline generally focuses on the amount that creditors would receive on account of their claims if the debtor's prebankruptcy assets were liquidated and distributed to creditors. Establishing this baseline requires legal rules specifying: (1) the property available for liquidation, (2) the amount of creditor claims, and (3) the order and manner in which such claims will receive payment. Importantly, as the scope of the property available for liquidation (i.e., the price of discharge) broadens, the net benefit of discharge theoretically decreases. In other words, the more property a debtor must relinquish in exchange for the discharge, the less robust the discharge becomes. Accordingly, the fresh-start principle is very much implicated in defining the scope of property available for liquidation.

Once the law has identified the property that should be available for liquidation, the focus shifts to identifying the types of claims that can be made by creditors against such property. This process necessarily entails defining what constitutes a claim and then establishing a framework for determining the allowance or disallowance of such claims against the debtor's estate. Bankruptcy law generally strives to provide a comprehensive, all-encompassing definition of a claim. For example, the US Bankruptcy Code broadly defines a claim as a "right to payment, whether or not such right is reduced to judgment, liquidated, unliquidated, fixed, contingent, matured, unmatured, disputed, undisputed, legal, equitable, secured or unsecured" (U.S. Code, 2012: Vol. 11, sec. 101(5)(A)) and imposes the temporal limitation that the claim should have existed as of or prior to the commencement of the debtor's bankruptcy case. Such an approach has the effect of consolidating all of the debtor's prebankruptcy liabilities for resolution *en masse* in the bankruptcy case – an instrumental step for effectuating the shift from individualistic creditor remedies outside of bankruptcy to a collective proceeding seeking to redistribute the debtor's limited resources for the benefit of all creditors.

Importantly, bankruptcy law generally defers to nonbankruptcy law for purposes of determining which substantive entitlements constitute a right to payment giving rise to a claim. Were bankruptcy law to do otherwise, the possibility arises that some creditors would receive more favorable treatment in bankruptcy rather than outside of bankruptcy – that is, in those instances where bankruptcy law would provide a more robust substantive entitlement than nonbankruptcy law. This, in turn, would prompt forum shopping by creditors who would receive preferred status within bankruptcy, thereby creating an opportunity for self-interested behavior (e.g., the preferred creditors attempting to push the insolvent debtor into an involuntary bankruptcy proceeding) that undermines the collective interests of the group of creditors as a whole. In other words, bankruptcy law would perpetuate the problem that it attempts to solve (Jackson, 1986: pp. 25–27).

The claim allowance process determines who may receive payment from the debtor's estate. Establishing the validity and the amount of the claim is at the heart of this process. In the context of the common-pool problem, reductions in the number and amount of claims are likely to produce a greater return to the remaining claimants in the proceeding. The possibility of claim disallowance thus provides an opportunity for creditors to knock out competing claimants from the competition for the debtor's limited resources. Of course, in many instances, this outcome may only have marginal benefit because of the *pro rata* approach to distributing the debtor's resources in bankruptcy (as discussed below) and because of the costs incurred in establishing the grounds justifying the

disallowance of a claim. Given these dynamics, pragmatism dictates an overinclusive approach toward the allowance of claims. In the United States, for example, the law presumes the validity and amount of claims and narrows the set of circumstances under which a claim will be disallowed. By adopting an allowance approach that complements the broad definition of a claim for bankruptcy purposes, bankruptcy law evinces hostility toward claim litigation and instead aims to facilitate an expeditious and efficient resolution of creditor claims (Pardo and Lacey, 2009: pp. 186–187).

Once the identity of claimants entitled to receive distributions from the debtors' estate has been established, bankruptcy law shifts its attention to the order of such distributions. For a creditor fortunate enough to have created a prebankruptcy lien in property of the debtor to secure repayment of the debt owed to it, such a creditor will generally be deemed to have a secured claim within the bankruptcy proceeding. The primary advantage of that status is that it generally entitles the creditor to receive the value of the property securing its claim. Provided that the lien is given effect within bankruptcy (although the possibility exists that it will not, for reasons discussed below), the creditor will be assured of receiving some distribution in bankruptcy, thus emphasizing the manner in which the lien serves as form of insurance against the risk of nonpayment by the debtor (both outside and inside of the bankruptcy forum). On the other hand, bankruptcy law will generally relegate unsecured creditors (i.e., those without a lien in estate property) to sharing ratably with one another the assets that remain in the common pool after the estate's encumbered assets have been administered. Importantly, not all unsecured claimants are treated alike. Instead, bankruptcy law often confers preferred status upon some unsecured claimants, such as government creditors with tax claims or the claims of the debtor's employees. In these sorts of provisions, one witnesses the public policy rationales at work that certain types of claimants deserve better treatment in bankruptcy than other competing claimants. For example, in the context of government creditors with tax claims, the idea is that bankruptcy law ought to be tailored in such a way that it protects the public fiscus (Morgan, 2000: p. 463), even though doing so may undermine the welfare of general contract creditors.

As the discussion above illustrates, bankruptcy law does not treat all creditors alike, but rather assigns various statuses based upon the nature of the claim. This differential treatment features prominently in the distribution scheme in bankruptcy. The order in which claimants receive payment on account of their claims reflects society's judgment as to the relative merit of claimants. For example, some bankruptcy regimes entitle secured creditors to receive distributions first, whereas other regimes entitle tax claimants to receive distributions prior to certain secured creditors (Morgan, 2000: pp. 474–499). Regardless of the order, creditors will be segregated into various classes based upon the status that bankruptcy law assigns to them. Once the separation has occurred, distributions will be made to the class with the highest priority, followed by distributions to the class with the second-highest priority, and so on and so forth. When insufficient funds remain to pay all members of a class in full, the members will receive a *pro rata* distribution. For those creditors that have been assigned to classes with a lower priority than the class that received a *pro rata* distribution, such creditors will not receive any distribution on account of their claims.

In liquidation proceedings, the debtor's assets will be reduced to money, which in turn will be used to repay creditor claims pursuant to the prescribed distribution scheme. In reorganization proceedings (for legal entities, such as a corporation) and rehabilitation proceedings (for natural persons), the debtor's future income will be used primarily (if not exclusively) to repay creditor claims. In the abstract, creditors as a group should theoretically receive a greater payout when the debtor's income is used as the source of repayment. If the debtor is a legal entity and has a reasonable prospect of being restored to financial health, the income generated by the debtor as a going concern ought to be greater than the amount that would be realized by liquidating the debtor's assets. And, if the debtor is a natural person, the disposable income that he or she generates over time will likely exceed the liquidation value of those assets that the debtor cannot exempt from the bankruptcy process and that are thus available as a source of repayment in a liquidation proceeding. Theory aside, bankruptcy law aims to ensure that creditors involved in a reorganization or rehabilitation proceeding are no worse off, from the repayment perspective, than had the debtor sought relief through a liquidation proceeding. In other words, when future income is the source of repayment, creditors should receive at least as much as they would have received from the proceeds of the debtor's liquidated assets. The latter amount serves as a baseline for confirming that creditors receive equal (if not better) treatment in reorganization and rehabilitation proceedings. Because claim repayment based on future income will involve a deferred stream of future payments, the aggregate amount projected to be paid to creditors must be discounted to its present value (thereby accounting for the time-value of money) when making the comparison to the liquidation baseline.

Given bankruptcy law's preoccupation with vindicating the equality principle, a concern arises when a debtor transfers property or an interest therein (e.g., a consensual lien) to a creditor on the eve of bankruptcy. Specifically, the transfer may render the creditor better off than similarly situated creditors (i.e., those with like legal entitlements), thereby upending what would have been the ordered arrangement for distribution in bankruptcy in the absence of the transfer. Additionally, the debtor may have made the transfer to put the property beyond the reach of its creditors, thereby depleting the estate of that which would have existed in the absence of the transfer. Bankruptcy law responds to these concerns through various avoidance provisions that restore the debtor's estate and the transferee to their prebankruptcy positions by allowing certain prebankruptcy transfers (e.g., preferential transfers and fraudulent transfers) to be recovered for the benefit of all creditors. Such provisions can be viewed as "stabilizing forces" intended to preserve the "scheme of social cooperation" (Rawls, 1971: p. 6) – that is, *ex post* mechanisms for unraveling *ex ante* behavior that would undermine the equality principle.

Implementing the Fresh-Start Principle

When thinking about the manner in which bankruptcy law seeks to implement the fresh-start principle, it is important to keep in mind that the amount of debt that bankruptcy law

forgives is completely unrelated to the distributional treatment that creditors receive on account of the status of their claims (Jackson, 1986: p. 225). Accordingly, rules governing the scope of discharge are distinct from and can coexist peacefully (and not in tension) with distributive rules governing the repayment of creditor claims *within* bankruptcy. On the other hand, implementing other aspects of the fresh-start principle (i.e., provisions unrelated to the discharge of debt) may have the effect of reducing creditor recoveries. For example, any provision that permits the debtor to claim certain prebankruptcy property as exempt (and thus unavailable for distribution) will reduce the amount that creditors will be repaid. Moreover, a discharge of prebankruptcy debt generally precludes a creditor from seeking postbankruptcy repayment from the debtor on account of such debt. One can see from these examples that certain aspects of the fresh-start principle and the equality principle are inversely related such that enlarging the reach of one necessitates narrowing the other's scope (Pardo and Lacey, 2005: p. 415).

Bankruptcy law's core mechanism for implementing the fresh-start principle is the discharge. In its most robust form, the bankruptcy discharge would forgive all prebankruptcy debts irrespective of the identity of the claimants to whom such debts are owed or the circumstances under which such debts were incurred. The law, however, does not afford such generous treatment to debtors and instead curtails the scope of relief to exclude certain debts from being forgiven (e.g., tax debts, fraudulent debts), imposing such restrictions predominantly on natural persons rather than legal entities. Such provisions reflect societal consensus that the fresh start should not be limitless: In other word, society's interest in the repayment of certain types of debts outweighs its interest in forgiving debtors. Discharge exceptions inevitably encroach upon a debtor's fresh start, threatening the ability of the debtor to return to economic productivity. Nonetheless, by excepting certain debts from discharge, bankruptcy law reflects the inherent belief that a debtor will be able to repay its prebankruptcy creditors from postbankruptcy income and/or assets (Boshkoff, 1982: p. 89).

Upon the grant of discharge, bankruptcy law seeks to protect the debtor's fresh start by enjoining efforts by creditors to collect discharged debts as a personal liability of the debtor. Unhampered by the burden of past debts, the debtor can focus his or her attention on reintegrating into the marketplace with the opportunity to make new financial commitments.

Conclusion

Since time immemorial, financial failure has been part and parcel of the debtor–creditor relationship. Inevitably, there will always be those who cannot fully repay their debts. Society must, broadly speaking, determine three things when choosing to address such failure through bankruptcy law: (1) whether to absolve the debtor of past debts; (2) the price that the debtor must pay in exchange for such absolution, and (3) how to allocate the price paid among existing creditors. These determinations necessarily entail intuitions and normative judgments regarding the causes of financial failure, the future ability of debtors to repay their debts, and the worthiness of competing claimants in light of their relative entitlements. Accordingly, when examining the content of bankruptcy law, a rich story unfolds about the manner in which society values remedying financial distress.

See also: Business Law; Ethical Codes, Professional: Business Codes.

Bibliography

Boshkoff, D.G., 1982. Limited, conditional, and suspended discharges in Anglo-American bankruptcy proceedings. University of Pennsylvania Law Review 131, 69–125.

Jackson, T.H., 1986. The Logic and Limits of Bankruptcy Law. Beard Books, Washington, DC.

Morgan, B.K., 2000. Should the sovereign be paid first? A comparative international analysis of the priority for tax claims in bankruptcy. American Bankruptcy Law Journal 74, 461–507.

Pardo, R.I., Lacey, M.R., 2005. Undue hardship in the bankruptcy courts: an empirical assessment of the discharge of educational debt. University of Cincinnati Law Review 74, 405–530.

Pardo, R.I., Lacey, M.R., 2009. The real student loan scandal: undue hardship discharge litigation. American Bankruptcy Law Journal 83, 179–236.

Rawls, J., 1971. A Theory of Justice. Belknap Press, Cambridge, MA.

U.S. Code, 2012. Vol. 11, sec. 101(5)(A).

Barthes, Roland (1915–80)

Dustin B Garlitz, University of South Florida, Tampa, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Roland Barthes was a twentieth-century French literary critic, philosopher, and cultural theorist important in the trajectories of structuralism and poststructuralism. This article begins by examining Barthes' formative years and influences, and highlights his contributions made to structuralism and poststructuralism in France and beyond. The article then discusses the role that Barthes' work has played in the social and behavioral sciences, including how his writings have been appropriated by disciplines such as sociology and anthropology. Barthes' writings on history are also examined in a disciplinary context.

Barthes, Roland (1915–80)

Roland Barthes, born 12 November 1915 in Cherbourg, France, died 26 March 1980 in Paris, was a French literary critic, philosopher, and cultural theorist. He made contributions to the intellectual movements of structuralism and poststructuralism, and to the fields of linguistics, semiotics, and social theory. Elements of Barthes' theory construction have also been employed in social and cultural anthropology, especially his discussion and theorizing of myth. Additionally, Barthes' study of the phenomena of popular media has influenced communication studies, popular culture studies, as well as intellectual niches such as the sociology of fashion, visual sociology, the sociology of art, aesthetics (including aesthetics of photography and music), music criticism, and music sociology.

Formative Years

Roland Barthes was born in Cherbourg, France in 1915. His father was a naval lieutenant killed in a battle 1 year later. By 1924, at the age of nine, Barthes moved with his mother to Paris. It was in Paris that Barthes received the majority of his education. He began by attending the Lycée Montaigne, and then continued his studies at the Lycée Louis-le-Grand, from 1930 to 1934. Ultimately, Barthes studied classical letters, grammar, and philology at the Sorbonne, receiving degrees in 1939 and 1943. He also studied French Literature and Greek tragedy there. It was between 1935 and 1939 that Barthes initially attended the Sorbonne, receiving a license of classical letters in the latter year.

Barthes' subsequent doctoral studies at the Sorbonne were marked by ill health. He suffered from pulmonary tuberculosis, spending time in sanatoriums in the years 1934–35 and 1942–46. The latter years coincided with the French occupation, at which time he had been exempted from military service for health reasons. Despite his medical condition, Barthes managed to read and write, established a theatrical group named the *Groupe de Théâtre Antique*, and also taught.

On 21 October 1939, Barthes was appointed by the Bordeaux education authority to a lycée in Biarritz on a temporary and provisional basis. Formally, Barthes had only been appointed a supply teacher in Biarritz. On return to Paris, Barthes was unsuccessful in securing a similar post, so he became a study and recreation supervisor at two schools, the Lycée Voltaire and the Lycée Carnot. Despite the limited capacity of his formal appointment there, Barthes gained experience as a tutor and professor at the Lycées Voltaire and Carnot.

Barthes' academic career was interrupted by a second attack of tuberculosis in 1941, and a 6-year absence from active work. During this time, he received a license in grammar and philology, published his first article, and, shortly engaged in premedical study with the intention of pursuing a career in psychiatric medicine. It was not until 1948 that he returned to academic work, first at the Institute Français in Bucharest, Romania, and, in 1949–50, at the University of Alexandria in Egypt.

From 1947 to 1950, Barthes contributed a number of articles on literary criticism to the leftist French paper *Combat*. These articles formed the basis of Barthes' first book, which was an 80-page volume entitled *Writing Degree Zero*, and was published in 1953. From 1950 to 1952, Barthes served as Direction Générale des Affaires Culturelles. His teaching career expanded by research positions with the Centre National de la Recherche Scientifique from 1952 to 1959, where he investigated sociology and lexicology. Barthes would go on to hold a directorship of studies at the École Pratique des Hautes Études from 1960 to 1976. He was initially appointed Chef de travaux in the sixth section of the École Pratique des Hautes Études, the Department of Economics and Social Sciences, or Chairman of the Section of the Social and Economic Sciences, and then 2 years later, Directeur d'études in the Sociology of Signs, Symbols, and Representations. He remained in the latter post for 14 years, until his election to the Collège de France. On 14 March 1976, following a recommendation made by French philosopher Michel Foucault (1926–84), the committee of professors at the Collège de France elected Barthes to a Chair of *Sémiologie Littéraire* (Literary Semiology), which he accepted and held until his death in 1980.

In 1952, when Barthes had joined the Centre National de la Recherche Scientifique, he first engaged in research in lexicology, and then followed in sociology. The earliest product of these studies was a series of amusing examinations and demystifications of contemporary French popular culture myths, the languages of mass culture such as wrestling, soap powders and detergents, toys, Einstein's brain, popular attitudes toward science, wine and milk, and steak and chips.

These analyses were initially contributed each month, from 1954 to 1956, to *Les Lettres Nouvelles*, and were finally gathered and published in a volume entitled *Mythologies* in 1957. This early work on popular culture by Barthes was concerned with the processes of signification, the mechanisms by which meanings were produced and distributed, or how they emerged and were put into circulation. His guiding principle throughout remained to interrogate the falsely obvious and to make explicit what too often remained implicit in the texts and practices of popular culture. One of his purposes was political, with his target being what he considered the bourgeois norm of modern French society.

Mythologies also served as Barthes' most significant attempt to bring the methodology of semiology to bear on popular culture. The volume concluded with the important theoretical essay, 'Myth Today,' and in the essay, Barthes outlined a semiological model for reading popular culture as though it were a text. *Mythologies* was widely read, and intellectual life in France was entering a new period of inquiry when the volume of essays first appeared in 1957. It soon formed part of the literature of the nation's new intellectual movement that of structuralism.

Barthes was one of the early proponents of structuralism in France in the 1950s, and would go on to develop post-structuralism in the 1960s. The primary influence in Barthes' intellectual development was Swiss linguist Ferdinand de Saussure (1857–1913), namely his *Course in General Linguistics*, which had been posthumously published in 1916. In general, Barthes was aware, and, to a certain degree, a disciple, of the structuralism of French anthropologist Claude Lévi-Strauss (1908–2009), including his structural study of myth. To a lesser acknowledged extent, nineteenth-century French sociologist Émile Durkheim (1858–1917), namely his social theories of collective representations and collective conscience, also were influential in the development of Barthes' cultural theory and criticism. Barthes was aware of Durkheim's writings well before the work of Saussure. Barthes was originally labeled one of the four major figures of structuralism in France, along with Lévi-Strauss, psychoanalyst Jacques Lacan (1901–81), and Foucault, with all but Lévi-Strauss ultimately renouncing structuralism and moving into phases of poststructuralism in the 1960s. The other two thinkers that influenced Barthes' structuralism, especially its semiological dimension, were Russian–American linguist Roman Jakobson (1896–1982) and Danish linguist Louis Hjelmslev (1899–1965).

Semiology and Myth

Barthes' semiology took up Saussure's claim that his linguistic theories could be used to study the wider life of signs in society. Accordingly, Barthes applied those theories to the postwar culture of goods, objects, and message in France to understand these as signs within a system of meaning generated in both their collective, structurally interrelated and individual, internally structured relationships, including their second-order cultural connotation, which Barthes referred to as myth. Although the term semiology had been coined at the beginning of the century by Saussure in his *Course in General Linguistics*, and despite it being defined by him systematically as the science, which studies the existence of the signs at the heart of social life, Barthes believed that a science had still not been conventionally introduced to the academic world.

In the mid-1960s, Barthes, therefore, found himself at the forefront of a major theoretical debate, which would place him against the advocates of orthodox linguistics. Barthes started down such a path in *Writing Degree Zero*, where he had reflected on the historical condition of literary language, and on the fact that all language is constricted by its previously ascribed meaning, and that it exists in a specific culture and, therefore, is always full of implicit assumptions about a given social reality. In addition to opposing the orthodoxies of linguistics at the university level, Barthes also complained of academic literary criticism, and did so in *Writing Degree Zero*. One of his complaints was that the academic criticism was psychologically naïve and deterministic. Moreover, Barthes felt it operated with a model of the human personality that was made obsolete by the discoveries of psychoanalysis.

In 1963, in the publishing of his *Elements of Semiology*, Barthes went on to argue how language and speech in their reciprocal relation complement Durkheim's concept of a collective conscience. Barthes, as a structuralist, wanted to locate the importance of the unspoken, unconscious language in the act of writing. He examined the desire and emotion that formed part of written texts, and did so in relation to politics and social life. In *Mythologies*, he focused on the hidden, unconscious messages by the mass media that promote capitalist ideologies. In his unmasking of all ideologies to destroy their effectiveness, Barthes had found that both capitalist and revolutionary languages perpetuate their own myths.

Barthes' method of reading myths like a text derived from his experience as a mythographer. The method also derived from linguistics, or, more precisely, semiotics, the latter understood as the study of sign systems and signification. Barthes made a distinction between denotation of a sign and its connotation, where denotation was the literal meaning, and the connotation was its mythical meaning. Connotation was classified as a symbolism, since connotations were additional meanings present along with the literal meaning of the sign that was in question. Barthes found in 'Myth Today' that myth was constituted by the loss of the historical quality of things, and in it, things themselves lost the memory of their own fabrication.

In *Elements of Semiology*, Barthes referred to the term denotation as primary signification and the term connotation as secondary signification, where the first system of denotation became the plane of expression or signifier of the second system of connotation. The signifiers of connotation were comprised of signs, signifiers, and signifieds united of the denoted system. Barthes argued that it was at the level of secondary signification or connotation that myth was produced for consumption. Myth in such an intellectual and social context referred to ideology as though it were understood as a body of ideas and practices, whereby actively promoting the values and interests of dominant groups of society, defended customary structures of power.

In cultural theory, Barthes' work provided a definition of ideology. Barthes argued that ideology, or myth, operated mainly at the level of connotation, the secondary, often unconscious meanings that texts and practices carried, or could be made to carry. At the time, it was the very concept of

connotation, taken from the work of Hjelmslev, which enabled Barthes to move between the realms of the scientific and the political. In 'Myth Today,' Barthes located ideology on the level of connotation, in what not was said explicitly but in what was implied and thus carried an even greater force. Where he came to propound a theory of semiotics in order to justify his practice of it, Barthes relied heavily on the work Hjelmslev, most notably his brief remarks that distinguished semiotic systems, which denote from those which also connote. Barthes himself was a scholar of connotations, not of denotations. He believed semiotics has to take its lead from linguistics.

Also in cultural theory, it has been argued that Barthes was one of the leading practitioners of semiotics as a political reading of society. Such an argument in this field treats Barthes' *Mythologies* as a series of cutting essays in the interpretation of a number of French cultural phenomena from a near-Marxian perspective. For the Barthes of *Mythologies*, the practice of semiosis was considered hygiene, and it brought to light the abusive way in which signs could be used by one powerful social group at the expense of others. Barthes employed the term myth there in a deliberately tendentious way to refer to such systematic abuse of the process of signification. According to Barthes, there were no limits as to which objects may be so abused.

In the volume *The Rustle of Language*, a collection of 45 essays written between 1967 and 1980, Barthes paid homage to Jakobson and the young Bulgarian–French philosopher Julia Kristeva (1941–), namely their contributions to semiotics. Nonetheless, it was Barthes' *Elements of Semiology* that represented his formal and systematic study in the mechanics of semiotics. The latter publication gave Barthes a solid reputation as a semiotician in France and beyond. This and many other works by Barthes were measured reflections on language sign systems, texts, and the very nature of what they were able to tell us about the human condition. In particular, *Elements of Semiology* was a prime example of scholarship in continental philosophy of language, a study that was partially indebted to the French tradition in the human sciences on one hand, and, on the other hand, an exercise in literary studies. Barthes' work, including *Elements of Semiology*, gained him a reputation in contemporary theory in the humanities and social sciences for searching for common forms taken by sign systems, myths or connotations across all the domains studied, and looking for oppositions, or binary relationships, between domains such as nature and culture, as well as essence and history. Barthes made it clear that the structural study of myth could only supply one half of the truth, and that the other half had to be supplied by empirical means, or by sociology, which involved the practice of quantification.

It was 'Myth Today,' along with *Elements of Semiology* and his work *The Fashion System*, which made up Barthes' second semiological period, and which immediately followed *Writing Degree Zero*. The publication of *Mythologies*, including 'Myth Today,' marked a period of Barthes' career in which he was beginning to elaborate his theoretical framework, one that became increasingly preoccupied by political issues. His Marxism started to push him toward the Communist Party in France, in a collegial manner, but his well-known fainthearted and unsure demeanor ensured that he remained at a distance from their affairs. He discussed his doubts with colleagues in Paris in the 1950s, explaining his worries about that party, and how he was tired of seeing intellectuals paralyzed, unable to decide where they stood in relation to communism. Of most significance, Barthes was depressed by the diminished importance left-wing intellectuals held in France. He was confronted with many contradictory opinions from people he respected, and such confrontations gave him the feeling that his interests in French intellectual and social life had shifted, going awry in the process.

Structuralism

Barthes accepted the label of structuralist at least as a historical label for his work produced in the late 1950s and the majority of the 1960s. Although it is difficult to mark with accuracy the beginnings and end of Barthes' structuralist phase, intellectual histories commonly associate the rise of structuralism as a modern intellectual movement in France with the publication of Lévi-Strauss' *Tristes Tropiques* in 1955. Barthes used the label more cautiously than the majority of others, and associated it exclusively with Saussure's linguistics, which it appears he had not studied until just before composing the earliest essays of *Mythologies*, not any earlier than 1952. The titles of Barthes' two most explicitly structuralist studies suggest brevity in duration and are 'The Structuralist Activity' (1963) and 'An Introduction to the Structural Analysis of Narrative' (1966). The latter was originally published in the journal *Communications*, of which Barthes was a founding editor. *Elements of Semiology* has also been referred to by Barthes scholars as a structuralist study, mostly because of its Saussurean basis. In Barthes' 'Introduction to the Structural Analysis of Narratives,' he began by reminding the reader that narrative was as old as human civilization, and that, therefore, there were simply countless narratives, which could be analyzed. He found that the structural analysis of narratives modeled itself on linguistics. It was linguistics, Barthes found, that stopped at the level of the sentence. Accordingly, he found that it took the sentence at its highest term and proceeded to study all the rules of combination and opposition, which allowed for sentences to be constructed.

Barthes wrote pieces for the literary magazine *Esprit*, in addition to *Les Lettres Nouvelles* and *Combat*. In 1960, a new review called *Tel Quel* was launched, with Barthes collaborating and forming part of its group of intellectuals in France in the 1960s, which included Kristeva, Foucault, and philosopher Jacques Derrida (1930–2004). *Tel Quel* was essentially an avant-garde literary magazine, and Barthes published many essays there. According to him, *Tel Quel* represented the dialectical articulation of a revolutionary, Marxian political stance and an avant-garde cultural activity. Nonetheless, Barthes spent the majority of protests during May 1968 at *Tel Quel* meetings, far removed from the demonstrations and the lecture theaters, as he believed that structures did not 'take to the streets.' By the end of May 1968, the intellectual movement had all but extinguished itself in the pages of *Tel Quel*, the very journal it and its luminaries such as Barthes had established for reasons of promotion.

Outside of France, the reaction against structuralism that May 1968 symbolized went largely unnoticed and many across the rest of the world at this time were only beginning to hear

about structuralism and about one of its proponents named Roland Barthes. Until 1967, *On Racine* was the only one of Barthes' books that had been translated into English. However, later that year English translations of *Writing Degree Zero* and *Elements of Semiology* appeared, and Barthes' reputation as a structuralist began to spread widely. It is in such a fashion that Barthes has been identified as one of the French theorists' who most changed the intellectual landscape of the American academy across the arts, humanities, and social and behavioral sciences.

Barthes' identification as a structuralist was helped by his visits to Japan and the United States in 1966. It was his trip to the United States in October, at Johns Hopkins University in Baltimore, Maryland, that he delivered a paper at an International Symposium on 'The Languages of Criticism and the Sciences of Man,' the events of which were later published as *The Structuralist Controversy* (1970). Nevertheless, by 1968, primarily in his seminar at the École Pratique, Barthes was already engaged in the analysis of a novella by the nineteenth-century French novelist Honoré de Balzac (1799–1850), *Sarrasine*, the result of which appeared in print by Barthes as *S/Z* and marked what, for lack of a better term, has been referred to as his poststructuralist phase. In *S/Z*, Barthes developed a structural analysis of narrative in the context of the fiction of Balzac's *Sarrasine*, with such analysis widely considered a high point of his poststructuralist phase.

Poststructuralism

Barthes himself described his poststructuralist phase as a period marked by his concern with textuality. The textual, in Barthes' sense, is not to be identified with the state of a physical text, as he believed the texts that writers work with are held 'only in language, not in the hand.' Barthes' first three books of his poststructuralist phase each supplied an example of this conviction. In *S/Z*, published in 1970, the text was not Balzac's *Sarrasine*, but Barthes' reading of that story as a language. In *Empire of Signs*, which was published the same year, the text was Japan, a country, which Barthes managed to read. The following year, Barthes published *Sade, Fourier, Loyola*, which took as its basis three texts, or three languages. In contrast to his earlier structural analyses, these three textual studies of Barthes emphasized the process of the text itself, not a closed structure.

In May 1966, Barthes went to Tokyo at the invitation of the director of the Franco-Japanese Institute. He gave a seminar on the structural analysis of narrative. Between 1966 and 1967 he went there three times, bringing back with him from the few weeks he spent there the material for the book on Japan he would eventually write entitled *Empire of Signs*. After the last of Barthes' travels to Japan, he published *Empire of Signs* in 1970. In *Empire of Signs*, Barthes brought to the critical forefront the problem of analyzing history and culture, or the cultural history of Japan, from an ostensibly objective point of view. Barthes' semiotic method seemed to transform history into historiography, or the writing of history. It was at such juncture that Barthes made the reader aware of the way in which the authority of history in the West is grounded in the logic of what the scholars of ancient Greek referred to as the metaphysical *logos*. Barthes found that the *logos* functioned as the mystical power of the word of God, to that which incarnated meaning and reason in the word. In such a sense, Barthes found that history itself functioned as the word, leading him to argue that the reason of history was the reason of the word. Such a presupposition of Barthes enabled him to generate a vision of culture that was recorded as history, with history also recorded as culture, or the writing of cultural history through which the meaning of culture was codified in a system and represented as the truth. Accordingly, *Empire of Signs* represented a prime example of Barthes' teasing out the practice of writing history and representing culture, or writing cultural history, within the logic of poststructuralism. On such grounds, *Empire of Signs* reflected Barthes' attempt to transgress the cultural boundaries of historical writing and semiology.

Whereas the early work of Barthes intended to unlock the secrets of representational forms such as mythology, his later work complicated the simplicity of that notion, and even questioned its relevance with respect to the cultural production of meaning. Signs, codes, structures, and systems were all still important in his later work, such as *Empire of Signs*, but mainly as conventions to be played with and enjoyed. Gone were the inventories and labels that enabled Barthes, as semiologist, to produce a metalanguage of scholarly inquiry. In *S/Z*, *The Pleasure of the Text*, and *A Lover's Discourse*, Barthes was not so much interested in what was represented and how, as he was in generating new ways of reading for their own sake.

Barthes' essay 'The Death of the Author' (1968) represented a prime intellectual influence in the field of literary criticism and theory. His 'From Work to Text' was an extremely important piece in the area of continental aesthetics, and his book *Camera Lucida: Reflections on Photography* (1980) was important to aesthetics as a scholarly field. His volume of collected essays translated and published in English as *Image-Music-Text* (1977) included 'The Death of the Author' and 'From Work to Text,' and represented an important compilation of essays on aesthetics in the tradition of poststructuralism. Works of his such as 'The Plates of the *Encyclopedia*' (1964) from his *New Critical Essays*, translated and published in English in 1980, are representative of the core of continental aesthetics in the poststructuralist tradition. Barthes also wrote on fashion, with his *The Fashion System* published in 1967, representing an analysis of fashion and its semiotics, which marked an early contribution to visual sociology and the sociology of fashion.

Barthes turned extensively to literary studies in the 1970s with the publication of *S/Z: An Essay* in 1970, and *The Pleasure of the Text* in 1975. His *A Lover's Discourse: Fragments*, published in 1978, sold 60 000 copies in France. He employed the methodologies of structuralism and poststructuralism to understand the phenomena of love. His literary study *On Racine* has been published in France in 1963. He would go on to practice the art of writing autobiography with the publication of his *Roland Barthes by Roland Barthes* in 1975. He published more literary criticism and theory in 1980, marked by *Sade, Fourier, Loyola*. A notable theory from this work surrounds the way he studies the different modes of reception and evaluation of a text, and proposes a distinction between a 'writerly' (*scriptable*) text, where the reader can rewrite, or desires to write, and a 'readerly' (*lisible*) text, which can only be read. With the publication of *The Pleasure of the Text*, Barthes examined the relationship between pleasure, ecstasy, the untranslatable French term *jouissance*, and desire. The day

after his mother's death in October 1977, he began a diary of mourning that would come to be published as *Mourning Diary*. Recently, Barthes' lectures at the Collège de France from the late 1970s have been translated into English for the first time, including *The Neutral* (1977–78) and *The Preparation of the Novel* (1978–79 and 1979–80), as well as his first course at the Collège after being elected to his Chair there, *How to Live Together* (1976–77). In 1974, Barthes went to China with Kristeva and others, agreeing to write an article about his experience for *Le Monde* on his return to France. Recently, his *Travels to China* have been published in English. His posthumously published Collège de France lectures and courses reveal how he went about incorporating influences from the orient in his literary criticism and theory.

Barthes' Influence in the Social and Behavioral Sciences

In the social and behavioral sciences, Barthes' influence primarily rests on social theory, sociological theory, the sociology of consumer culture, the sociology of culture/cultural sociology, and theory in social and cultural anthropology. Much of the early work done in the field of contemporary cultural studies, including its theoretical framework, rests on Barthes' writings. This is especially the case of his essays that comprise *Mythologies*, such as 'The World of Wrestling.' The work forms an integral part of analysis in cultural theory and popular culture studies, and bridges the humanities with the social and behavioral sciences. Some of his work has been influential in the development of the sociology of literature, and in general, social and cultural criticism, including the sociology of everyday life and the sociology of leisure. Some of his essays have formed a basis for the sociology of media as well.

For Barthes, the human sciences were a reformation of traditional humanism along lines of several modern intellectual movements, especially those in linguistics, which now were employed as a tool for analyzing any social practice as a language. As a result, the imprecise concept of the human could be demystified and studied scientifically. In his essay 'From Science to Literature' (1967), Barthes begins with a discussion of the function the social and human sciences serve in the French university system, and what the term 'science' actually signifies. Barthes held the conviction that structuralism would never be anything more than another science, and expressed so in 'From Science to Literature.' Barthes saw the opposition between science and literature of being of particular importance to structuralism. Barthes characterizes structuralism as paying special attention to classifications, orders, and arrangements. Further, he finds it is in such a fashion that it comprises forms of discourse. Barthes was interested in the general mythology that each literary work participated in, and that which affected the contents of structuralism. On his account, the contents included the inner logic that governs each literary work as myth.

Barthes' work *Michelet* was published in France in 1954, and was a critical analysis of the French historian Jules Michelet (1798–1874). The work inspired a revival of interest in the eminent nineteenth-century historian. Another of Barthes' essays, 'The Discourse of History' (1967), examined semiotics and continental philosophy of language among history, as traditionally understood in the framework of the human sciences. In historical writing, including the discourse of history, Barthes essentially scanned a historical text for traces of the writing subject. He at first found 'shifters,' a term borrowed from Jakobson. These were divided among testimonial shifters, organizing shifters, and shifters that marked interventions of the writer or that which appealed to the listener, whereas testimonial shifters marked the presence of the historian as someone who gathered and weighed up evidence from other sources, and organizing shifters represented where narration diverges from chronological sequence and reorganized itself according to the needs of the discourse, ultimately establishing its own structure of a temporal nature.

'The Discourse of History' attacked that discourse's pretension to deliver 'the facts' guaranteed by 'the real.' Barthes' essay argued that the discourse of history performed a sleight of hand whereby a discursive operator, the referent, was projected into a realm supposedly beyond signification, from which position it could be thought to precede and determine the discourse, which posited it in its very capacity as referent. Barthes claimed that such a paradox governed the entirety of the question of the distinctiveness of historical discourse, as he found the fact could only have a linguistic existence, as a term in a discourse, and yet it was exactly as if this existence were merely the 'copy,' purely and simply, of another existence situated in the extra-structural domain of what was the 'real.' This type of discourse was doubtless the only type in which the referent was targeted for something as external to the discourse without it ever being possible to attain it outside its confines. Barthes went on to propose an eminently structuralist account of such an operation, where he argued that the discourse of history is guilty of reducing the three-term structure of signification (signifier-signified-referent) to a two-term structure (signifier-referent), or rather, smuggling into such an ostensibly two-term structure an illicit signified, a 'real in itself,' which furtively transformed into an awkward signified. Historical discourse did not follow what was 'the real' for Barthes, as he felt it could do no more than signify 'the real,' in the capacity of constantly repeating that it happened, without such an assertion amounting to anything more than the signified reverse side of the entirety of the process of historical narration itself.

Barthes' interests in popular media and events were due in part to what he saw as an abuse in such phenomena of ideology. He conflated elements of what had been perceived as 'high' or 'low' culture. In November 1964, in the journal *Communications*, Barthes published his essay 'Rhétorique de l'image' ('Rhetoric of the Image'), which was comprised of an analysis of one poster advertising Panzani pasta. This and other writings examined the semiology and sociology of advertising. Barthes' *The Fashion System* was part sociology of fashion. His 'History and Sociology of Clothing' was published in French in the *Annales* in 1957, and his 'For a Sociology of Clothing' was published in French there in 1960. *The Fashion System* included these essays, and the entirety of the manuscript was labored over by Barthes from 1957 to 1963. He decided not to publish it until 3 years later, at which time it merely represented a document in the history of semiology.

Barthes also wrote music criticism, and other of his writings involved the aesthetics of music, and aesthetics, in general, and to a lesser degree, music sociology. The full extent of Barthes'

music criticism, aesthetics of music, and music sociology are collected in the latter part of a volume of his essays entitled *The Representation of Forms*, which was published in English in 1991. Barthes did much to shift the textual study of popular music away from the traditional concerns of musicologists and onto semiotic questions of representation and meaning, as well as emphasizing the pleasures of the text, or musical work. Barthes' essay 'The Grain of the Voice' is generally regarded in cultural theory as the work that laid the ground for a textual study of popular music. Barthes complained in this essay that music critics were obsessed with adjectives and that such intimidation of detail actually served to prevent an adequate understanding of the text. In order to grasp the meaning of a musical experience, Barthes argued that one ought to dispense with trying to capture the minutiae but instead respond directly to its surface, the so called 'grain of the voice' that was expressed in the materiality of its body.

See also: Consumption, Sociology of; Continental Philosophy of Language; Durkheim, Emile (1858–1917); Fashion, Sociology of, Foucault, Michel (1926–84); Jakobson, Roman (1896–1982); Leisure, Sociology of; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Mass Media and Cultural Identity; Postmodernism in Sociology; Postmodernism: Philosophical Aspects; Representations, Social Psychology of; Saussure, Ferdinand de (1857–1913); Sociological Theory; Visual Sociology; Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Allen, Graham, 2003. Roland Barthes. Routledge, London.
Batchen, Geoffrey (Ed.), 2009. Photography Degree Zero: Reflections on Roland Barthes's "Camera Lucida". The MIT Press, Cambridge, MA.
Brown, Andrew, 1992. Roland Barthes: The Figures of Writing. Oxford University Press, Oxford.
Calvet, Louis-Jean, 1994. Roland Barthes: A Biography. translated by Sarah Wykes. Indiana University Press, Bloomington.
Culler, Jonathan, 2001. Roland Barthes: A Very Short Introduction. Oxford University Press, Oxford.
Lavers, Annette, 1982. Roland Barthes: Structuralism and After. Harvard University Press, Cambridge, MA.
Lombardo, Patrizia, 2010. The Three Paradoxes of Roland Barthes. University of Georgia Press, Athens.
McQuillan, Martin, 2011. Roland Barthes. Palgrave Macmillan, Basingstoke and New York.
Moriarty, Michael, 1991. Roland Barthes. Stanford University Press, Stanford.
O'Meara, Lucy, 2013. Roland Barthes' Lectures at the Collège de France. Liverpool University Press, Liverpool.
Rabate, Jean-Michel (Ed.), 1997. Writing the Image after Roland Barthes. University of Pennsylvania Press, Philadelphia.
Rylance, Rick, 1994. Roland Barthes. Harvester Wheatsheaf, New York.
Shawcross, Nancy M., 1996. Roland Barthes on Photography: The Critical Tradition in Perspective. University Press of Florida, Gainesville.
Stafford, Andy, 1998. Roland Barthes, Phenomenon and Myth: An Intellectual Biography. Edinburgh University Press, Edinburgh.
Thody, Philip, 1984. Roland Barthes: A Conservative Estimate. The University of Chicago Press, Chicago.
Ungar, Steven, 1983. Roland Barthes: The Professor of Desire. University of Nebraska Press, Lincoln.
Ungar, Steven, McGraw, Betty R. (Eds.), 1989. Signs in Culture: Roland Barthes Today. University of Iowa Press, Iowa City.
Wasserman, George R., 1981. Roland Barthes. Twayne Publishers, Boston.
Wiseman, Mary B., 1989. The Ecstasies of Roland Barthes. Routledge, London.

Bayesian Graphical Models and Networks

David Heckerman, Microsoft Corporation, Redmond, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 2, pp. 1048–1052, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

This article describes a class of graphical models known as directed acyclic graphical (DAG) models. These models have numerous uses including data analysis, modeling of causal relationships, and representing and reasoning about uncertainty in expert systems and decision analysis. In this article, we define the model class and describe its basic properties. In addition, we discuss how DAG models are used in expert systems and decision analysis, and describe methods for inferring the structure and distributions of DAG models from data.

Introduction

A graphical model represents the probabilistic relationships among a set of variables. Nodes in the graph correspond to variables, and the absence of edges corresponds to conditional independence. Graphical models are becoming more popular in statistics and in its applications in many different fields for several reasons. For example, they provide a language to facilitate communication between a domain expert and a statistician, provide flexible and modular definitions of families of probability distributions, and are amenable to scaleable computational techniques (e.g., Pearl, 1988; Lauritzen, 1996; Whittaker, 1990).

The subject of this article is directed acyclic graphical (DAG) models. These models have numerous uses including data analysis (e.g., Whittaker, 1990; Spiegelhalter and Thomas, 1998), modeling of causal relationships (e.g., Pearl, 2000; Spirtes et al., 2001), and representing and reasoning about uncertainty in expert systems (e.g., Cowell et al., 1999). Specific applications of graphical models include diagnosis and troubleshooting, medical monitoring, genetic counseling, information retrieval, natural language processing, weather forecasting, manufacturing, digital communication, and machine vision.

In the section 'Definition and Basic Properties,' we define the DAG model and describe its basic properties. In the section 'DAG Models for Expert Systems and Decision Analysis,' we describe the use of DAG models in expert systems and decision analysis. In the section 'Learning DAG Models from Data,' we describe methods for constructing the structure and distributions of DAG models from data. In the section 'Pointers to the Literature' we conclude with additional pointers to the literature.

Definition and Basic Properties

Let us begin with some notation. We denote a variable by a capitalized token (e.g., X, X_i, Θ, *Age*), and the state or value of a corresponding variable by that same token in lower case (e.g., x, x_i, θ, *age*). We denote a set of variables by a bold-face capitalized token (e.g., $\mathbf{X}$, $\mathbf{X}_i$, $\mathbf{Pa}_i$). We use a corresponding bold-face lower-case token (e.g., $\mathbf{x}$, $\mathbf{x}_i$, $\mathbf{pa}_i$) to denote an assignment of state or value to each variable in a given set. We use $p(x|y)$ to denote the probability or probability density that $X = x$ given $Y = y$. We also use $p(x|y)$ to denote the probability distribution for X given Y. Whether $p(x|y)$ refers to a probability, probability density, or a probability distribution will be clear from context. Finally, we use $\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}$ to mean that $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent given $\mathbf{Z}$.

A DAG model for a set of variables or *domain* $\mathbf{X} = \{X_1, \ldots, X_n\}$ consists of (1) a network structure $\mathbf{s}$ that encodes a set of conditional independence assertions about variables in $\mathbf{X}$, and (2) a set $\mathbf{p}$ of local probability distributions associated with each variable. Together, these components define the joint probability distribution for $\mathbf{X}$. The network structure $\mathbf{s}$ is a directed acyclic graph, that is, each edge is directed and there are no paths along these directed edges that lead from a variable to itself. The nodes in $\mathbf{s}$ are in one-to-one correspondence with the variables $\mathbf{X}$. We use X_i to denote both the variable and its corresponding node, and $\mathbf{Pa}_i$ to denote the parents of node X_i in $\mathbf{s}$ as well as the variables corresponding to those parents. The *lack* of possible directed edges in $\mathbf{s}$ encode conditional independencies. In particular, given structure $\mathbf{s}$, the joint probability distribution for X is given by

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|\mathbf{pa}_i) \qquad [1]$$

The *local probability distributions* P are the distributions corresponding to the terms in the product of eqn [1]. Consequently, the pair $(\mathbf{s}, \mathbf{p})$ encodes the joint distribution $p(\mathbf{x})$.

As an example, consider **Figure 1**, which illustrates the structure of a DAG model for troubleshooting a printing problem on the Microsoft Windows 95 operating system. This model structure was constructed from expert knowledge using judgments of cause and effect. Note that, in general, the probability distributions in a DAG model may be subjective or objective.

In the remainder of this section, we discuss several basic issues concerning DAG models. Equation [1] implies that, for all X, the nondescendants of X are independent of X given the parents of X. Many additional independencies are entailed by this basis of independencies. In **Figure 1**, for example, we have that *Print Output OK* and *Application Output OK* are

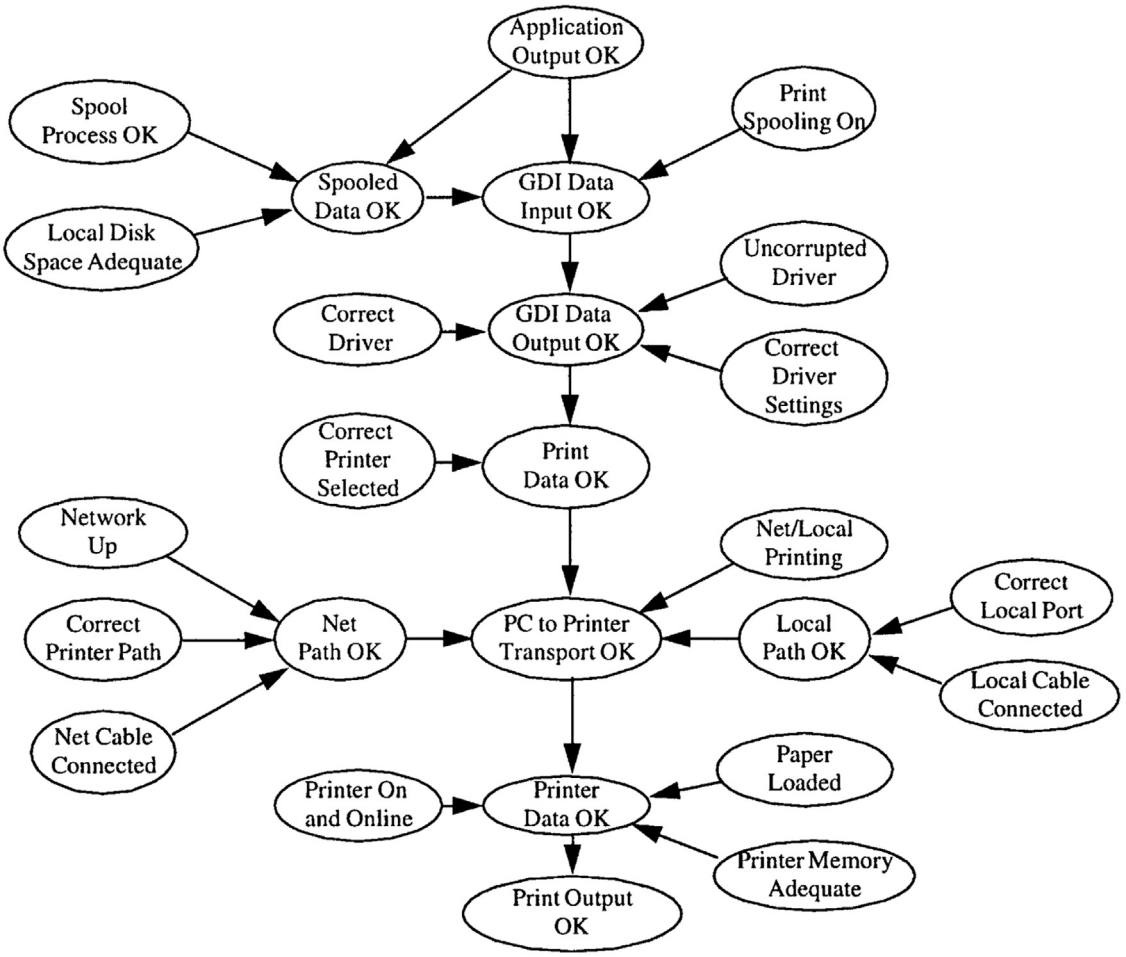

Figure 1 A DAG model for troubleshooting a printing problem.

conditionally independent given *Print Data OK*. Pearl (1988) describes a condition, called *d-separation*, that allows one to read off any independency from a DAG model in a computationally efficient manner.

Another commonly used graphical model is the undirected graphical (UG) model, also known as a Markov Random Field or Markov Network (e.g., Whittaker, 1990; Lauritzen, 1996). The independencies that can be encoded by DAG and UG models overlap, but there are independencies that can be encoded by UGs and not DAGs, and vice versa. For example, consider the UG model in **Figure 2(a)** for the domain (X, Y, Z, W). This graph encodes the independencies $X \perp Z | (Y, W)$ and $Y \perp W | (X, Z)$. There is no DAG model for these four variables that can encode both of these independencies. In contrast, consider the DAG model in **Figure 2(b)** for the domain (X, Y, Z). This graph encodes the marginal independence of X and Z and the possible dependence of X and Z given Y. No UG model for these variables can encode this set of relationships.

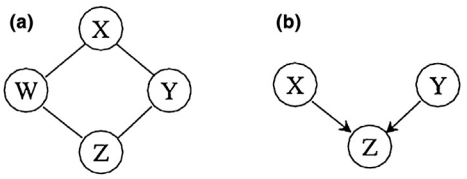

Figure 2 Undirected (a) and directed (b) graphical-model structures.

The relationships encoded by the 'v structure' or 'collider' in **Figure 2(b)** has an intuitive explanation. For example, consider the subgraph in **Figure 1** for the variables *Correct Local Port*, *Local Cable Connected*, and *Local Path OK*, which has this structure. Before one observes the value of *Local Port OK*, the remaining two variables are marginally independent. When we observe that *Local Port OK* is false, however, the remaining two variables become dependent. If – say – *Correct Local Port* subsequently is observed to be false, then *Local Cable Connected* is more likely to be true, because the faulty local path has been explained by the first observation. In general, this pattern of relationships is sometimes referred to as the 'explaining away' phenomenon (Pearl, 1988).

Another interesting point about DAG models is that two or more DAG models can be equivalent in the sense that they represent the same independencies. Given the domain (X, Y, Z), for example, the DAG models $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \rightarrow Z$, and $X \leftarrow Y \leftarrow Z$ encode the same independency: $X \perp Z | Y$. Chickering (1995) describes a simple and computationally

efficient algorithm for determining whether two DAG models are equivalent. First, he defines a directed edge from X to Y to be *covered* if the parents of X and Y are the same with the exception that X is also a parent of Y. He then proves that two DAG models are equivalent if and only if there exists a sequence of reversals of covered directed edges that transforms one model to the other.

Finally, we note that the definition of DAG model does not restrict the form of the local distributions $p(x_i|\mathbf{pa}_i)$. Later, we describe several commonly used models for local distributions. Here, we mention that, when a variable and its parents are continuous, a commonly used local distribution is the linear regression with Gaussian error. When all variables in a domain are continuous and the error terms are mutually independent, the resulting DAG model is equivalent to a structural equation model with independent errors (e.g., Pearl, 2000).

DAG Models for Expert Systems and Decision Analysis

The modern use of DAG models began with researchers in Management Science and AI communities (e.g., Howard and Matheson, 1981) who used them to encode prior or 'expert' knowledge for decision making and expert-system inference. These communities refer to this class of model as Knowledge Maps and Bayesian Networks, respectively. In this section, we examine this use of DAG Models.

First, let us consider how a person – who we will call the decision maker – constructs a DAG model for a given domain $\mathbf{X}$ from their knowledge. One method for construction starts with a total ordering on the variables provided by the decision maker. For simplicity, let us relabel the variables $\mathbf{X} = (X_1, \ldots, X_n)$ so that X_i is the ith variable in this ordering. From the chain rule of probability, we have

$$p(\mathrm{x}) = \prod_{i=1}^{n} p(x_i|x_1, \ldots, x_{i-1}) \qquad [2]$$

For every X_i there will be some subset $\Pi_i \subseteq \{X_1, \ldots, X_{i-1}\}$ such that X_i and $\{X_1, \ldots, X_{i-1}\}\backslash\Pi_i$ are conditionally independent given Π_i, that is, for any $\mathbf{x}$

$$p(x_i|x_1, \ldots, x_{i-1}) = p(x_i|\pi_i) \qquad [3]$$

Combining eqns [2] and [3], we obtain

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|\pi_i) \qquad [4]$$

Comparing eqns [1] and [4], we see that the variables sets $(\Pi_1, \ldots, \Pi_n)$ correspond to the DAG-model parents $(\mathbf{Pa}_1, \ldots, \mathbf{Pa}_n)$, which in turn fully specify the directed edges in the network structure $\mathbf{s}$. Consequently, to determine the structure of a DAG model, a decision maker (1) orders the variables somehow and (2) determines the variables sets that satisfy eqn [3] for $i = 2, \ldots, n$. Finally, the decision maker assesses the local distributions $p(x_i|\mathbf{pa}_i)$.

The approach for constructing model structure just described has an important drawback. If one chooses the variable order carelessly, then the resulting network structure may fail to reveal many conditional independencies among the variables. Fortunately, there is another often-used technique for constructing DAG models that does not require an ordering. The approach is based on two observations: (1) people can often readily assert causal relationships among variables and (2) absent causal relationships typically correspond to conditional independencies. Thus, to construct a DAG model for a given set of variables, one simply draws arcs from cause variables to their immediate effects. In almost all cases, doing so results in a network structure that satisfies the definition eqn [1]. The DAG model in **Figure 1** was constructed in this manner.

Once a DAG model is constructed (whether from prior knowledge, data, or both), we can use it to answer probabilistic queries. In the context of decision analysis, the answers to these queries can be used for decision making. In the context of expert systems, these answers convey 'expert advice.' For example, given the DAG model in **Figure 1**, we may be interested in the probability that *Printer On* and *Online* is true, given *Print Output OK* is false. We can obtain this probability by applying the rules of probability to the joint distribution defined by the DAG model. Many researchers refer to this process as *probabilistic inference*.

Direct computation via the joint distribution is expensive. Investigators, however, have developed probabilistic-inference algorithms that use the conditional independencies in the DAG model to speed up inference, sometimes dramatically. Many of these algorithms are limited to DAG models in which all variables are discrete. For example, Howard and Matheson (1981) and Shachter (1988) developed an algorithm that reverses arcs in the network structure until the answer to the given probabilistic query can be read directly from the graph. In this algorithm, each arc reversal corresponds to an application of Bayes's theorem.

Pearl (1988) developed a message-passing scheme that updates the probability distributions for each node in a DAG model in response to observations of one or more variables. Lauritzen and Spiegelhalter (1988) created an algorithm that first transforms the DAG model into a tree where each node in the tree corresponds to a subset of variables in $\mathbf{X}$. The algorithm then exploits several mathematical properties of this tree to perform probabilistic inference. Lauritzen (1992) extended this algorithm to include DAG models that encode multivariate-Gaussian or Gaussian-mixture distributions.

When the edges of a DAG model structure form a tree, these methods for exact probabilistic inference scale linearly in the number of domain variables. In general, however, these algorithms are NP-hard (Cooper, 1990). The source of the complexity lies in undirected cycles in the DAG model – cycles in the structure where we ignore the directionality of the edges. When a DAG model contains many undirected cycles, inference is intractable. For many applications, however, structures are simple enough (or can be simplified sufficiently without sacrificing much accuracy) so that inference is efficient. For those applications where generic inference methods are impractical, researchers are developing techniques that are custom tailored to particular network topologies or to particular inference queries. In addition, approximate algorithms involving Monte Carlo techniques (e.g., Spiegelhalter and Thomas, 1998) and variational methods (e.g., Jordan, 1998) are proving useful in many applications.

Learning DAG Models from Data

There are a variety of methods for constructing DAG models from data or a combination of data and prior knowledge. We concentrate here on the Bayesian approach, but briefly consider other approaches.

First, suppose the structure $\mathbf{s}$ of the DAG model is known and we are interested in inferring the local distributions from our prior knowledge and a set of data $\mathbf{d} = (\mathrm{x}^1, \ldots, \mathrm{x}^N)$ considered to be a sample from an infinitely exchangeable sequence. For simplicity, we assume that each local distribution is parametric so that the data form a random sample from

$$p(\mathbf{x}|\boldsymbol{\theta}_s, \mathrm{s}) = \prod_{i=1}^{n} p(x_i|\mathbf{pa}_i, \boldsymbol{\theta}_i, \mathbf{s}) \qquad [5]$$

where $\boldsymbol{\theta}_s = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ are the parameters of the local distributions.

If the parameters in $\boldsymbol{\theta}_s$ are mutually independent and there are no missing data, then the parameters in $\boldsymbol{\theta}_s$ are independent *a posteriori* and we can infer each local distribution separately. Because a local distribution can be thought of as a probabilistic classification or regression model, learning the local distributions of a DAG model under the assumptions above is a familiar task. Examples of classification/regression models that have been used in conjunction with DAG models include probabilistic decision trees, generalized linear models, neural networks, and probabilistic support-vector machines.

One class of local distributions used frequently are distributions from the exponential family. For example, when X_i and its parents are discrete, $p(x_i|\mathbf{pa}_i, \boldsymbol{\theta}_i, \mathbf{s})$ can be a set of mutually independent multinomial distributions, one for each instance of the parents of X_i. When X_i is continuous, $p(x_i|\mathbf{pa}_i, \boldsymbol{\theta}_i, \mathbf{s})$ can be a set of linear regressions with mutually independent error terms, one for each instance of the discrete parents of X_i. In these cases, the posterior uncertainty can be determined and integrated over in closed form (e.g., Heckerman and Geiger, 1995).

When the parameters in $\boldsymbol{\theta}_s$ are not mutually independent, when there are missing data (and/or latent variables), when local distributions from outside the exponential family are used, or when complex expectations over posterior uncertainty are required, closed-form expressions are unavailable and Markov Chain Monte Carlo methods are often employed. The BUGS system (Spiegelhalter and Thomas, 1998) is a popular software tool designed for these situations. If maximum-likelihood or maximum *a posteriori* (MAP) approximations are sufficient, then optimization-based algorithms such as Expectation–Maximization and gradient descent typically are used.

Now, let us suppose that we are also uncertain about the structure of the DAG model. Again concentrating on the Bayesian approach, we model this uncertainty with a probability distribution $p(\mathbf{s})$ over the possible structures (see also e.g., Bayesian Statistics). Given data $\mathbf{d}$, we update these uncertainties using Bayes's rule

$$p(\mathbf{s}|\mathbf{d}) \propto p(\mathbf{s})\, p(\mathbf{d}|\mathbf{s}) = p(\mathbf{s}) \int p(\mathbf{d}|\boldsymbol{\theta}_s, \mathbf{s})\, p(\boldsymbol{\theta}_s|\mathbf{s}) d\boldsymbol{\theta}_s \qquad [6]$$

where $p(\mathbf{d}|\mathbf{s})$ is the marginal likelihood of the data given the model structure. Given the posterior uncertainties, we can either 'construct' a single DAG model by – say – selecting the most likely structure, or average over some or all of the model structures.

Under the assumptions described above – (1) mutual independence of the parameters in $\boldsymbol{\theta}_s$, (2) complete data, and (3) local distributions from the exponential family – the marginal likelihood can be computed efficiently and in closed form (e.g., Heckerman and Geiger, 1995). Furthermore, simple heuristic search methods, such as greedy hill climbing, can be used to find a structure whose posterior is relatively large. In applications where computational efficiency is important such as data mining, where both the number of domain variables and the number of data samples is large, these techniques are used extensively (Heckerman, 1996).

There are several non-Bayesian methods for learning structure from data. One set of techniques mimics Bayesian model selection, but incorporates non-Bayesian criteria for selection such as Akaike's information criterion (AIC). Another, developed independently by Pearl (2000) and Spirtes et al., (2001) uses local non-Bayesian tests for conditional independence together with search to identify model structures compatible with the data.

Pointers to the Literature

Perhaps the first person to use DAG models was Wright (1921), who employed what are now called structural equation models for the analysis of crop failure. Good (1961) also developed the representation. In both cases, researchers attributed causal semantics to DAG models. In fact, this model class remains quite popular for encoding and reasoning about causal interactions (Glymour and Cooper, 1999; Pearl, 2000; Spirtes et al., 2001; see also Latent Structure and Causal Variables).

In addition to directed models, researchers have developed and used graphical models containing undirected edges, bidirectional edges, and mixtures of directed and undirected edges (e.g., Wermuth, 1976; Whittaker, 1990; Lauritzen, 1996; Jordan, 1998).

Finally, there are numerous software packages available for the knowledge-based and data-based construction of DAG models. A summary is provided at http://http.cs.berkeley.edu/murphyk/Bayes/bnsoft.html.

See also: Graphical Methods: Presentation.

Bibliography

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, New York, NY.

Chickering, D., 1995. A transformational characterization of equivalent Bayesian network structures. In: Besndard, P., Hanks, S. (Eds.), Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Mateo, CA, pp. 87–98.

Cooper, G., 1990. Computational complexity of probabilistic inference using Bayesian belief networks (research note). Artificial Intelligence 42, 393–405.

Cowell, R., Dawid, A.P., Lauritzen, S., Spiegelhalter, D., 1999. Probabilistic Networks and Expert Systems (Statistics for Engineering and Information Science). Springer Verlag, New York.

Glymour, C., Cooper, G. (Eds.), 1999. Computation, Causation, and Discovery. MIT Press, Cambridge, MA.

Good, I., 1961. A causal calculus (I). British Journal of Philosophy of Science 11, 305–318.
Good, I., 1983. Good Thinking: The Foundations of Probability and Its Applications. University of Minnesota Press, Minneapolis, MN, pp. 197–217.
Heckerman, D., 1996. Bayesian networks for data mining. Data Mining and Knowledge Discovery 1, 79–119.
Heckerman, D., 1998. A Tutorial on Learning with Bayesian Networks. Springer, New York, NY.
Heckerman, D., Geiger, D., 1995. Learning Bayesian networks: a unification for discrete and Gaussian domains. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, pp. 274–284.
Technical Report TR-95-16, Microsoft Research, Redmond, WA (February 1995).
Howard, R., Matheson, J., 1981. Influence diagrams. In: Howard, R., Matheson, J. (Eds.), Readings on the Principles and Applications of Decision Analysis, vol. II. Strategic Decisions Group, Menlo Park, CA, pp. 721–762.
Jordan, M. (Ed.), 1998. Learning in Graphical Models. Kluwer, Dordrecht.
Lauritzen, S., 1992. Propagation of probabilities, means, and variances in mixed graphical association models. Journal of the American Statistical Association 87, 1098–1108.
Lauritzen, S., 1996. Graphical Models. Clarendon Press, Oxford, UK.
Lauritzen, S., Spiegelhalter, D., 1988. Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society B 50, 157–224.
Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge, MA.
Nielsen, T.D., Jensen, F.V., 2009. Bayesian Networks and Decision Graphs. Springer, New York.
Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, CA.
Pearl, J. (Ed.), 2000. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, UK.
Shachter, R., 1988. Probabilistic inference and influence diagrams. Operations Research 36, 589–604.
Spiegelhalter, D., Thomas, A., 1998. Graphical modeling for complex stochastic systems: the BUGS project. IEEE Intelligent Systems and Their Applications 13, 14–15.
Spirtes, P., Glymour, C., Scheines, R., 2001. Causation, Prediction, and Search, second ed. MIT Press, Cambridge, MA.
Wermuth, N., 1976. Analogies between multiplicative models in contingency tables and covariance selection. Biometrics 32, 95–108.
Whittaker, J., 1990. Graphical Models in Applied Multivariate Statistics. Wiley, New York.
Wright, S., 1921. Correlation and causation. Journal of Agricultural Research 20, 557–585.

Bayesian Models in Neuroscience

Will Penny, Wellcome Trust Centre for Neuroimaging, University College, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The last two decades have seen a revolution in the statistical sciences with classical inferential approaches becoming slowly replaced by Bayesian ones. This has impacted on many fields in science and engineering, including neuroscience. A Bayesian perspective on neuroscience includes the idea that human behavior in many domains is close to statistical optimality and that the structure and machinery of the brain itself is reflected in message-passing algorithms that implement Bayesian inference.

This article presents a review of Bayesian models in Neuroscience. The work in this field splits broadly into two categories: (1) Bayesian modeling of behavior, which develops normative behavioral models based on statistical optimality principles, and (2) Bayesian models of the brain, which describe the internal operations of the brain that would give rise to such behaviors. In the latter category, neuroscientific data from a variety of sources are used including invasive electrophysiological recordings from mice, rats, and monkeys. It also focuses on detailed neurophysiological mechanisms over a range of spatial scales, from plasticity in dendrites to changes in connectivity among brain regions. Given the nature of this encyclopedia, our review focuses most strongly on the former category (behavior), as this is more closely related to psychology and sociology.

The review is structured in two main sections. The first section provides a mathematical description of Bayesian inference, introducing the basic concepts that will be referred to later on. We use the example of a game of tennis to make the main ideas more concrete. The second section reviews Bayesian models of behavior, focusing on a range of topics. Under 'sensory integration,' we review how, for example, information from visual and tactile senses are optimally combined. Under 'visual processing,' we review models of visual search and the processing of ambiguous or conflicting visual information. Under 'sensorimotor integration,' we describe how dynamic Bayesian models accurately describe how the consequences of movements can be predicted and perceptually downweighted. Under 'collective decision making,' we review how Bayesian inference has been used to describe how groups of people interact to make decisions.

Bayesian Inference

Consider some quantity, x. Our beliefs about the likely values of x can be described by the probability distribution $p(x)$. If we make a new observation y that is related to x, then we can update our belief about x using Bayesian inference. In statistics the optimal way of updating your beliefs is via Bayes rule.

First we need to specify the likelihood of observing y given x. This is specified by a probability distribution called the likelihood, $p(y|x)$. It tells us, if we know x, what are the likely values of y. Our updated belief about x, that is, after observing the new data point y is given by the posterior distribution $p(x|y)$.This can be computed via Bayes rule:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \qquad [1]$$

The denominator ensures that $p(x|y)$ sums to 1 over all possible values of x, i.e., that it is a probability distribution. It can be written as

$$p(y) = \int p(y|x')p(x')\mathrm{d}x' \qquad [2]$$

Equations [1] and [2] describe the basic computations underlying Bayes rule. These are multiplication and normalization (eqn [1]) and marginalization (eqn [2]). Following Wolpert and Ghahramani (2004), we will use the game of tennis to illustrate key points. Imagine that you are receiving serve. One computation you need to make before returning serve is to estimate x, the position of the ball when it first hits the ground, as depicted in **Figure 1**.

It is possible to make an estimate solely on the basis of the balls trajectory, i.e., via the data y. We can find the value of x which maximizes the likelihood, $p(y|x)$. This is known as Maximum Likelihood (ML) estimation. It is also possible to

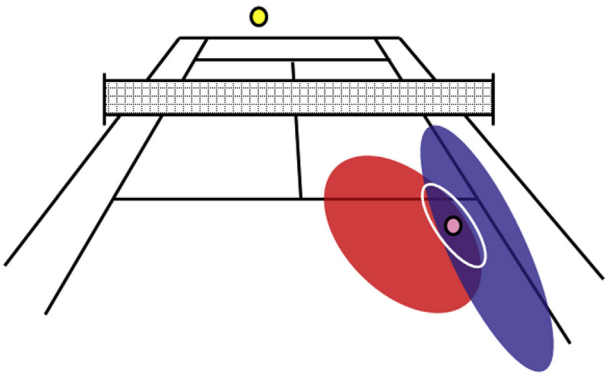

Figure 1 Estimating the position of the ball when it first lands. The prior is shown in blue, the likelihood distribution in red, and the posterior distribution with the white ellipse. The maximum posterior estimate is shown by the magenta ball. This estimate can be updated in light of new information about the ball's trajectory (yellow). Adapted from Wolpert, D., Ghahramani, Z., 2004. Bayes rule in perception, action and cognition. In: Gregory, I.L. (Ed.), Oxford Companion to the Mind. Oxford Univ.

estimate the uncertainty in this estimate. The ML estimate and the uncertainty in it together give rise to the likelihood distribution shown in Figure 1.

But before our opponent hits the ball we may have a good idea as to where they will serve. It may be the case, for example, that when they serve from the right the ball tends to go down the line. We can summarize this belief by the prior distribution $p(x)$ (shown in blue in Figure 1). We can then use Bayes rule to estimate the posterior distribution. This is the optimal combination of prior knowledge ('down the line') and new data (visual information from the ball's trajectory). Our final single best estimate of where the ball will land is then given by the maximum of the posterior density. This is known as MAP estimation (from 'Maximum a Posteriori').

As we continue to see the ball coming toward us we can refine our belief as to where we think the ball will land. This can be implemented by applying Bayes rule recursively such that our belief at time point n depends only on our belief at the previous time point, $n-1$. That is,

$$p(x_n|Y_n) = \frac{p(y_n|x_n)p(x_n|Y_{n-1})}{p(Y_n)} \quad [3]$$

where $Y_n = \{y_1, y_2, \ldots, y_n\}$ denotes all observations up to time n. Our prior belief, that is, prior to observing data point y_n is simply the posterior belief after observing all data points up to time $n-1$, $p(x_n|Y_{n-1})$. Colloquially, we say that "today's prior is yesterday's posterior."

The variable x is also referred to as a hidden variable or hidden state because it is not directly observed. If the hidden state were a discrete variable, such as whether the ball landed in or out of the service box, one can form a likelihood ratio:

$$\mathrm{LR} = \frac{p(x_n = \mathrm{IN}|Y_n)}{p(x_n = \mathrm{OUT}|Y_n)} \quad [4]$$

Decisions based on the likelihood ratio are statistically optimal in the sense of having maximum sensitivity for any given level of specificity. In contexts where LR is recursively updated, these decisions correspond to a sequential likelihood ratio test (Bogacz et al., 2006). There is a good deal of evidence showing that the firing rate of single neurons in the brain report evolving log LR values (Gold and Shadlen, 2001).

Gaussians

If our random variables x and y are normally distributed, then Bayesian inference can be implemented exactly using simple formulae. These are most easily expressed in terms of precisions, where the precision of a random variable is its inverse variance; a precision of 10 corresponds to a variance of 0.1. We first look at inference for a single variable (e.g., distance from side of tennis court).

For a Gaussian prior with mean m_0 and precision λ_0 and a Gaussian likelihood with mean m_d and precision λ_d, the posterior distribution is Gaussian with mean m and precision λ where

$$m = \frac{\lambda_0}{\lambda} m_0 + \frac{\lambda_d}{\lambda} m_d \quad [5]$$

So, precisions add and the posterior mean is the sum of the prior and data means, but each weighted by their relative precision. This relationship is illustrated in Figure 2. Though fairly simple, eqn [5] shows how to optimally combine two sources of information. As we shall see in the following section, various aspects of human behavior from sensory integration to instances of collective decision making have been shown to conform to this 'normative model.' Similar formulae exist for multivariate (instead of univariate) Gaussians (Bishop, 2006) where we have multidimensional hidden states and observations, e.g., three-dimensional position of the ball, and two-dimensional landing position on court surface.

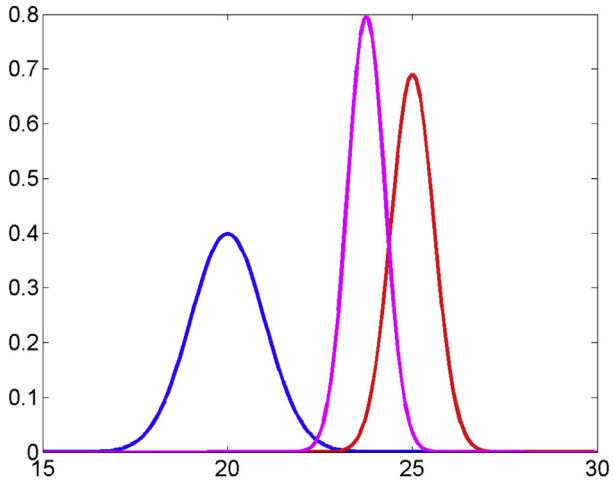

Figure 2 Bayes rule for Gaussians. For the prior $p(x)$ (blue) $m_0 = 20$, $\lambda_0 = 1$ and the likelihood $p(y|x)$ (red) $m_d = 25$, $\lambda_d = 3$, the posterior $p(x|y)$ (magenta) shows the posterior distribution with $m = 23.75$, $\lambda = 4$. The posterior is closer to the likelihood than the prior because the likelihood has higher precision. Bayes rule for Gaussians has been used to explain many behaviors from sensory integration to collective decision making.

Behavioral Models

An attractive feature of Bayesian models of behavior is that they provide descriptions of what would be optimal for a given task. They are often referred to as 'ideal observer' models because they quantify how much to update our beliefs in light of new evidence. Departures from these normative models can then be explained in terms of other constraints such as computational complexity or individual differences.

One way to address individual differences is to use an empirical Bayesian approach in which parameters of priors and their parametric forms are estimated from data. See Stocker and Simoncelli (2006) for an example of this approach in modeling visual motion processing.

What follows is a review of Bayesian models of sensory integration, visual processing, sensorimotor integration, and collective decision making. As we shall see, the priors that we have about, for example, our visual world most readily show themselves in situations of stimulus ambiguity or at low signal-to-noise ratios. Much of the phenomonology of these perceptual illusions is long established (Gregory, 1998) but Bayesian modeling provides new quantitative explanations and predictions. A more introductory review of much of this

material is available in Frith's outstanding book on mind and brain (Frith, 2007).

Sensory Integration

Ernst and Banks (2002) considered the problem of integrating information from visual and tactile (haptic) modalities. If vision v and touch t information are independent given observation of an object x, then Bayesian fusion of sensory information produces a posterior density:

$$p(x|v,t) = \frac{p(v|x)p(t|x)p(x)}{p(v,t)} \quad [6]$$

For a uniform prior $p(x)$ and for Gaussian likelihoods, the posterior will also be a Gaussian with precision λ_{vt}. From Bayes rule for Gaussians (eqn [5]), we know that precisions add:

$$\lambda_{vt} = \lambda_v + \lambda_t \quad [7]$$

where λ_v and λ_t are the precision of visual and haptic senses alone, and the posterior mean is a relative-precision weighted combination:

$$m_{vt} = w_v m_v + w_t m_t \quad [8]$$

where the weights are $w_v = \lambda_v/\lambda_{vt}$ and $w_t = \lambda_t/\lambda_{vt}$.

Ernst and Banks (2002) asked subjects which of two sequentially presented blocks was the taller. Subjects used either vision alone, touch alone or a combination of the two. They recorded the accuracy with which discrimination could be made and plotted this as a function of difference in block height. This was repeated for each modality alone and then both together. They also used various levels of noise on the visual images. From the single modality discrimination curves they then fitted cumulative Gaussian density functions, which provided estimates of the precisions λ_t and $\lambda_v(i)$ where i indexes visual noise levels.

In the dual modality experiment, the weighting of visual information predicted by Bayes' rule for the ith level of visual noise is

$$w_v(i) = \frac{\lambda_v(i)}{\lambda_v(i) + \lambda_t} \quad [9]$$

This was found to match well with the empirically observed weighting of visual information. They observed visual capture (strong weighting) at low levels of visual noise and haptic capture at high levels. Inference in this simple Bayesian model is consistent with standard signal detection theory. However, Bayesian inference is more general as it can accommodate, for example, nonuniform priors over block height.

There have been numerous studies of the potential role of Bayesian inference for integration of other senses. For example, object localization using visual and auditory cues (Alais and Burr, 2004) has supported a Bayesian integration model with vision dominating audition in most ecologically valid contexts. This visual capture is the basis of the 'ventriloquism' effect, but is rapidly degraded with visual noise. This literature has considered only simple inferences about single variables such as block height or spatial location. Nevertheless these studies have demonstrated a fundamental concept; that sensory integration is near Bayes-optimal.

The representations in the brain that are used to support the above computations are a subject of ongoing study. One proposal is that populations of neurons represent probability distributions such that higher neuronal firing rates reflect higher precisions. Moreover, if the distribution of cell activities is approximately Poisson then Bayesian inference for optimal cue integration, for example, can be implemented with simple linear combinations of neural activity (Ma et al., 2006). Recent recordings in which monkeys were trained to estimate direction of motion from visual and vestibular cues showed neuronal activity consistent with this theory (Fetsch et al., 2011).

Visual Perception

Kersten et al. (2004) reviewed the problem of visual object perception and argued that much of the ambiguity in visual processing, for example concerning occluded objects, can be resolved with prior knowledge. This idea is naturally embodied in a Bayesian framework (Knill and Richards, 1996) and has its origins in the work of Helmholtz who viewed perception as 'unconscious inference.' An example is how the inference of shape from shading is informed by a 'light-from-above' prior. This results in circular patches that are darker at the bottom being perceived as convex. The adaptability of this prior, and subsequent perceptual experience, has been demonstrated by Adams et al. (2004).

An example of such a Bayesian modeling approach is the work of Yu et al. (2009) who propose a normative model for the Eriksen Flanker task. This simple decision-making task was designed to probe neural and behavioral responses in the context of conflicting information. On each trial, three visual stimuli are presented and subjects are required to press a button depending on the identity of the central stimulus. The flanking stimuli are either congruent or incongruent. Yu et al. (2009) proposed a discrete time ideal observer model that qualitatively captured the dynamics of the decision-making process. This used the recursive form of Bayes rule in eqn [3]. In later work, a continuum time limit of this model was derived, which produced semi-analytic predictions of reaction time and error rate that accurately predicted subject behavior. They also proposed an algorithm for how these models could be implemented in the brain.

Weiss et al. (2002) proposed that many motion illusions arise from the result of Bayes-optimal processing of ecologically invalid stimuli. Their model was able to reproduce a number of psychophysical effects based on the simple assumptions that measurements are noisy and the visual system has a prior which expects slower movements to be more likely than faster ones. For example, the model could predict the direction of global motion of simple objects such as rhomboids, as a function of contrast and object shape. This model was later refined (Stocker and Simoncelli, 2006) by showing the prior to be non-Gaussian and subject specific, and that measurement noise variance was inversely proportional to visual contrast.

Najemnik and Geisler (2005) developed an ideal Bayesian observer model of visual search for a known target embedded in a natural texture. Prior beliefs in target location were updated to posterior beliefs using a likelihood term that reflected the foveated mapping properties of visual cortex. When this likelihood was matched to individual subjects discrimination

ability, the resulting visual searches were nearly optimal in terms of the median number of saccades. Later work showed that fixation statistics were also similar to the ideal observer.

An important idea to emerge in recent years is that Bayesian inference operates at the level of cortical macrocircuits (Rao and Ballard, 1999). These circuits are arranged in a hierarchy that reflects the hierarchical structure of the world around us, e.g., scenes are comprised of objects that are in turn comprised of features that are in turn comprised of simpler elements. The connections among regions have also been proposed to have a rather specific anatomy. For example, top-down connections originate from cortical pyramidal cells lying deep in the gray matter (i.e., close to white matter) whereas bottom-up connections originate more superficially (Mumford, 1992). The signals sent up and down these cortical hierarchies then map onto the message-passing algorithms that implement Bayesian inference in dynamic, hierarchical models (Friston, 2003).

If the world we perceive is the result of hierarchical processing in cortical networks then, because this processing may take some time (of the order of 100 ms), what is perceived to be the present could actually be the past. As this would be disadvantageous for the species, it has been argued that our perceptions are based on predictive models. A 50 ms delay in processing could be accommodated by estimating the state of the world 50 ms in the future. There is much experimental evidence for this view (Nijhawan, 1994). However, a purely 'predictive' account fails to accommodate recent findings in visual psychophysics. The flash-lag effect, for example, is a robust visual illusion whereby a flash and a moving object that are located in the same position are perceived to be displaced from one other. If the object stops moving at the time of the flash no such displacement is perceived. This indicates that the position of the object after the flash affects our perception of where the flash occurred. This 'postdictive' account explains the phenomenon (Eagleman and Sejnowski, 2000) and related data where the object reverses its direction at the flash time. A simple Bayesian model has been proposed to account for the activity of V4 neurons in this task (Sundberg et al., 2006) and later experimental work found evidence for a linear combination of both predictive and postdictive mechanisms.

Sensorimotor Integration

Wolpert et al. (1995) have examined the use of dynamic Bayesian models, also referred to as forward models, for sensorimotor integration. These models describe relations among a current state variable x_n, motor commands u_n, and sensory observations y_n. The state might correspond, for example, to a multivariate vector comprising a list of joint angle positions and velocities. These dynamic models have a transition density $p(x_n|x_{n-1}, u_{n-1})$ describing how the hidden state evolves over time. The posterior distribution over hidden states can then be computed as described in standard texts (Bishop, 2006). First, the dynamical equation describing state transitions is integrated to create an estimate of the next state. This requires as input a copy of the current motor command (so-called 'efference copy') and the current state. This is referred to as a time update step. A prediction of sensory input can then be made based on the predicted next state and the mapping from x_n to y_n. Finally, a measurement update or correction step can be applied, which updates the state estimate based on current sensory input.

Wolpert et al. (1995) cite a number of key features of dynamic Bayesian models including the following. First, they allow outcomes of actions to be predicted and acted upon before sensory feedback is available. This may be important for rapid movements. Second, they use efference copy to cancel the sensory effects of movement ('re-afference'), e.g., the visual world is stable despite eye movements. Third, simulation of actions allows for mental rehearsal which can potentially lead to improvements in movement accuracy.

This mathematical framework was applied to the estimation of arm position using proprioceptive feedback and a forward model based on a linear dynamical system (Wolpert et al., 1995). Inference in this model was then implemented using a Kalman filter. The resulting bias and variance in estimates of arm position were shown to closely correspond to human performance, with proprioceptive input becoming more useful later on in the movement when predictions from the forward model were less accurate.

One of the core ideas behind these forward models is that, during perceptual inference, the sensory consequences of a movement are anticipated and used to attenuate the percepts related to these sensations. This mechanism reduces the predictable component of sensory input to self-generated stimuli, thereby enhancing the salience of sensations that have an external cause. This has many intriguing consequences. For example, it predicts that self-generated forces will be perceived as weaker than externally generated forces. This prediction was confirmed in a later experiment (Shergill et al., 2003), thereby providing a neuroscientific explanation for force escalation during conflict; children trading tit-for-tat blows will often assert the other hit him harder.

Kording and Wolpert (2004) have investigated learning in the sensorimotor system using a visual reaching task in which subjects moved their finger to a target and received visual feedback. This feedback provided information about target position that had an experimentally controlled bias and variance. Subjects were found to be able to learn this mapping (from vision to location) and integrate it into their behavior in a Bayes-optimal way.

Returning to our tennis theme, an analysis of 3 years of Wimbledon games has indicated that the outcome of the current point depends on the outcome of the previous point (Klaassen and Magnus, 2001). There are multiple potential sources of correlation here. It could be that a player intermittently enjoys a sweet parameter spot where his internal sensorimotor model accurately predicts body and ball position and is able to hit the ball cleanly, or perhaps a player finds a new pattern in his opponent's behavior such as body position, or previous serve, predicting current service direction.

Collective Decision Making

Sorkin et al. (2001) have applied Bayes rule for Gaussians (see eqn [5]) in their study of collective decision making. Here the optimal integration procedure involves each group members' input to the collective decision being weighted proportionally by the member's competence at the task. Statistically, 'competence' corresponds to precision. This model

of group behavior was shown to be better than a different model which assumed members made individual decisions which were then combined into a majority vote. This latter model better described collective decision making when members did not interact.

Bahrami et al. (2010) investigated pairs of subjects (dyads) making collective perceptual decisions. Dyads with similarly sensitive subjects (similar precisions) were found to produce collective decisions that were close to optimal, but this was not the case for dyads with very different sensitivities. These observations were explained by a Bayes-optimal model under the assumption that subjects accurately communicated their confidence. This confidence sharing proved essential for the group decision to be better than the decision of the best subject.

See also: Bayesian Graphical Models and Networks; Bayesian Statistics; Bayesian Theory, History, Applications, and Contemporary Directions; Cognitive Modeling: Connectionist Approaches; Computational Neuroscience; Motor Control Models: Learning and Performance; Multisensory Integration, Principles of; Signal Detection Theory.

Bibliography

Adams, W., Graf, E., Ernst, M., 2004. Experience can change the 'light-from-above' prior. Nature Neuroscience 7, 1057–1058.
Alais, D., Burr, D., February 2004. The ventriloquist effect results from near-optimal bimodal integration. Current Biology 14 (3), 257–262.
Bahrami, B., Olsen, K., Latham, P., Roepstor, A., Rees, G., Frith, C., August 2010. Optimally interacting minds. Science 329 (5995), 1081–1085.
Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.
Bogacz, R., Brown, E., Moehlis, J., Holmes, P., Cohen, J., 2006. The physics of optimal decision making: a formal analysis of models and performance in two-alternative forced-choice tasks. Psychological Review 113 (4), 700–765.
Doya, K., Ishii, S., Pouget, A., Rao, R. (Eds.), 2007. Bayesian Brain: Probabilistic Approaches to Neural Coding. MIT Press.
Eagleman, D., Sejnowski, T., March 2000. Motion integration and postdiction invisual awareness. Science 287 (5460), 2036–2038.
Ernst, M., Banks, M., 2002. Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415 (6870), 429–433.
Fetsch, C., Pouget, A., DeAngelis, G., Angelaki, D., 2011. Neural correlates of reliability-based cue weighting during multisensory integration. Nature Neuroscience 15 (1), 146–154.
Friston, K., 2003. Learning and inference in the brain. Neural Networks 16 (9), 1325–1352.
Frith, C., 2007. Making Up the Mind: How the Brain Creates Our Mental World. Wiley-Blackwell, Malden.
Gold, J., Shadlen, M., 2001. Neural computations that underlie decisions about sensory stimuli. Trends in Cognitive Sciences 4, 10–16.
Gregory, R., 1998. Eye and Brain: The Psychology of Seeing. Oxford University Press.
Kersten, D., Mamassian, P., Yuille, A., 2004. Object perception as Bayesian inference. Annual Review of Psychology 55, 271–304.
Klaassen, F., Magnus, J., 2001. Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. Journal of the American Statistical Association 96, 500–509.
Knill, D., Richards, W., 1996. Perception as Bayesian Inference. Cambridge.
Kording, K., Wolpert, D., 2004. Bayesian integration in sensorimotor learning. Nature 427, 244–247.
Ma, W., Beck, J., Latham, P., Pouget, A., November 2006. Bayesian inference with probabilistic population codes. Nature Neuroscience 9 (11), 1432–1438.
Mumford, D., 1992. On the computational architecture of the neocortex II: the role of cortico-cortical loops. Biological Cybernetics 66, 241–251.
Najemnik, J., Geisler, W., March 2005. Optimal eye movement strategies in visual search. Nature 434 (7031), 387–391.
Nijhawan, R., 1994. Motion extrapolation in catching. Nature 370, 256–257.
Rao, R.P., Ballard, D.H., January 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive field effects. Nature Neuroscience 2 (1), 79–87.
Shergill, S., Bays, P., Frith, C., Wolpert, D., July 2003. Two eyes for an eye: the neuroscience of force escalation. Science 301 (5630), 187.
Sorkin, R., Hays, C., West, R., January 2001. Signal-detection analysis of group decision making. Psychological Review 108 (1), 183–203.
Stocker, A., Simoncelli, E., April 2006. Noise characteristics and prior expectations in human visual speed perception. Nature Neuroscience 9 (4), 578–585.
Sundberg, K., Fallah, M., Reynolds, J., February 2006. A motion-dependent distortion of retinotopy in area v4. Neuron 49 (3), 447–457.
Weiss, Y., Simoncelli, E., Adelson, E., June 2002. Motion illusions as optimal percepts. Nature Neuroscience 5 (6), 598–604.
Wolpert, D.M., Ghahramani, Z., Jordan, M.I., 1995. An internal model for sensorimotor integration. Science 269 (5232), 1880–1882.
Wolpert, D., Ghahramani, Z., 2004. Bayes rule in perception, action and cognition. In: Gregory, I.L. (Ed.), Oxford Companion to the Mind. Oxford Univ.
Yu, A., Dayan, P., Cohen, J., June 2009. Dynamics of attentional selection under conflict: toward a rational Bayesian account. Journal of Experimental Psychology Human Perception and Performance 35 (3), 700–717.

Relevant Websites

http://cnl.salk.edu/.
http://www.eecs.berkeley.edu/Faculty/Homepages/jordan.html.
http://www.fil.ion.ucl.ac.uk/~wpenny/.
http://www.gatsby.ucl.ac.uk/.
http://en.wikipedia.org/wiki/Bayesian_approaches_to_brain_function.

Bayesian Statistics

Hal S Stern, University of California Irvine, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Bayesian statistics is an approach to statistical inference that is characterized by the use of probability distributions to describe the state of knowledge about unknown quantities and the use of Bayes' theorem to update the state of knowledge to account for observed information. The basic ideas date to the late eighteenth century but broad application of the approach did not occur until computational developments in the late twentieth century allowed for simulation-based inference in a wide range of scientific applications. This article describes the key steps in carrying out a Bayesian analysis: model specification, calculation of the posterior distribution, and model assessment.

Bayesian statistics refers to an approach to statistical inference that is characterized by the use of probability distributions to describe the state of knowledge or belief about unknown quantities (e.g., parameters in a statistical model) and the updating of such distributions based on observed information (e.g., data). The approach takes its name from the use of Bayes' theorem (Bayes and Price, 1763), a mathematical result relating conditional probabilities, to perform the updating of probability distributions. As commonly practiced there are three key steps in a Bayesian analysis: specification of a probability model for the quantities of interest, calculation of the posterior distribution of unknown quantities conditional on observed quantities, and assessment of the appropriateness of the model for the questions at hand. These are discussed in detail below. The last of these topics, model assessment, is not always incorporated in introductory discussions of Bayesian statistics but it is argued below that it is an essential part of good statistical practice. Naturally, the discussion of the Bayesian approach here is brief. Additional details can be found in a wide range of modern statistical treatments including Carlin and Louis (2008), Christensen et al. (2010), and Gelman et al. (2013).

It is useful to begin with a brief history of the development of Bayesian ideas. Early work on probability focused on what might be called 'pre-data' probability calculations. Given assumptions about a random process, it is possible to assign or calculate the probability of various outcomes. For example, one can calculate the probability of obtaining an even number in a roll of a die or the probability of obtaining 10 or more heads in 20 tosses of a coin. Bayes and Laplace receive independent credit for 'inverting' probability statements in the late eighteenth century to make probability statements about parameter values given observed data values (Stigler, 1986). Thus, for example, having observed a given number of successes in a series of repeated trials, it is possible to make probability statements about the unknown probability of success in the trials. Modern Bayesian inference developed in the period around and after World War II. The name Bayesian inference replaces 'inverse probability' only in this era. Savage (1954) is an influential early book using decision theory to justify Bayesian methods. de Finetti (1974) contributed crucial work concerning the role of exchangeability, which plays a role analogous to that of independent identically distributed observations in the traditional frequentist approach to inference. Raiffa and Schlaifer (2000), Lindley (1972, 1990), and Box and Tiao (1973) contributed greatly to the popularization of the Bayesian approach. Computational advances in the late twentieth century, especially the development of fast, ubiquitous computers and the discovery (and rediscovery) of computational algorithms, made it possible to apply the Bayesian approach to a wide array of scientific fields and scientific problems. Indeed, one can now find Bayesian treatments for a variety of different application areas including the social sciences (Jackman, 2009), econometrics (Koop et al., 2007), marketing (Rossi et al., 2006), and biostatistics (Lesaffre and Lawson, 2012).

Model Specification

The basic ideas needed to describe Bayesian statistics can be easily explained in terms of some relatively simple notation. Let y denote data that is to be collected, such as a series of measurements or the outcomes of a series of test questions. Take θ to denote unknown quantities associated with the collection of the data. Often, θ represents the parameters in a probability model for the data y. This could include the mean measurement of a population of interest or the true ability of an individual test taker. There are of course many situations that are not adequately represented by this simple setup. For example, in some cases, like regression analysis, there may be data x (often called covariates) that can help explain y but are not of explicit interest themselves. Sometimes the unknown θ may include missing values (data that were intended to be collected but could not) or latent variables (additional unknowns introduced in developing a probability model) in addition to standard parameters. Here the simple notation y and θ are used throughout but it is straightforward to accommodate more complex situations like those described above. For details, see the references provided earlier.

The specification of a complete probability model for the data y requires two probability distributions: a probability distribution for the data given (or conditional) on parameters θ and a probability distribution for the parameters θ. The distribution of the data conditional on the parameters is usually denoted $p(y|\theta)$ and identified as the data distribution,

the data model, the sampling distribution, or the likelihood function. The distribution for the parameters θ, denoted $p(\theta)$, is known as the prior distribution. It can be thought of as describing the relative frequencies for different values of the model parameters before considering any of the data values. The notation above violates standard conventions in that the same letter p is used to represent two different probability distributions. It is useful for the purposes discussed here, however, in that it avoids the introduction of other letters that are not central to this discussion. The notation above is simple but is capable of accommodating the full range of Bayesian methods. For example, it is sometimes desirable to avoid specific parametric forms for one or both of the component distributions. The development of so-called nonparametric Bayesian methods has made this possible; in practice, they are still of the form $p(y|\theta)$ and/or $p(\theta)$ but rely on high-dimensional θ to avoid being overly restrictive in the choice of the distribution $p(.)$.

A simple example helps to show how the basic notation works. Suppose that a student is asked to answer a number of mathematics questions believed to measure his or her knowledge of algebra. Assuming there are n questions believed to be of roughly the same level of difficulty, it may be natural to let the number of correct responses y follow a binomial distribution with parameter θ measuring the probability that the student answers a question correctly (which can be taken as a measure of their knowledge of algebra). Then $p(y|\theta)$ is the binomial probability of getting exactly y correct responses, $\binom{n}{y}\theta^y(1-\theta)^{n-y}$. The prior distribution $p(\theta)$ should reflect any information about θ that is available prior to collecting the data. This choice is discussed further next.

For many users of statistical methods, the largest question about the use of Bayesian methods concerns the choice of prior distribution. There are a number of approaches to this choice. The subjective approach to choosing a prior distribution is to take the prior distribution to be an honest assessment of prior beliefs about the values of the model parameters. Although people are sometimes hesitant to supply such subjective prior distributions, it is often the case that some prior information is available. For the example discussed, a great deal may not be known about the specific student but previous experience may suggest that typically students in the class are able to answer between 60 and 90% of such questions correctly. This may be enough to suggest a prior distribution such as a beta distribution (appropriate for quantities that take values between 0 and 1) with parameters 3 and 1, which takes the form $p(\theta) = 3\theta^2$. This beta distribution has a mean of 0.75 with about half of the probability in the interval from 0.60 to 0.90. If it is possible to specify a prior distribution, like the beta distribution suggested here, then the Bayesian paradigm provides the way to update that prior information given observed data. The updating is discussed below.

Occasionally, the choice of a prior distribution is made primarily because of considerations of computational convenience. Conjugate families are families of prior distributions that combine with a given data distribution to produce posterior distributions in the same family as the prior distribution. The beta distribution developed above is an example; it is the conjugate prior distribution for binomial data. Conjugate prior distributions are capable of supporting a variety of prior opinions (e.g., by making different choices within the conjugate family) but of course not all.

The desire to avoid using subjective prior information and/or arbitrary distributional forms has led to research on using vague or 'noninformative' prior distributions. Formally, a vague prior distribution is one that assigns roughly equal probability to a wide range of possible values. In the binomial example, it might be asserted that there is no information at all about θ other than that it is between 0 and 1 and thus a uniform distribution between 0 and 1 may be chosen. (In fact, this is a particular form of the beta distribution.) The idea is that a vague prior distribution will not have a strong influence on the results and that the data will play the central role in determining the conclusions. In this way, the use of vague or 'noninformative' prior distributions represents the pursuit of a form of objective Bayesian inference. One difficulty is that, in the limit, especially for parameters that are not known to lie in a fixed range, vague prior distributions may become so vague as to no longer be proper distributions (they may not integrate to one!). It is 'legal' to use such improper prior distributions as long as it is verified that the resulting posterior distribution is a proper distribution. Improper prior distributions are popular because sometimes they appear to be 'noninformative' in the sense that they rely only on the likelihood. It is probably best to think of improper prior distributions as approximations to real prior distributions. If the improper prior distribution leads to a proper posterior distribution and sufficiently accurate conclusions, then one might use the results from this analysis without needing to work any harder to select a more appropriate prior distribution.

Some people find the need to explicitly choose a prior distribution problematic. For these individuals, it is disturbing that individuals with different prior distributions will obtain different results. There are a number of important points to make here. First, it is important to recognize that situations like this occur every day in the real world when individuals with different information make different decisions from the same data. Second, it is possible to show mathematically that the prior distribution does not have a large impact on the results of the statistical analysis when a great deal of data has been collected. This again matches everyday experiences where a large group of people may all come to agree on a course of action given a large amount of evidence. Perhaps the most important point to make is that prior information is not necessarily as rare as might be thought. This is especially true in problems that are more sophisticated than the simple math test. Consider a situation in which the same set of algebra questions are given to a number of different individuals. Each individual has his or her own probability of getting a correct answer. Although it may not be easy to select a prior distribution for each individual, it may be natural to treat the set of parameters as a sample from a common population. This type of model is known as a hierarchical model (*see* Hierarchical Models: Random and Fixed Effects) because the prior distribution itself may depend on additional parameters (e.g., the mean and variance of algebra knowledge parameters in the population). It is the success of Bayesian methods in multiparameter problems of this type that has led to the increased use of such methods.

Calculation of the Posterior Distribution

The model specification makes it clear that there are two sources of information about θ. There is information about θ in the prior distribution and there is information about θ in the data distribution. This information can be combined using the laws of probability, namely, Bayes' theorem, to obtain the updated probability distribution on θ after observing data y. The latter is known as the posterior distribution of θ and denoted $p(\theta|y)$. Bayes' theorem is

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y) = p(y|\theta)p(\theta) \Big/ \int p(y|\theta)p(\theta)\mathrm{d}\theta$$

The denominator of Bayes' theorem is known as the marginal probability of observing the data y. Bayes' theorem allows one to update his or her opinion about the possible values of θ so that values of θ that are most consistent with the observed data (i.e., values that have large values of the likelihood) have more weight in the posterior distribution then they do in the prior distribution.

The posterior distribution is fundamental to Bayesian inference for θ. The Bayesian approach makes probability statements concerning the unknown quantities after fixing the things that have been observed (i.e., the data) at their known values. Note that Bayesian methods differ from traditional methods, which are justified by repeated sampling considerations – only the sample at hand is considered relevant in the Bayesian approach.

The posterior distribution describes the current state of knowledge about θ after observing the data y. It is possible to draw a wide range of inferential conclusions from the posterior distribution. A point estimate is a one-number summary of the posterior distribution. Of course, a single estimate does not describe the full state of knowledge. To obtain a point estimate, it is necessary to specify a loss function that makes explicit the loss that is expected to be incurred because of errors in the point estimate. Given a loss function, it is possible to derive the optimal point estimate as the value that minimizes the expected loss (computed from the posterior distribution). An interval estimate specifies a range of plausible values for a parameter. The posterior distribution allows one to identify posterior intervals that contain a given parameter with any specified probability. Such intervals are analogous to standard confidence intervals. One difference is that the posterior intervals have a natural probabilistic interpretation in that they summarize the posterior state of knowledge regarding plausible values of the parameter. By contrast, the probability statements underlying traditional confidence intervals depend on behavior in a large number of repeated samples. In traditional inference, another common form of statistical inference is a hypothesis test about the value of a parameter or a set of parameters (e.g., it is common to test whether two means are equal). The information about such tests is equivalent to the information contained in posterior intervals, i.e., it can be addressed whether two means are plausibly the same by examining the posterior interval for the difference between the two means. Tests of some hypotheses are also possible using the Bayes factor, a ratio that measures the degree to which two competing models explain the observed data. This is discussed further below.

In relatively simple problems it is possible to analytically determine the form of the posterior distribution and subsequently obtain statistical inferences. For the example described earlier in which a binomial random variable measuring the number of correct responses y in n questions was observed and a conjugate beta prior distribution was used (with parameters 3 and 1), it turns out that the posterior distribution is also a beta distribution. The posterior distribution is a beta distribution with parameters $y + 3$ and $n - y + 1$. The form of the posterior distribution in this case shows explicitly how prior information combines with the data. The mean of this posterior distribution is $(y + 3)/(n + 4)$, which is a compromise between the students performance on the test y/n and the mean of the prior distribution $3/4 = 0.75$. For example, a student who answered 9 of 10 questions correctly would have posterior mean 0.86. Posterior intervals can also be easily obtained if the form of the posterior distribution is known.

In many realistic models, especially those containing a large number of parameters, it can be quite difficult to determine the form of the posterior distribution analytically. It is sometimes possible in such cases to develop a useful approximation to the posterior distribution, e.g., a normal approximation, but this too is difficult in many realistic problems. For many years, the difficulty in calculating the posterior distribution kept people from applying Bayesian methods. The wide availability of high-speed computing, beginning in the late twentieth century, made it possible to develop general-purpose simulation algorithms that allowed researchers to approximate the posterior distribution by generating a sample of simulations from the posterior distribution. The most common class of algorithms are known as Markov chain Monte Carlo (MCMC) algorithms; these algorithms work by building and then simulating a Markov chain, which has the posterior distribution as its stationary distribution. If the Markov chain is simulated for a long time, then the draws from the simulation can be treated as draws from the posterior distribution. Although there can be a number of challenges associated with Bayesian computation (e.g., developing efficient algorithms, judging when the simulation has run long enough to approximate the posterior distribution), there is little doubt that it has allowed researchers to apply Bayesian methods in diverse settings and with large data sets. Recent references describing MCMC algorithms and other relevant computational strategies include Chen et al. (2001), Liu (2008), and Brooks et al. (2011).

Simulation-based inference is a powerful tool. A set of simulations from the posterior distribution provides an empirical approximation to the posterior distribution. Any desired inference is then estimated using this empirical approximation. For example, the posterior mean is estimated by the sample mean of the simulations and a posterior interval for a parameter is obtained using the relevant percentiles of the simulations. Although the simulations are in fact providing estimates of the various quantities of interest, it is generally possible to obtain enough simulations to reduce the Monte Carlo or simulation error to acceptable levels. Simulations provide a great deal of flexibility – for example, it is

quite straightforward to estimate the posterior probability that a given parameter is positive since one needs only to calculate the proportion of simulations in which it is positive.

Model Assessment

It is important when performing statistical inference based on a probability model to examine the validity of the assumptions made in the model. For the Bayesian approach to inference, the assumptions include the form of the prior distribution, the form of the data model, and any other assumptions embodied in the data model (independence, constant variance, etc.). There are a wide range of activities that are useful for assessing the Bayesian model being used in a given situation. These include sensitivity analysis, model checking, and model comparison. Brief descriptions of these approaches are provided here. The general references provided earlier give additional details.

Sensitivity Analysis

The prior distribution is an important part of the Bayesian model. One can argue that there is no need to check the prior distribution in that as long as it is an honest reflection of the prior state of knowledge about unknown quantities, the Bayesian machinery will provide the appropriate updated posterior state of knowledge. At the same time, it is important to be able to describe whether the inferences are likely to change under other reasonable prior distributions. This suggests a form of sensitivity analysis where the prior distribution (or other aspect of the model) is varied and the effect on the posterior distribution is assessed. The conclusions of a Bayesian analysis are strengthened if it is found that key posterior inferences (e.g., the sign of an important parameter) are unchanged under other plausible model specifications.

Model Checking

In regression analysis, it is common to check the form of the model by examining residuals. The residuals are functions of the data that are expected to have no particular structure if the model is correct. Evidence of structure in the residuals suggests a possible problem with the model. The same basic strategy can be applied more generally to assess the fit of a Bayesian model or analysis. A function of the data (and perhaps parameters) can be defined that is expected not to have any structure if the model is correct. Examining the posterior predictive distribution of such functions (i.e., the distribution of the function in future data sets that would be expected if the model is correct) can allow one to detect difficulties in the model. These are known as posterior predictive model checks.

Model Comparison

In some applications, there are a number of models that are contemplated. In this case, researchers may want to compare or combine the models. Model comparison is frequently carried out using Bayes factors (*see* Model Selection and Model Averaging). It can also be carried out using predictive measures (e.g., cross-validation) or information criteria. There are several important points here. If model selection is used to identify a best-fitting model, it is still important to use model checking and sensitivity analysis as described above to ensure that the selected model is appropriate for the questions of interest. Also, one can argue that there are alternatives to model comparison and model selection that make more sense in the Bayesian framework. It may be possible to embed one model into a bigger family of models and then average over the entire family. For example, a normal model might be replaced by the Student-t family of models that includes the normal distribution as a limiting case. Then the posterior distribution for the degrees of freedom in the Student-t model provides information about whether the normal distribution is an appropriate choice.

Summary

The Bayesian approach to statistics outlined here is now used extensively in a wide range of application areas, including the social sciences. Historically, there has been controversy about the use of Bayesian methods. This remains true to a limited extent (Efron, 1986). The arguments favoring the use of the Bayesian approach for data analysis follow from its use of probability distributions to describe uncertainty about unknown quantities. There is a natural probability-based interpretation for Bayesian results (e.g., interval estimates) and great flexibility in the types of inferences that can be obtained (e.g., one can easily obtain a posterior distribution on the ranks of a set of parameters). In addition, the framework is extremely powerful in its ability to accommodate increasing degrees of complexity in a data analysis. For example, accommodating missing data values is straightforward in the Bayesian framework because the missing values become unknown random variables. In the same vein, censoring or truncation of observed values are easily incorporated through modifications of the underlying probability model. The reliance on formal probability distributions also means that it is possible to draw valid Bayesian inferences in finite samples without relying on large sample results.

See also: Bayesian Graphical Models and Networks; Bayesian Models in Neuroscience; Bayesian Theory, History, Applications, and Contemporary Directions.

Bibliography

Bayes, T., Price, R., 1763. An essay towards solving a problem in the doctrine of chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S. Philosophical Transactions of the Royal Society of London 53, 370–418. http://dx.doi.org/10.1098/rstl.1763.005.

Box, G.E.P., Tiao, G.C., 1973. Bayesian Inference in Statistical Analysis. John Wiley and Sons, New York.

Brooks, S., Gelman, A., Jones, G., Meng, X.-L. (Eds.), 2011. Handbook of Markov Chain Monte Carlo. CRC/Chapman and Hall, London.

Carlin, B.P., Louis, T.A., 2008. Bayes Methods for Data Analysis, third ed. CRC/Chapman and Hall, London.

Chen, M.-H., Shao, Q.-M., Ibrahim, J.G., 2001. Monte Carlo Methods in Bayesian Computation. Springer, New York.
Christensen, R., Johnson, W.O., Branscum, A.J., Hanson, T.E., 2010. Bayesian Ideas and Data Analysis. CRC/Chapman and Hall, London.
Efron, B., 1986. Why isn't everyone a Bayesian? The American Statistician 40, 1–11 (including discussion).
de Finetti, B., 1974. Theory of Probability. Wiley, New York.
Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis, third ed. CRC/Chapman and Hall, London.
Jackman, S., 2009. Bayesian Analysis for the Social Sciences. John Wiley, New York.
Koop, G., Poirier, D.J., Tobias, J.L., 2007. Bayesian Econometric. John Wiley, New York.
Lesaffre, E., Lawson, A., 2012. Bayesian Biostatistics. John Wiley, New York.
Lindley, D.V., 1972. Bayesian Statistics, a Review. Society for Industrial and Applied Mathematics, Philadelphia, PA.
Lindley, D.V., 1990. The 1988 Wald Memorial Lectures: the present position of bayesian statistics. Statistical Science 5, 44–89 (including discussion).
Liu, J.S., 2008. Monte Carlo Strategies in Scientific Computing, second ed. Springer, New York.
Raiffa, H., Schlaifer, R., 2000. Applied Statistical Decision Theory. Wiley-Interscience, New York.
Rossi, P.E., Allenby, G.M., McCulloch, R., 2006. Bayesian Statistics and Marketing. John Wiley, New York.
Savage, L.J., 1954. The Foundations of Statistics. Dover, New York.
Stigler, S.M., 1986. The History of Statistics. Harvard University Press, Cambridge, MA.

Bayesian Theory, History, Applications, and Contemporary Directions

Scott M Lynch, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D.A. Schum, volume 2, pp. 1056–1060, © 2001, Elsevier Ltd.

Abstract

Bayes' theorem is a simple method for reversing conditional probabilities in straightforward probability problems. The theorem's application was expanded to become a full-fledged paradigm of statistics that provides a coherent, but subjective, method for updating scientific knowledge. For more than a century, Bayesian methodology was controversial in academia but used extensively in practical applications in the military and business. Improved computing capabilities, coupled with advances in estimation methods and increased applicability to contemporary statistical problems have made it a mainstream methodology. In recent years, Bayesian applications have become too numerous to discuss in detail. Instead, general directions are discussed.

Introduction

Bayesian statistics is an approach to – or paradigm of – statistics that involves updating prior beliefs with new data, or more succinctly, 'learning' from data. Contemporary Bayesian statistics takes a subjective view of probability; that is, Bayesians argue that probability represents uncertainty, and therefore the language of probability can be used to represent any quantity that is unknown, including model (or distribution) parameters, even if such a quantity is, in fact, fixed. As more data are gathered, we can update our knowledge of the quantity of interest until, in theory, we eventually can learn its true (or current) value. This perspective on probability contrasts with the more classical view that data are random realizations of a process governed by a fixed parameter, that probabilistic language can only represent random quantities, and so probability distributions cannot be used to represent parameters.

Versions of Bayes' Theorem

Bayesian statistics begins with Bayes' theorem, which is simply a rule for reversing conditional probabilities. Using discrete sample spaces, Bayes' rule can be represented as:

$$p(\mathrm{B}|\mathrm{A}) = \frac{p(\mathrm{A}|\mathrm{B})p(\mathrm{B})}{p(\mathrm{A})} \qquad [1]$$

In this equation, $p(\mathrm{B}|\mathrm{A})$ is the probability of event B, given that we know event A has occurred, $p(\mathrm{A}|\mathrm{B})$ is the probability that event A occurs given that we know event B occurred, $p(\mathrm{B})$ is the probability that event B occurs, and $p(\mathrm{A})$ is the total probability of event A. In a simple case in which there are two possible states of the world, B and not B, $p(\mathrm{A}) = p(\mathrm{A}|\mathrm{B})p(\mathrm{B}) + p(\mathrm{A}|\neg\mathrm{B})\,p(\neg\mathrm{B})$. In the more general case in which there are multiple possible states of the world,

$$p(\mathrm{A}) = \sum p(\mathrm{A}|\mathrm{B}_i)p(\mathrm{B}_i), \qquad [2]$$

where B_i indicates one of the possible states of the world, $p(\mathrm{B}_i)$ is its probability of occurrence, and $\Sigma p(\mathrm{B}_i) = 1$.

As a contrived, but concrete example of the application of Bayes' rule, suppose I am interested in computing the probability that a weaker-than-average student cheated off an 'A' student on a weekly quiz, given that he received an A on the quiz. Suppose that I have given a total of 10 quizzes prior to the current one, and this average student has made an A on only two of them. Also suppose that, if the student cheats, there is a probability of .9 that he will obtain an A. Finally, suppose I thought, prior to the exam, that there is only about a 10% chance that the student would cheat. So, 'B' is 'cheated' and 'A' is 'received an A on the exam.' The table below shows the information we have for computing the desired probability.

Term in Bayes' rule	*Meaning*	*Probability*
$p(\mathrm{A}\|\mathrm{B})$	Probability of earning an A if cheated	.9
$p(\mathrm{B})$	Prior probability of cheating	.1
$p(\mathrm{A}\|\neg\mathrm{B})$	Probability of earning an A if no cheating	.2
$p(\neg\mathrm{B})$	Probability of no cheating	.9

Substituting these quantities into Bayes' rule gives:

$$p(\mathrm{B}|\mathrm{A}) = \frac{(.9)(.1)}{(.9)(.1) + (.2)(.9)} = .333.$$

Thus, before this exam, my assumed 'prior' probability that the student would cheat was 1/10; after he received an A on the most recent exam, I think there is a probability of 1/3 that he cheated (this is called a 'posterior probability'). Why did my 'prior' probability increase so drastically? The reason is that, for that student, obtaining an A on an exam is uncharacteristic: based on previous exams, the probability that he would obtain an A is only .2. In contrast, the probability he would obtain an A if he cheated is rather high. Thus, the A provides more evidence for the position that he cheated than the position that he did not.

Bayes' theorem is easily proven, involving simply a double application of the joint probability rule: $p(\mathrm{A},\mathrm{B}) = p(\mathrm{A}|\mathrm{B})p(\mathrm{B})$. Given that $p(\mathrm{B},\mathrm{A}) = p(\mathrm{A},\mathrm{B})$, $p(\mathrm{B}|\mathrm{A})p(\mathrm{A}) = p(\mathrm{A}|\mathrm{B})p(\mathrm{B})$. Dividing both sides of this latter equality by $p(\mathrm{A})$ yields the theorem. The mathematics underlying the theorem is therefore indisputable. However, the theorem's use as a method for representing subjective beliefs and updating them via probabilistic language is controversial, and this controversy permeates the history of Bayesian statistics, as discussed below.

A Brief History

Bayes' theorem was developed by Reverend Thomas Bayes in the mid-1700s, albeit not in its modern symbolic representation. Bayes' initial motivation involved a thought experiment in which a billiard ball is thrown onto a table and its location marked. Then, another ball is thrown n times onto the table, and a record is kept reflecting the proportion of times the ball lands on each side of the original ball. The question is: how does each additional ball's position (i.e., the acquisition of new data) update our knowledge of where the initial ball landed? Bayes' paper on the topic was published posthumously in 1763 (Bayes died in 1761) by his friend and fellow Unitarian minister, Richard Price. Approximately half a century after Bayes wrote his seminal paper, Laplace rediscovered the theorem. From that time until more than a century later, the theorem was commonly known as 'inverse probability' in reference to the theorem's key contribution: the reversal of a conditional probability (de Finetti, 1974; McGrayne, 2011).

This theorem for point probabilities forms the basis of Bayesian statistics, but Bayesian statistics involves extending the components of the original theorem to handle more than single, point probabilities like that demonstrated in the previous section. First, the 'data' in the above example was the student's performance on a single quiz. In fact, however, we may have a series of quiz grades with which to update our prior belief that he or she cheats. In that case, the first term in Bayes' theorem may be replaced with a likelihood function – a joint distribution for quiz grades – rather than a single quiz grade. Second, depending on how well we know the student's history, say, in other courses, we would probably not be confident enough to assign an exact value of .1 as the prior probability of cheating. Instead, we may place a distribution on this prior probability that gives a (potentially) differential weight to different prior probabilities of cheating. For example, we may believe a priori that the student's probability of cheating is about .1, give or take a few percent, with higher and lower probabilities of cheating being less likely. The prior distribution, then, may be bell-shaped with a peak centered over the value .1. Bayes' theorem can therefore be extended to continuous sample spaces and probability distributions as:

$$f(\theta|\gamma) = \frac{f(\gamma|\theta)\, f(\theta)}{f(\gamma)}. \qquad [3]$$

In this equation, $f(\theta|\gamma)$ is the 'posterior' *distribution* of θ, given the data γ. $f(\gamma|\theta)$ is the *distribution* for γ, given θ – the likelihood function in classical terminology, $f(\theta)$ is the 'prior' *distribution* for θ, and $f(\gamma)$ is the marginal distribution for γ. Here, $f(\gamma) = \int_S f(\gamma|\theta) f(\theta) d(\theta)$, where S is the sample space for θ. This integral is simply the continuous analog to the summation presented in eqn [2]. It is worth reiterating that the product of two distributions – one for the prior and one for the likelihood – yields a distribution for θ, rather than a single point estimate for θ, and this distribution represents our updated level of knowledge (and uncertainty) about θ.

More important than the shift from discrete point probabilities to continuous probability distributions is that this representation of the theorem includes both data (γ) and probability distribution parameters (θ) rather than data alone. Assigning probability distributions to parameters essentially represents the degree of uncertainty we have regarding theoretical hypotheses about the nature of some phenomenon. Consider the example above. The 'prior' probability of .1 that the student cheated is essentially just a 'guess' about the student's propensity to cheat prior to my having observed any data. This guess may be informed by my previous experience with students or my experience with this particular student. But it is essentially a hypothesis. The posterior probability that the student cheated is a revision of that hypothesis that takes into account the new data – the A on the most recent test. Thus, we could rewrite Bayes' theorem as:

$$f(\text{Hypothesis}|\text{Data}) = \frac{f(\text{Data}|\text{Hypothesis})\, f(\text{Hypothesis})}{f(\text{Data})}, \qquad [4]$$

where f(Data) is the marginal probability distribution for the data under all possible hypotheses. Thus, this formulation of Bayes' theorem suggests that we can evaluate the probability of a hypothesis conditional on observed data. Furthermore, given that f(Data) generally contains no information regarding the hypothesis that is not contained in the numerator and is simply a normalizing constant that makes the posterior distribution a proper probability distribution, the theorem is often reduced to:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}, \qquad [5]$$

where '$\propto$' means 'is proportional to.'

This specification for Bayes' theorem has invoked the greatest criticism historically and relegated Bayesian statistics to a marginal status for decades, if not centuries. From a 'classical' statistics perspective, a parameter is a fixed, but generally unknown quantity, and therefore it makes no sense to place a probability distribution on a parameter (and therefore a scientific hypothesis). From that view, only data are random: a given sample is a stochastic realization from a distribution with fixed parameters. Under that approach, hypotheses about the true state of the world are assessed by (1) posing the hypothesis and (2) evaluating the likelihood of observing the sample data we in fact observed. If the data are unlikely to have occurred if the hypothesis were true, we conclude that the hypothesis is improbable (Fisher, 1960). Another way to view the classical criticism is that hypotheses and data cannot share the same sample space, and so it makes no sense to talk about the joint distribution of hypotheses and data, a condition that is required to derive Bayes' rule (Edwards, 1972).

It is important to note that probabilistic reasoning is a deductive process. The classical approach to statistics maintains this approach by evaluating the probability of data under a given hypothesis regarding the parameters' value. The Bayesian approach, in placing a probability distribution on parameters and updating it with data, turns the statistical process into an inductive one, violating historical views of the scientific method (Fisher, 1960).

Bayesians historically have had a number of responses to this criticism, ranging from the philosophical to the pragmatic. First, and perhaps foremost, Bayesian statisticians take a completely different view than classical statisticians on the nature of probability, and the difference between the classical and Bayesian view of probability has been, and continues to be, discussed at

the very onset of the introduction of Bayesian thinking in almost any text (Box and Tiao, 1973; Gill, 2002; Lynch, 2007; Robert, 1994). Specifically, Bayesians argue that, although a parameter of a distribution may, in fact, be a fixed quantity, the fact that we do not know its true value justifies placing a probability distribution on it, or otherwise discussing it using the language of probability, and we do so in almost every real-world scenario. For example, when we discuss whether it will rain tomorrow, the fact is that it either will or will not rain, and we will know the 'parameters' of the world in that case 2 days from now. However, this does not prevent us from discussing today whether it will rain tomorrow in probabilistic terms. It is commonly understood, in that context, that the language of probability is used to represent uncertainty about a state of the world, regardless of whether that state can be known exactly in the future. Thus, in science, it is perfectly reasonable, even if we will eventually know a parameter's value with certainty in the future, that we use the language of probability to summarize our current knowledge of it.

Second, it often does not matter whether we treat parameters of distributions as fixed or random quantities. Put another way, the conclusions that we reach regarding a scientific hypothesis are often substantively the same, regardless of whether we take a classical or Bayesian approach. To be sure, whether conclusions are similar may depend on our choice of prior distribution for θ. Indeed, a long-standing criticism of the Bayesian approach has been that, even if one agrees with a subjective view of probability – that probability is a reflection of uncertainty – that subjectivity can carry over into the process of constructing prior distributions and thereby distort any conclusion that data alone may produce. For example, eqn [3] shows that the posterior distribution for a parameter is essentially a weighted average of the prior distribution for the parameter and the data (this is known as Bayesian 'shrinkage'). Yet, many, if not most Bayesian applications have historically used 'noninformative' (also called 'flat,' 'reference,' and 'uniform') priors specifically with the intent of avoiding injecting much if any prior knowledge into a model and thereby overruling the data in the posterior distribution (Gill, 2002; Gelman et al., 1995; Jeffreys, 1961).

Third, in a less defensive and more proactive manner, Bayesians argue that the Bayesian paradigm is superior to the classical paradigm, because it is, in fact, the way science progresses in a coherent fashion. The classical approach does not allow for the formal accumulation of knowledge; instead, the same ('null') hypotheses are tested again and again and knowledge accumulates only via literature reviews in scientific papers. This approach does not formalize the advance of science. In contrast, under the Bayesian approach, the current state of knowledge regarding a scientific hypothesis is contained in the prior distribution, and our knowledge is updated with each successive study. Although a prior study may begin as subjective construct, scientific consensus eventually molds the prior into an objective claim about reality.

Since the 1990s, Bayesian statistics has made considerable headway into mainstream statistics and science, because many statistical and scientific problems are very amenable to the Bayesian approach. For example, in recent decades, computer science has found the Bayesian approach to be very useful in machine learning and artificial intelligence more generally. Similarly, in social science, the large-scale collection of data following a hierarchical structure, either in terms of spatial hierarchies like students nested within schools, communities, states, or repeated measures on individuals collected over time (i.e., panel data), has made Bayesian analysis increasingly popular because of its ability to handle such hierarchical structuring easily. In short, in both 'hard' and 'soft' sciences, the explosion of computing power has enabled the use of Bayesian methods in areas in which it was not possible to use earlier and for which the classical approach was ill suited.

Applications Prior to 1990

Bayesian statistics held a relatively marginal status in statistics and science more generally prior to 1990, but there were numerous applications, nonetheless, many of them outside of the world of statistics itself. In general, the term 'Bayesian statistics' was not used in early applications. Instead, Laplace's term 'inverse probability' was preferred. After Bayes and Laplace, the theorem essentially disappeared – with much criticism – for roughly half a century. Near the turn of the twentieth century, the theorem began to reemerge in practical, applied (but not theoretical) settings. For example, the French army, under Joseph Bertrand used Bayes' theorem to update their artillery positioning (i.e., reaim it), and the mathematician Poincare used Bayes' theorem in a defense of Alfred Dreyfus in the 'Dreyfus affair' (McGrayne, 2011). In the US, Bayes' theorem found use in business, both in communications and the insurance industry. However, Bayes' theorem remained largely out of favor in academia, in part due to long-standing criticism by Fisher and Neyman who had developed their own approach (now called the 'classical' approach) to addressing scientific–statistical hypotheses. To be fair, however, Jeffreys, working at the same time as Fisher, kept Bayesianism alive.

During World War II (WWII), Bayes' theorem emerged in several areas. Turing used Bayesian logic to solve the 'enigma' code, and others used Bayesian logic to help the Allies to locate U Boats, and thereby protect commercial and military ships. Following WWII, Bayesian approaches continued to be used in private industry and expanded into medical research, albeit without the 'Bayesian' label (McGrayne, 2011). Additionally, Bayesian statistics reemerged in academia, with a key book by Savage published in 1954 (Savage, 1972).

The period between the 1950s and the 1980s saw only a slight increase in the use of Bayesian statistics in academia, and the stagnation was due to two key factors. First, *within* the Bayesian 'community' there was a disagreement over the nature of prior distributions. Specifically, some argued for an explicit subjective view of probability and prior distributions, while others argued for more objective, less informative priors like Jeffreys' priors.

Second, because Bayesian analysis requires heavy computation to either approximate integrals or compute them via 'brute force,' limitations of computing power restricted realistic and large-scale Bayesian applications. The evolution of personal computers and the accompanying explosion in computing power in the 1980s enabled more realistic Bayesian applications to be undertaken.

In short, the application of Bayesian thinking and methods, while spanning more than two centuries, has been marked by general disdain with periodic revivals in academia. Outside of academia, the Bayesian approach has found a more steady application, but generally under the guise of 'inverse probability' and not under the 'Bayesian' name (McGrayne, 2011; for a very detailed, nonmathematic history of Bayesian statistics). In the 1990s, the tide turned.

Since 1990

The number of Bayesian applications exploded in the 1990s, and so I limit the discussion here to the key issues that enabled this explosion and the direction of Bayesian research since then. Prior to 1990, solutions to most Bayesian applications were analytically intractable and required either computer-based numerical methods approaches and/or approximations based on normal distributions and normal mixture distributions. These approaches required computational and other skills that are specialized knowledge, especially prior to 1990, and so Bayesian analysis was restricted to those who both appreciated the paradigm and could implement it. Additionally, the use of approximation methods obviated one of the key benefits of Bayesian analysis: the ability to conduct more realistic analyses that did not require simplifying assumptions and appeals to asymptotic normality.

In 1990, Gelfand and Smith published a truly seminal paper introducing sampling methods as an approach useful for constructing summaries of posterior distributions. The logic behind sampling methods is that summaries of samples from a distribution can provide equivalent information as analytic summaries of a distribution. For example, sample quantiles are asymptotically equivalent to what is obtained via integral calculus or numerical integration (Gelfand and Smith, 1990).

The difficulty with simulation methods is that it is difficult to sample from distributions of high dimensional and/or unknown forms. Gelfand and Smith showed, perhaps most important, that high dimensional problems can be broken into lower dimensional, conditional ones, and repeated, sequential sampling from conditional distributions produce samples that, for all intents and purposes, can be treated as samples from the joint, multivariate distribution. These samples can then be summarized using basic descriptive methods, just as one would summarize sample data. This approach is called 'Gibbs sampling' and is one of a number of methods of 'Markov chain Monte Carlo' (MCMC) simulation methods.

A more general sampling approach – the Metropolis–Hastings (MH) algorithm – emerged subsequently as an alternative to Gibbs sampling. To be sure, the MH algorithm predates Gelfand and Smith's paper by more than three decades, but its application prior to the 1990s was limited to the applications in physics. Following the publication of Gelfand and Smith's paper, it was shown that Gibbs sampling is a special case of MH sampling. Together, MH and Gibbs sampling have been used in literally thousands of applications, leading to a tremendous increase in the application of Bayesian statistics.

Importantly, in 1996, an edited volume, entitled *Markov Chain Monte Carlo in Practice*, was published. This volume not only illustrated the use of MCMC methods for a variety of substantive problems, but also presented methodological advances in MCMC computation, including evaluation of model fit and model comparison, foreshadowing subsequent advances in Bayesian statistics. That is, since the mid-1990s, most advances in Bayesian statistics have been in terms of (1) improvement in MCMC design and implementation for various, often idiosyncratic applications, and (2) model evaluation and comparison. Regarding both areas, the free package WinBugs has played a key role in helping to routinize the application of Bayesian statistics (Lunn et al., 2000), enabling the number of scholarly papers adopting a Bayesian approach to explode.

Since the mid-1990s, the number of applications of Bayesian methods is too numerous to detail. From January 1971 to January 1991, a 'Web of Science' search found that there were 2160 papers involving the key word 'Bayesian.' Over the following 20 years, there were 40 242. Thus, publications involving Bayesian statistics – by this cursory measure – increased by a factor of roughly 19. To put this in context, given that the number of scientific journals and publications therein in general increased over that time period, I also obtained counts of articles for several other key words, including 'tornado,' 'beef,' 'fish,' and 'climate.' For 'tornado,' citations increased by a factor of 2. For 'beef' and 'fish,' citations increased by a factor of 3 and 7, respectively. Finally, for 'climate,' citations increased by a factor of 21. In sum, the growth in the key word 'Bayesian' has kept pace almost with the growth in the key word 'climate,' an impressive feat, given that interest in climate change has increased tremendously over the last 20 years.

Conclusion

In conclusion, Bayesian statistics has a long, storied history. For the better part of two centuries, Bayesian statistics was mired in theoretical disputes that rendered its use in academia negligible. However, its practicality was recognized outside of academia in several areas in which updating knowledge systematically was (and is) critically important. In recent decades, expansion in computing capability has led to the widespread adoption of Bayesian methods, and its growth in use shows no deceleration. Indeed, the Bayesian approach may become the dominant paradigm in statistics within another generation.

See also: Computational Approaches to Model Evaluation; Probability Density Estimation; Probability Theory: Formal; Probability Theory: Interpretations.

Bibliography

Bayes, T., 1763. An essay toward solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society 53, 370–418.

Box, G.E.B., Tiao, G.C., 1973. Bayesian Inference in Statistical Analysis. Addison-Wesley, Reading, MA.

Edwards, A.W.F., 1972. Likelihood. Cambridge University Press, Cambridge, MA.

Finetti, De, 1974. Theory of Probability, vols. 1 and 2. John Wiley & Sons, NY.

Fisher, R.A., 1960. Statistical Methods and Scientific Inference. Oliver and Boyd, Edinburgh.
Gelfand, A.E., Smith, A.F.M., 1990. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85, 398–409.
Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman & Hall, London.
Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds.), 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall, Boca Raton, FL.
Gill, J., 2002. Bayesian Methods: A Social and Behavioral Sciences Approach. Chapman & Hall, Boca Raton.
Jeffreys, H., 1961. Theory of Probability, third ed. Oxford University Press, NY.
Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS—a Bayesian modeling framework: concepts, structure, and extensibility. Statistics and Computing 10, 325–337.
Lynch, S.M., 2007. Introduction to Applied Bayesian Statistics and Estimation for Social Scientists. Springer, NY.
McGrayne, S.B., 2011. The Theory That Would Not Die. Yale University Press, New Haven, CT.
Robert, C.P., 1994. The Bayesian Choice: A Decision-Theoretic Motivation. Springer, NY.
Savage, L.J., 1972. The Foundations of Statistics, second revised edition. Dover, NY.

Beauty and the Labor Market

Eva M Sierminska, CEPS/INSTEAD Research Institute, Esch-sur-Alzette, Luxembourg
Xing (Michelle) Liu, The University of Arizona, Tucson, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

An important underlying determinant of wage discrimination, as well as the gender wage gap, is the way the labor market rewards individual physical attractiveness. This article surveys the extensive empirical literature of the effect of physical attractiveness on labor market outcomes. Particular attention is given to the channels through which attractiveness may affect individual labor market outcomes and the way attractiveness affects gender wage differentials. Further research is needed on the effect of attractiveness within occupations in order to provide more evidence on its productivity-enhancing channel of transmission and the effect this has on the gender wage gap.

Introduction

The labor economics literature has exhibited a long lasting interest in the examination of wage discrimination (*see* Discrimination, Economics of). This article has generated huge amount of research by economists. So far, a large sum of careful empirical studies of wage discrimination based on gender, race, and disabled workers, etc. have been produced. In recent years, discrimination in the labor market against a unique group – those physically unattractive/attractive – has brought increasing attention primarily because the number of employment-related discrimination claims based on employees' appearance has continued to increase (Malos, 2007). In July 2001, the city and county of San Francisco released compliance guidelines to prohibit weight and height discrimination in the local labor market. In 2008, the District of Columbia enacted a series of protection measures for employees, by making it illegal to discriminate on the basis of outward appearance for the purpose of recruitment, hiring, or promotion (Hamermesh, 2011) All these local legislations are aimed at protecting those that could find themselves in the physical appearance disadvantaged group. Although, not everyone in the past agreed that these type of efforts should take place (Barro, 1998).

The first economic study on the relationship between appearance and labor market outcomes, which laid a solid foundation for the studies afterward, was provided by Hamermesh and Biddle (1994). The focal point in this strand of the literature is to determine (1) whether labor market outcomes differentials based on physical appearance exist; and (2) if they exists, what are their potential explanations. Identifying the magnitude of the effect of physical appearance on labor market outcomes (known also as the beauty effect) and the channels through which they work can be critical as policy implications may vary across the various explanations. The effect of appearance may also vary across genders, which, in turn, may affect individual labor market responses (as well as the opportunity in the marriage market), differently. Therefore, studying the beauty effect across genders may help explain the observed gender wage inequality (or gap) to some extent. Further, beauty-based labor market outcome premiums (or penalties) may vary in across cultures and across countries, thus the study of country differences may shed light on the establishment and development of antidiscrimination laws in other countries.

This article reviews the large literature on the impact of physical attractiveness on labor market outcomes. The focus is on beauty and it is not the goal of the survey to provide an exhaustive review of other physical appearance attributes (such as height and weight (*see* Obesity and Labor Market Performance), etc.). Section What Is Beauty? Beauty Measure and Alternative Measures briefly introduces the definition of beauty and the beauty measures. Section Potential Explanations of Beauty Premium/Plainness Penalty summarizes the potential explanations of beauty premium/plainness penalty. Section Empirical Estimation Methods reviews the empirical models used to estimate the effect of beauty. Section Gender Differences discusses the beauty effect across genders. Section Country Findings lists the findings of the beauty effects across countries. Section Conclusion concludes.

What Is Beauty? Beauty Measure and Alternative Measures

Seldom have researchers offered a conceptual definition of physical attractiveness primarily because a physical attractive person is hard to define precisely. In Hatfield and Sprecher (1986) the authors define physical attractiveness as someone "(…) which represents one's conception of the ideal in appearance; (…) which gives the greatest degree of pleasure to the senses." By characterizing most previous research, Morrow (1990) proposes a definition of physical attractiveness from an inductive perspective: physical attractiveness is the degree to which one's facial image elicits favorable reaction from another. Although these two definitions of physical attractiveness are fairly consistent, there are few consistent standards of attractiveness across cultures (Hatfield and Sprecher, 1986; Hamermesh and Biddle, 1994) Fortunately, it has been shown from previous studies that, within a culture at a point in time, there is tremendous agreement on beauty standards and these standards change fairly slowly over time (Hamermesh and

Biddle, 1994) These conclusions made from a psychological and sociological view indicate that attractiveness is at least potentially measurable.

In psychological research, the most common way in which the concept of attractiveness is measured is by asking raters to judge the physical attractiveness of people in portrait photographs or to use self-reported appearance ratings. According to the theory mentioned above, physical attractiveness measured in this manner should have high reliability because people within a given culture tend to agree with each other regarding whether a person's facial appearance is attractive or not (Hatfield and Sprecher, 1986; Umberson and Hughes, 1987; Hamermesh and Biddle, 1994). (Although Doorley and Sierminska (2014) find that interview measured attractiveness can exhibit a 'personality'effect and vary from beginning to the end of the interview.) In addition to the attractiveness measures induced from evaluating photographs, others have used alternative measures emphasizing other aspects of physical attractiveness. These alternative attractiveness measures include height/stature (Frieze et al., 1991; Loh, 1993; Steckel, 1995; Persico et al., 2004; Deaton and Arora, 2009; Loureiro et al., 2012), weight/obesity/BMI (*see* Obesity and Labor Market Performance) (Larkin and Pines, 1979; Haskins and Ransford, 1999; Register and Williams, 1990; Frieze et al., 1991; Loh, 1993; Averett and Korenman, 1996; Maranto and Stenoien, 2000; Cawley, 2004; Loureiro et al., 2012; Caliendo and Gehrsitz, 2014), appropriateness of dress (Lambert, 1972), hair color (Johnston, 2010), and physical disability (Loureiro et al., 2012). Among all these alternative measures, height and weight are commonly used in economics literature. Some researchers point out that these two measures are less subject to measurement error, do not depend on subjective ratings from observers, and also play a fundamental role in the perception of attractiveness. However from a point of view of validity (or construct validity) these two measures are still not sufficient as the physical attractiveness of a person is more than standard height and weight.

Potential Explanations of Beauty Premium/Plainness Penalty

A large body of literature has documented that above-average looking workers receive a wage premium while below-average looking ones receive a wage penalty. The quandary, then, is why does there exist beauty premium/plainness penalty and through which channel can physical attractiveness affect earnings? Possible explanations for the channels cited in the literature include employer discrimination; customer discrimination/productivity; occupational sorting; and attractiveness affecting individuals' cognitive/noncognitive skills that are important for job performance. While the empirical evidence in support of some of these explanations is relatively thin, the attention they receive in the literature leads us to discuss each of these in turn.

Employer Discrimination

There are two conceptually distinct reasons employers might wish to hire more attractive workers. The first is based on a belief that good-looking workers are more productive or more capable even when workers ability is not observed. The second is that employers simple prefer to work with good-looking workers even when they do not have a biased belief about the workers actual ability. The former reason is attributed to an employers' biased belief or stereotype, while the later is Becker type discrimination (*see* Discrimination, Economics of).

People respond differently to attractive and unattractive individuals in a work setting. In general, people tend to be favorably biased toward those who are more attractive (Hatfield and Sprecher, 1986). One may wonder why these attractiveness effects occur. Research by Miller (1988), Dion et al. (1972), and Snyder (1984) provides the most straightforward answer from a psychological point of view. In their studies, they show that one's impression/belief based on the observation of someone's physical attractiveness affects the way one responds to that person. This type of belief assumes that physically attractive people have more socially pleasing personalities and that they are more likely to be successful in their careers. This biased belief can be particularly important when employers are making hiring decisions during face-to-face interviews. In fact, earlier psychological studies have confirmed the stereotype hypothesis, which predicts that attractive people are more likely to be hired for a variety of jobs (Dipboye et al., 1977; Riggio and Throckmorton, 1988). (Hamermesh (2012) finds that decision-makers are still more likely to respond to absolute rather than relative differences among physically attractive people.) The 'taste-based discrimination,' on the other hand, is developed by Becker (1971). In his book, he introduces the first economic model of discrimination. Becker's model assumes that employers have disamenity value to employing minority workers. In the beauty context, minority workers represent the unattractive workers. Employers may have unbiased beliefs about workers' performance but prefer hiring comparatively attractive people. Therefore, unattractive workers may have to accept a lower wage for identical productivity or, equivalently, be more productive given identical wage. Hamermesh and Biddle (1994) and Harper (2000) both provide strong support for employer discrimination against unattractive or short workers.

Although the effect of attractiveness on earnings is shown to be attributed to employer discrimination, it is still difficult to determine whether employer discrimination is due to biased belief or distaste. The reason for this is that household survey data usually do not provide information regarding workers' job performance realization, employers' estimate of a worker's ability, and employers' reaction in wage adjustment if workers' true performance is observed. Fortunately, this difficulty can be overcome under experimental settings in which worker and employer's wage negotiation process can be observed as well as workers' task-solving skill that is unrelated to physical attractiveness. Examples of laboratory experiments used to examine the beauty effect channels include Mobius and Rosenblat (2006) and Deryugina and Shurchkov (2013a). Both studies have found that employers' biased belief on the ability of attractive workers (through visual and oral stereotype) explains a large proportion of the beauty premium. Specifically, the latter one points out that the beauty premium can only be obtained from specific tasks (e.g., bargaining instead of data analysis or data entry). In the experiment, the authors

repeat the wage bidding for several rounds and find that when the workers' true ability is revealed, the beauty premium completely vanishes due to the employers' quick learning process.

Customer Discrimination/Productivity

It is possible that physically attractive workers are more productive than unattractive ones in some occupations, such as actors/actress, sales assistant, and waiters/waitress, etc. (*see* Erotic Capital). These types of occupations typically involve extensive interaction between workers and the customers. Therefore, such an advantage could arise from customer discrimination with customers preferring to interact with better-looking workers. Moreover, physical attractiveness may enhance workers' ability to engage in productive interaction with coworkers. Good working relationship with coworkers and the firm's clients can create a form of firm-specific human capital, which generates higher earnings for workers themselves and higher quasi-rents for their employers (Pfann et al., 2000).

It is been shown that attractive salesmen generally enjoy the benefit from customer discrimination. Reingen and Kernan (1993) suggest that buyers in the attractive-salesman condition are significantly more likely to grant a demonstration appointment and buy the product in a simulated buying–selling setting. Moreover, perceptions of attractiveness appear to induce buyers to behave more favorably toward attractive sellers. Using actual salespeople's data in shopping malls, Sachsida et al. (2003) results point out that women with good appearance receive a reward of approximately 9% in their earnings and this reward is entirely induced by the productivity effect and not discrimination. Hamermesh and Biddle (1994) and Biddle and Hamermesh (1998) also examine whether productivity accounts for the beauty premium using household survey data and a sample of law school graduate data. However, their results only provide weak support for productivity-related discrimination. Hamermesh and Parker (2005) continue to test this hypothesis using a sample of university instructors. Although they are able to uncover a positive association between standardized beauty scores and students' class ratings, they are still not able to disentangle the effect that stems from productivity difference from the one that stems from discrimination. Using an advertising industry sample, Pfann et al. (2000) show that executive's beauty raises firm's revenue, and the increase in firm's revenue far exceeds the extra wages that good-looking executives obtained. Therefore, executives' beauty enhances firm's human capital and the wage premium obtained is primarily induced by productivity. In sum, the evidence provided from above suggests that productivity-related discrimination is highly occupational specific.

Occupational Sorting

It is reasonable to believe that one's physical attractiveness can help to determine the occupational choice that one makes as a worker. As it is been mentioned above the effect of attractiveness could be fairly large in occupations, which involve extensive interaction with customers (like sales assistants). However, one's occupational choice is not only based on the advantages that one believes he/she has, but also on one's preferences regarding different activities and one's ability to perform different types of work. To put it differently, an individual does not choose to enter occupations solely based on one's appearance and the potential payoff to one's appearance in various occupations. The occupation decision also depends on individual characteristics other than appearance such as education level, past work experience, marital status, and parents' occupations, etc. That is why attractive workers are observed in greater proportion in occupations in which attractiveness is rewarded, but not all individuals who work in occupations where beauty pays are attractive (and we find attractive people in occupations where beauty does not pay off). The fact that people see less variation in looks within an occupation than in the workforce as a whole is attributed to the fact that segregation in appearance is incomplete. On average, better-looking individuals choose occupations where looks are important, and below-average looking individuals avoid these types of occupations.

The occupational sorting channel has been investigated by several studies. Hamermesh and Biddle (1994), Biddle and Hamermesh (1998), Harper (2000), and Doorley and Sierminska (2012) all have found empirical support for the standard implication that workers who are well endowed with beauty choose the type of occupations where the payoff to appearance is high. Liu (2014) shows that more attractive women are more likely to participate in managerial/administrative jobs while less likely to choose operative/craftsmen jobs. In particular, Biddle and Hamermesh (1998) extend the traditional sorting model to account for dynamic sorting of beauty using a longitudinal survey data. Using a sample of law school graduates, they first split the sample into individuals who work in private sector and public sectors 5 and 15 years after graduation. They then investigate the sector-switching pattern based on attractiveness. Their results show that lawyers who left the private sector between year 5 and year 15 are less attractive than those who stayed (controlling for other characteristics), while lawyers who switched from the public sector to private sector are more attractive than those who stayed. Their results provide evidence of dynamic sorting in directions consistent with changes in the relative returns to the individual characteristics.

Beauty Correlated with Cognitive/Noncognitive Skills

Another explanation for the existence of the beauty premium is that beauty may in fact be positively correlated with certain cognitive/noncognitive skills such as test score, communication skill, confidence, leadership skills, personality, etc., which are rewarded in the labor market or enable workers to enhance labor productivity (Feingold, 1992). These cognitive/noncognitive skills are usually acquired during adolescent period through socialization or activities in high school. It is been shown that attractiveness at a younger age may contribute human capital investment during adolescence because good-looking children are given more attention by teachers and peer (Lerner et al., 1990; Langlois et al., 2000). An extensive body of research indicates the importance of cognitive/noncognitive skills obtained during high school in future labor

market outcomes, such as test scores, college enrollment, social behavior, employment, occupational attainment, and wages (French et al., 2009; Cobb-Clark and Tan, 2011; Carneiro et al., 2007; Heckman and Rubinstein, 2001; Heckman et al., 2006; Jacob, 2002; Kuhn and Weinberger, 2005). Persico et al. (2004) investigate the wage premium resulting from being tall. They point out that the height premium actually comes from teenage height instead of adult height. Being tall during adolescent period increases the likelihood of participation in high school sports and clubs. Participation in these activities promotes the acquisition of confidence, which in turn increases future wages. Mobius and Rosenblat (2006) and Scholz and Sicinski (2014) reach similar conclusion using attractiveness measures. They find that more attractive individuals exude better communication skills, more confidence, and more extroversion, which all translate into higher wages. Mocan and Tekin (2010) evaluate the attractiveness effect on criminal activities and finds that unattractiveness increases a young adult's propensity for criminal activity. They argue that this is primary due to the fact that unattractiveness hinders student human capital development during high school. Han et al. (2011) examine the direct and indirect effect of teenage body weight on adult wages and find that education is the main pathway for the indirect BMI wage penalty. They find that the BMI wage penalty is underestimated by about 18% for women and is not statistically significant for men.

Individual's cognitive skill such as test score and study performance may also be correlated with attractiveness. Von Bose (2012) and Deryugina and Shurchkov (2013a) both show that physical attractiveness is positively correlated with high school Grade Point Average (GPA) and university overall GPA. Others have shown that beauty and ability are complements at high levels of beauty, but substitutes at low levels (Fletcher, 2009).

Empirical Estimation Methods

The primary method used in the literature to model the beauty effect on labor market outcomes is a Mincer type human capital model or wage equation. This type of model regresses individual earnings (or log hourly wage) on categories of attractiveness (e.g., above-average and below-average), a set of productivity-enhancing characteristics (e.g., years of schooling and potential experience, etc.), and a vector of individual characteristics (e.g., race, marital status, union status, self-reported health status, etc.). As it has been mentioned above, attractiveness may also be correlated with cognitive/noncognitive skills. If these determinants are not controlled for, the estimated effect of attractiveness could be biased. Therefore, previous studies sometimes include IQ scores, high school activities, and if the data allow the 'big five' personality traits (openness, conscientiousness, extraversion, agreeableness, neuroticism). The coefficients on the attractiveness measure represent the marginal variation in wage in relation to beauty. Almost all the previous literature (including the psychological, sociology, and economic literature) have adopted this type of linear model to construct baseline beauty effect estimates (*see* Wage Differential and Structure).

A less commonly used approach is introduced by Doorley and Sierminska (2012), who use distribution regression followed by an Oaxaca–Blinder style decomposition (Oaxaca, 1973; Blinder, 1973). The primary reason to adopt this methodology is that the authors posit the effect of beauty may vary across the wage distribution. To proceed, they start by estimating the conditional distribution of wages in different attractiveness group (e.g., beautiful and plain) given individual characteristics. They then construct counterfactual wage distributions. The final wage differentials between beautiful and plain workers can be decomposed into the characteristics effect and the unexplained difference. The unexplained wage differentials are interpreted as the beauty premium/plainness penalty.

One limitation of the Mincer type model and the distribution regression is that these methods do not allow for the potential channels to be separately identified. The empirical model proposed by Hamermesh and Biddle (1994) solves the problem to some extent. They propose the following regression:

$$w_i = \beta_0 + \beta_1 X_i + \beta_2 \theta_i + \beta_3 \mathrm{OCC}_i + \beta_4 \theta_i \mathrm{OCC}_i + \varepsilon_i$$

where w_i represents log hourly earnings. $\mathbf{X}_i$ is a vector of productivity-enhancing characteristics. θ_i is a binary variable that equals to 1 if the worker is attractive and 0 otherwise. $\mathrm{OCC}_i = 1$ if the worker's occupation is as one where looks are productive. The ε_i are residuals. This regression nests a model of occupational sorting, employer discrimination as well as a productivity differential. The occupational sorting model implies $\beta_3 > 0$. The productivity/customer discrimination model implies $\beta_4 > 0$ and $\beta_2 = \beta_3 = 0$. The employer discrimination model implies $\beta_2 > 0$ and $\beta_3 = \beta_4 = 0$. Examples of studies using the Hamermesh and Biddle (1994) model to examine the beauty effect channels include Harper (2000), Sachsida et al. (2003), Loureiro et al. (2012), Doorley and Sierminska (2012), and Johnston (2010), etc.

Although the above model provides a convenient method to test the existence of potential transmission channels of the beauty effect, it is not without its caveat. One difficulty with using this model is to determine the set of occupations where looks are likely to enhance productivity. Clearly, there is no simple rule of thumb. In Hamermesh and Biddle (1994), the authors use three different sources to identify these occupations. These sources include the *Dictionary of Occupation Titles* (1977), three-point scale ratings on whether appearance might be important to a specific occupation by eight raters, and a survey conducted by Holzer (1993). According to their classification, the occupations in which attractiveness could possibly enhance productivity include sales occupations, cashiers, receptionists, and waitresses, etc. These occupations evidently involve more interaction between workers and customers. Doorley and Sierminska (2012) divide occupations into dressy and nondressy taking advantage of a survey question asking individuals' opinion on the importance of appearance in their professional life. In their dressy category, within which beauty may enhance workers' productivity, they identify occupations such as supervisors/managers, intellectual professions, administrative employees, service, and sales employees, etc. The occupations included in their dressy category are shown to be generally consistent with the original 1994 Hamermesh and Biddle paper.

Another difficulty in using this model is to disentangle the productivity channel from other sources. Although Hamermesh and Biddle (1994) are able to provide some support for productivity-related discrimination, the evidence is fairly weak. A related explanation, they provide, is that there are still inherent productivity differences that the data do not capture because of omitted variables. Due to this reason, alternative variations need to be incorporated to better identify this channel. One approach is to use data from workers who are from a relatively homogeneous group (a specific occupation). Biddle and Hamermesh (1998) intend to overcome this difficulty by using a sample of graduates of a particular law school. To measure the wage differentials from consumer discrimination or productivity, they compare earnings between attorneys practicing in the private sector and those in the public sectors. The reasoning behind this research design is that, in the private sector, the duties of lawyers often include a marketing component in that they have to attract new clients and keep the existing ones. In public sectors, however, the ability to market one's services is not an essential part of the job. Therefore, the earning difference between private and public sectors also reveals customers' discrimination. In order to further separate the effect stemming from customers preferring to spend time with better-looking attorneys and away from the effect stemming from customers' belief that better-looking attorneys are more productive (obtain greater financial gains), the authors examine how mean attorneys attractiveness differ by legal specialization within the private sector. This test helps to determine whether attorneys' attractiveness can assist them in generating more favorable judgments from clients in each legal specialty. The results from these comparisons, however, only provide weak evidence for the belief that beauty produces more advantageous outcomes.

Even with the ability to estimate the impact of physical attractiveness on individual earnings, the literature still leaves open the question of how to separately identify productivity-related discrimination. Due to the fact that measures of productivity are fairly few and they are aggregated, given the heterogeneity of respondents' occupations in publicly available household survey, data on relatively homogeneous workers in specific occupation is definitely called for. The literature studying customer discrimination within occupation is still relatively thin, making this a valuable area for future research.

Gender Differences

The effect of attractiveness on labor market outcomes may vary with respect to gender. This is primarily due to the fact that attractiveness may result in different labor market opportunities for men and women. Although the participation gap has been narrowing, in general, women are still more likely than men to stay out of the labor force (additional information can be found in Family and Work; Earnings Inequality; Income Inequality.). This observation is particularly common for the 1970s when women's stated goal was 'career then family' (Goldin, 2004). During the 1970s, about 65% of married women were in the labor force at age 30 (Goldin, 2004), while overall about 97% of men were employed while only 57% of women were in the labor force (Liu, 2014). What determines women's labor force participation decision apart from marital status, number of children and education? Beauty endowment may be one of these factors. As discussed in Section Potential Explanations of Beauty Premium/Plainness Penalty, it is possible that more attractive women will be able to gain more from a job than a less attractive counterpart. More attractive women may be more likely to work because beauty brings them more confidence (Mocan and Tekin, 2010) and they may be more likely to take advantage of their attractiveness if it enhances their productivity at work. On the other hand, the disincentives associated with bad looks may lead women to avoid the labor market (Hamermesh, 2011). Hamermesh and Biddle (1994) investigate this point and show that married women with below-average looks are 3–11% less likely to participate in the labor market than above-average looking women. Harper (2000) finds that the probability of unemployment falls from 2 to 0.9% for attractive women. Borland and Leigh (2014) suggest a relatively large negative effect (about 18 percentage points) of below-average beauty on women's probability of employment. Women's self-selection into the labor force based on beauty endowments alters the distribution of observed female beauty in the labor market. Since more attractive women are more likely to appear in the labor market and vice versa, one observes less female beauty variation and, therefore, a smaller beauty premium. However, selection based on beauty endowment is less prevalent for men given men's fairly high labor force participation rate, thus, one expects to observe larger beauty premium for men than for women.

Although women's self-selection explains the beauty premium gender differences, this explanation is only appropriate when women's labor force participation rate is low. From the 1970s to 2000s, married women's labor force participation continued to expand (Goldin, 2006) and a narrowing of the participation gap between men and women has occurred (Goldin, 2014). In 2012, the labor force participation rate for men between 15 and 54 years of age was about 89%, which is only 14% higher than that of their female counterpart (BLS, 2013). Due to the converging roles of men and women in the labor market and at home over the past 40 years, the observed female beauty variation in the labor market should be largely expanded. Therefore, the beauty premium gap across genders is expected to decrease.

In the end, gender differences in the beauty premium/penalty is an empirical issue. Careful studies on gender difference in the effects of attractiveness on earnings show that the above hypothesis is general true. By stacking two American and one Canadian data set, Hamermesh and Biddle (1994) report that men who are assessed as being homely are penalized about 9% in hourly earnings while those who are viewed as having above-average looks receive an earning premium of 5%. Among women, there is a 4% attractiveness premium and a 5% plainness penalty in hourly earnings. Harper (2000) shows that men who are rated as unattractive at both ages 7 and 11 incur a large and significant pay penalty of 14.9%. This pattern also applies to their female counterpart with a 10.9% unattractiveness penalty. One point that needs to be noticed is that Harper (2000) attractiveness measures are obtained from evaluating adolescents' appearances by their teachers. Since adolescents' appearance may change dramatically when people get older and

may also contain relatively large measurement error, it is possible that the estimates suffer from attenuation bias. Borland and Leigh (2014) show that, in the 1980s, above-average beauty for men is associated with a premium of 11.6% and below-average beauty with a penalty of 10.4% with the latter effect being statistically insignificant. For women, however, no statistically significant attractiveness premium or unattractiveness penalty is found. They conclude that over time (1984 and 2009) the returns to beauty from the effect on a worker's hourly wage are more or less constant. Mocan and Tekin (2010) also confirm a similar pattern using a sample of young adults. Their results suggest that, using all three survey waves, the beauty premium is 4.5% for women, and being unattractive is associated with a 7% reduction in wages. The wage premium to be very attractive for men is 10.8%, and the penalty for being unattractive is the same amount as women (7%).

The above hypothesis regarding gender differences still holds even within specific occupations. Frieze et al. (1991) investigate the attractiveness premium using a sample of 737 MBA graduates between 1973 and 1982. They first conduct a Chow Test to confirm that the salary relationships for men and women are different. They then estimate the beauty effect separately by gender. They find that more attractive men obtained higher starting salaries and continued to earn more over time. For women, however, there exists no effect of attractiveness on starting salaries although more attractive women begin to earn more with greater job experience. Biddle and Hamermesh (1998) examine the payoff of law school graduates who graduated in the 1970s and 1980s. Their results indicate that the attractiveness effect of men is statistically significant 5 years after graduation with two standard deviation increases in attractiveness score associated with about a 10% increase in earnings. The estimated attractiveness effects for women, however, are indistinguishable from zero. Clearly, the measure of labor market outcomes may not be limited to individual earnings. Any measure that evaluates the performance in an occupation can be used as an outcome variable. In the case of university teaching, course/instructor evaluations are reasonable measures of teaching performance. A number of studies have shown that better teaching evaluations increase teachers' salaries (Moore et al., 1998; Kaun, 1983). Therefore, teachers' attractiveness may have an indirect impact on their salary via students' class ratings. Hamermesh and Parker (2005) evaluate beauty in the classroom. They find that a one standard deviation increase in the attractiveness score is associated with a 0.38-point increase in male teachers' class ratings while the beauty effect for female teachers is less than half compared to that of males.

Although the (women's) self-selection hypothesis is generally supported by previous studies, it is not uncommon to see studies that reach different conclusions. French (2002) is one such example. In this article, the sample data used is collected from two worksites (a nonprofit hospital and a large school district) in a Midwestern community in 1995–97. Using self-reported appearance as a beauty measure, the author finds that women with above-average appearance obtain about 8% higher earnings relative to women with an average appearance. There is no evidence that shows that women suffer from the plainness penalty. For men, neither the beauty premium nor the plainness penalty is detected in the regression results. Liu (2014) finds that gender differences in the beauty effect vary with the individual's age. The results suggest that men in their mid-1930s have no significant beauty premium. For women, on the other hand, a one standard deviation increase in the attractiveness rating is associated with a 5.4% increase in the hourly wage. At a later stage of the career (early-1950s), men's attractiveness premium develops with one standard deviation increase in attractiveness score associated with 2.65% increase in hourly wage while women's attractiveness premium fades. However, their findings are consistent with the narrowing beauty premium gap hypothesis mentioned above.

The gender wage gap literature has indicated several sources of wage inequality across genders. These sources include workers' observed characteristics, gender-specific workforce composition change (Mulligan and Rubinstein, 2008; Ruffle and Shtudiner, 2010), collective bargaining effect (Blau and Kahn, 2003; Sachsida et al., 2003; Antonczyk et al., 2010; Scholz and Sicinski, 2014), competition (Weichselbaumer and Winter-Ebmer, 2007), and work-time flexibility (Goldin, 2014). However, these studies do not consider the potential effect of physical attractiveness on individual earnings across genders. The literature reviewed in this section above confirms that gender differences in the beauty premium do exist and indicates that the role attractiveness plays in explaining the gender wage gap could be bigger than previously assumed.

Country Findings

It might be futile to search for a universal standard of beauty since they are determined within cultures (Hatfield and Rapson, 2000) and the variation of cultures around the world results in few consistent standards of beauty (Hatfield and Sprecher, 1986). One question one may want to ask is whether people in different culture all favor beauty? Is the beauty premium/plainness penalty consistent across different countries? In order to answer this question, we compare findings from previous international studies. The comparisons are shown in **Table 1**.

Evidently, physical attractiveness is rewarded in almost all country findings, except for men in Luxembourg. Among all the countries listed, the beauty premium in Germany and China are the highest (from 17.9 to 20%), especially for women. The largest plainness penalty shows up in studies investigating British and Australian data (from −10.9 to −14.9%). Interestingly, Harper (2000) and Borland and Leigh (2014) estimates show that individuals in the United Kingdom and women in Australia do not receive any attractiveness premium. This similarity in the payoff of attractiveness may be due to the similarity in the historical economic structure, as well as culture in both countries. The reader should notice that **Table 1** presents selected country findings from the literature. Some studies that investigate the beauty premium have adopted either an alternative attractiveness measures (e.g., height, weight, hair color, etc.) or used alternative labor market outcomes (e.g., whether individual gets called back during hiring process). Ruffle and Shtudiner (2010) and López Bóo et al. (2013) are examples of this kind of research. Both papers examine the link between beauty and hiring practices but use sample from different countries. Ruffle and Shtudiner (2010) investigate

Table 1 Comparison of the effect of beauty on wages across various countries

Country	Paper	Gender	Occupation	Wage effect: Above-average looks (%)	Wage effect: Below-average looks (%)	Notes
Canada and U.S.	Hamermesh and Biddle (1994)	Men	General	5.4	−8.9	Stacked estimates
		Women		3.9	−5.5	
U.S.	Mocan and Tekin (2010)	Men	General	10.8	−7	
		Women		4.5	−7	
U.K.	Harper (2000)	Men	General	Not significant	−14.9	
		Women		Not significant	−10.9	
Netherlands	Pfann et al. (2000)	Both	Advertising firm	18 800 DFL increase in wage with average beauty changes from 10th to 90th percentile (assuming a 7.5% effect on wages averaging 150 000 DFL per year)		Wage effect inferred from extraneous estimates
China (Shanghai)	Hamermesh et al. (2002)	Men	General	–	–	–
		Women		17.9	–	
Brazil	Sachsida et al. (2003)	Men	Salesmen	Not significant	Not significant	
		Women		9	Not significant	
Germany	Doorley and Sierminska (2012)	Men	General	14	–	
		Women		20	–	
Luxembourg	Doorley and Sierminska (2012)	Men	General	−3	–	
		Women		10	–	
Australia in 1984	Borland and Leigh (2014)	Men	General	11.6	Not significant	
		Women		Not significant	Not significant	
Australia in 2009	Borland and Leigh (2014)	Men	General	Not significant	−12.9	
		Women		Not significant	Not significant	

DFL = Dutch Florin (Dutch currency prior to 2002).

data from Israel while López Bóo et al. (2013)look at data from Buenos Aires, Argentina. Both studies provide evidence that attractive people receive more responses (callbacks) from employers than unattractive people. Also, the attractive ones tend to receive callbacks sooner than their unattractive counterparts. (An interesting discussion on the possible way to reduce unfair advantages related to physical appearance can be found in Krause et al., 2012.)

Conclusion

The attention given to beauty-based discrimination has increased in recent years. As this article has shown, most previous studies in this area focus on estimating the magnitude of the effect of physical attractiveness on labor market outcomes and trying to identify channels through which attractiveness can affect these outcomes using various survey data sources. This body of literature generally confirms the existence of a positive association between physical attractiveness and various labor market outcomes such as individual earnings and employment opportunity (callback rate). Moreover, several explanations (e.g., employer discrimination, customer discrimination/productivity, occupational sorting, etc.) of how beauty affects labor market outcomes are provided.

As it is mentioned above, all explanations are supported by previous studies to some extent; the empirical evidence in support of some of the potential explanations is still relative thin. This is particularly true for the productivity channel. Therefore, researchers could contribute in this area by using within occupation specific data, where workers can be viewed as relatively homogeneous. Additionally, this would partly remove the unobserved factors of productivity.

Another area where researchers could help shed light is the endogeneity of beauty. In previous studies, physical attractiveness is generally treated as exogenously determined. However, perceived attractiveness may be correlated with clothing/cosmetics spending, plastic surgery, and parents' attractiveness; these are usually not observed from survey data. Failing to consider the endogeneity issue may cause the estimated beauty effect to be biased upward.

See also: Discrimination, Economics of; Earnings Inequality; Erotic Capital; Income Inequality; Wage Differential and Structure.

Bibliography

Antonczyk, Dirk, Fitzenberger, Bernd, Sommerfeld, Katrin, 2010. Rising wage inequality, the decline of collective bargaining, and the gender wage gap. Labour Economics 17 (5), 835–847.

Averett, Susan, Korenman, Sanders, 1996. The economic reality of the beauty myth. Journal of Human Resources 31 (2), 304–330.

Barro, Robert, 16 March 1998. So you want to hire the beautiful. Well, why not? Business Week.

Becker, Gary S., May 1971. The Economics of Discrimination. Economics Books, University of Chicago Press.

Biddle, Jeff E., Hamermesh, Daniel S., January 1998. Beauty, productivity, and discrimination: lawyers' looks and lucre. Journal of Labor Economics 16 (1), 172–201.

Blau, Francine D., Kahn, Lawrence M., January 2003. Understanding international differences in the gender pay gap. Journal of Labor Economics 21 (1), 106–144.

Blinder, Alan S., 1973. Wage discrimination: reduced form and structural estimates. The Journal of Human Resources 8 (4), 436–455.

BLS, 2013. Labor force projections to 2022: the labor force participation rate continues to fall. Available at: http://www.bls.gov/opub/mlr/2013/article/labor-force-projections-to-2022-the-labor-force-participation-rate-continues-to-fall.htm.

Borland, Jeff, Leigh, Andrew, March 2014. Unpacking the beauty premium: what channels does it operate through, and has it changed over time? Economic Record 90 (288), 17–32.

Caliendo, Marco, Gehrsitz, Markus, February 2014. Obesity and the Labor Market: A Fresh Look at the Weight Penalty. IZA Discussion Papers 7947. Institute for the Study of Labor (IZA).

Carneiro, Pedro, Crawford, Claire, Goodman, Alissa, 2007. The Impact of Early Cognitive and Non-cognitive Skills on Later Outcomes, Centre for the Economics of Education, London School of Economics, London.

Cawley, John, 2004. The impact of obesity on wages. Journal of Human Resources 39 (2), 451–474.

Cobb-Clark, Deborah A., Tan, Michelle, 2011. Noncognitive skills, occupational attainment, and relative wages. Labour Economics 18 (1), 1–13.

Deaton, Angus, Arora, Raksha, 2009. Life at the top: the benefits of height. Economics and Human Biology 7 (2), 133–136.

Deryugina, Tatyana, Shurchkov, Olga, August 2013a. Does Beauty Matter in Undergraduate Education? MPRA Paper 53582 University Library of Munich, Germany.

Deryugina, Tatyana, Shurchkov, Olga, May 2013b. When Are Appearances Deceiving? the Nature and Evolution of the Beauty Premium.

Dion, Karen, Berscheid, Ellen, Walster, Elaine, 1972. What is beautiful is good. Journal of Personality and Social Psychology 24 (3), 285.

Dipboye, Robert L., Arvey, Richard D., Terpstra, David E., 1977. Sex and physical attractiveness of raters and applicants as determinants of resumé evaluations. Journal of Applied Psychology 62 (3), 288.

Doorley, Karina, Sierminska, Eva, June 2012. Myth or Fact? The Beauty Premium across the Wage Distribution. IZA Discussion Papers 6674. Institute for the Study of Labor (IZA).

Doorley, Karina, Sierminska, Eva, March 2014. First Impressions Count but Can Personality Affect Wages?.

Feingold, Alan, 1992. Good-looking people are not what we think. Psychological Bulletin 111 (2), 304–341.

Fletcher, Jason M., December 2009. Beauty vs. brains: early labor market outcomes of high school graduates. Economics Letters 105 (3), 321–325.

French, Michael T., 2002. Physical appearance and earnings: further evidence. Applied Economics 34, 569–572.

French, Michael T., Robins, Philip K., Homer, Jenny F., Tapsell, Lauren M., August 2009. Effects of physical attractiveness, personality, and grooming on academic performance in high school. Labour Economics 16 (4), 373–382.

Frieze, Irene H., Olson, Josephine E., Russell, June, 1991. Attractiveness and income for men and women in management. Journal of Applied Social Psychology 21 (13), 1039–1057.

Goldin, Claudia, 2004. The long road to the fast track: career and family. The Annals of the American Academy of Political and Social Science 596 (1), 20–35.

Goldin, Claudia, May 2006. The quiet revolution that transformed women's employment, education, and family. American Economic Review 96 (2), 1–21.

Goldin, Claudia, April 2014. A grand gender convergence: its last chapter. American Economic Review 104 (4), 1091–1119.

Hamermesh, Daniel S., 2011. Beauty Pays: Why Attractive People Are More Successful. Princeton University Press.

Hamermesh, Daniel S., 2012. Tall or taller, pretty or prettier: is discrimination absolute or relative. IZA Journal of Labor Economics 1 (1), 1–17. http://dx.doi.org/10.1186/2193-8997-1-2.

Hamermesh, Daniel S., Biddle, Jeff E., December 1994. Beauty and the labor market. American Economic Review 84 (5), 1174–1194.

Hamermesh, Daniel S., Meng, Xin, Zhang, Junsen, 2002. Dress for success – does primping pay? Labour Economics 9 (3), 361–373.

Hamermesh, Daniel S., Parker, Amy, 2005. Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity. Economics of Education Review 24 (4), 369–376.

Han, Euna, Norton, Edward C., Powell, Lisa M., 2011. Direct and indirect effects of body weight on adult wages. Economics & Human Biology 9 (4), 381–392.

Harper, Barry, 2000. Beauty, stature and the labour market: a British cohort study. Oxford Bulletin of Economics and Statistics 62, 771–800.

Haskins, Katherine M., Ransford, H. Edward, 1999. The relationship between weight and career payoffs among women. In: Sociological Forum, vol. 14. Springer, pp. 295–318.

Hatfield, E., Rapson, R.L., 2000. Physical Attractiveness, vol. 3. John Wiley & Sons.

Hatfield, Elaine, Sprecher, Susan, 1986. Mirror, Mirror. State University of New York Press, Albany.

Heckman, James J., Rubinstein, Yona, 2001. The importance of noncognitive skills: lessons from the GED testing program. American Economic Review, 145–149.

Heckman, James J., Stixrud, Jora, Urzua, Sergio, 2006. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. Journal of Labor Economics 24 (3), 411–482.

Holzer, Harry, 1993. Multi-city Study of Urban Inequality (Unpublished manuscript). Michigan State University.

Jacob, Brian A., 2002. Where the boys aren't: non-cognitive skills, returns to school and the gender gap in higher education. Economics of Education Review 21 (6), 589–598.

Johnston, David W., 2010. Physical appearance and wages: do blondes have more fun? Economics Letters 108 (1), 10–12.

Kaun, David E., 1983. Faculty advancement in a nontraditional university environment. Industrial and Labor Relations Review 37, 592.

Krause, Annabelle, Rinne, Ulf, Zimmermann, Klaus F., 2012. Anonymous job applications in Europe. IZA Journal of European Labor Studies 1 (5).

Kuhn, Peter, Weinberger, Catherine, 2005. Leadership skills and wages. Journal of Labor Economics 23 (3), 395–436.

Lambert, S., 1972. Reactions to a stranger as a function of style of dress. Perceptual and Motor Skills 35 (3), 711–712.

Langlois, Judith H., Kalakanis, Lisa, Rubenstein, Adam J., Larson, Andrea, Hallam, Monica, Smoot, Monica, 2000. Maxims or myths of beauty? A meta-analytic and theoretical review. Psychological Bulletin 126 (3), 390.

Larkin, Judith C., Pines, Harvey A., 1979. No fat persons need apply experimental studies of the overweight stereotype and hiring preference. Work and Occupations 6 (3), 312–327.

Lerner, Richard M., Delaney, Mary, Hess, Laura E., Jovanovic, Jasna, Von Eye, Alexander, 1990. Early adolescent physical attractiveness and academic competence. The Journal of Early Adolescence 10 (1), 4–20.

Liu, Xing, 2014. Physical Attractiveness and Earnings: Evidence from a Longitudinal Survey. Technical report, Working paper. University of Arizona.

Loh, Eng Seng, 1993. The economic-effects of physical appearance. Social Science Quarterly 74 (2), 420–438.

López Bóo, Florencia, Rossi, Martn A., Urzua, Sergio S., 2013. The labor market return to an attractive face: evidence from a field experiment. Économics Letters 118 (1), 170–172.

Loureiro, Paulo, Sachsida, Adolfo, Mendonca, Mario C., February 2012. Links between physical appearance and wage discrimination: further evidence. International Review of Social Sciences and Humanities 1 (2), 1–16.

Malos, Stan, 2007. Appearance-based sex discrimination and stereotyping in the workplace: whose conduct should we regulate? Employee Responsibilities and Rights Journal 19 (2), 95–111.

Maranto, Cheryl L., Stenoien, Ann Fraedrich, 2000. Weight discrimination: a multidisciplinary analysis. Employee Responsibilities and Rights Journal 12 (1), 9–24.

Miller, Carol T., 1988. Categorization and the physical attractiveness stereotype. Social Cognition 6 (3), 231–251.

Mobius, Markus M., Rosenblat, Tanya S., 2006. Why beauty matters. The American Economic Review, 222–235.

Mocan, Naci, Tekin, Erdal, 2010. Ugly criminals. The Review of Economics and Statistics 92 (1), 15–30.

Moore, William J., Newman, Robert J., Turnbull, Geoffrey K., 1998. Do academic salaries decline with seniority? Journal of Labor Economics 16 (2), 352–366.

Morrow, Paula C., 1990. Physical attractiveness and selection decision making. Journal of Management 16 (1), 45–60.

Mulligan, Casey B., Rubinstein, Yona, 2008. Selection, investment, and women's relative wages over time. The Quarterly Journal of Economics 123 (3), 1061–1110.

Oaxaca, Ronald, October 1973. Male-female wage differentials in urban labor markets. International Economic Review 14 (3), 693–709.

Persico, Nicola, Postlewaite, Andrew, Silverman, Dan, October 2004. The effect of adolescent experience on labor market outcomes: the case of height. Journal of Political Economy 112 (5), 1019–1053.

Pfann, Gerard A., Biddle, Jeff E., Hamermesh, Daniel S., Bosman, Ciska M., 2000. Business success and businesses' beauty capital. Economics Letters 67 (2), 201–207.

Register, Charles A., Williams, Donald R., 1990. Wage effects of obesity among young workers. Social Science Quarterly 71 (1), 130–141.

Reingen, Peter H., Kernan, Jerome B., 1993. Social perception and interpersonal influence: some consequences of the physical attractiveness stereotype in a personal selling setting. Journal of Consumer Psychology 2 (1), 25–38.

Riggio, Ronald E., Throckmorton, Barbara, 1988. The relative effects of verbal and nonverbal behavior, appearance, and social skills on evaluations made in hiring interviews. Journal of Applied Social Psychology 18 (4), 331–348.

Ruffle, Bradley, Shtudiner, Ze'ev, 2010. Are Good-Looking People More Employable?. Discussion Paper, (10–06) Ben-Gurion University of the Negev.

Sachsida, Adolfo, Dornelles, Adriana C., Mesquita, Carlos W., 2003. Beauty and the Labor Market-Study One Specific Occupation. Available at: SSRN 505742.

Scholz, John K., Sicinski, Kamil, 7 January 2014. Facial attractiveness and lifetime earnings: evidence from a cohort study. Review of Economics and Statistics.

Snyder, Mark, 1984. When belief creates reality. Advances in Experimental Social Psychology 18, 247–305.

Steckel, Richard H., 1995. Stature and the standard of living. Journal of Economic Literature, 1903–1940.

Umberson, Debra, Hughes, Michael, 1987. The impact of physical attractiveness on achievement and psychological well-being. Social Psychology Quarterly 50 (3), 227–236.

Von Bose, Caroline, 2012. Child Stars vs. Ugly Ducklings: Does Adolescent Attractiveness Contribute to the Beauty Premium? Technical report, Working paper.

Weichselbaumer, Doris, Winter-Ebmer, Rudolf, 2007. The effects of competition and equal treatment laws on gender wage differentials. Economic Policy 22 (50), 235–287.

Beauvoir, Simone de (1908–86)

Elizabeth Fallaize[†], Oxford University, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 2, pp. 1060–1064, © 2001, Elsevier Ltd.

Abstract

As the author of *The Second Sex* (1949), a radical and multidisciplinary essay on women, Beauvoir has come to be considered as a seminal feminist theorist. In her essay she deconstructs the cultural myths that construct femininity, examines the way in which women have historically been positioned as the other or object to men as subject, and analyses the weight of social and material oppression to which women have been subjected. Combined with this socially-constructionist analysis of the workings of patriarchy, Beauvoir nevertheless retains a belief in the possibility of individual agency and a freely chosen identity that sets her apart from more recent postmodern accounts of gender.

Novelist, philosopher, essayist, and autobiographer, the author of one of the founding texts of feminism in the twentieth century, Simone de Beauvoir can be considered to be one of the major intellectual figures of her generation. Born in Paris on 9 January 1908 into a conservative middle-class milieu, she received the narrow primary and secondary education deemed appropriate for a marriageable girl of her class. However, after a sudden deterioration in her family's financial circumstances, she was permitted to study philosophy at the Sorbonne, and in 1929 she became the youngest person ever to succeed in the prestigious *agrégation de philosophie* examination, a success which enabled her to earn her own living from a series of teaching posts which she held first in Marseilles, then in Rouen and, from 1936, in Paris. Occupying a series of cheap hotel rooms, she led a remarkably independent life, free from domestic obligations, and, despatching her teaching in the minimum time necessary, she spent her days writing in cafes and participating in a complex web of relationships, at the center of which was the couple she formed with Jean-Paul Sartre. The two never set up home together, setting their face against the idea of marriage and children, but referred to their close circle of former pupils and ex-lovers as 'the family.' The outbreak of the Second World War disrupted both this network of relationships and Beauvoir's assumption that her freedom was virtually total. As she writes in her autobiography she had to 'face the weight of reality.' Continuing her life in occupied Paris without Sartre she completed the manuscript of her first novel, *She Came to Stay* (1943), an exploration of the Hegelian dictum 'Each consciousness pursues the death of the other,' and went on to publish in quick succession *The Blood of Others* (1945), exploring some of the ethical and political dilemmas posed by the Resistance, and *All Men Are Mortal* (1946) which deals with the problem of the historical dimension of human existence. Sartre, meanwhile, had completed his philosophical essay *Being and Nothingness*, also published in 1943; in postliberation Paris the couple became famous figures as existentialism, as Sartre's philosophy came to be known, dominated the Parisian intellectual scene. In 1946 Beauvoir turned her attention to the situation of women, and in 1949 published the two volumes of radical and multidisciplinary thinking which constitute *The Second Sex*. Its frank treatment of sexuality and its critique of the maternal role expected of women made her a figure of some notoriety, and though she went on to write bestselling novels, and won the Goncourt prize for *The Mandarins* in 1954, her name was forever to be associated with her seminal work of feminist theory. From 1958 onward she also began to publish volumes of autobiography, constructing an account of life as a twentieth-century intellectual woman followed by millions of readers worldwide. The 1950s also opened up a phase of active political engagement as she campaigned against French atrocities in Algeria during the Algerian war, supported the Maoist groups in the events of 1968, and, in the 1970s, joined with the newly formed MLF (Women's liberation movement) in campaigns on abortion, contraception, sexism, and violence against women. After Sartre's death in 1980, she published his letters to her, and her own letters both to Sartre and to her American lover Nelson Algren, were published in 1990 and 1997, respectively, after her own death in 1986. It is thus as philosopher, feminist, activist, novelist, memorialist, and correspondent that Beauvoir has made a deep mark on her era.

Nonetheless, it is principally as the author of *The Second Sex* that Beauvoir has come to be considered as one of the most significant feminist voices of the twentieth century. Beauvoir's study of women began life as an autobiographical question (what had it meant to her personally to be a woman), developed into an essay about the cultural myths which construct femininity and eventually became an ambitious multidisciplinary work, studying women's situation from a panoply of directions to demonstrate that 'one is not born, but rather becomes, a woman.' This crucial distinction between sex and gender was to open up the entire field of women's and gender studies. Beauvoir stresses in the introduction that her approach is ontological and existentialist: if there is no fixed human nature then neither can there be a fixed nature for women. This intellectual origin does indeed provide a strong impetus to the work, driven by the assumption that, like all human beings, women can and should shape their own destinies. However, the existentialist approach stresses individual responsibility rather than the way in which patriarchy functions as an institution and, taken to the extreme, it could suggest that women's oppression is a form of individual bad faith.

One of the strengths of *The Second Sex* is that although Beauvoir does not admit explicitly the extent to which social

[†] Deceased

and economic analyses complement her ontology, she does in fact offer a balance between the weight of social, cultural, and economic oppression on the one hand, and the free construction of the subject on the other. Her argument that women historically have been positioned as the object or 'other' to men as subject is supported by a socially constructionist view of patriarchy.

The book begins with a series of attacks on the notion that there is a fixed destiny for women. In part 1 she investigates what biology, psychoanalysis, and Marxism appear to prescribe about female destiny to conclude that women are not defined or inevitably doomed to oppression by the laws of nature, by their psychoanalytic situation, or by economic history. The section on psychoanalysis attacks Freud for his definition of female sexuality in terms of penis envy, an attack that was responsible for a long feminist mistrust of Freud. The section on biology as destiny has proved particularly controversial since, despite insisting that female physiology offers no justification for the hierarchy of the sexes, Beauvoir paints a dispiriting picture of the daily experience of the body for women. In her account the female body is organized to such a high degree to ensure the reproduction of the species that it constitutes a burden against which women must continually struggle in their bid for freedom. Part 2 is devoted to a wide-ranging survey of women's historical role from Ancient Greece and Rome up to the 1940s. It describes women as fundamentally handicapped by maternity from the earliest times, stresses women's role as a medium of exchange, and analyses how tools of patriarchy such as Christianity and the feudal system have corralled the vast majority of women into the margins of historical narratives. Moving on to the myths of femininity, in part 3, Beauvoir dissects a rich set of constructs ranging from the Christian images of Mary and Eve to the pervasive myth of feminine mystery. The section ends on a discussion of five male writers, showing how literature has served as a vehicle for the creation and circulation of myths about women. This analysis later became the inspiration for the work of Kate Millett and the branch of feminist literary criticism that reads male-authored texts with a deconstructive eye.

Parts 4 and 5 of the essay turn to the social and sexual roles which women are called upon to play, tracing the way in which the girl gradually becomes socialized into her role as wife and mother, under the eye of her own mother, the unwitting accomplice of patriarchy. Up to the age of puberty Beauvoir nevertheless credits the young girl with a strong sense of her own autonomy; it is at the moment when the traits of female sexuality install themselves in her body that the girl sees the social trap closing, according to Beauvoir. A section considering female sexual response and case histories of women's first experiences of sexuality follows. Drawing on the Kinsey Report, and on other studies of sexuality, Beauvoir gives serious and detailed consideration to female sexual pleasure, including a discussion of the supposed distinction between the vaginal and clitoral orgasm. She also devotes a chapter to lesbianism in which she rejects the idea that lesbianism is an anatomical or physiological destiny. On the contrary, she presents it as positive choice for women that can lead to an authentic exercise in freedom just as much as heterosexuality. Unsurprisingly, she does not produce a theory of lesbian identity, since she consistently presents identity as following from choice and acts, not as preceding action and being.

Part 5 examines the two central roles which patriarchy calls upon women to play: those of the wife and the mother. Romantic love is in itself a dangerous condition for women, Beauvoir argues, because it tends to encourage their dependency on men and discourage them from pursuing their own goals. Marriage she describes as based on an obscene principle, because it makes a duty of what should be a freely given exchange. Her attack on the role of wife is closely linked to the role of women as maintainor of the home – the thankless round of domestic tasks that have by their nature to be repeated endlessly without achieving any final aim are described graphically by Beauvoir, and the illusory joys of the gleaming interior are analyzed with penetrating insights. Motherhood, another concomitant of marriage, is presented as even more problematic. Pregnancy acts as a positive incitement to women to sink passively into a fixed destiny; after the birth, the mother–child relationship is described as a battlefield yielding few positive results for either party. She concludes that the upbringing of children care is unsafe within maternal hands and should be organized within a collective framework. Beauvoir's analyses of the dangers and social pressures of the roles of wife and mother bring sharply into focus her dual insistence on, on the one hand, the weight of social conditioning to which women are exposed, and, simultaneously, individual women's responsibility to resist such pressures. Women are both victims and complicitous.

The final two parts of the book are entitled 'Justifications' and 'Towards liberation.' In part 6, 'Justifications,' Beauvoir considers some of the false routes which women have taken in a doomed attempt to flee from their responsibility by glorying in their imprisonment. One route is that of narcissism, in which the woman cultivates her body in an impossible attempt to be simultaneously subject and object, simultaneously both source of the gaze and object of it. A second false route is that of mysticism, the attempt to achieve sovereign being through an identification with God – also, for Beauvoir, an impossibility. The third route is that of the *amoureuse*, the woman who makes a cult of her love for a man whom she has appointed as her savior. She aspires to authenticity by attempting to live vicariously through him. Beauvoir demonstrates both the philosophical impossibility of this stance and the inevitable destruction that it brings about of the relationship on which the *amoureuse* depends, as she becomes increasingly tyrannical.

Part 7, entitled 'Towards Liberation,' offers a synthesis of women's situation and a portrait of 'The Independent Woman.' Beauvoir shows how women's blocked situation even creates what we think of as the female character. The reader is urged to accept that femininity is a social product – if we can accept this, she argues, then we can change society to change the possibilities open to women. She insists on the importance of economic independence for women, and pleads for men and women to work together toward a better future, but she has no confidence that anything of substance can be achieved without a radical restructuring of society. The independent woman is

not in a position to achieve much without the socialist revolution that Beauvoir assumes, in 1949, would create a society in which everyone would be free. This rather unsatisfactory ending is indicative of the exploratory, theoretical rather than campaigning spirit in which Beauvoir wrote the book; it was not until over 20 years later, in the context of French second wave feminism, that Beauvoir declared the necessity of actively pursuing the women's struggle independently of any broader political movement.

On its publication in French in 1949 the book caused moral outrage; the Catholic novelist Francois Mauriac, led a newspaper campaign to have it banned as pornography, and it was in fact put on the Index. Its status as a daringly sexy book written by a woman seems to have partly obscured its political thrust, and the French postwar climate was particularly inimical to the language of women's rights. In America the book was translated in a truncated form by the zoologist Parshley, and it was this version, published in 1953 and still the only English language version available today, which a generation of American and English women including Kate Millet, Betty Friedan, Germaine Greer, and Anne Oakley read during the 1950s and 1960s. Through them the work began to exert its influence, although they did not always acknowledge their debt. Betty Friedan, for example, whose book *The Feminine Mystique* (1963), is often considered to have been the launch pad for the American women's movement, offered her readers a watered down version of Beauvoir's analysis of the housewife and has only one mention of her in the index, which refers to the charge by an American critic that Beauvoir 'didn't know what life was all about' and 'besides she was talking about French women.' Where Beauvoir offers a radical critique, insisting that the institution of the family must be destroyed, that women must become economically independent, and that a radical restructuring of society will need to take place, Friedan advises women to think about their personal growth. However, it is atleast arguable that in translating some of Beauvoir's analyses into terms acceptable to a more conservative audience, Friedan did act as a channel of communication between two cultures in which political radicalism historically have played very different roles.

The key phrase 'one is not born, but rather becomes, a woman' is, as Haraway (1991) writes, at the root of 'all the modern feminist meanings of gender' and has been developed in many different directions. Butler (1986), for example, in a reading of the phrase, stresses Beauvoir's use of the verb 'becomes,' arguing that it allows for a process of gender acquisition that is a daily act, an interplay between individual choice and acculturation of which 'styles of the flesh' are a constitutive part. This important article prepares the way for some of the influential theoretical texts which Butler has gone on to write in the 1990s: *Gender Trouble* (1990) and *Bodies That Matter* (1993) propose a performative theory of gender which build on and develop Beauvoir's original insight. Second-wave feminism in France had a complicated relationship with Beauvoir. The construction of 'French feminism' in the Anglo-Saxon world as being constituted uniquely by the theorists of difference, Cixous, Irigaray, and Kristeva, whose theories are at odds with Beauvoir's stress on individual agency and freedom, has tended to obscure the work of other French feminists such as Christine Delphy and Monique Wittig working in Beauvoir's radical tradition. Feminists reading *The Second Sex* 50 years after its first publication encounter a number of problems: the text's ethnocentricity, its tendency to regard masculinity as unproblematic, the distaste displayed for the female body, and the devaluation of the maternal are largely at odds with contemporary sensibility. Despite this, there is considerable evidence in the 1990s of a renewed interest in *The Second Sex*, as the feminism of difference finds itself in turn in difficulties. Thus, Moi (1994) writes: 'In an intellectual field dominated by identity politics, *The Second Sex* represents a real challenge to established dogmas: if we are to escape from current political and theoretical dead ends, feminism in the 1990s cannot afford to ignore Beauvoir's pioneering insights.'

The Second Sex, with its confidence in the possibility of individual agency and a freely chosen identity, combined with an awareness of how patriarchal structures are at work in the construction of female subjectivity, offers a third way through the choice between essentialism on the one hand and postmodern dissolution of the subject on the other. *The Second Sex* founded many of the feminist strands of scholarship that are being pursued – in women's history, in philosophy, in literary criticism, in social studies. It offers a model of the multidisciplinary approach that the study of women requires, and has made the discussion and theorization of gender possible. It was not intended originally as a manual of liberation or a political tool, but it became one through the impact it had on others. In her autobiography, Beauvoir describes her own astonishment at discovering, in her forties, that there was a way of looking at women with new eyes; that vision is one which has now become a part of our consciousness.

See also: Gender Differences in Personality and Social Behavior; Gender History; Gender and Feminist Studies in Geography.

Bibliography

de Beauvoir, S., 1943/1978. She Came to Stay (Y. Moyse, R. Senhouse, Trans.). Penguin, Harmondsworth, UK.
de Beauvoir, S., 1945/1978. The Blood of Others (Y. Moyse, R. Senhouse, Trans.). Penguin, Harmondsworth, UK.
de Beauvoir, S., 1947/1976. The Ethics of Ambiguity (B. Fretchman, Trans.). Citadel Press, New York.
de Beauvoir, S., 1948/1952. America Day by Day (P. Dudley, Trans.). Duckworth, London.
de Beauvoir, S., 1949/1986. The Second Sex (H. Parshley, Trans.). Penguin, Harmondsworth, UK.
de Beauvoir, S., 1954/1986. The Mandarins (L. Friedman, Trans.). Fontana, London.
de Beauvoir, S., 1958/1987. Memoirs of a Dutiful Daughter (J. Kirkup, Trans.). Penguin, Harmondsworth, UK.
de Beauvoir, S., 1960/1986. The Prime of Life (P. Green, Trans.). Penguin, Harmondsworth, UK.
de Beauvoir, S., 1990/1991. Letters to Sartre (Q. Hoare, Trans.). Arcade, New York.
de Beauvoir, S., 1998. Beloved Chicago Man. Letters to Nelson Algren. Gollancz, London.
Bergoffen, D., 1996. The Philosophy of Simone de Beauvoir. Gendered Phenomenologies. Erotic Generosities. State University of New York Press, Albany, NY.

Butler, J., 1986. Sex and Gender in Simone de Beauvoir's 'The Second Sex. Yale French Studies. pp. 35–49.
Fallaize, E. (Ed.), 1998. Simone de Beauvoir. A Critical Reader. Routledge, London.
Haraway, D.J., 1991. Simians, Cyborgs and Women. The Reinvention of Nature. Free Association Books, London, 131.
Lundgren-Gothlin, E., 1996. Sex and Existence. Simone De Beauvoir's The Second Sex. Athlone, London.
Moi, T., 1994. Simone de Beauvoir. The Making of an Intellectual Woman. Blackwell, Oxford, UK.
Simons, M., 1995. Feminist Interpretations of Simone de Beauvoir. Penn State University Press, Pennsylvania, PA.

Beck, Ulrich (1944–2015)

Adam J Pritchard[†] University of Central Florida, Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

German sociologist Ulrich Beck (Ludwig Maximilian University in Munich, London School of Economics and Political Science) is best known for his work on risk, globalization, and reflexive modernization. His groundbreaking book *Risk Society* (1992) sparked international interest, debates, and scholarship on the topic of risk. His most recent work has promoted moving beyond 'methodological nationalism' by developing an innovative social science made for a cosmopolitan twenty-first century.

Ulrich Beck is a German sociologist best known for his pioneering work on the concept of risk within the context of globalization and reflexive modernization. Beck's most influential work *Risk Society* (1992) presents a theoretical conceptualization of 'risk' as an organizing force in what he terms the 'risk society,' which has emerged as the values and institutions of early industrialization become increasingly self-critical at the same time when modern technologies unleashed potentially catastrophic, unintended consequences. In his subsequent work, the idea of a global 'risk society' is situated within a framework of 'reflexive modernization' wherein the social structures and institutions of the twentieth century come to be characterized by increasing complexity and uncertainty rather than fulfilling modernity's promise of scientific certainty about the natural and social world. This perspective argues that the structural and institutional configurations of early modernity (e.g., gendered division of labor, social class, nation-states) are becoming increasingly detached from individual experiences, giving rise to a plurality of idiosyncratic individual-level options for organizing social relations and managing risks. At the same time, the 'hazards' created by the forces of modernization such as pollution, climate change, and nuclear and biological technology are becoming increasingly global and not bounded by national borders or existing social categories. As such, these global risks in this second modernity cannot be adequately addressed by the institutions of nation-states as they are currently constituted. Beck's work has long highlighted these social changes and the crisis of reflexive modernization, proposing epistemological and political solutions for understanding and managing emerging global risks by advocating a more cosmopolitan vision for addressing the consequences of a global risk society.

Biographical Information

Ulrich Beck was born in 1944 in Stolp, Pomerania and grew up in Hanover, Germany. Beck earned his PhD in Sociology from Ludwig Maximilian University in Munich in 1974, and was awarded Habilitation in 1979. After working as professor of sociology at the University of Münster from 1979 to 1981, and at the University of Bamberg from 1981 to 1992, Beck worked as professor of sociology at Ludwig Maximilian University in Munich until 2009 when he became professor emeritus. He has also held the British Journal of Sociology Visiting Centennial Professorship at the London School of Economics and Political Science since 1997. Since 2008 he has been a Fellow at the Centre for Advanced Studies of the Ludwig Maximilian University in Munich, and since 2009 he has held the Senior Loeb Fellowship of Harvard University, and is a senior research fellow at Max Planck Institute, Gottingen, for the study of religious and ethnic diversity. In 2011, he became one of 15 internationally distinguished inaugural chairs in the College of Global Studies at Fondation Maison des Sciences de l'Homme, Paris. His work has earned him honorary doctorates from Sofia University St Kliment Ohridski and Varna Free University in Bulgaria, the University of Buenos Aires in Argentina, the University of Lausanne in Switzerland, the University of Eichstätt-Ingolstadt in Germany, the University of Madrid in Spain, the University of Macerata in Italy, and the University of Jyväskylä in Finland.

He has received numerous prestigious national and international awards including recognition by the German Sociological Association in 2004 for contributions to public sociology, the prestigious Schader Prize from the Schader Foundation in 2005, and a German–British Forum Award (along with Anthony Giddens) in 1999 for contributing to international dialogue between the two nations. His books have been translated into more than 35 languages, and his work has been cited and expanded upon by scholars in as many nations. At the time of writing this article, *Risk Society* has been cited over 20 000 times according to Google Scholar. The small selection of 69 of his journal articles that are indexed in Web of Science have been cited over 1600 times, including 248 citations in 2012 alone (the most recent full year indexed). From 1999 to 2009, he was the founding director of the Reflexive Modernization Research Center (SFB 536) funded by the German Research Foundation to develop interdisciplinary, collaborative research projects to empirically investigate the forces and consequences of social change. The scientific conversations started by Beck's pioneering works on risk have contributed to international scholarly discourse in a wide range of social science fields including sociology,

[†] After this entry went to press, we learned that Professor Ulrich Beck passed away on January 1, 2015. Thus, this biography was written while Beck was still living.

psychology, criminology, economics, management, and political science.

Most recently, Ulrich Beck was awarded an Advanced Investigator Grant by the European Research Council entitled "Methodological Cosmopolitanism: In the Laboratory of Climate Change," which pioneers research to address methodologies for advancing social sciences beyond the boundaries of 'methodological nationalism.' In this work, Beck is working to develop concepts and analytic techniques which are used to study global environmental risks that are more cosmopolitan and less dependent upon nation-states as the unit of analysis.

Major Works

Risk Society

The English language publication of Ulrich Beck's *Risk Society: Towards a New Modernity* (1992) directly contributed to a proliferation of research into the social and political consequences of risk narratives in the context of rapid social and technological change. In *Risk Society*, Beck argues that forces of rationalization and industrialization gave rise to what he terms 'early modern' (later called 'first modernity') social institutions, but rapid technological change quickly outpaced society's ability to control and check the unintended consequences, or risks, of these technologies. Late or 'second modernity' is characterized by a situation in advanced societies in which most of the basic material needs of citizens can been met, and yet in the process new hazards have been "unleashed to an extent that was previously unknown" (Beck, 1992: p. 19) by the very technologies that gave rise to this level of material comfort.

For Beck, risk is "defined as a systematic way of dealing with hazards and insecurities induced and introduced by modernization itself" (Beck, 1992: p. 21). He theorizes that in contrast to genuinely external hazards (e.g., natural disasters), risks are reflexive, politically driven narratives about the *potential* for hazards to arise from modern institutions or technologies. Further, he argues that the potential hazards of late modern technologies are (1) global in their potential consequences (e.g., nuclear disasters or pollution), (2) difficult to perceive without relying on the expertise of the very institutions that created the hazard (e.g., chemists must test for chemical pollutants), and (3) arise not from a lack of resources as with previous hazards (e.g., famine) but rather from an abundance of unwanted consequences arising from modern technologies (e.g., cancer due to exposure to air pollution).

Using examples primarily based in environmental sociology, *Risk Society* succinctly presents the beginnings of an empirical critique of the prevailing concepts of modernity, discusses the epistemological and political consequences of a truly reflexive social science, and situates the concept of risk and the 'risk society' as an emerging consequence and organizing force emerging from radically reflexive modernization processes in the industrial West.

Reflexive Modernization

Beck's critique of modernity, in many ways, paralleled and complemented the work of renowned British sociologist, Anthony Giddens. In 1994 Ulrich Beck, Anthony Giddens, and Scott Lash published *Reflexive Modernization: Politics, Tradition and Aesthetics in the Modern Social Order* and wed their theoretical critiques of rationalization, modernization processes, and the role of social science into a unified and influential theoretical work.

Reflexive Modernization argues that the driving principles behind modernization arising out of the Enlightenment (e.g., rationalization, science) which helped legitimate and give rise to the institutions of industrial and early modern societies have turned these same critical powers in upon themselves. The modernization process, they argue, was so successful that these well-developed critical technologies, in particular social science, began to shift their critique from deconstructing premodern institutions (e.g., religion, cultural tradition) to reflexively deconstructing the institutions to which they had themselves given rise (e.g., sexual division of labor, social class).

The political and social consequences of this reflexivity have resulted in a fundamental shift in late modernity to a world that is increasingly organized around government's ability to manage individuals' perceptions of risk, in contrast to the early industrial society organized around conflicts over access to material goods. In his contribution to this volume, Beck argues that a microlevel consequence of the risk society is the 'individualization' of risk, such that each person becomes radically disembedded and forced to develop idiosyncratic biographies based on an abundance of conflicting options. One of Beck's major points in this work is that political institutions within a self-reflexive society begin to cannibalize themselves, undermining the ability of individuals to trust institutions that are constantly being destroyed and reinvented by modernization processes. Empirically, *Reflexive Modernization* engages these political processes by examining the externalities of technology such as ecological hazards, highlighting the confluence of political, technological, and social technologies around potentially catastrophic environmental risks.

The Brave New World of Work

Beck's extensions of the ideas in *Risk Society* (1992) to the areas of work and political economy have also been influential; contributing to academic debates over precarious work and globalization. In *The Brave New World of Work* (2000a), Beck details the consequences of a transition away from class-based social relations to a social order organized around a risk society. He locates the historical class-based conflict between capital and labor within the boundaries of nation-states, recognizing that multinational corporations are operating increasingly independently from the restrictions of organized labor, or becoming 'deinstitutionalized.' The central thesis in *The Brave New World of Work* (2000a) is that a world risk society is responsible for giving rise to a proliferation of precarious work, where full-time employment is being replaced by contingent and flexible working biographies that are pieced together by individuals in constant states of transition and self-reinvention. Beck calls for the creation of a cosmopolitan civil society of 'world citizens' to organize work and human activity around transnational interest groups rather than nationality.

Other Works

Following the introduction of the key concepts in *Risk Society* (1992) and *Reflexive Modernization* (1994), Beck's work has continued to build upon, develop, and empirically investigate these ideas. In one sense, *Ecological Politics in an Age of Risk* (1995) and *World Risk Society* (1999) more completely develop the political, environmental, and sociological ideas while building an empirical case for the risk society. Other work has attempted to broaden the theory to explain individual and household-level consequences and evidence for the emergence of a risk society, for example, microlevel investigations of risk society concepts from the perspectives of work, family, and intimate relationship among multinational Europeans are taken up in *The Normal Chaos of Love* (1995), *Individualization: Institutionalized Individualism and its Social and Political Consequences* (2002), and *Distant Love* (2014), all coauthored by Elisabeth Beck-Gernsheim. *Individualization* (2002), in particular, expounds theoretically upon the shift from institutional to individual responsibility for managing risk in one's own life, explicitly linking some of the risk society's key tenets with other sociological work on identity politics.

Though not as widely influential as *Risk Society* (1992), *Individualization* (2002) outlines the microlevel and cultural consequences of reflexive modernity. Individualization as described here is a process where individual identity becomes radically decoupled from traditional social structures like work, family, geography, or even social class. This concept is made particularly clear through examples like gender and work, where trade-offs for women force them to choose between traditional institutions of work and family by examining the political and social risks of an increasingly abundant array of viable biographical choices. The connection of a theory emerging social conditions (e.g., the risk society) to empirically verifiable transformations in gender, work, and family configurations is the primary sociological contributions of this work.

The Reinvention of Politics: Rethinking Modernity in the Global Social Order (1997), *World Risk Society* (1999), *Power in a Global Age: A New Global Political Economy* (2005), and *World at Risk* (2009) all to a greater or lesser extent take up the practical and political tension of a global risk society through timely, empirical critiques of how nation-states have responded (often inadequately) to the risks and consequences emerging risk narratives around issues including: nuclear energy, political management of natural disasters, bovine-strain encephalitis (or 'mad cow disease'), terrorism, biotechnology, climate change, financial crises, and numerous other contemporary risk discourses. All of these works theoretically ground these contemporary crises and examples in terms of risk society, reflexive modernity, and the need for cosmopolitan political solutions. For instance, in *Power in a Global Age* (2005), Beck calls for a 'self-critical cosmopolitanism,' arguing that cross-national organizations, political parties, and multilateral institutions are necessary to address global economic, environmental, and security issues. *World at Risk* (2009) effectively and concisely draws together the core of Beck's earlier ideas in a concise and focused diagnoses of contemporary social issues, cutting through nationalist political debates about terrorism, climate change, and the global financial crisis.

Finally, in another major line of work, Beck explicitly explores the sociological and political consequences of Europe's continued reliance on early modern institutions, describing the prevalence and dangers of an enterprise based on a 'methodological nationalism' that takes as unproblematic the historical distinctions between countries, especially in Europe. *Power in a Global Age* (2005), *The Cosmopolitan Vision* (2006), *Cosmopolitan Europe* with Edgar Grande (2007), and most recently *Twenty Observations on a World in Turmoil* (2012) are Beck's efforts to build a practical cosmopolitan sociology. *The Cosmopolitan Vision* (2006) most clearly begins to distinguish between 'first modernity' and 'second modernity,' including the notion that many individuals are passive or coerced participants in the dialectical transition to a 'cosmopolitanized' world. Some of Beck's most influential and well-cited journal articles emerge from this area of research, where he repeatedly challenges sociology to rethink core analytic concepts and categories like class, power, and the 'global flows' of information, money, and popular culture from a new cosmopolitan imagination (Beck, 2000b, 2002; Beck and Sznaider, 2006).

Impact of Work

The global impact of Ulrich Beck's ideas is evident in the expansive proliferation of his work into more than 35 languages. A Google Scholar search for the term 'risk society' returns over 49 000 results; this concept alone has been applied to the study of topics ranging from terrorism to influenza outbreaks, or from electronic surveillance to the study of intimate partner violence. It is difficult to summarize the scope and breadth of Beck's influence in social science, in part, because of his high visibility as a public sociologist in Germany and Great Britain; for instance, he has written a number of editorials for *Die Zeit* in Germany and *The Guardian* in the UK. He has also been praised for his efforts to promote his cosmopolitan vision for sociology through international, interdisciplinary, and intergenerational collaborations (Calhoun, 2010). Ulrich Beck is perhaps the most famous contemporary sociologist in Europe, where his ideas have struck a chord within both the social sciences and the general public.

In the more than 20 years since the publication of *Risk Society* (1992), there has been a prolific expansion of work across disciplines and across continents exploring the conceptualization and consequences of risk in advanced contemporary societies. Beck's *Risk Society* was the first major work in sociology to theorize about risk in a systematic and empirical manner, reinventing and extending Max Weber's concept of modernization in a manner reminiscent of other major contemporary theorists like Jurgen Habermas. Some have gone so far as to credit Beck with pioneering a 'risk turn' similar to the 'cultural turn' of the 1970s and 1980s in social science (e.g., Mythen and Walklate, 2006); Beck himself calls instead for a 'cosmopolitan turn' that reinvents the orientation of social sciences for a globalized world (Beck and Grande, 2010). In this effort, Beck seems to have been successful; both inspiring and leading an effort to overcome 'methodological nationalism' in the social sciences (Chernilo, 2006).

Critics of the risk society perspective have largely focused their criticisms on Beck's definition of risk. Anthropologists and some cultural sociologists cite the work of Mary Douglas (Douglas, 1992; Douglas and Wildavsky, 1982), for instance, in arguing that risk is by no means a new social phenomenon. Instead, the cultural construction of risk narratives has long been a mechanism for boundary maintenance and social control within human societies. Along similar lines, other sociologists draw on Foucaultian theories of power and governance to critique Beck's assertion that second modernity is not only dominated by risks, but that this occurs because late modernity is actually more hazardous due to the inability of society to control the unintended consequences of technology. Dean (1999), for instance, argues that risk is simply "a way of representing events so they might be made governable in particular ways" (p. 131). They argue that the failures of neoliberal policies in European nation-states to respond to problems that transcend national boundaries do not necessarily represent an evolutionary failure of modernity, especially when one attempts to apply this perspective to other parts of the world (Calhoun, 2010). These critics problematize the proposed transition to second modernity that hinges upon the inability of first modernity's social institutions to be able to handle emerging risks.

The comprehensive nature of Beck's theory of a risk society, combined with Beck's tendency to incorporate contemporary political examples into his work has helped his work reach a wide audience within the social sciences. Most directly, Beck's work has helped to frame environmental sociology within the politics of globalization. Cross-national debates over hazards such as nuclear technology, chemical pollutants, and biotechnology easily lend themselves to analysis within the risk society framework, and indeed many of Beck's early works and later examples draw upon this scholarship. However, the ideas of the risk society have implications for the study of work and economy, marriage and family, youth, popular culture, labor relations, criminology, psychology, media, sexuality, and health policy to name a few (e.g., see Mythen and Walklate, 2006).

Beyond the risk society, Beck's most influential contribution to the social sciences has been his extensive work toward envisioning a cosmopolitan global social science. Though theorists had taken up the issue of 'methodological nationalism' previously, Beck is widely credited with renewing this now-vibrant debate within the social sciences (Chernilo, 2006). In *Cosmopolitan Vision* (2006) Beck explicitly distinguishes a 'social scientific cosmopolitanism' from cultural and political cosmopolitanism, calling for the reinvention of analytic categories developed for a global world. This effort builds upon the critique of what Beck has elsewhere called "zombie categories" (Beck, 2002: p. 27), referring to analytic categories rooted in the social conditions of early modernity (e.g., social class, religion, ethnicity, nuclear families) that have ceased to function yet remain the predominant analytic paradigms. Beck's ideas are visionary in the sense that they persistently seek to transcend academic and disciplinary boundaries in a radical way. If the critique of first modernity is that modern institutions (including social science) are organized around outdated social conditions, Beck's vision for sociologists in second modernity is that they will become active and intentional contributors to political discourse, particularly with regard to the framing and management of risks. The explicit recognition that in a reflexive society, the findings and perspectives of science are already being marshaled by political forces to create or direct public concern is Beck's clarion call to sociologists and other experts that social science must be more proactive in addressing and framing emerging global issues if we are to move beyond a risk society.

Beck has been criticized variously as a pessimist in his diagnosis of modernization (e.g., Braun, 1996; Rose, 2000) by focusing only on the risks to the exclusion of the benefits of modern technological advances, or as exceedingly optimistic about the promise and rise of alternative institutions within second modernity (e.g., Kalleberg, 2013). Some have pointed out that Beck's analysis tends to focus on intangible risks in areas such as work and employment (e.g., job insecurity, unemployment) to the exclusion of more serious immediate dangers like workplace injuries (Tombs and Whyte, 2006). Others have identified limits to Beck's work on developing cosmopolitan political alternatives to state-based strategies, noting that many individuals still only experiences global risks as mitigated through the actions and decisions of the state (Donoghue, 2011). These types of critiques in the context of a global recession and growing economic inequality also raise serious doubts about the extent to which advanced societies have actually developed beyond the point of conflicts over wealth, social class, and material security.

The importance of Ulrich Beck's scholarship within sociology and social science is undeniable. Regardless of his critical reception in specific fields, his works have nevertheless prodded social scientists across disciplines and around the world to consider the consequences of globalization, risk, and reflexive modernization within their own work. The debates led by Beck and his many collaborators and colleagues have helped to dramatically advance a renewed interest in social theory developed for the twenty-first century.

See also: Capitalism: Global; Citizenship, Sociological Aspects of; Class: Social; Cosmopolitanism; Economic Sociology; Giddens, Anthony (1938–); Modernity; Modernization, Sociological Theories of; Political Sociology; Social Changes: Models; Trust, Political; Work, Sociology of.

Bibliography

Beck, U., 1992. Risk Society: Towards a New Modernity. Sage, London.
Beck, U., Giddens, A., Lash, S., 1994. Reflexive Modernization. Polity, Cambridge.
Beck, U., Beck-Gernsheim, E., 1995. The Normal Chaos of Love (M. Ritter, J. Wiebel, Trans.). Polity Press, Cambridge.
Beck, U., 1995. Ecological Politics in an Age of Risk. Polity Press, Cambridge.
Beck, U., 1999. World Risk Society. Polity Press, Cambridge, UK.
Beck, U., 2000a. The Brave New World of Work. Polity Press, Cambridge, UK.
Beck, U., 2005. Power in the Global Age: A New Global Political Economy. Polity Press, Cambridge, UK.
Beck, U., 2006. The Cosmopolitan Vision. Polity Press, Cambridge, UK.
Beck, U., Grande, E., 2007. Cosmopolitan Europe. Polity Press, Cambridge, UK.
Beck, U., 2009. World at Risk. Polity Press, Cambridge, UK.
Beck, U., 2012. Twenty Observations on a World in Turmoil. Polity Press, Cambridge, UK.

Beck, U., Beck-Gernsheim, E., 2014. Distant Love. Polity Press, Cambridge, UK.
Beck, U., 2000b. The cosmopolitan perspective: sociology of the second age of modernity. British Journal of Sociology 51 (1), 79–105. http://dx.doi.org/10.1080/000713100358444.
Beck, U., 2002. The cosmopolitan society and its enemies. Theory, Culture & Society 19 (1), 17–44. http://dx.doi.org/10.1177/026327640201900101.
Beck, U., Beck-Gernsheim, E., 2002. Individualization: Institutionalized Individualism and Its Social and Political Consequences. Sage, London.
Beck, U., Grande, E., 2010. Varieties of second modernity: the cosmopolitan turn in social and political theory and research. British Journal of Sociology 61 (3), 409–443. http://dx.doi.org/10.1111/j.1468-4446.2010.01320.x.
Beck, U., Sznaider, 2006. Unpacking cosmopolitanism for the social sciences: a research agenda. British Journal of Sociology 57 (1), 1–23. http://dx.doi.org/10.1111/j.1468-4446.2006.00091.x.
Braun, J., 1996. Reflexive modernization: politics, tradition and aesthetics in the modern social order [book review]. Theory & Society 25 (5), 752–760.
Calhoun, C., 2010. Beck, Asia and second modernity. British Journal of Sociology 61 (3). http://dx.doi.org/10.1111/j.1468-4446.2010.01328.x.
Chernilo, D., 2006. Social theory's methodological nationalism: myth and reality. European Journal of Social Theory 9 (1), 5–22. http://dx.doi.org/10.1177/1368431006060460.
Dean, M., 1999. Risk, calculable and incalculable. In: Lupton, D. (Ed.), Risk and Sociocultural Theory: New Directions and Perspectives. Cambridge University Press, Cambridge, pp. 131–159.
Donoghue, G., 2011. Global risk and Beck's cosmopolitan politics. Social Alternatives 30 (3), 15–19.
Douglas, M., 1992. Risk and Blame: Essays in Cultural Theory. Routledge, London.
Douglas, M., Wildavsky, A., 1982. Risk and Culture: An Essay on the Selection of Technological and Environmental Dangers. University of California Press, Berkeley.
Kalleberg, A., 2013. Globalization and precarious work. Contemporary Sociology 42 (5), 700–706.
Mythen, G., Walklate, S., 2006. Beyond the Risk Society: Critical Reflections on Risk and Human Security. Open University Press, Berkshire, UK.
Rose, H., 2000. Risk, trust and skepticism in the age of the new genetics. In: Adam, B., Beck, U., van Loon, J. (Eds.), The Risk Society and beyond: Critical Issues for Social Theory. Sage, London, pp. 63–77.
Tombs, S., Whyte, D., 2006. Work and risk. In: Mythen, G., Walklate, S. (Eds.), Beyond the Risk Society: Critical Reflections on Risk and Human Security. Open University Press, Berkshire, UK, pp. 169–193.

Relevant Websites

http://www.opendemocracy.net/author/ulrich-beck – Ulrich Beck's contributions to 'openDemocracy.'
http://www.theguardian.com/profile/ulrichbeck – Ulrich Beck's editorials in 'The Guardian' (UK).
http://www.lse.ac.uk/sociology/whoswho/academic/beck.aspx – Ulrich Beck's faculty profile at the London School of Economics.
http://www.ulrichbeck.net-build.net/ – Ulrich Beck's professional Web site.

Becker, Howard (1928–)

Suzie Guth, Strasbourg University, Strasbourg, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Howard S. Becker's intellectual career can be summarized under five headings. The first addresses his background and intellectual influences. The second, his introduction of 'perspectives' as a tool to understand the medical academic world and public school teacher's relation to social status and career. His best-known work *Outsiders* relates to drug use, the process of becoming an outsider and the labeling theory, this is his third center of interest. His fourth interest is reflected in *Arts Worlds* published in 1982, which compares the professional artistic world to the mavericks and outsiders. Finally, some of the tricks of the sociological trade and cases end this article.

Intellectual Influences

Howard S. Becker was born on the 18 April 1928 (Chicago, Illinois), he was a scholar in what is now called the second Chicago school, although several sociologists (e.g., Jennifer Platt) deny that interpretation of the Chicago department of social sciences after World War II (Platt, 1995). In many interviews H.S. Becker alludes to the influence and friendship with Everett Hughes and to the influence of G.H. Mead's writings in his analysis of the self in the interaction process. He had been a student in criminology with E.W. Burgess. Like Saul Alinsky he does not really acknowledge the courses on criminal behavior, nevertheless he wrote an introduction to the *The Jack Roller: A Delinquent Boy's Own Story*. In criminology he more often mentions Donald R. Cressey and Edwin H. Sutherland and their work on white-collar crime. The anthropologist Lloyd Warner is quoted in his biographical interviews. At the time H.S. Becker was a student at Chicago, Warner was directing the extensive *Yankee City Studies*. This important piece of urban work gave many a student an opportunity to go in the field with other graduate students and led them to reflect on methodology and empirical work. He and Erving Goffman belonged to the same circle of students and quite often he relies on Goffman's image of society as a stage on which actions take place. Becker's MA thesis (achieved in 1949) was first published in a series of articles and then, in 1963, as the well-known book *Outsiders*. His PhD in the Willard Waller tradition (1951) was about Chicago teacher–student relations, it gave rise to many articles gathered in a book published by the London Open University Press in 1995 entitled *Howard Becker: On Education*.

In his interviews Becker often refers to Everett C. Hughes' influence. We could say that both Hughes and Becker had, while in the field, '*a sociological eye*' and often grasped the microdetails, which gave way to an understanding of the whole social setting. While going into a classroom most observers see only the backs of the pupils, the lesson going on, and the day-to-day school routine, but Becker would find out the tricks of the trade and while interviewing teachers he learns about questions of social status, race problems, and much more. Like Robert E. Park and E.C. Hughes who invented and made use of '*their own concepts*': bastard institutions, perspectives, student careers, and the back stage of institutions, so did Howard Becker when he began his master's thesis *Outsiders*. While in Germany in 1948, Everett Hughes mentions in his diary that he was a little awkward about it, he calls them *our funny American concepts* and was wondering if they would seem appropriate to German students (Guth, 2010) , and so it was. Howard S. Becker following E. Hughes' example and with the anthropologist Lloyd Warner learned from the field, furthermore one might say that he was the one who carried off the definition of the situation (W.I. Thomas) in its different aspects either in Medical School or in the collective use of marihuana. One influence, which is often underestimated, is that of William Foote Whyte (Whyte, 1943) who was giving courses to the Chicago students in the late 40s, and had among his students Howard S. Becker and Eliot Freidson. Whyte's pragmatic approach to the field, the people he met and describes, his attempt to theoretical description with set events, and his interaction frame analysis were seen as a good example of what a student could learn from the field. Whyte's approach became, for Howard Becker, a sociological perspective, an ideal of what the field experience is able to stand for. Thus the second Chicago school can be seen as a reflexive school, especially on methodology, grounded theory, and group interaction processes. We should not forget that at the time Becker graduated, the most important book was Moreno's sociometric studies (Moreno, 1953).

In the following we are going at first to see what Becker's study of medical students and medical institutions brought to sociological theories and then referring either to his PhD (Becker, 1951) on the one hand or to the collective and remarkable study *Boys in White* on the other hand (Becker et al., 1995a), we will show how teaching, careers, and students' perspectives illustrate the interaction process. Criminology and deviant behavior will follow with *Outsiders* (Becker, 1963) and his introduction to *The Jack Roller* (Shaw, 1966). Finally, we will end with his lifelong interest in art worlds, music, jazz photographs, and, the questions about field studies, will conclude this article.

Students' Perspectives, Teaching, and Social Status

Students' Perspectives

One of the key concepts in *Boys in White* relates to the *perspectives* that are seen as a collective and comprehensive

meaning of one's future as a physician in Medical School. The perspectives are observed through the learning process: personal time, turning points in (their) apperception of their future, and the final motivations at the end of the learning process in Medical School.

The student's choice of the medical profession is at first an idealistic one and it remains such all along and can still be seen at the end of the process, but it becomes more realistic and less holistic. The Freshmen students want to save lives and cure patients from disease, this is in their opinion the very purpose of the physician. The first problem they encounter in their studies is the huge amount of work and knowledge asked for, and the short time left to apprehend it all. For most of them it seems to be an almost impossible task, so they have to secure a scheme of learning, which will lead to a successful grade. Since the amount of knowledge that is asked for is so huge and time left is so short, they try to reduce the items using different indicators. What is more important and what is less? There are several answers to that question, first they may ask a sophomore and see what the former Freshmen students have been doing, this can be seen as an answer to the issue, an other way of answering that question is the comparison with the manual looked at as a normative reference, and a third way of answering the question is to try to guess what the professors want and what is more important. In these three ways of solving the problem they establish a comprehensive hierarchy of knowledge which they assign to two categories: what is making the grade, and what will be useless. Of course with these indigenous categories they closely examine the former exams and add statistical data to their handling of the cognitive hierarchy. This shows the gap between the idealistic and long-term perspectives and the short-term issues with which the students have to deal. The authors call them the *initial perspectives*. As the studies advance the perspectives, which belong to a collective interpretation of the process of social action, change. Students recognize that they cannot learn all that is asked for, they become realistic especially after a *shotgun* and a block examination (p. 108), it was a focus for students discussions, most of them arrived at the following conclusion (p. 111):

1. In spite of all our efforts we cannot learn everything in the time available.
2. We will work just as hard as ever, but now we will study in only the most effective and economical ways, and learn only the things that are important.
3. Some students said: We will decide whether something is important according to whether it is important in medical practice.
4. Other Students said: We will decide whether something is important according to whether it is what the faculty wants us to know.

This is the provisional perspective, the last perspective led them to the idea that their knowledge about medical practice could not be a help in deciding what was most needed and what was not, they finally came to the conclusion that they had to study what the faculty wanted them to know. They study a little less and spend 60 study hours a week, instead of the original 77 (p. 169), they think that they are more efficient in their work compared with the first months of the freshman year.

Teaching and Social Status

Howard S. Becker's PhD unpublished dissertation (1951) *Role and Career of the Chicago Public School Teacher* gave rise to a number of articles which were republished in 1995 by Robert G. Burgess: *Howard Becker on Education*. Becker's past work fits very well with the new sociology of education in Great Britain and, along with Woods, Hammersley, Hargreaves, Colin Lacey, etc., it gives a new impetus and an ethnographic view to the field. When H.S. Becker was working in the field Lloyd Warner was still publishing his Yankee City series. Becker made use of Warner's status scale and was sensitive as was W.F. Whyte, to the middle-class norms and values developed in the settlement house, finding similar references in school culture. The variations in pupil–teacher relations are observed through three adjustments: teaching, discipline, and moral acceptability. He distinguishes in the different classes a lower–lower stratum, a middle-class stratum, and an upper-middle-class stratum. He looks first at performance which is quite different for middle-class students and lower class students; the latter have another scale of prestige and reward in the slum areas in which they live. In these areas children do not have good working habits, they do not seem interested in learning. Therefore, the teachers in such neighborhoods know that some of the kids are just waiting to quit school; in primary school, teachers cannot do what is expected. For example, at K (a good school) they are able to have nine words a week in spelling although in D (slum school) they are only able to achieve six words. Children have poor reading ability in the poor neighborhood and the teacher has to deal with a lack of ability, that is why the level of expectation is not very high and teachers are satisfied if the children can read and do simple sums (p. 35). Discipline is difficult in the slum schools, there is always a latent tension and one has to be on the edge in order to prevent violence and aggression. In lower class status groups physical aggression is part of a daily behavior among males, it is a behavior approved of and sometimes encouraged in those areas. That is why teachers have to be able to deal with such hardships while teaching. Teaching in middle-class areas where pupils are docile and eager to learn can be a real pleasure, it is not the same with lower–lower classes nor with upper–upper classes. The children of higher status classes can be hard to handle they are seen as spoiled and overindulged, they do not take orders at home and would not take them at school; that is why some teachers begin their class year being tough the first day, but they have to be consistent over the year.

The third item of moral acceptability refers to the group status. The upper class kids have more often learned politeness and respect for their elders while the others lack cleanliness, hygiene, and moral habits; some girls are prostitutes on the side. Above all, the lower class students do not approve of education, they do not want to learn and cannot see the use of being in school, in that sense the moral values of the teachers (individual striving, success, and social mobility) are denied.

One of Becker's best papers in 1952 relates to public school teacher's careers. What happens in the long range? Which turning points in their career might be found? Do public school teachers become principals when they grow older? Do they go

from school to school, from low-status school to higher status schools? Becker calls those changes, adjustments in a career, in a hierarchy of ranked positions, but most of the Chicago career movements are horizontal, at least during a certain period, the teacher keeps teaching maybe for as much as 20 years. Among the teachers examined, the main movement was between the various schools in the Chicago district. Under the Chicago Board of Education Rules teachers had to stay a whole year in order to apply for a transfer, they could apply to as many as 10 positions. If an opening occurred the principal had to approve of the applicant. The young teachers began their careers in the less desirable schools and then moved to more desirable schools, mostly in better surroundings. The teacher's movement was a movement from the least valued surroundings to a better neighborhood or nearer to one's home. But the nicer schools may not have been what they seemed, an authoritarian principal and the school might be worse than the one the teacher had just left, some are just *faux amis*. The principal's choice could be a personal choice, if he did not want the new applicant, he let them know saying that the new teacher would not be happy and this might be a self-fulfilling prophecy. So it seems one had to learn what is hidden behind a vacancy, via an understanding of the neighborhood, the personality of the principal, the social status of the students. All those elements had to be given some consideration in order to make a move. After the first move, many thought they would stay a while in the neighborhood before making a decisive move nearer to home.

Howard S. Becker draws attention to the different changes that take place in the first or second school in which the teacher works. At first the teacher may get adjusted to this low-class environment, his skills and his comprehension of the situation are developing, he becomes acquainted with his colleagues, some might even become friends, he is developing his social system within a social system. His expectations, his perspectives become more realistic that explains why a small number of teachers spend their whole career in only one school. The changes that occur in a teacher's career are linked to their expectations, a young teacher wants to teach an enormous piece of knowledge but he will not be able to do this, a little later he understands that he will never be able to teach such a vast amount of cultural items, he lowers his level of expectations in order to meet the childrens' abilities. While such adjustments are operating, the teacher belongs more and more to a professional network and to the neighborhood network, it becomes more and more easy to have relations with the children's parents, as time passes the teacher becomes known in the neighborhood. In American cities such as Chicago, neighborhoods are changing, with new inhabitants, the movement of more recent immigrants into the transitional area. One of the teachers interviewed says: *The neighborhood used to be ninety 95% Jewish. Now I do not think there are over 40%Jews. The rest are Greek, Italian a few Irish, it's pretty mixed now. And the children are not that nice as they used to be.* So staying a long time in the same school may not be the best policy, in the example quoted, the school changed, its pupils were no longer the same and their parents had perhaps more problems understanding English and the way of life of an American School. The problem might be much more serious in the slum district where discipline is difficult to maintain. If discipline was kept under control by a principal with authority, if he retired or if he went to another place, the whole school might change: a lack of discipline may induce a high turnover. The new principal may want to have his own way and, on the one hand, small privileges based on common understanding between teachers might disappear when the principal leaves. On the other hand, it can be a new opportunity for young teachers who felt that they were on the periphery of the school, with a new principal they strive to be at the center of the institution. Becker wondered why so few were willing to take the exam for principal and found that the answer lies in the horizontal mobility. A teacher goes from a poor neighborhood to a better one, each change is a success in the mobility process, but if he succeeds in his principal's exams, he has to go back to were he came from: *at some low class colored school (Those schools are nearly always vacant, so you have a very good chance of being assigned there when you starts in)*. There is an upward mobility, but at the same time one may go back where one has started, that explains why many do not try to take the principal's exams.

Labeling Theory, Drugs, and Addiction

The epigraph of '*Outsiders*' quotes William Faulkner's I: *It ain't so much what a fellow does, but it's the way the majority of folks is looking at him when he does it* this is the major argument of the book. In a sense it is quite a Simmelian point of view with reciprocal interactions, what W.F. Whyte called: a set event. "I am labeled in the eyes of others," it is an interaction between the group and the individual seen as a law breaker. When H.S. Becker was writing his MA it was still the golden age of group studies. Deviance is, among others, a labeling process – one is deviant when one is seen such although there are social factors which put blame on one and not the other. An unmarried mother is a deviant, which it is not the case of the father; the African American is more often seen as an offender compared to the white American. Some of the rules belong to a single social group, if a Jew breaks the rules of the kashrut it has no effects on the rest of the society; some offenses are limited to a community and Robert E. Park saw deviant behavior as limited to a specific urban area.

The labeling process is a sequence of social action with steps leading to deviant behavior:

1. Marihuana is in use in the given social situation.
2. A personality variable is now at stake, the given person will use marihuana with the other members of the group, he might be even unaware of the lawlessness of his behavior.
3. Even if the first experiences might not be so positive, he goes on using it. The group is showing him how to use the dope and how to get it.
4. He is now a member of a deviant group in which he learns to carry on the deviant activity with the less difficulty. He has now performed the deviant career, most of the deviant careers are alike such as thieves (E. Sutherland) they need a social milieu in which deviance take place.
5. The last process is rationalizing. Up to that moment the drug user is engaged only with people having the same habits or having the same behavior. The marihuana user neutralizes conventional attitudes, and shows that his behavior and his addiction are perfectly normal and manageable; one finds out that he is able to work even while

being high. Drug use is no longer a form of deviance, it is a way of enjoyment and is even much cheaper than alcohol. The marihuana user is almost sure that there is no addiction, it differs in that sense from opium and so he believes that he may stop at any time.

In the Festschrift in honor of Everett C. Hughes *Institutions and the Person: 2010*, Becker compares the user of marihuana to the user of lysergic acid dithylamide (LSD). Becker wants to demonstrate that the LSD psychosis issue might be a wrong interpretation of the drug symptoms. Getting high, too high is for the newcomer in drugs associated with going crazy and might be followed by a reaction of panic. Becker compares the group of marihuana users in which the effects of drugs are socialized and explained to the new user, so he understands what is happening to him and what to do to prevent becoming *too high*. The single user of LSD who is not surrounded by a group of users, and without any drug culture, might be frightened by the effects of the drug and have a panic reaction, often associated, as Becker suggests, to psychosis. Such an interpretation through his historical review of articles was given to marihuana: he has listed interpretations of marihuana use from 1920 to 1940 when it was at first introduced, up to 1940 when it has spread, the psychosis explanation was commonly used, and then, when drug use became more common, it disappeared. Will this be the fate of the comprehension of the effects of LSD? At that time, unlike oriental drugs such as marihuana and opium, which have led to a drug culture and a set of surroundings for drug use, and groups of users, the acid users might be increasing, but they show a different relation to the drug. In both studies *Outsiders* and the comparison of marihuana users to LSD users, Becker insists on the relation to the group of users compared with the individual user. The former involves a socializing process and shares with the newcomer a positive identity and a process of group rationalization while the latter is left alone. Once adopted, the identity changes especially with friends, parents, and colleagues, norms are now at stake: the marihuana user becomes a dope fiend. The LSD user remains a lonesome user of acid, although in both cases users have to find ways to get the drugs.

In a chapter written in 1970, about drug use on campuses *Campus Power struggle: 1973* Becker takes a provocative turn, saying that the drug problem on campuses is not a drug use problem because more and more students are engaged in such a process, the campus problem is a problem of students getting caught while selling drugs or giving drugs to one another, or using drugs in the open. In his opinion, which he repeats several times, the use of marihuana is harmless and one of his proposals is to teach students how to take drugs secretly without getting caught. His provocative position led to several letters from psychiatrists and physicians saying that the results of experiments and clinical observations were not clear about the harmlessness of marihuana, and the problem of addiction was still an open question.

Art Worlds

It seems that *Art Worlds* is Becker's hidden sociological interest, he has written that at first he wanted to be a musician and then he became a sociologist. This book (Becker, 1982) is related to his first achievements in music and to that lifelong passion. The world of arts is a world in which the borders are not very clear, the differences between what is art and what is not art is quite difficult to draw, especially with the motion pictures. Becker invites us to look at the artistic process, the craftsmanship needed, the division of labor, which exists as in any other profession. Artwork is a whole world not only in the case of the symphonic orchestra, which involves instrumental techniques and the Opera building where the music is performed. It can be even for a Quattrocento painter a contract between the client and the painter, in which it is said which kind of blue he has to use, if he is allowed to use gold paint and who is going to paint the portraits. With this example in mind we may understand that painting in Holland or in Florence is not only craftsmanship but also belonged to Renaissance capitalism. Art worlds follow standards, norms, and conventions of societies; most of Becker's examples refer either to music, literature, film, painting and artistic photography, and poetry. Even if there are a few references to the Renaissance in Europe, the main topics relate to twentieth-century America. To sum up the author's position: the world of arts is a world of action, characterized by a division of labor, it is a social world, arts and crafts being complementary and exchanges between popular art and arts can be found.

In one of his most quoted chapters on *Integrated professionals, mavericks, folk artists and naïve artists*, which compares integrated professionals, on the one hand and, an American maverick Charles Ives on the other, Ives writes unplayable music that was only performed after 20 years. Much of his music was not finished. An example is the *Concord Sonata*; Ives changed his mind so many times that one may think he is unable to choose between several resolutions. This was certainly a psychological trait of the author, but with this example Howard Becker wants us to understand that the creative process needs *form*. As Georg Simmel (Simmel, 1918) had already mentioned, form is only part of the process in the art world, art works have still to go through conventions, editing, and performance in order to exist as such. The *salon des refusés* (exhibition of rejects) in painting gives a more immediate view of its achievements compared to music, which needs to be performed in order to exist.

Folk and naïve artists do not belong to the art world, but some of their works might become examples of aesthetic achievement. Quilt making is craftsmanship and belongs to American community work and American community togetherness. Some quilts are in native craft museums, but some seem to belong to the same art movement as nonfigurative painting. Where does the design come from? No one knows, it seems to follow channels of women's gatherings and visiting and moves from one community to the other. Certain geometric designs have names such as 'drunkard's path,' 'mill wheel,' 'log cabin' all tend to have an optical effect with small pieces of material cut the same size and stitched on a frame. Some of those quilts have in Gestalt language a good form but they always belong to the applied arts. This comparison shows the distance with the conventional art world although some quilts could enter in one of Raymonde Moulin's serial art categories (Raymonde Moulin, 1992). Idiosyncratic style is even more relevant in the Naïve art form of Grandma Moses' or *Facteur Cheval*. Naïve art, such as

the Watts Towers in Los Angeles, or the Ideal Palace of the postmaster Ferdinand Cheval in Southern France, is a totally individual performance. Art is performed as a hobby, as a passion, even as an obsession but without any knowledge of craftsmanship. With stones, garbage pieces, junk, the artist is giving a second life to cans, bottles, pebbles, etc. Thus a new usefulness is given to these items, the stones of Facteur Cheval now belong to a very popular castle. Naïve art has developed in many different directions, it is always far from professional training and art conventions, it may however enter the art world if an artist or a journalist is interested, although the gate is always narrow.

Howard Becker takes a comparative stance in order to answer the question what is art and what is not art. We understand that the art world is above all a technical and professional world, but it can be interested in applied arts and naive artists. Being as Charles Ives on the margin of art world, it is difficult, especially in music, to fully enter the art world and it needs a process of recognition. Those who do not belong to the world of arts such as the quilters or the naive artists maybe recognized by chance, as long as their artwork belongs to the *Zeitgeist*. Finally, Becker comes to the conclusion that art worlds can be compared to any occupational world in which changes may occur. Changes occur in Arts and in the World, rock and roll music changed not only teenagers' music but created new possibilities for music in general. One of the last questions the author asks is: What is lasting? The answer to that question cannot be simple and is usually formulated by historians and archaeologists.

Cases

What is a case? Becker answers the question in writing the introduction to the Jack Roller's own story (Shaw, 1966). A life story is a piece of a mosaic, related to many other stories such as Chicago's West Madison area which is related to N. Anderson's stories about hoboes, the high life. The boys experience is also related to Walter Reckless's *Vice in Chicago* (Reckless, 1969) or to Paul G. Cressey's *Taxi Girls* (Cressey, 1969). When Stanley goes to prison the reader thinks about Saul Alinsky who worked several years with youngsters in jail and learned to know the members of the gang he was to study later and to handle the specific language of the gangs (Guth, 2013). Becker relates the Stanley's experience of prison to the ceremonies of degradation described by Erving Goffman (1961). In that sense Chicago's sociology was a study showing the mosaic of the urban areas.

To sum up this work, one can understand that a case is only one element of the social lives lived in the slum areas and it should be compared with other cases and especially with negative cases. Thus the researcher would be able to apprehend the processes of causation and life action. The labeling theory mentioned above has shown how from one step to the other, a new behavior, identity, group control, and conformity induces personal and social changes. Biographic material is the only way of demonstrating how time matters for the individual and how the different adjustments occur. The case is of course the person's own story, the story that might be told in different ways and from different points of view. As Becker adds in his conclusion: one of the better services rendered by the life story is a public service, it gives some insight about social problems to the readers, and as such one may understand what the life of a youngster can be in the slums, particularly if he comes from a totally different background.

Howard Becker's work is not limited to sociology and he uses different tricks or artifacts in order to help us understand what society is like. It can be shown in a movie, in a painting, in a photograph, or in a play: sociology is, in a sense, a representation of society. Usually representation is given by statistics, charts, and maps, but since Jacob L. Moreno (1953), one is allowed to view it in terms of characters on stage in an interaction process.

See also: Crime, Sociology of; Culture, Sociology of; Delinquency, Sociology of; Dramaturgical Analysis; Field Observational Research in Anthropology and Sociology; Goffman, Erving (1922–1982); Interactionism, Symbolic; Labeling Theory, History of; Medical Education; Professions, Sociology of; Social Control and Deviance, History of; Social Justice: Historical and Theoretical Considerations; Sociological Theory; Sociology: Overview; Status and Role, Social Psychology of.

Bibliography

Andersen, Nels, 1923. The Hobo: The Sociology of the Homeless Man. The University of Chicago Press, Chicago.
Becker, Howard S., 1951. Role and Career Problems of the Chicago Public School Teacher, PhD. The University of Chicago.
Becker, Howard S., 1963. Outsiders. Studies in the Sociology of Deviance. The Free Press, New York.
Becker, Howard S., 1973. Campus Power Struggle. Transaction Books, Rutgers University, New Brunswick, NJ.
Becker, Howard S., Geer, Blanche, Hughes, Everett C., Strauss, Anselm L., 1995a. Boys in White. Student Culture in Medical School, fifth ed. Transaction Publishers, New Brunswick and London.
Becker, Howard S., Geer, Blanche, Hughes, Everett C., 1995b. Making the Grade. The Academic Side of College Life. Transaction Publisher, New Brunswick (US), London (GB).
Becker, Howard S., 1982. Art Worlds. University of California Press, Berkeley and Los Angeles.
Becker, Howard S., Geer, Blanche, Riesman, David, Weiss, Robert S., 2010. Festschrift in Honor of Everett Hughes. Institutions and the Person. Aldine Transaction, New Brunswick (US), London (GB).
Becker, Howard S., 2008. Outsiders. Studies in the Sociology of Deviance. Simon and Shuster, New York.
Becker, Howard S., 1998. Tricks of the Trade, How to Think about Your Research while Doing It. The University of Chicago Press, Chicago.
Becker, Howard S., 2007. Telling about Society. The University of Chicago Press, Chicago.
Burgess, Robert G., 1995. Howard Becker: On Education. Open University Press, Buckingham, PA.
Goffman, Irving, 1961. Asylums: Essays on the Social Situation of Mental Patients and Other Inmates. Anchor Books.
Guth, Suzie, 2013. Saul Alinsky, Conflit et démocratie locale. L'Harmattan, Paris.
Guth, Suzie, 2010. Everett Hughes' journey in occupied Germany(1948): black market, bastard institutions and dirty work. In: Schrecker, Cherry (Ed.), Transatlantic Voyages and Sociology. The Migration and Development of Ideas. Ashgate, Farnham (GB), pp. 255–266.
Cressey, Paul G., 1969. The Taxi Dance Hall. A Sociological Study in Commercialized Recreation in City Life. Patterson Smith, Montclair, NJ.
Moulin, Raymonde, 1992. De la valeur de l'art. Flammarion, Paris.

Moreno, Jacob L., 1953. Who Shall Survive? Foundations of Sociometry, Group Psychotherapy and Sociodrama, second ed. Beacon House Inc, Beacon, NY.
Platt, Jennifer, 1995. Research methods and the second Chicago School. In: Fine, Gary A. (Ed.), A Second Chicago School? The Development of a Postwar American Sociology. The University of Chicago Press, Chicago, pp. 82–107.
Ragin, Charles, Becker, Howard S., 1992. What Is a Case? Exploring the Foundations of Social Inquiry. Cambridge University Press, Cambridge.
Reckless, Walter C., 1969. Vice in Chicago. Patterson Smith, Montclair, NJ.
Simmel, Georg, 1918. Das problem des portraits. In: Die Neue Rundschau, XXIX jg der Freien Bühne, vol. 2, pp. 1336–1344.
Shaw, Clifford R., 1966. The Jack Roller. A Delinquant Boy's Own Story, first ed. 1930. The University of Chicago Press, Chicago.
Warner, Lloyd W., Lunt, Paul S., 1941. The Social Life of a Modern Community, vol. 1. Yale university Press, New Haven.
Warner, Lloyd W., Lunt, Paul S., 1942. The Status System of a Modern Community, vol. 2. Yale university Press, New Haven.
Warner, Lloyd W., Strode, Leo, 1945. The Social System of American Ethnic Groups, vol. 3. Yale University Press, New Haven.
Warner, Lloyd W., Low, Josiah O., 1947. The Social System of the Modern Factory. The Strike: A Social Analysis. Yale University Press, New Haven.
Warner, Lloyd W., 1959. The Living and the Dead. Yale University Press, New Haven.
Whyte, William F., 1943. Street Corner Society. The Social Structure of an Italian Slum. University of Chicago, Chicago.

Relevant Websites

http://media.pfeiffer.edu/iridener:dss /DEADSOC.HTML – Dead Sociologists Society.
www.sociosite.net/topics/sociologists.php.

Behavior Analysis, Applied

Ruben Ardila, National University of Colombia, Bogotá, Colombia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

One of the main developments of applied psychology during the last decades is applied behavior analysis (ABA). ABA employs basic behavioral processes, research methods, and derivative procedures, in order to prevent and alleviate problems of social importance. ABA is accountable on seven dimensions: (1) its procedures are applied to problems of relatively immediate social importance; (2) its behavioral measures are valid and reliable; (3) its procedures are described in sufficient technological detail for replication; (4) its research methods are analytic; (5) its effectiveness is socially significant; (6) its generality is demonstrated across time, settings, and behaviors; and (7) it is relevant to an overall conceptual system of behavior. This article describes the foundations of ABA, its relations to basic research, and the main areas of application, both traditional and nontraditional: behavior analysis applied to clinical problems; behavior analysis in education; industrial/organizational behavior analysis; community behavior analysis; forensic applications of behavior analysis; sport psychology and behavior analysis; violence, racism, productivity, sexism, drug abuse, and other social issues. Innovative procedures and techniques will be pointed out in this article.

Definition

Applied behavior analysis (ABA) has been defined as the application of the methods, procedures, and techniques derived from the experimental analysis of behavior (EAB) to a wide range of clinical and social institutions. The main assumption of behavior analysis is that behavior is a function of its consequences. Environmental variables are utilized in order to produce behavioral changes. A wide range of intervention techniques has been developed based on the principles of laboratory research, such as stimulus control, reinforcement, extinction, and punishment.

In the context of ABA, the term 'behavior,' is interpreted broadly to include overt responding and also covert responses (cognitions, emotions). Behavior is defined as everything that the organism does or says, in the case of human beings. Behavior is a broad term, with different ways of being interpreted. ABA includes muscular contractions, visceral secretions, cognitive factors, emotional responses, language, and social behavior.

ABA is mainly associated with operant ('Skinnerian') psychology. The term is contrasted with behavior therapy, behavior modification, behavioral engineering, cognitive behavior therapy, multimodal therapy, social learning, and others. The differences are in some cases just a matter of terminology, and there is not complete consensus in relation to the use of each term.

In ABA, the emphasis is on a body of experimental work dealing with the relationships between changes in the environment and changes in the subject's responses. Great importance is given to the experimentally established procedures and principles. The applications to socially relevant issues are considered to be derived from basic laboratory research. Clinical problems are studied by means of experimental research studies.

The model of ABA is psychological rather than medical. The emphasis on behavior (maladaptive in the case of clinical populations) implies that the therapist is not interested in underlying or disease factors that 'cause' the symptoms. In the case of clinical or maladaptive behavior, the term 'behavior therapy' centers on behavior that is observable and definable.

ABA can be traced back to several conceptualizations and experimental strategies. Particularly important is the Skinnerian analysis of behavior, leading to behavioral shaping, token economies, and environmental design. The emphasis upon the application of experimental methodology to individual cases is a distinctive feature of ABA. Social learning, modeling, and classical conditioning procedures are also taken into account.

It is possible to state that the main techniques used in ABA are operant conditioning, extinction, systematic desensitization, modeling, self-control procedures, and techniques for the modification of cognitions – a wide range of procedures that are tailored to the particular individual, the specific situations, and the behaviors to be modified.

From Laboratory to Applications

The natural science from which ABA is derived is behavior analysis. It comprises three subdisciplines: the EAB for basic research elucidating fundamental behavioral processes; ABA for the application of these processes, derivative technologies, and research methods for clinical and community problems; and the conceptual analysis of behavior, for historical, philosophical, theoretical, and methodological investigations.

The EAB involves conceptual analysis as a matter of course, that is, ontological assumptions. The EAB seeks: (1) to describe functional relations between classes of responses and classes of stimuli, and (2) to demonstrate the reliability and generality thereof, and thus their lawfulness. The dependent variable is the probability of an organism's response as a function of independent variables, which occur in real time. (Arantes et al., 2012).

On the other hand, ABA employs basic behavioral processes, research methods, and derivative procedures, in order to prevent and alleviate problems of social importance. ABA has derived from the EAB with their procedural emphases of reinforcement, punishment, and discriminative-stimulus contingencies as behavior-analytic environmental variables, their reliance on single-subject designs as the formats of analysis and proof, and their consistent use of the Skinner box as

their arena (Kazdin, 2000). On the other hand, in the last decades ABA has been characterized by seven key terms: applied, behavioral, analytic, technological, conceptual, effective, and capable of appropriately generalized outcomes.

The importance of a conceptual analysis has been pointed out since the beginning of the EAB. The conceptual analysis includes metatheory and philosophy, history and historiography, methodology, and system and theory.

History of the Area

ABA began in the 1950s. The first decade was characterized by polemics and paradigmatic clashes. The second decade was the time of consolidation. The third saw the development of a refined methodology and new conceptual models. The most recent period has been characterized by applications to new fields, even in the so-called nontraditional areas (Mattaini and Thyer, 1996; Cooper et al., 2007; Madden, 2012).

A very relevant problem refers to the role of conditioning in ABA and in behavior therapy. This area was sometimes called 'conditioning therapy,' and it was considered as the application of the laws of 'modern learning theory.' Discussion has arisen concerning the differences between conditioning in the animal laboratory, in the clinic, and in daily life. Conditioning as the explanatory foundation of all ABA has been very controversial. Some authors have insisted on the need to include knowledge drawn from psychobiology and social psychology, instead of relying only upon conditioning-learning theory.

The role of cognition in ABA has been a central topic of discussion. Due to the desire to eliminate all reference to mentalism and to inner processes, early behavioral analysts rejected all forms of cognitive influence. But in the 1980s the 'cognitive revolution' led to some reformulations and reconceptualizations, and there was discussion about the role of cognition in behavior change. The term 'cognitive behavior therapy' was used, rather than just behavior therapy. However, if behavior includes cognitive processes, the addition of 'cognitive' is redundant. The controversy continues during the second decade of the twenty-first century, and several positions and disputes exist concerning this issue. The precise relationship between behavior and cognition remains equivocal (Ardila, 2010).

Techniques

A large number of techniques have been developed in the area of ABA. Some of them are based on operant conditioning, others on classical conditioning, social learning, or a combination of them. Some involve principles derived from social psychology, developmental psychology, physiological, and cognitive psychology. Only the main techniques is presented here, and the list is not exhaustive.

Systematic Desensitization

This is a term applied to a class of methods used for gradually weakening maladaptive anxiety response habits through the use of responses that compete with anxiety. As an example, a physiological state inhibitory of anxiety is produced in the patient by means of muscle relaxation; the patient is then exposed for a few seconds at a time to a stimulus arousing weak anxiety. With repetition of the exposure, the stimulus progressively loses its ability to evoke anxiety. In a progressive way, stronger stimuli are presented.

The standard procedure in systematic desensitization consists of several sets of operations: (1) introduction to the subjective anxiety scale, (2) training in deep muscle relaxation, (3) construction of anxiety hierarchies, and (4) counterposing relaxation and anxiety-evoking stimuli from the hierarchies.

Counterconditioning is the basis of the change that follows systematic desensitization, according to Wolpe (1958), other researchers and clinicians.

Biofeedback

This term refers to a number of clinical techniques, which provide feedback to patients concerning changes in a particular physiological response, so that they can learn to modify the physiological response. Two methods of providing information feedback to the patient have been developed: binary feedback (yes/no information) and analogue feedback (proportional). A number of physiological responses have been shown to be modifiable using biofeedback. The most common physiological responses to be used clinically are electromyogram activity, skin temperature, blood pressure, electro-encephalogram, vasomotor, and heart rate.

Biofeedback is used in lowering the blood pressure of essential hypertensives, for achieving relaxation, for epilepsy, migraine headaches, and many other clinical problems.

Punishment

Punishment is an operant conditioning procedure in which the future probability of occurrence of a response is reduced as a result of response-dependent delivery (positive punishment) or removal (negative punishment) of a stimulus (punisher). An example of positive punishment is spraying water mist in the face of a self-injurious child each time an instance of self-injurious behavior is observed. An example of negative punishment is taking an earned token from a person each time he or she curses. In the first case the aim is to reduce the occurrence of self-injury; in the second case it is to reduce the rate of cursing.

Punishment may weaken, strengthen, or have no effect on other response classes. Its effects are controversial and subject to discussion. In the early stages of ABA, punishment procedures were used more frequently than now. Alternatives to punishment are usually preferred.

Escape Training

In this training procedure an individual learns to perform a response in order to terminate a punishing stimulus.

Avoidance Training

This is a training procedure in which an individual learns to keep away from a designated stimulus in order to prevent

punishment. The first step consists in pairing the designated stimulus with a punisher. Then the individual is negatively reinforced by performing a response that terminates the punishing stimulus. The individual subsequently learns to prevent punishment from occurring by either responding with an alternative behavior (active avoidance) or not responding at all (passive avoidance) prior to coming in contact with the designated stimulus.

Extinction

Extinction is a response-weakening procedure, which can involve consistent failure to deliver a reinforcer following an operant response, which had previously produced that reinforcement (operant extinction). It eventually results in nearly complete elimination of the operant response, although a transient rate increase is frequently observed soon after operant extinction is arranged.

This technique is widely used in behavior modification and is preferred to punishment, avoidance training, escape training, and similar aversive procedures.

Shaping

This technique consists of reinforcing successive approximations to the desired response. Usually a simple response is required initially, and criteria for reinforcement are gradually made more stringent so as to produce more complex or refined response. Initially a simple response (in the direction of the desired goal) would be sufficient to receive reinforcement. After this behavior is performed reliably, reinforcement is given only for a more complex or difficult response. The next step consists of giving reinforcement for the slightly more complex response. The pattern continues until the final desired behavior is achieved regularly.

Contingency Management

Contingency management involves the analysis and change of the functional contingencies in the environment that determine a person's behavior. As a result, the behavior changes from its initial rate to a different rate of occurrence. In clinical work, this involves the development of new or alternative forms of adaptative behavior. Within a certain environmental context, a contingency describes a functional relationship among stimuli antecedent to a particular behavior (the target) and the consequences that follow the occurrence of the behavior.

The functional environment for a particular person is often different from the functional environment of another person. Many stimuli or events that occur in a person's environment do not have any effect on the person's behavior. In those cases we can say that the particular part of the environment is not functional, it is just present.

Contingency management involves the analysis and modification of contingencies involving discriminative stimuli, reinforcement, punishment, and extinction. It focuses on public and observable behaviors and controlling environments. The environment of interest is primarily that part which functions to increase, decrease, or control the occurrence of certain behaviors. Contingency management in clinical practice typically is concerned with appropriate discrimination and generalization of adaptative behavior changes.

After a comprehensive functional analysis of environmental contingencies, the therapist plans a program of intervention to modify contingencies to change the rate of the target behavior. Contingency management involves a thorough functional analysis of controlling contingencies, the rearranging of the functional analysis of controlling contingencies, the rearranging of the functional environment, and a careful monitoring of outcome.

Environmental Design

This consists of the application of empirically derived principles of behavior for the modification and design of the environment. It involves the planning of a coherent program and set of procedures to affect the total human and nonhuman environment in ways that increase the probability that certain goals will be achieved.

The goal of environmental design relates to social behavior, for instance, planning a therapeutic or educational system. It includes linking ABA and environmental psychology. It involves training people to conceptualize the environment so that they can apply the general principles of behavior analysis to design planning. Social design and physical design should be considered. The larger context is the planning of a whole society, such as *Walden Two* (Skinner, 1948), or *Walden Three* (Ardila, 1990).

Token Economy

In the token economy intervention, reinforcers are delivered for specific target behaviors. Reinforcers are tokens (tickets, coins, points, stars) that can be exchanged for a variety of other rewards. Tokens help bridge the delay between the person's performance of a desired behavior and delivery of a reward.

As an example of the use of token economies in complex settings we can mention the University of Kansas Experimental Living Project, a behaviorally managed experimental community of 30 college students (Johnson et al., 1991). Token economies have also been used in prisons, in nursing homes, on psychiatric wards, in normal classroom settings from preschool to university classes, on wards for the treatment of drug addicts and alcoholics, in various work settings to increase safety behavior, in community settings to increase recycling of waste and improve energy conservation, and in many other contexts.

Premack Principle

This principle says: for any pair of responses the more probable one will reinforce the less probable one. The technique derived from the Premack Principle consists of a special class of reinforcing operations in which access to high probability behaviors is made contingent upon clinically targeted activities. For instance, children may be allowed to eat candy after (contingent upon) doing their homework.

Modeling

This consists of the presentation of a live demonstration, a filmed or pictorial presentation, or an imagined one. Based on the presentation, the viewer learns a method of responding or the contingencies under which the target response will fall.

Observational learning, imitation, and modeling are terms frequently used with respect to this phenomenon. These techniques are useful in the treatment of phobias, and in forming social skills and many other desired behaviors.

Cognitive Procedures

Cognitive therapy or cognitive ABA refers to a wide range of approaches, based on two explicit hypotheses: (1) cognitive factors such as thoughts, images, and memories, are intimately related to dysfunctional behavior, and (2) modification of such factors is an important mechanism for producing behavior change.

Behavioral Self-Control

This set of techniques involves training patients to control their own behavior through the systematic use of behavior principles. The therapist educates patients regarding the technical aspects of various behavior change procedures. In a self-reinforcement program, patients are taught to administer consequences to themselves that are contingent upon the performance of the target behavior.

In the majority of cases behavioral self-control involves self-monitoring, self-evaluation, and self-delivered rewards. There are many potential benefits to teaching patients self-control rather than treatment through external agents. Self-control therapy allocates a great responsibility to patients. It is clear that teaching individuals to be their own therapists will contribute to generalization of treatment effects because patients will be able to continue to administer the program or develop new treatment following the termination of therapy.

Principal Areas of Application

ABA is a broad term and covers many fields of applied psychology, both traditional and nontraditional. A brief description of some of these areas follows.

Behavior Analysis Applied to Clinical Problems

This field is also referred to as behavior therapy, and has a long history. Behavior therapy has been defined as follows:

> Behavior therapy is the attempt to utilize systematically that body of empirical and theoretical knowledge which has resulted from the application of the experimental method in psychology and its closely related disciplines (physiology and neurophysiology) in order to explain the genesis and maintenance of abnormal patterns of behavior; and to apply that knowledge for the treatment or prevention of those abnormalities by means of controlled experimental studies of the single case, both descriptive and remedial.
> (Yates, 1970: p. 18)

Lovaas (1966, 1977, 1993) has developed behavioral treatment for autistic children. He focused on strategies to teach social behaviors, eliminate self-stimulatory behaviors, and develop language skills. The behavioral treatment has produced long-lasting effects, and no alternative treatment for autistic children has been shown to be as successful as behavior modification (Bourret et al., 2012).

Behavior Analysis in Education

A number of applications have been carried out in several important educational aspects: behavior modifications in special populations (mental retardation, visual and auditory limitations, people with physical limitations, children of superior intelligence, and so forth). Also work has been done in order to maximize the efficiency of teaching techniques, learning in the classroom, and in general academic learning from preschool to university education. The techniques of reinforcement, contingency management, environmental design, token economy, and so forth, previously described in this article, have been applied to the educational area.

Industrial/Organizational Behavioral Management

In the early 1970s ABA entered the industrial arena. At the heart of this approach lies the belief that most complex human behaviors are learned, and by altering either the conditions preceding the behavior or especially the consequences following it, that behavior can be altered and new behavior patterns can be learned. A wide variety of industrial-relevant behavior has been successfully altered: teaching the skills required for the job, motivating the performer, reducing absenteeism, reducing turnover, learning to work in groups, improving communication skills in the work setting, reinforcing adequate behavior, extinguishing undesired behavior, increasing productivity, improving morale, overcoming resistance to change, and others (Hersey et al., 1996).

Behavioral Community Psychology

As a 'marriage' of the methodology of behaviorism and the strategies and conceptual framework of community psychology, the area of behavioral community psychology (Nietzel et al., 1996) uses classical conditioning, operant conditioning, social learning, and cognitive/self-control procedures. Examples of operant behavior analysis are prompts, feedback, reinforcement, and response cost. Examples of social learning and cognitive/self-control procedures are modeling, behavioral rehearsal and role-playing, self-instruction, and self-regulation, especially with respect to the development of skills and self-help competences in target populations.

Forensic Applications

Most applications of behavior analysis to psychology and the law have been in connection with criminal justice. Work has been carried out concerning the behavior of the delinquent individual, including juvenile delinquents and predelinquents; in the prevention of criminal behavior; and in the areas of legal

processing, the courts, the jury, witnesses, and so forth. Special importance has been given to the behavior modification of delinquents using behavior analysis procedures.

Sport Psychology

In this area the contribution of ABA has been centered on motivation of athletes, athletic performance, leadership, optimal coaching strategies, audience effects, sport violence, and other similar issues. The main behavioral technology procedures used in sport psychology are positive reinforcement, shaping, chaining, extinction, time out, response cost, fading, prompting, performance feedback, self-monitoring, self-instructions, deep muscle relaxation, modeling, imaginal rehearsal, behavioral rehearsal, thought stopping, and response-induced aids (Luiselli and Reed, 2011).

Larger Sociocultural Applications

A number of 'big issues' have been addressed with the procedures of behavior analysis: youth violence, racism, sexism, poverty, health problems, destruction of the environment, the breakdown of schools, lack of productivity in the workplace, child maltreatment. According to Mattaini and Thyer (1996) these behavioral problems "could ultimately threaten the survival of the human race and other species" (p. 1).

Bibliography

Arantes, J., Berg, M.E., Dien Le, D., Grace, R.C., 2012. Resistance to change and preference for variable versus fixed response sequences. Journal of the Experimental Analysis of Behavior 97, 1–21.

Ardila, R., 1990. Walden Three. Carlton Press, New York.

Ardila, R., 1998. Manual de análisis experimental del comportamiento (Handbook for the Experimental Analysis of Behavior). Editorial Biblioteca Nueva, Madrid, Spain.

Ardila, R., 2010. Psychology and behavior analysis: the nature of the controversy. In: Schwarzer, R., Frensch, P.A. (Eds.), Personality, Human Development and Culture. International Perspectives in Psychological Science, vol. 2. Psychology Press, Hove, UK, pp. 263–274.

Baer, D.M., Wolf, M.M., Risley, T.R., 1987. Some still-current dimensions of applied behavior analysis. Journal of Applied Behavior Analysis 20, 313–327.

Bourret, J.C., Iwata, B.A., Harper, J.M., North, S.T., 2012. Elimination of position-biased responding in individuals with autism and intellectual disabilities. Journal of Applied Behavior Analysis 45, 241–250.

Caballo, V.E., 1998. International Handbook of Cognitive and Behavioral Treatments for Psychological Disorders. Elsevier Science, Oxford.

Cooper, J.O., Heron, T.E., Heward, W.L., 2007. Applied Behavior Analysis, second ed. Prentice-Hall, Upper Saddle River, NJ.

Hersey, P., Blanchard, K.H., Johnson, D.E., 1996. Management of Organizational Behavior. Prentice Hall, Upper Saddle River, NJ.

Johnson, S.P., Welch, T.M., Miller, L.K., Altus, D.D., 1991. Participatory management: maintaining staff performance in a university housing cooperative. Journal of Applied Behavior Analysis 24, 119–127.

Kazdin, A.E., 2000. Behavior Modification in Applied Settings, sixth ed. Brooks/Cole Pub. Co., Pacific Grove, CA.

Labrador, F.J., Cruzado, J.A., Muñoz, M., 1993. Manual de técnicas de modificación y terapia de conducta. (A Handbook of Behavior Modification and Behavior Therapy Techniques). Ediciones Pirámide, Madrid, Spain.

Lattal, K.A., Perone, M., 1998. Handbook of Research Methods in Human Operant Behavior. Plenum Press, New York.

Luiselli, J.K., Reed, D.K. (Eds.), 2011. Behavioral Sport Psychology. Springer, New York.

Lovaas, O.I., 1966. A program for the establishment of speech in psychotic children. In: Wing, J.K. (Ed.), Early Childhood Autism. Pergamon, Elmsford, NY.

Lovaas, O.I., 1977. The Autistic Child: Language Development through Behavior Modification. Irvington, New York.

Lovaas, O.I., 1993. The development of a treatment-research project for developmentally disabled and autistic children. Journal of Applied Behavior Analysis 26, 617–630.

Madden, G.J. (Ed.), 2012. APA Handbook of Behavior Analysis, 2 vols. American Psychological Association, Washington, DC.

Masters, J.C., Burish, T.G., Hollon, S.D., Rimm, D.C., 1987. Behavior Therapy, third ed. Harcourt Brace Jovanovich, San Diego, CA.

Mattaini, M.A., Thyer, B.A., 1996. Finding Solutions to Social Problems. Behavioral Strategies for Change. American Psychological Association, Washington, DC.

Plaud, J.J., Eifert, G.H., 1998. From Behavior Theory to Behavior Therapy. Allyn and Bacon, Boston, MA.

Skinner, B.F., 1948. Walden Two. Macmillan, New York.

Tolan, P., Keys, C., Chertok, F., Jason, L., 1990. Researching Community Psychology. American Psychological Association, Washington, DC.

Wolpe, J., 1958. Psychotherapy by Reciprocal Inhibition. Stanford University Press, Stanford, CA.

Yates, A.J., 1970. Behavior Therapy. Wiley, New York.

Behavior Therapy: Background, Basic Principles, and Early History

Kyle E Ferguson, St. Peter Family Medicine Residency Program, WA, USA; and University of Washington, Seattle, WA, USA
William O'Donohue, University of Nevada, Reno, NV, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Behavior therapy has a complicated history and ever-changing web of intellectual allegiances and influences, including behaviorism, behavior science, cognitive science, folk psychology, and Eastern philosophical traditions, to name only a few. These schools of thought are extremely difficult if not impossible to encapsulate using a few simple properties that traverse and define every approach. Rather, behavior therapy is much more like a 'family resemblance' in which common essential properties play a role in a complex web of belief. Against this backdrop, therefore, it is the aim of this article to discuss the common essential properties related to the conceptual and theoretical background, basic principles and learning processes, and early history of first-generation or first-wave behavior therapy. The first-wave behavior therapists include such prominent figures as Joseph Wolpe, Ogden Lindsley, Nathan Azrin, and Hans Eysenck, to name only a few. The so-called 'second-wave' and, perhaps, 'third-wave' behavior therapy will be taken up in article Behavior Therapy: The Second and Third Waves.

It is hard to envision the field of behavioral health before behavior therapy, considering its effectiveness and its pervasiveness in so many health care arenas. But before the advent of behavior therapy, a variety of schools of psychoanalysis had the greatest clinical following (Eysenck, 1952). These had the widest application in the treatment of various neuroses or neurotic tendencies like phobias and anxiety-driven compulsions (David-Ménard, 1989). However, despite the widespread adoption, these could offer little by way of treatment for such chronic conditions as schizophrenia and autism. In addition, these offered little data to support their efficacy and were not particularly efficient – often being applied for years or decades if the client was willing and had sufficient resources. Psychoanalysis advanced the, perhaps, iatrogenic notion that pathology resulted from ego impairment due to inadequate child rearing (Kanner, 1943; Mahler, 1968; Tustin, 1981).

Unfortunately, these parent–child relationship theories had a devastating impact on many families – especially with the assertion of the so-called 'refrigerator' mothers who were ultimately held accountable for their children's abnormal development (Willick, 2001). Fortunately, particularly in the United States, psychoanalytic views of mother–child attachment dysfunction have been supplanted by more empirically substantiated behavioral theories (Eldevik et al., 2009; Patterson et al., 2010). Only the most ardent proponents of psychoanalysis would now claim that neurodevelopmental disorders are caused by a patient's relationship with his or her mother – at least in North America and throughout much of Europe (*see* Psychoanalysis: Current Status for a thorough discussion on modern psychoanalysis).

Behavior therapy developed, in part, out of dissatisfaction with psychoanalysis in treating certain types of intractable, chronic psychiatric conditions. Part of the dissatisfaction was simply due to the lack of outcome data meeting minimal scientific standards. Part of the dissatisfaction was based on the notion that what data were available at the time did not point to large treatment effects (Eysenck, 1952). And part of the dissatisfaction was due to what were seen as problems with psychoanalytic theory more fundamentally – vague constructs, unclear measurement operations, unclear predictions, and warring factions within psychoanalysis that could not agree on fundamentals.

Of course, behavior therapy was not the only contender vying for psychoanalysis' seat in mental health practice. Carl Rogers's client-centered therapy and other humanistic theories also emerged out of dissatisfaction with psychoanalytic approaches (Rogers, 1951). Like psychoanalysis, however, client-centered therapy was not without its critics. Among other things, client-centered therapy, too, has been criticized for its vagueness of principles, its disregard for diagnosis and hence any modifications for specific problems, and its emphasis on the client's self-evaluation as the principal way to assess therapeutic outcomes (Carter, 1948; Hathaway, 1948; Louttit, 1948). Regarding the last point, self-evaluation is almost always the principal challenge in persons with personality disorders, e.g., borderline personality disorder, schizotypal personality disorder, histrionic personality disorder, and narcissist personality disorder. Accordingly, if left to their own devices, patients with personality disorders and virtually everyone around them would likely perceive the same clinical outcome quite differently. For example, what might be regarded as a treatment success for someone with narcissism might be deemed a treatment failure for a family member who has routine contact with him or her.

Behavior therapy, in contrast, is based on very different assumptions. According to Goldberg, there are 10 underlying assumptions of behavior therapy (Goldberg, *Family Therapy*, 1996: p. 254).

1. All behavior, normal or abnormal, is acquired and maintained in identical ways (that is, according to the same principles of learning).
2. Behavior disorders represent learned maladaptive patterns that need not presume some inferred underlying cause or unseen motive.
3. Maladaptive behavior, such as symptoms, is itself the disorder, rather than being a manifestation of a more basic underlying disorder or disease process.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.21054-4

4. It is not essential to discover the exact situation or set of circumstances in which the disorder was learned; these circumstances are usually irretrievable anyway. Rather, the focus should be on assessing the current determinants that support and maintain the undesired behavior.
5. Maladaptive behavior, having been learned, can be extinguished (that is, unlearned) and replaced by new learned behavior patterns.
6. Treatment involves the application of the experimental findings of scientific psychology, with an emphasis on developing a methodology that is precisely specified, objectively evaluated, and easily replicated.
7. Assessment is an ongoing part of treatment, as the effectiveness of treatment is continuously evaluated and specific intervention techniques are individually tailored to specific problems.
8. Behavior therapy concentrates on the 'here and now' problems, rather than uncovering or attempting to reconstruct the past. The therapist is interested in helping the client identify and change current environmental stimuli that reinforce the undesired behavior, in order to alter the client's behavior.
9. Treatment outcomes are evaluated in terms of measurable changes. Research on specific therapeutic techniques is continuously carried out by behavior therapists.
10. Research on specific therapeutic techniques is continuously carried out by behavioral therapists.

The first generation or 'first wave' of behavior therapists developed out of experimental work in the laboratory with animals (O'Donohue, 2009). Between the 1940s and the 1950s, behavior therapists presented better, although often still flawed, empirical evidence on a number of clinical problems that eluded traditional therapies of that era, like schizophrenia and autism. For several decades, there was a proliferation of applied behavioral research, which applied these principles of behavior in earnest to countless clinical problems. In fact, Lindsley and Skinner – in what some regard as the first use of the term behavior therapy – successfully treated an electively mute schizophrenic woman (Lindsley and Skinner, 1954). Mowrer and Mowrer (1938) reported similar, dramatic improvements in treating enuresis in residential settings for developmentally delayed persons (*see* Behavior Analysis, Applied for a discussion on applied behavior analysis).

Before proceeding, some clarification about this notion of generations or waves is warranted. These notions are not akin to scientific revolutions as seen in basic science (cf physics, biology, or chemistry; Kuhn, 1962; Lakatos, 1970, 1978; Popper, 1959, 1962, 1972). The older paradigm is not replaced. Rather, these waves seem to be additive. For example, the first wave continues to this day in the field of applied behavior analysis (ABA), especially in the treatment of patient populations with highly limited behavioral repertoires such as those suffering from autism (e.g., ABA is still considered the gold standard for autism therapy). The second wave will be taken up in Behavior Therapy: The Second and Third Waves. Likewise, few would argue that cognitive behavior therapy has appreciably declined in popularity among therapists despite the emergence of a large interest in what some authors suggest as third-wave therapies (Hayes, 2004; Hayes et al., 2006). So-called third-wave therapies will be discussed in Behavior Therapy: The Second and Third Waves as well.

The purpose of this article is to provide an overview of the conceptual and theoretical background, basic principles, and early history of first-generation or first-wave behavior therapy. Readers should note that the authors define behavior therapy broadly to include behavior therapy, cognitive therapy, cognitive behavior therapy, and behavior analytic or contextual behavioral science as some would have it. The authors acknowledge that the heterogeneous nature of behavior therapy may make this broad definition too generic. Nonetheless, the present state of the literature does not allow for a more specific definition or paradigm, as the field has yet to come to a consensus about what behavior therapy actually is.

Brief History and the Principles of Behavior Therapy

The first wave extended the core principles of behavior discovered in the learning laboratory to many socially relevant problems in the decades that followed (O'Donohue, 2009). Two of the most influential figures during the first wave were Ivan P. Pavlov and B.F. Skinner.

The Russian physiologist, Pavlov, identified several basic learning mechanisms in the early twentieth century: habituation/sensitization and classical conditioning processes (Pavlov, 1927, 1928). Habituation and sensitization are the most fundamental types of behavioral processes (nonassociative or perceptual learning). They involve a two-term relation between stimulus and response (S-R). In the case of habituation, repeated presentations of an eliciting stimulus (i.e., S) result in a decrease in response (R) magnitude. Pavlov, for example, noted that the orienting reflex in subjects diminished over time in the experimental apparatus. Clinically, in humans, the 'white coat effect,' or increased blood pressure during office visits in patients whose daily ambulatory blood pressure is otherwise normal (i.e., $<120/<80$ mm Hg), illustrates this behavioral principle (Chrysant, 2000). Such patients are often asked to sit in the examination room for longer durations until their physiological distress has diminished (conditioned response (CR)) with increased exposure to medical stimuli (conditioned stimuli (CSs)).

Sensitization, in contrast, results in an increase in response (R) magnitude with repeated presentations of an eliciting stimulus. Generally speaking, sensitization results from presenting stimuli like loud noises, painful shock, or bright lights. A learning history involving intermittent shock, for example, might increase the orienting response to novel environments, as the organism might be more likely to engage in defensive responses. Likewise, humans with a history of trauma (i.e., high-intensity aversive stimuli) may display hypervigilance or an exaggerated startle response when introduced to novel environments (ApA, 2013).

Pavlov also demonstrated that neutral stimuli that produce no change in responding can be potentiated when paired with or associated with a biological reflex (i.e., unconditioned stimulus (US)). In the four-term relation

in classical conditioning (CS–US–UR–CR), the neutral stimulus, CS (e.g., an auditory or visual stimulus), is paired with a US (e.g., food or shock). Through repeated presentations, the CS will come to elicit responses *similar* to the UR (unconditioned response). Although topographically similar (e.g., salivation or pupillary reflex), the CR may differ from the UR with respect to magnitude and latency (i.e., the time it takes to elicit the response following the presentation of the stimulus). The CS can also serve an *excitatory* or *inhibitory* function, signaling the presence or absence of the US, respectively. This process is called stimulus discrimination (Pierce and Epling, 1999). Again, persons with a history of trauma may show an intense or prolonged physiological distress when exposed to internal or external cues (i.e., the CS serves an excitatory function) that were paired with certain aspects of the traumatic event (ApA, 2013). Some Vietnam veterans, for example, display intense physiological distress when smelling gasoline (which, reportedly, smells similar to napalm), and/or upon hearing fireworks on the 4th of July or helicopters hovering overhead (*see* Classical Conditioning Methods in Psychotherapy for a further discussion on classical conditioning).

John B. Watson and Rosalie Rayner were the first researchers to explore Pavlovian or classical conditioning processes in human subjects. Extending Pavlovian principles to human subjects, Watson and Rayner examined classical conditioning processes in a landmark study involving an 11-month-old baby (Watson and Rayner, 1920). During baseline, the infant freely played with a white rat, demonstrating no significant fear behavior. By repeatedly pairing the rat (CS) with a loud noise (US), however, the infant started showing fear and would escape the stimulus whenever it was present.

Expanding on Watson and Rayner's work, Mary Cover Jones (1924a,b), showed the reverse through the process of counterconditioning: by repeatedly pairing the feared object (CS) with a different stimulus associated with a positive stimulus or having the child who feared the stimulus associate with children who did not show the characteristic fear response while in the presence of the feared stimulus. The process of counterconditioning (or respondent extinction), of course, is the fundamental principle behind the treatment now known as exposure and response prevention (Foa and Kozak, 1996) (interested readers should consult Exposure Therapies and Stress Inoculation: A Brief Overview for a discussion on exposure therapies). Of course, while necessary, the process of counterconditioning is not sufficient for conceptualizing most anxiety conditions that present clinically. Anxiety conditions are often more complex than can be addressed by basic respondent processes.

Mowrer's (1947; 1951; 1956) two-factor theory of emotion has been the most influential theory of fear and avoidance even to this day. The two-factor theory of emotion is based on the work of Hull (1943), though unlike Hull, Mowrer explained conditioned or learned fear by way of classical conditioning processes (recall Hull's drive reduction theory). Namely, CSs are first paired with aversive unconditioned stimuli (USs) until there has been adequate transfer of function(s) (i.e., the CS elicits the characteristic response associated with the US). Based on this model, individuals with this conditioning history will emit overt behavior in the service of terminating the CS rather than avoid the US.

Individuals with posttraumatic stress disorder (PTSD) illustrate this point very well. Internal or external reminders of a traumatic event (CSs) might be avoided for years – even decades after the threat has subsided (e.g., it is very unlikely that debris on the side of the road in Salt Lake City is an improvised explosive device) (*see* Post-Traumatic Stress Disorder for an in-depth discussion on PTSD). Sadly, many veterans of war, who have seen combat, seldom leave their homes (avoidance behavior), as stimuli in the community (e.g., crowds, work crews, loud music, and planes overhead) often elicit intense fear.

Despite interest in these early experiments, clinicians were slow to adopt these principles of behavior in routine practice, still opting for psychoanalytic paradigms. By the 1940s and the 1950s the tide was starting to turn, however; behavior therapy was on the rise in South Africa (Joseph Wolpe), England (Hans Eysenck), and the United States (Ogden Lindsley). Interestingly, as pointed out by Öst (2008), these researchers developed their ideas almost independently of one another.

Wolpe based his treatment, systematic desensitization, and theory of reciprocal inhibition on the work of Pavlov (Wolpe, 1958). Systematic desensitization is a counterconditioning strategy designed to treat "neurotic-anxiety response habits" (Wolpe, 1990: p. 150). Through systematic desensitization, the therapist induces a deep state of relaxation throughout graduated exposure to a feared stimulus as it is believed to reciprocally inhibit and thus weaken learned anxiety habits (Wolpe, 1976). In treating a snake phobia, for example, one would have the client continue to relax at every successive approximation which, ultimately, eventuates in actually handling the snake.

Wolpe marshaled many studies demonstrating the efficacy of behavior therapy in the treatment of phobias and other anxieties (Poppen, 1995). Wolpe, arguably, spearheaded the modern behavior therapy movement, as his work established a sustainable program of research that had a notable impact on how clinicians practiced (O'Donohue, 2009). In fact, to this day, therapists continue to use Wolpe's therapeutic approach. Stress inoculation therapy is one such example, which has helped countless patients undergo highly distressing medical procedures (Meichenbaum, 1985; Meichenbaum and Deffenbacher, 1988; Meichenbaum and Jaremko, 1993).

Since Wolpe's time, classical conditioning mechanisms have played a central role in understanding the development of fear and other anxiety disorders (e.g., PTSD, panic disorder), in the two-factor theory of emotions (Mowrer, 1951, 1956) and more contemporary theories such as Barlow's alarm theory (Barlow, 1988). Respondent extinction (i.e., the eliciting of the CS without the UCS) is the hallmark intervention of Wolpe's systematic desensitization and in more contemporary approaches.

At around the same time in England, Eysenck's work on hysteria and anxiety was largely influenced by the drive reduction theory of Hull and Spence (Hull, 1952; Spence,

1956). Eysenck saw personality traits as being inherited and due to learned behavior. He saw the personality trait of neuroticism – a susceptibility to the effects of stress and anxiety – principally due to an individual's learning history (1967, 1994). Accordingly, treatment would therefore entail providing new learning opportunities to counter an individual's learning history.

In the United States, Lindsely applied operant techniques, developed in the laboratory of B.F. Skinner and colleagues (Ferster and Skinner, 1957; Skinner, 1938, 1953), to patients with extremely limited behavioral repertoires (e.g., patients with schizophrenia and severe mental retardation (Barrett and Lindsley, 1962; Lindsley, 1956, 1959, 1960; Mednick and Lindsley, 1958; Lindsley et al., 1953)). Among his many accomplishments, Lindsley is best known for applying operant methods to treating chronic schizophrenia (Lindsley, 1956).

Skinner coined the term 'operant' because responding is said to operate on the environment – roughly, change it – to produce certain consequences (Skinner, 1953). The consequence in turn exerts influence on the behavior that produced it (Skinner, 1957). Behavior thus increases, decreases, or maintains at its present level of intensity and frequency following the associated consequence(s). The behavior is believed to be caused by the consequence and this causal relationship between the environment and behavior is called a contingency (Sulzer-Azaroff and Mayer, 1991: p. 251). Skinner thought this capability had origins in the evolutionary history of the species and was akin to selection at the biological level (Skinner, 1953).

The three-term contingency is the simplest conceptual model of operant conditioning processes (Holland and Skinner, 1961). The three-term contingency consists of a discriminative stimulus (antecedent stimulus, S^D), behavior, and the consequences of behavior (Catania, 1992). According to this model, discriminative stimuli in the environment occasion or evoke a given class of behavior (a class is determined by the function of behavior). Certain behaviors are therefore more or less likely to be emitted in the presence of various discriminative stimuli. Discriminative stimuli are potentiated by being frequently paired with reinforcers. Responding comes under stimulus control when discriminative stimuli differentially increase the probability of a class of behaviors under characteristic environmental conditions, while decreasing the probability of other responses that are unrelated to the prevailing contingencies of reinforcement (Michael, 1993). For example, when selections are illuminated on a soda machine, the illuminated choices serve as discriminative stimuli, as placing the required money in the vending machine and making the selection will likely result in the selected item. The illuminated selections are potentiated by being paired with the consequences of (1) delivering the soda (conditioned reinforcer) and, ultimately, (2) consuming the soda (primary reinforcer).

Insofar as an S^D increases the probability that a given operant will be emitted by an organism, an S-delta (S^Δ) decreases the likelihood of responding (Pierce and Epling, 1999). An S-delta signals that a reinforcer will not follow behavior in its presence (signaling extinction). Using the example above, an out-of-order sign signals that placing money in the vending machine will not eventuate in the desired item. The behavior of placing coins in the vending machine in the presence of the out-of-order sign becomes extinguished in short order.

Aversive stimuli can also enter into a contingency. An operant that removes an aversive stimulus is called escape. A person with a needle phobia, for example, might escape the CSs of the doctor's office that are associated with having one's skin poked (unconditioned stimuli, UC; Mowrer, 1947, 1951, 1956). An operant that prevents contact with the aversive stimulus is called avoidance. The same individual might avoid going to the doctor for well visits entirely. In either case, the most critical feature is that the stimulus is added to the present environment and removed or prevented contingently after a certain behavior is emitted. Over time, if the behavior maintains at its current level or increases in regard to intensity and/or frequency, the escape or avoidance behavior is said to have been negatively reinforced. The escape or avoidance behavior of persons with agoraphobia, for example, has been negatively reinforced. The behavior of a person with a specific phobia that either escapes or avoids the aversive stimulus is similarly negatively reinforced.

Behavior may decrease in intensity and/or frequency by adding an aversive stimulus or stimuli to the current environment or taking something away. When contingent behavior decreases in frequency in response to something aversive being introduced into the environment, the behavior is said to have been positively punished. This principle is also referred to as Type I Punishment in the applied literature (Foxx, 1982). The reduction of positive punishment (e.g., criticisms, nagging, yelling, etc. – if these consequents eventuate in a reduction in behavioral intensity or frequency) plays a central role in integrative behavioral couple therapy (Jacobson and Follette, 1986). Problematic relationships, for example, invariably entail skewed ratios where negative comments from partners greatly outweigh positive ones.

When contingent behavior decreases in frequency in response to taking something away, negative punishment is said to be operating on behavior. Negative punishment is also called Type II Punishment, or response cost in applied circles (Sulzer-Azaroff and Mayer, 1991). Returning to the couples' example, the 'silent treatment' is one such example, as it invariably follows behavior that one partner finds distasteful or unpleasant (readers should consult Principle of Reinforcement for a further discussion on the principle of reinforcement).

It is important to note that other learning processes were also studied and applied. Conditioned inhibition, latent inhibition, operant–respondent interactions, choice, behavioral economics, optimization, adjunctive behavior, learned helplessness, rule-governed behavior, stimulus equivalence, observational learning, among other phenomena were studied in the laboratory and applied to clinical phenomena (Catania, 1992; Pierce and Epling, 1999). However, despite advancements, it seems fair to say that as the technical nature of learning phenomena increased, the interest or ability to apply it to clinical

phenomena decreased. This phenomenon could also partly be responsible for the rise of second-wave behavior therapy, which will be taken up in article Behavior Therapy: The Second and Third Waves.

Conclusions

Cognitive behavior therapy has a complex history and an ever-changing web of intellectual commitments and interests. It is difficult or impossible to characterize by a simple definition that attempts to enumerate a few essential properties. Rather, it is much more like a Wittgenstein (1952) 'family resemblance' in which common essential properties play a central role in a complex web of belief (Quine and Ullian, 1970). This article discussed the essential properties related to the conceptual and theoretical background, basic principles and learning processes, and early history of first-generation or first-wave behavior therapy.

Behavior therapy developed out of dissatisfaction with psychoanalysis in treating certain types of intractable, chronic psychiatric conditions. Even though it proved highly effective in treating a number of clinical conditions that had eluded clinicians for decades, practitioners eventually perceived limits with the first wave of behavior therapy. Perceived shortcomings in treating certain patient populations – especially higher functioning patients – gave rise to the second wave of behavior therapy. The second wave will be taken up in article Behavior Therapy: The Second and Third Waves.

See also: Behavior Analysis, Applied; Classical Conditioning Methods in Psychotherapy; Exposure Therapies and Stress Inoculation: A Brief Overview; Post-Traumatic Stress Disorder; Psychoanalysis: Current Status.

Bibliography

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Association, Washington, DC.

Barrett, B.H., Lindsley, O.R., 1962. Deficits in acquisition of operant discrimination and differentiation shown by institutionalized retarded children. American Journal of Mental Deficiency 67, 424–436.

Barlow, D., 1988. Anxiety and Its Disorders: The Nature and Treatment of Anxiety and Panic. Guilford Press, New York.

Barlow, D.H., 1996. The effectiveness of psychotherapy: science and policy. Clinical Psychology: Science and Practice 1, 109–122.

Carter Jr., J.W., 1948. Critical evaluation of nondirective therapy. Journal of Clinical Psychology 4, 225–226.

Catania, A.C., 1992. Learning, third ed. Prentice Hall, Englewood Cliffs, NJ.

Chrysant, S.G., 2000. Treatment of white coat hypertension. Current Hypertension Reports 2 (4), 412–417.

David-Ménard, M., 1989. Hysteria from Freud to Lacan: Body and Language in Psychoanalysis (C. Porter, Trans.). Cornell University Press, Ithaca.

Eldevik, S., Hastings, R.P., Hughes, J.C., Jahr, E., Eikeseth, S., Cross, S., 2009. Meta-analysis of early intensive behavioral intervention for children with autism. Journal of Clinical Child and Adolescent Psychology 38 (3), 439–450.

Eysenck, H.J., 1952. The effects of psychotherapy: an evaluation. Journal of Consulting Psychology 16, 319–324.

Eysenck, H.J., 1967. The Biological Basis of Personality. Thomas, Springfield.

Eysenck, H.J., 1994. Personality: biological foundations. In: Vernon, P.A. (Ed.), The Neuropsychology of Individual Differences. Academic Press, London, pp. 151–207.

Ferster, C.B., Skinner, B.F., 1957. Schedules of Reinforcement. Appleton-Century-Crofts, New York.

Foa, E.B., Kozak, M.J., 1996. Psychological treatments for obsessive-compulsive disorder. In: Mavissakalian, M.R., Prien, R.F. (Eds.), Long-term Treatments of Anxiety Disorders. American Psychiatric Press, Inc., Washington, DC, pp. 285–309.

Foxx, R.M., 1982. Increasing Behaviors of Persons with Severe Retardation and Autism. Research Press, Champaign, IL.

Goldberg, H., Goldberg, I., 1996. Family Therapy, an Overview, fourth ed. Brooks/Cole Publishing Company, Pacific Grove.

Hathaway, S.R., 1948. Some considerations relative to nondirective counseling as therapy. Journal of Clinical Psychology 4, 226–231.

Hayes, S.C., 2004. Acceptance and commitment therapy, relational frame theory, and the third wave of behavioral and cognitive therapies. Behavior Therapy 35, 639–666.

Hayes, S.C., Luoma, J., Bond, F., Masuda, A., Lillis, J., 2006. Acceptance and commitment therapy: model, processes and outcomes. Behaviour Research and Therapy 44, 1–25.

Holland, J.G., Skinner, B.F., 1961. The analysis of behavior: a program for self-instruction. McGraw-Hill, New York.

Hull, C.L., 1943. Principles of Behavior: An Introduction to Behavior Theory. Appleton-Century, New York.

Hull, C.L., 1952. A Behavior System: An Introduction to Behavior Theory Concerning the Individual Organism. Yale University Press, New Haven, CT.

Jacobson, N.S., Follette, W.C., 1986. Predicting who will benefit from behavioral marital therapy. Journal of Consulting and Clinical Psychology 54 (4), 518–522.

Jones, M.C., 1924a. A laboratory study of fear: the case of Peter. Pedagogical Seminary 31, 308–315.

Jones, M.C., 1924b. Elimination of children's fears. Journal of Experimental Psychology 7, 325–341.

Kanner, L., 1943. Autistic disturbances of affective contact. American Journal of Psychiatry 103, 242–246.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.

Lakatos, I., 1970. Falsification and the methodology of scientific research programs. In: Lakatos, I., Musgrave, A. (Eds.), Criticism and the Growth of Knowledge. Cambridge University Press, New York, pp. 91–195.

Lakatos, I., 1978. Falsification and the methodology of scientific research programs. In: Worrall, J., Currie, G. (Eds.), The Methodology of Scientific Research Programmes. Cambridge University Press, Cambridge, UK, pp. 8–101.

Lindsley, O.R., 1956. Operant conditioning methods applied to research in chronic schizophrenia. Psychiatric Research Reports 5, 118–139.

Lindsley, O.R., 1959. Reduction in rate of vocal psychotic symptoms by differential positive reinforcement. Journal of the Experimental Analysis of Behavior 2, 269.

Lindsley, O.R., 1960. Characteristics of the behavior of chronic psychotics as revealed by free-operant conditioning methods. Diseases of the Nervous System, (Monograph Supplement) 21, 66–78.

Lindsley, O.R., Skinner, B.F., 1954. A method for the experimental analysis of the behavior of psychotic patients [Abstract] American Psychologist 9, 419–420.

Lindsley, O., Skinner, B.F., Solomon, H.C., 1953. Studies in Behavior Therapy (Status Report I). Metropolitan State Hospital, Walthama, MA.

Louttit, C.M., 1948. Training for non-directive counseling: a critique. Journal of Clinical Psychology 4, 236–240.

Mahler, M., 1968. On Human Symbiosis and the Vicissitudes of Individuation. International Universities Press, New York.

Mednick, M.T., Lindsley, O.R., 1958. Some clinical correlates of operant behavior. Journal of Abnormal and Social Psychology 57 (1), 13–16.

Meichenbaum, D.H., 1977. Cognitive Behavior Modification: An Integrative Approach. Plenum, New York.

Meichenbaum, D., 1985. Stress Inoculation Training. Pergamon Press, Elmsford, NY.

Meichenbaum, D., Deffenbacher, J.L., 1988. Stress inoculation training. Counseling Psychologist 16, 69–90.

Meichenbaum, D., Jaremko, M.E. (Eds.), 1993. Stress Reduction and Prevention. Plenum Press, New York.

Michael, J.L., 1993. Concepts and Principles of Behavior Analysis. Association for Behavior Analysis, Kalamazoo, MI.

Mowrer, O.H., 1947. On the dual nature of learning—a reinterpretation of conditioning and problem solving. Harvard Educational Review 17, 102–148.

Mowrer, O.H., 1951. Two-factor learning theory: summary and comment. Psychological Review 58, 350–354.
Mowrer, O.H., 1956. Two-factor learning theory reconsidered, with special reference to secondary reinforcement and the concept of habit. Psychological Review 63, 114–128.
Mowrer, O.H., Mowrer, W.M., 1938. Enuresis – a method for its study and treatment. American Journal of Orthopsychiatry 8, 436–459.
O'Donohue, W.T., 2009. A brief history of cognitive behavior therapy: are there troubles ahead? In: O'Donohue, W.T., Fisher, J., Hayes, S.C. (Eds.), General Principles and Empirically Supported Techniques of Cognitive Behavior Therapy, second ed. John Wiley & Sons, Inc., New York, pp. 1–14.
Öst, L.G., 2008. Efficacy of the third wave of behavioral therapies: a systematic review and meta-analysis. Behaviour Research and Therapy 46, 296–321.
Patterson, S.Y., Smith, V., Jelen, M., 2010. Behavioural intervention practices for stereotypic and repetitive behavior in individuals with autism spectrum disorder: a systemic review. Developmental Medicine and Child Neurology 52, 318–327.
Pavlov, I.P., 1927. In: Anrep, G.V. (Ed. and Trans.), Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex. Oxford University Press, London.
Pavlov, I.P., 1928. Lectures on Conditioned Reflexes, vol. 1. International, New York.
Pierce, W.D., Epling, W.F., 1999. Behavior Analysis and Learning, second ed. Prentice Hall, Upper Saddle River, NJ.
Poppen, R., 1995. Joseph Wolpe. Sage, London.
Popper, K.R., 1959. The Logic of Scientific Discovery. Basic Books, New York.
Popper, K.R., 1962. Conjectures and Refutations: The Growth of Scientific Knowledge. Basic Books, New York.
Popper, K.R., 1972. Objective Knowledge. Clarendon Press, Oxford.
Quine, W.V., Ullian, J.S., 1970. The Web of Belief. Random House, New York.
Rogers, C., 1951. Client-Centered Therapy: Its Current Practice, Implications and Theory. Constable, London.
Skinner, B.F., 1938. The Behavior of Organisms: An Experimental Analysis. Appleton-Century, New York.
Skinner, B.F., 1953. Science and Human Behavior. Macmillan, New York.
Skinner, B.F., 1957. Verbal Behavior. Appleton-Century-Crofts, New York.
Spence, K.W., 1956. Behavior Theory and Conditioning. Yale University Press, New Haven, CT.
Sulzer-Azaroff, B., Mayer, G.R., 1991. Behavior Analysis for Lasting Change. Harcourt Brace College Publishers, New York.
Tustin, F., 1981. Autistic States in Children. Routledge, Boston.
Watson, J.B., Rayner, R., 1920. Conditioned emotional reactions. Journal of Experimental Child Psychology 3, 1–14.
Willick, M., 2001. Psychoanalysis and schizophrenia: a cautionary tale. Journal of the American Psychoanalytic Association 49, 27–56.
Wittgenstein, L., 1952. Philosophical Investigations. Basil Blackwell, Oxford.
Wolpe, J., 1958. Psychotherapy by Reciprocal Inhibition. Stanford University Press, Stanford, CA.
Wolpe, J., 1976. Theme and Variations: A Behavior Therapy Casebook. Pergamon Press, New York.
Wolpe, J., 1990. The Practice of Behavior Therapy, fourth ed. Pergamon Press, New York.
Yoo, H.J., Ahn, S.H., Kim, S.B., Kim, W.K., Han, O.S., 2005. Efficacy of progressive muscle relaxation training and guided imagery in reducing chemotherapy side effects in patients with breast cancer and in improving their quality of life. Support Care Cancer 13 (10), 826–833.

Behavior Therapy: Evidence-Based Practice, Clinical Applications, Unresolved Issues

Kyle E Ferguson, Providence St. Peter Family Medicine Residency, Renton, WA, USA
William O'Donohue, University of Nevada, Reno, NV, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Behavior therapy is inextricably tied to the health care environment and its practice is subject to the same market forces as other therapeutic orientations like psychoanalysis and psychodynamic therapy, humanistic therapy, and theoretical eclecticism. Of course, as most practitioners who have had to navigate through current procedural terminology (CPT) codes and countless insurance plans would attest, the health care arena is changing at a blistering pace. Given a strong commitment to empiricism (at least ostensibly), behavior therapists are especially well posed to survive, and perhaps, even flourish in a constantly changing health care market, which is only becoming more complex and demanding as health care costs rise while Medicare and Medicaid dollars continue to dwindle. This article provides a general overview of evidence-based practices in psychology, specific examples of behavioral approaches to clinical disorders, and several important unresolved issues, which touches on possible reasons why behavior therapy's growth seems to have slowed considerably since its inception over half a century ago.

Behavior therapy does not exist in a vacuum. Behavior therapy, rather, exists within a health care environment that is constantly evolving. Behavior therapy, unlike other areas in psychotherapy (e.g., psychodynamic approaches) is, arguably, best suited to work within the health care environment of the future, considering its pragmatic aim and commitment to empirical science. The market increasingly wants to buy evidence-based assessments and interventions that have been shown to be effective and safe. Among other things, insurance carriers are most concerned about the 'bottom line' as captured in Gordon Paul's (1967) oft-quoted remark: "What treatment, by whom, is most effective for this individual with that specific problem, under which set of circumstances" (p. 44). Behavior therapy can readily answer this question, unlike other approaches that are more nebulous like insight-oriented approaches. Moreover, behavioral approaches also lend themselves well to continuous quality improvement, which is pervasive in the ever-evolving health care market (Leonard et al., 2004) (*see* Managed Care and Psychotherapy in the United States for a discussion on the impact of managed care on psychotherapy in the United States).

Behavior Therapy and Evidence-Based Practices in Psychology

Clinical practice guidelines (CPGs) were introduced into the medical community several decades ago. These guidelines were designed to assist health care providers in clinical decision-making and attempted to nudge clinicians toward standardization of practice (Shekelle et al., 1999). An emphasis on empirical research and standardization served toward quality improvement of medical surgical services by reducing treatment variability among providers (Carnett, 1999; Handley et al., 1994). Most importantly, through standardization and its minimization of unwanted variance, medical errors could be reduced, thus obviating iatrogenesis, patient morbidity, and mortality (e.g., dispensing the wrong medicine).

The American Psychiatric Association (APA) began drafting their own CPGs in 1991 to assist providers in managing the care of persons with psychiatric illnesses (APA, 2006). Naturally, the emphasis of APA's guidelines were on psychotropics, not nonpharmacological approaches like psychotherapy despite the fact that those approaches were either comparable or superior for certain clinical conditions (Barlow, 1996). Comprised of a panel of about a dozen experts, the American Psychological Association assembled a task force of its own to promote nonpharmacological approaches: Division 12 Task Force on Promotion and Dissemination of Psychological Procedures (Chambless et al., 1998).

Among other things, the Task Force bifurcates efficacious therapies into two groups: 'Well-established treatments' and 'probably efficacious treatments.'

Well-established treatments. Well-established treatments meet the following criteria (I or II, III, IV, and V):

I. At least two good between group design experiments demonstrating efficacy in one of two ways:
   A. The treatment in question is statistically significant to a placebo condition (pill or psychological) or to another treatment.
   B. The treatment in question is equivalent to an already established treatment in experiments with adequate sample sizes.
   or
II. More than a series of nine single-case experiments demonstrating efficacy. These experiments must employ a good experimental design and compare the intervention in question to another placebo or treatment, as is the case in I.A.
III. Experiments must utilize treatment manuals.
IV. Participant characteristics must be explicitly stated.
V. Results must come from at least two different research teams.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.21091-X

Probably efficacious treatments. Probably efficacious treatments share criteria IV and V though differ with respect to criteria I–III.

I. Two experiments that demonstrate that the treatment in question is statistically significant to a waiting list control group.
or
II. One or more experiments meeting all the criteria save criterion V.
III. Three or fewer single-case experiments otherwise meeting all the criteria for well-established treatments.

Behavioral approaches are well represented on the well-established treatments and probably efficacious treatments, as shown in **Table 1** below.

Although the commitment to evidence-based practice (EBP) helped establish empirically based therapies in health care settings, there were inherent shortcomings in how EBPs were selected for inclusion on the list. The four major weaknesses of EBPs with respect to their applicability to clinical settings include the following: (1) Studies on the list were selected because of statistical significance not clinical significance (i.e., treatment outcomes might not be meaningful to patients and key stakeholders); (2) They were selected based on efficacy not effectiveness (a given therapy might be efficacious under ideal randomized control trial (RCT) conditions but not effective under routine care); (3) Selection criteria likely did not reflect more typical heterogeneous patient populations, as many patients with comorbid conditions are excluded from study participation (patients with a high degree of suicidality, e.g., are routinely excluded from depression studies); and there is (4) an obvious bias toward group design methodology and inferential statistics relative to single-case experimental designs and visual inspection methods akin to many behavioral approaches. Accordingly, although the deck has been stacked against behavior therapy, its participation on the EBPs list is quite impressive, nonetheless (*see* Evidence-Based Psychotherapy for an in-depth discussion on evidence-based psychotherapy).

Table 1 Well-established and probably efficacious treatments with a behavioral focus

Well-established treatments
Behavior therapy for depression: Jacobson et al. (1996); McLean and Hakstian (1979)
Behavioral modification for enuresis: Houts et al. (1994)
Behavior therapy for headache: Blanchard et al. (1980); Holroyd and Penzien (1990)
Parent training programs for children with oppositional behavior: Walter and Gilmore (1973); Wells and Egan (1988)
Behavioral marital therapy: Azrin et al. (1980a); Jacobson and Follette (1986)
Probably efficacious treatments
Exposure treatment for social phobia: Feske and Chambless (1995); Foa et al. (1991)
Systematic desensitization for abnormal phobia: Kirsch et al. (1983); Öst (1978)
Systematic desensitization for public speaking anxiety: Paul (1967); Woy and Efran (1972)
Systematic desensitization for social anxiety: Paul and Shannon (1966)
Behavior therapy for cocaine abuse: Higgins et al. (1993)
Community reinforcement approach for alcohol dependence: Azrin (1976); Hunt and Azrin (1973)
Cue exposure and adjunctive to inpatient treatment for alcohol dependence: Drummond and Glautier (1994)
Social skills training adjunctive to inpatient treatment of alcohol dependence: Eriksen et al. (1986)
Multicomponent operant-behavioral therapy for chronic pain: Turner and Clancy (1988); Turner et al. (1990)
Behavior modification of encopresis: O'Brien et al. (1986)
Exposure for simple phobia: Menzies and Clarke (1993)
Behavioral modification for sex offenders: Marshall et al. (1991)
Habit reversal and control techniques: Azrin et al. (1980b, c)
Social skills training for improving social adjustment for schizophrenic patients: Marder et al. (1996)

With that as background, let us turn to a few examples of behavior therapy as used with several clinical populations. Of course, these examples by no means represent all of the clinical applications of behavior therapy, considering its enormous breadth. The authors intentionally selected less familiar clinical contexts to many readers to demonstrate the flexibility of behavioral approaches.

Specific Examples of Behavioral Approaches to Clinical Disorders

Classical Conditioning and Anticipatory Nausea and Vomiting. Anticipatory nausea and vomiting (ANV) occurs by the fourth cycle in 25% of patients undergoing chemotherapy (Morrow and Roscoe, 1997; Morrow et al., 1998). ANV is best understood from a classical conditioning perspective (Bovbjerg, 2006). Neutral stimuli involving the clinic (e.g., sights, smells, sounds, etc.) come to elicit the cognitive restructuring (CR) (i.e., ANV) after a sufficient number of pairings with chemotherapy infusions (unconditioned stimulus (UCS)). As indicated, four pairings are sufficient to elicit ANV in about one-fourth of cancer patients. Interestingly, younger patients who report severe, frequent nausea, and/or vomiting after treatment are at elevated risk for developing ANV related to older individuals (Roscoe et al., 2011). The infusions act as the UCSs because they elicit nausea and vomiting (unconditioned response (UCR)) without prior conditioning.

Systematic desensitization using relaxation is most often used in psychological interventions in cancer care (Fawzy et al., 1995). Behavioral treatment for ANV involves blocking the association between potential CSs (conditioned stimuli) and the UCSs. Counterconditioning using systematic desensitization and relaxation strategies have proven effective in this regard (Morrow and Morrell, 1982; Lyles et al., 1982; Yoo et al., 2005). The relaxation response is believed to reciprocally inhibit and therefore block the transfer of function from UCSs to neutral stimuli (Wolpe, 1976).

Biopsychosocial Approach to Symptom Management in Schizophrenia. Schizophrenia, as is the case with any chronic condition, is best treated from a team-based biopsychosocial perspective (Bartels et al., 1998). The biopsychosocial model includes biological (e.g., genetic and neurological), psychological, and social variables in multimodal assessment and treatment. Antipsychotic medication is routinely the first line of treatment in order to manage the acute psychotic phase of the illness (Bradford et al., 2002; Marshall et al., 2005; Spencer et al.,

2001). Unfortunately, as many as 80% of patients with schizophrenia experience a relapse within 5 years of the initial episode (Álvarez Jiménez et al., 2011). Moreover, it has been estimated that over one-third of patients with schizophrenia experience medication-resistant symptoms (Garety et al., 2000). Accordingly, medication management and nonpharmacological interventions play key role in long-term symptom management. In a recent meta-analytic review of psychosocial programs for schizophrenia in younger adults, supported employment, cognitive therapy, cognitive behavioral therapy (CBT), and social skills training (including social problem-solving) all reported positive findings (Bustillo et al., 2001).

Supported employment entails on-the-job training (e.g., occupational therapy). Cognitive therapy and CBT approaches directly modify specific thoughts or beliefs that supposedly mediate maladaptive behavioral and emotional responding. The goals of CBT treatments are to reduce the extent to which one 'buys' delusional thoughts (Gould et al., 2001). Social skills training involves teaching patients effective interpersonal strategies within targeted social contexts (e.g., assertiveness skills) (Bartels et al., 1998).

Acceptance and commitment therapy (ACT) has also been used in managing psychotic symptoms and the preliminary outcomes have been favorable. Bach and Hayes (2002), e.g., reported lower symptom believability in their participants with psychotic disorders. Moreover, their study found a 50% reduction in the rate of psychiatric rehospitalization compared to treatment-as-usual over a 4-month follow-up period. Interested readers in applying ACT strategies to psychosis should consult the recent edited book titled, *Acceptance and Commitment Therapy and Mindfulness for Psychosis* (Morris et al., 2013; also, *see* Social Intervention in Psychiatric Disorders for a discussion on social interventions in psychiatric disorders).

Sleep Disorders in the Elderly. Sleep disorders are among the most prevalent and poorly addressed problems of aging (Phillips and Ancoli-Israel, 2001). Poor sleep quality increases risk for falls (Beck and Ailshire, 2013) and placement in an assisted or skilled nursing facility (Hartholt et al., 2011). With prevalence rates as high as 44% in older persons (National Sleep Foundation, 2003), it is estimated that fewer than 15% of chronic insomnia sufferers receive treatment (Mellinger and Balter, 1995). Sleep disorders generally fall into one of three categories: (1) insomnia (i.e., complaints of difficulty falling asleep, staying asleep, and nonrestorative sleep), (2) daytime sleepiness (e.g., narcolepsy), and (3) disorders of arousal that result in disruptive behaviors during sleep (e.g., breathing-related problems like sleep apnea) (Pagel and Parnes, 2001). Among these, insomnia, particularly staying asleep or sleep maintenance, is the most common problem in older persons with sleep disturbances (McCall, 2004). Problems with sleep maintenance can result from age-related physiological changes in the amount and pattern of various stages of sleep, psychological stressors, general medical conditions (e.g., congestive heart failure), psychiatric disorders (e.g., generalized anxiety disorder), and substance use (concurrent and recent discontinuation of use) (Prinz, 2004: p. 1739).

Nonpharmacological treatments of sleep disorders include sleep-hygiene, CBT, or a combination of the two (Martin et al., 2000). Meta-analyses reveal that these interventions are efficacious in the general population, with an average of 70–80% of participants demonstrating significant improvements in their sleep problems (e.g., restoring sleep latency close to normative levels) (Morin et al., 1994; Murtagh and Greenwood, 1995). Given the fact that pharmacologic interventions almost always entail significant side effects (particularly, hypnotic medications, such as benzodiazepines), and, thus, increase the risk of daytime impairment, whenever possible nonpharmacological treatments are always indicated first.

Stimulus control and sleep restriction involves reconditioning the bedroom environment to elicit sleep instead of arousal. This may entail directly altering the stimuli associated with sleep or the patient's response to these stimuli. Turning down the temperature, reducing noise (e.g., wearing earplugs), and dimming lights an hour before bed are several ways of directly modifying stimuli that promote sleep (or eliminate or reduce stimuli that disrupt sleep). Sleep restriction is designed to increase sleep efficiency. Bootzin et al. (1991) recommend the following as a means of establishing new routines with respect to sleep-associated stimuli: only go to bed when sleepy; use the bed only for sleep and sex; get up if it takes longer than 10 min to fall asleep and only return to bed when ready to fall asleep (e.g., after reading something nonstimulating, such as a technical manual); get up at the same time everyday, including weekends (this helps maintain circadian rhythms); and no napping during the day (as this disrupts circadian rhythms).

Relaxation is designed to lower arousal at bedtime and during nighttime awakenings. Diaphragmatic breathing (DB), self-hypnosis, and progressive muscle relaxation (PMR) are common examples. As indicated, CR is the core strategy in many CBTs. With respect to cognitive distortions, persons with insomnia tend to engage in all-or-nothing thinking (Lovato et al., 2013). Namely, they either perceive that they had a great night's sleep or a bad night's sleep nothing in between. Such black-or-white thinking makes it difficult appreciating slight improvements in sleep. Regarding misattributions, individuals with insomnia tend to blame all of their problems on their inability to achieve restful sleep. In other words, they catastrophize their sleep disturbances. Catastrophizing sleep problems can interfere with sleep initiation (*see* Sleep Disorders across the Life Span for further discussion on sleep disorders).

Anger and Coronary Heart Disease. Regarding long-term outcomes, in a recent meta-analysis examining 45 studies, Chida and Steptoe (2009) found that anger and hostility were associated with a 20% increased risk of incident coronary heart disease (CHD) in, otherwise, healthy individuals and a similar increased risk of poor prognosis in patients already diagnosed with CHD. Anger and hostility also increase acute risk for vascular events. For example, Mostofsky et al. (2014) reported that within 2 hours following an angry outburst, vulnerable individuals' risk of myocardial infarction (MI) or acute coronary syndrome, the risk of stroke, and risk of ventricular arrhythmia increased considerably relative to other times when they were not angry.

Ellis's (1962, 1994) rational emotive therapy "concentrates on people's current beliefs, attitudes and self-statements as contributing to or 'causing' and maintaining their emotional

and behavioral disturbances" (Ellis et al., 1988: pp. 1–2). What follows, therefore, is the notion that appropriate emotions are preceded by rational beliefs and inappropriate emotions like hostility are preceded by irrational beliefs (Ellis, 1971). Accordingly, irrational core beliefs ought to be disputed rationally. Cognitive disputations are attempts at changing erroneous beliefs through persuasion, didactic presentations, Socratic dialogue, vicarious experiences, among other approaches (Ellis, 1962, 1971, 1994; Walen et al., 1992).

Many CBT anger management programs involve some form of relaxation. The rationale for employing relaxation is largely based on reciprocal inhibition theory (Wolpe, 1958). Namely, arousal is inhibited after pairing relaxation with anger-provoking stimuli. The stress response and relaxation response are mutually exclusive (Benson, 1975). PMR, using Bernstein and Borkovec's (1973) standardized protocol, and DB are often recommended. PMR teaches patients how to discriminate between feelings of relaxation and arousal states, by having them notice physiological changes as they systematically tense and release various muscle groups in their bodies. By improving self-awareness, patients are better able to recognize arousal and, therefore, employ effective arousal-reducing coping strategies.

DB is also recommended, as it is relatively easy to learn and can be used in a variety of settings (see Hazelett-Stevens and Craske, 2003; for a review of breathing techniques). For example, patients can employ DB after changing lines for the second time at the department of motor vehicles (DMV).

A recent RCT of 427 with cardiovascular disease (CVD) indicated that at an average 94-month follow-up, the CBT group had a 41% lower rate of fatal and nonfatal first recurrent CVD events, 45% fewer recurrent acute MIs, and a nonsignificant 28% lower all-cause mortality (0.72 [0.40–1.30]; $P = 0.28$) (Gulliksson et al., 2011). Regarding general outcomes, several meta-analytic reviews of anger management programs (i.e., cognitive therapy, CBT, social skills training, and relaxation therapy – group and individual modes of delivery) have appeared in the literature (Beck and Fernandez, 1998; Edmondson and Conger, 1996; Tafrate, 1995; Vecchio and O'Leary, 2004). All reviews demonstrated medium to large effect sizes for some aspects of anger, while small for other aspects. Regarding explosive anger as often seen in patients with CHD, CBT, coupled with relaxation, appears to be the treatment of choice for anger expression problems (Edmondson and Conger, 1996; Vecchio and O'Leary, 2004).

Unresolved Issues: Behavior Therapy Has Drifted from Its Roots

Upon reviewing the list of empirically supported psychotherapies, one might be impressed with the significant contributions of behavior therapy, particularly when compared to the numerous alternative schools of therapy (e.g., psychodynamic, Gestalt, feminist, etc.). However, it is also somewhat striking that most of these contributions came from first-generation behavior therapy, with few exceptions (O'Donohue et al., 2001). Moreover, many of these contributions pertain to basic first-generation learning research around contingency management – not from the more recent advances in learning research such as matching, behavioral momentum, and response deprivation, to name only a few (see Pierce and Epling, 1999, for a discussion on these).

It might be useful and instructive to examine what has happened to CBT over the past 50 years; to take stock in what has been gained and, perhaps, even to see what might have been lost. It might be useful to start these with a more fulsome and historically relevant explication of the conceptual basis of early behavior therapy. In one of the first reviews of 'behavior therapy' published in the *Annual Review of Psychology*, Leonard Krasner (1971), an important first-generation behavior therapist, argued that behavior therapy had 15 major streams of development:

1. The concept of behaviorism (e.g., Skinner, 1974).
2. The research in instrumental (operant) conditioning of Thorndike (1931) and Skinner (1938).
3. The development of the technique of reciprocal inhibition, based on Holliman learning theory, as a 'treatment' procedure (Wolpe, 1958).
4. The experimental studies based on learning principles of a group of clinical investigators at Causley Hospital in London under the direction of H.J. Eysenck (*Experiments in Behavior Therapy*, 1964).
5. The application of conditioning and learning concepts to human behavior problems in the United States, from the 1920s to the 1950s, by such investigators as John B. Watson, Mary Cover Jones, and E. Lain Phillips.
6. The interpretation of psychoanalysis in terms of learning theory (e.g., Dollard and Miller, 1950), enhancing learning theory as a respectable basis for clinical work.
7. The conditioning paradigm of classical conditioning derived from Ivan Pavlov as the basis for explaining and changing both normal and deviant behavior.
8. Theoretical concepts and research studies of social role learning and interactions in social psychology and sociology (e.g., George Humans, George H. Mead, Alcott Parsons, and Theodore Sabin).
9. Research in developmental and child psychology emphasizing various learning and modeling (e.g., Albert Bandura, Sidney Bijou, and John Dollard).
10. Social influence studies of demand characteristics, experimenter bias, hypnosis, and placebo.
11. An environmentally based social learning model as an alternative to the 'disease' model of human behavior (Bandura, *Principles of Behavior Modification*, 1969, Krasner and Ullmann, 1965).
12. Dissatisfaction with psychotherapy and the psychoanalytic model, as evidenced by several strong critiques (e.g., Eysenck, 1952).
13. The development of the idea of the clinical psychologist within the scientist–practitioner model (also known as, 'clinician–scientist').
14. A movement within psychiatry away from the orthodox focus of internal dynamics and pathology, toward concepts of human interaction and environmental influence (e.g., Adolph Meyer and Harry Stack Sullivan).
15. A utopian emphasis on the planning of social environments to elicit and maintain the best of human behavior (e.g., Skinner's *Walden Two*, 1948).

What seems to have changed over the past 50 years? Here is a partial list:

1. *Less alliance and interest in behaviorism.* It appears now that CBT is less 'deep' or, to put it more positively, simply oriented to a form of science in which randomly controlled trials are used to evaluate the effectiveness of certain techniques or packages of techniques. Skinner (1972), e.g., attempted to develop an integrated scientific and theoretical framework in which radical behaviorism served as a foundation for his experimental analysis of behavior, which in turn served as a foundation for applied behavior analysis. Modern cognitive behaviorists seem less concerned with these kinds of conceptual foundations and perhaps see nothing unusual in finding substitutes for these in such areas as Buddhism, feminism, or postmodern theories. This is somewhat of a change in the sociology and philosophy of the profession overall.
2. *More acceptance of the* 'medical model.' Early behaviorists argued against the reification of disease. They fought against the view that abnormality resulted from different processes than normality. They saw that what separates normal from abnormal in psychology was often a value judgment. They even objected to the word 'patient' and wanted to use terms like 'client' to differentiate what clinical psychologists were doing from what medical professionals such as psychiatrists were doing. However, for a variety of reasons, perhaps partly due to the power of the medical establishment, perhaps partly due to the assumptions of grant funding, contemporary behaviorists seem less concerned with the conceptual implications and problems of the medical model.
3. *Less interest in reciprocal inhibition and systematic desensitization.* Wolpe's work was quite influential in the late 1950s and early 1960s. However, it is less important now. Little research is being done on systematic desensitization or concepts like reciprocal inhibition. It is interesting to note that even in its origins in behavior therapy, it had several different subschools – there were Skinnerians emphasizing operant analysis and Wolpeans interested in Hullian learning constructs and reciprocal inhibition.
4. *Less emphasis on learning principles.* Principles are being derived from an eclectic mix of sources: Buddhism, Hegelian dialectical philosophies, postmodernism, semantics, Greek virtue ethics (positive psychology), etc. While certainly the pragmatic view of use-anything-and-everything-that-might-work has some merit, some intellectual coherence might be lost. Learning principles as powerful principles from the psychological laboratory had been evaluated by scientific means and were partly chosen because of their record of powerful effects – a very good approach to clinical problems. It is not so clear if these more contemporary sources of influence have the same sort of credentials.
5. *More emphasis on package treatments where process variables are less clear.* Early researchers often explicitly investigated the role of an experimental principle for modifying a clinical problem. Titles of journal articles were often along the lines of, "The use of contingency management to reduce tantruming" or "The use of shaping to increase vocalizations in an electively mute woman." Modern titles differ in two ways: (1) These more often do not refer to criterion behavior but rather to some other abstraction (e.g., acceptance or self-efficacy or to a diagnostic and statistical manual of mental disorders (DSM) diagnostic category like major depression); and (2) the intervention is less a clear process variable and much more a package: e.g., dialectical behavior therapy (DBT) or an anger management package consisting of a number of different elements. While certainly some of this may be derived from the hope that these packages will result in increased effectiveness, the process and mechanisms of change are less clear. At the minimum, this necessitates further research using dismantling designs to try to understand the mechanisms of change in these packages (*see* Psychotherapy Process Research for a further discussion on psychotherapy process research.)
6. *Less concern with psychoanalysis as a rival.* Although psychoanalysis has not completely died, it is not longer a strong competitor. The competitors for CBT now are much more likely to be psychotropic interventions, or vaguely defined supportive interventions (*see* Prophylaxis in Psychiatry for a discussion on medications used in psychiatry).
7. *More microcommunities (Kuhn) of behavior therapists.* Behavior therapy was never monolithic but has become increasingly divided into microcommunities. Sometimes these divisions occur along theoretical lines (the Skinnerians vs the cognitivists); sometimes along interest in a particular therapy approach (ACT vs DBT vs traditional CBT), or sometimes along a particular problem interest (those interested in depression vs those interested in Oppositional Defiant Disorder). But, increasingly, two cognitive behavior therapists, if they differ on these dimensions, may have little in common and there is seldom if any cross-pollination of ideas.
8. *Less utopianism.* Early behavior therapists, probably both due to Skinner's utopian interests and influenced by the Zeitgeist of the 1960s, thought 'big' – they often wanted to solve large societal problems such as pollution, the threat of nuclear war, or problems in the educational system. More contemporary behavior therapists think in terms of coming up with more effective treatments for a particular DSM diagnosis.
9. *Less concern with the possible tensions between cognitive and behavior therapy.* This was for many years a relatively 'hot' debate. There seems to have been a lot of quelling of the heat around this issue.
10. *A restricted scholarly scope.* One can see from Krasner's analysis that first-generation behavior therapists were synthesizing a lot of information from a number of scholarly domains: from learning research; from social psychology; from radical psychology and its critiques of the medical model, and the construct of mental illness; from philosophy of mind; and from a number of other disciplines. Contemporary behavior therapy seems less integrative, although certainly there is some interest in certain influences such as Buddhism.
11. *The emergence of third wave behavior therapies, particularly mindfulness.* The contemporary interest in mindfulness would probably come as a surprise to first-generation

behavior therapists. Nonmentalistic constructs were thought of as essential to this group, generally seeing that this was consistent with their philosophy of science – i.e., some form of behaviorism.

12. A loss of some of the uniqueness of behavior therapy methodological preference: e.g., the increased use of questionnaires instead of measuring criterion behaviors. Behavior therapy research in the first generation was somewhat unique – certainly the single-subject designs of proponents of Skinnerian behaviorism were. But their general intellectual commitments also wanted them to deal with criterion behavior directly – e.g., they wanted to count cigarettes and not look at a questionnaire purportedly measuring 'self-efficacy to stop smoking' or 'avoidance.' However, it also might be argued that the research methods used by cognitive behavior therapists have become more 'sophisticated.' There certainly has been increased sophistication in methodologies and data analysis involving rather elaborate statistical applications. Arguably, increased complexity in statistical techniques makes interpretation of outcomes more obscure. Few consumers of research and researchers for that matter fully understand the complexities of the multivariate analytic technique of structural equation modeling. In fact, more often than not, statisticians and quantitative psychologists are recruited on large-scale studies, as principal and coinvestigators lack the necessary skill set to run those analyses appropriately themselves. Accordingly, if the investigators cannot understand the analytic strategies, then how can one assume that these studies are within the grasp of consumers of that research who are likely less familiar with the content area?
13. *The influence of health care insurance companies has increased.* Insofar as the behavior of research participants and patients can be influenced by prevailing contingencies of reinforcement, so can the behavior of clinicians and researchers. Naturally, financial contingencies play a significant role on the behavior of researchers and practitioners. Federal and private grant agencies almost invariably fund projects based on DSM diagnostic categories (e.g., NIH RO1 for depression). Similarly, clinicians are reimbursed by insurance companies using international classification of diseases (ICD) and DSM diagnoses. In fact, there are special diagnostic categories, which garner higher reimbursement rates – so-called risk-adjusted diagnoses (e.g., a diagnosis of major depression versus depression not otherwise specified (NOS) garners an additional $3200 per patient annually), which, ultimately, shape the diagnostic practices of clinicians. Coding specialists and practice-review specialists employed by health care agencies also exert influence on what is actually done in session (*see* Psychiatric Care, Economics of for a discussion on the economics of behavioral health care).

Conclusions

CBT has a complex history with evolving intellectual allegiances and interests. Complicating matters further, being inextricably tied to the health care environment, behavior therapy has evolved in keeping with the same market forces that have greatly influenced how medicine and other therapeutic orientations are practiced. Given pragmatic aims and a strong commitment to empiricism, behavior therapists are especially well posed to meet the complexity and demands of the ever-changing market as health care costs continue to rise exponentially while national dollars allotted to health care dwindle. Upon reviewing the list of empirically supported psychotherapies, it is striking just how many significant contributions of behavior therapy appear on the list. Moreover, it is equally as striking that most of these contributions came from first-generation behavior therapy, with few exceptions (O'Donohue et al., 2001). Behavior therapy is, ultimately, a growing and dynamic enterprise that should understand its roots to better appreciate what is lost and what is gained as it evolves.

See also: Behavior Therapy: Background, Basic Principles, and Early History; Behavior Therapy: The Second and Third Waves; Managed Care and Psychotherapy in the United States; Psychotherapy Process Research; Sleep Disorders across the Life Span; Social Intervention in Psychiatric Disorders.

Bibliography

Álvarez Jiménez, M., Parker, A.G., Hetrick, S.E., McGorry, P.D., Gleeson, J.F., 2011. Preventing the second episode: a systematic review and meta-analysis of psychosocial and pharmacological trials in first-episode psychosis. Schizophrenia Bulletin 37 (3), 619–630.

American Psychiatric Association, 1994. Practice Guidelines. American Psychiatric Association, Washington, DC. From: http://www.psych.org/psych_pract/treatg/pg/prac_guide.cfm (retrieved 03.06.06.).

APA Presidential Task Force on Evidence-Based Practice, 2006. Evidence-based practice in psychology. American Psychologist 61, 271–285.

Azrin, N.H., 1976. Improvements in the CR approach to alcoholism. Behaviour Research and Therapy 14, 339–348.

Azrin, N.H., Bersalel, A., Bechtel, R., Michalicek, A., Mancera, M., Carroll, D., Shuford, D., Cox, J., 1980a. Comparison of reciprocity and discussion-type counseling for marital problems. American Journal of Family Therapy 8, 21–28.

Azrin, N.H., Nunn, R.G., Frantz, S.E., 1980b. Habit reversal vs. negative practice treatment of nailbiting. Behaviour Research and Therapy 18, 281–285.

Azrin, N.H., Nunn, R.G., Frantz-Renshaw, S., 1980c. Habit reversal treatment of thumbsucking. Behaviour Research and Therapy 18, 395–399.

Bach, P., Hayes, S.C., 2002. The use of acceptance and commitment therapy to prevent the rehospitalization of psychotic patients: a randomized controlled trial. Journal of Consulting and Clinical Psychology 70 (5), 1129–1139.

Barlow, D.H., 1996. The effectiveness of psychotherapy: science and policy. Clinical Psychology: Science and Practice 1, 109–122.

Bartels, S.J., Mueser, K.T., Miles, K.M., 1998. Schizophrenia. In: Hersen, M., Van Hasselt, V.B. (Eds.), Handbook of Clinical Geropsychology. Plenum Press, New York, pp. 173–194.

Beck, P., Ailshire, J., 20 November 2013. Sleep well: the effect of sleep quality on falls in older adults. In: Gerontological Society of America 66th Annual Meeting. New Orleans, Louisiana.

Beck, R., Fernandez, E., 1998. Cognitive-behavioral therapy in the treatment of anger: a meta-analysis. Cognitive Therapy and Research 22 (1), 63–74.

Benson, H., 1975. The Relaxation Response. Harper Collins, New York.

Bernstein, D.A., Borkovec, T.D., 1973. Progressive Relaxation Training: A Manual for the Helping Professions. Research Press, Champaign, IL.

Blanchard, E.B., Andrasik, F., Ahles, T.A., Teders, S.J., O'Keefe, D., 1980. Migraine and tension headache: a meta-analytic review. Behavior Therapy 11, 613–631.

Bootzin, R.R., Epstein, D., Wood, J.M., 1991. Stimulus control instructions. In: Hauri, P.J. (Ed.), Case Studies in Insomnia. Plenum Medical, New York, pp. 19–28.

Bovbjerg, D.H., 2006. The continuing problem of post chemotherapy nausea and vomiting: contributions of classical conditioning. Autonomic Neuroscience 129, 92–98.

Bradford, D., Stroup, S., Lieberman, J., 2002. Pharmacological treatments for schizophrenia. In: Nathan, P.E., Gorman, J.M. (Eds.), A Guide to Treatments that Work, second ed. Oxford University Press, Oxford, pp. 169–199.

Bustillo, J.R., Lauriello, J., Horan, W.P., Keith, S.J., 2001. The psychosocial treatment of schizophrenia: an update. American Journal of Psychiatry 158, 163–175.

Carnett, W.G., 1999. Clinical practice guidelines: a tool to improve care. Quality Management in Health Care 8, 13–21.

Chambless, D.L., et al., 1998. Update on empirically validated therapies, II. The Clinical Psychologist 51, 3–16.

Charles, K., 2011. Transforming Healthcare: Virginia Mason Medical Centers Pursuit of the Perfect Patient Experience. Productivity Press, New York.

Chida, Y., Steptoe, A., 2009. The association of anger and hostility with future coronary heart disease: a meta-analytic review of prospective evidence. Journal of the American College of Cardiology 53 (11), 936–946.

Dollard, J., Miller, N.E., 1950. Personality and Psychotherapy. McGraw-Hill, New York.

Drummond, D.C., Glautier, S., 1994. A controlled trial of cue exposure treatment in alcohol dependence. Journal of Consulting and Clinical Psychology 62, 809–817.

Edmondson, C.B., Conger, J.C., 1996. A review of treatment efficacy for individuals with anger problems: conceptual, assessment, and methodological issues. Clinical Psychology Review 16, 251–275.

Ellis, A., 1962. Reason and Emotion in Psychotherapy. Lyle Stuart, New York.

Ellis, A., 1971. Growth through Reason. Wilshire, North Hollywood, CA.

Ellis, A., 1975. A New Guide to Rational Living. Wilshire Book Company.

Ellis, A., 1994. Anger: How to Live with and without It. Carol Publishing Group, New York.

Ellis, A., McInerney, J.F., DiGiuseppe, R., Yeager, R.J., 1988. Rational-Emotive Therapy with Alcoholics and Substance Abusers. Allyn and Bacon, Boston.

Eriksen, L., Bjornstad, S., Gotestam, K.G., 1986. Social skills training in groups for alcoholics: one-year treatment outcome for groups and individuals. Addictive Behaviors 11, 309–329.

Eysenck, H.J., 1952. The effects of psychotherapy: an evaluation. Journal of Consulting Psychology 16, 319–324.

Eysenck, H.J., 1964. Experiments in Behaviour Therapy. Pergamon Press, London.

Fawzy, I.F., Fawzy, N.W., Arndt, L.A., Pasnau, R.O., 1995. A critical review of psychosocial interventions in cancer care. Archives of General Psychiatry 52, 100–113.

Feske, U., Chambless, D.L., 1995. Cognitive behavioral versus exposure only treatment for social phobia: a meta-analysis. Behavior Therapy 26, 695–720.

Foa, E.B., Rothbaum, B.O., Riggs, D.S., Murdock, T.B., 1991. Treatment of posttraumatic stress disorder in rape victims: a comparison between cognitive-behavioral procedures and counseling. Journal of Consulting and Clinical Psychology 59, 715–723.

Garety, P.A., Fowler, D., Kuipers, E., 2000. Cognitive-behavioral therapy for medication resistant symptoms. Schizophrenia Bulletin 26, 73–86.

Gould, R.A., Mueser, K.T., Bolton, E., Mays, V., Goff, D., 2001. Cognitive therapy for psychosis in schizophrenia: an effect size analysis. Schizophrenia Research 48, 335–342.

Gulliksson, M., Burell, G., Vessby, B., et al., 2011. Randomized controlled trial of cognitive behavioral therapy v. standard treatment to prevent recurrent cardiovascular events in patients with coronary heart disease: Secondary Prevention in Uppsala Primary Health Care project (SUPRIM). Archives of Internal Medicine 171, 134–140.

Handley, M.R., Stuart, M.E., Kirz, H.L., 1994. An evidence-based approach to evaluating and improving clinical practice: implementing practice guidelines. HMO Practice 8, 75–83.

Hartholt, K.A., van Beeck, E.F., Polinder, S., van der Velde, N., van Lieshout, E.M.M., Panneman, M.J.M., et al., 2011. Societal consequences of falls in the older population: injuries, healthcare costs, and long-term reduced quality of life. The Journal of Trauma 71, 748–753.

Hazelett-Stevens, H., Craske, M.G., 2003. Breathing retraining and diaphragmatic breathing techniques. In: O'Donohue, W., Fisher, J.E., Hayes, S.C. (Eds.), Cognitive Behavior Therapy: Applying in Empirically Supported Techniques in Your Practice. John Wiley & Sons, New Jersey, pp. 59–64.

Higgins, S.T., Budney, A.J., Bickel, W.K., Hughes, J.R., Foeg, F., Badger, G., 1993. Achieving cocaine abstinence with a behavioral approach. American Journal of Psychiatry 150, 763–769.

Holroyd, K.A., Penzien, D.B., 1990. Pharmacological versus nonpharmacological prophylaxis of recurrent migraine headache: a meta-analytic review of clinical trials. Pain 42, 1–13.

Houts, A.C., Berman, J.S., Abramson, H., 1994. Effectiveness of psychological and pharmacological treatments for nocturnal enuresis. Journal of Consulting and Clinical Psychology 62, 737–745.

Hunt, G.M., Azrin, N.J., 1973. A community reinforcement approach to alcoholism. Behaviour Research and Therapy 11, 91–104.

Jacobson, N.S., Dobson, K.S., Truax, P.A., Addis, M.E., Koerner, K., Gollan, J.K., Gortner, E., Prince, S.E., 1996. A component analysis of cognitive-behavioral treatment for depression. Journal of Consulting and Clinical Psychology 64, 295–304.

Jacobson, N.S., Follette, W.C., 1986. Predicting who will benefit from behavioral marital therapy. Journal of Consulting and Clinical Psychology 54 (4), 518–522.

Kirsch, I., Tennen, H., Wickless, C., Saccone, A.J., Cody, S., 1983. The role of expectancy in fear reduction. Behavior Therapy 14, 520–533.

Krasner, L., 1971. Behavior therapy. Annual Review of Psychology 22, 438–532.

Krasner, L., Ullmann, L.P. (Eds.), 1965. Research in behavior modification: new developments and implications. Holt, Rinehart, & Winston, New York.

Leonard, M., Frankel, A., Simmonds, T., Vega, K., 2004. Achieving Safe and Reliable Healthcare Strategies and Solutions. AONE, USA, IHI.

Lovato, N., Lack, L., Wright, H., Kennaway, D.J., 2013. Predictors of improvement in subjective sleep quality reported by older adults following group-based cognitive behavior therapy for sleep maintenance and early morning awakening insomnia. Sleep Medicine 14, 888–893.

Lyles, J.N., Burish, T.G., Krozely, M.G., et al., 1982. Efficacy of relaxation training and guided imagery in reducing the aversiveness of cancer chemotherapy. Journal of Consulting and Clinical Psychology 50 (4), 509–524.

Marder, S.R., Wirshing, W.C., Mintz, J., McKenzie, J., Johnston, K., Eckman, T.A., Lebell, M., Zimmerman, K., Liberman, R.P., 1996. Two-year outcome of social skills training and group psychotherapy for outpatients with schizophrenia. American Journal of Psychiatry 153, 1585–1592.

Marshall, W.L., Jones, R., Ward, T., Johnston, P., Barbaree, H.E., 1991. Treatment outcome with sex offenders. Clinical Psychology Review 11, 465–485.

Marshall, M., Lewis, S., Lockwood, A., Drake, R., Jones, P., Croudace, T., 2005. Association between duration of untreated psychosis and outcome in cohorts of first episode patients: a systematic review. Archives of General Psychiatry 62, 975–983.

Martin, J., Shochat, T., Ancoli-Israel, S., 2000. Assessment and treatment of sleep disturbances in older adults. Clinical Psychology Review 20, 783–805.

McCall, W.V., 2004. Sleep in the elderly: burden, diagnosis, and treatment. Primary Care Companion Journal of Clinical Psychiatry 6, 9–20.

McLean, P.D., Hakstian, A.R., 1979. Clinical depression: comparative efficacy of outpatient treatments. Journal of Consulting and Clinical Psychology 47, 818–836.

Mellinger, G.D., Balter, M.B., 1995. Insomnia and its treatment. Archives of General Psychiatry 42, 225–232.

Menzies, R.G., Clarke, J.C., 1993. A comparison of in vivo and vicarious exposure in the treatment of childhood water phobia. Behaviour Research and Therapy 31, 9–15.

Morin, C.M., Culbert, J.P., Schwartz, S.W., 1994. Nonpharmacological interventions for insomnia: a meta-analysis of treatment efficacy. American Journal of Psychiatry 151, 1172–1180.

Morris, E.M.J., Johns, L.C., Oliver, J.E., 2013. Acceptance and Commitment Therapy and Mindfulness for Psychosis. John Wiley & Sons, Inc., Malden, MA.

Morrow, G.R., Morrell, C., 1982. Behavioral treatment for the anticipatory nausea and vomiting induced by cancer chemotherapy. New England Journal of Medicine 307 (24), 1476–1480.

Morrow, G.R., Roscoe, J.A., 1997. Anticipatory nausea and vomiting: models, mechanisms and management. In: Dicato, M. (Ed.), Medical Management of Cancer Treatment Induced Emesis. Martin Dunitz, London, pp. 149–166.

Morrow, G.R., Roscoe, J.A., Kirshner, J.J., Hynes, H.E., Rosenbluth, R.J., 1998. Anticipatory nausea and vomiting in the era of 5-HT3 antiemetics. Support Care Cancer 6, 244–247.

Mostofsky, E., Penner, E.A., Mittleman, M.A., 2014. Outbursts of anger as a trigger of acute cardiovascular events: a systematic review and meta-analysis. European Heart Journal. http://eurheartj.oxfordjournals.org/content/early/2014/03/03/eurheartj.ehu033.full.pdf+html (accessed 13.03.14.).

Murtagh, D.R.R., Greenwood, K.M., 1995. Identifying effective psychological treatments for insomnia: a meta-analysis. Journal of Consulting and Clinical Psychology 63, 79–89.

National Sleep Foundation, 2003. 2003 Sleep in America poll. Retrieved from http://sleepfoundation.org/sites/default/files/2003SleepPollExecSumm.pdf(accessed 07.22.14.).

O'Brien, S., Ross, L.V., Christophersen, E.R., 1986. Primary encopresis: evaluation and treatment. Journal of Applied Behavior Analysis 19, 137–145.

O'Donohue, W.T., Henderson, D.A., Hayes, S.C., Fisher, J.E., Hayes, L.J. (Eds.), 2001. A History of the Behavioral Therapies: Founders' Personal Histories. Context Press, Reno, NV.

Öst, L.G., 1978. Fading vs. systematic desensitization in the treatment of snake and spider phobia. Behaviour Research and Therapy 16, 379–389.

Pagel, J.F., Parnes, B.L., 2001. Medications for the treatment of sleep disorders: an overview. Primary Care Companion Journal of Clinical Psychiatry 3, 118–125.

Paul, G.L., 1967. Insight vs. desensitization in psychotherapy two years after termination. Journal of Consulting Psychology 31, 333–348.

Paul, G.L., Shannon, D.T., 1966. Treatment of anxiety through systematic desensitization in therapy groups. Journal of Abnormal Psychology 71, 123–135.

Phillips, B., Ancoli-Israel, S., 2001. Sleep disorders in the elderly. Sleep Medicine 2, 99–114.

Pierce, W.D., Epling, W.F., 1999. Behavior Analysis and Learning, second ed. Prentice Hall, Upper Saddle River, NJ.

Prinz, P.N., 2004. Age impairments in sleep, metabolic and immune functions. Experimental Gerontology 39, 1739–1743.

Quine, W.V., Ullian, J.S., 1970. The Web of Belief. Random House, New York.

Roscoe, J.A., Morrow, G.R., Aapro, M.S., Molassiotis, A., Olver, I., 2011. Anticipatory nausea and vomiting. Supportive Care in Cancer 19 (10), 1533–1538.

Shekelle, P.G., Woolf, S.H., Eccles, M., Grimshaw, J., 1999. Clinical guidelines: developing guidelines. British Medical Journal 318, 593–596.

Skinner, B.F., 1938. The Behavior of Organisms: An Experimental Analysis. Appleton-Century, New York.

Skinner, B.F., 1948. Walden Two. Macmillan Co., New York.

Skinner, B.F., 1972. Cumulative Record, third ed. Appleton-Century-Crofts, New York.

Skinner, B.F., 1974. About Behaviorism. Vintage, New York.

Spencer, E., Birchwood, M., McGovern, D., 2001. Management of first-episode psychosis. Advances in Psychiatric Treatment 7, 133–140.

Tafrate, R.C., 1995. Evaluation of treatment strategies for adult anger disorders. In: Kassinove, H. (Ed.), Anger Disorders: Definition, Diagnosis, and Treatment. Taylor & Francis, Washington, DC, pp. 109–129.

Thorndike, E.L., 1931. Human Learning. Century, New York.

Turner, J.A., Clancy, S., 1988. Comparison of operant behavioral and cognitive-behavioral group treatment for chronic low back pain. Journal of Consulting and Clinical Psychology 56, 261–266.

Turner, J.A., Clancy, S., McQuade, K.J., Cardenas, D.D., 1990. Effectiveness of behavioral therapy for chronic low back pain: a component analysis. Journal of Consulting and Clinical Psychology 58, 573–579.

Vecchio, T.D., O'Leary, K.D., 2004. Effectiveness of anger treatments for specific anger problems: a meta-analytic review. Clinical Psychology Review 24, 15–34.

Walen, S.R., DiGiuseppe, R., Dryden, W., 1992. A Practitioner's Guide to Rational-Emotive Therapy, second ed. Oxford University Press, Oxford.

Walter, H.I., Gilmore, S.K., 1973. Placebo versus social learning effects in parent training procedures designed to alter the behavior of aggressive boys. Behavior Therapy 4, 361–377.

Wells, K.C., Egan, J., 1988. Social learning and systems family therapy for childhood oppositional disorder: comparative treatment outcome. Comprehensive Psychiatry 29, 138–146.

Wittgenstein, L., 1952. Philosophical Investigations. Basil Blackwell, Oxford.

Wolpe, J., 1958. Psychotherapy by Reciprocal Inhibition. Stanford University Press, Stanford, CA.

Wolpe, J., 1976. Theme and Variations: A Behavior Therapy Casebook. Pergamon Press, New York.

Woy, R.J., Efran, J.S., 1972. Systematic desensitization and expectancy in the treatment of speaking anxiety. Behaviour Research and Therapy 10, 43–49.

Yoo, H.J., Ahn, S.H., Kim, S.B., Kim, W.K., Han, O.S., 2005. Efficacy of progressive muscle relaxation training and guided imagery in reducing chemotherapy side effects in patients with breast cancer and in improving their quality of life. Supportive Care in Cancer 13 (10), 826.

Behavior Therapy: Psychiatric Aspects

Lynne M Drummond, South West London and St George's Mental Health NHS Trust, London, UK; and St George's, University of London, London, UK
Isaac Marks, King's College, University of London, London, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D. DiLillo, volume 2, pp. 1086–1090, © 2001, Elsevier Ltd.

Abstract

Behavior therapy and cognitive therapy describe ways to help patients change their responses in order to reduce distress and other problems. The problems include anxiety, social skills and depressive disorders, couple difficulties, learning disabilities, and aspects of psychoses. Guidance may be given face to face, by phone, book, or online. Attaining enough enduring improvement may need days, weeks, or months of help depending on the duration, severity, and complexity of the problem. Added antidepressant medication can be beneficial, if mood is low.

Behavior therapy refers to a range of methods to change a person's responses to specific situations. Altering those responses often alleviates distress and psychiatric problems. An early behavioral treatment of phobic disorders was systematic desensitization, which teaches patients to relax and then imagine their anxiety-evoking scenes repeatedly. Systematic desensitization is a form of graded exposure. Later, patients were asked to gradually face their feared situations until the fear began to subside, gradually increasing the fear intensity of those situations up their hierarchy. This is now the treatment of choice for phobic, obsessive-compulsive, and other anxiety disorders.

Other behavioral methods teach social skills to sufferers with problems such as social skills deficits, marital disharmony, and sexual dysfunction. Further behavioral methods reward desired behaviors and ignore or suppress undesirable ones – such methods are used most often in chronically institutionalized patients.

Behavior therapy is commonly used together with cognitive therapy, which helps patients to examine their thinking patterns and challenge unhelpful ways of thinking. Behavior therapy is also frequently combined with antidepressant drugs if mood is low.

A useful definition of behavior therapy is that of Meyer and Chesser (1970): Behavior therapy aims to modify current symptoms and focuses attention on their behavioral manifestations in terms of observable responses. The techniques used are based on a variety of learning principles. Although behavior therapists adopt a developmental approach to the genesis of symptoms, they do not think it is always necessary to unravel their origin and subsequent development.

In psychiatric practice, behavior therapy has improved anxiety and obsessive-compulsive disorders, habit disorders, and challenging or antisocial behaviors. Cognitive therapy that helps people examine their thinking patterns and alter maladaptive thoughts can often be used effectively together with behavior therapy. Combined cognitive behavior therapy improved depression, panic, generalized anxiety, posttraumatic stress, and other psychiatric conditions.

History of Behavior Therapy

Past historical documents described treatments of children and adults with techniques, which are now called behavior therapy, though anxiety disorders and other neuroses have often been treated by psychoanalytic psychotherapy since the early twentieth century. The latter assumes that psychiatric disorders are due to unconscious conflicts, which can take years of psychoanalysis to resolve. An early challenge to this belief was the work of Watson and Rayner with Little Albert (Watson and Rayner, 1920). Little Albert was a young boy who had shown no fear of animals. Watson and Rayner placed him in a room with a white rat. Whenever Albert approached the rat, the researchers made a loud aversive noise. Through such classical conditioning Albert quickly developed a fear of white rats, which generalized within hours to other furry objects and animals.

A few years later, Jones (1924) tried reward techniques in children with preexisting phobias. She used food with hungry children and brought the feared object toward the child as it ate. Any fear response caused her to move the frightening object further away until the child started eating again. This was repeated until the child was happy even when close to the feared object.

Despite such early examples of the use of behavioral principles, behavior therapy was not defined as a discipline until the 1950s when Joseph Wolpe popularized systematic desensitization for phobic disorder. Systematic desensitization assumes that an individual or group cannot simultaneously feel relaxed and also experience anxiety. Fear of a phobic stimulus can be reduced and eventually extinguished by learning to relax while imagining a series of increasingly frightening scenes. A hierarchy of increasingly fear-evoking stimuli was obtained from the patient and initially, after she/he had induced relaxation, they were asked to imagine a slightly frightening situation. Once this was tolerated, the patient was asked to move a step at a time to imagine increasingly threatening images. If the patient became anxious, the therapist asked the patient to relax again and then imagine

a less-frightening stimulus (Wolpe, 1958). Since that time there have been many studies of systematic desensitization. For example, Gelder et al. (1967) compared systematic desensitization with individual psychotherapy and with group psychotherapy in patients with phobic disorders. Systematic desensitization was consistently better than individual or group psychotherapy.

Although systematic desensitization is effective for specific phobic disorders, it involves a long laborious process of very gradual exposure. Quicker benefit was obtained with real-life exposure methods, one of which was called flooding (Stampfl and Levis, 1967). Real-life exposure involves asking the patient to face the feared stimulus without using relaxation or distraction. Even though this causes anxiety, if the patient remains in contact with the stimulus for sufficiently long (up to 1 h), then the anxiety does gradually reduce. Such exposure too is done up a hierarchy, starting with a situation that causes moderate anxiety before progressing to situations which the patient, unlike most other people, finds very frightening. The hierarchy is less structured than in systematic desensitization and is usually effective after only a total of a few hours in all of exposure.

Real-life exposure methods were pioneered with specific phobias (Watson and Marks, 1971). Specific phobics were treated with rapid exposure in real life with great success. Fears by some therapists that anxious patients would be made worse and other symptoms would appear proved to be unfounded. Exposure treatment was then tried for agoraphobia and shown to be effective (Mathews et al., 1976). From 1970 to 1980, exposure treatments for agoraphobia were refined and crucial treatment factors were isolated. For example, longer duration of exposure was better for improving agoraphobia (Stern and Marks, 1973). Real-life exposure was more effective than even very rapid imaginal exposure (Emmelkamp and Wessels, 1975).

Over the next few years, refinement of psychological treatment for agoraphobia continued. It was thought that if treatment could be done in groups, it would be more cost effective. Hand et al. (1974) compared 'structured' with 'unstructured' groups, and found progress was greater in the structured group. Only the structured group continued to improve after treatment and three month follow-up. Group exposure required less professional time and so was an advance in treatment. However, in practice it is often hard to get a group of similar phobic patients together at the same time, and progress of the group as a whole is linked to the speed of progress of its slowest member.

More recent work in the treatment of anxiety disorders has been in the development of self-exposure where the patient is given detailed instructions about how to perform exposure treatment but the therapist is not present during the exposure therapy. Marks et al. (1983) showed that brief therapist-aided exposure was a useful adjunct to self-exposure homework instructions, and even self-exposure homework instruction alone, unaccompanied by a therapist, is a potent treatment for agoraphobia. Experience to date supports the use of this approach, although involvement of family members as cotherapists can be beneficial.

The treatment of obsessive-compulsive disorder was developed at the same time as that for agoraphobia. Marks et al. (1975) reported the results over 2 years of 20 patients with chronic compulsive rituals treated by real-life exposure. This treatment reduced the rituals as it did agoraphobia. The only difference between the real-life exposure used in obsessive-compulsive disorder and in phobic disorder is that with OCD, the patient is also asked not to perform the anxiety-reducing rituals, which are salient in this disorder. These anxiety-reducing rituals would otherwise interfere with the natural habituation to anxiety, which occurs with prolonged exposure. In patients with rituals, the therapist often initially modeled exposure and ritual prevention.

A similar treatment for obsessive-compulsive rituals had been described in 1903 by the French psychiatrist, Janet. He wrote: "The person who assists in the performance of these actions, has a very complicated part to play. He must aid in the performance of the action without actually doing it himself, although the latter would be very much easier; and he must do his utmost to conceal his own contribution to the action, for it is essential that the patient should feel he does the action himself and does it unaided. The guide has chosen the action, has overcome the patient's hesitations, and has taken the responsibility … by continual repetition to perform the action; by words of encouragement at every sign of success however insignificant, for encouragement will make the patient realize these little successes, and will stimulate him with the hope aroused by glimpses of greater successes in the future."

The principles outlined by Pierre Janet were refined in later studies (Levy and Meyer, 1971; Rachman et al., 1971). Now a truly effective therapy can be offered to most patients with compulsive rituals.

As well as being a treatment for phobic, obsessive-compulsive, and other anxiety disorders, behavior therapy has also been applied to other complaints as varied as schizophrenia and marital/sexual dysfunction. The use of behavior therapy in schizophrenia and learning disability started when Ayllon and Azrin (1968) used operant reinforcement principles to help patients with chronic schizophrenia. Nowadays, behavior therapy is used to counteract secondary handicaps of schizophrenia and to aid patients discharged from hospital to the community. In 1970, the pioneering work of Masters and Johnson (1966, 1970) led to a range of behavioral techniques for sexual dysfunction.

There are many other conditions where behavior therapy has been effective but these cannot be detailed in this short article. There is a trend now for behavior therapy to be used for an increasing range of disorders, and much progress has been made since the first tentative exposure techniques.

Modern Behavior Therapy

This section has three main parts. The first deals with the current use and application of exposure methods of treatment, which are used mostly for anxiety disorders. The second examines methods to teach new skills where these have not been learned, e.g., forms of social skills training for sexual dysfunction, marital therapy, and problem solving. Third, there will be a description of the use of reinforcement techniques for learning disability, chronic patients, and in child psychiatry.

Exposure Treatments

Exposure treatment is effective in two-thirds of agoraphobics (Mathews et al., 1981) and up to 85% of obsessive-compulsives (Foa and Goldstein, 1978; Rachman et al., 1979), and is highly effective (Marks, 1981) in specific and social phobics.

There is often fear of using exposure treatment outside specialist centers. This seems to stem from erroneous views about its applicability, its success rates, the time commitment required by the therapist, and fear of the unknown. In fact, graded exposure is a quick and cost-efficient treatment, which can be easily applied in many general practice and hospital settings (Marks, 1981, 1986; Stern and Drummond, 1991). Basic training can easily be obtained by reading about clinical techniques (Hawton et al., 1989; Stern and Drummond, 1991), reading the self-help Chapter-12 section of Living With Fear (Marks, 2001), and by obtaining supervision from a trained behavioral psychotherapist (Drummond, 2014).

The most effective exposure is:

1. Longer rather than shorter in duration (Stern and Marks, 1973).
2. Done in real life rather than in fantasy (Emmelkamp and Wessels, 1975).
3. Practiced regularly with self-exposure homework tasks (McDonald et al., 1978).

A past concern about exposure treatment was the mistaken idea that it requires considerable professional input to accompany an anxious patient into fear-evoking situations. It has been shown that instruction in self-exposure techniques is all that is needed for many patients with phobic anxiety and obsessive-compulsive disorder (Ghosh et al., 1988; Marks et al., 1988). The efficacy of self-exposure led to the development of self-help manuals. Some patients can complete self-treatment successfully without professional guidance. In fact, many sufferers treat themselves with online guidance from programs such as FearFighter, OCFighter, and Cope. Other people need to be seen initially for education about anxiety and its treatment and for help in devising treatment targets. Subsequent sessions are used to monitor progress, give encouragement, and advise on any difficulties, which may arise (Ghosh et al., 1988).

Exposure is the cornerstone of treatment of obsessive-compulsive disorder when it is combined with ritual prevention – patients are encouraged to refrain from performing rituals and compulsions. Rituals are overt behaviors or internal thought patterns, which are used to counteract obsessive fears. Rituals can usually be prevented or reduced substantially by showing the patient how they interfere with exposure. In exposure therapy, the aim is to produce prolonged periods of contact with the feared situation until the ensuing anxiety reduces (habituation). Compulsions (rituals) immediately reduce the anxiety – this serves to reinforce the ritual. However, the reduction in anxiety produced by a ritual tends to be small and temporary. In effect, rituals prevent therapeutic exposure and instead further increase the tendency to ritualize.

Exposure methods have also been used to treat other anxiety disorders to good effect, including post-traumatic stress disorder (Keane et al., 1989; Foa et al., 1991) and hypochondriasis (Warwick et al., 1996).

Teaching New Skills

Some people lack appropriate skills to meet certain challenges of the environment in which they find themselves. This is, of course, true to some extent for most of us. However, with some people, specific training can resolve a crisis and remove psychological distress.

Skills training begins with the therapist establishing in detail with the patient what skills are at fault and what this leads to. This analysis is followed by sensitive discussion about how the skills could be improved. Frequently the therapist will then model the skill to be achieved. With a complex skill or where the patient has a very low level of skill in the area, the skill may have to be broken down into a series of stages, which are presented one at a time. The therapist then asks the patient to rehearse the new skill during the session and gives feedback. It is important that this feedback is encouraging, highlights positive aspects of the performance, and suggests how it might be improved. In some cases, video or audiotape recordings can also be used to give the patient additional feedback about progress. The patient is then asked to practice the skill. This continues until the patient is able to use the new skill in everyday life.

These techniques are frequently used in social skills training, which are often applied to groups. In couple therapy, the methods can be used in behavioral exchange therapy to encourage the couple to do positive things for each other using a 'give to get' principle. The methods may also be used to encourage positive communication between couples. Similar skills are also used in sex therapy, which is based partly on the exposure principle in anxious individuals and also aims to improve basic sexual knowledge and skill. In the case of sex therapy, the therapist does not demonstrate the skill but describes it using diagrams and the couple is asked to practice at home.

Problem-solving techniques can be used as an adjunct to other therapies for many patients. The patient is taught how to describe a problem and to weigh up the pros and cons of a decision. A discussion then follows of what an appropriate response might be to yield the best result. This is followed by rehearsal of the required skills.

Studies of sexual skills training showed it to be an effective treatment for a range of sexual problems (Masters and Johnson, 1970; Bancroft, 1983). Social skills training has had favorable results for people with social skills problems (e.g., Stravynski et al., 1982).

Operant Techniques for Chronic Problems

In patients with chronic behavior problems, operant conditioning methods are often used. Operant methods employ rewards that the patient values. The Premack (1959) principle addresses this by using a high-frequency preferred activity to reinforce a lower frequency, nonpreferred activity. As an example, if a teenage girl spends most of her time surfing the net, this high-frequency preferred activity could be used to reinforce the lower frequency nonpreferred activity of tidying her bedroom.

Use of this approach to reduce undesired behaviors and increase socially acceptable behavior has grown in recent years

since the closure of many older psychiatric institutions and a move toward community care. Discharge to the community or admission to a psychiatric unit of a district general hospital often means that bizarre or socially unacceptable behavior is tolerated less well than in older institutions, which were often segregated from the rest of society.

By far the most commonly used form of reinforcement is positive – reward. Negative reinforcement or even punishment is hardly ever used and usually only in dangerous or life-threatening situations due to the obvious ethical dilemmas presented by such methods. Ethical problems may also arise with positive reinforcement, especially if essential items such as food are used as rewards and the patient has not earned a meal that day.

Reinforcers Which Increase Desirable Behaviors

Positive Reinforcers

1. Social approval, e.g., therapist's approval of a phobic patient who has completed an exposure task.
2. Higher frequency preferred activities.
3. Feedback reinforcement, e.g., in a social skills group.
4. Food reinforcers.
5. Tokens awarded for certain behaviors, which the patient can exchange for a range of rewards.

Negative Reinforcers

This is the removal of an aversive event after a specific required response is obtained. It poses ethical problems and so is rarely used except with a consenting adult. For example, a man with an illegal sexual preference can be trained to imagine aversive scenes following arousing deviant images. He is also taught to gain relief from these aversive experiences by switching to neutral thoughts.

Reinforcers Which Reduce Undesirable Behaviors

Punishment

1. Time out, i.e., removal of the patient from a rewarding environment for up to 3 min.
2. Overcorrection – e.g., if someone smashes his cup to the floor, he is asked to not only clear up the mess but then to also wash the entire ward floor.
3. Positive punishment – e.g., if a child puts its finger toward an electric socket and receives a sharp slap on the hand.

Response Cost

1. Penalty involving some time and effort in response to certain behaviors.
2. A positive reinforcer is removed if certain undesired activities are indulged in.

The Future of Behavior Therapy

Although behavior therapy in the form of exposure remains the psychological treatment of choice for phobic, obsessive-compulsive, and other disorders, it is often combined with cognitive techniques. Many recent studies in conditions as varied as schizophrenia (Tarrier et al., 1993; Kingdon et al., 1994) and hypochondriasis (Warwick et al., 1996) describe cognitive behavior therapy. Further work is needed to identify the specific therapeutic factors in these treatment packages.

The treatment of anxiety disorders has developed rapidly since the 1980s. Anxiety can be treated using behavioral or drug treatments, and sometimes both, e.g., antidepressants as an adjunct when mood is low.

Overall, behavior therapy is of known efficacy for many psychiatric symptoms. It can be combined with cognitive techniques to improve the patient's compliance or to improve the outcome. Antidepressant drugs may also be used together with behavior and cognitive therapy. More research is needed into which ingredients of treatment packages are the most beneficial for patients.

See also: Health Behaviors; Obsessive-Compulsive Disorder; Pavlov, Ivan Petrovich (1849–1936); Psychological Treatment, Effectiveness of; Skinner, Burrhus Frederick (1904–90); Watson, John Broadus (1878–1958).

Bibliography

Ayllon, T., Azrin, N., 1968. The Token Economy: A Motivational System for Therapy and Rehabilitation. Appleton Century Crofts, New York.
Bancroft, J.H.J., 1983. Human Sexuality and its Problems. Churchill Livingstone, New York.
Drummond, L.M., 1993. The treatment of severe, chronic, resistant obsessive-compulsive disorder: an evaluation of an inpatient-programme using behavioural psychotherapy in combination with other treatments. British Journal of Psychiatry 163, 223–229.
Drummond, L.M., 2014. CBT for Adults. Royal College of Psychiatrists, London.
Emmelkamp, P.M.G., Wessels, H., 1975. Flooding in imagination vs. flooding *in vivo* – comparison with agoraphobics. Behavior Research & Therapy 13, 7–15.
Fineberg, N., Drummond, L.M., 1995. Anxiety: drug treatment or behavioural-cognitive psychotherapy. CNS Drugs 3, 448–466.
Foa, E.B., Goldstein, A., 1978. Continuous exposure and complete response prevention in the treatment of obsessive-compulsive neurosis. Behavior Therapy 9, 821–829.
Foa, E.B., Rothbaum, B.O., Riggs, D.S., Murdoch, T.B., 1991. Treatment of posttraumatic stress disorder in rape victims: a comparison between cognitive-behavioral procedures and counselling. Journal of Consulting and Clinical Psychology 59, 715–723.
Gelder, M.G., Marks, I.M., Wolff, H.H., 1967. Desensitization and psychotherapy in the treatment of phobic states: a controlled enquiry. British Journal of Psychiatry 113, 53–73.
Ghosh, A., Marks, I.M., Carr, A.C., 1988. Therapist contact and outcome of self-exposure treatment for phobias: a controlled study. British Journal of Psychiatry 152, 234–238.
Hand, I., Lamontagne, Y., Marks, I.M., 1974. Group exposure (flooding) *in vivo* for agoraphobia. British Journal of Psychiatry 124, 588–602.
Hawton, K., Salkovskis, P.M., Kirk, J., Clark, D.M., 1989. Cognitive Behaviour Therapy for Psychiatric Problems: A Practical Guide. Oxford University Press, Oxford, UK.
Janet, P., 1903. Les Obsessions et la Psychasthenie. Bailliere., Paris.
Jones, M.C., 1924. Elimination of children's fears. Journal of Experimental Psychology 7, 328.
Keane, T.M., Fairbank, J.A., Caddell, J.M., Zimering, R.T., 1989. Implosive (flooding) therapy reduces symptoms of PTSD in Vietnam combat veterans. Behavior Therapy 20, 245–260.
Kingdon, D., Turkington, D., John, C., 1994. Cognitive behaviour therapy of schizophrenia: the amenability of delusions and hallucinations to reasoning. British Journal of Psychiatry 164, 581–587.
Levy, R., Meyer, V., 1971. New techniques in behaviour therapy. Ritual prevention in obsessional patients. Proceedings of the Royal Society of Medicine 64, 115.
Marks, I.M., 1981. Cure and Care of Neurosis: Theory and Practice of Behavioural Psychotherapy. Wiley, New York.
Marks, I.M., 1986. Behavioural Psychotherapy: Maudsley Pocket Book of Clinical Management. Wright, Bristol, UK.
Marks, I.M., 2001. Living with Fear: Understanding and Coping with Anxiety. McGraw-Hill, New York.

Marks, I.M., Gray, S., Cohen, D., Hill, R., Mawson, D., Ramm, E., Stern, R., 1983. Imipramine and brief therapist-aided exposure in agoraphobics having self-exposure homework. Archives of General Psychiatry 40, 153–162.

Marks, I.M., Hodgson, R., Rachman, S., 1975. Treatment of chronic obsessive-compulsive neurosis by in-vivo exposure. Two-year follow-up and issues in treatment. British Journal of Psychiatry 127, 349–364.

Marks, I.M., Lelliot, P., Basoglu, M., Noshirvani, H., Monteiro, W., Cohen, D., Kasvikis, Y., 1988. Clomipramine, self-exposure and therapist-aided exposure for obsessive-compulsive rituals. British Journal of Psychiatry 52, 522–534.

Masters, W.H., Johnson, V.E., 1966. Human Sexual Response. Churchill, London.

Masters, W.H., Johnson, V.E., 1970. Human Sexual Inadequacy. Churchill, London.

Mathews, A.M., Gelder, M.G., Johnston, D.W., 1981. Agoraphobia: Nature and Treatment. Guilford Press, New York.

Mathews, A.M., Johnston, D.W., Lancashire, M., Munby, M., Shaw, P.M., Gelder, M.G., 1976. Imaginal flooding and exposure to real phobic situations – treatment outcome with agoraphobic patients. British Journal of Psychiatry 129, 362–371.

McDonald, R., Sartory, G., Grey, S.J., Cobb, J., Stern, R., Marks, I.M., 1978. Effects of self-exposure instructions on agoraphobic outpatients. Behaviour Research and Therapy 17, 83–85.

Meyer, V., Chesser, E.S., 1970. Behaviour Therapy in Clinical Psychiatry. Penguin, London.

Premack, D., 1959. Toward empirical behavior laws: 1. Positive reinforcement. Psychological Review 66, 219–233.

Rachman, S.J., Cobb, J., Grey, S., McDonald, R., Mawson, D., Sartory, G., Stern, R., 1979. The behavioural treatment of obsessive-compulsive disorders with and without clomipramine. Behavior Research and Therapy 17, 462–478.

Rachman, S., Hodgson, R., Marks, I.M., 1971. Treatment of chronic obsessive-compulsive neurosis. Behavior Research Therapy 9, 237–247.

Stampfl, T.J., Levis, D.G., 1967. Essentials of implosive therapy: a learning theory based psychodynamic behavior therapy. Journal of Abnormal Psychology 72, 496–503.

Stravynski, A., Marks, I.M., Yule, W., 1982. Social skills problems in neurotic outpatients. Archives of General Psychiatry 39, 1378–1385.

Stern, R.S., Drummond, L.M., 1991. The Practice of Behavioural and Cognitive Psychotherapy. Cambridge University Press, Cambridge, UK.

Stern, R.S., Marks, I.M., 1973. Brief and prolonged flooding: a comparison in agoraphobic patients. Archives of General Psychiatry 28, 270–276.

Tarrier, N., Beckett, R., Harwood, S., Baker, A., Yusupoff, L., Ugarteburu, I., 1993. A trial of two cognitive-behavioural methods of treating drug-resistant residual psychotic symptoms in schizophrenic patients: I. Outcome. British Journal of Psychiatry 162, 524–532.

Warwick, H.M.C., Clark, D.M., Cobb, A.M., Salkovskis, P.M., 1996. A controlled trial of cognitive-behavioural treatment of hypochondriasis. British Journal of Psychiatry 69, 189–195.

Watson, J.P., Marks, I.M., 1971. Relevant and irrelevant fear in flooding – a crossover study of phobic patients. Behavior Therapy 2, 275–293.

Watson, J.B., Rayner, P., 1920. Conditioned emotional reactions. Journal of Experimental Psychology 3, 1.

Wolpe, J., 1958. Psychotherapy by Reciprocal Inhibition. Stanford University Press, Stanford, CA.

Behavior Therapy: The Second and Third Waves

Kyle E Ferguson, St. Peter Family Medicine Residency Program, WA, USA; and University of Washington, Seattle, WA, USA
William O'Donohue, University of Nevada, Reno, NV, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

First-generation or first-wave behavior therapy developed out of dissatisfaction with psychoanalysis. Though it showed much promise in treating rather intractable conditions that eluded psychoanalysts at the time, it, too, was not above criticism – which gave rise to the second wave of behavior therapy. Cognitive therapy and cognitive behavior therapy (CBT) emerged out of the second wave. Cognitive therapy and CBT were successful in treating numerous complex psychiatric disorders including recurrent depression, posttraumatic stress disorder, panic disorder, and obsessive-compulsive disorder.

The third wave supposedly developed out of dissatisfaction with the second wave of behavior therapy. Acceptance and commitment therapy, dialectical behavior therapy, functional analytic psychotherapy, and behavioral activation are several of the most influential third-wave therapies. The third wave introduced a number of new concepts and terms into the behavioral canon including, metacognition, cognitive fusion, emotional regulation, acceptance, mindfulness, and dialectics. As will be taken up later, the third wave might have been oversold and may not actually fit within the general properties of an evidence-based orientation of behavior therapy.

First-generation or first-wave behavior therapy grew, in part, out of dissatisfaction with psychoanalysis in treating various intractable, chronic psychiatric conditions. Although it showed promise in treating a number of these, practitioners eventually became dissatisfied with the first wave of behavior therapy – particularly when those earlier technologies were applied to clients with more complex behavioral repertoires. Perceived shortcomings in treating certain patient populations gave rise to the second wave of behavior therapy. The second wave introduced concepts and terms from the basic laboratory of cognitive psychology and from folk notions regarding rationality and positive thinking, emphasizing the role of cognitive processes in clinical disorders. Cognitive therapy (CT; Beck, 1976; Beck et al., 1979) and cognitive behavior therapy (CBT; Ellis, 1961, 1962, 1975; Meichenbaum, 1977) were born out of this movement and were also successful in treating a variety of complex psychiatric disorders including recurrent depression, posttraumatic stress disorder (PTSD), panic disorder, and obsessive-compulsive disorder (OCD) (*see* Concepts and Methods of Cognitive Therapies for an in-depth discussion on various concepts and methods of CTs and CBTs).

The third wave supposedly developed out of dissatisfaction with second-wave cognitive approaches. The third wave is composed largely of acceptance and commitment therapy (ACT), dialectical behavior therapy (DBT), functional analytic psychotherapy (FAP), and behavioral activation (BA), among others (Hayes et al., 2006). Interestingly, as a side note, Marsha Linehan – the founder of DBT – reportedly, does not consider DBT a third-wave therapy. Rather, in her view, DBT is a "form of classic CBT that includes acceptance strategies" (David and Hofmann, 2013: p. 116). Accordingly, she would see DBT as a member of the second wave. The third wave introduced a number of new concepts and terms to the behavioral canon including, though not limited to, metacognition, cognitive fusion, emotional regulation, acceptance, mindfulness, dialectics, and commitment to meaning (Kahl et al., 2012).

The purpose of this article is to provide a general overview of second- and third-wave therapies. Readers should note that these schools of thought are nearly impossible (if not impossible) to neatly summarize using a few simple properties that pertain to each and every member. Behavior therapies, rather, share what Wittgenstein (1952) referred to as a 'family resemblance.' A family resemblance shares common, essential properties that play a central role in a complex web of belief (Quine and Ullian, 1970). Wittgenstein (1958) elucidated this concept using the notion of 'game' as a case in point:

> Compare chess with noughts and crosses. Or is there always winning and losing, or competition between players? Think of patience. In ball games there is winning and losing; but when a child throws his ball at the wall and catches it again, this feature has disappeared. Look at the parts played by skill and luck; and at the difference between skill in chess and skill in tennis… I can think of no better expression to characterize these similarities than "family resemblances" (PI § 66–67)

Accordingly, insofar as one can find family resemblances in games, commonalities can be readily seen emerging and disappearing in second- and third-wave therapies.

Second-Wave Behavior Therapy

There was a watershed event in the history of psychology during the late 1940 and the 1950s, where psychologists took one of two major paths: there were those who remained faithful to the behavioral tradition and there were those who broke ties with that tradition, pursuing cognitive psychology (Baars, 1986; Hergenhahn, 1997). Two important dates that were attributed to this breach of the behavioral tradition were 1948 and 1956 (Leahey, 1992). The former year marked the Hixon Symposium on Cerebral Mechanisms in Behavior (Gardner, 1985), and the latter marked the Symposium on Information Theory at the Massachusetts Institute of Technology (Baars, 1986). Basic research in cognitive psychology

had a significant influence on behavior therapy throughout the 1960s, even until the present day, in cognitive neuroscience. Major cognitive and cognitive behavioral approaches of the second generation of behavior therapists are Beck's (1976) CT, Ellis's rational emotive therapy (1962; though it has since been renamed rational emotive behavior therapy), Meichenbaum's (1977) self-instructional training, and, more recently, metacognitive therapy (Wells and Matthews, 1994).

CBT employs both cognitive and behavioral techniques in designing treatments. There are several assumptions underlying all CBT interventions. First, thoughts or beliefs, emotions, and behavior are interrelated. Behavior therefore affects thoughts and feelings and thoughts and feelings in turn affect behavior. Thoughts, feelings, and behavior are therefore inexplicably tied to one another. Second, changing 'automatic thoughts,' core beliefs, and/or behavior affects mood. Identifying and modifying negative or dysfunctional beliefs and/or engaging in more adaptive behavior, therefore, can eventuate in improvement in a person's mood. Three, all CBT interventions employ similar techniques built around one core strategy, cognitive restructuring, to which the authors now turn (Edmondson and Conger, 1996).

Cognitive restructuring is the core strategy in CT and most CBTs (Mahoney, 1977). It is a treatment technique that is designed to directly modify specific thoughts or beliefs that are believed to mediate maladaptive behavioral and emotional responding (Foa and Rothbaum, 1998). The patient, for example, might be assisted in identifying the so-called distorted schemas and related automatic thoughts that moderate anxiety responses (e.g., 'I'm inherently a bad person'). The therapist might assist the patient in replacing those distorted messages with positive, more realistic cognitive beliefs and messages (e.g., 'I sometimes behave badly in specific circumstances, but I also behave well in other situations').

There has been some dissatisfaction within CT regarding the bases of understanding of the general role of cognition in normal and abnormal human behavior. Ellis and Beck have generally not been influenced by the literature in basic experimental psychology. Instead, they have relied on folk notions of rationality (e.g., do not overgeneralize) and informal notions of what constitutes harmful irrational beliefs (e.g., Ellis' assertion that "Everyone must like me or else I am not a good person"). Others such as Dimidjian and Hollon (2010), Hollon et al. (1993), and McFall (1991, 1996) have suggested that this is a poor basis for clinical theories and strategies. They, in contrast, argue that CT ought to be based on experimental cognitive science – in much the same way that behavior therapy is based on the findings of learning and conditioning in the experimental laboratory.

Third-Wave Behavior Therapy

The behavioral perspective on cognitive phenomena has engendered a relatively new psychological therapy called acceptance and commitment therapy, which holds promise treating individuals with intractable conditions like chronic pain, substance abuse, and chronic suicidality (Hayes et al., 1999). Readers should note, however, that the term *promise* is used here, as ACT has yet to show superiority over extant approaches like CBT (see recent meta-analyses: Levin and Hayes, 2009; Powers et al., 2009; Ruiz, 2010, 2012). Moreover, although ACT has made some lists of empirically supported therapies, often these lists are constructed by ardent ACT supporters. Some critics, in contrast, conclude that "no third wave therapy fulfills the criteria for empirically supported treatments" (Öst, 2008: p. 315). Specifically, "The review of the RCTs for ACT shows that there is no study that is methodologically good enough and also fulfills the first of the EST criteria" (p. 313), i.e., "At least *two* good between-group design experiments must demonstrate efficacy in one or more of the following ways … superiority to pill or psychotherapy control, or to other treatment … equivalence to already established treatment with adequate sample sizes" (Chambless et al., 1998).

A third-wave perspective on cognitive phenomena as shown in ACT places a premium on how language (verbal behavior) interrelates with nonverbal behavior in a bi-directional manner (Hayes and Gifford, 1997). This mechanism is called rule-governed behavior (Hayes, 1989). Rules are special classes of verbal behavior (Baum, 1994), defined in functional terms, according to their involvement as a special class of discriminative stimuli (Catania et al., 1989: p. 120). Rules thus serve as contingency-specifying stimuli (Pierce and Epling, 1999). That is to say, rules describe what behaviors, under what conditions, will likely eventuate in positive reinforcement or punishment. For example, the rule, always showing up to work on time and pleasing one's boss will likely result in further paychecks. In contrast, the rule, always arriving late to work, and making clients wait will likely eventuate in termination.

ACT is a principles-driven treatment designed to promote behavior and thinking patterns that are more flexible and adaptive as individuals strive to keep their commitments and pursue their values (Wilson and Roberts, 2002). ACT modifies psychological functioning and behavior by altering the social and verbal contexts that are believed to link thoughts and feelings to overt actions. According to proponents of ACT, psychological inflexibility results from cognitive fusion or "excessive or improper regulation of behavior by verbal processes" (Hayes et al., 2006: p. 6). Cognitive fusion arises when words like 'I' and 'me' merge with statements like ('I' + 'am no good' or 'I' + 'do not deserve to live').

Cognitive fusion is oftentimes seamless, thus creating the illusion that no such verbal processes actually transpired (Hayes, 2004; i.e., the 'I' + what followed has always been the case). Accordingly, people are inclined to accept those self-evaluative statements prima facie, as if they were statements of facts or literal truths, not simply a thought or interpretation about how the world is arranged (Hayes and Wilson, 1994). ACT uses the terms literalization and the neologism 'deliteralization' to get at this process (Hayes et al., 1999). Experiential avoidance is more likely following cognitive fusion (Hayes et al., 1999). Individuals use drugs of abuse, increasingly isolate themselves, and, worse still, take their own lives to avoid contact with extremely painful thoughts and emotions that have fused with an individual's self-concept or sense of self-worth (Chiles and Strosahl, 2005).

Psychological flexibility – the opposite of psychological inflexibility – is promoted through six core processes of change, including acceptance, defusion, mindfulness, self as context, values, and committed action (see Strosahl et al., 2004; for an in-depth discussion). Employing mindfulness, acceptance, and cognitive defusion therapeutic strategies, ACT targets the role that experiential avoidance plays in maintaining and exacerbating psychological problems. Furthermore, contrary to traditional cognitive and CBT, ACT does not attempt to eliminate or suppress symptoms. Rather, ACT "seeks to ameliorate avoidance in the service of increasing the client's capacity to engage in a rich and meaningful life" (Wilson and Roberts, 2002: p. 238). That is, the goal of ACT is to live in accordance with one's core values as effectively as possible. Although symptom reduction often coincides with ACT therapy, symptom reduction is not the goal of ACT by any means but, perhaps, a therapeutic 'by-product' of sorts.

Although ACT is not as well established as other CBT (see Öst, 2008), preliminary findings suggest that ACT is effective for a number of conditions, including chronic pain (Dahl et al., 2004; Gutierrez et al., 2004; Johnston et al., 2010; McCracken et al., 2005; Vowles et al., 2008; Wicksell et al., 2009), depression (Forman et al., 2007), work-related stress and burnout (Bond and Bunce, 2000, 2003; Hayes et al., 2004), diagnosis acceptance (Blackledge and Hayes, 2006), panic disorder with agoraphobia (Carrascoso López, 2000), panic-related symptoms (Eifert and Heffner, 2003), generalized social anxiety disorder (Dalrymple and Herbert, 2007), self-harm (Gratz and Gunderson, 2006), relationship problems (Peterson et al., 2009), comorbid substance abuse and PTSD (Batten and Hayes, 2005), and comorbid depression and alcohol use disorders (Petersen and Zettle, 2009). Regarding meta-analytic reviews, as indicated, although ACT is superior to control and treatment as usual conditions (Hayes et al., 2006; Öst, 2008), it has not consistently been shown to be more effective than other more established treatments like CBT (Levin and Hayes, 2009; Powers et al., 2009; Ruiz, 2010).

DBT – another key player in the so-called third wave (according to some) – is arguably the most eclectic of the third-wave therapies. DBT appears much like applied behavior analysis with respect to behavioral chain analyses and basic case conceptualization (Linehan, 1993: pp. 258–264). Similarly, dialectical strategies, another core feature of DBT, are essentially cognitive reframing as seen in CT and CBT. Synthesizing theses and antitheses is one such example that bears a number of similarities to other cognitive approaches (e.g., recognizing that one's partner has both a beneficial and adverse effect on one's life; Linehan, 1993: pp. 120–124). DBT has the most empirical support with patients who engage in suicidal and parasuicidal behavior (e.g., cutting on or burning oneself) (Klonsky and Muehlenkamp, 2007). It, however, has recently been applied to other areas like eating disorders and substance abuse (Hill et al., 2011; Rizvi et al., 2011).

FAP – purportedly based on the principles of radical behaviorism – has not enjoyed the same degree of success as ACT or DBT (Kohlenberg and Tsai, 1991). Moreover, critics argue that there is little empirical support for FAP as a stand-alone therapy (Öst, 2008). Similar to the notions of transference and countertransference in psychoanalysis, FAP assumes that clinically relevant behavior (CRB) emerges in session within the context of the therapist–patient relationship. A CRB1 is the behavior that is problematic for the person (e.g., the client avoids intimacy or is quick to anger). After patients are taught to recognize CRB1s in therapy, other more adaptive, socially acceptable behavior is evoked and shaped by the therapist (CRB2). Eventually, the client or patient is taught to not only recognize and describe CRB1s and CRB2s but also elucidate the contingencies that are believed to influence them (CRB3s).

Ultimately, it is the aim of FAP to generalize this skill set shaped in the therapeutic milieu to daily life. As a side note, FAP is oftentimes used as a means of 'enhancing' or 'augmenting' CBT and ACT, as opposed to its use as a stand-alone therapy (e.g., Callaghan et al., 2004; Kohlenberg and Tsai, 1994; Kohlenberg et al., 2002). Its promise at 'enhancing' or 'augmenting' other therapies, notwithstanding, despite decades of theoretical work, FAP has produced only a modest amount of data showing its efficacy or effectiveness as a stand-alone therapy (Öst, 2008).

Charles Ferster (1973), a contemporary and major collaborator of B.F. Skinner (Ferster and Skinner, 1957), developed a functional analysis model of depression based on learning theory. Ferster maintained that individuals experience mood symptoms and engage in depressive behavior (e.g., crying, irritability, and hypersomnolence) when there is too little positive reinforcement and too much negative reinforcement (or punishment). Specifically, negative contingencies tend to narrow behavioral repertoires which, in turn, results in fewer opportunities of accessing positive reinforcers or, in common parlance, 'pleasant events.' For example, an abused spouse who avoids leaving her partner due to fear of retaliation and/or 'being alone' may never find a more suitable partner with whom she might build a more healthy relationship. Moreover, the constant threat of verbal and physical abuse that is contingent on her behaving or not behaving in accordance with her partner's rigid expectations (i.e., aversive control) narrows her behavioral repertoire, such that she almost invariably focuses only on 'doing what she is told.' The exploratory behavior of seeking a meaningful profession, advancing her education, or establishing new relationships in the area, in contrast, would most likely decrease in frequency under these prevailing contingencies. In those circumstances, of course, these women are at elevated risk for developing depression. In a recent meta-analysis, for example, women exposed to intimate partner violence were at significantly elevated risk for major depressive disorder, dysthymia, mild depression, or suicide attempts (Devries et al., 2013). Although one could not say definitively what caused their depression as these are correlational studies, one might assume that their mood symptoms are due in part to increased rates of negative reinforcement (and punishment) and decreased access to positive reinforcement.

Lewinsohn and colleagues expanded upon and laid the empirical foundations for Ferster's model (Lewinsohn et al., 1976). Lewinsohn and colleagues proposed that depressed individuals experience low rates of what they called response-contingent positive reinforcers (i.e., the behavior is causally linked to the positive reinforcer or class of reinforcers). There are several pathways by which this might

occur. First, the environment might simply change such that the reinforcer becomes less available (e.g., a best friend at work is transferred to another facility). Two, reinforcers might lose their effectiveness over time (e.g., perhaps, through habituation or other classical conditioning processes). Three, though the reinforcer is still available, the individual's behavior loses potency at accessing it, especially in the context of competitive arrangements (e.g., a potential love interest becomes more interested in a new employee whose behavior is more reinforcing of his attention). The goal of Lewinsohn et al.'s treatment is, therefore, to reinstate the rate of response-contingent positive reinforcers to previous baseline levels (or, ideally, foster higher rates). Lewinsohn and colleagues developed the *Pleasant Events Schedule* for this reason: to assist patients in identifying potential reinforcers (MacPhillamy and Lewinsohn, 1982).

Influenced by the work of Ferster and Lewinsohn and colleagues, and based on a rather compelling component analysis of CBT conducted by Neil Jacobson and colleagues (Jacobson et al., 1996) – where they found that CBT produced comparable outcomes when compared to BT alone (also see Dimidjian et al., 2006; Gortner et al., 1998, for similar findings) – Jacobson and colleagues developed a new treatment called BA therapy (Jacobson et al., 2001). Unfortunately, Jacobson passed away suddenly before his work reached fruition and was subsequently published. Accordingly, his colleagues continued the work that culminated in the book titled *Depression in Context: Strategies for Guided Action* (Martell et al., 2001).

The BA approach views depression from a contextual and functional perspective. It maintains that life events can disrupt regular behavioral patterns or, more technically, prevailing schedules or reinforcement (e.g., job loss, divorce, death of friends or family members). While some individuals tend to bounce back from adversity, other individuals – particularly those who employ passive coping styles – are more likely to develop depression. Passive coping strategies includes, though is not limited to, excessive sleeping, social isolation, excessive alcohol, or cannabis use. Although effective in the short term, these approaches often compound problems later on. Excessive sleeping might result in chronic insomnia due to disruption of one's normal sleep–wake schedule. Social isolation can eventuate in losing friends over time. Of course, chronic alcohol or cannabis use can result in abuse or dependence.

BA focuses on increasing access to reinforcement and developing more active behavioral coping strategies. In other words, BA sees avoidance behavior as the primary therapeutic target. Among other strategies, BA employs the use of various heuristics in assisting patients in first identifying problematic avoidance patterns (TRAP – Trigger, Response, Avoidance Pattern) and, second, discovering active, more effective coping styles (TRAC – Trigger, Response, Alternative Coping) (Martell et al., 2001).

While there are fewer empirical studies on BA relative to other third-wave therapies (DBT and ACT in particular), it has been shown to be effective in clinical settings, especially with depressed patients (Coffman et al., 2007; Cuijpers et al., 2007; Dimidjian et al., 2006; Jacobson et al., 2001).

Some authors, however, have argued that some of these therapies have been oversold and might not fit within the general properties of an evidence-based orientation of behavior therapy (O'Donohue et al., 2014). For example, the practice of mindfulness does not lend itself well to most behavioral interpretations. Mindfulness is neither a classical conditioning nor an operant conditioning process. Rather, mindfulness, as framed in ACT, is used to "pierce the illusion of language … like an audience that learns how a magic trick is accomplished" (Hayes, 2002). That is to say, it is a therapeutic strategy by which patients learn to contact contingencies of reinforcement and extricate themselves from obstructive rule following (e.g., "panic attacks are deadly"; Hayes, 1989; Zettle, 1990). It seems fair to say that first-generation behavior therapists were much more interested and allied with various behaviorisms (see O'Donohue and Kitchener, 1997) and this kind of speaking about the mind, for them, was quite problematic.

Conclusions

The purpose of this article was to provide a broad outline of second- and third-wave therapies. First-generation or first-wave behavior therapy developed out of dissatisfaction with psychoanalysis. Although first-wave behavior therapy showed promise in treating intractable conditions that eluded psychoanalysts at the time, it eventually gave rise to the second wave of behavior therapy. Second-wave behavior therapists were successful at treating a number of complex psychiatric disorders including recurrent depression, PTSD, panic disorder, and OCD.

The third wave supposedly developed out of dissatisfaction with the second wave of behavior therapy. The third wave introduced a number of new concepts and terms into the behavioral canon including metacognition, cognitive fusion, emotional regulation, acceptance, mindfulness, and dialectics. The third wave, of course, might have been oversold and may not actually fit within the general properties of an evidence-based orientation of behavior therapy. Moreover, while the third wave – especially ACT and DBT – have risen in stature within a short while (since the 1990s), many clinicians identify themselves as eclectic and thus employ a hybrid approach, cobbling principles and practices from CBT, ACT, and DBT (Jensen et al., 1990). However, it is fair to say that these waves also produce heterogeneity in the practice of behavior therapy as individual behavior therapists have various interests and allegiances to each of the waves.

See also: Behavior Therapy: Evidence-Based Practice, Clinical Applications, Unresolved Issues; Concepts and Methods of Cognitive Therapies.

Bibliography

Baars, B., 1986. The Cognitive Revolution in Psychology. Guilford, New York.

Batten, S.V., Hayes, S.C., 2005. Acceptance and commitment therapy in the treatment of comorbid substance abuse and post-traumatic stress disorder: a case study. Clinical Case Studies 4 (3), 246–262.

Baum, W.M., 1994. Understanding Behaviorism. Harper Collins College Publishers, New York.

Beck, A.T., 1976. Cognitive Therapy and the Emotional Disorders. Penguin Books, New York.
Beck, A.T., Rush, A.J., Shaw, B.F., et al., 1979. Cognitive Therapy of Depression. John Wiley & Sons, New York.
Blackledge, J.T., Hayes, S.C., 2006. Using acceptance and commitment training in the support of parents of children diagnosed with autism. Child and Family Behavior Therapy 28 (1), 1–18.
Bond, F.W., Bunce, D., 2000. Mediators of change in emotion-focused and problem-focused worksite stress management interventions. Journal of Occupational Health Psychology 5, 156–163.
Bond, F.W., Bunce, D., 2003. The role of acceptance and job control in mental health, job satisfaction, and work performance. Journal of Applied Psychology 88, 1057–1067.
Callaghan, G.M., Gregg, J.A., Marx, B., Kohlenberg, B.S., Gifford, E., 2004. FACT: the utility of an integration of functional analytic psychotherapy and acceptance and commitment therapy to alleviate human suffering. Psychotherapy: Theory, Research, Practice, Training 41, 195–207.
Carrascoso López, F.J., 2000. Acceptance and commitment therapy (ACT) in panic disorder with agoraphobia: a case study. Psychology in Spain 4 (1), 120–128.
Catania, A.C., Shimoff, E., Matthews, B.A., 1989. An experimental analysis of rule-governed behavior. In: Hayes, S.C. (Ed.), Rule-Governed Behavior: Cognition, Contingencies, and Instructional Control. Plenum Press, New York, pp. 119–150.
Chambless, D.L., et al., 1998. Update on empirically validated therapies, II. The Clinical Psychologist 51, 3–16.
Chiles, J.A., Strosahl, K.D., 2005. Clinical Manual for Assessment and Treatment of Suicidal Patients. American Psychiatric, Washington, DC.
Coffman, S., Martell, C.R., Dimidjian, S., Gallop, R., Hollon, S.D., 2007. Extreme non-response in cognitive therapy: can behavioral activation succeed where cognitive therapy fails? Journal of Consulting and Clinical Psychology 75, 531–541.
Cuijpers, P., van Straten, A., Warmerdam, L., 2007. Behavioral activation treatments of depression: a meta-analysis. Clinical Psychology Review 27, 318–326.
Dahl, J., Wilson, K.G., Nilsson, A., 2004. Acceptance and commitment therapy and the treatment of persons at risk for long-term disability resulting from stress and pain symptoms: a preliminary randomized trial. Behavior Therapy 35, 785–802.
Dalrymple, K.L., Herbert, J.D., 2007. Acceptance and commitment therapy for generalized social anxiety disorder: a pilot study. Behavior Modification 31, 543–568.
David, D., Hofmann, S.G., 2013. Another error of Descartes? Implications for the "third wave" cognitive-behavioral therapy. Journal of Cognitive and Behavioral Psychotherapies 13 (1), 115–124.
Devries, K.M., Mak, J.Y., Bacchus, L.J., Child, J.C., Falder, G., Petzold, M., Astbury, J., Watts, C.H., 2013. Intimate partner violence and incident depressive symptoms and suicide attempts: a systematic review of longitudinal studies. PLoS Med 10 (5), e1001439.
Dimidjian, S., Hollon, S.D., 2010. How would you know if psychotherapy was harmful? American Psychologist 65, 21–33.
Dimidjian, S., Hollon, S.D., Dobson, K.S., Schmaling, K.B., Kohlenberg, R., Addis, M., Jacobson, N.S., 2006. Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. Journal of Consulting and Clinical Psychology 74, 658–670.
Edmondson, C.B., Conger, J.C., 1996. A review of treatment efficacy for individuals with anger problems: conceptual, assessment, and methodological issues. Clinical Psychology Review 16, 251–275.
Eifert, G.H., Heffner, M., 2003. The effects of acceptance versus control contexts on avoidance of panic-related symptoms. Journal of Behavior Therapy and Experimental Psychiatry 34, 293–312.
Ellis, A., 1961. A Guide to Rational Living. Prentice-Hall, Englewood Cliffs, NJ.
Ellis, A., 1962. Reason and Emotion in Psychotherapy. Lyle Stuart, New York.
Ellis, A., 1975. A New Guide to Rational Living. Wilshire Book Company, North Hollywood.
Ferster, C.B., 1973. A functional analysis of depression. American Psychologist 28, 857–870.
Ferster, C.B., Skinner, B.F., 1957. Schedules of Reinforcement. Appleton-Century-Crofts, New York.
Foa, E.,B., Rothbaum, B.O., 1998. Treating the Trauma of Rape: Cognitive-Behavioral Therapy for PTSD. The Guilford Press, New York.
Forman, E.M., Herbert, J.D., Moitra, E., Yeomans, P.D., Geller, P.A., 2007. A randomized controlled effectiveness trial of acceptance and commitment therapy and cognitive therapy for anxiety and depression. Behavior Modification 31 (6), 772–799.
Gardner, H., 1985. The Mind's New Science: A History of the Cognitive Revolution. Basic Books, New York.
Gortner, E.T., Gollan, J.K., Dobson, K.S., et al., 1998. Cognitive behavioral treatment for depression: relapse prevention. Journal of Consulting and Clinical Psychology 66, 377–384.
Gratz, K.L., Gunderson, J.G., 2006. Preliminary data on an acceptance-based emotion regulation group intervention for deliberate self-harm among women with borderline personality disorder. Behavior Therapy 37, 25–35.
Gutierrez, O., Luciano, C., Fink, B.C., 2004. Comparison between an acceptance-based and a cognitive-control-based protocol for coping with pain. Behavior Therapy 35, 767–784.
Hayes, S.C. (Ed.), 1989. Rule-Governed Behavior: Cognition, Contingencies, and Instructional Control. Plenum, New York.
Hayes, S., 2002. Acceptance, mindfulness, and science. Clinical Psychology: Science and Practice 9 (1), 101–106.
Hayes, S.C., 2004. Acceptance and commitment therapy, relational frame therapy, and the third wave of behavioural and cognitive therapies. Behavior Therapy 35, 639–665.
Hayes, S.C., Bissett, R., Roget, N., Padilla, M., Kohlenberg, B.S., Fisher, G., et al., 2004. The impact of acceptance and commitment training on stigmatizing attitudes and professional burnout of substance abuse counselors. Behavior Therapy 35, 821–836.
Hayes, S.C., Gifford, E.V., 1997. The trouble with language: experiential avoidance, rules, and the nature of verbal events. Psychological Science 8 (3), 170–173.
Hayes, S.C., Luoma, J.B., Bond, F.W., Masuda, A., Lillis, J., 2006. Acceptance and commitment therapy: model, processes and outcomes. Behaviour Research and Therapy 44, 1–25.
Hayes, S.C., Strosahl, K.D., Wilson, K.G., 1999. Acceptance and Commitment Therapy: An Experiential Approach to Behavior Change. Guilford, New York.
Hayes, S.C., Wilson, K.G., 1994. Acceptance and commitment therapy: altering the verbal support for experiential avoidance. Behavior Analyst 17, 289–303.
Hergenhahn, B.R., 1997. An Introduction to the History of Psychology, third ed. Brooks/Cole, Pacific Grove.
Hill, D.M., Craighead, L., Safer, D.L., 2011. Appetite-focused dialectical behavior therapy for the treatment of binge eating with purging: a preliminary trial. International Journal of Eating Disorders 44 (3), 249–261.
Hollon, S.D., Shelton, R.C., Davis, D.D., 1993. Cognitive therapy for depression: conceptual issues and clinical efficacy. Journal of Consulting and Clinical Psychology 61, 270–275.
Jacobson, N.S., Dobson, K.S., Truax, P.A., Addis, M.E., Koerner, K., Gollan, J.K., Gortner, E., Prince, S.E., 1996. A component analysis of cognitive-behavioral treatment for depression. Journal of Consulting and Clinical Psychology 64, 295–304.
Jacobson, N.S., Martell, C.R., Dimidjian, S., 2001. Behavioral activation treatment for depression: returning to contextual roots. Clinical Psychology: Science and Practice 8, 255–270.
Jensen, J.P., Bergin, A.E., Greaves, D.W., 1990. The meaning of eclecticism: new survey and analysis of components. Professional Psychology: Research and Practice 21 (2), 124–130.
Johnston, M., Foster, M., Shennan, J., Starkey, N.J., Johnson, A., 2010. The effectiveness of an acceptance and commitment therapy self-help intervention for chronic pain. Clinical Journal of Pain 26, 393–402.
Kahl, K.G., Winter, L., Schweiger, U., 2012. The third wave of cognitive behavioural therapies: what is new and what is effective? Current Opinion in Psychiatry 25 (6), 522–528.
Klonsky, E., Muehlenkamp, J., 2007. Self-injury: a research review for the practitioner. Journal of Clinical Psychology 63 (11), 1045–1056.
Kohlenberg, R.J., Kanter, J.W., Bolling, M.Y., Parker, C., Tsai, M., 2002. Enhancing cognitive therapy for depression with functional analytic psychotherapy: treatment guidelines and empirical findings. Cognitive and Behavioral Practice 9 (3), 213–229.
Kohlenberg, R.J., Tsai, M., 1991. Functional Analytic Psychotherapy: Creating Intense and Curative Therapeutic Relationships. Plenum, New York.
Kohlenberg, R.J., Tsai, M., 1994. Improving cognitive therapy for depression with functional analytic psychotherapy: theory and case study. The Behavior Analyst 17, 305–319.
Last, C.G., 1985. Cognitive restructuring. In: Bellack, A.S., Hersen, M. (Eds.), Dictionary of Behavior Therapy Techniques. Pergamon, New York, pp. 59–60.
Leahey, T.H., 1992. The mythical revolutions of American psychology. American Psychologist 47, 308–318.
Levin, M., Hayes, S.C., 2009. Is acceptance and commitment therapy superior to established treatment comparisons? Psychotherapy & Psychosomatics 78, 380.
Lewinsohn, P.M., Biglan, A., Zeiss, A.M., 1976. Behavioral treatment of depression. In: Davidson, P.O. (Ed.), The Behavioral Management of Anxiety, Depression and Pain. Brunner/Mazel, New York, pp. 91–146.

Linehan, M.M., 1993. Cognitive Behavioral Treatment of Borderline Personality Disorder. Guilford Press, New York.

MacPhillamy, D.J., Lewinsohn, P.M., 1974. Depression as a function of levels of desired and obtained pleasure. Journal of Abnormal Psychology 83, 651–657.

MacPhillamy, D.J., Lewinsohn, P.M., 1982. The pleasant events schedule: Studies on reliability, validity, and scale intercorrelation. Journal of Consulting and Clinical Psychology 50 (3), 363–380.

Mahoney, M.J., 1977. Reflections on the cognitive learning trend in psychotherapy. American Psychologist 32, 5–13.

Martell, C.R., Addis, M.E., Jacobson, N.S., 2001. Depression in Context: Strategies for Guided Action. W.W. Norton, New York.

McCracken, L.M., Vowles, K.E., Eccleston, C., 2005. Acceptance-based treatment for persons with complex, long-standing chronic pain: a preliminary analysis of treatment outcome in comparison to a waiting phase. Behaviour Research and Therapy 43, 1335–1346.

McFall, R.M., 1991. Manifesto for a science of clinical psychology. Clinical Psychologist 44, 75–88.

McFall, R.M., 1996. Elaborate reflections on a simple manifesto. Applied and Preventive Psychology 9, 5–21.

Meichenbaum, D.H., 1977. Cognitive Behavior Modification: An Integrative Approach. Plenum, New York.

O'Donohue, W., Kitchener, R.F. (Eds.), 1997. Philosophy of Psychology. Sage, London.

O'Donohue, W., Snipes, C., Dalto, G., Soto, C., Maragakis, A., Im, S., 2014. The ethics of enhanced interrogations and torture: a reappraisal of the argument. Ethics and Behavior 24 (2), 109–125.

Öst, L.G., 2008. Efficacy of the third wave of behavioral therapies: a systematic review and meta-analysis. Behaviour Research and Therapy 46, 296–321.

Peterson, B.D., Eifert, G.H., Feingold, T., Davidson, S., 2009. Using acceptance and commitment therapy to treat distressed couples: a case study with two couples. Cognitive and Behavioral Practice 16, 430–442.

Petersen, C.L., Zettle, R.D., 2009. Treating inpatients with comorbid depression and alcohol use disorders: a comparison of acceptance and commitment therapy and treatment as usual. The Psychological Record 59, 521–536.

Pierce, W.D., Epling, W.F., 1999. Behavior Analysis and Learning, second ed. Prentice Hall, Upper Saddle River, NJ.

Powers, M.B., Zum Vörde Sive Vörding, M.B., Emmelkamp, P.M., 2009. Acceptance and commitment therapy: a meta-analytic review. Psychotherapy and Psychosomatics 78, 73–80.

Quine, W.V., Ullian, J.S., 1970. The Web of Belief. Random House Basil Blackwell, New York.

Rizvi, S.L., Dimeff, L.A., Skutch, J., Carroll, D., Linehan, M.M., 2011. A pilot study of the DBT coach: an interactive mobile phone application for individuals with borderline personality disorder and substance use disorder. Behavior Therapy 42, 589–600.

Ruiz, F.J., 2010. A review of acceptance and commitment therapy (ACT) empirical evidence: correlational, experimental psychopathology, component and outcome studies. International Journal of Psychology and Psychological Therapy 10, 125–162.

Ruiz, F., 2012. Acceptance and commitment therapy versus traditional cognitive behavioral therapy: a systematic review and meta-analysis of current empirical evidence. Journal of Psychology and Psychological Therapy 12, 333–357.

Strosahl, K.D., Hayes, S.C., Wilson, K.G., Gifford, E.V., 2004. An ACT primer: core therapy processes, intervention strategies, and therapist competencies. In: Hayes, S.C., Strosahl, K.D. (Eds.), A Practical Guide to Acceptance and Commitment Therapy. Springer, New York, pp. 31–58.

Vowles, K.E., Loebach-Wetherell, J., Sorrell, J.T., 2008. Targeting acceptance, mindfulness, and values based action in chronic pain: findings of two preliminary trials of an outpatient group-based intervention. Cognitive and Behavioral Practice 16, 49–58.

Wells, A., Matthews, G., 1994. Attention and Emotion: A Clinical Perspective. Erlbaum, Hove, UK.

Wicksell, R.K., Melin, L., Lekander, M., Olsson, G.L., 2009. Evaluating the effectiveness of exposure and acceptance strategies to improve functioning and quality of life in longstanding pediatric pain – a randomized controlled trial. Pain 141, 248–257.

Wilson, K.G., Roberts, M., 2002. Core principles in acceptance and commitment therapy: an application to anorexia. Cognitive and Behavioral Practice 9, 237–243.

Wittgenstein, L., 1952. Philosophical Investigations. Basil Blackwell, Oxford.

Wittgenstein, L., 1958. Philosophical Investigations (G.E.M. Anscombe, Trans.). Basil Blackwell, Oxford.

Zettle, R.D., 1990. Rule-governed behavior: a radical behavioral answer to the cognitive challenge. The Psychological Record 40, 41–49.

Behavioral Economics

Sendhil Mullainathan, Massachusetts Institute of Technology, Boston, MA, USA
Richard H Thaler, University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 2, pp. 1094–1100, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Behavioral economics is the combination of psychology and economics that investigates what happens in markets in which some of the agents display human limitations and complications. We begin with a preliminary question about relevance. Does some combination of market forces, learning, and evolution render these human qualities irrelevant? No. Because of limits of arbitrage, less than perfect agents survive and influence market outcomes. We then discuss three important ways in which humans deviate from the standard economic model. Bounded rationality reflects the limited cognitive abilities that constrain human problem solving. Bounded willpower captures the fact that people sometimes make choices that are not in their long-term interest. Bounded self-interest incorporates the comforting fact that humans are often willing to sacrifice their own interests to help others. We then illustrate how these concepts can be applied in two settings: finance and savings. Financial markets have greater arbitrage opportunities than other markets, so behavioral factors might be thought to be less important here, but we show that even here the limits of arbitrage create anomalies that the psychology of decision making helps to explain. Since saving for retirement requires both complex calculations and willpower, behavioral factors are essential elements of any complete descriptive theory.

It says something interesting about the field of economics that there is a subfield called behavioral economics. Surely all of economics is meant to be about the behavior of economic agents, be they firms or consumers, suppliers or demanders, bankers or farmers. So, what is behavioral economics, and how does it differ from the rest of economics?

Economics traditionally conceptualizes a world populated by calculating, unemotional maximizers that have been dubbed *Homo economicus*. In a sense, neoclassical economics has defined itself as explicitly 'antibehavioral.' Indeed, virtually all the behavior studied by cognitive and social psychologists is either ignored or ruled out in a standard economic framework. This unbehavioral economic agent has been defended on numerous grounds: some claimed that the model was 'right'; most others simply argued that the standard model was easier to formalize and practically more relevant.

Behavioral economics blossomed with the realization that neither point of view was correct. Empirical and experimental evidence mounted against the stark predictions of unbounded rationality. Further work made clear that one could formalize psychological ideas and translate them into testable predictions. The behavioral economics research program has consisted of two components: (1) identifying the ways in which behavior differs from the standard model and (2) showing how this behavior matters in economic contexts.

This article gives a flavor of the program. We begin by discussing the most important ways in which the standard economic model needs to be enriched. We then illustrate how behavioral economics has been fruitfully applied to two important fields: finance and savings. However, first, we discuss why the market forces and learning do not eliminate the importance of human actions.

Is Homo Economicus the Only One Who Survives?

Many economists have argued that a combination of market forces (competition and arbitrage) plus evolution should produce a world similar to that described in an economics textbook: do only the rational agents survive? Or, do the workings of markets at least render the actions of the quasi-rational irrelevant? These are questions that have been much studied since the 1980s, and the early impressions of many economists that markets would wipe out irrationality were optimistic.

Consider a specific example: human capital formation. Suppose that a young economist, named Sam, decides to become a behavioral economist, perhaps because he mistakenly thinks this will lead to riches, or because he thinks it is going to be the next fad, or because he finds it interesting and lacks the willpower to study 'real' economics. Whatever the reason for the choice, let us assume for the sake of argument that this decision was a mistake for Sam by any rational calculation. So, what will market forces do? First, Sam may be poorer because of this choice than if he had sensibly chosen to study corporate finance, but he will not be destitute. Sam might even realize that he could switch to corporate finance and make much more money but is simply unable to resist the temptation to continue wasting his time on behavioral economics. So, markets per se do not necessarily solve the problem: they provide an *incentive* to switch, but they cannot force Sam's hand.

What about arbitrage? In this case, like most things we study in economics outside the realm of financial markets, there is simply no arbitrage opportunity available. Suppose a wise arbitrageur is watching Sam's choices, what bet can he or she place? None. The same can be said if Sam saves too little for retirement, picks the wrong wife, or buys the wrong car. None of these irrational acts generates an arbitrage opportunity for anyone else.

Indeed, economists now realize that even in financial markets there are important limits to the workings of arbitrage. First, in the face of irrational traders, the arbitrageur may privately benefit more from trading that helps push prices in the wrong direction than from trading that pushes prices in the

right direction. Put another way, it may often pay 'smart money' to follow 'dumb money' rather than to lean against it (Haltiwanger and Waldman, 1985; Russell and Thaler, 1985). For example, an extremely smart arbitrageur near the beginning of the tulip mania would have profited more from buying tulips and further destabilizing prices than by shorting them. Second, and slightly related, arbitrage is an inherently risky activity and consequently the supply of arbitrage will be inherently limited (De Long et al., 1990). Arbitrageurs who did decide to short tulips early would probably have been wiped out by the time their bets were proven to be 'right.' Add to this the fact that in practice most arbitrageurs are managing other people's money and therefore judged periodically, one can see the short horizons that an arbitrageur will be forced to take on. This point was made forcefully by Shleifer and Vishny (1997), who essentially foresaw the scenario that ended up closing Long-term Capital Management.

So, markets per se cannot be relied upon to make economic agents rational. What about evolution? An old argument was that individuals who failed to maximize should have been weeded out by evolutionary forces, which presumably operated during ancient times. Overconfident hunters, for example, presumably caught less prey, ate less, and died younger. Such reasoning, however, has turned out to be faulty. Evolutionary arguments can just as readily explain over-confidence as they can explain appropriate levels of confidence. For example, consider individuals playing a war of attrition (perhaps in deciding when to back down during combat). Here over-confidence will actually help. Seeing the overconfidence, a rational opponent will actually choose to back down sooner. As can be seen from this example, depending on the initial environment (especially when these environments have a game theoretic component to them), evolution may just as readily weed out rational behavior as it does weed out quasi-rational behavior. The troubling flexibility of evolutionary models means that they can just as readily argue for bounds on rationality.

The final argument is that individuals who systematically and consistently make the same mistake will eventually learn the error of their ways. This kind of argument has also not stood up well under theoretical scrutiny. First, the optimal experimentation literature has shown that there can be a complete lack of learning even in infinite horizons. The intuition here is simple: as long as there are some opportunity costs to learning or to experimenting with a new strategy, even a completely 'rational' learner will choose not to experiment. This player will get stuck in a non-optimal equilibrium, simply because the cost of trying something else is too high. Second, work on learning in games has formally demonstrated Keynes' morbid observation on the 'long run.' The time required to converge to an equilibrium strategy can be extremely long. Add to this a changing environment and one can easily be in a situation of perpetual nonconvergence. In practice, for many of the important decisions we make, both arguments apply with full force. The number of times we get to learn from our retirement decisions is low (and possibly zero). The opportunity cost of experimenting with different ways of choosing a career can be very high.

The upshot of all these theoretical innovations has been clear. One cannot defend unbounded rationality on purely theoretical grounds. Neither arbitrage, competition, evolution, nor learning necessarily guarantees that unbounded rationality must be an effective model. In the end, as some might have expected, it must ultimately be an empirical issue. Does 'behavior' matter? Before evaluating this question in two different fields of application, we explore the ways in which real behavior differs from the stylized neoclassical model.

Three Bounds of Human Nature

The standard economic model of human behavior includes (at least) three unrealistic traits: unbounded rationality, unbounded willpower, and unbounded selfishness. These three traits are good candidates for modification.

Herbert Simon (1955) was an early critic of modeling economic agents as having unlimited information processing capabilities. He suggested the term 'bounded rationality' to describe a more realistic conception of human problem solving capabilities. As stressed by Conlisk (1996), the failure to incorporate bounded rationality into economic models is just bad economics – the equivalent to presuming the existence of a free lunch. Since we have only so much brainpower, and only so much time, we cannot be expected to solve difficult problems optimally. It is eminently 'rational' for people to adopt rules of thumb as a way to economize on cognitive faculties. Yet the standard model ignores these bounds and hence the heuristics commonly used. As shown by Kahneman and Tversky (1974), this oversight can be important since sensible heuristics can lead to systematic errors.

Departures from rationality emerge both in judgments (beliefs) and in choice. The ways in which judgment diverges from rationality is long and extensive (Kahneman et al., 1982). Some illustrative examples include overconfidence, optimism, anchoring, extrapolation, and making judgments of frequency or likelihood based on salience (the availability heuristic) or similarity (the representativeness heuristic).

Many of the departures from rational choice are captured by prospect theory (Kahneman and Tversky, 1979), a purely descriptive theory of how people make choices under uncertainty (see Starmer (2000) for a review of literature on non-EU theories of choice). Prospect theory is an excellent example of a behavioral economic theory in that its key theoretical components incorporate important features of psychology. Consider three features of the prospect theory value function. (1) It is defined over changes to wealth rather than levels of wealth (as in EU) to incorporate the concept of adaptation. (2) The loss function is steeper than the gain function to incorporate the notion of 'loss aversion'; the notion that people are more sensitive to decreases in their well being than to increases. (3) Both the gain and loss function display diminishing sensitivity (the gain function is concave, the loss function convex) to reflect experimental findings. To describe fully choices prospect theory often needs to be combined with an understanding of 'mental accounting' (Thaler, 1985). One needs to understand when individuals faced with separate gambles treat them as separate gains and losses and when they treat them as one, pooling them to produce one gain or loss.

A few examples can illustrate how these concepts are used in real economics contexts. Consider overconfidence. If investors are overconfident in their abilities, they will be willing to make

trades even in the absence of true information. This insight helps explain a major anomaly of financial markets. In an efficient market when rationality is common knowledge, there is virtually no trading, but in actual markets there are hundreds of millions of shares traded daily and most professionally managed portfolios are turned over once a year or more. Individual investors also trade a lot: they incur transaction costs and yet the stocks they buy subsequently do worse than the stocks they sell.

An example involving loss aversion and mental accounting is Camerer et al.'s (1997) study of New York City cab drivers. These cab drivers pay a fixed fee to rent their cabs for 12h and then keep all their revenues. They must decide how long to drive each day. A maximizing strategy is to work longer hours on good days (days with high earnings per hour such as rainy days or days with a big convention in town) and to quit early on bad days. However, suppose cabbies set a target earnings level for each day, and treat shortfalls relative to that target as a loss. Then, they will end up quitting early on good days and working longer on bad days, precisely the opposite of the rational strategy. This is exactly what Camerer et al. found in their empirical work.

Having solved for the optimum, *Homo economicus* is next assumed to choose the optimum. Real humans, even when they know what is best, sometimes fail to choose it for self-control reasons. Most of us at some point have eaten, drunk, or spent too much, and exercised, saved, or worked too little: such is human nature. People (even economists) also procrastinate. We are completing this article after the date on which it was due, and we are confident that we are not the only guilty parties. Although people have these self-control problems, the are at least somewhat aware of them: they join diet plans and buy cigarettes by the pack (because having an entire carton around is too tempting). They also pay more withholding taxes than they need to (in 1997, nearly 90 million tax returns paid an average refund of around $1300) in order to assure themselves a refund, but then file their taxes near midnight on April 15 (at the post office that is being kept open late to accommodate their fellow procrastinators).

Finally, people are boundedly selfish. Although economic theory does not rule out altruism, as a practical matter economists stress self-interest as the primary motive. For example, the free rider problems widely discussed in economics are predicted to occur because individuals cannot be expected to contribute to the public good unless their private welfare is thus improved. In contrast, people often take selfless actions. In 1993, 73.4% of all households gave some money to charity, the average dollar amount being 2.1% of household income. Also, 47.7% of the population does volunteer work with 4.2h per week being the average time volunteered. Similar selfless behavior is observed in controlled laboratory experiments. Subjects systematically often cooperate in public goods and prisoners dilemma games, and turn down unfair offers in 'ultimatum' games.

Finance

If economists had been polled around 1980 and asked to name the domain in which bounded rationality was least likely to find useful applications the likely winner would have been finance. The limits of arbitrage arguments were not well understood at that time, and one leading economist had called the efficient markets hypothesis the best established fact in economics. Times change. As we begin the twenty-first century finance is perhaps the branch of economics where behavioral economics has made the greatest contributions. How has this happened?

Two factors contributed to the surprising success of behavioral finance. First, financial economics in general, and the efficient market hypothesis in particular, generated sharp, testable predictions about observable phenomena. Second, there are many data readily available to test these sharp predictions. We briefly summarize here a few examples.

The rational efficient markets hypothesis makes two classes of predictions about stock price behavior. The first is that stock prices are 'correct' in the sense that asset prices reflect the true or rational value of the security. In many cases this tenet of the efficient market hypothesis is untestable because intrinsic values are not observable. However, in some special cases the hypothesis can be tested by comparing two assets whose relative intrinsic values are known. One class of these is called 'Siamese Twins': two versions of the same stock that trade in different places.

A specific well-known example is the case of Royal Dutch Shell as documented in Froot and Dabora (1999). The facts are that Royal Dutch Petroleum and Shell Transport are independently incorporated in The Netherlands and the UK, respectively. The current firm emerged from a 1907 alliance between Royal Dutch and Shell Transport in which the two companies agreed to merge their interests on a 60:40 basis. Royal Dutch trades primarily in the USA and The Netherlands and Shell trades primarily in London. According to any rational model, the shares of these two components (after adjusting for foreign exchange) should trade in a ratio of 60:40. They do not; the actual price ratio has deviated from the expected one by more than 35%. Simple explanations such as taxes and transactions costs cannot explain the disparity (see Froot and Dabora, 1999). This example illustrates that prices can diverge from intrinsic value because of limits of arbitrage. Some investors do try to exploit this mispricing, buying the cheaper stock and shorting the more expensive one, but this is not a sure thing, as many hedge funds learned in the summer of 1998 (when, at the time hedge funds were trying to get more liquidity, the pricing disparity widened).

The Royal Dutch Shell anomaly is a violation of one of the most basic principles of economics: the law of one price. Another similar example is the case of closed-end mutual funds (Lee et al., 1991). These funds are much like typical (open-end) mutual funds except that to cash out of the fund, investors must sell their shares on the open market. This means that closed-end funds have market prices that are determined by supply and demand, rather than set equal to the value of their assets by the fund managers as in an open-end fund. Since the holdings of closed-end funds are public, market efficiency would lead one to expect that the price of the fund should match the price of the underlying securities they hold (the net asset value or NAV). Instead, closed-end funds typically trade at substantial discounts relative to their NAV, and occasionally at substantial premiums. Most interesting from a behavioral perspective is that closed-end fund discounts are correlated

with one another and appear to reflect individual investor sentiment. (Closed-end funds are owned primarily by individual investors rather than institutions.) Lee et al. found that discounts shrink in months when shares of small companies (also owned primarily by individuals) do well, and in months when there is lots of IPO activity, indicating a 'hot' market. Since these findings were predicted by their theory, they move the research beyond the demonstration of an embarrassing fact (price not equal to NAV) toward a constructive understanding of how markets work.

The second principle of the efficient market hypothesis is 'unpredictability.' In an efficient market it is not possible to predict future stock price movements based on publicly available information. Many early violations of this had no explicit link to behavior. Thus it was reported that small firms, firms with low price–earnings ratios earned higher returns than other stocks with the same risk. Also, stocks in general, but especially stocks of small companies, have done well in January and on Fridays (but poorly on Mondays).

An early study by De Bondt et al. (1985) was explicitly motivated by the psychological finding that individuals tend to over-react to new information. For example, experimental evidence suggests that people tended to underweight base-rate data (or prior information) in incorporating new data. De Bondt et al. hypothesized that if investors displayed this behavior, then stocks that had performed fairly well over a period of years will eventually have prices that are too high. Individuals over-reacting to the good news will drive the prices of these stocks too high. Similarly, poor performers will eventually have prices that are too low. This yields a prediction about future returns: past 'winners' ought to underperform while past 'losers' ought to outperform the market. Using data for stocks traded on the New York Stock Exchange, De Bondt et al. (1985) found that the 35 stocks that had performed the worst over the past 5 years (the losers) outperformed the market over the next 5 years, while the 35 biggest winners over the past 5 years subsequently underperformed. Follow-up studies have shown that these early results cannot be attributed to risk (by some measures the portfolio of losers is actually less risky than the portfolio of winners), and can be extended to other measures of over-reaction such as the ratio of market price to the book value of equity.

More recent studies have found other violations of unpredictability that have the opposite pattern from that found by DeBondt et al., namely under-reaction rather than over-reaction. Over short periods of time, e.g., 6 months to 1 year, stocks display momentum – the stocks that go up the fastest for the first 6 months of the year tend to keep going up. Also, after many corporate announcements such as large earnings changes, dividend initiations and omissions, share repurchases, splits, and seasoned equity offerings, there is an initial price jump on the day of the announcement followed by a slow drift in the same direction for as long as 1 year or more (Shleifer, 2000).

These findings of under-reaction are a further challenge to the efficient markets hypothesis, but also to behavioral finance. Do markets sometimes over-react, and sometimes under-react? If so, can *any* pattern be explained, at least *ex post*? Is there a unifying framework that can bring together these apparently opposing facts and ideas?

Work is only now beginning to attack this extremely important question. Several explanations have beenoffered recently (Shleifer (2000) summarizes them). All rely on psychological evidence (in one way or another) in unifying the facts. They all explain the anomalies by noting that under-reaction appears at short horizons whereas over-reaction appears at longer horizons, but each paper provides its own distinctive explanation. Which (if any) of these is the best one has yet to be decided. But the facts discovered so far, combined with these models, demonstrate the changing nature of finance. Rigorous empirical work building on psychological phenomena has given us new tools to unearth interesting empirical facts. Rigorous theoretical work, on the other hand, has been trying to address the challenge of incorporating these empirical facts into a psychologically plausible model.

The research discussed so far has been about asset prices and the controversy about the efficient market hypothesis. There is another stream of research that is just about investor behavior, not about prices. One example of this stream is motivated by mental accounting and loss aversion. The issue is whether investors are reluctant to realize capital losses (because they would have to 'declare' the loss to themselves). Shefrin and Statman (1985) dubbed this hypothesis the 'disposition effect.' The prediction that investors will be less willing to sell a loser than a winner is striking since the tax law encourages just the opposite behavior. Nevertheless, Odean (1998) finds evidence of just this behavior. In his sample of the customers of a discount brokerage firm, investors were more likely to sell a stock that had increased in value than one that had decreased in value. While around 15% of all gains were realized, only 10% of all losses are realized. This hesitancy to realize gains came at a cost. Odean shows that the stocks the loser stocks held under-performed the gainer stocks that were sold.

Savings

If finance was the field in which a behavioral approach was least likely, *a priori*, to succeed, saving had to be one of the most promising. The standard life-cycle model of savings abstracts from both bounded rationality and bounded willpower, yet saving for retirement is both a difficult cognitive problem and a difficult self-control problem. It is then, perhaps less surprising that a behavioral approach has been fruitful here. As in finance, progress has been helped by the combination of a refined standard theory with testable predictions and many data sources on household saving behavior. One crisp prediction of the life-cycle model is that savings rates are independent of income. Suppose twins Tom and Ray are identical in every respect except that Tom earns most of his money early in his life (e.g., he is a basketball player), whereas Ray earns most of his late in life (e.g., he is a manager). The life-cycle model predicts that Tom the basketball player ought to save his early income to increase consumption later in life, whereas Ray the manager ought to borrow from his future income to increase consumption earlier in life. This prediction is completely unsupported by the data, which show that consumption very closely tracks income over individuals' life cycles. Furthermore, the

departures from predicted behavior cannot be explained merely by people's inability to borrow. Banks et al. (1998) show, for example, that consumption drops sharply as individuals retire and their incomes drop. They have simply not saved enough for retirement. Indeed, many low- to middle-income families have essentially no savings whatsoever. The primary cause of this lack of saving appears to be self-control. One piece of evidence supporting this conclusion is that virtually all saving done by Americans is accomplished in vehicles that support what are often called 'forced savings,' e.g., accumulating home equity by paying the mortgage and participation in pension plans. Coming full circle, one 'forced' savings that individuals may choose themselves may be high tax withholdings, so that when the refund comes they can buy something they might not have had the willpower to save up for.

One of the most interesting research areas has been devoted to measuring the effectiveness of tax-subsidized savings programs such as IRAs and 401(k)s. The standard analysis of these programs is fairly simple. Consider the original IRA program from the early 1980s. This program provided tax subsidies for savings up to a threshold, often $2000 per year; there was no marginal incentive to save for any household that was already saving more than $2000 per year. Therefore, those saving more than the threshold should not increase their total saving, they will merely switch some money from a taxable account to the IRA for the tax gain. Moreover, everyone who is saving a positive amount should participate in this program for the infra-marginal gain. The actual analysis of these programs has shown that the reality is not so clear. By some accounts at least, these programs appear to have generated considerable new savings. Some argue that almost every dollar of savings in IRAs appears to represent new savings. In other words, people are not simply shifting their savings into IRAs and leaving their total behavior unchanged. Similar results are found for 401(k) plans. The behavioral explanation for these findings is that IRAs and 401(k) plans help solve self-control problems by setting up special mental accounts that are devoted to retirement savings. Households tend to respect the designated use of these accounts, and their self-control is helped by the tax penalty that must be paid if funds are removed prematurely. (Some issues remain controversial. See the debate in the Fall, 1996 issue of the *Journal of Economic Perspectives*.)

An interesting flip side to IRA and 401(k) programs is that they also generated far less than the full participation one would have expected. Many eligible people do not participate, foregoing in effect a cash transfer from the government (and, in some cases, from their employer). O'Donoghue and Rabin (1999) present an explanation based on procrastination and hyperbolic discounting. Individuals typically show very sharp impatience for short-horizon decisions, but much more patience at long horizons. This is often referred to as hyperbolic discounting by contrast with the standard assumption of exponential discounting, in which patience is independent of horizon. In exponential models, people are equally patient at long and short horizons. O'Donoghue and Rabin argue that hyperbolic individuals will show exactly the low IRA participation that we observe. Although hyperbolic people will want eventually to participate in IRAs (because they are patient in the long run), something always comes up in the short run (where they are very impatient) that provides greater immediate reward. Consequently, they may delay joining the IRA indefinitely.

If people procrastinate about joining the savings plan, then it should be possible to increase participation rates simply by lowering the psychic costs of joining. One simple way of accomplishing this is to switch the default option for new workers. In most companies, when employees first become eligible for the 401(k) plan, they receive a form inviting them to join; to join they have to send the form back and make some choices. Several firms have made the seemingly inconsequential change of switching the default: employees are enrolled into the plan unless they explicitly opt out. This change has often produced dramatic increases in savings rates. For example, in one company studied by Madrian and Shea (2000), the employees who joined after switching the default to being in the plan were 50% more likely to participate than the workers in the year prior to the change. (They also found that the default asset allocation had a strong effect on workers' choices. The firm had made the default asset allocation 100% in a money market account, and the proportion of workers selecting this allocation soared.)

Along with these empirical facts, there has also been substantial theoretical work. To cite a few examples, Laibson (1997) and O'Donoghue and Rabin (1999) both examined the effects of hyperbolic discounting on the savings decisions. They highlighted that hyperbolic individuals will demand commitment devices (savings vehicles which are illiquid) and generally fail to obey the Euler equation.

Other Directions

We have concentrated on two fields here in order to give a sense of what behavioral economics can do, but we do not want to leave the impression that savings and financial markets are the only domains in which a behavioral approach has been or could be effective. In labor economics, experimental and empirical work has underlined the importance of fairness considerations in setting wages. For example, wages between industries differ dramatically, even for identical workers and, most interestingly, even homogeneous workers (such as janitors) earn higher wages when they work in industries where other occupations earn more. In law and economics, we have seen the importance of 'irrelevant' factors in a jury's decision to sentence or in the magnitude of awards they give. In corporate finance, one can fruitfully interpret a manager's decision to acquire or diversify as resulting from overconfidence. One could go on. There is much to be done.

Editor's Note

Thirty years ago, behavioral economics did not exist as a systematic field of inquiry; today, it is one of the most active and exciting areas of research within economics. Readers seeking useful entry points into this large and growing literature will be well served by any of the following: Camerer et al., 2004; Kamenica, 2012; Croson, 2014. Also useful: Smelser and

Swedberg, 2005. A great deal of research in this field is published in Elsevier's Journal of Behavioral and Experimental Economics.

See also: Bounded Rationality; Economic Sociology; Stock Market Predictability; Tversky, Amos (1937–96).

Bibliography

Banks, J., Blundell, R., Tanner, S., 1998. Is there a retirement-savings puzzle? American Economic Review 88, 769–788.

Camerer, C., et al., 1997. Labor supply of New York City cabdrivers: One day at a time. Quarterly Journal of Economics 112, 407–441.

Camerer, C.F., Loewenstein, George, Rabin, Matthew, 2004 (new edition 2011). Advances in Behavioral Economics. Princeton University Press, Princeton, NJ.

Conlisk, J., 1996. Why bounded rationality? Journal of Economic Literature 34, 669–700.

Croson, R.T.A., 2014. Behavioral Economics for Environmental Decisions. Available at Review of Environment, Energy and Economics. http://ideas.repec.org/a/fem/femre3/2014.02-01.html.

De Bondt, F.M., Werner, Thaler, R., 1985. Does the stock market overreact? Journal of Finance 40, 793–805.

DeLong, B., Shleifer, A., Summers, L., Waldman, R., 1990. Noise trader risk in financial markets. Journal of Political Economy 98, 703–738.

Froot, K.A., Dabora, E.M., 1999. How are stock prices affected by the location of trade? Journal of Financial Economics 53, 189–216.

Haltiwanger, J., Waldman, M., 1985. Rational expectations and the limits of rationality: An analysis of heterogeneity. American Economic Review 75, 326–340.

Kahneman, D., Tversky, A., 1974. Judgement under uncertainty: Heuristics and biases. Science 185, 1124–1131.

Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. Econometrica 47, 263–291.

Kahneman, D., Slovic, P., Tversky, A., 1982. Judgement Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, UK.

Kamenica, E., 2012. Behavioral economics and psychology of incentives. Annual Review of Economics 4, 427–452.

Laibson, D., 1997. Golden eggs and hyperbolic discounting. Quarterly Journal of Economics 112, 443–477.

Lee, C.M.C., Shleifer, A., Thaler, R.H., 1991. Investor sentiment and the closed-end fund puzzle. Journal of Finance 46, 75–109.

Madrian, B., Shea, D., 2000. The Power of Suggestion: Inertia in 401(k) Savings Behavior. Mimeo, University of Chicago, Chicago.

Odean, T., 1998. Are investors reluctant to realize their losses? Journal of Finance 53, 1775–1798.

O'Donoghue, E., Rabin, M., 1999. Procrastination in preparing for retirement. In: Aaron, H. (Ed.), Behavioral Dimensions of Retirement Economics. Brookings Institution, Washington, DC, pp. 125–156.

Russell, T., Thaler, R., 1985. The relevance of quasi rationality in competitive markets. American Economic Review 75, 1071–1082.

Shefrin, H., Statman, M., 1985. The disposition to sell winners too early and ride losers too long: Theory and evidence. Journal of Finance 40, 777–790.

Shleifer, A., Vishny, R., 1997. The limits of arbitrage. Journal of Finance 52, 35–55.

Shleifer, A., 2000. Inefficient Markets: An Introduction to Behavioral Finance. Clarendon Lectures. Oxford University Press, Oxford, UK.

Simon, H.A., 1955. A behavioral model of rational choice. Quarterly Journal of Economics 69, 99–118.

Smelser, N.J., Swedberg, R., (Eds.), 2005. The Handbook of Economic Sociology. Princeton University Press, Princeton, NJ.

Starmer, C., 2000. Developments in non-expected utility: The hunt for a descriptive theory of choice under risk. Journal of Economic Literature 38, 332–382.

Thaler, R., 1985. Mental accounting and consumer choice. Marketing Science 4, 199–214.

Behavioral Economics, History of

Michiru Nagatsu, University of Helsinki, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the historical development of behavioral economics, with an emphasis on how it has become part of mainstream economics. First I describe the separation between economics and psychology as a historical background against which behavioral economics emerged. Then I introduce critiques of expected utility theory (EUT), and explain how behavioral decision research developed using EUT as an experimental paradigm. I then distinguish 'old,' 'new,' and 'second-wave' behavioral economics, and briefly discuss criticisms of the last.

Introduction

Behavioral economics, an increasingly influential research program in economics, tries to improve economic theory and policy by drawing mainly on psychological or behavioral insights on how real people, as opposed to the idealistically rational agent, think and behave (Mullainathan and Thaler, 2001; Camerer and Loewenstein, 2004). Behavioral economics thus defined has now become part of mainstream economics, in terms of the numbers of behavioral economists hired in research institutions, journal articles, and textbooks published, and prestigious prizes awarded to behavioral economists. Its practical relevance is increasing as both public and private sectors use more findings from behavioral economics to efficiently and effectively change people's behavior. Finally, several popular books (e.g., Ariely, 2008; Poundstone, 2010; Kahneman, 2013) made behavioral economics well known among the educated general public. Accordingly, an interest in its history has arisen as well. Sent (2004) is one of the earliest studies on this subject, and Angner and Loewenstein (2012) is an informative review. Heukelom (2014) gives a book-length treatment (see also Heukelom, 2011, 2012). One of the questions these authors explicitly and implicitly ask is why behavioral economics has been so successful. There are several contributing factors, and the reader is referred to the above mentioned literature to get a fuller account. In this article, I will highlight one important reason for the success of behavioral economics, namely that it adopted mainstream rational economic modeling tools such as utility maximization and equilibrium analysis of games, instead of denying them and proposing an altogether different approach.

The article proceeds as follows: Section The Historical Context briefly introduces the historical background against I proceed as follows: after briefly introducing The Historical Context in which behavioral economics emerged, I describe Early Criticisms of EUT and describe 'Old' Behavioral Economics. Then I identify Behavioral Decision Research (BDR) as a basis of 'New' Behavioral Economics, and describe two important results in BDR, which gave rise to 'New' Behavioral Economics. Then I illustrate The 'Second-Wave' Behavioral Economics and its critics, followed by Conclusion.

The Historical Context

Economics as a Separate Science

If behavioral economics is characterized by the use of rich psychological insights, then its history has to go back at least to David Hume (1711–76) and Adam Smith (1723–90). In fact, Ashraf et al. (2005) argue that Smith's *Theory of Moral Sentiments* (Smith, 2002, first published in 1759) proposes what is now called the dual-process cognitive model; anticipates modern findings such as loss aversion, willpower, and fairness; and even suggests new directions for research in behavioral economics. Even though these claims might well be true, it is important to remember that contemporary behavioral economics has emerged against a backdrop of the sharp separation of economics from psychology that had been gradually established over 200 years since Smith. This separation started from John Stuart Mill's (1806–73) deductive and a priori methodology, and was achieved by successive neoclassical theoretical innovations such as marginal utility theory, indifference curve analysis, and revealed preference theory. As a result, economists became able to model an agent's choice behavior in terms of her preference satisfaction without referring to the utility concept connected with psychological valence. Instead, the concept of utility was reinterpreted simply as a mathematically useful way of representing preference orderings as defined by preference theory.

Subjective Expected Utility Theory

Preference theory made it possible to explain choice behavior in terms of preference orderings. In turn, the utility concept in economics became 'ordinal,' i.e., there was no need to talk about subjective and unobservable intensities of psychological utilities for different goods. This situation was somewhat complicated by the invention of expected utility theory (EUT). In general, desirability of a course of action depends not only on the value that action intends to deliver, but also on (one's belief about) the way the world is; for instance, the desirability of an action of bringing or not bringing an umbrella depends on what (one thinks) the weather is going to be like. In order to model decision-making under such uncertainty, the British philosopher Frank Ramsey (1903–30) sketched, in 1926, a way of quantifying and operationalizing degrees of beliefs and

desires solely from the observation of the agent's simple choices over lotteries, adopting an idealized folk psychological scheme in which "we act in the way we think most likely to realize the objects of our desires, so that a person's actions are completely determined by his desires and opinions"(Ramsey, 1931: 173). The mathematician John von Neumann (1903–57), while working on the theory of strategic interactions and its application to economics, axiomatized this system as EUT (von Neumann and Morgenstern, 2004, originally published in 1944). (The full axiomatization is given in the second edition published in 1947.) von Neumann and Morgenstern (2004) proposed the objective interpretation of probability (i.e., as relative frequencies of events in the long run), but anticipated subjective EUT, admitting that the subjective view of probability also leads to a satisfactory numerical concept of utility (3.3.3: fn. 2). The mathematician Leonard Savage (1917–71), a war-time assistant of von Neumann at the Institute for Advanced Study, Princeton, completed the subjective version of EUT, the foundation of modern decision theory (Savage, 1954). (The idea of expected utility itself was already proposed by the Dutch-Swiss mathematician Daniel Bernoulli (1700–82) in 1738. Another important founder of subjective EUT is the Italian statistician Bruno de Finetti (1906–85), whose work is acknowledged in Savage (1954).) As a result, it became possible to talk about the relative intensity of preferences in an operationally meaningful and principled way, bringing the 'cardinal' utility back in economic models of decision-making in a legitimate fashion. At the same time, Savage's theory made it possible to measure subjective beliefs *qua* probabilities in an equally rigorous manner. As we will see below, the birth of EUT proved to be crucial for the emergence of behavioral economics. (In what follows, I simply use the term EUT, but this should be read as a subjective version of the theory unless otherwise stated.)

Early Criticisms of EUT

Savage's EUT is not behavioristic as preference theory presumably aspired to be, because the former explicitly refers to psychological constructs such as subjective beliefs and expected utility. Nevertheless, economists welcomed it because the axiomatic approach was thought to guarantee the operationally meaningful and nonarbitrary measurement of these quantities. EUT thus soon found its application in economics. As early as 1948, Milton Friedman (1912–2006) collaborated with Savage, putting forward EUT as an economic hypothesis to explain the coexistence of insurance and lotteries (Friedman and Savage, 1948). But no sooner than EUT was completed as an axiomatic system, several economists expressed serious doubts on its normative acceptability and empirical plausibility (Markowitz, 1952; Allais, 1953; Ellsberg, 1961). These criticisms, however, were based on introspection and casual observations, rather than the rigorous experimental studies. This is only natural because back in the days experimental methods were not part of economists' toolbox yet. Instead, early experimental tests of EUT, with small but real monetary incentives, were conducted by the Harvard statistician Frederick Mosteller (1916–2006) and his student Philip Nogee (1951), and by the psychologist Sidney Siegel (1916–61), philosophers Donald Davidson (1917–2003), and Patrick Suppes (1922 to present) at Stanford (Stanford Value Theory Project) (1955). Another student of Mosteller, Ward Edwards (1927–2005), played a particularly important role in the rise of behavioral economics, as we will see in the next section. Before discussing that, I will turn to the so-called 'old' behavioral economics below.

'Old' Behavioral Economics

Herbert Simon (1916–2001), one of the earliest critics of models of rational choice in economics, was more radical than his contemporary critics of EUT. While these critics formulated problems as implausibility of specific axioms of EUT, such as the independent axiom, Simon offered a more programmatic critique, as part of the broader contribution he made to the so-called cognitive revolution: a paradigm shift in the postwar psychology from behaviorism to the computational modeling of internal cognitive processes. Simon argued that utility maximization models such as EUT, interpreted as representational models of cognitive processes of economic agents, are empirically unsound, drawing on the distinction between *substantive* and *procedural* rationality. The former concerns an intelligent system's adjustment or adaptation to its outer environment (e.g., an economic agent's reaction to the market), whereas the latter concerns the system's ability determined by its inner knowledge and computation that constrains the former (Simon, 1996: 25). A key idea is that we need only substantive rationality to predict behavior of the system *if* both its goal and environment are simple; but since these factors and their interaction are in fact complex, Simon argued, we have to explicitly consider the procedural rationality of the system. From this *bounded rationality* perspective, economic models of choice are deficient because they assume substantive rationality (individuals' utility maximization and organizations' profit maximization in the market environment) when in fact these conditions are not satisfied, requiring the explicit modeling of the inner workings of the agent, be it a human being or a firm.

Simon's impact on mainstream economics, however, was limited. Although he received the Nobel Memorial Prize in Economics in 1978 "for his pioneering research into the decision-making process within economic organizations," his proposal that *individual* procedural rationality should be explicitly modeled was not accepted by mainstream economics. Neither were his contemporary economists' alternative methods and theories that drew on psychology and other social sciences. Although already in the 1960s the term 'behavioral economics' was used to refer to these economists' rather heterogeneous research programs (Sent, 2004), their efforts did not have much impact on mainstream economics, despite the launch of *The Journal of Behavioral Economics* in 1972 (continued as *The Journal of Socio-Economics* since 1991) and such publications as Katona and Morgan (1980), Gilad and Kaish (1986) (but see Hosseini, 2003). For this reason, Sent (2004) calls these lines of research 'old' behavioral economics. (Given contemporary research such as Altman (2006), perhaps it is more

appropriate to call it heterodox rather than old behavioral economics.)

In retrospect, one reason why 'old' behavioral economics never caught on in the mainstream is that it started from the dismissal of the main methodological tenets of mainstream economics, such as positivism, deductivism, static equilibrium analysis, and optimizing models of economic agency (see Hosseini, 2003: 394). In contrast, as we will see below, 'new' behavioral economics has its deep root in BDR, which exploited maximization models such as EUT as experimental paradigms.

Behavioral Decision Research

Ward Edwards was among the first psychologists to notice the relevance of EUT for psychology. He introduced it to psychologists as *behavioral decision theory* (Edwards, 1954, 1961; Edwards et al., 1963), a rational model of individual probability judgment (subjective Bayesianism) and choice under risk. Edwards, the son of an economist, studied psychology at Harvard under the influence of Mosteller, who introduced Edwards to EUT (Phillips and von Winterfeldt, 2007). In 1958, Edwards moved to the psychology department at the University of Michigan, Ann Arbor, a then-rapidly expanding research center in mathematical psychology and decision-making (see Heukelom, 2010). But it was the following younger generation, including his former students at Michigan, who developed a subfield in cognitive psychology called BDR, by demonstrating systematic violations of the predictions of EUT in a series of experiments, and by proposing alternative models of individual judgment and decision-making. Mainstream economic rational choice models such as EUT were crucial for the development of BDR, since these models provided 'clear and crisp' predictions to be experimentally explored (Sent, 2004; Angner and Loewenstein, 2012); the deviations from the predictions could then be exploited to develop alternative models. In particular, two developments in BDR are important for the rise of 'new' behavioral economics.

Preference Reversals and the Construction of Preferences

One of the first systematic violations of EUT was demonstrated by Sarah Lichtenstein (1933 to present) and Paul Slovic (1938 to present), Edwards's former students, then working at the Oregon Research Institute. They conducted a series of experiments on people's risk assessments of hypothetical gambling, and discovered a serious anomaly for EUT called *preference reversal* (Lichtenstein and Slovic, 1971; see also Lindman, 1971), reported essentially the same result). Preference reversal is a phenomenon in which the decision-maker's preference over a pair of lotteries seems to be flipped depending on the method of preference elicitation. For example, when asked to choose one bet from two, many subjects in Lichtenstein and Slovic's experiments chose a low-payoff, high-probability bet (the so-called P bet, e.g., 35/36 chance of winning $4) over a high-payoff, low-probability bet ($ bet, e.g., 11/36 chance of winning $16); but when asked to state how much they were willing to pay, the subjects priced the $ bet higher than the P bet. The subjects seemed to have violated the axiom of transitivity (as their responses suggested they prefer A to B, and B to A at the same time), but the problem was deeper than that since it challenged the *procedural invariance* principle, a fundamental assumption of rational choice that preferences over an identical set of alternatives should not change depending on how you measure them (e.g., letting subjects rate, choose between, or state reserve prices for options). This result was replicated in a field experiment using real gamblers incentivized with high-stake money in the Four Sisters Casino in Las Vegas (Lichtenstein and Slovic, 1973).

Preference reversals caught economists' attention: two Caltech experimental economists challenged this result, with an explicit attempt to make the phenomenon go away, only to find it persisted (Grether and Plott, 1979). This paper was published in *The American Economic Review*, stimulating economists to produce alternative models to EUT that abandon or modify axiom(s) of EUT. In contrast, Slovic and other behavioral decision researchers proposed procedural models in which preferences are constructed in the process of decision-making (see articles collected in Section III of Lichtenstein and Slovic, 2006).

Framing Effects and Prospect Theory

Another series of serious challenges to EUT came from a team of Israeli-American psychologists Amos Tversky (1937–96) and Daniel Kahneman (1934 to present). Tversky completed his PhD at Michigan in 1965 under the supervision of Clyde Coombs (1912–88) and Edwards. Tversky's early research was on both the representational theory of measurement in mathematical psychology and empirical investigation into subjective EUT (Heukelom, 2009: 48). In the late 1960s, Tversky went back to Israel to teach at Hebrew University, where Daniel Kahneman had been establishing himself as a leading researcher in the field of mental effort. Tversky's faith in subjective Bayesianism was 'severely shaken' by conversations with Kahneman (Kahneman, 2003). The first series of their collaborative work thus studied individual probability judgment under uncertainty; how individuals use mental shortcuts called 'heuristics' such as representativeness, availability, and anchoring and adjustment; and how these heuristics lead to 'biases,' i.e., deviations from the subjective Bayesian model, part of subjective EUT. This research program, known as 'heuristics and biases,' was launched with their first collaborative paper (Tversky and Kahneman, 1971), and further developed in their *Science* paper (Tversky and Kahneman, 1974) and other papers collected in Kahneman et al. (1982) and Gilovich et al. (2002).

Another, perhaps even more important program by Tversky and Kahneman comprises a series of experiments that demonstrated *framing effects* (Tversky and Kahneman, 1981, 1986), and *Prospect Theory* (Kahneman and Tversky, 1979) as an alternative to EUT (other important papers are collected in Kahneman and Tversky, 2000). Broadly speaking, framing effects refer to another type of preference reversals, in which an individual's preference over an identical set of alternatives seems to be flipped depending on how they are described, in particular as a gain or loss. Framing effects are as serious an anomaly to EUT as Lichtenstein and Slovic's preference reversals, because they challenged another invariance principle

in rational choice called the *descriptive invariance*, which states that the agent's preferences over an extensionally identical set of alternatives should not be affected by how they are intensionally described. An innovation of Prospect Theory consisted in its introduction of the concept of *reference points* to explain framing effects and other related anomalies *within the framework of utility theory*. One of the central ideas of Prospect Theory is that an individual evaluates gains and losses relative to a certain reference point, not absolute quantities, an application of psychophysics to the perception of money and other goods. Different frames with different reference points shift the perception of alternatives as a gain or loss, and affect individuals' choice. With the assumptions that people are risk-averse toward unlikely loss and likely gain, and risk-seeking toward likely loss and unlikely gain, Prospect Theory explained not only framing effects but also other early problems such as Allais's paradox, and people's simultaneous purchasing of lotteries and insurances that had been discussed earlier by Friedman and Savage (1952) and Markowitz (1952). (For a critical assessment of Prospect Theory and related models of risky choice, see Friedman et al. (2014).)

Tversky and Kahneman's 1979 *Econometrica* paper (Kahneman (2003) recalls: "The choice of venue turned out to be important; the identical paper, published in *Psychological Review*, would likely have had little impact on economics. But our decision was not guided by a wish to influence economics. *Econometrica* just happened to be the journal where the best papers on decision-making to date had been published, and we were aspiring to be in that company.") and their 1981 *Science* paper were hailed as having "altered the intellectual history of economics; they brought the behavioral economics research program into the mainstream" (Laibson and Zeckhauser, cited in Angner and Loewenstein, 2012: 662). This reputation of Prospect Theory is due not only to its empirical success, but also to its unique hybrid nature: on the one hand, it is a psychologically informed cognitive process model in which alternatives are mentally 'edited' as gains and losses before choice; on the other, it is presented as a utility maximization model with a peculiar utility curve, representing different subjective valuations of likely and unlikely outcomes in gain and loss domains. Unlike 'old' behavioral economics and other BDR, Prospect Theory demonstrated an appealing, paradigmatic way of incorporating insights from cognitive psychology into economic models of choice.

'New' Behavioral Economics

The success of Prospect Theory encouraged a number of young economists to challenge the standard models of rational choice. In *Toward a Positive Theory of Consumer Choice* (1980), for example, Richard Thaler (1945 to present) discussed, drawing on Tversky and Kahneman's work, various aspects of consumer behavior (e.g., underweighting of opportunity costs, failure to ignore sunk costs, regret, and self-control) that would become central themes in behavioral economics. Thaler spent the 1984–85 academic year visiting the University of British Columbia to work with Kahneman, who recalls that "That was the year that behavioral economics began" (Poundstone, 2010: chapter 17). Thaler also contributed to the dissemination of empirical findings challenging economic models of choice, and helped mainstream economists to see the relevance of these results, through publishing, in collaboration with psychologists and economists, 'Anomalies' columns in the *Journal of Economic Perspectives*, between 1987 and 1991 (later collected in Thaler, 1992), and 1995–2001 (see Heukelom, 2012: fn. 16).

Two private foundations, the Sloan Foundation and Russell Sage Foundation, also played a key role in developing behavioral economics as a thriving subfield of economics, through financially supporting the behavioral economics program (1984–92), including projects such as nonresidential working groups, the visiting researchers program, and the Russell Sage Foundation Behavioral Economics Books Series (Heukelom, 2012). Eric Wanner (1942 to present), a former Harvard cognitive psychologist, who had a clear goal to apply cognitive science to economics, played a pivotal role as a director of the behavioral economics program at both foundations. Wanner involved Tversky, Kahneman, and Thaler from early on at the preparatory stage of the program. Although Wanner initially understood the program as a development of Simon's behavioral economics that was highly critical about the mainstream neoclassical economics, he quickly adopted several strategies to bring new behavioral economics to the mainstream. For example, he emphasized applications of behavioral insights to the study of economically important domains such as financial markets. Wanner also solicited prominent and promising economists for applications, including Kenneth Arrow (1921 to present), Vernon Smith (1927 to present), George Akerlof (1940 to present), and Robert Shiller (1946 to present) (Heukelom, 2012). (Smith, who shared the 2002 Nobel Prize with Kahneman, however, was critical about the direction of the behavioral economics program by the 1990s; his prize was awarded for a distinct research program called experimental economics.) Simon, one of Wanner's 40 invitees in 1985, criticized Wanner's program for taking "too seriously the premises of contemporary economic methodology that theories ('models') come first and empirical work afterward"; for Simon, the problem was that "mainline economists continue to ignore vast bodies of relevant evidence in their preferred pursuit of armchair model building" (Simon's 1986 letter to Wanner cited in Heukelom (2012)). Wanner and his advisors, however, seem to have been aware that a key to the success of behavioral economics was to accommodate psychological findings *without* giving up economists' way of model building.

The 'Second-Wave' Behavioral Economics

Matthew Rabin (2002: 658) observed the rise of 'second-wave behavioral economics,' which "moves beyond pointing out problems with current economic assumptions, and even beyond articulating alternatives, and on to the task of systematically and formally exploring the alternatives with much the same sensibility and mostly the same methods that economists are familiar with." As examples, he mentions the models of time preferences and social preferences. (In the

domain of risk preferences, Tversky and Kahneman's (1992) Cumulative Prospect Theory can be seen as a 'second-wave' upgrade of their original Prospect Theory.) In the following two subsections, I illustrate these models in turn, and in the last subsection I briefly discuss recent criticisms of these 'second-wave' models.

Modeling Dynamically Inconsistent Preferences

Since important economic behavior such as saving concerns consumption that are to be made over an extended period of time, economic theory needs to explicitly model how consumers value delayed as well as immediate consumptions. Paul Samuelson (1915–2009) proposed the discounted utility model (DUM), which had become the standard model of time discounting (Samuelson, 1937). One of the fundamental assumptions of DUM is that "[d]uring any specified period of time, the individual behaves so as to maximie the sum of all future utilities, they being reduced to comparable magnitudes by suitable time discounting" (p. 156). As a 'suitable time-discounting' function, Samuelson proposed, without any commitment to its empirical or normative plausibility, the following exponentially declining discount function:

$$d(t) = \delta^t \qquad [1]$$

where $0 < \delta < 1$. An exponential discount function is characterized by a *constant discount rate* for all future events. Thaler (1981), however, using a survey method, showed that the per-period discount factor δ declined over time. Later studies have confirmed this result (see Frederick et al., 2002, for a review). Intuitively put, this is because people value the present and future very differently. People obviously value present consumption (at time t) more than future consumption at $t+1$, but an exponential discount function suggests that the discount rate between t and $t+1$ is the same as that between $t+1$ and $t+2$. But since the former case is a comparison between the present event and a future one, while the latter between two future events, the former should show more pronounced discounting. To capture this feature of time discounting, Laibson (1997) adopted a 'quasi-hyperbolic' discount function, which had been used in a model of intergenerational altruism by Phelps and Pollak (1968):

$$d(t) = \beta\delta^t \qquad [2]$$

where $0 < \delta < 1$, and $0 < \beta < 1$. Hyperbolic discount functions imply discount rates that decline as the discounted event is moved further away in time, capturing familiar problems of impulsive behavior and procrastination. Note that hyperbolic discounting generalizes the standard assumption of exponential discounting, i.e., (2) reduces to (1) if $\beta = 1$, while retaining Samuelson's assumption about the existence of a time-additive utility function. Laibson (1997) adopted a discrete, rather than continuous, version of a hyperbolic discount function (hence the 'quasi' qualification), in order to "[mimic] the qualitative property of the hyperbolic discount function, while maintaining most of the analytical tractability of the exponential discount function" (p. 450). This is the kind of 'sensibility' and the methods that economists are familiar with.

Modeling Prosocial Preferences

Another growing domain of research in behavioral economics is the study of prosocial behavior. Although the default assumption in the theory of public goods since Samuelson (1954) had been selfish individuals, viz., that individuals care only about self-interest narrowly defined in terms of own material gains and losses, the field evidence such as successful voluntary provision of public goods and charitable giving prompted economists to propose alternative models of nonselfish preferences that better explain the observations (e.g., Becker, 1974; Sugden, 1984). Experimental economists and psychologists have also tested since 1980s so many different variations of prisoner's dilemma and public goods games that there is now substantial body of evidence against the selfishness assumption (see Ledyard, 1995; Chaudhuri, 2011).

However, the business of modeling prosocial preferences boomed with the inventions of other games from which one can more straightforwardly infer players' beliefs and preferences. The *ultimatum game* (Güth et al., 1982) is a game in which one player (the Proposer) makes a take-it-or-leave-it offer, as a division of a sum of money between herself and another player. If the second player (the Responder) accepts the division, then both players earn the specified amounts. If the Responder rejects it, they both get nothing. The standard prediction (rational play plus selfish preference) is that the Proposer will offer a minimum divisible sum of the money, which will be accepted by the Responder: for the Responder, any positive sum is better than zero (money maximization); since the Proposer knows this, she proposes the smallest amount to maximize her own money. Evidence from experiments shows that Proposers offer on average 30–40% of the money (modal and median offers are 40–50%); and Responders reject small offers below 20% about half the time. These results clearly falsify the standard prediction. Even in the game in which the Responder has no choice but to accept the offer (the *dictator game*, Kahneman et al. (1986)), the Dictator offers on average about 20% of the sum (see Camerer, 2003).

Models of *social preferences* try to accommodate the data across these experimental games, as well as prisoner's dilemma and public goods games. Rabin (1993), employing *psychological game theory*, proposed a model of *reciprocal fairness*, where one player matches another's 'intention,' which is inferred from the payoff structure of the game. For example, a responder in an ultimatum game may act 'mean' (reject) if she thinks the proposer to be 'mean' for offering, say 20%, because he could have offered her 40%, if not 50%. Fehr and Schmidt (1999) offered another model of *inequality aversion*, in which players have a preference for egalitarian distribution of the payoffs. For example, an inequality-averse player may give a substantial portion of the pie to the second player in a dictator game.

Recent Criticisms of the 'Second-Wave' Models

In recent years, these models of 'second-wave' behavioral economics attracted mainly two types of methodological criticism. The first concerns its single-minded modification of functional forms instead of explicitly modeling plausible

psychological *processes* (e.g., Rubinstein, 2003). In particular, social preference theorists' conviction that "[f]or modeling purposes, behaviorally relevant emotions can be captured by appropriate formulations of the utility function" (Camerer and Fehr, 2004: 80) have been challenged on several conceptual and methodological grounds (Heap and Varoufakis, 2004; Guala, 2006; List, 2007). Accordingly, models that draw more directly on cognitive and social psychology have been proposed (e.g., Bicchieri, 2006).

The second type of criticism concerns the specific ways in which behavioral economists interpret the data from experiments. In particular, Ross (2014) criticizes their strategy to identify central tendencies in a small sample and explain them in terms of common psychological dispositions of individuals, or utility functions that represent such tendencies (pp. 156, 233, 250). Ross instead advocates the approach to econometrically estimate the structure of *heterogeneity* in the data-generating processes in the large data set, in order to predict how such heterogeneity respond to policy interventions at the population level (see Andersen et al., 2008).

Conclusion

In this article, I briefly sketched the historical development of behavioral economics, which is nicely summarized as what Camerer and Loewenstein (2004: 7) call a 'recipe' of behavioral economics:

> First, identify normative assumptions or models that are ubiquitously used by economists, such as Bayesian updating, expected utility, and discounted utility. Second, identify anomalies – i.e., demonstrate clear violations of the assumption or model, and painstakingly rule out alternative explanations (such as subjects' confusion or transactions costs). Third, use the anomalies as inspiration to create alternative theories that generalize existing models. A fourth step is to construct economic models of behavior using the behavioral assumptions from the third step, derive fresh implications, and test them.

My narrative has along the way emphasized a key to the academic success of behavioral economics, namely its adoption of standard economic modeling techniques such as utility maximization and equilibrium analysis. But the prominence of behavioral economics might not be as stable as it appears now, for it faces challenges from *both* psychology and economics, as briefly discussed in Section Recent Criticisms of the 'Second-Wave' Models: one the one hand, the psychological critique suggests the limits of modeling *individual psychological processes* using existing economic methods; on the other, the economic (or econometric, to be more specific) critique points to the limits of modeling *aggregate-level data-generating processes* using a single representative functional form. It remains to be seen how behavioral economists will address these distinct challenges, and how their disciplinary identity will be affected as a consequence.

See also: Behavioral Economics; Behavioral Theories of Organization; Decision and Choice: Random Utility Models of Choice and Response Time; Expectations, Economics of; Experimental Economics; Microeconomics, History of; Neuroeconomics; Personality and Economics.

Bibliography

Allais, M., 1953. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'ecole americaine. Econometrica 21 (4), 503–546.
Altman, M., 2006. Handbook of Contemporary Behavioral Economics: Foundations and Developments. M.E. Sharpe, Armonk, NY.
Andersen, S., Harrison, G.W., Lau, M.I., Rutström, E.E., 2008. Eliciting risk and time preferences. Econometrica 76 (3), 583–618.
Angner, E., Loewenstein, G., 2012. Behavioral economics. In: Mäki, U. (Ed.), Philosophy of Economics, Handbook of the Philosophy of Science, vol. 13. Elsevier, pp. 641–689.
Ariely, D., 2008. Predictably Irrational: The Hidden Forces That Shape Our Decisions, first ed. Harper, New York.
Ashraf, N., Camerer, C.F., Loewenstein, G., 2005. Adam smith, behavioral economist. Journal of Economic Perspectives 19 (3), 131–145.
Becker, G.S., 1974. A theory of social interactions. The Journal of Political Economy 82 (6), 1063–1093.
Bicchieri, C., 2006. The Grammar of Society. Cambridge University Press, Cambridge, England.
Camerer, C.F., 2003. Behavioral Game Theory. Princeton University Press, Princeton, NJ.
Camerer, C.F., Fehr, E., 2004. Measuring social norms and preferences using experimental games: a guide for social scientists. In: Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. (Eds.), Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-scale Societies. Oxford University Press, Oxford, pp. 55–95 (Chapter 3).
Camerer, C.F., Loewenstein, G., 2004. Behavioural economics: past, present, future. In: Camerer, C.F., Loewenstein, G., Rabin, M. (Eds.), Advances in Behavioral Economics. Princeton University Press, Princeton, NJ, pp. 3–51 (Chapter 1).
Chaudhuri, A., 2011. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. Experimental Economics 14 (1), 47–83.
Davidson, D., Siegel, S., Suppes, P., 1955. Some Experiments and Related Theory on the Measurement of Utility and Subjective Probability. Technical Report Prepared for Office of Naval Research. Stanford University, Stanford.
Edwards, W., 1954. The theory of decision making. Psychological Bulletin 41, 380–417.
Edwards, W., 1961. Behavioral decision theory. Annual Review of Psychology 12, 473–498.
Edwards, W., Lindman, H., Savage, L.J., 1963. Bayesian statistical inference for psychological research. Psychological Review 70, 193–242.
Ellsberg, D., 1961. Risk, ambiguity, and the savage axioms. The Quarterly Journal of Economics 75 (4), 643–669.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. The Quarterly Journal of Economics 114 (3), 817–868.
Frederick, S., Loewenstein, G., O'Donoghue, T., 2002. Time discounting and time preference: a critical review. Journal of Economic Literature 40 (2), 351–401.
Friedman, D., Isaac, R.M., James, D., Sunder, S., 2014. Risky Curves: On the Empirical Failure of Expected Utility. Routledge, New York.
Friedman, M., Savage, L.J., 1948. The utility analysis of choices involving risk. Journal of Political Economy 56 (4), 279–304.
Friedman, M., Savage, L.J., 1952. The expected-utility hypothesis and the measurability of utility. Journal of Political Economy 60 (6), 463–474.
Gilad, B., Kaish, S., 1986. Handbook of Behavioral Economics. JAI Press, Greenwich, CT.
Gilovich, T., Griffin, D.W., Kahneman, D., 2002. Heuristics and Biases: The Psychology of Intuitive Judgement. Cambridge University Press, Cambridge, UK.
Grether, D., Plott, C.R., 1979. Economic theory of choice and the preference reversal phenomenon. American Economic Review 69, 623–638.
Guala, F., 2006. Has game theory been refuted? The Journal of Philosophy 103 (5), 239–263.
Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. Journal of Economic Behavior & Organization 3 (4), 367–388.
Heap, S.H., Varoufakis, Y., 2004. Game Theory: A Critical Text, second ed., rev. ed. Routledge, London.
Heukelom, F., 2009. Kahneman and Tversky and the Making of Behavioral Economics. Ph.D. thesis. University of Amsterdam.

Heukelom, F., 2010. Measurement and decision making at the University of Michigan in the 1950s and 1960s. Journal of the History of the Behavioral Sciences 46 (2), 189–207.
Heukelom, F., 2011. Building and defining behavioral economics. In: Biddle, J.E., Emmett, R.B. (Eds.), Research in the History of Economic Thought and Methodology, vol. 29. Emerald Group Publishing Limited, pp. 1–29.
Heukelom, F., 2012. A sense of mission: the Alfred P. Sloan and Russell Sage foundations' behavioral economics program, 1984–1992. Science in Context 25, 263–286.
Heukelom, F., 2014. Behavioral Economics: A History. Cambridge University Press, Cambridge.
Hosseini, H., 2003. The arrival of behavioral economics: from Michigan, or the Carnegie School in the 1950s and the early 1960s? Journal of Socio-Economics 32 (4), 391–409.
Kahneman, D., 2003. Biographical. Nobelprize.org.
Kahneman, D., 2013. Thinking, Fast and Slow, first pbk. ed. Farrar, Straus and Giroux, New York.
Kahneman, D., Knetsch, J.L., Thaler, R.H., 1986. Fairness and the assumptions of economics. The Journal of Business 59 (4), S285–S300.
Kahneman, D., Slovic, P., Tversky, A., 1982. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge.
Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47 (2), 263–292.
Kahneman, D., Tversky, A., 2000. Choices, Values, and Frames. Russell Sage Foundation, New York.
Katona, G., Morgan, J.N., 1980. Essays on Behavioral Economics. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI.
Ledyard, J., 1995. Public goods: a survey of experimental research. In: Kagel, J.H., Roth, A.E. (Eds.), Handbook of Experimental Economics. Princeton University Press, pp. 111–194.
Laibson, D., 1997. Golden eggs and hyperbolic discounting. The Quarterly Journal of Economics 112 (2), 443–478.
Lichtenstein, S., Slovic, P., 1971. Reversals of preferences between bids and choices in gambling decisions. Journal of Experimental Psychology 89, 46–55.
Lichtenstein, S., Slovic, P., 1973. Response-induced reversals of preference in gambling – extended replication in Las-Vegas. Journal of Experimental Psychology 101, 16–20.
Lichtenstein, S., Slovic, P., 2006. The Construction of Preference. Cambridge University Press, Cambridge.
List, J.A., 2007. On the interpretation of giving in dictator games. Journal of Political Economy 115 (3), 482–493.
Lindman, H.R., 1971. Inconsistent preferences among gambles. Journal of Experimental Psychology 89 (2), 390–397.
Markowitz, H., 1952. The utility of wealth. Journal of Political Economy 60 (2), 151–158.
Mosteller, F., Nogee, P., 1951. An experimental measurement of utility. Journal of Political Economy 59 (5), 371–404.
Mullainathan, S., Thaler, R.H., 2001. Behavioral economics. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Elsevier, Amsterdam, pp. 1094–1100.
Phelps, E.S., Pollak, R.A., 1968. On second-best national saving and game-equilibrium growth. The Review of Economic Studies 35 (2), 185–199.
Phillips, L.D., von Winterfeldt, D., 2007. Reflections on the contributions of Ward Edwards to decision analysis and behavioral research. In: Edwards, W. (Ed.), Advances in Decision Analysis: From Foundations to Applications. Cambridge University Press, pp. 71–80 (Chapter 5).
Poundstone, W., 2010. Priceless: The Myth of Fair Value (and How to Take Advantage of It), first ed. Hill and Wang, New York.
Rabin, M., 1993. Incorporating fairness into game theory and economics. The American Economic Review 83 (5), 1281–1302.
Rabin, M., 2002. A perspective on psychology and economics. European Economic Review 46 (4–5), 657–685.
Ramsey, F.P., 1931. Truth and probability. In: Braithwaite, R.B. (Ed.), The Foundations of Mathematics and Other Logical Essays. K. Paul, Trench, Trubner & Co., Ltd, pp. 156–198 (Chapter 7).
Ross, D., 2014. Philosophy of Economics. Palgrave Macmillan, New York.
Rubinstein, A., 2003. "Economics and psychology"? The case of hyperbolic discounting. International Economic Review 44 (4), 1207–1216.
Samuelson, P.A., 1937. A note on measurement of utility. The Review of Economic Studies 4 (2), 155–161.
Samuelson, P.A., 1954. The pure theory of public expenditure. The Review of Economics and Statistics 36 (4), 387–389.
Savage, L.J., 1954. The Foundations of Statistics. Wiley, New York.
Sent, E.-M., 2004. Behavioral economics: how psychology made its (limited) way back into economics. History of Political Economy 36 (4), 735–760.
Simon, H.A., 1996. The Sciences of the Artificial, third ed. MIT Press, Cambridge, MA.
Smith, A., 2002. The Theory of Moral Sentiments. Cambridge University Press, Cambridge, UK.
Sugden, R., 1984. Reciprocity: the supply of public goods through voluntary contributions. The Economic Journal 94 (376), 772–787.
Thaler, R., 1980. Toward a positive theory of consumer choice. Journal of Economic Behavior & Organization 1 (1), 39–60.
Thaler, R., 1981. Some empirical evidence on dynamic inconsistency. Economics Letters 8 (3), 201–207.
Thaler, R.H., 1992. The Winner's Curse: Paradoxes and Anomalies of Economic Life. Princeton University Press, Princeton, NJ.
Tversky, A., Kahneman, D., 1971. Belief in the law of small numbers. Psychological Bulletin 76 (2), 105–110.
Tversky, A., Kahneman, D., 1974. Judgement under uncertainty: heuristics and biases. Science 185, 1124–1131.
Tversky, A., Kahneman, D., 1981. The framing of decision and the psychology of choice. Science 211, 453–458.
Tversky, A., Kahneman, D., 1986. Rational choice and the framing of decisions. The Journal of Business 59 (4), S251–S278.
Tversky, A., Kahneman, D., 1992. Advances in prospect theory: cumulative representation of utility. Journal of Risk and Uncertainty 5, 297–323.
von Neumann, J., Morgenstern, O., 2004. Theory of Games and Economic Behavior, sixtieth anniversary ed. Princeton University Press, Princeton, NJ.

Behavioral Genetics: Psychological Perspectives

Valerie S Knopik, Rhode Island Hospital, Providence, RI, USA; Brown University, Providence, RI, USA; and Bradley Hasbro Children's Research Center, Providence, RI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Plomin, volume 2, pp. 1100–1105, © 2001, Elsevier Ltd.

Abstract

Behavioral genetics is concerned with the study of genetic influences on behavior; however, it also tells us about the importance of the environment. Through multiple levels of analyses ranging from quantitative genetic to molecular genetic studies of DNA to epigenetics and gene expression, this field examines the roles of genes and environment on individual differences in an outcome of interest. This report, focused on human behavior, outlines the foundations and future of this growing field.

Behavioral genetics is concerned with the study of genetic influences on behavior. Despite this definition, behavioral genetics also tells us about the importance of the environment. Behavioral genetics encompasses a wide range of levels of analyses, ranging from quantitative genetics to molecular genetic studies of DNA as it relates to behavior. More recently, efforts also include investigation of epigenetics as well as gene expression, described later in this article. Despite this wide range of methodological approaches, the core focus for most behavioral genetic research remains the same: to examine the roles of genes and environment on individual differences in an outcome of interest. For example, why do people differ with regard to personality, addiction, psychopathology, or cognitive ability? The words 'nature' and 'nurture' are often used in the field of behavioral genetics. The history of the nature versus nurture debate is beyond the scope of this article; however, the words nature and nurture will be used throughout as broad categories representing genetic and environmental influences, respectively.

Gregor Mendel (1822–84) is considered the father of heredity. Mendel studied inheritance in pea plants in the garden of his monastery in what is now known as the Czech Republic. On the basis of his many experiments, Mendel concluded that there are two 'elements' of heredity for each trait in each individual, and that these two elements separate, or segregate, during reproduction (Plomin et al., 2013). These elements are now known as genes. In the study of heredity, the focus is on genes that have different forms (called alleles) across individuals, animals, plants, or other model organisms. Each individual's combination of genes is called their genotype. The observed behavior or trait is called a phenotype. The fundamental issue in behavioral genetics is the extent to which differences in genotype account for differences in phenotype (or the observed differences among individuals).

Quantitative genetics – used, for example, in twin and adoption studies (described later) – rely on Mendel's laws of heredity but do not require knowledge of the biological basis of heredity. Molecular genetics, in contrast, looks at specific variation in individual genes or across the genome and, as described later, the epigenome and transcriptome. The human genome consists of about three billion base pairs, and 99.9% of this DNA sequence is identical for all human individuals. This leaves about three million base pairs that vary among individuals. There is no single human genome. In fact, we each have a different genome, with the exception of identical twins. It is the variation in the genome that can tell us, in part, about the origin, or etiology, of diseases and disorders. These variations are the focus of behavioral genetics.

More specifically, the DNA sequence defines our genes. Genes code for sequences of amino acids that form the thousands of proteins of which organisms are made. Proteins create the skeletal system, muscles, the endocrine system, the immune system, the digestive system, and most important for behavior, the nervous system. Genes do not code for behavior directly, but DNA variations that create differences in the abundance or function of the proteins that comprise these physiological systems can affect behavior (Plomin et al., 2013). It is important to note that while it is the variation in the genome that is of interest to the field of human behavioral genetics, many of the conserved genetic regions that are most important from an evolutionary perspective are nonvarying. In humans, it is not possible to study the effects of nonvarying genes; however, in animal models, differences in such genes can be created by 'knocking out 'the genes 'functioning (Plomin, 2001; Plomin and Crabbe, 2000). There is an ever-growing literature capitalizing on the power that animal models bring to behavioral genetic questions. This article, however, is limited to research on human behavior. Behavioral genetics uses methods of both quantitative and molecular genetics to study behavior.

Quantitative Genetics

Quantitative genetics offers a powerful theory as well as various methods for investigating the genetic and environmental etiology of any measurable phenotype or outcome. Most behavioral, psychological, or psychiatric traits are much more complex than single-gene disorders such as Huntington disease. Such complex phenotypes are influenced by multiple genes and likely multiple environmental factors. Because behavioral genetics is concerned with the study of individual differences, or detecting the factors that make individuals in a population different from one another, the first step is

typically to try to gauge the relative importance of genetic and environmental factors that cause individual differences. Quantitative genetics does this without specifying what the specific genes or environmental factors are. (It is now possible to identify what those specific genes are by using molecular genetic methods, which are briefly outlined later in this article.) Rather than using genetically defined populations, such as inbred mouse strains (strains that have been developed via brother–sister matings to result in nearly genetically identical individuals), or manipulating environments experimentally, human quantitative genetics focuses on naturally occurring genetic and environmental variation.

Although an outline of the model that forms the basis of quantitative genetics work is beyond the scope of this article, it is important to understand that by incorporating the basic laws of heredity and the expected genetic relationships among different kinds of relatives, quantitative genetics provides methods that focus on family data or variations of the family model. Most behavioral genetic analysis contrasts phenotypic similarity between related individuals with their genetic similarity. If individuals who are more closely related genetically also tend to be more similar on a measured trait, there is a suggestion for that trait being heritable, or influenced by genetic factors. Twin and adoption models are widely used for their ability to examine genetic and environmental influences on complex traits, like behavioral and psychological outcomes.

In genetic research, the word 'environment' refers to everything that is nongenetic. This is a much broader definition than is typically encountered in social and behavioral science. More specifically, environment denotes all nonheritable factors, including possible biological events such as early childhood illness, not just psychosocial factors that are usually considered in the behavioral and social sciences (Plomin, 2001). The environment impacts the development of behavioral and psychiatric disorders and the environment influences genetics as well (Li, 2010). There is a clear need to carefully measure the environment and to be mindful that many environmental measures show genetic influence (Kendler and Baker, 2007). Thus, genetically sensitive designs that incorporate and investigate interactions and correlations between genes and environments are important to the field moving forward (Knopik, 2011).

The Adoption Design

Family resemblance can be due to either genes or environment, or some combination of both. Adoption designs provide a way to disentangle these genetic and environmental sources of familial resemblance. Adoption creates genetically related individuals who do not share a common family environment because they were adopted apart, such as a birth mother and her biological child who was placed into an adoptive family at birth. Adoption also creates adopted-together family members who share a common family environment but are not genetically related, such as the adoptive mother and the adopted child. Adoption studies often yield evidence for genetic influence on behavioral traits, although results depend on the trait examined and the age of the adopted child (Plomin et al., 2013).

The first adoption study of schizophrenia, reported by Heston in 1966, was highly influential. At that time, it was hypothesized that early life and family experiences, such as poor parenting, caused schizophrenia. The Heston adoption study (as well as many other subsequent adoption studies) showed that the risk for developing schizophrenia was similar in adopted offspring of schizophrenic birth mothers and in children reared by their schizophrenic birth parents. These findings suggested not only that the rearing environment had little effect, but also that genetics makes a major contribution to this disorder.

Adoption designs do, however, have certain assumptions and limitations. First, there is question about the representativeness of adoption samples. That is, if biological parents, adoptive parents, or adopted children are not representative of the population, then this could affect the generalizability of results from adoption designs. Second, there is question of the prenatal environment. Resemblance between the birth mother and the adopted child might reflect prenatal environmental influences rather than, or in addition to, genetic influences. Third, due to an increase in 'open' adoptions, where birth parents and the adoptive parents know each other or share information, there is a concern about the extent to which the adopted child's rearing environment is truly independent of genetic influences from the birth parents. Finally, there is also the possibility that adopted children might be selectively placed with adoptive parents matched to the birth parents. Thus, there might be an issue with correlated environments. Many of these issues can be examined empirically (Plomin et al., 2013) and have been found to differentially influence different outcomes.

The Twin Design

The majority of quantitative genetic studies have focused on twins. Genetically, fraternal (dizygotic, or DZ) pairs, who develop from separately fertilized eggs, are no more alike than siblings who share, on average, 50% of their genetic material. Identical (monozygotic, or MZ) pairs are genetically identical because they came from the same fertilized egg. If genetic factors are important for a trait, these genetically identical pairs of individuals should be more similar than DZ twins who are only 50% similar genetically. If environmental factors are important for a trait, DZ twins should be as similar for the behavior as are MZ twins. Following the example of schizophrenia, twin studies show that genetics are an important contribution to familial resemblance for schizophrenia (Sullivan et al., 2003; Cardno et al., 2012).

The twin method has a different set of assumptions and potential limitations. First, it is possible that greater similarity in MZ twins is caused by environment rather than genetics because the twins are the same age, the same sex, and look alike. The *equal environments assumption* assumes that environmentally caused similarity is roughly the same for both MZ and DZ twins reared together. This assumption has been tested and appears reasonable for most traits (e.g., Bouchard and Propping, 1993; Derks et al., 2006). Second, as with the adoption design, generalizability is an issue for the twin method. Although twins are often born premature and may experience delays in language development, among other

things, the statistical distributions for most behaviors for twins and nontwins are generally similar, which suggests that twins are reasonably representative of nontwin populations (e.g., Robbers et al., 2011).

Estimating Genetic Influences on Behavior

Complex phenotypes, such as behavioral and psychological outcomes, are influenced by multiple genes and likely multiple environmental factors. The first step in quantitative genetics is typically to try to gauge the relative importance of genetic and environmental factors that cause individual differences. The next question might be to determine *how much* genetics contributes to a trait. This can also be described as the proportion of the phenotypic variance that is accounted for by genetic influences, or heritability. Heritability can be estimated from the phenotypic correlations between relatives. For example, in the adoption design, if the correlation for 'genetic' (adopted apart, such as birth mother and her biological child who was placed in an adoptive family at birth) relatives is zero, then heritability is suggested to be zero. If identical and fraternal twin correlations are the same, heritability is zero. If identical twins correlate 1.0 and fraternal twins correlate 0.50, a heritability of 100% is implied (Plomin et al., 2013). A rough estimate of heritability can be obtained by doubling the difference between MZ and DZ twin correlations.

Unfortunately, the results of quantitative genetic studies are often misunderstood and thus misinterpreted. It cannot be reiterated enough that heritability is the genetic contribution to individual differences (variance), not to the phenotype of a single individual. Heritability is a statistic that describes the contribution of genetic differences to the observed differences among individuals in a particular population at a particular time. In different populations or at different times, environmental or genetic influences may differ, and thus heritability estimates in such populations might differ (e.g., differences in obesity attributable to Western obesogenic environments might lead to different heritability estimates in different eras in US history and might result in very different heritability estimates for obesity in cultures with non-Western diets). Moreover, the causes of individual differences within a group have no implication to the cause of average differences between groups (e.g., men and women). Heritability refers to the genetic contribution to differences among individuals within a group. Heritability also does not imply genetic determinism. That is, just because a trait is heritable, or shows genetic influences, does not mean that nothing can be done to change it. Considering a behavioral example, if reading disability is 50% heritable, that does not mean that if one or more of your parents has a reading difficulty that you are 50% at risk. Rather, it means that of the reading scores available for a particular population at a particular time, 50% of the variance of those scores can be attributed to genetic influences. If we also consider height, which has a heritability of 90%, this does not mean you grew to 90% of your height for genetic reasons and the other inches are due to your environment. That said, even for a highly heritable trait such as height, changes in the environment could make a big difference.

A new direction in quantitative genetics uses measured genotypes to directly estimate genetic influences (rather than indirectly via comparison of correlations between sets of relatives, such as MZ and DZ twin pairs). This technique is called genome-wide complex trait analysis (GCTA; Yang et al., 2011). This method compares pairwise genetic similarity across DNA markers called single nucleotide polymorphisms (SNPs) for each pair of unrelated individuals. SNPs involve the mutation of a single nucleotide and are by far the most common type of DNA polymorphism. The GCTA method makes it possible to estimate heritability directly from DNA markers in a manner that is analogous to more traditional quantitative genetic methods, such as the twin design. More specifically, GCTA can estimate the total amount of phenotypic variance captured by all SNPs on the current generation of commercial genotyping arrays (Yang et al., 2011). This method comes with its own set of limitations; however, the value is that is does not require special samples such as twins or adoptees. Essentially, in any large sample with genotypes on hundreds of thousands of SNPs, GCTA can be used to estimate genetic influences for behavioral traits, such as the personality dimensions of neuroticism and extraversion (Vinkhuyzen et al., 2012) as well as the genetic overlap between traits, such as longitudinal measures of cognitive ability from childhood to old age (Deary et al., 2012).

Behavioral genetics not only gives us an idea of whether and how much genetics contributes to a trait, it also tells us about the environment. In fact, behavioral and psychiatric genetics research provides the best possible evidence for the importance of the environment. That is, if the heritability of a certain trait is 50%, the remaining 50% is due to the environment. For complex genetic traits such as those behavioral phenotypes described here, genes do not equal destiny.

Identifying Genes and the Pathways between Genes and Behavior

Quantitative genetics continues to identify *whether* there are genetic influences on behavioral traits, as well as *how much* those genetic factors account for. However, heritability estimates tell us nothing about specific genes. Since the 1980s, quantitative genetics and molecular genetics have been working together to identify specific genes responsible for genetic influence on behavior. These are exciting times because the identification of specific genes, or sets of genes, can provide information that will guide the exploration of pathways between genes and behavior.

For complex human behaviors, genetic influences are likely comprised of multiple genes of varying but small effects rather than a single gene or a few genes of large effect. The goal is to find some, but likely not all, of the genes that make contributions of varying effect size to the variance of a trait. There are multiple molecular genetic strategies that aid in this goal, for example: (1) whole-genome sequencing (Lander, 2011) detects most DNA polymorphisms across the genome and will enable researchers to look at both coding and noncoding parts of the genome in an attempt to capture genetic effects on behavior, but this remains a very expensive strategy; (2) exome

sequencing, or sequencing the 2% of the genome that contains protein-coding information (Bamshad et al., 2011), is becoming widely used as the cost is less prohibitive than whole-genome sequencing; and (3) microarrays continue to be used to genotype millions of SNPs on a chip, and have allowed for genome-wide association to dominate the attempts to identify genes in recent years.

Genome-wide association results have been less promising than once expected in that they have yielded genes accounting for much less of the genetic variance than suggested by heritability estimates. For most traits, the associated SNPs from genome-wide association studies only explain a small fraction of the heritability, leading to the 'missing heritability' problem (Maher, 2008). There are multiple explanations for this missing heritability (see Maher, 2008; Manolio et al., 2009), including the possibility that the heritability is not missing but has not previously been detected because the individual effects are too small to pass stringent significance tests in genome-wide association analyses (Yang et al., 2011). Using attention-deficit hyperactivity disorder (ADHD) as an example, how can we as a field reconcile the fact that the variance explained by successfully located genetic variants for ADHD account for only 1% of the variance (Franke et al., 2009), while heritability estimates from twin and family studies suggest approximately 80% (Knopik et al., 2005)? This concept of 'missing heritability,' that is, the presumed inconsistency between heritability estimates and the ability to explain this genetic variance with genetic markers (Maher, 2008), is nontrivial. New technologies, such as exome or whole-genome sequencing, may begin to shed light on this issue. In addition, in the interim it is possible to aggregate the small effects of many DNA variants associated with a trait (Wray et al., 2008). Such polygenic composites have been called polygenic susceptibility scores (Pharoah et al., 2002), genomic profiles (Khoury et al., 2004), SNP sets (Harlaar et al., 2005), genetic risk scores (Morrison et al., 2007), and aggregate risk scores (Purcell et al., 2009). With the advent of rare variant genotyping, new approaches combine the effects of rare and common variants, including variants that are risk inducing as well as protective (Neale et al., 2011). The missing heritability problem needs to be remedied or at least relieved before the potential of DNA can be actualized in behavioral and psychiatric genetic research (Plomin, 2013). Once we find genes associated with behavior, the next step is to understand the pathways between genes and behavior, or the mechanisms by which genes affect behavior.

Genetic information is transcribed from DNA into messenger RNA and then translated into protein. DNA is the same at all ages, and at all times (Plomin et al., 2013). RNA, on the other hand, degrades quickly and is tissue-specific, age-specific, and state- specific. RNA also regulates transcription and translation of protein-coding DNA. Epigenetic mechanisms are relevant to gene expression in that epigenetic changes do not alter the DNA sequence itself but do affect certain mechanisms of gene expression (Bird, 2007). The epigenome can be thought of as molecular material that sits on top, or outside, of the genome. Epigenetic marks, such as methylation or histone modification, can tell your genes to switch on or off (or gene expression). Thus, it may be through epigenetic marks that environmental factors such as diet, stress, and prenatal nutrition can influence gene expression from one cell to another (Zhang and Meaney, 2010; Knopik et al., 2012; Maccani and Knopik, 2012). Just as in the case of DNA, microarrays can be used to assess the degree of expression of all genes in the genome (the transcriptome) as well as DNA methylation of all coding genes in the genome (the methylome). In addition to the methylome, other forms of epigenetic characterization, such as noncoding RNA and histone modification, comprise the epigenome. The importance of these techniques for behavioral genetics lies in the fact that the epigenome and the transcriptome are the first steps in the correlation between genes and behavior.

Genetics in the Behavioral Sciences

Schizophrenia has been used as an example in both the adoption and the twin design sections earlier in the article. In fact, schizophrenia is the most highly studied area in behavioral genetic research on psychopathology. Other areas about which the most is known are cognitive abilities and disabilities, mood and anxiety disorders, autism, ADHD, personality, and substance use. Health psychology, aging, and evolutionary psychology are also beginning to see an upswing in genetically informed research. Moderate genetic influences are found for most behavioral outcomes that have been studied (Plomin et al., 2013). In general, there is a fairly balanced view that accepts the role of nature as well as nurture for many behavioral disorders. Researchers are also making progress in identifying specific genes.

Some of the most far-reaching genetic findings concern genetic comorbidity. For example, when considering psychopathology, people with one disorder have about a 50% chance of having more than one disorder in a 12-month period (Kessler et al., 2005). Behavioral genetics allows us to ask whether there are really disorders that co-occur or whether those comorbidities are seen simply as a function of the current diagnostic systems that tend to be based on phenotypic description rather than cause. In essence, behavioral genetics allows the questions of whether genes that affect one trait also affect another trait. For example, there is genetic overlap between major depressive disorder and bipolar disorder as well as the possibility of genetic overlap between bipolar disorder and schizophrenia. There is also genetic overlap between anxiety and depression, so much so that they are largely the same disorder from a genetics perspective (Kendler et al., 1992; Middeldorp et al., 2005).

Behavioral genetic human research has maintained a strong focus on the areas of behavioral and social sciences that have historically been interested in individual differences. These have been touched on at points throughout this article and include psychopathology, developmental psychopathology, cognitive ability and disability, personality, and substance use. There are certainly other areas of the social and behavioral sciences that have historically focused on the behavior of groups – these areas can be considered a rich area for behavioral genetics. This includes areas of study such as relationships, self-esteem, attitudes, and behavioral economics. For example, to what extent do genetic influences on parents and offspring contribute to relationships, especially the attachment

between infant and caregiver, or empathic emotional responses? To what extent do genetic factors influence the types of friends that we select (or who select us)? How does genetics influence mate selection or relationship quality? A rapidly growing area is also focused on political psychology, the study of political attitudes and behaviors (Hatemi and McDermot, 2011). Recent efforts also show promise in areas of behavioral economics, such as investor behavior (Barnea et al., 2010), financial decision making (Cesarini et al., 2010), and economic risk-taking (Le et al., 2010; Zhong et al., 2009; Zyphur et al., 2009). The field is moving quickly toward molecular genetic research (Beauchamp et al., 2011; Koellinger et al., 2010).

Behavioral genetics is moving beyond psychology and psychiatry into the fields of economics, neuroimaging, and evolutionary psychology. The questions of *whether* and *how much* genetics influences traits will move forward to begin to answer the question of *how* – that is, determining the mechanisms by which genes have their effects (Plomin et al., 2013). In order to answer questions about how genes influence behavior, gene identification efforts need to find specific genes responsible for the widespread influence on behavior. The pace at which information about genetic and epigenetic modification is being produced has created a strain on our available analytic tools when we try to consider the impact or significance of this genetic information. This poses a significant challenge to the field as far as integrating the effects of these individual differences to better understand the biological basis of behavior (Knopik, 2011). In many respects, the grandest implication for science is that DNA will serve as an integrating force across diverse disciplines (Plomin, 2001; Plomin et al., 2013). Just as molecular genetics might lead to personalized genomics, so too is the possibility that genetic findings can be used to research interventions that might prevent a disorder – a direction that behavioral scientists need to be ready to take.

As more genetic data becomes available, it is critical that behavioral and social scientists be prepared to use this information in their research. In fact, some funding agencies are highlighting the importance of considering biological influences, including genetics, in human behavioral research. As specific genes are implicated in behavioral and psychological disorders, social and behavioral scientists should participate in discussions (and inform their students) about the ethical concerns surrounding these types of findings (Plomin, 2001). There is a fear that the results will be used to justify social inequality, to select individuals for education or employment, or to enable parents to pick and choose among pregnancies for the most genetically desirable fetus. More recently, confidentiality concerns have been raised over the ability to identify individuals based solely on their genetic sequence. Many of these concerns stem from misunderstandings of how genes (and the environment) affect complex traits (Rutter and Plomin, 1997). Behavioral genetics has clearly demonstrated that both genes and the environment are important. Thus, behavioral and social scientists need to be informed and prepared to consider both nature and nurture when considering behavior.

See also: Bioethics in the Post-genomic Era; Epigenetics of Brain Plasticity and Behavior; Familial Studies: Genetic Inferences; Gene–Environment Interactions in Health and Well-being; Genetic Studies of Behavior: Methodology; Genetics and Anthropology; Genetics and Economic Behavior; Human Evolutionary Genetics; Human Genome Diversity Project: History; Human Genome Project: History and Assessment; Intelligence, Genetics of: Cognitive Abilities; Mental Illness, Genetics of; Occupational Health; Personality Differences and Development: Genetic and Environmental Contributions; Sexual Orientation: Genetic Aspects; Twin Studies and Epigenetics.

Bibliography

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., Shendure, J., 2011. Exome sequencing as a tool for Mendelian disease gene discovery. Nature Reviews Genetics 12, 745–755.

Barnea, A., Cronqvist, H., Siegel, S., 2010. Nature or nurture: what determines investor behavior? Journal of Financial Economics 98, 583–604.

Beauchamp, J.P., Cesarini, D., Johannesson, M., van der Loos, M.J.H.M., Koellinger, P.D., Groenen, P.J.F., Christakis, N.A., 2011. Molecular genetics and economics. Journal of Economic Perspectives 25, 57–82.

Bird, A., 2007. Perceptions of epigenetics. Nature 447, 396–398.

Bouchard Jr., T.J., Propping, P., 1993. Twins as a Tool of Behavioral Genetics. John Wiley & Sons, Chichester, UK.

Cardno, A.G., Rijsdijk, F.V., West, R.M., Gottesman, I.I., Craddock, N., Murray, R.M., McGuffin, P., 2012. A twin study of schizoaffective-mania, schizoaffective-depression, and other psychotic syndromes. American Journal of Medical Genetics. B: Neuropsychiatric Genetics 159B, 172–182.

Cesarini, D., Johannesson, M., Lichtenstein, P., Sandewall, Ö., Wallace, B., 2010. Genetic variation in financial decision-making. Journal of Finance 65, 1725–1754.

Deary, I.J., Yang, J., Davies, G., Harris, S.E., Tenesa, A., Liewald, D., Visscher, P.M., 2012. Genetic contributions to stability and change in intelligence from childhood to old age. Nature 482, 212–215.

Derks, E.M., Dolan, C.V., Boomsma, D.I., 2006. A test of the Equal Environment Assumption (EEA) in multivariate twin studies. Twin Research and Human Genetics 9, 403–411.

Franke, B., Neale, B.M., Faraone, S.V., 2009. Genome-wide association studies in ADHD. Human Genetics 126 (1), 13–50.

Harlaar, N., Butcher, L.M., Meaburn, E., Sham, P., Craig, I.W., Plomin, R., 2005. A behavioural genomic analysis of DNA markers associated with general cognitive ability in 7-year-olds. Journal of Child Psychology and Psychiatry 46, 1097–1107.

Hatemi, P.K., McDermot, R. (Eds.), 2011. Man is by Nature a Political Animal: Evolution, Biology, and Politics. University of Chicago Press, Chicago, IL.

Kendler, K.S., Baker, J.H., 2007. Genetic influences on measures of the environment: a systematic review. Psychological Medicine 37 (5), 615–626.

Kendler, K.S., Neale, M.C., Kessler, R.C., Heath, A.C., Eaves, L.J., 1992. Major depression and generalized anxiety disorder. Same genes, (partly) different environments? Archives of General Psychiatry 49, 716–722.

Kessler, R.C., Chiu, W.T., Demler, O., Merikangas, K.R., Walters, E.E., 2005. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry 62, 617–627. http://dx.doi.org/10.1001/archpsyc.62.6.617.

Khoury, M.J., Yang, Q.H., Gwinn, M., Little, J.L., Flanders, W.D., 2004. An epidemiologic assessment of genomic profiling for measuring susceptibility to common diseases and targeting interventions. Genetics in Medicine 6, 38–47.

Knopik, V.S., 2011. Grand challenge in behavioral and psychiatric genetics: quantitative challenges to keeping up with molecular advances. Frontiers in Genetics 2 (9).

Knopik, V.S., Maccani, M., Francazio, S., McGeary, J.E., 2012 Nov. The epigenetics of maternal cigarette smoking during pregnancy and effects on child development. Development and Psychopathology 24 (4), 1377–1390.

Knopik, V.S., Sparrow, E.P., Madden, P.A., Bucholz, K.K., Hudziak, J.J., Reich, W., Slutske, W.S., Grant, J.D., McLaughlin, T.L., Todorov, A., Todd, R.D., Heath, A.C., 2005 May. Contributions of parental alcoholism, prenatal substance exposure, and genetic transmission to child ADHD risk: a female twin study. Psychological Medicine 35 (5), 625–635.

Koellinger, P., van der Loos, M., Groenen, P., Thurik, A., Rivadeneira, F., van Rooij, F., Hofman, A., 2010. Genome-wide association studies in economics and entrepreneurship research: promises and limitations. Small Business Economics 35, 1–18.

Lander, E.S., 2011. Initial impact of the sequencing of the human genome. Nature 470, 187–197. http://dx.doi.org/10.1038/nature09792.

Le, A.T., Miller, P.W., Slutske, W.S., Martin, N.G., 2010. Are attitudes towards economic risk heritable? Analyses using the Australian Twin Study of Gambling. Twin Research and Human Genetics 13, 330–339.

Li, M.D., 2010. Grand challenges and opportunities for molecular psychiatry research: a perspective. Frontiers in Psychiatry 1, 2.

Maccani, M., Knopik, V.S., 2012. Cigarette smoke exposure-associated alterations to noncoding RNA. Frontiers in Genetics 3, 53.

Maher, B., 2008. Personal genomes: the case of the missing heritability. Nature 456, 18–21.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F., McCarroll, S.A., Visscher, P.M., 2009 Oct 8. Finding the missing heritability of complex diseases. Nature 461 (7265), 747–753.

Middeldorp, C.M., Cath, D.C., Van Dyck, R., Boomsma, D.I., 2005. The co-morbidity of anxiety and depression in the perspective of genetic epidemiology. A review of twin and family studies. Psychological Medicine 35, 611–624.

Morrison, A.C., Bare, L.A., Chambless, L.E., Ellis, S.G., Malloy, M., Kane, J.P., Boerwinkle, E., 2007. Prediction of coronary heart disease risk using a genetic risk score: the atherosclerosis risk in communities study. American Journal of Epidemiology 166, 28–35.

Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Daly, M.J., 2011. Testing for an unusual distribution of rare variants. PLoS Genetics 7, e1001322.

Pharoah, P.D.P., Antoniou, A., Bobrow, M., Zimmern, R.L., Easton, D.F., Ponder, B.A.J., 2002. Polygenic susceptibility to breast cancer and implications for prevention. Nature Genetics 31, 33–36.

Plomin, R., 2001. Behavioral genetics: psychological perspectives. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of Behavioral and Social Sciences, first ed. Pergamon, Oxford, UK.

Plomin, R., 2013. Child development and molecular genetics: 14 years later. Child Development 84 (1), 104–120.

Plomin, R., Crabbe, J.C., 2000. DNA. Psychological Bulletin 126, 806–828.

Plomin, R., DeFries, J.C., Knopik, V.S., Neiderhiser, J.M., 2013. Behavioral Genetics, sixth ed. Worth Publishers, New York.

Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., Moran, J.L., 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748–752.

Robbers, S.C.C., van Oort, F.V.A., Polderman, T.J.C., Bartels, M., Boomsma, D.I., Verhulst, F.C., Huizink, A.C., 2011. Trajectories of CBCL attention problems in childhood. European Child and Adolescent Psychiatry 20, 419–427.

Rutter, M., Plomin, R., 1997 Sept. Opportunities for psychiatry from genetic findings. British Journal of Psychiatry 171, 209–219.

Sullivan, P.F., Kendler, K.S., Neale, M.C., 2003. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Archives of General Psychiatry 60, 1187–1192.

Vinkhuyzen, A.A., Pedersen, N.L., Yang, J., Lee, S.H., Magnusson, P.K., Iacono, W.G., McGue, M., Madden, P.A., Heath, A.C., Luciano, M., Payton, A., Horan, M., Ollier, W., Pendleton, N., Deary, I.J., Montgomery, G.W., Martin, N.G., Visscher, P.M., Wray, N.R., 2012 Apr. Common SNPs explain some of the variation in the personality dimensions of neuroticism and extraversion. Translational Psychiatry 17 (2), e102.

Wray, N.R., Goddard, M.E., Visscher, P.M., 2008. Prediction of individual genetic risk of complex disease. Current Opinion in Genetics and Development 18, 257–263.

Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M., 2011. GCTA: a tool for genome-wide complex trait analysis. American Journal of Human Genetics 88, 76–82.

Zhang, T.Y., Meaney, M.J., 2010. Epigenetics and the environmental regulation of the genome and its function. Annual Review of Psychology 61, 439–466. http://dx.doi.org/10.1146/annurev.psych.60.110707.163625.

Zhong, S., Chew, S.H., Set, E., Zhang, J., Xue, H., Sham, P.C., Israel, S., 2009. The heritability of attitude toward economic risk. Twin Research and Human Genetics 12, 103–107.

Zyphur, M.J., Narayanan, J., Arvey, R.D., Alexander, G.J., 2009. The genetics of economic risk preferences. Journal of Behavioral Decision Making 22, 367–377.

Behavioral Inhibition and Social Withdrawal across Cultures

Jonathan D Lane, Harvard University, Cambridge, MA, USA
Ju-Hyun Song, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by K.H. Rubin, volume 21, pp. 14055–14058, © 2001, Elsevier Ltd.

Abstract

Behavioral inhibition (BI) refers broadly to a temperament trait of avoiding interaction with novel stimuli because such stimuli provoke anxiety or distress. *Social withdrawal* (SW) is a related construct that refers more specifically to a tendency to avoid interaction with people. Research on these behaviors has flourished over the past 30 years. Such research has demonstrated that biological, cognitive, and social factors contribute in shaping children's BI and SW, and mutually influence the developmental outcomes of these behaviors. Recent research has also shed important light on cultural variability in the social and psychological implications of these behaviors.

The study of children's behavioral inhibition (BI) and social withdrawal (SW) has flourished over the past 30 years, and research over the past decade has shed important light on the social and psychological implications of these behaviors across cultural contexts. BI refers broadly to a temperament trait of avoiding interaction with novel stimuli – including objects, places, and people – because such stimuli provoke anxiety or distress (Fox et al., 2005; Kagan, 1997). Such behavior is found in roughly 10–20% of infants, although the prevalence of such behavior varies across cultures (Kagan, 1997; Chen and French, 2008). Through the 1980s and the 1990s, Jerome Kagan and his colleagues pioneered work on the development and physiological correlates of BI during infancy and early childhood. Meanwhile, the laboratories of Kenneth Rubin, Mary Rothbart, and Jens Asendorpf explored behaviors theoretically linked to BI – SW and *shyness* – which, in contrast to BI, refer more specifically to a tendency to inhibit interactions with people. These research groups focused particularly on the development of SW and shyness during early childhood and identified factors that influence and are influenced by their development. The works of Kagan, Rothbart, Rubin, and Asendorpf, and of their colleagues and students, have had a lasting and profound influence on the field.

Throughout the literature, the terms 'social withdrawal' and 'shyness' have often been used interchangeably, and different researchers have used these terms to refer to different constructs. Recently, theorists have conceptualized SW as a category of behaviors, encompassing several distinct subtypes of withdrawal, one of which is *shyness* (Asendorpf, 1990; Coplan and Rubin, 2010). The following section describes these subtypes. Next, common measures of BI and SW are described. Subsequent sections review biological, cognitive, and social correlates of BI and SW, and describe how the development of BI and SW can be credited to transactional processes among these factors. Importantly, the social evaluation and developmental outcomes of BI and SW differ depending on proximal family and peer contexts as well as broader sociocultural contexts (Chen and French, 2008; Chen et al., 2011). Thus, throughout this article, important, recent research is highlighted, which documents the development and consequences of these behaviors across cultures. The article concludes with an outlook on future directions in the field.

Subtypes of SW

Various subtypes of SW have been identified, each with different causes, correlates, and outcomes. Broadly, subtypes of SW can be categorized along two dimensions – a motivation to approach and a motivation to avoid (Asendorpf, 1990). The term *shyness* has been used to refer to a form of SW characterized by anxiety or self-consciousness in novel social situations, especially situations in which one believes he or she is being evaluated (Asendorpf, 1990; Coplan et al., 2013). Shy children often exhibit *reticent* behavior – watching others interact from the periphery but not directly engaging in social interaction (Coplan et al., 2013). These children may loom near or play beside other children, but tend to not initiate interaction with unfamiliar people and are wary when approached by unfamiliar others. Thus, these children are often described as experiencing an *approach–avoidance conflict* – they want to interact with others, but remove themselves from direct involvement because of the anxiety or distress that social situations provoke.

Socially avoidant children tend to experience extreme fear and anxiety in the face of social interaction and thus have an especially strong avoidance motivation (Asendorpf, 1990; Coplan et al., 2013). Unlike shy children, socially avoidant children have a low approach motivation and less often demonstrate reticent behavior. Their aversion to social interaction leads some socially avoidant children to behave aggressively and impulsively in social situations, and because of their disruptive behavior and negative affect, these children are at higher risk for peer rejection. *Unsociable* or *socially disinterested* children prefer to play alone, but are not particularly anxious or socially disruptive, and are comfortable interacting with others if others initiate the interactions (Asendorpf, 1990; Coplan et al., 2013). Shy children are often attentive of others' social interactions, whereas unsociable children are more focused on objects,

such as toys. Thus, these children can be described as having neither a strong motivation to approach nor a strong motivation to avoid social situations.

SW traditionally refers to children's own (internal) motivation to disengage from direct social involvement. This should be distinguished from *active isolation* (also referred to as *active withdrawal*), which describes social isolation resulting from peer rejection (Asendorpf, 1990; Harrist et al., 1997; Rubin et al., 2009). Although active isolation does not necessarily reflect an individual's motivation to socially disengage, SW – especially SW characterized by high levels of negative affect and socially disruptive behavior – may eventually lead to peer rejection and thus active isolation (Coplan and Rubin, 2010; Rubin et al., 2009).

In the following sections, the term 'social withdrawal' will be used to refer generally to removal of oneself from social interaction, for one reason or another. When the empirical literature explicitly indicates which specific subtypes of SW were examined (e.g., shyness), those subtypes are noted.

Measures of BI and SW

Researchers use a variety of methods to measure BI and SW (Coplan and Rubin, 2010; Rubin et al., 2009). These methods include questionnaires on which parents or teachers rate the frequency or magnitude of children's behavior along several dimensions, including BI or SW. Parents' and teachers' evaluations have the advantage that they reflect children's behavior across a wide time span (e.g., a year in the classroom, or a lifetime) and, especially in the case of parent questionnaires, across a variety of contexts (e.g., at home, in public). Yet, such questionnaires are subject to biases; parents especially may rate their children more positively and more consistent with social norms. Researchers have additionally used peer ratings, whereby children in a classroom identify which of their peers they consider to be shy, friendly, aggressive, and so forth. Both teacher and peer evaluations may be less biased than parent ratings, but are limited in that only classroom behavior is typically considered in the evaluations.

Researchers also directly observe children's behavior in their homes, with peers in their classrooms, or with other children and adults in research laboratories. One of the most widely used measures of BI, developed by Kagan and colleagues, involves having toddlers or preschoolers and their caregivers in a room with novel stimuli (e.g., toys, people); investigators record children's reactions – their fear toward, approach toward, and interaction with the novel stimuli and their tendency to seek their caregivers for comfort. These measures are more objective than parent ratings, but are limited in that children's behavior is often observed across a narrow time span and within a limited context (e.g., an hour in the laboratory). Recently, self-report measures have been used to gauge SW among older children; these measures have the potential to more fully capture the personal motivations behind individuals' SW, but are subject to self-report biases (e.g., to portray oneself positively). Studies that have used several of these methods have revealed modest to strong correlations among measures (Rubin et al., 2009). Importantly, such studies have also identified that the measured inhibition varies depending on the context in which it is assessed and depending on who performs the evaluations.

Development of BI and SW: A Transactional Process

Developmental precursors to BI and SW can be found in early infancy. Reactivity – one's physiological and behavioral response to stimulation – is a strong predictor of BI and SW (Kagan, 1997; Rothbart et al., 2011). Kagan and colleagues found that more reactive 4-month-olds in the United States – that is, those who moved and cried more when they were stimulated by toys, faces, and voices – demonstrated greater fear toward unfamiliar events at the age of 1 and 2 years, when compared to nonreactive children (Kagan, 1997; Kagan and Snidman, 1991). Children who had been classified as fearful/inhibited toddlers talked and smiled much less than their peers at 4 years; at this age, many of these children would be considered socially withdrawn. Moreover, inhibition during infancy and early childhood is predictive of SW in late childhood, adolescence, and early adulthood; and this has been found in a variety of cultures (Fox et al., 2005; Rubin et al., 2009). Extremely withdrawn children are particularly likely to demonstrate a stable pattern of SW across development (Rubin et al., 2009). However, only a small portion of children maintain an extreme form of a reactive/inhibited profile across development; instead, it is more common for children to regress toward the mean – for example, for extremely reactive infants to develop into somewhat timid children (Kagan, 1997). Thus, early BI is predictive of later SW, and SW in early childhood is predictive of SW throughout development. Moreover, SW is relatively stable across contexts – children with extreme SW in particular tend to exhibit such behavior toward their peers in a variety of settings (e.g., at home, in school, at community events).

Findings that developmental precursors to BI and SW are present very early in life, findings that inhibited individuals often remain so throughout development, and genetic studies demonstrating the heritability of inhibited behavior all point to the strong biological bases of BI and SW (Fox et al., 2005; Kagan, 1997; Kagan and Snidman, 1991). However, relations between biology and behavior are not straightforward – early cognitive and behavioral dispositions interact to predict SW. For example, among US children, Pérez-Edgar et al. (2011) found an association between early BI (at 2–3 years) and later SW (at 5 years) only among children who paid the greatest attention to threatening stimuli – those who stared longest at angry faces during a laboratory task.

Children's environments, including the broader culture and the more proximate family and peer contexts, also shape the development of inhibited behavior as well as the psychosocial consequences of such behavior (Chen and French, 2008; Chen et al., 2011; Oh et al., 2008). Indeed, recent research has demonstrated that biological, cognitive, and environmental factors interact to shape children's BI and SW and mutually influence the developmental outcomes of these behaviors. Thus, the development of these behaviors and their outcomes can best be described as functions of complex transactions between internal factors (biology, cognition) and

environmental factors (other people) – infants' biologically driven reaction to novel entities influences how other people respond to infants, which in turn shapes children's biological, emotional, and cognitive development, which then further shapes children's environments (Rubin et al., 2009). This reciprocal, transactional process continues throughout the life span. Thus, even though there is moderate stability in BI and SW across development (Kagan, 1997; Rubin et al., 2002), environmental and cognitive factors influence the prevalence and outcomes of these behaviors.

Biological Correlates of BI and SW

Inhibited behavior is associated with several neurophysiological systems and processes, including the limbic system, autonomic nervous system (ANS), hypothalamic-pituitary-adrenal (HPA) axis, and prefrontal cortex (PFC). The amygdala is a medial temporal lobe structure involved in the limbic system. This structure plays an important role in processing emotional sensory information, including others' facial expressions, especially negative expressions, such as fear, and auditory stimuli associated with danger or anger. How the amygdala processes potentially threatening stimuli influences regulation of both the ANS – via the amygdala's projections to the hypothalamus – and the HPA axis; this has implications for BI and SW (Fox et al., 2005; Kagan, 1997; Lopez et al., 2004).

The ANS is composed of the sympathetic nervous system (which regulates 'flight-or-flight' responses to threat) and the parasympathetic nervous system (which calms the body down following a threat). The high reactivity typical of shy individuals might be partly attributable to overactivation of the sympathetic nervous system or to poorer functioning of the parasympathetic nervous system (Kagan, 1997). Indeed, studies have demonstrated that behaviorally inhibited toddlers and young children tend to have higher resting heart rates – an indicator of ANS activation – and resting heart rate is a predictor of SW in the preschool years (Fox et al., 2005; Kagan, 1997).

The HPA axis is also responsible for regulating humans' response to stress. When an individual encounters stressors, the HPA axis produces hormones, which activate other neural systems, eventually leading to the release of cortisol by the adrenal glands; cortisol, in turn, regulates the physiological response to the stressors. Thus, the level of cortisol found in saliva, urine, or blood is commonly used as a marker of individuals' level of physiological stress. Many studies have found a positive association between children's daily resting cortisol levels and SW (Lopez et al., 2004). For example, Blair et al. (2004) found that children who were rated as having both high approach and high withdrawal motivations (i.e., a typical characterization of shy children caught in an approach–avoidance conflict) had both higher resting heart rates and higher cortisol levels. One prominent hypothesis is that some withdrawn children experience consistently high levels of stress (reflected in their heart rates and cortisol), leaving them with a relatively low threshold for additional social stress. Thus, rather than socially engaging, these children withdraw from social situations in an attempt to avoid additional stress (Fox et al., 2005; Lopez et al., 2004).

The PFC is involved in a host of cognitive processes. Critical to SW is the role of the PFC in regulating emotion and behavior (Fox et al., 2005; Lopez et al., 2004). Generally, the left regions of the PFC are involved in processing positive emotions and approach-oriented emotional responses, whereas the right regions are involved in processing negative emotions and withdrawal-oriented emotional responses. Cerebral asymmetry in processing emotions, specifically more marked activation of the right hemisphere relative to the left hemisphere (i.e., 'right frontal asymmetry'), is linked to BI and SW. Inhibited infants, preschoolers, and older children show greater right frontal asymmetry, whereas outgoing infants and children show greater left frontal asymmetry (Fox et al., 2005). Lopez et al. (2004) have proposed that right frontal asymmetry might reduce signals that the PFC sends to decrease activation of the amygdala, which would in turn maintain HPA axis activation. The PFC also plays a role in regulating individuals' behavioral responses to physiological and environmental stimulation and thus could directly influence whether and how individuals' SW manifests (Rothbart et al., 2011).

BI and SW in Social Contexts

Cultural Influences on BI and SW

Different behaviors are favored in different cultural contexts, and social expectations about shy or reserved behavior vary markedly between cultures (Chen and French, 2008). These social expectations influence the prevalence of BI and SW, the ways inhibited individuals are treated by others, and their psychological health.

Cultural values influence people's attitudes toward social behavior. Broadly, many Western societies, such as the United States and Canada, highly value individualist behaviors including autonomy and assertiveness, while many East Asian and Latin American societies place greater emphasis on collectivist behaviors, such as self-control and actions that facilitate group harmony (Triandis, 1995). Individualist and collectivist values coexist in most cultures, but the relative balance of these values within cultures can help us understand the development, prevalence, evaluation, and adjustment outcomes of BI and SW across cultures.

Cultural values can influence the development and consequences of children's BI and SW through several pathways. First, cultural values shape parents' socialization goals for their children. Second, children adopt existing cultural norms for social behavior and apply them when they respond to peers' SW (Chen and French, 2008). Thus, the prevalence of BI and SW will vary depending on cultural values, as parents and peers either facilitate or discourage their development. For example, during laboratory free play sessions, toddlers in more collectivist East Asian countries, such as Korea and China, show more inhibited behavior than toddlers in more individualist cultures, such as Italy and Australia (Rubin et al., 2009).

BI and SW in Parent–Child Relationships

The quality of the parent–child attachment relationship is strongly associated with the development of BI and SW. In the

United States, infants with insecure attachments to their primary caregivers are more likely to exhibit BI during toddlerhood and are more likely to exhibit SW in preschool and school age (Hastings et al., 2010). Similarly, urban Chinese school-aged children's self-reported insecure attachment with their caregivers is associated with SW, whereas secure attachment is associated with greater social initiative (Chen, 2012). These findings suggest that feelings of mistrust and fear derived from an early insecure attachment might influence the development of SW. Children's SW can also affect the parent–child attachment relationship. For example, inhibited children's high arousal and wariness tend to elicit extremely solicitous (i.e., physically close and controlling) parenting, which may contribute to an insecure attachment relationship.

Mothers' affect, personalities, and parenting styles are closely linked to shy children's socioemotional adjustment. For instance, mothers' negative affect and neuroticism are linked to increased internalizing problems and social dissatisfaction among shy preschoolers. Parents with neuroticism are more likely to adopt an overprotective parenting style especially when their children are shy because such parents are more likely to see their children's behavior as problematic and in need of correcting (Rubin et al., 2009). Overprotective parenting can discourage independence and reduce opportunities for children to develop coping skills and problem-solving strategies, which can be especially problematic for withdrawn children. Overprotective parenting is predictive of inhibited preschoolers' shyness and reticence in the United States and Korea, and is predictive of peer difficulties and internalizing problems among Canadian kindergarteners (Hastings et al., 2010; Rubin et al., 2009). Other studies have found that parents' perceptions of their toddlers' shyness predicts overprotective parenting in the following years. Collectively, the findings just reviewed indicate that there are bidirectional relations between children's shyness and parental behavior across development (Rubin et al., 2009).

Certain qualities of parents and parenting styles can facilitate the positive development of shy children. For example, Canadian mothers' agreeableness – their competence in regulating their own emotions – is associated with fewer internalizing problems and peer difficulties among shy kindergarteners (Rubin et al., 2009). Agreeable mothers tend to display higher parental warmth and responsiveness and are able to assist in their shy children's emotion regulation; such parenting mitigates the links between shyness and negative adjustment outcomes. Thus, parental reactions to children's early dispositions influence children's developmental trajectory to SW and influence adjustment outcomes for withdrawn children.

How parents evaluate and respond to children's SW varies across cultures. In general, parents in individualist cultures tend to view SW negatively and associate such behavior with submissiveness, dependency, and timidity (Rubin et al., 2009), whereas parents in collectivist cultures evaluate SW more positively as a sign of respectfulness and self-control (Chen and French, 2008). There is a growing recognition, however, that strong individualist and collectivist values coexist within some societies (e.g., Cheah and Park, 2006; Hastings et al., 2010). This seems to be particularly the case in Westernized Asian cultures such as South Korea, urban China, and India (e.g., Bowker and Raja, 2011; Kim et al., 2008). In these changing societies, parental attitudes, socialization goals, and parenting strategies seem to reflect both traditional Asian values and Western values. For example, whereas Korean culture is rooted in a traditional collectivist ideology, the Korean War and industrialization brought with them an emphasis on the success of one's own family, which encouraged greater competition. These recently adopted Western values are reflected in contemporary Korean mothers' negative view of their children's SW (Cheah and Park, 2006). Moreover, Korean mothers prioritize child-centered goals when trying to promote their social behaviors (e.g., making children feel good about themselves, trying to understand children's perspectives), which also reflects individualist values. Korean mothers, however, attribute their children's SW to external causes, which reflects the traditional collectivist ideology that personalities are largely products of environmental factors rather than innate factors. Thus, in the context of sociocultural shifts, changing cultural norms are reflected in parents' attitudes about children's SW as well as in their parenting goals and strategies (Hastings et al., 2010).

BI and SW in Peer Relationships

Peer relationships are one of the main socializing settings for children, and children's behavior among peers provide a meaningful indicator of their social competence. Withdrawn children are more likely than sociable children to experience neglect and rejection by peers, and to be victimized through chronic bullying by their peers (Rubin et al., 2009). One potential reason for this rejection is that peers view withdrawn children as not complying with age-specific behavioral norms. Another possibility is that withdrawn children lack age-appropriate social skills and are seen as easy targets for victimization (Rubin et al., 2009).

Peer interactions deserve special attention because they are not only affected by the unsociable behavior of withdrawn children but also can influence the developmental trajectory of children's SW. Oh et al. (2008) identified groups of US children with different developmental trajectories for SW across the period from fifth grade to the final year of middle school: *low stable* (children who were not classified as withdrawn at any point in the study), *increasing SW*, and *decreasing SW*. Children who experienced high levels of peer exclusion and victimization were more likely to be categorized as withdrawn in fifth grade. Increased SW over the subsequent years was predicted by friendship instability, the absence of a mutual best friend, or having a best friend who was highly socially withdrawn. Decreased SW was predicted by decreased rejection and peer victimization during the transition to middle school. These findings suggest that having a positive peer context and having friends who model more outgoing behavior may encourage socially withdrawn children to engage in social interaction, effectively reducing their SW.

Norms for peer interaction vary across cultures, and thus how peers respond to BI and SW differs across cultures. For example, shy children in China are viewed less negatively by peers than shy children in the United States and Canada (Chen and French, 2008). In addition, cultural shifts across history

have yielded cross-cohort differences in how adaptive these behaviors are. For example, researchers have found that social adjustment outcomes vary between cohorts of shy elementary school children in urban China (Chen and French, 2008). Shyness was associated with positive outcomes – such as peer acceptance and leadership – in a 1990 cohort, but with negative peer reputation and school maladjustment in a 2002 cohort. Similar to what has been found in other Asian countries, recent social and economic changes in urban China have introduced individualist values such as self-assertion, and this has led people to view inhibited behaviors as suboptimal in the new competitive environment. In rural areas of China, however, where less dramatic economic and cultural changes have taken place, strong traditional collectivist values persist and shy children continue to experience a positive reputation among peers (Chen et al., 2011).

Each of the subtypes of SW may be evaluated differently by peers. Coplan et al. (2007) examined how young Canadian children perceived SW when presented hypothetical vignettes describing the behaviors of shy, unsociable, aggressive, and socially competent children. They found that as early as kindergarten, children distinguished between the different motivations behind SW, and preferred a shy playmate over an unsociable playmate. Similarly, Chen et al. (2011) found differences in the social adjustment of unsociable and shy school-aged children in rural China; unsociability was linked with peer rejection, whereas shyness was associated with peer- and teacher-rated social competence.

Emotional and Cognitive Outcomes of BI and SW

Self-Concept and Coping

Children identified as extremely shy or avoidant tend to experience many social difficulties, such as neglect and peer rejection. Due to frequent negative peer interactions, withdrawn children tend to attribute their social failures to internal factors rather than to their peers or to contextual factors (Wichmann et al., 2004). Consequently, these children are at a greater risk for developing a poor self-concept. Moreover, withdrawn children commonly use avoidant coping strategies – such as doing nothing rather than active problem solving – when they are faced with challenging social situations (Markovic et al., 2013). Researchers suspect that the main obstacles to proactive coping for these children are overarousal in social situations and self-defeating appraisals of their social skills. Such negative appraisals and low self-esteem can contribute to the development of internalizing problems.

Internalizing Problems

Many studies across cultures have found that early BI and SW are associated with internalizing problems, such as anxiety, depression, and loneliness. The relation between SW and anxiety can be considered transactional – one can reinforce the other and therefore the association between the two often persists longitudinally. SW and depression have also been associated in many studies; this might be due to withdrawn children's lower self-esteem, loneliness, or their use of avoidant coping strategies. For example, a longitudinal project in Canada found that childhood SW predicted loneliness and depression in adolescence (Rubin et al., 2009). In a study that followed up Norwegian children from infancy to early adolescence, SW was contemporaneously and longitudinally linked to anxiety and depression (Karevold et al., 2012).

Psychological outcomes vary depending on the subtype of SW and the cultural context. Among the different subgroups of withdrawn school-aged children in Canada, Coplan et al. (2013) found that avoidant children reported the highest levels of social anxiety and depressive symptoms, shy children reported moderately high levels of social anxiety and depressive symptoms, and unsociable children reported levels comparable to nonwithdrawn children. A different pattern was found among school-aged children in rural China, where strong traditional, collectivist values are held: unsociability, but not shyness, was associated with depression (Chen et al., 2011). Thus, whether a specific subtype of SW is adaptive depends on the cultural context.

Interesting associations between SW and psychological adjustment are found in Westernized Asian cultures, where people engage in some individualist behaviors within traditional collectivist contexts. Among Korean college students, shyness during adolescence is only weakly associated with emotional distress, reflective of the collectivist context; unsociability is not related to relational difficulties or to emotional distress in college, which is consistent with the pattern found in individualist cultures (Kim et al., 2008). These findings suggest that as long as children's preference for solitude does not disrupt group harmony, it might not compromise their socioemotional adjustment in such societies. Bowker and Raja (2011) found a similar pattern for unsociability among Indian adolescents, but additionally found that social avoidance was associated with loneliness. Avoidance strongly contrasts with the collectivist value of group cohesion and thus might be especially maladaptive in such cultural contexts.

Finally, evidence suggests that links between SW and psychological adjustment vary by gender in Western societies. Withdrawn boys are at a greater risk for negative outcomes; they are more likely to be excluded by peers and experience socioemotional difficulties across the life span compared to withdrawn girls (Rubin et al., 2009). This gender difference might be partly due to the differences in cultural and parental expectations for boys and girls; BI and SW are seen as less socially acceptable for boys than for girls in Western cultures.

Social Cognition

From an evolutionary standpoint, researchers have recently proposed that a reserved approach to novelty, including social novelty, might have provided the foundation on which humans were capable of making significant social cognitive advances, including advances in their understanding of how others think, their theory-of-mind (ToM). Indeed, human toddlers are more cautious toward novel stimuli compared to humans' closest primate cousins – bonobos, chimpanzees, and orangutans (Hermann et al., 2011) – all of which young children surpass in social cognition. Researchers have started to directly examine relations between SW and the ontogeny of

social cognition, and have demonstrated that some forms of SW are related to more advanced social cognition in children.

Harrist et al. (1997) used teacher ratings of US children's behavior to identify a group of non-withdrawn children, as well as subgroups of withdrawn children, including those who were unsociable, shy (whom Harris et al. term *passive-anxious*), and actively isolated. Social cognition was measured by presenting children hypothetical situations in which they were rejected or ignored when trying to enter a social group or in which another child bothered them. For each situation, children were asked, among other things, to recall the situation and to judge whether the story characters were acting mean (such judgments are called 'hostile attributions'). Kindergarteners who were categorized as *actively isolated* – who were more impulsive, angry, and defiant – had the poorest recall of the situations compared to other withdrawn and non-withdrawn children; and during elementary school, these children made more hostile attributions than the other withdrawn children. In contrast, unsociable children and shy children made fewer hostile attributions during elementary school, even when compared to non-withdrawn children.

Wellman et al. (2011) examined relations between dimensions of US preschoolers' temperament – including shyness, perceptual sensitivity, and aggression – and their ToM. When children were aged 3 years and 5 years, parents completed questionnaires about their children's behavior, and children were given ToM tasks in the laboratory. For the ToM tasks, children were presented scenarios where a character holds a false belief and children were asked to predict and explain characters' beliefs and behavior. Wellman et al. (2011) found that children who were shyer, more perceptually sensitive and less aggressive at the age of 3 had a more advanced ToM at the age of 5. Thus SW may have advantages; shy children, especially those who are nonaggressive and socially attentive, may learn more about others by watching from the periphery of social interaction, compared to their non-shy peers.

Lane et al. (2013) assessed relations between different forms of SW and ToM in two cultures that differed in their social norms concerning shyness: the United States and China. Two types of children were identified using parent ratings of SW and data on children's HPA axis reactivity (gauged via salivary cortisol). Withdrawn children with lower HPA reactivity evidenced more advanced ToM (gauged with false-belief tasks). However, withdrawn children with relatively high HPA reactivity evidenced poorer ToM. Parents' ratings of children's behavior along other dimensions helped to further distinguish the two groups. Withdrawn/high-reactive children had poorer perceptual sensitivity, poorer inhibitory control, and lower levels of smiling and laughter than other children; thus these children would fit the profile of disruptive, socially avoidant children. In contrast, withdrawn/low-reactive children did not differ from non-withdrawn children on these other dimensions; thus these children would fit the profile of shy children.

Lane et al. (2013) found that the social cognitive differences between these two groups were especially marked for Chinese children – withdrawn/low-reactive (i.e., shy) children were especially likely to outperform withdrawn/high-reactive (i.e., disruptive and avoidant) children. As discussed earlier, this likely reflects cultural differences in behavioral norms and expectations. In China, reactive behavior is more discouraged than in the United States, whereas shy or reticent behavior is more encouraged (Chen and French, 2008). Chinese children in this latter withdrawn group might be more welcomed in or near social situations, providing more and better opportunities to learn about others, enhancing their social cognition (Lane et al., 2013).

Suway et al. (2011) assessed how US 2-year-olds' BI and peer interaction jointly predict their ToM at 3 years. BI was measured by observing children's behavior while they were introduced to novel stimuli. Children's peer interactions were measured by pairing each child with an unfamiliar peer; the two engaged in collaborative tasks and the child's negative behaviors (e.g., grabbing, demanding) were recorded. Among children with fewer negative peer interactions, BI at the age of 2 was unrelated to their ToM (essentially, false-belief understanding) at the age of 3. However, among children with more negative peer interactions, BI at the age of 2 predicted poorer ToM at the age of 3. Consistent with the findings of Harrist et al. (1997) and Lane et al. (2013), these latter children appear to fit the profile of disruptive, socially avoidant children who have poorer social cognition.

To summarize, a shy, reticent temperament appears to promote social cognitive development, perhaps because shy children spend more time observing others interact and reflecting upon what they observe. In contrast, withdrawal that reflects greater reactivity and that is characterized by more aggressive, disruptive behavior might undermine social-cognitive development by compromising the quality of the social interactions in which children engage and observe.

Future Directions

Research over the past several decades has focused heavily on social and emotional difficulties associated with BI and SW. Yet, recent studies suggest that there are also benefits associated with SW. For example, BI in infancy predicts better effortful control during the preschool years (e.g., Rothbart et al., 2011), and SW is associated with greater prosocial behavior during the elementary school years (Oh et al., 2008). Recently, research has documented more advanced, or at least precocious, social cognition among shy preschoolers compared to their nonshy peers. Such work should inspire additional research on cognitive, emotional, and social advantages associated with BI and SW. Given the importance of social cognition in our daily interactions with others, it is worthwhile for researchers to examine whether precocious social cognitive abilities translate to greater social competence for these shy children. Research is also needed to examine whether relations between shyness and competencies such as effortful control and social cognition are specific to the preschool years or whether there are more long-term advantages to having a shy temperament.

Studies on cultural variability in the prevalence and developmental outcomes associated with BI and SW have increased over the past decade. Yet, as is the case with much cross-cultural research, these studies are limited to a selected group of countries – namely, those in North America, Western Europe, and East Asia – that have the academic and research infrastructures

to easily support these sorts of investigations. This means that still relatively little is known about BI and SW within, for example, cultures in South America, Africa, and the Middle East. Research with other cultures would serve to further evaluate hypotheses regarding cultural influences on social development. A particularly interesting avenue for research involves tracking the development of BI and SW in countries that are undergoing sociocultural shifts and increasingly incorporating the values and norms of other cultures. Finally, more research is needed to examine the influence of demographic variables (e.g., parents' education, family income) on parents' perceptions of and reactions to children's behavior; this research would likely shed light on intracultural differences in the development of BI and SW.

See also: Cross-Cultural Psychology; Cultural Influences on Interpersonal Relationships; Neural Correlates of Social Development; Parenting Attitudes and Beliefs across Cultures; Personality, Trait Models of; Social and Emotional Development in the Context of the Family; Temperament Development, Theories of; Temperament and Human Development; Temperament.

Bibliography

Asendorpf, J.B., 1990. Beyond social withdrawal: shyness, unsociability, and peer avoidance. Human Development 33, 250–259. http://dx.doi.org/10.1159/000276522.

Blair, C., Peters, R., Granger, D., 2004. Physiological and neuropsychological correlates of approach/withdrawal tendencies in preschool: further examination of the behavioral inhibition system/behavioral activation system scales for young children. Developmental Psychobiology 45, 113–124. http://dx.doi.org/10.1002/dev.20022.

Bowker, J.C., Raja, R., 2011. Social withdrawal subtypes during early adolescence in India. Journal of Abnormal Child Psychology 39, 201–212. http://dx.doi.org/10.1007/s10802-010-9461-7.

Cheah, C.S.L., Park, S., 2006. South Korean mothers' beliefs regarding aggression and social withdrawal in preschoolers. Early Childhood Research Quarterly 21, 61–75. http://dx.doi.org/10.1016/j.ecresq.2006.01.004.

Chen, B., 2012. The association between self-reported mother–child attachment and social initiative and withdrawal in Chinese school-aged children. The Journal of Genetic Psychology: Research and Theory on Human Development 173, 279–301. http://dx.doi.org/10.1080/00221325.2011.609847.

Chen, X., French, D.C., 2008. Children's social competence in cultural context. Annual Review of Psychology 59, 591–616. http://dx.doi.org/10.1146/annurev.psych.59.103006.093606.

Chen, X., Wang, L., Cao, R., 2011. Shyness-sensitivity and unsociability in rural Chinese children: relations with social, school, and psychological adjustment. Child Development 82, 1531–1543. http://dx.doi.org/10.1111/j.1467-8624.2011.01616.x.

Coplan, R.J., Girardi, A., Findlay, L.C., Frohlick, S.L., 2007. Understanding solitude: young children's attitudes and responses toward hypothetical socially withdrawn peers. Social Development 16, 390–409. http://dx.doi.org/10.1111/j.1467-9507.2007.00390.x.

Coplan, R.J., Rose-Krasnor, L., Weeks, M., Kingsbury, A., Kingsbury, M., Bullock, A., 2013. Alone is a crowd: social motivations, social withdrawal, and socioemotional functioning in later childhood. Developmental Psychology 49, 861–875. http://dx.doi.org/10.1037/a0028861.

Coplan, R.J., Rubin, K.H., 2010. Social withdrawal and shyness in childhood: history, theories, definitions, and assessments. In: Rubin, K.H., Coplan, R.J. (Eds.), The Development of Shyness and Social Withdrawal. Guilford, New York, pp. 3–22.

Fox, N.A., Henderson, H.A., Marshall, P.J., Nichols, K.E., Ghera, M.M., 2005. Behavioral inhibition: linking biology and behavior within a developmental framework. Annual Review of Psychology 56, 235–262. http://dx.doi.org/10.1146/annurev.psych.55.090902.141532.

Harrist, A.W., Zaia, A.F., Bates, J.E., Dodge, K.A., Pettit, G.S., 1997. Subtypes of social withdrawal in early childhood: sociometric status and social-cognitive differences across four years. Child Development 68, 278–294. http://dx.doi.org/10.1111/j.1467-8624.1997.tb01940.x.

Hastings, P.D., Nuselovici, J.N., Rubin, K.H., Cheah, C.S.L., 2010. Shyness, parenting, and parent-child relationships. In: Rubin, K.H., Coplan, R.J. (Eds.), The Development of Shyness and Social Withdrawal. Guilford Press, New York.

Herrmann, E., Hare, B., Cissewski, J., Tomasello, M., 2011. A comparison of temperament in nonhuman apes and human infants. Developmental Science 14, 1393–1405. http://dx.doi.org/10.1111/j.1467-7687.2011.01082.x.

Kagan, J., 1997. Temperament and the reactions to unfamiliarity. Child Development 68, 139–143. http://dx.doi.org/10.1111/j.1467-8624.1997.tb01931.x.

Kagan, J., Snidman, N., 1991. Infant predictors of inhibited and uninhibited profiles. Psychological Science 2, 40–44. http://dx.doi.org/10.1111/j.1467-9280.1991.tb00094.x.

Karevold, E., Ystrom, E., Coplan, R.J., Sanson, A.V., Mathiesen, K.S., 2012. A prospective longitudinal study of shyness from infancy to adolescence: stability, age-related changes, and prediction of socio-emotional functioning. Journal of Abnormal Child Psychology 40, 1167–1177. http://dx.doi.org/10.1007/s10802-012-9635-6.

Kim, J., Rapee, R.M., Oh, K.J., Moon, H.-S., 2008. Retrospective report of social withdrawal during adolescence and current maladjustment in young adulthood: cross-cultural comparisons between Australian and South Korean students. Journal of Adolescence 31, 543–563. http://dx.doi.org/10.1016/j.adolescence.2007.10.011.

Lane, J.D., Wellman, H.M., Olson, S.L., Miller, A.L., Wang, L., Tardif, T., 2013. Relations between temperament and theory of mind development in the United States and China: Biological and behavioral correlates of preschoolers' false-belief understanding. Developmental Psychology 49, 825–836. http://dx.doi.org/10.1037/a0028825.

Lopez, N.L., Vazquez, D.M., Olson, S.L., 2004. An integrative approach to the neurophysiological substrates of social withdrawal and aggression. Development and Psychopathology 16, 69–93. http://dx.doi.org/10.1017/S0954579404044414.

Markovic, A., Rose-Krasnor, L., Coplan, R.J., 2013. Shy children's coping with a social conflict: the role of personality self-theories. Personality and Individual Differences 54, 64–69. http://dx.doi.org/10.1016/j.paid.2012.08.002.

Oh, W., Rubin, K.H., Bowker, J.C., Booth-LaForce, C., Rose-Krasno, L., Laursen, B., 2008. Trajectories of social withdrawal from middle childhood to early adolescence. Journal of Abnormal Child Psychology 36, 553–566. http://dx.doi.org/10.1007/s10802-007-9199-z.

Pérez-Edgar, K., Reeb-Sutherland, B.C., McDermott, J.M., White, L.K., Henderson, H.A., Degnan, K.A., Hane, A.A., Pine, D.S., Fox, N.A., 2011. Attention biases to threat link behavioral inhibition to social withdrawal over time in very young children. Journal of Abnormal Child Psychology 39, 885–895. http://dx.doi.org/10.1007/s10802-011-9495-5.

Rothbart, M.K., Sheese, B.E., Rueda, M.R., Posner, M.I., 2011. Developing mechanisms of self-regulation in early childhood. Emotion Review 3, 207–213. http://dx.doi.org/10.1177/1754073910387943.

Rubin, K.H., Burgess, K.B., Hastings, P.D., 2002. Stability and social-behavioral consequences of toddlers' inhibited temperament and parenting behaviors. Child Development 73, 483–495. http://dx.doi.org/10.1111/1467-8624.00419.

Rubin, K.H., Coplan, R.J., Bowker, J.C., 2009. Social withdrawal in childhood. Annual Review of Psychology 60, 141–171. http://dx.doi.org/10.1146/annurev.psych.60.110707.163642.

Suway, J.G., Degnan, K.A., Sussman, A.L., Fox, N.A., 2012. The relations among theory of mind, behavioral inhibition, and peer interactions in early childhood. Social Development 21, 331–342. http://dx.doi.org/10.1111/j.1467-9507.2011.00634.x.

Triandis, H.C., 1995. Individualism & Collectivism. Westview Press, Boulder, CO.

Wellman, H.M., Lane, J.D., LaBounty, J., Olson, S.L., 2011. Observant, nonaggressive temperament predicts theory of mind development. Developmental Science 14, 319–326. http://dx.doi.org/10.1111/j.1467-7687.2010.00977.x.

Wichmann, C., Coplan, R.J., Daniels, T., 2004. The social cognitions of socially withdrawn children. Social Development 13, 377–392. http://dx.doi.org/10.1111/j.1467-9507.2004.00273.x.

Behavioral Medicine

Claus Vögele, Institute for Health and Behaviour, Research Unit INSIDE, University of Luxembourg, Walferdange, Luxembourg

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Behavioral Medicine is the interdisciplinary field of study of behavior in health and disease. Based on often-experimental investigations of behavior, behavioral medicine contributes to a better understanding of etiological factors and mechanisms, but also to clinical applications aimed at systematically improving health in clinical and at-risk populations. As the literature summarized in this article illustrates, behavioral medicine has shown tremendous progress in achieving these goals since its inception in 1977. Nevertheless, the successful translation of research results into clinical practice remains a challenge for the future.

Definition and Conceptual Framework

Definition

In 1977, behavioral medicine was defined as "the interdisciplinary field concerned with the development and integration of behavioral and biomedical science knowledge and techniques relevant to health and illness and the application of this knowledge and these techniques to prevention, diagnosis, treatment, and rehabilitation" (Schwartz and Weiss, 1977, 1978). Over the last 30 years, the emergence of several overlapping disciplines may have contributed to a reorientation in the definition of behavioral medicine. The medical specialty of psychiatry adopted the term, often using it as a synonym for its traditional focus on psychopathology. An increase in the promotion and use of therapeutic approaches in complementary and alternative medicine, which is typically based on the interaction of biological, psychological, social, and spiritual influence on health, also involves behavioral approaches and expanded the boundaries of behavioral medicine.

A broader, and more current, definition comes from the charter of the International Society of Behavioral Medicine, which states that behavioral medicine is the "interdisciplinary field concerned with the development and integration of psychosocial, behavioral and biomedical knowledge relevant to health and illness and the application of this knowledge to prevention, etiology, diagnosis, treatment and rehabilitation" (International Society of Behavioral Medicine).

The primary targets of behavioral medicine are physical conditions, although there is increasing recognition of the importance of mental disorders if they are comorbid with physical conditions, whether they are involved in the development of the physical condition or are a consequence thereof (Evans et al., 2005; Prince et al., 2007).

The Biopsychosocial Model

The definition of behavioral medicine implies that it is based on a biopsychosocial model of health and disease. The biopsychosocial model is a general model positing that biological, psychological (which includes thoughts, emotions, and behaviors), and social (e.g., socioeconomical, socio-environmental, and cultural) factors, all play a significant role in health and disease. It follows, that health and disease are best understood in terms of a combination of biological, psychological, and social factors rather than purely in biological terms. This is in contrast to the biomedical model of medicine that suggests every disease process can be solely explained in terms of a deviation from normal function such as a physiological processes, infections, genes, developmental abnormalities, or injuries. In addition to behavioral medicine, the model is used in fields such as medicine, nursing, health psychology and sociology, and particularly in more specialist fields such as psychiatry, family therapy, clinical social work, and clinical psychology.

The model was developed by George L. Engel, a psychiatrist at the University of Rochester, and putatively discussed in an article published in 1977 in Science, where he posited "the need for a new medical model" (Engel, 1977). In an article published a few years later (Engel, 1980) he illustrates this model in detail by applying it to the fate of a hypothetical patient with coronary heart disease. In contrast to a classical biomedical approach, which focuses on the investigation of pathophysiological processes and the intervention at organ-, tissue-, and cell-level, biopsychosocial care would comprise a systematic (behavioral) analysis of typical personality traits and coping behaviors, which may impede admission to hospital in good time or impair medical treatment in any other way. For example, if personal ways of responding to a failed arterial puncture are not taken into consideration (e.g., fainting due to sympathetic hyperarousal), cardiac arrest may be the result. At the social level, the stabilizing role of partners and other forms of social support are also taken into account.

While behavioral medicine aims to include biopsychosocial aspects in research and treatment, there is often such a breadth of information at the individual (clinical) level that it is difficult to derive specific action strategies from it. Indeed, the biopsychosocial model has been criticized extensively for its lack of specificity that renders it untestable, according to some critics. Psychiatrist Niall McLaren summarizes his critique as follows: "Since the collapse of the 19th century models (psychoanalysis, biologism, and behaviorism), psychiatrists have been in search of a model that integrates the psyche and the soma. So keen has been their search that they embraced the so-called 'biopsychosocial model' without ever bothering to check its details. If, at any time over the last three decades, they had done so, they would have found it had none" (McLaren, 2006). In his view the biopsychosocial model should be viewed in its historical context as an alternative to

biological reductionism, which was (and still is) dominating the medical sciences.

The Evolution of Behavioral Medicine

Historical Development

Behavioral medicine emerged in the 1970s as an interdisciplinary field aimed at integrating knowledge from the behavioral and biomedical sciences to improve the understanding of biopsychosocial processes in the development of disorders and the maintenance of health. The term behavioral medicine was used first in the title of a book called 'Biofeedback: Behavioral Medicine' (Birk, 1973). This was predated by earlier experimental results by Neal E. Miller on instrumental learning in paralyzed rats, providing the basis for research investigating the effects of behavioral processes (e.g., learning) on autonomic processes. Although later attempts to replicate his main findings were unsuccessful, these results sparked the inquiry into behavioral, physiological, and biochemical interactions with health and illness, and eventually gained prominence under the rubric of behavioral medicine. In 1976, in recognition of this trend, the National Institutes of Health created the Behavioral Medicine Study Section to encourage and facilitate collaborative research across disciplines.

The 1977 Yale Conference on Behavioral Medicine and a meeting of the National Academy of Sciences were explicitly aimed at defining and delineating the field to guide future research. Based on outcomes of the Yale conference, Schwartz and Weiss proposed the biopsychosocial model, emphasizing the new field's interdisciplinary roots and calling for the integration of knowledge and techniques broadly derived from behavioral and biomedical science. Shortly thereafter, Pomerleau and Brady (1979) published a book, in which they offered an alternative definition focusing more closely on the particular contribution of the experimental analysis of behavior in shaping the field.

Additional developments during this period of growth included the establishment of learned societies (the Society of Behavioral Medicine and the Academy of Behavioral Medicine Research, both in 1978) and of journals (the Journal of Behavioral Medicine in 1977 and the Annals of Behavioral Medicine in 1979). In 1990, at the inaugural International Congress of Behavioral Medicine in Uppsala, Sweden, the International Society of Behavioral Medicine was founded to provide an international focus for professional and academic development, through its currently more than 20 daughter societies and through its own peer-reviewed journal (the International Journal of Behavioral Medicine).

Current State of the Art

Behavioral medicine has firmly established itself over the last three decades as an interdisciplinary field, integrating knowledge from the behavioral and biomedical sciences relevant to health and disease and the application of this knowledge to prevention, assessment, treatment, and rehabilitation. Since its inception in the 1970s it has seen enormous growth and diversification, with new disciplines such as genetics, life course epidemiology, and health economics, and new techniques such as neuroimaging and ambulatory assessments now contributing and adding to the more traditional approaches of behavioral medicine, such as the investigation of the role of personality for disease processes and purely laboratory based approaches. This development was paralleled by the emergence of new, global health concerns such as obesity and metabolic disorders that present new challenges for the integration of the behavioral sciences and public health (Steptoe, 2010).

The more traditional behavioral medicine approaches such as the laboratory based investigation of psychological factors and their effects on psychophysiological responses have been expanded to include the assessment of intracellular processes, cell signaling molecules, and gene expression. Psychophysiology, psychoneuroimmunology, and psychoneuroendocrinology have converged in the shared exploration of behavioral processes across a range of medical conditions. Assessment techniques have been further developed to now include sophisticated measures of ambulatory assessment such as ecological momentary assessment and the day reconstruction method enabling the investigation of behavioral processes in study participants' own environment thus contributing to better ecological validity. Behavioral medicine intervention and prevention research has embraced new information technologies such as the Internet, and new perspectives such as public health, while an increasing number of meta-analyses of intervention trials targeting a range of conditions and patient populations provide the basis for improved care through the integration of single trial results.

Defining the Context

More people than ever die today from preventable chronic diseases of the cardiovascular system, respiratory disorders, or metabolic conditions. Psychological and social factors play an important role in the etiology and the maintenance of these disorders. Mental disorders, which often accompany physical conditions and which have been shown to negatively affect their prognosis, account for the largest part of the global burden of disease. If the risk factors for these disorders were to be reduced through behavior change, life expectancy would rise by approx. 10 years across the globe. The following section gives a brief outline of the context in which behavioral medicine is placed and thus the challenges that it faces.

Morbidity and Mortality in the Twenty-First Century

Epidemiology

Ever since the end of the eighteenth century a significant change in the prevalence of certain disorders and causes of death in the populations of larger countries seems to have taken place. If only 150 years ago the main causes of death were related to infectious diseases, this has changed at the end of the twentieth and the beginning of the twenty-first century to non-communicable, chronic-degenerative disorders. For example, in the European Region of the World Health Organization (i.e., European member states, the former member states of the Soviet Union, Turkey, and Israel) 86% of all deaths and 77% of the total disease burden are attributable to chronic disorders, such as cardiovascular disorders, cancer, mental

problems, diabetes, lung disorders, and musculoskeletal disorders. Among these, cardiovascular disorders rank highest, with more than half of all deaths being accounted for by cardiac disease or stroke.

At least partly responsible for this development is the successful prevention and treatment of infectious diseases, which – together with better hygiene and nutrition – has led to an ever increasing life expectancy. This has resulted in a higher percentage of older people in the population, who are also more likely to suffer from age-associated chronic-degenerative disorders. Nevertheless, there are probably also other reasons, which have contributed to the increasing prevalence of chronic conditions.

Behavioral Risk Factors

Seven leading risk factors have been identified as contributing to almost 60% of the disease burden in the European Region: high blood pressure, tobacco smoking, alcohol, high cholesterol, obesity, a diet low in fruit and vegetables, and physical inactivity. For mortality, high blood pressure is the most important risk factor, while for disease burden it is tobacco smoking. Alcohol is the most important risk factor in terms of mortality and disease burden in younger age groups.

Willett (2002) estimated the percentage rates of four prevalent conditions that could be prevented by long-term behavior modification. According to his estimates more than 80% of type 2 diabetes cases could be avoided, and 80% of coronary heart disease cases; for colon cancer and stroke the corresponding percentage is 70% each. These figures mean that type 2 diabetes, coronary heart disease, colon cancer, and stroke would be avoidable in a substantial number of cases if people did not smoke, engaged in a minimum of 30 min per day of physical activity, were normal-weight, consumed fewer than three alcoholic drinks per day, ate red meat fewer than three times per week, and took folic acid in sufficient amounts.

These figures illustrate in an impressive way the effects of behavior (change) on the etiology and prognosis of physical disorders, including prevention. Ezzati et al. (2003) estimated for which conditions behavior change would have the biggest impact, and concluded that ischemic heart disease, depression, and stroke are ranking first among preventable conditions, particularly in developed countries. If risk factors for these disorders were significantly reduced life expectancy would rise on average by 9.3 years globally.

Consequences for Health Care and Social Systems

Chronic disorders pose a serious challenge for health care systems and consume substantial resources. For example, in Denmark it is estimated that currently 70–80% of health care costs are accounted for by chronic conditions, whereas in the United Kingdom 8 out of 11 hospital-admissions are due chronic disorders.

Nevertheless, the economic consequences of chronic disorders go far beyond the immediate costs for the health care systems. Indirect costs, e.g., losses in productivity due to ill-health are equally as high and sometimes even supersede direct health care costs. In Sweden, for example, over 90% of the total costs for musculoskeletal disorders are indirect costs: 31.5% are accounted for by absenteeism and 59% by early retirement.

Premature death of principal earners and skilled workers not only affects the income of private households but also the national economies of entire states. In 2005, for example, chronic disorders were responsible for a loss of 1% of the gross domestic product across the member states of the Russian Federation; according to these estimates stroke and diabetes produced a loss in income of 11 billion US dollars.

In the middle (e.g., Baltic Sates, Romania, Ukraine, Belarus) and eastern countries of the European Region (e.g., member states of the Russian Federation) people die at a much younger age from chronic disorders compared to Western countries. An additional problem arises due to the aging population in most countries. Currently it is estimated that in 2050 more than a quarter of the global population will be older than 65 years. Age-related multimorbidity is already the rule rather than the exception (Agborsangaya et al., 2012).

Socioeconomic Status and Health

Within countries chronic physical and mental disorders are much more prevalent in lower-income sections of the population. People with lower socioeconomic status (SES) often have twice the risk to fall seriously ill compared to people with higher SES. In addition, high-SES groups benefit more from health behavior change interventions (Marmot et al., 2012).

Psychological Factors and Physical Disorders

Concepts and Disciplines

The notion that psychological factors contribute to the development of physical disorders has a long tradition in history, and in the health sciences in the twentieth century. Based on mostly psychoanalytical concepts psychosomatic approaches stipulated a list of disorders where psychological factors were considered to be particularly important or even crucial for their etiology. This notion, however, has long been abandoned based on empirically supported evidence showing that psychological factors can play an important role for all disorders, contributing either to their etiology, their maintenance, or their prognosis. In disciplines such as health psychology and clinical psychology, or interdisciplinary fields such as behavioral medicine the dichotomy of 'psychogenic' versus 'somatogenic' has been given up in favor of an integrative understanding of psychological and physiological processes in the etiology, treatment, and prevention of chronic physical conditions.

Interactions

(1) Psychological factors can contribute to the etiology of disorders, (2) psychological problems can be the consequence of physical disorders, and (3) they can affect their prognosis.

1. *Causal role in the etiology*. Several possibilities, how psychological factors exert their influence on the etiology of physical disorders exist, and they are not mutually exclusive. First, psychological factors, problems, or disorders can have direct, pathophysiological consequences. For example, depression is associated with neurobiological changes, which increase the risk for the development of chronic pain syndromes (via changes in the serotonergic system; Covic et al., 2006), coronary heart disease (for example through sympathetic overactivity; Brotman et al., 2007), or breast

cancer (e.g., reduced immune function; Oerlemans et al., 2007). Second, psychological factors, problems or disorders as well as adverse socioeconomic circumstances can increase the risk for health-damaging behaviors, e.g., increases in tobacco and alcohol consumption, unhealthy dietary behavior, physical inactivity, and sexual risk taking. Third, the chronic physical condition and the psychological disorder have shared, but independent risk factors ('shared underlying etiology model'). Cerebrovascular changes, for example, are characteristic for depression and Alzheimer's Disease. The underlying process may be a long-term inflammatory mechanism, which is mediated by immune cells. Fourth, the (medical) treatment of a mental disorder may increase the risk for physical conditions through toxic side effects, as is the case, for example, for antidepressants, which contribute to weight gain and have cardiovascular side effects.

2. *Psychological problems are the consequence of the physical disorder*. First, some physical conditions affect the brain directly, as is the case, for example, in some infectious disorders (e.g., HIV), cerebrovascular conditions (e.g., stroke because of high blood pressure or atherosclerosis), basal ganglia disease (e.g., Parkinson's Disease), lung disorders (e.g., neurocognitive abnormalities because of lack of oxygen in chronic obstructive pulmonary disease). Second, receiving the diagnosis of an incurable, chronic condition is a highly stressful event for most patients, in particular if the condition is life threatening or requires significant lifestyle changes. This can increase the risk for mental disorders, for example adjustment disorder in HIV, hypoglycemia anxiety in type 1 diabetes or panic disorder with agoraphobia in chronic obstructive pulmonary disease (Vögele and von Leupoldt, 2008). Third, psychological problems can arise as a consequence of treatment side effects. Certain drug treatments, for example, can contribute to anxiety disorders and depression, as is the case for pain medication (e.g., NRSA), antihypertensive medication, cell division inhibitors (cytostatic drugs), and corticosteroids (cortisone).
3. *Psychological factors affect the prognosis of the physical condition*. Once a physical condition has been established, the future development of this disorder is significantly affected by the patient's beliefs or explanatory model, the patient's self-management of the disorder (e.g., medication adherence), risk behaviors (e.g., tobacco smoking, alcohol consumption, dietary behavior), stress and comorbid mental disorders. There is a large body of literature demonstrating the negative impact of psychological problems or mental disorders on the prognosis of physical conditions. For example, depression doubles mortality risk in patients with coronary heart disease, and this effect is independent of other risk factors such as obesity, tobacco smoking, blood lipids, diabetes, and high blood pressure (Barth et al., 2004).

Methods and Applications in Behavioral Medicine

As an integrative discipline, behavioral medicine involves diverse methodologies and research paradigms that converge on health and disease. Major health problems such as cardiovascular disorders, cancer, lung disorders, obesity, and HIV are, therefore, explored from multiple perspectives. Rather than providing a list of disease-oriented sections, the following gives a brief summary of methods and applications with a focus on clinical practice (for a comprehensive review see Steptoe, 2010).

Assessment

Assessment in behavioral medicine encompasses such diverse approaches as the conceptualization and measurement of health behaviors (e.g., dietary behavior, physical activity, sexual behavior, treatment adherence), self-reported well-being and other psychological processes and outcomes (e.g., mood and negative affect, lay representations of illness and treatment, neurocognitive outcomes), social and interpersonal processes (e.g., social interactions, social support, social networks), socioeconomic and sociodemographic variables, genetic and biological markers (e.g., candidate gene and genome-wide association studies, laboratory stress testing, neuroendocrine measures, noninvasive assessment of autonomic function), and brain function and neuroimaging. Even statistical procedures (e.g., structural equation modeling, mediation analyses) and forms of data reporting (e.g., meta-analyses) have recently been discussed with a focus on behavioral medicine and the specific challenges posed by the wide range of different research paradigms and measures (see Steptoe, 2010).

While assessments in research will depend on the research questions asked, and may in principle cover the entire range of possible methods outlined in the previous paragraph, assessments in clinical practice are often limited due to practical reasons and ethical considerations. The most frequently employed assessment methods in clinical behavioral medicine practice are questionnaire-based self-report measures on mood, well-being, and health behaviors; semistructured and structured interviews for the assessment of mental disorders (e.g., Structured Clinical Interview for DSM Disorders); and routine medical data provided in patient records. Despite increasing recognition among clinicians of the importance of assessing psychological factors in patients with physical conditions, the systematic evaluation and application of disease- and patient-oriented assessment strategies in clinical behavioral medicine practice is still some way off.

The following section gives a brief summary of the most important elements of clinical assessment in behavioral medicine.

Lay Representations of Illness and Treatment

Lay representations of illness and treatment (or subjective explanatory models) are of crucial importance for the course of disease, the patient's way of coping with the disorder, the patient's medical adherence, and their limitations in a range of life areas. Even patients with the same disorder and comparable disease status, can differ in their view of treatability, prognosis, limitations in life, or need for retirement. Important elements of lay representations of illness and treatment are the patient's personal view of the causes of the disorder, which physical symptoms are signs of the disorder, what the course of the disorder will be like, which limitations are associated with the disorder, and possible ways of taking control of the course of

the disorder. An important question concerns the patient's expectation of the degree of his or her own role for improvement or cure. If coping strategies are recognized on the part of the patient, this generally has a positive effect on recovery, independent of their actual efficacy.

For example, lay representations of illness and treatment predict to a significant degree, if and when coronary patients return to work (Petrie et al., 1996). Also for neurological disorders such as multiple sclerosis, patients' explanatory models are the most important predictors of social and psychological impairments, exhaustion, and self-esteem (Jopson and Moss-Morris, 2003). Interventions designed to optimize patients' illness representations have been shown to significantly benefit recovery after myocardial infarction (MI; Petrie et al., 2002).

Cognitive Styles

In addition to subjective explanatory models, other cognitive processes, such as catastrophizing or trivialization of physical symptoms, can impact treatment outcome or prognosis. In a longitudinal study on cardiac patients, for example, 'threat minimization' and 'search for affiliation' assessed only days after a MI both predicted lower depression scores 2 and 6 months after the MI, respectively, suggesting an important prognostic role for cognitive appraisal processes in the early period after a cardiac event (Vögele et al., 2012). The assessment of cognitive styles of coping with the disorder is, therefore, important and should be taken into account when planning interventions (Van Damme et al., 2004).

Coping and Adherence to Treatment

Ways of coping with the experience of (chronic) physical illness range from denial, to collaboration, and constant worrying. These different coping styles evidently affect illness behavior, and, therefore, the course of the illness and quality of life. A particularly sensitive and important area affected by coping styles concerns adherence to treatment, e.g., taking medication reliably, cooperation with invasive medical procedures (chemotherapy, injections etc.), and regular checkups. These areas are, therefore, very important to include in the assessment and monitoring of the patient over the course of treatment.

Avoidance

Avoidance behaviors are a frequent form of illness behavior, which often interfere with treatment goals and, therefore, overall improvement. This is particularly true for chronic pain conditions where avoidance of, e.g., painful movements all-too-often replace the initial source of pain, as, for example, in musculoskeletal disorders involving back pain. It is, therefore, necessary to assess avoidance behavior in the beginning of the therapeutic process, as it is important to motivate the patient to accept a possible, but transient deterioration in physical symptoms in the short-term, in favor of a mid- and long-term improvement.

Resources

Even patients with chronic physical conditions that involve serious limitations (e.g., amyotrophic lateral sclerosis, cancer at an advanced stage) can experience good quality of life, dependent on a number of environmental factors (e.g., good social support, adequate care) and psychological strengths (e.g., self-regulation strategies). Following Antonovsky's approach of salutogenesis (Antonovsky, 1987) and also current conceptualizations of health (Jadad and O'Grady, 2008) such resources are crucial for patients' well-being and quality of life, and should, therefore, be part of any clinical assessment with the aim to integrate these resources into individual behavioral medicine treatment formulations.

Interventions

Treatment of Comorbid Mental Disorders

Distress, demoralization, and comorbid mental disorders such as anxiety and depression negatively affect the prognosis of physical conditions, either via direct, physiological consequences or indirectly, through behaviors such as treatment adherence or health-risk behaviors, or both. If depressed, cardiac patients, for example, have a significantly worse prognosis than nondepressed cardiac patients (Barth et al., 2004). Nevertheless, mental disorders not only deserve adequate treatment because of their negative impact on the physical condition; mental disorders must be addressed in their own right, given their negative impact on patients' quality of life.

1. *Clinical-psychological interventions*. The clinical psychological treatment of mental disorders comorbid with physical conditions principally encompasses all interventions, which are empirically supported with regards to their efficacy and are, therefore, currently used in the treatment of those with mental disorders. Cognitive behavioral techniques are demonstrably among the most successful intervention methods in the treatment of depression and anxiety disorders, i.e., the most frequent comorbid mental conditions. Nevertheless, cognitive behavior therapy approaches need to be frequently adapted to the specific conditions that are determined by the physical disorder. For example, physical symptoms usually considered being typical symptoms for a panic attack (e.g., light headedness, shortness of breath, racing heart, sweating) can also be a consequence of the physical disease, i.e., not only as a result of catastrophizing cognitions. In the cognitive reframing of physical symptoms, patients are, therefore, required to learn to differentiate between symptoms that occur unrelated to external causes (i.e., panic attack) from those that are associated with an acute deterioration of their physical condition (e.g., stronger shortness of breath because of an infection, trembling and sweating because of hypoglycemia). An important role concerns the identification of unrealistic, catastrophizing cognitions (e.g., "Now I am going to die," "Because of my shortness of breath, I am going to instantly suffocate"). The aim of modified cognitive behavior therapy in the treatment of comorbid mental disorders in patients with chronic physical conditions is to enable the patient to better assess how realistic such cognitions are, in particular in relation to the severity of the disorder (e.g., blood glucose awareness training for diabetic patients with hypoglycemia anxiety; Cox et al., 2006), and/or to emotionally disengage from them (mindfulness-approaches in the treatment of anxiety disorders and depression in chronic obstructive pulmonary disease; Fjorback et al., 2011).

2. *Drug treatment*. Comorbid mental disorders in patients with physical conditions are also treated using psychopharmacological interventions, however with mixed success. The more successful drugs include Serotonin Reuptake Inhibitors, in particular in the treatment of depression in Parkinson's Disease, coronary heart disease, HIV, and rheumatoid arthritis. Nevertheless, drug treatments necessitate the consideration of specific side effects (e.g., cardiovascular side effects, weight gain, etc.). In addition, after terminating drug treatment, symptoms often reoccur. In some cases, for example in the treatment of depression in diabetes, drug treatment and clinical-psychological interventions are combined with the aim to initially improve mood and increase drive and motivation through medication to enable the patient to fully benefit from cognitive behavior therapy. The antidepressive medication can then be reduced in a systematic fashion to prevent relapse.

Health Promotion and Support in Disease Management

1. *Stress management programs*. Stress management programs positively affect immune parameters, neuroendocrine responses, and further recovery outcomes. They are, for example, successfully employed in the treatment of people with HIV (Antoni, 2003) and cancer (Phillips et al., 2008). Their success, however, is related to the professional quality of their implementation. Stress management programs are usually conducted in a group format and combine any of the following elements: relaxation trainings (e.g., progressive muscle relaxation), cognitive reframing (changing dysfunctional cognitive styles, adjustment of realistic treatment goals), education (effects of stress on immune function, drug effects, and side effects), and communication training (anger management, assertiveness training). The aim of these programs is the improvement of self-efficacy and perceived control, and the acquisition of functional coping strategies as well as the training of strategies to improve social support.
2. *Patient education programs*. Patient education programs are conceptually closely linked with stress management programs. They were inspired initially by patient education programs for diabetics (Shelley et al., 2004), and then adapted for other chronic physical conditions, for example cardiovascular disorders (Dusseldorp et al., 1999). Central to these programs is not only information and education, but also the empowerment of patients. Techniques used to achieve these therapeutic goals include principles of motivational interviewing, as they have been used in health promotion programs for other chronic conditions such as drug and alcohol dependence.
3. *Relaxation trainings*. While relaxation techniques include a range of practices, e.g., progressive muscle relaxation, guided imagery, self-hypnosis, and deep breathing exercises, they all have the aim of inducing a state of relaxation characterized by slow breathing, lower blood pressure and heart rate, and a feeling of calm and well-being. As is the case for stress management programs, the success of relaxation trainings depends to a large extent on the quality of their implementation. Factors that have been shown to affect the efficacy of relaxation trainings include the structuredness of their implementation, the professional competency of the trainer, and flexibility in adapting the training to the specific requirements of the patient. They are successfully implemented in the psychological treatment of chronic pain, the reduction of side effects of invasive medical procedures, and in the treatment of neuropsychological deficits (for a meta-analytical review see Carlson and Hoyle, 1993).
4. *Biofeedback*. Biofeedback aims at gaining greater awareness of and, thereby, learning to influence physiological functions that are usually not under voluntary control (e.g., heart rate, blood pressure, brain waves, muscle tension). This is achieved by monitoring these physiological processes with psychophysiological equipment and feeding the activity back to the participant using graphic displays on a computer screen or auditory information (e.g., tones). Recent meta-analyses demonstrate the success of biofeedback in the noninvasive treatment of a range of chronic physical conditions, such as migraine, tension headache, chronic back pain, hypertension, incontinence, epilepsy, and tinnitus (for an overview see Schwartz and Andrasik, 2003). The Cochrane Library alone currently lists 41 entries on meta-analyses with the term biofeedback appearing in their title, abstract, or keywords. Nevertheless, like for many behavioral medicine interventions, the future challenge is to further systematize biofeedback implementation and to develop guidelines for its use in clinical practice for reasons of quality assurance.
5. *Interventions to reduce instrumentally reinforced problem behavior*. The aim of this behavior therapy approach is to reduce problem behavior (e.g., avoidance behavior in chronic pain or after stroke) by interrupting the reinforcement chain, which leads to avoidance of everyday activities and problem behavior. Positive and negative reinforcers are identified in a functional assessment to change these contingencies using contingency management techniques. Frequently the collaboration of carers, family, or friends is required as they often, inadvertently reinforce the patient's problem behavior. Particularly successful areas for this approach include the treatment of chronic pain, some neurological conditions (Parkinson's Disease, stroke, acquired brain injury; Alderman and Wood, 2013) and substance misuse.

Conclusion

Behavioral medicine is the interdisciplinary field concerned with the integration of knowledge from the behavioral and biomedical sciences relevant to health and disease, and the application of this knowledge to prevention, assessment, treatment, and rehabilitation. Since behavioral medicine's inception more than 30 years ago, enormous progress has been made in the understanding of biopsychosocial interactions. The inclusion of new techniques has expanded to the assessment of intracellular processes, cell signaling molecules, and gene expression. Assessment techniques now include sophisticated measures of ambulatory assessment such as ecological momentary assessment and the day reconstruction method enabling the investigation of behavioral processes in study participants' own environment thus contributing to better

ecological validity. Behavioral medicine intervention and prevention research has embraced new information technologies such as the Internet, and new perspectives such as public health, while an increasing number of meta-analyses of intervention trials targeting a range of conditions and patient populations provide the basis for improved care through the integration of single trial results. This development, however, was paralleled by the emergence of new health challenges such as HIV, obesity, metabolic disorders, and chronic obstructive pulmonary disease, and a growing recognition of the importance of mental problems and mental disorders comorbid with physical conditions. In contrast to this progress in scientific knowledge, the translation of study results into clinical practice is lacking. It remains a challenge for the future to develop treatment and assessment guidelines and to systematically evaluate their effectiveness thereby contributing to the continuing optimization of patient care.

See also: Cancer: Psychosocial Aspects; Cardiac Disease, Coping with; Chronic Illness, Psychosocial Coping with; Co-morbidity of Mental and Physical Conditions; Coronary Heart Disease: Psychosocial Aspects; Depression, Pessimism, and Health; Diabetes: Psychosocial Aspects; Explanatory Style and Health; HIV Risk Interventions; Health Behaviors, Assessment of; Health Behaviors; Health Psychology; Health Risk Appraisal and Optimistic Bias; Hypertension: Psychosocial Aspects; Mindfulness; Patient Adherence to Health Care Regimens; Personalized Medicine; Psychoneuroimmunology; Respiratory Disorders: Psychosocial Aspects; Rheumatoid Arthritis: Psychosocial Aspects; Sexual Risk Behaviors; Smoking and Health; Social Support and Recovery from Disease and Medical Procedures; Socioeconomic Status and Health; Stress, Coping and Health.

Bibliography

Agborsangaya, C.B., Lau, D., Lahtinen, M., Cooke, T., Johnson, J.A., 2012. Multimorbidity prevalence and patterns across socioeconomic determinants: a cross-sectional survey. BMC Public Health 12, 201. http://dx.doi.org/10.1186/1471-2458-12-201.

Alderman, N., Wood, R.L., 2013. Neurobehavioural approaches to the rehabilitation of challenging behaviour. NeuroRehabilitation 32, 761–770.

Antoni, M.H., 2003. Stress management effects on psychological, endocrinological, and immune functioning in men with HIV infection: empirical support for a psychoneuroimmunological model. Stress 6, 173–188.

Antonovsky, A., 1987. Unraveling the Mystery of Health. Jossey-Bass Inc, San Francisco.

Barth, J., Schumacher, M., Herrmann-Lingen, C., 2004. Depression as a risk factor for mortality in patients with coronary heart disease: a meta-analysis. Psychosomatic Medicine 66, 802–813.

Birk, L. (Ed.), 1973. Biofeedback: Behavioral Medicine. Grune & Stratton, New York.

Brotman, D.J., Golden, S.H., Wittstein, I.S., 2007. The cardiovascular toll of stress. Lancet 370, 1089–1100.

Carlson, C.R., Hoyle, R.H., 1993. Efficacy of abbreviated progressive muscle relaxation training: a quantitative review of behavioral medicine research. Journal of Consulting and Clinical Psychology 61, 1059–1067.

Covic, T., Tyson, G., Spencer, D., Howe, G., 2006. Depression in rheumatoid arthritis: demographic, clinical and psychological predictors. Journal of Psychosomatic Research 60, 469–476.

Cox, D.J., Gonder-Frederick, L., Ritterband, L., Patel, K., Schächinger, H., Fehm-Wolfsdorf, G., Hermanns, N., Snoek, F., Zrebiec, J., Polonsky, W., Schlundt, D., Kovatchev, B., Clarke, W., 2006. Blood glucose awareness training: what is it, where is it, and where is it going? Diabetes Spectrum 19, 43–49.

Dusseldorp, E., van Elderen, T., Maes, S., Meulman, J., Kraaij, V., 1999. A meta-analysis of psychoeduational programs for coronary heart disease patients. Health Psychology 18, 506–519.

Engel, G.L., 1977. The need for a new medical model: a challenge for biomedicine. Science 196, 129–136.

Engel, G.L., 1980. The clinical application of the biopsychosocial model. American Journal of Psychiatry 137, 535–544.

Evans, D.L., et al., 2005. Mood disorders in the medically ill: scientific review and recommendations. Biological Psychiatry 58, 175–189.

Ezzati, M., Hoorn, S.V., Rodgers, A., Lopez, A.D., Mathers, C.D., Murray, C.J.L., 2003. Estimates of global and regional potential health gains from reducing multiple major risk factors. The Lancet 362, 271–280.

Fjorback, L.O., Arendt, M., Ørnbøl, E., Fink, P., Walach, H., 2011. Mindfulness-based stress reduction and mindfulness-based cognitive therapy – a systematic review of randomized controlled trials. Acta Psychiatrica Scandinavica 124, 102–119.

International Society of Behavioral Medicine. www.isbm.info/html/charter.html (accessed 09 08.14.).

Jadad, A.R., O'Grady, L., 2008. How should health be defined? British Medical Journal 337, a2900.

Jopson, N.M., Moss-Morris, R., 2003. The role of illness severity and illness representations in adjusting to multiple sclerosis. Journal of Psychosomatic Research 54, 503–511.

Marmot, M., Allen, J., Bell, R., Bloomer, E., Goldblatt, P., 2012. WHO European review of social determinants of health and the health divide. Lancet 380, 1011–1029.

McLaren, N., 2006. Interactive dualism as a partial solution to the mind-brain problem for psychiatry. Medical Hypotheses 66 (6), 1165–1173.

Oerlemans, M.E.J., van den Akker, M., Schuurman, A.G., Kellen, E., Buntinx, F., 2007. A meta-analysis on depression and subsequent cancer risk. Clinical Practice and Epidemiology in Mental Health 3, 29. http://www.cpementalhealth.com/content/3/1/29 (accessed 03.03.08.).

Petrie, K.J., Weinman, J., Sharpe, N., Buckley, J., 1996. Role of patients' view of their illness in predicting return to work and functioning after myocardial infarction. British Medical Journal 312, 1191–1194.

Petrie, K.J., Cameron, L.D., Ellis, C.J., Buick, D., Weinman, J., 2002. Changing illness perceptions after myocardial infarction: an early intervention randomized controlled trial. Psychosomatic Medicine 64, 580–586.

Phillips, K.M., Antoni, M.H., Lechner, S.C., Blomberg, B.B., Llabre, M.M., Avisar, E., Glück, S., DerHagopian, R., Carver, C.S., 2008. Stress management intervention reduces serum cortisol and increases relaxation during treatment for nonmetastatic breast cancer. Psychosomatic Medicine 70, 1044–1049.

Pomerleau, O.F., Brady, J.P. (Eds.), 1979. Behavioral Medicine: Theory and Practice. Williams & Wilkins, Baltimore.

Prince, M., Patel, V., Saxena, S., Maj, M., Maselko, J., Phillips, M.R., Rahman, A., 2007. No health without mental health. The Lancet 370, 859–877.

Schwartz, G.E., Weiss, S.M., 1977. What is behavioral medicine? Psychosomatic Medicine 39, 377–381.

Schwartz, G.E., Weiss, S.M., 1978. Behavioral medicine revisited: an amended definition. Journal of Behavioral Medicine 1 (3), 249–251.

Schwartz, M.S., Andrasik, F. (Eds.), 2003. Biofeedback. A Practitioner's Guide, third ed. Guilford Press, New York.

Shelley, E.E., Speroff, T., Dittus, R.S., Brown, A., Pichert, J.W., Elasy, T.A., 2004. Diabetes patient education: a meta-analysis and meta-regression. Patient Education and Counseling 52, 97–105.

Steptoe, A. (Ed.), 2010. Handbook of Behavioral Medicine. Springer, New York.

Van Damme, S., Crombez, G., Eccleston, C., 2004. Disengagement from pain: the role of catastrophic thinking about pain. Pain 107, 70–76.

Vögele, C., von Leupoldt, A., 2008. Mental disorders in chronic obstructive pulmonary disease. Respiratory Medicine 102, 764–773.

Vögele, C., Christ, O., Spaderna, H., 2012. Cardiac threat appraisal and depression after first myocardial infarction. Frontiers in Psychology – Psychology for Clinical Settings (Special Issue on Psychocardiology) 3, 365.

Willett, W.C., 2002. Balancing life-style and genomics research for disease prevention. Science 296, 695–698.

Relevant Websites

http://www.thecochranelibrary.com.
http://www.isbm.info/.
http://www.sbm.org/.
http://www.scid4.org/.

Behavioral Problems, Effects of Parenting and Family Structure on

Helena R Slobodskaya, Novosibirsk State University, Novosibirsk, Russia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Parental behavior aims to provide care and socialize the child according to cultural norms and expectations, and has an independent effect on the child's behavior. Research has provided evidence that parental hostility/rejection, lack of warmth/support, and failure to provide adequate control is associated with behavioral problems such as aggression, defiance, and antisocial behavior. However, the underlying causal mechanisms are still in doubt. Contemporary models of the development of behavioral problems emphasize reciprocal parent–child interactions; family structure has an indirect effect on behavioral problems through parenting, family relationships, and socioeconomic factors. Evidence from behavioral genetics and remaining questions are presented.

Behavioral problems (destructiveness, aggression, defiance, rule breaking, stealing, and vandalism) are also described as externalizing problems. Disruptive behavior is the most common form of child and adolescent maladjustment; moreover, early and persistent behavioral problems are associated with multiple adverse outcomes in later life, including compromised mental and physical health, impaired educational attainment, unemployment, poor social relationships, substance abuse, and criminality, and create a substantial burden to society in terms of suffering and financial cost.

Although different sorts of behavioral problems are interrelated, there is accumulating evidence for distinct subtypes (Tremblay, 2010). It is useful to distinguish between overt (physical aggression, temper tantrums, opposition, and defiance) and covert (rule breaking, stealing and vandalism) behavioral problems. Physical aggression includes behaviors directed toward another individual with the intent to cause physical damage, e.g., hitting, kicking, pushing, beating, etc. Overt nondestructive behaviors include the following: losing temper, disobeying adults, openly refusing to comply with rules, annoying and blaming others. Covert behavioral problems include antisocial acts such as truanting from school, running away from home, use of alcohol and drugs, stealing, burglary, and destruction of property.

Development of Behavioral Problems

It is important to recognize developmentally normative level of behavioral problems and to distinguish it from an abnormal behavior. Longitudinal community studies have shown that overt behavioral problems tend to increase until the second or third year of life and peak between 2 and 4 years (Tremblay, 2010). Thus angry, aggressive, oppositional, noncompliant, and impulsive behaviors are common in toddlerhood and so are often considered developmentally normative, although this period is described as 'the terrible twos.' After the preschool years, overt behavioral problems decrease steadily to the end of adolescence and throughout adulthood, as children learn to use socially acceptable behavior, some considerably better than others.

Covert behavioral problems appear later than overt problems; they are very rare before middle childhood and increase with age until about 17 years. Thus antisocial behaviors (rule breaking, running away, substance abuse, vandalism, and theft) are very common in adolescents, suggesting that adolescence-limited delinquency is normative rather than abnormal. Most adolescents desist from antisocial behavior in young adulthood – possibly because they have learned that it leads to punishment, – and become increasingly motivated in fulfilling adult role expectations (Moffitt, 1993).

There are substantial differences in frequency and seriousness of behavioral problems and these differences appear remarkably stable from early childhood onward. Longitudinal studies show that children who display aggressive, oppositional, noncompliant, and impulsive behaviors most frequently in toddlerhood, will continue to do it most frequently throughout childhood. Furthermore, the small group of children (around 7%) who continue to show high levels of behavioral problems during middle childhood, are those most at risk of aggression, violence, and antisocial behavior during adolescence, and later in life too. For these children, behavioral problems remain stable (lifetime persistent) and lead to more serious maladaptive outcomes (Moffitt, 1993). Therefore, psychiatric classifications in the late 1970s recognized 'Conduct Disorder' as a medical condition, and nowadays it is one of the commonest psychiatric disorders in children and adolescents.

It is important to explain why and how some young children develop persistent behavioral problems whereas others develop normally. There is good evidence that early development of behavioral problems has its origin in the child's continuous interactions with the immediate environment, particularly with their primary caregivers.

The Diversity of Family Structures

Family is a domestic group of close people and the principal institution for childrearing. There is tremendous diversity of family structures both within and between cultures and over time (Georgas et al., 2006; Roopnarine and Gielen, 2005). In recent decades most countries have seen significant structural changes in families as a result of sociodemographic changes and globalization, and these changes are similar across traditional, developing, and developed countries. Worldwide, birth rates have been declining steadily, reducing the number

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23047-X

of children in the family. This is accompanied by a decrease in marriage rates and increases in rates of single parenthood, divorce, separation, remarriage, and cohabitation.

Family structure around the world has been gradually changing from the traditional extended and multigenerational families that include parents, children, and other relatives, to more nuclear families comprising parents and children. However, the conventional nuclear family consisting of two married parents (the working father and the housekeeper mother) and their children has become less common in the developed world. More children are born to unmarried parents or a single mother and they may experience changes in family structure throughout childhood and adolescence. These processes, along with a higher proportion of working mothers and elderly people, lead to a wide variety of family arrangements in which nuclear families, including single-parent, cohabiting, reconstituted, nonresidential-father, and same-sex families, are embedded in a network of extended kin that serve the purpose of childrearing (Georgas et al., 2006).

Although rates of specific family structures differ markedly across cultures and ethnicities, marriage still has a key role in family formation, and women are the primary child caregivers worldwide (Georgas et al., 2006; Roopnarine and Gielen, 2005). Recent surveys (e.g., Chapple, 2009) reported the highest rates of two-parent families in Asian, Middle Eastern, and Southern European countries that are strongly influenced by traditional and religious values (more than 80% in China, India, Turkey, Spain, and Greece) and the lowest rates in America, Australia, and Northern Europe (less than 60% in the United States and around two-thirds in Denmark and Iceland). Rates also vary by ethnicity, e.g., in the United States in 2010, approximately 75% of non-Hispanic white children and 35% of black children aged below 17 years, lived with two parents.

Overall, in developed countries, about three-quarters of the children live with both parents, about 15% with a single parent, about 9% in a stepfamily, and 2% in other arrangements. Although point estimates suggest that most children live in two-parent families, life course estimates indicate that more than half of all children will spend some time in another form of the family (two-fifths in a cohabiting family). In the Nordic countries the rates of cohabiting families are particularly high (45% of all births in Sweden), and this is accompanied by institutionalization.

Lone-parent families are increasing around the world and 80–90% of them are headed by women. In the developed countries, the rise is mainly due to increases in divorce, separation, and the number of single mothers who never married or cohabited; in the developing countries paternal labor migration, imprisonment, or death are more important. However, for many children raised in lone-parent families, nonresident fathers or other father figures may play a prominent role. Although labor migration may alienate fathers from their children (Roopnarine and Gielen, 2005), limited evidence suggests increases in paternal involvement over recent decades.

The Concept of Parenting

Parenting refers to childrearing behavior, as well as related beliefs, knowledge, values, and attitudes (Bornstein, 2002). The primary parenting goals involve providing appropriate care and socializing the child according to cultural norms and expectations. Parental attitudes may emerge long before the child's birth, but after birth, parenting develops in reciprocal interaction between child and caregiver in the wider sociocultural context.

Therefore, parenting varies markedly with the child's age, gender, and other physical and behavioral characteristics, and also across different caregivers, and from culture to culture. Parental personality plays an important role and behavioral genetic studies suggest a genetic influence. Parental mental health, family relationships, socioeconomic status, social network, and other factors are also important. Parental attitudes, values, and beliefs are likely to influence parental behavior, which, in turn, has an independent effect on the children. However, parental attitudes may differ significantly from actual behaviors; for example, parents who are under stress may treat their children in ways that they consider undesirable or inappropriate. In extreme cases, adverse parenting may take the form of child maltreatment, such as abuse and neglect.

Decades of research and clinical studies have led to multidimensional and multilevel concepts of parenting. Several typologies of parenting have been suggested and a large number of specific dimensions have been identified. However, at present there is no widely accepted theory of parenting (O'Connor, 2002) and no well-validated system for classifying parental behaviors. For research and applied purposes, parenting has been assessed by direct observations, interviews, or questionnaires. Observations are made in the playroom or in the home, are expensive and time consuming, and may not be appropriate with older children and also for some behaviors (e.g., corporal punishment). Standardized questionnaires or rating scales completed by parents and older children can be useful in community surveys. Despite a large number of measures developed over recent decades (Locke and Prinz, 2002), there is still a need for assessment techniques that might be relevant to the origin and development of behavioral problems and can be used with families from diverse backgrounds.

Dimensions of Parenting

Two main dimensions of warmth and control have consistently emerged in many studies across childhood and adolescence and across cultures and methods of measurement. The warmth dimension reflects the affective component of parenting, characterized by sensitivity, responsiveness, involvement, and support. There is some evidence that it is better to consider warmth/support and hostility/rejection as separate dimensions, since they are conceptually different and uniquely related to the child's behavior.

The control dimension reflects parental efforts to influence the child's behavior, such as enforcing rules, applying discipline, and supervision. Different types of parental control have very different implications for child development. Schaefer in the 1960s and later Steinberg distinguished between psychological control and behavioral control. Psychological control aims to influence the child's feelings and thoughts by verbal disapproval, love withdrawal, conditional regard, and guilt induction. Behavioral control, in contrast, aims to influence the

child's behavior by setting limits, enforcing rules, and monitoring maturity demands.

Parental disciplinary practices have received considerable attention both in research and practice. Positive discipline involves induction, i.e., reasoning, explanation and reminding of rules, and reinforcement of desired behavior through praise and rewards. Punishment is used to reduce undesired behavior through unpleasant consequences, such as expression of disappointment, verbal criticism, shaming, the raised voice, withdrawal of privileges, 'time out,' or physical punishment. Physical punishment ranges from the mild occasional light smack, to controlled conditional or customary spanking, to severe corporal punishment (e.g., hitting or beating), to physical abuse.

Some other dimensions of parenting have been widely used in research and practice. Coercive control refers to parental efforts to influence the child by use of force, physical manipulation, or commands. Harsh parenting involves coercive control, corporal punishment, and verbal aggression (e.g., yelling or scolding). The structural dimension concerns parental efforts to create a predictable and organized environment for the child, enforce house routines and a stable set of rules, supervision, and consistency in discipline.

Beside the dimensional approach, much research on parenting has used categorical classifications proposed by Baumrind in late 1960s and then modified by Maccoby and Martin in 1983. Based on the levels of two main dimensions of warmth and control, four parenting styles have been identified: authoritative (high warmth, high control), authoritarian (low warmth, high control), indulgent/permissive (high warmth, low control), and neglectful/uninvolved (low warmth, low control). It was supposed that parenting style conveyed the parents' general attitude toward the child and the emotional atmosphere in which parenting behavior occurred. However, because of doubts about the validity of parenting typologies, the focus of recent research moves toward specific parenting behaviors and cognitions and reciprocal interaction patterns between child and caregiver (Bornstein, 2002). Beyond the traditional parenting research exploring the role of mothers, an increasing number of studies consider fathers and other caregivers.

Associations between Parenting and Behavioral Problems

Research has provided abundant evidence that behavior problems in children and adolescents are related to the parenting they receive in their families. Studies have particularly emphasized the role of ineffective or negative parenting, such as hostility/rejection, lack of warmth/support, and failure to provide adequate control. By contrast, responsiveness, involvement, support, and positive control, such as approval, inductive reasoning, appropriate limit setting, firm and consistent discipline, and monitoring is associated with fewer behavioral problems.

Lack of supervision, and harsh and inconsistent discipline have all been linked to aggression, delinquency, and other disruptive behaviors. Neglect and extremely harsh, abusive parenting practices significantly increase the risk of aggression, criminality, and violence. There is continuing controversy over the effects of physical or corporal punishment. Historically, it has been widely used and accepted as a disciplinary method; one meta-analysis provided evidence for negative long-term consequences. However, compared with other disciplinary practices, controlled conditional spanking was most effective for reducing child noncompliance or antisocial behavior; physical punishment led to worse outcomes only when it was too severe or was the primary disciplinary method. Harsh verbal discipline such as yelling or scolding is uniquely related to child problem behavior, as well as psychological control. The link between inconsistent discipline and behavioral problems appears to be direct and independent of child's effects. Although parental monitoring has been linked to less delinquency, recent evidence indicates that this association reflects parental knowledge due to adolescent free disclosure and not to active parenting behavior.

Numerous studies have shown that an authoritative parenting style is associated with less behavioral problems, while authoritarian, permissive, and neglectful parenting styles can all contribute to the child's disruptiveness, aggression, delinquency, and antisocial behavior. There is some evidence that a psychologically controlling parenting style is associated with delinquency and with relational social and indirect aggression. Overall, the main findings on parenting and behavioral problems have been confirmed across contexts, ages, and cultures, although the cultural context may have an important role (Bornstein, 2002).

Theoretical Framework

Several theoretical models have been proposed to account for the association between parenting and behavioral problems. For much of the twentieth century, psychoanalytic and learning theories have been particularly influential. Psychoanalytic theories emphasized the importance of caregiving experiences received in early childhood, but provided inadequate explanatory models of development. Early learning theories have been formulated in the behaviorist tradition suggesting that all behaviors are learned and that parents increase desired behavior through rewards and reduce undesired behavior through punishments.

Social learning theory, developed by Bandura in the 1960s, suggests that the child can learn behavior through imitation by observing other people and emphasizes the importance of parents as role models. According to social learning theory, children acquire disruptive behavior such as aggression through direct modeling when parents use harsh verbal discipline, corporal punishment, and abusive practices. Learning theories considered the child as a passive recipient of parenting influences, but Bell in 1968 pointed out that children had effects on their parents, just as parents had effects on their children. Understanding of reciprocal parent–child interactions then led to the elaboration of transactional models (Dodge and Pettit, 2003).

The Coercion model, developed by Patterson in the 1980s, suggests that children's disruptive behaviors tend to elicit verbal aggression and physical punishment and if parents do not encourage appropriate behavior and do not consistently apply discipline, children's disruptive behavior may escalate because this enables them to get more attention, avoid

unpleasant demands, or get their own way. Thus parents may unintentionally reinforce tantrums, noncompliance, aggression, and antisocial behaviors, which in turn lead to less parental involvement, warmth and consistency, and harsher discipline. An escalating cycle of coercive parent–child interactions is supposed to be the main cause of high trajectories of behavioral problems from early childhood to adulthood, and the empirical evidence has supported this causal mechanism.

There are other possible mechanisms through which parenting is linked to child behavioral problems. Internalization of social norms is enhanced by positive parenting, which fosters attachment, provides inductive reasoning, and promotes autonomy and intrinsic motivation for good behavior; and it is compromised by harsh parenting which prevents the formation of secure attachment, provides strong pressure for compliance, fosters an external locus of control, and precludes the development of internal attributions. Cognitive processing of social information may also be important, so that children who have experienced hostile and harsh parenting are more likely to selectively attend to hostile cues, attribute hostile intent to others, access aggressive responses, and believe that violence and antisocial behavior is beneficial. Parenting may also influence self-regulatory capacities, such as effortful control and emotional regulation and the underlying physiology.

Contemporary models of the development of behavioral problems suggest that a specific combination of early temperament and parenting may lead to hostile coercive exchanges from toddlerhood onward. Dadds and Salmon (2003) emphasize interaction between parental disciplinary practices and the child's sensitivity to punishment. A dynamic cascade model designed by Dodge with colleagues, suggests that biologically based difficulties in impulse control and behavioral regulation initiate the chain of development of antisocial behavior from early childhood through adolescence, in which parents, family, and the wider environment play the crucial roles.

Effects of the Direction of Parenting-Behavioral Problem

Though the links between parenting and behavioral problems in children are well established, the effects are typically mild or moderate and the underlying causal mechanisms are still in doubt. Several research approaches have been suggested for studying the effects of parenting (O'Connor, 2002). Prospective studies showed that parenting predicts subsequent behavioral problems in children and adolescents, even when the initial level of the problem is taken into account. Overall, warm and responsive parenting has been linked with decreasing behavioral problems, whereas harsh parenting predicted smaller decrease in physical aggression in schoolchildren and increasing delinquency in adolescents.

More recent studies have demonstrated that change in parenting over time is associated with subsequent change in behavioral problems. For example, decrease in parental support, increase in rejection, and inconsistent discipline predicted increase in the child's behavioral problems. Sequential analyses of parent–child interactions have shown that parental use of coercive control and corporal punishment increases the likelihood of disruptive behavior in the child. These findings provide stronger support for the effect of parenting on behavioral problems but the evidence is indirect.

Experimental interventions designed to enhance parental effectiveness and to improve behavioral problems have the possibility of proving causal effects. Parenting interventions promote mutually rewarding activities, and teach parents to reward good behavior, to provide clear instructions, and consistently apply discipline by a firm and calm response such as 'time out.' Numerous randomized controlled trials have shown the effectiveness of parent training for childhood conduct disorders. Limited evidence from community-based studies suggests that altering parental behavior can lead to changes in the child's behavioral problems. For example, in one study of recently divorced mothers and their school-age sons, intervention produced benefits to coercive discipline and positive parenting, which in turn prevented noncompliance from growing over 2.5 years, in contrast to a control group. Although change in parenting behavior leads to improved child behavior, this does not necessarily mean that poor parenting is the cause of behavioral problems, just as symptomatic medication alleviates physical illness but it does not cause the disease itself.

The Role of Genetic Factors

Over recent decades, behavioral genetic studies provided evidence of an important role for genetic factors in both parenting and children's behavioral problems. At the same time, twin and adoption studies have shown that the effects of shared family environment are often weak or absent. These findings have led to the conclusion that parenting has little effect on child development. Although most studies showed a greater contribution from shared environment to disruptive behavior in the child, it has been argued that the associations between parenting and behavioral problems were almost entirely due to the genetic transmission. Developmental research addressed these misconceptions (Collins et al., 2000) and designed approaches that examine the causal relations and the processes involved (Moffitt, 2005).

Meta-analyses suggest that genetic factors account for about for 40–50% of individual differences in antisocial behavior and the contribution of shared environment is about 15–20%; the rest of the variance is accounted for by environmental influences that are not shared by children in the same family and by measurement error. The smaller effect of shared family environment as compared to nonshared, does not mean that family influences are unimportant; it only means that they tend to make siblings more different than similar to each other. The influence of shared family environment may be even larger because estimates of genetic influence include not only the direct effect of genes, but also the indirect effects of environmental factors. There is limited evidence that genetic effects may vary depending on the type of behavioral problems and age: genetic contribution to physical aggression in children from 1.5 through 6 years was higher (around 60%) and more stable than genetic contribution to disregard for rules (around 30%). Individual differences in parenting behaviors are also influenced by genes, with estimates ranging from 25 to 60%.

Genetic factors may have a role in linking parenting to children's behavior problems. Genetic transmission can explain up to half of the association between bad parenting and child aggression (Moffitt, 2005). Genetic child effects are also important. For example, children's genes accounted for most of the correlations between corporal punishment and children's aggression and between negative parenting and adolescent antisocial behavior, indicating that children's behavioral difficulties may evoke negative parenting and corporal punishment, but the reverse does not apply.

Nonshared environmental influences that make children growing up in the same family different have received much research attention. While early behavioral genetic studies estimated environmental effects as a residual, more recent studies used direct measures of family environment. For example, the 'Nonshared Environment in Adolescent Development' project used multiple measures of parenting and adolescent antisocial behavior longitudinally in identical and fraternal twins, full siblings, half-siblings, and genetically unrelated siblings. Studies using the full multivariate genetic design found significant, albeit small, nonshared environmental effects of parenting, such as parent–child mutuality and conflict, parental negativity, criticism, corporal punishment, and physical abuse, on children's behavioral problems, independent of genes and shared environment. A recent review concludes that important sources of nonshared environment may be epigenetics and children's differential experiences. Recent twin and adoption studies suggest that stability of children's behavioral problems is mainly accounted for by genetic (60%) and shared environmental influences (34%); the effect of nonshared environment is much smaller (6%).

Interactive Effect of Genes and Parenting on Behavioral Problems

Although developmental researchers have long been aware that genes and environment constantly interact, it was difficult to detect statistically significant interactions. Recent findings suggest that parents of children at genetic risk for disruptive behavior can effectively reduce the level of behavioral problems if they provide warmth, support, and positive control, and suppress negative and harsh parenting. For example, one study showed that parental monitoring and supervision protected adolescents from the genetic risk for conduct disorder.

Animal studies suggest that early parenting can influence gene expression and lead to a stable and lasting change in genetic effects. In addition, early parenting may affect biological programming of brain development, cognitive and emotional processing, stress responses, and social interactions. These mechanisms suggest that parenting effects may be transmitted intergenerationally; therefore, preventive interventions for serious behavioral problems should begin earlier than is current practice – before a child is born (Tremblay, 2010) – and target high-risk parents and the larger family.

Interactions between Parenting and the Child's Temperament/Personality

In the 1950s Thomas and Chess proposed that goodness-of-fit between parenting and children's individual characteristics plays a key role in the emergence of behavioral problems. Although this idea lacked empirical validation for many years, recent studies provide clear evidence that similar parenting may have different effects on children with different temperaments/personalities. Many studies have focused on difficult temperament characterized by irritability, low adaptability, and negative mood. The evidence suggests that the combination of a child's difficult temperament and poor parenting carries a particularly high risk for behavioral problems, while temperament, parenting, and family structure in themselves may have little effect. However, the findings were not entirely consistent – possibly because the concept of difficult temperament was not well defined.

Other studies focused on specific dimensions of parenting and specific temperament/personality traits and found a number of interactive effects on behavioral problems. In many instances, interactions corresponded to a dual-risk pattern whereby certain traits amplified the negative effects of parenting. For example, harsh and hostile mothering was a stronger predictor of behavioral problems in irritable children than in placid ones. Many studies have shown that children with low effortful control, high impulsivity, low conscientiousness, and low agreeableness are more affected by parenting than children with better self-control. By contrast, high effortful control protected children from the negative effects of maternal hostility, inconsistent discipline, and physical punishment. Thus certain temperaments/personalities can protect children against the negative effects of poor parenting.

Belsky's differential susceptibility hypothesis suggests that children may differ in the degree to which rearing environment may affect their development. Pluess and Belsky (2010) concluded that children who seem most susceptible to poor parenting may also benefit the most from developmentally supportive rearing across infancy, toddlerhood, and later childhood. For example, children high in negative emotionality were most vulnerable to the effects of unsupportive, rejecting, or intrusive parenting, but also benefited most from intervention programs.

Other interaction patterns were described by O'Connor and Dvorak in 2001. There is limited evidence that some children develop problems only if they receive ineffective parenting, whereas other children have high levels of conduct problems irrespective of the kind of parenting they received. One such trait is child psychopathy that includes low fear, callousness, and constricted emotionality. More generally, specific parenting dimensions and child temperament/personality traits operated differently, depending on the particular combination. In one study, physical punishment predicted increases in behavioral problems in boys high in anxiousness and low in frustration, but had the opposite effect in boys low in anxiousness and high in frustration.

Family Structure and Children's Behavioral Problems

Contextual models of child development particularly emphasize family effects on behavioral problems. Because an increasing proportion of children around the world live with the lone mother, the father's presence in the home has received considerable attention. Recent meta-analyses indicate that

children growing up in two-parent families tend to do better across a range of domains than children raised by a lone parent and the effects continue from the childhood and teenage years into adulthood (Chapple, 2009). The largest effect, as estimated by standardized mean difference, was found for behavioral problems (around 0.30).

Although these differences are well documented, the direction of effects is still unclear. A lone parent family may have a causal role in initiating and maintaining child behavioral problems, the influence may be reverse, and both lone parenthood and child behavioral problems may reflect the operation of a third factor, whether psychosocial or genetic. Several research designs are useful for estimating causal effects, however the evidence is not conclusive yet. One study found that children benefited from a father's presence only if he exhibited low levels of antisocial behavior; when fathers were antisocial, the longer they lived with their children, the more behavioral problems their children had. Overall, the effect of the father's presence on child behavioral problems depends on the father's support, involvement, parenting practice, and also on family relationships and socioeconomic factors.

Divorce may be followed by aggressive and antisocial behavior, but this is more due to prior and continuing marital discord than to separation from one of the parents. Generally, consequences of divorce tend to wane within a few years, the remaining negative effects result from a decrease in available resources and the quality of parenting, especially supervision. There is some evidence that each family transition carries additional risk, so that children are far more likely to have behavioral problems when they experience two or more parental divorces and/or remarriages. Although nationwide surveys consistently show that children in stepfamilies and single-parent families have on average higher probabilities of behavioral problems than children growing up with two married biological parents, the effect is small and varies markedly. The risk is mostly accounted for by socioeconomic factors, psychological stress, and family relationships, not by family structure per se.

Gender Differences

After the age of 4, behavioral problems are more common in males than females; the differences are most pronounced for physical aggression. Studies do not show marked gender differences in early childhood, therefore, different socialization of boys and girls is thought to play a key role. Socialization models emphasize that from birth, parents differentially encourage and punish boys' and girls' behaviors. Although there is no strong evidence for gender differences in parenting during infancy, in childhood boys tend to receive less warmth, less monitoring, harsher discipline, and more corporal punishment than girls. However, these differences may be accounted for by child effects. Overall, associations between parenting and behavioral problems do not differ between boys and girls; however, recent research suggests that interactive effects of temperament such as effortful control, may have a role.

There are clear differences between mothers' and fathers' parenting in many cultures. Mothers are more involved in caregiving in the early years of life, while fathers' involvement with children has been more limited until fairly recently. Research evidence suggests that children's behavioral problems are more closely associated with maternal than with paternal parenting; however, the underlying causal mechanisms are far from certain. Evidence is accumulating that both the child's and the parent's gender may be important. Studies show greater parental involvement with same-gender children, and recent meta-analysis found that paternal parenting was more likely to differ by child gender than maternal parenting. Researchers have hypothesized that parenting and behavioral problems are more closely associated in same- than in opposite-gender parent–child dyads, because children tend to imitate a same-gender model more, but the evidence is contradictory. Meta-analyses do not support the suggestion that boys from single-mother families are more likely than girls to have behavioral problems because of the absence of male role model, but they do show that in the 1990s, divorce had stronger effects on boys than girls.

Age Differences

Both parenting and the child's behavior changes with age; thus there may be developmental changes in the effects of parenting and family structure on behavioral problems. Studies show that consistent with a cumulative reciprocal model of parent–child influence, the association between parenting and behavioral problems is larger in older children than in toddlers and preschoolers; however, the link between corporal punishment and aggression is most pronounced in school-age children and monitoring becomes increasingly important in adolescence.

Parental age may also be important. Adolescent mothers tend to be more physically intrusive and punitive, and their children have been shown to be at higher risk of adverse outcomes, although recent evidence suggests that these effects are due to the third factors rather than early childbearing. The role of grandparents is controversial. Although they may provide help and support for teenage mothers, several western studies have shown that the multigenerational family structure may have negative effect on adolescent mothers' parenting; findings for children's behavioral problems are inconclusive.

Behavioral Problems in Siblings and Differential Parenting

Many studies have shown that siblings may influence a child's behavioral problems such as aggression, delinquency, and conduct disorder, independent of other risk factors. Within-family differences between siblings are also important. While earlier parenting research included only one child per family, newer approaches tend to focus on sibling differences. One study found that about two-thirds of the total variance in parenting is within families. Differential parenting refers to differences in parental treatment of siblings or to differences in siblings' perceptions of the parenting they receive. Research has shown that parents do treat their children differently, for example, in the 'Nonshared Environment in Adolescent Development' study, between-sibling correlation of parenting was 0.70 for parent reports, 0.25 for child reports, and 0.20 for observations.

Differential parenting is partly explained by sibling differences in gender, age, temperament, and disruptive behavior.

Family structure is also important, with differential parenting being much more common in stepfamilies than other family forms, for both mothers and fathers. One study showed that for mothers, biological relatedness is strongly associated with differential parenting, independent of the child's gender, age, or aggressive/disruptive behavior. Recent evidence suggests that differential parental treatment is related to child and adolescent behavioral problems beyond the absolute level of parenting.

The association between differential parenting and behavioral problems may vary depending on gender, birth order, and sibling relationship. Thus, one study found that in the association between differential parental warmth and adolescent delinquency 1 year later (but only for younger children in same-sex dyads), disfavored boys were reported to have higher rates of vandalism and violence, and disfavored girls were reported to have higher rates of theft; poor sibling relationships enhanced these effects. The role of differential parenting has been explained in terms of the social comparison model and the distributive justice framework, suggesting that siblings actively shape their experiences. Genetically informative studies showed that differential parenting primarily reflects genetically influenced differences between siblings; and differential effects of parenting on behavioral problems are due to genotype–environment correlation. This implies that both family and child effects have a role to play in the development of behavioral problems.

Conclusion

Current research on parenting, family structure and behavioral problems is increasingly benefiting from advances in genetics, neurobiology, psychology, and social science. Contemporary conceptual models emphasize bidirectional processes in parent–child interactions, the role of the family and wider sociocultural context and developmental trajectories. Newer research designs and multivariate analyses of longitudinal data are useful for testing causal influences. However, there is a lack of comprehensive measurement framework for examining the development of parenting and behavioral problems from early childhood through adolescence. There is also a need for comprehensive classification of family structures that applies to diverse backgrounds. It may be valuable to include positive parenting and to examine protective family influences in future research. Finally, a growing understanding of the effects of parenting and family structure on behavioral problems will benefit from cross-cultural research.

See also: Antisocial Behavior; Behavioral Problems, Effects of Parenting and Family Structure On; Defiant Behavior During Adolescence across Cultures; Defiant Behavior During Infancy and Early Childhood.

Bibliography

Bornstein, Marc H. (Ed.), 2002. Handbook of Parenting, second ed., vols. 1–5. Lawrence Erlbaum, Mahwah, NJ.

Chapple, Simon, 2009. Child Well-Being and Sole-Parent Family Structure in the OECD: An Analysis. OECD Social Employment and Migration Working Papers No. 82. Organisation for Economic Co-operation and Development Publishing, Paris.

Collins, W. Andrew, Maccoby, Eleanor E., Steinberg, Laurence, Hetherington, E. Mavis, Bornstein, Marc H., 2000. Contemporary research on parenting: the case for nature and nurture. American Psychologist 55, 218–232.

Dadds, Mark R., Salmon, Karen, 2003. Punishment insensitivity and parenting: temperament and learning as interacting risks for antisocial behavior. Clinical Child and Family Psychology Review 6, 69–86.

Dodge, Kenneth A., Pettit, Gregory S., 2003. A biopsychosocial model of the development of chronic conduct problems in adolescence. Developmental Psychology 39, 349–371.

Georgas, James, Berry, John W., van de Vijver, Fons J.R., Kagitcibasi, Cigdem, Poortinga, Ype H. (Eds.), 2006. Families across Cultures: A 30-Nation Psychological Study. Cambridge University Press, Cambridge, UK.

Locke, Lisa M., Prinz, Ronald J., 2002. Measurement of parental discipline and nurturance. Clinical Psychology Review 22, 895–929.

Moffitt, Terrie E., 1993. "Life-course-persistent" and "adolescence-limited" antisocial behavior: a developmental taxonomy. Psychological Review 100, 674–701.

Moffitt, Terrie E., 2005. The new look of behavioral genetics in developmental psychopathology: gene–environment interplay in antisocial behaviors. Psychological Bulletin 131, 533–554.

O'Connor, Brian P., Dvorak, Troy, 2001. Conditional associations between parental behavior and adolescent problems: a search for personality-environment interactions. Journal of Research in Personality 35, 1–26.

O'Connor, Thomas G., 2002. Annotation: the effects of parenting reconsidered: findings, challenges, and applications. Journal of Child Psychology and Psychiatry 43, 555–572.

Pluess, Michael, Belsky, Jay, 2010. Children's differential susceptibility to effects of parenting. Family Science 1, 14–25.

Roopnarine, Jaipaul L., Gielen, Uwe P. (Eds.), 2005. Families in Global Perspective. Allyn & Bacon, Boston, MA.

Tremblay, Richard E., 2010. Developmental origins of disruptive behaviour problems: the original sin hypothesis, epigenetics and their consequences for prevention. Journal of Child Psychology and Psychiatry 51, 341–367.

Behavioral Problems, Effects of Socio-Economic Status on

Patrick J McGrath, IWK Health Centre and Capital District Health Authority, Dalhousie University, Halifax, NS, Canada
Frank J Elgar, Institute for Health and Social Policy, McGill University, Montreal, QC, Canada

Published by Elsevier Ltd.

Abstract

The association between socioeconomic status (SES) and behavior problems in children is long established and well accepted. Across levels of SES, behavioral problems are generally more common among lower-SES children. However, less is understood about both the mechanisms that underlie these differences and their developmental origins. This article reviews the research regarding material and psychosocial paths that account for socioeconomic differences in child behavioral problems and how these differences are reinforced by social contextual influences within the family, school, and community. The article ends with a discussion about implications of SES differences for clinical care and policy interventions.

What Is Socioeconomic Status?

Socioeconomic status (SES) is a concept that summarizes an individual's social class position in society. Several conceptualizations of SES have been used, including material measures of income, occupation, and education as well as gradient-focused measures of relative status, relative deprivation, or class differences (American Psychological Association, Task Force on Socioeconomic Status, 2007). Although absolute levels of affluence (e.g., household income) are among the most commonly used measures of SES in research, over the past decade there has been increasing emphasis on the ability of social class gradients and perceived SES to better predict various social and behavioral health outcomes (Adler and Snibbe, 2003).

SES in children is typically measured using data on family income, parental occupation, parental education, or a combination of these factors. Although no standard measurement method exists, some methods are more commonly used than others. For example, the Hollingshead Index (Hollingshead, 1975), which was devised initially in the 1940s, is a four-factor index that comprises occupation, education, sex, and marital status. Although widely criticized, the occupational and education scales remain in widespread use. The classification of occupations may be made less useful by the proliferation of job titles that bear little resemblance to job status or income. Annual family income is often used in research; however, this factor may shift dramatically from year to year depending on employment status, overtime hours worked as well as changes in government income transfers and tax credits.

Similar challenges exist in gradient measures of SES. Although assessments of subjective SES share a robust relation to child behavioral symptoms, they also vary over time and context, are influenced by emotional styles, personality traits, or sense of personal control, and depend on how the individual evaluates social position (e.g., based on education, income, housing, or meeting one's potential). Researchers have now begun to explore the hypothesis that relative deprivation, measured in terms of individual income (or a proxy for it) relative to one's reference group, may relate more closely to child behavioral symptoms than absolute difference in income (Elgar et al., 2013). Despite the sensitivity of relative deprivation to both SES and the distribution of affluence in the reference group, currently no consensus exists about the appropriate formulation of relative deprivation and reference group characteristics (Reagan et al., 2007).

Childhood Behavioral Problems

Behavioral problems in children are usually categorized into internalizing and externalizing problems. Externalizing behavioral problems manifest outwardly as aggression, impulsivity, coercion, and noncompliance. Internalizing behavior problems are described as inward occurrences, displaying as an inhibited style described as withdrawn, lonely, depressed, and anxious. Comorbidity occurs both within and across these problems. Accordingly, children with aggression may also be suffering from anxiety or depression. In addition, externalizing and internalizing problems underlie various health and social problems that are more likely to initiate clinical intervention (e.g., physical fighting, body dysmorphism, self-harm).

Why Is SES Important in Relation to Childhood Behavioral Problems?

SES positively relates to behavioral health in every age-group and social context in which it has been studied. The relation is monotonic – meaning that at every level of SES, health and well-being are usually better at the level above and poorer at the level below. Youths from higher SES backgrounds exhibit fewer internalizing and externalizing problems, fewer social skills deficits, and higher life satisfaction (Tremblay, 1999). Because the gradient is found in countries with and without universal access to health services and the effect of SES does not plateau, even among middle- and upper-class families, SES differences in behavioral problems are mostly a psychosocial phenomenon and not attributable to absolute deprivation or poverty (Wilkinson and Marmot, 2003).

The association between SES and behavioral problems in children has received some attention in the epidemiological literature using cross-sectional, cohort, and longitudinal

studies. A recent systematic review by Reiss (2013) examined studies published in English or German between 1990 and 2011 that included at least one measure of SES and identified mental health problems in children and youth using validated measures. Reiss (2013) found that 52 of the 55 studies included in her sample showed an inverse relation between SES and mental health problems. Furthermore, children from families with lower SES were about two to three times more likely to have mental health problems than children from more advantaged families. Although there was some variability, low SES was more strongly associated with externalizing behavior disorders, such as opposition defiant disorder or attention-deficit hyperactivity disorder, than with internalizing disorders such as anxiety and depression. Additionally, SES related more closely to problems in early childhood than in adolescence. Gender differences were not clearly evident. Based on these findings, Reiss (2013) highlighted the need for early childhood interventions to reduce mental health problems in families with low SES.

Singh and Ghandour (2012) found that the odds of exhibiting a serious behavior disorder in children increased 1.9 times if the child lived in an unsafe neighborhood with poor housing, garbage in the streets, and high rates of vandalism; 1.9 times if the parents had less than high school education; and 3.7 times if the child lived in poverty (**Figure 1**). The authors concluded that improvements in neighborhood conditions and household income could reduce child behavioral problems.

The relation between SES and child behavior problems was also illustrated by Vine et al. (2012), who examined SES and anxiety in a cohort study of 498 U.S. middle school children in the age-group of 11–13 years. Annual household income and neighborhood median income were used to measure SES. Anxiety was measured using the four anxiety subscale scores from the Multidimensional Anxiety Scale for Children (MASC) (i.e., physical symptoms, harm avoidance, social anxiety, and separation/panic anxiety). The results were complex and showed that household income negatively related to scores on two of the four MASC subscales – physical symptoms and separation/panic anxiety. In contrast, adolescents living in higher-income neighborhoods with the same level of household income displayed higher physical and harm avoidance symptom scores. This study highlights the need for multiple measures of SES, as simple studies correlating just one measure of SES with global measures of problems can miss potentially important relationships.

Why Are Behavior Problems in Children Important?

According to the National Research Council and Institute of Medicine (O'Connell et al., 2009), about 17% of children and youth suffer from diagnosable mental health problems, with internalizing and externalizing behavior problems being the most common at a prevalence rate of 13 and 11%, respectively.

Behavior problems in children may impact adolescence and adulthood. Although some children recover, many develop chronic and debilitating mental health problems in adulthood, encounter academic problems in school, have difficulty finding long-term employment, and become involved in criminal activity (e.g., Sourander et al., 2005, 2008).

What Are the Mechanisms by Which SES Influences Behavioral Problems?

There are many potential mechanisms by which SES might influence the occurrence of behavioral problems. The influence is complex and challenging to untangle. Developmental research has traced the origins of SES differences in behavioral problems to prenatal exposures to maternal stress; early childhood experiences; and developmental processes that shape neuroendocrine responses to stress and neuroregulatory systems in the brain that govern emotion, attention, and social interactions (Repetti et al., 2002).

Epigenetics, or changes in gene expression due to environmental influences, has been investigated to explain the

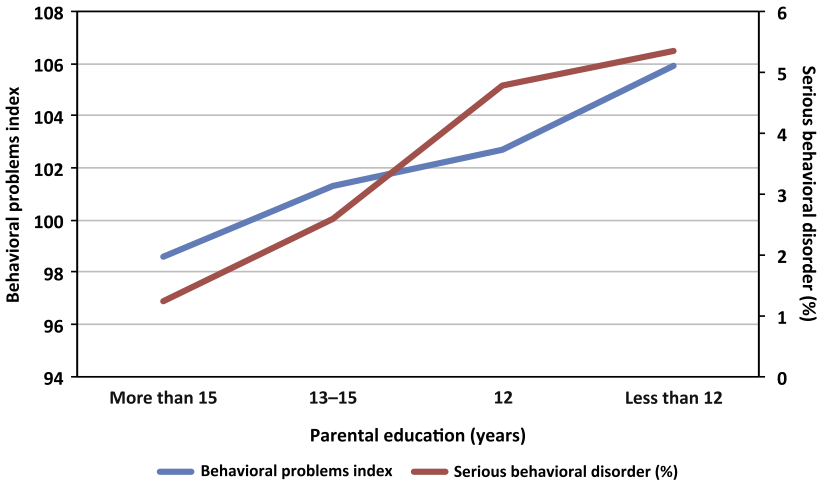

Figure 1 Social gradient in behavioral problems (blue line) and prevalence of a serious behavioral disorder (red line) in 91 642 American children. Adapted from data published by Singh, G.K., Ghandour, R.M., 2012. Impact of neighborhood social conditions and household socioeconomic status on behavioral problems among US children. Maternal and Child Health Journal 16 (1), S158–S169.

persistent genomic embedding of early-life experiences, most commonly, socioeconomic stress (Borghol et al., 2012). Mechanisms such as increased cortisol output and inflammatory responses are being proposed and tested (Lam et al., 2012). However, serious methodological challenges in this area abound, and all results must therefore be regarded as preliminary (Heijmans and Mill, 2012).

Research into the causal mechanisms that underlie SES differences in behavioral problems focuses on the direct consequences of material deprivation and the indirect psychosocial consequences of relative socioeconomic position (Adler and Snibbe, 2003). The material path involves an unequal distribution of material resources that can be used to support health (e.g., access to mental health services). In contrast, the psychosocial path involves the notion that *feeling poor* in comparison with others elicits class anxiety and psychological stress, erodes social supports, and thus contributes to increased mental distress (Wilkinson and Marmot, 2001). Although these mechanisms are interrelated, only the psychosocial path explains why children at the same level of household income or parental education might differ in their mental health when one is surrounded by more affluent people and the other by less affluent people (Vine et al., 2012). What differentiate these children are feelings of deprivation from a desirable standard of living that is established by society (Sweet, 2011).

Thus, SES relates to child behavioral problems in part through upward social comparisons of social class. Epidemiological studies of income inequality have found that societies with larger SES differences between the rich and poor (and therefore more relative deprivation) have higher prevalence rates of mental illness and lower well-being in children and adolescents, as indicated by rates of teenage pregnancy (Pickett et al., 2005), school bullying (Elgar et al., 2009), alcohol misuse (Elgar et al., 2005), and school dropout (Pickett and Wilkinson, 2007).

In addition, a third variable might influence both SES and behavioral problems. Genetics, for example, might influence both parental and child behavior and link SES to behavior problems. For example, in a study of 1069 11-year-old twins and their biological mothers and fathers, Bornovalova et al. (2010) demonstrated that parental genetic factors accounted for a large percentage of the variance in disorders in these children.

Some variables may produce intergenerational effects (Yoshikawa et al., 2012). For example, low levels of education and adolescent parenthood are associated with low SES, and these may perpetuate poverty, resulting in effects on behavior problems generations later.

Several other contextual variables exist that may serve to mediate SES with their effect on behavior problems, including abuse; neglect; exposure to violence; poor nutrition; exposure to toxins such as lead; poor neighborhood environment; inadequate and cramped housing; less cognitive stimulation; chaotic home environment; poorer access to good schools; harsh or inconsistent parenting style; parental stress, including depression, anxiety, and isolation; and poor access to effective interventions. All of these variables are more common exposures among lower-SES children.

Although SES is an individual or family attribute, the environments in which low-SES families live differ from those environments inhabited by advantaged families. Poor neighborhoods are more likely to be violent, congested, littered and with inadequate housing stock that is more likely to have toxins, more crowded, and in poor repair. Poor neighborhoods have fewer social resources, poorer schools, and fewer safe play and green spaces. These factors impact on the rate of behavior problems (Singh and Ghandour, 2012) and may also exert an indirect effect via parental behavior (e.g., harsh parenting, low monitoring) or other mechanisms.

What Are the Implications for Research, Clinical Activities, and Policy?

Recent, more sophisticated studies have found that the relation between SES and behavior problems is nuanced and may depend on several individual and contextual factors, including the measure of SES used, subtype of behavior problem examined, child's age, and ethnic or national factors. Thus, in the future, studies may need to take into account the child's context and the specific type of behavior problem. Interdisciplinary collaboration and increasing levels of methodological expertise in this research are critical components. Standardization of measures via guidelines and consensus across the research community might prove effective.

It is also important to consider policy approaches for reducing relative poverty. The corrosive social impacts of economic inequality suggest that children in areas of larger SES differences – not just from low-SES families or deprived neighborhoods – are more vulnerable and a legitimate focus of community and school-based interventions. However, it is critical to realize that behavior problems occur at all levels of society. These problems are not restricted to the poor. In fact, the majority of children with behavior problems are not poor. Consequently, efforts to reduce behavior problems should not be restricted to families of lower SES.

Different strategies have been employed to insure such efforts. Family-focused programs, such as Head Start in the U.S. (Garces et al., 2002), Sure Start in Great Britain (Melhuish et al., 2008), and Educa a Tu Hijo (Educate Your Child) in Cuba (Hertzman et al., 2010), have all improved behavioral and academic outcomes in disadvantaged youths through improving the quality of postnatal health, early childcare, and parenting. Other programs target parental discipline and monitoring, and parent–child communication to treat disruptive behavioral problems and improve health and social development of disadvantaged youths (Dishion et al., 2008; Webster-Stratton and Taylor, 2001). Universal school-based programs have been shown to effectively reduce behavior problems in children (Durlak et al., 2011). Additionally, interventions targeted at changing neighborhood conditions (e.g., garbage, safety) have shown similar positive results (Yoshikawa, 1994).

Standard mental health interventions present considerable access barriers to families who have low SES. These barriers include needing to take time off work to attend sessions, costs of child care, and travel expenses, including parking and meals. In addition, there is often stigma in attending a mental health clinic. Distance treatment systems such as Strongest Families™ (McGrath et al., 2011) that do not require travel schedules and

schedule appointments at times convenient for parents can help overcome these barriers. Olds' program of nurse home visiting (Kitzman et al., 2010) has also been shown to result in long-lasting outcome improvements in a variety of domains, including behavior problems.

Reducing SES differences with respect to child behavior problems is a policy priority in many countries. In 2008, the World Health Organization's Commission on the Social Determinants of Health called for urgent and coherent government action to 'close the gap' in health between the rich and the poor, with a particular focus on early childhood development and education (Irwin et al., 2007; WHO, 2008). This commitment was reinforced in 2011, when the WHO adopted resolution WHA62.14 (Rio Political Declaration on Social Determinants of Health), which compels member states to monitor the social determinants of child mental health. Indeed, social programs that reduce poverty, such as income support for poor families, economic development, and improvements to living arrangements, may decrease behavior problems in children (Yoshikawa, 1994). However, programs that strive to reduce SES differences in child behavior disorders must align with broader social and economic policy in order to change the social conditions that give rise to these differences in the first place.

See also: Academic Performance, Effects of Socio-Economic Status on; Language Development: Influence of Socio-Economic Status; Personality and Economics; Psychiatric Care, Economics of; Social and Economic Change: Psychological Challenges for Individuals; Socioeconomic Status and Social-Emotional Development of Children from Birth to 36 Months of Age: A Systematic Narrative Review.

Bibliography

Adler, N.E., Snibbe, A. Conner, 2003. The role of psychosocial processes in explaining the gradient between socioeconomic status and health. Current Directions in Psychological Science 12, 119–123.

American Psychological Association, Task Force on Socioeconomic Status, 2007. Report of the APA Task Force on Socioeconomic Status. Washington: as author.

Borghol, N., et al., 2012. Associations with early-life socio-economic position in adult DNA methylation. International Journal of Epidemiology 41, 62–74.

Bornovalova, M.A., et al., 2010. Familial transmission and heritability of childhood disruptive disorders. American Journal of Psychiatry 167, 1066–1074.

Dishion, T.J., et al., 2008. The family check-up with high-risk indigent families: preventing problem behavior by increasing parents' positive behavior support in early childhood. Child Development 79, 1395–1414.

Durlak, J.A., et al., 2011. The impact of enhancing students' social and emotional learning: a meta-analysis of school-based universal interventions. Child Development 82, 405–432.

Elgar, F.J., et al., 2005. Income inequality and alcohol use: a multilevel analysis of drinking and drunkenness in adolescents in 34 countries. European Journal of Public Health 15 (3), 245–250.

Elgar, F.J., et al., 2009. Income inequality and school bullying: multilevel study of adolescents in 37 countries. Journal of Adolescent Health 45 (4), 351–359.

Elgar, F.J., et al., 2013. Absolute and relative family affluence and psychosomatic symptoms in adolescents. Social Science and Medicine 91, 25–31.

Garces, E., et al., 2002. Longer-term effects of head start. The American Economic Review 92, 999–1012.

Heijmans, B.T., Mill, J., 2012. Commentary: the seven plagues of epigenetic epidemiology. International Journal of Epidemiology 41, 74–78.

Hertzman, C., et al., 2010. Bucking the inequality gradient through early child development. British Medical Journal 340, c468.

Hollingshead, A.B., 1975. Four Factor Index of Social Status (Unpublished working paper). Department of Sociology, Yale University, New Haven. Available at: http://psy6023.alliant.wikispaces.net/file/view/Hollingshead+SES.pdf (accessed 25.10.13.).

Irwin, L.G., et al., 2007. Early Child Development: A Powerful Equalizer. Final report for the WHO's Commission on Social Determinants of Health. Geneva: as author. Available at: http://www.who.int/maternal_child_adolescent/documents/ecd_final_m30/en/ (accessed 08.11.13.).

Kitzman, et al., 2010. Enduring effects of prenatal and infancy home visiting by nurses on children: follow-up of a randomized trial among children at age 12 years. Archives of Pediatric Adolescent Medicine 164 (5), 412–418.

Lam, L.L., et al., 2012. Factors underlying variable DNA methylation in a human community cohort. Proceedings of the National Academy of Sciences of the United States of America 109 (2), 17253–17260.

McGrath, P.J., et al., 2011. Telephone-based mental health interventions for child disruptive behavior or anxiety disorders: randomized trials and overall analysis. Journal of the American Academy of Child and Adolescent Psychiatry 50, 1162–1172.

Melhuish, E., et al., 2008. National evaluation of sure start research team. Effects of fully-established sure start local programmes on 3-year-old children and their families living in England: a quasi-experimental observational study. Lancet 372, 1641–1647.

O'Connell, M.E., Boat, T., Warner, K.E. (Eds.), 2009. Preventing Mental, Emotional, and Behavioral Disorders among Young People: Progress and Possibilities. National Academies Press, Washington, DC.

Pickett, K.E., Mookherjee, J., Wilkinson, R.G., 2005. Adolescent birth rates, total homicides, and income inequality in rich countries. American Journal of Public Health 95 (7), 1181–1183.

Pickett, K.E., Wilkinson, R.G., 2007. Child wellbeing and income inequality in rich societies: ecological cross sectional study. British Medical Journal 24 (7629), 1080.

Reagan, P.B., Salsberry, P.J., Olsen, R.J., 2007. Does the measure of economic disadvantage matter? Exploring the effect of individual and relative deprivation on intrauterine growth restriction. Social Science and Medicine 64, 2016–2029.

Reiss, F., 2013. Socioeconomic inequalities and mental health problems in children and adolescents: a systematic review. Social Science and Medicine 90, 24–31.

Repetti, R.L., Taylor, S.E., Seeman, T., 2002. Risky families: family social environments and the mental and physical health of offspring. Psychological Bulletin 128 (2), 330–366.

Singh, G.K., Ghandour, R.M., 2012. Impact of neighborhood social conditions and household socioeconomic status on behavioral problems among US children. Maternal and Child Health Journal 16 (1), S158–S169.

Sourander, A., et al., 2005. Childhood predictors of psychiatric disorders among boys: a prospective community-based follow-up study from age 8 years to early adulthood. Journal of the American Academy of Child and Adolescent Psychiatry 44, 756–767.

Sourander, A., et al., 2008. Changes in psychiatric problems and service use among 8-year-old children: a 16-year population-based time-trend study. Journal of the American Academy of Child and Adolescent Psychiatry 47, 317–327.

Sweet, E., 2011. Symbolic capital, consumption, and health inequality. American Journal of Public Health 111, 260–264.

Tremblay, R.E., 1999. When children's social development fails. In: Keating, D.P., Hertzman, C. (Eds.), Developmental Health and the Wealth of Nations : Social, Biological, and Educational Dynamics. Guildford Press, New York, pp. 55–71.

Vine, M., et al., 2012. Associations between household and neighborhood income and anxiety symptoms in young adolescents. Depression and Anxiety 29, 824–832.

World Health Organization, Commission on Social Determinants of Health, 2008. Closing the Gap in a Generation: Health Equity through Action on the Social Determinants of Health: Final Report. Geneva: as author.

Webster-Stratton, C., Taylor, T., 2001. Nipping early risk factors in the bud: preventing substance abuse, delinquency, and violence in adolescence through interventions targeted at young children (0–8 years). Prevention Science 2, 165–192.

Wilkinson, R., Marmot, M., 2001. Psychosocial and material pathways in the relation between income and health: a response to Lynch. British Medical Journal 322, 1233–1236.

Wilkinson, R., Marmot, M. (Eds.), 2003. The Solid Facts, second ed. Regional Office for Europe of the World Health Organization, Copenhagen.

Yoshikawa, H., 1994. Prevention as cumulative protection: effects of early family support and education on chronic delinquency and its risks. Psychological Bulletin 115, 28–54.

Yoshikawa, H., et al., 2012. The effects of poverty on the mental, emotional, and behavioral health of children and youth: implications for prevention. American Psychologist 67, 272–284.

Behavioral Theories of Organization

Henrich R Greve, INSEAD, Singapore
Linda Argote, Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Behavioral theories of organization are a major portion of the field of organizational studies, as behavioral reasoning is found across a range of theoretical approaches. We review the role of behavioral theory in the major research traditions of organization theory, starting with the original behavioral theory of the firm and continuing to evolutionary theory, population ecology, institutional theory, and transaction cost theory. We also address the organizational problems investigated by each theory, and we summarize by noting a trend toward greater integration of the field around behavioral reasoning.

Introduction

Behavioral theories of organization are typically traced back to Cyert and March, "A Behavioral Theory of the Firm," which sought to "develop an empirically relevant, process-oriented, general theory of decision making by a business firm" (Cyert and March, 1963: 3). The 'Behavioral' article was not part of their guidelines for constructing theory; but in order for a theory to be oriented toward the process of making decisions and to be empirically relevant, it had to focus on actual decision-making behavior. Hence, Cyert and March proposed a behavioral theory of the firm (BTOF) as a counterpoint to theories that were not grounded in observation of decision-making processes and were not proven to be empirically relevant for firm-level decisions, although they might still yield useful macro-level predictions.

Modern reviews and reformulations have been divided between attempts to reaffirm this definition of 'behavioral' and attempts to extend it. A recent review reflects this divide, as it is open about the definition in keeping with the general growth and diversification of the behavioral literature, but it also notes that behavioral theories have a common foundation in ideas of individual bounded rationality and organizational procedures for decision making (Gavetti et al., 2012).

Within the field of management, the idea that theories should be built on observation of decision-making processes is currently so conventional that its opposite can seem to have mostly historical value. However, the classical definition of 'behavioral' still has some bite left. Theories of intraorganizational decision making based on equilibrium ideas are not process theories and hence fail the process part of the definition, meaning that agency or contract theory, for example, is not behavioral. If a theory conforms to empirical observation, it may fit the other part of the definition, empirical relevance. Some theories that are phrased at a very high level of analysis also lack a clear process theory of decision making, or they would be consistent with multiple process theories of decision making, and such theories are not behavioral either. Population ecology did not start out as a behavioral theory, although we will argue in this article that it has become behavioral.

The process-oriented view of decision making in organizations necessarily implies a focus on multiple persons, and potentially multiple organizational units, being engaged in decision making. The earliest formulation of the BTOF emphasized this through its theory of problemistic search, which started locally to a problem but expanded if the search did not have a local solution, along with its theory of coalitions in decision making (Cyert and March, 1963). This focus on multiple persons anchors behavioral theories to the meso level of analysis. Although specific treatments can cross levels of analysis and hence look at how the meso level interacts with micro-level or macro-level outcomes, the meso level is unavoidable. Through its emphasis on observable processes, behavioral theory is also heavily focused on organizational mechanisms that produce decisions, rather than decisions as a result of incentives or pressures.

Behavioral theories of organizations need to take into account that behavior is carried out in an organization. Although scholars debate the exact definition of an organization, most definitions agree that organizations are collections of individuals and groups who perform interdependent activities that are oriented toward the accomplishment of goals. Other key characteristics of organizations are that they exhibit some permanence over time and are embedded in an external environment, but theorists differ in whether they see these characteristics as defining or as simply frequent. The defining characteristics have implications for what should be included in a behavioral theory of organization in order to be a complete theory. We think that organizational theory could address at least the problems that are listed in **Table 1**. Indeed, we find that a selection of major theories in the current literature does address these problems to differing degrees. We indicate this with stars designating the theoretical perspective that focuses on their solution. Multiple stars are used to designate that the theoretical perspective emphasizes the particular issue, while a single star indicates that it provides some treatment.

Going down the rows:

1. New organizations and forms spread when organizations of certain kinds become seen as legitimate or efficient means of fulfilling specific purposes. Population ecology predicts when new organizations will appear as a function of the density and concentration of existing firms (Carroll and Hannan, 2000). Transactions cost theory predicts when transactions will occur in the market or in organizations, and hence when new organizations are likely (Williamson, 1991).

Table 1 Theoretical perspectives and organizational problems

Organizational problem	*Behavioral theory of the firm*	*Evolutionary theory*	*Population ecology*	*Institutional theory*	*Transaction cost theory*
1. New organizations and forms			**	*	*
2. Forming goals and establishing tasks	*				
3. Dividing and coordinating tasks	*	*			**
4. Motivating members	*				
5. Learning and adaptation	**	**	*	*	*

2. Behavior is motivated toward the achievement of goals. The goals can be divided into subgoals (Cyert and March, 1963), which can be further broken into tasks for individuals to perform. Cyert and March (1963) addressed the development of goals through the concept of aspirations, which are a function of the organization's past performance and the performance of similar organizations. The goals are also affected by the organization's purpose, which is established at its founding and renegotiated periodically. The tasks that individuals perform in organizations are interdependent: accomplishing one task affects the accomplishment of another.
3. These interdependent tasks must be coordinated, which can be achieved through routines, a concept that grew out of the standard operating procedures of the BTOF (Cyert and March, 1963) and was developed fully in the evolutionary theory of the firm (Nelson and Winter, 1982).
4. A behavioral theory of organizations should take into account that organizations are composed of individuals whose participation and performance require motivation.
5. Organizations learn from their experience. This learning is embedded in routines so that knowledge acquired from past experience can affect future performance (Levitt and March, 1988; Nelson and Winter, 1982). Sometimes, the learning is adaptive and results in the organization fitting its environment better; at other times, the learning is maladaptive, and the organization's fit with the environment is not improved. Population ecology also examines the extent to which the organization fits its environment, including adaptive processes that can lead to organizational failure.

Thus, current theories of organizations focus on different organizational problems. Population ecology focuses on the interplay between the organization and its environment and studies how the founding and the death of organizations are affected by the ecology of other organizations and the nature of the environment. Population ecology theories treat organizations as unitary actors, and the early statements of the theory did not examine processes within organizations or permit learning. By contrast, the BTOF explicates the internal processes of organizations by including issues such as coordination, motivation, and goal formation. The BTOF also permits learning. Evolutionary theory also includes issues of coordination and learning from experience, which contributes to path dependence in organizations. Transaction cost theory focuses on which alternatives will be internalized within an organization and which will be achieved through market transactions. However, as we will argue in this article, these traditional boundaries between the theories are increasingly becoming blurred as each theory is addressing new questions. This is a development that seems to be moving the field toward a new integration as a connected (but not single) behavioral theory of the firm, where there had earlier been several separate theories and research traditions with limited communication between each.

Review

Behavioral Theory of the Firm

The BTOF is a research tradition that traces its origins directly to "A Behavioral Theory of the Firm" (Cyert and March, 1963), and its emphasis on organizational learning is closely related to organizational learning theory (Levitt and March, 1988). There is a recent review of the BTOF (Gavetti et al., 2012), so we review this research briefly and with an emphasis on key contributions and recent work.

One branch of this research has taken the statements on managerial decision making as a key feature of this theory. A key feature of the stylized decision maker in the BTOF is a focus on goal variables and adaptation of aspiration levels, and verification of this assumption is a feature of both older and current work. Adjustment of risk-taking propensity as a result of performance relative to aspiration levels was not part of the original formulation of the theory, but it has since become central in work on individual decision making (March and Shapira, 1992), in part as a result of cross-fertilization of ideas with psychology such as prospect theory.

The focus on goals in the BTOF also carries over to research on top management team and organizational levels of analysis. Because behaviors differ as a result of which goals are examined and which parts of the environment are salient, the question of organizational attention comes to the fore (Ocasio, 1997). The effects of attention on organizational action are explored in research on outcomes such as the attention on top management teams and public statements of the organization. This work is closely related to research that integrates the influences of managerial information seeking, managerial characteristics, and responses to performance on decision making.

The BTOF predicts that organizations respond to performance below the aspiration level by engaging in problemistic search. This has been developed into a research stream on performance feedback, which examines how organizational changes result from performance that is below aspiration levels, which in turn adapts as a result of social and historical comparisons (Greve, 2003). Much of this work is at the organizational level of analysis, and it shows effects on

decisions that seem tailored to quickly recover from poor performance such as market entry, investment, and innovation launches (Greve, 2003). These same decisions are risky, however, so the research also shows an increase in risk taking when performance is below aspiration levels.

An important insight in the BTOF is that organizations learn from their experience. The knowledge or intelligence they acquire is embedded in standard operating procedures or routines, which affect future performance. Thus, rather than every action an organization takes being a result of calculation, many actions are a result of enacting routines.

Organizational routines are interdependent task sequences performed by multiple members. Early conceptions of routines emphasized that they conserve lessons from the past and thereby improve efficiency in the future. Routines were conceived as a source of stability in organizations and as representing political truces of how tasks should be performed in organizations. Recent work on routines has emphasized that they are a source of change as well as stability (Feldman and Pentland, 2003) and has conceived the gradual development of organizational routines as an engine of evolution in organizations (Nelson and Winter, 1982).

Two strands of research on organizational learning have taken different stances on the role of organizational experience. One strand, exemplified by research on organizational learning curves, finds that although there is considerable variation in the rate of improvement, experience generally improves dimensions of organizational performance (Argote, 2013). Another strand emphasizes the challenges of appropriately sampling and learning from experience in organizations (Denrell and March, 2001). These two strands of research have become intertwined to some extent, with researchers explaining the variation in the rate at which organizations learn from experience and identifying conditions under which experience is harmful or helpful for organizational performance.

Levitt and March (1988) theorized that organizations not only learn from their own direct experience but also learn from the experience of other organizations. This latter form of learning has been referred to as vicarious learning or knowledge transfer. Considerable research has been conducted over the last couple of decades on the mechanisms through which and the conditions under which one organization learns from the experience of others. Knowledge transfer across organizations plays a key role in evolutionary theories as well as learning theories.

The BTOF has, since the start, shown a concern for the consequences of the behavioral rules that empirical research uncovered. This has led to a rich modeling tradition that shows how different behavioral rules affect organizational performance and adaptation in stable or changing environments (March et al., 2000). Classic findings include work showing the danger of rapid learning of routines or precise reactions to noisy feedback, both of which can cause an inherently inferior strategy to be chosen. This argument was further developed into the famous tradeoff between exploration of new strategies and exploitation of current ones (March, 1991), an argument that has since been shown to be stable across modeling frameworks (Denrell and March, 2001). Difficulties in finding the best strategy or structure comprise an enduring topic in this modeling tradition, and many treatments focus on how the structure of the environment presents traps or how the organizational structure affects decision making and learning (Fang et al., 2010).

Evolutionary Theory

Evolutionary theory is a branch of research that traces its origins back to "An Evolutionary Theory of Economic Change" (Nelson and Winter, 1982), a treatment that shared with Cyert and March (1963) a rejection of maximization assumptions and full rationality. It differed in emphasis by examining the gradual development of organizational routines as the engine of evolution, reasoning through a parallel with human skill development, and thus lacked the emphasis on stability and discrete change events found in the theory of problemistic search (Cyert and March, 1963).

Like learning theory, evolutionary theory has since developed many branches, and a brief review cannot do full justice to this work. A common feature of the work is that it emphasizes path dependence. Evolutionary processes only differ meaningfully from systems with equilibrium features when they are not teleological – they do not evolve toward fixed points that are determined by some maximization process. Rather, multiple end states are possible, and which one is reached depends on early steps that in turn can be the result of chance events or the agency of actors who do not fully foresee the consequences of their actions. Technological paths that get 'locked in' by early events are classical examples of evolutionary processes.

One branch of evolutionary theory examines the evolution of firms. It views the firm as taking on a trajectory determined by its current stock of knowledge and its rules for altering this stock, and it is thus closely related to learning models, although it sometimes blends intraorganizational learning mechanisms with external selection mechanisms (Levinthal and Posen, 2007). The behavioral assumptions are that managers have some ability to choose (short-term) superior options, but are prone to make incorrect judgments when short- and long-term payoffs must be traded against each other. Finally, firm-level evolutionary work often treats issues of how to design the firm as an effective adaptive system. It typically assumes that organizations vary in their designs, but their managers do not fully understand the adaptive consequences of this variation.

Empirically, the research focuses on the transformation and replication of knowledge. Direct investigation of knowledge transformation is challenging, but there has been theoretical and empirical work on how organizations can influence the knowledge acquisition process and thus shape the evolution of routines (Zollo and Winter, 2002). Likewise, there is work showing that organizations can 'fix' routines that have reached a high level of performance in order to prevent experimentation that might lead to lower performance levels and that replicating routines exactly can be associated with higher performance than attempts to adapt routines to local conditions (Knott, 2001). Finally, because evolutionary processes are influenced by the level of variation available to select from, manipulating the proportion of organizational units that are allowed to vary influences the overall rate of change. This can be done through the form of governance or through managerial routines with some local autonomy.

Replication is particularly important because knowledge replication is required in order for an evolutionary theory of the firm to contain something approaching a genetic code (Winter and Szulanski, 2001). Research has shown that firms are indeed able to make routines replicable across units, either with or without planned variation according to local conditions, but in so doing they face the dilemma that replicable routines are also more easily observed and copied by competitors. Researchers have suggested that this dilemma can be resolved by embedding knowledge in transactive memory systems (Argote and Ren, 2012).

Transactive memory systems are collective systems for encoding, storing, and distributing information. Colloquially known as 'knowledge of who knows what,' transactive memory systems capture knowledge of who knows what and who is best at what. Transactive memory systems develop as organizational members gain experience working together, and they are idiosyncratic to organizations because they depend on the particular individuals who are members of an organization. Transactive memory systems are hard for competitors to imitate. Thus, embedding knowledge in transactive memory systems is more effective at preventing knowledge spillover to other organizations than embedding knowledge in routines.

Another branch of evolutionary theory examines the evolution of industries. Research on technological paths has examined how technological choices are shaped by economic factors that favor making the same technological choices as other firms in the same industry, which means that initial leads in adoptions become magnified. Transitions between generations of technologies have also spurred interest among evolutionary theorists, with a particular focus on whether the new technology has characteristics that favor incumbent firms or newly founded firms (Tushman and Anderson, 1986). These empirical studies are built on evolutionary reasoning in which a central assumption is some form of behavioral myopia in which firms pursue immediate incentives, even if these lead them to commit to a technology that will lose in the long run.

Research on the structure of transactions in industries has examined the reconfiguration of transaction interfaces between firms over time. This work has shown how new transaction structures are gradually formed from old ones rather than created anew. This process is seen when industries form additional transactional links, as in the creation of a less integrated industrial structure. It is also observed more generally during periods of industrial formation or change of boundaries between actors, such as those that occur when major new technologies create opportunities (Jacobides and Winter, 2005) or market actors seek to rearrange exchange relations.

Population Ecology

Population ecology theory focuses on explaining organizational foundings and deaths as a function of characteristics of the organization and other organizations in their environment. Population ecologists theorized that selection was the mechanism through which organizations changed (Hannan and Freeman, 1977); organizations that fit their environments were selected and prospered, and those that did not fit their environments perished. Organizational change was seen as rare and likely to harm rather than help an organization's survival prospects (Carroll and Hannan, 2000).

Population ecology has produced a large body of empirical research. Carroll and Hannan (2000) reviewed the population ecology literature and identified important empirical regularities in it. First, studies generally find a liability of smallness: as organizational size increases, failure rates decrease. Second, although a liability of newness was once theorized, recent empirical work that controls for organizational size has found that mortality rates increase with organizational age. Third, the relationship between births and deaths of firms, on the one hand, and the number of competing firms in the environment, on the other, is nonmonotonic. At low levels of firm density, founding rates increase, and mortality rates decrease with increases in density. At high levels of density, however, founding rates decrease and mortality rates increase. The former effect has been attributed to legitimation, while the latter has been attributed to competition.

Although the original statement of population ecology theory argued that selection and not learning or adaptation was the mechanism through which organizations change, more recent treatments have acknowledged that learning occurs, although it is regarded as rare (Carroll and Hannan, 2000). Learning and selection have been productively integrated in models of organizational performance and survival. Another connection between population ecology and learning can be found in studies on the success of new entrepreneurial ventures. Research has found that entrepreneurial firms that spin out of other firms or whose members have previous experience in other firms perform better than those lacking previous experience.

Population ecology theory intersects with institutional theory on the concepts of legitimacy and organizational form. As noted in this article, there is evidence that as the density of a population increases, survival prospects of organizations first increase and then decrease. Population ecologists attribute the initial increase in survival to the increasing legitimacy of the organizational form. An organizational form is perceived as legitimate when it is taken for granted as the appropriate way to accomplish certain collective tasks. An active, current stream of research focuses on organizational form as a specific kind of social identity (Hsu and Hannan, 2005).

Institutional Theory

Institutional theory is a research tradition that traces its origins back to foundational articles that discussed how organizational founding and change were driven less by functional considerations and more by symbolic actions and external influences than the theory at the time assumed (Meyer and Rowan, 1977). These articles drew on concepts of bounded rationality that are central to behavioral theories and sketched a broad range of potential research questions, but much subsequent research drew away from the firm focus on behavioral theories of organization by emphasizing environmental influences such as the diffusion of new institutionalized practices among firms.

As institutional theory has grown, some branches have moved closer to behavioral theory. Direct dialogue between the perspectives has been started by researchers who have noted that the organizational change processes examined by

behavioral theory are influenced by the institutional context (Wezel and Saka-Helmhout, 2006). A growing subfield of institutional theory concerns institutional logics, which are broadly (but not universally) shared assumptions and action patterns (Thornton, 1995). At the organization level, institutional logics can be seen as sources of managerial decision-making rules, and hence institutional logics research is related to research on the BTOF and evolutionary theory.

Institutional theory has also moved into examining the founding conditions for new firms (Tolbert et al., 2011). This work questions the conventional assumption that entrepreneurs are rationally able to locate opportunities, and it instead posits that key sources of organization founding activities are institutional features of the social group to which entrepreneurs belong or the symbolic environment they face. Like population ecology, this work moves the concerns for decision-making processes and bounded rationality to the stage of organizational founding.

Transaction Cost Theory

Focusing on firm boundaries, transaction cost theory aims to answer the question of when activities would occur within the market and when they would occur within the firm (Williamson, 1991). More specifically, transaction cost theory predicts when the governance forms of hierarchies, markets, or hybrids (e.g., alliances) will be used. Williamson, who was recognized with a Nobel Prize for his work on transaction costs, theorized that whether activities would be internalized within a firm depended on their transaction costs. He saw transactions broadly as transfers of goods or services across interfaces, and argued that when transaction costs were high, internalizing the transaction within a hierarchy was the appropriate decision. Conversely, when transaction costs were low, buying the good or service on the market was the preferred option. Three dimensions were developed for characterizing transactions: uncertainty, frequency, and asset specificity, or the degree to which transaction-specific expenses were incurred. Transaction cost theory is built on assumptions of bounded rationality and opportunism, defined as self-interest with guile.

A review of the empirical literature on transaction cost theory concluded that findings regarding asset specificity were generally supportive of the theory, while findings for uncertainty were mixed (David and Han, 2004). Although only a small number of studies focused on the frequency of transactions, those studies were generally supportive of the theory. David and Han (2004) further concluded that considerable empirical support existed for predictions about hierarchies versus markets, but less support existed for predictions about the choice of hybrids over markets or hierarchies.

A debate has raged in recent years about whether transaction costs or organizational capabilities are the most important determinants of firm boundaries. Emphasizing the importance of capabilities, researchers have argued that firms internalize activities that they perform with greater capability than external providers (Jacobides and Winter 2005). This debate connects transaction cost theory to the BTOF and evolutionary theory because organizational learning, which figures prominently in those theories, contributes to the development of routines and organizational capabilities.

Recent theorizing in transaction costs includes capabilities as well as transaction costs as determinants of whether firms internalize activities. For example, Argyres and Zenger (2012) noted that transaction costs and capabilities are intertwined over time: past governance decisions affect the development of current capabilities, and current capabilities affect future governance decisions. Current research focuses on articulating the relationship between capabilities and transaction costs.

Discussion

Behavioral theories of organizations – in plural – are important for understanding how organizations actually work, and they represent the bulk of empirical work performed on this topic. We are unable to give a full review within a manageable number of pages, but we have given a taste of the work. The range of topics is wider than our table suggests, although new organizations and forms, forming goals and establishing tasks, motivating members, dividing and coordinating tasks, and learning and adaptation are still an impressive list of organizational problems. Likewise, the behavioral theory of the firm, evolutionary theory, population ecology, institutional theory, and transaction cost theory are important lines of investigation under the umbrella of behavioral theories of organizations, but they are not the only ones.

Although this field of research is too large and fast moving to make the identification of trends easy, we have observed that theories that start out behavioral remain so, and theories that initially have little behavioral content incorporate it over time. The behavioral concepts of organizational learning and knowledge transfer have been central to the BTOF and evolutionary theory since their inception and remain central, and these concepts are playing increasing roles in population ecology, institutional theory, and transaction costs theory. In the population ecology tradition, specific cases of such reasoning include work examining learning from other organizations, including work arguing that the transfer of routines from parent firms contributes to the success of their progeny. In institutional theory, early work on the diffusion of institutional practices emphasized their symbolic value, but recent work includes considerations of goal-oriented behavior that aligns the reasoning with behavioral theory. Transaction costs researchers acknowledge that an organization's capabilities, which are developed by learning from experience, affect decisions about whether to internalize tasks.

It is too early to tell the strength of the theoretical movement toward a common ground described here. If it does gain momentum, it will produce a unique result. Research traditions that started out with little relations to each other will coalesce into an overarching field of study joined by an interest in creating behavioral theories of organizations that are informed by ideas of bounded rationality. We think such an outcome would produce an era of dialogue between the different research traditions, to the benefit of our understanding of organizations.

See also: Ecology: Organizations; Hierarchies and Markets; Learning: Organizational; Modularity and Organizations;

Organization: Overview; Organizational Decision Making; People in Organizations; Strategic Management.

Bibliography

Argote, L., 2013. Organizational Learning: Creating, Retaining and Transferring Knowledge, second ed. Springer, New York.

Argote, L., Ren, Y., 2012. Transactive memory systems: a micro foundation of dynamic capabilities. Journal of Management Studies 49 (8), 1375–1382.

Argyres, N.S., Zenger, T.R., 2012. Capabilities, transaction costs, and firm boundaries. Organization Science 23 (6), 1643–1657.

Carroll, G.R., Hannan, M.T., 2000. The Demography of Corporations and Industries. Princeton University Press, Princeton, NJ.

Cyert, R.M., March, J.G., 1963. A Behavioral Theory of the Firm. Prentice-Hall, Englewood Cliffs, NJ.

David, R.J., Han, S.K., 2004. A systematic assessment of the empirical support for transaction cost economics. Strategic Management Journal 25 (1), 39–58.

Denrell, J., March, J.G., 2001. Adaptation as information restriction: the hot stove effect. Organization Science 12 (5), 523–538.

Fang, C., Lee, J., Schilling, M.A., 2010. Balancing exploration and exploitation through structural design: the isolation of subgroups and organizational learning. Organization Science 21 (3), 625–642.

Feldman, M.S., Pentland, B.T., 2003. Reconceptualizing organizational routines as a source of flexibility and change. Administrative Science Quarterly 48 (1), 94–118.

Gavetti, G., Greve, H.R., Levinthal, D.A., Ocasio, W., 2012. The behavioral theory of the firm: assessment and prospects. Academy of Management Annals 6, 1–40.

Greve, H.R., 2003. Organizational Learning from Performance Feedback: A Behavioral Perspective on Innovation and Change. Cambridge University Press, Cambridge, UK.

Hannan, M.T., Freeman, J., 1977. The population ecology of organizations. American Journal of Sociology 82 (5), 929–964.

Hsu, G., Hannan, M.T., 2005. Identities, genres, and organizational forms. Organization Science 16 (5), 474–490.

Jacobides, M.G., Winter, S.G., 2005. The co-evolution of capabilities and transaction costs: explaining the institutional structure of production. Strategic Management Journal 26 (5), 395–413.

Knott, A.M., 2001. The dynamic value of hierarchy. Management Science 47 (3), 430–448.

Levinthal, D.A., Posen, H.E., 2007. Myopia of selection: does organizational adaptation limit the efficacy of population selection? Administrative Science Quarterly 52 (4), 586–620.

Levitt, B., March, J.G., 1988. Organizational learning. In: Scott, W.R., Blake, J. (Eds.), Annual Review of Sociology, vol. 14. Annual Reviews, Palo Alto, CA, pp. 319–340.

March, J.G., 1991. Exploration and exploitation in organizational learning. Organization Science 2, 71–87.

March, J.G., Schulz, M., Zhou, X., 2000. The Dynamics of Rules: Change in Written Organizational Codes. Stanford University Press, Stanford.

March, J.G., Shapira, Z., 1992. Variable risk preferences and the focus of attention. Psychological Review 99 (1), 172–183.

Meyer, J.W., Rowan, B., 1977. Institutionalized organizations: formal structure as myth and ceremony. American Journal of Sociology 83, 340–363.

Nelson, R.R., Winter, S.G., 1982. An Evolutionary Theory of Economic Change. Belknap, Boston.

Ocasio, W., 1997. Towards an attention-based theory of the firm. Strategic Management Journal 18 (Summer), 187–206.

Thornton, P.H., 1995. Accounting for acquisition waves: evidence from the U.S. college publishing industry. In: Scott, W.R., Christensen, S. (Eds.), The Institutional Construction of Organizations: International and Longitudinal Studies. Sage, Thousand Oaks, pp. 199–225.

Tolbert, P.S., David, R.J., Sine, W.D., 2011. Studying choice and change: the intersection of institutional theory and entrepreneurship research. Organization Science 22 (5), 1332–1344.

Tushman, M.L., Anderson, P., 1986. Technological discontinuities and organizational environments. Administrative Science Quarterly 31, 439–465.

Wezel, F.C., Saka-Helmhout, A., 2006. Antecedents and consequences of organizational change: 'institutionalizing' the behavioral theory of the firm. Organization Studies 27 (2), 265–286.

Williamson, O.E., 1991. Comparative economic organization: the analysis of discrete structural alternatives. Administrative Science Quarterly 36, 269–296.

Winter, S.G., Szulanski, G., 2001. Replication as strategy. Organization Science 12 (6), 730–743.

Zollo, M., Winter, S.G., 2002. Deliberate learning and the evolution of dynamic capabilities. Organization Science 13 (3), 339–351.

Belief in a Just World

Claudia Dalbert and Matthias Donat, Martin Luther University of Halle-Wittenberg, Halle/Saale, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

According to just world research, people need to believe that they live in a world in which everyone gets what they deserve and deserves what they get. This belief in a just world is conceptualized as an interindividually varying disposition and has important adaptive functions. The stronger this belief is, the more people trust in justice and the more they defend justice in the world, either behaviorally or cognitively, and the better their well-being is.

Conceptualization of the Belief in a Just World

Societies are full of inequalities and injustices – the uneven distribution of wealth or the inequality of access to health care and education to name just a few. Individuals react differently to observed or experienced injustice. Some feel moral outrage and seek to restore justice (e.g., Montada et al., 1986). Others show disdain for the victims (for a review, see Lerner and Miller, 1978) or adopt belief systems that serve to justify existing social, economic, and political arrangements (Jost et al., 2004). In other words, people confronted with injustices restore justice in reality or they may try to restore injustices – that are difficult to redress in reality – cognitively by blaming the victim or justifying the status quo. The more people believe in a just world the more they care for justice.

The Just World Hypothesis

Several psychological theories propose explanations for justice-driven reactions. One of the most influential is the just world hypothesis introduced by Lerner (e.g., Lerner and Simmons, 1966). The just world hypothesis states that people need to believe in a just world in which everyone gets what they deserve and deserves what they get. This belief enables them to deal with their social environment as though it were stable and orderly and thus serves important adaptive functions. As a result, people are motivated to defend their belief in a just world when it is threatened by injustices, either experienced or observed. If possible, justice is restored in reality (e.g., by compensating victims). If the injustice seems unlikely to be resolved in reality, however, people restore justice cognitively by reevaluating the situation in line with their belief in a just world. This cognitive process is called the assimilation of injustice.

This just world dynamic was first evidenced by Lerner and Simmons (1966). These researchers confronted their participants with an 'innocent victim,' a young woman participating in a learning task who was punished for each mistake by being administered seemingly painful electric shocks. When led to believe that the experiment would continue in the same way, the participants showed disdain for the victim on an adjective measure; when led to believe that the victim would be compensated for the pain of the electric shocks by receiving money for each correct answer in a second part of the experiment, they stopped disdaining the victim. Finally, nearly all participants who were given the choice between continuing the shock condition and switching to the compensation condition voted for the latter. Note, however, that merely voting to award the victim compensation did not stop participants from derogating the victim. It was only when they were certain that compensation would be given that the injustice was no longer assimilated. This innocent victim paradigm remains the most influential in modern experimental just world research; it is only the type of innocent victim that has changed (e.g., Correia et al., 2007). In sum, the confrontation with an innocent victim triggers a motivation to restore justice.

The Belief in a Just World as a Disposition

A substantial amount of research on belief in a just world has been experimental in nature (for a review, see Hafer and Bègue, 2005), focusing primarily on the maladaptive functions of the belief in a just world, such as disdain for the victim. Since the 1970s, however, another strand of research has examined individual differences in the belief in a just world and found that it also serves important adaptive functions (for a review, see Furnham, 2003). This research agenda was triggered by the introduction of the first belief in a just world scale by Rubin and Peplau (1973), which assessed individual differences in the belief that the world is generally a just place. This approach allowed the role of the belief in a just world to be investigated within the framework of personality dispositions, and positive associations were found particularly with authoritarianism and internal locus of control (for a review, see Furnham and Procter, 1989).

Since the 1990s, more studies have investigated the positive as well as the negative social consequences of the belief in a just world, and the focus of these investigations has been extended to cover the consequences of holding a belief in a just world for the believers. Based on suggestions originating from earlier research (e.g., Furnham and Procter, 1989), these studies have shown that it is necessary to distinguish the belief in a personal just world, in which one is usually treated justly, from the belief in a general just world, in which people in general get what they deserve (e.g., Dalbert, 1999). In line with the self-serving bias in general and in fairness reasoning in particular (Messick et al., 1985), research evidenced that people tend to endorse the personal more strongly than the general belief in a just world and that the two constructs have a different meaning. The personal belief in a just world is a better predictor of adaptive outcomes (e.g., subjective well-being), and the belief in a just

world in general is a better predictor for example of harsh social attitudes (e.g., Bègue and Muller, 2006).

Of course, other differentiations of the just world construct have also been proposed. For the general just world belief, e.g., a general belief in immanent justice has been distinguished from a general belief in ultimate justice, and a general belief in distributive justice has been distinguished from a general belief in procedural justice (e.g., Lucas et al., 2007). Finally, the general belief in a just world has been differentiated from the general belief in an *un*just world (Dalbert et al., 2001). This latter research showed that general belief in a just world should not be seen as a bipolar construct, but as a two-dimensional one. However, the differentiation between a more general and a more personal just world belief thus far seems to be the most widespread and well-examined distinction. Thus, the present summary focuses on research on general and personal just world beliefs.

Belief in a Just World and Other Personality Dispositions

One of the first associations observed between the belief in a just world and other personality dispositions was the positive correlation between just world belief and religiosity (e.g., Rubin and Peplau, 1973). Research on the differences between the two has confirmed that they are distinct dispositions, and cross-cultural research has found few differences in the just world belief across cultures with contrasting religious and political backgrounds (e.g., Furnham, 1993). A positive and sometimes substantial association has also been found between authoritarianism and just world belief (for a review, see Furnham and Procter, 1989). Analyses of the common factor structure of the two constructs support the two-factor hypothesis and their differential meaning, with the belief in a just world providing a more positive outlook than authoritarianism (e.g., Dalbert, 1992). The positive associations repeatedly observed between just world belief and internal locus of control have prompted speculation about an overlap between these two constructs as well (for a review, see Furnham and Procter, 1989). The two constructs should be distinct from a theoretical perspective, however. The belief in personal agency is consistent with the belief in a just world as long as the justice principle endorsed is the equity norm. Other ideas of justice (e.g., the equality or the need principle of justice or the belief in a just God) are not consistent with the belief in internal control. Finally, there is some evidence to suggest that the belief in a just world as a personality trait is correlated with global personality dimensions. In particular, empirical findings indicate a negative relationship between just world belief and neuroticism, consistent with the positive outlook that the belief in a just world provides (e.g., Lipkus et al., 1996). Nevertheless, studies controlling for neuroticism evidenced the incremental validity of the just world belief (e.g., Dalbert and Dzuka, 2004). Taken collectively, research supports the differential validity of the belief in a just world within the network of personality dispositions.

Justice Motive versus Justice Motivation

In the context of just world research and theory, scholars often speak of the justice motive (e.g., Ross and Miller, 2002). The shift from the experimental to the individual differences approach to the belief in a just world made it necessary to differentiate between a justice motive and justice motivation. Motives are individual dispositions reflecting individual differences in the tendency to strive for a specific goal. A justice motive is thus an individual disposition to strive for justice as an end itself. According to Lerner (1977), the individual belief in a just world can be interpreted as an indicator of such a justice motive. The belief in a just world indicates a personal contract; the more people want to rely on being treated justly by others, the more obligated they should feel to behave justly themselves. Thus, the stronger the belief in a just world is, the stronger the justice motive.

Experimental just world research typically does not assess individual differences, however, but interprets experimental reactions in the light of just world reasoning. Such research thus addresses justice motivation, and not the justice motive as an individual differences disposition. Motivation can be defined as a person's orientation toward a specific goal in a specific situational state; thus, justice motivation means the orientation toward justice in a given situation. Justice motivation is triggered by specific situational circumstances in interaction with personal dispositions. In the case of justice motivation, that personal disposition may be the justice motive or other dispositions (e.g., Lind and van den Bos, 2002). As an extension of this reasoning, belief in a just world is an indicator of a justice motive; justice motive theory (Dalbert, 2001) moreover perceives it as an implicit justice motive that operates on an intuitive level of information processing and predicts intuitive reactions to injustice (for a review, see also Hafer and Bègue, 2005).

Consequences of the Belief in a Just World

Today belief in a just world represents a disposition that can easily be measured. Thus, a growing number of research in the last decade was able to show important consequences of that belief. In line with justice motive theory, this research indicates that belief in a just world leads to intuitive justice-driven reactions as the striving for just behavior and the avoidance of unjust behavior, assimilation of injustice, and trust in justice. Therefore, belief in a just world serves adaptive functions and can thus be seen as a resource that sustains subjective well-being (Dalbert, 2001).

Belief in a Just World and Social Behavior

In a just world, a positive future is not the gift of a benevolent world, but a reward for the individual's behavior and character. Consequently, the more individuals believe in a just world, the more compelled they should feel to strive for justice themselves. The just world belief is thus indicative of a personal contract (Lerner, 1977), the terms of which oblige the individual to behave justly. Therefore, strong just world believers are more likely to help people in need (Bierhoff et al., 1991), at least as long as the victims are seen as 'innocent' or as member of the in-group (e.g., Correia et al., 2007). In addition, the belief in a just world has been shown to be one of the important correlates of social responsibility (Bierhoff, 1994), commitment to just means (e.g., Hafer, 2000), and, inversely,

delinquent and cheating behavior (e.g., Donat et al., 2013), and bullying behavior (e.g., Donat et al., 2012). Moreover, the obligation for reciprocity has been found to be stronger among individuals with a strong just world belief (Edlund et al., 2007). Finally, a laboratory study revealed that own unjust behavior is censured by a decrease in self-esteem only for those with a strong belief in a just world (Dalbert, 1999).

Belief in a Just World and the Assimilation of Injustice

When individuals with a strong just world belief experience an injustice that they do not believe can be resolved in reality, they try to intuitively assimilate the experience to their just world belief. This can be done for example by justifying the experienced injustice as being at least partly self-inflicted, by playing down the injustice, by avoiding self-focused rumination, or by forgiving (e.g., Dalbert, 1997). As a result of these mechanisms, positive relationships have been observed between the belief in a just world and justice judgments in various domains of life. Most research into this assimilation has dealt with blaming the victim and justice judgments.

Blaming the Victim

A wealth of evidence from traditional research into the just world construct shows that individuals confronted with injustice are motivated to defend their just world belief. When observers are given the opportunity to adequately compensate an 'innocent' victim (e.g., Berscheid and Walster, 1967) and thus restore justice in reality, nearly all choose to do so (Lerner and Simmons, 1966). If they are not in a position to secure compensation for the victim, observers tend to defend their belief in a just world by psychological means. Two of these means have been examined in detail in just world research. Observers can either show disdain for victims, reasoning that their fate is a deserved punishment for a bad character (characterological attribution), or they can blame victims for having inflicted their fate upon themselves – after all, a self-inflicted fate is not unjust (behavioral attribution). Just world research has shown that observers prefer to blame the victim rather than to show disdain. The more a fate is seen as self-inflicted, the less disdain is observed (e.g., Lerner and Matthews, 1967). In sum, when people are confronted with the victim of an unjust fate, blaming the victim seems to be a crucial element in the defense of their belief in a just world.

Similar mechanisms can be assumed to operate for the victims of injustice themselves. Comer and Laird (1975) have shown experimentally that internal attributions seem to be a way of reevaluating one's fate as just. The significance of causal attributions, and especially of internal attributions, has thus been a subject of much discussion in the context of the just world hypothesis (e.g., Lerner and Miller, 1978). People with a strong just world belief are expected to be motivated to defend their belief by making internal attributions of negative outcomes, thus maintaining their subjective well-being. Although some research evidenced the hypothesized positive association between just world belief and internal attributions of the victims themselves (e.g., Hafer and Correy, 1999), other studies found no association between them (e.g., Fetchenhauer et al., 2005). Overall, the pattern of results for the belief in a just world and victims' internal attribution is rather mixed.

Justice Judgments

As a consequence of the assimilation process, individuals with a strong just world belief are expected to evaluate observed events and events in their own life as being more just. For example, school students with a strong belief in a just world have been found to be more likely to evaluate their school grades and their teachers,' peers,' and parents' behavior toward them as just (e.g., Peter and Dalbert, 2013), and feel less excluded from school (Umlauft et al., 2013). Similarly, prisoners with a strong personal just world belief were more likely to evaluate the justice of the legal proceedings leading to their conviction, the treatment by their prison officers, and decisions on prison affairs as more just (e.g., Dalbert and Filke, 2007).

The personal just world belief is usually seen as a personal disposition, but results indicating an additional, reverse effect of justice experiences on the belief in a just world qualify this assumption. Research has shown that justice experiences in the school and the family modify the personal just world belief (Dalbert and Stoeber, 2006), and factors such as length of imprisonment (Otto and Dalbert, 2005), monotony at work, and mobbing experiences at work (e.g., Dzuka and Dalbert, 2007) are negatively related to the personal just world belief. Thus, the belief in a personal just world must be seen as a partly experiential construct. Nevertheless, an unambiguous pattern of results clearly indicates that a strong personal just world belief leads to events being evaluated as just. Cubela Adoric and Kvartuc (2007) have suggested that injustice experiences only impact the belief in a just world when they reach a specific degree of adversity. Further studies are needed to determine under which conditions the just world belief fosters the assimilation of injustice and under which conditions injustice can no longer be assimilated, but instead undermines the belief in a just world and leads to adverse consequences as embitterment (Dalbert, 2010).

Belief in a Just World and the Trust in Justice

People with a strong belief in a just world are thought to be confident in being treated justly by others, and it is this trust in particular that is hypothesized to give the just world belief the character of a resource in everyday life. Assuming that people get what they deserve, they will be punished for deceiving others. Accordingly, in a just world, people are expected to be honest with one another, and people who have been deceived may conclude that they deserved it in some way. It can thus be hypothesized that people with a strong just world belief prefer not to think they have been deceived or taken advantage of. In line with this reasoning, research has shown a positive association of just world belief with general interpersonal trust (e.g., Bègue, 2002), trust in societal institutions (Correia and Vala, 2004), and young adolescents' trust in the justice of their future workplace (Sallay, 2004). This trust in future justice has a number of implications.

Risk Perception

Individuals with a strong just world belief are convinced that good things happen to good people and bad things happen to bad people. Because individuals tend to think of themselves as good people (e.g., Messick et al., 1985) the belief in a just world can be expected to give them an optimistic outlook on the

future. This buffering effect is expected to be particularly evident when people are threatened by injustice. Lambert et al. (1999) were the first to study the meaning of the belief in a just world for risk perception and showed that the just world belief seems to enable fearful individuals (i.e., those high in authoritarianism) to be confident of avoiding an unjust fate. It is particularly important for individuals exposed to external risks – i.e., those perceived to be controlled by others or by fate (e.g., robbery) – rather than to internal risks – i.e., those that are under their personal control (e.g., suicide) – to be able to rely on the environment being just. Indeed, Dalbert (2001) found that the buffering effect of the just world belief for fearful individuals held only for external risks, and not for internal risks. Finally, Hafer et al. (2001) found that individuals with a strong just world belief, but low in interpersonal control, seem to put themselves at greater risk, presumably as a consequence of a lowered risk perception. In sum, just world belief appears to function as a buffer against the perception of external risk for those who tend to need such a buffer, but this mechanism may result in higher exposure to risks in reality.

Investment in One's Future

The belief in a just world enables individuals to rely on their good deeds being rewarded at some point in the future. The certitude that everyone will ultimately get what they deserve encourages individuals to invest in their future. In contrast, those who do not believe in a just world doubt the value of such an investment, because the return on it is uncertain. Zuckerman (1975) was the first to observe that people with a strong just world belief may choose to invest in their future, when in a state of need to trust in the justice of their own future. Hafer (2000) corroborated these findings and demonstrated experimentally that individuals with a particular need to believe in a bright future defended their just world belief more strongly in the face of threat. In the same vein, questionnaire studies with students facing the school-to-work transition (Dette et al., 2004), young male prisoners (Otto and Dalbert, 2005), and young adults living in assisted accommodation (Sutton and Winnard, 2007) have shown that just world belief is positively associated with confidence that personal goals will be attained.

Achievement

Individuals with a strong belief in a just world show more trust in their future and in others' behavior toward them. It is thus hypothesized that they expect to be confronted with fair tasks in achievement situations, and for their efforts to be justly rewarded. They can thus be hypothesized to feel less threatened and more challenged by the need to achieve, to experience fewer negative emotions, and to achieve better results. Tomaka and Blascovich (1994) conducted a laboratory study to test the basic hypotheses outlined above and confronted their participants with two rapid serial subtraction tasks. Participants with a strong just world belief felt more challenged and less threatened, and performed better than those with a weak belief.

Extending this laboratory research to the school and work setting, studies have revealed a positive correlation between just world belief and school achievement, and self-rated performance at work (e.g., Kahileh et al., 2013). Finally, Allen et al. (2005) have observed that nations whose citizens have stronger just world beliefs show a faster pace of workforce modernization and GNP/capita growth.

Belief in a Just World and Subjective Well-Being

Because the main properties of the belief in a just world – indicating commitment to a personal contract, endowing trust in the justice of the world, and providing a framework for the interpretation of the events in one's life – have a variety of adaptive implications, the belief in a just world can be expected to positively impact subjective well-being, either directly or indirectly, mediated by these implications. There is ample evidence of a positive relationship between just world beliefs and subjective well-being. Moreover, research has shown that the belief in a personal just world is more important than the general just world belief in explaining well-being (e.g., Otto et al., 2006), and that this positive association between just world belief and well-being is true for nonvictims (e.g., Dzuka and Dalbert, 2006) and various groups of victims (e.g., Otto et al., 2006). In addition, Dzuka and Dalbert (2007) demonstrated that teachers' well-being was positively associated with their belief in a just world and that this relationship held when exposure to student violence was controlled. This study is one of the few to have found evidence for a buffering effect of the just world belief: It was only among teachers with a weak just world belief that exposure to violence was associated with more negative affect; exposure to violence did not explain negative affect among those with a strong just world belief.

A personal resource can be defined as a personal disposition that helps people to cope with the events of their daily life. The stronger the resource, the better equipped they are to cope. A personal resource thus implies a main effect hypothesis. A personal buffer, in contrast, is usually seen as a resource that takes effect only under specific adverse conditions. A buffer thus implies a moderator hypothesis and it moderates the association between strain and outcome. Overall, research findings are very much in line with the resource hypothesis and do not support the buffer hypothesis. The just world belief should thus be seen as personal resource helping to sustain the well-being of people of all ages in diverse situations, victims and nonvictims alike.

Conclusion

Just world research has shown that people need to believe in justice, and that they strive for justice in order to maintain their basic belief in a just world (e.g., Lerner and Miller, 1978). This justice motive is reflected by an interindividually varying belief in just world and this disposition explains the differences in people's striving for justice as an end in itself, including their own behavior and assimilation of observed or experienced injustices. In return, the justice motive endows trust in the justice of the world and in being treated justly by others.

The basic idea of the just world hypothesis is that people confronted with injustices suffer and feel the unconscious need to restore justice (e.g., Lerner and Simmons, 1966). As a consequence, the belief in a just world particularly impacts intuitive justice-driven reactions such as, for example, the assimilation of injustice. Thus, research suggests that the belief

in a just world is an essential but unconscious source of responses to injustice, in line with the role of other implicit human motives (McClelland et al., 1989). Justice motive theory (Dalbert, 2001) thus interprets the belief in a just world as indicating an implicit justice motive. Lerner and Goldberg (1999) argue that conscious and intuitive justice-driven reactions coexist, and may be elicited simultaneously in the same situation. The belief in a just world seems to operate on an unconscious level and can thus be expected to better explain intuitive than conscious reactions to injustice (Dalbert, 2001). Important challenges for future research on the just world construct include integrating just world research within such a broader framework of human motives and differentiating between the explanation of more controlled versus more intuitive justice-driven reactions in the light of just world reasoning.

See also: Academic Achievement Motivation, Development of; Aggression, Social Psychology of; Authoritarian Personality; Classroom Climate; Justice: Social Psychological Perspectives; Neuroticism; Prosocial Behavior and Empathy; Religiosity; Risk: Social Psychological Perspectives; Social Psychology; Subjective Wellbeing, Psychology of.

Bibliography

Allen, M.W., Ng, S.H., Leiser, D., 2005. Adult economic model and values survey: crossnational differences in economic beliefs. Journal of Economic Psychology 26, 159–185.

Bègue, L., 2002. Beliefs in justice and faith in people: just world, religiosity and interpersonal trust. Personality and Individual Differences 32, 375–382.

Bègue, L., Muller, D., 2006. Belief in a just world as moderator of hostile attributional bias. British Journal of Social Psychology 45, 117–126.

Berscheid, E., Walster, E., 1967. When does a harm-doer compensate a victim? Journal of Personality and Social Psychology 6, 435–441.

Bierhoff, H.W., 1994. Verantwortung und altruistische Persönlichkeit [Responsibility and altruistic personality]. Zeitschrift für Sozialpsychologie 25, 217–226.

Bierhoff, H.W., Klein, R., Kramp, P., 1991. Evidence for the altruistic personality from data on accident research. Journal of Personality 59, 263–280.

Comer, R., Laird, J.D., 1975. Choosing to suffer as a consequence of expecting to suffer: why do people do it? Journal of Personality and Social Psychology 32, 92–101.

Correia, I., Vala, J., 2004. Belief in a just world, subjective well-being and trust of young adults. In: Dalbert, C., Sallay, H. (Eds.), The Justice Motive in Adolescence and Young Adulthood: Origins and Consequences. Routledge, London, UK, pp. 85–100.

Correia, I., Vala, J., Aguiar, O., 2007. Victim's innocence, social categorization, and the threat to the belief in a just world. Journal of Experimental Social Psychology 43, 31–38.

Cubela Adoric, V., Kvartuc, T., 2007. Effects of mobbing on justice beliefs and adjustment. European Psychologist 12, 261–271.

Dalbert, C., 1992. Der Glaube an die gerechte Welt: Differenzierung und Validierung eines Konstrukts [The belief in a just world: differentiation and validation of a construct]. Zeitschrift für Sozialpsychologie 23, 268–276.

Dalbert, C., 1997. Coping with an unjust fate: the case of structural unemployment. Social Justice Research 10, 175–189.

Dalbert, C., 1999. The world is more just for me than generally: about the personal belief in a just world scale's validity. Social Justice Research 12, 79–98.

Dalbert, C., 2001. The Justice Motive as a Personal Resource: Dealing with Challenges and Critical Life Events. Kluwer Academic/Plenum Publishers, New York.

Dalbert, C., 2010. Embitterment – from the perspective of justice psychology. In: Linden, M., Maercker, A. (Eds.), Embitterment: Societal, Psychological, and Clinical Perspectives. Springer, New York, pp. 30–41.

Dalbert, C., Dzuka, J., 2004. Belief in a just world, personality, and well-being of adolescents. In: Dalbert, C., Sallay, H. (Eds.), The Justice Motive in Adolescence and Young Adulthood: Origins and Consequences. Routledge, London, UK, pp. 101–116.

Dalbert, C., Filke, E., 2007. Belief in a just world, justice judgments, and their functions for prisoners. Criminal Justice and Behavior 34, 1516–1527.

Dalbert, C., Lipkus, I.M., Sallay, H., Goch, I., 2001. A just and an unjust world: structure and validity of different world beliefs. Personality and Individual Differences 30, 561–577.

Dalbert, C., Stoeber, J., 2006. The personal belief in a just world and domain-specific beliefs about at school and in the family: a longitudinal study with adolescents. International Journal of Behavioral Development 30, 200–207.

Dette, D., Stöber, J., Dalbert, C., 2004. Belief in a just world and adolescents' vocational and social goals. In: Dalbert, C., Sallay, H. (Eds.), The Justice Motive in Adolescence and Young Adulthood: Origins and Consequences. Routledge, London, UK, pp. 11–25.

Donat, M., Umlauft, S., Dalbert, C., Kamble, S.V., 2012. Belief in a just world, teacher justice, and bullying behavior. Aggressive Behavior 38, 185–193.

Dzuka, J., Dalbert, C., 2006. The belief in a just world's impact on subjective well-being in old age. Aging and Mental Health 10, 439–444.

Dzuka, J., Dalbert, C., 2007. Student violence against teachers: teachers' well-being and the belief in a just world. European Psychologist 12, 253–260.

Edlund, J.E., Sagarin, B.J., Johnson, B.S., 2007. Reciprocity and the belief in a just world. Personality and Individual Differences 43, 589–596.

Fetchenhauer, D., Jacobs, G., Belschak, F., 2005. Belief in a just world, causal attributions, and adjustment to sexual violence. Social Justice Research 18, 25–42.

Furnham, A., 1993. Just world beliefs in twelve societies. Journal of Social Psychology 133, 317–329.

Furnham, A., 2003. Belief in a just world: research progress over the past decade. Personality and Individual Differences 34, 795–817.

Furnham, A., Procter, E., 1989. Belief in a just world: review and critique of the individual difference literature. British Journal of Social Psychology 28, 365–384.

Hafer, C.L., 2000. Investment in long-term goals and commitment to just means drive the need to believe in a just world. Personality and Social Psychology Bulletin 26, 1059–1073.

Hafer, C.L., Bègue, L., 2005. Experimental research on just-world theory: problems, development, and future challenges. Psychological Bulletin 131, 128–167.

Hafer, C.L., Bogaert, A.F., McMullen, S.L., 2001. Belief in a just world and condom use in a sample of gay and bisexual men. Journal of Applied Social Psychology 31, 1892–1910.

Hafer, C.L., Correy, B.L., 1999. Mediators of the relation of beliefs in a just world and emotional responses to negative outcomes. Social Justice Research 12, 189–204.

Jost, J.T., Banaji, M.R., Nosek, B.A., 2004. A decade of system justification theory: accumulated evidence of conscious and unconscious bolstering of the status quo. Political Psychology 25, 881–919.

Kahileh, R., Peter, F., Dalbert, C., 2013. Gerechte-Welt-Glaube, Gerechtigkeitserleben und Schulerfolg [Belief in a just world, justice experience, and academic achievement]. In: Dalbert, C. (Ed.), Gerechtigkeit in der Schule. Springer VS, Wiesbaden, pp. 55–71.

Lambert, A.J., Burroughs, T., Nguyen, T., 1999. Perceptions of risk and the buffering hypothesis: the role of just world beliefs and right-wing authoritarianism. Personality and Social Psychology Bulletin 25, 643–656.

Lerner, M.J., 1977. The justice motive: some hypotheses as to its origins and forms. Journal of Personality 45, 1–52.

Lerner, M.J., Goldberg, J.H., 1999. When do decent people blame victims? the differing effects of the explicit/rational and implicit/experiential cognitive system. In: Chaiken, S., Trope, Y. (Eds.), Dual-Process Theories in Social Psychology. Guilford, New York, pp. 627–640.

Lerner, M.J., Matthews, J., 1967. Reactions to suffering of others under conditions of indirect responsibility. Journal of Personality and Social Psychology 5, 319–325.

Lerner, M.J., Miller, D.T., 1978. Just world research and the attribution process: looking back and ahead. Psychological Bulletin 85, 1030–1051.

Lerner, M.J., Simmons, C.H., 1966. The observer's reaction to the "innocent victim": compassion or rejection? Journal of Personality and Social Psychology 4, 203–210.

Lind, E.A., van den Bos, K., 2002. When fairness works: toward a general theory of uncertainty management. Research in Organizational Behavior 24, 181–223.

Lipkus, I.M., Dalbert, C., Siegler, I.C., 1996. The importance of distinguishing the belief in a just world for self versus for others: implications for psychological well-being. Personality and Social Psychology Bulletin 22, 666–677.

Lucas, T., Alexander, S., Firestone, I., LeBreton, J., 2007. Development and initial validation of a procedural and distributive just world measure. Personality and Individual Differences 43, 71–82.

McClelland, D.C., Koestner, R., Weinberger, J., 1989. How do self-attributed and implicit motives differ? Psychological Review 96, 680–702.

Messick, D.M., Bloom, S., Boldizar, J.P., Samuelson, C.D., 1985. Why are we fairer than others? Journal of Experimental Social Psychology 21, 480–500.

Montada, L., Schmitt, M., Dalbert, C., 1986. Thinking about justice and dealing with one's own privileges: a study of existential guilt. In: Bierhoff, H.W., Cohen, R.L., Greenberg, J. (Eds.), Justice in Social Relations. Plenum Press, New York, pp. 125–143.

Otto, K., Dalbert, C., 2005. Belief in a just world and its functions for young prisoners. Journal of Research in Personality 39, 559–573.

Otto, K., Boos, A., Dalbert, C., Schöps, D., Hoyer, J., 2006. Posttraumatic symptoms, depression, and anxiety of flood victims: the impact of the belief in a just world. Personality and Individual Differences 40, 1075–1084.

Peter, F., Dalbert, C., 2013. Die Bedeutung der LehrerInnengerechtigkeit: Klimaerleben oder persönliches Erleben? [The meaning of teacher justice: climate experience or personal experience?] In: Dalbert, C. (Ed.), Gerechtigkeit in der Schule. Springer VS, Wiesbaden, pp. 33–53.

Ross, M., Miller, D.T. (Eds.), 2002. The Justice Motive in Everyday Life. Cambridge University Press, Cambridge, UK.

Rubin, Z., Peplau, L.A., 1973. Belief in a just world and reaction to another's lot: a study of participants in the national draft lottery. Journal of Social Issues 29 (4), 73–93.

Sallay, H., 2004. Entering the job market. Belief in a just world, fairness and well-being of graduating students. In: Dalbert, C., Sallay, H. (Eds.), The Justice Motive in Adolescence and Young Adulthood: Origins and Consequences. Routledge, London, UK, pp. 215–230.

Sutton, R.M., Winnard, E.J., 2007. Looking ahead through lenses of justice: the relevance of just-world beliefs to intentions and confidence in the future. British Journal of Social Psychology 46, 649–666.

Tomaka, J., Blascovich, J., 1994. Effects of justice beliefs on cognitive, psychological, and behavioral responses to potential stress. Journal of Personality and Social Psychology 67, 732–740.

Umlauft, S., Dalbert, C., Schröpper, S., 2013. Die Bedeutung des Gerechtigkeitserlebens für das Exklusionsempfinden [The meaning of justice experience for the feeling of exclusion]. In: Dalbert, C. (Ed.), Gerechtigkeit in der Schule. Springer VS, Wiesbaden, pp. 109–125.

Zuckerman, M., 1975. Belief in a just world and altruistic behavior. Journal of Personality and Social Psychology 31, 972–976.

Belief, Anthropology of

Byron Good, Faculty of Arts and Sciences, Harvard Medical School, Boston, MA, USA
Andrew McDowell, McGill International TB Centre, McGill Anthropology, Montreal Chest Institute, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Belief,' along with 'believe, beliefs,' and 'belief systems,' has served as a kind of 'odd job' word for anthropologists: a word commonly used for the analysis of a society's culture, religion, or ideas about the world, but seldom defined or explicitly theorized. Belief is often included in omnibus definitions of culture, such as Tylor's classic "Culture, or civilization, … is that complex whole, which includes knowledge, belief, art, law, morals, custom, and any other capabilities and habits acquired by man as a member of society" (1871, from Kroeber and Kluckhohn, 1952, p. 43) or Sapir's "culture, that is, … the socially inherited assemblage of practices and beliefs that determines the texture of our lives …" (1921, Kroeber and Kluckhohn, 1952, p. 47). The term often refers explicitly to religion, as in Tylor's minimum definition of religion as 'the belief in Spiritual Beings' (1873), but is used equally for claims people make about the empirical world or their social institutions.

Accounts of central tenets of culture or worldview in a society are often reported as beliefs. Lawrence's (1964) discussion of cargo cults in New Guinea is typical: "The Ngaing believed that their world was brought into being by their deities" (p. 16); "all these peoples believed in a creator God" (p. 21); "the natives believed they could obtain cargo … largely by ritual" (p. 235). As Needham (1972, p. 3) points out, Lawrence did not find it necessary to define belief or describe how he knows "that men in New Guinea believe what they say"; he simply assumed that this "psychological category in the English language denotes a common human capacity." This is typical of anthropological writing at the time. Despite its ubiquity, 'belief' seldom appears in the Index of ethnographies, and the *International Encyclopedia of the Social Sciences* (Sills, 1968) has no entry for Belief, simply referring the reader to '*see* Attitudes; Myth and Symbol; Religion, Anthropology of; Values.'

Despite a long history of analyzing other societies through a systematic description of their 'beliefs and behaviors,' by the 1980s ethnographic accounts had begun to reflect ironically on belief as a shifting anthropological category. "Take an ethnographer," writes Jean Favret-Saada (1980, p. 4):

> She has spent more than thirty months in the Bocage in Mayenne, studying witchcraft … 'Tell us about the witches,' she is asked again and again when she gets back to the city … confirm that *out there* there are some people who can bend the laws of causality and morality, who can kill by magic and not be punished; but remember to end by saying that they do not really have that power: they only believe it because they are credulous, backward peasants … No wonder that country people in the West are not in any hurry to step forward and be taken for idiots …

Favret-Saada concludes: "To say that one is studying beliefs about witchcraft is automatically to deny them any truth: it is just a belief, it is not true."

'Belief' is thus a ubiquitous term in much of the anthropological corpus, a commonsense category used for the analysis of the tenets of culture or religion. It is, however, a term that is increasingly contested, and its use has declined significantly in favor of focus on practice, narrative, and metaphor. A review of the meaning and history of the term belief is needed to make sense of its complex genealogy in anthropological usage.

Defining 'Belief'

The Oxford English Dictionary (1989) defines 'belief' as:

1. The mental action, condition, or habit, of trusting to or confiding in a person or thing; trust, dependence, reliance, confidence, faith.
2. Mental acceptance of a proposition, statement, or fact, as true, on the ground of authority or evidence …
3. The thing believed; the proposition or set of propositions held true …

Hahn (1973, p. 208) attempts to clarify anthropological understandings of belief, defining beliefs as "general propositions about the world (consciously) held to be true." Tooker (1992, p. 808) points out that such anthropological conceptions of belief typically have two elements: 'propositionality' ('a mental state or conviction in which a doctrine or proposition concerning one's worldview is affirmed as true as opposed to false') and an assumption that such a 'prepositional relationship to tradition is an 'interiorized' one because of its reference to mental states.'

The *OED* indicates that the definition of belief as a proposition held to be true, a definition apparently taken over by anthropologists, rests on an older definition of belief as the condition of trusting in a person or thing, as 'faith.' At the same time, while a 'belief' is defined as a 'mental acceptance of a proposition … as true,' Favret-Saada and others suggest that describing a proposition as a 'belief' implies doubt or even falsehood, that "studying beliefs about witchcraft is automatically to deny them any truth." A review of the history of the semantics of 'belief' suggests that this complexity is grounded in changes in the meaning and use of the terms 'believe' and 'belief' in English (and their counterparts in other European languages).

A History of 'Belief' in Euro-American Language Use

Anthropologists often talk with members of other societies about some aspect of 'their' world which does not exist in 'ours' and which anthropologists are comfortable asserting is not part of empirical reality. An examination of anthropological texts suggests that such phenomena – the witches of the Bocage or

the Azande, the 'cargo' of the Papua New Guinea cults – are discussed in the language of belief – and that the beliefs of others are juxtaposed to 'our' knowledge of the empirical world (see Good, 1994, Chapter 1). Western civilization has long given great importance to 'beliefs,' with wars, church schisms, persecutions, and martyrdom revolving around credos and demands for the renunciation of false beliefs. How is it that this same term, 'belief,' came to be so central to anthropological analysis, and what is implied by the juxtaposition of belief and knowledge?

The richest discussion of the history of the concept belief is to be found in two books by Wilfred Cantwell Smith, the historian of religion, which explore the relation between 'belief' and 'faith' historically and across religious traditions. Through historical and linguistic analysis, Smith comes to the startling conclusion that "the idea that believing is religiously important turns out to be a modern idea," and that the meaning of the English words 'to believe' and 'belief' have changed dramatically, since around 1700, leading to misunderstanding of both the Christian tradition and the religious faith of others.

Smith demonstrates that the Old English words that evolved into the modern 'believe' (*geleofan*, *geliefan*) meant 'to belove,' 'to hold dear,' 'to cherish,' 'to regard as lief,' and reflect the Latin root *libet*, 'it pleases,' or *libido*, 'pleasure' (Smith, 1977, pp. 41–46; Smith, 1979, pp. 105–127). In Chaucer's *Canterbury Tales*, the words 'accepted by bileve' mean simply "accept my loyalty; receive me as one who submits himself to you." Thus, Smith argues that "belief in God" originally meant "a loyal pledging of oneself to God, a decision and commitment to live one's life in His service" (Smith, 1977, p. 42). Its counterpart in the medieval language of the Church and in the Credo was "I renounce the Devil," belief and renunciation being parallel and contrasting performatives or actions, rather than states of mind. Today, however, Smith argues, the ritual acclamation "I believe in God" means for many people "Given the uncertainty as to whether there be a God or not, as a fact of modern life, I announce that my opinion is 'yes, I judge God to be existent" (1977, p. 44).

Smith argues that this change in the language of belief can be traced in the grammar and semantics of English literature and philosophy, as well as in popular usage. Three changes serve as indicators of the changing semantics of the verb 'to believe.' First, Smith finds that grammatically, the *object* of the verb 'to believe' shifted from a person (whom one trusted or had faith in), to a person and his word (his virtue accruing to the trustworthiness of his word), to a proposition. This latter shift began to occur by the end of the seventeenth century, with Locke, and was firmly represented by the mid-nineteenth century in John Stuart Mills' philosophy (Smith, 1977, p. 48). The twentieth century has seen a further shift, as beliefs have come to mean 'presuppositions,' and it was in this sense that 'belief systems' became nearly synonymous with 'culture' in some strands of anthropology (see Black, 1973).

A second shift has occurred in the *subject* of the verb 'to believe,' from an almost exclusive use of the first person – 'I believe' – to the predominant use of the third person, 'he believes' or 'they believe.' In anthropology, the impersonal 'it is believed that' parallels the discussion of culture as belief system or system of thought. This change in subject subtly shifts the nature of the speech act involved – from existential or performative ('I believe') to descriptive ('the Ngaing believe that').

Third, Smith observes that an important and often unrecognized change has occurred in 'the relation of belief to truth and knowledge,' as these are historically conceived. Bacon wrote in 1625 of "the belief of truth," which he defined as the "enjoyment of it." Belief maintains its sense here of holding dear, of appropriating to oneself that which is recognized as true. By the nineteenth century, however, 'to believe' had come to connote doubt, and today it suggests outright error or falsehood. Knowledge requires both certitude and correctness; belief implies uncertainty, error, or both. Smith's favorite illustration of the juxtaposition of belief and knowledge is an entry in the Random House dictionary which defines 'belief' as "an opinion or conviction," and at once illustrates this with 'the belief that the earth is flat'! It is virtually unacceptable in popular English usage to say that members of some society 'believe' the earth is round; if this is part of their worldview, then it is knowledge, not belief.

Smith goes on to argue that failure to recognize this shift in meaning has led to mistranslation of texts in the Christian tradition and ultimately to 'the heresy of believing,' the deeply mistaken view that belief in this modern sense is the essence of the religious life rather than faith. Equally important, for Smith, the misplaced focus on belief as the primary dimension of religious life has led to mistranslations and misunderstandings of other religious traditions. It is this claim that has special relevance for anthropology.

'Beliefs' and 'Believing' in Diverse Anthropological Traditions

The term 'belief' functions quite differently in diverse anthropological traditions. A brief review of several traditions will indicate why 'belief' has emerged as a contested term in recent anthropological writing.

The Intellectualist Tradition and Rationality Debates

'Beliefs' and 'rationality' are central terms in the Intellectualist tradition most closely associated with E.B. Taylor and J.G. Frazer, as well as with a series of 'Neo-Tylorians' and a set of rationality debates in the 1970s and 1980s. Tylor and Frazer set out to trace the evolution of beliefs from our 'rude forefathers' to the present, investigating forms of rationality, errors in reasoning, and the history of the evolution of culture from magic to religion and the emergence of science. An example from Frazer's Gifford Lectures, published as *The Belief in Immortality* (Frazer, 1913), will illustrate the logic of this program, as well as the evolutionist language that led to its rejection:

> … the question of the validity or truth of religious creeds cannot, perhaps be wholly dissociated from the question of their origin. If, for example, we discover that doctrines which we had accepted with implicit faith from tradition have their close analogies in the barbarous superstitions of ignorant savages, we can hardly help suspecting that our own cherished doctrines may have originated in the similar superstitions of our rude forefathers; … The doubt thus

> cast on our old creed is perhaps illogical, since even if we should discover that the creed did originate in more superstition, … this discovery would not really disprove the beliefs themselves, for it is perfectly possible that a belief may be true, though the reasons alleged in favor of it are false and absurd …

Beliefs in this tradition were investigated as rational propositions and forms of explanation for natural and religious phenomena. The intellectualist paradigm was grounded in philosophical empiricism, and judgments concerning validity of propositions and the evolution of explanatory frames from those which were false, even absurd, to those which correctly represent the natural world were essential to the analysis.

British structural–functionalism explicitly rejected the search for origins, the evolutionary claims, and the focus on individual beliefs and their rationality as a basis for social analysis. Nonetheless, classic texts, such as Evans-Pritchards' *Witchcraft, Oracles and Magic among the Azande* (1937) maintained the basic structure of inquiry, investigating witchcraft as a form of rational explanation for misfortune and distinguishing those explanations which attribute 'supra-sensible qualities' to phenomena (described as beliefs) from those which accord to contemporary science (described as knowledge) (cf Good, 1994, pp. 11–13).

In the 1970s and 1980s, the intellectualist tradition was revived in British anthropology under the rubric of Neo-Tylorianism and the rationality debates (e.g., Wilson, 1970; Hollis and Lukes, 1982). Reacting to the perceived vacuousness of functionalist explanations, a group of anthropologists returned to investigations of the rational structure of non-Western religious systems and their similarities to science as systems of explanation for the natural world. At the heart of these debates was a central question: "when I come across a set of beliefs which appear prima facie irrational, what should be my attitude toward them?" (Lukes, 1970, p. 194). For example, "what does one do with the exotic utterances of primitives, who may say such things as 'my brother is a cockatoo,'" asks Turner (1979, p. 401). "Some things, such as a flat contradiction, could not be 'believed,' in any useful sense of 'believe.' … Can the fact that these utterances cannot be translated into anything that we would call 'rational' warrant a claim that, e.g., these primitives possess a non-logical mentality?" 'Beliefs' in this paradigm were understood as rational propositions, held to be true by members of a society, and analysis focused on the identification, translation, and rational structure of belief systems, examining characteristics of systems 'open' to falsification, and thus to change, in contrast to 'closed systems,' which remain impervious to contradictory data that would lead to change and evolution in the direction of more rational explanation. Needham (1972), however, criticized this view of belief, arguing that belief lacked cross-cultural equivalents. His empiricist perspective also suggested that anthropology has no tools by which to know if informants actually believe what they report to believe. Despite his rejection of the concept, his criticism maintains a separation between belief and knowledge, holding belief as having only an incidental relationship to the known or knowable.

Structural Functionalism

Reacting against intellectualists' search for origins and focus on individual states of minds, Radcliffe-Brown and generations of British anthropologists placed investigations of the structure of social institutions at the center of anthropological theorizing. In this context, individual beliefs are of little interest, and a focus on the relationship between rational propositions and the natural world is replaced by an interest in the structural relation between the symbolic and the social order. Leach made this explicit: "To ask questions about the content of beliefs which are not contained in the content of ritual is nonsense … Ritual action, and belief are alike to be understood as forms of symbolic statement about the social order" (1954, p. 14).

The Boasian Tradition

Inheritors of the German philosophical romantics' theorizing about culture and its *Geist*, American anthropologists of the Boasian tradition were more interested in the 'spiritual possessions of the group rather than the individual' (Sapir, 1924, p. 405), in the role of language and culture in constituting phenomenal experience and the 'behavioral environment' (Hallowell, 1955) of members of a society. Although the term 'belief' appears in this tradition, it has little theoretical salience. Sapir notes this explicitly: " … the cultural conception we are now trying to grasp aims to embrace in a single term those general attitudes, views of life, and specific manifestations of civilization that give a particular people its distinctive place in the world. Emphasis is put not so much on what is done and believed by a people as on how what is done and believed functions in the whole life of that people, on what significance it has for them" ([1924]1970, p. 83). The broad rejection of 'belief' as a common category for cultural analysis carries over into the Geertzian interpretive tradition, into studies of narrative, embodiment, and experience, as well as into the most psychological anthropology, particularly in the American context.

Cognitive Anthropology

The term 'belief' appeared as a core analytic category in American anthropology beginning in the 1950s among anthropologists intent on creating a new, more scientific paradigm for studying culture, who gathered under the banner of 'The New Ethnography,' ethnosemantics, or cognitive anthropology. For cognitivists, "cultures … are not material phenomena; they are cognitive organization of material phenomena. … The object of study is not these material phenomena but the way they are organized in the minds of men" (Tyler, 1969, p. 3). In this tradition, culture is conceived of as generative models, and beliefs are those cognitive structures which take propositional form: "Beliefs are propositions about the relations among things to which those who believe have made some kind of commitment … for pragmatic or emotional reasons" (Goodenough, 1990, p. 597). Although 'beliefs' assume a core analytic function in this tradition, many cognitivists share more with the Boasian tradition than with the intellectualist tradition, in that they are interested in investigating how semantic categories and

linguistic structures shape perception of the phenomenal world rather than in normative claims about rationality. (They differ sharply in methodology and forms of ethnographic writing.) 'Beliefs' and 'knowledge' are often used interchangeably in this tradition, rather than explicitly juxtaposed (see Black, 1973). Nonetheless, the term 'belief' often subtly connotes counterfactuality in this tradition as in others.

Medical Anthropology, Public Health, and Health Psychology

Studies of health beliefs was a key element of early forms of medical anthropology, particularly those devoted to applied work in the public health field, although this has not been true in recent years (see Good, 1994). However, analyses of beliefs about illness etiologies, the risks of particular behaviors, or the benefits of particular treatments are still ubiquitous in public health work and health psychology. Popular or folk health beliefs were explicitly juxtaposed to medical knowledge, and educational campaigns are devoted to correcting mistaken ideas with the hope that this will produce more rational behavior.

A review of recent medical anthropology literature finds this kind of discussion of health beliefs increasingly uncommon. Those who talk about health beliefs tend to do so in an explanatory model format and understand belief as a template for action (Erickson, 2007). Most medical anthropologists today focus instead on narrative, political, and moral economies of health, and experience rather than on health beliefs. Moral anthropology and studies of care giving (Kleinman, 1999, Fassin, 2012) too have moved away from a beliefs-related framework to one which emphasizes personal and social priorities exhibited not by statements of conviction but through practice.

Anthropology of Religion

Although other fields of anthropology have slowly dismissed analyses of belief, anthropology of religion has sustained a healthy debate. Many pick up Needham's skepticism, and question the empirical utility of studying belief (Lindquist and Coleman, 2008). Yet, proponents of belief see it as an opportunity to analyze tensions created by uncertainty, syncretism, and change (Carlisle and Simon, 2012). Convictional belief, for them, becomes a map of action and values in lived experience (Handelman, 2008), and as such a perspective on subjectivity (Simon, 2012). Still others consider convictional belief from Weberian and Durkheimian perspectives focusing on the intellectual and emotional origins and experience of belief, respectively (Lindholm, 2012). Kirsch, however, suggests that belief is best understood as a performative – necessary for ritual efficacy but neither a conviction nor permanent (2004). Engagements with Islam tend to follow Asad and examine practices of piety rather than belief (1993), but this emphasis on practice over convictional belief is not wholesale. Belief as conviction has been a focus of anthropology of religion, but propositional belief seems hard to find, leaving the question of separating knowledge and belief for the faithful to answer.

Critical Reflections

The language of belief has been widely used for ethnographic descriptions of culture, dating back to at least the nineteenth century. In some anthropological traditions it has been a key analytic concept; in others, it has been used infrequently and in a commonsense way. In some of the behavioral sciences, 'beliefs' continues to be used with little self-consciousness. However, in each of these domains, 'beliefs' are explicitly or implicitly juxtaposed to 'knowledge,' and the use of 'believe' or 'beliefs' connotes assertions to be counter to what is 'known' about empirical reality. This is particularly true in fields such as medicine, where the knowledge claims of the natural sciences run headlong into the historicist convictions of anthropology (Good, 1994).

In recent years, the claims implicit in this formulation have been sharply challenged. The representation of others' culture as 'beliefs' ('their' beliefs and 'our' knowledge) authorizes the position and knowledge claims of the anthropological observer. Such authority has been challenged by a variety of forms of critical theory – post-structuralism, feminist writing, postcolonial theorizing, and subaltern studies. The place of the ethnographer as an objective scientific observer, both in research and in ethnographic texts, is increasingly contested, and along with it, the use of the language of 'belief.' At the same time, developments in the history, philosophy, and sociology of sciences challenge any views of science as a progressive, normative 'mirror of nature,' the arbiter between belief and knowledge. As a consequence, the labeling of others' assertions about the world as 'beliefs' is increasingly challenged as pejorative, as failing to recognize the 'local' character of knowledge, and as hegemonic. Thus, although belief is still a word used unreflexively by some, its primacy as a mode of understanding culture, social life, and human action is increasingly uncommon.

See also: Anthropology and History; Attitudes and Behavior; Boas, Franz (1858–1942); Cognitive Anthropology; Cognitive Archaeology; Common Sense, Anthropology of; Cultural Relativism, Anthropology of; Evans-Pritchard, Sir Edward E. (1902–73); Interpretation in Anthropology; Knowledge, Anthropology of; Performance: Anthropological Aspects; Psychological Anthropology; Religion, Anthropology of; Religion, Sociology of; Ritual; Values, Anthropology of; Worldview.

Bibliography

Asad, T., 1993. Genealogies of Religion: Discipline and Reasons of Power in Christianity and Islam. Johns Hopkins University Press, Baltimore.

Black, M.B., 1973. Belief systems. In: Honigmann, J.J. (Ed.), Handbook of Social and Cultural Anthropology. Rand McNally College Publishing Co., Chicago, pp. 509–577.

Carlisle, S., Simon, G., 2012. Believing selves: negotiating social and psychological experiences of belief. Ethos 40 (3), 221–236.

Erickson, B., 2007. Toxin or medicine? Explanatory models of radon in Montana health mines. Medical Anthropology Quarterly 21 (1), 1–21.

Fassin, D., 2012. A Companion to Moral Anthropology. Wiley-Blackwell, Chichester, UK.

Favret-Saada, J., 1980. Deadly Words. Witchcraft in the Bocage. Cambridge University Press, Cambridge, UK.
Frazer, J.G., 1913. The Belief in Immortality, vol. 1. Macmillan and Co., London.
Good, B.J., 1994. Medicine, Rationality, and Experience: An Anthropological Perspective. Cambridge University Press, Cambridge, UK.
Goodenough, W.H., 1990. Evolution of the human capacity for beliefs. American Anthropologist 92, 597–612.
Hahn, R.A., 1973. Understanding beliefs: an essay on the methodology of the statement and analysis of belief systems. Current Anthropology 14, 207–229.
Handelman, D., 2008. Returning to cosmology – thoughts on the positioning of belief. Social Analysis 52 (1), 181–195.
Hallowell, I., 1955. Culture and Experience. University of Pennsylvania Press, Philadelphia.
Hollis, M., Lukes, S., 1982. Rationality and Relativism. Basil Blackwell, Oxford, UK.
Kleinman, A., 1999. Experience and its Moral Modes: Culture, Human Conditions, and Disorder. The Tanner Lectures. University of Utah Press, Salt Lake City.
Kirsch, T., 2004. Restaging the will to believe: religious pluralism anti-syncretism, and the problem of belief. American Anthropologist 106 (4), 699–709.
Kroeber, A.L., Kluckhohn, C., 1952. Culture: A Critical Review of Concepts and Definitions. Papers of the Peabody Museum of American Archaeology and Ethnology, Harvard University, vol. 47, no. 1, Cambridge, MA.
Lawrence, P., 1964. Road Belong Cargo. Manchester University Press, Manchester, UK.
Lindholm, C., 2012. "What is bread?" The anthropology of belief. Ethos 40 (3), 341–357.
Lindquist, G., Coleman, S., 2008. Against belief. Social Analysis 52 (1), 1–18.
Lukes, S., 1970. Some problems about rationality. In: Wilson, B.R. (Ed.), Rationality. Harper & Row, New York, pp. 194–213.
Needham, R., 1972. Belief, Language, and Experience. University of Chicago Press, Chicago.
Sapir, E., 1924. Culture genuine and spurious. American Journal of Sociology 29 (4), 401–429.
Sills, D., 1968. International Encyclopedia of the Social Sciences. Macmillan Co. and The Free Press, New York.
Simon, G.M., 2012. Conviction without being convinced: maintaining Islamic certainty in Minangkabau, Indonesia. Ethos 40 (3), 237–257.
Simpson, J.A., Weiner, E.S.C., 1989. The Oxford English Dictionary. Clarendon Press, Oxford, UK.
Smith, W.C., 1977. Belief and History. University of Virginia Press, Charlottesville, VA.
Smith, W.C., 1979. Faith and Belief. Princeton University Press, Princeton, CT.
Tooker, D.E., 1992. Identity systems of Highland Burma: belief, Akha Zan, and a critique of interiorized notions of ethno- religious identity. Man 27, 299–819.
Turner, S.P., 1979. Translating ritual beliefs. Philosophy of the Social Sciences 9, 401–423.
Tyler, S., 1969. Cognitive Anthropology. Holt, Rinehart and Winston, New York.
Tylor, E.B., 1873. Primitive Culture, 2 vols. John Murray, London.
Wilson, B.R., 1970. Rationality. Harper and Row, New York.

Bell, Daniel (1919–2011)

Howard Brick, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Daniel Bell's books on social theory, cultural criticism, and political affairs included *The End of Ideology* (1960), *The Coming of Post-Industrial Society* (1973), and *The Cultural Contradictions of Capitalism* (1976). A world-renowned sociologist, he was also recognized as a leader of 'the New York intellectuals,' who began their careers in left-wing movements of the 1930s and achieved prominence by the 1950s as proponents, mostly, of liberal anticommunism. Bell's support of Cold War containment policy and political moderation, alongside his critique of countercultural hedonism, made his work the target of left-wing criticism, which often obscured his actual long-standing social-democratic convictions.

Daniel Bell (10 May 1919 to 25 January 2011) was both a paramount figure in US sociology after World War II and one of the most distinguished 'public intellectuals' of the postwar period. Best known for his description and analysis of 'postindustrial society,' Bell was, earlier, central to the debate over 'the end of ideology,' a common phrase of the 1950s and 1960s citing the apparent decline in the appeal that extreme left-wing and right-wing political doctrines held for Western European and American intellectuals. At the very beginning of his career, he had been an expert labor journalist and an influential historian of American socialism. After his major works had appeared on postindustrial society and what he regarded as the cultural distemper of the 1960s and 1970s, Bell continued to publish essays through the 1980s and 1990s on technological trends, problems of community and civil society, the relation between religion and politics, the nature of modernity and its discontents. His books continued to appear in new editions into the twenty-first century, and his death in 2011 drew widespread attention in the press and eulogies by much younger writers who hailed him as a mentor. His highly nuanced views on politics, morality, culture, and society left his work open to diverse interpretations.

Early Life

Raised in New York City by his mother, a widowed garment worker, Bell spent parts of his childhood in a 'day orphanage' when times were hard or night work kept his mother away from home. An uncle changed the family name from Bolotsky to Bell around the time, in the depths of the Great Depression, that he joined the Young People's Socialist League, youth wing of the Socialist Party (SP). Even at a young age, his political development manifested the kind of idiosyncratic syntheses that always marked his thought: influenced by an anarchist relative who warned him against Bolshevism (and related the Soviet suppression of the Kronstadt sailors' revolt of 1920), Bell veered to the SP's more conservative, anticommunist faction, dominated by an 'Old Guard' of orthodox, evolutionary Marxists historically most hostile to anarchism. Known after 1936 as the Social Democratic Federation (SDF), the Old Guard faction became Bell's political home. Among his mentors was the Russian-Jewish émigré Samuel (Sol) Levitas, who ran the SDF weekly, the *New Leader*, and later introduced Bell to the circle of exiled Mensheviks in New York City, including Boris Nicolaevsky, David Dallin, and Raphael Abramovich.

Attending the City College of New York (which charged no tuition), Bell frequented the gathering places of the student left, where he was familiar with Trotskyists and other dissidents, many of whom held views to the left of his own. With those circles he shared ardent antiwar views, which motivated annual student strikes and rallies for peace in early April (marking the anniversary of the day in 1917 when Woodrow Wilson brought the United States into the European war). His position savored more of pacifist morals (which he recognized in works by the German exile Ernst Toller) than the typical antiimperialism proclaimed by the far left. He moved toward support of US involvement in World War II only after 1940, under the influence of Sidney Hook, the country's leading Marxist philosopher at the time, who had just broken with SP leader Norman Thomas in order to urge US entry on antifascist grounds. Bell's early familiarity with Hook and the City College radicals gave him a firm grounding in Marxism, which remained a permanent deposit in his thought even after he came, roughly by the mid-1940s, to repudiate it. In the meantime, having completed a master's degree in sociology at Columbia University in 1939, Bell also became acquainted with Max Horkheimer and Theodor Adorno, leaders of the Institute of Social Research (Frankfurt) in exile.

Bell's formal academic training ended with his Columbia master's degree. Levitas brought 21-year-old Bell on the *New Leader* staff in summer 1940 and 2 years later made him managing editor. At the paper, Bell astutely balanced his personal allegiance to the anticommunism of the orthodox Old Guard with friendships among more radical young writers (some still part of a diversifying Trotskyist milieu), residual suspicions of the state in war, and links with other political currents ranging from prolabor and middle-class reformers to Dwight Macdonald's emerging antiwar anarchosocialism. Bell's *New Leader* kept up a drumbeat of criticism against American capitalists, first for dragging their feet in the war mobilization, then for fattening on war profits and thus threatening to concentrate corporate power ever more in league with trends toward state centralization. Thus his political views turned slightly leftward again, paradoxically influenced both by the labor agitation that stirred unionist A.

Philip Randolph and the social-democratic economist Lewis Corey to anticipate the postwar creation of a new American labor party and by the hothouse milieu of self-consciously marginal radical-intellectual estrangement from the corporate war regime. He left the *New Leader* for another editing job in 1944, and in 1945 assailed his old newspaper for promoting a 'war-with-Russia' program and helping to fuel what would soon be called the Cold War, which Bell at first called "a paranoid situation where two delusional systems come into conflict" (quoted in Brick, 1986: p. 149).

The war's end marked the beginning of Bell's transition to academic work. Recruited by the economist (and SP activist) Maynard Krueger, he left New York to spend 3 years teaching at the University of Chicago as an instructor in the general-education social science program that David Riesman was helping to shape. Bell moonlighted, teaching US labor history at Roosevelt University, Chicago's city college, and he had some ties, via Edward Shils, with the incipient disarmament milieu around the *Bulletin of Atomic Scientists*. Most of all, the Chicago social science core program attracted young intellectuals for whom social theory and public affairs were assumed to be intertwined; they included Benjamin Nelson, Philip Rieff, Bert Hoselitz, Milton Singer, and Lewis Coser. While he continued to do a bit of political writing, Bell told Dwight Macdonald that his 'ideas are slowly changing,' and by the time he returned to New York in 1948, his compass had shifted: hired to write the 'Labor' column for *Fortune* magazine, Bell also edited for a short time *Modern Review*, a journal led by the Menshevik Raphael Abramovich, worked on a history of American socialism, and took up a lecturer's position in Columbia University's Sociology department (Brick, 1986: pp. 105, 155–157).

The *Fortune* column, which Bell maintained for 10 years, offered vigorous reporting on the big trends in union organizing and labor's relations with corporate business, as well as serving as a study of labor as a social movement. *Modern Review* soon closed, with remarks by Bell on the loss of vitality in the left's intellectual life, while – writing on both politics and social sciences for the new *Commentary* magazine – Bell had staked out an analysis recognizing Harry Truman's election victory on a 'Fair Deal' platform as the sign of a new social and political formula, one in which Business and Labor both had a seat at the table as government frankly fashioned social policy on a national plane, responsive to large institutionalized interest groups but acting at least to some degree in a directive role that was more than a broker. It was a highly modified version of the 'state monopoly capitalism' he had toyed with at the end of the war, but one he now thought showed more promise for enhancing social equity and political inclusion than either Frankfurt School theorists of 'administered society' or his erstwhile friend C. Wright Mills, in his later critique of 'the power elite,' would acknowledge. Here was his verdict, at the conclusion of his first book, *Marxian Socialism in the United States*:

> The old simplistic theories no longer hold. We seem to be evolving toward some form of technical-military-administrative state, especially as the pressures of a permanent war economy bring into focus a priority of needs which are national in character and override the demands of any particular interest group. The growth of a federal budget from four billion in 1930 to more than forty billion in 1950 (apart from the wartime peak of over ninety billion in 1944) was an unplanned and crescive fact, and yet these new enormous magnitudes are of decisive import in shaping the economy. Along the way, the nascent state capitalism has had to enlarge its social budgets and provide for the welfare of large masses; quondam socialists, now in high positions in labor and government, have tended to instill a sense of social responsibility. But it is not primarily a social welfare state which has developed. In the dimly-emerging social structure, new power sources are being created and new social divisions are being formed. Whatever the character of that new social structure may be—whether state capitalism, managerial society, or corporative capitalism—by 1950 American socialism as a political and social fact had become simply a notation in the archives of history.
>
> Bell, 1996a: p. 193

Thus Bell pronounced his valedictory for the left-wing politics he had known. His history of American socialism, moreover, reflected the Weberian perspective that he, like other young American sociologists of the time, had adopted. Max Weber pinned sociological analysis, as his paramount American exponent, Talcott Parsons, put it, on 'the subjective point of view' of actors: hence it was the mentality or perspective of the socialist movement that drew Bell's attention. Borrowing further from the notion of 'otherworldly asceticism' and from Weber's promotion of an 'ethic of responsibility' in political action as opposed to an 'ethic of ultimate ends,' Bell concluded that American socialism was hobbled, and finally defeated, by its posture of living 'in, but not of the world,' so committed to ideal ends and a misguided moral purity that it could never adapt to practical politics.

The End-of-Ideology Debate

Also, after 1950, Bell helped lead the American Committee on Cultural Freedom (ACCF), the US affiliate of the international Congress for Cultural Freedom (CCF), a liberal anticommunist group intended to rally intellectuals to a common Cold War front against Soviet power and influence. There was some span of opinion among ACCF members, from David Riesman on the 'left' of the group (he resigned, protesting that ACCF had not done enough to defend physicist J. Robert Oppenheimer against red-baiting) to those on the 'right,' like Sidney Hook, who vigorously argued for excluding Communist Party members from teaching positions, and Irving Kristol, who notoriously wrote, "There is one thing that the American people know about Senator McCarthy: he, like them, is unequivocally anti-Communist. About the spokesmen for American liberalism, they feel they know no such thing" (quoted in Pells, 1985: p. 296). Again, Bell occupied a distinctive position on that spectrum. Although friendly with Hook and Kristol, he also collaborated in the mid- and late-1950s with Erich Fromm to promote nuclear disarmament and diplomatic recognition of the People's Republic of China (seeking to defuse Cold War tensions in East Asia), far from conservative views in American life at the time (Friedman and Schreiber, 2013: p. 199).

The CCF's 1955 conference in Milan on 'The Future of Freedom' brought into focus the 'end of ideology,' a theme that had been growing among disenchanted left-wing and liberal intellectuals from the 'discovery of totalitarianism' in the late 1930s to the high pitch of Cold War confrontation in 1950. By

then, Bell, like others, had surrendered his stance of neutrality between 'two delusional systems' and decisively chose to support the world power of the United States and its allies. Writing shortly before the Milan conference, the philosopher of science and conference organizer Michael Polanyi set the tone. His essay in the CCF's English-language organ, *Encounter*, dispensed with the idea that totalitarianism stemmed from economic planning (as free-market 'neo-liberal' Friedrich von Hayek had argued in his 1944 polemic, *The Road to Serfdom*). Rather, Polanyi wrote, totalitarianism arose as a purely political phenomenon when 'civility' (a sense of public responsibility and affective identity nurtured by common traditions) was lacking and when speculative hopes of social reconstruction on a grand scale outran a community's inherited cultural resources. Polanyi was gratified by 'examples of civic reconciliation' in Western Europe since the late 1940s. He noted the signs that "the intensity with which violent political illusions are held had undoubtedly decreased," though he complained that overly excited 'anticapitalists' and 'antisocialists' "are still glaring at each other through the angry masks of obsolete ideologies" (Polanyi, 1955).

Those themes governed the Milan conference. Thus, Edward Shils reported in an essay entitled 'The End of Ideology?' that the last of the doctrinaire formulae in Western politics had indeed fallen: "The full awareness that nationalization is no universal solution to economic problems and that British socialism has not resulted in tyranny have materially weakened the ideologies of thorough-going socialism and thorough-going neo-liberalism" (Shils, 1955). At the same time, Milan signaled a desire among some participants to assuage the ideological passions of the Cold War itself, seeking to make the most of the 'thaw' that appeared possible a few years after Stalin's death. Polanyi argued that the Soviet 'planned economy' in fact functioned "through a system of more or less regulated, and more or less legal, markets." The problems of both world camps were essentially those of a common 'industrial society' (Polanyi, 1955). This was precisely the theme of the conclusion to Raymond Aron's *The Opium of the Intellectuals* (1955), 'The End of the Ideological Age?' Market and plan, Aron wrote, were pure types existing nowhere in reality: "Mixed systems . . . are the normal thing." He wondered why intellectuals still yearned for some abstract, universal truth to realize in action, and he hoped for "the advent of the skeptics" who would "doubt all the models and utopias" (Aron, 1962).

Taking a year's leave from *Fortune*, Bell worked in Paris, 1956–57, planning CCF international seminars on topics like democratic economic planning and the assessment of liberalizing trends in the Soviet Union, as he worked closely with Polanyi, Aron, and C.A.R. Crosland, the leading theorist of 'revisionism' in the British Labor Party. The year's sojourn also secured his personal link to French culture and (centrist) intellectual life, which would lead, years later, to his position as president of the elite Tocqueville Society (1992–2000). His time in Paris coincided with the Hungarian Revolution and suppression of it in fall 1956, and he helped facilitate the escape of dissident intellectuals from the repressive Kadar regime installed by Soviet tanks (Bell, 2000a).

In Bell's judgment, it was the Labor Party's 'New Right' – the current borne by Crosland, Hugh Gaitskell, and Roy Jenkins – that gave the CCF its distinct 'political coloration.' After World War II, Crosland argued in *The Future of Socialism* (1956), "capitalism had been transformed almost out of recognition." Given the reform accomplishments of the post-war Labor government, Crosland claimed, "the dominant emphasis [in public policy] ceases to be on the rights of property, private initiative, competition, and the profit motive; and is transferred to the duties of the state, social and economic security, and the virtues of cooperative action." The demotion of these capitalist standards was complemented by the obsolescence of old, archetypal principles of 'socialism.' Neither more nationalization of production nor complete economic planning at the 'commanding heights' held much conviction among Crosland's like-minded reformers, since high-output mass industry, corresponding mass consumption, Keynesian fiscal management, and fairer norms of meritocracy promised the steady reduction of poverty and a gradual approach to genuine equality of opportunity – a large part of older socialist aims that could be won by nonsocialist, indeed market-based means (Crossman, 1952: p. 42; Crosland, 1957). It was this social-democratic program that informed Daniel Bell's work as it achieved greatest renown following the 1960 publication of *The End of Ideology*. Committed to determinedly moderate reform, he declared that ideologies – totalizing systems of ideas wedded to feverish, mass-based desires for a messianic reinvention of society – had fostered the great calamities of the twentieth century and, thankfully, had now subsided. The same year, Bell's friend Seymour Martin Lipset echoed the theme as a conclusion to his book *Political Man*: "The ideological issues dividing left and right had been reduced to a little more or a little less government ownership and economic planning" (Lipset, 1955, 1960: pp. 170–171).

Both books appeared, fortuitously, at a moment that was believed to open a new era. In his presidential campaign, John Kennedy cleverly seized on the prevailing sense that the 1950s were a sluggish, stolid time and promised rejuvenation and 'Greatness for the Sixties.' Although Bell's antiideological arguments filtered into Kennedy's rhetoric of reform, the coincident surge of civil rights activism and an increasingly vocal 'ban the bomb' sentiment would signal a renewal of dissent reaching a crescendo of protest within a few years. Thus some observers considered Bell and Lipset's 'end of ideology' – insofar as the phrase connoted a historically unique cessation of sharp political conflict between interest groups or between dueling visions of a new future – to be an idea that came 'too late,' obsolete as soon as it appeared in print (Steinfels, 1979: p. 162). Also characteristic of that moment was boom in higher education – the dramatic rise in the numbers and proportion of young people in college, the expansion of university systems, and the consolidation of the disciplinary structure of modern scholarship. American social science gained in prestige and resources, as government funders joined private foundations in research grants and large-scale survey-research centers blossomed at Columbia, Chicago, and Michigan. This heyday in advanced social science research spawned the conviction, fairly prevalent among liberal scholars, that a new Enlightenment had arrived, promising to deliver sound knowledge of social affairs in value-neutral terms (Brick, 2000: pp. 24–33).

These conditions set the stage for a flourishing debate over the end of ideology, focused on the value of consensus or dissent in social and political life, the epistemological status of the social sciences as they become embroiled in contentious political affairs, and the degree to which academic affairs could or should be shielded from political controversy. A leader of the left-wing Students for a Democratic Society, Alan Haber, wrote in 1962 that ideology need not be doctrinaire, millenarian, or violent: rather, ideology comprised a set of moral values that sustained political criticism, a definition of the 'good society,' and a strategic plan of action dedicated to realize it (Waxman, 1969: pp. 182–205). The young Yale political scientist Joseph LaPalombara defined ideology as "a philosophy of history, a view of man's present place in it, some estimate of probable lines of future development, and a set of prescriptions regarding how to hasten, retard, and/or modify that developmental direction" (Waxman, 1969: p. 320). Defined in these terms, ideology's 'end' appeared not at all to be a benign outcome. In fact, C. Wright Mills argued, the end-of-ideology doctrine "stands for the refusal to work out an explicit political philosophy"; it amounted to an "intellectual celebration of apathy" based upon "a disbelief in the shaping by men of their own futures." Moreover, 'end-of-ideology' notions forgot that every social theory presupposed an 'ideological' or evaluative stance toward the given realities of society and the uses of power within it. By pretending, allegedly, that intellectual work could be free of political commitment, the end-of-ideology thesis simply masked its proponents' embrace of 'society as it is, a going concern' – and bowed to the narrow-minded empiricism that prevented American social science from broaching big questions about the nature of social order and pressing demands for social change (Mills, 1963: pp. 247–259).

Almost none of Bell's critics, however, directly challenged his central argument that socialism – as the concerted program of a dedicated political party advocating revolutionary social transformation – had lost relevance to the problems of Western industrial society. In fact, in this respect, there was a great deal of agreement on such central 'end-of-ideology' arguments between its main exponents and their most vigorous critics. The sociologists Stephen Rousseas and James Farganis, who like Mills reproached Bell for his alleged 'apotheosis of a noncommitted scientism,' claimed only to advocate 'long-range planning on a governmental level' that would "make the capitalist system viable in a power world [i.e., a world marked by systems of highly organized, centralized power that made old market ideology of decentralized competition obsolete]." In these terms, the end-of-ideology debate was not much of a debate at all, but a combination of misunderstandings, claims to marginal differentiation, or a contest between distinct *styles* of responding to conditions of welfare-state capitalism, which looked very much the same to most of the contenders. Much of the debate actually steered away from substantive argument over the structural character of the contemporary social organization of political economy and led to chest-thumping and self-regarding recriminations over the status of what a later generation would call 'public intellectuals.' Writing in the social-democratic journal *Dissent*, Dennis Wrong agreed that the threat of totalitarianism and successes of the welfare state had reduced the salience of independent socialist movements; yet, the end of ideology made a cult of moderation and accepted a consensus providing little room for debate. In an age marked by the end of ideology, he warned, intellectuals fail "to perform their roles as unattached critics and visionaries" (Mills, 1959; Waxman, 1969: pp. 116–125, 206–207, 221–222, 338).

Bell indeed abjured the role of 'visionary,' though he retained a deep interest, stemming from his Marxist origins, in 'forecasting' changes in social structure of an 'epochal' character – hence his work, beginning at that time, on 'postindustrial society.' He nonetheless took umbrage at other attacks, claiming his critics knew the title but not the contents of *The End of Ideology*. In fact, he evinced a strong interest in the radical critique of bureaucratic order. In one of the classic essays in the book, Bell marked the 'cult of efficiency' in a mass-consumption economy, embraced not only by corporate management but by the workers who suffered the hierarchical regimentation of mass production, as a proper target for criticism. Combating this 'cult' in his essay, 'Work and Its Discontents,' Bell mocked the 'scientific' approach of industrial engineers to conflict resolution and called instead for 'new, humanistic' approaches to work that would break down centralized controls and attack the alienation or estrangement induced by the gulf between labor and meaningful activity. "For the unions to challenge the work process would require a radical challenge to society as a whole," Bell wrote, adding that 'the exploration of such a road is necessary.' And Bell needed no lessons about the value-laden character of social analysis. "'The idea that a scientist simply studies 'what is,' as John Dewey has argued, is a parochial conception of science.… As Dewey said forcefully on another occasion: 'Anything that obscures the fundamentally moral nature of the social problem is harmful'" (Bell, 2000b: pp. 391, 386–387, 249, 237). Nor, contrary to those who saw 1960s protest as disproof, did his idea of the end of ideology ever deny the persistence of political conflict. In fact, he asserted that the new conditions of a politicized economy would, if anything, stir sharp new controversies (American Academy of Arts and Sciences, 1967: p. 7).

Bell's analysis and evaluation of American or 'Western' society during the period of postwar reconstruction and the 'end of ideology' had always been ambivalent, signaled by the subtitle of his book, 'The Exhaustion of Political Ideas in the Fifties,' which not only lacked any celebratory ring but came close to sharing his critics' disheartened view of the status of acute political thinking. On the one hand, as his conclusion to *Marxian Socialism* and the later work on postindustrialism indicated, Bell considered most inherited doctrines of social science and political theory irrelevant to addressing the real problems posed by the new shape of things. On the other, he recognized the shortfall of contemporary political debate:

> The pacifist and Socialist elements have been unable to make the peace issue salient. The radicals have been unable to develop a comprehensive critique of the social disparities in American life—the urban mess, the patchwork educational system, the lack of amenities in our culture. Among the liberals, only the exhaustion of the "received ideas," such as they were, of the New Deal remains. It

is a token of the emptiness of contemporary intellectual debate that from the viewpoint of the radical right, the Americans for Democratic Action constitutes the 'extreme left' of the American political spectrum.

He *bemoaned* the fact that "a viable left [does not] exist in the United States today" (Bell, 2002: p. 41).

Still, given his staunch anticommunism, his feud with C. Wright Mills and his contemptuous view of Mills's 'Power Elite' thesis, and his highly ambivalent, uncertain view of American social structure, Bell's writing featured few strongly worded sallies against the status quo. In line with his view that aspirations for social change and thorough-going reform – the principle of 'utopia' he still considered valid – must offer an 'empirical' path forward, specifying "*where* one wants to go, *how* to get there, the cost of the enterprise, and some realization of, and justification for the determination of *who* is to pay," Bell cofounded the journal, *The Public Interest* in 1965 to offer social–scientific knowledge and judgment on behalf of liberal social policy. While turning a skeptical eye toward elements of Lyndon Johnson's Great Society programs, such as the principle of 'maximum feasible participation' by community members in antipoverty efforts – and an opposition, generally, to the 'populist' spirit of the times – the *Public Interest* also provided the venue for liberal economists who developed the basic-income proposals that George McGovern advanced in his failed 1972 presidential campaign.

For much of the 1950s and 1960s, Bell and his associates devoted attention to the threat of 'the radical right,' which had gained prominence in US political life in the guise of Joseph McCarthy's alarmist conspiracy mongering and later, after McCarthy's censure, the John Birch Society, militia groups such as the Minutemen, and the Goldwater tide within the Republican Party. In a collection of essays first published in 1955 and then reissued as The Radical Right in 1962, Bell collaborated with historian Richard Hofstadter (as well as sociologists Talcott Parsons and Seymour M. Lipset) to venture a new analysis of political movements based on 'status anxiety' or 'status politics.' That is to say, right-wing surges like McCarthyism stemmed less from 'interests' based in one's social class than from anxieties over the prestige or moral standards one's group held, ways of life perceived to be threatened by social change or even, perhaps, dislocated by the experience of economic growth and prosperity. For Bell, this approach was wedded to his critique of orthodox Marxian views on the primacy of 'class' (and his view that this was a 'crude lens' for social and political analysis); for Hofstadter, it was closely related to his critique of Populism and Progressivism for seeking to hold on to outmoded views that placed simple farmers or the moral middle class at the heart of the Republic. Critics argued, variously, that the 'status' analysis neglected the 'real' economic grievances of middling farmers, unjustly suggested a link between right-wing McCarthyism and the old reform movements, missed the evidence of conventional business- or upper-middle-class conservatism behind McCarthy's campaigns, and mistakenly reduced the Right to a psychic pathology or a 'backward' perspective resisting, but bound to be outmoded by, ongoing 'progress.' Thus, according to its critics, the 'status' analysis missed the potency and future success of the conservative movement. Many of these criticisms were ill-informed or misguided, and Bell would later claim that *The Radical Right* effectively introduced the idea of 'cultural politics' that would prove in fact a key element of conservatism in the last quarter of the twentieth century (Bell, 2002).

Postindustrial Society and Cultural Contradictions

The main drift of Bell's work, however, led toward his portrait of postindustrial society. He did not invent the term, but as he worked the theme throughout the 1960s, he became the writer most closely associated with it. He began plying the theme around 1959–62, and a well-developed version of it debuted in a two-part article published in *The Public Interest* in 1967, prior to a major draft in 1969–70 that led to *The Coming of Post-Industrial Society* in 1973. Several traits marked the advent of this new order, including the rising weight of the 'tertiary' or service sector in the economy, but Bell more importantly defined postindustrial society as "one in which the intellectual is predominant." By that he meant not that intellectuals became a new ruling class but that the expanding social functions of science, the development of computer-based techniques of modeling and simulation, and growth in both public and private funding of research and development all contributed to making intellectual institutions (most notably the university) as socially 'central' in contemporary society as the business corporation had been in recent times. Bell always balanced skepticism about grand claims of postindustrial abundance through automation with an emphasis on new, emergent social principles that could transform human relations. The preeminence of 'theoretical knowledge' (basic, rather than applied, science) demanded an 'orientation toward the future' that was the watchword of postindustrialism. Only such forethought could provide long-term guidance to the generation of knowledge as a social and economic resource. A society organized to anticipate long-range needs was perforce focused on central, national institutions; at the moment he wrote, recent civil rights laws lent some plausibility to his claim that American life had entered a new period in which "the conscious direction of social change [is undertaken] by the federal government" (Bell, 1999, 1966: pp. 36, 70–88).

In a way, Bell's view of postindustrial society forecast a future in which the evolutionary development of the welfare state enabled a democratic polity to shoulder tasks of social planning. Here, his skeptical and progressive inclinations worked together, for only the assumption of persistent scarcity made public planning an imperative, resting on the need to choose between alternative allocations of resources that (like knowledge itself) were essentially and profoundly social. If, in Bell's mind, postindustrial society could not be considered 'posteconomic' in the sense of having escaped scarcity, it nonetheless pressed beyond the logic of markets: he argued that postindustrial trends elevated a new 'sociologizing' mode of decision-making over an 'economizing' mode. The first comprehended integral social processes involving nonprivate resources and goods such as education and environmental quality, while the second demanded only the efficient

allocation of discrete inputs to given production processes. Bell consistently resisted the inference some readers drew that his postindustrial society could no longer be understood as 'capitalist.' Still, his forecast suggested a society that looked significantly different from most conventional definitions of capitalism: the United States, he wrote, was "moving away from a society based on a private-enterprise market system toward one in which the most important economic decisions will be made at the political level"; "production and business decisions will be subordinated to, or will derive from, other forces in society" (Bell, 1999: pp. 344, 481).

As always, a certain degree of self-conscious ambiguity prevailed in Bell's analysis. The trait other writers put at the heart of postindustrial visions – the apparent rise of consumer culture supplanting 'productivist' values – as well as the persistence of capitalism in contemporary society provided the key themes of Bell's next book, *The Cultural Contradictions of Capitalism*. It had been drafted in 1969–70 as part of the same manuscript that yielded his *Post-Industrial Society*; the timing was telling. His part in a faculty committee trying to negotiate a resolution to the Columbia student protests of 1968 had deepened his distaste for the New Left (and, on the other side, for hard-liners who summoned club-wielding police to roust the radicals from their building occupations); he left Columbia to assume an endowed chair at Harvard University the next year. The kind of analysis he had applied to the Radical Right now focused on the New Left. According to Bell, 1960s left-wing radicalism bore the curse of 'mass politics,' manifesting a simplistic 'populism' – that is, a 'leveling' disposition – that challenged elites and repudiated any kind of legitimate authority. He also considered the 1960s 'counterculture' to be quite dated, a repeat of the 'modernist' rejection of social constraint and moral standards, now turned into a mass attitude. If his complaint over mass hedonism (signaled, he often asserted, by the flaunting of sexuality) struck some ears as prudery, his stance actually rested on a concern closely wedded to what Emile Durkheim originally meant by 'anomie' – not so much egoism as such but rather a kind of disorientation, conceivably stirred by prosperity itself, that so unhinged social norms as to leave impulses entirely unregulated.

Bell's *Cultural Contradictions* saw in these recent trends an antinomian dynamic of self-gratification that undermined the communal ethos needed for society to take advantage of the new modes of social planning made available by postindustrial trends. And that dynamic arose from the individualistic, consumerist passions spawned by capitalism, which thus appeared as the force spoiling the prospects of social-democratic advance. While his book title indicated the ways that new mass-culture impulses contradicted the moral foundations (in a Protestant work ethic) on which capitalism originally legitimated itself, Bell had little interest in defending capitalism as such. In terms that evoked, instead, his old left-wing origins, he argued that as the consumer market sundered its ties to the moral imperatives of work, capitalism had lost its 'transcendent ethic.' Bell claimed "the social order lacks either a culture that is a symbolic expression of any vitality or a moral impulse that is a motivational or binding force." Capitalism could offer no way forward; it was 'ideologically impotent.' What the emerging postindustrial society needed was what Bell called a 'public household' – 'an effort, in the realm of the polity, to find a social cement for the society,' or better put, the value-laden principles of decision that put a 'pilot' in the seat of government, 'the cockpit' of the postindustrial order in his view, to steer socioeconomic allocations as the market lost primacy and precedence (Bell, 1996b: pp. 84, 278).

In effect, Bell in the 1970s turned right in his recoil from the New Left and simultaneously turned left in his revitalized skepticism of capitalist norms. Surely, he did not share the Marxist or neo-Marxist assumptions of writers such as James O'Connor and Jürgen Habermas. He denied that there was a general systemic dynamic that made a coming social crisis inevitable. Having earlier dismissed his youthful confidence that socialism was destined to replace capitalism, Bell was prepared to address the crosscurrents of the 1970s as a manifestation of what he now called 'the disjunction of realms,' the argument that societies did not hold together according to a central governing logic but incorporated several different axes of development. In the current case, these axes were a social realm governed by postindustrial trends, a political realm determined by liberal principles of individual rights and representative government, and a cultural realm dominated by modernist or postmodernist mass consumer desire (Bell, 1996b: pp. 3–30).

Cultural Contradictions was one of the earliest American works to examine postmodernism – the peculiar amalgam of styles and attitudes, which fused the modernist avant-garde with mass culture while spurning elements of modernity associated with the 'rational cosmology' of the Enlightenment. Among the features of this rebellion Bell identified an 'eclipse of distance' manifested not only in the avant-garde idea of fusing life and art (evident, among other things, in the way 'happenings' and other action-oriented art forms sought to erase the gap between the work as such and the spectator) but also in the breaking of boundaries between public and private life and a general loss of moral reserve. Boundaries were essential, he believed: in writings on religion, he rejected the old view of modernity fostering secularization – he highlighted the resurgence of religious identity and motivations in his essays from the 1970s to the 1990s – but argued that the distinct phenomenon of 'profanation' indeed characterized popular culture that, in denying 'distance,' disregarded the realm of 'the sacred' (Bell, 1996b, 1980: pp. 99–119, 324–354).

Much of Bell's sensibility stemmed from his participation during the late 1940s in a kind of religious revival among Jewish intellectuals, who shared a new interest in theology, respect for religious experience, and the sentiment of belonging to a people having a transhistorical identity. Without claiming a personal conviction of faith, Bell came to voice a deep appreciation for the value of tradition, a sense of depending on larger things, and the need to accept constraint on impulse and personal gratification. This disposition carried over to social affairs as he insisted on the role that authority – that is, respect due to professional, political, and cultural elites – played in granting society continuity and stability, a precious achievement, he believed, in a world of many atrocities. Hence Bell's call in the 1970s and 1980s for the 'restoration of authority' in American life, combined with his bitter

denunciations of 1960s radicals and their legacy, constituted a genuinely conservative element of his thought (Bell, 1985).

Still, it was a mistake to associate Bell, as many critics did, with neoconservatism (Steinfels, 1979). He refused to follow the rightward current that led some postwar liberals, like his friend Irving Kristol, to join first the Nixon campaign of 1972 and later the Reagan coalition of the 1980s. It was Kristol's endorsement of Nixon (while Bell backed McGovern) that highlighted the political gap between them and led Bell to resign from the editorial board of the journal he cofounded, *The Public Interest*. Reaganism, for Bell, connoted an 'ideologizing' of capitalism, an outburst of crabbed religious moralism that threatened rights of privacy, and a desecration of the public household. Against claims that he joined Kristol's promotion of neoconservatism, Bell ventured his famous self-definition as "a socialist in economics, a liberal in politics and a conservative in culture." The formula mirrored his theoretical view on 'the disjunction of realms' and accurately summed up his dispositions. The 1980s found him writing in *Dissent*, and in its pages – and at a 1990 Mexico City conference drawing many of his old CCF colleagues together to reflect on the fall of the Berlin Wall – he wrote that "socialism, in the classic sense, has not failed, for it was never tried. It remains, like all ethical creeds, an ideal, a measure of potentiality and possibility, against reality. That has always been the function of utopia. The mistake was to assume that utopia could come down from the mountain and assume human form" (Bell, 1990: p. 188). He insisted, nonetheless, that such an ethical creed, devoted to 'equity, justice, and dignity' as against 'egoism' remained his own benchmark. In essays of the same period, Bell continued to probe the social and political implications of recent technological and economic innovations, particularly the 'problem of scale' that rendered the modern nation-state unfit either to manage a new economy of global range or to foster local participatory institutions that could give people a stake in civil society, understood in his liberal view as "tolerant and pluralist" (Bell, 1990: p. 188).

While the fall of the Soviet Union led observers to claim that Bell's *End of Ideology* had been vindicated, his perspective was falsely assimilated to the like-sounding title of Francis Fukuyama's *The End of History and the Last Man* (1991). According to Fukuyama, the end of the Cold War marked the 'end' of fundamental conflict between opposed visions of the good society, and it signaled the establishment of a broad consensus that only liberal democracy could truly fulfill the aspirations implicit in human reason and desire for prosperity and the recognition of personhood and fundamental rights. Ostensibly like the end-of-ideology writers in the 1950s, who saw Western society passing a threshold in which old 'ideological' illusions 'are dissolving and collapsing like a dream picture' (as Fukuyama quoted Hegel), Fukuyama celebrated a now world-ranging consensus on liberal values. Seemingly like end-of-ideology writers, he stated that "in most advanced democracies the big issues concerning governance of the community have been settled, reflected in the steady narrowing of the already narrow policy differences between political parties in the United States and elsewhere" (Fukuyama, 1992: p. 317). Yet in fact, Fukuyama's perspective was light-years apart from the end-of-ideology thesis, not only because Fukuyama's supposed final resolution of fundamental social and political disputes entailed an unambiguous settlement on the virtue and efficacy of free-market capitalism (never avowed by the end-of-ideology social democrats), but more importantly because Fukuyama built his argument by resurrecting Hegel's Universal History, whereby the 'cunning of reason' finally led to the end of all basic questions and struggles of social and political existence. In other words, Fukuyama's 'History,' having ended, evoked precisely the overarching schemes and systematic philosophy that the end of ideology had considered the curse and cause of the worst modern nightmares.

It was due to the distance between his own views and those like Fukuyama's that the new forward Bell wrote for the year 2000 edition of *The End of Ideology* was entitled, 'The Resumption of History in the New Century,' and why he worked over a manuscript entitled 'The Rebirth of Utopia' in his last decade, telling a young sociologist in 2004 that the forward-looking virtues of 'utopian' thought must still be valued and distinguished from the 'messianic' fervor for a brave new world coming in our time (what he always meant by 'ideology') (Bell, 2000b: pp. xi–xxviii; Beilharz, 2006). As always, however, the campaign against 'messianism' drew more of his attention than a focused discussion of viable 'utopias' (apart from his affirmation of a socialist ethic); also, while he assailed the Reagan ascendancy for granting full legitimacy to the Radical Right and its 'distinct tones of divisiveness and dark incitements of conspiracy,' he welcomed any sign of social 'stability' and civility at the political center (which he mistakenly perceived at the beginning of the George W. Bush years) (Bell, 2002: pp. 478–479). As a consequence, while he was more or less a 'man of the left' all his life, his death brought competing interpretations ranging from 'Daniel Bell, American Menshevik' to 'Daniel Bell, Conservative' (Beilharz, 2013; Murphy, 2013). He was in any case the American sociologist best known around the world, whose works were more widely read than those of the grand theorist, Talcott Parsons, who dominated the discipline when Bell entered it in the 1950s. By all accounts of those who knew him well, he would have welcomed continuing argument over the social, political, and moral implications of his ideas.

See also: Anomie; Aron, Raymond (1905–83); Cold War: Impact on Social and Behavioral Sciences; Intellectuals: History of the Concept; Liberalism: Historical Aspects; Mills, Charles Wright (1916–62); Social Democracy; Social Scientists as Experts and Public Intellectuals; Socialism: Historical Aspects; Societies, Types of; Sociology, History of; The American Conservative Movement; Totalitarianism; Weberian Social Thought, History of; Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

American Academy of Arts and Sciences, Bell, D., 1967. Toward the Year 2000: Work in Progress. [By Daniel Bell, et al., Summer 1967. Daedalus, Boston].

Aron, R., 1962. The end of the ideological age? In: The Opium of the Intellectuals. Norton, New York, pp. 305–324.

Beilharz, P., 2006. Ends and rebirths: an interview with Daniel Bell. Thesis Eleven, 93–103.

Beilharz, P., 2013. Daniel Bell – American Menshevik. Thesis Eleven 118, 64–71.

Bell, D., 1966. The Reforming of General Education: the Columbia College Experience in Its National Setting. Columbia University Press, New York.
Bell, D., 1980. The Winding Passage: Essays and Sociological Journeys 1960–1980. Basic Books, Inc. [s.l.], New York.
Bell, D., 1985. The Revolt against Modernity. Public Interest, 42–63.
Bell, D., 1996a. Marxian Socialism in the United States. Cornell Paperbacks, Cornell University Press, Ithaca, NY.
Bell, D., 1996b. The Cultural Contradictions of Capitalism. Basic Books, New York.
Bell, D., 1999. The Coming of Post-Industrial Society: A Venture in Social Forecasting. Basic Books, New York.
Bell, D., 2000a. Souvenirs: Paris, 1956–57. Tocqueville Review 21, 19–31.
Bell, D., 2000b. The End of Ideology: On the Exhaustion of Political Ideas in the Fifties: with a New Afterword. Harvard University Press, Cambridge, MA.
Bell, D., 2002. The Radical Right. Transaction Publishers, New Brunswick, NJ.
Bell, D., 1990. On the fate of communism. Dissent 37, 187–188.
Brick, H., 1986. Daniel Bell and the Decline of Intellectual Radicalism: Social Theory and Political Reconciliation in the 1940s, History of American Thought and Culture. University of Wisconsin Press, Madison, WI.
Brick, H., 2000. Age of Contradiction: American Thought and Culture in the 1960s. Cornell University Press, Ithaca, NY.
Brick, H., 2006. Transcending Capitalism: Visions of a New Society in Modern American Thought. Cornell University Press, Ithaca.
Crosland, A., 1957. The Future of Socialism. Macmillan, New York.
Crossman, R.H.S., 1952. New Fabian essays. Books That Matter. Praeger, New York.
Friedman, L.J., Schreiber, A.M., 2013. The Lives of Erich Fromm: Love's Prophet. Columbia University Press, New York.
Fukuyama, F., 1992. The End of History and the Last Man. Free Press; Maxwell Macmillan Canada; Maxwell Macmillan International, New York, Toronto.
Lipset, S.M., 1955. The state of democratic politics. Canadian Forum 35.
Lipset, S.M., 1960. Political Man; the Social Bases of Politics. Doubleday, Garden City, New York.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, New York.
Mills, C.W., 1963. Power, Politics, and People; the Collected Essays of C. Wright Mills. Oxford University Press, New York.
Murphy, P., 2013. Daniel Bell, conservative. Thesis Eleven 118, 72–82.
Pells, R.H., 1985. The Liberal Mind in a Conservative Age: American Intellectuals in the 1940s and 1950s. Harper & Row, New York.
Polanyi, M., 1955. On liberalism and Liberty. Encounter 4, 29–34.
Shils, E., 1955. The end of ideology? Encounter 5, 52–58.
Steinfels, P., 1979. The Neoconservatives: The Men Who Are Changing America's Politics. Simon and Schuster, New York.
Waxman, C.I., 1969. The End of Ideology Debate. Funk & Wagnalls, New York.

Bellah, Robert N (1927–2013)

Matteo Bortolini, Università di Padova, Padova, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Robert N. Bellah (1927–2013) was an American sociologist, public intellectual, and educator. A student of Talcott Parsons, he became known in the 1950s as a Japanologist and a theorist in the sociology of religion. During the 1970s, he started the 'American civil religion' debate and contributed to the interpretive/hermeneutical turn in the social sciences. In the 1980s, he was one of the coauthors of one of the most successful books in the history of sociology, *Habits of the Heart*. Toward the end of his life he published his *magnum opus*, *Religion in Human Evolution*.

Becoming a Social Scientist

Robert Neelly Bellah was born in Altus, Oklahoma, on 23 February 1927, to a Presbyterian family of Scotch-Irish and English descent. He grew up in West Los Angeles and attended Los Angeles High School, where he edited the school daily newspaper, *Blue and White*, and was involved in student government. A high achiever, Bellah formed a wide circle of friends, including Jewish students, with whom he shared an interest in liberal politics, classical music, and literature. One of them, Melanie Hyman, would become his wife in 1949. In 1945, Bellah entered Harvard College, where he received a BA in social anthropology in 1950 from the Department of Social Relations. His honors thesis, *Apache Kinship Systems*, won the Phi Beta Kappa Prize and was published by Harvard University Press in 1952.

Established in the summer of 1946, the Department of Social Relations was the institutional base of an ambitious interdisciplinary project envisioned by Talcott Parsons, Clyde Kluckhohn, Gordon Allport, and Henry A. Murray. The Department aimed at integrating the 'basic social sciences' (sociology, cultural anthropology, clinical and social psychology) as the foundations upon which any scientific study of human behavior should be built (Gilman, 2003). Parsons and his colleagues created an ambitious graduate studies program in which different disciplinary streams of research would interact in intense mutual interchanges. While the Department never reached the scientific and professional integration Parsons had hoped for, its results were remarkable: during its first 10 years it recruited George Homans, Samuel Stouffer, Jerome Bruner, David Riesman, and Erik Erikson as faculty members, and granted doctoral degrees to scholars like Harold Garfinkel, Clifford Geertz, and Neil Smelser.

In 1951 Bellah began a joint doctoral program in Sociology and Far Eastern Languages and started working on a dissertation inspired by Max Weber's sociology of religion. Applying Weber's thesis linking the Protestant ethic and the spirit of capitalism to Tokugawa Japan (1600–1868), he aimed at describing the emergence of religious movements analogous to the Protestant sects as carriers of the rational and methodical ethic needed for the takeoff of modernization. With Parsons and John Pelzel as his advisors, Bellah worked on the dissertation until the summer of 1954, when he experienced the troubles of McCarthyism on campus.

As an undergraduate inspired by the ideals of social democracy, Bellah had been a member of the John Reed Club, a Marxist organization at Harvard College, and identified as a Communist. Despite his personal disillusionment with Stalinism and his subsequent break with the party in 1950, in the summer of 1954 Bellah was asked by the dean of the Faculty of Arts and Sciences, McGeorge Bundy, to cooperate with the FBI, but he refused to denounce his old comrades. Some months later, when Parsons proposed Bellah for an instructorship in the Department of Social Relations, the interrogation began again. While Parsons and other colleagues backed him, in 1955 Bellah decided to leave Harvard and accepted a 2-year fellowship at the Institute of Islamic Studies established by Wilfred Cantwell Smith at McGill University in Montreal, Canada.

Bellah at Harvard

Declining offers from McGill and Columbia University, Bellah went back to Harvard on a fellowship at the Center of Middle Eastern Studies which was converted into a 4-year appointment to the Faculty of Arts and Sciences and the Harvard Divinity School in 1958. His immediate circles abounded with eminent figures. Paul Tillich, Krister Stendhal, and, after 1964, Cantwell Smith were among his colleagues at the Divinity School, while the Department of Social Relations was becoming America's most influential sociology department. In area studies, Bellah was mentored by Edwin Reischauer, H.A.R. Gibb, Serge Elisséeff, and John Fairbank. He regularly taught classes on Islam, China, Japan, and courses and seminars on the history and sociology of religion, Weber, and Émile Durkheim with Talcott Parsons. Bellah spent a year (1960–61) in Japan and was a fellow at the Center for Advanced Studies in the Behavioral Sciences at Stanford in 1964–65. He was appointed associate professor in 1961 and full professor in 1966.

During his first decade as a scholar, Bellah was an integral member of two of the most influential networks of the postwar social sciences, those promoting structural functionalism and modernization theory. Led by Parsons and Robert K. Merton, structural functionalism was characterized by a naturalistic understanding of the social sciences; an analytical conception of sociological theory; an interest in European theorists; and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61321-1

the search for functional explanations of social phenomena in terms of normative structures and moral schemata. Influenced by America's Cold War response to Marxist (and thus Sovietic) depictions of the rise of industrialism, modernization theory was an evolutionary theory of societal change loosely founded on a Parsonian systemic framework. In particular, the work of modernization theorists was informed by a sharp distinction between traditional and modern societies; a systemic depiction of societies as differentiated but quasi-coherent wholes; an explanation of social change as a unilinear sequence of evolutionary stages; an allegedly technocratic and postideological understanding of public life; and an identification of the United States as the apex of the modernization process.

A participant in the institutions where these networks intersected – the Department of Social Relations, the Committee for the Study of the New Nations at Chicago, and the MIT Center for International Studies – Bellah developed structural-functionalism and modernization theory as an analytically informed macrosociology focused on the problem of meaning in modern and modernizing societies. In typical Social Relations fashion, he carefully combined the study of symbolic systems, psychological motives, and social structures, while his deep engagement with Weber and Durkheim extended Parsonian analysis by incorporating the historical side of their writings. In *Religious Aspects of Modernization in Turkey and Japan*, Bellah (1958) hypothesized that the differentiation between religion and political ideology, a property of modern societies, could only be brought about by movements able to challenge the religious foundations of 'old,' undifferentiated traditional societies. In *Durkheim and History*, he vindicated the French sociologist's (and, by extension, structural functionalism's) interest and ability to study social change (Bellah, 1959).

Bellah's second book, *Tokugawa Religion: The Values of Pre-Industrial Japan*, was published in 1957. It laid on a synthesis of structural-functionalism, modernization theory, and his more culturalist orientation. Bellah explained Japan's successful modernization pointing to the paramount importance given to political loyalties and relationships within the country's 'central value system.' Using a complex theoretical framework built on Weberian concepts and Parsons' pattern variables and four-function (AGIL) paradigm, Bellah (1957) compared Japan's 'political values' to America's 'economic values' and China's 'integrative values,' and described their different paths to modernization. As the capital needed to catch up with Western capitalist countries was too great to be supplied by the economy alone, the particular orientation of its value system was instrumental in giving Japan an edge over its main Eastern competitor: during the Tokugawa period, the government had succeeded in mobilizing political loyalties toward industrialization and in orienting the population toward national development. Religious commitments made sure that political values and allegiances were 'meaningful in an ultimate sense' and gave a metaphysical justification to the Restoration of the Emperor, itself a major disruption of existing social and cultural structures. Following Weber's lead, Bellah focused on a spiritual movement, Shingaku, as a proxy for the inner-worldly asceticism that the German sociologist had deemed crucial for the emergence of the spirit of capitalism. The book closed on a critique of materialist versions of modernization theory and pleaded for a multidimensional analysis of industrialization that would not consider the United States the sole example of successful modernization.

Reviewers praised *Tokugawa Religion* for its use of the Parsonian framework to shed light on some of the most important questions of Japanese economic and political history. The book was reviewed by Japan's foremost political scientist and philosopher, Maruyama Masao, who criticized Bellah's interpretation of the role of Shingaku and the Restoration of the Emperor, but also compared him to Ruth Benedict. In his subsequent work, Bellah abandoned the search for Protestant ethic analogies in Asia, focusing instead on those twentieth-century intellectuals, such as Ienaga Saburo and Watsuji Tetsuro, who had tried to make sense of Japanese modernization (see Bellah, 2003). In the early 1960s, Bellah's involvement in the Conference for Modern Japan marked his entry into the pantheon of American Japanologists.

In his more theoretical work, Bellah sought to create a multidimensional analytical framework based on Parsons' AGIL model. In *The Systematic Study of Religion*, an unpublished paper widely circulated in the mid-1950s, he assigned the label 'religion' to the most abstract subsystem of the cultural system, its main functions being the maintenance of meaning complexes and the management of psychological tensions. As Parsons himself had done in a number of essays and articles, Bellah ruled out the very possibility of a fully secularized society: from an analytical point of view, no society could do without 'religion,' so that if traditional religions could not fulfill the functions of pattern maintenance and tension management, other symbols, actions, and institutions would emerge and replace them. This position informed Bellah's major theoretical piece of the early 1960s, *Religious Evolution*. Bellah had begun his reformulation of one of classical sociology's main themes while in Montreal, and worked out his ideas in a seminar he led with Parsons and Shmuel Einsenstadt at Harvard – the three men each published an essay on evolution in a special issue of the *American Sociological Review* in 1964.

Starting from a Tillichian definition of religion as "a set of symbolic forms and acts that relate man to the ultimate conditions of his existence," Bellah (1970a) described the evolution of religious symbolism within the context of growing differentiation, rationalization, and emancipation of humankind from its material, social, and metaphysical environments. Religious evolution went through five stages: (1) *primitive religion*, where mythic symbol systems combine with ritual participation, and religious organizations correspond with whole societies; (2) *archaic religion*, characterized by complex social differentiation and the development of differentiated gods, who are nonetheless ancestors or kin of the high chiefs or noble classes who rule; (3) *historic religion*, in which there is a radical break between two different realms, mundane and transcendent, and religious action becomes oriented to gain salvation from worldly evils; (4) *early modern religion*, which comprises only a single case, the Protestant Reformation, based on the direct relationship between the individual and transcendent reality, a devaluation of ritual, and the diffusion of 'religion' through everyday life; (5) *modern religion*, a post-Protestant phase where monist and dualist representations give way to an 'infinitely multiplex' world and life is an 'infinite

possibility thing.' The search for ultimate meaning turns into an experimental, ever-changing activity, while religious organizations become deeply individualized – up to the point that each individual becomes a 'sect himself.'

The Civil Religion Debate

Bellah's ambivalent conclusions to 'Religious Evolution' – "No doubt the possibilities for pathological distortion in the modern situation are enormous (…) Yet the very situation that has been characterized as one of the collapse of meaning and the failure of moral standards can also, and I would argue more fruitfully, be viewed as one offering unprecedented opportunities for creative innovation in every sphere of human action" (Bellah 1970a) – were an anticipation of a major change in his intellectual and personal standing as a scholar. As the 1960s came to an end, Bellah was slowly distancing himself from the ideas and the networks that had made him a well-known Japanologist and theorist of religion. First, he was discarding functionalism in favor of an original synthesis between historical sociology and a deep hermeneutic of the symbolic dimension of social and cultural phenomena. Second, in his language and argumentation Bellah was creating an evocative and metaphorical vocabulary which bore no resemblance to Parsons' esoteric jargon. Third, from an organizational point of view, while his mentor was the foremost proponent of a differentiated sociology, Bellah assumed a cross-disciplinary posture and started to reflect on the artificiality, and ultimately the unsuitability, of scholarly boundaries. Last but not least, he abandoned Parsons' detached, professional stance and became increasingly involved from a personal point of view. Bellah's intellectual shift was also accompanied by a major existential change. After declining offers from the University of Michigan, UCLA, Columbia University, and the University of Chicago, he finally left Harvard in the spring of 1967 to become Ford Professor of Sociology and Comparative Studies and chairman of the Center for Japanese and Korean Studies at the University of California, Berkeley, and an adjunct professor at the Berkeley Graduate Theological Union.

Civil Religion in America was probably the most significant instance of this profound change. In October 1965, Bellah participated in a conference on 'Religion in American Culture' at the American Academy of Arts and Sciences. Although not a specialist, Bellah had written occasionally on American religion in the 1950s and 1960s. His paper for the AAAS, *Heritage and Choice in American Religion*, received notice for its use of the Rousseauian concept of 'civil religion.' The final version of the paper, titled *Civil Religion in America*, was published on the winter 1967 issue of *Daedalus*. Bellah's main point was spelled out unambiguously in its very first lines: "There actually exists alongside of and rather clearly differentiated from the churches an elaborate and well-institutionalized civil religion in America" (Bellah, 1970a). According to Bellah, a set of shared symbols, beliefs, and rituals rooted in the language and imagery of the Old Testament, but also 'genuinely American and genuinely new,' had historically substituted for the lack of an established religion. This American civil religion had been shaped by two crucial moments of national history, the Revolution and the Civil War, which had introduced 'a new theme of death, sacrifice, and rebirth' symbolized by Abraham Lincoln, the 'American Jesus.' As the late 1960s emerged as a 'third time of trial,' the civil religious tradition could serve as an inspiration for the United States to contribute to the creation of a world civil religion. Just like any tradition, Bellah concluded, the American civil religion could never transcend moral ambiguity, but it might be a guide for righteous action when correctly understood and enacted.

Although the theme was not new, Bellah's conceptualization of the American civil religion immediately caught on and generated a wide and raucous debate which involved sociologists, historians, theologians, and political scientists on both sides of the Atlantic. Bellah decided to participate in the discussion out of his political concerns after the escalation of the Vietnam War and his disappointment in the scholarly status of coeval sociology of religion. Between 1968 and 1980, Bellah participated in countless conferences, symposia, and public meetings with academic and nonacademic audiences where critics discussed the value and the uses of the concept. As a newcomer to American studies, he had to immerse himself in a crash course in the history of the United States and he profoundly changed his own perspective on the civil religion in the process. A first reformulation of his ideas was presented in the fall of 1971 in the Frank L. Weil Memorial Lectures in Cincinnati, where he said that the original American civil religion had always been unable, by itself, to raise the moral standards of the nation, and only exceptional presidents like Jefferson or Lincoln had been able to give it the force and appeal it needed to function as the critical conscience of the American people. Moreover, the civil religion had always been accompanied by the public-spirited interventions of theologians and pastors, so that it had been 'the churches who spoke with the combined voice of the believers and citizens.' In subsequent essays, Bellah considered the civil religion as one of two competing cultures – biblical-republican and liberal-individualistic – which had struggled for the soul of the American people ever since the founding of the early Bay colonies.

These ideas found a comprehensive formulation in Bellah's (1975) book, *The Broken Covenant: American Civil Religion in Time of Trial*, which expanded the Weil Lectures in a critical appreciation of the symbolic and mythical life of the United States from John Winthrop's speech on the Arbella to the Vietnam War. In the conclusion, Bellah (1975) famously wrote not only that "today the American civil religion is an empty and broken shell," but also that it had been, "from the beginning an external covenant," that is, a *modus vivendi* and not a shared *conscience collective*. Bellah feared that the utilitarian-individualistic tradition would finally impose itself on the more solidaristic, biblical-republican one, thus undermining the very bases of American society and transforming the latter into a bureaucratic authoritarian nightmare. The book was widely acclaimed and won the Sorokin Award of the American Sociological Association for 1976. At the same time, it was heavily criticized by Parsons before and after publication. Bellah's old mentor was disappointed at both the book's grim diagnosis of the moral malaise of American society and its passionate, almost prophetic tone. In *Law as an Intellectual*

Stepchild, Parsons (1977) portrayed *The Broken Covenant* as excessively Manichean and called Bellah 'a moral absolutist.'

Interpretive Social Science and the 'Bellah Affair'

As the American civil religion debate was turning him into a star of American studies, Bellah managed to publish a host of methodological essays that challenged the received wisdom of the sociology of religion in mid-century America. Embarrassed by being too close to theology and not 'scientific' enough, the sociology of religion had enthusiastically embraced the naturalistic positivism of the 1950s and set out to come of age as a modest, unpretentious, and 'objective' subfield of social science. In *Transcendence in Contemporary Piety* Bellah abandoned his previous functionalism for the concept of *ultimate symbols* – defined as those symbols which represent the unity of all beings and that overcome ordinary dichotomies bringing together 'the coherence of the whole experience.' In a short autobiographical piece, aptly titled *Confessions of a Former Establishment Fundamentalist*, he repudiated the naturalistic logic of the postwar social sciences – and with it his own Parsonian upbringing – in favor of a hermeneutical approach to the study of religion (Bellah, 1970a, 1970b).

Bellah publicly discussed his new stance at a joint session of the Society for the Scientific Study of Religion and the American Academy of Religion held at the Sanders Theater, Harvard University, on 25 October 1969. There he attacked both the reductionist rationalism of Enlightenment social science and the unreflective, first-order naïveté of traditional religious education. Both perspectives had to be abandoned in favor of a new framework, 'symbolic realism,' which would recognize the *sui generis* reality of religious symbols, and would blur the boundary lines between social science, the humanities, and theology. According to Bellah, religious symbols were attempts to grasp "the totality which [included] subject and object and [provided] the context in which life and action finally [had] meaning." As such, they could be interpreted and understood only if scholars would embrace a particular attitude, which he called, following Paul Ricoeur, a 'second naïveté,' according to which any interpretation of reality, even science and rational thought, should be treated as 'only an interpretation, and not reality itself.'

Bellah's first collection of essays, *Beyond Belief*, was published in 1970. It included all his major papers along with a dense introduction where he worked out the connection between his personal experiences and his professional work, and made a plea for the establishment of interdisciplinary departments of religion in all universities, whether public or private. Ideas about symbolic realism were further elaborated in Bellah's introduction to his edited collection of Durkheim's writings on morality and society, and, more radically, in his many interventions about the teaching of religion, where he called for a new integration between knowledge, commitment, and faith. While the perspective of symbolic realism was reworked in a host of essays and articles, it was Bellah's forceful phrase – 'to put it bluntly, religion is true' – that continued to echo most memorably in the minds of his interlocutors. This made him the most radical proponent of what came to be called 'interpretive social science,' a loose movement which included former Parsons students, and Bellah's personal friends, David Schneider and Clifford Geertz, along with Victor Turner, Peter Berger, and Thomas Luckmann.

A major episode for Bellah's scholarly career came in 1973, when he was invited to take part in creating a school of social science at the Institute for Advanced Study at Princeton. Geertz, who had been appointed as the Institute's first permanent member of the social sciences in 1970, invited Bellah to join him in 1972, but the nomination aroused strong opposition from the Institute's mathematicians and historians. The Institute's director, former Harvard economist Carl Kaysen, pushed forward the nomination, which was approved by the Board of Trustees, against the majority of the Institute's permanent members. A harsh public quarrel ensued when confidential evaluations of Bellah's work were forwarded to the national press, threatening his scholarly reputation. The so-called 'Bellah affair at Princeton' abruptly ended when his eldest daughter, Tammy, committed suicide. Bellah and his family decided to leave the Institute and go back to Berkeley. The project of creating a recognizable school of interpretive social science was put aside, to be revived only when Bellah led a National Endowment for the Humanities residential seminar on 'Tradition and Interpretation: The Sociology of Culture' at Berkeley in 1976 and other similar enterprises (Rabinow and Sullivan, 1979; Bellah et al., 1983).

Habits of the Heart and Public Sociology

In 1976 Bellah and Charles Glock concluded a 5-year research project on new and 'chosen' religions and published *The New Religious Consciousness* (Glock and Bellah, 1976). Although the book was a diverse collection of ethnographies of religious and expressive groups, only partially reflecting the principles of symbolic realism, it constituted a diversion from the American civil religion debate, which Bellah found too focused on conceptual quarrels. His disappointment was partly due to the fact that his change of mind, spelled out in *The Broken Covenant*, had been barely noticed by his critics, who continued to attack an overly integrative, Durkheimian version of the civil religion thesis to which he had never subscribed.

Bellah then started working on a study titled 'The Moral Basis of Social Commitment in America' with four younger colleagues: Richard Madsen, a sociologist specializing in China; William Sullivan, a moral philosopher who had participated in Bellah's NEH seminar; Ann Swidler, a sociologist interested in cultural theory; and Steven Tipton, author of *Getting Saved From the Sixties* and the initiator of the project. In the meantime, Bellah's stature as a public intellectual continued to grow, as his many interviews and passionate interventions on American politics and religion show. In July 1979, he was invited along with other intellectual and spiritual leaders to advise President Jimmy Carter a few days before his famous 'Crisis of Confidence' speech (Horowitz, 2005). One month later, he delivered an intense

commemoration of Talcott Parsons, who had died in May 1979, and claimed his heritage for interpretive social science (Bellah, 1980). From 1979 to 1985 Bellah also served as Chair of Berkeley's Department of Sociology.

The results of the collective research were published in 1985 as *Habits of the Heart*, a phrase taken from Alexis de Tocqueville's *Democracy in America* (Bellah et al., 1985). The book, written for the general public, aimed at starting a 'conversation' on what it meant 'to create or preserve a morally coherent life' and focused on Bellah's cultural *bête noire*, individualism. Starting from a conception of culture as a multivocal conversation about three 'ultimate goals' – success, freedom, and justice – *Habits* presented a complex picture of the American symbolic realm, one that put individuals and their dignity at the center of moral discourse, albeit in very different ways. Bellah et al. identified four main cultural vocabularies: the republican and the biblical traditions, centered on a solidaristic understanding of moral life, and two individualistic traditions, which they labeled as 'utilitarian' and 'expressive' individualism.

From a sociological point of view, *Habits of the Heart* focused on four main topics: intimate relationships, civic and political involvement, psychotherapy, and religion. It attempted to create a continuous dialogue between the cultural heritage of American political and social history and the data collected interviewing some 200 average, middle-class Americans. In most of the cases, the authors' 'active interviews' with their fellow citizens revealed that the hegemonic language of individualism, whether utilitarian or expressive, was unfit to either describe or to make sense of the social and cultural relations in which ordinary people lived their daily life. Despite its clear Durkheimian roots and its penchant for an ecumenical version of the biblical – republican tradition, *Habits* did not portray America as an atomized, Hobbesian *Gesellschaft*. Bellah and his coauthors saw modern society as a combination of multiple 'communities of memory' – that is, groups of people who share a common narrative of their past, which they use to make sense of their lives, and who engage in continuous exchanges with each other.

The book abounded with evocative catchwords, among which 'Sheilaism' probably became Bellah's most celebrated intervention in the sociology of religion. Taken from an interview with the fictionally named Sheila Larson, it indicated a kind of do-it-yourself spirituality rooted 'in the effort to transform external authority into internal meaning,' a purely individualized creed assembled using bits of existing traditions. In the book's methodological appendix, 'Social Science as Public Philosophy,' Bellah attacked the allegedly value-free attitude of social scientists and pleaded for a publicly responsible social science, aware of the impossibility of separating cognitive and moral reasoning and committed to engage the lay public in dialogue. In his critique of existing social science Bellah was definitively abandoning his postwar upbringing and was translating his early symbolic realism in a more accessible version.

Habits of the Heart was a huge success. It was reviewed in *The New York Times Book Review*, *Newsweek*, *The Los Angeles Times*, and *The Times Literary Supplement* and became one of the all-time best sellers by sociologists (Gans, 1997). As its authors had hoped, it spawned a vibrant discussion among scholarly and lay publics, and was the object of countless reviews, meetings, and symposia. In 1986, *Habits* was nominated for the Pulitzer Prize for General Non-fiction; one year later, the team started working on a sequel, *The Good Society*, which was published in 1991. The new book focused on institutions and the cultural meanings that underlie them: the market, the government, schools and universities, and 'the public church' (Bellah et al., 1991). A more sociological book than *Habits*, *The Good Society* ended with a plea for democracy as 'paying attention.' As the senior member of the team, Bellah enjoyed most of the recognition for the two books and became one of the most cited public intellectuals in America (Posner, 2003). Despite his aversion to the term, in the 1990s he was often labeled a 'communitarian' – in part due to the fact that he was an early signer of Amitai Etzioni's 'Responsive Communitarian Platform.'

Back to Religion and Evolution

In 1997 Bellah retired from the University of California, but continued his scholarly work, publishing widely on American politics and society, the legacy of the classics, and education. His many interventions as an outspoken critic of unethical versions of individualism and his well-known sermons as a member of the Episcopal Church rounded out his public persona. In 2000 he was awarded the United States National Humanities Medal 'for his efforts to illuminate the importance of community in American society.' In the epilogue to his *Festschrift*, Bellah hypothesized the emergence of a new, post-Protestant society, one in which the 'axial' rupture between transcendence and the mundane order was somehow supplanted by a new form of 'disembedded immanentism' which threatened to turn all human action into technical calculation. As many of the essays collected in his eponymous *The Robert Bellah Reader* show, he found the roots of this grim development in some unredeemable 'flaw in the Protestant code' (Madsen et al., 2002; Bellah and Tipton, 2006).

More importantly, after retirement Bellah resumed a research project he had left dormant for nearly 30 years: writing a book on religion and evolution. In the early 1960s he decided to devote his scholarly life to a grandiose project, a rewriting of Max Weber's comparative sociology of religion in the terms of a theory of societal evolution. The American civil religion debate and the huge popularity of *Habits of the Heart* had 'distracted' him from this line of research, although the fivefold typology he had first spelled out in 'Religious Evolution' continued to be the backbone of his lectures on the sociology of religion at Berkeley. Preceded by some preparatory essays and a conference on 'The Axial Age and its Consequences for Subsequent History and the Present,' held at the University of Erfurt, Germany, in 2008 – in which Charles Taylor, Jürgen Habermas, Shmuel Eisenstadt, Jan Assman, and David Martin, among others, participated (see Bellah and Joas, 2012) – *Religion in Human Evolution* was finally published in 2011 by Harvard University Press.

In order to reconstruct the interactions between religion, social order, and the world of everyday life 'from the

Paleolithic to the Axial Age,' Bellah (2011) combined a Durkheimian definition of religion, centered on the notion of the sacred, with a more phenomenological, meaning-centered one, derived from the work of Clifford Geertz. From a theoretical point of view, religion was considered as a particular symbolic and institutional sphere that combined with biological, psychological, social, and cultural structures to create emerging structures and 'capabilities.' Conceptually, the most important advance with respect to Bellah's previous work was the combination of Jerome Bruner's four levels of representation with the four kinds of culture singled out by the evolutionary psychologist Merlin Donald: unitive symbols/episodic culture; enactive representation/mimetic culture; symbolic representation/mythical culture; and conceptual representation/theoretic culture. Each of these categories was considered as an autonomous evolutionary stage, emergent from previous ones and able to react on them. In fact, according to Bellah, 'Nothing is ever lost': as evolution proceeds, earlier stages are not undone or superseded, but only reorganized under new conditions.

The book also presented a threefold typology of societies, each with its own kind of religion, seen as the symbolic realm that secures ultimate meaning to social and political structures. Tribal societies are based on 'reverse dominance hierarchies,' in which a fundamental equality is maintained in the face of repeated attempts at imposing a full-fledged, vertical hierarchy. Upstarts are successfully kept at bay by mimetic religious rituals in which all the members of the tribe participate in order to reimpose again and again the power of society over its individual members and their disposition to dominate. Archaic societies are characterized by strong political centers, articulated systems of stratification, and complex cosmologies which make sense of the relationships between ever-growing pantheons of deities and society. The shift from tribal to archaic societies is possible only when divine kings succeed in having the whole of society focus its attention on them as the sole mediators between the divine and the mundane realms.

Axial religions are marked by the so-called 'axial break,' which Bellah reconceptualized as the emergence of 'second-order thinking,' understood as 'thinking about thinking' – that is, a reflection on the very bases of rational reflection. Often, but not always, axial religions are characterized by a radical cleavage between a transcendent and a mundane order, so that the status of truth and the authority of divine kings are put into question: 'Who is the (true) king, the one who really reflects divine justice?' Through the analysis of four case studies – Ancient Israel, India, China, and Greece – Bellah described the emergence of second-order thinking in religious and philosophical thought, and focused on those prophets, philosophers, and free-floating intellectuals who put themselves at the center of the axial break as the carriers of the higher truths of some universal ethics. In the concluding chapter, Bellah praised the critical and universalistic heritage of the Axial Age, but warned against the modern pretension to detach allegedly universal truths and ethics from practical and concrete forms of life.

Upon its publication, *Religion in Human Evolution* was acclaimed as a milestone in the historical sociology of religion. It received praise and critiques from scholars in various disciplines and was the subject of symposia and meetings in America, Europe, China, and Japan (see Stausberg, 2014). When Robert Bellah died on 31 July 2013 from the complications of heart surgery, he was writing one last book, tentatively titled *The Modern Condition in the Light of Human Evolution*. His work on civil religion, Japan, and religious evolution, and his understanding of the social sciences as public philosophy remain influential in both sociology and religious studies alike. In his last major intervention – the Paul Tillich Lecture at Harvard's Memorial Church in May 2013 – Bellah restated his critique of utilitarian individualism and the perils of the 'Protestant Principle.' But he concluded, speaking of Tillich notion of *kairos* as the possibility of something new in history, that "the *kairos* is now, always now. It is now that we are called to hope and to act."

See also: Communitarians; Geertz, Clifford (1926–2006); Parsons, Talcott (1902–79); Riesman, David (1909–2002).

Bibliography

Bellah, R.N., 1957. Tokugawa Religion. The Free Press, Glencoe.

Bellah, R.N., 1958. Religious aspects of modernization in Turkey and Japan. American Journal of Sociology 64 (1), 1–5.

Bellah, R.N., 1959. Durkheim and history. American Sociological Review 24 (4), 447–461.

Bellah, R.N., 1970a. Beyond Belief. Harper & Row, New York.

Bellah, R.N., 1970b. Confessions of a former establishment fundamentalist. Bulletin of the Council for the Study of Religion 1 (3), 3–6.

Bellah, R.N., 1975. The Broken Covenant. Seabury Press, New York.

Bellah, R.N., 1980. The world is the world through its theorists. American Sociologist 15 (2), 60–62.

Bellah, R.N., 2003. Imagining Japan. University of California Press, Berkeley-Los Angeles.

Bellah, R.N., 2011. Religion in Human Evolution. Harvard University Press, Cambridge, MA.

Bellah, R.N., Glock, C.Y. (Eds.), 1976. The New Religious Consciousness. University of California Press, Berkeley-Los Angeles.

Bellah, R.N., Haan, N., Rabinow, P., Sullivan, W. (Eds.), 1983. Social Science as Moral Inquiry. Columbia University Press, New York.

Bellah, R.N., Joas, H. (Eds.), 2012. The Axial Age and Its Consequences. Harvard University Press, Cambridge, MA.

Bellah, R.N., Madsen, R., Sullivan, W.M., Swidler, A., Tipton, S.M., 1985. Habits of the Heart. University of California Press, Berkeley-Los Angeles.

Bellah, R.N., Madsen, R., Sullivan, W.M., Swidler, A., Tipton, S.M., 1991. The Good Society. Alfred A. Knopf, New York.

Bellah, R.N., Tipton, S.M. (Eds.), 2006. The Robert Bellah Reader. Duke University Press, Durham.

Gans, H.J., 1997. Best-sellers by sociologists: an exploratory study. Contemporary Sociology 26 (2), 131–135.

Gilman, N., 2003. Mandarins of the Future. The Johns Hopkins University Press, Baltimore.

Glock, C.Y., Bellah, R.N. (Eds.), 1976. The New Religious Consciousness. University of California Press, Berkeley, LA.

Horowitz, D. (Ed.), 2005. Jimmy Carter and the Energy Crisis of the 1970s. Bedford/St. Martin's, Boston-New York.

Madsen, R., Sullivan, W.M., Swidler, A., Tipton, S.M. (Eds.), 2002. Meaning and Modernity. University of California Press, Berkeley-Los Angeles.

Parsons, T., 1977. Law as an intellectual stepchild. Sociological Inquiry 47 (3–4), 11–58.

Posner, R., 2003. Public Intellectuals. Harvard University Press, Cambridge, MA.

Rabinow, P., Sullivan, W.M. (Eds.), 1979. Interpretive Social Science. University of California Press, Berkeley-Los Angeles.

Stausberg, M., 2014. Religion in human evolution: a post-review. Numen 61, 281–299.

Bendix, Reinhard (1916–91)

John Bendix, Universität Zürich, Zürich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Reinhard Bendix was a German-born political sociologist whose work focused on industrial relations, nation-building, the history of ideas, the role of the social sciences in society, and the transformation of political legitimacy over time and across nations. He drew his inspiration from the work of Max Weber, one of sociology's founders, and was instrumental during the 1960s and 1970s in making Weber's vast and complex work accessible to English-speaking scholars. Bendix was a key figure in what has come to be called the 'first wave' of historically and comparatively oriented sociology in the United States, together with Seymour Martin Lipset (with whom he collaborated extensively; see Lipset and Bendix, 1953, 1959), and Barrington Moore.

Born in 1916, Reinhard Bendix grew up in Berlin. Thrown out, at age 17 years, of the academic high school he had attended, ostensibly because his classmates could not be expected to tolerate someone in their midst whose father had been arrested and taken into 'protective custody,' he spent the next years working as a clerk in a Jewish export firm in a Berlin growing increasingly oppressive to those of Jewish heritage. Emigrated colleagues of his father Ludwig, a well-known and widely published Berlin lawyer, were able in 1938 to obtain a visa for him that allowed him to emigrate to the United States. Despite his truncated secondary education, one of those colleagues secured admission for him to, and a tuition fellowship from, the University of Chicago, a university that would subsequently award Bendix his BA (1941), MA (1943), and Ph.D. (1947) in sociology. In 1947, he was offered a position at the University of California at Berkeley in the Sociology Department. He remained with the sociologists until 1972, at which point he decided to move a few floors up in Barrows Hall to join the political scientists.

His father Ludwig (Louis) Bendix left the Jewish community at age 20 years in 1897, embraced the German cultural norms of what it meant to be a cultivated, educated citizen, and did his best at home to pass this *Bildungsbürgertum* orientation, as well as his interest in social and political reform, on to his son. But Ludwig's utter commitment to the law (Bendix, L., 1968), which included defending communists in court, along with his outspoken opposition to what the Nazis were doing to the law and to the German state, brought about his repeated imprisonment after 1933, and his eventual permanent expulsion (including being stripped of his German citizenship) from Germany in 1937.

During his father's enforced absences in the difficult years between 1933 and 1938, Reinhard Bendix sought answers in meetings and discussions held by *Neu Beginnen* (a socialist-communist underground movement) and by *Hashomer Hatzair* (a socialist-Zionist secular Jewish youth movement), both Berlin groups that would be subsequently banned and their participants persecuted. As he reflected in the last two books he wrote, *Unsettled Affinities* (1993) and *From Berlin to Berkeley: German-Jewish Identities* (1986), neither left-wing ideology nor Zionism provided satisfactory answers to his questions, and throughout his life, Bendix remained acutely aware of his marginal status as an intellectual in an often antiintellectual American society, and as an heir to a religious tradition he did not and could not identify with. That personal disquiet was the starting point for rational analysis; both father and son shared a penchant for abstracting and objectifying their own situations (Bendix, 2004).

At Chicago, Reinhard Bendix found himself among professors (including Louis Wirth, Charles Merriam, and Leonard White) who were 'partisans of impartiality,' which he found exciting, as he wryly put it many years later, "because only a short time before I had experienced enough partisanship under the Nazis to be ready for a more rational approach to seeking knowledge" (Bendix, 1993: 11). Yet he found the sociological approach he learned at Chicago troubling, as it was inspired by the natural sciences and concerned with formulating hypotheses and subsequently trying to empirically verify or falsify them. While he was not averse to drawing on quantitative data (Bendix, 1949, 1956; Lipset and Bendix, 1959), for his taste such an approach seemed to rely too heavily on demonstration and not nearly enough on discovery and initial exploration, or even consideration of what was worth knowing. Reason mattered in social inquiry, but it should not be overshadowed by methodology or a blind adherence to forms of scientism. From his father, Bendix had also inherited an interest in the irrational foundations of rationality, and these various strands led to his MA thesis (Bendix, 1943) as well as to a number of sweeping essays later on the purpose of knowledge, the distrust of reason, the image of man, and the role of ideology (Bendix, 1984, 1988). Many of his concerns were expressed in his Presidential Address to the American Sociological Association in 1970 when he argued that an abiding faith in reason served as a counter to taking mindless action, and that it was important to protect a scholarly community that prized free discussion and independent inquiry.

When his *Work and Authority in Industry* (1956), a comparative and historical study of managerial ideologies, was republished in 1974, he wrote in a retrospective introduction that his impulse to conduct this study had arisen to counter the Marxist notion that ideas were simply epiphenomena of material production processes, and had not come about only due to his penchant for using wider frameworks than were common among his American colleagues at the time. At the

same time, the reasons employees subordinated themselves to the authority of employers were worthy of analysis, not least due to ideas prevalent in postwar America about rationality and efficiency in large-scale, hierarchically organized, economic enterprises. The comparative and historical vantage point employed in *Work and Authority in Industry* thus challenged received notions about rationality and subordination, and from a non-Marxist position, illustrated the relative autonomy of ideas and their use in managerial contexts.

Reinhard Bendix's experiences as a young man of the use and misuse of authority in a highly ideologically charged atmosphere were intellectually nurtured by Karl Mannheim's *Ideologie und Utopie* (1929) and Hans Barth's *Wahrheit und Ideologie* (1945). Coupled with his study of civil servants (Bendix, 1949), and then of managerial ideologies, this drew him to study Max Weber's work on domination and authority more intensively than he had previously. While not always agreeing with him (see Bendix, 1998), Bendix discovered an 'elective affinity' (*Wahlverwandschaft*) to Weber's work as well as to the comparative and historical methods Weber had used, and this deepened interest would lead him to publish *Max Weber: An Intellectual Portrait* (1960).

English translations of parts of Max Weber's vast oeuvre were available (Gerth and Mills, 1946; Henderson and Parsons, 1947), but no general road map, overview, or introduction to the entire body of Weber's work existed at the time in English. Reviewers were both grateful as well as effusive in their praise, with H. Stuart Hughes in the American Historical Review calling it "the full scale study that has long been needed" and an "extremely useful key" for those who wanted an introductory guide, as well as a "godsend" for students, while another reviewer in the American Anthropologist praised it as "an admirably straightforward exposition of the whole body of Weber's work." In his own introduction, Bendix self-deprecatingly noted that "the plain fact is that Weber's work is difficult to understand." Though he preferred to see himself as more of an interpreter than a proponent (Bendix, 1993: 12–14), publication of his intellectual portrait rapidly elevated him into one of the leading Weberians among sociologists.

But he was unwilling to be limited to only interpreting Weber, or to solely focus on industrial sociology, and thus turned to an examination of the global changes to political authority then underway: decolonization had meant 50 new states were added to the global map between 1944 and 1964. Many efforts to create different or new forms of state and society were underway, and such 'modernization' or 'economic development' processes were analyzed at length in works by Daniel Lerner, Rupert Emerson, Clifford Geertz, and Alexander Gerschenkron. Walt Rostow's stages of growth model, a five-phase scheme that ran from 'traditional society' through 'takeoff' to 'the age of high mass consumption,' was particularly influential. Bendix, however, remained skeptical of such interpretations, and was troubled by what often appeared to be neoevolutionary (if not teleological) assumptions in such works. Once a certain kind of social and political change had taken place, he argued, it could never again occur in exactly the same manner, so the path taken by European nations should not be seen as normative or determinative for the new nations of Africa and Asia. Historical experience was unique, limited, individual, and bound by space, time, and circumstance. His influential article 'Tradition and Modernity Reconsidered' (1967) was an early, explicit objection to the belief in the inevitability of 'modernization' trajectories.

His interest in transformation prompted him, in *Nation-Building and Citizenship* (1964), to examine these assumed 'Western' precursors of 'Third World' economic development in order to better understand what the conditions for political and social transformation might be. Drawing on the work of T.H. Marshall, and writing in collaboration with Stein Rokkan, he identified a key element of such transformations in the extension of civil, social, and political rights to previously excluded groups in European societies. This altered earlier relationships of dependence and fundamentally changed the roles the state and its civil servants, authority itself, would play. In *Nation-Building and Citizenship*, these insights were first used to discuss historical transformations in England, Western Europe, and Russia. Bendix then drew a more detailed comparison of the preconditions for the development of public authority in Japan and Germany, following it with a longer treatment of how public authority had emerged in the 'developing' political community of India. This work is clearly quite Weberian in inspiration, methods, and case selection.

In a very real sense, *Nation-Building and Citizenship* was a practice run for *Kings or People: Power and the Mandate to Rule* (1978). Published after 10 years' research, and a National Book Award finalist (1979; History), this was a sweeping examination of the development of political authority in England, France, Germany, Russia, and Japan. The first part of the book examined the sacred and secular basis for kingly rule through the early modern period; the second part, starting with the transformation of Western societies in the sixteenth century, examined the emergence of the modern nation-state in each of these five countries (there were also briefer consideration of the paths China and the Muslim world had taken). The book was exactly what the subtitle promised: an examination of what it meant to have a mandate to rule, both historically and comparatively, in the key countries that helped shape the world in which we now live.

Reinhard Bendix regarded intellectual work as a conversation rather than as a battle of egos and wills for dominance, an ongoing dialogue about matters of importance that was conducted between past, present, and future generations. The vital issues of the day certainly mattered to him, but so did the older, deeper discourses to which they belonged. That long, measured, European view prompted one of his long-time colleagues to once describe him as "an academic's academic."

When he liked a particular book (and many produced by Albert Hirschman fit that category), he would speak of "books to conjure with": it is safe to say that he produced several books like that himself. Bendix was also deeply involved in the university community that was his second home, not only in numerous administrative roles, but also as a committed and effective teacher whose graduate students would go on to hold significant professional positions in sociology, history, and political science. Yet in his view, the comity among colleagues and the productive intellectual ferment he had encountered in the United States in the postwar decades (as well as in the Western Europe he frequently returned to teach in as well as

earn several honorary doctorates from), had begun to wane and when he retired in 1986 he was grateful to gracefully exit from debates and intellectual discourses he no longer felt an affinity for. What he left behind were lasting intellectual bridges between Germany and the United States that emphasized an awareness of historical contingency as well as of the demonstration effects nation-states exerted on one another, and of the endless caution that is needed when attempting to generalize.

See also: Authority, Social Theories of; Nation, Sociological Aspects of; Nations and Nation-States in History; Political Sociology; Social Scientists as Experts and Public Intellectuals; Weber, Max (1864–1920).

Bibliography

Bendix, John, 1998. Comparison in the work of Reinhard Bendix. Sociological Theory 16 (3), 302–312.

Bendix, John, 2004. The lived analytic: layers of meaning (fullness) in the context of the Holocaust. History of the Human Sciences 17 (2/3), 125–145.

Bendix, Ludwig, 1968. In: Weiss, Manfred (Ed.), Zur Psychologie der Urteilstätigkeit des Berufsrichters und andere Schriften. Luchterhand, Berlin.

Bendix, Reinhard, 1943. The Rise and Acceptance of German Sociology. Thesis (M.A.). University of Chicago.

Bendix, Reinhard, 1949. Higher Civil Servants in American Society. Thesis (Ph.D.). University of Chicago.

Bendix, Reinhard, 1956. Work and Authority in Industry. University of California Press, Berkeley.

Bendix, Reinhard, 1960. Max Weber: An Intellectual Portrait. Doubleday & Co., New York.

Bendix, Reinhard, 1964. Nation-Building and Citizenship. John Wiley & Sons, New York.

Bendix, Reinhard, April 1967. Tradition and modernity reconsidered. Comparative Studies in Society and History IX, 292–346.

Bendix, Reinhard, 1978. Kings or People: Power and the Mandate to Rule. University of California Press, Berkeley.

Bendix, Reinhard, 1984. Force, Fate and Freedom: On Historical Sociology. University of California Press, Berkeley.

Bendix, Reinhard, 1986. From Berlin to Berkeley: German-Jewish Identities. Transaction Publishers, New Brunswick.

Bendix, Reinhard, 1988. Embattled Reason: Essays on Social Knowledge, vol. 1. Transaction Publishers, New Brunswick.

Bendix, Reinhard, 1993. Unsettled Affinities. Transaction Publishers, New Brunswick.

Bendix, Reinhard, Roth, Guenther, 1971. Scholarship and Partisanship: Essays on Max Weber. University of California Press, Berkeley.

Gerth, Hans, Mills, C. Wright, 1946. From Max Weber: Essays in Sociology. Oxford University Press, New York.

Henderson, A.M., Parsons, Talcott, 1947. Max Weber: The Theory of Social and Economic Organization. Oxford University Press, New York.

Lipset, Seymour Martin, Bendix, Reinhard (Eds.), 1953. Class, Status and Power. Free Press, Glencoe.

Lipset, Seymour Martin, Bendix, Reinhard (Eds.), 1959. Social Mobility in Industrial Society. University of California Press, Berkeley.

Benedict, Ruth (1887–1948)

Judith [Modell] Schachter, Carnegie Mellon University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A US cultural anthropologist Ruth Fulton Benedict is best known for two books, the 1934 *Patterns of Culture* and the 1946 *The Chrysanthemum and the Sword*. Each contributes importantly to the central premises of 'culture and personality' studies in US anthropology. Both books demonstrate the close connection between descriptive intensity and ethnographic reliability. Both, too, treat the theories and methods of a science in terms of their significance for ensuring human rights in any society. A foremost anthropologist, Benedict also wrote for a general audience, addressing the Great Depression, racism, and the rise of Fascism and Nazism in the 1930s.

A US cultural anthropologist Ruth Fulton Benedict is best known for two books, the 1934 *Patterns of Culture* and the 1946 *The Chrysanthemum and the Sword*. Each contributes importantly to the central premises of 'culture and personality' studies in US anthropology. Both books demonstrate the close connection between descriptive intensity and ethnographic reliability. Both, too, treat the theories and methods of a science in terms of their significance for ensuring human rights in any society. *Patterns of Culture* gave a wide audience a new concept of culture: an integrated and distinctive whole that molds the temperaments and talents of individuals. *The Chrysanthemum and the Sword* penetrated into the mechanisms through which 'culture' and 'personality' reinforce and replicate a dominant ethos. Both books assert the importance of individual creativity, a bow to the humanitarian impulse Benedict never lost.

The Life

Ruth Fulton was born on 5 June 1887, in New York City, the daughter of a Vassar graduate and a doctor of medicine. Two years after her birth, her father died and the vision of his lost opportunity for serving humankind haunted her for the rest of her life (Modell, 1983). After a childhood spent moving from city to city, in 1905 Ruth and her younger sister Margery entered Vassar College. After graduation, Ruth returned to her mother's house in Buffalo, where she worked for the Charity Organization Society of New York. The experience increased her sensitivity about her partial deafness and exposed her to the struggles of immigrant families in an industrial city. Dissatisfaction persisted, unresolved by teaching in a private school and living with Margery and her husband Robert Freeman in Pasadena, California. A marriage proposal from Stanley Rossiter Benedict seemed to resolve the doubts about her future. Like other women of her generation, Ruth hoped the role of wife would give her life a purpose.

A month after their marriage, World War I broke out. The crisis accentuated Benedict's despair at the limitations set for women, nonconformists, and outsiders in her milieu. She tackled the problem through essays on three 'rebellious' women, Mary Wollstonecraft, Margaret Fuller, and Olive Schreiner; never published, the manuscript reveals the writer's impatience with the conventions that restricted her own options. In 1919, Ruth Benedict enrolled in courses at the New School for Social Research, where encounters with sociologist Elsie Clews Parsons and anthropologist Alexander Goldenweiser transformed her thinking. A meeting with Franz Boas at Columbia University completed the process, and in 1921 Benedict enrolled in his anthropology program (*see* Boas, Franz (1858–1942)). A remarkable 2 years later, she received her PhD for a dissertation on 'The Concept of the Guardian Spirit in North America.'

During the 1920s, Ruth Benedict viewed her commitment to the discipline in the light of her quest for identity as a woman in the US society. Personal concerns provided an impetus for anthropological inquiries, while at the same time intervening on the path to a conventional academic career. During this period, too, the strains between her and Stanley increased, and in 1930 they agreed on a formal separation. By then Benedict had formed strong bonds with her student Margaret Mead and with Edward Sapir, her colleague. Conversations about the poetry all three wrote prompted Benedict to prove that anthropology could incorporate the voice she had developed as a poet. Her aesthetic principles contributed as much to her anthropology as the advice about culture and person offered by Boas.

Ruth Benedict took on teaching duties at Columbia, trained students for the field, and in 1924 assumed editorship of the *Journal of American Folk-Lore*, a position she held until 1940. External obstacles to success in academia became apparent when, after Boas retired in 1937, she did not receive the headship of Columbia's Anthropology Department. The appointment of the anthropologist Ralph Linton surprised her colleagues, seeming to many to indicate the continued discrimination against women in US society. In reaction to Linton's appointment, to her abiding restlessness, and to her equally abiding belief in the scientist's duty as citizen, Ruth Benedict moved to Washington to serve her government during the war. She died on 17 September 1948, in New York City, where she had returned. She was 61 years old and in the midst of formulating a major theory of self and society.

The Anthropology

The 1934 publication of *Patterns of Culture* placed Ruth Benedict in the forefront of American anthropology (*see*

Anthropology: Overview). The book pulled together ideas she had developed in earlier pieces: the significance of coherence and of integrating customs, the social construction of marginality and deviance, the uses of a comparative method for establishing a relativistic stance. In the book's five discursive and three ethnographic chapters, the anthropologist set the foundation for her theoretical and methodological contribution to contemporary social sciences.

The book is best remembered for the notion that 'culture is personality writ large.' The theoretical claim receives support in the ethnographic chapters, which treat three distinct cultures as unitary, integrated, and patterned entities with characteristic personalities. Benedict's concept of culture did not exclude individuals, but explained the links between person and group. The links, she wrote, could be harmonious, leading to the 'successful' individual, or they could be discordant, driving an individual into nonconformity, marginality, or madness. Three cultures form the triptych for an application of her theory, the Pueblos of the American Southwest, the Dobu Islanders, and the northwest coast Kwakiutl. Each of these cultures receives a 'tag,' the famous Apollonian, paranoiac, and megalomaniac designations that for many readers obscured Benedict's concern with the processes by which a cultural configuration comes into being, persists over generations, bends to change, or collapses under crisis.

Methodologically *Patterns of Culture* introduces interpretive techniques that remain significant to anthropological analysis today. Ruth Benedict's reluctance to spend time in the field went hand in hand with an imaginative reading of available evidence from disparate details of behavior reported to her by informants, stored in ethnographies, or enshrined in items of material culture. Constant reference to US culture further extends her method, drawing comparison away from a simple juxtaposition of traits and toward the assessment of diverse 'arrangements' that became central to her work of the next two decades. In *Patterns of Culture*, Benedict shows brilliantly how the social scientist can maintain relativism in description while taking a stand on the implications of an 'arc' of traits (1934: p. 24).

Over the next decade, she published articles on classic anthropological topics in the 1933 *Encyclopedia of the Social Sciences* and simultaneously probed into the causes and consequences of the Great Depression for a general audience. In periodicals like the *American Scholar* and the *Partisan Review*, she brought an anthropological perspective to the subjects of, for example, unemployment, the 'problem of youth,' and the value of freedom of speech. Drawing on her discipline's comparative method, she mounted a stern critique against the United States. She criticized her country for wasting the resources of its population, for discriminating against particular groups, and for nurturing neuroses in its citizens. Professional and popular pieces preached an enlightened social engineering that drew lessons from cultures in which no individual was denied social participation or deprived of fulfillment.

Benedict's inquiry into the patterns of American culture led her to take on issues of race and class – related, in her framework, to democracy and social justice. Consequently, she examined the spread of totalitarianism under the lens of a comparative analysis of institutions and ideologies that deny the humanity of selected groups. From her perspective, understanding the rise of Nazism and Fascism meant discovering the conditions under which individual citizens detect no disjunction between their commonplace assumptions and the dictates of a political ideologue.

On the eve of World War II, Ruth Benedict reconsidered the implications of cultural relativism and its nonjudgmental stance (*see* Cultural Relativism, Anthropology of). Like her mentor Franz Boas, she had argued that every culture should be treated in terms of its own values and purposes. By the end of the 1930s, she sought a scheme for evaluating cultures that respected difference while also disclosing the danger some arrangements clearly posed. Danger came in terms of the suppression of individual creativity, fulfillment, and satisfaction; danger came, more dramatically, in acts of murder and genocide. Reacting to world events, Benedict fashioned the material for a theory of self and society in three separate endeavors: an article on child rearing, a book on race and racism, and a series of lectures on the theme of 'synergy' (the effective integration of all aspects of society).

In 1938, Ruth Benedict published a seminal piece on child rearing, entitled 'Continuities and Discontinuities in Cultural Conditioning.' The article appeared in *Psychiatry*, a journal that fused psychiatry and social science (*see* Cross-Cultural Psychology). The piece reveals Benedict's ambivalence toward Freudian theory and her conviction that social conditions provide the framework for personality development. Her argument then sounds commonplace now: cultures should foster consistency between childhood and adulthood, so that traits learned early are appropriate to behaviors required later. At every turn, 'Continuities' criticizes US culture where, by contrast, dominant values result in sharp discontinuities between childhood and adulthood. The outcome, the anthropologist concluded, is the 'neurotic personality' Karen Horney described. Under ideal conditions in any culture, Benedict wrote, children are conditioned into responsible social participation, their capacities fully utilized.

Simultaneously Benedict struggled to satisfy the terms of an invitation to lecture on 'the individual and society' at Bryn Mawr College in February 1941. An opportunity to expand her theories, the forum challenged her to delineate systematically the relationships between individual capacity, favored temperaments, social institutions, and cultural values. By then, too, the anthropologist explicitly recognized the urgency of calculating "the consequences for human life of different social inventions" (1970: p. 322).

Only notes remain from these crucial and complicated lectures to show how far Benedict had moved from *Patterns of Culture*. The move took three forms: greater attention to the psychodynamics of personality development; an acute analysis of the institutions through which individual behaviors are organized; and an effort to develop a plan for social change that acknowledged the significance of customary beliefs and behaviors. She chose the word 'synergy' to represent the effective working together of parts, and she compared low synergy with high synergy societies. The lecturer praised 'high synergy' societies, in which cooperation is at a maximum and all acts are mutually reinforcing. In such societies, individuals achieve a strong sense of responsibility and well-being. The unanswered question appears as well: what factors produce high

synergy, which low synergy in a society? Her wartime writings pushed her closer to an answer.

A third endeavor refined Benedict's theory of self and society. In the 1940 *Race: Science and Politics*, Benedict analyzed discrimination and racism in her own society. She mounted her strongest attack against a presumed scientific fact of innate endowment, the rationale for prejudices that condemned 'Negroes' (in the language of the times) to marginal participation in US society (*see* Racism, History of). Using comparative examples as well as historical references, she accused her country of fostering divisiveness, crushing the will of individuals, and undermining its own stamina.

The book acknowledged a debt to Boas in insisting that anthropology turn its methods and theories to the solution of contemporary problems. War made the obligation crucial for Benedict, and in the winter of 1942 she assumed a position at the Office of War Information (OWI) in Washington, DC. Obituaries she wrote when Boas died, in December 1942, convey the viewpoint that brought her into the government: "He never understood how it was possible to keep one's scientific knowledge from influencing one's attitudes and actions in the world of affairs" (quoted in Modell, 1983: p. 274). In the OWI, Benedict pursued 'national character' studies; the last assignment put on her desk was 'Japan.'

"The Japanese were the most alien enemy the United States had ever fought in an all-out struggle." So reads the first sentence of Benedict's 1946 publication, *The Chrysanthemum and the Sword*. The wartime study validates the place Benedict already had in her discipline and indicates the directions she would have pursued had she lived longer than 2 years after the book's publication. The study of Japan continues the themes she raised throughout the 1930s: how to ensure the well-being of members of diverse cultures without violating the integrity of each culture, how to turn anthropology's comparative method to the service of international organizations devoted to peace, how to exercise responsibility as a social scientist without transgressing the rules of evidence and neutrality. The failings of the book, like its strengths, can be attributed to the complexity of Benedict's goals.

Despite thickly textured descriptions, the book is less an ethnography of Japanese culture than an expansion of anthropological method and theory. Benedict turned the impossibility of doing fieldwork into a virtue, stressing the importance of interpretation over immersion. The anthropologist, she wrote, "can draw up his hypotheses and get his data in any area of life with profit. He can learn to see the demands any nation makes, whether they are phrased in political, economic, or moral terms, as expressions of habits and ways of thinking which are learned in their social experience" (1946: p. 13). Her data are the 'commonplaces' any Japanese would recognize the behaviors and beliefs an ordinary person deems natural.

The result is a persuasive display of the dominant 'patterns' of Japanese culture without the homogenizing descriptive tag of her earlier efforts. Benedict shuns the isomorphism implied by 'culture and personality,' while recognizing the importance of consistency between 'society' and 'self' for maintaining high synergy. Japan in World War II was notably synergistic: every act and every attitude seemed to the observer–interpreter to flow harmoniously, the Emperor's commands coordinating with the motives and feelings of Japanese citizens. This was Benedict's test case, exemplar of the high synergy society she had praised in 1941. The book explores the processes through which personal impulses, tightly constructed cultural demands, and the needs of a nation are channeled smoothly together, without frustration or conflict.

Critics of *The Chrysanthemum and the Sword* gloss over Benedict's delineation of process, and mistakenly consider the designation of Japan as a shame culture the book's major thesis. In the study of Japan, however, *emotion* is an analytic concept that allows the anthropologist to link the drives of individuals to the maintenance of social order and cultural consistency. 'Shame' pushed Benedict's ongoing exploration of self and society into its most radical formulation yet, as she demonstrated the delicate and daily ways in which persons internalize the pressures of their surroundings without experiencing lethargy, anger, suffocation, or despair. Chapters display the abundance of reinforcing activities and attitudes that make shame both a powerful sanction and a barely perceptible burden for the Japanese individual.

Ruth Benedict bequeathed to social science a theory of the social construction and the social uses of an emotion. She argues that mechanisms for instituting and fostering an emotional pattern crucially support the contemporaneous conditions in which a society finds itself; change in the conditions can render the emotion dysfunctional – a conclusion she reached about Japan, as the nation encountered devastating defeat in the war. She vividly evokes the circumstances under which an emotional pattern wastes the resources and erodes the capacity for meaningful social action by individuals and institutions.

A comparison between Japan and the United States runs through the book, permitting a sophisticated analysis of shame and guilt that anticipates later recognition by social scientists of the twinning of these emotions. At the same time, Benedict's concern with the collapse in Japanese society after the war biases the comparison, so that shame comes to seem an immature and confining emotion. On a theoretical level, however, her discussion of emotion is value free. She does not posit a hierarchy, in terms either of human or social development, but rather assesses the function of an emotional configuration as it releases or represses the expression of human creativity – the humanism in her science.

An emphasis on the social construction of emotion completed Benedict's critique of Freudian theory. If individual emotions are socially constructed, then they are alterable through planning and careful calculation of conditions. Recognizing the complexities of this assumption, Benedict developed strategies for testing her theoretical premises in new settings. She undertook several projects after the war, including child rearing studies for UNESCO, participation in the US Navy's 'cultures at a distance' project, and further comparative inquiry into emotional configurations (*see* National Character).

The Accomplishment

Ruth Benedict's contribution to social science comes from her ability to combine critical theory with cultural relativism. She accomplished this by constructing a scheme for assessing

cultures in terms of the integrity of purpose and coherence of 'traits.' The scheme included inserting the 'human being' into an analysis of social structures in a way that foreshadows a theoretical framework dominated by the concepts of 'actor' and 'agency.' Benedict, however, did not simplify those concepts or use them to signal a desirable relation between 'person' and 'culture.' Rather, she construes 'actor' as constituted at once by idiosyncratic drives and by external shaping forces; she connects 'agency' to the possibilities of self-fulfillment and individual creativity. Her attempts to explore systematically the conditions under which the ensuing tension between self and society is productive rather than destructive banished the simplistic 'culture and personality' equation of earlier works. A founder of that subfield of US anthropology (*see* Anthropology, History of), Benedict insisted on the humanism that is key to contemporaneous research in diverse social scientific disciplines.

Inasmuch as her comparative analyses have at core the human being – creative, cooperative, resistant, rebellious – they impose a rigorous critique on all cultures. The body of her work asks, "does a particular social arrangement suppress or exploit diverse individual capacities?" The question pushes Ruth Benedict into the vanguard of critical social theory in two respects: she does not abstract the individual from the material conditions that govern social responsibility, and her emphasis on the diversity of purposes and temperaments in any setting anticipates a twenty-first century deconstruction of the concept of culture. Prefiguring her successors in the discipline, Benedict replaced the definition of culture as a static collection of traits with the notion of an open-ended, dynamic, and continuously created mode of life. Her bold move toward emphasizing the possibilities for an integrating culture, in which self and society evolve according to a dominant purpose, makes her work a model for today's interdisciplinary critical social theory. Moreover, the openness of her concept of culture broke down the boundaries between 'us' and 'them,' bringing the United States onto the same plane as the smallest tribal society.

Behind these innovative contributions lay Benedict's persistent commitment to the application of her discipline. While she embraced cultural relativism, she also insisted that the anthropologist interpret diverse cultures in terms of their benefits to the individual. She perceived no contradiction between 'neutral' research and a judgment of 'worth.' From her point of view, the findings of any science ought to be devoted to improving the human condition: preventing war, ensuring peace, and, above all, granting anyone in any milieu access to resources and to rights. Stressing the importance of directed change, Benedict did not preach an imperialistic social engineering. She argued instead that the social scientist had unique skills with which to delineate the cultural values and social institutions upon which change must be based in order to last.

In Ruth Benedict's writings, humanism extends from theory to methods of inquiry and styles of presentation. She knew, as poet and as anthropologist, the importance of writing, and her words carry her argument through evocation and nuance. In this, too, she anticipated a twenty-first century concern with disciplinary writing and the extension of anthropological findings to a wide audience. For Benedict the value of a social science lay in its persuasive embodiment of an ethical stance. Here again she makes a major contribution: writings that appeal to a general audience, to policy makers and to soldiers, to social scientists and to artists, presage a contemporary scholarship that refuses to remain in an ivory tower.

See also: Anthropology, History of; Boas, Franz (1858–1942); Cross-Cultural Psychology; Cultural Relativism, Anthropology of; Mead, Margaret (1901–78); National Character; Racism, History of; Sapir, Edward (1884–1939).

Bibliography

Benedict, R.F., 1934. Patterns of Culture. Houghton Mifflin, Boston, MA.

Benedict, R.F., 1938. Continuities and discontinuities in cultural conditioning. Psychiatry 1, 161–167.

Benedict, R.F., 1940. Race: Science and Politics. Modern Age Books, New York.

Benedict, R.F., 1946. The Chrysanthemum and the Sword. Houghton Mifflin, Boston, MA.

Benedict, R.F., 1970. Synergy: some notes of Ruth Benedict. Selected by Maslow A., Honigman J.J. American Anthropologist 72, 320–333.

Modell, J. Schachter, 1983. Ruth Benedict: Patterns of a Life. University of Pennsylvania Press, Philadelphia, PA.

Benhabib, Seyla (1950–)

Leandro Aramburu, Instituto de Investigaciones Gino Germani, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Seyla Benhabib is currently one of the most prestigious intellectuals across the globe. She was born in Istanbul, Turkey, in 1950 and she was brought up within a multilingual context. Later, she moved to the United States of America to pursue her PhD and then to start her academic career. She received the intellectual heritage mainly from German thinkers such as Kant, Hegel, Arendt, and Habermas. Thus, Benhabib's work is an attempt to perform a review of the unfinished project of the Modernity. The article aims to draw an overview about the personal, intellectual, and professional life of Seyla Benhabib.

Introduction

Seyla Benhabib is currently one of the most prestigious and interesting intellectuals across the globe. She comes from Turkey where she spent her childhood and adolescence. She was brought up within a multilingual context, due to the fact that her nuclear family spoke several languages including Turkish, French, Italian, and Ladino. She was educated in English and American schools, thus English became her language of instruction. At the age of 20 she earned a scholarship to move to the United States of America (USA) in order to carry on with her university studies. Further on in time, she received a PhD at Yale University with a dissertation on Hegel's philosophy of right. Her works have been influenced mainly from German thinkers such as Immanuel Kant, Georg Hegel, Hannah Arendt, and Jürgen Habermas. In fact, she defines herself as a specialist on European and German thought (Kreisler, 2004).

Benhabib has taught in several universities in the USA including Yale, Harvard, and Boston, among others; she has also lectured in many different countries worldwide on relevant topics for both academic and general audiences. She also held visiting professorships and research fellow positions at many well-known universities and research institutions around the world. Furthermore, she has a broad intellectual production, which is mostly formed by scientific journal articles, books, and editions of books. To date, her work has been translated into at least 11 languages in which we can find German, Spanish, French, Chinese, Japanese, Turkish, Russian, Italian, Servo-Croatian, Hebrew, and Swedish. Along 35 years (and it continues), she has been performing a brilliant academic career, working as a professor at the most highly renowned institutions in the USA.

According to some scholars (Campillo, 2006; Sánchez-Muñoz, 2009; Fernández de Sevilla, 2011), Benhabib's work is an attempt to perform a review of the unfinished project of the Modernity. In this regard, her work shows long-standing concern and interest with the same basic question. That question addresses the reconciliation of fundamental universalistic principles, such as human rights, autonomy, and freedom, with the existence of specific identities as members of particular communities separated by language, ethnicity, and religion (Kreisler, 2004). Nevertheless, this main question was studied by Benhabib through different relevant topics, which were appearing to her throughout the years. She has been focusing on issues, such as Critical Theory, Deliberative Democracy, Law, Globalization, Cosmopolitanism, Feminism, Hannah Arendt's thought, Immigration, Ethnicity, and Culture, among others.

Indeed her work can be classified into three main stages due to those main topics of study (Sánchez-Muñoz, 2009). The first stage of her work began during the 1980 decade as a critique and reformulation of Critical Theory, especially around the theory of communicative action of Jürgen Habermas. The second one takes place in the 1990 decade when her work turned to a more empirical approach. In this stage, the main question is how that contextualized subject can be included in the specific liberal democracies without losing their own cultural identity. The last and current stage began in the beginning of the twenty-first century, when Benhabib started to analyze the existing tensions between the rights of the people who live in a globalized world but whose citizenships are, at the same time, fixed to particular territorial spaces.

Therefore, the aim of this article is to give an overview of the personal, intellectual, and professional life of Seyla Benhabib until today. The article uses the following structure. The present Introduction, i.e., this first section, contains the general ideas of the whole article. This is followed by the second section, which reviews Benhabib's family background and her childhood and youth in Turkey. Taking a close look at her personal and family background suggests how different issues could fit well into her research interests and political concerns later when she became an intellectual with a strong political commitment. The third section describes her education at Brandeis University and Yale University in the USA, and the academic positions that she held afterward. The fourth section deals with the issue about the political commitment of doing political philosophy and with the role of Benhabib as a committed intellectual. The fifth section is concerned with her research interests and her intellectual production, which are classified into three stages.

Multicultural Family Background, Childhood, and Youth in Turkey

Seyla Benhabib was born in Istanbul, Turkey, in 1950 within a Jewish family. Indeed she was raised within a multicultural and multilingual environment. At that time, Istanbul was (and still is) one of the largest cities in Europe, where a rich cultural diversity has assembled over several centuries. In fact, it was continuously the capital city of four empires, from the Roman Empire (330) until the end of the Ottoman Empire (1922).

Benhabib's family has a rich and interesting cultural background. Her ancestors were Sephardic Jews who came to Turkey in 1492 when the Alhambra Decree of the Spanish Kingdom forced the Jewish community either to convert to Christianity or to leave Spain. The Ottoman Empire was somehow open to receive that immigration in order to seize part of the eastern Mediterranean commerce.

Her nuclear family also has a quite remarkable story that could have strongly influenced Benhabib in several ways. On her father's side, though her grandfather was a rabbi her father could not complete his studies because he had to join the Turkish Army as soon as World War II started. However, he was released soon because Turkey remained neutral in the conflict until 1945. Then her father went into business. On her mother's side, her grandmother could neither read nor write, as she never had the possibility to attend school. Despite her grandmother's lack of education, Benhabib's mother was educated under French and Italian languages, due to the fact that she attended a French elementary school and later an Italian High School. After finishing high school, her mother got a scholarship in order to go to Italy to study History of Arts, but for family reasons she had to refuse the scholarship and stay in Turkey. A year later in 1939 her parents were married. Benhabib believes that her mother could not study further because of the circumstances of being a woman in Turkey at that time (Kreisler, 2004). Certainly, that fact could have had some remarkable influence in later Benhabib's concerns about the women's rights.

With this parental background, several languages were in use during Benhabib's childhood and youth. Her mother could speak French and Italian besides Turkish, of course. And Ladino, a Romance language spoken mostly by Sephardic Jews, was also used among the members of the family. So, as Benhabib said, they could easily change into each of those languages in different everyday-life situations (Kreisler, 2004). The English language appeared as soon as Benhabib went to high school. Benhabib was educated in an English speaking context since she attended the English High School for Girls for secondary school or junior high, and the American College for Girls for senior high education, where she received a Bachelor's degree in Philosophy in 1970.

The institutions she attended in Istanbul during her youth gave her the chance to get in contact with classmates from different countries, cultures, and religions as both schools were attended by children of an 'elite,' such as children of diplomats or teachers who were working in Turkey, among others. There were also locals with similar non-Turkish family backgrounds, including Armenian and Greek. Furthermore, there were teachers in the American College for Girls, who came to teach in Istanbul instead of going to the Vietnam War. They also brought with them political discussion during the 'lunch hours.' It was around 1968 and that time has been described by Benhabib as "a speedy process of politicization" (Kreisler, 2004). A bunch of events were taking place at that moment, the Vietnam War and the antiwar movement, the student movement with the 'French May' in France, and a remarkable workers mobilization in Turkey around the emergence of an independent left.

As a consequence, she was brought up within a tension between oriental and occidental values. In fact she was Jewish, that is, part of a minority within a primarily Muslim society, but she was also part of the European cultural heritage gathered through her formal and informal education. Therefore, the diversity in Cosmopolitan Istanbul, her family background, and her own personal life must have played a key role in shaping her views of the world, making her sensitive of issues and concerns highly relevant in her later career (Kreisler, 2004). Her personal background, along with the political mobilization around particular worldwide events during the 1960s and 1970s, such as the Vietnam War, the student's and the worker's movements, were surely fundamental experiences that affected her political thought and her intellectual production.

Higher Education Abroad and Academic Positions

After receiving her Bachelor's degree in Humanities at the American College for Girls in 1970, she moved to the USA to carry on with her university studies. She attended Brandeis University where she received a Bachelor's degree in Philosophy in 1972. Brandeis University closely related to the Jewish community and founded in 1948, was at that time one of the youngest private research universities in the USA. The possibility to study in that institution was given by a scholarship granted by the Wien International Scholarship Program (Kreisler, 2004), which is a need-based scholarship awarded to exceptionally international applicants who show strong academic achievement, significant community involvement, and demonstrated financial needs. This scholarship program had been running since 1958 with the aim of funding students from all over the world, who pursue undergraduate studies at Brandeis University.

This scholarship also represented the front door through which Benhabib entered to the American educational system. It would become a platform from which to jump to the highest university level worldwide, reaching Yale University to perform PhD studies and later to start her teaching career. However, when she arrived at Brandeis, it must have been a big shock for her since she was coming from one of the biggest and most cosmopolitan cities of Europe to study in a University located in the small town of Waltham, Massachusetts, with only 60 000 inhabitants.

Once she obtained her BA degree from Brandies, she moved to Yale University to pursue her PhD studies. After 5 years she received a PhD degree in Philosophy at Yale University in 1977. Her dissertation was on Hegel's Philosophy of Right under the title 'Natural Right and Hegel: An Essay on Modern Political Philosophy.' She was interested in Hegel's thought because the German philosopher was "one of the first systematic thinkers in modernity" (Kreisler, 2004), who pointed out the three

principles of the modern State: morality, legality, and ethical life. Those three principles and their particular interrelations have traced the path in which the thought of Benhabib has developed.

Benhabib has stated that she, since her PhD dissertation, has been involving the same central question in her further work. In this regard, she said that this question might not be related only to the research interests that a philosopher might have, but also to 'an existential question' that comes along in their philosophical work. To Benhabib the existential issue that already appeared in her dissertation and seems to remain through the years, deals with the reconciliation between basic universalistic principles, such as human rights, autonomy, and freedom, and specific identities as members of particular communities (Kreisler, 2004). Though an existential question, there are different ways to study the same research object, as she has shown when she changed the direction of her work during the 1990 decade (Wahl-Jorgensen, 2008; Sánchez-Muñóz, 2009). Her work will change the direction again during the 2000 decade (Sánchez-Muñóz, 2009), which will be further discussed in the last section.

Immediately after receiving her PhD in 1977, Benhabib began her teaching career at Yale University, where she joined the Department of Philosophy for 2 years, as an Assistant Professor of Philosophy. Then between 1979 and 1981, she was an Alexander von Humboldt Research Fellow at the Max Planck Institute in Starnberg, Germany. After that she had a position for 4 years (1981–85) at Boston University as an Assistant Professor of Philosophy at the Department of Philosophy. Subsequently she became Associate Professor of Political Theory at Harvard University between 1987 and 1989, Associate Professor of Philosophy and Women's Studies at State University of New York at Stony Brook for the period 1989–91, and finally Professor of Political Science and Philosophy at the New School for Social Research during 1991–93.

Then between 1993 and 2001, she returned to Harvard University, however, this time, she became a Professor of Government in the Department of Government at that institution. At the same time, she was a Senior Research Fellow at the Center for European Studies at Harvard. Also in 1996, she was a Visiting Senior Fellow at the Institute for Human Sciences (Institut für die Wissenschaft vom Menschen, IWM) in Vienna, Austria. From 1997 until she left Harvard in 2001, she further chaired The Committee on Degrees in Social Studies at Harvard. That committee promotes the interdisciplinary concentration of social theory, history, and culture at undergraduate level in order to analyze social, political, and economic problems confronting contemporary societies.

In 2001, she was endowed with a Eugene Meyer Professorship of Political Science and Philosophy from the Department of Political Science at Yale University. Moreover, between 2002 and 2008 she was the Director of the Program in Ethics, Politics, and Economics at Yale. This Program promotes the interplay among empirical, analytical, and normative concerns from different disciplines related to Social Sciences and Humanities. Taking into account the current complex social realities, the program was designed to promote well-founded debate about public policy in view of the reflection on the essentials of human association. From January to July 2009, she was also a Senior Fellow at the Institute for Advanced Studies of Berlin (Wissenschaftskolleg zu Berlin).

Furthermore, Benhabib has been recently awarded with honorary degrees and several prizes. She holds honorary degrees from the University of Utrecht (2004), University of Valencia (2010), and Bogazici University in Istanbul (2012). During 2011–2012, she was a Guggenheim Fellow and Straus Fellow at New York University's Straus Institute for the Advanced Study of Law and Justice. Due to her contributions to ethics and political theory, Benhabib was also awarded in 2013 with the Meister Eckhart Prize of the University of Cologne and the Identity Foundation in Germany. In May 2014, she received an Honorary Doctorate of Humane Letters from the Graduate School of Arts and Science at Georgetown University.

The Political Commitment of Doing Political Philosophy

The aim of this section is to consider Seyla Benhabib as a committed intellectual of our time. Indeed, she has shown a strong political commitment in her career, which is widely demonstrated throughout her intellectual work, civic activities, and lectures given around the world. An important example of her civic engagement is the organization of the 'Istanbul Seminars: Close encounters across all divides,' in which Benhabib, together with other scholars and intellectuals (Giancarlo Bosetti, Alessandro Ferrara, Abdou Filali-Ansary, Nina zu Fürstenberg, Nilüfer Göle, Ferda Keskin, Nadia Urbinati, and David Rasmussen) has been participating in as a member of the Executive Committee since 2007. These seminars take place every year in Istanbul with the support of the 'Foundation Reset-Dialogues on Civilizations (Reset-DoC),' in which Benhabib also carries out the Chair of the Scientific Committee. The aim of the Istanbul Seminars is to bring together thinkers from different cultural backgrounds in order to discuss and debate the prospects for democracy, human rights, and religion within the current global society.

Recently, Benhabib has been awarded with two remarkable prizes due to her dedication and contribution to promote intercultural dialogue and understanding in the context of the global civilizations. The first one is the Ernst Bloch prize that was given to her in 2009 by the *Ernst Bloch Zentrum* of the city of Ludwigshafen am Rhein, Germany. The second one is the Leopold Lucas prize that was awarded by the Faculty of Protestant Theology at the University of Tubingen (Germany) in 2012. Furthermore, Benhabib was invited by the United Nations Democracy Fund (UNDEF) to serve as an individual member on the UNDEF Advisory Board for the period 2010–11. She accepted the honored invitation under the belief that democracy as a global project must be strengthened through the work of civil society organizations worldwide.

As Benhabib has already said, her interest in philosophy was always related to the fact that doing philosophy is a way of thinking about how certain abstract principles are currently "at work, or play themselves out in the course of history" (Kreisler, 2004). In this regard, it is worth mentioning, on the one hand, what being an intellectual means and what is the role of intellectuals in the (world) society in which they live, according to some relevant scholars to the subject. On the other hand, it

takes into account what Benhabib considers about her own work and function as a political philosopher.

According to Bauman (1987: p. 2), the term intellectual is used to describe those who "rise above the partial preoccupation of one's own profession or artistic *genre* and engage with the global issues of truth, judgment and taste of the time." In this sense, Shils (1971) argues that an intellectual is someone who, when participating in a tradition becomes the spokesperson of a particular knowledge. One of the main tasks of the intellectuals, following Shils, is to criticize the traditional beliefs upon which the social organization has been built. In addition, Coser (1968; Isola, 2013) points out that being an intellectual implies the manipulation of ideas and symbols, with the aim to criticize and modify them, to rearticulate them later in a specific *ethos* in order to share and defend them, and certainly to deny them to *others*.

A few years ago in an interview given to Harry Kreisler (2004) in the context of the 'Conversations with History' series of the Institute of International Affairs at the University of California (Berkeley), Benhabib clearly stated that in her opinion, being a political philosopher is 'more a vocation than a career.' In that sense, political philosophers can work at different places doing a variety of activities, such as teaching and researching at universities, being journalists, or even human right activists. Following this distinction, we have to take into account the difference between the terms intellectuals and experts. The latter refers to a 'socioprofessional category,' which means that an expert is someone who can solve specific problems in the path to reach a purpose, according to a particular technique, which is linked to a specific field of knowledge. On the contrary, the term intellectual refers to someone who defines himself some ideals and objectives related to the good of a community. Those objectives are built on its 'symbolic knowledge,' which is based on universal principles. According to that, the main function of an intellectual would be to participate in the public sphere to criticize everything missing in order to reach those ideals (Isola, 2013).

In this regard, following Benhabib's points of view, a political philosopher might have 'a vocation for thinking' about the 'phenomenon of the political.' She points out here that there is an important current of thinking within political science that understands the phenomenon of the political mainly as 'a strategic bargaining of various games,' through which various groups try to get different benefits of any kind, either power, money, or authority. Instead, Benhabib claims for a somewhat more anthropological definition of the phenomenon of the political whereby the political must be understood as the way that "all communities of any degree of complexity organize themselves according to certain principles of justice, equality, reciprocity, and authority" (Kreisler, 2004).

Benhabib stated that doing political philosophy is to deal with particular realities in terms of 'what we are doing' or 'where we are going' according to certain abstract political principles. It is necessary, following Benhabib, to have dedication and passionate involvement in order to bring those principles to the concrete world (Kreisler, 2004). Just a look at all her work can make one realize that she has been pursuing that kind of political philosophy along her entire career. All her writings and lectures are contributions to think about our current democracies from an innovative and critical prism, which is what an intellectual must do.

According to Benhabib (1992; Fernández de Sevilla, 2011), philosophy has the special task to criticize the concepts upon which the beliefs and actions of our society are built. In this regard, Benhabib points out the necessity to perform a theoretical exercise, known as 'social exile,' through which philosophy could have the point of view from outside the mauls and borders of a particular society. It is arguable that her family background and her personal experience during her childhood and youth that made her politically conscious at an early age, and this has certainly been relevant to her intellectual and philosophical activity. Benhabib herself has stated that according to her biographical background, she always had "the sense of being in Turkey but not really being Turkish in the ethnic or national sense, and negotiating those conflicts as [she] was growing up" (Kreisler, 2004).

In the opinion of Fernández de Sevilla (2011), the philosophy of Benhabib is built upon "a model of public moral and political dialogue, free and egalitarian," in which "the elements of uncertainty, change or mobility" perform a noteworthy role. In that sense, Benhabib points out the impossibility to take into account fixed certainties, due to the fact that the fragility of the political institutions and the unpredictability of the practices of governance have frequently been shown throughout the history. Therefore, following Fernández de Sevilla, the political philosophy of Benhabib is based on the concepts of 'deliberation and democratic dialogue.' Indeed, Benhabib believes that only if we understand the conditions of others who might be fundamentally different from us, can we widen the boundaries of our sympathy with them (Wahl-Jorgensen, 2008) in order to build a pluralistic society.

Research Interest and Intellectual Production

Seyla Benhabib's intellectual production is broad, including journal articles, books, and edited and coedited volumes. Some of Benhabib's journal articles can be downloaded for free from her personal Web site at Yale University; two public lectures and an interview (the latter was already cited in this article) can also be seen online. Through her entire work, Benhabib has focused on issues such as Critical Theory, Deliberative Democracy, Law, Globalization, Cosmopolitanism, Feminism, Hannah Arendt's thought, Immigration, Ethnicity, Culture, among others. In the opinion of some scholars (Campillo, 2006; Sánchez-Muñoz, 2009; Fernández de Sevilla, 2011), Benhabib's work can be seen as an attempt to perform a review of the unfinished project of the Modernity. In this regard, her intellectual production deals with the challenges, with which the concept of universalism conceived by the deliberative democracy theory must confront within the new frame of the Globalization phenomenon.

According to Sánchez-Muñóz (2009), Benhabib belongs to the 'third generation' of Critical Theory, which reelaborates and widens the contributions of the 'second generation' mainly around the ideas of Jürgen Habermas. In terms of Shils (1971), Benhabib embraces that 'tradition' of this philosophical current of thinking, within which she shaped her own theoretical and analytical perspective. Throughout her works, she had been

continuously improving that 'inherited pattern of beliefs' by intending to solve new problematic phenomena derivative from the Globalization phenomenon and its varied consequences.

In this regard, the works of Benhabib have criticized and revised fundamental concepts forming the discourse of ethics and the theory of communicative action (Campillo, 2006; Bracci, 2002). To Benhabib (1994) the principle of universalism implies that human beings, due to their humanity, are entitled to 'moral respect from others,' which means they possess the undisputed entitlement to equally enjoy basic human, civil, and political rights. Therefore, her work shows a long-standing concern and interest about the same fundamental question. That question addresses the reconciliation of essential universalistic principles, such as human rights, autonomy, and freedom, with the existence of specific identities as members belonging to particular communities separated by language, ethnicity, and religion (Kreisler, 2004).

Following Sánchez-Muñoz (2009), the intellectual production of Benhabib can be classified into three stages. The first stage of her work began during the 1980s as a critique and reformulation of the Critical Theory, especially around the theory of communicative action of Habermas. Then she provides the Habermasian model with a 'contextualized subject.' The second one takes place in the 1990s when her work turned into a more empirical one. In this stage the main question is how that contextualized subject can be included in the specific liberal democracies without losing their own cultural identity. The last and current stage began in the beginning of the twenty-first century, when Benhabib started to analyze the existing tensions between the rights of the people who live in a globalized world but whose citizenships at the same time are fixed to particular territorial spaces. Her aim is to include the culturally excluded voices through the practice of the citizenship rights in the public sphere.

Three fundamental books of Benhabib belong to the first stage. Her first book is 'Critique, Norm and Utopia,' which appeared in 1986 and contains a strong critique to Critical Theory, focusing on Habermas' work and the concept of universalism. The second book, published in 1992, is 'Situating the Self' and proposes a reformulation of the Habermasian model, providing it with a 'contextualized subject.' In this regard, this work intends to perform a dialogue between the universalism of Modernity and its critiques coming from different currents of thinking, such as the communitarianism, feminism, and postmodernism, in order to produce a new form of universalism: the 'interactive universalism.' The third book is 'The Reluctant Modernism of Hannah Arendt' and was published in 1996. This book also represents a bridge between the first and the second stage of Benhabib's intellectual production.

According to Sánchez-Muñoz (2009), Benhabib finds in Arendt's works the foundations to carry out the reformulation of the Habermasian model of deliberative democracy. A communicative action is also a narrative action, which means that the identity of the subject and its actions are inserted in a 'web of narratives,' i.e., a 'web of interpreted relationships and stories.' Benhabib introduces a 'narrative turn' to the Habermasian model, conceiving human beings as 'story-telling animals.' Thus, Benhabib (1994) proposes the model of the 'moral conversation' in order to achieve reversible perspectives, which permit to understand the conditions of others who might be fundamentally different than one. According to that, Fernández de Sevilla (2011) states that the concepts of "deliberation and democratic dialogue" are the cornerstones, upon which the thinking of Benhabib is built.

During the 1990s Benhabib's work has experienced a significant change of direction, from studying the history of political thought to applying political philosophy to the present time. Therefore, she took a more empirical focus by looking at, and participating in, some concrete and contemporary discussions around new dilemmas that universalism must confront. She began working with issues, such as immigration, citizenship, women's rights, and multiculturalism, which became both academic and politically relevant with regard to the emergence of the European Union (Wahl-Jorgensen, 2008; Sánchez-Muñoz, 2009).

At the beginning of the twenty-first century another fundamental book was published: 'The claims of culture.' Here Benhabib (2002) observes that the unstoppable movement of people crossing state borders and experiencing different European cultures has a counterpart in the revitalization of local cultures. As a consequence, the claims to identity of different cultural groups have forced a rethinking of the concept of citizenship within the context of liberal democracies. How to handle the mixture and exchange of cultures and conflicting interpretations is one of the current challenges that European democracy must face in order to ensure a common future (Gutmann, 2003).

Thus, the third stage in the intellectual production of Benhabib takes place within the current globalized scenario, which involves the appearance of new dilemmas to the reformulation of the deliberative democracy model carried out by Benhabib. In this regard, the new challenges, to which this model must face, are the demands of a 'globalized universalism,' the spread of the 'moral conversation' to the entire humanity, and the building of a 'postnational solidarity project.' All of that is represented by those who are considered foreign elements to a particular nation or society, among which one can find immigrants, refugees, stateless people, and foreigners in general. Thus from a postnational approach, Benhabib outlines the principles of a cosmopolitan theory of justice, which centers on the principles and practices as a member of a political and legal community (Sánchez-Muñoz, 2009; Osborn, 2010).

During that period, the following two relevant books were published. The first one is 'The Rights of Others: Aliens, Residents and Citizens,' appeared in 2004. Here Benhabib states an interesting concept of citizenship as a 'social practice.' In this regard, following Fernández de Sevilla (2011), citizenship is more a pragmatic ability than an essential quality of the human beings. Thus, citizenship is the process and result of the political exercise carried out by the individuals in the public sphere. As a consequence, the notion of citizenship would no longer be attached to a fixed identity, because this identity could change through political practice. In the second book, 'Another Cosmopolitism,' Benhabib (2006; Sánchez-Muñoz, 2009) continues with her analysis concerning the 'processes of democratic iterations,' through which different members of the civil society can renegotiate the universal rights of belonging in both, legal and political

institutions, and civil society associations, within a new cosmopolitan social context.

Conclusion

Throughout this article Seyla Benhabib can be seen as one of the most important living intellectuals of our time. This has been widely demonstrated by her distinguished academic career, her strong political commitment, and her broad intellectual production. Benhabib's original point of view would be partly explained by her own condition as a person who was raised within and was influenced by a rich multicultural context. This background inevitably gave her a somewhat privileged critical view. On the one hand, she performed a reformulation of the theoretical framework derivative from the 'third generation' of Critical Theory and its critics. On the other hand, she deals with highly relevant issues to the building of a pluralistic democracy system in the context of the globalization phenomenon. Benhabib's works should surely be taken into account to better understand current democratic societies, especially those, like the European democracies, that are facing a relevant flux of people coming from different political and cultural environments.

See also: Arendt, Hannah (1906–75); Citizenship, Historical Development of; Critical Theory; Democracy, History of; Democracy: Normative Theory; Ethnicity and Ethnic Groups: Historical Aspects; Habermas, Jürgen (1929–); Hegel, Georg Wilhelm Friedrich (1770–1831); Human Rights, History of; Identity and Identification; Integration and Multiculturalism: Focus on Western Europe; Kant, Immanuel (1724–1804); Multiculturalism; Political Thought, History of; Postmodernism: Philosophical Aspects; Public Sphere: Nineteenth and Twentieth-Century History; Transnational Citizenship; Welfare State, History of.

Bibliography

Bauman, Z., 1987. Legislators and Interpreters: On Modernity, Post-Modernity and Intellectuals. Polity Press, Cambridge.

Benhabib, S., 1986. Critique, Norm and Utopia: A Study of the Foundations of Critical Theory. Columbia University Press, New York.

Benhabib, S., 1992. Situating the Self: Gender, Community and Postmodernism in Contemporary Ethics. Polity Press, Cambridge.

Benhabib, S., 1994. In defense of universalism: yet again! A response to critics of situating the self. New German Critique 62 (Spring-Summer), 173–189.

Benhabib, S., 1996. The Reluctant Modernism of Hannah Arendt. Sage, Thousand Oaks.

Benhabib, S., 2002. The Claims of Culture: Equality and Diversity in the Global Era. Princeton University Press, Princeton.

Benhabib, S., 2004. The Rights of Others: Aliens, Residents and Citizens. Cambridge University Press, Cambridge.

Benhabib, S., 2006. Another Cosmopolitanism: Hospitality, Sovereignty and Democratic Iterations. Oxford University Press, Oxford. With Commentaries by Jeremy Waldron, Bonnie Honig and Will Kymlicka.

Bracci, S.L., 2002. Seyla Benhabib's interactive universalism: fragile hope for a radically democratic conversational model. Qualitative Inquiry 8 (4), 463–488.

Campillo, N., 2006. Seyla benhabib. In: Guerra, M.J., Hardisson, A. (Eds.), Veinte pensadoras del siglo XX. Nobel, Oviedo.

Coser, L., 1968. Hombres de ideas: El punto de vista de un sociólogo. FCE, México. [Original in English: Coser, L., 1965. Men of Ideas: A Sociologist's View. Free Press, New York.]

Fernández de Sevilla, L., 2011. Ciudadanía y democracia: análisis comparativo feminista de dos propuestas diferentes (Norberto Bobbio vs. Seyla Benhabib). La pregunta por el quién, el qué y el cómo de nuestras democracias actuales (Ph.D. thesis). Universidad Complutense de Madrid, Madrid. http://eprints.ucm.es/13134/1/T33114.pdf (last accessed 28.02.14.).

Gutmann, A., 2003. Identity in Democracy. Princeton University Press, Princeton.

Isola, N.J., 2013. Intelectuales, Estado y Universidad: Algunas consideraciones sobre este vínculo complejo. Diálogos Latinoamericanos 20, 134–161.

Kreisler, H., 2004. Philosophic Iterations, Cosmopolitanism, and the 'Right to Rights': Conversation with Seyla Benhabib. Conversations with History. Institute of International Studies at University of California, Berkeley. http://globetrotter.berkeley.edu/people4/Benhabib/ (last accessed 28.02.14.).

Osborn, R., 2010. Seyla Benhabib, Wendell Berry, and the question of migrant and refugee rights. Humanitas 23 (2), 118–138.

Sánchez-Muñoz, C., 2009. Seyla Benhabib: hacia un universalismo interactivo y globalizado. In: Máiz-Suárez, R. (Ed.), Teorías Políticas Contemporáneas. Tirant lo Blanch, Valencia, pp. 271–310.

Shils, E., 1971. Tradition. Comparative Studies in Society and History 13 (2), 122–159. Special Issue on Tradition and Modernity.

Wahl-Jorgensen, K., 2008. On the public sphere, deliberation, journalism and dignity: an interviewed with Seyla Benhabib. Journalism Studies 9 (6), 962–970 (Also available at Reset-Dialogues on Civilizations (Reset-DoC), 08.04.2008). http://www.resetdoc.org/EN/Benhabib-interviewed-by-Karin-Wahl-Jorgensen.php (last accessed 28.02.14.).

Relevant Websites

http://www.gf.org/fellows/16973-seyla-benhabib – Guggenheim Fellowship.

http://nyustraus.org/fellows/11-12-fellows/seyla-benhabib.html – New York University's Straus Institute.

http://socialstudies.fas.harvard.edu/history – Harvard University.

http://www.iwm.at/the-institute/former-visiting-fellows/seyla-benhabib/ – Institut für die Wissenschaften vom Menschen.

http://prelectur.stanford.edu/lecturers/benhabib/ – Standford University.

http://www.yale.edu/polisci/sbenhabib/, http://politicalscience.yale.edu/people/seyla-benhabib, http://philosophy.yale.edu/benhabib, http://www.yale.edu/opa/arc-ybc/v29.n31/story5.html – Yale University.

http://www.un.org/democracyfund/sites/dr7.un.org.democracyfund/files/UU6.pdf – The United Nations Democracy Fund.

Multimedia Sources

Are Democracies still Sovereign? The Challenge of Transnational Institutions. Reset-Dialogues Istanbul Seminars 2012. http://www.resetdoc.org/story/00000021985.

Challenges to cosmopolitanism in Europe. ARENA Lecture 05.02.2012, ARENA Center for European Studies, Faculty of Social Sciences at the University of Oslo. http://www.sv.uio.no/arena/english/services/podcast/arena-lecture-2012-benhabib.html.

Philosophic iterations, cosmopolitanism, and the 'Right to Rights. Seyla Benhabib interviewed by Harry Kreisler. In: Conversations with History, 2004. Institute of International Studies at the University of California, Berkeley. http://conversations.berkeley.edu/content/seyla-benhabib.

Seyla Benhabib interviewed by Nina zu Fürstenberg Interactive Universalism and the Rights of the Others, 29 May 2013. Reset-Dialogues on Civilizations (Reset-DoC). http://www.resetdoc.org/story/00000022249.

Seyla Benhabib interviewed by Nina zu Fürstenberg Migration and Human Rights, 01 March 2010. Reset-Dialogues on Civilizations (Reset-DoC). http://www.resetdoc.org/story/00000021116.

Bentham, Jeremy (1748–1832)

William Twining, University College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Jeremy Bentham (1748–1832) was educated at Westminster, Oxford, and Lincoln's Inn. From an early age he rebelled against the complacency, complexities, and injustices of English law. After a short period in practice, he devoted himself to reasoned criticism and reform of the law on the basis of 'the greatest happiness principle.' Bentham made significant contributions to philosophy, logic, penal reform, social policy, and democratic theory. But nearly all his main ideas stem from juristic concerns. Apart from unsuccessful attempts to promote his scheme for a model prison (panopticon) and his services as 'legislator for the world,' Bentham concentrated on writing. He was enormously prolific, but much of his work was not published in his lifetime. The *Collected Works* will fill a projected 70 volumes, of which 30 had been published by mid-2014. Work on Bentham texts has been accelerated since 2010 by an international project for transcribing and digitizing the hitherto unpublished articles. Bentham's jurisprudence is based on five pillars: the principle of utility; the theory of fictions (ontology and epistemology); the command theory of law; a comprehensive set of codes (the pannomion); and his constitutional theory. He was a follower and critic of Adam Smith in economics, a pioneer in promoting the collection of statistics, a strong legal positivist, a lifelong critic of natural law and natural rights, and a radical democrat. The picture of Bentham that is emerging from recent scholarship presents him as more complex, subtle, and genuinely radical than earlier representations.

Life and Writings

Bentham's father was a prosperous attorney. Hoping that his precocious child would have an illustrious legal career, he had him educated at Westminster, the Queen's College, Oxford, and Lincoln's Inn (called to the Bar in 1769). For Bentham each of these three institutions came to epitomize aspects of the prevailing orthodoxy: conservatism, complacency, hierarchy, superstition, insincerity, and the elevation of authority over reason. As a child, he conceived a passionate antipathy to the inadequacies and injustices of English law and procedure. Later, "English common law and equity seemed to Bentham archaic, uncodified, incomprehensible, arbitrary, irrational, cruelly vindictive, tortuously dilatory, and so ruinously expensive that nine out of ten men were literally outlawed" (Mack, 1968: p. 55). He very soon abandoned legal practice in disgust. Unmarried, and comfortably off, he devoted himself to criticism and reform of legal, political, and social institutions on the basis of utility or 'the greatest happiness principle.'

Bentham's most important diversion from writing was his attempt to lobby the government for a contract to design and manage a model prison, the panopticon. His proposals to William Pitt in 1791 received some encouragement, but then encountered repeated delay and opposition. The project failed, in the process nearly bankrupting him, alienating him from the establishment, harming his reputation with posterity, and pushing him toward democracy. He was eventually granted a sum in compensation for his efforts and expenses in 1813, but the project was effectively dead by 1803. Thereafter, he stayed mainly at home, or in various retreats, writing, conversing with his friends, acolytes and other visitors, and conducting an extensive international correspondence. When he died, he left approximately 70 000 pages of article, mostly unpublished.

Bentham was a secular child of the Enlightenment. The principle of utility came mainly from Helvetius and he owed much to Bacon, David Hume (*see* Hume, David (1711–76)), John Locke, Priestley and European thinkers, including Voltaire, Beccaria, and Charles Montesquieu. But he differed from his predecessors in several respects. He insisted that 'the art-and-science' of legislation required as preliminaries exhaustive taxonomy based on precise concepts and exact analysis of social facts. He had an acute awareness of the importance and limitations of language as the main instrument for expressing abstract ideas, and applied his version of utility and relentlessly detailed analysis to a remarkably wide range of subjects. He adopted clear, often extreme, positions on some issues. For example, he maintained that the only criterion of good and bad, of right and wrong, is the principle of utility; all other criteria are either meaningless or perversions of utility or utility disguised or pure subjectivism; that it is easier to justify torture than punishment; that talk of natural rights is both nonsensical and dangerous; that no artificial rule of evidence can promote rectitude of decision; that the common law is not really law, but a fiction based on judicial usurpation of power; and that the interests of the legal profession and of those in power are in general opposed to those of the community.

Despite this seeming clarity, many of Bentham's ideas invite conflicting interpretations. 'Benthamic ambiguity' is an important thread in reading Bentham. For example, he is often treated as a forerunner of both *laissez faire* and the modern welfare state, yet he does not fit either; he was a genuine political radical, but he placed a high value on security, and on the protection of existing property entitlements; he favored both strong government and strong democratic control; he virulently rejected most features of the common law, yet defended some salient institutions such as orality and cross-examination; the great codifier opposed all binding rules of evidence and procedure. There is no more ambiguous symbol than 'the auto-icon,' Bentham's skeleton, fully clothed and

topped with a wax head, that sits enigmatically in public in a cupboard in University College London, stimulating endless speculation about its possible meanings. Some seeming ambiguities can be resolved textually; some suggest honest and perceptive puzzlement; others hint that behind the benign exterior lay some deep inner conflicts.

Bentham as Jurist

Bentham explicitly aspired to be to jurisprudence what Luther was to Christianity and Newton was to science. His contributions to philosophy, political theory, and social reform were offshoots of his concern with law. Utility was a tool for both criticizing existing laws and institutions and guiding legislation. His ideas on linguistic analysis, on evidence and inference, and on democracy grew out of specifically juristic concerns.

Jurisprudence is the most extensive, original, and neglected part of Bentham's legacy. Yet he is better known as a philosopher, political theorist, and social reformer, for three reasons: few Bentham scholars have been jurists; many of the legal writings have only recently started to become accessible; and Bentham's disciple, John Austin, was for a century treated as 'the Father of English Jurisprudence.' Austin's command theory of law is inferior to Bentham's, but Austin was preferred because he was more sympathetic to the common law, less critical of the legal profession, and less dangerously radical than his master. His general ideas were also easier to grasp and to criticize.

Bentham's science of legislation is based on five main pillars: the greatest happiness principle, the theory of fictions, his positivist command theory of law, the pannomion, and, as a late but crucial addition, his constitutional and democratic theory.

Bentham as Utilitarian: The Greatest Happiness Principle

Bentham's best-known expositions of utility are in two early works, *A Fragment on Government* (1977, hereafter *Fragment*), first published in 1776, and *An Introduction to the Principles of Morals and Legislation* (1996, hereafter *IPML*), first published in 1789. The former is primarily an attack on the complacency of William Blackstone's idealization of the common law and his espousal of natural law and social contract theory at the start of his *Commentaries*. *IPML* was conceived as a prolegomenon to a penal code dealing especially with the classification of offences and the principles of punishment. The opening words of *IPML* are "Nature has placed mankind under two sovereign masters, *pain* and *pleasure*. It is for them alone to point out what we ought to do, as well as to determine what we shall do."

As a utilitarian Bentham is generally regarded as more representative than original. The rhetorical opening of *IPML* is often treated as epitomizing classical utilitarianism, a vehicle for considering standard criticisms on, for example, promising, punishment of the innocent, torture, and egoism. However, not all such objections apply to Bentham. For example, pleasure refers to all human values, not just to sensual or hedonistic pleasures or to wealth; the felicific calculus was not intended as an exact measure, but rather as providing a checklist of factors relevant to weighing the costs and benefits of any action; every individual equally counts for one, but the ultimate test is aggregate happiness (i.e., the general welfare). 'Extent' (whose happiness is to be taken into account?) applies not just to humans, but to all sentient beings, for in respect of animals "the question is not, Can they *reason*?, nor can they *talk*?, but can they *suffer*?" (*IPML*, 1996, Chapter XVII: pp. 282–286).

The first six chapters of *IPML* are not fully representative of Bentham's utilitarianism, as it developed over time. They are pedagogically convenient, because they are easy to read and to criticize. Here Benthamic ambiguity seems to begin at the very root of his thought: Does pleasure refer to satisfaction or to preference? Was he a direct or an indirect utilitarian? How different was Bentham's utilitarianism from Hume's? Does rooting utility in human nature make Bentham a closet natural lawyer? Do Bentham's utilitarian actors maximize their own pleasure or is every human choice, in individual morality and collective decision making, to be judged by its tendency to maximize the general welfare? Is the primary concern of sovereigns, their subordinates, and individuals rightly confined to one's own community, subject to certain exceptions?

Such questions are much debated. It is dangerous to impose on Bentham ideas that emerged in later debates about utilitarianism, for example, modern distinctions between act-utilitarianism and rule-utilitarianism, between preference and satisfaction, or between aggregate and average utility. Bentham's ideas developed over time and he was not always consistent. He might also have rejected sharp versions of these distinctions as false dichotomies.

Bentham's utility is a principle of both individual psychology and of social and political morality, of morals and legislation. For the jurist, the latter is both more important and less vulnerable to criticism. Humans generally seek to maximize their own interests, and the primary task of legislation is to harmonize individual interests with the general welfare by a system of rewards and punishments. Similarly, the interests of the governors need to be harmonized with the interests of the governed, especially through systematic design and implementation of securities against misrule. Bentham rejected any idea of principles independent of utility, but gave an important place to four principles subordinate to utility: security, subsistence, abundance, and equality. Of these security, or settled grounds of expectation, was the principal one because it implies extension in point of time and is a precondition for achieving the other ends of government. Bentham gave a high priority to both property and order, and hence for liberty in the sense of absence of coercion. Security is also the basis for the important motto of the good citizen, "To obey punctually; to censure freely" (*Fragment*, Preface). But security can be overridden by utility in given circumstances. Security and subsistence generally have a higher priority than equality, but in the calculus each individual counts for one, the idea of diminishing marginal utility serves as a distributive function, and Bentham believed that as a psychological fact "the nearer the actual proportion approaches to equality, the greater will be the total mass of happiness" (*Civil Code*; Harrison, 1983: pp. 243–250).

The originality and interest of Bentham's utilitarianism lie mostly in his *applications* of the basic ideas. Utility gives coherence and bite to his detailed treatment of evidence, classification of crimes, the rationalization of punishment, and numerous other specific topics. His utilitarianism should be

judged as much by its potential for illumination as an analytical and critical tool applied to concrete issues as by its philosophical defensibility.

Fictions

Bentham's theory of logic and language anticipated key later developments in English analytical philosophy, including the idea that sentences rather than words are the primary unit of meaning. To clarify basic abstract legal conceptions, such as 'obligation,' 'right,' and 'property,' he substituted for the traditional definition *per genus et differentiam* the techniques of 'paraphrasis' and 'phraseoplerosis.' Abstract nouns with no direct counterpart in the world of fact are names of fictitious entities. They can be elucidated by placing them in a whole sentence, such as "X has a right against Y," and then asking under what conditions this sentence is true. It is true if Y has a duty. "Y has a duty" is true if the sovereign of an existing legal system has commanded that there shall be a sanction (pain) for nonperformance in given circumstances, and Y's situation is subsumed under this command. Thus the terms 'right' and 'duty' are linked through 'sovereign' and 'sanction' to pains and pleasures, things that can be directly experienced.

Bentham's theory of fictions contains his basic ideas on ontology, epistemology, and logic. We construct our knowledge of the world through the lens of language, itself a human construct, which is both necessary because it is our window on the world and dangerous because it is a distorting lens. Language is the best tool we have for knowledge and understanding. This constructivist epistemology based on utility has affinities with later developments in pragmatism and the sociology of knowledge.

Bentham often used 'fiction' pejoratively. He attacked natural law, social contract theories, and technical legal fictions as dangerous mystifications. One example is his famous attack on natural rights as pestilential nonsense: to say that "X has a right" logically implies some existing general rule or command. "X has a legal right" presupposes a law commanded by a sovereign and backed by a sanction. Legal rights are creatures of positive laws. But since natural law does not exist nor do natural rights. The idea of natural rights is nonsensical, like the idea of a son without a father. To talk of imprescriptible (i.e., unalterable) rights is also meaningless, 'nonsense upon stilts,' for the sovereign's power cannot be legally limited. Since rights conflict, talk of 'absolute' rights is contradictory. Talk of nonlegal rights is also mischievous, raising expectations that cannot be fulfilled, inviting anarchy. Bentham's objections relate to the mode of discourse rather than to particular values. There may be moral claims and legal rights based on utility. He would not necessarily have opposed all modern legal bills of rights. His objection was to talk of nonlegal rights, not to the enactment into positive law of enforceable legal rights. A modern Benthamite can argue that talk of "a right to food" or "a right to work" refers to important aspirations, but begs questions about the allocation of correlative duties, about enforcement if not legally binding, and about the feasibility of implementing such 'rights.' They represent wishful thinking. As Bentham put it: "Want are not means: hunger is not bread." (Bentham, 2002: p. 330). Bentham started a line of analysis of 'rights,' 'powers,' and 'obligations' that was developed by Austin, Hohfeld (*see* Hart, Herbert Lionel Adolphus (1907–92)), Hart, and others. He still poses a challenge to loose rights talk.

Theory of Law

Bentham's legal theory operates on at least three levels: first, his general theory of law and its philosophical underpinnings; second, subtheories on, for example, codification and nomography, constitutional law, punishment and reward, and adjective law; third, the pannomion, a comprehensive body of codes of general application, subject to particular, essentially minor, modifications according to time, place and social context.

Bentham's jurisprudence starts with a strong positivist distinction between law as it is and law as it ought to be. "To the province of the *Expositor* it belongs to explain to us what, as he supposes, the Law is: to that of the *Censor*, to observe what he thinks it ought to be" (*Fragment*, 397). Exposition can be particular or general; censorial jurisprudence is general. Bentham's main interest was in universal censorial jurisprudence.

The fullest account of Bentham's general theory of law, *Of Laws in General* (1970 hereafter *OLG*), was first published in a definitive edition, edited by H.L.A. Hart, after Hart's *The Concept of Law* had become the leading text of English legal positivism, partly through its critique of Austin's imperative theory. Ironically, Hart conceded in a late essay that Bentham's version contained an important core of truth, "the notion of a content-independent, peremptory reason for action" as part of the general notion of authority (Hart, 1982, Chapter 10: p. 261). *OLG* involved a deeper exploration of difficulties that he had dimly recognized earlier, especially the distinction between the civil and the penal branches of jurisprudence and individuation (What constitutes a single law?). The outcome was a dense, but incisive work, which includes original ideas on the logic of the will, a precursor of modern deontic logic.

Bentham elaborated his conception of law in terms later made familiar by John Austin: commands backed by threats made by a sovereign who is subject to a habit of obedience. Yet Bentham's ideas were subtly different. Bentham allowed for limited and divisible sovereignty, whereas Austin did not. The habit of obedience might be partial. Commands, such as judicial orders, need not be general; some laws can be backed by moral or religious sanctions, some laws by rewards. The idea of command has a more complex and limited role in Bentham's theory.

Burns, Hart, and others have explored aspects of Bentham's jurisprudence in detail, including universal censorial jurisprudence; legal positivism and sovereignty; and his analysis of fundamental legal concepts, such as rights and obligations. Less abstract and less well known, but more extensive, are his theoretical writings on codification, adjective law (evidence and procedure), judicial organization, civil law, criminal law, indirect legislation, international law, and many other topics, whose importance is often underestimated. For example, Bentham's writings on evidence represent the most original and wide-ranging contribution to evidence theory in the Anglo-American tradition.

Codification and the Pannomion

From the early 1780s, Bentham's main aspiration was to produce a complete body of laws – a pannomion – to

rationalize and systematize English law and, indeed, any legal system. At first he thought that the main weaknesses of English law were lack of organization, incompleteness, and inaccessibility, linked to the prejudice that legislation was an inferior form of law to common law – defects that might be remedied by a comprehensive Digest. By 1782, he was convinced of the necessity for a systematic structure of codes based on utility and satisfying the tests of coherence, completeness, and comprehensibility. The pannomion encompassed integrated codes on civil, penal, procedural, and constitutional law, backed by detailed legislation on specific topics.

He started with penal law, but was diverted by theoretical concerns into 'the metaphysical maze' of *OLG*. By the late 1780s, he had completed important work on the general theory of law and legislation, but no actual code. Though codification remained his long-term aim, he only returned to the work of actually preparing codes much later. Starting in 1811 he offered his services as a codifier, first for several jurisdictions in the United States, then for Russia, Poland, and Geneva, and latterly for Spain, Portugal, Tripoli, and various Latin American countries. He now prepared some important general works on codification and started drafting several codes. His ideas excited considerable interest, but did not lead to the adoption of a single code. In 1830, he published the first volume of his magisterial *Constitutional Code* (1983). This, together with a procedure code, was only completed after his death by a former amanuensis, Richard Doane. Shortly before his death, Bentham was still 'codifying like any dragon.' He believed strongly that a code should be the product of a single mind, but his failure to complete a single draft reinforced the view that the pannomion was impracticable.

Constitutional Law and Democratic Theory

Bentham was not an instinctive democrat. When young, he was mainly concerned with how legislators should govern. He was prepared to deal with aristocrats, enlightened despots, revolutionaries, and democrats in promoting his ideas. His constitutional theory was a late development. The timing and reasons for his 'conversion to democracy' have been much debated. He was sympathetic to democratic ideas at the start of the French Revolution, but he only became involved in the movement for parliamentary reform after 1809, stimulated by his experience of general resistance and indifference to reform, especially the panopticon. Bentham firmly located sovereignty in the people. He recognized the problem of the tyranny of the majority, but he never satisfactorily solved it, perhaps because of his virulent antipathy to talk of nonlegal rights.

His elaborate constitutional design is based on three axioms: maximization of the happiness of all as the proper end of government; the maximization of the self-interest of the governors as the actual end of government in every political community; and the harmonization of the interests of the rulers and the ruled. His scheme involves vesting real and sustained power in the sovereign people to locate, monitor, and dislocate officials through a system of securities against misrule, especially publicity.

Several of the constitutional writings have only recently become available. The main elements in his code were 'expense minimized, aptitude maximized,' universal manhood suffrage (subject to a literacy test), secret ballots, frequent elections, a unicameral legislature, limited judicial power, accountability and transparency, and a power of recall of office holders, including the military, vested in the sovereign people. He supported ideas backed by other radical reformers of the time. In principle, he favored votes for women, but considered it inexpedient to press for this. Bentham conceived of a constitutional code as a logical and engineered product of a single mind, based on articulated principles, and designed to provide popular control over misrule, patronage, and other forms of corruption.

Bentham and Social Science

Bentham did not draw sharp distinctions between jurisprudence, psychology, political economy, and other fields that have since become semiautonomous social sciences. Utilitarianism requires a concern for social facts. Legislation is the social science. To make calculations the legislator needs reliable information about the existing situation, the likely consequences of a given course of action, and causal relations. He emphasized 'the statistic function' of government; his stepbrother, Charles Abbott, played a leading role in the first population census (Population Bill, 1800); and Bentham himself tirelessly argued for better social data, especially regarding the Poor Laws. This Benthamite enthusiasm was satirized in Charles Dickens's *Hard Times* in Mr Gradgrind, who believed that 'Facts alone are wanted in life.' Bentham believed no such thing. Utility was his guiding principle and he had a sophisticated awareness of the complexities of concepts such as 'fact' and 'causation.' To his credit, he was among those who first recognized and campaigned for the systematic collection of reliable social data as a necessary, but not a sufficient, condition for policy making and legislation.

On 'the art and science of political economy,' Bentham was a follower of Adam Smith, but criticized his defense of usury laws, and on some issues differed from him in either being more averse or more amenable to interventionism. Bentham saw the main role of government as promoting abundance indirectly through ensuring security, the basic presumption being against government involvement in economic affairs 'without some special reason.' But his proposals included provision for many areas later associated with the welfare state, such as public health, education, and the relief of indigence. He argued for the emancipation of colonies both on economic and constitutional grounds and, rather than being a doctrinaire supporter of *laissez faire*, consistently applied cost-benefit analysis to particular issues. Bentham had some impact on the economic writings of James Mill, John Stuart Mill (*see* Mill, John Stuart (1806–73)), David Ricardo, and later W.S. Jevons, Alfred Marshall, and F.Y. Edgeworth (*see* Edgeworth, Francis Ysidro (1845–1926)). There is a useful, but incomplete, edition of Bentham's economic writings (1952–1954).

One cannot enter into detail here about Bentham's psychology, which is regarded by some as his Achilles heel, nor to his extensive and detailed writings on education, poor law, social administration, and other aspects of social policy. There are some interesting affinities with Marx's (*see* Marx, Karl (1818–83)) ideas, for example, the clash between the interests of the few

and the many (though with less emphasis on social class), interest-begotten prejudice, and demystification. But Marx, like Foucault, criticized Bentham without having studied him carefully. Marx treated him as the epitome of 'commonplace bourgeois intelligence.' Foucault interpreted the Panopticon as a symbol of repressive authoritarian control, ignoring the fact that Bentham's concern with transparency applied to those with power as well as those subject to it. Both images are, at best, half-truths. Bentham was also a pioneer in the study of bureaucracy. His standing as a social theorist is ripe for reappraisal.

Influence

Bentham first gained recognition through the efforts of a Genevan, Etienne Dumont (1759–1829), who between 1802 and 1828 produced five elegant French 'recensions' from articles and printed texts. These were more lucid, succinct, and diplomatic than the originals. Some were translated into several languages, including English. The first, *Traité de legislation civile et penale* (1802), made Bentham's reputation in Europe, whence it spread to Latin America and eventually back to Britain.

In the late nineteenth century, Bentham was hailed as the most influential reformer in English history. Subsequently, there has been a continuing debate about the relationship between his ideas and particular reforms. Bentham had both committed disciples and more pragmatic followers who became influential, including Romilly, Brougham, Chadwick, James Mill, and John Stuart Mill; his ideas contributed to the climate of reform in the nineteenth century both in Britain and internationally; many changes, for example, in the Anglo-American law of evidence, moved in directions he had advocated, but in a more piecemeal fashion than he would have approved. He hated jargon, but coined neologisms. Some – codify, maximize, international – have become part of the language. Others, fortunately, have not. Many of his proposals, for instance on codification, have not yet been implemented. How far responsibility for specific changes should be attributed to him or his followers remains controversial.

Bentham Studies

The Bentham Project was begun in 1959 with the objective of producing a definitive edition of Bentham's *Collected Works* (*CW*) in 68–70 volumes. Progress was slow in the early years mainly because of shortage of funds, but also because of the extent of Bentham's articles (more than 70 000 pages) and the extraordinary editorial difficulties. By 2000, 20 substantial volumes had been published and considerable editorial work had been done on others. Between 2000 and 2014 ten further substantial volumes were published with work progressing on the preparation of several more. Since September 2010, the process of initial transcription of hitherto unedited articles has been greatly accelerated by the Bentham Papers Transcription Initiative (generally known as 'Transcribe Bentham'). Using modern technology and employing volunteers, who are recruited internationally and supervised by members of the Bentham Project, almost 10 000 manuscript pages had been transcribed in digital form by mid-2014 and were in course of preparation for an open-source Consolidated Bentham Papers Repository. This will make the bulk of the hitherto unpublished documents available on line as reliable texts before final editing for publication. Transcribe Bentham won an Award of Distinction in the Digital Communities category of the Prix Ars Electronica 2011. Previous holders include *Wikipedia* and *Wikileaks*. This combination of popular participation, new technology, open access, and prestigious recognition would surely have pleased Mr Bentham.

In addition to its main task of producing *The CW*, the Bentham Project has stimulated a range of ancillary activities: the journal *Utilitas* (formerly *The Bentham Newsletter*), the International Society for Utilitarian Studies, conferences, courses, a large and proliferating secondary literature, a growing international community of Bentham scholars, especially in France, Italy, China, and the United States, and some more frivolous/frolicsome activities all of which can be accessed by the Project's very extensive web site. The main outlines of Bentham's life and ideas are well known, and there have been some important specialist studies. In recent years, much more has become available about his detailed views on many topics (including sex, religion, freedom of speech, the Poor Laws, logic and language, fallacies, indirect legislation, and critical jurisprudence). Moreover, recent volumes in *The CW* and the hitherto unpublished articles contain second thoughts, refinements, and new departures that are likely to confirm the view that Bentham was less doctrinaire, and more subtle, original, and genuinely radical than has been traditionally supposed. Revisionism in Bentham studies is now well under way.

Questions arise about his contemporary 'significance.' It is anachronistic to apply early nineteenth-century ideas directly to twenty-first-century issues. Yet, Bentham's views are still relevant to topics as varied as human rights, sovereignty, constitutional reform, freedom of the press, sex, penal policy, torture, corruption, patronage, transparency, accountability, freedom of information, political rhetoric, the legal profession, access to justice, inheritance, free markets, globalization, and 'the farther uses of the dead to the living.' On these and many other issues he is still worth consulting, not least because he often takes provocatively radical positions. As Herbert Hart (*see* Hart, Herbert Lionel Adolphus (1907–92)) remarked, "where Bentham fails to persuade, he still forces us to think" (Hart, 1982: p. 39).

See also: Constitutions; Economics, History of; Enlightenment; Justice and Law; Law and Democracy; Law: History of Its Relation to the Social Sciences; Natural Law; Positivism, History of; Prisons and Imprisonment; Ricardo, David (1772–1823); Smith, Adam (1723–90); Utilitarian Social Thought, History of.

Bibliography

Works by Bentham

Bentham, J., 1838–1843. In: Bowring, J. (Ed.), The Works of Jeremy Bentham, 11 vols. W. Tait, Edinburgh.

Bentham, J., 1952–1954. In: Stark, W. (Ed.), Economic Writings: Critical Edition based on his printed works and unpublished manuscripts. Allen & Unwin, London.

Bentham, J., 1968–1981. Correspondence, vols. 1–5. Athlone Press, London.
Bentham, J., 1970. In: Hart, H.L.A. (Ed.), Of Laws in General. Oxford University Press, Oxford, UK.
Bentham, J., 1977. In: Burns, J.H., Hart, H.L.A. (Eds.), A Comment on the Commentaries and a Fragment on Government. Oxford University Press, Oxford, UK.
Bentham, J., 1983. In: Burns, J.H., Rosen, F. (Eds.), Constitutional Code, vol. I. Oxford University Press, Oxford, UK.
Bentham, J., 1984–2006. Correspondence, vols. 6–12. Oxford University Press, Oxford, UK.
Bentham, J., 1989. In: Schofield, P. (Ed.), First Principles Preparatory to the Constitutional Code. Oxford University Press, Oxford, UK.
Bentham, J., 1995. In: Schofield, P. (Ed.), Colonies, Commerce and Constitutional Law: Rid Yourselves of Ultramaria and Other Writings on Spain and Spanish America. Oxford University Press, Oxford, UK.
Bentham, J., 1996. In: Burns, J.H., Hart, H.L.A. (Eds.), An Introduction to the Principles of Morals and Legislation. Oxford University Press, Oxford, UK.
Bentham, J., 1997. In: Schofield, P., Clero, J.-P., Laval, C. (Eds.), De l'ontologie et autres textes sur les fictions Paris: Editions du Seuil.
Bentham, J., 1998. In: Schofield, P., Harris, J. (Eds.), Legislator of the World: Writings on Codification, Law, and Education. Oxford University Press, Oxford, UK.
Bentham, J., 1999. In: James, M., Blamires, C., Pease-Watkin, C. (Eds.), Political Tactics. Oxford University Press, Oxford, UK.
Bentham, J., 2001. In: Quinn, M. (Ed.), Writings on the Poor Laws, vol. 1. Oxford University Press, Oxford, UK.
Bentham, J., 2002. In: Schofield, P., Pease-Watkin, C., Blamires, C. (Eds.), Rights, Representation, and Reform – Nonsense upon Stilts and Other Writings on the French Revolution. Oxford University Press, Oxford, UK.
Bentham, J., 2010a. In: Schofield, P. (Ed.), Of the Limits of the Penal Branch of Jurisprudence. Oxford University Press, Oxford, UK.
Bentham, J., 2010b. In: Quinn, M. (Ed.), Writings on the Poor Laws, vol. 2. Oxford University Press, Oxford, UK.
Bentham, J., 2011a. In: Crimmins, J., Fuller, C. (Eds.), Church-of-Englandism and its Catechism Examined. Oxford University Press, Oxford, UK.
Bentham, J., 2011b. In: Engelmann, S.G. (Ed.), Selected Writings. Yale University Press, New Haven, CT.
Bentham, J., 2012. In: Pease-Watkin, C., Schofield, P. (Eds.), On the Liberty of the Press, and Public Discussion and Other Legal and Political Writings for Spain and Portugal. Oxford University Press, Oxford, UK.
Bentham, J., 2014. In: Schofield, P., Pease-Watkin, C., Quinn, M. (Eds.), Of Sexual Irregularities, and Other Writings on Sexual Morality. Oxford University Press, Oxford.

Secondary Writings

Dinwiddy, J., 2003. In: Twining, W. (Ed.), Bentham: Selected Writings of John Dinwiddy. Stanford University Press, Stanford, CA.
Harrison, R., 1983. Bentham. Routledge and Kegan Paul, London.
Hart, H.L.A., 1982. Essays on Bentham. Oxford University Press, Oxford, UK.
Mack, M., 1968. In: Bentham, Jeremy (Ed.), International Encyclopedia of Social Science, vol. 2. Macmillan, New York, pp. 55–58.
Postema, G.J., 1989. Bentham and the Common Law Tradition. Oxford University Press, Oxford, UK.
Schofield, Philip, 2006. Utility and Democracy: The Political Thought of Jeremy Bentham. Oxford University Press, Oxford, UK.
Schofield, Philip, 2009. Bentham: A Guide for the Perplexed. Continuum, London.

Anthologies of Secondary Writings

Parekh, Bhikhu (Ed.), 1993. Jeremy Bentham: Critical Assessments, 4 vols. Routledge, London.
Postema, Gerald J. (Ed.), 2002. Bentham: Moral, Political and Legal Philosophy, 2 vols. Ashgate, Aldershot, UK.

Relevant Website

http://www.ucl.ac.uk/Bentham-Project – Bentham Project.

Bereavement

Margaret Stroebe, Utrecht University, Utrecht, The Netherlands
Henk Schut, Utrecht University, Utrecht, The Netherlands
Kathrin Boerner, Jewish Home Lifecare, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article outlines and reflects on the current state of scientific knowledge of bereavement. Bereavement reactions and associated health consequences are described, and the factors influencing vulnerability and resilience are summarized. We explore ways of coping and links to (mal)adjustment. We outline the current understanding of complicated grief and the effectiveness of psychological interventions. Theoretical developments, particularly those relating to grief complications, are considered. The bereavement research area is rapidly growing and evolving. Many firmly held views have been put to test and sometimes discarded. Therefore, throughout the article, we try to pinpoint key, sometimes controversial, issues and identify directions for future research.

Bereavement, Grief, and Mourning

The Latin derivations of the three closely related basic terms bereavement, grief, and mourning help pinpoint the overlap and differences between them, in addition to suggesting continuity in usage across the centuries. The term 'bereavement' comes from the Latin word *rumpere* (to break, to carry, or tear away). It refers to the objective situation of a person who has recently suffered the loss of a significant person through death. Some types of bereavement such as the loss of a close family member (child, spouse, or parent) have received more scientific attention than others (siblings, grandparents/children). There is room for expansion of research on other meaningful relationships too, including those that sometimes involve secrecy and private suffering (e.g., abortions, death of an ex-marital partner).

'Grief' is derived from the Latin word *gravare* (to weigh down). It refers to the emotional experience of psychological, cognitive/behavioral, social, and physiological reactions to one's loss; it is considered to be a complex emotional syndrome (see Table 1), and it is often regarded as the price we pay for love (thus linking the phenomenon to bonding and attachment to others; see Parkes, 2006). However, it is important to note that not all of these reactions are evident in any particular case and that considerable variation may occur across time and place.

'Mourning' is derived from the Latin word *memor* (mindful). It denotes actions and manners expressive of grief that are shaped by social and cultural practices and expectations that can differ greatly (e.g., the funeral as celebration of life vs as expression of pain; wearing of white vs black to funerals; spreading vs keeping of ashes; an open vs closed casket; removing vs cultivating reminders in one's living context). More research is needed to examine precisely how mourning customs (such as the rituals illustrated above) impact on grief (and vice versa).

The Impact of Bereavement

Although bereavement is associated with detrimental health consequences, it is important to emphasize that it is a normal, natural human experience, one that most people manage to come to terms with over the course of time and without the aid of professional intervention for their grief reactions. Indeed, it is a life event that, sooner or later, affects nearly everyone. Although comparatively rare in childhood, among children

Table 1 Reactions to bereavement

Affective	Depression, despair, dejection, distress
	Anxiety, fears, dreads
	Guilt, self-blame, self-accusation
	Anger, hostility, irritability
	Loneliness
	Yearning, longing, pining
	Shock, numbness
Cognitive	Preoccupation with thoughts of deceased, intrusive ruminations
	Sense of presence of deceased
	Suppression, denial
	Lowered self-esteem
	Self-reproach
	Helplessness, hopelessness
	Suicidal ideation
	Sense of unreality
	Memory, concentration problems
	Meaning making efforts
Behavioral	Agitation, tenseness, restlessness
	Fatigue, exhaustion
	Overactivity
	Searching
	Weeping, sobbing, crying
	Social withdrawal
Physiological–somatic	Loss of appetite
	Sleep disturbances
	Energy loss, exhaustion
	Somatic complaints
	Physical complaints similar to deceased
Immunologic and endocrine changes	Susceptibility to illness, disease, mortality

Adapted from Stroebe, M., Schut, H., Stroebe, W., 2007. Health consequences of bereavement: a review. The Lancet 370, 1960–1973.

below 18 years, approximately 3–4% can be expected to experience the death of a parent (in Western societies). By contrast, among elderly populations, approximately 45% of women and 15% of men over 65 years can expect to become widowed.

Bereavement is associated with a period of intense suffering for many people and with an increased risk of suffering mental and physical health problems. Adjustment is not confined to the calendar year that is commonly talked of as the expected duration of intense grief; sometimes it can take years. For the most part, researchers have abandoned the use of the word 'recovery' to describe the course of adaptation, acknowledging the experience of many bereaved persons that "one does not get over it, in time one gets used to it." There are substantial variations between individuals and across cultures not only in duration, but also in intensity of reactions. Some people show great resilience in response to loss. For others, typically a minority of bereaved persons, health consequences are extreme and lasting.

Examination of Phenomena and Manifestations Associated with Bereavement

Many questions have arisen and been given scientific attention regarding the links between bereavement and ill-health, for example: Is grief a disease? When might grief make you ill? Some persons develop nongrief-related mental and/or physical health difficulties, and some succumb to complications in the grieving process itself: Who suffers from what and why? Why do some have extreme and lasting effects, while others do not? We begin by addressing the most extreme health consequence, frequently posed as the question: Is there such a phenomenon as the 'broken heart'? Is it myth or reality?

The Mortality of Bereavement

There is considerable evidence that bereavement is associated with an increased risk of mortality from many causes, including suicide. Well-controlled studies now support the conclusion that the heightened risk of dying is attributable in large part to the so-called 'broken heart' phenomenon, if we understand this to imply psychological distress due to the loss, such as loneliness, and secondary consequences of it, such as changes in social ties, living arrangements, eating habits, and economic support. While most research has focused on heightened risk among widows and widowers, recently, it has extended to systematic investigation of mortality following the death of a child. For example, Espinosa and Evans (2013) reported increased risk of mothers dying after their child's death, with indications that the excess risk was concentrated in the first 2 years of bereavement. However, while the myth of the broken heart seems to be confirmed, it is important to emphasize that the investigations also show that in terms of absolute numbers, few people can be expected to succumb to this drastic outcome of losing a loved person; it is a rare occurrence.

The Range of Physical and Psychological Health Outcomes

Given the variety of reactions covered in **Table 1**, it is perhaps not surprising that bereavement is associated with a wide range of physical and psychological ailments and illnesses. Physical health problems are more frequently found among the bereaved than among matched nonbereaved control persons. The former also have higher rates of disability, medication use, and hospitalization than their nonbereaved counterparts. Widowed people consult with their physicians more often, most likely due to symptoms of anxiety and tension, but results of research also suggest that many with intense grief fail to seek medical help when they need to.

Studies have also confirmed elevated rates of psychological symptoms and ill-health, for example, of suicidal ideation, loneliness, insomnia, anxiety, depression, somatic symptoms, and social dysfunction. For a few people, depression or anxiety can reach levels that are of clinical importance. To indicate prevalence: According to some estimates, 25–45% of bereaved people have mild levels of depressive symptoms, while 10–20% reach clinical levels (Stroebe et al., 2007). In some cases, particularly following horrific death circumstances, the bereaved may develop posttraumatic stress disorder.

Recent research has attested to the need for differentiation regarding the range of health consequences. For example, in the past, a generic depression measure was frequently used instead of a specific grief one as an index of the impact of bereavement on mental health. Now it is recognized that grief and depression may, for example, represent distinct, though related, clusters of reactions to bereavement. They may be associated with different risk factors and difficulties and need separate investigation. Nevertheless, similarities have been emphasized. In an influential review, Zisook and Kendler (2007) concluded "The prevailing evidence more strongly supports similarities than differences between Bereavement Related Depression and Major Depression" (p. 779). Another issue concerns the scope of reactions from the typically negative ones listed, to more positive aspects. To illustrate, it is still a matter for consideration where such positive phenomena as personal growth fit in, since research on traumatic losses (particularly) has expanded the understanding of people's remarkable, not uncommon, resilience to enormous losses.

In some cases, the grieving process itself can become disturbed or complicated. Complicated grief (CG) has been variously defined, but most theorists understand it as a form of grief characterized by persistent, intense longing and yearning for the deceased (separation distress), intrusive thoughts or images, emotional numbness, anger or guilt related to the loss, a sense of emptiness, and reactivity in response to cues. At the same time, persons with CG often avoid people and places they associate with the loss, because of the intense distress those reminders evoke. Thus, CG tends to involve the oscillation between anxious preoccupation with and an avoidance of memories of the deceased. In addition, complicated grievers commonly have difficulty in redefining themselves (Mancini and Bonanno, 2006), and often experience difficulties in forming satisfying new relationships or engaging in potentially rewarding activities. The scenario of an initially absent grief reaction followed by a delayed response also seems possible, although empirical evidence on this phenomenon is scarce.

Estimates of the occurrence of CG vary widely, with some reporting 9% of a bereaved adult population and others 20%. In a study of parents who had lost a child through suicide or accident 18 months previously, Dyregrov et al. (2003) reported that as many as 78% scored above a suggested cut-off level on criteria for CG reactions – raising the question whether this

cut-off point is a good marker of CG or whether the norm scores for parents need to be reset (intense grief may not necessarily reflect 'complications').

In recent years, conceptualization and understanding of CG have been strongly channeled by the new Diagnostic and Statistical Manual of Psychiatric Disorders (DSM-5) of the American Psychiatric Association. We return to this below.

Risk and Protective Factors

The term 'risk factor' signifies the situational, intrapersonal, and interpersonal characteristics associated with increased vulnerability associated with the range of bereavement outcomes. Protective factors are those that promote resilience and lower the risk of adverse health consequences.

The situation and circumstances of death contribute to adjustment in complex ways. For example, the broader circumstances of death, including the cause and caregiver strain, must be taken into account. Also, combinations of situational factors (with other types, such as personality factors) can jointly account for the impact on adjustment. While some conclude that sudden unexpected death leads to worse bereavement consequences, others have found that caring for a terminally ill spouse can increase difficulties subsequent to death. Characteristics of the bereaved person or of the predeath relationship with the deceased also play a part. For example, research on attachment has strongly suggested that the quality and nature of the lost relationship is critical to outcome. Childhood loss of a parent has various long-term effects, but importantly, evidence suggests that the adequacy of the remaining parental care after the death of one parent and the resilience of the child him/herself are more powerful predictors of later adjustment than the loss of the parent per se.

Regarding interpersonal factors, a lot has been written over the years about the importance of social support; yet this has now been identified as a general risk factor, i.e., one that affects the health and well-being of nonbereaved to the same extent as bereaved persons. This suggests that others cannot easily take the place of the deceased, or make up for the loneliness following the death. More recently, finer-grained interpersonal processes have been researched and uniquely, intergenerational relationships were subjected to closer inspection. Carr and Boerner (2013) examined the extent to which adaptation to late-life widowhood was not only conditional upon qualities of the late marriage (positive and negative interactions, dependence), but also with parent–child relations in terms of support, criticism, and dependence. Importantly, it was found that widowed persons report significantly less criticism from and more dependence upon children than married controls. Furthermore, effects were conditional upon characteristics of the late marriage: among the bereaved, marital warmth is associated with high levels of support from children post-loss, whereas marital strains were associated with less emotional support from and dependence on children. Widow(er)s who were highly dependent on their late spouse reported lower levels of criticism from children. The importance of including interpersonal variables in risk factor research becomes evident. This extends to interpersonal processes in coping, which have been largely neglected.

Ways of going about dealing with bereavement, referred to as coping strategies or processes (including the interpersonal aspects just mentioned), can impede or facilitate adjustment, and as such can be considered risk factors. Indeed, and unlike many other risk factors, coping is a particularly important factor, because it can potentially be changed (e.g., amenable to interventions). The concept covers many dimensions. Doing one's 'grief work,' for example, used to be considered focal in coming to terms with loss, but more recent research has shown that people who do not work through their grief frequently recover as well as those who do. It was also widely believed that adaptive grieving involved 'letting go' of the deceased. Nevertheless, research over the last decade has shown a potentially healthy role of 'continuing bonds,' although the conditions under which continuing bonds are helpful or not still need to be established. Continuing bonds can reflect a lack of acceptance of the death of the loved one and an effort to maintain a concrete tie, rather than a more internalized, symbolic connection reflecting relinquishment and acceptance of the physical absence of the deceased. There are also considerable cultural differences in their meaning and influence.

So what additional directions should risk factor research take? Many potential risk factors have been underresearched, which is critical, particularly because it creates difficulties for scientists to develop valid screening instruments for use by practitioners. In this respect, studies of multiple risk factors and the interaction between such factors is advisable, rather than focusing (as in earlier studies) on single factors. It is important to establish the relative and the joint impact of risk factors. Finer-grained examinations of maladaptive (interpersonal) coping processes is another important approach that is already gaining momentum, as is the study of positive and negative cognitions and the regulation of emotion in grieving. Theory-guided research along these lines is essential. Finally, further investigation of the ways that risk factors relate to the different health outcomes described above is in order: Why does one bereaved person suffer complications in the grieving process, while another falls physically ill?

CG: Intervention and Diagnosis

We have already indicated that, for a minority of bereaved persons, grieving can take a difficult course, perhaps best summarized as a 'derailing' of the normal, if painful, process of adapting to the loss of a significant person. Given that there are complications, research also needs to address questions about the advisability and effectiveness of intervention. Linked to this are questions about the mental health status of bereaved persons who encounter difficulties in coping with their loss: should these bereaved persons be diagnosed with a mental illness?

Intervention Efficacy

Should intervention be recommended and if so, who benefits from it? More specific research questions include establishing the relative effectiveness of different types of intervention and exploring which type is effective for a particular mental and/or physical health difficulty. In fact, most research has investigated psychological, not medical or pharmacological, interventions, so we focus on the former type of treatment. It is also important

to note that adequate scientific investigation involves the use of randomized control trials (RCTs) for prevention or treatment. To establish efficacy, it is not enough to demonstrate that bereaved persons are satisfied with an intervention program: it needs to be shown that bereaved persons who are assigned to an intervention condition adjust to their loss relatively better than those not receiving treatment.

An approach that has frequently been adopted by bereavement researchers to try to establish who in general benefits from intervention has been to divide the programs into primary, secondary, and tertiary interventions. Primary interventions are those in which (professional) help is available to all bereaved persons, irrespective of whether intervention is indicated. Secondary interventions are designed for those who, through screening or assessment, can be regarded as more vulnerable to the adverse consequences of bereavement. Tertiary interventions refer to programs that provide therapy for CG, grief-related depression, or posttraumatic disorders, which are usually evident longer after bereavement (since pathological processes usually take longer to develop). Conclusions reached have been reasonably unanimous: there is little evidence that all bereaved people will benefit from counseling or therapy. Primary prevention can, however, be helpful when the initiative to get help is left to the bereaved person (indicating need). Secondary preventive interventions for risk groups seem to be an important provision, but as noted above, improvement in risk factor assessment is essential both for better and lasting results of interventions and for the adequate assessment of efficacy.

With regard to tertiary interventions, a wide variety of treatments for grief complications, now including remote modality ones, are available. Systematic comparisons of the relative effectiveness of different therapeutic approaches are still needed to understand what is most effective for whom, but there is evidence that tertiary interventions in general are effective. Wittouck et al. (2013) have reviewed studies investigating the short- and long-term effects of both preventive and treatment interventions specifically for CG. Results of their systematic analyses of the pattern of results of the available RCTs indicate that treatment interventions can effectively reduce CG symptoms. This indication of the efficacy of professional treatment for CG brings us to the related question about the diagnostic status of CG.

Diagnostic Status

Are (severe) complications associated with grief to be considered a mental disorder included in manuals of mental disorders? There has been considerable discussion, and opinions vary greatly regarding the definition and diagnosis of grief complications (cf Stroebe et al., 2013). In the new edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; APA, 2013), CG, termed persistent complex bereavement disorder (PCBD), is not listed among established disorders but among conditions requiring further research investigation. (While the focus was recently on the DSM-5 (particularly in the U.S.A.), the ICD-11 is currently being drafted and is due for publication in 2015. The ICD system is a core instrument of the WHO, which is in its constitution and is ratified by all the WHO member states.) A further change in the DSM-5 since the previous edition has been the removal of the exclusion of bereavement in the first 2 months from the criteria for the diagnosis of major depressive disorder (MDD). Although the article dictates caution in making a diagnosis of MDD for a bereaved person, concerns remain that in practice, the removal of this exclusion criterion may lead to overdiagnosis of recently bereaved persons as having MDD; early on, they may very well reach the levels of symptomatology defined in the diagnostic criteria for MDD.

An additional concern, given the DSM-5 developments outlined above, is the potential neglect of subtypes of CG, described in classic sources (particularly the clinical literature). Chronic grief, which is similar to PCBD (and as such covered in the DSM-5 to some extent), is characterized by the long-lasting presence of symptoms associated with intense grief. Delayed or inhibited grief have been taken to occur in cases where there is little or no sign of grieving early on in bereavement, but rather, at a later time (clearly, pathology is not always indicated by early lack of symptoms). Absent grief has been understood to be problematic if the person manifests other symptoms, atypical of grief. However, the existence and impact of delayed, inhibited, and absent grief subtypes is currently debated; more empirical research is needed.

Theoretical Understanding

Theoretical approaches to bereavement provide explanations for phenomena and manifestations of grief and grieving. For example, some provide insights into mal/adaptive coping mechanisms, which are potentially useful for guiding research, and ultimately, for applied use in helping the bereaved. Theories vary widely with regard to basic principles, levels of analysis, and degree of specificity. They also range from application to bereavement of general psychological theories such as psychoanalytic and attachment theories or theories of loss and trauma, to bereavement-specific models of coping.

Much theorizing in the bereavement field has been influenced by Freud's (1917) paper 'Mourning and Melancholia.' According to this psychoanalytic approach, when a loved one dies, the bereaved person is faced with the struggle to sever ties and detach the energy invested in the deceased person. The psychological function of grief is to free the person from the ties to the deceased, achieving a gradual detachment by means of the process of grief work. Grief work remained a central concept not only in subsequent theories but also in principles of counseling and therapy. It denotes a cognitive process of going over events that occurred before and at the time of death and focusing on memories and working toward detachment from the deceased. According to Freud, a major cause of pathological grief was the existence of ambivalence in the relationship with the deceased preventing the normal transfer of libido (energy) from that person to a new object.

The notion of grief work has been one of the most persistent features of psychoanalytic theory to influence subsequent research and thinking about coping with bereavement. While its theoretical interpretation is different, it has been incorporated in the major contribution of Bowlby, in his attachment theory (e.g., Bowlby, 1980). Bowlby focused on the biological rather than the psychological origins of grief. According to his

approach, the biological function is to regain proximity to the attachment figure, separation from which has caused anxiety. In the case of permanent loss, regaining proximity is not possible; the response is dysfunctional because reunion cannot be achieved. But an active working through of the loss still needs to be done, as it is an essential part of grief and grieving. The focus on working through grief led Bowlby to formulate phases or stages of grieving (shock, numbness and denial, yearning and protest, despair, and gradual recovery). However, there has been a move away from a phased approach, particularly in view of the evidence that grief and grieving do not follow a predictable, sequential order from initial high distress to return to normal (although it is important to remember that the phases were initially conceptualized as descriptive, i.e., indicating regularities, rather than being prescriptive). One major shift was brought about by Worden's identification of tasks rather than phases of grief (e.g., Worden, 2009), in terms of acceptance of the reality of loss, processing the pain, adjusting to a world without the deceased, and retaining connection with the deceased while embarking on a new life. Worden's tasks have provided clinicians with a 'tool' to help clients work through their grief. A further major development has been Bonanno and colleagues' (e.g., Bonanno et al., 2008) identification of trajectories of grief. Their research showed that there are qualitatively distinct pathways across the months following the loss of a loved one, supporting the conclusion that the grieving process cannot be well described in terms of a single set of stages or tasks. Bonanno's research suggested that resilience is actually common and that delayed grief is rare.

In other respects, attachment theory has also remained enormously influential in contemporary bereavement research. Early childhood experiences with attachment figures are still considered critical. These experiences are understood in terms of the development of either secure or insecure bonding with the caregiver, and this emerging style of attachment has a lasting influence on later relationships. To illustrate, frequent separation from attachment figures in childhood can lead to anxious attachment in later relationships, which is associated with chronic grief. Mikulincer and Shaver (e.g., 2013) have conducted sophisticated empirical research, confirming the importance of attachment security in the prediction of adjustment to bereavement and providing a fine-grained understanding of many associated phenomena.

Diverse trauma and stress theories have influenced the understanding of the phenomena and manifestations of bereavement. One major line of work is represented by the work of Horowitz and colleagues (e.g., Horowitz, 1986), applicable to traumatic events in general. A basic assumption of this approach is that stressful life events (SLEs) play an important role in the etiology of various somatic and mental disorders. A further line of research derived from the related field of trauma was that of Janoff-Bulman (1992), particularly through identification of shattered beliefs, which need to be rebuilt. This has been expanded to the study of 'meaning making' particularly by Niemeyer and collaborators (e.g., Neimeyer, 2001), giving centrality to the need for 'making sense' and 'finding meaning' after the loss of a loved person. A basic idea is that the reconstruction of meaning about the self and the world is critical to adjustment. Difficulties in establishing the role of meaning making in adjustment remain (e.g., studies have not always succeeded in separating the process from the outcome, beliefs from adjustment, or establishing the direction of causality among these factors). Others have distinguished two components of meaning-making. Davis et al. (1998) identified two distinct processes, making sense of the loss and finding benefit, which entail distinguishable psychological concerns for the bereaved person, with, for example, the former diminishing in importance in time, while the latter growing stronger as time goes on.

Stress theorizing received independent impetus through the work of Lazarus and Folkman (1984), who provided detailed analyses of coping processes, through which SLEs may precipitate poor physical or mental health outcomes. A major premise of their cognitive stress theory is that the intensity of stress created by an SLE depends on the extent to which the perceived demands of the situation tax or exceed an individual's coping resources, given that failure to cope leads to negative outcomes. Recently, research has identified neurophysiological mechanisms linking stress with various detrimental consequences to the immune, gastrointestinal, and cardiovascular systems (O'Connor, 2013). Folkman (2001) applied cognitive stress theory specifically to the stressor of bereavement, significantly extending the perspective to include positive emotions/appraisals as a component in coping with bereavement. Drawing partly on this theoretical approach, Stroebe and Schut (1999) developed their dual process model (DPM), a bereavement-specific coping model, to try to capture the complexity, diversity, or idiosyncratic nature of grieving. They postulated two types of stressor, the first called loss orientation, which relates to the deceased person (e.g., ruminating about the death), and the second restoration orientation, which has to do with secondary stressors that come about indirectly but as a consequence of the death (e.g., dealing with new financial concerns). Rather than assuming phases or tasks, the DPM includes an emotion regulation process, labeled oscillation. Adaptive coping entails oscillation between the two types of stressor and 'time out,' given the necessity to rest from the arduous process of dealing with the loss.

Conclusions

Loss of a loved one can cause deep suffering and the risk of mental and physical health detriments, but it is important to remember that grief is a normal, natural response to bereavement. Most reactions are not complicated and for most persons, family and friends, religious and community groups, and various societal resources will provide the necessary support. Professional psychological intervention is generally neither justified nor effective for uncomplicated forms of grief.

Although advancements across all the areas of research reviewed above have been evident in recent years, improvement in the design of research projects is often indicated, for example, in the design of intervention programs and for strict assessment of their relative effectiveness, following evidence-based criteria. Furthermore, gaps in knowledge still remain. So what directions should future research take? Both micro- and macrolevel extensions can usefully be pursued. At microlevel, one line of research that is currently expanding is fine-grained cognitive/emotional (regulatory) process analyses,

including physiological measurement. Laboratory techniques are beginning to answer questions about distinctions between adaptive grief work and maladaptive rumination and to link these to confrontation versus avoidance processes.

At macrolevel, more research is needed that not only focuses on bereaved individuals, but also on families, organizations, and cultural groups, to investigate broader changes that may be triggered by a loss. Inclusion of underacknowledged, disenfranchised groups within this wider scope seems essential and many examples come to mind, such as those with learning difficulties, expartners, nursing home and home care staff's grief for their patients. Finally, much is still to be learned about – and to be gained scientifically – by extending consideration to losses without physical death, such as organ donation, euthanasia, dementia, or irreversible coma.

See also: Co-morbidity of Mental and Physical Conditions; Health and Illness: Mental Representations in Different Cultures; Mental Health; Religion and Health: Clinical Considerations and Applications; Social Support and Recovery from Disease and Medical Procedures.

Bibliography

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Association, Washington, DC.

Bonanno, G., Boerner, K., Wortman, C., 2008. Trajectories of grieving. In: Stroebe, M.S., Hansson, R.O., Schut, H., Stroebe, W. (Eds.), Handbook of Bereavement Research and Practice: Advances in Theory and Intervention. American Psychological Association, Washington, DC, pp. 287–307.

Bowlby, J., 1980. Attachment and loss. In: Loss: Sadness and Depression, vol. 3. Hogarth, London.

Carr, D., Boerner, K., 2013. The impact of spousal loss on parent-child relations in later life: are effects contingent upon the quality of the late marriage? In: Stroebe, M., Schut, H., Boerner, K. (Eds.), Bereavement: A Family Perspective. Family Science, 4, 37–49 (Special Issue).

Davis, C., Nolen-Hoeksema, S., Larson, J., 1998. Making sense of loss and benefitting from the experience: two construals of meaning. Journal of Personality and Social Psychology 75, 561–574.

Dyregrov, K., Nordanger, D., Dyregrov, A., 2003. Predictors of psychosocial distress after suicide, SIDS and accidents. Death Studies 27, 143–165.

Espinosa, J., Evans, W., 2013. Maternal bereavement: the heightened mortality of mothers after the death of a child. Economics and Human Biology 11, 371–381.

Folkman, S., 2001. Revised coping theory and the process of bereavement. In: Stroebe, M., Hansson, R.O., Stroebe, W., Schut, H. (Eds.), Handbook of Bereavement Research: Consequences, Coping, and Care. American Psychological Association Press, Washington, DC, pp. 563–584.

Freud, S., 1917/1957. Mourning and melancholia. In: Strachey, J. (Ed.), Standard Edition of the Complete Work of Sigmund Freud. Hogarth Press, London, pp. 240–263.

Horowitz, M., 1986. Stress Response Syndromes. Aronson, Northvale, NJ.

Janoff-Bulman, R., 1992. Shattered Assumptions: Towards a New Psychology of Trauma. Free Press, New York.

Lazarus, R., Folkman, S., 1984. Stress, Appraisal, and Coping. Springer, New York.

Mancini, A., Bonanno, G., 2006. Resilience in the face of potential trauma: clinical practices and illustrations. Journal of Clinical Psychology: In Session 62, 971–985.

Mikulincer, M., Shaver, P., 2013. Attachment insecurities and disordered patterns of grief. In: Stroebe, M., Schut, H., van den Bout, J. (Eds.), Complicated Grief: Scientific Foundations for Health Care Professionals. Routledge, London, pp. 190–203.

Neimeyer, R., 2001. Meaning Reconstruction and the Experience of Loss. American Psychological Association Press, Washington, DC.

O'Connor, M.-F., 2013. Physiological mechanisms and the neurobiology of complicated grief. In: Stroebe, M., Schut, H., van den Bout, J. (Eds.), Complicated Grief: Scientific Foundations for Health Care Professionals. Routledge, London, pp. 204–218.

Parkes, C.M., 2006. Love and Loss: The Roots of Grief and Its Complications. Routledge, London.

Stroebe, M., Schut, H., 1999. The dual process model of coping with bereavement: rationale and description. Death Studies 23, 197–224.

Stroebe, M., Schut, H., Stroebe, W., 2007. Health consequences of bereavement: a review. The Lancet 370, 1960–1973.

Stroebe, M., Schut, H., van den Bout, J. (Eds.), 2013. Complicated Grief: Scientific Foundations for Health Care Professionals. Routledge, London.

Wittouck, C., van Autreve, S., de Jaegere, E., Portzky, G., van Heeringen, K., 2013. The prevention and treatment of complicated grief: a meta-analysis. Clinical Psychology Review 31, 69–78.

Worden, J., 2009. Grief Counselling and Grief Therapy: A Handbook for the Mental Health Practitioner. Springer, New York.

Zisook, S., Kendler, K., 2007. Is bereavement-related depression different than non-bereavement-related depression? Psychological Medicine 37, 779–794.

Bereavement and Complicated Grief across the Lifespan

Hans Jörg Znoj, Institut für Psychologie, Abteilung Klinische Psychologie und Psychotherapie, Universität Bern, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Loss can be seen as a normative life event; nevertheless for most individuals, the experience of the loss of a loved one is extremely stressful. The grief reaction is universal but shaped by cultural norms and individual expectations; grief may also become a serious health problem. The aim of this article is to review the consequences of grief as well as explanations and treatment of complicated grief disorder. Five areas are identified to be of importance to clinicians and researchers: (1) grief reactions, (2) models of grief, (3) bereavement across the life span, (4) diagnostic criteria, and (5) treatment for complicated grief.

Introduction

Bereavement after a loss is a normative but threatening life event that can lead to serious consequences, such as anxiety, sleep disorders, problems to adapt, depression, and even death. A loss of a person is often shared among many individuals who have developed different bonds with the deceased person. Henceforth, grief is often a shared experience but bereavement is rarely the same for different individuals. Besides individual coping with loss, there are cultural aspects of grief and bereavement that also have to be accounted for. Evidence exists that grief can become a disorder, similar to other emotional disturbances such as anxiety disorders or disruptive aggression (Shear et al., 2011). Therefore, the question of therapeutic intervention must be seen in this context.

In this article, the focus will be on coping with loss across the life span, the question how grief can be conceptualized as a normal reaction to loss and on risk factors for developing a loss-related disorder. The proposed diagnostic criteria will be presented in this article.

Grief Responses

Humans share as other social animals the capacity to build long-lasting social relationships. Attachment is perceived as a natural basic need and we are attached to others in various ways, as dependent children and lovers, or because we share the same living circumstances. Following the work of Bowlby (e.g., 1980), the loss of an attachment figure may lead to massive stress, even to death. In a biological sense, grief is the price humans pay for the capacity to form social bonds, groups, and communities. Grief is the individual response to separation distress, which is caused by a genetically provided subcortical network (Panksepp and Watt, 2011) leading to emotional responses such as panic, sadness, and a diminished strive for dopamine-driven activities. Indeed, many bereaved parents expressed their grief in metaphors relating to bodily loss (Znoj, 2004). Following the work of Stroebe and coworkers (Stroebe and Schut, 2001), the intensity of the grief response is more related to the relationship of the deceased person and less to other circumstances of the loss. This corroborates the notion that grief and related disorders are consequences of a breakdown of the bonding system and constitute a specific form of threat.

Models of Grief

A variety of models aim at explaining grief. Phase models of bereavement are still popular by providing a system that is both logical and dynamic. After initial shock and disbelief, the bereaved is assumed to go to further stages of adaptation, namely to realization of the loss and grief work that includes the gradually disengagement from the target person and finally the resumption of emotional life and the capacity to form new intimate relationships. These phases are generally conceptualized as sequences. However, most researchers agree that this view is oversimplified.

Some models focus on coping with losses. The task model (Worden, 1996) suggests four tasks to adjust to bereavement. In contrast to just passively experiencing the loss, it focuses on active coping with the challenges, namely accepting the reality of the loss, experiencing the pain, adjusting to an environment without the deceased person, and finally finding a new understanding with the deceased. Coping models often focus on various coping styles, such as avoidant coping or rumination (e.g., Nolen-Hoeksema and Jackson, 1996) that often lead to elevated depression and higher distress levels later on. In contrast, a problem-oriented coping leads to better adaptation and includes distractive coping, even cognitive avoidance (e.g., Bonanno, 2004). Positive meaning and meaning reconstruction following bereavement have been linked to positive outcomes following loss; the capacity to display positive emotions during bereavement predicts better outcome after the loss of a spouse (Bonanno and Keltner, 1997). Taken together, these results may indicate that too much avoidance as well as too much confrontation is detrimental to adaptation following loss. Stroebe and Schut (1999) proposed the dual-process model that integrates both stress and coping theories as well as psychosocial models, such as the two-track model (Rubin, 1981) addressing both, the biosocial responses to bereavement and the transformation of the attachment that may include a new still ongoing relationship with the deceased person. The integrative dual-process model of grief takes into account the empirical evidence and postulates an oscillation

between loss-oriented coping and restoration-oriented coping. Loss-oriented coping includes positive reappraisal versus rumination, wishful thinking, revisions of personal goals, positive and negative event interpretation, and expressing emotions and mood states such as dysphoria or positive effect toward the deceased. On the other hand, restoration-oriented coping is focused on attending to life changes, doing new things, distracting from grief, and finding new roles and identities.

For stress–response disorders in general, Horowitz' (1978) model of working through a traumatic event posits an oscillation between phases of intrusion and avoidance as necessary process for adaptation. A good fit can be found between the dual-process model (Stroebe and Schut, 1999) and deepened investigation of risk factors as has been shown for cognitive–emotional changes after bereavement (Znoj, 2004). For instance, anger about the circumstances of the death of a loved one could lead to more severe grief, specifically when the death is perceived as unjust, such as in the case of the death of a child. For example, Znoj et al. (2004) investigated bereaved parents and found high correlations of the feeling that fate is unjust and increasing psychopathology. Orth and Maercker (2009) demonstrated that anger, in addition to post-traumatic stress disorder (PTSD) symptoms, leads to further aggravation of symptomatology.

In Western societies, the belief exists that grief should be resolved within 12 months (Wortman and Silver, 1989). Zisook and Schuchter (1986) described persons within the first year of widowhood. While the majority had adapted well after 4 years, some individuals still felt responsible for the loss and reported symptoms like preoccupation with thoughts of the spouse, or had clear visual images; moreover 40% reported elevated levels of depressive symptoms. Overall, the symptoms degraded slowly over time. Obviously, the process of grief lasts longer than usually assumed by the general population. Some of the symptoms clearly have a negative impact on well-being, such as intrusive longing, pain, bouts of crying, and worrying. However, depressive thoughts may also hint toward personal growth and the stimulation of human resilience (e.g., Znoj, 2006).

Loss Experience and Bereavement across the Life Span

Vivid memories, emotional turmoil, high levels of anxiety, and depression are seen as normal reactions to the loss of a beloved person in adults. Information about grief reactions in children or young adults is scarce. Depending on age and developmental stage, a child or young adult will have different concerns. Following Christ et al. (2002), between 3 and 5 years of age, the concept of death is hard to grasp or even absent for most children. Therefore the child will ask about the deceased person as if he or she will come back or is just temporarily unavailable. Magic thinking and fantasy are predominant and the child will experience a great deal of distress when separated from the primary caregiver but eventually accept a competent substitute. Between 6 and 8 years of age, the child begins to understand that death is a universal fact and therefore the parent will not return (Oltjenbruns, 2001). However, the causes and effects are often not well understood, and parents are an important source of emotion regulation and self-esteem. Fear, anger, and aggression, or wishes to be together with the deceased parent can be harmful (e.g., danger of suicide). Between 9 and 11 years, the child becomes slowly independent, the improved logical thinking supports understanding the cause and effect, but even then child needs concrete details (e.g., about illness) to come to a conclusion. Emotion regulation relies on distraction strategies and seeks actively support with peers and relatives. Bereaved adolescents often have a strong sense of the dead parent's presence and accept him or her as a gender role model. In older adolescents between 15 and 17 years of age, grief has adult characteristics with overwhelming sadness and painful memories but usually the duration of intensive grief is shorter than among adults.

In each of these developmental stages, grief can become overwhelming and a source of intense distress. Complicated grief (CG) can be observed in any of these stages and is usually characterized by deterioration of social aspects as well as depression, including regressive states, phobias, loss of emotional and behavioral control, or reduced cognitive ability.

Because the cognitive and emotional apparatus is changing, Oltjenbruns (2001) observed the phenomena of regrief, meaning that at a later developmental stage, the loss can be reexperienced and more intense than before. In this perspective, the loss of a mother with the 4 years of age can have a different but no less powerful impact on a young woman of 17 years of age, when she would need her motherly advice and support to develop her own life in the transition to adulthood.

Silverman and Worden (1992) investigated 125 children from 6 to 17 years of age within 4 months of the death of a parent. After 4 months, the effective response stabilized fairly for most children; however, 74% reported headaches or other somatic problems, one-third had sleeping problems they attributed to their parent's death. Further problems included difficulties concentrating in school, hassling by peers, or uneasiness at the absence of one parent. To remain connected with the deceased parent seemed to be a strong motive for children and youths at least in this study. In sum, "the deceased parent continued to be a real presence for many of the children who carried on a relationship similar to the one prevailing before the death" (Silverman and Worden, 1992: p. 100). Balk and Corr (2001) resumed that the risk of depression following the death of a parent is seven times higher than the norm. In an effort to shed more light on this underinvestigated area, Schwartz (2003) interviewed 39 youngsters between 10 and 22 years of age in the Bern study within 6 months to 5 years after the loss. For more than 50% of these youngsters, the death came sudden, unheralded. Eighteen percent of the bereaved children and youngsters had clinical relevant levels of depression. Avoidance and retreat were closely related to depression ($r = 0.50$, $p < 0.01$), whereas seeking support was negatively related to depression. Although 50% of the bereaved reported to be not religious, almost all wanted their parents to be well-off in Heaven and almost three-fourth uttered the opinion that their deceased parent can hear them, and therefore tried to contact them via praying. Almost half of the bereaved children indicated having vivid dreams and feelings of presence of their deceased parent. One-third even heard their deceased parent as if he or she would still be alive. Level of spirituality and religion were positively correlated with social support, especially by

grandparents ($r = 0.49$, $p < 0.01$), and also with feelings of presence ($r = 0.33$, $p < 0.05$).

In terms of long-term consequences of childhood adversities, Cuijpers et al. (2011) also looked at bereavement in a large population sample in The Netherlands. While childhood adversities in general had very high impact on the disease burden in adulthood, the death of a parent had no effect. This suggests that bereavement in childhood might not necessarily lead to long-term negative consequences. However, this finding needs to be replicated.

Child loss is one of the most challenging losses a person can experience. In the Bern study of parental bereavement, we (e.g., Znoj, 2004, 2012) found extremely high levels of intrusions as measured with the Impact of Event Scale (Maercker and Schützwohl, 1998). Bereaved parents also reported high levels of depression; they often experienced the loss as if they were physically amputated and in physical pain. The loss of a child can be seen as a serious risk factor for prolonged grief.

For elderly people, the loss of a close person is a normative event and especially old age is a time of multiple sequential losses. After Moss and Moss (1986), two-fifth of persons over 60 years of age experienced already the loss of a sibling. Conjugal bereavement, the death of a spouse, is also an important topic and often the bond persists and continues beyond the death. Following Lopata (1996), loneliness is among the most mentioned consequences of widowhood. Most widows and widowers live alone but many acquire new skills and develop in positive ways. The impact of the loss may be less intense than in younger years but the emotional distress and grief decline more slowly among older persons (Moss et al., 2001). Mortality risk is elevated, especially for widowed men (e.g., Verbrugge, 1989). Moreover, a special situation arises when there is a generational violation, especially when a child dies before an elderly parent. In addition to intense emotional distress, such a loss may imply a breakdown of the supportive system and may have serious health consequences. Research is rare, despite the fact that adult children play a major part in the social and emotional world of parents throughout their lifetime (Moss et al., 2001). The death of an old person – on the other hand – is often experienced as a relief (the suffering has ended) and may lower the emotional impact of the loss for close persons.

Diagnostic Criteria for Disordered Grief

Emotional and behavioral consequences of a loss are intensive sadness, anxiety, aggression toward the deceased or persons related to the loss, and, sometimes, dissociative states. In contrast to other traumatic experiences, the experience of a loss is related to the capacity of building lasting bonds to significant others, including parents, spouses, children, and close friends.

Nevertheless, the reaction to loss is structurally similar to traumatic events (Horowitz, 1978) and often results in emotional dysregulation. For most people, grief intensity weakens to a manageable degree within several months. After the most intensive period, grief is still present but the loss becomes gradually integrated and no longer hinders the way of ongoing life. Similarly, most immediate stress reactions following life-threatening events recede over time. Only when stress symptoms exist beyond 4 weeks, a diagnosis PTSD can be given.

Horowitz who provided the first description of the PTSD criteria and symptoms had pointed early to similarities shared by psychotrauma (PTSD) and grief. He and colleagues called this family of failure to adapt to stress the 'stress–response syndromes' (e.g., Horowitz, 1978). This concept is becoming increasingly recognized.

Up to today, grief patients receive a variety of International Classification of Diseases (ICD) or Diagnostic and Statistical Manual of Mental Disorders (DSM) diagnoses, for example, PTSD, depressive, anxiety, adjustment, and personality disorders (Enright and Marwit, 2002). The clinical utility of these labels, however, has to be called into question and mostly they do not describe the difficulties experienced after the loss of a loved one adequately.

The point at which the state of a mourning person becomes a disorder has been debated from many strands. The debate centers on the extent to which complicated or prolonged grief represents a unique pathological entity when contrasted with normal grief, but also with PTSD or major depression. However, consensus is that grief can be complicated, much as wound healing can be complicated (Shear et al., 2011). One easy accessible indicator is to listen to clients or patients. Self-statements like "I fear, I go crazy if I fully realize the death of my loved one" is very specific to CG but not to depression (Boelen et al., 2006).

Diagnostic Criteria for Complicated Grief Disorder

Only recently have Prigerson, Horowitz and other proponents of a separate classification for complicated grief disorder (CGD) reached a consensus on its clinical criteria (Prigerson et al., 2009). The consensus results from a reanalysis of field trial data from the Yale Bereavement Study (317 participants). These participants were interviewed repeatedly at different time points after the loss of a close person. The symptoms of CGD were assessed using an extended rater version of the Inventory of Complicated Grief-Revised (Prigerson et al., 1995). Analyses aimed to derive a set of informative, unbiased symptoms that allow proposing a complete set of 'DSM-style' diagnostic criteria.

The resulting set of diagnostic criteria specifies that a bereaved person with CGD must experience yearning and at least five of nine additional symptoms (**Table 1**). These symptoms must persist for at least 6 months following the loss and must be associated with functional impairment.

Finally, the study showed that individuals given a CGD diagnosis 6–12 months after a death have a 2.4 times higher subsequent risk of mental health and functional impairment than people not diagnosed with CGD.

Treatment of Complicated (Prolonged) Grief Disorder

In the following, current treatment approaches to CGD will be outlined. Shear's approach of treating CGD has become widely recognized (Shear et al., 2005). This approach was essentially informed by the imagined and *in vivo* exposure techniques used for PTSD, including the confrontational technique of 'revisiting' the deceased loved one. With this technique, the therapist asks patients to close their eyes and tell the story of the death.

Table 1 Prolonged grief disorder criteria proposed for inclusion in *Diagnostic* and Statistical Manual of Mental Disorders V (DSM-V) (by the American Psychiatric Association)

1. Event: the person has experienced bereavement, i.e., the loss of a significant other.
2. Separation distress: the bereaved person experiences separation distress most days and to a disabling degree, as manifest by yearning, longing, craving, or pining for, or preoccupation with the deceased person.
3. Additional cognitive, emotional, behavioral symptoms of grief: the bereaved person must have five (or more) of the following symptoms experienced most days and to a disabling degree:
   a. Confusion about one's role in life or diminished sense of self (i.e., feeling that a part of oneself has died)
   b. Difficulty in accepting the loss
   c. Avoidance of reminders of the reality of the loss or avoidance of thoughts, activities, or situations that arouse intense emotions related to the loss
   d. Inability to trust others or feeling alone or detached from others since the loss
   e. Bitterness or anger related to the loss
   f. Difficulty in moving on with life (e.g., making new friends, pursuing interests, feeling life no longer holds the potential for satisfaction or joy)
   g. Numbness (absence of emotion) since the loss
   h. Feeling that life is unfulfilling, empty, meaningless, or unbearable since the loss
   i. Feeling stunned, dazed, or shocked by the loss
4. Timing: diagnosis should not be made until at least 6 months have elapsed since the death.
5. Impairment: the disturbance causes clinically significant impairment in social, occupational, or other important areas of functioning (e.g., domestic responsibilities).
6. Relation to other mental disorders: the disturbance may co-occur with but is not better accounted for by major depressive disorder, generalized anxiety disorder, or post-traumatic stress disorder.

Prigerson, Horowitz, and 17 coauthors, 2009; adapted from Maercker, A., Znoj, H.J., 2010. The younger sibling of PTSD: similarities and differences between grief and posttraumatic stress disorder. European Journal of Psychotraumatology 1 (5558), 1–9.

The therapist tape-records the story and periodically asks the patient to report distress levels. The patient is given the tape to listen to at home during the week. Distress related to the loss (e.g., yearning, or fears of losing the deceased forever) is targeted using techniques to promote a sense of connection to the deceased. These include an imagined conversation with the deceased. This technique is complemented by preceding psychoeducation and a subsequent 'restoration of life goals' phase. Shear et al. (2005) compared this treatment to standard interpersonal psychotherapy. The results were much in favor of Shear's CG treatment, so that this treatment approach is now being disseminated and applied around the world.

Boelen et al. (2007) studied a modified form of this treatment and examined different sequences of exposure and cognitive restructuring (CR). Exposure began with the writing of distressing memories and included in sensu exposure during the sessions. CR focused on individual dysfunctional thoughts (e.g., guilt, anger). Doing exposure first, followed by CR, yielded the best results.

Because no other approaches have been shown to be effective aside from those tested in these randomized trials, the Web-based approach for CGD (Wagner et al., 2007) might be of importance. This approach also begins with a technique of self-confrontation similar to that used for PTSD consisting of a written assignment. Patients are asked to write four texts where they describe the circumstances of their bereavement and their thoughts and feelings at that time. The therapist facilitates the nonavoidance of fear related to these memories. The CR is then based on these accounts focusing mostly on the feelings of guilt or responsibility in the death of the loved one. Patients are then asked to write a supportive letter to a friend who finds himself or herself in exactly the same situation. A further intervention consists in establishing rituals or activities to commemorate the deceased. The aim is to give the deceased a place in everyday life, to reorganize priorities, and to see how the patient can consider reconnecting with friends and social life. Within 10 weeks, patients write a total of 10 assignments, to which they receive an individual feedback from the therapist via e-mail within one workday.

Due to the Web-based intervention, the group of patients was highly selective. There were a number of exclusion criteria that were necessary due to the virtual nature of the relationship between patient and therapist. In a randomized controlled trial with a wait-list control condition, the effect sizes were in the range of 1.2–1.5, and 1.2–1.6 for pretreatment to 3-month follow-up. Follow-up measures at 18 months confirmed the stability of these effects (Wagner and Maercker, 2007). Further results confirmed functional change in the five domains of posttraumatic growth (new possibilities, relating to others, appreciation of life, personal strength, spiritual change), but no significant increase in optimism was observed (Wagner et al., 2007).

Overall, evidence shows that good therapeutic success can be achieved with exposure and CR. The efficacy of these two standard modules, also found in therapy for PTSD, will perhaps be enhanced with a third module consisting of social sharing. The sharing of pain with others in an appropriate way seems to be an important issue. Social sharing as treatment goal may also be specifically suited for group treatments of CGD.

Outlook

In PTSD, people typically fail to assimilate their experiences and have prevailing fundamental beliefs and specific experiential readiness. The consequences of PTSD are strong negative emotions and a constant cognitive preoccupation to detect dangerous inconsistencies. In contrast, in CGD the predominant feeling is not threat but loss-related distress. The persisting inconsistency concerns lack of affiliation. A failure to adapt following a major loss of an attachment figure may not only lead to a complicated or prolonged grief disorder but also to other forms of psychopathology such as depression or panic disorder. It is probably of highest importance not only to look at grief-specific symptoms but also to the individual processes of coping. Our own work (e.g., Znoj, 2008) has shown that processes of emotion regulation may lead to completely different outcomes after the loss of a loved one. Grief and the experience of loss and death is part of the human condition, social and cultural aspects shape the individual's response. The question whether we perceive a grief response as part of the social life or whether we focus on individual suffering and

consider the necessity of individual treatment requires careful indication and extensive future developments.

See also: Bereavement; Coping across the Lifespan; Death and Dying, Psychology of; Death and Dying, Sociology of; Depression; Mass Trauma: Psychopathological Effects across the Life Span; Post-Traumatic Stress Disorder.

Bibliography

Balk, D., Corr, C., 2001. Bereavement during adolescence. In: Stroebe, M.S., Hansson, R.O., Stroebe, W., Schut, H. (Eds.), Handbook of Bereavement Research: Consequences, Coping, and Care. American Psychological Association, Washington, DC, pp. 199–218.

Boelen, P.A., Keijser, J., van den Hout, M.A., van den Bout, J., 2007. Treatment of complicated grief: a comparison between cognitive-behavioral therapy and supportive counseling. Journal of Consulting and Clinical Psychology 75, 277–284.

Boelen, P.A., van den Bout, J., van den Hout, M.A., 2006. A cognitive-behavioral conceptualization of complicated grief. Clinical Psychology: Science and Practice 13, 109–128.

Bonanno, G.A., 2004. Loss, trauma, and human resilience. Have we underestimated the human capacity to thrive after extremely aversive events? American Psychologist 59 (1), 20–28.

Bonanno, G.A., Keltner, D., 1997. Facial expressions of emotion and the course of conjugal bereavement. Journal of Abnormal Psychology 106, 126–137.

Bowlby, J., 1980. Attachment and Loss. In: Loss: Sadness and Depression, vol. 3. Hogarth Press & Institute of Psychoanalysis, London.

Christ, G.H., Siegel, K., Christ, A.E., 2002. Adolescent grief "it never really hit me … until it actually happened". JAMA 288 (10), 1269–1278.

Cuijpers, P., Smit, F., Unger, F., Stikkelbroek, Y., Ten Have, M., de Graaf, R., 2011. The disease burden of childhood adversities in adults: a population-based study. Child, Abuse and Neglect 35, 937–945.

Enright, B.P., Marwit, S.J., 2002. Diagnosing complicated grief: a closer look. Journal of Clinical Psychology 58, 747–757.

Grawe, K., 1998. Psychologische Therapie. Hogrefe, Göttingen.

Horowitz, M., 1978. Stress Response Syndromes. Jason Aronson, New York.

Lopata, H.Z., 1996. Current Widowhood: Myths and Realities. Sage, Thousand Oaks, CA.

Maercker, A., Schützwohl, M., 1998. Erfassung von psychischen Belastungsfolgen: die impact of event Skala-revidierte version. Diagnostica 44, 130–141.

Moss, M.S., Moss, S.Z., 1986. Death of an adult sibling. International Journal of Family Psychiatry 7, 397–418.

Maercker, A., Znoj, H.J., 2010. The younger sibling of PTSD: similarities and differences between grief and posttraumatic stress disorder. European Journal of Psychotraumatology 1 (5558), 1–9.

Moss, M.S., Moss, S.Z., Hansson, R.O., 2001. Bereavement and old age. In: Stroebe, M.S., Hansson, R.O., Stroebe, W., Schut, H. (Eds.), Handbook of Bereavement Research: Consequences, Coping and Care. American Psychological Association, Washington, DC, pp. 241–260.

Nolen-Hoeksema, S., Jackson, B., August 1996. Ruminative Coping and Gender Differences in Depression. Paper presented at the 104th Annual Convention of the American Psychological Association, Toronto, Ontario, Canada.

Oltjenbruns, K.A., 2001. Development context of childhood: grief and regrief phenomena. In: Stroebe, M.S., Hansson, R.O., Stroebe, W., Schut, H. (Eds.), Handbook of Bereavement Research: Consequences, Coping, and Care. American Psychological Association, Washington, DC, pp. 169–197.

Orth, U., Maercker, A., 2009. Posttraumatic anger in crime victims: directed at the perpetrator and at the self. Journal of Traumatic Stress 22, 158–161.

Panksepp, J., Watt, J., 2011. Why does depression hurt? Ancestral primary-process separation-distress (PANIC) and diminished brain reward (SEEKING) processes in the genesis of depressive affect. Psychiatry 74, 5–14.

Piaget, J., 1963. The Origins of Intelligence in Children. Norton, New York.

Prigerson, H.G., Horowitz, M.J., Jacobs, S.C., Parkes, C.M., Aslan, M., Goodkin, K., et al., 2009. Prolonged grief disorder: psychometric validation of criteria proposed for DSM-V and ICD-11. PLoS Medicine 6 (8), e1000121.

Prigerson, H.G., Maciejewski, P.K., Newsom, J., Reynolds, C.F.I., Frank, E., Bierhals, E.J., et al., 1995. The inventory of complicated grief: a scale to measure certain maladaptive symptoms of loss. Psychiatric Research 59, 65–79.

Rubin, S., 1981. A two-track model of bereavement: theory and application in research. American Journal of Orthopsychiatry 51, 101–109.

Schwartz, L., 2003. Bedeutung und Bewältigung des Todes eines Elternteils im späten Kindesalter und in der Adoleszenz (Unpublished Master's thesis). University of Bern.

Shear, K., Frank, E., Houck, P.R., Reynolds, C.F., 2005. Treatment of complicated grief: a randomized controlled trial. JAMA 293, 2601–2608.

Shear, M.K., Simon, N., Wall, M., Zisook, S., Neimeyer, R., Duan, N., et al., 2011. Complicated grief and related bereavement issues for DSM-5. Depression and Anxiety 28 (2), 103–117.

Silverman, P., Worden, J., 1992. Children's understanding of funeral rites. Omega 25, 319–331.

Stroebe, W., Schut, H., 2001. Risk factors in bereavement outcome: a methodological and empirical review. In: Stroebe, M., Hansson, R.O., Stroebe, W., Schut, H. (Eds.), Handbook of Bereavement Research: Consequences, Coping and Care. American Psychological Association Press, Washington DC, pp. 349–372.

Stroebe, M.S., Schut, H., 1999. The dual process model of coping with bereavement: rationale and description. Death Studies 23, 197–224.

Verbrugge, L.M., 1989. The twain meet: empirical explanations of sex differences in health and mortality. Journal of Health and Social Behavior 30, 282–304.

Wagner, B., Maercker, A., 2007. 1.5 year follow-up of an internet-based intervention of complicated grief. Journal of Traumatic Stress 20 (4), 625–629.

Wagner, B., Knaevelsrud, C., Maercker, A., 2007. Posttraumatic growth and optimism as outcomes of an internet-based intervention for complicated grief. Cognitive Behaviour Therapy 36, 156–161.

Worden, W.J., 1996. Tasks and mediators of mourning: a guideline for the mental health practitioner. In Session: Psychotherapy in Practice 2 (4), 73–80.

Wortman, C., Silver, R., 1989. The myths of coping with loss. Journal of Consulting and Clinical Psychology 57, 349–357.

Zisook, S., Schuchter, S.R., 1986. The first four years of widowhood. Psychiatric Annals 16 (5), 288–294.

Znoj, H.J., 2012. Trauer und Trauerbewältigung. Psychologische Konzepte im Wandel. Kohlhammer, Stuttgart.

Znoj, H.J., 2008. Regulation emotionaler Prozesse in Psychotherapie und Verhaltensmedizin. Peter Lang, Bern.

Znoj, H.J., 2006. Bereavement and posttraumatic growth. In: Calhoun, L., Tedeschi, R. (Eds.), The Handbook of Posttraumatic Growth: Research and Practice. Lawrence Erlbaum, Mahwah, NJ, pp. 176–196.

Znoj, H.J., 2004. Komplizierte Trauer. Leitfaden für Therapeuten. Hogrefe, Göttingen.

Znoj, H.J., Morgenthaler, C., Zwingmann, C., 2004. Mehr als nur Bewältigen? Religiosität, Stressreaktionen und Coping bei elterlicher Depression nach dem Verlust eines Kindes. In: Zwingmann, C., Moosbrugger, H. (Eds.), Religiosität: Messverfahren und Studien zu Gesundheit und Lebensqualität. Neue Beiträge zur Religionspsychologie. Waxmann, Münster, pp. 277–298.

Berlin, Isaiah (1909–97)

Hans-Peter Müller, Humboldt-University of Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Isaiah Berlin (1909–97) was a British philosopher, social and political theorist, essayist, and intellectual. He is most renowned for his work in the history of ideas, a field that he made attractive again for philosophy and the social sciences against the current of analytical philosophy. His major analytical contributions concern the distinction between positive and negative freedom and the problem of value pluralism in modern societies. Upon his death, he was above all venerated for his personality, his conversational art, his brilliant style, his steadfast opposition to totalitarianism and his unflinching commitment to liberalism.

Isaiah Berlin was a British philosopher and an ardent liberal who defended political freedom in modern society against totalitarian temptations of all kinds. He started his academic career as a conventional philosopher preoccupied with logical positivism and language philosophy but soon began to work in the field of the history of ideas. He focused on the enlightenment and developed a keen interest in the thought of the counterenlightenment; he also dealt with romanticism and the Russian thinkers, above all Alexander Herzen. The main thrust of his multiple and voluminous essays is not easy to capture. Berlin never made a serious attempt to systematically elaborate his own approach in the history of ideas. An exception from this might be his narrative 'My intellectual path' (Berlin, 2000: 1–23) in which he laid out the development of his thought. When pressed to display his intellectual agenda, he used to answer laconically: "I am an intellectual taxi: people flag me down and give me destinations and off I go" (Ignatieff, 1998: 7).

A strong aversion to theoretical systems was part and parcel of Berlin's intellectual habitus. He preferred to work from a kind of latent 'systématique' without spelling out its logic. When reconstructing the systematic approach behind his articles, it should thus be clear that we are dealing with an interpretation of Isaiah Berlin's work open for refutation. In my view, there is neither a coherent intellectual system nor one *idée directrice* to be found throughout Berlin's work (though liberty and value pluralism would be possible candidates). Instead, Berlin employs what – with Arthur O. Lovejoy (1942) – might be called an ensemble of unit-ideas, certain core thoughts or elements in the history of philosophical doctrines. We can use this concept of unit-ideas for the reconstruction of Isaiah Berlin's approach to the history of ideas as well as the evolution of this approach: In his social and political philosophy the unit-idea he started from was the rational monism of the Enlightenment. This unit-idea is characterized by a tripartite dogma: (1) in epistemology and ethics there is but one answer to all genuine problems; (2) if found, this answer represents the truth; and (3) all true propositions cannot conflict with one another because there is only one truth.

While Berlin acknowledges the liberating ideals of perfectibility and progress that this notion entails, he is also alert to the danger that when realized recklessly by 'enlightened elites' these very ideals might lead to oppression and totalitarianism instead of a liberal and open society. The antimonist stance Berlin develops from this insight also leads him to his other three unit-ideas: The modern condition is above all characterized by value pluralism, the quest for liberty, and political liberalism. In working out these unit-ideas Berlin draws on Counter-Enlightenment and Romanticist thought as well as the Russian intellectual tradition. Although it remains somewhat uncertain what ultimately holds Berlin's *oeuvre* together, in my view his intellectual agenda can be identified as the attempt to decipher the constellation of values and institutions of modernity, which ensure that liberty is not only paid lip service by governments and states but can be enjoyed by ordinary people in everyday life.

Life and Career

Isaiah Berlin was born on 6 June 1909, in Riga, Latvia, which belonged to the Russian Empire at the time. His father, Mendel Berlin, was a prosperous timber merchant. The family moved to St Petersburg in 1917 where young Isaiah would witness the violence of the Russian Revolution. This early experience made him a stern anticommunist throughout his life. He came to see the 'true freedom' communists promised as a mere pretext for an oppressive rule based on violence and totalitarian ideology. Soon after the return to Riga in 1920, the family decided to emigrate to Britain.

In London, Isaiah Berlin visited St Paul's School and went to Corpus Christi College, Oxford. He studied Greats (philosophy, ancient history, and classical languages) and philosophy, politics, and economics. He entered All Souls upon election to a Prize fellowship and was the first Jew to be nominated. Berlin started a career as a philosopher at Oxford. But after he had given a paper on 'Other Minds' in Cambridge followed by a fierce discussion with the young Ludwig Wittgenstein (who politely dismissed him by saying "Very interesting discussion. Thank you" (Ignatieff, 1998: 95)), he gave up philosophy in favor of the history of ideas.

Oxford turned out to be his *alma mater* for a lifetime. Upon his death, his All Souls colleague William Waldegrave remarked in the *Daily Telegraph*, "if you had asked me to show you what I meant by the ideal of Englishness, I would have taken you to see a Latvian, Jewish, German, Italian mixture of all the cultures

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61160-1

of Europe. I would have taken you to see Isaiah Berlin" (Ignatieff, 1998: 300).

Yet, despite this story of successful integration into a renowned English institution, his social environment would not forget that he was a Jewish-Russian *émigré*. During the Second World War, Berlin worked as a correspondent first in New York City, then in Washington DC in the British Embassy. He was charged with delivering weekly reports from the United States. Due to his brilliance, those reports soon gained notoriety. Prime Minister Churchill wrote to his Foreign Secretary in 1944: "The summaries are certainly well written. I have a feeling that they make the most of everything and present a somewhat perfervid picture of American affairs." Eden commented in his own hand: "I agree. There is perhaps a too generous Oriental flavor" (Ignatieff, 1998: 125).

The years after the war elevated Berlin's prominence to a new level. His reports from Washington had gained him a reputation in Whitehall and the upper society was curious to come to know this young and gifted philosopher from Oxford. The British Broadcasting Corporation invited him to broadcast and Churchill let Berlin review his memoirs. The journal *Foreign Affairs* invited him to contribute and his two articles earned him the reputation of being an expert on Soviet Russia. As a consequence, he was invited to US universities, among them Harvard, Bryn Mawr, and Princeton.

Yet, his worldliness and frivolity – he was afraid to become an intellectual 'chatterbox' – raised the question of his own intellectual agenda. He found it in Russian intellectual history, the history of socialism and Marxism, the Enlightenment and its critics. In 1950, he returned to All Souls with a fellowship and devoted his time exclusively to research. In 1957, he became the Chichelle Professor of Social and Political Theory at Oxford, the same year he had married Aline Halban (née de Gunzbourg) and was knighted. His inaugural lecture in 1958 developed the famous distinction of the 'Two concepts of Liberty.' Berlin resigned his Chair when he was nominated to be the founding President of Wolfson College in 1967. From 1966 to 1971 he served as a visiting Professor of Humanities at the City University of New York. He became the President of the British Academy in 1974 and resigned from this post in 1978. Berlin received numerous prizes – among them the Agnelli, Erasmus, and Lippincott Prizes – for the history of ideas and the Jerusalem prize for defending civil liberties throughout his life. He was awarded with a number of honorary degrees.

In retrospect, Berlin's intellectual identity was the result of a number of different forces, of which three bear special mention: a Russian, British, and Jewish one. The Russian thinkers, above all Herzen and Turgenev, furnished in him the fascination for ideas and the proper moral role of the intellectual. The British environment instilled in him a sober sense of reality and the matter-of-fact mentality of empiricism as well as liberalism and tolerance. His Jewish heritage led him to pursue a moderate Zionism, probably an expression of his own need for belonging. He steadfastly defended this position against the critique of cosmopolitans.

During his lifetime, Isaiah Berlin published just one monograph, his classic on Karl Marx in 1939. Starting in 1978, Henry Hardy, a student of Berlin, set out to publish his other writings. Thanks to Hardy's editorial work, an amazing oeuvre has emerged in recent years comprising 13 books, among them an anthology, 'The Proper Study of Mankind' (Berlin, 1997) and a collection of letters in three volumes (2004, 2009, 2013a). Berlin died on 5 November 1997, in Oxford.

The Early Achievement: Isaiah Berlin's 'Karl Marx'

Berlin, at the time just 30 years old, wrote a remarkable intellectual biography of Karl Marx, which is still in print. The subject was somewhat inopportune at the time because soon after his first book was published, the Second World War broke out. The original manuscript was twice as long. But Berlin was pressed by his publishing house to cut one-half in order to bring the book to the agreed-upon length. The young author rewrote his manuscript more as an intellectual biography than an exegesis of Marx's theory of history, his economic and sociological analysis.

What is the trademark of Berlin's Marx? Three critical insights are decisive: (1) Berlin's empathy for Marx, (2) the problem of historical inevitability, and (3) the role of ideas. Berlin draws a balanced and nuanced picture of Marx's accomplishments and weaknesses. In his first major work, Berlin already exhibits a unique ability to bring historical persons and constellations to life. Berlin's great virtue is precisely his 'empathy for the other.'

It goes without saying that Karl Marx embodies for Berlin the 'other.' The project in this earliest work is to make Marx the adversary look as strong as possible and then demonstrate why he is dangerously wrong.

Marx's historical materialism set out to determine the laws of social development based upon a model of economy and society, the famous 'base' and 'superstructure.' Yet, for Berlin, Marx made much too strong a point in claiming this sort of 'historical inevitability,' the necessary transformation of bourgeois capitalism into proletarian socialism. In Berlin's eyes, the history of mankind is contingent, the development of society open, and the future unpredictable. Socialism may be mankind's future but this is dependent upon the ideas people have, the course of action they take, and critically, the amount of blood they are willing to spill.

The third point pertains to the role of ideas in Marx's theory of history. According to Marx, ideas are ideological illusions. What determines the course of history and society are the iron laws of capitalist production and contradiction. Ironically, Marx's own ideas belie this interpretation, since those ideas made a great deal of world history themselves.

In his struggle with one of the most important thinkers of the nineteenth century, Berlin tapped into problems and questions he would deal with throughout his academic career: the philosophy of history, historical inevitability, progress and perfectibility, the role of ideas. In retrospect, Berlin credited Marx with having contributed three main ideas: "The idea of Marx that was genuinely original was the idea of the influence of technology on culture, that technological change influences culture to a profound degree. (…) The other idea was about big business and what it did. (…) The third idea is that of the class war. Class warfare has been exaggerated, but nevertheless exists as a phenomenon. There is such a thing as conflict between classes" (Berlin and Lukes, 1998: 71).

Two Concepts of Liberty

Berlin's most famous work is the inaugural lecture 'Two Concepts of Liberty' given in 1958 after he had taken the Chair of Social and Political Theory at Oxford. Berlin tries to turn the manifold meanings of this notion into a coherent concept by drawing a distinction between negative and positive freedom. Negative liberty (Berlin used to employ the terms liberty and freedom interchangeably) refers to a private realm of the individual where he or she is free to choose: "By being free in this sense I mean not being interfered with by others. The wider the area of non-interference the wider my freedom" (Berlin, 1969: 123). In particular, political liberty requires a strong frontier between the public and private realm. Political authorities, therefore, need to show respect for this private domain of the individual. According to Berlin, this notion of negative liberty or 'freedom from' refers to a widespread, 'normal' understanding of the term and coincides with the main thrust of English philosophy. Freedom of choice, noninterference, and the distinction between public and private define the core of this meaning of liberty. Berlin is fully aware that however high the value of liberty, there are other values making legitimate claims too: equality, justice, solidarity, etc. Yet, it would be a dangerous 'confusion of values' to confound liberty with those other values. "Everything is what it is: liberty is liberty, not equality or fairness or justice or culture, or human happiness or a quiet conscience" (Berlin, 1969: 125).

Berlin holds that the value of freedom has its all-embracing meaning predominantly in civilized societies and even there most of all for established citizens. "Freedom for an Oxford don (…) is a very different thing from freedom for an Egyptian peasant" (Berlin, 1969: 124). Berlin is ready to admit that first things come first. If these 'first things' are medicine or clothes, they may have a higher priority in the life of an Egyptian peasant. But even here, Berlin assumes, the peasant will not be able to carry on without a minimum of freedom. He is not willing to follow the argument – noble as it may seem – that as long as there is oppression and exploitation in the world, as long as some parts of the world are still in chains, one has to sacrifice one's freedom. It may well prove necessary to restrict freedom in order to avoid misery. But such a restriction is a sacrifice and not an enlargement of freedom.

Negative freedom as defined by Berlin is a rather recent accomplishment. "This is liberty as it has been conceived by liberals in the modern world from the days of Erasmus (some would say Occam) to our own. Every plea for civil liberties and individual rights, every protest against exploitation and humiliation, against the encroachment of public authority, or the mass hypnosis of custom or organized propaganda, springs from this individualistic, and much disputed, conception of man" (Berlin, 1969: 127f.). Berlin weaves together liberty, liberalism, and individualism and notes three facts about their relationships. First, while truth and character may happen to flourish particularly well under a political regime of liberty, this is a contingent and not a necessary truth. For Berlin (1969: 128, footnote 1), the common contrary belief is "but another illustration of the natural tendency of all but a very few thinkers to believe that all the things they hold good must be intimately connected or at least compatible, with one another." A negatively free human is equally free to pursue truth or error.

Secondly, this doctrine is rather modern. For the ancient world of the Greeks and Romans, the idea of individual liberty was as unfamiliar as it was to all the other ancient civilizations. Freedom itself is a comparatively recent accomplishment of the Renaissance and Reformation movements.

Thirdly, this doctrine could well be conceived to be compatible with some sort of autocracy and does not automatically entail a political democracy. "Just as a democracy may, in fact, deprive the individual citizen of a great many liberties which he might have in some other form of society, so it is perfectly conceivable that a liberal-minded despot would allow his subjects a large measure of personal freedom" (Berlin, 1969: 129). Indeed, judged historically, benevolent despots are no rarer than democratic demagoguery.

Since Berlin's understanding of the negative side of freedom is so natural and self-evident to him, he spends much more time spelling out the positive side. Positive liberty refers to self-mastery. "I wish to be a subject, not an object; to be moved by reasons, by conscious purposes, not by causes which affect me, as it were, from the outside. I wish to be somebody, not nobody; a doer – deciding, not being decided for, self-directed and not acted upon by external nature or by other men as if I were a thing, or an animal, or a slave incapable of playing a human role, that is, of conceiving goals and policies of my own and realizing them" (Berlin, 1969: 131).

Put together, the negative and positive aspect of freedom, noninterference and self-mastery, seem to be but two sides of the same coin. 'Freedom from' is the defensive; 'freedom to' is the offensive. Despite this apparent inner relationship, Berlin notes a deep internal ambivalence in these two parts of liberty, which tend to clash historically in the 'magical transformation' of freedom into oppression. How can this be explained? The goal of self-mastery is an ideal accomplished by only a fraction of mankind. If leaders step up with a mission to free mankind as a whole, to bring out the better selves of man, the script for liberation is written. Paradoxically enough, however, this process of liberation often turns out to be a monstrous act of oppression: "Once I take this view, I am in a position to ignore the actual wishes of men or societies, to bully, oppress, torture them in the name, and on behalf, of their 'real' selves, in the secure knowledge that whatever is the true goal of man (happiness, performance of duty, wisdom, a just society, self-fulfillment) must be identical with his freedom – the free choice of his 'true,' albeit often submerged and inarticulate, self" (Berlin, 1969: 133).

Berlin analyzes this paradoxical formula, 'liberation as oppression,' in two forms: negatively, as a form of self-abnegation that he calls "the retreat to the inner citadel"; and positively, as 'self-realization' proper in the pursuit of an ideal. The negative way of self-liberation is to shed desires one cherishes yet is unable to fulfill in the environment of one's life. For instance, in an autocratic regime it might be 'wise' to restrict one's own 'free speech.' To shake off the needs that one cannot realize seems a commendable avenue to gain inner peace and serenity. "This is the traditional self-emancipation of ascetics and quietists, of stoics or Buddhist sages, men of various

religions or of none, who have fled the world, and escaped the yoke of society or public opinion, by some process of deliberate self-transformation that enables them to care no longer for its values, to remain, isolated and on its edges, no longer vulnerable to its weapons" (Berlin, 1969: 135f.). This form of 'inner emigration' is typical for the individual and social life under harsh political regimes and in hard times. "Ascetic self-denial may be a source of integrity or serenity and spiritual strength, but it is difficult to see how it can be called an enlargement of liberty" (Berlin, 1969: 140).

The positive form of self-realization relies upon critical reason. The liberating force of knowledge teaches us to distinguish between necessity and contingency. The laws of nature are the case in point. Whoever obeys these laws is not a slave, but, given intelligence and creativity, may use these laws for a mastery of nature. What holds true for nature may equally be valid for society. An enlightened society rests upon the design of institutions no rational person can resist. To obey rational rules and regulations is the only true practice of freedom. "That is the programme of enlightened rationalism from Spinoza to the latest (at times unconscious) disciples of Hegel. *Sapere aude*. What you know, that of which you understand the necessity – the rational necessity – you cannot, while remaining rational, want to be otherwise. For to want something to be other than what it must be is, given the premises – the necessities that govern the world – to be *pro tanto* either ignorant or irrational. Passions, prejudices, fears, neuroses, spring from ignorance, and take the form of myths and illusions" (Berlin, 1969: 142).

If this project of self-mastery is transferred from the individual level to the social or 'organic' level, Berlin says we are entering 'The Temple of Sarastro.' In Mozart's famous opera 'The Magic Flute,' Sarastro's Temple is the embodiment of light, goodness, kindness, decency, and an appropriate conduct of life. Based upon the idea of 'scientism,' all political and moral questions are, in principle, soluble and hence there will be one rational solution. "On this assumption, the problem of political liberty was soluble by establishing a just order that would give to each man all the freedom to which a rational being was entitled" (Berlin, 1969: 145). If there is an objective reason established in a just social and political order, the law becomes an embodiment of liberty. To abide by the law, then, is to be free. Sarastro's Temple turns out to be an example for benign despotism.

But even that is not the end of the story. Berlin discusses recognition and sovereignty as two forces that enter the discourse on liberty. The "great cry for recognition" (Berlin, 1969: 157) has been called 'social freedom.' Is it therefore a third form beyond negative and positive liberty? No. Berlin sees the enormous meaning of mutual recognition in times of decolonization and the movement for independence of formerly colonized states. In his eyes, much in line with his dictum that 'everything is what it is,' recognition is not a question of freedom but a matter of justice, solidarity, and integration. Berlin accepts this need to belong as a very important ingredient in the modern world but rejects the term 'social freedom' as a misnomer.

A similar case can be made with respect to sovereignty. Berlin (1969: 162) reads the French Revolution as "an eruption of the desire for 'positive' freedom of collective self-direction." But even in a Republican democracy the doctrine of absolute sovereignty turned out to be an oppressive new form of benign despotism. The freeing of 'the people' impaired the freedom of ordinary people. This is why for Benjamin Constant, Rousseau was an enemy of freedom despite his republican outlook. For liberals like Constant, Mill, and Tocqueville, democracy can foster freedom only if it ensures that no power but only rights are regarded as absolute and that there are frontiers guarding individual privacy.

"One belief, more than any other, is responsible for the slaughter of individuals on the altars of the great historical ideals – justice or progress or the happiness of future generations, or the sacred mission or emancipation of a nation or race or class, or even liberty itself, which demands the sacrifice of individuals for the freedom of society" (Berlin, 1969: 167). This is the belief in "the notion of a final harmony" (Berlin, 1969: 168). Yet, this idea of a 'final harmony' is a grand illusion. It assumes that all values are compatible and can be fulfilled simultaneously. In fact, Berlin holds, values chronically conflict. We live in a world of 'trade-offs' and cannot have it all and at once. Therefore, choices have to be made that are hard rather than harmonious, conflictual rather than consensual, tragic rather than conciliatory. In his eyes this conclusion "cannot be escaped by those who, with Kant, have learnt the truth that out of the crooked timber of humanity no straight thing was ever made" (Berlin, 1969: 170).

The History of Ideas

Upon first sight, the multifaceted preoccupation of Isaiah Berlin's essays in the history of ideas gives no clue of a systematic approach. His disdain for grand theory or systems, his preference for papers and not books, the wide array of thinkers he dealt with, his own modesty saying he and his work had always been overestimated, all those peculiarities of his oeuvre give the reader the impression that this is the work of an intellectual thinker with idle curiosity. In addition, his elegant style, his conversational tone, the character of his narratives, and his parsimony of conceptual discussion – if compared to the complexity and difficulty of Hegel's or Marx's thought – seem to exhibit the traits of an essayist. What could be the underlying thread in his analysis of Vico, Machiavelli, Herder, Hamann, de Maistre, Marx, Herzen, Belinski, and many others? Apparently, there is no thread.

Yet this impression may prove wrong because it is possible to reconstruct the ensemble of unit-ideas Berlin worked with. Upon closer look, he seeks to understand the three-dimensional constellation of modernity: Historically, he tries to explain the experience of revolution, war, and barbarism in the twentieth century by tracing its intellectual *longue durée* in the eighteenth and nineteenth century; politically, he studies the phenomenon of totalitarianism; philosophically, he seeks to gauge the chances and risks of a critical liberalism with its dual components of liberty and pluralism. Taken together, these three dimensions of the history of ideas, the political problem and the philosophical questions define Berlin's underlying research program.

According to Berlin, the Western discourses in science and ethics may be characterized by three underlying assumptions that have ruled them from Plato to Kant to Hegel and Marx. First, to all genuine questions there is one and only one true answer. Differing answers are deviations from the truth and hence false. This pattern holds true for science and ethics, pertains to scholarly problems as well as to the question of how to lead a good life. Second, these true answers, in principle, can be known even if we do not know them yet. Third, once found the true answers are compatible with one another since true propositions cannot conflict with each other. The realm of true answers forms an organic whole, a totality, a system based upon a logical order devoid of contradictions.

In its history of more than 2000 years, this model was more and more undermined by three rifts. These rifts are turning points in the history of political thought that bring about a revaluation of values in its wake. After such an event, nothing remains the same and a new frame of reference begins to emerge. The first rift in the foundation of this model, according to Berlin (1996: 168–170), came into being by the emergence of Stoicism after the death of Aristotle. Instead of understanding the individual as a public person, it is conceived as an inner being. Instead of conceptualizing the social politically, it is interpreted as ethical. Instead of dealing with a republican social order, anarchy reaches center stage.

The second rift is due to Machiavelli. For Benedetto Croce, Machiavelli's patrimony is, above all, the crucial distinction between immoral politics and moral culture. In the wake of Machiavelli's legacy, there is the talk of 'Realpolitik,' violence as the *raison d'être* of the State and, in times of crisis, the suspension of the institute of *habeas corpus*. According to Berlin, this famous interpretation of Croce – the antinomy between politics and ethics – puts forward 'a false antithesis.' What Machiavelli achieved is not just "the emancipation of politics from ethics"; "what he institutes is something that cuts deeper still – a differentiation between two incompatible ideals of life, and therefore two moralities" (Berlin, 1998: 289). He confronts the two moral worlds of Christian and Pagan morality and claims their irreconcilability. "One can save one's soul, or one can found or maintain or serve a great and glorious State; but not always both at once" (Berlin, 1998: 294).

This insight triggered such a revolution in moral reasoning that the distinction was later reiterated in different forms: Max Weber (1958) separates the 'ethics of responsibility' from the 'ethics of conviction.' Berlin prefers to distinguish between a 'morality of consequence' and a 'morality of motive.' In the-first case, an ethic or morality reckons with the consequences of one's action; in the second case – as, for instance, a true Christian or a revolutionary would believe – the consequences are up to God or society since values are pursued for their own sake no matter what their consequences are. This central development replaced the monism of the old model of Western rationalism by a dualism that is irreducible, incompatible, and incommensurable. In Berlin's eyes, the discovery of plural moral worlds put the nail into the coffin of the 'great chain of being.'

The third rift was brought about by romanticism. Studying Vico, Hamann, and Herder, Berlin (1990, 1993, 1996, 2002a, 2006, 2013b) isolated three path-breaking movements with another revaluation of values in its wake: Populism underlines the belief in the value of belonging to a group, class or nation; expressivism holds that human action in general and art in particular express the personality of a person or group; pluralism relates to a plethora of values and its incompatibility. No culture is alike; all cultures are different. This genuine heterogeneity cannot be transformed into inequality or hierarchy in the sense that one culture could be regarded as better, higher, or more valuable than another. According to Berlin, Romanticism thus also tackles the Enlightenment idea that nature binds together truth, happiness, and virtue. Once the diversity and incompatibility of values and ideals, ideas, and objectives have penetrated into human consciousness, scientifically grounded concepts of the perfect human being or the perfect social order are no longer possible. To be sure, human beings will always strive for values and ideals in order to improve their lives or the social order they live in. But these attempts at betterment of self and society have to do without the backing of scientific truths or a scientifically certified morality.

Conclusion

In his study of 'Russian Thinkers,' Berlin cited Archilochus' famous line: "The fox knows many things, but the hedgehog knows one big thing" (Berlin, 1978a: 22). Berlin was a hedgehog who championed the idea of liberty. He did so in three respects: as an ideal, a value, and a yardstick; as an ideal that he himself held but that had to be interpreted in order to find out its meaning and snares; as a value that is only one among many competing and conflicting values and whose status and standing has to be checked; and as a yardstick for his history of ideas because the preeminent question for Berlin was always how a complex of ideas fares with liberty. But Berlin was also a fox knowing many things. The concept of liberty was embedded in a critical liberalism and a value pluralism ready to acknowledge the complexity of social and individual life. Berlin was very sensitive to the imminent dangers of totalitarian rule that sprang from ideological attempts to realize a collective vision of the good life, be it communism or fascism. His ardent plea for liberty did not fare well with his concept of value pluralism. How could the universal value of freedom be reconciled with the relativism of value pluralism? People free to choose might prefer equality or justice over liberty, which often enough only means freedom and power for the elites. People might also prefer the sense of belonging over freedom. As a versatile fox, Isaiah Berlin knew all that. But his critical liberalism with its aversion to theoretical system building could not and would not ultimately reconcile these inconsistencies.

See also: Attitudes, Political: Authoritarianism and Tolerance; Civil Liberties and Human Rights; Democracy: Normative Theory; Enlightenment; Ethics and Values; Freedom: Political; Historicism; Individuals in Society: History of the Concept; Kant, Immanuel (1724–1804); Liberalism: Political Doctrine and Impact on Social Science; Marx, Karl (1818–83); Objectivity, History of; Pluralism and Tolerance; Political Theory, History of; Value Pluralism.

Bibliography

Baum, B., Nichols, R. (Eds.), 2013. Isaiah Berlin and the Politics of Freedom: 'Two Concepts of Liberty' 50 Years Later. Routledge, New York and London.

Berlin, I., 1969. Four Essays on Liberty. Oxford University Press, Oxford.

Berlin, I., 1978a. Russian Thinkers. Penguin, London.

Berlin, I., 1978b. Concepts and Categories: Philosophical Essays. Hogarth Press, London.

Berlin, I., 1979. Against the Current: Essays in the History of Ideas. Hogarth Press, London.

Berlin, I., 1990. The Crooked Timber of Humanity: Chapters in the History of Ideas. John Murray, London.

Berlin, I., 1993. The Magus of the North. J.G. Hamann and the Origins of Modern Irrationalism. John Murray, London.

Berlin, I., 1995. Karl Marx: His Life and Environment. Princeton University Press, Princeton.

Berlin, I., 1996. The Sense of Reality: Studies in Ideas and Their History. Chatto and Windus, London.

Berlin, I., 1997. The Proper Study of Mankind: An Anthology of Essays. Chatto and Windus, London.

Berlin, I., 1998. Personal Impressions. Pimlico, London.

Berlin, I., 1999a. The First and the Last. New York Review Books, New York.

Berlin, I., 1999b. The Roots of Romanticism. Chatto and Windus, London.

Berlin, I., 2000. The Power of Ideas. Chatto and Windus, London.

Berlin, I., 2002a. Freedom and Its Betrayal: Six Enemies of Human Liberty. Chatto and Windus, London.

Berlin, I., 2002b. Liberty. Oxford University Press, Oxford and New York.

Berlin, I., 2004. Flourishing: Letters 1928–1946. Chatto and Windus, London.

Berlin, I., 2006. Political Ideas in the Romantic Age: Their Rise and Influence on Modern Thought. Chatto and Windus, London.

Berlin, I., 2009. Enlightening: Letters 1946–1960. Chatto and Windus, London.

Berlin, I., 2013a. Building: Letters 1960–1975. Chatto and Windus, London.

Berlin, I., 2013b. Three Critics of the Enlightenment: Vico, Hamann, Herder (1960–65). Princeton University Press, Princeton.

Berlin, I., Lukes, S., 1998. Isaiah Berlin in conversation with Steven Lukes. Salmagundi 120, 52–134.

Cherniss, J.L., 2013. A Mind and Its Time: The Development of Isaiah Berlin's Political Thought. Oxford University Press, Oxford.

Cherniss, J.L., Hardy, H., 2013. Isaiah Berlin. Stanford. Encyclopedia online.

Crowder, G., 2004. Isaiah Berlin: Liberty and Pluralism. Polity, Cambridge.

Crowder, G., Hardy, H. (Eds.), 2007. The One and the Many: Reading Isaiah Berlin. Prometheus, Amherst.

Dubnov, A.M., 2012. Isaiah Berlin: The Journey of a Jewish Liberal. Palgrave Macmilian, London.

Galipeau, C.J., 1994. Isaiah Berlin's Liberalism. Clarendon Press, Oxford.

Gray, J., 1995. Isaiah Berlin. Princeton University Press, Princeton.

Gray, J., 2002. Two Faces of Liberalism. Polity, Cambridge.

Hardy, H. (Ed.), 2009. The Book of Isaiah: Personal Impressions of Isaiah Berlin. Boydell, Woodbridge.

Hausheer, R., 1997. Introduction. In: Berlin, I. (Ed.), The Proper Study of Mankind: An Anthology of Essays. Chatto and Windus, London, pp. xxiii–xxxvi.

Honneth, A., 1999. Negative freedom and cultural belonging: an unhealthy tension in the political philosophy of Isaiah Berlin. Social Research. An International Quarterly 66, 1063–1077.

Ignatieff, M., 1998. Isaiah Berlin: A Life. Metropolitan, New York.

Jahanbegloo, R., 1992. Conversations with Isaiah Berlin. Peter Halban, London.

Lilla, M., Dworkin, R., Silvers, R.B. (Eds.), 2001. The Legacy of Isaiah Berlin. New York Review Books, New York.

Lovejoy, A.O., 1942. The Great Chain of Being. Harvard University Press, Cambridge MA.

Lukes, S., 1994. The singular and the plural: on the distinctive liberalism of Isaiah Berlin. Social Research. An International Quarterly 61, 687–718.

Margalit, E.,A. (Ed.), 1991. Isaiah Berlin: A Celebration. Hogarth Press, London. University of Chicago Press, Chicago.

Müller, H.-P., 1999. Romantischer Realismus. Isaiah Berlins Botschaft. Merkur 599, 157–162.

Ryan, A. (Ed.), 1979. The Idea of Freedom: Essays in Honour of Isaiah Berlin. Oxford University Press, Oxford.

Walicki, A., 2011. Encounters with Isaiah Berlin. Story of an Intellectual Friendship. Peter Lang, Frankfurt.

Weber, M., 1958. Politik als Beruf. In: Gesammelte Politische Schriften. Mohr-Siebeck, Tübingen, pp. 494–548.

Best Worst Scaling: Theory and Practice

AAJ Marley, University of Victoria, Victoria, BC, Canada; and University of Technology Sydney, Sydney, NSW, Australia
Terry N Flynn, University of Technology Sydney, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Best worst scaling (BWS) can be a method of data collection, and/or a theory of how respondents provide top and bottom ranked items from a list. The three 'cases' of BWS are described, followed by a summary of the main models and related theoretical results, including an exposition of possible theoretical relationships between estimates from two of the cases. This is followed by the theoretical and empirical properties of 'best minus worst scores.' The entry ends with some directions for future research.

Introduction

Louviere and Woodworth (1990) and Finn and Louviere (1992) developed a discrete choice task in which a person is asked to indicate the 'least preferred' item in a choice set, in addition to indicating the (traditional) 'most preferred' item; the approach that Louviere pioneered is now called best worst scaling (BWS). Louviere's initial focus was on 'objects,' such as attitudes, general public policy goals, brands, or anything that did not require a detailed description, such as would be required for a consumer product such as a soft drink or a car. Thus, Finn and Louviere (1992) used BWS to examine the degree of concern the general public had for each of a set of food safety goals, including irradiation of foods and pesticide use on crops. Figure 1 contains a BWS question similar to ones used in that study.

Almost immediately, Louviere began applying BWS to more complex items such as attribute levels describing a single alternative (profile), or complete alternatives (profiles) of the type familiar to choice modelers. The former case, requiring respondents to identify the best attribute level and worst attribute level within an alternative, was relatively unfamiliar to choice modelers, whereas the latter case is a 'natural' extension of a standard discrete choice experiment (DCE) in which the decision maker selects the most preferred item (Louviere et al., 2000; Hensher et al., 2005).

BWS was initially used mainly as a method of collecting data in a cost-efficient manner, though Finn and Louviere (1992) did introduce the maximum difference (maxdiff) model for best–worst choice (see below). After a 10-year delay, Marley and Louviere (2005) completed an extensive analysis of BWS as a theory, explaining the processes that individuals might follow in providing best and worst data, and presenting plausible mathematical forms for such processes; summaries of those, and more recent, results are presented in Section 3. Louviere has recently returned to the use of BWS as a method of data collection (Louviere et al., 2008), in which case its main purpose, via repeated rounds of best–worst choices, is to obtain a full ranking of items in a manner that is 'easy' for respondents and can then be analyzed in various ways.

The remainder of the entry is as follows. Section 2 illustrates the three types ('cases') of BWS that Louviere developed. Section 3 summarizes the basic models, with Section 3.3 summarizing models of ranking by repeated best and/or worst choices; Section 4 presents properties of simple 'scores'; and Section 5 summarizes and discusses future research.

BWS: the Three Cases

Case 1 (Object Case)

Case 1 BWS (as already illustrated in Figure 1) is appropriate when the researcher is interested in the relative values associated with each of a list of objects. These might be brands, public policy goals, or any set of objects that can be meaningfully compared. Once the researcher has chosen the list of objects, (s)he must present choice sets of these to respondents to obtain best and worst data. Choice sets here serve a similar purpose to those in traditional DCEs: statistical designs are implemented that include (some or all) subsets of all possible items which, with suitable assumptions, facilitate

Most	Issue	Least
	Pesticides used on crops	
	Hormones given to livestock	
	Irradiation of foods	✓
	Excess salt, fat cholesterol	
✓	Antibiotics given to livestock	

Please consider the food safety issues in the table above and tick which concerns you most and which concerns you least.

Figure 1 A completed example BWS 'object case' question. Reproduced with permission from **Figure 1** of Flynn, Valuing citizen and patient preferences in health: recent developments in three types of best–worst scaling, *Expert Review of Pharmacoeconomics & Outcomes Research*, 10, 2010, 259–267.

Most	Appointment #1	Least
	You will have to wait two months for your appointment	
	The specialist has been treating skin complaints part-time for 1-2 years	✓
	Getting to your appointment will be quick and easy	
✓	The consultation will be as thorough as you would like	

Please imagine being offered the appointment described above and tick which feature would be best and which would be worst.

Figure 2 A completed example BWS 'profile case' question.

inferences about the value associated with the wider list of objects.

Case 2 (Profile Case)

Case 2 BWS, which was introduced by McIntosh and Louviere (2002), is illustrated in Figure 2 using a question from a dermatology study (Coast et al., 2006; Flynn et al., 2008).

The choice set has the structure of a single profile (multiattribute alternative, see Case 3). However, the respondent is not required to consider the value of the profile as a whole, but has to consider the attribute levels that describe it, and choose the one that is best (most attractive) and the one that is worst (least attractive). Case 2 BWS is popular in health because health care goods/services can be complicated and even pairs of specifications (in a simple DCE) may lead to unacceptable cognitive burden, particularly among vulnerable patient groups. Also, recent theoretical results show that a Case 2 study, in combination with a Case 3 study of choices between the same (or suitably related) profiles, may allow the separate measurement of attribute weight and level scale value (Flynn and Marley, 2012); should such a study be successful, it would solve this classic measurement problem (Anderson, 1970; Marley et al., 2008).

Case 3 (Multiprofile Case)

Case 3 BWS requires a respondent to choose the worst (least attractive) profile as well as the best (most attractive) one in the current choice set. Figure 3 shows an example question from a mobile phone study (Marley and Pihlens, 2012). The additional data provided by worst choices are valuable in many marketing applications, both in terms of better estimation of parameters and in terms of identifying attribute levels that make a product 'unacceptable.'

	Phone 1	Phone 2	Phone 3	Phone 4
Phone Style				
	Clam or flip phone	Candy Bar or straight phone	Swivel flip	PDA phone with touch screen input
Handset Brand	A	B	C	D
Price	$49.00	$199.00	$249.00	$129.00
Built-in Camera	No camera	5 megapixel camera	2 megapixel camera	3 megapixel camera
Wireless Connectivity	No Bluetooth or WiFi connectivity	Bluetooth and WiFi connectivity	WiFi connectivity	Bluetooth connectivity
Video Capability	No video recording	Video recording (up to 1 hour)	Video recording (more than 1 hour)	Video recording (up to 15 minutes)
Internet Capability	Internet Access	Internet Access	No Internet access	No Internet access
Music Capability	No music capability	MP3 Music Player only	FM Radio only	MP3 Music Player and FM Radio
Handset Memory	64 MB built-in memory	2 GB built-in memory	512 MB built-in memory	4 GB built-in memory

Figure 3 An example BWS 'multiprofile case' question. Reproduced with permission from Figure 1 of Marley and Pihlens, Models of best–worst choice and ranking among multiattribute options (profiles), *Journal of Mathematical Psychology*, 56, 2012, 24–34.

Models of Best and/or Worst Choice

We use notation that applies to all Cases (1, 2, and 3) and talk of 'choice options' (or 'options') without distinguishing between objects (Case 1), attribute levels of a profile (Case 2), and profiles (Case 3). We also present the results in terms of a numeric 'utility' value associated with each choice option (and, as relevant, with each of its attribute levels), rather than in terms of the utility coefficients ('beta weights') that are standard in the discrete choice literature (Louviere et al., 2000); we do this because various theoretical results on BWS can only be stated and proved in the former notation (Marley and Louviere, 2005; Marley et al., 2008; Marley and Pihlens, 2012).

Notation

Let P with $|P| \geq 2$ denote the finite set of potentially available choice options, and let $D(P)$ denote the *design*, i.e., the set of (sub) sets of choice alternatives that occur in the study. For example, participants might be asked about their preferences for cars by repeatedly asking them for choices amongst sets of four different options: P represents the collection of car types in the study, and each element of the set $D(P)$ represents the set of options provided on one particular choice occasion. For any $Y \in D(P)$, with $|Y| \geq 2$, $B_Y(y)$ denotes the probability that alternative y is chosen as best in Y, $W_Y(y)$ the probability that alternative y is chosen as worst in Y, and $BW_Y(x,y)$ the probability that alternative x is chosen as best in Y and the alternative $y \neq x$ is chosen as worst in Y.

Multinomial Logit Models of Best and/or Worst Choice

Many models of choice are based on extensions of the *multinomial logit* (*MNL*) *model*. The *best choice MNL model* assumes there is a difference scale u such that for all $y \in Y \in D(P)$,

$$B_Y(y) = \frac{e^{u(y)}}{\sum_{z \in Y} e^{u(z)}}. \quad [1]$$

The value $u(y)$ for an option y is interpreted as the utility for that option.

The parallel *worst choice MNL model* assumes there is a difference scale v such that for all $y \in Y \in D(P)$,

$$W_Y(y) = \frac{e^{v(y)}}{\sum_{z \in Y} e^{v(z)}}.$$

Marley and Louviere (2005) present a theoretical argument for the case where $v = -u$, i.e., we have

$$W_Y(y) = \frac{e^{-u(y)}}{\sum_{z \in Y} e^{-u(z)}}. \quad [2]$$

Note that eqn [1] with [2] implies that the probability that $y \in Y$ is selected as *best* in a set Y with scale values $u(z)$, $z \in Y$ is equal to the probability that $y \in Y$ is selected as *worst* in a set Y with scale values $-u(z)$, $z \in Y$.

The MNL models for best choices and for worst choices can be combined in three simple ways to model best–worst choice:

First, assuming the best choices satisfy eqn [1] and the worst choices satisfy eqn [2], the *best, then worst, MNL model* assumes that for all $x, y \in Y \in D(P)$, $x \neq y$,

$$BW_Y(x,y) = B_Y(x) W_{Y-\{x\}}(y),$$

Second, the *worst, then best, MNL model* assumes that for all $x, y \in Y \in D(P)$, $x \neq y$,

$$BW_Y(x,y) = W_Y(y) B_{Y-\{y\}}(x).$$

Repeated choices satisfying a common one of the above models lead naturally to models of rank order data (Section 3.3).

Third, the *maxdiff model for best–worst choice* assumes that the value of a choice alternative in the selection of a best option (u) is the negative of the value of that option in the selection of a worst option ($-u$), and that for all $x, y \in Y \in D(P)$, $x \neq y$,

$$BW_Y(x,y) = \frac{e^{[u(x)-u(y)]}}{\sum_{\substack{\{p,q\} \in Y \\ p \neq q}} e^{[u(p)-u(q)]}}. \quad [3]$$

Models of Ranking by Repeated Best and/or Worst Choice

Thus far, we have discussed models for the choice of the best and the worst option in a set of available options. To the extent that those choices are reliable, it is of interest to consider the use of repeated choices of the best and/or worst option to generate a rank order of the available options, and to develop models for such rankings. For four options, there are eight distinct patterns of best and/or worst questions, any one of which an experimenter can ask a participant to use for the generation of a rank order on those options; and, also, at least eight possible (MNL-based) models of best and/or worst choice for the final rank order. Here, we introduce notation for two such models, illustrated with four-element sets.

Let Y denote a typical choice set; $R(Y)$ the set of rank orders of Y; $\rho \in R(Y)$ a typical rank order (from best to worst) of Y; and $p_{R(Y)}(\rho)$ the probability of that rank order occurring. Let r, s be two indices, each of which can be fixed (throughout a rank order) at the value b (for ***b***est) or w (for ***w***orst). Let $P^{r,s}_{R(Y)}(\rho_1\rho_2\rho_3\rho_4)$ denote the probability of the rank order $\rho = \rho_1\rho_2\rho_3\rho_4$ (from best to worst) when the rank order is obtained by: the first choice is of type r; the second of type s; the third, again, of type r. Then we can have:

1. repeated best:

$$P^{b,b}_{R(Y)}(\rho) = B_Y(\rho_1) B_{Y-\{\rho_1\}}(\rho_2) B_{\{\rho_3,\rho_4\}}(\rho_3). \quad [4]$$

2. repeated best, then worst:

$$P^{b,w}_{R(Y)}(\rho) = B_Y(\rho_1) W_{Y-\{\rho_1\}}(\rho_4) B_{\{\rho_2,\rho_3\}}(\rho_2). \quad [5]$$

The natural first assumption in testing these models is to assume that the best (respectively, worst) choice probabilities satisfy the MNL model (1) (respectively, (2)), and, as needed

by data, generalizations of those models that include a scale factor that depends on the current choice set in some identifiable manner; this scale factor relates to the possibly changing variability ('consistency') of the choices across sets. An example of such a generalization of the MNL model for best choices, (1) is: there is a nonnegative scale factor s defined for each integer 2,3,…, and a difference scale u such that for all $y \in Y$,

$$B_Y(y) = \frac{e^{s(|Y|)u(y)}}{\sum_{z \in Y} e^{s(|Y|)u(z)}}. \quad [6]$$

We then define a *generalized rank ordered logit model* as a set of rank orders that satisfy eqn [4] with the best choice probabilities satisfying eqn [6]. The form [6] is a special case of what Vermunt and Magidson (2005, Section 2.4) call an MNL with *replication-specific scale factor* and of Fiebig et al. 's (2010) *generalized multinomial logit model*; the latter model also includes scale (variance) heterogeneity across individuals.

Scarpa et al. (2011) collected ranking data by repeated best, then worst, choices and fit that data quite successfully with a model based on repeated best choice, i.e., eqn [4], with those choices satisfying a generalization of the model in eqn [6] that included properties of the scale factor s on aspects of the design that took account of the difference between the data collection method (repeated best, then worst) and the model (repeated best); it would be interesting to see if their data could be better fit by a model that matched their data collection method, i.e., eqn [5]. Collins and Rose (2011) fit related models to stated preference data on dating choices.

Maximum Random Utility Models of Choice and Response Time

When treated as a single model, the three models [1]–[3], satisfy an *inverse extreme value maximum random utility model* (Marley and Louviere, 2005; Def. 11). That is, for $z \in P$ and $p, q \in P$, $p \neq q$, there are independent random variables ε_z, $\varepsilon_{p,q}$ with the extreme value distribution. (This means that: for $-\infty < t < \infty$ $\Pr(\boldsymbol{\varepsilon}_z \leq t) = \exp - e^{-t}$ and $\Pr(\boldsymbol{\varepsilon}_{p,q} \leq t) = \exp - e^{-t}$.) such that for all $y \in Y \in D(P)$,

$$B_Y(y) = \Pr(u(y) + \varepsilon_y = \max_{z \in Y}[u(z) + \varepsilon_z]), \quad [7]$$

$$W_Y(y) = \Pr(-u(y) + \varepsilon_y = \max_{z \in Y}[-u(z) + \varepsilon_z]), \quad [8]$$

and for all $x, y \in Y \in D(P)$, $x \neq y$,

$$BW_Y(x, y) = \Pr\left(u(x) - u(y) + \varepsilon_{x,y} = \max_{\substack{p,q \in Y \\ p \neq q}} \left[u(p) - u(q) + \varepsilon_{p,q}\right]\right). \quad [9]$$

Standard results (summarized by Marley and Louviere, 2005) show that the expression for the choice probabilities given by eqn [7] (respectively, eqns [8] and [9]) agrees with that given by eqn [1] (respectively, eqns [2] and [3]).

These random utility models are particularly interesting in the present context because they can be rewritten, and extended, in such a way as to also predict response time (Hawkins et al., 2012).

Extensions of the Models to Case 2 and Case 3

The notation, models, and results (such as for scores, Section 4) are easily extended to Case 2 and Case 3 (Marley et al. (2008) for Case 2, Marley and Pihlens (2012) for Case 3, and Flynn and Marley (2012) for a summary of all cases).

Properties and Uses of Scores for the Maxdiff Model

We now summarize theoretical and empirical results for best minus worst scores (defined below) for the maxdiff model of best–worst choice. Although the theoretical results are not exact for, say, the best, then worst, MNL model, to date no identifiable differences have been found between fits of that model and the maxdiff model (see Flynn et al., 2008; for an example of such fits); hence the score measures are useful for preliminary analyses of the data independent of the model that is eventually fit to the data.

For each option x in the design, the *score* for x (in this particular design) is the number of times option x is chosen as best in the study minus the number of times option x is chosen as worst in the study; we call refer to 'the scores' for these values across the options in the design.

Scores: Theoretical Property 1 (for Cases 1, 2, and 3)

Marley and Islam (2012) state the following terms and results exactly. Assume that one is interested in the rank order, only, of the (utility) scale values in the maxdiff model. An *acceptable loss function* is a 'penalty' function with a value that remains constant under a common permutation of the scores and the scale values, and that increases if the ranking is made worse by misordering a pair of scale values. Let P be a master set with $n \geq 2$ elements and assume that, for some k with $n \geq k \geq 2$, every subset of P with exactly k elements appears in the design (Further work is needed to extend the theoretical result to, say, balanced incomplete block designs. See Marley and Pihlens (2012) for related discussions of *connected designs*.) $D(P)$. Then, given the maxdiff model, ranking the scale values in descending order of the (best minus worst) scores, breaking ties at random, has 'minimal average loss' amongst all ('permutation invariant') ranking procedures that depend on the data only through the set of scores.

The above result actually holds for the class of *weighted utility ranking models*, which includes the MNL for best; MNL for worst; and the maxdiff model for best–worst choice (Marley and Islam, 2012).

Scores: Theoretical Property 2 (for Cases 1, 2 and 3)

The set of (best minus worst) scores is a sufficient statistic. As with the first theoretical property, this one holds for the class of *weighted utility ranking models* (Marley and Islam, 2012; Theorem 3).

Scores: Empirical Properties (for Cases 1, 2 and 3)

The best minus worst scores have been found to be linearly related to the maximum likelihood estimates of the best, then worst, MNL model, and of the maxdiff model, in virtually every empirical study to date. For example, Flynn (2010) obtained this result in a Case 2 quality of life study, and Marley and Islam (2012) present such linear relations for both profiles and attribute levels in a Case 3 study of

consumer attitudes to the installation of solar technology for household electricity production. Such linearity is probably a manifestation of the linear portion of the logistic (cumulative distribution) function; thus, a nonlinear relationship is likely only when the researcher is plotting the scores for a single highly consistent respondent, or for a sample of respondents each of whom is highly consistent and the choices are highly consistent across the sample.

The scores also enable considerable insights to be drawn at the level of the individual respondent. For example, taxonomic (clustering) methods of analysis have been applied to the scores (Auger et al., 2007) and Flynn and colleagues have used the scores to decide which solutions from latent class analyses 'make sense' (Flynn et al., 2010).

Summary and Future Research

Further research is needed to understand the factors affecting the scale factor at different choice depths and the conditions under which the data require different functional forms (models) for best and worst choices. Also, research to date suggests that the class of models with natural process interpretations is different from the class of models with useful score properties. Nonetheless, the scores may give the average applied researcher more confidence, not least in terms of better understanding heterogeneity in preferences and/or scale factors in their data.

Data pooling has normative issues that are particularly pertinent to health economists: for best–worst choice, it is only if all individuals satisfy the maxdiff model that the average of their utility estimates represents the preferences of the 'representative individual' used in economic evaluation. Also, appropriate data pooling exercises (likely involving Case 2 and Case 3 data) may allow the separation of attribute weight and level scale value (Marley et al., 2008; Flynn and Marley, 2012). While the distinction between these two measures is not recognized by traditional economic welfare theory, it offers benefits to health researchers constructing multidimensional health outcome instruments from individual symptom scales: for example, rejection of a simple 'sum-score' aggregation rule necessitates decisions as to what importance weights are to be applied to the individual symptom scales (Fayers and Machin, 2007, p. 218).

See also: Decision and Choice: Luce's Choice Axiom; Decision and Choice: Random Utility Models of Choice and Response Time; Experimental Laboratories: Psychological; Experimental Laboratories: Social and Economic; Measurement Theory: Conjoint Analysis Applications; Panel Surveys: Uses and Applications; Psychometrics: Preference Models with Latent Variables.

Bibliography

Anderson, N.H., 1970. Functional measurement and psychophysical judgement. Psychological Review 77 (3), 153–170.

Auger, P., Devinney, T.M., Louviere, J.J., 2007. Using best-worst scaling methodology to investigate consumer ethical beliefs across countries. Journal of Business Ethics 70, 299–326.

Coast, J., Salisbury, C., de Berker, D., Noble, A., Horrocks, S., Peters, T.J., Flynn, T.N., 2006. Preferences for aspects of a dermatology consultation. British Journal of Dermatology 155, 387–392.

Collins, A.T., & Rose, J.M. (2011). Estimation of a stochastic scale with best-worst data. Paper presented at the Second International Choice Modelling Conference, Leeds, UK.

Fayers, P.M., Machin, D., 2007. Quality of Life: The Assessment, Analysis and Interpretation of Patient-reported Outcomes. John Wiley, Chichester.

Fiebig, D.G., Keane, M.P., Louviere, J., Wasi, N., 2010. The generalized multinomial logit model: accounting for scale and coefficient heterogeneity. Marketing Science 29, 393–421.

Finn, A., Louviere, J.J., 1992. Determining the appropriate response to evidence of public concern: the case of food safety. Journal of Public Policy & Marketing 11, 12–25.

Flynn, T.N., 2010. Valuing citizen and patient preferences in health: recent developments in three types of best-worst scaling. Expert Review of Pharmacoeconomics & Outcomes Research 10, 259–267.

Flynn, T.N., Marley, A.A.J. Best-worst scaling: theory and methods. Invited chapter. In: Hess, S., Daly, A. (Eds.), Handbook of Choice Modelling. Edward Elgar Publishing, in press.

Flynn, T.N., Louviere, J.J., Peters, T.J., Coast, J., 2008. Estimating preferences for a dermatology consultation using Best-Worst Scaling: comparison of various methods of analysis. BMC Medical Research Methodology 8 (76).

Flynn, T.N., Louviere, J.J., Peters, T.J., Coast, J., 2010. Using discrete choice experiments to understand preferences for quality of life. Variance scale heterogeneity matters. Social Science & Medicine 70, 1957–1965. http://dx.doi.org/10.1016/j.socscimed.2010.03.008.

Hawkins, G.E., Marley, A.A.J., Heathcote, A., Flynn, T.N., Louviere, J.J., Brown, S.D., 2012. Accumulator Models for Best-worst Choices. Department of Psychology, University of Newcastle, Australia, Newcastle, NSW, Australia.

Hensher, D.A., Rose, J.M., Greene, W.H., 2005. Applied Choice Analysis: A Primer. Cambridge University Press, Cambridge.

Louviere, J.J., Hensher, D.A., Swait, J.D., 2000. Stated Choice Methods. Cambridge University Press, Cambridge.

Louviere, J.J., Street, D.J., Burgess, L., Wasi, N., Islam, T., Marley, A.A.J., 2008. Modelling the choices of single individuals by combining efficient choice experiment designs with extra preference information. Journal of Choice Modelling 1, 128–163.

Louviere, J.J., Woodworth, G., 1990. Best-worst scaling: A model for largest difference judgments. Working Paper, Faculty of Business. University of Alberta.

Marley, A.A.J., Flynn, T.N., Louviere, J.J., 2008. Probabilistic models of set-dependent and attribute-level best-worst choice. Journal of Mathematical Psychology 52, 281–296.

Marley, A.A.J., Islam, T., 2012. Conceptual relations between expanded rank data and models of the unexpanded rank data. Journal of Choice Modelling 5, 38–80.

Marley, A.A.J., Louviere, J.J., 2005. Some probabilistic models of best, worst, and best-worst choices. Journal of Mathematical Psychology 49, 464–480.

Marley, A.A.J., Pihlens, D., 2012. Models of best-worst choice and ranking among multi-attribute options (profiles). Journal of Mathematical Psychology 56, 24–34.

McIntosh, E., Louviere, J.J., 2002. Separating Weight and Scale Value: An Exploration of Best-attribute Scaling in Health Economics. Health Economists' Study Group, Brunel University.

Scarpa, R., Notaro, S., Raffelli, R., Pihlens, D., Louviere, J.J., 2011. Exploring scale effects of best/worst rank ordered choice data to estimate benefits of tourism in alpine grazing commons. American Journal of Agricultural Economics 93, 813–828.

Vermunt, J.K., Magidson, J., 2005. Technical Guide for Latent GOLD Choice 4.0: Basic and Advanced. Belmont Massachusetts: Statistical Innovations.

Relevant Websites

http://www.censoc.uts.edu.au/ – Centre for the Study of Choice.
http://www.jocm.org.uk/index.php/JOCM –Journal of Choice Modelling.
http://www.sjdm.org/ – Society for Judgment and Decision Making.
http://emlab.berkeley.edu/∼mcfadden/ – Daniel L. McFadden.
http://datasearch.uts.edu.au/censoc/members/detail.cfm?StaffID=7884 – T.N. Flynn.
htttp://datasearch.uts.edu.au/censoc/members/detail.cfm?StaffID=158 – J. Louviere.
http://www.uvic.ca/psyc/marley/ – A.A.J. Marley.

Bielefeld School of History

Lutz Raphael, Universität Trier, Trier, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Bielefeld School' has become the collective name of a group of West German historians who developed a specific form of social history defined as 'history of society' or 'Historische Sozialwissenschaft.' This version of social history was theoretically inspired by the work of Max Weber and critically used modernization theory as their points of reference for the historical analysis of modern German, American, and European History since 1800. History of modern society was understood as the dynamic interplay of autonomous subsystems, mainly economy, culture, and power ('Herrschaft'). Thus the history of processes and structures was given priority over the history of political events, of individuals or symbolic (cultural) representations. The explanation of the German 'Sonderweg' leading to the rise of Nazism and the Holocaust was a driving political motive behind the theoretical and methodological program of the School. It was mainly formed by historians trained by Hans-Ulrich Wehler and Jürgen Kocka at Bielefeld University and had their high time during the 1970s and 1980s.

Bielefeld School

'Bielefeld School' has become the collective name of a group of West German historians who developed a specific form of social history defined as 'history of society.' This program was mainly formed by Hans-Ulrich Wehler (born in 1931) and Jürgen Kocka (born in 1941), both teaching at Bielefeld University during the formative years of this school (1971–96 and 1974–88, respectively). This version of social history was theoretically inspired by the work of Max Weber and critically used modernization theory as their points of reference for the historical analysis of modern German, American, and European History, since 1800. History of modern society was understood as the dynamic interplay of three autonomous subsystems: economy, culture, and power ('Herrschaft'). Thus the history of processes and structures is given priority over the history of political events, of individuals or symbolic (cultural) representations. The explanation of the German 'Sonderweg' leading to the rise of Nazism and the Holocaust has been a driving political motive behind the theoretical and methodological program of the school.

The Formative Moment

The specificities of the social history the 'Bielefeld School' developed as their own style of historiography are strongly related to the intellectual situation of German historiography after the defeat of Nazism and German imperialism in 1945. The division of Germany and the political confrontation of the Cold War contributed to the survival of a (both politically and methodologically) conservative mainstream in West German historiography. As a result the international rise of new themes and methods of historical research had no or very limited effect on West German scholarship after 1945 (Berger, 2011). Neither the Annales School nor the American or British social history found many followers in both Germanies. Instead, orthodox Marxist–Leninist historiography (in East Germany) and a 'Strukturgeschichte' informed by the program of the Völkisch historiography of the interwar years were the main varieties of social history available in Germany till the early 1960s (Etzemüller, 2002).

The 1960s were the formative years of the future Bielefeld School whose early members were all born between 1929 and 1932 (with the exception of Jürgen Kocka born in 1941 but joining this group very early in his career in 1968). Wehler and Kocka, the two leaders of the school, and their colleagues were contemporaries of the pre-1968 intellectual situation in West Germany; they got their new intellectual orientations from the US and the UK. Both the fascination of the new social sciences and Western democracy were their starting points and lead to a sharp break with traditional German historiography and its theoretical backbone, a historicism that was practically reduced to a professionally sophisticated history of politics, history of (political) ideas, the history of institutions, and strongly linked with liberal and conservative nationalism (Berger, 2011). The younger historians chose a critical look at the German history of the nineteenth century and particularly of the Empire (1871–1918) focusing on the sharp class antagonisms and the negative role of the Prussian nobility in combination with the authoritarian traditions of the Prussian State. The so-called Fischer controversy starting in 1961 about the responsibility of the German politicians for the outbreak of World War I and the continuities of an aggressive imperialism among the German elites from the Kaiserreich to the Nazi period condensed all these new elements into a new master narrative of contemporary German history, revitalizing the thesis of a German 'Sonderweg,' a German way to modernity different from the Western liberal path, but now judged very critically as a path into moral and political catastrophe (Sheehan, 2005). The 'Bielefeld' School transposed this initially political impulse into a program of a new kind of history based on a broader approach toward the past including mainly the history of social groups and classes and theoretically inspired by the theories of modern societies developed by the social sciences. The rediscovery of Max Weber and his return back into German historical and social theory via Talcott Parsons and American historical sociology were the second disruptive event that gave start to the Bielefeld School.

The political upheaval of the years 1967 and 1968 and the rise of a new leftism (largely inspired by Neo-Marxism, later by French poststructuralism) did not change the intellectual profile and theoretical orientation of the Bielefeld School, now

on their way to organize themselves as a special group with their own program. However, the political turn to the left and the push toward democratic reforms in West Germany largely helped to pave the way for the sudden and quick breakthrough of the new group at the beginning of the 1970s. Their common political background brought most of them in more or less strong relations to the social democratic party and the social liberal coalition, and many of them engaged in the heated public debates of the time about the German (Nazi) past and the interpretation of the Holocaust (the so-called 'Historikerstreit'). They established themselves as public historians gaining high profile in the intellectual debates of West German democracy (Kocka and Budde, 2001; Wehler, 1988).

The Program

The group (at this starting point still rather small but using their networks within a rapidly expanding minority of social democratic or liberal historians) has very early defined their intellectual aims and their theoretical positions. They deliberately used the Kuhn model of scientific revolution through paradigm change as their own strategy for innovation and symbolically burnt all bridges with the 'historicist' past and symbolically distanced themselves from the mainstream of both professionally and politically conservative historians. This symbolic revolution was underpinned by the new critical interpretation of the historicist German tradition established in the Anglo-American world (Iggers, 1968). Confrontation and controversy were used to give profile and unity to the group in the making. Wolfgang J. Mommsen gave the tone in 1970 in his inaugural lecture at Düsseldorf university in announcing a historiography (Geschichtswissenschaft) "beyond historicism"(Cornelißen, 2010). Mommsen criticized the cult of empathic hermeneutical understanding (Verstehen) as opposed to explanation (Erklären), the idol of the biography and the priority of politics (Mommsen, 1971). To transcend these shortcomings, the social sciences and social theory were the new remedies. This turn lead to a series of attempts to better define the new agenda of social history. These were the two headings that suited best the Bielefeld program in the making.

'Gesellschaftsgeschichte' translated the term 'History of Society' coined in 1971 by Eric Hobsbawm but deleted the link to a Marxian view (Hobsbawm, 1974). It expressed the conviction of the Bielefeld School that social history should not be understood as a subfield of historical research – like military, cultural, or political history – but must be developed as a comprehensive approach to history. The analytical unit of inquiry and the theoretical object is 'society' understood as the integrated ensemble of different subsystems (economy, politics, culture) whose past can only be explained by historical processes leading to what social sciences may describe as 'structures' of society. Thus 'Gesellschaftsgeschichte' as a research practice has the aim to identify, describe, and explain the bundle of such processes giving form and unity to a given society. This program concentrated on modern Western societies and their complexities and therefore it heavily depended on different macro theories about the dynamics of the Western World since the eighteenth century. It were mainly American macro sociology, particularly Weberian historical sociology, structural functionalism, and modernization theory that informed the research program of the school, but less as overall big theories of development in sharp competition with Marxism but more as the converging point of a set of theories of middle range dealing with different processes like urbanization, professionalization, industrialization, secularization, and so on. The perspective on these processes should be one of critical inquiry mainly interested in problems of social inequality and of political power. From the beginning, the Bielefeld School investigated the particularities of each process in each region or nation they were studying, seeing their own historical practice as complementary to sociological theory or ongoing empirical research of the social sciences in contemporary society. It was Max Weber who now got highest esteem as the leading social theorist of modernity and therefore often was the reference point when it came to larger models of explanation or choosing of concepts (Wehler, 1973b, 1975; Kocka, 1977, 2011a).

This approach opened the door for the use of comparative methods in historical research as it allowed to define the common (overarching) element (a social structure, a process) linking two or more particular historical processes or phenomena transforming them into 'cases' open for a comparative analysis (Welskopp, 1995; Haupt, 2001; Haupt and Kocka, 2004). It was Jürgen Kocka who from the beginning engaged in comparative research in combination with the use of theoretical approaches (Kocka, 1981). The debate about the German 'Sonderweg' gave further political and moral incentives to comparative studies along national lines (Kocka, 1993). Since the 1980s, the comparative approach got stronger and in 1993 a center of comparative historical research was founded at the Freie Universität Berlin (now 'Berliner Kolleg für Vergleichende Geschichte Europas') (Haupt and Kocka, 2009).

In consequence of this opening to sociological and economic research of the past, the second heading 'Historische Sozialwissenschaft' (history as social science) delineated the ambitious program of an interdisciplinary scholarship or a new discipline integrating the traditions of the social sciences and history. It was Werner Sombart who had first coined the term in 1916 (Lenger, 2010). But what at the beginning of the twentieth century was a well-established tradition in social and economic research turned out to be a new program in the 1970s that was far away from academic reality. The relationship between the social sciences and history did not develop toward the merging of the disciplines but on the contrary toward segregation and mutual distance. The success of the social sciences since the 1970s pushed toward a further specialization and an even more pronounced focus on contemporary society and their problems. Unlike in the USA, historical sociology never established itself as a subfield in West German sociology.

The program of 'Historische Sozialwissenschaft' implies a new understanding of intellectual references and the Bielefeld School created their own intellectual tradition by adopting a long list of scholars from the field of economics and social sciences as founding fathers of the new kind of social history they were creating. The German 'historical school of national economics' (like Gustav Schmoller, Max Weber, Werner Sombart, Joseph A. Schumpeter, or Lujo Brentano) was the most prominent among the new

gallery of forerunners and founding fathers presented in a series of essays on "Deutsche Historiker"(Wehler, 1972–1982).

Their own rapid success during the 1970s and early 1980s changed deeply the intellectual landscape of German historiography. Whereas many historians in the fast-growing academic field of history still regarded the full program with suspicion and refused the style of antagonistic confrontation typical of the early years, they tacitly shared many positions with the Bielefeld School. Social history, the use of social theories of middle range, and the concept 'society' were widely accepted by scholars in modern and contemporary history. But this kind of moderate but general acceptance was accompanied by new challenges. During the 1980s, the international trends of the history workshop movement and of microhistory came to Germany and gave birth to 'Alltagsgeschichte' (history of everyday life) and 'Historische Anthropologie' ('historical anthropology'). Although these new groups shared many assumptions with the Bielefeld School, the leading members of the school refused these new trends engaging in a new polemic controversy. In this case, they could marginalize the new group institutionally but they failed to stop the spread of these new approaches better prepared to the 'cultural turn' as they refused the dichotomy of culture and society (Berdahl, 1982).

Therefore from the late 1980s on, the Bielefeld School had to adapt their own theoretical framework to a fast-changing national and international context (Wehler and Hettling, 1991; Kocka, 2010b). The Weberian classicism of the school tended to look old-fashioned and dogmatic to many young historians, their – modestly critic – defense of modernization theory (Wehler, 1975) came under heavy attack from different sides; adherents of postmodernism, postcolonialism, of gender studies, microhistorians, and historians reading Bourdieu or Foucault, all refused the framework of a theory that seemed as the last bulwark of intellectual defense of the West in a kind of never ending cold war. Younger historians trained in the Bielefeld School started to distance themselves from the earlier positions of the school. They proposed a variety of revisionisms giving birth to a pluralism that blurred the differences toward other positions inside the discipline. Firstly, modernization theory was given up in favor of concepts of modernity. Another option was to accept the 'cultural turn' and to give up any priority for socioeconomic explanations and macro theories (Daniel, 1993). A third option was to remodel the theoretical approaches. Bourdieu's concept of capital, habitus, and field, Giddens' macrotheory of social structure and agency were largely used to renew the program of 'Gesellschaftsgeschichte' (Wehler and Hettling, 1991; Mergel and Welskopp, 1997). A fourth option was to shift the thematic focus away from society toward other units of historical research. Thus a part of the group engaged in taking up the concept of civil society internationally at the forefront since 1990. The latest revision opened the perspective toward global history renewing the exchange with newer versions of historical sociology and macrotheory (Osterhammel et al., 2006). But these revisions of the program did not stop its overall decline. The marginalization of social history has played an important role in this loss of visibility and attraction because historians following the new trends (toward a 'new cultural history' of the political sphere, discourse analysis, or the history of concepts) in general did not follow the theoretical framework of the Bielefeld School, neither in its first nor in its revised versions.

The Group and Its Institutional Backings

The term Bielefeld School itself alludes to the institutional and local backing of this intellectual program. In this, as in other cases of academic groups, 'school' is a very strong term. Seen from within, the Bielefeld School has always been a federation of individuals around a small set of common ideas but a growing list of common institutional interests. It was the expansion of the German educational system and the creation of new universities and new posts in academe that accelerated the formation of the school giving professorships or other independent research positions to the still 'young' historians inspired by the new ideas and the examples of Anglo-American social history. Wehler was nominated professor at the Freie Universität Berlin in 1971, Kocka at Bielefeld in 1973, W. Mommsen (born in 1934) in 1970 at Düsseldorf, R. Rürup (born in 1934) in 1975 at Technische Universität Berlin, Wolfgang Schieder (born in 1935) at Trier in 1971. From the beginning, the Bielefeld School was relying on a much broader circle of historians sympathetic with the social history and the critical revisions of German historiographical traditions but not sharing all theoretical and methodological positions of the group. Thus elder traditions of social history survived and merged into a common practice of research.

From the beginning they were coopting younger scholars like Kocka (born in 1941), H.A. Winkler (born in 1939), or Klaus Tenfelde (born in 1944) at a very early stage of their career. All were about 40 years or younger when they got their first chairs and when they joined their efforts for the creation of the new school. Both the universities of Bielefeld and of Berlin became academic strongholds of the new school, in Bielefeld, Kocka and Wehler were teaching side by side since 1973, coopting new colleagues sympathetic with or openly supporting their program. In Berlin, first R. Rürup, later H. Kaelble, J. Kocka, and H.A. Winkler were teaching. The other members were holding chairs at other universities like Gießen, Düsseldorf, Trier, Freiburg, or Darmstadt. Thus the recruitment of students engaging in the kind of new social history started at many academic points in the 1970s. As long as the academic field of historical sciences was polarized along the internal political and theoretical controversies of the 1960s and early 1970s, adherents of the Bielefeld School could be easily identified. But with the expanding of the field and the fading away of the controversies, the field tended to a new kind of informal consensus and a broad pluralism blurring all fixed lines among schools.

From the beginning the group established strong ties to publishing houses, first Kiepenheuer & Witsch, then Vandenhoeck & Ruprecht, and Suhrkamp, later C.H. Beck ,and got high visibility in an expanding market for academic books (Blaschke and Raphael, 2007). The book series 'Kritische Studien' launched in 1972 soon became the most prestigious place for younger historians joining the group and their program to get published. In 2011, 200th volume was published (a reprint of articles of J. Kocka). In 1975 the first issue of the journal

Geschichte und Gesellschaft appeared. This journal was the most visible and most durable institutional support for the Bielefeld School. Its editorial board brought together the core group of 11 historians more or less closely connected to the program, and five social scientists (two political scientists, two sociologists, and one economist) whose presence underlined the interdisciplinary ambitions of the program (Raphael, 1999). The editorial board was meant to play an active role, but in the end it was a smaller board of insiders who did most of the editorial work. Editor-in-chief from 1973 to 1996 was Hans-Ulrich Wehler. He was mainly supported by H. Berding, Wolfgang Mommsen, Hans J. Puhle, Wolfgang Schieder, and Jürgen Kocka who were frequently present at the board meetings and took over the responsibility for a single thematic issue. Their names reappear among the list of authors with the highest number of contributions to the journal from 1975 to 1994. The journal quickly gained national and international recognition and established itself as one of the leading German historical journals in modern and contemporary history. It was never conceived as a partisan journal and opened its pages to critical voices and debates but in the end became the most visible institution of the Bielefeld School presenting like a showcase the ongoing research and the debates that were judged important among the school members. The editorial board was more or less stable for more than 15 years and it was only during the late 1980s that a first exchange of members took place. The cooptation of eight new members between 1987 and 1989 opened the journal for younger scholars (mainly born between 1942 and 1944) who had established more or less strong ties of collaboration with the core group since the 1970s. A generational break did not happen before the turn of the century when all elder members of the group left the editorial board and when the journal was handed over to historians who had started their academic careers in the 1970s and 1980s often under the supervision of the first generation of the Bielefeld School. This change in the editorial board can be seen as the moment when the journal, now in the hand of the heirs, was not any longer regarded as the flagship of a school. Themes and theoretical points of references have got more pluralistic corresponding to the new intellectual orientations and positions of power among German historians in modern and contemporary history.

Topics of Research

'Gesellschaftsgeschichte' or 'Historische Sozialwissenschaft' defined a program that in itself was open for many versions of historical research. It did not imply a turn to a defined set of themes, not even the 'social' had by definition in itself any priority. The search for processes and their interplay allow a very flexible choice of topics. But in practice, the Bielefeld school took up a set of themes and problems that got first priority: their geographical focus was Germany so that the bibliographies of German social history of the 1980s and 1990s can be read as research records of the school itself (Wehler, 1976, 1993).

The history of social classes and groups in Western societies since 1800 may be defined as the first and largest theme of common research. From the 1960s to the 1980s it was the core field of social history. From the beginning, two tendencies coexisted among the Bielefeld School: one that may be defined as the 'single-issue' approach concentrating at the study of one single social group or class and tending toward a detailed description of the different dimensions of group or class existence (often at a regional level), and the other that prefers a 'relational' perspective seeing classes or groups defined by their relative position and their antagonisms. The second perspective is closer to Marxist approaches and the original definition of 'history of society' by Hobsbawm. As always, many studies combined both approaches in more or less consistent ways. The Bielefeld scholars invested heavily in the study of upper classes in nineteenth-century Germany (both nobility and bourgeoisie 'Bürgertum'), the new middle class professions (like physicians, academics, journalists) and the white collar workers (Kocka, 1981, 1995).

But at the same time, the group shared the general trend of the 1960s and 1970s to link social history to the history of the modern working classes and the labor movement. In itself this subfield can be seen as a cluster of many single issues: the history of the organizations, the class formation as a social and cultural process, the study of industrial labor, and the history of working class culture and consumerism. As a whole, the history of the labor movement was strongly leaning toward political history.

A third topic, often nearer to political than to social history attracted much of the research activities of the 'Bielefeld School.' This field may be called the social historical study of national politics. It focuses on the transformation of social forces, economic interests into political will and power. Hans-Ulrich Wehler has given the model book for this new program of a political history in 1973 with his study on the German Empire (Wehler 1973a, 1985). Later he engaged himself in a long lasting effort to write the history of German society since 1800. It took 20 years to finish the book – more than 5500 pages – interpreting the history of modern Germany as the interplay of three independent but strongly related subsystems (economy, power structure 'Herrschaft,' and culture and as the interfering zone of all three subsystems: the social) (Wehler, 1987–2008). More often, the approach leads to the study of particular political institutions, groups, or processes by including the study of the social background and the economic interests behind the scene. This kind of 'materialistic' approach was particularly at fashion during the rise of the Bielefeld School and ended with a large body of case studies about the impact of social and economic phenomena on the political history of Germany from 1800 to 1945. A particular subfield concerns the political and social history of modern ideologies. Again it were the German versions of liberalism, nationalism, socialism, or antisemitism that were analyzed in a perspective integrating social, cultural, and political aspects.

A quantitative study of articles published in the journal *Geschichte und Gesellschaft* allows to specify the main trends. Political phenomena or political power were the single dominant theme of articles published during the two 5-year periods compared (1975–79 vs 1995–99); more than 43% of the 149 articles were dealing with this issue. Economic processes (14 and 8%), social change (12–14%) were less prominent. Cultural phenomena were more often treated (27 and 30%) respectively. These two snapshots underline the continuities in

the research agenda and they show that the Bielefeld School was much more open to themes of cultural history as the controversies of the 1990s may suggest (Raphael 1999: p. 23). The historical world of the Bielefeld school was mainly a German world (66% of all articles from 1975 to 1999 in *Geschichte und Gesellschaft*) and secondly an Atlantic world including Western Europe and North America. Southern and Eastern Europe came as third and the history of Asia or Africa was at the margins (7 from 336 articles with a geographically identified subject).

Since the turn of the century, the new editorial board has tried hard to turn away from this Germano-centrism inherited by the original program and the new team was deliberately opening the agenda of the journal toward the history of modern Asia, Africa and put global and transnational history to the list of top priorities. But this shift implied a further step away from the original theoretical and empirical program of the school.

International Relations and Perspectives

The strategy of polemic break with German historicism has, as its counterpart, a turn toward internationalization. The 1960s and 1970s saw the establishment of an international rise of social history and the Bielefeld School participated actively in this international trend (Lewis, 2013). British and American social historians played an important role in establishing new contacts and their transfers of ideas were crucial for the integration of the Bielefeld school into this transnational field of research. The younger generation was systematically invited to study abroad and to create their own international networks.

The internationally most visible historian of the Bielefeld group is Jürgen Kocka whose publications were translated since the 1970s in English, French, and many other languages and who actively participated in international debates (Kocka, 1999, 2010a,b, 1980). He has been an academic guest and fellow at Chicago, Stanford, Princeton, Los Angeles, Jerusalem, Budapest, Paris, and Oxford. His high international reputation lead to his election as president of the International Committee of Historical Sciences from 2000 to 2005.

But at the same time, the Bielefeld school shared with most of the other national schools of social history a leaning toward methodological nationalism. In this respect they shared a particularity of social history in general: more than colleagues in other subfields of historical research like intellectual or (international) political history social historians were and still are strongly embedded in the debates on social problems of their own nations making it often very difficult to bridge the gap between the different national perspectives informing the way social historians were constructing their national pasts. The German nation-state was the preferred point of reference and modern 'society' was understood as structured by this political unit. Therefore many members of the Bielefeld school specialized in problems of social history defined in strictly national terms.

Past or Present?

In 2013, the Bielefeld School seems to be more a phenomenon belonging to the past than to the present. As a collective, the group has dissolved into different subgroups and most of their members see themselves now as single parts in a professional field that is no longer divided along clear cut programmatic or political lines. But the intellectual heritage of their program and the many lasting results of their research activities still inform many fields of German modern and contemporary historiography.

See also: Comparative History; Comparative Method in Anthropology; Historical Thought and Historiography: Current Trends; Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; Historicism; History and the Social Sciences; Modernity: History of the Concept; Modernity; Modernization and Modernity in History; Social History; Weber, Max (1864–1920).

Bibliography

Berdahl, Robert M., 1982. Klassen und Kultur. Syndikat, Frankfurt am Main.

Berger, Stefan, 2011. From the search for normality to the search for normality: German historical writing. In: Schneider, Axel, Woolf, Daniel (Eds.), The Oxford History of Historical Writing, Historical Writing since 1945, vol. 5. Oxford UP, Oxford, pp. 220–241.

Blaschke, Olaf, Raphael, Lutz, 2007. Im Kampf um Positionen. Änderungen im Feld der französischen und deutschen Geschichtswissenschaft nach 1945. In: Eckel, Jan, Etzemüller, Thomas (Eds.), Neue Zugänge zur Geschichte der Geschichtswissenschaft. Wallstein, Göttingen, pp. 69–109.

Cornelißen, Christoph, 2010. Wolfgang J. Mommsen – der Repräsentant einer Historiker generation? In: Cornelissen, Christoph (Ed.), Geschichtswissenschaft im Geist der Demokratie. Wolfgang J. Mommsen und seine Generation. Akademie Verlag, Berlin, pp. 11–41.

Daniel, Ute, 1993. "Kultur" und "Gesellschaft". Überlegungen zum Gegenstandsbereich der Sozialgeschichte. Geschichte und Gesellschaft 19, 69–99.

Etzemüller, Thomas, 2002. Sozialgeschichte als politische Geschichte. die Etablierung der sozialgeschichte in der westdeutschen Geschichtswissenschaft. Comparativ 12 (1), 12–33.

Haupt, Heinz-Gerhard, 2001. Comparative history. International Encyclopedia of the Social and Behavioral Sciences 4, 2397–2403.

Haupt, Heinz-Gerhard, Kocka, Jürgen, 2004. Comparative history. Methods, aims, problems. In: Cohen, Deborah, O'Connor, Maura (Eds.), Comparison and History. Europe in Cross-national Perspective. Routledge, New York, pp. 23–39.

Haupt, Heinz-Gerhard, Kocka, Jürgen, 2009. Comparative and Transnational History. Central European Approaches and New Perspectives. Berghahn Books, New York.

Hobsbawm, Eric, 1974. From social history to the history of society. In: Flinn, Michael W., Smout, T.C. (Eds.), Essays in Social History. Clarendon Press, Oxford, pp. 1–22.

Iggers, Georg G., 1968. The German Conception of History. Wesleyan University Press, CT.

Kocka, Jürgen, 1977. Sozialgeschichte. Begriff, Entwicklung, Probleme. Vandenhoeck und Ruprecht, Göttingen (Kleine Vandenhoeck-Reihe, 1434).

Kocka, Jürgen, 1980. White collar Workers in America, 1890–1940. A Social-Political History in International Perspective. Sage Publications, London, Beverly Hills (Sage studies in 20th century history, 10).

Kocka, Jürgen (Ed.), 1981. Angestellte im europaeischen Vergleich. D. Herausbildung angestellter Mittelschichten seit dem. späten 19. Jh. Vandenhoeck und Ruprecht, Göttingen (Geschichte und Gesellschaft : Sonderheft, 7).

Kocka, Jürgen, 1993. Comparative historical research: German examples. International Review of Social History 38 (03), 369–379.

Kocka, Jürgen (Ed.), 1995. Bürgertum im 19. Jahrhundert. Deutschland im europäischen Vergleich : eine Auswahl. Vandenhoeck & Ruprecht, Göttingen (Kleine Vandenhoeck-Reihe, 1574).

Kocka, Jürgen, 1999. Industrial Culture and Bourgeois Society. Business, Labor, and Bureaucracy in Modern Germany. Berghahn Books, New York.

Kocka, Jürgen, Budde, Gunilla-Friederike, 2001. Interventionen. Der Historiker in der öffentlichen Verantwortung. Vandenhoeck & Ruprecht, Göttingen.

Kocka, Jürgen, 2010a. Civil Society and Dictatorship in Modern German History. University Press of New England, Hanover [N.H.] (The Menahem Stern Jerusalem lectures).

Kocka, Jürgen, 2010b. History and social sciences today. In: Joas, Hans, Klein, Barbro (Eds.), The Benefit of Broad Horizons. Intellectual and Institutional Preconditions for a Global Social Science. BRILL, Leiden, pp. 53–67 (International Comparative Social Studies).

Kocka, Jürgen, 2011a. Zwischen Dogmatismus und Dezision: Wirklichkeit und Methode bei Karl Marx und Max Weber. In: Kocka, Jürgen (Ed.), Arbeiten an der Geschichte. Gesellschaftlicher Wandel im 19. und 20. Jahrhundert. Vandenhoeck & Ruprecht, Göttingen (u.a.), pp. 15–39 (Kritische Studien zur Geschichtswissenschaft, 200).

Kocka, Jürgen (Ed.), 2011b. Arbeiten an der Geschichte. Gesellschaftlicher Wandel im 19. und 20. Jahrhundert. Vandenhoeck & Ruprecht, Göttingen (u.a.) (Kritische Studien zur Geschichtswissenschaft, 200).

Lenger, Friedrich, 2010. "Historische sozialwissenschaft": Aufbruch oder Sackgasse? In: Cornelissen, Christoph (Ed.), Geschichtswissenschaft im Geist der Demokratie. Wolfgang J. Mommsen und seine Generation. Akademie Verlag, Berlin, pp. 115–132.

Lewis, Brian, 2013. Social history: a new kind of history. In: Partner, Nancy F., Foot, Sarah (Eds.), The SAGE Handbook of Historical Theory. SAGE, Los Angeles, CA, pp. 93–104.

Mergel, Thomas, Welskopp, Thomas (Eds.), 1997. Geschichte zwischen Kultur und Gesellschaft. Beiträge zur Theoriedebatte. Originalausg. C. H. Beck, München (Beck'sche Reihe, 1211).

Mommsen, Wolfgang J., 1971. Die Geschichtswissenschaft jenseits des Historismus. Erw. Fassung einer am 3. Feb. 1970 an der Univ. Düsseldorf gehaltenen Akad. Antrittsvorlesung. Droste, Düsseldorf.

Osterhammel, Jürgen, Langewiesche, Dieter, Nolte, Paul (Eds.), 2006. Wege der Gesellschaftsgeschichte. Vandenhoeck & Ruprecht, Göttingen.

Raphael, Lutz, 1999. Nationalzentrierte Sozialgeschichte in programmatischer Absicht: Die Zeitschrift "Geschichte und Gesellschaft. Zeitschrift für Historische Sozialwissenschaft" in den ersten 25 Jahren ihres Bestehens. Geschichte und Gesellschaft 25, 5–37.

Sheehan, James J., 2005. Paradigm lost? The "Sonderweg" revisited. In: Budde, Gunilla-Friederike, Conrad, Sebastian, Janz, Oliver (Eds.), Transnationale Geschichte. Themen, Tendenzen und Theorien. Vandenhoeck & Ruprecht, Göttingen, pp. 150–160.

Wehler, Hans-Ulrich (Ed.), 1972–1982. Deutsche Historiker. Vandenhoeck & Ruprecht, Göttingen.

Wehler, Hans-Ulrich, 1973a. Das Deutsche Kaiserreich, 1871–l9l8. Vandenhoeck & Ruprecht, Göttingen (Deutsche Geschichte, 9).

Wehler, Hans-Ulrich, 1973b. Geschichte als historische Sozialwissenschaft, 1. Aufl. Suhrkamp, Frankfurt (am Main) (Edition Suhrkamp, 650).

Wehler, Hans-Ulrich, 1975. Modernisierungstheorie und Geschichte. Vandenhoeck und Ruprecht, Göttingen (Kleine Vandenhoeck-Reihe, 1407).

Wehler, Hans-Ulrich, 1976. Bibliographie zur modernen deutschen Sozialgeschichte. (18.–20. Jh.). Vandenhoeck und Ruprecht, Göttingen (Uni-Taschenbücher, 620).

Wehler, Hans-Ulrich, 1985. The German Empire, 1871–1918. Berg Publishers, Leamington Spa, Warwickshire, UK, Dover, N.H.

Wehler, Hans-Ulrich, 1987–2008. Deutsche Gesellschaftsgeschichte, 5 vol.. C.H. Beck, München.

Wehler, Hans-Ulrich, 1988. Entsorgung der deutschen Vergangenheit? Ein polem. Essay zum "Historikerstreit". Büchergilde Gutenberg, Frankfurt (Main) (Die kleine Reihe).

Wehler, Hans-Ulrich, 1993. Bibliographie zur neueren deutschen Sozialgeschichte. Verlag C.H. Beck, München (C.H. Beck Studium).

Wehler, Hans-Ulrich, Hettling, Manfred, 1991. Was ist Gesellschaftsgeschichte? Positionen, Themen, Analysen. Beck, München.

Welskopp, Thomas, 1995. Stolpersteine auf dem Königsweg. Methodenkritische anmerkungen zum internationalen Vergleich in der Gesellschaftsgeschichte. Geschichte und Gesellschaft 25, 339–367.

Big Five Factor Model, Theory and Structure

Boele de Raad, University of Groningen, Groningen, The Netherlands
Boris Mlačić, Institute of Social Sciences Ivo Pilar, Zagreb, Croatia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The psycholexical origin of the Big Five model is sketched, theoretical considerations that have guided the development of the model are given, different structural representations are described, critical notions reported, and cross-cultural findings and its usefulness reviewed. The importance of a paradigm for furthering the field is emphasized, and the lexical Big Five factors are compared with the questionnaire-based Five-Factor Model. Also alternative models are reviewed. Moreover, specifications of the Big Five model are described, both in terms of factor facets, and in terms of representation in the form of circular and hierarchical arrangements of personality traits.

Introduction

The 'Big Five' or Five-Factor Model (FFM) is the personality trait model constituted by the five factors or dimensions Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Intellect or Openness to Experience. This model emerged from applying the principles of the psycholexical approach to personality (De Raad, 2000). That approach theoretically started with the so-called lexical hypothesis, holding that *our common stock of words, our lexicon, embodies the distinctions, men have found worth drawing. Observations of individual differences that people have found of interest, utility, or importance get encoded by words and expressions into the substance of language*. This hypothesis has been expressed in different ways by the personality psychologists Cattell (1943) and Goldberg (1981), and independently by the philosopher of language Austin (1970), the poet and novelist Themerson (1974), and the psycholinguist Miller (1991).

The early success of the Big Five model was caused, among others, by Norman (1963), and by the seminal psycholexical project led by Goldberg (e.g., 1981). This made Costa and McCrae (1992a) add the two factors Agreeableness and Conscientiousness to their own model until then consisting of Neuroticism, Extraversion, and Openness to Experience (NEO): thus, the NEO-PI came into existence to measure the FFM dimensions. A host of new lexical studies in various languages, often supportive of the Big Five model, and the productive work of Costa and McCrae provided the grounds for McCrae and Costa (2008) to build their Five-Factor Theory, and for claiming that the Big Five factors have causal status.

The Importance of Paradigm and Taxonomy in Building Personality Theory

While the Big Five model is now shared in personality psychology (John et al., 2008), it received frequent criticism from perspectives presenting alternative structures (e.g., Ashton and Lee, 2007; Saucier, 2003; Zuckerman et al., 1993) and from studies focusing on cross-cultural replicability of trait factors (De Raad et al., 2010).

For long, personality psychology had been criticized for lack of a paradigm or a shared model. Eysenck (1991) argued that the development of a taxonomy is a prerequisite for furthering any scientific field. John et al. (1988), who defined taxonomy as "a systematic framework for distinguishing, ordering, and naming types and groups within a subject field" (p. 172), compared the importance of a taxonomy in personality with the Linnaean taxonomy that spurred the development of biology as a natural science. Stelmack (1997) paralleled the importance of personality taxonomy with Mendeleev's periodic system of elements in chemistry, stressing the predictive aspects of paradigmatic processes.

The rise of the Big Five model in the early 1990s was applauded as a possible solution regarding the lack of a shared model (Costa and McCrae, 1992a; Goldberg, 1993), yet with stern opposition especially from Eysenck (1992). Now, the Big Five model has secured its place as the dominant model in personality starting in the late 1990s (DeYoung et al., 2007; John et al., 2008). This could be concluded by the sheer number of publications, but, more importantly, by the models' significance in predicting many important life outcomes (Roberts et al., 2007).

Despite the accomplishments in organizing the scattered field of personality, the Big Five was criticized as an atheoretical approach (Block, 2010; Eysenck, 1992). While some claim that the theory underlying the Big Five model is still in development (DeYoung et al., 2010), McCrae and Costa (2008) presented their Five-Factor Theory, aiming to explain the genetic, biological, and causal underpinnings of each dimension. Regardless of the current status of Big Five theory, one can conclude that in just two decades, personality psychology came a long way from a disjointed field to taxonomy building, a shared model, and a prospective personality theory.

The Lexical Hypothesis in the Context of Everyday Differentiation

Everyday conversations relate for a large part to behavior and traits of self or other. Ideas about people are put into words, and when there is lack of words, new words are invented. The practical usage provides feedback on the effectiveness of such

words. Terms like *egoistic*, *aggressive*, or *shy* that have proven useful for their capacity to represent and communicate experiences may end up in the lexical storehouse. The more often a trait or disposition is observed in behavior, the greater the chance that the trait is labeled and that it becomes a communicative commodity.

That lexical archive with its enormous differentiating potential forms the rough material for the construction of a scientifically acceptable medium to describe personality traits. A problem is that everyday language is not a neutral language. The lingual elements that stood the survival test, are loaded in various respects.

Trait words are often contaminated by their context of development and application. For example, the meaning of *devout* and *compassion* is particularly articulated under the influence of Christian values. *Jovial* stems from astrology, *radical* has political connotation, and *neurotic* is from psychology. The extravert was once described as spiritually poor and superficial and the introverted as high and inwardly rich. Now, the introverted is rather the eccentric and individualistic and the extroverted is energetic and sociable. Some words are used only in specific contexts – they are region specific – belong to a certain jargon, or serve a fashionable purpose. As invectives words can become the verbal vitriol with which sometimes reputations are ruined and people are brought to stand apart. So, any effort in exploiting everyday language for scientific purposes needs to show awareness of the fact that many trait words are imbued with contextual information.

The Use of the Dictionary for the Development of a Shared Model of Personality

It took a long time for the scientific study of personality to recognize the value of natural language for the development of a shared model. The psycholexical approach usually involves the use of a dictionary as the tangible representation of the lexicon. The first who pointed to a dictionary as a valid resource was possibly Galton (1884), estimating that roughly 1000 words might be used to describe character. Allport and Odbert (1936) selected all (almost 18 000) personality-relevant words from Webster's unabridged dictionary, 50 years later. Those words were classified into four categories, differing in importance for the description of personality for everyday conversation and differentiation, thus taking into account the contextual issues mentioned above. Those categories were 'Neutral terms designating possible personality traits,' 'temporary moods or activity,' 'evaluations,' and a 'miscellaneous' category. The first group contained most clearly 'real' traits of personality (pp. 24–38).

Block (1995) criticized the ordinary language origin of the trait concepts. Earlier, Costa and McCrae (1992b) and Eysenck (1992) disputed the criteria to arrive at basic dimensions of personality. Much of these disputes relate to personality dimensions found elsewhere in the personality literature, and which were hardly visible in the Big Five framework. Some of those dimensions covered specific areas of interest, such as locus of control (Rotter, 1954) and ego development (Loevinger, 1976); other dimensions were part of alternative multidimensional systems.

John et al. (1988) summarized the main criticisms, including the laypeople origin, the contextual loading, and the additional complication of communicability across languages. Saucier and Goldberg (1996) agreed that the natural language has shortcomings, but they argued that it represents the best approximation of the total population of personality variables. Wiggins (1973) concluded that "we can hardly afford to ignore such a storehouse of accumulated wisdom as a natural starting-point for the study of behavioral attributes" (p. 329). In addition, the lexical approach minimizes a prestructuring of variables according to theorists' preconceptions of personality (Saucier and Goldberg, 1996; De Raad, 2000). Hofstee (2003) argued that one can dispute that ordinary language is not subtle enough for scientific purposes, yet, in studies with questionnaires built on expert language (e.g., Digman and Inouye, 1986), the Big Five were also recovered.

Working on Allport and Odbert's first category, the neutral or 'real' personality traits, including about 4500 terms, Cattell (1943, 1947) built a system with 16 intercorrelated primary factors. None of the later studies using Cattell's variables, however, was successful in replicating that many factors but rather arrived at a much simpler structure with five independent dimensions (see Norman, 1963). Since Norman's (1963) study, those five factors had been referred to as the 'Norman five,' but dubbed the 'Big Five' by Goldberg (1981) in whose work the fifth factor shifted labels from Culture to Intellect (Goldberg, 1993). In the late 1980s and 1990s, the lexically based taxonomy approach was followed in many languages, mainly in Europe, with the replication of the Big Five being most successful in Germanic and Slavic languages (Saucier, 2009).

Intellect or Openness to Experience: A Historical Difference of Opinion

John (1990) observed that there seemed to be no *single* Big Five. Referring to different Big Five studies, the variation in factor naming was striking, with the clearest differences for factor five, referred to as Intellect (or Culture), but also as Openness to Experience (Costa and McCrae, 1985). The Intellect label proceeded from lexical studies, and emphasized 'intellectual' traits such as intelligent and insightful, and Openness to Experience proceeded from a questionnaire approach and emphasized 'open' traits such as imaginative and artistic. The question about the precise meaning of the fifth factor had led to a special issue of the *European Journal of Personality* (De Raad and Van Heck, 1994), showcasing the difference in perspective of the lexical and the questionnaire approach.

The difference of opinion was about what should be considered as important differentiating traits. Costa and McCrae (1985) adopted Openness to Experience from Coan (1974), as the concept that should capture the contents of a cluster of traits in their own work, they had referred to as an Experiential Style dimension. Coan had developed the Openness to Experience scale on the basis of a study using a battery of existing instruments. Coan distinguished various facets for his Openness to Experience scale, and Costa and McCrae (1985) adopted six of those facets for their own purpose. That facet-differentiation for Openness to Experience seemed

to have dictated a similar extension for the other Big Five factors in Costa and McCrae's (1992c) instrument.

Although the difference in perspective between the proponents of the 'lexical Big Five' and the FFM are frequent and sometimes heated, the adoption by Costa and McCrae (1992b) of the two 'lexical' personality dimensions, Agreeableness and Conscientiousness, was incremental in popularizing the Big Five model and it can be construed as a test of its usefulness. The work of Costa and McCrae expanded the five factors from the domain of language and trait-descriptive words to the domain of questionnaires and led to accept the model by more researchers and ultimately formed a step forward to building a theory embedding the Big Five.

The questionnaire origin in Coan's work may be responsible for the expansion of semantics of factor five, in comparison to what the lexical approach had generated. In line of the above critics, McCrae (1990) argued that Openness to Experience is inadequately represented in natural language trait terms. However, Mulder (2006) demonstrated that, in Dutch (De Raad and Barelds, 2008), there is no problem finding a sufficient number of descriptors in the natural language to reliably define both the domain of Openness to Experience and the corresponding six facets.

Interestingly, DeYoung et al. (2009), who used the fMRI in relating working memory with Intellect and Openness, found that Intellect was correlated with working memory accuracy and brain activity, while Openness was not. Moreover, the authors interpreted the neural regions involved in the association between working memory and Intellect as a neural substrate of Intellect.

From Factor Model to Big Five Theory

Regarding the Big Five theoretical status today, there are two groups of thought: one sees that the transition from the Big Five model to theory is still in progress while the other finds that shift has already been achieved. For the Big Five framework to be considered a theory, the first group (e.g., DeYoung, 2010a; John and Naumann, 2010; Srivastava, 2010) specified conditions to be fulfilled, such as an explained evolutionary basis, an account of developmental aspects, an explanation of the underlying psychobiology, a specification of the hierarchy in the system, cross-cultural replicability, and more. The second group (e.g., McCrae, 2010; McCrae and Costa, 2008) postulated the Five-Factor Theory claiming all or most of the conditions above to be answered, with the five dimensions being universal, and with strong biological and genetic bases. Maybe their strongest claim is that traits constituting the five factors are not just descriptive, but represent causal influences that have effects on lives of all persons. The key components of their theory include the five factors as biologically based tendencies, characteristic adaptations covering the developmental aspects and learning experiences, the self-concept as a sort of derivation from the adaptations, and external influences. With the exclusion of the biological bases, these different components are assumed to be connected through dynamic processes.

One major criticism of the theory, as acknowledged by the authors (McCrae, 2010), concerns the dynamic processes that are so prominent in the theory, yet not explained in detail. In order to have a model to be shared by others, more substantial research regarding the major issues has to be undertaken. This has in fact begun in some areas such as personality neuroscience or linking the Big Five with the brain structure (e.g., Adelstein et al., 2011; DeYoung et al., 2010), and in the biology of traits (cf DeYoung, 2010b). Regarding developmental aspects, much research has been done already (cf Mlačić and Milas, in press), as well on the evolutionary basis of the five factors (cf Buss and Hawley, 2010).

The Big Five/FFM Dimensions and Lower-Level Constructs

Just as they were firm in claiming that the Five-Factor Theory is indeed a theory, Costa and McCrae (1992c) claimed that each of the Five Factors consists of exactly six facets, or lower-level constructs, thus forming a set of 30 facets. Openness to Experience, for example, comprises facets of Fantasy, Aesthetics, Feelings, Actions, Ideas and Values.

Contrary to FFM, the work on the lexical Big Five subcomponents or facets was far from being in agreement. The number of facets ranged from only a few to nine per dimension (Goldberg, 2014). Those nine facets came about in the work of Hofstee et al. (1992), generating a total of 45 facet-scales. Peabody and De Raad (2002) analyzed the psycholexical work in six languages and reported six facets for Extraversion, four for both Agreeableness and Conscientiousness, three for Emotional Stability, and five for Intellect. DeYoung et al. (2007) collected data on the 30 facet-scales from Costa and McCrae, and the 45 facets-scales originating in Hofstee et al. (1992), thus producing 15 facet-scales per Big Five factor domain. DeYoung et al. (2007) proposed a two-level hierarchy where the 15 narrower facets make 2 broader facets for each of the Big Five. For Openness to Experience, for example, these findings were roughly in agreement with Griffin and Hesketh (2004) and Mulder (2006), who distinguished a broader facet Openness to internal experiences (poetic, fantasy, and sensitive) and a second broader facet Openness to external experiences (ideas, broad interests).

Apart from such matters of detailing the Big Five domains, the contents of the Big Five factors may well be summarized as follows: Extraversion generally relates to the approach of the individual to the social and material world in an energetic manner (John and Srivastava, 1999). Agreeableness mostly relates to prosocial behavior (Graziano and Eisenberg, 1997). Conscientiousness mostly relates to impulse control and conformity (Hogan and Ones, 1997). Emotional Stability, representing the one pole of a dimension that has Neuroticism as the other pole, is the factor that is omnipresent in psychological research, especially in clinical psychology. De Raad et al. (2010) found that the Emotional Stability/Neuroticism factor appeared relatively weak in cross-cultural lexical studies, which contrasts with its historical prominence. It is possible that Emotional Stability has lesser representation in most natural languages, in comparison to clinical psychology, where its appreciation has led to many nuances of emotional experiences. Finally, the fifth of the Big Five relates to such traits as imagination, fantasy, creativity, and talent (cf Goldberg, 1993; John et al., 2008; Peabody and De Raad, 2002).

Big Five Alternatives

Because of its number of factors, the most obvious alternative for the Big Five seems to be Zuckerman's alternative five-factorial model (e.g., Zuckerman, 2002). This alternative started from the assumption that important traits are assumed to have a biological-evolutionary basis, and it had its origin in studies of temperament and of psychobiological studies. Some other studies with a similar origin are Eysenck's (1967) Giant Three (Extraversion, Neuroticism, and Psychoticism), and Strelau's (1983) Pavlovian theory-based model of temperament.

Assuming the comprehensiveness of the Big Five and their trait semantic coverage, an important question for competing systems is whether they account for more or less the same phenomena. According to Zuckerman (1984) human traits are best studied in comparison to animal behavior and he has made strong efforts in demonstrating that *sensation seeking* is a basic trait underlying both human and animal behavior. In a study using Costa and McCrae's five factors, Eysenck's Giant three, and Zuckerman's five factorial alternative, it turned out that those alternative factors had strong relations to the Big Five, except to Openness to Experience. Hogan (1983), who shares evolutionary elements with Zuckerman's model, combines it in his Socioanalytic theory with elements of interpersonal theory, focusing on interpersonal competence and effectiveness. In a study where seven Hogan Personality Inventory scales were combined with measures of the Big Five, the different measures covered largely the same field of trait semantics (Hogan and Hogan, 2002).

Cloninger (2009) listed 36 often cited, well-known personality theorists, such as James, Freud, Maslow, Kelly, Bandura, and Buss. This historical flock of theorists represent different theoretical perspectives from which many differential dimensions have been generated, such as achievement motivation, field dependence (dealing with information), sensation seeking, dispositional optimism, and locus of control (control over the environment). Some of those topics have faded away, and others were expected or even predicted to disappear from the stage, but reappear sometimes vigorously (cf Pervin, 1990). It is important to understand that the Big Five system does not tell the whole story of personality. It is a comprehensive and differential descriptive system that is understood to form a rich starting point for further specification. Issues, such as change, unconscious processes, and heredity of traits are not contained in the lexical trait vocabularies. Yet trait taxonomy helps to understand semantically the different features of those issues; moreover, empirical research around those issues often make use of instruments based on the Big Five.

Psycholexically Based Models Departing from the Big Five

Lexical Big Five models across languages are far from identical. Especially since Goldberg (1981) talked first of the 'Big Five,' lexical studies in other languages were usually undertaken with the expectation of the Big Five in mind. While names of Big Five factors showed variations on the same themes, explicit diversions occurred with different psycholexical ideas concerning coverage of the trait domain, different procedures in arriving at a full trait-vocabulary, and different cultures or languages.

De Raad and Szirmák (1994) reported on a six-factor trait structure for Hungarian, that included a factor Integrity (*veracious*, *just*, *trustworthy*) beyond the Big Five. A similar factor (called Trustworthiness) was found in Italian by Di Blas and Forzi (1999), and in Korean (Hahn et al., 1999). Ashton et al. (2004) found further support for this factor in French, Dutch, Italian, and Polish and thus proposed Honesty (capturing Integrity and Trustworthiness) to be the Sixth factor of personality. Almagor et al. (1995) untightened the rules of selection of trait terms that aimed to exclude evaluative terms and state terms. Following an unrestricted approach, they demonstrated the possibility of seven broad personality factors that included versions of the Big Five, and two additional factors called Negative Valence (e.g., *dangerous*, *vicious*, *malicious*) and Positive Valence (e.g., *refined*, *noble*, *brilliant*) (cf Benet and Waller, 1995). De Raad and Barelds (2008) virtually exhausted the (Dutch) lexicon using trait-descriptive adjectives, nouns, verbs, adverbs, and standard expressions, thus forming an extremely long list of 2365 trait-descriptive items. Principal Components Analysis yielded an eight-factorial structure in self- and other ratings, that included the Big Five or versions thereof, and three additional factors namely Virtue, Competence, and Hedonism.

In a psycholexical research program in Filipino (Church et al., 1997), seven factors were identified in two studies. Four of the factors were quite comparable to Big Five factors. In addition, Negative Valence was identified, Self-Assurance (most similar to Big Five Neuroticism), and Temperamentalness, which had a complex relation to different Big Five factors. A lexically based structure, most deviating from the Big Five was found in Hindi (Singh et al., 2013), where six factors were identified in self- and peer ratings. Three smaller factors had some relationship with the Big Five Extraversion, Agreeableness, and Conscientiousness. The larger three, however, had complex relations to Big Five factors, but were recognized especially to reflect a conception of personality that dates back far into history, but is still alive in everyday interactions, the so-called *triguna*.

Cross-Cultural Findings

Cross-cultural psychologists have often endorsed the universality of psychological characteristics, as can be seen in cross-cultural studies on dimensions such as achievement motivation, anxiety, and authoritarianism (Church, 2000). Mayer et al. (2011) studied the possibility of universality of personality conceptions in cultural traditions as different as Confucianism, Buddhism, and Judaism. They concluded that judging personality was an important aspect of all those different traditions.

The Big Five model has found cross-cultural support generally in two ways; these two are often linked to the distinction between emic and etic (Berry, 1969). The emic approach in this case aims at finding a trait structure that best summarizes the trait domain of a particular language or culture. Linked to this first approach has been the repeated finding that independent psycholexical studies in various

Western languages led to the Big Five structure. Studies comparing lexical Big Five structures from different languages and leading to the conclusion that the Big Five is replicable across those languages are often done through the analysis of the contents of the structures. Notwithstanding the recurrence of the Big Five, the structure is reproduced better in some languages than in others (Saucier and Goldberg, 2001).

The etic approach is typically followed by constructing a trait system (a questionnaire) in one language, and translate it to and apply it in another language or culture. Studies with the Five Factor Personality Inventory (FFPI; Hendriks et al., 1999) in 13 languages gave evidence of the FFPI to be a reliably and valid measure in a large variety of countries (Hendriks et al., 2003). Similarly, studies with the NEO-PI-R (Costa and McCrae, 1992c) showed replicability of the five factors in most cultures (McCrae et al., 2005). A lesson drawn by Allik et al. (2013) from studies like these was that it is apparently easy to transcend language barriers using such personality instruments. Notwithstanding such excellent cross-cultural findings, criticism has been expressed repeatedly, especially from the side of the cross-cultural methodologists (e.g., Berry et al., 2002). An important issue is that translated instruments tend to be relatively insensitive to detect individual differences of interest in the target language. Ashton and Lee (2001), for example, found that certain FFM-Openness to Experience facets were not well applicable in many Asian samples.

Cheung et al. (2011) proposed an approach that would combine etic and emic into an integrated approach. Such a combined approach can be found in De Raad et al. (2010), who compared 14 independently developed trait structures. For each structure, the starting-point was taken in the lexicon of the pertaining language. The trait terms of that language were used to obtain ratings, and those were factored to arrive at a trait structure meaningful to that language. The 14 taxonomies were pairwise compared, after finding a common part of the factor structures on the basis of acceptable translations of items into the languages of a pair. On average, the results indicated that not five, but rather three factors were well replicable across the languages under study. Beyond the first three factors (with traits that are typical of Extraversion, Agreeableness, and Conscientiousness, respectively) the equivalence of factors across languages tend to divert.

Dimensional, Circular, and Hierarchical Characteristics of the Big Five Model

Studies on the Big Five, but also studies of other trait models, are usually based on analyses that produce orthogonal, or independent factors. That structure is imposed upon the data through the application of the psychometric technique, most typically Principal Components Analysis. The underlying ideal that is pursued is the *simple structure*, in which each trait variable loads substantially on only one factor and the loadings on the other factors are close to zero. In such a conception, the factors are rather narrow in meaning, with high-internal consistency and a rather fixed angular position in the trait space (Hofstee, 2003). A simple conception of Big Five trait factors is thus the representation with five independent dimensions. Although there are trait variables like that, in practice, however, the vast majority of trait variables tend to have substantial loadings on two factors.

Because of the fact that clear and meaningful traits usually load on one or two factors, traits are well represented in two-dimensional arrangements. When all variables from a set of variables that load on a particular pair of factors are located in such a two-dimensional arrangement, using the pair of loading as coordinate values, a circular configuration occurs, called a circumplex (Guttman, 1954). In contrast to a simple structure understanding of traits with narrow meaning, a circular arrangement of traits demonstrates the breadth of semantic coverage through the dispersion of the trait items along the circle (cf Gurtman, 1997). Hofstee et al. (1992) integrated the Big Five model and the circumplex understanding of traits in a comprehensive representation.

Several researchers in the field of personality have emphasized the hierarchical representation (e.g., Cattell, 1947; Eysenck, 1970) with relations among traits running from more abstract to more specific. The behavior 'laughing at jokes' is more specific than 'being lively,' which in turn is more specific than 'Extraversion.' Most typically, hierarchies are studied in two ways, a bottom-up approach and a top-down approach (see Goldberg, 2006). The more traditional ones are the bottom-up hierarchies in which one starts with individual items, which are then clustered into semantic groups; those clusters in turn are clustered into larger groups of traits, until the highest level is reached. The more recently developed top-down approach has been applied in Principal Components analyses of traits, in which structures are considered with one factor or component, two factors, up to a structure with the maximum number of factors considered relevant. The factors from adjacent levels of factor extraction are then correlated yielding the hierarchical configuration (cf De Raad and Szirmák, 1994; Mlačić and Ostendorf, 2005; Zuckerman et al., 1988).

Nomological Network and Practical Usefulness of the Big Five

The Big Five model has led to the development of quite a number of instruments to measure the five dimensions, including the NEO-PI-R and the FFPI. An enormous number of studies have been performed relating the Big Five with other concepts. Of special interest here are the relationships that have been demonstrated between all or some of the Big Five factors and a great variety of criteria. Zeidner (2009), for example, concluded that Conscientiousness and Emotional Stability, and to a lesser extent also Openness to Experience, are predictors of learning and achievement (cf De Raad and Schouwenburg, 1996). Barrick and Mount (1991) studied the relationship of the Big Five factors with job performance. The strongest predictor across a range of occupational groups was Conscientiousness. Maladaptive variants of the Big Five (minus Openness to Experience) are argued to play an important role in relation to abnormal personality (Watson et al., 2008), especially also as a more adequate replacement of the categorical DSM-IV scheme. Roberts et al. (2007) reviewed the role of personality traits as predictors of

important life outcomes. Conscientiousness, Extraversion, and Emotional Stability were found to be related to longevity, and Neuroticism is a strong and consistent predictor of relationship dissatisfaction and divorce.

The availability of the Big Five conceptual system, especially in its integration with the circumplex configuration, forms a fine-grained system that can be used for the purpose of understanding the semantics of other conceptual systems. De Raad and Doddema-Winsemius (1999) used it successfully for the classification of a list of instincts. Kohnstamm et al. (1998) found that virtually all free descriptions of persons in their study could be classified in the Big Five system. Passakos and De Raad (2009) described the traits of the many characters in Homer's Iliad using the Big Five system. Johnson and Ostendorf (1993) used the facet-nuances of the Big Five system to clarify controversies regarding the interpretation of the Big Five Factors.

As described in previous paragraphs, the Big Five conceptual framework has played and still does play an important role in theory building. Digman's (1997) study is a good example where Big Five factors from various studies were factored leading to two higher order factors, which, in turn, permit linkage to various existing theoretical systems in personality (cf DeYoung, 2006). The Big Five as part of a hierarchy of factors enhances the cross-cultural understanding of personality traits (De Raad et al., 2014; DeYoung, 2006).

There is still growing potential for the Big Five as organizer in research and in applied fields. Further research on the separate Big Five factors is needed to sharpen their understanding. Critical approaches to the Big Five may turn existing issues resolved. Especially new psycholexical studies with a model, like the Big Five in mind in many nonexplored languages, may render certain concepts more important in the one culture than in the other. It also leads to new tests of the validity of the Big Five system.

Future Directions

The development of the Big Five model into a possible Big Five theory, outlined above, shows that, after a century of meandering, personality psychology, or at least the trait approach, is close in finding a common stream bed. Although the Big Five model has been criticized both within the lexical approach and from other competing approaches, and alternatives were offered, it was successful in surviving as a shared model in personality psychology for more than two decades. That kind of a consensus was rare in the last 100 years since personality psychology emerged as a scientific discipline. The Big Five successfully organized a scattered field pointed to the natural language as a source for the total space of personality variables, generated numerous publications and was linked to many important outcomes. The Big Five model was adopted by researchers and theorists, not just from the trait approach, and was proven nomologically and practically useful.

However, to be considered a theory in the full sense of the word, there are many bridges the Big Five has to cross. The most important pertain to the various levels of the hierarchy and to the consensus on facets and higher order factors, to longitudinal studies of personality in a life cycle perspective, to the brain-behavior connection, to the link between personality dynamics, traits and outcomes, etc. Last, but not least, since the Big Five is mostly the product of the Westernized world, future research should make an important step in conducting the large-scale studies in other cultures before reaching the conclusion on the universality of the model and the prospective theory.

See also: Agreeableness; Conscientiousness; Cross-Cultural Psychology; Cultural Psychology; Extraversion; Factor Analysis and Latent Variable Models in Personality Psychology; Five Factor Model of Personality, Assessment of; Five Factor Model of Personality, Universality of; Honesty and Humility; Indigenous Psychology; Neuroticism; Openness to Experience; Personality Assessment: Overview; Personality Assessment; Personality Changes During Adolescence Across Cultures; Personality Differences and Development: Genetic and Environmental Contributions; Personality and Adaptive Behaviors; Personality, Biological Models of; Personality, Evolutionary Models of; Personality, Trait Models of; Personality: Historical and Conceptual Perspectives; Sensation Seeking: Behavioral Expressions and Biosocial Bases.

Bibliography

Adelstein, J.S., Shehzad, Z., Mennes, M., DeYoung, C.G., Zuo, X.-N., Kelly, C., Margulies, D.S., Bloomfield, A., Gray, J.R., Castellanos, F.X., Milham, M.P., 2011. Personality is reflected in the brain's intrinsic functional architecture. PLoS ONE 6, e27633.

Allik, J., Realo, A., McCrae, R.R., 2013. Universality of the five-factor model of personality. In: Widiger, T.A., Costa Jr., (Eds.), Personality Disorders and the Five-Factor Model of Personality. American Psychological Association, pp. 61–74.

Allport, G.W., Odbert, H.S., 1936. Trait-names. A psycho-lexical study. Psychological Monographs 47 (Whole No. 211).

Almagor, M., Tellegen, A., Waller, N., 1995. The Big Seven model: a cross-cultural replication and further exploration of the basic dimensions of natural language of trait descriptions. Journal of Personality and Social Psychology 69, 300–307.

Ashton, M.C., Lee, K., 2001. A theoretical basis for the major dimensions of personality. European Journal of Personality 15, 327–353.

Ashton, M.C., Lee, K., 2007. Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. Personality and Social Psychology Review 11, 150–166.

Ashton, M.C., Lee, K., Perugini, M., Szarota, P., De Vries, R.E., Di Blas, L., Boies, K., De Raad, B., 2004. A six-factor structure of personality-descriptive adjectives: solutions from psycholexical studies in seven languages. Journal of Personality and Social Psychology 86, 356–366.

Austin, J.L., 1970. Philosophical Papers. Oxford University Press, London.

Barrick, M.R., Mount, M.K., 1991. The Big Five personality dimensions and job performance: a meta-analysis. Personnel Psychology 44, 1–26.

Benet, V., Waller, N.G., 1995. The Big Seven factor model of personality description: evidence for its cross-cultural generality in a Spanish sample. Journal of Personality and Social Psychology 69, 701–718.

Berry, J.W., 1969. On cross-cultural comparability. International Journal of Psychology 119–128.

Berry, J.W., Poortinga, Y.H., Segall, M.H., Dasen, P.R., 2002. Cross-cultural Psychology: Research and Applications. Cambridge University Press, New York.

Block, J., 1995. A contrarian view of the five-factor approach to personality description. Psychological Bulletin 117, 187–215.

Block, J., 2010. The five factor framing of personality and beyond: some ruminations. Psychological Inquiry 21, 2–25.

Buss, D.M., Hawley, P.H., 2010. The Evolution of Personality and Individual Differences. Oxford University Press, New York.

Cattell, R.B., 1943. The description of personality: basic traits resolved into clusters. Journal of Abnormal and Social Psychology 38, 476–507.

Cattell, R.B., 1947. Confirmation and clarification of primary personality factors. Psychometrika 12, 197–220.
Cheung, F.M., Van de Vijver, F.J.R., Leong, F.T.L., 2011. Toward a new approach to the study of personality. American Psychologist 66, 593–603.
Church, A.T., 2000. Culture and personality: towards an integrated cultural trait psychology. Journal of Personality 68, 651–703.
Church, A.T., Reyes, J.A.S., Katigbak, M.S., Grimm, S.D., 1997. Filipino personality structure and the Big Five Model: a lexical approach. Journal of Personality 65, 477–528.
Cloninger, S., 2009. Conceptual issues in personality theory. In: Corr, P.J., Matthews, G. (Eds.), The Cambridge Handbook of Personality Psychology. Cambridge University Press, Cambridge, pp. 3–26.
Coan, R.W., 1974. The Optimal Personality. Columbia University Press.
Costa Jr., P.T., McCrae, R.R., 1985. The NEO Personality Inventory Manual. Psychological Assessment Resources, Odessa, Fl.
Costa Jr., P.T., McCrae, R.R., 1992a. Normal personality assessment in clinical practice: the NEO personality inventory. Psychological Assessment 4, 5–1 3.
Costa Jr., P.T., McCrae, R.R., 1992b. Four ways five factors are basic. Personality and Individual Differences 13, 653–665.
Costa Jr., P.T., McCrae, R.R., 1992c. Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual. Psychological Assessment Resources, Odessa, Fl.
De Raad, B., 2000. The Big Five Personality Factors: The Psycholexical Approach to Personality. Hogrefe & Huber Publishers, Göttingen.
De Raad, B., Barelds, D.P.H., 2008. A new taxonomy of Dutch personality traits base on a comprehensive and unrestricted list of descriptors. Journal of Personality and Social Psychology 94, 347–364.
De Raad, B., Doddema-Winsemius, M., 1999. Instincts and personality. Personality and Individual Differences 27, 293–305.
De Raad, B., Szirmák, Z., 1994. The search for the "Big Five" in a non-Indo-European language: the Hungarian trait structure and its relationship to the *EPQ* and the. PTS. European Review of Applied Psychology 44, 17–24.
De Raad, B., Van Heck, G.L. (Eds.), 1994. The fifth of the Big Five. European Journal of Personality, vol. 8 (special issue).
De Raad, B., Barelds, D.P.H., Levert, E., Ostendorf, F., Mlačić, B., Di Blas, L., et al., 2010. Only three factors of personality description are fully replicable across languages: a comparison of 14 trait taxonomies. Journal of Personality and Social Psychology 98, 160–173.
De Raad, B., Barelds, D.P.H., Timmerman, M.E., De Roover, K., Mlacic, B., Church, A.T., 2014. Towards a pan-cultural personality structure: input from 11 psycho-lexical studies. European Journal of Personality 28, 497–510.
De Raad, B., Schouwenburg, H.C., 1996. Personality in learning and education: a review. European Journal of Personality 10, 303–336.
DeYoung, C.G., 2006. Higher-order factors of the Big Five in a multi-informant sample. Journal of Personality and Social Psychology 91, 1138–1151.
DeYoung, C.G., 2010a. Toward a theory of the Big Five. Psychological Inquiry 21, 26–33.
DeYoung, C.G., 2010b. Personality neuroscience and the biology of traits. Social and Personality Psychology Compass 4, 1165–1180.
DeYoung, C.G., Quilty, L.C., Peterson, J.B., 2007. Between facets and domains: 10 aspects of the Big Five. Journal of Personality and Social Psychology 93, 880–896.
DeYoung, C.G., Shamosh, N.A., Green, A.E., Braver, T.S., Gray, J.R., 2009. Intellect as distinct from openness: differences revealed by fMRI of working memory. Journal of Personality and Social Psychology 97, 883–892.
DeYoung, C.G., Hirsh, J.B., Shane, M.S., Papademetris, X., Rajeevan, N., Gray, J.R., 2010. Testing predictions from personality neuroscience: brain structure and the Big Five. Psychological Science 21, 820–828.
Di Blas, L., Forzi, M., 1999. Refining a descriptive structure of personality attributes in the Italian language: the abridged Big Three circumplex structure. Journal of Personality and Social Psychology 76, 451–481.
Digman, J.M., 1997. Higher-order factors of the Big Five. Journal of Personality and Social Psychology 73, 1246–1256.
Digman, J.M., Inouye, J., 1986. Further specification of the five robust factors of personality. Journal of Personality and Social Psychology 50, 116–123.
Eysenck, H.J., 1967. The Biological Basis of Personality. Charles C. Thomas, Springfield, Il.
Eysenck, H.J., 1970. The Structure of Human Personality, third ed. Methuen, London.
Eysenck, H.J., 1991. Dimensions of personality. 16, 5 or 3? – criteria for a taxonomic paradigm. Personality and Individual Differences 12, 773–790.
Eysenck, H.J., 1992. Four ways five factors are not basic. Personality and Individual Differences 13, 667–673.
Galton, F., 1884. Measurement of character. Fortnightly Review 36, 179–185.
Goldberg, L.R., 1981. Language and individual differences. The search for universals in personality lexicons. In: Wheeler, L. (Ed.), Review of Personality and Social Psychology, vol. 2. Sage Publications, Beverly Hills, CA, str. 141–165.
Goldberg, L.R., 1990. An alternative "description of personality". The Big-Five factor structure. Journal of Personality and Social Psychology 59, 1216–1229.
Goldberg, L.R., 1993. The structure of phenotypic personality traits. American Psychologist 48, 26–34.
Goldberg, L.R., 2006. Doing it all bass-ackwards: the development of hierarchical factor structures from the top down. Journal of Research in Personality 40, 347–358.
Goldberg, L.R., 2014. International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences. http://ipip.ori.org/. Internet Web Site (accessed 18.02.14.).
Graziano, W.G., Eisenberg, N., 1997. Agreeableness: a dimension of personality. In: Hogan, R., Johnson, J.A., Briggs, S.R. (Eds.), Handbook of Personality Psychology. Academic Press, San Diego, pp. 795–824.
Griffin, B., Hesketh, B., 2004. Why openness to experience is not a good predictor of job performance. International Journal of Selection and Assessment 12, 243–251.
Gurtman, M.B., 1997. Studying personality traits: the circular way. In: Plutchik, R., Conte, H.R. (Eds.), Circumplex Models of Personality and Emotions, Washington, DC, pp. 81–102.
Guttman, L., 1954. A new approach to factor analysis: the radex. In: Lazarsfeld, P.F. (Ed.), Mathematical Thinking in the Social Sciences. Free Pres, Glencoe, IL, pp. 258–348.
Hahn, D.-W., Lee, K., Ashton, M.C., 1999. A factor analysis of the most frequently used Korean personality trait adjectives. European Journal of Personality 13, 261–282.
Hendriks, A.A.J., Perugini, M., Angleitner, A., et al., 2003. The five-factor personality inventory: cross-cultural generalizability across 13 countries. European Journal of Personality 17, 347–373.
Hendriks, A.A.J., Hofstee, W.K.B., De Raad, B., 1999. The five-factor personality inventory (FFPI). Personality and Individual Differences 27, 307–325.
Hofstee, W.K.B., 2003. Structures of personality traits. In: Millon, T., Lerner, M.J. (Eds.), Handbook of Personality, vol. 5. John Wiley & Sons, Inc, Hoboken, NJ, pp. 231–254.
Hofstee, W.K.B., De Raad, B., Goldberg, L.R., 1992. Integration of the Big Five and circumplex approaches to trait structure. Journal of Personality and Social Psychology 63.
Hogan, R., Hogan, J., 2002. The hogan personality inventory. In: De Raad, B., Perugini, M. (Eds.), Big Five Assessment. Hogrefe & Huber Publishers, Göttingen, pp. 329–351.
Hogan, R., 1983. A socioanalytic theory of personality. In: Page, M.M. (Ed.), 1982 Nebraska Symposium on Motivation. University of Nebraska Press, Lincoln, pp. 55–89.
Hogan, J., Ones, D.S., 1997. Conscientiousness and integrity at work. In: Hogan, R., Johnson, J.A., Briggs, S.R. (Eds.), Handbook of Personality Psychology. Academic Press, San Diego, pp. 849–870.
John, O.P., 1990. The "Big Five" factor taxonomy: dimensions of personality in the natural language and in questionnaires. In: Pervin, L.A. (Ed.), Handbook of Personality: Theory and Research. The Guilford University Press, New York, pp. 66–100.
John, O.P., Angleitner, A., Ostendorf, F., 1988. The lexical approach to personality. A historical review of trait taxonomic research. European Journal of Personality 2, 171–203.
John, O.P., Naumann, L.P., 2010. Surviving two critiques by block? The resilient Big Five have emerged as the paradigm for personality trait psychology. Psychological Inquiry 21, 44–49.
John, O.P., Naumann, L.P., Soto, C.J., 2008. Paradigm shift to the integrative Big Five trait taxonomy: history: measurement, and conceptual issues. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research. Guilford Press, New York, pp. 114–158.
John, O.P., Srivastava, S., 1999. The Big Five trait taxonomy: history, measurement, and theoretical perspectives. In: Pervin, L.A., John, O.P. (Eds.), Handbook of Personality Psychology: Theory and Research, second ed. Guilford Press, New York, pp. 102–138.
Johnson, J.A., Ostendorf, F., 1993. Clarification of the five-factor model with the abridged big five dimensional circumplex. Journal of Personality and Social Psychology 65, 563–576.
Kohnstamm, G.A., Halverson Jr., C.F., Mervielde, I., Havill, V.L., 1998. Analyzing parental free descriptions of child personality. In: Kohnstamm, G.A., Halverson Jr., C.F., Mervield, I., Havill, V.L. (Eds.), Parental Descriptions of Child Personality: Developmental Antecedents of the Big Five? Erlbaum, Mahwah, NJ, pp. 1–9.

Loevinger, J., 1976. Ego Development. Jossey-Bass, San Francisco.
Mayer, J.D., Lin, S.C., Korogodsky, M., 2011. Exploring the universality of personality judgments: evidence from the great transformation (1000 BCE-200 BCE). Review of General Psychology 15, 65–76.
McCrae, R.R., 1990. How well is openness represented in natural languages? European Journal of Personality 4, 119–129.
McCrae, R.R., 2010. The place of the FFM in personality psychology. Psychological Inquiry 21, 57–64.
McCrae, R.R., Costa Jr., P.T., 2008. The five-factor theory of personality. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research. Guilford, New York, pp. 159–181.
McCrae, R.R., Terracciano, A., et al., 2005. Universal features of personality traits from the observer's perspective: data from 50 cultures. Journal of Personality and Social Psychology 88, 547–561.
Miller, G.A., 1991. The Science of Words. Scientific American Library, New York.
Mlačić, B., Milas, G. Personality changes during adolescence and cultural variations. In: Wright, J.D. (Ed.), The International Encyclopedia of Social and Behavioral Sciences, second ed. Elsevier, Oxford, in press.
Mlačić, B., Ostendorf, F., 2005. Taxonomy and structure of Croatian personality-descriptive adjectives. European Journal of Personality 19, 117–152.
Mulder, B.F., 2006. Openheid Voor Ervaringen, Lexical Benaderd (Openness to Experience, Lexically Approached). thesis. University of Groningen.
Norman, W.T., 1963. Toward an adequate taxonomy of personality attributes. Replicated factor structure in peer nomination personality ratings. Journal of Abnormal and Social Psychology 66 (6), 574–583.
Passakos, C.G., De Raad, B., 2009. Ancient personality: trait attributions to characters in Homer's Iliad. Ancient Narrative 7, 75–95.
Peabody, D., De Raad, B., 2002. The substantive nature of psycholexical personality factors: a comparison across languages. Journal of Personality and Social Psychology 83, 983–997.
Pervin, L.A., 1990. A brief history of modern personality theory. In: Pervin, L.A. (Ed.), Handbook of Personality: Theory and Research. The Guilford Press, New York, pp. 3–18.
Roberts, B.W., Kuncel, N.R., Shiner, R., Caspi, A., Goldberg, L.R., 2007. The power of personality: the comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. Perspectives on Psychological Science 2, 313–345.
Rotter, J.B., 1954. Social Learning and Clinical Psychology. Prentice-Hall.
Saucier, G., 2003. An alternative multi-language structure for personality attributes. European Journal of Personality 17, 179–205.
Saucier, G., 2009. What are the most important dimensions of personality? Evidence from studies of descriptors in diverse languages. Social and Personality Psychology Compass 3 (4), 620–637.
Saucier, G., Goldberg, L.R., 2001. Lexical studies of indigenous personality factors: premises, products, and prospects. Journal of Personality 69, 847–879.
Saucier, G., Goldberg, L.R., 1996. The language of personality: lexical perspectives on the five-factor model. In: Wiggins, J.S. (Ed.), The Five-Factor Model of Personality. Theoretical Perspectives. The Guilford Press, New York, London, pp. 21–50.
Singh, J.K., Misra, G., De Raad, B., 2013. Personality structure in the trait lexicon of Hindi, a major language spoken in India. European Journal of Personality 27, 605–620.
Srivastava, S., 2010. The five-factor model describes the structure of social perceptions. Psychological Inquiry 21, 69–75.
Stelmack, R.M., 1997. Toward a paradigm in personality. Comments on Eysenck's view. Journal of Personality and Social Psychology 73 (6), 1238–1241.
Strelau, J., 1983. Temperament-Personality-Activity. Academic Press, London.
Themerson, S., 1974. Logic, Labels, and Flesh. Gaberbocchus, London.
Watson, D., Clark, L.A., Chmielewski, M., 2008. Structures of personality and their relevance to psychopathology: II. Further articulation of a comprehensive unified trait structure. Journal of Personality 76, 1545–1585.
Wiggins, J.S., 1973. Personality and Prediction: Principles of Personality Assessment. Addison-Wesley, Reading, MA.
Zeidner, M., 2009. Personality in educational psychology. In: Corr, P.J., Matthews, G. (Eds.), The Cambridge Handbook of Personality Psychology. Cambridge University Press, Cambridge, UK, pp. 733–747.
Zuckerman, M., 1984. Sensation seeking: a comparative approach to a human trait. Behavioral and Brain Sciences 7, 413–471.
Zuckerman, M., 2002. Zuckerman-Kuhlman personality questionnaire (ZKPQ): an alternative five-factorial model. In: De Raad, B., Perugini, M. (Eds.), Big Five Assessment. Hogrefe & Huber Publishers, Goettingen, pp. 377–396.
Zuckerman, M., Kuhlman, D.M., Camac, C., 1988. What lies beyond E and N? Factor analyses of scales believed to measure basic dimensions of personality. Journal of Personality and Social Psychology 54, 96–107.
Zuckerman, M., Kuhlman, D.M., Joireman, J., Teta, P., Kraft, M., 1993. A comparison of three structural models of personality: the Big Three, the Big Five, and the alternative five. Journal of Personality and Social Psychology 65, 757–768.

Big Man, Anthropology of

Rena Lederman, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Referring to achieved leadership (often contrasted with inherited 'chiefly' rank), 'big man' has come to stand for a type of political organization closely associated with, although not limited to, Melanesia. This article locates the term's early uses, tracks developing analyses, and notes the shifting place of these studies vis-à-vis larger scholarly arguments. Contemporary trends include waning interest in classic comparativist typology building in favor of historically and culturally situated understandings of personhood and social action: attentive to gendered and emergent class relations, to nation-making and governance dilemmas, and to the impacts of international agencies and corporate interests on local communities and environments.

Referring to an informal style of achieved leadership, the term *big man* has come to stand for a type of polity distinguished, for example, from political systems with formally elected offices or inherited ranks and statuses. It is closely associated with, although not limited to, the ethnography of Melanesia (islands in the southwestern Pacific located just north and east of Australia, for example Papua New Guinea (PNG), Solomon Islands, Vanuatu, and Fiji). While prominent in regional ethnography, political anthropology, and social evolutionary theorizing from the 1950s through the 1980s, this construct has become less visible as a focus of research in recent years. Contemporary work reflects a methodological turn away from typology building (and the functional and evolutionary/developmental theories it served). Emphasis has shifted toward historically and culturally situated understandings of personhood and social action: attentive to gendered and emergent class relations, to nation-making and governance dilemmas, and the impacts of international agencies and corporate interests on local communities and environments.

Early Work and Developing Analyses

Big man is the Anglicization of a descriptive phrase *bikpela man* (and variants) – meaning 'prominent man' – common in Melanesian lingua francas (like *Tok Pisin* and *Bislama*) and analogous phrases or terms of reference in that region's hundreds of local languages, e.g., *nema asori* on Tanna (Lindstrom, 1981: p. 902) and *ol onda* in the Mendi Valley (Lederman, 1986b). The term was widely adopted in post–World War II (WWII) Melanesianist anthropology (e.g., Oliver, 1955; Read, 1959; Meggitt, 1967) to refer to male leaders whose political influence is achieved by means of public oratory, informal persuasion, and the skillful conduct of both private and public wealth exchanges. The anthropology of big men comprises both regional ethnography (the sociocultural interpretation of case materials derived from extended field research) and – as this ethnography is puzzling from a comparative perspective – theoretical debate.

Ethnographic Foundations

While a variety of political systems – involving both inherited and achieved positions, and both ritual performances and political/economic displays – have been observed in Melanesia (as throughout the island Pacific) from the early colonial period to the present day, the big man style of leadership is a notable feature of political life in highland New Guinea (comprising independent PNG and the Indonesian province of Papua, also known as West Papua), Bougainville, parts of the Solomon Islands (e.g., Guadalcanal), and elsewhere. Highland New Guinea in particular is distinctive, in anthropology's cross-cultural ethnographic record, for high population densities and communities loosely organized in terms of kinship groups (e.g., 'clans' and 'tribes' associated with named territories). Their horticultural economies are characteristically based on intensive tuber (sweet potato, taro) cultivation and pig raising. Big man societies intrigued anthropologists and theorists of development for opening up the possibility that economic intensification might be possible in the absence of formal political hierarchies (see Section Political Typologies and Developmental Puzzles below).

Ethnographic research began in earnest in the New Guinea highlands only after WWII, when colonial authority was extended throughout the area and tribal warfare suppressed. It continued apace during the 1970s – when most parts of Melanesia made the political transition to independence – and, with shifting analytical and topical concerns, through the final decades of the twentieth century into the new millennium. Early on, it became evident that highland Melanesian socioeconomic life was relatively egalitarian; that is, access to garden land was universal (a flexible concomitant of kin relationships) and systematic differences in access to subsistence means, spouses, and valuables were muted or absent.

Direct participation in clan events was a more or less exclusively male prerogative and a key means by which men advanced their local and regional social standing; however, it was not obligatory. Even for men, collective clan interests did not necessarily take precedence over personal relationships with kin and other exchange partners, relationships in which

women were also involved. Finally, communities did not typically have institutionalized decision-making councils, nor did political or administrative offices exist to be filled. Instead, men earned leadership status by organizing events held in the names of clans and tribal alliances: they motivated their communities to participate by means of oratory, informal politicking, the cultivation of large personal networks of exchange partnerships, and their own personal and household production work.

Despite the lack of formal political roles and structures, during precolonial and early colonial times clans affiliated with one another in regional alliances for warfare and for elaborate ritual performances associated with initiations and fertility cults (some of which involved multiyear planning). People also frequently came together to sponsor competitive displays and exchanges of customary valuables like pigs and pearl shells: festive events in which thousands participated in a coordinated fashion as recipients and donors. Periodic exchange festivals of different scales and degrees of social and political complexity have driven intensive garden production for generations in the highlands and some other parts of Melanesia.

During the colonial period and – after PNG independence in 1975 – continuing into the present day, exchange events also came to incorporate money (e.g., PNG national currency, *kina*) and nontraditional goods of all sorts (e.g., cows, trucks, and cartons of beer). The desire to participate in these events even helped to prompt early participation in wage work and commercial endeavors. Held to mark important life cycle events like funerals and marriages and to constitute or reorder regional alliances between groups, exchange festivals require skillful organizing and long-term planning. They also require the intensive cultivation of sweet potatoes to feed large herds of pigs, which, in turn, are slaughtered to mark important events (not simply for household consumption) and also invested as wealth in the elaboration of personal exchange relationships.

It is in these contexts that the activities of the big man have been critical. Indeed, investigations into the dynamics of political agency associated with big manship were a centerpiece in the matrix of anthropological investigations into gendered social and political forms and, relatedly, into indigenous concepts of personhood and of the morality of social action.

Early in Highlands research, Kenneth Read (1959) noted a tension between culturally sanctioned assertions of personal autonomy and of group purpose. In Euro-American cultures, an apparently analogous tension counterposes naturalized, 'individual' interests to morally obligatory collective interests (e.g., the state). In contrast, Melanesian cultures implicate both differentiating and collectivizing interests as different but equally sociable possibilities inhering in a distinctive (and distinctively gendered) concept of personhood. Personal autonomy is enacted in the creative elaboration of networks of kin and exchange partners (affines, friends). Each person's social network makes possible the enactment of interests that converge partially – and may also regularly conflict – with the collective (and, in Highland New Guinea, typically male) projects of clans. While the composition and size of exchange networks reflect the contingencies of personal effort (as in the case of friendship circles in Western societies), clan affiliation and rights to land are givens. One generally belongs to one's parent's group, inheriting homesteads and garden land through one's father or mother, although conventions vary across Melanesia and involve considerable flexibility on account of a pervasive recognition of multiple bases for membership and land claims (e.g., Lederman, 1986b).

Against that background, big men make their names not simply by mobilizing wealth for personal network ends – like staging impressive funerary wealth distributions and generously supporting their exchange partners when the latter stage events – but also by successfully orienting their own and their clansmen's respective network interests to collective projects. They persuade fellow clan members to organize their personal, often divergent network affairs to advance common clan events. By enabling events like large-scale wealth distributions between tribal and clan alliances – strengthening the reputations of these named collectivities by demonstrating their organizational capacities – big men augment their group's reputation ('name') in the very act of augmenting their own renown.

In all this, their influence is personal and ephemeral relative to that in societies with inherited rank or with formal councils. This absence of structurally reproduced power has given Melanesian communities a reputation for egalitarianism (among men, if not between men and women). Men with leadership ambitions work to develop personal access to resources within and outside their communities. They achieve fame and influence by using local and regional social networks as bases for organizing collective wealth prestations. By means of both public oratory and private persuasion, big men work to add a certain collective value to actions that would otherwise be everyday and personal.

These events therefore make and remake clans and tribal alliances, which were means for collective defense in the precolonial era. In recent generations, they have been important bases for launching commercial endeavors – for example, cattle herding, coffee plantations – and have also constituted social springboards for provincial and national political careers.

Political Typologies and Developmental Puzzles

In post–WWII anthropology, ethnographic accounts from Melanesia and Polynesia were influential in modeling sociopolitical ideal types (respectively, big man societies and chiefdoms). Such typologies were central to theories of cultural evolution and economic development, of special interest to anthropological archaeologists (who reconstruct the long-term social histories of human populations from material evidence: see Section Anthropology of Big Manship in a Wider Field below) and also to social theorists both within and outside of sociocultural anthropology. For the most part, theories of social evolution assumed that less economically productive societies characterized by limited social differentiation and political decentralization tend to give way, over time, to more productive types characterized by elaborate divisions of labor and centralized, hierarchical political systems. They posited a series of functional interdependencies among variables like population density, technology, the organization of production, 'surplus' production, and sociopolitical stratification (see Lederman, 1986a). Against this background, the big man societies of Melanesia are intriguingly anomalous.

Marshall Sahlins's (1963) typological comparison of the Polynesian chief and Melanesian big man – perhaps the single most influential argument concerning Pacific polities – established the figure of the leader as key to arguments concerning their societies' historical fortunes. With Polynesian chiefs as a developmental standard, Sahlins highlighted the limited coercive power of the big man in mobilizing wealth for public prestations, and identified these limits in the refusal of his clansmen (understood as his political supporters in regional interclan prestations) to put up with his increasingly unreciprocal behavior.

Viewed from this perspective, evidence of large-scale, intensifying, and expanding systems of wealth exchange in Melanesia was puzzling. Ethnographically grounded arguments by Mervyn Meggitt (1967), Andrew Strathern (1971), and others with extensive fieldwork experience in highland New Guinea demonstrated that – while relations between men (dominant in wealth exchange) and women (mainstays of horticultural and pig production) were conflictual – relations between leaders and their fellow clansmen were not predominantly extractive (as Sahlins had supposed). Instead, big men were able to tap labor and other resources in groups other their own by means of personal exchange networks.

Meanwhile, a convergent line of research – bent more on understanding economic change than sociopolitical systems – developed a comparison between Melanesian exchanges and capitalist markets. Dubbing Melanesians and particularly their leaders as 'primitive capitalists,' this work emphasized apparent similarities between local orientations and Euro-American cultural values like personal achievement, competition, material wealth, and financial reinvestment. For example, Scarlett Epstein (1968) labeled Tolai leaders 'primitive capitalists'; and Ben Finney (1973) argued that these similarities culturally and psychologically 'preadapted' Goroka big men for capitalist development.

Indeed, during the colonial period, highland Papua New Guineans were notably entrepreneurial. They reinvested wealth in commercial enterprises, like trade stores, rather than simply spending cash incomes on consumer goods.

These lines of research came to an ambivalent resolution in the 1980s and 1990s. Specifically with respect to the anthropology of big men, Maurice Godelier (1986) centered attention on the structure of marriage exchanges in an argument meant to suggest that the big man type – with its emphasis on the manipulation of wealth – is not as typical of Melanesia as earlier ethnography implied. In communities where marriage conventions de-emphasize bride wealth (transfers of wealth for persons or their capacities) in favor of 'sister exchange' (transfers of persons for persons), he argued that one observes not big men but 'Great Men,' varieties of male prominence (notably initiation cult leaders) founded on ritual expertise. Some recent work in areas of Melanesia whose local political systems are conventionally characterized in terms of inherited rank have found Godelier's Great Man construct surprisingly apt (Barker, 2012).

Nevertheless, Godelier's position followed the classic pattern of using ethnographic cases to construct political economic types as elements in hypothetical developmental sequences. It made diverse cases comparable and capable of being organized as a progression but at the cost of assuming that these cases shared a common commitment to the maintenance of male collective (clan) interests. The Great Man idea also situated big manship on the frontier of market capitalism. However, this strategy of comparative analysis was already under siege ethnographically and theoretically by the mid/late 1980s.

Contemporary Trends

The relative decline in research on big manship as an ethnographic focus in its own right, during the 1990s and 2000s, is part of a long-standing disciplinary trend away from deploying ethnography for comparativist projects of developmental modeling and toward treating ethnography as a means for understanding culture histories.

Rethinking Anthropology, Rethinking Big Manship

Typology building and developmental theorizing, common in early to mid-twentieth century sociocultural anthropology (and political anthropology particularly), treated cultures as objectively identifiable, historically stable, socially boundable units of comparison. Over the past 30 or 40 years, however, a widespread theoretical reorientation in sociocultural anthropology from (as Clifford Geertz put it in his well-known 1983 essay, 'Blurred genres') a 'laws and instances ideal of explanation' modeled on the natural sciences 'toward a cases and interpretations' ideal drawn from the humanities undermined the very idea of bounded, stable, objective cultural units in favor of situated, reflexive accounts of cultural process more amenable to mutual critical translation than to positivist comparative analysis.

On the whole, Melanesian anthropology has not only reflected but also spearheaded these transformations in comparativism and the 'culture' concept (e.g., Roy Wagner's 1975 *Invention of Culture*). In this new staging, big manship has moved out of the analytical limelight, although not out of sight. Hardly relics, big men are everyday facts of life in the rural and urban communities and the provincial and national governments of twenty-first century Melanesia. The shifting array of contradictory values, beliefs, and rhetorics about big men are equally contemporary. They constitute postcolonial transformations of what Sahlins (1963) referred to as the 'Melanesian contradiction': the already delicate ethicopolitical balance that successful leaders had to maintain between reciprocity and exploitative self-advancement in rural communities a generation ago, now made orders of magnitude more difficult by its articulation with market logics and the politics of nation making.

During the 1990s, investigations into big manship became absorbed into a wide variety of academic, policy-oriented, and community-based projects concerned to understand and intervene in the 'emergent forms of life' (in the sense Michael Fischer developed in his 1999 *Annual Review* essay by that name) characterizing contemporary Melanesian experience. It therefore makes little sense to separate the anthropology of big men from these other literatures. Nevertheless, while these projects do not identify themselves with the 'anthropology of big men' as such, they do still acknowledge portions of that

foundational literature as historical grounding for a remarkably wide range of freshly conceptualized research foci, including gender relations and 'engendered' violence, socioeconomic transformations (e.g., emergent class relations and the social and environmental impacts of globalization and extractive industries), and the dilemmas of governance. The following paragraphs sample these themes.

Melanesian gender studies have long directed attention to divergent and conflictual perspectives and relations within communities. While 'sexual antagonism' was a well-developed theme by the 1970s, the past two decades' work historicizes regional experience and sets it in a global context (e.g., Jolly and Stewart, 2012). It renders the male-centered (and certainly male leader-centered) typification of whole cultures analytically unusable and big manship a decidedly qualified value (e.g., Lederman, 1986b, 1990, 1991; Godelier and Strathern, 1991). This research has attended to gender ideologies and gendered practices as grounds for making sense of conflicts between and among men and women, particularly in relation to emergent forms of female political and economic action and leadership, like the 'Wok Meri' movement in the Eastern Highlands (Sexton, 1990) and Kulka women's club intervention in a tribal war in the western Nebilyer Valley (Rumsey and Merlan, 1991). These gendered engagements are transforming local ideas about 'maleness' that long underpinned big manship and associated ideas about collective social forms (e.g., male initiation 'cults').

In recent years, anthropologists have also encountered mirror images of the big man that deploy its metaphoric value for new ends. For example, describing local perspectives on the past generation's social transformations among Dano speakers in the upper Asaro valley of PNG – which foreground increased crime and economic inequality – Thomas Strong (http://anthropology.nuim.ie/research/why-are-men-papua-new-guinea-highlands-shrinking-0) writes, "People point to men's bodies as further evidence of social decay – they claim that the present generation of men are smaller, weaker, and more fragile than the ancestors" and seek to reinvent both themselves and their culture by various means (including evangelical churches and commercial enterprises).

This theme of 'troubled masculinities' – in the comprehensive sense that Laura Zimmer-Tamakoshi develops in her chapter, "Troubled masculinities and gender violence in Melanesia" (in Jolly and Stewart, 2012) – is by now reasonably well developed in anthropological writing on gender and contributes to a contemporary cross-disciplinary interest in the dark side of economic and political change.

Complementary work dissects incipient class relations as Melanesians situate themselves relative to a roiling global economy (e.g., Gewertz and Errington, 1999) and analyzes struggles over the meaning and value of 'environmental resources' – struggles that mutually implicate transnational and local social agents and frames of reference (West, 2006). Big men have long brokered or actively invited these entanglements, particularly the articulation of gift- and market-oriented projects. While early writers interpreted customary big man activities through a capitalist lens (recall Section Political Typologies and Developmental Puzzles), ethnography of the 1990s and 2000s developed subtle understandings of the moral valences of 'gift' and 'market' as distinct, often opposed, kinds of relations even as their mutual entanglements have ramified.

These subtleties echo in ambivalent and contradictory commentary about the ethics of big men in everyday talk, popular media, and policy discourse. Because customary exchange relations are as much political as economic, references to 'big men' become convenient concretizations of otherwise abstract arguments about both development challenges and governance dilemmas (Sullivan, 2005; see also Section Anthropology of Big Manship in a Wider Field below).

For example, Highland PNG peoples have personified the state itself as a 'big man,' an idiom that collapses the scale distinction between 'local' and 'national' in order to insist on a relationship of equivalence and mutual responsibility between them (Clark, 1992; compare Joseph Ketan's 2004 book, *The Name Must Not Go Down: Political Competition and State-Society Relations in Mt Hagen*). Observations like Clark's make evident the salience of idealized references to 'big men' as political rhetoric. Such references can therefore be productive ethnographic foci for understanding the stakes for Melanesians as the wealth-mediated but community-centered competition (that is, competition premised on a wealth 'commons') with which big manship has customarily been associated confronts – however ambivalently, unevenly, and contingently – structural inequalities associated with neoliberal development (that is, premised on private property and individualist constructs of interests) in the context of ineffectual or absent state regulation.

Similarly, Laura Zimmer-Tamakoshi's (1997) instructive case study of the life and times of Ruge Angiva – a "self-proclaimed 'last Big Man'" of southern Madang Province – reads as a contemporary morality tale about the incompatibilities of traditional big manship and Western style capitalist interests, particularly their respective valuations of community and kin loyalties. Related tensions were evident in Tim Sharp's fieldwork on large-scale betel nut traders based in Mt Hagen. Despite the impressive geographical and financial scale of their enterprises, Sharp (2013) reports that relations among traders recall the interplay of cooperation and competition characteristic of clansmen and big men (see also Mosko, 2013). On the other hand, in a 2012 University of St Andrews dissertation, *The Pattern Changes Changes: Gambling Value in Papua New Guinea*, Anthony Pickles charts the rise of new forms of exchange relations and renown associated with slot machine gambling, which favor "big shots" (see also, e.g., Pickles, 2013). The latter term – which Keir Martin's provocatively titled 2010 article, 'The death of the big men,' analyzes as a linguistic index of the emergence of a new category of unambiguously problematic political elite – is defined in explicit counterpoint to 'big men' of the sort Zimmer-Tamakoshi describes.

Anthropology of Big Manship in a Wider Field

Even though the anthropology of big men is no longer as evident as it once was in Melanesian studies or as discrete

a literature as it used to be within political anthropology, it can be located in a number of other notable contexts.

It is alive and well in recent editions of introductory textbooks in sociocultural anthropology, like James Peoples and Garrick Bailey's 2011 *Humanity: An Introduction to Cultural Anthropology*, where it is used to illustrate leadership styles possible in 'egalitarian societies,' in contrast with leadership typical of 'ranked' societies, but with contemporary cautions concerning the disconnection between such typologies and "the lives of actual people" (p. 293). Big manship is also featured in comprehensive surveys of regional anthropology; for example, Paul Sillitoe's 1998 book, *An Introduction to the Anthropology of Melanesia: Culture and Tradition*, devotes a full chapter to big manship, focusing on customary practices of the Siuai people of Bougainville.

The big man/great man/chief typology also appears (as John Barker noted, *pers. com*. 10 July 2012), "in contexts in which we are discussing/updating the insights of our own anthropological 'big men'." Indeed, two recent critical appreciations of influential Melanesianist scholars – Barker's own 2007 edited collection, *The Anthropology of Morality in Melanesia and Beyond* (a substantive engagement with Kenelm Burridge's work) and *The Scope of Anthropology: Maurice Godelier's Work in Context*, a 2012 collection edited by Laurent Dousset and Serge Tcherkézoff – offer chapters that rethink big manship ethnographically and analytically. In keeping with contemporary theoretical trends, these works contribute to a more thoroughly historical perspective on the politics of shifting contexts for meaningful action.

Over the past several decades, anthropological archaeologists have recognized the general value of 'ethnographic analogies' in reconstructing the living social practices likely to have produced the material remains discovered in their excavations. They have begun to recognize the particular value of ethnographies of big manship in expanding their interpretive imaginations beyond conventional correlations between intensifying systems of production and exchange and hierarchizing political structures (see Section Political Typologies and Developmental Puzzles above) to more historically nuanced scenarios. Encouraging more sophisticated mutual engagement between archaeologists and ethnographers and more sophisticated theorizing and modeling of political action, Paul Roscoe (2000) reviewed the past decades' ethnography of New Guinea leadership. This comprehensive and admirably critical review articulates important cautions concerning comparativism. (By the by, its bibliography also testifies to the attenuation of big manship as a topical focus in the regional literature after the early 1990s, compared with previous decades.)

References to big manship also appear from time to time in contemporary ethnographic reports from outside of the island Pacific, especially in analyses of local political dynamics in the context of 'absent' or contested state influence. For example, in a 2012 Stanford University talk concerning post-civil war Lebanon, Marten Boekelo highlighted 'neighborhood big men' in an account of reworked forms of governance emergent after older mechanisms for the local mediation of the state and central political actors had broken down with the eruption of violence.

The term and its classic anthropological referents appear occasionally in popular media outside Melanesia, where it is used (with a positive valence) to signify decentralized or nonhierarchical influence, e.g., of Internet bloggers, and (with a negative valence) to refer to "crony capitalism" and other distortions of democratic process. For example, in an extended metaphor, an *Anchorage Daily News* article ('Juneau's Big Men should be scrutinized,' 18 May 2012) written by Kenai Peninsula College anthropologist Alan Boraas compared state legislators visiting their districts after state budget negotiations to Viking raiders returning home with the spoils of war, the point being that local leaders reinforce their local political clout by personally redistributing state wealth to their constituents. Referring to Sahlins's Melanesian exemplar, the article asserts that Alaskan legislators "function as 'Big Men'" insofar as their actions bypass the constitutional process mandating state agency control over local budget requests. Legislator-big men act as "gatekeepers to whom citizens come hat in hand." However, while "the state funding system is, at least in theory, transparent, non-government requests via the Big Man system are not" because that system relies on networks of "who knows whom" fueled by bribelike gifts ("a late-evening phone call, a fishing trip").

As if to echo such popular media representations, political and economic policy scholars' diagnoses of the causes of political 'dysfunction' in PNG and other Pacific island nations locate big manship prominently in this picture, although in inconsistent ways. On one side, in a 2007 report, Francis Fukuyama (http://www.sais-jhu.edu/faculty/fukuyama/Governance_PNG.doc) situated the big man as an anchor of indigenous democratic process and as an obstacle to the consolidation of state power. He argued that without a well-developed administrative bureaucracy, the national government has a limited ability to provide basic services like education and health care, especially in rural areas where the continued strength of kinship-based social structures and big manship – "highly unstable coalitions based on personality and patronage networks" – act as checks on the development of strong state control (p. 6). He linked the weaknesses of state agencies to the fact that "Papua New Guinea is arguably one of the most inherently democratic societies in the world." That is, conventions of "prolonged discussion and deliberation" and the 'big man' style of achieved leadership – "constantly subject to challenge and change" (p. 9) – all work against consolidation of strong state institutions. Fukuyama noted, however, that these same decentralizing sociopolitical dynamics also short circuit the development of powerful regional 'warlords' who might be in a position to challenge the state (as in Somalia or Liberia). Fukuyama concluded that governance reform in PNG may therefore depend on the prior cultivation of common national identifications that transcend the local loyalties with which big men are strongly identified.

Then again, other observers of PNG development trends have situated big men at the center of dysfunctional governance itself in diagnoses of the reasons for 'stalled' economic development. Writing in 2012 in the East–West Center's *Pacific Islands Report* (http://archives.pireport.org/archive/2012/April/04-18-an.htm), Helen Hughes associated that

country's "political stagnation" with elections that yield only "a changing roster of 'Big Men' to exploit parliamentary power. Senior public servants share in the spoils." Asserting that "Since the 1970s the Pacific has received the world's highest aid per head of population," Hughes concluded that foreign aid has not promoted rationalized economic development. Instead "aid is regarded as 'cargo'" (a term implying material wealth acquired by mystical means) to be redistributed strategically to curry favor with constituents. Similarly, in a 2005 Australian National University 'State, Society, and Governance in Melanesia' discussion paper (http://ips.cap.anu.edu.au/ssgm/papers/discussion_papers/05_04_dp_patience.pdf), former University of Papua New Guinea political science professor Allan Patience noted shifts in Australian aid policy that were motivated by "security fears" concerning "mutations of traditional ('wantok' or 'bigman') structures of power into virulent forms of crony politics," a political form "resistant to democratic institutions, transparency in government, and conventional state-making strategies" and therefore vulnerable to exploitation by international crime syndicates and terrorist networks.

These core themes of regional development policy analysis are echoed, debated, and countered in opinion pieces and letters to the editor of *The National*, *The Post Courier*, and other Melanesian media, where denunciations or praise of big manship figure prominently in commentary concerning the impacts of Western development agendas on historically Melanesian political and economic forms.

In contemporary discourse among urban elites and rural villagers, the figure of the big man is something of a key symbol. It powerfully condenses several tensions in twenty-first century Melanesian sociopolitical life at the heart of contemporary policy debates about governance. As Nancy Sullivan (2005) and the impressively varied collection of papers on 'the challenges of governance' in the Pacific that she introduces make clear, it stands for a neotraditional vision of properly reciprocal or redistributive sociality in counterpoint to both the alienating impacts of international consumer culture and the disappointing performances of national and provincial state governments. It also stands for the corrupting influence of excessively narrow loyalties and the oppressive dominance in domestic and public life of practices and values historically associated with maleness.

Holding contradictory impulses simultaneously in view is a characteristically anthropological move, whereas both popular and policy discourse tend to simplify or resolve them in some way. Loyalty to kin and expectations concerning material reciprocities can appear as nepotistic corruption or as a bulwark against capitalist individualism. Similarly, calls for the cultivation of cross-sectional national identity can appear as a ploy to undermine the very social resources most reliably available to ordinary people, under present conditions, in their disorienting encounters with multinational corporations and agents of neoliberal political and economic reform. However, contradictions among these points of view are themselves social facts with material impacts on the outcomes of projects that communities and experts care about.

Conclusion

In short, big manship is present both in public discourse at the national and local levels in Melanesia and among national and international development scholars and policy analysts with regional expertise. It is, therefore, a live issue in need of fresh ethnographic attention; this constitutes one rationale for a new generation of anthropological investigations into twenty-first century transformations in this leadership form, its associated social relations, and its shifting rhetorical values among diverse social actors.

Those investigations may be enriched in unexpected ways by continued or refreshed familiarity with an earlier generation's contributions to the anthropology of big manship. Similarly, contemporary anthropological archaeologists may be persuaded to draw on that literature as an exceptionally well-developed source of insights into the workings of decentralized political systems, on which to model their interpretations of the material evidence of past sociopolitical and exchange systems.

New fieldwork and ethnographic writing about big men are more likely to identify themselves with contemporary theoretical and topical rubrics than with the 'anthropology of big men' as such. However, so long as big manship, its variants, and echoes remain a focus of lay and expert attention in Melanesia, anthropological work will continue to enlighten by situating particular perspectives and experiences sympathetically in culture-historical context, and their commonalities and differences critically in juxtaposition.

See also: Anthropology and History; Comparative Method in Anthropology; Development: Social-Anthropological Aspects; Elites, Anthropology of; Exchange in Anthropology; Gender: Anthropological Aspects; Inequality: Comparative Aspects; Oceania, Sociocultural Overviews: Melanesia; Political Anthropology; Postcoloniality; State: Anthropological Aspects; Tribe.

Bibliography

Barker, J., 2012. The enigma of Christian conversion: exchange and the emergence of new great men among the Maisin of Papua New Guinea. In: Dousset, L., Tcherkézoff, S. (Eds.), The Scope of Anthropology: Maurice Godelier's Work in Context. Berghahn Books, New York, pp. 46–66.

Clark, J., 1992. Imagining the state: idioms of membership and control in the Southern Highlands Province of Papua New Guinea. In: Paper Presented at the Conference "Nation, Identity, and Gender". Australian National University, Canberra.

Epstein, T.S., 1968. Capitalism, Primitive and Modern: Some Aspects of Tolai Economic Growth. Michigan State University, East Lansing.

Finney, B., 1973. Big-Men and Business: Entrepreneurship and Economic Growth in the New Guinea Highlands. Australian National University Press, Canberra.

Gewertz, D., Errington, F., 1999. Emerging Class in Papua New Guinea: The Telling of Difference. Cambridge University Press, Cambridge.

Godelier, M., 1986. The Making of Great Men: Domination and Power among the New Guinea Baruya. Cambridge University Press, Cambridge.

Godelier, M., Strathern, M. (Eds.), 1991. Big Men and Great Men: Personifications of Power in Melanesia. Cambridge University Press, Cambridge.

Jolly, M., Stewart, C. (Eds.), 2012. Engendering Violence in Papua New Guinea. Australian National University Press, Canberra.

Lederman, R., 1986a. Changing times in Mendi: notes toward writing Papua New Guinea history. Ethnology 33, 1–30.
Lederman, R., 1986b. What Gifts Engender: Social Relations and Politics in Mendi, Highland Papua New Guinea. Cambridge University Press, Cambridge.
Lederman, R., 1990. Big men large and small? Towards a comparative perspective. Ethnology 29, 3–15.
Lederman, R., 1991. Interests in exchange: Mendi big-men in context. In: Godelier, M., Strathern, M. (Eds.), Big Men and Great Men: The Development of a Comparison in Melanesia. Cambridge University Press, Cambridge.
Lindstrom, L., 1981. Big man: a short terminological history. American Anthropologist 83 (4), 900–905.
Martin, K., 2010. The death of the big men: depreciation of elites in New Guinea. Ethnos 75 (1), 1–22.
Meggitt, M., 1967. The pattern of leadership among the Mae-Enga of New Guinea. Anthropological Forum 2 (1), 20–35.
Mosko, M., 2013. Dividuals, individuals, or possessive individuals? Recent transformations of North Mekeo commoditization, personhood, and sociality. In: Mccormack, F., Barclay, K. (Eds.), Engaging with Capitalism: Cases from Oceania, Research in Economic Anthropology, vol. 33. Emerald Group Publishing Limited, Bingley, pp. 167–198.
Oliver, D., 1955. A Solomon Island Society. Cambridge University Press, Cambridge.
Pickles, A., 2013. 'One-man one-man': how slot-machines facilitate Papua New Guineans' shifting relations to each other. In: Cassidy, R., Loussouarn, C., Pisac, A. (Eds.), Qualitative Research in Gambling: Exploring the Production and Consumption of Risk. Routledge, London.
Read, K., 1959. Leadership and consensus in a New Guinea society. American Anthropologist 61, 425–436.
Roscoe, P., 2000. New Guinea leadership as ethnographic analogy: a critical review. Journal of Archaeological Method and Theory 7 (2), 79–126.
Rumsey, A., Merlan, F., 1991. Ku Waru: Language and Segmentary Politics in the Western Nebilyer Valley. Cambridge University Press, Cambridge.
Sahlins, M., 1963. Poor man, rich man, big man, chief: political types in Melanesia and Polynesia. Comparative Studies in Society and History 5, 285–303.
Sexton, L., 1990. Pigs, pearl shells, and 'women's work': collective response to change in Highland Papua New Guinea. In: Lockwood, V.S., et al. (Eds.), Contemporary Pacific Societies. Prentice Hall, Englewood Cliffs, pp. 117–134.
Sharp, T., 2013. Baias, bisnis, and betel nut: the place of traders in the making of a Melanesian market. In: Mccormack, F., Barclay, K. (Eds.), Engaging with Capitalism: Cases from Oceania, Research in Economic Anthropology, vol. 33. Emerald Group Publishing Limited, Bingley, pp. 227–256.
Strathern, A., 1971. The Rope of Moka: Big Men and Ceremonial Exchange in Mount Hagen, New Guinea. Cambridge University Press, Cambridge.
Sullivan, N., 2005. Introduction. In: Sullivan, N. (Ed.), Governance Challenges for PNG and the Pacific Islands. Divine Word University in Association with the State, Society and Governance in Melanesia Project, Australian National University, Madang, pp. 1–28.
West, P., 2006. Conservation is Our Government Now: The Politics of Ecology in Papua New Guinea. Duke University Press, Durham and London.
Zimmer-Tamakoshi, L., 1997. The last big man: development and men's discontents in the Papua New Guinea highlands. Oceania 68, 107–122.

Bilingual Aphasia

Onésimo Juncos-Rabadán, University of Santiago de Compostela, Santiago de Compostela, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The manifestations of aphasia in bilingual or multilingual persons are an important source of data for the study of different recovery patterns, for the elaboration of hypothesis about the organization of the two or more languages in the brain, and in sum for a better understanding of the language representation and processing. It is critical that all languages of a person with aphasia are assessed by means of a cross-linguistically equivalent instrument and that the rehabilitation be done taking into account all of the involved biological and ecological variables.

Introduction

Bilinguals are persons who use two or more languages (or dialects) in their everyday life (Grosjean, 2010) at different levels of communicative competence and proficiency. Many factors need to be taken into account when characterizing bilinguals such as the languages actually known and used, the structural relationships between them, the history of acquisition and use, and the functions of their languages including context and purposes of use. All these factors determine the verbal communicative competence of bilinguals for each of their languages. Verbal communicative competence comprises: (1) implicit linguistic competence that refers to the knowledge inferred from individuals' systematic verbal performance, it is acquired incidentally by practice, it is stored implicitly, and it is used automatically; (2) explicit metalinguistic knowledge that refers to knowledge of which individuals are aware and that they are capable of representing to themselves and verbalizing, it is learned consciously, it is stored explicitly, and recalled voluntarily; (3) pragmatic competence is the ability to select the appropriate linguistic and paralinguistic choices, given the context, the corresponding purposes, and the required effect on the audience; and (4) motivational drive that refers to the complex network of motivations and affects involved in each language acquisition, learning, and use (Paradis, 2004).

Verbal communicative competence of bilingual individuals can be affected as result of different brain damage that produces linguistic and communicative deficits, broadly known as aphasia, with consequences on the bilingual person's life. As bilingualism is a worldwide phenomenon and half of the world's population is bilingual (Grosjean, 2010), bilingual individuals suffering from aphasia are not rare cases but rather a majority of clinical cases (Paradis, 1998a). The study of manifestations, assessment, and rehabilitation of aphasia in bilinguals constitutes one important issue of the neurosciences and applied linguistics, and allows researchers and clinicians to better understand how languages are represented in the brain, how verbal communicative competence may be affected at several levels, and to different extents in the way that languages can be rehabilitated.

Interest in bilingual aphasia appeared in the nineteenth century concurrently with the classical studies of language impairment in brain-damaged patients and with the controversies about the representation of language in various brain centers. Some reports of bilingual aphasic patients appeared after the publication of the Paul Broca's (1861) seminal paper on aphasia but the first classical study on aphasia in bilinguals was written by Albert Pitres in 1895 (Pitres, 1895). In his study, Pitres reviewed several case histories described by contemporary scientists (Bastian, Oré, Rinkenbach, Bianchi, Charcot, and Bernard) and analyzed seven cases that he gathered himself. From these cases, especially from his case number seven, Pitres designed the first theoretical proposal on bilingual aphasia based on the following principles: (1) differential loss, "when a polyglot patient is stricken with aphasia, he does not necessarily lose to an equal extent the use of all the languages he knew before"; (2) differential recovery based on familiarity, "the patient first begins to understand, then to speak the language that was most familiar to him. Ultimately he is able to understand, then to speak the other languages that he knew"; (3) inhibition but not destruction, "the systematic return of the languages occurs in cases where language centres have been shocked or inhibited but not destroyed by lesions…, gradually take up their functional activities again after having lost them for a period of time"; and (4) participation of the same cortical language centers as in monolinguals, "the temporary inertia of the cortical language centres accounts well enough for the sequence of phenomena observed in polyglot aphasics that it is not necessary to bring in the absolutely hypothetical existence of new centres that are especially assigned to each of the languages learned successively by these subjects" (Pitres, 1895, translation in Paradis, 1983, p.48).

From the Pitres' classical paper the most important issues that have been addressed by researchers and clinician are the following: (1) description and explanation of different patterns of recovery; (2) representation of the languages in the brain; (3) assessment of bilingual aphasia; and (4) rehabilitation of bilingual aphasia.

Recovery Patterns

Study of the manifestations of aphasia in bilinguals has served for the elaboration of hypotheses about the cognitive representation and processing of languages and their cerebral organization. Descriptions of verbal behavior of bilingual aphasic

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54041-0

patients reported since the end of the nineteenth century and more extensively over the twentieth century allow for the identification of several patterns of recovery. In extensive surveys of the world literature on aphasia in bilinguals, Michel Paradis (1977, 1983, 1989, 2001) identified the following basic patterns of spontaneous recovery after an injury that caused language impairments: parallel, differential, successive, antagonist, selective, blended recovery, and selective aphasia. Parallel recovery is when both or all languages are recovered simultaneously. Differential recovery refers to cases where one language is recovered much better than the others. Successive recovery is said when the recovery of one language does not begin until another has been maximally recovered. Antagonist recovery occurs when the languages are not concurrently available; only one language is available to the patient, and the language that is recovered first ceases to be available and is replaced by the language subsequently recovered. This pattern may occur recursively and is called alternating antagonism when patients have access to only one language for alternating periods of time. The recovery is selective when patients only recover one of the languages but not the others. Blending recovery occurs when patients inappropriately mix or blend phonological, morphological, syntactic, lexical, and/or semantic features of their languages. Selective aphasia refers to the presence of aphasic symptoms in one language without measurable deficits in the other(s). The incidence of the various recovery patterns based on 138 single cases published from 1843 to 2000 (Paradis, 2004) indicate that parallel recovery is the most common pattern (76%), followed by differential (12.3%), blended (6.5%), selective (4.3%), and antagonistic (0.73%), taking into account that cases of parallel recovery are not often reported.

All these patterns are merely descriptions of patients' verbal behavior in a given situation at a particular time. They are not mutually exclusive and they can be seen as different manifestations on a continuum of time and/or severity. Theoretical explanations have been proposed over the years based on single or multiple factors to account for recovery patterns. Ribot (1881) predicted that the first acquired language, the native language, should be better recovered according to his regression rule which states that the most recent memory traces are the first destroyed and the oldest acquisitions are the most resistant. Pitres (1895) proposed that patients preferentially recover the language that was the most familiar to them at the time of the insult. Pick (1921) suggested that the language most automatized seems to be the first that patients recovered. Minkowsky (1928, 1965) speculated that the recovery responds to the instinctive needs of the patients such as the emotional and affective forces particularly linked to their developmental history. Other authors observed that recovery may be facilitated if the language had been learned in written form (Halpern, 1941; Paradis and Goldblum, 1989). Goldstein (1948) stated that the language first recovered was the most appropriate for the patient's purposes, for example, for helping him/her to communicate in the hospital environment. Lambert and Fillenbaum (1957) and Mastronardi et al. (1991) hypothesized more selective disturbances for coordinate bilinguals, who have acquired each language in a different context than for compound bilingual, who acquired both language in the same context. This hypothesis includes neurological separated structures underlying language for coordinate bilinguals. Albert and Obler (1978) suggested that bilinguals might have different cerebral lateralization for the two languages and from that suggestion it was assumed that nonparallel recovery might result from one language being less lateralized than the other. However, none of the above factors accounts for all the reported cases. Some patients have first recovered their first or second language; some patients first recovered the most familiar language and in some cases, the least familiar; some patients recovered the language of their surroundings and many did not recovered the language of their environment. Recovery seems to be produced by different factors at different times and a theoretical explanation which could account for all the observed recovery patterns is needed. To date the most integrated theory that can account for all the recovery patterns is that proposed by Michel Paradis (2004) reported below.

Paradis' Neurofunctional Model

This model incorporates several hypotheses proposed by Paradis from 1977 to 2004: the Three-Store Hypothesis, the Subsystems Hypothesis, the Activation Threshold Hypothesis, the Direct Access Hypothesis, and the contribution of implicit competence, metalinguistic knowledge, pragmatic capacity, and motivation.

The Three-Store Hypothesis proposes to separate the nonlinguistic cognitive store from the two (or more) language stores of bilinguals. It postulates that bilinguals possess two language subsystems constituted by grammar and lexicon for each language, and one nonlinguistic system that contains conceptual representations which interact with the two linguistic subsystems. The bilingual persons might organize the mental representations of their nonlinguistic system slightly differently in each language. The conceptual system is not language-specific but the meaning of the words and utterances are language-specific because each language specifies what mental representation is included in the meaning of each word or sentence combining features of mental representations in different ways.

The Direct Access Hypothesis proposes that a word or a sentence in either language is perceived directly as a word or a sentence integrating sound and meaning, irrespective of which language it belongs to and without the listener's having to first identify the language affiliation of that word or sentence. The access to the meaning is automatic and the awareness of the language identity of a word or sentence is a metalinguistic phenomenon that occurs after the meaning has been accessed.

The Subsystems Hypothesis proposed an independent neurofunctional language system with two or more subsystems or subsets corresponding to each of the bilingual's languages. The two or more specific subsystems (say, English and Spanish) share the properties of the linguist system whose principles they instantiate along different parameters (rules or forms). Structural distance between two languages can be measured in terms of the number of shared parameters. The two subsystems are subserved by one subset of neural connections for each language while at the same time, because they partake in the general neurofunctional language system, elements of either language can be drawn at any time.

The Activation Threshold Hypothesis proposes that a language or an item of this language (word or morphosyntactic feature) is activated when a sufficient amount of positive neural impulses have reached its neural substrate. The amount of impulses necessary to activate the language or item constitutes its activation threshold. The activation of a language or item requires that this activation exceeds that of the other alternatives that must be simultaneously inhibited. Activation depends on the frequency and recency of use of a particular language or item.

Implicit competence and metalinguistic knowledge constitute two independent systems and two neurofunctional subsystems of communication. These two systems participate in the appropriation of the two languages in bilinguals. If the acquisition is accomplished incidentally by practice, the implicit competence is mainly involved, whereas metalinguistic knowledge is more activated if the appropriation is made consciously and formally.

Motivation and affect are important factors in the acquisition, learning, and use of the first and second languages in bilinguals and contributes to lowering the activation threshold of languages in the same way that frequency and recency of use do.

The Paradis' integrated model allows to explain all the various described patterns of recovery in bilingual aphasia. Parallel impairments can be interpreted as the result of damage in the linguistic neurofunctional system as a whole. Differential, successive, antagonistic, and selective aphasia can be interpreted as the result of damage to one of the language subsystems. Each language, as a subsystem, is susceptible to selective pathological inhibition depending on its activation threshold that can be related to the frequency and recency of use, the acquisition and learning mode (implicit or explicit), and the emotional factors involved in its acquisition and use.

Representation of the Languages in the Brain

The study of the representation of multiple languages in the human brain has been one of the most important issues in the history of research on bilingual aphasia. Several hypotheses have been proposed (for a review see Fabbro, 1999; Paradis, 2001). (1) All languages are represented in the same areas, the classical language areas. This proposal includes several formulations that go from the broad hypothesis suggested by Pitres, participation of the same cortical language centers as in monolinguals, to the more clinically and neurolinguistically based explanation proposed by Paradis (2004), who claimed that the two languages are represented as distinct microanatomical and neurofunctional subsystems located in the same gross anatomical areas. (2) Each language is represented in different areas in the brain. The strongest formulation of this hypothesis is that known as differential lateralization that proposes that the two languages of bilinguals are differently lateralized and one of them is more bilaterally represented. This hypothesis was first suggested by V. Gorlitzer von Mundy (1959/1983), who stated that language centers are originally in both hemispheres, remain in both hemispheres in illiterates but become lateralized in literates due to the instruction and acquisition of unimanual writing skills. Albert and Obler (1978) from an analysis of the literature on case studies and experimental neuropsychology on bilingual aphasia made their famous speculative suggestion that "the right hemisphere …may be dominant for one of the languages of the bilinguals" (p.254) that led to many experimental studies with contradictory results. Many researchers have carried out behavioral studies on cerebral lateralization on language in brain-intact bilinguals, neurological and clinical research on crossed aphasia, but their findings are controversial and constrained by methodological weaknesses and have not advanced our knowledge of the lateralization of language in bilingual speakers (see Hull and Vaid, 2007, 2008; Paradis, 2008). All the evidence gathered to date points to both languages being represented in the language areas of the left hemisphere in the same proportion as in monolinguals. However, many studies on communication disorders in adults with right hemisphere damage confirm that the right hemisphere is crucially involved in the pragmatic aspects of language rather than form or content (Cummins, 2009; Paradis, 1998b). Paradis (1998c) proposed the term dyshyponoia to refer to impairments of the use of linguistic pragmatics produced by right hemisphere damage; e.g., the inability to draw appropriate inferences from the context or from general knowledge, leading to problems in the interpretation of indirect speech acts, metaphors, and in general in the unspoken components of an utterance. In this sense, a greater involvement of the right hemisphere in bilinguals may be necessary to apply pragmatic features when using the appropriate language in certain contextual circumstances. Pragmatics is also necessary to compensate lacunae in implicit linguistic competence.

Neuroimaging Studies of the Bilingual Brain

The advent of noninvasive neuroimaging techniques such as positron emission tomography (PET), functional magnetic resonance imaging, event-related brain potentials, and magnetoencephalography, has permitted the investigation of language organization in bilinguals. From the first PET study by Klein et al. (1994), who used a word repetition task in a group of Canadian bilinguals who were native English speakers and French speakers after age 5 to elucidate whether production in a second language involved the same neural substrates as in the first language, many different studies have been carried out (see for reviews Abutalebi et al., 2001; Vaid, 2008; Vaid and Hull, 2002). To date we have a large amount of puzzling and sometimes contradictory findings from different neuroimaging techniques, different linguistic tasks, and different types of healthy bilingual individuals (see for a review Sebastian et al., 2011). From these results one may derive some patterns: (1) the activation of the left perisylvian area increases with degree of proficiency or implicit linguistic competence in the second language suggesting that brain activity converges toward that of the first language (Chee, 2009); (2) activation of the left prefrontal cortex, anterior cingulate gyri, and basal ganglia increases as fluency in the language decreases, indicating an increased use of controlled processes as additional support to processing the weaker language (Abutalebi and Green, 2007, 2008); (3) activation of

the right temporal lobe increased also in late bilinguals (Dehaene et al., 1997; Perani and Abutalebi, 2005) suggesting the involvement of pragmatic processes to compensate their weaker linguistic competence; (4) activation of parahippocampal gyri associated with declarative memory was observed during the use of the second language. It is important to note that studies using single-word stimuli find no differences between Language 1 and Language 2 cerebral representations, whereas studies that use sentences do (see Paradis, 2004).

The neuroimaging studies of bilingual aphasia patients are scarce and focused on examining the effects of specific language treatment on the representation of language areas during the recovery. Although they are limited to particular language tasks, picture-naming tasks and word semantic judgment task, and to case studies, they provide interesting findings. (1) Improved performance in the treated language was associated with increased activation in language area and in regions associated with control (anterior cingulate cortex, the left caudate, and prefrontal cortex), which form together a whole network with stronger connections for the treated language and with weaker connections for the untreated language. This network is completed by the left basal ganglia and left prefrontal circuitry that may subserve language planning and inhibition (Abutalebi et al., 2009). These results provide neurobiological evidence for the activation threshold theory and for the contribution of metalinguistic competence postulated by Paradis (2004). (2) The improvement of the treated language is accompanied by bilateral activation of the left and the right superior temporal, in frontal (inferior and medial frontal areas) and inferior parietal areas, while the activation pattern of the untreated language remains unchanged (Meinzer et al., 2007). (3) The activation for the weaker language is greater than for the more fluent one and includes both left hemisphere and right hemisphere regions, superior temporal gyrus, postcentral gyrus, supramarginal gyrus, and subinsular lobe, indicating that bilingual patients need to recruit a more distributed network of regions during the processing of the less proficient and used language. There is also a great variability in activation pattern which indicates that less proficient bilinguals engage greater cognitive resources when processing the less used/less proficient language (Sebastian et al., 2012). (4) When the level of proficiency and use is high in both languages, the same cerebral regions are recruited for both languages before and after training. Increasing activations are observed perilesionally and in homologous contralesional areas (Marangolo et al., 2009).

The Assessment of Bilingual Aphasia

Assessment of language proficiency, history and contexts of acquisition, and use of the two or all of multilingual's languages is essential for purposes of diagnosis, research, and treatment. Literature on different recovery pattern has shown that many factors are responsible for parallel or differential recovery such as the history and contexts of acquisition, the history and contexts of premorbid uses, and the structural relationships of the languages. Postlesional proficiency of bilingual aphasics must be assessed with an equivalent instrument in each language which could test phonology, morphology, grammar, lexicon, and semantics so that the results obtained may be validly compared. Different standardized test batteries are available for assessing bilingual aphasia. Most of them are originally constructed in English and then translated into other languages, but it is very important that all languages of a person with aphasia are assessed by means of a cross-linguistically equivalent instrument and not by a mere translation of a standardized test from another language (Paradis, 2004). The criteria of cross-language equivalence must include the cultural appropriateness of items, the equivalence in complexity of syntactic structures rather than the same translated constructions, the selection of phonological minimal pairs which are depending of the structure of each language and the equivalent frequency, and use of the selected words. In order to provide a systematic and cross-linguistically comparable instrument to assess bilingual aphasia, Michel Paradis developed the Bilingual Aphasia Test (BAT) which was designed to be equivalent across the 65 languages in which it is currently available (Paradis, 2011; Paradis and Libben, 1987).

The BAT examines language performance in all four modalities (hearing, speaking, reading, and writing). It is linguistically multidimensional because it explores, for each modality, language performance along three dimensions: (1) linguistic level – phonological, morphological, syntactic, lexical, semantic; (2) linguistic task – comprehension, repetition, judgment, lexical access, propositionizing; and (3) linguistic unit – word, sentence, paragraph. The BAT is composed of three parts. Part A assesses the patient's history of bilingualism and it is devoted to obtaining information about the context of acquisition and use of all the languages acquired by the patients, exploring the patient's linguistic environment at home, education, and work from their childhood to the present. Part B provides information relating to the language background, spontaneous speech, and linguistic competence (comprehension and production) for each of the languages tested. Part C examines bilingual skills in a specific pair of languages and comprises four tasks in each direction: recognition of translation-equivalent words, production of translation-equivalent words, translation of sentences, and grammatical judgments. Within each task, the patients' performance in one language may be compared with their performance in the other.

The Rehabilitation of Bilingual Aphasia

As pointed out by Paradis (2000), the study of aphasia rehabilitation in bilingual subjects is one of the major challenges facing the neurolinguistics of bilingualism in the twenty-first century. There are numerous questions that need to be addressed. In what language should the therapy be performed, in all languages that the patient spoke previously, or in only one language? If it is necessary to perform the therapy in all languages, how should this be done: simultaneously or successively, and why? If one of the languages must be selected for therapy, what criteria should be applied, the native language, the one that was the most fluent premorbidly, the best recovered, the language of the environment? Can the benefits obtained from therapy in one language transfer to the untreated language? If so, on what factors does this transfer depend, the structural distance between

the languages, the age or acquisition, the dominance pre- or postonset, the type of aphasia, the pattern of recovery, or the type of therapy? (Paradis, 1993, 2004).

The number of studies on rehabilitation in bilingual aphasia is limited even though it is increasing since the beginning of the twenty-first century, especially those focused on the effect of treatment in one language on the untreated language(s) (see as examples in Galvez and Hinckley, 2003; Goral et al., 2012; Hinckley, 2003; Kiran and Roberts, 2010; Kohnert, 2004; Miller Amberber, 2012). The differences between them, including variation in the assessment measures, treatment approaches, type and severity of aphasia, proficiency in each of the languages, and acquisition history of participants, make interpretation of results problematic (see for reviews, Faroqi-Shah et al., 2010; Juncos-Rabadán et al., 2002; Roberts, 2008). Notwithstanding the enormous variability among the reported cases, and their methodological limitations, several hypotheses have been proposed from the findings: (1) Chlenov's (1948) hypothesis that therapy should initially be carried out only in one language, the language in which spontaneous recovery typically occurs and which may or may not be the mother tongue, thus avoiding the probable inhibitory effect of therapy in two languages. Only when significant recovery has been achieved in the basic language should treatment be attempted in the other. (2) Transfer from the treated language to untreated language might be affected by several factors such as structural distance between the two languages, age of acquisition, and the level of proficiency. (3) Structural distance is the main factor for transfer when therapy is based on morphology and syntax; however, age of acquisition and the level of proficiency are more important when the therapy is focused on the lexicon. (4) Pragmatic and social factors such as the history of use of a particular language and the language of the environment during the treatment have been proposed to account for the presence and absence of cross-language generalization. (5) No clear trends emerged regarding the potential interaction between these factors and the studies highlight the inherent problem of variability in research on bilingual aphasia therapy.

In the light of the Paradis' neurofunctional model, several processes and strategies have been proposed to apply for treatment of bilingual aphasics (Juncos-Rabadán et al., 2002). Activation, competition, priming, and inhibition can be applied to the therapy of linguistic fragments, levels, or skills of the treated language. For instance, the activation of the lexical unit corresponding to a target word in one of the languages will be facilitated supplying contextual and phonological primes specific for this language and restricting the number of competing alternatives from the other language. If the lesions affect basically implicit memory the focus of the therapy should be on explicit learning using metalinguistic knowledge and strategies such as semantic–conceptual and semantic–syntactic therapy; if the lesions affect explicit memory, the focus will be on implicit learning such as identification and associative priming.

A recent approach to addressing rehabilitation in bilingual aphasia is to use computational modeling to complement our understanding of how language recovery occurs in a bilingual individual with aphasia (Kiran et al., 2012). However, this approach is constrained by the necessity of simplifying models to particular tasks such as naming, by a particular architecture of the computational model, and by the limitation of factors that are considered, for example, age of acquisition and level of proficiency.

Conclusion

Bilingual or multilingual individuals with aphasia recover their languages at different times and to different extents. This recovery pattern depends on several biological and ecological factors. For a better understanding of these factors, more case and group studies – carefully assessed with an equivalent instrument across different languages and administered before and after controlled therapy programs – are needed. Experimental and neuroimaging studies have as challenges for the future to find linguistic tasks that form part of the natural use of language and to use reliable and replicable techniques with comparable subject populations in order to get data that allow valid inferences within a theoretical framework on bilingual aphasia. From these theoretical explanations new strategies and methodologies must be designed to rehabilitate bilingual/multilingual persons with aphasia in order to improve their communicative competence and quality of life.

See also: Aphasia; Brain Damage: Neuropsychological Rehabilitation; Cerebral Cortex; Code Switching: Linguistic; Communicative Competence: Linguistic Aspects; Declarative Memory, Neural Basis of; Functional Brain Imaging of Language processes; Hippocampus and Related Structures; Implicit Memory; Interhemispheric Interaction in Language and Cognitive Processes; Language Acquisition; Language Disorders, Treatment and Remediation of; Language Processing, Functional Magnetic Resonance Imaging of; Lateralization of Language as Demonstrated by Brain Imaging Procedures; Learning and Memory, Neural Basis of; Multilingualism; Neurolinguistics, History of; Neurological Approaches to Agrammatism; Non Declarative (Procedural) Memory; Phonology; Pragmatic Disorders; Pragmatics, Linguistic; Sign Language; Sociolinguistics; Speech Production, Psychology of.

Bibliography

Abutalebi, J., Green, D., 2007. Bilingual language production: the neurocognition of language representation and control. Journal of Neurolinguistics 20, 242–275.

Abutalebi, J., Green, D., 2008. Control mechanisms in bilingual language production: neural evidence from language switching studies. Language and Cognitive Processes 23, 557–582.

Abutalebi, J., Capp, S.F., Perani, D., 2001. The bilingual brain as revealed by functional neuroimaging. Bilingualism: Language and Cognition 4, 179–190.

Abutalebi, J., Rosa, P.A.D., Tettamanti, M., Green, D.W., Cappa, S.F., 2009. Bilingual aphasia and language control: a follow up fMRI and intrinsic connectivity study. Brain and Language 109, 141–156.

Albert, M.L., Obler, L.K., 1978. The Bilingual Brain. Neuropsychological and Neurolinguistc Aspects of Bilingualism. Academic Press, New York.

Broca, P., 1861. Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie. Bulletin de la Société Anatomique XXXVI, 330–357.

Chee, M.W.L., 2009. fMR-adaptation and the bilingual brain. Brain & Language 109, 75–79.

Chlenov, 1948/1983. On aphasia in polyglots. In: Paradis, M. (Ed.), Readings on Aphasia in Bilinguals and Polyglots. Didier, Montreal, pp. 445–454.

Cummins, L., 2009. Clinical Pragmatics. Cambridge University Press, Cambridge.
Dehaene, S.D., Dupoux, E., Mehler, J., Cohen, L., Paulesu, E., Perani, D., van de Moortele, P.F., Lehéricy, S., Le Bihan, D., 1997. Anatomical variability in the cortical representation of first and second languages. NeuroReport 8, 3809–3815.
Fabbro, F., 1999. The Neurolinguistics of Bilingualism. An Introduction. Psychology Press, Hove.
Faroqi-Shah, Y., Frymark, T., Mullen, R., Wang, B., 2010. Effect of treatment for bilingual individuals with aphasia: a systematic review of the evidence. Journal of Neurolinguistics 23, 319–341.
Galvez, A., Hinckley, J.J., 2003. Transfer patterns of naming treatment in a case of bilingual aphasia. Brain and Language 87, 173–174.
Goldstein, K., 1948. Language and Language Disturbances. Grune and Stratton, New York.
Goral, M., Rosas, J., Conner, P.S., Maul, K.K., Obler, L.K., 2012. Effects of language proficiency and language of the environment on aphasia therapy in a multilingual. Journal of Neurolinguistics 25, 538–551.
Gorlitzer von Mundy, V., 1959/1983. A 94-year-old with one German language center and probably two Slovenian centers. In: Paradis, M. (Ed.), Readings on Aphasia in Bilinguals and Polyglots. Didier, Montreal, pp. 624–625.
Grosjean, F., 2010. Bilingual. Life and Reality. Harvard University Press, Cambridge.
Halpern, L., 1941/1983. Restitution in polyglot aphasia with regard to Hebrew. In: Paradis, M. (Ed.), Readings on Aphasia in Bilinguals and Polyglots. Didier, Montreal, pp. 418–422.
Hinckley, J.J., 2003. Picture naming treatment in aphasia yields greater improvement in L1. Brain and Language 87, 171–172.
Hull, R., Vaid, J., 2007. Bilingual language lateralization: a meta-analytic tale of two hemispheres. Neuropsychologia 45, 1987–2008.
Hull, R., Vaid, J., 2008. Bilingual laterality and the matter of degree: a response to Paradis (2008). Neuropsychologia 46, 1591–1593.
Juncos-Rabadán, O., Pereiro, A.X., Rodríguez, M.J., 2002. Treatment of aphasia in bilingual subjects. In: Fabro, F. (Ed.), Advances in the Neurolinguistics of Bilingualism. Essays in Honor of Michel Paradis. Forum, Udine, Italy, pp. 275–298.
Kiran, S., Grasemann, U., Sandberg, Ch, Miikkulainen, R., 2012. A computational account of bilingual aphasia rehabilitation. Bilingualism: Language and Cognition 16, 325–342. http://dx.doi.org/10.1017/S1366728912000533.
Kiran, S., Roberts, P.M., 2010. Semantic feature analysis treatment in Spanish-English and French-English bilingual aphasia. Aphasiology 24, 231–261.
Klein, D., Zatorre, R., Milner, B., Meyer, E., Evans, A., 1994. Left putaminal activation when speaking a second language: evidence from PET. NeuroReport 5, 2295–2297.
Kohnert, K., 2004. Cognitive and cognate-based treatments for bilingual aphasia: a case study. Brain and Language 91, 294–302.
Lambert, W.E., Fillenbaum, S., 1957. A pilot study of aphasia among bilinguals. Canadian Journal of Psychology 13, 28–34.
Marangolo, P., Rizzi, Ch, Peran, P., Piras, F., Sabatini, U., 2009. Parallel recovery in a bilingual aphasic: a neurolinguistic and fMRI study. Neuropsychology 23, 405–409.
Mastronardi, L., Ferrante, L., Celli, P., Acqui, M., Fortuna, A., 1991. Aphasia in polyglots: report of two cases and analysis of the literature. Neurosurgery 29, 621–623.
Meinzer, M., Obleser, J., Flaisch, T., Eulitz, C., Rockstroh, B., 2007. Recovery from aphasia as a function of language therapy in an early bilingual patient demonstrated by fMRI. Neuropsychologia 45, 1247–1256.
Miller Amberber, A., 2012. Language intervention in French–English bilingual aphasia: evidence of limited therapy transfer. Journal of Neurolinguistics 25, 588–614.
Minkowsky, M., 1928. Sur un cas d'aphasie chez un polyglotte. Revue Neurologique 49, 361–366 [English translation in Paradis, M. (Ed.), Readings on Aphasia in Bilinguals and Polyglots. Didier, Montreal, 1983, pp. 274–279].
Minkowsky, M., 1965. Considérations sur l'aphasie des polyglottes. Revue Neurologique 112, 486–495 [English translation in Paradis, M. (Ed.), Readings on Aphasia in Bilinguals and Polyglots. Didier, Montreal, 1983, pp. 717–726].
Paradis, M., 1977. Bilingualism and aphasia. In: Whitaker, H., Whitaker, H.A. (Eds.), Studies in Neurolinguistics, vol. 3. Academic Press, New York, pp. 65–121.
Paradis, M., 1983. Readings on Aphasia in Bilinguals and Polyglots. Didier, Montreal.
Paradis, M., 1989. Bilingual and polyglot aphasia. In: Boller, F., Grafman, J. (Eds.), Handbook of Neuropsychology, vol. 2. Elsevier, Amsterdam, pp. 117–140.
Paradis, M., 1993. Language therapy in bilingual aphasia. In: Paradis, M. (Ed.), Foundations of Aphasia Rehabilitation. Pergamon Press, Oxford, pp. 413–420.
Paradis, M., 1998a. Language and communication in multilinguals. In: Stemmer, B., Whitaker, H.A. (Eds.), Handbook of Neurolinguistics. Academic Press, San Diego, pp. 418–431.
Paradis, M., 1998b. Pragmatics in Neurogenic Communication Disorders. Pergamon, Oxford.
Paradis, M., 1998c. The other side of language: pragmatic competence. Journal of Neurolinguistics 11, 1–10.
Paradis, M., 2000. The neurolinguistics of bilingualism in the next decades. Brain and Language 771, 178–180.
Paradis, M., 2001. Bilingual and polyglot aphasia. In: Berndt, R.S. (Ed.), Handbook of Neuropsychology, second ed., vol. 3. Elsevier, Amsterdam, pp. 69–91.
Paradis, M., 2004. A neurolinguistic theory of bilingualism. J.Benjamins Pub, Amsterdam.
Paradis, M., 2008. Bilingual laterality: unfounded claim of validity. A comment on Hull and Vaid. Neuropsychologia 46, 1588–1590.
Paradis, M., 2011. Principles underlying the Bilingual Aphasia Test (BAT) and its uses. Clinical Linguistics & Phonetics 25, 427–443.
Paradis, M., Goldblum, M.C., 1989. Selective crossed aphasia followed by reciprocal antagonism in a trilingual patient. Brain and Language 36, 62–65.
Paradis, M., Libben, G., 1987. The Assessment of Bilingual Aphasia. Lawrence Erlbaum Associates, Hillsdale, NJ.
Perani, D., Abutalebi, J., 2005. Neural basis of first and second language processing. Current Opinion of Neurobiology 15, 202–206.
Pick, A., 1921/1983. Concerning the explanation of certain exceptions to the so-called rule of Ribot. In: Paradis, M. (Ed.), Readings on Aphasia in Bilinguals and Polyglots. Didier, Montreal, pp. 156–168.
Pitres, A., 1895/1983. Etude sur l'aphasie des polyglottes. Revue de Médecine 15, 873–899 [English translation in Paradis, M. (Ed.), Readings on aphasia in bilinguals and polyglots. Didier, Montreal, 1983, pp. 26–49].
Ribot, Th., 1881. Les maladies de la mémoire, 2e édition. Baullère, Paris.
Roberts, P.M., 2008. Aphasia assessment and treatment for bilingual and culturally diverse patients. In: Chapey, R. (Ed.), Language Intervention Strategies for Aphasia and Related Neurogenic Communication Disorders, fifth ed. Lippincott, Williams & Wilkins, New York, pp. 245–276.
Sebastian, R., Kiran, S., Sandberg, C., 2012. Semantic processing in Spanish–English bilinguals with aphasia. Journal of Neurolinguistics 25, 240–262.
Sebastian, R., Laird, A., Kiran, S., 2011. Meta-analysis of the neural representation of first language and second language. Applied Psycholinguistics 32, 799–819.
Vaid, J., 2008. The bilingual brain: what is right and what is left? In: Altarriba, J., Heredia, R.R. (Eds.), An Introduction to Bilingualism: Principles and Processes. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, USA, pp. 129–144.
Vaid, J., Hull, R., 2002. Re-envisioning the bilingual brain using functional neuroimaging: methodological and interpretative issues. In: Fabbro, F. (Ed.), Advances in the Neurolinguistics of Bilingualism. Essays in Honor of Michel Paradis. Forum, Udine, Italy, pp. 315–356.

Bilingual Education: International Perspectives

Rosemarie Tracy, University of Mannheim, Mannheim, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article deals with the objectives and types of bilingual education, from traditional models such as *immersion*, *submersion*, and *dual language curricula* to current initiatives. Among the goals, assimilation of linguistic minorities, heritage-language maintenance, and linguistic empowerment are prominent. Reasons for the success or failure of various approaches are discussed, and it is shown how controversies and attitudes reflect dominant ideologies. It is argued that evidence from research on bilingualism/multilingualism in general and, in particular, on bilingualism in early childhood underscores the superb human ability to deal with more than one language at a time and, therefore, encourages efforts of implementing bilingual language programs throughout the educational system.

The majority of the world's population knows, speaks, and sometimes even writes in more than one language. Given the human cognitive capacity of managing multiple linguistic systems fairly effortlessly, educational curricula employing two or more languages of instruction, with schools profiting in turn from the linguistic and cultural resources of bilingual and multilingual individuals, should be quite common and find wide acceptance. However, to this day, bilingual or multilingual education is a controversial topic, with positions ranging from active promotion to more or less subtle resistance – the latter often camouflaged by liberal and pluralist rhetoric. While researchers no longer question that children can be raised with and schooled in more than one language without negative consequences for their linguistic, emotional, and cognitive development, prejudice, xenophobia, or simply ignorance raise the stakes for bilingual or multilingual programs. Their fate is intricately linked to attitudes toward bilingualism in general and also toward the languages involved. The European Commission, for instance, encourages European citizens to acquire two languages besides their first language(s), and it celebrates the cultural and linguistic diversity of its community as an added value, sure to guarantee a return on investment within the global market. At the same time, the multitude of immigrant languages and cultures originating from other parts of the globe does not appear to count, and the linguistic competence of children growing up with non-European languages and cultures is dismissed or treated as an obstacle and handicap. As pointed out by Stubbs (1991 (1995: 30)), "there is likely to be an enormous waste: allowing languages already spoken in the home and community to be eroded, whilst starting from scratch to teach other languages in schools and colleges." Even though my contribution can only marginally deal with contradictions and double standards of this kind, they nevertheless form a part of the backdrop for what follows.

After an explication of relevant terminology, an overview of the objectives and major types of bilingual education, such as the 'classics' Immersion, Submersion, Transition, Dual Language/Two Way models, and a recent offspring of immersion, Content and Language Integrated Learning (CLIL), are provided, and some parameters determining their effectiveness are identified. A brief historical sketch, mainly based on developments within the United States, then illustrates how the political climate swings attitudes in favor of or against the goals and types of bilingual education. The final section brings together what we have learned from more than five decades of documentation and assessment of bilingual programs with current research on language acquisition and bilingualism.

Definitions

In a strict sense, for educational programs to qualify as 'bilingual,' the following conditions should be met: (1) at least two languages serve as media of instruction in subjects other than language classes, and (2) bilingualism, biliteracy, and biculturalism are explicit goals and emphatically supported by all stakeholders.

These criteria are not always adhered to by schools and programs claiming the label 'bilingual' for their various efforts in language teaching. Throughout the literature, *bilingual education* is also used in connection with bilingual families raising children with two or more first languages from birth. While my main focus here is on the fate of languages in schools, this article also mentions which findings from the research addressing bilingualism in early childhood and the study of bilingualism and multilingualism in general could be particularly pertinent to bilingual education as well.

For the purpose of this contribution, reference to bilingualism is meant to subsume multilingualism, and *acquisition* and *learning* are employed synonymously. *L1* refers to a child's first language(s), hence includes the simultaneous acquisition of two or more first languages, and *L2* is used as a shorthand expression for any language(s) acquired after age 3–4. Two additional terms need explication because usage is, to some extent, counter-intuitive: *majority* and *minority* (languages). Regardless of actual numbers, *majority* refers to linguistically 'powerful' speakers whose language rights are not in question. Their L1 typically reflects a society's accepted default, the socially *dominant*, and possibly, the constitutionally guaranteed *official* language. For majority children – even though one or two foreign languages, that is, languages not part of the local environment, may be mandatory subjects within their educational system – bilingualism is a matter of choice, in the ideal case a culturally and cognitively enriching experience, and it typically involves other, equally high-status majority

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92044-0

languages. The result is considered *additive*, not *subtractive* (Cummins, 1991; Baker, 2011). This means that becoming bilingual entails no threat to majority children's L1, their sense of identity and their social and cultural networks. Minority children, on the other hand, have no real choice since their L1 competence is irrelevant within a majority-language context. If they are to have a future within the mainstream educational system and beyond, they have to acquire the linguistic currency needed within the majority-dominated symbolic market.

Motivation and Aims of Bilingual Education

Over the past 50 years, bilingual curricula have emerged in many parts of the world (cf the overviews in Baker and Jones, 1998; García, 2009). Alongside globalization and internationalization, the following factors have been conducive to this development: (1) the coexistence of two or more official languages, as in Belgium, Canada, Luxembourg, Northern Italy, and Switzerland; (2) the coexistence of any number of indigenous languages alongside a non-indigenous, former colonial language functioning as a 'neutral' official language, such as English in Ghana and Ethiopia, French in the Ivory Coast, and Portuguese in Mozambique; (3) immigration and dramatic dropout rates of minority-language children in schools originally designed for majority children; and (4) grassroot movements rekindling interest in the revival of ethnic cultures and heritage languages.

Against this general background, three different educational objectives can be identified:

1. assimilation of immigrant and indigenous minority groups into mainstream society;
2. maintenance and development of the first language of minority children (immigrant and indigenous); and
3. linguistic and cultural enrichment and empowerment for any learner, that is, of minority and majority children alike.

Assimilation acknowledges at least a phase of transitional bilingualism. For Skutnabb-Kangas, assimilation reflects a linguistic variant of racism, 'linguicism' (1988 (1995: 40)), since it results in the "dominance of one language at the expense of others," for her a clear violation of linguistic human rights (1988 (1995: 45)). In transitional bilingualism "[l]anguage loss appears to be recommended" (Stubbs, 1991 (1995: 33)), which explains the critical stance taken by most, though not all, linguists (Hayakawa, 1992). For those favoring the second goal, maintenance and further development of children's L1, the following arguments play a particularly prominent role:

1. Children's sense of identity, awareness of their family's cultural roots, and their self-esteem are strengthened.
2. A curriculum with contents taught through their L1 allows minority children to keep in step with majority children as far as school subjects, world knowledge, and – if literalized in the L1 – cultural tools like reading and writing skills are concerned.
3. Language maintenance helps revitalizing minority languages threatened by extinction. This aspect matters particularly from the point of view of indigenous minorities.

The third general objective mentioned above, namely the enrichment and empowerment of all, is very much in line with current insights into cognitive advantages of bilingualism, such as enhanced metalinguistic awareness, theory of mind development, flexibility, executive control, and self-confidence with respect to language learning in general.

Prototypes of Bilingual Education: a Brief Overview

According to Baker (2011) and Baker and Jones (1998), educational policies can be positioned along a continuum from *weak* to *strong*, depending on how seriously and effectively they endorse bilingualism.

Strong Candidates

One intensively studied form of bilingual education consists in the *immersion* of whole classes of majority children in a second language, such as English L1 speakers in French L2 programs in Canada, which may begin in kindergarten (*early immersion*) or later grades (*late immersion*). Teaching in the L2 may be *partial* in the beginning, starting with a small selection of subjects (such as math and geography) or *total*, encompassing all subjects. In the case of *early total immersion*, a shift toward teaching at least some or half of the subjects in the L1 follows in later years. *Maintenance* or *Heritage* programs have also been counted among the strong educational options. Minority children are taught in their L2 and also in their L1 at least 50% of the time. Successful programs have led to the revitalization of Navajo in the United States, of Catalan, Gaelic, Finnish, Welsh, Ladinian, and other minority languages and dialects in Europe, of Maori in New Zealand, of immigrant languages in Israel, and of aboriginal languages in Australia.

Two Way or *Dual Language* education teaches through both minority and majority languages. Ideally, balanced numbers of native speakers of each language share a classroom, as in English–Spanish schools or in tandem with English and other languages in North American contexts. Languages employed may alternate, either by subject, day, or some other consistent principle. Baker (1993 (1995)) describes Welsh–English programs where native speakers of both, aged 11–16, are jointly taught in Welsh for at least 70% of the curriculum. Because equal attention is given to the languages, cultures, and speakers involved, *Dual Language* curricula have also been implemented as part of the peace process in ethnic conflict zones (Tankersley, 2001).

Instruction via two or more parallel languages, sometimes offering different choices of language pairs, is an important characteristic of *International Schools* and *European Schools*. The languages to choose from are, in most cases, high-status majority languages. Predominantly enrolled are children of internationally mobile families (e.g., diplomats, academics, members of international companies) and of local socioeconomic and intellectual elites.

With its explicit double focus on subject content and medium of instruction, *Content and Language Integrated Learning* (CLIL) has turned into a very popular European immersion variant over the past 20 years (cf Coyle et al., 2010). It deserves a place among the strong forms of bilingual

education, even though it may also be implemented on a small scale, that is, by individual teachers and within the microcosm of specific subjects classes in schools that don't offer a long-term or in-depth bilingual program. CLIL pedagogy stimulates authentic L2 communication among students in task-based interactions, and explicitly acknowledges the bilingual potential available in the classroom. Thus, the purist dogma of keeping languages strictly separate is replaced by a communicative style welcoming code-switching, which makes CLIL approaches very congenial to the goal of fostering the resourcefulness of bilingual speakers.

Weak Candidates

A 'sink or swim' educational policy known as *submersion* is the most widespread way of schooling individual minority children, both immigrant and indigenous. The child is placed in mainstream classrooms, sometimes assisted by additional instruction in the majority language, in the North American context called *withdrawal* or *pull-out* classes. Classes containing only minority children, while superficially looking very much like the immersion of majority children (often referred to as *structured* immersion) differ in the outcome, a point I return to in the following section.

An equally 'weak' form in terms of results and – by definition – goals consists in *transitional* models, which start by teaching in the minority child's L1, but as soon as possible, move over to instruction via L2.

Further Options along the Strong–Weak Continuum

Throughout childhood and beyond, language learners may well experience a mix of different teaching strategies and types of exposure to new languages or mother tongue instruction. Language policies aiming at mainstreaming may well tolerate bilingualism as a personal privilege, confining the 'other' languages to the invisible, private domain (cf also Stubbs, 1991).

Without support of minority languages at school, L1 instruction may be sought in other ways. This was, and sometimes still is, the case for children and grandchildren of the 'Gastarbeiter' (immigrant worker, 'guest worker') families, who first came to Germany about 50 years ago. Consulates of countries of origin, parental initiatives, and religious organizations may provide L1 classes in addition to submersion in the majority language.

Apart from motivation and long-term goals, implementation of the types of bilingual education mentioned here may differ significantly in conceptual, organizational, and administrative detail (cf Mehisto, 2012). Since terminology may differ too, it is not always easy to recognize where on the weak–strong continuum individual programs should be placed. In many countries, the French and English expression *immersion* has been borrowed as a superordinate term for what, especially in preschool settings, is considered to be the ideal assimilatory and submersive strategy for dealing with large groups of young minority children and, independently of any migration background, with children from socioeconomically disadvantaged milieus.

Additionally, there has been a rapidly increasing demand on the part of middle-class parents requesting foreign language programs for daycare centers and kindergartens. In response to this pressure, more and more preschool institutions offer various foreign language or second language activities, from an occasional, minimal encounter in the form of songs, games, and communicative routines to total immersion programs. Sometimes, bilingual kindergartens emulate the *one person–one language strategy* employed by bilingual families, with teachers, ideally native speakers, strictly requested to stick to only one language. Quite understandably, many middle-class bilingual parents who have consciously decided to raise their children with two simultaneous first languages are very eager to find family–external support for a non-dominant language spoken at home. The European Community also explicitly encourages parents from monolingual families to expose their children to additional (especially European) languages as a key to success in a globalized world (cf www.piccolingo.com). What is not often communicated (or simply not known), however, is that an offer of isolated lexical items (as in picture naming), or even recurrent songs and rhymes, while fun and harmless, does not provide effective keys for cracking the many codes of a new language unless these activities are supported by authentic communicative experience and guaranteed continuity, ideally within the educational system.

Findings and Challenges

In his overview of European initiatives Baetens Beardsmore (1993 (1995: 147)) concludes that "[n]o single model of bilingual education is universally applicable and no single existing model should be transplanted to a totally different context." He also cautions against unrealistic expectations: "The message to the educational planner, then, is not to expect bilingual education to produce native-like competence in two languages if the contextual variables do not allow for this to develop" (Baetens Beardsmore, 1993 (1995:148)). This last point stresses the basic unfairness of expecting an outcome simply impossible under unfavorable circumstances, such as gaps and a reduced spectrum of registers and styles in the input.

The weak–strong metaphor proposed in the literature and employed above is based on the relative success of forms of bilingual education, especially in the light of their aspirations. One of the most important findings emerging from the pioneering Canadian immersion experience, justifying its position at the 'strong' pole, was that L1 English majority children enrolled in French L2 immersion programs did not lag behind control groups of monolingual native speakers of English in their L1 proficiency. This means that in the case of majority speakers, length of exposure to L1-taught classes did not seem to matter. Neither did teaching science or math through children's incipient L2 interfere with their academic progress in these subjects. With respect to their emerging L2 proficiency, best results were achieved in early total immersion.

Maintenance programs, with instruction in both L1 and L2, also proved successful, but while they support children's L1 and ethnic identity, they may also segregate them from their majority peers. This is where Two Way/Dual Language schools have the clear advantage of bringing together relatively balanced numbers of students from both linguistic and cultural backgrounds and by giving equal status and rights to both

languages. Since Two Way/Dual Language schools typically rely on parental involvement and on the support of their communities, they offer – quite literally – a 'common ground' for families as well.

These various positive findings are in stark contrast with the results for learners participating in the other, hence 'weaker' group. To this day, it is puzzling that the average minority child performs below the norm in submersion classes and responds less well to what superficially look like challenges identical to those mastered by majority children, namely the simultaneous exposure to a new language and new contents, and that they do not even perform well in transition classes. To make things worse: it turned out that minority children underperformed not just in their L2 but also in their L1. To be sure, there are many individuals from minority cultures who manage to thrive even under submersion conditions, and who consequently believe that everybody else should be able to succeed as well, and that any special language-related treatment of minorities is some new racism in disguise (for instance Hayakawa, 1992). All in all, available evidence allowed the conclusion that simply more majority language instruction did not pay off, at least not to the extent expected, and not for the average minority child.

Cummins (1991) and in many other publications explains this majority–minority achievement discrepancy by proposing that cognitive benefits of bilingualism could only emerge once both languages had reached a certain *threshold* or level of proficiency, and that an advanced level of L1 before L2 exposure was a good predictor for successful L2 acquisition as well, an assumption known as the *interdependence hypothesis*. That a relatively advanced L1 should prove more resistant to the growing pressure from a prestigious L2, and that skills and concepts already mastered in L1, including literacy-related expertise, should be transferrable from L1 to L2, sounded entirely reasonable and has had a huge impact on pedagogical studies. World knowledge acquired via L1 as well as domain-general cognitive skills relating to memory organization, problem solving, inference, etc. should remain accessible. However, the attempt to identify concrete linguistic system properties which could offer themselves as a basis for transfer, quickly leads to an impasse since in many cases the result would hardly be helpful and would have to be 'undone' again. To mention just one concrete example, an L1 without articles (such as Russian or Turkish) does not offer itself as a fruitful model for a language which requires them (e.g., French, English, and German). In this case, ignoring L2 articles results in negative transfer (interference). As we know from older L2 learners of any target language, negative transfer due to solidly entrenched L1 properties which represents a problem rather than a solution.

One important overall message to be gained from years of studying the effectiveness of bilingual education is that positive effects require time. While learners very quickly (within 2–3 years) pick up elementary communicative routines, which Cummins (1991) refers to as *Basic Interpersonal Communicative Skills* (BICS), decontextualized registers relying heavily on written styles (*Cognitive/Academic Language Proficiency*, CALP) require more time (5–7 years; numbers proposed differ across studies). As argued by Baker (2011), BICS and CALP cannot be distinguished in a clear-cut manner. From a linguistic point of view, the distinction underestimates the inferential know-how and the intention-reading needed for the interpretation of informal speech. That is, in terms of their pragmatics, BICS are far from simple. One additional drawback Baker stresses is that "[s]uch terms may be used to label and stereotype students, especially if BICS are seen as inferior to CALP […]" (2011: 171). Learners who have never been exposed to academically oriented varieties simply have no reason to even miss them.

One is forced to conclude that the outcome of bilingual education depends on many interacting variables, some of them poorly understood to this day. What complicates matters is that developmental L1 norms are available for only a tiny fraction of the 6000 to 7000 existing languages known today. In the case of L2 development, there are hardly any norms based on large samples taking into account the age of onset and the duration of exposure. This also means that most assessments of bilingual children's competence is based on L1 norms, which makes poorer results highly predictable and raises, once more, questions of fairness.

One particularly relevant variable stressed in virtually all evaluations concerns the teachers and their professional competence, commitment, and, ideally, their function as convincing role models. For some researchers, the involvement of bilingual teachers is an absolute essential. Skutnabb-Kangas goes as far as to suggest that "[a] teacher, monolingual in L2, can never be a really good L2 teacher! A good L2 teacher knows both languages" (1988 (1995: 51)). While this is a desirable goal in general, providing all minorities with teachers who also know their L1 may well be too much to ask in the case of linguistically heterogeneous classrooms with children from many different first languages. Not only is the spectrum of first languages in contact within the same classroom a daunting pedagogical and linguistic challenge. Children born into families with an immigrant history may have most likely been exposed to some linguistic variety at home which already underwent significant change due to loss of close contact with a pre-immigration dialect of origin and due to intensive contact with new languages. Polinsky and Kagan (2007) argue that it is this variety that should be used as an appropriate and fair *baseline* against which a learner's competence ought to be compared, not some idealized norm that he or she has never been exposed to. It is crucial therefore that teachers involved in L1 instruction have a rough idea at least of how the variety he or she takes to be an appropriate target, differs from the heritage language internalized by learners. Polinsky and Kagan (2007: 374) describe what typically happens when even highly motivated heritage speakers who had shifted to some L2 early on, attempt to build on their rudimentary L1 competence: "[…] the heritage learner's competence is always suspect, and both instructors and curriculum designers take stock only of their deficiencies" and "[t]he problem is acerbated in cases where language instructors are biased against the language variety that served as the home language" (Polinsky and Kagan, 2007).

The question of which variety to pick as language of instruction in a bilingual educational setting and how to make sure that individual learners can relate what he/she knows to the instructional medium is far from trivial. It raises the question of how one can do justice to all children's linguistic rights in the absence of bilingual curricula and material, appropriate tests, and competent bilingual teachers for each and all

language combinations already spoken by children in schools and preschools all over the world. Educational contexts involving learners with diverse language backgrounds and differing degrees of proficiency in their heritage languages therefore pose a tremendous challenge for any educational system, but it also holds a great promise. In the light of this diversity it would seem sensible to start with a strong focus on a general subject, *Language* or *Language awareness*, as an appropriate forum where all learners could be encouraged to reflect on their linguistic repertoires and, in particular, on their bilingual resourcefulness.

Bilingualism and Education: a Brief Historical Excursion

In their overview of the history of bilingual education in the United States, Baker and Jones (1998) (see also Baker, 2011; Lyons, 1990) distinguish four overlapping periods characterized by what they call (1) permissiveness, (2) restrictiveness, (3) opportunity, and (4) dismissal. Before World War I, tolerance of linguistic diversity prevailed. Many ethnic communities taught children in languages other than English. But the Nationality Act of 1906 required immigrants to learn English in order to become naturalized citizens. World War I and new immigration waves led to calls for assimilation and Americanization, a development strongly reinforced by World War II, and to the restriction of public school instruction in languages other than English. This attitude relaxed in the wake of the Civil Rights and Equal Education movements in the 1960s and a growing interest in ethnic values and traditions. In addition, the persisting academic failure of children considered *Limited English Proficient Children* raised the issue of equal opportunity and led to calls for compensatory measures. Courts mandated bilingual education for minorities, and amendments to Education Acts made federal funds available. A next period, starting in the 1980s, took a turn in favor of submersion and transition programs, vigorously supported by pressure groups such as US English, English First, or English for the Children and, most recently, No Child Left Behind (see respective websites). Currently, yearly assessments are supposed to guarantee that schools and teachers make sure that students move into mainstream classes as soon as possible. As a consequence, L1 maintenance is typically dismissed.

The overall history of bilingual education for both immigrants and indigenous minorities in North America clearly reveals an intricate connection between ideology, power, and educational policy. Without doubt, conservative forces anywhere will continue to consider bilingualism in the public domain "a costly and confusing bureaucratic nightmare" (Hayakawa, 1992: 44) and a source of political disloyalty and destabilization. While nobody seriously objects to teaching high-status foreign languages to elites, bilingual education involving low-status minority languages is seen as the cause of scholastic underachievement. It may be a small comfort to supporters of bilingual (maintenance) programs that *Dual Language education/Two Way immersion* programs continue gaining recognition. Not only do they represent 'the antithesis of assimilation' (Baker, 2011: 229), but they also demonstrate that empowerment of minority and majority children is possible. It will be interesting to observe how the European commitment to 'unity in diversity' and Europe's vision of sustainable multilingualism and, moreover, a proclaimed 'consolidating' factor will fare in the future. The attempt to motivate children and adults to go for the languages of the smaller European countries is also meant to soothe concerns within these countries that their languages and cultures may be 'assimilated' as well and hence marginalized and wiped out in the long run. Given the rate at which threatened language die out, this expectation is not unrealistic. At the same time it can be shown that interest in the revival and survival of ethnic languages can be rekindled through well-thought-out educational planning in which strong bilingual programs play an important role.

Bilingual Resilience

In the past, supporters of bilingual education have had to contend not just with those who fear that multilingualism and cultural diversity lead to societal disintegration and social unrest, but also with those who, in principle, value the mastery of additional languages, especially those carrying prestige, as a cultural ideal and an important cognitive asset but who are nevertheless concerned about negative consequences for young children's cognitive, linguistic, and emotional development. Hence, in addition to concerns about an early shift to some majority language and the subsequent deterioration and attrition of minority languages, there is the more general question of an optimal time of first exposure, regardless of status differences. While myths fed by ideology and xenophobia are hard to eradicate, anxieties about children's ability to cope with coexisting languages can take comfort from decades of experience with bilingual education, and, in addition from our current understanding of bilingualism in general and in early childhood. Psycholinguistic experiments involving both children and adults have by now revealed a number of advantages of bilingualism in comparison with monolinguals with respect to executive control, metalinguistic awareness, divergent thinking, and self-esteem, to mention but a few (cf Bialystok et al., 2004); but some drawbacks can be identified as well, for instance delayed lexical access and somewhat slowed-down lexical decisions (cf de Bot et al., 2005; Grosjean, 2008).

Bilingualism is a natural state of the human brain, and due to its demands on self-monitoring and control – after all, the language not needed at any particular moment has to be inhibited – it keeps us mentally agile and appears to delay cognitive aging (cf Bialystok et al., 2004). Nevertheless, the long-term coexistence of more than one linguistic system within a single head hardly remains without consequences. Grosjean (2008: 9; see also 1982) criticizes what he calls a 'fractional view' according to which bilinguals are expected to behave like two monolinguals in one body and to keep their languages neatly apart. A more realistic view of bilingualism acknowledges the existence of tight cross-linguistic networks, with coactivation and competition a natural result of the coexistence of different linguistic systems within a single head. Perfectly balanced repertoires are unlikely, especially in bilinguals employing their languages according to some

division-of-labor principle, aligning their languages with different contexts, topics, or functions. Cross-linguistic coactivation and priming enhance the smooth and fluent transitions between languages known as *code switching* (cf Bullock and Toribio, 2009; Gardner-Chloros, 2009). Of course, what starts as formally and functionally highly skilled performance due to coactivation may eventually lead to a convergence of subsystems and, hence, to individual language change. The result would be some new, hybrid system, hopefully acquired as an L1 or an L2 by future generations.

Finally, it is instructive to look at very early bilingualism, especially at children growing up with two simultaneous first languages (cf Genesee and Nicoladis, 2007; Meisel, 2004). Since it is unlikely that their input is evenly distributed across different languages, we would not expect children to come up with perfectly balanced and equivalent systems. Longitudinal case studies show that children exposed to simultaneous first languages are very good at constructing parallel albeit separate representation systems. Evidence comes, for instance, from the asynchronic development of grammatical subsystems, with different languages in the lead for specific subsystems (Tracy and Gawlitzek-Maiwald, 2005).

Clearly, children are not thrown off track by the very presence of two (or more) linguistic systems in their environment, even though language pairs may make it easier or more difficult for them, or for learners of any age, to work out similarities and differences. It must also be recognized, though, that not all children exposed to two languages from birth will embrace them actively, speak them equally fluently, or maintain them throughout their lifetime. Whatever happens in the long run will depend both on opportunity (provided by the environment) and relevance (possibly a very personal decision). Crucially, and relevant within the current context, early bilingualism is living proof that there is no need for one system to reach a certain threshold before another is allowed to enter the stage. The ability of preschool children to cope with more than one language at a time is also stressed in recent studies of early L2 exposure, that is, of cases where the encounter with a second language takes place around age 3 to 4 within kindergarten settings (cf Rothweiler, 2006).

The relevance of all these findings to the topic of bilingual education should be clear. Bilingualism from birth and the early successive acquisition of a second language show that the cognitive and emotional capacity for handling multiple linguistic systems simultaneously exists from early on. On the basis of what we know about bilingual competence and performance, that is, about the cross-linguistic coactivation of resources in adults and in children alike, we should be able to develop a more appreciative perspective of what learners can achieve in bilingual education programs.

Conclusion

There is a vast body of publications on bilingual education. This short overview could not even begin to do justice to the overall spectrum of approaches. The topic is naturally complex because language and language behavior are complex, and because bilingual education involves a multitude of variables, not all of them well understood. Nevertheless, there is enough evidence to support the conclusion that bilingual education in the strict sense, as presented in the first section works, if given a fair chance. Whether one wants it to work, given its ambitious goal of sustaining bilingualism, biliteracy, and biculturalism, is an entirely different matter. The fact that bilingual education, in particular for minority children, still has a questionable reputation is remarkable and means that it is still not appreciated that "[b]ilingual education, when effectively implemented, is not the cause of underachievement; rather it is the cure" (Baker, 2011: 201).

In the past and sometimes well into the present, schools have prohibited and punished the use of minority languages on their grounds (Grosjean, 1982: 27ff). What we should have learned from bilingual education programs is that banning or ignoring resources children already bring with them to schools is not just wasteful and unfair but also counterproductive. Bilingual education can be much more than an educational objective or a gateway for careers in a tightly interconnected world. If done well, it explicitly allows all children to be proud and confident of what they know and to build on the natural fascination of young learners with language-related puzzles. Finally, it is to be hoped that more and more bilingual children grow into adults willing to cast their votes – quite literally, and in whatever language it is called for – in favor of high-quality education welcoming all languages and enriching their speakers.

See also: First Language Acquisition, Linguistic Theory of; Language Acquisition; Second Language Acquisition.

Bibliography

de Bot, K., Lowie, W., Verspoor, M., 2005. Second Language Acquisition. Routledge, London.

Baetens Beardsmore, H., 1993. European models of bilingual education: practice, theory and development. In: Jones, G.M., Ozog, A.C.K. (Eds.), Bilingualism and National Development. Multilingual Matters, Clevedon, pp. 103–121. García, O., Baker, C. (Eds.), 1995. Policy and Practice in Bilingual Education. Multilingual Matters, Clevedon, pp. 139–151.

Baker, C., 1993. Bilingual education in Wales. In: Baetens Beardsmore, H. (Ed.), European Models of Bilingual Education. Multilingual Matters, Clevedon, pp. 7–29. García, O., Baker, C. (Eds.), 1995. Policy and Practice in Bilingual Education. Multilingual Matters, Clevedon, pp. 152–165.

Baker, C., 2011. Foundations of Bilingual Education and Bilingualism, fifth ed. Multilingual Matters, Clevedon, UK.

Baker, C., Jones, S.P., 1998. Encyclopedia of Bilingualism and Bilingual Education. Multilingual Matters, Clevedon, UK.

Bialystok, E., Craik, F.I.M., Klein, R., Viswanathan, M., 2004. Bilingualism, aging, and cognitive control: evidence from the Simon task. Psychology and Aging 19 (2), 290–303.

Bullock, B., Toribio, A.J. (Eds.), 2009. The Cambridge Handbook of Linguistic Code-switching. CUP, Cambridge.

Coyle, D., Hood, Ph, Marsh, D., 2010. Content and Language Integrated Learning. Cambridge University Press, Cambridge.

Cummins, J., 1991. Interdependence of first- and second-language proficiency in bilingual children. In: Bialystok, E. (Ed.), Language Processing in Bilingual Children. Cambridge University Press, New York, pp. 70–89.

García, O., 2009. Bilingual Education in the 21st Century: A Global Perspective. Blackwell, Oxford.

Gardner-Chloros, P., 2009. Code-switching. Cambridge University Press, Cambridge.

Genesee, F., Nicoladis, E., 2007. Bilingual first language acquisition. In: Hoff, E., Shatz, M. (Eds.), Blackwell Handbook of Language Development. Blackwell, Oxford, pp. 324–342.

Grosjean, F., 1982. Life with Two Languages. Harvard University Press, Cambridge, MA.
Grosjean, F., 2008. Studying Bilinguals. Oxford University Press, Oxford.
Hakuta, K., 1986. The Mirror of Language. The Debate on Bilingualism. Basic Books, New York.
Hayakawa, S.I., 1992. Bilingualism in America: English should be the only language. In: Goshgarian, G. (Ed.), Exploring Language. Harper Collins, New York, pp. 42–47.
Lyons, J.J., 1990. The past and future directions of federal bilingual education policy. Annals of the American Academy of Political and Social Science 508, 66–80. García, O., Baker, C. (Eds.), 1995. Policy and Practice in Bilingual Education. Multilingual Matters, Clevedon. pp. 1–14.
Mehisto, P., 2012. Excellence in Bilingual Education. A Guide for School Principals. Cambridge University Press, Cambridge.
Meisel, J.M., 2004. The bilingual child. In: Bhatia, T.K., Ritchie, W.C. (Eds.), The Handbook of Bilingualism. Blackwell, Malden, MA, pp. 91–113.
Polinsky, M., Kagan, O., 2007. Heritage Languages: In the 'wild' and in the classroom. Language and Linguistics Compass 1 (5), 368–395.
Rothweiler, M., 2006. The acquisition of V2 and subordinate clauses in early successive acquisition of German. In: Lleó, C. (Ed.), Interfaces in Multilingualism. John Benjamins Publishing Company, Amsterdam, pp. 91–113.
Skutnabb-Kangas, T., 1988. Multilingualism and the education of minority children. In: Skutnabb-Kangas, T., Cummins, J. (Eds.), Minority Education: From Shame to Struggle. Multilingual Matters, Clevedon, pp. 9–44. García, O., Baker, C. (Eds.), 1995. Policy and Practice in Bilingual Education. Multilingual Matters, Clevedon, pp. 40–62.
Stubbs, M., 1991. Educational language planning in England and Wales: multicultural rhetoric and assimilationist assumptions. In: Coulmas, F. (Ed.), A Language Policy for the European Community. Mouton de Gruyter, Berlin, pp. 215–239. García, O., Baker, C. (Eds.), 1995. Policy and Practice in Bilingual Education. Multilingual Matters, Clevedon, pp. 25–39.
Tankersley, D., 2001. Bombs or bilingual programs? dual language immersion, transformative education and community building in Macedonia. International Journal of Bilingual Education and Bilingualism 4 (2), 408–424.
Tracy, R., Gawlitzek-Maiwald, I., 2005. The strength of the weak: asynchronies in the simultaneous acquisition of German and English. Zeitschrift für Literaturwissenschaft und Linguistik 139, 28–53.

Relevant Websites

http://ec.europa.eu/languages/index_en.htm – European Language Policy.
www.piccolingo.com – Fostering Bilingualism in Early Childhood.
http://www2.ed.gov/nclb/landing.jhtml – No Child Left Behind.
http://www.us-english.org/ – Promotion of English in the U.S.

Binary Response Models and Logistic Regression

Garrett M Fitzmaurice and Nan M Laird, Harvard School of Public Health, Harvard University, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The focus of this article is on regression models for a binary response. In most fields of application, logistic regression has become the standard method for relating a binary response to a set of covariates. In this article the main features of logistic regression are described and some aspects of interpretation of logistic regression are illustrated with an example. The logistic regression model is first introduced for the simple case where there is only a single predictor variable. This model is compared and contrasted with the classical linear regression model. Later the generalizations to more than one predictor variable are considered. A major emphasis of this article is placed on how logistic regression is used in practice and how the logistic regression coefficients should be interpreted. An example, based on data from the US General Social Survey, is used to illustrate and reinforce the main concepts. Finally, three important extensions of standard logistic regression are briefly reviewed: conditional logistic regression, exact logistic regression, and logistic regression for clustered data.

This article describes regression models for the relationship between a binary response variable (e.g., 'success' or 'failure') and a set of predictor variables. Among the many binary response models that have been proposed in the statistical literature, logistic regression has become the standard model for analyzing the effects of a set of predictor variables; logistic regression is ubiquitous in most fields. In common with linear regression (e.g., *see* Linear Hypothesis: Regression (Basics)), the primary objective of logistic regression is to model the mean of the response variable, given a set of predictor variables. However, what distinguishes logistic regression from linear regression is that the response variable is binary rather than continuous in nature. This has a number of consequences for modeling the mean of the response variable. In this article the main features of logistic regression are described and some aspects of interpretation of logistic regression are illustrated with an example.

Introduction to Binary Response Models

In this section, we introduce some notation and consider regression models for a binary response variable. Let Y denote a binary response variable, that is, a discrete response variable having only two categories, for convenience often referred to as 'success' or 'failure.' For example, Y might indicate an individual's political party preference (Democrat or Republican, in the United States), an individual's opinion on the death penalty (favor or oppose), or an individual's self-reported happiness (happy or unhappy). Denoting the two possible outcomes for Y by 0 and 1, the probability distribution of the response variable is the Bernoulli distribution, with $\Pr(Y=1)=\pi$ (and, correspondingly, $\Pr(Y=0)=1-\pi$). The primary goal is to describe the effect on π of changes in a set of predictor variables. For ease of exposition, we will first consider the simple case where there is only a single predictor variable, x. Generalizations to more than one predictor variable will be considered later in the article.

Given that the main analytic goal is to investigate the relationship between π and x, there are a number of alternative binary response models that can be considered. Recall that with a binary response, the mean of Y is $\pi=\Pr(Y=1)$, i.e., the mean of Y is simply the probability that Y takes on the value 1. Since linear models (e.g., *see* Linear Hypothesis: Regression (Basics); Analysis of Variance and Generalized Linear Models) play such an important and dominant role in applied statistics, it may at first seem natural to assume a linear model relating the mean of Y to x,

$$E(Y|x) = \pi = \beta_0 + \beta_1 x \qquad [1]$$

However, this linear model for the probabilities has one obvious difficulty. Expressing π as a linear function violates the restriction that probabilities must lie within the range from 0 to 1, unless constraints are placed on x or on β_0 and β_1. That is, for sufficiently large or small values of x (when $\beta_1 \neq 0$), the linear model given by eqn [1] will yield predicted probabilities outside of the range from 0 to 1. A further difficulty with the linear model for the probabilities is that we often expect there to be a nonlinear relationship between π and x. For example, a 0.2-unit increase in π might be considered more 'extreme' when $\pi=0.1$ than when $\pi=0.5$. In terms of ratios, the change from $\pi=0.1$ to $\pi=0.3$ represents a threefold (0.3/0.1) or 200% increase, whereas the change from $\pi=0.5$ to $\pi=0.7$ represents only a 40% (or 0.7/0.5) increase. In a sense, the units of measurement for a probability or proportion are often not considered to be constant over the range from 0 to 1. The linear probability model given by eqn [1] simply does not take this into consideration when relating π to x.

Because it is more natural to compare values of π in *relative* rather than *absolute* terms, this suggests fitting a linear model relating the logarithm of π to x

$$\log(\pi) = \beta_0 + \beta_1 x \qquad [2]$$

The model given by eqn [2] is often referred to as 'relative risk' regression. It can be easily shown that β_1 has interpretation as the log *relative risk* (associated with a unit increase in x). The relative risk is defined as the ratio of the probability of success ($Y=1$) for a single unit increase in x. For example, when the predictor variable x is dichotomous, taking values of 0 and 1 representing the absence ($x=0$) or presence ($x=1$) of some factor, β_1 has interpretation as the *difference* between the log of the probability of success ($Y=1$) when the factor is present

$(x=1)$ and the log of the probability of success when the factor is absent $(x=0)$, $\beta_1 = \log\{\pi(x=1)\} - \log\{\pi(x=0)\} = \log\left\{\frac{\pi(x=1)}{\pi(x=0)}\right\}$, where $\pi(x=1)$ and $\pi(x=0)$ denote the respective success probabilities (and recall from the rules of logarithms that $\log(a/b) = \log(a) - \log(b)$). Therefore, e^{β_1} has interpretation as the *ratio* of the probability of success when the factor is present $(x=1)$ to the probability of success when the factor is absent $(x=0)$, $e^{\beta_1} = \frac{\pi(x=1)}{\pi(x=0)}$. Relative risks are often considered the preferred measures of association in prospective studies, especially when the binary response of interest is common. In particular, many empirical researchers prefer relative risks because they find them to be simple and intuitively interpretable. However, the relative risk regression model given by eqn [2] shares some of the structural issues associated with the linear probability model. Although the predicted probabilities given by $\pi = e^{\beta_0+\beta_1 x}$ cannot be negative, they can still exceed one for sufficiently large or small values of x (when $\beta_1 \neq 0$).

To circumvent the aforementioned difficulty, alternative nonlinear transformations can be applied to π where the constraints on the predicted probabilities are automatically satisfied. In particular, a transformation of π, say $g(\pi)$, can be chosen so that it maps the range of π from $(0, 1)$ to $(-\infty, \infty)$. Since there are many possible transformations, $g(\pi)$, that achieve this goal, this leads to an extensive choice of models that are all of the form

$$g(\pi) = \beta_0 + \beta_1 x \qquad [3]$$

and where the transformed probabilities are related linearly to x. However, the most commonly used in practice are

1. logit or logistic function: $g(\pi) = \log[\pi/(1-\pi)]$;
2. probit or inverse normal function: $g(\pi) = \Phi^{-1}(\pi)$, where Φ is the standardized Gaussian cumulative distribution function; and
3. complementary log–log function: $g(\pi) = \log[-\log(1-\pi)]$.

We note that all of these transformations are very closely related when $0.2 < \pi < 0.8$, and in a sense only differ in the degree of 'tail-stretching' outside of this range. Indeed, for most practical purposes it is not possible to discriminate between a data analysis that is based on, for example, the logit and probit functions. To discriminate empirically between probit and logistic regression would, in general, require very large numbers of observations. However, the logit function does have a number of distinct advantages over the probit and complementary log–log functions that probably account for its more widespread use in practice. Later in this article we will consider some of the advantages of the logit function.

Logistic Regression

When the transformed probabilities are related linearly to x by the logit function,

$$\log[\pi/(1-\pi)] = \beta_0 + \beta_1 x \qquad [4]$$

this is known as the *logistic regression* model. Recall that if π is the probability of success, then $\frac{\pi}{1-\pi}$ is known as the *odds* of success; the latter term will be familiar to readers belonging to the gambling fraternity. For example, if the probability of success is 0.8 then the odds of success is 4 (or 0.8/0.2) to 1. That is, the probability of success is 4 times as large as the probability of failure. Thus, the logistic regression model assumes a linear relationship between the log odds (of success) and x. Note that this simple model can be expressed equivalently in terms of π,

$$\pi = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \qquad [5]$$

While the latter expression may appear to be somewhat more daunting, we simply wish to emphasize that eqns [4] and [5] are completely equivalent ways of expressing the logistic regression model. Expression [4] describes how the log odds, $\log\left(\frac{\pi}{1-\pi}\right)$, has a linear or straight-line relationship with x. Expression [5], on the other hand, describes how π has an S-shaped relationship with increasing values of $\beta_1 x$; although, in general, this relationship is approximately linear within the range $0.2 < \pi < 0.8$ (see **Figure 1** for a plot of π versus x when $\beta_0 = 0.5$ and $\beta_1 = 0.9$). Observe that the expression on the right of eqn [5] cannot yield a value that is either negative or greater than 1. That is, the logistic transformation ensures that the predicted probabilities are restricted to the range from 0 to 1. Finally, note that

$$1 - \pi = \frac{1}{1 + \exp(\beta_0 + \beta_1 x)}$$

so that the odds, $\frac{\pi}{1-\pi}$, is simply $\exp(\beta_0 + \beta_1 x)$, and the log odds can be expressed as $l = l(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x$.

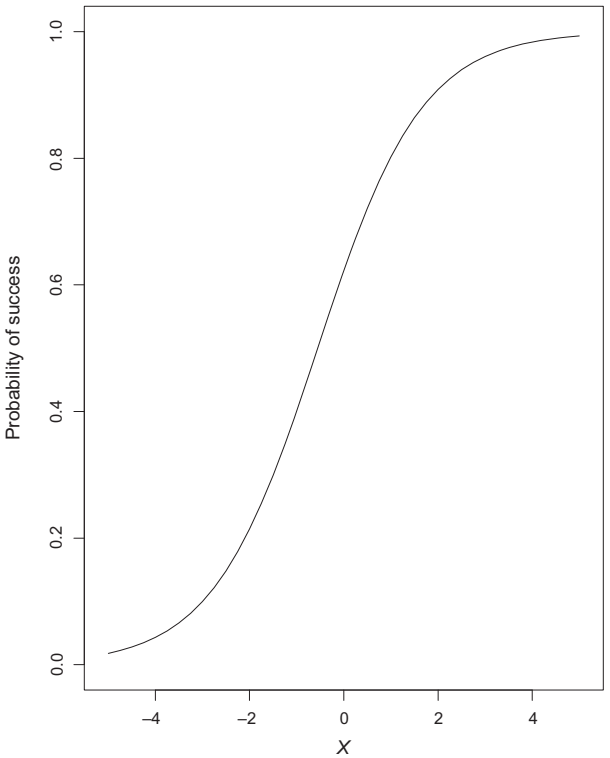

Figure 1 Plot of logistic regression function for logistic model: $\log\left(\frac{\pi}{1-\pi}\right) = 0.5 + 0.9X$.

Interpretation of Logistic Regression Coefficients

Next we consider the interpretation of the logistic regression coefficients, β_0 and β_1, in eqn [4]. In simple linear regression recall that the interpretation of the slope of the regression is in terms of changes in the mean of Y for a single unit change in x. Similarly, the logistic regression slope, β_1, in eqn [4] has interpretation as the change in the log odds (of success) for a unit change in x. Equivalently, a unit change in x increases or decreases the odds of success *multiplicatively* by a factor of $\exp(\beta_1)$. Also, recall that the intercept in simple linear regression has interpretation as the mean value of the response variable when x is equal to zero. Similarly, the logistic regression intercept, β_0, in eqn [4] has interpretation as the log odds (of success) when $x = 0$; alternatively, if we exploit the expression on the right of eqn [5], $\frac{\exp(\beta_0)}{1+\exp(\beta_0)}$ is the probability of success when $x = 0$. Although eqn [5] can be used to back-transform from the logit scale to the probability scale, we note that it is not possible to back-transform the regression slope, β_1, to the probability scale. This is due to the fact that a constant difference on the logit scale does not translate to a constant difference on the probability scale. For example, a single unit change in the log odds, say from 0 to 1, corresponds to a risk difference of 0.23 (or $\frac{e^1}{1+e^1} - \frac{e^0}{1+e^0} = 0.73 - 0.50$); in contrast, a single unit change in the log odds from 2 to 3 corresponds to a risk difference of 0.07 (or $\frac{e^3}{1+e^3} - \frac{e^2}{1+e^2} = 0.95 - 0.88$).

For the special case where the predictor variable x is dichotomous, taking values of 0 and 1, the logistic regression slope, β_1, has a simple and very attractive interpretation. Consider the two possible values for π when $x = 0$ and $x = 1$,

$$\pi(x = 0) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

(obtained by plugging-in or substituting the value 0 for x in the expression on the right of eqn [5]), and

$$\pi(x = 1) = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$$

(similarly obtained by plugging-in the value 1 for x in the expression on the right of eqn [5]). Then, we can construct a 2×2 table of the cross-classification of Y and x (see **Figure 2**). Notice that the odds, $\pi/(1 - \pi)$, has a very simple expression, $e^{\beta_0+\beta_1 x}$, and the log odds is just the linear predictor, $\beta_0 + \beta_1 x$. Thus it follows that the log of the odds ratio (or cross-product ratio) in this 2×2 table is equal to β_1, i.e.,

$$\begin{aligned}\log\,(\text{odds ratio}) &= \log\left\{\frac{\pi(x = 1) \times [1 - \pi(x = 0)]}{\pi(x = 0) \times [1 - \pi(x = 1)]}\right\} \\ &= l(x = 1) - l(x = 0) = (\beta_0 + \beta_1) - \beta_0 \\ &= \beta_1.\end{aligned}$$

Thus, $\exp(\beta_1)$ has interpretation as the odds ratio of the response for the two possible values of the predictor variable.

The odds ratio has many appealing properties that probably account for the widespread adoption of the logistic function (and, hence, logistic regression) over other functions such as the probit or complementary log–log functions. First, note that the odds ratio does not change when rows and columns of the 2×2 table are interchanged. This implies that it is not necessary to distinguish which variable is the response and predictor variable in order to estimate the odds ratio. Furthermore, a very appealing feature of logistic regression is that the odds ratio, $\exp(\beta_1)$, is equally valid whether the study design is prospective, cross-sectional, or retrospective; indeed, this property of the odds ratio holds for arbitrary x (e.g., a predictor variable x that is nondichotomous or quantitative). That is, logistic regression provides an estimate of the same association between Y and x regardless of whether we (1) randomly sample individuals having $x = 0$ and $x = 1$ and subsequently classify them according to whether $Y = 1$ or $Y = 0$ (prospective design); or (2) randomly sample individuals from a population and cross-classify them according to whether $x = 0$ or $x = 1$ and whether $Y = 1$ or $Y = 0$ (cross-sectional design); or (3) randomly sample individuals having $Y = 0$ and $Y = 1$ and subsequently classify them according to whether $x = 1$ or $x = 0$ (retrospective or case-control design). The case-control design is popular in epidemiological studies when the outcome Y is rare, and the other designs are impractical. In all three study designs logistic regression provides an estimate of the same association between Y and x. We note that the odds ratio is probably the only measure of association that has this property; although the *relative risk* is considered by many to be the preferred measure of association due to its simple and intuitive interpretation, it cannot be estimated from retrospective study designs.

		Y = 1	Y = 0	
x	1	$\pi(x=1) = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}$	$1-\pi(x=1) = \frac{1}{1+\exp(\beta_0+\beta_1)}$	1.0
	0	$\pi(x=0) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$	$1-\pi(x=0) = \frac{1}{1+\exp(\beta_0)}$	1.0

Figure 2 Probabilities of success and failure for logistic regression of Y on x.

Finally, in medical studies where Y typically denotes the presence or absence of a disease or disorder, the odds ratio is often interpreted as an approximation to the *relative risk* of disease. The relative risk of disease is simply the ratio of the probability of disease ($Y=1$) when a risk factor is present ($x=1$) to the probability of disease when a risk factor is absent ($x=0$), $\frac{\pi(x=1)}{\pi(x=0)}$. When the probability of disease is rare, and π is reasonably close to 0 in both of the risk groups (often known as the 'rare disease' assumption), the odds ratio provides a close approximation to the relative risk. Retrospective designs (case (3) above) are especially common in health and medical studies where the possible outcomes of interest are very rare. Although the relative risk cannot be estimated from a retrospective study, the odds ratio can be used to provide an approximation to the relative risk. However, extra care is necessary when interpreting the odds ratio as an approximation to the relative risk in prospective studies. In many prospective studies the binary event is relatively common (say greater than 10%) and the 'rare disease' assumption no longer holds; in these settings, the odds ratio can be a very poor and unreliable approximation to the relative risk and should not be given such an interpretation.

Multiple Logistic Regression

So far, we have only considered the simple case where there is a single predictor x. Next, we consider the extensions of eqns [4] and [5] to the case where there are two or more predictors. When there are many predictor variables, the logistic regression model becomes,

$$\log\left[\pi/(1-\pi)\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad [6]$$

where $x_1, x_2, \ldots, x_p$ are the outcomes of p predictor variables. Note that $x_1, x_2, \ldots, x_p$ may represent p distinct predictor variables; alternatively, they may represent q distinct predictor variables (where $q<p$) in addition to a subset of interactions among the predictors. The logistic regression coefficients in eqn [6] have the following interpretations. The logistic regression intercept, β_0, now has interpretation as the log odds (of success) when all predictor variables assume the value zero, i.e., when $x_1=x_2=\cdots=x_p=0$. Alternatively, $\frac{\exp(\beta_0)}{1+\exp(\beta_0)}$ is the probability of success when $x_1=x_2=\cdots=x_p=0$. Each of the logistic regression slopes, β_k (for $k=1, \ldots, p$), has interpretation as the change in the log odds (of success) for a unit change in x_k given that all of the other predictor variables remain constant. We note that this is completely analogous to the interpretation of the regression coefficients in multiple linear regression. Thus, holding the remaining predictors at some fixed set of values and not allowing them to vary with any changes in x_k, a single unit increase in x_k, is predicted to increase or decrease the log odds of success by an amount β_k. Equivalently, a single unit increase in x_k increases or decreases the odds of success *multiplicatively* by a factor of $\exp(\beta_k)$. Finally, we remark that the multiple logistic regression model given by eqn [6] can also be expressed equivalently in terms of π,

$$\pi = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right)}{1+\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right)} \qquad [7]$$

In Section Interpretation of Logistic Regression Coefficients, we remarked that one of the appealing properties of logistic regression is that the same odds ratio can be estimated from either a prospective or retrospective study design. This result readily generalizes when x is quantitative rather than dichotomous, and also when there are two or more predictor variables. Thus, in a retrospective study $\beta_1, \ldots, \beta_p$ can be estimated by analyzing the data as if they had been obtained from a prospective study. That is, the data from a retrospective study can be treated as prospective data in a logistic regression analysis where the binary outcome of interest is modeled as the response variable. A formal proof of this result can be found in Farewell (1979) and Prentice and Pyke (1979). The only parameter that cannot be estimated with data from a retrospective study is the intercept, β_0, since it is determined by the proportions of 'successes' ($Y=1$) and 'failures' ($Y=0$) selected into the study. In many studies, however, the intercept is not of direct scientific interest.

Estimation

In this section we very briefly discuss estimation of the logistic regression coefficients. This section is somewhat more technical in nature than the previous sections and can been skipped over on a first reading of this article. Recall that in linear regression we estimate the linear regression coefficients using the method of least squares. That is, the least squares criterion chooses values for the linear regression coefficients that minimize the sum of squared deviations of the observed Y from the predicted values of Y, say $\widehat{Y}$, under the assumed regression model. The least squares method yields estimates of the regression coefficients that are also the *maximum likelihood* estimates when the errors have a normal distribution with constant variance. We use the more general method of maximum likelihood estimation for logistic regression.

The essential idea behind maximum likelihood is quite simple, and in fact is conveyed by its name. The method chooses values of the logistic regression coefficients that are most likely (or most probable) to have generated the observed data. This is achieved by maximizing the *likelihood function* for the data. The likelihood function is simply the probability of the response variables evaluated at the observed values, regarded as a function of the unknown logistic regression coefficients. Construction of the likelihood function requires that we must make an assumption about the probability distribution of Y. Fortunately, with a binary response, the selection of a probability distribution is straightforward as there is only a single distribution to choose from, namely the Bernoulli distribution.

In the following we assume that there are independent observations of the response variables Y_i and the predictor variables $x_{i1}, x_{i2}, \ldots, x_{ip}$ available on N individuals. The N binary responses, Y_i ($i=1, \ldots, N$), can be regarded as independent Bernoulli random variables, with $\pi_i = E(Y_i) = \Pr(Y_i=1)$. Since the N binary responses are independent the joint probability of ($Y_1, \ldots, Y_N$) is the product of the N Bernoulli functions. Thus, the likelihood function can be expressed as the product,

$$\prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

It is this function, or equivalently, the logarithm of the likelihood function, that must be maximized. Note that the likelihood is a function of the unknown logistic regression coefficients, since

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})}$$

The maximum likelihood estimates of the logistic regression coefficients are obtained by substituting the above expression for π_i into the likelihood function and finding the value of the logistic regression coefficients that produce the largest value for the likelihood function. Ordinarily, the likelihood function has only a single maximum. Maximization of the likelihood, or the log likelihood, requires an iterative procedure and this has been implemented in many of the commercially available statistical software packages (e.g., SAS, STATA, and SPSS). A description of the iterative procedure can be found in, for example, McCullagh and Nelder (1989: pp. 40–43). Finally, estimates of the standard errors of the estimated logistic regression coefficients can readily be obtained using the method of maximum likelihood estimation.

Example

Next we consider an application of logistic regression to illustrate how the model can be used in practice. **Table 1**, adapted from Clogg (1982), contains data from 1518 individuals who participated in the US General Social Survey (Davis, 1977). The data are composed of a pair of 2×4 cross-classifications on the relationship between reported happiness (not too happy versus pretty or very happy) and completed years of schooling (less than 12, 12, 13–16, 17 or more), by the gender of the respondent. The goal of the analysis is to describe how reported happiness is related to years of schooling and the gender of the respondent. In particular, it is of interest to determine whether reported happiness is related to the amount of schooling received and whether this relationship is different for males and females.

We first consider a logistic regression model that includes effects of years of schooling, gender, and their interaction (model 1):

$$\log[\pi/(1-\pi)] = \beta_0 + \beta_1 G + \beta_2 \mathrm{YSC}_1 + \beta_3 \mathrm{YSC}_2 + \beta_4 \mathrm{YSC}_3 + \beta_5 G \times \mathrm{YSC}_1 + \beta_6 G \times \mathrm{YSC}_2 + \beta_7 G \times \mathrm{YSC}_3$$

Table 1 Cross-classification of U.S. sample according to reported happiness and years of schooling, by gender

		Years of school completed			
Gender	Reported happiness	<12	12	13–16	≥17
Males	Not too happy	40	21	14	3
	Pretty or very happy	213	177	167	54
Females	Not too happy	62	26	12	3
	Pretty or very happy	242	283	171	30

where $G = 1$ if female and $G = 0$ if male, and YSC_1, YSC_2, and YSC_3 are three *indicator variables* for the 4 levels of years of schooling. That is, $\mathrm{YSC}_1 = 1$ if years of schooling <12 and $\mathrm{YSC}_1 = 0$ otherwise; $\mathrm{YSC}_2 = 1$ if years of schooling is 12 and $\mathrm{YSC}_2 = 0$ otherwise; $\mathrm{YSC}_3 = 1$ if years of schooling is 13–16 and $\mathrm{YSC}_3 = 0$ otherwise. Maximum likelihood estimates of the regression coefficients for model 1 are given at the top of **Table 2**.

Note that since all of the predictor variables are discrete, this is a *saturated model* and will perfectly fit or predict the data in **Table 1**. Examination of **Table 2** reveals that the three estimated coefficients for the interaction of gender and years of schooling are relatively small when compared with their respective standard errors. A formal test of the interaction can be constructed by fitting a logistic regression model that omits the interaction and comparing the likelihoods for the two models (e.g., *see* Hypothesis Testing in Statistics). Next, we consider a logistic regression model that includes only the effects of years of schooling and gender (model 2):

$$\log[\pi/(1-\pi)] = \beta_0 + \beta_1 G + \beta_2 \mathrm{YSC}_1 + \beta_3 \mathrm{YSC}_2 + \beta_4 \mathrm{YSC}_3$$

Maximum likelihood estimates of the logistic regression coefficients in model 2 are given in the middle of **Table 2**. Comparing model 2 to model 1 in terms of their likelihoods, we can formally test the null hypothesis of no interaction, $H_0{:}\beta_5 = \beta_6 = \beta_7 = 0$. The likelihood-ratio chi-squared statistic is 2.99 on 3 degrees of freedom ($p \approx 0.4$). Thus, we cannot reject $H_0{:}\beta_5 = \beta_6 = \beta_7 = 0$, and model 2, without an interaction between gender and years of schooling, is defensible. This implies that the relationship between reported happiness and years of schooling is the same for males and females. Examination of the estimated regression coefficients for years of schooling in model 2 suggest that a further simplification of the model might be possible. The effect of years of schooling on reported happiness appears to be greatest for those individuals with less than 12 years versus those having 12 or more years of education. Thus, we next consider the following logistic regression model (model 3):

$$\log[\pi/(1-\pi)] = \beta_0 + \beta_1 G + \beta_2 \mathrm{YSC}_1$$

Note that in this model we have combined the levels of years of schooling completed to form a dichotomy: less than 12 years versus 12 or more years. Maximum likelihood estimates of the regression coefficients in model 3 are given at the bottom of **Table 2**. Comparing model 3 to model 2 in terms of their likelihoods provides a formal test of the null hypothesis, $H_0{:}\beta_3 = \beta_4 = 0$. The likelihood-ratio chi-squared statistic is 1.46 on 2 degrees of freedom ($p \approx 0.5$) and thus the null hypothesis cannot be rejected. The estimated logistic regression coefficients for model 3 indicate that the overall reported rates of happiness are very similar for males and females but that the rates of reported happiness are substantially lower for those with less than 12 years of education. In particular, the odds of reported happiness for those with 12 or more years of schooling is 2.5 (or $\exp[0.9178] = 2.50$) times greater (with 95% likelihood-based confidence interval: 1.83, 3.44) than for those with less than 12 years of schooling. That is, the ratio of odds of reported happiness is 2.5 when those with more than 12 years of schooling are compared to those with less than 12 years of schooling. Finally, the nature of this

Table 2 Logistic regression parameter estimates and standard errors under three models for the probability of reporting happiness

Model	*Parameter*	*Estimate*	*Standard error*	*Wald* χ^2	*Wald* P	*LR* χ^2	*LR* P
Model 1	Intercept	2.890 4	0.593 2	23.74	<.000 1	55.51	<0.000 1
	G	−0.587 8	0.847 7	0.48	0.488 0	0.48	0.490 5
	YSC_1	−1.218 0	0.617 7	3.89	0.048 6	5.24	0.022 1
	YSC_2	−0.758 7	0.636 5	1.42	0.233 2	1.67	0.196 5
	YSC_3	−0.411 4	0.655 2	0.66	0.530 0	0.43	0.514 3
	$G \times YSC_1$	0.277 2	0.876 6	0.10	0.751 9	0.10	0.752 1
	$G \times YSC_2$	0.843 5	0.902 1	0.87	0.349 8	0.86	0.353 7
	$G \times YSC_3$	0.765 6	0.940 8	0.66	0.415 8	0.66	0.418 0
Model 2	Intercept	2.672 8	0.427 2	39.14	<0.000 1	78.37	<0.000 1
	G	−0.089 8	0.162 5	0.31	0.580 5	0.31	0.580 0
	YSC_1	−1.127 8	0.437 5	6.65	0.009 9	8.75	0.003 1
	YSC_2	−0.336 2	0.451 2	0.56	0.456 2	0.60	0.439 8
	YSC_3	−0.061 8	0.469 6	0.02	0.895 2	0.02	0.894 7
Model 3	Intercept	2.473 2	0.149 1	274.98	<0.000 1	453.71	<0.000 1
	G	−0.108 5	0.161 6	0.45	0.502 2	0.45	0.501 4
	YSC_1	−0.917 8	0.160 6	32.64	<0.000 1	32.91	<0.000 1

Note LR: Likelihood-ratio.

relationship between years of schooling and happiness is the same for males and females.

Extensions of Logistic Regression

In this section we briefly review a number of more advanced topics that can be considered extensions of the standard logistic regression model: conditional logistic regression, exact logistic regression, and logistic regression for clustered data.

Conditional Logistic Regression

The previous sections showed that logistic regression is a very versatile model for assessing the associations between a binary response and a set of predictors. This section introduces a related technique known as *conditional logistic regression*, which extends many of the benefits of logistic regression to studies with *matched* designs. In matched study designs, individuals are stratified on the basis of variables thought to be related to the outcome variable of interest. For example, age and socioeconomic status are two variables commonly used to define strata in many studies in the social and behavioral sciences. The conditional logistic regression model used to analyze matched data is

$$\log\left[\pi_j/(1-\pi_j)\right] = \alpha_{j0} + \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_p x_{jp} \qquad [8]$$

Note that this is similar to the standard logistic regression model given by eqn [6], except that the probability of success and the predictors are now indexed by j to indicate that they apply to individuals (or units) from the jth stratum (e.g., matched pair), $j = 1, \ldots, S$. There can be any number of individuals in the jth stratum; however, for estimation (discussed later) there must be at least one success ($Y = 1$) and one failure ($Y = 0$) in each stratum. Note also that the common intercept β_0 in standard logistic regression has been replaced in eqn [8] by S stratum-specific intercepts denoted by α_{j0} ($j = 1, \ldots, S$), but the remaining regression coefficients are assumed constant over the strata.

Although the conditional logistic regression model bears much similarity to the standard (or unconditional) logistic regression model, there are some subtle but important differences. First, the stratum-specific intercepts, α_{j0}, are considered nuisance parameters and cannot be estimated when there are few observations in each of the strata. This is usually not of any concern since, as for standard logistic regression, these intercepts are generally not of scientific interest. Second, because the model includes stratum-specific intercepts, the β's now have stratum-specific interpretations in terms of changes in the log odds of success for within-stratum changes in the predictors. That is, the β's have interpretation assuming the level of the matching factors is held constant. For example, β_k has interpretation in terms of changes in the log odds of success for a single unit change in x_{jk} *within the* jth *stratum* (i.e., comparing two individuals within the same stratum that happen to differ by one unit in the kth predictor). Third, the associations between any variables used for matching (or any other covariates that are constant within strata) and the outcome cannot be quantified. This is because the method of estimation is based entirely on variation within a stratum; conditional logistic regression cannot be used to estimate the effect of a covariate that varies only between strata (but not within a stratum).

Next we consider estimation of the conditional logistic regression model parameters. One naive approach to fitting this model would be to attempt to estimate all of the parameters, including the stratum-specific intercepts. However, for matched designs, the number of strata grows as the sample size increases, which means that the number of parameters would be large relative to the sample size no matter how large a sample was collected. For example, in a matched-pair design with N pairs (i.e., two subjects in each stratum), such an analysis would require the estimation of $N + p$ parameters from a sample of only $2N$ observations. It should not be too surprising that this proliferation of stratum-specific intercepts

causes problems for estimation; when the number of these nuisance parameters is comparable to the number of observations, the maximum likelihood estimates of the β's may be far from their true values. In general, the standard properties of maximum likelihood break down when there are few observations in each strata. To circumvent this, a variant of the approach, *conditional maximum likelihood*, can be used in these settings. Conditional logistic regression maximizes a likelihood for the conditional distribution (and hence the term 'conditional' logistic regression) that eliminates these stratum-specific intercepts and bases estimation of the associations between the predictors and response entirely on information from within the strata. As such, conditional logistic regression allows for estimation of the β's, the parameters of real scientific interest, at the expense of ignoring those parameters of far less interest. In order to estimate the β's it is necessary to have at least one success and one failure in each stratum. The term conditional means we condition our results on the successes and failures recorded in each stratum. This enables the use of logistic regression with matched designs.

Exact Logistic Regression

Recall that in simple logistic regression with a single dichotomous predictor, x, a test for no association ($\beta_1 = 0$) can be constructed based on the *Wald test* statistic, given by:

$$Z = \frac{\widehat{\beta}_1}{\sqrt{\widehat{\mathrm{Var}}\left(\widehat{\beta}_1\right)}} = \frac{\widehat{\beta}_1}{\widehat{\mathrm{SE}}\left(\widehat{\beta}_1\right)}$$

which, in large samples, has an approximate standard normal distribution under the null hypothesis of no association; equivalently, the squared Wald test statistic, Z^2, has an approximate chi-square distribution with 1 degree of freedom. Alternatively, the *likelihood-ratio test* statistic can be used and this is simply twice the difference in the log-likelihood under the alternative ($\beta_1 \neq 0$) and null ($\beta_1 = 0$) hypotheses. In large samples, the likelihood-ratio chi-square statistic has an approximate chi-square distribution with 1 degree of freedom. For logistic regression, the Wald test is known to exhibit unreliable and aberrant behavior that leads to very low power in certain practical circumstances; as a result, the likelihood-ratio test is preferred. Although most statistical software packages routinely report Wald tests, many provide options for producing likelihood-ratio tests (and likelihood-based confidence intervals).

If the sample size, N, is relatively small, these asymptotic or large sample approximations cannot be relied upon. In particular, a rule of thumb in statistical folklore is that the asymptotic approximations cannot be relied upon if one (or say 25%) of the cells in the 2×2 table of the cross-classification of Y and x has an estimated expected count less than 5. When at least one of the estimated expected counts is less than 5, and it is of interest to make inferences about the association, a common technique is to fix both margins of the 2×2 table and use so-called 'exact' tests and confidence intervals. That is, for a prospective study (where the row margins in Figure 2 are fixed), we further condition on the column margins; for a retrospective study (where the column margins are fixed), we further condition on the row margins; or for a cross-sectional design (where only N is fixed), we condition on both row and column margins. In all of these cases it can be shown that the counts in the resulting table with fixed margins have a noncentral hypergeometric distribution. Under the null hypothesis H_0: $\beta_1 = 0$, the noncentral hypergeometric becomes a central hypergeometric distribution, which forms the basis of Fisher's exact test of no association in a 2×2 contingency table (see, for example, Fisher, 1934). This test is appropriate in small samples; the noncentral hypergeometric can also be used to obtain an estimate of the odds ratio that has better small sample properties than the usual odds ratio estimate given by $e^{\widehat{\beta}_1}$ in eqn [2]. A potential drawback with Fisher's exact test, however, is that it can be 'conservative' in the sense that the true significance level is often far smaller than the nominal level (e.g., 0.05), thereby making it more difficult to reject the null hypothesis.

In general, multiple logistic regression analysis with many predictors relies on large sample theory for the validity of its results. In particular, maximum likelihood estimation of the logistic regression parameters can perform poorly when the sample size is small, the probability of success is near one or zero, or when there is an insufficient number of successes or failures for certain combinations of the predictor variables. *Exact logistic regression* (see, for example, Mehta and Patel, 1995) is an alternative method for fitting logistic regression models that produces valid estimates, test statistics, and confidence intervals when data are sparse. The relationship of exact logistic regression to maximum likelihood logistic regression is similar in certain ways to the relationship of Fisher's exact test to large-sample methods for 2×2 contingency tables. However, whereas Fisher's exact test conditions on the row and column totals in order to derive the distribution of the test statistic, exact logistic regression conditions on the so-called *sufficient statistics* for the remaining parameters in the model when estimating each logistic regression parameter. The sufficient statistics for the parameters are determined by the number of successes for different values of the corresponding predictors. Similar to Fisher's exact test and other exact methods, exact logistic regression guarantees that tests conducted at the α significance level have a type I error rate less than or equal to α, and that 95% confidence intervals have at least 95% coverage even for small sample sizes and sparse data. Parameter estimates and confidence intervals have interpretations identical to those for maximum likelihood logistic regression.

Although the theory underlying exact logistic regression has been well established since the 1970s, the computations required of the method were so demanding that its implementation was considered impractical. More recently, however, with the development of efficient algorithms coupled with increasingly fast computers, exact logistic regression routines are implemented in many statistical software packages.

In summary, exact logistic regression overcomes problems due to sparse data. In principle, exact logistic regression can be applied in settings with multiple predictors; however, greater care is required when attempting to fit complex models. The method has some potential disadvantages. In particular, it may be overly conservative in settings when maximum likelihood logistic regression performs adequately. In addition, it can be

computationally intensive, especially when quantitative predictors or a large number of categorical predictors are included in the logistic regression model. When feasible, exact logistic regression is an attractive alternative to maximum likelihood logistic regression in small sample and sparse data settings.

Logistic Regression for Clustered Data

So far, we have considered logistic regression for a single binary response or in settings where all responses are independent. However, multivariate binary response data commonly arise in a number of applications in the social and behavioral sciences. That is, two or more measurements of the response are often obtained in a block or *cluster* and the binary responses within a cluster are expected to be positively correlated. When this occurs the responses from any pair of members of the same cluster are expected to be more closely related than the responses from a pair belonging to different clusters. Some common examples where data arise in clusters include repeated measures or longitudinal studies and studies on families, communities, or other naturally occurring groups. The important aspect of all of these studies is that the binary responses within a cluster cannot be regarded as independent of one another.

There may be many reasons for the correlation among cluster members. For example, when the cluster is composed of all the siblings within a family the correlation among siblings may be due to shared (or at least similar) genetic, environmental, and social conditions. In a longitudinal study, where the responses within a cluster represent measurements taken at different occasions, the binary responses are expected to be positively correlated simply because they have been obtained on the same individual (or cluster). Whatever the underlying reasons for the correlation, failure to account for it in the analysis can lead to invalid inferences. That is, the standard application of logistic regression in this setting is no longer appropriate.

There are two general approaches for handling the analysis of clustered binary data. The first is to consider models for the joint distribution of the cluster of binary responses that explicitly account for the within-cluster correlation. There is an extensive statistical literature on this topic and the interested reader in referred to a review article by Pendergast et al. (1996). For the most part, these models can be computationally demanding and few have been implemented in commercially available statistical software. An alternative approach is to simply ignore the correlation among members of a cluster. That is, the analysis proceeds naively as though the binary responses within a cluster can be regarded as independent observations, but later a correction is applied to ensure that valid standard errors are obtained. Note that in this 'naive' approach that ignores the within-cluster correlation the estimated logistic regression coefficients are valid, but their nominal standard errors are not. However, valid standard errors can be readily obtained using the well-known empirical variance estimator, first proposed by Huber (1967). Thus, the analysis proceeds in two stages. In the first stage, the correlation among binary responses within a cluster is simply ignored and standard logistic regression is used to obtain estimates of the logistic regression coefficients. In the second stage, valid standard errors for the estimated logistic regression coefficients are obtained using an alternative, but widely implemented, variance estimator that properly accounts for the correlation among the binary responses. The chief advantage of this approach is that it can be readily implemented using standard statistical software for logistic regression (see, for example, Hosmer et al., 2013: Chapter 9).

Conclusion

In this article we have considered regression models for a binary response. In most fields, logistic regression has become the standard model for relating a binary response to a set of covariates. We have described the salient features of logistic regression and considered various aspects of interpretation of logistic regression coefficients. We remarked that the many appealing properties of the odds ratio probably account for the widespread adoption of the logistic function, and hence logistic regression, over alternative functions such as the log, probit, or complementary log–log functions. A more detailed discussion of some of the issues presented here can be found in the textbooks by Agresti (1996), Kleinbaum (1993), and Hosmer et al. (2013).

We conclude this article by noting that logistic regression happens to be a special case of a broad and unified class of statistical models known as *generalized linear models*. Other familiar examples of generalized linear models include linear regression and analysis of variance (ANOVA) models for a continuous response, log-linear models for count data, and probit regression for a binary response. It is beyond the scope of this article to give a detailed description of generalized linear models; the interested reader is referred to the textbooks by Dobson (1990) and McCullagh and Nelder (1989). Instead, we will simply outline the main features of generalized linear models and indicate how logistic regression is related to alternative models for binary responses.

Generalized linear models are characterized by a three-part specification: (1) the *random component*, where the distribution of the response variable is assumed to be from the exponential family (the exponential family includes, among others, the normal, Bernoulli and Poisson distributions); (2) the *systematic component*, where a set of predictor variables, say $x_1, \ldots, x_p$, produce a *linear predictor*, $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$; and (3) the *link function* relating the random and systematic components. The link function in (3) describes how a particular transformation of the mean of the response variable is related *linearly* to the predictor variables. In summary, a generalized linear model describes how the transformed mean of a response variable from the exponential family of distributions is linearly related to a set of predictors. Thus, within the framework of generalized linear models, logistic and probit (or complementary log–log) regression share the same specification of the random and systematic components. What distinguishes logistic regression from probit regression is solely the choice of link function in (3). In logistic regression a logit function is chosen to link the random and systematic components, i.e., a logit function relates the probability of success, π, to the linear predictor. In probit regression the inverse normal function is chosen to relate π to the linear predictor.

Finally, we remark that the logit function also happens to be the so-called *natural* or *canonical* link function, i.e., the logit is a special link function that can be derived from the Bernoulli distribution. Although the choice of canonical link function does result in some minor mathematical simplifications, the selection of a suitable link function should primarily be made on scientific grounds. That is, the choice of link function should be based on the goodness of fit of the resulting model and the ease of interpretation of the model parameters. In the case of a binary response, we have already seen that for most practical purposes it is not possible to discriminate empirically, in terms of goodness of fit, between probit and logistic regression. As a result, logistic regression is often the method of choice for analyzing a binary response due to the more appealing interpretation and properties of the logistic regression coefficients.

See also: Multivariate Analysis: Discrete Variables (Loglinear Models).

Bibliography

Agresti, A., 1996. An Introduction to Categorical Data Analysis. Wiley, New York.

Clogg, C.C., 1982. Some models for the analysis of association in multiway cross-classifications having ordered categories. Journal of the American Statistical Association 77, 803–815.

Davis, J.A., 1977. Codebook for the 1977 General Social Survey. National Opinion Research Center, Chicago.

Dobson, A.J., 1990. An Introduction to Generalized Linear Models. Chapman and Hall, New York.

Farewell, V.T., 1979. Some results on the estimation of logistic models based on retrospective data. Biometrika 66, 27–32.

Fisher, R.A., 1934. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied Logistic Regression, third ed. Wiley, New York.

Huber, P.J., 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, pp. 221–233.

Kleinbaum, D.G., 1993. Logistic Regression: A Self-Learning Text. Springer-Verlag, New York.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman and Hall, New York.

Mehta, C.R., Patel, N.R., 1995. Exact logistic regression: theory and examples. Statistics in Medicine 14, 2143–2160.

Pendergast, J.F., Gange, S.J., Newton, M.A., Lindstrom, M.J., Palta, M., Fisher, M.R., 1996. A survey of methods for analyzing clustered binary response data. International Statistical Review 64, 89–118.

Prentice, R.L., Pyke, R., 1979. Logistic disease incidence models and case-control studies. Biometrika 66, 403–412.

Binet, Alfred (1857–1911)

Robert J Sternberg, Cornell University, Ithaca, NY, USA
Linda Jarvin, Paris College of Art, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Alfred Binet's impact on the field of psychology is unequaled, with the possible exception of Sigmund Freud. In retrospect, it appears that quite ironically Alfred Binet's most outstanding contribution is the Stanford–Binet Intelligence Scale, which owes its popularity (though not its conceptualization) largely to the work of Lewis Terman rather than to Binet himself. The intelligence scale became the tree that hid the forest: Much of Binet's exceptional, and often revolutionary, research was largely ignored on account of the overwhelming visibility that his work on intelligence testing achieved. A thorough review of his bibliography shows that his research interests spread a wide range of topics, including but not limited to anatomy, animal magnetism, audition, chess players, court witnesses, double consciousness, fetishism in love, graphology, intellectual exhaustion, intelligence, intelligence assessment, hallucinations, hysteria, inhibition, language, literary creation, memory, mental alienation, mental images, moral responsibility, movement, pedagogy, perception, phrenology, physiology, reasoning, retardation and academic underachievement, the soul, and suggestibility. Though at first glance these topics may seem unrelated, there really are two consistent threads throughout Binet's life and work. The first is a commitment to experimental psychology. The second is his pursuit to understand the complexities of the human mind in order to establish 'an individual psychology.'

A Brief Biography

Alfred Binet was born in the seaport town of Nice on what was to become France's Mediterranean coast on 11 July 1857. According to later reports by his daughter Madeleine (Avanzini, 1974), Alfred was a bright child who succeeded so well in school that his mother decided to send him away to the capital Paris when he was barely 12 years old, in order that he might study in one of the best schools in France. Upon graduation from high school he completed a law degree (1877), but then decided to pursue studies in medicine and biology rather than practice law. Under the supervision of the entomologist and embryologist Balbiani, who was later to become his father in law, he started writing a dissertation on "The sub-intestinal nervous system of insects." At the same time he wrote plays for the theater, already displaying his characteristic eclectic range of interests. He finally found a discipline that allowed him to combine his different areas of interests: Psychology. To simplify, one can distinguish three phases in his career as a psychologist, although his publication record shows that the stages were overlapping rather than sequential (Delay, 1958): Psychopathology, experimental psychology, and child psychology.

In the second half of the nineteenth century, French psychology was mainly focused on psychopathology, and Binet's early interests were no exception. His early influences were the memory researcher Théodule-Armand Ribot and neurologist Jean-Martin Charcot, and Binet's first major writings explored the domains of sensations, hysteria, hypnotism, and personality disorder. Binet worked closely with Ribot between 1880 and 1907, and published over 40 articles in Ribot's *Revue Philosophique*. During this period, he also began studying hypnotism under Charcot at the Salpetrière hospital in Paris.

In 1891, Binet received tenure at the physiological psychology laboratory at the Sorbonne University in Paris, and his research focus shifted from clinical to experimental psychology. The major contemporary influences in psychophysiology and experimental psychology were German, namely Wilhelm Wundt whose laboratory in Leipzig served as the model for the one Binet founded in Paris. Binet completed his thesis in biology in 1895, the same year in which he took the head of the physiological psychology laboratory where he had tenure, and published a seminal *Introduction to Experimental Psychology*. 1895 was also the year in which he founded the journal *L'Année Psychologique*, still published and considered a reference today. In 1899, Binet took on Theodore Simon, who was to become his closest and most constant collaborator, as a doctoral student and that same year he joined the *Free Society for Psychological Child Studies*, of which he became the president in 1902. Like many researchers of his time, Binet decided to follow in the footsteps of neurologist Paul Broca, and explored intelligence through phrenology, exploring the direct relation between the size of the cranium and a person's level of intelligence. Binet's empirical investigations consisted in measuring the crania of school children deemed bright by their teachers and to compare these with the measurements obtained from academically low-achieving students. In 1901, Binet conceded that "the idea of measuring intelligence by measuring the head … seemed ridiculous" (Binet, 1901: p. 403). Later, however, he did discover that there were noticeable differences in the cranial size and shape of children from contrasted groups, with very bright students having larger skulls than very low-achieving students. Binet himself reached the conclusion that the initial measures showing cranium size differences between low- and high-achieving students were probably due to his own suggestibility. With a rare sense of honesty, he admitted that he had been looking for differences and had found them where he had expected them, and with his faith in phrenology shaken, he abandoned it. He similarly discarded hypnosis as a method of study, and moved instead to explore the human psyche through tests and puzzles that he initially

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61008-5

created for, and validated on, his two daughters (Binet, 1890). As was the case with Jean Piaget many years later, most of his hypotheses regarding cognitive development stemmed from the observation of his own children.

After a phase devoted the psychopathology, followed by one centered on experimental psychology, the third phase of Binet's life focused primarily on child psychology. The start of this period can be traced to 1905, with the publication of several articles by Binet and Simon devoted to the diagnosis of abnormality (Binet and Simon, 1905a,b,c) and the establishment of a laboratory in experimental psychology. In the laboratory Binet worked closely with classroom teachers, and it is here that he made his major contributions to the study of academic underachievement, but also to psychometrics, a term coined by Sir Francis Galton in 1879. Another of Binet's major contributions to child psychology was promoting the idea that children are not, as previously thought, miniature adults, but follow different principles of cognitive functioning (Zazzo, 1958). The rest of this article will focus on Alfred Binet's contributions to our understanding of abnormal children and the development of the assessment tools to which this interest led.

Alfred Binet passed away unexpectedly on 28 October 1911, at age 54, and was laid to rest in the Montparnasse cemetery in Paris, France.

The Study of Abnormal Children

Starting with the beginning of the nineteenth century, Binet increasingly focused his research on the study of 'abnormal' children (Binet and Simon, 1907). The *Free Society for Psychological Child Studies*, presided by Binet, lobbied the French government to start assessing school children, with the aim of early identification of those children that would not benefit from regular schooling in order to provide them with special education. Though this is an established practice in many countries today, at the time it was quite novel and controversial. In 1904, the French minister of education officially appointed Binet to devise a means of identifying 'retarded' school children (see also Sternberg and Jarvin, 2003) and out of these efforts came the publication in 1905 of Binet and Simon's Metric Intelligence Scale (Avanzini, 1974). The scale identifies students with learning impairments who were nevertheless thought able to profit from education and to allow them into special education classes. The first 'perfectioning class' was opened in 1907, and two more such specialized classes followed that same year. Though Binet admitted starting out with the belief that all children could progress and 'abnormality' to a certain extent be overcome, he never lost sight of the importance of empirically testing and validating his ideas. In that spirit, he asked that the usefulness of special education classes be empirically established before any legislative measures to generalize their use be taken.

Through his experience with developing assessments for selecting students into special education classes and working with teachers on adapted pedagogical methods, Binet strengthened his belief in psychometric assessment and further refined his theoretical ideas about 'abnormality.'

Binet and Simon's (1907) publication *Abnormal Children* proposed a new view of mental retardation in which an abnormal child is not a normal child who has either halted or slowed down in its development, but rather a child who follows a different developmental pattern. This pattern is unbalanced and uneven, with some aspects resembling the developmental patterns of 'normal' children and others being quite different. It is precisely this lack of balance between the different developmental aspects that constitutes the abnormality (Binet and Simon, 1907, 1922). In the 'abnormal' child, there might be an important delay in the acquisition of certain aptitudes, while other aspects, be they cognitive or somatic, follow a normal pattern of development.

Binet and Simon were also precursors in arguing that the age at which one acquires a skill has specific impacts and implications. For example, learning to read at age six is different from learning to read at age 12. The abnormal child who learns to read at age 12 has a more extensive vocabulary and more extrascholastic interests than the regular 6-year old when he or she learns to read. Intellectual impairment can be characterized by three major components: a global developmental delay, the imbalance of this delay depending on what cognitive functions are considered, and the resulting lack of coordination in cognitive functioning. Because abnormality is a distinctive developmental pattern rather than mere retardation in a normal developmental pattern, special education is possible and needed.

The Metric Intelligence Scale

Binet and Simon's Intelligence Scale published in 1905 contained 30 tests, presented in order of increasing difficulty, and designed to measure everything from basic skills such as coordination of movement to more complex processes such as comprehension, judgment, and abstract reasoning. Some of the tests were specifically created for the Scale and some were adaptations of existing cognitive assessments, like those created by French physicists Blin and Damaye. The entire battery of tests was piloted on samples of 'normal' and 'retarded' children from ages 2 to 12.

Performance on the easiest tests distinguished severely and profoundly retarded children (then labeled 'idiots') from the rest; tests of intermediate difficulty distinguished between the severely and the moderately retarded (then labeled 'imbeciles'); and the most complex tests lead to different performances for mildly retarded (then labeled 'morons') and normally functioning children of the same age. The tests also distinguished between younger and older children.

As a member of the ministerial committee on education, Binet's main purpose was to identify those mildly retarded children thought most likely to benefit from special education in order to enhance their cognitive functioning. The 1905 version of the Intelligence Scale stressed that the results of the tests only be used as an indication of a child's cognitive level at the time of administration, and that this level was likely to change with appropriate and adapted education. A revised version of the Scale, published in 1908, offered the possibility to establish a child's mental age (a notion first introduced by Chaillé in 1887) based on his or her performance. If 65–75%

of children in a given age group respond correctly to a given test, that test is considered to be appropriate to measure mental performance for that age group. A child can thus succeed on tests that are classified as being lower, equal to, or higher than that child's biological age. It should be noted that the mental age thus derived simply corresponds to the performance norm of a given age group, and does not indicate the developmental stage reached by the child, contrary to what the American adapters of the Scale later thought. In other words, having a given mental age does not imply that you function like a child of that biological age, but only that your individual performance on a given test corresponds to what the majority of children at a given age will achieve.

Binet explicitly warned against the use of performance on an IQ test as a fixed measure of intelligence: by essence, the test is constructed in such a manner that 25–35% of children of a given age group will fail on a test intended for their age (Lippman, 1922).

Binet also made it clear that the results on the Scale were just one among several indicators of the child's level of functioning, and that the careful observation of the global person and the test taker's reactions to the assessment situation were as important to take into consideration as the numerical tests results per se. As a researcher, Binet was less interested in the end results than in the paths by which they were reached by different children.

To paraphrase an article by Sarason (1976), the fate of Alfred Binet and school psychology was unfortunate, and the misinterpretations and misuses of the Metric Intelligence Scale were many. Binet's aim was to develop and validate an instrument that would help identify those children, who would benefit from special education and allow them to make the most of their potential in adapted pedagogical settings. His aim was to prevent further rejection of these children by providing them the special education that would allow them to thrive and eventually to return to the regular classroom. This second aspect, however, is all too often neglected, and the Scale used to weed out students rather than a means to strive for greater inclusion.

It is important to remember that Binet did not view intelligence as a fixed quantity set at birth, but rather as incremental. It is precisely because of the elastic nature of intelligence and the possibility of developing it through education that we must identify children with special needs to provide them the support that will allow them to develop their intellectual abilities and to blossom. The researchers who imported the Intelligence Scale to the United States and Great Britain, however, had different views and different goals in mind. Lewis Terman, who in 1916 introduced the Scale as the Stanford–Binet assessment instrument in the United States, stressed the importance of hereditary factors in explaining performance on intelligence tests. If intelligence is largely hereditary, specialized education becomes less important, and in that his use of the Scale goes against that originally intended by Binet. Terman considered the genetic influence to be too strong to be altered through education. For Terman, children with low scores on the tests should be put in separate classes, not because they would there receive the specialized and adapted education that would allow them to thrive and then return to the regular classroom, but because there was no possibility of them ever being able to integrate into a regular classroom and benefit from regular instruction. Low-achieving students should thus be set aside and the State would have to assume guardianship of those with intelligence levels in the mentally retarded range (Terman, 1916).

Strengths and Weaknesses of Binet's Contributions to Social and Behavioral Sciences

Binet's strengths and weaknesses are very much the two sides of the same coin: His strength is the applied nature of his research and the lasting social cultural impact of his work, his weakness the lack of a global, coherent theoretical framework. Key to his research is the pragmatic approach and attempt to directly connect empirical research results to social and educational concerns. His research on abnormal children and cognitive assessment started not as a theory, but in response to a concrete problem, that of educating more children than what were at the time benefiting from regular instruction.

That is not to say that his work had no theoretical basis, and he is perhaps best remembered for two major theoretical contributions. First, he broke with the Galtonian tradition of studying intelligence through psychophysical processing. He suggested instead that intelligence be studied through the expression of higher order processing in complex everyday tasks. Though this viewpoint may seem trite today it was novel at the time, and almost provocative to many of his contemporaries. Before Binet, most psychologists followed in Wundt's footsteps and thought that intelligence was best explored by measuring lower order, very simple, processes. Binet not only decided to assess higher order cognitive functions, he decided to do so with the same experimental rigor that had previously been reserved for lower order processes. Second, he introduced a new methodology for studying intelligence.

Though it is difficult to find a global theory of intelligence in Binet's work (Reuchlin, 1977), it is clear that he did think in theoretical terms. In his publications, intelligence is variously assimilated to common sense, faculty to adapt, initiative, judgment, and practical sense. At other times, intelligence is decomposed into four 'functions': comprehension (understanding), direction, invention, and self-censorship. Neither Binet's head nor his heart was in theorizing about intelligence, however, and he went so far as to state that he was only interested in understanding the specific facts that he gathered and not in any general theory of intelligence.

Psychological Assessment Today

The importance of psychometrically sound assessment scales to measure levels of cognitive functioning was soon generalized beyond the population of abnormal school-aged children and constitutes one of Binet's greatest legacies to the field of cognitive evaluation. The strong positive correlation between children's scores on the *Binet–Simon Metric Intelligence Scale* and their school performance, largely due to the very scholastic nature of any of the assessment battery's tests, was

seen as a proof that the scale was indeed measuring intelligence, and contributed to its widespread success. The Stanford–Binet is still used today, although its prominence has been eclipsed by the Wechsler scales.

The debates triggered by Binet's work are alive and well; however, intelligence and intelligence testing remaining hotly debated topics, as demonstrated by the steady crop of new books to appear on the subject every year.

See also: Cattell, Raymond Bernard (1905–98); Galton, Sir Francis (1822–1911); Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Avanzini, G., 1974. A. Binet: Ecrits psychologiques et pédagogiques (Writings on Psychology and Pedagogy). Privat, Toulouse, France.

Binet, A., 1890. Recherches sur les mouvements de quelques jeunes enfants (Research on the movements of some young children). Revue Philosophique 29, 297–309.

Binet, A., 1901. Recherches sur la technique de la mensuration de la tête vivante (Research on the technique of measuring a living head). L'Année Psychologique 7, 314–368.

Binet, A., Simon, T., 1905a. Sur la nécessité d'établir un diagnostic scientifique des états inférieurs de l'intelligence (On the importance of a scientific measure of lower levels of intelligence). L'Année Psychologique 11, 163–190.

Binet, A., Simon, T., 1905b. Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux (New methods for diagnosing intellectual performance in abnormal persons). L'Année Psychologique 11, 191–244.

Binet, A., Simon, T., 1905c. Application des méthodes nouvelles au diagnostic du niveau intellectuel des anormaux d'hospice et d'école primaire (Application of new methods for diagnosing intellectual performance in hospitalized abnormal persons and primary school). L'Année Psychologique 11, 245–336.

Binet, A., Simon, T., 1907. Les enfants anormaux (Abnormal children). Privat, Paris.

Binet, A., Simon, T., 1922. La mesure de développement de l'intelligence chez les jeunes enfants. Société Pour L'etude Psychologique de l'Enfant, Paris.

Delay, J., 1958. La vie et l'oeuvre d'Alfred Binet (The life and work of Alfred Binet). Psychologie Française 3, 85–95.

Lippman, W., 25 October 1922. The mental age of Americans. The new republic. In: Jacoby, R., Glauberman, N. (Eds.), 1995. The Bell Curve Debate: History, Documents, Opinions. Random House, New York, pp. 561–565.

Reuchlin, M., 1977. Psychologie. Presses Universitaires de France, Paris.

Sarason, S.B., 1976. The unfortunate fate of Alfred Binet and school psychology. Teachers College Record 77, 579–592.

Sternberg, R.J., Jarvin, L., 2003. Alfred Binet's contributions as a paradigm for impact in psychology. In: Sternberg, R.J. (Ed.), Anatomy of Impact: What has Made the Great Works of Psychology Great? American Psychological Association, Washington, DC, pp. 89–108.

Terman, L.M., 1916. The Measurement of Intelligence. Houghton Mifflin, Boston, MA.

Zazzo, R., 1958. Alfred Binet et la psychologie de l'enfant (Alfred Binet and child psychology). Psychologie Française 3, 113–121.

Bioarchaeology

Clark S Larsen, The Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Bioarchaeology is the study of human remains from archaeological contexts. The field focuses on a range of circumstances that shape who we are today, including health and well-being, diet and nutrition, environment and human impacts, and the influence of social, economic, and political forces on the lives and livelihoods of past societies in recent human evolution.

Bioarchaeology is the study of human remains from archaeological settings, viewing these remains in relation to biocultural adaptation, life history, lifeway reconstruction, population and demographic history, and the influence of social, economic, and political forces on the lives and livelihoods of earlier societies (Agarwal and Glencross, 2011; Buikstra and Beck, 2006; Larsen, 1997). Human remains can include any body tissue, ranging from artificially or naturally preserved bodies – such as mummies from Egypt or bog bodies from Denmark – to fragments of bones and teeth. Because skeletons are preserved far more often than other types of remains, they comprise most of the subject matter of bioarchaeology.

Origins and Paradigm Shifts: From Racial Classification to Biological Process

Bioarchaeology originated in the field of osteology, the anatomical study of bones and teeth. Much of the history of study of ancient skeletons focused on classification and racial typology. Biological differences between populations from different time periods – such as differences in skull form – were interpreted to represent replacement of one 'race' of people by another. Various lines of evidence reveal the adaptability of bone tissue, especially with regard to influence of physical activity on shape and structure. For example, skulls representing earlier populations who lived exclusively on wild plants and animals tend to be larger and more robust than skulls representing later farmers (e.g., Carlson and Van Gerven, 1977). In consideration of the extrinsic factors that influence skull form (e.g., shift from chewing hard- to soft-textured foods), the temporal variation for many settings is best interpreted in light of underlying biological processes that influence musculoskeletal development and not racial history.

The reinterpretation of temporal changes in skull form is emblematic of the paradigmatic shift that took place in bioarchaeology in the last 40 years (Larsen and Walker, 2010). Namely, ancient bones and teeth are no longer viewed as objects for classification only. Rather, bioarchaeologists seek to understand ancient skeletons as though the populations they represent are alive today, subject to the myriad of environmental, cultural, and social influences that affect human biology.

Skeletal and dental tissues are remarkably sensitive to a variety of environmental conditions affecting their form and composition. Owing to this sensitivity during life, human remains recovered from archaeological settings are informative about quality of life (health) and lifestyle (activity and work) in particular, and behavior in general. Skeletons also contain a fund of information about biological relatedness and demographic history.

Quality of Life and Health

Quality of life refers to a variety of factors that influence physical well-being – health is the most important. Health is assessed from human remains via chemical, stress, pathological, and morphological indicators.

Reconstructing Diet and Inferring Nutrition: Stable Isotopes and Bone Chemistry

The food humans eat and the nutrition that it provides is fundamental to health. Until the 1980s, most knowledge about ancient diets was based on plant and animal remains from archaeological contexts. In the late 1970s, Vogel and Van der Merwe (1977) identified the timing and intensity of the shift from diets based on wild plants to domesticated plants (maize) by analyzing stable isotope ratios of carbon ($^{12}C/^{13}C$) extracted from archaeological human bone samples. This pathbreaking study provided a new basis for reconstructing past diet and set the stage for analysis of stable isotopes of other elements (e.g., nitrogen, strontium, and oxygen) and drawing inferences about variation in diet, food use, weaning practices, climate, mobility, and social structure of past societies (Dupras and Tocheri, 2007; Katzenberg, 2008).

Stress and Deprivation

The cells responsible for tooth and bone development are easily disrupted if physiological stress caused by malnutrition or disease occurs while these tissues are forming. In teeth, macroscopically visible areas of enamel deficiency (hypoplasia) or microscopically visible defects (accentuated Retzius lines) indicate growth disruption. In addition to having elevated frequencies of indicators of growth arrest, populations that experience chronic illness or malnutrition tend to be short in stature and have less bone mass in comparison with healthy, well-nourished populations.

Many human populations experience nutritional deficiencies, infection and infectious disease, or poor living

circumstances generally (e.g., Walker et al., 2009). Anemia often develops, resulting in increased production of a person's red blood cells. The skull bones of anemic persons are often thickened and display porosity (porotic hyperostosis). The incidence of growth arrest indicators, anemia, and other related morbid conditions increased in the last 10 000 years as human populations shifted to agriculture and became sedentary.

Illness and Infection

Throughout the entire 5 million or so years of human evolution, human beings have been exposed to many different microbes, many of which cause disease. Skeletons record the effects of a number of chronic diseases, such as the treponematoses (the disease group that includes the nonvenereal and venereal forms of syphilis), leprosy, and tuberculosis. Identification of deoxyribonucleic acid (DNA) in archaeological skeletons has verified that the lesions found in ancient remains result from the disease-causing microbes (e.g., *Mycobacterium tuberculosis*; Klaus et al., 2010; Salo et al., 1994).

Leg (tibia) and, to a lesser extent, other bones in archaeological skeletal samples sometimes display periosteal reactions, areas of pathological bone apposition. These lesions are nonspecific because they can be caused by different infectious agents or by trauma to the lower leg. For many settings, these lesions are likely due to population crowding and poor hygiene, conditions conducive to the maintenance and spread of infectious disease. Sedentary, crowded populations in prehistory have generally higher frequencies of nonspecific and specific infectious disease than mobile, dispersed populations.

Violence and Injury

All human populations experience physical confrontation of one type or another at some point in time. In the Santa Barbara Channel Island region of the southern California Pacific coast, numerous skulls display indentations (depressed fractures) caused by being struck on the head with a club (Lambert and Walker, 1991). In late prehistory in this setting, the frequency of cranial fractures increased, and lethal projectile wounds became more common than before. This temporal shift coincides with deterioration in climate and reduced availability of food resources, suggesting that as environmental stress increased, people became more competitive and more violent. The skeletal evidence derived from the victims of violence may be more informative about the causes and consequences of confrontation than other archaeological evidence (e.g., fortifications, iconography, and weaponry) (see Tung, 2012).

Lifestyle and Activity

Physical activity is a defining characteristic of different human adaptive regimes. Hunter-gatherers are often characterized as highly mobile, living demanding lifestyles that require a great deal of work in order to acquire food. Agriculturalists, on the other hand, are often thought of as being more 'advanced,' leading lives of leisure. Study of ancient skeletons allows us to address these assumptions and to test hypotheses about labor, workload, and physical activity in past societies.

Activity and Degenerative Pathology

The articular surfaces of the joints of the skeleton are adapted to withstand mechanical stress, such as those that derive from lifting or walking, or a combination of activities. Over the course of a person's lifetime, the cartilage and bone in the articular joints (e.g., knees) begins to erode in response to mechanical loading. In addition, spicules of bone form along the joint margins, which may impede movement. These bone changes are part of a disorder called osteoarthritis. In archaeological settings, there is a tendency for skeletons of prehistoric hunter-gatherers to have more osteoarthritis than those of prehistoric farmers, and males generally have more osteoarthritis than females (see Bridges, 1992; Larsen, 1995). However, the pattern and frequency of osteoarthritis is highly variable from one region to another, suggesting that the disorder is influenced by local circumstances involving a complex interplay between culture and lifestyle.

Activity and Bone Structure

Bone tissue remodels itself in response to mechanical stimulation (Wolff, 1892/1986). In areas of the skeleton (or a particular bone) that are subjected to high levels of physical demand, more bone tissue is added in order to add strength and to resist forces that would cause fracture. Breadths and circumferences (e.g., for leg and arm bones) can be used to characterize the size of bones, and to infer levels and types of physical activity. Engineers have developed a more precise way of inferring levels or types of physical activity. Civil and mechanical engineers measure the strength of beams that are used in the construction of buildings. The distribution of material in cross sections of these beams is analyzed in relation to how they are able to resist bending, twisting (torsion), and other loading modes. Similarly, the human skeleton provides the body with a superstructure for the support of body weight and the action of muscles. From cross-sectional measurements of bones (derived from computed tomographic scans or by direct cutting to reveal the cross sections), bioarchaeologists apply these engineering principles and methods to determine the strength of bones (Ruff, 2008). Biomechanical analysis of skeletons from around the world shows a general reduction in bone strength over the course of human evolution, especially in the last 10 000 years, suggesting that physical activity has declined.

Activity above the Neck

The size and morphology of the face and jaws is influenced by mechanical loading and activities involving mastication of food, and in the use of teeth in nondietary roles (e.g., use of front teeth for preparing animal hides for clothing). Human faces and jaws have become smaller in recent human evolution, reflecting the increased tendency for eating softer, more processed foods. Consumption of softer foods and more carbohydrates has also contributed to increase in dental caries (cavities) in recent humans. Patterns of tooth wear are

informative about how people use their teeth. Scanning electron microscopy of the chewing surfaces of teeth reveals microwear patterns relating to diet and tooth use (Teaford, 1991).

Estimating Biological Relatedness: Bone Shapes, Tooth Traits, and Ancient DNA

The shapes and structures of bones and teeth contain key information about population history, since these shapes and structures are at least partly genetically determined. There are some 30 different dental (discrete morphological) traits that are especially useful for identifying inter- and intrapopulation biological relatedness (Scott and Turner, 1997).

Because bone shape is influenced by environmental factors (see above), and many dental traits are obliterated by wear on chewing surfaces of teeth, identification of biological relatedness is sometimes problematic. A potentially more informative approach to reconstructing biological history came about in the mid-1980s with the discovery that DNA is often preserved in ancient bones. The development of polymerase chain reaction made it possible to amplify short segments of DNA into millions of copies that could be used to identify key parts of the genome (O'Rourke et al., 1996; Stone, 2008). The analysis of DNA has allowed the testing of hypotheses about biological history, such as long-standing issues relating to population origins and continental settlement (e.g., the origin of native Americans and migration to the Americas), in ways not possible by study of morphology and traits in bones and teeth alone.

Population Profiles and Demographic History

Physical anthropologists have developed various methods for identifying sex and estimating age at death from skeletons (Ubelaker, 1989). Identification of sex (a biological attribute) allows insight into gender (a social and cultural attribute) and its relationship to health, activity, and social and cultural behavior (Grauer and Stuart-Macadam, 1998). Mortality and fertility history of a past population can be reconstructed from age composition of collections of skeletons from archaeological cemeteries. In a stable population with minimal in-migration or out-migration, age composition is especially informative about birth rates of past populations. A skeletal series having many infants and young children and relatively fewer old adults suggests relatively high birth rate. Conversely, a series containing few children and many older adults suggests low birth rate (e.g., Buikstra et al., 1986). In some regions of the world, the increase in population size that accompanied the shift from foraging to farming appears to be related to decreased birth spacing and increased fertility.

Challenges

Research results derived from the study of ancient skeletons – once mostly relegated to the appendices of obscure, unpublished archaeological reports – are now viewed as integral to archaeology. Reflecting this increased role is the growing visibility of bioarchaeology in the popular press and in the scientific literature. Past human biology provides an essential context for understanding present-day diseases, health, and lifestyle. As the discussion of cultural 'repatriation' continues to build in the United States, Canada, Australia, Israel, and elsewhere, the archaeological community is increasingly called upon to explain to the larger public why ancient remains are important, why they should be studied, and why they are relevant to understanding biology, health, and lifestyle in living people today. Public education about bioarchaeology is important and should be pursued if an informed understanding of the past using the human biological component of the archaeological record is to progress.

See also: Plagues and Diseases in History.

Bibliography

Agarwal, S.C., Glencross, B., 2011. Social Bioarchaeology. Wiley-Blackwell, Hoboken.

Bridges, P.S., 1992. Prehistoric arthritis in the Americas. Annual Review of Anthropology 21, 67–91.

Buikstra, J.E., Beck, L.A., 2006. Bioarchaeology: The Contextual Analysis of Human Remains. Academic Press, San Diego.

Buikstra, J.E., Konigsberg, L.W., Bullington, J., 1986. Fertility and the development of agriculture in the prehistoric Midwest. American Antiquity 51, 528–546.

Carlson, D.S., Van Gerven, D.P., 1977. Masticatory function and post-Pleistocene evolution in Nubia. American Journal of Physical Anthropology 46, 495–506.

Dupras, T.L., Tocheri, M.W., 2007. Reconstructing infant weaning histories of Roman period Kellis, Egypt using stable isotope analysis of dentition. American Journal of Physical Anthropology 134, 63–74.

Grauer, A.L., Stuart-Macadam, P., 1998. Sex and Gender in Paleopathological Perspective. Cambridge University Press, Cambridge, UK.

Katzenberg, M.A., 2008. Stable isotope analysis: a tool for studying past diet, demography, and life history. In: Katzenberg, M.A., Saunders, S.R. (Eds.), Biological Anthropology of the Human Skeleton. Wiley-Liss, Hoboken, pp. 413–442.

Klaus, H.D., Wilbur, A.K., Temple, D.H., Buikstra, J.E., Stone, A.C., Fernandez, M., Wester, C., Tam, M.E., 2010. Tuberculosis on the north coast of Peru: skeletal and molecular paleopathology of late pre-Hispanic and postcontact mycobacterial disease. Journal of Archaeological Science 37, 2587–2597.

Lambert, P.M., Walker, P.L., 1991. Physical anthropological evidence for the evolution of social complexity in coastal Southern California. Antiquity 65, 963–973.

Larsen, C.S., 1995. Biological changes in human populations with agriculture. Annual Review of Anthropology 24, 185–213.

Larsen, C.S., 1997. Bioarchaeology: Interpreting Behavior from the Human Skeleton. Cambridge University Press, Cambridge, UK.

Larsen, C.S., Walker, P.L., 2010. Bioarchaeology, health, lifestyle, and society in recent human evolution. In: Larsen, C.S. (Ed.), A Companion to Biological Anthropology. Wiley-Blackwell, Malden, pp. 379–394.

O'Rourke, D.H., Carlyle, S.W., Parr, R.L., 1996. Ancient DNA: a review of methods, progress, and perspectives. American Journal of Human Biology 8, 557–571.

Ruff, C.B., 2008. Biomechanical analysis of archaeological human skeletons. In: Katzenberg, M.A., Saunders, S.R. (Eds.), Biological Anthropology of the Human Skeleton. Wiley-Liss, New York, pp. 183–206.

Salo, W.L., Aufderheide, A.C., Buikstra, J., Holcomb, T.A., 1994. Identification of *Mycobacterium tuberculosis* DNA in a pre-Columbian Peruvian mummy. Proceedings of the National Academy of Sciences of the United States of America 91, 2091–2094.

Scott, G.R., Turner II, C.G., 1997. The Anthropology of Modern Human Teeth: Dental Morphology and Its Variation in Recent Human Populations. Cambridge University Press, Cambridge, UK.

Stone, A.C., 2008. DNA analysis of archaeological remains. In: Katzenberg, M.A., Saunders, S.R. (Eds.), Biological Anthropology of the Human Skeleton. Wiley-Liss, New York, pp. 461–483.

Teaford, M.F., 1991. Dental microwear: what can it tell us about diet and dental function? In: Kelley, M.A., Larsen, C.S. (Eds.), Advances in Dental Anthropology. Wiley-Liss, New York, pp. 341–356.

Tung, T.A., 2012. Violence, Ritual, and the Wari Empire: A Social Bioarchaeology of Imperialism in the Ancient Andes. University Press of Florida, Gainesville.

Ubelaker, D.H., 1989. Human Skeletal Remains: Excavation, Analysis, Interpretation, second ed. Taraxacum, Washington, DC.

Vogel, J.C., Van der Merwe, N.J., 1977. Isotopic evidence for early maize cultivation in New York state. American Antiquity 42, 238–242.

Walker, P.L., Bathurst, R.R., Richman, R., Gjerdrum, T., Andrushko, V.A., 2009. The causes of porotic hyperostosis and cribra orbitalia: a reappraisal of the iron-deficiency-anemia hypothesis. American Journal of Physical Anthropology 139, 109–125.

Wolff, J., 1892/1986. The Law of Bone Remodelling (P. Maquet, R. Furlong, Trans.). Springer-Verlag, Berlin.

Biobanking: Ethical Issues

Klaus Hoeyer, University of Copenhagen, Copenhagen, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article outlines the main ethical issues debated in relation to population-based genetic research biobanks. There are many types of biobanking, but few have attracted as much ethical attention as population-based ones. Some possible reasons for this are discussed along with considerations of the political landscape of biobank ethics before the themes dominating the literature are presented: consent to participation; good governance; fairness and reciprocity; commercialization; trust and trustworthiness; and unintended effects such as breaches of confidentiality and privacy, surveillance, and social changes.

Introduction

Many forms of medical research rely on access to various forms of human biological material. For more than a century, it has been common in middle- and high-income countries to gather such material in structured repositories along with information about the person from whom the material was taken. Today such repositories are typically referred to as biobanks. Every day, millions of samples are stored throughout the world for diagnostic, therapeutic, forensic, or other research purposes (*see* Forensic Genetic Databases: Ethical and Social Dimensions). When scholars refer to the *ethical issues* surrounding biobanks, however, they typically refer to a relatively recent and quite specific form of biobanking: population-based genetic research biobanks.

One event, in particular, framed biobank ethics. In 1996 Kári Stefánsson, an Icelandic professor at Harvard University, founded the company deCODE genetics, Inc., and returned from the United States to Iceland proposing the construction of three interrelated databases: a genealogical database, a health sector database of electronic medical records, and a genetic biobank. The genetic biobank was planned to rely on informed consent (i.e., provision of information about the biobank followed by a voluntary signature of consent to participation), but for the electronic patient record to be of benefit to the health services, it would need to include all residents and inclusion could therefore not be optional. Though deCODE would not conduct any genetic research without informed consent, the presumed consent principle for the health sector database was widely misunderstood by international observers as covering all deCODE's research, including genetic research. International ethical commentary on population-based genetic research biobanks became relatively focused on the consent issue, and this interest has in turn affected the reception of subsequent population-based biobanks in, for example, Canada (CARTaGENE), Estonia (Estonian Genome Project), the United Kingdom (UK Biobank), Japan (BioBank Japan Project), and Sweden (LifeGene).

Other major research biobanks existed prior to this wave of activity, in particular in Scandinavia, where central population registries facilitate easy follow-up in cohort studies; but these biobanks were intended primarily for other types of research and only gradually did they become resources for genetic research. Large-scale biobanks are also maintained by military authorities (Eiserman and Castillo, 2000) and the pharmaceutical industry (Lewis, 2004), even though little is known about them and they are rarely granted much attention in relation to the general ethical debate about biobanks. Therapeutic, diagnostic, and forensic biobanks have been used for research, too, but the field of biobank ethics generally concentrates on well-known population-based genetic biobanks. Granted that systematic research on stored human biological material has taken place for more than a century, that other biobanks have long been common but deemed ethically uncontroversial, and that genetic research preceded population-based biobanking, it raises a question: why did the advent of deCODE and other new population-based biobanks in the first decade of the new millennium suddenly raise 'ethical issues'?

There is no easy answer to that question, but it seems that the increase in scale was of some importance. During the 1990s, it became obvious that cohorts needed to be very large to achieve statistically significant results in exploration of genetic dispositions. This implied a transition akin to the one known in physics as the advent of Big Science, when experimental physics necessitated the development of novel forms of collaboration to attract the massive funding needed for the new large-scale experiments (Burton et al., 2009). The increased attention granted by scale interacted with more general public concerns about genetics as well as with related cultural redefinitions of the stored tissue, which now became talked about as 'part of' persons rather than mere 'samples.'

Biobanks did not come with a fine nested set of ethical problems. Diverse groups have negotiated how and why population-based genetic biobanking may constitute ethical problems in reflection of many divergent ideas about what they are, what they should be, and the very point of discussing them. A serious contemplation of biobank ethics must therefore involve reflections on the processes through which ethical problems get named and framed. This article begins with a short reflection of this kind before outlining the issues typically discussed in the literature on biobank ethics.

The Naming and Framing of Biobank Ethics

Ethics, when thought of as an academic discipline, comprises diverse and partly incompatible or at least competing modes of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.82057-7

analyzing moral problems. The ethics literature dealing with biobanks, however, rarely articulates classic differences such as that between utilitarian and deontological positions or between virtue ethics, pragmatic ethics, or feminist ethics. Here, 'ethics' takes a more implicit form as articulation of normative, yet practical, concerns related to the operation and organization of a biobank and its relation to its donating population. Hoffman, Solbakk, and Holm have pointed out how we first need to make sense of biobanks before we figure out what type of problem they might entail. Such sense making usually involves analogies: depending on our chosen analogies, we identify different types of problems (Hofmann et al., 2006). In an analysis of deCODE, for example, Árnason and Simpson showed that when the national media make sense of the company through historical comparisons to the fishing industry, the research endeavor is represented as a contemporary equivalent to well-established Icelandic sources of income: Iceland used to extract natural resources from the sea, and now it is only natural to extract them from the gene pool. The analogy provides the endeavor with enhanced public legitimacy. In other cases, biobanking has been compared with surveillance, whereby biobanks come to represent examples of intrusions of privacy now seen as operating at a genetic level (Williams and Johnson, 2004). Also, analogies are drawn to existing solutions whereby the challenge becomes to align biobanking with standard operating procedures in other settings, such as informed consent procedures used in clinical medicine.

Once ethics develops into concrete suggestions for policies, it also comes to constitute a form of regulation, and as such it interacts with a wider cultural context (Hoeyer, 2005a). Brekke and Sirnes (2006), among others (Hoeyer, 2005b; Metzler, 2010), have argued that biobank ethics entails political consequences by deflecting attention from commercial aspects of biobank exploitation. Biobanking interacts with the cultural context in relation to, for example, the idea of 'genetic exceptionalism,' meaning that everything involving genetics is treated according to its own standards because of popular meanings attached to the word 'gene.' Genetic exceptionalism has been criticized as ethically unfounded, and some have argued that genes are not particularly 'personal' but are shared not only within the family but often across species (Beskow et al., 2001). Still, genetic biobanks have been seen as attracting other regulatory instruments than, for example, nutritional biobanks. Ethicists therefore must reflect on regulatory and cultural implications of their endeavors.

I now turn to what is typically discussed as the ethical issues relating to population-based genetic biobanking beginning with good governance. This is a good place to start because most ethics debates in one way or another feed into thoughts about how one ought to regulate biobanking. Good governance always in some way aims at striking a balance and achieving fairness for the involved actors. The next topic is therefore fairness and reciprocity, followed by an outline of debates of commercial utilization of biobanks, which is a topic deemed particularly controversial and challenging to the aims of fairness. Following the discussion of commercialization, I turn to the issue of trust, which is both represented as a precondition for and an expected consequence of good governance, and something potentially threatened by commercialization. This takes us to issues grouped here as unintended effects comprising concerns that good governance is meant to protect against, but cannot provide guarantees against. I end with the one issue that has gained more attention than any other ethical issue relating to biobanks: informed consent. Informed consent can be seen as the most prominent governance tool in biobank ethics, and I outline the prevailing arguments surrounding this organizational recipe, studies of how it is practiced, and some discussions of its effects.

Ethical Issues

Good Governance

Concrete governance models relate either to the organizational structure, as with suggestions outlining a 'stewardship model' or a 'charitable trust model' (Ashburn et al., 2000; Winickoff and Winickoff, 2003), or to the mode of creating incentives (as with suggestions for biobank impact measurement or taxation of tissue usage; see below). Usually governance models mirror ideals of either universalism or local adaption and optimization. Universalism is more common and occurs in various forms, for example in calls for harmonization to facilitate international collaboration (see, e.g., www.bbmri.eu). The concept of harmonization has also been challenged and accused of covering disagreements between divergent constituencies about what is desirable and of ignoring donor concerns in favor of researcher ambitions (Maschke and Murray, 2004). In the technicality of governance initiatives, basic value differences can be forgotten.

In a reflection on trends and perspectives in biobank governance, Gottweis and Zatloukal (2007) remark that good governance is a response to complex sociocultural challenges and therefore never can be reduced to a technical matter of finding the right tool for the task. Defining the 'good' is basically a normative endeavor. Hume defined public goods as nonrivalrous in consumption and nonexcludable, and this might help us ascertain when a biobank is a 'public good' (Knoppers, 2005); but a definition does not necessarily settle the normative issue of figuring out when and why biobanks produce something desirable, why this outcome is desirable, and for whom. In an attempt to dissect biobanks as public goods, Widdows and Cordell make a distinction between aggregative and corporate community goods, where the former is produced by individuals but compiled by society (as with taxes and government spending), and the latter stems from society and floats to individuals (as with citizen rights) (Widdows and Cordell, 2011). They argue that a community is coproduced with what it perceives as good and in the process of defining the good we seek to identify who shares it. This is also the case with biobanks, they suggest, because biobanks are interesting arenas for public negotiation of value and belonging. Negotiations of contribution (aggregate good) and entitlement (corporate good) become entangled in one another, making it an important task for ethicists to describe the type of community construed through biobanking.

Good governance presupposes a need for identification of the people who should be governed. Ambiguities relate to both users of and contributors to a biobank: are all collaborators of a biobank and the people using the published results biobank

users? Are only the donors, or also their relatives (about whom most biobanks contain information) and other carriers of the same genetic setup, to be seen as research subjects? Should a national biobank consider the interests of all citizens in a given context of collection or limit itself to consideration of the actual donors? Wilson and Chadwick, for example, argue that because biobanks constitute public goods benefitting all members of a society, all citizens are morally obliged to contribute samples to make them as efficient as possible (Wilson and Chadwick, 2007). The difficulties associated with identifying stakeholders appear also when, in the process of establishing good governance, biobankers consider issuing a public consultation. Who, then, is the public? Who should be consulted, and how?

When scholars disagree with Wilson and Chadwick's position, it often reflects contestation of whether or not biobanks really contribute productively to society. Are biobanks *useful* research tools – and for what (Häyry and Takala, 2007b)? Good governance debates therefore involve reflections on what usefulness implies and how it is enhanced in light of very different ideas about what a desirable future holds in store. Debates about usefulness sometimes translate into discussions about incentives for enhancing utility and for sharing biobank resources. Cambon-Thomsen has pointed out how individual researchers who have established biobanks rarely benefit from sharing their material with other researchers, and governance structures that do provide proper acknowledgment therefore need to develop for society to reap the potential benefits of biobanking (Cambon-Thomsen, 2004). Such work, however, builds on the premise that biobanks are worth investing in, and this proposition is contested by researchers thinking that the potential gain from biobanking does not equal the investment (e.g., Watson, 2005). But what constitutes a fair return on a research investment?

Fairness and Reciprocity

Ethical analysis of biobank fairness relates to striking the right balance between state investment and commercial interests (Potts, 2002), or striking a fair balance in relation to what donors get in return for their research contribution. The latter is discussed in the form of both benefit sharing (sharing of medical or commercial benefits) and feedback of research results. Since the ethics debate is heavily centered on the relationship between donor and research institution, I too will focus on the latter set of issues here.

Two competing ethical frameworks construe the relationship between donor and biobank very differently partly because they, as discussed above, view the potential benefits of biobanking differently: Either the donor is seen as deprived of profit in analogy to a Marxian 'exploited worker' (Corrigan and Williams-Jones, 2006), or the citizen who opts out of biobanking is described as a 'free rider' scrounging on the community biobank (Eriksson, 2005). Obviously, the two frameworks construe the biobank community differently. If we limit ourselves to the relationship between actual donors and the biobank and begin with fairness reached through feedback of research results, some scholars have argued for a right to know about research results in general as well as specified individual feedback (Fernandez et al., 2003); while others have argued that unwarranted individual feedback could undermine donor interests and cause insurance complications and unnecessary worries, and that feedback of aggregate research results could hamper donor anonymity and cause public confusion through dissemination of unvalidated propositions (Renegar et al., 2006). Even if it is accepted that a biobank does have a feedback obligation, it remains unclear *how* to deliver feedback without turning it into an obligation on behalf of the donor to keep him- or herself informed; just as it remains undecided which pieces of information will be relevant for whom (Knoppers et al., 2006): Feedback might very well deliver knowledge that is still too uncertain to be of clinical value, and it might relate to susceptibilities that the individual prefers not knowing.

Feedback constitutes a form of direct benefit sharing through which donors are expected to achieve fairer treatment. The sharing of benefits can also be construed as a natural consequence of societal progress, whereby the key ethical issue becomes one of establishing criteria for 'progress': Do biobanks enhance human health, or do they consume resources with little prospect of substantial medical benefit besides, perhaps, more individualized treatment options available mainly for those whose health needs are already well attended to? To deal with these issues, one must contemplate the range of the biobank constituency and its diverse interests *prior* to actual biobank research. Another way of construing benefit sharing relates to the potential sharing of profit (rather than medical benefits), which becomes an issue *after* biobank research has proven a success, also in commercial terms.

Hayden has argued that ethical emphasis on benefit-sharing programs reflects an urge to break with superficial portrayals of 'altruistic donations,' which serve to cover up donor exploitation; but she also points out that such programs represent a form of politics that might generate effects other than the desired 'fairness' (Hayden, 2007). She claims that benefit sharing as a form of politics creates new collectives of those identified as entitled to a share; but also that risks, not only benefits, become shared in new ways because the terms of research participation become construed as a matter of individual negotiation. Furthermore, when putting emphasis on benefit sharing, biobanks might make potential donors think of research participation as a future source of income despite the very limited chances of such profits ever materializing.

Commercialization

The role of the private sector in relation to population-based genetic biobanks is highly controversial for reasons other than just benefit sharing. In some cases, such as UK Biobank, there is an emphasis on the national in the rhetoric used to mobilize donors, an emphasis that is seen by some as at odds with downstream commercial appropriation and patenting of research results (Busby and Martin, 2006) (*see* Anthropology, Genomics, and Human Variation: National Roots). Commercial usage of biobank resources places emphasis on the creation of marketable research results, such as diagnostic tools or pharmacogenetics, potentially at the expense of public health research into prevention or health promotion.

Commercialization is related to a set of legal issues with ethical repercussions having to do with entitlements to

donated tissue. As a general rule, the human body and its parts cannot be bought and sold as plain property. Ownership is therefore not an appropriate category to describe entitlements to tissue. Still, stakeholders may claim a range of other entitlements such as the right to give an informed consent, dispositional rights, rights of utilization, and so on. When the sample leaves the donor, it is typically construed as fully beyond property rights, but further down the line biobank materials may give rise to patentable knowledge that is appropriated by researchers, research institutions, or collaborating companies. In most instances, the actual legal entitlements to donated tissue are quite ambiguous, with several competing stakeholders regarding themselves as having the ultimate dispositional rights concurrently. Some scholars have viewed the incremental downstream commercialization as illustrative of a more general 'commodification of the body'; while others find such concerns misplaced and suggest providing donors with straightforward property rights (see Wilkinson, 2003). Individual property rights to one's body are also seen by some legal scholars as a means for protecting donor privacy (Laurie, 2002), a route that has been criticized by philosophers for misconstruing the key values associated with privacy (Ursin, 2010). Though ideas about commodification might not stand closer analytical scrutiny, strong moral ideals of the separation of bodies and commodities reflect a set of societal values that ethical analysis must also take into account when assessing the stakes in biobanking, in particular if biobankers want to conserve the public trust needed for their endeavor.

Trust and Trustworthiness

O'Neill has pointed out the need to make a basic distinction between trust and trustworthiness (O'Neill, 2002). A biobank may very well be surrounded by trust without deserving it, and it might act honorably and trustworthily without actually being trusted by donors. From a philosophical perspective, she finds it more useful to discuss criteria for trustworthiness than for the existence of 'trust' as an empirical phenomenon, and yet trust continues to receive more attention in biobank debates than trustworthiness.

Trust is discussed both as a means for biobanking and as a desirable end in its own right (Caulfield and Outerbridge, 2002). Trust is an elusive concept. From a sociological perspective, Luhmann argues that trust is a prerequisite of modern times where familiarity is no longer possible in our everyday operations (Luhmann, 2000). As a necessary mechanism in complex networks, trust implies an element of calculated risk that the confidence it expresses might be unfounded. It is often said that there is a 'crisis of trust,' and the concept has also been used in relation to biobanks, but one might also say that biobanks as large-scale, anonymous, and open-ended social networks produce the very sort of social relation in which trust in the sense described by Luhmann emerges. Petersen has argued that claims of a biobank 'crisis of trust' hinge on an illusion of some undefined historical era in which everybody trusted biobankers (Petersen, 2007). There is no reason to assume that it ever was the case, and without any clear measures of trust it is hard to say whether there is less or more of it today. The Eurobarometer surveys indicate that more people are becoming aware of biobanks, and the countries in which most people know about biobanks are also the countries in which biobanking has caused public controversies and gains the highest support. Based on these findings, Gaskell and colleagues argue that perhaps public controversy creates public trust, rather than eroding it, by creating general awareness of biobanking (Gaskell and Gottweis, 2011). Breaches of trust can ensue from breakdowns in confidentiality or abuse of information (Sutrop, 2007), or stem from mismatches between donor expectations and biobank practices, for example with respect to the issues of commercialization mentioned above or the type of disease investigated, which takes us to the issue of unintended effects.

Unintended Effects

Trust can be said to be unfounded if a biobank leads to unwanted, unintended effects. Such effects are what good governance should protect against, and ethicists have pointed to potential dangers relating to risk and safety; breaches of confidentiality and privacy; surveillance; and social implications relating to identity, stigma, and group membership.

Risks associated with biobank research participation are different from those typically in focus in clinical medicine. The living body faces little or no physical risk because the research is done on stored samples (Eriksson, 2005). The major risks are informational in nature, rather than physical, and they therefore reflect the potential harm accruing to the individual through breaches of confidentiality. Such breaches can be analyzed as causing harm if they affect the individual's chances of health insurance, employment, or even marriage. They can also be seen as interfering with the right to privacy. From the latter perspective, it is not the harmful effect as such constituting the problem; it is the violation of intrinsic values relating to human life (using deontological rather than utilitarian premises for the analysis). Not all languages have any word corresponding to the English 'privacy,' and a pan-European project on the protection of privacy in biobanking found considerable cross-national variation in the values attached to a private life, integrity, and intimacy (Taylor and Townend, 2010). Accordingly, different donors will define breaches of privacy differently, and good governance must take specific local approaches in reflection of the prevalent concerns in the concrete biobank context.

One way of protecting donor privacy is through anonymity, but for biobanks to be effective it must be possible to link up samples with registries continuously over time. Hence, protection of privacy comes at a cost of utility. Furthermore, a sample cannot be fully 'anonymized' in as far as it always contains unique biological information, and therefore the merit of anonymization policies as such remains disputed. Also, it has been pointed out that it is sometimes possible to identify otherwise anonymized individuals in published articles through their unique single nucleotide polymorphisms (SNPs) (Homer et al., 2008; Lin et al., 2004).

Social scientists have pointed to unintended effects operating at social levels besides those mainly affecting individual donors. Biobanks could be associated with potential increases in surveillance, for example in conjunction with criminal investigations or public health monitoring or for commercial

purposes. Any increase in available information involves increases in potential oversight and data mining for purposes other than the intended. Besides surveillance, social scientists have discussed how biobank data may generate new genetic identities, reshape group affiliations, and challenge established conceptions of ancestry (Prainsack, 2007; Tutton, 2004). If conditions relating to sexuality, mental state, or behavior become redefined as genetically determined, it may affect how they are perceived and thereby their potential for stigma and/or public acceptance. Genetic research sometimes challenges oral stories of descent, and in some contexts indigenous ancestry is associated with particular legal entitlements making genetics politically hazardous for donors (Reardon, 2005). In short, research participation may have so-called looping effects for identity, entitlements, and societal groupings; and individuals may thereby also facilitate research having unintended effects for people other than the donors who provided their informed consent in the first place.

Consent

Nevertheless, informed consent is beyond a doubt the topic that has been most intensely debated in relation to biobanks, and thereby also the governance tool most commonly suggested as a solution when addressing the various problems (or potential problems) listed above. Informed consent involves full disclosure of relevant aspects of the planned research and voluntary cognizant consent from the research subject. Because population-based genetic biobanks usually work as long-term cohort studies, people donate to a research facility that is expected to run for 10, 30, or 50 years or even longer, and to outlive individual donors and researchers. It is very difficult to predict the type of research that will be deemed most valuable in 30 years, but it is reasonable to say as much as very few people hope it will be the same as today. Research is expected to create new questions. If donors are to be informed about the exact nature of the research they facilitate with their donation, they either have to renew their consent continuously (and perhaps bequeath this right to relatives upon their death) or the biobank value must be seriously reduced. Biobanks therefore have to decide the level of specificity in the information sheets used during the consent process, and it has given rise to an intense debate about the pros and cons of specific versus general consent as well as whether a general consent can be said to be informed at all (Diest and Savulescu, 2002). Also, because genetic traits are to some extent shared, it is discussed whether it is the individual, the family, the community (geographical or disease specific), or the nation who should consent. Such issues have caused a multiplication of concepts (e.g., group consent, community consent, broad consent, soft presumed consent, open consent, general consent, and alternatives such as written authorization). Topics such as proxy consent for minors and incapacitated adults are known from other areas of research, but they have also been treated in relation to biobanks where inclusion in a biobank cohort may take place at birth but not necessarily be followed up with confirmation at the age of majority (Hens, 2010; Samuël et al., 2008). Instead of ensuring consent to a researcher-defined agenda, others have suggested focusing more on genuine donor influence on the research, for example through deliberative democracy models (Secko et al., 2009) or by emphasizing participation rather than consent. Following the latter inclination, it has been suggested to avoid talking about donors and instead think of the people delivering material as ongoing research participants and co–decision makers (Tutton, 2007).

The intense debates about consent (or participation) tend to focus on *what* to do, rather than the ethical reasons *why* it should be done. Is consent desirable because it allows people to follow their personal preference, because it provides the means for individual protection against harm, or because it facilitates desirable relations between researcher and research subject? Often, it is said simply that informed consent serves to protect respect for autonomy, while the concrete meaning of autonomy is left undefined (Häyry and Takala, 2007a). Ethical analysis serves an important purpose in identifying confusions of this sort and helping the involved actors realize their own stakes and values.

Numerous studies have shown that in actual practice, donors do not remember or understand the information provided, and even when they do it rarely relates to the issues they find most important (e.g., Barr, 2006; Ducournau, 2007; Hoeyer, 2003; Skolbekken et al., 2005). The prospect of continued consenting has stimulated terms such as 'consent fatigue' (Cambon-Thomsen, 2004). Also, the emphasis on consent has been criticized for sidestepping values of solidarity (consolidating the before-mentioned free-rider problem) and for practical reasons for hampering good research and adding to the cost of administration (Furness and Sullivan, 2004). Certain discoveries could not have been made had researchers not been allowed to go back and use old samples to identify disease agents that no one would have considered including in an information sheet at the point of sample collection.

Conclusion

Do informed consent practices enhance good governance? There is no way to reach agreement on that issue when it is raised as a simple yes or no question, even though it is an observable fact that such hopes have been attached to the informed consent ideal. Instead, ethical analysis of population-based biobanking must move beyond simple 'what should we do now' questions and allow time to reflect on basic quandaries such as 'who are part of that "we"' and 'what is good' for those involved in and affected by biobanking. Biobanking has become a way of negotiating boundaries for communities and for reflecting on the direction of contemporary medical research. This article has outlined some of the topics through which such negotiations manifest themselves, namely, in debates about public goods, fairness, forms of commercialization, criteria for trustworthiness, and potential unintended effects. Answers remain to be found, and they must be argued in relation to the specific contours of concrete biobank initiatives.

See also: Forensic Genetic Databases: Ethical and Social Dimensions.

Bibliography

Ashburn, T., Wilson, S., Eisenstein, B., 2000. Human tissue in the genomic era of medicine. Archives of Internal Medicine 160, 3377–3384.

Barr, M., 2006. 'I'm not really read up on genetics': biobanks and the social context of informed consent. BioSocieties 1, 251–262.

Beskow, L.M., Burke, W., Merz, J.F., Barr, P.A., Terry, S., Penchazadeh, V.B., Gostin, L.O., Gwinn, M., Khoury, M., 2001. Informed consent for population-based research involving genetics. Journal of the American Medical Association 286 (18), 2315–2321.

Brekke, O.A., Sirnes, T., 2006. Population biobanks: the ethical gravity of informed consent. BioSocieties 1, 385–398.

Burton, P.R., Hansell, A.L., Fortier, I., Manolio, T.A., Khoury, M.J., Little, J., Elliott, P., 2009. Size matters: just how big is BIG? International Journal of Epidemiology 38 (1), 263–273.

Busby, H., Martin, P., 2006. Biobanks, national identity and imagined communities: the case of UK Biobank. Science as Culture 15 (3), 237–251.

Cambon-Thomsen, A., 2004. The social and ethical issues of post-genomic human biobanks. Nature Review Genetics 5, 6–13.

Caulfield, T., Outerbridge, T., 2002. DNA databanks, public opinion and the law. Clinical and Investigative Medicine 25 (6), 252–256.

Corrigan, O.P., Williams-Jones, B., 2006. Pharmacogenetics: the bioethical problem of DNA investment banking. Studies in History and Philosophy of Biological and Biomedical Sciences 37, 550–565.

Diest, P.v., Savulescu, J., 2002. For and against: no consent should be needed for using leftover body material for scientific purposes. British Medical Journal 325, 648–651.

Ducournau, P., 2007. The viewpoint of DNA donors on the consent procedure. New Genetics and Society 26 (1), 105–116.

Eiserman, E., Castillo, J.J., 2000. RAND Monograph Report. Handbook of Human Tissue Sources – A National Resource of Human Tissue Samples. pp. 1–256.

Eriksson, S., 2005. Potential harms, anonymization, and the right to withdraw consent to biobank research. European Journal of Human Genetics 13 (9), 1071–1076.

Fernandez, C.V., Kodish, E., Weijer, C., 2003. Informing study participants of research results: an ethical imperative. IRB: Ethics & Human Research 25 (3), 12–19.

Furness, P., Sullivan, R., 2004. The human tissue bill: criminal sanctions linked to opaque legislation threaten research. British Medical Journal 328, 533–534.

Gaskell, G., Gottweis, H., 2011. Biobanks need publicity. Nature 471, 159–160.

Gottweis, H., Zatloukal, K., 2007. Biobank governance: trends and perspectives. Pathobiology 74, 206–211.

Hayden, C., 2007. Taking as giving: bioscience, exchange, and the politics of benefit-sharing. Social Studies of Science 37 (5), 729–758.

Häyry, M., Takala, T., 2007a. American principles, European values and the mezzanine rules of ethical genetic databanking. In: Häyry, M., Chadwick, R., Árnason, V., Árnason, G. (Eds.), The Ethics and Governance of Human Genetic Databases. Cambridge University Press, Cambridge, pp. 14–36.

Häyry, M., Takala, T., 2007b. Bioethical analysis of the results: how well do laws and regulations address people's concerns? In: Häyry, M., Chadwick, R., Árnason, V., Árnason, G. (Eds.), The Ethics and Governance of Human Genetic Databases. Cambridge University Press, Cambridge, pp. 249–256.

Hens, K., 2010. Banking on the Future. An Ethical Inquiry into the Storage and Use of Bodily Material of Minors for Genetic Research. Leuven University Press, Leuven.

Hoeyer, K., 2003. "Science is really needed – that's all I know": informed consent and the non-verbal practices of collecting blood for genetic research in Sweden. New Genetics and Society 22 (3), 229–244.

Hoeyer, K., 2005a. Studying ethics as policy: the naming and framing of moral problems in genetic research. Current Anthropology 46, 71–90.

Hoeyer, K., 2005b. The role of ethics in commercial genetic research: notes on the notion of commodification. Medical Anthropology 24 (1), 45–70.

Hofmann, B., Solbakk, J.H., Holm, S., 2006. Analogical reasoning in handling emerging technologies: the case of umbilical cord blood biobanking. The American Journal of Bioethics 6 (6), 49–57.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., et al., 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genetics 4 (8), e1000167. http://dx.doi.org/10.1371/journal.pgen.1000167.

Knoppers, B.M., 2005. Of genomics and public health: building public "goods"? Canadian Medical Association Journal 173 (10), 1185–1186.

Knoppers, B.M., Joly, Y., Simard, J., Durocher, F., 2006. The emergence of an ethical duty to disclose genetic research results: international perspectives. European Journal of Human Genetics 14, 1170–1178.

Laurie, G., 2002. Genetic Privacy. A Challenge to Medico-Legal Norms. Cambridge University Press, Cambridge.

Lewis, G., 2004. Tissue collection and the pharmaceutical industry: investigating corporate biobanks. In: Tutton, R., Corrigan, O. (Eds.), Genetic Databases. Socio-Ethical Issues in the Collection and Use of DNA. Routledge, London, pp. 181–202.

Lin, Z., Owen, A.B., Altman, R.B., 2004. Genomic research and human subject privacy. Science 305, 183.

Luhmann, N., 2000. Familiarity, confidence, trust: problems and alternatives. In: Gambetta, D. (Ed.), Trust: Making and Breaking Cooperative Relations. Blackwell, Oxford, pp. 94–107.

Maschke, K.J., Murray, T.H., 2004. Ethical issues in tissue banking for research: the prospects and pitfalls of setting international standards. Theoretical Medicine 25, 143–155.

Metzler, I., 2010. Über "Moralapostel" und "smooth operators": Die Praxis der Bioethik im Feld eines österreichischen Biobankenprojekts. In: Griessler, E., Rohracher, H. (Eds.), Genomforschung – Politik – Gesellschaft. VS Verlag, Wiesbaden.

O'Neill, O., 2002. Autonomy and Trust in Bioethics. Cambridge University Press, Cambridge.

Petersen, A., 2007. Biobanks' "engagements": engendering trust or engineering consent? Genomics, Society and Policy 3 (1), 31–43.

Potts, J., 2002. At least give the natives glass beads: an examination of the bargain made between Iceland and deCODE genetics with implications for global bio-prospecting. Virginia Journal of Law and Technology 7 (8), 1–40.

Prainsack, B., 2007. Research populations: biobanks in Israel. New Genetics and Society 26 (1), 85–103.

Reardon, J., 2005. Race to the Finish: Identity and Governance in an Age of Genomics. Princeton University Press, New Jersey.

Renegar, G., Webster, C.J., Stuerzbecher, S., Harty, L., Ide, S.E., Balkite, B., Rogalski-Salter, T.A., Cohen, N., Spear, B.B., Barnes, D.M., Brazell, C., 2006. Returning genetic research results to individuals: points-to-consider. Bioethics 20 (1), 24–36.

Samuël, J., Ries, N.M., Malkin, D., Knoppers, B.M., 2008. Biobanks and longitudinal studies: where are the children? GenEdit 6 (3), 1–8.

Secko, D.M., Preto, N., Niemeyer, S., Burgess, M.M., 2009. Informed consent in biobank research: a deliberative approach to the debate. Social Science & Medicine 68, 781–789.

Skolbekken, J.-A., Ursin, L.Ø, Solberg, B., Christensen, E., Ytterhus, B., 2005. Not worth the paper it's written on? Informed consent and biobank research in a Norwegian context. Critical Public Health 15 (4), 335–347.

Sutrop, M., 2007. Trust. In: Häyry, M., Chadwick, R., Árnason, V., Árnason, G. (Eds.), The Ethics and Governance of Human Genetic Databases. Cambridge University Press, Cambridge, pp. 190–198.

Taylor, M., Townend, D., 2010. Issues in protecting privacy in medical research using genetic information and biobanking: the PRIVILEGED project. Medical Law International 10 (4), 253–268.

Tutton, R., 2004. "They want to know where they came from": population genetics, identity, and family genealogy. New Genetics and Society 23 (1), 105–120.

Tutton, R., 2007. Constructing participation in genetic databases: citizenship, governance and ambivalence. Science, Technology and Human Values 32 (2), 172–195.

Ursin, L.Ø, 2010. Privacy and property in the biobank context. HEC Forum 22, 211–224.

Watson, J., 2005. Beset by practical hurdles, UK biobank moves at sluggish pace. Nature Medicine 11 (7), 696.

Widdows, H., Cordell, S., 2011. Why communities and their goods matter: illustrated with the example of biobanks. Public Health Ethics 4 (1), 14–25.

Wilkinson, S., 2003. Bodies for Sale: Ethics and Exploitation in the Human Body Trade. Routledge, London.

Williams, R., Johnson, P., 2004. Circuits of surveilliance. Surveillance and Society 2 (1), 1–14.

Wilson, S., Chadwick, R., 2007. Pursuing equality: questions of social justice and population genomics. In: Häyry, M., Chadwick, R., Árnason, V., Árnason, G. (Eds.), The Ethics and Governance of Human Genetic Databases. Cambridge University Press, Cambridge, pp. 150–157.

Winickoff, D.E., Winickoff, R.N., 2003. The charitable trust as a model for genomic biobanks. New England Journal of Medicine 349 (12), 1180–1184.

Relevant Website

www.bbmri.eu.

Biocitizenship

Carlos Novas, Carleton University, Ottawa, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Biological or genetic citizenship is a term used to describe the complex intersection between novel developments in the life and biomedical sciences, the growing involvement of lay citizens in scientific research and health activism, the formation of individual and group identities organized around a genetic or biological classification, and the reconfiguration of the political economy relating to health care provision, as well as the generation of wealth and health from living processes.

In 2008, Roger Cooter (2008) described biological citizenship or biocitizenship as a keyword-in-the-making since this term seemed to emerge in tandem with the phenomenon it sought to describe. As with many keywords in the social and natural sciences, this term was developed almost simultaneously by different authors: Adriana Petryna (2002) in her monograph titled *Life Exposed: Biological Citizenship after Chernobyl*, by Deborah Heath et al. (2004) in their essay titled "Genetic Citizenship," and by Nikolas Rose and Carlos Novas (2005) in their essay titled "Biological Citizenship." These different renderings of what can be termed biocitizenship share an interest in exploring the complex intersection between the life sciences, citizen activism, health and social welfare provision, the regulation and governance of life, and the broader political economy. This article begins by exploring biocitizenship through exploring how the aforementioned authors discuss the relationship between biopolitics and biological or genetic citizenship, the intersection between identity practices and biologization, the political economic relations informing this conception of citizenships, and the role of information and communication technologies (ICTs) in shaping this form of citizenship.

Biopolitics and Biocitizenship

As the terms biological and genetic citizenship are oriented toward charting and describing novel developments in the life sciences, the transformation of health care systems, and the forms through which biomedical research is conducted, they are in effect concerned with examining biopolitics and the government of life in the contemporary present. In the case of the Ukraine, as Petryna (2002) explains, the state, failed both intentionally and inadvertently to produce knowledge about its population that would enable the state to provide for the welfare of its individual citizens and population in the years following the Chernobyl nuclear accident. Petryna's explanation focuses on the state's shifting field of medical categorizations relating to radiation sickness, and its uncertain criteria for compensating those affected by this disaster. Given this context of uncertainty and unknowability in post-socialist Ukraine, for Petryna (2002), biological citizenship is about the efforts of citizens to use their embodied experiences, their suffering, and their acquisition of biomedical and legal knowledge to create social welfare institutions that not only benefit those affected by radiation exposure, but much more broadly, to provide for the health and well-being of this population.

Heath et al. (2004) also draw analytically on a version of biopolitics in their elaboration of genetic citizenship, focusing in particular around struggles regarding the provision of health and social welfare to United States citizens. Although the bulk of their ethnographic analysis focuses on the efforts of genetic advocacy groups, scientists, clinicians, and legislators to transform the organization of health care and the biomedical research establishment around rare diseases in the United States, they reference the important role of the AIDS and women's movement in transforming the institutional and therapeutic modalities through which life is governed in this particular national context. For Heath et al. (2004), biopolitics is not simply a matter of the state, but includes health activists, physicians, politicians, scientists, and nonhuman actors such as genetic databases and bioengineered medical devices as part of what they term genetic governmentality.

Rose and Novas' (2005) account of biological citizenship also seeks to extend the analytical purchase of biopolitics by attempting to describe how citizenship projects have been linked to beliefs about the biological existence of humans as individuals, as a population, and as a species. In contrast to Petryna (2002) and Heath et al. (2004) whose work focuses on the Ukraine and the United States, Rose and Novas (2005) employ the term to discuss the intersection between biology and citizenship in the history of European and North American state formations. The authors build upon Michel Foucault's (1978) account of biopolitics by emphasizing that biological citizenship has individualizing and collectivizing facets. It has an individualizing dimension in the sense that individuals in the past and in the present use medical and genetic knowledge to understand and shape certain aspects of their selves. It is collectivizing in the sense that biological beliefs about the human species and 'races' in the past served as a means to differentiate and exclude certain groups of citizens. As Rose and Novas (2005) briefly discuss, biological citizenship in the early part of the twentieth century assumed a racialized form in Europe, North America, South America, and Australasia. For Rose and Novas (2005), the eugenics movement provides one exemplar of biological citizenship as it sought to act upon the biological characteristics of citizens and populations in the name of racial purity, mental hygiene, the elimination of 'undesirable' genetic diseases, and the augmentation of public health using a variety of compulsory or voluntary measures

 http://dx.doi.org/10.1016/B978-0-08-097086-8.85059-X

depending on national context. Shifting to the present, Rose and Novas (2005) argue that biological citizenship does not assume a nationalized or racialized form, but is collectivizing in the sense that biological understandings can form the basis for a shared identity around a particular genetic or chromosomal abnormality, that may lead to the formation of organizations oriented around challenging medical expertise or that seek to promote greater medical research on a particular genetic condition.

Identity and Biocitizenship

A defining feature of biological citizenship is the claim that new connections are being forged between biology and identity. Building upon Paul Rabinow's (1996) concept of biosociality, biological citizenship makes reference to the growing role of patients' organizations and health-based social movements in contemporary biomedicine. In the case of Rose and Novas (2005), as well as Petryna (2002), they put forward the claim that the types of relationships being forged between identity and biology are distinct from those organized around the eugenics movement in the early part of the twentieth century, which were based on ill-founded conceptions of human genetics and race. A notable difference from the past, as these authors suggest, is that the association between identity and biology is said to give rise to illness-based social movements. As Petryna (2002) points out in her work on the case of post-Chernobyl Ukraine, suffering, in both its biological and social sense gave rise to legitimate claims on the state for medical care, for health care reform, for scientific research, and for democratic political reform. For Rose and Novas (2005) the link between identity and biological citizenship refers to the organization of collectivities around a biological conception of shared identity; collectivities who work to gain specialized medical knowledge about their condition, advocate for greater research funding for their illness, create novel communities of support through Internet forums and Web sites, and seek to develop novel modalities of patient participation in medical research. Heath et al. (2004) understand the ties between identity and genetic citizenship to be located at the multiple intersections of embodied experience, genetic difference, and practices of care associated with political activism, coalition building, and participatory knowledge making.

Economies of Life: The Political Economy of Biocitizenship

In Petryna's (2002) ethnographic account of post-Chernobyl Ukraine, the economic hardships faced by this emerging democratic nation-state had an impact on its health care system and policies regarding compensation for those affected by radiation sickness. As Petryna (2002) documents in her study, the individual and collective survival strategies of biological citizens to seek compensation for their injuries takes place in a context where basic survival is at stake and in conditions of economic and social exclusion.

In their rendering of genetic and biological citizenship, Heath et al. (2004) also are concerned with mapping the novel economic relations emerging at the intersection of health activism and the life sciences. This also is the case for Rose and Novas (2005). Heath et al. (2004) document how the health advocates they studied often had to form complex political economic relations with the state and the market in order to gain access to medical treatment, to ensure the provision of social services, and to further biomedical research on rare orphan diseases. The forms of the activism of such genetic advocacy groups are engaged in often transcend the boundaries of a single genetic condition, and are oriented toward imagining 'other possible futures' and an 'ethics of care' in US public discourse.

In their account of biological citizenship in the contemporary present, Rose and Novas (2005) argue that is oriented around the generation of biovalue, a term that they borrow from Catherine Waldby (2000). According to Rose and Novas (2005), advances in molecular biology and biomedicine, alongside the development of national biobanks in the early part of the twenty-first century make it possible to harness the blood, tissue samples, and health data of national populations, in order to generate potential revenue streams for biotechnology companies, in addition to possibly augmenting the health of national populations. For Rose and Novas (2005), an animating feature of what they term a 'political economy of hope' is that the generation of biovalue extends to patients' organizations who, through the establishment of blood and tissue registries, are able to generate economic returns in the form of patent royalties, and, as in the case of PXE International, health benefits through the development of diagnostic techniques or therapies. And, lastly, this political economy of hope generates novel ethical values regarding the participation of patients in biomedical research.

Biocitizenship and Digital Life

Another theme that unites accounts of biological (Rose and Novas, 2005) and genetic citizenship (Heath et al., 2004) is the emphasis accorded to the role of ICTs in enabling socially diverse and geographically dispersed persons affected by genetic disease to share life experiences, to provide social support to one another, to disseminate biomedical knowledge, to engage in participatory knowledge making, and lastly, to coordinate advocacy efforts. The Internet not only is a tool used for communication among patients and families affected by genetic disease; it also facilitates cooperation with clinicians and scientists, and can be used as a marketing tool through which pharmaceutical and biotechnology companies engage with a diverse range of audiences.

See also: Biomedical Sciences, History and Sociology of; Citizenship, Anthropology of; Citizenship, Historical Development of; Citizenship, Sociological Aspects of; Citizenship: Political; Foucault, Michel (1926–84); Governmentality; Science and Social Movements; Science and Technology Studies: Experts and Expertise; Science, New Forms of; Science, Public Engagement with; Sexual Minorities and Sexual Citizenship; Stem Cell Research and Cloning.

Bibliography

Cooter, R., 2008. Biocitizenship. The Lancet 372 (9651), 1725.

Foucault, M., 1978. The history of sexuality. In: The Will to Knowledge, vol. 1. Penguin, London.

Heath, D., Rapp, R., Taussig, K.S., 2004. Genetic citizenship. In: Nugent, D., Vincent, J. (Eds.), A Companion to the Anthropology of Politics. Blackwell, Oxford, pp. 152–167.

Petryna, A., 2002. Life Exposed: Biological Citizenship after Chernobyl. Princeton University Press, Princeton, NJ.

Rabinow, P., 1996. Artificiality and enlightenment: from sociobiology to biosociality. In: Rabinow, P. (Ed.), Essays on the Anthropology of Reason. Princeton University Press, Princeton, NJ, pp. 91–111.

Rose, N., Novas, C., 2005. Biological citizenship. In: Ong, A., Collier, S. (Eds.), Global Assemblages: Technology, Politics, and Ethics as Anthropological Problems. Blackwell, Malden, MA, pp. 439–463.

Waldby, C., 2000. The Visible Human Project: Informatic Bodies and Posthuman Medicine. Routledge, New York.

Bioethics

Felix Thiele, Europaeische Akademie, Bad Neuenahr-Ahrweiler, Germany; and Institute for Philosophy, University of Duisburg-Essen, Essen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Bioethics is a relatively young field of learning drawing on many established academic disciplines, such as philosophy, jurisprudence, sociology, life sciences, and others. Bioethics is concerned with genuine problems from philosophical ethics as well as practical questions arising from the medical and nonmedical life sciences and affecting public policy and the direction and control of science. This article aims at providing an overview on the meaning and scope of bioethics, the history of bioethics, the international and intercultural dimension of bioethics, the contribution of related disciplines, and major substantive issues in bioethics – such as decisions at the beginning and end of life, allocation of resources, and moral problems connected to research in life sciences.

The History of Bioethics

Bioethics is a relatively young field of learning drawing on many established academic disciplines, such as philosophy, jurisprudence, sociology, life sciences, and others. The main stimuli for the increase of research in bioethics around the 1960s arose from the technical progress in medicine, societal changes in the western world, and a revitalized interest in the discipline of philosophical ethics (Jonsen, 2003). Around that time, progress in medical technology – e.g., kidney dialysis, artificial respirators, organ transplantation, medically safe abortions, and the contraceptive pill – radically changed the practice of medicine and created the demand to apply ethical reasoning to concrete situations. In addition, the first steps toward a powerful gene technology were made. Roughly at the same time, profound changes occurred in western societies. The civil rights movements in the United States and elsewhere, as well as the transformation of traditional institutions such as the family and the church, caused moral questions concerning an individual's rights and duties in modern society. These developments were accompanied by a rising concern about humanity's impact on the world sustaining us and the future generations, leading to the development of new subspecialties of bioethics such as animal ethics and environmental ethics.

Finally, the philosophical discipline of ethics received a fresh impetus. For a long time, there had been consensus among most philosophers that judgments about moral problems could not be rational in principle. In the second half of the twentieth century, much work has been done, which successfully aimed at providing scientific methods for moral reasoning (Arrington, 1997; Stingl, 1997). In combination, these developments formed the basis of the extraordinary rise of bioethics in the second half of the twentieth century. Meanwhile, bioethics has become institutionalized in academic departments, professional journals, and international associations.

The Scope of Bioethics

Although medical ethics remains to be the most known feature of bioethics, the latter has a broader scope and is understood to encompass questions arising in the context of the life sciences, that is, medicine and the biological disciplines, as well as in some fields of the environmental and social sciences.

The scope of bioethics can be divided roughly into three fields of research. First, there is the project of a rational justification of moral norms. It forms the methodological core of bioethics, since it supplies the strategies of moral reasoning that are applied to the respective problems at issue.

Second, there is what might be called clinical ethics; this group encompasses moral problems that individuals, for example, nurses, physicians, or researchers, are confronted with in their daily practice. Examples of this kind involve decisions about the end of life, for example, decisions regarding abortion or physician-assisted suicide.

Finally, there is a group of problems that should be dealt with on the regulatory or policy level. Examples of problems of this kind involve the necessity to develop operationalizable juridical guidelines for awarding patents on biological materials, policies for the allocation of medical resources, or social challenges from neuroenhancement use of the animals for medical purposes or access to the benefits of the advancement of science and technology.

The Method of Bioethics

Morality and Philosophical Ethics

The task of professionalized ethics is not to be seen in gaining insights from some special source of knowledge – insights that would entitle the ethicist to demand a certain course of action or insist on certain conduct being prohibited. Rather, the proper task of ethics is the critical assessment of existing moral convictions and norm systems established from them. This assessment should be done by using a method that is in principle accessible to everyone. Moral convictions are subject to such an assessment when they generate conflicts, for example, when they are incompatible with other moral convictions. In the case of moral conflicts there is a need for an ethical criterion that allows us to establish which conviction should be followed. A simple example for such a criterion is the 'Golden Rule': "We should do unto others as we would wish them to do

unto us." This rule does not provide us with a material action-guide, but provides a formal criterion that any action guide has to fulfill in order to be morally acceptable.

Traditionally, it has been the domain of philosophical ethics to design strategies for the solution of moral conflicts. In this sense bioethics is only a new area because the conflicts to be solved were unknown until progress in medical technology and societal changes directed public attention to the ethical problems of the life sciences.

It should be the role of a philosopher wishing to contribute to the solution of bioethical issues to provide a clarification of the moral concepts involved and to reconstruct the arguments used in the discussion. On this basis he can develop strategies of moral reasoning that help determine which moral convictions one ought to accept.

It should be made clear, however, that these strategies are developed on the basis of an underlying ethical theory and that a variety of different ethical theories has been developed during the history of ethics (for an introduction, see Frankena, 1973): consequentialism claims that, if anything determines the moral status of an action, then it is the consequences of that action (Singer, 2011). Contractarian theories analyze what it means for a society to endorse principles of fairness and justice (Daniels, 2008). In revitalizing ancient approaches, virtue ethics aims at developing action guides that tell agents what they ought to do in concrete situations (Pellegrino and Thomasma, 1993).

Principle-based approaches hold that some general moral norms are central to moral reasoning; contrary to this (Beauchamp and Childress, 2013), case-oriented approaches advocate starting with the examination of concrete cases and proceeding from that to more general norms (Jonsen and Toulmin, 1990). An example of the latter is the ethics of care: it aims at an expressly practice-oriented approach to bioethics by engaging emotionally in the moral problems at issue (Gilligan, 1982). The above list of ethical theories is neither complete nor does it present a proper classification of such theories, but it names some of those theories most relevant in the recent debate on bioethics.

Such theories often share certain presuppositions, for example, the demand for 'universalizability' and 'consistency,' but differ in others, so that – depending on the theory applied – different strategies of moral reasoning can be developed in the discussion of a certain moral problem at issue.

It is often used as an argument against the use of moral reasoning for practical questions that there are no indisputable ultimate arguments for moral norms achievable on a rationally understandable basis. Although true, this should lead to modesty, not resignation, concerning the power of moral argumentation, since we should not demand in the field of ethics what is not available in other areas also: for example, in applying the laws, a judge is always left with a degree of discretion within which he has to examine carefully the special circumstances of the case at issue.

The ability to analyze the moral concepts involved and to apply one or the other ethical theory to bioethical problems will help the physician, policy maker, or whoever else is confronted with these questions to think about them in a sophisticated way and to develop moral convictions acceptable to all parties concerned.

Bioethics and the Natural and Social Sciences

The methodology supplied by philosophical ethics is only one, albeit essential, component in the attempt to solve bioethical problems. The philosopher is trained in examining the arguments proposed, whether they are valid relative to some standards of rationality or whether they fall back upon metaphysical or emotion-based beliefs. Substantial arguments about bioethical problems should be based, however, on knowledge from those sciences in the domain in which those problems originated. A discussion of the ethical problems of genetic counseling, for example, without the participation of physicians and others involved in this counseling would be futile. Similarly, it would be pointless to establish the impact that the results of a new technique, such as the genetic manipulation of food, will have on society without the participation of those social sciences that have the necessary empirical methods at their disposal. Finally, it would be vain to develop recommendations for societal regulations, for example, economically effective and juridically sound allocation procedures for medical resources, without the backing of health economics and jurisprudence. Bioethics is embedded in an inter- and transdisciplinary context and supports the discussions of other scientific disciplines and politics. Moreover, it helps to raise public awareness, social involvement, and democratic governance of the further development of science and technology.

The International and Intercultural Aspects of Bioethics

The experience of a conflict between existing moral convictions in view of the consequences of new techniques is not confined to any cultural or geographical frontier: it is a universal experience. The actual object of the conflict, however, may be highly dependent on, for example, the culture in which the conflict arises – examples are the debates on whether informed consent to, say, research participation must be given strictly by the (autonomous) individual concerned, or can be replaced by family or community consent, or the controversies on just benefit sharing in exploiting bioresources around the globe.

This context dependency of bioethical problems is often taken as the prime argument against there being a single method of bioethics, since solutions to moral problems are – so it is claimed – valid only relative to the particular situation for which they were developed. But, although it is often the case that material moral convictions do differ between cultures, this does not in itself exclude the possibility of there being a formal criterion acceptable across cultures – in the same way that the differing views of the plaintiff and the defendant do not in themselves exclude the possibility of an acceptable judgment.

Major Issues in Contemporary Bioethics

The sheer number and variety of issues discussed in contemporary bioethics compels us to make a rigorous selection. The following section throws a spotlight on some topics that were intensely discussed in recent debates and that furthermore aptly illustrate an interdisciplinary approach to bioethics (see Chadwick, 2012; Kuhse and Singer, 2006; Post, 2003).

Medical Ethics

Since the time of Hippocrates and the ancient Greek medico-philosophers, the emphasis of medical ethics has been on the doctor–patient relationship. Since the 1960s, this has been debated mainly in terms of the improvement of patient autonomy in view of an allegedly prevailing paternalism on the physician's side (Katz, 1984; Faden and Beauchamp, 1986). Today medical ethics has developed into a highly differentiated research area with rich conceptual debates, for example, on the meaning of human dignity, as well as thoroughgoing analyses of structural changes in medical practices and of new medical technologies for the sake of solving existing and avoiding future moral conflicts. Topics of ongoing debate are among other moral problems arising at the beginning of life, the end of life, clinical practice, and the allocation of resources in and to the health care system, access to health care, biobanking, and organ transplantation.

Beginning of life

The moral acceptability of abortion as well as the fate of handicapped newborns is one of the most extensively and intensely discussed issues in bioethics. Basic to this discussion is what quality an entity must possess in order to be a moral subject endowed with certain rights, for example, a right to life, and how we know when an entity actually possesses this quality.

For some authors, this quality is simply that of belonging to the species *Homo sapiens* – a position called speciesism and criticized because affiliation to a certain biological class is seen as morally irrelevant as the affiliation to a certain race, nation, or social class.

According to another position, this quality is that of being a person– where being a person is usually understood to involve some kind of self-consciousness. A third position claims that being an agent, that is, causing one's own actions, makes an entity a moral subject. Characteristic of the latter two positions is that, according to them, there is the possibility that some members of the species *H. sapiens* are not persons or agents (e.g., a human fetus) while some persons or agents are not members of species *H. sapiens* (e.g., gorillas or dolphins). Once what counts as a moral subject has been defined, it still has to be clarified what rights this moral subject is endowed with. It is, for example, not plausible to claim that a fetus is endowed with an absolute right to life, when even adults are usually not awarded such a far-reaching right: for example, in the event of war and in self-defense (Tooley, 1972).

Over and above this, some criterion has to be found which is applicable in practice and which allows us to determine whether an entity belongs to the class of moral subjects or not. Here an interdisciplinary approach is necessary that combines the above moral considerations concerning the quality characterizing a moral subject with the state-of-the-art knowledge from the natural sciences about when an entity actually possesses this quality.

End of life

It is widely accepted today that patients have the moral right to refuse medical treatment even if this may result in their death. In some cases, patients not only refuse further treatment but explicitly want to end their life. Even if there is no consensus on the moral evaluation of suicide, there is widespread understanding and legal toleration when patients commit suicide as a radical means for ending what they deem to be intolerable suffering. It is controversially debated, however, whether a physician may assist a patient in committing suicide (physician-assisted suicide) or even actively bringing the patient's life to an end (euthanasia). It has long been (and still is) a rule of most professional codes in medicine to forbid active euthanasia, whereas passive euthanasia is sometimes allowed. This distinction between a morally reprehensible action (active euthanasia) and a morally acceptable action (passive euthanasia) is based on an alleged moral difference between 'killing' (act) and 'letting die' (omission). There have been numerous attempts to point out the morally relevant differences between these – without much success so far (Steinbock and Norcross, 1994). Although active euthanasia may under certain conditions be acceptable from the point of view of ethics, academic as well as political and public debates on this issue remain controversial. Only very few countries, for example, the Netherlands, have legalized active euthanasia. The reason for this might be that performing (or even only advocating) active euthanasia can jeopardize the patient's trust in his doctor and the health-care system.

Another controversially debated topic in bioethics is the brain-death criterion: modern medical techniques allow us to sustain the blood circulation of patients, who could not breathe spontaneously due to the loss of certain basal brain functions. According to the traditional definition of death – the standstill of the cardiovascular system – these patients are not dead (although they depend on the continuous support of a respirator). On the one hand, these patients have warm bodies and rosy skin, like a sleeping person, and unlike a corpse, on the other hand, they are irrevocably unable to show any behavior expected from a 'normal' living human being. Since the traditional death criterion was deemed to be inconclusive and unsuitable for medical practice a new criterion was developed. The Ad Hoc Committee of Harvard Medical School to Examine the Definition of Brain Death in 1968 developed the brain-death criterion from which modern procedures for determining brain death are derived. It is important, however, to notice that the brain-death criterion is not the only possible definition of death – indeed the traditional criterion of complete cessation of blood circulation is another one. In a certain sense, death is not naturally given but defined by convention.

Clinical Practice

The interaction of patient and doctor (and other health care professionals) undoubtedly forms the core of medical practice. A salient feature of this relation is that the health care professional due to his training and experience has knowledge that the patient usually has not. It is especially the professional experience that allows the medical professional to collect and interpret a multitude of parameters relevant for evaluating a patient's situation, and for making suggestions for her treatment. Moreover, any patient is for the health professional one out of many (severely) ill patients he has to professionally take care of. For the patient, however, her (severe) illness frequently constitutes a crisis radically changing her life. These at first only

descriptive differences in knowledge and judgmental capacities are interpreted differently in view of their normative implications and precipitate in different conceptions of the patient–doctor relation – a frequently used scheme is the differentiation between a paternalistic model, a contract model, and a partnership model (Emanuel and Emanuel, 1992). These models are used, for example, in the debate on medical paternalism, that is, the question whether a doctor or other health care professionals are morally justified to limit a patient's freedom of decision for the sake of the patient's well-being (Katz, 1984).

Further moral challenges are caused by structural changes in the health care system. The traditional ideal of a doctor–patient relationship, where the doctor commits himself exclusively to his patient is hardly realistic under the conditions of modern health care systems. Doctors as other health-care professionals are part of complex institutions such as hospitals and have, therefore, obligations, for example, toward their employer or the health insurance of their patients – that may be in conflict with the aim to provide the best health care for the patient.

Finally, as research in medical ethics differentiates increasing attention is paid to the specific needs and vulnerabilities of patient groups such as minors, and the mentally ill. A core problem in these groups is that the now commonly accepted duty to respect a patient's autonomy is not easily applied in patient groups whose capacities for autonomous choice are limited or even not existing (Bloch and Green, 2009).

Allocation of Resources

Whenever limited resources are allocated (rationing), refraining from taking desirable actions of one kind becomes the price of performing actions of another kind. Health care is no exception to this universal truth. Fixing a budget in the health-care sector means that an increase in spending for, say, preventive medicine has to be accompanied necessarily by decreasing spending in another sector, say, transplantation centers or, on another level, higher spending in a state's education or military sector necessitates lower spending in other sectors, for example, the health system. The lack of even a basic public health-care system makes allocation of resources to the health care system an important topic in many developing countries. But increasing costs of health care have brought the topic of allocation of resources in the health-care system onto the general agenda even in those countries that are equipped with a public health-care system. The moral questions arising from this are among other things: if, as it widely accepted, there is a human right to health care, what has to be provided by the state: a decent minimum or optimal provision? Who shall allocate the scarce resources and by what criteria? Is there a moral obligation to redistribute health-care resources on a global level (Daniels, 2008)?

Human Nature: Understanding and Intervening

Medical innovation without research involving human subjects seems hardly possible. Much research is done in test tubes, cell cultures, and in nonhuman animals, but at least clinical research investigating effectiveness and safety of medical innovations in human subjects is without alternative. Moreover, using nonhuman animals for medical purposes is burdened with scientific and moral challenges: scientific, because the transferability of results gathered in nonhuman animals such as rodents to humans is limited; moral, because inflicting suffering and death on nonhuman animals for human purposes is itself contested (Sandø and Christiansen, 2008).

Fundamental for performing research in human subjects is that the test person has given her informed consent (Berg et al., 2001). Over and above the controversies on the adequate conceptualization of 'informed consent' and its application in the practice of research, even more complicated challenges arise when it comes to research in patient groups such as minors and the mentally ill whose capacity for informed consent is frequently limited. Another vulnerable patient group is, for example, that of pregnant women and their fetuses. Where in the past decades the common view was that vulnerable populations are best protected from harm through research by excluding them from risky clinical research, it is now increasingly argued that this restrictive practice bears the danger of depriving vulnerable populations from research-driven medical innovation.

Ethics of Genome Research and Manipulation

Human genome research – best known via the International Human Genome Project, but actually consisting of a multitude of different research programs – has experienced an extraordinary boom during the past decades. Although this research provides promising unprecedented options for the prevention, diagnosis, and therapy of diseases, this bears considerable risks. In this context it is useful to differentiate medico-technical problems of human genome research (e.g., the safety of gene therapy protocols) from moral problems (e.g., the moral acceptability of intervening into the human genome). Only the latter will be considered in the following.

Genetic tests performed before and after birth as well as the genetic screening of populations are debated controversially because of the potential misuse of the information acquired. Even if the test is used only for medical purposes, the results might lead to stigmatization and discrimination of individuals or ethnic groups. Although not based on the state-of-the-art knowledge from the biomedical sciences, this geneticization of medicine should be countered by very careful implementation of genetic diagnostics in the medical practice, including a stronger than usual emphasis on the protection of the privacy and autonomy of the tested (Buchanan et al., 2000).

Another possible domain of use of genetic diagnostics is the testing of individuals applying for insurance coverage (Sorell, 1998): it is feared that individuals with an increased genetic risk might have to pay higher premiums or, worse, might be unable to obtain insurance coverage at all.

It is hoped that the incorporation of human or nonhuman genes in body cells (somatic gene therapy) or the human germline (germline gene therapy) will be a major contribution to the therapy of diseases with a genetic component – hereditary as well as nonhereditary. Somatic gene therapy is seen as a relatively unproblematic method from the ethical (not medico-technical) point of view, since the transferred genes are not inherited by the descendant of the treated individual.

In stark contrast to this, germline gene therapy is controversially debated. Since the transfer of genes to germline cells affects all offspring, the question arises whether such an intervention in the genetic constitution of an unborn individual can be legitimated and, on a more general level, whether germline therapy leads to a new form of eugenics or genetic enhancement. Since most authors hold that if germline gene therapy were to be applied, it cannot be exclusively confined in the long run to medical application; the debate, therefore, centers on the question whether the intervention into the lives of future persons by genetic enhancement is different in any morally relevant way from classic interventions such as education (Agar, 1998; Habermas, 2003).

The first cloning – i.e., asexual reproduction – of a mammal in 1996 raised moral questions concerning the application of this technique to nonhuman animals and humans (Harris, 1997). Animal cloning is debated mainly in view of the legitimation of manipulating the genome of wild-type animals and the effect of a change in the genetic pool of the biosphere that might have considerable effects on biodiversity.

The debate on human cloning has focused on the ethical problems of producing copies of existing human beings (reproductive cloning). Concerning reproductive cloning, the emphasis of the moral debate lies on whether there are morally acceptable aims for the actual application of this technique, whether the genetic identity of a clone and its donor amounts to a morally problematic personal identity, and whether human cloning leads to an infringement of the dignity of the clone or the donor.

The cloning only of parts of humans (cells or organs) is technically still in its infancy but successful cloning of human cells has been performed leading to pluripotent stem cells. This has caused an intense debate concerning the moral status of the embryonic cells (see Lauritzen, 2001).

Neurosciences

Rapid advance in the neurosciences over the past decades has not only greatly improved our understanding of the brain function and disease, but also created new powerful means for modulating brain function. Examples for the former are among other imaging techniques, neurogenetics and computational models of the brain. Prominent examples for the latter are techniques such as deep brain stimulation and also psychotropic drugs.

These developments reactivated debates on the fundamental relation of mind and brain, for example, on whether the mind can be reduced to, or is noting else than the deterministic brain function (Bennett et al., 2009). Related are debates on the justification and exact meaning of concepts like 'free will' and 'moral responsibility.' Far from being theoretical only these debates may have implications for education, and the law, especially, the penal system (Illes and Sahakian, 2011).

Another area arousing great interest in both academia and the general public are attempts to not only restore restricted brain function, but to improve brain function beyond the normal. This neuroenhancement raises questions on what 'normal' means, and whether and to what extent it may be justified to change the normal human nature that hitherto was largely something humans had to accept as given, and not something that can be changed by them. In a radical form support for enhancing human capacities leads to transhumanism, that is, the program of not only pushing human capabilities to their natural limits, but of dissolving these limits themselves, that is, of transcending human nature (Hansell and Grassie, 2011).

Animal Ethics

The debate on the moral status of nonhuman animals and the moral boundaries of using them for human purposes has become a prominent topic in bioethics over the decades. Challenging in this context is not only the way we treat those animals (primates and dolphins) that possibly have higher mental capacities in some respects akin to those humans have, but also the conditions for using animals for food production, and for research purposes.

Central to animal ethics is the debate on what moral status animals do have: according to some authors it is some kind of higher mental capacity that conveys moral status to certain animals. According to another position, it is not higher mental capacities that only very few animals possess likely, but the ability to suffer which seems to be widespread in the animal kingdom. Further arguments tie the moral status of animals to their having specific needs or simply to the fact that they are living organisms.

Besides these attempts to clarify the foundations of animal ethics, the practice of using animals for human purposes finds increasing interest in academic moral discourse as well as in public debates. Topics among others are the moral implications of industrial livestock farming, that is, the justification of customizing animals using classical breeding methods and gene technology for production needs and consumer preferences. In this context ethical questions on, for example, 'naturalness' and 'integrity' as moral criteria for establishing moral boundaries for the intentional manipulation of life forms are raised. In addition, risks for biodiversity and the stability of ecosystems are discussed. Another hotspot of animal ethics is the concern with using animals for purposes of medical research (Sandø and Christiansen, 2008).

Synthetic Biology

Synthetic biology is a young and dynamic branch of modern biology that focuses on understanding the conceptual design of living systems. What makes up the specific difference between classic molecular biology and synthetic biology is debated, but in a first approximation synthetic biology in contrast to molecular biology aims at changing organisms by redesigning large parts of them, or even by introducing entirely new functions. In a radical form synthetic biology is aiming at creating new, that is, synthetic life forms. On the application level synthetic biology is hoped to facilitate the search for new effective medicines, cheap biofuels, and new methods to produce petrochemicals. The salient challenges of synthetic biology are biosafety and biosecurity risks. The biggest problem in the assessment of biosafety risks of synthetic organisms is their strong divergence from well-known natural systems. When there is no natural reference system, many proven methods of technology assessment are not applicable (Gutmann, 2011). The de-skilling agenda of synthetic biology

and the possibilities of dual use also lead to enormous biosecurity risks, since states, or terrorists may exploit synthetic biology for hostile purposes.

See also: Animal Rights in Research and Research Application; Bioethics: Philosophical Aspects; Direct-to-Consumer Personal Genetic Testing; Ethical Dilemmas, Research, and Access to Treatment; Ethics for Biomedical Research Involving Humans: International Guidelines; Euthanasia; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genomics, Ethical Issues in; Informed Consent of Research Subjects; Intellectual Property Rights: Ethical Aspects; Research Ethics, Cross-Cultural Dimensions of; Research Ethics: Research; Risk Screening, Testing, and Diagnosis: Ethical Aspects.

Bibliography

Agar, N., April 1998. Liberal eugenics. Public Affairs Quarterly 12 (2), 137–155.

Arrington, R.L., 1997. Ethics I (1945 to the present). In: Canfield, J.V. (Ed.), Routledge History of Philosophy, vol. X. Routledge, London.

Beauchamp, T.L., Childress, J.F., 2013. Principles of Biomedical Ethics. Oxford University Press, New York.

Bennett, M., et al., 2009. Neuroscience and Philosophy. Brain, Mind, & Language. Columbia University Press, New York.

Berg, J.W., Appelbaum, P.S., Lidz, C.W., Parker, L.S., 2001. Informed Consent. Legal Theory and Clinical Practice. Oxford University Press, Oxford.

Bloch, S., Green, S.A. (Eds.), 2009. Psychiatric Ethics. Oxford University Press, Oxford.

Buchanan, A., Brock, D.W., Daniels, N., Wikler, D., 2000. From Chance to Choice. Genetics and Justice. Cambridge University Press, Cambridge.

Chadwick, R. (Ed.), 2012. Encyclopedia of Applied Ethics, 4 vols. Academic Press, San Diego, CA.

Daniels, N., 2008. Just Health Care. Meeting Health Needs Fairly. Cambridge University Press, Cambridge.

Emanuel, E.J., Emanuel, L.L., 1992. Four models of the physician-patient relationship. Journal of the American Medical Association 267 (16), 2221–2227.

Faden, R.R., Beauchamp, T.L., 1986. A History and Theory of Informed Consent. Oxford University Press, New York.

Frankena, W.K., 1973. Ethics. Prentice-Hall, Englewood Cliffs, NJ.

Gilligan, C., 1982. In a Different Voice. Psychological Theory and Women's Development. Harvard University Press, Cambridge.

Gutmann, A., 2011. The ethics of synthetic biology: guiding principles for emerging technologies. Hastings Center Report 41 (4), 17–22.

Habermas, J., 2003. The Future of Human Nature. Blackwell Publishers, Oxford, UK.

Hansell, G.R., Grassie, W. (Eds.), 2011. Transhumanism and Its Critics. Metanexus, Philadelphia.

Harris, J., 1997. Goodbye Dolly? The ethics of human cloning. Journal of Medical Ethics 23, 353–360.

Illes, J., Sahakian, B. (Eds.), 2011. Oxford Handbook of Neuroethics. Oxford University Press, Oxford.

Jonsen, A., 2003. The Birth of Bioethics. Oxford University Press, New York.

Jonsen, A., Toulmin, S., 1990. The Abuse of Casuistry. A History of Moral Reasoning. University of California Press, Berkely.

Katz, J., 1984. The Silent World of Doctor and Patient. The Johns Hopkins University Press, Baltimore.

Kuhse, H., Singer, P. (Eds.), 2006. Bioethics. An Anthology. Blackwell Publishers, Oxford, UK.

Lauritzen, P. (Ed.), 2001. Cloning and the Future of Human Embryo Research. Oxford University Press, New York.

Pellegrino, E.D., Schulman, A., Merill, T.W. (Eds.), 2009. Human Dignity and Bioethics. University of Notre Dame Press, Notre Dame.

Pellegrino, E.D., Thomasma, D.C., 1993. The Virtues in Medicine. Oxford University Press, New York.

Post, S.G. (Ed.), 2003. Encyclopedia of Bioethics, 5 vols. Macmillan, New York.

Sandø, P., Christiansen, S.B., 2008. Ethics of Animal Use. Blackwell Publishers, Oxford.

Singer, P., 2011. Practical Ethics. Cambridge University Press, Cambridge.

Smith, D.H., Veatch, R.M., Miller, R.B. (Eds.), 2003. Children, Ethics, and Modern Medicine. Indiana University Press, Bloomington.

Steinbock, B., Norcross, A. (Eds.), 1994. Killing and Letting Die. Fordham University Press, New York.

Stingl, M., 1997. Ethics I (1900–45). In: Canfield, J.V. (Ed.), Routledge History of Philosophy, vol. X. Routledge, London.

Sorell, T. (Ed.), 1998. Health Care, Ethics and Insurance. Routledge, London.

Tooley, M., 1972. Abortion and infanticide. Philosophy and Public Affairs 2 (1), 37–65.

Bioethics: Genetics and Genomics

Ruth Chadwick, Centre for Social Ethics and Policy, The University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The history of bioethics is closely related to that of genetics and genomics. As genetic issues have moved out of the clinical setting and become increasingly concerned with genomics, biobanks, and public health, there has been a move away from the primacy of individualism and autonomy to include a focus on other bioethical principles such as solidarity and the common good.

Introduction – the Relationships between Bioethics and Genomics

The development of bioethics has been closely connected to that of genetics and genomics, at least for parts of their respective histories. The history of bioethics in relation to genetics and genomics has been marked by varying emphases, at different times, on environmental and biological explanations of human health and behavior, and by changing perceptions of the relationship between the individual's interests and the public good. Thinking about the ethical issues in genetics of course preceded the development of bioethics as a field of study. Thus in the early years of the twentieth century, in relation to the sterilization of Carrie Buck, it was claimed "three generations of imbeciles are enough" by the judge in the case of *Buck v Bell*. The history of eugenics in the early twentieth century, and responses to that, had a significant influence on how debates became framed in the second half of the twentieth century, in particular in relation to the emergence of the nondirective ethos in genetic counseling. Reaction against forced sterilization of those deemed 'unfit,' such as Buck, led not only to concerns in succeeding decades about the improper interpretation of scientific information, but also to worries about abuse of genetic science; to an emphasis on free choice in the genetics clinic; and a preference for environmental explanations and interventions at a social level. In the second half of the twentieth century, however, in the light of the success and high profile of genetic and genomic research, genetic explanations and promises regained popularity.

Bioethics came to prominence as a distinct field in the 1970s. Although there are some differences of opinion about its origins, the first use of the term is frequently attributed to Van Rensselaer Potter, who used it to refer to issues relevant to the biosphere as a whole – in other words, including concern for environmental as well as biomedical issues. In its earliest manifestations in the 1970s, however, it was primarily concerned with relationships between health care professionals and individual patients, and with allocation of resources in situations of scarcity, such as the allocation of new technologies developing at the time, including dialysis machines. Reacting against the perceived paternalism prevalent in the professions, and reflecting social changes at the time such as civil rights, feminism, and a liberalization of laws relating to homosexuality and abortion, an individualistic ethic predominated, with autonomy as a central ethical principle. Toward the end of the twentieth century, however, bioethics expanded its scope of concern. This coincided with rapid developments in genomics, and with the initiation and progress of the Human Genome Project (HGP), which allocated a dedicated budget to address ethical, legal, and social issues (ELSI). This was followed by an ELSA programme (ethical, legal, and social *aspects*) in Europe. Although ELSI and ELSA have become recognized as brands in themselves (subsequently succeeded by Responsible Research and Innovation), the broader field of bioethics was also itself greatly affected by the huge explosion in work on the ethical dimensions of genomics. Bioethics, according to the definition of the International Association of Bioethics, examines not only ethical, legal, and social issues but also philosophical ones, and thus includes considerations of the philosophical aspects of identity, for example, which have also been at stake in debates about genomics.

Over time, the issues giving rise to the most scrutiny in bioethics and genomics have changed, as genomics has moved out of the clinic and into society at large. What have also been noticeable are the ways in which scientific developments have posed challenges for ways of thinking about bioethical issues such as informed consent, leading to a considerable amount of work on ethical frameworks (Knoppers and Chadwick, 2005).

Issues in Clinical Genetics: Genetic Testing and Counseling

When the focus of discussion was within the genetics clinic, and concerned with diseases recognized as having an identifiable genetic component – for example, Huntington's disease, the ethical issues turned largely on the autonomy and confidentiality of the client in relation to testing, which may be diagnostic of a current condition or predictive of a late onset one, and relating not only to adults but also TO their children and fetuses. In the clinical context a genetic *test* (as opposed to *screening*) is typically carried out where there is some reason to think there may be a problem. A good example is the situation where there is a strong family history of breast cancer. For adults, it has been argued that there is a right not to know (see below) the result of a predictive test which reveals a risk factor but not certainty about onset or severity of a disease.

Where children are concerned, for parents with a child with an apparent disorder, the result of a *diagnostic* test may bring relief in knowing what is wrong. In the case of prenatal testing,

however, there might be difficult questions concerning termination decisions, in cases where a fetus is found to be suffering from a genetic or chromosomal abnormality (relevant factors also include the context of social support, or lack of it, for persons with disabilities which might affect a mother's decision). Preimplantation genetic diagnosis and embryo selection avoid the issues of termination of a fetus but the issues of the criteria for selection of embryos, resource issues, and what should be done with 'spare' embryos remain controversial.

Nondirectiveness

As mentioned above, following the abuses of genetics in the early twentieth century and the rise in individualism in medical ethics and bioethics, the prevailing ethos in counseling was that of nondirectiveness, facilitating but not directing the decision of the client in relation to decisions such as termination. Nondirectiveness may of course be difficult to achieve, in situations where there is a power differential between counselor and client; and some clients may actually want to be directed or at least advised, but for present purposes the main point is that the ethos reflects the primacy of autonomy.

In some situations, there is a clear tension between the interests of the client and those of third parties. Some of the most discussed questions in the early 1990s concerned possibilities of disclosure to family members of the results of genetic tests on blood relatives, and the conflict between the interests of those family members and the confidentiality of the person tested – for example, is it right to disclose a finding of non-paternity where a genetic test on the fetus reveals that the mother's partner could not actually be the genetic father? The balance between competing interests here is normally struck by the position that individuals should be encouraged to share their information, but that there may be some cases at the extreme, where disclosure is permissible.

Children

The genetic testing and screening of children is acknowledged to be a special case. In relation to children, the issues of autonomy are more complicated, where it might be the parents' wish to have a child tested, because *they* want to know. The possibilities of stigmatization and self-fulfilling prophecy are real issues for a child who is symptom free but identified as being at risk of a late onset disorder. Hence there are strong arguments for the view that children should not be tested for late onset disorders until they are capable of deciding for themselves whether or not to be tested. The situation is different where there are potentially life threatening or life limiting disorders which can be treated in childhood, where the expectation would be that parents should opt for treatment as they would with any other illness.

Genetic Exceptionalism

The fact that blood relatives may share genetic predispositions to particular diseases was one of the principal factors driving discussion of the thesis of 'genetic exceptionalism,' the thesis that there is something special or different about the ethical issues in genetics, the other factors being that genetic information is predictive and independent of time: hence a test in childhood has the potential to reveal an individual's risk status with regard to certain late onset diseases. The thesis of genetic exceptionalism, however, has been criticized on the grounds that other types of medical information may share these features in varying degrees: although the ethical issues in genetics may be particularly complicated, they are not in a wholly different category from those relevant in other areas of medicine, so that there is no need to develop a bioethics specific to genetics per se. Different issues need to be considered on their own merits and in specific contexts.

While the issues in clinical genetics and genetic counseling remain live, they are no longer the primary focus of concern or debate in bioethics. What has changed is that genetic information has become relevant not only in the genetics clinic but throughout medical practice, as the biological factors underlying not only genetic disorders but also the common diseases such as heart disease and the cancers are identified. This is perhaps one of the most marked changes in relation to genetics – the 'geneticization' of health and health care.

Screening

In contrast to genetic testing, where there are factors relating to an individual which suggest that testing is indicated, a screening program is carried out on a population where there is no reason to think, for any given individual, that he or she is at higher risk (Nuffield Council on Bioethics, 2003a). In the 1990s there was considerable discussion about the principles underlying the introduction of a genetic screening program – adapting the 1968 Wilson and Jungner principles, for screening in general, to the particular context of genetic screening. The first criterion is that the condition screened for must be important, or serious; another is that there must be an acceptable treatment – which was changed in the genetic context to 'scope for action,' as the intervention in question may be a termination or change of lifestyle. There was considerable but to some extent inconclusive discussion about what constituted a condition sufficiently 'serious' to warrant a screening program. A classic example is phenylketonuria, for which newborns are screened and which can be controlled by a diet low in phenylalanine.

There have been cases of interventions at the population level to counteract high incidence of genetic disease, for example, in the case of thalassemia in Cyprus. The prevention program changed over time with developing technologies, but in 1980 premarital screening was made compulsory by law. The incidence of thalassemia has declined and patients with the condition have a better life expectancy as treatment improves. This example makes clear, however, that genetics retains its connection with reproduction decisions of those who may be themselves unaffected, but carriers, and is not only concerned with individuals who either have a condition or are diagnosed as being at risk.

Gene Therapy

One of the reasons why issues in genetics have been so high profile lies in the nature of genetic disorder and the options for

treatment. It is in the nature of genetic disorders that they cannot typically be cured in the way that infectious diseases can, for example, by prescribing drugs (notwithstanding the current anxieties about antibiotic resistance). The boundary between genetic and nongenetic disease is not, of course, clear cut: infectious diseases have a genetic component, not only in relation to the genome of the infecting organism but in its interaction with the genome of the persons affected. People will differ in their susceptibility and response to infectious disease in accordance with their own genomic variation.

The early prospects of gene therapy, of introducing a functioning gene into the body of a patient who lacked a functioning one, by using a vector such as a virus, led to disappointment arising partly from the results of tests and partly because of ethical concerns about some trials. The focus of interest turned away somewhat to the new promise of stem cell therapy. In the twenty-first century, however, some forms of gene therapy are seeing encouraging results. In January 2014, for example, positive news was reported in relation to a trial using gene therapy to treat eye disease. There remain concerns about the difference between somatic gene therapy (where the new genetic information affects the individual patient only) and germline therapy (where the germline is also affected). In the latter case, future generations will be affected who have no opportunity to consent. To some, however, the latter argument is overstated, as we cannot avoid decisions that will affect future generations in any case.

Debates about clinical genetic testing, screening programs, and therapy, however, were overtaken by genomics moving center stage with the advent of the HGP.

The HGP and Genomics

The aim of the HGP was to produce a complete map and sequence of all the genes in the human genome, so genomics marked a shift of attention from genes to genomes. Until quite recently most of the bioethical discussion related to genomics concerned the nuclear genome, which was what was referred to in the HGP. The sequencing of the mitochondrial genome, completed in the 1980s, passed almost unattended by any ethical comment. presumably because of how the role of the mitochondrial genome, which exists within the cell outside of the nucleus, was perceived. The analogy of a battery has been used, suggesting that the role of mitochondria is just to provide energy to the cell, in contrast to the nuclear genome which, together with other factors, influences the characteristics of the phenotype.

It is nevertheless the case that a defect in the mitochondria, which are inherited through the female line, can have a devastating effect on the phenotype. The battery analogy, however, has been influential in making the positive case for mitochondrial replacement therapy, which involves removing the diseased mitochondria from a woman's egg and replacing them with those from donor. The egg can then be fertilized by the woman's husband or partner. This has been described in some sections of the press as producing babies with three parents, but the supporters of the technique argue that it prevents potentially very great suffering by the prevention of mitochondrial disease which cannot otherwise be cured.

With regard to the nuclear genome, it was thought by some that to know the genome of an individual would enable us to know everything about that person – that a human being was identified with the sum total of his genome. Genetic determinism, the thesis that a person's genes determine the outcome in the phenotype, had considerable support. It therefore came as a shock to some when the HGP was completed and revealed that there are far fewer genes in the human genome than had previously been thought – about 22 000 rather than 100 000. This led to the claim that genetic determinism was dead. If there are so few genes in the human genome, genes cannot possibly explain the huge variety and complexity that we see in human beings. Genetic *factors*, as opposed to whole genes, are clearly influential in relation to development and phenotypic expression, but so are many other factors, such as gene–environment interaction, copy number variation – variation in the number of copies of a genetic factor that an individual has – and epigenetics, or factors over and above the genome that affect gene expression.

Biobanks

In addition to trying to make sense of the information provided by the outcome of the HGP itself, attention turned in the last decade of the twentieth century to population wide genomic research (Gottweis and Petersen, 2008). Biobanks were established in a number of countries, such as Iceland and Estonia. The purpose of a biobank was to collect both DNA samples and phenotypic information from participants, in an effort to establish links between genetic factors and the development of disease, and not only diseases classifiable as 'genetic diseases.' For example, the UK Biobank aimed to clarify the factors at work in the common diseases that typically manifest themselves in midlife, such as heart disease and the cancers.

Attention turned to the study of variation in the genome – the ways in which individuals differed from each other at the genetic level, leading to effects which may be due to a single base pair in a particular part of the genome. There had been earlier interest in variation, in the Human Genome Diversity project, but that had attracted considerable amount of hostility among different ethnic groups who had concerns about the motivation behind the study of diversity.

Biobanks, particularly in Iceland, were also initially very controversial and gave rise once more to discussions about whether there is anything special about the ethical issues involved. One prominent example concerned informed consent. The doctrine of informed consent had become a cornerstone of biomedical ethics in the twentieth century, following not only the concerns about eugenics mentioned above but also malpractice in relation to research subjects, such as had occurred both in Europe in World War II and in the United States.

Participants asked to donate DNA samples to biobanks were asked to give their informed consent. The problem that emerged was that it was not always clear what they were being asked to consent to. Biobank operators were proposing to collect and store their samples, but for what? It might not be clear even to the scientists and researchers what they might want to do with the samples a few years down the line. New

scientific discoveries would almost certainly reveal new lines of inquiry. Thus it was argued by some that genuine informed consent could not be provided in relation to a donation to a biobank. Again, this did not support a thesis of genetic exceptionalism, because there is a view that fully informed consent is an ideal that can never be fully reached in any branch of medicine. The new context, however, gave rise to the bioethical debate about the difference between 'narrow' and 'broad consent.' A person who gives narrow informed consent agrees to specific, detailed research purposes, whereas broad consent involves agreement to research for purposes that may not be specified or specifiable at the time. The UK Biobank adopted a form of broad consent, asking participants to consent to being in the Biobank – or not.

There are different arguments in favor of broad consent in the biobank genomics context. First, the purposes for which biobanks are established are clearly to some extent undermined if there are restrictions on the research that may be done on the samples. This is not just an argument about scientific freedom: it is argued that this kind of population research is essential for future population health, in addressing the incidence of the common diseases. Another argument is that the kind of harm which may befall contributors to a biobank is very different from that involved in conventional medical research such as a drug trial. The physical intervention involved is limited to the taking of a blood sample: the potential harm arises largely from the potential from access to and misuse of a participant's information, especially if the samples are linked to phenotypic information of various kinds.

As noted above, in the clinical genetics context confidentiality and disclosure had been an issue. Biobank research however gave rise to a considerable amount of new debate about the handling of genomic information, both as regards storage and access. Storage issues included to what extent and in what ways samples were linked, or not, to phenotypic information, by techniques of coding and anonymization. If samples could not be linked at all, then the value of the research was decreased. On the other hand, the greater the possibilities of identifiability, the greater the risk to the privacy of the individual participant. Increasingly it was realized, in any case, that where DNA samples are collected and stored the possibility of maintaining complete privacy is undermined, and it was argued by some that the better course would be to acknowledge this fact. This led to a call for 'open consent' where participants would donate in the knowledge that there was no guarantee of privacy, in the interests of research.

The development of biobank research and population wide genomic research has been associated with a move toward an emphasis on public health in bioethics, marking a widening in scope of this field, although it would not be accurate to say that bioethics had no public health aspect at the beginning, as resource allocation issues constituted a driving force behind its development. It would be a mistake to characterize the issues here as a trade-off between a personal interest in privacy on the one hand and a public interest in health on the other: there is also a public interest in the maintenance of privacy protections, and the balance of interest has to be negotiated.

Nevertheless, given the predominance of the public good argument in relation to biobanks, it is not surprising that there has been a shift in the bioethical framing of the issues. In addition to the arguments for broad consent, there has been a move toward a greater emphasis on principles other than the individualistic ones at play in the earlier days of bioethics, including equity and justice. Another obvious example is the principle of solidarity, which has been appealed to in this context. The principle of solidarity, which has been defined in the report prepared by Barbara Prainsack and Alena Buyx for the Nuffield Council of Bioethics on solidarity (Prainsack and Buyx, 2011) as the willingness to bear costs for another's good, can be used, for example, to provide a reason for contributing to a biobank. Autonomy nevertheless remains important – the individual must be free to give consent, whether broad or narrow, but other principles are increasingly emphasized.

Feedback and the Right (Not) to Know

Much of the discussion over genetic testing and screening concerned an individual's right to access to and control over their genetic data. The ways in which this played out in the clinical context of genetic counseling have already been alluded to.

In the biobank context, these issues took on a particular focus in relation to feedback of results of biobank research. Opposing arguments have been advanced in relation to this on the one hand it is argued that participants should not be given information other than the general results of the research. In many cases, the raw data that emerge will not be meaningful to individuals in any case. On the other hand, it is argued that if information specific to an individual (sometimes called incidental findings) emerges which could be relevant to their health, even potentially life saving, then there is an obligation to tell them. At the time of writing this debate is far from resolved, and there are ongoing debates about what information participants in biobanks and genomic research should have access to.

Debates in the 1990s had already considered whether there could be a right not to know genetic information (Chadwick et al., 2014). The historical predominance of autonomy, choice, and informed consent in the second half of the twentieth century perhaps produced a default position that individuals had a right to information: what had to be justified was withholding it. Different understandings of autonomy, however, suggested that an individual can make an autonomous decision to live their life without certain information which might not only cause distress but undermine their conception of themselves and their life plan.

Personalized Medicine

The potential irony in the situation where principles such as solidarity and concepts of the common good are appealed to in the ethical discussions of population wide genomic research, is that the interest in the variations in the genome between individuals, the discovery of which was an aim of this research, was to drive the quest for a new form of individual-centered medicine, so called personalized medicine, whereby treatment could be prescribed in accordance with the particular genome of the patient.

The argument for personalized medicine made much of the fact that traditional medical practice and drug prescribing had a considerable problem of disease caused by medical examination or treatment, or iatrogenic disease. There was a large burden of not only morbidity but also mortality arising from the side effects of pharmaceutical products. While it was known that any given drug would not be appropriate for every individual, in the past for the most part the only way to find out would be for that individual to take the drug – a trial and error approach predominated in prescribing. The theory behind pharmacogenomics was that it should be possible through genomic research to discover the associations between genetic factors and drug response. Implementation in clinical practice would of course have to have access to genetic information about the individual patient, and this would then inform the choice of drug, or the appropriate dosage. The promise was safer and more effective medicine – and also a reduction in time wasted through trying drugs that did not work for a given patient (Nuffield Council on Bioethics, 2003b).

The label 'personalized medicine' has not proved totally beneficial. First, it is not always clear what 'personalization' means. Initially personalized medicine might have meant assigning patients to groups of good or poor responders to a particular drug on the basis of a targeted test – a fairly limited sense of personalization, which might more appropriately be called patient stratification. Developments in science and technology, from genome-wide association studies to next generation sequencing, have given a new meaning to personalization, ultimately involving investigating the whole genome sequence of an individual. (Some approaches to personalized medicine also include wider ranges of nongenetic data; e.g., NAS, 2011; ESF, 2012). The metaphor of 'tailoring' has been used to describe the fit between the individual's genome sequence and the prescription of medical and lifestyle advice. The cost for such sequencing has also fallen rapidly and continues to fall.

There are also new challenges, however. While on the one hand, it fits into an individualistic paradigm in society that employs the rhetoric of choice, this has also led to one of the main criticisms: that on a wide scale it is only feasible in wealthy societies (despite the falling costs). Indeed, it has been called 'boutique' medicine by Abdallah Daar and Peter A Singer (Daar and Singer, 2005). They claimed, that the underlying research could be beneficial for other purposes – what has been described as 'drug resuscitation' of products taken off the market because of adverse events in the west, to help underserved populations in developing countries, where differences in genomic factors might avoid the adverse effects. This discussion is part of a much wider debate about the sharing of the benefits of genomic research. Benefit sharing, to be understood in a wide sense beyond commercial benefit, has become a standard ethical dimension of proposals for genetic and genomic studies, one that points to the greater consideration of the equity and justice (HUGO, 2000).

Direct-To-Consumer Testing

Following the HGP companies emerged offering testing direct-to-consumers, where the findings could include information about genetic ancestry, and health-related information. As regards the latter they gave rise to many concerns about the bypassing of professional counseling and advice in the delivery of the information, as well as criticisms regarding the accuracy and usefulness of the information provided. Whereas some people regarded it as an aspect of the 'democratization' of science to facilitate this kind of accessing of their genomic information, there were worries about the possible adverse effects of individuals acting on misleading information about themselves and their families. In December 2013 in the USA, for example, the Food and Drug Administration ordered the company 23andMe to halt its disclosure of health-relevant results to new direct-to-consumer testing clients.

'Recreational' and Behavioral Genomics

While the main argument for genetic testing, screening, and now sequencing has always been the medical benefits, other uses have increasingly become apparent, some of which are regarded as recreational. Ancestry tracing has already been mentioned. Behavioral and forensic genomics raise ethical issues about responsibility, freedom, and punishment – could there be a legitimate defense in court relating to one's genetic predisposition, for example. Also giving rise to concern are the promotion of tests related to character and natural talents of children, and to pick out budding sports stars so that they can be given appropriate training. In some cases this overlaps with medical matters, as in the case of a predisposition to sudden cardiac death which is clearly important for someone considering a sporting career. The main issues here, however, remain the quality of the scientific basis of the information on offer and the tension between the marketplace and the professions as the medium through which genetic information should be conveyed or bought and sold.

An area which also crosses boundaries is that of nutrigenomics. At first sight it seems analogous to pharmacogenomics. Just as the latter facilitates personalized prescribing on the basis of genetic information, so the former, in principle, makes possible personalized dietary advice which can also be extremely important in the health care setting, especially in the light of a purported obesity epidemic. However, nutrigenomics is very different from pharmacogenomics in that foodstuffs, with multiple ingredients, are far more complex and work on more aspects of the body than do drugs, which are designed to act on specific targets and pathways in the body. Establishing the associations on the basis of which useful advice can be given is therefore more challenging.

Genomics and Identity

The ethical debates about genetics and genomics have raised identity issues in a number of ways. First, there has been considerable controversy about what counts as a disorder. This is of course a live issue even absent the genetic element: mental disorders are a case in point, and testing for predisposition to such disorders is particularly sensitive. In the genetics clinic the question of what counts as a disorder has been a particular issue because of the fact that termination was one potential outcome of prenatal genetic counseling, giving the impression

that the birth of babies with disabilities was regarded as an outcome to be avoided. Disability rights groups objected to the identity and worth of people being reduced to a disorder, questioned the very label of disorder, and feared the return of a form of eugenics. Arguments on the other side attempted to draw a distinction between the conditions and the people who experienced them: between discrimination against a disorder and discrimination against persons with disabilities, with varying degrees of success.

For a given individual or community, the results of genomic research can disturb long-held beliefs about their ancestry or relatedness, leading to considerable distress and upset in some well-publicized cases, such as that of research on the Havasupai Indians, who gave consent to the use of their samples in research on diabetes, which was a serious problem in their community. However, they subsequently learned that their samples had been used to study other things, and that publications on the results contradicted their history as they saw it and a lawsuit led to the payment of significant compensation.

Another identity issue concerns that of the human species as a whole. Although the majority of the ethical writing concerns human genomics, the genomes of other species have been, and continue to be sequenced, with results about the number of genes in some of those species that have been surprising to some. Information about both the complexity of the genomes of other species and their degree of relatedness to the human has the potential to undermine some long-held beliefs about the identity of the human species in relation to those others.

International Perspectives and Cultural Issues

International issues in genomics may take different forms – first there are questions both ethical and practical about international research and the transfer of samples and data across borders – issues of ownership are at stake here in both subnational research projects and national initiatives. Whereas biobanks were initially set up in different countries such as Iceland and Estonia, as mentioned above, arguments for linking between different national efforts appeal to the need for large amounts of data to underpin statistical validity.

The second type of international issue, which is related to the first, concerns cultural differences in various parts of the world where not only ethical values but also the cultural meaning of genetic material may be different. In these situations ethical principles may at least need to be revisited, and informed consent may need to be approached in different ways. Public engagement has become increasingly important, partly as a necessary condition of a process of benefit sharing. The Human Heredity and Health in Africa initiative (H3Africa), for example, which includes a number of different genomics projects funded by the Wellcome Trust and the National Institutes of Health in the US, is addressing issues of public engagement and informed consent in multiple populations in different social contexts, both rural and urban.

In terms of international approaches to the issues in genomics, these include declarations such as the Universal Declaration on the Human Genome and Human Rights issued by UNESCO (UNESCO, 1997), which states in Article 1 "The human genome underlies the fundamental unity of all members of the human family, as well as the recognition of their inherent dignity and diversity. In a symbolic sense, it is the heritage of humanity." While this Declaration lays down universal principles, other international instruments incorporate rules. At the level of the European Union, for example, the proposed new Regulation on Data Protection lays down parameters for transfer of data across the Union and beyond.

Conclusion

Despite the close interaction between bioethics and genomics, and the ways in which bioethical principles have been subject to reinterpretation in the light of the new challenges, issues are far from settled. The ways in which bioethics has been done have also changed during the period, to include much more empirical research and reflection on modes of multidisciplinary working. The rapid advances in technology and prospects for implementation offer the promise of considerable development in the relationship between science and bioethical debate in the years to come.

See also: Anthropology, Genomics, and Human Variation: National Roots; Biobanking: Ethical Issues; Bioethics in the Post-genomic Era; Bioethics: Philosophical Aspects; Direct-to-Consumer Personal Genetic Testing; Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute; Eugenics as a Basis of Population Policy; Eugenics in Asia; Eugenics in Europe; Eugenics in the Americas; Eugenics, History of; Genetic Ancestry Testing; Genetic Screening for Disease; Genetics and Indigenous Communities: Ethical Issues; Genetics and Social Justice; Genetics and Society; Genetics and the Media; Genetics: Legal Aspects; Genetics: The New Genetics; Genographic Project; Genomics, Ethical Issues in; Genomics, Ethical Issues in; Human Genome Project: History and Assessment; Human Genomics in Asia; New Genetics and Race.

Bibliography

Chadwick, R., Levitt, M., Shickle, D. (Eds.), 2014. The Right to Know and the Right Not to Know. Cambridge University Press, Cambridge.

Daar, A.S., Singer, P.A., 2005. Pharmacogenetics and geographical ancestry: implications for drug development and global health. Nature Reviews Genetics 6, 241–246.

European Science Foundation (ESF), 2012. Personalised Medicine for the European Citizen – towards More Precise Medicine for the Diagnosis, Treatment and Prevention of Disease. ESF, Strasbourg.

Gottweis, H., Petersen, A. (Eds.), 2008. Biobanks: Governance in Comparative Perspective. Routledge, London.

Human Genome Organisation, 2000. Statement on Benefit-sharing. HUGO.

Knoppers, B.M., Chadwick, R., 2005. Human genetic research: emerging trends in ethics. Nature Reviews Genetics 6, 75–79.

Nuffield Council on Bioethics, 2003a. Genetic Screening: Ethical Issues. Nuffield Council on Bioethics, London.

Nuffield Council on Bioethics, 2003b. Pharmacogenetics: Ethical Issues. Nuffield Council on Bioethics, London.

Prainsack, B., Buyx, A., 2011. Solidarity: Reflections on an Emerging Concept in Bioethics. Nuffield Council on Bioethics, London.

UNESCO, 1997. Universal Declaration on the Human Genome and Human Rights.

[US] National Academy of Sciences (NAS), 2011. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. NAS, Washington, DC.

Bioethics in the Post-genomic Era

Alena Buyx, University College London, London, UK; and University of Münster, Münster, Germany
Barbara Prainsack, King's College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Bioethics is a rapidly developing field. It draws on conceptual and methodological tools from a number of academic disciplines to tackle ethical issues that arise within medical practice, the life sciences, health care provision and management, and related areas. From its early inception in the form of medical ethics, where it was mainly concerned with moral conflicts that played out within the relationship between doctors and patients, it has evolved into a multidisciplinary field that utilizes a wealth of theoretical and methodological approaches to address a variety of moral and practical questions. This article focuses on the most recent developments of bioethics in the so-called postgenomic age.

Introduction

Bioethics is a rapidly developing field. It draws on conceptual and methodological tools from a number of academic disciplines, including philosophy, sociology, theology, jurisprudence, psychology, and others, to tackle ethical issues that arise within medical practice, the life sciences, health care provision and management, and related areas. From its early inception in the form of medical ethics, where it was mainly concerned with moral conflicts that played out within the relationship between doctors and patients, it has evolved into a multidisciplinary field that utilizes a wealth of theoretical and methodological approaches to address a variety of moral and practical questions. These range from purely theoretical philosophical work, to various types of empirical studies, to endeavors to influence and shape practical policy making. This article will focus on the most recent developments of bioethics in the so-called post-genomic age; for a description of the field of bioethics as such, its history, and the development of its scope throughout the twentieth century, see the cross-references at the end of this article.

There is no coherent use of the term 'postgenomics.' One common use refers to a cluster of research areas that were spin-offs of the Human Genome Project (HGP). In the late 1990s, the US-led yet international HGP set out to identify all genes on the human genome, to determine the sequence of the base pairs that make up human DNA, and to render this information available to the international research community. A draft sequence of the first human genome was published with much media fanfare at the dawn of the new millennium.

Another use of the term reflects some of the insights following the HGP and the subsequent changes to the field of genetics. Although the human genome sequence published initially was incomplete and riddled with problems and open questions that scientists are still struggling to solve, the inherent promise of the completion was that humans could now understand the basic units of life and would soon be able to develop new disease treatments. Many of these expectations have turned out to be unrealistic. While the HGP has doubtlessly triggered an unprecedented advance in the development of high-throughput tools in molecular biology, the capacities for interpreting these data and for making sense of what they mean for health and disease lag behind. One of the biggest lessons learned in the aftermath of the HGP was that the information that can be obtained from DNA sequence information alone is seriously limited (Maher, 2008). Researchers have therefore started to move beyond studying the sequence of nucleotides on the human genome (structural genomics) toward understanding their function (functional genomics), gene expression under various physiological states (transcriptomics), and modifications of the genome (epigenomics). They also look at processes within other areas of human biology, and how they intersect with our DNA. For example, researchers in proteomics study the structure and function of proteins, those studying metabolomics focus on substances and processes involved in metabolism, others look at the interaction between proteins and other molecules ('interacteomics'), and others study the interactions of humans with their gut microbial flora ('microbiomics'). The suffix '-omics' signifies that substances and processes are not studied individually, but at a large scale, often using mass spectrometry tools. For example, in proteomics, although the function of proteins and their interaction with other molecules may be looked at in particular individuals, the aim of the field is to learn about the role of proteins in human organisms in general.

The exact meaning of the term 'postgenomics' also depends on the discipline and institutional context of its use. In its widest meaning, the term 'postgenomics' comprises endeavors to understand the function of DNA in the context of other molecules; in this understanding, anything that moves beyond mapping the structure of human genomes is included. This means that all '-omics' disciplines, including the use of environmental (e.g., phenotypical and lifestyle related) data to refine and complement biological '-omic' data, are included in this concept. In this article, the term 'postgenomics' is used in this wide sense.

Changes in Methodological Approaches

In its early days, bioethics relied heavily on using theories and arguments from the fields of theology, anthropology, law, and, in particular, philosophy. Indeed, for a long time, many viewed bioethics as a subdiscipline of philosophy that applied philosophical theories such as utilitarianism, virtue ethics, deontological theory, or a set of midrange principles to bioethical

questions (often called the 'applied ethics approach'). Until the early 2000s, the methodology chapters of many bioethics textbooks focused on these and a few other philosophical approaches (such as feminist philosophy, political philosophy, or care ethics theory). Other theories and methodological approaches were referred to only briefly, if at all. This has changed over the last two decades. With the discipline growing in size as well as in academic and political relevance, it began to draw in people from more diverse academic backgrounds, which in turn has contributed to the inclusion of a wider range of concepts and theories. The following section will give a brief overview of the most prominent expansions and additions to bioethics' theoretical toolbox.

At least for the last decade, the social sciences have played an important role in bioethics (Fox and Swazey, 2008). With the growing need for both hard and soft law and policy and regulation of bioethical issues that lie outside of clinical medicine, the legal, social, and political sciences have increasingly contributed important theoretical and empirical work to bioethics. They have done so both by theorizing governance and regulation and through the development of practical policy proposals using methodological tools from their respective theoretical background, such as the analysis of cases and judgments (Mason and Laurie, 2010), or policy and contextual analysis (Gottweis and Prainsack, 2006; Moreno, 2012). Moreover, psychology has informed recent bioethical work more explicitly than in the past, for example on coping with illness, on the role of carers, or within neuroethical debates (see, e.g., the journal *Philosophy, Psychiatry, & Psychology*). Economics has also contributed to the growing body of bioethical theory. Health economists have long been in dialogue with ethicists about issues of allocating health care resources, such as the construction, application, and consequences of cost–benefit analysis (see, e.g., Nord, 1999). More recently, behavioral economics – itself drawing on psychological research – has introduced the concepts of 'choice architecture' and 'nudging' as factors that powerfully contribute to shaping our decision making (Thaler and Sunstein, 2008), which have been taken up vigorously, for example, in public health ethics (see, e.g., the UK's Centre for the Study of Incentives in Health).

Due to their disciplinary origins, these and other recent theoretical contributions to bioethics bring with them particular terminologies and traditions, and this can sometimes lead to translational misunderstandings. The multi- and interdisciplinary dialogue is not always successful. However, the integration of different academic perspectives has yielded many substantive and innovative results. At its best, bioethics is now a discipline that has a rich arsenal of both well-established and innovative theories at its disposal that, if applied well and appropriately, can generate genuinely novel theoretical insights and inform practical solutions for complex ethical challenges.

Partly due to the contribution of a broader range of disciplines to bioethics, and partly in response to an increasing discontent within the discipline, bioethics has accommodated empirical research at an ever-growing rate over the last two decades. Many bioethicists as well as outside commentators – often those involved in practical decision making, such as clinicians or policy makers – have voiced dissatisfaction with the foundational approach of applied ethics as lacking a recognition of the facts and dilemmas of the 'real world.' In parallel, the growing influence of the evidence-based paradigm in medicine stimulated interest in empirical studies aimed to inform bioethical reasoning (Borry et al., 2005). Many clinical ethicists became engaged in empirical research due to their strong integration in medical settings. In addition, the aforementioned growing contributions from disciplines such as psychology, sociology, and political science have led to a stronger incorporation of qualitative and quantitative methodology into bioethics. Last but not least, some of the debates that emerged over the last two decades on issues such as the social determinants of health, brain mapping, and systematic reviews of reasons have been underpinned by discussions of empirical methodology and/or data, and have required in-depth knowledge of the empirical evidence.

Robust knowledge of the empirical evidence is now regarded as essential for the formulation of good normative arguments that are subsequently applied in practice, for example, on questions such as: What shapes the needs and desires of people who are dying? What values, goals, and risks do women and couples consider when deciding whether or not to donate their unused *in vitro* fertilization (IVF) embryos for research? Nevertheless, the relationship between empirical and normative perspectives in bioethics has not always been smooth (Hurst, 2010; Goldenberg, 2005). This is largely due to the difficulty of determining how empirical data could and should appropriately inform normative reasoning without violating the fact–value distinction, displacing sound normative argument, or producing trivial insights that could have easily been gained by other means. Empirical researchers, on the other hand, have complained that bioethics has incorporated empirical research without also fully taking on board the rigorous quality standards to which those designing empirical studies are held within their respective disciplines. There is no question that bioethics as a field has greatly benefitted from the inclusion of empirical data in its core debates. Examples of well-designed studies abound that inform normative thought in an appropriate and illuminating way. However, the discussions between the two perspectives on how to marry normative reasoning and empirical research are far from over.

Both described changes to bioethics' methodology can also be regarded as part of a trend toward a more professionalized discipline with rigorous research standards and tools for evaluating the quality and impact of resulting publications. By the end of the twentieth century, the first big push toward professionalization was more or less over, at least in the developed world, with research ethics committees, national ethics councils, large bioethics research projects, and devoted institutions and journals firmly in place in most Western countries, and increasingly in the developing world as well. This trend is continuing; the postgenomic era of bioethics also marks the time when bioethics has arrived in the scientific mainstream insofar as its output is now also judged, in most countries and university systems, by the same methods as other research – that is, the impact factors of the journals where they are published; how influential they have been in the clinical, societal, and policy-making domains; and so on. This has to do, in part, with changes in how academic research is

assessed in general, and also with the interrelation between science and bioethics, for example in the large ELSI (ethical, legal, and social implications) projects around the time of, and following, the HGP, as described below.

Emerging Subfields, Topics, and Debates

Beyond Gen-Ethics

From the beginning of the HGP, bioethics has played a particularly important role. When the HGP was set up, it was decided that a particular percentage of the public funding involved (3–5%) would need to be dedicated to the study of the ethical, legal, and social aspects/implications (ELSA, aka ELSI) of genomics. This, and subsequent similar obligations by, for example, the European Commission, have given rise to the establishment of a multi- and interdisciplinary yet distinct field studying such aspects. There can be no doubt that the large ELSA projects of the last two decades have contributed important findings and supported bioethics' higher visibility in scientific research and with the public. However, the ELSA develop-ment has also spurred considerable critique. ELSA researchers in genomics and postgenomics have been criticized for playing the 'handmaidens' of scientists by granting legitimacy to scientific advances (e.g., Zwart and Nelis, 2009). From another direction, ELSA research has been criticized for doing the opposite, namely, for overemphasizing potential pro-blems and thus hampering research that could potentially be lifesaving for people (e.g., Taylor, 2008). These controversies are bound to continue.

As mentioned in this article, bioethics in the postgenomic era is increasingly addressing challenges in areas of bioscience and biomedicine that integrate knowledge and practices from more than one discipline. For example, the field of personalized medicine aims to use data on people's biological and environmental (including social and lifestyle-related) characteristics in order to render the classification, understanding, treatment, and prevention of disease more precise (i.e., more 'personalized') (ESF, 2012). Its remit includes, at least potentially, every area of biomedical research and every medical specialty. Its success will rely on the participation of patients and citizens who, for example, collect and provide data for their own diagnosis and treatment, and often also release it for research use. The involvement of the commercial sector in the development and provision of personalized medicine is seen by some as a desirable goal, and by others as a necessary evil; everybody agrees, however, that commercial companies will play an important role here. Such strong integration of scientific and medical disciplines and specialties, with actors from public and private domains, presents bioethics with complex challenges, as decision making increasingly takes place in a decentralized and networked manner. Moreover, understanding and tracing the implications of decisions have become very difficult endeavors. The result is that bioethicists can say very little about issues in integrated fields such as personalized medicine *in abstract terms* (other than reiterating some concerns lying at the core of the discipline, such as justice, equity, and the prevention of harm). Unlike in most other areas of bioethical inquiry, bioethical scholarship in these integrated areas tends to address particular concrete practices (such as pharmacogenomics testing of a particular patient group in a specific institutional and national setting).

Other examples of growing domains of bioscientific and biomedical practice that integrate a multitude of different disciplines include epigenetics, which merges molecular biology expertise with the need to understand the influence of environmental factors on the genetic dimension, and synthetic biology. This latter field applies an engineering approach to building biological systems.

What these integrated fields have in common is that they are underpinned by large-scale databases and repositories. These databases and repositories – depending on the purpose for which they are used – comprise annotated biological and other kinds of data (epidemiological, lifestyle related, technical characterizations, etc.). Being seen by many scholars as key infrastructures of the life sciences in our times, their establishment and governance pose complex problems in their own right (Kaye et al., 2012). Core challenges include the need to balance the protection of privacy and confidentiality of personal data, on the one hand, and the quest for ever larger, better characterized databases that could yield new insights when adequately mined, on the other.

The question of when to inform people of incidental findings based on their genetic and genomic data is another hotly debated issue. Some scholars also critically question the considerable resources needed for the establishment and maintenance of large-scale databases that, if employed for other means, could lead to faster health outcomes and/or address the pressing health needs of larger parts of the world's population.

Neuroethics

An important newer subfield of bioethics that emerged in the postgenomic age focuses on the ethical, legal, and social dimensions of neuroscientific research. The increasing attention that bioethicists pay to neuroscience was triggered by the new possibilities and scenarios that emerge with the advance of new imaging techniques (particularly magnetic resonance imaging (MRI)). Neuroimaging is said to provide 'glimpses into the workings of the brain' through 'brain mapping.' New developments in neuroengineering, cognitive enhancement, neurosurgery, and so on have also been important drivers of bioethical interest in this field. The interface of neuroscience and bioethics is now known as 'neuroethics' and has attracted considerable funding in its own right (e.g., the National Core for Neuroethics, University of British Columbia, Canada, and the Oxford Center for Neuroethics, United Kingdom). Simultaneously, it has been called into question whether contemporary neuroscience poses any genuinely new bioethical challenges to warrant so much attention. Scholars who have put forward this argument hold that most of the ethical issues addressed within neuroethics are in fact *not* specific to applications of neuroscience and have been taken up previously, for example in relation to the ethics of genetics (Illes and Racine, 2005). Notwithstanding such questions, the usage of 'neuroethics' to describe a subdiscipline of bioethics is now firmly established, not least through the foundation of research institutions bearing it in their name as well as a journal of the same name (*Neuroethics*, published by

Springer). And regardless of where one stands on the question of whether neuroethics tackles any genuinely novel issues, one of the benefits of this young area in bioethics is that it showcases the integration of expertise emerging from interdisciplinary work that is initiated and led either by ethicists and social scientists, or by neuroscientists and policy makers. Many of the now-classic articles on ethical issues arising from developments in neuroscience have not been written by ethicists, but by neurologists, neuropsychologists, or neurosurgeons drawing on bioethical concepts and methodologies (e.g., Farah, 2002).

Neuroethics is often broadly divided into two categories – 'the ethics of neuroscience and the neuroscience of ethics' (Roskies, 2002) – with obvious overlap between the two. The latter examines the neural bases of moral reasoning, for example by trying to establish whether there are two different systems for moral judgments at work in human decision making, one 'utilitarian' and one 'deontological' (Greene et al., 2004). What are the neural bases of agency, responsibility, and free will, and could these concepts be illusions? Some scholars also subsume the examination of the neural correlates of decision making in general under the label 'neuroscience of ethics,' whether in economic or consumer choice settings, or in relation to love, religion, or sex. The former area – the ethics of neuroscience – addresses, among other things, the potential and actual consequences of interventions in the brain through a variety of methods, such as psychopharmacological drugs, psychosurgery, tissue engineering, and brain–computer interfaces. Discussions focus, for example, on questions such as whether it is ethically appropriate to use drugs to enhance cognition and mood beyond what is considered 'normal,' whether and what brain interventions that change personality are permissible, and what effects neuroscientific interventions and research have on society (Glannon, 2011).

Population-Level Bioethics, Public Health Ethics, and International Bioethics

Arguably the biggest newer development in bioethics is the inclusion of a population perspective. Much of traditional medical and bioethics focused on ethical issues to do with individuals and their rights and obligations – patients, doctors, transplant and stem cell donors, and so on – or with the relationship between individuals and society, present in discussions such as on the rationing of health care or the regulation of reproductive technology. As recently as 2007, however, the field and the profession could still be accused of ignoring the many ethical issues arising from a consideration of population-level data and problems (Brock and Wikler, 2007). This is currently changing; population-level bioethics is growing at high speed.

One of the most important initial drivers of this was an increased awareness of the population-level health data that epidemiology and other public health disciplines had been amassing over the last half-century. The analysis of these data has made it increasingly obvious that there are, for example, significant socially determined health inequalities along the whole socioeconomic gradient within countries, and even greater ones between the populations of different countries (e.g., Marmot, 2004). These findings, replicated multiple times and across the whole globe, have sparked an intense discussion around questions such as which socially determined health inequalities are unjust, and how these can be addressed or compensated for (Powers and Faden, 2006; Daniels, 2008). Many bioethical arguments in these discussions have borrowed from philosophical debates about distributive justice and egalitarianism, invigorating these in the process, as well as from the more technical discussions in public health about the role of the social determinants of health and well-being, thus constituting truly interdisciplinary and often innovative work with impact on national policy making and international institutions (e.g., the World Health Organization (WHO)).

Population-level bioethics has also brought some new impetus to the long-standing controversy around the behavioral bases of health. The contested questions of whether some health inequalities that are linked to behavior should be compensated for or whether they fall under the personal responsibility of those engaging in these behaviors to some degree mirror the older debate about whether rationing should take personal health behavior into account (Buyx and Prainsack, 2012; Prainsack and Buyx, 2011). Population-level data on the social determinants of health have challenged some of the old arguments in favor of a stronger recognition and expectation of personal responsibility. On the other hand, the necessity to positively influence both socially determined health behavior and other socially determined inequalities has become ever clearer. An interesting contribution to how such influence could be wielded on a policy level has been brought to the field from behavioral economics (Thaler and Sunstein, 2008), and a lot of recent work in population-level bioethics and public health ethics focuses on the permissibility and form of 'nudges' based on the idea of 'libertarian paternalism.'

In parallel with the stronger focus on population-level health data, an awareness of the international dimensions of bioethics has also grown. This was and is particularly the case, again, with health inequalities, as these are most starkly obvious between the developed world and poorer countries, raising the question of the richer world's obligations to address them. Other topics that international, population-level bioethics addresses are – to name just a few – the 'brain drain' of health professionals from poor to rich countries, research priorities in developing countries, and other ethical issues in international research (e.g., Eyal and Hurst, 2008).

To a large degree, population-level bioethics and another newer bioethics subfield – public health ethics – overlap thematically. In addition to those already mentioned in this article, they also focus on topics such as pandemic preparedness and vaccinations. Public health ethics addresses some additional issues that are more specific to the discipline, such as the ethical appropriateness of particular public health programs. There are also a number of topics that have recently risen to prominence in bioethics that tackle even wider, international, and interstate ethical concerns, such as the effects of climate change, the consequences of new approaches to bioenergy, or the ever-present yet worsening challenges of food crises (e.g., Buyx and Tait, 2011; Nuffield Council on Bioethics, 2011). Due to

their complexity and relevance, these latter issues can be expected to receive even more attention in years to come.

The Dual-Use Problem

An issue that receives increasing attention in bioethics in the postgenomic era is the problem of dual use. 'Dual use' means that science (and applications) developed for civil purposes can also be put to other (e.g., military or even terrorist) uses. This dilemma is by no means new; it has been a bone of contention within policy and public debates in connection with nuclear science, behavioral genetics, and recombinant DNA research, for example. Bioethicists, however, have only more recently turned toward discussing dual-use problems in the life sciences (e.g., Ehni, 2008). A 2008 paper in *Bioethics* identified five obligations that should be met by scientists working in areas affected by dual-use issues (Kuhlau et al., 2008): these include engaging in response activities, considering the negative implications of research, and not sharing information. Especially the latter obligation – not to share or publish sensitive information – has been at the heart of a controversy in connection with papers reporting on experiments with avian H5N1 influenza virus transmission in ferrets in 2012 (Herfst et al., 2012; Imai et al., 2012). When both manuscripts were reviewed, The US National Science Advisory Board for Biosecurity (NSABB) suggested that only the general conclusions, but not the methodological details, should be published to prevent anyone from replicating these experiments with bad intentions (some even suggested that the research itself should be halted). Eventually, both papers were published, in a modified form, and both advisory committees and scientists agreed that oversight of such research should be intensified (Yong, 2012).

Dual-use considerations have also been very prominent in the ethical debate about synthetic biology (see, e.g., Douglas and Savulescu, 2010).

Human Enhancement

Medicine has been used to improve or enhance 'normal' human function or features for a very long time. Doping in sports and cosmetic surgery are well-established examples of this. However, over the last two decades, developments in medical technology and increasing demands to be 'better than well' have sparked interest in what has been summarily termed 'enhancement.' Examples range from the use of growth hormone in children with a predicted small height absent of any pathology, to hormone treatment to delay signs of aging, to mood, cognitive, and personality enhancement afforded through the use of psychopharmacological drugs and brain stimulation (e.g., Parens, 1998).

While the evidence of 'net' enhancing properties of medical treatments is, to some degree, still contested, a lively discussion has developed around a number of theoretical and ethical issues. The distinctions between concepts such as treatment and enhancement, illness and health, normality and pathology, and naturalness and artifice have been greatly contested, while others have focused on the ethical permissibility – or, indeed, the moral obligation – of such enhancements, particularly in minors; their potential effects on equality, society, and our understanding of what makes us human; if and how the medical profession should respond to requests for enhancement; and what policy options are available to address these issues (e.g., Savulescu and Bostrom, 2009). The debate, very active in the early 2000s, has recently quieted down slightly but is, nonetheless, still relevant.

Vision for the Future

Without any claim to comprehensiveness, these are some of the more recent trends that have had considerable impact on the way postgenomic bioethics has developed. One feature that is apparent in all of them is the increasingly fruitful interdisciplinary integration that has taken place.

While these debates and subfields can be expected to grow still further, there are of course a number of issues that have been of long-standing interest to bioethics, such as ethical questions at the beginning of life and in end-of-life care, as well as other areas of clinical ethics; the stubborn ethical challenges around policies for transplantation and donation of organs, gametes, tissue, and so on; ethical issues in research; questions of fair rationing in modern health care systems; ethical provision of modern reproductive services; and appropriate regulation of genetically modified organisms and related research activity, to mention just some of the most contentious ones (for detailed discussions of these, see the cross-references below). Bioethics, benefitting from the changes and developments that have been laid out in this article, will continue to seek answers and solutions to these knotty problems.

See also: Bioethics: Genetics and Genomics; Bioethics: Philosophical Aspects; Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute; Genetic Engineering; Genomics, Ethical Issues In; Human Genome Project: History and Assessment.

Bibliography

Borry, P., Schotsmans, P., Dierickx, K., 2005. The birth of the empirical turn in bioethics. Bioethics 19 (1), 49–71.

Brock, D.W., Wikler, D., 2007. Population-level bioethics: mapping a new agenda. In: Dawson, A., Verweij, M. (Eds.), Ethics, Prevention and Public Health. Oxford University Press, Oxford, pp. 78–94.

Buyx, A., Prainsack, B., 2012. Lifestyle-related diseases and individual responsibility through the prism of solidarity. Clinical Ethics 7, 79–85.

Buyx, A., Tait, J., 2011. Ethical framework for biofuels. Science 332 (6029), 540–541.

Daniels, N., 2008. Just Health. Cambridge University Press, Cambridge.

Douglas, T., Savulescu, J., 2010. Synthetic biology and the ethics of knowledge. Journal of Medical Ethics 36 (11), 687–693.

Ehni, H.J., 2008. Dual use and the ethical responsibility of scientists. Archivum Immunologiae et Therapiae Experimentalis 56 (3), 147–152.

European Science Foundation (ESF), 2012. Personalised Medicine for the European Citizen – Towards More Precise Medicine for the Diagnosis, Treatment and Prevention of Disease. Strasbourg.

Eyal, N., Hurst, S., 2008. Physician brain drain – can nothing be done? Public Health Ethics 1 (2), 180–192.

Farah, M.J., 2002. Emerging ethical issues in neuroscience. Nature Neuroscience 5 (11), 1123–1129.

Fox, R., Swazey, J., 2008. Observing Bioethics. Oxford University Press, Oxford.

Glannon, W., 2011. Brain, Body, and Mind: Neuroethics with a Human Face. Oxford University Press, Oxford.
Goldenberg, M.J., 2005. Evidence-based ethics? On evidence-based practice and the 'empirical turn' from normative bioethics. BMC Medical Ethics 6 (11), E1–E9.
Gottweis, H., Prainsack, B., 2006. Emotion in political discourse: contrasting approaches to stem cell governance: the US, UK, Israel, and Germany. Regenerative Medicine 1 (6), 823–829.
Greene, J.D., Nystro, L.E., Engell, A.D., Darley, J.M., Cohen, J.D., 2004. The neural bases of cognitive conflict and control in moral judgment. Neuron 44, 389–400.
Herfst, S., Schrauwen, E.J.A., Linster, M., Chutinimitkul, S., de Wit, E., Munster, V.J., Sorrell, E.M., Bestebroer, T.M., Burke, D.F., Smith, D.J., Rimmelzwaan, G.F., Osterhaus, A.D.M.E., Fouchier, R.A.M., 2012. Airborne transmission of influenza A/H5N1 virus between ferrets. Science 336 (6088), 1534–1541.
Hurst, S., 2010. What 'empirical turn in bioethics'? Bioethics 24 (8), 439–444.
Illes, J., Racine, E., 2005. Imaging or imagining? A neuroethics challenge informed by genetics. The American Journal of Bioethics 5 (2), 5–18.
Imai, M., Watanabe, T., Hatta, M., Das, S.C., Ozawa, M., Shinya, D., Zhong, G., Hanson, A., Katsura, H., Watanabe, S., Li, C., Kawakami, E., Yamada, S., Kiso, M., Suzuki, Y., Maher, E.A., Neumann, G., Kawaoka, Y., 2012. Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. Nature 486, 420–428.
Kaye, J., Gibbons, S.M., Heeney, C., Parker, M., Smart, A., 2012. Governing Biobanks: Understanding the Interplay between Law and Practice. Hart Publishing, Oxford.
Kuhlau, F., Eriksson, S., Evers, K., Höglund, A.T., 2008. Taking due care: moral obligations in dual use research. Bioethics 22 (9), 447–487.
Maher, B., 2008. The case of missing heritability. Nature 456 (7218), 18–21.
Marmot, M., 2004. The Status Syndrome. Owl Books, New York.
Mason, J.K., Laurie, G.T., 2010. Law and Medical Ethics, eighth ed. Oxford University Press, Oxford.
Moreno, J.D. (Ed.), 2012. Progress in Bioethics: Science, Policy, and Politics. MIT Press, Cambridge, MA.
Nord, E., 1999. Cost-Value Analysis in Health Care: Making Sense Out of QALYS. Cambridge University Press, Cambridge.
Nuffield Council on Bioethics, 2011. Biofuels: Ethical Issues. London.
Parens, E. (Ed.), 1998. Enhancing Human Traits: Ethical and Social Implications. Georgetown University Press, Washington, DC.
Powers, M., Faden, R., 2006. Social Justice: The Moral Foundations of Public Health and Health Policy. Oxford University Press, New York.
Prainsack, B., Buyx, A., 2011. Solidarity: Reflections on an Emerging Concept in Bioethics. Nuffield Council on Bioethics, London.
Roskies, A., 2002. Neuroethics for the new millennium. Neuron 35, 21–23.
Savulescu, J., Bostrom, N. (Eds.), 2009. Human Enhancement. Oxford University Press, Oxford.
Taylor, P., 2008. Personal genomes: when consent gets in the way. Nature 456, 32–33.
Thaler, H.T., Sunstein, C.R., 2008. Nudge: Improving Decisions about Health, Wealth, and Happiness. Yale University Press, New Haven and London.
Yong, E., 2012. Mutant-flu paper published: controversial study shows how dangerous forms of avian influenza could evolve in the wild. Nature 485, 13–14.
Zwart, H., Nelis, A., 2009. What is ELSA genomics? EMBO Reports 10 (6), 540–544.

Bioethics: Philosophical Aspects

Richard E Ashcroft, School of Law, Queen Mary University of London, London, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by U. Schuklenk, volume 2, pp. 1195–1201, © 2001, Elsevier Ltd.

Abstract

This article describes the relationship between philosophy and bioethics. It begins by defining bioethics. It then explains how philosophy became the central academic discipline within bioethics, and discusses the relationship between philosophy and other disciplines in bioethics. The article then discusses the main contributing theories within philosophical bioethics, and concludes by briefly discussing the developing range of topics within bioethics.

Definition

The definition of bioethics is somewhat controversial. Different definitions are proposed, which focus on different accounts of its scope and of the methods used in its arguments and inquiries. However, for most practical purposes, a broad and inclusive definition is appropriate. We may define bioethics as the "field of normative problems and issues arising from biomedicine and medical sciences and their applications." This very broad definition is intentionally silent on the question of whether bioethics is a discipline, or even an emerging profession, and on the question of whether bioethics includes a concern with animal welfare, environmental ethics, climate change, or issues of law and social policy which have an impact on health without being directly related to the practice of medicine. Some writers in the field of bioethics would see gun control or sexual violence as being public health issues, and therefore claim that the ethical questions associated with these issues, and with interventions addressed to these issues, as falling within bioethics.

Philosophical bioethics is much more easily defined. It is the application of the methods, concepts, and theories of philosophy to the problems of bioethics. There are two ways of looking at this: one is to see philosophical bioethics as one part of bioethics, and not by any means the whole of bioethics. The other is to see philosophical bioethics as one part of a more general enterprise of 'applied philosophy.' Not all of applied philosophy is concerned with bioethics: indeed, not all of applied philosophy is concerned with ethics (or its sibling, normative political philosophy).

Philosophy and Bioethics

The relationship of philosophy to bioethics is complex. Philosophers have often written about issues that might now be considered characteristically 'bioethical' in nature: suicide is a clear instance of a topic of central interest to bioethics, which has been at the heart of philosophical thought and debate. Nevertheless, it would be absurd to argue that Plato or Hume were 'doing bioethics' when they wrote about the rationality or ethics of suicide. The description was unavailable to them; their conception of what they were doing and their reasons for doing it were quite different from the conceptions of modern 'bioethicists'; their intellectual, cultural, and religious environments were quite different and so on. If we consult the standard 'internalist' histories of bioethics, we often find the following historical train of thought. Classical philosophy from Socrates up to the nineteenth-century moral philosophers (including Friedrich Nietzsche, Auguste Comte, and Henry Sidgwick) was essentially concerned with how to live well. Problems of ethics were therefore central to philosophy. Once philosophy came to be dominated by logic, epistemology, and the philosophy of language (from Gottlob Frege and Edmund Husserl onward), ethics was displaced from its central role within philosophy. To a greater or lesser extent, not only did ethics stop being a prestigious 'research program' (to borrow Imre Lakatos's term from the philosophy of science) but also it started to become an object of outright skepticism within philosophy. By the mid-twentieth century the dominant paradigms for moral philosophy were the existentialist school, which emphasized authenticity and commitment to a project but eschewed any concrete moral judgments, and the post-positivist 'emotivist' school, which argued that the language of ethics had no objective content and was merely rhetorical in nature. There remained an active research program in moral *theology*, especially within the Roman Catholic Church, but philosophy, newly professionalized in the early twentieth century, took great pains to distinguish itself from theology.

Although, as we shall see, philosophy began to take ethics very much more seriously again from the 1960s onward, it remains true of most professional philosophers, and most philosophy departments as institutions, that although some parts of ethics – mainly those concerned with theoretical problems of mind and language – are 'proper' philosophy, most of what nonphilosophers think of as ethics is 'merely' … something; that something might be 'applications,' 'journalism,' 'critical thinking,' or 'public engagement.' To the extent that ethics (be it labeled bioethics, applied ethics, or practical ethics) is taken seriously, it is as a source of test cases. For instance, there is a growing literature on the ethics of human-induced climate change, and a good deal of this literature is concerned with fixing technical problems in the theory of utilitarianism, or with addressing the question 'why should I be moral,' with policy or practical recommendations coming into focus almost as a secondary consideration. This is not necessarily a criticism of philosophy or philosophers: it is a description of the nature of academic professional philosophy at this

point in its disciplinary history. Moreover, it can hardly be a criticism of philosophy to note that philosophers tend to work on problems that they recognize as distinctively philosophical, and to use the methods and paradigms with which they are most familiar and skilled. That said, an influential paper in the 1970s surveyed the field of philosophical *medical* ethics and concluded that medicine had 'saved the life of ethics' (and, he might have said, philosophy itself, Toulmin, 1982). It did this by throwing up a whole series of problems which were – to a philosopher – recognizably philosophical and intellectually challenging, and which forced moral philosophers to think more deeply both about these problems and about the moral theories they were using.

Some philosopher bioethicists have mistakenly taken this to be the whole story of bioethics. This is especially evident when one looks at the literature from the late 1980s onward arguing that the social sciences should play a greater role in bioethics. Philosophers critical of this position often argued as if philosophy was what was distinctive of bioethics, as if it were philosophy, which had saved the life of medicine. In part this reflects the then current perception that philosophy (at least in its Anglo-American versions) was mainly concerned with language and the clarification of concepts, so that the contributions of the social sciences would at most be of 'data,' which would be of rather limited use or relevance in analytical argument (a useful critique can be found in Hedgecoe, 2004.)

It is certainly true that philosophy made, and makes, a great contribution to bioethics. We will see this later in this article. But the attempt to claim bioethics for philosophy is quixotic. Historians and sociologists of bioethics as a social movement and as an intellectual enterprise have consistently shown that the philosophers were relatively late entrants to the field: the emerging field of bioethics was shaped to a far greater extent by theologians, lawyers, and scientists and medical professionals. Nevertheless it is true that philosophers came to play a major part in the shaping of bioethics, and its questions, methods, intellectual standards, quasi-professional entry criteria, and so on continue to be dominated by philosophy. One exception to this general rule is where bioethics has been taken up as a 'clinical' specialty (clinical bioethics), where the influence of the methods and norms of academic medicine is much more pronounced.

Philosophy and Social Problems in Bioethics: Introducing Principles

One way to account for the dominance of philosophy within bioethics is to point to the superficially obvious fact that in as much as bioethics is about moral problems, philosophy is the intellectual discipline that focuses centrally on morality and its structure. The difficulty with sustaining this line of thought is that other disciplines also lay claim to this territory. If we frame 'moral problems' slightly differently, as social problems, then it would appear that sociology and social policy might be the dominant disciplines (particularly given the concentration of mid-twentieth-century US sociology on 'social problems'). Equally, we might turn to history (in Germany, the history of medicine occupied a much more central role in thinking about 'bioethics' than did philosophy), theology, or law. Indeed, bioethics has periodically drawn on all these sources. A better explanation of the domination of philosophy in bioethics is needed.

Sociologist John H. Evans has argued that the role of philosophy in bioethics became central because the issues at the heart of bioethics were initially posed using language and concepts that were 'thick.' Thick concepts are concepts that combine description and evaluation. A good commonplace example is the concept of rudeness: if I say someone is rude I am both stating a fact which I believe to be true, and evaluating their conduct and character, and I am inviting you to agree with both the evaluation and the judgment of fact. Moreover, you cannot have the moral judgment without the factual judgment, or vice versa. The difficulty in complex societies is that thick concepts rest on background assumptions about social norms and values that may not be universally understood or shared. Evans argues that theological bioethics depends on a strong way on thick concepts (not only doctrinal requirements about articles of faith or religious observance, but also on thick conceptions of 'the nature and value of humanity' or 'human dignity'), in ways which make theological arguments very vulnerable to challenge in pluralistic societies. In its place, he argues, bioethics came to require methods of argument that did not strongly depend on thick concepts, and developed methods of consensus formation. Moreover, the social force of such consensus had to depend, or be seen to depend, on the authority of something other than custom, practice, doctrine, or tradition (which could be identified as the values of some specific social subgroup), but rather on a style of reasoning and a set of concepts that could be identified as common across cultural boundaries. Philosophy was seen to offer these tools. In particular, the emphasis in philosophy of scrutinizing the form and presuppositions of argument could be seen as immunized against the inbuilt moral judgments of 'common sense' – those very 'thick concepts' that were taken to be the problem in the first place (Evans, 2012).

It should be noted that this vision of philosophy as the means to overcome value conflict in pluralist societies involves a very specific kind of philosophy. In the second part of this article we will review the main philosophical methodologies applied in philosophical bioethics, but here we need to make the acquaintance of the most well known: the method of principles. This is most closely associated with the philosopher Thomas Beauchamp and theologian James Childress. Beauchamp was a staff philosopher supporting the preparation of the 'Belmont Report,' the report of the (US) National Commission on Protection of Subjects of Biomedical and Behavioral Research (published in National Commission (1978)). The National Commission had been established to report to the US President on the protection of research subjects in the aftermath of the revelations of long-term exploitation of African-American men who were participants in the 'Tuskegee Syphilis Study' from the early 1930s until 1972. The intellectual method of the Commission was to identify basic principles that underpinned the ethics of human subjects research, which could both account for the various concrete rules found in law and policy, guide formulation of new rules to cover situations currently not provided for, and assist in resolving differences in moral judgment. During the preparation of the Commission report, Childress joined Beauchamp in writing what became

Principles of Biomedical Ethics (first published in 1978, now on its seventh edition, 2013). While the Belmont report identified three principles (respect for persons, beneficence, and justice), Beauchamp and Childress identify four (respect of autonomy, beneficence, nonmaleficence, and justice) together with a process of 'specification' that ties principles to the context of application.

A key feature of the 'principlist' methodology is that it is not, and never was intended to be, a *sui generis* comprehensive moral theory. Rather, it begins from 'common morality,' the ordinary, everyday moral maxims that are taken to be more or less universal in social life, around which there is a rough consensus. The role of principles is to provide an organizing structure, or 'mid-range moral theory' (akin to sociologist Robert K. Merton's account of a 'mid-range sociological theory'), which both explains and justifies the lower level common moral doctrines, and enables these to be tested, broadened, and applied in novel or controversial contexts. The roots of this approach are various, drawing on moral philosopher W.D. Ross's theory of duties as prima facie principles, John Rawls's theory of 'reflective equilibrium,' American philosophical pragmatism, among others. But the crucial feature of this approach to thinking about bioethics is its emphasis on consensus formation as a method of overcoming moral conflict. In this regard, it is quite different from both piecemeal moral theories that argue from cases (in much the way common law reasoning works in Common Law legal systems), and critical moral theories that seek rational reconstructions of morality which overcome what proponents see as defects in common morality.

The 'principlist' approach in bioethics continues to be paradigmatic, and it is certainly the most widely taught method in bioethics. In part, this is because it is easily summarized, and can be taught without needing much – if any – prior background in philosophy. As such it is particularly appealing in medical and health care education and training, where bioethics is taught as a 'clinical skill,' often over the course of just a handful of classes, and often where the assessment may be in the form of automatically-marked multiple choice questions (or similar). The form in which the principlist approach is taught, however, neglects all the subtleties of Beauchamp and Childress's own writing, and all the patient attention to common morality and its detail. Instead, the focus tends to be on the Four Principles themselves, treated as a list and without much attention to how they interrelate. The crudest versions emphasize 'respect for autonomy' to such an extent that the other three principles are either neglected altogether or treated as mere restatements of the autonomy principle. Critics of bioethics commonly note that justice almost always seems to disappear from discussion, or is treated in narrow terms as concerning economic distributive justice. One way out of this set of criticisms is to restate the basic ideas of Beauchamp and Childress with new, alternative principles. This is often done where the original principles are taken to be 'American' or 'Anglo-Saxon' in nature. Versions of 'European' principles have been stated, for instance, where it is unclear whether they are meant to be better principles, or simply local principles. The universalism of principlist theory is something its proponents do insist on, yet it is at least arguable that both in terms of its basis in common morality and in its application to specific contexts it remains much more 'culture-bound' than it would at first appear. Equally, the criticism that it is falsely universalist, and insufficiently focused on culture, structure, power, and agency is also leveled at principlist bioethics (for discussion, see Macklin, 1999). The problem here is that these criticisms can be leveled at any philosophical theory, and indeed at philosophy itself, and the best answer to these criticisms must be philosophy's own reflexive self-awareness and self-criticism of its presuppositions and conditions of possibility.

Philosophical Methods in Bioethics

We introduced principlism to explain how philosophy came to take up a central role in bioethics as a method for resolving value conflicts in public discussion and policy making. Another feature of principlism is the way it generated an efficient method for generating policy in bioethics. An important feature here is the nature of philosophical expertise. At least in the early years of philosophical bioethics, the expertise of philosophy was seen as on the one hand its lack of doctrinal presuppositions (in contrast with law or theology) and on the other its methodological rigor (in terms of commitments to analytical clarity and logical consistency). As with other emergent disciplines, bioethics came to have its own doctrines, canons of literature, methodological rules, and so on. A good example here is the way bioethics took up the doctrine of consent from law, but has developed that doctrine in ways that now mean that when bioethicists and lawyers speak to one another about consent they often do not understand one another. Yet the overall image of philosophy as providing the 'view from nowhere' (in Thomas Nagel's memorable phrase), dispassionate, analytic, and logical persists. From the point of view of intellectual history, however, this image links philosophy in bioethics to a very specific time and place, namely to mid-century Anglo-Saxon analytical philosophy (Nagel, 1986). This remains the dominant tradition within philosophical bioethics, although, as we will see, there are other important traditions which are of growing importance.

Conceptual analysis plays an important role within philosophical bioethics in its own right. The idea is that before we can carry out normative reasoning, we must be clear about the meaning and scope of the terms under discussion. A central example in discussions of end-of-life care, for example, is that we must understand correctly what 'killing' means. Are withdrawing life support, failing to offer life-sustaining treatment in the first place, and administering a poison all correctly identified as 'killing'? We can, presumably, get clear on this question without taking up a position on whether killing is ever morally permitted or even required. Related to this, can we clearly distinguish an act from an omission? Philosophers often proceed on the basis that we have ways of identifying the meaning of linguistic terms in terms of concepts, and ways of clarifying our understanding of concepts, which are normative for their proper use, and contrast with the ways words are actually used by individuals or groups (for instance, by lawyers in a legal system). Not all philosophers think this is a sensible or even intelligible way of proceeding – for example, Ludwig Wittgenstein, who argues that we should look for meaning in use, or Gilles Deleuze, who argues that philosophy is about

creating, not finding, concepts. Nevertheless, the analytic move is often a useful one in trying to set out what the nature of the debate is, even just as the first, and not the last, word.

Normative moral philosophy is the branch of philosophy which considers theories of right action. It is related to *metaethics*, which considers the metaphysical and epistemological basis of rightness and the good. In metaethics, we are concerned with the questions "Why should I be moral?" and "Is anything objectively good?" Bioethics rarely engages directly with metaethics, although bioethics and bioethicists rarely subscribe to radical skepticism in metaethics (versions of which include egoism, subjectivism, and relativism). Occasionally, metaethical issues do become salient to bioethics, sometimes in connection with the status of bioethics itself (see the discussion of the universality or otherwise of principles, above), sometimes more concretely in discussions about the nature of 'quality of life' or the value of beginning or ending a human life. Here, for reasons of space, we concentrate on normative ethics.

The main doctrines in normative ethics are *utilitarianism*, *Kantianism*, *communitarianism*, *virtue ethics*, and *casuistry*. Each of these theories has a distinctive history within moral philosophy that predates the appearance of bioethics and continues to evolve alongside and sometimes in dialogue with it. Before we see brief descriptions of each, we need to consider the nature of theory in normative ethics and its implications for bioethics in practice. The relationship between bioethics and normative ethics runs two ways. From normative ethics to bioethics, a writer who is a defender of a particular theory in normative ethics may well draw a conclusion relating to a topic in bioethics, and propose it as the right thing to do. And from bioethics to normative ethics, a writer who thinks that the proffered conclusion is obviously wrong, wicked or absurd, will argue that this shows a flaw in the normative theory in question. In principle, such disputes can be resolved by seeking to optimize coherence between moral intuitions, robustly supported conclusions in analogous cases, evidence about people's moral beliefs, and logical consistency in theoretical construction to come to a conclusion about where the (moral) truth lies. In practice, these debates rarely run so smoothly or conclude within any finite time horizon. This is one reason why debates come down so quickly to debates about personal autonomy: it is often argued that where consensus on a moral matter cannot be arrived it, then it will be oppressive to impose a moral view on someone who does not hold it, unless to prevent avoidable harm to nonconsenting or vulnerable others. However, even this strategy is not beyond criticism, linked closely as it is to the political doctrine of (Millian) liberalism, in itself frequently criticized. Another way in which theory and bioethics come together is rather different. One way in which a bioethical proposal can be tested is to see whether it can be demonstrated as holding true even when we start with different theoretical starting points. Thus, if Kantians, utilitarians, and communitarians all reach the same conclusion, then we have some confidence in saying that the proposition is true *independently* of the theoretical route we took to get there. (Obviously, the argument is more complicated than that, but this is the idea). The difficulty here is in avoiding 'begging the question' (i.e., implicitly or explicitly assuming what we are setting out to prove).

Utilitarianism is arguably the most influential normative theory in bioethics other than principlism. It is the best known version of 'consequentialism,' which is the theory that actions are to be judged good or bad with reference only to the goodness or badness of their consequences. There are various reasons for this. One is that it is easily stated: act so that the action you perform promotes the most good, compared with all the other actions you could perform in the circumstances. Another is that utilitarianism has a natural representation in classical economics, so that in problems involving the allocation of resources, the utilitarian answer and the economists' answer will be strikingly similar. A third is that utilitarianism often produces strikingly interesting, not to say controversial, answers. Outside the principlist mainstream, the most influential work in bioethics was that done in the 1970s by Australian philosopher Peter Singer, a well-known utilitarian, whose papers on animal welfare, euthanasia, and the permissibility of infanticide of severely disabled newborns were both remarkably clear and troubling to the 'common morality' of the majority. Utilitarianism is the best example of a 'critical moral theory,' which uses clearly stated premises and logical argument to challenge widespread assumptions – both by positive advocacy of alternative views and by critique of the coherence of the mainstream assumptions. Utilitarianism has been criticized within and beyond philosophy on a number of grounds. The first is the way it generates unpalatable conclusions; utilitarians often respond by saying that it is insufficient to say that you dislike their conclusions, you must argue why they are wrong, and they propose a critique of the role of 'intuitions' in moral argument to overcome objections. Unfortunately in most moral arguments, in the end, intuitions are all we have. Another problem in utilitarianism is specifying the nature of the good we are to promote and how we are to measure and aggregate it. While the philosophical literature is very rich in discussions of these issues, this can make practical applications hard to justify. To take an example, in an area where utilitarianism is naturally strong (distribution of resources), it is often argued that distribution should take account not only of economic efficiency but also of equity (a variety of fairness). Utilitarians, in theory, can deal with this by allowing for how much equity matters to individuals (or, how unhappy does an unfair distribution make us) to offset against the welfare or pleasure conferred by the distribution independent of equity considerations. But in practice this is very difficult to do. Hence, the apparent clarity of utilitarianism (an easy principle that can be applied simply) seems to melt away, and undermine its value as a decision tool that guides action.

Kantianism is a shorthand for a variety of duty-based theories of ethics, which contrast with consequentialist theories. The most well known is the theory of ethics developed by Immanuel Kant in the late eighteenth century. The details of how duties are identified and justified are somewhat technical; in practice, the interesting question is not so much how we identify duties – principlism gives one answer and may be considered a Kantian theory – as how we reconcile conflicts or uncertainties governing their application. One form in which duty-based theories of ethics are relevant in bioethics is in the form of codes of medical ethics, which specify lists of obligations and prohibitions and which are intended to be followed by professionals. In one sense this is outside philosophy, in that they have a quasi-legal status, but in debates about their reform and extension philosophical argument can come into

prominence. A good example is the Declaration of Helsinki that governs the ethics of research in medical research as its major international statement. This has been revised several times in recent years, always with attendant controversy, largely because of its attempt to grapple with specifying the duties researchers have to research subjects and to communities where they carry out research, specifically in avoiding unethical exploitation. Unlike utilitarianism, such duty-based theories as are applied to this kind of problem lack clear decision rules that account for solutions.

Virtue ethics has its roots in ancient philosophy (notably Aristotle and St Thomas Aquinas), but in its modern forms responds both to the challenges of consequentialist and duty-based theories noted above and to the importance common morality places on motivation and intention in ethics. By focusing on the moral character of the agent, rather than on his or her actions, it is hoped that we can describe the kind of personal characteristics that make for right action. Virtue ethics is often linked to *communitarianism* that focuses on the characteristics of good societies (rather than good personalities) but shares a common focus on traits, rather than actions, and dispositions to behave in certain ways. It is also notable that traits found praiseworthy in virtue ethics (such as courage, integrity, and honesty) are examples of the 'thick concepts' we mentioned above, and which depend on a social context for both their definition and for the criteria by which they are judged. Both virtue ethics and communitarianism are strongly critical of the 'rationalism' they perceive in both duty-based ethics and utilitarianism, and are often (although not always) associated with forms of political conservatism. Critics of virtue ethics and of communitarianism focus on this perceived conservatism, as also on their lack of focus on actions and rules, which they allege make for relativism and uncertainty.

Casuistry identifies the weakness of all the above theories as lying in their attempts to be excessively abstract and insensitive to the specifics of situations. Casuistry focuses on the moral issues in specific cases, and seeks to resolve them by carefully identifying the cases that are relevantly similar, inferring a rule that covers all these cases (but not the cases which are in relevant respects different) and deducing from that rule what is morally required in the case at hand. To some extent all moral reasoning involves an element of casuistry (in defining and describing the situation of moral agency well enough to allow the moral problem to be identified), but casuistry avoids theoretical argument, instead proceeding in a quasi-inductive manner. Casuistry has a connection to virtue ethics and communitarianism in the form of narrative ethics, which seeks to analyze moral problems in the form of story structures that seek to build coherent narratives, which 'make sense' of the situation. Virtues map onto character traits that make sense of the characters in the story as seeking to tell coherent stories about themselves, which are consistent with our concepts of good lives (and good stories).

Other Traditions in Philosophical Bioethics

Bioethics has become thoroughly internationalized in the past 30 years, and while one of the factors that has enabled it to travel is the dominance of the English language (and hence, up to a point, English-language ways of doing things like philosophy), the influence of other philosophical traditions is increasingly important. Traditions in European philosophy (notably phenomenology and hermeneutics), focusing on meaning and experience, rather than conceptual analysis and normative moral theory, have been particularly influential in work in ethics in mental health and – especially – in work that seeks to elaborate the nuances of complex relationships of care. Unsurprisingly, perhaps, this approach in philosophy is closer in spirit to certain kinds of qualitative social inquiry, and is one area where philosophical bioethics works most closely with the social sciences. This kind of philosophical engagement with health and illness is sometimes considered to be external to bioethics, in part because of its caution about making explicit moral judgments and emphasis on understanding, and in Europe there is a strong tradition of 'philosophy of medicine and health care,' which draws on hermeneutics and philosophy of science as much as on moral theory.

Another complementary field to philosophical bioethics is theological bioethics; philosophy tends to resist strong engagement with theology in this domain, in part because of its social history as a discipline that seeks to avoid close identification with any particular moral, social, or political group in the name of wide social consensus, in part because of a strong tradition of secularism in bioethics, and in part because of the politics of bioethics.

Because of the social controversies that give rise to many prominent issues in bioethics (embryo research, *in vitro* fertilization, euthanasia, and human enhancement, to name but a few examples), social authority to speak on these topics is often contested. Bioethics is faced with a paradox of seeking to be 'beyond politics' and at the same time central to an intensely political field. Religious affiliation is just one dimension of this: to one party 'Catholic bioethics' may simply be 'bioethics, done rightly'; to another it may be 'not bioethics at all, but merely the doctrine of some group I don't care for.' To some extent this is overheated: utilitarian bioethics is a respected subfield, and individuals who work on it can move, often seamlessly, between contexts in which this is what they are publishing on to situations where they are expected to be more 'ecumenical' and back. The same can apply to Catholic, or Marxist, or feminist bioethics. On the other hand, it is not so simple as saying "there is no such thing as hyphenated bioethics, only good bioethics and bad bioethics." The flourishing field of feminist bioethics starts from an explicit premise that is both ethical (women are equal in dignity and importance and value to men) and political (in order to overcome patriarchal oppression of women and the silencing of women in public and private discourse, we should privilege women's voices and women's issues in bioethics as a strategic matter). It is not a question of leaving one's politics (or faith, or conscience) at the door.

The role of politics in bioethics has become more important not only because of its social context, but also as a focus for bioethical research. One illustration of this is the rise to prominence in recent years of 'public health ethics,' which focuses on health policy, resource allocation, population health, and – increasingly – global justice. In this focus it is inevitably engaged with questions of social structure, power, and collective choice, and as such it has increasingly

acknowledged the importance of normative political philosophy alongside normative moral philosophy.

Conclusion: The Topics of Bioethics

There is not space here to examine any bioethical topics in detail. Bioethicists work on such a wide range of topics within health, medicine, and biomedical research that it is impossible to do that breadth justice. However, in broad terms we can distinguish the following subfields of bioethics: clinical bioethics, which focuses on the ethics of clinical practice in medical and nursing care of patients (usually in hospital settings, but not always); public health ethics (with a major emphasis on the ethics of preventive medicine, health promotion, infectious disease control, and resource allocation in health systems); so-called 'regulatory bioethics' (which focuses on the normative principles that govern the regulation and supervision of specific technologies in medicine, including genetic technologies, applications of neuroscience, and assisted conception, and also on the normative underpinnings of health law as applied to abortion and euthanasia). One thing that all these fields have in common, it may be argued, is that there is a more or less prevalent tendency to treat these topics ahistorically and with limited attention to social context. Bioethics is commonly criticized for a kind of socio-technical determinism, particularly when engaged in prospective regulation: if we permit cloning, then this will happen … In a sense this is the 'original sin' of bioethics. If we ask philosophers to think about the moral implications of technology, then it is unsurprising if they do exactly that, rather than challenging the assumption that it is the technology, which will make the difference. On the other hand, there are encouraging signs that the development of bioethics, in closer contact with the social sciences and wider humanities, will continue to enrich debate and – with due humility and reflexivity – enhance it.

See also: Biobanking: Ethical Issues; Bioethics in the Post-genomic Era; Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute; Genetics and Indigenous Communities: Ethical Issues; Genetics: Legal Aspects; Genomics, Ethical Issues in.

Bibliography

Ashcroft, R.E., Dawson, A., Draper, H., McMillan, J.R., 2007. Principles of Health Care Ethics, second ed. John Wiley, Chichester.

Beauchamp, T.L., Childress, J.F., 2013. Principles of Biomedical Ethics, seventh ed. Oxford University Press, New York.

Evans, J.H., 2012. The History and Future of Bioethics: A Sociological View. Oxford University Press, New York.

Hedgecoe, A., 2004. Critical bioethics: beyond the social science critique of applied ethics. Bioethics 18, 120–143.

Macklin, R., 1999. Against Reglativism: Cultural Diversity and the Search for Ethical Universals in Medicine. Oxford University Press, New York.

Nagel, T., 1986. The View from Nowhere. Oxford University Press, Oxford.

National Commission on Protection of Subjects of Biomedical and Behavioral Research, 1978. Ethical Principles and Guidelines for the Protection of Human Subjects of Research. United States Department of Health, Education and Welfare, Washington, DC. Available at. http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html (last visited 10.16.14.).

Toulmin, S., 1982. How medicine saved the life of ethics. Perspectives in Biology and Medicine 25, 736–750.

Biography and Society

Roswitha Breckner, University of Vienna, Institute of Sociology, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Bertaux, volume 4, pp. 1210–1213, © 2001, Elsevier Ltd.

Abstract

Within sociology, biographical research has expanded during the last 20 years. The common ground is still the interest in understanding the interrelation between societal developments (including those on a global scale) and the experiential world of people. How experiences are organized in specific social contexts in the perspective of a lifetime and its generational relations to get oriented in everyday life and in the world, is the core topic in biographical research. This article gives an insight into recent methodological developments in biographical research and its variety of contributions to major issues of globalizing societies, such as history and memory; collective identities; migration and transnational biographies; gender and intersectionality; and the search for new concepts of inclusion.

Since the last article on 'Biography and Society' to this encyclopedia written by Daniel Bertaux (Bertaux, 2001), the founder of the ISA Research Committee 38 on Biography and Society, much has changed in the field of sociologically oriented biographical research. In his article Bertaux pointed at the obstacles that sociological–biographical research had to face in its long history, which started in the 1920s with the major investigation by Thomas and Znaniecki on *The Polish Peasant in Europe and America* (Thomas and Znaniecki, 1918–1920/1923) and the famous studies from the Chicago School (see references in Bertaux, 2001). After flourishing in the 1930s and 1940s in the US and a little longer in Poland, research based on biographical material subsequently was pushed to the background of mainstream sociology. It was only in the 1980s that it was revived again, still against considerable criticism and skepticism. In contrast to the period in which this approach was marginal, it has undergone dynamic development during the last 20 years. The most remarkable development is that biographical research has turned into a success story. It has continued to be broadly institutionalized, and nowadays there is more biography in sociology than ever before. This shows, for example, in the emphasis on situated knowledge brought into theoretical discussion. The early dictum of Martin Kohli to talk about ourselves because we as scholars are biographically shaped, has become an indispensable part of the teaching of classical sociological theory, through the introduction of the biographies of respected classical theorists. Furthermore, a constantly growing body of scholarly literature on a huge variety of topics investigated through biographical perspectives and methodological approaches, appears worldwide in many different languages (ISA RC 38 Newsletters, 2008–2013). These wide-ranging theoretical and methodological approaches were presented in extended compilations (Miller, 2005; Harrison, 2009). Last but not least, there is a growing area of research by PhD students and senior scholars that succeeds, though still rarely, in obtaining funding on a large scale, be it by the EU (Apitzsch and Kontos, 2008; Chamberlayne et al., 2002; Miller and Day, 2012), or by national funding institutions (Rosenthal et al., 2013).

Even though the field is growing and diversifying internationally so that it is almost impossible to cover it in one short article, there is still common ground. The core question (originally put up by Thomas and Znaniecki), which focuses on how changes in society are connected with varying organizations of the self and the processes of reorganizing family, work life, and generational relations as well as institutional settings, is still the basic orientation of biographical research. In short, one of the main purposes of biographical research still is to understand how processes of personal and societal change are intertwined and inseparable. The assumption is that 'making sense' and orienting our actions in the modern era increasingly requires reference to the construction of biographies and institutionalized life courses. People are more and more organizing their experiences and general patterns of orientation in biographical frameworks; other institutional contexts based on strong normative patterns such as 'family,' 'nation,' 'class' and their interrelation, have been changing more or less profoundly (Fischer-Rosenthal, 2000; Kohli, 2007). The social concept of 'biography' therefore is regarded as a relevant field of interference between institutionalized societal reality and the experiential world of individuals in modern societies. It is assumed that the construction of biographies is a social practice in which the societal and the personal are closely interrelated. These processes of making sense are especially intense in periods of radical social change, when institutionalized ways of interpreting and orienting in social situations and the world are called into question and new organizations must be built. In general, 'biography' is regarded as a field in which social processes of continuation and transformation constantly take place. To explore the implicit as well as explicit practices of 'doing biography' in different social fields is the general purpose of analysis, in order to understand society from within, which also means understanding it from the perspective of the actors.

To illustrate the specific contribution of a biographical perspective for understanding actual developments in present societies and also with regard to their historical backdrop, some research areas are presented in more detail. Their selective choice is based on the topics, which were and still are, prominently discussed in the context of the ISA Research Committee

38 on Biography and Society. First, developments in methodology and methods are briefly outlined. Secondly, some insights in scholarly contributions to the research on major topics of globalizing societies are provided, e.g., from the area of biographical research on history, memory, and politics; on collective identities; on migration and transnational biographies; on gender and intersectionality; and on social inclusion.

Developments in Methodology and Methods

There is not just a single biographical approach, but a variety focusing on different aspects, e.g., on the life course (Kohli, 2007); biographical trajectories (Schütze and Riemann, 1990); turning points (Hackstaff et al., 2012); relations between the lived life and the life story (Rosenthal, 1993); gendered biographies (Davis and Lutz, 2002); generational transmission (Inowlocki, 1993; Rosenthal, 2010); specific strands of experience like the work life (Brose, 1989); health careers (Bell, 2009; Frank, 2002; Ettorre, 2010); or different life phases like childhood, adolescence, and older age (see citations in Kohli, 2007). Likewise, theoretical concepts as well as methodological and methodical paths to access the complex evolvement of experience and action patterns in specific local and global contexts have diversified. One strand, which is related to phenomenology, the sociology of knowledge, and partly also to critical theory, is mainly based in narration theory (Bruner, 1987; Kohler Riessman, 2008; Rosenthal, 1993; Schütze, 2008; Frank, 2010) focusing on how a story is told and what this can reveal about patterns of action and experience and their structural meaning in a biography. The narrative form itself is regarded as a means to organize experiences in a temporal and thematic order that makes sense of what we have lived through, and which explicitly as well as implicitly, guide our actions. Biographies in this perspective are not understood as mirrors or copies of lived lives but as narrations produced in a current situation, relating to past experiences, the present situation, and also future expectations in communication with others. The experiences and processes, along which orientations and patterns of action develop and change in relation to specific social contexts, are reconstructed mainly by means of in-depth and comparative analyses of individual cases. Another strand, based more in materialist and structural theories, relies not so much on single-case analyses of biographical narrations, but on a broader ethnographic approach aiming at reconstructing the contexts in which narratives take shape. Recurrent participant observation over a longer period of time, the integration of a variety of empirical material aimed at creating an overall picture of a specific social context, are here the methods of choice. Last not least, scholars in this area also follow theoretical assumptions and criticism of discourse analysis, questioning the concepts of the 'individual,' the 'personal,' or 'subjectivity' in general, which in part was sometimes naively emphasized in biographical research, but not throughout theoretical discussions (Stanley and Temple, 2008).

Given lively discussion between different schools of biographical research, further advancement of methodologies and methods is vital. Narrative approaches are combined with ethnographic research (Delcroix and Pape, 2010) as well as with discourse analysis (Völter et al., 2005), and new approaches are created by linking to developments in other research areas. One of the most intriguing lines is to take the corporal and emotional experiences in processes constituting social life seriously (Plummer, 2001) also taking in feminist theory, which has pointed at this dimension already for some time (Davis, 2007). The study of the biographical significance of the body and especially passion (aspects which are not easy to grasp from the sociological point of view) is new ground for methodological and theoretical reflection (Davis, 2014b). The incorporation of visual sociology into biographical research (Bell, 2010; Breckner, 2013; Dussel and Gutiérrez, 2006) also provides new impulses. The importance of the visual dimension in actual interaction (which increasingly is organized via the social media) and the creation of (new) images of ourselves and of families, generations, and milieus has just started to be explored from a genuine biographical perspective. Its potential contribution to the rapidly growing field of research on the use of globalizing media and its impact on social life is to provide case-based, in-depth analysis on the biographical significance of visual self-presentation intertwined with oral and written communication. Theoretically promising are explorations how patterns of a 'narrated self' interact with patterns of a 'showing self,' and whether the increasing dominance of the visual dimension in everyday life in general is changing biographical concepts.

Globalizing Societies from a Biographical Perspective

Apart from promising developments in methodology and methods, which always have been and still are, closely connected to the field of qualitative and interpretive methodology, scholars are dealing with topics highly relevant in globalizing societies. In a brief overview the author mentions just a few to give an impression of what has been extensively investigated and published during the last 15 years, with focus on the major fields of migration and transnational biographies.

History, Memory, and Politics

Together with oral History, biographical research became prominent mainly in the 1990s and 2000s in revealing how World War II, and especially the Holocaust, affected and shaped the lives of millions of people, making them perpetrators or victims, or both (Inowlocki, 2000; Kazmierska, 2012; Rosenthal, 2010). The aftermath of the still not wholly graspable atrocities became accessible to subsequent generations through the life narratives of previous generations and examination of the processes of generational transmission. In this line, the creation of memory and, more generally, the reshaping of the past from a present perspective in conjunction with the complexity of lived lives, collective narratives, and discourses are points of interest. The knowledge gathered in this research area, revealing the close interrelation of biographical processes, collective trajectories, and major historical upheaval (Schütze, 1992), became fundamental. It expanded through the researching of the crash of the Soviet system and the consequent restructuring

of East European societies after 1989 (Humphrey et al., 2003; Breckner et al., 2000; Bertaux et al., 2004; Semenova, 2002; Andrews, 2007), as well as the end of the Apartheid regime and its aftermath in South Africa (Coetzee, 2011; Stanley, 2006; Coetzee et al., 2004). Against this backdrop the experiential background of political trajectories and actions has become a more broadly discussed topic in recent biographical research. In the German context, for example, the situation of two societies, the GDR and FRG, fusing into one after 1989 with major transformation processes in institutions (mainly in the disappearing GDR), political trajectories, and social movements has called for systematic biographical research. Likewise, the considerably growing right wing activities of women in East and West Germany during the last few years raised questions concerning the biographical background and impact of extremist political orientations (see exemplary Köttig et al. (2012) and the quoted literature there).

In general, the effort to understand how ordinary people deal with situations of societal and global crisis, ethnicized social conflicts, war and trauma, and violence, abuse, and disappearance (Germano, 2013) has been, and still is, a core topic in biographical research. Insights on how these experiences shape people's lives, how they are part of social contexts by the way in which those affected orient and act based on their experiences (which are formed and change over a longer period of time), and how they mold different generations and their mutual relations are specific contribution from biographical research. Thus, biographical research is closely part of and connected with general processes of societies have to deal with the aftermath of war, genocide, and collective violence in different parts of the world.

Collective Identities

Associated with these research topics, interest in collective identities and how they develop in situations of transforming societies has expanded during the last 10 years. The concept of ethnicity and its significance in biographies has been widely scrutinized (Apitzsch, 2012b; Rosenthal and Bogner, 2009). Seen from a biographical perspective, ethnicity is no longer conceptualized as an ascribed characteristic of a specific group, even if a group identifies itself ethnically, but is seen as something which is created and revised in biographical processes in which the significance of ethnicity is negotiated in majority–minority relations, in family and generational relations, and within a social community. Its meaning thus can vary and change, e.g., from a marginal social position due to an ascribed belonging, which hinders social mobility, to a biographical resource to overcome discrimination.

In this line of thought and based in research on the formation of European identities in a globalizing world, the intriguing concept of a 'European mental space' has been developed (Schütze and Schröder-Wildhagen, 2012). The narrative interviews conducted with several groups who systematically experienced international and transnational encounters within and outside of Europe over a longer period of time, revealed that one cannot assume a European identity comparable to national or other collective identities, as a specific form of 'we-relation' in the sense of Alfred Schütz. Instead, a 'European mental space' is developing as a frame of reference in which comparisons of lifestyles based in experiences is possible; biographical opportunities as well as obstacles in mutual recognition of competences are becoming visible; and arenas for younger generations to work through historically burdened relations is evolving – to name just a few relevant aspects of this concept. A presently developing European mental space might serve in the future as basis to build more coherent 'we-relations,' which are only effective if rooted in common experiences, mutual comparison, critique, and recognition. Even though national identities still prevail, in the frame of a European mental space they become more liberal and loose.

Migration and Transnational Biographies

Research on ethnicity and collective identities is, not surprisingly, closely connected with investigations of the biographical meaning of migration processes. Migration research in sociology traditionally has been an area of quantitative research focusing on the flows of migration, on push and pull factors, on structural conflicts and changes in immigrant societies, and on social stratification connected mainly to work migration, etc. The experience of migrants and their patterns of action only recently have become a broader topic in scholarly discussion, particularly in connection with the criticism on the 'methodological nationalism' inherent in macro-sociological perspectives in former mainstream migration research. There have been pioneering and innovative conceptual contributions from biographical research to this shift in migration research (Apitzsch and Siouti, 2007; Breckner, 2007; Lutz, 1993).

One of the main insights has been that the considerable family- and ethnic milieu-orientation in migrant communities can be a forward-oriented strategy supporting the second and third generation to socially advance in the immigrant society and cope with exclusion, rather than a backward link to the society from where the migrants came. This insight resulted from in-depth biographical studies and led to criticism of concepts of modernization which were widely applied in migration research mostly from a macro-sociological perspective. The latter presupposes that the attitudes, especially of working migrants, as shown in religious habits and traditional family structures, are characterized by backward traditionality and a gap in modernization compared to secularized life worlds, and thus were the main obstacles in a normative perspective favoring assimilation as solution to social tensions. In contrast, biographical as well as ethnographic research, which included families and members of different generations, revealed that seemingly obvious (because visible) signs of 'premodern backwardness,' like symbols of religiosity and emblems of traditionality, are due to reflexive processes in dealing with a current social situation rather than reminiscences of a past that has to be overcome in a modern society in order to adapt. A more adequate and differentiated understanding of gender roles, ethnic communities, and religiosity could evolve through the complex interplay between the knowledge developed in one society orienting everyday life in the perspective of the whole lifespan and greater generational

transmission, with that developed newly in another society and milieu from the position of 'foreigner' or 'stranger' (Apitzsch and Kontos, 2008; Inowlocki, 2006).

In this line of thought the general image of 'migrants' is changing. Within the well-known concept of Robert Parks 'marginal man' as a prototype of the migrant, the experience of migration necessarily seems to create personal disorder. Looking at migration experiences and how they are situated in biographies, one can see that personal crises and trajectories of suffering triggered by migration can, but do not necessarily happen. Experiences of crises and strangeness are not caused by the fact of migration itself, but by its being embedded in an overall biographical construct (Riemann, 2003), by the shape of the border which has been crossed in respect of the possibility of return, by the conditions of immigration, and not least by the historical relations between the respective societies of departure and arrival, which have the potential to sharply shape experiences of discrimination or acceptance of migrants. The latter can be structured by relations built on cooperation versus hierarchical subordination, domination, or even by hostility, with a broad range in between, including ambivalence on a collective level (Breckner, 2007; Waniek, 2012).

Altogether, with a biographical approach specific aspects and dimensions of the complex field of migration can be grasped. Migration can be reconstructed as a process with an open beginning and an open end. Here the concrete social contexts in which migration is embedded and experienced by those involved become visible. The perspective of migrants which develops from their experiences in the contexts of both departure and arrival is part of the subject matter. The extent to which migration implies or requires changes of orientation and action patterns and the extent to which it allows or supports the continuation of orientation is revealed. Furthermore, with a biographical perspective addressing an entire lifetime and the interrelations between different life spheres, it can be shown how experiences of migration and its trajectories are connected with the dynamics and trajectories in other life spheres (e.g., the professional or the family sphere); how migration experiences interact with specific life phases, careers, or status passages; how the meaning of migration can change over the course of time and in relation to different biographical and societal contexts; and how members of the same immigrant community experience migration differently. Immigrant communities no longer appear as homogeneous or even subject to monolithic 'otherness,' but can be seen as a space for interactive negotiation between experiences connected to life in the societies which were left and the current position of being a migrant in a receiving society (Breckner, 2007: 118).

Finally, the contributions from biographical research to the broad discussion of different concepts of transnationality are remarkable and outstanding. I can outline here just a few. Biographical and anthropological research in Russian–Israeli and Russian–Jewish–German communities on the meaning of food and its transformation in a new societal context, while it still is perceived and practically kept as a strong link to the society and culture left behind, shows how transnationality is literally and materially created by food as an important means for creating family and community bonds (Bernstein, 2010). Other research focuses on how the second generation of work migrants in Germany do not completely transcend their national belongings, but negotiate them biographically to serve as enlarged opportunity e.g., for having educational careers in two different countries trying to overcome the obstacles of each. However, the challenges in transnational lives become apparent when opportunities are not backed by the still-prevailing institutional structures organized within the frame of nation states, as e.g., the social security and educational systems even within Europe, Bologna notwithstanding. As a result from this research, sensitive concepts of transnationality are claimed (Siouti and Ruokonen-Engler, 2014).

Gender and Intersectionality

Following up from the research on transnational migration, investigations of its impact on gendered experiences, biographies, and the gender order in general, are major contributions to this broad field of scholarly research (Apitzsch, 2012a; Lutz, 2011). The life world and biographical paths of the 'new maids' in Germany (mostly women working in private households, in general without work permission or social security arrangements, coming mainly from Poland, the Ukraine, and other Eastern European Countries), has been examined (Lutz, 2011). This research gives close insights into a field which publicly is barely visible; and it shows how global shifts in the distribution of work concretely take place and how their consequences are experienced by those involved on the side of the employers as well as by the jobholders in an unregulated transnational job market. It becomes apparent, that the concept of 'nuclear family' as one based in face-to-face close relationships is changing, because mothers work abroad leaving their children with grandparents or other members of a wider family, and keep in touch with them almost exclusively by phone or Skype, bridging the long distances and periods of absence.

In connection with recent feminist discourses, biographical research in this area expands theoretically as well as methodologically by elaborating 'critical sociological theory' and contributing to its development with concepts of intersectionality (Lutz et al., 2011; Davis, 2014a). These are based on empirically consolidated findings about the way in which gendered experiences and situations concretely connect with class positions and ethnic belonging or ascription in biographical trajectories (Lutz and Davis, 2009). Approached from a biographical angle, gender, class, race, and ethnicity are not perceived just as categories which have to be related in an abstract theoretical model, but as intertwined experiences and situations in which one dimension can get more relevant as another, or drops to the background at all, only to come to the fore in another biographical context, period of time, or social space. The challenge present in almost all biographical research, to precisely reconstruct the connection between personal experiences, orientations, and action with global societal transformation, undermining the distinction between micro-and macro-approaches in sociology, is explicitly addressed in this area. How social inequality among women develops on a global scale and at the same time is negotiated in concrete interactive settings, how gender roles in everyday life practices and their generational transmission

are negotiated, and how an increasingly global economy touches and transforms local gender arrangements are core topics to understand how structural global changes go along with the transformation of locally bound social settings.

Social Exclusion and Inclusion

Finally, issues of social exclusion and inclusion, social inequality, policy, and social work constitute other areas of research in which biography strongly matters. Practitioners in these areas always have been well aware of the fact that biographical processes and trajectories can be deeply shaped by experiences of discrimination and even exclusion resulting from an ascribed or felt belonging to a social group or category. Likewise, to find ways to overcome critical life situations, the biographical as well as collective resources have to be taken into account. Research on how to deal with crises and risks by understanding the individual perspective of those affected without individualizing them has been at the core of various research projects (Chamberlayne et al., 2002; Santos, 2010). Additionally, the consequences of professional intervention on the lives of subjects are investigated in order to develop, together with practitioners in social work, a self-reflexive attitude which is crucial, also as part of professional formation (Dausien et al., 2008; Siouti and Ruokonen-Engler, 2014). To understand oneself to a certain extent is regarded as necessary for understanding and helping others to cope with difficult situations. This area is also connected with educational sciences where biographical approaches are even more widespread than in sociology. Especially questions concerning education and lifelong learning have been at the center of interest over a long period of time (Alheit and Dausien, 2007; Olesen, 2012).

Conclusion

In a globalizing world, different concepts of biography and life story rooted in different societal organizations and cultural traditions of narratives appear and start to interact. The inter- and transcultural encounters of different social concepts on the organization of life experiences in a meaningful and orienting frame will bring about changes in the concept of biography and its narrative patterns as understood and examined in so-called western societies. In what way the social concept of 'biography' will travel or disseminate globally in the future, will remain an intriguing project for scholars to examine.

In conclusion, 'biography' and 'society' still matter and will matter in the future. The further development of methodologies and methods has been a relevant field of biographical research integrating narrative, ethnographic, and discourse approaches. Innovative potential for further development of methodologies and methods has evolved by addressing the body and emotions as a relevant part of social life that is closely related to biography, and by taking in the visual dimension of constructing biographies together with narratives in the context of changing media technologies.

Furthermore, biographical research matters because of its contributions in analyzing and understanding major issues in changing societies. Biographical approaches have the potential to concretely show how societal upheaval, shifts in social systems, and in general social structures affect the lives of those involved, and how actors contribute to shaping the social contexts they are living in. Especially the aftermath of social crises, historical shifts, war, and genocidal atrocities lived and transmitted through different generations become visible through reconstructing biographies and family histories by eliciting and analytically understanding their narratives. This applies also to understanding the complexity of migration processes. Viewpoints on specific groups of migrants in public discourse, which often bear discriminatory clichés and prejudices, are confronted with the perspectives of those who are ascriptively ethnicized. By means of biographical approaches, a more sophisticated picture of the aspirations, habits, orientations, and prospects of individuals who are part of different ethnic or religious communities and negotiate different expectations, belongings, rules, and knowledge in organizing everyday life, can evolve.

How will these areas develop in the future? Given the situation of an increased visibility of an unequal global and local distribution of material resources, and of a multitude of separations and conflicts between different groups along class, gender, and ethnic lines, the search for new concepts of social inclusion will become essential and indispensable. Biographical research might focus more on these vital issues showing how ordinary people develop strategies to deal with social conflicts and inequality, and how, through institutional settings, they can be supported in their attempts to construct peaceful lives in the diversity of social contexts we all are involved in. The examination of passion as a strong feeling and its biographical power might expand to, and connect with, the political field thus showing that strong emotions like passion are a driving force, in positive as well as destructive ways, in the social construction of biographies. Another topic on which one can expect further biographical research, is the dynamic change in media technologies and their impact on the processes of constructing biographies by different means of communication. Conceptional considerations of these processes have just started. These new directions ensure that biographical research will continue to be a success story.

See also: Biography: Historical; Discourse Studies; Ethnography in Applied Social Research; Migration: Anthropological Perspectives; Visual Sociology.

Bibliography

Alheit, P., Dausien, B., 2007. Lifelong learning and biography: a competitive dynamic between the macro- and the micro level of education. In: West, L., Alheit, P., Andersen, A.S., et al. (Eds.), Using Biographical and Life History Approaches in the Study of Adult and Lifelong Learning: European Perspectives. Peter Lang, Frankfurt am Main, pp. 57–70.

Andrews, M., 2007. Shaping History. Narratives of Political Change. Cambridge University Press, Cambridge.

Apitzsch, U., 2012a. Transnational migration, gender and European citizenship. Aleksanteri Papers. In: Saarinen, A., Calloni, M. (Eds.), Builders of a New Europe. Women Immigrants from the Eastern Trans-regions. Kikimora Publications, Helsiniki, pp. 122–131.

Apitzsch, U., 2012b. In: Bekerman, Z., Geisen, T. (Eds.), The Concept of Ethnicity and Its Relevance for Biographical Learning. Springer-Science, Heidelberg, London, New York, pp. 53–66.

Apitzsch, U., Kontos, M., 2008. Self-employment Activities of Women and Minorities. Their Success or Failure in Relation to Social Citizenship Policies. VS-Verlag, Wiesbaden.

Apitzsch, U., Siouti, I., 2007. Biographical Analysis As Interdisciplinary Research Perspective in the Field of Migration Studies. http://www.york.ac.uk/res/researchintegration/Integrative_Research_Methods/Apitzsch%20Biographical%20Analysis%20April%202007.pdf.

Bell, S.E., 2009. DES Daughters. Embodied Knowledge and the Transformation of Women's Health Politics. Temple University Press, Philadelphia.

Bell, S., 2010. Visual methods for collecing and analysing data. In: Bourgeault, I., Dingwall, R., Vries, R.D. (Eds.), Qualitative Methods in Health Research. Sage, Los Angeles, London, etc, pp. 513–535.

Bernstein, J., 2010. Food for Thought. Transnational Contested Identities and Food Practices of Russian-Speaking Jewish Migrants in Israel and Germany. Campus Verlag, Frankfurt am Main. Chicago University Press, New York.

Bertaux, D., 2001. Biography and society. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Elsevier, Amsterdam, pp. 1210–1213.

Bertaux, D., Thompson, P., Rotkirch, A., 2004. On Living in Soviet Russia. Routledge, London.

Breckner, R., 2007. Case-oriented comparative approaches. The biographical perspective as opportunity and challenge in migration research. In: Schittenhelm, K. (Ed.), Concepts and Methods in Migration Research. Study Group "Cultural Capital during Migration" – Conference Reader, pp. 113–152. www.cultural-capital.net/reader/Concepts-and-Methods.pdf.

Breckner, R., 2013. Bild und Biographie - ein Kaleidoskop von Selbstbildern? In: Heinze, C., Hornung, A. (Eds.), Medialisierungsformen des (Auto-)Biographischen. UVK, Konstanz.

Breckner, R., Kalekin-Fishman, D., Miethe, I., 2000. Biographies and the division of Europe. Experience, action and change on the 'Eastern Side'. Leske & Budrich, Opladen.

Brose, H.-G., 1989. Coping with instability – the emergence of new biographical patterns. Life Stories 5, 3–26.

Bruner, J., 1987. Life as narrative. Social Research 54, 11–32.

Chamberlayne, P., Rustin, M., Wengraf, T., 2002. Biography and social exclusion in Europe. Policy Press, Bristol, UK.

Coetzee, J.K., 2011. Narrating memory: weighing up the testimony. Qualitative Sociology Review VII, 31–43.

Coetzee, J.K., Gilfillan, L., Hulec, O.H., 2004. Fallen Walls: Prisoners of Conscience in South Africa and Czechoslovakia. Transaction Publishers, Piscataway.

Dausien, B., Hanses, A., Inowlocki, L., et al., 2008. The analysis of professional practice, the self-reflection of practitioners, and their way of doing things. Resources of biography, analysis, and other interpretative approaches. FQS – Forum Qualitative Social Research 9.

Davis, K., 2007. The Making of Our Bodies, Ourselves. How Feminism Travels across Borders. Duke University, Durham and London.

Davis, K., 2014a. Intersectionality as critical methodology. In: Lykke, N. (Ed.), Writing Academic Texts Differently: Intersectional Feminist Methodologies and the Playful Art of Writing. Routledge, New York.

Davis, K., Lutz, H., 2002. Traveling theorists. On biographical border crossing and feminist imagination. In: Davis, K.D., Brinkgreve, C., van Heerikhuizen, B., et al. (Eds.), Levensverhalen. Sociologisch Tijdschrift, Amsterdam, pp. 38–58.

Davis, K., 2014b. Passionate encounters, dancing tango in a globalizing world. New York University Press, New York.

Delcroix, C., Pape, E., 2010. Conducting field research abroad – a socio-anthropological approach. Zeitschrift für qualitative Forschung 11. http://www.budrich-journals.de/index.php/zqf/article/view/5778.

Dussel, I., Gutiérrez, D., 2006. Educar la mirada. Políticas y pedagogías de la imagen. FLACSO/Fundación OSDE, Buenos Aires.

Ettorre, E., 2010. Autoethnography: making sense of personal illness journeys. In: Bourgeault, I., Dingwall, R., de Vries, R. (Eds.), Handbook on Qualitative Methods in Health Research. Sage, London, pp. 478–496.

Fischer-Rosenthal, W., 2000. Address lost: how to fix lives. Biographical structuring in the European modern age. In: Breckner, R., Kalekin-Fishman, D., Miethe, I. (Eds.), Biographies and the Division of Europe. Leske & Budrich, Opladen, pp. 55–75.

Frank, A.W., 2002. At the Will of the Body: Reflections on Illness. Houghton Mifflin, Boston.

Frank, A.W., 2010. Letting Stories Breathe: A Socio-Narratology. University of Chicago Press, Chicago.

Germano, G., 2013. Ausiencias. Available at: http://www.gustavogermano.com/.

Hackstaff, K.B., Kupferberg, F., Négroni, C., 2012. Biography and Turning Points in Europe and America. Policy Press, Bristol, UK.

Harrison, B., 2009. Life Story Research, four-volume set. Sage, London.

Humphrey, R., Miller, R., Zdravomyslova, E., 2003. Biographical Research in Eastern Europe. Ashgate, Aldershot/Burlington.

Inowlocki, L., 1993. Grandmothers, mothers, and daughters: intergenerational transmission in displaced families in three Jewish communities. In: Bertaux, D., Thompson, P. (Eds.), Between Generations. Family Models, Myths and Memories. Oxford University Press, Oxford, pp. 139–154.

Inowlocki, L., 2000. Doing being Jewish: constitution of normality in families of Jewish displaced persons in Germany. In: Breckner, R., Kalekin-Fishman, D., Miethe, I. (Eds.), Biographies and the Division of Europe. Experience, Action, and Change on the 'Eastern Side'. Leske+Budrich, Opladen, pp. 159–178.

Inowlocki, L., 2006. Stretching it so it fits real well. Biographical, gendered, and intergenerational dimensions of turning to religion. In: Geyer, M., Hölscher, L. (Eds.), Die Gegenwart Gottes in der modernen Gesellschaft, Transzendenz und religiöse Vergemeinschaftung in Deutschland. Wallstein, Göttingen.

Kazmierska, K., 2012. Biography and Memory: The Generational Experience of the Shoah Ssurvivors. Academic Studies Press, Brighton, MA.

Kohli, M., 2007. The institutionalization of the life course: looking back to look ahead. Research in Human Development 4, 253–271.

Köttig, M., Müller-Botsch, C., Schiebel, M., 2012. Biography and politics. FQS – Forum Qualitative Social Research 10.

Kohler Riessman, C., 2008. Narrative Methods for the Human Sciences. Sage, Los Angeles.

Lutz, H., 1993. In between or bridging cultural gaps? Migrant women from Turkey as mediators. New Community 19 (3), 485–492.

Lutz, H., 2011. The New Maids. Transnational Women and the Care Economy. Zed Books, London.

Lutz, H., Davis, K., 2009. Geschlechterforschung und Biographieforschung: Intersektionalität als biographische Ressource am Beispiel einer außergewöhnlichen Frau. In: Völter, B., Dausien, B., Lutz, H., et al. (Eds.), Biographieforschung im Diskurs. VS Verlag, Wiesbaden, pp. 228–247.

Lutz, H., Teresa Herrera Vivar, M., Supik, L., 2011. Framing Intersectionality. Debates on a Multi-Faceted Concept in Gender Studies. Ashgate, Abingdon.

Miller, R., 2005. Biographical Research Methods, vols. 4. Sage, London.

Miller, R., Day, G., 2012. The Evolution of European Identities. Biographical Approaches. Palgrave Macillan, Houndmills, UK.

Newsletters, 2008–2013. Reports on biographical research in different countries (e.g. UK, France, Hungary, Latin America, Denmark, Brazil, Greece). International Sociological Association, RC 38: Biography and Society http://www.isa-sociology.org/rc38.htm.

Olesen, H.S., 2012. Cultural analysis and in-depth hermeneutics – psycho-societal analysis of everyday life, culture, interaction, and learning. FQS – Forum Qualitative Social Research 13.

Plummer, K., 2001. Sexualities: Critical Concepts in Sociology. Routledge, London.

Riemann, G., 2003. Doing biographical research. FQS – Forum Qualitative Social Research 4.

Rosenthal, G., 1993. Reconstruction of life stories. Principles of selection in generating stories for narrative biographical interviews. Narrative Study of Lives 1, 59–91.

Rosenthal, G., 2010. The Holocaust in Three-Generations. Families of Victims and Perpetrators of the Nazi-Regime. Barbara Budrich, Opladen (New edition of 1998 by Cassell).

Rosenthal, G., Bogner, A., 2009. Ethnicity, Belonging and Biography. Ethnographical and Biographical Perspectives. LIT, Münster and New Brunswick.

Rosenthal, G., Sagy, S., Dajani Daoudi, M.S., 2013. Belonging to the Outsider and Established Groupings: Palestinians and Israelis in Various Figurations. Newsletter of the ISA RC 38, pp. 45–46 http://www.isa-sociology.org/pdfs/rc38newsletter_july2013.pdf.

Santos, H., 2010. Adolescents and experiences with violence: making sense of subjective interpretations of life-world. ISSN: 1809–4341. Vibrant. Florianópolis 7, 135–165.

Schütze, F., 1992. Pressure and guilt: war experiences of a young German soldier and their biographical implications. International Sociology 7, 187–208, 347–368.

Schütze, F., 2008. Biography analysis on the empirical basis of autobiographical narratives: how to analyse autobiographical narrative interviews. Part I, II. European Studies on Inequalities and Social Cohesion 1, 2, 153–242, 155–177.

Schütze, F., Riemann, G., 1990. Trajectory as a basic theoretical concept for analyzing suffering and disorderly social processes. In: Maines, D.R. (Ed.), Social Organization and Social Processes. Aldine, Hawthorne, NY, pp. 333–357.

Schütze, F., Schröder-Wildhagen, A., 2012. European mental space and its biographical relevance. In: Miller, R. (Ed.), The Evolution of European Identities. Biographical Approaches. Palgrave Macmillan, Houndmills, UK, pp. 255–278.

Semenova, V.V., 2002. Two cultural worlds in one family. The History of the Family 7, 259–280.

Siouti, I., Ruokonen-Engler, M., 2014. "Doing biographical reflexivity" As a Key Tool in transnational research settings. Transnational Social Review: A Social Work Journal 3.

Stanley, L., 2006. Mourning Becomes… Post/Memory and the Concentration Camps of the South African War. Manchester University Press, Manchester.

Stanley, L., Temple, B., 2008. Narrative methodologies: subjects, silences, re-readings and analyses. Qualitative Research 8, 275–281.

Thomas, W.I., Znaniecki, F., 1918–1920/1923. The Polish Peasant in Europe and America. Monograph of an Immigrant Group, vols. 5. Richard G. Badger, Boston.

Völter, B., Dausien, B., Lutz, H., et al., 2005. Biographieforschung im Diskurs. VS Verlag, Wiesbaden.

Waniek, K., 2012. Polish Immigrants to Germany. Biographical Analysis of Narrative Interviews with Young Polish People Who Left for Germany between 1989 and 1999. Wydawnictwo Uniwersytetu Łódźkiergo, Łódź.

Biography: Historical

Birgitte Possing, The Danish National Archives, Copenhagen, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article defines the historical biography as encompassing both a reconstruction of a human life and a representation of a human individual. It shows that the historical biography is a genre that verges on a series of genres and disciplines. The history of the historical biography reaches from antiquity to present days, and the genre is characterized by variety and diversity, both in historical outlook, media, and methodology, and in a cornucopia of categories and forms. Different aims, archetypes, and methodologies of the historical biography have developed the genre into a prism for a multitude of preoccupations, until 'the biographical turn' in history during the first decade of the 2000s augured the genre of the historical biography as an interdisciplinary analytic field in its own right, placed in a battlefield of multifarious narrative forms.

Definition

Historical biography (from the Greek *historia*: inquiry, knowledge; *bios*: life; and *grafein*: writing) is a reconstruction of a human life and a representation of a historical individual.

Thus, the notion of biography encompasses more than a pure life depiction. It encompasses the events of a life, the narrative of a life, and the interpretation of its characteristics. The historical biography represents one of many ways of telling history. Traditionally, biography places the individual at the center of the narrative, instead of larger analysis of dynamics, structures, and events. As such, biography has been a controversial genre within the scholarly writing of history for centuries.

The historical biography is a genre characterized by variety and diversity, both in historical outlook, media, and methodology, and in a cornucopia of categories and forms. In terms of genre, historical biography verges on autobiography, literary biography, the traditional tale, the life story, and the biographical novel. In specialist professional terms, the historical biography relates to psychology, sociology, anthropology, history, and literature. In terms of protagonist, the historical biography is open to everyone and anyone, statesmen and farmers, generals and artists, philosophers and scientists, heroes and villains, women and men. As regards time scale, historical biography has existed since antiquity and up to the present. All this makes it an extremely complex and deeply fascinating genre. In present times, the historical biography is a multimedia phenomenon from the highest academic publications in different disciplines to the popular journalist biography and the electronic life stories of the entertainment industry. It diverges from fiction, but in its narrative form it shares similarities with the modern novel, which has long since broken with linear narrative in recognition that the passage of real life is fragmented.

2000 Years of the Historical Biography

Since antiquity, Western historical biography has been regarded as a branch of historiography and has developed from being an ethical-humanistic genre to being a genre consisting of various methodologies, forms, and styles within twenty-first-century specialized scholarship. As a genre, biography in the West is considered to have been established by the Greek Plutarch (AD 45–120), who published the comparative lives of Greek and Roman statesmen in *Bioi Parallelloi* (AD 125). This work, together with Tacitus' *De vita Agricolae* (AD 98), Diogenes Laërtius' biographies of Greek philosophers (AD third century), and Suetonius' *De Vita Caesarum* (AD 121), characterized what has been called the classical Hellenistic biography, built upon the fundamental principles of ethics: the central figures were either commended for having fulfilled their duty or censured for falling into the trap of ambition or arrogance. The aim was didactic, inspiring the general audience to become ethical acting subjects. Also the Christian gospels of the New Testament are seen as ideal biographies of antiquity. The classical biography type was maintained throughout the Middle Ages. A sidelong consideration must here be given to China's great historian Sima Qian, who very early in history developed a biographical form that belongs to a much later Western epoch: *Shiji* (Sima Qian, 145–185 BC). This work contains biographies, not just of eminent statesmen and soldiers, but also of individuals such as fortune tellers, courtiers, and murderers. A modern approach, in present Western terms, was demonstrated by Sima Qian's contemporary, the poet and literary historian Liu Xiang, whose work *Lienü zhuan* (78–79 BC) shows that female personalities in China were already at this stage considered worthy of biographical studies.

The Greek-Roman golden age of biography vanished with its empires, and in accordance with the requirements of the church and spiritual need, the Middle Ages saw the development of the martyr biography (e.g., John Foxe's *Book of Martyrs*; Foxe, 1563), and the manifold lives of holy men and women, known as hagiographies. The purpose was didactic and the central character was presented as a model of Christian propriety and public virtue: God's creation was portrayed as an individual who, in the course of a lifetime, developed as a moral example to others, or whose destiny was first fulfilled in death. The borderline between biography and hagiography was broad during the Middle Ages. Christian funeral sermons that focused on the devotional everyday life belief in God included biographical life stories of men and women not belonging to the educated people of culture. Christianity and the Christian biography opened the door to the reception of the hagiographical tradition.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62132-3

At the same time, though, a nonhagiographic model was also used for biographies of princes and emperors, with Eginhard Einhard's *Vita Caroli Magni* (AD 829) frequently cited as being the most important. The secular approach to historical biography was introduced during the Italian Renaissance, a change that continued further afield from the seventeenth up to the nineteenth century. As stated in Christopher Klein (2009), the decisive biographical innovation came from Italy with Francesco Petrarcha's (1304–1374) *De Viris Illustribus* and developed from Giovanni Boccaccio's (1313–1375) sense for the specific, the classical times, and the secular in *De Claris Mulieribus* (1355–1359) and *Vita di Dante Alighieri* (1354–1355). Petrarcha's and Boccaccio's glorification of brilliant individualities connected to the much later liberal individualism in society. Early encyclopedias of women from Germany confirm this tendency. Biographical literature became so extensive during the eighteenth and nineteenth centuries that E. M. Ottinger's (1854) bibliography of biographical literature was only rudimentary, even though it was wide-ranging.

From the end of the eighteenth century, J. Boswell's *Life of Samuel Johnson* (1791) is still singled out as the pioneering biography of the era, as it was based on empirical material in the form of letters, private papers, conversations, interviews, and personal observations of Samuel Johnson's comportment. This was the first biography to construct a nuanced, candid personality focusing on the character of the protagonist. In the historical biography of the time, the history of human achievement was fundamentally the story of imposing male heroes. In the biographical discourse of the nineteenth century, also Thomas Carlyle's lectures *On Heroes, Hero-Worship and the Heroic in History* (1841) have a paradigmatic position claiming 'the History of the World' to be the biographies of great men. Thus claiming biography as a branch of history, and employing similar processes of research and scholarship in the making of biography, he claimed historical biography as essentially the chronicle of an individual's life journey, whereas Boswell and Johnson had claimed it an art of human portraiture. This difference of perception illustrates the eternal question on history versus biography that has hung over historical biography from antiquity to the present days. But still, at the end of the nineteenth century, it was the exceptional (male) person and the developmental process that characterized the 'life-and-letters' biography. Even though ideas and people from the community at large were incorporated into the biographies via reference to letters, and via learned, illustrious, and strange women, there was generally no trace of historical reflection on the individual and society or any change of the skewed gender balance in the genre of historical biography. Here, the Russian G. V. Plechanov was a pioneer introducing the issue of the dynamics between the individual personality and the society as a historical transformation factor with *The Role of the Personality in History* (Plechanov, 1898). The impact of his work in Western historical biography, however, seems to be limited.

The modern life story, as twentieth-century historical biography has been called, came about as the result of the crisis of humanism, Christianity, and rationalism that followed in the wake of Darwinism and psychoanalysis as developed during the nineteenth century. The modern life story in literary history was introduced with L. Strachey's *Eminent Victorians* (1918), which came to play the same revolutionary role for historical biographers of all professions as Boswell had played for biography in the eighteenth and nineteenth centuries: the introduction of the artistic, interpretative biography, given form via selection, concentration, and interpretation of the sources. The idea was to get behind the myth of the subject of the biography and pave the way for the writing of biographies of men *and* women. Another member of the famous Bloomsbury group, Virginia Woolf, went even further in her book *The Art of Biography* (1939). Rather than attempting to escape the limitations of the genre as Strachey had, Woolf urged biographers to endorse them by taking the front position, "testing the atmosphere, detecting falsity, unreality, and the presence of obsolete conventions." The biographer's sense of truth had to be alive, because biography was "only at the beginning of its career; it has a long and active life before it, we may be sure – a life full of difficulty, danger, and hard work" (Woolf, 1939). Thus, both Strachey and Woolf positioned themselves as pioneers in the twentieth-century paradigm of *literary* historical biography. L. Edel (1957) published the most influential postwar study of the biography, which he honed throughout his life's work under the motto "A writer of lives is allowed the imagination of form but not of fact" (Edel, 1984). Literary biography became a well-established genre in its own right (Friedson, 1981). Within this genre, the postmodernists I. B. Nadel (1984), W. H. Epstein (1987), and P. Bourdieu (1989) have sought for a poetics of biography by pointing out that definitive life portraits are not to be found.

Biography could be 'collective,' too, involving a comparison of several lives or a tighter analysis of a number of lives linked through a central theme, such as Plutarch's *Parallel Lives*, Strachey's *Eminent Victorians*, and, in a modern understanding, P. Rose's *Parallel Lives* (1983) on a number of famed English married couples in the nineteenth century, and L. W. Banner's *Intertwined Lives* on Margaret Mead and Ruth Benedict (Banner, 2003).

Biography as a Methodological Approach

By contrast, within the historical and sociological disciplines, the historical biography remained a methodology in understanding the significance of the historical individual and the subjective experience, rather than becoming a genre in itself. Nevertheless, the historical biography vividly developed during the twentieth century, displaying a multitude of biographical methodologies, as documented in P. France and W. St Clair (2002), G. Fromm (1986), J. Le Goff (1989), M. Rhiel and D. Suchoff (1996) and M. Shortland and R. Yeo (1996). Furthermore, the number of published historical biographies increased. Most important was the emergence of the *life-work-and-times* biography, in which the individual was contextualized in the light of contemporary society. A characteristic of the twentieth-century historical biography was that male heads of state, with a few exceptions, still were the focus of interest throughout the century, in spite of the first- and second-wave feminism of the nineteenth and twentieth centuries. Another characteristic was that biographical critique and a theoretical debate on biography as a genre intensified at regular intervals without being carried out systematically. A new aspect was incorporated into the genre: demythologization and unmasking of the protagonist; as Oscar Wilde said, "Formerly we used

to canonize our heroes; today we vulgarize them." Finally, historical biography was influenced by both academics and artists. Thus, serious biographical literature covered the entire spectrum from not very accessible treatises to stylistically inviting reading.

As mentioned, the sociological and historical professions regarded the biography critically throughout the twentieth century even though the publications of historical biographies were manifold (Possing, 2001, 2012b; Rosengren and Östling, 2007; Klein, 2009; Caine, 2010). Historical-sociological biographers and literary biographers began to draw a little closer together after the 1930s. This drawing together continued throughout the rest of the century: literary-framed biography became more scholarly framed, and historical-, sociological-, anthropological-, and natural science–framed biography became more literary framed, claiming that the more objective a biography was, the more lifeless and hollow it became (Kendall, 1965; Madélenat, 1984). The psycho-historical biography did not really establish a firm footing within the humanities and the social sciences, as did the life narratives growing into a paradigmatic trend of mixing autobiography and biography in literature and history at the turn of the millennium.

Sociologists have used historical biography as a methodology for understanding the significance of subjective experience. Pioneers in this respect were F. Znaniencki and W. Thomas (1918–1920), who presented the individuals and groups of individuals as both creative and created in social evolution, thus introducing prosopography. The life history perspective was followed up by the Italian sociologist F. Ferrarotti (1983), who saw the individual as an active pivot in respect to the structures and history of society and B. Merrill and B. West (2009) who reflected on the use of biographical methods in social research. Life histories have also interested late-twentieth-century anthropologists wishing to promote subjective experience as a factor in the creation of personal identity (Crapanzano, 1980). Anthropologists have considered biography to be a participant observer in a cross-cultural and cross-epochal dialogue with the historical subject.

The mentalities-historians also influenced the historical biography, especially the French Annales-school and its preoccupation with personification and its hermeneutic, mentalities-historical studies. In her innovative study of sixteenth-century *Martin Guerre* (1983), N. Zemon Davis used new ideas, concepts, and methodologies from anthropology, ethnography, and literary criticism to examine the dilemma of truth and doubt in historical research. Perception of a central figure became the result of a communicative process between two cultures and two people, not an objective description. Jean-Paul Sartre's concept of interaction between the life lived and the written text inspired the development of the dialectic biography. In Denmark and the Scandinavian countries, the historical figure within this tradition was interpreted as both the bearer of a cultural convention and a cultural agitator (Possing, 1992, 1997, 2001; Klein, 2009). Sociologists, ethnologists, and historians have all used reminiscence, the oral-history tradition, interviews, and private papers as empirical source material via which to understand social action in a wider perspective as a confrontation with the positivist, Marxist, and structuralist thinking that has starved the biography of life. The objective was a demythologization of major figures: "A modern biographer may or may not choose to reveal the intimate, the amorous details of a life, but he must, if he is good at what he does, probe beneath its public, polished self" (Pachter, 1979).

Nonetheless, academic historians were ambivalent about the genre of biography. It remained the professions' unloved stepchild or courtesan (Possing, 2012a), because many historians were skeptical of the capacity of biography to convey the kind of analytically sophisticated interpretation of the past that academics expected. At the turn of the millennium, historians still characterized biography as a degraded form of historical writing.

The Budding Renaissance of Biography

In the light of a quarter century's interest in social history, mentalities-history, and historical narrative, the historical biography, with its focus on the individual above and as part of community, class, and social group, underwent a renaissance. The role of historical biography as a point of orientation in human life was one of several reasons for its popularity as a genre at a time when belief in the great utopias disappeared with the end of the Grand Narratives. Historical biography became the prism for a multitude of preoccupations. The 1990s saw a desire for images of women and men as creative, reflective, doubting, and determining individuals, and a wish to leave behind the tendency to reduce specific singularity to social regularity. More historians, with G. Levi (1989) in the lead, expressed this need for open reflection by pointing out that, more than ever, the ambiguous biography was a central preoccupation for historians. The biography became a sanctuary for the dramatic, old-fashioned narrative amidst a deconstructing and fragmented era. It facilitated reflection on the human ideal in the perspective of the flawed living individual, preventing biography from becoming eulogy.

The renaissance of biography was a manifestation of a genuine renewal of the genre, taking its inspiration from advanced literature and the new gender studies, and from the critical question of the extent and manner in which history is created by people and how a life can be decoded. On the one hand, historical biography made a justifiable demand to set the agenda via a rehumanization of the humanities, which meant that passion, irrationality, and human idiosyncrasy were drawn into an understanding of an individual's life's work (Caine, 2010; Hamilton, 2007, 2008; Klein, 2009). On the other hand, the genre's respectability was contested time and again, and still, it was not secured a serious, scholarly profile. Some maintained that biography chose them while they were engaged in other research. Ambivalence in relation to placing biography as a genre within historiography seemed to be the result of an ongoing opposition to a consideration of single characters. Individual lives, especially the private, were considered as irrelevant or inappropriate to the understanding of history, scholarship, or art by the structuralist, positivist, and postmodern antibiographers.

But in spite of being left out in the cold, the genre imposed itself again and again. This was further illustrated by the reaction of postmodern and structuralist critics such as Jacques

Derrida, Michel Foucault, and Roland Barthes, who called the historical biography "impossible to use as reference," "spurious," "profit-mongering in intimacy," or, conversely, a thanatography – an account of a person's death. Bourdieu used the expression "the biographical illusion" (1986), by which he meant that a life story had no direction and thus could not be construed in a chronological order. Against such criticism, social constructivist historians stressed that a human individual had to be conceived as a part of the surrounding society; that biography did not represent a life as a continuum, but as a construction of many identities; and that historical biography could and should de- and reconstruct a life (Macey, 1993; Margadent, 2000). The biographical turn in history, literature, other sciences, the media, and the entertainment industry became reality at the turn of the millennium.

The Biographical Turn

During the first decade of the 2000s, historical biography underwent a renewal. It exploded in a cornucopia of publications, and historians identified the biographical turn in history. Historical biography changed from an approach in history to an inherently interdisciplinary and self-assertive genre calling itself the 'new biography' (Wagner-Martin, 1994; Margadent, 2000). This was especially influenced by feminists, postmodernists, and race theorists. Feminists and postmodernists focused on the diversity and complexity of personalities, movements, and institutions, decrying 'essentialism',. Theorists of race focused on the marginalized black and Afro-American groups who had suffered oppression and discrimination, raising broader issues on transnationalism in postcolonial studies. Already during the twentieth century, historical biography as a field of interest had developed from a conscious mythologization of prominent, historical figures to a critical, manifold storytelling of public, national, transnational, and anonymous historical lives.

In the light of the abundance of publications that emerged during the first decade of the third millennium, the biographical genre can be understood from two perspectives. *Firstly*, historical biography has been despised and looked upon with condescension, despite the fact that in the international forum it is a 2000-year-old genre (Egeland, 2000; Hamilton, 2008). Although not gaining respect or dignity, paradoxically scholars and journalists have produced a cornucopia of biographical works during the last 20 years. Historical biography seems to be much appreciated by the general audience. It is like a courtesan, in constant demand and much visited, but taboo and referred to with disrespect by academic society. As the courtesan is visited by men, so too the biography has been preoccupied with men. Statesmen, kings, financiers, scientists, scholars, and authors of the male sex have populated the genre, which has only within the last decade become properly accessible to women and ordinary, anonymous people (Larsson, 2001; Banner, 2003; Huismann et al., 2012; Possing, 2012b). The well-known gender imbalance of the historical biography is changing and has altered with the biographical turn in history, with more historical biographies having female protagonists or female biographers. This phenomenon was pioneered by authors and scholars with experience from the modern fields of women's history and gender studies (Alpern, 1992; Possing, 1992; Moi, 1994; Larsson, 2001; Banner, 2003; Benhabib, 2003; Hirdmann, 2006).

Secondly, historical biography underwent a major renaissance at the end of the twentieth century – in the Scandinavian countries, all over Europe, and in most of the Western world. It exploded as part of the popular 'experience economy' and in academic culture. Since the turn of the millennium, the spirit of our times has been colored by 'the biographical turn' (Hamilton, 2008; see also Caine, 2010) and "le retour de la biographie" (Arnaud, 1989; Klein, 2009). This can be explained as an element of the individualistic or individualizing zeitgeist following the demise of the Grand Narrative. In present days, there is hardly a conference or a symposium in the area of the humanities and social sciences that does not have a session or a paper on the subject of biography. National and international networks of autobiographical and biographical authors and scholars are blooming. Also, autobiography and the myriad forms of life stories and narratives have exploded as a sphere of interest.

The parameter for the evaluation of a historical biography is most often the dialectical *life-work-and-times* biography with a balanced analysis of the individual situated in the larger context, or with an analysis of the power structure and the dynamic interplay between the individual and the society. Most are based on archival material and some even on detective-like research. Some historical biographies skew the depiction perspective at the expense of historical analysis; others focus on more simple life depiction at the expense of analysis (Rosengren and Östling, 2007; Banner, 2009).

The historical biography can also be written as a *prism* in which the light of history is refracted and the perspective raises the central figure as representative of a time, a historical situation, a type, a social phenomenon, or a culture. Or it can be written as a *collective or relational biography* comparing more lives as in the tradition of Strachey (1918) and Rose (1983), and as seen in Banner (2003) or in Anneke Ribberink's comparison of prime ministers Margaret Thatcher and Gro Harlem Brundtland (Huismann et al., 2012).

Challenges in the Wake of the Biographical Turn

At the turn of the millennium, the historical biography was in the process of redeveloping 'l'ecriture historique,' historical writing. After the biographical turn in history, the challenge of the historical biography changed. Despite the biographical flood, there was no well-developed critical tradition. Neither universities nor schools of higher education offered specific training in the all-round study of biography: "Its ethics, like the history and theory behind it, thus go largely unaddressed, while at a practical level there are still relatively few courses offered to those who wish to write a biography, whether big or small. Would-be biographers are thus left largely to their own devices, scrabbling for advice and examples in every direction," the British biographer and historian Nigel Hamilton wrote in 2008. Biographers from Plutarch to Ian Kershaw, the biographer of Hitler (Kershaw, 2008), have hard-won experiences in depicting real lives, and in representing their stories in historical biographies. Experiences from the reconstructions of human life and the representations of historical individuals

through 2000 years have shaped biographical conventions. Traditionally, every biography is a life from cradle to grave; every biography has an intention, a plot, an agenda, or a motive beyond that of telling a life; every biography is shaped by the biographical triangle (the relation between the biographer, the protagonist, and the audience); and every biographer holds the fate of someone else in their hands. At the biographical turn in history, all these four conventions are being both challenged and used by present historical biographers. The biographical turn not only happened because of the post-Freudian Western world's inclination to psycho-historical analysis. It also happened because the historical biography became part of the democratic project in an increasingly globalized community. Historical biography expresses interest in the individual, and his or her place in culture and society.

The new knowledge of the historical biography is to be found in the *interdisciplinary analytic field* known as biographical studies. This field distinguishes between the *genre*, this being the empirical narratives of individual lives as they can be read from antiquity to the present day, and the *academic discipline*, which this genre comprises. The biographical sphere is by its very nature interdisciplinary, because a single historical biography borders on many other genres, subjects, and narrative traditions; it raises the issues of politics, social relations, economics, culture, ancestry, kinship, family, sexuality, psychology, and friendship (Klein, 2009).

Thus, the historical biography has turned into an interdisciplinary sphere and a genre, placed in a battlefield of multifarious narrative forms. Only lately have biographers started a theoretical work of building bridges between types, intentions, methodologies, and narrative techniques. Historical biography is in the frontline of tackling an unwieldy ethical accountability vis-à-vis an increasing public pressure on the right to invade the private lives of public personalities (King, 2001; Rollyson, 2005; Baets, 2009, 2010). And last, but not least, the historical biography is in the vanguard of understanding the influence of the inclusion of women in the public arena to twenty-first-century changes in the genre that during the twentieth century was called 'the science of man' (Edel, 1957; Huismann et al., 2012).

See also: Ancient Greece and Rome, History of; Biography and Society; Comparative Studies: Method and Design; Enlightenment; Ethics and Values; History and Literature; History and the Social Sciences; Life Course in History; Middle Ages, The; Modernization and Modernity in History; Postmodernism: Methodology; Privacy: Theoretical and Legal Issues; Prosopography (Collective Biography); Social Constructivism.

Bibliography

Alpern, Sara (Ed.), 1992. The Challenge of Feminist Biography: Writing the Lives of Modern American Women. Illinois.
Ambjörnsson, R., et al. (Eds.), 1997. Att skriva människan. Essäer om biografin som livshistoria och vetenskaplig genre Umeå.
Arnaud, Claude, 1989. Le Retour de la biographie: d'un tabou á l'autre. no. 54. Le Débat, 40–47.
Ascher, C., et al. (Eds.), 1993. Between Women. Biographers, Novelists, Critics, Teachers and Artists Write About Their Work on Women Boston.
Baets, A De, 2009. Responsible History New York & Oxford.
Baets, A De, 2010. Posthumous Privacy, 21st International Congress of Historical Sciences, August 2010, Who Owns History? Sources Past and Present Amsterdam.
Banner, L.W., 2003. Intertwined Lives: Margaret Mead. Ruth Benedict, and Their Circle, NY.
Banner, L.W., 2009. Biography as history. The American Historical Review 114, 579–587. n 3, June.
Benhabib, S., 2003. The Reluctant Modernism of Hannah Arendt New York.
Bertaux, D. (Ed.), 1981. Biography and Society. The Life History Approach in the Social Sciences. Sage Studies in International Sociology 23. ISA, Beverly Hills.
Blanning, T.C.W., Cannadine, D. (Eds.), 1996. History and Biography. Essays Cambridge.
Boccaccio, G., 1354–1355. Vita di Dante Alighieri Das Leben Dantes, transl. Taube O., Leipzig.
Boccaccio, G., 1355–1359. De Claris Mulieribus, Concerning Famous Women. transl. G. Guarino, 1964, Rutgers.
Boswell, J., 1791. The Life of Samuel Johnson London.
Bourdieu, P., 1986. L'Illusion biografique, Actes de la recherche en sciences sociales. no. 62/63, June. Editions de Minuit, Paris.
Caine, B., 2010. Biography and History, Theory & History Hampshire.
Carlyle, T., 1841. On Hero, Hero-Worship and the Heroic in History London.
Crapanzano, V., 1980. Tuhami. A Portrait of a Maroccan Chicago, London.
Davis, N.Z., 1983. The Return of Marten Guerre Cambridge.
Edel, L., 1957. Literary Biography London.
Edel, L., 1984. Writing Lives. Principia Biographia New York.
Egeland, M., 2000. Hvem bestemmer over livet? Biografi en historisk og litterær genre Oslo.
Einhard, E., (AD 829): Vita Caroli Magni, The Frankish Kingdom.
Epstein, W.H., 1987. Recognizing Biography Philadelphia.
Ferrarotti, F., 1983. Biography and the Social Sciences 56–80, Social Research.
Foxe, J., 1563. Books of Martyrs.
France, P., St Clair, W. (Eds.), 2002. Mapping Lives. The Uses of Biography Oxford.
Friedson, A.M. (Ed.), 1981. New Directions in Biography Honolulu.
Fromm, G., 1986. Essaying Biography. A Celebration for Leon Edel Honolulu.
Garraty, J.A., 1957. The Nature of Biography New York.
Hamilton, N., 2007. Biography. A Brief History Harvard.
Hamilton, N., 2008. How to do Biography – A Primer Harvard.
Hirdman, Y., 2006. Det tänkende Hjärtat. Bokan om Alva Myrdal Stockholm.
Huismann, M., et al. (Eds.), 2012. Life Writing Matters in Europe Amsterdam.
Kendall, P.M., 1965. The Art of Biography London.
Kershaw, I., 2008. Hitler, the Germans and the final solution, New Haven.
King, M., 2001. Tread Softly for You Tread on My Life Auckland, New Zealand.
Klein, C. (Ed.), 2009. Handbuch Biographie. Methoden, Traditionen, Theorien Stuttgart.
Laërtius, D (AD third century): Vitae Philosophorum
Larsson, L., 2001. Sanning och Konsekvens Stockholm.
Levi, G., 1989. Les usages de la biographie. Annales ESC 6, 1325–1336.
Liu Xiang, 78–79 BC. Lienü zhuan, Biographies des Femmes Illustres, transl.1963 M. Kaltenmark, Peking.
Macey, D., 1993. The Lives of Foucault London.
Madelenat, D., 1984. La Biographie Paris.
Margadent, J.B. (Ed.), 2000. The New Biography. Performing Femininity in Nineteenth Century France Berkeley.
Merrill, B., West, B. (Eds.), 2009. Using Biographical Methods in Social Research London.
Moi, T., 1994. Simone de Beauvoir: The Making of an Intellectual Woman Oxford 1994.
Nadel, I.B., 1984. Biography: Fiction, Fact and Form London.
Ottinger, E.M., 1854. Bibliographie biographique 1–2, Brussels.
Pachter, M., 1979. Telling Lives: The Biographer's Art Washington, DC.
Plechanov, G.V., 1898. K Voporsu o Roli Lichnosti v Istorii [The Role of the Individual in History, transl. 1946]. In: Nauchnoe Obozrenie, 3–4.
Plutarch, AD 125. Bioi Parallelloi. Reprinted from: Parallel Lives, 11, (1958–1968), London.
Possing, B., 1992 and 1997. Viljens Styrke. Natalie Zahle. En Biografi, [with English summary: Strength of Will: Natalie Zahle. A Biography on Gender, Education and Empowerment], vol. 1–2, Copenhagen.
Possing, B., 2001. The historical biography. In: Encyclopedia of Social & Behavioral Sciences, vol. 2. Elsevier.
Possing, B., 2004. The historical biography: genre, history and methodology. In: Bale, J., et al. (Eds.), Writing Lives in Sports. Aarhus University Press.
Possing, B., 2012a. Portraiture and re-portraiture of a political individual. In: Lauring-Knudsen, A., Gram-Skjoldager, K. (Eds.), Living Political Biography. Narrating Ideas of Europe. Aarhus.

Possing, B., 2012b. Biography: An Unloved, but Much Pursued Courtesan. In: Pehrlinge, A., Sjögren, H. (Eds.), Biogrpahies of the Fiancial World. Stockholm.
Rhiel, M., Suchoff, D. (Eds.), 1996. The Seductions of Biography New York & London.
Rollysson, C., 2005. A Higher Form of Cannibalism? Adventures in the Art and Politics of Biography Chicago.
Rose, P., 1983. Parallel Lives New York.
Rosengren, H., Östling, J. (Eds.), 2007. Med livet som insats. Biografin som humanistisk genre Lund.
Scheuer, H., 1979. Biographie: Studien zur Funktion und zum Wandel einer literarischen Gattung vom 18. Jahrhundert bis zur Gegenwart Stuttgart.
Shortland, M., Yeo, R. (Eds.), 1996. Telling Lives in Science. Essays on Scientific Biography Cambridge.
Sima Quian (145-185 BC): Shiji [Historical Memoirs]
Strachey, L., 1918. Eminent Victorians London.
Svetonius, T., AD 121. De Vitae Caesarum (The Twelve Caesars), transl. Robert Graves, Penguin Classics, Harmondsworth 1957.
Tacitus, G.C., (AD 98): De Vita Agricolae
Wagner-Martin, L., 1994. Telling Women's Lives. The New Biography New Jersey.
Woolf, V., 1939. The Art of Biography. In Collected Essays, vol. 4 London 1967, 221–228.
Znaniencki, F., Thomas, W., 1918–1920. The Polish Peasants in Europe and America. Monograph of an Immigrant Group, vol. 1–5 Chicago.

Biological Preconditions for Language Development

Jutta L Mueller, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany; and University of Osnabrück, Institute for Cognitive Science, Osnabrück, Germany
Claudia Männel and Angela D Friederici, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Language, an elaborate system of discrete units and combinatorial rules, builds on complex neurocognitive foundations. Language development results from both maturation and learning. Specifically, learning mechanisms are implemented in brain networks that are still in the process of structurally and functionally maturing during the first year of life, which is the focus of this review. We will first sketch basic language-learning skills that provide the biologically grounded foundation of language development and expound their neural implementation. Subsequently, we will outline neural maturation in language-relevant brain areas and how this might relate to the way language development proceeds.

Early Speech and Language Abilities

Discrimination of the Sounds of Language

At the core of language acquisition research is the question of whether we are biologically prepared to process human language in a special way. Early behavioral research has shown that newborns can discriminate their native language from a rhythmically different language (Mehler et al., 1988) based on prosodic information. Findings showing that they can also discriminate between two rhythmically different nonnative languages that are unrelated to their native language (Nazzi et al., 1998) indicate that this is a universal perceptual ability. Not only can newborns differentiate between language and similarly complex nonlanguage stimuli (Vouloumanos and Werker, 2007), they can also discriminate phonemes of any human language in a categorical manner (Eimas, 1975). This ability develops from an early general ability to discriminate all phonemes, even from unfamiliar languages, into the ability to discriminate only native language phonemes toward the end of the first year of life (Werker and Tees, 1984).

Given that early sensitivities with regard to auditory linguistic input have been demonstrated, the question arises as to whether infants engage a similar neural substrate to adults when they listen to language. Neuroimaging studies during the first 3 months of life comparing language stimuli with acoustically matched nonlanguage stimuli have shown left-hemispheric activation for language in similar fronto-temporal brain regions as was found for adult language processing (Dehaene-Lambertz et al., 2002; Peña et al., 2003). Furthermore, the native language was specifically shown to already elicit left-hemispheric activation a few days after birth (Sato et al., 2012). Nonetheless, the infant language network may be at an immature stage. Perani et al. (2011) conducted a functional connectivity analysis in adults and infants, who were only 3 days old, focusing on 'cross-talk' between different brain regions while the participants listened to language, and reported marked differences between the adult and the infant brain. While the adult brain showed strong intrahemispheric fronto-temporal connectivities, the infant brain showed only interhemispheric connectivities (Perani et al., 2011). Yet, the involvement of similar brain regions in infants and adults speaks in favor of a biological preparedness to process familiar linguistic stimuli in a specialized network. The crucial signal invoking this network may be contained in the acoustic, rather than linguistic, properties of the speech signal. A study looking at the processing of temporal acoustic information reported more left-hemispheric contributions to the processing of fast acoustic modulations compared to slow modulations (Telkemeyer et al., 2009). This supports the notion that the speech signal, which is characterized by fast acoustic modulations between phonemes, triggers left-hemispheric processing mechanisms that might be innately specified. Yet, as there is evidence that there is already some fetal language perception (Smith et al., 2003) and memory for language (Decasper and Spence, 1986) in the last trimester of pregnancy, these results do not completely rule out that early left-hemispheric specializations are due to language experience in the womb.

In a similar way, later developmental steps taking place during the first year of life may also be subject to both biological maturation and linguistic experience. Recent studies with preterm infants, who show comparable postnatal linguistic experience but delayed neural maturation due to their lower gestational age, have provided interesting evidence concerning the interplay between the two. Peña et al. (2010, 2012) reported that the onset of linguistic abilities, such as the discrimination between the native language and a rhythmically similar language, and discrimination between nonnative phonemes, is determined by maturational status, rather than by experience. However, there are converse findings from a study on the perception of consonantal patterns, which reports clear effects of experience in preterm infants (Gonzalez-Gomez and Nazzi, 2012), supporting the idea that such 'experience vs maturation effects' are highly dependent on the exact linguistic feature under scrutiny. However, it is not only global neuronal maturation, as measured by gestational age, which can be used as a measure for the impact of maturation on the timing of language learning. Also, specific physiological parameters, for example hormonal measures, have been shown to affect basic language perception mechanisms. Friederici et al. (2008) were able to show that boys' discrimination of phonemes at the age of 2 months was influenced by their blood testosterone level, higher testosterone levels being linked to reduced ability to discriminate

phonemes. Such studies show that global measures, as well as specific maturational ones, can be related to the timing of the appearance of basic mechanisms of language perception.

Word Learning

Infants produce their first words around their first birthday and it is only later, at the age of 18–24 months, that the well-known sudden increase in vocabulary, the so-called vocabulary spurt occurs. Yet, indications of word recognition are reported much earlier. Infants at the age of only $4^1/_2$ months recognize highly frequent words such as 'mommy' and 'daddy' (Mandel et al., 1995) and from 6–9 months of age infants are able to grasp some aspects of referential meaning (Parise and Csibra, 2012; Bergelson and Swingley, 2012). The question is, which learning mechanisms allow young infants to detect and segment words from ambient speech and how are they rooted in the neural system? Two basic kinds of cues have been suggested to possibly assist segmentation of words. These are statistical and acoustic cues. Statistical learning of word boundaries, for example, the use of transitional probabilities between syllables to segment words has been demonstrated to start from the age of 7 months, in studies using behavioral measures (Saffran et al., 1996; Thiessen and Saffran, 2003), and to start from the age of 2 days, in studies using electrophysiological measures (Teinonen et al., 2009). Although the phenomenon of statistical word segmentation is undisputed, it is still unclear what role it plays in word learning during natural language acquisition. Recent findings show that word forms that have been segmented via statistical learning are preferentially used for the mapping of meaning (Estes et al., 2007; Hay et al., 2011). This supports the idea that statistical learning has the potential to play a role during natural word learning: Because statistical learning is present from the onset of the language acquisition process, it might well be one of the driving forces of language learning. Future studies face the challenge of specifying the neural basis of statistical learning in early infancy.

Besides mere statistical cues, the acoustic features inherent in spoken language, such as word stress, have been shown to be important cues for word segmentation in infants (Jusczyk et al., 1999). Infants are sensitive to language-specific prosodic realizations of words from early on. For example, a study with 4–5-month-old French and German infants found that they already show differing electrophysiological brain responses as a function of language-specific word-stress patterns (Friederici et al., 2007). An intriguing related finding is the observation that infants' cry patterns seem to reflect their native language's stress pattern (Mampe et al., 2009), indicating that even newborns are sensitive to their language's stress pattern.

Despite all these findings about astonishing language abilities in early infancy, not much is known about the brain structures supporting early learning of word forms. Benavides-Varela et al. (2011, 2012) tested the learning of simple CVCV word forms (e.g., 'mita,' 'pelu') in a familiarization-test paradigm in 1–4-day-old infants, while they recorded their hemodynamic brain activation using near-infrared spectroscopy. The infants could retain the sound of a spoken word form for a couple of minutes, which was reflected in a decrease of activation in a widespread brain network including frontal, parietal, and temporal areas (Benavides-Varela et al., 2011). A subsequent study extended these findings by showing that neonates' early word form memories rely on vowel information and not on consonantal information (Benavides-Varela et al., 2012).

As a next step in word learning, infants are prompted to understand word forms as referents to the world surrounding them, for example, as labels for objects. Initially, it was assumed that infants can only do so at 9–10 months of age, once they start to grasp that speakers have referential intentions (Carpenter et al., 1998). However, Bergelson and Swingley (2012) found that even 6–7-month-old infants know the meaning of some words, because they looked longer at named than unnamed pictures. It has also been shown that after only a few presentations, children this age can map new word forms onto visual objects (Friedrich and Friederici, 2011; Junge et al., 2012), with prosodic information supporting this process (Shukla et al., 2011). Interestingly, Parise and Csibra (2012) were able to show that young children's word knowledge does not merely reflect audiovisual associations, but also referential understanding.

Together, these findings suggest that word-like linguistic units can be segmented and retained for some time starting at early infancy and that the brain may use domain-general statistical mechanisms in concert with language-specific acoustic cues to achieve this task. In the second half of their first year, infants start to grasp the referential meaning of words and use them as labels for the world surrounding them.

Grammar Learning

Production of grammatical utterances starts after the production of the first words. Yet, similar to word recognition, infants show sensitivity to properties of the language input coding grammatical aspects of language already within the first year of life. Those initial sensitivities may play a crucial role in detecting the basic grammatical patterns in the native language onto which grammatical functions are mapped at a later stage. As already pointed out, newborns are very sensitive to prosodic aspects of speech, such as word stress. Word stress is not confined to assisting word segmentation; it also systematically varies according to grammatical word class. Function words, such as 'the,' 'in,' and 'on,' differ acoustically from content words such as 'window,' 'head,' and 'pen,' with respect to stress and duration, for example. Newborns have been shown to be able to differentiate between the two word classes based on their acoustic differences (Shi et al., 1999). Not much later, at the age of 2 months, sentence intonation can be used to remember word order for a short time interval (Mandel et al., 1994).

Concurrently with their sensitivity to acoustic aspects of language, young infants are already sensitive to structural rules that go beyond mere word segmentation. Neonates were shown to process sequences that contain syllable repetitions (e.g., 'mu ba ba') differently from sequences without repetitions (e.g., 'mu ba ge') (Gervain et al., 2012, 2008). However, syllable repetitions may only play a minor role in natural language processing. Grammatical relations in natural language are often carried by dependencies between distant units of speech, so-called non-adjacent dependencies, as, for example, in 'she smiles.' Behavioral evidence has shown that the learning of

such nonadjacent dependencies occurs only after 1 year of life (Lany and Gomez, 2008; Gòmez and Maye, 2005). Electrophysiological studies have provided evidence that nonadjacent dependencies can be extracted from nonnative sentences, as well as from artificial syllable sequences, at a much earlier age, namely between 3 and 4 months of age (Mueller et al., 2012; Friederici et al., 2011). Interestingly, the ability to extract nonadjacent dependencies has been linked to the developmental status of basic auditory processing (i.e., pitch discrimination) in 3-month-old infants. Only infants who showed electrophysiological indication of more mature auditory processes showed evidence of nonadjacent dependency learning (Mueller et al., 2012). This finding underlines the scaffolding role of acoustic cues and auditory functions for the acquisition of grammatical structures. The reviewed studies have shown that infants process some grammatical aspects of language, so long as they are carried by salient auditory features of the input, from early on. What emerges during the first year of life is most probably a kind of a 'raw grammar' based on phonologically represented rules onto which syntactic functions, which guide thematic interpretation, can be mapped at a later stage during development.

There is an ongoing debate over the kind of learning mechanism needed to extract grammatical rules in language. Specifically, there is disagreement on the question whether one single associative/statistical learning mechanism suffices to explain both adjacent and nonadjacent dependency learning (Pacton and Perruchet, 2008; Seidenberg et al., 2002; Laakso and Calvo, 2011) or whether an additional rule-based mechanism has to be assumed (Peña et al., 2002; Endress et al., 2007; Marcus et al., 1999). It has been argued that some positional effects (Endress et al., 2007), as well as generalizations (Peña et al., 2002; Marcus et al., 1999), cannot be captured by statistical learning alone. Opposing this view, Pacton and Perruchet (2008) were able to show that in adults the joint focus of attention, which is thought to relate the to-be-learned elements to each other, is both a necessary and sufficient factor, irrespective of temporal contiguity of the elements. This speaks in favor of the more parsimonious assumption of a single associative learning mechanism. Whether such an attention-based mechanism is at work in very young infants remains to be demonstrated. If there is a separable rule-learning mechanism, it seems to be present at birth (Gervain et al., 2008).

What is clear is that the language learning mechanism(s) operate(s) in concert with developing perceptual abilities on the acoustically structured linguistic input. This leads to the emergence of quite sophisticated representations of linguistically relevant units and rules of the native language during the first year of life.

Neural Substrate for Language Development

The general idea that brain maturation is closely linked to language development has a long tradition, which has recently gained a new momentum as methods of brain imaging have started to be used extensively in studies of infancy and childhood (Lenneberg, 1967; Pulvermuller and Schumann, 1994; Thompson-Schill et al., 2009; Friederici, 2009; Penfield and Roberts, 1959; Amunts et al., 2003; Pujol et al., 2006; Paus et al., 1999). Proposals regarding the relation between brain maturation and language development can be roughly split up with respect to their explanatory focus. While some theories, especially the older ones, have drawn on brain maturation as an explanation for children outperforming adults as language learners (Penfield and Roberts, 1959; Lenneberg, 1967; Pulvermuller and Schumann, 1994; Thompson-Schill et al., 2009), the majority of brain imaging-based ideas assume that proceeding maturation of certain brain areas causes, or at least parallels, important developmental steps in language acquisition (Friederici, 2009; Dubois et al., 2008; Amunts et al., 2003; Paus et al., 1999). If one accepts the idea that language recruits a specific neural network, which most researchers do, it logically follows that the maturation of that network has an impact on language development. Yet, showing the relation between structural maturation and language processing mechanisms is not trivial because the former cannot be experimentally manipulated. For this reason, experimental evidence is necessarily correlative, at best. As our present review focuses on language development during the first year of life, we will outline what is known about structural brain development in language-relevant areas at and around birth, although brain maturation continues well into adolescence and even adulthood. Up to a few decades ago, brain structure could only be visualized in postmortem studies. Recently developed brain imaging methods have made it possible to quantify properties of white and gray matter *in vivo* and noninvasively in participants of different age groups. While sophisticated analyses of postmortem brains continue to contribute important information about brain maturation, *in vivo* measurements have made the study of structural brain maturation easier than ever been before.

In the mature brain, a mainly left-hemispheric perisylvian network, including frontal, anterior, and posterior temporal brain areas, hosts language processing. These cortical areas are connected via two major fiber pathways: one running dorsally, connecting the pars opercularis of Broca's area, the motor cortex, and the mid-posterior temporal cortex, and the other running ventrally, connecting the pars triangularis of Broca's area with temporal cortical areas (cf Friederici and Gierhan, in press; for review, Gierhan, in press). In adults, one can find a hemispheric asymmetry with the language-dominant left hemisphere showing enhanced white matter density (Pujol et al., 2002; Catani et al., 2007). Coming from a developmental perspective, we will discuss two questions: the maturation of language-relevant cortical brain areas and pathways, and hemispheric lateralization during development. As pointed out in Section Discrimination of the Sounds of Language, perisylvian language areas have been shown to be selectively active during language processing from shortly after birth, which is suggestive of the language system of the brain already being in place before the onset of extensive linguistic experience. Yet, in order to fully characterize the biological basis of language, it is important to know whether the functional and behavioral sensitivities are building on a mature and specialized brain structural basis or, alternatively, whether the structural basis develops later with increasing linguistic experience.

Frontal cortical regions reach adult levels of maturation relatively late, as anatomical studies using postmortem brains

have shown (Huttenlocher and Dabholkar, 1997; Amunts et al., 2003). Thus, it came as a surprise when a study found that frontal cortices were activated while infants who were only 3 months old listened to language (Dehaene-Lambertz et al., 2002) and these activations could even be observed in younger infants (e.g., Perani et al., 2011). This indicates that the relative neuroanatomical immaturity does not make the frontal cortices functionally silent, but suggests that they may not yet serve the same function that they do in adults. The auditory cortex, which shows much faster maturation compared to the frontal cortex, also continues to mature until late childhood (Moore and Linthicum, 2007). Brain imaging studies that quantify brain maturation based on anatomical brain scans sometimes come to different conclusions than postmortem studies. One of the basic processes of brain maturation, myelination (i.e., the progressive coverage of axons with an insulating sheath of glia cells), seems to proceed at a similar speed in temporal and frontal white matter during the first year of life, but fiber bundles connecting different brain regions mature at a different speed (Pujol et al., 2006). Interestingly, the arcuate fascicle, that connects Broca's area in the inferior frontal gyrus and the posterior temporal cortex, develops relatively late. The study of Pujol et al. (2006) is one of the few studies of early development that shows a correlation with a measure of linguistic performance, in this case vocabulary size. The vocabulary spurt coincided with reaching a high percentage of myelination in temporo-frontal areas, which leads to speculation about the relation between perisylvian myelination and the course of language development.

A recent MRI study quantifying gray matter maturation (including, for example, myelination, and axonal and dendritic growth) in 1–4-month-old infants reported correlated maturation of posterior temporal cortex and Broca's area, with the temporal cortex lagging behind (Leroy et al., 2011). The significant correlation between those gray matter results and previous discoveries concerning white matter data (Dubois et al., 2009) led the authors to conclude that the dorsal language pathway is mature from early on (Leroy et al., 2011), although the frontal portion of the pathway was not identifiable. Another study, which focused on the dorsal pathway as a whole, was able to show that the fiber bundle that anatomically connects posterior temporal regions to Broca's area is not yet present in newborns (Perani et al., 2011), but another dorsal fiber tract that connects the posterior temporal cortex with premotor cortex is. Moreover, with respect to functional connectivities, the aformentioned study by Perani et al. (2011) demonstrated that infants, in contrast to adults, showed no intrahemispheric long-range functional connections. This finding strongly suggests that, despite the functional responsiveness of the single parts of the language network at and shortly after birth, the anatomical and functional connectivities within the network are far from adult-like and still have to be established during the course of language development.

What does the relative immaturity of the connections to the frontal cortex mean for language development? Possibly, as has been suggested above, it means that language learning in early infancy is a rather sensory-guided enterprise. The auditory cortex has been shown to be able to extract quite complex and even abstract auditory patterns (Bendixen et al., 2008), and it can be reasonably assumed that many of the early language learning abilities that have been outlined above can be managed by the auditory cortex alone. Efficient interplay between frontal and temporal language areas becomes especially important when syntactically complex sentences have to be processed (Friederici, 2009). However, this ability is not fully developed until later childhood, because even 6-year-olds (Schipke et al., 2012) and 7-year-olds (Dittmar et al., 2008) have difficulties processing, for example, object-initial sentences. In this context, it is interesting to note that in 7-year-olds the arcuate fascicle and the superior longitudinal fascicle, connecting Broca's area and the temporal cortex, are still not fully matured (Brauer et al., 2011). This is relevant in so far as this fiber connection has been shown to support the processing of syntax (Wilson et al., 2011).

One of the basic features of the brain's language areas is their structural and functional lateralization. As lateralized functional activation has been reported to exist early on, whether this builds on an already lateralized anatomical substrate is an important question. Lateralization clearly undergoes maturation. The left–right asymmetry in Broca's area, as measured by cytoarchitectonic profiles, was shown to be fully adult-like only at the age of 11 years (Amunts et al., 2003), whereas left-lateralized white matter maturation in the dorsal pathway has been shown across childhood (Paus et al., 1999). Despite these findings of an extended period of lateralized maturation, a left–right asymmetry in the gray matter was shown for both the frontal and the temporal language regions, even in preterm infants (Dubois et al., 2010), suggesting an anatomical prerequisite of lateralization that might set the basis for later functional lateralization. This corresponds to earlier postmortem studies on gray matter lateralization that reported asymmetries in temporal language areas during the last trimester of gestation (Chi et al., 1977).

These findings show that gross anatomical properties of some white matter pathways for speech and language, and some degree of lateralization, are already present from birth. However, the protracted maturation, especially of the dorsal pathway, may underlie some of the differences observed in language learning between infants and adults. It has been suggested that infants learn in a much more bottom-up, sensory-guided way than adults, who instead engage prefrontally guided processes of cognitive control. For infants, this might be, with respect to the final outcome, the more efficient way to learn conventional patterns such as language (Thompson-Schill et al., 2009). Although there is not much direct evidence for this hypothesis, the anatomical substrate found at and after birth supports this notion, because the relative immaturity of the fiber pathways speaks for initial sensory-guided language development.

Conclusion

The infant brain is predisposed to learn language starting from birth and even earlier. Behavioral and neurophysiological methods have revealed infants' astounding sensitivities to linguistic stimuli, as well as impressive abilities to learn aspects of words and grammar from early on. While important neural

characteristics of the language network, such as functional specification of fronto-temporal language regions and hemispheric specialization, are present from early on, they continue to mature in important aspects. Although the whole language network is active during language processing in early infancy, the relative immaturity of the dorsal fiber pathway might contribute to the observation that the earliest language learning processes happen in a bottom-up, sensory-guided manner.

See also: First Language Acquisition, Linguistic Theory of; Functional Brain Imaging of Language Processes; Human Development, Theories of; Interhemispheric Interaction in Language and Cognitive Processes; Language Acquisition; Language Processing, Functional Magnetic Resonance Imaging of; Neural Basis of Language Development; Prenatal and Infant Development: Overview, Current Trends, Future Directions; Self-Development in Childhood and Adolescence; Speech Perception; Speech Production, Neural Basis of.

Bibliography

Amunts, K., Schleicher, A., Ditterich, A., Zilles, K., 2003. Broca's region: cytoarchitectonic asymmetry and developmental changes. Journal of Comparative Neurology 465, 72–89.

Benavides-Varela, S., Gomez, D.M., Macagno, F., Bion, R.A.H., Peretz, I., Mehler, J., 2011. Memory in the neonate brain. PLoS One 6 (11), e27497.

Benavides-Varela, S., Hochmann, J.R., Macagno, F., Nespor, M., Mehler, J., 2012. Newborn's brain activity signals the origin of word memories. Proceedings of the National Academy of Sciences of the United States of America 109, 17908–17913.

Bendixen, A., Prinz, W.G., Horvath, J., Trujillo-Barreto, N.J., Schroger, E., 2008. Rapid extraction of auditory feature contingencies. NeuroImage 41, 1111–1119.

Bergelson, E., Swingley, D., 2012. At 6–9 months, human infants know the meanings of many common nouns. Proceedings of the National Academy of Sciences of the United States of America 109, 3253–3258.

Brauer, J., Anwander, A., Friederici, A.D., 2011. Neuroanatomical prerequisites for language functions in the maturing brain. Cerebral Cortex 21, 459–466.

Carpenter, M., Nagell, K., Tomasello, M., 1998. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. Monographs of the Society for Research in Child Development 63 (4), serial no. 255.

Catani, M., Allin, M.P.G., Husain, M., Pugliese, L., Mesulam, M.M., Murray, R.M., Jones, D.K., 2007. Symmetries in human brain language pathways correlate with verbal recall. Proceedings of the National Academy of Sciences of the United States of America 104, 17163–17168.

Chi, J.G., Dooling, E.C., Gilles, F.H., 1977. Left-right asymmetries of temporal speech areas of human fetus. Archives of Neurology 34, 346–348.

Decasper, A.J., Spence, M.J., 1986. Prenatal maternal speech influences newborns perception of speech sounds. Infant Behavior and Development 9, 133–150.

Dehaene-Lambertz, G., Dehaene, S., Hertz-Pannier, L., 2002. Functional neuroimaging of speech perception in infants. Science 298, 2013–2015.

Dittmar, M., Abbot-Smith, K., Lieven, E., Tomasello, M., 2008. German children's comprehension of word order and case marking in causative sentences. Child Development 79, 1152–1167.

Dubois, J., Benders, M., Lazeyras, F., Borradori-Tolsa, C., Leuchter, R.H.V., Mangin, J.F., Huppi, P.S., 2010. Structural asymmetries of perisylvian regions in the preterm newborn. NeuroImage 52, 32–42.

Dubois, J., Dehaene-Lambertz, G., Perrin, M., Mangin, J.F., Cointepas, Y., Duchesnay, E., Le Bihan, D., Hertz-Pannier, L., 2008. Asynchrony of the early maturation of white matter bundles in healthy infants: quantitative landmarks revealed noninvasively by diffusion tensor imaging. Human Brain Mapping 29, 14–27.

Dubois, J., Hertz-Pannier, L., Cachia, A., Mangin, J.F., Le Bihan, D., Dehaene-Lambertz, G., 2009. Structural asymmetries in the infant language and sensori-motor networks. Cerebral Cortex 19, 414–423.

Eimas, P.D., 1975. Auditory and phonetic coding of cues for speech – discrimination of [R-L] distinction by young infants. Perception and Psychophysics 18, 341–347.

Endress, A.D., Dehaene-Lambertz, G., Mehler, J., 2007. Perceptual constraints and the learnability of simple grammars. Cognition 105, 577–614.

Estes, K.G., Evans, J.L., Alibali, M.W., Saffran, J.R., 2007. Can infants map meaning to newly segmented words? Statistical segmentation and word learning. Psychological Science 18, 254–260.

Friederici, A.D., 2009. Pathways to language: fiber tracts in the human brain. Trends in Cognitive Sciences 13, 175–181.

Friederici, A.D., Friedrich, M., Christophe, A., 2007. Brain responses in 4-month-old infants are already language specific. Current Biology 17, 1208–1211.

Friederici, A.D., Gierhan, S.M.E., 2013. The language network. Current Opinion in Neurobiology 23, 250–254.

Friederici, A.D., Mueller, J.L., Oberecker, R., 2011. Precursors to natural grammar learning: preliminary evidence from 4-month-old infants. PLoS One 6, e17920.

Friederici, A.D., Pannekamp, A., Partsch, C.J., Ulmen, U., Oehler, K., Schmutzler, R., Hesse, V., 2008. Sex hormone testosterone affects language organization in the infant brain. NeuroReport 19, 283–286.

Friedrich, M., Friederici, A.D., 2011. Word learning in 6-month-olds: fast encoding-weak retention. Journal of Cognitive Neuroscience 23, 3228–3240.

Gervain, J., Berent, I., Werker, J.F., 2012. Binding at birth: the newborn brain detects identity relations and sequential position in speech. Journal of Cognitive Neuroscience 24, 564–574.

Gervain, J., Macagno, F., Cogoi, S., Peña, M., Mehler, J., 2008. The neonate brain detects speech structure. Proceedings of the National Academy of Sciences of the United States of America 105, 14222–14227.

Gierhan, S.M.E., in press. Connections for auditory language in the human brain. Brain and Language.

Gòmez, R.L., Maye, J., 2005. The developmental trajectory of nonadjacent dependency learning. Infancy 7, 183–206.

Gonzalez-Gomez, N., Nazzi, T., 2012. Phonotactic acquisition in healthy preterm infants. Developmental Science 15, 885–894.

Hay, J.F., Pelucchi, B., Estes, K.G., Saffran, J.R., 2011. Linking sounds to meanings: infant statistical learning in a natural language. Cognitive Psychology 63, 93–106.

Huttenlocher, P.R., Dabholkar, A.S., 1997. Regional differences in synaptogenesis in human cerebral cortex. Journal of Comparative Neurology 387 (2), 167–178.

Junge, C., Kooijman, V., Hagoort, P., Cutler, A., 2012. Rapid recognition at 10 months as a predictor of language development. Developmental Science 15, 463–473.

Jusczyk, P.W., Houston, D.M., Newsome, M., 1999. The beginnings of word segmentation in English-learning infants. Cognitive Psychology 39, 159–207.

Laakso, A., Calvo, P., 2011. How many mechanisms are needed to analyze speech? A connectionist simulation of structural rule learning in artificial language acquisition. Cognitive Science 35, 1243–1281.

Lany, J., Gomez, R.L., 2008. Twelve-month-old infants benefit from prior experience in statistical learning. Psychological Science 19, 1247–1252.

Lenneberg, E., 1967. Biological Foundations of Language. Wiley, New York.

Leroy, F., Glasel, H., Dubois, J., Hertz-Pannier, L., Thirion, B., Mangin, J.F., Dehaene-Lambertz, G., 2011. Early maturation of the linguistic dorsal pathway in human infants. Journal of Neuroscience 31, 1500–1506.

Mampe, B., Friederici, A.D., Christophe, A., Wermke, K., 2009. Newborns' cry melody is shaped by their native language. Current Biology 19, 1994–1997.

Mandel, D.R., Jusczyk, P.W., Kemler Nelson, D.G., 1994. Does sentential prosody help infants organize and remember speech information. Cognition 53, 155–180.

Mandel, D.R., Jusczyk, P.W., Pisoni, D.B., 1995. Infants recognition of the sound patterns of their own names. Psychological Science 6, 314–317.

Marcus, G.F., Vijayan, S., Bandi Rao, S., Vishton, P.M., 1999. Rule learning by seven-month-old infants. Science 283, 77–80.

Mehler, J., Jusczyk, P.W., Lambertz, G., Halsted, N., Bertoncini, J., Amiel-Tison, C., 1988. A precursor of language acquisition in young infants. Cognition 29, 143–178.

Moore, J.K., Linthicum, F.H., 2007. The human auditory system: a timeline of development. International Journal of Audiology 46, 460–478.

Mueller, J.L., Friederici, A.D., Männel, C., 2012. Auditory perception at the root of language learning. Proceedings of the National Academy of Sciences of the United States of America 109, 15953–15958.

Nazzi, T., Bertoncini, J., Mehler, J., 1998. Language discrimination by newborns: toward an understanding of the role of rhythm. Journal of Experimental Psychology. Human Perception and Performance 24, 756–766.

Pacton, S., Perruchet, P., 2008. An attention-based associative account of adjacent and nonadjacent dependency learning. Journal of Experimental Psychology: Learning, Memory and Cognition 34, 80–96.

Parise, E., Csibra, G., 2012. Electrophysiological evidence for the understanding of maternal speech by 9-month-old infants. Psychological Science 23, 728–733.

Paus, T., Zijdenbos, A., Worsley, K., Collins, D.L., Blumenthal, J., Giedd, J.N., Rapoport, J.L., Evans, A.C., 1999. Structural maturation of neural pathways in children and adolescents: in vivo study. Science 283, 1908–1911.

Peña, M., Bonatti, L.L., Nespor, M., Mehler, J., 2002. Signal-driven computations in speech processing. Science 298, 604–607.

Peña, M., Maki, A., Kovacz, D., Dehaene-Lambertz, G., Koizumi, H., Bouquet, F., Mehler, J., 2003. Sounds and silence: an optical topography study of language recognition at birth. Proceedings of the National Academy of Sciences of the United States of America 100, 11702–11705.

Peña, M., Pittaluga, E., Mehler, J., 2010. Language acquisition in premature and full-term infants. Proceedings of the National Academy of Sciences of the United States of America 107, 3823–3828.

Peña, M., Werker, J.F., Dehaene-Lambertz, G., 2012. Earlier speech exposure does not accelerate speech acquisition. Journal of Neuroscience 32, 11159–11163.

Penfield, W., Roberts, L., 1959. Speech and Brain Mechanisms. Athenaeum, New York.

Perani, D., Saccuman, M.C., Scifo, P., Anwander, A., Spada, D., Baldoli, C., Poloniato, A., Lohmann, G., Friederici, A.D., 2011. Neural language networks at birth. Proceedings of the National Academy of Sciences of the United States of America 108, 16056–16061.

Pujol, J., Lopez-Sala, A., Deus, J., Cardoner, N., Sebastian-Galles, N., Conesa, G., Capdevila, A., 2002. The lateral asymmetry of the human brain studied by volumetric magnetic resonance imaging. NeuroImage 17, 670–679.

Pujol, J., Soriano-Mas, C., Ortiz, H., Sebastian-Galles, N., Losilla, J.M., Deus, J., 2006. Myelination of language-related areas in the developing brain. Neurology 66, 339–343.

Pulvermuller, F., Schumann, J.H., 1994. Neurobiological mechanisms of language-acquisition. Language Learning 44, 681–734.

Saffran, J.R., Aslin, R.N., Newport, E.L., 1996. Statistical learning by 8-month-old infants. Science 274, 1926–1928.

Sato, H., Hirabayashi, Y., Tsubokura, H., Kanai, M., Ashida, T., Konishi, I., Uchida-Ota, M., Konishi, Y., Maki, A., 2012. Cerebral hemodynamics in newborn infants exposed to speech sounds: a whole-head optical topography study. Human Brain Mapping 33, 2092–2103.

Schipke, C.S., Knoll, L.J., Friederici, A.D., Oberecker, R., 2012. Preschool children's interpretation of object-initial sentences: neural correlates of their behavioral performance. Developmental Science 15, 762–774.

Seidenberg, M.S., MacDonald, M.C., Saffran, J.R., 2002. Does grammar start where statistics stop? Science 298, 553–554.

Shi, R.S., Werker, J.F., Morgan, J.L., 1999. Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. Cognition 72, B11–B21.

Shukla, M., White, K.S., Aslin, R.N., 2011. Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. Proceedings of the National Academy of Sciences of the United States of America 108, 6038–6043.

Smith, S.L., Gerhardt, K.J., Griffiths, S.K., Huang, X.Y., Abrams, R.M., 2003. Intelligibility of sentences recorded from the uterus of a pregnant ewe and from the fetal inner ear. Audiology and Neuro-otology 8, 347–353.

Teinonen, T., Fellman, V., Näätänen, R., Alku, P., Huotilainen, M., 2009. Statistical language learning in neonates revealed by event-related brain potentials. BMC Neuroscience 10, 21.

Telkemeyer, S., Rossi, S., Koch, S.P., Nierhaus, T., Steinbrink, J., Poeppel, D., Obrig, H., Wartenburger, I., 2009. Sensitivity of newborn auditory cortex to the temporal structure of sounds. Journal of Neuroscience 29, 14726–14733.

Thiessen, E., Saffran, J., 2003. When cues collide: use of stress and statistical cues to word boundaries by 7–9-month-old infants. Developmental Psychology 39, 706–716.

Thompson-Schill, S.L., Ramscar, M., Chrysikou, E.G., 2009. Cognition without control: when a little frontal lobe goes a long way. Current Directions in Psychological Science 18, 259–263.

Vouloumanos, A., Werker, J.F., 2007. Listening to language at birth: evidence for a bias for speech in neonates. Developmental Science 10, 159–164.

Werker, J.F., Tees, R.C., 1984. Cross-language speech-perception – evidence for perceptual reorganization during the 1st year of life. Infant Behavior and Development 7, 49–63.

Wilson, S.M., Galantucci, S., Tartaglia, M.C., Rising, K., Patterson, D.K., Henry, M.L., Ogar, J.M., DeLeon, J., Miller, B.L., Gorno-Tempini, M.L., 2011. Syntactic processing depends on dorsal language tracts. Neuron 72, 397–403.

Biomarkers in Demographic Research

Thomas McDade, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The integration of biological measures into demographic research bridges the gap between field- and clinic-based approaches to understanding human health, and it has the potential to transform the study of human biology. This article discusses the benefits, as well as potential costs, of collecting biomarker data in large-scale, social science surveys. It reviews methodological options for field-based biomarker data collection, and highlights recent applications to demonstrate their utility for promoting innovative research into the causes and consequences of human biological variation.

Introduction

Social, cultural, economic, and other environmental factors are widely recognized as critical determinants of health and disease, and the causes and consequences of variation in health have long been topics of central interest to demography. But disciplinary traditions, and challenges associated with collecting biological data from large numbers of people, are such that the vast majority of demographic research on health relies on vital records or self-report methods. In contrast, biomedical research employs sophisticated measures in controlled clinical or laboratory settings, but typically relies on small clinic-based samples and rarely evaluates social contexts beyond standard measures of socioeconomic status or self-reported health behaviors.

Integrating biological measures into demographic research bridges this gap, and promotes efforts to understand the biological pathways through which social, economic, and community factors shape population health (Seeman and Crimmins, 2001). The expansion of methodological options for collecting biological samples in nonclinical settings has greatly facilitated this effort, and encouraged a growing number of social scientists to include biological measures in large-scale surveys (Finch et al., 2001; Weinstein et al., 2008).

The objective of this article is to discuss the benefits, as well as potential costs, of integrating biomarkers into demographic research. The types of biological data that can be collected in community-based settings, as well as methods for collecting them, are briefly reviewed. A handful of demographic studies are discussed in order to showcase the potential of integrating biological measures into large-scale, social science surveys.

Why Biomarkers?

Biomarkers are directly measured traits that provide information on the functioning of biological systems. They cast light on normal biological processes, as well as pathological processes that indicate the presence of clinically identifiable disease, or an increased risk of disease in the future. Many biomarkers require the collection and subsequent analysis of biological specimens (e.g., blood, saliva, urine), while others rely on physical and functional measures (e.g., blood pressure, anthropometry, grip strength).

The vast majority of research on human biology and health is conducted in clinical or laboratory settings, or with focused epidemiological cohorts. Indeed, much of what is assumed to be known about human biology does not even involve humans, and is based on research with animal models. By collecting biological data from large, representative samples in diverse community-based settings, investigators can better understand the population distribution of health and disease, shed light on the individual, household, and community-level variables that account for this distribution, and identify trends that can be used to predict future patterns of morbidity and mortality. Several conceptual models have been proposed to facilitate our understanding of the biological pathways through which contextual and behavioral factors shape population health, and the measurement of biomarkers is central to this effort (Seeman and Crimmins, 2001; Glass and McAtee, 2006; McDade et al., 2007). Applications representing particular theoretical, disciplinary, or practical interests will focus on a subset of these pathways, with much higher resolution on the exposures and outcomes of primary interest.

There are at least five reasons why direct measurement of biomarkers has the potential to advance population-based research on human biology and health. First, self-reports of health (general health, as well as more specific reports of symptoms, diagnoses, and functional limitations), as well as information contained in vital and clinical records, have comprised the primary sources of data for health-related research in demography. These sources provide information on the distribution of important health outcomes, but biomarkers are necessary to illuminate the specific biological pathways through which socioeconomic, demographic, and psychosocial factors lead to these outcomes. In addition, biomarkers can help identify which aspects of psychosocial and physical environments are most damaging to health, as well as point toward resiliency factors that buffer subsets of individuals from the damaging effects of these exposures. Conversely, biomarkers may provide insight into the extent to which biological processes shape individual life course trajectories and inform selection into various social and physical environments.

Second, biomarkers provide direct information on pathophysiological processes that contribute to the development of disease, before clinically diagnosable problems are evident. For

example, prospective studies of cardiovascular disease (CVD) indicate that relative levels of CVD biomarkers (e.g., blood pressure, lipids) tend to remain stable, or 'track' from childhood into adulthood (Berenson et al., 1995; Li et al., 2004). While biomarker measurement earlier in life may not uncover many clinical cases of CVD, attention to predisease pathways will identify individuals most at risk for the future development of disease. Along these lines, biomarkers can help identify which physiological pathways in particular (e.g., aspects of neuroendocrine, cardiovascular, metabolic, and/or immune function) are most involved in the health outcomes of interest.

Third, biomarkers typically represent more direct and objective measures of health, and are therefore not susceptible to many of the limitations and biases associated with more subjective self-reported health measures. Biological measures typically do not rely on participants' ability to report on relevant aspects of their health, or their willingness to share this information. Biomarkers may be useful for validating self-reports of health, but most often their utility comes from offering access to embodied information that is below the threshold of perception, but that is nonetheless meaningful for physical health. This is a particular advantage for settings where accurate information on self-reported health may be especially difficult to obtain, as in research with children, or across international settings where linguistic and/or cultural factors may contribute to variation in the perception, experience, and/or reporting of health (Hahn, 1995; Kleinman, 1986).

Fourth, the methodological options for field-based biomarker research are rapidly expanding (see below), further encouraging the collection of biological data from large numbers of participants in diverse community-based settings. Larger, more representative samples increase the generalizability and potential policy impact of research findings, and help identify subgroups of individuals, or subsets of environments, that merit special attention. Population-based biomarker research also encourages an epistemological shift that reframes the study of human biology as a holistic, integrative endeavor that foregrounds contextual factors as potentially important contributors to variation in human physiological function and health across the life course (Stinson et al., 2012).

Lastly, the integration of biological measures into demographic research promotes productive collaboration among social, life, and biomedical scientists. Health is a complex, multidimensional phenomenon that articulates with human experience in dynamic ways over the life course. Collective efforts that bridge disciplinary boundaries can be significant sources of innovation that advance our understanding of human biology and health, as well as the psychosocial, behavioral, and cultural contexts within which it is embedded.

While the scientific payoff to including biomarkers in demographic research is potentially very high, important caveats should be kept in mind, particularly for investigators involved in designing data collection efforts. First, biological measures add to respondent burden, and may impose additional risks to research participation. Burdens and risks vary across methods of sample collection, and in most field-based applications they are relatively minimal. But even minimal risks need to be taken seriously, and for some participants the actual or perceived risk may be higher. It is therefore important to consider the implications of biomarker protocols for sample recruitment and retention. It is worth noting, though, that for many research participants burdens associated with biomarker collection may seem less onerous or problematic than the psychological discomfort associated with disclosure of embarrassing or otherwise sensitive information. For example, in the Add Health study (discussed below), 94% of participants were willing to provide a finger-stick blood sample, but only 93% would report their income.

Second, biomarkers can be costly, and they add to the logistical challenges associated with data collection. Some measures, like height or waist circumference, can be relatively easily implemented by survey interviewers. The collection of finger-stick blood or saliva samples can also be done by interviewers, but only after adequate training in sample collection and handling to make sure collection is done safely and effectively. Biological samples also need to be handled and transported in ways that maintain sample integrity prior to analysis in the lab, which adds to the logistical complexity and costs of data collection. And once in the lab, analysis costs typically range from approximately $10 to 30 per biomarker per sample, which can add up quickly, particularly for demographic surveys with large numbers of participants.

Lastly, as with any human subjects research involving potentially sensitive information, biomarkers raise important ethical issues that require careful consideration (Botkin, 2001). Strategies for addressing these issues should be informed by the biomarkers being measured, the populations involved in the research, and the requirements of institutional review boards, which can vary substantially.

Methodological Challenges and Innovations

Demographic studies often have to trade off numbers of participants against depth of measurement, particularly with respect to biological measures. Few field-based studies will be able to achieve the resolution that is possible in controlled clinical or laboratory settings, but the rapidly expanding range of 'field-friendly' biomarker methods is bridging this gap. This section reviews the methods and measures that have been applied in demographic research, and their potential contribution for illuminating biological processes of relevance to population health. More comprehensive reviews on this topic provide more details on these methods, as well as references to validation studies and prior applications (McDade et al., 2007, Lindau and McDade, 2007).

Anthropometric and Functional Measures

Anthropometric measures of body size and composition have been used for more than a century to capture energy balance, nutritional history, and risk for diseases associated with under- and overnutrition (World Health Organization, 1995; Gibson, 2005). Measures like height, weight, skinfold thicknesses, limb lengths, and waist and arm circumferences can be taken quickly with portable equipment at minimal cost, using standardized methods and reference data that enhance reliability, comparability, and interpretation.

Grip strength – a related functional measure of muscle strength – has recently been incorporated into several population-based studies, particularly studies of health and aging. Grip strength can be assessed with a small handheld dynamometer, and has been shown to predict general health, as well as bone density, mobility, physical function, and long-term mortality. Additional functional measures that can be relatively easily implemented in the field, and that predict physical decline and health, include the chair stand and timed get-up-and-go exercises.

Anthropometric measures like body mass index and waist circumference provide integrated measures of energy balance, but accelerometry and heart rate monitoring can be used for more precise estimates of physical activity and energy expenditure on an hourly or daily basis. Accelerometers are of the size of wristwatches that are worn on the wrist, ankle, or waist, and heart rate monitors are worn on the chest. Both methods record the frequency, intensity, and duration of physical activity as participants go about their normal routines, which may be particularly useful for research on obesity and sleep.

Pulmonary function can be assessed in nonclinical settings using portable spirometry devices that provide measures of expiratory volume, expiratory flow, and lung capacity. Measures of lung function are used clinically to diagnose and monitor asthma and obstructive pulmonary diseases, and in population-based settings impaired lung function predicts functional decline and mortality and may therefore be of relevance to studies of health and aging.

Impairments in hearing and vision affect quality of life and predict increased mortality risk. Self-reports are poor indicators of hearing, and audiometry provides more accurate measures of hearing loss that have been successfully implemented in population-based studies. Field-based assessments of visual acuity using standardized charts are feasible, but face challenges associated the need for sufficient space and lighting during implementation.

Blood Pressure

Arterial blood pressure provides an example of a biomarker with direct clinical relevance that can be applied to field-based research with relative ease. Blood pressure is measured clinically to diagnose hypertension, and to predict subsequent risk of coronary heart disease, stroke, dementia, and kidney failure. Smoking, physical activity, diet, and psychosocial stressors are all determinants of systolic and diastolic blood pressure. Noninvasive measures of blood pressure using manual or automated sphygmomanometers can be implemented by interviewers in field-based settings, with results that approximate the clinical standard. Measurements are typically taken two or three times with participants sitting comfortably, and results averaged for a measure of resting systolic and diastolic blood pressure (the first reading is discarded in some protocols). Pulse rate is often recorded along with blood pressure, and can also be used as a measure of cardiovascular risk.

Ambulatory monitoring of blood pressure is also possible, using portable battery-powered instruments worn by participants as they go through their normal daily routines. Frequent measures over a 24-h period can provide a more accurate measure of average blood pressure, assessment of blood pressure variability, differences in blood pressure across the day and night, and information on associations between blood pressure and specific events, contexts, or emotional states.

Measures Involving Biological Specimens

Anthropometric measures, physical exams, and functional tests provide valuable information on human biological function and health, but many important biomarkers are accessible only in biological specimens. In most cases, blood collected via venipuncture represents the gold standard for biomarker measurement, but the costs, invasiveness, and logistical challenges associated with venipuncture are major impediments to blood collection in field-based research. Fortunately, there are alternatives: finger-stick whole blood, saliva, urine, and hair all provide opportunities for collecting biological specimens in demographic surveys. While each type of specimen has its own challenges and limitations, in certain applications these methods may facilitate biomarker measurement in studies where it would not otherwise be possible.

Dried blood spots – drops of capillary whole blood collected on filter paper following a simple prick of the finger – represent a viable alternative to venipuncture that has recently been incorporated into a large number of population-based studies. Dried blood spots were originally developed in the 1960s for neonatal screening programs, and over the past 20 years they have been increasingly used to facilitate blood collection for research purposes in diverse community-based settings. Sample collection is relatively straightforward, and can be implemented in field settings by nonmedical interviewers. Once collected, samples are allowed to dry and can be stacked for storage and shipment, and most biomarkers remain stable for days, if not weeks, at room temperature. The logistics, burdens, and costs of sample collection and transport are greatly reduced in comparison with venipuncture, lowering the barriers to blood collection in large-scale studies. Protocols for over 100 biomarkers have been validated for use with dried blood spots, including important indicators of endocrine, immune, reproductive, and metabolic function, as well as measures of nutritional status and infectious disease (McDade et al., 2007).

Finger-stick blood sampling can also be used for the onsite analysis of biomarkers with small, portable 'point-of-care' devices. As with blood spot sampling, a sterile lancet is used to collect blood from the finger, and a drop of blood is placed into a cartridge or cuvette that is then inserted into the instrument. Results are available seconds or minutes later. Hemoglobin has been widely measured using this approach, and point-of-care devices have also been developed for the measurement of glucose, HbA1c, C-reactive protein (CRP), lipids, blood chemistry, or liver function. A single finger prick can be used to provide capillary whole blood for onsite analysis, as well as for collection on filter paper for subsequent analysis in the laboratory.

Saliva is an attractive alternative to blood sampling since collection is noninvasive, participants can collect samples by themselves during their normal daily routines, and repeat sampling is possible. For these reasons saliva sampling has

been successfully implemented in a large number of experimental, naturalistic, and population-based studies (Adam and Kumari, 2009). However, salivary biomarkers are typically of interest only if they represent circulating concentrations in blood, and the number of biomarkers for which this is the case is limited. Cortisol – a steroid hormone that is centrally involved in the body's response to stress – can be quantified in saliva and is the focus of studies investigating biological stress responses in relation to contextual factors and downstream health outcomes. Methods for measuring aspects of reproductive function, immune function, infectious disease, and drug use have also used saliva.

As with saliva, many biomarkers that otherwise require blood can be measured in urine samples. Urine collection has been successfully implemented in population-based studies with younger and older participants, men and women, and in a variety of cultural settings. While sample collection is relatively straightforward, many biomarkers require urine samples to be refrigerated for frozen soon after collection and during transport to the lab. Urine samples can be analyzed for biomarkers of renal function, stress hormone activity, reproductive hormones, sexually transmitted infections, toxic substances, and drug metabolites.

Analyzing cortisol in hair samples has emerged as an alternative to saliva and urine, with the benefit of providing a retrospective measure of cortisol exposure (Sauve et al., 2007; Kirschbaum et al., 2009). Human hair grows approximately 1 cm per month, and samples close to head represent a cumulative record of recent exposure. More distal samples provide information on exposures up to several months earlier. For many research purposes such an integrated measure of cortisol may be appropriate, although a disadvantage of this approach is that it cannot capture aspects of diurnal cortisol rhythms (e.g., cortisol awakening response, slope of decline across the day) that have been shown to provide valuable information in addition to overall cortisol level (Adam and Kumari, 2009).

Recent Applications of Biomarkers in Population-Based Surveys

The number of sociodemographic studies incorporating biological measures has grown rapidly, in large part due to methodological innovations that have reduced the burdens and costs associated with biomarker data collection. A comprehensive review of these studies is well beyond the scope of this article, but a handful of studies are highlighted to demonstrate the range of applications and their potential to advance scientific understandings of the dynamic relationships among society, biology, and health. In addition to their scientific contributions, studies such as these are excellent sources of information on how to collect biomarker data since their protocols are often publicly available.

National Health and Nutrition Examination Survey

Since the 1960s, the National Center of Health Statistics has implemented the National Health and Nutrition Examination Survey (NHANES), representing a major series of community-based studies designed to evaluate the health and nutritional status of adults and children in the United States (http://www.cdc.gov/nchs/nhanes.htm). Since 1999 NHANES has become a continuous research program that includes a nationally representative sample of about 5000 people every year, with changes in specific health and nutrition measures across surveys depending on emerging issues in population health.

Interviews to collect demographic, socioeconomic, and self-reported health information are conducted in participants' homes. Physical exams and biological specimen collection are implemented using specially designed mobile examination centers that travel to selected research locations around the country. By bringing these centers into community locations NHANES is able to collect an impressive set of biological measures that have clinical as well as public health significance. Recent surveys have collected whole blood, plasma, serum, urine, hair, DNA, saliva, and nasal and vaginal swabs, and have implemented laboratory analyses for measures of micro- and macronutritional status, cardiometabolic function, reproductive function, infectious disease, and exposure to environmental contaminants.

Given the size, scope, and national representation of NHANES, data are often used to determine the prevalence of risk factors and major diseases across sociodemographic groups in the US, and to inform policy accordingly. The study is widely regarded as the gold standard for assessing health in community-based settings, and findings are applied as benchmark comparisons for clinical, epidemiological, and demographic studies collecting overlapping measures. NHANES does not include a comprehensive panel of psychosocial or contextual measures, and the serial cross-section research design precludes longitudinal, within-person analyses of change over time. Nonetheless, the in-depth measurement of a wide range biomarkers in a large representative sample is a major strength, and the data make important contributions to demographic research on population health in the US.

The Health and Retirement Study

The Health and Retirement Study (HRS) is a nationally representative, longitudinal study of aging in the US that was launched in 1992. The original sample consisted of 12 652 individuals born between 1931 and 1941, and the study now includes more than 26 000 Americans over the age of 50 years (http://hrsonline.isr.umich.edu/). Surveys are implemented every 2 years, and focus on physical and mental health, cognitive functioning, financial and employment status, retirement planning, social/family environments, health care access and usage, and insurance coverage.

The HRS first collected dried blood spots in 2003 from a subsample of participants, for the measurement of HbA1c. In 2006, a larger panel of measures was collected during face-to-face interviews with a random one-half of the sample, followed by the other half of the sample in 2008. Samples and measures included dried blood spots, saliva for DNA, blood pressure, anthropometrics (height, weight, waist circumference), and a series of physical performance tests (e.g., grip strength, balance, peak flow, timed walk). Blood samples were

analyzed for HbA1c, total cholesterol, high-density lipoprotein cholesterol, CRP, and cystatin C. The collection of biomarker data is also part of the 2010 and 2012 surveys.

Health is a defining feature of the HRS, and in many ways it reflects – and has helped define – a new generation of population-based, social science surveys that include objective biological measures in addition to more commonly used self-report measures (Weir, 2007). Studies of aging and health in particular have been important in advancing integrative biosocial survey research (Seeman and Crimmins, 2001), and the HRS experience has encouraged parallel studies of aging in several countries in Europe and Asia to harmonize data collection efforts to enhance comparability.

National Longitudinal Study of Adolescent Health

The National Longitudinal Study of Adolescent Health (Add Health) is a longitudinal study of the social, behavioral, and biological linkages defining trajectories of health from adolescence to adulthood in the US (Harris, 2011). Wave 1 was conducted in 1994–95 with a nationally representative sample of 20 745 adolescents in grades 7–12, with follow-up surveys in 1996, 2001–02, and 2007–08. In-home interviews collect information on self-reported health and health behaviors, access to health care, peer and family contexts, psychosocial stressors and mental health, educational aspirations and attainments, employment experience, socioeconomic resources, and romantic relationships. The study also includes information from a parental interview at wave 1, neighborhood and school information from administrative sources, and peer/friendship network data. A strategic oversample of identical and fraternal twins, biological siblings, and nonrelated siblings is an innovative feature of the dataset that allows for analyses investigating genetic and contextual influences on health outcomes.

Initially, Add Health included only self-reported and parent-reported health information. Wave 3 introduced the collection of saliva and urine for HIV and sexually transmitted infection testing, and for DNA genotyping of candidate polymorphisms in the genetic subsample. Biomarker collection was expanded during wave 4 to cover a wider range of measures relevant to health in young adulthood. Anthropometric measures included height, weight, and waist circumference, blood pressure was recorded, and dried blood spots were collected and subsequently analyzed for cardiometabolic (total cholesterol, high-density lipoprotein, HbA1c, glucose) and immune (CRP, Epstein–Barr virus antibodies) function. All participants were also asked to provide buccal cell samples for DNA genotyping. Biomarker collection protocols were implemented in the home as part of the wave 4 interview.

The Add Health study has the potential to make important contributions to population health due to its large, nationally representative sample, the collection of longitudinal data that includes a wide range of measures of social contextual factors during important periods of development, and the integration of biological measures that represent the function of key physiological systems. In addition, the study is unique in focusing on the transition to young adulthood, which is a time when the cohort is settling into new work, family, and social roles that matter to health, and a time when predisease pathways can be measured to gain insight into how these contexts may influence future disease risk.

Moving to Opportunity

In addition to adding objective information on health and biological function in observational studies, biomarkers provide unique opportunities to evaluate the impact of policy-relevant social experiments. The Moving to Opportunity (MTO) demonstration project represents an example of this type of application. The study was initiated in 1994 with the goal of understanding how residential contexts influence employment opportunities, educational attainments, income, and general well-being. The focus of MTO was on families in Baltimore, Boston, Chicago, Los Angeles, and New York living in the most impoverished urban neighborhoods. Families with children in public housing were randomly assigned to one of three groups: (1) rent subsidy vouchers that could only be used if the family moved to a low-poverty neighborhood; (2) traditional vouchers with no restrictions on residence; and (3) a control group that was not offered new assistance.

In 2008–10, 10–15 years after randomization, participants were followed up to assess the long-term impacts of the intervention on health, as well as other outcomes. The panel of biological measures included height, weight, waist circumference, and blood pressure, as well as dried blood spot samples for the analysis of HbA1c, CRP, Epstein–Barr virus antibodies, total cholesterol, and high-density lipoprotein. Intention-to-treat analyses indicate that assignment to the low-poverty voucher group resulted in a 13–19% reduction in obesity in comparison with the control group, and a 22% relative reduction in diabetes, as defined by HbA1c greater than 6.5% (Ludwig et al., 2011).

The magnitude of these effects is large, and the experimental study design increases confidence in concluding that they are indeed causal. The inclusion of biomarkers in social experiments identifies the specific biological pathways and health outcomes that are affected by contextual factors like neighborhood poverty, which may inform future interventions. In addition, biological measures can provide more objective outcome measures since subjective health measures are vulnerable to biased reporting that is associated with assignment to particular experimental groups. Lastly, implementation of biological outcome data may be particularly effective in mobilizing the attention of policy makers and informing interventions around important social issues.

Study on Global AGEing and Adult Health

Biological measures have been successfully implemented into population surveys all around the world, and in fact, many of the 'field-friendly' biomarker methods that are now being integrated into demographic studies in the US were originally developed to meet the challenges of data collection in remote community-based settings (Campbell, 1994; Worthman and Stallings, 1997; Ellison, 1988). Anthropometric measures like height and weight have been central parts of global surveillance efforts to identify undernutrition, particularly in childhood (World Health Organization, 1995; Gibson, 2005), while the ongoing Demographic and Health

Surveys (http://www.measuredhs.com/) are increasingly turning to the collection of biological specimens to detect micronutrient deficiencies and infectious diseases. Biological measures in lower-income nations tend to focus on outcomes related to undernutrition and infectious disease, but rising rates of overweight/obesity globally – even in the context of undernutrition and high burdens of infectious disease (Boutayeb, 2006) – are expanding the range of biomarkers that are relevant in international research.

The Study on Global AGEing and Adult Health, (SAGE) led by the World Health Organization, represents an ambitious cross-national data collection effort that will provide useful comparative data for demographic studies of health in the US (http://www.who.int/healthinfo/sage/cohorts/en/index.html). A baseline cohort of more than 60 000 adults age 18 years and older was established in 2002–04, with nationally representative samples of participants in six countries: China, Ghana, India, Mexico, Russia, and South Africa. Wave 1 was implemented in 2007–10, and included self-reported measures of health, as well as a series of tests of physical performance (timed walk, spirometry, vision, grip strength), cognitive function, blood pressure, anthropometric measures, and dried blood spot samples. Follow-up surveys are planned for 2013 and 2015.

Biomarkers can provide practical information for surveillance efforts and public health interventions in under-resourced lower-income nations. For research purposes, comparisons across populations living in diverse social, cultural, and ecological settings are powerful tools for generating hypotheses regarding the role of contextual factors in shaping human physiological function and health. Developmental plasticity and ecological sensitivity are defining characteristics of human biology, and as a result, global environmental variation produces substantial variation in human biology around the world (Stinson et al., 2012). Health research conducted exclusively in affluent industrialized settings like the US cannot capture the full range of socioecological environments that may be important for gaining insights into the development and function of critical physiological systems. A field-based, comparative approach to research on human biology in diverse communities around the world therefore has the potential to advance our understanding of human biological variation, with implications for understanding health in the US and internationally (McDade, 2012).

Conclusion

The future is bright for research that integrates conceptual and methodological tools from the biomedical and social/behavioral sciences. Demography provides a solid foundation for modeling patterns of fertility, morbidity, and mortality over the life course in diverse populations around the world. A new generation of research is taking advantage of methodological innovations that facilitate the direct, objective measurement of biological processes in naturalistic settings, greatly expanding the range of possibilities for future research. By moving beyond vital records and self-reported measures, this research has the potential to have an immediate impact on demography, and to answer the call for more integrative, transdisciplinary approaches to population health (Weinstein et al., 2008). In the long run, this approach will transform the study of human development, biology, and health in relation to the diverse social and physical contexts within which human lives are embedded.

See also: Aging and Health in Old Age; Bioethics; Demographic Measurement: General Issues and Measures of Fertility; Early Life Influences on Health and Mortality in Adulthood; Ethics for Biomedical Research Involving Humans: International Guidelines; Health Economics; Health Outcomes, Assessment of; Human Development and Health; Life Course Epidemiology; Medical Sociology; Mortality, Biodemography of; Public Health as Social Science; Social Epidemiology; Social Integration, Social Networks, and Health; Socioeconomic Status and Health.

Bibliography

Adam, E.K., Kumari, M., 2009. Assessing salivary cortisol in large-scale, epidemiological research. Psychoneuroendocrinology 34, 1423–1436.

Berenson, G.S., Wattigney, W.A., Bao, W., Srinivasan, S.R., Radhakrishnamurthy, B., 1995. Rationale to study the early natural history of heart disease: the Bogalusa Heart Study. American Journal of Medical Science 310, S22–S28.

Botkin, J.R., 2001. Informed consent for the collection of biological samples in household surveys. In: Population, C.O., Finch, C.E., Vaupel, J.W., Kinsella, K. (Eds.), Cells and Surveys: Should Biological Measures be Included in Social Science Research? National Academy Press, Washington, DC, pp. 276–302.

Boutayeb, A., 2006. The double burden of communicable and non-communicable diseases in developing countries. Transactions of the Royal Society of Tropical Medicine and Hygiene 100, 191–199.

Campbell, K.L., 1994. Blood, urine, saliva and dip-sticks: experiences in Africa, New Guinea, and Boston. Annals of the New York Academy of Sciences 709, 312–330.

Ellison, P., 1988. Human salivary steroids: methodological considerations and applications in physical anthropology. Yearbook of Physical Anthropology 31, 115–142.

Finch, C.E., Vaupel, J.W., Kinsella, K.G., 2001. National Research Council. Cells and Surveys: Should Biological Measures Be Included in Social Science Research? National Academies Press, Washington, DC.

Gibson, R.N., 2005. Principles of Nutritional Assessment. Oxford University Press, Oxford.

Glass, T.A., McAtee, M.J., 2006. Behavioral science at the crossroads in public health: extending horizons, envisioning the future. Social Science & Medicine 62, 1650–1671.

Hahn, R.A., 1995. Sickness and Healing: An Anthropological Perspective. Yale University Press, New Haven.

Harris, K.M., 2011. Design Features of Add Health. http://www.cpc.unc.edu/projects/addhealth/data/guides/design%20paper%20WI-IV.pdf.

Kirschbaum, C., Tietze, A., Skoluda, N., Dettenborn, L., 2009. Hair as a retrospective calendar of cortisol production-increased cortisol incorporation into hair in the third trimester of pregnancy. Psychoneuroendocrinology 34, 32–37.

Kleinman, A., 1986. Social Origins of Distress and Disease. Yale University Press, New Haven, CT.

Li, S., Chen, W., Srinivasan, S.R., Berenson, G.S., 2004. Childhood blood pressure as a predictor of arterial stiffness in young adults: the Bogalusa Heart Study. Hypertension 43, 541–546.

Lindau, S.T., McDade, T.W., 2007. Minimally invasive and innovative methods for biomeasure collection in population-based research. In: Weinstein, M., Vaupel, J.W., Wachter, K.W., Council, N.R. (Eds.), Biosocial Surveys. The National Academies Press, Washington, DC.

Ludwig, J., Sanbonmatsu, L., Gennetian, L., Adam, E., Duncan, G.J., Katz, L.F., Kessler, R.C., Kling, J.R., Lindau, S.T., Whitaker, R.C., McDade, T.W., 2011. Neighborhoods, obesity, and diabetes – a randomized social experiment. New England Journal of Medicine 365, 1509–1519.

McDade, T.W., 2012. Early environments and the ecology of inflammation. Proceedings of the National Academy of Sciences 109, 17281–17288.

McDade, T.W., Williams, S., Snodgrass, J.J., 2007. What a drop can do: dried blood spots as a minimally invasive method for integrating biomarkers into population-based research. Demography 44, 899–925.

Sauve, B., Koren, G., Walsh, G., Tokmakejian, S., Van Uum, S.H., 2007. Measurement of cortisol in human hair as a biomarker of systemic exposure. Clinical and Investigative Medicine 30, E183–E191.

Seeman, T.E., Crimmins, E., 2001. Social environment effects on health and aging: integrating epidemiologic and demographic approaches and perspectives. Annals of the New York Academy of Sciences 954, 88–117.

Stinson, S., Bogin, B., O'rourke, D. (Eds.), 2012. Human Biology: An Evolutionary and Biocultural Perspective. Wiley-Blackwell, New York.

Weinstein, M., Vaupel, J.W., Wachter, K.W., 2008. National Research Council. Biosocial Surveys. National Academies Press, Washington, DC.

Weir, D., 2007. Elastic powers: the integration of biomarkers into the health and retirement study. In: Weinstein, M., Vaupel, J.W., Wachter, K.W., Council, N.R. (Eds.), Biosocial Surveys. The National Acadamies Press, Washington, DC.

World Health Organization, 1995. World Health Organization: Physical Status: The Use and Interpretation of Anthropometry Report of a WHO Expert Committee. World Health Organization Technical Report Series, 14.

Worthman, C.M., Stallings, J.F., 1997. Hormone measures in finger-prick blood spot samples: new field methods for reproductive endocrinology. American Journal of Physical Anthropology 104, 1–22.

Biomedical Sciences, History and Sociology of

Robin W Scheffler, Program in Science, Technology, and Society, Massachusetts Institute of Technology, MA, USA
Bruno J Strasser, University of Geneva, Geneva, Switzerland

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A. Cambrosio and P. Keating, volume 2, pp. 1222–1226, © 2001, Elsevier Ltd.

Abstract

The biomedical sciences have become a prominent site of the production of knowledge about health, disease, and behavior. Here, we provide an overview of the genealogy of the concept of biomedicine, the philosophical and methodological assumptions upon which it came to rest, an overview of the growing scholarship in the history and sociology of biomedicine, and a vision of some future directions for scholarship.

Introduction

The History and Sociology of the Biomedical Sciences (HSBS) might seem to be a logical point of overlap between the history of biology and medicine. However, in the recent past, historiographic trends conspired to divert the attention of both these fields from their common ground. During the 1980s and 1990s many historians of medicine turned away from accounts focusing on great ideas and heroic doctors. Instead, they emphasized the social and ideological facets of medicine, especially through a close focus on the experiences and agency of patients, a 'history from below.' While salutary in many respects, this 'social turn' left out the production of biomedical knowledge in the clinic as well as the laboratory (Cambrosio and Keating, 2001). Historians of biology, meanwhile, generally directed their attention toward the development of experimental, especially molecular, knowledge in the laboratory, taking for granted its 'application' in the clinic (de Chadarevian and Kamminga, 1998).

Since the 1990s, HSBS diverged from these trends in seeking to understand the practices and knowledge produced at the intersection of the laboratory and clinic, but until recently this area of scholarship has been a patchwork rather than a field. The years since the first edition of this encyclopedia have been fruitful for studies of HSBS (Strasser, 2014). The emphasis on practice in HSBS has produced many studies whose narratives unite the laboratory and clinic, demonstrating the entangled relationship of laboratory and clinical practice well as the links between both of these sites and society as a whole (Creager, 2002; de Chadarevian, 2002; Gaudillière, 2002; Keating and Cambrosio, 2003; Landecker, 2007; Lindee, 2005; Löwy, 2010; Rader, 2004; Wailoo and Pemberton, 2006). As Ilana Löwy (2011) recently argued, "biomedicine studies" promise to develop conceptual tools capable of integrating the social, cultural, and material accounts of practice, explore the definition of normal and pathological bodies, and understand the role of sites such as the clinic and activities such as computation in the production of knowledge about health and disease.

Defining Biomedicine

Genealogy of the Term

Biomedicine as a term first arose in the early twentieth century. The terms 'biomedicine' (and 'biomedical') have different origins and trajectories in English, German ('Biomedizin'), and French ('biomédecine'), but very similar meanings today. 'Biomedical' first appeared in the writing of American and British authors in the 1920s, and was followed a decade later by 'biomedicine' (or 'bio-medicine') (Keating and Cambrosio, 2003: Chapter 3). From the onset, biomedicine and biomedical research referred to a kind of medicine that was closely associated with experimentation and the laboratory rather than the experience of physicians and the clinic. The American Illustrated Medical Dictionary defined biomedicine as "clinical medicine based on the principles of physiology and biochemistry" rather than the expertise physicians gained through clinical observation (Dorland, 1923: 172).

After the Second World War, biomedicine increasingly came to signify 'modern medicine' based on the expertise of the laboratory as opposed to 'traditional medicine' associated with the clinical experience and acumen of individual physicians. Although the laboratory had served as a rallying standard for medical reformers since the late nineteenth century, it was only after 1945 that laboratory-based medicine, identified with biomedicine, began to have a strong impact on clinical practice (Warner, 1991; Cunningham and Williams, 1992; Quirke and Gaudillière, 2008). As biomedicine grew socially and politically to a dominant practice, it became associated with 'orthodox medicine' (opposed to 'alternative medicine'), 'scientific medicine' (opposed to 'clinical' or 'traditional medicine'), and 'Western medicine' (opposed to 'non-Western' or 'indigenous medicine'). A key assumption of biomedicine, to which we will return, is that the difference between the normal and the pathological is one of degree, not of kind. This in itself represents a relatively recent development in the history of medicine and a key point of entry for biomedical approaches to disease.

Philosophical Themes

The study of health and disease in the laboratory draws upon a number of implicit and explicit philosophical assumptions. Chief among these is 'reductionism,' held to be a critical tenet of biology in opposition to more holistic clinical approaches. Reductionism consists of at least three distinct claims. First, an ontological claim that a whole organism is 'nothing more' than the sum of its parts; second, an epistemological claim that the organism is best explained by reference to its parts; and finally a methodological claim that the organism is best investigated

by the study of its parts (Sarkar, 1998). The reductionist position is in opposition to various 'holistic' interpretations of disease, which encompass a range of positions from nineteenth century Romanticism to twenty-first century systems biology and ecology (Lawrence and Weisz, 1998; Rosenberg, 2007: Chapter 8). In discussions of biomedicine, many of these claims have been associated with the rise of molecular approaches to biology and medicine after 1945 (Morange, 2000). However, this recent wave of 'molecularization' is only one episode in a longer history of reductionist approaches to biology (de Chadarevian and Kamminga, 1998).

Modeling is another key assumption of biomedical research closely associated with reductionism. The use of models or exemplars is found in both medicine and biology. In medicine, physicians often use exemplary clinical case descriptions for understanding particular illnesses (Anderson, 2013). In biology, researchers focus on model organisms or *in vitro* model systems because they believe in the universality of living processes. From the assumption that similar processes in different organisms operate by the same biological mechanisms, as claimed by the nineteenth century French physiologist Claude Bernard, experiments aiming to elucidate a particular process in humans may be carried out using organisms or systems which are better suited to laboratory conditions. (Prominent model organisms are mice, flies, worms, yeasts, bacteria, and viruses. See Endersby (2007).) Although they are the product of human artifice, these systems are held to be 'model systems' or 'model organisms' which are capable of standing in or substituting for more complex phenomena (Rheinberger, 1999). As the molecular biologists Jacques Monod and François Jacob famously put it in 1961, "[What is] true of *E. coli* must also be true of Elephants" (Morange, 2000). These modeling practices have been immensely powerful for biological research and its applications to human diseases. The use of model organisms has become a dominant practice in biomedical research in the twentieth century (Creager et al., 2007).

Biomedicine: Grounded in Laboratory Science?

Although there is no universally agreed upon definition of 'biomedicine,' it is often conceptualized as the search for therapeutic innovations in the laboratory. In the early nineteenth century physicians such as German pathologist Rudolf Virchow had grounded the autonomy of medicine on the 'ontological' presumption that disease was caused by processes distinct from those taking place in the healthy body, but in the late nineteenth century, other physicians started to challenge this view (Temkin, 1977). Most notoriously, Claude Bernard advocated a 'physiological view' of disease, where pathology was a gradual deviation from normal biological processes. Medicine could thus be studied as a subset of physiology and other sciences. As he asserted in his Introduction to Experimental Medicine (1865) "for a man of science, there is no separate science of medicine or physiology, there is only a science of life…the true sanctuary of scientific medicine is the laboratory…every scientific physician should have, therefore, a physiology laboratory" (Bernard, 1957: 146–147). Even if Bernard's vision proved difficult to realize in practice, the study of 'normal' and 'pathological' states could both be pursued by the same laboratory-based means (Canguilhem, 1978). The pursuit of therapeutic aims in the laboratory received further support later in the nineteenth century from Louis Pasteur's and Robert Koch's innovations in bacterial culture and the rise of microbiological theories of disease etiology (Geison, 1995; Gradmann, 2009).

Under this understanding of biomedicine, laboratory research on basic biological mechanisms in almost any organism has potential medical relevance. This argument has been enthusiastically championed by molecular biologists since the mid-twentieth century. However, as historian of science Jean-Paul Gaudillière (2002) has shown, many of these laboratory researchers sought to distance themselves from the practices of clinical research even while claiming that they were contributing to it through their inquiries into the 'basic' or 'fundamental' process of life. The limitations of the distinction between 'basic' (laboratory) research and 'applied' (clinical) practice are immediately apparent in the case of biomedicine. Indeed, the production of knowledge about therapeutics always requires a clinical phase involving patients, where research and therapy take place at the same time on the same subjects (Löwy, 1996). We return to the epistemic questions raised by this observation below.

Biomedicine as Practice

Diagnosis and Etiology

Early work on biomedical practices sought to integrate biomedicine into the framework of a narrow range of laboratory practices rather than examining its diversity. More recent scholarship on biomedical research into the diagnosis and etiology of disease has questioned this assumption by following particular practices to open up our understanding of how biomedical knowledge is produced. By following the flow of materials and information between laboratory and clinic, for example, Keating and Cambrosio have suggested that biomedicine rests upon the construction of 'platforms' capable of coordinating the action in the laboratory with the clinic rather than the one-way application of laboratory studies to therapeutics. For example, the diagnosis of lymphoma through the development of immunological tests in the 1990s did not displace morphological analysis of bone marrow samples. Instead, teams of oncologists, technicians, and immunologists seeking to make a diagnosis had to consider the evidence from many forms of practice (Keating and Cambrosio, 2003: 4–7, 50–56).

The single most prominent feature of contemporary biomedical diagnosis is the increasing power attributed to genes. From cancer to political leaning and sex drive to criminality, genes have been invoked to explain a growing number of traits (Nelkin, 2004). However, the identification of genetic factors involved in the etiology of hereditary disease has often relied not only on molecular biology but also on long-term clinical observation, popular knowledge of heredity, and family histories. The history and experience of individual populations, such as the Pennsylvania Amish or Ashkenazi Jews, are a vital part of the production of genetic and genomic knowledge in the laboratory (Comfort, 2012; Gibbon et al., 2014; Lindee, 2005; Wailoo and Pemberton, 2006).

These practices have progressively blurred the boundaries between the normal and pathological and between health and illness. Disease, once defined by qualitative symptoms, is now more often defined in terms of quantitative metrics. For example, the diagnosis of cardiovascular disease or cancer has been transformed from distinct moments of illness into lifelong risks for healthy individuals, defined by ranges of 'normal' values for biomedical markers (tension, cholesterol, genes, antigens, etc.) rather than observable symptoms (Aronowitz, 1998; Greene, 2007).

Clinical Trials

Recent scholarship suggests that clinical research has been no less important as a site for the development of biomedical knowledge about disease. The single most important event in integrating hospitals and patients into the biomedical enterprise was the development of the randomized clinical trial (RCT) after the Second World War (Marks, 1997). Before the RCT, questions of drug safety and therapeutic efficacy were mainly adjudicated through the observation of patients by a relatively small number of experienced physicians. The efficacy of a particular treatment rested on the professional credibility of the physicians involved rather than on a standardized method. The adoption of statistical methods promised to provide an 'objective' means of carrying out these evaluations. When state agencies such as the US Food and Drug Administration started to take a more active role in regulating the drug market, they also promoted the widespread use of RCT as the primary standard for determining safety and efficacy (Carpenter, 2010; Daemmrich, 2004).

Advocates of RCT and some historians have assumed that the development of RCT provided an important basis for testing of the efficacy for human subjects of compounds developed in laboratory settings. However, it appears equally true that RCTs were an important site for the production of knowledge about diseases themselves. Ilana Löwy has shown that clinical trials in hospitals are no less effective in producing knowledge about the workings of diseases and their possible treatments than experiments in laboratories (Löwy, 1996). Other recent scholarship has argued that RCTs also produced knowledge about diseases themselves. In this view, clinical trials contribute to the production of biomedical knowledge and are the preferred site for defining and demonstrating the relationship between therapeutic agents and diseases. Moreover, scale and organization also played a vital role in the production of such knowledge. With thousands and sometimes tens of thousands of participants, RCTs during the 1970s and 1980s started to produce their own classifications of disease, such as types of breast cancer based on trial results rather than laboratory diagnostics (Keating and Cambrosio, 2012). This means that clinical trials may yield new definitions of diseases and target populations as well as more conventional verdicts on efficacy.

Risk and Enhancement

Biomedicine has blurred the boundary between the restoration of health and the enhancement of individuals. New methods of diagnosis have profoundly affected the mission of medicine and our understanding of what disease is. Rather than restoring health, biomedical treatments offer ways to improve an individual's physical and mental performance in an endless "pursuit of perfection" (Rothman and Rothman, 2003). Hormone replacement therapy, for example, started off as therapy for some of the most severe symptoms of menopause (which itself can hardly be characterized as a 'disease'), only to become a preferred treatment to keep the skin of postmenopausal women supple (Watkins, 2007). Enhancement and pharmaceutical marketing have intersected with the development of 'risk' interpretations of disease. Originating in the focus of public health on disease incidence in aggregate populations, the risk notion of disease has become integrated into biomedical discourse as a key concept through which individual patients understand their health or illness (Rothstein, 2003). By shifting its aim from the elimination of disease to the management of disease risks, biomedicine has opened the door to understanding health as the continuous pursuit of risk reduction (Aronowitz, 1998; Jones, 2013).

Biomedicine and Society: The Biomedical Complex

In addition to studies of practice, the HSBS has produced a body of work seeking to understand biomedicine's political, institutional, and social development. Scholarship here has focused on tensions arising from the sponsorship of laboratory biology by organizations with a commitment to public health rather than the creation of biological knowledge for its own sake. While originally consisting of universities and private foundations, the range of actors involved in the patronage of biomedical research expanded over the course of the twentieth century to include state public health authorities, patient advocacy groups, and various industries. Reconciling the actual activities of biomedical researchers with the aspirations of their patrons marks an ongoing process, one that yields rich material for studies of biomedicine's social relations in many contexts.

Biomedicine, Philanthropy, and the State

Philanthropic initiatives played a critical role in developing the intellectual and institutional agendas of biomedicine. At the beginning of the twentieth century, biomedicine's emphasis on laboratory research, especially biological research, accorded universities a central role in its development. National governments, by and large, gave modest support to biological and medical research. These early conditions amplified the influence of philanthropies, the one set of organizations that actively sought to support and develop biomedical research in the early twentieth century. Private foundations such as the Carnegie Institute of Washington or the Rockefeller Foundation supported the application of the biological sciences to understanding human health (Kohler, 1991). These foundations were able to exert considerable influence over the direction of biomedical research as a field, although the extent to which their influence altered the research agendas of individual scientists is still a matter of debate among historians. The Rockefeller Foundation, for example, supported a number of research projects, such as the physical chemist Linus Pauling's

study of the molecular basis of biological processes (which formed the basis of his later work on sickle-cell anaemia) hoping to create a new 'science of man' and possibly bring eugenics on a sound scientific basis (Kay, 1993). (For the view that Foundations played a major role, see Kay (1993); for the view that Foundations did not, see Abir-Am (2002).)

These foundations aimed at a better understanding of the biological basis of heath, disease, and behavior, but declined to support more 'holistic' approaches. Critically, philanthropic activity strongly supported the early rise of molecular biology at a time when its approaches fell outside traditional academic disciplines; this rise subsequently played a leading role in shaping biomedicine as a laboratory science. Just after the Second World War, the Rockefeller Foundation played a key role in supporting the reconstruction of European science by encouraging the mobility of researchers, the organization of international conferences, and crucially, the purchase of expensive scientific instruments for biomedical research (Gemelli et al., 1999).

After the Second World War, state patronage for academic research vastly expanded. Although not as expensive as 'big science' projects in physics or astronomy, the biomedical sciences were indebted to this patronage for their expansion. However, this patronage was not for science on its own terms. Health and medicine, like national security, emerged as areas where the state was expected to cultivate and guard the safety of its citizens. In the United States and Europe, in practice this meant providing support to university-based biological research (Appel, 2000; Quirke and Gaudillière, 2008; Wright, 1994: Chapter 1). Funding for research accordingly tended to focus on diseases rather than disciplines, as exemplified by the disease specific Institute structure of the US National Institutes of Health or initiatives such as the 'War on Cancer' of the 1970s (Patterson, 1987; Mukherjee, 2011). In some cases, such support benefited research on particular model organisms, such as mice, or the creation and promotion of particular methods for the study of biological processes, such as radioisotopes. This state patronage played a decisive role in the production of biomedical knowledge (Creager, 2013; Fujimura, 1996; Rader, 2004).

Patient/Disease Advocacy

The early historiography of biomedicine often regarded patients as passive objects of biomedical expertise. However, the growth of contemporary disease advocacy movements has encouraged scholarship addressing patients as active subjects in the production of biomedical knowledge. Following the model of public philanthropy, patient organizations have played an especially important role in this regard, influencing science policy and research agendas and providing substantial financial support for biomedical research on specific diseases. Most famously, the National Foundation for Infantile Paralysis, founded in 1938, raised money from the American public to support polio research, leading to the Salk vaccine in 1955 (Oshinsky, 2005). During the 1970s, the American Cancer Society played a dominant role as an advocate for the 'War on Cancer.' In the 1980s and 1990s, the fundraising telethons of the Association Française Contre les Myopathies supported the study of the genetic basis of muscular dystrophy, leading to the identification of the first gene related to Duchenne muscular dystrophy (Rabinow, 1999; Rabeharisoa and Callon, 2002). A significant shift in the orientation of patient organizations toward biomedical research occurred with the onset of the AIDS epidemic. As the sociologist Steven Epstein has shown, patient organizations such as the AIDS Coalition to Unleash Power (ACT UP) not only provided financial support for scientific research on AIDS, but also claimed a form of 'lay expertise' that contributed to the production of scientific knowledge about the disease (Epstein, 1996).

Biomedicalization

The development of the biomedical sciences has also been entangled with the expanding scope of medical authority. This is the latest episode in what scholars have referred to as the progressive 'medicalization' of society since the nineteenth century (Conrad, 2007). The concept of medicalization has recently been extended by scholars to the concept of 'biomedicalization,' a specific kind of medical transformation. Biomedicalization emphasizes technological and scientific intervention, based less on the 'clinical gaze' than the 'molecular gaze,' as well as the creation of new individual and collective identities ('biosocialities') grounded upon shared biomedical traits such as risk factors or disease genes (Clarke et al., 2010; Burri and Dumit, 2007; Rabinow, 1996; Rose, 2007; Lynch et al., 2010). These new etiologies and identities of diseases occupy contentious social ground because they suggest ways to diagnose and treat particular diseases while sidelining others. In particular, critics charge that genetic causes or predispositions have been granted exclusive status in these explanations, to the detriment of explanations examining the social, economic, environmental, or behavioral determinants of health (Krimsky and Gruber, 2013). Debates surrounding the causation of conditions such as obesity, autism, or fatigue continue to offer numerous further opportunities for tracing the process of biomedicalization.

Biomedicine and Political Economy

The biomedical sciences have been deeply shaped by the intellectual property regimes concerning drugs and biological materials. In Europe, and especially in France, drugs were originally excluded from the protections of intellectual property law because of their importance for human health. However, after the Second World War, the United Kingdom (1949), France (1959), and Germany (1965), for example, passed legislation authorizing the patenting of therapeutic substances (Gaudillière, 2008). According to historian of science Jean-Paul Gaudillière (2002), such legislation should be understood as part of post-war efforts by states to modernize by fostering scientific research in service of economic growth.

The extension of patent protection played a particularly significant role in stimulating the rise of the biotechnology industry in which new and experimental therapeutic technologies were based, not on chemical synthesis or extraction, but on the production of substances through genetically modified organisms. Although the focus on the therapeutic possibilities

of 'biological' substances was far from novel (von Schwerin et al., 2013), following the development of recombinant DNA technology between 1972 and 1974, it became theoretically possible to produce almost any protein in unlimited quantities with high purity once a corresponding gene had been isolated. The production of insulin and other therapeutically useful proteins was envisioned as a major market for the application of recombinant DNA technologies by biotechnology startups. The expansion of the biotechnology industry accelerated after the 1980 US Supreme Court decision in Diamond versus Chakrabarty established that a "live, human-made micro-organism is patentable subject matter." In 1988, the US Patent and Trademark Office extended this decision to include transgenic mammals, approving a patent filed by Harvard University for 'OncoMouse,' a mouse specifically modified to serve as a model organism for cancer research (Kevles, 1994).

This ruling, coupled with the greater ease of patenting state-supported research fostered by the Bayh-Dole Act in 1980, promoted a close relationship between the pharmaceutical industry and academic research in the United States. The pressure to produce economically valuable research has subsequently been criticized for distorting the research priorities of universities and scientists. The rhythm of biomedical discovery, for example, has become profoundly entangled with the promissory environment of financial markets (Hughes, 2011; Rajan, 2006). The impact of these policies on the direction of biomedical research marks an exciting area for the future study of biomedical practices and institutions within the history of innovation and intellectual property.

Future Opportunities for Scholarship

While the completion of the Human Genome Project in 2003 seemed to show that molecular biology had realized its ambition of providing explanations of health and disease in terms of discrete 'master molecules,' subsequent experience has suggested that the process is far more complicated (Reardon, 2005; Keller, 2010; Paul, 1998). The contemporary relationship between biology and medicine remains unsettled; a circumstance that offers rich ground for the further growth of scholarship in HSBS, particularly lines of work that overlap with disciplines such as anthropology in efforts to examine contemporary historical events (Lock and Nguyen, 2010). Following the practices of biomedicine beyond the axis of laboratory and clinic into society offers many opportunities for study by the social and behavioral sciences. In closing, we will suggest a few of these areas.

The emphasis in this article on the United States and Western Europe evident reflects the scope of scholarship in HSBS. Analyzing the development of biomedicine as a global phenomenon offers many opportunities. Most RCT for drug development, for example, now occur in developing rather than industrial countries (Petryna, 2009). Indeed, the global dimension of clinical trials cannot be regarded as a recent development, as it has been a longstanding feature of cancer research (Keating and Cambrosio, 2012). Biomedical definitions of disease often appear very different when they are examined in distinct populations or localities. For example, the biomedical idea that sickle-cell anemia is a 'molecular' disease has been complicated by examining how this genetic discourse was received by Senegalese patients suffering from the disease and their physicians (Fullwiley, 2011).

Tracing the circulation of materials – from chemicals to tissues – has provided a powerful tool for analyzing the practices and social relations of biomedicine. Following the history of individual drugs from the laboratory to market has been used successfully by a number of recent studies (Bud, 2007; Goodman and Walsh, 2001; Pieters, 2005; Rasmussen, 2008; von Schwerin et al., 2013; Watkins, 2007). In following these substances, these studies manage to integrate a number of the perspectives on laboratory, clinic, and society in biomedicine discussed above, especially through the state regulation of therapeutic substances (Carpenter, 2010; Gaudillière and Hess, 2012; Gradmann and Simon, 2010; Greene and Watkins, 2012). Research that follows the movement of materials has also suggested that the individual body is not always the appropriate scale for understanding the production, reception, or uses of biomedical knowledge. This insight opens ground for the fruitful collaboration between HSBS and environmental studies, to gain insight into how the causation of disease is understood in relationship to ecological and economic change from the study of areas as diverse as allergy, pollution, food additives, agriculture, or occupational safety (Mitman, 2007; Langston, 2010).

The changing scale of biomedical research, the means developed to create and process growing amounts of biological data ('big data biology'), and the increasing integration of biological and medical information offer numerous opportunities for further study. The role of computers in biomedical work, for example, has only recently started to receive historical attention, despite the ubiquity of information technology in current biomedical research (November, 2012; Stevens, 2013). The deep continuities between current modes of collecting, comparing, and computing data and earlier scientific practices, in natural history and the clinic for example, have also just begun to be explored (Strasser, 2012).

Finally, the relationship between neoliberalism and the biomedical sciences appears to be closely linked with the process of biomedicalization (Clarke et al., 2010; Cooper, 2008; Rajan, 2006). The development of lay expertise has gone hand in hand with the development of consumer markets for biomedical knowledge. Before the Food and Drug Administration restricted the disclosure of genetic information in 2013, direct-to-consumer companies such as 23andMe offered cheep genetic testing and broad promises for a "personalized medicine" (Prainsack and Vayena, 2013). How much this promise will hold true, and the extent to which it represents a departure from the universalist ideal that has been so important for the construction of biomedicine offer rich terrain for further study.

See also: Anthropology, Genomics, and Human Variation: National Roots; Biobanking: Ethical Issues; Bioethics: Genetics and Genomics; Community Genetics; Direct-to-Consumer Personal Genetic Testing; Ethical, Legal, and Social

Implications Program at the National Human Genome Research Institute; Eugenics as a Basis of Population Policy; Eugenics as an International Movement; Eugenics, History of; Familial Studies: Genetic Inferences; Genetic Engineering; Genetic Screening for Disease; Genetics and Forensics; Genetics and Indigenous Communities: Ethical Issues; Genetics and Society; Genetics and the Media; Genetics, Disease, and Reproduction; Genetics: Legal Aspects; Genetics: The New Genetics; Human Genome Diversity Project: History; Human Genome Project: History and Assessment; Human Genomics in Asia; New Genetics and Race.

Bibliography

Abir-Am, P.G., 2002. The Rockefeller Foundation and the rise of molecular biology. Nature Reviews Molecular Cell Biology 3, 65–70.

Anderson, W., 2013. The case of the archive. Critical Inquiry 39, 532–547.

Appel, T.A., 2000. Shaping Biology: The National Science Foundation and American Biological Research, 1945–1975. Johns Hopkins University Press, Baltimore, MD.

Aronowitz, R.A., 1998. Making Sense of Illness: Science, Society, and Disease, Cambridge History of Medicine. Cambridge University Press, Cambridge and New York.

Bernard, C., 1957[1865]. An Introduction to the Study of Experimental Medicine. Dover, New York.

Bud, R., 2007. Penicillin: Triumph and Tragedy. Oxford University Press, Oxford.

Burri, R.V., Dumit, J. (Eds.), 2007. Biomedicine as Culture: Instrumental Practices, Technoscientific Knowledge, and New Modes of Life. Routledge, New York.

Cambrosio, A., Keating, P., 2001. Biomedical sciences and technology: history and sociology. In: Baltes, P., Smelser, N. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Pergamon, Oxford, pp. 1222–1226.

Canguilhem, G., 1978. On the Normal and the Pathological. D. Reidel Pub. Co., Dordrecht.

Carpenter, D.P., 2010. Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA. Princeton University Press, Princeton.

de Chadarevian, S., Kamminga, H. (Eds.), 1998. Molecularizing Biology and Medicine: New Practices and Alliances, 1910s–1970s. Harwood Academic Publishers, Amsterdam.

de Chadarevian, S., 2002. Designs for Life: Molecular Biology after World War II. Cambridge University Press, Cambridge.

Clarke, A., Fosket, L., Ruth, J. (Eds.), 2010. Biomedicalization: Technoscience, Health, and Illness in the U.S. Duke University Press, Durham, NC.

Comfort, N.C., 2012. The Science of Human Perfection: How Genes Became the Heart of American Medicine. Yale University Press, New Haven.

Conrad, P., 2007. The Medicalization of Society: On the Transformation of Human Conditions into Treatable Disorders. Johns Hopkins University Press, Baltimore.

Cooper, M., 2008. Life as Surplus: Biotechnology and Capitalism in the Neoliberal Era. University of Washington Press, Seattle.

Creager, A.N.H., 2002. The Life of a Virus: Tobacco Mosaic Virus as an Experimental Model, 1930–1965. University of Chicago Press, Chicago.

Creager, A.N.H., Lunbeck, E., Wise, M.N. (Eds.), 2007. Science Without Laws: Model Systems, Cases, Exemplary Narratives, Science and Cultural Theory. Duke University Press, Durham.

Creager, A.N.H., 2013. Life Atomic: A History of Radioisotopes in Science and Medicine. University of Chicago Press, Chicago.

Cunningham, A., Williams, P. (Eds.), 1992. The Laboratory Revolution in Medicine. Cambridge University Press, Cambridge.

Daemmrich, A.A., 2004. Pharmacopolitics: Drug Regulation in the United States and Germany. University of North Carolina Press, Chapel Hill.

Dorland, W.A.N., 1923. The American Illustrated Medical Dictionary, twelfth ed. W.B. Saunders, Philadelphia & London.

Endersby, J., 2007. A Guinea Pig's History of Biology. Harvard University Press, Cambridge.

Epstein, S., 1996. Impure Science: AIDS, Activism, and the Politics of Knowledge, Medicine and Society. University of California Press, Berkeley.

Fujimura, J.H., 1996. Crafting Science: A Sociohistory of the Quest for the Genetics of Cancer. Harvard University Press, Cambridge, MA.

Fullwiley, D., 2011. The Enculturated Gene: Sickle Cell Health Politics and Biological Difference in West Africa. Princeton University Press, Princeton.

Gaudillière, J.-P., 2008. How pharmaceuticals became patentable: the production and appropriation of drugs in the twentieth century. History and Technology 24, 99–106.

Gaudillière, J.-P., 2002. Inventer La Biomédecine: La France, l'Amérique et la Production des Savoirs du Vivant, 1945–1965. La Découverte, Paris.

Gaudillière, J.-P., Hess, V. (Eds.), 2012. Ways of Regulating Drugs in the 19th and 20th Centuries. Palgrave Macmillan, London.

Geison, G.L., 1995. The Private Science of Louis Pasteur. Princeton University Press, Princeton.

Gemelli, G., Picard, J.-F., Schneider, W.H., 1999. Managing Medical Research in Europe: The Role of the Rockefeller Foundation: 1920s–1950s. CLUEB, Bologna.

Gibbon, S., et al., 2014. Breast Cancer Gene Research and Medical Practices. Routledge, London.

Goodman, J., Walsh, V., 2001. The Story of Taxol: Nature and Politics in the Pursuit of an Anti-Cancer Drug. Cambridge University Press, Cambridge.

Gradmann, C., 2009. Laboratory Disease: Robert Koch's Medical Bacteriology. Johns Hopkins University Press, Baltimore.

Gradmann, C., Simon, J., 2010. Evaluating and Standardizing Therapeutic Agents, 1890–1950. Palgrave Macmillan, New York.

Greene, J.A., 2007. Prescribing by Numbers: Drugs and the Definition of Disease. Johns Hopkins University Press, Baltimore.

Greene, J.A., Watkins, E.S., 2012. Prescribed: Writing, Filling, Using, and Abusing the Prescription in Modern America. Johns Hopkins University Press, Baltimore.

Hughes, S.S., 2011. Genentech: The Beginnings of Biotech. University of Chicago Press, Chicago.

Jones, D.S., 2013. Broken Hearts: The Tangled History of Cardiac Care. Johns Hopkins University Press, Baltimore.

Kay, L.E., 1993. The Molecular Vision of Life: Caltech, the Rockefeller Foundation, and the Rise of the New Biology. Oxford University Press, New York.

Keating, P., Cambrosio, A., 2003. Biomedical Platforms: Realigning the Normal and the Pathological in Late-Twentieth-Century Medicine. MIT Press, Cambridge, MA.

Keating, P., Cambrosio, A., 2012. Cancer on Trial: Oncology as a New Style of Practice. The University of Chicago Press, Chicago and London.

Keller, E.F., 2010. The Mirage of a Space Between Nature and Nurture. Duke University Press, Durham.

Kevles, D.J., 1994. Ananda Chakrabarty wins a patent: biotechnology, law, and society, 1972–1980. Historical Studies in the Physical and Biological Sciences 25, 111–135.

Kohler, R.E., 1991. Partners in Science: Foundations and Natural Scientists, 1900–1945. University of Chicago Press, Chicago.

Krimsky, S., Gruber, J., 2013. Genetic Explanations: Sense and Nonsense. Harvard University Press, Cambridge.

Landecker, H., 2007. Culturing Life: How Cells Became Technologies. Harvard University Press, Cambridge.

Langston, N., 2010. Toxic Bodies: Hormore Disruptors and the Legacy of DES. Yale University Press, New Haven.

Lawrence, L., Weisz, G., 1998. Greater Than the Parts: Holism in Biomedicine, 1920–1950. Oxford University Press, New York.

Lindee, M.S., 2005. Moments of Truth in Genetic Medicine. Johns Hopkins University Press, Baltimore.

Lock, M., Nguyen, V.-K., 2010. An Anthropology of Biomedicine. Wiley-Blackwell, Dordrecht.

Löwy, I., 1996. Between Bench and Bedside: Science, Healing, and Interleukin-2 in a Cancer Ward. Harvard University Press, Cambridge, MA.

Löwy, I., 2010. Preventive Strikes: Women, Precancer, and Prophylactic Surgery. Johns Hopkins University Press, Baltimore.

Löwy, I., 2011. Historiography of biomedicine: "bio," "medicine," and in between. Isis 102, 116–122.

Lynch, M., Cole, S.A., McNally, R., Jordan, K., 2010. Truth Machine: The Contentious History of DNA Fingerprinting. University of Chicago Press, Chicago.

Marks, H.M., 1997. The Progress of Experiment: Science and Therapeutic Reform in the Unites States, 1900–1990. Cambridge University Press, Cambridge.

Mitman, G., 2007. Breathing Space: How Allergies Shape Our Lives and Landscapes. Yale University Press, New Haven, CT.

Morange, M., 2000. A History of Molecular Biology. Harvard University Press, Cambridge, MA.

Mukherjee, S., 2011. The Emperor of All Maladies. Simon and Schuster, New York.

Nelkin, D., 2004. The DNA Mystique: The Gene as a Cultural Icon, second ed. University of Michigan, Ann Arbor.

November, J.A., 2012. Biomedical Computing: Digitizing Life in the United States. Johns Hopkins University Press, Baltimore.

Oshinsky, D., 2005. Polio: An American Story. Oxford University Press, Oxford.

Patterson, J.T., 1987. The Dread Disease: Cancer and Modern American Culture. Harvard University Press, Cambridge, MA.
Paul, D.B., 1998. The Politics of Heredity. Essays on Eugenics, Biomedicine and the Nature-Nurture Debate. State University of New York Press, New York.
Petryna, A., 2009. When Experiments Travel: Clinical Trials and the Global Search for Human Subjects. Princeton University Press, Princeton.
Pieters, T., 2005. Interferon: The Science and Selling of a Miracle Drug. Routledge, London.
Prainsack, B., Vayena, E., 2013. Beyond the clinic: "direct-to-consumer" genomic profiling services and pharmacogenomics. Pharmacogenomics 14, 403–412.
Quirke, V., Gaudillière, J.-P., 2008. The era of biomedicine: science, medicine, and public health in Britain and France after the Second World War. Medical History 52, 441–452.
Rabeharisoa, V., Callon, M., 2002. The involvement of patients' associations in research. International Social Science Journal 54, 57–63.
Rabinow, P., 1996. Making PCR: A Story of Biotechnology. University of Chicago Press, Chicago.
Rabinow, P., 1999. French DNA: Trouble in Purgatory. University of Chicago Press, Chicago, IL.
Rader, K., 2004. Making Mice: Standardizing Animals for American Biomedical Research, 1900–1955, first ed. Princeton University Press, Princeton.
Rajan, K., 2006. Biocapital: The Constitution of Postgenomic Life. Duke University Press, Durham.
Rasmussen, N., 2008. On Speed: The Many Lives of Amphetamine. New York University Press, New York.
Reardon, J., 2005. Race to the Finish: Identity and Governance in an Age of Genomics. Princeton University Press, Princeton, NJ.
Rheinberger, H.-J., 1999. Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube. Stanford University Press, Stanford, CA.
Rose, N.S., 2007. The Politics of Life Itself: Biomedicine, Power, and Subjectivity in the Twenty-First Century. Princeton University Press, Princeton, NJ.
Rosenberg, C.E., 2007. Our Present Complaint: American Medicine, Then and Now. Johns Hopkins University Press, Baltimore.
Rothman, D.J., Rothman, S., 2003. The Pursuit of Perfection: The Promise and Perils of Medical Enhancement. Pantheon Books, New York.
Rothstein, W.G., 2003. Public Health and the Risk Factor: A History of an Uneven Medical Revolution, Rochester Studies in Medical History. University of Rochester Press, Rochester, NY.
Sarkar, S., 1998. Genetics and Reductionism, Cambridge Studies in Philosophy and Biology. Cambridge University Press, Cambridge, UK and New York.
von Schwerin, A., Stoff, H., Wahrig, B., 2013. Biologics, a History of Agents Made from Living Organisms in the Twentieth Century. Pickering & Chatto, London.
Stevens, H., 2013. Life Out of Sequence: A Data-Driven History of Bioinformatics. University of Chicago Press, Chicago.
Strasser, B.J., 2012. Collecting nature: practices, styles, and narratives. Osiris 27 (1), 303–340.
Strasser, B.J., 2014. Biomedicine: Meanings, Assumptions, and Possible Futures. Swiss Science and Innovation Council, Bern.
Temkin, O., 1977. The Double Face of Janus and Other Essays in the History of Medicine. Johns Hopkins University Press, Baltimore.
Wailoo, K., Pemberton, S.G., 2006. The Troubled Dream of Genetic Medicine: Ethnicity and Innovation in Tay-Sachs, Cystic Fibrosis, and Sickle Cell Disease. Johns Hopkins University Press, Baltimore.
Warner, J.H., 1991. Ideals of science and their discontents in late nineteenth-century American medicine. Isis 82 (3), 454–478.
Watkins, E.S., 2007. The Estrogen Elixir: A History of Hormone Replacement Therapy in America. Johns Hopkins University Press, Baltimore.
Wright, S., 1994. Molecular Politics: Developing American and British Regulatory Policy for Genetic Engineering, 1972–1982. University of Chicago Press, Chicago.

Relevant Websites

Archives at Cold Spring Harbor Laboratory. http://library.cshl.edu/archives (accessed 03.05.14.).
The DNA Exchange. http://thednaexchange.com/ (accessed 03.05.14.).
Genetopia. http://genotopia.scienceblog.com/ (accessed 03.05.14.).
Medical Museion. http://www.museion.ku.dk/ (accessed 03.05.14.).
National Library of Medicine, History of Medicine Division. https://www.nlm.nih.gov/hmd/ (accessed 03.05.14.).
Office of History, National Institutes of Health. http://history.nih.gov/ (accessed 03.05.14.).
UCLA Institute for Society and Genetics. http://socgen.ucla.edu/ (accessed 03.05.14.).
Wellcome Library: The Library at Wellcome Collection. http://wellcomelibrary.org/ (accessed 03.05.14.).

Biosocial Approaches: Crime

Michelle A Coyne and John P Wright, University of Cincinnati, Cincinnati, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Over the past two decades, scholars have discovered a host of information regarding genes, brain activity, and other biological processes. Many of these biological discoveries are related to behavior. Biosocial approaches to crime examine the way that biological factors such as genes and the brain affect environmental factors and influence criminality over the life course. This article provides an overview of the evolution, main approaches, and contributions of biosocial criminology to linking the way that both nature and nurture influence criminality.

Introduction

The biosocial approach to understanding crime is relatively new, emerging only recently as a prominent and respected area of research. The rapid ascendancy of biosocial criminology owes much of its recognition to the sheer volume of evidence biosocial scholars have produced. Using a variety of genetically informed research designs, including twin studies, biosocial criminologists have produced a broad array of evidence that reveals and implicates the complex interplay between 'nature,' those biological and genetic factors that influence human development and behavior, and 'nurture,' often defined as our experiences with certain social variables, in the development of criminal behavior. The evidence linking biological factors to crime is now extensive. However, one of the major insights garnered from the biosocial approach is that criminal behavior is often produced when biological factors and social factors interlock – that is, when they correlate and when they interact. Today, findings from the biosocial approach are now being used to develop more effective treatments for criminal behavior, and they are gaining a broader audience of scholars.

Although biosocial criminology has only recently penetrated mainstream criminological thinking, the roots of biosocial criminology extend to Darwin's writings on human origins and evolution and to Gregor Mendel's discovery of molecular genetics. Interestingly, Darwin and Mendel each understood that many traits and behaviors were produced by environmental and biological factors. Even so, the road that led to the emergence of biosocial criminology has been long and perilous – ending the careers of earlier scholars who dared to breach the wall of imposed disciplinary ignorance that would eventually make the biological study of human behavior, especially criminal behavior, taboo. Indeed, few areas in the social sciences have been subject to as much conflict, prejudice, and political ideology as has been the connection of biology to behavior. For reasons that had little to do with science, contemporary biosocial criminology almost did not emerge and, consequently, the insights provided by biosocial scholars almost did not appear. Despite the recent ascendancy and acceptance of the paradigm, it remains perched perilously close to intellectual obscurity. In the following pages, we thus trace the journey of biosocial criminology, from its early emergence in criminology, through the history of its discrimination, to its recent revitalization, and finally to what it offers the future of criminology and the social sciences.

The Evolution of the Biosocial Approach

Biological explanations of crime predate sociological explanations of crime and at one time were the dominant approach to examining and understanding criminality. From its inception, the biological study of crime has always been rooted in scientific positivism. Scientific positivism involves the application of the scientific method, where traits and characteristics are operationalized, measured, and analyzed in an effort to detect patterns of covariation. The movement toward biological explanations of crime was most notably spurred by the nineteenth century physician and Italian positivist named Cesare Lombroso.

Often adorned as the father of positivism, Lombroso studied thousands of individuals incarcerated in Italian prisons, documenting many physical and environmental differences he thought potentially related to criminal behavior. Lombroso is often reduced to his thoughts on the 'atavistic man,' – that is, his beliefs that a minority of criminals were biologically less advanced than others and that atavists shared many physical characteristics with primitive humans. Less recognized is that Lombroso also thought that most criminal offending could be explained by environmental and biological factors and that his 'atavists' represented a very small fraction of the offending population. Nonetheless, Lombroso examined thousands of individual criminals and noted several shared physical characteristics such as muscular builds, low foreheads, broad noses, larger jaws, and smaller cranial vaults. He concluded that criminals were often, but not always, biologically different from others and that some were born with a natural inclination toward pleasure-seeking and criminal conduct. These observations and conclusions were chronicled in his famous work, *The Criminal Man* (1911). Lombroso's early studies inspired other scholars to conduct similar research regarding the correlations between physical characteristics and criminality (Hooten, 1939; Jacobs, 1965).

The Italian positivists focused on specific physical characteristics of offenders and attempted to correlate those differences with criminal behavior. Throughout the 1800s and early 1900s, biological explanations of crime were well received and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45086-5

supported as scientific advancements in the study of human nature. However, as society began to rapidly change in both Europe and the United States, political movements began to paint a very different picture of the biological approaches to crime. The rise of totalitarian governments, namely Fascist Italy and Nazi Germany, gave way to some of the most horrifying political movements in modern history. These totalitarian governments, scholars would later point out, sometimes used biological explanations to justify their repressive state actions. The same was true in the United States with the rise of the 'Progressive Movement' and their related acceptance of eugenic social policies. The progressive political movement, often linked to the ascendancy to modern-day liberalism, advocated the forced sterilization of men and women as a method to control the social and financial costs of socially undesirable populations. The eugenic movement went on to inform Nazi efforts to eradicate unwanted groups.

Collectively, the Nazis and the Progressives shared a view that they could socially engineer a better society if government controlled sexual reproduction. To advance their agendas, they used evidence and theories found in the biological sciences to justify state intervention in sexual reproduction. This movement reached its peak during World War II through the early 1950s, but forced sterilizations continued in the United States through the 1970s. Nonetheless, biological explanations of human behavior and human differences quickly fell out of favor as they became politically fused to some of the worst atrocities the world had witnessed.

Even after totalitarian, governments were disbanded and the eugenics movement ended, biological approaches to crime remained taboo. In the United States, there was a radical change in culture in the 1960s with the emergence of civil rights movements. While the civil rights movements brought instrumental and much needed change in society, it also radicalized the field of sociology. Criminology at the time was part of sociology, and sociology had a long history of excluding biology and the study of individual differences. Many of the emerging scholars during this time period were radicals who were highly influenced by civil rights, feminism, and anticapitalism campaigns. The radicalization of sociologists during the 1960s is important because these scholars became the gatekeepers of ideas and they did not hesitate to sanction individual scholars interested in the biological bases of behavior. Thus, liberal ideology focusing on social processes and environmental experiences became the dominant paradigm in criminology, specifically, and in much of the rest of the social sciences. In rapid succession, social process and social structural theories, such as social disorganization theory, differential association theory, and social bond theory, emerged as the sole focus and explanation of crime. Thus, from the 1960s until recently, the social sciences was fundamentally opposed to explanations of individual differences, rejecting biological explanations on ideological rather than scientific grounds. Scholars attempting to revitalize the study of personality, biology, and individual differences were painted as racist, classist, or sexist. Biological theorizing, moreover, was seen not only as undesirable but also as something to be sanctioned, and scholars who conducted research in this area were essentially black listed and socially ostracized.

The study of biological influences on crime would have completely disappeared from criminology had it not been for the scholars who courageously continued to study and highlight the importance of biological influences – sometimes at great personal cost. One individual was E.O. Wilson. E.O. Wilson's (1975) book, *Sociobiology: The New Synthesis*, stirred tremendous controversy. Wilson's book discussed ants and applied basic evolutionary perspectives of ants to human beings. As Wilson discussed individual differences and evolution rather than social processes, his brilliant insights were heavily criticized by scholars, especially those in the social sciences.

Nearly a decade later, Wilson and Herrnstein's (1985) book, *Crime and Human Nature: The Definitive Study of the Causes of Crime*, was released, showcasing individual differences in traits such as aggression and intelligence as strong predictors of a variety of negative social outcomes. Wilson and Herrnstein's (1985) work brought on a firestorm of scathing criticisms from a myriad of social science scholars. The influential book was rejected by the criminological community and is still seen today by radical criminologists as a dark mark in the history of social science literature. The 'nature versus nurture debate raged on, however, and despite the controversy surrounding these works and the study of individual differences, authors such as Wilson and Herrnstein inspired a handful of other brave scholars. Individuals such as Lee Ellis, Anthony Walsh, David Rowe, and Adrian Raine adopted and maintained allegiance to the biological perspective.

Influenced by these multidisciplinary scholars and excited by new lines of research, an ambitious crop of scholars emerged at the turn of the century. These scholars sought to integrate biological discoveries into modern criminological theories. From this group of researchers, the biological approach to crime evolved into a broader research agenda including not only findings from biology, neurology, psychology, and medicine but also the way these individual differences interact with traditional sociological principles surrounding the individual's environment.

The modern biosocial approach emphasizes the connection between biology and environment, asserting that both nature and nurture are important and that both can interact across time. Because scholars study the way that biology influences social outcomes and the socialization process, the term biological was replaced with the term biosocial, integrating both biological explanations and sociological explanations of criminality. This new biosocial approach has grown exponentially in the last decade, producing more research and interested young scholars than ever before. While many criminology programs across the nation have yet to include biosocial criminology in their curriculum, several programs are beginning to add biosocial classes, attracting cohorts of young scholars eager to learn about the discoveries of biosocial criminology. While the biosocial approach is still met with resistance and criticism, it has emerged as a powerful and pervasive presence in modern criminology.

The biosocial paradigm involves a broad range of topics focusing on individual differences and their connection to criminality. This includes a myriad of topics surrounding personality traits, antisocial behavior, aggression, self-control, psychology, brain structure and function, hormones, and

gene x correlations and interactions. In general, the many subjects falling under the umbrella of the biosocial approach can be categorized into four main areas of study: evolutionary perspectives, behavioral genetics, molecular genetics, and brain-based studies.

Evolutionary Perspectives

Why are men more prone to aggression and violence than women? Evolutionary theory asserts that the reason lies in our evolutionary history (Fergusson and Horwood, 2002; Krueger et al., 1994). Darwinian evolution is built on the principles of natural selection and reproductive processes. Natural selection is the process by which certain traits are passed on through time because they are beneficial and aid in adaptation to the environment, while other traits are selected against and filtered out over time. When humans first evolved, they were faced daily with dangerous tasks of hunting predators, securing shelter, and locating a source of water and nourishment. Males were particularly adept at hunting and at protecting the family unit or tribe, while females were more adept at gathering food, and nurturing and protecting children. These tasks required different traits. Males needed high levels of aggression to track down and to kill prey, and they had to be able to ward off competing tribes. Additionally, men needed to secure females for reproductive purposes. Males who were larger and more muscular in stature were, the theory goes, viewed by females as the most desirable and capable mates. As such, natural selection preserved traits most beneficial to each sex, creating pronounced physiological and psychological differences between males and females.

Crime is often concentrated in families. Several family studies indicate that criminal family members, especially parents, are an important correlate of criminality (Farrington et al., 2001; Farrington, 2002; McCord, 2007). Why do we often see cycles of criminality in families, generation after generation? Evolutionary psychology points to assortive mating as a likely culprit. Assortive mating is the process by which individuals select mates that are similar to themselves. Research indicates that individuals particularly choose partners similar in traits such as intelligence, personality, antisocial behavior (Krueger et al., 1998; Rhule-Louie and McMahon, 2007), and even similar political orientation (Alford et al., 2005; Kandler et al., 2012). Essentially, this means that antisocial individuals often gravitate toward one another. Children of two antisocial individuals thus inherit not only their parents' genes but also the antisocial environment produced by the parent. Owing to this insight we can better understand how problem behaviors and criminality become segregated within some families, but not others, and why crime can be transmitted across generations.

Behavioral Genetics

Can criminality be inherited? Behavioral genetics is a branch of genetic studies, which seeks to find answers to this question. Behavioral genetic studies analyze the contributions of genes and environment to a variety of traits such as intelligence, self-control, aggression, and impulsivity. Behavioral genetic studies attempt to quantify the extent to which genetic factors influence a specific trait. This estimate is known as heritability. Nearly, all traits are heritable to some degree, and behavioral genetics examines the degree of heritability in a variety of traits associated with criminality. Behavioral genetic scholars emphasize that genes do not cause people to become criminal but rather that genes can facilitate dispositions to respond to environmental stimuli in criminal ways. There is no specific gene for criminality. However, some genes may lead to traits such as low self-control and low empathy, which, when combined with a dysfunctional environment, may increase the probability of criminal behavior.

This line of research also examines continuity and change of traits over time. Many studies find that traits such as self-control and aggression remain relatively stable across the life course (Beaver and Wright, 2007; Haberstick et al., 2006; Huesmann et al., 1984; Vazsonyi and Huang, 2010). Studies have analyzed individuals in very early childhood up through adulthood finding that antisocial behavior remains stable across time and is therefore identifiable very early in life (Caspi, 2000; Caspi and Moffitt, 1995; Rose et al., 1989). Behavioral genetics research also examines the sources of stability and change in traits. Specifically, behavioral genetics studies examine the amount of variance in a trait such as impulsivity that is attributed to genetics and the amount of variance that is attributed to environmental factors. The two main studies that help answer these questions are adoption studies and twin studies.

Adoption studies compare adopted children to their adoptive and to their biological parents. Hutchings and Mednick (1977) conducted a study of male children who had been adopted early in life. They gathered criminal records of both the adoptive and biological parents and followed the sample of boys into adulthood. The criminal records of the biological parents were statistically predictive of the children's criminal records – more so than were those of the adopted parents. These findings support genetic influences in criminality. However, some scholars point out that adoptive parents may be matched by agencies and be similar to the biological parents. In response to these criticisms, another line of studies emerged. Researchers examined twins that were raised apart from one another in unique environments. A study done at the University of Minnesota examined twins who were separated at birth and reared apart. Researchers found that twins were very similar in terms of personality, antisocial tendencies, and criminal behavior even when coming from very different social environments (Bouchard et al., 1990). These findings indicate that genetic factors play a role in criminality.

Extending beyond the observational nature of adoption studies, twin studies allow researchers to statistically estimate genetic and environmental contributions to specific phenotypes by comparing monozygotic (identical) twins to dizygotic (fraternal) twins. Monozygotic twins share 100% of their genes, while dizygotic twins share, on average, 50% of their genes. If there is a genetic effect, monozygotic twins should be more similar than dizygotic twins. If twin 1 scores high on impulsivity and twin 2 scores high on impulsivity, their outcome is concordant. If one is high and the other is low, the outcome is discordant. Higher concordance rates among monozygotic

twins as compared to dizygotic twins indicate a genetic effect as environmental influences, such as family, are expected to be similar. However, this method does not perfectly tease apart the variation in a trait as some concordance among twins may also be due to other environmental experiences such as similar peer groups.

Another statistical method in behavioral genetics is using the standard ACE model. The ACE model compares monozygotic twins and dizygotic twins but rather than simply calculating concordance rates, this statistical technique breaks down the variation in a trait into three estimates. The 'A' term is equal to the proportion of variation attributed to genetic factors, which is also known as heritability. The environmental contributions are broken down into shared/common environmental factors (C), such as the family, and nonshared/unique environmental factors (E), such as peer groups or work groups. The ACE model compares the correlations between monozygotic and dizygotic twins and decomposes those correlations into genetic, shared environmental, and nonshared environmental contributions, with the sum of the squared estimates adding to 1.00. Many studies that have employed ACE models indicate that genetic effects for traits such as low self-control, aggression, and antisocial behavior are dominated by genetic influences, with estimates typically ranging from 50 to 95% (Rhee and Waldman, 2002). Another trend in longitudinal studies is that while the genetic effects grow stronger across time, the effect of shared/common environmental influences decreases over time. As children age, nearly all of the variation in their behavioral traits is attributed to genetic and nonshared environmental factors. Behavioral genetics research has identified consistent predictors of criminality that are identifiable early in life, that remain stable across the life span, and that are predominantly genetically influenced. This line of research has helped highlight the importance of early identification and treatment of problem behaviors.

Molecular Genetics

Genes make us who we are. Genes are behind virtually all traits, from the color of our eyes, our height, and other physical characteristics to our personality, intelligence, and our behavioral tendencies. We also know that genes interact with our environment and in turn, our environment can have reciprocal effects on our genetic makeup. While behavioral genetics focuses on heritability of traits, molecular genetics examines specific genes, proteins, enzymes, and genetic combinations and how they interact with environmental stimuli. Genes do not directly affect our behavior. Rather, genes interact with other stimuli such as abuse, to indirectly affect a behavioral outcome, such as aggression. The completion of the Human Genome Project has allowed researchers to identify and study the indirect effects of particular genetic material on behavior and social outcomes. A particular gene of interest is the MAOA gene. Caspi et al. (2002) conducted an extensive longitudinal study of the interaction between variants of the MAOA gene and abuse/neglect during childhood on violent and antisocial behavior in adulthood. Specifically, the authors measured whether the subjects had been arrested for a violent crime in adulthood. In general, low-activity variants of the MAOA gene are a risk factor for violent behavior, while the high-activity variant is generally a protective factor. Results indicated that neither the environmental risk factor of abuse/neglect nor the genetic risk factor of the low-activity MAOA gene was predictive of arrests for violent crime. However, those individuals with the combination of the low-activity MAOA gene and abuse/neglect in childhood were nearly 10 times more likely to have been arrested for a violent crime than those who had neither the genetic nor the environmental risk factor. Additionally, while subjects with both genetic and environmental risk factors comprised just 12% of the sample, they were responsible for 44% of the violent crime arrests (Caspi et al., 2002). Other studies have also confirmed interactions between maltreatment and the MAOA gene on antisocial behavior (Haberstick et al., 2005; Kim-Cohen et al., 2006). Certain gene sequences and variants have also been associated with alcohol abuse and related behaviors (Heath et al., 1997; Wang et al., 2008).

Other important discoveries involve the effects of differential levels of hormones such as testosterone and estrogen. Research indicates that men with elevated levels of testosterone and other androgens are more prone to risk-taking behavior and aggression, both of which are linked to criminality (Ellis, 2005). Neurotransmitters, chemicals that pass electric signals between neurons, also play an important role in regulating emotions and decision making. In particular, low levels of serotonin, an important chemical in both cognitive and emotional processing, is associated with aggression, impulsivity, and criminal behavior (Blumensohn et al., 1995; Coccaro et al., 1997; Ellis, 2005). The identification of hormones, neurotransmitters, and specific gene variants and their interactions with the environment has been an important step in further understanding criminality (Raine, 1997).

Brain Studies

The brain is a remarkably complex, vital, and fragile component of human life. The brain controls everything from bodily functions to cognitions to emotions. Scholars from medicine, neurology, biology, and other hard sciences have always recognized the importance of the brain in shaping human behavior. It is just recently that criminological scholars are beginning to recognize the role of the brain in explanations of criminality. Damage to the brain can result in a myriad of problems involving cognition, perception, sensation, memory, and impulse and emotional regulation. The limbic system sits atop of the brain stem and is within the cerebrum. Damage to particular areas of the brain has been linked to criminality. Problems with the frontal lobe are consistently associated with criminal behavior. The frontal lobe is located directly behind the forehead and contains the prefrontal cortex, an area of the brain responsible for executive functioning. Executive functioning includes tasks such as planning, decision making, delaying gratification, impulse control, and anticipation of consequences. Chronic offenders typically display executive dysfunctions. Another area of the brain, the limbic system, communicates with other parts of the brain and helps balance emotional and rational impulses. In particular, a structure

called the amygdala is responsible for emotional regulation. Problems with the amygdala and limbic system often manifest as individuals with high levels of impulsivity and aggression (Ogilvie et al., 2011).

Research indicates that damages and dysfunction in the frontal lobe are linked to criminality. Many of these functional and structural issues are highly heritable (Rafter, 2008). Just as genes and environment interact, the environment may also have important effects on the brain. Longitudinal studies of Romanian orphans, Chugani et al. (2001) and Eluvathingal et al. (2006), found that those orphans who had experienced severe emotional deprivation had brain abnormalities and were cognitively underdeveloped relative to the comparison group. Other studies have highlighted the effects that alcohol and drug abuse have on the brain, noting that prolonged abuse can physically alter an individual's brain structure (Volkow and Ting-Kai, 2004). Differences in individual brain structure and function have also been observed in individuals with conduct disorder, schizophrenia, and tendencies toward risk-taking behaviors (Decety et al., 2008; Gogtay et al., 2004). Clearly, the brain structure and function plays a crucial role in criminality.

Moving Forward

The biosocial approach to crime has had a unique and critical impact on criminology. Biosocial approaches have revitalized and highlighted the importance of individual differences, an insight that was ignored for several decades. Differences in personality, intelligence, self-control, aggression, executive functioning, and a host of other biological and physiological factors explain differences in criminality. These individual differences emerge early in life and can be identified in early childhood. Furthermore, traits such as personality, aggression, and self-control remain relatively stable across childhood, through adolescence, and into adulthood. Focusing prevention and rehabilitation efforts on only adolescents and young adults is inefficient. These findings implicate the need for early detection of problem behaviors and early intervention strategies before serious social outcomes manifest.

While bringing the focus of criminology back to the individual, biosocial approaches do not refute the importance of the individual's environment. Genes and biology interact with environmental stimuli. Specifically, findings indicate that environmental factors such as parenting styles, attachment, and abuse/neglect can affect the expression of genes and the manifestation of behavioral outcomes. The more we learn about how biology, especially our genes and brain chemistry, is related to criminal behavior, the more we come to understand the importance of environmental influences. There is no need for the nature versus nurture debate to continue. Rather, we may look at the connection between biology and criminality as nature via nurture. Inadequate parents, poor socialization, deviant peer groups, and other negative environmental stimuli play important roles in individual development, but criminological theories that exclude the effects of biological factors are incomplete. As stated by former American Society of Criminology and former American Criminal Justice Society president, Dr Francis T. Cullen, at the 2010 American Society of Criminology Sutherland Address:

> As we move forward, however, we can no longer pretend that biology is not intimately implicated in human behavior and thus in criminal behavior. Knowledge about genes and how brains operate is exploding, and it is influencing other social scientific fields—including even sociology… In the end, we truncate reality if we do not study what we all now know: Babies emerge from the womb with individual differences, which then influence the lives that they will help to create.

There is an overwhelming amount of evidence across disciplines that point to the direct and indirect influences biology has on behavior. Genes and neurological deficits matter. Study after study has found that genes have a profound impact on human development and behavior and that genetic influences account for the majority of variation in criminality, followed by nonshared environmental influences such as delinquent peer groups. Scientific advancements and discoveries have allowed scholars to conduct more sophisticated analyses of heritability estimates, effects of brain structure and function, the presence of certain gene and enzyme combinations, and the gene × environment interactions related to criminality. While limitations remain, the biosocial approach has helped advance the methodological standards of empirical research in criminology. Despite the ideological obstacles, skepticism, and controversies surrounding biosocial criminology, the growing interest and productivity of this line of research is unparalleled in the field of criminology. Regardless of those who oppose the application of science in the study of criminality, the biosocial approach is the future of criminology. Biosocial criminology has and will continue working to integrate theories from a variety of fields using empirically sound research to enrich our understanding of criminality.

See also: Genetic Studies of Behavior: Methodology; Life Course Criminology; Lifespan Development: Evolutionary Perspectives; Personality, Biological Models of; Self-Control Theory and Crime; Sociobiology: Overview.

Bibliography

Alford, J.R., Funk, C.L., Hibbing, J.R., 2005. Are political orientations genetically transmitted. American Political Science Review 99 (2), 153–167.

Beaver, K.M., Wright, J.P., 2007. The stability of low self-control from kindergarten through first grade. Journal of Crime and Justice 30, 63–86.

Beaver, K.M., Wright, J.P., DeLisi, M., Vaughan, M.G., 2008. Genetic influences on the stability of low self-control: results from a longitudinal sample of twins. Journal of Criminal Justice 36, 478–485.

Blumensohn, R., Ratzoni, G., Weizman, A., Apter, A., Tyano, S., Israeli, M., Greuner, N., Apter, A., Tyano, S., Biegon, A., 1995. Reduction in serotonin 5HT2 receptor binding on platelets of delinquent adolescents. Psychopharmacology 118 (3), 354–356.

Bouchard, T.J., Lykken, D.T., McGue, M., Segal, N.L., Tellegen, A., 1990. Sources of human psychological differences: the Minnesota study of twins reared apart. Science 250 (4978), 223–228.

Chugani, H.T., Behen, M., Muzik, O., Juhasz, C., Nagy, F., Chugani, D.C., 2001. Local brain functional activity following early deprivation: a study of postinstitutionalized Romanian orphans. NeuroImage 14, 1290–1301.

Caspi, A., McClay, J., Moffitt, T.E., Mill, J., Martin, J., Craig, I.W., Poulton, R., 2002. Role of genotype in the cycle of violence in maltreated children. Science 297 (5582), 851–854.

Caspi, A., Moffitt, T.E., 1995. The continuity of maladaptive behavior: from description to understanding in the study of antisocial behavior. In: Cicchetti, D., Cohen, D. (Eds.), Manual of Developmental Psychopathology, vol. 2. Wiley, New York, pp. 472–511.

Caspi, A., 2000. The child is the father of the man: personality continues from childhood to adulthood. Journal of Personality and Social Psychology 78, 158–172.

Coccaro, E.F., Kavoussi, R.J., Cooper, T.B., Hauger, R.L., 1997. Central serotonin activity and aggression: inverse relationship with prolactin response to d-fenfluramine, but not CSF 5-HIAA concentration, in human subjects. American Journal of Psychiatry 154 (10), 1430–1435.

Decety, J., Michalska, K.J., Akitsuki, Y., Lahey, B.B., 2008. Atypical empathetic responses in adolescents with aggressive conduct order: a functional MRI investigation. Biological Psychology. http://dx.doi.org/10.1016/j.biopsycho.2008.09.004.

Ellis, L., 2005. A theory explaining biological correlates of criminality. European Journal of Criminality 2, 287–315.

Eluvathingal, T.J., Chugani, H.T., Behen, M.E., Juhasz, C., Muzik, O., Maqbool, M., Chugani, D.C., Makki, M., 2006. Abnormal brain connectivity in children after early severe socioemotional deprivation: a diffusion tensor imaging study. Pediatrics 117 (6), 2093–2100.

Farrington, D.P., 2002. Families and crime. In: Wilson, J.Q., Petersilia, J. (Eds.), Crime: Public Policies for Crime Control. ICS Press, Oakland, CA, pp. 129–148.

Farrington, D.P., Jolliffe, D., Loeber, R., Stouthamer-Loeber, M., Kalb, L.M., 2001. The concentration of offenders in families, and family criminality in the prediction of boys' delinquency. Journal of Adolescence 24 (5), 579–596.

Fergusson, D.M., Horwood, L.J., 2002. Male and female offending trajectories. Development and Psychopathology 14 (1), 159–177.

Gogtay, N., Giedd, J.N., Lusk, L., Hayashi, K.M., Greenstein, D., Vaituzis, C.A., Nugent III, T.F., Herman, D.H., Clasen, L.S., Toga, A.W., Rapoport, J.R., Thompson, P.M., 2004. Dynamic mapping of human cortical development during childhood through early adulthood. PNAS 101 (21), 8174–8179.

Haberstick, B.C., Schmitz, S., Young, S.E., Hewitt, J.K., 2006. Genes and developmental stability of aggressive behavior problems at home and school in a community sample of twins aged 7–12. Behavioral Genetics 36, 809–819.

Haberstick, B.C., Lessem, J.M., Hopfer, C.J., Smolen, A., Ehringer, M.A., Timberlake, D., Hewitt, J.K., 2005. Monoamine oxidase A (MAOA) and antisocial behaviors in the presence of childhood and adolescent maltreatment. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 135 (1), 59–64.

Heath, A.C., Bucholz, K.K., Madden, P.A.F., Dinwiddie, S.H., Slutske, W.S., Bierut, L.J., Martin, N.G., 1997. Genetic and environmental contributions to alcohol dependence risk in a national twin sample: consistency of findings in women and men. Psychological Medicine 27 (6), 1381–1396.

Hooten, E., 1939. Crime and the Man. Harvard University Press, Cambridge, MA.

Huesmann, L.R., Eron, L.D., Lefkowitz, M.M., Walder, L.O., 1984. The stability of aggression over time and generations. Developmental Psychology 20, 1120–1134.

Hutchings, B., Mednick, S.A., 1977. Criminality in adoptees and their adoptive and biological parents: a pilot study. In: Mednick, S.A., Christiansen, K.O. (Eds.), Biosocial Bases of Criminal Behavior, Gardner, Newyork, pp. 127–141.

Jacobs, P., 1965. Aggressive behavior, mental subnormality, and the YY male. Nature 208, 1351–1352.

Kandler, C., Bleidorn, W., Riemann, R., 2012. Left or right? Sources of political orientation: the roles of genetic factors, cultural transmission, assortative mating, and personality. Journal of Personality and Social Psychology 102 (3), 633.

Kim-Cohen, J., Caspi, A., Taylor, A., Williams, B., Newcombe, R., Craig, I.W., Moffitt, T.E., 2006. MAOA, maltreatment, and gene–environment interaction predicting children's mental health: new evidence and a meta-analysis. Molecular Psychiatry 11 (10), 903–913.

Krueger, R.F., Moffitt, T.E., Caspi, A., Bleske, A., Silva, P.A., 1998. Assortative mating for antisocial behavior: developmental and methodological implications. Behavior Genetics 28 (3), 173–186.

Krueger, R.F., Schmutte, P.S., Caspi, A., Moffitt, T.E., Campbell, K., Silva, P.A., 1994. Personality traits are linked to crime among men and women: evidence from a birth cohort. Journal of Abnormal Psychology 103, 328–338.

Lombroso, G., 1911. Criminal Man: According to the Classification of Cesare Lombroso. Putnam, New York.

McCord, Joan, 2007. In: Sayre-McCord, G. (Ed.), Crime and Family: Selected Essays of Joan McCord. Temple Univ. Press, Philadelphia, PA.

Ogilvie, J.M., Stewart, A.L., Chan, R.C.K., Shum, D.H.K., 2011. Neuropsychological measures of executive function and antisocial behavior: a meta-analysis. Criminology 49 (4), 1063–1107.

Rafter, N.H., 2008. The criminal brain: understanding biological theories of crime. New York University Press, New York.

Raine, A., 1997. The Psychopathology of Crime: Criminal Behavior as a Clinical Disorder. Gulf Professional Publishing.

Rhee, S.H., Waldman, I.D., 2002. Genetic and environmental influences on antisocial behavior: a meta-analysis of twin and adoption studies. Psychological Bulletin 128 (3), 490.

Rhule-Louie, D.M., McMahon, R.J., 2007. Problem behavior and romantic relationships: assortative mating, behavior contagion, and desistance. Clinical Child and Family Psychology Review 10 (1), 53–100.

Rose, S.L., Rose, S.A., Feldman, J.F., 1989. Stability of behavior problems in very young children. Development and Psychopathology 1, 5–19.

Vazsonyi, A.T., Huang, L., 2010. Where does self-control come from: on the development of self-control and its relationship to deviance over time. Developmental Psychology 46, 245–257.

Volkow, N.D., Ting-Kai, L., 2004. Drug addiction: the neurobiology of behavior gone awry. Perspectives: Science and Society 5, 963–970.

Wang, J.C., Grucza, R., Cruchaga, C., Hinrichs, A.L., Bertelsen, S., Budde, J.P., Goate, A.M., 2008. Genetic variation in the CHRNA5 gene affects mRNA levels and is associated with risk for alcohol dependence. Molecular Psychiatry 14 (5), 501–510.

Wilson, E.O., 1975. Sociobiology: The New Synthesis. Harvard University Press, Cambridge, MA.

Biotechnology

Stephen Hilgartner, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article surveys the rise of biotechnology since the closing decades of the twentieth century, examining both scientific developments and social and institutional change. As an area of science and technology with the explicit goal of intervening in the machinery of life, biotechnology often disrupts traditional ways of distinguishing 'nature' from 'culture,' calling into question settled social arrangements. As a result, biotechnology poses difficult challenges of governance. This article examines the rise of biotechnology and considers its technological and epistemic dimensions, then turns to its institutional aspects and problematic position in contemporary politics. The role of risk management and bioethics in legitimation are also discussed.

In the closing decades of the twentieth century, biotechnology emerged as a site of rapid change in science and technology and as an arena of social and institutional transformation. The development of new techniques for studying, manipulating, and redesigning living things produced important applications in medicine and agriculture, and generated massive investment. In the world of research, biotechnology was often at the forefront of change in scientific institutions and practices. More broadly, biotechnology – widely perceived as having 'revolutionary' implications – inspired both intense enthusiasm and determined opposition. As an area of science and technology with the explicit goal of intervening in the machinery of life, biotechnology often disrupts traditional ways of distinguishing 'nature' from 'culture,' calling into question settled social arrangements. The potential for change associated with emerging biotechnologies thus has presented societies with normative choices about the kinds of futures they seek to create, sparking controversies about new hopes, risks, subjectivities, and ethical dilemmas. As a result, biotechnology poses difficult challenges of governance. This article surveys the rise of biotechnology, examining its technological and epistemic dimensions, its institutional aspects, and problematic position in contemporary politics.

The Term 'Biotechnology'

Defining biotechnology poses challenges, for the word is less a tightly defined, technical term than a loose umbrella category, or even a slogan, that conveys – sometimes simultaneously – visions of unbounded progress and unregulated tampering with nature. Many authors have tried to capture biotechnology within their own well-crafted definitions, but these attempts cannot neatly contain this expanding network of activities and increasingly dense connections to diverse social worlds. Although the word has a long history (Bud, 1993), in most contemporary contexts biotechnology refers to a novel and growing collection of techniques, grounded in molecular and cell biology, for analyzing and manipulating the molecular building blocks of life. The term also designates products, such as pharmaceuticals or genetically modified (GM) foods, created using these techniques. At times, it refers not to products or techniques but to an economic sector or area of research. Biotechnology acquired these intertwined meanings toward the end of the 1970s, coming into widespread use in the early 1980s, as molecular biology was increasingly understood not only as a 'science' for learning about nature but also as a 'technology' for altering it.

The Biotechnology 'Revolution'

One of the ironies of the rise of biotechnology is that a sense of revolutionary potential energizes both the enthusiasm and the opposition it engenders. Supporters and critics alike often fit biotechnology into a narrative of radical discontinuity (e.g., Carlson, 2010; Conway, 1997; Kevles and Hood, 1992; Rifkin, 1983; Kass, 2002). Biotechnology advocates claim that it will completely transform medicine, spawn entirely new industries, and supply adequate food to the growing world population. Critics – who warn of unanticipated consequences and the hazards of altering human identity – also cast the new biology as a revolutionary force, arguing that genetic engineering is ushering in a risky new stage in evolution, as nature itself comes under technological control.

Alongside the notion of revolution, the discourse on biotechnology also features an opposing frame that stresses continuity with the biological and historical past. Thus, supporters of biotechnology sometimes downplay its novelty, defining it inclusively (e.g., as the harnessing of biological agents to provide goods and services) and portraying it as merely the latest twist on age-old methods of plant breeding, animal husbandry, and brewing. Such moves help to make the exotic familiar, to suggest that ancient precedents justify apparently novel practices, and to undermine the notion of an untouched 'natural' order that must be protected from human intervention. Critics of biotechnology, for their part, often contend that predictions of unprecedented progress overstate the benefits, suggesting that biotechnology will perpetuate, rather than eliminate, long-standing inequalities in agricultural and health care systems.

As these observations suggest, narratives of continuity and discontinuity provide opposing frames that participants in debates about biotechnology can selectively deploy to mobilize

 http://dx.doi.org/10.1016/B978-0-08-097086-8.85046-1

support for their views. But for social scientists, the relevant question is whether the concept of a biotechnology revolution has analytic utility. Although some might be tempted to dismiss this notion as overheated rhetoric or media 'hype,' revolutionary rhetoric and mass media coverage play a constitutive role in shaping biotechnology (Fortun, 2008; Hilgartner, 2012). Like information technology and computers, biotechnology seems perpetually to stand at the threshold of the next qualitative transformation (*see* Science and Technology, Social Study of: Computers and Information Technology). Emerging technologies exist not merely in material form, but also as expected or imagined futures built in part of hopes, fears, and promises that drive speculation and have effects in the present (Brown et al., 2000; Fortun, 2008; Nelson et al., 2008; Rabinow and Dan-Cohen, 2005). Biologists who pause to marvel at the pace of progress, like ordinary citizens who reflect on the stream of announcements from the research front, experience biotechnology as an area of ongoing change where old boundaries are continually broken and the unprecedented rapidly becomes the mundane.

A New Technological and Epistemic Space

Whether framed as revolutionary change or incremental evolution, there is no doubt that biotechnology has grown increasingly prominent since the 1970s, inspiring the growth of an increasingly global, 'big biology' (Davies et al., 2013). Biotechnology emerged from a complex constellation of phenomena, among them the development of increasingly powerful means for representing and manipulating the molecular building blocks of life. At the level of infrastructure, the routinization of practices for culturing biological tissues transformed cells into tools for investigating their own operation (Landecker, 2007). Beginning in the early 1970s, the development of a cascade of novel techniques – especially recombinant DNA (rDNA) technology, monoclonal antibodies, *in vitro* fertilization, the polymerase chain reaction (PCR), and high-throughput DNA sequencing – dramatically extended human control over living organisms. Although the arrival of each of these techniques was widely heralded as 'revolutionary,' the growing power of biotechnology cannot be attributed to any single tool, but stems from the ability to combine and recombine such tools, creating new assemblages capable of building and taking apart molecules, rearranging the genetic code of organisms, and reading, analyzing, and writing the 'language' of DNA with an increasingly integrated set of laboratory techniques and computational tools. No less significantly, such techniques have consolidated and accelerated epistemic change in the life sciences, increasing the centrality of metaphors from information theory and cybernetics (e.g., code, control, and text) for understanding living things (Kay, 2000; Keller, 1995, 2002). More precisely, these tools provide practical means that, quite literally, lend substance to these metaphors, giving them material form and making them into epistemic things embedded in experimental systems (Rheinberger, 1997) and sociotechnical networks (*see* Experiment, in Science and Technology Studies).

The field of molecular biology expanded rapidly after World War II (de Chadarevian, 2002), and by 1970, scientists knew how genes specify protein structures and had found enzymes that cut, join, and extend DNA molecules. This set the stage for rDNA technology – the emergence of which is often taken as the starting point for modern biotechnology. rDNA techniques enabled scientists for the first time to insert or delete genes from the genetic code of living things, for example, by removing a piece of DNA from one organism and splicing it into the DNA of a distantly related species. The development of these methods generated great excitement and provoked controversy (discussed below) about the wisdom and safety of genetic engineering. Beyond opening up a wide range of new research strategies, this technology allowed scientists to design genetically modified organisms (GMOs) to produce commercial products. In the late 1970s, researchers began using genetically engineered bacteria as microscopic manufacturing plants for making valuable proteins, such as human insulin and human growth hormone.

In the last two decades of the twentieth century, the power of the biotechnology tool kit continued to increase. The development of monoclonal antibodies, which permitted scientists to construct extremely accurate assays for detecting specific proteins, yielded applications in research and medical diagnostics (Cambrosio and Keating, 1995). New methods for transferring genes into developing embryos allowed scientists to alter the genetic code of animals, thus extending genetic engineering from microorganisms to higher forms of life. Parallel techniques enabled scientists to produce transgenic plants (Nelson, 2012), and research aimed at improving the genomes of agriculturally important crops became a major focus of the plant sciences. Molecular biology and biotechnology also played an increasingly central role in the development of pharmaceuticals, including vaccines, antibiotics, and drugs used to combat heart disease, cancer, and AIDS.

Tools and practices for representing and analyzing DNA molecules were also rapidly developed (Fujimura, 1996). Methods for mapping the location of genes grew increasingly powerful; PCR became important in research, diagnostics, and forensics (Rabinow, 1996b); and DNA sequencing – a technique for representing the nucleotide sequences of DNA molecules as written inscriptions (strings of the letters A, C, G, and T) – rapidly developed. In the 1990s, the Human Genome Project, along with other concerted mapping and sequencing efforts, produced an exponentially increasing volume of biomolecular data on a variety of organisms of medical and agricultural importance (Cook-Deegan, 1994; Gaudillière and Rheinberger, 2004). New 'factory-style' laboratories dedicated to sequencing also emerged during the human genome project, and in the years since the completion of a 'first draft' of the human sequence was announced in 2001, a new generation of sequencing machines spurred a continuing exponential increase in the production of sequence data (Hutchison, 2007). These data – along with increasingly automated means of producing, standardizing, ordering, and analyzing them – constitute the core resource of genomics and bioinformatics, fields that have further united information science and biology. In effect, by opening up new computational approaches to biology, genome data have created a new technological and epistemic 'space,' allowing researchers to perform

'experiments' *in silico* or in a domain linking computational work to the 'wet lab.'

Developments in cellular biology also expanded the reach of biotechnology. In 1997, a team of Scottish researchers announced that they had cloned a sheep, who they named 'Dolly,' using a technique called somatic cell nuclear transfer (SCNT). Put simply, Dolly was a 'copy' of an adult ewe created by transferring the nucleus of one of the adult's cells into an enucleated sheep egg and implanting it in another ewe using *in vitro* fertilization. This feat – the first successful cloning of a mammal – suggested that cloning of human beings could also be achieved, sparking controversy around the world about the ethics of human cloning. At about the same time, the discovery of embryonic stem cells (ESCs) suggested a novel strategy for attacking a variety of human diseases. Because ESCs can differentiate to form a wide range of tissue types, they could potentially be used to create replacement tissues to restore function to damaged or diseased organs or systems. Stem cell research inspired great hope and significant investment, but like SCNT, human ESC research also provoked controversy. In many countries, the destruction of human embryos to 'harvest' stem cells generated determined opposition, although the cross-national differences were significant (Prainsack, 2006).

In recent years, biologists and biotechnologists have (often quite self-consciously) advanced new 'paradigms' for research and technology development. One notable development is the rise of 'systems biology,' which seeks to holistically examine the interactions of systems of biological components, tracking their operation quantitatively using computational and mathematical techniques (Calvert, 2013). Another is 'synthetic biology' – an effort that seeks to apply engineering principles to make biotechnology into a true engineering discipline (Campos, 2012; Carlson, 2010; Pottage, 2006). Rather than modifying existing living organisms as biotechnologists presently do, synthetic biologists imagine creating entirely new biology-based devices and systems. Several strategies for achieving this have been proposed, including a scheme based on creating standardized biological 'parts' (or 'biobricks') analogous to electronic components; the idea of using a 'minimal genome' as a 'chassis' for constructing organisms; and the possibility of engineering novel nucleic acid chemistries of human design. Synthetic biology has generated much excitement, but the field is still in its infancy and how and whether it will mature remains an open question.

Institutional Dimensions

Large infusions of capital, promotional state policies, and novel institutional arrangements played a significant role in spurring these developments (Thackray, 1998). By the end of the 1970s, government, industry, and academic elites in both Europe and the United States saw investment in 'high technology' – especially computers and biotechnology – as the long-term prescription for economic growth. On both sides of the Atlantic, policymakers sought to spur innovation and speed the translation of basic science into marketable products. The result: a significant shift in the political economy of research that eased the commodification of knowledge (*see* Patenting in the Biological Sciences). Governments encouraged universities and corporations to form new kinds of academic–industry alliances, and the scope of intellectual property protection grew considerably broader, expanding to encompass, for example, genetically engineered organisms (Boyle, 1996; Jasanoff, 1995; Hilgartner, 2009). These developments – along with increasing globalization of not only the economy but also research and development – have contributed to the emergence of new forms of economic activity and what some have termed 'biocapital' (Sunder Rajan, 2006; Helmreich, 2009).

In the United States during the 1980s, significant numbers of academic biologists became entrepreneurs for the first time, seeking to launch new biotechnology companies that many imagined at the time might soon grow into enormous firms. The most dramatic institutional innovation was the start-up company founded by venture capitalists and university professors – many of who kept their academic posts (Kenney, 1986). Increasingly, biotechnology research took place in spaces that were not neatly lodged in either an academic or a corporate milieu, but in complex hybrids of university and industry, public and private, and basic and applied research (*see* Universities and Science and Technology: United States). Some observers warned that these hybrids would threaten the independence of academic biologists, weakening the research system and compromising their ability to offer a credible, critical voice in public decision-making. Debate continues about whether these developments represent a threat to the moral and intellectual integrity of university science (Mirowski, 2011; Shapin, 2008), but ultimately these institutional innovations proved irresistible, given competitive pressures and the scientific and economic opportunities.

Owing to international variation in research systems, the commercialization of biotechnology followed distinct paths in different nations. In Europe and Japan, the large multinational corporations (MNCs) that controlled production of pharmaceuticals, chemicals, and agricultural inputs tended to build biotechnology directly into their in-house research operations. But in the United States, with its unique pattern of venture capital financing, a separate 'biotechnology industry' composed of freestanding firms took shape (Kenney, 1998). The first major wave of biotechnology firms appeared between 1979 and 1981, and in the following three decades, entrepreneurs founded successive waves of firms, whipping up investor enthusiasm around a series of 'revolutions,' including monoclonal antibodies, gene therapy, genomics, stem cell research, and synthetic biology.

Even in the United States, however, MNCs – which poured huge sums into genetic engineering, rational drug design, and genomics – became the dominant players in efforts to exploit the new biology commercially. Indeed, given the structural connections between the biotechnology firms and the MNCs, the image of a separate biotechnology industry is somewhat misleading: Biotechnology start-ups often derive the bulk of their revenue from contracts with giant companies; MNCs often acquire successful firms; and many

biotechnology firms are best understood as research shops that hope to sell intellectual property to large corporations. But if efforts to exploit the new biology are increasingly the province of both large and small firms, the culture of the biotechnology start-ups that emerged in the United States captured the imagination of science policymakers elsewhere, some of whom have sought ways to stimulate parallel developments. More broadly, policy analysts are promoting the goal of stimulating the growth of the 'bioeconomy' – a term defined inclusively to reach beyond health and agriculture to include the production of fuels, the recycling of materials, and the manufacture of plastics, paper, and industrial chemicals (OECD, 2009).

The Politics of Biotechnology

Since the early 1970s, biotechnology has posed difficult problems of governance, confronting societies with a stream of potentially controversial issues. The expanding applications of biotechnology – which moved quickly from being mainly a research tool into being an important part of medicine and agriculture – significantly increased the numbers of actors with normative ideas about its proper use. The development of DNA forensics (Lynch et al., 2008; Hindmarch and Prainsack, 2010), the advent of genetic tests that yielded predictive information about the onset of future disease (Parthasarathy, 2007; Hedgecoe, 2004), and the rise of 'biobanks' of samples of people's DNA (Gottweis and Petersen, 2008) all raised questions of governance. Biotechnology also got entangled in the politics of race, as scientific research projects intended to study human diversity raised persistently troubling questions about how to represent human similarity and difference (Reardon, 2005; Koenig et al., 2008). In the aftermath of the terrorist attacks of 11 September 2001 and the anthrax attacks only days later, the military potential of biotechnology grew into a prominent security concern (Balmer, 2012; Vogel, 2012). Indeed, as biotechnology grew increasingly entangled in a variety of social institutions, as it was integrated into more and more aspects of everyday life, societies throughout the world confronted a stream of normative questions about how to fit biotechnology into their ways of living in the world.

Finding legitimate settlements to these controversies sometimes proved difficult. Because biotechnology often seems to undermine basic categories of social order, such as the 'natural' or the 'human,' it tends to pose problems that do not fit neatly into the cosmologies and routines that guide public action. Democratic societies, thus face the challenge of building new discursive regimes, regulatory mechanisms, and forums for public discussion that can narrow disputes and achieve legitimacy. In this respect, the institutionalization of risk management and bioethics have been particularly important. Although, the precise techniques for engineering consent have varied across nations (Gottweis, 1998; Jasanoff, 2005), states have consistently responded to opposition in ways that allowed the development of biotechnologies to proceed. Despite a cascade of visible controversies, states have generally managed to convert volatile mixtures of concerns into more stable forms. In culturally specific ways, unstable blends of technical uncertainties, science fiction scenarios, critiques of corporate power, and misgivings about hubris and slippery slopes, have largely been discursively distilled into categories amenable to bureaucratic management and expert decision-making (Jasanoff, 2012).

Risk Management

The process of 'thinning' multidimensional controversies into manageable technical problems susceptible to regulatory solutions is well illustrated by what has come to be known as the rDNA debate (*see* Risk, Sociology and Politics of). During the 1960s and early 1970s, scientists and other observers often construed the arrival of genetic engineering – viewed, once again, through the frame of radical discontinuity – as a portentous development that raised many troubling choices for society. As researchers became convinced that molecular techniques would soon make it possible to modify the genomes of living organisms, a number of prominent scientists warned that developments in biology could be dangerously misapplied, and they called for broad public discussion of such possibilities as the genetic engineering of humans. However, soon after researchers began splicing genes into bacteria and viruses, the focus of discussion shifted from broad concerns about the long-term implications of genetic engineering to much narrower and immediate questions about the laboratory hazards of rDNA research (Wright, 1994). In particular, scientists familiar with the early gene splicing experiments worried about accidentally creating dangerous microorganisms, resulting in catastrophic epidemics or disrupting ecosystems. The problem of providing credible scientific arguments that rDNA research could be safely conducted posed significant challenges, given limited evidence and technical debate about the magnitude of the hazards, and in 1974, 10 prominent biologists published a call for a partial moratorium on certain rDNA experiments.

The debate about the 'moratorium' and the physical hazards of rDNA research threw the future of genetic engineering into doubt, but it also allowed broader ethical and political questions to slip into the background (Gottweis, 1998; Wright, 1994). Moreover, scientific experts soon developed a discursive framework that provided a technical strategy for controlling the laboratory risks. This strategy, which took shape at a famous international meeting in Asilomar, California, relied on a system of classification (experiments would be grouped according to their expected level of hazard) and a gradient of controls (e.g., physical barriers, such as negative air pressure, and biological controls, such as 'disarmed' bacterial strains unable to survive outside the laboratory). On both sides of the Atlantic, the strategy of requiring increasingly stringent containment for increasingly risky experiments formed the basis for guidelines or regulatory controls, designed by expert committees that allowed research to proceed.

Containing the microbes within the laboratory became a means of containing fears and allowing genetic engineers to pursue the unbounded possibilities they believed to lie before them. As the 1970s progressed, these controls were

progressively relaxed as a growing number of scientists concluded that the danger of devastating epidemics was remote. Ultimately, the resolution of the controversy inspired a second-order debate about whether scientists, policymakers, and publics had responded appropriately or excessively to the risks of novel organisms. In contemporary debates over biotechnology, the rDNA debate has acquired an almost mythic status (Hurlbut, 2014): at times, it serves as a cautionary tale about the dangers of irrational fear of technology; more often, it provides a commendable example of scientists taking responsibility for the social implications of their work.

In the 1980s and 1990s, new regulatory challenges emerged as agricultural biotechnologists sought to move GMOs from the laboratory to the field – first on an experimental basis and later, in full-scale production (*see* Agricultural Science and Technology). In this context, risk management shifted from containment to deliberate release into the environment, a move that engaged new sources of expertise, such as ecologists, in risk assessment. Nevertheless, fears grew that agricultural biotechnology might accidentally produce new superweeds, alter ecosystems, and expose unsuspecting consumers to the risks of allergic reactions. By the late 1990s, the efforts of US companies, notably Monsanto, to introduce GM crops into Europe were encountering intense opposition from activists warning about the dangers of 'Frankenfoods' and calling for their ban. Framing their arguments in terms of the consumer's right to know, opponents of agricultural biotechnology also demanded that all GM foods be explicitly labeled as such. Soon, the technology was embroiled in a global debate turning not only on issues of safety, narrowly construed, but also on questions of corporate power in agriculture and the danger that reliance on GM seeds would trap the poor farmers of the developing world into dependency. Supporters of agricultural biotechnology contended that GM crops are a green and pro-poor technology, reducing risks from conventional pesticides and conveying useful traits such as drought resistance or nutritional enhancements.

The reaction to GM foods in different countries over the last two decades is too diverse for easy generalization. Extensive controversy broke out in Europe and India, and GM-free movements mobilized in Brazil, South Africa, and elsewhere (Stone, 2010). Jasanoff (2005) provides an instructive comparison of Europe (especially Germany and the United Kingdom) with the United States, where agricultural biotechnology provoked relatively less public reaction. The US Food and Drug Administration was able to 'naturalize' GM foods using a regulatory strategy that defined them as 'substantially equivalent' to non-GM foods, a move that not only allowed GM products to be brought to market but also forestalled labeling. In the United Kingdom, however, the issue could not be contained within the domain of regulatory deliberation. Environmental and consumer groups actively opposed the introduction of GM crops and foods, leading major supermarket chains to withdraw GM products (Jasanoff, 2005), and spurring the government to undertake an inquiry into the public acceptability of GM, including an unprecedented 2-year public participation exercise known as 'GM Nation' in 2001–02 (Horlick-Jones et al., 2007). Opposition movements also took hold on the European continent, and in 1999, five European Union (EU) member states decided to block the introduction of GM products until a system for identifying and labeling them was put into place. In contrast to the United States, the EU decided that rather than treating GM foods and crops (absent evidence to the contrary) as equivalent to their 'natural' counterparts, it would treat them as a class of biological entities requiring regulatory action. Within a few years, the EU had put in place a regulatory framework for labeling GM products and making them 'traceable' as they move through the food system (Lezaun, 2006). The market for GM foods and the percentage of land used for GM crops remain much smaller in Europe than in the United States (Stone, 2010).

Observers have often attributed the contrasting reception of GM foods in Europe and the United States to public ignorance or simple material interests, but careful analysis suggests that the deeper reasons stem from variations in collective modes of public knowledge making and public reasoning – what Jasanoff (2005) calls 'civic epistemologies.' The ongoing debate over GM foods and crops thus suggests that the ability of risk management techniques to depoliticize conflicts over biotechnology is far from universal, depending deeply on specific cultural and institutional contexts, especially at the level of the nation.

Bioethics

The policy narratives and practices of risk management offered one means for translating the unruly politics of biotechnology into governable forms; bioethics offered another – especially, regarding human biotechnology and molecular medicine. The potential of the new biology to destabilize, redefine, or deliberately alter the identities of individuals, groups, or 'human nature' often generated apprehension or resistance, even as it also inspired excitement and opened possibilities for constructing new subjectivities (Rabinow, 1996a; Rose, 2007; *see* Human Sciences, History of; Biomedical Sciences, History and Sociology of). Encounters with genetic medicine often cast people into the problematic role of 'moral pioneers' (Rapp, 2000). Some observers worried that predictive genetic testing would threaten privacy and deprive people of uncertainty about fate; others feared that genetics would reinforce racial prejudices or produce new stigmatized groups; still others objected to the desacralization and commodification of human life. Since the late 1990s, the creation of large collections of human DNA, so-called 'biobanks,' inspired concerns about privacy, control, and governance (e.g., Corrigan and Tutton, 2009). Such worries that did not neatly fit into the technical and economistic discourse of risk management were typically addressed through the interpretive frameworks and institutions of bioethics.

During the 1970s and 1980s, bioethics emerged as an institutionalized area of inquiry and technique for decision-making. The bioethics phenomenon, like the growth of risk management, was by no means confined to biotechnology. In medicine, health care, and clinical research, bioethics – performed in such diverse sites as academic literature, national commissions, hospital ethics committees, and,

most recently, biotechnology corporations – has grown into a prominent feature of public life (DeVries and Subedi, 1998; Kleinman et al., 1999; Rothman, 1991). The rise of biotechnology contributed to the rise of bioethics, fueling its growth with a cascade of controversies about genetic testing, gene therapy, and reproductive technology – matters that increasingly came to be seen as part of the professional jurisdiction of a new cadre of bioethics experts (*see* Science and Technology Studies: Experts and Expertise). Indeed, by the late 1980s, when the Human Genome Project was proposed, the notion that bioethical analysis could play a major role in addressing the societal dimensions of the new biology was sufficiently widespread that national governments flocked to build 'ethics' into their programs of genomics research (Cook-Deegan, 1994).

The peculiar relationship between the domains of bioethics and biotechnology is nicely illustrated by the politics of governmental programs on the ethical, legal, and social implications (ELSI) of genomics. (In Europe, these programs tended to avoid the arguably deterministic word 'implications' in favor of the more open term 'aspects,' and hence were known as ELSA programs.) A broad consensus supporting ELSI/ELSA programs often coexisted with deep disagreement about what these programs should aim to accomplish. The US ELSI program, for example, was celebrated as an important innovation in science policy and presented as proof that social concerns would not be neglected, yet the program was also beset by criticism from many directions. It was accused of being both hypercritical and unduly promotional of genetic technology; of taking an overly academic and long-term view and of narrowly focusing on the immediate; of being too small to address the pace of change and of being a waste of valuable research funds. But such debates about the shape of these programs missed the more pervasive, if subtle, ways that bioethics functions as a quasiregulatory institution in contemporary states. Bioethics provided an idiom and institutional space for defining controversies about human genetics and genomics as ethical issues, better addressed through moral reflection than mobilization. Thus bioethics commissions often served to narrow debate, focusing attention on 'thin' questions of means rather than 'thicker' debate about ends (Evans, 2002).

To be sure, the ability of bioethics to narrow and contain public debate was incomplete, as the controversy about ESCs in the United States nicely illustrates. In this case, debate over ESC research, tightly connected to religious beliefs about the sanctity of life and the divisive national controversy over abortion, ended up being fought in the arena of electoral politics, and the ongoing debate featured attacks on national bioethics commissions for having 'politicized bioethics.' But despite its limitations, bioethics offered a tool for shoring up the legitimacy of political institutions, suggesting that states could find rational and secular means for drawing moral boundaries to replace the apparently natural ones that the biotechnology revolution erased. In this way, bioethics – like other forms of expert advice (Jasanoff, 1990; Hilgartner, 2000) – has become an important means not merely for making decisions but also for reordering societies challenged by rapid technological and social change.

See also: Bioethics in the Post-genomic Era; Bioethics: Genetics and Genomics; Bioethics: Philosophical Aspects; Biomedical Sciences, History and Sociology of; Cultural Evolution: Overview; Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute; Forensic Genetic Databases: Ethical and Social Dimensions; Genetic Engineering; Genomics, Ethical Issues in; Human Genome Project: History and Assessment; Medicine, History of Western.

Bibliography

Balmer, B., 2012. Secrecy and Science: A Historical Sociology of Biological and Chemical Warfare. Ashgate, Farnam, UK.
Boyle, J., 1996. Shamans, Software, and Spleens: Law and the Construction of the Information Society. Harvard University Press, Cambridge, MA.
Brown, N., Rappert, B., Webster, A. (Eds.), 2000. Contested Futures: A Sociology of Prospective Techno-Science. Ashgate, Farnam, UK.
Bud, R., 1993. The Uses of Life: A History of Biotechnology. Cambridge University Press, Cambridge.
Calvert, J., 2013. Systems biology, big science and grand challenges. BioSocieties 8 (4), 466–479.
Cambrosio, A., Keating, P., 1995. Exquisite Specificity: The Monoclonal Antibody Revolution. Oxford University Press, New York.
Campos, L., 2012. The BioBrickTM road. BioSocieties 7 (2), 115–139.
Carlson, R.H., 2010. Biology Is Technology: The Promise, Peril, and New Business of Engineering Life. Harvard University Press, Cambridge, MA.
de Chadarevian, S., 2002. Designs for Life: Molecular Biology after World War II. Cambridge University Press, Cambridge.
Conway, G., 1997. The Doubly Green Revolution: Food for All in the 21st Century. Cornell University Press, Ithaca, NY.
Cook-Deegan, R., 1994. The Gene Wars: Science, Politics, and the Human Genome. Norton, New York.
Corrigan, O., Tutton, R., 2009. Biobanks and the challenges of governance, legitimacy, and benefit. In: Atkinson, P., Glasner, P.E., Lock, M. (Eds.), Handbook of Genetics and Society: Mapping the New Genomic Era. Routledge, New York, pp. 302–318.
Davies, G., Frow, E., Leonelli, S., 2013. Bigger, faster, better? rhetorics and practices of large-scale research in contemporary bioscience. BioSocieties 8 (4), 386–396.
DeVries, R., Subedi, J. (Eds.), 1998. Bioethics and Society: Constructing the Ethical Enterprise. Prentice-Hall, Upper Saddle River, NJ.
Evans, J.H., 2002. Playing God? Human Genetic Engineering and the Rationalization of Public Bioethical Debate. University of Chicago Press, Chicago.
Fortun, M., 2008. Promising Genomics: Iceland and DeCode Genetics in a World of Speculation. University of California Press, Berkeley, CA.
Fujimura, J.H., 1996. Crafting Science: A Sociohistory of the Quest for the Genetics of Cancer. Harvard University Press, Cambridge, MA.
Gaudillière, J.-P., Rheinberger, H.-J. (Eds.), 2004. From Molecular Genetics to Genomics: The Mapping Cultures of Twentieth-Century Genetics. Routledge, New York.
Gottweis, H., 1998. Governing Molecules: The Discursive Politics of Genetic Engineering in Europe and the United States. MIT Press, Cambridge, MA.
Gottweis, H., Petersen, A., 2008. Biobanks: Governance in Comparative Perspective. Routledge, Abingdon, UK.
Hedgecoe, A., 2004. The Politics of Personalized Medicine: Pharmacogenetics in the Clinic. Cambridge University Press, Cambridge.
Helmreich, S., 2009. Alien Ocean: Anthropological Voyages in Microbial Seas. University of California Press, Berkeley, CA.
Hilgartner, S., 2000. Science on Stage: Expert Advice as Public Drama. Stanford University Press, Stanford.
Hilgartner, S., 2009. Intellectual property and the politics of emerging technology: inventors, citizens, and powers to shape the future. Chicago-Kent Law Review 84, 197–224.

Hilgartner, S., 2012. Staging high-visibility science: media orientation in genome research. In: Rödder, S., Franzen, M., Weingart, P. (Eds.), The Sciences' Media Connection: Public Communication and its Repercussions (Sociology of the Sciences Yearbook). Springer, Dordrecht, pp. 189–215.

Hindmarch, R., Prainsack, B. (Eds.), 2010. Genetic Suspects: Global Governance of Forensic DNA Profiling and Databasing. Cambridge University Press, Cambridge.

Horlick-Jones, T., Walls, J., Rowe, G., et al., 2007. The GM Debate: Risk, Politics and Public Engagement. Routledge, Abingdon, UK.

Hurlbut, B., 2014. Remembering the future: science, law, and the legacy of asilomar. In: Jasanoff, S., Kim, S.-H. (Eds.), Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power. University of Chicago Press, Chicago.

Hutchison, C., 2007. DNA sequencing: bench to bedside and beyond. Nucleic Acids Research 35, 6227–6237.

Jasanoff, S., 1990. The Fifth Branch: Science Advisers as Policymakers. Harvard University Press, Cambridge, MA.

Jasanoff, S., 1995. Science at the Bar: Law, Science, and Technology in America. Harvard University Press, Cambridge, MA.

Jasanoff, S., 2005. Designs on Nature: Science and Democracy in Europe and the United States. Princeton University Press, Princeton, NJ.

Jasanoff, S., 2012. Product, process, or programme: three cultures and the regulation of biotechnology. In: Jasanoff, S. (Ed.), Science and Public Reason. Routledge, New York, pp. 23–41.

Kass, L.R., 2002. Life, Liberty and the Defense of Dignity: The Challenge for Bioethics. Encounter Books, San Francisco.

Kay, L.E., 2000. Who Wrote the Book of Life? A History of the Genetic Code. Stanford University Press, Stanford, CA.

Keller, E.F., 1995. Refiguring Life: Metaphors of Twentieth-Century Biology. Columbia University Press, New York.

Keller, E.F., 2002. The Century of the Gene. Harvard University Press, Cambridge, MA.

Kenney, M., 1986. Biotechnology: The University-Industrial Complex. Yale University Press, New Haven, CT.

Kenney, M., 1998. Biotechnology and the creation of a new economic space. In: Thackray, A. (Ed.), Private Science: Biotechnology and the Rise of the Molecular Sciences. University of Pennsylvania Press, Philadelphia, pp. 131–143.

Kevles, D.J., Hood, L. (Eds.), 1992. The Code of Codes: Scientific and Social Issues in the Human Genome Project. Harvard University Press, Cambridge, MA.

Kleinman, A., Fox, R.C., Brandt, A.M., 1999. Biothics and beyond. Special Issue Daedalus: Journal of the American Academy of Arts and Sciences 128 (4), v–325.

Koenig, B.A., Lee, S.S.-J., Richardson, S.S., 2008. Revisiting Race in a Genomic Age. Rutgers University Press, New Brunswick, NJ.

Landecker, H., 2007. Culturing Life: How Cells Became Technologies. Harvard University Press, Cambridge, MA.

Lezaun, J., 2006. Creating a new object of government: making genetically modified organisms traceable. Social Studies of Science 36, 499–531.

Lynch, M., Cole, S.A., McNally, R., Jordan, K., 2008. Truth Machine: The Contentious History of DNA Fingerprinting. University of Chicago Press, Chicago.

Mirowski, P., 2011. Science Mart: Privatizing American Science. Harvard University Press, Cambridge, MA.

Nelson, N., 2012. Shooting genes, distributing credit: narrating the development of the biolistic gene gun. Science as Culture 21, 205–232.

Nelson, N., Geltzer, A., Hilgartner, S. (Eds.), 2008. The Anticipatory State: Making Policy-Relevant Knowledge about the Future. Special Issue. Science and Public Policy, vol. 8, pp. 546–606.

OECD, 2009. The Bioeconomy to 2030: Designing a Policy Agenda. Organization for Economic Cooperation and Development. www.oecd.org/publishing/corrigenda (accessed 05.11.13.).

Parthasarathy, S., 2007. Building Genetic Medicine: Breast Cancer, Technology, and the Comparative Politics of Health Care. MIT Press, Cambridge, MA.

Pottage, A., 2006. Too much ownership: bio-prospecting in the age of synthetic biology. BioSocieties 1, 137–158.

Prainsack, B., 2006. Negotiating life: the regulation of embryonic stem cell research and human cloning in Israel. Social Studies of Science 36, 173–205.

Rabinow, P., 1996a. Essays on the Anthropology of Reason. Princeton University Press, Princeton, NJ.

Rabinow, P., 1996b. Making PCR: A Story of Biotechnology. University of Chicago Press, Chicago.

Rabinow, P., Dan-Cohen, T., 2005. A Machine to Make a Future: Biotech Chronicles. Princeton University Press, Princeton, NJ.

Rapp, R., 2000. Testing Women, Testing the Fetus: The Social Impact of Amniocentesis in America. Routledge, New York.

Reardon, J., 2005. Race to the Finish: Identity and Governance in an Age of Genomics. Princeton University Press, Princeton, NJ.

Rheinberger, H.-J., 1997. Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube. Stanford University Press, Stanford, CA.

Rifkin, J., 1983. Algeny: A New Word – A New World. Viking, New York.

Rose, N., 2007. The Politics of Life Itself. Princeton University Press, Princeton, NJ.

Rothman, D.J., 1991. Strangers at the Bedside: A History of How Law and Bioethics Transformed Medical Decision Making. Basic Books, New York.

Shapin, S., 2008. The Scientific Life: A Moral History of a Late Modern Vocation. University of Chicago Press, Chicago.

Stone, G.D., 2010. The anthropology of genetically modified crops. Annual Review of Anthropology 39, 381–400.

Sunder Rajan, K., 2006. Biocapital: The Constitution of Postgenomic Life. Duke University Press, Durham, NC.

Thackray, A. (Ed.), 1998. Private Science: Biotechnology and the Rise of the Molecular Sciences. University of Pennsylvania Press, Philadelphia.

Vogel, K., 2012. Phantom Menace or Looming Danger? A New Framework for Assessing Bioweapons Threats. Johns Hopkins University Press, Baltimore.

Wright, S., 1994. Molecular Politics: Developing American and British Regulatory Policy for Genetic Engineering, 1972–1982. University of Chicago Press, Chicago.

Bipolar (Manic-Depressive) Disorder

Leonardo Tondo, Harvard Medical School, Boston, MA, USA; McLean Hospital, Belmont, MA, USA; and Lucio Bini Mood Disorders Centers, Cagliari and Rome, Italy
Ross J Baldessarini, Harvard Medical School, Boston, MA, USA; and Mailman Research Center, McLean Hospital, Belmont, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Bipolar (manic-depressive) disorder is a severe syndrome involving episodic shifts of mood into mania or depression, with variations in severity, frequency, clinical presentation, and complexity of episodes. Associated risks are high rates of psychosis, anxiety syndromes, and substance abuse. Mortality is increased due to suicide, accidents, and medical complications of substance abuse, as well as stress-sensitive medical disorders. Lifetime prevalence for relatively severe cases is probably more than 2% of the general population, affecting men and women similarly; half of cases begin before age 20. There is a strong familial predisposition but no identified genetics or etiology. Mania sometimes arises in association with neuromedical disorders or reactions to drugs. The long-term course is quite predictable in individuals from early morbidity patterns, but unpredictable across patients. Treatment of mania involves antipsychotics, anticonvulsants, or sedative medicines for mania. Treatment of bipolar depression remains unsatisfactory and the place of antidepressants in such treatments remains controversial. Long-term maintenance treatment with lithium salts or certain anticonvulsants seeks to avoid recurrences, with only partial control, especially of depression. Psychosocial and rehabilitative interventions also are useful components of comprehensive care.

Introduction and Historical Background

Bipolar disorder, a form of manic-depressive illness, is a typically episodic, often severe, and sometimes disabling psychiatric illness characterized by marked changes in mood, thinking, and behavior (Goodwin and Jamison, 2007). Bipolar disorder is a major public health problem, though sometimes underrecognized, and commonly diagnosed as major depressive or other disorders before manic episodes occur; hypomania sometimes is not perceived by patients or recognized by physicians as abnormal mood elevation. Characteristically, symptomatic changes in the disorder present in extremes of mania and depression, or with simultaneous excited and depressive ('mixed') features.

Clinical descriptions of mania and melancholia date at least to Hippocratic physicians (*c*.400 BCE), although early cases probably included neuromedical disorders with associated excitement or withdrawal. Aretæus of Cappadocia (*c*.CE 150) is credited with early descriptions of associations of mania and melancholia in the same person at different times (Baldessarini et al., 2014a). Elaborations of such associations continued during the second millennium, including description of a *manic-melancholic syndrome* by Piquer-Arrufat in the 1700s. In the 1800s, Baillarger and Falret added descriptions of *double or circular insanity*, and Kahlbaum and Hecker described *cyclothymia* (Baldessarini et al., 2014a). At the end of the nineteenth century, Kraepelin conceived a more comprehensive and influential description of *manic-depressive insanity* to include patients with recurrent melancholic episodes and others with manias as well as depressions, along with similar, milder conditions (Trede et al., 2005).

Contemporary distinctions of bipolar (manic and depressive) and unipolar (depressive) forms of recurrent major mood disorders evolved in the writings of Kleist and Leonhard from the 1950s, and later, Angst, Perris and Winokur (Baldessarini et al., 2014a). In the 1970s, bipolar disorder was subdivided into types I (with mania) and II (with only hypomania), both with recurrent depression (APA, 2013). In the 1980s and 1990s, an even broader 'bipolar spectrum' disorder that included recurrent depression with mild hypomania (Goodwin and Jamison, 2007), leading to total lifetime prevalence-estimates of bipolarlike disorders that are similar to those of unipolar major depressive disorders (Angst et al., 2010).

Mania

Episodes of mania present with an increased rate of thinking and speaking, increased physical and sexual energy, decreased need for sleep, changes in appetite, and possible aggressive behavior. Thinking is typically rapid, pressured, overproductive and semilogical, often with a grandiose flavor. As a result, attention and concentration are impaired. Manic persons are often distractible, with continuous shifting of loosely connected topics that can be hard to follow. Attempts to interrupt or restrain pressured behavior often provoke irritable or angry responses. Mania can be associated with paranoid or grandiose delusions. Overactivity is incessant but typically disorganized and unproductive, often without clear or well sustained aims. Abnormal behavior appears to derive from grossly excessive self-confidence, and includes lack of social inhibition, intrusiveness, impulsivity, recklessness, irritability and sometimes aggressiveness. Risk-taking is common, and may include unrealistic plans or implausible business schemes, impulsive overspending or gambling, disinhibited or increased sexual activity, or reckless driving. Impulsivity and impaired judgment can lead to accidents and injuries, especially when complicated by intoxication.

Bipolar Disorders with Hypomania

Fully expressed mania usually involves disruptive or dangerous behavior, usually requiring emergency treatment. *Hypomania* is a milder form of sustained, unnatural mood elevation that may not disrupt occupational and interpersonal relationships, and may require intervention or treatment, but rarely on an emergency basis. In hypomania as well as in mania, judgment and critical self-awareness often are impaired. Hypomania is characteristic of type II bipolar disorder, which presents clinically mainly as recurring episodes of major depression. This syndrome is not a lesser form of bipolar disorder, but includes high rates of disability and excess mortality similar to those encountered in type I bipolar disorder (Novick et al., 2010; Merikangas and Lamers, 2012).

Even milder forms of hypomanic features may arise in persons with recurrent depression, and have been considered part of a 'bipolar spectrum' of disorders related to bipolar II disorder (Angst et al., 2010). *Cyclothymia* is a form of bipolar disorder of limited severity whose diagnostic criteria are not well established and overlap those of type II bipolar disorder as well as certain personality disorders (Goodwin and Jamison, 2007). Antecedents, personality types or temperaments sometimes associated with bipolar disorder can be marked by sustained high energy, productivity, and social interaction ('hyperthymic' type) or long-term moodiness with moderate fluctuations between elevated and depressed mood and behavior ('cyclothymic' type).

Bipolar Depression

Bipolar disorder includes depressive episodes in forms ranging from profound inertia and withdrawal to severe, and sometimes psychotic major episodes, or sustained *dysthymia* of moderate intensity, as well as *dysphoric* episodes in which depressed mood is mixed with increased energy or irritable agitation (Baldessarini et al., 2010b). Characteristically, major episodes of bipolar depression present with psychomotor retardation or slowing of thought and activity with loss of interest in usual or pleasurable activities and impaired self-care. Along with a low level of energy, lassitude, and loss of sexual libido, abnormalities of sleep and appetite are usually found. These may include increased appetite and sleep (so-called 'atypical' features) or the anorexia and insomnia more typical of major depressive illnesses. Sluggish thinking, impaired concentration, and inability to make decisions are common, as well as feelings of worthlessness, guilt and self-blame may be present. Self-assessments in all forms of severe depression often are severely distorted, unrealistically negative and pessimistic and can alter recollections of previous phases of illness. Sometimes, somatic complaints such as migrainelike headaches, aching muscles or joints, gastrointestinal, respiratory, cardiac, and other somatic symptoms may be more prominent than psychic symptoms. Despair and hopelessness are common, and there is a greatly increased risk of suicide during depressive or more energized, dysphoric-agitated phases of the illness, with much less risk of suicide during hypomania or mania. Nevertheless, self-injuries and accidents can occur in all phases of the disorder.

Other Psychopathological Features

Bipolar disorder can present with mixed manic-depressive states that include both depressive and manic features at the same time or in rapid alternation. Both mania and hypomania can be associated with dysphoria and agitation, or may fluctuate rapidly within minutes or hours, between predominantly manic and depressive states ('emotional lability') and so can lead to the diagnosis of a bipolar 'mixed state.' Indeed, many, if not most episodes of mania are marked by emotional and behavioral lability and mixtures of depressive as well as excited features.

Bipolar disorder as well as severe major depressive disorder is associated with evidence of cognitive impairment that tends to be most prominent during acute episodes or illness, particularly depression, but may also be sustained between recurrences (Harvey et al., 2010). Long-term pharmacological treatment of bipolar disorder may contribute to reduced cognitive functioning, psychomotor slowing, and depressed mood. Such effects may not always be adequately evaluated and are usually considered to be minor (Baldessarini, 2013).

In addition, psychotic symptoms are found in perhaps three-quarters of bipolar disorder patients at some time. Bipolar disorder is, arguably, the most frequent form of idiopathic psychotic illness, exceeding even schizophrenia in prevalence (Goodwin and Jamison, 2007). Most commonly the delusions of bipolar disorder are paranoid, with either grandiose or persecutory themes (often pertaining to religious or other fame- or power-based exaggerations), particularly in manic or mixed episodes. Nihilistic or guilty delusions and other pessimistic exaggerations may occur in depressive phases of the illness. Self-referential misinterpretations of random events, as well as auditory hallucinations, may occur in manic and, less often, in depressive episodes. Psychotic features typically are congruent with the current dominant mood. Mood-incongruent delusions may involve impressions that thoughts are being externally controlled, inserted by, or broadcast to other persons. Some patients who have been diagnosed with bipolar disorder are later considered to have a *schizoaffective* disorder, as they show an excess of sustained, psychotic over affective features. They often become chronically ill with severe disability. Such disorders are currently considered as related to schizophrenia in current nosological systems, and their differentiation from severe bipolar disorder can be challenging (Pagel et al., 2013). Schizoaffective disorders also represent a challenge for the dualistic, Kraepelinian, division of severe or psychotic mental illnesses into a mainly mood-disordered episodic (manic-depressive) group and a more psychotic and chronic group (dementia præcox or schizophrenia). In general schizoaffective disorders remain inadequately defined or studied (Pagel et al., 2013; Salvatore et al., 2013a).

Co-Occurring Psychiatric Illnesses

Persons diagnosed with bipolar disorder often show features of other psychiatric syndromes, particularly major anxiety disorders (in about half of patients), including panic disorder, obsessive-compulsive disorder, or lesser forms of generalized

anxiety or particular phobias, and may meet contemporary, categorical diagnostic criteria for more than one disorder. It is uncertain whether such co-occurring conditions represent separate illnesses or are part of the spectrum of symptoms ('phenotype') of bipolar disorder. To some extent, the concept of 'comorbidity' may be an artifact of contemporary, categorical diagnostic systems (Vázquez et al., 2014).

In addition, abuse of alcohol or drugs occurs often in patients with bipolar disorder, and can severely complicate clinical management and worsen the overall prognosis (Goodwin and Jamison, 2007; Farren et al., 2012). Whether substance abuse is a contributing cause of bipolar disorder, or the disorder favors addictive behavior remains unresolved.

Psychodynamic Characteristics

Psychodynamic characterizations of bipolar disorder patients recognize prominent narcissistic features, lack of empathy, instability of interpersonal relationships, use of denial and activity defenses in hypomania, and guilt-driven self-aggressiveness. Mania has been considered an attempt to avoid or to compensate for depressive phases. Depression has been interpreted as a manifestation of self-aggression in response to the loss of a loved object (psychoanalytic perspective), as a consequence of learned helplessness after repeated, unavoidable, adverse experiences (developmental-behavioral model), or an expression of guilt for excesses during manic episodes (phenomenological view), or as an expression of negative self-evaluation activated by stress and learning (cognitive approach). Both polarities of bipolar disorder share common traits that may include emotional dependence, deficient awareness of internal emotional responses in oneself or others, and a lack of appreciation of differences between actual and idealized situations.

Diagnosis

The presence of at least one episode of mania is sufficient for a diagnosis of type I bipolar disorder, though most cases involve a series of manic and depressive recurrences over years of illness. Recurrences of clinical depression and hypomanic episodes (sometimes simultaneously) are considered to have type II bipolar disorder. Both types are recognized in the World Health Organization's *International Classification of Diseases* (ICD-10; WHO, 2000) and the American Psychiatric Association's *Diagnostic and Statistical Manual* (DSM-5; APA, 2013). The latter also recognizes 'rapid-cycling' bipolar disorder, with at least four episodes of mania or depression within 12 months of clinical assessment, as well as 'mixed-states' as a qualifying characterization that can be added to mania, hypomania, or bipolar depressive episodes (APA, 2013).

Sometimes mania arises in association with exposure to particular drugs or toxins, or with neurological or general medical disorders. Such responses are commonly associated with antidepressant treatment for depression or anxiety, or with stimulants for attention deficits, less commonly during treatment with anti-inflammatory corticosteroids, and rarely with use of quininelike antimalarial drugs, and may arise in dementia or following brain injury (Krauthammer and Klerman, 1978; Offidani et al., 2013).

Course of the Illness

The onset of bipolar disorder is often unrecognized as such, and several years may precede diagnosis and appropriate treatment. Many cases start in adolescence, and some before puberty when their distinction from attention or conduct disorders of childhood is often difficult, and they may arise from a background of depressive or anxiety disorders (Offidani et al., 2013; Salvatore et al., 2013b). More than half of bipolar cases present initially with an episode of mania or hypomania, particularly in men, whereas initial depressions are more common in women. Women also tend to have more type II bipolar illness and prominent depressive episodes, often with relatively frequent recurrences. Some cases present with recurrent depression, and mania or hypomania emerge later, sometimes during treatment with an antidepressant. Such cases are prevalent in adolescents and young (Baldessarini et al., 2013).

The natural duration of untreated episodes of mania or bipolar depression is about 6–9 months, with somewhat longer depressions, and approximately yearly recurrences, on average. With modern psychopharmacological treatments, acute symptomatic episodes of mania last about 1–2 months, and depression, 2–4 months, but full recovery may require several more months, and full functional recovery may not occur (Tohen et al., 2003). Without treatment, episodes of mania are slightly more frequent than depressive episodes, but total time in depression exceeds that in mania. This pattern currently may reflect generally inferior long-responses to available treatments for depressive versus manic phases of bipolar disorder (Baldessarini et al., 2010a,b).

Single episodes of mania with or without depressions, as well as single episodes of depression with one or more hypomanic or manic episodes occur in small proportions of cases. Instead, episodic recurrences are the rule and their course over years tends toward randomness among patients, without consistent progression toward shortening cycles, provided that analyses are adjusted for years of illness (Oepen et al., 2004; Baldessarini et al., 2012a). Although cycling frequency and patterns may be random among bipolar disorder patients, there is considerable predictability about the predominant types of recurrent episodes in individual patients, often with the same or similar polarity of initial episodes (mania predicting an excess of later mania, psychosis, and hypomania; depression predicting later depression; anxiety predicting later depression and anxiety; mixed-states predicting more depression or future mixed states of agitated dysphoria), as well as polarity-sequences similar to initial illness-cycles (Baldessarini et al., 2010a, 2014b). In addition, there is a generally worse long-term course and less responsiveness to mood-stabilizing treatments among bipolar disorder patients who usually present in depression and then cycle into mania (depression-mania-interval course-pattern), compared to those who usually present in mania and later become depressed (mania-depression-interval pattern) (Koukopoulos et al., 2013).

Pediatric-onset bipolar disorder remains a controversial topic, including its likely adult outcomes and its optimal and

safe treatment (Geller and DelBello, 2003). Such illnesses typically are marked by emotional turmoil and disturbances of behavior and sleep in childhood and involve highly unstable mood and behavior over periods of months or years. A pattern of discretely recurring episodes of mania and depression becomes established during adolescence and is usual in adults (Goodwin and Jamison, 2007). Long-term studies of outcomes among patients considered to have had 'juvenile-onset bipolar disorder' are emerging, and suggest that some may not progress to typical adult bipolar disorder (Geller et al., 2008; Baldessarini et al., 2013).

Mania beginning after age 60 years often follows a severe course, high recurrence rate, poor responsiveness to treatment, and a high mortality rate. Some such cases may be associated with early dementia (Dols et al., 2013a).

An unfavorable illness-course and inferior response to treatment are associated with rapid-cycling, usually defined as four or more recurrences of mania or depression within a single year, as is somewhat more common among women and persons with type II bipolar disorder (Tondo et al., 2003a). This pattern occurs in about 15% of cases, though only a fraction of such persons sustain high rates of recurrence over many years. An uncertain proportion of cases of bipolar disorder follow a rather consistent hemicircannual or *seasonal* course, commonly but not always, with depressions in fall or winter and manias in spring or summer (Faedda et al., 1993). This recurrence pattern is more likely with greater distance from the earth's equator, among women, and in younger persons.

A majority of bipolar disorder patients experience some degree of recovery between episodes of acute illness, and euthymic or stable periods often allow them to re-establish previous activities and relationships. Nevertheless, even with modern treatment, bipolar disorder can lead to sustained occupational and social dysfunction, particularly following youthful onset, psychotic symptoms, and early hospitalization (Tohen et al., 2003). Poorer outcomes tend to associate with very early or very late onset-age, frequent mixed episodes or psychotic features, severe and sustained depression, with a tendency toward sustained relatively rapid cycling, and as a consequence of inadequately treated episodes (Goodwin and Jamison, 2007; Farren et al., 2012; Swann et al., 2013). Long-term outcome in bipolar disorder reflects the secondary, adverse psychological and social consequences of the illness, associated stigmatization, and their often profound impact on self-esteem, independent functioning, and interpersonal relations.

Increased Mortality

Premature mortality is a strikingly common outcome in bipolar disorder, in part owing to its strong association with suicide. The *standardized mortality ratio* for suicide compared to the general population matched for age and sex may exceed 20:1 among bipolar disorder patients. More than one-third of bipolar disorder patients make at least one serious suicide attempt, and at least 10% of their deaths are ascribed to suicide indicating that the illness represents a nearly necessary, but not a sufficient, condition for suicide (Tondo et al., 2003b, 2007; Goodwin and Jamison, 2007; Manchia et al., 2013). Suicidal risks are similarly high in types I and II bipolar disorder (Novick et al., 2010). Suicide attempts have a particularly ominous implication in bipolar disorder (and in severe major depression), in that the *index of lethality* (ratio of attempts to fatalities) is probably less than 5:1, whereas in the general population the ratio is approximately 20:1, indicating greater lethality in bipolar disorder (Baldessarini et al., 2012a). The depressive and mixed phases of bipolar disorder carry the highest suicide risk, and account for at least three-quarters of all attempts and fatalities. Mixed-states with both agitated turmoil and dysphoria or depression are particularly risky (Swann et al., 2013). Risk of completed suicide in bipolar disorder is greater for men than women, whereas attempts are more frequent among women (Tondo et al., 2003b, 2007). Other risk factors include previous attempts, older age, abuse of alcohol or drugs, and probably a family history of psychiatric illness that includes suicide (Tondo et al., 2003b; Baldessarini et al., 2012a).

In addition to suicide, mortality rates are elevated in bipolar disorder owing to effects of risk-taking behaviors and to substance abuse leading to accidental deaths, medical complications of prolonged substance-abuse, and the impact of stress-sensitive cardiovascular, pulmonary, and endocrine disorders, in which mortality is perhaps three times higher than in the general population, and possibly greater than in many unipolar depressed patients (Tondo et al., 2007; Ösby et al., 2001). Owing to the large proportion of medically ill, elderly persons in the population, the number of deaths per year among bipolar disorder patients due to co-occurring medical disorders is very similar to the number associated with suicides, accidents and other violent forms of death (Ösby et al., 2001; Crump et al., 2013).

Epidemiology

Bipolar disorders have a lifetime prevalence ranging from at least 1.0% (for mania), to as many as 2–5% of the general population if type II and cyclothymic cases are included, and may approach 10% if bipolar spectrum cases with recurrent depression and very mild hypomania are included, and so may be similar to rates for nonbipolar major depressive disorders (Angst et al., 2010). The annual incidence of type I bipolar disorders is at least 0.05%. Women have a moderately excessive risk for bipolar disorder (sex ratio about 1.4:1) that is much less than in nonbipolar major depressive disorder (Goodwin and Jamison, 2007). The median age-at-onset is about 20 years (Baldessarini et al., 2012b). There have also been secular trends for all major mood disorders to be recognized more commonly and at younger onset-ages over the past century. However, the contributions of improved case finding, broadening conceptualizations of mood disorders, and the probable impact of substance abuse and other environmental factors to these trends are not well defined.

Biology

The risk of bipolar disorder, depression, psychosis or suicide in first-degree relatives of index cases of bipolar disorder is

approximately 20–25%. The same risk is found in fraternal twins, but it is about 70% in identical twins, leading to a monozygotic:dizygotic risk ratio of more than 3:1 (Goodwin and Jamison, 2007). Rare adoption studies carried out to evaluate the role of genetic versus environmental factors have found more affective illness and suicides among biological relatives of index cases of major affective disorders than in the adopting family (Goodwin and Jamison, 2007; Craddock and Sklar, 2013). Despite this substantial epidemiological evidence of a familial contribution to the risk of affective illness, a specific mode of inheritance, let alone a discrete genetic marker or molecular defect specific to bipolar disorder has yet to be identified (Bondy, 2011; Craddock and Sklar, 2013).

Other biological findings also are far from definitive (Kupferschmidt and Zakzanis, 2011; Yang, 2011; Cousins and Grunze, 2012). Structural and functional brain imaging techniques are receiving much attention, but the findings remain inconclusive, and risk lack of specificity in comparisons with other forms of severe mental illness. Older findings of changes in biochemical measurements of stress-sensitive neurotransmitters and neurohormones have proved nonspecific by diagnosis, or incidental to altered general metabolic activity and dependent on the state of being ill. There is also a tendency toward early appearance of the rapid-eye-movement phase of sleep in mania as well as in severe depression, with other alterations of ultradian or circadian biorhythms (Salvatore et al., 2013b). Some of these may persist even in euthymic or well intervals between episodes of acute illness, suggesting potential value as 'trait' markers (Yang, 2011; Salvatore et al., 2013b). Relative cerebral deficiency of the monoamine neurotransmitter serotonin has been associated with aggression and suicide, but this is not specific to bipolar disorder. Cognitive and other neuropsychological deficits can also be identified in bipolar disorder, and some may be sustained between acute episodes of illness (Vöhringer et al., 2013).

Clinical Management

Mania and severe episodes of bipolar depression require careful clinical management. Manic patients need protection if they show impulsive, hostile or excessively uninhibited, risky behavior, or if they refuse treatment. Bipolar disorder patients risk exploitation by others who may take advantage of their hypersexuality and inappropriate generosity in hypomanic states. In extreme cases, patients may need to be committed to specialized hospital units where they can be treated in structured, calm, secure settings that can provide expert care. Depressed patients may require hospitalization because of a high risk of suicide or inability to care for themselves. In both circumstances, early transfer to partial-hospital, day-treatment, or intensive outpatient treatment following brief hospitalization is often recommended.

Psychotherapeutic efforts are much less extensively evaluated in bipolar than in unipolar depressive disorders, in part reflecting traditional views that bipolar patients are often less able to accept their condition as a disorder requiring treatment. Nevertheless, psychotherapy methods developed and scientifically evaluated for depressive disorders may also have utility in bipolar disorder when acute symptoms are adequately controlled medically. These include interpersonal psychotherapy, based on improving coping strategies for social and interpersonal relationships. Cognitive-behavioral approaches attempt to modify irrational or exaggerated concepts by encouraging development of more flexible schemas, and rehearsal of new cognitive and behavioral responses. Psychoeducational and group-based or individual supportive treatments are advisable, and usually feasible with medication (Lolich et al., 2012).

Patients and their families also benefit from education about the illness, its treatment and expected course. Counseling is also useful to plan for the possible impact of the illness on career, financial planning, and family relations. Genetic counseling regarding risks to offspring can be offered. Also appropriate are recommendations to avoid pregnancy by women during treatment intended to be mood-stabilizing (especially with potentially teratogenic anticonvulsants), and it is necessary to discuss of maternal risks of abrupt discontinuation of treatment in anticipation of, or at the start of pregnancy, as well as anticipation of the high risk of recurrences early in pregnancy (especially of depressive and mixed states), and even greater risks in the postpartum period (including mania, psychosis, and risk of violence) without treatment (Viguera et al., 2007).

Treatment

Treatment of acute episodes of mania primarily involves antipsychotics or other central depressants (including some anticonvulsants, such as carbamazepine and divalproex) and potent sedatives such as lorazepam (Yildiz et al., 2011; Baldessarini, 2013).

Treatment of bipolar depression remains difficult and controversial (Baldessarini et al., 2010b). Particularly controversial are the effectiveness and safely of antidepressants for bipolar depression. Many experts believe that antidepressants are not only less effective but also dangerously destabilizing in bipolar disorder patients, short-term, and lacking in prophylactic effectiveness long-term (Pacchiarotti et al., 2013). However, emerging evidence suggests that many depressed bipolar disorder patients do respond to antidepressant treatment, especially if cases involving current agitation are avoided (Tondo et al., 2013; Vázquez et al., 2013). Agents with mood-stabilizing effects, such as lithium and lamotrigine, are probably ineffective in acute bipolar depression, and only partially protective long-term (Vázquez et al., 2013). Lithium and anticonvulsants, though not clearly more effective than antidepressants for bipolar depression, are associated with low risks of 'mood-switching' from bipolar depression to mania. Even during treatment with an antidepressant, the added risks may be quite small (5% or less), though risks of spontaneous switching into mania without exposure to mood-elevating drugs are quite high (10–15%) (Tondo et al., 2010). Such risks, including risk of inducing rapid-cycling lead many clinicians to avoid use of antidepressants to treat depressed bipolar disorder patients, especially those diagnosed with type I disorder or who have

switched into mania previously when exposed to a mood-elevating agent (Undurraga et al., 2012). It is widely assumed that combining antidepressants with antimanic agents or use of antidepressants other than tricyclics or venlafaxine given in low doses and briefly may limit risks of mood-switching, although these points remain unproved (Tondo et al., 2010; Pacchiarotti et al., 2013).

Despite current controversies about the value and safety of antidepressants for bipolar depression, they are among the most commonly prescribed drugs for bipolar disorder patients – probably reflecting strong efforts by patients to avoid this painful condition (Baldessarini et al., 2008). Modern antidepressants, including serotonin-reuptake inhibitors and bupropion are most commonly employed because of their far greater safety than tricyclic antidepressants on acute overdose with suicidal intent, as well as lesser risks of inducing mood-switches into mania. Electroconvulsive therapy may be a temporary alternative in severe cases and can be beneficial in both manic and depressive episodes. Available evidence supports the value of the modern antipsychotic agents quetiapine, lurasidone, and olanzapine (if combined with the antidepressant fluoxetine) for treatment of acute bipolar depression, and additional experimental agents are emerging (Pacchiarotti et al., 2013; Vázquez et al., 2013).

Since a key feature of bipolar disorder is recurrence of acute episodes, a principal aim of treatment is prevention of recurrences of both depression and mania. Lithium (as the carbonate or citrate) is the best established option for this purpose (Cade, 1949; Baldessarini, 2013). Lithium treatment also is the only psychiatric treatment with compelling evidence of reduction of suicidal behavior (Tondo et al., 2003b; Baldessarini et al., 2006; Yerevanian and Choi, 2013). Lithium treatment requires more than 1 week at safe doses to gain appreciable antimanic effects, and alone is usually insufficient as well as less rapidly effective than most antipsychotics and some anticonvulsants such as valproate and carbamazepine. Lithium has complex effects on interneuronal signal transduction and on other molecular regulatory events in nerve cells (Malhi et al., 2013). The principal value of lithium is for preventing, delaying, or lessening severity and duration of manic recurrences in a majority of bipolar disorder patients over many years, with lesser effects against recurrences of bipolar depression (Bauer et al., 2006; Baldessarini, 2013).

Alternative treatments compete strongly for the formerly leading position of lithium for the purpose of long-term mood-stabilization. These include the anticonvulsant lamotrigine (which is moderately effective against recurrences of bipolar depression, with little short- or long-term antimanic efficacy). Antimanic–anticonvulsants including valproate and carbamazepine (alone or with lithium) are also employed empirically in long-term treatment, even though such 'off-label' applications lack regulatory approval. In addition, modern antipsychotic agents, including quetiapine, aripiprazole, and olanzapine may have long-term benefits in bipolar disorder, although with the possible exception of quetiapine, best demonstrated for limiting recurrences of mania, whereas lurasidone appears to be selectively effective in acute bipolar depression.

Despite the growing number of treatment options, many bipolar disorder patients continue to experience some illness each year, including minor hypomania and dysthymia as well as major illness recurrences. In several long-term observational studies, including from first-episodes, bipolar disorder patients were ill approximately half of the time despite use of modern treatments aimed at mood-stabilization; moreover fully three-quarters of this unresolved morbidity was accounted for by depressive illness (Baldessarini et al., 2010a). Long-term effectiveness of lithium, anticonvulsants or antipsychotic drugs in bipolar disorder patients appears to be independent of past history, including the delay from illness-onset to the start of sustained treatment and the number of previous recurrences (Baldessarini et al., 2007).

Owing to its limited safety margin, daily minimum serum concentrations of lithium are monitored regularly to guide dosing, and are safe and effective in the range of 0.6 to 1.0 $mEq\,l^{-1}$, although levels of 0.60–0.75 $mEq\,l^{-1}$ are adequate for most patients and better tolerated than higher levels. Increased tremor, fatigue, dizziness, hyperreflexia, confusion, lethargy or convulsions indicate intoxication with lithium. Other adverse effects include tremor, polyuria, diarrhea, nausea, weight-gain, hypothyroidism, and occasional impairment of renal function with uremia. Lithium is also associated with Ebstein's teratogenic malformation of the cardiac tricuspid valve with septal defects (prevalence less than 1/1000), although such outcomes are uncommon, usually can be diagnosed prenatally, and are often repairable surgically (Giles and Bannigan, 2006). A serious problem of long-term treatment with lithium and most other psychotropic drugs is that abrupt or rapid discontinuation is often followed by temporarily greatly increased risk for recurrences of mania, depression, or suicidal behavior, even after years of effective treatment (Baldessarini et al., 1999; Bauer et al., 2006; Baldessarini, 2013).

Most antipsychotic drugs are rapidly effective in acute mania, and most modern antipsychotics have relatively low risks of many of the neurological adverse effects of first-generation neuroleptic drugs, but some (notably olanzapine and quetiapine) are very likely to induce weight-gain, with risks of 'metabolic syndrome' that can include type 2 diabetes mellitus, dyslipidemia, and hypertension – all major risk factors for potentially life-threatening cardiovascular disease (Centorrino et al., 2012; Czepielewski et al., 2013).

Adverse effects of anticonvulsants include headache, rashes and life-threatening Stevens-Johnson dermal and mucosal necrosis, especially with lamotrigine, as well as increasing the metabolic clearance of other drugs, especially by carbamazepine (Baldessarini, 2013). Some anticonvulsants including valproate and carbamazepine also have very high risks of severe teratogenic effects during pregnancy, including crippling and irreversible spina bifida, which can arise within the initial weeks of pregnancy; lamotrigine has been associated with mid-line facial clefts (Baldessarini 2013; Dols et al., 2013b). These risks require special care in the use of such drugs by women of child-bearing age. Concern about these risks is a common basis for discontinuing mood-stabilizing treatments, often with severe, adverse psychiatric consequences to pregnant women (Viguera et al., 2007; Baldessarini, 2013).

Disclosures

Neither author nor any immediate family member has relationships with commercial organizations that might or might appear to represent potential conflicts of interest in the material presented here.

Acknowledgments

Supported, in part, by an award from the Aretæus Foundation of Rome and the Lucio Bini Private Donors Research Fund (to LT); a grant from the Bruce J. Anderson Foundation and by the McLean Hospital Private Donors Research Fund (to RJB).

See also: Borderline (Patient) Personality; Cognitive Behavioral Therapy; Depression; Mania; Mental Illness: Family and Patient Organizations in Europe.

Bibliography

American Psychiatric Association (APA), 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5), fifth ed. American Psychiatric Association, Arlington, VA.

Angst, J., Cui, L., Swendsen, J., Rothen, S., Cravchik, A., Kessler, R.C., Merikangas, K.R., 2010. Major depressive disorder with subthreshold bipolarity in the National Comorbidity Survey Replication. American Journal of Psychiatry 167, 1194–1201.

Baldessarini, R.J., 2013. Chemotherapy in Psychiatry, third ed. Springer Press, New York.

Baldessarini, R.J., Faedda, G.L., Offidani, E., Vázquez, G.H., Marangoni, C., Serra, G., Tondo, L., 2013. Rate of new-onset mania or hypomania in patients diagnosed with unipolar major depression. Journal of Affective Disorders 148, 129–135.

Baldessarini, R.J., Henk, H.J., Sklar, A.R., Chang, J., Leahy, L.F., 2008. Psychotropic medications for bipolar disorder patients in the United States: polytherapy and adherence. Psychiatric Services 59, 1175–1183.

Baldessarini, R.J., Pérez, J., Salvatore, P., Trede, K., Maggini, C., 2014a. History of bipolar manic-depressive disorder. In: Yildiz, A., Nemeroff, C., Ruiz, P. (Eds.), Bipolar Disorder: Millennium Update. Oxford University Press, New York.

Baldessarini, R.J., Salvatore, P., Imaz-Etxeberria, H., Khalsa, H.-M.K., Gonzalés-Pinto, A., Tohen, M., 2012a. Episode cycles with increasing recurrences in first-episode bipolar-I disorder patients. Journal of Affective Disorders 136, 149–154.

Baldessarini, R.J., Salvatore, P., Khalsa, H.M., Gebre-Medhin, P., Imaz-Etxeberria, H., González-Pinto, A., Pérez, J., Cruz, N., Maggini, C., Tohen, M., 2010a. Morbidity in 303 first-episode bipolar I disorder patients. Bipolar Disorders 12, 264–270.

Baldessarini, R.J., Tondo, L., Baethge, C., Lepri, B., Bratti, I.M., 2007. Effects of treatment latency on response to maintenance treatment in manic-depressive disorders. Bipolar Disorders 9, 386–393.

Baldessarini, R.J., Tondo, L., Davis, P., Pompili, M., Goodwin, F.K., Hennen, J., 2006. Decreased risk of suicides and attempts during long-term lithium treatment: meta-analytic review. Bipolar Disorders 8, 625–639.

Baldessarini, R.J., Tondo, L., Vázquez, G.H., Bolzani, L., Khalsa, H.-M.K., Lai, M., Lepri, B., Lolich, M., Maffei, P.M., Salvatore, P., Tohen, M., 2012b. International study of functional and symptomatic outcome versus onset-age in 1437 bipolar-I disorder patients. World Psychiatry 11, 40–46.

Baldessarini, R.J., Tondo, L., Viguera, A.C., 1999. Effects of discontinuing lithium maintenance treatment. Bipolar Disorders 1, 17–24.

Baldessarini, R.J., Tondo, L., Visioli, C., 2014b. First-episode types in bipolar disorder: predictive associations with later illness. Acta Psychiatrica Scandinavica 129, 383–392.

Baldessarini, R.J., Undurraga, J., Vázquez, G.H., Tondo, L., Salvatore, P., Ha, K., Khalsa, H.M.K., Pepri, B., Ha, T.H., Chang, H.S., Tohen, M., Vieta, E., 2012c. Predominant recurrence polarity among 928 adult international bipolar-I disorder patients. Acta Psychiatrica Scandinavica 125, 293–302.

Baldessarini, R.J., Vieta, E., Calabrese, J.R., Tohen, M., Bowden, C.L., 2010b. Bipolar depression: overview and commentary. Harvard Review of Psychiatry 18, 143–157.

Bauer, M., Grof, P., Müller-Oerlinghausen, B. (Eds.), 2006. Lithium in Neuropsychiatry: The Comprehensive Guide. Taylor & Francis, London, pp. 465–481.

Bondy, B., 2011. Genetics in psychiatry: are the promises met? World Journal of Biological Psychiatry 12, 81–88.

Cade, J.F.J., 1949. Lithium salts in the treatment of psychotic excitement. Medical Journal of Australia 2, 349–352.

Centorrino, F., Masters, G., Baldessarini, R.J., Öngür, D., 2012. Metabolic syndrome in psychiatrically hospitalized patients treated with antipsychotics and other psychotropics. Human Psychopharmacology 27, 521–526.

Cousins, D.A., Grunze, H., 2012. Interpreting magnetic resonance imaging findings in bipolar disorder. CNS Neuroscience & Therapeutics 18, 201–207.

Craddock, N., Sklar, P., 2013. Genetics of bipolar disorder. Lancet 381, 1654–1662.

Crump, C., Sundquist, K., Winkleby, M.A., Sundquist, J., 2013. Comorbidities and mortality in bipolar disorder: Swedish national cohort study. JAMA Psychiatry 70, 931–939.

Czepielewski, L., Daruy Filho, L., Brietzke, E., Grassi-Oliveira, R., 2013. Bipolar disorder and metabolic syndrome: systematic review. Revista Brasileira de Psiquiatria 35, 88–93.

Dols, A., Kupka, R.W., van Lammeren, A., Beekman, A.T., Sajatovic, M., Stek, M.L., August 6, 2013a. Prevalence of late-life mania: review. Bipolar Disorders [Epub ahead of print].

Dols, A., Sienaert, P., van Gerven, H., Schouws, S., Stevens, A., Kupka, R., Stek, M.L., 2013b. Prevalence and management of side effects of lithium and anticonvulsants as mood stabilizers in bipolar disorder from a clinical perspective: review. International Clinical Psychopharmacology 28, 287–296.

Faedda, G.L., Tondo, L., Teicher, M.H., Baldessarini, R.J., Gelbard, H.A., Floris, G.F., 1993. Seasonal mood disorders: patterns of seasonal recurrence in major affective disorder. Archives of General Psychiatry 50, 17–23.

Farren, C.K., Hill, K.P., Weiss, R.D., 2012. Bipolar disorder and alcohol use disorder: review. Current Psychiatry Reports 14, 659–666.

Geller, B., DelBello, M.P. (Eds.), 2003. Bipolar Disorder in Childhood and Early Adolescence. Guilford Press, New York.

Geller, B., Tillman, R., Bolhofner, K., Zimerman, B., 2008. Child bipolar I disorder: prospective continuity with adult bipolar I disorder; characteristics of second and third episodes; predictors of 8-year outcome. Archives of General Psychiatry 65, 1125–1133.

Giles, J.J., Bannigan, J.G., 2006. Teratogenic and developmental effects of lithium. Current Pharmaceutical Design 12, 1531–1541.

Goodwin, F.K., Jamison, K.R., 2007. Manic-Depressive Illness, second ed. Oxford University Press, New York.

Harvey, P.D., Wingo, A.P., Burdick, K., Baldessarini, R.J., 2010. Cognition and disability in bipolar disorder: lessons from schizophrenia research. Bipolar Disorders 12, 364–375.

Koukopoulos, A., Reginaldi, D., Tondo, L., Visioli, C., Baldessarini, R.J., 2013. Course sequences in bipolar disorder: depressions preceding or following manias or hypomanias. Journal of Affective Disorders 151, 105–110.

Krauthammer, C., Klerman, G.L., 1978. Secondary mania: manic syndromes associated with antecedent physical illness or drugs. Archives of General Psychiatry 35, 1333–1339.

Kupferschmidt, D.A., Zakzanis, K.K., 2011. Toward a functional neuroanatomical signature of bipolar disorder: quantitative evidence from the neuroimaging literature. Psychiatric Research 193, 71–79.

Lolich, M., Vázquez, G.H., Alvarez, L.M., Tamayo, J.M., 2012. Psychosocial interventions in bipolar disorder: review. Actas Espanolas de Psiquiatria 40, 84–92.

Malhi, G.S., Tanious, M., Das, P., Coulston, C.M., Berk, M., 2013. Potential mechanisms of action of lithium in bipolar disorder. CNS Drugs 27, 135–153.

Manchia, M., Hajek, T., O'Donovan, C., Deiana, V., Chillotti, C., Ruzickova, M., Del Zompo, M., Alda, M., 2013. Genetic risk of suicidal behavior in bipolar spectrum disorder: analysis of 737 pedigrees. Bipolar Disorders 15, 496–506.

Merikangas, K.R., Lamers, F., 2012. The 'true' prevalence of bipolar II disorder. Current Opinion in Psychiatry 25, 19–23.

Novick, D.M., Swartz, H.A., Frank, E., 2010. Suicide attempts in bipolar I and bipolar II disorder: review and meta-analysis. Bipolar Disorders 12, 1–9.

Oepen, G., Salvatore, P., Baldessarini, R.J., 2004. On the periodicity of manic-depressive insanity by Eliot Slater (1938): translation and commentary. Journal of Affective Disorders 78, 1–9.

Offidani, E., Fava, G.A., Tomba, E., Baldessarini, R.J., 2013. Excessive mood-elevation with juvenile antidepressant treatment in depressive versus anxiety disorders: systematic review. Psychotherapy and Psychosomatics 82, 132–141.

Ösby, U., Brandt, L., Correia, N., Ekbom, A., Sparén, P., 2001. Excess mortality in bipolar and unipolar disorder in Sweden. Archives of General Psychiatry 58, 844–850.

Pacchiarotti, I., Bond, D.J., Baldessarini, R.J., Task Force Members, September 13, 2013. International Society for Bipolar Disorders (ISBD) task-force report on antidepressant use in bipolar disorders. American Journal of Psychiatry [Epub ahead of print].

Pagel, T., Baldessarini, R.J., Franklin, J., Baethge, C., 2013. Characteristics of patients diagnosed with schizoaffective disorder compared with schizophrenia and bipolar disorder. Bipolar Disorders 15, 229–239.

Salvatore, P., Baldessarini, R.J., Khalsa, H.-.M.K., Amore, M., DiVittorio, C., Ferraro, G., Maggini, C., Tohen, M., 2013a. Predicting diagnostic change among first-episode patients diagnosed with DSM-IV major depression with psychotic features. Journal of Clinical Psychiatry 74, 723–731.

Salvatore, P., Baldessarini, R.J., Khalsa, H.-M.K., Vázquez, G.H., Pérez, J., Maggini, C., Tohen, M., 2014. McLean-Harvard International First Psychotic Episode Study: antecedents in 263 bipolar I disorder patients. Acta Psychiatrica Scandinavica 129, 275–285.

Salvatore, P., Baldessarini, R.J., Tohen, M., Khalsa, H.-M., Sanchez-Toledo, J.P., Zarate Jr., C.A., Vieta, E., Maggini, C., 2009. McLean-Harvard International First-Episode Project: two-year stability of DSM-IV diagnoses in 500 first-episode psychotic disorder patients. Journal of Clinical Psychiatry 70, 458–466.

Salvatore, P., Indic, P., Murray, G., Baldessarini, R.J., 2013b. Biological rhythms and mood disorders. Dialogues in Clinical Neuroscience 14, 369–379.

Swann, A.C., Lafer, B., Perugi, G., Frye, M.A., Bauer, M., Bahk, W.M., Scott, J., Ha, K., Suppes, T., 2013. Bipolar mixed states: an international society for bipolar disorders task force report of symptom structure, course of illness, and diagnosis. American Journal of Psychiatry 170, 31–42.

Tohen, M., Zarate Jr., C.A., Hennen, J., Kaur Khalsa, H.M., Strakowski, S.M., Gebre-Medhin, P., Salvatore, P., Baldessarini, R.J., 2003. The McLean-Harvard First-Episode Mania Study: prediction of recovery and first recurrence. American Journal of Psychiatry 160, 2099–2107.

Tondo, L., Baldessarini, R.J., Vázquez, G.H., Lepri, B., Visioli, C., 2013. Clinical responses to antidepressants among 1036 acutely depressed patients with bipolar or unipolar major affective disorders. Acta Psychiatrica Scandinavica 127, 355–364.

Tondo, L., Hennen, J., Baldessarini, R.J., 2003a. Meta-analysis of treatment responses of rapid-cycling and non-rapid-cycling bipolar disorder patients. Acta Psychiatrica Scandinavica 104, 4–14.

Tondo, L., Isacsson, G., Baldessarini, R.J., 2003b. Suicide in bipolar disorder: risk and prevention. CNS Drugs 17, 491–511.

Tondo, L., Lepri, B., Baldessarini, R.J., 2007. Risks of suicidal ideation, attempts and suicides among 2826 men and women with types I and II bipolar, and recurrent major depressive disorders. Acta Psychiatrica Scandinavica 116, 419–428.

Tondo, L., Vázquez, G.H., Baldessarini, R.J., 2010. Mania associated with antidepressant-treatment: comprehensive meta-analytic review. Acta Psychiatrica Scandinavica 121, 404–414.

Trede, K., Salvatore, P., Baethge, C., Gerhard, A., Maggini, C., Baldessarini, R.J., 2005. Manic-depressive illness: evolution in Kraepelin's Textbook, 1883–1926. Harvard Review of Psychiatry 13, 155–178.

Undurraga, J., Baldessarini, R.J., Valenti, M., Pacchiarotti, I., Tondo, L., Vázquez, G.H., Vieta, E., 2012. Bipolar depression: clinical correlates of receiving antidepressants. Journal of Affective Disorders 139, 89–93.

Vázquez, G.H., Baldessarini, R.J., Tondo, L., 2014. Co-morbidity of anxiety and bipolar disorders: clinical and therapeutic overview. Depression Anxiety 31, 196–206.

Vázquez, G.H., Tondo, L., Undurraga, J., Baldessarini, R.J., 2013. Overview of antidepressant treatment in bipolar depression: critical commentary. International Journal of Neuropsychopharmacology 16, 1673–1685.

Viguera, A.C., Whitfield, T., Baldessarini, R.J., Newport, D.J., Stowe, Z., Cohen, L.S., 2007. Recurrences of bipolar disorder in pregnancy: prospective study of mood-stabilizer discontinuation. American Journal of Psychiatry 164, 1817–1824.

Vöhringer, P.A., Barroilhet, S.A., Amerio, A., Reale, M.L., Alvear, K., Vergne, D., Ghaemi, S.N., 2013. Cognitive impairment in bipolar disorder and schizophrenia: systematic review. Frontiers in Psychiatry 4, 87–98.

World Health Organization (WHO), 2000. Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death. Tenth Revision (ICD-10). World Health Organization, Geneva.

Yang, H., 2011. Decoding the biology of bipolar disorder: update on recent findings in genetics, imaging, and immunology. FOCUS Psychiatry 9, 423–427.

Yerevanian, B.I., Choi, Y.M., 2013. Impact of psychotropic drugs on suicide and suicidal behaviors. Bipolar Disorders 15, 594–621.

Yildiz, A., Vieta, E., Leucht, S., Baldessarini, R.J., 2011. Efficacy of antimanic treatments: meta-analysis of randomized, controlled trials. Neuropsychopharmacology 36, 375–389.

Birdsong and Vocal Learning During Development

Sarah W Bottjer, University of Southern California, Los Angeles, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 2, pp. 1245–1252, © 2001, Elsevier Ltd.

Abstract

Songbirds are one of the few groups of organisms other than humans that learn sounds for vocal communication during development, and the neural substrate that controls vocal learning and behavior is highly localized, making these circuits amenable to experimental manipulation and analysis. The song system is a model one for studying the neural and hormonal bases of vocal learning, and for examining plasticity of structure and function in neural systems generally. Similarities between vocal learning in birds and humans suggest that many aspects of the learning process have evolved to meet demands imposed by vocal communication (Marler, 1976; Doupe and Kuhl, 1999). Therefore, an understanding of the mechanisms underlying vocal learning in songbirds may have immediate application to problems in human vocal learning. In this article, the author concentrates on what is known about neural and hormonal mechanisms of vocal learning in zebra finches (*Taeniopygia guttata*), a well-studied species in which juvenile males gradually learn to produce a specific vocal pattern during a sensitive period of development, and by adulthood are normally incapable of altering that pattern or learning new ones.

Songbirds are one of the few groups of organisms other than humans that learn sounds for vocal communication during development, and the neural substrate that controls vocal learning and behavior is highly localized, making these circuits amenable to experimental manipulation and analysis. The song system is a model one for studying the neural and hormonal bases of vocal learning, and for examining plasticity of structure and function in neural systems generally. Similarities between vocal learning in birds and humans suggest that many aspects of the learning process have evolved to meet demands imposed by vocal communication (Marler, 1976; Doupe and Kuhl, 1999). Therefore, an understanding of the mechanisms underlying vocal learning in songbirds may have immediate application to problems in human vocal learning. In this article, the author concentrates on what is known about neural and hormonal mechanisms of vocal learning in zebra finches (*Taeniopygia guttata*), a well-studied species in which juvenile males gradually learn to produce a specific vocal pattern during a sensitive period of development, and by adulthood are normally incapable of altering that pattern or learning new ones.

A Brain-Behavior System

Zebra finches learn the sounds used for vocal communication during a sensitive period of development. Specific auditory experience is necessary for vocal learning to proceed normally; juveniles must hear a 'tutor' song model (normally their father) from approximately 20–50 days of age in order to reproduce an accurate copy of that song later in development (e.g., Böhner, 1990). Zebra finches begin to produce their first song-related vocalizations around 30–35 days, and gradually refine their utterances until they achieve a close match to the external tutor sounds they have heard earlier (**Figure 1**). During this period of auditory-motor integration, birds must hear their own

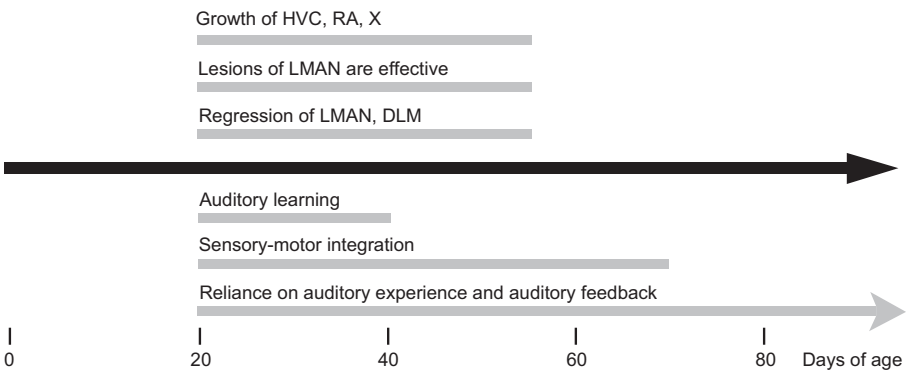

Figure 1 A time line of zebra finch vocal development showing approximate timing of some aspects of song learning. Birds fledge from the nest around 20 days of age and start singing soon thereafter. They are dependent on their parents to feed them up until 35–40 days of age. The exact timing of auditory learning from an external model (tutor song) and reliance on auditory feedback of self-produced sounds are not completely known. Juvenile birds use auditory feedback to refine subsong until it matches the template of the tutor song. Once song learning is complete, adult birds also require auditory feedback to maintain production of stable, stereotyped song. Large-scale growth and regression of brain regions involved with vocal learning and behavior occurs during the sensitive period (see text).

vocalizations; auditory feedback is necessary for juvenile males to learn to adjust the motor patterns giving rise to vocal output so that the latter gradually comes to match the tutor song (Price, 1979). By 80–90 days, zebra finches are sexually mature and produce a stereotyped song pattern that is maintained without changes throughout adult life. Once a stereotyped song is produced, adults also rely on auditory feedback to maintain that stable song pattern (Nordeen and Nordeen, 1992). Surprisingly, although it has been known for many years that there is a sensitive period for hearing external auditory sounds during early stages of vocal learning, it is not known whether there is also a sensitive period for hearing self-produced sounds (i.e., for experiencing auditory feedback).

Song behavior is highly sexually dimorphic: normal adult females never sing, even if treated with testosterone (T). However, females treated with sex hormones soon after hatching produce male-typical songs as adults (Pohl-Apel, 1985; Simpson and Vicario, 1991). The behavioral dimorphism reflects a profound neural sex difference: the sizes of cortical song-control regions are several times larger in males than in females (Nottebohm and Arnold, 1976). The neural substrate for vocal learning and behavior is highly localized in males, comprising several interconnected brain regions (**Figure 2**) that are highly amenable to investigation of structure–function relationships. For example, high vocal center (HVC) and robust nucleus of the archistriatum (RA) are part of a direct efferent system linking the cortex with descending motor circuitry that activates vocal musculature, and lesions of this pathway disrupt stereotyped song behavior in adult males

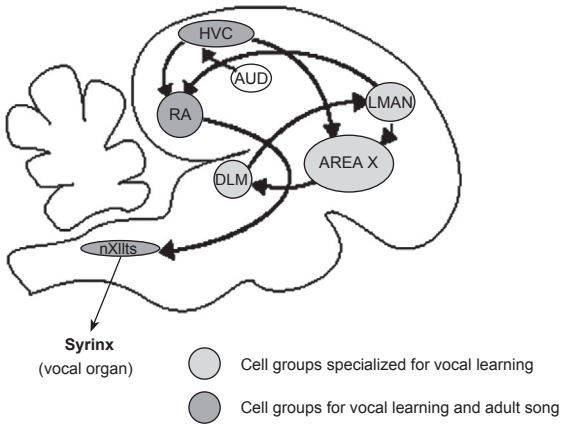

Figure 2 A side view of adult male zebra finch brain showing a highly simplified schematic of the major circuits controlling vocal learning and behavior. HVC contains two separate populations of projection neurons, one sends axons to RA and the other to Area X. The HVC → RA pathway regulates production of already-learned songs in adult birds, and is assumed to be involved with later stages of song learning in juveniles. The HVC → X → DLM → LMAN → RA pathway is necessary for normal song production during early stages of vocal learning, but is clearly not on the main motor pathway for vocal behavior in older juveniles and adults (see text). RA projects to the motor neurons (nXIIts) that activate the vocal muscles. LMAN, lateral magnocellular nucleus of the anterior neostriatum; X, Area X of the avian basal ganglia; HVC, high vocal center; AUD, auditory cortex; RA, robust nucleus of the archistriatum; DLM, medial dorsolateral nucleus of the thalamus; nXIIts, tracheosyringeal portion of the hypoglossal nucleus.

(Nottebohm et al., 1976; Simpson and Vicario, 1990; Wild, 1997). The activity of HVC neurons in awake, singing birds correlates precisely with production of specific song syllables, indicating that the HVC → RA pathway is part of the online circuitry for producing learned vocal patterns in adult birds (McCasland and Konishi, 1981). Furthermore, the activity of RA neurons in singing birds correlates with subsyllabic elements of vocal production, whereas HVC neurons apparently encode larger chunks of information, indicating a hierarchical organization in the neural control of vocal behavior (Yu and Margoliash, 1996; Vu et al., 1994).

There is also a multirelay route from HVC to RA: a distinct population of neurons in HVC projects to Area X in the basal ganglia, which relays through the thalamic nucleus medial dorsolateral nucleus of the thalamus (DLM) to the cortical region lateral magnocellular nucleus of the anterior neostriatum (LMAN), which projects onto the same RA neurons that receive inputs from HVC. RA-projecting neurons in LMAN send an axon collateral to Area X, which creates a forebrain loop connecting X, DLM, and LMAN (Vates and Nottebohm, 1995). Lesions of the X → DLM → LMAN pathway have no effect on song production in adult birds, but profoundly disrupt song behavior in juveniles during early stages of vocal learning (Bottjer et al., 1984; Sohrabji et al., 1990; Scharff and Nottebohm, 1991). There is a dramatic decrease in the effectiveness of LMAN lesions in juvenile males around 55–60 days of age, which seems to correlate with the development of song as a motor pattern in the sense that the temporal sequence of notes becomes fairly regular at this point. These results suggest that LMAN circuitry is necessary for learning during early stages of vocal development, but may play no role in stereotyped vocal production. The decrease in the effectiveness of lesions within this pathway suggests a change in its function, whatever that may be. One possibility is that the function subserved by this circuit is either no longer important for vocal learning and production, or is taken over by other song-control circuits, such as the HVC → RA pathway. However, an alternative interpretation of the developmental decline in the ability of LMAN lesions to disrupt behavior is that the function of this pathway is conserved, but somehow masked or suppressed at later stages of vocal learning (see below). Either way, this pathway may be similar to mammalian basal ganglia pathways in terms of its involvement in motor aspects of learning, planning, and coordination of movement and motivation (cf Alexander, 1994; Graybiel et al., 1994; Bottjer and Johnson, 1997).

Steroid Hormones and Vocal Development

Much evidence suggests that sex hormones are involved in regulating one or more aspects of behavior during the sensitive period for song learning (Bottjer and Johnson, 1997). Blocking the action of androgens in juvenile zebra finches prevents the development of stereotyped vocal behavior (but has no effect in adults). That is, syllable order fails to become stereotyped in juvenile zebra finches that are castrated and treated with antiandrogens, such that the overall structure of song is unstable and abnormal. Interestingly, subsequent exposure to physiological levels of T in adulthood induces stereotyped song production over a period of weeks, suggesting that the final

stages of auditory-motor learning can be carried out in delayed fashion upon exposure to high levels of T. The enhanced behavioral and neural plasticity associated with the sensitive period for vocal learning appears to depend, at least in part, on the relatively low levels of circulating hormones that normally occur during early stages of juvenile development. In accord with the idea that reduced levels of plasma testosterone are actually necessary for normal song development, chronic exposure to normal adult levels of T during juvenile development severely impairs vocal learning in zebra finches. Juvenile males receiving systemic T exposure at the onset of vocal learning (20–40 days) develop fewer song syllables than controls, suggesting that the ability to memorize tutor syllables is impaired by exposure to adult levels of T during this interval. These data indicate that the proper timing and amount of exposure to hormones are important factors in the normal development of vocal behavior. Because many song-control neurons contain hormone receptors (particularly in HVC and LMAN), it is likely that T may exert direct effects on song-control circuitry to alter functional properties of neurons during vocal learning.

One possible speculation suggested by this pattern of results is that blocking the action of androgens extends a sensitive period for vocal learning, whereas normal exposure to rising titers of androgens closes one or more windows of opportunity for learning. This pattern of results has strong parallels to studies of human language acquisition, which have shown that the capacity for language learning decreases sharply following puberty (e.g., Johnson and Newport, 1989). It is also known that the sensitive period for song learning can be extended by rearing zebra finches in isolation. Although birds that remain isolated produce highly abnormal vocalizations, birds that are provided with tutors in adulthood can learn some song syllables long after the normal sensitive period has ended (Eales, 1987; Morrison and Nottebohm, 1993). Interestingly, juvenile birds that have been socially and acoustically isolated have abnormally low levels of T, although hormonal titers recover by the time birds reach late stages of song learning (Livingston et al., 2000). Thus the extended capacity for learning produced by isolation may be due to a prolonged period of low hormonal titers, although lack of experience may also be a contributing factor. It is possible that any conditions that prevent learning tend to keep the sensitive period open. This idea suggests the hypothesis that the loss of plasticity associated with the closing of the sensitive period is due to learning; if vocal learning is prevented then the sensitive period remains open.

Developmental Changes in Song-Control Nuclei: Growth and Regression

The neural substrate for song control undergoes large-scale morphological changes during the period of vocal learning in juvenile male zebra finches (Nordeen and Nordeen, 1997; Bottjer and Arnold, 1997). Substantial (e.g., threefold) changes in the overall size of song-control nuclei can include changes in neuron number and/or size and complexity of axon terminals and dendritic arbors. For example, HVC and RA grow substantially in males during song learning, more than doubling in volume, whereas LMAN shows substantial regression (**Figure 1**). Regression of LMAN correlates with the loss in effectiveness of LMAN lesions, whereas the growth of HVC and RA are consistent with the idea that the HVC → RA pathway assumes greater control of vocal behavior as song learning progresses. The profound growth and regression of song-control brain regions suggest highly dynamic roles for vocal circuitry in one or more aspects of song learning, and must be accompanied by substantial remodeling of synaptic connectivity. An interesting feature of these changes is that their occurrence during development is delayed relative to circuits that are not related to song control. Thus, the major neural scaffold controlling song behavior is created as song is being learned.

The gross changes in volume of brain regions observed during song learning suggested the possibility that experiential factors might play a role in inducing such morphological changes in the substrate for song, such that these large-scale changes actually reflect vocal learning. However, the volume and neuronal number of telencephalic song-control nuclei is the same in hearing and deaf birds, showing that auditory experience has no influence on regulation of neuron number, and that normal patterns of growth and regression occur even in birds that do not engage in normal song learning (Burek et al., 1991). This result indicates that gross morphological changes in song-control circuits are in fact *not* a product of the learning process, but rather may be prerequisite to learned changes in behavior. In accord with this idea, Brenowitz et al. (1995) demonstrated that tutoring two groups of marsh wrens with either many syllables or few syllables led to the development of large versus small learned song repertoires, respectively, although the size (and neuronal number) of HVC and RA in the two groups did not vary. However, within the group tutored with many syllables, only birds with a relatively large HVC learned large repertoires. Thus, a large HVC may be permissive for a larger vocal repertoire, but is clearly not a consequence of such. It seems likely that the basic neural scaffold underlying vocal learning is dependent on innate factors such as endogenous levels of hormones and cytokines (Johnson et al., 1997; Rasika et al., 1999). However, the idea that large-scale changes in song-control nuclei are a necessary prerequisite for vocal learning has not been tested directly (i.e., the occurrence of these changes during the sensitive period for vocal learning is correlational; cf Bottjer and Arnold, 1997).

Neural Signatures of Learning and Development

If large-scale patterns of growth and regression in the neural substrate for vocal learning are not a reflection of learning, then what changes in the brain signify learning and the close of the sensitive period? Several studies have described normative developmental changes that occur during the sensitive period, and may underlie specific aspects of the learning process. For example, the axonal projections of the DLM → LMAN → RA pathway show exuberant growth during early stages of song learning followed by pruning (Nixdorf-Bergweiler et al., 1995; Iyengar et al., 1999). In addition, NMDA (*N*-methyl-D-aspartate) receptors decrease in density within LMAN during song learning, and the duration of NMDA-R-mediated synaptic currents in individual LMAN neurons becomes significantly

shorter (Aamodt et al., 1995; Livingston and Mooney, 1997). These changes indicate that NMDA-R carry a greater proportion of the synaptic current in juveniles than adults. NMDA-R figure prominently in activity-dependent adjustment of synaptic connections because they require both a ligand (glutamate) and depolarization of the postsynaptic cell, they serve as molecular detectors of coincident pre- and postsynaptic activation and therefore provide a mechanism for Hebbian learning (e.g., Cline, 1991). Thus, it is possible that NMDA-R in LMAN neurons could detect correlated activity of auditory and motor aspects of song to reinforce synapses that correspond to the acoustic structure of tutor song. In general, highly precise patterns of synaptic connectivity as well as the behaviors they subserve are established during sensitive periods guided by activity-dependent mechanisms based on sensory experience (Knudsen, 1999), so it is therefore seems likely that experience-dependent synaptic rearrangements in song-control circuits, such as refinement of the DLM → LMAN → RA pathway, may underlie the emergence of specific vocal patterns.

If developmental changes in synaptic rearrangements and NMDA-R currents reflect learning, then one prediction is that they should be experience-dependent. This is at least partially true; isolation of juvenile birds to prolong the sensitive period (see above) does delay the shortening of NMDA-R currents, as well as loss of the modulatory subunit that gates the corresponding slow kinetics normally seen in juveniles (White et al., 1999; Singh et al., 2000). In addition, the number of dendritic spines on LMAN neurons decreases during song learning in normal birds, but remains high in age-matched birds that have been reared in isolation (Wallhaüsser-Franke et al., 1995), and the exuberant growth of the LMAN → RA connection is prolonged in birds that have been deafened or raised in white noise (Iyengar et al., 1999). However, isolation does not prevent these normative changes permanently. For example, although NMDA-R currents are longer in isolate-reared birds compared to normal birds at early stages of song learning, they recover to normal levels at later stages (Livingston et al., 2000). Thus, although LMAN neurons of isolate-reared birds seem to remain in a more juvenile state initially, they eventually achieve normal levels despite continued isolation. Furthermore, because isolated birds can learn some new syllables as adults, this pattern indicates that the enhanced NMDA-R currents seen in normal juveniles during the sensitive period are not necessary for at least some types of learning to occur. However, it is possible that the greater contribution of NMDA-R to synaptic currents in young birds could subserve some early aspect of learning that is necessary for later song development. For example, longer NMDA-R currents might enable young birds to learn an association between specific articulatory gestures and the vocal sounds they produce.

One way to assess this latter idea would be to prevent the longer NMDA-R currents that normally occur at the onset of the sensitive period. Exposing juvenile males to adult levels of testosterone starting at 20 days, severely disrupts vocal learning and causes premature shortening of NMDA-R currents as well as an accelerated loss of the slow modulatory subunit of the NMDA-R (White et al., 1999; Singh et al., 2000). This pattern is consistent with the idea that enhanced NMDA-R synaptic currents are necessary for some, perhaps unknown, aspects of vocal learning. The delay seen in several neural parameters caused by experiential and hormonal manipulations may provide neural correlates that contribute to the extended sensitive period for song learning in isolated zebra finches. Furthermore, the relative decrease in the contribution of NMDA-R currents could be responsible for closing one or more aspects of the sensitive period.

One particularly interesting neural signature of vocal learning is the development of auditory tuning in song-control neurons (Doupe and Solis, 1997; Margoliash, 1997). For example, LMAN neurons in adult birds that project to RA are selectively responsive to playback of each bird's own song, and this selectivity emerges at about 60 days of age, around the time when birds are beginning to produce a relatively stable song pattern and LMAN lesions are losing the ability to disrupt vocal behavior (Solis and Doupe, 1999; Rosen and Mooney, 2000). This correlation suggests that the development of auditory tuning coincides with a decreased role for LMAN neurons in vocal development, and is consistent with an idea raised above – namely that learning itself acts to curtail the sensitive period. Perhaps auditory selectivity of LMAN neurons reflects the fact that the bird has achieved a match to the tutor song (or developed their own individual template), such that LMAN no longer contributes actively to regulating the song pattern. Indeed, Solis and Doupe (2000) found that birds that were prevented from developing a good match to the tutor song (via damage to the vocal motor nerve) showed lower auditory selectivity in Area X than did birds that were able to mimic the tutor song. Thus, it seems as if the development of selectivity to bird's own song is dependent on learning, and perhaps a reflection of it. As indicated above, refinements of axonal connectivity such as that seen in the DLM → LMAN → RA pathway may represent an important neural signature of learning; remodeling of synaptic connections in this pathway may be necessary for the increased auditory selectivity of LMAN neurons.

Recent Findings

As indicated above, adult male zebra finches normally are incapable of altering their previously learned, stereotyped vocal pattern. This behavioral pattern correlates well with neural data showing that lesions of the X → DLM → LMAN pathway are effective only during a restricted period of vocal learning, and not in adult birds. These results encouraged hypotheses of vocal learning that stressed the idea that 'age-limited learners' such as zebra finches develop a fixed central motor program, and that the basal ganglia-thalamocortical circuit from Area X to DLM to LMAN is no longer able to regulate behavioral and neural plasticity in adult birds. However, traditional hypotheses have been challenged by recent data. (a) Exposing adult birds to delayed auditory feedback causes gradual deterioration of stable song, but original song patterns may be recovered when normal auditory feedback is restored (Leonardo and Konishi, 1999). (b) Although lesions of Area X and LMAN do not impair adult song production, the act of producing 'undirected' song (i.e., not directed to a female) induces strong expression of the immediate early gene *zenk* in X and LMAN of adult males (Jarvis et al., 1998), and multiunit activity recorded

in LMAN and Area X is consistently higher during production of undirected song (Hessler and Doupe, 1999). (c) Although lesions of LMAN in normal adults do not cause behavioral disruption, LMAN lesions in adult birds that are not producing a normal stereotyped song pattern do produce changes in song behavior, and prevent delayed learning from a tutor in isolate-reared birds (Morrison and Nottebohm, 1993). (d) Williams and Mehta (1999) demonstrated that vocal plasticity induced by injury to the vocal motor nerve (nXIIts) in adult birds is blocked by lesions of LMAN, and similarly Brainard and Doupe (2000) showed that LMAN lesions in adult deafened birds prevent gradual deterioration (i.e., the song remains stereotyped despite the absence of auditory feedback, suggesting that LMAN may monitor auditory feedback and respond if an error signal is detected; if LMAN is gone, the error signal produced by altered auditory feedback would not be detected in deafened birds).

In aggregate, this pattern of findings suggests that the X → DLM → LMAN pathway is not irrevocably 'shut down' in adults after vocal learning is complete, despite the fact that lesions of this pathway are ineffective in normal adults. Furthermore, one or more functions of this circuit may be 'reinvoked' by certain types of changes in learned song behavior. In addition, these results encourage the idea that adult zebra finches might be capable of new learning following the sensitive period, if appropriate conditions can be found (e.g., if the bird 'unlearns' his initial song pattern; Zevin et al., 2000). Loss of a stable vocal pattern may be a prerequisite to (re)inducing a greater degree of neural and behavioral plasticity in adulthood, perhaps by 'unmasking' an apparently latent function of LMAN circuitry that may be conserved in adult birds.

Conclusions and Questions

Vocal learning, like other developmentally acquired skills, is shaped by the interplay of innate factors and experiential influences during a sensitive period. The substantial growth and regression of song-control nuclei that occur during the period of vocal learning are not dependent on experience, and are likely regulated by innate factors that specify the amount of brain space allocated to vocal learning, thereby influencing how much can be learned (Nottebohm et al., 1981). Experience-independent specification of the basic neural scaffolding for vocal learning may also provide specific innate constraints on vocal learning, such as predispositions to hear or produce certain vocal sounds (Marler and Pickert, 1984; Marler and Sherman, 1985; cf Seidenberg, 1997). In addition, the fact that large-scale patterns of growth and regression occur as learned vocal behavior is being acquired may set temporal constraints on learning. Perhaps this is why humans, like songbirds, are much more adept at vocal learning as juveniles than as adults. If large-scale neural changes are necessary concomitants of vocal learning, then we must ask how experiential factors influence the fine scale of song-control circuitry during such changes. Is it possible that certain types of experience-dependent mechanisms, such as refinement of specific synaptic patterns, can operate only in the context of the overall growth and regression that occurs during song learning? Answers to such questions will begin to provide a better mechanistic understanding of sensitive periods in general.

See also: Associative Learning in the Cerebellum; Learning and Memory, Neural Basis of; Neural Plasticity; Speech Production, Neural Basis of.

Bibliography

Aamodt, S.M., Nordeen, E.J., Nordeen, K.W., 1995. Early isolation from conspecific song does not affect the normal developmental decline of N-methyl-D-aspartate receptor binding in an avian song nucleus. Journal of Neurobiology 23, 76–84.

Alexander, G.E., 1994. Basal ganglia-thalamocortical circuits: their role in control of movements. Journal of Clinical Neurophysiology 11, 420–431.

Böhner, J., 1990. Early acquisition of song in the zebra finch. Animal Behavior 39, 369–374.

Bottjer, S.W., Arnold, A.P., 1997. Developmental plasticity in neural circuits for a learned behavior. Annual Review of Neuroscience 20, 459–481.

Bottjer, S.W., Johnson, F., 1997. Circuits, hormones, and learning: vocal behavior in songbirds. Journal of Neurobiology 33, 602–618.

Bottjer, S.W., Miesner, E.A., Arnold, A.P., 1984. Forebrain lesions disrupt development but not maintenance of song in passerine birds. Science 224, 901–903.

Brainard, M.S., Doupe, A.J., 2000. Interruption of a basal ganglia-forebrain circuit prevents plasticity of learned vocalizations. Nature 404, 762–766.

Brenowitz, E.A., Lent, K., Kroodsma, D.E., 1995. Brain space for learned song in birds develops independently of song learning. Journal of Neuroscience 15, 6281–6286.

Burek, M.J., Nordeen, K.W., Nordeen, E.J., 1991. Neuron loss and addition in developing zebra finch song nuclei are independent of auditory experience during song learning. Journal of Neurobiology 22, 215–223.

Cline, H., 1991. Activity-dependent plasticity fin the visual systems of frogs and fish. Trends in Neurosciences 14, 104–111.

Doupe, A.J., Kuhl, P.K., 1999. Birdsong and human speech: common themes and mechanisms. Annual Review of Neuroscience 22, 567–631.

Doupe, A.J., Solis, M.M., 1997. Song- and order-selective neurons develop in the songbird anterior forebrain during vocal learning. Journal of Neurobiology 33, 694–709.

Eales, L.A., 1987. Song learning in female-raised zebra finches: another look at the sensitive phase. Animal Behavior 35, 1356–1365.

Graybiel, A.M., Aosaki, T., Flaherty, A.W., Kimura, M., 1994. The basal ganglia and adaptive motor control. Science 265, 1826–1831.

Hessler, N.A., Doupe, A.J., 1999. Social context modulates singing-related neural activity in the songbird forebrain. Nature Neuroscience 2, 209–211.

Iyengar, S., Viswanathan, S.S., Bottjer, S.W., 1999. Development of topography within song control circuitry of zebra finches during the sensitive period for song learning. Journal of Neuroscience 19, 6037–6057.

Jarvis, E.D., Scharff, C., Grossman, M.R., Ramos, J.A., Nottebohm, F., 1998. For whom the bird sings: context-dependent gene expression. Neuron 21, 775–788.

Johnson, F., Hohmann, S.E., DiStefano, P.S., Bottjer, S.W., 1997. Neurotrophins suppress apoptosis induced by deafferentation of an avian motor-cortical region. Journal of Neuroscience 17, 2101–2111.

Johnson, J.S., Newport, E.L., 1989. Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. Cognitive Psychology 21, 60–99.

Knudsen, E., 1999. Early experience and critical periods. In: Zigmond, M.J., Bloom, F.E., Landis, S.C., Roberts, J.L., Squire, L.R. (Eds.), Fundamental Neuroscience. Academic Press, San Diego, CA, pp. 637–654.

Leonardo, A., Konishi, M., 1999. Decrystallization of adult bird-song by perturbation of auditory feedback. Nature 399, 466–470.

Livingston, F.S., Mooney, R., 1997. Development of intrinsic and synaptic properties in a forebrain nucleus essential to avian song learning. Journal of Neuroscience 17, 8997–9009.

Livingston, F.S., White, S.A., Mooney, R., 2000. Slow NMDA-EPSCs at synapses critical for song development are not required for song learning in zebra finches. Nature Neuroscience 3, 482–488.

Margoliash, D., 1997. Functional organization of forebrain pathways for song production and perception. Journal of Neurobiology 33, 671–693.

Marler, P., 1976. Sensory templates in species-specific behavior. In: Fentress, J.C. (Ed.), Simpler Networks and Behavior. Sinauer, Sunderland, MA, pp. 314–329.

Marler, P., Pickert, R., 1984. Species-universal microstructure in the learned song of the swamp sparrow (*Melospiza georgiana*). Animal Behavior 32, 673–689.

Marler, P., Sherman, V., 1985. Innate differences in singing behaviour of sparrows reared in isolation from adult conspecific song. Animal Behavior 33, 57–71.

McCasland, J.S., Konishi, M., 1981. Interaction between auditory and motor activities in an avian song control nucleus. Proceedings of the National Academy of Sciences of the United States of America 78, 7815–7819.

Morrison, R.G., Nottebohm, F., 1993. Role of a telencephalic nucleus in the delayed song learning of socially isolated zebra finches. Journal of Neurobiology 24, 1045–1064.

Nixdorf-Bergweiler, B.E., Wallhausser-Franke, E., DeVoogd, T.J., 1995. Regressive development in neuronal structure during song learning in birds. Journal of Neurobiology 27, 204–215.

Nordeen, K.W., Nordeen, E.J., 1992. Auditory feedback is necessary for the maintenance of stereotyped song in adult zebra finches. Behavioral and Neural Biology 57, 58–66.

Nordeen, K.W., Nordeen, E.J., 1997. Anatomical and synaptic substrates for avian song learning. Journal of Neurobiology 33, 532–548.

Nottebohm, F., Arnold, A.P., 1976. Sexual dimorphism in vocal control areas of the songbird brain. Science 194, 211–213.

Nottebohm, F., Kasparian, S., Pandazis, C., 1981. Brain space for a learned task. Brain Research 213, 99–109.

Nottebohm, F., Stokes, T.M., Leonard, C.M., 1976. Central control of song in the canary, *Serinus canarius*. Journal of Comparative Neurology 165, 457–486.

Pohl-Apel, G., 1985. The correlation between the degree of brain masculinization and song quality in estradiol treated female zebra finches. Brain Research 336, 381–383.

Price, P.H., 1979. Developmental determinants of structure in zebra finch song. Journal of Comparative and Physiological Psychology 93, 260–277.

Rasika, S., Alvarez-Buylla, A., Nottebohm, F., 1999. BDNF mediates the effects of testosterone on the survival of new neurons in an adult brain. Neuron 22, 53–62.

Rosen, M.J., Mooney, R., 2000. Intrinsic and extrinsic contributions to auditory selectivity in a song nucleus critical for vocal plasticity. Journal of Neuroscience 20, 5437–5448.

Scharff, C., Nottebohm, F., 1991. A comparative study of the behavioral deficits following lesions of various parts of the zebra finch song system: implications for vocal learning. Journal of Neuroscience 11, 2896–2913.

Seidenberg, M.S., 1997. Language acquisition and use: learning and applying probabilistic constraints. Science 275, 1599–1603.

Simpson, H.B., Vicario, D.S., 1990. Brain pathways for learned and unlearned vocalizations differ in zebra finches. Journal of Neuroscience 10, 1541–1556.

Simpson, H.B., Vicario, D.S., 1991. Early estrogen treatment alone causes female zebra finches to produce learned, male-like vocalizations. Journal of Neurobiology 22, 755–776.

Singh, T.D., Basham, M.E., Nordeen, E.J., Nordeen, K.W., 2000. Early sensory and hormonal experience modulate age-related changes in NR2B mRNA within a forebrain region controlling avian vocal learning. Journal of Neurobiology 44, 82–94.

Sohrabji, F., Nordeen, E.J., Nordeen, K.W., 1990. Selective impairment of song learning following lesions of a fore-brain nucleus in the juvenile zebra finch. Behavioral and Neural Biology 53, 51–63.

Solis, M.M., Doupe, A.J., 1999. Contributions of tutor and bird's own song experience to neural selectivity in the songbird anterior forebrain. Journal of Neuroscience 19, 4559–4584.

Solis, M.M., Doupe, A.J., 2000. Compromised neural selectivity for song in birds with impaired sensorimotor learning. Neuron 25, 109–121.

Vates, G.E., Nottebohm, F., 1995. Feedback circuitry within a song-learning pathway. Proceedings of the National Academy of Sciences of the United States of America 92, 5139–5143.

Vu, E.T., Mazurek, M.E., Kuo, Y.C., 1994. Identification of a forebrain motor programming network for the learned song of zebra finches. Journal of Neuroscience 14, 6924–6934.

Wallhausser-Franke, E., Nixdorf-Bergweiler, B.E., DeVoogd, T.J., 1995. Song isolation is associated with maintaining high spine frequencies on zebra finch IMAN neurons. Neurobiology of Learning and Memory 64, 25–35.

White, S.A., Livingston, F.S., Mooney, R., 1999. Androgens modulate NMDA receptor-mediated EPSCs in the zebra finch song system. Journal of Neurophysiology 82, 2221–2234.

Wild, J.M., 1997. Neural pathways for the control of birdsong production. Journal of Neurobiology 33, 653–670.

Williams, H., Mehta, N., 1999. Changes in adult zebra finch song require a forebrain nucleus that is not necessary for song production. Journal of Neurobiology 39, 14–28.

Yu, A.C., Margoliash, D., 1996. Temporal hierarchical control of singing in birds. Science 273, 1871–1875.

Zevin, J., Seidenberg, M.S., Bottjer, S.W., 2000. Song plasticity in adult zebra finches exposed to white noise. Society for Neuroscience Abstracts 26, 723.

Birth, Anthropology of

Lucile Newman, Brown University, Providence, RI, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 2, pp. 1252–1255, © 2001, Elsevier Ltd.

Abtract

Conception from successful female ovulation and male insemination results in nine months' pregnancy and birth. The outcome of shorter gestation is premature birth of a low birthweight infant. Fertility regulation, or contraception, limits conception by male methods such as the condom or surgical vasectomy, and by females by use of the oral contraceptive pill, the female condom, or long-term breastfeeding of the previous infant. Prenatal care refers to learning how to maintain health and prepare for the birth of a healthy infant.

An anthropology of birth is based on the common human experiences of insemination, conception, pregnancy, and birth of an infant. Characteristic of the human species and central to the origins of culture are the immaturity of the newborn human and its long-term dependence on caretakers for survival. These experiences exist in every society, but details vary in different cultural environments.

Cultural Anthropological Perspectives

Anthropological perspectives on birth are concerned with the context, variety, and complexity of reproductive life, modes of caring for a newborn infant, and cultural factors including rules for behavior that are learned, shared, and transmitted within families and cultural groups. There is cross-cultural variation among societies during these universal human processes. Because birth and recruitment to society are important to all human groups, customs and rules surround fertility behavior in all societies.

The influence of culture appears in definitions, for example, of which individuals may become parents (maternal age or married status), or the beginning of life (the point of conception, pregnancy to a viable birthweight, or delivery of a live infant). Throughout history and across cultures, certain variations have occurred in the definition of when life begins – whether it is at birth, or from motion or other indication of consciousness of an infant, or at a formal tenth day ceremony. Naming an infant provides a universal identification of the beginning of social life. Membership in society is defined and protected by both rules and traditions of social behavior. For example, international differences in mate selection include monogamy, or single partners, in most of the Christian Western world, and polygamy, or multiple wives in religious groups such as Mormons in the United States, or in some Islamic countries.

Differences occur also in sexual behavior, timing of pregnancy, and reproduction of society over time. It is the basis of identity, including how many people will be members of each society. The concept of 'legitimacy' refers to the acceptability of infants born to married mothers in most societies. 'Illegitimacy' refers to birth to an unmarried mother. Despite the language of legality, this is not generally an issue of law, but of social or religious acceptability.

Additional international and intercultural differences are discussed in the following sections, including fertility and fertility regulation, prenatal care, a comparative view of birth rates in different countries (Table 1), infant and under 5 child mortality (Table 2), and finally, the impact of the modern international infectious disease of HIV (human immunodeficiency virus, Table 3).

Fertility

Fertility refers to the capability of reproduction in the human species from sexual relations of male and female to conception and pregnancy in the female. Sex and social continuity are of concern in all societies. This concern is a part of the ancient and universal desire to bring order to membership in society as well as the behavior of individual members. 'Infertility' of either partner indicates an incapacity to reproduce.

Pregnancy occurs as a result of the connection of male sperm to female ovum initiating cell division leading to conception and beginning development of an embryo, until around the third month of gestation when it begins to develop body structures recognizable as human, and enters the fetal

Table 1 Total population and birth rate

Nation	*Total population (millions)*	*Birth rate (per woman)*
Australia	18.6	1.8
Brazil	165.9	2.3
China	1238.6	1.9
France	58.8	1.8
India	979.7	3.2
Indonesia	203.7	2.7
Mexico	95.8	2.8
Nigeria	120.8	5.3
Russian Federation	146.9	1.2
South Africa	41.4	2.8
United Kingdom	59.1	1.7
United States	270.3	2.0

Source: The World Bank, 2000. 1998 World Development Indicators. Washington, DC.

Table 2 Infant and under 5 child mortality

Nation	*Infant mortality (per 1000 live births)*	*Under 5 child mortality (per 1000)*
Australia	5	6
Brazil	33	40
China	31	36
France	5	5
India	70	83
Indonesia	43	52
Mexico	30	35
Nigeria	76	119
Russian Federation	17	20
South Africa	51	83
United Kingdom	6	7
United States	7	7

Source: The World Bank, 2000. 1998 World Development Indicators. Washington, DC.

Table 3 International prevalence of HIV infection

Nation	*Percent of adults*	*Estimated number infected (in 000's)*
Australia	0.14	11
Brazil	0.63	580
China	0.06	400
France	0.37	110
India	0.82	4100
Indonesia	0.85	52
Mexico	0.35	180
Nigeria	4.12	2300
Russian Federation	0.05	40
South Africa	12.91	2900
United Kingdom	0.09	25
United States	0.76	820

stage. Cells differentiate and the fetus then grows and develops the human form of an infant.

Gestation is the period of intrauterine development. Human variation occurs in the length of gestation for live birth. The biologically expectable period of gestation is around nine months of pregnancy. However, there can be differences in the timing, location, and length of the birth process. In some societies infants are delivered by a midwife or nurse, or a birth specialist either at home, or in the institutional setting of a hospital or nursing home. In modern urban societies, delivery of an infant is usually conducted in a hospital by a nurse or a physician specializing in obstetrics. Even here, however, there are significant cultural differences. For example, in urban areas, the surgical procedure of caesarean section is used for an impeded delivery, or for selective timing of birth.

In developed countries medical care may include prenatal care with a medical specialist in obstetrics or a nurse practitioner in a professional medical environment. In many areas of the world, reproductive health and training for motherhood are carried out by a nurse midwife or an experienced female family member, although there is increasing professionalization of birth attendants internationally. Attention to the process of gestation is perceived as important to the quality of life of the potential human being in all societies.

Fertility Regulation

Fertility regulation refers to intentional methods of avoidance of pregnancy used by either male or female. There have been methods of fertility regulation at least as far back as the beginnings of recorded history, as families have sought to control the number and timing of births. Male methods of family planning can range from a barrier method such as the condom, or withdrawal, or the surgical procedure of vasectomy. Female methods include the oral contraceptive pill or the barrier method of the female condom. Extended breastfeeding of the previous infant has long been known to suppress ovulation and enable spacing between births.

Social concerns regarding birth and reproduction vary internationally, with age, ethnic identity, or religion. Cultural factors, such as assuring virginity of marriageable females in some societies, religious opposition to fertility regulation in others, or in a few societies, female genital surgery for removal of the clitoris to reduce female enjoyment of sexual relations, have an impact on quality of life as well as reproductive capacity.

Ending unwanted pregnancy is one of the most contentious issues of reproduction. It may be terminated by induced clinical abortion in those countries where it is legal and medically provided. The psycho-social effects of unwantedness of a newborn infant are difficult to measure, but outcomes appear consistently negative in terms of later parental disinterest, child abuse, or abandonment. However, while clinical abortion is legal in such countries as Sweden, Japan, or the United States, it is illegal in most Muslim and Catholic areas. In these countries, family planning is more likely to include abstinence during the expectable time of ovulation (sometimes difficult to determine), and birth rates may be comparatively higher.

Prenatal Care

Most developed societies have some form of preparation for birth. The importance of prenatal care is in testing for health of the fetus, and in teaching the pregnant woman how to maintain the best possible health for herself and her expected infant. This includes protective health care, measuring regular weight gain, providing training for the birth process, and social support.

An important component is also prevention of behaviors detrimental to infant development. For example, prenatal smoking, alcohol and drug use, or starvation to limit weight gain, can all have damaging effects on the developing fetus. These can range from decreased blood flow to the fetus from smoking to fetal alcohol syndrome (FAS), which can result in abnormal physical or mental development (Espy et al., 2000). As risks from these behaviors have become better known over the last several decades, preventing risks of learning impairment in the resulting child is an important component of effective prenatal care before birth (Newman, 1996).

Premature Birth

Not all pregnancies continue to a viable birth. Loss can occur at any time. Abortion refers to expulsion of a malformed or

incompletely developed embryo within the first three months of gestation. Later loss can result from fetal death. For example, if loss occurs in the first weeks of pregnancy, it is considered a 'miscarriage' of a non-viable fetus. Early delivery of a viable infant at a birth weight of 2500g (about 5.5 16) can happen as early as seven months gestation, and is termed 'low birth weight.' 'Very low birth weight' refers to the infant born around 1500 g.

In developed and industrialized countries, extensive efforts are undertaken to ensure survival of the lowest birth weight infants. A viable low birth weight infant in developed countries is kept at a neonatal intensive care nursery in an incubator for clear observation, to maintain body heat, and to provide oxygen, if necessary to breathing, as well as to enable constant surveillance by special care staff. The purpose of enclosure in all countries is to keep the premature infant safe from infection. In developing countries without intensive care technology, such babies are seen as likely to suffer mental and physical limitations, and may not be given special care so that family and community resources can be focused on those infants perceived as more viable.

Birth Rate

Birth rates differ from country to country based on religion, cultural expectation, issues such as female participation in the workforce, or religious and political views of fertility limitation through birth control. It can also be influenced by political programs such as China's 'one child family' policy previously undertaken by the nation with the highest population in the world.

Table 1 indicates total population for 12 nations and the birth rate, or births per woman, for each country.

Infant Mortality

Infant mortality refers to death in the first year of life. While it has lessened worldwide in the twentieth century, the rates vary in different countries, ranging from 5 per 1000 live births in developed countries such as Australia and France, to as many as 76 per 1000 live births in Nigeria. Rates of Under 5 Child Mortality, while higher, are comparable nationally. Table 2 compares Infant and under 5 child mortality in 12 countries.

While infant mortality can occur from prenatal causes, or in the process of delivery, it can also occur through accident or by endemic infectious disease later within the first year of life.

In many areas of the world, fertility regulation and birth control methods have been used to increase the time between births and enable optimum developmental opportunity for each infant born. Exceptions occur in those countries where fertility regulation is forbidden by religious law. Birth spacing is accomplished in some countries by later marriage and in some by abstinence during the fertile period of ovulation.

Sexually Transmitted Disease

The most pervasive sexually transmitted diseases in the world have been syphilis, gonorrhea, and chlamydia. However, a powerful new influence on the culture of pregnancy and birth is the spread of the new infectious disease of the modern world, human immunodeficiency virus, or HIV infection, identified in 1980. While HIV can be transmitted by sharing needles among drug users, it can be transmitted sexually to both males and females. It is also transmitted from mother to infant prenatally, during delivery, or postnatally through breast feeding (WHO and UNAIDS, 1998). This has had a devastating effect internationally on the number of infants born with HIV infection or being infected within the first year of life. While the prevalence of HIV differs geographically, and in many countries is less than 1 percent, the rate in some African countries is higher, ranging from 4 percent to 12 percent, as indicated in Table 3.

The fear of becoming HIV positive is beginning in some countries to change behavior toward careful choice of sexual partners, and using barrier methods of protection. The failure to change in other areas has led to devastating epidemics of HIV.

Conclusion

Anthropological perspectives on birth enable cross- cultural comparison of healthy birth practices and effective family planning methods based on cultural acceptability. Effective fertility regulation and prenatal care enable continuing reproductive health. Cultural variation in reproductive rights, values, and education demonstrate differences in fertility, contraceptive use, and birth practices, as well as expected family size internationally.

See also: Fertility Theory; Fertility and Culture: Anthropological Insights; Fertility of Single and Cohabiting Women; Infant and Child Mortality in Industrialized Countries; Infant and Child Mortality in the Less Developed World.

Bibliography

Andrews Espy, K., Francis, D.J., Riese, M.L., 2000. Prenatal cocaine exposure and prematurity: Neurodevelopmental growth. Journal of Developmental and Behavioral Pediatrics 21 (4).

Newman, L.F., 1996. Preventing Risks of Learning Impairment. Education Commission of the United States, Denver, CO.

Rosen, J.E., Shanti, R.C., 1998. Africa's Population Challenge: Accelerating Progress in Reproductive Health, Country Study Series No. 44. Population Action International, Washington, DC.

UNICEF, 2000. The Progress of Nations: The Nations of the World Ranked According to their Achievements in Fulfilment of Child Rights and Progress for Women. The World Bank, Washington, DC. UNICEF, New York World Bank 2000 World Development Indicators, 2000.

World Health Organization, 2000. Reproductive health. Bulletin of the World Health Organization 78 (5), 563–714.

WHO and UNAIDS, 1998. HIV and Infant Feeding. 1. Guidelines for Decision-Makers. 2. A Guide for Health Care Managers and Supervisors. 3. A Review of HIV Transmission through Breastfeeding.

Bleuler, Eugen (1857–1939)

Paul Hoff, Psychiatric University Hospital, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Swiss psychiatrist Eugen Bleuler (1857–1939) became known worldwide through his theory of the 'group of schizophrenias' and their psychopathological formation, a conception that differed markedly from Emil Kraepelin's prognostically pessimistic 'dementia praecox theory' and that incorporated specifically psychoanalytical elements. Bleuler sought an understanding of psychosis that, while fully acknowledging the biological determinants, brought in the individual psychological, social, and anthropological dimension. Furthermore, he outlined a natural philosophy in his later work that was to a large extent independent of contemporary philosophical trends, but that was, however, barely taken up by either psychiatry or philosophy.

Eugen Bleuler is a central figure in the history of European psychiatry. His importance is founded on the fact that he proposed and practiced a critically mediating link between two basic psychiatric positions that stood at two controversially opposite poles at the beginning of the twentieth century (and that still do in part up to the present day) – that is, generally speaking, the tension between objective and subjective methods in clinical assessment as well as in research, or, more specifically, the psychoanalytical depth psychological school on the one hand, which was situated predominantly outside the universities, and academic psychiatry on the other, which was oriented principally toward the descriptive somatic approach. In this respect Eugen Bleuler – to put it in layman's terms – stood, as far as his basic psychiatric conviction was concerned, between the two similarly aged exponents of the aforementioned competing approaches, namely, Sigmund Freud (*see* Freud, Sigmund (1856–1939)) and Emil Kraepelin (*see* Kraepelin, Emil (1856–1926)) (Hell et al., 2001; Hoff, 2009).

Born in Zollikon, near Zurich, on 30 April 1857, Bleuler came from an old-established Swiss farming family on both his father's and his mother's side. After completing his medical studies between 1875 and 1881, he worked first as an intern at the psychiatric clinic Waldau, near Bern. Study visits took him to Paris (with Charcot and Magnan) and to Munich (with von Gudden, who had been the first director of the Psychiatric University Hospital in Zurich, named the 'Burghölzli,' from 1870 to 1872). In 1885, Bleuler went to work with Forel at the Zurich 'Burghölzli' clinic, which had been founded in 1870, only to be appointed director of the Psychiatric Hospital Rheinau in the canton of Zurich the following year, a position which he held for 12 years, that is, until 1898. The professional position that followed also covered an unusually long period of time and brought with it more far-reaching repercussions and to a large extent had a formative influence on the way Bleuler was to conceive the whole field of psychiatry. He took over the chair in psychiatry at the University of Zurich in 1898, succeeding Auguste Forel (1848–1931), a position that was attached to the directorship of the 'Burghölzli.' Bleuler held this position for almost three decades until 1927.

At the end of his academic life – he was at that time also the Rector of the University of Zurich – he was said, in an account cited by the Swiss psychiatrist J. Klaesi, to have remarked, not without a hint of maliciousness, "that the state forces one to be an old man at 72" (Klaesi, 1970). Eugen Bleuler died in Zollikon at 83 years old on 15 July 1939.

Bleuler's work is characterized by its tireless attempt, influenced by a strong sense of respect for the individual patient, to approach the enigma of psychotic illnesses on as many levels as possible, while at the same time taking into consideration general or normal psychological findings. On the one hand, he wanted to avoid making any one-sided dogmatic assertions, such as in the sense of Karl Jaspers' polemic term 'brain mythologies,' which, with no further reflection, identified psychic phenomena with neurobiological processes along ingenuously materialistic lines. On the other hand, Bleuler was concerned with regarding the psychotic person *not only* as ill, as deviant, or as disturbed, but rather he directed his attention toward the fundamental comparability to healthy mental life, a standpoint that interested him above all because of the possibility for its therapeutic exploitation. This did not mean, however, that Bleuler in any way placed little value on the then-prevailing 'brain pathological' – in twenty-first century terminology, neurobiological – school of thought and the research approaches arising from it. On the contrary, like the vast majority of contemporary psychiatrists, he too was convinced that the brain function in the etiology and pathogenesis of endogenously psychotic illnesses – predominantly, therefore, schizophrenic and (bipolar) affective disorders – played a decisive role. Unlike the many 'simplifiers,' however, Bleuler always insisted that it was of great importance for the handling of patients, for the diagnosis, and above all for the therapy to *understand* the clinical picture and each individual, concrete symptom, independent of the suspected etiology.

Bleuler's handling of the basic concepts of association theory was not unified and hence was liable to give rise to misunderstandings and controversial statements. On the one hand, he repeatedly took up a clear position against the 'association psychology' in the narrow sense advocated earlier in the nineteenth century, in which (healthy as well as ill) human mental life was merely the sum total of individual intellectual associations prompted by sensory perception. Against this, he set an affective psychology that aimed to comprehend the personality in its individual totality, but that

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61009-7

nonetheless was coupled substantially with elements from association psychology in the field of the schizophrenias. Indeed, not only did Bleuler see no contradiction in this, he even saw it as objectively necessary. It was here that the (then as now) topical controversy surrounding the quantifying scientific *explanation* and the interpretative humanistic *understanding* of mental and, in particular, psychotic manifestations arose. Bleuler also attempted in this context to overcome a rigid distinction between the two approaches, quite unlike Karl Jaspers, who strongly criticized him over this on several occasions (e.g., in Jaspers, 1923).

Unlike most of his university colleagues in his field, Bleuler saw in general terms *no* fundamental contradiction between (explanatory) brain research in the broadest sense and the interpretative access to the psychotic of depth psychology, which at that time was almost exclusively advocated for by the still-young Sigmund Freud. On the contrary, Freud's theory seemed to Bleuler to be virtually predestined for critical implementation into clinical psychiatry. For Freud himself, who fought a lifelong struggle to achieve 'academic' recognition for his theory, such a competent and influential researcher like Bleuler naturally took on a particularly great importance. Bleuler, however, emphasized the aspect of the *critical* adoption of psychoanalysis already mentioned so much over the years that it led to a rift (Bleuler, 1913; Dalzell, 2007; Küchenhoff, 2001). Much has been written about the escalating debate between Freud and his particularly creative and original disciples, such as Carl Gustav Jung (*see* Jung, Carl Gustav (1875–1961)) or Alfred Adler. It is less well known that Eugen Bleuler also must be mentioned in this connection. The psychiatry historians Alexander and Selesnick (1969) went so far as to call him the 'first victim' of this controversy, as Bleuler resigned from the International Psychoanalytic Society in 1910 because of his discontent at its, in his view, authoritarian style of leadership. The fundamental importance of the depth psychology approach for psychiatry was for him in no way diminished because of this (Küchenhoff, 2009; Wieser, 2001). Scherbaum (1992) pointed out the paradoxical situation that Bleuler was 'too psychoanalytical' for academic colleagues in his field, whereas the psychoanalysts for their part viewed him as being too critical, even 'disloyal,' and he posed the question of whether Bleuler's understanding of psychoanalysis might not also have been shaped by the same 'scientistic self-misconception,' which Habermas (1979) had observed in Freud himself.

One book chapter was of decisive importance for the development of *clinical psychiatry*, a chapter of 420 pages that Bleuler (1911) had written for the *Handbuch der Psychiatrie (Handbook of Psychiatry)*, edited by Gustav Aschaffenburg, entitled *Dementia praecox oder Gruppe der Schizophrenien* (1911; *Dementia praecox, or the Group of Schizophrenias)* (Bleuler and Bleuler, 1986; Kaplan, 2008; Kuhn, 2004; Scharfetter, 2001, 2006; Stotz-Ingenlath, 2000). The agenda was already set out in its title. Unlike Kraepelin, who was searching for 'natural disease units' and through whose influential textbook the concept of 'dementia praecox' had become both popular and influential in the German-speaking world (Hoff, 1994), Bleuler did not speak of schizophrenia as a *single* unit, but rather he spoke of the *group* of schizophrenias. He was concerned, however, with the question of which characteristics the otherwise-heterogeneous schizophrenic patients had in common. Bleuler was describing first and foremost *psychopathological* phenomena and was avoiding any all-too-narrow (and speculative) analogies to the brain function.

Two psychopathological distinctions that Bleuler introduced are important with regard to schizophrenic psychoses and frequently are confused: the basic and accessory symptoms on the one hand, and the primary and secondary symptoms on the other. *Basic symptoms* were, according to Bleuler, peculiar to every schizophrenic disease process, independent of its other clinical manifestation. Among these he counted disturbed associative processes (a kind of dissociation relating not only to formal thought processes but also to all other mental functions – hence the term *schizo*phrenia meaning *splitting/dissociation* of mental processes), schizophrenic autism (not to be confused with the syndrome of the same name in child and adolescent psychiatry), ambivalence (a volitional disorder), and various types of affective disturbances. *Accessory*, that is, not necessarily existing and also frequently observed in nonschizophrenic and above all exogenous psychoses, meant symptoms such as hallucinations, delusion, or catatonic states. Although this differentiation is still of great heuristic value for psychopathological research in the twenty-first century, the differentiation between the *primary symptoms* – which result directly from the suspected organic disease process – and the *secondary symptoms* – which constitute a subjective psychological reaction to the illness by the person affected – has lost considerable importance.

In contrast to the (as the term implies) prognostically pessimistic concept of 'dementia praecox,' such as in Kraepelin's work, Bleuler constantly stressed that there were also benign schizophrenic courses – and by no means only as an exception – that after one or more episodes turned into an extensive or complete recovery.

The *Lehrbuch der Psychiatrie* (*Textbook of Psychiatry*), which first appeared in 1916, achieved a virtually legendary reputation. In this work, Bleuler (1916) put forward his deeply personal view of this field in a systematically, vividly, and clinically oriented manner. He had worked 5 years on this book. It went through many editions, for which he himself took on the task of updating as necessary up to and including the sixth edition. After his death (1939), his son Manfred Bleuler took over the task.

Bleuler's (1925) writings on general psychology and philosophy have, on the other hand, all but been forgotten in the twenty-first century. In his later work – such as in the work *Die Psychoide als Prinzip der organischen Entwicklung* (1925; *Psychoids as a Principle of Organic Development*) – he attempted to develop a comprehensive natural philosophical concept that, however, was not adopted on any great scale at the time. His basic idea of a memory that was independent of conscious processes and did not only refer to the individual person but also – and thus in a way an evolutionary biological perspective – to the whole species ('Mneme'; Bleuler, 1934) was, in its idiosyncratic embracing of biogenetic vitalistic ideas of the late nineteenth century, of central importance for Bleuler's whole scientific understanding. Although Bleuler considered inappropriate the classification of his (later) work as 'vitalistic,' from the point of view of the history of twenty-first century psychiatry, the parallels are unmistakable. This is true in

particular if one looks at Bleuler's objective of a comprehensive 'science of life,' which did not separate the categories of physical, biological, mental, and even social phenomena from each other, but rather sought to understand them as scientifically equal forms to express a single integrative (life) principle (Möller and Hell, 2000; Möller, 2001). A weakening or even lifting of the aforementioned categorical dividing lines (although not *this* particular way) is a highly topical subject in psychiatric research, which, indeed, appears less and less willing to maintain the classical differentiation of 'organic' and 'endogenous' as well as 'psychogenous' illnesses.

The approaches in his later work, irritating and speculative as they seem to twenty-first century readers, may be regarded a critical part of Bleuler's thinking. Yet another aspect is even more problematic: As many contemporary academic psychiatrists, Bleuler – especially regarding questions of forensic psychiatry that are also in his work – repeatedly referred to the concept of 'degeneration' (Entartung). Even his explicit awareness that the term 'degeneration' was vague and scientifically questionable did not prevent him from discussing and applying it in several contexts. A considerable number of research activities addressed this field in the past decades, often focusing on classical authors like Emil Kraepelin, Richard von Krafft-Ebing, or Eugen Bleuler (Pick, 1989; Hermle, 1986; Hoff, 2008).

In view of the whole picture, however, Eugen Bleuler was one of the most original and consistent psychiatric clinicians and researchers of his time. His thought is also of interest to current research into psychosis and has been receiving increased attention in recent literature (Hoff, 2010, 2012; Moskowitz and Heim, 2011). What is particularly gratifying about this is the fact that research into the history of psychiatry has turned away extensively from the bad habit, which one often came across in the early years, of uncritically hagiographic and therefore unscientific excessiveness from eminent advocates in the field. It now is concerning itself discernibly with a comprehensive understanding of the history of the field supported by the original literature and other sources, and is embedded in contexts from social history and the history of ideas.

See also: Dementia: Psychiatric Aspects; Freud, Sigmund (1856–1939); Kraepelin, Emil (1856–1926); Normality and Mental Health; Nosology in Psychiatry; Philosophy, Psychiatry, and Psychology; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Schizophrenia.

Bibliography

Alexander, F.G., Selesnick, S.T., 1969. Geschichte der Psychiatrie. Konstanz, Germany: Diana.

Bleuler, E., 1911. Dementia praecox oder die gruppe der schizophrenien. In: Aschaffenburg, G. (Ed.), Handbuch der Psychiatrie. Spezieller Teil, sect. 4. Deuticke, Leipzig, Germany, pp. 1–420.

Bleuler, E., 1913. Kritik der Freud'schen theorien. Allgemeine Zeitschrift für Psychiatrie und Psychisch-Gerichtliche Medizin 70, 665–718.

Bleuler, E., 1916. Lehrbuch der Psychiatrie. Springer, Berlin.

Bleuler, E., 1925. Die Psychoide als Prinzip der Organischen Entwicklung. Springer, Berlin.

Bleuler, E., 1934. Mnemismus, psychoide. Schweizer Archiv für Neurologie und Psychiatrie 33, 177–191.

Bleuler, M., Bleuler, R., 1986. Books reconsidered. Dementia praecox oder die gruppe der schizophrenien: Eugen Bleuler. British Journal of Psychiatry 149, 661–664.

Dalzell, T.G., 2007. Eugen Bleuler 150: Bleuler's reception of Freud. History of Psychiatry 18, 471–482.

Habermas, J., 1979. Erkenntnis und Interesse, fifth ed. Suhrkamp, Frankfurt/Main, Germany.

Hell, D., Scharfetter, C., Möller, A. (Eds.), 2001. Eugen Bleuler – Leben und Werk. Hans Huber, Bern Göttingen.

Hermle, L., 1986. Die degenerationslehre in der psychiatrie. Fortschritte der Neurologie und Psychiatrie 54, 69–79.

Hoff, P., 1994. Emil Kraepelin und die psychiatrie als klinische wissenschaft. Ein beitrag zum selbstverständnis psychiatrischer forschung. In: Hippins, H., Janzanik, W., Müller, C. (Eds.), Monographien aus dem Gesamtgebiete der Psychiatrie, vol. 73. Springer, Berlin.

Hoff, P., 2008. Kraepelin and degeneration theory. European Archives of Psychiatry and Clinical Neuroscience 258 (Suppl. 2), 12–17.

Hoff, P., 2009. Historical roots of the concept of mental illness. In: Salloum, I.M., Mezzich, J.E. (Eds.), Psychiatric Diagnosis – Challenges and Prospects. Wiley-Blackwell, Chichester, pp. 3–14.

Hoff, P., 2010. Sicht der klassiker: der affekt in den psychosekonzepten von Emil Kraepelin und eugen bleuler. Psychiatrie und Psychotherapie 6, 98–103.

Hoff, P., 2012. Eugen Bleuler's concept of schizophrenia and its relevance to present-day psychiatry. Neuropsychobiology 66, 6–13.

Jaspers, K., 1923. Allgemeine Psychopathologie, third ed. Springer, Berlin.

Kaplan, R.M., 2008. Being Bleuler: the second century of schizophrenia. Australasian Psychiatry 16, 305–311.

Klaesi, J., 1970. Eugen Bleuler (1857–1939). In: Kolle, K. (Ed.), Große Nervenärzte, second ed., vol. 1. Thieme, Stuttgart, Germany, pp. 7–16.

Küchenhoff, B., 2001. Die auseinandersetzung Eugen Bleulers mit Sigmund Freud. In: Hell, D., Scharfetter, C., Möller, A. (Eds.), Eugen Bleuler – Leben und Werk. Huber, Bern Göttingen, pp. 57–71.

Küchenhoff, B., 2009. Bleuler, Freud and Jung on dementia praecox (schizophrenia) in 1908. The Letter 40, 69–82.

Kuhn, R., 2004. Eugen Bleuler's concepts of psychopathology. History of Psychiatry 15, 361–366.

Möller, A., Hell, D., 2000. Prinzipien einer naturwissenschaftlich begründeten ethik im werk eugen bleulers. Nervenarzt 71, 751–757.

Möller, A., 2001. Grundpositionen im spätwerk. In: Hell, D., Scharfetter, C., Möller, A. (Eds.), Eugen Bleuler – Leben und Werk. Huber, Bern Göttingen, pp. 104–112.

Moskowitz, A., Heim, G., 2011. Eugen Bleuler's dementia praecox or the group of schizophrenias (1911): a centenary appreciation and reconsideration. Schizophrenia Bulletin 37, 471–479.

Pick, D., 1989. Faces of Degeneration: A European Disorder 1848–1918. Cambridge University Press, Cambridge.

Scharfetter, C., 2001. Eugen Bleuler's schizophrenias – synthesis of various concepts. Schweizer Archiv für Neurologie und Psychiatrie 152, 34–37.

Scharfetter, C., 2006. Eugen Bleuler (1857–1939). Polyphrenie und Schizophrenie. vdf Hochschulverlag, Zürich.

Scherbaum, N., 1992. Psychiatrie und psychoanalyse—Eugen Bleulers dementia praecox oder gruppe der schizophrenien (1911). Fortschritte der Neurologie und Psychiatrie 60, 289–295.

Stotz-Ingenlath, G., 2000. Epistemological aspects of Eugen Bleuler's conception of schizophrenia in 1911. Medicine, Health Care and Philosophy 3, 153–159.

Wieser, A., 2001. Zur frühen psychoanalyse in Zürich. In: Hell, D., Scharfetter, C., Möller, A. (Eds.), Eugen Bleuler – Leben und Werk. Huber, Bern Göttingen, pp. 78–84.

Bloch, Mark Léopold Benjamin (1886–1944)

Peter Schöttler, Centre National de la Recherche Scientifique, Institut d'histoire du temps présent, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The French historian Marc Bloch (1886–1944) is one of the authors most frequently cited by those working in contemporary historical disciplines around the world. This immense influence is founded on both scholarly and biographical grounds. As one of the founders, together with Lucien Febvre, of the journal '*Annales d'histoire économique et sociale*' in 1929, and as a pioneer in the field of the history of society and mentality, a historical theoretician, and a critical intellectual who was executed by the Gestapo in 1944, Bloch has become an inspiration and model for generations of researchers working in a diverse range of scientific fields and from a variety of different political perspectives.

The French historian Marc Bloch is one of the authors most frequently cited by those working in contemporary historical disciplines around the world. This immense influence is founded on both scholarly and biographical grounds. As one of the founders of the field of the history of society and mentality, a historical theoretician, and a critical intellectual who was executed by the Gestapo in 1944, Bloch has become an inspiration and model for generations of researchers working in a diverse range of scientific fields and from a variety of different political perspectives.

Marc Bloch was born in Lyon on 6 July 1886 in a family of practicing Jews of Alsatian heritage. Bloch himself was an agnostic, and in his final will and testament he explicitly requested a secular funeral. Because his father was Professor for Roman History at the Sorbonne, Bloch grew up in Paris, where he visited the elite secondary school *Louis-le-Grand*. From 1904 to 1908, he studied history and geography at the *École Normale Supérieure* in Paris; following this period he spent two semesters in Berlin and Leipzig, where he attended lectures by scholars such as Gustav Schmoller, Karl Bücher, and Karl Lamprecht. From 1909 to 1912, he was the recipient of a *Fondation Thiers* scholarship, before becoming a teacher at secondary schools in Montpellier and Amiens. Bloch fought on the front as an enlisted soldier in World War I, and in the course of being repeatedly wounded in action and decorated for his service he was promoted to the rank of captain. In 1919, he married and was appointed lecturer for medieval history at the newly founded French University of Strasbourg. In 1921, he was appointed associate professor, and in 1927, he received his full professorship. In 1929, together with his Strasbourg colleague Lucien Febvre (1878–1956), Bloch founded the journal *Annales d'histoire économique et sociale*, which rapidly grew to become a widely influential forum for innovative historiography. In 1936, after several unsuccessful attempts to apply for a position in Paris, Bloch was granted a Sorbonne professorship in economic history. Immediately following the outbreak of World War II, he volunteered for military service to defend his homeland against Nazi Germany. As an officer sent to reinforce the First Army, he experienced both the anticipation of the '*Sitzkrieg*' and then the unexpectedness of the '*Blitzkrieg*' in the French coastal canal region. Like thousands of other French soldiers, Bloch was rescued from Dunkirk in June of 1940 in the emergency evacuation across the channel. From England, he returned immediately to France, where he was able to avoid imprisonment. Because Nazi law prohibited him from reassuming his teaching position at the Sorbonne, the Vichy government sent him to rejoin the faculty of the University of Strasbourg, which had been relocated to Clermont-Ferrand. At the same time, Bloch sought, and eventually attained, an appointment at the New School of Social Research in New York. Because of the complicated provisions for obtaining a visa, however, he would have been compelled to leave behind most of his family in France – a condition Bloch, the father of six children, was entirely unwilling to accept. In December of 1940, Bloch was barred from government service by the 'Jewish Statute,' only to be exempted several weeks later along with a number of other prominent Jewish scholars, and allowed to reassume his teaching duties. For the sake of his wife's health, Bloch applied a year later for a transfer to Montpellier, even though this meant that he would no longer be permitted to hold public lectures. It was at this point in time that Bloch's active involvement in the *Résistance* began. Following the occupation of the Vichy zone by the German army in November 1942, Bloch left Montpellier with his family for his countryside home near Guéret in the Creuse district of France. From this time on he worked and lived for the clandestine resistance. Operating under pseudonyms like 'Chevreuse,' 'Arpajon,' and 'Narbonne' he belonged to the masterminds of the left-wing '*Franc-Tireur*' movement in Southern France. In 1943, he moved to Lyon, where he belonged to the regional leadership of the '*Mouvements Unis de la Résistance*.' On 8 March 1944, Bloch was arrested by the Gestapo and brutally tortured before being thrown in prison. On 16 June, several days after the Allied troops landed at Normandy, Bloch and 29 other prisoners were taken to a field on the outskirts of Lyon and shot.

The thematic emphasis of Bloch's scholarly production lies in the social and economic history of the medieval period. Although, Bloch was initially concerned primarily with the topic of serfdom, in the early 1920s he began to devote his attention to what appeared to be a highly specific question in the field of political theology: namely the issue of the widespread belief in feudal France and England in the healing power of the king; in particular, when he laid his 'holy hands' upon the sick who had gathered before the cathedral following the coronation ceremony. This anthropological approach,

however, so clearly inspired by Durkheim's sociology of religion, was one Bloch would only use intermittently in later years. Instead, he turned his focus to a comparative historiography of (1) European agricultural structures and (2) societal structures that spanned the period from late antiquity to early modern times. As he wrote in his 1934 application for a teaching position at the *Collège de France*, he no longer considered himself to be first and foremost a 'medievalist,' but rather a general historian of 'European societies.' His two-volume work *Feudal Society* (1939–40) can be understood as an exemplary attempt at putting this perspective of a '*histoire totale*' into practice. In what amounted to an intellectual testament, Bloch himself summed up his book as follows: "I have given an example of something […] that I chose to call 'the dissection of a social structure.' In the course of the development of Western society there was a phase, which we call feudal society, which had a certain social tone. It was this tone that I wanted to investigate. […] Furthermore, I have attempted to bring into play in a European context the multifarious experiences, which the comparative method allows us to grasp. If my work is truly original in one regard, then in my opinion its originality lies in both of these efforts: structural analysis and the incorporation of comparative experiences" (*Cahiers Marc Bloch*, No. 2, 1995, p. 16).

Bloch's originality lies indeed in this 'structural' and comparative approach, but also in the manner in which he conducts his scholarly interventions. Precisely because he took the approach of Durkheim disciples, such as François Simiand, seriously, who in their work had critically engaged the traditional praxis of political history, Bloch had been confronted with the necessity of making a fundamental change in the theory and practice of historiography. In countless essays and reviews he called not only for a new, interdisciplinary, and internationally oriented historiography, which would address itself to new themes by means of new methods, but also for new ways of educating historians and new forms of communication within the historian's guild. An important instrument, a sort of editorial 'lever,' in his efforts to this end was the journal *Annales*, which he published together with Lucien Febvre. It was in the course of this cooperative venture, and also as a result of the ensuing friendly competition between him and Febvre that Bloch developed his remarkable ability to quickly and succinctly formulate new scholarly perspectives. That is why the fight between the two coeditors was all the more painful, when, in the aftermath of the French defeat in 1940, they were forced to decide if and how the journal should continue to be published. Whereas Bloch argued in favor of discontinuing the publication of the *Annales* until the end of the War, Febvre was prepared to adapt – at least on the surface level – to the new conditions under occupation in order to continue to have a public voice. In the end, Bloch accepted this strategy and continued to publish in the *Annales* under the pseudonym 'Fougères' until the very end. At the same time, he continued to work on two book-length manuscripts, which were edited by Febvre after the War: an analysis of the French defeat (*Strange Defeat*) and a reflective work on the intellectual preconditions of historiography as: science, craft, and art (*The Historian's Craft*).

The beginning of Bloch's considerable reputation in French as well as international scholarly circles can be traced back to the 1930s. It was really after World War II, however, particularly in the closing decades of the twentieth century, that Bloch's ascendancy to the ranks of one of the world's most quoted historians took place. His books have been translated into all of the world's major languages: the number of printed copies of *The Historian's Craft* extends well into the six-figure range, one-third of which have appeared in Latin America. Bloch's heroic death has elevated him to the status of an icon, adored by intellectuals and academics representing a broad spectrum of disciplines and political convictions alike. Accordingly, one encounters a range of conflicting images of 'Marc Bloch': the Marxist or the conservative, the Jew or the nationalist, and recently even the 'postmodernist' Bloch, who supposedly drew on Nietzsche as a source of inspiration. Whereas such portrayals threaten to bury the author's historical work under ever-increasing layers of reception, recent years have seen an increased effort toward precise scholarly contextualization of this work in order to better reconstruct Bloch's contribution to the historical discipline of the twentieth century.

See also: Anthropology and History; Comparative History; Economic History; Feudalism; Historical Thought and Historiography: Current Trends; History and the Social Sciences; History: Theories and Methods; Middle Ages, The; Social History.

Bibliography

Works by Marc Bloch

Bloch, Marc, 1924. Les Rois thaumaturges. Étude sur le caractère surnaturel attribué à la puissance royale, particulièrement en France et en Angleterre. Gallimard, Paris, 1983. Engl. Trans. The Royal Touch, London, 1973.

Bloch, Marc, 1939–1940. La Société féodale. Albin Michel, Paris, 1994. Engl. Trans. Feudal Society, 2 vols., London, 1961.

Bloch, Marc, 1949. In: Bloch, E. (Ed.), Apologie pour l'histoire ou Métier d'historien. Armand Colin, Paris, 1993. Engl. Transl. The Historian's Craft, New York, Knopf, 1953.

Bloch, Marc, 2006. In: Boch, E., Becker, A. (Eds.), L'Histoire, la guerre, la résistance. Gallimard, Paris.

Work on Marc Bloch

Bloch, E., 1997. Marc Bloch 1886–1944. Une biographie impossible – An Impossible Biography. Culture & Patrimoine en Limousin, Limoges (bilingual).

Fink, C., 1989. Marc Bloch. A Life in History. Cambridge University Press, Cambridge.

Friedman, S.W., 1996. Marc Bloch, Sociology and Geography. Encountering Changing Disciplines. Cambridge University Press, Cambridge.

Schöttler, P., Rheinberger, H.-J., 2011. Marc Bloch et les crises du savoir. Max Planck Institut für Wissenschaftsgeschichte, Berlin.

Bloomfield, Leonard (1887–1949)

Manfred Bierwisch, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Leonard Bloomfield, the most important American linguist in the first half of the twentieth century, began his studies in comparative historical philology, which he extended right from the beginning beyond the Indo-European languages by investigating Austronesian and Indian languages. While he initially adopted Wundt's notion of *Völkerpsychologie* as theoretical background of linguistics, Bloomfield developed a strictly antimentalistic, essentially behavioristic view in his groundbreaking book *Language* (1933). According to this objective approach, linguistic analysis has not to deal with what the speakers have in mind, but only with acoustic properties and principles of distribution of the basic elements of their utterances. In spite of this restrictive doctrine of '*Distributionalism*', Bloomfield relied on complex rule systems of necessarily mentalistic nature to formulate important phonological insights in a seminal paper on *Menomini Morphophonemics* (1939).

Biography

Leonard Bloomfield, one of the key figures of American linguistics, was born in Chicago on 1 April 1887. He was the son of Sigmund and Carola Bloomfield, who immigrated from Austrian Silesia, and the nephew of the renowned Indologist Maurice Bloomfield. After completing secondary school in Chicago, Bloomfield went for undergraduate studies to Harvard College in 1903, which he left with the Bachelor degree in 1906. During his graduate studies from 1906 to 1908 at the University of Wisconsin, he was impressed by Eduard Prokosch (1876–1938), a philologist at the German department, who motivated Bloomfield's decision for a career in linguistics. In 1908, he went to the University of Chicago, where he received his PhD, submitting a dissertation in Germanic historical linguistics. In March 1909, he got married to Alice Sayers.

Bloomfield began his academic career in 1909 as Instructor of German at the University of Cincinnati. He moved to the University of Illinois in 1910, where he got a position as Assistant Professor of Comparative Philology and German in 1913. He left for 1 year of graduate studies with the leading Indo-Europeanists Karl Brugmann and August Leskien at the University of Leipzig and the Sanskrit specialist Jabob Wackernagel in Göttingen. Returning to the University of Illinois in 1914, he published his textbook *An Introduction to the Study of Language*. Even though Bloomfield's primary occupation was German philology, he was convinced that the study of language must not be restricted to the historically documented Indo-European languages. He therefore began to explore Tagalog, an Austronesian language, which he described and analyzed with a native speaker of Tagalog, who was a student of the University of Illinois. Parts of this research were published in 1917. With the same motivation, Bloomfield undertook fieldwork on different Algonquian languages in 1920 and 1921, to which he returned on various later occasions. In 1921, he moved to the Ohio State University, where he was Professor of German and Linguistics until 1927. During this period, he was in close contact with the psychologist Albert Paul Weiss (1879–1931), an early advocate of behaviorism. In 1925, Leonard Bloomfield and George M. Bolling from Ohio together with Edgar H. Sturtervant from Yale initiated the foundation of the Linguistic Society of America that soon grew into the most influential organization of the field. The initial Call, drafted by Bloomfield, appeared as the opening paper of the Society's regular journal *Language*. During the summer of 1925, Bloomfield continued his Algonquian fieldwork under the auspices of the Canadian Department of Mines. In 1927, Bloomfield moved to the University of Chicago, where he held a position as Professor of Germanic Philology until 1940. His most important colleague there was the anthropologist Edward Sapir, one of the prominent students of Franz Boas. During Bloomfield's time in Chicago, the publication of his extremely influential book *Language* (1933), his growing influence on the field of linguistics in general, and his international reputation made him one of the leading representatives of American linguistics. He was President of the Linguistic Society of America in 1935, his courses at the Society's Linguistic summer institute 1938–41 were of unusual influence. Although he was not a charismatic person, easily dominating a large audience, but rather of reserved and unimposing personality, he still had an enormous impact on the field. He had few students, but many linguists regarded themselves as members of a school bearing his name. His unique effect was completely due to sober argument and the strength and clarity of his writings. In 1940, he became Sterling Professor of Linguistics at Yale University as the successor of Edward Sapir, who had moved there from Chicago 6 years before. During World War II, Bloomfield participated actively in the Intensive Language Program of the American Council of Learned Societies, applying his theoretical insights to practical problems of textbooks and courses for strategically important languages. In May 1946, he suffered a heavy stroke, from which he never fully recovered. He died in New Haven on 18 April 1949.

The Background of American Linguistics

When Bloomfield began his career, language was largely the subject matter of philological departments. Although grounded in the great tradition of European historical and comparative linguistics, which had revealed the systematic relationship among the Indo-European languages, American linguistics

gradually developed its own, independent orientation during the first decades of the twentieth century. Step by step, independent linguistic departments got established, so that a strong and influential discipline flourished, when Bloomfield left the field. Besides the philological European heritage, during this development two major factors played an important role, in both of which Bloomfield was decisively involved. First, there was a characteristic interest to document and analyze a wide variety of languages outside the Indo-European family, to which linguistic analysis was traditionally constrained. Elementary motivation for the extended perspective came from the concern for language and culture of American Indians, the systematic study of which was initiated by the great anthropologist Franz Boas (1858–1942). Among the active followers of Boas were his student Edward Sapir (1884–1939), and Leonard Bloomfield, who shared this orientation and extended it by his Tagalog studies to further language families. The interest in languages without written records raised methodological issues unfamiliar to European historical linguistics. While the neogrammarian Hermann Paul (1846–1921) in his influential book *Prinzipien der Sprachgeschichte* (1880) emphatically denied the possibility of a scientific approach to language other than the historical one, the American-Indian languages required exactly this, namely the analysis of speech independent of any written historical records. The systematic foundation of synchronic description as a scientific approach to language leads to the second characteristic trait of American linguistics, viz., the development of a specific descriptive methodology, whose characteristic, rigorous form became known as 'distributionalism'. This interest in methodological clarification of descriptive linguistics corresponds to a large extent with similar developments in European linguistics in the first decades of the twentieth century, articulated programmatically by Ferdinand de Saussure (1857–1913) in his *Cours de linguistique générale* (1916). In spite of shared intentions, American linguistics made its own specific contribution to the overall development called 'structural linguistics'. Many of the inspiring proposals, taking up Saussurean ideas in the American context, were due to Edward Sapir. But the classical formulation of the framework, which by its systematic character eventually shaped the development and spelled out its principles, was due to Leonard Bloomfield. In spite of their mutual respect and scientific appreciation, there is an interesting tension between the personality as well as the intellectual approach of these two important men. While Sapir was burning with brilliant ideas, Bloomfield organized the field by careful and systematic argument. The tension is inherent, moreover, in Bloomfield's own contribution, where interesting friction between theoretical foundation and factual insight is to be observed in several respects.

From Bloomfield's training and intense experience in Indo-European philology, two important consequences emerged for the orientation of his further research. First, a basic doctrine of the neogrammarians with whom Bloomfield had studied in Leipzig and Göttingen was the claim that sound laws have no exceptions. According to this doctrine, apparent exceptions to phonetic regularities are to be explained by additional laws, a commitment that had led to various spectacular insights. Bloomfield emphatically stressed this principle in *A note on sound change* (1928), taking it as a genuine step in his more general attempt to clarify the scientific nature of his field of interest, which had to be based on objective principles, a goal that concerned him from the very beginning. The second, equally important insight, which Bloomfield, like Sapir, adopted from comparative historical studies, was the principle of internal reconstruction that explains the differences between related dialects by means of different changes applying to common historical predecessors. This principle could be shown to apply with equal success to languages of completely different structure and history. On the basis of extensive fieldwork on Fox, Cree, Ojibwe, and Menomini, four separate Algonquian dialects, Bloomfield (1946), an important, posthumously published investigation, explained their relationship by essentially the same principles as that among the Indo-European languages, with the reconstructed dialect of Proto-Algonquian as the common historical source. This was a remarkable insight with respect to linguistic methodology as well as empirical facts. Generally speaking, the development of American linguistics, in which Bloomfield participated and which he decisively shaped, depends on European sources in various respects, but it extends and modifies this heritage in genuine and sometimes spectacular ways.

Behaviorist Orientation of Linguistics

Starting his career within the tradition of Indo-European philology, Bloomfield recognized right from the beginning the need for an overall framework, within which the study of language has its natural place and context. In his first monograph, *An Introduction to the Study of Language* (1914), he adopted Wilhelm Wundt's overall notion of 'Völkerpsychologie' to this effect. According to this conception, language belongs to the spirit or soul of its respective nation, it is, in any case, a mental entity outside the range of experimental investigation. In spite of the neogrammarians' concern for exception-less sound laws, the study of language essentially differed from the natural sciences.

A radical change in Bloomfield's attempt to provide linguistics an objective scientific foundation by general methodological standards is marked by the axiomatic approach he adopts in *A set of postulates for the science of language* (1926). In this programmatic approach, the psychological orientation of European linguistics is replaced by a strictly operationalist conception that he became familiar with through the work of his colleague Albert Paul Weiss at the Ohio State University. This perspective, systematically elaborated and convincingly illustrated in Bloomfield's most influential chef d'œvre *Language* (1933), is ambivalent or even paradoxical in characteristic ways.

On the one hand, Bloomfield criticizes Hermann Paul and other European linguists not only for their exclusively historical orientation, but even more for their imprecise psychological notions on which their linguistic concepts are based, a perspective that Bloomfield himself had accepted in his *Introduction* in 1914. He now considered the scientific analysis of language as a strictly objective matter, without any commitment to concepts taken from psychology. Under this perspective, he considered it as completely irrelevant for descriptive linguistics which sort of psychology one adheres to.

Linguistics should be established as an autonomous discipline that need not rely on any concepts other than its own.

On the other hand, Bloomfield was strictly opposed to one particular type of psychology, insisting on a rigorous exclusion of all mentalistic concepts from scientific description and explanation. He considered any resort to internal states and mental accounts of linguistic (or other types of) behavior as unwarranted at the present state of knowledge, and presumably unnecessary and hence impermissible for scientific accounts in principle. In fact, Bloomfield advocated what he called a mechanistic explanation of behavior, opting more specifically for Weiss' version of behaviorism. Both aspects of this ambivalent position – the behaviorist account of language as well as the principle of autonomous, formal analysis – are carefully worked out and applied to all kinds of concrete problems in the 28 chapters of *Language*, supporting his unusual influence during the decades to come.

Bloomfield's proposal to extend the standard behavioristic stimulus–response schema to the conditions of linguistic behavior is original, instructive, and misguided. Assuming that the relation between a conditioned response R, caused by a stimulus S, as represented in the standard schema S → R, where the arrow indicates the 'internal' causal connection between two types of activity, can be mediated by a linguistic act, Bloomfield arrives at the following proposal: An external stimulus S causes – instead of (or in addition to) the practical response R – a substitute, linguistic response r of the speaker, which now acts as a substitute, linguistic stimulus s on the hearer, causing the practical response R. Thus, the basic schema S → R is turned into the complex schema S → r…s → R, where the dotted line between r and s indicates the fact that the linguistic response r functions as the linguistic stimulus s. For the sake of illustration, Bloomfield imagines a situation, where S is the condition of being hungry and seeing an apple on a tree, while R is the act of fetching and eating it. This basic connection, S → R is now to be mediated by the substitute reaction r of making some appropriate noise, to be transcribed as, e.g., *I am hungry*, and the perception of that very noise acts as the substitute stimulus s. Thus the dotted line between r and s represents the interindividual relation connecting the production and perception of a given phonetic form, while the arrows between S and r and between s and R represent the intra-individual association of this form with the practical stimulus imposed on the speaker and the reaction produced by the hearer, respectively. What is misleading in this proposal is the fact the features of r and s, the produced and perceived signal of the utterance in question, must be identical in the relevant respects. Otherwise the assumed connection would not result. This identity cannot in general be the case for the features of S and R, as the stimulus S and the response R are supposed to be causally connected, although not, of course, to be identical. But Bloomfield explicitly considers S and R as the meaning of the utterance (s…r). In Bloomfield's example, seeing and wanting an apple and fetching do not have the same properties, and there is hardly an utterance that is causally connected to both of them. Hence even if there were a systematic, causal relation between the stimulus S and the substitute response r, this relation could not in general hold between s and R. This observation is by no means restricted to Bloomfield's original example, it rather applies to the relation between language and the conditions of use in a rather principled way. In other words, the behavioristic attempt to account for linguistic behavior in terms of conditioned stimulus–response connections was doomed to fail, even though it was taken for granted by a whole generation of American linguists and psychologists, including Skinner with his programmatic *Verbal behavior* (1957).

The paradoxical fact is that in spite of advocating this misleading theoretical orientation, Bloomfield was extremely successful not only by his impact on the field, but also in his own work on methodological issues as well as in rich descriptive and analytical work in many domains. This in turn was completely in line with his claim that linguistics should be self-contained and independent of commitments to any psychological theory. But one characteristic consequence of this theoretical orientation was the sober, albeit frustrating conviction that a systematic theory of linguistic meaning could only be developed on the basis of a complete, explicit account of all possible aspects of speaker's and hearer's environment and the physiological mechanisms determining their pertinent reactions. For this reason, Bloomfield considered the scientific treatment of semantics out of reach, at least for a period of time, and perhaps in principle, although he had no doubts about the central role of meaning for any satisfactory conception of language. For three decades, skepticism toward the possibility of a serious theory of meaning became a characteristic feature of American as opposed to European linguistics. There are explicit residues of this skepticism even in Chomsky's *Syntactic Structures* (1957), which initiated a radically different approach and paved the way toward an explicitly mentalistic theory of language.

Structural Analysis of Language

Restricting the central task of linguistics to the descriptive analysis of linguistic form indicated as (r…s), i.e., to the properties that determine the production and perception of linguistic expressions, Bloomfield developed a systematic and comprehensive account of the properties linguistic expressions might possibly exhibit. The central idea of the framework carefully laid out in *Language* is the assumption, that all relevant aspects of verbal utterances can be captured on the basis of strictly formal criteria, identifying their parts in terms of articulatory and perceptual distinctions and their subsequent classification according to their possible occurrences. More specifically, basic linguistic forms, called morphemes, are characterized as sequences of phonemes, where phonemes are classes of minimal segments identified by their articulatory properties (roughly corresponding, incidentally, to letters in alphabetic writing); morphemes can then be grouped into form classes on the basis of their distribution, i.e., observable combination, giving rise to the properties of so-called constructions that are the invariant skeleton of increasingly complex utterances.

Methodologically, this framework was based on systematic segmentation of utterances and classification of their components in terms of strictly acoustic and distributional criteria, applied to actual utterances in the sense of well defined types of physical events. In this way, segments of the

acoustic stimulus are classified as phonemes, sequences of phonemes are identified as morphemes, morphemes are grouped into several classes, particular sequences of which are classified as constructions. The details of this elegant framework – the relevant criteria for segmentation and classification, the properties of primitive and complex elements, and the specificity of the pertinent levels of structural analysis – have been intensively discussed and carefully worked out by a whole generation of linguists in the decades following the publication of Bloomfield's influential book. A rigorous and comprehensive version of the overall approach was arrived at in Harris' (1951) *Methods in Structural Linguistics*, where the canon of distributionalism is formulated in the most systematic and radical form. Bloomfield's own exposition developed the major ideas with remarkable sensitivity and respect for traditional grammatical analyses, preserving their insights where possible and correcting them where necessary, identifying their appropriate place under the new, systematic perspective. This holds for traditional notions of descriptive grammar like parts of speech, morphological categories like gender, number, tense, etc., but particularly for the important discoveries establishing the historical relationship among different Indo-European languages. In fact, more than one-third of Bloomfield's seminal book deals with topics of comparative and historical linguistics, with different types of language change, and problems of dialect geography, i.e., issues that go beyond purely descriptive analysis. Bloomfield's intimate familiarity with the history of the field, combined with his capacity to fundamentally reconstruct its aims and methodology, explains to a large extent his unique impact.

In order to accommodate the standard grammatical phenomena traditionally treated in linguistic descriptions, Bloomfield relies on the behaviorist notion of habits controlling the speaker's behavior. A habit manifests itself in a pattern of linguistic forms. Thus speakers of English have a habit according to which finite verb forms agree with their subject, as in *the boy walks* vs *the boys walk*. The paradigms of inflectional forms like *walk*, *walks*, *walked*, *walking* involved in these habits can be accounted for by systematic combinations of 'free' and 'bound' morphemes, such as *walk* and *walks*, *walked*, or *walking*. The patterns set up on this basis can then be extended to 'irregular' forms like *sing*, *sings*, *sang*, *sung*, *singing*, where the free form *sing* is combined with an alternation of the vowel instead of the affix *–ed*, yielding the form *sang*, and to even further irregularities like *go*, *goes*, *went*, *gone*, *going*, where *went* substitutes the combination *go–ed*.

In a similar way, the systematic character of habits is supposed to control the patterns of variation in the phonetic realization of linguistic forms. Thus Bloomfield argues e.g., that the masculine and feminine form of French adjectives like *plat* vs *platte*, phonetically (pla) vs (plat) 'flat', *gris* vs *grise*, phonetically (gri) vs (gri:z) 'gray', *frais* vs *fraîche*, phonetically (frɛ) vs (frɛ:š) 'fresh', and many others, are best described by taking the feminine as the fundamental form plus "the simple statement that the masculine form is derived from the feminine by means of a minus feature, namely, loss of the final consonant" (*Language*: p. 217). A second operation must then be assumed to shorten the long vowel in the resulting forms (gri:) or (frɛ:), etc., in order to account for the actual masculine (gri), (frɛ), etc. In other words, an additional vowel shortening is necessary after the previous deletion of the final consonant.

Beyond the Limits of Distributionalism

These apparently minor details of descriptive technology are worth mentioning because of the far-reaching consequences emerging from this type of analysis. The indicated description, according to which the French masculine adjective is to be construed as a regular reduction of its feminine counterpart, is certainly an elegant account of the speakers' behavior, but the assumption of an underlying form plus operations deriving its surface realization clearly stretches the notion of a habit determining linguistic behavior, because the actual behavior must now be taken to depend on preceding operations applying to underlying forms that control the articulatory patterns at best indirectly, if at all. Assuming underlying forms plus operations or rules of realization is obviously an essential step toward the recognition of internal, which means mental, representations and operations, a notion that Bloomfield rigorously wanted to exclude from scientific methodology. But he goes even further in the actually prohibited direction. In his ingenious *Menomini morphophonemics* (1939), an analysis of the sound system of an Algonquian dialect, he set up underlying forms and various operations of realization with by far more intricate conditions on their relative dependency than those required for the simple examples of French adjectives. This dependency, which Bloomfield was very much concerned with, requires an ordering of the operations up to five steps. It is obvious that internal operations of this complexity cannot be accommodated to the notion of stimulus–response habit in any serious sense. What is fascinating here, is first the depth of analytic insight in Bloomfield's analysis, and second his clear decision in case of a conflict between behaviorist methodology and descriptive adequacy. Although he was presumably not aware of the conflict, he clearly discarded the orthodoxy in favor of the insight.

In the same spirit, Bloomfield implicitly argued for an ultimately mentalistic theory of sound change, summarizing the core of his opinion by the slogan "Phonemes change" (*Language*: p. 351). This means in fact that the underlying pattern, rather than the acoustic realization, is relevant for the processes of linguistic change. In a far-sighted conclusion, based on his *Menomini morphophonemics*, he even envisaged the possibility that sound change may consist in the introduction of rules of alternation: "our basic forms do bear some resemblance to those which would be set up for a description of Proto-Algonquian, some of our statements of alternation … approximate the historical development from Proto-Algonquian to present-day Menomini" (1939). Chomsky and Halle (1968) have later developed this surprising and fascinating observation systematically in generative phonology.

Bloomfield had vividly argued for a strictly anti-mentalistic program of linguistic analysis, pointing out "Non-linguists … constantly forget that a speaker is making noise, and credit him, instead, with the possession of impalpable 'ideas'. It remains for the linguist to show, in detail, that the speaker has no 'ideas' and that the noise is sufficient" (*Language or ideas?*: p. 92). But while he complied

with this dogma by dispensing e.g., semantics from the actual linguistic research agenda, he was, surprisingly, not able to exclude mentalistic concepts from the domain of his very central concern – the analysis of linguistic form: underlying forms, derivational rules, precedence order of operations, and even phonemes as the elements that underlying forms consist of, are concepts that go beyond the mechanistic elements to be extracted from noise. This conflict between strict methodological principles on the one side and explanatory insights on the other mirrors one aspect of the above-mentioned tension between Bloomfield and Sapir, in the following sense: Although both scholars agreed on the general principles of linguistics as a scientific enterprise and also on the insights in relevant detail such as the description of Indian languages, the reconstruction of their historical relations, or the nature of sound change, Sapir always insisted on the psychological nature of linguistic facts. This is clear from his programmatic statement *Sound patterns in language* (1925), and even more so from *The psychological reality of phonemes* (1933). As a matter of fact, the orientation of Sapir's inspiring essays amounted to a mentalistic research strategy.

Thus, although Bloomfield in 1940 was Sapir's successor as Sterling Professor of Linguistics at Yale University, it was Sapir's notion of language as a psychological reality that dominated the subsequent period of post-Bloomfieldian linguistics, as shown e.g., by the title and orientation of *The Sound Pattern of Russian* (1959) by Morris Halle and *The Sound Pattern of English* (1962) by Chomsky and Halle.

But the tension persists: While Sapir (1933) and Jakobson (1941) articulated the ideas and the orientation of this development, it is still Bloomfield's radical clarification of the conceptual framework and descriptive technology that provides the indispensable foundation for practically all branches of modern linguistics. And perhaps more significantly, Bloomfield's *Menomini morphophonemics* is the paradigm case not only of a complete descriptive analysis in terms of an explicit, complex rule system, but also of an impressive victory of clear insight over methodological orthodoxy. In line with this creative tradition and conceding the amount of mentalistic machinery implied in this analysis of sound structure, strong theories of syntax and even semantics have in the meanwhile been shown to be possible and productive – in spite of Bloomfield's antimentalistic bias, but keeping to his standards and principles of linguistic analysis.

See also: Generative Grammar.

Bibliography

Bloch, Bernard, 1949. Leonard Bloomfield. Language 25, 87–98.

Bloomfield, Leonard, 1914. An Introduction to the Study of Language. Henry Holt, New York.

Bloomfield, Leonard, 1917. Tagalog Texts with Grammatical Analysis. University of Illinois, Urbana, Illinois.

Bloomfield, Leonard, 1926. A set of postulates for the science of language. Language 2, 153–164.

Bloomfield, Leonard, 1928. A note on sound change. Language 4, 99–100.

Bloomfield, Leonard, 1933. Language. Henry Holt, New York.

Bloomfield, Leonard, 1936. Language or ideas? Language 12, 89–95.

Bloomfield, Leonard, 1939. Menomini morphophonemics. Traveaux du cercle linguistique de Prague 8, 105–115.

Bloomfield, Leonard, 1946. Algonquian. In: Harry, Hojer, et al. (Eds.), Linguistic Structures of Native America. Wenner Gren Foundation, New York, pp. 85–129.

Chomsky, Noam, 1957. Syntactic Structures. Mouton, 's-Gravenhage.

Chomsky, Noam, Halle, Morris, 1968. The Sound Pattern of English. Harper and Row, New York.

Hall, Robert A., 1987. Bloomfield the man. In: Historiographia Linguistica XIV:1/2. Benjamins, Amsterdam, pp. 61–70.

Halle, Morris, 1959. The Sound Pattern of Russian. Mouton, 's-Gravenhage.

Harris, Zellig S., 1951. Methods in Structural Linguistics. University of Chicago Press, Chicago.

Hockett, Charles F., 1970. A Leonard Bloomfield Anthology. Indiana University Press, Bloomington, Indiana.

Jakobson, Roman, 1941. Kindersprache, Apahsie und allgemeine Lautgesetze. In: Jakobson, Selected Writings I. Mouton, 's-Gravenhage, pp. 328–401.

Paul, Hermann, 1880. Prinzipien der Sprachgeschichte. Niemeyer, Halle.

Sapir, Edward, 1925. Sound patterns in language. Language 1, 37–51.

Sapir, Edward, 1933. La Réalité psychologique des phonèmes. Journal de Psychologie Normale et Pathologique 30, 247–265.

Saussure, Ferdinand de, 1916. Cours de linguistique générale. Payot, Paris.

Skinner, Burrhus Frederic, 1957. Verbal Behavior. Appleton-Century-Crofts, New York.

Weiss, Albert Paul, 1924. A Theoretical Basis of Human Behavior. R. Adams, Columbus, Ohio.

Boas, Franz (1858–1942)

Georg Elwert, Free University of Berlin, Berlin, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 2, pp. 1266–1270, © 2001, Elsevier Ltd.

Abstract

Franz Boas was born on 8 July, 1858, in Minden, in the German province of Westphalia. He was the son of freethinkers of Jewish origin who still cherished the ideals of the democratic revolution that had taken place in Germany in 1848. Coming from such an educated bourgeois background (*Bildungsbürgertum*), he had access to books on the natural sciences and geography, and this environment had a profound impact on his future aspirations. From 1877 onwards, he studied natural sciences and geography, first in Heidelberg and then in Bonn and Kiel. After completing his PhD in physics (on the color of water) he went to Berlin, where he began to move in the academic and social circles characteristic of much of the German scientific community at that time. A major role model for him was Rudolf Virchow, the professor of medicine and anthropology.

Biography

Franz Boas was born on 8 July, 1858, in Minden, in the German province of Westphalia. He was the son of freethinkers of Jewish origin who still cherished the ideals of the democratic revolution that had taken place in Germany in 1848. Coming from such an educated bourgeois background (*Bildungsbürgertum*), he had access to books on the natural sciences and geography, and this environment had a profound impact on his future aspirations. From 1877 onwards, he studied natural sciences and geography, first in Heidelberg and then in Bonn and Kiel. After completing his PhD in physics (on the color of water) he went to Berlin, where he began to move in the academic and social circles characteristic of much of the German scientific community at that time. A major role model for him was Rudolf Virchow, the professor of medicine and anthropology.

The most decisive period for Boas was that from 1881 to 1887. In Berlin he frequented the meetings of a learned society, the *Berliner Gesellschaft für Anthropologie, Ethnologie, und Urgeschichte*, and sought to raise funds for research in North America while preparing intellectually for such a trip. In 1883–4, after a short period of service in the German army, he began to conduct field research among the Inuit of Baffin Island. This prompted him to shift his focus of interest from the landscape to its inhabitants. He found that the geographical determinism of his time did not hold up against his observations of the central Inuit, in that some of the behavior he observed was exhibited in spite and not because of the environment. After a stay in New York, he returned to Berlin to become assistant at the Museum of Ethnology in 1885. There he met a group of Bella Coola Indians who had been brought to Berlin as a living exhibit (*Völkerschau*). This encounter prompted his decision to study the American Indian peoples of the northwest coast. In 1886 he embarked on the first of his many research trips to this region, where he gained rich insights into the social organization and self-expression of the Kwakiutl. On his way back he stopped in New York, where he met the editor of *Science* magazine and was immediately appointed assistant editor. Thus, he now had the means to marry Marie Krackowitzer, a woman from New York's liberal German-speaking community, to whom he had been introduced by his uncle. He decided to remain in America, despite the fact that in 1882 he had written that 'his dearest aim' was to become a German professor. Even then, he had known that if the USA offered him 'a better chance ... I would go without hesitation.' A strong motivation for staying was that American science was, as he saw it, 'so far behind in comparison to Europe that a young man as I am, is able to accomplish much more' (Liss, 1996). There was less teaching of anthropology or ethnology than in German universities, but the strong role of the natural sciences was to impact on the development of anthropology as an academic discipline. Boas' understanding of anthropology was that of an empirical science, part of a wider trend in the social sciences, oriented towards the works of Mach, Dewey, and James (Lowie, 1956, p. 1015).

Boas' experiences in the field resulted in a rather rare intellectual approach, that of a commuter who kept travelling between different nations and their intellectual environments. He thus saw different milieus as an outsider and observer, sometimes integrated and engaged, sometimes detached. Over the course of his life he showed a strong sympathy for other such commuters. For example, he greatly encouraged George Hunt, brought up among the Kwakiutl by a Tlingit mother and a British father, to do his own fieldwork among the Kwakiutl, and also provided the funding for this work. Later, too, he was active in support of those moving between different environments, be they immigrants to the USA or people living on the fringes of their home community. Boas himself experienced the advantages of being an outside observer, maintaining that such a status sharpens one's perception of those features of a culture which are self-evident to those living within it. He was very conscious of diversity, variation, and individual differences.

Boas' basic research motivation was one for which no assumption of cultural heritage was needed: curiosity combined with ambition. For Boas, it was a particular thrill to find complex answers to simple-looking phenomena that others considered to be obvious. He stated that he was 'motivated by the effective appeal of a phenomenon that impresses us as a unit, although its elements may be irreducible to a common chance. In other words the problem that attracted me primarily was the intelligent understanding of a complex phenomenon (cf Herskovits, 1953, p. 10).

Boas assumed an entrepreneurial role in the organization of US anthropology when, in 1896, he became assistant curator of the American Museum of National History in New York (curator, 1901–5), and lecturer in physical anthropology at Columbia University (professor of anthropology, 1899–1936). Among other activities, he modernized the learned journal *The American Anthropologist* in 1898, and helped to found the American Anthropological Association in 1902. He was not only interested in founding institutions (indeed, others tried to do this as well), but in insuring their professionalization. It was this specific approach (which also brought him into conflict with those who had hoped for a broader amateur membership of the AAA), which resulted in the lasting success of his institutional activities. It should not be forgotten that the period between the Civil War and the 1920s was a great era for entrepreneurs in and around the sciences. Many institutions of higher learning, new disciplines, scientific journals, and learned societies were founded, and funding raised, but most of these efforts failed in the long run. Today, it is difficult to appreciate that Boas' efforts were unusually successful, as we only notice those institutions which have survived.

Contributions to Anthropology

Boas' major and lasting contribution lies in the methodological domain. He was always on the look out for differences and, as such, was less interested in the contemporary transformations of Kwakiutl Indian culture than in what he perceived to be its remembered past. His quest for difference also made him sensitive to variations within the ensemble of aspects which the informants or outsiders described as a culture. Any boundaries and standards were suspect to him. He raised certain issues 90 years before they were advocated by the postmodernists of the 1980s (Jacknis 1996). He was particularly skeptical about classification, and naive subjective classification aroused his anger (see for e.g., 'The principles of ethnographic classification'). A systematic discussion of the problem can be found in his linguistic paper on 'Alternating sounds'. Boas contended that people tend to classify according to familiar categories, but that anthropologists should be concerned with what was later called emic distinctions – the distinctions made by the people themselves.

Boas' empiricism shielded his work from the speculative thinking which dominated the discipline in the German-speaking countries. This methodological rigor contributed to the strong and lasting reputation acquired by American anthropology, particularly in comparison to other social and natural sciences. It was manifest in the way Boas and many of his pupils discussed and validated the results of their fieldwork. More than any other author of his time, he was conscious of two influences on anthropological sources: that of the method chosen and that of the observers' perspective, including artifacts created by the methods. He gave preference to texts emanating from the culture itself, and asked his interlocutors to produce and record texts of all kinds, not only ceremonial ones. Moreover, he carefully recorded the effects of his methods, e.g., how the phonograph was used, and how different informants influenced the style and content of dictated or written texts. The aim of this never-ending endeavor was to achieve objectivity. For Boas, a command of the native languages was essential in order to 'understand directly, what the people ... speak about, what they think, and what they do' (Boas, 1906, pp. 183–8). His concept of objectivity introduced a clear frame of reference, distinguishing his approach from that of his German and American contemporaries – a frame of reference rooted within the culture itself.

The results of Boas' fieldwork may suggest to many readers that he spent prolonged periods within the cultures under investigation, participating intensively in their everyday activities. The reality was different. His most important results were yielded by repeated short stays in the field – up to three weeks in one place. As the letters to his family show, he profited from the contrasts he experienced, seeing the surprising and the colorful in what was self-evident to the local population. The value of his reports is rooted in his insistence that each phenomenon should be viewed from different perspectives, and evidence gathered from different informants. He also integrated the latest technological methods into his approach – photography as early as 1883, and sound recording in 1887.

One of Boas' research topics was largely disregarded by the subsequent generation, namely, the place of the individual within society. Boas' interest in this subject may be seen as a legacy of the liberal German milieu in which he grew up. It was in front of the German Society of New York that he stated in 1880: 'The invention is not difficult. Difficult is the retention and further development ... . It is important to observe the fight of individuals against tribal customs. The same kind of struggle the genius has to undergo ... in his battle against dominant ideas' (cf Herskovitz, 1953, p. 75). He continued to address this question later on, in his interpretation of esoteric cults, insisting that cults should be studied in connection with the broader base of belief in society as a whole. It would be an error, he argued, to see esoteric beliefs 'as the only true form of the inner life. They might as well be considered as a reaction ... to the general cultural environment.'

The anthropological subject most intimately linked to Boas' name, and the one which prompted most references to his work in the second half of the twentieth century, was the *potlatch*. 'Potlatch' is a Chinook word for festivities, which Boas used to describe a group of redistributive ceremonies performed by the Kwakiutl. Redistribution was not the dominant feature in all cases, however. As he emphasized in his very first discussion of the subject in 1899, the creation of credit and obligation was also a central element. Moreover, the redistribution of wealth created prestige as well as obligation.

To judge from the attention he gave to the potlatch, Boas was clearly fascinated. His curiosity about otherness was captured by a phenomenon which, with historical distance, can be seen to be a travesty of one particular feature of (non-native) American civil life – too conspicuous consumption. (Both private universities and museums were dependent on such redistributions of wealth, and on the fact that benefactors were, in turn, endowed with prestige.) The Canadian administration did not recognize society's need to have ways of creating prestige. Nor did it understand that the indebtedness created by such ceremonies was, in fact, a creation of wealth and capital not entirely different from that of the modern banking system, in which the capital mobilized as credit by far outweighs the

liquid capital deposited. Boas brought these arguments into the foreground in order to defend the Kwakiutl and neighboring groups against intrusion by the Canadian administration.

Later critics observed that the potlatch, as portrayed in Boas' works, was just one momentary impression in a rapidly changing situation (Mauzé, 1986, pp. 21–63), and that Boas neglected other aspects of the economy, mainly land tenure and the livelihood of classes other than the nobility. However, these criticisms cannot diminish the importance of his discovery of a system of attributing merit which was, in itself, economically organized.

Boas was also skeptical about the evolutionism of his times, insofar as it implied a ranking of more or less progressive cultures. How was one to classify a culture which had a seemingly simple technology, but an extremely intricate social structure? For Boas, it was not acceptable to label these as 'early forms' of humanity. Assumptions about the coincidence of linguistic communities, biological inheritance, and culture were implicit in evolutionism and explicit in historical reconstructions of the 'genealogies of cultures.' Yet using material gathered from American Indians, Boas was able to show that linguistic communities, groups of descent, and culture do not necessarily coincide, and that they have in fact varied independently of each other in many cases. Thus the evolution of a given language or language group may run across different groups of descent which, over the course of time, may become members of different linguistic families.

In the first half of the twentieth century, Boas' contribution to science was most strongly identified with his work on physical anthropology. Here, he introduced the perspective of regarding differences in the human races as a phenomenon of domestication, rather than in analogy to differences in animal species. He maintained that nutrition, the instrumental and ornamental use ofthe body, crossing, and selection are the social facts shaping the variant physical appearance of humankind. Of his many empirical investigations, a study published in *The American Anthropologist* in 1912 is particularly notable. In this study, based upon a huge sample, he compared the head shapes of immigrants and their children. The shape of the head had, until then, been considered the best indicator of hereditary lines. Yet Boas was able to show that these measurements changed considerably in the children's generation. This challenged the dogma that the physical type was determined by heredity alone.

Boas did as much as he could to collect texts, especially songs, tales, and myths, and to encourage others to do the same. He maintained that such materials have the merit of bringing out 'those points which are of interest to the people themselves' (Boas 1916, p. 393). He objected to other authors classifying genres according to their subjective standpoint or the standpoint prevalent in their own culture, and referred instead to the categories used by the people themselves (emic categories) – a procedure which, even today, is not always followed. He rejected explanations of symbolism which were not based on explicit data from the culture itself as pure speculation, whether they be such as those of the German *Kulturhistorische Schule*, or Freudian. In contrast, he placed particular weight on the recognition of myths and tales as artistic creations, arguing that these should not only be a source of anthropological research, but valued as a product of human creativity.

Boas' attitude to politics was interested but rather detached. The liberal German milieu in which he grew up advocated the separation of science from religion and politics. The sociologist Max Weber, who later matured in the same milieu, formulated the idea of value freedom in science. Having experienced the repression of such scientific freedom by the institutions of power and religion in the nineteenth century, the liberal scientific community saw this conception of science as an important basis for the growth of scientific endeavor and discourse – a neutral medium for the way society saw itself and for societal development. Detachment did not mean indifference, however. Boas had already described the goal of science to be 'the ice-cold flame of the passion for seeking the truth.' Moreover, observations of prejudice called for scientists to speak out. On several occasions, Boas spoke out against racial prejudice, defending the American 'Negro,' underlining the evidence of high cultural achievement in the African past, criticizing the US and Canadian policies which discriminated against Indian customs (namely the potlatch), censuring war and, last but not least, condemning anti-Semitism. His texts against anti-Semitism were printed in a widely read German newspaper (*Frankfurter Zeitung*) in 1926 and later distributed as leaflets by the German anti-Nazi resistance.

Influences

Boas' legacy cannot be identified with a particular theory. In contrast to other authors, after his investigation of the Inuit he produced few texts which attempted to provide a general ethnography. The organization of findings in his texts often obscures information rather than elucidating it. Central theoretical arguments are typically found neither in the introduction nor in the summary, but somewhere in the middle of the text, sometimes as footnotes. He did, however, introduce some important methodological approaches. It was Boas who cross-bred physical anthropology with ambitious statistical methods. He and his pupils were the first to take linguistic approaches, looking for the categories by which language was organized in the mind. This later lead to the study of phonemics with his pupil Edward Sapir, and later still, to the idea of the categorial grammar which takes Boas' perceptive of trying to find by induction categories immanent to the language itself, rather than transplanting the categories of Indo-German grammars.

In the decade after his death, more was written about Boas than about any American anthropologist before him. This may have created the impression of a Boas School. Indeed, his pupils included such anthropologists as Ruth Benedict, Melville Herskovits, Robert Lowey, Margaret Mead, Ashley Montague, and Edward Sapir. Later, however, harsh criticisms such as those of Lesley White exposed some limitations in Boas' work. His didactical methods apparently did not meet the standards of the North American universities. Moreover, considering that those who studied under him represented 'the greatest variety of interests, methods and theoretical orientation,' one cannot but agree with Herskovits' early statement (1953, p. 23) that 'the term "Boas' school" ... is a misnomer.'

However, one cannot overlook the fact that functionalism and its later variations such as conflict theory owe their success to the methodological rigor introduced in Boas' aversion to speculative anthropology, and to the institutional foundations he laid. Evidently, Boas and Malinowski – the leading Anglo-Saxon anthropologist of the time – tended to ignore each other, although both were trained in the German approach to the natural sciences. In the few remarks he made on Malinowski and functionalism, Boas named the close observation of daily life as a merit, though underlining that fieldwork was already an established practice in US anthropology before the beginning of the twentieth century. He criticized the explanation of human behavior on the basis of the actual social environment alone, the approach taken in earlier British functionalism. Boas felt reference to historical data to be indispensable for an understanding of ongoing processes (cf Boas 1938).

It was thanks to Boas that exotic presentations of unexplainable behavior and the reduction of unfamiliar cultural forms to an irreducible genetical heritage were gradually excluded as forms of scientific explanation. He brought the assumption of rationality to the foreground, making this the starting point of anthropological explanation. This does not imply that all human behavior is rational, or that all social institutions are functional (in fact, Boas abhorred such generalizations). However, it does mean that explanations of individual behavior and culturally encoded behavior should relate to the goals, cultural strategies, and environmental conditions of the informants. Of course, behavior may still fail to achieve its goals or may even be erratic in view of all possible goals.

The 1980s and 1990s saw a return to the skepticism which distinguished Boas' approaches from those of most earlier theorists. His legacy is an antidote to reification (*Verdinglichung*), reminding us that any culture is multifaceted and continuously changing, and that individuals may row against its currents.

See also: Anthropology, History of; Diffusion: Anthropological Aspects; Diffusion: Geographical Aspects; Ethnography; Evolution: Diffusion of Innovations; Exchange in Anthropology; Exchange: Social; Potlatch in Anthropology.

Bibliography

Boas, F., 1885. Baffin-Land: geographische Ergebnisse einer in den Jahren 1883 und 1884 ausgeführten Forschungsreise. Petermanns Mitteilungen 17, Supplement 80. Perthes, Gotha, Germany.

Boas, F., 1888. The Central Eskimo, 6th Annual Report, US Bureau of American Ethnology, Washington, DC, pp. 399–669.

Boas, F., 1906. Some philological aspects of anthropological research. In: Stocking, G.W. (Ed.), 1974 A Franz Boas Reader: The Shaping of American Anthropology 1883–1911. University of Chicago Press, Chicago, pp. 267–281.

Boas, F., 1909. The Kwakiutl of Vancouver Island. Yesup North Pacific Expedition. Brill, Leiden, Germany. Vol. 5, pt. 2.

Boas, F., 1911. The Mind of Primitive Man. Macmillan, New York.

Boas, F., 1912. Changes in bodily form of descendants of immigrants. American Anthropologist 14, 530–562.

Boas, F., 1916. Tsimshian Mythology, 31st Annual Report, 1909–1910, US Bureau of American Ethnology, Washington, DC.

Boas, F. (Ed.), 1938. General Anthropology. Heath, Boston and New York.

Boas, F., 1940. Race, Language and Culture. Macmillan, New York.

Herskovits, M.J., 1953. Franz Boas: The Science of Man in the Making. Scribner, New York.

Lowie, R., 1956. Reminiscences of anthropological currents in America half a century ago. American Anthropologist, 995–1015.

Mauzé, M., 1986. Boas, Les kwagul et le potlatch. L'homme 100: 21–63 [including comments by Claude Meillassoux and other authors].

Ryding, J., 1975. Alternatives in 19th century German ethnology: A case-study in the sociology of science. Sociologus 25, 1, 1–28.

Stocking, G.W. (Ed.), 1974. A Franz Boas Reader: The Shaping of American Anthropology 1883–1911. University of Chicago Press, Chicago.

Stocking jr., G.W., 1996. In: Volksgeist as Method and Ethic. Essays on Boasian Ethnography and the German Anthropological Tradition. University of Wisconsin Press, Madison, WI.

Bobbio, Norberto (1909–2004)

Richard Bellamy, School of Public Policy, University College London, London, UK; and European University Institute, Florence, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Norberto Bobbio was the leading Italian legal and political philosopher of the second half of the twentieth century. His life and work were conditioned by the vicissitudes of Italian democracy both before and after the Second World War. The experience of Fascism, the ideological divisions of the Cold War, and the transformation of Italian society during the 1960s and the 1970s, which he described so evocatively in his *Ideological Profile Of Italy in the Twentieth Century* (1969a, 1995a), prompted and enriched his passionate defense of the constitutional 'rules of the game' against those who denied their relevance or would overturn them for reasons of pragmatic convenience.

Biography

Norberto Bobbio was born in 1909 into an affluent, professional, upper-middleclass Turin family. He attended the local university, where he took degrees in both Jurisprudence and Philosophy. His main influence was as a legal and political philosopher. He brought to both a keen civic as well as intellectual engagement, while never sacrificing academic rigor or independence. His academic significance in both fields lay in combining conceptual analysis with a positivist orientation that focused attention on the institutional sources and embodiment of law, on the one hand, and the social and procedural preconditions of democratic politics, on the other. As a public intellectual, he was a lifelong proponent of social democracy. He played a key role in seeking to move the Italian Communist Party (PCI) toward a principled rather than purely pragmatic acceptance of the norms and forms of liberal democracy and in the promotion of a centre Left capable of taking on the centre Right coalition that ruled Italy for much of the postwar period. He was also a prominent campaigner for nuclear disarmament and a pioneer in the field of international political thought.

Bobbio taught jurisprudence first at the University of Camerino and then at Sienna and was appointed to a Chair at Padua in 1940. In 1948, he replaced his teacher Gioele Solari as Professor of Legal Philosophy in Turin, a post he held until 1972. He then took the chair in the newly created Faculty of Political Science in Turin, where he remained until the then statutory age of retirement of 75 in 1984. He was a member of the Accademia dei Lincei, a Corresponding Fellow of the British Academy from 1966, and longtime editor (together with the philosopher Nicola Abbagnano) of the *Rivista di Filosofia*. Bobbio received, among other awards and honors, the Balzan Prize in 1994 (for Law and Political Science: government and democracy) and diplomas *honoris causa* from the Universities of Paris (Nanterre), Madrid (Complutense), Bologna, Chambéry, Madrid (Carlos III), Sassari, Camerino, Madrid (Autónoma), and Buenos Aires.

Bobbio played an important intellectual role within Italian political life and commentated regularly on Italian politics for the Turin daily *La Stampa*. In 1942, he joined the then illegal radical liberal party, the Partito d'Azione ('Party of Action'), and was briefly imprisoned in 1943 and 1944. He ran unsuccessfully in the elections for the Constituent Assembly of Italy in 1946. Throughout the 1950s, the 1960s, and the 1970s, he was a powerful advocate of liberal democratic values against both the Marxists in the PCI and the New Left and extremist groups outside it, while also condemning the corrupt electoral practices of successive governments to exclude the first and their often illiberal measures to tackle the second. The year of his retirement, Bobbio was nominated by President Sandro Pertini to one of the five life senatorships, and sat in the upper house as an independent socialist. In 1992, he came close to being elected President as a compromise candidate. However, he confessed to finding decision making difficult – his talent was always for spotting problems rather than solutions – and he was relieved the bid failed. He died as he had lived his life, with great dignity – instructing his doctors not to intervene when taken into hospital just after Christmas 2003. He died on 9 January 2004, his status marked by the Italian President's immediate departure for Turin to be among the first mourners.

Major Works and Publications

Bobbio grew up in Turin in a well-off bourgeois family. He has characterized his parents' sympathies as 'filo-fascist,' regarding Fascism as a necessary evil to guard against the supposedly greater danger posed by Bolshevism – a threat that seemed all too present following the occupation of the Turin Fiat works in 1921–22 by the *Ordine Nuovo* group gathered around the Marxist theorist and future leader of the Communist Party of Italy, Antonio Gramsci (Bobbio, 2002: pp. 15–16). At school and university, however, he became acquainted with many of the leading lights of the largely antifascist Turin intelligentsia. These included the novelists Cesare Pavese and Carlo Levi, his future publisher Giulio Einaudi, the critic Leone Ginsburg, and the radical politician Vittorio Foa – the last two being particularly close friends at the time.

Bobbio never met the two best-known martyrs of Turin's antifascist movement – Gramsci and the 'revolutionary liberal' Piero Gobetti, though he became Director of the center dedicated to the latter's memory and his own papers have now been housed there along with Gobetti's. However, some of his

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61164-9

teachers and friends had been involved with them and contributed to their journals (Bobbio, 2002: pp. 17–21). Despite associating with certain key opponents of the regime, a number of whom were imprisoned, he largely kept his head down. Indeed, there was something of a furor when in 1992 the Italian weekly *Panorama* published a somewhat groveling letter he had written to Mussolini in 1935. Earlier that year Bobbio had been arrested and detained for 7 days on suspicion of antifascist activities and sympathies. The resulting record had led to his exclusion from the promotion process for a permanent professorship. The letter gave examples of his devotion to the Fascist cause in an attempt to get a police caution removed – something later achieved when a family friend within the Fascist hierarchy also intervened on his behalf. Bobbio confessed that the letter had brought him face to face "with another self who I thought I had defeated forever" (Bobbio, 2002: p. 31). His defenders, though, among them the surviving friends he had apparently disavowed, pointed out that the letter had been written at the height of Fascism's power simply to keep his job. It had revealed not only that Bobbio clearly was not considered a dutiful Fascist but also the debasement the authorities imposed on all their subjects.

In his autobiography (1997), Bobbio described this first half of his life as belonging to his prehistory. Yet one can see Bobbio's own deep commitment to democracy and the rule of law as flowing from this humiliating experience. Far from detracting from his opposition to authoritarianism in all its forms, the incident – that encapsulates life under a dictatorship – makes it all the more comprehensible and profound. The fall of Mussolini on 8 September 1943 catapulted him along with so many others of his generation from a state of total exclusion from political life into active involvement with it. Bobbio had been close to the 'liberal socialist' circle of intellectuals since the late 1930s. This group became part of the Party of Action, the main noncommunist resistance movement. Bobbio played a minor role, but did engage in some clandestine activity against the German occupation and was briefly imprisoned from 1943 to 1944. Although intellectually influential, the *azionisti* lacked a popular base. A candidate in the 1946 elections for the Constituent Assembly, he was not elected and returned to academic life. However, the party's slogan 'Justice and Liberty' captures the central theme of much of Bobbio's subsequent work – how to unite the liberties beloved of liberals with the socialist demand for social and economic justice. It was his commitment to these twin ideals that was to give him such an influential place within Italian political life, rendering him the perfect critical interlocutor between the Communist Party, on the one hand, and the various governmental parties gathered around the Christian Democrats, on the other.

Bobbio's first book, *The Phenomenological Turn in Social and Legal Philosophy* of 1934, was followed in 1938 by a monograph on *The Use of Analogy in Legal Logic* and in 1944 by a critical study of existentialism, *The Philosophy of Decadence* – the first of his books to be translated into English (Bobbio, 1948a). Over this period, Bobbio gradually dissociated himself from the broadly idealist approach to philosophy then dominant in Italian universities. He was friendly with the philosopher of science, Ludovico Geymonat – also based at Turin – and together with him helped establish the interdisciplinary Centre for Methodological Studies. Bobbio now set himself the task of elaborating a general theory of the practice and validity of law, breaking with the attempts of most contemporary Italian philosophers to offer a speculative philosophy of the idea and morality of law. In elaborating his version of legal positivism, he drew on the writings of Hans Kelsen, whose work he had come across as early as 1932. This research ultimately bore fruit in a number of books based on his Turin lectures, of which the most important are *A Theory of Judicial Norms* (1958) and *A Theory of the Legal Order* (1960) – both still in print – and studies of Locke, Kant, and legal positivism. Between 1955 and 1970, he also published three collections of his most important essays. These writings had a similar place in Italian academic legal circles as the work of H.L.A. Hart in British universities, and both men held each other in high esteem.

Bobbio's legal studies fed into his political writings. Influenced again by Kelsen, he adopted a procedural view of democracy as consisting of certain minimal 'rules of the game,' such as regular elections, free competition between parties, equal votes, and majority rule. His theory of politics was additionally enriched by a strong realist current deriving partly from Hobbes, and partly from his studies of the Italian pioneers of political science, such as Gaetano Mosca and Vilfredo Pareto (collected as 1969b, 1972), whose reputation Bobbio did much to resurrect. Bobbio had produced the first Italian edition of Hobbes' *De Cive* (1948b), earning the plaudits of the German scholar Carl Schmitt, who he had briefly met in Berlin in 1937. Bobbio later dedicated numerous studies to the English philosopher, a collection of which were published in 1989 and appeared in English a couple of years later (Bobbio, 1989b, 1993a). He drew on Hobbes to modify what he now saw as unsatisfactory elements of his earlier Kelsenism. Bobbio regarded Kelsen as caught uncomfortably between a purely formal account of law, on the one hand, and a substantive position grounded in what he called the 'basic norm' underlying all law, on the other. The missing dimension was the institutional context of law making and its relationship to the exercise of power. Unlike earlier legal positivists, such as John Austin, Bobbio did not thereby equate law with the commands of the sovereign. His point was rather that law and rights were best conceived as a historical achievement belonging to a particular form of constitutional and democratic state.

Bobbio's shift from a pure theory of law to a concern with its political embodiment was marked by his moving to the newly created chair in Politics at Turin in 1972. The essays from this period were later collected as *The Future of Democracy: A Defense of the Rules of the Game* (1984a, 1987a), perhaps the most original of his books; *State, Government and Society* (1985, and published in English as *Democracy and Dictatorship* 1989d); and *The Age of Rights* (1990, 1996a). Bobbio's linking of the rule of law and rights to the distribution of power produced by liberal democracy informs his contributions to the political debates of the period. His prime concerns from the 1950s onward were to enter into dialogue with the Communist Party and build a social democratic opposition in Italy. Indeed, the latter could only be achieved if the Communists, the largest grouping on the left in Italy, could be weaned away from the Soviet Union and converted to liberalism. It is no accident that he published the first (and for

some years the only) Italian study of Karl Popper's *The Open Society and its Enemies* as early as 1946, in a journal appropriately entitled *Il Ponte* (The Bridge). Bobbio was a founding member of the European Society of Culture, which had this critical dialogue as a goal. His first book of political essays, *Politics and Culture* (1955), consisted largely of a debate with the Marxist philosopher Galvano Della Volpe over whether socialist legality could be based on anything other than traditional liberal rights – a discussion that ultimately prompted the intervention of Palmiro Togliatti, the leader of the PCI.

This theme resurfaced in Bobbio's next major foray into politics in the 1970s. The spur this time came from the 'historic compromise' between the PCI and the Christian Democrats, whereby the PCI, which had a strong foothold in local government and was the main opposition party, was given access to positions in the state while being denied participation in central government. PCI leaders were worried that terrorist violence, some of it certainly promoted by the security services, might be used as an excuse for a right-wing coup should they appear to be too strong – a fear reinforced by the fate of Allende in Chile. They sought to make themselves nonthreatening to the status quo, therefore, while strengthening their position within the Italian political system – a tactic they associated with Gramsci. Bobbio's interventions were to challenge the coherence of this Euro-communist strategy of a 'third' way between liberalism and Soviet communism. In a series of essays, he criticized Marxism for lacking a theory of the state or democracy, and implicitly urged the PCI to become a social democratic party. These essays were published as *Which Socialism?* in 1976 with an English edition in 1987b. It should be emphasized, however, that Bobbio was as harsh a critic of the corruption of Italian politics and of the role of the noncommunist Socialist Party under Bettino Craxi in upholding that system and taking it to new depths. It was entirely in keeping with his stance that his first article following the collapse of the former Soviet bloc should not be a piece of liberal triumphalism but a reminder that the cause of social justice that had inspired communism remained as pressing as ever, and that liberals could not afford to ignore it (1989e,f). He later reiterated this thesis in his long essay *Left and Right* (1995b, with an English edition published as 1996b), which entered the Italian bestseller lists, in which he argued that the search for a reconciliation between the claims of liberty and equality still provided the key issue of modern politics and the main dividing line between political parties.

The other political issue associated with Bobbio was the peace movement. This too bears a direct relation to his academic work. His view of the political character of law led him to recognize the need for a political theory of international relations. In a series of pathbreaking essays, he explored the possibility for global forms of democracy to give meaning to international law. He was a passionate critic of nuclear weapons, which he saw as making war intrinsically unjust, and a member of the Bertrand Russell Foundation. His writings on this issue were collected in the volumes *The Problem of War and the Roads to Peace* (1979) and *The Absent Third* (1989a). He was not a pacifist, though many of his supporters were surprised when he supported the Gulf War – a position he defended in his book *A Just War?* (1991).

An esteemed political commentator, who wrote regularly for the Turin-based daily *La Stampa*, he kept aloof from a direct involvement in party politics and refused invitations to stand as a senator. He had a healthy disdain for academic politics and, apart from a stint as Dean from 1973 to 1976, kept clear from positions of power and deplored the way so many Italian professors built up extensive patronage networks to secure posts for their pupils. He took his teaching duties extremely seriously, and sympathized with that element of the 1960s' student movement (of which his eldest son was one of the leaders) that complained about the large numbers of Italian academics who engaged in extracurricular activities to the detriment of their university responsibilities. Unbelievably prolific, many of his books had their origins in his lectures and he put on new courses for almost every session. He was deeply disappointed by the failure of the centre Left to establish its hold on Italian politics and became an outspoken critic of Berlusconi, lamenting in a book-length interview on *The Idea of the Republic* (Bobbio and Viroli, 2001, 2003), how the second republic appeared to lack any of the idealism of the first. Significantly, in a series of essays *In Praise of Meekness* (1994, 2000), he turned his attentions to the nonpolitical virtues and the issue of how to respond to evil in a corrupt world. His last works included a subtle study of the meaning of old age in contemporary societies, *De senectute* (1996c, 2001).

The Paradoxes and Broken Promises of Democracy

As the foregoing discussion shows, Bobbio was a prolific author who wrote on a wide range of topics in legal and political philosophy. He was not a systematic thinker, tending to react to the views of others and the events of the day. His forte was to ask pertinent questions rather than offering answers. Nevertheless, a core concern runs through his writings: namely, the desire to link liberalism, socialism, and democracy and to show how each entailed the other. He firmly believed that there could be no democracy without both liberal rights and social justice, yet he also contended that these were grounded in their turn in a democratic division of power in which citizens were treated as political equals. In common with the Italian political tradition more generally (Bellamy, 2014), his idealism was grounded in and constrained, and occasionally subverted, by realist considerations concerning the nature and limits of politics. For example, he saw rights as a historic and political achievement rather than as universal principles that lie outside or provide the preconditions for politics.

As a result, tensions always exist in Bobbio's thought between the various ideals he espouses and their realization, with the mutually entailing notions of liberalism, socialism, and democracy being each presupposed by while at the same time in potential conflict with the others. A good example of this feature of Bobbio's thinking in his famous account of the 'paradoxes' and 'broken promises' of democracy. Bobbio advocates a minimal definition of democracy that consists of the following basic 'rules of the game':

1. That all citizens who have reached the age of majority, regardless of sex, race, creed, or economic condition, possess political rights and can vote on collective issues or elect someone to do so for them.

2. That everyone's vote has equal weight, counting for only one.
3. That all citizens can vote according to their own freely-arrived-at opinion, that is in a free competition between rival political groups, which vie with each other to aggregate demands and transform them into collective decisions.
4. That they have a free choice in the sense of having real alternatives to pick from.
5. That they are bound by the majority decision (whether relative, absolute, or qualified).
6. That no majority decision can limit the rights of the minority to become in their turn, and on an equal basis, the majority (Bobbio, 1987b: p. 66).

Bobbio gives three reasons for preferring democratic government to other forms. The first, ethical defense, derives from Rousseau's formula that liberty consists in obeying laws that we have prescribed to ourselves – a condition most nearly approximated by democracy. The second, political justification, regards it as the best available protection against the abuse of power, since it shares sovereignty among the people as a whole. Finally, from a utilitarian standpoint, democracy is preferable to autocracy on the grounds that the people are the best interpreters of their own collective interests. This point, as Bobbio notes, is the most debatable, since the collective interest is not identical with the sum of individual interests. However, provided one assumes that everyone has agreed to cooperate to find a mutually beneficial solution, and that each person has a better-than-even chance of being right, then (as Condorcet proved) the majority will be more often right than a single voter (Barry, 1967: pp. 112–126).

The above arguments have a distinctly Rousseauvean tenor, and hence might seem to provide a prima facie case for preferring 'direct' forms of democracy to 'indirect' or representative models. For the latter, as practiced in most Western countries, reduces the autonomy and sovereignty of the people by delegating large areas of decision making and encourages self-interested voting for different factional groups, rather than a disinterested adjudication on the common good. Indeed, Rousseau himself gave the most comprehensive account of why democracy as he understood it could not work without modification within modern societies:

> How many conditions that are difficult to unite does [a democratic republic] pre-suppose! First, a very small state, where the people can readily be got together and where each citizen can with ease know all the rest: second, great simplicity of manners, to prevent business from multiplying and raising thorny problems [which empowers those 'in a position to expedite affairs']; next, a larger measure of equality in rank and fortune, without which equality of rights and authority cannot long subsist; last, little or no luxury – for luxury either comes of riches or makes them necessary; it corrupts at once rich and poor, the rich by possession and the poor by covetousness; it sells the country to softness and vanity, and takes away from the state all its citizens, to make them slaves one to another, and one and all to public opinion.
>
> Rousseau, 1973: p. 217

For Bobbio these discrepancies are indicative of the four main 'paradoxes of democracy' facing us today. The first paradox arises from the need for ever more democracy in conditions that are increasingly inimical to its functioning. The 'iron law of oligarchy' operates the larger the state becomes, so that direct democracy degenerates into the rubber-stamping of executive decisions. The second paradox derives from the growth not only in the size of the state but also of its functions. This process leads to the development of bureaucratic structures in which power is organized hierarchically, descending from above, rather than democratically, ascending from below. Even worse, this phenomenon increases in step with the expansion of democracy itself, since the latter produces ever greater demands on the state, which in turn requires ever newer mechanisms to satisfy them. The third paradox refers to the effects of technology on industrial societies, which means more and more decisions call for specialized knowledge to solve them, so that democracy is subverted by technocracy (quite the reverse of what Marx and Lenin supposed). The final paradox belongs to the contradiction between democracy and mass society. Although mass involvement potentially means more democracy, the pressures toward social conformity produced by the former undermine the sense of individual responsibility presupposed by the latter. For example, the growth of information technology potentially gives greater access to the workings of government and could facilitate a more active and informed participation of citizens in making decisions previously left to politicians or administrators. However, it has also given the state, through the storage of private files and the use of the media, more persuasive and subtle ways of manipulating arid spying on us (Bobbio, 1987a,b: p. 72, pp. 96–97).

All the above paradoxes reflect the double-edged nature of the processes at work in contemporary society. Summarizing these effects, Bobbio remarks how 'state' and 'civil society' have become increasingly intertwined. While the state has been 'socialized' through the influence of greater democratic control, democracy in turn has induced increasing intervention by the state in society:

> These two processes are represented by the two aspects of citizenship, that of the citizen qua participant and that of the citizen qua state protected subject, which often conflict with each other within the same person. For through participation the active citizen calls for greater protection from the state and thereby reinforces the very state which he or she wishes to control.
>
> Bobbio 1985: pp. 41–42.

State and society are thus two separate but interdependent moments of the modern social and political system. 'Direct' democracy, in so far as it aspires to absorb state functions within society through schemes for self-administration through collective decision making, falls victim to the dialectic of modern politics. Only representative democracy, albeit in modified form, can meet this challenge. Yet a representative system will always fall short of the radical expectations of democracy and is itself subject to all four paradoxes: being an elite and, via its association with parties, a bureaucratic system that is often itself overly dependent on technocrats in the state administration and business, and both an exploiter of and often at the mercy of the mass media. As such, democracy is destined to always 'break its promises.' It

provides an ideal to which one is moved to aspire, but in reality always disappoints. Real democratic systems will always fail to deliver on the democratic ideal. Indeed, he came to regard Silvio Berlusconi and Italian democracy post 1989 more generally as epitomizing these paradoxes and broken promises of real democratic systems.

Conclusion

Norberto Bobbio was the model of the engaged intellectual. He was deeply involved in the politics and culture of his time yet saw his role as much as an analyst as an advocate. Bobbio's thought has been criticized as mired in contradictions, not least due to his attempt to reconcile liberalism and socialism (e.g., Anderson, 1988). Yet in many respects, his awareness of these tensions was a source of his distinctiveness and strength as a political theorist. He never looked for easy solutions or sought to sacrifice principle for a pragmatic advantage. Rather, he insisted that we should strive to uphold certain basic ideals, while being ever aware of the difficulties standing in the way of their realization. He believed the good should not be sacrificed either for the sake of an elusive and unreachable best or for something worse than is achievable. His absence from contemporary debates is a major loss not only to political thought but also to Italian politics, where he served as his country's political conscience.

See also: Democratic Theory; Legal Positivism; Liberalism: Political Doctrine and Impact on Social Science; Marxism–Leninism: The Ideology of Twentieth-Century Communism; Philosophy of Law; Positivism, History of; Rule of Law (and Rechtsstaat); Social Justice: Historical and Theoretical Considerations; Socialism.

Bibliography

Anderson, P., July–August 1988. The affinities of Norberto Bobbio. New Left Review I (170), 3–36.
Barry, B., 1967. The public interest. In: Quinton, A. (Ed.), Political Philosophy. Oxford University Press, Oxford, pp. 112–126.
Bellamy, R., 2014. Croce, Gramsci, Bobbio and the Italian Political Tradition. ECPR Press, Colchester.
Bellamy, R., 1987. Modern Italian Social Theory: Ideology and Politics from Pareto to the Present. Polity Press, Cambridge.
Bobbio, N., 1934. L'indirizzo fenomenologico nella filosofia sociale e giuridica. Istituto giuridico della R. Università, Turin.
Bobbio, N., 1938. L'analogia nella logica del diritto. Law Institute of the University of Turin, Turin.
Bobbio, N., 1944. La filosofia del decadentismo. Chiantore, Turin.
Bobbio, N., 1948a. The Philosophy of Decadenceism: A Study in Existentialism (D. Moore, Trans.). Blackwell, Oxford.
Bobbio, N., 1948b. In: Bobbio, Norberto (Ed.), Preface to T. Hobbes, Elementi filosofici sul cittadino. UTET, Turin.
Bobbio, N., 1955. Politica e cultura. Einaudi, Turin.
Bobbio, N., 1958. Teoria della norma giuridica. Giappichelli, Turin.
Bobbio, N., 1959. In: Bobbio, N. (Ed.), Introduction to T. Hobbes, Opere politiche, vol. 1. Utet, Torino.
Bobbio, N., 1960. Teoria dell'ordinamento Giuridico. Giappichelli, Turin.
Bobbio, N., 1969a. Profilo ideologico del Novecento. In: Cecchi, E., Sapegno, N. (Eds.), Storia della letteratura italiana, vol. 9. Garzanti, Milan, pp. 119–228.
Bobbio, N., 1969b. Saggi sulla scienza politica in Italia. Laterza, Bari.
Bobbio, N., 1972. On Mosca and Pareto. Librairie Droz, Genève.
Bobbio, N., 1976. Quale socialismo: Discussione di un'alternativa. Einaudi, Turin.
Bobbio, N., 1979. Il problema della Guerra e le vie della pace. Il Mulino, Bologna.
Bobbio, N., 1984a. Il futuro della democrazia: Una difesa delle regole del gioco. Einaudi, Turin.
Bobbio, N., 1984b. Maestri e compagni. Passagli, Florence.
Bobbio, N., 1985. Stato, governo e società: per una teoria generale della politica. Einaudi, Torino.
Bobbio, N., 1987a. In: Bellamy, Richard (Ed.), The Future of Democracy: A Defence of the Rules of the Game (R. Griffen, Trans.). Polity, Cambridge.
Bobbio, N., 1987b. In: Bellamy, Richard (Ed.), Which Socialism? Marxism, Socialism and Democracy (R. Griffen, Trans.). Polity, Cambridge.
Bobbio, N., 1989a. In: Polito, Pietro (Ed.), Il terzo assente: saggi e discorsi sulla pace e la Guerra. Sonda, Turin.
Bobbio, N., 1989b. Thomas Hobbes. Einaudi, Torino.
Bobbio, N., 1989c. 'Dalla priorità dei doveri alla priorità dei diritti' in Bobbio 1999.
Bobbio, N., 1989d. Democracy and Dictatorship: The Nature and Limits of State Power (P. Kennealy, Trans.). Polity Press, Cambridge.
Bobbio, N., June 9, 1989e. L'utopia capovalta. La Stampa. A. 123, n. 128, 9 June, 1.
Bobbio, N., September–October 1989f. The upturned Utopia. The New Left Review 177, 37–39.
Bobbio, N., 1990. L'eta dei diritti. Einaudi, Turin.
Bobbio, N., 1991. Una Guerra giusta? Sul conflitto del golfo. Marsilio, Venice.
Bobbio, N., 1993a. Thomas Hobbes and the Natural Law Tradition (D. Gobetti, Trans.). The University of Chicago Press, Chicago & London.
Bobbio, N., 1993b. Il dubbio e la scelta: Intellectualli e potere nella societa contemporanea. NIS, Rome.
Bobbio, N., 1994. Elogio della mitezza e altri scritti morali. Linea d'Ombra, Milan.
Bobbio, N., 1995a. Ideological Profile of Twentieth Century Italy (L. Cochrane, Trans.). Princeton University Press, Princeton, NJ.
Bobbio, N., 1995b. Destra e sinistra. Donzelli editore, Rome.
Bobbio, N., 1996a. The Age of Rights (A. Cameron, Trans.). Polity, Cambridge.
Bobbio, N., 1996b. Left and Right (A. Cameron, Trans.). Polity, Cambridge.
Bobbio, N., 1996c. De senectute e altri scritti autobiografici. Einaudi, Turin.
Bobbio, N., 1997. In: Papuzzi, Alberto (Ed.), Autobiografia. Laterza, Bari.
Bobbio, N., 1999. In: Bovero, Michelangelo (Ed.), Teoria generale della politica. Einaudi, Turin.
Bobbio, N., 2000. In Praise of Meekness: Essays on Ethics and Politics (T. Chataway, Trans.). Polity, Cambridge.
Bobbio, N., 2001. Old Age and Other Essays (A. Cameron, Trans.). Polity, Cambridge.
Bobbio, N., 2002. In: Papuzzi, Alberto (Ed.), A Political Life (A. Cameron, Trans.). Polity, Cambridge.
Bobbio, N., Viroli, M., 2001. Dialogo intorno alla repubblica. Laterza, Rome-Bari.
Bobbio, N., Viroli, M., 2003. The Idea of the Republic (A. Cameron, Trans.). Polity, Cambridge.
Rousseau, J.J., 1973. The Social Contract and Discourses (G.D.H. Cole, Trans.). J. M. Dent and Sons Ltd, London.

Body: Anthropological Aspects

Thomas J Csordas, University of California, San Diego, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The theoretical status of 'the body' in cultural anthropology has evolved in roughly four stages: (1) in earlier decades of the twentieth century, it was an implicit, taken-for-granted background feature of social life; (2) beginning in the 1970s, the body became an explicit topic of ethnographic concern; (3) from the 1980s, the body began to become a problem to be accounted for with respect to its cultural and historical mutability; and (4) by the century's end, the body offered a theoretical locus for rethinking various aspects of culture and self. This conception of the changing status of the body in anthropology guides the exposition that follows.

Early Reflections on the Body

Although the body was traditionally implicit in anthropological writing, a rereading of older sources is nevertheless likely to offer a surprising richness of insights on bodiliness. Paul Radin, for example, explicitly discusses indigenous philosophy of the body in a chapter on 'The Ego and Human Personality' (1927). In what we would today call an attempt to formulate an ethnopsychology of the person, he emphasized the Maori distinction between material (substance) and immaterial (form) aspects of the body as the resting place of various components of the person. In its anatomical aspect, the Maori body was associated with psychical rather than physiological functions. After going on to discuss the person among the Oglala Dakota and the Batak, he concludes that these ethnotheories display an 'inability to express the psychical in terms of the body,' but must project it onto the external world (1927: pp. 273–274). Robert Lowie, in a chapter on 'Individual Variability' in religious experience, dealt with issues relevant to embodiment under the heading of 'sensory types' (1924: pp. 224–231). His data for this discussion were instances of visionary, or revelatory imagery. Noting the absence of studies in this area by anthropologists, he was concerned with the manner, in which imagination is concretely engaged with the various sensory modalities. Drawing primarily on Crow material supplemented by evidence from other North American tribes and from the African Bushmen, he suggested that the cross-cultural data conformed with the then-current psychological findings that visual imagery predominates in frequency over imagery in other sensory modalities, though certain individuals might show a propensity toward auditory or even motor, tactile, and kinesthetic imagery. Maurice Leenhardt (1947/1979) reports a conversation between himself and an elderly New Caledonian philosopher about the impact of European civilization on the indigenous cosmocentric world. His interlocutor said that "What you've brought us is the body." For Leenhardt, this pronouncement upends a stereotype that presumes the body lies on the side of nature (the primitive) and spirit on the side of culture (the civilized). Leenhardt vividly suggests that the very possibility of individuation has as its condition of possibility, a particular mode of inhabiting the world as a bodily being.

Some of the earliest sustained treatments of the body in its cultural and social dimensions were Robert Hertz's ([1909] 1960) study of the symbolic preeminence of the right hand and Marcel Mauss' ([1934]1950) justly famous article on techniques of the body. Mauss' work is almost universally cited as a precursor of the contemporary interest in the body. It anticipates the notions of practice and habitus later elaborated by Pierre Bourdieu (1977), though the notion of habitus had already passingly been used with a similar sense by Max Weber in his discussion on religions of ethical salvation. Important historically is the relation between this article on the body and Mauss' fragmentary but influential discussion of the person, appearing only 4 years later (1938/1950). In the latter article, he suggested that all humans have a sense of spiritual and corporal individuality, and saw the person as associated with the distinction between the world of thought and the material world as promulgated by Descartes and Spinoza.

Anthropology of the Body

In later decades, attention to the body largely took the form of studies of gesture, nonverbal communication, kinesics, and proxemics (Benthall and Polhemus, 1975). In these studies, interest in the body in its own right was subordinated to interest in communication as a cultural process, with the body serving as the means or medium of communication. In other words, rather than beginning with a concern for bodiliness per se, these analyses took language as their model, using a linguistic analogy to study various types of languages of the body. Perhaps for this reason, they are rarely cited in the literature on the body, and embodiment that has appeared in abundance since the late 1980s.

Two books by Mary Douglas define the threshold of a true anthropology of the body. In 'Purity and Danger' (1966), she made her famous argument that the dietary rules in the biblical book of Leviticus summarized the categories of Israelite culture, and made the general claim that social system controls induce consonance between social and physiological levels of experience. In 'Natural Symbols' (1973), Douglas made her equally famous argument that the bodily state of trance in different societies corresponds to the state of social organization characterized in terms of the 'grid' of cultural classifications and the 'group' control over the ego as an individual actor. Building on Mauss, she rejected both Edward

Hall's work on nonverbal communication and Claude Levi-Strauss' work on mythical thought, on the grounds that they lacked hypotheses to account for variation across cultures. In presenting such a hypothesis, she committed herself to understanding the body as a 'medium of expression,' arguing that a drive to achieve consonance across the levels of experience requires the use of the body to be 'coordinated with other media' so as to produce distinct bodily styles, and that 'the physical body can have universal meaning only as a system which responds to the social system, expressing it as a system' (1973: p. 112). Overall, Douglas' legacy to the anthropology of the body is encapsulated in her conception of 'two bodies,' the physical and social, 'the self and society' (1973: p. 112).

Blacking's (1977) introduction to his edited volume on anthropology of the body offers a programmatic outline that is both more concerned with the body per se and its contribution to social processes than with the manner, in which it reflects or expresses those processes, and which is more explicitly concerned with the relation between the biological and cultural. Its 'chief concern is with the cultural processes and products that are externalizations and extensions of the body in varying contexts of social interaction,' and it rejects the distinction between the biological and cultural anthropology on the grounds that culture has shaped the physical body, while features of culture such as language are biologically based (1977: p. 2). Blacking sketches four premises for the anthropology of the body: that society is a biological phenomenon, all humans possess a common repertoire of somatic states, nonverbal communication is fundamental, and the mind cannot be separated from the body (1977: p. 18). Blacking emphasized the liberatory potential in anthropology of the body insofar as it could contribute to dealienation and 'ownership of our senses.'

Once the body emerged from theoretical anonymity to become a recognized topic in its own right, its omnipresence in social life led to a multiplication of ways to organize its study. Following Douglas' inspiration to recognize 'two bodies,' Nancy Scheper-Hughes and Margaret Lock (1987) suggest that they instead consider 'three bodies,' including the individual body, social body, and body politic. John O'Neill (1985) suggested thinking in terms of 'five bodies,' including the world's body (i.e., the anthropomorphized cosmos), the social body, body politic, consumer body, and medical body.

The Body as a Theoretical Problem

It was not long between the emergence of the body as a topic and the transformation of the body into a problem. This aspect of the body's career in anthropology is deeply bound up with the career of anthropology in interdisciplinary studies. Scholars from virtually every branch of the human sciences have been influenced by reformulations of basic understandings of the body through the work of Michel Foucault on the discursive formations that have constituted the hospital (1973), the prison (1977), and sexuality (1978); the work of Pierre Bourdieu (1977) on practice and habitus; and the work of Maurice Merleau-Ponty (1962) on perception and embodiment. Equally, scholarly understandings of the body have been transformed by feminist works including Luce Irigaray's (1985) extended critique of psychoanalysis and Donna Haraway's (1991) studies of the encounter between gendered bodies and technology. Although it is a simplification to formularize the scope and nature of the theoretical sea change effected through such works, two aspects stand out. First, though work such as Blacking's called into question the distinction between mind and body, and between biological and cultural anthropology, subsequent work has gone even farther in calling into question the degree, to which biological nature can be considered a stable substrate of human existence. Second, though Douglas encouraged us to think in terms of two bodies corresponding to self and society, subsequent work has gone beyond her tendency to treat the body-self as a passive lump of clay or tabula rasa, upon which society imposes its codes, toward understanding it as a source of agency and intentionality, taking up, and inhabiting the world through processes of intersubjective engagement. This radical shift has made the body a central problem, and one of some urgency, across a range of disciplines. Thus, Emily Martin, whose own work on gender, science, and technology (1987, 1994) has contributed significantly to the body's move to center stage in social theory, has posed the question of the 'end of the body' as we have known it.

In addition, there has been a chorus of statements about the historical malleability of the body. The radical 'situatedness' of the body extends even to the domain of biology itself, as is evident in recent feminist theory that eliminates 'passivity' as an intrinsic characteristic of the female body, and reworks the distinction between sex and gender (Haraway, 1991: pp. 197–198), as well as the decoupling of female sexual pleasure with the act of conception. With biology no longer a monolithic objectivity, the body is transformed from object to agent (Haraway, 1991: pp. 197–198).

The contemporary cultural transformation of the body can be conceived not only in terms of revising biological essentialism and collapsing conceptual dualities, but also in discerning an ambiguity in the boundaries of corporeality itself. Some have suggested that in contemporary civilization, the human body can no longer be considered a bounded entity due to the destabilizing impact of social processes of commodification, fragmentation, and the semiotic barrage of images of body parts. Others have explicitly problematized bodily boundaries between animal and human, animal/human and machine, and human and divine. Exploring these cultural boundaries can be incredibly rewarding and remarkably problematic given the circumstances of corporeal flux and bodily transformation sketched above.

Embodiment and Culture

This radical rethinking has also created an opportunity for a rethinking of culture and self from the standpoint of the body and embodiment. If the body was often a background feature in traditional ethnographies, so it has often remained implicit in anthropological theories of culture, which historically have been cast in terms such as symbols, meanings, knowledge, practices, customs, or traits. The problematizing of the body and its movement to the center stage in social theory has also led to the emergence of studies that do not claim to be about the body per se, but instead suggest that culture and self can

be understood from the standpoint of embodiment as an existential condition, in which the body is the subjective source or intersubjective ground of experience.

This development can be seen taking place concretely over time in the work of a variety of anthropologists. For example, in the early 1970s, Strathern and Strathern produced a conventional monograph on body decoration in Mount Hagen, New Guinea. More recently, each author has developed their thought in light of a reformulated understanding of embodiment (A. Strathern, 1996) and its gendered nature (M. Strathern, 1992). Terence Turner (1995) published an influential essay on body decoration among the Brazilian Kayapo, and since then has moved from this conception of the 'social skin' to a more thorough development of the place of 'bodiliness' among the Kayapo and more generally in anthropological thinking (1995). James Fernandez, following his powerful treatment of metaphor and body symbolism in the Bwiti religion among the Fang in Gabon, later turned his attention to the place of the body in Bwiti explicitly with respect to developing theorizations of embodiment (Fernandez, 1990). In work on ritual healing, ritual language, and the cultural constitution of self among Catholic Charismatics in contemporary North America, Thomas Csordas elaborated earlier arguments cast in terms of rhetoric with analyses made from the standpoint of embodiment (Csordas, 1997). In her work on South Africa, Jean Comaroff (1985) refocused the traditional anthropological interest in body symbolism in the context of contemporary approaches to the analysis of political economy, and ritual healing, tracing these themes across the precolonial, colonial, and postcolonial periods. In two brief but compelling passages, Michelle Rosaldo (1984) did much to open the way for an understanding that collective symbols acquire 'power, tension, relevance, and sense' through embodiment (p. 141), and that emotions are 'embodied thoughts' (p. 143).

The theoretical crux of these new syntheses is a critique of tenacious conceptual dualities such as those between mind and body, subject and object, sex and gender, and body and embodiment. These developments must be seen in the theoretical context of the prominence since the 1970s of notions such as the interpretive turn, the linguistic turn, the move to cultures defined as systems of symbols, and concepts such as textuality, discourse, representation, and semiotics broadly conceived. Understandings of the body itself in terms of the 'metaphor of the text' is evidenced in common scholarly phrases like 'the body as text,' 'the inscription of culture on the body,' or 'reading the body.' Recent work on embodiment offers the opportunity to balance what is thus, to some, an overemphasis on representation. When this tack is chosen, a key theoretical term that comes to take its place alongside representation is 'being-in-the-world.' The latter notion, drawn from phenomenology, does not supplant representation, but offers it a dialogical partner: in brief, semiotics offers the notion of textuality in order to understand representation, and phenomenology offers the notion of embodiment in order to understand being-in-the-world. 'The body' per se, can then be construed both as a source of representations and as a ground of being-in-the-world (Csordas, 1994). Meanwhile, 'embodiment' becomes an avenue of approach to culture and self, just as textuality is an approach to the study of culture and self. Thus, to work in a 'paradigm of embodiment' is not to study anything new or different, but to address familiar topics – healing, emotion, gender, or power – from a different standpoint. There is not a special kind of data or a special method for eliciting such data, but a methodological attitude that demands attention to bodiliness even in purely verbal data such as written text, or oral interview.

Interdisciplinary Consolidation

Since the beginning of the twenty-first century, there has been an intellectual consolidation of studies on the human body, embodiment, and bodily experience across disciplines with anthropology playing a strong role. This is evident in the historically oriented reader by Lock and Farquhar (2007) with its gesture toward the increasing awareness of materiality as relevant to understanding the body, and in the collection of original and cutting edge works compiled by Mascia-Lees (2011) with its recognition of the distinction between the body and embodiment across issue ranging from cultural phenomenology to virtuality and from colonialism to impairment. Closely related volumes with a strong contribution from sociology, philosophy, and cultural studies include those edited by Forth and Crozier (2005), Moore and Kosut (2010), Pirani (2011), Weiss and Haber (1999), and Turner (2012), with critical individual articles by Gatens (1996) and Sheets-Johnstone (2009). Further evidence of the consolidation of body studies is the massive six volume, *Cultural History of the Human Body* (Crozier, 2010; Crozier and Beccalossi, 2010; Garrison, 2010; Kalof, 2010; Kalof and Bynum, 2010; Reeves, 2010; Sappol and Stephen, 2010) spanning the period from antiquity to the contemporary moment. From a number of disciplinary standpoints, significant works have appeared on embodiment and religion (Csordas, 2002, 2004), aesthetics (Johnson, 2007), and gesture (Kendon, 2004; McNeill, 2005; Noland, 2009). Particularly, intriguing is a new rethinking of the relation between mind and body that actively seek a synthesis between phenomenology and cognitive science (Chemero, 2009; Gallagher, 2005; Gibbs, 2005). Finally, an impressive body of work has been emerging from Latin America and Iberia (Esteban, 2004; Muniz, 2010, 2011; Caon, 2012; Steil and Carvalho, 2012), including a number of works emphasizing dance as a locus of embodiment studies (Citro, 2009, 2007; Citro and Aschieri, 2012). Although consolidation is an apt term for the manner in which these works define an empirical and methodological terrain, this in no way indicates a rigidification or decline into redundancy. Instead, it indicates a vibrance and surge of intellectual creativity that can be expected to yield continued scholarly advances. This can be seen as the concern with human corporality increasingly engages the concern with animality and materiality, as phenomenological and cognitive styles of understanding embodiment achieve a rapprochement, as interdisciplinary discourse continues to expand, and as an ever-growing number of topics come to be approached from the standpoint of embodiment.

Summary

In summary, the body's career in anthropology has been on an upswing. From an early anonymity as a taken-for-granted

background feature of social life, it has emerged in recent decades first as an explicit topic of anthropological research, then as a problem as its cultural and historical instability as a natural object became increasingly evident. More recently, embodiment has presented itself as an opportunity for reformulating earlier interpretations and rethinking fundamental concepts of culture and self. Through the 1990s, the problem posed by the body of the relation between representation and being-in-the-world has been an increasingly prominent site, at which anthropologists have become engaged in the wider interdisciplinary discourse of the human sciences. Relevant studies can be found under the rubrics not only of anthropology of the body, but also of medical anthropology and anthropologies of the senses, space, dance, violence, and science and technology. On a wide variety of fronts, the body continues to advance.

See also: Body Image; Body, History of; Cultural Relativism, Anthropology of; Erotic Capital; Gender and Feminist Studies in Geography; Mind–Body Dualism; Psychological Anthropology; Reflexivity in Anthropology.

Bibliography

Benthall, J., Polhemus, T. (Eds.), 1975. The Body as a Medium of Expression. E.P. Dutton, New York.

Blacking, J. (Ed.), 1977. The Anthropology of the Body. Academic Press, London.

Bourdieu, P., 1977. Outline of a Theory of Practice (R. Nice, Trans.). Cambridge University Press, Cambridge, UK.

Caon, Paulina Maria, 2012. Construir Cuerpos, Tecer Histórias: Educação e Cultura Corporal em Duas Comunidaded Paulistas. Annblume Editora, São Paulo.

Chemero, Anthony, 2009. Radical Embodied Cognitive Science. MIT Press, Cambridge.

Citro, Silvia (Ed.), 2007. Cuerpos Plurales: Antropología de y desde Los Cuerpos. Editorial Biblos, Buenos Aires.

Citro, Silvia, 2009. Cuerpos Significantes: Travesías de una etnografía dialéctica. Editorial Biblos, Buenos Aires.

Citro, Silvia, Aschieri, Patricia (Eds.), 2012. Cuerpos en Movimiento: Antropología de y desde las Danzas. Editorial Biblos, Buenos Aires.

Comaroff, J., 1985. Body of Power, Spirit of Resistance: The Culture and History of an African People. University of Chicago Press, Chicago, IL.

Crozier, Ivan, Beccalossi, Chiara (Eds.), 2010. A Cultural History of the Human Body. A Cultural History of the Human Body in the Modern Age (1920–2000+), vol. 6. Berg Publishers (Edited by Ivan Crozier, University of Edinburgh, and Chiara Beccalossi, University of Queensland).

Csordas, T.J. (Ed.), 1994. Embodiment and Experience: The Existential Ground of Culture and Self. Cambridge University Press, Cambridge, UK.

Csordas, T.J., 1997. Language, Charisma, and Creativity: The Ritual Life of a Religious Movement. University of California Press, Berkeley, CA.

Csordas, Thomas J., 2002. Body/Meaning/Healing. Palgrave, New York.

Csordas, Thomas J., 2004. Asymptote of the ineffable: embodiment, alterity, and the theory of religion. Current Anthropology 45, 163–185.

Douglas, M., 1966. Purity and Danger. Routledge and Kegan Paul, London.

Douglas, M., 1973. Natural Symbols. Vintage, New York.

Esteban, Maria Luz, 2004. Antropología del Cuerpo: Género, Itinerarios corporals, identidat y Cambio. Edicions Bellaterra, Barcelona.

Fernandez, J., 1990. The body in Bwiti: variations on a theme by Richard Werbner. Journal of Religion in Africa 20, 92–111.

Forth, Christopher, Crozier, Ivan (Eds.), 2005. Body Parts: Critical Explorations in Corporeality. Lexington Books, Lanham.

Gallagher, Shaun, 2005. How the Body Shapes the Mind. Oxford University Press, Oxford.

Garrison, Daniel (Ed.), 2010. A Cultural History of the Human Body: A Cultural History of the Human Body in Antiquity (1000 BC - 500 AD), vol. 1. Berg Publishers (Northwestern University).

Gatens, Moira, 1996. Imaginary Bodies: Ethics, Power, and Corporeality. Routledge, London.

Gibbs Jr., Raymmond W., 2005. Embodiment and Cognitive Science. Cambridge University Press, New York.

Haraway, D., 1991. Simians, Cyborgs, and Women: The Reinvention of Nature. Routledge, New York.

Hertz, R., 1960[1909]. The Preeminence of the Right Hand: Death and the Right Hand (R. Needham, C. Needham, Trans.). Aberdeen University Press, Aberdeen, UK.

Irigaray, L., 1985. This Sex Which Is Not One (C. Porter, C. Burke, Trans.). Cornell University Press, Ithaca, IL.

Johnson, Mark, 2007. The Meaning of the Body: Aesthetics of Human Understanding. University of Chicago Press, Chicago, IL.

Kalof, Linda (Ed.), 2010. A Cultural History of the Human Body: A Cultural History of the Human Body in the Medieval Age (1000–1400), vol. 2. Berg Publishers, Oxford (Edited by Linda Kalof, Michigan State University).

Kalof, Linda, Bynum, William (Eds.), 2010. A Cultural History of the Human Body: A Cultural History of the Human Body in the Renaissance (1400–1650), vol. 3. Berg Publishers, Oxford (Edited by Linda Kalof, Michigan State University and William Bynum, University College London).

Kendon, Adam, 2004. Gesture: Visible Action as Utterance. Cambridge University Press, Cambridge.

Leenhardt, M., 1947/1979. Do Kamo: Person and Myth in a Melanesian World (B. Miller, Trans.).

Lock, Margaret, Farquhar, Judith (Eds.), 2007. Beyond the Body Proper: Reading the Anthropology of Material Life. Duke University Press, Durham.

Lowie, R.H., 1924. Primitive Religion. Grosset and Dunlap, New York.

Martin, E., 1987. The Woman in the Body: A Cultural Analysis of Reproduction. Beacon Press, Boston, MA.

Martin, E., 1994. Flexible Bodies: The Role of Immunity in American Culture from the Days of Polio to the Age of AIDS. Beacon Press, Boston, MA.

Mascia-Lees, Frances (Ed.), 2011. Companion to the Anthropology of the Body/Embodiment. Wiley-Blackwell, Chichester, UK.

Mauss, M., 1950[1934]. Les Techniques du Corps. Sociologie et Anthropologie. Presses Universitaires de France, Paris.

McNeill, David, 2005. Gesture and Thought. University of Chicago Press, Chicago, IL.

Merleau-Ponty, M., 1962. Phenomenology of Perception (J. Edie, Trans.). Northwestern University Press, Evanston, IL.

Moore, Lisa Jean, Kosut, Mary (Eds.), 2010. The Body Reader: Essential Social and Cultural Readings. NYU Press, New York.

Muñiz, Elsa (Ed.), 2010. Disciplinas Y Prácticas Corporales: Una Mirada a Las Sociedades Contemporáneas. Anthropolos Editorial, Barcelona.

Muñiz, Elsa, 2011. La Cirurgía Cosmética: Un Desafío a la "Naturaleza"? Universidad Autónoma Metropoliltana, México.

Noland, Carrie, 2009. Agency and Embodiment: Performing Gestures/Producing Culture. Harvard University Press, Cambridge.

O'Neill, J., 1985. Five Bodies: The Shape of Modern Society. Cornell University Press, Ithaca, IL.

Pirani, Bianca Maria (Ed.), 2011. Learning from Memory: Body, Memory, and Technology in a Globalizing World. Cambridge Scholars, Newcastle upon Tyne.

Radin, P., 1927. Primitive Man as Philosopher. Dover Publications, New York.

Reeves, Carole (Ed.), 2010. A Cultural History of the Human Body. A Cultural History of the Human Body in the Enlightenment (1650–1800), vol. 4. Berg Publishers (Edited by Carole Reeves, Wellcome Trust Centre for the History of Medicine, University College London).

Rosaldo, M., 1984. In: Shweder, R., Levine, R. (Eds.), Toward an Anthropology of Self and Feeling: Culture Theory. Cambridge University Press, Cambridge, UK, pp. 137–157.

Sappol, Michael, Stephen, P., 2010. A cultural History of the human body. In: A Cultural History of the Human Body in the Age of Empire (1800-1920), vol. 5. Berg Publishers (Edited by Michael Sappol, National Library of Medicine in Washington, DC, and Stephen P. Rice, Ramapo College of New Jersey).

Scheper-Hughes, N., Lock, M., 1987. The mindful body: a prolegomenon to future work in medical anthropology. Medical Anthropology Quarterly 1, 6–41.

Sheets-Johnstone, Maxine, 2009. The Corporeal Turn: An Interdisciplinary Reader. Imprint Academic Philosophy Documentation Center, Charlottesville, VA.

Steil, Carlos Alberto, Carvalho, Isa bel Cristina de Moura (Eds.), 2012. Cultura, Percepção, e Ambeinte: Diálogos com Tim Ingold. Editora Terceiro Nome, São Paulo.

Strathern, A.J., 1996. Body Thoughts. University of Michigan Press, Ann Arbor, MI.

Strathern, M., 1992. Reproducing the Future: Essays on Anthropology, Kinship, and the New Reproductive Technologies. Routledge, New York.

Turner, Bryan (Ed.), 2012. Routledge Handbook of Body Studies. Routledge, London.

Turner, T., 1995. Social body and embodied subject: bodiliness, subjectivity, and sociality among the Kayapo. Cultural Anthropology 10, 143–170.

Weiss, Gail, Haber, Honi (Eds.), 1999. Perspectives on Embodiment. Routledge, New York, pp. 143–162.

Body, Evolution of

Christopher B Ruff, Johns Hopkins University, Baltimore, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The earliest stages in the evolution of the human body (2–6 Ma) involved a gradual adaptation to terrestrial bipedality through various structural modifications. Body size in these early ancestors was smaller, but sexual dimorphism larger than in modern humans. With the evolution of *Homo erectus* about 1.5–2.0 Ma, essentially modern body proportions and levels of sexual dimorphism were established. Body size reached a peak between about 50 000 and 500 000 years ago and has declined on average since then. Body shape variation among both earlier *Homo* and modern humans is strongly conditioned by climatic adaptation.

Major changes in body form have taken place during human evolution, related to the initial adoption and increasing commitment to bipedal locomotion and continuing adaptations to climatic, dietary, and other environmental factors. Although all modern humans share a similar body plan, there is still substantial variability in body size and body shape within and between populations that is in part a reflection of environmental adaptation. This article reviews the evidence for both the evolution and modern distribution of body form in humans and its behavioral and ecological significance.

The Bipedal Transformation

It is widely agreed that the transition from quadrupedal to bipedal locomotion was one of the most important innovations in human evolution; indeed, bipedality defines our lineage better than any other single characteristic (Clark, 1964; Aiello and Dean, 1990; Fleagle, 1999). Efficient bipedal locomotion requires changes in a number of anatomical characteristics, including restructuring of the lower limb and vertebral column, and other alterations. Recent evidence indicates that some form of bipedality may have been practiced by hominins as early as 6 Ma (Senut et al., 2001; Richmond and Jungers, 2008), and possibly even 7 Ma (Zollikofer et al., 2005), very close to the time when humans split off from African apes.

However, it is also becoming increasingly clear that the transition to a fully modern form of terrestrial bipedality occurred much later in hominin evolution, perhaps only with the appearance of *Homo erectus* about 1.5–1.8 Ma (Walker and Leakey, 1993; Ruff, 2008). Hominins dating from about 2.0 to 4.5 Ma, including *Ardipithecus ramidus*, *Australopithecus anamensis*, *Australopithecus africanus*, *Australopithecus garhi*, and *Australopithecus sediba*, as well as some extending more recently (between 1.2 and 2.0 Ma), including *Paranthropus robustus*, *Paranthropus boisei*, and *Homo habilis*, exhibit a mosaic of locomotor characteristics that indicate either an altered form of terrestrial bipedality and/or retained arboreal (tree climbing) capabilities. These include, in some taxa, a foot better adapted for climbing than in modern humans (Clarke and Tobias, 1995; Lovejoy et al., 2009a; Zipfel et al., 2011; DeSilva et al., 2012; Haile-Selassie et al., 2012), long and robust upper limbs and a more superiorly oriented shoulder joint, again for climbing (Clark et al., 1984; Heinrich et al., 1993; White et al., 1993; Kimbel et al., 1994; Leakey et al., 1995; McHenry and Berger, 1998; Lovejoy et al., 2009b; Ruff, 2009; Green and Alemseged, 2012), and variations in foot, hip, and pelvic structure that indicate some differences in posture during walking compared to modern humans (Ruff, 1998; Lovejoy et al., 2009c; Zipfel et al., 2011; Ruff and Higgins, 2013).

Thus, the adoption of terrestrial bipedality was gradual, and probably involved increasing reliance on this form of locomotion while still retaining the ability to use trees, e.g., for escape from predators or for food procurement. There are many theories on why and how human bipedalism evolved, including ecological and dietary shifts, and changes in social organization (for a review, see Fleagle, 2013). While distinctively human proportions of the body, such as forelimb to hindlimb size, were established by 1.5 Ma in *H. erectus* (Walker and Leakey, 1993; Ruff, 2008), within this basic body plan variations in other aspects of body size and shape are apparent throughout the more recent fossil record and among living humans.

Variation in Body Size

Body mass (weight) for fossil specimens has been estimated using various parts of the skeleton, including craniodental dimensions, but it is clear that postcranial dimensions provide the least biased and potentially most accurate means for reconstructing body size, especially in an evolutionary lineage where changes in relative tooth and cranial size have been dramatic (Pilbeam and Gould, 1974). Two basic types of methods can be distinguished: a mechanical approach, which uses features that are related to the mechanical load-bearing function of the skeleton, such as hip joint (femoral head) size, and a morphometric approach, in which features that actually reflect the size and shape of the postcranial skeleton are used (Auerbach and Ruff, 2004). For the latter, it has been shown that stature, estimated from limb bone length, and body breadth, measured across the pelvis, together provide a good estimate of body mass (Ruff, 2000; Auerbach and Ruff, 2004; Ruff et al., 2005). Correlations between

body masses determined using both techniques in the same specimens are good, indicating little bias in results (Ruff et al., 1997; Auerbach and Ruff, 2004).

Changes in estimated body mass over 6 My of human evolution are shown in Figure 1, along with a set of values for modern living humans for comparison. A total of 148 premodern individual specimens are included. In most of these, body mass was estimated from hip joint size (femoral head breadth measured directly or in a few cases estimated from acetabular breadth), using equations derived from modern humans (Ruff, 2010). In 26 cases, it was also possible to estimate body mass from stature and pelvic breadth, again using equations derived from modern humans (Ruff et al., 2005). In these specimens, an average of the two estimates was used (as in previous studies, the two techniques produced very similar results). Unlike in some previous studies (e.g., Ruff et al., 1997), no pelvic breadths were estimated, due to recent evidence that the patterning of relative pelvic breadth among fossil hominins was more complex than previously modeled (Trinkaus, 2009; Ruff, 2010). Because of the much higher density of data available for more recent periods, the data are plotted on a logarithmic temporal scale. Specimens are divided into *Homo* (*Homo* sp., *H. erectus*, and *Homo sapiens*, including Neandertals) and non-*Homo* (see earlier text for taxa). The modern living data are 53 sex-specific sample means (Ruff, 2010). Both modern and earlier data are from the Old World (except for two modern Inupiat samples) and include samples or individuals from both higher and lower latitudes as well as both sexes; all are adults.

Several trends are apparent in Figure 1. First, body size was relatively small (21–53 kg) in the earliest period of human evolution and markedly increased with the appearance of *H. erectus* at about 2 Ma. This likely coincided with a number of other biological and behavioral changes, including a complete commitment to terrestrial bipedality (also see earlier text), increased diurnal travel through open country, and possibly an increased reliance on hunting (or at least foraging range) (Wheeler, 1992; McHenry, 1994; Ruff, 2009). Note that one specimen of uncertain attribution – the pelvis from Gona, Ethiopia (Simpson et al., 2008) – falls well under the known body size range of *Homo*, further supporting its assignment to a non-*Homo* species, most likely *Australopithecus boisei* (Ruff, 2010).

The second general observation apparent from Figure 1 is that body size in premodern *Homo* was generally above the average for modern (living) human populations. The median body mass for the modern humans is 57 kg; 87% of the earlier *Homo* specimens are heavier than this, with an overall median of 68 kg (19% heavier). Part of this may be attributable to a male bias in the fossil record: of the 87 out of 113 earlier *Homo* specimens that can be reliably sexed, 63% are male. However, this pattern is evident even within sex: earlier *Homo* males average 20% heavier and females 13% heavier than modern males and females. Another possible factor is latitude: as discussed later in the text, among modern humans (and many other mammalian species) there is an ecogeographic cline of increased body mass in colder climates, and a majority of the premodern specimens in Figure 2 (especially among the later ones) are from higher latitudes. Figure 2 is a plot of temporal trends in body mass over the past 35 000 years among only higher-latitude males (mainly European),

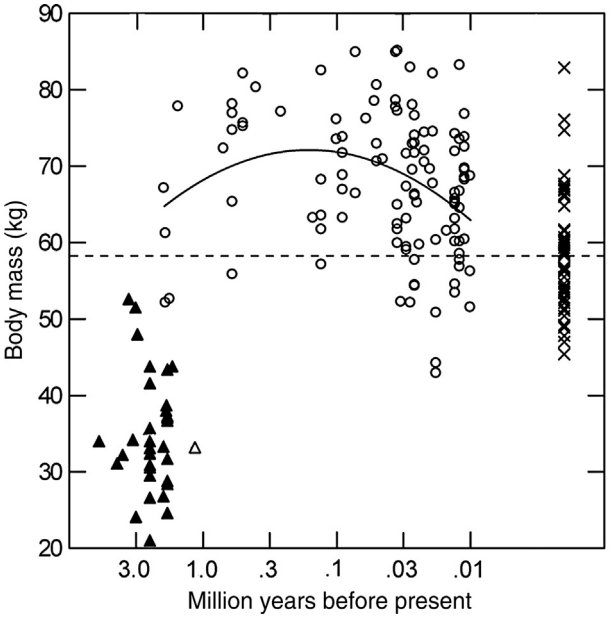

Figure 1 Temporal changes in body mass (weight) over the past 6 My of hominin evolution (note logarithmic time scale). Body masses in fossil specimens determined primarily from femoral head size (see text). Filled triangles: non-*Homo* species; open triangle: Gona specimen, of uncertain taxonomy (see Ruff, 2010); open circles: *Homo;* x's: modern humans (worldwide sample of male and female population means). Quadratic equation used to fit regression line through premodern *Homo*. Dotted line indicates modern median.

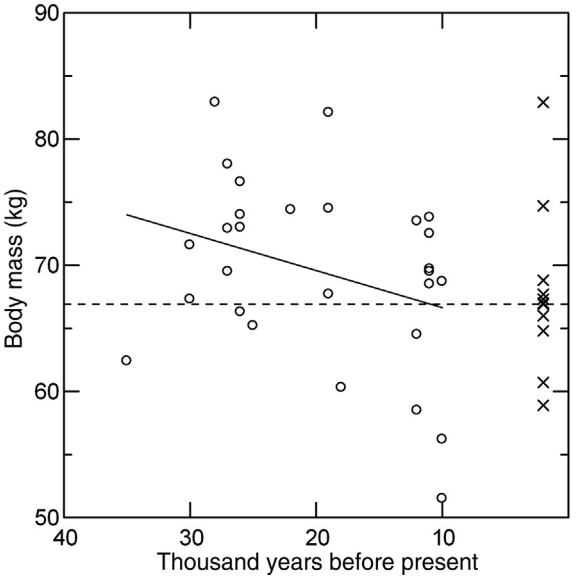

Figure 2 Change in body mass among higher-latitude males over the past 35 000 years. Body masses and symbols as in **Figure 1**. Linear regression fit through premodern sample. Dotted line indicates modern median.

compared to modern higher-latitude males. The linear regression on time (not including the modern data) is negative and near-significant ($p = .09$). Higher-latitude males in the Early-Middle Upper Paleolithic (16 000–35 000 years ago) average about 9% heavier than modern higher-latitude males. Median body mass in males from the Late Upper Paleolithic (10 000–15 000 years ago) is very similar to that of the living sample.

This highlights another aspect of temporal trends in body size: maximum body mass was apparently reached in the late Middle and early Late Pleistocene, i.e., between about 500 000 and 50 000 years ago, after which it declined to modern human values by the end of the Pleistocene (10 000 years ago). Thus, a quadratic curve best fits the premodern *Homo* data in Figure 1 (as determined by the akaike information criterion (AIC)). Increased body size during the late Middle and early Late Pleistocene may be related to foraging strategies such as an increase in big game hunting (e.g., Thieme, 1997) as well as more common occupation of high-latitude environments (see below). The later decrease in body size may be attributable to changing technology, resource exploitation, and demographic shifts in the terminal Pleistocene that no longer selected as strongly for large body size (Frayer, 1984; Richards et al., 2001; Milisauskas, 2002; Lagerlöf, 2007). There is evidence that average body size continued to decline through the Neolithic, at least in Europe (Frayer, 1984; Niskanen et al., 2012), perhaps because of decreased nutritional quality and/or increased disease transmission associated with the agricultural revolution (Cohen and Armelagos, 1984). Body size has increased dramatically over the past century in many populations, probably due to improvements in the environment (Eveleth and Tanner, 1990; Katzmarzyk and Leonard, 1998), although the opposite trend has been observed in some disadvantaged populations (Tobias, 1985).

Sexual dimorphism (i.e., differences between sexes) in body size is another important characteristic of a species with many potential behavioral and ecological implications (e.g., McHenry, 1994; Plavcan and Van Schaik, 1997). Recent analyses of sexual dimorphism in body mass, using methods similar to those described earlier, indicate elevated levels of dimorphism in hominin taxa prior to the appearance of *H. erectus*, with an average male/female ratio of about 1.3 or higher (Ruff, 2002, 2010). Species within *Homo* are characterized by reduced levels of dimorphism, similar to those of modern humans, i.e., an average male/female ratio of about 1.1–1.2. This implies changes in ecology and behavior with the evolution of *Homo*. It also implies that in at least some aspects of social organization, early *Homo* may not have differed markedly from modern hunter-gatherers.

Variation in Body Shape

One aspect of the environment that has probably had a profound effect on human body shape is climate. Systematic relationships between climate and body form among homeothermic animals have been recognized for more than a century and have been codified as Bergmann's and Allen's ecogeographical 'Rules' (Mayr, 1963). Bergmann's Rule states that within a species or closely related group of species, populations living in colder climates will be larger in body mass than those in warmer climates, while Allen's Rule states that colder climates will be associated with shorter extremities and warmer climates with longer extremities. Both of these observations are actually consequences of a more general relationship between body mass and body surface area, in which both a larger body mass and relatively shorter extremities decrease the ratio of surface area to body mass, thus conserving heat, and vice versa. The same principle can also be applied to the overall shape of the human trunk: it can be shown that an absolutely wide trunk (and thus wide body) will lead to a decreased surface area/body mass ratio, regardless of body height, while an absolutely narrow trunk will lead to an increase in the ratio, explaining the narrower bodies of tropical populations and wider bodies of arctic populations (Ruff, 1994).

Ecogeographical clines that conform to physiological (climatic) expectations have been demonstrated for many aspects of living human body form, including body mass, body breadth, relative limb length, and direct estimates of surface area to body mass (Roberts, 1978; Ruff, 1994; Cowgill et al., 2012). It is more difficult to evaluate such morphological variability in our fossil ancestors, but recent paleoanthropological discoveries combined with developments in methodology have provided some new insights.

All early hominins apparently had relatively wide bodies compared to humans living in the same areas today (Ruff, 2010), which may be the result of an altered birth process in which the neonate's head did not rotate during birth as it does today (Tague and Lovejoy, 1986; Ruff, 1995, 2010). However, even among earlier human ancestors, there is evidence for latitudinal variation in body breadth similar to that observed among modern humans, with specimens from higher latitudes exhibiting wider bodies than those from lower latitudes (Ruff, 2010). Relative limb length is also affected by locomotion (*see* Section The Bipedal Transformation), so this aspect of morphology is more difficult to evaluate in terms of climatic adaptation per se in early hominins. However, among *H. erectus* and later taxa, where terrestrial bipedality was well established, lower-latitude specimens are characterized by relatively longer extremities, and higher-latitude specimens by shorter extremities, again matching ecogeographic predictions (Trinkaus, 1981; Ruff and Walker, 1993; Ruff, 1994; Holliday, 1997; Rosenberg et al., 2006). In fact, in some ways earlier humans appear to exhibit extreme morphological adaptations to climate, i.e., 'hyperarctic' and 'hypertropical' body forms (Ruff and Walker, 1993; Ruff, 1994; Holliday, 1997; Weaver, 2003), possibly as a result of less efficient cultural buffering against the environment.

Such observations can also inform interpretations of past population movements and in situ adaptations. Occupants of Western Europe during the Middle and early Late Pleistocene were subjected to periodic episodes of extreme cold, i.e., glaciations; cool conditions prevailed even during some interglacial periods (Trinkaus et al., 1999). Body breadths were correspondingly wide, either absolutely or relative to stature, and limb lengths relatively short (Ruff, 1994; Holliday, 1997; Arsuaga et al., 1999; Ruff, 2010; Walker et al., 2011a,b). About 25 000–45 000 years ago, archaic humans (Neandertals) in Europe were gradually replaced by 'early anatomically

modern humans' (EAM) (Finlayson et al., 2006; Benazzi et al., 2011; Higham et al., 2011). These EAM populations were much taller, relatively less wide-bodied, and had relatively longer limbs (Ruff, 1994; Holliday, 1997), all evidence of a more tropical origin. Over time, the body proportions of European EAMs became less 'tropical,' increasing in relative body breadth and decreasing in relative limb length; until by the Late Upper Paleolithic (10 000–19 000 years ago), they were essentially indistinguishable from those of modern Europeans (Ruff, 1994; Holliday, 1997). This change can be seen, at least in part, as evidence of environmental adaptation to colder climatic conditions following migration from farther south, i.e., northeast African and/or the Middle East (White et al., 2003). Anatomical evidence of previous climatic adaptation to cold conditions followed by a migration over the Bering Sea Land Bridge is also apparent among early inhabitants of the New World, shedding new light on the origins of these populations (Ruff, 1994; Auerbach, 2012).

Body Size as a Baseline for Other Comparisons

In addition to its biological and cultural significance, body size is important in human evolutionary studies because it is often used as a 'denominator' against which to evaluate other physical traits. One prime example is brain size, which is commonly expressed as a ratio against body mass, or encephalization quotient (EQ) (e.g., Pilbeam and Gould, 1974). EQ increased in early *Homo* (1.5–2.0 Ma) relative to earlier taxa, but then remained constant for at least a million years (to 0.5 Ma), after which it increased again, exponentially, to modern values (McHenry, 1994; Ruff et al., 1997). Skeletal robusticity has also been evaluated relative to body size, and shown to follow an inverse trend relative to brain size (Ruff et al., 1993), which may reflect the increasing importance of technology and decreasing selective advantage of physical strength among modern humans.

Finally, the body size of earlier humans can be used as a baseline for contextualizing the health of modern populations. Comparing present-day measures of stature and body mass with those of ancestral populations may give better estimates of potential body size limits and provide guidance in setting nutritional and other health standards (WHO, 1995).

See also: Adaptation, Fitness, and Evolution; Birth, Anthropology of; Body, History of; Body: Anthropological Aspects; Brain, Evolution of; Human Behavioral Ecology; Primates, Evolution of; Technology, Evolution of.

Bibliography

Aiello, L., Dean, C., 1990. An Introduction to Human Evolutionary Anatomy. Academic Press, London, p. 596.

Arsuaga, J.-L., Lorenzo, C., Carretero, J.-M., Gracia, A., Martinez, I., Garcia, N., Bermudez de Castro, J.-M., Carbonell, E., 1999. A complete human pelvis from the Middle Pleistocene of Spain. Nature 399, 255–258.

Auerbach, B.M., 2012. Skeletal variation among early Holocene North American humans: implications for origins and diversity in the Americas. American Journal of Physical Anthropology 149 (4), 525–536.

Auerbach, B.M., Ruff, C.B., 2004. Human body mass estimation: a comparison of "morphometric" and "mechanical" methods. American Journal of Physical Anthropology 125, 331–342.

Benazzi, S., Douka, K., Fornai, C., Bauer, C.C., Kullmer, O., Svoboda, J., Pap, I., Mallegni, F., Bayle, P., Coquerelle, M., Condemi, S., Ronchitelli, A., Harvati, K., Weber, G.W., 2011. Early dispersal of modern humans in Europe and implications for Neanderthal behaviour. Nature 479 (7374), 525–528.

Clark, J.D., Asfaw, B., Assefa, G., Harris, J.W.K., Kurashina, H., Walter, R.C., White, T.D., Williams, M.A.J., 1984. Paleoanthropological discoveries in the Middle Awash Valley, Ethiopia. Nature 307 (5950), 423–428.

Clark, W.E.L., 1964. The Fossil Evidence for Human Evolution. University of Chicago Press, Chicago.

Clarke, R.J., Tobias, P.V., 1995. Sterkfontein member 2 foot bones of the oldest South African hominid. Science 269, 521–524.

Cohen, M.N., Armelagos, G.J., 1984. Paleopathology at the Origins of Agriculture. Academic Press, New York.

Cowgill, L.W., Eleazer, C.D., Auerbach, B.M., Temple, D.H., Okazaki, K., 2012. Developmental variation in ecogeographic body proportions. American Journal of Physical Anthropology 148 (4), 557–570.

DeSilva, J.M., Proctor, D.J., Zipfel, B., 2012. A complete second metatarsal (StW 89) from Sterkfontein Member 4, South Africa. Journal of Human Evolution 63 (3), 487–496.

Eveleth, P.B., Tanner, J.M., 1990. Worldwide Variation in Human Growth. Cambridge University Press, Cambridge, p. 397.

Finlayson, C., Pacheco, F.G., Rodriguez-Vidal, J., et al., 2006. Late survival of Neanderthals at the southernmost extreme of Europe. Nature 443 (7113), 850–853.

Fleagle, J.G., 2013. Primate Adaptation and Evolution, 3rd ed. Academic Press, Inc., New York.

Frayer, D.W., 1984. Biological and cultural change in the European Late Pleistocene and Early Holocene. In: Smith, F.H., Spencer, F. (Eds.), The Origins of Modern Humans: A World Survey of the Fossil Evidence. Wiley-Liss, New York, pp. 211–250.

Green, D.J., Alemseged, Z., 2012. *Australopithecus afarensis* scapular ontogeny, function, and the role of climbing in human evolution. Science 338 (6106), 514–517.

Haile-Selassie, Y., Saylor, B.Z., Deino, A., Levin, N.E., Alene, M., Latimer, B.M., 2012. A new hominin foot from Ethiopia shows multiple Pliocene bipedal adaptations. Nature 483 (7391), 565–569.

Heinrich, R.E., Rose, M.D., Leakey, R.E., Walker, A.C., 1993. Hominid radius from the middle Pliocene of Lake Turkana, Kenya. American Journal of Physical Anthropology 92 (2), 139–148.

Higham, T., Compton, T., Stringer, C., et al., 2011. The earliest evidence for anatomically modern humans in northwestern Europe. Nature 479 (7374), 521–524.

Holliday, T.W., 1997. Body proportions in Late Pleistocene Europe and modern human origins. Journal of Human Evolution 32, 423–447.

Katzmarzyk, P.T., Leonard, W.R., 1998. Climatic influences on human body size and proportions: ecological adaptations and secular trends. American Journal of Physical Anthropology 106, 483–503.

Kimbel, W.H., Johanson, D.C., Rak, Y., 1994. The first skull and other new discoveries of *Australopithecus afarensis* at Hadar, Ethiopia. Nature 368 (6470), 449–451.

Lagerlöf, N.-P., 2007. Long-run trends in human body mass. Macroeconomic Dynamics 11, 367–387.

Leakey, H.G., Feibel, C.S., McDougall, I., Walker, A., 1995. New four-million-year-old hominid species from Kanapoi and Allia Bay, Kenya. Nature 376, 565–571.

Lovejoy, C.O., Latimer, B., Suwa, G., Asfaw, B., White, T.D., 2009a. Combining prehension and propulsion: the foot of *Ardipithecus ramidus*. Science 326 (5949), pp. 72e71–72e78.

Lovejoy, C.O., Suwa, G., Simpson, S.W., Matternes, J.H., White, T.D., 2009b. The great divides: *Ardipithecus ramidus* reveals the postcrania of our last common ancestors with African apes. Science 326 (5949), 100–106.

Lovejoy, C.O., Suwa, G., Spurlock, L., Asfaw, B., White, T.D., 2009c. The pelvis and femur of *Ardipithecus ramidus*: the emergence of upright walking. Science 326 (5949), pp. 71e71–71e76.

Mayr, E., 1963. Animal Species and Evolution. Harvard University Press, Cambridge.

McHenry, H.M., 1994. Behavioral ecological implications of early hominid body size. Journal of Human Evolution 27, 77–87.

McHenry, H.M., Berger, L.R., 1998. Body proportions in *Australopithecus afarensis* and *africanus* and the origin of the genus *Homo*. Journal of Human Evolution 35, 1–22.

Milisauskas, S. (Ed.), 2002. European Prehistory: A Survey. Kluwer Academic/Plenum, New York.

Niskanen, M., Ruff, C., Holt, B., et al., 2012. Temporal trends and regional differences in body size and shape of Europeans from the Late Pleistocene to recent times (abstract). American Association of Physical Anthropologists.

Pilbeam, D., Gould, S.J., 1974. Size and scaling in human evolution. Science 186, 892–901.

Plavcan, J.M., Van Schaik, C.P., 1997. Interpreting hominid behavior on the basis of sexual dimorphism. Journal of Human Evolution 32, 345–374.

Richards, M.P., Pettitt, P.B., Stiner, M.C., Trinkaus, E., 2001. Stable isotope evidence for increasing dietary breadth in the European mid-Upper Paleolithic. Proceedings of the National Academy of Sciences of the United States of America 98 (11), 6528–6532.

Richmond, B.G., Jungers, W.L., 2008. *Orrorin tugenensis* femoral morphology and the evolution of hominin bipedalism. Science 319 (5870), 1662–1665.

Roberts, D.F., 1978. Climate and Human Variability, second ed. Cummings, Menlo Park.

Rosenberg, K.R., Lü, Z., Ruff, C.B., 2006. Body size, body proportions and encephalization in a Middle Pleistocene archaic human from northern China. Proceedings of the National Academy of Sciences 103, 3552–3556.

Ruff, C.B., 1994. Morphological adaptation to climate in modern and fossil hominids. The American Journal of Physical Anthropology Yearbook Series 37, 65–107.

Ruff, C.B., 1995. Biomechanics of the hip and birth in early *Homo*. American Journal of Physical Anthropology 98, 527–574.

Ruff, C.B., 1998. Evolution of the hominid hip. In: Strasser, E., Rosenberger, A., McHenry, H., Fleagle, J. (Eds.), Primate Locomotion: Recent Advances. Plenum Press, Davis, CA, pp. 449–469.

Ruff, C.B., 2000. Body mass prediction from skeletal frame size in elite athletes. American Journal of Physical Anthropology 113, 507–517.

Ruff, C.B., 2002. Variation in human body size and shape. Annual Review of Anthropology 31, 211–232.

Ruff, C.B., 2008. Femoral/humeral strength in early African *Homo erectus*. Journal of Human Evolution 54, 383–390.

Ruff, C.B., 2009. Relative limb strength and locomotion in *Homo habilis*. American Journal of Physical Anthropology 138, 90–100.

Ruff, C.B., 2010. Body size and body shape in early hominins – implications of the Gona pelvis. Journal of Human Evolution 58 (2), 166–178.

Ruff, C.B., Higgins, R., 2013. Femoral neck structure and function in early hominins. American Journal of Physical Anthropology 150, 512–525.

Ruff, C.B., Niskanen, M., Junno, J.A., Jamison, P., 2005. Body mass prediction from stature and bi-iliac breadth in two high latitude populations, with application to earlier higher latitude humans. Journal of Human Evolution 48 (4), 381–392.

Ruff, C.B., Trinkaus, E., Holliday, T.W., 1997. Body mass and encephalization in Pleistocene *Homo*. Nature 387, 173–176.

Ruff, C.B., Trinkaus, E., Walker, A., Larsen, C.S., 1993. Postcranial robusticity in *Homo*, I: temporal trends and mechanical interpretation. American Journal of Physical Anthropology 91, 21–53.

Ruff, C.B., Walker, A., 1993. Body size and body shape. In: Walker, A., Leakey, R. (Eds.), The Nariokotome *Homo erectus* Skeleton. Harvard University Press, Cambridge, pp. 234–265.

Senut, B., Pickford, M., Gommery, D., Mein, P., Cheboi, K., Coppens, Y., 2001. First hominid from the Miocene (Lukeino Formation, Kenya). Comptes Rendus De L Academie Des Sciences Serie Ii Fascicule a-Sciences De La Terre Et Des Planetes 332 (2), 137–144.

Simpson, S.W., Quade, J., Levin, N.E., Butler, R., Dupont-Nivet, G., Everett, M., Semaw, S., 2008. A female *Homo erectus* pelvis from Gona, Ethiopia. Science 322 (5904), 1089–1092.

Tague, R.G., Lovejoy, C.O., 1986. The obstetric pelvis of A.L. 288-1 (Lucy). Journal of Human Evolution 15, 237–255.

Thieme, H., 1997. Lower Paleolithic hunting spears from Germany. Nature 385, 807–810.

Tobias, P.V., 1985. The negative secular trend. Journal of Human Evolution 14, 347–356.

Trinkaus, E., 1981. Neanderthal limb proportions and cold adaptation. In: Stringer, C.B. (Ed.), Aspects of Human Evolution. Taylor and Francis, London, pp. 187–224.

Trinkaus, E., Stringer, C.B., Ruff, C.B., Hennessy, R.J., Roberts, M.B., Parfitt, S.A., 1999. Diaphyseal cross-sectional geometry of the Boxgrove 1 Middle Pleistocene human tibia. Journal of Human Evolution 37, 1–25.

Trinkaus, E.T., 2009. The human tibia from Broken Hill, Kabwe, Zambia. Paleoanthropology, 145–165.

Walker, A., Leakey, R. (Eds.), 1993. The Nariokotome *Homo erectus* Skeleton. Harvard University Press, Cambridge.

Walker, M.J., Ortega, J., Lopez, M.V., Parmova, K., Trinkaus, E., 2011a. Neandertal postcranial remains from the Sima de las Palomas del Cabezo Gordo, Murcia, Southeastern Spain. American Journal of Physical Anthropology 144 (4), 505–515.

Walker, M.J., Ortega, J., Parmova, K., Lopez, M.V., Trinkaus, E., 2011b. Morphology, body proportions, and postcranial hypertrophy of a female Neandertal from the Sima de las Palomas, southeastern Spain. Proceedings of the National Academy of Sciences of the United States of America 108 (25), 10087–10091.

Weaver, T.D., 2003. The shape of the Neandertal femur is primarily the consequence of a hyperpolar body form. Proceedings of the National Academy of Sciences of the United States of America 100 (12), 6926–6929.

Wheeler, P.E., 1992. The influence of the loss of functional body hair on the water budgets of early hominids. Journal of Human Evolution 23, 379–388.

White, T.D., Asfaw, B., DeGusta, D., Gilbert, H., Richards, G.D., Suwa, G., Howell, F.C., 2003. Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. Nature 423 (6941), 742–747.

White, T.D., Suwa, G., Hart, W.K., Walter, R.C., Woldegabriel, G., Deheinzelin, J., Clark, J.D., Asfaw, B., Vrba, E., 1993. New discoveries of *Australopithecus* at Maka in Ethiopia. Nature 366 (6452), 261–265.

WHO, 1995. Physical Status: The Use and Interpretation of Anthropometry. World Health Organization, Geneva.

Zipfel, B., DeSilva, J.M., Kidd, R.S., Carlson, K.J., Churchill, S.E., Berger, L.R., 2011. The foot and ankle of *Australopithecus sediba*. Science 333 (6048), 1417–1420.

Zollikofer, C.P., Ponce de Leon, M.S., Lieberman, D.E., Guy, F., Pilbeam, D., Likius, A., Mackaye, H.T., Vignaud, P., Brunet, M., 2005. Virtual cranial reconstruction of *Sahelanthropus tchadensis*. Nature 434 (7034), 755–759.

Body, History of

Jakob Tanner, Forschungsstelle für Sozial- und Wirtschaftsgeschichte, Universität Zürich, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The history of the body continues to be a sensitive subject. Although the body was considered to be a 'stable experience' (and as such, not part of cultural evolution and historical change), bodily experiences of humans were never absent from historiography. Nonetheless, a continuous, conscious, explicit reflection on the role of the body only began four decades ago across the social sciences and humanities. It was only when the static, biological, and naturalistic idea of the body started being criticized that body-related questions were addressed among historians and constituted a new field of research. As such, it is sometimes seen as an interdisciplinary meeting point of several different approaches (discourse analysis, semiotics, gender studies, cultural studies, science and technology studies, etc.); others consider it an independent subdiscipline of history. From a historico-scientific perspective, there exist a number of arguments in favor of highlighting the shift of paradigms that took place in the 1970s, and asking how a new, theoretically based examination of the body has influenced older fields of research like demography and social history. This article addresses the problem of corporeality and the two dominant approaches to the history of the body, the continuities and the shift of paradigms, current fields of research, and current controversies and prospects.

The Problem of Corporeality and the Two Approaches to the History of the Body

The pathbreaking essay 'Dialectic of Enlightenment,' first published in 1947, includes a text fragment entitled 'The interest in the body.' There, Max Horkheimer and Theodor W. Adorno argued that the 'love–hate relationship to the body' (2000: 289) and the rejection of corporeality were the hidden subtexts of modern society. This oblivion of the body was related to the long-lived dichotomy between body and soul, i.e., with the Cartesian opposition of *res extensa* and *res cogitans*, which met with such success for many centuries. The consequence was the exclusion of the body from social processes and its naturalization (Danto, 1999). As the progress of secularization accelerated in the eighteenth century, mankind learned to consider society, its institutions, and modes of representation as historical realities which man himself had called into existence. The 'modern body' however was taken for granted and considered to be outside historical evolution: As a 'stable experience' (Armstrong, 1983) it could even be defined as the antithesis of cultural change. Because it was not man-made, it could be of no interest as an object of historical interpretation. In the nineteenth and twentieth centuries, the body became a privileged object of experimental research, labor techniques, and medical interventions. In epistemological terms, it remained an immutable biological given, a bastion of nature that opposed a limit to the general historicizing of thought.

Since the middle of the nineteenth century, this 'natural body' had become the stronghold of evolutionary biology, while a number of other disciplines (anatomy, physiology, biomedicine, physical anthropology, occupational and nutritional science, hygiene, molecular biology, genetics, etc.) individually began to research this seemingly inalterable physical entity of the *Homo sapiens*, using quantitative measures, exact methods, and laboratory experiments in order to produce identification techniques (Groebner, 2007) and medical knowledge (Foucault, 2008). Safeguarding collective security and the optimization of personal fitness were the main objectives of their efforts. Individuals should be protected, rationalized, normalized, healed, or maintained in good health. The fact that the body was thought to be independent of culture and that, at the same time, the modern age was literally obsessed by the impact of carnal corporeality on both individuals and society, explains the existence of a 'love–hate relationship' toward the body, diagnosed by Critical Theory.

Against this background, the history of the body proves to be a sensitive subject. However, it does not deal exclusively with the presence of the body in history. At this point, it is necessary to define two different concepts:

- On the one hand, it must be stressed that the body was never absent from historiography. "Either 'the body' is not really a subject of its own, or it contains nearly all other subjects" (Bynum, 1996). Historical works are full of phenomena which would be impossible to describe without frequent references to the human body: Demographers work with vital statistics to produce indicators on birth rate, morbidity or mortality; some of these studies reduce the drama of life to quantitative relationships. Other authors (Imhof, 1992) combined empirical findings on the extension of the average life expectancy and the tendency toward a decline in the number of children with historico-cultural questions, in which the experience of the body and the meaning of life were omnipresent. Neither the history of political philosophy nor the history of societies and mentalities, as suggested by the Annales historians in France since the 1920s, would make any sense without the presence of the body. As early as 1938, Lucien Febvre suggested writing a history of sensitivity and of the changing 'systems of emotions' in order to understand human agency (Febvre, 1941).
- On the other hand, it is clear that a continuous, conscious, explicit reflection on the role of the body only began four decades ago in social sciences and humanities. For it is only since then that a static, biological, and naturalistic idea of the body started being criticized in a theoretically

productive way. Since then, the idea of the historicity of the body developed a transformative momentum in a newly constituted field of research. As such, it is sometimes seen as an interdisciplinary meeting point of several different approaches (discourse analysis, semiotics, gender studies, cultural studies, historical anthropology, science and technology studies, etc.), whereas others consider it an independent subdiscipline of history.

These two concepts of 'body history' give rise to frequent misunderstandings. While one party insists on the innovative potential that body history has been proved to have since the 1970s, the other party, mostly traditional social historians, dismisses it as a fashion trend that conceals the evidence that history as a science focusing on the socio-cultural interaction of human beings, has always been body history. From a historico-scientific perspective, however, there exist a number of arguments in favor of highlighting the shift of paradigms that took place in the 1970s, and asking how this new, theoretically based examination of the body has influenced older fields of research like demography, social and medical history as well as gender and environmental studies (Corbin et al., 2005).

Continuities, Shift of Paradigms, and New Topics

The historicizing of the body can be split up into five phases (which are not clearly distinct, but overlap frequently with one another).

In the first phase, which lasted until the early 1970s, a number of innovative studies looked at the human body and the imagery that had crystallized around it as objects of theoretically elaborated analyses. Later, they became 'classics' in the domain, although their initial intention had nothing to do with the new specialty named 'history of the body,' which started emerging in the 1970s. The historical studies concentrated principally on the Middle Ages and the early modern era, whereas anthropology, philosophy, and psychoanalysis dared to delve into contemporary phenomena. Some of the most significant works of this period are Sigmund Freud's *Civilization and Its Discontents* (1929/30), whose central theme is the extension of the body with the help of technical artifacts and the promotion of the human being to a 'prosthetic God' – a thesis further developed by Marshall McLuhan with his understand of media as 'extensions of man'; Marcel Mauss' *Techniques of the Body* (1935), in which the body is analyzed as the first and most natural instrument at man's disposal; Paul Valéry's *Problem of the Three Bodies* (1943), which looks into the various, parallel ways of speaking of the body; Maurice Merleau-Ponty's *Phenomenology of Perception* (1945); as well as Samuel Todes *The Human Body as Material Subject of the World* (written in 1963), which deal with the body as the condition of awareness, in compliance with a philosophy of finiteness. Ernst H. Kantorowicz' in *The King's Two Bodies* (1957), examines the political theology of the Middle Ages and looks at the mystical fiction of an immortal 'political body' existing in parallel to the mortal 'natural body' of the King.

A second phase began in the 1960s and was blossoming in the 1970s. Theoretical thinking was accompanied by the removal of the body's taboos in the alternative culture of the time. Remaining in a traditional trajectory, social history focused on the medical profession whose power was imposed on the bodies of broad strata of the population. This view of a 'medicalization' (analyzed both from the top down and from the bottom up) was challenged by Michel Foucault, who epistemologically dissipated the 'medical,' as perceived by social historians, into a 'microphysics of power.' Particularly with *The Birth of the Clinic* (1963) and *Discipline and Punish: The Birth of the Prison* (1975) Foucault became the deciding authority of a new approach. Another seminal contribution was Mary Douglas' *The Two Bodies* (1970), in which the author examines the metaphorical relationships and cognitive interactions between the body as a social construction and the individual physical body. The new studies of the human body were based on the thesis of a constant interaction between knowledge and power. In the late 1970s, Foucault's concepts of a bio-power and of a political economy of the body were set up against a liberal-Marxist scenario of disciplinary force, which represses the body. Contrary to the assumption of a 'disciplinary society' (in which the subject is oppressed by power), in Foucault's view the body is produced by knowledge-based power, which shapes everyday practices and generates regimes of visibility and discourses. In the concept of a 'normalizing society,' the body is the medium which binds together the macroorganization of power and the micro-practices through which social structures are stabilized. This perspective is directly connected to Norbert Elias' opus magnum, *The Process of Civilization*, which was published in the late 1930s but met with significant success only three decades later. Elias promoted the concept of the individual internalization of social constraints ('force to force oneself'). The 'cultural production' and the 'social shaping' of a modern human being is defined by Elias in terms of an ever-growing, individual self-control, mainly through the decoupling of emotions and bodily expressions. The 1970s were far away from being homogeneous. They also saw the publication of interesting studies in various domains of social sciences, notably by Helmuth Plessner (philosophical anthropology of the senses); Heinrich Schipperges (outlines for a philosophy of the body); Dietmar Kamper and Christoph Wulf (return of the body and fading of the senses); Rudolf zur Lippe (on the own body) which deals with 'the economy of life'; and many others.

The third phase, the displacement of the 'social' by the 'cultural' and the concomitant attention to the body, gained momentum. This shift was strongly influenced by gender studies and the feminist criticism of science, which admitted gender as a category for analysis and introduced new standards into body history. Thanks to Yvonne Verdier, Carolyn Bynum, Ludmilla Jordanowa, Marie-Christine Pouchelle, and many others, new interpretations of the body appeared at the end of the 1970s, emphasizing its relation to gender history and clearing the ground for a far-reaching, methodologically highly interesting discussion. In Germany, these endeavors are connected above all to the name of Barbara Duden, who published an innovative study, *The Woman Beneath the Skin: A Doctor's Patients in Eighteenth-Century Germany*, in 1987. In this book about the interior, invisible corporeality of women in Eisenach (Germany) and their doctor Johann Storch, the historian and sociologist Barbara Duden examines the sociogenesis of the modern body in relation to a history of the body's receptivity to

experiences. She analyzes the 'loss of control over the body' as a gate toward physical experience and speaks in favor of a "hermeneutic interpretation of the body's reality in the past": Although we cannot use 'our body' (which we have 'possessed' since the eighteenth century), as a bridge to interpret the experiences of people before the Enlightenment, the corporeality of every human existence gives us access to the context in which women and men experienced feelings in the past.

The fourth phase is characterized by a turn toward discursive, semiotic, and performative approaches and marks the beginning of a new era in the history of the body, which also led to a post-structuralist reassessment of key authors of the 1960s, especially of Foucault, whose impact on cultural theory was deepened considerably after his death in 1984 (Fassin and Memmi, 2004). The 90s saw new authors coming up: in particular, Thomas W. Laqueur's *Making Sex* (1990) and Judith Butler's *Bodies That Matter* (1993) fulfilled the linguistic turn in body history and provoked the supporters of the traditional, experience-based study of corporeality. Laqueur examines the constitution of the body through discourse and action; from this point of view, sex is "situational; it is explicable only within the context of battles over gender and power." In contrast, Butler – who, after her book was published, was wrongly considered to be an advocate of 'disembodiment' – criticized the subject-centered constructionist approach by relating the materiality of the body to the performativity of gender: "What I would propose in place of these conceptions of construction is a return to the notion of matter (…) as a process of materialization that stabilizes over time to produce the effect of boundary, fixity, and surface we call matter" (Butler, 1993; Lorenz, 2000). Notably the debate on AIDS provoked a crisis of representation and fostered a new sensibility of the materiality of the human body (Cooter and Stein, 2008).

Since the turn of the century (the fifth phase) a 'somatic turn' has questioned the hitherto antiessentialist and representationalist understanding of the body. As Chapter 5 will argue, the growing tendency to criticize the "dissolving of the body into speech" (Bynum, 1996) is however not equipollent to a relapse in an unreflecting neoessentialism of corporeality but rather the recognition of the uncircumventable materiality of the body (Cooter, 2010).

Fields of Research and Types of Analysis

For a long time, researchers in the field of body history were not sure whether they were looking into a completely new domain or whether they had simply discovered a new and interesting perspective, which offered the possibility of an innovative approach toward well-known social phenomena. The question becomes more complex: from the very beginning, body history was the fruit of interdisciplinary efforts, which can have two consequences – namely losses of clarity due to frictions between the disciplines; and the positive effects of synergy (cf among others the *Fragments for a History of the Human Body*, Feher, 1984). Authors who had already attempted a systematic approach to body history in the mid-1980s presented an overview of the broad portion of research based on social sciences. From a phenomenological point of view, Bryan Turner (1984) describes four tasks with which any social organization confronts bodies (and vice versa): (1) the reproduction of populations in time, (2) the regulation of bodies in space, (3) the constraints of the 'interior' body through control and discipline, and (4) the representation of the 'exterior' body in the symbolic space of the society. John O'Neill's *Five Bodies* (1985) provides an alternative, more institution-based conceptualization, distinguishing (1) the world's body, (2) social bodies, (3) the political body, (4) consumer bodies, and (5) medical bodies.

In the narrower field of historiography, although there have been frequent variations in the main fields of interest, one can observe an obvious concentration of research on the Middle Ages and the Early Modern Era. Starvation and fasting, posture, body language, clothing habits, disciplining of the human body through the art of dancing, training of soldiers, seduction, sexuality, pregnancy and childbearing: these and other themes were examined in the light of cultural anthropology and iconography and resulted in new views on everyday life and bodily experiences in past centuries. The *Education of the Senses* (Peter Gay), the history of *Sensory Awareness* (Rudolf zur Lippe), and an *Anthropology of the Senses* (Febvre, 1941) turned out to be especially interesting fields of research. The conflict about the hierarchy of the senses, which began in antiquity, the theses of a 'mythopoetical constitution of the individual' (Hartmut Böhme) and of a 'privilege of the eye' in the present time (Virilio, 1996) remain of symptomatic value. The history of sensory experience is closely linked with the history of the 'wisdom and the remembering of the body': as a materialized matrix for inscriptions of cultural experience, the body is a part of the memory of a society. Step by step, performativity relates the signification process with the materiality of the body. It is the reiteration of practices and the regulation of discourses which generate the feeling of continuity and reliability of the 'real world.'

Another area that offers many possibilities for research is the history of the perception of the body and of body-related imagery and esthetics. In the course of secularization, the image of the body as a 'wheelwork,' whose perfect functioning was the proof of God's Creation, was replaced by the idea of a machine working according to the laws of nature (with a male connotation: the man is considered to be the standard human being). Along with this evolution came the growing conviction that the body is a closed entity and the medieval image of a 'grotesque,' open body (Bakhtin, 1981/1930) was gradually replaced by the modern, closed 'Skin-Self' (Anzieu, 1989). This evidence depended on new schemes of visibility: Barbara Stafford describes this historical shift toward visualization with the help of somatic metaphors and she points out the important role played by the visual arts when it came to solving the representational problem of the embodiment or personification of knowledge. Crucial for this development was the eighteenth century, which saw the breakthrough of a 'possessive individualism' in the perception of the body. This resulted in a "degradation of the notion of a self extended into a unique and inviolable corporeal volume, to one in which the self only loosely possessed a body" (Karl Figlio). The problem for the historical investigation into this body, which we possess and which is analyzed by science, "is not how something that is obvious today had remained hidden for so long, but how the body had become so evident in the first place" (Armstrong,

1983). Thanks to the modern perceptions of the body, new ways of seeing collective bodies were also brought to the fore. Jacob Rogozinski (2010) showed that the incarnation of the individual in an all-embracing global body is an old phenomenon, but in the nineteenth and twentieth centuries, the political hypostatization of this collective body became still stronger. While society depended more and more on technology, strengthening the thesis that social cohesion results from the linking-up of the elements of society by a powerful communicative 'nervous system,' nationalism encouraged the phantasm of a 'body of the people' and of a 'blood community.' For example, the imagery of the *Body of the Nation* (Goltermann, 1998) proved to be mediated by the practice of male gynmastics, which aims at the appropriation of the national order through individuals and social groups.

Medicine, dietetics inspired from humoral theory and the modern conception of hygiene, which emerged in the middle of the nineteenth century, also offer fertile ground for research in body history. Traditional, epidemiological approaches of research on disease are completed by analyses of illness, suffering, pain, and death of human beings, who were increasingly considered as patients in the course of the nineteenth century, and thus submitted to medical experts' power of definition.

A similar process of professionalization and scientization took place in the areas of occupational and nutritional science. The interaction of physiology and the industrial society show, for instance, how epistemic regimes, technological artifacts, and social interests are intertwined in the production of knowledge; and how the knowledge thus produced facilitates the appearance of individuals who consider their bodies as 'instruments of self-perfection' (Sarasin, 2001). In this context, the modification of individuals' self-images under the influence of scientific knowledge is studied in parallel with the unwillingness of the body faced with the exigencies of a working society, which is subject to the paradigms of efficiency and rationalization. In the leisure society of the postwar period, in which the body functions as a site of consumption, the questions arising from an invariably incomplete integration of the individual body into the social structure remain to be answered by further historical research.

Body History Reloaded in 2000

Body history was always entangled by normative judgments and was the object of emotional stakes. Generally, in social sciences and humanities, experience-based concepts of the body implying a lived corporeality are being challenged by poststructuralist–deconstructivist models that consider the body primarily as a discoursive effect or a product of the knowledge-power-complex accumulated since the Enlightenment. Nonetheless, around 2000, there was a consensus that it was essential to overcome the opposition of 'discourse' and 'experience' (Canning, 1994). These debates frequently overlook the fact that the history of the body is in the midst of the fundamental discussion on the blurring of the boundaries between nature and culture. The ontological status of the traditional categories 'natural' and 'social' has become a critical question and this makes it obvious to consider the products of technology – and therefore the world we live in – as hybrid artifacts.

From this point of view, the human body, too, would be a hybrid, i.e., a 'crossbreed' in which social, technological, cultural, and natural components cannot be held apart with certainty. The materiality of the body plays a crucial role in this debate. This is obvious in disciplines, such as biomedicine, cognitive sciences (CS), and artificial intelligence (AI), which lie in a paradoxical area of tension between hyperembodiment and disembodiment: while scientists in the domain of AI start out from the assumption that intelligence can be modeled only on a material, body-bound, base, which implies that machines that have 'emerging properties' and therefore the ability of learning must by definition be 'techno-bodies' (Lakoff et al., 1999), and other authors (e.g., Flusser, 1994) follow the principle 'from the subject to the project' to develop visions according to which the concept of the body has become a creative activity of mankind, escaping the 'blind contingencies of nature.' It is essential to realize that information technology, genetic engineering, and xenotransplantation have made natural boundaries permeable and put mankind in a position to gain control over the evolutionary process of life, which started billions of years ago.

It is precisely because of these perspectives, which are technocratic and playful at once, that the findings of body history have gained capital importance. Although they are incomparable for various reasons, the concepts of the body suggested by biomedicine, information technology, and genetic engineering are in compliance with the more modern approaches of the social sciences and humanities concerning the thesis that the body is not an invariable, unchangeable entity, independent of time and history. The willingness to reassess the body as a product of society and the readiness to think about how bodies are culturally constructed or constituted, became a salient feature of historical research on the body. If historiography demonstrates that bodies were always modeled by culture, this means that scientific and technological interventions on the physiology and (for the time to come) in the genome of the human being are nothing new in principle: they are simply an intensification of a traditional way of using bodies.

In this sense, Haraway (1991) pleads in favor of a curious and innovative application of the new possibilities offered by technology; she sees the creation of natural–cultural hybrids as a promising path toward the creative self-realization of mankind. This 'Cyborg-optimism' has been contradicted. Distancing herself from technocratic feasibility fantasies, Barbara Duden proposes that although human beings are partially formed by culture, the somatic fundaments of life are an inescapable condition of their existence. Based on the paradigm of a 'historical somatology,' the examination of the 'time-bound forms of experiencing corporeality' concentrates on the relation of human beings to their body, which changes in the course of history, as a *sine qua non* of experience. Such an approach considers the current colonization of the body by science and technology as an unprecedented process and an attack on the human condition itself. Duden remains skeptical of the *mythopoesis* of a technology which models the desires of individuals in such a way that they anticipate the possibilities of science and technology.

New Trends and Topics in the Twenty-First Century

Since the advent of the twenty-first century the problems of how to understand the intricate process of the cultural construction of reality and investigate the historicity of the human body have been radicalized. Two tendencies are at the center of attention.

On the one hand the neoliberal mode of the subjectivization strengthened both the evaluation of the educational attainments and the state of the body in terms of 'Capital' (and consequently, as a problem of 'pay-back periods' and 'return on investment'). The popular resonance of the human capital theory, which was launched in economic science in the early 1960s, became tremendously successful after 1980, paralleled by an increased personal 'will to health.' The human being became cognitively and corporeally an investment sphere, a development which has been problematized in different and conflicting ways. In the first decade of the twenty-first century, the transformation of the Pursuit of Happiness (as the rationale of the 'American Way of Life,' which was imitated on a worldwide scale after World War Two) into a Pursuit of Perfection (Rothman and Rothman, 2003) became tangible, encompassing medical-enhancement techniques using surgical procedures and drugs (Serlin, 2004). The phantom of perfectibility embraces the body which becomes a malleable entity subjected to a history of its own. A large number of studies in the history of science and in art history have tracked down the endeavor to realize an optimal physiological functioning and the best possible esthetic performance of the human body (e.g., Hau, 2003).

On the other hand, and thus related, a comeback of an essentialist approach can be observed. This somatic turn was a swing reaction against the linguistic turn. Against the representationalism, put forward by the constructivist theories of the 1980s and the 1990s, emphasis was now laid on physicality, related with materiality, experience, and practice. Such a view was already proposed (earlier than its time) in the 1980s by Barbara Duden and Caroline Bynum who criticized the trend toward de-essentialization. The correlate of the emergence of a new regime of truth based in corporeality, was the interest in the 'really lived body' (Ruggiero, 2004).

These two shifts toward subjectivation and soma have theoretically valorized Foucault's concepts of the technologies of the self and of biopolitics (Foucault, 2008). Since the 1960s, Foucault has advocated a constructivist perspective on society; mean while he has criticized the reduction of reality to language and symbolic operations. He thus combined in a seminal way a constructivist approach to society and an ontological view of history (Cooter, 2010; Smith, 2007; Hacking, 2002).

This also allowed the historicizing of the relationship between personal identity and body concepts. Experienced corporeality is no longer considered to be an *explanans* to self-perceptions, but an *explanandum* i.e., something that has to be explained. The nature–nurture dichotomy and the normative juxtaposition of salvific progress and frightful threat is subverted by a new understanding of the ambiguous effects of technologies. Nicolas Rose' concept of a 'Biocitizenship' should not be confounded with biological determinism. Instead of being a natural determinant, biology turns out to be a moral commitment. People are expected to comply self-paced with the possibilities and promises of modern medicine and genetics. The crux of such utilitarian behavior is that on an aggregate level, the mode of the free choice produces flexible standards and moving norms, thereby retroacting on personal decision-making. When citizens are required to take self-governed responsibility for their genetic constitution and their neurochemical selves, then every individual who rejects that onus becomes a burden on society. Under such conditions, the individual person is transformed into a competitive entrepreneur of him/herself and can no longer argue the unavailability of his/her body. If he/she is not willing to fulfill his/her duty, he/she will be socially discriminated or penalized financially. On a normative level, a fusion of body, esthetics, and morality will be brought into operation.

As to the areas of historical and interdisciplinary research, five lines of research have become important since the beginning of the twenty-first century. First, there is a focus on the ecology of the human body that is analyzed in its multifarious interactions with its environment. Architecture, settlement patterns, urban designs, social structures, mobility behavior, and environmental pollution become important categories of an analysis of changes in perceptions of the body and bodily performances in history (Floyd-Wilson and Sullivan, 2007). A second focal point is post-colonial studies, especially the concept of 'critical whiteness' which refers to body-related perceptions of both the colonists and the colonized (Lopez, 2005; Olsén, 2011). An analogue problem is addressed with the term 'subaltern,' suggested by Gayatri Spivak to denominate the ambiguous forms of body-mediated political power. Spivak (2010) is critical toward an unhistorical essentialism. Nonetheless she advocates the use of a 'strategic essentialism,' i.e., a form of self-assertion in asymmetric power relations that propagate one 'We,' thereby trying to accept the diversity of voices in each group. In the third field we find disability studies examining the changing social interaction and the normative reciprocity between 'normal' and 'handicapped' individuals from the perspective of the latter (Watson et al., 2012). The growing acceptance of sign language demonstrates a sustainable shift in the definition of disability and a normalization of the status of people hitherto judged to be 'disabled.' In including categories like 'monsters' and 'freaks' (Stammberger, 2011), the field offers deep insights in the changing tangible and invisible restrictions, respective limitations contained in the norms and institutions of a society, as well as the dynamics of access and exclusion in all social institutions.

The fourth and fifth foci are closely related. The fourth relates to Robotics, which represents both an expanding field of current electronic-mechanical research and a topic of historical inquiry. The thesis that human intelligence and feeling are 'embodied,' has led the designers of robots away from 'clever algorithms' to morphological designs and materials. The 'Companions for Citizens'-project is aimed at the development of humanoids, which combine a high degree of thinking with (simulated) emotional traits and become reliable and bearable helpers in many everyday situations,

especially for older people. Concomitantly, the history of androids, robots, and golems has received a new boost and the long-lasting visions of a man-made man or an artificial invader (from the Golem via Faust/Frankenstein to the Blade Runner replicants, etc.) have achieved an upturn (e.g., De la Peña, 2003; Muri, 2007).

The most provoking (the fifth) research area is neuroscience which includes not only highly sophisticated Magnetic Resonance Imaging (MRI), but also the proliferation of colorful brain pictures in mass media and popular culture. From the outset, these widespread and inherently contradictory research approaches have been confronted with a historically and philosophically informed reflection on a materialistic monism (or a monistic materialism) (Danto, 1999; Hagner, 2008). A strong argument is bound to the fact that the most popular criticism of Cartesian dualism lost its innocence long ago. Already in the late nineteenth century, an antirationalist, rather occult 'holistic approach,' combined with a dull rhetoric of 'unity' has threatened Cartesian rationalists. At the turn of the twenty-first century, reductionist concepts have taken its place. In *Mind Time* (2004) the neuroscientist Benjamin Libet argues against his false followers that there is no sound reason for reducing consciousness on a neural substrate which is a necessary, though not a sufficient condition for symbolic interaction and language. Cultural evolution cannot be wrapped up either in neural encoding/decoding nor in neural circuits. The assumption that cognitive capacities are nothing more than 'ghosts in the machine' reproduces instead in a paradoxical way, the Cartesian dualism.

The body history problem needs a more subtle sublation of this binary opposition which also aims at overcoming the deeply ingrained nature–culture distinction. Already now, two points are clear: in the first place, the plasticity and malleability of the human brain supports the assumption that the neuronal substrate is shaped by social practices, cultural techniques, and symbolic codes. Learning and exercising changes cerebral networks to a degree that makes the idea of a 'natural brain' fallacious. Research projects using functional imaging techniques may show 'what happens where' in the brain. But neither the relationship between structure and function nor the temporal alteration of such operations can be adequately apprehended so far. There is no reason to believe that forthcoming neurobiological research will end up in a biological essentialism. Secondly it is obvious that new approaches of a body history can only be developed via a crosstalk involving philosophy, cultural anthropology, theory of emotions, and brain research.

Altogether the body and its 'polyphonic history' (Peter Burke) remains in the twenty-first century also, a preferential site for interdisciplinary studies and cultural analysis. The love–hate relationship mentioned by Horkheimer/Adorno is still present in many practices and discourses. There is however an endeavor to accept challenge, both epistemologically and empirically, of writing the history of the body in a more serene way, thereby focusing on the concrete, mortal human body, without forgetting that the corporeal existence of human beings will never be independent of the empire of signs, which is permanently created and transformed by cultural change.

See also: Affect and Emotion, Anthropology of; Body Image; Body: Anthropological Aspects; Culture and Emotion; Emotion: History of the Concept; Emotions, History of; Emotions, Sociology of; Senses, Anthropology of the; Social and Emotional Development in the Context of the Family.

Bibliography

Anzieu, D., 1989. The Skin Ego. Yale University Press, New Haven.
Armstrong, D., 1983. Political Anatomy of the Body: Medical Knowledge in Britain in the Twentieth Century. Cambridge University Press, Cambridge.
Bakhtin, M., 1981/1930. In: Holquist, Michael (Ed.), The Dialogic Imagination: Four Essays. University of Texas Press, Austin and London.
Butler, J., 1993. Bodies That Matter. On the Discursive Limits of Sex? Routledge, New York/London.
Bynum, C.W., 1996. Warum das ganze Theater mit dem Körper? Die Sicht einer Mediävistin. In: Historische Anthropologie (6. Jg. No. 1), pp. 1–33.
Canning, K., 1994. Feminist history after the linguistic turn: historicizing discourse and experience. Signs 19, 368–404.
Cooter, Roger, mayo-junio, 2010. The turn of the body. History and the politics of the corporeal. ARBOR Ciencia, Pensamiento y Cultura 436, Nr. 743, S. 393–405.
Cooter, R., Stein, C., 2008. Coming into focus: posters, power, and visual culture in the history of medicine. Medizinhistorisches Journal 42, 180–209.
Corbin, A., et al. (Eds.), 2005–06. Histoire du corps, 3 vols. Éditions du Seuil, Paris.
Danto, A.C., 1999. The Body/Body Problem. Selected Essays. University of California Press, Berkeley.
De la Peña, Carolyn Thomas, 2003. The Body Electric: How Strange Machines Built the Modern American. New York University Press, New York.
Douglas, M., 1973. Natural Symbols. Exploration in Cosmology. Barrie & Jenkins, London.
Duden, B., 1987. Geschichte unter der Haut. Ein Eisenacher Arzt und seine Patientinnen um 1730. Klett-Cotta, Stuttgart.
Elias, N., 1969. Über den Prozess der Zivilisation: soziogenetische und psychogenetische Untersuchungen. Francke, Bern.
Elkins, J., 1999. Pictures of the Body Pain and Metamorphosis. Stanford University Press, Stanford.
Fassin, D., Memmi, D. (Eds.), 2004. Le Gouvernement des corps. EHESS, Paris.
Febvre, L., 1941. La sensibilité et l'histoire: comment reconstituer la vie affective d'autrefois? In: Annales d'histoire sociale Nr. 3, S. 5–20.
Feher, Michel (Ed.), 1984. Fragments for a History of the Human Body, vol. 3. Urzone, New York.
Floyd-Wilson, M., Sullivan Jr., G.A. (Eds.), 2007. Environment and Embodiment in Early Modern England. Palgrave Macmillan, Basingstoke.
Flusser, V., 1994. Vom Subjekt zum Projekt: Menschwerdung. Bollmann, Bensheim/Düsseldorf.
Foucault, M., 1977. Discipline and Punishment: The Birth of the Prison, 1979-ed. Penguin Books, London.
Foucault, M., 1978. The History of Sexuality, 1979-ed. Penguin Books, New York.
Foucault, M., 2008. In: Senellart, Michel (Ed.), The Birth of Biopolitics: Lectures at the Collège de France, 1978–79. Palgrave Macmillan, Basingstoke.
Goltermann, S., 1998. Körper der Nation: Habitusformierung und die Politik des Turnens 1860–1890. Vandenhoeck und Ruprecht, Göttingen.
Graduiertenkolleg Sozialgeschichte Bielefelder (Ed.), 1999. In: Körper Macht Geschichte – Geschichte Macht Körper. Verlag für Regionalgeschichte, Bielefeld, pp. 35–57.
Groebner, V., 2007. Who Are You? Identification, Deception, and Surveillance in Modern Europe. Zone Books, New York.
Hacking, I., 2002. Historical Ontology. Harvard University Press, Cambridge, MA.
Hagner, M., 2008. Homo cerebralis. Der Wandel vom Seelenorgan zum Gehirn. Frankfurt a.M.
Haraway, D., 1991. Simians, Cyborgs, and Women. The Reinvention of Nature. Routledge, New York.
Hau, M., 2003. The Cult of Health and the Beauty in Germany. A Social History, 1890–1930. University of Chicago Press, Chicago.
Horkheimer, Max, Adorno, Theodor W., 2000 [1947]. Das Interesse am Körper. In: Dialektik der Aufklärung: philosophische Fragmente, first ed. Frankfurt a.M.: Fischer, pp. 288–293.

Imhof, A.E. (Ed.), 1992. Leben wir zu lange? Die Zunahme unserer Lebensspanne seit 300 Jahren – und die Folgen. Böhlau, Köln.
Jones, C., Porter, R., 1994. Reassessing Foucault. Power, Medicine and the Body. Routledge, London/New York.
Kamper, D., Wulf, C. (Eds.), 1982. Die Wiederkehr des Körpers. Suhrkamp, Frankfurt.
Kantorowicz, E.H., 1957. The King's Two Bodies. A Study in Mediaeval Political Theology. Princeton University Press, Princeton.
Lakoff, G., Johnson, M., Rothstein, E., 1999. Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. Basic Books, New York.
Laqueur, T., 1990. Making Sex, Body and Gender from the Greeks to Freud. Harvard University Press, Cambridge, MA.
Libet, B., 2004. Mind Time. The Temporal Factor in Consciousness. Harvard University Press, Cambridge, MA.
Lopez, A.J., 2005. Postcolonial Whiteness. A Critical Reader on Race and Empire. State University of New York Press, Albany, NY.
Lorenz, M., 2000. Leibhaftige Vergangenheit. Einführung in die Körpergeschichte. Edition Diskord, Tübingen.
Mauss, M., 1989. Die Techniken des Körpers, first ed. 1935. Soziologie und Anthropologie, vol. 2. Fischer-Taschenbuch-Verlag, Frankfurt/M, pp. 197–220.
Merleau-Ponty, M., 1966. Phänomenologie der Wahrnehmung. de Gruyter, Berlin.
Muri, A., 2007. The Enlightenment Cyborg: A History of Communications and Control in the Human Machine, 1660–1830. University of Toronto Press, Toronto; Buffalo, NY.
Öhlschläger, C., 1997. Körper-Gedächtnis-Schrift. Der Körper als Medium kultureller Erinnerung. Schmidt, Berlin.
Olsén, J.E., 2011. The body voyage as visual representation and art performance. Nuncius (1), 222–241.
O'Neill, J., 1985. Five Bodies. Cornell University Press, Ithaca, NY.
Perrot, P., 1984. Le corps féminin, XVIIIme–XIXme siècle. Le travail des apparences. Edition du Seuil, Paris.
Revel, J., Peter, J.-P., 1974. Le corps. L'homme malade et son histoire. In: LeGoff, J., Nora, P. (Eds.), Faire de l'histoire. Nouveaux objects. Gallimard, Paris, pp. 169–191.
Rogozinski, J., 2010. The Ego and the Flesh: An Introduction to Egoanalysis. Stanford University Press, Stanford.
Rose, N., 2007. The Politics of Life Itself. Biomedicine, Power, and Subjectivity in the Twenty-First Century. Princeton University Press, Princeton/Oxford.
Rothman, S.M., Rothman, D.J., 2003. The Pursuit of Perfection. The Promise and Perils of Medical Enhancement. Pantheon Books, New York.
Ruggiero, K., 2004. Modernity in the Flesh. Medicine, Law and Society in Turn-of-the-Century Argentina. Stanford University Press, Stanford, California.
Sarasin, Ph, 2001. Reizbare Maschinen. Eine Geschichte des Körpers 1765–1914. Suhrkamp, Frankfurt/M.
Schreiner, K., Schnitzler, N. (Eds.), 1992. Gepeinigt, begehrt, vergessen. Symbolik und Sozialbezug des Körpers, im späten Mittelalter und in der frühen Neuzeit. Fink, Munich.
Serlin, D., 2004. Replaceable You. Engineering the Body in Postwar America. University of Chicago Press, Chicago.
Smith, R., 2007. Being Human: Historical Knowledge and the Creation of Human Nature. Manchester University Press, Manchester.
Spivak, G. Ch, 2010. In: Morris, Rosalind C. (Ed.), Can the Subaltern Speak?: Reflections on the History of an Idea. Columbia University Press, New York.
Stafford, B.M., 1993. Body Criticism. Imaging the Unseen in Enlightenment Art and Medicine. MIT Press, Cambridge, MA.
Stammberger, B., 2011. Monster und Freaks: eine Wissensgeschichte außergewöhnlicher Körper im 19. Jahrhundert. Transcript, Bielefeld.
Todes, S., 2001. Body and world. Cambridge/Mass./London (with introductions by Hubert L. Dreyfus and Piotr Hoffman), rev. ed. (1990). In: The Human Body as Material Subject of the World. Garland, New York (Thesis 1963).
Turner, B., 1984. The Body and Society. Blackwell, Oxford.
Virilio, P., 1996. Die Eroberung des Körpers: vom Übermenschen zum überreizten Menschen. Fischer-Taschenbuch-Verlag, Frankfurt/M. München/Wien.
Watson, N., et al. (Eds.), 2012. Routledge Handbook of Disability Studies. Routledge, London [u.a.].

Body Image

Gemma L Tatangelo, Marita P McCabe, and Lina A Ricciardelli, School of Psychology, Deakin University, Burwood, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Body image is a multifaceted construct which is central to an individual's self-concept, and critical to mental health and general well-being. Given our increasingly appearance-focused culture, body image concerns are now reported by both women and men across the life span. This article examines the different components of body image and prevalence of body dissatisfaction, and provides examples of measures that have been developed to assess these constructs. It also examines the main factors that shape body image related to age and gender. The nature of programs to prevent body image problems, particularly at school and policy levels, is also discussed.

Introduction

Body image is a multidimensional construct encompassing perceptions, attitudes, beliefs, feelings, and behaviors regarding one's appearance (Cash, 2004). Over a century of research on the topic has advanced the field of body image from its initial conception as a means of describing neuropathological observations in brain-injury patients, to current conceptualizations that comprise perceptual, attitudinal, cognitive, and behavioral facets. Perceptual body image refers to the accuracy with which individuals perceive their body size, whereas the attitudinal or affective component of body image relates to the satisfaction and feelings that individuals have toward their body and appearance. The cognitive dimension refers to the thoughts and beliefs that individuals have concerning their body shape and appearance. Lastly, the behavioral component examines specific behaviors, such habitual body checking, and avoidance of mirrors and public situations (Thompson and van den Berg, 2002).

Body image is not an objective evaluation of one's appearance or body. Rather, it is highly subject to change, and is shaped by the acquisition of new information (Grogan, 2008). As such, the 'image' held of one's body may be quite different from the reality of a person's body shape. A disparity between an individual's perceived body shape and his or her ideal body shape or size is proposed to result in body image dissatisfaction. It can range from a mild dislike of a particular body characteristic to intense distress associated with a desire to change the body. It may focus on a particular part of the body, or may relate to the overall appearance, weight, shape, or muscularity of the body.

To fully understand the multifaceted nature of body image, it is necessary to look beyond the individual, and understand the social context underlying the development of body image. In Western cultures, a great deal of importance is placed on physical appearance. This has been the case for women for many years, and in the last 20 years has increasingly become the case for men. Individuals are constantly faced with sociocultural pressures that promote and reinforce a narrow ideal of what is considered attractive and desirable. These sociocultural messages create unattainable standards for most individuals to achieve. This is a major reason why a substantial proportion of women and girls, and increasingly men and boys, develop negative feelings about their bodies. It is also a reason why body dissatisfaction is now one of the most significant health concerns for many individuals living in Western countries.

Body image plays a critical role in the development of both physical and psychological health. While positive body image fosters mental health and general well-being, poor body image has been implicated as a major contributor to a number of negative behavioral and psychological problems. Body dissatisfaction is an important contributor to various body-change behaviors ranging from minor weight control methods, such as dieting and increased exercise, to complete restriction of food intake. In fact, body dissatisfaction has emerged as the most consistent contributor to disordered eating. Several other potentially harmful behaviors have also been associated with body image concerns, including use of anabolic steroids and invasive cosmetic surgery. Additionally, body image concerns are strongly related to poor self-esteem, the effects of which can be significantly detrimental to overall well-being.

Prevalence of Body Dissatisfaction

Recent studies have shown that in Western societies such as the United States, as many as 80% of women experience body image dissatisfaction in regard to some aspects of their appearance. It is estimated that over 70% of women and adolescent girls desire a thinner body shape and over 60% report having taken steps toward losing weight. Even preadolescent girls as young as 5 years of age demonstrate body dissatisfaction, with studies showing that about 40–50% of them report a desire to be thinner.

The last two decades have seen a growing awareness that body image concerns are not unique to women and girls. Recent research indicates that around 15–60% of adolescent boys and young men also experience body dissatisfaction. The level of body dissatisfaction among boys and men has often been underestimated, as many studies have only examined body image concerns that are more relevant to women with the focus being on weight and lower body parts. A critical aspect of research with men's and boys' body image is their focus on muscle as well as on weight. Specifically, among adolescent boys, it is estimated that 21–47% have used strategies to increase muscle and/or weight, and 13–48% have used

strategies to lose weight. Studies have also indicated that preadolescent boys demonstrate similar body concerns to adolescent boys and men. Ricciardelli et al. (2006) found that 47% of boys aged 8–11 years desired a thinner body size and 21% desired a larger body size. In addition, close to one-third of boys frequently engaged in weight loss strategies, while almost half of the boys frequently engaged in muscle gain strategies.

Measurement of Body Image

One of the most common methods to measure body dissatisfaction is by using figure rating scales. Figure rating scales consist of a series of human silhouettes, which range from large to small weights, shapes, and sizes. Although different sets are available, they typically consist of figures ranging from very thin to obese. Individuals are instructed to select the figure which they believe best represents their current body size and shape, and then select the figure which they would ideally like to look like. The discrepancy between the two selections reflects the degree of body dissatisfaction. An important feature of figure rating scales is that they have been adapted to suit the target population and also the attribute under inquiry. For instance, there are different versions for women, men, adolescents, and children. Child-versions of figure rating scales are the most common tools for the assessment of body dissatisfaction among children. They are particularly well suited for use with children as the instructions can be easily explained and understood, and they can be completed in a short time. The problem with these rating scales is that although they may demonstrate a difference between the person's ideal and perceived actual body size, they do not indicate how the person feels about this discrepancy (i.e., his or her level of negative affect in relation to the discrepancy). Further, it is difficult with these scales to adequately represent all of the different dimensions of body image (e.g., weight, shape, muscularity).

The measurement of perceptual body image involves assessing the accuracy with which individuals perceive their body size. Contemporary body estimation methods involve the use of computer software that allows participants to manipulate the size of their body and individual body parts to reflect how they perceive themselves. These computer programs have demonstrated excellent psychometric properties; however, the main limitation is that data can be collected from only one participant at a time.

Several questionnaires have also been developed to measure body dissatisfaction. Questionnaires such as the Body Esteem Scale for adolescents and adults (Mendelson et al., 2001) assess dissatisfaction with the body as a whole and tap into how an individual feels about his or her physical appearance and/or body weight. This measure was originally developed for children, and has been validated for children as young as 8 years of age.

Questionnaires that measure the affective component of body image involve the assessment of feelings of anxiety, stress, and shame about specific aspects of the body, or weight and appearance in general. For example, the Body Shame subscale of the Objectified Body Consciousness scale (McKinley and Hyde, 1996) measures feelings of shame as a result of failing to meet culturally defined appearance ideals. Measures focusing on the cognitive dimension assess beliefs, thoughts, and attritions that contribute to body dissatisfaction. The Assessment of Body Image Cognitive Distortions (Jakatdar et al., 2006) is one such questionnaire, which assesses cognitive errors or distortions relating to body image thoughts across different contexts and situations.

Fewer measures have been developed to assess the behavioral component of body image. One self-report measure is the Body Checking Questionnaire (Reas et al., 2002), which assesses body checking behaviors in three domains: overall appearance, specific body parts, and idiosyncratic checking.

In more recent years, several measures have also been designed specifically for men. These include the Male Body Attitudes Scale developed by Tylka et al. (2005), which assesses men's body attitudes across three dimensions: body fat, muscularity, and height. Other measures such as the Drive for Muscularity Scale (McCreary et al., 2004) and the Muscle Appearance Satisfaction Scale (Mayville et al., 2002) provide a more comprehensive assessment of men's concerns with their muscles.

Determinants of Body Image

Psychological models have been fundamental in developing an understanding of body image and its determinants. While several models have been proposed over the last two decades, a common understanding is that body image development is a function of a combination of sociocultural influences and individual factors.

Sociocultural Factors

In Western cultures, a slender figure is considered the ideal for girls and women. The ideal for boys and men is characterized by a muscular yet lean figure, with muscular arms and broad shoulders. Such appearance ideals become influential when they are transmitted and reinforced by sociocultural agents such as the media, family, and peers. The Tripartite Influence Model proposed by Thompson et al. (1999) conceptualizes this relationship by proposing that media, peer, and familial influences predict the development of body dissatisfaction via two mechanisms: internalization of appearance ideals and social comparisons. Internalization refers to the extent to which an individual considers cultural beauty ideals to be important and relevant to him- or herself (Grogan, 2008). Due to the inevitable discrepancy between beauty ideals and one's own appearance, internalization is often associated with body dissatisfaction. Social comparisons are also an important factor in the development of body image concerns. Appearance-related comparisons refer to the process of comparing one's own body and that of another in order to gather information about one's own physical attractiveness (Myers and Crowther, 2009). Studies show that individuals who have a bias toward making appearance-related comparisons are at an increased risk for body dissatisfaction.

The media is frequently recognized as a powerful influence in the development of body image. Research demonstrates

a consistent and vast overrepresentation of idealized thin and muscular bodies in the media. In addition to being underrepresented, overweight characters are commonly portrayed in a negative and stigmatizing manner. Thin characters are usually portrayed as intelligent, happy, attractive, and clean, whereas overweight characters are typically depicted as lazy, self-indulgent, in poor health, lacking self-control, and unattractive. It is therefore unsurprising that body dissatisfaction is associated with exposure to idealized images. For instance, recent *meta*-analyses have demonstrated that media exposure to idealized body shapes is associated with eating disorder symptomatology (Hausenblas et al., 2013). This trend has been found predominantly among women and adolescent girls, but recently, also among men, adolescent boys, and young children. In a review of research on the effects of television and magazines on body image and disordered eating attitudes among women and girls, López-Guimerà et al. (2010) concluded that there is increasing evidence that exposure to media constitutes a direct risk for body dissatisfaction, and also indirectly via peers and parents. The way in which peers and the family reinforce media messages has not been widely researched. However, there is increasing evidence that indicates that the effect of negative messages from a combination of the media, peer, and family greatly increases the risk of body dissatisfaction.

Appearance ideals are conveyed and reinforced by more immediate social agents such as peers. Peers exert their influences on body image via appearance conversations, direct comments, modeling, and social comparisons. For example, appearance conversations with peers involving looks, image, and attractiveness create a social environment where appearance is focused on and valued, thereby reinforcing the importance of the thin/muscular appearance ideals. Lawler and Nixon (2011) found that among Irish adolescents, appearance-related conversations with friends were a predictor of body dissatisfaction and that internalization mediated this relationship. Peers are also an important source of social comparison as perceived similarity is a strong motivating force when choosing a comparison target. A tendency to engage in social comparisons with peers is also an important contributor to body dissatisfaction. This was demonstrated by Halliwell and Harvey (2006) who found that social comparisons were related to body dissatisfaction among both female and male British adolescents aged 11–16 years.

Parents also play an important role in shaping the attitudes and behaviors of their children in regard to weight and appearance. A primary way parents exert this influence is via the modeling of values, feelings, dieting, exercising, and other weight-related behaviors. Research shows that mothers' body dissatisfaction and dieting behaviors are related to their daughters' and sons' body dissatisfaction (van den Berg et al., 2010a). Additionally, parents can directly influence their children's body image by pressuring their children in regard to their weight or appearance, in the form of appearance-related criticism. Parental comments about children's weight, teasing, and encouragement to lose weight have been associated with higher levels of body dissatisfaction among adolescents and children of both genders. For example, Helfert and Warschburger (2011) found that for German adolescents, parental encouragement to control weight and shape was a strong predictor of weight concerns in boys and girls.

Individual Factors

Weight status is the most consistent and robust predictor of body image. The classification of weight status is based upon body mass index (BMI), an index of weight and height; whereby an individual's body mass is divided by the square of his or her height. Studies show that overweight and obese individuals have poorer body image than average-weight individuals, with obese individuals having an even poorer body image than overweight individuals. Furthermore, this relationship has been found to be consistent across gender and age. For example, a 10-year longitudinal study with participants from diverse racial and socioeconomic backgrounds in the United States examined changes in body dissatisfaction from adolescence to young adulthood (Bucchianeri et al., 2013). The researchers found that for males and females, an increase in BMI over time was associated with an increase in body dissatisfaction. Similarly, an Australian study conducted by Olive et al. (2012) with preadolescents aged 7–9 years found that increases in percentage of body fat among girls and boys were associated with greater body dissatisfaction. The relationship between body image and BMI has also been found in other countries, including Spain, Portugal, China, Malaysia, Palestine, and Tonga. For example, McCabe et al. (2012) found that among Tongan adolescents, dissatisfaction with weight and shape was significantly higher among overweight and obese adolescents than among normal-weight adolescents.

The relationship between body image and self-esteem is also well established. Self-esteem is critical for the development of a positive self-image and a sense of self-worth, which are vital in developing a positive body image and promoting general well-being. Childhood and adolescence are particularly significant periods for the development of body image and self-esteem. Several studies have demonstrated the link between the two. For example, in a study conducted with Australian adolescents in grades 7–12, van den Berg et al. (2010b) found that body dissatisfaction and self-esteem were strongly related across different age-groups and for both males and females. These relationships between body dissatisfaction and self-esteem are not only important for young people, but are significant issues throughout the life span. For example, Mellor et al. (2010) examined the relationship between body image and self-esteem across time and age in Australian men and women aged between 20 and 86 years. Self-esteem was related to body dissatisfaction for both men and women, and did not change over a 2-year period or across adulthood. The researchers argued that the finding is unsurprising giving that body image represents a large portion of an individual's self-concept, regardless of gender and age.

Having a perfectionistic personality type is often associated with eating disorders and body image concerns. Perfectionism is characterized by rigid personal expectations, driven by exceedingly high personal standards and critical self-evaluations. Theorists have defined two dimensions underlying the concept of perfectionism, which have been described

as maladaptive and adaptive perfectionisms. Adaptive perfectionism is related to high achievement standards and is generally not associated with psychological distress, whereas maladaptive perfectionism is related to negative reinforcement and self-defeating behaviors (Slade and Owens, 1998). Maladaptive perfectionism has been shown to be related to a number of negative health outcomes, including anxiety and depression, and is an important factor in the development and maintenance of body concerns and eating pathology. Although adaptive perfectionism is thought to be less problematic, body dissatisfaction may be one context where adaptive perfectionism may also be detrimental, as some studies have found that both adaptive and maladaptive perfectionisms are associated with body dissatisfaction. For example, Wade and Tiggemann (2013) found that both adaptive and maladaptive perfectionisms were associated with body dissatisfaction among women aged 28–40 years.

Body Image across Gender and the Life Span

Childhood is a significant time for social and cognitive development. A number of different aspects of children's socialization have implications for their body image development. Social comparisons are a fundamental aspect of children's cognitive, social, and emotional developments. Children learn about their own abilities, differences, strengths, and weaknesses via social comparisons. Around the age of 7–8 years, children's social comparisons become increasingly evaluative and more frequently used. This is also a significant age for the development of body dissatisfaction. Holt and Ricciardelli (2002) found that 30% of boys and girls engaged in social comparisons in regard to aspects of their body, including their weight and muscles. These comparisons were related to problem in eating attitudes, lower levels of self-esteem, and higher levels of negative affect.

For both adolescent and preadolescent boys, performance in sport is closely related to their self-esteem and body image. Sport plays an important socializing role for boys, but it is also a way for boys to develop and demonstrate their muscularity. Sport also provides a context for boys to more openly discuss what they like and do not like about their bodies, and many aspects of the body that boys want to improve are those that would make them more successful at sport. These include a focus on strength, endurance, and toughness. Thus, whereas girls focus on their appearance, body image for boys is primarily focused on the function of their bodies. Even in qualitative interviews with boys aged 5 and 6 years, Birbeck and Drummond (2006) found that boys associated a larger body with greater physical performance. Boys maintained that if one was 'bigger,' one could run faster, play football better, and have other physical advantages.

One of the most critical milestones for individuals is puberty. For girls and boys, puberty is accompanied by major changes in their body, their self-concept, mood, and social interactions. There are important differences for girls and boys during this time. Boys often have a positive pubertal experience as developmental changes, which include increases in stature and muscle, move them closer toward achieving the ideal male body shape. On the other hand, girls often find that physical changes that accompany puberty, for example, weight gain, move them further away from beauty ideals for women. For girls and boys, early or late onset of puberty can lead to feelings of alienation and depression. In particular, for girls, research has shown that pubertal development predicts body dissatisfaction, whereas for boys the relationship appears to be less consistent. In one study, McCabe and Ricciardelli (2004) found that early pubertal onset was a risk factor for body dissatisfaction in girls, whereas early maturing boys demonstrated higher levels of body satisfaction. Despite these higher levels of body satisfaction, early maturing boys were more likely to adopt strategies to decrease weight and increase muscle, perhaps because of their focus on achievement in sport.

Pregnancy is another significant change which impacts on adult women's body image. Women often report that they are satisfied with their bodies during pregnancy, and in some cases research has demonstrated that pregnant women are actually happier with their bodies than they were prior to being pregnant. It is thought that this is due to women's greater acceptance of bodily changes as a reflection of the healthy development of the growing baby. For example, Clark et al. (2009) found that during pregnancy, women's perceived attractiveness remained stable, whereas women reported greater body dissatisfaction after the baby was born.

Studies have shown that women in their 40s and 50s are as likely to desire a thinner figure or engage in weight loss strategies as are women in their 20s. Appearance ideals are often at odds with natural changes that occur in a woman's body with age. For example, women experience pressure associated with maintaining a youthful appearance. In efforts to keep faces and bodies smooth, taut, slim, and wrinkle-free, women are increasingly turning toward invasive cosmetic procedures such as face lifts and collagen treatments. One developmental factor significant for this age-group is the onset of menopause. Studies have shown that the changes which often accompany menopause are associated with increased body dissatisfaction. This is often related to weight gain, which is a common characteristic of menopause. The progression into older adulthood is accompanied by major physical changes including wrinkling of the skin, loss of muscle, graying of the hair, and loss of height. In a study of body image among American women aged 25–89 years, Runfola et al. (2013) found that body dissatisfaction significantly decreased across older aged women; however, the association was no longer significant after stratifying for BMI. In addition, other studies have shown that over 60% of women aged between 60 and 85 years report being dissatisfied with their body weight and/or appearance.

There is little research currently available that has examined the development of body image during adulthood for men. Some studies show that older men display more body dissatisfaction than younger adult men, while other studies show no age-related trends. Studies which have examined body image among older men are rare; however, one study conducted by Peat et al. (2011) found significant differences in body dissatisfaction between younger and older men, and found that younger men demonstrated greater dissatisfaction despite having a lower BMI. A similar finding was obtained by McCabe and McGreevy (2011). The researchers argue that this finding may reflect differential media pressures on younger

and older males, with media influences becoming less relevant for older men.

Body Image in Different Cultural Groups

The study of body image across cultures is also important given there are cultural differences that can give rise to diverse beauty ideals and body modification practices. These have been studied both within different cultural groups in Western countries, and between Western and non-Western countries.

Black American women are one cultural group who have been found to report low levels of body dissatisfaction. Black women in the United States tend to have larger body sizes; yet relative to White women, they are often found to be more satisfied with their bodies. In a *meta*-analysis of 98 studies conducted between 1960 and 2004, Black women were found to consistently report less body dissatisfaction than White or Hispanic women (Kronenfeld et al., 2010). Body ideal differences are often attributed to differences in BMI. For instance, Blacks in the United States are much more likely to be obese or overweight than Whites. However, family and peers are viewed as providing protective barriers against cultural messages, which reinforce the thin ideal. Granberg et al. (2009) found that greater positive appearance-related feedback from family and peers was associated with being more comfortable with a larger body shape.

In Hispanic communities, there may also be more acceptance of a wider range of body types than found among Whites. For Hispanic women, the ideal figure is represented as curvaceously slender with fat distributed in specific areas such as the breasts, buttock, and hips. However, research comparing the rate of body dissatisfaction between Hispanics and Whites has provided mixed findings. Some studies indicate that Hispanic girls and women experience greater body satisfaction and less concern with weight compared to White girls and women. Other studies have found that Hispanic girls and women are as dissatisfied as White girls and women. Hispanic males' body image does not seem to differ from White males' body image, as the majority of studies have found no differences between Hispanics and Whites (Ricciardelli et al., 2007).

In stark contrast to Western appearance ideals, which are defined by thinness, non-Western cultures have traditionally valued and striven toward achieving a larger body shape. Research suggests that ideal body shapes are slimmer in countries with higher socioeconomic status, such as Westernized societies. In less socioeconomically developed countries, a larger body shape is more desirable as it signifies personal wealth, social status, fertility, and sexuality among women. This may provide some understanding to findings of cross-cultural differences such as those found by McCabe et al. (2012) who examined body dissatisfaction across eight different countries. The researchers found that Fijian and Tongan men and women were among the highest for body size status, yet also expressed the highest level of body satisfaction. The nature of the messages that are received regarding the value of a large body size is important in shaping the body image idea and hence the levels of body dissatisfaction. In these countries, a large body is a sign of status, care, and respect. Researchers have highlighted the effects of globalization and the spread of Westernized media imagery as a major reason why even in developing nations, body image concerns and the desire for thinness are becoming increasingly prevalent.

Preventing Body Image Concerns

School-based intervention programs are increasingly being adopted as a means of preventing and reducing body image concerns and eating problems among children and adolescents. There have been different approaches to addressing body image within the classroom. Self-esteem approaches have found some success by drawing upon Social Cognitive Theory in order to tap into students' self-beliefs and attitudes associated with their body image. The 'Everybody is Different' program adopted this approach with Australian adolescents in grades 7 and 8. With a specific focus on the development of self-esteem and positive self-image, this 9-week program demonstrated improvements in body image for both girls and boys, immediately post program and also at 12-months follow-up. Media literacy approaches have also become a common way to address the negative influence of the media via the education of students to understand the ways in which media images create social norms. This involves critically analyzing media images, and recognizing the portrayal of stereotypes, to discover how media images are 'perfected' with the use of airbrushing and retouching. Recently, increasing awareness of body image concerns among preadolescent and adolescent boys has led to the development of school-based body image programs specifically for boys. Stanford and McCabe (2005) developed a program aimed to enhance adolescent boys' self-identity and self-esteem, promote acceptance of different body shapes, and develop media literacy. The program resulted in increased self-esteem and satisfaction with muscles, and reduced negative affect, particularly among those boys who already experienced high levels of body dissatisfaction.

In addressing the various influences that impact on children's and adolescents' body image, prevention efforts need to target the broader school environment, well beyond the confines of the classroom. A whole-school approach involves the collective effort of the school and community to promote health and positive self-image. Policies such as antibullying and greater tolerance of individual differences have potential for the most wide-reaching and powerful effects.

At the public policy level, initiatives have also been developed by a number of governments in efforts to address body image concerns and problem eating. In some countries, government legislation has been developed to reduce exposure, internalization, or pressure associated with thin ideals. For example, the Spanish Government has banned the broadcasting of television commercials that promote beauty products, cosmetic surgery, and chemical options for achieving an ideal appearance. Several other countries, which include Australia, Italy, France, and Canada, have also developed codes of conduct for media and related industries. These generally aim to influence the presentation of media images in a way that is less likely to contribute to the development of body image concerns and problem eating. For example, in Australia, the Federal Government has promoted voluntary media and industry codes of conduct in which media and fashion

organizations are encouraged to not digitally manipulate images and to represent a diverse range of body shapes.

Conclusion

The concept of body image has had a long history of development over the last century. It is shaped by biological, psychological, and sociocultural factors. These include the spread of Western media, as portrayed in magazines, television, and films. Body image is strongly associated with psychological adjustment, as well as health risk behaviors related to eating and building muscles. Current research demonstrates that body image and body image concerns are no longer topics that are exclusively relevant to women and adolescent girls, but rather they are of relevance to males and females, regardless of age or cultural background. Prevention and intervention programs to address these body image concerns need to be designed so that they address the specific concerns of the target group.

See also: Adolescent Health and Health Behaviors; Culture as a Determinant of Mental Health; Eating Disorders; Health Behavior, Psychosocial Theories of; Health Education and Health Promotion; Health Interventions, Community-based; Health Promotion in Schools; Health Psychology in Childhood; Health Psychology; Human Development and Health; Men's Health; Mental Health; Physical Activity and Health Promotion; Physical Activity and Health; Race: Ethnicity and Health; Weight Regulation.

Bibliography

Birbeck, D., Drummond, M., 2006. Understanding boys' bodies and masculinity in early childhood. International Journal of Men's Health 5, 238–250.

Bucchianeri, M.M., Arikian, A.J., Hannan, P.J., et al., 2013. Body dissatisfaction from adolescence to young adulthood: findings from a 10-year longitudinal study. Body Image 10, 1–7.

van den Berg, P.A., Keery, H., Eisenberg, M., Neumark-Sztainer, D., 2010a. Maternal and adolescent report of mothers' weight-related concerns and behaviors: longitudinal associations with adolescent body dissatisfaction and weight control practices. Journal of Pediatric Psychology 35, 1093–1102.

van den Berg, P.A., Mond, J., Eisenberg, M., et al., 2010b. The link between body dissatisfaction and self-esteem in adolescents: similarities across gender, age, weight status, race/ethnicity, and socioeconomic status. Journal of Adolescent Health 47, 290–296.

Cash, T.F., 2004. Body image: past, present, and future. Body Image 1, 1–5.

Clark, A., Skouteris, H., Wertheim, E.H., et al., 2009. The relationship between depression and body dissatisfaction across pregnancy and the postpartum: a prospective study. Journal of Health Psychology 14, 27–35.

Granberg, E.M., Simons, L.G., Simons, R.L., 2009. Body size and social self-image among adolescent African American girls: the moderating influence of family racial socialization. Youth & Society 41, 256–277.

Grogan, S., 2008. Body Image : Understanding Body Dissatisfaction in Men, Women and Children, second ed. Routledge, East Sussex, UK.

Halliwell, E., Harvey, M., 2006. Examination of a sociocultural model of disordered eating among male and female adolescents. British Journal of Health Psychology 11, 235–248.

Hausenblas, H.A., Campbell, A., Menzel, J.E., Doughty, J., Levine, M., Thompson, J.K., 2013. Media effects of experimental presentation of the ideal physique on eating disorder symptoms: a meta-analysis of laboratory studies. Clinical Psychology Review 33, 168–181.

Helfert, S., Warschburger, P., 2011. A prospective study on the impact of peer and parental pressure on body dissatisfaction in adolescent girls and boys. Body Image 8, 101–109.

Holt, K., Ricciardelli, L.A., 2002. Social comparisons and negative affect as indicators of problem eating and muscle preoccupation among children. Journal of Applied Developmental Psychology 23, 285–304.

Jakatdar, T.A., Cash, T.F., Engle, E.K., 2006. Body-image thought processes: the development and initial validation of the assessment of body-image cognitive distortions. Body Image 3, 325–333.

Kronenfeld, L.W., Reba-Harrelson, L., Von Holle, A., et al., 2010. Ethnic and racial differences in body size perception and satisfaction. Body Image 7, 131–136.

Lawler, M., Nixon, E., 2011. Body dissatisfaction among adolescent boys and girls: the effects of body mass, peer appearance culture and internalization of appearance ideals. Journal of Youth and Adolescence 40, 59–71.

López-Guimerà, G., Levine, M.P., Sánchez-Carracedo, D., et al., 2010. Influence of mass media on body image and eating disordered attitudes and behaviors in females: a review of effects and processes. Media Psychology 13, 387–416.

Mayville, S.B., Williamson, D.A., White, M.A., et al., 2002. Development of the muscle appearance satisfaction scale: a self-report measure for the assessment of muscle dysmorphia symptoms. Assessment 9, 351–360.

McCabe, M.P., Fuller-Tyszkiewicz, M., Mellor, D., et al., 2012. Body satisfaction among adolescents in eight different countries. Journal of Health Psychology 17, 693–701.

McCabe, M.P., McGreevy, S.J., 2011. Role of media and peers on body change strategies among adult men: is body size important? European Eating Disorders Review 19 (5), 438–446.

McCabe, M.P., Ricciardelli, L.A., 2004. A longitudinal study of pubertal timing and extreme body change behaviors among adolescent boys and girls. Adolescence 39, 145–166.

McCreary, D.R., Sasse, D.K., Saucier, D.M., et al., 2004. Measuring the drive for muscularity: factorial validity of the drive for muscularity scale in men and women. Psychology of Men & Masculinity 5, 49.

McKinley, N.M., Hyde, J.S., 1996. The objectified body consciousness scale. Psychology of Women Quarterly 20, 181.

Mellor, D., Fuller-Tyszkiewicz, M., McCabe, M., et al., 2010. Body image and self-esteem across age and gender: a short-term longitudinal study. Sex Roles 63, 672–681.

Mendelson, B.K., Mendelson, M.J., White, D.R., 2001. Body-esteem scale for adolescents and adults. Journal of Personality Assessment 76, 90–106.

Myers, T.A., Crowther, J.H., 2009. Social comparison as a predictor of body dissatisfaction: a meta-analytic review. Journal of Abnormal Psychology 118, 683–698.

Olive, L.S., Byrne, D.G., Cunningham, R.B., et al., 2012. Effects of physical activity, fitness and fatness on children's body image: the Australian LOOK longitudinal study. Mental Health and Physical Activity 5, 116–124.

Peat, C.M., Peyerl, N.L., Ferraro, F.R., et al., 2011. Age and body image in Caucasian men. Psychology of Men & Masculinity 12, 195–200.

Reas, D.L., Whisenhunt, B.L., Netemeyer, R., et al., 2002. Development of the body checking questionnaire: a self-report measure of body checking behaviors. International Journal of Eating Disorders 31, 324–333.

Ricciardelli, L.A., McCabe, M.P., Lillis, J., et al., 2006. A longitudinal investigation of the development of weight and muscle concerns among preadolescent boys. Journal of Youth and Adolescence 35, 177–187.

Ricciardelli, L.A., McCabe, M.P., Williams, R.J., Thompson, J.K., 2007. The role of ethnicity and culture in body image and disordered eating among males. Clinical Psychology Review 27 (5), 582–606.

Runfola, C.D., Von Holle, A., Trace, S.E., et al., 2013. Body dissatisfaction in women across the lifespan: results of the UNC-SELF and gender and body image (GABI) studies. European Eating Disorders Review 21, 52–59.

Slade, P.D., Owens, G.R., 1998. A dual process model of perfectionism based on reinforcement theory. Behavior Modification 22, 372–390.

Stanford, J.N., McCabe, M.P., 2005. Evaluation of a body image prevention programme for adolescent boys. European Eating Disorders Review 13, 360–370.

Thompson, J.K., Coovert, M.D., Stormer, S.M., 1999. Body image, social comparison, and eating disturbance: a covariance structure modeling investigation. International Journal of Eating Disorders 26, 43–51.

Thompson, J.K., van den Berg, P., 2002. Measuring body image attitudes among adolescents and adults. In: Cash, T.F., Pruzinsky, T. (Eds.), Body Image: A Handbook of Theory, Research, and Clinical Practice. Guilford Press, New York, pp. 142–154.

Tylka, T.L., Bergeron, D., Schwartz, J.P., 2005. Development and psychometric evaluation of the male body attitudes scale (MBAS). Body Image 2, 161–175.

Wade, T.D., Tiggemann, M., 2013. The role of perfectionism in body dissatisfaction. Journal of Eating Disorders 1, 2–8.

Bologna Process

Annamaria Campanini, Università di Milano Bicocca, Milano, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes the Bologna Process in higher education in Europe that started with the Bologna Declaration signed by 29 Ministers of Education of different European countries in 1999, and aimed to create the European Higher Education Area. Based on cooperation between ministries, higher education institutions, students and academic staff, this initiative grew to encompass 47 countries and various international organizations by 2012. This article describes the European institutions involved and analyzes the effects of the Bologna Process on social work education and practice to highlight strengths and weaknesses in this approach.

Introduction

Education plays an important role in integration processes aimed at developing a common identity in a subcontinent composed of different nation-states with their own histories, cultures, policies, and traditions. The concept of identity is a slippery one, including the European identity in relation to social work (Frost, 2008). Formal European agencies were involved in creating the Bologna framework. Crucial among these is the Council of Europe, established under the 1949 London Treaty to promote democracy, protect human rights, and enforce the rule of law in Europe among the 47 countries that form this organization. During this period, different agencies created to preserve peace in Europe had the political and economic aims of strengthening integration between member countries. In 1958 they established the European Economic Community (EEC). Since then, a huge single market has been created and continues to develop its full potential as what began as a purely economic union evolves into an organization that spans all policy areas including development aid and the environment. The Maastricht Treaty signed on 7 February 1992 was enacted on 1 November 1993 when the EEC changed its name to the European Union (EU) that currently has 27 member states. The Bologna Process was a European intervention in higher education. I consider its role in devising a more unified identity for social work below.

There was an aspiration for closer cooperation between member states at all levels of education since the first Education Action Program in 1976. In the1980s, with the emergence of 'social Europe,' the European Commission (a body representing the interests of the EU as a whole) proposed new legislation to the European Parliament with the key objectives of developing and improving the European component within teaching; and promoting geographical mobility among students. Two measures, the Socrates and Erasmus Mobility Programs, have had significant impact upon the creation of a European identity and development of better understandings of many common problems that affect European societies (Campanini, 2007). These have encouraged the evolution of a social work shared knowledge base and identified a need to strengthen the intercultural dimension within student learning, and prepare younger generations for life in societies increasingly characterized by cultural and linguistic diversities where capacity to fight racism and xenophobia is important. Social work educators participated actively in many such initiatives including: the Tempus Programme that supported the development of social work education in Eastern European countries; and Thematic Network Scheme (ECSPRESS, led by Walter Lorenz from 1996 to1999, examined the social professions in Europe; EUSW, coordinated by Annamaria Campanini, from 2002 to 2008 followed the Bologna Process).

What Is the Bologna Process?

The Bologna Process is a route aimed at creating the European Higher Education Area (EHEA) through cooperation between ministries, higher education institutions, students, and staff from 47 countries, with the participation of international organizations. The name comes from the first Declaration signed in Bologna (Italy) on 19 June 1999 by ministers in charge of higher education from 29 European countries. The idea had emerged in 1998, during the celebration of the 800th anniversary of the University of Paris when education ministers from France, Germany, the United Kingdom, and Italy decided that the segmentation of the higher education sector in Europe was outdated and harmful. To counteract this, they signed the Sorbonne Joint Declaration to create a common frame of reference within the intended EHEA to promote student, teaching staff, and graduate mobility and develop qualifications suited to the labor market.

This voluntary process, initiated with the Bologna Declaration (1999), was to harmonize diverse European Higher Education Systems by 2010 through a European Space of Higher Education.

Bologna's specific aims were to increase the employment prospects and geographical mobility of European citizens; and enhance the reputation of European Higher Education globally. To realize this project it was necessary to plan long-term actions that would achieve:

- Easily, comprehensible and comparable systems of degrees.
- An educational structure founded upon two cycles (BA and MA), using the European Credits Transfer System.
- Geographical mobility for students, teachers, researchers, and administrative staff, and recognition of experiences within a European context.
- European cooperation in quality assurance and the development of comparable criteria and methodologies.

These established the first steps of the Bologna Process. During subsequent biannual Ministerial Conferences this educational agenda was broadened through further communiqués. Increasing intercountry participation (47 out of the 49 countries that ratified the European Cultural Convention of the Council of Europe in 1954 are now involved) resulted in further evolution of the main objectives and tools. Each meeting produced further steps down the harmonization path and highlighted different aspects of the new higher education policy. While acknowledging the strengths of this process, dissenting voices were expressed in the Nordic countries and Portugal by educators concerned about the dilution of standards (Dominelli, 2007).

The Prague Communiqué in 2001 had 33 member states expanding these objectives by stressing the importance of lifelong learning as an essential strategy within the EHEA. Additionally, students were configured as active partners in enhancing the attractiveness and competitiveness of European Higher Education. Other important dimensions were the introduction of the 'social' dimension expressed as equality of opportunities in higher education; facing the challenges of competitiveness and new technologies; and improving social cohesion, equal opportunities for all citizens, and quality of life. Participating ministers committed to ensuring the further development of quality assurance mechanisms and national qualification frameworks.

The Berlin Communiqué, enlarging the number of countries to 40 members, was adopted in 2003 and emphasized research, research training, and interdisciplinarity, in maintaining and enhancing the attractiveness of higher education in Europe. Another aspect it highlighted was that of developing an agreed set of standards, procedures, and guidelines on quality assurance to secure implementation of an adequate peer review system. Connected with the topic of comparability, the communiqué invited member states to develop a framework of comparable and compatible qualifications for their higher education systems, which would describe qualifications in terms of workload, level, learning outcomes, competences, and profile. The different levels of degree were to have different orientations and profiles, to accommodate a diversity of individual, academic, and labor market needs.

These Communiqués, predicated on the Lisbon Recognition Convention (1998), argued that first cycle degrees (Bachelors) should provide access to the second cycle programs (Masters), which would lead to doctoral studies. The Berlin Communiqué (2003) recommends the introduction of a universal 'diploma supplement' (transcript with description of the types of institution and level of the degree) to be annexed to each certificate by 2005. This was to facilitate recognition of titles across all European countries, in accordance with the Lisbon Convention (1998).

The Bergen Communiqué of 2005 emphasized the significance of higher education in enhancing research; underpinning higher education; bringing economic and cultural development within European societies; and promoting social cohesion. It also asserted that efforts to introduce structural change and improve the quality of teaching should not detract from efforts to strengthen research and innovation. And, it emphasized the centrality of partnerships among diverse stakeholders – students, HEIs, academic staff, and employers. Regarding the third cycle (doctoral studies), ministers highlighted the need for doctoral level qualifications to be fully aligned within the overarching EHEA framework for qualifications using the outcomes-based approach. Participants in doctoral programs were to be deemed both students and early stage researchers that advanced knowledge through original research. The document emphasized the need to increase the number of doctoral candidates and doctoral programs that promoted interdisciplinary training, developed transferable skills, and guaranteed transparent supervision and assessment.

In 2007, the London Communiqué engaged 46 countries and evaluated achievements in mobility, degree structure, recognition, qualification frameworks (both overarching and national), lifelong learning, quality assurance, and the 'social' dimension alongside the economic one. The inclusion of the 'social' was very important for social work as it is embedded within this domain. An increasing awareness that a significant outcome of the process would be a move toward student-centered higher education and away from teacher-driven provision became evident. This meeting also set priorities for 2009: mobility and the 'social' dimension, which were defined for the first time, data collection, employability, EHEA in a global context, and stock taking. For 2010 and beyond, it stressed further collaboration and an opportunity to reformulate the visions and values of this important process.

The last meeting, before the official end of the first phase of the Bologna Process was held in Leuven/Louvain-la-Neuve in 2009. It aimed to define the working areas for the next decade – to enhance the Bologna Process through an in-depth realization of the reforms. Emphasis was on the 'social' dimension, lifelong learning, employability, student-centered learning, the teaching mission of education, international openness, mobility, education, research and innovation, data collection of the implementation of the Bologna Process among participating countries, funding of the HE, and multidimensional transparency tools.

The Impact of the Bologna Process on Social Work Education

In March 2010, in Budapest–Vienna, a decade of the Bologna Process was celebrated with an Anniversary Conference, for the official launching of the EHEA. The Bologna Process has given social work an incredible opportunity to engage with educational systems across Europe. As Lorenz (2006) and Lyons and Lawrence (2006) among others claim, social work is strictly related to specific cultural and contextual dimensions of the different countries in which it is operating. Historical backgrounds, social policy trends, pedagogical and disciplinary relationships, and 'politics' all have implications for the way in which training in social work has been established in different countries and for the specific features in which social work roles have been interpreted. Local factors and European policy developments are significant elements in determining how social work education evolves its structures, curricula, and pedagogical issues and the roles that social workers will play in the social scene. In the book, *European Social Work. Commonalities and Differences* (Campanini and Frost, 2004) representatives of 24 different European nations presented the 'state of

the art' of social work in their countries, giving a clear image of commonalities and differences in social work education at different levels. Other research, e.g., by Martinez (2007) and Hamburger et al. (2004–2007), focused specifically on social education in the wake of the Bologna Process and structures across Europe and on fieldwork education/placements in 15 European countries (Campanini, 2009). These studies reveal that despite the Bologna Process, which aimed to harmonize higher education in Europe, meaningful differences between countries persist, including variety in educational providers and manner of delivery (Campanini, 2011).

Alongside countries that have recently entrusted social work education to the university sector are those that have developed learning routes in higher private or public institutes. Nations such as Poland, Lithuania, and Portugal have the simultaneous presence of various educational contexts, e.g., public and private universities and colleges, in others, employing agencies are at the fore as in the work-based route in the United Kingdom. The Bologna Process has also recognized institutions like Fachhochschule (Germany) or Hogeschoolen (Netherlands, Denmark) as universities of applied or professional sciences with the power to organize BAs. However, the second education level (Masters) remains the prerogative of universities that retain sole responsibility for providing the third level (doctoral studies).

Despite many changes that have already taken place, the need to restructure educational paths in accordance with the Bologna Process of 180 credits for the first cycle of 3 years (the BA) and 120 credits for the second cycle (MA), or alternatively 240 credits for 4 years for the BA, and 60 credits for 1 year for the MA remains, while differences have not always been successfully harmonized and first and second cycles can be of varying lengths. The Bologna Process stipulates 300 credits and 5 years of study for a Master's degree, but countries have responded differently, e.g., Sweden, has settled on 3.5 years plus 1.5 years pattern, while Spain defined its curriculum as 4 years plus 1, and Italy 3 years plus 2.

There are also differences in the disciplinary components of the curricula and internship (practice placement) organization. The foundational base of social work education includes disciplines like the human and social sciences of sociology, psychology, social pedagogy, social policy, and legal studies. Curricula vary in the subjects offered and prepare students to undertake distinct roles, thereby contributing to professional differentiation. Additionally, preparation for professional practice – variously referred to as 'placement,' 'fieldwork,' or 'practice education,' – differs in the number of hours dedicated to social work theory and supervised practice and the modalities within which the training is conducted. These can vary from 50% of the entire period of study as in the United Kingdom, to 210 h in Estonia.

Teaching methodologies also vary. The Bologna Process places the student as the central actor in the learning process. However, the role of teachers on social work courses can be quite different and is determined by the 'learning process' that dominates a particular national context. Active learning has a very strong background and presence in Scandinavian countries, with specific tools and evaluation processes more oriented to helping the student develop a critical and self-reflective attitude (Campanini et al., 2012; Frost et al., 2012). In contrast, in Italy, e.g., more conventional methods based on the traditional notion of an expert lecturer imparting knowledge to educable and eager students are used. The EU has funded web-based courses since the late 1990s (Dominelli and Thomas Bernard, 2003), but Information and Communications Technologies (ICTs) technologies are only slowly entering social work education. As an example, the Norwegian-led initiative of a virtual classroom, 'Virclass' involving teachers and students from different countries, has proved useful for organizing pan-European modules (Larsen and Hole, 2007; viewable on www.virclass.net).

As foreseen by the Berlin communiqué, research is a pillar of the knowledge society and important for enhancing social work education (Campanini, 2007). The inclusion of research within education varies considerably among countries. Nations such as Sweden and the United Kingdom have a long tradition of PhD studies linked to extensive publication of social work research. In Finland, the creation of a joint venture between the universities through SOSNET allowed intensive cooperation in research including conducting several research projects concerned with questions about social work education, expertise, and practical training. Many countries are now organizing PhD programs specifically in social work or as a pathway within other disciplines. For example, in 2013, France added a title of 'social work' in sociology and pedagogy PhD courses. There are still problems linked with the positioning of social work education inside and outside the university, both in relation to having students undertake research degrees, and the expectation that teachers will be research active (Lyons, 2006). Yet, more and more social work teachers are required to have a PhD and be evaluated in relation to research, as occurs regularly in the United Kingdom.

Involvement and participation in European projects and activities is high in nearly all countries and social work has been active in opening itself to comparisons in social work education in various ways on an inter-European level (Campanini and Frost, 2004; Lyons and Lawrence, 2006; Friesenhan et al., 2012) and with North America since the late 1990s; and more recently to Asian-Pacific countries. Educators in many countries use the opportunities made available by the European Community for student mobility and teaching staff exchanges (Dominelli and Bernard, 2003).

These can involve intensive periods of study on specific topics through concentrated programs and particular initiatives that are open to contributions from teachers of various nations including during an international week. Opening education to a wider vision is also demonstrated by the insertion of modules that address various comparative aspects into social work while offering experiences through summer schools, e.g., Fachhochschule Alice Salomon in Berlin, Rovaniemi University of Applied Science in Finland, and those held in Parma (Italy) and Dornbirn (Austria) by the European Social Work Thematic Network (EUSW-TN). Some offer international Master's courses: e.g., Master of Arts in Comparative European Studies (MACESS) in Maastricht; Goteborg University; and Instituto Superior de Ciencias do Trabalho de Emrpesa (ISCTE) Lisbon. Others attempt to create specific courses that award the title of 'Bachelor of Intercultural and International Social Work' as in Copenhagen. In spite of these initiatives, difficulties remain. These include students and staff having insufficient knowledge

of languages, especially in France, Great Britain, Italy, and Spain; problems related to different scheduling and timetabling of courses; and the economic costs of studying abroad. These can all limit student mobility (Dominelli and Bernard, 2003). Reluctance to operate transnationally might be indicative of an inward facing attitude, whereby the European dimensions of engagement are not perceived as opportunities. Students then fail to develop intercultural and international professional competences that are now seen as core competencies in many parts of Europe.

The Bologna Process gave quality assurance considerable importance. 'The European Network for Quality Assurance in Higher Education (ENQA)' was established in 2000 to promote European cooperation in the field of quality assurance. In November 2004, the General Assembly of this network renamed the ENQA. The idea of the Association originated from the European Pilot Project for Evaluating Quality in Higher Education (1994–95) that demonstrated the value of sharing and developing experience in the area of quality assurance.

The Bologna Process has developed a substantial number of exchanges, contacts, and reflections. However, more remain to be done to harmonize social work education and balance the sharing of commonalities while respecting the particularities and differences of specific local contexts.

Bologna Process: A New Phase

A new phase of consolidation and operationalization among European countries has led to the formation of closer links with higher education systems around the world, including an increasing number of non-EU countries becoming involved in dialogue with countries participating in the Bologna Process to enhance worldwide cooperation in higher education. In 2009, a Bologna Policy Forum was organized in Leuven/Louvain-la-Neuve between Ministers of the 46 Bologna countries and colleagues from different parts of the world, and a wide range of NGOs to facilitate global dialogue for the first time. This event focused on the key role that HE plays in the development of society, based on lifelong learning and equitable access to learning opportunities at all levels of society; and importance of public investment in higher education. In spite of the economic crisis, transnational exchanges in higher education governed by academic values, a balanced exchange of teachers, researchers, and students between countries continue to promote fair and fruitful 'brain circulation,' as an alternative to the 'brain drain.' This ministerial initiative has become a regular event at the EHEA Ministerial Conferences.

In the second Bologna Policy Forum in Vienna, in March 2010, a specific Statement was adopted referring to the manner in which higher education systems and institutions responded to growing demand for their services, multiple expectations, and the balance between cooperation and competition in international higher education. This Forum's Statement included concrete feedback to be adopted by participants, e.g., nominating contact persons for each participating country to act as liaison points for the better flow of information and joint activities, and support global student dialogue.

The overarching theme of the third Bologna Policy Forum held in Bucharest in April 2012 was "Beyond the Bologna Process: Creating and connecting national, regional and global higher education spaces." It emphasized four topics: public responsibility for higher education within national and regional contexts; global academic mobility; incentives, barriers, balances, and imbalances in global and regional approaches to quality enhancement in higher education; and the contribution of HE reforms in enhancing graduate employability.

The Bologna Process is ongoing. The Leuven Communiqué adopted during the first Bologna Policy Forum established 10 priorities for the decade to 2019 to ensure more comparable, compatible, and coherent systems of higher education in Europe. These were the social dimension; lifelong learning; employability; student-centered learning; education, research, and innovation; international openness; mobility; data collection; multidimensional transparency tools; and funding. Some working groups and networks like the EHEA Information and Promotion Network; Network for Experts in Student Support in Europe; Network for National Qualifications Frameworks Correspondents were formed to monitor progress and report on the implementation of the Bologna Process. After the launching of the EHEA, the Bologna Process moved toward a new phase, a more in-depth one aimed at reducing discrepancies in the countries forming the EHEA. To support these efforts with joint action at European level, the ministers entrusted the Bologna Follow-up Group to prepare a work plan for the period 2012–15.

During the third Bologna Policy Forum in Bucharest in April 2012, the Ministerial Conference underlined the importance of higher education reform to generate sustainable growth and jobs in the face of the economic crisis. Its three key aims were to provide quality higher education to more students; better equip students with employable skills; and increase student mobility. A strategy was formulated to increase mobility with the specific target that at least 20% of those graduating in Europe in 2020 would have been on a study or training period abroad. This last requirement is very important in social work education, given its local base in practice. Having students study in a foreign country can encourage and test their ability in languages other than their mother tongue, capacity to understand a new reality, become more culturally aware, and recognize commonalities and differences in each situation. It also speeds up the abandonment of rigid premises, preconstituted certainties, and stereotypes; helps to understand the links between the theoretical frames of reference and concrete actions; engage with opportunities and constraints in social policies and the choices of social services organization (Dominelli and Thomas Bernard, 2003; Payne and Askeland, 2008; Campanini, 2009).

The Effects of the Bologna Process: From Education to Practice

Despite many initiatives aimed at exploring social work education in different countries undertaken through the activities of the EUSW Thematic Network, teacher exchanges, participation in other measures through the Erasmus–Socrates framework, and seminars and conferences organized by the European Association of Schools of Social Work, it has been impossible to detail and agree a comprehensive description of

core competences for the social work profession thus far. There are many reasons for this. The principal one is related to diverse conceptions of the social work profession in different European countries. To tackle this problem, more time and money would have to be invested in meetings, including the formation of a new Thematic Network on this topic, if the European Commission would have funded it. Unfortunately, neither a dissemination project nor a new proposal for a Thematic Network was accepted.

Nonetheless, the Bologna Process remains an important framework for drawing up general guidelines and providing specific tools to support this process. For example, an important aspect within the harmonization process is the definition of the 'Dublin descriptors'; the product of an informal group of specialists originating from different countries, who participated in the 'The Joint Quality Initiative network.' These descriptors clarified what is expected from students – knowledge and understanding, application of this knowledge and understanding, making judgments, communication skills, and learning skills. These specific qualities are adapted and matched to the demands of the perceived successful completion of the first, second, and third Bologna cycles, and highlight progression between these cycles (www.jointquality.org).

Another interesting initiative in the harmonization process is the project 'Tuning Educational Structures in Europe,' submitted to the European Commission at the end of 2000. It is still working toward an overall objective of implementing discussion and reflection upon the changes that Bologna has brought to educational structures and study content. The process of reforming requires a 'tuning' of curricula, in terms of structures, programs, and teaching. Although it is very important to protect the rich diversity within European education, it is also necessary to identify competences, learning outcomes, and design, and construct and assess the qualifications provided through this education (www.unideusto.org/tuning/).

The final reports by González and Wagenaar (2003), and González and Wagenaar (2005), from the 'Tuning' project, contain guidelines and practical examples. A specific group ran with the ambition to link the Dublin descriptors of the European Qualification Framework (EQF) for the EHEA with the eight levels of the EQF for Lifelong Learning in the Social Sciences. Operating during the period 2008–10, representatives of social work education participated in this group alongside colleagues from psychology, education, and occupational therapy.

The final document published on the 'Tuning' Web site has remained more a 'cultural' exercise than a starting point that identifies and discusses concrete applications for social work as has occurred for nurses or occupational therapists. As explored by Lyons and Lawrence (2006), the past two decades have witnessed significant changes, not only in the construction of Europe, but also within education in the social professions including social work, but with different consequences. Although the Bologna Process provides a strategic opportunity to augment professional growth and progress within social work, Juliusdottir (2006) argues that a reflexive process of becoming more academic creates conditions for a dialogue to emerge between social work and academic cultures that could lead to interdisciplinary understanding and mutual respect. Being a part of this culture facilitates fresh motivations for training and active participation in social workers' own knowledge production through research and other theoretical activities, raising social workers' confidence and status in society. And, it may lead to increased compatibility and mobility associated with increasing globalization and shared labor markets. Christine Labonté (2005) emphasizes the possibility, of retaining a generic program at Bachelor's level, then providing a greater degree of specialization, required particularly for the new professional fields, by offering specially tailored Masters programs as a by-product of a tiered academic structure. Dominelli (2007) argues that this also facilitates the linking of a ladder of qualifications to career structures whereby lecturers and managers hold doctorates, although transitional arrangements may be needed in some countries. The 'advanced' study program can be conceptualized as part of specific 'skilling' or reskilling and lifelong learning. However, a simultaneous disadvantage of the two cycles structure is "the danger of reducing the professional element of training, by subjecting it to purely academic quality criteria, and what is more, academic criteria that are largely not defined from within an 'indigenous' discipline" (Lorenz, 2006: p. 54).

Other problems arise from differing levels of importance accorded to the BA and MA in different countries; connected to the historical and cultural context of education in each country. One trend promotes the awarding of the social work qualification after an overall 5-year academic package, whereas another views the MA as an area for specialization, often connected with 'advanced' (e.g., managerial) tasks. The Bologna Process introduced vibrant discussions on the changes needed to revise the educational systems to meet its demands and implement it, as occurred in Portugal, where the social work qualification was achieved in 5 years or in Spain that was going in the same direction. Paradoxically, in Italy, where the document was first signed, the aims were only met at a structural level. Its philosophical premises and the two different levels recognized in the Professional Register were not given the same recognition in the labor market (Campanini, 2009).

From a connected but different point, Walliman (2004) emphasizes the dangers of an educational structure too exclusively responsive to, and tied up with market needs. He links this without and out competition between education providers, which might lead to a process of privatization within education, analogous to outcomes driven by the General Agreement on Trade in Services. This could create university programs for elites with higher costs and fewer opportunities of equal access for everyone.

The Bologna Process has motivated academics to reflect upon and discuss topics connected with both the problems and opportunities created by the new structures within social work education, and contemporary challenges regarding teaching methods and new identities for students in social work. In this context, the effort to intensify cross-national analytical comparisons is oriented toward clarifying current situations; and stepping back from daily activities to "regain a degree of initiative and autonomy and develop quality criteria that are accountable to the users of social services" (Lorenz, 2006: p. 57).

I conclude that the Bologna Process has given an important input to developing a strong interest in participating in mutually identifying processes that recognize common

concerns alongside ethical principles and promoting a European social work identity. According to Frost (2008: p. 350):

> Examples of social work constituencies in European countries forging links and associations, projects and cooperative activities to engender European social work identities are burgeoning, with the EU itself financing some of this work. In social work education and academia particularly, there is perceptible evidence of conscious attempts to develop Europeanization.

Another important consideration is whether social work in Europe educates for 'Bologna competences' as well as national expectations. 'European' social workers must be educated to meet the general Bologna competencies, but importantly in every country, education is expected to equip students to fulfill fundamental social work values around situating understanding and interventions in their specific cultural and ethnic contexts (Campanini and Frost, 2004).

See also: Social Work Education.

Bibliography

Campanini, A., Frost, E., 2004. European Social Work: Commonalities and Differences. Carocci Editore, Roma.

Campanini, A., 2007. Introduction. Educating social workers in the context of Europe. In: Frost, E., Freitas, M.J. (Eds.) with Campanini, A. (guest ed.), Social Work Education in Europe. Carocci Editore, Roma, pp. 1–22.

Campanini, A., 2009a. Opportunities in social work education in Europe. In: Kantowic, E. (Ed.), Role of Research in Education for Social Work in Europe. UWM, Olsztyn, pp. 114–117.

Campanini, A., 2009b. Social work in Italy. Problems and perspectives. Locus Soci@l 3, 36–47.

Campanini, A., 2011. Social work in Europe. In: Healy, L.M., Link, R.J. (Eds.), Handbook of International Social Work: Human Rights, Development, and the Global Profession. Oxford University Press, New York, pp. 388–392.

Campanini, A., Frost, E., Höjer, S., 2012. Educating the new practitioner: the building of professional identities in European social work. Revista de Asistentă Socială, Social Work Review 1, 33–47.

Dominelli, L., Thomas Bernard, W. (Eds.), 2003. Broadening Horizons: International Exchanges in Social Work. Ashgate, Aldershot.

Dominelli, L., 2007. Contemporary challenges to social work education in the United Kingdom. Australian Social Work 60 (1), 29–45.

Frost, E., 2008. Is there a European social work identity? European Journal of Social Work 4, 341–354.

Frost, E., Höjer, S., Campanini, A., 2012. Readiness for practice: social work students' perspectives in England, Italy, and Sweden. European Journal of Social Work 16 (3), 327–343. Published on line 13 September 2012.

Friesenhahn, G.J., Seibel, F.W., Kniephoff-Knebel, A., 28 November 2012. Europeanising the social professions – networking in practice and education. Social Work and Society. www.socmag.net/?p=769.

González, J., Wagenaar, R. (Eds.), 2003. Tuning Educational Structures in Europe. Final Report. Phase One. University of Deusto, Deusto.

González, J., Wagenaar, R. (Eds.), 2005. Tuning Educational Structures in Europe. Final Report. Phase Two. University of Deusto, Deusto, Spain.

Hamburger, F., Hirschler, S., Sander, G., Wöbcke, M. (Eds.), 2004–2007. Ausbildung für soziale Berufe in Europa, 4 Bände. ISS-Eigenverlag, Frankfurt.

Juliusdottir, S., 2006. The emerging paradigm shift in social work – in the context of the current reform of social work education. Social Work and Society 3 (2). www.socwork.net.

Labontè-Roset, C., Marynowicz-Helka, E., Szmagalski, J. (Eds.), 2003. Social Work Education and Practice in Today's Europe. Slask, Katovice, Poland.

Labonté Roset, C., 2005. The European higher education area and research-oriented social work education. European Journal of Social Work 8 (3), 285–296.

Larsen, A.K., Hole, G.O., 2007. The role of the virtual classroom in opening up the European curriculum. In: Frost, E., Freitas, M.J. (Eds.) with Campanini, A. (guest ed.), Social Work Education in Europe. Carocci Editore, Roma, pp. 73–88.

Lyons, K., 2006. Research and professional education. In: Lyons, K., Lawrence, S. (Eds.), Social Work in Europe: Educating for Change. Venture Press, Birmingham.

Lyons, K., Lawrence, S. (Eds.), 2006. Social Work in Europe: Educating for Change. Venture Press, Birmingham.

Lorenz, W., 2006. Perspectives on European Social Work. Barbara Budrich Publishers, Opladen.

Martinez-Román, M.A., 2007. Social work education in Europe: the Bologna process and the challenges for the future of social work. In: Frost, E., Freitas, M.J. (Eds.) with Campanini, A. (guest ed.), Social Work Education in Europe. Carocci Editore, Roma, pp. 23–37.

Payne, M., Askeland, G.A., 2008. Globalization and International Social Work. Ashgate, Aldershot.

Walliman, I., 2004. Bologna e GATS: Ripercussioni in prospettiva a livello universitario. www.aidoss.org (accessed 04.12.12.).

Relevant Websites

http://www.a3es.pt – A3ES (Agência de Avaliação e Acreditação do Ensino Superior del Portogallo).
www.ehea.info – Bologna Process European Higher Education Area.
www.coe.int – Council of Europe.
www.EHEA.info – European Higher Education Area.
www.europa.eu – European Union.
http://www.ond.vlaanderen.be/hogeronderwijs/bologna – The Bologna Process.
www.jointquality.org – The Joint Quality Initiative.
http://www.unideusto.org/tuning/ – Tuning.
www.virclass.net – Vircampus.

Boltanski, Luc (1940–)

Simon Susen, City University London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides a brief introduction to the life and work of Luc Boltanski, who is widely regarded as one of the most influential French sociologists of the late twentieth and early twenty-first centuries. The article is divided into two main parts. The first part offers a summary of Boltanski's biographical trajectory and academic career. The second part gives a comprehensive overview of Boltanski's major works and intellectual contributions. The concluding section presents a short assessment of Boltanski's overall impact upon the contemporary social sciences, particularly in terms of the international relevance of his writings. The bibliography contains *key* primary and secondary sources, reflecting not only the prolific development of Boltanski's oeuvre, but also the strong influence his work has had, and continues to have, on cutting-edge forms of social and political analysis.

Biographical Facts

Luc Boltanski is widely regarded as one of the most influential French sociologists of the late twentieth and early twenty-first centuries. He is one of the leading figures of the 'pragmatic' tradition within contemporary social and political thought. More specifically, he – along with Laurent Thévenot – is one of the founding figures of an approach that he himself characterizes as the 'pragmatic sociology of critique.'

Boltanski was born in 1940. He studied social sciences at the University of Paris, La Sorbonne, and completed his *Thèse de troisième cycle* in 1968. This dissertation – entitled *Prime éducation et morale de classe* – was supervised by Raymond Aron; it was published in 1969 and subsequently translated into Italian and Spanish. Boltanski was awarded a *Doctorat d'État* in 1981 for his thesis entitled *Les cadres. La formation d'un groupe social*; this study, completed under the supervision of Pierre Ansart, was published in 1982.

Throughout his career Boltanski has been based at the *École des Hautes Études en Sciences Sociales* (EHESS), Paris, France. At the EHESS, he has held three major academic positions: *Chef de travaux* (1965–69), *Maître de conférences* (1970–81), and *Directeur d'études* (since 1982).

Between 1965 and 1984, he was a member of the *Centre de Sociologie Européenne* (EHESS/CNRS), directed by Pierre Bourdieu. In 1985, he – together with Laurent Thévenot – co-founded the *Groupe de Sociologie Politique et Morale* (GSPM, EHESS/CNRS), of which he was the director between 1985 and 1992. At the GSPM, he carried out several research projects and led numerous research programs until its closure in 2013. He has been a Visiting Professor at various universities, both in Europe and in the United States, and he was a member of the Institute for Advanced Study at Princeton University during the academic year 1991–92. Currently, he is a member of the *Institut de Recherche Interdisciplinaire sur les Enjeux Sociaux (Sciences Sociales, Politique, Santé)* (IRIS, EHESS).

In the early 1970s, Boltanski was involved in launching the journal *Actes de la recherche en sciences sociales*, when his research was still profoundly influenced by the works of his academic mentor, Pierre Bourdieu. In the mid-1980s, however, Boltanski dissociated himself from Bourdieu's 'critical sociology,' in order to create his own research program, commonly described as the 'sociology of critique' or, more recently, as the 'pragmatic sociology of critique.'

Between 1965 and 1982, Boltanski's key research interests were directed toward the sociology of social classes and social stratification (mainly within the following areas: bodily and medical practices, education, social classifications, and moral norms). Between 1983 and 2009, his sociological investigations were concerned, primarily, with two areas of interest: first, *different notions of justice*, particularly in relation to disputes and critique; and, second, *transformations of capitalism*, especially those taking place between the early 1960s and the late 1990s. In relation to these research foci, Boltanski has sought to develop a 'sociology of critique,' based on empirical fieldwork undertaken in a number of domains, such as the media, state policies, management, as well as new forms of work and organization.

In 2008, Boltanski delivered the Adorno Lectures at Frankfurt, which were subsequently published as *De la critique. Précis de sociologie de l'émancipation* (Paris: Gallimard, 2009). In 2012, he was awarded the *Lauréat du 1er prix Pétrarque de l'essai France Culture/Le Monde* (2012) for his study *Énigm es et complots. Une enquête à propos d'enquêtes* (Paris: Gallimard, 2012).

As reflected in the themes examined in *Énigmes et complots*, Boltanski has recently embarked upon a critical study of the construction of the modern European nation-state, notably in terms of its systemic capacity to reduce the multiple uncertainties permeating social life. One key issue with which he has been grappling in this context is the question of the extent to which the tension-laden project of the European nation-state has triggered the emergence of 'new forms of representation' in the humanities and social sciences.

Major Works and Contributions

Boltanski has produced a large number of single-authored and co-authored books, edited and co-edited volumes, book chapters, and journal articles. In addition, he has written and published poetry, and more recently, theater plays. For the sake of brevity, the summary provided in this section shall focus on his most influential sociological works.

I

Les cadres. La formation d'un groupe social (Paris: Minuit, 1982) [*The Making of a Class. Cadres in French Society* (Trans. Arthur Goldhammer, Cambridge: Cambridge University Press, 1987)] (Boltanski, 1982; see also Boltanski, 1987[1982]).

As mentioned above, this book is based on the thesis for which Boltanski – under the supervision of Pierre Ansart – was awarded a *Doctorat d'État* in 1981. It provides an in-depth study of *les cadres*, that is, of a powerful social group made up of business leaders, managers, directors, chiefs, supervisors, and executives. One of the defining features of this group is that it projects the image of *a new class*, which is neither bourgeois nor proletarian. Its members may be described as 'highly competent,' 'highly skilled,' 'highly motivated,' and both politically and economically 'highly influential.' Yet, far from portraying them as a homogenous cluster of social actors, Boltanski stresses their *internal diversity*. He does so by drawing upon the information provided in numerous interviews conducted with representatives of this group, enabling him to deconstruct the myth that the emergence of *les cadres* can be regarded as a quasi-natural outcome of social, economic, and technological progress.

One of Boltanski's most significant achievements in this study is to have demonstrated the immense internal heterogeneity, along with the profound structural fragility, characterizing *les cadres*. His fine-grained analysis illustrates that the portrayal of this social group as a uniform and homogenous collective force, as well as its triumphalist celebration as the protagonist of a new meritocratic era based on prosperity and progress, must be rejected as a reductive misrepresentation of what is – in reality – a highly complex, heterogeneous, and volatile assemblage of actors.

Critics may have plausible reservations about the Franco-centric – and, hence, geographically and sociopolitically limited – scope of this enquiry. Indeed, the English translation of the original French *La formation d'un groupe social* (*The Making of a Social Group*) into *The Making of a Class* may – contrary to Boltanski's intentions – convey the misleading impression that *les cadres* form a social class, rather than a social group. Such an assumption seems untenable, given the fragmented and unstable constitution of their material and symbolic resources for action, of their internal organizational structure, and of their members' trajectories. In the contemporary era, a significant sociological challenge consists in exploring the extent to which *les cadres* continue to play a pivotal role in shaping social, economic, political, and ideological developments both in and beyond France.

II

L'amour et la justice comme compètences. Trois essais de sociologie de l'action (Paris: Métailié, 1990) [*Love and Justice as Competences* (Trans. Catherine Porter, Cambridge: Polity, 2012)] (Boltanski, 1990; see also Boltanski, 2012[1990]).

This book is of crucial importance in that it is one of the first works marking Boltanski's unambiguous *rupture* with the sociological approach developed by his academic mentor, Pierre Bourdieu. To be precise, it is Boltanski's first major single-authored study that makes an explicit attempt to challenge the arguably scientistic, positivistic, and fatalistic presuppositions underpinning Bourdieu's 'critical sociology.' In essence, this treatise accomplishes this by arguing that people care a great deal about justice. To be exact, Boltanski starts from the assumption that ordinary actors engage – *enthusiastically and critically* – in everyday disputes over different – *and, in many respects, competing* – conceptions of justice. It is particularly in situations in which people's sense of justice is challenged, affronted, or disturbed that they are likely to engage in subtle or open confrontations with others. Unlike Bourdieu, however, Boltanski posits that – instead of acting mainly in strategic, instrumental, utility-driven, or calculating ways – subjects capable of reflection and justification are able to engage in intersubjectively constituted processes of discourse and argumentation, thereby raising claims to validity in relation to different sets of normativity.

People's practical construction of, intuitive immersion in, and critical engagement with '*regimes of justice*' can be considered central to the value-laden unfolding of social life. No less important, in this respect, is the socio-ontological role of what Boltanski refers to as the '*regime of peace*' and the '*regime of love*' (*agapè*), whose existence is due to the fact that some actions are selfless and gratuitous. What all of these grammatically structured regimes of action – between which people, in their everyday lives, move back and forth – have in common is that they require its protagonists to refuse to draw on their *capacity for violence*, which can manifest itself in various – notably, physical, symbolic, and structural – forms of power.

As members of a 'common humanity,' we are equipped with the normative capacity to establish 'orders of worth' in the pursuit of the 'common good,' which transcends the divisive logic of competitive position-taking and merely strategic performativity. One of the main contributions of Boltanski's *Love and Justice as Competences*, therefore, is to have shed light on the *moral foundations of society* by taking seriously people's ability to engage in the construction of *everyday forms of normativity*. In other words, this book is a powerful reminder that our capacity to mobilize the reflexive resources embedded in our critical capacity permits us to build meaningful social relations based on a genuine concern with justice, love, and reciprocity.

III

De la justification. Les économies de la grandeur (avec Laurent Thévenot, Paris: Gallimard, 1991) [*On Justification. Economies of Worth* (with Laurent Thévenot, trans. Catherine Porter, Princeton: Princeton University Press, 2006)] (Boltanski and Thévenot, 1991; see also Boltanski and Thévenot, 2006[1991]).

This study provides numerous useful insights into the pivotal role that processes of justification play in the construction of social life. These can be synthesized in terms of the following levels of analysis:

1. *Ordinary actors are equipped with critical, moral, and judgmental capacities.* Owing to their ability to participate – actively and reflexively – in the meaning- and value-laden construction of different forms of sociality, their claims to objective, normative, or subjective validity are irreducible to mere epiphenomena of an interest- and power-laden struggle for legitimacy in field-specific – and, hence, positionally determined and dispositionally reproduced – realities.

2. *Any attempt to construct a hierarchy between 'ordinary knowledge' and 'social-scientific knowledge'* is epistemologically erroneous, methodologically counterproductive, sociologically untenable, politically patronizing, and philosophically fatalistic. To be sure, the point is not to deny that there are substantial qualitative differences between 'scientific analysis' and 'common sense.' It is crucial, however, to concede that these two levels of epistemic engagement with specific aspects of reality are not as far apart as they may appear at first glance. Instead of undertaking a clear-cut 'epistemological break' with the doxic illusions of common sense, the challenge consists in exploring the extent to which ordinary people's critical capacity constitutes a precondition for, rather than an obstacle to, the possibility of reflection and justification in all normatively codified settings of social interaction.
3. *All activities of justification have both grammatical and processual dimensions*, which can be empirically studied and conceptually grasped. Given their *grammatical* constitution, activities of justification are structured by context-specific logics of rationalization, argumentation, and valorization. Given their *processual* constitution, the underlying objective, normative, or subjective parameters mobilized in order to justify a belief or an action are not only in a constant state of flux, but also contingent upon the changing sets of circumstances in which they are applied by those making claims to validity and aiming to obtain empowering degrees of legitimacy. Different *cités* (polities) may be regarded as idiosyncratic *mondes* (worlds) capable of establishing different *grandeurs* (orders of worth) with different conceptions of *bien comun* (common good), whose validity can be confirmed or undermined by means of different *épreuves* (tests). Irrespective of the spatiotemporal specificity of a social situation, there are no practices of meaning- and value-laden interaction without both grammars and processes of justification.
4. *There are multiple normative orders with corresponding regimes of justification and modes of evaluation*. Six 'worlds', with corresponding 'orders of worth,' are particularly important: 'the inspired world,' 'the domestic world,' 'the civic world,' 'the world of opinion and fame,' 'the world of the market,' and 'the industrial world.' These 'worlds' possess both a '*quotidian*' and a '*metaphysical*' dimension:
   a. Their '*quotidian*,' and thus '*ordinary*,' constitution is reflected in the fact that these 'worlds' permeate the normative structure of *people's everyday practices*, as they find themselves immersed in different *regimes of action and justification* when navigating their way through the social universe. The experiences of passion ('inspired'), trust ('domestic'), solidarity ('civic'), recognition ('fame'), exchange value ('market'), and productivity ('industry') are built into 'orders of worth' by means of which actors engage with, and attribute meaning to, reality on a day-to-day basis.
   b. Their '*metaphysical*' or simply '*philosophical*' constitution is expressed in the fact that the systematic concern with the ontological significance of these 'worlds' can be traced back to the writings of classical, social, and political thinkers: St. Augustine ('the inspired world'), Jacques-Bénigne Bossuet ('the domestic world'), Jean-Jacques Rousseau ('the civic world'), Thomas Hobbes ('the world of fame'), Adam Smith ('the world of the market'), and Henri de Saint Simon ('the industrial world').
5. *Social actors are obliged to possess a certain degree of realism when engaging in disputes*. Put differently, people's participation in the normative construction of social life is inconceivable without their competence to *assess* what is possible, and what is not possible, when faced with a given set of materially and symbolically organized circumstances. People's capacity to be *realistic* in terms of what they can, and cannot, achieve within particular situations is a praxeological precondition for their ability to make judgments about – and, if required, take decisions in relation to – specific issues at stake in changing settings of interaction. Just as they are obliged to make compromises, they are expected to be able to cope with the fact that overt or hidden conflicts form an ineluctable component of social existence. Since, in their everyday lives, they are constantly required to position themselves in relation to normatively codified forms of action, they cannot escape the need to engage in processes of justification.

IV

La souffrance à distance. Morale humanitaire, medias et politique (Paris: Métailié, 1993; Paris: Gallimard, 2007 (Extended version)) [*Distant Suffering: Morality, Media and Politics* (Trans. Graham D. Burchell, Cambridge: Cambridge University Press, 1999)] (Boltanski, 1993; see also Boltanski, 1999[1993]).

This book provides an interdisciplinary analysis of 'distant suffering,' that is, of the experience and effects of perceiving processes of human grief and misery 'from a distance.' Perhaps, the most fundamental sociological issue with which Boltanski grapples in this study is the question of how human actors react when exposed to spectacles of suffering, while being geographically remote from the locations in which tragic or catastrophic events occur. Seeking to respond to this question, Boltanski unearths various sociological, political, moral, psychological, and fictional accounts concerned with the impact of 'distant suffering' upon those who experience it.

The book comprises three main parts. In Part I, entitled 'The Question of the Spectator,' Boltanski explores the normative issues arising from a set of principles and practices to which he refers as 'the politics of pity.' In Part II, entitled 'The Topics of Suffering,' Boltanski draws on literacy sources to examine several intermediary elements that influence the spectator's rational and emotional reactions to gruesome media portrayals. In Part III, entitled 'The Crisis of Pity,' Boltanski reflects on the implications of the fact that spectators can be converted into moral and political actors, particularly when passing value judgments on the alleged facts and happenings to which they are exposed via the media.

In the digital age, those who have direct and regular access to the mass media are habitually exposed to horrific images – such as starving children, bombed villages, war, genocide, and mass graves. It is far from clear, however, what it means to respond accurately and responsibly to such scenarios. For Boltanski, there is no doubt that, in order to overcome

a paralyzing state of sensationalism and voyeurism, spectators must rise above mere feelings of empathy and compassion. The 'politics of pity' is unsatisfactory in the sense that it encourages consumers of mediated misery to focus on the spectacle of suffering, thereby requiring them to observe the unfortunate, rather than to scrutinize – let alone to act upon – the socio-historical roots behind their deprivation. Boltanski, therefore, urges his readers to imagine possibilities of action and thereby challenge the constraining limitations, and detrimental consequences, of the largely passive consumption of information. Granted, the shift from a potentially disempowering 'world of representation' to a genuinely empowering 'world of action' is complex. Yet, the Boltanskian idea of a political and moral sociology cannot be dissociated from the conviction that subjects capable of reflection and justification are able to mobilize their critical resources in order to engage in normatively defensible and performatively empowering forms of action.

V

Le nouvel esprit du capitalisme (avec Ève Chiapello, Paris: Gallimard, 1999) [*The New Spirit of Capitalism* (with Ève Chiapello, Trans. Gregory Elliott, London: Verso, 2005)] (Boltanski and Chiapello, 1999; see also Boltanski and Chiapello, 2005[1999]).

This book provides a cutting-edge analysis of the emergence of what Boltanski and Chiapello describe as the 'new spirit of capitalism' in the late twentieth century. Shortly after its original publication in 1999, this study became a bestseller in France. According to Boltanski and Chiapello's analysis, we can distinguish *three spirits of capitalism*:

- The *first spirit*, prevalent in early modern societies, can be characterized as *family capitalism*, in the sense that it prioritizes the individual figure of the bourgeois proprietor and finds its ideological justification, above all, in the 'domestic city.' It is intimately interrelated with the *productive ethos* of Weber's famous Protestant Ethic. Sweeping away the rigid social, political, and economic structures of feudal-absolutist formations, the constitutive component of the 'first spirit of capitalism' is *productivism*.
- The *second spirit* can be referred to as *industrial or organizational capitalism*, epitomized in the protagonist role of 'the manager,' whose societal function is associated with 'organization man.' Emerging in response to the crisis of 1929–30, it is composed of a combination of *Fordist industrialism* and *Keynesian interventionism*, which may be interpreted as a trade-off between Rousseau's 'civic city' and St. Simon's 'industrial city.' The societal constellation generated by this historic settlement had two major consequences: (a) it contributed to enhancing the acquisitive power of the working classes in particular and people's chances to benefit from upward social mobility in general; and (b) it contributed to the rise of a relatively autonomous salaried professional labor force, especially in the liberal professions, arts and sciences, and public sector.
- The *third spirit* manifests itself, most clearly, in the *city of projects*, in which market-driven principles – such as 'flexibility,' 'adaptability,' 'creativity,' and 'mobility' – play a pivotal role in developing an ever-more elastic, and seemingly forward-looking, capitalist system. Also described as the *new spirit of capitalism*, it is inextricably linked to the rise of neoliberalism and neomanagerialism, especially from the 1970s onward, indicating the restoration of large-scale *market discipline* along with a shift toward the increasing *financialization* of capital flows. One of the paradoxical achievements of this 'new spirit' is to have succeeded in appropriating the subversive forces that sought to undermine the legitimacy of capitalism for its own purposes. The elastic and flexible nature of this 'new spirit' emanates from capitalism's capacity to promote and integrate discursive processes of debate and critique, thereby ensuring that, as a politico-economic system, it is both structurally and ideologically highly adaptable. The idea of *dominating by change* is essential to contemporary forms of social domination.

The rise of the 'network man' illustrates not only the emergence of a new systemic and ideological *modus operandi* of capitalism, but also its new spirit's capacity to take seriously four sources of indignation: *inauthenticity*, *oppression*, *misery and inequality*, and *egoism*. The former two were central objects of different versions of *artistic critique* and the latter two were principal matters of concern under the umbrella of *social critique*. It is one of Boltanski and Chiapello's noteworthy achievements to have demonstrated, with considerable empirical evidence and conceptual precision, that most contemporary forms of capitalism possess the capacity to *incorporate* normative processes based on critical discourse into their mode of functioning. In this sense, categorical openness to debate, controversy, and constant re-assessment has been converted into one of the normative cornerstones underlying the 'new spirit of capitalism.'

VI

La condition fœtale. Une sociologie de l'engendrement et de l'avortement (Paris: Gallimard, 2004) [*The Foetal Condition* (Trans. Catherine Porter, Cambridge: Polity, 2013)] (Boltanski, 2004; see also Boltanski, 2013[2004]).

Undoubtedly, this is one of Boltanski's most controversial books – possibly, because it deals with one of the most contentious issues in contemporary society: *abortion*. In this study, Boltanski draws upon accounts and statements collected from hospital settings as well as upon in-depth interviews conducted with women who have undergone abortions. In his insightful socio-philosophical interpretation of the discursive data upon which this enquiry is based, he directs his readers' attention to the profound *ambivalence* that appears to be built into abortion as a social practice. To be precise, for Boltanski, abortion exposes a contradiction that is inherent in all human life forms: on the one hand, we assume that individual human beings are *unique* and *distinctive*; on the other hand, we are confronted with their *replaceable* and *disposable* nature, without which there would be no demographic renewal and no societal regeneration.

Boltanski, therefore, proposes to examine the ways in which human beings are engendered by dissecting the symbolically mediated controls and constraints that are imposed upon them by society, of which they can become fully-fledged members

only to the extent that they are both willing and able to share its – normatively charged – conception of species-constitutive existence. On this view, a fetus is not a human being 'in itself,' ensconced within the female body, but, rather, a human being 'for itself,' to the degree that it is *symbolically constructed* and *discursively considered* as such by the members of a given society. For Boltanski, one twofold categorization is particularly important in this regard: the *project fetus* and the *tumoral fetus*. The former is desired by its parents, who attribute positive characteristics – such as 'meaningfulness,' 'fulfilment,' 'love,' and 'life plans' – to its existence. The latter is deprived of the privilege of forming an integral element of a parental endeavor and, in extreme-case scenarios, may be reducible to a nameless, replaceable, and undesired form of being, whose non-existence is preferred to its existence by those who have the power to decide over its future.

In the human world, then, sexual reproduction is never simply a *biological* affair but always also a process of *social* construction, especially in terms of how it is both interpreted and regulated by members of particular cultural life forms. Boltanski demonstrates, in a neo Durkheimian fashion, that every *social order* constitutes a *moral order*: a set of interrelated practices performed by ethically responsible actors whose decisions – irrespective of whether these are made consciously and unconsciously – have normative implications both for those who undertake them and for those who are, directly or indirectly, affected by them. Whatever one makes of Boltanski's analysis, owing to the contentious nature of this subject, *The Foetal Condition* cannot fail to challenge – and, in some cases, irritate – those contributing to contemporary controversies concerning abortion.

VII

De la critique. Précis de sociologie de l'émancipation (Paris: Gallimard, 2009) [*On Critique: A Sociology of Emancipation* (Trans. Gregory Elliott, Cambridge: Polity, 2011)] (Boltanski, 2009; see also Boltanski, 2011[2009])

Arguably, *On Critique* is Boltanski's most philosophical book. It provides an in-depth analysis of the conceptual underpinnings of the 'pragmatic sociology of critique,' focusing on the following six key dimensions:

1. Boltanski reflects on the task of *critical theories*. One of their vital concerns is the sustained effort to scrutinize the causes, symptoms, and consequences of *power relations* within particular historical contexts, especially those that are entrenched in societal systems of domination. In this respect, a fundamental difference between Bourdieu and Boltanski becomes evident. According to the former, ordinary people are largely *unconscious* of the workings, and essentially *naive* about the implications, of power relations. According to the latter, ordinary people are not only *conscious* of, and *realistic* about, power relations but also able to problematize the tangible implications of their existence. For Bourdieu, it is the task of 'critical sociologists' to uncover the underlying mechanisms that determine the asymmetrical structures permeating the interest-laden practices of *strategic agents*, who compete for material and symbolic resources. For Boltanski, by contrast, it is the mission of 'sociologists of critique' to recognize that human beings are *moral and reflexive actors*, whose critical capacity permits them to assess – and, if necessary, justify – the normative validity of their performances.
2. Boltanski aims to *reconcile Bourdieu's critical sociology with his own pragmatic sociology of critique*. Thus, he seeks to combine and cross-fertilize two seemingly antagonistic approaches. The former appears to advocate *social determinism* and *positivist scientism*, favoring the epistemic capacities of scientists over those of everyday actors, who seem to be deluded by doxic preconceptions based on common sense. The latter appears to endorse *social pragmatism* and *interpretivist normativism*, proposing to take ordinary people seriously in terms of both their performative capacity to shape the world and their discursive capacity to provide reasonable justifications for their beliefs and actions. Rather than conceiving of these two sociological approaches as diametrically opposed and irreconcilable, Boltanski aims to demonstrate that useful insights can be gained not only from comparing and contrasting, but also from combining and integrating, these two paradigmatic frameworks.
3. Boltanski grapples with the principal functions of *social institutions*. Their most essential task, it seems, consists in producing solidified – or, at least, seemingly solidified – realms of social interaction, enabling humans to cope with the uncertainty inherent in all worldly life forms. Boltanski draws an important distinction between 'world' ('*monde*') and 'reality' ('*réalité*'). While the former encompasses 'everything that is the case,' the latter comprises 'everything that is constructed.' Put differently, the world is 'everything that happens *to* people,' while reality is 'everything that is constructed *by* people.' To the extent that institutions convert our encounter with the world into an experience founded on the illusion of relative certainty, they can be regarded as a *conditio sine qua non* of the material and symbolic construction of reality. Institutions, therefore, constitute 'bodiless beings' that fulfill the task of defining what Boltanski calls 'the whatness of what is' or, to be exact, 'the whatness of what appears to be.'
4. Boltanski examines the role of *critique* in the normative consolidation of social life. Critique constitutes a driving force of historical change: it permits both individual and collective actors to shape the development of society in accordance with their discursively articulated search for principles that are defensible in terms of their practical worth and normative validity. For Boltanski, two registers of action are crucial in this respect. On the one hand, the *practical register* is characterized by relatively *weak and rudimentary levels of reflexivity*, presupposing a considerable tolerance for differences and discrepancies, as well as sustaining a set of codified arrangements that guarantee the reproduction of society. On the other hand, the *metapragmatic register* is marked by rather *elevated and differentiated levels of reflexivity*, involving an implicit or explicit reference to the normative force of critique and, at the same time, allowing for the articulation of two metapragmatic forces: *confirmation* and *justification*. People's ability to confirm and justify the legitimacy of their actions is central to their capacity to participate in the construction of normatively regulated constellations.

Confronted with *hermeneutic contradictions* emanating from the potential discrepancies between 'world' and 'reality,' human actors are in a position to question the apparent givenness of objectivity by facing up to the genuine arbitrariness of all forms of normativity. Illustrating the 'pragmatic' dimension of Boltanski's framework, it is crucial to recognize that processes of critique cannot be dissociated from three types of 'test' (*épreuve)* undertaken either to *reinforce* or to *undermine* the legitimacy of a specific ensemble of social constellations: (a) 'truth tests' (*épreuves de vérité*), (b) 'reality tests' (*épreuves de réalité*), and (c) 'existential tests' (*épreuves existentielles*). From a Boltanskian perspective, the emancipatory transformation of society is inconceivable without a critical engagement with the normative constitution of reality.

5. Boltanski discusses the concept of *domination*. More specifically, he draws a distinction between two fundamental types of domination: *simple domination* or *primitive domination*, on the one hand, and *complex domination* or *managerial domination*, on the other.

   '*Simple forms of domination are monolithic* in the sense that, under their authority, control over a particular population is monopolized by a state or overarching institution. Here, people are deprived of fundamental liberties (such as freedom of speech, expression, and communication) as well as of basic rights (such as civil, political, social, economic, and human rights). Under regimes of 'simple domination,' the exercise of power is relatively *arbitrary* and unambiguously *asymmetrical*. Obvious historical examples of this type of domination include absolutism, fascism, and any kind of dictatorial government whose exercise of power is motivated by normative principles based on political authoritarianism.
   '*Complex forms of domination are polycentric* – or, in a more radical sense, even *centreless* – in the sense that, under their influence, power structures are circular, amorphous, volatile, and in a constant state of flux, lacking an institutional or ideological epicenter. Here, people's essential liberties and rights are not only largely respected, or even defended, but also instrumentalized in order to foster the legitimacy of the hegemonic political and economic system in place. Under regimes of 'complex domination,' the exercise of power is – at least in principle – *democratic* and – albeit, admittedly, to varying degrees – *criticizable*. Contemporary scenarios that can be described in these terms are *democratic–capitalist societies*, shaped by cultures and institutions based on political pluralism and, hence, by the fact that critique is incorporated into the routines of everyday life. For Boltanski, then, the emergence of the 'new spirit of capitalism' is inextricably linked to the possibility of *dominating by change*, which is based on categorical openness to criticism and, thus, on the capacity to obtain legitimacy by advocating the aforementioned neomanagerial ideals, such as 'flexibility,' 'adaptability,' 'creativity,' and 'mobility.'

6. Boltanski expresses his own view of the conditions underlying processes of human *emancipation*. In his eyes, these processes are based on individual or collective practices that *promote the critical project of a reduction in the material and symbolic privileges of dominant social groups* and thereby *contribute to a more balanced distribution of capacities for action*. To be sure, Boltanski does not propose a utopian blueprint envisaging the construction of a perfect society. Owing to the anthropological optimism that undergirds his writings, however, he dares to believe that the construction of a world based on emancipatory life forms is both desirable and possible. Such a world would not be determined by constraining sources of social domination, such as privilege, status, and authority. Rather, it would be shaped, above all, by people's purposive, cooperative, and creative capacities allowing for individual and collective experiences of self-realization.

VIII

Enigmes et complots. Une enquête à propos d'enquêtes (Paris: Gallimard, 2012) [*Mysteries and Conspiracies: Detective Stories, Spy Novels and the Making of Modern Societies*, trans. Catherine Porter (Cambridge: Polity, 2014)] (Boltanski, 2012; see also Boltanski, 2014[2012]).

In this book, Boltanski draws an analogy between two domains of modern writing, which, at first glance, do not appear to have anything significant in common: on the one hand, the development of two *literary genres*, namely *detective stories*, which are based on methodical enquiries, as well as *spy novels*, which are built around plots and conspiracies; on the other hand, the development of the *human and social sciences*, which are founded not only on systematic investigations but also on what may be described as the 'hermeneutics of suspicion.' Particularly important, in this regard, are the following three scientific disciplines: *psychiatry*, known for fabricating seemingly evidence-based conceptions of paranoia; *sociology*, inspired by the enlightening mission to uncover the underlying causal forces that determine both the constitution and the evolution of the social world; and *political science*, seeking to explain the origins of major historical events by reference to conspiracy theories.

Thus, what *detective stories* and *spy novels* have in common with the *human and social sciences* is not only the fact that they emerged – and underwent profound paradigmatic transitions – in the same *historical context*, that is, in the late nineteenth and the early twentieth centuries, but also the fact that they are driven by the ambition to shed light on the *mysteries and conspiracies* whose existence escapes people's ordinary perception of reality. Hence, they seek to call taken-for-granted assumptions about the world into question, by providing logically coherent accounts, and evidence-based explanations, of the multiple factors influencing different patterns of human action.

According to Boltanski, the most powerful institutional expression of the attempt to organize and unify reality by regulating and controlling the behavior of a population living within a given territory is the modern nation-state. Central to his socio-historical analysis in this book is the assumption that, in the modern age, speculations and suspicions about *conspiracies* became a motivational driving force behind both popular and academic conceptions of the exercise of power. Inevitably, the search for hidden sources of influence involved the ideological construction of a dichotomously constituted reality: on the one hand, an *official reality*, based on appearances, public performances, and superficial impressions; on the other hand, an *unofficial reality*, founded on underlying structures, hidden

causal mechanisms, and concealed social forces. What crime and spy fiction have in common with positivist conceptions of science, then, is that they presuppose a *discrepancy* between these two levels of reality: 'appearance' *versus* 'substance,' 'interpretation' *versus* 'explanation,' 'imagination' *versus* 'observation,' 'fiction' *versus* 'authenticity.'

By putting the works of major detective and spy novelists – such as G.K. Chesterton, Arthur Conan Doyle, John Le Carré, and Graham Greene – under sociological scrutiny, Boltanski demonstrates that their writings reveal fundamental features not only of fiction-based genre, but also of modern society, especially with respect to the reciprocal relationship between modern institutions and modern science. According to Boltanski, the nation-state's ambition to exercise unlimited control over the reality constructed within the boundaries of its territory is aimed at the *stabilization* of volatile sets of social relations. The binary distinction between 'normal' and 'deviant,' which undergirds this socio-political project of the state, is as central to detective and spy novels as it is to the functionalist spirit permeating the works of the founding figures of sociology – that is, the writings of Karl Marx, Émile Durkheim, and Max Weber.

Boltanski emphasizes that *detective stories* and *spy novels* are fundamentally different in the following sense: within the former, the state tends to be portrayed as essentially 'apolitical'; within the latter, the state tends to be conceived of as deeply 'political' or even as a 'war state.' Given the profound uncertainty permeating capitalist societies, which are characterized by processes of constant and rapid technological and demographic transformation, spy novels and the social sciences serve a complementary function: in the early modern era, the widespread diffusion of the term 'paranoia' is inextricably linked to the paradigms of 'conspiracy' and 'suspicion,' which inform the exploratory spirit pervading both spy novels and positivistically inspired social science – especially psychiatry, sociology, and political science. Boltanski has illustrated – with great skill and considerable eloquence – that the *hermeneutics of investigation*, which one encounters in detective stories, and the *hermeneutics of conspiracy*, which is central to most spy novels, contain significant historical and presuppositional similarities with the *hermeneutics of suspicion*, which lies at the heart of the founding disciplines of the human and social sciences.

Conclusion

Given the wide-ranging scope and scholarly originality of Boltanski's writings, it may hardly be surprising that he is generally regarded as one of the most prominent contemporary French sociologists. In fact, his considerable influence manifests itself in the emergence of an extensive secondary literature concerned with the multifaceted aspects of his oeuvre. His influence spans far beyond Francophone (see, e.g.: Bénatouïl, 1999a; Gadrey et al., 2001; Nachi, 2006; Negri, 1994; Susen, 2012; Thévenot, 2006) spheres of social and political thought. Indeed, his international impact on current academic debates is reflected, particularly, in recent and ongoing Germanophone (see, e.g.: Bogusz, 2010; Boltanski and Honneth, 2009; Celikates, 2009: esp. 136–157) and Anglophone (see, e.g.: Bénatouïl, 1999b; Blokker, 2011; Honneth, 2010; Susen, 2007: 7, 146n.8, 147n.31, 167n.5, 202n.89, 202n.93, 223–224, 227n.25, 228n.50, 229n.51, 229n.52, 271n.24, 319, 322, and 325; Susen, 2011b: 370; Susen, 2011a: esp. 447–450, 453–456, and 459–461; Susen and Turner, 2014; Wagner, 1999) controversies concerned with both the empirical and the conceptual significance of his various contributions to the humanities and social sciences. While it would be erroneous to reduce Boltanski's project to a mere – albeit sophisticated – response to the work of his academic 'patron,' Bourdieu, there is no doubt that his proposed paradigm shift from 'critical sociology' to a 'pragmatic sociology of critique' has opened hitherto unexplored intellectual avenues in the attempt to do justice to the pivotal role that critical capacity plays not only in the pursuit of sociology, but also, more fundamentally, in the daily construction of society.

Note

A more detailed version of this article was published in Susen and Turner (2014); see Susen (2014).

See also: Bourdieu, Pierre (1930–2002); Capitalism; Class Consciousness; Control, Social; Critical Theory: Contemporary; Culture, Sociology of; Economic Sociology; Inequality, Social; Institutions; Knowledge, Sociology of; Legitimacy, Sociology of; Methodological Individualism in Sociology; Positivism, Sociological; Power in Society; Social System; Sociological Theory; Sociology, Epistemology of.

Bibliography

Bénatouïl, Thomas, Mars–Avril 1999a. Critique et pragmatique en sociologie. Quelques principes de lecture. Annales HSS 2, 281–317.
Bénatouïl, Thomas, 1999b. A tale of two sociologies: the critical and the pragmatic stance in contemporary French sociology. European Journal of Social Theory 2 (3), 379–396.
Blokker, Paul, 2011. Pragmatic sociology: theoretical evolvement and empirical application. European Journal of Social Theory 14 (3), 251–261.
Bogusz, Tanja, 2010. Zur Aktualität von Luc Boltanski. Einleitung in sein Werk. VS Verlag für Sozialwissenschaften, Wiesbaden.
Boltanski, Luc, 1982. Les Cadres. La formation d'un groupe social. Minuit, Paris.
Boltanski, Luc, 1987[1982]. The Making of a Class: Cadres in French Society (Arthur Goldhammer, Trans.). Cambridge University Press, Cambridge.
Boltanski, Luc, 1990. L'amour et la justice comme compétences. Métailié, Paris.
Boltanski, Luc, 1993. La souffrance à distance. Métailié, Paris.
Boltanski, Luc, 1999[1993]. Distant Suffering: Morality, Media and Politics (Graham Burchell, Trans.). Cambridge University Press, Cambridge.
Boltanski, Luc, 2004. La condition fœtale. Une sociologie de l'engendrement et de l'avortement. Gallimard, Paris.
Boltanski, Luc, 2009. De la critique. Précis de sociologie de l'émancipation. Gallimard, Paris.
Boltanski, Luc, 2011[2009]. On Critique: A Sociology of Emancipation (Gregory Elliott, Trans.). Polity, Cambridge.
Boltanski, Luc, 2012. Énigmes et complots. Une enquête à propos d'enquêtes. Gallimard, Paris.
Boltanski, Luc, 2012[1990]. Love and Justice as Competences (Catherine Porter, Trans.). Polity, Cambridge.
Boltanski, Luc, 2013[2004]. The Foetal Condition: A Sociology of Engendering and Abortion (Catherine Porter, Trans.). Polity, Cambridge.
Boltanski, Luc, 2014[2012]. Mysteries and Conspiracies: Detective Stories, Spy Novels and the Making of Modern Societies (Catherine Porter, Trans.). Polity, Cambridge.
Boltanski, Luc, Chiapello, Ève, 1999. Le nouvel esprit du capitalisme. Gallimard, Paris.
Boltanski, Luc, Chiapello, Ève, 2005[1999]. The New Spirit of Capitalism (Gregory Elliott, Trans.). Verso, London.

Boltanski, Luc, Honneth, Axel, 2009. Soziologie der Kritik oder Kritische Theorie? Ein Gespräch mit Robin Celikates. In: Jaeggi, Rahel, Wesche, Tilo (Eds.), Was ist Kritik? Suhrkamp, Frankfurt am Main, pp. 81–114.

Boltanski, Luc, Thévenot, Laurent, 1991. De la justification. Les économies de la grandeur. Gallimard, Paris.

Boltanski, Luc, Thévenot, Laurent, 2006[1991]. On Justification: Economies of Worth (Catherine Porter, Trans.). Princeton University Press, Princeton.

Celikates, Robin, 2009. Kritik als soziale Praxis. Gesellschaftliche Selbstverständigung und kritische Theorie. Campus Verlag, Frankfurt am Main.

Gadrey, Jean, Hatchuel, Armand, Boltanski, Luc, Chiapello, Ève, 2001. Symposium sur *Le nouvel esprit du capitalisme* de Luc Boltanski et Ève Chiapello. Sociologie du travail 43 (3), 389–421.

Honneth, Axel, 2010. Dissolutions of the social: on the social theory of Luc Boltanski and Laurent Thévenot. Constellations 17 (3), 376–389.

Nachi, Mohamed, 2006. Introduction à la sociologie pragmatique. Armand Colin, Paris.

Negri, Toni, 1994. Relire Boltanski et Thévenot: sociologie et philosophie politique. Multitudes(Janvier).

Susen, Simon, 2007. The Foundations of the Social: Between Critical Theory and Reflexive Sociology. Bardwell Press, Oxford.

Susen, Simon, 2011a. *Kritische Gesellschaftstheorie* or *kritische Gesellschaftspraxis*? Robin Celikates, *Kritik als soziale Praxis. Gesellschaftliche Selbstverständigung und kritische Theorie* (Frankfurt am Main, Campus Verlag, 2009). Archives Européennes de Sociologie/European Journal of Sociology 52 (3), 447–463.

Susen, Simon, 2011b. Afterword: concluding reflections on the legacy of Pierre Bourdieu. In: Susen, Simon, Turner, Bryan S. (Eds.), The Legacy of Pierre Bourdieu: Critical Essays. Anthem Press, London, pp. 367–409.

Susen, Simon, 2012. Une sociologie pragmatique de la critique est-elle possible? Quelques réflexions sur *De la critique* de Luc Boltanski. Revue Philosophique de Louvain 110 (4), 685–728.

Susen, Simon, 2014. Luc Boltanski: his life and work – an overview. In: Susen, Simon, Turner, Bryan S. (Eds.), The Spirit of Luc Boltanski: Essays on the 'Pragmatic Sociology of Critique'. Anthem Press, London, pp. 3–28.

Susen, Simon, Turner, Bryan S. (Eds.), 2014. The Spirit of Luc Boltanski: Essays on the 'Pragmatic Sociology of Critique'. Anthem Press, London.

Thévenot, Laurent, 2006. L'action au pluriel. Sociologie des régimes d'engagement. La Découverte, Paris.

Wagner, Peter, 1999. After *Justification*: repertoires of evaluation and the sociology of modernity. European Journal of Social Theory 2 (3), 341–357.

Borderline (Patient) Personality

Otto F Kernberg, Personality Disorders Institute, New York Presbyterian Hospital, New York, NY, USA; Joan and Sanford I. Weill Medical College of Cornell University, New York, NY, USA; and Columbia University Center for Psychoanalytic Training and Research, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes the borderline personality disorder (BPD) as defined in the DSM-IV and DSM-V classification system and points to the present major shift in the conceptualization of personality disorders. The present, dimensionally oriented classification of personality disorders stresses the pathology of the self and of the relationship with others as central organizing concepts regarding the severity of these disorders. Borderline personality organization is described as the structural intrapsychic organization of experience that explains the pathology of self and object relations characteristic of the BPD. The clinical differential diagnosis of this disorder, its etiological features, and treatment approaches with psychotherapeutic methods and psychopharmacological intervention completes this overview.

The borderline personality disorder (BPD) is a prevalent personality disorder in clinical psychiatric practice; its prevalence is approximately 1–2% of the general population, at least in the advanced English-speaking democracies. It shares with all personality disorders an enduring pattern of rigid and maladaptive behaviors and inner experiences that deviate significantly from cultural norms and lead to significant distress and/or impairment in patients' social life, work or profession, and intimate relations. The Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR) and DSM-5 (to be introduced in 2014) describe the following criteria for the specific features of the BPD (Table 1).

A Major Conceptual Shift

The committee developing the new classification of personality disorders for the forthcoming DSM-V started from several assumptions: first, that the DSM-IV classification had proved unsatisfactory because of the high comorbidity found regarding the various personality disorders, and the fact that the most frequent diagnostic conclusion, in clinical practice, was personality disorder NOS, that is, 'not otherwise specified.' Second, within the committee there continued an old dynamic tension between empirical researchers interested in developing classification systems for the personality characteristics of normal populations, on the one hand, and clinicians who were concerned about developing a classification system that would do justice to the personality disorder constellations found in clinical settings.

The dynamic tension between the researchers interested in the empirical studies of normal populations, relating them to the predominant prototypes of personality disorders, and clinical psychiatry trying to preserve what they saw as their confirmed experience of the validity of major categories described in DSM-IV, evolved in the direction of a compromise, that included an important, major new development, namely, the agreement on a common basic factor of all personality disorders, a factor that represented a major criterion for the assessment of the severity of any personality disorder: the integration or lack of integration of the *self*, that is, of normal identity, and the degree of normality or pathology in the relations of the individual with others (Bender et al., 2011). Here, what might be called the common sense observation that patients with personality disorders have difficulty in their comprehension and management of themselves and with their comprehension and management of the relationship with

Table 1 DSM-IV and DSM-5 diagnostic criteria for borderline personality disorder

A pervasive pattern of instability of interpersonal relationships, self-image, and affects marked by impulsivity beginning by early adulthood and present in a variety of contexts, as indicated by five (or more) of the following criteria:

1. Frantic efforts to avoid real or imagined abandonment. Note: Do not include suicidal or self-mutilating behavior covered in criterion 5.
2. A pattern of unstable and intense interpersonal relationships characterized by alternating between extremes of idealization and devaluation.
3. Identity disturbance: markedly and persistently unstable self-image or sense of self.
4. Impulsivity in at least two areas that are potentially self-damaging, such as spending, sex, substance abuse, reckless driving, binge eating. Note: Do not include suicidal or self-mutilating covered in criterion 5.
5. Recurrent suicidal behavior, gestures, or threats, or self-mutilating behavior.
6. Affective instability due to a marked reactivity of mood (for example, intense episodic dysphoria, irritability, or anxiety usually lasting a few hours and only rarely more than a few days).
7. Chronic feelings of emptiness.
8. Inappropriate intense anger or difficulty controlling anger (for example, frequent displays of temper, constant anger, recurrent physical fights).
9. Transient, stress-related, paranoid ideation, or severe dissociative symptoms.

Source: American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Association, Washington, DC, p. 663.

significant others was recognized, for the first time, as a basic characteristic of personality disorders. The fact that this dimension could be explicated, operationalized, and clinically evaluated in terms of the degree of its disturbance satisfied both empirical dimensionalists, and, particularly, psychodynamic psychotherapists, who, for over 30 years, have observed, described, and utilized this dimension in their assessment and therapeutic approaches to personality-disordered patients (Kernberg, 1975, 1980).

In summary, the proposed DSM-V personality disorder model includes two fundamental assessments: criteria A the general level of personality functioning scale, including the assessment of self and of interpersonal functioning; and criteria B the diagnostic criteria for six selected specific categories of personality disorders.

The assessment of degrees of pathology of the experience of self includes the following components: (1) Identity: the experience of oneself as unique, with clear boundaries between self and others; self-esteem and accuracy of self-appraisal; capacity for, and ability to regulate the range of emotional experience. (2) Self-direction: the pursuit of coherent and meaningful short term and life goals; utilization of constructive and prosocial internal standards of behavior; ability to self-reflect productively. The assessment of interpersonal functioning includes: (1) Empathy: comprehension and application of others' experiences and motivations; tolerance of different perspectives; understanding of the effects of own behavior on others. (2) Intimacy: depth and duration of positive connection with others; desire and capacity for closeness; mutuality of regard reflected in interpersonal behavior. In short, the components proposed to be central to a personality functioning continuum are identity, self-direction, empathy, and intimacy and the corresponding scale differentiates five levels of impairment, going from nonimpairment (healthy functioning), to extreme impairment.

The diagnostic criteria for six specific personality disorders types: antisocial, avoidant, borderline, narcissistic, obsessive compulsive, and schizotypal, are defined by the level of personality functioning reflected by the self (the A criteria) and by the corresponding specific pathological personality traits (the B criteria).

Basic Underlying Structure of the BPD: Borderline Personality Organization

The term borderline personality organization corresponds to the common features of lack of integration of the self-concept and lack of integration of the concepts of significant others that characterize all severe personality disorders, including the BPD as a typical expression of this pathology.

'Borderline personality organization' is characterized by pathological identity formation (identity diffusion), primitive defensive operations, and varying degrees of pathology of internalized value systems in the setting of maintained but somewhat reduced reality testing marked by a decreased capacity for subtle and tactful evaluation of interpersonal processes, particularly in the setting of more intimate relations. This level of personality organization includes all the severe personality disorders seen in clinical practice. Typical personality disorders included here are the BPD, the schizoid, and schizotypal personality disorders, the paranoid personality disorder, the hypomanic personality disorder, hypochondriasis (a syndrome which has many characteristics of a personality disorder proper), the narcissistic personality disorder (including the malignant narcissism syndrome (Kernberg, 1992)), and the antisocial personality disorder. The antisocial personality, the syndrome of malignant narcissism, and many of the narcissistic personalities are characterized in addition by significant pathology of internalized value systems (Kernberg, 1989, 1992).

From a clinical standpoint, the syndrome of identity diffusion represents the dominant characteristics of borderline personality organization and, thus, of the severe personality disorders as a group, including the BPD. In particular, in the context of pathological aggression, we see a poorly integrated, superficial, and unstable sense of self and others, and limited affect dispositions marked by a combination of intensity and superficiality in the setting of the predominance of negative affects. These core features of the severe personality disorders reflect the splitting or primitive dissociation of an idealized 'good,' segment of experience from a 'bad,' paranoid one. Splitting mechanisms are naturally reinforced by other primitive defensive operations intimately connected with splitting (projective identification, denial, primitive idealization, devaluation, omnipotence, and omnipotent control). This entire constellation of defensive operations serves to distort interpersonal interactions, to create chronic disturbances in interpersonal relations, and to interfere with the capacity to assess other people's behavior and motivations in depth, particularly under the impact of intense affect activation. The lack of integration of the concept of the self interferes with a comprehensive integration of one's past and present into a capacity to predict one's future behavior and decreases the capacity for commitment to professional goals, personal interests, work and social functions, and intimate relationships.

The lack of integration of the concept of significant others interferes with the capacity for realistic assessment of others, for selecting partners harmonious with the individual's actual expectations, and for investment in others. The predominance of negative affect dispositions leads to an infiltration of sexual intimacy by excessive aggressive components. The outcome is, frequently, an exaggerated and chaotic interest in polymorphous perverse sexual practices as part of the individual's sexual repertoire. In more severe cases, we see a primary inhibition of the capacity for sensual responsiveness and erotic enjoyment. Under these latter circumstances, overwhelming negative affect states eliminate the very capacity for erotic response, leading to the severe types of sexual inhibition that are to be found in the most severe personality disorders.

The lack of integration of the concept of self and of significant others also interferes with the internalization of the early layers of the system of internalized values. This leads to a particularly exaggerated quality of the idealization of positive values and ideals and to an extremely persecutory quality of prohibitions. These developments lead, in turn, to a predominance of splitting mechanisms at the level of internalized value systems, with excessive projection of internalized prohibitions. At the same time, excessive, idealized demands for perfection further interfere with the integration of a normal superego.

The BPD and the schizoid personality disorder may be described as the simplest forms of severe personality disorders. These disorders reflect identity diffusion in the setting of the predominance of splitting mechanisms and can be seen as the 'purest' expression of the general characteristics of borderline personality organization.

Fairbairn (1954) described the schizoid personality as the prototype of all personality disorders and provided an understanding of the psychodynamics of these patients unsurpassed to this day. He described the splitting operations separating 'good' and 'bad' internalized object relations, the motivated self and object representations that comprise the split-off object relations, the consequent impoverishment of interpersonal relations, and their replacement by a defensive hypertrophy of fantasy life. In fact, in the course of psychoanalytic exploration, the apparent lack of affect display seen in the schizoid personality turns out to reflect severe splitting operations; extreme splitting leads to a fragmentation of affective experience, which 'empties out' interpersonal experience. At the same time, the internalized object relations of the schizoid personality have the split, persecutory, and idealized characteristics typical of the BPD (Kernberg, 1975).

The BPD presents structural and dynamic features similar to those seen in the schizoid personality, but in the borderline personality we see expression of this pathology predominantly in impulsive interactions in the interpersonal field (Akhtar, 1992; Stone, 1994). In contrast to the schizoid personality, where internal object relations are expressed in conscious fantasy life in the setting of social withdrawal, in the BPD the same widely split, internalized object relations are enacted in the interpersonal field. In fact, in the BPD, repetitive, powerfully motivated interpersonal behaviors often replace self-awareness. Episodic, intense, overwhelming affect states ('affect storms') and poor impulse control in the setting of identity diffusion are typical of the BPD, in marked contrast with the schizoid personality where we see apparent lack of affect and good impulse control.

It may well be that the descriptive differences between the schizoid and borderline disorders reflect temperamental differences. In particular, the borderline and schizoid personalities appear to differ across the dimension of extroversion and introversion, one of the important temperamental factors that emerges under different names in various models of classification.

The intensity of affect activation and the lack of affect control seen in the borderline personality, along with the high incidence of affective illness that characterizes this group, suggest the presence of a temperamental factor relating to affect regulation as a predisposing factor for development of BPD. At the same time, it is impressive how the degree to which the integration of negative and positive affect states obtained in the course of psychodynamic treatments brings about a marked toning down and modulation of affect response. The increase of impulse control and affect tolerance in the borderline personality seen as a result of successful treatment illustrates that splitting mechanisms play a central role in the pathology of affects seen in borderline personality (**Figure 1**).

Differential Diagnosis

The most frequent mistake, in our experience, consists in confusing the chronic emotional instability and affect storms of personality-disordered patients with a truly hypomanic or manic behavior. In the case of manic behavior, the differentiation is easier; here the clear loss of reality testing, the presence of hallucinations and/or delusions, or inappropriate social behavior usually leads to intervention by others to control the patient, interventions that are typical enough to confirm loss of reality testing and to warrant the diagnosis

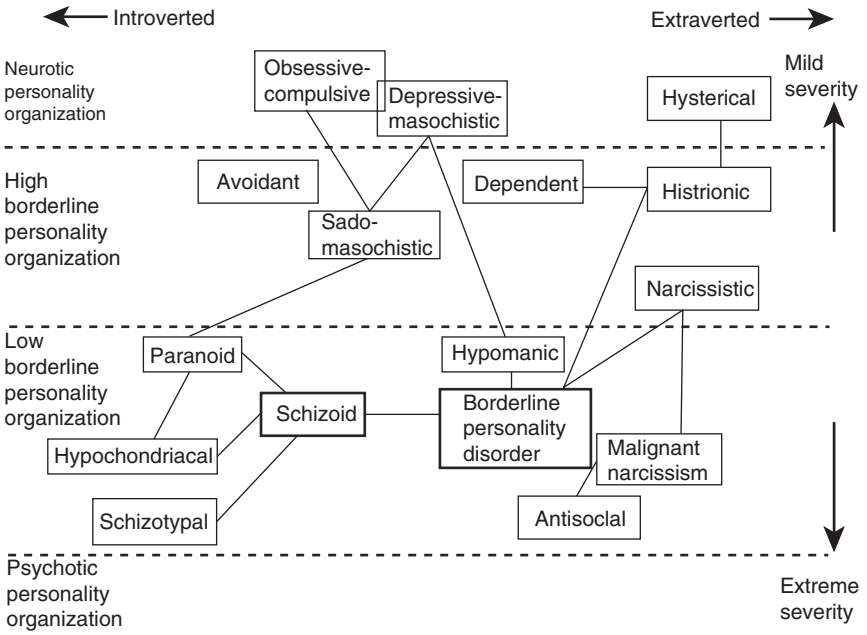

Figure 1 Summarizes the relationships we have outlined that can be seen among the various personality disorders.

of a bipolar disorder. Therefore, the confusion between bipolar illness and BPD is usually reduced to cases of assumed hypomanic behavior used as the basis to diagnose bipolar II in patients.

In about 19% of patients with BPD, however, a comorbidity with bipolar disorder may be present, and the patient shows both severe, chronic affective instability and clear hypomanic episodes (Gunderson et al., 2006). To ascertain the presence or absence of BPD in these cases, it is helpful to evaluate the general nature of the patient's relationships with significant others. Cases of pure bipolar symptomatology do not show severe pathology of object relations during periods of normal functioning, and even chronic bipolar patients, who suffer from both manic episodes and major depressive episodes, maintain the capacity for relationships in depth, stability in their relations with others, and the capacity for assessing themselves and most significant persons in their life appropriately (Stone, 2006).

In contrast, in severe personality disorders with the syndrome of identity diffusion, there is a marked incapacity to assess others in depth, a lack of integration of the concept of self, with severe, chronic discrepancies in the assessment of self and others, and chronic interpersonal conflicts, together with the difficulty of maintaining stable commitments to work and profession, as well as to intimate relationships.

The differential diagnosis between an episode of major depression and a chronic dysthymic reaction in BPD is more difficult, but eminently feasible if enough time is available to clarify the four major areas of symptoms.

First are the psychic symptoms of a depressive spectrum of illness. In major depressions, there is a significant slowdown of the patient's thought processes and the patient's psychomotor behavior, severe depression of mood that varies between profound sadness to the total unavailability of any subjective sense of feeling – a sense of total freezing of all emotional experience in the most severe cases. Typically, thought processes are severely self-demeaning and self-accusatory – rather than focused on accusing and blaming others.

A second area of exploration of the differential diagnosis is the evaluation of the personality structure that predated the beginning of the depressed episode. Patients with severe narcissistic personality disorder, BPD, histrionic personality disorder, and masochistic/depressive personality disorder are prone to severe dysthymic reactions characterized by frequent days with symptoms of depression without reaching the intensity, consistency, and duration of major depressive episodes.

A third area of inquiry facilitating the differential diagnosis between major depression and characterologically based dysthymic reactions involves the following neurovegetative symptoms that point to major depressions: severe insomnia, particularly consistent early awakening hours before the usual waking time; loss of appetite with severe weight loss; consistent loss of sexual desire; possibly impotence in men and suspension of menstrual periods in women; chronic, severe constipation (considering, naturally, that this may be secondary to the use of antidepressive medication); a heightened sensitivity to cold temperature; and, in severe cases, a typical 'masklike' facial expression of severe depression.

A fourth area of diagnostic relevance for the evaluation of depression is the analysis of environmental triggers that may have preceded a depressive episode. Typically, in chronic dysthymic, characterological reactions, environmental conditions may trigger depressive reactions, and these environmental conditions are often remarkably minor, while the patient pays a disproportionate attention to their symbolic value. Major depressions usually do not show such a direct relationship between environmental triggers and depression, although the combination of strong genetic disposition and environmental triggers can occur.

In contrast to this picture in personality disorders, suicide attempts in the context of symptoms of severe depression are typical of major depressive disorders and require a careful diagnostic assessment of the conditions under which suicidal behavior occurred. The types of suicidality generally found in patients with personality disorders that we have just discussed can most often be treated with outpatient psychotherapy.

One other relatively frequent and often difficult differential diagnosis is that between a severe personality disorder, particularly a BPD or a narcissistic personality disorder functioning on an overt borderline level with antisocial features, and an attention deficit/hyperactivity disorder in adolescent or adult patients.

Etiology

BPD, in common with all personality disorders, is the outcome of a complementary series of neurobiological, psychodynamic, and psychosocial etiological features. While, at this time, competing theories of personality development and psychopathology would accentuate different features of this series, there probably is general agreement on the participation of all of them in determining the final outcome of the crystallization of a BPD. These features include

- Genetic disposition to abnormal activation of one or several affect determining neurotransmitter systems, particularly the dopaminergic, serotonergic, and/or noradrenergic systems.
- Genetic disposition (related or not to the alternation of affective neurotransmitter systems), of the structural aspects and/or functions of the prefrontal cortical control system, and reactive hyperactivity of the amygdala to negative stimuli.
- Temperamental hyperreactivity to negative affective stimuli, related to the genetic dispositions referred to.
- Insecure attachment, related to abnormal caregiver–infant interaction in the first few months and years of life.
- Exposure to and/or victimization by physical and/or sexual abuse.
- Severe familial disorganization and unpredictability, blurring of generational boundaries, and early chaotic interactions with close family members.
- Strong predominance of the negative segment of intrapsychic experience, in contrast to a weak segment of internalized satisfactory emotional experiences, leading to persistence and dominance of primitive splitting, and related defensive mechanisms.
- Fixation, as consequence of all above features, of the internalized world of object relations and the self at the pathological organization of identity diffusion, the basic characteristic of severe personality disorders.

Present day research on the pathology of the neurobiological features and the psychopathological constellations of BPD is attempting to assess their relative importance and interrelatedness, their modifications in the course of treatment, and their implication for the strategies of treatment indications and techniques.

Treatment

The treatment of the BPD is essentially psychotherapeutic. There are several specialized psychotherapies available with evidence-based efficacy, outlined below. Psychopharmacological treatment has an auxiliary function and is only indicated in combination with one of these psychotherapeutic approaches.

Psychotherapies for this condition may be classified into psychodynamic and cognitive behavioral treatments.

Psychodynamic psychotherapies	*Cognitive behavioral psychotherapies*
● Supportive psychotherapy (SP) ● Mentalization-based psychotherapy (MBT) ● Transference-focused psychotherapy (TFP)	● Dialectic behavior therapy (DBT) ● Schema-focused psychotherapy (SFT) ● systems training for emotional predictability and problem solving (STEPPS)

Additional treatments include the integrative general psychiatric management, and several other approaches with less empirical research support. Recent comparative meta-analyses of these treatments – the Cochrane collaboration – conclude: "There are indications of beneficial effects for both comprehensive psychotherapies as well as non-comprehensive psychotherapeutic interventions for BPD core pathology and associated general psychopathology. DBT has been studied most intensely, followed by MBT, TFP, SFT, and STEPPS. However, none of the treatments has a very robust evidence base, and there are some concerns regarding the quality of individual studies. Overall, the findings support a substantial role for psychotherapy in the treatment of people with BPD but clearly indicate a need for replicatory studies."

From a practical viewpoint, clinical experience indicates that not all patients respond to all treatments, and selected patient subgroups may respond to some treatments but not to others. In general, when the predominant symptoms are circumscribed to one or several specific problems, such as, suicidal or para-suicidal attempts, specifically focused emotional crises, impulsivity or dissociative reactions, cognitive behavioral approaches, particularly dialectic behavior therapy, would seem the first treatment of choice. When general life problems dominate the clinical picture, such as, failure or breakdown in work or profession, in intimate love relations or a patient's social relationships, a psychodynamic approach may become the first priority. Mentalization-based psychotherapy focuses on patients' learning realistic ways to appreciate their own mental states and the mental states of significant others, and evinces a broad spectrum of patients able to benefit from it. Transference-focused psychotherapy is the only treatment geared to modify the basic pathological personality structure of BPD patients, and may present the optimal indication for patients with severe distortion of all their psychological functions in the areas of work, or profession, love and sex, and social relations. This approach is focused directly on the syndrome of identity diffusion, the basic pathology of the self system and internalized relations with significant others. At this time, research is proceeding regarding the process, mechanisms of change, indications and contraindications of BPD psychotherapies, as well as of the underlying psychological and neurobiological determinants of BPD pathology, and patients' response to treatment.

Regarding psychopharmacologies contributions to the treatment of BPD patients, no one specific medication represents a treatment of choice. Probably about one-third of patients respond to medications over a period of months or years. The main group of medications that have proven helpful to reduce BPD symptoms include antidepressive medications, particularly the Selective Serotonin Re-uptake Inhibitors SSRI group; low doses of atypical antipsychotic medication, particularly risperidone, olanzapine, and aripiprazole; and mood stabilizers, particularly lithium carbonate, valproate, and carbamazepine. SSRI medication has particular indications for affective dysregulation and impulsive behavioral symptoms. Low dose atypical antipsychotic medication also targets these same symptoms, and is particularly indicated for cognitive perceptual symptoms. Mood stabilizers have a relatively secondary indication for some patients with affective dysregulation and impulsivity.

See also: Bipolar (Manic-Depressive) Disorder; Cognitive Behavioral Therapy; Family and Psychiatric Illness: Family Therapy; Personality Disorders; Women and Psychiatry.

Bibliography

Akhtar, S., 1992. Broken Structures. Jason Aronson, Northvale, NJ.

Bender, D.S., Morey, L.C., Skodol, A.E., 2011. Toward a model for assessing level of personality functioning in DSM-5, part I: a review of theory and methods. Journal of Personality Assessment 93, 332–346.

Fairbairn, W., 1954. An Object-Relations Theory of the Personality. Basic Books, New York.

Gunderson, J.G., Weinberg, I., Daversa, M.T., Kueppenbender, K.D., Zanarini, M.C., Shea, M.T., et al., 2006. Descriptive and longitudinal observations on the relationship of borderline personality disorder and bipolar disorder. American Journal of Psychiatry 163 (7), 1173–1179.

Kernberg, O.F., 1975. Borderline Conditions and Pathological Narcissism. Jason Aronson, New York.

Kernberg, O.F., 1980. Internal World and External Reality: Object Relations Theory Applied. Jason Aronson Inc., New York.

Kernberg, O.F., 1989. The narcissistic personality disorder and the differential diagnosis of antisocial behavior. In: Kernberg, O.F. (Ed.), Narcissistic Personality Disorder: Psychiatric Clinics of North America. Saunders, Philadelphia, pp. 553–570.

Kernberg, O.F., 1992. Aggression in Personality Disorder and Perversions. Yale University Press, New Haven.

Stoffers, J.M., Völlm, B.A., Rücker, G., Timmer, A., Huband, N., Lieb, K., 2012. Psychological Therapies for People with Borderline Personality Disorder (Review). The Cochrane Collaboration. John Wiley & Sons.

Stone, M., 1994. Characterologic subtypes of the borderline personality disorder: with a note on prognostic factors. In: Share, I. (Ed.), Borderline Personality Disorder: Psychiatric Clinics of North America. Saunders, Philadelphia, pp. 773–784.

Stone, M.H., 2006. Relationship of borderline personality disorder and bipolar disorder [Editorial]. American Journal of Psychiatry 163 (7), 1126–1128.

Borders, Anthropology of

Hastings Donnan, Queen's University, Belfast, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the frontiers and borders in which anthropologists have been interested, from cultural borders which separate and connect different worlds of meaning and identity to borders that mark out geopolitical space and borders that order social relations and define membership of 'community.' It shows how these cultural, territorial, and social dimensions often overlap and how all three have been integral to an anthropology of borders. The article suggests that the social sciences have increasingly converged in their emphasis on borders as relational and dynamic, no longer confined to political margins but embedded throughout everyday life, a perspective that anthropology has long been adept at exploring.

Introduction

The anthropology of borders has been characterized by three perspectives which, though they overlap, can be identified by their relative emphasis on the cultural, territorial, and social dimensions of borders. These three dimensions were sometimes seen as different types of border, and at other times as aspects of a single border. Indeed, anthropologists are rarely interested in only one type of border or the other, and in many cases the borders they analyze have all three dimensions. Much discussion has thus revolved around how these dimensions intersect, diverge, or coincide. Building on these three perspectives, borders have increasingly been interpreted as a dynamic process, shifting through time and space, and made and remade through everyday practice. Many contemporary border researchers thus no longer take borders for granted and then examine their effects, but instead ask how, why, and when borders are shaped, endure or recede through the agency of those who live and cross there. Answers to these questions not only contribute to wider social and cultural theory but also address current debates about global security and mass mobility.

Anthropologists have a long-standing interest in borders. In the past, this was reflected in a concern with bounding their field of study. Cultural variation was considered a function of geographic and social isolation and early ethnographers advocated analyzing this diversity by treating each social unit as a discrete organic whole. But as the mobility of people dramatically increased (with urbanization, industrialization, etc.), culture contact rather than cultural isolation became the focus of the day. Diffusionists and acculturationists had always been interested in culture contact, but their work was overshadowed and ultimately displaced by the rapid rise and supremacy of functionalism. The belated 'rediscovery' of the value of their insights for contemporary theoretical questions about global cultural interconnectedness indicates a major shift within the discipline: from a concern with what borders encompass to an interest in the borders themselves and the processes through which they emerge.

Frontiers

One way in which anthropologists originally approached the study of culture contact was through the concept of frontier. Etymologically, frontier means 'in front' and came to signify the process of expansion of a political entity (Febvre, 1973). Among its most influential proponents was Turner (1977: p. 3), whose 'frontier hypothesis' – first formulated in 1893 – claimed that "the existence of an area of free land, its continuous recession, and the advance of American settlement westward explain American development." Here, the frontier was the unequal meeting ground between 'civilization and savagery' whose domestication, Turner argued, fashioned the pioneering and independent spirit of the nation. Largely discredited now, Turner's frontier hypothesis was once among the most-debated problems in American historiography, not least because of its attempt to link space and time to the formation of national character. Other early approaches to frontier as a zone of cultural transition and overlap, or that examined the view from the frontier's 'other side' (which Turner had ignored), or that emphasized the magicoreligious aspect of frontier crossing (Van Gennep, 2010; chapter 2), have perhaps been more congenial to anthropology. For example, Bohannan (1967: p. xiii) saw frontier as a 'device' to view cultural and social changes, while Kopytoff (1987) used it to critique Eurocentric views of ethnogenesis in Africa by arguing that ethnic heterogeneity on the continent is better understood as a consequence of the cyclical proliferation of internal frontiers rather than as the result of colonial intrusion. More recently, Roitman's (2005) analysis of financial regulation in the Chad Basin sees this part of Central Africa, which is characterized by the multiple intersections of colonial, postcolonial, and other boundaries, as a discontinuous and shifting frontier zone of fiscal possibility and ambiguity predicated on the illegal border crossings of a 'population flottante.' According to Tsing (2005: pp. 27–29), such 'resource frontiers' are generated in the interstitial space between the legitimate and illegitimate where deregulation means that anything is possible, creating frontiers of 'unmapping' and 'wildness' that have become central to contemporary capitalism and a feature of neoliberal sovereignty (Chalfin, 2010).

Despite these differences and recent refinements, frontier often continues to be used as a synonym for borders, and 'borders, boundaries, and frontiers' are frequently listed together without obvious distinction unless the discussion is explicitly about definitions. Sometimes, it is also used to emphasize a zone of transition characterized by cultural

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12028-8

juxtaposition and intermingling, where differences meet, merge, blur, or coexist. It has been proved particularly effective as a trope in contexts of diaspora and transnational mobility, with their 'fuzzy frontiers' of identity and belonging, a usage anticipated in 1966 by Bohannan (1967: p. xviii) who noted how the frontier is no longer characterized solely by space but is 'all around us,' 'marked by the cultural selectivity of peoples and interest groups.' The frontier 'all around us' is the primary focus of many contemporary anthropologists as outlined below.

Some anthropologists have thus been principally interested in cultural borders (some would say cultural frontiers) in which different worlds of meaning and identity both collide and connect; while others have focused on territorial borders which mark out geopolitical space; and yet others on borders which order social relations and indicate membership of 'community.' All three have been integral elements in the emergence of an anthropology of borders and with the 'cultural turn' in the human sciences generally have been influential in the study of borders in other disciplines.

Cultural Borderlands

Borders have become a metaphor for the cultural flux and indeterminacy of much contemporary life. In anthropology, the border metaphor was introduced by critics of the classic anthropological view of culture as shared, consensual, and discrete. Often with personal experience of cultural contradictions – as members of sexual, ethnic, or other minorities – these critics sought ways to study the differences within and the spaces between cultures, ways that could incorporate the changes, inconsistencies, and incommensurabilities of everyday life. These intercultural spaces are often referred to as 'borderlands,' a usage that evokes the geopolitical and the metaphorical, the literal and the conceptual. In this view, borders and borderlands exist not just at the edges of the nation-state, but anywhere cultures meet (Rosaldo, 1989).

Borderlands are zones of cultural overlap characterized by a mixing of cultural styles. They are liminal spaces, simultaneously dangerous and sites of creative cultural production open to cultural play and experimentation as well as domination and control. The fusion of registers in borderlands may be ultimately empowering for those who inhabit these zones, but it need not always be so, as testified by the lives of many Mexican migrants in the American Southwest.

Not everyone agrees that it is helpful to extend the use of 'border' and 'borderland' in this way. When cultural encounters share some of the specific sociopolitical processes characteristic of borders between states, such as unequal access to official forms of power, the borderland metaphor may be appropriate. But where there is no strong analytical connection to state border processes, understanding of the border can become reductive and delocalized (Heyman, 1994).

Cultural differences and the juxtaposition of different worlds of meaning come to be emphasized at the expense of stressing inequalities of power. Heyman and others feel uneasy about this differential emphasis, reflecting wider tensions within the discipline over the relative prominence to be given to culture and power. The force of the image, they feel, takes over from the analysis.

In a sense, of course, borders are always metaphors, since they are arbitrary constructions based on cultural convention. Metaphors, moreover, are part of the 'discursive materiality of power relations' and in this respect no less concrete in their consequences than state borders (Brah, 1996: p. 198) as Vila (2000) has compellingly demonstrated in his narrative ethnographies of the Mexico–US border. Metaphorical and state borders are not, therefore, as far apart as some studies imply. In fact, the anthropology of borders has benefited from the productive interaction of approaches that emphasize one or the other. The following sections outline other anthropological analyses of borders from which the border-as-metaphor draws its force and resonance.

Territorial and Political Borders

State borders entail a mapping out in geographic space and recognition in international law. They mark the limits of sovereignty and of state control over citizens and subjects, limits that may be upheld by force or by the threat of force. They are often highly visible physically, which has led some scholars to refer them as 'real' borders in contrast to borders without territorial counterparts. This can be misleading, and the materiality of state borders should not blind us to their cultural and symbolic dimensions (just as we should not assume that cultural and symbolic borders automatically have no materiality).

Apart from the Mexico–US border, state borders have not until recently been subject to systematic comparative scrutiny by anthropologists. Where they did appear in ethnographic accounts, they often figured only as a backdrop to some other line of inquiry. Prior to the 1970s, even studies of the Mexico–US border rarely included the border as a variable in the analysis. As a result, the anthropology of state borders was slow to develop.

A fledgling 'school' of anthropological border studies briefly emerged at the University of Manchester, UK, in the 1960s, but subsequently fizzled out. During this period, several books appeared specifically with borders in their titles; each influenced by Max Gluckman who was Departmental Chair at that time (e.g., Cohen, 1965; Harris, 1972). Taken together, they began something that took another 20 years to bear fruit. But though these books did not problematize, the state borders near which their authors did research, and though they did not cross-refer to one another, they nevertheless identified some common themes in the political anthropology of borders which laid a solid foundation for contemporary anthropological analyses of nations and states. They pointed out, for instance, how proximity to a state border can intensify local conflicts, which can easily escalate into conflicts over nationality, a threat which only cross-cutting linkages (of kinship, common residence, and shared social and economic interests) help to keep in check. Understanding these borders thus requires local ethnographic knowledge, and not just the knowledge of state-level institutions and international relations, a fine-tuned sensitivity that is alert to cultural meanings and practice that remains central to the anthropology of borders today.

The value of localized studies for understanding how cultural landscapes are superimposed across social and political divides was first systematically developed by Cole and Wolf

(1974) whose field site in the Italian Tyrol was specifically chosen because its successive historical partitioning allowed them to explore the transformation of local political loyalties in relation to nation-building. What particularly interested Cole and Wolf about the South Tyrol was the durability of a cultural frontier long after the political borders of state and empire had shifted. National boundaries had clearly survived the demise of state borders, and remained important in everyday life. Despite their similarities, the two villages studied had followed a different political and cultural course since World War I. Villagers minimized these ideological and cultural differences in public encounters, but in private quickly resorted to ethnic stereotypes to explain the actions of the others. Cole and Wolf thus reiterate Barth's (1969) observation that ethnic boundaries may be maintained despite relations across them. Their major contribution, however, was in demonstrating the need to combine the study of local and extralocal influences to understand and explain this process.

Here, then, is an example of where ethnic boundaries result from, and evolve with, the rise and demise of state borders. Each can only be understood by reference to the other. In this respect, Cole and Wolf combined symbolic border studies (see Section Social and Symbolic Borders) with a political economy perspective that situates ethnographic knowledge of local boundaries within wider historical and political processes, a novel combination that marked an important transition in the anthropology of borders and heralded the beginning of a new form of inquiry. Subsequent anthropological research on national and international borders was to draw explicitly or implicitly upon this groundbreaking work (see Donnan and Wilson, 1999, for example). Anthropologists began to use their field research at state borders as a means of widening perspectives in political anthropology to encompass the formal and informal ties between local communities and the larger polities of which they are a part. Some studied border areas as a way of examining how proximity to an international border could influence local culture or could create the conditions that shape new rural and urban communities. Others focused on the voluntary and involuntary movement of people across borders as traders, migrants, and refugees. Yet others concentrated on the symbols and meanings which encode border life. The Mexico–US border, where the issues of underdevelopment, transnationalism, and the globalization of power and capital, among other aspects of culture, increasingly concern the growing number of historically informed and wide-ranging ethnographic accounts, offers excellent examples of all of these trends (see Alvarez, 1995, 2012).

Regardless of theoretical orientation or locale, however, most of these border studies have focused on how social relations, defined in part by the state, transcend the territorial limits of the state, and, in so doing, transform the structure of the state at home and in its relations with its neighbors.

Anthropological interest in how local border developments can have an impact on national centers of power and hegemony was partly stimulated by historical analyses of localities and the construction of national identities (e.g., Sahlins, 1989). Cole and Wolf's insistence on the need to view the anthropology of borders as historical anthropology is recalled. As the South Tyrol case so clearly shows, borders are spatial and temporal records of relationships between local communities and between states. Ethnographic explorations of the intersection of symbolic and state borders have salience beyond anthropology because of what they may reveal about the history of cultural practices and about the role of border cultures and communities in policy-making and diplomacy. This kind of research highlights the growing importance of a border perspective in political anthropology, a perspective in which the dialectical relations between border areas and their nations and states take precedence over local culture viewed with the state as a backdrop.

The next section outlines the work of some of the scholars whose innovative theorizing of the symbolic dimensions of social boundary making has provided a cornerstone to the anthropology of state borders worldwide.

Social and Symbolic Borders

One of the most influential theorists in the anthropology of borders has been Barth (1969), who argues that people may traverse group boundaries and maintain regular relations across them without affecting the durability and stability of the boundaries themselves. Barth's focus is on ethnic groups, which he argues are social constructions, the result of individuals strategically manipulating their cultural identity according to context, rather than the predetermined outcome of some objective list of culture traits. Membership of ethnic groups should thus be understood as based on self-ascription and ascription by others: individuals claim membership in a particular group and others confirm or question that claim. Attention is thereby drawn to the boundary between groups: to how and why people distinguish themselves from others, and to the rules of behavior that sustain such boundaries in the face of relations that transcend them. Barth's boundaries are above all 'social' boundaries: they may or may not have a territorial dimension, and cultural differences become significant only in so far as they are useful for organizing social relations.

Barth's work both revolutionized the study of ethnic groups and stimulated further research on borders. Some of this research demonstrates that borders mark affective and identificatory as well as structural and organizational disjunctures: those on the inside not only inhabit a separate social system from those on the outside but they identify themselves differently. Any social border is thus a consequence of the various possible relationships between these two dimensions on each side of the border as well as across it (Wallman, 1978). This highlights the relational nature of social boundaries, drawing attention to what they mean to people and to how they are marked. These are issues developed by Cohen (1985). Echoing Barth, Cohen suggests that borders are constructed by people in their interactions with others, from whom they wish symbolically to distinguish themselves. However, we cannot predict what the distinguishing features of these symbolic borders will be, or exactly where the lines will be drawn. Moreover, they may mean different things to different individuals, both to those within a border as well as those outside it. In fact, borders recognized by some may be invisible to others. The anthropologist's task is to uncover

these borders and their meanings in order to grasp what inclusion or exclusion entails.

Cohen has been criticized for devoting more attention to one side of the border than the other. For him, the significance of symbolic borders is that they allow people to retain a sense of distinctiveness in a world where the diversity of local communities is increasingly threatened by structural incorporation into wider political formations and global cultural styles. This inevitably leads him to emphasize what goes on within a border, since this is how he can identify the symbolic beliefs and practices that people deploy to mark and celebrate their sense of difference. When he does consider what lies beyond the border, it is often only to show how external events can be manipulated to symbolic advantage at the local level.

Similar criticisms have been leveled at Barth. As we have seen, Barth also emphasizes that boundary making involves both self-ascription and ascription by others. But he too tends to focus on only one side of the border, emphasizing internal identification rather than external categorization and the shaping influence of wider structures. This minimizes the relative power relations upon which the external ascriptive process depends, and underplays the fact that some people can impose their categorizations on other people. Relationships of domination and subordination can only be restored to theoretical centrality in border studies by recognizing those on one side of a border may be better able than those on the other side to determine where and how the line is drawn (cf. Jenkins, 1997). Indeed, we saw earlier that some anthropologists who focus on state borders have attempted to build this dimension into their analyses.

Conclusion and Convergence

Three reasonably distinct but mutually enriching streams in the anthropology of borders have been identified. All three show that anthropologists have brought to the study of borders an appreciation of what borders mean to those whose lives they frame. It is on this emphasis on how borders are constructed, negotiated, and viewed from 'below' that the three perspectives outlined here converge and on which the value and distinctiveness of an anthropology of borders largely depends. It is this that anthropology brings to the eclectic and interdisciplinary field that border studies have become, one increasingly characterized by a convergence of question, concept, and comparative approach. Anthropology's longstanding sensitivity to comparative worlds of identity and meaning, and to symbolic boundaries and cultural borderlands, arguably anticipated some of the questions, now engaging disciplines once principally preoccupied with 'realist' perspectives that emphasize borders of space, territory, and geopolitics. This has opened up these disciplines to new conceptual directions.

New perspectives have also opened up within the anthropology of borders, as well as new field sites in parts of the world previously underrepresented in borders research. Border ethnographies have proliferated, with case studies of colonial, precolonial, indigenous and pre- and postsocialist borders in eastern Europe, Africa, South and Southeast Asia, Central Asia, and North and South America bringing a welcome set of comparators to the borders of the European Union, Middle East, and Mexico–US that had long been paradigmatic in border studies. Thematic focus is now similarly diverse (for overview, see Wilson and Donnan, 2012) and while post-9/11 research is dominated by the relationship between borders, security, surveillance, and mobility (e.g., Chalfin, 2012; Cunningham, 2012; Cunningham and Heyman, 2004; Kalir and Sur, 2012; Smart and Smart, 2012), other novel ethnographic fields, such as health, bio- and eco-borders, maritime boundaries, and fiscal frontiers have also emerged.

But perhaps above all it has been the epistemological shifts, rather than the expansion of geographical and thematic focus, that characterize the last decade in the anthropology of borders. These have entailed an increasing emphasis on relationality as the critical feature of borders, echoing Wallman's (1978) approach and developing related strands within the other perspectives outlined above which emphasize the ubiquity and placelessness of borders rather than their locational fixity in space. The relative prominence attributed to territory, culture, and the social as elemental aspects of 'border' has altered, and the relationship between states and borders become loosened or detached altogether in some analyses even as the significance of the state for social boundaries is forcefully argued by others. Borders are now more likely to be seen as 'process' than as 'product,' the outcome of a fluctuating tension between dynamic relationships of inclusion and exclusion. This shift has been variously conceptualized within and beyond the anthropology of borders: as borderscape (Rajaram and Grundy-Warr, 2007), as bordering or b/ordering (Van Houtum et al., 2005) and as tidemark and borderness (Green, 2012). What these reconceptualizations of 'border' share in common is their insistence that 'the border' is not (or is not just) a particular physical space. Rather, the border is historically contingent, always in flux, and the momentary outcome of an "inherent contestability of the meaning of the border between belonging and non-belonging" (Rajaram and Grundy-Warr, 2007: p. xxviii). Border is thus better seen as 'bordering,' as verb rather than noun, whose traces should be understood as an ever-changing tidemark rather than as a fixed and predetermined line. Conceptualized in this way, 'border' is increasingly studied not as a once-and-for-all but as continuously emergent through the performative making and remaking of difference in everyday life through text, discourse, practice, and event. No longer just the limits of the political (Balibar, 2002: p. 92), borders are now 'all around us,' embedded in the practice of the ordinary life that anthropology has been so adept at exploring.

This current openness in border studies to transdisciplinary dialogue and convergence of perspective and approach suggests a new willingness to engage global comparison and the work of scholars in other fields, thereby rethinking disciplinary boundaries alongside our reconceptualizing of borders themselves.

See also: Ethnicity, Sociology of; Ethnicity: Anthropological Aspects; Ethnocentrism; Frontiers in History; Groups, Sociology of; Liminality; Nationalism: General; State: Anthropological Aspects; Symbolic Boundaries; Symbolism in Anthropology.

Bibliography

Alvarez, R.R., 1995. The Mexican–US border: the making of an anthropology of borderlands. Annual Review of Anthropology 24, 447–470.

Alvarez, R.R., 2012. Reconceptualizing the space of the Mexico–US borderline. In: Wilson, T.M., Donnan, H. (Eds.), A Companion to Border Studies. Blackwell, Oxford.

Balibar, E., 2002. Politics and the Other Scene. Verso, London.

Barth, F., 1969. Introduction. In: Barth, F. (Ed.), Ethnic Groups and Boundaries: The Social Organization of Culture Difference. George Allen and Unwin, London.

Bohannan, P., 1967. Introduction. In: Bohannan, P., Plog, F. (Eds.), Beyond the Frontier: Social Process and Cultural Change. The Natural History Press, New York.

Brah, A., 1996. Cartographies of Diaspora: Contesting Identities. Routledge, London.

Chalfin, B., 2010. Neoliberal Frontiers: An Ethnography of Sovereignty in West Africa. Chicago University Press, Chicago, IL.

Chalfin, B., 2012. Border security as late-capitalist 'fix'. In: Wilson, T.M., Donnan, H. (Eds.), A Companion to Border Studies. Blackwell, Oxford.

Cohen, A., 1965. Arab Border-Villages in Israel: A Study of Continuity and Change in Social Organization. Manchester University Press, Manchester, UK.

Cohen, A.P., 1985. The Symbolic Construction of Community. Tavistock, London.

Cole, J.W., Wolf, E., 1974. The Hidden Frontier: Ecology and Ethnicity in an Alpine Valley. Academic Press, New York.

Cunningham, H., 2012. Permeabilities, ecology and geopolitical boundaries. In: Wilson, T.M., Donnan, H. (Eds.), A Companion to Border Studies. Blackwell, Oxford.

Cunningham, H., Heyman, J.McC. (Eds.), 2004. Introduction: Mobilities and Enclosures at Borders. Special Issue on Borders. Identities: Global Studies in Culture and Power 11 (3), 289–302.

Donnan, H., Wilson, T.M., 1999. Borders: Frontiers of Identity, Nation and State. Berg, Oxford.

Febvre, L., 1973 [1928]. Frontière: the word and the concept. In: Burke, P. (Ed.), A New Kind of History: From the Writings of Febvre. Routledge and Kegan Paul, London.

Green, S., 2012. A sense of border. In: Wilson, T.M., Donnan, H. (Eds.), A Companion to Border Studies. Blackwell, Oxford.

Harris, R., 1972. Prejudice and Tolerance in Ulster: A Study of Neighbours and 'strangers' in a Border Community. Manchester University Press, Manchester, UK.

Heyman, J., 1994. The Mexico–United States border in anthropology: a critique and reformulation. Journal of Political Ecology 1, 43–65.

Jenkins, R., 1997. Rethinking Ethnicity: Arguments and Explorations. Sage, London.

Kalir, B., Sur, M. (Eds.), 2012. Transnational Flows and Permissive Polities: Ethnographies of Human Mobilities in Asia. Amsterdam University Press, Amsterdam.

Kopytoff, I., 1987. The internal African frontier: the making of African political culture. In: Kopytoff, I. (Ed.), The African Frontier. Indiana University Press, Bloomington, IN.

Rajaram, P.K., Grundy-Warr, C., 2007. Introduction. In: Rajaram, P.K., Grundy-Warr, C. (Eds.), Borderscapes: Hidden Geographies and Politics at Territory's Edge. University of Minnesota Press, Minneapolis, MN.

Roitman, J., 2005. Fiscal Disobedience: An Anthropology of Economic Regulation in Central Africa. Princeton University Press, Princeton, NJ.

Rosaldo, R., 1989. Culture and Truth: The Remaking of Social Analysis. Beacon Press, Boston, MA.

Sahlins, P., 1989. Boundaries: The Making of France and Spain in the Pyrenees. University of California Press, Berkeley, CA.

Smart, A., Smart, J., 2012. Biosecurity, quarantine and life across the border. In: Wilson, T.M., Donnan, H. (Eds.), A Companion to Border Studies. Blackwell, Oxford.

Tsing, A.L., 2005. Friction: An Ethnography of Global Connection. Princeton University Press, Princeton, NJ.

Turner, F.J., 1977 [1920]. The significance of the frontier in American history. In: Turner, F.J. (Ed.), The Frontier in American History. The Franklin Library, Franklin Center, PA.

Van Gennep, A., 2010 [1960]. The Rites of Passage. Routledge, London.

Van Houtum, H., Kramsch, O., Zierhofer, W. (Eds.), 2005. B/ordering Space. Ashgate, Aldershot, UK.

Vila, P., 2000. Reinforcing Borders: Social Categories, Metaphors, and Narrative Identities on the US–Mexico Frontier. University of Texas Press, Austin, TX.

Wallman, S., 1978. The boundaries of 'race': processes of ethnicity in England. Man 13, 200–217.

Wilson, T.M., Donnan, H. (Eds.), 2012. A Companion to Border Studies. Blackwell, Oxford.

Boserup, Ester (1910–99)

Gale Summerfield, University of Illinois, Champaign, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ester Boserup (1910–99) was a strong proponent of an interdisciplinary approach to problems of international development. She made significant contributions through her studies of agriculture, population growth, informal work, rural–urban migration, polygamy, property rights, education, and urban employment. Boserup's book, *Woman's Role in Economic Development*, generally is acknowledged as initiating the formal field of study: Women in International Development. By demonstrating that women are often marginalized in the course of development rather than gaining from the process, her work helped shift the focus of development research from growth of income toward people's well-being and agency.

Ester Boserup (1910–99) summed up her research as focusing "… on the interplay of economic and noneconomic factors in the process of social change, both today and in the past, viewing human societies as dynamic relationships between natural, economic, cultural, and political structures, instead of trying to explain them within the framework of one or a few disciplines" (1999). Boserup's book, *Woman's Role in Economic Development* (1970), generally is acknowledged as initiating the formal field of study: Women in International Development (WID). While the WID field is interdisciplinary and built on the research and practice of numerous women and men, Boserup's contribution is substantial and helped shift research and policies to address women's work more thoroughly.

Boserup was born Ester Borgesen in Copenhagen, Denmark, 18 May 1910. She studied economics at the University of Copenhagen during the Great Depression (1929–35), which influenced her search for theoretical explanations of economic processes that went beyond the equilibrium approach. Her major field of study stressed theoretical economics but also allowed her to take courses in sociology and agricultural policy. While at the university, she married Mogens Boserup. After graduating, she worked for 12 years for the Danish government writing on regulations and agricultural policy in Europe. Her three children were born in this phase. From 1947 to 1957, Boserup worked in Geneva for the United Nations' Economic Commission for Europe and began collaborative work with the UN Food and Agriculture Organization (FAO). During this period, her research focus expanded to developing countries.

In 1957, Boserup and her husband left the UN and went to India for the *Asian Drama* research project with Gunnar Myrdal. Ester Boserup drafted chapters on agriculture, Gandhism, and development planning. Her travels in South and Southeast Asia exposed her to different types of farming and land tenure systems. She began to formulate her own concepts of the process of development that countered much of the received wisdom of the time; for example, she observed that labor was still a constraint in traditional agricultural systems so that the marginal product of labor could not be zero (or close to it) as commonly assumed and that rural–urban migration was often caused more by pull factors, such as attractive money incomes, than by push factors. In 1960, Boserup and her husband resigned from the *Asian Drama* project because their ideas about the development process had grown too far from Myrdal's.

For the rest of her life, Boserup worked as a consultant for organizations such as the FAO Trade Division in Rome and the Center for Industrial Development (later called UN Industrial Development Organization (UNIDO)). This was the period when she published her most influential works: *The Conditions of Agricultural Growth* (1965/1993) and *Woman's Role in Economic Development* (1970). She played a central role in the UN World Population Conference in Bucharest in 1974 and the first UN World Conference on Women in Mexico City in 1975. In the 1990s, she stopped participating in conferences because of her health, but she continued to be active in research and writing, stressing dynamic, interdisciplinary approaches to development issues, until her death on 24 September 1999 in Ascona, Switzerland.

Boserup's Theoretical Contributions

Boserup's works reflect her background in classical and neoclassical economics as well as interdisciplinary studies. In 1936, she began building her analysis with a comparison of Marx's theory of underconsumption and Keynes's theory of propensity to consume (1999). Boserup established her reputation as a researcher with *The Conditions of Agricultural Growth* in the mid-1960s. Her main thesis in this work is that technological change and property rights are endogenous. The book was published at a time when growth theory was dominated by neoclassical models that treated technology as exogenous. Boserup hypothesized that increasing population density led to transformation of the environment and introduction of new technology, in particular in a change of agricultural systems from shifting (slash and burn) techniques to use of the plow and irrigation. This view countered the Malthusian and neo-Malthusian positions on population growth that took the carrying capacity of the Earth and technology as given and predicted that continued population growth would lead to soaring prices for grain, wars, disease, and at best a return to subsistence survival for most people.

While Boserup's analysis was a bold break with much of the received wisdom of the time, it still contained elements of the linearity present in the works of writers like Marx and Rostow and implicitly posited a view of development as progress from family production to modernity characterized by labor specialization

and monetization of the economy. Statements that Boserup made at times in her writing, however, demonstrate that her views were actually more complex; she noted, for example, that reversals can occur such as the fall of the Roman Empire (1999). The complexity showed up more consistently in her interdisciplinary, dynamic approach rather than in addressing issues of nonlinearity and variety in development processes. Although there are differences in current views about what motivates technological change, the conceptualization of technology as endogenous has been a significant contribution to development theory. It is the basis of much of the new growth theory of the 1980s and 1990s, but new growth theorists are more likely to explain the technological adaptation in terms of human capital and positive externalities rather than population pressure.

Her book on agricultural growth gave Boserup the credentials in the academic community necessary to stir interest in her next major work, *Woman's Role in Economic Development* (1970). The timing was also good since activism among women was growing and interest in changing public policy was increasing among men and women. Boserup brought gender into development analysis as an integral factor that has remained central to much subsequent development work (Jaquette and Staudt, 2006; Tinker, 2001). She pointed out that both colonialism and modern development programs had frequently marginalized women from rights that they previously had.

She brought in her ideas from the agricultural growth book to identify shifting agriculture with low population density, tools such as the hoe, common property, and women taking an active role in subsistence farming. This system, common in Africa, she identified as a female farming system. She contrasted this with the system often found in more densely populated Asia of plowed fields, irrigated crops, and private property, which she termed a male farming system. She noted that the gender roles in economic activity were distinct in all the societies she studied and although considered 'natural' by members of each group, they varied across cultures – illustrating that the gender roles were constructed socially rather than being biological. Despite her awareness of these variations in occupation and time use by gender, Boserup focused on constructing a general view and downplayed the variation among agricultural systems. She stated that the agricultural system would tend to progress from the female farming, shifting system to the male farming, plowed system as population density increased. Continuing growth of population would then lead to more intensive farming of irrigated land by both women and men (a mixed system) (1970). Much of the research that had focused on women and gender at that time had been done by anthropologists who frequently used in-depth case studies as a methodological approach. As a trained economist, Boserup used the approach of collecting large amounts of data on the roles of women and men from many countries (mainly in sub-Saharan Africa and South/Southeast Asia) and making inferences – thus, the 12 figures and 64 tables in her book (1970, 1999).

Continuing Influence of Boserup's Pioneering Effort

Woman's Role in Economic Development has been cited often and has generated debate and controversy that has deepened our understanding of human development. The publication of the book in 1970 marks the beginning of the multidisciplinary WID field that brought together researchers, practitioners, and policy makers/advocates initially to work to integrate women into the development planning process (see Tinker, 1990). Boserup's work clearly demonstrated that she recognized that women had long been actively participating in the actual development process (Jaquette, 2001).

The WID field has evolved gradually. In the late 1980s, advocates concerned with global gender issues put forward a new gender and development (GAD) model to focus more on relative gender analysis of how women and men fare in the costs and opportunities associated with economic transformation policies and the process of globalization and how differences in age, class, religion, and location influence the results (Jaquette and Staudt, 2006). The distinction between WID and GAD has blurred in practice, and the field has centered more on issues of globalization and crisis as well as development policy. Significant critiques of *Woman's Role* were published by Huntington (1975) and Beneria and Sen (1981). Huntington challenged Boserup's classification of female and male farming systems and suggested that women in Africa had been in subordinate relations with men, more like peasants and aristocrats, long before the colonial period. Beneria and Sen acknowledged Boserup's contributions but critiqued the lack of clear theoretical framework, the dominance of the capitalist model of development in her work, and the lack of feminist analysis of women's subordination and their reproductive roles (1981). Boserup's publications also motivated the *Persistent Inequalities* volume (Tinker, 1990) that remains a leading contribution to the WID/GAD field. Debate continues both over the rhetoric of what to call this evolving field and appropriate theories and policies that recognize diversity and give voice to the people affected throughout the world.

Boserup's pioneering publications have contributed significantly to our understanding of how economic policies and processes affect people's well-being and agency – all people, not just the men of the society. While best known for her work on gender and agricultural systems, Boserup also wrote about the 'bazaar and service occupations' (1970) now called the informal sector that has gained much attention with economic restructuring policies since the 1980s (Tinker, 2001). In addition, she wrote on rural–urban migration, polygamy, property rights, education, and urban employment. By demonstrating that women are often marginalized in the course of development rather than gaining from the process, her work helped shift the focus of development research from growth of income toward people's well-being and agency. Boserup argued that better educational opportunities could help reduce the inequality resulting from this marginalization (Turner and Fischer-Kowalski, 2010). Rodgers (2010) notes that Boserup's work has been useful in policy efforts to reduce the gender gap in education, especially in Asia and South America.

Boserup's work is global and interdisciplinary in perspective, addressing the links between culture, environment, technology, population, employment, and family. She framed the argument for women's rights in terms of development that was more acceptable to the mainstream at the time (1999). She brought efficiency as well as equity into the argument for women's rights and a just society, but she did not depict women as victims

(Jaquette, 1990). Boserup used her publications to make women's work visible and stressed the contributions of women as well as men to economic growth.

See also: Agriculture, Economics of; Gender and Feminist Studies in Geography; Gender and Women's Studies, Applied Research On; Gender, Economics of; Household Economics.

Bibliography

Beneria, L., Sen, G., 1981. Accumulation, reproduction, and women's role in economic development: Boserup revisited. Signs 7 (2), 279–298.

Boserup, E., 1965/1993. The Conditions of Agricultural Growth: The Economics of Agriculture under Population Pressure. G. Allen and Unwin, London.

Boserup, E., 1970. Woman's Role in Economic Development. St. Martin's Press, New York.

Boserup, E., 1990. Economic change and the roles of women. In: Tinker, I. (Ed.), Persistent Inequalities: Women and World Development. Oxford University Press, New York.

Boserup, E., 1996. Development theory: an analytical framework and selected applications. Population and Development Review 22 (3), 505–515.

Boserup, E., 1999. My Professional Life and Publications 1929–1998. University of Copenhagen. Museum Tusculanum Press, Denmark.

Huntington, S., 1975. Issues in woman's role in economic development: critique and alternatives. Journal of Marriage and the Family 37 (4), 1001–1011.

Jaquette, J., 1990. Gender and justice in economic development. In: Tinker, I. (Ed.), Persistent Inequalities: Women and World Development. Oxford University Press, New York.

Jaquette, J., 2001. From academia to the WID office: crossing the line. In: Tinker, I., Fraser, A. (Eds.), Women Affecting International Development: The Personal and Political. Feminist Press, New York.

Jaquette, J., Staudt, K., 2006. Women, gender, and development. In: Jaquette, J., Summerfield, G. (Eds.), Women and Gender Equity in Development Theory and Practice: Institutions, Resources and Mobilization. Duke University Press, Durham, NC.

Rodgers, Y., 2010. Review of woman's role in economic development, by Ester Boserup. International Journal of Social Economics 37 (4), 339–340.

Tinker, I. (Ed.), 1990. Persistent Inequalities: Women and World Development. Oxford University Press, New York.

Tinker, I., 2001. Ester Boserup – a Tribute to Ester Boserup: Utilizing Interdisciplinarity to Analyze Global Socio-economic Change. Available: http://irenetinker.com/publications-and-presentations/ester-boserup (accessed 05.04.12.).

Turner II, B., Fischer-Kowalski, M., 2010. Ester Boserup: An Interdisciplinary Visionary Relevant for Sustainability. Available: http://www.pnas.org/content/107/51/21963.full (accessed 05.04.12.).

Boundaries and New Organization Forms

Walter W Powell and Birthe Soppe, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W.W. Powell, volume 2, pp. 1295–1298, © 2001, Elsevier Ltd.

Abstract

Organizational boundaries, the demarcation between an organizational entity and its external environment, have undergone considerable remaking in recent years. Long thought to be relatively stable, boundaries were once redrawn only to gain more leverage over external factors or increase scale. But boundaries have become much more fluid and permeable as organizations experimented with a wide variety of new organizational forms to access critical knowledge, skills, and resources. Particularly in knowledge-intensive contexts, there is ample evidence of growth in novel collaborative forms of organizing. In this article, several arguments that offer insight into the creation of novel forms are highlighted, the limits of an atomistic view of organizations are emphasized, and avenues for future research are offered.

Introduction

New systems of organizing do not arrive on the scene readymade and announce their availability. The historian David Hounshell (1984) illustrated how the mass production model emerged piecemeal in the United States in the later half of the nineteenth century, beginning with the use of interchangeable parts in rifles. Subsequently, the manufacture of sewing machines and then bicycles, and later, meatpacking and beer brewing all played a critical role in the eventual development of the assembly line by Henry Ford. Similarly, new forms of organizing in the late twentieth century emerged incrementally, in fits and starts, and became visible in a variety of guises. For much of the second half of the twentieth century, arguments from economics and sociology drew on considerations of efficiency and power to stress that organizational boundaries sharply demarcated an organization from its external environment. Indeed, some organizations operated like medieval castles, walled off and protected from external influences. When boundaries were redrawn, considerations of scale and scope loomed large as decisions to shift boundaries were based on weighing the costs of internalizing a transaction versus conducting it in the market (Williamson, 1985). In another line of work, emphasizing the ability of an organization to maintain autonomy and control over its environment, boundaries were recast to either reduce uncertainty or dependency on external parties (Pfeffer and Salancik, 1978) or to reshape the production chain in order to enhance strategic position within an industry. These lines of research were subsequently challenged by a wide proliferation of alternative governance arrangements and new types of organizational forms.

Today, organizations worry more about their adaptability to a changing landscape than their adaptive fit to a specific environment (Smith-Doerr and Powell, 2005; Stark, 2009). These new forms enhance absorptive capacity (Cohen and Levinthal, 1990), that is, the capability to access skills and knowledge located outside the organization, as well as the ability to reconfigure organizational practices and structures to respond to a rapidly changing environment in which internal activities are increasingly interdependent with external parties (Carlile, 2004; Meyer et al., 2005).

The unraveling of the old system of bureaucratic employment in the large firm is now widely recognized, but developing a set of metrics to assess the spread of new forms of work and organizing is challenging. The developments underway are multilevel in nature, involving a transformation in the ordering of work at the point of production, increased porousness of the boundaries of many organizations, and a remaking of relations with competitors and collaborators (Fleming et al., 2007; O'Mahony and Bechky, 2008). These developments are discontinuous because there is no clear stopping point in the process and no road back to the previous system. Performance has replaced seniority as the condition of employment. Learning and speed have replaced quantity as the criteria for evaluating organizations. These shifts bring new actors and new business models to the fore, and push aside incumbents.

New Forms of Workplace Organization

Jobs emerged in the late nineteenth and early twentieth centuries as a way to package work in settings where the same task was done repeatedly. But work has changed into short-term projects often organized around a team or work group charged with responsibility (Hargadon and Bechky, 2006). Consequently, the future organization of work is likely to be much less frequently honeycombed into a pattern of highly specified jobs.

Sabel (1994) terms this process of joint exploration 'learning by monitoring.' The activities of work teams are coordinated by a process of iterated goal setting. General projects, such as the design of a new car, are initially determined by study of best practices and prospects for competing alternatives. Then broad plans are in turn decomposed into tasks for work groups, located inside, outside of, and across organizational boundaries. The goals are subsequently modified as work teams gain experience in executing the required tasks. Through these revisions, changes in the parts lead to modifications in the conception of the whole, and vice versa. The same procedure of monitoring decentralized learning, moreover, allows each party to

observe the performance of the other collaborators closely enough to determine whether continued reliance on them, and dedication of resources to the joint projects, are warranted.

This form of production integrates conception and execution, with design and production running on parallel tracks. With concurrent design and development, participants constantly evaluate each other's work. If project groups decide who supplies their inputs, they need not choose the traditional internal unit, but instead turn to outside suppliers if they provide better value. This reconceptualization of work reduces and exposes fixed costs, making the expenses of all units dependent on their contribution, and fuses the knowledge housed in different parts of the organization and the environment.

These new arrangements are deeply corrosive of the old system of sequential steps, linear design, and vertical integration that provided the worker and manager alike with security. One consequence of this proliferation of collaborative work is the democratization of the product development process at the expense of internal lines of authority (von Hippel, 2005). Another consequence is that important tasks no longer need to be performed inside the organization. This change remakes not only the organization of work but also the boundaries of organizations.

Flattening of Hierarchies and the Spread of Networks

In tandem with a conception of work as organized around projects that transforms firms internally, the growing involvement of organizations in an intricate latticework of collaborations with 'outsiders' also blurs the boundaries of the firm, making it difficult to know where the firm ends and where the market or another firm begins. The former redraws internal lines of authority, while the latter spreads the core activities of the firm across a wider array of participants, with an attendant loss of centralized control. As one of the first astute observers of these developments, Rosenbloom and Spencer (1996) suggested that industrial competition was resembling less a horse race and more a rugby match in which players frequently change uniforms.

Various forms of organizational collaborations have grown rapidly. So intensive are these partnering efforts that the interorganizational network is increasingly the relevant unit of production and the locus of innovation (Powell et al., 1996; Carlile, 2004). Rivalry among single firms has shifted toward networks of allied organizations competing against one another (Gomes-Casseres, 2006). Such alliances do not dampen rivalry, but instead shift the playing field to both competition and collaboration among rival networks with fluid membership.

The growth of alliances entails complex collaborations with suppliers, end users, distributors, and even competitors, ranging from informal ties to formally defined partnerships. This upsurge in collaboration is particularly striking in knowledge-intensive industries (Heckscher and Adler, 2006). Networks have become an essential component in such contexts because they channel information and resources, and afford rapid organizational learning.

Collaboration in Research and Development

In research and development (R&D)-intensive settings, forms of organizing for technological innovation have undergone a significant transformation. Among the many challenges of the early twenty-first century are much shorter product development cycles, more sophisticated technological infrastructures, geographically distributed knowledge, and pressing global issues such as climate change. These challenges have provoked deep changes in the organizational structure of R&D. The once prevalent form of centralized in-house R&D has vanished, while the boundaries of organizations have become less fixed and more open-ended as the development of novel products and services requires greater coordination across various types of organizations (Lingo and O'Mahony, 2010). The wellspring for new ideas is no longer the individual firm; the generation of novelty takes place at the nexus of different kinds of organizations.

Collaboration has also notably increased with organizations in the public sphere, such as universities, nonprofit organizations, and governmental agencies. With this upsurge, organizational 'castles' with clear boundaries, strong fences, and protection from outside forces have been superseded by numerous types of partnerships and affiliations. Formerly canonical decisions, such as make or buy or centralized versus decentralized R&D, are inadequate in a high-velocity environment. As more types of alliances appear, organizational boundaries become more fluid and porous, which poses very different questions and challenges for both contemporary corporations and scholars.

In recent decades, teaming up with competitors has become a frequent form of organizing. Particularly in fast-paced industries, linkages among direct competitors – also coined coopetition – offers access to valuable resources and knowledge for established and young firms alike (Bengtsson and Kock, 2000). Competitors typically hold highly relevant knowledge and complementary resources. Such interfirm arrangements therefore often lead to enhanced reciprocal learning and higher levels of absorptive capacity (Cohen and Levinthal, 1990). Partnerships open up access to new ideas, opportunities, critical information, and an expanding array of informal contacts. Such relationships can take various forms, ranging from strategic alliances and joint ventures to less formally defined partnerships.

Collaborative arrangements with suppliers and subcontractors have gained prominence along every stage in the value chain. In vertical interfirm linkages, corporations typically engage in the exchange of specific resources and services rather than the pooling of complementary assets. Relying on subcontractors or external suppliers reflects an effort to reshape the contours of production by substituting outside procurement for in-house production. Suppliers and subcontractors work under short time frames, provide considerable variety of designs, spend more on R&D, and deliver higher quality, while the 'lead' firm affords relation-specific investments and reciprocal access through data sharing and security through longer term relationships. Procurement relationships are no longer just a source of high quality and critical components, but generate mutual benefits including reduced uncertainty and performance advantages (Dyer, 2000).

Among the broad array of collaborators are also customers and consumers of the product or service under development. Co-creating with end users and entire user communities enables firms to tap sources of new ideas and opportunities, develop applications with immediate commercial potential, and enhance their position in an emerging technological field (von Hippel, 2005). User–firm arrangements have become particularly visible in the production of software. As opposed to the closed model of R&D in which firms protect their prototypes from outsiders, these forms of organizing pursue development in an open manner (Chesbrough, 2005). In such settings, knowledge gains stem from diverse contributors of a community and potential buyers of a product.

Public and nonprofit organizations also increasingly redefine their boundaries. Growing linkages among public, private, and for-profit entities lead to the rise of hybrid structures. Such public–private 'hybrids' have formed at the crossroads of the academy and commerce, between business firms and nonprofit organizations, and with the government (e.g., Murray, 2010; Jay, 2013). The various collaborative arrangements in which external partners take an active role reshape organizational boundaries, making them more porous and requiring new skills to manage.

Ramifications

Technological progress today takes place at the nexus of various kinds of organizations. These interstices between organizations are the locus where knowledge is generated and new products and services developed. But such fluidity and porousness poses challenges for managers and scholars alike. How can organizational boundaries be analyzed if they are in flux? If there is no natural stopping point in the chain of decisions to devolve centralized control, fixing the boundaries of an organization becomes a daunting task, as relationships with suppliers, subcontractors, customers, and competitors evolve in unexpected ways. When network ties proliferate and deepen, it becomes more sensible to exercise voice rather than exit. A mutual orientation between parties may be established, based on knowledge that the parties assume each has about the other and upon which they draw in communication and problem solving. Fixed contracts are thus ineffectual, as expectations, rather than being frozen, change as circumstances dictate. At the core, then, of this form of relational contracting are the entangling strings of reputation, interdependence, and learning.

Much sophisticated technical knowledge is tacit in character – an indissoluble mix of design, process, and expertise. Such information is not easily transferred by license or purchase. Moreover, passive recipients of new knowledge are less likely to appreciate fully its value or be able to respond rapidly. Learning is closely linked to the conditions under which knowledge is gained (Powell et al., 1996; Brown and Duguid, 2002). Thus, regardless of whether collaboration is driven by calculative motives, such as filling in missing pieces of the value chain; by strategic considerations to gain access to new knowledge; or by prior social relations, network ties become admission tickets to high-velocity races. Connectivity to an interorganizational network and competence at managing various types of partnerships have become key drivers of this new logic of organizing.

Motives and goals for the formation of such linkages are as varied as the forms they take. The growth of collaborative linkages is driven by divergent factors and pursued in a different manner by a wide array of organizations. Larger organizations are making their boundaries more permeable in order to procure key components or critical R&D. Subcontracting and outsourcing are steps taken to reduce fixed overheads. Organizations cooperate with competitors and users in order to take on projects too risky or complex for one entity to pursue alone. Clusters of small organizations collaborate, and so do public and private organizations, cohering into a network to create what no single entity could on its own. In sum, organizations are coming to resemble a network of treaties because these multifaceted relationships encourage learning from a broad array of collaborators and promote experimentation with new forms of organizing, while at the same time reducing the cost of expensive commitments. One tangible outcome of these network initiatives is enhanced technological innovation.

Novelty and Multiple Networks

Clearly, innovation is increasingly likely to occur at the crossroads of various types of organizations, but explaining when and how this occurs is undertheorized. A conceptual framework offered by Padgett and Powell (2012), which sheds light on how organizational novelty arises, takes shape, and persists, combines research on multiple social networks with the biochemical literature on the origins of life, using the concept of autocatalysis. New organizational forms do not rise like a phoenix. The mechanisms out of which new genres develop and evolve are fundamentally relational and dynamic. At the intersection of multiple, interlinked social networks, ideas and practices are transposed across multiple social worlds and repurposed for new uses.

Organizational forms can be conceived of as assemblages of elements and practices. For newness to occur, existing templates and ideas must flow together and be recombined in novel ways, and integrated into the social world. Although individual skills and competencies are certainly important ingredients in this process, innovation is most often the result of collective experimentation. Such generative dynamics frequently occur when otherwise distant social arenas are brought together. Structural spaces between multiple networks have a high potential for foreign material and elements to flow together and become repurposed and rearranged to form a fresh constellation. Multiple networks are linked when people and organizations, either stemming from diverse domains or simultaneously occupying different roles in distinct networks, begin to interact. These intensified interactions can spur coordination and feedback across networks and may result in creative outcomes. Such interactions can take place in more formal spaces, in occasional in situ gatherings, or may be prompted by necessity or serendipity.

Other structural conditions that catalyze multiple networks involve the presence of brokers (Fleming et al., 2007; Lingo and O'Mahony, 2010), influential multivocal leaders (Padgett and Ansell, 1993), and open elites (Powell et al., 2005). These individuals and organizations facilitate the flow of ideas, and

strengthen interaction and interdependencies across realms. Brokers can be effective in connecting otherwise detached parties and social worlds, serving to transfer the creative integration of distant ideas. Multivocal actors utilize language and symbols that appeal to members of diverse networks. Open elites spur novel connections by sponsoring newcomers to join networks. The establishment of such diverse new linkages accelerates the spread of practices and promotes innovation.

The genesis of novel forms and approaches is seldom anticipated. Emerging forms can be either incremental in nature or disruptive with profound ramifications. A contrast that illuminates these differential effects is between innovation and invention (Padgett and Powell, 2012). Innovations improve on existing ways of doing things, whereas inventions change the ways things are done. Thus, when organizational inventions occur, they can have transformative effects on the interlinked social context and alter social boundaries. In contrast, most organizational innovations do not have such wide-ranging effects. But when new forms of organizing alter the social context, and reverberate back and have systemic repercussions for the social networks that originally produced them, significant organizational change occurs.

As one example, the transposition of entrepreneurship into the domain of the university – the emergence of academic entrepreneurship – was initially contested, but subsequently had profound ramifications as commercial engagement became integrated into the academic realm. The novel constellation of commercializing university research had even more systemic effects as it spread into the interconnected domains of the university, industry, and government. The growing role of entrepreneurship in academia spurred new university careers as scientific entrepreneurs became rewarded and encouraged, which fostered industry–university collaborations with academics on industry advisory boards and industry-sponsored research. With the spread of university–industry interfaces, a new organizational regime of proprietary science took root in academia, a social context in which ownership of one's research findings was previously considered illegitimate (Murray, 2010; Colyvas and Maroulis, 2012).

Similarly, community principles such as collective production and voluntarism have been repurposed into commercial settings, promoting open collaboration in the commercial software development domain (O'Mahony and Lakhani, 2011). Previously the home of hackers and hobbyists, community forms such as open source software gained acceptance in the commercial world and in the eyes of governments. The growth of community forms also thrived beyond software and reshaped many existing configurations in the public and private sectors. Prime examples range from voluntary online communities such as Wikipedia and crowdsourcing platforms, to increasingly popular forms of shared consumption including cars and accommodations, to the upsurge of worldwide accessible online learning platforms spreading across universities.

Conclusion

The examination of organizational boundaries was long dominated by efficiency and power considerations (Pfeffer and Salancik, 1978; Williamson, 1985). These approaches led to an atomistic conception of organizations, with research efforts to clearly demarcate one organization from another, and from its environment (Santos and Eisenhardt, 2005). In a more complex and interconnected world, the image of organizations as either dyads or at the center of an ego-centered network neglects the upsurge of collaborations in organizational life, and the interconnectedness of social and economic life in general. Studying organizations as independent in their boundary decisions and detached from multiple relational influences has become inadequate.

The authors encourage future research to systematically examine the social processes out of which novel forms and collaborations arise and are sustained. What are the places where members from distinct domains meet and interact? What structural conditions foster the coevolution of multiple networks? Conceptual as well as empirical work is needed that highlights antecedent conditions that both constrain and enable novel constellations. Work in this area could adopt a multilevel approach to explore both macro- and microlevel attributes affecting the formation and coevolution of interconnected social arenas. There is also profound need for research that focuses on organizational failures and attempts that do not come to fruition. Such work could significantly contribute to our understanding of the social mechanisms and structural forces that defer or impede organizational innovation. Another promising line of investigation examines the role of boundary work and boundary organizations in enabling collaborations across distinct social arenas (O'Mahony and Bechky, 2008). Research focused on the intersection of boundaries and multiple social networks will afford insight into the shifts taking place in the nature and position of boundaries in today's interconnected organizational landscape.

See also: Corporate Culture; Design and Form: Organizational; Development: Organizational; Ecology: Organizations; Modularity and Organizations; Narrative Approaches to Organizations; Network Paradigm: Applications in Organizational Science; Organization: Overview; Organizations, Sociology of.

Bibliography

Bengtsson, M., Kock, S., 2000. "Coopetition" in business networks—to cooperate and compete simultaneously. Industrial Marketing Management 29 (5), 411–426.

Brown, J.S., Duguid, P., 2002. The Social Life of Information. Harvard Business School Press, Boston.

Carlile, P., 2004. Transferring, translating, and transforming: an integrative framework for managing knowledge across boundaries. Organization Science 15 (5), 555–569.

Chesbrough, H., 2005. Open Innovation. Harvard Business School Press, Boston, MA.

Cohen, W., Levinthal, D., 1990. Absorptive capacity: a new perspective on learning and innovation. Administrative Science Quarterly 35, 128–152.

Colyvas, J., Maroulis, S., 2012. Academic laboratories and the reproduction of proprietary science: modeling organizational rules through autocatalytic networks. In: Padgett, J.F., Powell, W.W. (Eds.), The Emergence of Organizations and Markets. Princeton University Press, Princeton, pp. 496–544.

Dyer, J.H., 2000. Collaborative Advantage: Winning through Extended Enterprise Supplier Networks. Oxford University Press, Oxford, New York.

Fleming, L., Mingo, S., Chen, D., 2007. Collaborative brokerage, generative creativity, and creative success. Administrative Science Quarterly 52, 443–475.

Gomes-Casseres, B., 2006. How alliances shape competition. In: Shenkar, O., Reuer, J.J. (Eds.), Handbook of Strategic Alliances. SAGE Publications, Thousand Oaks, CA, pp. 39–53.

Hargadon, A.B., Bechky, B.A., 2006. When collections of creative become creative collectives: a field study of problem solving at work. Organization Science 17 (4), 484–500.

Heckscher, C., Adler, P.S., 2006. The Firm as a Collaborative Community: The Reconstruction of Trust in the Knowledge Economy. Oxford University Press, New York.

von Hippel, E., 2005. Democratizing Innovation. MIT Press, Cambridge, MA.

Hounshell, D.A., 1984. From the American System to Mass Production 1800–1932: The Development of Manufacturing Technology in the US. Johns Hopkins University Press, Baltimore, MD.

Jay, J., 2013. Navigating paradox as a mechanisms of change and innovation in hybrid organizations. Academy of Management Journal 56 (1), 137–159.

Lingo, E.L., O'Mahony, S., 2010. Nexus work: brokerage on creative projects. Administrative Science Quarterly 55, 47–81.

Meyer, A.D., Gaba, V., Colwell, K.A., 2005. Organizing far from equilibrium: nonlinear change in organizational fields. Organization Science 16 (5), 456–473.

Murray, F., 2010. The Oncomouse that roared: hybrid exchange strategies as a source of distinction at the boundary of overlapping institutions. American Journal of Sociology 116 (2), 341–388.

O'Mahony, S., Bechky, B.A., 2008. Boundary organizations: enabling collaboration among unexpected allies. Administrative Science Quarterly 53 (3), 422–459.

O'Mahony, S., Lakhani, K.R., 2011. Organizations in the shadow of communities. Research in the Sociology of Organizations 33, 3–36.

Padgett, J.F., Ansell, C., 1993. Robust action and the rise of the Medici, 1400–1431. American Journal of Sociology 98 (6), 1259–1319.

Padgett, J.F., Powell, W.W., 2012. The problem of emergence. In: Padgett, J.F., Powell, W.W. (Eds.), The Emergence of Organizations and Markets. Princeton University Press, Princeton, pp. 1–29.

Pfeffer, J., Salancik, G.R., 1978. The External Control of Organizations: A Resource Dependence Perspective. Harper and Row, New York.

Powell, W.W., Koput, K., Smith-Doerr, L., 1996. Interorganizational collaboration and the locus of innovation: networks of learning in biotechnology. Administrative Science Quarterly 41, 116–145.

Powell, W.W., White, D.R., Koput, K., Owen-Smith, J., 2005. Growth of interorganizational collaboration in the life sciences. American Journal of Sociology 110 (4), 1132–1205.

Rosenbloom, R.S., Spencer, W.J., 1996. The transformation of industrial research. Issues in Science and Technology 12 (3), 68–74.

Sabel, C.F., 1994. Learning by monitoring: the institutions of economic development. In: Smelser, N.J., Swedberg, R. (Eds.), Handbook of Economic Sociology. Princeton University Press, Princeton, NJ, pp. 137–165.

Santos, F.M., Eisenhardt, K., 2005. Organizational boundaries and theories of organization. Organization Science 16 (5), 491–508.

Smith-Doerr, L., Powell, W.W., 2005. Networks and economic life. In: Smelser, N.J., Swedberg, R. (Eds.), Handbook of Economic Sociology, second ed. Princeton University Press, Princeton, NJ, pp. 379–402.

Stark, D., 2009. The Sense of Dissonance: Accounts of the Worth in Economic Life. Princeton University Press, Princeton, NJ.

Williamson, O.E., 1985. The Economic Institutions of Capitalism. The Free Press, New York.

Bounded Rationality

John Bendor, Stanford University and Center for Advanced Study in the Behavioral Sciences, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 2, pp. 1303–1307, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The research program of bounded rationality is built on three premises: (1) humans are cognitively constrained; (2) these constraints impact decision making; and (3) difficult problems reveal the constraints and highlight their significance. Thus, strategies may be optimal in easy domains (e.g., tic-tac-toe) but suboptimal in hard ones (e.g., chess). Understanding the most important mental constraints, such as selective attention and limited working memory, helps explain why people use certain kinds of heuristics (i.e., rules of thumb) in hard domains, hence why and how performance is suboptimal. However, the Simonian branch of this research program emphasizes why many highly trained specialists manage to perform *relatively* well despite working under these universal cognitive limitations. By contrast, the Tversky–Kahneman branch focuses more on errors made by everyone, experts and novices alike. Theories of bounded rationality have been applied to diverse political contexts such as elections and budgeting; these models are briefly described.

Introduction

The central ideas of bounded rationality (BR) are straightforward. First, humans are cognitively constrained in various ways, e.g., we can consciously attend to only one choice problem at a time. Second, these mental properties have behavioral consequences: they significantly affect decision-making. Third, the harder the problem the more likely it is that a decision-maker's information-processing constraints will matter (Simon, 1981).

The first claim is uncontroversial. Cognitive psychologists have demonstrated, for example, that our short-term or 'working' memory holds only a few (between about five and nine) chunks of information (Miller, 1956). The second claim follows close on the heels of the first. Since our working memory is small, our conscious attention must be selective. Hence, e.g., an agency head can work on only one problem at a time, no matter how many his subordinates present to him.

The third claim is important for demarcating domains in which BR theories have empirical 'bite', i.e., their predictions differ from those of other theories – especially those of BR's main competitor, rational choice (RC) theories. Consider Simon's formulation: "the capacity of the human mind for formulating and solving complex problems is very small compared with the size of the problems whose solution is required for objectively rational behavior in the real world – or even for a reasonable approximation to such objective rationality" (1957, p. 198). Thus, bounded rationality describes a relation between a person's mental abilities and the complexity of the problem he or she faces. It is not a claim about the brilliance or stupidity of human beings, independent of their task environments. It is easy to miss this central point and to reify the idea of bounded rationality into an assertion about our absolute capacities (e.g.,'people are dumb'). The fundamental notion is of cognitive limits, and as is true of any constraint, if cognitive constraints do not bind in a particular situation, then they do not matter: they will not affect the outcome. And whether they bind depends partly on the demands placed on decision-makers by the problem at hand. Simon has called the joint effects of "the structure of task environments and the computational capabilities of the actor … a scissors [with] two blades" (1990, p. 7). Theories of bounded rationality have cutting power – especially relative to RC theories – only when both blades operate.

Since this is the program's foundation, it is worthwhile exploring its implications via an example. Consider a group of normal adults, randomly paired up to play either chess or tic-tac-toe. A pair can either play their assigned game or they can stipulate a particular outcome, take their money for participating in the experiment and leave. The intuitive prediction is that more people assigned to tic-tac-toe would stipulate an outcome (a draw): after all, that game is trivial.

However, so – in a sense – is chess: Zermelo proved long ago that there exists an optimal way to play chess, and if played optimally then, as in tic-tac-toe, the same outcome always occurs: either white wins or black does or they draw. Indeed, for classical game theory, which ignores cognitive constraints, chess and tic-tac-toe belong to the same class: they are zero-sum, finite games of perfect information. Hence they are game theoretically equivalent. In the real world, of course, these games are not equivalent. Normal adults do not play tic-tac-toe; it is pointless. But they do play chess; indeed, for some it is a career. The point is simple but vital: our mental abilities are a binding constraint in chess but not in tic-tac-toe. Accordingly, BR and RC theories make observationally equivalent predictions about the latter but not the former. Hence BR theories have cutting power in chess – knowing the players' cognitive capacities gives us predictive leverage – but not in tic-tac-toe.

The example provides a more subtle point. Chess simplifies toward a game's end, but the players' cognitive capacities remain the same. Since the task's demands are falling while mental resources stay fixed, at some point *those resources may no longer be a binding constraint*. If that happens, the players will play optimally, and so a standard RC theory will accurately predict their behavior from that point on: a chess master burdened by a big disadvantage in the endgame will resign, for he knows it is hopeless. Once the players can identify the optimal strategies, the game becomes completely predictable, hence continuing is as pointless as playing tic-tac-toe.

This example reveals the subtlety of the contest between BR and RC theories. Theories from these two research programs predict different behaviors in chess only when the game is sufficiently complex so as to make players' mental capacities binding constraints. For expert players, this condition holds early in the game but not at the end. Thus, the two types of theories make observationally equivalent predictions about the endgame behavior of experts.

The lesson is that the theoretical significance of bounded rationality turns on the *difference* between cognitive resources and task demands, not on the former's absolute level.

The Main Branches in Political Science of BR

Bounded rationality should not be confused with a theory (e.g., of satisficing), much less with a specific formal model (e.g., Simon, 1957). It is best considered a research program: a sequence of theories with overlapping sets of assumptions, aimed at solving similar problems. Political scientists have often conflated BR with the specific theory of satisficing, but that conflation produces a serious underestimation of the program's substantive content. In principle, the program's empirical domain is vast – it is as imperialistic as the rational choice program – and so its set of possible theories is also very large.

Many research programs contain multiple branches. So it is for BR: political science alone exhibits two main orientations toward bounded rationality. The first orientation sees the glass as half full, emphasizing how people manage to do 'reasonably well' *despite* cognitive limitations, *even in complex tasks*. Here belong Simon's line of work and also Lindblom's theory of incrementalism (Lindblom, 1959; Braybrooke and Lindblom, 1963; Bendor, 1995). In the second branch, the glass is half empty: it emphasizes how people make mistakes *even in simple tasks*. Here is the Tversky–Kahneman (T–K) research tradition on heuristics and biases. (For a superb overview see Davis et al., 1966.)

The Glass Is Half Full

As noted above, one of Simon's key premises is that a decision maker's 'inner environment' (Simon, 1981) of information processing will become manifest only when the task is sufficiently difficult. This implies that in order to discover the important types of mental limits, one should study people facing hard problems, such as chess.

But analyzing chess revealed more than mental limits: it generated new questions for the Simonian line, especially about performance. People's performance in chess varies tremendously, and this variation demanded an explanation. Thus, a new research focus emerged: how do some decision-makers *do so well*, given that they face similar cognitive constraints? Thus, scholars in this branch came to see the glass as half full: on some hard problems some humans perform very well, relative to an empirically sensible aspiration level. Given that objectively optimal play in chess-like problems is clearly impossible, what is interesting is not whether people behave fully rationally – we know they don't – but how do relatively competent agents work around their cognitive limits. (See, e.g., Simon, 1990, and the associated references.)

The Glass Is Half Empty

In sharp contrast to the Simonian program's focus on 'good' performance in difficult domains, the prototypical experiment of the T–K tradition shows that even highly trained subjects answering simple questions can perform suboptimally. Kahneman and Tversky's goal – clearly stated in their seminal paper (Tversky and Kahneman, 1974, p. 1124) – was to map subtle mental processes that can cause cognitive illusions, similar to the study of perceptual illusions. This project's results are more striking if they are shown to occur even in simple tasks, just as perceptual illusions are more striking if demonstrated with simple stimuli. The results are still more striking if one shows that even experts are vulnerable to illusions.

Subsequently, however, the field has tended to emphasize more the errors themselves, rather than the underlying cognitive mechanisms: "The … approach triggered an explosion of research on inferential error, and the list of illusions, foibles, flops, and bloopers to which ordinary people were apparently prone became rather long … soon, if there was a mistake to be made, someone was making it in the presence of a social psychologist" (Gilbert, 1998). (After reading many of these studies, the glass probably does look half empty – or worse.) This focus on mistakes triggered a scholarly backlash which has tried to show that humans are better decision makers than the T–K program apparently claims (e.g., Gigerenzer, 1991; for a reply see Kahneman and Tversky, 1996). Unfortunately, this debate has clouded Tversky and Kahneman's original intention, which was not to show that *Homo sapiens* are dolts but rather to uncover fundamental cognitive mechanisms that leave their imprint nearly everywhere. Given the primary objective of demonstrating that certain mental processes are ubiquitous, T–K's approach made sense. But it had the unfortunate side effect of making the program largely inattentive to variations in performance. With a few exceptions (e.g., the study of how people calibrate subjective probabilities), specialists in the T–K branch have shown little interest in performance variation; in particular, genuinely expert performance, at the level of chess masters, has been neglected. Instead, performance evaluations in this branch mostly use dichotomous theoretical standards: do people make choices in accordance with the axioms of expected utility theory? Do they revise beliefs in a Bayesian manner? The answers – generally no, they don't – are less informative than would be answers based on a quantitative scale measuring degrees of sophistication, even if the empirical best was far short of the theoretical ideal.

However, these differences in how the branches' study performance reveal that in one respect the T–K research agenda is more ambitious: it is harder demonstrating that cognitive limits show through on simple problems than on hard ones. Thus, the T–K program has worked on *pushing out the boundaries of BR* by showing that even quite subtle problem representations ('framing') can induce suboptimal performance. (Establishing which problem-representations are transparent – i.e., cognitively obvious – and which are opaque has been important throughout the T–K program (Kahneman and Tversky, 1996).) One needn't go all the way to chess to uncover our mental limitations: humans are sufficiently sensitive (hence vulnerable) to framing so that judgmental or

decisional imperfections appear even when a task exhibits no combinatorial explosion. Thus, a main finding of the T–K program is that BR has predictive bite in a larger domain than once thought.

Essential Properties of Humans as Information-Processors

What are our essential properties as information-processors? Synthesizing Hogarth (1987) and Simon (1990) gives us the following list.

1. Selective perception of information: our objective environments always contain far more information than we can perceive or attend to. Thus, perceptions must be guided by anticipations and other forms of top-down processing (e.g., schemas).
2. High order information-processing, especially conscious thinking and attention, is largely serial. This has significant implications for the real-time behavior of busy officials.
3. Compared with modern computers, humans process many kinds of information slowly, partly due to physiological limits (neurons are much slower than electronic circuits).
4. Compared with computers, people are poor at calculation.
5. Memory is actively reconstructive, not photographic. It therefore lacks photographic fidelity.
6. Although there is no known limit to long-term memory, working memory is very small (Miller, 1956). Thus, because everything that enters the former goes through the latter, short-term memory is a key bottleneck in information-processing.

The other crucial type of cognitive property are ways in which humans work around the above constraints. Simon calls these mechanisms for procedural rationality (1981; 1990). Three are especially worth noting.

1. Recognition processes. After long experience, experts have stored a great many patterns about their tasks in long-term memory, both about situations ('this is situation type x') and actions ('in situation x, do y'). For familiar problems a fundamental trick is to *substitute recognition for search*.
2. Heuristic search. When recognition fails, experienced decision-makers fall back on search. For reasonably complex problems search is heuristic: optimality is not guaranteed. Instead, heuristics make complicated problems manageable: they cut them down to (human) size and, in the hands of experts, often yield 'good' solutions. There are two types of heuristics. If the task is highly and recognizably structured, experts will use powerful task-specific heuristics. When these are unavailable they use general (but weak) heuristics such as satisficing (stop searching for alternatives when you find one that satisfices – exceeds your aspiration level – rather than requiring an optimal solution). Note the place of satisficing here. Far from being the heart of BR, it is merely one of several general-but-weak heuristics.
3. Heuristic search occurs in *problem spaces*: mental representations of the task at hand. Experts learn to use economical and sometimes powerful problem representations.

Applications: Theories of Politics

Budgeting

Building on Lindblom, Wildavsky and his colleagues constructed an incremental model of budgeting (Davis et al., 1966) which received substantial empirical support. Crecine (1969) argued that the general heuristic of decomposition – when faced with a hard problem, decompose it into smaller, easier subproblems (Simon, 1981) – is central to budgeting, and developed a model of city budgeting based on that idea. Padgett (1980) created a model of federal budgeting that continued to use the BR notion of heuristic search but which implied that allocations will sometimes be nonincremental.

Organizational Reliability

Landau (1969) has shown that certain types of structural redundancy can make organizations more reliable than any of their subunits. (Consider several R&D teams working in parallel.) Relatedly, Ladha (1992) and others have shown how using Condorcet's jury theorem can reduce both type 1 errors (e.g., the Food and Drug Administration's approving bad drugs) and type 2 errors (rejecting good drugs). Significantly, these results hold even if individual decision-makers face quite severe cognitive constraints.

The Evolution of Cooperation

Axelrod (1981) and Bendor and Swistak (1997) have used evolutionary game theory to analyze how cooperation might emerge and stabilize when central authority cannot solve Hobbes' problem. (Evolutionary game theory assumes that people learn to use strategies that work and to discard those that fail; agents need not optimize.)

Subgovernments

Bendor and Moe (1985) constructed a model of subgovernments with myopic actors (legislators, an agency and interest groups). The simulation outcomes were typically 'pluralist equilibria,' characterized by a balance of power between the interest groups.

Electoral Competition

In Kollman et al. (1992), adaptively rational political parties compete for the support of voters in two-party elections. In unidimensional policy spaces the parties converge to the median voter (Pages, personal communication, September 1999); this result appears to hold for a wide variety of search heuristics. Hence Downs' famous result can be derived with weak cognitive assumptions. In a multidimensional policy space, the parties behave 'sensibly' but can get hung up on local optima. This suboptimization depends crucially on nonobvious aspects of voters' preferences.

Prospect Theory

The key premise of prospect theory, Tversky and Kahneman's most important theoretical contribution, is that choices are

evaluated relative to a reference point, e.g., the status quo. The second assumption is that people are risk-averse about gains (relative to the reference point) but risk-seeking about losses. The third premise is loss-aversion: losing x hurts more than gaining x helps. This simple and elegant structure has yielded suprisingly powerful predictions in the study of bargaining (Neale and Bazerman, 1991), international politics (Levy, 1997), and voting (Quattrone and Tversky, 1988).

For an excellent and accessible review of the BR program in economics see Conlisk (1996). Rubinstein (1998) is very instructive but much more technical.

Editor's Note

The research program on "bounded" or "limited" rationality continues in both political science and behavioral economics. For an update of bounded rationality thinking in the latter, see Camerer et al., 2004. Within the discipline of political science, readers will find the following of interest: Poulsen, 2014; Ostrom, 2012; Vollaard and Martinsen, 2014; Botterill and Hindmoor, 2012; Binswangera and Prüfer, 2012; Bendor, 2003.

See also: Rational Choice Explanation: Philosophical Aspects; Rational Choice Theory in Sociology; Rational Choice and Organization Theory; Rationality in Society.

Bibliography

Axelrod, R., 1981. The emergence of cooperation among egoists. American Political Science Review 75, 306–318.
Bendor, J., Moe, T., 1985. An adaptive model of bureaucratic politics. American Political Science Review 79, 755–774.
Bendor, J., Swistak, P., 1997. The evolutionary stability of cooperation. American Political Science Review 91, 290–307.
Bendor, J., 1995. A model of muddling through. American Political Science Review 89, 819–840.
Bendor, 2003. Herbert A. Simon: political scientist. Annual Review of Political Science 6, 433–471.
Binswangera, Johannes, Prüfer, Jens, September 2012. Democracy, populism, and (un)bounded rationality. European Journal of Political Economy 28 (3), 358–372.
Botterill*, Linda Courtenay, Hindmoor, Andrew, 2012. Turtles all the way down: bounded rationality in an evidence-based age. Policy Studies 33 (5), 367–379.
Braybrooke, D., Lindblom, C., 1963. A Strategy of Decision. Free Press, New York.
Camerer, Colin F., Loewenstein, George, Rabin, Matthew, 2004. Advances in Behavioral Economics. Princeton University Press, Princeton, NJ.
Conlisk, J., 1996. Why bounded rationality? Journal of Economic Literature 34, 669–700.
Crecine, J., 1969. Governmental Problem-Solving. Rand McNally, Chicago, IL.
Davis, O., Dempster, M., Wildavsky, A., 1966. A theory of the budgetary process. American Political Science Review 60, 529–547.
Dawes, R., 1998. Behavioral decision making and judgment. In: Gilbert, D., et al. (Eds.), The Handbook of Social Psychology, vol. I. McGraw-Hill, Boston, MA, pp. 497–548.
Gigerenzer, G., 1991. How to make cognitive illusions disappear: beyond heuristics and biases. European Review of Social Psychology 2, 83–115.
Gilbert, D., et al. (Eds.), 1998. Ordinary personology. In: The Handbook of Social Psychology, vol. II. McGraw-Hill, Boston, MA, pp. 89–150.
Hogarth, R., 1987. Judgement and Choice, second ed. Wiley, New York.
Kahneman, D., Tversky, A., 1996. On the reality of cognitive illusions: a reply to Gigerenzer's critique. Psychological Review 103, 582–591.
Kollman, K., Miller, J., Page, S., 1992. Adaptive parties in spatial elections. American Political Science Review 86, 929–938.
Ladha, K., 1992. The condorcet jury theorem, free speech, and correlated votes. American Journal of Political Science 36, 617–634.
Landau, M., 1969. Redundancy, rationality, and the problem of duplication and overlap. Public Administration Review 29, 346–358.
Levy, J., 1997. Prospect theory, rational choice, and international relations. International Studies Quarterly 41, 87–112.
Lindblom, C., 1959. The science of 'muddling through. Public Administration Review 19, 79–88.
Miller, G., 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review 63, 81–97.
Neale, M., Bazerman, M., 1991. Cognition and Rationality in Negotiations. Free Press, New York.
Ostrom, Elinor, 2012. Coevolving relationships between political science and economics. Rationality. Markets and Morals 3, 51–65.
Padgett, J., 1980. Bounded rationality in budgetary research. American Political Science Review 74, 354–372.
Poulsen, Lauge N. Skovgaard, March 2014. Bounded rationality and the diffusion of modern investment treaties. International Studies Quarterly 58 (1), 1–14.
Quattrone, G., Tversky, A., 1988. Contrasting rational and psychological analyses of political choice. American Political Science Review 82, 719–736.
Rubinstein, A., 1998. Modeling Bounded Rationality. MIT Press, Cambridge, MA.
Simon, H., 1957. Models of Man. Wiley, New York.
Simon, H., 1981. The Sciences of the Artificial, second ed. MIT Press, Cambridge, MA.
Simon, H., 1990. Invariants of human behavior. In: Rosenzweig, M., Porter, L. (Eds.), Annual Review of Psychology, vol. 41. Annual Reviews, Palo Alto, CA, pp. 1–19.
Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. Science 185, 1124–1131.
Vollaard, Hans, Martinsen, Dorte Sindbjerg, 2014. Bounded rationality in transposition processes: the case of the European patients' rights directive. West European Politics 37 (4), 711–731.

Bourdieu, Pierre (1930–2002)

Gisèle Sapiro, École des hautes études en sciences sociales, Paris, France; and CNRS (Centre européen de sociologie et de science politique), Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

One of the most important thinkers of the second half of the twentieth century, French sociologist Pierre Bourdieu (1930–2002) developed a theory of social domination and reproduction based on the concepts of 'habitus', 'capital' (economic, cultural, social), 'field', and 'symbolic violence'. This theoretical framework is rooted in extensive empirical research in the sociologies of education, culture, work, and social stratification. He also dealt with gender, political sociology, and economic sociology. Following the evolution of his thought and career, the article presents his theory of practice, social classes, field, and symbolic violence, ending with his engagement as a public intellectual.

Bourdieu's theory of the social world combines a structural approach to power relations with a historical perspective on reproduction and change. It provides an original synthesis of the classical theories of Marx, Durkheim, and Weber, while striving to overcome the traditional oppositions between structure and individuals, objectivism and subjectivism, continuity and discontinuity. Agents are determined by the class and gender conditions in which they were socialized (habitus), but still have latitude to maneuver in different situations according to their dispositions and the various types of capital (economic, cultural, social) they are endowed with. Their actions and strategies are oriented by the rules and values of the fields in which they occur and by the position they occupy in these fields. Contrary to rational action theory, Bourdieu's theory of action distinguishes between different kinds of interests with regard to the "illusio", the belief in the issues at stake in the field (political, economic, legal, professional, artistic, literary, and so on). The focus on belief allows him to highlight the symbolic aspects of domination and the mechanism of symbolic violence.

From Philosophy to Anthropology: The Theory of Practice and the Concept of Habitus

Pierre Bourdieu was born on 1 August 1930, in Denguin, France, (Pyrénées-Atlantiques department) to a father who was first a farmer, then a postman, and then a postmaster, and a post office employee mother. He attended the Lycée de Pau before becoming a student at Lycée Louis-le-Grand in Paris, from which he entered the prestigious École normale supérieure on the Rue d'Ulm in 1951. In 1954, he obtained the philosophy *agrégation* (a competitive certificate providing access to high school teaching positions) and became a teacher at the Lycée de Moulins. In 1955, in the middle of the war for Algerian independence, Bourdieu was called to military duty in Algeria. His research on Algerian society began at this point. Partway between Orientalism and ethnology, it resulted first in a small book entitled *Sociologie de l'Algérie*, which was published in 1958, the same year he was named as Assistant Professor at the University of Algiers. In order to understand the transition from precapitalist societies to capitalist economies, he then conducted three studies: the first, carried out in 1958, dealt with traditional Kabyle society; the second (1960) with Algerian workers in urban spaces (*Travail et travailleurs en Algérie*, 1963; *Algérie 60*, 1977a); and the third, also from 1960, dealt with regrouping centers and the effects of colonialism more generally (*Le Déracinement*, 1964, with Abdelmalek Sayad).

Together with a study on celibacy in the sociologist's native province of Béarn, which he undertook upon his return to France in 1960, these studies form the basis for Bourdieu's theory of practice and habitus (*Esquisse d'une théorie de la pratique*, 1972; *Le Sens pratique*, 1980). The theory was born out of a confrontation with Lévi-Strauss' structural anthropology and draws from Bourdieu's reading of Merleau-Ponty, Husserl, Heidegger, Weber, and Wittgenstein. His reflection on the concept of practice has its roots in a double critique of subjectivism and objectivism. On the one hand, from Sartre's conception of freedom to rational choice theory, subjectivist approaches assume that action is produced by the will of a conscious subject, capable of projecting herself into the future and anticipating the consequences of her actions, both of which are assumptions Bourdieu rejected. On the other hand, structuralist objectivism turns agents into quasi-mechanical conveyors for structures. The polysemic use of the notion of the rule in structural anthropology, at times appearing as a norm that agents follow consciously, and at other times as a principle reconstructed by the researcher in order to account for the regularity of practices, alerted Bourdieu to the breaches and biases of this approach, and he criticized its intellectualist bias (he later termed it a 'scholastic bias'). Moreover, the inscription in time of exchanges such as the gift and the counter-gift, or the challenge and the counterattack, renders insufficient any mechanistic explanation of the logic governing these interactions. The theory of practice emphasizes the specificity of the logic of practice, which is carried out in particular situations, and often with immediacy, as opposed to the time and reflexivity that characterize scholarly activity.

Scholarly reason fails to account for practice because, as Marx said, it takes the things of logic for the logic of things. However, the logic of practice is radically different from logic: it is a fuzzy logic that evades attempts to apply systematically the structural method in order to reconstruct the system of cultural oppositions, because of the ambiguous signification of certain

elements that may vary according to the principles of oppositions they refer to, meaning that there is no perfect structural homology between these different principles; Bourdieu observed this with Kabyle ritual practices. The practical sense ('sens pratique') is at once a sense of orientation and a sense of the game: that is what makes it possible for agents to adjust to an infinite number of situations without explicitly following norms, rules, or codes (and the mechanistic structural approach cannot explain this), and at the same time without actually obeying that which their thought freely decrees (as subjectivist or rational choice theories would have it).

How is this adjustment possible? The answer to this question lies in two key concepts from Bourdieu's theory of practice: habitus and strategy. Bourdieu began to theorize the concept of habitus in the afterword to his French translation of Panofsky's book, *Architecture gothique et art scolastique* (1968), then took it further in *L'Esquisse d'une théorie de la pratique* (1971), defining it as the schemes of perception, evaluation, and action that individuals internalize during their primary (familial) and secondary (school-based) education in the form of dispositions. These 'structured structures', which frame the agents perception, action, and judgment, underlie the adjustment between subjective perception and the objective structures of the social world. They explain the agents' ability to develop strategies that are more or less adapted to the situations they encounter. The notion of strategy refers to the extent to which the agents are able to improvise according to their dispositions.

Examples of perfect adjustment between strategies and situations represent, however, only one particular case among different possibilities. In his studies on Kabyle society and the peasant society in Béarn, Bourdieu observed the contrary: individuals had difficulty adjusting to new situations in contexts of social transformation, that is, when social structures no longer resembled those they had internalized, and when the value system had changed. Cases such as these reveal the gap between subjective perception and objective structures in a world undergoing change. The inability to adjust to change results from a contradiction between traditional values (such as joint ownership of property, the logic of honor which rejects calculations and rationalizations of economic exchanges, the notion of time) and capitalist logic. The inertia of the past is felt through deeply ingrained habits. Bourdieu's ethnographic description of a ball where celibate males, rooted resolutely in the land, stay grouped together in a corner while country girls dance with boys from the city reveals the demonetization of peasant values in the Béarn. This description of the body of peasants, published in 1962, contains the first appearance of the term habitus in Bourdieu's work. Here it is associated with hexis, referring to Marcel Mauss and his analysis of 'techniques of the body'. But the body is also the site of practice, invention, and improvised behavior, more or less adjusted to given situations. This faculty of improvisation remains nevertheless limited by the conditions of its socialization, that is, by habitus, as incorporated habits often resist – the expression of their inertia – individuals' most conscious attempts to master their bodies.

This theory of practice therefore establishes a general theory of action, ranging from the most ritualized actions to the most inventive (creation, scientific innovation). All the while taking this faculty of improvisation into account, the concept of habitus thereby outlines the limits of free action, in contradistinction to Sartre's philosophy of freedom and Goffman's sociology of roles. The concept of habitus therefore allows Bourdieu to integrate the achievements of different sociological traditions: the Durkheimian tradition of studying forms of classification, especially in its structuralist adaptation, the Marxist tradition, which relates forms of consciousness to conditions of existence, and the Weberian critique of Marxist materialism, which re-establishes the place of social groups' vision of the world in the orientation of their action.

The concept of habitus is also at the heart of the Bourdieuan theory of domination, which relies on 'symbolic power'. Bourdieu's conception of symbolic power is in line with the neo-Kantian rationalist tradition, which holds, against the empiricist tradition, that the vision of the world is structured by categories of perception. In sociology, this tradition is represented by Durkheim and Mauss, and in philosophy, by Cassirer. Bourdieu reminds us, however, following Marx and Weber, that these categories are not neutral, but rather fulfill ideological functions and thus underlie the mechanisms of domination (Bourdieu, 1977a,b). For Bourdieu, symbolic power derives its efficacy from being internalized in bodies, in the form of dispositions that constitute the habitus. Symbolic power is observed, for example, in linguistic exchanges, which result from the encounter between a linguistic habitus and a market that confers meaning upon them (Bourdieu, 1982a,b). Indeed, language is at once a site of expression of common sense, that is, of the schemes of perception and classification instilled in education that underlie 'logical conformism', and the site of separation between social groups, between sacred and profane knowledge. A place of interaction between thought and the body, language is both a medium of communication and the vehicle for the daily exercise of domination via symbolic violence.

Bourdieu's reflection on the symbolic dimension of exchanges and power relations in society originated in his observations of precapitalist societies. In precapitalist economies, where there is no systematic reproduction of the mode of domination, political authority comes from the asymmetry of exchanges, that is, from ostentatious redistribution, but also from an ensemble of strategies oriented toward the domination of people via the mechanisms of symbolic violence (Bourdieu, 1972, 1980). This observation leads the sociologist to a refutation of what he calls 'economism', that is the propensity to reduce exchanges to material interest. Bourdieu contests the relevance of this analysis even in capitalist societies, as he argues later in his book *Les Structures sociales de l'économie* (Bourdieu, 2000a,b). Symbolic capital is credit, broadly speaking, based on a group's belief system. In Kabylie, symbolic capital takes the form of honor, but in capitalistic societies, cultural capital is the paradigmatic embodiment of this symbolic capital, as Bourdieu showed in his studies of contemporary French society.

The Concept of 'Cultural Capital' and the Theory of Social Classes

These studies were undertaken upon Bourdieu's return to France at the Centre de sociologie européenne, founded by Raymond Aron, and of which the young sociologist became co-director in

1964, the year he was hired as a Directeur d'études (equivalent to a full professorship) at the École pratique des hautes études (after having been Maître de conférences in Lille from 1961 to 1964). He pursued these studies at the Centre de sociologie de l'éducation et de la culture, which he and his team founded in 1970, and where he remained director until 1984 (in 1997, the center became the Centre de sociologie européenne, and in 2010, the Centre européen de sociologie et de science politique). The results of these studies were published in the series Bourdieu launched in 1964, and edited until 1992, at Éditions de Minuit, 'le sens commun' (from 1997 on, the series was passed on to the 'Liber' series at Seuil). Results were also published in the interdisciplinary journal *Actes de la recherche en sciences sociales*, founded by Bourdieu in 1976.

Never separating theory and empirical work, Bourdieu argued for a rationalist epistemology that stood at equal distance from naive inductivism and relativistic constructivism: he called for epistemological vigilance at the crucial step of the construction of the object, which depended on well-controlled hypotheses (Bourdieu et al., 2005). For this reason, he considered research as a craft, a practice, that should be transmitted by training, a kind of 'apprenticeship', like in the arts, rather than through theoretical teaching (see Bourdieu and Wacquant, 1992). The epistemology he defended was also critical, since it questioned common sense and doxa.

With the tools of empirical sociology, Bourdieu dealt with 'burning' and current topics such as student social inequalities in the face of culture as higher education underwent 'democratization' (*Les Héritiers*, 1964), new cultural practices such as photography (*Un art moyen*, 1965), and museum visitation patterns at a time when the Culture minister, André Malraux, was implementing a policy that aimed at making culture accessible to all (*L'Amour de l'art*, 1966).

In *Les Héritiers*, co-written with Jean-Claude Passeron, he revived the sociology of education, proposing a more suitable way of calculating a student's chances for academic success. In order to show social inequality at school, it is indeed not enough to calculate the percentage of students enrolled in university according to the socio-professional category of the parents. Rather this enrollment rate must be linked back to the representation of the categories within the population. It then becomes clear that a student's chances of attending university are less than 1% for a child of farm workers, 1.4% for a child of blue-collar workers, and close to 60% for a child of learned professionals and upper-level executives. Why do students from the dominant classes do better at school? First, because the culture taught at school is closer to the culture of the dominant classes. But this is not all. In *La Reproduction* (1970), Bourdieu and Jean-Claude Passeron analyze the arbitrariness of pedagogic action. Indeed, teachers tend to value the least scholastic aptitudes (such as distance from educational culture) as well as the inherited culture of the family, to the detriment of the scholastic education given at school. The unequal power relation is thereby rendered legitimate by the ideology of the 'gift', wherein academic success is explained by innate capacities and personal merit. Contrary to the Marxist conception, which bases class differences on the mere possession of material means of production, Pierre Bourdieu sheds new light on the weight of inherited 'cultural capital' as a condition for academic achievement, which therefore contributes to reproducing the dominant classes. On the eve of May 1968, this research met with a strong echo in circles of students and teachers, who were experiencing a boom in demands for education, and were dealing with the system's incapacity to respond favorably to the expectations it had engendered.

The results of this study of the modes of domination were expanded and generalized in the survey on tastes and lifestyles that is at the core of Bourdieu's major work, *La Distinction* (1979). Dominant classes distinguish themselves from dominated ones in affirming their distance from necessity in their cultural tastes and lifestyles. Cultural capital, and more generally symbolic capital, appears as the mode of legitimization of social domination. Bourdieu's approach is different from the Marxist theory of social class in its analysis of lifestyles that differentiates fractions of classes not just according to the overall amount of capital, but also according to the capital's composition. The empirical study of lifestyles reveals indeed that when revenues are equal, lifestyle (cultural practices, type of expenditures and investments, sport activities) varies depending on the cultural capital the agents own.

Consequently, Bourdieu analyzes the structure of the social space as a chiasm. Dominant classes are set apart from dominated classes by the overall volume of capital they possess. But this hierarchic space is also structured by the type of capital owned. Dominant classes are split into fractions depending on those with predominantly economic and political capital, and those with mostly cultural capital (learned and intellectual professions), who are dominated within the larger 'field of power'. This same principle of differentiation can be found within the dominated classes, mainly between the educated petite bourgeoisie (schoolteachers, librarians, etc.) and the petite bourgeoisie of merchants. This differentiation of the types of capital correlates with differences in habitus between class fractions.

This chiasm is the result of the process of affirmation of cultural capital, which was imposed and made relatively autonomous with the ideology of the meritocracy. The system of direct reproduction, prevalent under the Ancien Régime, where social hierarchy depended on birth and the direct designation of family inheritors (the eldest was the inheritor, the second was destined for priesthood, girls were exchanged between families or sent to the convent, etc.), as well the tradition of father-to-son transmission of trades, was replaced by a system of reproduction with an 'educational component'. Instead of the old feudal nobility, a 'state nobility' emerged, born of a selection process that is meritocratic in appearance, but contributes in reality to legitimizing social reproduction. The social overselection of students in the *grandes écoles* (elite schools) in France, revealed in a study Bourdieu dedicated to the subject, bears witness to this process. The study (Bourdieu, 1989) articulates reflections on the educational system and the field of power.

It was during this same period that Bourdieu theorized the notion of social capital, which was to become a central concept in the sociology of networks during the 1990s (Bourdieu, 1980). He defined social capital as a network of long-lasting relations between persons who know each other and have economic and symbolic exchanges. In order to maintain social capital, the agents who own it must develop strategies. Though related

to economic and cultural capital, social capital is relatively autonomous and can have a multiplier effect.

The Theory of Field

Beginning in the 1970s, in his seminars (Bourdieu, 2013a) and in the articles collected in the volume *Questions de sociologie* (Bourdieu, 1980), Bourdieu developed his theory of field. The process of affirmation of cultural capital corresponds to the historical process of differentiation and specialization of the work of domination, which, as Max Weber analyzed in the case of religion, occurs through the formation of a group of specialists (Bourdieu, 1971a). The same is true in the case of law, which replaces the religious principle in establishing the legitimacy of political authority, and for other spaces of intellectual production – science, literature, art, music – which Bourdieu defines as fields because of their relative autonomy and specific rules of functioning. The fields of cultural production are spaces of competitive struggles over the monopoly of the specific capital related to a particular activity. These struggles occur according to specific rules (for instance, argumentation in the scientific field), and investment in the game supposes a belief in the game – illusio. Like Bourdieu's conception of social classes, his field theory is based on a relational and topographical approach: agents occupy social positions which are objectively defined in relation one to the other, independently of their interactions and subjective perception. Fields are thus structured around two main oppositions, homologous to the structure of social space. The first places the dominant and the dominated in opposition to one another according to the overall volume of specific capital. This opposition often overlaps with the opposition between the old or the 'established', who tend to defend the 'orthodoxy', and the young or the 'newcomers', who try to subvert this 'orthodoxy' and thus appear as heretical. The religious field, especially the Weberian opposition between priest and prophet, is paradigmatic here (Bourdieu, 1971b). The second opposition confronts the defenders of specific capital, guaranteed via peer judgment (autonomy), with those who subject their production to external, economic or ideological criteria (heteronomy). After the intellectual and religious fields (which served as a matrix), Bourdieu began to study the scientific field (Bourdieu 1976, 1997b) and the political field, which he criticized for its functioning and its fencing in through favoring 'professionals' and excluding the 'profane' (Bourdieu, 1971c, 1977c, 1981a,b, 2002b; Wacquant, 2005). These reflections were to make up an entire research program that would spark projects in different disciplines, namely in political science and the sociology of literature, in France and abroad.

In 1981, Bourdieu was elected on a Chair of Sociology at the Collège de France, the country's most prestigious academic institution (his inaugural lesson, which he gave on 23 April 1982, was published as *Leçon sur la leçon*, 1982). Adopting a reflexive approach, he carried out an empirical study on the academic field. He showed that the field is structured generally by the opposition between faculties that dominate in terms of temporal power (law and medicine) and those that are dominated (sciences and humanities), where symbolic capital prevails. In each discipline, however, there is an opposition between a heteronomous pole tending toward accumulating temporal power through the control of reproducing bodies, and an autonomous pole tending toward accumulating symbolic capital. In the book that was born out of this research (*Homo academicus*, 1984), Bourdieu also develops a model of analysis for historical sociology through the study of the 'critical moment' of May 1968, the structural causes of which he attempts to understand. The historic 'event' results from the interaction of independent causal series, which creates a conjuncture, and from the synchronization (or harmonization) of the different fields through their politicization.

Bourdieu's inquiry on Heidegger (1988) illustrates, through a close reading of his work, what he calls the 'field effect', that is, the mechanisms by which ethical and political dispositions are euphemized through philosophical concepts and are embedded within philosophical problems, following the specific rules of the field.

In *Les Règles de l'art* (1992b), Bourdieu systematized his theory of fields, gathering all of the research he had done on the literary field since the 1960s. In the first part, he retraces the historical process by which the literary field gained autonomy in the nineteenth century, with the development of a pole of small-scale circulation based on peer recognition, as opposed to the pole of large-scale circulation, governed by the law of the market and sales figures. In imposing the primacy of symbolic value over commercial value, this small-scale circuit managed to reverse the principles that govern social space and the economic world. Bourdieu speaks of an 'inversed economy'. Instead of the immediate profitability of best-sellers, small-scale production makes bets on long-term recognition, which can lead to 'canonization', that is, the integration of works turned 'classics', into a national and universal heritage, namely through the educational system.

In the second part of the book, Bourdieu lays the foundations of a science of works. The concept of field, he argues, helps to overcome the alternative between an internal analysis of works, as structuralism suggests, and an external analysis that reduces them, as in the Marxist reflection theory, to their conditions of production. For Bourdieu, the science of works requires three operations: (1) an analysis of the position of the literary field within the field of power (its degree of autonomy) and its evolution in different sociohistorical conjunctures; (2) an analysis of the structure of the literary field according to objective relations between the individuals and groups who compete in the field for the acquisition of specific capital; (3) an analysis of the habitus of these individuals and their social trajectories, since the field effect is not exerted mechanically, but rather it is mediatized by the space of possibilities such as agents perceive it according to their dispositions.

This approach may be applied to other fields of cultural production. Bourdieu used it, for example, in his research on Manet, which he began in the 1980s and to which he dedicated his 1998–99 a series of courses at the Collège de France. These courses, as well as the unfinished manuscript or a book written with his wife, Marie-Claire Bourdieu, deal with the conditions under which symbolic revolutions take place (Bourdieu, 2013b). The sociologist reconstructs the academic system, which exerted a monopoly on artistic formation and the art market during Manet's time. The system functioned as a body organized around its reproduction. Painters careers

followed a predefined path: competition for access to the École des Beaux-Arts, apprenticeship in studios, prizes, exhibition of their works at the Salon through a selective process that gave them access to the art market, and, for the very best, State commissions. The academic aesthetic proceeded from this system: it valued technical mastery (perspective, relief, clair-obscur), readability (or the message), reference to history (at the basis of the hierarchy of genres), and the 'finish'. This aesthetic code, backed up by a state monopoly, was to be destroyed by Manet. However, against a heroic, idealistic vision of Manet's revolution, Bourdieu aims to inscribe this symbolic revolution in the social history it results from, without reducing the aesthetic stakes to socioeconomic and political ones. Such a social history combines technical revolutions (such as the invention of tubes of paint, making it possible to paint outdoors) and morphological factors, particularly the growth of a population of artists who had been excluded from the system and who were, with their anti-academic sentiment, to invent a bohemian lifestyle and an alternate, parallel system with societies of artists selling their own works. Manet's dispositions, the different types of capital (economic, cultural, social) he owned, and his 'cleft' habitus, split between the two poles of field of power, economic and cultural, explain the painter's ability to carry out this revolution.

The Theory of Symbolic Violence

Bourdieu's work on education and the academic system of painting are part of a more general reflection on the State and its role in inculcating the categories of thought that underlie social order. This was the theme of his Collège de France courses between 1989 and 1991 (Bourdieu, 2012). Completing Max Weber's definition of the State, Bourdieu defines it not just in terms of the monopoly of legitimate violence, but also by the monopoly of legitimate symbolic violence – and the latter monopoly suffices, since it legitimatizes physical violence. Contrary to the Marxist tradition, which reduced the State to a superstructure, for Bourdieu the State is the space where 'legitimate identities' are produced and justify procedures of inclusion and exclusion. While the monopolization of symbolic violence must be studied within a perspective of historical sociology, focusing on the role of certain social groups such as jurists, one may also observe this violence on an everyday level in ordinary acts such as purchasing a house, as the ethnological approach illustrates. As a place where the 'public' and the 'official' are created, the State is defined moreover by its monopoly on the universal. It has the power to universalize certain practices or characteristics at the expense of others, which are in turn referred to as particularisms, such as provincial accents versus those of the Parisian elite, or culture said to be popular versus legitimate culture sanctioned by the educational system, or qualities said to be feminine versus masculine attributes. In this way, the State produces the taxonomies and hierarchies that are internalized by the individuals it governs, including those who rebel against it. For Bourdieu, however, the State is not an abstraction: he argues in favor of a sociology of institutions and individuals that make up the meta-field of the State.

The concept of symbolic violence aims to overcome the alternative between coercion and consent in order to understand the mechanisms of domination. Symbolic violence functions via three components acting simultaneously: ignorance of the arbitrariness of domination; recognition of this domination as legitimate; internalization of domination by the dominated. This functioning is at work in masculine domination, a subject to which Bourdieu (1998) devoted a book.

Based on his fieldwork in Kabylie, Bourdieu shows that the masculine/feminine opposition is one of the fundamental principles of the world's di-vision, which organizes the mythical-cosmological categories of perception of space and time: high/low, above/below, right/left, outside/inside, discontinuous/continuous, extraordinary/ordinary, event/duration. Underlying these are the main oppositions that organize social life: culture/nature, public/private, active/passive. Far from being the expression of biological differences, masculine and feminine identities are socially constructed identities (gender) instilled in individuals via bodily training. For boys, expressions of virility, honor, and domination are taught; instilled in girls are expressions of femininity, effacement, and submission (discretion, restraint, lowered eyes, making oneself small, etc.). Biological differences serve to justify the sexual division of work that results from these gendered identities. As such, socially constructed identities are naturalized, and cultural arbitrariness takes on natural appearances. In contemporary societies, these identities are reproduced not only by familial education, but also by institutions – first of which are the Church, the State, and School. In posing the exchange of women as the basis for communication between men, Lévi-Strauss obscured the political dimension of this foundational act of social life, in which violence toward women, turned into objects, is institutionalized. This approach makes it possible to overcome the alternative between coercion and consent that often underlies debates on masculine domination. Symbolic violence implies complicit submission on the part of the dominated, which is not the result of a free, voluntary act, but rather results from the internalization of an androcentric world vision. Nothing reveals these mechanisms better than the logic of self-exclusion of women from the public sphere. Access to education and birth control may have given women the means to independence, but the presence of women in the working world has not sufficed to abolish inequality between men and women, an inequality that results by and large from gendered representations. In equal-level positions, women are often less well paid than men. In the same way, though technical progress may have liberated women from some household chores, the emergence of working women has not sufficed to challenge the sexual division of domestic tasks. In an appendix to the book on masculine domination, Bourdieu published a text in support of the gay and lesbian movement, which is part of the sociologist's engagement as a 'public intellectual'

Reflexivity and Engagement: A Global Intellectual

The last part of Bourdieu's life was marked by growing international recognition and political commitment. Until the 1980s, the international reception of Bourdieu's work, though not

insignificant, had remained limited to the sociology of education, the sociology of art and culture (Bourdieu's work sometimes helped these subdisciplines to emerge, for example in the United States) and anthropology (for the theory of practice). The reception of Bourdieu's work became unified after the English translation of *La Distinction*, published by Harvard University Press in 1984. The publication was met with a loud echo, raising Bourdieu to the level of a theoretician of the social world. In the context of globalization, such a reception led Bourdieu to reflect upon the international circulation of ideas and the conditions of comparativism, which requires, he argues, a comparison from system to system of the principles of differentiation of types of capital (Bourdieu, 1994). This comparative approach also stimulates reflexivity with regards to our own societies.

Increasingly solicited during the 1990s, Bourdieu intervened more and more on current issues. In 1993, the same year he received the gold medal from the CNRS, France's main research organization (and he was the first sociologist to receive it), Bourdieu published *La Misère du monde*, a collection of interviews, realized by a team of researchers, that gave voice to those without a right to public speech. The book met with great public success. It established a new mode of collective political intervention based on scientific work, which was fully theorized in 1995, when Bourdieu took a stand in favor of a large social movement in France, and slightly later, upon the creation of the Raisons d'agir collective and the Liber-Raisons d'agir publishing house.

First in the series was Bourdieu's book *Sur la télévision* (1996a), a critique of the media that set down the bases for an analysis of the field of journalism. Translated into 25 languages, the slim volume sparked a violent campaign against the sociologist in France. From 1995 on, Bourdieu rallied against neo-liberalism (Bourdieu, 1998, 2001a, 2010). Relating his political commitments to a more developed scientific reflection, he published *Les Structures sociales de l'économie* (2002) at the same time, a work showing the necessity of articulating economic analysis together with social analysis, and ending with a description of the process of globalization in the economic field. Bourdieu was also concerned with the effects of growing concentration in the publishing industry with the merging-acquisition process, which challenged the autonomy of intellectual life by reinforcing the logics of profitability: this is why he dedicated his last empirical study to the French publishing field (Bourdieu, 1999). Amplified by the wide media reception of *Sur la télévision*, Bourdieu's international notoriety guaranteed that his engagement would have a worldwide impact, turning him into a true 'global intellectual'.

In the tradition of French intellectuals since Voltaire, Zola, Sartre, and Foucault, Bourdieu put his symbolic capital to the service of political causes in the name of universal principles. But unlike Sartre, who embodied the criticized figure of the 'total intellectual', Bourdieu acted as a 'specific intellectual', in Foucault's term, that is, he limited his interventions to his areas of expertise as a sociologist. Furthermore, he went beyond Foucault in forging the notion of the 'collective intellectual', which takes the division of scientific labor into consideration.

Bourdieu's final course at the Collège de France dealt with science (*Science de la science et réflexivité*, 2001d). Is it possible to account sociologically for the way the scientific field functions without tipping over into relativism? This was Bourdieu's guiding question for his reflexive approach to the practice of research. Proposing to study a 'science in the making' (*science se faisant*) instead of a 'ready-made science' (*science toute faite*), Bourdieu saw in the scientific field – and not in individual scholars – the 'subject' of science. By the orchestration of competitive struggles according to specific rules, the scientific field confers upon science its validity, despite the social determinants and power relations that introduce heteronomous factors. Against the drifting of the postmodern relativists, Bourdieu therefore affirmed a 'historicist rationalism' that refused to rely on an enchanted vision of the scientific world. In his final course, he engaged in an exercise of self-analysis, attempting to explain from a sociological point of view his own trajectory based on his 'cleft habitus' He intended to develop this reflection in a book, but was not able to bring it to term before his premature death on 23 January 2002; the unfinished version was published under the title *Esquisse pour une autoanalyse* (2004). Since his death, the interest in his theory and research program has increased in the world, and he has already become a classical author.

See also: Durkheim, Emile (1858–1917); Field Theory; Foucault, Michel (1926–84); Gender: Gendering of Categories; Goffman, Erving (1922–1982); Habitus: History of a Concept; Levi-Strauss, Claude (1908–2009); Life-style, History of the Concept; Marx, Karl (1818–83); Relativism: Philosophical Aspects; Social Inequality and Schooling; Social Inequality in History (Stratification and Classes); Social Structure: History of the Concept; State, History of; Weber, Max (1864–1920); Wittgenstein, Ludwig (1889–1951).

Bibliography

Works by Bourdieu (Selection)

Bourdieu, P., 1963. Travail et travailleurs en Algérie. Paris-La Haye, Mouton.

Bourdieu, P., 1971a. Genèse et structure du champ religieux. Revue française de Sociologie XII, 295–334. English trans.: 1991. Genesis and structure of the religious field. In: Calhoun, C. (Ed.), Comparative Social Research (13), 1–44.

Bourdieu, P., 1971b. Une interprétation de la théorie de la religion selon Max Weber. Archives européennes de sociologie XII (1), 3–21. English trans. (revised version): 1987. Legitimation and structure interests in Weber's sociology or religion. In: Whimster, S., Lash, S. (Eds.), Max Weber, Rationality and Modernity. Allen & Unwin, London, pp. 119–136.

Bourdieu, P., 1971c. L'opinion publique n'existe pas. Les Temps modernes (318), 1292–1309. English trans.: 1979. Public opinion does not exist. Communication and Class Struggle 1, 124–130.

Bourdieu, P., 1972. Esquisse d'une théorie de la pratique. Minuit, Paris. In: Paperback (Ed.), 2000. Seuil 'Points', Paris. English trans.: 1977. Outline of a Theory of a Practice (R. Nice, Trans.). Polity Press, Cambridge.

Bourdieu, P., 1976. Le champ scientifique, Actes de la recherche en sciences sociales (2–3), 88–104.

Bourdieu, P., 1977a. Algérie 60. Structures économiques et structures temporelles. Minuit, Paris. English trans.: 1979. Algeria 1960 (R. Nice, Trans.). Polity Press-Éditions de la Maison des sciences de l'homme, Cambridge, Paris.

Bourdieu, P., 1977b. Sur le pouvoir symbolique. Annales Économies, Sociétés, Civilisations 32 (3), 405–441.

Bourdieu, P., 1977c. Questions de politique. Actes de la recherche en sciences sociales (16), 55–89.

Bourdieu, P., 1979. La Distinction. Critique sociale du jugement. Minuit, Paris. English trans.: 1984. Distinction. A Social Critique of the Judgement of Taste (R. Nice, Trans.). Harvard University Press, Cambridge (Massachusetts).

Bourdieu, P., 1980. Le Sens pratique. Minuit, Paris. English trans.: 1990. The Logic of Practice (R. Nice, Trans.). Polity Press, Cambridge.
Bourdieu, P., 1981a. La représentation politique. Actes de la recherche en sciences sociales (36–37), 3–24. English trans.: 1991. Political Representation: Elements for a Theory of the Political Field, in Language & Symbolic Power. Polity Press, Cambridge, pp. 171–202.
Bourdieu, P., 1981b. Décrire et prescrire. Note sur les conditions de possibilité et les limites de l'efficacité politique. Actes de la recherche en sciences sociales (38), 69–73. English trans.: 1991. Description & Prescription: The Conditions of Possibility & The limits of Political Effectiveness, in Language & Symbolic Power. Polity Press, Cambridge, pp. 127–136.
Bourdieu, P., 1982a. Ce que parler veut dire. L'économie des échanges linguistiques. Fayard, Paris. English trans.: 1991. Extracts published in Language & Symbolic Power. Polity Press, Cambridge.
Bourdieu, P., 1982b. Leçon sur la leçon. Minuit, Paris. English trans.: 1990. Leçon sur la leçon. In: In Others Words. Essays Towards a Reflexive Sociology. Polity Press, Cambridge, pp. 177–198.
Bourdieu, P., 1984. Questions de sociologie, first ed. 1984. Minuit, Paris. English trans.: 1993. Sociology in Questions (R. Nice, Trans.). Sage Publications, London.
Bourdieu, P., 1987. Choses dites. Minuit, Paris. English trans.: 1990. In Others Words. Essays Towards a Reflexive Sociology. Polity Press, Cambridge.
Bourdieu, P., 1988. L'Ontologie politique de Martin Heidegger. Minuit, Paris. English trans.: 1991. The Political Ontology of Martin Heidegger (P. Collier, Trans.). Polity Press, Cambridge.
Bourdieu, P., 1989. La Noblesse d'État. Grandes écoles et esprit de corps. Minuit, Paris. English trans.: 1996. The State Nobility, Elite Schools in the Field of Power (L.C. Clough, Trans.). Polity Press, Cambridge.
Bourdieu, P., 1992a. Homo academicus, first ed. 1984. Minuit, Paris. English trans.: 1988. Homo academicus (P. Collier, Trans.). Polity Press, Cambridge.
Bourdieu, P., 1992b. Les Règles de l'art. Genèse et structure du champ littéraire. Seuil, Paris. In: Paperback (Ed.), 1998. Seuil 'Points', Paris. English trans.: 1996. The Rules of Art (S. Emanueltrans, Trans.). Polity Press-Stanford University Press, Cambridge-Stanford.
Bourdieu, P., 1993a. La Misère du monde. Seuil, Paris. English trans.: 1999. The Weight of the World. Social Suffering in Contemporary Society (P. Parkhurst Ferguson, S. Emanuel, J. Johnson, S.T. Waryn, Trans.). Polity Press, Cambridge.
Bourdieu, P., 1993b. The Field of Cultural Production. Edited and Introduced by Randal Johnson. Polity Press, Cambridge.
Bourdieu, P., 1994. Raisons pratiques. Sur la théorie de l'action. Seuil, Paris. In: Paperback (Ed.), 1996. Seuil 'Points', Paris. English trans.: 1998. Practical Reasons. On the Theory of Action (R. Johnson, Trans.). Polity Press, Cambridge.
Bourdieu, P., 1996a. Sur la télévision. Liber-Raisons d'agir, Paris. English trans.: 1998. On Television & Journalism (P. Parkhurst Fergusontrans, Trans.). The New Press/Pluto Press, New York/London.
Bourdieu, P., 1996b. Contre-feux. Liber-Raisons d'agir, Paris. English trans.: 1998. Act of Resistance. Against the Tyranny of Market. Polity Press, Cambridge; 1999. Act of Resistance. Against the Tyranny of Market. The New Press, New York.
Bourdieu, P., 1997a. Méditations pascaliennes. Seuil, Paris. In: Paperback (Ed.), 2003. Seuil 'Points', Paris. English trans.: 2000, Pascalian Meditations (R. Nice, Trans.). Polity Press, Cambridge.
Bourdieu, P., 1997b. Les usages sociaux de la science. INRA, Paris.
Bourdieu, P., 1998. La Domination masculine. Seuil, Paris. English trans.: 2001. Masculine Domination (R. Nice, Trans.). Polity Press-Stanford University Press, Cambridge-Stanford.
Bourdieu, P., 1999. Une révolution conservatrice dans l'édition. Actes de la recherche en sciences sociales (126/127), 3–28. English trans.: 2008. A conservative revolution in publishing (Eng. transl.). Translation Studies 1 (2), 123–153.
Bourdieu, P., 2000a. Propos sur le champ politique. Presses universitaires de Lyon, Lyon.
Bourdieu, P., 2000b. Les structures sociales de l'économie. Seuil, Paris. English trans.: 2005. The Social Structures of the Economy (C. Turner, Trans.). Polity Press, Cambridge.
Bourdieu, P., 2001a. Sociologie de l'Algérie, first ed. 1958. Que sais-je? PUF, Paris. English trans.: 1962. The Algerians (A.C.M. Ross, Trans.), preface Aron, A. Beacon Press, Boston.
Bourdieu, P., 2001b. Contre-feux 2. Pour un mouvement social européen. Liber-Raisons d'agir, Paris. English trans.: 2003. Firing Back. Act of Resistance. Against the Tyranny of Market 2 (L. Wacquant, Trans.). The New Press, New York.
Bourdieu, P., 2001c. Langage et pouvoir symbolique. Seuil, Paris. English trans.: 1991. Language & symbolic power. In: Thompson, J.B. (Ed.), (M. Adamson, G. Raymond, Trans.). Polity Press, Cambridge.
Bourdieu, P., 2001d. Science de la science et réflexivité. Liber-Raisons d'agir, Paris. English trans.: 2004. Science of Science and Reflexivity (R. Nice, Trans.). Polity Press, Cambridge.
Bourdieu, P., 2002a. Le Bal des célibataires. Crise de la société en Béarn. Seuil 'Points', Paris. English trans.: 2008. The Bachelor's Ball (R. Nice, Trans.). Polity Press, Cambridge.
Bourdieu, P., 2002b. In: Poupeau, F., Discepolo, T. (Eds.), Interventions. 1961–2001. Science sociale et action politique. Agone, Paris. English trans.: 2008. Political Interventions: Social Science and Political Action (David Fernbach, Trans.). Verso, London.
Bourdieu, P., 2004. Esquisse pour une auto-analyse. Liber-Raisons d'agir, Paris. English trans.: 2008. Sketch for a Self-Analysis (R. Nice, Trans.). Polity Press, Cambridge.
Bourdieu, P., 2008. Esquisses algériennes. Seuil, Paris. English trans.: 2013. Algerian Sketches. Polity Press, Cambridge.
Bourdieu, P., 2010. Sociology Is a Martial Art. Political Writings by Pierre Bourdieu. edited and introduced by Gisèle Sapiro. The New Press, New York.
Bourdieu, P., 2012. Sur l'État. Cours au Collège de France (1989–1992). Seuil, Paris. English trans.: 2014. On the State. Lectures at the College de France (1989–1992). Polity Press, Cambridge.
Bourdieu, P., 2013a. Séminaires sur le concept de champ, 1972–1975. Actes de la recherche en sciences sociales (200), 4–37.
Bourdieu, P., 2013b. Manet. Une révolution symbolique. Seuil-Raisons d'agir, Paris.
Bourdieu, P., Haacke, H., 1994. Libre-échange. Seuil-Les presses du réel, Seuil. English trans., Minuit, Paris, 1995. Free Exchange. Polity Press-Blackwell Publishers, Cambridge-Oxford.
Bourdieu, P., Passeron, J.C., 1966. Les Héritiers. Les étudiants et la culture, first ed. 1964. Minuit, Paris. English trans.: 1979. The Inheritors, French Students & their Relation to Culture (R. Nice, Trans.). The University of Chicago Press, Chicago-London.
Bourdieu, P., Passeron, J.C., 1989. La Reproduction. Éléments pour une théorie du système d'enseignement, first ed. 1970. Minuit, Paris. English trans.: 1977. Reproduction in Education, Society & Culture (R. Nice, Trans.). Sage Publications, London-Beverly Hills.
Bourdieu, P., Sayad, A., 1964. Le Déracinement. La crise de l'agriculture traditionnelle en Algérie. Minuit, Paris. English trans.: 2014. Uprooting. Polity Press, Cambridge.
Bourdieu, P., Wacquant, L., 1992. Réponses. Pour une anthropologie réflexive. Seuil, Paris. English trans.: 1992. An Invitation for a Reflexive Sociology (L. Wacquant, Trans.). University of Chicago Press-Polity Press, Chicago-Cambridge.
Bourdieu, P., Boltanski, L., Castel, R., Chamboredon, J.C., 1970. Un Art moyen. Essai sur les usages sociaux de la photographie, first ed. 1965. Minuit, Paris. English trans.: 1990. Photography. A Middle Brow Art (S. Whiteside, Trans.). Polity Press, Cambridge.
Bourdieu, P., Chamboredon, J.C., Passeron, J.C., 2005. Le Métier de sociologue, first ed. 1968. Éditions de l'EHESS, Paris. English trans.: 1991. The Craft of Sociology (R. Nice, Trans.). Walter de Gruyter, New York.
Bourdieu, P., Darbel, A., Schnapper, D., 1966. L'Amour de l'art. Les musées et leur public. Minuit, Paris. English trans.: 1990. The Love of Art. European Art Museums and Their Public (C. Beattie, N. Nerriman, Trans.). Polity Press, Cambridge.

Works on Bourdieu (Selection)

Bouveresse, J., Roche, D. (Eds.), 2004. La Liberté par la connaissance. Pierre Bourdieu (1930–2002). Odile Jacob, Paris.
Calhoun, C., LiPuma, E., Postone, M., 1993. Bourdieu: Critical Perspectives. The University of Chicago Press, Chicago.
Gorski, Ph L. (Ed.), 2013. Bourdieu and Historical Analysis. Duke University Press, Durham.
Pinto, L., Sapiro, G., Champagne, P. (Eds.), 2004. Pierre Bourdieu, Sociologue. Fayard, Paris.
Shusterman, R., 1999. Bourdieu: A Critical Reader. Blackwell Publishers, Oxford-Malden.
Swartz, D., 1997. Culture & Power: The Sociology of Pierre Bourdieu. University of Chicago Press, Chicago.
Wacquant, L. (Ed.), 2005. Bourdieu & Democratic Politics: The Mystery of Ministry. Polity Press, Cambridge.

Relevant Websites

http://www.arss.fr/.
http://www.univ-paris1.fr/centres-de-recherche/cessp/.

Bourgeoisie and Middle Classes, History of

Hannes Siegrist, University of Leipzig, Leipzig, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 2, pp.1307–1314, © 2001, Elsevier Ltd.

Abstract

The history of the bourgeoisie and the middle classes concerns the development and the transformation of social formations that are made up of those occupying intermediary social positions. These social formations and their members take on mediating or leading functions in society, economics, politics, and culture. The historiography of the bourgeoisie and middle classes takes as its point of departure the development of the modern bourgeoisie out of the older urban bourgeoisie (burghers) of Europe. It concentrates on Western and Central Europe and on the USA, the history of which has, from the eighteenth through the twentieth centuries, been shaped to a large degree by the bourgeoisie and the middle classes. It asks how the functions and meaning of the 'bourgeoisie' and the 'middle classes' have changed in these core areas, how, since the late nineteenth century, the conception has diffused to the periphery of Europe and worldwide, and how, in the process, it has transformed and run up against certain limits. Finally, it analyses the attempts to eliminate or transform the bourgeoisie in socialist countries, and the reconstruction of modern middle classes in postsocialist societies.

Introduction

The history of the bourgeoisie and the middle classes concerns the development and the transformation of social formations that are made up of those occupying intermediary social positions. These social formations and their members take on mediating or leading functions in society, economics, politics, and culture. The historiography of the bourgeoisie and middle classes takes as its point of departure the development of the modern bourgeoisie out of the older urban bourgeoisie (burghers) of Europe. It concentrates on Western and Central Europe and on the USA, the history of which has, from the eighteenth through the twentieth centuries, been shaped to a large degree by the bourgeoisie and the middle classes (Kocka 1995a,b, Zunz et al., 2001). It asks how the functions and meaning of the 'bourgeoisie' and the 'middle classes' have changed in these core areas, how, since the late nineteenth century, the conception has diffused to the periphery of Europe and worldwide, and how, in the process, it has transformed and run up against certain limits. Finally, it analyzes the attempts to eliminate or transform the bourgeoisie in socialist countries and the reconstruction of modern middle classes in post socialist societies (Hoepken and Sundhaussen, 1998).

Historical research on the middle classes investigates symbolic and social structures of long duration, historical continuities and ruptures, and regional and national similarities and differences. Of special interest are the particular characteristics of different formations and actors in various historical contexts. Research topics include the social, economic, professional, political, cultural, and legal characteristics and structures of the middle strata, whether in hierarchically differentiated societies, under conditions of social and cultural inequality, or in processes of class formation. Historians are also concerned with symbolically mediated practices and processes, in which individuals and groups come together in historically identifiable middle classes such as the traditional 'middle estates,' the exclusive 'modern bourgeoisie,' the economic and political 'bourgeoisie,' and the 'broad and inclusive middle classes.' Common to both historical and systematic concepts of the bourgeoisie and the middle classes, however diverse they are otherwise, is the preoccupation with the central region of the society in question, that is with that which takes shape between the extremities. The fundamental assumption is that society is formed and integrated not 'from above' (aristocracy, upper class, oligarchy), or 'from below' (working class, the propertyless, the uneducated), or 'from outside' (foreign domination).

Definitions

The concepts 'bourgeoisie,' 'middle class,' 'middle classes,' and many similar terms applying to particular eras and cultures, have more or less in common, depending on the historical period, the society in question, and the historical situation. A typical characteristic of the historical scholarship on these topics is the tension between exclusive and inclusive concepts of the bourgeoisie and the middle classes.

Definitions vary with regard to the composition of the bourgeoisie, the criteria for inclusion and exclusion, the social and symbolic processes of integration and delimitation, and with regard to the functions and aspirations of the middle classes in society and history. In the fields of social and economic history, scholars may employ classifications that are descriptive, historically and culturally specific, and statistical; or they may, for heuristic purposes, adopt abstract definitions, which are derived from particular theories of the bourgeoisie or from general theories of society. In either case, the focus is on the correspondence between social position and interests. In contrast, cultural historians and culturally oriented social historians understand the bourgeoisie and the middle classes primarily in terms of images of self and other, strategies, performances, discourses, rituals, institutions, sociocultural practices, habitus, and mentalities. In the history of ideas and of philosophy, as well as in theories of the state, law, and constitutions, scholars usually rely on normative and systematized conceptions of the bourgeoisie and middle classes. Social historians also draw on *Begriffsgeschichte*, the history of

ideas, and the history of mentalities in order to show how, under what conditions, and why the concepts 'middle class,' 'bourgeoisie,' and 'middle classes' have been employed for the self-definition of individuals and groups, or for the purposes of establishing boundaries between self and other (Kocka 1995a, p. 12; Koselleck et al., 1991). Of special interest, in this regard, is how such classificatory activity affects consciousness and social action. A central element in all definitions and research agendas is the question of exclusion and inclusion.

Exclusive Formations: Middle Class and Bourgeoisie

Exclusive formations of the middle classes include wealthy property owners, entrepreneurs, those with high salaries, the highly educated and highly qualified, as well as administrative, political, and cultural elites. These groups are characterized by similar economic and political situations. They are also linked through social networks and have similar cultural orientations and life-styles. Historical research has been directed toward eliciting the meaning of the concepts and categories just cited in concrete historical situations, and discovering the historically and regionally specific terms that allow them to be grouped together in a single formation. In England and the USA, this social phenomenon has, since the eighteenth and nineteenth centuries, been referred to as the 'upper middle class' or 'middle class' (singular). In francophone Europe, the relevant terms are *notables*, *bourgeoisie*, or *classe moyenne* (singular). The German-speaking peoples of Europe have employed the terms *Bürgerstand*, *Mittelstand*, or *Bürgertum*. The corresponding labels in Italy are *borghesia* or *ceti medi*, and in Spain *burguesía* (Kocka, 1995a; Zunz et al., 2001). In the nineteenth century, the bourgeoisie in this narrow sense made up approximately 5% of the population, although this percentage was somewhat higher in cities.

In contemporary scholarship, these formations are understood in terms of Max Weber's sociology of classes and status groups, and are referred to using the historical-systematic concept of the 'modern bourgeoisie' or 'middle class' (Kocka, 1995a; Zunz et al., 2001). Understood in this way, the bourgeoisie is a heterogeneous formation, which includes economically independent and dependent classes (both entrepreneurs and civil servants, both professionals and employees), the members of which have similar values and styles of life. The concept of the 'modern bourgeoisie' differs from the Marxian concept of the 'economic bourgeoisie.' The latter is defined in terms of the ownership of capital, access to the means of production, kinds of revenue, and strategies of capital deployment and commercialization. The economic bourgeoisie, which is the counterpart of the working class, is the central category of political-economic development theory. Historical scholarship that has been influenced by Karl Marx and Marxian theory distinguishes among the precapitalist bourgeoisie (speculative capital, merchant capital, usury) those land owners and agriculturalists who have a dynamic relationship to property and capital, the petite bourgeoisie of artisans, the bureaucratic bourgeoisie, the modern commercial and industrial bourgeoisie, and the monopolist bourgeoisie (Radandt et al., 1981). Since the late nineteenth century, the Marxian conception of the bourgeoisie has, in many variations, become established in scholarship and in public discussion. From the late nineteenth century and up until the 1980s, Marxian theory was also very influential in non-Marxian sociology and social theory as well.

More recently, the sociological and economic conceptions of the middle classes and the bourgeoisie have been challenged by cultural historians, who insist upon the central importance of phenomena which, in classical Marxian theory, are part of the superstructure. From a cultural historical perspective, membership in the middle classes depends primarily on particular ideas, values, attitudes, and sociocultural practices. Emphasis falls on the symbolic mediation of that which is bourgeois or middle class through discourses, images, institutions, and places of socialization and sociability such as the family, the school, the public sphere, the theater, literature, voluntary associations, coffee houses, festivals, architecture, and monuments. Research of this type leads to further important insights into the bourgeoisie. If, however, bourgeois culture is not linked back to the social groups that foster it, this kind of approach may become merely a socially diffuse cultural history of modernity.

Inclusive Middle Classes

The history of the exclusive middle class has been shaped by the need to establish distinct boundaries with reference to traditional elites, the working class, and the 'inclusive' or 'broad' middle classes. The 'inclusive middle classes' may seem to be less homogeneous than the exclusive bourgeoisie in terms of profession, status, education, income, style of life, attitudes, and values. Nevertheless, from the viewpoint of both historical actors and historians, they have often been thought to belong together. Their unity is often characterized with qualifying expressions such as 'finely graduated' or 'leveled.'

The factors that unite the inclusive middle classes vary from one land, historical period, and social situation to another. These common factors may include professional and economic interests, similar legal claims, traditions and mentalities, political viewpoints and voting behavior, similar hopes and fears with regard to upward or downward mobility, or attitudes with regard to style of life, which may in turn be oriented toward those of the exclusive bourgeoisie. The social spectrum of the inclusive middle classes extends from mid-level entrepreneurs to independent professionals, mid-level and low-level civil servants and employees, mid-level farmers and tenants, and even to workers with higher salaries and qualifications. When social, professional, and economic criteria are used to assess the size of this group in the Western and Central European countries of the nineteenth century, it may be estimated to have made up 15–30% of the total population. When other criteria are employed, this percentage may rise or sink.

In the USA and in France, the inclusive middle class is referred to as 'middle classes' or *classes moyennes* (plural), respectively. The corresponding term in Germany is *Mittelstand*, and in Switzerland *Mittelstand* or *Bürgertum* in the broadest sense of the term. In the German-speaking countries, the term *Volk* was also used in this sense, especially in the socially and politically most inclusive societies of the early twentieth century (Kocka, 1995a; Tanner, 1995; Savage, 2001; Charle, 2001). In the second half of the twentieth century, the concept of the inclusive middle classes has been extended even further.

As it gains in influence and in integrative potential, the older concept of the exclusive middle class seems to be falling increasingly into disuse.

It is common for both historical actors and historians to distinguish between historically older and newer groups of the broad middle classes. Master artisans and small businessmen are referred to as the *alter Mittelstand* (old middle estate) or the 'old middle class' (Haupt and Crossick, 1998). From the early nineteenth century and up to the 1970s, they have often been viewed as a historical relict and as potential victims of modernization. Employees i.e., members of the strata of white-collar workers, which has been expanding since the late nineteenth century and which continues to become increasingly differentiated, are referred to as the 'new middle class.' They are often thought to have developed a distinctive class consciousness at a relatively late date, due especially to their intermediate position between the bourgeoisie and the working class. In different societies and in different historical periods, employees may be understood either as agents or as victims of modernization (Mills, 1956; Kocka, 1977).

History of the Bourgeoisie and the Middle Classes

Recent scholarship on the eighteenth and nineteenth centuries has concentrated on the conditions and forms of the genesis of the modern exclusive bourgeoisie. The historiographical literature on the twentieth century focuses especially on the transformation, the crisis, and the declining significance ofthe exclusive bourgeoisie. With reference to the development of the modern bourgeoisie in Central and Northwestern Europe, Jürgen Kocka distinguishes between three phases, including an ascendant phase from the eighteenth to the early nineteenth century, a culminating phase from the middle of the nineteenth century to World War I, and a period of differentiation, dissolution, and even renaissance, which lasts into the late twentieth century (Kocka, 1995a,c). Attempts to assign the history of the broadly inclusive middle classes to historical periods usually posit a period of decline in the nineteenth century and a new period of ascendance in the twentieth century, which culminates in the inclusive middle-class society. In the following summary of the history of the bourgeoisie and the middle classes, reference is made to periods of 'genesis,' 'expansion,' and 'contraction.'

The Genesis of the Modern Bourgeoisie and Middle Classes (Eighteenth to Nineteenth Centuries)

The modern bourgeoisie took shape in the eighteenth and nineteenth centuries, though the precise time varied by region and by nation (Kocka, 1995b; Schrader, 1996; Gall, 1993; Charle, 2001). It included the following social groups: (1) the more dynamic part of the older urban bourgeoisie of the cities and communes; (2) those civil servants of the modern territorial state who were not part of the feudal hierarchy; (3) merchants, bankers, industrial entrepreneurs, and those members of the independent artisanry who were able to remove themselves from the order of the corporative estates; (4) professionals, literati, and artists; (5) enlightened or economically innovative portions of the landholding nobility, as well as non-noble estate owners and agriculturalists (in Southern and Western Europe the landowners often lived in cities, which is why they were counted among the bourgeoisie); (6) participants in emancipatory movements that were cultural or confessional, liberal or democratic, or regional or national in character.

Communication, symbolic representation, and social integration among the ranks of the bourgeoisie were promoted through institutional changes, which were implemented partly by the bourgeoisie itself, and partly by the monarchs and the state in the context of reforms and revolutions (freedom of trade, religious freedom, freedom of association, freedom of speech and of the press, political participation and accountability). The bourgeoisie developed its selfunderstanding and its language in the fields of the economy, culture, politics, and law. Bourgeois praxis developed in the representative organs of the estates, in free associations, in scientific academies, in literary circles and salons, in the print media, and in the reformed institutions of higher education, administration, law, legislation, and government. In these contexts, 'bourgeois' notions of collective and individual autonomy, representation, and the free play of public opinion were formed. In both the workplace and in public debates, new understandings of property and interest, ability and performance, self-cultivation and virtue, individual and society, contract and constitutions, and tradition and progress took shape. New experiences and discourses allowed the emerging middle class to criticize existing conditions, conceive of new social and symbolic orders, and lay claim to leading or mediating positions as well as to central functions and competences in society.

The bourgeoisie improved its chances for gaining favorable positions in a system of material, social, political, and cultural inequality by generalizing its own understandings of law and removing traditional privileges. It distanced itself from the traditional 'upper estates' of the court, the landed nobility, and the urban patriciate by adopting the role of a modern, dynamic, and achievement-oriented cultural and functional elite. It constituted itself in opposition to the nobility and the clergy as a 'third estate,' which represented the general public and was, in principle, open to everyone. The 'bourgeoisie' and the 'middle class' are concepts that imply a program and that serve to mobilize and justify certain forms of social action in the transition from estate society to class society. In principle, they ignore traditional differences of status, profession, and confession.

Typically, however, new forms of differentiation and new tensions arise fairly rapidly within the bourgeoisie, for example between the economic bourgeoisie (*Wirtschaftsbürgertum*) and the educated bourgeoisie (*Bildungsbürgertum*),between professional groups and status groups, between rural and urban groups, along local, regional, and national lines, among different confessions, or between political traditionalists and progressives or liberals and democrats. The integration of the bourgeoisie is a complicated and lengthy process, which usually succeeds only partially. The unity of the bourgeoisie is more a postulate than a reality (Banti, 1996; Siegrist, 1996). It is most evident in critical historical situations or in conflicts with others, to whom the bourgeoisie feels itself to be superior, whether morally, economically, politically, or in terms of

performance (nobility, manual laborers, the propertyless, the uneducated, the disenfranchised, and strangers or foreigners).

Contraction and Expansion: Bourgeosie and Middle Classes in the Nineteenth and Twentieth Centuries

In the process of the formation of the modern bourgeoisie, some groups of artisans and merchants, who previously belonged to the inclusive urban bourgeoisie, were economically marginalized and socially downgraded. In contrast, the educated bourgeoisie, merchants, and entrepreneurs belonging to ethnic and confessional minorities, and rural industrialists acquired new rights and gained in status. In England, the concept of the 'middle class' served to mobilize business people, industrialists, and mine owners from the provinces in the context of the electoral reform movement of 1832. It helped to integrate in a single group all of those who had not had the right to vote for the Lower House and who had felt themselves to be neglected by the upper middle class (London wholesalers and financiers, aristocratic landowners). Since the 1830s, in England as on the European continent, the interconnection among political, social, and legal rights possessed by the bourgeoisie had begun to emerge (Koselleck et al., 1991). This bundle of rights was viewed by bourgeois philosophers and social theorists as the very basis of political virtue in a well ordered society. In the course of the nineteenth century, in Europe and in the USA, perceptions of the 'bourgeois' or 'middle-class' male person became closely linked with the concept of the politically enfranchised 'citizen' (*citoyen*, *Staatsbürger*). In those parts of Central and Eastern Europe in which the political rights of citizens remained limited, property owners and the educated often shared political power at the local level. The towns and cities, in which various property-based or tax-based voting laws remained in effect until World War I, were, in many parts of Europe, the domain of the bourgeoisie. In Central Europe, the concept of the bourgeoisie lost much of its political content after the failure of the 'bourgeois revolutions' of 1848/49. Under these conditions, emphasis shifted to the cultural and aesthetic connotations of the concept. Throughout Europe, the bourgeoisie experienced significant gains in processes of 'inner' nation-building and cultural nationalization in the old nation-states, and through nationalist movements and the formation of new nation-states in Northern, Southern, and Eastern Europe.

Since the late nineteenth century, the bourgeoisie has continually renewed itself by incorporating members of new and expanded professional, functional, or status groups such as managers, engineers, planning and transportation experts, journalists, and high-level employees. These groups proliferated under conditions of intensified industrialization, commercialization, and urbanization or with the concentration of capital, the expansion of the role of science in society, increasing bureaucratization, and the rise of new communication media. The exclusive bourgeoisie allied itself increasingly, though not unreservedly, with social groups whose middle-class status was dubious, but who offered themselves as supporters of the established order. Tenants, agriculturalists with medium-sized holdings, and independent artisans and merchants were offered greater protection and drawn into the bourgeois alliance. Employees gained social and legal privileges over workers and were, thus, prevented from cooperating more closely with the socialist and trade-union organizations of the working class (Kocka, 1977).

At no time has the bourgeoisie been unchallenged. In the nineteenth century, the critique of the bourgeoisie was formulated from aristocratic, corporativist, Catholic, anti-Semitic, and socialist points of view. Since the late nineteenth century, the traditional, exclusive conception of the bourgeoisie has come increasingly under attack from members of social movements that arose within the bourgeoisie and the middle classes themselves. The leading concepts of the bourgeois progressives, social reformers, life-style reformers, and democrats are science, achievement, rationality, political equality, and social justice. Progressive and democratic critiques of the bourgeoisie contrast, however, with those of the authoritarian and conservative-elitist movements for renewal ofthe early twentieth century. The latter often attracted members of the bourgeoisie and petite bourgeoisie, even though they cultivated a specifically 'antibourgeois' discourse, in which the established and increasingly pragmatic bourgeoisie was accused of decadence and the inability to lead. In the twentieth century, antibourgeois sentiment formed a broad spectrum that ranged between conservative-corporativist and fascist programs, on the one hand, and communist ideology, on the other.

The concept of the exclusive bourgeoisie and antibourgeois discourses were supplemented in the twentieth century by conceptions of a broad and leveled-off middle class, which is typical of advanced industrial society. In the history of Europe and the USA, this perspective drew upon older ideas, such as those of social democratic, social-liberal, Christian-social, egalitarian, and communitarian traditions. After World War II, the USA made the concept of the inclusive middle class into a guiding principle for the reconstruction of the defeated states of Europe and Japan (Zunz et al., 2001). This concept unites socially, culturally, politically, and economically heterogeneous groups in a common vision of political and social citizenship in a democratic system and a market economy. During the era of the conflict between the East and the West, this vision was also motivated by a broadly based anticommunism i.e., by a rejection of a model of society, economy, and government that explicitly excludes the middle and claims to constitute society from below, under the leadership of the party of the working class.

In the noncommunist countries of Europe, in the first decades after 1945, those who wanted to conserve or renew the exclusive bourgeoisie of the past mounted an attack against the concept of the leveled and encompassing middle-class society. The exclusive bourgeoisie had lost prestige because of the Great Depression and due to the role that it had played in the fascist states. Now, however, its representatives mourned the waning and the demise of the bourgeoisie. Up until 1970 they warned of the decline of the bourgeoisie and the resulting threat posed to the nation and to Europe by 'mass culture,' 'consumer society,' 'Americanization,' 'materialism,' the 'welfare state,' the 'excesses' of democracy, the 'weakening of bourgeois values,' the 'loss of individuality,' and the 'leveled middle-class society' (Siegrist, 1994; Savage, 2001). The debate over the bourgeoisie experienced a new surge around 1968, which had, however, other causes, as the 'establishment,' on one hand, and students, intellectuals, socialists, and communists, on the others, battled

one another in the language of anti Marxism and Marxism. Since the 1970s, the discussion about the bourgeoisie has been superseded by debates over 'elites' and over traditional cultural, ethnic, and gender-based differences and discriminations within and outside the middle classes.

Middle Classes, Functional Equivalents, and the Limits of the Concept

The development of the middle classes in Eastern and Southeastern Europe, Latin America, Japan, and in the colonies and postcolonial societies of Africa and Asia differs considerably, both temporally and qualitatively, from that of core regions in the history of the middles classes. The concepts have their limits.

In many parts of the world in the nineteenth century, the bourgeoisie was hardly present. The social groups and the corresponding discourses, mentalities, and institutions were largely lacking. When concepts are imported and extended to phenomena that are more or less similar to the bourgeoisie, their meaning changes significantly (Kocka, 1995a,c). Marxian or economic and social historical studies, which, on the basis of theoretical premises and European concepts, identify bourgeois and middle-class groups in these areas, rarely do justice to their materials.

Recent approaches in the cross-national and cross-cultural comparative history of the middle classes focus on cultural and historical differences, and also on the specific meaning of similarities. Comparisons may show, for example, that institutional, economic, and technical innovations, which in Western and Central European history are associated with the bourgeoisie, are planned and carried out by other actors and groups in those areas without a bourgeoisie. In such cases, it is not very helpful to apply the concept of the bourgeoisie to such groups. Modernizing landowners in estate societies, civil servants in authoritarian states without a bourgeoisie, and rural and urban oligarchies and military leaders in preindustrial societies may be understood, at best, as 'functional equivalents' of the bourgeoisie. Latin American estate owners and the mercantile elite of large cities and ports may, in the course of the nineteenth century, have come to approximate the economic bourgeoisie. In societies that are sharply differentiated by status group and ethnicity, however, they form a political-social 'oligarchy,' rather than a 'modern bourgeoisie.' With the growth in the twentieth century of new agrarian, mercantile, bureaucratic, and industrial middle strata, the middle classes in such societies have expanded, and the traditional oligarchies have been forced to include them, at least to a degree, into the political elite (Florescano, 1985). Even then, however, the situation differs significantly from that in Europe.

In Eastern Europe in the nineteenth century, progressive groups of nobles, members of ethnic and confessional minorities, and Western-educated intellectuals and civil servants assimilated to the pattern of the bourgeoisie of Western and Central Europe. The formation of the bourgeoisie did not, however, proceed beyond the initial stages. In Southeastern Europe, modern middle classes began to emerge in the late nineteenth century, when, following the demise of the Ottoman empire, the segmented and not extremely hierarchical society changed into a national, centralized, and bureaucratic one. The national functional and cultural elites made the state into a resource and became a specific type of bourgeoisie (Höpken and Sundhaussen, 1998). In the communist societies of Eastern Europe, both genuinely and ostensibly bourgeois and middle-class groups were marginalized in the context of the communist program of the classless society under the leadership of the working class. Those who remained were, along with the new functionaries, redefined as members of the functional middle strata. The leading middle stratum was termed 'nomenclatura,' whereas the others were known as the 'service class' and the 'socialist intelligentsia.' Given this historical background, the problems involved in forming postsocialist middle classes are quite significant.

In twentieth-century Japan, where, previously, a historical bourgeoisie had hardly existed, there was a dramatic development of the inclusive middle classes. In the early twentieth century, the word *sarariiman* (salary-man), which at first meant scribe, office employee, and civil servant, became a catch-all term for members of the qualified, salaried middle classes (Gordon, 2001). Since the 1960s, the concept of *churyu* has become popular. This means 'mainstream' and includes all those who consider themselves to be modern, average, and normal with regard to attitudes toward work, family, consumption, and life-style. They make up the statistical and social 'middle mass' (Kelly, 2001), with which, in public opinion polls, 90% of the population aligns itself. The broad conception of the middle classes in Japan has its roots both in national ideas of community and in the inclusive view of the middle class that has been imported from the USA in the postwar period.

Historiography and Emphasis in Current Research

The exclusive bourgeoisie and (to a lesser degree) the broad middle classes have, from the very beginning, employed historical writing as a means of self-presentation, legitimation, community building, and self-reflection. History serves the bourgeoisie as a means of locating itself in space and time, and orienting itself in a modern, dynamic society, which it seeks constantly to transform and to stabilize. The historiography of the middle classes is characterized, first, by teleological notions of progress and modernization; second, by a narrative of crisis and decadence and of danger and dissolution, which draws upon older religious and aristocratic motifs but is subsequently transformed and takes on a life of its own; and, third, through the motifs of conservation and stabilization. Central elements of this history are elevated by the bourgeoisie and middle classes to the status of norms and, thereby, dehistoricized. They then take on the form of myths or theories of law, society, and development, which are insensitive to variations in era and context. The bourgeoisie displays a tendency toward idealizing, sacralizing, and naturalizing its own past. For this reason, its historiography has often been the target of historians of various ideological or social provenance, who are critical of or opposed to the bourgeoisie.

Until the 1970s, the history of the bourgeoisie and the middle classes was, by and large, written from the perspective

of either proponents or opponents. Since the 1980s, historical research on the bourgeoisie has, in some important senses, become more impartial, more independent, more scientific, and more innovative. Traditional versions of history at local and national levels have been relativized and revised with reference to interdisciplinary, international, and comparative approaches in historical research (Kocka 1995b; Siegrist, 1996). As a result of a stronger focus on the dimensions of 'gender' and 'religion' (Hall and Davidoff, 1987) as well as 'ethnicity,' new perspectives have been achieved with regard to the periodization and evaluation of historical materials. In Europe and, to a somewhat lesser degree, the USA, the bourgeoisie and the middle classes rank among the best researched topics in historical scholarship.

See also: Business History; Capitalism; Elites: Sociological Aspects; Industrialization, Typologies and History of; Marx, Karl (1818–83); Social Stratification; Weber, Max (1864–1920).

Bibliography

Banti, A.M., 1996. Storia della borghesia italiana. L'età liberate, Donizelli, Rome.

Charle, C., et al., 2001. The middle class in France: Social and political functions of semantic pluralism 1870–2000. In: Zunz, O (Ed.), Social Contracts Under Stress. Russell Sage, New York.

Davidoff, L., Hall, C., 1987. Family Fortunes: Men and Women of the English Middle Classes 1780–1850. Routledge, London.

Florescano, E. (Ed.), 1985. Origines y desarollo de la burguesía en America Latina. D. F., Mexico.

Gall, L. (Ed.), 1993. Stadt und Bürgertum im Übergang von der traditionalen zur modernen Gesellschaft Oldenbourg. Munich, Germany.

Gordon, A., et al., 2001. The short happy life of the Japanese middle class. In: Zunz, O., et al. (Eds.), Social Contracts Under Stress. Russell Sage, New York.

Haupt, H.-G., Crossick, G., 1998. Die Kleinbürger. Eine europäische Geschichte des 19. Jahrhunderts. C. H. Beck, Munich, Germany.

Hoepken, W., Sundhaussen, H. (Eds.), 1998. Eliten in Südosteuropa. Rolle, Kontinuitäten, Brüche in Geschichte und Gegenwart. Südosteuropa-Gesellschaft, Munich, Germany.

Kelly, W.W., et al., 2001. At the limits of new middle class Japan: Beyond mainstream consciousness. In: Zunz, O (Ed.), Social Contracts Under Stress. Russell Sage, New York.

Kocka, J., 1995a. Das europäische Muster und der deutsche Fall. In: Kocka, J. (Ed.), Bürgertum im 19. Jahrhundert, vol. 1. Vandenhoeck and Ruprecht, Göttingen, Germany.

Kocka, J. (Ed.), 1995b. Bürgertum im 19. Jahrhundert. Vandenhoeck and Ruprecht, Göttingen, Germany.

Kocka, J., 1995c. The middle classes in Europe. The Journal of Modern History 67, 783–806.

Kocka, J., 1977. Angestellte zwischen Faschismus und Demokratie. Zur politischen Sozialgeschichte der Angestellten. USA 1890–1940 im internationalen Vergleich. Vandenhoeck and Ruprecht, Göttingen, Germany.

Koselleck, R., Spree, U., Steinmetz, W., 1991. Drei bürgerliche Welten? Zur vergleichenden Semantik der bürgerlichen Gesellschaft in Deutschland, England und Frankreich. In: Puhle, H.-J. (Ed.), Bürger in der Gesellschaft der Neuzeit. Vandenhoeck and Ruprecht, Göttingen, Germany.

Mills, C.W., 1951. White Collar. The American Middle Classes. Oxford University Press, New York.

Radandt, H., et al., (Ed.), 1981. Deutscher Verlag der Wissenschaften, Berlin 692–700, 856–61, 968–76.

Savage, M., et al., 2001. Professional dominance and middle class culture in Britain. In: Zunz, O (Ed.), Social Contracts Under Stress. Russell Sage, New York.

Schrader, F., 1996. Die Formierung der bürgerlichen Gesellschaft 1550–1850. Fischer, Frankfurt, Germany.

Siegrist, H., 1994. Ende der Bürgerlichkeit? Die Kategorien 'Bürgertum' und 'Bürgerlichkeit' in der westdeutschen Gesellschaft und Geschichtswissenschaft der Nachkriegs periode. Geschichte und Gesellschaft 20, 549–583.

Siegrist, H., 1996. Advokat, Bürger und Staat. Sozialgeschichte der Rechtsanwälte in Deutschland, Italien und der Schweiz (18.–20.Jahrhundert). Klostermann, Frankfurt, Germany.

Tanner, A., 1995. Arbeitsame Patrioten—wohlanständige Damen: Bürgertum und Bürgerlichkeit in der Schweiz 1830–1914. OrellFüssli, Zurich, Switzerland.

Zunz, O., et al., 2001. Social Contracts Under Stress. Russell Sage, New York.

Bowlby, John (1907–90)

Joan Stevenson-Hinde, University of Cambridge, Cambridge, UK
Robert A Hinde, St. John's College, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The son of Sir Anthony and Lady Bowlby, John was the fourth born of six children. During daily life, both parents were rather distant, with child care given over to nursemaids and then boarding school at 11. Bowlby's later clinical observations and research indicated that separations could have lasting effects. He found an explanation not in the analytical thinking of the time, but in ethology. This enabled a view of attachments as an integral part of human nature, with attachment behavior selected for during the course of evolution along with its complement, caregiving behavior. When effective, attachment behavior leads to a feeling of security, and when ineffective to acute or chronic anxiety, as indeed happens with separations. The development of attachment theory and its implications are outlined.

Edward John Mostyn Bowlby was born in London on 26 February 1907 to Mary Bridget Mostyn and Anthony Alfred Bowlby. They had six children, who were usually referred to in groups of two: 'the girls' (Winnie and Marion), 'the boys' (Tony and John), and 'the babies' (Jim and Evelyn). Only 13 months apart, 'the boys' had a close relationship, including intense rivalry. But in addition to keeping his end up with a slightly older brother, John was caring toward a younger brother with chronic problems due to a thyroid deficiency, discovered too late to be cured. John was born at a time when it was customary for an upper-class mother to hand over her children to a nanny and nursemaids. The nanny in charge practiced a firm, regular, and disciplined routine. A particular nursemaid took care of John on a daily basis. He loved her, and years later she said that she was always fondest of John. However, she left the family when John was about four. The effect of this on John is unknown, as he never discussed it (van Dijken, 1998). The Bowlby children saw their mother for a short time each day, and their father on Sundays. Their father, who was knighted in 1911 as a result of being a surgeon to royalty, was absent even more during World War I. However, there were long family holidays in the country to the New Forest in July and to Scotland in late summer. This remained a family tradition with John, his wife Ursula, their four children, and later their grandchildren retreating to the Isle of Skye each summer. John did much of his writing there, as well as walking and enjoying the 'season,' including the Skye Ball with its Scottish reels. Skye is where John died on 2 September 1990. John, and later Ursula, are buried there in a remote churchyard above the cliffs and the sea.

When John was 11, he and Tony were sent off to boarding school. From there, John went to the Royal Naval College, Dartmouth. But at 17, he decided that the Navy was not for him, writing to his mother that he wanted a job, which would "improve the community as a whole" (van Dijken, 1998: p. 46). At Trinity College, Cambridge, John followed his father by studying medicine, with a departure into psychology in his final year, graduating in 1928. Rather than going straight into clinical school, John spent a year teaching in two boarding schools, including one for disturbed children. Their histories of disrupted childhoods impressed Bowlby, and he decided to combine his medical training at University College Hospital with psychoanalytic training at the Institute of Psycho-Analysis. After becoming medically qualified in 1933, Bowlby went to the Maudsley to train in adult psychiatry, and then to the London Child Guidance Clinic in 1936. He became an Army psychiatrist in 1940, and after the war he went to the Tavistock Clinic, where he remained.

Having begun analysis with Joan Riviere, Bowlby went on to qualify as an analyst in 1937 and then to train in child analysis with Melanie Klein. Differences grew concerning the importance of real-life experiences over unconscious fantasies. For example, when treating a hyperactive and anxious 3-year-old boy, Bowlby noted that the boy's mother was having a breakdown and had been admitted to mental hospital. Klein, however, did not allow Bowlby to treat the mother, devaluing what was actually happening in the child's life. Yet Bowlby was keenly interested in what happened, thanks to his love of nature, his scientific training, and his clinical observations (Holmes, 1993).

The Theme of Separation

Bowlby was particularly interested in what happened around separation. This interest may have sprung from his own childhood experience, as well as his work with maladjusted children in boarding school and later in the clinic. "I was struck by the high incidence of severely disrupted relationships with mother-figures in the early histories of children and adolescents who had been referred to the clinic on account of repeated and apparently incorrigible stealing." Bowlby's research for the WHO gave him the opportunity to survey the field and make personal contact with leading psychiatrists and psychologists in the United States and the continent, culminating in his influential WHO report (1952). His study with Joyce and James Robertson on children going to hospital indicated, in Bowlby's own words, that "First-hand observations of young children admitted to residential nursery or hospital and not visited by parents revealed the intense and prolonged distress they suffer. Moreover, visits home showed that the child's relationship with the mother is seriously disturbed for weeks or longer after

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61014-0

reunion. These observations lent considerable support to the hypothesis that the experience of separation and loss of mother-figure in the early years could have adverse effects on development, at least in some cases. That raised a question: if the disruption of a child's relationship with mother-figure in the early years creates much distress and anxiety, what is so special about the relationship that has been disrupted?" (1991a: pp. 302–303). The prevailing answer was that bond formation stemmed from the association of mother with the provision of food, thereby satisfying a primary need. But in Bowlby's view, this "cupboard love theory" (p. 303) was insufficient. Both experimental observations and a theoretical framework were lacking (for a full account of the transition from psychoanalysis to ethology, see van der Horst, 2011).

An Ethological Perspective

In the summer of 1951, Bowlby came across Lorenz's concept of imprinting and was encouraged by Julian Huxley to pursue ethology further. He recommended two new books, Lorenz's *King Solomon's Ring* and Tinbergen's *The Study of Instinct*. Bowlby spent some months 'mugging up on ethology' as well as making personal contact with the authors. In early 1954, Bowlby met Hinde at a meeting on ethology and psychiatry in London, where they were invited to speak in place of Lorenz and Tinbergen, who could not come. This led to Bowlby inviting Hinde to join his weekly seminar, held in a dingy room in the old Tavistock in Beaumont Street. The group was entirely heterogeneous: a Freudian and a Kleinian analyst, a Hullian and a Skinnerian learning theorist, a Piagetian, an ethologist, sometimes an antipsychiatrist, psychiatric social workers, having in common only an interest in parent–child relations. Bowlby's intellectual dynamism and judicious enthusiasm held the group together. During that period, Bowlby formulated his ideas on 'a child's tie to his mother' and 'separation anxiety.'

From an ethological perspective, the occurrence of species-characteristic behavior suggests that such behavior may have been selected for during the course of evolution. Bowlby applied this thinking to attachment behavior defined as any form of behavior, which attains or maintains proximity to a caregiver in times of need or stress. He argued that individuals who exhibited attachment behavior would have been more apt to survive and leave offspring who in turn would reproduce, compared with those who did not. Obviously selection for attachment behavior could not have happened without a similar pressure on its complement, caregiving behavior. "During the course of time, the biologically given strategy of attachment in the young has evolved in parallel with the complementary parental strategy of responsive caregiving, the one presumes the other" (Bowlby, 1991b: p. 293).

Bowlby further suggested that attachment behavior reflects a distinct motivational system in its own right, not dependent on prior association with any other motivational system such as hunger. Bowlby postulated that the biological function of attachment behavior is protection from harm, by keeping a child in touch with one or more caregivers. With this approach, activation of a fear behavior system would lead to activation of the attachment behavior system. Attachment behavior would lead to proximity to the caregiver, which would in turn deactivate the fear system, enabling activation of an exploratory or social system (e.g., Bowlby, 1969/1982).

Bowlby realized that an evolutionary argument could provide insight into behavior, which otherwise appeared abnormal. "It is against this picture of man's environment of evolutionary adaptedness that the environmentally stable behavioural equipment of man is considered. … not a single feature of a species' morphology, physiology, or behavior can be understood or even discussed intelligently except in relation to that species' environment of evolutionary adaptedness" (Bowlby, 1969/1982: p. 64). Bowlby used this reasoning to understand as 'natural' the occurrence of 'irrational fears' as well as the occurrence of attachment behavior throughout the life span. In considering the fears of childhood, such as fear of strangeness, separation, or darkness, Bowlby disagreed with influential analysts such as Klein or Freud who had labeled them as irrational. Instead he argued that our tendency to fear what had been dangers in the environments in which we evolved is "to be regarded as a natural disposition of man … that stays with him in some degree from infancy to old age … Thus it is not the presence of this tendency in childhood or later life that is pathological, pathology is indicated either when the tendency is apparently absent or when fear is aroused with unusual readiness and intensity" (Bowlby, 1973: p. 84).

Similarly, attachment behavior in times of stress should not be viewed as a sign of weakness, even beyond childhood. Long before research began on the role of attachment in adulthood, Bowlby wrote "such tendencies … are present not only during childhood but throughout the whole span of life. Approached in this way, fear of being separated unwillingly from an attachment figure at any phase of the life-cycle ceases to be a puzzle, and instead, becomes classifiable as an instinctive response to one of the naturally occurring clues to an increased risk of danger" (Bowlby, 1973: p. 86).

In one of his final contributions, Bowlby summed up thus, "Once we postulate the presence within the organism of an attachment behavioural system regarded as the product of evolution and having protection as its biological function, many of the puzzles that have perplexed students of human relationships are found to be soluble. … an urge to keep proximity or accessibility to someone seen as stronger or wiser, and who if responsive is deeply loved, comes to be recognised as an integral part of human nature and as having a vital role to play in life. Not only does its effective operation bring with it a strong feeling of security and contentment, but its temporary or long-term frustration causes acute or chronic anxiety and discontent. When seen in this light, the urge to keep proximity is to be respected, valued, and nurtured as making for potential strength, instead of being looked down upon, as so often hitherto, as a sign of inherent weakness" (Bowlby, 1991b: p. 293).

Development of an Attachment Bond

Thus, an attachment behavior system, organized as a distinct motivational system, gives rise to attachment behavior when activated by stress, particularly threats or perceived threats of separation. An infant is born with "a marked bias to respond in special ways to several kinds of stimuli that commonly

emanate from a human being" (Bowlby, 1969/1982: p. 265). As attachment behavior develops, it forms the basis for an inferred attachment bond. Bowlby described particular phases in its development: preattachment (from birth to about 2 months), involving signaling without discriminating one person from another; attachment-in-the-making (2–6 months) where signals become directed to particular persons; clear-cut attachment (0.5–4 years) with locomotion and goal-corrected behavior; and finally a goal-corrected partnership (4 years onward) with perspective taking, communication skills, and sharing mutual plans. Although additional attachments may develop throughout life, these early attachments endure. What is not common to all individuals is the quality of attachment, which tends to be consistent throughout life, and is reflected in the three basic patterns: avoidant, secure, and ambivalent. Although Bowlby has been mistakenly quoted as stating that such patterns are indelibly 'fixed' over the first year, he in fact stressed that any continuity must depend on a continuous flow of interactions. "It is evident that the particular pattern taken by any one child's attachment behavior turns partly on the initial biases that infant and mother each bring to their partnership and partly on the way that each affects the other during the course of it" (Bowlby, 1969/1982: p. 340). Bowlby continues, "By the time the first birthday is reached both mother and infant have commonly made so many adjustments in response to one another that the resulting pattern of interaction has already become highly characteristic" (Bowlby, 1969/1982: p. 348). In addition to the immediate characteristics of the infant, the behavior of the caregiver is determined by "the particular sequence of environments, from infancy onwards, within which development takes place" (Bowlby, 1969/1982: p. 378). Thus, while a consistent cycle of interactions tends to maintain a given quality of attachment, if the cycle were to change, then so would the pattern of attachment.

As to which pattern of attachment is desirable, Bowlby was concerned with what might be called 'psychological desiderata.' However, other perspectives may prevail, such as 'biological desiderata' for increasing inclusive fitness, or 'cultural desiderata.' Although influencing each other, the three desiderata, biological, psychological, and cultural, may differ, especially in modern industrialized societies (Hinde and Stevenson-Hinde, 1991). Nevertheless, it was 'psychological well-being,' which Bowlby as a practicing psychiatrist regarded as basic. Making an analogy with 'physical well-being,' Bowlby argued that 'psychological well-being' had an absolute meaning, with security of attachment an essential ingredient. Security in turn depends on antecedent interactions with a 'sensitively responsive' mother, as found in Mary Ainsworth's pioneering Baltimore study (Ainsworth et al., 1978) and subsequently in many others. In a review chapter, van IJzendoorn and Bakermans-Kranenburg (2004) conclude that correlational, experimental, and cross-cultural studies have replicated the association between sensitivity and attachment numerous times. Thus the main tenets of attachment theory appear to be universal: (1) the phenomenon of infants showing attachment behavior from birth, leading to (2) the formation of an attachment bond with one or more specific caregivers; and (3) the security of that attachment bond dependent upon antecedent caregiver responsiveness to infant's signals.

Assessing the Quality of Attachment

Bowlby's collaboration with Mary Ainsworth (1913–99) remains a model for creativity in science. Their friendship spanned 40 years, from 1950 when Ainsworth joined Bowlby at the Tavistock. After her longitudinal study of mother–infant behavior in Uganda, Ainsworth joined the faculty at the Johns Hopkins University in 1955. Appreciating that assessments of attachment must involve how a child uses mother as a 'secure base' when the attachment behavior system is activated, Ainsworth developed a Strange Situation procedure for infants (12–18 months old; Ainsworth et al., 1978). This was modified for older children, with her guidance (2.5–4.5 years: the Cassidy and Marvin system; 6 years: the Main and Cassidy system; reviewed by Solomon and George, in Cassidy and Shaver, 1999/2008). Ainsworth's procedure paved the way for empirical studies to explore the framework provided by Bowlby.

The Strange Situation procedure involves a series of episodes in an unfamiliar laboratory setting, including two separations and two reunions from an attachment figure. The coding emphasis is on the reunion episodes, which show how the child organizes behavior to mother following the stress of separation. A secure pattern involves the child-greeting mother with full gaze and positive affect. Interactions are calm, while also intimate and indicative of a special relationship. The two main insecure patterns may be contrasted with this, and indeed with each other. Whereas the avoidantly attached child shows minimal responses and maintains neutrality, the ambivalently attached child emphasizes dependence on mother, with angry/whiny resistance and/or immature behavior. A disorganized pattern reflects confusion and apprehension. Beyond infancy, one may see a transition from disorganization to a controlling pattern, reflecting an effort to reduce uncertainty by taking charge.

The move beyond behavioral observations to the level of representation was largely inspired by one of Mary Ainsworth's former PhD students, Mary Main. Once children reach Bowlby's 'goal-corrected partnership' stage, their verbal behavior may be used to index representations of attachment (reviewed by Solomon and George, in Cassidy and Shaver, 1999/2008). Mary Main and colleagues developed the Adult Attachment Interview (AAI), based around 18 questions designed to access one's current state of mind with respect to attachment. Correspondence between parents' AAI classifications and their infants' Strange Situation classifications has now been reported in many studies (reviewed by Hesse, in Cassidy and Shaver, 1999/2008).

Emotional Communication

In Bowlby's own words, "Thus what is happening during these early years is that the pattern of communication that a child adopts towards his mother comes to match the pattern of communication that she has been adopting towards him. Furthermore, from the findings stemming from the Adult Attachment Interview, it seems clear that the pattern of communication a mother adopts towards her infant and young child is modeled on the pattern characteristic of her own intrapsychic communications. Open and coherent intrapsychic communication in a mother is associated with open and

coherent two-way communication with her child, and vice versa" (1991b: pp. 295–296). To Bowlby, "the concern of a psychoanalyst is not only with the extent to which a patient feels free to express his emotions openly, but with the prior questions of whether he knows what his feelings are and what has aroused them" (1991b: p. 294).

One reason for the success of attachment theory is that it lies at the heart of our emotional life. "Many of the most intense emotions arise during the formation, the maintenance, the disruption and the renewal of attachment relationships. The formation of a bond is described as falling in love, maintaining a bond as loving someone, and losing a partner as grieving over someone. Similarly, threat of loss arouses anxiety and actual loss gives rise to sorrow; while each of these situations is likely to arouse anger. The unchallenged maintenance of a bond is experienced as a source of security and the renewal of a bond as a source of joy" (1991a: p. 306).

Implications

Normative Development and Parenting

The implications of attachment theory for the development and expression of emotions within close relationships, including how patterns may be transmitted across generations, are clear from the above. Furthermore, the sense of security, which must underlie any appreciation of emotions within one's self and their expression to others depends upon having attachment figures who are sensitively responsive to emotional needs. Bowlby postulated that experiences within early close relationships lead to an 'internal working model' of the self in relation to others that guides future interactions. With a secure working model comes the expectation of gaining support when it is needed, along with feeling worthy of such support. Empirical studies have related security to particular developmental outcomes, including resilience, empathy, and social competence (reviewed in Weinfield et al., in Cassidy and Shaver, 1999/2008). A classic study – the Minnesota Longitudinal Study of Risk and Adaptation from Birth to Adulthood – involved a variety of assessments in multiple contexts of 180 children born into poverty. Security in infancy was a predictor of adaptation in playgroup, which along with security was an additional predictor of social competence in middle childhood. The Minnesota study illustrates how early close relationships may set the stage for later ones, which along with changing circumstances may affect the outcome, in a probabilistic way. With its range of assessments and outcomes over time, the Minnesota study supports the view that attachment status should not and does not predict every aspect of development.

Finally, cross-cultural studies highlight the need to take a child's network of early relationships into account in predicting the outcome (van IJzendoorn & Sagi-Schwartz, in Cassidy and Shaver, 1999/2008). A further challenge involves defining and assessing attachment-related outcomes in culturally appropriate ways (Chen and Rubin, 2011).

Psychopathology

Attachment theory has direct implications for disorders related to separation or loss, such as abnormal grief, neurotic depression, or agoraphobia (see Holmes, 1993). Furthermore, particular patterns of insecure attachment may shed light on aspects of parenting that might be modified to promote security. For example, the disorganized pattern, often seen in clinical samples, appears to be associated with fearfulness in either parent or child or both, thereby suggesting a window for intervention (Lyons-Ruth & Jacobvitz, in Cassidy and Shaver, 1999/2008). Concerning the early onset of disruptive behavior problems, Deklyen & Greenberg (in Cassidy and Shaver, 1999/2008) propose a 4-factor risk model, which includes attachment. In their own words, "attachment theory provides a critical developmental frame for understanding *how* caregiving relationships influence processes thought to be central to emerging psychopathology – for example, the construction of cognitive-affective expectancies, the capacity for emotional and behavioral regulation, and strategies for coping with stress" (p. 637).

Social Policy Issues

When a child goes to hospital, we now take it for granted that parents may visit or even live in. However, before Bowlby's influence, hospital practice involved leaving caregiving to the experts and keeping attachment figures away. Obstetric practice followed similar lines. Bowlby (1953) did not hesitate to challenge this: "reflect for a moment on the astonishing practice which has been followed in obstetric wards of separating mothers and babies immediately after birth and ask … whether this is the way to promote a close mother–child relationship. It is hoped that this madness of western society will never be copied by so-called less developed societies." Such concern over separation from attachment figures remains relevant to current provision of care, whether for children placed in multiple foster homes with long delays until adoption or for adults, such as the elderly and mentally ill.

Regarding day care, Bowlby was keen to put right the misinterpretation of what he wrote. In stressing the importance of early caregiving, he did not mean that a parent must be with the child all the time. From the above it will be clear that attachment theory is about the quality of a relationship rather than amount of time spent together. Similarly with adoption, attachment theory does not say that early adverse experience cannot be overcome. Indeed the implication is that adoptive parents, if sensitively responsive to the child's needs, may do a great deal to set interactions in a cycle that will promote security.

Finally, attachment theory has been applied to the level of society. As a professor of social planning has urged, "We need also to institute a style of governing our relationships with each other which takes as its first principles reciprocity of commitment, predictability, and respect for the unique structure of meaning and attachment which makes life worthwhile for each of us." Such an argument implies the "need to create a world in which we dare become attached" (Marris, 1991: p. 89).

Conclusion

Bowlby's framework for understanding a core aspect of development is set out in his trilogy, *Attachment and Loss*

(1969/1982, 1973, 1980). The fruits of his theory are gathered together in the first and second editions of the *Handbook of Attachment* (Cassidy and Shaver, 1999/2008) dedicated to John Bowlby and Mary Ainsworth. Each edition is organized into six sections, with the present (2008) edition consisting of: Overview, Biological perspectives, Attachment in infancy and childhood, Attachment in adolescence and adulthood, Psychopathology and clinical applications, and Systems, culture, and context.

The last major publication of Bowlby himself was a biography of Charles Darwin (Bowlby, 1991c). With a convincing application of attachment thinking, Bowlby explained Darwin's lifelong psychosomatic symptoms in terms of unmourned loss of mother at the age of 8. Additionally, a keen reader may find parallels between the two men. Like Darwin, Bowlby had a distant father who was a doctor, both were younger sons, both were open to evidence, and both lacked rancor toward their detractors. Finally, it might be said of their theories that they have the quality of immediacy and 'obviousness.' In retrospect it seems obvious that species have evolved by natural selection, that people are attached to one another and suffer when they separate "but it took child-like simplicity of vision combined with mature determination and attention to detail to root out the obvious and to create for it a secure theoretical base" (Holmes, 1993: pp. 34–35; see also van Dijken, 1998).

See also: Childhood and Adolescence: Developmental Assets; Emotion and Expression; Infancy and Childhood: Emotional Development; Klein, Melanie (1882–1960); Love and Intimacy, Psychology of; Psychoanalysis, History of.

Bibliography

Ainsworth, M.D.S., Blehar, M.C., Waters, E., Wall, S., 1978. Patterns of Attachment. Erlbaum, Hillsdale, NJ.

Bowlby, J., 1952. Maternal Care and Mental Health. World Health Organization, Geneva.

Bowlby, J., 1953. Child Care and the Growth of Love. Penguin, Melbourne, Australia.

Bowlby, J., 1969/1982. Attachment and Loss. In: Attachment, vol. I. Hogarth Press, London.

Bowlby, J., 1973. Attachment and Loss. In: Separation, Anxiety and Anger, vol. II. Hogarth Press, London.

Bowlby, J., 1980. Attachment and Loss. In: Loss, Sadness and Depression, vol. III. Hogarth Press, London.

Bowlby, J., 1991a. Ethological light on psychoanalytical problems. In: Bateson, P. (Ed.), Development and Integration of Behaviour. Cambridge University Press, Cambridge, UK, pp. 301–313.

Bowlby, J., 1991b. Postscript. In: Parkes, C.M., Stevenson-Hinde, J., Marris, P. (Eds.), Attachment across the Life Cycle. Tavistock-Routledge, London, pp. 293–297.

Bowlby, J., 1991c. Charles Darwin: A New Life. Norton, New York.

Cassidy, J., Shaver, P.R. (Eds.), 1999/2008. Handbook of Attachment: Theory, Research, and Clinical Applications. Guilford Press, New York.

Chen, X., Rubin, K.H. (Eds.), 2011. Socioemotional Development in Cultural Context. Guilford Press, New York.

van der Horst, F.C.P., 2011. John Bowlby: From Psychoanalysis to Ethology. Wiley-Blackwell, London.

van Dijken, S., 1998. John Bowlby: His Early Life. Free Association Books, London.

van IJzendoorn, M.H., Bakermans-Kranenburg, M.J., 2004. Maternal sensitivity and infant temperament in the formation of attachment. In: Bremner, G., Slater, A. (Eds.), Theories of Infant Development. Blackwell, London, pp. 233–258.

Hinde, R.A., Stevenson-Hinde, J., 1991. Perspectives on attachment. In: Parkes, C.M., Stevenson-Hinde, J., Marris, P. (Eds.), Attachment across the Life Cycle. Routledge, London, pp. 52–65.

Holmes, J., 1993. John Bowlby and Attachment Theory. Tavistock-Routledge, London.

Marris, P., 1991. The social construction of uncertainty. In: Parkes, C.M., Stevenson-Hinde, J., Marris, P. (Eds.), Attachment across the Life Cycle. Tavistock-Routledge, London, pp. 77–90.

Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences

William J Hoyer, Syracuse University, Syracuse, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W.J. Hoyer, volume 2, pp. 1318–1321, © 2001, Elsevier Ltd.

Abstract

Normal or typical age-related changes in brain mechanisms have consequences for cognitive and personality function during the adult years and in old age. In recent years, much has been learned about cellular and molecular details of brain aging from advances in basic sciences. Advances in the understanding of the neurobiological bases of complex human behavior have come about through improved brain-imaging technologies and improved quantitative modeling of the dynamic relations between brain aging and behavior. New findings and theoretical models that attempt to link advances in the understanding of brain aging with observations of age-related changes in behavior, cognition, and personality point to the dynamic interplay between behavior and brain. Concepts of brain plasticity, cognitive plasticity, cognitive reserve, and compensation are emphasized in much of the recent research on the relations between brain aging and its consequences for cognitive and personality. Progress in this area informs the understanding of the potentials and limits of human aging and conceptualizations of interventions for healthy aging.

Introduction

Throughout the life span, a complex interplay of developmentally dynamic neurobiological mechanisms, genetic factors, environmental factors, and the individual's actions and experiences yields a range of manifestations that are expressed in the cognitive and personal attributes of individuals (e.g., see Lövdén et al., 2010). Generally, healthy older adults show age-related cognitive deficits in situations that require the management and coordination of multiple tasks or that involve delayed recollection of details about events or experiences (for a review, see Verhaeghen, 2013). Age-related cognitive deficits are most prominent in tasks that depend on hippocampal and prefrontal brain circuitry. In terms of personality, healthy individuals usually show only minor changes in personality traits across the adult years. Although the consequences of brain aging on personality across the adult life span are usually not readily observable, recent evidence suggests that particular personality traits influence brain aging mechanisms and age-related differences in neuroanatomical integrity (e.g., Hill et al., 2011). Further, it is well-known that some personality traits are associated with greater mortality risk, whereas other personality traits are associated with greater longevity and increased adherence to health-promoting activities (e.g., medical adherence, dietary behavior, exercise). Cognitive and social behavior and personality attributes have consequences for healthy aging, morbidity, and mortality risk.

Scientific knowledge about the specific relations between brain aging, cognition, and personality has increased tremendously in recent years, due to advances in brain science, epigenetics, and in application and resolution of noninvasive techniques for assessing brain–cognition–personality relations, such as structural and functional magnetic resonance imaging (fMRI) and positron emission tomography (PET). Recent advances in neuroimaging allow identification of the brain regions associated with age-related changes in selected cognitive functions. One relatively new technique, magnetic resonance diffusion tensor imaging (DTI), is useful for characterizing the integrity of anatomical connections and white-matter circuitry and provides a quantitative assessment of white-matter microstructure. For example, DTI studies show that the greatest amount of age-related alterations in white matter occur in the prefrontal cortex and the anterior corpus callosum. The kinds of cognitive tasks that show the largest age-related deficits depend on exactly these brain regions.

This article emphasizes the idea that the aging of brain–cognition–personality relations is a consequence of the interplay between multiple etiologies and moderators that involve not only neural mechanisms and structural changes in the brain, but also the influences of genetic factors, environmental factors, and health and lifestyle experiences of the individual. We begin this article with a brief summary of brain aging and plasticity of function, and cognitive reserve. Then, we discuss recent advances in the understanding of the relations between brain aging, cognition, and personality.

Brain Aging

Brain aging reflects time-based alterations in a variety of neuroanatomical and biochemical structures and functions. Structurally, MRI data using voxel-based morphometry show age-related shrinkage in the volume of the ventricles of the brain, and show that the degree of shrinkage varies somewhat for different brain regions. The ventricles are interconnected cavities within the brain that contain cerebrospinal fluid. Because the ventricular cavities are interconnected, changes in the ventricular system can be considered to be a fair index of the general health of the brain. As the brain shrinks in size with aging, the ventricles expand to take up more space.

Neurogenesis refers to the production of new neurons originating from stem cells and progenitor cells. Stem cells are unprogrammed cells that divide and can change into cells with specialized functions. A host of factors (including individually controlled actions such as exercising and maintaining a healthy diet) can stimulate or deter neurogenesis. The hippocampus, a brain region that is centrally involved in associative learning and memory, is known to produce new neurons throughout life (e.g., Deng et al., 2010). A steady production of new neurons supports effective hippocampal function, whereas

deterred neurogenesis eventually results in memory failure. Cognitive health is associated with increased neurogenesis and the strengthening of particular pathways in the hippocampus in line with functional demands (e.g., Hertzog et al., 2009).

Neural integrity and neural functions are impaired by the accumulation of beta amyloid and by inflammations within the environment of neurons. Within-brain environmental factors affect the vitality of neurons and their physiology. Epigenesis refers to the premise that development throughout life roughly follows a genetic plan, and that the plan is modifiable by the turning-on and turning-off of genes in response to experiences and environmental factors. For example, either acute or chronic exposure to toxins can accelerate inflammation in the central nervous system. Thus, the course and rate of genetically programmed aging depend on a host of within-brain factors and external factors.

Aging of Brain–Cognition–Personality Relations

Intraindividual changes in cognition depend on the interplay between neural mechanisms and brain structures, genetic influences, and environmental influences. For example, the Catechol-*O*-Methyltransferase (COMT) genotype affects age-related variability in decision time in younger and older adults (Li et al., 2010). The COMT gene affects the degradation of dopamine in the prefrontal cortex. Dopamine is manufactured in the cells of the substantia nigra and travels in the pathways connected to the frontal cortex. Dopaminergic systems normally decline with aging across various brain regions. Age-related reductions in dopamine are known to have detrimental consequences for cognitive and psychomotor functions including the planning and execution of actions (Bäckman et al., 2010). Compared to older individuals with a genetic predisposition for efficient dopamine signaling in the prefrontal cortex, older individuals with a genetic predisposition for less-efficient dopamine signaling in the prefrontal cortex show greater variability in performance and an overall lower rate of information processing when carrying out a simple two-choice perceptual decision task (Li et al., 2010).

A relation between the Cholesterol Ester Transfer Protein (CEPT) genotype and cognitive function has also been demonstrated (Barzilai et al., 2006). A variant of the CEPT genotype is associated with preserved cognitive function in very old adults. This variant is associated with lower CEPT levels and a favorable lipoprotein profile. These and other findings of associations between particular genetic variants and polymorphisms, and differences in relative efficiency of neuromodulation and cognitive function point to the necessity of a multicausal view in building a comprehensive understanding of developmental changes in brain–cognition relations.

Recent research suggests that a fundamental issue in the understanding of brain aging has to do with change in the effectiveness of switching between two brain networks, the default network and the attention network (e.g., Spreng and Schacter, 2012). The default network is an integrated system of regions that includes the medial prefrontal cortex, posterior cingulate cortex, superior frontal gyrus, medial and lateral temporal lobes, and the posterior part of the inferior temporal lobe. The default network is active in internally focused tasks such as autobiographical memory, imagining the future, and social cognition. In contrast, the attention network is an integrated system of regions that includes the dorsolateral prefrontal cortex, frontal eye fields, inferior precentral sulcus, middle temporal motion complex, and superior parietal lobe. The attention network is active in externally driven tasks, such as processing and making responses to visual-spatial information. A third neural network, referred to as the control network, expedites the shift between the default network and the attention network in response to changing internal and external states. The regions of the control network include rostrolateral prefrontal cortex, middle frontal gyrus, precuneus, the anterior extent of the inferior parietal lobe, dorsal anterior cingulate, and the anterior insula. Thus, the control network is anatomically interposed between the default network and the attention network, and serves to facilitate the ability to rapidly adapt thoughts and behaviors to changing internal states and external stimuli. In a recent study examining the interactivity of these networks during autobiographical memory and visual-spatial planning tasks in young adults and older adults, it was found that young adults could efficiently deactivate the default network and that older adults failed to do so. The failure was not due to default network dysfunction, but rather to a failure to a lower degree of flexible network interactivity.

Brain and Cognitive Plasticity

Maintenance as well as change in effective cognitive function depends on the adaptability of brain functions (within constraints) in response to experience. Normal neurobiological aging brings with it across-the-board compromise in a wide range of cognitive functions. However, cognitive performance can be enhanced or at least maintained by experience via whatever allowances and compensations there might be in the neural circuitry of individuals as they grow older. Concepts of experience-based plasticity are now prominent in neuroscience.

Neuronal viability is significantly involved in normal brain aging. Shrinkage of dendritic branching and other changes in neuronal viability and in the neural architecture with aging are central to the understanding of age-related changes in cognitive function. Various aspects of biological aging conspire to produce age-related losses in behavioral and cognitive function. Current theories of biological aging emphasize genetically programmed deficits, the cumulative negative consequences of mutation, and the cumulative negative effects of wear and tear on function. Among the brain areas most affected by aging are the frontal-striatal circuits of the caudate. Other areas that are known to be substantially affected by normal brain aging are the prefrontal cortex and the dense network of projections connecting the prefrontal cortex and subcortical monoaminergic nuclei. The extent of age-related deterioration is also substantial for the posterior association areas and the neostriatum. The primary sensory areas of the brain are largely unaffected by aging. In light of findings of preserved brain function in the sensory areas, reliable declines in sensory efficiency with aging are attributed to receptor function and peripheral neural function (e.g., see Bennett et al., 2009).

The main, general source of behavioral deficits associated with aging is known to be general slowing in the speed of mental processes. Age-related slowing in the speed of encoding, perceiving, processing, retrieving information accounts for a substantial part of the variance in comparisons between the performance of younger and older adults across a wide range of tasks (e.g., for a comprehensive meta-analytic review, see Verhaeghen, 2013). Specifically, in 1354 young-old data pairs, Verhaeghen (2013) found that the performance of older adults could be quantitatively described in terms of a ratio of between 1.2 and 1.7 of the performance of young adults. Variation in the size of ratio was associated with type of task; the slowing factor for tasks associated with the temporal lobes was about 1.2; the slowing factor for tasks associated with the frontal lobes was about 1.3; the slowing factor for tasks associated with the occipital lobes was about 1.5; and the slowing factor for tasks associated with the parietal lobes was 1.8.

Age-related slowing in cognitive performance across tasks may be associated with a number of brain aging phenomena such as changes in neurotransmitter substances at the synapse, neural demyelination, disruptions in neural circuitry-related vascular lesions, and increases in the extent or amount of recruitment and activation of brain volume. With regard to possible age differences in the amount of recruitment of brain areas, the results of several recent fMRI studies suggest that more cognitive resources are required for older adults than for younger adults to carry out cognitive tasks (e.g., see Reuter-Lorenz and Park, 2010). In these studies, younger persons showed a relatively well-defined pattern of activation involving the inferior prefrontal-orbitofrontal cortex during the encoding of episodic information, and involving the right prefrontal cortex during the retrieval of episodic information. In contrast, older adults showed more widespread activation patterns at encoding and retrieval.

Much of the work in the cognitive neuroscience of aging has focused on age-related memory loss. Some aspects of memory function show large deficits associated with normal aging, whereas other aspects of remembering appear to be relatively unimpaired in healthy older adults (e.g., semantic priming, implicit memory). Relatively substantial age-related declines in cognitive function can be observed in tasks that entail the active and simultaneous processing and storing of information. This kind of processing, referred to as working memory, is a key component of cognitive tasks that require juggling lots of information at once. Research bearing on the neural substrates underlying age-related deficits in working memory suggests that working memory requires cooperation among scattered areas of the brain. Which specific areas are activated depends on whether the task requires spatial or semantic memory. Working memory has been associated with the prefrontal cortex. Structural imaging studies indicate shrinkage of the prefrontal cortex with aging. However, functional imaging studies indicate increased activation in the prefrontal areas for older adults than for younger adults. These findings can be interpreted as suggesting that older adults use more cognitive resources to carry out demanding cognitive tasks.

It has been proposed that there are two kinds of reserve that protect and support function in the face of brain pathology; brain reserve and cognitive reserve. Brain reserve refers to measures of brain vitality, brain size. Cognitive reserve is based on observations that individuals differ in their ability to sustain effective cognitive function in the presence of brain pathology. Evidence supporting the cognitive reserve hypothesis demonstrates that individuals with higher cognitive reserve show better maintenance of effective cognitive function than individuals with lower cognitive reserve (Ewers et al., 2013).

Although brain aging phenomena do not have obvious consequences for personality changing during the adult years, there is clear evidence showing that selected personality attributes have consequences for brain aging and for health. In a recent MRI study of the relations between personality and regional brain volume in individuals between 44 and 88 years of age, individuals who scored high in the personality trait of neuroticism from the NEO Personality Inventory showed smaller volumes of gray matter in the frontal and medial temporal regions and greater age-related decreases in regional volume. Further, individuals who scored high in the personality trait of conscientiousness from the NEO Personality Inventory showed higher volumes of gray matter and less decline in regional volume (Jackson et al., 2011). These findings suggest that the personality attributes of neuroticism and conscientiousness moderate age-related cross-sectional decline in the prefrontal and medial temporal regions of the brain.

Developmental studies of health–personality relations have found that mortality risk is lower for individuals who score high on conscientiousness and that mortality risk is higher for individuals who score high on neuroticism. Further, it is well-known that personality traits affect maintenance of health-promoting behavior (dietary behavior, exercise), and have consequences for physical health outcomes, including cardiovascular disease, elevated blood pressure, diabetes, and mortality.

In describing the relations between personality and brain aging, it is important to recognize that the brain is the central mediator of ongoing, system-wide physiological adaptation to environmental challenges (e.g., see Ganzel et al., 2010). Indeed, the healthy brain is dynamically adaptive in mediating between ongoing environmental challenges and the biological self, and in regulating the neurophysiological consequences of prolonged and acute stress.

Conclusions

Age-related changes in behavior, cognition, and personality during adult years and in later life depend on the dynamic interplay between brain aging mechanisms and behavioral factors (i.e., learning, regularly occurring adaptive organism–environment transactions). Age-related losses occur for many aspects of brain aging, but there is also growth and stability for some aspects of brain function throughout life. Concepts of brain plasticity and brain reserve capacity give emphasis to the potential of individuals to continue to improve, maintain, or optimize behavioral function in response to insidious losses associated with normal aging, and to relatively abrupt losses associated with disease.

See also: Aging Mind: Facets and Levels of Analysis; Aging and Health in Old Age; Aging and Memory; Aging, Theories of; Cognitive Aging; Genetic Approaches to Memory.

Bibliography

Alosco, M.L., Spitznagel, M.B., Raz, N., Cohen, R., Sweet, L.H., van Dulmen, M., Colbert, L.H., Josephson, R., Waechter, D., Hughes, J., Rosneck, J., Gunstad, J., 2012. Cognitive reserve moderates the association between heart failure and cognitive impairment. Journal of Clinical and Experimental Neuropsychology 34, 1–10.

Bäckman, L., Lindenberger, U., Li, S.-C., Nyberg, L., 2010. Linking cognitive aging to alterations in dopamine neurotransmitter functioning: Recent data and future avenues. Neuroscience and Biobehavioral Reviews 34, 670–677.

Barzilai, N., Atzmon, G., Derby, C.A., Bauman, J.M., Lipton, R.B., 2006. A genotype of exceptional longevity is associated with preservation of cognitive function. Neurology 67, 2170–2175.

Bennett, I.J., Madden, D.J., Vaidya, C.J., Howard, D.V., Howard, J.H., 2009. Age-related differences in multiple measures of white matter integrity: a diffusion tensor imaging study of healthy aging. Human Brain Mapping 31, 378–390.

Bennett, I.J., Rypma, B., 2013. Advances in functional neuroanatomy: a review of combined DTI and fMRI studies in healthy younger and older adults. Neuroscience & Biobehavioral Reviews 37 (7), 1201–1210.

Cacioppo, J.T., Decety, J., 2009. What are the brain mechanisms on which psychological processes are based? Perspectives in Psychological Science 4, 10–18.

Deng, W., Aimone, J.B., Gage, F.H., 2010. New neurons and new memories: how does adult hippocampal neurogenesis affect learning and memory? Nature Reviews Neuroscience 11, 339–350.

Ewers, M., Insel, P.S., Stern, Y., et al., 2013. Cognitive reserve associated with FDG-PET in preclinical Alzheimer's disease. Neurology 80, 1194–1201.

Ganzel, B.L., Morris, P.A., Wethington, E., 2010. Allostasis and the human brain: integrating models of stress from the social and life sciences. Psychological Review 117, 134–174.

Hertzog, C., Kramer, A.F., Wilson, R.S., Lindenberger, U., 2009. Enrichment effects on adult cognitive development: can the functional capacity of older adults be preserved and enhanced? Psychological Science in the Public Interest 9, 1–65.

Hill, P.L., Turiano, N.A., Hurd, M.D., Mroczek, D.K., Roberts, B.W., 2011. Conscientiousness and longevity: an examination of possible mediators. Health Psychology 30, 536–541.

Jackson, J., Balota, D.A., Head, D., 2011. Exploring the relationship between personality and regional brain volume in healthy aging. Neurobiology of Aging 32, 2162–2171.

Li, S.-C., Lindenberger, U., Bäckman, L., 2010. Dopaminergic modulation of cognition across the life span. Neuroscience and Biobehavioral Reviews 34, 625–638.

Lövdén, M., Bäckman, L., Lindenberger, U., Schaefer, S., Schmiedek, F., 2010. A theoretical framework for the study of adult cognitive plasticity. Psychological Bulletin 136, 659–676.

Mõttus, R., McNeill, G., Jia, X., Craig, L.C., Starr, J.M., Deary, I.J., 2013. The associations between personality, diet and body mass index in older people. Health Psychology 32, 353–360.

Mroczek, D.K., Spiro, A., 2007. Personality change influences mortality in older men. Psychological Science 18, 371–376.

Park, D.C., Bischof, G.N., 2013. The aging mind: neuroplasticity in response to cognitive training. Dialogues in Clinical Neuroscience 15, 110–119.

Reuter-Lorenz, P.A., Park, D.C., 2010. Human neuroscience and the aging mind: a new look at old problems. Journal of Gerontology: Series B: Psychological Sciences 65B, 405–415.

Spiro, A., Brady, C.B., 2011. Integrating health into cognitive aging: toward a preventive cognitive neuroscience of aging. Journal of Gerontology: Series B: Psychological Sciences 66B, 117–125.

Spreng, R.,N., Schacter, D.L., 2012. Default network modulation and large-scale network interactivity in healthy young and older adults. Cerebral Cortex 22, 2610–2621.

Sullivan, E.V., Pfefferbaum, A., 2006. Diffusion tensor imaging and aging. Neuroscience & Biobehavioral Reviews 30, 749–761.

Verhaeghen, P., 2013. Essentials of Cognitive Aging. Oxford University Press, New York.

Brain and Behavioral Lateralization in Animals

Lesley J Rogers, School of Science and Technology, University of New England, Armidale, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Differential control of behavior and information processing by the left and right hemispheres of the brain is now well known to be characteristic of a wide number of vertebrate species, thus overturning the long-held myth that it is unique to humans. Perhaps, even more unexpected has been the discovery that the pattern of lateralization in vertebrate species is fundamentally the same as that in humans. Research on animals is revealing important information about the development, evolution, causation, and function of this key aspect of brain organization.

Definition of Lateralization

Lateralization refers to left–right structural or functional differences. Structural differences may be present between the left and right sides of the body or the brain. Here, the focus is on left–right differences in brain and behavior. In other words, the question addressed is whether the left and right sides of the brain control different functions and process information in different ways. Such side differences in behavior may exist without apparent, or known, differences in structure.

Within a population or species lateralization may be present in two ways: at the individual or population level. Individual-level lateralization refers to examples in which most, if not all, individuals show lateral biases but the number biased to the left is the same as the number biased to the right. Population-level lateralization refers to directional bias in which the majority of animals are biased in the same direction.

History

Lateralization used to be considered a unique characteristic of the human brain, and therefore a specialization making us different from, and superior to, all other animals. This view was held for over a century, ever since the discovery in the 1860s of lateralization in humans. Particular importance was made of the left hemisphere's specialization for speech and language and the right-hand preference for fine manipulation required in using tools and writing (Corballis, 1991).

In the 1970s and 1980s, the misconception of human exclusiveness in this characteristic was laid bare by the discovery of functional lateralization in avian and rodent species. In a songbird species, it was shown that control of singing is a function of the higher vocal center in the left hemisphere but not by the equivalent area in the right hemisphere (Nottebohm et al., 1976). In domestic chicks, it was shown that the left but not the right hemisphere is used when searching for grains made difficult to find by scattering them among pebbles of similar shapes, sizes and colors, and adhered to the floor (Rogers and Anson, 1979). The same study also found that the left hemisphere is used in auditory habituation learning. In rats, hemispheric differences were shown for control of activity in the open field (Denenberg et al., 1978) and the right hemisphere was found to be specialized for control of spatial and affective (emotional) behavior (Denenberg, 1981, 1984). These examples of lateralization were all directional biases present at the population level.

Since these initial discoveries, the field of lateralization in vertebrate species has expanded with increasing rapidity and it is now clear that lateralization of brain and behavior is widespread among vertebrate species and expressed in a broad spectrum of behavior (Rogers and Andrew, 2002; Rogers et al., 2013b). It has even been revealed recently in some invertebrate species (see below). Among the vertebrates, certain species have emerged as models for studying the development and nature of lateralization of brain and behavior: these model species include the domestic chicken, the domestic pigeon, and the zebra fish (Halpern et al., 2005). Among the invertebrates, model species used to investigate lateralization include the nematode *Caenorhabditis elegans* (Hobert, 2014) and, as a recent addition, honeybees (see below).

Until about a decade ago, researchers working on lateralization in humans had been reluctant to accept the presence of lateralization of brain and behavior in nonhuman species but that is no longer the case. Some researchers still argue that lateralization of the brain is stronger in humans than in any other species but this view is, in essence, simply a broadening of the originally held belief of human exclusiveness of laterality and no evidence supports it. On the contrary, the footedness of some species of cockatoos is as strong as, or even stronger than, the right-hand preference of humans (Harris, 1989; Rogers, 1980) and just as many brain functions are known to be lateralized in birds, for example, as is the case in humans (Rogers, 2008).

Methods of Revealing Lateralization

These early examples of lateralization were shown by unihemispheric placement of lesions in songbirds (e.g., canaries), ablation of one or the other hemisphere in rats and, in chickens, unihemispheric administration of pharmacological agents that impair neural function or development (e.g., cycloheximide or glutamate). Behavior of animals with treatment of the left hemisphere was compared to that of animals with the right hemisphere treated in the same way. Following this early work, it was found that lateralization could be demonstrated in species with their eyes positioned on the sides

of their head simply by putting a patch over one or the other eye and testing their behavioral performance. This procedure reveals functional differences between the hemispheres because, in these species, input from one eye is processed mainly by the opposite side of the brain. The first use of the monocular technique, by Mench and Andrew (1986), confirmed the role of the left hemisphere (right eye) in chicks searching for grain on a background of distracting pebbles. Eye patching has become a standard technique for showing visual lateralization, and it has been used especially in domestic chicks (research summarized in Rogers et al., 2013b). Eye patching does not reveal visual lateralization in species with their eyes positioned frontally (i.e., with a large binocular field of vision) because their visual input, except that from the extreme left or right sides of the visual field, goes from each eye to both sides of the brain. In these species, as in humans, presentation of stimuli in the extreme left or right visual field can reveal lateralization and this has been used in the laboratory to test primates after sectioning of the commissure that connects the left and right hemispheres, known as split-brain preparations (Hamilton and Vermeire, 1988).

Another technique now used widely is to compare responses to stimuli seen in the left monocular visual field with responses to those seen in the right monocular visual field. Toads, for example, react more strongly to a model predator (mimicking an advancing snake) seen in their left monocular visual field than they do to the same stimulus seen in their right monocular visual field (Lippolis et al., 2002) and this asymmetry has also been shown in marsupial dunnarts (Lippolis et al., 2005) and horses (Austin and Rogers, 2007). Stronger aggressive and sexual responses are directed toward conspecifics seen on an individual's left side, as found in a number of species, including birds, lizards, frogs, baboons, and feral horses (summarized in Rogers, 2002; for evidence in horses see Austin and Rogers, 2012).

A variation on the latter way of revealing lateralization is to record head turning and the eye used to look at particular stimuli. As an example of the latter, lizards turn their head so that they can use their left eye to view a potential predator approaching from behind (Bonati et al., 2010) and Australian magpies tilt their head to look upward with the left eye to see aerial predators (Kaplan and Rogers, 2013). This method has excellent application in field studies of species with lateral eyes.

Measuring hand or limb preferences are other ways in which researchers have attempted to determine whether or not a species is lateralized. Determining whether primates show handedness has been of particular interest, stemming from a desire to see whether our closest relatives might show right-handedness, as do humans. The first reports appeared to confirm the uniqueness of humans in this respect: primates were seen to have individual hand preferences, which could have been acquired through practice, but no directional biases within species (Warren, 1980). However, since the data collected were rather piece-meal and published in a variety of journals, there was a need to examine it more comprehensively and in doing so MacNeilage et al. (1987) found convincing evidence that primates do show hand preferences as directional biases. Lower primates (prosimians) show left-hand preferences for capturing insects and grasping food while they support their body using the right hand. As postulated by the Postural Origins Theory, this shifted to right hand preference for manipulation of objects as primates evolved to become more bipedal and the right hand was less needed to support the body, culminating in the right-handedness of humans (MacNeilage et al., 1987). At first there was much controversy about whether or not chimpanzees are right-handed but more recent evidence from larger samples shows that they do have a significant preference to use the right hand when using tools or in bimanual coordination (Hopkins et al., 2004, 2011).

Although population bias of limb preference shows brain lateralization of motor control, its absence in a species does not mean that brain lateralization for perceptual processing is absent (explored further in Rogers, 2009). Indeed, the early evidence for brain lateralization in rats (discussed above) came at a time when it was known that individual rats had paw preferences but they had no directional bias at the population level to use the same preferred paw to reach for and grasp food items (summarized in chapter 3 of Bradshaw and Rogers, 1993). Those working on primates took little to no cognizance of this evidence and continued to focus on measuring hand preferences as the assumed marker of brain lateralization. For this reason, the battle about human uniqueness continued for longer than it might have done so had researchers considered the evidence for limb preferences in species other than primates.

Lateralization of brain activity can be revealed by using radioactive tags that reflect increased blood flow or early gene expression in regions of the left and right hemispheres. Since these techniques require killing of the animals followed by sectioning of their brain, functional magnetic resonance imaging of the brains of living animals is a preferred method of revealing lateralization. It is a recent development among the techniques used to study lateralization in animals. For example, it has been used to examine regional activity of the chimpanzee cortex during communicative signaling. In fact, a bias to use regions of the left hemisphere (e.g., the left inferior frontal gyrus) suggests that chimpanzees may possess neural precursors of human language (Taglialatela et al., 2008). As imagining techniques become more widely available, they are bound to become an important method of studying animal brain function, including lateralization.

Pattern of Lateralization in Vertebrates

The pattern of specialization of the left and right hemispheres of vertebrates is broadly similar across species and remarkably similar to that present in humans (summarized by MacNeilage et al., 2009; Rogers et al., 2013b; **Figure 1**). The left hemisphere is specialized to categorize objects and other stimuli, and to control learned or well-practiced behavior when the animal is in a relaxed state. It also focuses attention on particular features of objects or aspects of the environment. Use of the left hemisphere occurs when the animal has to sustain responses, as in pursuing prey or searching for food (e.g., chicks searching for grains, discussed above). It is used when choices have to be made according to categories, such as food type or preferred type of partner, as shown in male zebra finches by Templeton

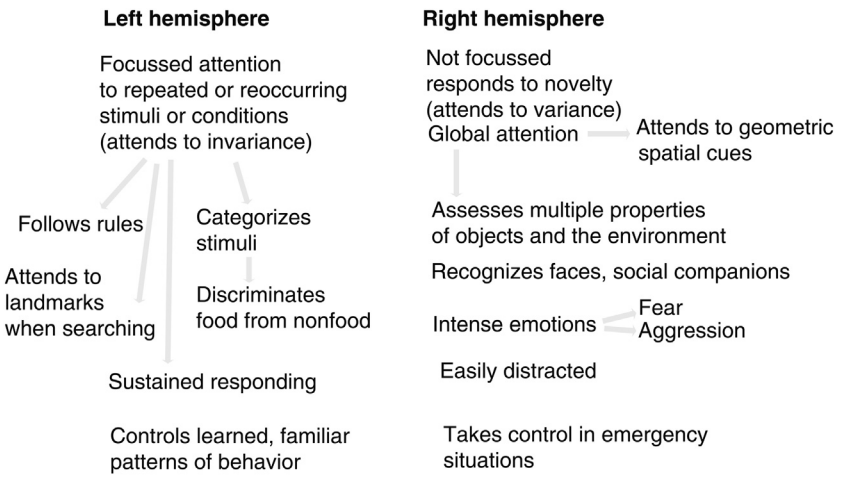

Figure 1 Summary of the different functions of the left and right hemispheres, as determined by testing a range of vertebrate species. The arrows link functions that may be related.

et al. (2014). (Note that use of the right eye and left hemisphere is expressed during *choice* between potential partners but that the right hemisphere controls the actual act of copulation.)

The right hemisphere is specialized for controlling rapid responses given in emergency situations, including attention to and escape from predators. It has a broad, global attention important in detecting novel stimuli. It assesses multiple properties of objects and the surrounding environment. The right hemisphere also processes spatial information relying on geometric cues, recognizes faces, and controls the expression of emotions, particularly of fear and aggression.

As an example, the dichotomy of hemispheric function was illustrated by testing chicks on their spatial ability used to find hidden food (Tommasi and Vallortigara, 2001): when the left eye was occluded (using the right eye and left hemisphere) chicks relied on landmarks to find the food and when the right eye was occluded (using the left eye and right hemisphere) they used geometrical cues (i.e., metric properties such as direction and distance from walls of an arena). Rats have also been shown to use the right hemisphere and left eye for spatial navigation (Cowell et al., 1997).

Humans show similar use of the right hemisphere for processing global patterns of information and the left for attention to detail (summarized in MacNeilage et al., 2009). Faces are also processed in the right hemisphere in humans and animals, as shown clearly in sheep by left visual hemifield superiority in discriminating faces of other sheep (Peirce and Kendrick, 2002). Chicks show a similar specialization of the right hemisphere in that they can discriminate between social companions when they use their right hemisphere but not when they use their left hemisphere (Vallortigara and Andrew, 1994) and the right hemisphere is used in establishing hierarchies, as needed to form their species-typical social order (Daisley et al., 2009).

As in humans, emotions, particularly strong negative ones, are controlled by the right hemisphere. Rats, for example, rely on the right hemisphere to learn the fear response required for aversion of particular tastes (Denenberg et al., 1980). Chicks also express higher levels of emotion when they view fear-inducing stimuli with their left eye (right hemisphere) than they do when they view the same stimuli with their right eye (Andrew, 1983) and a similar right-hemisphere bias to respond more strongly to fear-inducing stimuli has been shown in dogs for stimuli they see (Siniscalchi et al., 2010), those that they hear (Siniscalchi et al., 2008) and those that they smell (Siniscalchi et al., 2011).

The strong negative emotions expressed by the right hemisphere can be suppressed by the left hemisphere, and this is needed to allow pursuit of prey without distraction of attention. In mammals, this suppression of the right by the left hemisphere is affected via the corpus callosum, a set of connections between the hemispheres. In other species it is mediated via commissures, some of which occur at levels of brain organization below the hemispheres (e.g., the tectal and posterior commissures of birds).

It is important to note the popular concept of 'left-brained' or 'right-brained' individuals is incorrect. Both hemispheres are in communication with each other and the balance of control can change frequently depending on external circumstances and internal states. Although individuals might show a trend toward using one hemisphere more often than the other, this is not an exclusive bias.

Evolution of Lateralization in Vertebrates

Lateralization of behavior was already present in the earliest vertebrates (Andrew, 2002; summarized in Rogers et al., 2013b). Ancestral chordates were structurally asymmetrical. An organism still living today, the lancelet, *Branchiostoma*, provides a key to early vertebrate evolution. Its larval form has its mouth located asymmetrically on the left side and the associated structures in the nervous system used for feeding and controlling the mouth are also on the left side. Although asymmetry is no longer present in adult lancelets, the asymmetry of the larval stages seems to have been key to the evolution of lateralization in vertebrates.

As vertebrates evolved, they acquired paired eyes and these eyes, it seems, had differently specialized functions (see chapter 3 in Rogers et al., 2013b): the left side of the brain continued to control detection and capturing of prey and this meant specialization of the right eye to perform this behavior since the right eye sends its inputs to the left side of the brain.

The general pattern of lateralization was already present in early vertebrates and it was conserved throughout vertebrate evolution. Although some substantial modifications may have taken place during the evolution of brain laterality, it was already a firmly established characteristic of the vertebrate brain well before humans evolved. In other words, lateralization in animals was a precursor to the hemispheric specializations of humans (MacNeilage, 2008; MacNeilage et al., 2009). Specialization of the left hemisphere for language and speech in humans seems to be an elaboration of this hemisphere's superiority in sequential processing (Bradshaw and Nettleton, 1981) and for communication, present well before humans evolved (MacNeilage et al., 2009; Meunier et al., 2013; Meguerditchian et al., 2013). Furthermore, it seems that dominance of the left hemisphere for gestural communication was followed by the addition of vocal control and thence to speech and language.

Lateralization in Invertebrates

Recently, researchers have begun to look for evidence of lateralization in invertebrates and some convincing examples are coming to light (reviewed by Frasnelli et al., 2012). In bees, for example, the left and right antennae have different roles in learning and recall of olfactory associations. First, honeybees were found to be able to learn to associate a particular odor with a sugar reward provided they used their right antenna but not when they used their left antenna (Letzkus et al., 2006). Recall of such a memory is initially dependent on the right antenna but then, after a few hours, the memory can be recalled only if the left antenna is used (Rogers and Vallortigara, 2008). The same lateralized access to memory has been shown in other social bees but not in a solitary species of bee (summarized in Frasnelli et al., 2014). This social versus nonsocial difference in lateralization at the population level is consistent with the hypothesis about directional bias and social behavior discussed further below.

Structural asymmetry in the brain also determines the ability of fruit flies, *Drosophila*, to form long-term memories (Pascual et al., 2004). Transition of memory from short- to long-term in this species depends on a structure usually present only in the right side of the brain and close to the region linking the left and right sides of the brain. Some fruit flies have a duplicate of this structure in the left side of their brain (i.e., they are symmetrical) and they are unable to make long-term memories.

Squids provide another example of lateralized behavior in invertebrate species. *Sepia officinalis* uses its left eye to guide escape swimming (Jozet-Alves et al., 2012). It is not known whether the deep-sea squid *Histioteuthis* has the same lateralization of behavior but marked asymmetry in the structure of its eyes would suggest so: the left eye is large and looks upward, presumably to detect predators, and the smaller right eye looks downward, presumably for prey (see Rogers et al., 2013b).

Whether lateralization in invertebrate species is in any way related to lateralized brain function in vertebrates is not known at this time. However, evidence of its presence has opened up new avenues of research on evolution and factors influencing development of lateralization (see Halpern et al., 2014). It remains possible that lateralization in invertebrates evolved quite independently of that in vertebrates but, regardless of whether or not this is the case, evidence suggests that invertebrate brains also benefit from being lateralized.

Development of Lateralization in Vertebrates

Early studies by Collins (1985) on paw preferences in mice demonstrated that, through genetic selection, it is possible to alter the strength of paw preference but not its direction (i.e., whether individuals prefer to use the left or right paw to reach into a tube to obtain food). Other behavioral lateralities appear to be partly genetically determined and partly influenced by environmental factors. In birds, for example, exposure of the eggs to light leads to lateralization of the visual pathways and of visual behavior (summarized by Rogers, 2008). This was shown first in chickens and later in pigeons. In the final stages of incubation avian embryos are positioned in the egg so that their right eye is next to the shell and can be stimulated by light, whereas their left eye is occluded by the embryo's body, thereby preventing light from reaching it. This asymmetry of stimulation by light before hatching initiates the development of visual lateralization. Embryos hatched from eggs incubated in the dark lack visual lateralization (chickens, Rogers, 1982; pigeons, Güntürkün, 2002). As little as 2 h of light exposure of chick embryos during the final days before hatching, when the visual system is becoming functional, is sufficient to generate visual lateralization (Rogers, 1990). In the pigeon, visual lateralization caused by this early effect of light persists into adulthood and it involves connections both to and from the forebrain (Valencia-Alfonso et al., 2009). In chicks, it is known that the effect lasts for the first 3 weeks of life but no studies have examined whether or not it persists in to adulthood. It should be noted, however, that light exposure of the embryo affects lateralization in different pathways of the visual system in chickens and pigeons although the outcome in behavior is quite similar in both species (Deng and Rogers, 2002; Manns and Güntürkün, 2009). Exposure to light during embryonic development also affects communication between the hemispheres (Manns and Römling, 2012).

Light stimulation also has a critical effect on the development of visual lateralization in zebra fish but, in this case, the influence of light occurs earlier in development. Zebra fish fry hatched from eggs exposed to the normal light/dark cycle show lateralization of avoiding a model predator presented on their left side and approaching it when presented on their right side, whereas fry raised in the dark do not show this asymmetry (Budaev and Andrew, 2009). This effect of light differs from that seen in birds because the zebra fish's eyes are not even functional at this early stage of development. Instead, light may activate particular genes important in the development of lateralization. Alternatively, the light-sensitive parapineal organs, which are known to be

asymmetrical in zebra fish (Halpern et al., 2003), may be involved (Budaev and Andrew, 2009).

Development of lateralization in sensory systems other than vision might also be affected by experience in early life but there has been very little research in this area. One example is, however, available and that concerns laterality of behavior in the cuttlefish, *Sepia officinalis*. Laterality in this species is enhanced by exposure of the eggs to the odor of a predator, that of the sea bass (Jozet-Alves and Hérbert, 2013). There is also recent evidence that changing carbon dioxide levels of the sea can affect lateralized turning behavior in a species of coral-reef fish (Domenici et al., 2014): it reverses the direction of lateralization. The relevance of this to potential effects of climate change on behavior deserves further research.

Steroid hormones also play a role in the development of lateralization and one way in which they might achieve this is by interacting with the effects of light on lateralization. Injecting estrogen, testosterone, or corticosterone into the developing egg nullifies the effect of light exposure on lateralization (Rogers and Rajendra, 1993). Whether or not similar processes operate in other species and whether they have any relation to sex differences in lateralization has not yet been investigated, although sex differences in lateralization have been documented in both animals and humans (discussed by Pfannkuche et al., 2009) and frequent claims are made about sex hormones determining these.

Function of Lateralization

Lateralized behavior has been shown in wild animals. Wild black-winged stilts peck at prey on their right side (using the right eye) and show copulatory behavior to conspecifics seen on their left side (Ventolini et al., 2005). Wild Australian magpies view a predator with the left eye before they withdraw from it (Koboroff et al., 2008) and wild horses view with the left eye before they attack (Austin and Rogers, 2012). Wild beluga whales show a preference to have their calf on their left side (right hemisphere) (Karenina et al., 2010). These and other examples show that lateralization of behavior is not simply an inconsequential aspect of brain function revealed only by testing animals in the laboratory. Since it is not obligatory that asymmetries of brain anatomy or neurochemistry will be expressed as side biases in behavior (i.e., they could be masked), these manifestations of laterality in natural behavior are important. They raise intriguing issues about the advantages and disadvantages of being lateralized.

Clearly it must be disadvantageous to have side biases in response to predators since predators may approach on either side and even a slight difference in speed of detection or response could have consequences for survival. Likewise, a similar argument could be made for side biases in responding to prey or other food items. Therefore, it seems logical to expect that these disadvantageous side biases are counterbalanced by advantages of another kind.

By comparing the behavior of animals with and without lateralization, it has been possible to determine some of the benefits of having a lateralized brain. The chick is an excellent model for testing this since visual lateralization can be modified by light exposure before hatching.

One of the hypothesized advantages of having a lateralized brain is to increase cognitive capacity and this might be manifested as more success in performing two or more tasks simultaneously. Results obtained by comparison of chicks with and without visual lateralization support this hypothesis. Chicks hatched from eggs exposed to light are able to monitor overhead for predators (right hemisphere) while they are searching for food (left hemisphere), whereas chicks hatched from eggs incubated in the dark are unable to do this (Rogers et al., 2004). The latter are less likely to detect the predator (a silhouette of an eagle moving overhead) and, when they do see it, they are so distracted by it that they are unable to find food scattered on a background of pebbles. A similar result has been obtained using a species of fish faced with capturing prey while a predatory fish watched on: fish with stronger lateralization performed more efficiently than those with weak lateralization (Dadda and Bisazza, 2006).

Pigeons too perform better in finding food among distracting grit (pebbles) when they are more strongly lateralized (Güntürkün et al., 2000): there is a positive correlation between strength of lateralization and success in pecking at grain. In chimpanzees, stronger hand preference is associated with more efficient 'fishing' for termites by inserting a piece of grass into the termite mound, although in this case the direction of hand preference is irrelevant (McGrew and Marchant, 1999). In line with this finding, a study of Australian parrots has reported that those with stronger lateralization, measured by foot preference for holding food, have better ability to solve cognitive problems (Magat and Brown, 2009).

These results provide evidence of cognitive advantage of individuals being lateralized but they do not explain why the majority of individuals are lateralized in the same direction. This seems to have something to do with social behavior. When a lateralized animal interacts with another, it could be advantageous if they are both lateralized in the same direction (Vallortigara and Rogers, 2005). Stable social hierarchies in chicks depend on this: groups of lateralized chicks form more stable hierarchies than do groups of chicks not so lateralized (Rogers and Workman, 1989).

Highly social honeybees show lateralization of social behavior (Rogers et al., 2013a). When honeybees are put in pairs in containers they may contact in positive social interaction, which includes proboscis extension, or negative social interaction, which includes C-responses as used in stinging. After removal of their left antennae, pairs of bees perform appropriate social behavior (i.e., proboscis extension if they are from the same hive and C-responses if they are from different hives). Pairing of bees after removal of their right antennae leads to inappropriate social behavior: same-hive pairs are slow to contact and they perform more C-responses and fewer proboscis extension responses. This finding demonstrates the importance of population-level lateralization in social behavior.

Conclusion

Lateralization of brain and behavior is widespread among vertebrate species and it is also present in invertebrates. Hence,

it does not depend on brain size or the overall neural capacity of a species and seems to be beneficial to all, or most, brains, possibly by enhancing the ability of individuals within a species to solve problems, form memories, and to perform more than one task simultaneously. In essence, it allows parallel processing of incompatible tasks requiring very different neural processing.

The left hemisphere is specialized to focus attention and categorize stimuli, such as required in choosing food (prey). The right hemisphere is used in emergency situations when rapid responses are required and it has a special role in stress responses, expression of intense emotions, and detection of novelty.

Individual- and population-level lateralization can be generated or modulated by experience in early life. Evidence for this comes mainly from research on avian embryos and in the visual modality. However, a limited number of examples, in mammals and for other sensory modalities indicate that sensory input may influence lateralization more generally. This provides a means by which lateralization might be adapted to specific environmental conditions. Stress in early life might, for example, influence the development of lateralization either directly by changing the embryo's hormone levels or indirectly by altering maternal behavior (e.g., by varying the amount of light exposure during a sensitive period of visual development).

Now that the idea of brain lateralization as a trait only of humans has been disproven, research is increasingly using animal models to address questions about the development, causation, function, and evolution of lateralization. This has implications for understanding not only animal behavior but also human behavior, particularly in cases of dysfunction that seem to be associated with reduced or reversed lateralization (e.g., depression, autism, or dyslexia).

See also: Asymmetry of Body and Brain; Brain Asymmetry; Interhemispheric Interaction in Language and Cognitive Processes; Lateralization of Language as Demonstrated by Brain Imaging Procedures.

Bibliography

Andrew, R.J., 1983. Lateralization of emotional and cognitive function in higher vertebrates, with special reference to the domestic chick. In: Ewert, J.P., Capranica, R.R., Ingle, D.J. (Eds.), Advances in Vertebrate Neuroethology. Oxford University Press, Oxford, pp. 477–510.

Andrew, R.J., 2002. The earliest origins and subsequent evolution of lateralization. In: Rogers, L.J., Andrew, R.J. (Eds.), Comparative Vertebrate Lateralization. Cambridge University Press, Cambridge, pp. 70–93.

Austin, N.P., Rogers, L.J., 2007. Asymmetry of flight and escape turning responses in horses. Laterality 12, 464–474.

Austin, N.A., Rogers, L.J., 2012. Limb preferences and lateralization of aggression, reactivity and vigilance in feral horses (*Equus caballus*). Animal Behaviour 83, 239–247.

Bonati, B., Csermely, D., López, P., Martin, J., 2010. Lateralization of escape behaviour in the common wall lizard (*Podarcis muralis*). Behavioural Brain Research 207, 1–6.

Bradshaw, J.L., Nettleton, N.C., 1981. The nature of hemispheric specialization in man. Behavioural Brain Research 4, 51–63.

Bradshaw, J.L., Rogers, L.J., 1993. The Evolution of Lateral Asymmetries, Language, Tool Use and Intellect. Academic Press, San Diego.

Budaev, S., Andrew, R.J., 2009. Patterns of early embryonic light exposure determine behavioural asymmetries in zebrafish: a habenular hypothesis. Behavioural Brain Research 200, 91–94.

Collins, R.L., 1985. On the inheritance of the direction and the degree of asymmetry. In: Glick, S.D. (Ed.), Cerebral Lateralization in Nonhuman Species. Academic Press, New York, pp. 41–71.

Corballis, M.C., 1991. The Lopsided Brain. Oxford University Press, New York.

Cowell, P.E., Waters, N.S., Denenberg, V.H., 1997. The effects of early environment on the development of functional laterality in Morris maze performance. Laterality 2, 221–232.

Dadda, M., Bisazza, A., 2006. Does brain asymmetry allow efficient performance of simultaneous tasks? Animal Behaviour 72, 523–529.

Daisley, J.N., Mascalzoni, E., Rosa-Salva, O., Rugani, R., Regolin, L., 2009. Lateralization of social cognition in the domestic chicken (*Gallus gallus*). Philosophical Transactions of the Royal Society B 364, 965–981.

Denenberg, V.H., 1981. Hemispheric laterality in animals and the effects of early experience. Behavioral and Brain Sciences 4, 1–49.

Denenberg, V.H., 1984. Behavioural asymmetry. In: Geschwind, N., Galaburda, A.M. (Eds.), Cerebral Dominance: The Biological Foundations. Harvard University Press, Cambridge, pp. 114–133.

Denenberg, V.H., Garbanati, J.A., Sherman, G., Yutzey, D.A., Kaplan, R., 1978. Infantile stimulation induces brain lateralization in rats. Science 201, 1150–1152.

Denenberg, V.H., Hofmann, M., Garabanati, J.A., Sherman, G.F., Rosen, G.D., Yutzey, D.A., 1980. Handling in infancy, taste aversion, and brain laterality in rats. Brain Research 200, 123–133.

Deng, C., Rogers, L.J., 2002. Factors affecting the development of lateralization in chicks. In: Rogers, L.J., Andrew, R.J. (Eds.), Comparative Vertebrate Lateralization. Cambridge University Press, Cambridge, pp. 206–246.

Domenici, P., Allan, B.J.M., Watson, S.-A., McCormick, M.I., Munday, P.L., 2014. Shifting from right to left: the combined effect of elevated CO_2 and temperature on behavioural lateralization in a coral reef fish. PLoS ONE 9 (1), e87969.

Frasnelli, E., Vallortigara, G., Rogers, L.J., 2012. Left-right asymmetries of behaviour and nervous system in invertebrates. Neuroscience and Biobehavioral Reviews 36, 1273–1291.

Frasnelli, E., Haase, A., Rigosi, E., Anfora, G., Rogers, L.J., Valloritgara, G., 2014. The bee as a model to investigate brain and behavioural asymmetries. Insects 5, 120–138.

Güntürkün, O., Diekamp, B., Manns, M., Nottlemann, F., Prior, H., Schwarz, A., Skiba, M., 2000. Asymmetry pays: visual lateralization improves discrimination success in pigeons. Current Biology 10, 1079–1081.

Güntürkün, O., 2002. Ontogeny of visual asymmetry in pigeons. In: Rogers, L.J., Andrew, R.J. (Eds.), Comparative Vertebrate Lateralization. Cambridge University Press, Cambridge, pp. 247–273.

Halpern, M.E., Güntürkun, O., Hopkins, W.D., Rogers, L.J., 2005. Lateralization of the vertebrate brain: taking the side of model systems. Journal of Neuroscience 9, 10351–10357.

Halpern, M.E., Hobert, O., Wright, C.V.E. (Eds.), 2014. Left-Right Asymmetry: Advances and Enigmas. genesis, The Journal of Genetics and Development, vol. 52 (6). Wiley, New Jersey.

Halpern, M.E., Liang, J.O., Gamse, J.T., 2003. Leaning to the left: laterality in the zebrafish forebrain. Trends in Neurosciences 26, 308–313.

Hamilton, C.R., Vermeire, B.A., 1988. Complementary hemispheric specialization in monkeys. Science 242, 1691–1694.

Harris, L.J., 1989. Footedness in parrots: three centuries of research, theory, and mere surmise. Canadian Journal of Psychology 43, 369–396.

Hobert, O., 2014. Development of left/right asymmetry in the *Caenorhabitis elegans* nervous system: from zygote to postmitotic neuron. Genesis 52, 528–543.

Hopkins, W.D., Phillips, K.A., Bania, A., Calcutt, S.E., Gardiner, M., Russell, J., Schaeffer, A., Lonsdorf, E.V., Ross, S.R., Shapiro, S.J., 2011. Hand preferences for coordinated bimanual actions in 777 great apes: implications for the evolution of handedness in Hominins. Journal of Human Evolution 60, 605–611.

Hopkins, W.D., Wesley, M.J., Izard, M.K., Hook, M., Schapiro, S.J., 2004. Chimpanzees are predominantly right-handed: replication in three colonies of apes. Behavioral Neuroscience 118, 659–663.

Jozet-Alves, C., Romagny, S., Bellanger, C., Dickel, L., 2012. Cerebral correlates of visual lateralization in *Sepia*. Behavioural Brain Research 234, 20–25.

Jozet-Alves, C., Hébert, M., 2013. Embryonic exposure to predator odour modulates visual lateralization in cuttlefish. Proceedings of the Royal Society B 280, 2012–2575.

Kaplan, G., Rogers, L.J., 2013. Stability of referential signalling across locations: testing alarm calls of Australian magpies (*Gymnorhina tibicen*) in urban and rural Australia and in Fiji. PeerJ 1, 5112. http://dx.doi.org/10.7717/peerj.112.

Karenina, K., Giljov, A., Baranov, V., Osipova, L., Krasnova, V., Malashichev, Y., 2010. Visual laterality of calf-mother interactions in wild whales. PLoS ONE 5, e13787.

Koboroff, A., Kaplan, G., Rogers, L.J., 2008. Hemispheric specialization in Australian magpies (*Gymnorhina tibicen*) shown as eye preferences during response to a predator. Brain Research Bulletin 76, 304–306.

Letzkus, P., Ribi, W.A., Wood, J.T., Zhu, H., Zhang, S.W., Srinivasan, M.V., 2006. Lateralization of olfaction in the honeybee *Apis mellifera*. Current Biology 16, 1471–1476.

Lippolis, G., Bisazza, A., Rogers, J., Vallortigara, G., 2002. Lateralization of predator avoidance responses in three species of toads. Laterality 7, 163–183.

Lippolis, G., Westman, W., McAllan, B.M., Rogers, L.J., 2005. Lateralization of escape responses in the striped-faced dunnart, *Sminthopsis macroura* (Dasyuridae: Marsupalia). Laterality 10, 457–470.

MacNeilage, P.F., 2008. The Origin of Speech. Oxford University Press, Oxford.

MacNeilage, P., Rogers, L.J., Vallortigara, G., 2009. Origins of the left and right brain. Scientific American 301, 60–67.

MacNeilage, P.F., Studdert-Kennedy, M.J., Lindblom, B., 1987. Primate handedness reconsidered. Behavioral and Brain Sciences 10, 247–303.

Magat, M., Brown, C., 2009. Laterality enhances cognition in Australian parrots. Proceedings of the Royal Society B 276, 4155–4162.

Manns, M., Güntürkün, O., 2009. Dual coding of visual asymmetries in the pigeon brain: the interaction of bottom-up and top-down systems. Experimental Brain Research 199, 323–332.

Manns, M., Römling, J., 2012. The impact of asymmetrical light input on cerebral hemispheric specialization and hemispheric cooperation. Nature Communications 3, 696.

McGrew, W.C., Marchant, L.F., 1999. Laterality of hand use pays off in foraging success for wild chimpanzees. Primates 40, 509–513.

Meguerditchian, A., Vauclair, J., Hopkins, W.D., 2013. On the origins of human handedness and language: a comparative review of hand preferences for bimanual coordinated actions and gestural communication in nonhuman primates. Developmental Psychobiology 55, 637–650.

Mench, J., Andrew, R.J., 1986. Lateralization of a food search task in the domestic chick. Behavioral and Neural Biology 46, 107–114.

Meunier, H., Fagard, J., Maugard, A., Briseno, M., Fizet, J., Canteloup, C., Defolie, C., Vauclair, J., 2013. Patterns of hemispheric specialization for a communicative gesture in different primate species. Developmental Psychobiology 55, 662–671.

Nottebohm, F., Stokes, T.M., Leonard, C.M., 1976. Central control of song in the canary, *Serinus canarius*. Journal of Comparative Neurology 165, 457–486.

Pascual, A., Huang, K.-L., Nevue, J., Préat, T., 2004. Brain asymmetry and long-term memory. Nature 427, 605–606.

Peirce, J.W., Kendrick, K.M., 2002. Functional asymmetry in sheep temporal cortex. Neuroreport 13, 2395–2399.

Pfannkuche, K.A., Bouma, A., Groothuis, T.G.G., 2009. Does testosterone affect lateralization of brain and behaviour? A meta-analysis in humans and other animal species. Philosophical Transactions of the Royal Society B 364, 929–942.

Rogers, L.J., 1980. Lateralization in the avian brain. Bird Behavior 2, 1–12.

Rogers, L.J., 1982. Light experience and asymmetry of brain function in chickens. Nature 297, 223–225.

Rogers, L.J., 1990. Light input and the reversal of functional lateralization in the chicken brain. Behavioural Brain Research 38, 211–221.

Rogers, L.J., 2002. Lateralization in vertebrates: its early evolution, general pattern and development. In: Slater, P.J.B., Rosenblatt, J., Snowdon, C., Roper, T. (Eds.), Advances in the Study of Behavior, vol. 31. Academic Press, San Diego, pp. 107–162.

Rogers, L.J., 2008. Development and function of lateralization in the avian brain. Brain Research Bulletin 76, 235–244.

Rogers, L.J., 2009. Hand and paw preferences in relation to the lateralized brain. Philosophical Transactions of the Royal Society B 364, 943–954.

Rogers, L.J., Andrew, R.J., 2002. Comparative Vertebrate Lateralization. Cambridge University Press, New York.

Rogers, L.J., Anson, J.M., 1979. Lateralization of function in the chicken fore-brain. Pharmacology Biochemistry and Behavior 10, 679–686.

Rogers, L.J., Rajendra, S., 1993. Modulation of the development of light-initiated asymmetry in chick thalamofugal visual projections by oestradiol. Experimental Brain Research 93, 89–94.

Rogers, L.J., Vallortigara, G., 2008. From antenna to antenna: lateral shift of olfactory memory recall by honeybees. PLoS ONE 3 (6), e2340.

Rogers, L.J., Workman, L., 1989. Light exposure during incubation affects competitive behaviour in domestic chicks. Applied Animal Behaviour Science 23, 187–198.

Rogers, L.J., Rigosi, E., Frasnelli, E., Vallortigara, G., 2013a. A right antenna for social behaviour. Nature Communications 3, 2045.

Rogers, L.J., Vallortigara, G., Andrew, R.J., 2013b. Divided Brains: The Biology and Behaviour of Brain Asymmetries. Cambridge University Press, Cambridge.

Rogers, L.J., Zucca, P., Vallortigara, G., 2004. Advantage of having a lateralized brain. Proceedings of the Royal Society of London. Series B 271, S420–S422.

Siniscalchi, M., Quaranta, A., Rogers, L.J., 2008. Hemispheric specialization in dogs for processing different acoustic stimuli. PloS ONE 3 (10), e3349.

Siniscalchi, M., Sasso, R., Pepe, A.M., Dimatteo, S., Vallortigara, G., 2011. Sniffing with the right nostril: lateralization of response to odour stimuli by dogs. Animal Behaviour 82, 399–404.

Siniscalchi, M., Sasso, R., Pepe, A.M., Vallortigara, G., Quaranta, A., 2010. Dogs turn left to emotional stimuli. Behavioural Brain Research 208, 516–521.

Taglialatela, J.P., Russell, J.L., Schaeffer, J.A., Hopkins, W.D., 2008. Communicative signalling activates 'Broca's' homolog in chimpanzees. Current Biology 18, 343–348.

Templeton, J.J., McCracken, B.G., Sher, M., Mountjoy, D.J., 2014. An eye for beauty: lateralized visual stimulation of courtship behavior and mate preferences in male zebra finches, *Taeniopygia guttata*. Behavioural Processes 102, 33–39.

Tommasi, L., Vallortigara, G., 2001. Encoding of geometric and landmark information in the left and right hemispheres of the avian brain. Behavioral Neuroscience 115, 602–613.

Valencia-Alfonso, C.E., Verhaal, J., Güntürkün, O., 2009. Ascending and descending mechanisms of visual lateralization in pigeons. Philosophical Transactions of the Royal Society B 364, 955–963.

Vallortigara, G., Andrew, R.J., 1994. Differential involvement of right and left hemisphere in individual recognition in the domestic chick. Behavioural Processes 33, 41–58.

Vallortigara, G., Rogers, L.J., 2005. Survival with an asymmetrical brain: advantages and disadvantages of cerebral lateralization. Behavioral and Brain Sciences 28, 575–633.

Ventolini, N., Ferrero, E.A., Sponza, S., Chiesa, A.D., Zucca, P., Vallortigara, G., 2005. Laterality in the wild: preferential hemifiled use during predatory and sexual behaviour in the black-winged stilt. Animal Behaviour 69, 1077–1084.

Warren, J.M., 1980. Handedness and laterality in humans and other animals. Physiological Psychology 8, 351–359.

Brain Asymmetry

Sebastian Ocklenburg, Ruhr-Universität Bochum, Bochum, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E. Zaidel, volume 2, pp. 1321–1329, © 2001, Elsevier Ltd.

Abstract

Structural and functional asymmetries between the left and the right hemisphere of the brain are ubiquitous in the vertebrate brain. In this article, a general introduction into brain asymmetry is given, followed by a discussion of handedness and other forms of behavioral preferences. Moreover, cognitive domains for which brain asymmetry has been reported are described in addition to structural brain asymmetries. The ontogenesis of brain asymmetry as well as the advantages of having an asymmetric brain is discussed. The article concludes with a description of factors modulating brain asymmetry, most notably sex and pathological processes.

Definition of the Term

The term brain asymmetry refers to structural, functional, or physiological asymmetries between the left and the right hemisphere of the brain. The hemisphere in which the relevant parts of the brain are relatively larger, superior in performance, or more active during a given task is called dominant. Brain asymmetry is a continuous phenomenon, with considerable interindividual differences in strength and direction, and it is common that for a given function or structure, a minority of individuals show the so-called atypical asymmetry (Van der Haegen et al., 2013), that is, a reversal of the direction of the asymmetry found on population level. This phenomenon can either be due to cortical reorganization processes after pathological events or due to an ontogenetic predisposition. It is also important to note that brain asymmetry is usually relative: While one hemisphere is more active or superior in performance than the other, it does not indicate that the subdominant hemisphere is not active at all. Essentially, three forms of brain asymmetry can be distinguished (Palmer, 1996; Vallortigara et al., 1999):

1. Absence of asymmetry (sometimes also called symmetry): All or most members of the population do not show a left- or right-hemispheric dominance.
2. Individual-level asymmetry (sometimes also called antisymmetry): Some individuals show a significant left-hemispheric dominance while others a significant right-hemispheric dominance, but there is no significant asymmetry at the population level.
3. Population-level asymmetry (sometimes also called directional asymmetry): The majority of the population shows a hemispheric dominance to one side.

While brain asymmetry has initially been discovered in humans and was thought to be uniquely human for several decades, it is now considered a general principle of vertebrate brain organization (Ocklenburg and Güntürkün, 2012; Rogers, 2009) and has been reported in dozens of different vertebrate species (Bisazza et al., 1998; Halpern et al., 2005; Ocklenburg et al., 2013b). Notable examples for non-human asymmetries have been found in the songbird song system (George, 2010), the zebrafish habenula (Bianco and Wilson, 2009), as well as in the visual systems of chickens (Rogers, 1996), and pigeons (Manns and Güntürkün, 2009). Interestingly, left–right asymmetries of behavior and nervous system have recently also been shown in several invertebrate species, e.g., honey bees, ants, and octopuses (Frasnelli et al., 2012).

The scientific interest in hemispheric asymmetries started with a seminal discovery by a French physician called Pierre Paul Broca, who published about a patient called Monsieur Tan because the only word he was able to utter was 'tan' (Broca, 1861). Postmortem analysis of the brain of this massively speech-impaired patient revealed a lesion in the left frontal lobe, demonstrating that the left hemisphere and especially an area in the left posterior inferior frontal gyrus now known as Broca's area is highly relevant for language production. After Broca's initial discovery, a dominance of the left hemisphere for many aspects of speech production and perception was demonstrated by several researchers. However, it was unclear for a long time whether the right hemisphere had any complementary specialization, until it was shown that in most individuals it is dominant for several aspects of spatial processing (Vogel et al., 2003). Subsequently, several different functional and structural asymmetries have been identified in humans and other vertebrates, all of which are discussed in this article. Starting in the 1960s, research in split brain patients and callosotomized animal models allowed for deeper insights into complementary hemispheric specialization and how the disconnected cerebral hemispheres work independently of each other (e.g., Sperry, 1961; Gazzaniga, 2005; *see* Split Brains and Human Consciousness for further details). These experiments in turn gave new impulses to asymmetry research in healthy participants, which, over the last decades, has been conducted with an impressive array of different neuroscientific approaches. Today, identifying the molecular and neurophysiologic mechanisms underlying brain asymmetry remains one of the key questions of modern cognitive neuroscience. Cerebral specialization has been reported for a large number of vertebrate brain structures and functions, ranging from basic perceptual systems in invertebrates (Hobert et al., 2002) to several higher cognitive functions in humans (Corballis, 2012). This article will focus

on these functional domains, as well as on the different techniques used to investigate lateralization within these cognitive systems and relevant brain structures and the phylogenetic and ontogenetic mechanisms underlying its development.

Handedness and Other Forms of Behavioral Preferences

The probably most salient form of brain asymmetry in everyday life is left- or right-handedness. About 90% of the population is right-handed, while the remaining 10% is left-handed (Grimshaw and Wilson, 2013), making human handedness the most pronounced behavioral left–right preference in vertebrates. This clear population-level asymmetry in hand use is not paralleled by any distinct morphological asymmetries between the two hands. This indicates that handedness is a manifestation of varying skills between the two hands and reflects a brain asymmetry in the motor cortex (Corballis, 2009). The easiest methodological approach to determine an individual's handedness is a dichotomous differentiation between left- and right-handers based on an assessment of hand preference for writing or drawing. However, this method can lead to erroneous results under some conditions, e.g., if left-handers who have been forced to learn to write with the right hand are assessed (Klöppel et al., 2010). Thus, several other methods to assess handedness have been developed, including observational methods (Bryden et al., 2007), assessment of handwriting speed (Provins and Magliaro, 1993), the pegboard test (Annett, 2002), as well as a number of different questionnaires that assess hand use during different activities (e.g., Annett, 1970; Bryden, 1977; Crovitz and Zener, 1962; Oldfield, 1971; Peters and Servos, 1989; Siefer et al., 2003). The most frequently used of these instruments is the Edinburgh Handedness Inventory (EHI) (Oldfield, 1971), a 10-item questionnaire, in which participants have to indicate their preferred hand for different everyday activities (e.g., writing, drawing, throwing, and striking a match). Based on the subjects' answers, a laterality quotient (LQ) ranging between −100 (consistently left handed) and +100 (consistently right handed) can be calculated, with positive values indicating a right-sided preference and negative values a left-sided preference. Based on the individual LQ values, subjects can then be grouped into different categories, e.g., subjects with an LQ larger than zero are considered right handed and participants with an LQ smaller than zero are considered left handed. Besides this simple model, there are also different categorization schemes in which subjects with intermediate LQ values are considered inconsistently handed or ambidextrous (e.g., Isaacs et al., 2006; Tzourio-Mazoyer et al., 2010). While humans show a clear population-level asymmetry in handedness, the situation is less clear for non-human primates. Great apes exhibit individual-level asymmetry in handedness and evidence for population-level asymmetries has been reported by several studies (Hopkins, 2006; Hopkins and Cantalupo, 2005; Hopkins et al., 2011) but there is also evidence arguing against this view (McGrew and Marchant, 2001; Palmer, 2002). These contradicting results indicate that handedness in primates is modulated by environmental factors to a larger extent than is human handedness (Ströckens et al., 2013).

In addition to handedness, behavioral preferences have also been identified for several other motor and sensory systems. For all of these, the majority of the population exhibit a right-sided preference, but typically with a somewhat lower rate than the 90% observed for handedness. For example, infants show a preference to turn their head to the right relative to their body (Hopkins et al., 1987; Ververs et al., 1994), and a right-sided head turning preference can also be observed in adults, e.g., during kissing (Güntürkün, 2003; Ocklenburg and Güntürkün, 2009). Moreover, approximately 80% of amateur soccer players show a preference to use the right foot when kicking the ball (Carey et al., 2009). This preference called footedness can be quantitatively assessed using the Waterloo Footedness Questionnaire (Elias and Bryden, 1998; Elias et al., 1998), a short questionnaire that allows for the calculation of an LQ similar to the EHI.

In addition to motor preferences as handedness or footedness, preferences in the usage of sensory organs have been observed for eyes and ears. For example, ear preference can be assessed by asking which ear an individual would use when listening to a conversation going on behind a closed door and it has been shown that about 65% of the population preferably uses the right ear in this context (Reiss and Reiss, 1999). A similar distribution has also been observed for eye preference which can be assessed by asking subjects to indicate which eye they would prefer when looking through a keyhole. Here, it has been reported that about 66% of the population prefer the right eye for this and similar activities (Reiss and Reiss, 1997). The question whether the different motor and sensory preferences have a common ontogenetic background, or not, has not been answered unequivocally. While some studies report strong positive correlations between handedness, footedness, ear preference, and eye preference (Dittmar, 2002), others only found an association between handedness and footedness, but do no significant association with lateral preference for eye or ear use suggesting that the ontogenesis of lateral preference in sensory and motor systems may at least partly rely on independent processes (Reiss et al., 1999).

Apart from this left–right biases in the use of bilaterally paired motor or sensory systems, several behavioral left–right preferences in interaction with the environment have been reported. For example, subjects tend to bump significantly more often into the right than into the left side of a narrow doorway when walking through it (Nicholls et al., 2007) or using an electric wheelchair to drive through it (Nicholls et al., 2010). Moreover, healthy subjects show a right-sided seating preference in cinemas (Karev, 2000).

Functional Hemispheric Asymmetries

The Left Hemisphere

While there are parts of language processing that are conducted by the right hemisphere (e.g., paralinguistic aspects; Lindell, 2006), most aspects of speech production and perception are mediated by the left hemisphere. Since Broca's initial discovery, several different methods have been applied to

determine the language-dominant hemisphere. For example, anesthesia of the left hemisphere by injecting sodium amorbital into the left carotid artery (a procedure called Wada test; Baxendale, 2009) leads to severe language production impairments in most subjects, thus indicating a language dominance of the left hemisphere just like in Broca's patient. However, since this procedure is too invasive to be used in basic research, several other methods have been developed to determine the language-dominant hemisphere. One of the most widely used behavioral tasks to do so is the dichotic listening technique (Kimura, 2011; Tervaniemi and Hugdahl, 2003; *see* Dichotic Listening and Language: Historical Review for an in-depth discussion of this paradigm). In this experimental paradigm, subjects wear headphones and two auditory stimuli (e.g., meaningless syllables like 'ba' or 'ga' or words) are presented simultaneously, one to the left and the other one to the right ear. Subjects are instructed to indicate (e.g., by pressing a button or by verbal response) which stimulus they heard. Most individuals report the stimulus that was presented to the right ear more often than the one that was presented to the left ear (the so-called right ear advantage), which is thought to reflect a dominance of the left hemisphere for the processing of verbal stimuli (Bryden, 1988). Another commonly used behavioral performance measure to determine brain asymmetry is the visual half-field technique. This method is based on the principle that visual input from the right visual hemifield is processed by the left hemisphere, while input from the left visual hemifield is processed by the right hemisphere (Kolb and Whishaw, 2003). By presenting stimuli tachistoscopically (very briefly, e.g., less than 200 ms) either within the right or the left visual hemifield it is ensured that the stimulus is only processed by the contralateral hemisphere, and that no interhemispheric transfer via the corpus callosum occurs during this process. Paradigms utilizing this technique have been developed for several different types of asymmetrically processed visual stimuli (e.g., spatial stimuli and faces). For example, language dominance can be determined using a lexical decision task. In this task, a word is presented in the center of the visual field, which is subsequently followed by another word presented either in the left or right visual hemifield. Subjects then have to indicate whether the two words were identical or not and on average a right visual field/left hemisphere advantage is observed (Hausmann and Güntürkün, 2000). In addition to behavioral paradigms, language dominance has been assessed with several different neuroimaging techniques. For example, a noninvasive and well-validated alternative to the Wada test is functional magnetic resonance imaging (fMRI) (Binder, 2011). Bethmann et al. (2007) recorded fMRI during a semantic language task and compared activations in several language-relevant regions of interest between the left and right hemisphere, including the inferior frontal sulcus, the inferior frontal gyrus, and the superior temporal sulcus. They observed stronger activations in the left compared to the right hemisphere in 92% of right-handers and similar numbers have also been obtained in studies measuring language lateralization with functional transcranial Doppler sonography during a word generation (Flöel et al., 2005; Knecht et al., 2000). Moreover, left-hemispheric language dominance has also been shown using electroencephalography (Sandmann et al., 2007) and magnetoencephalography (Della Penna et al., 2007). On the electrophysiological level, especially the N1 event-related potential (ERP) has been related to brain asymmetry for language. The N1 is the earliest ERP component reflecting recognition of verbal stimuli, and it has been found to be involved with the automatic lexical classification of words (Spironelli and Angrilli, 2009). In tasks using visually presented words as stimuli, a larger amplitude of the N1 has been observed for occipitotemporal brain areas in the left hemisphere than for the same areas in the right hemisphere (Spironelli and Angrilli, 2007).

The Right Hemisphere

The right hemisphere was considered being the 'inferior hemisphere' as recent as in the 1970s by some researchers (e.g., Lebrun, 1974), due to the fact that it was not clear whether it had any complementary specializations comparable to the left hemisphere's language dominance. Contrasting this view, a relative dominance of the right hemisphere has now been identified for a number of cognitive domains. For example, a strong right-hemispheric advantage has been reported for several visuospatial tasks (Vogel et al., 2003). On the behavioral level, brain asymmetry for visuospatial processing can be investigated with a figural comparison task using the visual half-field technique (Hausmann et al., 2002; Hausmann and Güntürkün, 1999, 2000). At the beginning of each trial, an irregular polygon is presented in the center of the visual field. Subsequently, a second stimulus is presented either in the left or right visual field. Subjects have to indicate whether the two stimuli were the same or not and on average a left visual field/right hemisphere advantage is observed in this task. In addition to behavioral indicators for a right-hemispheric dominance for spatial processing, fMRI studies indicate right-hemispheric brain activation asymmetries during spatial processing (Jansen et al., 2004) and spatial working memory (Nagel et al., 2013). Related to spatial processing is the pseudoneglect, the tendency of healthy subjects to exhibit a leftward bias when they are asked to mark the center of a horizontal line (the so-called line bisection task; Jewell and McCourt, 2000). This leftward bisection bias is thought to be due to right-hemispheric dominance for spatial processing which leads to a systematic overrepresentation of the left compared to the right visual half-field. A cognitive domain related to visuospatial processing for which a right-hemispheric dominance has consistently been found on the behavioral level is face processing (e.g., Godard and Fiori, 2010; Hausmann and Güntürkün, 1999). These findings are supported by neuroimaging studies that reported a stronger activation in the right fusiform face area compared to the same region in the left hemisphere during face recognition (Minnebusch et al., 2009; Rossion et al., 2000) as well as by electrophysiological studies investigating the N170, an ERP component related to face processing that shows higher amplitudes in the right hemisphere (George et al., 1996). Other cognitive domains for which a right-hemispheric dominance has been suggested include the perception of socially relevant stimuli (Brancucci et al., 2009), self-awareness, theory of mind and deception (Keenan et al.,

2005), as well as creative thinking (Mihov et al., 2010). Moreover, the processing of both positive and negative emotions (Borod et al., 1988; Demaree et al., 2005) or only negative emotions (Silberman and Weingartner, 1986) has been suggested to be cognitive functions with a right-hemispheric dominance.

Structural Brain Asymmetries

The scientific interest in structural brain asymmetries started with the seminal discovery that the left planum temporale, a region on the posterior surface of the temporal lobe which includes parts of Wernicke's area is considerably larger than the right one in the majority of the population (Geschwind and Levitsky, 1968). Three subtypes of structural brain asymmetries can be distinguished (for an overview see, Amunts, 2010): macroscopic asymmetries such as in shape or overall volume of a specific cortical area or pathway, microstructural asymmetries such as in the density or number of neurons within an area, and molecular asymmetries such as in gene expression. Besides the leftward asymmetry in the planum temporale, other macroscopic gray matter asymmetries have, for example, been observed in the Sylvian fissure, Heschls gyrus, Brocas area, as well as several sensory and motor regions, the limbic system, and several occipital and parietal areas involved in visuospatial processing (Amunts, 2010). Besides structural gray matter asymmetries, several left–right asymmetries in white matter pathways have been described (Catani, 2010). For example, relative fiber density of the arcuate fasciculus, a major white matter pathway that, besides other structures, connects Broca's and Wernicke's areas (Barrick et al., 2007), shows a pronounced leftward asymmetry in the most individuals. This leftward structural brain asymmetry was positively correlated with the strength of brain activation asymmetries during a verbal fluency task in right- but not left-handers (Vernooij et al., 2007) and has also been related to verbal recall ability (Catani et al., 2007). Another major pathway that has been found to be asymmetric is the uncinate fasciculus, a pathway connecting the anterior temporal lobe with inferior frontal cortex (Rodrigo et al., 2007; Yin et al., 2013); several more subtle structural asymmetries in smaller white matter pathways have also been reported (Barrick et al., 2007). On the molecular level, several genes that are asymmetrically expressed in the left and right hemispheres have been identified, such as LMO4 or DAPPER1 (Sun and Walsh, 2006). While the functional significance of this asymmetrical expression patterns is largely unclear, these genes constitute interesting candidate genes for different other forms of brain asymmetry.

Ontogenesis of Brain Asymmetry

The most advanced understanding on how brain asymmetry develops has probably been achieved in the zebrafish, in which a nodal signaling pathway has been identified that regulates neuroanatomical asymmetries in the habenular nuclei (Concha et al., 2000). In humans, both structural (Rentería, 2012) and functional asymmetries (Francks et al., 2007) are considered to be multifactorial brain phenotypes that are determined by several, possibly interacting, genetic and nongenetic factors. The vast majority of research on the ontogenesis of human brain asymmetry has been conducted regarding handedness. Genetic factors that have been related to handedness include the androgen receptor gene AR (Medland et al., 2005), LRRTM1 (Francks et al., 2007), and PCSK6 (Scerri et al., 2011). In addition, nongenetic factors like social modulation or perinatal asymmetrical perception induced by asymmetrical head position or parental care have been discussed to be relevant for individual handedness (Schaafsma et al., 2009). Apart from handedness not much research on the genetics base of other forms of brain asymmetry has been conducted, yet. There are two recent studies linking genetic variation in the *N*-methyl-D-aspartate receptor 2B subunit gene *GRIN2B* (Ocklenburg et al., 2011) and the cholecystokinin A receptor gene *CCKAR* (Ocklenburg et al., 2013a) to language lateralization measured with the dichotic listening task, but clearly a lot more research on the genetic and environmental factors determining language lateralization and all other forms of brain asymmetry is still needed.

Advantages of Brain Asymmetry

Since brain asymmetry is such a common phenomenon in vertebrates, several ideas regarding its evolutionary origin and advantage for survival have been brought forward. For example, Rogers et al. (2004) suggested that brain asymmetry is advantageous because it enhances the ability to perform two different tasks simultaneously. Also, a functional incompatibility between the logical demands associated with novelty detection and object categorization has been brought forward as a possible factor leading to brain asymmetry (Vallortigara et al., 1999), as well as an avoidance of unnecessary processing conflicts between the two hemispheres in general (Levy, 1969). Other suggested advantages include an increase in neural capacity due to an avoidance of unnecessary duplication of neural networks (Vallortigara, 2006) and the greater speed of uni- compared to bi-hemispheric processing since no interhemispheric transfer via the corpus callosum is needed (Ringo et al., 1994).

Brain Asymmetry and Sex

Various forms of brain asymmetries have been found to be sexually dimorphic, with several studies reporting that women are less lateralized then men (Draca, 2010). This sex difference is modulated by steroid hormones, as studies investigating brain asymmetry across the menstrual cycle reported that lateralization in naturally cycling women is decreased when the progesterone level is high (Hausmann et al., 2002; Hausmann and Güntürkün, 1999, 2000). Interestingly, a recent comprehensive analysis of over 1700 dichotic listening datasets (Hirnstein et al., 2013) reported that sex differences in language asymmetry as assessed by dichotic listening are age-dependent and rather small: While male

young adults showed greater asymmetry than female young adults, there were no sex differences in children and older adults. Also, this behavioral sex difference was not reflected by an fMRI activation asymmetry, as an analysis of a smaller sample of 104 dichotic listening fMRI datasets did not reveal any significant sex effects.

Brain Asymmetry and Pathology

Damage to corresponding regions in the left and the right hemispheres does not necessarily lead to the same detrimental effects. While left-hemispheric strokes often lead to language impairments, damage to the right hemisphere often leads to spatial and attentional deficits (Karnath and Rorden, 2012). Most notably, damage to the right perisylvan region leads to spatial neglect, a deficit in awareness of and attention toward the left side (Corbetta and Shulman, 2011; Karnath and Dieterich, 2006), a phenomenon that is only rarely observed after left-hemispheric strokes (Suchan and Karnath, 2011).

In a number of psychiatric or developmental disorders, atypical lateralization has been reported to be significantly more frequent than in the general population. For example, it has been reported that schizophrenic patients have – on average – a higher prevalence of non-right-handedness (Dragovic and Hammond, 2005) as well as reduced left-hemispheric language dominance (Collinson et al., 2009; Sommer et al., 2001), with the experience of positive symptoms and especially auditory hallucinations leading to a more pronounced reduction of the typical lateralization pattern (Ocklenburg et al., 2013c). In addition to schizophrenia, a greater frequency of atypical brain asymmetry has also been observed in autism or autism spectrum disorder (Knaus et al., 2010; Wan et al., 2012), attention deficit hyperactivity disorder (Shaw et al., 2009; Reid and Norvilitis, 2000), and dyslexia (Illingworth and Bishop, 2009).

See also: Asymmetry of Body and Brain; Cerebral Cortex; Dichotic Listening and Language: Overview; Emotion, Neural Basis of; Handedness and Language, Relationship Between; Human Language: Evolutionary Precursors; Interhemispheric Interaction in Language and Cognitive Processes; Lateralization of Language as Demonstrated by Brain Imaging Procedures; Split Brains and Human Consciousness.

Bibliography

Amunts, K., 2010. Structural indices of asymmetry. In: Hugdahl, K., Westerhausen, R. (Eds.), The Two Halves of the Brain: Information Processing in the Cerebral Hemispheres. The MIT Press, Cambridge, pp. 145–175.

Annett, M., 1970. A classification of hand preference by association analysis. British Journal of Psychology 61, 303–321.

Annett, M., 2002. Handedness and Brain Asymmetry: The Right Shift Theory. Psychology Press, Hove.

Barrick, T.R., Lawes, I.N., Mackay, C.E., Clark, C.A., 2007. White matter pathway asymmetry underlies functional lateralization. Cerebral Cortex 17, 591–598.

Baxendale, S., 2009. The Wada test. Current Opinion in Neurology 22, 185–189.

Bethmann, A., Tempelmann, C., De Bleser, R., Scheich, H., Brechmann, A., 2007. Determining language laterality by fMRI and dichotic listening. Brain Research 1133, 145–157.

Bianco, I.H., Wilson, S.W., 2009. The habenular nuclei: a conserved asymmetric relay station in the vertebrate brain. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 364, 1005–1020.

Binder, J.R., 2011. Functional MRI is a valid noninvasive alternative to Wada testing. Epilepsy and Behavior 20, 214–222.

Bisazza, A., Rogers, L.J., Vallortigara, G., 1998. The origins of cerebral asymmetry: a review of evidence of behavioural and brain lateralization in fishes, reptiles and amphibians. Neuroscience and Biobehavioral Reviews 22, 411–426.

Borod, J.C., Kent, J., Koff, E., Martin, C., Alpert, M., 1988. Facial asymmetry while posing positive and negative emotions: support for the right hemisphere hypothesis. Neuropsychologia 26, 759–764.

Brancucci, A., Lucci, G., Mazzatenta, A., Tommasi, L., 2009. Asymmetries of the human social brain in the visual, auditory and chemical modalities. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 364, 895–914.

Broca, P., 1861. Remarques sur le siége de la faculté du language articulé, suivies d'une observation d'aphémie (perte de la parole). Bulletin de la Société Anatomique 6, 330–357.

Bryden, M.P., 1977. Measuring handedness with questionnaires. Neuropsychologia 15, 617–624.

Bryden, M.P., 1988. An overview of the dichotic listening procedure and its relation to cerebral organization. In: Hugdahl, K. (Ed.), Handbook of Dichotic Listening: Theory, Methods and Research. Wiley, Weinheim.

Bryden, P.J., Roy, E.A., Spence, J., 2007. An observational method of assessing handedness in children and adults. Developmental Neuropsychology 32, 825–846.

Carey, D.P., Smith, D.T., Martin, D., Smith, G., Skriver, J., Rutland, A., Shepherd, J.W., 2009. The bi-pedal ape: plasticity and asymmetry in footedness. Cortex 45, 650–661.

Catani, M., 2010. The functional anatomy of white matter: from postmortem dissections to in vivo virtual tractography. In: Jones, D.K. (Ed.), Diffusion MRI: Theory, Methods, and Applications. Oxford University Press, Oxford, pp. 5–18.

Catani, M., Allin, M.P., Husain, M., Pugliese, L., Mesulam, M.M., Murray, R.M., Jones, D.K., 2007. Symmetries in human brain language pathways correlate with verbal recall. Proceedings of the National Academy of Sciences of the United States of America 104, 17163–17168.

Collinson, S.L., Mackay, C.E., O, J., James, A.C., Crow, T.J., 2009. Dichotic listening impairments in early onset schizophrenia are associated with reduced left temporal lobe volume. Schizophrenia Research 112, 24–31.

Concha, M.L., Burdine, R.D., Russell, C., Schier, A.F., Wilson, S.W., 2000. A nodal signaling pathway regulates the laterality of neuroanatomical asymmetries in the zebrafish forebrain. Neuron 28, 399–409.

Corballis, M.C., 2009. The evolution and genetics of cerebral asymmetry. Philosophical Transactions of the Royal Society London. Series B, Biological Sciences 364, 867–879.

Corballis, M.C., 2012. Lateralization of the human brain. Progress in Brain Research 195, 103–121.

Corbetta, M., Shulman, G.L., 2011. Spatial neglect and attention networks. Annual Review of Neuroscience 34, 569–599.

Crovitz, H.F., Zener, K., 1962. A group-test for assessing hand- and eye-dominance. American Journal of Psychology 75, 271–276.

Della Penna, S., Brancucci, A., Babiloni, C., Franciotti, R., Pizzella, V., Rossi, D., Torquati, K., Rossini, P.M., Romani, G.L., 2007. Lateralization of dichotic speech stimuli is based on specific auditory pathway interactions: neuromagnetic evidence. Cerebral Cortex 17, 2303–2311.

Demaree, H.A., Everhart, D.E., Youngstrom, E.A., Harrison, D.W., 2005. Brain lateralization of emotional processing: historical roots and a future incorporating "dominance". Behavioral and Cognitive Neuroscience Reviews 4, 3–20.

Dittmar, M., 2002. Functional and postural lateral preferences in humans: interrelations and life-span age differences. Human Biology 74, 569–585.

Draca, S., 2010. Gender-specific functional cerebral asymmetries and unilateral cerebral lesion sequelae. Reviews in the Neurosciences 21, 421–425.

Dragovic, M., Hammond, G., 2005. Handedness in schizophrenia: a quantitative review of evidence. Acta Psychiatrica Scandinavia 111, 410–419.

Elias, L.J., Bryden, M.P., 1998. Footedness is a better predictor of language lateralisation than handedness. Laterality 3, 41–51.

Elias, L.J., Bryden, M.P., Bulman-Fleming, M.B., 1998. Footedness is a better predictor than is handedness of emotional lateralisation. Neuropsychologia 36, 37–43.

Flöel, A., Buyx, A., Breitenstein, C., Lohmann, H., Knecht, S., 2005. Hemispheric lateralization of spatial attention in right- and left-hemispheric language dominance. Behavioural Brain Research 158, 269–275.

Francks, C., Maegawa, S., Laurén, J., et al., 2007. LRRTM1 on chromosome 2p12 is a maternally suppressed gene that is associated paternally with handedness and schizophrenia. Molecular Psychiatry 12, 1129–1139.

Frasnelli, E., Vallortigara, G., Rogers, L.J., 2012. Left–right asymmetries of behaviour and nervous system in invertebrates. Neuroscience and Biobehavioral Reviews 36, 1273–1291.

Gazzaniga, M.S., 2005. Forty-five years of split-brain research and still going strong. Nature Reviews Neuroscience 6, 653–659.

George, I., 2010. Hemispheric asymmetry of songbirds. In: Hugdahl, K., Westerhausen, R. (Eds.), The Two Halves of the Brain: Information Processing in the Cerebral Hemispheres. The MIT Press, Cambridge, pp. 91–120.

George, N., Evans, J., Fiori, N., Davidoff, J., Renault, B., 1996. Brain events related to normal and moderately scrambled faces. Brain Research Cognitive Brain Research 4, 65–76.

Geschwind, N., Levitsky, W., 1968. Human brain: left–right asymmetries in temporal speech region. Science 161, 186–187.

Godard, O., Fiori, N., 2010. Sex differences in face processing: are women less lateralized and faster than men? Brain and Cognition 73, 167–175.

Grimshaw, G.M., Wilson, M.S., 2013. A sinister plot? Facts, beliefs, and stereotypes about the left-handed personality. Laterality 18, 135–151.

Güntürkün, O., 2003. Adult persistence of head-turning asymmetry. Nature 421, 711.

Halpern, M.E., Güntürkün, O., Hopkins, W.D., Rogers, L.J., 2005. Lateralization of the vertebrate brain: taking the side of model systems. Journal of Neuroscience 25, 10351–10357.

Hausmann, M., Becker, C., Gather, U., Güntürkün, O., 2002. Functional cerebral asymmetries during the menstrual cycle: a cross-sectional and longitudinal analysis. Neuropsychologia 40, 808–816.

Hausmann, M., Güntürkün, O., 1999. Sex differences in functional cerebral asymmetries in a repeated measures design. Brain and Cognition 41, 263–275.

Hausmann, M., Güntürkün, O., 2000. Steroid fluctuations modify functional cerebral asymmetries: the hypothesis of progesterone-mediated interhemispheric decoupling. Neuropsychologia 38, 1362–1374.

Hirnstein, M., Westerhausen, R., Korsnes, M.S., Hugdahl, K., 2013. Sex differences in language asymmetry are age-dependent and small: a large-scale, consonant-vowel dichotic listening study with behavioral and fMRI data. Cortex 49, 1910–1921.

Hobert, O., Johnston Jr., R.J., Chang, S., 2002. Left–right asymmetry in the nervous system: the *Caenorhabditis elegans* model. Nature Reviews Neuroscience 3, 629–640.

Hopkins, B., Lems, W., Janssen, B., Butterworth, G., 1987. Postural and motor asymmetries in newlyborns. Human Neurobiology 6, 153–156.

Hopkins, W.D., 2006. Comparative and familial analysis of handedness in great apes. Psychological Bulletin 132, 538–559.

Hopkins, W.D., Cantalupo, C., 2005. Individual and setting differences in the hand preferences of chimpanzees (*Pan troglodytes*): a critical analysis and some alternative explanations. Laterality 10, 65–80.

Hopkins, W.D., Phillips, K.A., Bania, A., Calcutt, S.E., Gardner, M., Russell, J., Schaeffer, J., Lonsdorf, E.V., Ross, S.R., Schapiro, S.J., 2011. Hand preferences for coordinated bimanual actions in 777 great apes: implications for the evolution of handedness in hominins. Journal of Human Evolution 60, 605–611.

Illingworth, S., Bishop, D.V., 2009. Atypical cerebral lateralisation in adults with compensated developmental dyslexia demonstrated using functional transcranial doppler ultrasound. Brain and Language 111, 61–65.

Isaacs, K.L., Barr, W.B., Nelson, P.K., Devinsky, O., 2006. Degree of handedness and cerebral dominance. Neurology 66, 1855–1858.

Jansen, A., Flöel, A., Deppe, M., van Randenborgh, J., Dräger, B., Kanowski, M., Knecht, S., 2004. Determining the hemispheric dominance of spatial attention: a comparison between fTCD and fMRI. Human Brain Mapping 23, 168–180.

Jewell, G., McCourt, M.E., 2000. Pseudoneglect: a review and meta-analysis of performance factors in line bisection tasks. Neuropsychologia 38, 93–110.

Karev, G.B., 2000. Cinema seating in right, mixed and left handers. Cortex 36, 747–752.

Karnath, H.O., Dieterich, M., 2006. Spatial neglect—a vestibular disorder? Brain 129, 293–305.

Karnath, H.O., Rorden, C., 2012. The anatomy of spatial neglect. Neuropsychologia 50, 1010–1017.

Keenan, J.P., Rubio, J., Racioppi, C., Johnson, A., Barnacz, A., 2005. The right hemisphere and the dark side of consciousness. Cortex 41, 695–704.

Kimura, D., 2011. From ear to brain. Brain and Cognition 76, 214–217.

Klöppel, S., Mangin, J.F., Vongerichten, A., Frackowiak, R.S., Siebner, H.R., 2010. Nurture versus nature: long-term impact of forced right-handedness on structure of pericentral cortex and basal ganglia. Journal of Neuroscience 30, 3271–3275.

Knaus, T.A., Silver, A.M., Kennedy, M., Lindgren, K.A., Dominick, K.C., Siegel, J., Tager-Flusberg, H., 2010. Language laterality in autism spectrum disorder and typical controls: a functional, volumetric, and diffusion tensor MRI study. Brain and Language 112, 113–120.

Knecht, S., Dräger, B., Deppe, M., Bobe, L., Lohmann, H., Flöel, A., Ringelstein, E.B., Henningsen, H., 2000. Handedness and hemispheric language dominance in healthy humans. Brain 12, 2512–2518.

Kolb, B., Whishaw, I.Q., 2003. Fundamentals of Human Neuropsychology, fifth ed. Worth Publishers, New York. pp. 250–281.

Lebrun, Y., 1974. The inferior brain hemisphere and language. Nervenarzt 45, 510–513.

Levy, J., 1969. Possible basis for the evolution of lateral specialization of the human brain. Nature 224, 614–615.

Lindell, A.K., 2006. In your right mind: right hemisphere contributions to language processing and production. Neuropsychology Reviews 16, 131–148.

Manns, M., Güntürkün, O., 2009. Dual coding of visual asymmetries in the pigeon brain: the interaction of bottom-up and top-down systems. Experimental Brain Research 199, 323–332.

McGrew, W.C., Marchant, L.F., 2001. Ethological study of manual laterality in the chimpanzees of the Mahale Mountains, Tanzania. Behaviour 138, 329–358.

Medland, S.E., Duffy, D.L., Spurdle, A.B., Wright, M.J., Geffen, G.M., Montgomery, G.W., Martin, N.G., 2005. Opposite effects of androgen receptor CAG repeat length on increased risk of left-handedness in males and females. Behavior Genetics 35, 735–744.

Mihov, K.M., Denzler, M., Förster, J., 2010. Hemispheric specialization and creative thinking: a meta-analytic review of lateralization of creativity. Brain and Cognition 72, 442–448.

Minnebusch, D.A., Suchan, B., Köster, O., Daum, I., 2009. A bilateral occipito-temporal network mediates face perception. Behavioural Brain Research 198, 179–185.

Nagel, B.J., Herting, M.M., Maxwell, E.C., Bruno, R., Fair, D., 2013. Hemispheric lateralization of verbal and spatial working memory during adolescence. Brain and Cognition 82, 58–68.

Nicholls, M.E., Hadgraft, N.T., Chapman, H.L., Loftus, A.M., Robertson, J., Bradshaw, J.L., 2010. A hit-and-miss investigation of asymmetries in wheelchair navigation. Attention, Perception and Psychophysics 72, 1576–1590.

Nicholls, M.E., Loftus, A., Mayer, K., Mattingley, J.B., 2007. Things that go bump in the right: the effect of unimanual activity on rightward collisions. Neuropsychologia 45, 1122–1126.

Ocklenburg, S., Arning, L., Gerding, W.M., Epplen, J.T., Güntürkün, O., Beste, C., 2013a. Cholecystokinin A receptor (CCKAR) gene variation is associated with language lateralization. PLoS One 8, e53643.

Ocklenburg, S., Arning, L., Hahn, C., Gerding, W.M., Epplen, J.T., Güntürkün, O., Beste, C., 2011. Variation in the NMDA receptor 2B subunit gene GRIN2B is associated with differential language lateralization. Behavioural Brain Research 225, 284–289.

Ocklenburg, S., Güntürkün, O., 2009. Head-turning asymmetries during kissing and their association with lateral preference. Laterality 14, 79–85.

Ocklenburg, S., Güntürkün, O., 2012. Hemispheric asymmetries: the comparative view. Frontiers in Psychology 3, 5.

Ocklenburg, S., Ströckens, F., Güntürkün, O., 2013b. Lateralisation of conspecific vocalisation in non-human vertebrates. Laterality 18, 1–31.

Ocklenburg, S., Westerhausen, R., Hirnstein, M., Hugdahl, K., 2013c. Auditory hallucinations and reduced language lateralization in schizophrenia: a meta-analysis of dichotic listening studies. Journal of the International Neuropsychological Society 19, 410–418.

Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9, 97–113.

Palmer, A.R., 1996. From symmetry to asymmetry: phylogenetic patterns of asymmetry variation in animals and their evolutionary significance. Proceedings of the National Academy of Sciences of the United States of America 93, 14279–14286.

Palmer, A.R., 2002. Chimpanzee right-handedness reconsidered: evaluating the evidence with funnel plots. American Journal of Physical Anthropology 118, 191–199.

Peters, M., Servos, P., 1989. Performance of subgroups of left-handers and right-handers. Canadian Journal of Psychology 43, 3413–3458.

Provins, K.A., Magliaro, J., 1993. The measurement of handedness by preference and performance tests. Brain and Cognition 22, 171–181.

Reid, H.M., Norvilitis, J.M., 2000. Evidence for anomalous lateralization across domain in ADHD children as well as adults identified with the Wender Utah rating scale. Journal of Psychiatry Research 34, 311–316.

Reiss, M., Reiss, G., 1997. Ocular dominance: some family data. Laterality 2, 7–15.

Reiss, M., Reiss, G., 1999. Earedness and handedness: distribution in a German sample with some family data. Cortex 35, 403–412.

Reiss, M., Tymnik, G., Kögler, P., Kögler, W., Reiss, G., 1999. Laterality of hand, foot, eye, and ear in twins. Laterality 4, 287–297.

Rentería, M.E., Jun 2012. Cerebral asymmetry: a quantitative, multifactorial, and plastic brain phenotype. Twin Research and Human Genetics 15 (3), 401–413. http://dx.doi.org/10.1017/thg.2012.13. Review.

Ringo, J.L., Doty, R.W., Demeter, S., Simard, P.Y., 1994. Time is of the essence: a conjecture that hemispheric specialization arises from interhemispheric conduction delay. Cerebral Cortex 4, 331–343.

Rodrigo, S., Naggara, O., Oppenheim, C., et al., 2007. Human subinsular asymmetry studied by diffusion tensor imaging and fiber tracking. Amercian Journal of Neuroradiology 28, 1526–1531.

Rogers, L.J., 1996. Behavioral, structural and neurochemical asymmetries in the avian brain: a model system for studying visual development and processing. Neuroscience and Biobehavioral Reviews 20, 487–503.

Rogers, L.J., 2009. Hand and paw preferences in relation to the lateralized brain. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 364, 943–954.

Rogers, L.J., Zucca, P., Vallortigara, G., 2004. Advantages of having a lateralized brain. Proceedings in Biological Sciences 271 (Suppl. 6), 420–422.

Rossion, B., Dricot, L., Devolder, A., Bodart, J.M., Crommelinck, M., De Gelder, B., Zoontjes, R., 2000. Hemispheric asymmetries for whole-based and part-based face processing in the human fusiform gyrus. Journal of Cognitive Neuroscience 12, 793–802.

Sandmann, P., Eichele, T., Specht, K., Jäncke, L., Rimol, L.M., Nordby, H., Hugdahl, K., 2007. Hemispheric asymmetries in the processing of temporal acoustic cues in consonant vowel syllables. Restorative Neurology and Neuroscience 25, 227–240.

Scerri, T.S., Brandler, W.M., Paracchini, S., Morris, A.P., Ring, S.M., Richardson, A.J., Talcott, J.B., Stein, J., Monaco, A.P., 2011. PCSK6 is associated with handedness in individuals with dyslexia. Human Molecular Genetics 20, 608–614.

Schaafsma, S.M., Riedstra, B.J., Pfannkuche, K.A., Bouma, A., Groothuis, T.G., 2009. Epigenesis of behavioural lateralization in humans and other animals. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 364, 915–927.

Siefer, A., Ehrenstein, W.H., Arnold-Schulz-Gahmen, B.E., Sökeland, J., Luttmann, A., 2003. Population statistics and association analysis of sensorimotor lateral preferences and its relevance for various fields of professional activity. Zentralblatt für Arbeitsmedizin, Arbeitsschutz und Ergonomie 53, 346–353.

Shaw, P., Lalonde, F., Lepage, C., et al., 2009. Development of cortical asymmetry in typically developing children and its disruption in attention-deficit/hyperactivity disorder. Archives of General Psychiatry 66, 888–896.

Silberman, E.K., Weingartner, H., 1986. Hemispheric lateralization of functions related to emotion. Brain and Cognition 5, 322–353.

Sommer, I., Ramsey, N., Kahn, R., Aleman, A., Bouma, A., 2001. Handedness, language lateralisation and anatomical asymmetry in schizophrenia: meta-analysis. British Journal of Psychiatry 178, 344–351.

Sperry, R.W., 1961. Cerebral organization and behavior: the split brain behaves in many respects like two separate brains, providing new research possibilities. Science 133, 1749–1757.

Spironelli, C., Angrilli, A., 2007. Influence of phonological, semantic and orthographic tasks on the early linguistic components N150 and N350. International Journal of Psychophysiology 64, 190–198.

Spironelli, C., Angrilli, A., 2009. Developmental aspects of automatic word processing: language lateralization of early ERP components in children, young adults and middle-aged subjects. Biological Psychology 80, 35–45.

Ströckens, F., Güntürkün, O., Ocklenburg, S., 2013. Limb preferences in non-human vertebrates. Laterality 18, 536–575.

Suchan, J., Karnath, H.O., 2011. Spatial orienting by left hemisphere language areas: a relict from the past? Brain 134, 3059–3070.

Sun, T., Walsh, C.A., 2006. Molecular approaches to brain asymmetry and handedness. Nature Reviews Neuroscience 7, 655–662.

Tervaniemi, M., Hugdahl, K., 2003. Lateralization of auditory-cortex functions. Brain Research Brain Research Reviews 43, 231–246.

Tzourio-Mazoyer, N., Simon, G., Crivello, F., et al., 2010. Effect of familial sinistrality on planum temporale surface and brain tissue asymmetries. Cerebral Cortex 20, 1476–1485.

Vallortigara, G., 2006. The evolutionary psychology of left and right: costs and benefits of lateralization. Developmental Psychobiology 48, 418–427.

Vallortigara, G., Rogers, L.J., Bisazza, A., 1999. Possible evolutionary origins of cognitive brain lateralization. Brain Research Brain Research Reviews 30, 164–175.

Van der Haegen, L., Westerhausen, R., Hugdahl, K., Brysbaert, M., 2013. Speech dominance is a better predictor of functional brain asymmetry than handedness: a combined fMRI word generation and behavioral dichotic listening study. Neuropsychologia 51, 91–97.

Vernooij, M.W., Smits, M., Wielopolski, P.A., Houston, G.C., Krestin, G.P., van der Lugt, A., 2007. Fiber density asymmetry of the arcuate fasciculus in relation to functional hemispheric language lateralization in both right- and left-handed healthy subjects: a combined fMRI and DTI study. Neuroimage 35, 1064–1076.

Ververs, I.A., de Vries, J.I., van Geijn, H.P., Hopkins, B., 1994. Prenatal head position from 12–38 weeks. I. Developmental aspects. Early Human Development 39, 83–91.

Vogel, J.J., Bowers, C.A., Vogel, D.S., 2003. Cerebral lateralization of spatial abilities: a meta-analysis. Brain and Cognition 52, 197–204.

Wan, C.Y., Marchina, S., Norton, A., Schlaug, G., 2012. Atypical hemispheric asymmetry in the arcuate fasciculus of completely nonverbal children with autism. Annals of the New York Academy of Sciences 1252, 332–337.

Yin, X., Han, Y., Ge, H., et al., 2013. Inferior frontal white matter asymmetry correlates with executive control of attention. Human Brain Mapping 34, 796–813.

Brain Damage: Neuropsychological Rehabilitation

Bryan D Fantie, American University, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

One of the most significant changes in neuropsychological rehabilitation over the last two decades has been its evolution from a field focused almost exclusively on cognitive and memory dysfunction to one that is holistic and that is concerned with entire human beings as multifaceted individuals, members of relationships, and responsive organisms trying to survive and flourish within complex environments. This article reviews briefly some of the major ideas and theories that gave rise to contemporary neuropsychological rehabilitation as well as recent advances and innovations, and challenges for the future of the field.

Definition

Advances in emergency medicine, biotechnology, computers, and, in particular, basic neuroscience, which has resulted, most notably, in an improved understanding of multiple forms of neural plasticity (see Kolb and Muhammad (2014) for a review), have fueled substantial developments in applied neuropsychology over the past few decades. These, and other contributions from the cognitive, social, and behavioral sciences, have not only led to dramatic changes in the way we conceptualize the fundamental functional nature of the human brain but have enriched the repertory of tools and techniques available to ameliorate the effects of brain damage.

For centuries, people thought of the brain as a magnificent, logical, thinking machine, allowing us to solve mathematical problems and do things such as physics and architecture. The current zeitgeist, however, is based on the notion that the evolutionary pressure related to the need to make sense of both our often infuriatingly complex emotional lives and our large and convoluted social world is really what shaped our hefty, intricate brain and that those other intellectual abilities arose largely as by-products of the capacity to organize complexity and to create internal models that could help predict the behavior of others as well as the future consequences of our own actions.

Not that long ago, mainstream clinical neuropsychology itself was focused almost entirely on general intellectual ability, memory, problem-solving, executive function, and the like, while emotion and social interaction were in charge of other specialties. Many considered improving cognition the *sine qua non* of neuropsychological rehabilitation (Robertson, 2001). Concurrent with the growing consensus that cognitive and affective functioning are less dichotomous than believed previously (see Damasio, 1994, for example), a more holistic view of brain function has emerged coupled with the acknowledgment that the brain is the driving force for many more functions than just intellect, and neuropsychology has expanded its purview accordingly. There is also increased recognition of the value of quality of life and interpersonal experience to recovery, and that social relationships, directly and indirectly, are often strong predictors of outcome after brain injury (Cattran et al., 2011; Fantie and Kolb, 1991; Morton and Wehman, 1995).

In light of the connectedness of cognition, emotion, and psychosocial functioning, all should be the focus of rehabilitative efforts. Therefore, neuropsychological rehabilitation can be defined as the set of applied practices that involve "the systematic use of instruction and structured experience" (Robertson, 2001) concerned with the "amelioration of cognitive, emotional, psychosocial and behavioural deficits caused by an insult to the brain" (Wilson, 2011) that are guided by neuropsychological theory and neuropsychological assessment.

Brief History

Although neuropsychological rehabilitation has emerged as a formal entity only relatively recently, the study of the injured brain and its recovery has a long pedigree shared by medicine, anatomy, neurology, neuroscience, and neuropsychology. Space limitations preclude a comprehensive review of this history. Nevertheless, there are several excellent books on this topic that should more than whet the appetite of the interested reader (Corsi, 1991; Finger, 1994, 2000; Marshall and Magoun, 1998). What follows is a brief sample of some of the brain science pioneers who have made contributions especially relevant to neuropsychological rehabilitation. This list is incomplete, by necessity, and is not meant to be an exhaustive *dramatis personae*.

Paul Broca

Nineteenth-century French neurologist, Paul Broca (1824–80), known best for his discoveries related to the neurological basis of speech and language (Broca, 1861, 1863, 1866; Ryalls and Lecours, 1996), recommended that if one wished to teach an aphasic patient to speak again, one should approach the task with the "perseverance" and "indefatigable consistency of a mother who is teaching her infant to speak" (Broca, 1865: p. 390).

> Puis, comment sait-on que l'aphémique n'est pas capable d'apprendre à parler avec l'hémisphère droit qui lui reste? A-t-on essayé de faire son éducation? lui a-t-on donné ces leçons de tous les jours, de toutes les heures, de tous les instants à l'aide desquelles on finit à la longue par faire parler un enfant? Pour ma part, je suis convaincu que, sans rendre aux aphémiques la partie de leur intelligence qui a péri avec une partie de leur cerveau, on pourrait, en y mettant assez de persévérance, en les traitant avec l'infatigable constance de la mère qui apprend à parler à son enfant, on pourrait, dis-je, obtenir des résultats considérables.

Broca based this advice on his conclusion that, without the return of the part of their brain that had been lost, aphasics might best be taught to speak as though they were learning to do so for the first time because the processes involved when infants learn are different from those of an adult.

Shepherd Ivory Franz

New Jersey-born neuropsychologist Shepherd Ivory Franz, author of *Nervous and Mental Re-Education* (1923), investigated how well clients with aphasia benefitted from therapy. Franz, like Broca, also observed that the process of recovering from aphasia resembled learning a new skill more than merely reacquiring a habit established previously (Franz, 1905). Consequently, many subsequent rehabilitative treatment techniques designed for those with brain damage put an emphasis on learning new skills (Brewer-Mixon and Cullum, 2013; Witsken et al., 2008).

Franz made significant contributions to the science of rehabilitation in the laboratory as well as the clinic (Colotla and Bach-Y-Rita, 2002). He, along with Robert Ogden, also showed that forcing a monkey to use its affected arm after motor cortex ablation produced significant recovery of function that was not obtained in animals with similar lesions that were not forced to use the affected limb (Ogden and Franz, 1917), which had implications for people with hemiplegia, in particular, as well as other similar disabilities resulting from neural injury.

Otfrid Foerster

The German neurologist and neurosurgeon Otfrid Foerster (1873–1941), who had Vladimir Ilyich Lenin as a patient (Sarikcioglu, 2007), was a major influence on neurology and the emerging field of neurosurgery (Loring, 2010). His contributions to neuropsychological rehabilitation included a better understanding of the structure of the nervous system in general and the neuroanatomical substrata serving functional processes. This knowledge benefitted the design of more effective therapies and treatments.

Foerster (1930) conceptualized two fundamental processes as the bases of functional recovery after neurological insult. These are (1) regeneration and (2) reorganization. Regeneration involves the regrowth or replacement of the injured brain or nerve tissues so that any function that is restored is the result of a direct replacement of the damaged circuitry. During reorganization, on the other hand, it is the undamaged neural structure that alters its structure or actions to restore the original function subserved by the lost tissue. Therefore, surgery and prolonged exercise can both promote recovery of function by different means, each effecting one of the processes of regeneration and reorganization, respectively (Benton and Tranel, 2000).

More than any other branch of psychology, neuropsychology has been involved with the "cruel experiments of nature," a phrase used by G. Stanley Hall in his preface to Freud's *A General Introduction to Psychoanalysis* (1920). It is not surprising, then, that two World Wars would result in a surge in new knowledge about recovery from traumatic brain injury.

Walther Poppelreuter

At the beginning of World War I, Walther Poppelreuter (1886–1939), a German psychologist and neurologist, was stationed at the Psychiatric Hospital of Cologne, where he specialized in working with soldiers who had suffered head injuries. This work continued after the War at the Institute for Applied Psychology in Bonn and eventually the University of Bonn (Boake, 2003). Poppelreuter is associated with the following: (1) the notion that one should concentrate on the mastery of accuracy first and speed second when trying to teach a skill requiring both, known as Poppelreuter's law (but see Hellebrandt (1972)), and (2) the introduction of the Overlapping Figures test (Della Sala et al., 1995; Lezak et al., 2004: p. 403).

Kurt Goldstein

A contemporary and compatriot of Poppelreuter during World War I, Kurt Goldstein (1878–1965) also treated German soldiers with war-ravaged brains. Notably, both Goldstein and Poppelreuter developed and promoted the use of quantitative neuropsychological assessment methods to isolate and delineate deficits, to identify brain lesion–behavioral dysfunction relationships, and to inform treatment (Sullivan and Tapert, 2013). In addition to developing procedures for treating impaired reading, writing, and speech, which guided subsequent rehabilitative efforts, Goldstein also devised new methods that could be used to help shape specific target behaviors during recovery (Brewer-Mixon and Cullum, 2013; Witsken et al., 2008).

Goldstein developed a theoretical model for neuropsychological rehabilitation that, on the surface, appears similar to that of Foerster (see earlier entry), comprising two major mechanisms of recovery, one direct and one indirect, but their theories differ at their philosophical core. Addressing the scientific 'facts' associated with a brain injury (Prigatano, 2013), Goldstein thought that repetitive stimulation might be able to restore a cognitive function that had been disrupted by injury only if a sufficient portion of the related neuroanatomy was still intact. In the case of the complete loss of the relevant tissue, rehabilitative efforts would be best directed toward recuperation (Robertson, 2001). As a neurologist and *psychiatrist*, as opposed to a neurosurgeon such as Foerster,

Goldstein focused his theory less on the specific neuroanatomy involved in the injury and more on the phenomenological effects that trauma imposed on the injured person as an entire organism striving to function within an environment (Goldstein, 1939, 1963). For example, Goldstein (1942) emphasized how functional deficits could thwart an individual's resumption of a purposeful everyday life. According to Goldstein, brain injury changes the entire person, forever, and this experiential 'fact' produces what Goldstein referred to as a 'catastrophic reaction,' which has widespread effects and poses the major obstacle to social adjustment (Prigatano, 2013; Sullivan and Tapert, 2013). According to Prigatano (2000: p. 116), the first principle of neuropsychological rehabilitation is that the "clinician must begin with patient's subjective or phenomenological experience to reduce their frustrations and confusion in order to engage them in the rehabilitation process," and Goldstein also developed perceptive methods to help patients cope with these issues (Prigatano, 2013).

Alexander Romanovich Luria

World War II also produced an abundance of opportunities for gaining new knowledge about the sequelæ of brain trauma. Most notable, perhaps, was the work of A.R. Luria (1902–77), a psychologist and physician, who treated wounded Russian soldiers. He was the author of *The Man with a Shattered World: The History of a Brain Wound* (1972), arguably the best case study written about a person's recovery from a devastating brain injury. Luria had a profound influence on the development of neuropsychology as a result of a career treating patients with dysfunctions that ranged from the simplest to most complex neuropsychological functions (Brewer-Mixon and Cullum, 2013; Christensen and Caetano, 1996). In particular, his creative (and often improvised) assessment techniques provided a functional dissection of the precise nature and component features of the neurocognitive deficits of concern. Luria's approach provided insight into understanding both the consequences of the injury and the normal functioning of the intact human brain (Luria, 1966, 1973). This analysis and identification of the basis of the manifesting deficit was at the heart of what Luria thought were the primary goals of neuropsychological rehabilitation, namely, that recovery could be maximized if therapies were custom-designed to the individual and targeted the specific problem either to restore the disrupted root function or, if that was not possible, to at least mitigate the adverse effects of the loss of that basic capacity (Christensen, 1989; Luria, 1963; Prigatano, 2013). Robertson (2001) listed Luria's four main principles of rehabilitation, which are as follows: (1) recognition that each person's brain and functional neural circuitry are unique, (2) employing the function of undamaged neural systems to substitute for those that have been compromised, (3) using speech or other aids to assist the conscious performance of behaviors that were automatic before the injury until those behaviors are again internalized and spontaneous, and (4) providing consistent feedback to patients, so they can monitor their progress (Luria, 1963).

William Ritchie Russell, Freda Gladys Newcombe, and William Alwyn Lishman

Work with wounded World War II veterans in Great Britain supported the creation of three seminal works in neuropsychology and neuropsychiatry: *The Traumatic Amnesias* (1971) by W. Ritchie Russell (1903–80); *Missile Wounds of the Brain: A Study of Psychological Deficits* (1969) by Freda Newcombe (1925–2001); and *Organic Psychiatry: The Psychological Consequences of Cerebral Disorder* (1978) by Alwyn Lishman (1931–present). In light of the relatively recent increased interest concerning the importance of rehabilitating noncognitive functions in a holistic therapeutic approach, these works were prescient. As Newcombe (1969) observed, "Few of the subjects complained of loss of intellectual efficiency when asked to describe in detail any symptoms or difficulties" and "There were more complaints of affective symptoms: a tendency to irritability, depression, or loss of confidence " (p. 53 cited in Prigatano (2013): p. 752).

Oliver Zangwill

In contrast to the respective models presented by Foerster and Goldstein, British neuropsychologist Oliver Zangwill (1913–87) described three cardinal processes of rehabilitation following brain injury: (1) *compensation*, which is "a reorganization of psychological function so as to minimize or circumvent a particular disability" (p. 63); (2) *substitution*, defined as "the building up of a new method of response to replace one damaged irreparably by a cerebral lesion." (p. 64); and (3) *direct training*, which is the restoration of the affected abilities through conscious practice in cases where the neural substrate has not been destroyed completely (Zangwill, 1947).

Issues and Recent Advances

After consulting a number of leaders in the field, Barbara Wilson (2011) described what she considered were 10 of the most important developments in neuropsychological rehabilitation over the previous 5 years. Her list follows, and these topics will serve as a guide for a brief review of what neuropsychological rehabilitation has been and what it is becoming.

Recognition of the Need to Evaluate Rehabilitation in More Appropriate Ways

Recognition of the need to integrate practical clinical experience with evidence-based measures and techniques is, in itself, a significant advance. Recognizing that the standard single- and double-blind randomized control trials with placebos are not always the best research designs in rehabilitation, or even appropriate, is leading to the development of new paradigms promising to strengthen future therapeutic efforts, for example, the development of single-case experimental design (SCED) (Perdices and Tate, 2009; Tate et al., 2008). In 2014, the journal *Neuropsychological Rehabilitation* published a special issue devoted to SCED (Evans et al.,

2014). This issue contained articles covering topics that included Bayesian analysis of data from single-case designs (Rindskopf, 2014), randomization tests of effect size (Heyvaert and Onghena, 2014), measuring procedural fidelity (Ledford and Gast, 2014), and using multilevel analysis to integrate results within and across SCED studies (Baek et al., 2014), among others.

Errorless Learning for People with Language Disorders

The goal of errorless learning (EL) is to reduce the number of mistakes made during the acquisition of new skills or facts as much as possible so that erroneous responses are not reinforced unconsciously or unintentionally, presumably by processes related to implicit memory. Wilson, Baddeley, and others (e.g., Baddeley and Wilson, 1994; Evans et al., 2000; Wilson et al., 1994) have found EL especially effective in treating people with severe memory problems in contrast to trial-and-error learning, and EL appears about as effective as errorful learning with language disorders, although EL is greatly preferred by the patients, who find it less frustrating (Raymer et al., 2010).

Theories and Models Influencing Rehabilitation

A single model is inadequate to accommodate the complexity and multidimensional nature of neuropsychological rehabilitation. "For assessment, we need to refer to theories and models of cognitive functioning, localization, ecological validity and others; for treatment, we need to refer to theories and models of learning, cognition, emotion, social functioning and many others; we also need to know about recovery, plasticity, social systems and other aspects" (Wilson, 2011: p. 37). Developments and improvements in the models and theories from many relevant fields have enriched neuropsychological rehabilitation and continue to inform rehabilitative theories, goals, and techniques.

Theory-Driven Rehabilitation Interventions for Executive Dysfunctions

Problems with executive functions are common after brain injury. Wilson (2011) reported that "some new and effective treatment strategies" have emerged based on Duncan's theoretical model that describes disordered executive function as "a defect … in the role of purpose: control of action by its intended result" (Duncan, 1986: p. 271). Robertson (1996) based goal-management training on this model, which has served as the basis of some promising results (e.g., Krasny-Pacini et al., 2014; Levine et al., 2000; Miotto et al., 2009).

Restitution of Working Memory Deficits

Working memory is a fundamental component of executive functioning. There have been multiple reports, from different groups, working with a variety of populations that have reported improvements in working memory through training, which generalize to other tasks and everyday situations (e.g., Serino et al., 2007 and Westerberg et al., 2007 in adults after traumatic brain injury (TBI) and stroke, respectively).

The Perception of Emotion and Its Rehabilitation

Survivors of TBI frequently experience changes in their ability to perceive and express emotion (e.g., Biszak and Babbage, 2014; Bornhofen and McDonald, 2007, 2008; Fantie et al., 2001). The processes related to the perception of emotional stimuli, such as facial expression, affective prosody, body posture, and other aspects of interpersonal and social communication, have finally achieved deserved attention after being neglected for decades by the majority of the mainstream scientific community, including behavioral psychologists, largely due to the constraints of behaviorism. Given how important this aspect of behavior is to functioning successfully in our daily lives in a social world in addition to the quality of life, this recognition is long overdue and will drive sorely needed developments in the remediation of deficits within this domain.

Brain–Computer Interaction

Faster computers, able to monitor brain-related activity recorded through techniques such as EEG, fMRI, or PET in real time, continue to support the development of a better understanding of how the brain works and the effects of brain damage as well as new tools to assist with diagnosis and rehabilitation. Wilson (2011) singled out two advances in the use of brain–computer interaction (BCI) related to neurofeedback: In the first, BCI controlled devices such as prosthetic limbs and wheelchairs (e.g., Tan et al., 2014) and in the second, BCI served as means to communicate with people who appear to be in a vegetative state, but who show good evidence of active cognitive function when their brain activity is monitored during certain situations. For example, these patients were able to modulate their brain activity differentially to provide meaningful yes–no responses to questions (e.g., Monti et al., 2010; Owen et al., 2006).

Virtual Reality

Neuropsychologists have been using computers to administer tests during neuropsychological assessments for about three decades, with benefits over paper-and-pencil measures that include improved procedural consistency across assessments as well as the ability to record response-related reaction times and to control the duration of stimulus presentations more precisely and accurately. Another promising and exciting recent application of computers involves the use of virtual reality (VR), which is producing novel and inventive procedures that benefit both assessment and therapy (Weiss and Klinger, 2009). As discussed earlier, executive functions are of particular importance to the ability to function in everyday life but are often compromised by brain injury. Assessing executive functions with traditional paper-and-pencil tasks is not completely satisfactory (Chan et al., 2008). VR can create more realistic and ecological valid simulations of the complex, sequential, multistage, and often simultaneous behaviors that are frequently compromised after brain trauma. For example, Raspelli and colleagues (Raspelli et al., 2010, 2009, 2012, 2011) developed a VR version of the Multiple Errands Task (Shallice and Burgess, 1991) that simulates a supermarket. Similarly, larger VR environments (e.g., a shopping mall, small neighborhoods) are being developed that can be used to assess

and train navigation, wayfinding, and spatial orientation abilities after brain damage (Carelli et al., 2009a,b; Koenig et al., 2009).

Cognitive Enhancement: Crosswords, Computers, and Video Games

Amy Jak et al. (2013) reviewed and evaluated outcomes from a sample of 21 cognitive training programs that used electronic activities and that included participants with traumatic brain injury, brain tumor, stroke, severe mental illness, as well as healthy individuals. Although these programs typically produced improvements on the training tasks themselves, there was not often evidence of lasting generalizability to other tasks or everyday activities. The computerized cognitive training did, however, appear to have a positive effect on the participants' perception of their social connectivity and self-efficacy. Jak and colleagues expressed concern about the poor methodology they found employed by many of the studies that they evaluated and remained optimistic about the potential for these activities to be used more effectively in the future.

Identity Issues following Acquired Brain Injury

There are issues related to a person's concepts of self and identity that must be resolved before brain trauma survivors can achieve their optimal potential (Ben-Yishay and Daniels-Zide, 2000; Prigatano, 1999). These concepts are central to the holistic approach to rehabilitation, albeit, they have not been defined well operationally (Wilson, 2011). The journal *Neuropsychological Rehabilitation* published a special issue dedicated to "The self and identity in rehabilitation" (Gracey and Ownsworth, 2008). This issue contained reports that show some promising work toward the resolution of this shortcoming by providing models that can help us think about the interactions of personal and social identity, cognition, and neural processes (e.g., Cloute et al., 2008; Haslam et al., 2008; Wilson et al., 2008; Yeates et al., 2008).

Support in the Early Stages of Dementia

Wilson (2011) noted the need for psychosocial (i.e., 'non-pharmacological') interventions to support those affected by early-stage dementia because of the relevant scarcity and limited efficacy of available medications as well as the need for outcome measures that do a better job of assessing the everyday, real-world benefits that interventions purportedly provide. A contemporaneous study by Clare et al. (2010) reported improved goal performance and satisfaction ratings in a group of people with early-stage Alzheimer's disease as the result of goal-oriented cognitive rehabilitation, suggesting the usefulness and value of this intervention to those with early-stage dementia and their families.

Computational Modeling

Within cognitive neuropsychology, a computational model uses sets of mathematical formulae to represent and simulate the processes underlying linguistic and cognitive abilities. These formulae are derived from a target theory and form the basis for a computer program that not only can perform the cognitive task of interest, but does so, presumably, in exactly the same way as a human would, at the functional, information-processing level (Caramazza and Coltheart, 2006). Computational modeling has done quite well replicating and predicting language and cognitive deficits, but has been less successful predicting recovery and rehabilitation in response to interventions (see Dell and Caramazza, 2008). There is a growing interest in trying to solve this problem (e.g., Welbourne and Ralph, 2005). Nevertheless, Wilson (2011) reports a personal communication from Matthew Lambon Ralph in which he noted, "the model does not get depressed after its stroke, it does not have a grieving family, it does not have to worry about the mortgage, and does not have to contend with other real-life issues" (p. 34). This observation underlines the need to adopt a big picture view when dealing with clinical populations and the challenge of rehabilitating a human being back to as normal a life as is possible.

The Holistic Approach to Rehabilitation

Perhaps the most significant development in neuropsychological rehabilitation over the past decade or so has been the increased acceptance of the holistic approach, the conceptual basis of which was Kurt Goldstein's 'organismic' theory of human nature (Goldstein, 1939, 1963). Put simply, in contrast to the notion that brain injuries produce cognitive or behavioral deficits that can be treated by focusing on that specific dysfunction within its domain, the holistic approach sees "the failure to function normally after a brain injury as the consequence of a combination of (or the interactions among) biological and behavioral causes" (Ben-Yishay and Diller, 2008: p. 79). The biological causes relate to the actual physical injury and the consequences of the damaged neural circuitry. These 'behavioral' causes, however, go well beyond the specific deficits manifested by damaged tissue and can include the full gamut of conscious and unconscious emotional, intellectual, and experiential responses to changes that have taken place in that person as an individual, as an emotional and social being, and as an organism interacting with its environment (see Ben-Yishay and Diller (1983) and Prigatano et al. (1986) for more comprehensive reviews). Thus, to maximize a person's adjustment to a brain injury, holistic approaches to neuropsychological rehabilitation require psychotherapy in addition to interventions directed at ameliorating cognitive and behavioral difficulties (Prigatano, 1999).

Future Directions

Barbara Wilson (2011) suggested the possible future developments in neuropsychological rehabilitation, which include the following: establishing stronger links with basic neuroscience, better use of imaging procedures, more and better collaboration with pharmaceutical companies, better evaluation of rehabilitation programs, and the need to educate researchers and practitioners as to the meaning and value of rehabilitation.

Prigatano (2013: p. 756, Table 2) presented the following as opportunities for advancing the field of neuropsychological rehabilitation:

1. The clinical utility as well as size effects of cognitive rehabilitation need to be thoroughly documented.

2. More time-efficient and 'user-friendly' memory compensation techniques need to be developed.
3. The characteristics of the psychotherapist and of the patient that lead to effective symptom reduction and ultimate maintenance of employment following severe traumatic brain injury need to be thoroughly described.
4. The characteristics of the therapist and the patient that allow for the development of the therapeutic or working alliance need to be thoroughly studied.
5. Long-term cost–benefit analysis of neuropsychologically oriented rehabilitation needs to be conducted.
6. Techniques for improving family education and involvement in the rehabilitation process of children and adults need to be developed.
7. The neuroimaging correlates (structural and functional) for improved cognitive and behavioral functioning in adults and children following severe traumatic brain injury need to be better understood.
8. The role of deep brain stimulation and the recovery of certain cognitive functions (for example, memory) after severe post-acute traumatic brain injury should be explored.
9. Clarification of what are the 'direct' versus 'indirect' effects of brain injury needs to be specified and studied in both clinical and experimental research.
10. Barriers for conducting effective neuropsychological you militate in of children need to be eliminated.

See also: Parietal Lobe; Prefrontal Cortex; Recovery of Function: Dependency on Age; Temporal Lobe.

Bibilography

Baddeley, A., Wilson, B.A., 1994. When implicit learning fails: amnesia and the problem of error elimination. Neuropsychologia 32, 53–68.

Baek, E.K., Moeyaert, M., Petit-Bois, M., Beretvas, S.N., Van Den Noortgate, W., Ferron, J.M., 2014. The use of multilevel analysis for integrating single-case experimental design results within a study and across studies. Neuropsychological Rehabilitation 24, 590–606.

Ben-Yishay, Y., Daniels-Zide, E., 2000. Examined lives: outcomes after holistic rehabilitation. Rehabilitation Psychology 45, 112–129.

Ben-Yishay, Y., Diller, L., 1983. Cognitive remediation. In: Rosenthal, M., Griffith, R.R., Bond., M.R., Miller, J.D. (Eds.), Rehabilitation of the Head Injured Adult. FW Davis, Philadelphia, PA, pp. 167–183.

Ben-Yishay, Y., Diller, L., 2008. Kurt Goldstein's holistic ideas—an alternative, or complementary, approach to the management of traumatically brain-injured individuals. US Neurology 4, 79–80.

Benton, A.R., Tranel, D., 2000. Historical notes on reorganization of function and neuroplasticity. In: Levin, H.S., Grafman, J. (Eds.), Cerebral Reorganization of Function after Brain Damage. Oxford University Press, New York, pp. 3–23.

Biszak, A.M., Babbage, D.R., 2014. Facial affect recognition difficulties in traumatic brain injury rehabilitation services. Brain Injury 28, 97–104.

Boake, C., 2003. Stages in the history of neuropsychological rehabilitation. In: Wilson, B.A. (Ed.), Neuropsychological Rehabilitation: Theory and Practice, first ed. CRC Press, pp. 23–52.

Bornhofen, C., Mcdonald, S., 2007. Treating deficits in emotion perception following traumatic brain injury. Neuropsychological Rehabilitation 18, 22–44.

Bornhofen, C., Mcdonald, S., 2008. Emotion perception deficits following traumatic brain injury: a review of the evidence and rationale for intervention. Journal of the International Neuropsychological Society 14, 511–525.

Brewer-Mixon, K.K., Cullum, C.M., 2013. Historical principles and foundations of neuropsychological rehabilitation. In: Noggle, C.A., Dean, R.S., Barista, M.T. (Eds.), Neuropsychological Rehabilitation. New York.

Broca, P., 1861. Nouvelle observation d'aphémie produite par une lésion de la moitié postérieure des deuxième et troisième circonvolution frontales gauches. Bulletin de la Société Anatomique 36, 398–407.

Broca, P., 1863. Localisations des fonctions cérébrales. Siège de la faculté du langage articulé. Bulletin de la Société d"Anthropologie 4, 200–208.

Broca, P., 1865. Sur le siège de la faculté du langage articulé. Bulletins de la Société d'anthropologie de Paris 377–393.

Broca, P., 1866. Sur la faculté générale du langage, dans ses rapports avec la faculté du langage articulé. Bulletin de la Société d"Anthropologie 1, 377–382.

Caramazza, A., Coltheart, M., 2006. Cognitive neuropsychology twenty years on. Cognitive Neuropsychology 23, 3–12.

Carelli, L., Rusconi, M.L., Mattioli, F., Stampatori, C., Morganti, F., Riva, G., 2009a. Neuropsychological and virtual reality assessment in topographical disorientation. In: Wiederhold, B.K., Riva, G. (Eds.), Annual Review of Cybertherapy and Telemedicine 2009: Advanced Technologies in the Behavioral, Social and Neurosciences.

Carelli, L., Rusconi, M.L., Mattioli, F., Stampatori, C., Morganti, F., Riva, G., 2009b. Neuropsychological and virtual reality assessment in topographical disorientation. Cyberpsychology & Behavior 12, 596–597.

Cattran, C.J., Oddy, M., Wood, R.L., Moir, J.F., 2011. Post-injury personality in the prediction of outcome following severe acquired brain injury. Brain Injury 25, 1035–1046.

Chan, R.C.K., Shum, D., Toulopoulou, T., Chen, E.Y.H., 2008. Assessment of executive functions: review of instruments and identification of critical issues. Archives of Clinical Neuropsychology 23, 201–216.

Christensen, A.-L., 1989. The neuropsychological investigation as a therapeutic and rehabilitative technique. In: Ellis, D.W., Christensen, A.-L. (Eds.), Neuropsychological Treatment after Brain Injury. Kluwer Academic Publishers, Norwell, MA, USA.

Christensen, A.-L., Caetano, C., 1996. Alexander Romanovich Luria (1902–1977): contributions to neuropsychological rehabilitation. Neuropsychological Rehabilitation 6, 279–304.

Clare, L., Linden, D.E.J., Woods, R.T., Whitaker, R., Evans, S.J., Parkinson, C.H., Van Paasschen, J., Nelis, S.M., Hoare, Z., Yuen, K.S.L., Rugg, M.D., 2010. Goal-oriented cognitive rehabilitation for people with early-stage Alzheimer disease: a single-blind randomized controlled trial of clinical efficacy. American Journal of Geriatric Psychiatry 18, 928–939.

Cloute, K., Mitchell, A., Yates, P., 2008. Traumatic brain injury and the construction of identity: a discursive approach. Neuropsychological Rehabilitation 18, 651–670.

Colotla, V., Bach-Y-Rita, P., 2002. Shepherd Ivory Franz: his contributions to neuropsychology and rehabilitation. Cognitive, Affective, & Behavioral Neuroscience 2, 141–148.

Corsi, P., 1991. The Enchanted Loom: Chapters in the History of Neuroscience. Oxford University Press, New York.

Damasio, A., 1994. Descartes' Error: Emotion, Reason, and the Human Brain. Putnam Publishing, New York.

Dell, G.S., Caramazza, A., 2008. Introduction to special issue on computational modelling in cognitive neuropsychology. Cognitive Neuropsychology 25, 131–135.

Della Sala, S., Laiacona, M., Trivelli, C., Spinnler, H., 1995. Poppelreuter-Ghent's overlapping figures test: its sensitivity to age, and its clinical use. Archives of Clinical Neuropsychology 10, 511–534.

Duncan, J., 1986. Disorganisation of behaviour after frontal lobe damage. Cognitive Neuropsychology 3, 271–290.

Evans, J.J., Gast, D.L., Perdices, M., Manolov, R., 2014. Single case experimental designs: Introduction to a special issue of neuropsychological rehabilitation. Neuropsychological Rehabilitation 24, 305–314.

Evans, J.J., Wilson, B.A., Schuri, U., Andrade, J., Baddeley, A., Bruna, O., Canavan, T., Della Sala, S., Green, R., Laaksonen, R., Lorenzi, L., Taussik, I., 2000. A comparison of "errorless" and "trial-and-error" learning methods for teaching individuals with acquired memory deficits. Neuropsychological Rehabilitation 10, 67–101.

Fantie, B.D., Bhatt, S., Bailey, C., Kolb, B., 2001. Reduced salience of emotional facial expressions following closed head injury. Proceedings and Abstracts of the Annual Meeting of the Eastern Psychological Association. Washington, DC.

Fantie, B.D., Kolb, B., 1991. The problems of prognosis. In: Dyman, J., Kaplan, R.D., Pirozzolo, F.J. (Eds.), Neuropsychology and the Law. Springer-Verlag, New York, pp. 186–238.

Finger, S., 1994. Origins of Neuroscience: A History of Explorations into Brain Function. Oxford University Press, New York.

Finger, S., 2000. Minds behind the Brain: A History of the Pioneers and Their Discoveries. Oxford University Press, Oxford; New York.

Foerster, O., 1930. Restitution der motilitat: restitution der sensibilitat. Deutsche Zeitschrifi fiir Nervenheilkunde 115, 248–314.

Franz, S.I., 1905. The re-education of an aphasic. Journal of Philosophy, Psychology, and Scientific Methods 2, 589–597.

Franz, S.I., 1923. Nervous and Mental Re-education. The Macmillan Company, New York.

Goldstein, K., 1939. The Organism: A Holistic Approach to Biology Derived from Pathological Data in Man. American Book Publishing, Salt Lake City, UT.

Goldstein, K., 1942. Aftereffects of Brain Injuries in War. Grune and Stratton, New York.

Goldstein, K., 1963. Human Nature in the Light of Psychopathology. Schocken Books, New York.
Gracey, F., Ownsworth, T., 2008. Editorial. Neuropsychological Rehabilitation 18, 522–526.
Hall, G.S., 1920. Preface. In: Freud, S. (Ed.), A General Introduction to Psychoanalysis. Boni and Liveright, New York.
Haslam, C., Holme, A., Haslam, S.A., Iyer, A., Jetten, J., Williams, W.H., 2008. Maintaining group memberships: social identity continuity predicts well-being after stroke. Neuropsychological Rehabilitation 18, 671–691.
Hellebrandt, F.A., 1972. The physiology of motor learning. In: Singer, R.N. (Ed.), Readings in Motor Learning. Lea & Febiger, Philadelphia, PA.
Heyvaert, M., Onghena, P., 2014. Analysis of single-case data: randomisation tests for measures of effect size. Neuropsychological Rehabilitation 24, 507–527.
Jak, A., Seelye, A., Jurick, S., 2013. Crosswords to computers: a critical review of popular approaches to cognitive Enhancement. Neuropsychology Review 23, 13–26.
Koenig, S.T., Crucian, G.P., Dalrymple-Alford, J.C., Dunser, A., 2009. Virtual reality rehabilitation of spatial abilities after brain damage. In: Wiederhold, B.K., Riva, G. (Eds.), Annual Review of Cybertherapy and Telemedicine 2009: Advanced Technologies in the Behavioral, Social and Neurosciences, pp. 105–107.
Kolb, B., Muhammad, A., 2014. Harnessing the power of neuroplasticity for intervention. Frontiers in Human Neuroscience 8.
Krasny-Pacini, A., Limond, J., Evans, J., Hiebel, J., Bendjelida, K., Chevignard, M., 2014. Context-sensitive goal management training for everyday executive dysfunction in children after severe traumatic brain injury. Journal of Head Trauma Rehabilitation 29 (5).
Ledford, J.R., Gast, D.L., 2014. Measuring procedural fidelity in behavioural research. Neuropsychological Rehabilitation 24, 332–348.
Levine, B., Robertson, I.H., Clare, L., Carter, G., Hong, J., Wilson, B.A., Duncan, J., Stuss, D.T., 2000. Rehabilitation of executive functioning: an experimental-clinical validation of goal management training. Journal of the International Neuropsychological Society 6, 299–312.
Lezak, M.D., Howieson, D.B., Bigler, E.D., Tranel, D., 2004. Neuropsychological Assessment. Oxford University Press, Oxford, New York.
Lishman, W.A., 1978. Organic Psychiatry: The Psychological Consequences of Cerebral Disorder. Blackwell Scientific, Oxford.
Loring, D.W., 2010. History of neuropsychology through Epilepsy eyes. Archives of Clinical Neuropsychology 25, 259–273.
Luria, A.R., 1963. Restoration of Brain Function after Brain Injury. Pergamon Press, London.
Luria, A.R., 1966. Higher Cortical Functions in Man. Basic Books, New York.
Luria, A.R., 1972. The Man with a Shattered World: The History of a Brain Wound. Basic, New York.
Luria, A.R., 1973. The Working Brain; an Introduction to Neuropsychology. Basic Books, New York.
Marshall, L.H., Magoun, H.W., 1998. Discoveries in the Human Brain: Neuroscience Prehistory, Brain Structure, and Function. Humana Press, Totowa, NJ.
Miotto, E.C., Evans, J.J., Souza De Lucia, M.C., Scaff, M., 2009. Rehabilitation of executive dysfunction: a controlled trial of an attention and problem solving treatment group. Neuropsychological Rehabilitation 19, 517–540.
Monti, M.M., Vanhaudenhuyse, A., Coleman, M.R., Boly, M., Pickard, J.D., Tshibanda, L., Owen, A.M., Laureys, S., 2010. Willful modulation of brain activity in disorders of consciousness. New England Journal of Medicine 362, 579–589.
Morton, M.V., Wehman, P., 1995. Psychosocial and emotional sequelae of individuals with traumatic brain injury - a literature-review and recommendations. Brain Injury 9, 81–92.
Newcombe, F., 1969. Missile Wounds of the Brain: A Study of Psychological Deficits. Oxford U.P, London.
Ogden, R., Franz, S.I., 1917. On cerebral motor control: the recovery from experimentally produced hemiplegia. Psychobiology 1, 33–49.
Owen, A.M., Coleman, M.R., Boly, M., Davis, M.H., Laureys, S., Pickard, J.D., 2006. Detecting awareness in the vegetative state. Science 313, 1402.
Perdices, M., Tate, R.L., 2009. Single-subject designs as a tool for evidence-based clinical practice: are they unrecognised and undervalued? Neuropsychological Rehabilitation 19, 904–927.
Prigatano, G., 2000. A brief overview of four principles of neuropsychological rehabilitation. In: Christensen, A.-L., Uzzell, B.P. (Eds.), International Handbook of Neuropsychological Rehabilitation. Springer US.
Prigatano, G.P., 1999. Principles of Neuropsychological Rehabilitation. Oxford University Press, New York.
Prigatano, G.P., 2013. Challenges and opportunities facing holistic approaches to neuropsychological rehabilitation. Neurorehabilitation 32, 751–759.
Prigatano, G.P., Fordyce, D.J., Zeiner, H.K., Roueche, J.R., Pepping, M., Wood, B.C., 1986. Neuropsychological Rehabilitation after Brain Injury, Johns Hopkins University Press, Baltimore.
Raspelli, S., Carelli, L., Morganti, F., Poletti, B., Corra, B., Silani, V., Riva, G., 2010. Implementation of the multiple errands test in a Neuro VR-supermarket: a possible approach. Annual Review of Cybertherapy and Telemedicine 2010: Advanced Technologies in Behavioral, Social and Neurosciences 154, 115–119.
Raspelli, S., Carelli, L., Morganti, F., Riva, G., Weiss, P.L., Kizony, R., Katz, N., IEEE, 2009. Implementation of the Multiple Errand Test in a NeuroVR-supermarket, Virtual Rehabilitation International Conference, 2009. IEEE, Haifa Israel, pp. 210.
Raspelli, S., Pallavicini, F., Carelli, L., Morganti, F., Pedroli, E., Cipresso, P., Poletti, B., Corra, B., Sangalli, D., Silani, V., Riva, G., 2012. Validating the Neuro VR-based virtual version of the multiple errands test: preliminary results. Presence-Teleoperators and Virtual Environments 21, 31–42.
Raspelli, S., Pallavicini, F., Carelli, L., Morganti, F., Poletti, B., Corra, B., Silani, V., Riva, G., 2011. Validation of a neuro virtual reality-based version of the multiple errands test for the assessment of executive functions. Studies in Health Technology and Informatics 167, 92–97.
Raymer, A., Strobel, J., Prokup, T., Thomason, B., Reff, K.L., 2010. Errorless versus errorful training of spelling in individuals with acquired dysgraphia. Neuropsychological Rehabilitation 20, 1–15.
Rindskopf, D., 2014. Bayesian analysis of data from single case designs. Neuropsychological Rehabilitation 24, 572–589.
Robertson, I.H., 1996. Goal Management Training: A Clinical Manual, PsyConsult, Cambridge, UK.
Robertson, I.H., 2001. Brain damage: neuropsychological rehabilitation. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Pergamon, Oxford.
Russell, W.R., 1971. The Traumatic Amnesias. Oxford University Press, London.
Ryalls, J., Lecours, A.R., 1996. Broca's first two cases: from bumps on the head to cortical convolutions. In: Code, C., Wallesch, C.-W., Lecours, J.Y., Lecours, A.R. (Eds.), Classic Cases in Neuropsychology. Psychology Press, Hove, pp. 235–242.
Sarikcioglu, L., 2007. Otfrid Foerster (1873–1941): one of the distinguished neuroscientists of his time. Journal of Neurology, Neurosurgery & Psychiatry 78, 650.
Serino, A., Ciaramelli, E., Santantonio, A.D., Malag, S., Servadei, F., Làdavas, E., 2007. A pilot study for rehabilitation of central executive deficits after traumatic brain injury. Brain Injury 21, 11–19.
Shallice, T., Burgess, P.W., 1991. Deficits in strategy application following frontal lobe damage in man. Brain 114, 727–741.
Sullivan, E.V., Tapert, S.F., 2013. Introduction to the special issue of neuropsychology review on cognitive enhancement and rehabilitation. Neuropsychology Review 23, 10–12.
Tan, L.F., Dienes, Z., Jansari, A., Goh, S.Y., 2014. Effect of mindfulness meditation on brain-computer interface performance. Consciousness and Cognition 23, 12–21.
Tate, R.L., Mcdonald, S., Perdices, M., Togher, L., Schultz, R., Savage, S., 2008. Rating the methodological quality of single-subject designs and n-of-1 trials: introducing the single-case experimental design (SCED) Scale. Neuropsychological Rehabilitation 18, 385–401.
Weiss, P.L., Klinger, E., 2009. Moving beyond single user, local virtual environments for rehabilitation. In: Gaggioli, A., Keshner, E.A., Weiss, P.L., Riva, G. (Eds.), Advanced Technologies in Rehabilitation: Empowering Cognitive, Physical, Social and Communicative Skills through Virtual Reality, Robots, Wearable Systems and Brain-computer Interfaces. Amsterdam. Ios Press.
Welbourne, S.R., Ralph, M.A.L., 2005. Using computational, parallel distributed processing networks to model rehabilitation in patients with acquired dyslexia: an initial investigation. Aphasiology 19, 789–806.
Westerberg, H., Jacobaeus, H., Hirvikoski, T., Clevberger, P., Östensson, M.L., Bartfai, A., Klingberg, T., 2007. Computerized working memory training after stroke–A pilot study. Brain Injury 21, 21–29.
Wilson, B.A., 2011. 'Cutting edge' developments in neuropsychological rehabilitation and possible future directions. Brain Impairment 12, 33–42.
Wilson, B.A., Baddeley, A., Evans, J., Shiel, A., 1994. Errorless learning in the rehabilitation of memory impaired people. Neuropsychological Rehabilitation 4, 307–326.
Wilson, B.A., Kopelman, M., Kapur, N., 2008. Prominent and persistent loss of past awareness in amnesia: delusion, impaired consciousness or coping strategy? Neuropsychological Rehabilitation 18, 527–540.
Witsken, D.E., D'amato, R.C., Hartlage, L.C., 2008. Understanding the past, present, and future of clinical neuropsychology. In: D'amato, R.C., Hartlage, L.C. (Eds.), Essentials of Neuropsychological Assessment: Treatment Planning for Rehabilitation, second ed. Springer Publishing Company, New York, pp. 3–29.
Yeates, G.N., Gracey, F., Mcgrath, J.C., 2008. A biopsychosocial deconstruction of "personality change" following acquired brain injury. Neuropsychological Rehabilitation 18, 566–589.
Zangwill, O.L., 1947. Psychological aspects of rehabilitation in cases of brain injury. British Journal of Psychology 37, 60–69.

Relevant Websites

http://wfnr.co.uk/.
http://www.assbi.com.au/index.html.
http://www.biausa.org/.
http://www.cdc.gov/traumaticbraininjury/.
https://www.nanonline.org/.
http://www.ninds.nih.gov/disorders/tbi/tbi.htm.
http://www.societyforcognitiverehab.org/index.php.
https://www.theaacn.org.
http://www.the-ins.org/.
www.internationalbrain.org/.

Brain Development: Ontogenetic Neurobiology of

Robbin Gibb, University of Lethbridge, Lethbridge, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.I. Posner and M.K. Rothbart, volume 2, pp. 1332–1338, © 2001, Elsevier Ltd.

Abbreviations

bRGC Basal radial glia cell
CR Cajal retzius
CNS Central nervous system
Cl^- Chloride
E Embryonic
GABA γ-Aminobutyric acid
ICI Inhibitory cortical interneuron
IPC Intermediateprecursor cell
Na^+ Sodium
NEC Neuroepithelial cell
OPC Oligodendrocyte progenitor cell
RGC Radial glial cell
RMS Rostral migratory stream
SNP Short neural precursor
SVZ Subventricular zone
VZ Ventricular zone

Abstract

The development of the human brain is a prolonged process that begins 3 weeks postconception and continues well into adulthood. The phases of brain development are outlined and discussed in relation to the prenatal and postnatal periods of life. The features of brain construction that promote neural plasticity are discussed in the context of the interaction of our inherited genes (nature) and our environmental experiences (nurture).

Introduction

Human brain development is a protracted process that begins shortly after conception and continues at least into the third decade of life. The brain is crafted by the lifelong interplay of generative (cell birth and synapse formation) and degenerative processes (cell death and synaptic pruning), which are modulated to varying degrees by an individual's experience. The process of brain development is intricately programmed by information carried on DNA. Although the process is fundamentally the same for each individual, environmental exposures can alter the manner in which the program becomes manifest. This environmental modulation of genetic expression is called epigenetic programming and its effects can be seen in the preconception period (e.g., Mychasiuk et al., 2013), the prenatal period, (e.g., Mychasiuk et al., 2012) and through the life span of the individual (e.g., Harker et al., 2013). Mounting evidence demonstrates that experiences of our forebears can and often do result in epigenetic changes in gene expression ultimately changing the expression of proteins. Because proteins are the building blocks of the brain, different proteins build different brains by modifying cell number, cell connectivity, brain size, and ultimately, behavior. Epigenetic programming thus provides an adaptive means for an organism to prepare its brain for the unique environmental challenges that it will face without changing its genetic blueprint.

Phases of Brain Development

The process of brain construction occurs in seven well-defined phases that extend over a prolonged period of brain development (Kolb and Whishaw, 2014). Some phases are more or less confined to restricted periods, whereas other phases are in play for extended periods of time. The seven phases of brain development are

1. cell birth (neurons, glia, microglia precursors),
2. cell migration,
3. cell differentiation,
4. cell maturation,
5. synaptogenesis,
6. cell death and synaptic pruning, and
7. myelination.

Each phase will be considered in depth in the developmental epoch(s) during which it predominately occurs.

Formation of the Brain: Embryonic Development (Conception to Week 8)

By embryonic day 13 (E13), the primitive node and the primitive streak appear during the process known as gastrulation, whereby the single-layer blastula is replaced by a trilayer structure. At the end of gastrulation (E20), the endodermal,

mesodermal, and ectodermal layers of the trilamina have formed. The ectodermal layer gives rise to both the skin and the central nervous system (CNS). The neuroepithelial cells (NECs), which eventually produce cells that make the neurons and glial that form the brain, are found within this layer. The neural plate forms on E21, and by E22 the neural groove becomes apparent. The neural groove fuses starting on E23 to form the neural tube (neurulation) with the central section closing first. The rostral portion of the neural tube is inhabited by the earliest migrating cells and becomes the brain, while the caudal portion receives later migrating cells and forms the spinal cord (Stiles and Jernigan, 2010). The rostral and caudal regions of the neural tube are the last to close. Neural tube defects are birth defects that affect the brain or spinal cord. Failure of the rostral section of the neural tube to close on E23–E26 results in a condition known as anencephaly, where the brain and skull fail to fully develop. Infants born with anencephaly usually lack sensory processing abilities and are unable to feel pain. They generally die shortly after birth. Spina bifida is the result of incomplete closure of some portion of the caudal end of the neural tube on about E28. The symptoms of spina bifida range in severity, but it is most commonly characterized by paralysis of the legs. Both anencephaly and spina bifida commonly result from a lack of the B vitamin, folate, in the diet. The cylindrical cavity in the neural tube eventually becomes the ventricular system and the NECs inhabit this region, also known as the ventricular zone (VZ) (*Phase 1-cell birth*). The NECs are aligned in the ventricular zone in an apicobasal fashion. At the ventricular (apical) surface, the NECs are bound together by both *tight* and *adherens junctions* and at the pial (basal) surface by *integrins*. During neurulation, the neural stem cell population undergoes rapid expansion by symmetric cell division of the NECs. A primitive map that provides a basic blueprint for brain organization emerges at this stage of development. In order to supply the demand for neurons to populate the brain, on E25 the NECs divide in a symmetrical fashion. Symmetrical cell division produces two NECs with each division and this process continues until E42. By E28, the rudimentary structures of the forebrain (prosencephalon), midbrain (mesencephalon), and hindbrain (rhombencephalon) are visible in the human embryo (Stiles and Jernigan, 2010). Just prior to neurogenesis, NECs begin to transform into radial glial cells (RGCs). Recent work on neural progenitors has led to an understanding that diverse populations of progenitors exist. Three new subclasses: the intermediate progenitor cell (IPC), the short neural precursor (SNP) with an apical attachment at the VZ but no basal attachment at the pial surface, and basal radial glial cells (bRGCs) with basal attachments at the pial surface but no apical attachment in the VZ, have been described (Franco and Muller, 2013). Traditionally, RGCs were thought to divide asymmetrically to self-renew and provide a non-RGC daughter cell (E42–E108). It turns out that some but not all of the daughter cells are postmitotic neurons. Most of these daughter cells are IPCs. IPCs undergo symmetrical divisions to self-renew but most supply pairs of neurons. The IPCs occupy a basal position in the ventricular zone and ultimately form the proliferative region known as the subventricular zone. SNPs are similar in function to the IPCs in that they produce neurons but they differ in that they maintain an apical ventricular attachment and remain in the VZ. The bRGCs that occupy the outermost region of the SVZ self-renew and generate IPCs and neurons. At the end of neurogenesis, (E108; Clancy et al., 2001) both RGCs and bRGCs transform into progenitors that produce astrocytes (**Figure 1**).

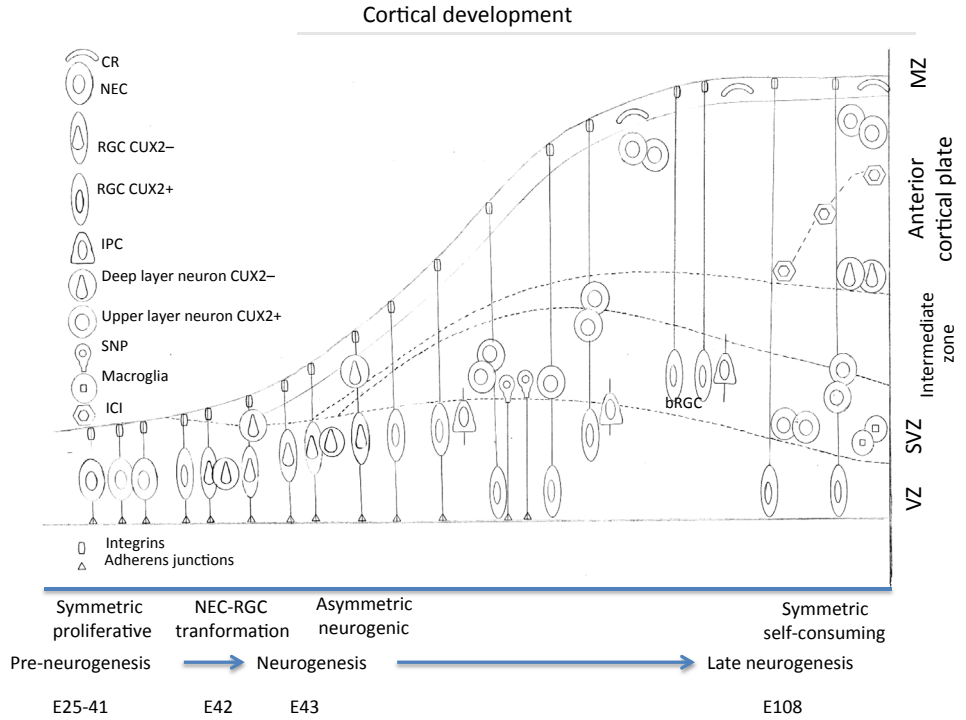

Figure 1 After Paridaen and Huttner, 2014.

Formation of the Brain: Fetal Development (Week 9 to Birth)

During this phase of brain development, the previously smooth brain acquires the folding pattern with gyri (bumps) and sulci (grooves) typically seen in the mature brain. After cells are born in the VZ or SVZ, they migrate away in a radial fashion with the cells that form the deepest layers of the cerebral cortex leaving first (*Phase 2-cell migration*). As cells continue to form, the layers gradually build with the newest arrivals seated atop those most recently placed in a so-called 'inside out' fashion. A notable exception to the inside out rule is the first wave of cells that leave the proliferative zone. These cells form a primordial structure known as the preplate (Bayer and Altman, 1991). Once formed, the preplate is split into two by the next wave of arriving cells producing a transient subplate and marginal zone (layer 1). The new region that these cells form is known as the cortical plate, and the first cells to populate the cortical plate are the cells destined to become layer 6 cells. The marginal zone contains a population of cells called Cajal–Retzius (CR) cells. These cells provide important signaling information for newly arriving neurons to position them in their correct layer in the cortex. The CR cells produce a substance called *reelin* that signals migrating neurons to stop their journey and take a position in the developing cerebral cortex. Animals with reelin deficiencies fail to develop laminar structure in the cerebral cortex. In humans, abnormalities in the reelin signaling cascade are implicated in a variety of neurodevelopmental disorders including schizophrenia, autism, major depression, and bipolar disorder (Folsom and Fatemi, 2013). RGCs that are anchored at the ventricular zone and the pial surface form the guideposts for the migrating cells. Newly born neurons climb along the radial glia fibers to reach their destination, and until recently, it was thought that all of the neurons that climb an RCG are progeny of that cell. Recent experimental evidence has cast doubt on this view. Because different cell types inhabit the different cortical layers, it has been surmised that the cortical progenitors have the ability to produce a broad class of neural cell types (*Phase 3-cell differentiation*). IPCs produce neurons whereas, RGCs give rise to both projection neurons and astrocytes (Franco and Muller, 2013). As successive layers of the cortical mantle are formed, the progenitors become more limited in the cell types that they are able to construct. This process is known as fate restriction and it is thought to account for how progenitors are able to manufacture the variety of cells required to build the brain. In a recent paper, Franco et al. (2012) describe work done in a murine model that demonstrates that birthdate alone does not account for the ultimate position of neurons in particular cortical layers. Rather, they have identified RGCs sublineages that are responsible for producing either the lower (V and VI) or upper (II, III, and IV) cortical lamina. IPCs and some RGCs appear to bear the same transcription factor (Cux2 and are thus Cux2+) as neurons destined for the upper cortical layers. RGCs that lack expression of Cux2 (Cux2−) generate neurons destined for lower cortical layers. Results of this work suggest that birthdate may not be the cause for layer specification of the cerebral cortex. It appears that IPCs and RGCs that produce upper-layer cortical neurons just happen to do so at a later time than the RGCs that generate the lower-layer neurons. This work suggests that cell fate specification drives birth order and not the reverse as was previously believed. Cux2− RGCs stop symmetrical divisions and start to produce neurons before Cux2+ cells, allowing expansion of the Cux2+ progenitor pool. The authors of this work note that upper cortical layers are expanded in primates and are necessary for sophisticated associative connectivity. They propose that defects in the generation of these cells could result in conditions such as autism and schizophrenia. Most radially migrating cells are glutamatergic (use glutamate as their principal neurotransmitter) and comprise approximately 80% of cortical neurons.

Inhibitory cortical interneurons (ICIs) arise from different origins. Three ganglionic eminences, the medial, lateral, and caudal are confirmed sources of ICIs, whereas the rostral migratory stream, the septal region, and the anterior region of the cortical plate may be additional generative sites (Wonders and Anderson, 2006; Al-Jaberi et al., 2013) ICIs have multiple phenotypes based on their morphology, neurochemical constituents, and physiology. A uniting feature of ICIs is their use of the inhibitory γ-aminobutyric acid (GABA) as their neurotransmitter. Postmitotic ICIs migrate in what is called a tangential fashion and rely on a number of molecular guides or tropic factors to reach their destination. The guidance molecules arise from local regions of the cortex and help the interneurons reach their final destination (Stiles and Jernigan, 2010). Proper functioning of the cerebral cortex depends on ICIs (20% of cortical neurons). In addition to the key role that ICIs play in modulating cortical output, they are also involved in developmental processes including regulation of neuron proliferation, migration, and plasticity of cortical circuitry (Wonders and Anderson, 2006). A fascinating feature of the developing cortex is the depolarizing effects of GABA in the immature brain. Although this phenomenon has been recently called into question (Bregestovski and Bernard, 2012), overwhelming evidence seems to support the finding that in immature ICI's, GABA has an excitatory action by depolarizing membranes and sometimes even causing sodium (Na^+) mediated action potentials (Ben-Ari et al., 2012). Excessive accumulation of the chloride ion (Cl^-) in immature neurons may mediate the excitatory action of GABA, and indeed the shift in GABA polarity is concurrent with a progressive reduction of intracellular Cl^-. It appears that this inverted excitatory action of GABA during development may have an organizational role in cell birth, migration (especially the tangentially migrating ICIs), synaptogenesis, and network formation. Tyzio et al. (2006) describe the influence of oxytocin during delivery as the cellular switch that changes the polarity of GABA function to inhibition. Failure to assume the fully mature GABA system is associated with the appearance of epileptic seizures (Aronica et al., 2011).

Once neurons reach their final destination, they extend dendrites and an axon (*Phase 4-cell maturation*) in an attempt to make connections with other cells and become an integral part of a communication network. Dendrites gather information from other neurons and the axon provides a mean of sending information to neurons further down the communication line. Many dendrites extend from a neuron to receive input from cells in the information network, but a single axon conveys

information processed by the cell. In order to make appropriate contacts, the axon has a growth cone on its leading edge. The growth cone is guided by sampling locally produced tropic molecules that ultimately assist the axon in finding its intended target. Once that target is identified, a connection called a synapse is formed (*Phase 5-synaptogenesis*). The synapse provides the means for cell-to-cell communication. In the context of the synapse, the axon is considered the presynaptic terminal and the dendrite, the postsynaptic terminal.

Interestingly, the formation of the brain also depends on degenerative processes that commence in the prenatal period of life. Programmed cell death or apoptosis (*Phase 6-cell death and synaptic pruning*) is instigated to reduce the number of cells in the brain that have failed to make useful connections or have connections that are underutilized (Chan et al., 2002). During this process, a cascade of death promoting genes are expressed that initiate a suicide program in the cell (for more details, see Section Formation of the Brain: Postnatal Development). Cells that have established stable and vibrant connections are protected from apoptosis by neurotrophic factors supplied by other cells in the circuit.

Formation of the Brain: Postnatal Development

The process of neurogenesis (*Phase 1-cell birth*) persists through the postnatal period in a restricted manner. Cells continue to form in the SVZ, which are destined to populate the olfactory bulb and they then migrate (*Phase 2-cell migration*) via the rostral migratory stream (RMS) to their target site. Postnatal neurons are also born in the dentate granule layer in the hippocampus. Neural proliferation in these two privileged areas continues throughout the life span. Proliferation and migration of cells destined to become glial cells begins in the prenatal period and continues for an extended time in the postnatal period. Glial progenitors multiply in the SVZ and then migrate out to the white matter eventually reaching the cortex, striatum, and hippocampus. The mode of emigration for glial progenitors is dependent on their ultimate destination. Cells destined for dorsal areas migrate in a radial fashion to the overlying corpus callosum, where they remain or continue on along RGCs to their target in cerebral cortex. Cells destined for lateral cortex follow white matter laterally and then exit to follow RGCs to their terminal location. Some glial progenitors track along the corpus callosum to reach the opposite hemisphere and others follow blood vessels to their target (Cayre et al., 2009). It appears that glial progenitors, particularly oligodendrocyte progenitor cells (OPCs) are retained in the cerebral cortex for an indefinite period of time. Even the adult brain has the ability to produce astroglia and oligodendroglia and both can arise from OPCs. Synaptogenesis (*Phase 5-synaptogenesis*) that began in the prenatal period continues in the postnatal period and throughout the life span of an individual. Dendrites on a single neuron can make tens of thousands of connections with other neurons and it appears that the number of connections varies with the cortical area. At the site of contact, the neuron usually elaborates a dendritic spine to enhance the effectiveness of the synaptic site. Fewer dendritic spines are observed in posterior regions (visual areas) than in anterior regions of the brain (frontal lobe) and in the most anterior areas (prefrontal cortex), estimates of 23 times more synapses than in primary visual cortex have been derived (Elston, 2003). Most dendritic spines are sites of excitatory neural transmission. Excitatory synapses, also known as Type I synapses, are morphologically distinct from inhibitory or Type II synapses (Gray, 1959). Type I synapses are characterized by round synaptic vesicles in the presynaptic terminal, a wide synaptic cleft or space between the presynaptic and postsynaptic membranes, a large active zone, and a dense network of proteins on the postsynaptic membrane known as the postsynaptic density. Inhibitory synapses are conversely identified by oval or flattened synaptic vesicles in the presynaptic terminal, a narrow synaptic cleft, a small active zone, and a sparse postsynaptic density (Kolb and Whishaw, 2014).

Apoptosis (*Phase 6-cell death and synaptic pruning*) continues to play a major role in brain development in the postnatal period. 'Use it or lose it' becomes the rationale for brain architecture at this juncture, and it does not just apply to dormant cells. The exuberant production of synapses is also subject to this philosophy. Underused synapses are removed in a process called synaptic pruning. Recently, the role of microglia in both apoptosis and synaptic pruning has been described (Paolicelli et al., 2011; Wake et al., 2012). Microglia are not derived through the same precursor cells as neurons and astroctyes, but rather originate from mesodermal myeloid progenitors that also produce macrophages and blood cells. Microglia immigrate to the CNS early in embryogenesis where they proliferate. During development, microglia typically display an ameboid appearance in contrast to the ramified phenotype that characterizes microglia in the adult CNS. Estimates are that approximately 50–70% of the cells born into the brain are removed through apoptosis (the number varying with the location within the CNS; Stiles and Jernigan, 2010; Rabinowcicz et al., 1996), and approximately 40–50% of synaptic connections are lost through synaptic pruning (Low and Cheng, 2006). Cell loss and synaptic pruning provides a means for tuning brain connections by ramping up those that are important and eliminating those that are not.

Although myelination (*Phase 7-myelination*) proceeds to a limited extent in the prenatal period, it ramps up after birth and continues well into the third decade of life. The process of myelination frequently heralds the maturation of cortical areas. Primary motor and sensory areas of the brain myelinate first and association areas myelinate last. For example, the first cortical area to myelinate is motor cortex and myelination of the dorsolateral prefrontal cortex occurs last (Flechsig, 1901). OPCs upon reaching their final destination differentiate. Processes are extended and begin to form myelin that wraps around axons and acts as insulation to speed up the transmission of electrical signals. For myelination to occur, an ample supply of an energy substrate needs to be readily available. In the absence of a readily available energy source, myelination is impaired and OPC's risk enacting the apoptotic pathway. Although glucose is often that substrate under certain conditions of intense neural activity, lactate may provide another resource for the energy required. In cases of anoxia during myelination resulting in cerebral palsy or periventricular leukomalacia, lactate may be an important factor in minimizing

the devastating outcomes by supporting integrity of immature oligodendrocytes (Rinholm et al., 2011). It is clear that the role of oligodendroglia extends far beyond that of just insulating axonal fibers. In a recent paper (Morrison et al., 2013), the role of oligodendroglia in supporting axonal health and function was examined. It appears that axons are metabolically costly to maintain and once myelinated, have limited access to the extracellular fluid to fulfill their energy requirements. Oligodendroglia may provide an important link to energy supplies for neurons. Another recent study compared the time course of myelination in humans and chimpanzees, our closest living relatives (Miller et al., 2012). The authors in this study discovered that while chimpanzees completed the process of cortical myelination by the time of sexual maturity, in humans this process extended well beyond late adolescence. They conclude that extended timing of myelination in the human brain may be in part responsible for the greater degree of cognitive development and cortical plasticity observed in humans. They also propose that this extended period of maturation makes the human brain more vulnerable to psychiatric disorders that have typical onset in adolescence or early adulthood like schizophrenia.

Experience and Brain Development

The rudimentary plan for constructing the human brain is encoded in the genome, but much of the connectivity is unspecified and depends on the input of experience to complete the wiring. Processes that proceed without the need of external experiences have been described as experience independent (Shatz, 1992). Such developmental processes rely on internally generated activity to proceed rather than externally generated environmental input. An example of such a process is the formation of layers in the lateral geniculate nucleus based on firing patterns of retinal ganglion cells. Experience-expectant plasticity (Greenough et al., 1987) occurs when the brain uses information derived from the environment to drive the normal development of a structure. If the expected experience is unavailable, the structure will develop in an abnormal fashion or will fail to develop. The development of ocular dominance columns in visual cortex is achieved through experience-expectant plasticity. Experience also modifies the construction of the brain through a process termed experience-dependent plasticity (Greenough et al., 1987). In this form of plasticity, structures in the developing brain do not require particular experiences but may be modified or changed by environmental exposures. Learning is a form of experience-dependent plasticity. All seven phases of brain development can be altered to a certain extent by experience. Experience-dependent plasticity allows the developing brain to change responsively to experiential inputs and thus adapt to the unique environmental challenges an individual may face.

Summary

Brain development has a protracted time course in humans beginning in the third week postconception and extending to the third decade of life. The brain develops in seven well-defined phases that overlap and are repeated over the course of development. Neurodevelopmental disorders arise when processes engaged during construction of the brain go awry. Experiences of an individual play a fundamental role in producing a brain uniquely constructed to complement the environmental niche that individual occupies.

See also: Epigenetics of Brain Plasticity and Behavior; Epilepsy; Neural Plasticity; Neurotransmitters; Schizophrenia: Metabolic Consequences.

Bibliography

Al-Jaberi, N., Lindsay, S., Sarma, S., Bayatti, N., Clowry, G.J., 2013. The early fetal development of human neocortical GABAergic interneurons. Cerebral Cortex. http://dx.doi.org/10.1093/cercor/bht254 bht254[pii].

Aronica, E., Iyer, A., Zurolo, E., Gorter, J.A., 2011. Ontogenetic modifications of neuronal excitability during brain maturation: developmental changes of neurotransmitter receptors. Epilepsia 52 (Suppl. 8), 3–5.

Bayer, S.A., Altman, J., 1991. Neocortical Development. Raven Press, New York.

Ben-Ari, Y., Woodin, M.A., Sernagor, E., et al., 2012. Refuting the challenges of the developmental shift of polarity of GABA actions: GABA more exciting than ever! Frontiers in Cellular Neuroscience 6, 35. http://dx.doi.org/10.3389/fncel.2012.00035.

Bregestovski, P., Bernard, C., 2012. Excitatory GABA: how a correct Observation may turn out to be an experimental Artifact. Frontiers in Pharmacology 3, 65.

Cayre, M., Canoll, P., Goldman, J.E., 2009. Cell migration in the normal and pathological postnatal mammalian brain. Progress in Neurobiology 88, 41–63.

Chan, W.Y., Lorke, D.E., Tiu, S.C., Yew, D.T., 2002. Proliferation and apoptosis in the developing human neocortex. Anatomical Record 267, 261–276.

Clancy, B., Darlington, R.B., Finlay, B.L., 2001. Translating developmental time across mammalian species. Neuroscience 105, 7–17.

Elston, G.N., 2003. Cortex, cognition and the cell: new insights into the pyramidal neuron and prefrontal function. Cerebral Cortex 13 (11), 1124–1138.

Flechsig, P., 1901. Developmental (myelogenetic) localisation of the cerebral cortex in the human subject. Lancet 158, 1027–1030.

Folsom, T.D., Fatemi, S.H., 2013. The involvement of Reelin in neurodevelopmental disorders. Neuropharmacology 68, 122–135.

Franco, S.J., Gil-Sanz, C., Martinez-Garay, I., Espinosa, A., Harkins-Perry, S.R., Ramos, C., Muller, U., 2012. Fate-restricted neural progenitors in the mammalian cerebral cortex. Science 337, 746–749.

Franco, S.J., Muller, U., 2013. Shaping our minds: stem and progenitor cell diversity in the mammalian neocortex. Neuron 77, 19–34.

Gray, E.G., 1959. Axo-somatic and axo-dendritic synapses of the cerebral cortex: an electron microscope study. Journal of Anatomy 93, 420–433.

Greenough, W.T., Black, J.E., Wallace, C.S., 1987. Experience and brain development. Child Development 58, 539–559.

Harker, A.F., Nakahashi, A., Kolb, B., Gibb, R., 2013. Paternal stress changes offspring behavior lifelong and reduces behavioral recovery after brain injury in rats. Society for Neuroscience Abstracts 38.

Kolb, B., Whishaw, I.Q., 2014. An Introduction to Brain and Behavior, fourth ed. Worth Publishers, New York, NY.

Low, L.K., Cheng, H.J., 2006. Axon pruning: an essential step underlying the developmental plasticity of neuronal connections. Philosophical Transactions of the Royal Society of London B: Biological Science 361 (1473), 1531–1544. http://dx.doi.org/10.1098/rstb.2006.1883, 463621T5G6131557 [pii].

Miller, D.J., Duka, T., Stimpson, C.D., et al., 2012. Prolonged myelination in human neocortical evolution. Proceedings of the National Academy of Sciences of the United States of America 109, 16480–16485.

Morrison, B.M., Lee, Y., Rothstein, J.D., 2013. Oligodendroglia: metabolic supporters of axons. Trends in Cell Biology 23 (12), 644–651. http://dx.doi.org/10.1016/j.tcb.2013.07.007. S0962–8924(13)00120-7 [pii].

Mychasiuk, R., Harker, A., Ilnytskyy, S., Gibb, R., 2013. Paternal stress prior to conception alters DNA methylation and behaviour of developing rat offspring. Neuroscience 241, 100–105.

Mychasiuk, R., Zahir, S., Schmold, N., Ilnytskyy, S., Kovalchuk, O., Gibb, R., 2012. Parental enrichment and offspring development: modifications to brain, behavior and the epigenome. Behavioural Brain Research 228, 294–298.
Paolicelli, R.C., Bolasco, G., Pagani, F., et al., 2011. Synaptic pruning by microglia is necessary for normal brain development. Science 333, 1456–1458.
Paridaen, J., Huttner, W., 2014. Neurogenesis during development of the vertebrate nervous system. EMBO Reports. http://dx.doi.org/10.1002/embr.201438447.
Rabinowicz, T., de Courten-Myers, G.M., Petetot, J.M., Xi, G., de los Reyes, E., 1996. Human cortex development: estimates of neuronal numbers indicate major loss late during gestation. Journal of Neuropathology and Experimental Neurology 55 (3), 320–328.
Rinholm, J.E., Hamilton, N.B., Kessaris, N., Richardson, W.D., Bergersen, L.H., Attwell, D., 2011. Regulation of oligodendrocyte development and myelination by glucose and lactate. Journal of Neuroscience 31, 538–548.
Shatz, C.J., 1992. The developing brain. Scientific American 267, 60–67.
Stiles, J., Jernigan, T.L., 2010. The basics of brain development. Neuropsychology Review 20, 327–348.
Tyzio, R., Cossart, R., Khalilov, I., Minlebaev, M., Hubner, C.A., Represa, A., Ben-Ari, Y., Khazipov, R., 2006. Maternal oxytocin triggers a transient inhibitory switch in GABA signaling in the fetal brain during delivery. Science 314, 1788–1792.
Wake, H., Moorhouse, A.J., Miyamoto, A., Nabekura, J., 2013. Microglia: actively surveying and shaping neuronal circuit structure and function. Trends in Neuroscience 36 (4), 209–217. http://dx.doi.org/10.1016/j.tins.2012.11.007. S0166-2236(12)00204-4 [pii].
Wonders, C.P., Anderson, S.A., 2006. The origin and specification of cortical interneurons. Nature Reviews Neuroscience 7, 687–696.

Relevant Websites

http://www.albertafamilywellness.org/.
http://www.nimh.nih.gov/health/publications/brain-development-during-childhood-and-adolescence/index.shtml.
http://www.ninds.nih.gov/disorders/cephalic_disorders/detail_cephalic_disorders.htm.
https://notes.utk.edu/bio/greenberg.nsf/3966e623d80328cc85257015006e192f/0b49219cfc7806eb85256d270057a91c?OpenDocument.
http://www.withthebraininmind.org/buildingbrains/unit_1/unit_1_01.php.
http://www.zerotothree.org/child-development/brain-development/.

Brain, Evolution of

Ralph L Holloway, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The evolution of the human brain has been a largely mosaic affair, in which changes of reorganization of brain components, increases of brain size through both hyperplasia and hypertrophy, underlain by neurogenomic changes, have occurred sometimes independently, sometimes together. The direct evidence for these changes relies on paleoneurology based on endocast studies (where casts are made of the inside of the cranial cavity), which provide the best evidence for volume changes through time and occasionally some convolutional details of the once-pulsating cerebral cortex. Both comparative neuroscience and neurogenomics provide essential information of brain and behavioral variations, which allow more accurate interpretation of the endocasts, which are often poor in detail even though the extant species studied are not direct ancestors to humans. Reorganizational changes include a reduction of primary visual cortex and relative enlargement of posterior association cortex, expanded Broca's regions, as well as cerebral asymmetries. The size of the hominid brain increased from about 450 ml, 3.5 Ma, to our current average volume of 1350 ml. These two changes through time were sometimes gradual but not always.

The evolution of the human brain has largely been a matter of integrating both increases in the size of the brain, and the brain's organization through the past 3–4 My. Three lines of evidence are used by paleoneurologists to ascertain how these events might have occurred: (1) Direct evidence is from the brain endocasts of fossil hominids. (2) The indirect evidence is from comparative neuroscience, where variations in brain structures can be related to variations in behavior. The latter evidence is 'indirect' because extant living animals are not ancestral to humans. Indeed, the last common ancestor for apes and the hominid line existed some 5–7 Ma. (3) Newer neurogenomics evidence also promises to provide important clues to how and when certain aspects of brain changes occurred during human evolution.

The human brain evolved from an early hominid 3–4 Ma, *Australopithecus afarensis*, having a size of roughly 400 ml to our present average of 1330 ml. These brain size increases, at different taxonomic levels, were mostly allometric, that is, related to body size, but not always. Integrated with these changes in brain size was reorganization of the cerebral cortex, as well as subcortical structures such as the hippocampus, amygdala, and so forth, to mention a few important structures relating to aspects of social behavior that cannot be seen on endocasts.

At least three areas of the cerebral cortex were affected at different times: (1) a relative reduction of primary visual striate cortex (V1, PVC) and an attending relative increase in posterior parietal association cortex; (2) a change in Broca's region, resulting in a more human-like pattern; and, (3) increasing degrees of cortical asymmetry, as well as increases in overall brain size, and number of neurons.

When did the human brain evolve, and how did it happen? Obviously, to answer this question will require a time machine, and thousands of generations of observations to ascertain both the variability and direction of selection pressures in the past. We can, however, flesh out an initial understanding of how we got to be the animal *par excellence* that utilizes its brain for intelligent rationalizations based largely on the use of arbitrary symbol systems. The evidence consists of two components: (1) the 'direct' evidence from the fossil record; and (2) the 'indirect' evidence of the comparative neuroscientific record of extant living animals, particularly those most closely related to us, such as the chimpanzee and bonobo. There is a third possibility: since the human genome project has sequenced almost all of the genetic code, the future study of evolutionary neurogenomics might provide more data about the actual genetic history of our genus through time, as well as that of the great apes mentioned previously. As this latter possibility is simply a mote in our eye at present, this article must concentrate on the evidence provided by the first two components.

Lines of Evidence

Direct Evidence

The term *paleoneurology* is used to describe evidence appraising the size and morphology of the casts made from the inside of actual fossil cranial remains. Occasionally, the casts are 'natural,' that is, where fine sediments have filled the inside of the cranial cavity, and become compacted through time. These casts often retain whatever morphological details were imprinted on the internal table of bone of the cranium. The famous australopithecine Taung child's skull, described by Dart (1925), is one of the best-known examples. Curiously, these 'natural' endocasts are only found in the S. African australopithecines, of which several exist, dating from about 3.0 My to about 1.5 My. Most often, the paleoneurologist makes a cast of the inside of the fossil skull using rubber latex, or silicone rubber, and extracts this from the cranium. The partial cast is reconstructed by adding Plasticine (modeling clay) to the missing regions. The whole is then measured by immersion into water, and the amount of water displaced is regarded as the volume of the once-living brain. Other measurements and observations are made on the original cast. More recently, 'virtual' endocasts have been made from CT scans of intact or partial crania, which have the advantage of being noninvasive, and as the method is computer driven, there are different algorithms for measuring the size of the endocast and other metrics (Weber et al., 2012).

During life the brain is surrounded by three dural tissues (dura mater, arachnoid tissue and its cerebrospinal fluid, and pia mater) that interface between the actual brain tissue (cerebral cortex, mostly) and the internal table of bone of the skull. The gyri and sulci (convolutions) of the once-pulsating cerebral cortex are rarely imprinted on the interior of the skull, and the degree of replication often varies in different regions, for example, sometimes the frontal lobe imprints more details than the parietal lobe, and also by age. The degree of replication also varies in different animals. Two extremely important considerations emerge from this: (1) the resulting imprints are never complete and thus 'data poor,' and never include subcortical structures; and (2) controversial interpretations of what the underlying brain once looked like are guaranteed. Nevertheless, these endocranial casts do provide extremely important information regarding size, shape, rough estimates of the lobar dimensions of the brain, and cortical asymmetries that have relationships to hemispheric specializations, including handedness. In addition, if the imprints of the underlying gyri and sulci (i.e., the characteristic shapes of the brain) are available, these can provide important information regarding the organization of the cerebral cortex, and whether the patterns of these are the same or different as in known extant primate brains. The infamous 'lunate sulcus' is a good example, as it is a demarcation boundary between purely sensory primary visual striate cortex (PVC) and multimodal association cortex. When the lunate sulcus appears in an anterior position, it is most similar to the condition known in modern apes. When it is found in a posterior position, it is more in a human-like condition. Ascertaining its correct position is essential in deciding whether or not such a fossil hominid had a brain organized along human or ape lines. In modern humans, the 'lunate' is only partially homologous with that found in apes, and is usually fragmented (Allen et al., 2006). Finally, meningeal arteries and veins that nourished the dura mater also imprint on the internal table of bone, and these sometimes show patterns that are useful for deciding taxonomic issues; these have no known relationship to behavioral functions of the brain. (See also Grimaud-Hervé in Holloway et al., 2004; for further discussion and illustrations.)

Indirect Evidence

This line of evidence is 'data rich,' providing comparative neurological information on living species, such as brain size (both absolute and relative, i.e., related to body size), the actual makeup of the brain from the gross to microscopic levels, including neural nuclei, fiber systems and interconnections, and distribution of neurotransmitters and neuroreceptors. Additionally, the brain can be studied ontogenetically, and neuroscientists can actually study the relationships between how the brain varies neurologically and how these variations relate to the behavioral variation. Neurogenomic information will also add considerable details as to how living brains vary and operate, both within and between different species. This richness is simply lost to the paleoneurologist as it is not available as direct evidence. However, it is necessary to realize that the extant living species – for example, bonobo, chimpanzee, macaque – are end points of their own evolutionary lines of development and are not our ancestors, however closely related (*see* Primates, Evolution of). It is the blending and complementation of these two approaches that provides the best evidence for when and how our brains evolved.

Characteristics of the Human Brain

Brain Size, Absolute and Relative

The human animal is obviously obsessed with size, and those who study the brain comparatively, perhaps more so. With an average brain weight of 1330 g (Tobias, 1971), the human species has the largest absolute brain size within the primate order, but is actually dwarfed by elephants and some of the whales, where brain weight can exceed 7500 g. Of course, the body weights are very much higher in elephants and whales. But even for its body weight, *Homo sapiens* does not have the largest relative brain weight, being outdone by several monkeys, some rodents, and even some fish. Normal modern human brain size varies between roughly 900 and 2000 g, although some very small number of exceptions does occur, with sizes ranging from about 750–900 and 2000–2200 g. There is human population variation as well as differences between the sexes. In general, people who live in the Arctic have larger brains than those who live in the tropics; the small-statured Ituri forest pygmies in Africa appear to have some of the smallest brains in the world. Males in all populations for which good autopsy data have been gathered show brain sizes on the average of 100–150 g greater than females, an amount roughly the same as the range of modern human racial variation. It should be pointed out that these differences, and their possible relationship to cognitive skills, are highly controversial and simple correlations are deceptive (Holloway, 1996, 2008; Nyborg, 2003). Table 1 provides a listing of the major fossil hominid taxa and their respective brain sizes. Notice that the range of values from the earliest australopithecine to modern *Homo* is roughly 1000 ml, or about the same amount as the normal range of variation within our species.

Encephalization (Encephalization Coefficient, EQ)

Nevertheless, the human animal does come out on top of the evolutionary heap when its absolute brain and body weights are considered together. When the log (base 10) of brain weight is plotted against the log10 of body weight for a group of relevant taxa, the result is a linear line, where (log10) brain weight = a + b (log10) body weight. For a large array of primate data (e.g., Stephan et al., 1981), the slope of the line (b in the equation above) is about .76, and the correlation coefficient is .98, indicating that the relationship is almost perfect. This relationship will naturally vary depending on the databases and the transformations used. This is known as an allometric equation, and these are used frequently in biology to assess the underlying relationships between the size of parts of the body and the whole. The slope sometimes has an interpretation suggesting functional relationships between the brain and other variables. For example, in the above example, the slope is .76, extremely close to .75 or 3/4, which often describes a metabolic relationship (Martin, 1983). The slope of .666, or 2/3, has been championed by some (e.g., Jerison, 1973) as indicating an important geometric relationship

Table 1 Some fossil hominid brain volumes

Group	Number	Location	Average Brain volume	Range	Dating (My)
A. afarensis	3	E. Africa	435	385–500+	3–4
A. africanus	8	S. Africa	440	420–500+	2–3
A. aethiopicus	1	E. Africa	410	na	2.5
A. garhi	1	E. Africa	ca. 450	na	2.5
A. sediba	1	S. Africa	ca. 420	na	2–3
A. robustus	6	E. & S. Africa	512	500–530	1.6–2.0
H. rudolfensis	2	E. Africa	775	752–800	1.8
H. habilis	6	E. Africa	612	510–687	1.7–2.0
H. georgicus	3	Georgia, Europe	677	600–775	1.7
H. ergaster	2	E. Africa	826	804–848	1.6
H. erectus	2	E. Africa	980	900–1067	1.0–1.6
H. erectus	8	Indonesia	925	780–1059	1.0
H. erectus	8	China	1029	850–1225	.6
Archaic *H. sapiens*	6	Indonesia (Solo)	1148	1013–1250	.13
Archaic *H. sapiens*	6	Africa	1190	880–1367	.125
Archaic *H. sapiens*	7	Europe	1315	1200–1450	.5–.25
H. sapiens (Neand.)	25	Europe, M. East	1415	1125–1740	.09–.03
H. sapiens sapiens	11	World	1506	1250–1600	.025–.01

Source: Holloway, R.L., 1997. Brain evolution. In: Dulbecco, R. (Ed.), Encyclopedia of Human Biology, vol. 2. Academic Press, New York, pp. 189–200.

between volume and surface area. It is important to realize that these slopes vary depending on the taxa examined. In general, as the taxa become more similar, the slope decreases. Species within a genus generally have a slope around .3; within a species the slope is smaller yet, being about .2, and the correlation coefficient is also reduced (see also Martin and Isler, 2010).

Just as the human animal is curious, it is also vainglorious, always trying to find a measure that places it at the top. Thus we can fabricate a device, the *Encephalization Coefficient*, or EQ, which shows that relative to any database, the human animal is the most encephalized animal living. The point for *H. sapien*s shows a clear positive residual above the expected regression line, and in fact the human value is about three times that expected for a primate with its body weight. Table 2 provides a number of different equations based on differing databases, which happily give *H. sapiens* the highest value. (Actually, young immature dolphins will provide a higher number, but when compared to an immature human, the value is higher in the latter.) Two additional points should be made: (1) EQs are relative to the databases used, and thus there is an inherent 'relativity' to relative brain sizes; and (2) EQs do not evolve, only brain weight/body weight relationships do, and EQs are simply a heuristic device enabling comparisons between taxa; they have no reality outside of the database chosen, or species within a taxa, and are not designed to discuss within-species variation. For example, female humans are 'more' encephalized than males, given their smaller body sizes and smaller brains but the relationship might be simply a statistical artifact with no known behavioral manifestation given their nearly equal overall intelligence.

We will discuss later how the processes of hypertrophy and hyperplasia have been positively selected for in the course of the last 2 My of hominid evolution. (Hypertrophy refers to increases in size of the neural components, e.g., neurons, dendritic branching, nuclei, fiber tracts; hyperplasia refers to

Table 2 Some examples of encephalization quotients

Species	Brain wt. (g)	Body wt. (g)	EQ Homo[a]	EQ Jerison[b]	EQ Primates[c]	EQ Stephan[d]
Lemur	23.3	1400	21	1.56 (22.6)	.94 (32.7)	5.66 (19.6)
Baboon	201	25 000	28	1.97 (28.5)	.90 (31.3)	7.94 (27.5)
Gorilla	465	165 000	23	1.56 (22.5)	.61 (21.2)	6.67 (23.2)
Orang	370	55 000	31	2.15 (31.1)	.91 (31.7)	8.90 (30.9)
Chimp	420	46 000	39	2.63 (28.1)	1.81 (41.1)	11.3 (39.3)
Human	1330	65 000	100	6.91 (100)	2.87 (100)	28.8 (100)

Note: Each formula is based on a different set of data. The EQ Homo equation simply uses the average brain and body weight for *Homo sapiens*, and assumes an intercept where both brain and body weights are zero. The value of whichever animal is calculated is then given as a direct % of modern *Homo sapiens*. EQ Jerison is based on data for almost 200 mammals, while the EQ Primates is based on Martin, R.D., 1983. Human Evolution in an Ecological Context. American Museum of Natural History James Arthur lecture (1982), New York, data set for primates only. The EQ Stephan equation is based on insectivores only. The numbers in the parentheses are the % of the *Homo sapiens* value.
[a]Formulae: EQ Homo = Brain wt/1.0 Body wt$^{.64906}$.
[b]EQ Jerison = Brain wt/.12 Body wt$^{.66}$.
[c]EQ Primates = Brain wt/.0991 Body wt$^{.76237}$.
[d]EQ Stephan = Brain wt/.0429 Body wt$^{.63}$.
Source: Holloway, R.L., 1997. Brain evolution. In: Dulbecco, R. (Ed.), Encyclopedia of Human Biology, vol. 2. Academic Press, New York, pp. 189–200.

increased production of cells through mitotic division.) It is most probably the case that these processes are controlled by regulatory genes, and one of the major differences between ourselves and our closest nonhuman primate relative, the chimpanzee (brain size = ca. 385 g), relates to the schedules by which hyperplasia and hypertrophy are turned on and off during ontogenetic development (Holloway, 1979, 1995; Miller et al., 2012).

Brain Organization and Reorganization

It is well known that the brains of most animals are extremely similar to one another in terms of their overall organization, by which are meant neural nuclei and fiber systems. The human animal does not appear to show any different structures when compared to Old World monkeys such as the macaque, or the Great Apes, including the bonobo, chimpanzee, gorilla, and orangutan. Even the neural fiber tracts that are involved in human language appear in these primates (Deacon, 1997). One might ask, then, given the obvious species-specific repertoires that exist in all animals, how can these behaviors differ without differences in the underlying nervous systems? This is one of the major challenges of studying brain evolution, and in particular understanding what neural organizations account for the specificity of, say, human behavior, the ability to use language composed of arbitrary symbols. In other words, all mammals have cerebral cortex, thalamus, cerebellum, hypothalamus, and so forth, and basically these structures possess almost identical divisions of nuclei and do the same neural tasks. Clearly, brain size alone will never explain species-specific behavior, and the relationships between neural nuclei and fiber tracts will only go so far in explaining behavioral differences.

Allometric equations showing the relationship between one bodily component and the whole are instructive here. If we were to plot the logs (base 10) of primary visual cortex (PVC) against brain volume, we would find that the human PVC is 121% less than predicted, and similarly, the lateral geniculate nucleus of the thalamus is about 144% less than expected for a primate of our brain size. In contrast, if one were to plot the amount of cerebral cortex against brain weight, the result is a straight line, and the human point lies almost exactly on the line. In short, the human cerebral cortex is as large as would be expected for a primate of its brain size. But do portions of the cerebral cortex vary in size between different primates? In humans, the residuals mentioned previously suggest that compared to chimpanzees, the amount of PVC is significantly smaller in humans, or alternatively put, the posterior association cortex of the parietal and temporal lobes is relatively larger in humans. Since there are no essential differences between chimpanzees and humans in their visual abilities and competencies, these differences reflect selection for expanded functioning of the association cortex in humans. This is precisely what is meant by 'reorganization.'

When used in a comparative or evolutionary context, reorganization means changes in the sizes and proportions thereof of neural nuclei and their fiber tracts. Given that chimpanzees and hominids last had a common ancestor some 5–7 My, and that chimpanzees appear to have large PVC cortices, we infer that one aspect of human brain evolution has been some reorganization of the cerebral cortex, namely an increase in posterior association cortex (or, equally, a reduction in PVC) involved in polymodal cognitive tasks, where visual, auditory, and motor information are brought together in a synthetic whole. The trick, of course, is to demonstrate objectively when, where, and why these changes took place. This example of PVC has been purposefully chosen because one of the sulcal landmarks of the cortex that defines the anterior border of PVC is the 'lunate' sulcus, named for its crescentic shape, and there is some hope of identifying its position on some of the early hominid brain endocasts. In this regard, endocasts are most often frustratingly mute on other convolutional details.

Neuroanatomists have been trying for many decades to demonstrate the major differences between humans and other primates, and aside from gross brain size, very little has been shown. The frontal lobe, particularly its prefrontal portion, has been a favorite target, and, indeed, Brodmann (1909) claimed it was proportionally larger in humans, a view most recently championed by Deacon (1997). Unfortunately, other works have shown that the human brain has just as much frontal lobe as would be expected for a primate of its brain weight (Semendeferi et al., 1997; Uylings and van Eden, 1990), although the picture regarding prefrontal cortex has yet to be determined objectively using cytoarchitectonic criteria, which is how prefrontal cortex is differentiated from the pure motor cortex behind it (Schenker et al., 2010; Sherwood et al., 2003; Rilling et al., 2008). Hominid brain endocasts do not, alas, provide any sulcal landmarks with enough reliability to determine the boundaries of prefrontal cortex, which is so important to impulse control, and higher cognitive functions such as planning and abstraction. Thus these regions cannot be accurately measured in a phylogenetic sequence. The Neanderthals, living from about 300 000 to about 24 000 years ago, have frequently been described as having smaller frontal lobes; this is not based on objective measurements, but rather a perception that the large brow ridges on these humans were constraining frontal lobe development. Studying the Neanderthal brain endocasts and comparing them to modern humans, I have failed to see any significant difference between these two groups (Bookstein et al., 1999).

Similarly, regions such as 'Broca's and Wernicke's areas,' anterior and posterior association cortical regions involved in motor (Broca's) and receptive (Wernicke's) aspects of speech, are determinable on most fossil endocasts, and we can determine, for example, that Broca's region is more human-like on the brain cast of early *Homo*, some 1.8 My. This is the famous KNM-ER 1470 endocast of *Homo rudolphensis* from Kenya, which had a brain volume of 752 ml, but which may not be a direct ancestor to our own line of *Homo*. We know that Broca's regions in modern *Homo* are asymmetrical both in overall size and cytoarchitectonic divisions between areas 44, 45, and 47 of Brodmann (Schenker et al., 2010). Interestingly, Neanderthal endocasts show similar asymmetry in Broca's regions.

While the concept of reorganization has a heuristic value in directing our attention to changing quantitative relationships between different neural nuclei and fiber tracts, we cannot yet describe behavioral differences between closely related

animals such as the chimpanzee, gorilla, and orangutan, or different species of the genus *Macaca*, or indeed different breeds of dogs with their different temperaments, aptitudes, and sociality. We simply do not know what magic level of neural description is necessary to describe species-specific behavior. More recent research on prairie and mountain voles suggests that the difference in the females' ability to retrieve pups back to the nest depends on the distribution and number of neuroreceptors for the hormone oxytocin found in several nuclei of the brain, particularly the thalamus. Otherwise, their brains appear identical (Insel and Shapiro, 1992). In addition, it is necessary to remember that the brain possesses aspects of plasticity that we did not appreciate except within the past decade, and that as the brain's organization unfolds ontogenetically, interactions with environmental stimuli are always occurring, and the brain builds its organization partly through its plasticity. It is difficult enough to study and understand such patterns in laboratory animals, let alone in our fossil ancestors! While what has just been discussed suggests a somewhat pessimistic tone, we should remember that advances in noninvasive technology such as MRI, fMRI, PET, and tensor diffusion scanning have enormously increased our understanding of how the brain works, and how neural systems integrate and dissect data from the environment, always providing us with newer paradigms for further exploration about our brains and behavior, and in time, those of our closest relatives, the apes, in particular the bonobo and chimpanzee (see in particular Semendeferi et al., 2010).

Human Brain Asymmetry

The cerebral cortices of the human brain are usually asymmetrical, and tend to grow in a torque manner, reflecting minor differences in maturation rates. The hemispheres are seldom, if ever, equipotential in terms of functioning. Our left hemisphere is often characterized as 'analytic' and involved with language tasks, while our right hemisphere appears most competent in visuospatial integration, and is often thought of as the 'intuitive' or 'gestalt' hemisphere. These characterizations, while crude, hold up fairly accurately for right-handers and many ambidextrals. From radiographic studies, it was possible for LeMay (1976) to ascertain different petalia patterns for right- and left-handed humans with a high degree of precision. These petalias are small extensions of cerebral cortex that extend farther in one part of a hemisphere than on the other side. For example, we speak of a left-occipital right-frontal torque pattern of petalias as occurring with high frequency in right-handed individuals. This means that the left occipital lobe bulges somewhat more posteriorly on the left hemisphere, while the right hemisphere is somewhat broader in width in the frontal lobe. In true left-handers, who make up about 8–10% of human populations, the pattern is reversed, meaning they exhibit a right-occipital left-frontal pattern. Petalia patterns for a large collection of apes indicated that while chimpanzees, gorillas, and orangutans sometimes demonstrated asymmetries, they did not show the particular torque pattern described before as frequently. The gorilla, incidentally, was the most asymmetrical of the apes (Holloway and de LaCoste-Lareymondie, 1982). On the other hand, brain asymmetries, particularly in the *planum temporale* (temporal cortex) of the chimpanzee, show a strong left-hemispheric size difference compared to the right (Gannon et al., 1998). This is simply puzzling as we do not have any evidence that chimpanzees use this structure in communication as do humans, and the fact that we share this difference with chimpanzees suggests that brain organizational features relating to complex cognitive functioning has been around for at least 5–7 My. As our noninvasive scanning techniques become more sophisticated, we can expect to learn how these asymmetries function in animals other than ourselves. In fact, asymmetries appear in many animals and are hardly unique to primates. It is probably the degree of asymmetry that might be important in attempting to distinguish humans from other primates (Balzeau and Gilissen, 2010).

Hominid brain endocasts, when complete for both sides (unfortunately, this is very rare) allow the paleoneurologist to assess the cerebral asymmetries, and indeed, even australopithecines appear to show beginnings of the right-handed torque pattern found in humans, and, as one progresses through time, the petalia patterns become more accentuated in the modern human direction. If we add to these observations those of Toth's (1985) studies on the early stone tools of about 2 My, which strongly suggest right-handedness, this underlines the fact that our early ancestors' brains, despite their small size (sometimes within the extant apes' range), were reorganized, and that they probably had some modes of cognition very similar to our own.

Synthesis: Putting Together Size, Organization, and Asymmetry during Human Evolution

As mentioned earlier, human brain evolution has clearly been a process of integrating neurogenomic processes that led to increased size of the brain (hyperplasia and hypertrophy), and these neurogenomic changes also played roles in the reorganization (quantitative shifts) of neural nuclei, fiber tracts, and cortical cytoarchitectonics. In addition, it is probable that other changes occurred at the neurochemical level, involving neurotransmitters and receptor sites, but these are not well known from the comparative record, let alone the fossil one. This integration was sometimes gradual, sometimes 'punctuated,' at least based on the fossil hominid record currently available. The only reliable evidence from paleoneurology suggests that Brodmann's area 17 (PVC) was reduced early in hominid evolution, signs of the reduction being clear in *A. afarensis* some 3–3.5 Ma. While this would have meant a relative increase in posterior parietal cortex (area 39) and peri- and parastriate cortex (areas 18 and 19 respectively), the faithfulness of sulcal impressions does not allow for unambiguous definition of these areas. Similarly, it is not possible at this time to measure and delineate remaining areas of the temporal cortex and superior parietal lobule unambiguously. What is suggested, however, is that visuospatial abilities were most probably cognitively enhanced early in hominid evolution. It is not until we come to *Homo rudolfensis c.* 1.8 Ma that a case can be made for some frontal lobe reorganization in the third inferior frontal convolution, Broca's area. Thus it would appear there was a gradient of cerebral reorganizational changes starting posteriorly, and progressing anteriorly. Table 3 outlines these changes.

Table 3 Summary of reorganizational changes in the evolution of the human brain

Brain changes, reorganizational	*Taxon*
1. Reduction of primary visual striate cortex, area 17, and a relative increase in posterior parietal and temporal cortex, Brodmann areas 37, 39, 40, as well as 5 and 7	*Australopithecus afarensis, Australopithecus africanus*
2. Reorganization of frontal lobe (3rd inferior frontal convolution, Broca's areas 44, 45, 47)	*Homo rudolfensis;* early *Homo*
3. Cerebral asymmetries, left-occipital right frontal petalias	Australopithecines and early *Homo*
4. Refinements in cortical organization to a modern *Homo sapiens* pattern	*Homo erectus* to present

Note: (4) is inferred, as brain endocasts cannot provide that level of detail necessary to demonstrate the refinements in cortical organization from surface features alone. Areas 18, 19 are peri- and parastriate cortex just anterior to area 17, and are included in posterior association cortex here.
Source: Holloway, R.L., 1997. Brain evolution. In: Dulbecco, R. (Ed.), Encyclopedia of Human Biology, vol. 2. Academic Press, New York, pp. 189–200.

Table 4 outlines the major size changes in the human brain during its evolutionary odyssey. Paleoneurological data simply are not detailed enough to integrate the two tables of size and reorganizational changes into one holistic sequence of events. Basically, the paleontological record supports an early reorganizational change resulting in an increase in posterior cortex associated with visuospatial processing, perhaps accompanied by a relative small allometric increase in brain size from *A. afarensis* to *Australopithecus africanus*. This would correlate well with geological and paleontological evidence that shows that early hominids were expanding their ecological niches and becoming more diverse in their subsistence patterns in mixed habitats. We know this based on the fact that stone tool types are becoming standardized in form, tool inventories grow larger, and right-handedness is highly probable. With the advent of *Homo*, we find strong evidence for a major increase in brain size, both allometric (related to body size) and nonallometric, and a reorganized frontal lobe, broader, and showing a more modern human-like Broca's area. This suggests that indeed there had been some strong and dramatic selection pressures for a somewhat different style of sociality, one most probably based on a primitive proto-language that had some arbitrary symboling elements as the standardization of stone tools (e.g., Acheulean hand axes) increases, suggesting social cohesion and control mediated through symbolically based communication (Holloway, 1981). Needless to say, this is but one speculative account of the evidence. But from about 1.8 to roughly .5 Ma, we think there were minor allometric brain size increases to the earliest *Homo erectus* hominids of Indonesia and China, where brain sizes ranged from 750 to 1250 ml in volume. We have very little evidence for body sizes, but we believe, on the basis of the KNM-WT of 15 000 Nariokotome youth from Kenya at *c.* 1.6 Ma, that these did not differ significantly from our own.

This is also a time during which cerebral asymmetries are becoming more strongly pronounced. With the advent of Archaic *H. sapiens*, about .15–.2 Ma, we find brain sizes well within modern human values, and no evidence for further allometric increases, except possibly for the Neanderthal humans in which it can be argued that their larger brain and body sizes (lean body mass: bone and muscle) were adaptations to colder conditions (*see* Body, Evolution of). If further changes took place in cerebral and/or subcortical organization, they are simply not apparent from a paleoneurological perspective. Yet, the Upper Paleolithic is the time when cave art makes its appearance, and one cannot help but wonder whether the explicit use of art involving symbolization might not also have been the time for the emergence of language. In fact, however, there is nothing in the direct fossil evidence, and, in particular paleoneurology, that provides any evidence for such views. Claims for a single mutation are extremely speculative, and while some genes have been identified, such as FOXP2 (also found in Neanderthals), these also involve more general aspects of cognition. It is more likely that stone tool making and its underlying cognitive elements are very similar to

Table 4 Brain size changes in hominid evolution

Brain changes	*Taxon*	*Time (My)*	*Evidence*
1. Small increase, allometric[a]	*A. afarensis* to *A. africanus*	3.5–2.5	Brain endocast increase from ca 400 to 450+ ml.
2. Major increase, rapid, both allometric and non-allometric	*A. africanus* to *H. habilis, rudolfensis*	2.5–1.8	KNM–1470, 752 ml (300 ml increase)
3. Modest allometric increase in brain size to 800–1000 ml	*H. habilis* to *H. erectus*	1.8–.5	*H. erectus* brain endocasts and postcranial bones
4. Gradual and modest size increase to archaic non-allometric	*H. erectus* to *H. sapiens neanderthalensis*	.5–.075	Archaic *Homo H. sapiens*, Neandertal endocasts 1200–1700+ ml.
5. Small reduction in brain size among modern allometric	*H. sapiens* to *H. sapiens sapiens*	.015–present	Modern endocranial volumes

[a]Related to increase in body size only.
Source: Holloway, R.L., 1997. Brain evolution. In: Dulbecco, R. (Ed.), Encyclopedia of Human Biology, vol. 2. Academic Press, New York, pp. 189–200, and more recent endocast data, Holloway, R.L., Broadfield, D.C., Yuan, M.S., 2004. The human fossil record. In: Schwartz, JH, Tattersall, I. (Eds), Brain Endocasts: The Paleoneurological Evidence, vol. 3. Wiley-Liss, New York.

language, if not partially homologous (Holloway, 1969, 2012). Finally, it would appear that there has actually been a small reduction in brain size, probably allometric in nature, from about .015 Ma to the present (Henneberg, 1988; Hawks, 2012).

The totality of evidence shows that the brain has always been evolving during our evolutionary journey, with myriad changes taking place at different tempos during different times. As suggested recently (Holloway, 1997: p. 200):

> In sum, the major underlying selectional pressures for the evolution of the human brain were mostly social. It was an extraordinary evolutionary 'decision' to go with an animal that would take longer to mature, reach sexual maturity later, and be dependent for its food and safety upon its caretakers (parents?) for a longer period of time. The benefits for the animal were many, including a longer learning period, a more advanced, larger, and longer-growing brain, and an increasing dependence on social cohesion and tool making and tool using to cope with the environments that they encountered. Needless to say, language abilities using arbitrary symbol systems were an important ingredient in this evolution.

> The fossil record shows us that there was a feedback between the complexity of stone tools (which must be seen as a part of social behavior) and increasing brain size and the expansion of ecological niches. The 'initial kick', however, the process that got the ball rolling, was a neuroendocrinological change affecting regulatory genes and target tissue–hormonal interactions that caused delayed maturation of the brain and a longer growing period, during which learning became one of our most important adaptations.

These ideas have been detailed elsewhere (Holloway, 1967, 1969, 1979,1996, 2010), where more details may be found.

And to the Future?

There appear to be two common presumptions about our future brain evolution. One is that our biological evolution has stopped. The second is that our brains will continue to grow in size, with bulging frontal lobes (sort of a cross between E.T. and X-Files etc.), to handle our growing dependence on technology. What we have witnessed from the past fossil record is that our brains and bodies work largely in allometric fashion, and given the high metabolic cost of operating bigger brains (about 20–25% of our metabolic resources go to supporting our brains, which constitute only 2% of our total body weight), the second scenario seems highly unlikely. The first scenario is simply untrue, but it would require vast amounts of information from each generation of many living populations, something feasible perhaps, but not currently being collected. Furthermore, it is quite controversial whether brain size has any close relationship to intelligence, however intelligence is actually defined and measured. Recent research based on MRI determinations of brain volume and selected batteries of cognitive tests have shown correlations between test scores and brain volume ranging from .4 to .6. (Andreasen et al., 1993; Anderson, 2003; Davies et al., 2011). But if protein resources were to nosedive throughout the world for a significant period of time, selection would probably favor smaller body sizes in our species, and that could result in smaller brains, given an allometric relationship of roughly .3 between stature and brain size, at least in males (Holloway, 1980). While genetic engineering may well provide some respite between the ever-increasing mass of humanity and ecological and nutritive degradation, this too is likely to be nothing more than short-term fending off of the unstoppable future. These degradations are part and parcel of the human brain's capacity to ignore warnings that should curtail greed and stupidity. The paleontological record for most mammals suggests that at the taxonomic level of the genus (such as *Pan*, *Homo*, *Canis*, *Notocherus*, etc.), one finds a recognizable record of that genus spanning approximately 5–10 My. Our genus has thus far a duration of about 2 My. We, as a genus, despite our largish highly encephalized brains, have another 3 My to go if we wish to be as successful in the paleontological longevity game.

See also: Body, Evolution of; Cognition, Evolution of; Primates, Evolution of.

Bibliography

Allen, J.S., Bruss, J., Damasio, H., 2006. Looking for the lunate sulcus: a magnetic resonance imaging study in modern humans. Anatomical Record A 288, 867–876.

Anderson, B., 2003. Brain imaging and g. In: Nyborg, H. (Ed.), The Scientific Study of General Intelligence: Tribute to Arthur R. Jensen. Pergamon Press, NY, pp. 29–40.

Andreasen, N.C., Flaum, M., Swayze, H.V., et al., 1993. Intelligence and brain structure in normal individuals. American Journal of Psychiatry 150, 130–134.

Balzeau, A., Gilissen, E., 2010. Endocranial shape asymmetries in *Pan paniscus*, *Pan troglodytes* and *Gorilla gorilla* assessed via skull based landmark analysis. Journal of Human Evolution 59, 54–69.

Bookstein, F., et al., 1999. Comparing frontal cranial profiles in archaic and modern *Homo* by morphometric analysis. Anatomical Record 6, 217–224.

Brodmann, K., 1909. Vergleichende Lokalizationzlehre der Grosshirnrinde. J.A. Barth, Leipzig, Germany.

Dart, R., 1925. *Australopithecus africanus*: the man-ape of South Africa. Nature 115, 195–199.

Davies, G., et al., 2011. Genome-wide association studies establish that human intelligence is highly heritable and polygenic. Molecular Psychiatry 16, 996–1005.

Deacon, T., 1997. The Symbolic Species: The Co-evolution of Language and the Brain. Norton, New York.

Gannon, P.J., Holloway, R.L., Broadfield, D.C., Braun, A.R., 1998. Asymmetry of chimpanzee planum temporale: humanlike pattern of Wernicke's brain language area homolog. Science 279, 220–222.

Hawks, John, 2012. Selection for Smaller Brains in Holocene Human Evolution arXiv:1102.5604.

Henneberg, M., 1988. Decrease of human skull size in the Holocene. Human Biology 60, 395–405.

Holloway, R.L., 1967. The evolution of the human brain: some notes toward a synthesis between neural structure and the evolution of complex behavior. General Systems XII, 3–19.

Holloway, R.L., 1969. Culture: a human domain. Current Anthropology 10, 395–412.

Holloway, R.L., 1979. Brain size, allometry, and reorganization: toward a synthesis. In: Hahn, M.E., Jensen, C., Dudek, B.C. (Eds.), Development and Evolution of Brain Size: Behavioral Implications. Academic Press, NY, pp. 59–88.

Holloway, R.L., 1980. Within –species brain-body weight variability: a re-examination of the Danish data and other primate species. American Journal of Physical Anthropology 53, 109–121.

Holloway, R.L., 1981. Culture, symbols, and human brain evolution. Dialectical Anthropology 5, 287–303.

Holloway, R.L., 1984. The Taung endocast and the lunate sulcus: a rejection of the hypothesis of its anterior position. American Journal of Physical Anthropology 64, 285–287.

Holloway, R.L., 1995. Toward a synthetic theory of human brain evolution. In: Changeux, J.P., Chavaillon, J. (Eds.), Origins of the Human Brain. Clarendon Press, Oxford, UK, pp. 42–54.

Holloway, R.L., 1996. Evolution of the human brain. In: Lock, A., Peters, C. (Eds.), Handbook of Human Symbolic Evolution. Oxford University Press, New York, pp. 74–116 (Chap. 4).

Holloway, R.L., 1997. Brain evolution. In: Dulbecco, R. (Ed.). Encyclopedia of Human Biology, vol. 2. Academic Press, New York, pp. 189–200.

Holloway, R.L., 2000. Brain. In: Delson, E., Tattersall, I., Van Couvering, J., Brooks, A.S. (Eds.), Encyclopedia of Human Evolution and Prehistory, second ed. Garland Publishing, New York, pp. 141–149.

Holloway, R.L., 2008. The human brain evolving: a personal retrospective. Annual Review of Anthropology 37, 1–19.

Holloway, R.L., 2010. Human brain endocasts, Taung, and the LB1 Hobbit brain. Chap. 4. In: Broadfield, D., et al. (Eds.), The Human Brain Evolving: Paleoneurological Studies in Honor of Ralph L. Holloway. Stone Age Institute Press, Gosport, Indiana.

Holloway, R.L., 2012. Language and tool making are similar cognitive processes. Behavioral and Brain Sciences 35 (4), 326.

Holloway, R.L., Broadfield, D.C., Yuan, M.S., 2004. The human fossil record. In: Schwartz, J.H., Tattersall, I. (Eds.), Brain Endocasts: The Paleoneurological Evidence, vol. 3. Wiley-Liss, New York.

Holloway, R.L., de LaCoste-Lareymondie, M.C., 1982. Brain endocast asymmetry in pongids and hominids: some preliminary findings on the paleontology of cerebral dominance. American Journal of Physical Anthropology 58, 101–110.

Insel, T., Shapiro, L.E., 1992. Oxytocin receptors and maternal behavior. Annals of the New York Academy of Sciences 652, 448–451.

Jerison, H.J., 1973. Evolution of Brain and Intelligence. Academic Press, New York.

LeMay, M., 1976. Morphological cerebral asymmetries of modern man, fossil man, and nonhuman primates. Annals of the New York Academy of Sciences 280, 349–366.

Martin, R.D., 1983. Human Evolution in an Ecological Context. American Museum of Natural History, New York. James Arthur lecture (1982).

Martin, R.D., Isler, K., 2010. The maternal energy hypothesis of brain evolution: an update. In: Broadfield, et al. (Eds.), The Human Brain Evolving: Paleoneurological Studies in Honor of Ralph L. Holloway. Stone Age Institute Press, Gosport, Indiana.

Miller, D.J., et al., 2012. Prolonged myelination in human neocortical evolution. Proceedings National Academy of Sciences 109 (41), 16480–16485.

Nyborg, H. (Ed.), 2003. The Scientific Study of General Intelligence: Tribute to Arthur R. Jensen.

Rilling, J.K., et al., 2008. The evolution of the arcuate fasciculus revealed with comparative DTI. Nature Neuroscience 11, 382–384.

Schenker, N.M., et al., 2010. Broca's area homologue in chimpanzees (Pan troglodytes): probabilistic mapping, asymmetry, and comparison to humans. Cerebral Cortex 20, 730–742.

Semendeferi, K., Damasio, H., Frank, R., Van Hoesen, G.W., 1997. The evolution of the frontal lobes: a volumetric analysis based on three-dimensional reconstructions of magnetic resonance scans of human and ape brains. Journal of Human Evolution 32, 375–388.

Semendeferi, K., et al., 2010. Brain reorganization in humans and apes. In: Broadfield, D.C., et al. (Eds.), The Human Brain Evolving: Paleoneurological Studies in Honor of Ralph L. Holloway. Stone Age Institute Press, Gosport, Indiana.

Sherwood, C.C., et al., 2003. Variability in Broca's area homologue in African great apes: implications for language evolution. Anatomical Record 271A, 276–285.

Stephan, H., Frahm, H., Baron, G., 1981. New and revised data on volumes of brain structures in insectivores and primates. Folia Primatologia 35, 1–29.

Stout, D., 2006. Oldowan tool making and hominin brain evolution: theory and research using positron emission tomography (PET). In: Toth, N., Schick, K. (Eds.), The Oldowan: Case Studies into the Earliest Stone Age. Stone Age Institute Press, Gosport, Indiana.

Tobias, P.V., 1971. The Brain in Hominid Evolution. Columbia University Press, New York.

Toth, N., 1985. Archaeological evidence for preferential right-handedness in lower and middle Pleistocene, and its behavioral implications. Journal of Human Evolution 14, 607–614.

Uylings, H.B.M., van Eden, C.G., 1990. Qualitative and quantitative comparison of the prefrontal cortex in rats and primates, including humans. Progress in Brain Research 85, 31–62.

Weber, G.W., et al., 2012. Digital South African fossils: morphological studies using reference-based reconstruction and electronic preparation. In: Reynolds, S.C., Gallagher, A. (Eds.), African Genesis: Perspectives on Hominin Evolution. Cambridge, New York, pp. 298–316.

Brain: Response to Enrichment

Janice M Juraska and Leslie M Wise, University of Illinois, Champaign, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M.C. Diamond, volume 2, pp. 1352–1358, © 2001, Elsevier Ltd.

Abstract

Enriched environments with opportunities for social and object interactions improve performance on complex behavioral tasks and the structure of the brain in rodents when compared with individuals housed alone or in small social groups. Although the largest changes can be seen in young animals, enrichment can induce plasticity at any point in the lifespan. Recent studies indicate that enrichment ameliorates neural problems and that the effects can be extended to primates, including humans.

In his now classic book *Organization of Behavior*, Donald Hebb described taking young lab rats home and letting his children raise them as pets. When these pet rats were later brought back to the laboratory for extensive training in the Hebb–Williams maze, they clearly outperformed cage-reared rats (Hebb, 1949). This charming anecdote (now it would be illegal) was meant to demonstrate the environmental effects on intelligence, which was one of the major points of Hebb's book. This led to a more systematic examination of environmental effects that can be accomplished in the laboratory through what is frequently known as an enriched environment or condition (EC). Although this environment can vary in different labs, there are common elements: group housing (rats and mice are social animals) in a larger than standard cage and the addition of objects ('toys') that are changed on a regular basis (**Figure 1**). These rats are then compared with either rats housed alone in standard cages (IC for individual, isolated or impoverished condition) or to socially housed (SC for social condition) rats in groups that usually contain fewer animals than in EC.

One could argue that the enriched environments are not really 'enriched' in comparison to the myriad environments that wild rats encounter. In general this is true, but the environment is being compared with laboratory conditions. However, this objection has led to the use of the term 'environmental complexity' for EC by some laboratories, including my own.

Figure 1 Rats exploring an enriched environment. In the Juraska laboratory 'toys' are changed daily. Note that the photograph was taken before wood was disallowed as part of the environment. The rats were healthy nonetheless.

The literature concerning use of enriched environments is large, and different uses have been found for the paradigm from demonstrating the malleability of the brain to exploring the therapeutic benefits of environmental stimulation. The current article cannot list all of these studies but is meant to trace the history of enrichment effects and cover the issues surrounding the paradigm to the present.

Behavior

Hebb's idea was followed up by many investigators who found that rats raised in stimulating conditions outperform rats in more standard environments in many types of mazes (Forgays and Forgays, 1952; Brown, 1968; Greenough et al., 1972; Juraska et al., 1984; Seymoure et al., 1996; Pietropaolo et al., 2006), discrimination reversals (Krech et al., 1962) and sequential alternations (Nyman, 1967). It is notable that rats from EC showed better performance on the Hebb–Williams maze even when rats from IC were handled every day (Greenough et al., 1972). This indicates that it is not the habituation to handling, generally part of EC, that is causing the superior performance. Several investigators found evidence that stimulated rats used spatial cues to solve the Hebb–Williams maze because their performance was more disrupted upon rotation of the maze compared with standard caged rats (Forgays and Forgays, 1952; Brown, 1968). This does not mean that stimulated rats inflexibly use distally spatial cues. The 17-arm radial maze involves baiting every arm with food (in the all arms-baited version), and ideally the rat visits each arm once in a daily session. Any arm that is revisited within a session is considered an error of working memory. Prior work from our lab found that rats from EC were more likely to solve the maze by using an adjacent arm strategy (response chaining) compared with rats from IC. This led to perfect performance by the end of training in some EC rats (Juraska et al., 1984; **Figures 2** and **3**). Thus, it appears that EC rats are flexible in the type of cues used depending on the

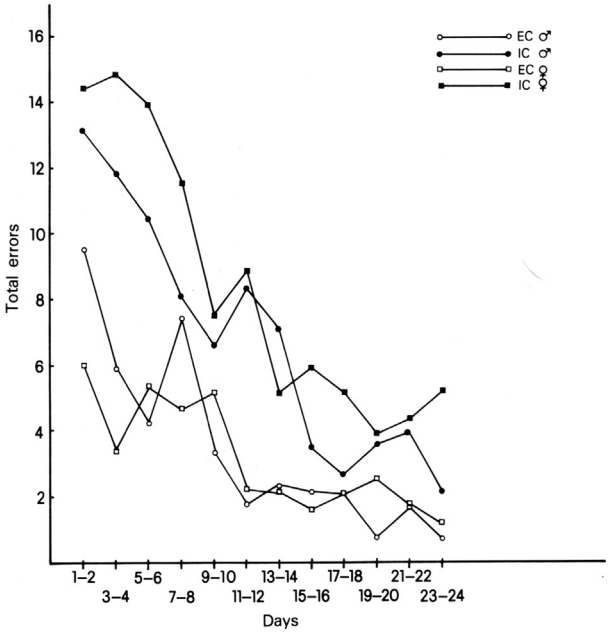

Figure 2 The errors for 2-day blocks of training on the 17-arm radial maze that were averaged for each environmental group and sex. There were significant differences between groups ($p < .001$) and days ($p < .001$) but no sex differences. Reproduced from Juraska, J.M., Henderson, C., Müller, J., 1984. Differential rearing experience, gender, and radial maze performance. Developmental Psychobiology 17, 209–215.

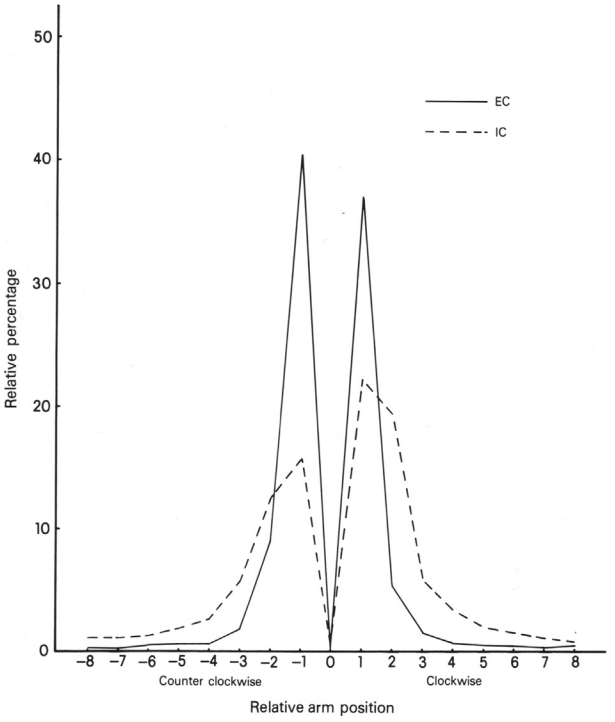

Figure 3 A graph illustrating that rats exposed to EC were more likely to choose an adjacent arm strategy in the 17-arm radial maze than were rats from IC. The probable position of the next arm chosen when point 0 is the given arm is plotted for rats from EC and IC (collapsed across sex). 75% of choices by EC rats were to an adjacent arm while the adjacent arm choice for IC rats was 36%. This led to fewer errors (**Figure 2**) committed by rats from EC. Plotted from data presented in Juraska, J.M., Henderson, C., Müller, J., 1984. Differential rearing experience, gender, and radial maze performance. Developmental Psychobiology 17, 209–215.

contingencies of the maze. More recently, Woodcock and Richardson (2000) found that rats from EC showed more contextual and discriminative ability. In contrast, tasks that are easily learned, such as black/white visual discriminations, often do not result in differential performances probably due to floor effects (Krech et al., 1962; Nyman, 1967). Depending on details of the procedure and the species/strain of rodent, exposure to EC can result in inconsistent results on the performance of the spatial water maze (Pietropaolo et al., 2006; Speisman et al., 2013) unlike the consistent benefits to performance on the more difficult Hebb–Williams and radial arm mazes.

The Brain

Effects on the brain were first noted by a group in Berkeley, headed by Mark Rosenzweig, who found that the weight of the cerebral cortex was greater in rats from EC than IC (Rosenzweig et al., 1962). The largest cortical difference was in the visual cortex which also was thicker in rats from EC than IC (reviewed in Bennett et al., 1964). Although modest (6–7%), this difference in cortical thickness has been replicated many times by the Berkeley group and in other laboratories (e.g., Juraska et al., 1985). It was also found that exposure to EC for as little as 2 h day^{-1} was sufficient to induce these changes (Rosenzweig et al., 1968). It should be noted that these experiments were all performed on male rats and the rare exceptions will be dealt with in a later section.

Controls

The relative ease of using cortical weight or the thickness of the visual cortex as an indicator of effects of enrichment led to a series of experiments by the Berkeley group that examined what aspects of the enriched environment were important for these effects. These are summarized in Bennett et al. (1964) and in several book chapters (e.g., Rosenzweig, 1971; Bennett, 1976). For example, enriched environments often include a running wheel among other objects, but running wheel activity alone did not produce the effects of EC. This was replicated by Zolman and Morimoto (1965). In addition, social experience alone is not sufficient to induce the neural or behavioral effects of EC on most measures. Many experiments contain an SC group and generally find it to be between the EC and IC groups. Social housing was further explored by Rosenzweig et al. (1978) in which the SC rats were in a group comparable in size to EC but did not achieve the EC brain effects. Interaction with objects in the environment appears to be essential. Rats that were in individual cages within the enriched environment were unable to interact with either objects or conspecifics and did not exhibit environmental effects. Likewise rats housed alone, but with objects, did not differ from rats in IC.

Thus, the interaction of both social housing and objects for exploration are necessary for the cerebral effects. The hypothesis that social interaction is arousing and leads to object interactions was further explored by the Berkeley group. They found that if an excitant drug was given to rats just before experiencing EC alone every day for 30 days, they did have a thicker cortex than rats in IC that had been given the same drug every day. It appears that a drug that increases arousal can compensate to some degree for the arousing presence of other rats (reviewed in Rosenzweig, 1971).

Another explanation for the effects of EC is that it is hormonal either through differential growth or stress. If EC causes accelerated growth, all parts of the body should mature faster. However, the increased weight of the brain in rats from EC is accompanied by smaller body weight and bone length (Diamond et al., 1965; Black et al., 1989). These dissociations indicate that the rate of maturation per se is not what is different between the groups. Evidence of differential stress effects have been more mixed with some studies finding no difference in adrenal weights (Bennett et al., 1964; Riege and Morimoto, 1970; Mirmiran et al., 1982; Black et al., 1989), while others find the rats from IC have larger adrenals than rats from EC (Geller et al., 1965; Krech et al., 1966; Uphouse and Brown, 1981). This discrepancy in IC may be due to more severe stress through early weaning, severe isolation and shipping of their pregnant mothers that were part of these studies. IC rats born to dams mated in the laboratory vivarium that are weaned after 21 days and handled periodically do not appear to be more stressed than EC rats.

The Visual Cortex

The largest cortical difference between EC and IC is in the visual cortex as documented many times by the Berkeley group. What cellular differences underlie the weight and thickness changes in the visual cortex? **Table 1** lists the many differences that have been documented in the visual cortex through the 1980s. All of these studies were in male rats exposed to the environments starting at weaning age until they were young adults (day 55–60). This paradigm shows that enrichment increases neural connectivity and the milieu (e.g., astrocytes, myelination, capillaries) that supports it.

Table 1 Cellular measures in the rat visual cortex in which enriched condition (EC) > isolated condition (IC)

Cellular variable	*Reference*
Dendritic branching	Volkmar and Greenough (1972)
Dendritic spine density	Globus et al. (1973)
Number of synapses per neuron	Turner and Greenough (1983)
Size of synaptic contact	West and Greenough (1972)
Synaptic plate perforations	Greenough et al. (1978)
Percentage of total tissue volume	
Capillary vessels	Black et al. (1987)
Astrocytic nuclei	Sirevaag and Greenough (1987)
Oligodendrocytic nuclei	Sirevaag and Greenough (1987)
Mitochondria	Sirevaag and Greenough (1987)
Number of unmyelinated axons in splenial corpus callosum	Juraska and Kopcik (1988)
Size of unmyelinated axons in splenial corpus callosum	Juraska and Kopcik (1988)

The Hippocampus

Initially, work on the effects of EC did not concentrate on the hippocampus because its thickness changes were smaller (2–3%) than in the visual cortex and less reliable. This is partially due to the rolled structure of the hippocampus which does not allow for standardized gross measurements in the same way that the laminar cerebral cortex does. Quantification of dendritic growth also indicated the male EC rats have almost no change in the dendritic tree compared with male IC rats in two major portions of the hippocampus: the dentate gyrus and CA3 (Juraska et al., 1985, 1989), although there are increases in spine density in some parts of the dendritic tree in CA1 (Moser et al., 1997).

Renewed interest in the hippocampus and EC was generated when Kempermann et al. (1997) found that mice from EC had more neurogenesis in the dentate gyrus and consequently, more neurons. Their enriched environment contained a running wheel, and this has led to the question of whether the aerobic exercise that a running wheel can provide is essential for the increased neurogenesis associated with EC. Other investigators have found separable behavioral effects for EC and exercise with each being beneficial (Pietropaolo et al., 2006). Kempermann et al. (2010) have reviewed the current evidence, and they conclude that the effects of a running wheel and other aspects of EC are additive. Aerobic exercise is more effective at increasing neurogenesis while EC itself increases the rate of survival of new neurons, giving each a major role. This was further confirmed by Birch et al. (2013) who found, in addition to increased neuronal survival, that EC without aerobic exercise increased NGF (nerve growth factor) and synaptogenesis in the hippocampus including the dentate gyrus. It becomes apparent that the lack of increased dendritic growth in the dentate in male EC rats previously reported (Juraska et al., 1985) was due to synaptogenesis being associated with new neurons rather than augmentation of existing neurons as in the visual cortex.

Sex

Historically, the vast majority of studies looking at the effects of EC and IC have used male animals. Even in a recent survey that includes the many studies that employ enrichment to ameliorate neural problems (see **Table 2**, the majority only examine

Table 2 Examples of improvement from environmental stimulation

Trauma	*Reference*
Cortical injury	Kolb and Gibb (1991)
Developmental alcohol exposure	Klintsova et al. (1997)
Developmental lead exposure	Cao et al. (2008)
Psychiatric disorder models	Takuma et al. (2011)
Transgenerational effects of infantile stress	Leshem and Schulkin (2012)
Cocaine self-administration	Puhl et al. (2012)
Self-administration of methylphenidate	Alvers et al. (2012)
Models of brain disease	Pang and Hannan (2013)
Chronic neuropathic pain	Vachon et al. (2013)

male animals and only 17% look at both sexes (Simpson and Kelly, 2011)). However, the studies that have examined both males and females have often found different effects, both quantitative and qualitative, between the sexes. For example, the well-studied effect of increases in the size of the dendritic tree of the visual cortex in male rats in EC is not as pronounced in female rats, where neurons from two of the three cortical layers sampled had smaller EC-IC differences in females than males (Juraska, 1984; **Figure 4**). In the granule neurons in the dentate gyrus, on the other hand, females from EC had larger dendritic trees than those from IC, while the environment did not affect the dendrites in males (Juraska et al., 1985). The pattern was even more complicated in hippocampal CA3 which receives major input from the dentate granule neurons. Within these neurons, there were sex differences in reaction to the environment that varied between different portions of the dendritic tree (Juraska et al., 1989). One interesting correlate of these sex differences is that females in EC have a more pronounced rise in brain-derived neurotrophic factor (BDNF) than males in EC (Bakos et al., 2009).

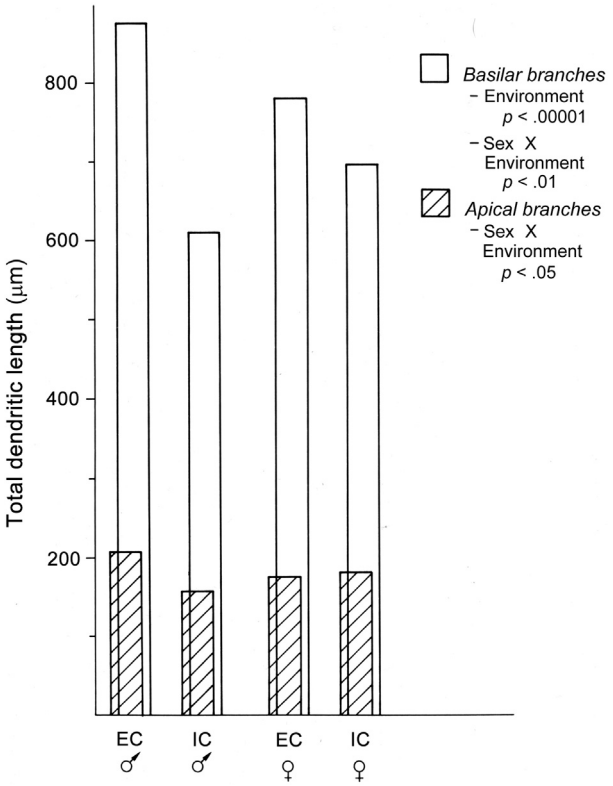

Figure 4 The dendritic length of two portions of the dendritic tree (basilar and apical oblique dendrites) of pyramidal neurons in the rat visual cortex. In both portions of the dendritic tree, there was a significant sex by environment interaction, so that males had larger differences due to the environment than females. The direction of this finding is the opposite of what was seen in the hippocampal dentate gyrus. Plotted from data in Juraska, J.M., 1984. Sex differences in dendritic response to differential experience in the rat visual cortex. Brain Research 295, 27–34.

The splenium of the rat corpus callosum carries axons between the two sides of the visual cortex almost exclusively (Kim et al., 1996). Given the greater plasticity in the rat visual cortex of males compared with females, it is interesting that in addition to the EC-IC differences listed in **Table 1**, female rats also had an increase in the number of myelinated axons following exposure to EC and male rats did not (Juraska and Kopcik, 1988). There are no simple generalizations about sex differences and the cellular effects of EC. It is obviously important to understand more of the mechanisms involved in the differences between the sexes in cellular plasticity to the environment.

Age

Both the starting age and the duration of enriched environments vary considerably between studies. However, cellular (dendritic) changes can be detected in as little as 4 days of exposure when rats have been exposed to EC immediately after weaning (Wallace et al., 1992). The greater plasticity of juvenile and adolescent rats may mean that the age of first exposure is relatively more important than duration. Most studies start enrichment at less than 2 months of age (Simpson and Kelly, 2011) which capitalizes on the growth and pruning that the adolescent brain continues to undergo (Juraska et al., 2013). It has been established that the dendritic tree is larger in the visual cortex in EC compared with IC rats when the environments are started in adulthood (after 4 months of age) (Uylings et al., 1978; Juraska et al., 1980), although the difference between EC and IC is not as great as in rats that are exposed starting at weaning (Volkmar and Greenough, 1972). Introducing adult males to group living is far more stressful than it is at weaning (Juraska, unpublished personal observations) which may contribute to this discrepancy.

The plastic cellular changes in adult rats in response to an enriched environment open the possibility that enrichment might help to ameliorate the effects of aging. The cortical thickness increases seen in younger animals in response to EC have also been documented in aged (2+ years old) rats (Diamond et al., 1985). Middle (15 months) to aged (24 months) rats also show increases in the cortical dendritic tree following EC (Green et al., 1983; Kolb et al., 2003). More direct comparisons of compensation for aging effects have been done. Through the use of an aged group examined at the time of initiation of the enriched environment, Greenough et al. (1986) demonstrated that the loss of dendritic branches on cerebellar Purkinje neurons during aging are halted in aged rats placed in EC. Likewise, Saito et al. (1994) found that a decrease in synaptic density in the cortex with age can be reversed by EC exposure, as can the age-related decrease in hippocampal neurogenesis, even if EC does not contain a running wheel (Speisman et al., 2013). Reviews covering details on behavioral and neurotransmitter effects of EC during aging are available (Frick and Benoit, 2010; Mora et al., 2007).

Applicability to Primates, Including Humans

At the time of Hebb's work in the 1940s and the Berkeley group's in the 1960s, the idea that enriched environments increased intelligence and even altered the structure of the

brain was considered radical. Neural plasticity in response to the environment was in its infancy. Today, the profound role of the environment on both the brain and behavior is widely accepted, including and importantly, for humans. At least some of this change in the attitude of both the scientific community and the general public is due to work on EC/SC/IC in rats. It has been bolstered by cellular evidence from primates exposed to more stimulating environments (Floeter and Greenough, 1979; Kozorovitskiy et al., 2005). The noninvasive structural magnetic resonance imaging (MRI) technique and the unfortunate circumstances of many infants and young children in institutionalized care have brought the evidence into the human realm as well (Sheridan et al., 2012). The importance of environmental stimulation has also become important as the human population ages (Volkers and Scherder, 2011). The systematic study of environmental stimulation changed how society views the human condition.

EC as Therapy

Now that the basic tenets of environmental stimulation and neural plasticity have been accepted, most current studies have applied EC as a therapy to improve both behavioral and neural outcomes in models of a wide array of disorders. A sampling of these is listed in **Table 2**. On the whole, environmental stimulation ameliorates every neural problem examined. The increases in several neurotrophin levels (i.e., growth factors) that result from exposure to enriched environments in many neural regions could be part of the explanation (Ickes et al., 2000). More details on the mechanisms whereby EC mitigates neural trauma can give clues for therapies.

From another vantage point, the improvement that almost always accompanies exposure to enrichment casts doubt on the validity of our animal models where rodents, in particular, live in barren, and often asocial, conditions. The lesson that the enriched condition taught us was that environmental experience is an important component in how the brain works and in behavioral performance. We need to apply this lesson to our animal models more widely.

See also: Hebb, Donald Olding (1904–1985); Learning and Memory, Neural Basis of; Neonatal Social Isolation: Effects on Neural Function; Neural Plasticity.

Bibliography

Alvers, K.M., Marusich, J.A., Gipson, C.D., Beckmann, J.S., Bardo, M., 2012. Environmental enrichment during development decreases intravenous self-administration of methylphenidate at low unit doses in rats. Behavioural Pharmacology 23, 650–657.

Bakos, J., Hlavacova, N., Rajman, M., Ondicova, K., Koros, C., Kitraki, E., Steinbusch, H.W.M., Jezova, D., 2009. Enriched environment influences hormonal status and hippocampal brain derived neurotrophic factor in a sex dependent manner. Neuroscience 164, 788–797.

Bennett, E.L., 1976. Cerebral effects of differential experience and training. In: Rosenzweig, M.R., Bennett, E.L. (Eds.), Neural Mechanisms of Learning and Memory. The MIT Press, Cambridge, pp. 279–287.

Bennett, E.L., Diamond, M.C., Krech, D., Rosenzweig, M.R., 1964. Chemical and anatomical plasticity of brain. Science 146, 610–619.

Birch, A.M., McGarry, N.B., Kelly, A.M., 2013. Short-term environmental enrichment, in the absence of exercise, improves memory, and increases NGF concentration, early neuronal survival, and synaptogenesis in the dentate gyrus in a time-dependent manner. Hippocampus 23, 437–450.

Black, J.E., Sirevaag, A.M., Greenough, W.T., 1987. Complex experience promotes capillary formation in young rat visual cortex. Neuroscience Letters 83, 351–355.

Black, J.E., Sirevaag, A.M., Wallace, C.S., Savin, M.H., Greenough, W.T., 1989. Effects of complex experience on somatic growth and organ development in rats. Developmental Psychobiology 22, 727–752.

Brown, R.T., 1968. Early experience and problem-solving ability. Journal of Comparative and Physiological Psychology 65, 433–440.

Cao, X., Huang, S., Ruan, D., 2008. Enriched environment restores impaired hippocampal long-term potentiation and water maze performance induced by developmental lead exposure in rats. Developmental Psychobiology 50, 307–313.

Diamond, M.C., Johnson, R.E., Protti, A.M., Ott, C., Kajisa, L., 1985. Plasticity in the 904-day-old male rat cerebral cortex. Experimental Neurology 87, 309–317.

Diamond, M.C., Rosenzweig, M.R., Krech, D., 1965. Relationships between body weight and skull development in rats raised in enriched and impoverished conditions. Journal of Experimental Zoology 160, 29–36.

Floeter, M.K., Greenough, W.T., 1979. Cerebellar plasticity: modification of Purkinje cell structure by differential rearing in monkeys. Science 206, 227–229.

Forgays, D.G., Forgays, J.W., 1952. The nature of the effect of free-environmental experience in the rat. Journal of Comparative and Physiological Psychology 45, 322–328.

Frick, K.M., Benoit, J.D., 2010. Use it or lose it: environmental enrichment as a means to promote successful cognitive aging. The Scientific World Journal 10, 1129–1141.

Geller, E., Yuwiler, A., Zolman, J.F., 1965. Effects of environmental complexity on constituents of brain and liver. Journal of Neurochemistry 12, 949–955.

Globus, A., Rosenzweig, M.R., Bennett, E.L., Diamond, M.C., 1973. Effects of differential experience on dendritic spine counts in rat cerebral cortex. Journal of Comparative and Physiological Psychology 82, 175–181.

Green, E.J., Greenough, W.T., Schlumpf, B.E., 1983. Effects of complex or isolated environments on cortical dendrites of middle-aged rats. Brain Research 264, 233–240.

Greenough, W.T., McDonald, J.W., Parnisari, R.M., Camel, J.E., 1986. Environmental conditions modulate degeneration and new dendrite growth in cerebellum of senescent rats. Brain Research 380, 136–143.

Greenough, W.T., West, R.W., DeVoogd, T.J., 1978. Subsynaptic plate perforations: changes with age and experience in the rat. Science 202, 1096–1098.

Greenough, W.T., Madden, T.C., Fleischmann, T.B., 1972. Effects of isolation, daily handling, and enriched rearing on maze learning. Psychonomic Science 27, 279–280.

Hebb, D.O., 1949. The Organization of Behavior. Wiley & Sons, New York.

Ickes, B.R., Pham, T.M., Sanders, L.A., Albeck, D.S., Mohammed, A.H., Granholm, A.C., 2000. Long-term environmental enrichment leads to regional increases in neurotrophin levels in rat brain. Experimental Neurology 164, 45–52.

Juraska, J.M., 1984. Sex differences in dendritic response to differential experience in the rat visual cortex. Brain Research 295, 27–34.

Juraska, J.M., Fitch, J.M., Henderson, C., Rivers, N., 1985. Sex differences in the dendritic branching of dentate granule cells following differential experience. Brain Research 333, 73–80.

Juraska, J.M., Fitch, J.M., Washburne, D.L., 1989. The dendritic morphology of pyramidal neurons in the rat hippocampal CA3 area. II. Effects of gender and the environment. Brain Research 479, 115–119.

Juraska, J.M., Henderson, C., Müller, J., 1984. Differential rearing experience, gender, and radial maze performance. Developmental Psychobiology 17, 209–215.

Juraska, J.M., Kopcik, J.R., 1988. Sex and environmental influences on the size and ultrastructure of the rat corpus callosum. Brain Research 450, 1–8.

Juraska, J.M., Greenough, W.T., Elliott, C., Mack, K.J., Berkowitz, R., 1980. Plasticity in adult rat visual cortex: an examination of several cell populations after differential rearing. Behavioral and Neural Biology 29, 157–167.

Juraska, J.M., Sisk, C.L., Don Carlos, L.L., 2013. Sexual differentiation of the adolescent rodent brain: hormonal influences and developmental mechanisms. Hormones and Behavior 64, 203–210.

Kempermann, G., Fabel, K., Ehninger, D., Babu, H., Leal-Galicia, P., Garthe, A., Wolf, S.A., 2010. Why and how physical activity promotes experience-induced brain plasticity. Frontiers in Neuroscience 4, 189.

Kempermann, G., Kuhn, H.G., Gage, F.H., 1997. More hippocampal neurons in adult mice living in an enriched environment. Nature 386, 493–495.

Kim, J.H., Ellman, A., Juraska, J.M., 1996. A reexamination of sex differences in axon density and number in the splenium of the rat corpus callosum. Brain Research 740, 47–56.

Klintsova, A.Y., Matthews, J.T., Goodlett, C.R., Napper, R.M., Greenough, W.T., 1997. Therapeutic motor training increases parallel fiber synapse number per Purkinje neuron in cerebellar cortex of rats given postnatal binge alcohol exposure: preliminary report. Alcoholism: Clinical and Experimental Research 21, 1257–1263.

Kolb, B., Gibb, R., 1991. Environmental enrichment and cortical injury: behavioral and anatomical consequences of frontal cortex lesions. Cerebral Cortex 1, 189–198.

Kolb, B., Gibb, R., Gorny, G., 2003. Experience-dependent changes in dendritic arbor and spine density in neocortex vary qualitatively with age and sex. Neurobiology of Learning and Memory 79, 1–10.

Kozorovitskiy, Y., Gross, C.G., Kopil, C., Battaglia, L., McBreen, M., Stranahan, A.M., Gould, E., 2005. Experience induces structural and biochemical changes in the adult primate brain. Proceedings of the National Academy of Sciences of the United States of America 102, 17478–17482.

Krech, D., Rosenzweig, M.R., Bennett, E.L., 1962. Relations between chemistry and problem-solving among rats raised in enriched and impoverished environments. Journal of Comparative and Physiological Psychology 55, 801–807.

Krech, D., Rosenzweig, M.R., Bennett, E.L., 1966. Environmental impoverishment, social isolation, and changes in brain chemistry and anatomy. Physiology & Behavior 1, 99–104.

Leshem, M., Schulkin, J., 2012. Transgenerational effects of infantile adversity and enrichment in male and female rats. Developmental Psychobiology 54, 169–186.

Mirmiran, M., Van Den Dungen, H., Uylings, H.B., 1982. Sleep patterns during rearing under different environmental conditions in juvenile rats. Brain Research 233, 287–298.

Mora, F., Segovia, G., Del Arco, A., 2007. Aging, plasticity and environmental enrichment: structural changes and neurotransmitter dynamics in several areas of the brain. Brain Research Reviews 55, 78–88.

Moser, M.B., Trommald, M., Egeland, T., Andersen, P., 1997. Spatial training in a complex environment and isolation alter the spine distribution differently in rat CA1 pyramidal cells. Journal of Comparative Neurology 380, 373–381.

Nyman, A.J., 1967. Problem solving in rats as a function of experience at different ages. Journal of Genetic Psychology 110, 31–39.

Pang, T.Y., Hannan, A.J., 2013. Enhancement of cognitive function in models of brain disease through environmental enrichment and physical activity. Neuropharmacology 64, 515–528.

Pietropaolo, S., Feldon, J., Alleva, E., Cirulli, F., Yee, B.K., 2006. The role of voluntary exercise in enriched rearing: a behavioral analysis. Behavioral Neuroscience 120, 787–803.

Puhl, M.D., Blum, J.S., Acosta-Torres, S., Grigson, P.S., 2012. Environmental enrichment protects against the acquisition of cocaine self-administration in adult male rats, but does not eliminate avoidance of a drug-associated saccharin cue. Behavioural Pharmacology 23, 43–53.

Riege, W.H., Morimoto, H., 1970. Effects of chronic stress and differential environments upon brain weights and biogenic amine levels in rats. Journal of Comparative and Physiological Psychology 71, 396–404.

Rosenzweig, M.R., 1971. Effects of environment on development of brain and of behavior. In: Tobach, E., Aronson, L.R., Shaw, E. (Eds.), The Biopsychology of Development. Academic Press, New York, pp. 303–342.

Rosenzweig, M.R., Bennett, E.L., Hebert, M., Morimoto, H., 1978. Social grouping cannot account for cerebral effects of enriched environments. Brain Research 153, 563–576.

Rosenzweig, M.R., Krech, D., Bennett, E.L., Diamond, M., 1962. Effects of environmental complexity and training on brain chemistry and anatomy: a replication and extension. Journal of Comparative and Physiological Psychology 55, 429–437.

Rosenzweig, M.R., Love, W., Bennett, E.L., 1968. Effects of a few hours a day of enriched experience on brain chemistry and brain weights. Physiology and Behavior 3, 819–825.

Saito, S., Kobayashi, S., Ohashi, Y., Igarashi, M., Komiya, Y., Ando, S., 1994. Decreased synaptic density in aged brains and its prevention by rearing under enriched environment as revealed by synaptophysin contents. Journal of Neuroscience Research 39, 57–62.

Seymoure, P., Dou, H., Juraska, J.M., 1996. Sex differences in radial maze performance: influence of rearing environment and room cues. Psychobiology 24, 33–37.

Sheridan, M.A., Fox, N.A., Zeanah, C.H., McLaughlin, K.A., Nelson III, C.A., 2012. Variation in neural development as a result of exposure to institutionalization early in childhood. Proceedings of the National Academy of Sciences of the United States of America 109, 12927–12932.

Simpson, J., Kelly, J.P., 2011. The impact of environmental enrichment in laboratory rats–behavioural and neurochemical aspects. Behavioural Brain Research 222, 246–264.

Sirevaag, A.M., Greenough, W.T., 1987. Differential rearing effects on rat visual cortex synapses. III. Neuronal and glial nuclei, boutons, dendrites and capillaries. Brain Research 424, 320–332.

Speisman, R.B., Kumar, A., Rani, A., Pastoriza, J.M., Severance, J.E., Foster, T.C., Ormerod, B.K., 2013. Environmental enrichment restores neurogenesis and rapid acquisition in aged rats. Neurobiology of Aging 34, 263–274.

Takuma, K., Ago, Y., Matsuda, T., 2011. Preventive effects of an enriched environment on rodent psychiatric disorder models. Journal of Pharmacological Sciences 117, 71–76.

Turner, A.M., Greenough, W.T., 1983. Synapses per neuron and synaptic dimensions in occipital cortex of rats reared in complex, social, or isolation housing. Acta Stereologica 2, 239–244.

Uphouse, L.L., Brown, H., 1981. Effect of differential rearing on brain, liver and adrenal tissues. Developmental Psychobiology 14, 273–278.

Uylings, H.B., Kuypers, K., Diamond, M.C., Veltman, W.A., 1978. Effects of differential environments on plasticity of dendrites of cortical pyramidal neurons in adult rats. Experimental Neurology 62, 658–677.

Vachon, P., Millecamps, M., Low, L., Thompsosn, S.J., Pailleux, F., Beaudry, F., Bushnell, C.M., Stone, L., 2013. Alleviation of chronic neuropathic pain by environmental enrichment in mice well after the establishment of chronic pain. Behavioral and Brain Functions 9, 22.

Volkers, K.M., Scherder, E.J.A., 2011. Impoverished environment, cognition, aging and dementia. Reviews in the Neurosciences 22, 259–266.

Volkmar, F.R., Greenough, W.T., 1972. Rearing complexity affects branching of dendrites in the visual cortex of the rat. Science 176, 1445–1447.

Wallace, C.S., Kilman, V.L., Withers, G.S., Greenough, W.T., 1992. Increases in dendritic length in occipital cortex after 4 days of differential housing in weanling rats. Behavioral and Neural Biology 58, 64–68.

West, R.W., Greenough, W.T., 1972. Effect of environmental complexity on cortical synapses of rats. Behavioral Biology 7, 279–284.

Woodcock, E.A., Richardson, R., 2000. Effects of multisensory environmental stimulation on contextual conditioning in the developing rat. Neurobiology of Learning and Memory 74, 89–104.

Zolman, J.F., Morimoto, H., 1965. Cerebral changes related to duration of environmental complexity and locomotor activity. Journal of Comparative Physiology 60, 382–387.

Brain Stimulation Reward

Peter Shizgal, Concordia University, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In 1953, Olds and Milner discovered that rats would readily learn to work for electrical stimulation of certain brain sites. Their findings inspired a large body of research on the neural basis of reward, motivation, and learning. Unlike consummatory behaviors, which satiate as a result of ingestion of and contact with the goal object, performance for rewarding brain stimulation is remarkably stable and persistent. Pursuit of the stimulation is enhanced by different classes of dependence-inducing drugs, suggesting that common neural mechanisms underlie the rewarding effects of drugs and electrical brain stimulation. Indeed, dopamine-containing neurons in the midbrain are implicated in both phenomena. Major schools of thought that have addressed brain stimulation reward (BSR) differ with regards to the roles played by hedonic experience and craving, although there is substantial overlap between the different viewpoints. A tradition that arose in the study of machine learning has been brought to bear on the role of dopamine neurons in reward-related learning in animals and on the phenomenon of intracranial self-stimulation. Neuroeconomic perspectives strive to integrate the processing of benefits, costs, and risks into an account of decision-making grounded in brain circuitry. Adjudication of the differences between the various viewpoints and progress toward identifying the relevant neural circuitry has been hindered by the lack of specificity inherent in the use of electrical stimulation to study central nervous system function. Many neurons in addition to those targeted are likely activated by such stimulation. Recently developed optogenetic methods may overcome this obstacle, providing much more specific means for stimulating or silencing populations of nerve cells selected on the basis of their gene expression, cell-body location, and projections. Coupled with behavioral methods of increasing sophistication and specificity and with quantitative modeling of signal flow in the relevant neural circuitry, the new optogenetic methods promise to bring us much closer to fulfillment of the hopes engendered long ago by the discovery of BSR.

The Discovery and Its Context

Brain stimulation reward (BSR) was discovered in 1953 by James Olds and Peter Milner (Olds and Milner, 1954; Milner, 1989; Olds, 1973). They had come to McGill University to work with D.O. Hebb, inspired by his groundbreaking theoretical work. In contrast to the dominant behaviorist orthodoxies of the day, which isolated psychology from the emerging disciplines of neuroscience and cognitive psychology, Hebb's views linked brain, mind, and behavior.

New findings and ideas about the neural bases of motivation and reinforcement provided a rich context for the discovery of BSR (Valenstein, 1973). For example, the demonstration that lesions of different hypothalamic nuclei could lead either to massive overeating or starvation contributed to a seminal, physiologically based theory of motivation (Stellar, 1954). During that era, English-speaking investigators became familiar with Hess' use of electrical stimulation to map the hypothalamic control of autonomic outflow and behavior (Hess, 1949). Among Hess' observations were stimulation-elicited responses suggesting states of rage or fear. Shortly before Olds and Milner's discovery, Delgado, Roberts, and Miller (Delgado et al., 1954) showed that cats would learn to escape electrical stimulation of certain deep forebrain structures. Another key component of the context was research linking brain stem circuitry to the control of arousal and vigilance (Moruzzi and Magoun, 1949). The pattern of connections between nerve cells in the brain stem region of interest was seen as netlike ('reticulated'); thus the region was dubbed the 'reticular formation.'

At the time that Olds joined Hebb's group, Milner and Seth Sharpless, another graduate student, had developed a theory linking reticular activation to positive reinforcement, the process by which desirable outcomes strengthen the behaviors that produce them. Their experimental tests of this idea proved disappointing: The rats tended to avoid locations in a maze where stimulation of the reticular formation had been delivered.

A strikingly different reaction was noted in a rat prepared by Olds. The electrode had somehow missed its reticular-formation target and ended up instead in the basal forebrain. This subject returned repeatedly to the location where stimulation had been delivered. As the experimenters altered the rewarded location, the animal's behavior changed accordingly, and it directed its searching toward the site where it had most recently hit pay dirt. Olds then succeeded in training the rat to press a lever that triggered delivery of the stimulation. This animal worked indefatigably to stimulate its own brain. Such behavior has since been termed 'intracranial self-stimulation' (ICSS), and the effect that causes the animal to seek out the stimulation is called 'BSR.'

In the initial report published by Olds and Milner (Olds and Milner, 1954), they presciently discuss their findings "as possibly laying a methodological foundation for a physiological study of the mechanisms of reward." Indeed, their paper was a wellspring for thousands of subsequent experiments.

Why did the report by Olds and Milner have such a large and enduring impact? Perhaps paramount is the promise their results offered that psychological phenomena such as learning, reinforcement, and motivation could be investigated fruitfully by physiological means. The authors also pointed out that the phenomenon of BSR could be used to distinguish between competing behavioral theories. For example, a prominent theory at the time held that reinforcement arose from the

reduction of a drive, such as the abatement of hunger. This view is not easily reconciled with the fact that rewarding stimulation of many sites arouses the animal rather than calming it and that stimulation of some BSR sites can also induce sated animals to engage in consummatory behaviors such as eating and drinking (Margules and Olds, 1962; Hoebel and Teitelbaum, 1962).

The Phenomenon

The behavior of the self-stimulating rat is striking to behold, particularly in the case of electrodes arrayed along the medial forebrain bundle (MFB). When the rewarding stimulation is strong and easily triggered, the rat attacks the lever with singular focus, pressing repeatedly and forcefully at high rates. This behavior is remarkably persistent. For example, a rat working for MFB stimulation maintained an average response rate above 2000 presses per hour over a continuous 26-hour period (Olds, 1958a). Rats have been shown to cross an electrified grid (Olds, 1958b) or to gallop uphill along a runway, leaping over hurdles (Edmonds and Gallistel, 1974) in order to access a lever that triggers delivery of the stimulation. When food was provided uniquely in an ICSS chamber in which strong stimulation was available, rats worked for the electrical reward in lieu of feeding and thus starved themselves (Routtenberg and Lindy, 1965).

Why would an animal work so assiduously for a reward that has no evident biological value? It has long been suspected that the neural activity triggered by the stimulation mimics some aspects of natural rewards. Indeed, MFB stimulation is as good an economic substitute for food and water as they are for themselves (Green and Rachlin, 1991), and the rewarding effect of such stimulation both competes and summates with the rewarding effects of sweet (Conover and Shizgal, 1994a) or salty (Conover et al., 1994) solutions. The incompleteness of the mimicry may give rise to some of the differences between BSR and natural rewards. Unlike performance for natural consummatory rewards, ICSS does not result in the accumulation of a substance in the body, which may explain why the appetite for stimulation does not satiate. Similarly, the electrical reward bypasses peripheral receptors for taste, smell, and touch, which may explain why the accumulation of nutrients in the gut undermines the rewarding effect of a gustatory stimulus (sucrose) more severely than the rewarding effect of MFB stimulation (Conover and Shizgal, 1994b).

The nature of the rewarding effect elicited by electrical stimulation may vary both across and within brain sites. For example, food or water deprivation can alter preference between stimulation of different brain sites (Gallistel and Beagley, 1971), and the potentiating effect of chronic food restriction is seen only at some MFB sites (Blundell and Herberg, 1968). Moreover, the fat-signaling hormone, leptin, has opposite effects on self-stimulation of different brain sites, depending on whether or not the rewarding effect is sensitive to chronic food restriction (Fulton et al., 2000).

Anatomy and Neurochemistry

The sites where BSR can be produced occupy a substantial proportion of the brain volume (Wise, 1996). Positive sites are distributed from the olfactory bulbs, at the front of the brain, to the nucleus of the solitary tract, at the back, and are found both in cortical and subcortical regions of the forebrain. The best-characterized sites lie along the trajectories of the MFB and of fibers coursing near the midline of the brain stem. In both cases, the neurons that carry the reward-related signal away from the electrode tip include cells with relatively fast-conducting, longitudinally oriented, myelinated axons (Shizgal, 1997; Miguelez and Bielajew, 2004). This characterization narrows the field of plausible candidates, but it leaves open multiple possibilities. The recent technical innovations described in the last section of this article hold great promise for the long-sought identification of these neurons.

Several neurotransmitter systems are implicated in BSR. There is a large body of evidence tying dopamine-containing neurons to the rewarding effect (Wise and Rompré, 1989). ICSS is potentiated by manipulations that enhance transmission in these cells and weakened by manipulations that decrease dopaminergic neurotransmission. Recently, Witten, Steinberg, and their coworkers demonstrated that rats will work for direct optical activation of midbrain dopamine neurons (Witten et al., 2011). The dopamine-containing neurons that project from the ventral tegmental area of the midbrain to the nucleus accumbens, a structure at the base of the anterior forebrain, are thought to play a particularly important role in BSR as well as in the effects of natural rewards such as food and water. Mice have been shown to self-stimulate for direct optical activation of excitatory inputs to the nucleus accumbens from glutamate-containing neurons in other forebrain structures (Stuber et al., 2011; Britt et al., 2012).

Although ventral tegmental dopamine neurons play a crucial role in BSR and project to forebrain structures via the MFB, their axons are difficult to excite under the conditions of ICSS experiments. Instead, or in addition, the rewarding effect may well arise from activation of inputs to the ventral tegmental dopamine neurons (Shizgal and Murray, 1989; Shizgal, 1997; Moisan and Rompré, 1998).

Acetylcholine-containing neurons in the pedunculopontine and laterodorsal tegmental nuclei play a role in BSR, are activated transsynaptically by rewarding MFB stimulation (Yeomans et al., 2000; Rada et al., 2000), and provide excitatory input to ventral tegmental dopamine neurons (Forster et al., 2002). Mice will work to disinhibit dopamine neurons in the ventral tegmental area via a disynaptic GABAergic pathway arising in the bed nucleus of the stria terminalis in the basal forebrain (Jennings et al., 2013), and there are multiple pathways employing glutamate as a neurotransmitter that provide excitatory input to the ventral tegmental area from forebrain and brain stem structures (Geisler et al., 2007).

Like ICSS self-administration and abuse of drugs entail the copious expenditure of behavioral resources in pursuit of an objective that appears worthless or detrimental to biological fitness. A full account of the neural circuitry responsible for ICSS is likely to provide insight into how dependence-inducing drugs attain their powerful influence over behavior and decision-making. Performance for BSR is enhanced by many drugs that are both self-administered by laboratory animals and subject to abuse by humans (Wise, 1996). The dominant hypothesis concerning these effects is that the drugs in question alter dopaminergic neurotransmission directly or indirectly.

Much recent work on ICSS and reward seeking is directed at inputs to ventral tegmental dopamine neurons or to their postsynaptic targets, which must be considered as possible substrates of drug action in addition to the dopamine neurons themselves.

Schools of Thought

Theoretical accounts of ICSS have emphasized different aspects of the phenomenon. However, the differences are often less marked than the contrasting labels for the theories suggest.

Hedonic

An early article on ICSS in *Scientific American* bears the provocative title: "Pleasure Centers in the Brain" (Olds, 1956). When someone first observes a self-stimulating rat, they typically speculate that the stimulation 'makes the rat feel good.' It seems commonsensical that one would tend to repeat an activity that produced pleasurable consequences. The converse, however, is far from clear. Does the fact that an activity is repeated necessarily imply that it generates a positive hedonic state? Will the activity be repeated if, and only if, it is pleasurable?

Pleasure implies awareness, a faculty believed to depend on resources in short supply, such as attention and working memory. It would be highly maladaptive to pass all signals controlling action through such a narrow processing bottleneck. Thus, contemporary depictions of the control of action place most of the burden on distributed processes operating below the waterline of awareness (Baars, 1988; LeDoux, 1996). On that view, the repeated pressing of the lever does not necessarily imply a positive hedonic state (Shizgal, 1999).

Theorists who have ventured onto this slippery terrain have typically been more circumspect than the labels for their theories might suggest (Wise, 1982, 2008; Kringelbach and Berridge, 2009). Their task is not easy. On the one hand, they cannot ignore reports from early studies of human patients who voluntarily self-administered electrical stimulation to sites homologous to those that support ICSS in laboratory animals. These accounts include descriptions of pleasurable experience (Sem-Jacobsen, 1959; Heath, 1964). On the other hand, some have contested the interpretation of the verbal subjective reports from these studies of human patients (Kringelbach and Berridge, 2009), and we lack the means to infer the subjective experience of nonhuman animals. Moreover, subjective pleasure and persistent operant responding for rewards have been dissociated in studies of human participants (Lamb et al., 1991).

Seeking and Craving

If a path to the lever is blocked by an obstacle, the self-stimulating rat will seek another. Such flexible approach to a goal is typical of what psychologists call appetitive behavior, and it contrasts with the more stereotyped patterns observed once the animal procures a natural goal object, such as food, and begins to consume it. Potentiation of flexible approach figures heavily in several different accounts of the ICSS phenomenon. Although Wise applied the psychiatric term, anhedonia, to his theory of why reductions in dopaminergic neurotransmission attenuate ICSS, drug self-administration, and operant performance for natural rewards, the core of his formulation concerns what he calls 'motivational arousal,' a state that promotes flexible reward seeking. His theory differs from several others in that he views subjective pleasure as a normal, but not obligatory, concomitant of the motivational arousal engendered by exposure to rewards. In contrast, the incentive-sensitization hypothesis of Robinson and Berridge (1993; 1998) portrays the ICSS phenomenon as a manifestation of what they call 'wanting,' a process that endows stimuli with the ability to seize attention and to promote anticipatory, flexible approach (Ikemoto and Panksepp, 1999). They argue that dopaminergic neurotransmission is essential to 'wanting' but not to a contrasting component of reward that they term 'liking.' When 'liking' impinges on consciousness, it is experienced as pleasure. It is also manifested in consummatory behavior so that we ingest foods that we like and eschew foods that we do not. The subjective states of rats are hidden from us, but their ingestive behavior is not. Thus, Berridge and his coworkers use 'liking' in quotation marks to describe behavior that they believe to arise from the neural underpinning of hedonic states, whether these states are experienced or not. Experimenter-delivered stimulation of ICSS sites fails to strengthen ingestive patterns and may weaken them, a finding that these theorists take to reflect a dissociation between 'wanting' and 'liking.' Although this interpretation has been contested (Wise, 2008), the distinction between wanting and liking has been highly influential, particularly because of its correspondence to the self-reports of long-time drug users who evince continued cravings but claim that the drug no longer generates euphoric effects. Indeed, in studies of human participants, partial blockade of dopamine synthesis reduced drug craving without systematically undermining self-reported euphoric effects (Leyton et al., 2007).

Reinforcement Learning

Self-stimulation of many brain sites is acquired quickly, and the behavior typically extinguishes rapidly once the reward is no longer delivered. This rapid reinforcement learning is among the most striking aspects of the ICSS phenomenon, and it has played a prominent role in theoretical accounts. During the 1990s, the confluence of work in machine learning and electrophysiology cast the reinforcement-learning aspects of ICSS in a new light.

In their quest to develop intelligent machines, researchers came to appreciate that much of the knowledge required to solve real-world problems cannot be specified in advance. The environment is not entirely predictable, and its complexity is effectively without limit. Under such conditions, it is desirable to imbue machines with the ability to learn about the consequences of their actions and the relationships between events. The algorithms developed to meet these objectives share important properties with formulations developed by students of animal learning while extending the quantitative rigor and scope of those ideas (Sutton and Barto, 1998). One such algorithm, called temporal-difference learning, entails moment-to-moment updating of predictions concerning future rewards and comparison of these predictions to the

rewards actually received. Discrepancies between predictions and outcomes, called temporal-difference errors, cause reward predictions and the associated reward-seeking actions to be modified and improved.

Montague, Dayan, and Sejnowski recognized a striking correspondence between the properties of temporal-difference errors and the activity of midbrain dopamine neurons observed by Schultz and colleagues during Pavlovian and operant conditioning experiments carried out in monkeys (Montague et al., 1996; Schultz et al., 1997). They proposed that this correspondence could account for the ICSS phenomenon. According to this view, the brief burst of ('phasic') firing in midbrain dopamine neurons that is triggered by electrical stimulation constitutes a temporal-difference error that alters weights in the neural networks encoding the value of the reward-seeking actions and the predictions associated with response-related cues. In effect, the self-stimulating rat is portrayed as the victim of a hallucination of value. Although the stimulation produces no homeostatic or reproductive benefit, cues that predict it and the actions that procure it come to be represented as increasingly valuable due to the reward-prediction errors it generates.

A problem with the original formulation of the prediction-error hypothesis of ICSS is that the weights increase progressively with delivery of successive stimulation trains and should saturate eventually regardless of the strength of any suprathreshold stimulation train; saturation would be achieved rapidly in the case of strong trains and more slowly in the case of weaker ones. This prediction does not jibe well with the finding that rats systematically choose higher strength trains in preference to weaker ones, even after extensive experience with the stimulation (Simmons and Gallistel, 1994). The evidence that midbrain dopamine neurons are activated transsynaptically by MFB stimulation (Shizgal, 1997) can rescue the hypothesis: an inhibitory input representing the reward prediction could come to counteract the transient activation produced by the stimulation train, reducing and eventually eliminating the prediction error. Such a nulling effect should be much less likely in the case of optical stimulation, which can excite the dopamine neurons directly. Substituting optical for electrical stimulation could provide a powerful test of the hypothesis that induced bursts of firing in dopamine neurons are the cause of the observed reward-related learning.

Neuroeconomic

The study of the neural basis of decision-making and the evaluation of the benefits, costs, and risks on which choices are based has come to be called neuroeconomics (Glimcher and Rustichini, 2004; Loewenstein et al., 2008). The first papers written from this perspective were accounts of ICSS (Shizgal and Conover, 1996; Shizgal, 1997), and it has been argued that there is a striking correspondence between the neuroeconomic approach to ICSS and some core ideas in economics, judgment, and decision-making (Shizgal, 2012). The neuroeconomic approach shares with the reinforcement-learning approach a commitment to quantitative models that can be tested by means of behavioral and neural measurements. One distinction between the approaches concerns the phase of the experiment that is of primary interest. The reinforcement-learning approach focuses primarily on how relationships between events and contingencies between actions and their consequences are learned. In contrast, neuroeconomic experiments typically begin with a subject that has already learned a task. In the case of ICSS, such work has focused (1) on the psychophysical transformations that translate objective variables, such as the strength and cost of rewards, into their subjective equivalents, and (2) on the stages of processing at which particular neural populations contribute to the computation of reward value (Hernandez et al., 2010). That said, the reinforcement-learning approach has been extended to address the issue of reward-related costs; whereas this approach links reward-related learning to phasic firing of dopamine neurons, effort costs have been related to more slowly varying, tonic levels of dopamine (Niv et al., 2007).

Methodological Issues and Emerging Directions

Biophysical Specificity

Studies employing electrical stimulation of the central nervous system have been plagued by a lack of biophysical specificity. The stimulation typically activates a set of neurons of unknown composition in addition to the population targeted by the experimenter. This is the principal reason why the directly stimulated neurons that give rise to the rewarding effect have proved so difficult to identify. The recent development of optogenetic methods promises to circumvent the lack of specificity inherent in electrical stimulation methods while retaining all of the principal advantages (Deisseroth, 2010; Yizhar et al., 2011).

Using optogenetic methods, neural activity can be silenced or induced on timescales as short as that corresponding to single nerve impulses and as long as the actions of many drugs. These effects can be restricted to specific neural populations targeted on the basis of conditional gene expression, cell-body location, and axonal projections. The optical component of optogenetic technologies consists of two elements: molecules that transduce light into cellular signals and a mechanism to deliver appropriate photic stimuli to the targeted cells. The transducer molecules are opsins, the class of proteins that render photoreceptors sensitive to light. Unlike vertebrate opsins, which compose part of a multistage chain of complex chemical reactions, microbial opsins are far simpler, functioning directly as light-gated ion channels or pumps that alter the movement of electrical charge across the cell membrane so as to either promote or suppress the generation of neural signals. The genetic component of optogenetic technologies consists of methods for inserting microbial opsin genes into the genome of neurons, restricting the expression of these genes to specifically targeted cell populations and promoting trafficking of the protein products of these genes to the cell membrane. Fine fiber-optic probes implanted in the brain are used to deliver light at a wavelength selected for optimal capture by the microbial opsin. Optogenetic methods are complemented by molecular neuroanatomical techniques for labeling and visualizing specific populations of neurons.

The initial attempts to train rodents to perform operant responses to obtain optical stimulation of ventral tegmental

dopamine neurons were only moderately successful. In one case (Adamantidis et al., 2011), mice could be trained to lever press only if food was delivered along with the laser illumination; however, operant performance was subsequently maintained by optical stimulation alone, albeit at low response rates. In a second study (Kim et al., 2012), mice were successfully trained to nose poke for optical stimulation of ventral tegmental neurons but again, response output was low. In contrast, rats expressing channelrhodopsin2, an excitatory opsin, in dopamine neurons respond in a vigorous, sustained manner for optical stimulation of the ventral tegmental area (Witten et al., 2011). This work and associated studies in which specific inputs to ventral tegmental dopamine neurons were manipulated optogenetically (Britt et al., 2012; Jennings et al., 2013) demonstrate the promise of the new methods to overcome technical obstacles to determining the neural underpinnings of reward seeking.

Behavioral Specificity

Another challenge to achieving an adequate neurobehavioral account of ICSS and of reward seeking in general arises from what can be called 'multistage causal convergence.' There are many possible reasons why a laboratory animal may choose to perform an operant response. The various processes that can support such responses are likely composed of multiple stages. For example, the decision to press a lever may arise from the separate computation of benefits, costs, and risks, the folding together of the results of these computations into a unitary estimate of value that is compared with estimates of the values of other available actions, and the translation of the winning estimate into reward-seeking actions. When the proclivity of the animal to press the lever is altered by a neural manipulation, how is one to know which of the possible causes of lever pressing has been affected and at what stage of processing? Answering such questions requires that the specificity and power of behavioral measurement methods match those of the promising new methods for optical stimulation, optical silencing, and neuroanatomical visualization.

Work employing a neuroeconomic approach provides an example of the increasing specificity of new behavioral methods and the multistage nature of the processes underlying ICSS. It has long been known that psychomotor stimulants boost performance for BSR, an effect that been attributed to a drug-induced increase in the sensitivity of reward-related circuitry (Crow, 1970). This hypothesis was tested by means of a multistage model of ICSS coupled to a behavioral testing paradigm that can distinguish the signatures of drug action at different stages of processing (Hernandez et al., 2010). According to this model, the effect of the volley of action potentials triggered by the rewarding stimulation is integrated according to an aggregate rate code; reward strength is represented by the number of firings within a time window, integrated across neurons. This count is transformed nonlinearly into a subjective representation of reward intensity (Simmons and Gallistel, 1994) and then combined with information about reward cost and probability so as to yield an estimate of the payoff from lever pressing. If psychomotor stimulants increase the sensitivity of the circuitry to electrical activation, these drugs should lower the number of firings required to produce a reward of given intensity. By measuring ICSS as a function of both the strength and cost of rewarding stimulation, Hernandez and colleagues showed that cocaine enhances reward seeking primarily by other means, such as a reduction in subjective reward cost or amplification of the output of the circuitry performing the integration (Hernandez et al., 2010).

Emerging Directions

Optogenetic experiments on reward seeking are typically complemented by electrophysiological measurement of neural activity. Integration of these approaches with increasingly specific behavioral testing methods and with quantitative modeling in the reinforcement learning, psychophysical and neuroeconomic traditions may address many of the problems that have hindered progress in the study of ICSS and of the neural basis of reward seeking. This would provide a solid foundation for testing competing accounts of how goals are evaluated, selected, and sought, and how information about benefits, costs, and risks is learned, used, and abused.

Acknowledgments

The author's research on brain stimulation reward has been supported by the Canadian Institutes of Health Research, the Natural Science and Engineering Research Council of Canada, the Fonds de Recherche du Québec (Santé), and the Concordia University Research Chairs program. Brian Dunn, Daniel Palacios, and Sarah Nolan-Poupart provided helpful comments on the article.

See also: Learning and Memory, Neural Basis of; Motivation and Actions, Psychology of; Neuroeconomics; Psychophysical Theory and Laws, History of; Reinforcement: Neurochemical Substrates.

Bibliography

Adamantidis, A.R., et al., 2011. Optogenetic interrogation of dopaminergic modulation of the multiple phases of reward-seeking behavior. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 31 (30), 10829–10835.

Baars, B.J., 1988. A Cognitive Theory of Consciousness. Cambridge University Press, Cambridge, England; New York.

Berridge, K.C., Robinson, T.E., 1998. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Research Brain Research Reviews 28 (3), 309–369.

Blundell, J.E., Herberg, L.J., 1968. Relative effects of nutritional deficit and deprivation period on rate of electrical self-stimulation of lateral hypothalamus. Nature 219 (5154), 627–628.

Britt, J.P., et al., 2012. Synaptic and behavioral profile of multiple glutamatergic inputs to the nucleus accumbens. Neuron 76 (4), 790–803.

Conover, K.L., Shizgal, P., 1994a. Competition and summation between rewarding effects of sucrose and lateral hypothalamic stimulation in the rat. Behavioral Neuroscience 108 (3), 537–548.

Conover, K.L., Shizgal, P., 1994b. Differential effects of postingestive feedback on the reward value of sucrose and lateral hypothalamic stimulation in rats. Behavioral Neuroscience 108 (3), 559–572.

Conover, K.L., Woodside, B., Shizgal, P., 1994. Effects of sodium depletion on competition and summation between rewarding effects of salt and lateral hypothalamic stimulation in the rat. Behavioral Neuroscience 108 (3), 549–558.

Crow, T.J., 1970. Enhancement by cocaine of intra-cranial self-stimulation in the rat. Life Sciences 9 (7), 375–381.

Deisseroth, K., 2010. Controlling the brain with light. Scientific American 303 (5), 48–55.

Delgado, J.M., Roberts, W.W., Miller, N.E., 1954. Learning motivated by electrical stimulation of the brain. The American Journal of Physiology 179 (3), 587–593.

Edmonds, D.E., Gallistel, C.R., 1974. Parametric analysis of brain stimulation reward in the rat: III. Effect of performance variables on the reward summation function. Journal of Comparative and Physiological Psychology 87 (5), 876–883.

Forster, G.L., et al., 2002. M5 muscarinic receptors are required for prolonged accumbal dopamine release after electrical stimulation of the pons in mice. Journal of Neuroscience 22 (1), RC190.

Fulton, S., Woodside, B., Shizgal, P., 2000. Modulation of brain reward circuitry by leptin. Science 287 (5450), 125–128. New York.

Gallistel, C.R., Beagley, G., 1971. Specificity of brain stimulation reward in the rat. Journal of Comparative and Physiological Psychology 76 (2), 199–205.

Geisler, S., et al., 2007. Glutamatergic afferents of the ventral tegmental area in the rat. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 27 (21), 5730–5743.

Glimcher, P.W., Rustichini, A., 2004. Neuroeconomics: the consilience of brain and decision. Science 306 (5695), 447–452. New York.

Green, L., Rachlin, H., 1991. Economic substitutability of electrical brain stimulation, food, and water. Journal of the Experimental Analysis of Behavior 55 (2), 133–143.

Heath, R.G., 1964. Pleasure response of human subjects to direct stimulation of the brain: physiologic and psychodynamic considerations. In: Heath, R.G. (Ed.), The Role of Pleasure in Behavior. Harper & Row, New York, pp. 219–243.

Hernandez, G., et al., 2010. At what stage of neural processing does cocaine act to boost pursuit of rewards? PLoS ONE 5 (11).

Hess, W., 1949. The Central Control of the Activity of Internal Organs. Nobelprize.org. Available at: http://www.nobelprize.org/nobel_prizes/medicine/laureates/1949/hess-lecture.html (accessed 03.10.13.).

Hoebel, B.G., Teitelbaum, P., 1962. Hypothalamic control of feeding and self-stimulation. Science 135 (3501), 375–377. New York.

Ikemoto, S., Panksepp, J., 1999. The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking. Brain Research Brain Research Reviews 31 (1), 6–41.

Jennings, J.H., et al., 2013. Distinct extended amygdala circuits for divergent motivational states. Nature 496 (7444), 224–228.

Kim, K.M., et al., 2012. Optogenetic mimicry of the transient activation of dopamine neurons by natural reward is sufficient for operant reinforcement. PLoS ONE 7 (4), e33612.

Kringelbach, M.L., Berridge, K.C., 2009. Towards a functional neuroanatomy of pleasure and happiness. Trends in Cognitive Sciences 13 (11), 479–487.

Lamb, R.J., et al., 1991. The reinforcing and subjective effects of morphine in post-addicts: a dose-response study. The Journal of Pharmacology and Experimental Therapeutics 259 (3), 1165–1173.

LeDoux, J.E., 1996. The Emotional Brain: The Mysterious Underpinnings of Emotional Life. Simon & Schuster, New York.

Leyton, M., et al., 2007. Mood-elevating effects of d-amphetamine and incentive salience: the effect of acute dopamine precursor depletion. Journal of Psychiatry & Neuroscience: JPN 32 (2), 129–136.

Loewenstein, G., Rick, S., Cohen, J., 2008. Neuroeconomics. Annual Review of Psychology 59, 647–672.

Margules, D.L., Olds, J., 1962. Identical "feeding" and "rewarding" systems in the lateral hypothalamus of rats. Science 135, 374–375. New York.

Miguelez, M., Bielajew, C., 2004. Mapping the neural substrate underlying brain stimulation reward with the behavioral adaptation of double-pulse methods. Reviews in the Neurosciences 15 (1), 47–74.

Milner, P.M., 1989. The discovery of self-stimulation and other stories. Neuroscience & Biobehavioral Reviews 13 (2–3), 61–67.

Moisan, J., Rompré, P.P., 1998. Electrophysiological evidence that a subset of midbrain dopamine neurons integrate the reward signal induced by electrical stimulation of the posterior mesencephalon. Brain Research 786 (1–2), 143–152.

Montague, P.R., Dayan, P., Sejnowski, T.J., 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 16 (5), 1936–1947.

Moruzzi, G., Magoun, H.W., 1949. Brain stem reticular formation and activation of the EEG. Electroencephalography and Clinical Neurophysiology 1 (1–4), 455–473.

Niv, Y., et al., 2007. Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology 191 (3), 507–520.

Olds, J., 1973. Commentary on Olds, J. and Milner, P. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. In: Valenstein, E.S. (Ed.), Brain Stimulation and Motivation; Research and Commentary. Scott Foresman and Company, Glenview, Ill, pp. 81–99.

Olds, J., 1956. Pleasure centers in the brain. Scientific American 195 (4), 105–117.

Olds, J., 1958a. Satiation effects in self-stimulation of the brain. Journal of Comparative and Physiological Psychology 51 (6), 675–678.

Olds, J., 1958b. Self-stimulation of the brain; its use to study local effects of hunger, sex, and drugs. Science 127 (3294), 315–324. New York.

Olds, J., Milner, P., 1954. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. Journal of Comparative and Physiological Psychology 47 (6), 419–427.

Rada, P.V., et al., 2000. Acetylcholine release in ventral tegmental area by hypothalamic self-stimulation, eating, and drinking. Pharmacology, Biochemistry, and Behavior 65 (3), 375–379.

Robinson, T.E., Berridge, K.C., 1993. The neural basis of drug craving: an incentive-sensitization theory of addiction. Brain Research Brain Research Reviews 18 (3), 247–291.

Robinson, T.E., Berridge, K.C., 1998. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Research Brain Research Reviews 28 (3), 309–369.

Routtenberg, A., Lindy, J., 1965. Effects of the availability of rewarding septal and hypothalamic stimulation on bar pressing for food under conditions of deprivation. Journal of Comparative and Physiological Psychology 60 (2), 158–161.

Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science 275 (5306), 1593–1599. New York.

Sem-Jacobsen, C.W., 1959. Depth-electrographic observations in psychotic patients: a system related to emotion and behavior. Acta Psychiatrica Scandinavica. Supplementum 34 (136), 412–416.

Shizgal, P., 1997. Neural basis of utility estimation. Current Opinion in Neurobiology 7 (2), 198–208.

Shizgal, P., 1999. On the neural computation of utility: implications from studies of brain stimulation reward. In: Kahneman, D., Diener, E., Schwarz, N. (Eds.), Well Being: Foundations of Hedonic Psychology. Russell Sage Foundation, New York, pp. 502–526.

Shizgal, P., 2012. Scarce means with alternative uses: Robbins' definition of economics and its extension to the behavioral and neurobiological study of animal decision making. Frontiers in Neuroscience 6, 20.

Shizgal, P., Conover, K., 1996. On the neural computation of utility. Current Directions in Psychological Science 5 (2), 37–43.

Shizgal, P., Murray, B., 1989. Neuronal basis of intracranial self-stimulation. In: Liebman, J.M., Cooper, S.J. (Eds.), The Neuropharmacological Basis of Reward. Oxford University Press, Oxford, pp. 106–163.

Simmons, J.M., Gallistel, C.R., 1994. Saturation of subjective reward magnitude as a function of current and pulse frequency. Behavioral Neuroscience 108 (1), 151–160.

Stellar, E., 1954. The physiology of motivation. Psychological Review 61 (1), 5–22.

Stuber, G.D., et al., 2011. Excitatory transmission from the amygdala to nucleus accumbens facilitates reward seeking. Nature 475, 377–380.

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction, first ed. MIT Press, Cambridge, MA.

Valenstein, E.S., 1973. History of brain stimulation: investigations into the physiology of motivation. In: Valenstein, E.S. (Ed.), Brain Stimulation and Motivation; Research and Commentary. Scott Foresman and Company, Glenview, IL.

Wise, R.A., 1996. Addictive drugs and brain stimulation reward. Annual Review of Neuroscience 19, 319–340.

Wise, R.A., 2008. Dopamine and reward: the anhedonia hypothesis 30 years on. Neurotoxicity Research 14 (2–3), 169–183.

Wise, R.A., 1982. Neuroleptics and operant behavior: the anhedonia hypothesis. The Behavioral and Brain Sciences 5, 39–97.

Wise, R.A., Rompré, P.P., 1989. Brain dopamine and reward. Annual Review of Psychology 40, 191–225.

Witten, I.B., et al., 2011. Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. Neuron 72 (5), 721–733.

Yeomans, J.S., et al., 2000. Brain-stimulation reward thresholds raised by an antisense oligonucleotide for the M5 muscarinic receptor infused near dopamine cells. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 20 (23), 8861–8867.

Yizhar, O., et al., 2011. Optogenetics in neural systems. Neuron 71 (1), 9–34.

Breast-feeding: Effects on Cognitive and Neural Development

Fahmida Tofail and Jena Hamadani, International Centre for Diarrhoeal Disease Research, Bangladesh, Dhaka, Bangladesh

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Breast milk is a complete, natural food for the baby that addresses both growth and brain development of the growing infant through many different pathways. Although research findings from both developed and developing countries reported comparatively smaller benefits of breastfeeding on children's brain development, almost all the studies consistently showed a beneficial trend. Assessing intelligence and brain development is not straight forward and can be influenced by a multitude of factors. In this article, we have reported the association of breastfeeding on neuro-cognition of children under the light of published research findings. The review highlighted that breastfeeding is indisputably important for neuro-cognitive development, along with other health and economic advantages. So, promotion of exclusive breastfeeding up to 6 months and its continuation until 24 months of age, along with appropriate complementary feeding need to be strongly encouraged.

Breast milk is a complete, natural food for the baby. Besides being of nutritional value, it plays a vital role in the neurocognitive development of the growing infant directly as a result of improved nutritional status of the infant and indirectly by providing psychosocial care and child-friendly environment for learning. Several longitudinal and cross-sectional studies have shown consistent beneficial effects of breastfeeding on children's cognitive development and intelligence; however, the effects are small and sometimes not significant. In this article, we report the association of breastfeeding on neurocognition of children under the light of published research findings.

Overall development of a child is widely influenced by factors like health, nutrition, exposure to toxicants, cognitive stimulation, mental state of the caregivers, and exposure to violence (Walker et al., 2011). With great disappointment, it is reported that over 200 million children from low- and middle-income countries of the world are failing to reach their full potentials due to multiple risk factors, e.g., stunting, micronutrient deficiency, and poor stimulation. Breastfeeding was identified as one of the protective factors for child development (Walker et al., 2011). The term 'child development' refers to progression of a number of domains that include physical, cognitive, motor, language, emotional, and behavioral development. None of these domains alone can describe a child. However, in this article we will mainly discuss association of breastfeeding with neurocognitive development of children.

Cognition and Neuronal Development

In general, cognition indicates a complex higher-order mental ability necessary for survival, i.e., understanding and adaptation with one's surrounding world. These abilities include mainly thinking, reasoning, learning how to manipulate the environment, and problem solving (Vasta et al., 2004; Goswami, 1998). Growth of the brain starts *in utero* and most of the brain cells are formed before birth. Development of the brain following birth is a prolonged process and interactions with the environment modify the connections between the brain cells or neurons. During early infancy, massive brain growth involves increase in neuronal connections (synapses), nerve fibers (dendrites), and fiber bundles along with their covering of fatty sheaths (myelination). All these neuronal development contribute to child development. There is a general consensus that by 2 years of age the brain structure simulates the overall appearance of the adult brain and by 3 years all main fiber tracts are visible. However, myelination of nerve fibers continues for a longer period of time and is related with nutrition (Matsuzawa et al., 2001; Paus et al., 2001).

Breastfeeding

The World Health Organization (WHO) and UNICEF recommended exclusive breastfeeding (EBF) up to the age of 6 months and continued breastfeeding along with appropriate complementary foods until 24 months of age (WHO/UNICEF, 2003).

The following rules were recommended by the WHO, as a joint WHO/UNICEF statement for every facility that provides maternity services and care for newborn infants.

Have a written breastfeeding policy that is routinely communicated to all health care staff.

1. Train all health care staff in skills necessary to implement this policy.
2. Inform all pregnant women about the benefits and management of breastfeeding.
3. Help mothers initiate breastfeeding within half an hour of birth.
4. Show mothers how to breastfeed and how to maintain lactation even if they should be separated from their infants.
5. Give newborn infants no food or drink other than breast milk, unless medically indicated.
6. Practice rooming-in – that is, allow mothers and infants to remain together – 24 h a day.
7. Encourage breastfeeding on demand.
8. Give no artificial teats or pacifiers (also called dummies or soothers) to breastfeeding infants.
9. Foster the establishment of breastfeeding support groups and refer mothers to them on discharge from the hospital or clinic.

Breastfeeding not only contributes to better growth and development of children, but also influences the physical and mental health of mothers. Breast milk is considered the best food for young infants because of its nutritional, anti-infective, and antiallergic properties. It also has contraceptive and economic advantages for the mother and family. It enhances mother–child

interactions and reduces maternal stress and depression by providing beneficial hormones. Combination of all these factors lead to improved childcare and facilitates child development. The possible beneficial effect of breastfeeding on the child's neurocognition remains a matter of great scientific interest.

Different Terminologies Used to Define Breastfeeding

In the 1960s and 1970s, 'exclusive breastfeeding' used to mean providing breast milk to the child as the only source of milk with or without foods other than milk.

The current, globally accepted definitions of breastfeeding, developed by the WHO and the UNICEF, are modifications of definitions earlier set by the 'WHO Global Data Bank on Breastfeeding' (WHO, 1996), which focuses on the well-being of the mother–infant pair (Table 1).

Mechanism of Action of Breastfeeding on Cognitive and Neurodevelopment

It is suggested that breastfeeding can affect cognition and neurodevelopment both directly and indirectly. The two main mechanisms and the possible pathways to improved cognition and neurodevelopment as a result of breastfeeding are discussed.

Nutritional Effects

Breast milk is natural first food for newborns and is unique in its composition as it contains all the macro and micronutrients, hormones, and immune components that an infant requires until 6 months of life for optimum growth and development. Thereafter, it continues providing half or more of nutritional requirement of child until the second half of infancy and up to a third during the second year of life.

Breast milk contains fat, protein, carbohydrate, and minerals. The calorie content is about 60–75 kcal/100 mL which is sufficient for the healthy growth of the child's organs including the brain. Casein is its main protein constituent. In addition, it contains alpha-lactalbumin, lactoferrin, immunoglobulin A (IgA), lysozyme, and serum albumin. Colostrum is rich in protein content and provides immunity to the newborn that protects the child from common childhood illnesses, e.g., measles, diarrhea, pneumonia (Wang and Wu, 1996). A healthy child has a greater chance of exploring the environment and learning through interaction with surroundings. Hence, the anti-infective and antiallergic properties of breast milk can indirectly promote neurocognitive development. Breast milk is an important source of essential fatty acids like linoleic acid and alpha-linolenic acid and provides polyunsaturated, long-chain fatty acid (PUFA). PUFA plays a major role in the structural development of retinal photoreceptors, neurones, and synaptic membranes. It carries out certain functions for improving neurogenesis and synaptogenesis like ion transport, receptor activity, enzyme action, and membrane fluidity (Banapurmath et al., 1996). Through the above actions it enhances visual and cognitive development (Mitmesser and Jensen, 2007). In addition, breast milk contains hormones like choline, taurine, and various peptides and polysaccharides. Choline helps in proper communication between nerves and muscles and therefore increases processing speed. Taurine is a membrane

Table 1 Different terminologies of infant feeding

Category of infant feeding	*Definition*
Breastfeeding	Child receives breast milk directly from the breast or expressed breast milk.
Exclusive breastfeeding	The infant receives only breast milk from mother or a wet nurse, or their expressed breast milk, but no other liquids or solids with the exception of drops or syrup formulations of vitamins, mineral supplements, or medicines.
Predominant breastfeeding	Breast milk is the major source of nourishment; however, child also receives water and/or water-based drinks (e.g., sweetened and flavored water and tea); fruit juice; oral rehydration solution (ORS), drop and syrup formulation of vitamins, minerals and medicines, and limited quantities of ritual fluids. With the exception of fruit juice and sugar water, no food-based fluid is allowed.
Full breastfeeding	Exclusive breastfeeding and predominant breastfeeding together constitute full breastfeeding.
Artificial feeding	Infant receives only artificial feeds and no breast milk at all.
Partial breastfeeding	Infant receives some breastfeeds and some artificial feeds, milk, cereal, or other foods.
Complementary feeding	Infant receives both breast milk and timely introduced appropriate semisolid or solid food(s).
Bottle-feeding	Infant receives liquid or semisolid food from a bottle with a nipple/teat.

WHO, 1996. Global Data Bank on Breastfeeding. World Health Organization, Geneva.

stabilizer and antioxidant and is involved in the development of visual and the central nervous system (Reynolds, 2001). Leptin, a hormone in breast milk, through its regulatory acts on hypothalamic–pituitary–adrenal axis, reduces stress in newborns. It is also hypothesized to contribute to learning and memory of young infants through the hippocampal development of brain (Walker et al., 2004). Recently, a small experimental study reported the possible role of a brain-derived neurotrophic factor (BDNF) in breast milk on neurons of central and the peripheral nervous systems. BNDF plays an important role in supporting the survival, growth, and differentiation of neurons and synapses. In the developing brain, it is active in areas vital for learning, memory, and higher thinking ability, particularly in cerebral cortex, hippocampus, and basal forebrain. BNDF also has an association with long-term memory and contributes to neurocognition (Nassar et al., 2011).

Psychosocial Effects

Breastfeeding indirectly promotes neurocognition of children through its effects on maternal well-being and mother–child interaction. The emotional warmth and consistency of the relationship with the mother, particularly the quality of attachment is important for emotional development of the child. Children develop attachment to their mothers during the first 6 months of life. Bowlby (1982) described attachment as the feeling of security that the infant seeks when under stress. A secure child, after a period of separation from its mother, demonstrates lesser resistance and can be easily calmed and comforted upon return of the mother. On the other hand, an insecure child may demonstrate avoidant or ambivalent response. The avoidant child after separation from the mother may not become distressed and does not seek to contact her when she returns. The ambivalent child wants to contact the mother but remains upset even after being reunited following a period of separation (Benoit, 2004). As the mother interacts closely with the child while breastfeeding, the child develops a secure attachment helping her/his emotional development and allows explorative behavior for learning (Gordon, 1998). Lactating mothers have higher levels of prolactin and oxytocin compared to nonlactating mothers. In animal models, these hormones have been associated with feelings of calmness, well-being, and nurturing behavior (Walker et al., 2004). Thus, benefits to mothers who breastfeed include reduced postpartum depression, workload, reduced chance to become pregnant, and increased emotional bonding. Breastfeeding also contributes to physical well-being of mothers by reducing the risk of ovarian and breast cancers. This, in turn, facilitates positive mother–child interactions leading to better childcare and child development.

Research Findings Related to Breastfeeding and Neurocognition

A good number of studies, mostly from developed countries, have examined the link between breastfeeding and child development. Here, we present important findings from relevant reviews and meta-analyses reported till 2011 and independent reports published after 2011.

In 1999, Grantham-McGregor et al. (2007) documented 22 articles published during 1929–98, which examined the developmental advantages of breast-fed over bottlefeed children. The studies demonstrated a small but consistent benefit of breastfeeding on children's development. Among older children aged 18–24 months, there were significant developmental advantages from breastfeeding even after adjustment for socioeconomic variables, particularly if they were exclusively breast-fed for at least 3 months. Studies that examined the long-term developmental effects in 4- to 18-year-old children observed consistent developmental benefits in breast-fed children but the effect quantum depended on other factors such as size at birth, poverty, and duration of breastfeeding. Interestingly, there was a detrimental effect of EBF for longer than 12 months, but not when breastfeeding was continued along with complementary feeds.

A meta-analysis on articles published between 1966 and 1996 found that even after adjusting for probable covariates, breastfeeding was significantly associated with higher cognitive score compared to formula feeding (Anderson et al., 1999).

Another review on studies published between 1966 and 1998 could not draw any firm conclusion on the cognitive benefits of breastfeeding although the trend of higher score was still in favor of breastfeeding (Drane and Logemann, 2000).

Jain et al. (2002) reviewed articles published between 1929 and 2001 and concluded that breastfeeding promotes intelligence.

Another review article by Rey in 2003 discussed breast feeding related cognitive findings from earlier retrospective studies, large birth cohorts, meta-analysis, critical reviews, and individual studies on full-term infants and low birth weight infants. The review highlighted a wide range of disparities among the studies and concluded that the beneficial effects of breast milk on infants' cognitive development could not be clearly recognized from these studies.

A meta-analysis by Der et al. (2006) on the studies conducted in developed countries on full-term breast-fed babies adjusting for maternal IQ and other socioeconomic factors did not find significant benefit of breastfeeding on children's cognition.

The latest *Lancet* series on child development however reported strong evidence on the benefits of breastfeeding on children's development and school achievement (Walker et al., 2011).

Since the above article that reviewed publications till 2010, we could locate seven papers published on the relationship of breastfeeding to neurocognitive development of children born at full term.

Out of these seven studies, only one was cross sectional, conducted among Irish school children at the age of 9 years using the propensity score matching (PSM) analysis (McCrory and Layte, 2011). The outcome measure was 'academic test scores' and information on breastfeeding was based on parental recall. Although the study showed a robust benefit, the finding can be confounded by recall bias.

Among the other six cohort studies, one was from a middle-income country that prospectively followed the Krakow birth

Table 2 Studies on breastfeeding and cognition in the years 2011 and 2012

Study, type, and location	*Study population and age*	*Developmental assessment tool*	*Result for adjusted models*
McCrory and Layte, cross-sectional study, Ireland (2011)	8226 school children at 9 years	Children's academic test scores	Breast-fed children had 3.24 ($p < .001$) and 2.23 ($p < .001$) percentage points higher scores on reading and mathematics, respectively, compared to those who were never breast-fed
Paterson et al., cohort of Pacific infants, New Zealand (2011)	1018 children, 2 years	Australian developmental screening test	35% had developmental delay. Breastfeeding <6 weeks was identified as one of the risk factors for developmental delay
Guxens et al., population-based cohort study, Spain (2011)	504 children assessed at 14 months	Bayley scale of infant development	Cumulative intensity of EBF during the first year of life was linearly and positively associated with 0.37 points; higher mental development score duration of EBF was also associated with higher mental development score of 0.37 points
Boutwell et al., early childhood longitudinal study, birth cohort, UK (2012)	3700 children (2700 breast-fed; 1000 nonbreast-fed) assessed at 9 months and 2 and 4 years	Bayley Short Form-Research Edition (BSF-R) mental scale	Significant difference between groups (mean difference = 1.92, $p < .05$)
Quigley et al., United Kingdom millennium cohort study, UK (2012)	11 101 term babies born at 37–42 weeks of gestational age assessed at 5 years	British ability scales tests (naming vocabulary, pattern construction, and picture similarities subscales)	A two-point increase in score for picture similarities (when breast-fed ≥4 months) and naming vocabulary (when breast-fed ≥6 months)
Jedrychowski et al., Krakow prospective birth cohort study, Poland (2012)	468 term babies (>36 weeks gestational age) assessed at 1, 2, 3, 6, and 7 years	Bayley scale of infant development up to 3 years age. Wechsler intelligence test (WISC-R) for 6 and 7 years	EBF for 3 months had 2.1 points; EBF for 4–6 months had 2.6 points; and EBF >6 months had 3.8 points; higher IQ scores compared to the others
Borra et al., Avon Longitudinal Survey of Parents and Children (ALSPAC study), UK (2012)	12 268 children assessed at 5, 7, 11, and 14 years	Key Stage 1 Battery: language, reading, reading comprehension, writing, spelling, and mathematics. Key Stages 2 and 3: English, mathematics, and science	Breastfeeding is associated with higher tests scores, about one-tenth of standard deviation

cohort in Poland at 1, 2, 3, 6, and 7 years of age. Developmental assessment was done using the Bayley Scale of infant development (BSID) for the younger group and the Wechsler intelligence scale for children-revised (WISC-R) for the older age group. After controlling for a large number of confounders, including maternal IQ, the effects of breastfeeding on cognitive development of children was sustained through 7 years of age (Jedrychowski et al., 2012).

The remaining five studies were all from developed countries: three from United Kingdom, one from Spain, and one from New Zealand (**Table 2**). Paterson et al. (2011) conducted an observational study on a cohort of children at 2 years of age. He found a large percentage (35%) of children to be developmentally delayed in one of the five domains of Australian Developmental Screening Test. It was noted that breastfeeding for less than 6 weeks was one of the risk factors for developmental delay among Pacific children. In a nonexperimental study using Bayley Short Form-Research Edition (BSF-R) mental scale on children at 9 months and 2 and 4 years, the cognitive function of breast-fed children was significantly higher than that of nonbreast-fed children by PSM analysis (Boutwell et al., 2012). A birth cohort follow-up in UK showed significant benefit in breast-fed children compared to those who were never breast-fed after adjusting for all possible confounders (Quigley et al., 2012). Breastfeeding was shown to be beneficial in Spanish children tested at 14 months using the Bayley scales (Guxens et al., 2011). In a large longitudinal cohort, the ALSPAC study, benefit of only 0.1 SD was found in children's cognitive development assessed at 5, 7, 11, and 14 years using a battery of English reading and writing, mathematics, and science (Borra et al., 2012).

Despite wide variations in the design, population, site, test measures, and confounders in the above-mentioned studies, all of them consistently show a trend of positive association of breastfeeding with neurocognition.

Conclusion

Breastfeeding is very important for providing ideal food for proper growth and neurocognitive development of young infants. Besides nutritional value, breast milk is a rich source of hormones, enzymes, and antibodies that are protective for the

child's health and development. Breastfeeding also contributes to mother–child interaction and to effective parenting by reducing maternal depression and improving maternal well-being. This helps in efficient bonding that allows enhanced neuronal connections.

Although the results from both developed and developing countries reported comparatively smaller benefits of breastfeeding on child development, all the studies consistently showed a beneficial trend. It is possible that the use of different assessment tools, their adaptation in different countries, adjusting for confounders, wide variations in the design, small sample size in some studies, operational definitions of breastfeeding, and age of assessment complicated the interpretations in many ways. Moreover, long-term consequences were seen in very few studies. Longitudinal studies are important as some of the developmental domains become measurable in later ages.

Moreover, intelligence can be influenced by a multitude of factors and it is not easy to control for those factors while analyzing data. Although several of the recent studies made efforts to control a number of known confounding factors, there could be other, as yet unidentified factors that could influence the outcome. It is also unethical to conduct randomized trials on breastfeeding and thus prospective, observational studies are the possible choices to determine the impact of breastfeeding on cognitive development of children. In view of the trends in benefits of EBF on cognition, and other advantages of breastfeeding including its cost-effectiveness, nutrient value, protective properties from infections and well-being of both mother–child dyads, promotion of EBF up to 6 months and its continuation until 24 months of age, along with appropriate complementary feeding, need to be done in developing countries.

See also: Neonatal Bonding; Neonatal Social Isolation: Effects on Neural Function.

Bibliography

Anderson, J.W., Johnstone, B.M., Remley, D.T., 1999. Breast-feeding and cognitive development: a meta-analysis. American Journal of Clinical Nutrition 70 (4), 525–535.

Banapurmath, C.R., Banapurmath, S., Kesaree, N., 1996. Developing brain and breastfeeding. Indian Pediatrics 33 (1), 35–38.

Benoit, D., 2004. Infant-parent attachment: definition, types, antecedents, measurement and outcome. Paediatric Child Health 9 (8), 541–545.

Borra, C., Iacovou, M., Sevilla, A., 2012. The Effect of Breastfeeding on Children's Cognitive and Noncognitive Development. Institute for the Study of Labor (IZA). IZA Discussion Papers 6697.

Boutwell, B.B., Beaver, K.M., Barnes, J.C., 2012. Role of breastfeeding in childhood cognitive development: a propensity score matching analysis. Journal of Paediatrics and Child Health 48 (9), 840–845.

Bowlby, J., 1982. Attachment and Loss, second ed., vol. 1. Basic Books, New York.

Der, G., Batty, G.D., Deary, I.J., 2006. Effect of breast feeding on intelligence in children: prospective study, sibling pairs analysis, and meta-analysis. BMJ 333 (7575), 945. Epub 2006 Oct 4.

Drane, D.L., Logemann, J.A., 2000. A critical evaluation of the evidence on the association between type of infant feeding and cognitive development. Paediatric and Perinatal Epidemiology 14 (4), 349–356.

Gordon, N., 1998. Some influences on cognition in early life: a short review of recent opinions. European Journal of Paediatric Neurology 2 (1), 1–5.

Goswami, U., 1998. Cognition in Children. Psychology Press, London.

Grantham-McGregor, S., Cheung, Y.B., Cueto, S., Glewwe, P., Richter, L., Strupp, B., International Child Development Steering Group, 2007. Developmental potential in the first 5 years for children in developing countries. Lancet 369 (9555), 60–70.

Guxens, M., Mendez, M.A., Moltó-Puigmartí, C., Julvez, J., García-Esteban, R., Forns, J., Ferrer, M., Vrijheid, M., López-Sabater, M.C., Sunyer, J., 2011. Breastfeeding, long-chain polyunsaturated fatty acids in colostrum, and infant mental development. Pediatrics 128 (4), e880–e889.

Jain, A., Concato, J., Leventhal, J.M., 2002. How good is the evidence linking breastfeeding and intelligence? Pediatrics 109 (6), 1044–1053.

Jedrychowski, W., Perera, F., Jankowski, J., Butscher, M., Mroz, E., Flak, E., Kaim, I., Lisowska-Miszczyk, I., Skarupa, A., Sowa, A., 2012. Effect of exclusive breastfeeding on the development of children's cognitive function in the Krakow prospective birth cohort study. European Journal of Pediatrics 171 (1), 151–158.

Matsuzawa, J., Matsui, M., Konishi, T., Noguchi, K., Gur, R.C., Bilker, W., Miyawaki, T., 2001. Age-related volumetric changes of brain gray and white matter in healthy infants and children. Cerebral Cortex 11 (4), 335–342.

McCrory, C., Layte, R., 2011. The effect of breastfeeding on children's educational test scores at nine years of age: results of an Irish cohort study. Social Science & Medicine 72 (9), 1515–1521.

Mitmesser, S.H., Jensen, C.L., 2007. Roles of long-chain polyunsaturated fatty acids in the term infant: developmental benefits. Neonatal Network 26 (4), 229–234.

Nassar, M.F., Younis, N.T., El-Arab, S.E., Fawzi, F.A., 2011. Neuro-developmental outcome and brain-derived neurotrophic factor level in relation to feeding practice in early infancy. Maternal & Child Nutrition 7 (2), 188–197.

Paterson, J., Iusitini, L., Gao, W., 2011. Child developmental assessment at two-years of age: data from the Pacific Islands Families Study. Pacific Health Dialog 17 (2), 51–63.

Paus, T., Collins, D.L., Evans, A.C., Leonard, G., Pike, B., Zijdenbos, A., 2001. Maturation of white matter in the human brain: a review of magnetic resonance studies. Brain Research Bulletin 54 (3), 255–266.

Quigley, M.A., Hockley, C., Carson, C., Kelly, Y., Renfrew, M.J., Sacker, A., 2012. Breastfeeding is associated with improved child cognitive development: a population-based cohort study. Journal of Pediatrics 160 (1), 25–32.

Rey, J., 2003. Breastfeeding and cognitive development. Acta Paediatrica Supplement 92 (442), 11–18.

Reynolds, A., 2001. Breastfeeding and brain development. Pediatric Clinics of North America 48 (1), 159–171.

Vasta, R., Miller, S.A., Ellis, S. (Eds.), 2004. Child Psychology, fourth ed. John Wiley & Sons, Inc., Hoboken, NJ.

Walker, C.D., Deschamps, S., Proulx, K., Tu, M., Salzman, C., Woodside, B., Lupien, S., Gallo-Payet, N., Richard, D., 2004. Mother to infant or infant to mother? Reciprocal regulation of responsiveness to stress in rodents and the implications for humans. Journal of Psychiatry & Neuroscience 29 (5), 364–382.

Walker, S.P., Wachs, T.D., Grantham-McGregor, S., Black, M.M., Nelson, C.A., Huffman, S.L., Baker-Henningham, H., Chang, S.M., Hamadani, J.D., Lozoff, B., Gardner, J.M., Powell, C.A., Rahman, A., Richter, L., 2011. Inequality in early childhood: risk and protective factors for early child development. Lancet 378 (9799), 1325–1338.

Wang, Y.S., Wu, S.Y., 1996. The effect of exclusive breastfeeding on development and incidence of infection in infants. Journal of Human Lactation 12 (1), 27–30.

WHO, 1996. Global Data Bank on Breastfeeding. World Health Organization, Geneva.

WHO/UNICEF, 2003. Global Strategy for Infant and Young Child Feeding. World Health Organization, Geneva, Switzerland.

British Cultural Studies

Graeme Turner, University of Queensland, Brisbane, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 2, pp. 1362–1365, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

One of the markers of contemporary studies in the humanities and social sciences is the close, critical attention directed onto the mass media and popular culture. British cultural studies, emerging out of the expanded English university system in the 1970s, were among the first to develop and broadly disseminate the methodologies and approaches which made such attention possible. Focusing initially on the relations between language and ideology through the close analysis of media texts, before diversifying into a broad range of interdisciplinary activities – cultural theory, cultural history, ethnography, media studies, and the analysis of the social construction of gender, race and identity – British cultural studies has played a fundamental role in establishing the conceptual and theoretical agenda for studies of popular cultural forms and practices.

The Analysis of Popular Culture

Academic disciplines in the humanities and social sciences took some time to accept the need to understand properly the power and influence of the mass media in Western societies. The spread of a transnational commercial popular culture, apparently disconnected from specific local cultures and disseminated through advertising and marketing, was seen by many in the academy as a threat to traditional or civilized values which should simply be repudiated. Within the British university system until the end of the 1960s, the task of analyzing and evaluating popular culture had been taken on by the discipline of English. However, the methods supplied by training in literary studies were suited for making esthetic and moral distinctions that proved inappropriate as a means of understanding the appeal of popular cultural forms such as film, television, or popular music. It was gradually acknowledged that popular cultural forms served a very different social function for their audiences than that served by high art for its audiences. Cultural studies was a product of the consequent recognition of the necessity of finding new ways of understanding the specific social functions of popular culture. Although it was partly a product of literary studies, the methods it developed had to be very different.

The starting points for this project are usually seen (Turner, 1996) to be Richard Hoggart's account of the clash between an 'organic' English working class culture and mass mediated popular culture, *The Uses of Literacy* (1958), and Raymond Williams' history of cultural change in an industrializing England, *Culture and Society 1780–1950* (1958). The debates generated by these books set the conditions within which Hoggart established the Centre for Contemporary Cultural Studies (CCCS) at the University of Birmingham in 1964. In 1968, he was succeeded as Director of the CCCS by Stuart Hall, who played a major role in developing methods of analysis for popular cultural forms over the next two decades. In particular, Hall's work during the 1970s provided a very important conduit through which theoretical developments from France, by Barthes around post-Saussurean semiotics and by Althusser around the workings of ideology (Hall, 1977), reached an English readership.

Initially, the analyses of popular culture which emerged from this tradition were highly political critiques of the ways ideology was embedded in language. Close analysis of media 'texts' (specific examples of media practice exposed to detailed study) was employed as means of explaining how social understandings were carried and reinforced by the mass media, and how these social understandings were structurally inclined to be those supported by the state or the interests of market capitalism (Hall et al., 1978). However, by the latter half of the 1980s, British cultural studies had distanced itself from the notion that the media was simply a means of reinforcing the ideological interests of capital or the state. Instead, it had acknowledged the alternative possibilities suggested by such writers as Dick Hebdige (1979) during the late 1970s and began to investigate how popular cultural forms might also serve the interests of those who consume them. This produced a more positive reading of popular culture, and a more nuanced and contingent set of expectations about the specific interests served by popular cultural forms at particular historical conjunctures (Fiske, 1987). Among the results of this focus over the 1980s was a significant expansion of the field, as a broad range of studies of television, popular music, the print media, fashion, media audiences, and youth subcultures were published, and as cultural studies established itself at undergraduate and graduate level in universities in the UK, the USA, Canada, Australia, and elsewhere. As the various forms of cultural studies proliferated (borrowing methods from history and anthropology as well as literary studies and sociology), the relative distinctiveness of the British tradition began to dissipate.

By the early 1990s, this reorientation of the field and modes of analysis had exposed cultural studies to the accusation of 'cultural populism' (McGuigan, 1992): of complicity with the media and cultural industries that reduced its capacity to provide a critical analysis of the interests these industries served. This provoked a further reassessment of the study of popular culture which served to reassert the complicated set of understandings required: to understand how popular cultural

forms worked as communication; to understand the competing political potentials implicit in the consumption of popular culture; and to understand how the political economy of the media and cultural industries operates as a determining frame around the relations between producers and consumers.

Crucial to the development of the analysis of popular culture within this tradition has been the focus on the media product as a 'text,' and the complementary insistence on the importance of understanding the processes through which these texts are consumed by audiences.

Texts and Audiences

The idea of treating media products as 'texts' comes from literary studies but the early means of doing so came from semiotics (see, for example, *Semiotics*). Semiotics was especially useful for two reasons. First, while it was interpretative as well as descriptive, it did not produce an aesthetic or moral evaluation of the text and was thus differentiated from the only other discipline where textual analysis was highly developed – literary studies. Second, it was the first method which offered the means for analyzing all forms of communication – written, verbal, and visual, either singly or in conjunction with other forms. As a result, semiotics enabled the detailed and comparatively objective analysis of popular cultural products such as advertising, television news, and newspaper layouts.

For most of the 1970s and much of the 1980s, British cultural studies busied itself with developing modes of textual analysis from a combination of semiotics, certain aspects of film theory, and Althusserian theories of ideology. The interest in the text was initially in critically unraveling its carriage of ideology. This slowly gave way to an attention that was more intrinsically interested in the text or genre of texts under examination – the appeal of soap opera, for example, or the construction of the news. Such work initially assumed that the readings generated by the cultural studies critic were immanent in the text, repeating an elitist assumption also made within literary studies that the critic had a privileged insight into what the 'true meanings' of popular culture might be.

The flaws behind this assumption were addressed by a number of interventions, the most significant being Hall's argument that the process of encoding or producing the text was separate and different from the process of decoding or 'reading' the text (1980). Indeed, he argued that there were at least three kinds of reading positions consumers might take when decoding the text. Consumers might simply accept what was described as the 'preferred' meaning; that is, the meaning apparently preferred by the producers or senders of the message or what is often referred to in the literature as the 'dominant' meaning. They might take an 'oppositional' position, which flatly rejects or subverts what they nevertheless would recognize as the preferred reading in order to construct their own, alternative reading of the message. Or they might produce a 'negotiated' reading, developed out of a relatively complicated consideration of the preferred reading and its alternatives. Hall's schema emphasized the contingency of the act of consumption, the varying ways in which audiences might understand any text. In practice, it was accepted, most would develop a negotiated reading rather than either of the other two options. This also emphasized how individualized the processes might be through which culture constructs its meanings. Consequently, cultural studies began to focus on these broader cultural processes rather than the processes of the texts themselves; this in turn was a shift away from the opinion of the critic and towards a serious consideration of the evidence from the audience.

Research into audiences' understanding of texts had occurred within the CCCS during the late 1970s, and the Centre's interest in subcultures had also produced a number of studies which used the methods of ethnography as a means of investigating how subcultural groups made sense of their social experiences (Hobson, 1980; Willis, 1977). However, it was not until Ien Ang's innovative analysis of fans of the US soap opera, *Dallas* (1985), that audience studies began the process of becoming a fundamental component of the practice of cultural studies. Ang's study used the letters from a group of *Dallas* fans both as texts to be analyzed and as ethnographic evidence which helped substantiate the nature of the soap opera's appeal to, for instance, intelligent and educated women who, at that time, would not have been reckoned to belong to the target audience for that kind of television. Something of an explosion of audience studies occurred, usually claiming to be in some way 'ethnographic' (that is, in this context, involving some direct observation of the group under analysis). The appropriation of what was a much reduced version of their research methodology provoked stern critiques from social science disciplines where the established protocols of ethnography were taken more seriously and employed more rigorously.

While the trend for what one might call 'thin' ethnography eventually waned, the usefulness of ethnography as a means of examining not only media audiences but other cultural formations has been accepted in cultural studies. David Morley (1986, 1992) moved audience studies away from largely providing a complementary strategy for those trying to understand the workings of media texts, and situated it as a means of, for instance, understanding the function of television within the home, and its implication in the web of social relations involving age and gender that made up families' everyday lives (*see* Audiences, Media).

Representation and Identity

While the above account might imply that cultural studies' interest in the text has been displaced by the focus on audiences, there is at least one area where that is not true. Textual analysis in cultural studies was initially almost an end in itself; the development of a process which could unpack the construction of meaning within, particularly, media texts was rightly seen as a significant methodological advance. As a result, for quite some time, the performance of textual analysis was a dominant genre of research and writing within the literature of the field. As the authority of the critic was challenged by audience studies, however, the point and practice of textual analysis had to change.

Cultural studies had always been interested in how the world was 're-presented' to us in other forms: as images, as sounds, as writing, and as narrative. The influence of post-structuralist theories of language and culture situated representation as our only access to the real: we understood the world through languages, and through the ways in which it is represented to us. The problem of culture for many, therefore, is overwhelmingly the problem of representation.

One of the shifts in approach emerging from the mid-1980s interest in the audience was a focus on how cultural identity was represented. Feminist cultural studies was the primary instigator here, with its emphasis on how women's lives were determined and constrained by the dominant representational patterns within which the feminine is constructed (Franklin et al., 1991). The interest in identity as a contingent product of representation, and of the need for minority groups to therefore regard their representation as a political issue, becomes a major concern for cultural studies theory and a major genre of textual analysis.

It is also probably the location where the two complementary streams of British cultural studies – on the production and textual side, and on the consumption or audience side – have been most effectively and productively combined (Hall, 1997). The focus on the representation of particular minority groups, and the interrogation of the effect of that representation upon particular examples of those groups, is precisely the territory that feminist cultural studies has developed. It has also been taken up by other groups – gays and lesbians, ethnic and racial minorities, diasporean communities – as a means of understanding the construction of their identities, and of contesting the political effects of that construction through a critique of the representational patterns involved. Finally, in the 1990s, the focus on representation has assisted in opening up the issue of the construction of national identity for British cultural studies.

The Politics of British Cultural Studies

One of the ways in which British cultural studies have conceptualized their project is to see it as inherently political. That is, British cultural studies set out to understand how culture works in order to intervene in that process and produce change. Williams, Hall, and other key figures were identified with New Left Marxism, and had played the roles of activitists within British politics. The process of developing the field, the work of 'theoretical clarification,' as it was called, has routinely addressed the problem of how to make academic knowledge matter within contemporary party politics. Indeed, when Hall (1988) addressed the problem of the Left's incapacity to combat the working class alignment with (then) Prime Minister Margaret Thatcher's policies, he presented it as an issue both for party policy and cultural theory.

The centrality of politics to the British cultural studies project is among the factors which may still differentiate it from those more recent cultural studies traditions which have developed, for instance, in the USA. Indeed, among the concerns expressed by Hall and others about the success British cultural studies achieved in exporting itself to the USA at the end of the 1980s was the likely displacement of politics from the theoretical field. The widespread pattern of institutionalization of cultural studies, in the UK and elsewhere, has also been proposed as a threat to this fundamental principle. Similarly, if a little paradoxically given its direct engagement with political institutions, the incorporation of cultural studies academics into the process of state policy formation, particularly in the UK and Australia, and the consequent interest in the related field of cultural policy studies (Bennett, 1998), has been criticized as a form of co-option.

The politics of British cultural studies itself is not without its blind spots, of course. It was reluctant to admit feminist concerns and paid little attention to issues of race, ethnicity or nationality until the 1990s. Its resistance to empirical research methods has exposed it to criticism from within the social sciences. However, its determination to make a highly theoretical enterprise prove itself in the material world of practical politics continues to mark it out as a distinctive intellectual and social project.

Editor's Note

As Turner's original entry makes clear, cultural studies is a field of critical thinking and literary criticism that emerged in the UK in the late 60s and early 70s and is focused largely on mass media and popular culture. Although work in the "cultural studies" tradition now goes on around the world, the British tradition remains an intellectually vibrant area of study. Contemporary works in this tradition would include: Brantlinger, 2013[1999]; Jin, 2011; Hughson, 2011. Turner's essay also notes that cultural studies was "reluctant to admit feminist concerns." Today, there is an identifiable literature of feminist cultural studies; see, e.g., Clarke and Olesen, 2013 [1999]; Torres 2013.

See also: Cultural Geography; Cultural History; Cultural Policy Regimes in Western Europe; Cultural Psychology; Cultural Studies of Science; Culture, Sociology of; Leisure Activities Choices among Adolescents; Mass Media and Cultural Identity; Popular Culture; Television: General.

Bibliography

Ang, I., 1985. Watching Dallas: Soap Opera and the Melodramatic Imagination. Methuen, London.

Bennett, T., 1998. Culture: A Reformer's Science. Allen and Unwin, Sydney, Australia.

Brantlinger, P., 2013[1999]. Crusoe's Footprints: Cultural Studies in Britain and America. Chapman and Hall, Routledge, Oxford.

Clarke, A.E., Olesen, V. (Eds.), 2013[1999]. Revisioning women, health and healing: Feminist, cultural and technoscience perspectives. Taylor and Francis Group, New York.

Fiske, J., 1987. Television Culture. Methuen, London.

Franklin, S., Lury, C., Stacey, J., 1991. Off-Centre: Feminism and Cultural Studies. Harper Collins, London.

Hall, S., Critcher, C., Jefferson, T., Clarke, J., Robinson, B., 1978. Policing the Crisis: Mugging, the State, and Law and Order. Macmillan, London.

Hall, S., 1977. Culture, the media and the 'ideological effect. In: Curran, J., Gurevitch, M., Woollacott, J. (Eds.), Mass Communication and Society. Edward Arnold, London, pp. 315–348.

Hall, S., 1980. Encoding/decoding. In: Hall, S., Hobson, D., Lowe, A., Willis, P. (Eds.), Culture, Media, Language. Hutchinson, London, pp. 128–138.

Hall, S., 1988. The Hard Road to Renewal: Thatcherism and the Crisis of the Left. Verso, London.
Hall, S. (Ed.), 1997. Representation: Cultural representations and Signifying Practices. Sage, London.
Hebdige, D., 1979. Subculture: The Meaning of Style. Methuen, London.
Hobson, D., 1980. Crossroads: The Drama of a Soap Opera. Methuen, London.
Hoggart, R., 1958. The Uses of Literacy. Penguin, London.
Hughson, J., 2011. Sport and the British, and British Cultural Studies. Sport in History 31 (2), 163–166.
Jin, H., 2011. British cultural studies, active audiences, and the status of cultural theory: an interview with David Morley. Theory, Culture and Society 28 (4), 124–144.
McGuigan, J., 1992. Cultural Populism. Routledge, London.
Morley, D., 1986. Family Television: Cultural Power and Domestic Leisure. Comedia, London.
Morley, D., 1992. Television, Audiences and Cultural Studies. Routledge, London.
Torres, E.E., 2013. Chicana without Apology: The New Chicana Cultural Studies. Routledge, Oxford.
Turner, G., 1996. British Cultural Studies: An Introduction. Routledge, London.
Williams, R., 1958. Culture and Society 1780–1950. Penguin, London.
Willis, P.E., 1977. Learning to Labour: How Working Class Kids Get Working Class Jobs. Saxon House, Farnborough, UK.

Broadbent, Donald Eric (1926–93)

Patrick Rabbitt, University of Oxford Department of Experimental Psychology, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Donald Broadbent's World War II experiences of deficiencies in engineering translated into a lifelong commitment to what was then the emerging field of Applied Psychology. After studying under Frederick Bartlett at Cambridge, Broadbent went on to an appointment with the UK Medical Research Council Applied Psychology Research unit at Cambridge and subsequently to an MRC appointment at Oxford. Throughout his career he addressed, with colleagues, a range of related topics including auditory attention, the effects of fatigue and noise as well as extroversion and introversion on attention, the roles of errors and lapses in production systems, and information processing in management systems. Perhaps his greatest achievement is reflected across these works in the respect that he garnered for the value of applied psychology.

Donald Broadbent described a boyhood troubled by sharply fluctuating family fortunes, the early loss of a father whose talents greatly exceeded his educational and social opportunities, and the determination of his mother that Broadbent should transcend similar problems by gaining a scholarship to a prestigious public school, Winchester. There he found the physical sciences dispiriting because of the volume of existing knowledge to be mastered before any original work was possible. He was called up to the Royal Air Force, where pilot training in the USA forced recognition that poor engineering design causes catastrophic human errors. Also that, in a far richer society than postwar Britain, these problems were being addressed by an emerging profession of Applied Experimental Psychology to which he, too, might aspire. A brief wartime course in engineering had arguably provided the skills and mind-set that developed his understanding and empathy with this particular slant on psychology. So, when he resumed his interrupted undergraduate career at Pembroke College, Cambridge, against the strong advice of his tutors he read psychology in Sir Frederick Bartlett's Department of Experimental Psychology.

This was then, arguably, the best possible place in the world for him. Bartlett, and the talented staff he had appointed, had distinguished themselves in applied human factors research and personnel selection during the war, and so gained respect from other scientists for what was then, misunderstood by Government circles as an insignificant pseudoscience. The Cambridge Department was uniquely in the mainstream of new ideas. Broadbent has described it as intellectually 'explosive,' similar to that in the Cavendish laboratories and the 'Bun Shop' pub just across Downing Street, where DNA was being discussed. Among the volatile ingredients were the immediate local legacy of Kenneth Craik's seminal applications of control theory to human performance, William Hick's use of Shannon and Weaver's new information theory to model decision-making and his interpretation of its use in the recent wartime work of Egan, Garner, and Pierce; the beginnings of speculative computational neuropsychology in Hebb's (1949) "Organisation of Behaviour," and demonstrations by Grey Walter and Ross Ashby that automata theory could simulate and explain the behavior of simple organisms. After gaining a first-class degree, Broadbent had to explain to Bartlett that he could not afford to follow his advice to support himself independently to work on any problem that interested him. So he was appointed as a member of scientific staff at the Medical Research Council Applied Psychology Research Unit (APRU) in Cambridge, which Bartlett had founded and Norman Mackworth then directed. The APRU was funded to solve applied problems for government agencies, principally the Royal Navy, who determined the research agenda. Mackworth asked Broadbent to work on noise, both as an obstacle to effective speech communication and as a factor in attentional distraction, arousal, and stress. Noise became a main theme of Broadbent's research, from his first paper, "Some effects of noise on visual performance" in 1951, to a concluding review, "Human Performance and Noise," published with Dylan Jones just before Broadbent's retirement 40 years later. It brought an exceptionally congenial and fruitful early collaboration with Peter Ladefoged on speech perception and a lifelong interest in auditory psychophysics and in the subjective and objective measurement of the effects of noise and other stressors on human comfort and performance. Most importantly, it provided applied problems that cued Broadbent's seminal work on auditory multitasking and dichotic listening, which crystallized in his pivotal achievement, his book *Perception and Communication* (Broadbent, 1958).

Modern readers can immediately appreciate that the remarkable impact of this book came partly from a wonderfully lucid style and inspired choice of metaphors that clarified difficult arguments. It is no longer possible to convey its fresh delight for his contemporaries. For all outside a small circle in Cambridge this was the first insight into how an entirely new style of conceptual models, derived loosely from information flow diagrams used by computer scientists, could make sense of exciting new discoveries in momentary and sustained selective attention, short-term memory, arousal, and the effects of stress. The subtext was that the dominant conceptual framework of S–R (stimulus–response) connectionism could not possibly account for these new results. Broadbent's next book, *Behaviour* (Broadbent, 1961) illustrates his recognition of the power of the zeitgeist that *Perception and Communication* had to exorcise. It showed how the behaviorist framework derived from animal learning experiments had become internally inconsistent and that much of the data it had engendered could now be better

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61015-2

interpreted in terms of simple cybernetic models such as those currently being suggested by scientists such as Tony Deutsch and Stuart Sutherland.

In 1958, the year of publication of *Perception and Communication*, Donald Broadbent was appointed director of the APRU and began to attack energetically the hard tasks he inherited, and the even harder ones that he set himself. The funding base of he Unit had to be widened beyond insecure reliance on Royal Navy contracts. Civilian clients, such as the Post Office, the Department of Transport, British Rail, and the Royal Mint were, arduously, brought to recognize practical benefits from human factors research. Part of a self-imposed task of improving public understanding and support in the UK for a misapprehended subject were frequent appearances on radio and television. Effective representation of human experimental psychology through and within the British Association for the Advancement of Science, work with committees and boards of the MRC and the other, newer, Research Councils enlarged academic respect, and secured better funding for psychology in Britain. Similar efforts in the British Psychological Society and the Experimental Psychology Society brought him presidency of both bodies – and membership of even more committees. There was also the day-to-day administration of the rapidly expanding, and increasingly demanding, APRU.

Broadbent always insisted that he could not have continued his personal research without the help of Margaret Gregory, who began to be his collaborator in 1961, and became his wife in 1972. Apart from a diversity of papers on auditory and visual selective attention, decision processes, and stress, Donald Broadbent and Margaret's main work during the 1960s was reinterpretation of data from vigilance tasks in terms of the new metric of Tanner and Swets' signal detection theory. They found that small shifts in operator bias, which could be related neatly to maintenance of arousal and motivation, brought about counterintuitively large changes in patterns of errors. This suggested fresh ways of analyzing the effects on sustained attention of stressors such as fatigue and noise, and of personality variables such as introversion and extraversion. It also led to speculations on distinctions between selective perceptual sensitization for particular signals and preparatory bias for particular responses that became central to his next book, the monumental *Decision and Stress* (Broadbent, 1971), and to much of his later research. However, at the APRU, administrative loads and personal problems became overwhelming. A sabbatical at All Souls College, Oxford allowed personal reassessment and completion of "Decision and Stress," which appeared in 1971. Perhaps election to the Royal Society also brought new confidence in public esteem for his individual contributions, independently of the background successes of the Unit that he had built up. He gave 3 years notice of resignation and in 1974, moved to the Department of Experimental Psychology in Oxford, which was then emerging into a period of preeminence in Britain under the leadership of Larry Weiskrantz. Here, until his retirement in 1991, he at last had his chance to follow Bartlett's advice to work full time on whatever problems he wished.

Decision and Stress immediately achieved widespread respect for its scholarship and its message. Its core is a model for the way information is processed, remembered, and retrieved, expressed metaphorically as a distinction between 'pigeonholing' and 'categorization.' At first sight this seems as potentially helpful as all of Broadbent's other heuristic devices, but generations of students have failed to discover precisely what it means. It seems to contrast control of attention by selective tuning of perceptual sensitivity and by response bias. It is stressed that these functional processes must be guided 'from the top down,' by the complex mental representations that people make of their environments, their momentary and continuing goals, and of the information that they need and the actions that they must perform to achieve these goals, rather than from 'the bottom up' by passive assimilation of sensory input. To do this Broadbent reviewed the entire body of experimental work in applied experimental psychology between 1958 and 1970. Hindsight suggests that the problem was the familiar and inevitable crisis in any rapidly developing science: the style of models that had been so successful in 1958 no longer worked for a new literature and a collection of new experimental results that had become too vast and diverse to be captured by any single guiding metaphor or by models as concise and fruitful as those that *Perception and Communication* was able to draw from the no more than 20 key papers that had been published. The problem of descriptive metaphors that were once new and brilliant being inevitably overwhelmed by new data, cuts as deep in applied cognitive psychology as in all other sciences.

After 1972, Broadbent continued his commentary on the problems of intellectually navigating between finicky debates about the increasingly sterile details of models based on single experimental paradigms, and theoretical frameworks that are so general that it is unclear what they explain or how they can be tested. This was one of the themes of his William James lectures at Harvard, published as *In Defence of Empirical Psychology* (Broadbent, 1973). A background provocation was a fashion, during the 1970s, for rejection of empiricism in favor of 'cognitivist' descriptions. Broadbent's response to the new discipline of 'cognitive psychology' that had been pompously named, but ill-defined, by Ulric Neisser in 1966, had been cautiously tepid. He feared that it might encourage less gifted practitioners to a psychology based on pseudo-philosophical speculation about phenomenology rather than on empirical evidence. This motivated an expression of concern at these trends in British psychology in a paper "In Defence of Empirical psychology" (Broadbent, 1970) published in the *Bulletin of the British Psychological Society*. Broadbent detested academic affectation, particularly grandiloquent posturing to win the generous hearts of students who had not yet been given the critical apparatus necessary to distinguish efflorescing cults of personality from the necessities of methodological rigor and hard-won information. He saw a threat to civilized values when, as he said, "If one doesn't argue from evidence it must come to bludgeons." Perhaps much of his distaste came from what he called his puritanical streak, which took him directly athwart the spirit of the 1970s: "I decided then, and still believe, that self-realisation and the development of personal experience are neither dignified nor respectable goals in life, that most of the world lives within extremely tight economic margins and that positions of privilege (such as the conduct of psychological research), demand obligations in return."

During the later 1970s and the 1980s he commented shrewdly on early connectionist models by McClelland and Rumelhart, on parallels between human experts and expert computer systems, and became concerned with levels, hierarchies, and the locus of control, and the effects of emotional states and personality traits on selective attention and decision-making. He warmed to the strong neuroscience bias that Larry Weiskrantz had brought to the Oxford University Department. He proposed a 'Maltese Cross' model to describe interactions between modality-based memory subsystems and their mutual control to achieve selective attention. His empirical research continued on a remarkable breadth of topics: how the emotional valence of words affects the speed with which they can be recognized; more general problems of word perception and visual perceptual selectivity under stress and distraction and, with Margaret Broadbent and Sue Gathercole, on the processing of rapid strings of successive visual stimuli, and with other distinguished collaborators on grouping effects in short-term memory.

Broadbent never lost his determination to apply psychology to economically and socially useful ends, and to enrich theory by studying complex real-life tasks rather than artificially simple experimental paradigms. With Diane Berry he investigated how managers develop and use mental representations of the systems that they try to control, providing empirical evidence that their decisions are affected strongly by assumptions, biases, and even by computational procedures that they do not articulate, and of which they may even remain unaware. Parallel work on errors and lapses took place in hospital wards and on automobile production lines, in collaboration with his student Peter Cooper, and with Dennis Gath of the Oxford University Department of Psychiatry. It is a pity that this work has gained far less attention than it deserves, because it teaches precisely those lessons of empirical methodology and acceptance of social responsibility that he most wished to pass on.

Broadbent felt that his self-termed 'puritanical streak' discomforted his younger colleagues, but this was not quite the issue. He could not endure gladly the company of people whom he took to be insincere, or perverse, in their determination to avoid recognition of logical weaknesses in their positions, but he was always as fair and generous to them as he could force himself to be. Any discomfort he provoked did not come from his humane responses but from glimpses of the considerable effort that they sometimes cost him. His intellectual acuity was often intimidating, as was his occasional looming impatience with remarks or behavior that he interpreted as histrionic or unkind. His genuine shyness was disconcerting in one so obviously, and exceedingly, clever and eminent, and sometimes compounded problems of communication, but his underlying kindliness was unmistakable, totally reliable, and amply assuaged occasional unease. He was frugal with praise, perhaps because he was suspicious of merely verbal gestures, but he was endlessly generous and sensitively proactive with practical help. It is excellent that he retired amply convinced of the intellectual esteem of his colleagues and students, but sad that he remained unaware that he had also earned far more affection than he ever realized.

Donald Broadbent's lifetime research legacy is documented in his 247 publications and four books. His death, distressingly soon after his retirement, deprived his family and friends of emotional and intellectual comfort and support, and curtailed further sharp insights on how the intellectual history of the subject of psychology may determine its future.

See also: Cognitive Neuropsychology, Methods of; Cognitive Psychology: History; Experimentation in Psychology, History of; Hebb, Donald Olding (1904–1985); Psychology, History of (Twentieth Century); Stress, Coping and Health; Stress: Psychological Perspectives.

Bibliography

Broadbent, D.E., 1958. Perception and Communication. Pergamon Press, Oxford, UK.
Broadbent, D.E., 1961. Behaviour. Basic Books, New York.
Broadbent, D.E., 1970. In defence of empirical psychology. Bulletin of the British Psychological Society 23 (79), 87–96.
Broadbent, D.E., 1971. Decision and Stress. Academic Press, London.
Broadbent, D.E., 1973. In Defence of Empirical Psychology. Oxford University Press, Oxford, UK.
Broadbent, D.E., 1984. The Maltese cross: a new simplistic model for memory. Behavioural and Brain Sciences 7 (1), 55–94.
Hebb, D., 1949. The Organisation of Behaviour, Methuen, London.

Broadcasting: Regulation

Eli Skogerbø, University of Oslo, Oslo, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

By broadcasting regulation, we refer to political decisions and legal and economic instruments by which broadcasting activities, mainly radio and television services directed at a mass audience, are governed. A main distinction between regulatory systems can be drawn between democratic and authoritarian states, another between commercial and public broadcasting. With digitization, convergence, and globalization, the borders between broadcasting and other communications have become increasingly blurred, but the systems have retained their main characteristics.

By broadcasting regulation, we refer to political decisions and legal and economic instruments by which broadcasting activities, mainly radio and television (TV) services directed at a mass audience, are governed. These instruments include laws, rules, statutes, concessions, licenses, and traditions, as well as international law, conventions, directives, and policy making. All these instruments make up the legal, economic, political, and social frameworks within which broadcasting activities are carried out. Naturally, the regulation of broadcasting varies in time and space. First and foremost, broadcasting regulations have varied considerably between countries, as they historically reflected national, political, cultural, social, and economic conditions. Historically, one of the most important structural conditions that led to regulation of broadcasting worldwide was scarcity of resources. Scarcity was technological as the amount of radio frequencies that carried transmission of radio and later TV was limited. In some countries, e.g., the United States, the first years of radio were characterized by chaos in the ether, i.e., by radio stations using the same frequencies and interfering and destroying each other's broadcasts. This situation called for regulations. Scarcity also referred to economic resources as broadcasting soon became a costly industry that in some countries attracted major business interests that had the capacity to monopolize the airwaves. In other countries, lack of capital led to increased state intervention. Both spurred regulations, although with different objectives and effects.

Types of Regulatory Practices – Major Distinctions

There are several ways of typologizing the way broadcasting is regulated but the general rule is that the type of political system that a state has has a fundamental bearing on the way it regulates the media. Accordingly, the first main distinction between regulatory systems can be drawn between democratic and authoritarian states. Another main distinction can be drawn between states that historically have regulated broadcasting as an industry, a commercial market, and those that have regulated broadcasting as a public service such as in the European welfare states where broadcasting historically has been regarded a culture service and regulated in the public interest.

Authoritarian and Democratic Regimes

Broadcasting and all other media will be regulated fundamentally differently with regard to freedom of the media and freedom of expression in authoritarian as opposed to democratic states. Concerning ownership and market behavior, there may not necessarily be large differences. In democratic countries, the media are guaranteed a high degree of autonomy, editorial freedom, and freedom from censorship, persecution, and sanctions. Authoritarian states generally have repressive politics toward the media: either the media are treated as instruments for the incumbent regimes (e.g., the former Soviet Union, contemporary North Korea) or they are censored or placed under laws that to larger or lesser degree restrict freedom to criticize the incumbent political leadership (e.g., Zimbabwe, Uganda). Restrictions on media freedom may take many forms in repressive regimes and may vary from total censorship to no censorship but with heavy fining or penalties for those voicing critical views as countries classified as authoritarian may be repressive to varying degrees. Many states tolerate some degree of media freedom, whereas others exercise heavy censorship on more or less all media outlets. There is not always correspondence between legislation and media freedom: Many countries formally adhere to media freedom and freedom of expression for citizens, but still media outlets are closed down and persecuted and journalists are subject to harassment, persecution, and killings when they publish critical articles about governments, corruption, organized crime, and the like.

Neither is the distinction between authoritarian and democratic systems coincidental with the degree of state intervention in the broadcasting or media sectors. Here, too, there is a fundamental difference between authoritarian and democratic states. The media in authoritarian states may be controlled by state ownership or takeovers, as well as by strict regulation, whereas in democracies, state ownership of broadcasters have been accompanied by an ethos of distance between authorities and media institutions, freedom of expression, and editorial freedom. To illustrate, there are many examples around the world of state-owned or government-regulated public service broadcasting institutions that have or have had a national monopoly and that simultaneously have enjoyed a high degree of editorial freedom. In the Nordic and North European countries, for instance, broadcasting institutions have been wholly or partly owned by the state, but the

relationship between the institutions and the state have been distant, following the principle of 'arms' length' between the state and the media, and similar if not identical policies were followed in the United Kingdom toward British Broadcasting Corporation (BBC) and in post-World War II Japan concerning the public broadcaster NHK. To the contrary, there are also examples of countries where privately owned and market-driven broadcasting institutions have been heavily censored and worked as instruments of authoritarian governments, such as was the case with Brazilian Globo during the military government in the 1960s.

The developments of the Internet, mobile communications, and smartphones with both video and audio playing and recording facilities, and newcomers such as the social media (e.g., Facebook, YouTube, Twitter) have in the past decades made it harder for authoritarian governments to suppress and restrain media freedom. What has been termed the Arab Spring, the revolutionary uprising in the Arab World, most notably Tunisia, Libya, and Egypt, in the course of 2011 and 2012, was by some commentators described as made possible by the social media. Although the media did not cause political upheaval, such examples serve to show that the increased access to mobile communications and many different media outlets facilitate and diversify political communication. The uprisings did not everywhere lead to fast throw-overs of authoritarian regimes and the media have remained strictly controlled in several Arab countries, thereby confirming the tight relationship between political systems and media regulation.

Within democracies, two models of broadcasting regulation have historically been clearly distinguished: the commercial and the public service broadcasting model. They have their roots in different parts of the world and developed out of the societies in which they were set up.

Commercial Broadcasting Regulation

The commercial broadcasting model, regulating of broadcasting as a competitive industry, has its deepest roots in the United States where a distinct broadcasting policy was never formulated. Here, the main objectives of broadcasting regulation were to avoid market failure such as 'chaos in the ether,' monopolization of ownership, and secure access to the airwaves for different opinions, interests, and groups. In the United States, the Communications Act of 1934 for decades was the backbone of broadcasting regulations. It set up the regime that was in power until the late 1990s, and more or less defined the framework within which the American media and broadcasting industry was to develop. Broadcasters were regulated according to dominance in the market by the restrictions on ownership in terms of how many radio and TV stations could be owned by one company or individual and by entry regulations that foreclosed the networks from dominating program production and that divided broadcasting and communications into separate markets. The basic principles remained until the Communications Act of 1996 removed most of the entry barriers between different markets and lifted or relaxed ownership restrictions on radio and TV stations. In effect, these changes were aimed at opening and establishing one communications market, taking into account that the changes in information technologies had undermined the rationale for implementing different regulations for, e.g., broadcasting and telecommunications. In 2014, more than a decade after the lifting of these restrictions, however, the communications markets remain rather separate, although a number of multimedia services are developed in the intersection between them.

Some variant of the market regulation model was adopted in other countries too. In Latin America, the commercial model was adopted in most countries, and in some respects, these countries placed few restrictions on the broadcasters. In Brazil and Mexico, for instance, the privately owned and funded Globo and Televisa, the dominant broadcasters in the two countries, respectively, were allowed to monopolize their national markets in the second half of the twentieth century. The broadcasters benefitted from close alliances with repressive governments in the 1960s and 1970s, and experienced few or no restrictions concerning market expansion, or obligations concerning programming. In regulatory terms, Brazil and Mexico thus can be regarded as historical examples where authoritarian regimes combined strong restrictions on media freedom and freedom of expression with few market restrictions. In Europe, the small state of Luxembourg adopted a regulation that allowed commercial broadcasters, among them well-known RTL, to transmit commercial and popular radio and TV across borders, reaching large and young audiences in many European countries (Paterson and Smith, 1998). Broadcasting regulation in Luxembourg thus represented a quite different tradition than the public service model that was dominant in most Western European countries for decades. The European tradition of vesting the responsibility for production of diverse programs in one or a few public institutions was adverse to the American conception that diversity and pluralism is a product of freedom of expression interpreted as freedom to establish media institutions.

Public Service Broadcasting

In most West European countries, broadcasting was for many decades regulated as a public service, meaning that broadcasting, as other media, have been included among other goods and services that wholly or partly have been provided by means of public policy. Public services are goods provided to the entire population on fairly equal conditions and that should be accessible for all at a relatively low cost. Most European countries centralized their broadcasting services in one or a few institutions either when broadcasting was introduced in the national context or shortly after. Sometimes but not always the broadcasters were wholly or partly state owned, and to begin with, there were few or no distinctions between the state as the regulator of broadcasting markets and the state as actor in the broadcasting sector. Instead, media policy and broadcasting policy were treated as central topics within the larger area of cultural policy, i.e., together with political issues related to maintaining and strengthening national identities, language, history, and culture. As such, in Europe, broadcasting has been and to some extent still is an area considered as a sector for production of cultural goods more than cultural commodities (Lunt and Livingstone, 2011).

The public broadcasting system thus does not refer to one particular type of institution; rather it can be described in

contractual terms as a set of privileges and obligations placed on certain types of service providers. The privileges included exclusive licenses to broadcast and exclusive rights to revenues such as license fees or broadcast advertising. In return, the licensed broadcasters were obliged to provide nationwide coverage and diverse programming that catered to the interests of the entire population (Syvertsen, 2003). This model, as set out by Trine Syvertsen, allowed for national variations in the public service model. The fact that regulating broadcasting as a public service did not mean that identical models were set up in several countries, rather that each country tried to design a system of publicly governed broadcasting within the frames of external constraints and national characteristics, is also a feature that has been emphasized during the 1990s. The model has also proven productive as it incorporates changes in the political developments of the public service systems. In some countries, the public broadcasters held a national monopoly and provided universal coverage and a program menu consisting of enlightenment, education, and entertainment that should appeal to different social and cultural groups. In some countries, license fees were the sole source of revenue for the public broadcasters; in others they drew their incomes from advertising or collected revenues from both these sources. Public broadcasters were designed to cater to a diverse number of national interests (management of scarce resources, security, education, nation-building, news) as they were set up in many European countries in the 1920s and 1930s.

In Europe, a large number of quite different national systems have historically been subsumed under the concept. Public service broadcasting regulations varied from country to country, reflecting the cultural and political characteristics of different states. The large, publicly owned, and centralized organization, such as the BBC, is often regarded as the 'mother' of all public service broadcasters, and a model later exported all over the world, whereas it also easy to find historical examples of public service broadcasters displaying quite different characteristics, e.g., the Dutch or Finnish systems, where different institutions were granted privileges and obligations on different conditions. The Dutch public service system, for instance, incorporated a number of different broadcasters that had in common that they are affiliated with the traditional social pillars of Dutch society, i.e., the religious and social cleavages that until the 1960s divided Dutch society into a number of strictly separate groups. Each of these groups or pillars established 'broadcasting societies' that were allowed the right to operate their own radio station and, at a later stage, produced programs for TV. These broadcasters had the right to collect license fees, and from 1967, they also had revenues from advertising, allocated according to a complicated system. In 1967, a national broadcaster providing news and sports was established in order to provide independent programming (Nederlandse Omroep Stichting, NOS), too. The allocation of, first, radio frequencies and later 'windows' on the national TV channels to 'broadcasting societies' took care of the needs to cater to different audience segments, whereas NOS provided the audience with common references to the national culture, politics, and society. Over time, the pillarized system disintegrated but the Dutch public service system had by 2012 retained many of its main characteristics (Donders, 2012). The model of public broadcasting has, as the examples above show, been constantly changing and adapted under various internal and external pressures. This means that the content and design of the privileges enjoyed by the public broadcasters and the obligations imposed on them have also changed over time, but not simultaneously. In the United Kingdom, BBC lost its monopoly in the mid-1950s, whereas in neighboring Norway, the NRK retained this privilege until the beginning of the 1990s. After 2000, some countries have abandoned the license fees, whereas others have retained it.

Although there have been many predictions of the demise of public broadcasting, the public service broadcasters have remained more or less intact all over Europe. As the examples above show, the public service broadcasting model has been continuously under pressure for decades, by challenges in the form of increasing competition, convergence, and globalization of markets. Nevertheless, there has been a growth not a decline of public broadcasters throughout the world lately. The model was exported to many African countries as a legacy of colonialism and several African countries too have adopted the public service model more recently. Post-1990, public broadcasting has been a model for the new democracies in Eastern Europe too. The model has in spite of the discourse of continuous crisis, proved viable, tenacious, resilient, and exportable, and it may prove to be so for the foreseeable future (Jakubowicz, 2011; Bardoel and d'Haenens, 2008; Storsul and Syvertsen, 2007).

Convergence, Globalization, and Competition

In recent decades, nevertheless, the borders between broadcasting, telecommunications, print, and online media in terms of technologies and market have increasingly become blurred as a result of digitization and convergence of production and transmission technologies. Broadcasting institutions no longer are restricted to transmitting their services over broadcast networks, they operate a number of online and mobile services as well, such as news Web sites, online radio and TV, streaming services, and mobile applications, only to mention a few. These developments have changed the relationship between broadcasters and their audiences. Audiences now have access to broadcast services over widely different platforms that are internally competitive, and they have also become important producers of and contributors to programs and services. All these factors, combined with internationalization and globalization of the media industry, have challenged traditionally models of broadcasting regulation. Digitization, globalization, and liberalized economic policies have spurred processes of reregulation and liberalization in communications policies all over the world, making the distinctions between interventionist and market systems less obvious. The public service and the commercial model historically could be different systems, whereas over the past decades, hybrid or mixed models have become common in Europe and elsewhere. Canada and Australia early set up mixed systems, Britain introduced its duopoly consisting of the BBC and the independent broadcasters in 1956, and most other European countries followed suit during the 1960s and 1970s. The Scandinavian countries and Belgium were the last countries to introduce nationwide commercial broadcasters in the early 1990s, the so-called

hybrid channels. Hybrid channels were private broadcasters granted the privilege of tapping the national advertising market in return for complying with public service obligations concerning programming and coverage. The group of mixed systems is large, diverse, and increasing, reflecting technological as well as political changes in the broadcasting sector.

In Europe, the opening of markets and equalization of competition that have taken place within the European Union have had important influence over national broadcasting and communications policies. Over the past decades, all member states in the European Union as well as most other countries have opened their broadcasting sectors so that the public broadcasters become one of several broadcasters that the national audiences can choose from. Further, the European Union has decided on a set of regulatory measures of which some are aimed at protecting the public broadcasters (e.g., the Amsterdam protocol 1997) and some are aimed at securing equal competition and a level playing field (the Audiovisual Media Services Directive, the AVMS Directive). By many commentators, the EU provisions are regarded as conducive to imposing commercial broadcasting regulation all over Europe.

Still, in 2014, when the globalization of the media industry has become conspicuous, it is fair to state that regulation of broadcasting to a considerable degree remains within the hands of individual states, although the mixed pressures of technological change, globalization, and market harmonization have added to leveling out some of the differences. This is true even if we consider the attempts of international organizations, such as the European Union and the United Nations, to establish international regulations for media freedom, trade in audiovisual products, and other issues concerning cross border broadcasting. A common and pressing issue for governments is, however, to find ways of meeting new issues and problems raised by converging media, i.e., the fact that the media no longer are confined within one technological platform for distribution and production. Broadcasters not only operate TV and radio channels but also publish Web sites and distribute news and other services on social media and on applications for mobile phones and tablets, and their programs may be viewed on TV screens, mobile media, or on a computer screen. Other media such as the traditional newspapers have become equally multimedial. While most continue to publish a print edition, they also publish online newspapers, Web TV, online services, and more.

Research on Broadcasting Regulations

As a field of systematic research, the politics and regulation of broadcasting have developed gradually from the 1970s onward. Although the broadcasting industry and the broadcasting institutions received much attention from the moment they were established, they were not generally subject to systematic analysis studies until the 1980s. There are several explanations for this. The first and most obvious one is that communication studies in many countries were not established as a separate academic field. To the degree that there existed a research tradition on media policy and politics in general and broadcasting regulation in particular, studies tended to be either descriptive country studies accounting for one specific broadcasting system or they were efforts to typologize regulatory regimes across the world.

One example of early typologies that gained much attention was the book *Four Theories of the Press* by Fred S. Siebert, Wilbur Schramm, and Theodore Peterson. Writing at the peak of the Cold War, they divided the world into two main categories of media regulatory regimes. Their main distinctive criterion was the type of state, democratic or authoritarian. Within each of these categories, they distinguished between two types of media systems. Democratic states could have 'liberal' or 'social responsible' media regulation, broadly referring to countries with commercial media markets and countries with a higher degree of state intervention. Within authoritarian regimes, a Soviet-Communist media system was set up as a specific variant (Siebert et al., 1956). Another, less well-known typology was presented by Raymond Williams. Williams distinguished between authoritarian, commercial, and democratic systems. This was also an effort to set up a universal typology over media regimes, with the exception that Williams argued that fully democratic regimes were nonexistent and needed to be established (Williams, 1968). Both typologies were prescriptive rather than descriptive and did not in any detail describe the conditions under which different regulatory systems developed. In 2004, Daniel Hallin and Paolo Mancini published their modern, already classic, typology of media systems. Unlike their predecessors, they had no global ambitions but limited their classification to Western Europe and North America (Hallin and Mancini, 2004). In their terms, the Polarized Pluralist Model applied to the Southern European countries, France, Portugal, Italy; the Democratic-Corporatist to the North European countries, and the Liberal to the United Kingdom and North America. The models characterize the media systems as a whole, not only broadcasting, and point to some of the main characteristics of media markets and regulatory characteristics of each of these types. The typology is constructed from comparisons of structural and systemic characteristics of media markets, journalism, and media institutions across countries, the authors relying on existing qualitative and quantitative data. It is intended as "a theoretical synthesis and framework for comparative research," providing hypotheses and concepts (Hallin and Mancini, 2004: 16). As such, the typology has been widely discussed in media research and has been criticized for its shortcomings, empirically and conceptually. As the types to a large extent rest on historical and structural data, the critics have been mainly occupied with the conceptual 'fit' of the types and the validity, less with empirical testing (Humphreys, 2012; Herkman, 2009).

The relative stability of national regimes and broadcasting institutions between 1950 and 1970 probably contributed to the lack of studies investigating, comparing, and contrasting national broadcasting systems in order to arrive at more precise models. In spite of the fact that TV in this period replaced radio as the dominant electronic mass medium in most of the Western world, the introduction of the new medium in general did not change the existing regulatory regimes. In countries that had a national monopoly on broadcasting, TV services were opened within the existing institution (e.g., Britain, Scandinavia). In countries that had a commercial system, TV services were provided by the existing actors (the United States), and in authoritarian states, TV was controlled and used as an

instrument for government propaganda in the same way as radio and newspapers.

In research, media politics and regulation had its breakthrough in the 1980s, as the old models came under pressure by driving forces originating in technological as well as political and economic changes. These changes also spurred renewed interest into describing and evaluating the media regulation and media institutions, among them the broadcasters. Increased interest among media researchers for studies of the political and regulatory aspects of the media to some degree replaced the preoccupation with power and influence of media institutions, media corporations, and media content that dominated research in political communication in the preceding periods. One of the outcomes of renewed interest into media regulation was an increase in the number of comparative studies of broadcasting regulation. Research into broadcasting typically had meant country studies, whereas the late 1980s and 1990s saw an upsurge of works that compared and contrasted national systems. These studies put different systems in relief to each other and reconceptualized the characteristics of national systems into general models, such as the coining of the 'public service broadcasting' model as an analytical concept in the early 1990s.

Technological changes, in the 1980s illustrated by the venture of transnational satellite TV and in the 1990s by the breakthrough of digital networks, brought a whole range of new products and services. This created a new situation of 'communicative abundance' Karppinen, 2012 in sharp contrast to 'scarcity,' which was one of the foundations for broadcasting regulation. At the same time, many Western countries saw deep-going political change, in the sense that liberal governments initiated reforms and liberalization of some of the previously protected sectors, such as telecommunications, energy, and infrastructure (i.e., New Zealand, United Kingdom, Norway). Reforms of the public sector (new public management) affected also broadcasting in particular by the breaking up of the remaining broadcasting monopolies. As a direct consequence, media politics and regulations in the last two decades of the twentieth century were not only moved high on the agenda for national policy makers but also discussed in regional and international forums such as the Council of Europe, the European Union, and the World Trade Organization.

Adapting Broadcasting Regulation to the Age of Digitization

The first decade of the twenty-first century has framed broadcasting within a media landscape where the processes of digital change, marketization, and globalization have continued to influence services, institutions, and policy making. Many countries have again undertaken profound revisions of their broadcasting legislation. In the United States, ownership regulation has been radically softened, and broadcasting is no longer as protected toward monopolization as before. In the United Kingdom, the 2003 Communications Act set up an entirely new way of regulating broadcasting and communications markets, implementing a common regulatory agency for broadcasting, telecommunications, and digital services. A large number of states (e.g., Finland, Norway) have either phased out analogue distribution of TV and radio or put up a deadline for the transition to a fully digital distribution.

The study of media politics and regulations is still a relatively young field within communication research. During the 1990s, the number of empirical studies comparing different national broadcasting systems increased markedly. This development signals that there is a methodological reorientation in the field, emphasizing descriptive and explanatory analyses of a few cases at the expense of normative evaluation of universal models. The comparison of national systems yields insights into the conditions that have shaped broadcasting institutions and provides knowledge necessary to understand the current political conflicts on how to regulate broadcasting under new conditions, both nationally and internationally. In media studies, there is a long tradition for adopting an implicit, not explicit, normative (critical or liberal) approach when phenomena are analyzed, which sometimes has created problems of analytical clarity. With the increased focus on comparisons, these problems may be overcome, as normative judgments will have to be made more explicit.

See also: Broadcasting: Regulation; Freedom of the Press; Mobile Communication; New Media and Democracy in the Arab World; Printing as a Medium; Public Broadcasting; Radio as a Medium; Social Media; Social Protest and New Media; Television: General; Television: History.

Bibliography

Bardoel, J., d'haenens, L., 2008. Reinventing public service broadcasting in Europe: prospects, promises and problems. Media, Culture & Society 30, 337–355.

Donders, K., 2012. Public Service Media and Policy in Europe. Palgrave Macmillan, Basingstoke.

Hallin, D.C., Mancini, P., 2004. Comparing Media Systems: Three Models of Media and Politics. Cambridge University Press, Cambridge.

Herkman, J., 2009. The structural transformation of the democratic corporatist model: the case of Finland. Javnost – The Public 16.

Humphreys, P., 2012. A political scientist's contribution to the comparative study of media systems in Europe: a response to Hallin and Mancini. In: Just, N., Puppis, M. (Eds.), Trends in Communication Policy Research: New Theories, Methods and Subjects. Intellect, Bristol, UK and Chicago, USA.

Jakubowicz, K., 2011. Public service broadcasting: product (and victim?) of public policy. In: Mansell, R., Raboy, M. (Eds.), The Handbook of Global Media and Communication Policy. Wiley-Blackwell.

Karppinen, K., 2012. Rethinking Media Pluralism. Fordham University Press, New York.

Lunt, P., Livingstone, S., 2011. Media Regulation: Governance and the Interests of Citizens and Consumers. Sage, London.

Paterson, R., Smith, A., 1998. Television: An International History. Oxford University Press, Oxford.

Siebert, F.S., Schramm, W., Peterson, T., 1956. Four Theories of the Press: The Authoritorian, Libertarian, Social Responsibility And Soviet Communist Concepts of What the Press Should Be and Do. University of Illinois Press, Urbana, IL.

Storsul, T., Syvertsen, T., 2007. The impact of convergence on European television policy. Convergence: The International Journal of Research into New Media Technologies 13, 275–291.

Syvertsen, T., 2003. Challenges to public television in the era of convergence and commercialization. Television & New Media 4, 155–175.

Williams, R., 1968. Communications. Penguin, Harmondsworth.

Broaden-and-Build Theory of Positive Emotions

Suzannah F Isgett and Barbara L Fredrickson, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The broaden-and-build theory of positive emotions stems from a set of twin hypotheses. First, the broaden hypothesis proposes that positive emotions momentarily expand our perception of the world in ways that facilitate global visual processing, better attentional flexibility, and larger thought–action repertoires. The build hypothesis purports that, over time, these fleeting experiences of expanded awareness that accompany positive emotions such as joy and excitement accumulate over time to facilitate growth of a person's social, cognitive, emotional, and physical resources. Empirical evidence supporting these hypotheses is discussed, as well as the theory's implications for behavior, psychological resilience, social interaction, and health.

Introduction

Emotions permeate the very fiber of our lives, affecting our judgment and behavior, as well as our health and happiness. Despite their importance to our everyday lives, they can be operationally difficult to define. Broadly described, emotions are brief, multisystem responses to the way individuals appraise their current circumstances (Fredrickson, 2013). These responses last for just seconds or minutes, and when natural and healthy, serve adaptive functions. Just as our bodies sense fluctuations in the physical environment and adapt accordingly by regulating our heart rate or hormones, emotions allow our minds to adapt to the surrounding social or physical environment. For example, the emotion of anger prompts us to approach the source of the emotion to resolve or better understand a conflict. Fear signals our body to flee from danger to avoid potential harm.

Positive emotions are a special class of emotions; they are responses to an individual's current beneficial circumstances. For example, joy is a positive emotion that emerges when one's current circumstances present unexpected good fortune. Serenity, sometimes called contentment, arises when one interprets his or her circumstances as utterly cherished, right, or satisfying. Love, the most frequently experienced positive emotion, is felt in the context of a safe interpersonal connection. Other positive emotions, such as gratitude, interest, hope, pride, amusement, inspiration, and awe, spring from different ways of appraising one's circumstances as beneficial in some way. Any event or situation can elicit multiple emotions, both positive and negative, which adds to the richness and depth of the human emotional experience.

Positive emotions, like those mentioned above, also serve adaptive functions, although scientific evidence for their importance – relative to negative emotions – has been slower to accumulate. During the mid-twentieth century, the burgeoning field of psychology was largely studied under the doctrine of behaviorism, in which observable behaviors are only attributable to the environment and not to mental states. Within this methodological and analytical approach, all emotions were largely deemed irrelevant and misleading epiphenomena (Skinner, 1974). Eventually, emotions began to be studied more rigorously as the zeitgeist of psychology shifted once again to include mental states. However, for decades the focus was still on negative emotions – like fear, anger, or disgust – that could be linked to serious societal issues such as violence, social ostracism, or mental illness. In the history of psychology, only recently have positive emotions been rigorously studied.

However, positive emotions differ from other types of positive states. Hedonic pleasure is a positive response to physical sensations such as eating food or having sex that requires no appraisal of meaning. Optimism is an enduring mental attitude that leads people to interpret events and situations as good. It is not in itself a positive emotion, although it does predispose people to experience more positive emotions. Even positive affect or mood, although components within positive emotions, refer more to an unfocused, diffused state of being. Positive emotions, though momentary or fleeting, are unique in that they cause patterns of thought that are flexible, inclusive, creative, and forward-looking (Fredrickson, 2013).

While it is intuitively easy to consider the adaptive functions of negative emotions such as fear, it is less clear as to how positive emotions serve us. One can imagine and even personally relate to how strong negative emotions alter our thoughts and behavior. On the other hand, what does a fleeting moment of serenity do for us? Do positive emotions exist solely to make us feel good, or do they serve another purpose? These questions, among others, are addressed by the broaden-and-build theory of positive emotions.

The broaden-and-build theory synthesizes two hypotheses that encapsulate the beneficial effects of positive emotions on cognition, behavior, as well as both mental and physical health. The first hypothesis – the broaden effect – proposes that positive emotions have momentary effects on cognition and behavior that expand our awareness and repertoire of what we want to do. The second – the build effect – posits that moments of broadened awareness, sparked by positive emotions, accumulate and compound to build people's personal resources, whether social, emotional, or even physical. Moreover, these effects are marked by reciprocal causality, in that many of the resources that positive emotions help to build also increase the odds that people experience positive emotions, leading to an 'upward spiral' of well-being and flourishing mental health (Fredrickson and Joiner, 2002). The empirical evidence for the broaden-and-build effects of positive emotions is explained below, as well as the theory's implications for resilience and health.

Evidence for the Broaden Effect

Global Bias in Global–Local Processing

Although we take in information through our five senses, we do not consistently process it in the same way. In fact, the details of our environment that enter our awareness may differ based on a stimulus's saliency, its relevance to a situation, or even depending on our mental state. Research in cognitive science has shown that we can process stimuli *globally*, by evaluating a stimulus holistically, or *locally*, by attending to the details. In a paradigm first developed by Navon (1977), participants are shown a geometrical figure that possesses both global and local features and are then asked to choose which of two comparison figures is most similar to the target figure. One comparison figure is similar to the target in its global aspects; the other figure, in its local aspects (**Figure 1**). Decisions on this task have been shown to correlate with certain individual differences. A local preference tends to be associated with anxiety and depression, whereas a global tendency is associated with subjective well-being and optimism (Basso et al., 1996).

Positive emotions lend a global preference to the way people process information. In a study by Fredrickson and Branigan (2005), people temporarily induced to feel the positive emotions of amusement or contentment (through video clips) chose the 'global' comparison figure significantly more than people in a neutral condition. Other studies have linked genuine smiles (Duchenne smiles), characterized by engagement of muscles that raise the corners of the mouth and those that raise the cheek near the eye, to global preferences in processing stimuli (Johnson et al., 2010). In line with the global orientation that positive emotions evoke, they appear to literally broaden the scope of what we see. In Wadlinger and Isaacowitz's (2006) study using an eye-tracking apparatus and visual arrays that included central and peripheral images, participants who were induced to feel a positive emotion before the experiment spent more time, relative to the control group, looking at peripheral images and more frequently changed the location of their gaze.

Supporting the notion of broadened visual attention, a functional magnetic resonance imaging study by Schmitz et al. (2009) examined brain activity in people while they experienced positive, negative, or neutral affective states. Specifically, the researchers were interested in how different emotional states influence visual processing. Participants were shown images of faces, which were superimposed on pictures of different houses. When participants were induced to be in a positive emotional state, they exhibited distinct patterns of neural activity in the parahippocampal place area, a brain region associated with processing of scenes or places. The authors concluded that positive states cognitively expand our field of vision, such that we process more peripheral visual information. Thus, positive emotions seem to literally broaden people's visual scope.

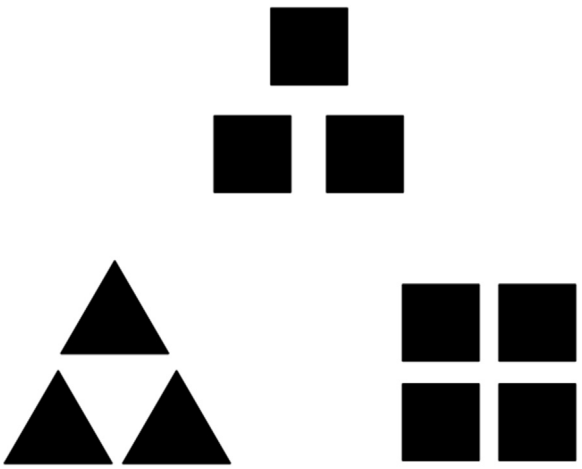

Figure 1 Sample global-local items used to test the broaden hypothesis.

Since positivity tends to increase our awareness of global features instead of just the details, people experiencing positive emotions have a temporarily expanded awareness of their surroundings. In addition, this shift in cognition has the potential to increase our semantic access and expand awareness, allowing more open and creative problem solving. Having a broadened scope of awareness may enable us to recognize opportunities more easily, and – colloquially speaking – not miss the forest for the trees.

Attentional Flexibility and Enlarged Thought–Action Repertoires

In addition to expanding awareness, positive emotions can enhance attentional flexibility. The ability to maneuver the direction of our attention with ease is essential for learning, problem solving, and healthy emotional functioning. Not being able to flexibly control attention can lead to rumination about a problem in our lives, or distractibility when focus is necessary. Johnson et al. (2010) used facial electromyography to detect Duchenne smiles in participants given an attentional orienting task, and found that a higher frequency of genuine smiles, after an experimentally induced positive emotion, predicted greater attentional flexibility in the task. Isen et al. (1985) demonstrated that experimentally induced positive affect led to more unusual words produced in a remote associates task, which is often used as an indicator of creative ability. Similarly, Rowe et al. (2007) showed that positive emotions, induced through music, increased the scope of semantic access and visual attention. By broadening our cognitive resources, such as attentional flexibility, positive emotions allow us to more easily access disparate ideas and flexibly direct our focus.

Positive emotions also broaden our repertoires of what we want to *do*. In a study by Fredrickson and Branigan (2005), participants had a larger breadth of action urges – or activities they wanted to do – after viewing video clips that induced positive emotions, whereas clips that induced negative emotions actually narrowed the breadth of action urges relative to neutral states. Although sometimes fleeting, positive emotions put us in a mental state that allows us to see more, think more broadly, and want to do more.

Inclusive Social Thinking

Finally, the broaden effect of positive emotions can be witnessed in the social domain. In everyday life we tend to think of people in terms of social groups: "Is he one of 'them' or one of 'us'?" "Would I consider her to be in my social circle?"

Although social in-groups can help facilitate social bonding and trust, they can also lead to exclusion and prejudice against out-group members. Because positive emotions broaden our scope of awareness, the way we think about social groups becomes more inclusive. For instance, while experiencing positive affect, people form more inclusive positive categorization, such that they are more likely to perceive a person as a member of an in-group rather than an out-group (Dovidio et al., 1998).

Additionally, positive emotions have even shown to soften the boundaries of racial categorization. A robust phenomenon studied in social psychology is the own-race bias in face recognition, in which people are better able to recognize individual faces belonging to members of their own race, relative to other races (Meissner and Brigham, 2001). Because faces are generally processed by their holistic properties as opposed to a collection of features, it is possible that the own-race bias partly emerges from less holistic face processing of cross-race faces (Tanaka et al., 2004). Because positive emotions enhance global processing, Johnson and Fredrickson (2005) sought to test the effect of positive emotions on facial recognition in own-race and cross-race faces. In a sample of Caucasian individuals, participants induced to feel joy showed nearly the same level of accuracy in recognizing black faces as white faces, whereas participants in the fear-induction or neutral condition still displayed the often-observed own-race bias in racial recognition. While the mechanism behind this effect is unclear – whether it is a change in global processing or general formation of more inclusive social categories – this study provides evidence of ways that positive emotions broaden within social domains.

Positive emotions expand the breadth of overall social connection, and in turn people exhibit more perspective-taking and compassion for dissimilar others, such as individuals from another cultural background (Fredrickson, 2013). Waugh and Fredrickson (2006) proposed that because positive emotions elicit greater social inclusiveness, they subsequently produce greater feelings of oneness between people in a new relationship. This was shown in a sample of first-year college students; compared to those who experienced relatively few positive emotions, participants who experienced relatively more positive emotions in daily life reported greater feelings of connectedness with their new roommate. When our social thinking is broadened, we may be less constrained by narrow social categories and therefore better able to recognize others and interact in ways that foster interpersonal connections.

Overall, evidence across multiple domains shows that positive emotions can temporarily broaden our scope visually, cognitively, and socially. When a positive emotion arises – whether it is joy, contentment, amusement, or love – our attention becomes more flexible, and processing becomes more global. Our visual scope is literally expanded, and we feel the urge to do more. Importantly, our social thinking expands as well. Taken as a whole, such broadened awareness appears to enhance people's ability to recognize a variety of opportunities.

Evidence for the Build Effect

Empirical evidence that supports the broaden effect of positive emotions provides the scientific foundation for the build effect, as well as the broaden-and-build theory as a whole. Although inextricably connected, the timescales of the broaden effect and the build effect are different. Broadened awareness happens in the moment; a brief experience of awe, for instance, allows us greater semantic access. However, the build effect happens more gradually, over the long term, as the broaden effects of positive emotions accumulate and compound over time. These micro-moments of positive emotions grow into stable affective dispositions that foster our physical health and render life more satisfying. While the effects are stronger for people with more frequent positive emotions, nearly everyone can harness the long-term benefits of positive emotions. The social, physical, and mental resources that are part of the build effect are typically shown in longitudinal studies, since a one-time laboratory visit cannot capture this long-term cumulative effect of positive emotions. When the broaden-and-build effect takes place, it can initiate an 'upward spiral' of positive well-being, in which positive emotions build enduring and consequential personal resources, such as social support, resilience, or better awareness of one's environment. These resources in turn beget future positive emotions. Empirical evidence for the build effect's impact on psychological, social, and physiological resources is described below.

A field experiment by Fredrickson et al. (2008) provided the first empirical evidence of the build effect, testing how increasing the daily amount of positive emotions can increase cognitive, psychological, and social resources. To experimentally achieve this in a sample of adults, participants were either assigned to take classes in loving-kindness meditation (LKM) or were assigned to a monitoring waitlist control group. Because LKM trains the mind to generate feelings of love and compassion for self and others, it is an effective technique of learning to self-generate positive emotions. The experimental group showed significant increases in their daily reports of positive emotions (**Figure 2**), which predicted subsequent increases in personal resources, including mindfulness, social support, purpose in life, and even fewer physical illness symptoms. In turn, the elevation in personal resources led to increased life satisfaction and lessened depressive symptomatology. Thus, by learning to cultivate more positive emotions, people can build personal resources that improve their lives.

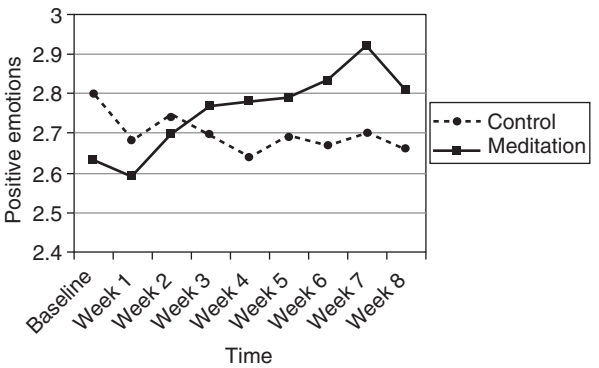

Figure 2 Week-by-week positive emotions by experimental condition. Note: Meditation training centered on loving-kindness meditation. Positive emotions were computed as the mean across all positive states on the mDES, rated on a scale from 0 to 4.

Psychological Resources

Being able to flexibly adjust to the trials and tribulations of life requires substantial psychological resources. People who are depressed tend to show bias toward negative faces, narrowed attention, and have difficulty coping with stressors. Unfortunately, the inability to cope only leads to worse psychological outcomes. Upward spirals can counter these *downward spirals* of narrow-minded thinking and negative emotions. Broad-minded coping, which captures the degree to which people can step back from their current problems and take a big-picture perspective, is predicted to increase over time with levels of positive emotionality (Fredrickson and Joiner, 2002). This reciprocal relationship between broad-minded coping and positive emotions speaks to how positivity builds psychological resources little-by-little over time.

Flourishers are individuals who rate higher than average on mental health and positive emotionality, and, in addition to experiencing bigger boosts from everyday positive emotions, they also show greater gains in resources, which in turn predicts increased levels of flourishing in an upward spiral dynamic (Catalino and Fredrickson, 2011). Luckily, the build effect is not only for those who start out positive and end up more positive. In a study exploring the benefits of mindfulness training (Garland et al., 2011), there was a reciprocal relationship found among positive reappraisal, trait mindfulness, and emotional well-being. Thus, if individuals learn how to self-generate positive emotions, as through mind-training techniques, an increase in those psychological resources may ensue.

Social Resources

People who express and experience more positive emotions also report being more socially connected. In a longitudinal study in secondary school teachers, positive emotions at work were reciprocally related to an index of social support. Positive coping also has a reciprocal relationship with an interpersonal trust (for a review, see Fredrickson, 2013).

Examining the longer-term effects of positive emotions, Kok et al. (2013) assessed how learning to self-generate positive emotions, via LKM, had effects on perceptions of social connectedness. A sample of community adults was either assigned to learn LKM for 6 weeks or serve in a monitoring waitlist control. Over the 9 weeks of the study, participants reported their daily positive emotions, as well as the degree of social connection they felt. Relative to the waitlist control group, participants in the LKM condition showed significant improvements in the amount of daily positive emotions they experienced, which in turn predicted greater social connectedness. The broaden effect of positive emotions enables people to feel more socially embedded within their community, as indexed by perceptions of social connectedness (Fredrickson, 2013).

Physiological Resources

Beyond self-report measures of positivity and subjective well-being, physiological data also support the build effect of positive emotions. Kok and Fredrickson (2010) explored the relationship between positive emotionality and a physical resource, cardiac vagal tone (VT). VT is the subtle functional arrhythmia in heart rate associated with respiration, marked by a decrease in heart rate during exhalation relative to inhalation (Porges, 2007). This physiological measure can be considered an important biological resource because higher VT (as measured by heart-rate variability or respiratory sinus arrhythmia) is generally associated with a range of mental and physical health outcomes: socioemotional well-being, better emotional and attentional regulation, as well as lowered risk for cardiovascular disease, diabetes, and obesity (Fredrickson, 2013; Porges, 2007). Like other personal resources, Kok and Fredrickson (2010) found that increases in VT were reciprocally related to increases in positive emotion.

However, it is important to determine whether or not changes in VT can be experimentally manipulated by using a positive intervention. Because of its ability to attune the mind to feelings of love and social connection, LKM training – as described earlier – can be used as a type of positive intervention. Kok et al. (in press) linked this type of positive intervention to changes in VT. By teaching people how to self-generate positive emotions through LKM, positive emotions increased over the duration of the study, and VT moderated this effect. That is, people higher in VT showed greater increases in positive emotions. Importantly, people in the LKM group showed larger increases in VT. This study experimentally linked how daily boosts of positive emotions can have effects on not only subjective well-being, but also objective markers of health as well.

Applications

Understanding how positive states initiate a series of cognitive and behavioral effects can lead to better approaches to improving human well-being in both clinical and nonclinical settings. In this section, the implications of the broaden-and-build theory for personal resilience and health will be explored.

Apart from the psychological and health benefits of the intertwined broaden-and-build effects, experiencing positive emotions may have implications for daily life and workplace environments. For instance, broadened visual awareness sparked by positive emotions has been linked to better steering performance in a driving simulator (Trick et al., 2012). Because positive emotions increase attentional flexibility relative to negative emotions or neutral states, they may aid with better task switching, especially when encountering novel information (Wang and Guo, 2008). In the world where both personal and work-related information are constantly updating and evolving, being able to encounter new information with attentional flexibility might aid in the fast-paced structure of the digital age.

The broaden-and-build effects of positive emotions also have special relevance in the workplace, where efficiency and flexibility can lead to a more productive work environment. Increasing the daily amount of positive emotions, which momentarily alter cognition in beneficial ways, might over time increase our personal resources in organizational contexts. In addition, some work tasks require creativity, integration of information, and seeing the big picture. In both cases, positive

states not only feel good, but may also foster more productivity as a by-product of these broaden-and-build effects.

Implications for Resilience

Resilience can be defined as one's ability to cope and 'bounce back' from life stressors, such as the death of a loved one, loss of a job, or a national crisis like the attacks on 11 September 2001. While some individuals are able to manage their negative emotions and eventually get back on track, others struggle to regain psychological well-being and are pulled into a downward spiral of negative emotions which can even slide into clinically significant emotional disorders. However, the broaden-and-build effects of positive emotions can buffer us, helping us climb over the wall of stress and hardship, back to a normal and healthier life.

The heart of resilience can be partly traced to the *undo effect* that positive emotions have on negative states. Biologically, positive emotions undo the cardiovascular effects of a negative emotion. In a study conducted by Fredrickson et al. (2000), anxiety-related cardiovascular reactivity was induced in a classic stress task, which instructed participants to prepare a speech to be recorded and evaluated by peers. However, they were also told that if a video clip began playing, they no longer needed to give the speech. The clips presented were intended to induce contentment, amusement, sadness, or a neutral state; speed of cardiovascular recovery was measured by the length of time it took for the person to return to his or her baseline cardiovascular levels. Viewing a positive video clip (i.e., contentment or amusement) quickened cardiovascular recovery, relative to sad or neutral clips. The authors demonstrated that positive emotions can provide your body and mind with the psychological resources it needs to recover from stressful events.

While the Fredrickson et al. (2000) study shows that positive emotions allow better recovery from a stressful task, how do these effects of positive emotions translate to the real world? Cohn et al. (2009) examined the relationship between emotions, life satisfaction, and resilience in a sample of college students. For 1 month, participants reported their daily emotions, and they evaluated their life satisfaction and ego resilience – the ability to flexibly respond to challenging or shifting circumstances – once before and once after the month of daily reporting. The authors found that, even though resilience was related to both life satisfaction and daily positive emotions, only positive emotions – not life satisfaction – predicted an increase in resilience over 1 month. Intriguingly, daily negative emotions did not significantly influence the benefits of positive emotions. Therefore, experiencing frequent positive emotions helps people lead more satisfying lives by generating greater resilience to the hardships and challenges we all face.

Additionally, evidence from brain scan studies shows that individuals with resilient personality styles exhibit less enduring neural activity to negative cues in the anterior medial prefrontal cortex and posterior cingulate cortex, a brain regions broadly associated with self-referential processing and rumination (Waugh et al., 2010). This study suggests that resilient people focus less on worrying about what is to come and take a more open-minded awareness approach. As mentioned earlier, 'broad-minded coping' is the ability to step back and evaluate a situation more broadly; this form of resilience has a reciprocal building effect with positive states.

However, resilience is not a completely static, individual trait. People can work on improving their resilience by cultivating more positive emotions. In another study by Tugade and Fredrickson (2004) in which participants were instructed to give a speech (a stress inducer), individuals without resilient personality styles who were randomly assigned to reframe the task as a positive, exciting challenge showed the quick physiological recovery characteristic of their more resilient peers. Additionally, as described above, a study using LKM to teach people how to self-generate emotions led to boosts in environmental mastery, which is an important component of resilience (Fredrickson et al., 2008).

Implications for Health

When we think of healthy behavior, we typically imagine eating fruits and vegetables, exercising, or refraining from smoking. While these behaviors are crucial to health, all of them focus on the body. After all, it is in our bodies that cancer, heart attacks, or strokes occur. However, these ordinary healthy behaviors do not address how a healthy mind can nourish a healthy body. At first blush, how we process information or feel about our relationships seems to be unrelated to these salient diseases. However, evidence for the broaden-and-build theory shows that positive emotions do not simply feel good, but also affect our physiological resources, such as cardiac VT.

This novel line of research is just the tip of the iceberg. The way we interact with others and our surroundings may influence more in our bodies than VT. Our cells constantly respond to signals from our body and communicate with one another. By regulating gene expression, cells are able to adapt to our ever-changing environment. The genes in our cells' DNA can tell our bodies to make more or less of certain proteins in order to grow, fight infection, or handle an injury, to name a few functions. New research is showing that our social environment can affect gene expression as well. A study by Cole et al. (2007) compared lonely individuals to nonlonely controls, and found a dramatic difference in their gene-expression profiles. In white blood cells – cells that are integral to our immune system – genes that promote inflammation were overexpressed in lonely individuals. Chronic inflammation can initiate a cascade of negative health effects, including an increased risk for cardiovascular disease, Alzheimer's, and rheumatoid arthritis. Therefore, being able to make social connections and feel social support has implications for health. As we have seen, when people initiate an upward spiral of positive emotions, they are more likely to make positive and meaningful connections with others. The broaden-and-build effect of positive emotions may thus enable us to forge meaningful social connections, which in turn may help keep chronic inflammation at bay.

Since the dawn of psychology, the functional understanding of positive emotions has come a long way. Despite being lighter and more fleeting than negative emotions, positive emotions are not just expendable pleasantries. The broaden-and-build theory of positive emotions explains that short-term changes in cognition and behavior elicited by positive emotions can, over time, facilitate people's ability to build

social, psychological, and behavioral resources. Newer lines of research will explore how the broaden-and-build effect influences health and well-being by integrating other areas of science, including brain imaging, immunology, genetics, and medicine.

See also: Academic Engagement; Affect-Regulation Motivation; Burnout, Psychology of; Emotion and Expression; Eudaemonism; Flow in Motivational Psychology; Interest, Psychology of; Optimism and Pessimism; Passion and Motivation; Self and Emotional Development in Adulthood and Later Life; Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model; Temperament and Motivation; Test Anxiety and Academic Achievement; Wanting vs Needing.

Bibliography

Basso, M.R., Schefft, B.K., Ris, M.D., Dember, W.N., 1996. Mood and global-local visual processing. Journal of the International Neuropsychological Society 2 (3), 249–255.

Catalino, L.I., Fredrickson, B.L., 2011. A Tuesday in the life of a flourisher: the role of positive emotional reactivity in optimal mental health. Emotion 11 (4), 938–950.

Cohn, M.A., Fredrickson, B.L., Brown, S.L., Mikels, J.A., Conway, A.M., 2009. Happiness unpacked: positive emotions increase life satisfaction by building resilience. Emotions 9 (3), 361–368.

Cole, S.W., Hawkley, L.C., Arevalo, J.M., Sung, C.Y., Rose, R.M., Cacioppo, J.T., 2007. Social regulation of gene expression in human leukocytes. Genome Biology 8 (9). Article R189.

Dovidio, J.F., Isen, A.M., Guerra, P., Gaertner, S.L., Rust, M., 1998. Positive affect, cognition, and the reduction of intergroup bias. In: Sedikides, C. (Ed.), Intergroup Cognition and Intergroup Behavior. Erlbaum, Mahwah, NJ, pp. 337–366.

Fredrickson, B.L., 2013. Positive emotions broaden and build. In: Devine, P., Plant, A. (Eds.), Advances in Experimental Social Psychology, vol. 47, pp. 1–53. http://dx.doi.org/10.1016/B978-0-12-407236-7.00001-2.

Fredrickson, B.L., Branigan, C., 2005. Positive emotions broaden the scope of attention and thought-action repertoires. Cognition and Emotion 19 (3), 313–332.

Fredrickson, B.L., Cohn, M.A., Coffey, K.A., Pek, J., Finkel, S.M., 2008. Open hearts build lives: positive emotions, induced through loving-kindness meditation, build consequential personal resources. Journal of Personality and Social Psychology 95 (5), 1045–1062.

Fredrickson, B.L., Joiner, T., 2002. Positive emotions trigger upward spirals toward emotional well-being. Journal of Personality and Social Psychology 65 (1), 45–55.

Fredrickson, B.L., Mancuso, R.A., Branigan, C., Tugade, M.M., 2000. The undoing effect of positive emotions. Motivation and Emotion 24 (4), 237–258.

Garland, E.L., Gaylord, S.A., Fredrickson, B.L., 2011. Positive reappraisal mediates the stress-reductive effects of mindfulness: an upward spiral process. Mindfulness 2, 59–67.

Isen, A.M., Johnson, M.M.S., Mertz, E., Robinson, G.F., 1985. The influence of positive affect on the unusualness of word associations. Journal of Personality and Social Psychology 48 (6), 1413–1426.

Johnson, K.J., Fredrickson, B.L., 2005. "We all look the same to me": positive emotions eliminate the own-race in face recognition. Psychological Science 16 (11), 875–881.

Johnson, K.J., Waugh, C.E., Fredrickson, B.L., 2010. Smile to see the forest: facially expressed positive emotions broaden cognition. Cognition & Emotion 24 (2), 299–321.

Kok, B.E., Coffey, K.A., Cohn, M.A., Catalino, L.I., Vacharkulksemsuk, T., Algoe, S.B., Brantley, M., Fredrickson, B.L., 2013. How positive emotions build physical health: perceived positive social connections account for the upward spiral between positive emotions and vagal tone. Psychological Science 24 (7), 1123–1132.

Kok, B.E., Fredrickson, B.L., 2010. Upward spirals of the heart: autonomic flexibility, as in indexed by vagal tone, reciprocally and prospectively predicts positive emotions and social connectedness. Biological Psychology 85 (3), 432–436.

Meissner, C., Brigham, J., 2001. Thirty years of investigating the own-race bias in memory for faces. Psychology, Public Policy, and Law 7, 3–35.

Navon, D., 1977. Forest before trees: the precedence of global features in visual perception. Cognitive Psychology 9 (3), 353–383.

Porges, S.W., 2007. The polyvagal perspective. Biological Psychology 74, 116–143.

Rowe, G., Hirsh, J.B., Anderson, A.K., 2007. Positive affect increases the breadth of attentional selection. Proceedings of the National Academy of Sciences 104 (1), 383–388.

Schmitz, T.W., De Rosa, E., Anderson, A.K., 2009. Opposing influences of affective state valence on visual cortical encoding. Journal of Neuroscience 29 (22), 7199–7207.

Skinner, B.F., 1974. About Behaviorism. Vintage, New York.

Tanaka, J., Kiefer, M., Bukach, C.M., 2004. A holistic account of the own-race effect in face recognition: evidence from a cross-cultural study. Cognition 93, B1–B9.

Trick, L.M., Brandigampola, S., Enns, J.T., 2012. How fleeting emotions affect hazard perception and steering while driving: the impact of image arousal and valence. Accident Analysis and Prevention 45, 222–229.

Tugade, M.M., Fredrickson, B.L., 2004. Resilient individuals use positive emotions to bounce back from negative emotional experiences. Journal of Personality and Social Psychology 86 (2), 320–333.

Wadlinger, H.A., Isaacowitz, D.M., 2006. Positive mood broadens visual attention to positive stimuli. Motivation and Emotion 30 (1), 87–99.

Wang, Y., Guo, D., 2008. The effects of positive emotions on task switching. Acta Psychologica Sinica 40, 301–306.

Waugh, C.E., Fredrickson, B.L., 2006. Nice to know you: positive emotions, self-other overlap, and complex understanding in the formation of a new relationship. Journal of Positive Psychology 1 (2), 93–106.

Waugh, C.E., Hamilton, J.P., Gotlib, I.H., 2010. The neural temporal dynamics of the intensity of emotional experience. NeuroImage 49 (2), 1699–1707.

Brokers and Brokerage, Anthropology of

Johan Lindquist, Stockholm University, Stockholm, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article deals with the ebb and flow of interest in the broker in anthropology since the 1950s. It attempts to situate this interest in relation to the changing world that anthropology has been concerned with, and in the context of shifting theoretical perspectives within the discipline. Beginning with the era of decolonization and the rise of the new nation-states, it tracks the position of the broker in relation to trends within political and economic anthropology, before pointing toward current concerns in anthropology.

Introduction

The broker is a classic figure in the anthropological literature, particularly with regard to the study of low-level political and economic relationships. It is important to note, however, that there has been an ebb and flow of interest in the broker and brokerage in anthropology. While appearing as a critical figure for anthropology in the context of decolonization and modernization theory in the 1950s and 1960s, and later with the rise of transactionalism, the broker largely vanished from view in the late 1970s before reemerging in the wake of current engagements with neoliberalism. The moral ambiguity of the broker as an individual who crosses social boundaries and whose motives and loyalties are questioned gained increasing force in this process. This article attends to this history, which illuminates some changing concerns in the discipline since the middle of the twentieth century, primarily in the United Kingdom and the United States. But first a statement should be made about definitions. In this article, the broker is a specific type of middleman, mediator, or intermediary. Most generally, the broker is a human actor who gains something from the mediation of valued resources that he or she does not directly control, which shall be distinguished from a patron who controls valued resources, and a go-between or a messenger, who does not affect the transaction. As shall become clear, these ideal-type distinctions are difficult to sustain in the face of a changing empirical reality and innovations in anthropological theory, particularly the concern with mediation, but will serve as a placeholder to attend to the history of the concept.

Historical Trajectories

Even if one can find essays that deal explicitly with middlemen in the anthropological canon, significant anthropological interest in brokers and brokerage does not emerge until the postcolonial era during the 1950s and 1960s, as a growing concern with social change challenged the largely ahistorical perspectives of functionalism and structural-functionalism. This was first made evident through studies that considered the modern village chief or head in Africa as a broker between overlapping fields of African and European political authority, particularly in the context of the Manchester school's focus on situational analysis (Gluckman et al., 1949; but see also Fallers, 1955). Redfield's (1956) concern with the relationship between parts and wholes, between little and great traditions in complex societies, also led to an interest in what he called "hinge groups." There was rarely, however, an explicit attempt to engage in a general cross-cultural or conceptual discussion concerning brokers.

In contrast, classic articles by Eric Wolf (1956) and Clifford Geertz (1960) explicitly developed the idea of the cultural broker as a way to describe shifting forms of political authority and the transforming relationship between village and metropole following decolonization in Mexico and Indonesia, respectively. The work of F.G. Bailey should also be noted, as he clearly identified brokers as 'agents of social change' (1963: p. 101) that allowed for the integration of villages into a wider society in Orissa in India. Generally speaking, these studies attempted to historicize the development of new forms of political relationships through the figure of the broker, particularly in relation to the rise of the new nation-states.

While sociologists for decades had been concerned with the 'marginal man' as a personality type that developed through migration and intercultural relations (Parks, 1928), in anthropology the cultural broker – generally an insider – appeared as a starting point for considering social change more broadly (Press, 1969). In this historical era, Wolf defined brokers as "groups of people who mediate between community-oriented groups in communities and nation-oriented groups which operate through national institutions" (Wolf, 1956: p. 1075). These brokers were 'Janus-like,' turned in two directions at once, and came to "stand guard over the crucial junctures or synapses of relationships which connect the local system to the larger whole" (Wolf, 1956: pp. 1075–1076).

Particular types of individuals – often multilingual and comfortable in multiple settings – emerged as ethnographic exemplars. The African chief has been mentioned, the traditional healer engaging with Western medicine another, but most well known is perhaps the *kijaji*, the Javanese Muslim leader described by Geertz (1960), whose local legitimacy was based on experiences in, and relationships with, the center of the Islamic world, Mecca. With the rise of the Indonesian nation-state, the *kijaji* was drawn into new forms of translocal relations. As Geertz phrased it, while before he was a "scholar, curer, and mystic teacher, he is now a politician," albeit "an amateur" one (1960: p. 247). For both Geertz and Wolf, a new form of ethnographic attention emerged that

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12178-6

was concerned with understanding complex systems and the relations between different levels of sociocultural integration. There were explicit disciplinary concerns as well. For Geertz, the focus on the broker was an attempt to find a space between the predominance of community-based studies in anthropology and political science's concern with elites.

The rise of modernization theory during the 1960s positioned the new nations within a developmental framework that considered the broker as a necessary but temporary actor that would disappear with the rise of new rational organizational forms. In the wake of Geertz and Wolf, there were attempts to develop and circumscribe the broker conceptually along these lines. Sydel Silverman, following Wolf, highlighted two criteria: first, 'critical functions' that are of 'direct importance to the basic structures of either or both systems,' and, second, 'exclusivity,' meaning that the links between two systems were monopolized by particular types of individuals (1965: p. 173). In her historical account of Central Italy, Silverman showed how certain patrons came to dominate a wide range of mediating functions between the rural community and the national system, but that after World War II many of these functions were taken over by the state, as the expansion of literacy and education allowed for locals to engage directly with bureaucracy. As such, 'critical junctures' were no longer guarded by one group. In Silverman's account, there is thus a historical relationship between the position of the patron and the broker, as the former takes on the role of the latter during a certain period. Deep-seated relations of inequality – not only based on control of economic resources and familial relations of debt, but also on access to education and the ability to engage with the state – thus formed the basis for new brokerage opportunities. In this context, and in line with modernization theory, the hypothesis was that the broker emerged in an early stage of the development of the nation-state, only to later become unnecessary with the process of integration of local and national levels of society.

The rise of transactionalism – a theoretical position first advanced by Fredrik Barth – focused on the strategic actions of individuals in contexts of social constraint. It was therefore not surprising that leading proponents such as F.G. Bailey, Robert Paine, and Jeremy Boissevain turned their attention to brokerage. In keeping with the methodological individualism of transactionalism, Bailey and Paine attempted to move beyond a model of brokerage based on inequality, such as that outlined by Silverman, to one that considered the nature of the transaction as a starting point. Bailey's *Strategems and Spoils* broadened the discussion considerably by understanding brokers as bridging a gap in communication in a 'situation of encapsulation' (1969: p. 167). In *Patrons and Brokers in the East Arctic* (1971), Robert Paine and his collaborators were, on the one hand, concerned with the Canadian government's expansion into the Arctic, and its effect on indigenous groups, and the role of brokers in this process. On the other hand, Paine attempted to develop a theory of patronage and brokerage that focused on roles rather than statuses based on inequality. He argued that roles were situational and that brokerage should be understood as the regulation of the circulation of values. Boissevain, in turn, identified the broker as "a professional manipulator of people and information who brings about communication for a profit" (1974: p. 148). He argued that centrality of location in networks, time to deal with social relations, and control over first-order resources were all critical in becoming a successful broker (pp. 154ff). The broker's capital lay in the communication channels he controlled, while his credit was what others believed his capital to be (p. 159).

Transactionalists' focus on communication shifted attention more directly toward the formal elements of brokerage as well as the figure of the entrepreneur, thus attending more to economic relations. This was perhaps unsurprising with the rise of rural capitalism in the areas in which anthropologists were concerned. Boissevain's claim that 'brokerage is business' (1974: p. 158) stated this most clearly. But transactionalism's concern with formal relationships also allowed for the potential of bringing the economic and the political into a common frame, thus highlighting the entrepreneurial aspects of politics rather than stable relationships. While the African headman had often been considered a victim of tensions between different groups, the strategic actions of the transactionalist broker highlighted a particular problem, namely that the broker is not able to justify his actions in normative terms. If manipulation rather than the resolution of conflict was at the center of the broker's craft, it was not surprising that he was often despised and considered dishonest. More generally, this pointed toward issues that became increasingly important as anthropologists shifted their attention away from villages toward complex societies, namely issues of trust and interpersonal relations across sociocultural borders.

The Death of the Broker

By the late 1970s, however, the broker had begun to disappear as a focus of concern for anthropology. The question of power had from the beginning been at the heart of well-known debates between transactionalists and Marxists. Even sympathetic critics such as Cohen and Comaroff claimed that although it was important to distinguish the broker from patron, the broker role embodied the essential attribute of power: "the capacity to construct and purvey meanings concerning a variety of relationships and interactions" (1976: p. 89). More importantly, they claimed that political relationships were often based on the creation or mediation of boundaries of meaning rather than transactions within coherent systems of meaning, a point that illuminates contemporary discussions in anthropology and a more sophisticated understanding of the workings of power.

Deborah James (2011) understands the disappearance of the broker in relation to what Jonathan Spencer (2007, Chapter 1) calls the 'strange death of political anthropology,' in which the analysis of local-level politics ran into a 'dead end' with the rise of structuralist (particularly Marxist) models of political analysis and a broad-based understanding of the postcolonial state as the primary site of power and domination in the Third World. There was perhaps also a more subtle shift of attention from the concern with particular persons, such as patrons and brokers, to particular situations that allowed for new forms of relations and brokerage (Vincent, 1978: p. 186). Arguably, however, a series of further trends reinforced the waning interest in the broker. As post-structuralism emerged

with full force during the 1990s, the focus was broadly placed on how power productively disciplined and shaped individual subjectivities, as questions concerning agency were effectively sidelined. In this world, the local-level broker appeared increasingly insignificant in both empirical and analytical terms. In the meantime, with the reflexive critique of anthropology that emerged primarily in the United States during the 1980s, the broker arguably lost relevance precisely because of the crisis of cultural holism, which came to complicate the question of brokerage and how it might be conceptualized. As hybridity came to replace identity, the cultural forms of encapsulation that had formed the basis for an analysis of brokerage were effectively critiqued and disbanded. Finally, during the same period, a technologically centered concern with mediation, primarily in the wake of Benedict Anderson's seminal *Imagined Communities* (1983), came to dominate scholarly debate, as an anthropology of media gained ground (Boyer, 2012). The question of the relationship between mediation and brokerage is one that will be further discussed at the end of this article.

It is worth noting contrary developments within the discipline of history. Richard White's (1991) influential *The Middle Ground: Indians, Empires, and Republics in the Great Lakes Region, 1650–1815* offered a nuanced understanding of relationships between settlers and Native Americans through a focus on brokers. Interestingly, White begins his book with an epigraph from the historian of anthropology, James Clifford, concerning the ongoing critique of the culture concept, thus offering a perspective on the American frontier that moved beyond the 'billiard ball' theory of cultural relations toward one highlighting struggles for accommodation and common meaning. More recently, historians influenced by or trained in anthropology have considered brokers as a starting point for considering imperial governance. Rothman (2011), in her study of brokers in sixteenth-century Venice, successfully combines anthropology and history by illustrating how brokers licensed by the Venetian state did not merely mediate between different categories of actors – that is, 'Venetians' and 'Turks' – but were in fact active in producing the categories of identity themselves by various means and arrangements.

The Return of the Broker

In the past decade, there has arguably been a renewed concern with the broker in anthropology. Transformations in the global economy since the collapse of Bretton Woods and ensuing neoliberal reform have increasingly problematized state-centered models of power and sovereignty and pushed for a reconceptualization of the relationship between state and market. It is in this context that we can note a great potential for the return of the broker in ethnographic terms as an actor that emerges and becomes positioned along these fault lines (James, 2011). Furthermore, technological innovations in transportation, communication, and media have created a transnational landscape that has formed the backdrop for anthropology. It has become increasingly clear that the metaphor of transnational flow needs to be specified empirically and that brokers are critical in this process, not only on the local level but also among the transnational elites that have increasingly become a familiar focus for anthropologists. As noted earlier, in the past few decades a broader focus on mediation has largely displaced the concern with brokers. With the rise of digital media and the internet, in particular, mediation has moved to the center of anthropological attention. As science and technology studies has come to replace cultural studies as a primary interlocutor for anthropological theorizing, however, perspectives such as actor–network-theory have led many anthropologists to problematize the distinction between human and nonhuman agency, thus allowing for a renewed discussion concerning the relationship between brokerage and mediation, and of agency, more generally.

In contrast to the fantasies of ideologues who imagine a world of disintermediation characterized by free markets without borders, in the contemporary moment it appears that brokers are proliferating rather than becoming obsolete. In fact, with neoliberal reform and economic and political deregulation, the broker appears as a critical figure, indeed, one might argue, as an ideal anthropological informant. In the global south, James notes as much in relation to transitional settings such as post-apartheid South Africa, where "state, market, and patrimonial/patriarchal-style political authority intersect" (2011: p. 318). She sees the reemergence of brokers who can mediate between these spheres, in her case in the context of land reform. In what follows, I will identify a few other domains in which the broker has emerged as a site of concern.

Current Trajectories

As tourism has become one of the world's largest industries in recent decades, it has also become a key site of interest for anthropologists. In the relationship between 'hosts' and 'guests,' various forms of brokers, most notably the tour guide, have attracted the attention of anthropologists. In the new anthropology of tourism there is, for instance, the recognition that brokers do not mediate between already existing cultures, but are rather active in producing and encapsulating cultural authenticity, and of commodifying and selling these experiences and images. Macdonald's (2006) work on tour guides in a 'difficult heritage' site, the former Nazi Party Rally Grounds in Nuremburg, highlights one productive site in ethnographic terms, in which guides attempt to control interpretation in different ways. But tourism itself is not easily circumscribed, and the rise of volunteer and retiree tourism points to relationships between leisure and labor, and the rise of new zones of contact and brokerage that are capturing the interest of anthropologists. The rapidly expanding global industry for tertiary education should also be noted in this context, as the recruitment of students depends on brokers, many of them former students themselves.

The global rise of nongovernmental organizations (NGOs), particularly in the context of the development aid industry, but also in transnational activism, has led scholars to consider the role of brokers in these processes. Although Olivier de Sardan (2005, Chapter 11) distinguishes between agents who work directly for donors and rent-seeking development brokers, it is perhaps more fruitful to highlight his point that there is an a priori normative distinction – which has limited empirical

support – between grassroots do-gooders and corrupt actors. Of particular interest in the emerging literature on development brokers is *Brokers and Translators* (Mosse and Lewis, 2006), which is an attempt to combine an actor-oriented approach rooted in the Manchester school of anthropology with a Latourian concern with translation in actor–network-theory. Rather than considering brokers as strictly mediating between aid donors and recipients, the authors ask how unpredictable development projects are actually implemented through the translation of interests, concepts, and practices, and how particular forms of meaning are established in this process.

The anthropology of transnational migration has tended to focus on sites of departure or destination. While social networks and diasporic communities have been widely researched, the commercially oriented migrant broker has been vilified since the advent of liberalism, an ethical position that has often been reproduced in the scholarly literature. This discourse is particularly evident in discussions concerning human trafficking in which there is an enduring distinction between 'victim' and 'perpetrator.' However, in the context of Asia and the Middle East, in particular, there is an ongoing turn from undocumented to documented transnational migration, in which the rise of licensed recruitment agencies has been matched by increasing demands for state documentation. In this process, brokers both convince people to become migrants and negotiate paperwork in relation to the state bureaucracy (Lindquist et al., 2012). As in ongoing discussions concerning the rise of NGOs, neoliberalization and the shifting relationship between the state and market are both creating the opportunity for new types of brokers – who engage both with local populations and bureaucracies – while reproducing an a priori ethical discourse. In an important sense, brokerage has been an ethnographic black box in both the development and migration industries, a fact of which anthropologists are increasingly cognizant.

In the examples noted earlier, we see how the moral ambiguity of the broker – as an actor whose primary commitment appears as maximizing individual gain and is therefore deemed untrustworthy – has remained constant across time and space (James, 2011: p. 319). This ethical position is perhaps most evident in the burgeoning work on the transnational market in organ transplants, in which Nancy Scheper-Hughes (2003), in her brand of militant anthropology, names particular brokers in the 'rotten trade.' On the other hand, it is also possible to see how this position has become problematic in ethnographic terms in relation to the earlier discussion. For instance, Lawrence Cohen (2003) speculates about the rise of 'bioethical brokers' who come to control the discourse surrounding organ transplants. These include scholars and policy-makers, who are certainly often sincere in their intentions, but who should nonetheless be considered as epistemic brokers. Cohen's ethnographic focus thus illustrates a broadened anthropological concern that makes evident multiple fields of brokerage across previously disconnected domains.

Another field in which one can note a renewed interest in the broker is in the anthropology of finance, which has shifted attention toward the economic elites that are at the center of the global economy. Cailtin Zaloom (2006), for instance, in her study of trading floors in Chicago's and New York's financial districts, pays close attention to various types of 'economic men' who broker trades, as financial markets have moved from floor-based to electronic trading. From an anthropological perspective, it is clear that the 'global economy' is made up of actors who are engaged in making actual trades and that this should be considered ethnographically. But the importance of technologies and algorithms in finance has also meant that it is within the anthropology of finance, as well as the anthropology of science, that actor–network-theory has been most influential. The attempt to do away with the distinction between human and nonhuman agents has opened up for a more radical understanding of brokerage. In this context, Latour makes a critical distinction between intermediaries and mediators. The intermediary "transports meaning without transformation" while "mediators transform, translate, distort, and modify the meaning or the elements they are supposed to carry" (2005: p. 39). In fact, it is upon this distinction that Latour makes a contrast to other forms of sociology. From the perspective of actor–network-theory, there exist an endless number of mediators, and thereby brokers, who thereby are positioned at the center of analysis and as a starting point for considering various processes of encapsulation and the production of social forms. A recent example of anthropologists using actor–network-theory is the discussion by Bill Maurer et al. (2013) on the rise of the 'mobile money agent' who functions as a kind of 'human ATM' by facilitating monetary transactions for users via mobile phones. As they note, there are ongoing discussions within the mobile money industry concerning whether the agent is an intermediary and mediator, a classic question in anthropology. Rather than focusing attention on the broker, the mobile money agent itself, Maurer and his colleagues use the industry's interest in the agency of the broker as an entry point for considering ongoing discussions about 'agency' and money, once again illustrating how anthropologists are increasingly considering particular kinds of experts as knowledge brokers.

Finally, it is worth noting the anthropology of media. In Boyer's (2012) review of the field, he notes a shift of attention away from media as a form to mediation as a process, which has in part led to an increasing focus on producers (2012: p. 386), thus suggesting that one question for anthropology in the future is precisely the relationship between brokerage and mediation. In this context, the anthropological interest in journalists is notable, as in Ulf Hannerz's (2004) ethnography of foreign correspondents, and in Zeynep Devrim Gürsel's (2012) work on 'image brokers' in news journalism in the era of digital media. In both cases, anthropologists are concerned with how journalists broker news and images in a process of world-making, thus once again showing how in the contemporary era brokers are less – once again borrowing F.G. Bailey's term – encapsulated than engaged in processes of encapsulation.

Final Thoughts: Ebb and Flow

From one perspective, the ebb and flow of interest in the broker can be considered in relation to the changing nature of the state. While former colonies were transformed into nation-states after World War II, since the 1980s neoliberalization has both transformed the state and modes of governance.

It is therefore worth considering if interest in the broker both reflects a crisis of political authority and a mode of anthropological theorizing in the face of such a crisis. But a major difference between the current era and the earlier one is that while previously it was generally clear what the broker was mediating between, the village and the national capital, or, more generally, encapsulated political structures, in the contemporary context of unbounded fieldsites this is not immediately obvious. This is arguably more of a possibility than a constraint as anthropologists are able to avoid the pitfalls of methodological individualism and consider the broker as an ethnographic entry point that illuminates broader contexts and processes from a particular position of mediation. As such, the renewed focus on the broker moves beyond a static conceptualization of the relationship between the 'local,' 'national,' or 'global,' thus understanding scale as emergent in relation to the ethnography at hand and aiming to set things in motion rather than to resolve or put an issue to rest. The problem of how to describe and conceptualize what the broker mediates between is thus primarily ethnographic, as is the more specific question if the broker is better understood as an intermediary or mediator – actors or entities that make a difference or no difference in the transaction, respectively (Latour, 2005).

See also: Anthropology: Overview; Big Man, Anthropology of; Borders, Anthropology of; Development: Social-Anthropological Aspects; Economic Anthropology; Exchange in Anthropology; Finance, Anthropology of; Globalization, Anthropology of; Inequality: Comparative Aspects; Migration: Anthropological Perspectives; Nongovernmental Organizations: Anthropological and Historical Aspects; Patron–Client Relations, Social and Anthropological Study of; Political Anthropology; Political Economy in Anthropology; State: Anthropological Aspects; Tourism.

Bibliography

Anderson, B., 1983. Imagined Communities. Verso, London.
Bailey, F.G., 1963. Politics and Social Change: Orissa in 1959. University of California Press, Berkeley.
Bailey, F.G., 1969. Stratagems and Spoils: A Social Anthropology of Politics. Blackwell, Oxford.
Boissevain, J., 1974. Friends of Friends: Networks, Manipulators and Coalitions. Basil Blackwell, Oxford.
Boyer, D., 2012. From media anthropology to the anthropology of media. In: Fardon, R., et al. (Eds.), Handbook of Social Anthropology. Sage, London, pp. 383–392.
Cohen, L., 2003. Where it hurts: Indian material for an ethics of organ transplantation. Zygon 38, 663–688.
Cohen, A., Comaroff, J.L., 1976. The management of meaning: on the phenomenology of political transactions. In: Kapferer, B. (Ed.), Transaction and Meaning: Directions in the Anthropology of Exchange and Symbolic Behavior. Institute for the Study of Human Issues, Philadelphia, pp. 87–107.
Fallers, I., 1955. The predicament of the modern African chief: an instance from Uganda. American Anthropologist 57, 290–305.
Geertz, C., 1960. The Javanese kijaji: the changing role of a cultural broker. Comparative Studies in Society and History 2, 228–249.
Gluckman, M., Mitchell, J.C., Barnes, J.A., 1949. The village headman in British Central Africa. Africa 19, 89–106.
Gürsel, Z.D., 2012. The politics of wire service photography: infrastructures of representation in a digital newsroom. American Ethnologist 39, 71–89.
Hannerz, U., 2004. Foreign News: Exploring the World of Foreign Correspondents. University of Chicago Press.
James, D., 2011. The return of the broker: consensus, hierarchy, and choice in South African land reform. Journal of the Royal Anthropological Institute 17, 318–338.
Latour, B., 2005. Reassembling the Social: An Introduction of Actor–Network-Theory. Oxford University Press, New York.
Lindquist, J., Xiang, B., Yeoh, B., 2012. Opening the black box of migration: brokers, the organization of transnational mobility and the changing political economy in Asia. Pacific Affairs 85, 7–19.
Macdonald, S., 2006. Mediating heritage: tour guides at the former Nazi Party Rally Grounds, Nuremburg. Tourist Studies 6, 119–138.
Maurer, B., Nelms, T., Rea, S., 2013. 'Bridges to cash': channelling agency in mobile money. Journal of the Royal Anthropological Institute 19 (1), 52–74.
Mosse, D., Lewis, D., 2006. Theoretical approaches to brokerage and translation in development. In: Lewis, D., Mosse, D. (Eds.), Development Brokers and Translators. Kumarian, Bloomfield, pp. 1–26.
Olivier de Sardan, J.P., 2005. Anthropology and Development: Understanding Contemporary Social Change. Zed Books, London.
Paine, R. (Ed.), 1971. Patrons and Brokers in the East Arctic. Social and Economic Papers, St John's, Newfoundland.
Parks, R.E., 1928. Human migration and marginal man. American Journal of Sociology 33, 881–893.
Press, I., 1969. Ambiguity and innovation: implications for the genesis of the culture broker. American Anthropologist 71, 205–217.
Redfield, R., 1956. Peasant Society and Culture: An Anthropological Approach to Civilization. University of Chicago Press, Chicago.
Rothman, N., 2011. Brokering Empire: Trans-Imperial Subjects between Venice and Istanbul. Cornell University Press, Ithaca.
Scheper-Hughes, N., 2003. Rotten trade: millennial capitalism, human values and global justice in organs trafficking. Journal of Human Rights 2, 197–226.
Silverman, S., 1965. Patronage and community-nation relationships in Central Italy. Ethnology 4, 172–189.
Spencer, J., 2007. Anthropology, Politics, and the State: Democracy and Violence in South Asia. Cambridge University Press, Cambridge.
Vincent, J., 1978. Political anthropology: manipulative strategies. Annual Review of Anthropology 7, 175–194.
White, R., 1991. The Middle Ground: Indians, Empires, and Republics in the Great Lakes Region, 1650–1815. Cambridge University Press.
Wolf, E., 1956. Aspects of group relations in a complex society: Mexico. American Anthropologist 58, 1065–1078.
Zaloom, C., 2006. Out of the Pits: Traders and Technology from Chicago to London. University of Chicago Press, Chicago.

Bruner, Jerome Seymour (1915–)

Keiichi Takaya, Faculty of Letters, Kokugakuin University, Tokyo, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Jerome Bruner is one of the most original and influential psychologists of all time. His work in psychology and cognitive science played a significant role in making the study of mind as both scientific and human. His interest and influence are also not limited to the field of psychology, but have been extended to related fields, most notably, education.

Life and Career

Jerome Bruner is one of the most original and influential figures in the history of psychology. Some of the key concepts and perspectives in today's psychology, cognitive science, and such related disciplines as education and anthropology are inspired by Bruner's work: for example, cognitive science, folk psychology, cultural psychology, and discovery learning. He has written numerous books and articles, and received awards and honorary degrees for his contributions to the understanding of the human mind. In addition to his work in psychology and cognitive science, he is known for his involvement in education, particularly in the curriculum reform movement in the 1960s and the early 1970s.

His work in psychology or cognitive science played a significant role in making the study of mind as both scientific and human. As part of the larger movement in overcoming the narrowly scientific – or scientistic – approach of psychology in the first half of the twentieth century, behaviorism and experimentalism, he liberally borrowed research methods and findings from various disciplines as well as devised ingenious psychological approaches. He also consistently inquired into the nature of the *human* mind – how it works, how it is formed, and how it can be helped to become what it is – from diverse perspectives.

Jerome Seymour Bruner was born on 1 October 1915, to an affluent Jewish family in New York. Except for his love of books, Jerome was not particularly precocious intellectually. Remembering back to his high-school days, he says that he loved French and mathematics, and also sailing with his close friend whom he calls 'water rats,' which gave him the 'first taste of collaboration' (1983b: p. 19).

He received his B.A. (psychology) from Duke University and Ph.D. (psychology) from Harvard. He had worked for a few governmental agencies, including the Foreign Broadcast Intelligence Service and the Office of Public Opinion Research, and engaged in surveys on propaganda and public opinions in the early 1940s before he turned to research and teaching in university. Later, toward the end of the War, he worked in Paris for the Anglo-American Psychological Warfare Division of Supreme Headquarters Allied Expeditionary Force in Europe (PWD-SHAEF). There he engaged in cultural relations and also witnessed the educational reform led by the physicist Paul Langevin. Through these experiences, he made connections with, and learned from, leading figures in various fields. An example of such encounters was his meeting with Robert Oppenheimer.

Since 1945, Bruner has taught at Harvard University, Oxford University, the New School for Social Research, and New York University (NYU) (he is currently University Professor in the School of Law, NYU). Through his expertise as psychologist and his extensive network of intellectuals and policy makers, he has been involved in various professional initiatives including the establishment of the Center for Cognitive Studies at Harvard in 1960. He also served in the Educational Panel of the President's Science Advisory Committee (in the Kennedy and Johnson years), and was involved, albeit indirectly, in the launching of Project Head Start.

For his contributions to the advancement of psychological research and to educational practice, he has been awarded the International Balzan Prize, the American Psychological Association's Award for Distinguished Scientific Contributions, and honorary doctorates from such universities as Yale, Columbia, and Sorbonne.

Bruner's Cognitive and Psychological Theories

Although Bruner is primarily interested in the formation and development of the human mind, the specific foci of his research have changed over the years: perception and thought in the 1940s and the 1950s (*A Study of Thinking*); intuition, invention, and mythology (*On Knowing: Essays for the Left Hand*); cognitive and linguistic development from the 1960s to the early 1970s (*Studies in Cognitive Growth*; *Child's Talk*); the nature of narrative and various forms of cultural activities (*Actual Minds, Possible Worlds*; *Acts of Meaning*; *The Culture of Education*; *Making Stories: Law, Literature, Life*) since the 1980s.

Borrowing from his Oxford colleague Isaiah Berlin's phrase, Bruner describes himself as "a fox rather than a hedgehog, preferring to know many things rather than one big thing" (1983b: p. 8). As a 'fox,' he does not limit his research to the field of psychology narrowly conceived. Instead, in pursuing the formation and operation of our mind, he liberally borrows insights, research findings, and alternative perspectives from various fields. Bruner writes, "I was … an intellectual first and a scientist in support. … I used psychology to pursue matters that existed for me in their own right. Psychology was (and remains) only one way to use mind in behalf of these pursuits" (1983b: p. 77). In return, his works are constantly referenced and taken as a source of inspiration in various other fields.

He also says that he has a strong 'left-handed' part, that is, a less purposeful, more intuitive character (1983b: p. 5). True

to his word, he has caused a few major reform movements in psychology and education. Through a series of research and publications in the latter half of the 1940s and the early 1950s, he and his colleagues (e.g., Postman and Bruner, 1946; Bruner and Goodman, 1947) launched a new approach to psychology called the 'New Look,' which paved the way for the Cognitive Revolution in the late 1950s. 'New Look' psychology and the Cognitive Revolution were revolts against the then-dominant behaviorism that focused exclusively on the correlation of sensory input and behavioral output without regard to the intermediate mental processes, and against forms of experimental psychology that specialized in animal study or decontextualized behavioral study of people in the laboratory. In other words, Bruner and the 'New Look' psychologists thought that behaviorism and experimentalism, by concentrating on the analysis of the input (stimuli) and the output (response) in order to maintain the scientific and objective appearance, lost sight of mind, while they were supposedly pursuing the science of mind. 'New Look' psychology joined forces with emergent innovative research of the late 1940s to the early 1950s in such fields as computer science (or cybernetics) and anthropology and brought about the Cognitive Revolution, restoring the mind as the center of interest. This movement for the revitalization of the study of mind led to the birth of cognitive science in the 1950s.

The term *cognition* means mental operations such as perception, thought, and memory, particularly of humans. We now tend to use *cognitive science* or *cognitive sciences*, instead of psychology, to imply that psychology alone, let alone the narrowly scientific approach of behaviorism, can enable us to understand the nature and development of our mind sufficiently. Thus, cognitive science is an interdisciplinary approach to the understanding of the human mind and explores what makes our mind work, including language, conceptual systems, cultural foundations of mental operations, whose scopes go well beyond the traditional field of psychology.

In the 1980s, however, Bruner became disenchanted with the Cognitive Revolution, because he felt that the movement lost the initial impetus that tried to restore the mind in the science of mind. He writes:

> The [cognitive] revolution was intended to bring 'mind' back into the human sciences after a long cold winter of objectivism … [But] that revolution has now been diverted into issues that are marginal to the impulse that brought it into being. Indeed, it has been technicalized in a manner that even undermines that original impulse. This is not to say that it has failed: far from it, for cognitive science must surely be among the leading growth shares on the academic bourse. It may rather be that it has become diverted by success, a success whose technical virtuosity has cost dear. Some critics, perhaps unkindly, even argue that the new cognitive science, the child of the revolution, has gained its technical successes at the price of dehumanizing the very concept of mind it has sought to reestablish in psychology, and that it has thereby estranged much of psychology from the other human sciences and the humanities.
>
> 1990: p. 1

A major cause of this growing disenchantment is the dominance of computer models in conceptualizing the mind. Bruner refers to 'an interesting turning point' in his life in the mid-1980s, when he 'got to thinking about agency more and more,' while, he thought, psychologists talked about mechanisms a lot without talking about human agency (Olson, 2007: p. 149).

Bruner started to focus on the study of cognitive development in the late 1960s; and when he moved to Oxford in 1972 he moved away from laboratories' controlled environments to more natural settings where real people interact. He launched a project in which he observed mother–infant interactions in the process of language acquisition (e.g., Scaife and Bruner, 1975; Bruner, 1983a).

Through his research in cognitive and linguistic development, he increasingly came to appreciate the role of culture. The significance of culture in equipping the human mind, especially the role of the narrative mode of language use, turned him to 'cultural psychology,' and some regards his cultural turn a second revolution in psychology (Bakhurst and Shanker, 2001: pp. 1–2). He now looks to cultural practice in understanding the essence of our humanness. Against the mechanical or computational conceptions of mind, he highlights agency and intentionality in our mental operations. But he also thinks that these traits are mere potentials in us that need to be enhanced by cultural tools and realized in cultural contexts. He says, "to understand man you must understand how his experiences and his acts are shaped by his intentional states, and … the form of these intentional states is realized only through participation in the symbolic systems of the culture" (Bruner, 1990: p. 33). Thus he has been inquiring into the way people understand themselves and the world, how they weave meanings through narratives, and the processes of their situated actions in cultural settings.

One may well recognize a thread in his seemingly divergent range of research from psychological experiments in laboratory to the observation of linguistic development in natural settings to the nature of narrative processes employed in legal procedures. It is, in short, his interest in what makes us human.

One may also notice that his style of work is very dialectic. As he himself admits, he likes starting up new programs without bothering to work out every detail to the end, and he leaves them to others. Barbara Rogoff, David Olson, and Patricia Greenfield are the examples of researchers who are working on such programs (Shore, 1997: pp. 17–20, 24; also, Bruner, 1980: p. 111). Just as he had earlier rejected the behavioral and experimental approaches to psychology and launched the Cognitive Revolution, a few decades later he attempted another new approach in cultural psychology. He keeps on finding new problems, new ways of looking at things, and keeps conversation going.

Theoretical Sources and Inspirations

In his autobiography, Bruner names John von Neumann (computer science), Herbert Simon (computer science), Noam Chomsky (linguistics), Jean Piaget (psychology), and Lev S. Vygotsky (psychology), among others as his intellectual heroes, because they prepared foundational ideas that recovered the study of mind from the dominance of the study of stimulus and response, conditioning, decontextualized behavior, and of environmental determinism (1983b: p. 60).

Besides these heroes, a special mention is due to his long-time colleague and friend George Miller and his idea of

cognitive constraints. In 1956, George Miller published an ingenious account of the constraints on our cognitive capacities. Miller's argument, along with others, such as Sir Frederick Bartlett's 'scheme theory' and Edward Tolman's idea of 'cognitive map,' gave Bruner the idea of the importance of strategy for humans, who had a very limited range of attention, perception, and memory, to process the overwhelming amount of information. Bruner discarded the traditional approach to the study of human cognition, whose typical example may be Hermann Ebbinghaus' study of memory that used the memorization of nonsense syllables. Bruner thinks that people like Ebbinghaus overlooked, first, the significance of the context in which such human cognitive functions as memory works, and second, the idea of agency seen in such aspects of human cognition as selection, organization, and strategy in memory and thought (1983b: pp. 97–98).

Bruner regards Jean Piaget as a great pioneer, even though he disagrees in details of Piaget's theory. He acknowledges Piaget's contribution in advancing our understanding of the developmental processes, because he replaced "[t]he old view that pictured the infant as 'random' in his actions and saw growth as constituting of becoming 'coordinated'" (1983a: p. 28).

One of the major ideas Bruner took from Piaget (and his collaborator Bärbel Inhelder) is that knowledge takes different forms at each developmental stage, and the acquisition of knowledge is not a process of incremental accumulation as behaviorism and associationism hold (Olson, 2007: p. 70). Unlike the Genevan psychologists, however, he does not think that the process of development occurs spontaneously. Here, Bruner sides with Vygotsky and Alexander Luria and thinks that the development is a result of enculturation; development happens outside-in rather than inside-out. And this view has important implications for Bruner's educational theory, as we will see later.

Somewhat similar to his attitude toward the Piagetian theory of development, Bruner develops his own theory of language acquisition by critically reviewing Noam Chomsky's nativist theory of linguistic development. Bruner believes that human competence in general is "both biological in origin and cultural in the means by which it finds expression" (1983a: p. 23). While he acknowledges, along with Chomsky, that there are some 'endowments' that provide "foundation processes that aid the child's language acquisition," he also thinks, unlike Chomsky, that none of these endowments generates language on its own. Thus Bruner advances the notion of the Language Acquisition Support System (LASS) to supplement Chomsky's Language Acquisition Device (LAD). Assuming that human beings are born with some propensity for linguistic communication, and that such a propensity may be efficient enough to be regarded as a LAD, Bruner says that it will not turn into a full-fledged language without a LASS, which is provided by the social world. "It is LASS that helps the child navigate across the Zone of Proximal Development to full and conscious control of language use" (Bruner, 1986: p. 77).

Bruner and Education

Outside the specialized field of psychology or cognitive science, people got to know Jerome Bruner through his educational theory and his involvement in the curriculum reform movement in the 1960s. His influence on education was most strongly felt in the 1960s and the early 1970s, and a good example is the fact that Project Head Start, which is now part of the American educational lexicon, is Bruner's brainchild, though he was not directly involved in the implementation of the program.

In the 1950s, a sense of educational crisis was already brewing in America. The center of the concern was declining academic standards in public schools. In the atmosphere of the Cold War and the scientific and technological rivalry between the United States and the Soviet Union, the American public felt that their public schools were not effective, or even failing, in educating citizens who were intelligent and knowledgeable enough to carry on the competition.

Among various factors that brought about the educational crisis, two should be mentioned. One was the dominance of progressivism, or a vulgarized form of progressivism that 'had degenerated into permissivism,' which allegedly caused the decline of academic content and standards (Olson, 2007: p. 65). The other had to do with the American tradition of local control of educational matters. That is, education professionals (teachers and local administrators), who were not particularly well informed in academic content areas or in learning theory, controlled the daily grind of schooling.

The so-called Sputnik Shock in 1957 added to their sense of crisis and led to the involvement of the federal government, researchers, and scientists in educational matters, and Bruner played a major role in this process. Bruner wrote *The Process of Education* as a chairman's report of the Woods Hole Conference (held in September 1959), which was organized by the National Academy of Science, and attended by a number of leading scientists and psychologists, to have a closer look at the new science curricula that had already been revised and implemented. The book was an instant success. While retaining the progressivist theme of students' involvement in their own learning, Bruner provided a theoretical framework to ensure also focused attention on academic content.

Bruner in *The Process of Education* presented a few innovative ideas, and they were (1) the importance of the structure of the discipline, (2) the idea of spiral curriculum, (3) the use of intuitive thinking along with analytic thinking, and (4) how students' desire to learn can be stimulated (the problem of motivation). In sum, the central idea that connects all these concrete proposals is seen in his remark "intellectual activity anywhere is the same, whether at the frontier of knowledge or in the third-grade classroom" (1977: p. 14).

Given the information explosion in the modern world, Bruner took it for granted that one had to acquire powerful learning and thinking skills and to acquire a taste for further, self-motivated, learning. The problem then is to figure out, first, how to motivate children for learning, and second, how effectively schools may equip students with powerful tools of knowing and thinking.

On the first task, Bruner thought that the traditional method of coverage was hopeless, and also thought that the Deweyan view of arousing children's motivation by connecting learning in the classroom with their daily experience outside school was not enough ("After John Dewey, what?" in Bruner, 1962). He

was also critical of the Piagetian view of readiness: waiting for children's motivation and capacity for learning to spontaneously emerge. He believed that the motivation can and should be stimulated and created, and that the best way to arouse students' desire to learn is to expose them to the world of real scientific or academic inquiry.

He thought that intellectual activity in young students was essentially the same as in first-rate scientists, and that the essential part of making study or inquiry self-rewarding was 'discovery.' Thus, the process of learning ought to be the process of discovery, and discovery of what he called the 'structure' of knowledge. In his view, mere accumulation of pieces of information will not give students either effective tools of thinking or motivation for learning. On the students' part, instead of being given disconnected, decontextualized pieces of information to memorize, students should better be encouraged to discover, as much as possible on their own, meaningful connections (structures) of knowledge. Each school subject or academic discipline has its own structure of knowledge, and structured knowledge has 'generativity' that would make the acquisition of further knowledge easier. Students were also encouraged to utilize different modes of thinking – 'intuitive' as well as 'analytic' – just like scientists, to first grasp a hypothesis by drawing on their hunches (albeit well-informed hunches; cf Bruner, 1977: pp. 56–57), and later confirm, revise, or reject the tentative conclusion by drawing on analytic thinking. These are essential features of real inquiry, which Bruner thought had unfortunately been neglected in school, and had better become the model for learning in school.

Young students, however, do not possess either the amount of knowledge or the skills of inquiry that research scientists do, and a sort of translation is necessary. Here, Bruner borrows from the Piagetian insight that the way children understand things is qualitatively different according to their age. So, the same topic can and needs to be presented to the child in different forms depending on what 'mode of representation' is dominant in the child, and the same topic will be covered at different timings (grade levels, etc.), using different 'modes of representation.' This is his famous idea of spiral curriculum; the same topic is learned by using different modes of representation and with increasing complexity as students' cognitive capacities grow.

In combining his beliefs that interest can be created and that there are ways to teach things to students effectively, Bruner makes a bold assertion in *The Process of Education*, "any subject can be taught to anybody at any age in some form that is honest." This is probably the most quoted and most sensational claim of his. I think that Bruner's point in this proposition is in the word 'honest.' It is true that progressive education in some cases did very well to engage children in school tasks, but the question still remained: are they really learning what's worth learning or what will enable them to understand the world and themselves better? Many answered this question in the negative and Bruner's proposals, encapsulated in this quote, seemed to offer the means to ensure the honesty part; that is, children not only have fun or excitement but also learn academically respectable content – in a word, academic 'excellence' (1960: pp. 9, 70; 1962: p. 119).

Bruner published a series of books on education in the 1960s and the early 1970s, starting with *The Process of Education*, which caused quite a sensation not only in the United States, but also around the world. His educational publications in those decades, in fact, inspired the curriculum reform movement. One of the projects in the movement was his famous *Man: A Course of Study* (MACOS), an innovative elementary social studies program. It was initially received favorably, but its anthropological approach stirred up some controversy, and MACOS was eventually removed from the public school curriculum (see Dow, 1991; for detailed accounts).

After a few decades, Bruner published in 1996 *The Culture of Education*. In this book, Bruner critically reviews some of his earlier positions by incorporating his later perspectives on cultural or folk psychology. For example, Bruner now rejects the idea of deprivation that underlined Head Start (the view that impoverished children lacked something that needed to be filled) though he still endorses the overall program (Bruner, 1996: p. 80; cf also, Geertz, 2001: p. 22).

Another change might be found in his reference to Dewey. Bruner's comments on Dewey were very critical in the 1960s and, many people regarded him as someone to replace Dewey. Now, however, he speaks more appreciatively of Dewey, and many, including Bruner himself, regard – I think rightly – him as an heir to the great American tradition of pragmatism (Olson, 2007: pp. 64, 161).

Today, Bruner's influence on educational discourse and practice does not seem as direct as it used to be in the 1960s. While in his earlier career, he was much more concerned with implementing his educational ideas (as in MACOS), his educational work in more recent years is rather descriptive and his influence indirect. Many educators continue to get inspirations and borrow insights from Bruner, but very few try to actually implement his ideas as he did in the 1960s and the 1970s (Takaya, 2013).

Conclusion

Today, Bruner's presence is not felt as strongly as it used to be. Nevertheless, in the hands of his former students and colleagues, the theoretical seeds he planted continue to grow in various ways. A few examples may be cultural psychology, the notion of school as a community of learners, and the significance of narrative in understanding the world and ourselves. Also his former students and colleagues have been making significant impacts on education and psychology (e.g., Howard Gardner, David Olson). And most amazingly, as Gardner, one of his former students, comments, Bruner, at well past his 90s mark, still is the "youngest and most eager child in his class" (Gardner, n.d.).

See also: Bruner's Theory of Cognitive Development; Cognitive Development, Educational Theories of; Cultural Psychology; First Language Acquisition, Developmental Psychology of; Piaget's Theory of Cognitive Development; Vygotsky's Sociocultural Theory; Vygotsky's Theory of Cognitive Development.

Bibliography

Bruner, J.S., Goodman, C., 1947. Value and need as organizing factors in perception. Journal of Abnormal and Social Psychology 42, 33–44.

Bruner, J.S., Postman, L., 1949. On the perception of incongruity: a paradigm. Journal of Personality 18, 206–223.

Bruner, J.S., Goodnow, J.J., Austin, G.A., 1956. A Study of Thinking. Transaction Publishers, New Brunswick, NJ.

Bruner, J.S., 1960/1977. The Process of Education. Harvard University Press, Cambridge, MA.

Bruner, J.S., 1962. On Knowing: Essays for the Left Hand. Harvard University Press, Cambridge, MA.

Bruner, J.S., 1966. Toward a Theory of Instruction. Harvard University Press, Cambridge, MA.

Bruner, J.S., 1971. The Relevance of Education. W.W. Norton and Company, New York.

Bruner, J.S., 1974. Beyond the Information Given: Studies in the Psychology of Knowing. Selected, Edited and Introduced by Anglin, J.M.. W.W. Norton & Company, New York.

Bruner, J.S., 1980. Jerome S. Bruner. In: Gardner, L. (Ed.), A History of Psychology in Autobiography, vol. VII. W.H. Freeman and Company, San Francisco, CA, pp. 75–151.

Bruner, J.S., 1983a. Child's Talk. W.W. Norton and Company, Inc., New York.

Bruner, J.S., 1983b. In Search of Mind: Essays in Autobiography. Harper and Row, New York.

Bruner, J.S., 1986. Actual Minds, Possible Worlds. Harvard University Press, Cambridge, MA.

Bruner, J.S., 1990. Acts of Meaning. Harvard University Press, Cambridge, MA.

Bruner, J.S., 1996. The Culture of Education. Harvard University Press, Cambridge, MA.

Bruner, J.S., 2002. Making Stories: Law, Literature, Life. Harvard University Press, Cambridge, MA.

Bruner, J.S., Amsterdam, A.G., 2002. Minding the Law. Harvard University Press, Cambridge, MA.

Bruner, J.S., 2006. In Search of Pedagogy: The Selected Works of Jerome S. Bruner, 2 vols. Routledge, London and New York.

Bakhurst, D., Shanker, S.G., 2001. Jerome Bruner: Language, Culture, Self. SAGE Publications, London.

Cole, M., Scribner, S., 1974. Culture and Thought: A Psychological Introduction. John Wiley and Sons, New York.

Dow, P.B., 1991. Schoolhouse Politics: Lessons from the Sputnik Era. Harvard University Press, Cambridge, MA.

Gardner, H., 1987. The Mind's New Science: A History of the Cognitive Revolution. Basic Books, New York.

Gardner, H., n.d. Reflections on Jerome "Jerry" Bruner. 29.09.14: http://insidetheacademy.asu.edu/reflections-on-jerome-jerry-bruner (retrieved on 29.09.14.).

Geertz, C., 2001. Imbalancing act: Jerome Bruner's cultural psychology. In: Bakhurst, D., Shaker, S.G. (Eds.), Jerome Bruner: Language, Culture, Self. SAGE Publications, London, pp. 19–31.

Miller, G., 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review 63, 81–97.

Miller, G., March 2003. The cognitive revolution: a historical perspective. Trends in Cognitive Sciences 7 (3), 141–144.

Olson, D.R., 2007. Jerome Bruner: The Cognitive Revolution in Educational Theory. Continuum, New York and London.

Postman, L., Bruner, J.S., 1946. The reliability of constant errors in psycho-physical measurement. The Journal of Psychology 21, 293–299.

Scaife, M., Bruner, J.S., 24 January 1975. The capacity for joint visual attention in the infant. Nature 253, 265–266.

Shore, B., 1997. Keeping the conversation going: an interview with Jerome Bruner. Ethos 25 (1), 7–62.

Takaya, K., 2013. Jerome Bruner: Developing a Sense of the Possible. Springer, Dordrecht, Heidelberg, New York, and London.

Bruner's Theory of Cognitive Development

Keiichi Takaya, Kokugakuin University, Shibuya-ku, Tokyo, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Jerome Bruner is one of the founding fathers of cognitive science, an interdisciplinary approach to the study of the human mind. Cognitive science originally grew out of researchers' discontent with the behavioral and experimental approaches in the early twentieth century. By integrating findings in various fields such as psychology, philosophy, anthropology, neuroscience, information theory, and linguistics, they now understand how the human mind develops and works from diverse and richer perspectives. This article examines Bruner's contribution to the development of cognitive science and major features of his theory of cognitive development.

The Birth of Cognitive Science

Cognitive science grew out of psychology in the mid-twentieth century, and its founders, including Jerome Bruner, generated it out of their discontent about the way psychology was conceived and practiced – what the central concerns of psychology ought to be and how psychological problems were conceptualized and examined. In criticizing existing forms of psychology, cognitive science developed as an interdisciplinary approach to 'long-standing epistemological questions' (Gardner, 1987: p. 6) in such diverse disciplines as psychology, philosophy, anthropology, computer science, information theory, linguistics, and neuroscience. Researchers interested in the study of the human mind today tend to use the designation *cognitive science* or *cognitive sciences*, instead of psychology, to imply that psychology alone, let alone the narrow approach of behaviorism, can enable us to understand the nature and development of the mind adequately.

When cognitive scientists discuss cognition, they usually focus on *human* cognition. When discussing cognition, Bruner, for example, means distinctively human forms and processes of perceiving, thinking, and learning (1983b: p. 122). Howard Gardner (1987), who was one of Bruner's assistants, adopts the same criteria in writing arguably the most comprehensive study available on the history of cognitive sciences; he writes, "Though the term cognitive science is sometimes extended to include all forms of knowledge – animate as well as inanimate, human as well as nonhuman – I apply the term chiefly to efforts to explain human knowledge" (p. 6). Michael Cole and Sylvia Scribner (1974) in their introductory textbook to cognitive science also write that their focus in researching and writing about cognition is the acquisition, transformation, and use of information by humans (Chapter 1). Cognitive scientists today go beyond the traditional scope of psychology to examine various elements that constitute our cognition, such as language acquisition, concept formation, metaphor, social and cultural foundations of our mental operations, etc.

The sources of cognitive scientists' discontent about the then-dominant type of psychological research – behaviorism and experimental psychology – were two; first their units of analysis were too small, and second, they were too close to environmental determinism. In short, they felt the absence of the mind, the very essence of humanness, left much of this work impoverished.

On the first point, behavioral and experimental approaches avoided questions of mind because it is not observable (a bit like a drunkard who is looking for his lost car key under the streetlamp precisely because he can see things there). Accordingly, those approaches missed essential features of human perception and thinking; an example of this was the idea of selectivity in perception or of strategy of thought (cf Bruner and Postman, 1949: p. 206; Bruner, 1983b: p. 276).

For Bruner, and his like-minded colleagues, a stimulus is not equivalent to perception, and human cognition is not reducible to the largely predetermined patterns of stimuli and responses observable in animal studies. Human beings have what may be appropriately called minds, and they can act on their environments to open up new possibilities.

Bruner began to use at least hints of such ideas as agency and purposiveness as early at his undergraduate years at Duke University. There he took courses (besides those taught by his mentor, William McDougall) from Donald Keith Adams and Karl Zener who had learned Gestalt psychology from Wolfgang Köhler, Max Wertheimer, and Kurt Lewin. Two books that influenced him at that time in this same direction were Frederic Bartlett's *Remembering* (1932) and Edward Tolman's *Purposive Behavior in Animals and Men* (1932).

Another inspiration came from Robert Oppenheimer in 1951, when Bruner was at the Institute for Advanced Study on sabbatical. Oppenheimer remarked, that "units of analysis in science usually turn out to be much larger and more inclusive than one at first expects." Bruner and his coauthors came to appreciate 'the appropriateness of his remark' as they worked on *A Study of Thinking* (Bruner et al., 1956: Preface, p. xix). The appropriate units of analysis they found were not stimulus and response in individual instances, but the 'overall patterns of responding' (see the quote from Gardner below) that is, the process of perception or thought infused with selectivity and strategy. In coming to this view, they

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23095-X

borrowed from such neurophysiological concepts like 'gating,' 'tuning,' 'models,' and 'intentions' (Bruner, 1983b: p. 99). Howard Gardner writes on the innovation that Bruner and his coauthors brought about in the book:

> [Bruner, Goodnow, and Austin] ended up analyzing the informational properties of long sequences of acts called "strategies." It had turned out that the best way to account for individual performances was in terms of these overall patterns of responding over many trials rather than of particular responses to a particular stimulus configuration. This was the most iconoclastic – and the most influential – aspect of the work. (1987: p. 94)

In a famous study on the perceived sizes of coins, Bruner and Goodman (1947) showed that the felt value of the coin affected the perception of the size of the coin; children from poor families tended to overestimate the sizes of coins, compared to children from affluent families. Perception, therefore, is not raw data. It is infused with selective mechanisms, intentions, and expectancies; in short, what we call the 'mental.' Bruner elsewhere remarks, "Stimuli … do not act upon an indifferent organism… The organism in perception is in one way or another in a state of expectancy about the environment" (Bruner and Postman, 1949). We may notice that positing things mental, subjective, or intentional is a critique of environmental determinism that behaviorism inherited from the classical empiricism or sensationism of such philosophers like John Locke and Claude Adrien Helvetius, distant predecessors of 'scientific' psychology against which Bruner and his colleagues were revolting.

For Bruner, it is obvious that '(m)ind in some operative form is there from the start,' and the more important task for psychology is to analyze 'the conditions that produce human minds that are richer, stronger, and more confident' (1983b: p. 152). As he went on with his research into the 1980s and the 1990s, he increasingly shifted his focus from the processes working inside the individual mind toward more cultural, dialectical types of mental formation and operation. The central thread of his research is, however, clear; he finds excessive objectivism, reductionism, straightforward causations, and determinism in contemporary psychological research, and tries to restore a richer conception of what it means to be human.

Bruner, looking back on the mid-century when he and George Miller started the Center for Cognitive Study, says:

> I suppose everyone has a somewhat different notion of what the "center" of psychology is. We had our idea. We thought it was concern for the distinctively human forms of gaining, storing, transforming and using knowledge of all sorts – what makes humans human. Call them the cognitive process. (1983b: p. 122)

Miller (2003) also remarks, by quoting Noam Chomsky, that they thought, "defining psychology as the science of behavior was like defining physics as the science of meter reading." People like Bruner, Miller, and Chomsky used 'cognitive' in their center's name "defiantly, when most respectable psychologists at the time still thought cognition was too mentalistic for objective scientists" (Bruner, 1983b: p. 124).

In searching for the essence of our humanness (i.e., in searching for mind as the title of his autobiography suggests), he has consistently highlighted such 'mentalistic' concepts as intentionality and strategy, but his focus in his early career seems to have been largely on the *intra*psychic functions, so to speak. In his later career, he shifts his focus toward more *inter*psychic aspects of the mind such as culture and intersubjectivity. Bruner's cognitive theory, particularly since the 1980s, is marked by its cultural approach, because he thinks that we become human by and through culture. He emphasizes that the human mind becomes what it is by internalizing cultural tools – among which language is prominent – available in the society where one is born and to live. He says, "To treat the world as an indifferent flow of information to be processed by individuals each on his or her own terms is to lose sight of how individuals are formed and how they function" (1990: p. 12). Influences of Vygotskyan psychology are evident here, but he also integrates views from anthropology, philosophy, linguistics, etc.

By emphasizing the importance of 'outside-in,' however, he does not go to the extreme of environmental determinism of Locke and Helvetius, or of James Watson and B.F. Skinner. Bruner also emphasizes, again without going to the other extreme of nativism which Jean Piaget's or Noam Chomsky's theories may do, that there are some predispositions that lead human infants naturally toward interaction and communication with people around them and learning from cultural resources. In Bruner's view, human nature and culture are seamlessly connected, and this is a view which David Bakhurst develops in recent years, albeit critically, by drawing on the philosophy of John McDowell as well as on Bruner, as an idea of culture as 'second nature' (Bakhurst, 2001; particularly, pp. 188–190; cf also, Bakhurst, 2011).

Major Topics in Bruner's Developmental Theory

In order to understand distinctive marks of Bruner's theory of cognitive development, it helps to have a look at his concrete ideas. In the following, I will review some of his major concepts and topics: (1) the idea of three modes of representation, (2) language acquisition, (3) the concepts he favors after his 'cultural turn,' and (4) constructivism.

Modes of Representation

Somewhat similar to Jean Piaget's idea of developmental stages, Bruner says that there are three distinctive modes of representation, enactive, iconic, and symbolic, and individuals acquire these modes one by one, in the order of enactive to iconic to symbolic, until in the end they become capable of all the three modes of understanding. Very young children who have not mastered their native language understand and communicate through behavior; then, they start using concrete

images; and finally they begin using symbols and can now engage in abstract thinking without necessarily relying on concrete objects or images. Bruner says:

> Human beings represent their knowledge of the world in three ways. One is through habit and action: knowing what to do. A second is through imagery: depiction of events and relations. Finally, we "know" things by representing them in a symbolic system like language or mathematics. Making progress in the mastery of any domain most often involved using all three modes of representation – typically progressing from enactive, through iconic, to symbolic. Indeed, understanding could be deepened by using all three modes, even (and sometimes especially) when they were in conflict with each other. If some modes were easier for the child than others, then the way of helping him across the "zone of proximal development" was to start with it and then move on to others. (1983b: p. 142)

The modes of representation have educational implications just like Piaget's stage development theory does. In teaching young children, teachers have to carefully observe which mode is dominant in the child they are teaching. Knowing the nature of these modes of representation and individual children's dominant modes would allow teachers to effectively convey what they have to teach.

Bruner's modes of representation, however, are not the same as Piaget's developmental stages, because while Piaget saw cognitive development as spontaneous, that is, as a result of internal or physiological maturation, Bruner sees it "as the consequence of learning to deal with the world in terms of progressively more advanced cultural forms and cultural modes of representation" (Olson, 2007: p. 69).

As to the causes of development, Bruner is unmistakably Vygotskyan, as clearly seen in his reference to the Vygotskyan concept of the zone of proximal development in the quote above, for he does not think that human cognitive development is 'spontaneous'; for him, psychological/cognitive development is not simply a function of physiological maturation. It is true that human beings are uniquely predisposed to picking up 'tools' from their surrounding culture to make them their own psychological tools, but, nonetheless, internalization of cultural tools must take place for psychological development to take place.

Having said that, however, it should be noted that Bruner shares an idea of cognitive development with Piaget and Bärbel Inhelder in that they consider development as a process of transformation rather than a mere accumulation of information. Bruner is in agreement with Piaget on the importance of this transformation, but unlike Piaget, he thinks that the moving force of the development is culture, in which language is prominent. Bruner wrote, "learning to encode the world linguistically and then to operate on language rather than on the world – that was the ultimate stage of cognitive development" (1983b: p. 161; cf also, Olson, 2007: p. 21). Bruner's theory of cognitive development is developmental in a Piagetian sense to the extent that it has a sense of direction in which development ought to go and involves a notion of qualitative changes, but it is also Vygotskyan in that the development is brought about by the acquisition of cultural tools.

Language Acquisition

While acknowledging that human beings have innate endowments for language acquisition, Bruner does not go so far as Noam Chomsky's nativism. Tomasello writes:

> In opposition [to Chomsky's Language Acquisition Device, LAD], Bruner posited that human beings come into a world that is already structured culturally and linguistically, and their long period of immaturity is designed precisely for them to acquire the particular cultural and linguistic conventions into which they are born. (2001: p. 36)

One of the endowments that human beings are born with is their propensity for shared attention. Scaife and Bruner (1975; also, Bruner, 1983b: p. 170) showed that infants, even in their first year of their lives, could direct their attention to common objects with their caretakers. They also found that in engaging in such a play like peekaboo, infants can quickly learn to reverse hider and seeker roles with their mothers. Clearly, human infants are predisposed to treating others as minded creatures, interacting with others, and communicating. Their findings are also quite inconsistent with Piaget's picture of very young children as very 'egocentric' and of 'egocentricity' being overcome by language acquisition (Tomasello, 2001: p. 36). Bruner says:

> In the end, I came to the conclusion that the need to use language fully as an instrument for participating in a complex culture (just as the infant uses it to enter the simple culture of his surround) is what provides the engine for language acquisition. The genetic "program" is only half the story. The support system is the other half. (1983b: p. 173; cf also, Geertz, 2001: p. 21)

Besides the traditional focus on syntax ('well-formedness') and semantics ('(language's) capacity to refer and to mean'), Bruner highlights the importance of pragmatics in studying language acquisition, that is, "how to get things done with words" (1983a: pp. 17–18). While he thinks that the three facets are interdependent and inseparable, he, by emphasizing the significance of pragmatics, highlights the point that "language acquisition 'begins' before the child utters his first lexico-grammatical speech" (ibid., p. 18). Language is one of the means of 'getting things done,' though it is a very powerful one.

Roughly speaking, in the 1970s (when he was at Oxford University), Bruner studied the process of linguistic development in infants. In the 1980s, he broadened his research from linguistic development in the individual to linguistic practices in wider culture (Tomasello, 2001: pp. 33–34). Since his youth, Bruner has consistently refused the classical view on the nature of human cognition that regards perception or sensory input as raw data and thought as an addition; for Bruner, psychology is not a mere addition to physiology. Human mind becomes what it is by participating in culture and developing its potential through culture. Thus, for Bruner, and as is acknowledged widely by cognitive scientists and philosophers today, psychology cannot be constructed on the basis of the individual alone (Bruner,

1990: p. 12). It then makes sense that, in the 1980s, after he had studied human cognition as a property solely of individuals, he shifted his attention increasingly toward culture's role in the formation of our cognition; how language functions in cultural practices. Thus, he now explores the uses and significance of narratives in meaning making (and his most recent research interest is in how narratives work in the context of legal procedures).

Another important turn his research brought about was a perspective on intersubjectivity, that is, how people understand what others are perceiving and thinking ('Did you see what I saw?'; cf Olson, 2007: p. 25). Thanks to the research on shared attention, we now know that human beings at a very young age already 'know' others are minded beings. This is again a refutation of the Piagetian belief on children's egocentricity and the notion of decentering as development, and shows how humans are predisposed to communication and participation in culture.

This perspective on intersubjectivity has important implications for education. Piagetian theory of children's egocentricity suggests that children must acquire knowledge of the world directly, unmediated by language, and on their own, unassisted by more experienced members of society. According to Bruner, this position underestimates the resources available for even very young children and makes the task of education unnecessarily difficult. Children are naturally prepared to enter the world of shared meanings, and they will later be assisted by language. Bruner says, "Most of our approaches to the world are mediated through negotiation with others" (1986: p. 68), and intersubjectivity is the key to entering the 'negotiation with others.'

Culture and Narrative

Even though Bruner has consistently appreciated how culture shapes the mind, he was more cautious about the constraints culture might impose on our thoughts, values, and perspectives in the 1960s and the 1970s. That's why he thought it best to incorporate the most sophisticated form of cultural activities – first-rate research in natural and social sciences – into school curricula when he led the curriculum reform movement in the United States in the 1960s (cf his famous ideas of 'discovery learning' and 'the structure of disciplines'; see my article 'Jerome Bruner' for more). In recent years, however, he is more trusting in the resources available in our familiar, everyday culture ('folk culture') for the development of our cognitive capacities. He writes in 1986:

> It would not be an exaggeration to say that in the last decade there has been a revolution in the definition of human culture. It takes the form of a move away from the strict structuralism that held that culture was a set of interconnected rules from which people derive particular behaviors to fit particular situations, to the idea of culture as implicit and only semiconnected knowledge of the world from which, through negotiation, people arrive at satisfactory ways of acting in given contexts. The anthropologist Clifford Geertz likens the process of acting in a culture to that of interpreting an ambiguous text. (p. 65)

While he used to appreciate the power of scientific and academic aspects of cultural activities for the purpose of cultural transcendence, he is now more trusting of more mundane aspects of culture as well.

The end of cognitive development, in his more recent formulations, is the acquisition of capacities with which people not only function well in the social and natural environment but also are able to, if necessary, transcend the limit that their culture or society imposes. He used to think that this kind of capacity was attained best through acquiring the knowledge and skills embodied in academic disciplines (what he calls the 'paradigmatic' mode of thought), but he now thinks that resources available in our folk culture (the 'narrative' mode) are no less potent.

Culture embodies 'a set of values, skills, and ways of life' and it is the 'toolkit' for sense making and communicating (Bruner, 1996: p. 3). Cultural tools enhance our presumably natural endowment in action, perception, sense making, and thought (Bruner, 1966: p. 126; cf also Bruner, 1971: p. 53). Culture provides tools and contexts with which individual members make sense of and appraise incidents and phenomena. As such, on the one hand, it imposes a constraint on endeavors of the individual mind; just like our native language imposes a certain limit on the way we think and communicate. On the other hand, culture does not lack resources for the individual to transcend to some extent the limits imposed either by the putative human native endowment or by culture itself. Bruner points out, for example, that all known languages have ways to indicate 'the might-or-might-not-be, the *possible*' (in Olson, 2007: p. 5. emphasis in original). He also says that people who are adept at various modes of language, such as metaphor and irony, seem to be much less constrained by the syntactical structure of the language (Bruner, 1996: p. 18).

It seems, then, very natural that he is now more interested in exploring the possibilities of the latter, particularly the uses of narrative in legal procedure.

Constructivism

The epistemological position of constructivism is not Bruner's invention. Bruner's name, however, is often associated with it. Constructivism, in a nutshell, holds that what we may call reality, truth, or knowledge is not something fixed, eternal, or objective in a strictest sense, but it is in fact what we construct according to our particular perspective or purpose. Bruner writes:

> The constructivist view, that what exists is a product of what is thought, can be traced to Kant, who first fully developed it. Kant, in turn, attributed his insight to Hume's discovery that certain relations among things in the real world could not be attributed to events but rather were mental constructions projected onto an "objective world." (1986: p. 96)

Unlike such predecessors' assumptions, however, constructivists today argue that our cognition cannot be neatly divided into raw sensory data and thought, and our knowledge of the world cannot be founded on incorrigible foundations – there is no such thing as raw data. Accordingly, knowledge is not a mere mirroring of things 'out there' (Bruner, 1971: p. xi). Rather, Bruner believes that knowledge

is 'justified belief' and a model we construct for specific purposes (1996: p. 59). Although he does not take the position of free-for-all relativism or extreme subjectivism (ibid., p. 61), he firmly believes that knowledge is revisable and accommodates multiple interpretations. Knowledge of the world or what we know as the 'reality' of the world is a tentative result of negotiations between our individual perspectives and more-or-less stable, established systems of meanings that traditions and cultures have established. Knowledge, therefore, should rather be understood in terms of meaning making, which is inseparable from communicating (1996: p. 3). In refusing the assumption that knowledge has incorrigible foundations and in adopting the criteria of perspective and purpose in determining what is right or wrong, Bruner inherits the pragmatism of William James and John Dewey, and is in agreement with such contemporary philosophers as Nelson Goodman (one of Bruner's colleagues at Harvard's Center for Cognitive Study) and Richard Rorty.

Piaget and Vygotsky

Although Bruner's theory, particularly his theory of cognitive development in later years, is more akin to Vygotsky's than to Piaget's, he does not discard Piaget's perspectives completely; Piaget is still one of Bruner's intellectual heroes (1983b: p. 60). In an essay in *Actual Minds, Possible Worlds* (1986), he critically appraises the psychological theory that pictures very young children as 'egocentric' and the process of development as a process of decentralization in the light of his findings on the emergence of intersubjectivity in early childhood (Chapter 4).

Considering, for example, the ease with which very young children master 'deictic shifters,' which require the grasp of the viewpoints of the person with whom one is interacting (e.g., 'here' used by me is close to me, while 'here' used by you is closer to you), the Piagetian picture of children as 'self-centered' is inadequate. Bruner then describes the problems of the major premises of Piagetian view, for example, its view that the individual mind encounters the world unmediated. Nevertheless, Bruner does not forget to add that these premises are rather arbitrary and partial than completely wrong. He says, "they are true under certain conditions, false under others, and their 'universalization' reflects cultural bias" (ibid., p. 61).

Unlike Piaget, Vygotsky does not think that cognitive development is automatic or something that happens without regard to cultural and social contexts. Bruner appreciates Vygotsky for his recognition of the role of cultural context in human development (1983b: p. 139). From a Vygotskyan perspective, what brings about children's cognitive development is cultural input and social interaction as much as internal, physical maturation. As such, it is crucial that children are engaged with the people and the world around them. You just do not wait for their interest to show up; you can and should create and stimulate interest (Bruner, 1962: p. 117).

Though Bruner is not without some disagreements with Vygotskyan psychology (Olson, 2007: p. 152), his cognitive and developmental theories are, on the whole, closer to Vygotsky's than to Piaget's, particularly in more recent years.

Cultural Turn

As Bakhurst and Shanker (2001) point out, Bruner was involved in two revolutionary movements in the study of the human mind (Bruner et al., 1956: Preface, p. xix). In the 1950s, when the dominant psychological approaches disregarded the mind almost altogether and focused on the correlations of stimuli and behaviors, Bruner and others incorporated conceptual schemes and research findings from various fields to restore the concept of the mind and launched cognitive science.

A few decades later, however, Bruner came to notice that cognitive science was now so caught up with computational models – ironically, given that these had earlier been among the moving forces to restore the notion of mind – and instigated a cultural turn by once again examining the characteristics that make up our minds. Bakhurst and Shanker write:

> [The Cognitive Revolution] restored to psychology the inner life of the mind after decades of arid behaviorist objectivism … But cognitivism had become so enamoured of conceptual computational models of the mind that it had replaced behaviorism's impoverished view of the person with one no better; human beings as information processors. In response, Bruner argued forcefully that meaning is not a given, but something made by human beings as they negotiate the world. (2001: pp. 1–2)

While Bruner has always been attracted to the notion of culture since the earliest days of his psychological research, he highlighted what he calls the 'paradigmatic' part of culture such as sciences as practiced by first-rate researchers who were equipped with well-structured knowledge of their fields. In more recent years, however, he finds at least equally potent resources in more mundane aspects of culture. He says:

> Well, I think the two principals that have changed are, first, I became much more mindful in the 1970s about alternative modes of thought. These modes of thinking are all narrative and they are all narrative not only in thinking but in the way we justify and structure our culture. And second, I have become much more deeply concerned about the nature of human culture and the way in which culture shapes individuality and the way in which individuality works back into the culture.
>
> Those are the two big changes as I shift toward narrative. I think I would add a third change, an emphasis on agency. Narrative, culture, and agency. (Bruner in Olson, 2007: p. 175)

Bruner has consistently highlighted the importance of the way the human mind goes beyond the information given as well as the importance of recognizing how our minds are shaped by social interaction and culture. In this way, he tries to overcome the limitations of behaviorism and its logical implication, that is, environmental determinism, and retain the idea of human agency.

Conclusion

Bruner characterizes his own approach to the study of mind as follows:

> … I have always thought of myself as an 'intellectual' first, and then a psychologist. An intellectual is somebody who follows his ideas where they may lead without regard to disciplinary boundaries and who, having done so, accepts the consequences of his conclusions in how he thinks, feels and acts personally. (1983b: p. 230)

In my view, this is precisely what brought about the cognitive revolution in the mid-twentieth century and made him a central figure in that movement that revolutionized psychology or the study of mind. As he describes himself as having a strong 'left-hand' – that is, intuitive rather than analytic – tendency, he likes to open up new ways of thinking and start up new programs of research without bothering to pursue every single detail and implication to the end; he simply leaves that to others. These topics are now in the hands of his former students, younger colleagues, and new generations of researchers (such as Howard Gardner, David Olson, and David Bakhurst), and they in turn keep the Brunarian spirit alive and explore the topics in a new light.

See also: Cognitive Development, Educational Theories of; Cultural Psychology; First Language Acquisition, Developmental Psychology of; Piaget's Theory of Cognitive Development; Vygotsky's Sociocultural Theory; Vygotsky's Theory of Cognitive Development; Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Bakhurst, D., 2001. Memory, identity, and the future of cultural psychology. In: Bakhurst, D., Shanker, S.G. (Eds.), Jerome Bruner: Language, Culture, Self. Sage Publications, London.

Bakhurst, D., 2011. The Formation of Reason. Wiley-Blackwell, Oxford.

Bakhurst, D., Shanker, S.G., 2001. Jerome Bruner: Language, Culture, Self. Sage, Publications, London.

Bruner, J.S., 1962. On Knowing: Essays for the Left Hand. Harvard University Press, Cambridge, MA.

Bruner, J.S., 1966. Toward a Theory of Instruction. Harvard University Press, Cambridge, MA.

Bruner, J.S., 1971. The Relevance of Education. W. W. Norton and Company, New York.

Bruner, J.S., 1983a. Child's Talk. W. W. Norton and Company, New York.

Bruner, J.S., 1983b. In Search of Mind: Essays in Autobiography. Harper and Row, New York.

Bruner, J.S., 1986. Actual Minds, Possible Worlds. Harvard University Press, Cambridge, MA.

Bruner, J.S., 1990. Acts of Meaning. Harvard University Press, Cambridge, MA.

Bruner, J.S., 1996. The Culture of Education. Harvard University Press, Cambridge, MA.

Bruner, J.S., Goodman, C., 1947. Value and need as organizing factors in perception. Journal of Abnormal and Social Psychology 42, 33–44.

Bruner, J.S., Postman, L., 1949. On the perception of incongruity: a paradigm. Journal of Personality 18, 206–223.

Bruner, J.S., Goodnow, J.J., Austin, G.A., 1956. A Study of Thinking. Transaction Publishers, New Brunswick, NJ.

Gardner, H., 1987. The Mind's New Science: A History of the Cognitive Revolution. Basic Books, New York.

Geertz, C., 2001. Imbalancing act: Jerome Bruner's cultural psychology. In: Bakhurst, D., Shaker, S.G. (Eds.), Jerome Bruner: Language, Culture, Self. Sage Publications, London.

Miller, G., March 2003. The cognitive revolution: a historical perspective. Trends in Cognitive Sciences 7 (3), 141–144.

Olson, D.R., 2007. Jerome Bruner: The Cognitive Revolution in Educational Theory. Continuum, New York and London.

Scaife, M., Bruner, J.S., January 24, 1975. The capacity for joint visual attention in the infant. Nature 253, 265–266.

Tomasello, M., 2001. Bruner on language acquisition. In: Bakhurst, D., Shaker, D. (Eds.), Jerome Bruner: Language, Culture, Self. Sage Publications, London.

Buddhism

David N Gellner and Richard Gombrich, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D.N. Gellner, volume 2, pp. 1378–1386, © 2001, Elsevier Ltd.

Abstract

Buddhism began a humanistic and individualist set of practices for salvation, a missionary religion which, in the course of its 2500-year history, has encompassed meditation, philosophy, ritual, art, ethnonational identities, and politics. The interaction and changing relationship between Buddhism's monastic leaders (monks, nuns) and the laity, and in particular kings or other secular authorities, has been a leitmotif of Buddhist history.

Buddhism is one of the world's three great missionary religions (the others being Christianity and Islam). It has assumed a great variety of social and ritual forms as it has spread throughout Asia and (since the end of the nineteenth century) the world. It began as a humanist and individualistic soteriology (religious doctrine of salvation) that preached the abandonment of ordinary householder life and its associated rituals. Monks and nuns renounce work and worldly activity; lay people support them in return for spiritual and ritual services. Whether the forms of ritual and religious life that developed for lay people should be seen as fully Buddhist or as some kind of 'second best' is an important part of debate over the interpretation of Buddhism between textual and more fieldwork-oriented scholars, as well as among Buddhists themselves. In both the modern and premodern eras, Buddhism has played an important but changing role in defining national identity in Tibet, Sri Lanka, Thailand, Cambodia, Laos, and Burma. In Japan, after 1868, the state attacked Buddhism as foreign and antinational, without being able to destroy it entirely. Everywhere revitalization movements in the modern period have tended to stress the rationalist and egalitarian parts of the Buddhist message; meditation has, for the first time, become a mass phenomenon practiced as much by lay people as by monastics.

Of the world's three great missionary religions (Buddhism, Christianity, and Islam), Buddhism is the oldest. It was one of a number of universalist soteriologies that emerged in the Gangetic Plain of north India in the sixth and fifth centuries BCE along with cities and kingdoms. It spread throughout Asia between 250 BCE and CE 1200. Thereafter, it began to go into decline in some places, eventually dying out in India, the land of its birth (except in the Kathmandu Valley, Nepal); in other places it went through cycles of decline and revival, culminating in newly invigorated and often innovative revivals in the modern era. In the nineteenth century – partly because of perceived similarities to and contrasts with Christianity – it proved to be powerfully attractive to many Westerners, an attraction that continued and grew in the twentieth century. But there are also underlying differences between Christianity and Buddhism which, when overlooked, lead to serious misunderstandings.

From a sociological viewpoint, Buddhism can be seen as an individualist and humanist soteriology (salvation religion) that developed a variety of communal forms and worldly orientations in different settings. It has legitimated many elaborate hierarchies; in modern times, many Asian intellectuals interpret it as an early form of egalitarian socialism. For many, meditation is the heart of Buddhism (Carrithers, 1983); in other places and times, elaborate ritual has been, and remains, more important. Some see vegetarianism as central to Buddhism; yet many Buddhists eat meat. Everywhere Buddhism is centrally concerned with death and salvation; but in some places, notably (but by no means only) in East Asia, Buddhist temples have become deeply involved in providing rituals for worldly protection and winning personal advantages. Most Buddhists take very seriously the notion of rebirth in accordance with the morality of one's actions; but in East Asia, the influence of Confucianism means that moral notions are usually divorced from the idea of rebirth.

The Buddha, His Teaching (Dharma), and His Community (Sangha)

Buddhist tradition is unanimous that the Buddha was born in Lumbini in southern Nepal and died, at the age of 80 years, in Kushinagar, India. The Theravada tradition maintains that he died in 543 BCE, and this date has become widely accepted by Buddhists: in 1956 the 2500th anniversary of the Buddha's death was commemorated in Nepal with delegations from around the world. Western historians of Buddhism have long rejected this date as too early and thus many textbooks give dates between 486 and 460 BCE. Recently, it has been argued that the correct date is probably around 400 BCE.

'The Buddha' means 'the awakened one'; the historical Buddha's family name was Gautama and according to later sources his personal name was Siddhartha ('he who has achieved his aim'). Buddhists believe that in principle anyone can become enlightened and teach others, and that there have been and will be many other Buddhas in different world ages and in different worlds. According to the tradition, the Buddha's father knew from a prophecy that his son might abandon his kingdom to become a great spiritual teacher and therefore isolated him within the palace, attempting to surround him with nothing but pleasure and ease. One day, however, he managed to escape and saw an old man, a sick man, and

a corpse. These 'three signs' alerted him to the existence of suffering and death, and he resolved to seek a way beyond them. He left the palace at night, cut off his own hair, abandoned his clothes, sent away his servant and horse, and set out to seek enlightenment. He spent 6 years experimenting with different teachers, before coming to see that mortification of the body was as bad as self-indulgence. He realized that a Middle Way was necessary – eating once a day, enough to maintain the body without overindulgence – and this led him to attain Enlightenment in Bodh Gaya, India, at the age of 35 years. Being enlightened – an aim that he shared with other renouncers of the time – means escape from the otherwise endless cycle of rebirths.

He gave his first sermon, to his first five disciples, in Sarnath, near Banaras in India, on the Four Noble Truths: the existence of suffering, the arising (cause) of suffering, the end of suffering (nirvana), and the path to achieve it. Nirvana is the 'blowing out' of the passion, hatred, and delusion with which we are on fire (and from which we suffer). The metaphor of fire refers to Brahmanical thinking: the Buddha took over much of the Brahmans' vocabulary while radically changing its meaning by ethicizing and universalizing it (Gombrich, 1988). The path to nirvana consisted of three stages: morality, meditation, and wisdom. In order to practice the latter two stages, and to attempt to attain enlightenment in this life, it was (until modern times) usually thought necessary to renounce the world, that is, to abandon, like the Buddha himself, the life of a householder, and live with other celibates: monks or nuns. The role of lay people was to practice morality and to support monks and nuns; by doing so they could improve their karmic state and attain better rebirths in future, eventually going on to attain enlightenment themselves. Later, many Buddhists came to believe that the world had declined to such an extent that even monks and nuns could not attain enlightenment, and they too should seek better rebirths.

At his death, the Buddha entered full nirvana and was not reborn. His remains were cremated and the ashes were divided between eight different lay groups and enshrined in *stupas*. Later, the remains of other prominent Buddhist clerics were similarly enshrined and these places, or smaller *caityas* modeled on them, became the characteristic places of Buddhist worship.

To facilitate a life of self-control and meditation, the Buddha founded his Sangha or Monastic Community. All Buddhists 'take refuge in' the Three Jewels, which are the Buddha, the Dharma (his Teaching), and the Sangha. Members of the Sangha are governed by 227 Vinaya rules (in the Theravada tradition; other traditions have 250 or more). These rules cover the correct behavior of monks; the nuns have their own Vinaya, with even more rules. This way of life appealed mainly to those of high status: although there were low-caste converts within the Sangha, high castes predominated. It was, as Max Weber (1958[1916–17]: p. 371) remarked, a 'genteel' (*vornehm*) soteriology.

Every fortnight monks are supposed to come together to confess any infractions of the rules and then to recite the rules together. However, in Theravada Buddhism there was originally little enforceable hierarchy. Monks do not promise to obey anyone. Deference is owed on the basis of seniority to monks who have been ordained for longer, and to one's teachers, but that is all. Ecclesiastical hierarchies have emerged in many Buddhist countries, but they are not prescribed in the Vinaya.

Despite the fact that, in practice, monks frequently maintain close ties with their kin and often carry on secular occupations, the institution of celibacy was to have weighty consequences for societies that adopted the Buddha's teaching. In Tibet it is estimated that before the 1950s 10–12% of the male population were monks, and a further 10–15% were noncelibate Buddhist practitioners. In Tibet and the Theravada countries, the history, as written in the premodern period, was the history of the Sangha. The issue of monastic property, and its transmission, is important in all Buddhist countries, with a variety of solutions being adopted.

By the standards of his day, the Buddha's teaching was individualistic, humanist, egalitarian, and rationalist. It was individualistic in that individual persons had their own karmic trajectory, which they could work to better by good actions or to harm by bad. There was no question of any higher power or other person granting salvation. It was humanist in that it debunked the Brahmans' claims to higher purity and spiritual status on the basis of their birth; it denied the efficacy of ritual (especially animal sacrifice), ecstatic trance, or worshipping gods in order to obtain salvation. Only one's own efforts will do. It was egalitarian in that all human beings have an equal chance to start out on this path (and humans are superior to gods in this respect: gods, though powerful, have to be reborn as humans to have a chance of achieving nirvana). It was rationalist in that the Buddha favored openness (in contrast to the secretiveness of the Brahmans, whose learning was available only to other Brahmans); and he stressed the need to try out doctrines for oneself and a rational ethic of helpfulness to others.

Thus, modernist reformers in the nineteenth and early twentieth centuries could relatively easily represent Buddhism as a rational and scientific belief system, as a philosophy and not a religion, as unconcerned with the existence of God or gods, and as implying social reform and development. However, most Buddhists have believed in gods, have practiced ritual in order to overcome misfortune and in order to benefit the dead, have regarded their faith as a religion, and have accepted socioreligious hierarchies. All the evidence shows that these practices are as old as Buddhism itself, and the modernist interpretation of Buddhism should be viewed as just that: as modernist and as an interpretation.

The Spread of Buddhism

At the beginning, Buddhism was just one of many different schools of renunciatory (i.e., renouncing earthly pleasures) philosophy. With its universalism and its teaching of careful attention, it seems to have appealed to townspeople and merchants, and to have spread with them throughout the Indian subcontinent. It emerged to prominence through the actions of the Emperor Ashoka (reigned c.268–239 BCE) who ruled most of present-day India. Though he protected all religions within his realm, his inscriptions advocate the Buddhist virtues of nonviolence, hard work, and consideration for others. He may well have seen Buddhism as a means of binding his enormous empire together and he sent out Buddhist

missionaries in all directions, including to West Asia (the Middle East). Buddhist tradition remembers Ashoka as the ideal Buddhist ruler and he became the model for many subsequent rulers attempting to reform the Sangha.

As Buddhism spread, different schools and ordination traditions emerged naturally, simply through geographical separation. The Buddha's teachings were preserved orally. This explains their repetitious character: they were only preserved because monks learned to recite them, and preservation of the canon quickly became a prime raison d'être ('reason for existence') of monastics. Writing probably did not yet exist in the Buddha's day. The earliest compilations of teachings were not systematically written down until the first century BCE, but long before that they had been organized into three 'baskets' (*pitaka*): the Sutta Pitaka, preserving the Buddha's teachings; the Vinaya Pitaka, preserving his rulings on the way of life of the Sangha; and the Abhidhamma Pitaka, which claims to present the teaching literally and systematically, and dates from after Buddhism was already established. (On internal evidence some of the Sutta Pitaka and much of the Vinaya Pitaka must also postdate the Buddha himself.) From time to time monks would come together to chant and check on each other's recitations. The early Buddhist councils were about fixing the Buddhist canon and debating minor differences that had arisen over monastic discipline. As far as the Buddhist laity was concerned, the central teachings were encapsulated in narratives about the Buddha's life, and particularly those concerning his 550 past lives contained in scriptures known as Jatakas and Avadanas; many of the latter were composed much later than the canon.

Among the many early Buddhist traditions, one was the Theravada, meaning 'doctrine of the elders,' so called because they considered themselves the most conservative. This is now found in Southeast Asia and Sri Lanka. Other schools had more adherents in India, especially in the northwest, from which China and Tibet received their missions, but Buddhism in what is now Pakistan, Kashmir, and Central Asia died out after the arrival of Islam. Consequently, Theravada Buddhism is the only form of pre-Mahayana Buddhism to survive into the modern world.

Although there are some small disagreements, the vast mass of Theravada scriptures, preserved in the Pali language, are shared and agreed upon in the Theravada countries. Pali was close to the local vernaculars of the Buddha's environment and the Buddha rejected use of the Brahmans' holy language, Sanskrit, both because of his policy of openness and because Sanskrit was held to be unchanging throughout time, which fundamentally contradicted the Buddha's teaching of impermanence. But as Buddhism spread, and as the spoken languages of north India evolved, Pali quickly became a purely scriptural and liturgical language, just like Sanskrit, and lay people became dependent on monks to expound the teaching to them.

Mahayana Buddhism

Around the turn of the Common Era an entirely new form of Buddhism arose in northwest India. It called itself the Mahayana. This is conventionally translated as 'The Great Vehicle' but equally means 'the great path or way.' It introduced a whole raft of new scriptures, now composed in a form of Sanskrit, which was supposed to date from the time of the Buddha himself; some claimed to have been hidden by him until people would be ready for them. The Mahayana reinterpreted the term *bodhisattva*, which for non-Mahayanists essentially means 'a future Buddha.' Mahayana Buddhism posited the existence of numerous powerful, celestial *bodhisattva*s. The most important of these was Avalokiteshvara, the embodiment of compassion, who would appear to people in their hour of need and save them from disaster. In China he came to be thought of as female and became the goddess Kwan Yin (Kannon in Japan). The Mahayana, unlike earlier forms of Buddhism, was most definitely based on a cult of the book: by this time scriptures were holy objects to be enshrined and worshipped as well as read.

The Mahayana introduced a revolution into the moral code of Buddhism, because it encouraged all Buddhists, whether monks or lay people, to aspire themselves to become Buddhas: this is called 'taking the *bodhisattva* vow.' Selfless service to others is part of this path; thus non-Mahayana Buddhists are said to be following the 'Hinayana' or Lesser Way. Great stress is laid on the Buddha's doctrine of 'skill in means,' i.e., adapting the teaching to the level of the hearers. This, together with the philosophical teaching of the essential 'emptiness' of all worldly phenomena, acted as a charter for developing new forms of ritual to attract lay Buddhists and non-Buddhists onto the path. These doctrines also served to legitimate the relaxation of some monastic rules, e.g., the ban on monks' handling gold and silver.

Mahayana Buddhism did not reject the pantheon, the scriptures, or the monastic discipline of earlier Buddhism; it simply added further scriptures and objects of worship. Two of its most influential scriptures were the Lotus Sutra (Saddharma-Pundarika) and the Perfection of Wisdom (Prajna-Paramita). It was primarily in this form that Buddhism then spread to China, Tibet, Mongolia, Japan, Korea, and northern Vietnam.

Vajrayana (Tantric) Buddhism

There was a still later current that is known as Vajrayana or Tantric Buddhism. This is based on scriptures known as Tantras, which appeared in India between CE 400 and 900 primarily within Hinduism and Buddhism. (In fact, there were several successive waves of Buddhist Tantras.) These scriptures were esoteric, placed great emphasis on ritual, and used magical and antinomian symbols and practices, which, they claimed, enabled the spiritual seeker under the guidance of a guru to achieve rapid results not possible by other methods. The earlier forms of Tantric Buddhism spread to China and Japan (and are still carried on in Japan by the Tendai and Shingon sects). Later forms of Tantrism, once found from India to China, survive today only as part of Tibetan and Nepalese (Newar) Buddhist traditions. The latest Tantric Buddhist scripture, the Kalachakra, can be dated to the early eleventh century and, very unusually, contains numerous references to Islam.

The development of Tantric (Vajrayana) Buddhism encouraged the emergence of Buddhist priests, that is to say, full-time ritual specialists devoted to the religious needs of lay people, who are frequently married and pass on their position to their sons. Such specialists are found in Tibet, Nepal, and Japan. The ideology of Tantric Buddhism attaches paramount importance

to the guru. He is frequently included alongside the Three Jewels of Buddhism or identified with the Buddha himself. In order to receive the full Tantric teachings it is necessary to take special initiations (Skt. *diksa*); celibacy is not a prerequisite; in fact, given the sexual nature of much of the symbolism and some of the practices, it may be an impediment. Thus Tantric Buddhism introduced a criterion of spiritual hierarchy that was sometimes in conflict with the monk/lay distinction, and at the very least served to undermine it. The necessity of a guru introduced a clear rationale for hierarchy among clerics. However, Tantric Buddhism, just like Mahayana Buddhism before it, did not reject what had gone before, but rather added to it; Tantric Buddhism therefore became the esoteric heart of a highly complex and layered set of traditions. Different schools and teachers disagreed on what exact place to give it. Monastic practice continued to be important in most forms of Buddhism. Tantrism has often been interpreted as a 'tribal,' peripheral, or 'shamanic' form of religion: however, whatever its origins, within Buddhism Tantrism was a highly sophisticated, learned, and scholastic tradition that attempted to tap some of the power of popular practices, both for salvation and for worldly benefits, while maintaining an elite distance from them.

Later Developments

In East Asia whole sects of Mahayana Buddhism grew up devoted to a particular scripture, a development that did not occur in India or Tibet. Thus, in Japan there are several sects whose primary allegiance is to the Lotus Sutra, practicing a devotional Buddhism, while others, e.g., the Shingon, have as their core practice the Tantric ritual of fire sacrifice, and yet others (the various Zen sects) focus on meditation. However, the Tendai attempts to teach and combine all the scriptures, while holding the Lotus Sutra to be the final and highest truth. In Japan, until the Meiji restoration, most Buddhist practitioners were celibate, except those of the Jodo Shinshu, founded by Shinran (1173–1263). This was one of the most radical sects focused on the worship of Amida, the Buddha of the Western paradise (Pure Land): simply uttering obeisance to Amida was all that was required for salvation; all other rules (including celibacy) were beside the point. A strong opponent of Shinran was Nichiren (1222–82), who advocated exclusive devotion to the Lotus Sutra and mounted vituperative attacks on Buddhists who did not agree with him.

In Tibet yet another innovation, dating from the thirteenth century, was the institution of reincarnate lamas (teachers), of whom the most famous is the Dalai Lama. The first Dalai Lama lived in the sixteenth century; the present Dalai Lama is the fourteenth. When a Dalai Lama dies, after the interbirth period, his successor is sought according to the indications of the state oracle. The boy chosen is believed to be both the reincarnation of the previous Dalai Lama and of the *bodhisattva* Avalokiteshvara, the special protector of Tibet. There are many other reincarnate lamas throughout Tibet and in neighboring areas where Tibetan Buddhism is practiced.

The Contemporary Situation

Currently, Theravada Buddhism is practiced in Burma, Thailand, Laos, Cambodia, and Sri Lanka. Tibet was and is a stronghold of both Mahayana and Vajrayana Buddhism, with both married and celibate practitioners. Chinese Buddhism was predominantly non-Tantric Mahayana of various sorts, as are Korean and Vietnamese Buddhism. In East Asia generally different sects grew up on the basis of different Mahayana scriptures; thus, in Japan there are both Tantric (Shingon, Tendai) and non-Tantric Mahayana schools, and even a tiny school (the Ritsu) based entirely on the practice of the monastic code. Thus when new religious movements arose in this century in Japan, naturally some were strongly Buddhist: some, like Soka Gakkai and Reiyukai, have spread through missionary activity to many different countries. The Tibetan diaspora has carried Tibetan Buddhism around the globe as well, and Theravada Buddhists have also proselytized widely. There have also been attempts to build a global ecumenical Buddhist movement, though it will prove difficult to decide who should lead it.

There is a striking contrast between the Theravada countries plus Tibet, where Buddhism became the dominant cultural tradition and came eventually to define a nationality, and East Asia, where Buddhism, though at periods dominant and favored by the elite, never achieved unquestioned acceptance. In China, Japan, and Korea the influence of Confucianism and Daoism meant that Buddhists often had to defend celibacy against the argument that it undermined society; by contrast, the spiritual value of avoiding marriage and reproduction is rarely questioned in South Asia and Buddhist Southeast Asia.

Buddhism and Political Authority

Although a minority of the first generations of monks chose to wander for 9 months of the year, most later monks and nuns settled in monasteries, working at preserving the scriptures and serving the laity. Buddhist monasteries rapidly acquired economic assets, and some were wealthy, owning villages, buying and selling land, and involved in trade. Property brought problems, requiring managers, both lay and monastic (in Theravada Buddhism monks are not supposed to handle money). On the other hand, monks and nuns are obliged to act as 'a field of merit' and accept whatever the laity offers them. Moreover, Theravada Buddhist monks are exempt from military service (in Mahayana countries, the government can make them disrobe temporarily for that purpose). Inevitably, this has often led to the suspicion that many are only interested in a comfortable and secure life, free from labor in the fields.

A king was meant to be the foremost layman of the kingdom and ensure the purity and safety of the Sangha. All ordinations took place under his authority. Following the example of Ashoka, kings were supposed to expel monks who broke the rules or fomented splits within the Sangha. It has been argued by Samuel (1993) that it was the relative strength of the monarchy in Southeast Asia that enabled Theravada to remain dominant and prevented other sects, based on the Mahayana scriptures, for example, from becoming established. By contrast, in Tibet the effective political control of Lhasa never spread far beyond the heartland, thus enabling a rich variety of Buddhist philosophies and schools to

flourish, including many that gave a much greater emphasis than the dominant Gelukpa sect to shamanic and noncelibate practice.

Buddhist kings were faced with the problem that simply in order to carry out their duty as a king – protecting the people, maintaining order – they had to commit sins (primarily killing) and were therefore disbarred from attaining salvation. At the same time, in both Theravada and Mahayana countries, the king is often regarded as a *bodhisattva* (Buddha-to-be). Ashoka is remembered as having killed many enemies in the war of Kalinga; then, after conversion, he became full of remorse and vowed not to do it again. But in Sri Lanka the heroes of the Sinhalese chronicles (composed by monks) are kings Duttha-gamani (101–77 BCE) and Parakkamabahu I (1153–86), who fought on behalf of Buddhism. The Sangha offered a way out: earning merit to counterbalance their sins though supporting the Sangha. At the same time, one of the most popular and (for Buddhists) most poignant Jataka (Rebirth) stories tells how heir apparent Prince Vessantara (the Buddha in a previous life) renounced and gave away, not only his kingdom and wealth, but his wife and children as well.

In short, the Sangha enabled the king to legitimate himself in the eyes of the people, and it refused to ordain fugitives from royal justice. In return, the king supported the Sangha, granting it land and other privileges, and when necessary carried out purifications, suppressed noncanonical innovations, and supported revivals. Only in Tibet did the Sangha eventually take over and *become* the state.

Religious Specialists and the Laity

Lay Buddhists take the Five Precepts: undertakings not to take life, not to steal, not to commit sexual wrongdoing, not to lie, and not to take intoxicants, which lead to carelessness and the breaking of the other four precepts. At fixed points in the lunar fortnight lay people take the Eight Precepts: this includes the Five (while interpreting the third to mean a complete cessation of sexual activity), and adds three others: not to eat after midday, not to adorn oneself or go to entertainments, and not to use luxurious beds (i.e., to sleep on the floor). In this way lay people, mostly women, take up the same discipline as a monastic novice, but on a temporary basis.

Traditionally, the Buddhist laity have never had to foreswear all other religious relationships. Providing they approach the gods merely for worldly matters, this is usually of no concern to the monks, and has nothing to do with their faith in Buddhism as a path to salvation. It is hard for Buddhist laity to continue to practice animal sacrifice, but even this is done by some Buddhists in Tibet, Nepal, and Thailand in what they define as non-Buddhist contexts.

The facts that Buddhism is centered so firmly on salvation and the afterlife, and that lay Buddhists almost inevitably have other, worldly interests, mean that "Buddhism in real life is *accretive*" (Gombrich, 1971: p. 49, original emphasis). Buddhism always coexists – often in a structured hierarchy that places the Buddha at the top of the pantheon – with other religious systems that satisfy the worldly needs of ordinary Buddhists. Thus, when Theravada Buddhists celebrate the harvest, seek a cure for illness, or get married, they do so in a non-Buddhist idiom with the help of non-Buddhist specialists (Tambiah, 1970). Hindu gods are included in the Buddhist pantheon as powerful spirits, subordinate to the Buddha; they traveled with Buddhism to other parts of Asia, where they were sometimes identified with local deities, all of whom became part of the Buddhist cosmology.

This soteriological focus explains why, traditionally, there is no such thing as a Buddhist wedding ceremony, although pious lay people may optionally invite a monk or monks to chant and be fed after the ceremony. Even in Mahayana countries the close relationship between Buddhism and death continues. In Japan, death ritual is almost exclusively a Buddhist concern, whereas birth and marriage are celebrated in a Shinto idiom. However, many Buddhist temples in East Asia also offer help for worldly needs; and in Nepal, among the Newars, Buddhist priests have even established a complete set of Buddhist life-cycle rituals, including marriage.

This interrelationship between Buddhism and other systems has frequently been mislabeled 'syncretism' by Western observers using an exclusivist definition of religion. But this is an ethnocentric judgment which ignores the fact that Buddhist lay people who worship non-Buddhist spirits or gods do nothing wrong, providing they do not do so in the hope of obtaining salvation. Ironically, however, modernist Buddhists have themselves adopted the Western-derived exclusivist definition of religion in their attempts to reform what they see as backwardness and superstition.

The radical individualism of the Buddha's message was socially and psychologically hard to accept. Most Buddhists, including monks, have wished to make some kind of offering to benefit their dead parents, for example. The ritualized offering of merit to others, which may include dead relatives, is common in all forms of Buddhism. In Theravada Buddhism the canonical explanation of the protective chanting ritual performed for the dead is that the dead can only benefit if they happen to be in the form of ghosts and happen to be nearby so that they can gain merit from listening. That such an abstruse justification might not be shared by all lay people can easily be imagined. In a similar way, many lay people have a straightforwardly magical attitude to the use of Buddhist amulets. Mahayana Buddhist specialists do not insist so unequivocally on maintaining a rationalist, commemorative explanation of Buddhist rituals; but they agree that the magical attitude corresponds to the lowest level of understanding.

Reforms, Revivals, and National Identity

Buddhism has always been subject to cycles of decline and reform. All forms of institutionalized Buddhism experience a tension between well-endowed institutions, which specialize in scholarship and ritual, on the one hand, and wandering or forest-dwelling monks or practitioners, on the other. Frequently, the laity (including the elite) perceive the freelance practitioners as closer to the Buddhist ideal and beat a path to their door. Over time, and thanks to the donations they cannot refuse, the freelancers themselves become well endowed, and inspire a further cycle of reform aimed at recapturing true asceticism and/or meditative practice.

The Rise of Buddhist Modernism

In the nineteenth and twentieth centuries, faced everywhere with competition from Christian missionaries and taking advantage of the new technology of the printing press, reform movements took a new and sometimes overtly political turn. Buddhist institutions experienced very different trajectories depending on the country. In Japan, Buddhism had been closely tied to the state in pre-Meiji times and Buddhist priests had been responsible for government record keeping; after 1868, Buddhism was forcibly separated from Shintoism (with mixed religious sites being forced to choose one or the other identity) and much land was expropriated (Ketelaar, 1990). In China and Tibet, Buddhism, like all autonomous civil institutions, suffered enormously under Communist rule, being permitted a limited revival in the 1980s and 1990s.

The Theravada countries have experienced a new form of rationalist, scripturalist Buddhism that has been dubbed Buddhist Modernism or Protestant Buddhism (Gombrich and Obeyesekere, 1988). In Sri Lanka this was a protest both against the Protestant missionaries and against traditional Theravada Buddhism: it criticized monks for being unlearned ritualists, for being inactive on issues of social and political concern, for not practicing meditation, and for encouraging superstition. More positively, it stood for universal literacy, Buddhist Sunday schools, and a middle-class morality of strict monogamy and hard work. It sought to base Buddhist teachings strictly on the Pali canon, and to purify practice of anything that could be dubbed Hindu or Mahayana influence. It stressed the scientific nature of Buddhist teachings, often claiming that major scientific advances were already to be found in the Pali canon. In Sri Lanka it also aimed to make Buddhists proud to be Buddhist, and to defend their cultural and religious heritage against the inroads of the Protestant missionaries who confidently predicted that in 50 years the Buddha would have gone the same way as Wodin and Thor. The rise of Protestant Buddhism went hand in hand, therefore, with the emergence of Sinhala nationalism. The Sinhalese race-nation came to be defined in terms of its Buddhism, and Buddhist historical chronicles were taken to demonstrate its long association with, and right to dominate, the island. The key figure in this Buddhist revival in Sri Lanka was Don David Hewawitarana (1864–1933), who took the title Anagarika Dharmapala, founded the Maha Bodhi Society, and campaigned ceaselessly to awaken the Sinhalese to their religious and nationalist mission.

Western Views of Buddhism and the Protestant/Catholic Parallel

The story of Buddhist revival in Sri Lanka intersects crucially with the story of the Western discovery of Buddhism. Many of the Victorians had a deep sympathy for Buddhism, because they thought it was, like Christianity, a simple and rational faith founded by a pure-hearted reformer. Edwin Arnold's poem, *The Light of Asia* (1879), had an enormous impact. Counterattacking Christian apologists dubbed Buddhism 'nihilistic' and 'antilife.' There was a pervasive tendency, which has lasted up to the present in some quarters, to view Mahayana and Theravada Buddhisms as analogous to Roman Catholicism and Protestantism within Christianity. (Hostile descriptions of Buddhism borrowed a well-worn vocabulary, honed in many Protestant descriptions of Catholicism, in which the words 'priestcraft,' 'superstition,' and 'mumbo jumbo' loomed large.) It is true that Mahayana Buddhism has a baroque pantheon, numerous elaborate rituals, and more elaborate priestly hierarchies; but the contrast was exaggerated because of the adoption of the modernist interpretation of Theravada Buddhism as a simple, rationalist philosophy whose only ritual was the commemoration of its founder. The parallel breaks down as a serious tool of analysis for several reasons:

1. Historically, Theravada Buddhism represents the earlier form, against which Mahayana rebelled;
2. Theravada Buddhism is traditionally built on monasticism, and therefore on spiritual hierarchy, and it is modernist Buddhism, which has introduced a Protestant-style belief in the equality of all believers;
3. The scriptures of the Theravada were preserved in Pali, and it is only in the modern period that they began to be translated into the vernaculars;
4. The worship of relics is very important in Theravada Buddhism, and sponsoring their worship has been an important source of legitimacy for rulers, both traditionally and in the modern period;
5. Traditionally, Theravada Buddhists believed that nirvana was impossible to attain at present, because of the degeneracy of the age, which made the quest for salvation considerably less urgent than in Protestantism.

These differences between Protestantism and actually existing Theravada Buddhism were obvious to those Westerners, such as Madame Blavatsky and Colonel Olcott, founders of the Theosophical Society, who traveled to India in search of the wisdom of the East. Olcott came to Sri Lanka in 1880 and set about reforming Buddhism, so that it could become what they claimed it had always been. It was Olcott who first set up Buddhist schools, drew up a Buddhist catechism, and encouraged the annual celebration of the Buddha's enlightenment with Buddhist carols. He was instrumental in the setting up of Buddhist Sunday schools and the Young Men's and Young Women's Buddhist Associations. He also designed the Buddhist flag, which is now flown around the world. It was Olcott and Blavatsky who first encouraged Dharmapala. Many of the early Buddhist revivalists were Theosophists as well. Dharmapala, along with D.T. Suzuki, who first popularized Zen in the West, and Vivekananda, the leader of neo-Hindu revivalism, attended the 1893 World Parliament of Religions in Chicago, a key event in the history of Buddhist modernism.

Government Control in Thailand

In Thailand, which avoided European colonialism, modernist reforms were, by contrast, largely a top–down process. The reforming King Rama IV (Mongkut) (reigned 1852–68) had spent 27 years as a monk before ascending the throne. He emphasized strict adherence to the Pali canon, excluding later, popular works that encouraged the laity to view the Buddha as a miracle worker. It was Mongkut's monastery that was the center of what was to become the Dhammayutika Nikaya, a new monastic order that stressed adherence to the Vinaya

rules. Increasing centralization led in Thailand to the Sangha Act of 1902, by which the government took effective control not just of the royal monasteries in the capital but, for the first time, of all monasteries and monks in Thailand. A national exam syllabus was established with textbooks written by Rama IV's son and the king's half brother, Prince Wachirayan, who himself became the Supreme Patriarch of the Thai Sangha in 1910. In Thailand, but not in Sri Lanka, it is the custom for all young men to spend at least 2–3 months as a monk, around the age of 20 years. (In Burma, temporary ordination tends to be taken much younger, and can last anything from a day to several years.) Giving up the status of monk and returning to lay life is not seen as in any way shameful; many men spend 10 or more years receiving an education as a monk, and then return to lay life. One consequence of tightening up of monastic education was that monasteries, which had previously had a monopoly on learning and literacy, had to stick to a more strictly religious curriculum, and many forms of knowledge – medicine, astrology, history writing – that had been passed on in them, now found a home outside.

Buddhism and National Identity

As in Sri Lanka, the reform of Buddhism in Thailand was intimately tied to the construction of a sense of national identity. Though this has now changed, Muslim and Christian minorities rarely used to be persecuted; but Buddhism receives official backing. However, the detailed government control of religion has meant that, when dissident movements arise, it is faced with the difficult choice between suppressing them in line with its traditional responsibility for the purity of Buddhism, and permitting them in the name of religious freedom.

In Burma, Thailand, and Sri Lanka the relationship between the ruling elite, the people, and Buddhism has been so close that it is possible to speak of protonationalism there even before the modern period; that is to say, the people were defined in terms of their adherence to Buddhism and any ruler who failed to protect Buddhism could not be accepted as legitimate. Certainly, when the modern period dawned and nation-states were created, it was hard to be accepted as a Burmese or a Thai unless one professed Theravada Buddhism, and many of the ethnic troubles of Sri Lanka have flowed from the fact that most Sinhalese Buddhists feel the same way. As Islam in some neighboring countries becomes more powerful and gains influence abroad, in Burma, Thailand, and Sri Lanka there is now considerable hostility toward Muslim minority communities.

There is a similarly tight connection between nationalism and Buddhism in Tibet, but in other Mahayana countries the position is very different. In Japan, Buddhism was persecuted as a foreign religion in the 1870s, and in Korea restrictions were continuous from 1392 until the twentieth century: the ruling Choson dynasty, which based itself on Confucian ideas, banned Buddhist monasteries from the major towns.

Lay Meditation and the Movement for Female Ordination

The rise of the lay meditation movement marked a major change in Buddhist practice in the twentieth century. Few monks meditated traditionally, most performed ritual services for the laity, and some studied the scriptures. The highest status went to learned monks in urban centers or to rare forest meditators. Lay life was thought to be incompatible with meditation. Part of the modernization of the religion has seen meditation being taught to the laity, often by laymen and women. For many middle-class Buddhists, this means that they bypass monasteries altogether and base their religious practice in meditation centers. Burma led the way in the development of meditation for the laity, and Vipassana (insight) meditation has spread throughout the world; for example, the efforts of an ethnic Indian born in Burma, S.N. Goenka, have been very influential well beyond the community of those born Buddhist. Zen Buddhism has perhaps been even more influential in the West, though whether the popular understanding of it accords with the way it is practiced in East Asia is open to question. In India, Bhimrao Ambedkar, leader of India's untouchables and one of the framers of India's Constitution, began a major movement of social liberation by urging his followers to embrace Buddhism and converting himself at the end of his life in 1956 (Zelliot, 1992).

Just as the Buddhist laity in general are not as submissive to monastic authority as they once were, so there has been a general movement for the emancipation of women in Buddhism. Historically, the order of nuns has always been subordinate to monks, the highest-ranking nun being considered, formally, as inferior to the most junior monk. Furthermore, within Theravada countries the nuns' ordination tradition died out in the eleventh century (Sri Lanka) and thirteenth century (Burma): since five nuns are needed to ordain a new one, the formal status of nun was no longer available to women. All that they could do was to take the Ten Precepts, like a monastic novice, on a permanent basis, and live a celibate life, a precarious and unprestigious possibility attractive mainly to widows (though often also done by elderly men). As part of the revival of Buddhism in the Theravada countries, many women started to adopt a more monastic way of life and to be treated as fully fledged nuns. In the 1980s and 1990s, Western feminist Buddhists have argued that Theravada women should take full ordination from the Chinese tradition where it has been preserved, and some have done so. This has been controversial; many of those (mainly men) who have opposed them doing so argue that they should thereby become Mahayanists or refrain from calling themselves nuns. While the theological objections are very different, the issue arouses similar passions to the ordination of women priests in the West.

'Humanistic' and 'Engaged' Buddhism

Another major strand in Buddhist modernism goes back to the Chinese monk Tai Xu (1890–1947). He founded 'humanistic Buddhism,' which emphasized better organization and a more rational outlook, in particular that Buddhism should attend to the needs of this life, not rituals for the dead. This new Buddhist tradition was renamed 'engaged Buddhism,' by the Vietnamese monk, Thich Nhat Hanh, who settled in France and America and had many Western followers.

At the end of World War II, Tai Xu and some of his chief pupils fled Maoism; though he died in Hong Kong, some

pupils reached Taiwan, which has become a major source of Buddhist revival and innovation. It is notable for having by far the highest proportion of Buddhist nuns of any country in the world; the foundation for this development was laid by the Japanese during their occupation of Taiwan between the two World Wars, after Tai Xu had played a minor role.

Two Taiwanese sects have spread worldwide in recent years. Fo Guang Shan, led by Master Hsing Yun, while relatively traditional, combines Pure Land and Zen elements, and has been active in promoting the nuns' revival overseas. Even more remarkable is Tzu Chi ('Compassion Relief'), founded and still headed by a laywoman, the Master Cheng Yen (b.1937) (who now however has an unorthodox ordination). Starting from nothing, in a remote town in eastern Taiwan, Cheng Yen in 1967 founded Tzu Chi as a medical charity; it now has six hospitals, a medical school, and the largest bone marrow bank in Asia. It also fights environmental pollution and undertakes disaster relief, notably even in mainland China. At the outset, most of Tzu Chi's members were women, who sustained the movement by unremitting fundraising; now that it has spread overseas (claiming 10 million members), probably half of the members are men, not necessarily Buddhists. Tzu Chi's vision of Buddhism is starkly simple: neither meditation nor doctrine plays any significant role; the emphasis is wholly on doing practical good, for which Tzu Chi refuses to work through any intermediary organization: every gift must be made by a Tzu Chi volunteer in person.

See also: Religion, Psychology of; Religion, Sociology of; Religion: Evolution and Development; Religions of East Asia.

Bibliography

Almond, P.C., 1988. The British Discovery of Buddhism. Cambridge University Press, Cambridge.
Bartholomeusz, T.J., 1994. Women under the Bo Tree: Buddhist Nuns in Sri Lanka. Cambridge University Press, Cambridge.
Bechert, Heinz, Gombrich, Richard (Eds.), 1984. The World of Buddhism. Thames and Hudson, London.
Buswell Jr., R.E., 1992. Zen Monastic Experience: Buddhist Practice in Contemporary Korea. Princeton University Press, Princeton, NJ.
Carrithers, Michael, 1983. The Forest Monks of Sri Lanka. Oxford University Press, Delhi, India.
Collins, Steven, 1982. Selfless Persons: Imagery and Thought in Theravada Buddhism. Cambridge University Press, Cambridge.
Gellner, D.N., 1992. Monk, Householder, and Tantric Priest: Newar Buddhism and its Hierarchy of Ritual. Cambridge University Press, Cambridge.
Gombrich, R.F., 1971. Precept and Practice: Traditional Buddhism in the Rural Highlands of Ceylon. Clarendon Press, Oxford, 1991. Reissue as Buddhist Precept and Practice. Motilal Banarsidass, Delhi, India.
Gombrich, R.F., 1988. Theravada Buddhism: A Social History from Ancient Benares to Modern Colombo. Routledge, London.
Gombrich, R.F., Obeyesekere, Gananath, 1988. Buddhism Transformed: Religious Change in Sri Lanka. Princeton University Press, Princeton, NJ.
Goodwin, J.R., 1994. Alms and Vagabonds: Buddhist Temples and Popular Patronage in Medieval Japan. University of Hawaii Press, Honolulu, HI.
Gunawardana, R.A.L.H., 1979. Robe and Plough: Monasticism and Economic Interest in Early Medieval Sri Lanka. The University of Arizona Press, Tucson, AZ.
Hardacre, Helen, 1984. Lay Buddhism in Contemporary Japan: Reyukai Kyodan. Princeton University Press, Princeton, NJ.
Ishii, Y., 1986. Sangha, State, and Society: Thai Buddhism in History (P. Hawkes, Trans.). University of Hawaii Press, Honolulu, HI.
Ketelaar, J.E., 1990. Of Heretics and Martyrs in Meiji Japan: Buddhism and Its Persecution. Princeton University Press, Princeton, NJ.
Lamotte, Etienne, 1988. History of Indian Buddhism: From the Origins to the Saka Era (S. Webb-Boin, Trans.). Oriental Institute. Université Catholique de Louvain, Louvain-la-Neuve, Belgium.
LeVine, Sarah, Gellner, D.N., 2005. Rebuilding Buddhism: The Theravada Movement in Twentieth-Century Nepal. Harvard University Press, Cambridge, MA.
Lopez Jr., Donald S., 1999. Prisoners of Shangri-La: Tibetan Buddhism and the West. The University of Chicago Press, Chicago, IL.
Madsen, Richard, 2007. Democracy's Dharma: Religious Renaissance and Political Development in Taiwan. California University Press, Berkeley, CA.
Malalgoda, Kitsiri, 1976. Buddhism in Sinhalese Society, 1750–1900. University of California Press, Berkeley, CA.
Mumford, S.R., 1989. Himalayan Dialogue: Tibetan Lamas and Gurung Shamans in Nepal. University of Wisconsin Press, Madison, WI.
Ortner, S.B., 1989. High Religion: A Cultural and Political History of Sherpa Buddhism. Princeton University Press, Princeton, NJ.
Samuel, Gregory, 1993. Civilized Shamans: Buddhism in Tibetan Societies. Smithsonian Institution Press, Washington, DC.
Schopen, Gregory, 1997. Bones, Stones, and Buddhist Monks: Collected Papers on the Archaeology, Epigraphy, and Texts of Monastic Buddhism in India. University of Hawaii Press, Honolulu, HI.
Silber, I.F., 1995. Virtuosity, Charisma and Social Order: A Comparative Sociological Study of Monasticism in Theravada Buddhism and Medieval Catholicism. Cambridge University Press, Cambridge.
Spiro, M.E., 1982[1970]. Buddhism and Society: A Great Tradition and Its Burmese Vicissitudes, second expanded ed. University of California Press, Berkeley, CA.
Tambiah, Stanley J., 1970. Buddhism and the Spirit Cults in North-East Thailand. Cambridge University Press, Cambridge.
Tambiah, Stanley J., 1976. World Conqueror and World Renouncer: A Study of Buddhism and Polity in Thailand against a Historical Background. Cambridge University Press, Cambridge.
Tambiah, Stanley J., 1984. The Buddhist Saints of the Forest and the Cult of Amulets. Cambridge University Press, Cambridge.
Tambiah, Stanley J., 1992. Buddhism Betrayed? Religion, Politics and Violence in Sri Lanka. University of Chicago Press, Chicago, IL.
Weber, Max, 1958[1916–17]. Gerth, H.H., Martindale, D. (Eds.), The Religion of India: The Sociology of Hinduism and Buddhism. The Free Press, Glencoe, IL.
Welch, Holmes, 1967. The Practice of Chinese Buddhism: 1900–1950. Harvard University Press, Cambridge, MA.
Yao, Yu-Shuang, 2012. Taiwan's Tzu-Chi as Engaged Buddhism: Origins, Organization, Appeal and Social Impact. Global Oriental/Brill, Leiden and Boston.
Zelliot, Eleanor, 1992. From Untouchable to Dalit: Essays on the Ambedkar Movement. Manohar, Delhi, India.

Relevant Websites

www.accesstoinsight.org.
www.globalbuddhism.org.
www.ocbs.org.
http://online.sfsu.edu/rone/Buddhism/introbuddhistbibiog.htm.

Budgeting and Anticipatory Management

Michel J Lebas, HEC School of Management, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Budgeting is a deliberate act of anticipatory interactive coordination of the use of a business' resources to take it in a strategically selected direction. This article describes the purposes and the principles of anticipatory management. It emphasizes the collaborative nature of this process around a shared and adaptive causal model. It asserts that practice and principles are often not aligned, thus delegitimizing the budgeting process. The author concludes by evoking the paths currently carried out so as to reinvent a relevant budgeting process and lists 10 key points any business must address to anticipate effectively and support quality strategy implementation.

Budgeting is a deliberate act of anticipatory interactive coordination of the use of an organization's resources to take the firm in a selected direction. It involves several actors in a business process in identifying alternative action plans leading to the achievement of the goal or intent, selecting one, and documenting it. An action plan allows a balancing of resources required and available or to be created. The 'best' plan (the one retained) has high probability of achieving success and has a compatible resource requirement. A budget defines (1) financial and nonfinancial measures of both outcome and intermediate milestones describing the path to success and (2) the resources required or generated by this plan.

Deviations (actual or forecasted) from the milestones are sources of learning. They are used to adapt the action plan to newly forecasted conditions so that the revised plan will remain coherent with the intent or the goal in the changing competitive environment. Traditional confusion between (1) anticipation (i.e., action plan selection and forecasting outcome and resources), (2) resource allocation to responsibility centers, and (3) the use of milestones of achievement as bases of reward systems often leads to dysfunctional behavior such as budget slack or inability to exploit new opportunities. The budgeting process serves to focus the persons in any organization on both the 'performance gap' (more and better effectiveness and efficiency) and the 'opportunity gap' (preparing the organization to handle future opportunities).

All organizations must:

1. coordinate the way resources will be used by the various (often specialized) decision-makers; and
2. learn from the observed deviations between intended and actual.

They need these two activities in order to create value for their stockholders by increasing the probability of satisfying their key stakeholders such as customers, employees, suppliers, and any outside agency granting the organization the right to continue to operate.

For the past 100 years or more, budgets have been the prime instrument in providing such a process of *ex ante* coordination and *ex post* analysis. They did so by defining, too often from 'above' (the leadership sends a 'command') in a complex and cumbersome process, the actions (and the resources attached) that are likely to lead to goal achievement, combined with *ex post* evaluation of variances or deviations from plan ('control' to bring conformance).

A budget is generally defined as a formal document that quantifies an organization's plan for achieving its goal (Jiambalvo, 2012). A budget, in other words, is a description of the financial implications of a sequence of coordinated actions and specialized targets that will allow an organization to achieve its objectives in a changing environment. Most budgets, until recently, assumed implicitly, once the budget had been defined – i.e., the actions (including resources attached) and targets selected – that uncertainties about the environment could essentially be 'ignored' for a period of implementation. Managers were expected to 'manage to the budget' more than to pay attention to the competitive environment and to new opportunities. The budget period used to be about 12 months and was supplemented by frequent updates (two or three times a year, depending on the turbulence of the competitive environment) but within the same old 12-month framework and, too often without changing the 'end of year' objectives. Despite these updates, many managers still try to 'achieve the budget,' thus still assuming an estimate, made some 6–18 months before, is still valid.

A budget is in itself a fundamental document, and budgeting is an essential process in the operation of any organization, however the way it has been used in practice has led to many abuses and negative effects.

This short article will define the principles of budgeting and 'budgets' before showing the limitations of current practice and will explore how the 'concept' (i.e., a tool to support ongoing adaptive anticipation or anticipatory management) ought to be used today to meet current economic and competitive challenges.

A Philosophy of Anticipation

"As for the future, your task is not to foresee it, but to enable it" (Antoine de Saint Exupery in *Le Petit Prince*) and, to do so means to be able to mobilize, continuously, all resources in the organization to adapt to and exploit new opportunities that appear. This means the path to 'success' cannot be completely defined beforehand.

The whole organization, and especially its management team collectively, must continuously update the 'best' way to

create value for stakeholders in continuously adapting to the competitive environment. Despite such an observation, 'command and control' are still the most commonly used form of management, essentially in contradiction to the reactivity and agility required for success in a competitive world.

Budgets and the budgeting activity (process of establishing and validating action plans) are processes and procedures that aim at contributing to the creation of value. A budget is a statement – a declaration – of actions intended to be carried out by several organizational decision-makers in a coordinated way. A budget reflects the answers a firm and its leadership give at one point in time (and it may possibly be a valid answer, but only for a short period of time) to one key question: 'How are we going to go from where we are to where we want to be?.' A budget and budgeting process is one of the key ways through which energies are mobilized and channeled for the successful implementation of the business strategic intent.

We will first identify the purposes and the process of creation of the budget, then the key requirements for the budget to be an effective tool that fulfills its promises, before reviewing criticisms of budget practices and formulating the key points we see developing in new interpretations of this fundamental tool of anticipation.

Purposes of Budgets and of Budgeting Processes

The budgeting process and budgets contribute to the firm's anticipation in that they:

1. Provide a forum in which the various organizational members (who do not have identical skills, experience, and expertise) can identify and share their views of opportunities and threats to the value creation process of the business.
2. Provide a locus for the construction of a shared business model (identify the key success factors and how they interact in creating a successful firm).
3. Allow senior management to discuss, select, and communicate values (to guide behavior), intent (defining what value creation or success means), and plans (coordinated sequence of interlocked actions and identification of resources created and consumed).
4. Create a forum for coordinated decision-making between actors that may otherwise act in a disconnected or centrifugal fashion that would result from their referring to different business models. For example, the sales manager might consider developing sales at a rate that might not be compatible with the action plan imagined by the manufacturing manager (who foresees difficulties in bringing the required new technology on line) or with the action plan of the financial manager (who would rather increase the dividends than finance the increase in working capital that the increased sales might require).
5. Provide for the development of transfunctional action plans describing how to achieve continuous alignment of the firm's 'behavior' with competitive pressures.
6. Identify what resources are needed at what time. Resources ought not be limited to tangible or financially measurable resources: know-how, intellectual capital, customer, supplier or employee loyalty, morale, image and style, etc., are resources that must be used appropriately to create and maintain long-term success.
7. Allow for the identification and allocation of the scarce resources required to carry out the action plans effectively and efficiently.
8. Reinforce a structure of delegation, devolution, and 'subsidiarity' in colocating 'rights to make decisions' (on the use of resources) and knowledge and in supporting initiative and autonomy while maintaining coordination. We use the term 'subsidiarity' to refer to the fundamental idea that decision-making should be as close as possible to the 'customer' or to the market, i.e., any unnecessary attempt at going up in the delegation chain to reach a decision is slowing down the organization's reaction and runs the risk of creating a solution which is maladapted to the need.
9. Help build commitment by individuals or groups of individuals to achieving time-based results that move the organization in the intended direction. The principle is to obtain commitment to a result and not necessarily to the way through which this result is to be achieved. This aspect is the source of great misuse of budgets in that many organizations confuse 'commitment to the use of resources as agreed upon' and 'commitment to results.' Often the source of confusion is due to a combination of lack of trust on the part of the supervisor and lack of self-confidence on the part of the subordinate.
10. Contribute to establishing benchmarks for performance measurement and evaluation (for both business units and individuals). Benchmarks need not be limited to financial measures of results or resource consumption.
11. Provide a basis for learning from deviations. This aspect has often been misunderstood to mean that conformance to the plan is expected. Since, by definition, no plan can be perfect, conformance for its own sake is a source of noncompetitiveness.
12. Create a context for the deliberate management of performance improvement.
13. Allow for the forecasting of cash flow needs on the basis of the action plans that have been retained. Cash is an essential scarce resource in any organization and must be monitored at all times.
14. Offer a process for an ongoing estimation of year-end financial results and more specifically earnings per share (EPS) and return on equity (ROE).

This whole array of purposes cannot be broken down and each aspect dealt with separately. They must all be pursued simultaneously.

The Budgeting Cycle

Generally, the budgeting process, once a strategy (direction) has been selected (by senior management), follows a preestablished sequence but many variations are possible with today's simulation instruments:

1. Identify and communicate estimated future demand, markets conditions, and business environment evolution to

establish, for every participant, the hypotheses all should assume in researching their contribution to the action plans implementing the strategy. Such a task should, most of the time, originate with upper management that holds a bird's-eye view of the whole business.
2. Jointly identify threats, opportunities, weaknesses, and strengths of the organization in the perceived (or declared) future environment.
3. Identify and evaluate the various possible action plans (sales, production, sourcing, human resources, innovation and investments, R&D, etc.) and verify their compatibility with both intent or goals and resources.
4. Repeat the process until compatibility is obtained and acceptance by everyone in the organization has been secured. This step may require the superior to act as a binding arbitrator to avoid subordinates wasting time in useless battles. An arbitrator listens to all parties, reflects on the best collective solution, and selects a preferred course of action that is both explained and justified to all parties who must adhere to it.
5. Once the budget or action plan is established, results are matched against expected (intermediate or final) milestone results and the analysis of differences (deviations or variances) leads to continuous learning and thus to a reevaluation of the appropriateness of the action plan and its possible updating.

Components of an Action Plan

An action plan reflects the answers of the concerned persons to at least the following questions:

1. Where are we and why do we need to do 'something' (understand the cause of the action plan by identifying a possible gap between where we are and where we are headed, and where we think we ought to be, given the intent or goals of the firm or organization)?
2. What is the 'intention' behind doing something?
3. How does whatever action that is decided relate to the firm's goal, target, or mission?
4. Who are the various actors who need to be involved (basis for both richness of analysis and delegation in the implementation process)? An action plan is always teamwork, even though each person on the team is personally responsible for one 'component' of the plan, often overlapping with someone else's. For example, an action plan aimed at 'generating orders that the firm can deliver to the customer's satisfaction and that help the firm grow' would require the cooperation of such specialized managers as: sales manager, sales people, marketing manager, production scheduling, assembly manager, legal department, sourcing manager, management accountant, etc. If the plan is not the result of an agreement between all these required skill holders, it has little chance of success.
5. What business model (i.e., causal model) do the 'team members' agree on to help them identify the 'action variables' they will activate through their decisions and behavior?
6. What possible actions can be envisioned? It is important to review several possible alternatives critically.
7. Which alternative will we choose, and why? What will we actually do in this action plan in each responsibility center?
8. How will we do it? What processes? What resources will be consumed and created and when? What delegation structure will be developed, etc.?
9. Are the required and generated resources coherent with the availability and needs of the organization? If the answer is negative, there must be a return for adjustment to any or all questions (1) through (8).
10. Why will we do it this way? What were the alternatives' pros and cons? How realistic is the alternative selected and how realistic were the other ones? What are the risks attached to each alternative and especially the one selected?
11. Coherence with strategic intent, strategy, and other action plans.
12. Expected impact of the selected actions (several may apply simultaneously):
    a. development of a competitive advantage;
    b. ambitious yet realistic: placing the organization under tension;
    c. challenging existing ways of thinking; and
    d. all of the above or other.
13. When will these actions be carried out (calendar of actions)?
14. How often will these actions need to be carried out (if plan requires recurring actions)? A multistage and conditional action plan allows for better learning and facilitates continuous adjustment.
15. To what do each of the actors actually commit: milestones, dates, intermediate or final results, nonfinancial indicators, relationships, etc.?
16. How will the achievement of the commitments be measured? There are at least three domains for which measures should be agreed upon beforehand by all parties involved in the action plan. These measures will serve as a basis for learning from deviations and for continuously updating the action plan so that it remains aligned with the intention selected, if it is still the right one:
    a. Process, i.e., the way the action plan should and does unfold during implementation (the equivalent aerospace term would be 'trajectory'):
        – evolution of the conditions that created the need for the action;
        – resources consumed and created;
        – resource productivity and effectiveness;
        – organization and relationships between actors;
        – mastery over risks identified in step (10);
        – efficiency of use of resources; etc.
    b. Progress (against the time scale selected by the actors in the plan to state by which date the target will be achieved):
        – timing: date of start, selection of milestones, date of expected completion;
        – positioning of progress on a PERT diagram (Program Evaluation and Review Technique, a tool used in project management that creates a dynamic

map the sequence of actions required to completion and records duration – theoretical and actual – and defines a critical path) or other time-based diagram(s);
  - validation of the critical path;
  - definition of the method of calculation of the percentage achievement; etc.
c. Impact:
  - effect of actions on objectives and key performance indicators (KPIs) for both short- and long-term; and
  - evaluation of the continued relevance of the intent.

17. What priority should be given to this action plan and how much of the firm will be involved (if applicable)?

During the twentieth century budgeting seems to have focused essentially on steps (8) and (9) and therefore has led to a very partial view of what is meant by an action plan. Essential to the effectiveness of the budgeting process is the measurement system evoked in step (16). The measures should be balanced between areas of expertise, perspectives, and time horizons and allow both effective implementation and continuous validation of the relevance of the whole action plan given the evolution of the competitive environment. Budgeting and processes of anticipation link, therefore, directly to the area of management now commonly described as the Balanced Score Card.

Budget Process Requirements

A Sense of Direction

There must be goals that preexist the beginning of the budgeting process. Each goal must have a time frame and be a milestone toward achieving the organization's mission or intent. In specifying its goals, the organization states which are the stakeholders that are taken into consideration.

A Continuously Updated Shared Causal Model

Decision-makers and personnel must be able to understand some key 'cause and effect' relationships that affect the organization's output or performance. The budget rests on the identification of a causal model that helps understand what consequences might result from what actions in the perceived competitive environment. Much time is needed in a business to identify and maintain the model(s) or causal representations that will help understand how it could go from point X to point Y on the map of opportunities.

A Monitoring System

A monitoring system must be able to tell managers how well they are 'getting there' from 'here.' Some of the measures must be anticipatory or 'leading,' in the sense that they allow the managers to anticipate inappropriate or unacceptable results and therefore head off the undesired trend through appropriate corrective actions, while other measures are more historical and serve to keep track of what has been achieved.

A Set of Criteria for Trade-Off Between Goal Achievement and Resource Consumption and Creation

Whether the trade-off will be carried out in economic terms only or using criteria such as market share, employee satisfaction, or citizenship of the firm, the firm needs to have some preference ordering system that allows it to identify a 'preferred way' each time a choice is to be made. It is very difficult in any multi-person setting to come to an agreement on criteria for preference ordering. The budgeting process allows managers to encourage participation and expression on the part of everyone in the organization and communicates values and preference orientations.

A Budget Requires Trust Between Participants

A budget is grounded in delegation. Each person holds responsibility for a 'piece of the action' and must be able to count on the others doing what they committed to do and letting their partners, who depend on the achievement of a milestone or of a result, know as early as feasible (through shared leading indicators) the actual changes in the probability of fulfilling the commitments. Budgets therefore require trust between participants (both vertically and laterally) and self-confidence combined with a reasonable willingness to take risks and not adhere to a prewritten script.

Budgets, 'Opportunity Gap,' and 'Performance Gap'

In projecting itself into the future, any organization faces two simultaneous challenges:

1. Improve the 'productivity' (effectiveness and efficiency) of resources consumed in carrying out currently retained action plans.
2. Identify new opportunities for continued growth that may require modification of current action plans and, possibly, the development of entirely new ones. Opportunities have a limited life span. If no new opportunities are identified and addressed, the organization will cease to grow and decline into oblivion. Single-purpose organizations have had to reorganize their mission completely to continue existing. For example, the March of Dimes, founded in 1938 by President F.D. Roosevelt to fight polio, shifted its mission to the prevention of premature birth and birth defects; the Tennessee Valley Authority, founded in 1933 initially to transform an agrarian region through the use of electricity is now mainly an interstate regional economic development agency. A long-lasting organization such as the Kirin Beer Brewery (incorporated in 1907) or Coca-Cola (formulated in 1888, incorporated in 1892) are always reinventing themselves because product attractiveness has a time limit and markets have limits.

We define the first challenge as that of managing the 'efficiency and effectiveness gap' while we will call the second one that of managing the 'opportunity gap.' As business environments grow turbulent, the continued relevance of any given action plan in delivering effectiveness is less and less likely.

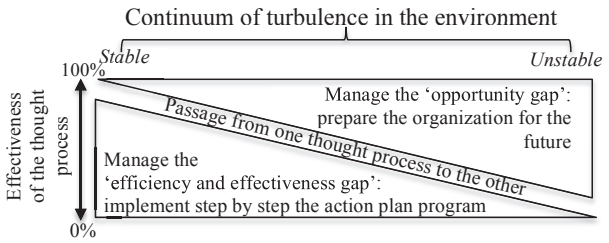

Figure 1 Continuum of turbulence in the environment.

Simultaneously, the need for continuous identification of new opportunities grows greater.

Figure 1 shows that, at any point in the continuum of turbulence, the required anticipation thought processes of an organization results from the combination of both 'managing the efficiency and effectiveness gap' and 'managing the opportunity' gap (preparing the organization for the future). At each point on this continuum, managers must hold simultaneously two very different mental maps, in differing proportions.

The effectiveness of the budgeting system regardless of the level of turbulence (i.e., competitive environment) requires that managers and supporting processes take into account the appropriate mix of attention paid to both the efficiency and effectiveness gap and to the opportunity gap.

Competitive turbulence affects each level of delegated responsibility differently in the organization. It is the responsibility of a superior to 'filter' some of the uncertainties and turbulence he/she faces for his/her subordinate.

The higher the person is in an organization, the higher the degree of uncertainty he/she has to deal with. Therefore, the higher the person is, the more his/her anticipation efforts should be dedicated to addressing the 'opportunity gap,' but addressing some form of efficiency and effectiveness gap cannot be delegated entirely. As responsibility is delegated down the organization, each successive level is more and more focused on the efficiency and effectiveness of the organization, but there is always a nonnegligible preoccupation with some form of attention paid to the opportunity gap, recognizing that new opportunities can be identified at all levels of the organization.

Current Budgeting Practice is not Always Aligned with the Principles

As mentioned previously, it is a challenge to build a budgeting process that considers the issue of competitiveness not only from the point of view of the efficiency and effectiveness gap (resource consumption control, process efficiency, delivery of promised results, etc.) but also from that of the opportunity gap. The causes for such a difficulty can be traced to the following elements:

- Objectives and processes to get to the target are often defined only *in* and *on* financial terms.
- Value creation is mostly defined in terms of financial profit.
- Partition of the processes to achieve the target tends to be carried along hierarchical or functional lines rather than being developed along the necessarily transfunctional business processes of the customer-value creation process.
- The processes of alignment, dialog, and team building through budget construction are frequently ignored and the exercise is too often reduced to a process of scarce resource allocation leading to competitive behaviors between organizational entities that ought to cooperate to serve the stakeholders best.
- The time horizon of action plans is not linked to the business cycle but is based, too often, on fiscal year constraints (which are defined by reporting needs and have little link with the need to have visibility of the future). Budgeting should be carried out as an ongoing process, with rolling forecasts. Today's information technology makes it economically feasible. Budgeting is a state of mind. Budget construction should be reengineered in most firms to take as little time as is possible (for example, by concentrating on the relationships in the causal model and on leading descriptors of the status of these relationships). If the budget were an ongoing process, frequent debates would take place that would continuously replace the firm on the most appropriate trajectory or path to its objectives and targets, thus developing confidence and commitment on the part of all in the organization.
- Many times the budget has been hijacked as a substitute for an anticipatory form of financial accounting used in estimating year-end accounting results of the firm so that management can comfort stockholders about forecasted profit.
- Focus is more on cost reduction than cost management (the 'right' cost is not always the least cost: costs are driven by product and process design, and by people eliminating non-value-adding activities).
- Most action plans assume, usually erroneously, that costs are driven by the volume of sales and are best controlled at the point where they can be measured (Lebas, 1999). This leads to conformance budget (focused on resource allocation) thus limiting initiative and customer satisfaction.
- Many budget processes merge and confuse anticipating and defining benchmarks for reward systems (especially if the latter are based on conformance with resource consumption) thus creating gaming behavior and budget slack (budget padding) which immobilize resources uselessly. Budget slack occurs when managers overestimate resource consumption or required effort and underestimate value created with the customers. Managers indulge in such behavior because most of the time they do not trust the leadership either to make additional resources available when new opportunities arise (added to the fear of the corollary, i.e., that resources might be taken away when anticipated opportunities do not materialize) or to allocate resources in support of action plans that are compatible with the manager's perceptions. Such behaviors immobilize resources and prevent the exploitation of unforeseen opportunities. Practices such as 'flexible budgeting' are designed to address this concern. This practice consists in:
  - separating clearly, before the fact, what areas of the business model each manager is responsible for (i.e., which parameters are under his/her control) and for which results commitments are made, from those parameters that are not under the person's control, and

- recalculating, after the facts are known, the elements of the action plan (consumption of resources as well as outcome or output) that would have been determined had the 'uncontrollable' elements been forecasted with pinpoint accuracy.

Such an approach allows managers' freedom to really anticipate and forecast because he/she knows no precautions are needed (through the building of budget slack) to cover risks he/she cannot control.

- Often, the budgeting process does not implicate everyone in contributing to the definition of goals, targets, and action plans, thus preventing buy-in of the decisions made. A frequent reaction is to consider that 'spending' resources allocated for an action plan becomes the objective rather than continuously validating the alignment, with the intent, of the planned actions that led to the original allocation.

The main weaknesses of many budget processes currently in place in too many firms are that they are not built on a well thought-out system of:

1. shared representations of causal relationships between actions and consequences;
2. subsidiarity and delegation which would rest on trust in people's ability to make the right decision at the right time, as close to the point of need as possible;
3. representation of the firm as a network of business processes; and
4. customer orientation.

Conclusion

Anticipatory tools today are not constrained by the fiscal time horizon; they are rolling and cover a period often closely related to product life cycles. Anticipatory practices recognize the preeminence of the human aspect in the success of businesses. Choosing the right people and motivating them in well thought-out and coordinated processes make for satisfied customers who, in the end, drive the financials of the firm. Such a sequence requires flexibility and continuous adaptation.

On the basis of the experience of many firms that have developed anticipatory practices that are more coherent with the principles we have listed above, we can come up with a set of good practices that define anticipatory management for the twenty-first century. Such firms include Svenska Handelbanken, Volvo Cars, or Skandia in Sweden; Schlumberger, Air Liquide, or Bull Group in France; Borealis in Denmark; Statoil and SpareBank1 in Norway; Boots the Chemist in the UK; or Elkey Manufacturing, Group Health Cooperative or Sprint in the US; Tenaga Nasional in Malaysia; or Grupo Auto Sueco in Portugal. Their practices, although case-specific, can be distilled to a common set of principles, which include the following elements:

1. Release individuals' creativity and initiative by generalizing entrepreneurial centers (business units that are large enough to be profit driven, but not so large they would lose agility).
2. Develop a 'customer-oriented' model of performance creation in each business unit.
3. Develop a shared causal model described by a set of multivariate key performance indicators specific to each business unit, nested in a firm wide business model that encompasses all units.
4. Nested devolution of strategy and responsibility, within and between units.
5. Time-bound autonomy: i.e., *the right to err and correct* as described, for example, in the "Management Principles" formulated by Percy Barnevik, then CEO of ABB, in 1988).
6. Implementation, throughout the firm, of the philosophy of continuous progress, i.e., the Deming PDCA model (plan–do–check–adjust) benchmarked on competition.
7. Rigorous commitment of managers and business units to understandable and controllable balanced 'indicators' (i.e., generalized use of Balanced Score Cards).
8. Rolling Forecasts and Action Plans with flexible and case-specific time horizons.
9. Disconnection of forecasts from the commitment/reward system.
10. Resources allocated on the basis of justified need: internal 'market' for resources and transparency in allocation.

These firms have moved from a twentieth century culture of seeing budgets as a set of contractual commitments within limited resources, with an emphasis on preventing centrifugal forces, to a twenty-first century view that emphasizes doing better than the competition and encouraging autonomy of and cooperation between units. They have shown that there are alternatives to the old mechanical model. The principles we listed here are not the definitive answer to what will be a work in progress for a long time.

The fact is we are entering an era of continuous adaptation and learning. Businesses need to adapt the tools too, and the 'budget' will not be spared here. Because they have to be completely reinvented, 'budget processes' (coordinated anticipatory processes) should be even more fascinating to students, management researchers, and managers alike than they were before.

See also: Corporate Finance: Financial Control; Corporate Governance; Entrepreneurship and Organizations; Learning: Organizational; Management Accounting; Marketing Strategies; Organizational Control; Organizational Decision Making; Organizations: Authority and Power; Strategic Adaptation; Strategic Management; Strategizing; Supply Chain Management.

Bibliography

Banham, R., September 2012. Freed from the Budget, CFO Magazine www.BBRT.org (USA) and www.beyondbudgeting.org (UK): two connected websites where the most provocative ideas about the evolution of budgets are openly discussed by professionals.

Bogsnes, B., 2008. Implementing beyond Budgeting: Unlocking the Performance Potential. John Wiley & Sons, London.

CAM-I Europe, 1994. A Journey to Advanced Management Systems: A Report on the Research and Findings of the CAM-i Advanced Budgeting Study Group. UK: R-94-AMS-0.1.1 Poole.

Hope, J.D., Player, R.S., 2012. Beyond Performance Management. Harvard Business Review Press, Boston.

Hope, J.D., Bunce, P., Röösli, F., 2011. The Leader's Dilemma: How to Build an Empowered and Adaptive Organization without Losing Control. Jossey Bass, London.

Hope, J.D., Fraser, J.R.T., Bunce, P., Röösli, F., 2006. Beyond Budgeting, BBRT White Paper. Beyond Budgeting Partnership LLP, Lymington.

Hope, J.D., Fraser, J.R.T., 2003. Beyond Budgeting: How Managers Can Break Free from the Annual Performance Trap. Harvard Business School Press, Boston.

Hope, J.D., Fraser, J.R.T., January 1999. Beyond Budgeting. Management Accounting (UK).

Jiambalvo, J., 2012. Managerial Accounting. John Wiley and Sons, New York.

Lebas, M., 1999. Which ABC? Accounting based on causality rather than activity-based costing. European Management Journal 17 (5), 501–511.

Budgeting: Political Science Aspects

James D Savage, University of Virginia, Charlottesville, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of public budgeting enables political scientists to theorize about the allocation of government resources, the development, structure, and dynamics of political institutions, and the formation of comparative fiscal and macroeconomic policies. Among the major theories either derived from or used to study budgeting are incrementalism, decrementalism, punctuated equilibrium, and rational choice theory.

Budgeting is a process by which governments allocate public resources to bureaucracies, other governmental units, and client groups. This process usually takes place on an annual basis, involves procedural rules and institutional arrangements, often invokes intense political conflict, and, together with taxation, forms the basis for fiscal and macroeconomic policy. Because it fundamentally determines political winners and losers through its allocative function, budgetary processes and outcomes serve as a distinct guide to a state's philosophy of government and public policy. The act of budgeting, therefore, lies at the heart of politics and government.

The Study of Budgeting

Budgeting may be studied from several perspectives. From the vantage point of professional schools of public administration, the practical elements of budgeting include the drafting and formatting of budget documents, the application of cost–benefit and other tools of program analysis, the projection of revenues and expenditures, the division of spending into its operational and capital components, and the management and implementation of programs funded through the budget. Economists principally focus on the macroeconomic consequences of budgeting, such as the effects of deficit spending on interest rates, inflation, the 'crowding out' or 'crowding in' of private investment, foreign exchange rates, and national debt management. At the microeconomic level, economists may also consider the effectiveness of specific programs funded in the budget, such as determining the influence of public jobs programs on the labor market. These questions may also be of interest to and addressed by political scientists. Yet, the centrality of political science research looks at budgeting as a guide to a broader theoretical analysis of politics, the examination of institutional arrangements, an understanding of the behavior of political actors within those institutions, and an exploration of specific public policies.

Budgeting as Incrementalism

Modern theoretical understandings of budgeting begin with the idea of 'incrementalism.' Formulated in the 1960s, incrementalism attempts to explain how budgetary actors carry out their roles, and what the consequences are for politics in structuring budgetary decisions in one form as opposed to alternative ones (Wildavsky, 1964; Fenno, 1966). Incrementalism is grounded in certain assumptions derived from social psychology about human nature. The theory contends that people are inherently limited by cognitive, time, and resource constraints from being truly rational and synoptic in their ability to make budgetary decisions. Budgetary actors 'muddle through' and 'satisfice' rather than 'optimize.' They are constrained from making decisions in the way that microeconomically rational consumers might, weighing all information sources and alternatives before a decision is reached. To compensate, these actors rely on 'aids to calculation' and institutional norms and roles to govern behavior. In particular, instead of reviewing every program and spending item in the budget to determine what should be funded, the focus of analytical and political decision making is the annual change, or increment, in spending. The budget base or the previous year's spending is taken as a given, and the politics of budgeting revolve around determining the size of incremental 'fair shares' allocated to budgetary claimants.

Incrementalism takes into account the role of institutions and actors within those institutions. Budgeting is regarded as a decentralized, 'bottom-up' process, where budget proposals emerge both from within executive branch agencies and from the demands of politically successful client groups. In the case of the United States, federal agencies serve as advocates for larger funding increments for their programs to the Appropriations Committees in the House of Representatives and the Senate. Committee members evaluate these funding claims and act according to a set of behavioral norms. The House Appropriations Committee, for example, operates as the 'guardian of the purse,' and the Senate Committee acts as a 'court of appeals.' Wildavsky and Fenno saw the appropriations process as a form of legislative oversight of the bureaucracy. Where agencies proposed ever-growing budgets to expand their administrative domain, appropriators regularly trimmed those requests to keep the bureaucracy under control. Fenno, for example, noted that agency budget requests were cut 77% of the time by the House Appropriations Committee to enforce bureaucratic discipline.

Meanwhile, new client groups gain their access to fair shares of the budget through the electoral process. These groups establish their political legitimacy and justify their budgetary claims by supporting the successful election of members of

Congress. Sympathetic and electorally beholden members then create programs that are administered by these agencies and included in the budget base. Once added to the base, client groups may then rely upon receiving annual incremental budgetary allocations.

Institutional norms also guide how committees produce fiscal coordination. The House and Senate Appropriations Committees act separately in making their spending decisions and they act independently from the revenue-generating House Ways and Means Committee and the Senate Finance Committee. Nevertheless, according to incremental theory, budgetary coordination stems from the overarching norm that budgets should be balanced. Consistent with this norm that is deeply rooted in American political history, Congress would grant spending increments just within the margin of revenue growth.

Incrementalism views budgeting, above all else, as a political activity that gives limited consideration to notions of administrative efficiency or economic rationality in the determination of budgetary outcomes. Both agencies and politicians instead rely upon incremental aids to calculation. For bureaucrats, incrementalism implies stability and long-term growth in resources. For politicians, incrementalism reduces societal tension and political conflict; because interested groups quickly learn that once their claims are included as part of the budget base, they may count on regular, and presumably, growing increments.

Challenges to Incrementalism

Critics of incrementalism soon argued that the model had become obsolete. First, alternative methods of budgeting came into vogue that directly challenged incrementalism's assumptions of limited human rationality and organizational capacity. Zero-based budgeting (ZBB) and program budgeting in the form of Planning, Program, Budgeting (PPB) were designed to improve the choices of decision makers (Schick, 1966). Supporters of ZBB declared that the entire budget, rather than just the increment, could be analyzed and evaluated in detail each year by breaking up the budget into discrete 'decision units.' Policy makers would then be given the opportunity to reprioritize and reassemble these units to make rational decisions about which elements of the budget would be funded. Defenders of PPB argued that multiyear planning, rather than annual incrementalism, enabled decision makers to set broad, strategic, programmatic goals and analyze the long-term fiscal implications of current budgetary choices. New analytical techniques and growing computer capacity contributed to the development of Baseline Budgeting, where current spending levels serve as the baseline for estimating future rates of programmatic spending, by taking into account inflation and changes in law. Thus, ZBB, PPB, and Baseline Budgeting, each in its own way, rejected the basic assumptions of incrementalism in favor of presumably more rational, analytical, synoptic, and multiyear practices that could lead to more effective decision making and more efficient government.

Second, where incrementalism depended upon budgetary growth for its conflict-reducing fair shares and growing budget bases, the rise of deficit spending, economic instability, and ideological demands for reductions in the size of the public sector caused governments to engage in 'decrementalism' rather than incrementalism. By the mid-to-late 1970s, governments confronted the politically difficult task of making decremental reductions in spending through 'cutback budgeting,' rather than by providing for orderly budgetary growth. Political scientists noted that governments throughout the world coped with fiscal stress by adapting similar techniques to control budget deficits (Schick, 1986, 1988). Budgeting became more centralized, 'top-down,' and 'front-loaded' to reduce the influence of budgetary claimants, as compared to incrementalism's bottom-up approach that assumed stability in the budget's base and ignored the executive branch in setting the budgetary agenda and shaping policy outcomes. The cuts in spending that took place in Australia, Canada, Denmark, New Zealand, and Sweden, for example, were accompanied by the top down centralization of budgetary decision making throughout the parliamentary formulation and administrative execution of the budget. Governments established macrobudgetary rules at the beginning of the budgetary process that constrained bottom-up decisions taking place during the remainder of the process; employed hard-spending targets, ceilings, and caps to limit spending; calculated long-term inflationary and program costs through baseline budgeting; created new legislative committees and support agencies to impose greater oversight on the budgetary process; strengthened the powers of finance ministers over line agency spending demands; and outsourced government functions to private contractors. Whereas incrementalism, its critics claimed, neglected to address the rise of entitlement spending, the new macrobudgetary rules imposed 'pay as you go' requirements that mandated offsetting revenues or spending cuts to pay for new entitlement programs.

Domestic and international politics often revolved around the budgetary and fiscal decisions that stemmed from new macrobudgetary rules. The United States employed a variety of formal rules in the 1980s, 1990s, and in 2011 to balance budgets, control deficits, and constrain spending (Savage, 1988; White and Wildavsky, 1989). The Balanced Budget and Emergency Control Act of 1985, better known as Gramm–Rudman–Hollings, mandated that the budget should be balanced by setting annual deficit targets over 5 years. Excessive deficits would be reduced by automatic budget cuts, called sequesters, in various categories of spending. American politicians learned they could not balance budgets in the midst of recession and expanding deficits, because the size of sequesters required, proved to be politically impossible to activate. The politicians changed their target from balancing the budget to controlling spending, and they replaced Gramm–Rudman–Hollings with the Budget Enforcement Act of 1990. This law set spending caps that would be forced by sequesters, and it imposed 'pay as you go' requirements. The Congress renewed a version of this law in 1997 and enacted the similar Budget Control Act of 2011.

The European Union (EU) created its own set of macrobudgetary rules as part of its continued efforts at political and economic integration. In 1972, the EU adopted the Treaty on European Union, also known as the Maastricht Treaty, which required that member states seeking to join the Economic and Monetary Union (EMU) bring their budget deficits within the 'reference value' of 3% of GDP and their debt within 60% of

GDP. To ensure continued compliance with these targets after member states successfully entered the monetary union, the EU adopted the Stability and Growth Pact in 1997. Despite the Pact's threat of financial sanctions, as shown in Table 1, numerous member states violated the targets, especially as the financial crisis of 2009 ravaged Europe's economy. The collective deficits of the 17 countries using the euro currency grew from 0.7% of GDP in 2007 to 6.4% in 2009, with the deficits of numerous states far exceeding this level. Deficits in all of the 27 EU countries grew from 0.9% of GDP in 2007 to 6.9% in 2009. Member states hid the true condition of their deficits to gain entry into the EMU, and then disguised their deficits to avoid the financial sanctions that could be imposed under the Pact. In response, the EU adopted additional surveillance procedures and rules that promoted deeper economic and fiscal integration in 2011 and 2012. Nevertheless, the EU avoided imposing the Pact's financial penalties on any member state for running excessive deficits, justifying the absence of sanctions on the basis that the financial crisis produced extenuating economic conditions. The EU repeatedly granted extensions to the member states, even as many of them attempted to bring their budgets into compliance, often through politically unpopular austerity programs (Savage, 2005; Hallerberg et al., 2010; Heipertz and Verdun, 2011). These measures did reduce the collective deficits in the 17 country euro area to 3.7% of GDP in 2012, and in the 27 EU countries to 4.0% of GDP.

Third, another challenge to incrementalism's assumption of ongoing budgetary stability came from punctuated equilibrium theory. Where incremental theory predicted stable, fair share budgetary allocations, competing analyses of long-term, cross-national spending patterns reveals a combination of incrementalism mixed with dramatically large changes in expenditures. These large-scale punctuations of the status quo stemmed from radical and urgent changes in public priorities and policies (Jones et al., 2009). As a result, from time to time, budgets experienced 'bursts of spending,' or, less likely, reductions in spending. After responding to these policy changes, spending reverted to long-term patterns more consistent with incrementalism. Although it does not account for these substantial variations in spending, incremental theory does note the presence of 'shift points' in expenditure patterns and deviant cases due to changes in public policy.

Table 1 European Union Euro Area Government surplus/deficits (−) as percent of GDP, 2007–2012

	2007	*2008*	*2009*	*2010*	*2011*	*2012*
Austria	−0.9	−0.9	−4.1	−4.5	−2.5	−2.5
Belgium	0.1	−1.0	−5.6	−3.8	−3.7	−3.9
Cyprus	3.5	0.9	−6.1	−5.3	−6.3	−6.3
Estonia	2.4	−2.9	−2.0	0.2	1.2	−0.3
Finland	5.3	4.4	−2.5	−2.5	−0.8	−1.9
France	−2.7	−3.3	−7.5	−7.1	−5.3	−4.8
Germany	0.2	−0.1	−3.1	−4.1	−0.8	0.2
Greece	−6.5	−9.8	−15.6	−10.7	−9.5	−10.0
Ireland	0.1	−7.4	−13.9	−30.8	−13.4	−7.6
Italy	−1.6	−2.7	−5.5	−4.5	−3.8	−3.0
Luxembourg	3.7	3.2	−0.8	−0.9	−0.2	−0.8
Malta	−2.3	−4.6	−3.7	−3.6	−2.8	−3.3
Netherlands	0.2	0.5	−5.6	−5.1	−4.5	−4.1
Portugal	−3.1	−3.6	−10.2	−9.8	−4.4	−6.4
Slovakia	−1.8	−2.1	−8.0	−7.7	−5.1	−4.3
Slovenia	0	−1.9	−6.2	−5.9	−6.4	−4.0
Spain	1.9	−4.5	−11.2	−9.7	−9.4	−10.6
Euro area 17	−0.7	−2.1	−6.4	−6.2	−4.2	−3.7
EU 27	0.9	−2.4	−6.9	−6.5	−4.4	−4.0

Source: Eurostat, April 2013.

Rational Choice Approaches to Budgeting

Since the early 1980s, many political scientists have come to rely upon economically inspired rational choice models, rather than incrementalism's social–psychological approach, to study budgeting. The basic assumption of rational choice is that political actors rationally seek to maximize their self-interest. Holding power is central to political self-interest and in democratically elected governments this means gaining reelection. Consequently, political institutions and public policies, including budgetary processes and outcomes, are structured by politicians for electoral purposes.

Symptomatic of reelection budgeting is distributive 'pork barrel' spending, where members direct appropriations to their states and districts to win political credit with their constituents. Pork barrel projects are also allocated as side payments to create coalitions of legislators in sufficient numbers to pass legislation. Depending upon the model, the size of these coalitions vary from 'minimum winning coalitions' that contain a majority plus one legislator, to 'universal coalitions,' where all legislators who seek a project receive one. The budget is therefore characterized by programs where projects are provided for each legislative district. As a result of distributive spending, the government expends funds on economically dubious projects that create large, systemic budget deficits, with what are presumed to be negative consequences for the macroeconomy (Shepsle and Weingast, 1981).

Like incrementalism, the basic rational choice model has not gone unchallenged. Although politicians do engage in distributive spending, the cost of these projects is insufficient to account for the huge budget deficits that occurred in the 1980s and 1990s. Entitlement spending, principally in the form of pensions and health care expenditures, rather than pork barrel spending, is clearly the primary source of budgetary growth. The rise of large-scale deficit spending in the United States, moreover, may also be traced to stagflation, recession, monetary policy, tax cuts, and the buildup in defense spending during the 1980s. Furthermore, politicians in many countries have risked their own reelection chances by pursuing politically costly fiscal policies, including cutbacks in distributive spending, in order to reduce these deficits.

The minimum winning and universal coalition elements of the rational choice models, which have often been expressed in formal, mathematical terms, have also been challenged on empirical grounds. Individual discretionary government projects appear to benefit only a limited number of legislative districts. Even when bundled together for purposes of creating voting coalitions, most program portfolios fail to provide universal benefits for all districts. While the rational model assumes that pork barrel budgeting takes place for reelection purposes, individual constituents overwhelmingly receive their

government benefits from entitlement and general benefit programs, and not from the lower-visibility, pork barrel funding (Stein and Bickers, 1995).

Institutions, Divided Government, and Budgeting

Studies suggest that divided governments, where competing political parties control the executive and legislative branches, experience greater difficulties in controlling budget deficits and expenditures than unified ones (Alesina and Rosenthal, 1995). Leftist parties spend more than rightist parties, and parties differ in their spending priorities. Changes in the American budgetary process since 1974 enhanced the power of the parties to set overall budget and fiscal priorities, though the detailed spending decisions remain the prerogative of the committee system. When control of the government is divided, priorities conflict and fiscal stalemate is likely to occur. Similarly, the more unified and conservative the government, the greater the priority to balance the budget (Kieweit and McCubbins, 1991).

Incrementalists add that a divided government also influences the advocate and guardian roles of budget makers. Whereas the early incremental models regarded specific roles as enduring, the behavior of budgetary actors during the rise of divided government indicated they were in fact mutable (Cox et al., 1993). When rightist parties control the executive branch, for example, the executive becomes less of an advocate of spending and more of a guardian or protector against perceived budgetary excesses. The more leftist-oriented the House Appropriations Committee becomes, the less its emphasis on guardianship, particularly in the presence of a conservative executive. A decreasing level of guardianship is associated with the growing size of deficits. The Senate Appropriations Committee, meanwhile, tends to remain in its role as a court of budgetary appeals.

Budgeting, Foreign Assistance, and State Building

Budgeting plays a role in international relations and political development. Democratic budgetary institutions and processes are regarded as key contributors to good governance, the rule of law, democratic deliberation, transparency in public affairs, political accountability, and the efficient provision of public services. The failure to budget effectively, on the other hand, contributes to dysfunctional government. Dysfunctional governments, especially those falling into the category of failed states, pose a potential threat to other states due to their tendencies toward political instability, unrest, and violence. Consequently, recent theories of political development and state building point to the need to provide foreign assistance and technical aid to promote state budget capacity, in order to encourage democratic development and economic growth, and to reduce international conflict in developing and post-conflict countries, including Afghanistan and Iraq (Ghani and Lockhart, 2008; Savage, 2013).

International organizations, donors, and national development agencies have indeed provided technical assistance to build state capacity in all the phases of budget formulation, adoption, and execution. Budget formulation refers to a government's technical ability to draft credible budgets that reflect coherent and detailed programmatic spending plans. Budget adoption refers to a government's political ability to evaluate and pass a budget in a deliberative manner. Budget execution refers to a government's administrative ability to spend the funds allocated in the budget effectively, honestly, and in a timely fashion consistent with international best practices.

Foreign assistance and democratic development projects, however, have experienced mixed success. Critics note the failure of international organizations and donors to coordinate their own assistance efforts, provide adequate oversight in the use of development funds, and offer adequate training of beneficiary personnel, to name just some of the limitations of these interventions. As a result, the beneficiary governments remain in a failed and dysfunctional condition, plagued by instability, violence, and corruption (Collier, 2008). In this regard, one ongoing debate is whether foreign assistance funds should be spent directly, 'off-budget' by the donors, or transferred to and spent 'on-budget' by the beneficiary governments. Donors often prefer to spend their funds directly through their own budgets to ensure that their project preferences are maximized and the mismanagement of their funds are minimized. Doing so, however, delegitimizes the beneficiary government, conflicts with state spending priorities, encourages governments to remain dependent on donor funding, and undermines efforts at building political accountability for government actions (del Castillo, 2008).

Budgeting and Political Science

Incrementalism and rational choice-based theories have greatly contributed to political science's analyses of the politics and the institutional forces that create budgets. These models and theories and the challenges presented to them offer important policy implications for budgeting. Incrementalism points to the norm of budget balance as a reoccurring goal of public policy. Challenges to incrementalism point to new procedures and rules to improve budgetary decisions. Rational choice accounts for the presence of often wasteful distributive spending that threatens the success of government programs and public policy. Political scientists have also studied budgeting in the context of international relations, comparative politics, and political development. The evolution of non-incremental budgeting macrobudgetary rules plays a central role in European integration. Budgeting is also seen as an essential factor in the development of good governance and effective administration for developing and postconflict countries.

See also: Administration in Organizations; Budgeting and Anticipatory Management; Bureaucracy and Bureaucratization; Organizational Decision Making; Organizations: Authority and Power; Public Administration: Organizational Aspects; Public Policy; Public Sector Organizations; Rational Choice and Organization Theory.

Bibliography

Alesina, A., Rosenthal, H., 1995. Partisan Politics, Divided Government, and the Economy. Cambridge University Press, New York.

Collier, P., 2008. The Bottom Billion: Why the Poorest Countries Are Failing and What Can Be Done about It. Oxford University Press, New York.

Cox, J., Hager, G., Lowery, D., 1993. Regime change in presidential and congressional budgeting: role discontinuity or role evolution? American Journal of Political Science 37, 88–118.

del Castillo, G., 2008. Rebuilding War-Torn States. Oxford University Press, New York.

Fenno Jr., R.F., 1966. The Power of the Purse: Appropriations Politics in Congress. Little, Brown, Boston, MA.

Ghani, A., Lockhart, C., 2008. Fixing Failed States: A Framework for Rebuilding a Fractured World. Cambridge University Press, New York.

Hallerberg, M., Strauch, R.R., von Hagen, J., 2010. Fiscal Governance in Europe. Cambridge University Press, New York.

Heipertz, M., Verdun, A., 2011. Ruling Europe: The Politics of the Stability and Growth Pact. Cambridge University Press, New York.

Jones, B.D., Baumgartner, F.R., Breunig, C., Wlezien, C., Soroka, S., Foucalt, M., Francios, M., Green-Pederssen, C., Koski, C., John, P., Mortensen, P.B., Varone, F., Wlagrave, S., 2009. A general empirical law of public budgets: a comparative analysis. American Journal of Political Science 53, 855–883.

Kieweit, D.R., McCubbins, M.D., 1991. The Logic of Delegation. University of Chicago Press, Chicago.

Savage, J.D., 1988. Balanced Budgets and American Politics. Cornell University Press, Ithaca, NY.

Savage, J.D., 2005. Making the Emu: The Politics of Budgetary Surveillance and the Enforcement of Maastricht. Oxford University Press, New York.

Savage, J.D., 2013. Reconstructing Iraq's Budgetary Institutions: Coalition State Building after Saddam. Cambridge University Press, New York.

Schick, A., 1966. Planning program budget system—symposium: the road to PPB: the stages of budget reform. Public Administration Review 26, 243–258.

Schick, A., 1986. Macro-budgetary adaptations to fiscal stress in industrialized democracies. Public Administration Review 46, 124–134.

Schick, A., 1988. Micro-budgetary adaptations to fiscal stress in industrialized democracies. Public Administration Review 48, 523–533.

Shepsle, K.A., Weingast, B.R., 1981. Political preferences for the pork barrel: a generalization. American Journal of Political Science 25, 96–111.

Stein, R.M., Bickers, K.N., 1995. Perpetuating the Pork Barrel. Cambridge University Press, New York.

White, J., Wildavsky, A., 1989. The Deficit and the Public Interest. University of California Press, Berkeley, CA.

Wildavsky, A., 1964. The Politics of the Budgetary Process. Little, Brown, Boston, MA.

Relevant Websites

http://www.abfm.org/ – Association for Budgeting and Financial Management.

http://www.gfoa.org/ – Government Financial Officers Association of the United States & Canada.

http://blog-pfm.imf.org/pfmblog/ – International Monetary Fund Public Financial Management Blog.

http://www.oecd.org/ – OECD: Public Finance Department.

http://www.cbo.gov/ – United States Congressional Budget Office.

http://www.omb.gov/ – United States Office of Management and Budget.

http://www.pefa.org/ – World Bank: Public Expenditure & Financial Accountability.

Burckhardt, Jacob (1818–97)

Wolfgang Hardtwig, Humboldt University, Berlin, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 2, pp. 1400–1402, © 2001, Elsevier Ltd.

Abstract

Swiss historian of art and culture, in conscious contrast to the contemporary professionalization of historical science he retained an older, personal understanding of learning and education. Historical education should not contribute to an alteration of the outer reality, but widen the mental horizon of those seeking knowledge. With this maxim, his independent cultural historical method, and his art of portrayal not only in fact influenced the educational horizon of those interested in history to a great degree, but also had a stimulating effect on the increasingly specialized areas of historical science (ancient history, modern history, history of art).

Jacob Burckhardt was born in Basle, Switzerland on 25 February 1818 as the son of the pastor of Basle cathedral. At first he studied theology in Basle, although after a religious crisis he changed to history and history of art, and continued his studies in Bonn and Berlin from 1841 to 1843. He was most influenced, especially in Berlin, by Franz Kugler, Johann Gustav Droysen, August Boeckh, and Leopold Ranke. He habilitated at Basle in 1844, held lectures in history and history of art there, and edited the 'Basler Zeitung' from 1843 to 1845. After the uprisings in Switzerland of 1844–45, an early manifestation of the midcentury revolutionary unrest in Europe, his previously liberal political outlook changed into a decided conservatism. With this he turned away from his political and publishing activities in favor of the 'education of old Europe.' In 1855 he was appointed to a chair at the technical university of Zurich, and in 1858 at Basle, where he taught history until 1886, and history of art from 1874 to 1893. In 1860 he stopped publishing and concentrated on his lectures, although he completed a number of manuscripts after this point. He died in Basle on 8 August 1897.

Jacob Burckhardt's historical writing can be understood as a critical diagnosis of the present. This found its most clear expression in his lectures on the history of the revolutionary era and *Über das Studium der Geschichte* (*On the Study of History*) as well as in references to the present throughout his work. Burckhardt rejected the liberal historians' optimism in progress, and saw a danger for his ideal of culture in the contemporary processes of mechanization, industrialization, and fundamental democratization. He also felt the modern development of large states to be a threat, particularly that of the new German Reich from 1871, and he made accurate prognoses of the political catastrophes of the twentieth century. However, his historical pessimism was in contradiction at least to his early religiously influenced opinions on and admiration of art. In numerous long tours through Italy he worked on the basis of his extensive works on art history, and these journeys through Italy, France, and The Netherlands remained an important part of his life until his old age. In 1855 he published *Cicerone. Eine Anleitung zum Genuß der Kunstwerke Italiens* (*Cicerone. An Introduction to Enjoying Italy's Art*), a work which visitors to Italy interested in art still use today. In 1867 his work, *Kunst der Renaissance* (*Art of the Renaissance*), was brought out by his friend Wilhelm Lübke. In this work he deliberately avoided a history of the artists themselves and instead concentrated on a 'history of art according to functions.' His late masterpiece *Erinnerungen an Rubens* (*Memories of Rubens*) was published from his papers after his death, in which the re-evaluation of the Baroque which he had completed in his later years found its expression. His portrayal of Renaissance art had an important influence on the educated German public's reception of the Renaissance and Italy. With his concept of a history of art ordered according to functions he guided the emerging study of art on to new methodological paths.

Burckhardt's basic historiographical principles are most clearly illustrated in his introductory lecture *On the Study of History*, which he held three times between 1868 and 1873, and which his nephew Jacob Oeri edited and published in 1905 under the now famous title *Welthistorische Betrachtungen* (*Reflections on History*). Burckhardt drew up his images of the historical world from a universal historical perspective, assuming three basic human needs: (1) the political (state), (2) the 'metaphysical' (religion), and the 'critical' (culture). In this scheme of 'potencies' he relativized the importance of the nation and the nation state in conscious contrast to German historicism, and allotted an equally important role to culture as the expression of human spontaneity as to the 'stable' powers of state and religion. As culture he understood 'the quintessence of all that which has come into existence spontaneously for the promotion of material life, and as the expression of the spiritual and emotional–traditional life, all social gathering, all technologies, arts, writing and sciences.' Burckhardt thereby rediscovered an enlightenment tradition of cultural history, and consolidated it in that he saw the polarity of freedom and violence in a developed form not only in state and religion, but also in culture. However, the potencies theory also enabled him to relativize the stress on a typological basis. The interaction of the three potencies adds depth and complexity to historical reality, as Burckhardt experienced it.

In his three historical masterpieces, *Die Zeit Constantins des Großen* (*The Age of Constantine*), (1853), *Die Kultur der Renaissance in Italien* (*The Civilisation of the Renaissance in Italy*) (1860) and *Die griechische Kulturgeschichte* (*The Cultural History of Greece*) (*3 volumes, 1898–1900*) Burckhardt varied the potencies scheme in that he added a further definition of

culture to state and religion. Culture is then the entirety of the achievements of a human creative urge, which is, however, permanently under threat of being destroyed. In all three works Burckhardt assumes a state of acute breakdown of authority and an increased societal and moral antagonism. In this latent chaos the respective forms of state and religion represent the relatively stable potencies, which, however, remain fragile because the 'spirit' strives persistently as a 'subversive,' as a destructive, and also creative ancient force, to bring forth new forms.

In *The Civilisation of the Renaissance* Burckhardt focused on Italy of the fifthteenth and early sixteenth century, in contrast to Sismondi and Michelet, and thereby lent the term 'Renaissance' the temporal and geographical boundaries it has retained up to the present day. Burckhardt drafted the image of an epoch in the sway of violence and crime, lawlessness and tyranny, in which, however, equally the rational calculation of seizing and retaining power became established and so created the specific modern form of the 'state as artform.' The Humanists established themselves alongside the tyrants as a temporally specific human type, in whom the modern individual was embodied just as it was in the rulers. For Burkhardt they were the 'most obvious examples and victims of unbridled subjectivity' with a new intensity and sophistication of emotional life as well as a new method of rationalizing their relationship to the world. In *The Cultural History of Greece* Burckhardt describes similarly to the image of the state in *The Civilisation of the Renaissance* the Greek state, the polis, as a place of 'state slavery of the individual' in which democracy degenerated into a tyranny of the majority. Burckhardt developed the principle of agon as an essential structural feature of life in the Greek polis, the competition between individuals, yet also the bitter struggle of the poleis against one another. This principle does have destructive effects, yet at the same time it brings about the preconditions for the development of cultural freedom and a singular creative wealth. In Burckhardt's view the Greek form of religion was, however, just as constitutive as the principle of agon, as it knew neither theology, theocracy nor the rule of priests, unlike the states of the old orient, and so enabled a free development of spiritual life. As leading experts on ancient history criticized sharply, Burckhardt made a decisive contribution to the demythologization of the neohumanist–idealist reception of Greece with his portrayal of Greek society.

Burckhardt's greatest success by far among his audience, apart from *Cicerone*, was with *The Civilisation of the Renaissance in Italy*. This is due to a consciously narrative and nonscientific diction, the consistency of his analysis and references to contemporary society kept continuously alive in the subtext, which consistently refer to the dialectic of the rationalization of the individual relation to the world and the development of modern subjectivity. Burckhardt hereby made an important contribution to a structural analysis of the modern age. He overcame the thought of individuality of German historicism with his typological thinking, which orders material according to the aspect of 'cultural importance,' and anticipated Max Weber's cultural historical method. The potencies theory is reminiscent of the conception of equal systems which, however, exercise an influence over each other, or of historical areas of reality in modern sociology (Talcott Parsons for example). Burckhardt's potencies theory differs though, in that economy and society are not taken into account, and in the important position of religion and the specific function of culture, among other factors. Nor did Burckhardt, who always portrayed himself as an unsystematic thinker, develop a 'historical method.' Yet he did make an essential contribution to both the theoretical and methodological reflection of historical research. He analyzed the relationship of prejudice and historical judgement and the effects of the emancipation of knowledge from political rule and religious belief more precisely than most of his contemporaries. In the accelerated social and cultural change of the 'Revolutionary Age' he discovered an elementary compensatory need for the preservation of traditions. He encountered every tendency to political–social instrumentalization of historical knowledge, out of progressive or conservative interest, with extreme skepticism. In conscious contrast to the contemporary professionalization of historical science he retained an older, personal understanding of learning and education. Historical education should not contribute to an alteration of the outer reality, but widen the mental horizon of those seeking knowledge. With this maxim, his independent cultural historical method, and his art of portrayal he not only in fact influenced the educational horizon of those interested in history to a great degree, but also had a stimulating effect on the increasingly specialized areas of historical science (ancient history, modern history, history of art).

See also: Art History; Civilization, Concept and History of; Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; Historicism; Scientific Disciplines, History of.

Bibliography

Burckhardt, J., Gesamtausgabe, Dürr, E., et al.
von Boch, S., Hartau, J., Hengeross-Dürkop, K., Warnke, M., (Eds.) 2000 Kritische Gesamtausgabe, vol. 6: Beiträge zur Kunstgeschichte von Italien, vol. 10: Aesthetik der bildenden Kunst.
Burckhardt, J., 1949–86. Briefe. Vollständige und kritisch bearbeitete Ausgabe, Burckhardt M (Ed.) vol. 10.
Burckhardt, J., 1974. Vorlesung über die Geschichte des Revolutionszeitalters in den Nachschriften seiner Zuhörer, Rekonstruktion des gesprochenen Wortlautes. Ernst Ziegler, Basel/Stuttgart.
Burckhardt, J., 1982. Über das Studium der Geschichte. Der Text der weltgeschichtlichen Betrachtungen aufgrund der Vorarbeiten von E. Ziegler, nach den Handschriften. Munich.
Gilbert, F., 1992. Geschichte, Politik oder Kultur? Rückblick auf einen klassischen Konflikt.
Große, J., 1997. Typus und Geschichte. Eine Jacob-Burckhardt-Interpretation.
Hardtwig, W., 1974. Geschichtsschreibung zwischen Alteuropa und moderner Welt.
Guggisberg, H.R. (Ed.), 1994. Umgang mit Jacob Burckhardt.
Hardtwig, W., Burckhardt, J., 1990. Weber M zur Genese und Pathologie der modernen Welt. In: Geschichstkultur und Wissenschaft, pp. 189–223.
Jaeger, F., 1994. Bürgerliche Modernisierungskrise und historische Sinnbildung. Kulturgeschichte bei Droysen, Burckhardt und Max Weber.
Kaegi, W., Burckhardt, J., 1947. Eine Biographie.
Löwith, K., Burckhardt, J., 1966. Der Mensch inmitten der Geschichte, second ed.
Schulin, S., 1983. Burckhardts Potenz- und Sturmlehre.
Siebert, I., Burckhardt, J., 1991. Studien zur Kunst- und Kulturgeschichtsschreibung.

Bureaucracy and Bureaucratization

Marshall Meyer, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 2, pp. 1402–1406, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Bureaucracy is an epithet, but it is also rationalization of administration. In contrast to pre-bureaucratic forms of organization, bureaucracy rationalizes reporting relationships, procedures, careers, and compensation. Even so, bureaucracy is fraught with dysfunctions, among them resistance to change, endemic power struggles, and the tendency to elaborate structure and process. Today, large firms and government agencies retain many of the elements of bureaucracy, anti-bureaucratic rhetoric notwithstanding. Conventional indexes show that the level of bureaucratization in Western societies to have remained essentially flat over the last decade. However, there has been a proliferation of new organizational forms, for example entrepreneurial organizations, network organizations, and virtual organizations, which attempt to achieve the benefits of rationalization without incurring the costs associated with bureaucratic administration.

Bureaucracy has several meanings, some positive, most negative. The sociologist Max Weber (1946) coined the term 'bureaucracy' and treated it as synonymous with rational organization: bureaucracy embodies the ideals of rational–legal authority, under which all decisions, except executive decisions, are based on rules that are internally consistent and stable over time. Political scientists tend to think of bureaucracy as governance by bureaus having the following characteristics: bureaus are large, are staffed by full-time employees who have careers within the organization, and rely on budget allocations rather than revenues from sales since their outputs cannot be priced in voluntary *quid pro quo* transactions in the market (Downs, 1967; Wilson, 1989). There is a third definition of bureaucracy, which is far less flattering: bureaucracy is inefficient organization, is inherently antidemocratic, cannot adapt to change, and, worse, exacerbates its own errors (Crozier, 1964). Discussions of bureaucracy tend to be ideologically tinged. The political left emphasizes the rationality and neutrality of government while downplaying the power of bureaucracy itself, while the right uses bureaucracy as an epithet or shibboleth and focuses on bureaucracy's antidemocratic tendencies and inefficiencies. Outside of academic circles, the rightist view of bureaucracy as inefficient and antidemocratic administration prevails. At least in the USA, very few positive references to bureaucracy can be found in the press. For most people, bureaucracy is unbusinesslike administration, an epithet or an accusation (Adler, 2012).

Bureaucracy versus Traditional Organizations

Weber's analysis of bureaucracy begins by comparing the structure of bureaucracies with traditional organizations. This comparison is made along four dimensions: differentiation, integration, constraints, and incentives.

Compared to traditional organizations, bureaucracies are highly differentiated. There is horizontal division of labor: jobs are specialized and responsibilities are strictly delimited. There is vertical division of labor as well: there are higher and lower offices, the latter subordinate to the former, all of which are ultimately accountable to the head of the organization. Perhaps most importantly, there is a clear differentiation of official duties from personal interests and obligations, what Weber calls separation of home from office. Traditional organizations, by contrast, are undifferentiated. They do not delimit responsibilities, do not separate higher from lower offices, and do not distinguish personal from official capacities-instead, the organization mirrors the stratification system of the larger community.

Compared to traditional organizations, bureaucracies have numerous integrating mechanisms, among them written rules and regulations, procedures for selection and advancement of officials, and a specialized administrative staff charged with maintaining these rules and procedures. Written rules, formal procedures, and specialized administrative staffs are largely absent from traditional organizations – actions taken by traditional organizations need not be consistent.

Compared to traditional organizations, bureaucracies constrain the conduct of officials while offering powerful incentives for compliance. There are several types of constraints: actions must be justified in terms of the larger purposes of the organization; actions must be guided by the norm of impersonality, that is, detachment and objectivity is required in all decisions; and advancement is contingent on contributing to the purposes of the organization. Traditional organizations have no distinctive purposes apart from the purposes of the people participating in them; norms of impersonality and objectivity are practically nonexistent because the organization and the community are indistinct; and there is no possibility of advancement within the organization because, again, stratification within the organization is determined by the stratification of the community.

The incentives offered by bureaucracies include the prospect of a lifetime career, salaries paid in cash rather than in kind, and (in Europe if not the USA) a modicum of prestige attached to the status of the official. Careers, salaries, and prestige based on position within the organization do not exist under traditional administration. The elements of differentiation, integration, constraints, and incentives render bureaucracies more powerful than traditional organizations yet more responsive to central

authority. The power of bureaucracies results from their capacity for coordinated action. Traditional organizations are incapable of coordinating large-scale action, except temporarily when the interests of elites coincide. The responsiveness of bureaucracies to central authority arises from top-down control of ultimate decision premises, the policies and objectives of the organization, which is impossible in traditional organizations where authority is fragmented. Paradoxically, the same elements that make bureaucracies powerful yet responsive to central authority render individual bureaucrats nearly powerless. Limited responsibilities, subordination to higher authority, and numerous rules and regulations, as well as the norm of impersonality deprive bureaucrats of latitude in decision-making. The powerlessness of individual bureaucrats, moreover, is exacerbated by their dependence on the organization for their income and social standing. Weber claims that the capacity for coordinated action, responsiveness to central authority, and dependence of individual bureaucrats on the organization promote administrative efficiency. 'Precision, speed, unambiguity, knowledge of the files, continuity, discretion, unity, strict subordination, reduction of friction and of material and personal costs – these are raised to the optimum point in the strictly bureaucratic administration...' (Weber, 1946: p. 214). The context in which Weber was writing is critical. Weber was not comparing modern businesses with bureaucracies. Rather, he was comparing bureaucracies capable of coordinated large-scale action with traditional organizations that were not.

Bureaucracy versus Business Administration

Many of the elements of bureaucracy identified by Weber are present in modern business organizations, but some are not. Like bureaucracies, large businesses are highly differentiated – jobs are specialized, a chain of command runs from higher to lower offices, and people are expected to ignore personal interest when acting in their official capacities. As in bureaucracies, many of the integrative mechanisms used are also present in large businesses, among them written regulations, procedures governing selection and advancement (today called human resource management), and specialized administrative staff. And similar constraints operate in bureaucracies and businesses – actions must be justified by the purposes (today called strategy) of the organization, norms of impersonality prevail, and advancement depends on contributing to the organization's purposes. The similarity of the administrative structures of bureaucracies and business firms occurs for several reasons. Many administrative practices arise from bounded rationality limits, people's limited capacity to make and execute complicated decisions. No .one person is capable of assembling an airplane or winning a war (Scott and Davis, 2007). Instead, complicated activities are coordinated through administrative hierarchies that transform broad decision premises (maximize shareholder value, win the war) into specific programs of action (marketing plans, battle plans) and guided by general rules that people accept as conditions of employment (Simon, 1976).

Research on organizational structures also shows that organizational size is the principal determinant of administrative structure for both bureaucracies and businesses (Blau and Schoenherr, 1971). Both bureaucracies and businesses, moreover, are constrained by norms of rationality, which means that people expect to find a chain of command, consistent rules and purposes, and rewards based on qualifications and accomplishments whether or not they are technically necessary (Meyer and Rowan, 1977).

In other respects, however, businesses and bureaucracies have evolved differently in the last 50 years. Public and private administration were similar at the time Weber was writing in the early 1900s. Much of the US public sector was modeled explicitly after the private sector at that time. It is not accidental that the reform movement in the USA, which called for governmental administration devoid of politics (or, to use another term, 'businesslike' administration), coincided with the emergence of scientific management, which called for rational selection, training, and supervision within firms. Nor is it accidental that in the 1940s, the same theory of administration was believed to apply to public- and private-sector enterprises (Gulick and Urwick, 1937). Public and private organizations have diverged in the last 50 years, however. Divergences have occurred in several domains, most notably in organizational design, accounting practices, and incentives and performance measurement. With respect to organizational design, firms now choose from among diverse organizational designs. Many larger firms have alternated between functional and divisional organizational designs, that is, between designs where the principal units are specialized by function (such as purchasing, manufacturing, and sales) so as to maximize scale economies, and designs where the principal units are self-contained businesses containing a full array of functions so as to focus attention on bottom-line objectives. Some firms have experimented with matrix organization, where people report to both product and functional managers, and still other firms have more complicated designs where some functions are centralized while others are dispersed across geographic or product units. Public bureaucracies, in contrast to business firms, exhibit very little variety in organizational design. Almost all remain organized by function as they were 90 years ago. To be sure, some central government functions have been decentralized to local governments (this is called 'loadshedding' in the USA), contracted to the private sector, or privatized outright, but the core organizational design for public bureaucracies remains functional.

With respect to accounting, public-sector agencies have departed substantially from business practice. At the beginning of the twentieth century, public entities issued consolidated financial reports and maintained capital accounts just like private businesses. Consolidated accounting gave way to much more complicated fund accounting during the 1920s, when it was believed necessary to segregate revenues and expenditures intended for different purposes into separate funds. Capital accounting has all but disappeared from the public sector – few governments distinguish between current and capital expenditures – though accounting for long-term indebtedness is retained because it is required by creditors.

With respect to incentives and performance measurement, public sector bureaucracies face more complicated issues than private businesses and, as a consequence, appear to lag substantially behind business. Businesses use several

approaches to measure and reward performance. To begin with, most publicly-held businesses define performance as total return to shareholders and reward performance so defined using techniques like value-based management. Under value-based management, executive compensation is based largely on gains in share prices, and the compensation of middle managers is based on a metric (for example, return on invested capital or economic value added) believed to drive shareholder value. The non-financial drivers of financial performance are captured in measures of internal process and customer satisfaction. Moreover, firms' internal operations are gauged against industry benchmarks assembled by third parties, usually consultants or trade associations. There is no analog to total shareholder returns in the public sector, at least none that can be easily measured – the electoral process is a coarser measure of performance than the stock market, and it is difficult to reward civil servants for electoral performance. Public agencies are taking some initial steps toward measuring non-financial performance by setting objectives and gauging progress toward them – performance measurement of this sort is part of the initiative, in the USA, to 'reinvent government' (Osborne and Gaebler, 1992) and also by measuring customer (constituent) satisfaction. Even so, competitive performance assessment hardly exists for government – the concept of benchmarking has not yet penetrated the public sector. In the USA, at least, performance comparisons across governmental units are strongly resisted, mainly on the grounds that the functions and outputs of governmental units cannot be compared across jurisdictions.

Liabilities of Bureaucracy

There is widespread agreement that bureaucracies suffer some liabilities in comparison with business. Many of these liabilities of bureaucracy are systemic in that they arise from the structure of bureaucracy rather than from people's predilections. Sociologists and political scientists have focused on what they call dysfunctions of bureaucracy, among them displacement of goals, so-called vicious circles in which dysfunctions feed on one another, and spiraling bureaucratic growth. Economists, by contrast, have drawn attention to the efficiency disadvantages of bureaucracies as compared to firms, asking whether, in general, nonmarket transactions are inefficient compared to market transactions, and, specifically, whether the funding of bureaucracies through budgets rather than market transactions results in overproduction of bureaucratic services. These potential liabilities of bureaucracy should be reviewed *seriatim*.

Displacement of Goals

Bureaucracies are known for rigid adherence to rules and procedures, even when rules and procedures appear to impede the objectives of the organization. The notion of goal displacement provides both a description and an explanation for this seemingly nonrational conduct. Goal displacement, following Robert K. Merton (1958), describes the process whereby means become ends in themselves, or 'an instrumental value becomes a terminal value'. The displacement of goals is especially acute in settings – such as bureaucracies – where the following conditions exist: the technical competence of officials consists of knowledge of the rules; advancement is contingent on adherence to the rules; and peer pressure reinforces the norm of impersonality, which requires rules and procedures to be applied with equal force in all cases. What is important is that goal displacement, at least as originally conceived, argues that bureaucracies are efficient in general – under conditions anticipated by their rules and procedures – but inefficient in circumstances that cannot be anticipated. The implications of goal displacement for innovation and new product development have been realized only gradually: bureaucracy can be antithetical to innovation.

Vicious Circles

A more thorough critique of bureaucracy argues that dysfunctions are normal rather than exceptional and, moreover, that dysfunctions accumulate over time such that organizational stasis is the expected outcome. The elements of the vicious circle of bureaucratic dysfunctions are impersonal rules that seek to limit the discretion of individual workers, centralization of remaining decisions, isolation of workers from their immediate supervisors as a consequence of limited decision-making authority, and the exercise of unofficial power in arenas where uncertainty remains.

Thus, as Michel Crozier observes (1964), maintenance people exercise undue influence in state-owned factories because their work is inherently unpredictable and cannot be governed by rules. The logic of vicious circles, it should be pointed out, yields several consequences. To begin, new rules will arise to eliminate whatever islands of power remain in the organization, but these rules will trigger further centralization, isolation, and power plays as new sources of uncertainty arise. Additionally, to the extent that the organization is opened to uncertainties arising externally, line managers have the opportunity to reassert power that would otherwise erode through the dynamics of vicious circles. External crisis, in other words, may be an antidote to bureaucracies' tendency toward rigidity over time.

Spiraling Growth

Bureaucratic systems also tend toward growth, other things being equal (Meyer, 1985). Until recently, growth of government and of administrative staff in private firms was endemic. The causes of growth lie in several factors, but chief among them are people's motives for constructing organizations in the first place. People construct formal organizations in order to rationalize or make sense of otherwise uncertain environments; organizations, in fact, succeed at making the world more sensible; as a consequence, there is continuous construction of bureaucracy and hence bureaucratic growth as people attempt to perfect their rationalization of an inherently uncertain world. Three comments are in order.

1. The logic of bureaucratic growth is built into neoclassical administrative theory developed by Simon (1976) and others. Irreducible uncertainty in the environment, in conjunction with the belief that administrative organization can rationalize uncertainty, will result in continuous growth in administration.

2. The growth imperative is so strong that deliberate campaigns to 'downsize' or 'restructure' organizations must be launched in order to achieve meaningful reductions in staff.
3. Inefficiency. Economists have asked persistently without resolution whether bureaucracies are inherently less efficient than private-sector enterprises. Several answers have been proffered, none fully satisfactory.

From the 1940s to the present time, the Austrian school of economics, von Mises (1944) and others, have argued that any departure from market principles yields both inefficient transactions and antidemocratic tendencies. This position has proved difficult to reconcile with contemporary transaction-cost theories (Williamson, 1975), which argue that hierarchies may be more efficient than markets under some circumstances. In the 1970s, the efficiency question was cast somewhat differently: might bureaus, which depend on budgets for their sustenance, overproduce compared to firms subject to the discipline of the market? (Niskanen, 1971) Here too the answer was equivocal, as analysis showed that rent-maximizing monopolists would have similar incentives to overproduce whether they were located in public bureaucracies or private firms. Despite the uncertain analytic underpinnings for the belief that bureaucracies are more apt to harbor inefficiencies than private-sector organizations, privatization of governmental functions is occurring rapidly and with positive results in many countries. It is unclear whether the liabilities of public bureaucracies are simply the liabilities of established organizations that have been shielded from extinction for too long, or whether bureaucracies suffer disadvantages in comparison with private organizations because of weak incentives and inconsistent objectives.

Research on Bureaucracy

Organizational research and research on bureaucracy were once synonymous or nearly so, as the bureaucratic model was believed descriptive of all organizations, for-profit, non-profit, and governmental. Case studies of bureaucracy written during the 1950s and 1960s encompassed government agencies and industrial firms alike as evidenced by titles like Gouldner's (1954) *Patterns of Industrial Bureaucracy*. Early quantitative research on organizations, such as the work originating at Aston University and the University of Chicago, focused mainly on relations among elements of organizational structure (size, hierarchy, administrative ratio, formalization, centralization, etc.) that flowed from the bureaucratic model implicitly if not explicitly. As attention shifted to external causes of organizational outcomes, however, the bureaucratic model has become less relevant to organizational theory. Thus, for example, the key causal variable in resource dependence models of organizations (Salancik and Pfeffer, 1978) is control of strategic resources, which is more germane to businesses than to government bureaus. The key dependent variables in organizational population ecology (Hannan and Freeman, 1984) are births and deaths of organizations, which are infrequent in the public sector (see Kaufman, 1976). The new institutional theory of organizations has downplayed Weber's notion of bureaucracy as rational administration and has substituted for it the notion that all organizations, bureaucratic and nonbureau- cratic alike but especially the former, pursue social approval or legitimation by appearing to be rational rather than actual efficiency outcomes. Almost alone among major organizational theorists, Philip Selznick argues that the problem of bureaucracy remains a central issue in organizational theory: '... the ideal of an effective, fair, and responsive bureaucracy remains elusive. Our society desperately needs organized ways of dealing with social problems; we cannot rely solely on market strategies. Yet the specter of bureaucracy still haunts and repels, still saps public confidence and weakens support for collective action' (Selznick, 1996: p. 276).

Several streams of research on bureaucracy remain, although they are at the periphery of organizational theory. One stream of research, largely in the field of public administration, focuses on the possibility of bureaucratic reform, of reducing the otherwise endemic dysfunctions of bureaucracy. This research does not promise solutions; quite the opposite, it finds bureaucratic dysfunctions intractable. For example, reform initiatives almost inevitably create new positions for people to oversee the reform, resulting in new layers of administration and 'thickening' of each layer (Light, 1997). Decentralizing managerial authority in public bureaucracies is offset by a corresponding centralization of political authority, with little net improvement in efficiency or responsiveness (Maor, 1999). Attempts to 'reinvent government' by reshaping bureaucracies into so-called performance-based organizations are usually frustrated by conflict over what the performance standards should be (Roberts, 1997).

Another stream of research, pursued mainly by economists, focuses on the impact of incentives and performance standards on performance outcomes in bureaucracies. The tentative conclusions of this research are, again unpromising (Carnis, 2011). Only weak incentives are available in bureaucracies. Not only do bureaucracies perform multiple tasks, but they also answerable to multiple constituencies or, in the argot of agency theory, principals. Multiple tasks weaken incentives in all organizations because the outcomes of some tasks can be more readily observed than the outcomes of others. Multiple constituencies or principals, however, are unique to bureaucracies and further weaken incentives because each constituency will bargain separately with the bureaucracy (Dixit, 1996). Specific performance standards built into legislation do discipline the behavior of bureaucrats somewhat, but measures of short-term results can work against long-term objectives and are frequently gamed (Heckman et al., 1997).

See also: Administration in Organizations; Bureaucracy, Sociology of; Corporate Culture; Corporate Governance; Hierarchies and Markets; Industrialization; Management Accounting; Modernization, Sociological Theories of; Organizations: Authority and Power; Public Bureaucracies.

Bibliography

Adler, P.S., 2012. The sociological ambivalence of bureaucracy: From weber via gouldner to marx. Organization Science 23 (1), 244–266.
Blau, P.M., Schoenherr, R., 1971. The Structure of Organizations. Basic Books, New York.
Carnis, L.A.H., 2011. New perspectives on the economic approach to bureaucracy. Quarterly Journal of Austrian Economics 13 (2), 53–78.

Crozier, M., 1964. The Bureaucratic Phenomenon. University of Chicago Press, Chicago.
Dixit, A., 1996. The Making of Economic Policy. MIT Press, Cambridge, MA.
Downs, A., 1967. Inside Bureaucracy. Boston, Little, Brown.
Gouldner, A.W., 1954. Patterns of Industrial Bureaucracy. Free Press, Glencoe, NY.
Gulick, L., Urwick, L., 1937. Paper on the science of administration. Institution of Public Administration, Columbia University, New York.
Hannan, M.T., Freeman, J., 1984. Structural inertia and organizational change. American Sociological Review 49, 149–164.
Heckman, J., Heinrich, C., Smith, J., 1997. Assessing the performance of performance standards in public bureaucracies. American Economic Review 87, 389–395.
Kaufman, H., 1976. Do Government Organizations Ever Die? The Brookings Institution, Washington, DC.
Light, P.C., 1997. Thickening Government: Federal Hierarchy and the Diffusion of Authority. The Brookings Institution, Washington, DC.
Maor, M., 1999. The paradox of managerialism. Public Administration Review 59, 5–18.
Merton, R.K., 1958. Bureaucratic structure and personality. In: Merton, R.K. (Ed.), Social Theory And Social Structure, second ed. Free Press, Glencoe, NY, pp. 195–206.
Meyer, J.W., Rowan, B., 1977. Institutionalized organizations: Formal structure as myth and ceremony. American Journal of Sociology 83 (2), 340–363.
Meyer, M.W., 1985. Limits To Bureaucratic Growth. de Gruyter, Berlin.
von Mises, L., 1944. Bureaucracy. Yale University Press, New Haven, CT.
Niskanen, W., 1971. Bureaucracy And Representative Government. Aldine, Chicago.
Osborne, D., Gaebler, T., 1992. Reinventing government: How the entrepreneurial spirit is transforming government. Adison Wesley Public Comp, Reading, MA.
Roberts, A., 1997. Performance-based organizations: Assessing the Gore plan. Public Administration Review 57, 465–478.
Salancik, G.R., Pfeffer, J., 1978. A social information processing approach to job attitudes and task design. Administrative Science Quarterly 23 (2), 224–253.
Scott, W.R., Davis, G.F., 2007. Organizations and Organizing: Rational, Natural and Open System Perspectives. Pearson Education, Upper Saddle River.
Selznick, P., 1996. Institutionalism 'old' and 'new'. Administrative Science Quarterly 41, 270–277.
Simon, H.A., 1976. Administrative Behavior, third ed. Free Press, New York.
Weber, M., 1946. Bureaucracy. In: Gerth, H., Wright Mills, C. (Eds.), Essays In Sociology. Free Press, Glencoe, NY, pp. 196–244.
Williamson, O.E., 1975. Markets and hierarchies. Free Press, New York, pp. 26–30.
Wilson, J.Q., 1989. Bureaucracy. Basic Books, New York.

Bureaucracy, Sociology of

Amanda K Damarin, Georgia Perimeter College, Dunwoody, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by M. Crozier, volume 2, pp. 1406–1410, © 2001, Elsevier Ltd.

Abstract

Bureaucracy is an organizational form that became dominant in business, government, and other arenas during the nineteenth and twentieth centuries. Since then, extensive debates among sociologists and others have centered on its nature, rationality, and social impact. This article reviews the major arguments, beginning with Max Weber's foundational ideal type and addressing subsequent work on bureaucracy's informal and cultural dimensions, external influences, internal contradictions, and pitfalls of societal bureaucratization. While some have proclaimed bureaucracy's demise amidst today's postindustrial economy, evidence suggests that it lives on – with consequences that are just as ambivalent as in Weber's time.

Bureaucracy can be understood as a form of organization that features elaborate rules and regulations, a division of labor, and a hierarchical command structure, all rationally designed to achieve established goals. Though such organizations have existed since ancient times, the term – based on *bureau*, French for 'desk' or 'office' – emerged in eighteenth century Europe, where bureaucracy proliferated alongside the expansion of large-scale commercial enterprise and centralized government agencies. The first English usages, in the 1800s, referred to bureaucracy as a 'complication of intrigues', a 'tyranny', and an 'evil', suggesting that the concept has never been without mixed evaluations and meanings (OED online, 2013; Williams, 1983 [1976]). This extends to the sociological literature on the topic, where relatively neutral analysis is accompanied by extensive critical discussion of the consequences (mostly negative) of bureaucratic organizations and the bureaucratization of modern society. Although some critiques target a straw man, most have been productive, enabling sociologists and others to better understand not only bureaucracy but also alternate forms and organizations in general. To see this, it makes sense to begin with sociology's early and appreciative treatments of bureaucracy and then examine the critical assessments that followed.

The Weberian Ideal Type

Most contemporary sociological discussion of bureaucracy refers explicitly or implicitly to the ideal type formulated by Max Weber in the early 1910s and first published in English translation in 1946. This conceptual model was based mostly on observations about government and commercial administration, but Weber and others have also used it to describe religious, educational, military, and other organizations. By this definition, the key feature of any bureaucracy is its basis in rational-legal authority, a form of domination that is accepted as legitimate by subordinates because it rests on legally enacted formal rules and on the commands of leaders chosen in accordance with them. In addition to providing legitimacy, rules – such as laws, administrative regulations, policies, or standard operating procedures – are rationally designed to facilitate achievement of the organization's goals. Thus, they permeate and define bureaucracy. Staff follow rules more than rulers: their work centers on carrying out officially prescribed duties rather than obeying superiors' personal commands, especially should the latter diverge from the former. Even the highest leaders are bound by rules; the 'chief' does not use his or her position to achieve private ends, but instead acts as the organization's 'first official' or even 'first servant' (Weber, 1978 [1921]: p. 957). The affairs of those subject to bureaucratic authority (e.g., citizens, customers, clients, etc.) are handled by the rules, not in accordance with their personal needs, connections, or identities. Rules structure bureaucratic organizations by specifying the duties and limits of particular jurisdictions within their typically elaborate divisions of labor, and by establishing hierarchies of authority and procedures for appealing commands from below. The importance of rules can also be linked to other defining features of bureaucracy, such as the separation of personal from official action and property, the appointment of staff based on technical expertise and formal qualifications, and the proliferation of records and files.

Weber saw the proliferation of bureaucratic organizations as part of a modern, Western trend toward rationalization, or increased use of means-ends reasoning in all areas of life. As with other aspects of this shift he was profoundly ambivalent about it. His assessment of the benefits and drawbacks of bureaucracy is probably best appreciated by understanding the contrast he draws between it and other organizational forms that prevailed in premodern times. Among the most common were patriarchal and patrimonial organizations, both based on traditional authority, or belief in the sacredness of age-old traditions and in the legitimacy of those ruling in accordance with them. While bureaucracy involves adherence to abstract rules, these organizations rest on personal loyalty to a traditionally chosen leader and obedience to his commands. It is the leader, not fixed rules, that determines the organization's goals, structures, and actions. The leader also chooses staff, typically based on personal relationships rather than formal qualifications. Further, official resources can legitimately be used for personal gain, which is not necessarily viewed as distinct from organizational goals so long as its pursuit does not violate tradition. Less common but also important historically were organizations based on charismatic authority, or devotion and

obedience to an individual leader due to his (or rarely, her) exceptional characteristics, such as genius or prophecy. Charismatic organizations tend to have fairly simple structures, with staff appointments based on personal connection or resemblance to the leader.

Given these reference points it may be easy to see why Weber thought bureaucracy to be technically superior to other organizational forms. Bureaucratic organizations are able to accomplish their goals with optimal "precision, speed, unambiguity … [and] reduction of friction and of material and personal costs" (Weber, 1978[1921]: p. 973) because there are staffed by individuals selected for their qualifications rather than their personal connections, and because officials treat their work as a vocation rather than an opportunity for personal benefit. Further, bureaucratic organizations can operate predictably, continuously, and indefinitely because they are built on rationally designed rules, while traditionalist and charismatic organizations are guided by the variable whims, abilities, and interests of their individual rulers. For these reasons, Weber thought bureaucratic organization to be almost necessary for commercial enterprise and government in modern, capitalist democracies. The impersonality of bureaucracy – its ability to operate without regard for specific persons or their ascribed statuses and privileges – makes it a fitting vehicle for "all pursuits of naked economic interests" (Weber, 1978[1921]: p. 975) and also for the exercise of mass democracy, which rejects privilege and demands formal equality.

However, Weber recognized that these very strengths of bureaucracy could also have negative consequences. First, what makes bureaucracy predictable and precise is that its power is vested in its rules and structure, rather than in people. This means that it disempowers its own staff: "the individual bureaucrat cannot squirm out of the apparatus into which he has been harnessed" (Weber, 1978[1921]: pp. 987–988). Also, the fixed rules, hierarchies, and divisions of labor that enable bureaucracies to operate continuously tend to become entrenched, making them difficult to alter or destroy. Further, a great irony of bureaucracy is that its emphasis on precise adherence to rules may lead to a situation in which formal rationality – Weber's term for technical, procedural calculation – gets in the way of substantive rationality, or reason directed at achieving the organization's ultimate values and goals. The relationship between bureaucracy and democracy is also fraught with contradictions. Because appointments are made based on formal qualifications, bureaucracy does entail a democratic leveling of ascribed inequalities such as feudal privileges. However, it also creates a new governing class of trained, qualified experts and thus, in its pure form at least, is not fully compatible with popular rule. Together, bureaucracy's impersonality, emphasis on formal rationality, and disempowerment of both staff and subjects combine to give it a "cool 'matter-of-factness'" that eliminates "love, hatred, and all purely personal, irrational, and emotional elements with escape calculation" (Weber, 1978[1921]: pp. 980, 975). The more bureaucracy becomes a technically perfect administrative machine, the more it becomes meaningless, and thus its growth helped transform modern society into what Weber described as a rationalized and disenchanted 'iron cage'.

Critical Approaches

Aside from this last commentary, Weber attempted to remain neutral in defining bureaucracy and weighing its consequences. Although his ideas were widely accepted, many subsequent authors have taken a more explicitly negative stance. Some critiques center on the shortcomings of Weber's ideal type, while many target bureaucracy itself, challenging its rationality by pointing to general flaws or specific conditions in which it becomes inefficient. Finally, a number of authors have addressed the negative effects of growing bureaucratization on the societies in which it occurs. We will review these themes in turn.

Departures from Weber's Model

Critics of Weber's ideal type usually do not claim that it is mistaken about the essential characteristics of bureaucracy (e.g., rules, hierarchy, division of labor, etc.), but rather that it fails to include features that are empirically evident in real bureaucratic organizations. In particular, following a common typology originated by Gouldner, Weber's definition can be understood as a 'rational systems' model which emphasizes the conscious, calculated manipulation of organizational rules and structures to achieve established goals. Partisans of the competing 'natural systems' approach see organizations as spontaneously developed, organic entities that evade such deliberate control. They fault Weber – and/or his rationalist successors in organization studies – for overemphasizing formal structure and neglecting informal relations and culture.

Recognition of the importance of informal organizational life – the unofficial relationships, groups, norms, and prestige rankings that emerge spontaneously through day-to-day workplace interactions – came early and has been a persistent feature of ethnographic research on bureaucracies. Melville Dalton's (1959) *Men Who Manage* showed that the everyday operation of bureaucratic firms was governed more by informal cliques, rivalries, and favor-swapping than by formal command hierarchies, and that some managers had informal influence that rivaled the power of their formal superiors. These findings were echoed in Robert Jackal's (1988) subsequent study of corporate managers, *Moral Mazes*, which claimed that formal bureaucracy may prevail at lower managerial ranks, but at the highest levels it is blended with a patrimonialism that features fealty relations to superiors, promotion based on connections rather than achievement, and clique-based power struggles. Both works suggest that the informal side of bureaucracy is just as important as formal structure in explaining organizational action. Gouldner's earlier (1954) *Patterns of Industrial Bureaucracy* went further claiming that whether and how official bureaucratic rules are enforced is in fact decided through informal negotiations between workers and managers. Here, informal relations actually determine formal structure rather than simply coexisting with it.

Culture – the meanings, beliefs, values, norms, and symbols that are shared within organizations – has also been emphasized in recent research, where it is alternately seen as a cause or an effect of structure. Some scholarship focuses on how organizational cultures originate and vary at the societal level (Crozier, 2009[1964]; Dore, 1983), while others see

intranational diversity in organizational cultures as a deliberate result of managerial choices or an unintended effect of structural variation (Morrill, 1995); still others emphasize the distinctive subcultures that emerge within particular organizational subunits via day-to-day interaction (Jackall, 1988; Morrill, 1995). Regardless of origin, organizational culture is thought to have important consequences. Many have emphasized its positive functions, such as bringing members together and giving them a shared sense of purpose, though less sanguine outcomes are also possible (Smircich, 1983). One effect that has received extensive attention is 'normative control', in which organizational members' actions are directed by cultural norms and values that encourage their spontaneous assent to and compliance with organizational objectives. How this relates to bureaucratic control is a matter of some dispute. Ouchi (1980) sees them as alternatives, arguing that normatively controlled 'clan' organizations emerge only when bureaucratic structures fail, typically due to ambiguity in evaluating individual contributions and consequent inability to elicit cooperation among members. An important counterpoint is provided by Kunda's (1992) ethnography of a high technology company where control was achieved less through typically bureaucratic formal mechanisms than with strong norms and values that made employees 'want to' achieve corporate goals. Though this might seem to exemplify culture replacing bureaucracy, Kunda instead argues that culture had itself become bureaucratized: "The essence of bureaucratic control – the formalization, codification, and enforcement of rules and regulations – does not change in principle under a system of normative control; it merely shifts its focus … from the organizational structure to the organizational culture, from the members' behavior to their experience" (p. 220). Thus, strong culture and bureaucracy may be complements rather than substitutes.

Bureaucratic Contradictions

While natural systems theorists fault the Weberian bureaucratic model for overemphasizing formal structure, they typically do not contest the claim that this structure is, in itself, rational and efficient. Other scholars do, and here two streams of thinking are particularly important. One suggests that the (in)effectiveness of bureaucratic structures depends on their external environments, while the other focuses on structural irrationalities intrinsic to all bureaucracy.

Though he acknowledged social context, Weber's focus was largely intraorganizational. This emphasis became even stronger in successive scholarship, and thus it was not until the popularization of 'open systems' approaches in the 1960s that scholars began in earnest to examine how organizations are affected by their environments (Scott, 2004). Importantly, the resulting theories do not focus exclusively on bureaucracy; instead, they view the adoption of any organizational form, bureaucratic or otherwise, as influenced by contextual conditions such as levels or types of change, trust, information, power, competition, and legitimacy. One of the earliest open systems models was contingency theory, which suggests that the efficiency and success of any organizational type depends on its ability to successfully address environmental characteristics such as rate of technological change and degree of certainty about competitive conditions; bureaucratic forms are thought to be best suited to stable and certain environments (Burns and Stalker, 1961; Lawrence and Lorsch, 1967). The economic theory of transaction costs uses a similar logic of necessity to draw different conclusions: bureaucracy is not a response to but a way of creating environmental stability. When economic coordination can no longer be achieved through market exchange due to problems of bounded rationality and possible malfeasance, bureaucratic organizations emerge to reduce uncertainty and limit opportunism (Williamson, 1981). This occurred in America and elsewhere following the technological revolutions of the late nineteenth and early twentieth centuries, as increasingly complex and long-distance commercial transactions came to require the careful coordination of bureaucratic managers (Chandler, 1984). Resource dependence theory also focuses on coordination across organizational boundaries, but rather than costs it emphasizes the power asymmetries that can arise when, for instance, one organization depends on another for vital inputs that it cannot obtain elsewhere. Though such imbalances need not lead to bureaucracy they do suggest that organizational decisions, including those regarding structure, may be profoundly shaped by pressures from external parties (Pfeffer and Salancik, 2003[1978]). Similar conclusions are drawn by network theorists who examine how organizational characteristics, including bureaucratic structure, are affected by positioning in networks of competitors, suppliers, customers, workers, or other entities (Granovetter, 1985). A more cultural approach is found in the new institutionalist school's explanation of bureaucratization as the result of isomorphism, a process in which external institutional influences – demands from government and other powerful actors, uncertainty-driven imitation of other organizations, pressures from professional organizations – cause organizations to adopt similar bureaucratic structures due to high legitimacy (in the modern industrial West, at least) rather than because they yield efficiency gains (DiMaggio and Powell, 1991[1983]). Finally, organizational ecologists focus on how macrolevel trends in organizational structure, including bureaucratization, result from organizational births and deaths, which are in turn shaped by environmental conditions of legitimacy and competition (Hannan and Freeman, 1989). New organizational forms initially have few births and many deaths because they lack legitimacy; if they manage to proliferate and gain it, their spread will then be limited as increased competition for scarce resources leads to fewer births and more deaths. Overall, open systems accounts vary considerably but share the claim that no organizational form is likely to prevail under all conditions. Bureaucracy grows when mass production and distribution creates complex coordination needs, when industrial stability is high, and/or when rational-legal legitimacy dominates the cultural landscape; under different circumstances, other organizational forms become more efficient, acceptable, and common.

However, another line of critique claims that even under the best circumstances, all bureaucracies suffer from internal flaws that contradict any claims to optimal efficiency. First, while Weber emphasized impersonality, later authors have characterized bureaucratic rationalization as downright dehumanizing – particularly when extended onto the factory floor through 'scientific management' of the labor process (Braverman, 1974). The human costs of this movement toward rigid,

mindless tasks and worker disempowerment became evident in the strikes it provoked; the organizational limitations were illustrated by the famous Hawthorne experiments, which showed that factory workers who were granted some discretion, interaction opportunities, and a sense of organizational importance outproduced those in more strictly rationalized settings. These findings became foundational for the human relations school of management, which focused on increasing productivity by treating workers as fully human, emotional, and social beings, and criticized bureaucratic tendencies to turn the workplace into an alienating machine (Kanter, 1977). Further, bureaucracy may endow its inhabitants with 'trained incapacity', Veblen's (1914) term for the inability of highly specialized workers to function beyond their narrow task areas. This results from extensive division of labor, which many have seen as central to bureaucratic rationality; in contrast, Veblen noted that staff members' blindness to business needs outside their immediate purviews could lead to "discrepancies, obstructive tactics, and disserviceable practices" (p. 248). Just as threatening to bureaucratic efficiency, especially at the highest levels, is the "Peter Principle" which suggests that "in a hierarchy every employee tends to rise to his level of incompetence" (Peter and Hull, 1969).

In addition to negative effects on employees, bureaucracy also involves more systemic problems. Echoing popular complaints about 'red tape', scholars ranging from early contingency theorists to recent students of postindustrialism have assailed the rigidity of the bureaucratic form, which prevents it from successfully innovating or adapting to a changing environment (Burns and Stalker, 1961; Saxenian, 1994). Perhaps most damning, bureaucracies may carry out actions quite distant from those they were designed to achieve. Merton (1949) used the term 'goal displacement' to describe how emphasis on formal procedure can turn rules and regulations into the ends rather than means of bureaucratic action; in such cases, "primary concern with conformity to the rules interferes with the achievement of the purposes of the organization" (p. 254). Crozier (2009[1964]) claimed that formal regulations and centralization of control result in 'vicious circles' which prevent bureaucracies from learning from their mistakes and thus interfere with the accomplishment of organizational goals. With these observations, we come full circle from Weber's claims about bureaucracy's technical superiority to arguments that it has quite irrational outcomes.

Finally, many critics are less concerned with bureaucracy's internal (ir)rationalities than with the consequences of widespread bureaucratization for the societies in which it occurs. First, the tension between bureaucracy and democracy was identified by Weber but examined with far greater pessimism by his contemporary Robert Michels. Based on a study of political parties and labor groups, Michels (1915) claimed that all democratic movements require organizations, but that as such organizations grow they develop bureaucratic structures with specialized leadership cadres who limit participation from below and use organizational power to benefit themselves rather than the masses they ostensibly serve. According to this 'iron law of oligarchy', bureaucracy is the inevitable but inevitably antidemocratic outcome of democratizing trends. In addition, the relationship of bureaucracy to ascriptive inequalities has received much attention. To some, bureaucracy vastly improves on the sexism and racism of patrimonial organization because its formal hiring procedures limit managers' ability to use ascriptive biases in selecting candidates (Reskin and McBrier, 2000) and because it actively promotes logics of universalism and meritocracy (Baron et al., 2007). However, others argue that bureaucracy furthers ascriptive inequality. For instance, extensive divisions of labor permit bureaucratic organizations to marginalize minorities and women in areas that are low-status, low-paying, and often racialized or gendered, e.g., affirmative action or ethnic marketing positions, thereby perpetuating employment inequality (Collins, 1989). Kanter (1977) takes this further, noting that bureaucratic hierarchies elicit different behaviors from staff with different levels of power and opportunity. Women and minorities have long been clustered in low-power positions and often exhibit associated behaviors such as rules-mindedness and territoriality, but through stereotyping these are attributed to their ascribed identities rather than their positions. As such behaviors are not deemed suitable for leadership positions, the stereotypes hinder these groups' upward movement within organizations. Acker (1990) makes an even more damning argument, claiming that bureaucratic organizations are inherently gender-biased because their constitutive structures, symbols, and behavioral expectations are built on the androcentric logics of surrounding societies. For instance, the bureaucratic concept of an ideal worker rests on an ostensibly genderless but actually masculine image of an individual "whose life centers on his full-time, life-long job, while his wife or another woman takes care of his personal needs and his children"; women, who cannot achieve such qualities because "to do so is to become like a man", are marginalized (pp. 149–150).

Lastly, bureaucratization has been famously assailed for its impact on the character and the temperament of the societies in which it has occurred. This began at least with Weber, for whom bureaucratization and other forms of rationalization replaced purposeful, value-oriented actors with 'specialists without spirit' and 'sensualists without heart' (1930[1904]: p. 124). The critique was extended to the American context by C. Wright Mills, whose *White Collar* (1951) described a shift in national outlook from the independence and entrepreneurialism of the nineteenth century farmer or businessmen to the dependence and powerlessness of the twentieth-century bureaucrat, the unheroic 'little man' "living out in slow misery his yearning for the quick American climb" (pp. ix–xx). William H. Whyte's more famous *Organizational Man* (1956) faulted bureaucratization for the demise of American individualism and its replacement by a 'social ethic' that emphasized group conformity and discouraged adherents from taking their conscience and destiny into their own hands.

Bureaucracy Today

More recently, numerous social scientists have suggested that bureaucracy's era is over and that it is rapidly being replaced by other organizational forms. The primary culprit (or hero, to some) is increased economic turbulence since the mid-1970s,

a product of new technologies and increasingly international, specialized, and competitive markets. Drawing on the logic of contingency theory, many scholars have surmised that under such conditions, both mass production and the bureaucratic forms that typically accompany it have grave disadvantages because their large size and complex hierarchical structures make them slow to innovate and adapt to change (Piore and Sabel, 1984; Saxenian, 1994). Thus, the world described by Whyte and Mills has given way to a new economy comprised of 'postbureaucratic', 'postindustrial', 'post-Fordist', 'flexible', and/or 'networked' organizations (Heydebrand, 1989; Piore and Sabel, 1984; Powell, 2001; Wood, 1989). Regardless of nomenclature, these new forms share some characteristics of the nonbureaucratic 'organic' organizations originally identified by contingency theorists Burns and Stalker (1961): flexible and multiskilled jobs rather than rigid divisions of labor, control via collaboration in a relatively 'flat' networked structure rather through a command hierarchy, and an emphasis on worker participation and commitment. They also have features that Burns and Stalker did not anticipate, notably including heavy use of contingent labor such as temporary, freelance, and subcontracted work (Heydebrand, 1989; Powell, 2001). Whether these turns away from bureaucracy are positive, particularly for workers, is the subject of rather polarized debate, with some linking them to increases in skill and autonomy while others emphasize the potential for labor intensification, neopatrimonialism, and insecure employment (see Powell, 2001; Smith, 1997; Vallas, 1999; Wood, 1989 for reviews).

Further, an even more recent line of scholarship suggests that bureaucracy's downfall has been overestimated. Ritzer (1993) claims that small eddies of postindustrialism have done little to stem the rising tide of McDonaldization, in which logics of bureaucratization and rationalization are greatly intensified through the application of the fast-food principles – efficiency, predictability, quantification, and control – to all areas of social life, including workplaces but also health, education, recreation, and the family. Similarly, Head (2003) points out that rather than embracing flexibility and collaboration, many firms have responded to economic distress by adopting new, hi-tech productivity practices such as business process reengineering and enterprise resource planning, which extend bureaucratic regulation to new levels of exactitude and new categories of workers. Others argue that bureaucracy may have waned but is today making resurgence, albeit often in new, more flexible hybridized forms, and that contrary to popular opinion this is not necessarily for the worse (du Gay, 2005; Piore, 2011). For workers, the insecurity, particularism, and insidious normative and peer controls found in 'flexible' firms may make formal structure look enticing (Barker, 1993; Kunda, 1992). More generally, after roughly three decades of experience with the consequences of ever-less-regulated market capitalism – Enron, the Dotcom bubble, the mortgage crisis and Great Recession, spiraling economic inequality, the Bangladeshi factory collapse, ongoing environmental degradation – many see bureaucratic emphases on planning, structure, accountability, transparency, and universalism as virtues rather than vices. For better *and* for worse, bureaucracy remains a persistent feature of the contemporary world.

See also: Authority, Social Theories of; Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Hierarchies and Markets; Management: General; Modernity; New Institutionalism in the Analysis of Complex Organizations; Oligarchy, Iron Law of; Organization: Overview; Organizations, Sociology of; Organizations: Unintended Consequences; People in Organizations; Public Bureaucracies; Weber, Max (1864–1920).

Bibliography

Acker, J., 1990. Hierarchies, jobs, bodies: a theory of gendered organizations. Gender & Society 4 (2), 139–158.
Barker, J.R., 1993. Tightening the iron cage: concertive control in self-managing teams. Administrative Science Quarterly 38, 408–437.
Baron, J.N., Hannan, M.T., Hsu, G., Kocak, O., 2007. In the company of women: gender inequality and the logic of bureaucracy in start-up firms. Work and Occupations 23 (1), 35–66.
Braverman, H., 1974. Labor and Monopoly Capital. Monthly Review Press, New York.
Burns, T., Stalker, G.M., 1961. The Management of Innovation. Tavistock, London.
Chandler Jr., A.D., 1984. The emergence of managerial capitalism. The Business History Review 58, 473–501.
Collins, S.M., 1989. The marginalization of black executives. Social Problems 36 (4), 317–331.
Crozier, M., 2009[1964]. The Bureaucratic Phenomenon. Transaction Publishers, New Brunswick, NJ.
Dalton, M., 1959. Men Who Manage: Fusions of Feeling and Theory in Administration. John Wiley & Sons, Inc., New York.
DiMaggio, P.J., Powell, W.W., 1991[1983]. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. In: Powell, W.W., DiMaggio, P.J. (Eds.), The New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago, pp. 63–82.
Dore, R., 1983. Goodwill and the spirit of market capitalism. The British Journal of Sociology 34 (4), 459–482.
du Gay, P., 2005. The values of bureaucracy: an introduction. In: du Gay, P. (Ed.), The Values of Bureaucracy. Oxford University Press, Oxford, UK, pp. 1–13.
Gouldner, A.W., 1954. Patterns of Industrial Bureaucracy. Free Press, New York.
Granovetter, M., 1985. Economic action and social structure: the problem of embeddedness. American Journal of Sociology 91 (3), 481–510.
Hannan, M., Freeman, J., 1989. Organizational Ecology. Harvard University Press, Cambridge, MA.
Head, S., 2003. The New Ruthless Economy. Oxford University Press, Oxford.
Heydebrand, W., 1989. New organizational forms. Work and Occupations 16 (3), 323–357.
Jackall, R., 1988. Moral Mazes: The World of Corporate Managers. Oxford University Press, New York.
Kanter, R.M., 1977. Men and Women of the Corporation. Basic Books, New York.
Kunda, G., 1992. Engineering Culture: Control and Commitment in a High-Tech Corporation. Temple University Press, Philadelphia, PA.
Lawrence, P.R., Lorsch, J.W., 1967. Organization and Environment. Harvard University Press, Cambridge, MA.
Merton, R.K., 1949. Social Theory and Social Structure. The Free Press, New York.
Michels, R., 1915. Political Parties (E. Paul, C. Paul, Trans.). Hearst's International Library Co., New York.
Mills, C.W., 1951. White Collar. Oxford University Press, New York.
Morrill, C., 1995. The Executive Way: Conflict Management in Corporations. The University of Chicago Press, Chicago.
OED Online, 2013. Bureaucracy, n. Oxford University Press. Available at: http://www.oed.com/view/Entry/24905?redirectedFrom=bureaucracy.
Ouchi, W.G., 1980. Markets, bureaucracies, and clans. Administrative Science Quarterly 25, 129–141.
Peter, L.J., Hull, R., 1969. The Peter Principle. Morrow, New York.
Pfeffer, J., Salancik, G.R., 2003[1978]. The external control of organizations: a resource dependence perspective. In: Handel, M.J. (Ed.), The Sociology of Organizations. Sage, Thousand Oaks, CA, pp. 233–242.
Piore, M., 2011. Beyond markets: sociology, street-level bureaucracy, and the management of the public sector. Regulation & Governance 5 (1), 145–164.

Piore, M., Sabel, C., 1984. The Second Industrial Divide. Basic Books, New York.
Powell, W.W., 2001. The capitalist firm in the twenty-first century: emerging patterns in Western enterprise. In: DiMaggio, P. (Ed.), The Twenty-First-Century Firm. Princeton University Press, Princeton, NJ, pp. 33–68.
Reskin, B.F., McBrier, D.B., 2000. Why not ascription? Organizations' employment of male and female managers. American Sociological Review 65 (2), 210–233.
Ritzer, G., 1993. The McDonaldization of Society. Pine Forge Press, Thousand Oaks, CA.
Saxenian, A., 1994. Regional Advantage. Harvard University Press, Cambridge, MA.
Scott, W.R., 2004. Reflections on a half-century of organizational sociology. Annual Review of Sociology 30, 1–21.
Smircich, L., 1983. Concepts of culture in organizational analysis. Administrative Science Quarterly 28, 339–358.
Smith, V., 1997. New forms of work organization. Annual Review of Sociology 23, 315–339.
Vallas, S.P., 1999. Rethinking post-fordism: the meaning of workplace flexibility. Sociological Theory 17 (1), 69–101.
Veblen, T., 1914. The Instinct of Workmanship and the State of the Industrial Arts. Macmillan, New York.
Weber, M., 1930[1904]. The Protestant Ethic and the Spirit of Capitalism (T. Parsons, Trans.). Routledge, London.
Weber, M., 1978[1921]. In: Roth, G., Wittich, C. (Eds.), Economy and Society. University of California Press, Berkeley, CA.
Whyte, W.H., 1956. The Organization Man. University of Pennsylvania Press, Philadelphia, PA.
Williams, R., 1983[1976]. Bureaucracy. Keywords: A Vocabulary of Culture and Society, revised ed. Oxford University Press, New York. 49–50.
Williamson, O.E., 1981. The economics of organization: the transaction cost approach. American Journal of Sociology 87 (3), 548–577.
Wood, S., 1989. The transformation of work? In: Wood, S. (Ed.), The Transformation of Work? Unwin Hyman, London, pp. 1–43.

Relevant Websites

http://criticalmanagement.org/ – Critical Management Website.
https://www.facebook.com/EconSociology – Economic Sociology and Political Economy Blog.
http://omtweb.org/omt-blog – Organization and Management Theory, a Division of the Academy of Management (Blog).
http://oowsection.org/ – Organizations, Occupations, and Work: A Section of the American Sociological Association.
http://orgtheory.wordpress.com/ – Orgtheory.net.
http://organizations.nfshost.com/ – Sociology of Organizations: Research Committee 17 in the International Sociological Association (Blog).
http://workinprogress.oowsection.org/ – Work in Progress: Blog of the American Sociological Association's Organizations, Occupations, and Work Section.

Bureaucratization and Bureaucracy, History of

Rolf Torstendahl, Uppsala University, Uppsala, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Bureaucratization as a term is discussed, and its relation to the history of bureaucracies is developed. A main distinction is made between societies that have differentiated the political power from the administrative power and those that have not. Colonial rule has formed an intermediate type, as colonies were almost entirely governed by appointed administrators, while this was not possible in Europe or the United States. Bureaucratization is not just a state phenomenon, and big business in the private sector has become more and more dependent on an efficient administration. The interrelation between welfare and bureaucracies has been observed, and a current debate is going on about the relation between democracy, bureaucracy, and welfare.

Bureaucratization is a process, and the term in itself thus refers to a historical development. It is, however, used in several distinct meanings, some of which refer to moral and political values, which are endangered by a growth of bureaucracy according to a common belief. As a descriptive term, apart from its evaluative content, bureaucratization may mean (1) a growth of numbers of administrators and administrative organs between points of reference, (2) a growth in administrative complication in the handling of matters, and (3) an increase in administrative competence and independence and in the functional division of administration. These three definitions may be interrelated, although this is not necessarily the case. The third one comes close to the analysis first made by Max Weber of the development of a bureaucratic principle (Weber, 1922).

There is also a great difference in the use of the term in different contexts. Here one may distinguish first (1) a reference to state administration only or (2) a reference to any sort of administration, whether public or private.

A third distinction may also be made. Sometimes the term bureaucratization is used in regard to states and societies irrespective of their political system. On other occasions the use is restricted to states and societies where a separation is made, at least in principle, between political and administrative duties and responsibilities. The latter limited scope of the term is, of course, much more related to the near history than to past centuries, with the nineteenth century as an intermediate period when this separation first came to be more commonly observed.

The different aspects of the concept of bureaucratization and their close relationship to political ideals regarding the scope of state interference in relation to citizen's rights and the division of responsibility between politicians and bureaucrats make it difficult to use the concept in historical and social science analysis. However, in regard to changes in power and control, it covers an important reality, which has to be differentiated from the use of administration in general. The following account goes out from Weber's analysis combined with the distinction between political and administrative responsibilities.

From Antiquity to the Nineteenth Century

There is a striking difference between the state of Egypt in Antiquity, the Mesopotamian empires of Babylon and Assur, or Rome compared to the Medieval states of Europe including the vast empire of Charles the Great. This difference has to do with the ambitions of the state. While delegation to subprinces was the rule in the feudal system that emerged in medieval Europe, the state was in principle centralized as to the important decisions in the early empires. It is often said that these empires were highly bureaucratized. The Russian historian of Antiquity Michail I Rostovcev in the first half of the twentieth century wrote several books in which he tried to show that the process of bureaucratization was going on in Egypt, the Hellenistic Empires, and the Roman Empire (Rostovcev, 1936–38, 1941, 1957). In spite of his brilliance his views are not current among ancient historians of today. Some would rather stress the limits of state power in Antiquity and others the specific structure of the ancient system of government in different parts of the world. What was common to these regions was the fusion of politics with administration. No real difference was made, and only the king or emperor exerted a political power meant to express a deliberate policy. He headed an organization where the highest officials had some political duties, but their policy making had mostly to be concealed under administrative clothing. Further down in the hierarchy, the administrative chores were the real occupation of the employees. Often, administrators were neither full-timers nor independent, and regulations tended to vary with the will of the ruler. Further, it is important that the state in Antiquity was highly centralized and that no clear difference was made between public law and private law until the late Roman Empire and, thus, no real limit was set between state obligations and rights and a private sector. Mostly the kings and emperors commanded only 'their' servants and demanded only what belonged to them, although whether these relations were of private or public nature was often not quite clear. This 'bureaucratic' state was, thus, vague in its boundaries, but, as a centralized state, it needed an efficient bureaucracy. It is hardly evident that bureaucracy should have grown or extended its procedures from one period to another or from one empire to another during the period before AD 400.

China was in many ways the same sort of bureaucratic political state as imperial Rome, but the Chinese system started earlier and continued in all the dynasties up to the end of the empire in 1911. The Confucian class of state servants was hierarchically ordered under the emperor, who had the political power in his hands. No real delegation of political power

was used, but rather a blend of political and administrative obligations, which included a strict rule regimen. The bureaucracy was organized in a refined system already in the Ch'in and Han dynasties (221 BC–AD 220). There were ministries and ministers, but even these, and some senior officials or 'excellencies' who were the highest in rank, could influence policy only via their channels to the emperor. Their responsibilities did not include the task to form a policy even for the field for which they were responsible. Hierarchy was enormously developed both within each ministry or office and also in the country as a whole. Local administration took place via organs in at least five different levels. However, it will seem that all political decision making stayed with the emperor in spite of a system of checks and balances within the bureaucracy (Loewe, 1986; Bielenstein, 1986).

The politico-bureaucratic system of China, Egypt, and imperial Rome under Antiquity did not change in principle, which makes it difficult to use the term bureaucratization in each of the cases. The fundamental structure of bureaucratic government did not develop even though specific modifications were introduced in the division of tasks and responsibilities. Except in China, Korea, and Japan, a fundamental debureaucratization took place in the centuries after AD 400. In the nomad empires of which that of the Mongols was by far the most extensive, virtually no bureaucracy existed. When the Mongols had taken China, the Chinese system of government became theirs but not in the parts of the empire that broke away from Mongol China. Bureaucracy was also kept small in the Arab and Islamic states, which were rapidly formed in the seventh to tenth centuries. European feudal princes had very few officials who were employed to administer; such officials seldom had full-time employment for their administrative work, and they could almost never rely on a rule system, which might ensure a field for decision making without interference and with some autonomy. Within the Catholic church a system resembling a bureaucracy survived.

Only with the rise of central power in the fifteenth to seventeenth centuries did administrative institutions grow again in Western Europe; in Eastern Europe the same system existed for much longer, in Russia until 1905 and partly until 1917. There are many similarities between this system of bureaucratic forms of government and the one in ancient China and Rome. A specific instruction for future bureaucrats emerged. Universities played a role in this, and China had such educational institutions from an early date. In Europe two different types of reforms increased the theoretical competence of the bureaucracy. First, in centralized monarchies there was a need for educated secretaries and fiscal employees, who could master the subtleties of foreign and internal relations and international and national law. This need was partly met through recruitment of low-estate university men, who were sometimes promoted into the nobility (the 'noblesse de robe'). Second, nobles who wanted to preserve for nobles influential posts in the immediate vicinity of central power, arranged specific educational institutions for young nobles, to make them fit for bureaucratic leadership. It also became customary to send young nobles to universities in a tour over Europe to polish their education in a fitting way. Centralized monarchy in this manner again made bureaucracy important in the European way of handling public affairs. The well-functioning of the administration of the state was a precondition for efficiency in foreign policy and warfare, and these were the ultimate aims of most of these monarchies (Raeff, 1983; Yeroshkin, 1997). In the middle of the seventeenth century military entrepreneurs were finally made servants of the state and a military bureaucracy evolved. Later 'cameralism,' the 'science' of civil administration, developed mainly in the German states (Mann, 1986/1993). The relatively poor Swedish state became a power of significance partly because of the efficiency of its bureaucratic apparatus, organized in 'collegia' defined according to functional criteria, and with (in principle) salaried employees. This organization impressed Peter I of Russia, so that he reformed the Russian bureaucracy according to the same model, although it was soon modified by his successors (Peterson, 1979; Shepelyov, 1999).

The new European bureaucracies represented some innovations in the process of bureaucratization of state and society. Some of the state bureaucrats of the seventeenth and eighteenth centuries had a real full-time employment and were salaried for their services, and rules were established to ensure that the administrations worked in an efficient manner and on a daily basis.

In the seventeenth- and eighteenth-century rush to make non-European regions dependent in the form of colonies, European states transferred part of their bureaucratic system to their colonies. When the agent was not a state but a company, like the British East India Company (EIC), it in fact acted as a quasi state and formed its own 'public' bureaucracy (Stern, 2011). This exceptional form of bureaucratization, affecting both the EIC and vast regions in India, had no counterpart in Europe.

Bureaucratization and Parliamentary Politics from the Early Nineteenth to the Early Twenty-First Century

The development of bureaucratization took a new turn from the nineteenth century in Europe and the United States, and because of European imperialism and the domination of these countries in world politics and economy this new turn made a global impact. The first half of the century was filled with parliamentary and bureaucratic reforms, and the second half saw different stages of democratization take place. Politics and bureaucracy thus became centers of societal change. Constitutional rule was the first political demand from the 'left' (relative to the political scale of the time) all over Europe, and its fundamental issue was to get rid of 'despotism' in the form of monarchic rule through bureaucracy. This demand meant to give a decisive influence to elected bodies at least in fiscal and educational matters. It became usual to contrast bureaucratic and political responsibilities.

In this situation, during the first half of the nineteenth century, a long-standing request for reform of the civil service was finally agreed to in several countries. In Germany, France, and Britain these demands led to results, and extensive reforms of their civil service took place (Mann, 1986/1993, Gillis, 1971; Charle, 1980; Thuillier, 1980; Campbell, 1955; Mackenzie and Grove, 1957). These reforms had as their main objectives to reduce or get rid of nepotism in the appointment of personnel and to abolish corruption in order to establish a certain independence and objectivity among bureaucrats. Those who have investigated these reforms have not contended that they were

entirely successful, but it is generally said that a huge improvement took place. Since these days West European bureaucracy has established a reputation for being normally impartial and honest.

During part of the nineteenth century when the ideas of economic liberalism flourished, state intervention was normally confined to a restricted area, although practice never conformed to theory. When the security of the state or vital interests of common nature to the whole society were at stake, the state should intervene. This meant that military purposes were generally excluded from the restriction for the state to engage in economic activities and also some public education for the supply of civil servants, and, in fact, common interests could be extended into new fields. In principle, however, the civil servants had to protect state interests through legal means, and the military had to protect such interests through means of force and violence. The bureaucracy in the strict sense thus was an instrument for a 'legal domination,' one of the forms of power acknowledged by Max Weber. During most of the nineteenth century bureaucracy was regarded with distrust by the liberal forces in Europe and America and its growth was seen as an evil. Conservatives, with another evaluation of law and legal forms, tended to appreciate bureaucracy as a counterweight to democratic rule and as a guarantor of equal treatment and long-term obligations of the state system. This did not imply, however, that conservatives generally were favorably inclined to any growth of bureaucracy but, rather, that they wanted it to be strong and limited.

What holds for the development of state bureaucracies in Europe and the United States is not valid for the European colonies. The European powers pursued another policy in their colonies, which were ruled almost entirely through bureaucracy. Real political bodies existed only in the 'mother' countries, and the bureaucrats in the colonies were responsible to the governments in their political center and, indirectly, to their political mandators. Political elements in the colonies were few, elected bodies had little or no influence on the administration, and, most important, administration and military power were closely united in terms of command and responsibility. For example, in India the heads of bureaucracy (the civil administration, Indian Civil Service) in each district were intertwined with the military forces in the districts. Although they were separate organizations they were used together, if required, in command and operations, and the military personnel was used to enforce the implementation of the regulations of the civil service. This system, only slightly different in the colonies of different European powers, was based on bureaucratic rule in the colony. Thus, it differed completely from what was tolerated in the colonial states at home, and it meant a bureaucratization, which was never only a legal domination but always also domination by violence.

At the close of the nineteenth century the economy of most European countries and North America started to grow rapidly after a protracted slump. Investments increased, and trade and industry grew. As shown by Jürgen Kocka, the structure of the industrial and commercial firms changed. Very big companies arose, and these companies met new organizational problems. The structure of the Siemens and Halske Company had been, earlier, one of a patrimonial type with an allegiance of the salaried employees to the owners directly. Their loyalty was ensured through bonuses and other benefits given on a personal basis from the members of the Siemens family. In the 1880s, a total reorganization took place. The Company replaced the family. General regulations were given instead of personal instructions from the leadership. Benefits were regulated and depended on the firm (Kocka, 1969). Although this may have been an unusually clear case, several other examples show that this type of impersonalized bureaucratization took place in many industrial and commercial enterprises during the late nineteenth and early twentieth centuries (Torstendahl, 1991) (*see* Business History).

Not only did bureaucratization get an inroad in each company but industries also started to organize themselves into federations both as employers and as producers and traders. Such employers' associations and industrial organizations were intended to give support to their members through a general knowledge of industries and branches. This required an administrative staff. In this way private sector enterprises, both as individual firms and as organizations of firms, took a decisive step into a formal bureaucratization of their activities in the late nineteenth or early twentieth century. Thereby they also forced the states to act. It was impossible for states to rely only on their traditional administration, when organizations of industry made demands on the state and its policy.

States had to engage competent personnel, in the first place technically educated persons who could meet the demands from industry and trade. These new bureaucrats should also assist states in their demands on industry, initially in questions of security – inspectors of ships, elevators, constructions, bridges, railroads, etc. were engaged and organized into new administrative units. Security of labor became another source of state activity (Winkler, 1974; Bruguière et al., 1985; Torstendahl, 1991). Voluntary organizations and, not least, parties were also drawn into the organizational mainstream and got a bureaucratic structure (Michels, 1911).

It is well known that World War I (*see* First World War, The) meant an enormous growth of state intervention in the economy all over Europe. In order to make resources suffice for the most urgent needs the state had to regulate vast areas of the economy entailing an extensive bureaucratization. After the war most of the new administrations were abolished. The recollection of the war years and the large-scale bureaucratization of society during these years was, however, alive with many people, and it formed a source of inspiration to some both on the left and the right (Rials, 1977).

The reshaping of Europe after World War II (*see* Second World War, The) brought a new stage to bureaucratization: the social welfare administrations. They were foreshadowed in Scandinavia and France before the war but became a general phenomenon in West Europe only after the war. States then took the lead in the economies of European societies and, with or without nationalization of industries, wanted to carry through extensive planning. The variations between West European states were considerable in regard to political will. Enthusiasm for planning and welfare was great in the socialist and social democratic camps in countries like France, Britain, and Sweden (Rousso, 1986, 1987; Morgan, 1984) (*see* Socialism: Historical Aspects). In Germany and Italy, for a long period dominated by their Catholic and conservative parties, governments also accepted some state intervention (Baring,

1982; Mammarella, 1985), accentuated with political change in the late 1960s. All over Western Europe administrations grew in size and importance.

Around the end of the 1960s complaints about bureaucratization had become numerous, and bureaucracies had begun to behave independently of politicians, no longer implementing political demands but presenting more and more of their own policies (Aberbach et al., 1981; Mayntz and Scharpf, 1975). This new phase of bureaucratization meant that politicians began a drive to reduce the influence of bureaucracy on policy making. The cutting up of big administrations became usual in Europe, and efforts to strengthen the influence of politicians in relation to bureaucracies were frequent in the 1970s and 1980s. Curtailed bureaucracies, however, showed their tenacity through branching off and reestablishing at a local or regional level. Political initiatives may have been transferred more to the politicians, but in pursuing their policy politicians were dependent on negotiations with different interest groups and different bureaucratic bodies within the public sector. In the new negotiating economy bureaucracies played a distinctive role, although mostly in the background or as participants in the groups forming policy solutions in complete packages. Bureaucratization was no longer as obvious as earlier, but it did not mean that its development was reversed (Torstendahl, 1991).

Outside of Western Europe strong presidential systems have often given rise to a process of bureaucratization different from the one in Western Europe. In the United States the president is regarded as the head of the administration as well as the head of politics, while in Western Europe the two spheres have in principle been kept apart. Bureaucratization in Europe has meant that the superiority of politics over administration has been challenged through bureaucratic expansion, while in the United States bureaucratization has meant, rather, the increasing influence of federal (public) decisions in the private sector of economy and life, balanced by the equally strong judicial sector. Strong presidential systems have arisen in many other countries in the period after World War II but often without the strong judiciary that exists in the United States. Russia is one recent example.

The parallelism of the development of bureaucratization in Europe and North America, and in the twentieth century also Australia and New Zealand, thus became less visible after around 1950. Russia, on the other hand, has not done away with large parts of its bureaucracy from Soviet times after *perestroika* starting in 1985, which might have been expected. The strong presidential rule in new Russia relies heavily on the administrative organs, and these are still intertwined with the leadership of the privatized huge industrial conglomerates. In domestic matters the president is regarded and seems to regard himself more as the leader of an administration than as a political leader, and he has no party to rely on. The Duma and the Federation Council (and their predecessors in the previous constitution) have little influence over the government and the administration. Bureaucratic rule is still prevailing, as in Soviet and Czarist times (Shevtsova, 1999, 2005).

In the first decade of the twenty-first century the discussion over public administration and managerialism has been lively, but the term bureaucratization has become rare. Different schools of thought stand against one another: new public management (NPM) advocates hierarchy and efficiency (Pollitt and Bouckaert, 2011) in a so-called neo-Weberian state, while public value management has launched ethics and integrity as rivals to efficiency and economic criteria (Moore, 2005). Further, a participatory model has been put forward as a major alternative to NPM (Gay, 2010: Chapter 10). Ultimately, this debate is concerned with the relation between bureaucracy, democracy, and welfare, and the idea of representative bureaucracy has become important (Raadschelders, 2003). However, the whole debate is concerned only with bureaucracy in the democratic states of the present world. Globally other varieties exist, and the links to the past are different.

Bureaucratization as a General Phenomenon and Object of Theory Making

The interest in the problem of bureaucratization seems to have been directly dependent on the relation between politics and administration. With regard to the interest in bureaucracies surprisingly few books have been devoted to the process of bureaucratization as a general phenomenon (primarily Jacoby, 1969; Rizzi, 1939).

A wide literature has been produced on the bureaucratization of different parts of modern Western societies. Eisenstadt, with several investigations on bureaucratic developments, gave an overview of the literature in 1958, now somewhat outdated (Eisenstadt, 1958). A general approach has been tried in regard to the bureaucratization of socialism (Hodges, 1981), but most discussions of bureaucratization have been particular in regard to the selection of societies and time periods discussed.

Max Weber brought forward the basic criteria for a discussion of bureaucratization as a social phenomenon, answering to the need for rationalization in the sphere of government. His standpoint is often rendered as an inevitability of bureaucratization and an impossibility to evade the 'iron cage' of bureaucracy. This hardly does justice to Weber's sophisticated argument about men's possibility to counteract what seem to be probable predictions (Mommsen, 1974, 1989; Schluchter, 1989). However, several critics of contemporary culture in post-World War II Europe and America have taken bureaucratization as their point of departure for a demand to change a development into a civilization run by administration.

See also: Administration in Organizations; Bureaucracy and Bureaucratization; Civil Service; Democracy, History of; Development and the State; Parliaments, History of; Public Bureaucracies; State Formation, Theory of; Weber, Max (1864–1920); Weberian Social Thought, History of; Welfare State, History of.

Bibliography

Aberbach, J.D., Putnam, R.D., Rockman, B.A., 1981. Bureaucrats and Politicians in Western Democracies. Harvard University Press, Cambridge, MA.

Baring, A., 1982. Im Anfang war Adenauer: Die Entstehung der Kanzlerdemokratie. DTV, Munich, Germany.

Bielenstein, H., 1986. The Institutions of Later Han, in Cambridge History of China. In: The Ch'in and Han Empires, 221 BC–AD 220, vol. 1. Cambridge University Press, Cambridge, UK, pp. 491–519.
Bruguière, M., Clinguart, J., Guillaume-Hofnung, M., Machelon, J.-P., Plessis, A., Rials, S., Thuillier, G., Tulard, J., 1985. Administration et Controle de l'Economie 1800–1914. Libr. Droz, Geneva, Switzerland.
Campbell, G.A., 1955. (new ed. 1965) The Civil Service in Britain. Duckworth, London.
Charle, C., 1980. Les Hauts Fonctionnaires en France au XIXe Siecle. Gallimard/Julliard, Paris.
Eisenstadt, S.N., 1958. Bureaucracy and bureaucratization: a trend report and a bibliography. Current Sociology 7, 2.
Gay, P.B., 2010. The Politics of Bureaucracy: An Introduction to Comparative Public Administration, sixth ed. Routledge, London & New York.
Gillis, J.R., 1971. The Prussian Bureaucracy in Crisis, 1840–1860: Origins of an Administrative Ethos. Stanford University Press, Stanford, CA.
Hodges, D.C., 1981. The Bureaucratization of Socialism. University of Massachusetts Press, Amherst, MA.
Jacoby, H., 1969. Die Bürokrati sierung der Welt: Ein Beitrag zur Problemgeschichte, Soziologische Texte. In: Neuwied [1973], The Bureaucratization of the World, vol. 64. University of California Press, Berkeley, CA.
Kocka, J., 1969. Unternehemensverwaltung und Angestelltenschaft am Beispiel Siemens 1847–1914. Zum Verhaltnis vom Kapitalismus und Burokratie in der Deutschen Industrialisierung. Stuttgart, Germany.
Loewe, M., 1986. The Structure and Practice of Government, in Cambridge History of China. In: The Chin and Han Empires, 22 BC–AD 220, vol. 1. Cambridge University Press, Cambridge, UK, pp. 463–90.
Mackenzie, W.J.M., Grove, J.W., 1957. Central Administration in Britain. Longmans Green, London.
Mammarella, G., 1985. L'Italia Contemporanea (1943–1985). Il Mulino, Bologna, Italy.
Mann, M., 1986/1993. The Sources of Social Power, vols. 1–2. Cambridge University Press, Cambridge, UK.
Mayntz, R., Scharpf, F.W., 1975. Policy-Making in the German Federal Bureaucracy. Elsevier, Amsterdam.
Michels, R., 1911. Zur Soziologie des Parteiwesens in der Modernen Demokratie. W. Klinkhardt, Leipzig, Germany.
Mommsen, W.J., 1974. The Age of Bureaucracy: Perspectives on the Political Sociology of Max Weber. Blackwell, Oxford, UK.
Mommsen, W.J., 1989. The Political and Social Theory of Max Weber: Collected Essays. Polity Press, Oxford, UK.
Moore, M.H., 2005. Creating Public Value: Strategic Management in Government. Harvard University Press, Cambridge, MA.
Morgan, K.O., 1984. Labour in Power, 1945–1951. Clarendon Press, Oxford, UK.
Peterson, C., 1979. Peter the Great's Administrative and Judicial Reforms: Swedish Antecedents and the Process of Reception. In: Rattshistoriskt Bibliotek, vol. 29. Nordiska bokhandeln, Stockholm.
Pollitt, C., Bouckaert, G., 2011. Public Management Reform: A Comparative Analysis, third ed. Oxford University Press, Oxford, UK.
Raadschelders, J.N.C., 2003. Government: A Public Administration Perspective. Sharpe, New York & London.
Raeff, M., 1983. The Well-Ordered Police State: Social and Institutional Change Through Law in the Germanies and Russia, 1600–1800. Yale University, New Haven, CT.
Rials, S., 1977. Administration et Organisation. De L'organisation de la Bataille d la Bataille de L'organisation dans L'administration Française. Beauchesne, Paris.
Rizzi, B., 1939. La Bureaucratisation du Monde. Paris [1985], The Bureaucratization of the World, first Am. ed. Free Press, New York.
Rostovcev, M.I., 1936–1938. A History of the Ancient World, vols. 1–2. Clarendon Press, Oxford, UK.
Rostovcev, M.I., 1941. The Social and Economic History of the Hellenistic World, vols. 1–3. Oxford University Press, Oxford, UK.
Rostovcev, M.I., 1957. The Social and Economic History of the Roman Empire, vols. 1–2. Clarendon Press, Oxford, UK.
Rousso, H. (Ed.), 1986. De Monnet k Masse. Enjeux Politigues et Objectifs Economiques dans le Cadre des Quatre Premiers Plans (1946–65). Editions du Centre National de la Recherche Scientifique, Paris.
Rousso, H. (Ed.), 1987. La Planification en Crises (1965–1985). Editions du Centre National de la Recherche Scientifique, Paris.
Schluchter, W., 1989. Rationalism, Religion, and Domination: A Weberian Perspective. University of California Press, Berkeley, CA.
Shepelyov, L.E., 1999. Chinovnii mir XVIII-Nachalo XX v [The World of the Civil Service from the 18th to the Beginning of the 20th Century]. Iskusstvo-SPB, Saint-Petersburg, Russia.
Shevtsova, L., 1999. Yeltsin's Russia: Myths and Reality. Carnegie Endowment for International Peace, Washington, DC.
Shevtsova, L.E., 2005. Putin's Russia. Carnegie Endowment for International Peace, Washington, DC.
Stern, P.J., 2011. The Company-State: Corporate Sovereignty and the Early Modern Foundations of the British Empire in India. Oxford University Press, Oxford & New York.
Thuillier, G., 1980. Bureaucratic et Bureaucrats en France au XIXe Siecle. Droz, Geneva, Switzerland.
Torstendahl, R., 1991. Bureaucratisation in Northwestern Europe, 1880–1985: Domination and Governance. Routledge, London.
Weber, M., 1922. Wirtschaft und Gesellschaft, Mohr, Tübingen [1978]. In: Economy and Society, vols. 1–3. Bedminster Press, New York.
Winkler, H.A. (Ed.), 1974. Organisierter Kapitalismus: Voraussetzungen and Anfange. Vandenhoeck & Ruprecht, Gottingen, Germany.
Yeroshkin, N.P., 1997. Istoria gosudarstvennykh uchrezhdenii dorevolyutsionnoi Rossii [A History of State Administration in Prerevolutionary Russia]. Tretii Rim, Moscow.

Burke, Peter (1937–)

Stephen Mennell, School of Sociology, University College Dublin, Dublin, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article contains a brief biography of the British cultural historian Peter Burke, and an outline of his very diverse and extensive publications, which focus principally on early modern Europe but sometimes stray backward to the late Middle Ages and forward into the nineteenth and twentieth centuries. Burke is especially well known among social scientists for his textbook on *Sociology and History*, revised under the title *History and Social Theory*.

Peter Burke (1937–) is a distinguished, wide-ranging, and highly productive British cultural and social historian. His specialist area is early modern Europe, although his writing is not narrowly confined to that. He is noted for his command of and research in several European languages – Latin, French, Italian, Spanish, Portuguese, Dutch, and German – and also for his extensive knowledge of social theory, on which he frequently draws in his writings. Social scientists are likely to be especially aware of his textbook *Sociology and History*, first published in 1980 and subsequently revised and expanded in 1992 and 2005 under the title *History and Social Theory*.

Burke was born in London; his father was of Irish origin, but his mother was the daughter of Jewish émigrés from Eastern Europe. In his essay in intellectual autobiography (2009a), he recalls that, for reasons that are in retrospect not quite explicable, from quite early in his childhood he wanted to be a historian. From St Ignatius's College, a Jesuit grammar school in Stamford Hill, north London, at 18 years he won an Open Scholarship to St John's College, Oxford. But in 1955, conscription was still in force in the United Kingdom, and the college advised Burke to do his 2 years of National Service in the army before going up to university. He became a pay clerk – a private soldier, not an officer – in the Royal Corps of Signals, and was posted to Singapore, where he served for 19 months. He was attached to the multiethnic, mainly locally recruited, Singapore District Signal Regiment.

> To me, an innocent fresh from an English school, the world of the regiment was excitingly scandalous. Everything was for sale. Drivers siphoned petrol out of their vehicles and sold it. Objects from the military stores turned up in the so-called 'thieves market' in the centre of town, offered for sale for a fraction of the official price.

The regiment, he writes, "bears some responsibility for making me into a particular breed of historian, fascinated by cultural encounters and misunderstandings and the distance between official and unofficial rules, a topic central to social history." Singapore also awakened an anthropological interest in other cultures.

Burke finally arrived in Oxford as an undergraduate in 1957, where he was tutored principally by Keith Thomas and the architectural historian Howard Colvin, and where he attended the lectures of other outstanding historians such as Christopher Hill, Lawrence Stone, Richard Southern, and John Hale. The Oxford history syllabus of those days was, he recounts, rather narrow, focusing predominantly on political and to a lesser extent economic history, with compulsory examination papers in English history and English constitutional documents. But Burke's interests were already turning to the social and cultural: he chose the Italian Renaissance as his third-year Special Subject, because "this was the only Special with room for cultural history." He spent his time reading Castiglione and Guicciardini – an anticipation of his lasting interest in the history of historical thought – and also recalls attending the inspirational lectures of the art historian Edgar Wind.

In 1960, having graduated with first-class honors, Burke moved to St Antony's College, then one of Oxford's newest, a postgraduate college with a mainly non-British student body. It was during this period that he began to be strongly influenced by the *Annales* school of historians, among them Lucien Febvre, Marc Bloch, and Fernand Braudel. He also began to take serious note of sociology, starting with attending a series of seminars on 'Alienation' – a topic much discussed in that period – organized by Norman Birnbaum and Iris Murdoch at Nuffield College. For his DPhil thesis, he chose to study the friar Paolo Sarpi's *History of the Council of Trent* (1619). The thesis, supervised by Hugh Trevor-Roper, was to have been entitled 'New trends in European historical writing, 1500–1700,' but it was never finished. Asa Briggs, who was just moving from the chair of history at the University of Leeds to become the founding Professor of History at the brand-new University of Sussex, gave a lecture in Oxford on sociology and history, and mentioned that Sussex was looking to recruit staff with interdisciplinary interests. Burke abandoned the DPhil and in the autumn of 1962 arrived in Sussex as an Assistant Lecturer in the School of European Studies.

During the next decade, Sussex's star burned brightly in the British intellectual firmament. Burke remembers that his first ever lecture was on Freud, as part of a course on 'The European Mind.' The sociologists Zev Barbu and Tom Bottomore; the art historian Hans Hess; the intellectual historians John Burrow, Stefan Collini, along with Donald Winch, Gabriel Josipovici, and Stephen Medcalf, were among Burke's friends and collaborators in a stellar interdisciplinary milieu that helped to shape his interests and career. By the late 1970s, however, the decline of university funding in Britain was beginning to take its toll at Sussex (as elsewhere), and Burke moved to Cambridge as a Fellow of Emmanuel College and as University Lecturer in History, eventually Professor of Cultural History. Since the 1970s, Burke has published at such pace that it is necessary to distinguish several interrelated strands: books reporting his

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61169-8

own substantive historical research, his historiographical writings, and his activity as editor and translator. Yet Burke did not publish his first learned article until he was 28, and his first 'substantial' book, *Culture and Society in Renaissance Italy* (Burke, 1972a), not until he was 35. As he points out, such dilatoriness would not now be possible in an age when young scholars are endlessly evaluated, with any failure to churn out a quota of publications always threatening to bring an end to their careers.

Nor had he, until his early thirties, worked in an archive. His earliest research had been based on printed sources. This too, for a historian, might now prove a serious obstacle to an academic career. The subsequent flood of books and essays from Burke's pen, many derived from archival research in several languages, shows how mistaken it may prove to be if artificial hurdles are erected for young scholars.

It was Tom Bottomore who invited Burke to write the first version of the series of books that are probably most widely familiar to nonhistorians. Appearing in 1980, *Sociology and History* – note the order of the words – can be seen in hindsight as marking the end of two decades in which sociology seemed an exciting new discipline that offered useful new ideas relevant to the work of several other university disciplines, notably history. Not that historians and sociologists always enjoyed the best of relations: as Burke wrote,

> In Britain at least, many historians still regard sociologists as people who state the obvious in a barbarous and abstract jargon, lack any sense of place and time, squeeze individuals into rigid categories and, to cap it all, believe that these activities are 'scientific'. Sociologists, for their part, see historians as amateurish myopic fact-collectors without a method, the vagueness of their data matched only by their incapacity to analyse them.
>
> 1980: 14

Burke stressed that he was facing two ways, addressing professions and writing about major contributors to each discipline for the benefit of the other. On the one hand, his book makes a powerful case for the relevance of the ideas of social and cultural theorists past and present to the practice and writing of history, ideas that met with resistance from an older generation of more traditional professional historians. On the other hand, less explicitly but no less convincingly, it demonstrates what the discipline of sociology has lost through what Norbert Elias (2009) called its 'retreat into the present' or what Goudsblom (1977) termed 'hodiecentrism.'

This masterly little book – it contained only just over 100 pages of printed text – opened with a chapter that sketched the major landmarks in the growth since the Enlightenment of a historically orientated sociology and a socially orientated history, in both Europe and the Americas. The longest of the three chapters was entitled 'Social Structures,' and introduced some key ideas: the comparative method; history, models, and types; survey analysis (including by historians); structure and function; social role; socialization, deviance, and control; social class and social stratification; social mobility; bureaucracy; patrons, clients, and factions; and mentality and ideology. The third chapter, on 'Social Change,' was constructed upon discussions of Herbert Spencer's and Karl Marx's models, and concluded with a consideration of 'four monographs in search of a theory.' These were Fernand Braudel's *The Mediterranean and the Mediterranean World in the Age of Philip II* (1972[1949]), William H. McNeill's *Europe's Steppe Frontier* (1964), Emmanuel Le Roy Ladurie's *The Peasants of Languedoc* (1974), and Nathan Wachtel's *Vision of the Vanquished* (1977), a study of the Spanish conquest of Peru that is perhaps now a little less widely known than the first three. All four monographs, it will be observed, are by historians, and, though he certainly faced two ways, this confirms that Burke was a little more concerned with explaining the value of sociological ideas to his fellow historians than with stressing the necessity of historical perspectives for sociology.

When Burke was invited, by a different publisher, to produce a revised and significantly enlarged second edition (and later a third) of the book, it appeared under the new title of *History and Social Theory* (1992). He had originally subsumed a good deal of social anthropology under the umbrella term 'sociology,' which reflected the fact that until the 1980s the sociological syllabus in British universities – and perhaps to a lesser extent in the United States – had always drawn a good deal upon anthropological studies. The original book had also said a little about economics and political science too. By the 1990s, however, there had been a veritable explosion of 'theory' across the board of the humanities and social sciences. As Burke remarked,

> a discussion of social theory might reasonably be expected to include much more, including such disciplines or sub-disciplines as communications, geography, international relations, law, linguistics (especially socio-linguistics), psychology (especially social psychology) and religious studies. It is also virtually impossible to exclude interdisciplinary enterprises such as critical, cultural and feminist theory, or indeed philosophy (which might be defined as the theory of theory).
>
> 1992: vii

Burke does not say so, but the changed title may be taken also to reflect the diminished scope of the discipline and the word 'sociology,' which has gradually lost the encompassing general meaning of 'the study of society,' and become one social science among many. More disturbingly, despite the renewed vigor of 'historical sociology' in recent decades, that has been a minority trend; the broader trend in the sociological profession has been toward what Burke would hope to have countered: an increasingly present-centered, ahistorical, and technicized discipline.

If the 1980 book reflected the preoccupations of sociological theory in the 1960s and 1970s, the two later versions incorporate many of the developments and the writers who rose to prominence in various disciplines in the subsequent decades. Much of the original text remained, but it was radically reorganized and supplemented. For instance, by 1992 the four monographs had become six, and now included Norbert Elias's *On the Process of Civilisation* (2012[1939]), Michel Foucault's *Discipline and Punish* (1977), and Marshall Sahlins's *Islands of History* (1985). Figures such as Clifford Geertz and Pierre Bourdieu received much greater attention. By 2005, the 'cultural turn' in social theory was recognized in fuller discussions of Bakhtin and Gombrich (T.S. Kuhn makes a surprisingly delayed appearance too), and of Pierre Bourdieu who by then had become the world's most cited sociologist. In face of

the avalanche of theoretically significant writing, the six monographs now disappear, but are replaced by a whole concluding chapter on 'Modernity and Postmodernism,' with discussion of globalization and the development of a global history that begins to surmount the Eurocentrism of the past.

Burke as a historian displays a breadth of knowledge and theoretical command that would be the envy of most specialist 'theorists' in sociology. What is possibly more impressive is the broad tolerance he exhibits toward so many varieties of 'theory.' Most specialists feel the need to favor some and reject other varieties. Some of what is called 'theory' is widely seen among social scientists as philosophoidal pretension with little relevance to the conduct of historical or social scientific research, and often redescribing old problems with new jargon. What seems to enable Burke to show such broad sympathies and tolerance is that over the course of all three editions of his book, he has filtered out and retained precisely the ideas that have relevance to the writing of history. And writing history has always been, to put it colloquially, Peter Burke's 'day job.'

Culture and Society in Renaissance Italy (Burke, 1972a) was an instance of the prosopographical method, focusing on 600 artists and humanists of the Italian Renaissance. The method, which involves using a great many only partially known biographies to establish a collective picture of the life of members of a group, had been championed a year earlier by Lawrence Stone in an influential article (1971), but the work of Sir Lewis Namier on eighteenth-century parliamentarians (1957[1929]), Francis Haskell's on patronage and taste (1963), as well as the writings of Arnold Hauser (1951–62, 1982) and Raymond Williams (1958) were prior influences. Prosopography as a historical method is also analogous in a limited way to survey research in sociology.

Prosopography continued in Burke's first thoroughly archival project, *Venice and Amsterdam* (1974), a comparison of the two cities' patricians in the long seventeenth century, 1580–1720. Both were rich mercantile cities dominated by merchant elites, rather than the warrior and subsequently courtly nobility found in the larger western European states. The Amsterdam *Regenten* were commoners, while the Venetian patricians were noblemen, although Burke notes that the Venetian elite's style of life resembled not that of *noblesse de l'épée* but that of the *noblesse de robe*, or even of the Chinese mandarins (1974: 62). Both lived in their cities rather than on country estates, although later in the eighteenth century both tended to be transformed into rentiers.

This book also led later to a small but interesting offshoot, a study of Antwerp (1993a) in its brief period of world significance, roughly the first six decades of the sixteenth century. Its dominance followed that of Venice, and was succeeded by that of Genoa and Amsterdam. After that, what Braudel termed 'world economies' were not cities but nation-states. The comparative sweep of this short book is striking: the chapter on 'centers of painting' discusses Florence, Venice, Rome, Antwerp, Amsterdam, and Paris, as well as the lesser centers of Nuremberg and London. Chapters follow on 'centers of performance,' dealing with public and private rituals as well as theaters, 'centers of printing,' and finally a study of three Asian cities, Istanbul, Peking, and in Japan the Kyoto/Osaka/Edo complex. Burke concludes that what makes a city a cultural metropolis is absence of censorship, more centralized states (especially the deracination or courtisation of the warrior nobility), and economic prosperity.

After *Venice and Amsterdam*, Burke turned his attention from elite to popular culture. This book displayed a tendency he later identified as a characteristic of his career, that the choice of a research topic for the next one "in some sense compensated for what was lacking in my earlier work" (2009a: 276). *Popular Culture in Early Modern Europe* (1978) was also influenced by Burke's 'discovering (belatedly) the work of Edward Thompson,' and a growing friendship with Raphael Samuel and involvement in the History Workshop movement. One of Burke's key theses was that until Renaissance, European elites were 'bicultural' or 'amphibious': they participated in popular culture as well as high culture. At the close of the Middle Ages, popular culture was everyone's culture, in the sense that the educated participated in it as a second culture and it was the only culture for everyone else. Peasant and priest, artisan and noble shared to a significant extent a common culture of intellectual and religious interests, moral and political assumptions, but over the next three centuries this common ground dwindled very markedly. Printing was a powerful force promoting the withdrawal of the upper classes from popular culture. There was a sharpening of boundaries between city dwellers and peasants, and between the educated urban elite and the illiterate urban laborer. This happened in part because the effect of wider literacy, amplified by printing, made it possible for the literate, learned culture to evolve far more rapidly than the traditional culture of the less educated classes. Printing also very much facilitated the process of social emulation of the habits, styles, and interests of one class or stratum by another, yet it should not be thought that printing led to rapid cultural homogenization. In early modern Europe it reduced certain social divisions but enhanced others. Today, many boundaries are being eroded, but, as Burke notes, "there is now a danger of underestimating the importance of cultural boundaries in the past" (1978: 9 (third ed.)).

Thus far, Burke had never made politics central to his writings – a prolonged reaction perhaps to its overprominence in the traditional history that prevailed in his student days in Oxford – and *The Fabrication of Louis XIV* (1992a) was in part intended to compensate for that. The topic of the construction of Louis XIV's image combined politics with culture. The first draft was written at the Wissenschaftskolleg in Berlin in 1989–90, a time and place when politics was inescapable, and the book is notable for its comparisons of official image making in the seventeenth and twentieth centuries. Loosely linked to *Fabrication* by the subject matter of courts and courtiers was Burke's book *The Fortunes of the Courtier* (1995), a study of the Europe-wide impact of Castiglione's *Il Cortegiano*.

One whole group of Burke's books is that concerned with the social history of language. *The Art of Conversation* (1993b) brought together essays by Burke on language early modern Europe: postmedieval Latin; language and identity in Italy; the art of conversation in early modern Europe; and the social history of silence. There followed *Languages and Communities in Early Modern Europe* (2004), in which he explores the social and cultural history of the languages spoken or written in Europe between the invention of printing and the French Revolution, arguing that, from a linguistic point of view, 1450 to 1789

should be regarded as a distinct period. One major theme of the book is the relation between languages and communities (regions, churches, occupations, and genders, as well as nations) and the place of language as a way of identifying others, as well as a symbol of one's own identity. A second, linked theme is that of competition: between Latin and the vernaculars, between different vernaculars, dominant and subordinate, and finally between different varieties of the same vernacular, such as standard languages and dialects. But Burke's own writings on language have to be considered along with numerous books that he has edited, notable among them three important volumes in collaboration with Roy Porter (1987, 1991, 1995). His introduction to the first of these begins, "It is high time for a social history of language, a social history of speech, a social history of communication. In the last generation or so, as the rise of feminist and regionalist movements shows, dominated groups have become more sharply aware of the power of language as well as the involvement of language with other forms of power." In the third volume, his essay on 'The Jargon of the Schools' demonstrates how far back goes the accusation that 'unintelligibility is at a premium' among intellectuals staking out their territory through the use of their own jargon. A fourth volume, *Cultural Translation in Early Modern Europe* (2007), was edited with R. Po-chia Hsia. Burke's own paper there is concerned with translations from the vernacular languages into Latin; as he remarks, that seems counterintuitive – translation in the 'wrong' direction – but it was not unusual.

Burke answered his own call for a 'social history of communication' in a two-volume work of his own and another coauthored with Asa Briggs. Together, the two volumes of *A Social History of Knowledge* (2000–12) are perhaps Burke's most ambitious project to date. The first, subtitled *From Gutenberg to Diderot*, is firmly in his own early modern territory, but in the second, *From the Encyclopédie to Wikipedia*, he has been 'eased out' of that familiar landscape. He notes that he began the work consciously in the tradition of Karl Mannheim and the sociology of knowledge. Topics covered in the book include the 'establishment' of knowledge in institutional arrangements such as academies, centers and peripheries in the location of knowledge (the idea of center and periphery being a recurrent theme in Burke's writings), the control of knowledge by churches and states, and the sale of knowledge through markets and the press. The final chapter of the first volume concerns 'trusting and distrusting knowledge,' and concludes with a quotation from Mannheim: "Basically, it was in political struggles that for the first time men became aware of the unconscious motivations which had always guided the direction of thought …. The discovery of the social-situational roots of thought first, therefore, took the form of unmasking" (Mannheim, 1936: 35).

In between those two volumes appeared *A Social History of the Media* (Briggs and Burke, 2002). Burke was primarily responsible for the first three chapters – the Introduction, The Print Revolution in Context, and The Media and the Public Sphere in Early Modern Europe – and Briggs for the later period, but the two authors read, revised, and took joint responsibility for all chapters.

One other book in which Peter Burke journeys beyond his early modern homeland is *Cultural Hybridity* (2009b), which takes its departure from a remark by Perry Anderson (1998) that we live in a postmodern age in which intersections, hybridity, potpourri are esteemed. Burke recognizes that today we live in era of 'culture wars' in which there is conflict over precisely such cultural hybrids.

Besides his many works of substantive history, Peter Burke has written very extensively on matters of historiography. His very first book, which he does not deign to mention in his essay in intellectual autobiography, was *The Renaissance Sense of the Past* (1969). An opening chapter on 'Medieval historical thought,' in which he shows that medieval historians, besides taking the Bible as an unquestionable source, lacked a sense of the past being different in quality from the present, resulting in what would later seem incongruous anachronisms in their writings. And their attitude toward evidence was uncritical, so that medieval history writing is full of myths, secular and religious. The body of the book then traces the emergence of a 'sense of the past' through among other things the antiquarian movement, developments in the law, and the critical study of documents, culminating in a look forward 'Towards a sociology of historiography.' *The Renaissance* (1987) too is essentially historiographical, charting modern historians' view that the European Renaissance of the fifteenth to seventeenth centuries was neither historically unique nor so abrupt a departure from the medieval period as Jacob Burckhardt (1860) depicted it. It was more a movement than a period. *Eyewitnessing* (2001) is concerned with the use of images as historical evidence. One striking point that it makes is related to the discussion of anachronism in the much earlier *Renaissance Sense of the Past*: that painters, unlike historians, could not dodge the question of whether to dress their subjects in a historically accurate way or in the style of their own day.

Throughout his career, Peter Burke has felt an affinity with the *Annales* school. He wrote a history of the school (1990), but, as editor and translator, has also contributed mightily to making the work of the French historical school better known in the Anglophone world. In *A New Kind of History* (1973), he made available several of the most important essays by Lucien Febvre who, despite being the founder of the school along with Marc Bloch, and in particular the champion of the idea of *mentalités* as one of its central objects of study, had remained until then somewhat less familiar in English than Bloch and the next generation of *Annalistes*. *Economy and Society in Early Modern Europe* (1972b) presented essays published in the *Annales Économies–Sociétés–Civilisations* in the period 1955–66, by inter alia Fernand Braudel, Carlo Cipolla, and Emmanuel Le Roy Ladurie. Burke's French connection is also seen in his edition of a collection of critical writings on Michel Foucault (1992b) who, though his origins lay in philosophy and though he was not exactly a member of the *Annales* school, had an immense impact upon the writing of history that would not have been possible without the prior influence of the school. A final editorial contribution to historiography is *New Perspectives on Historical Writing* (1991), which contains a catholic selection of contributions on history from below, women's history, overseas history, microhistory, oral history, history of reading, visual history, history of political thought, history of the body, and environmental history. Burke's own final chapter is on 'History of Events and the Revival of Narrative' (1991: 283–300). The recognition that, in the end,

events still matter is further evidence of the breadth of Burke's historical vision. Too often 'history from below' and the 'history of events' – especially the kind of narrative of national political events that once monopolized the school and undergraduate syllabuses – have been set up as mutually exclusive opposites. But events – the reigns of kings and queens, wars, battles, and political conflicts – have lasting social and cultural consequences that cannot be ignored by the social and cultural historian.

In fitting conclusion, it should be mentioned that Burke, together with his wife Maria Lucia Pallares-Burke, has written a major study (Burke and Pallares-Burke, 2008) of the Brazilian intellectual Gilberto Freyre (1900–87). It is fitting because Freyre was also a multifaceted thinker who is difficult to pigeonhole. Wikipedia has described him as "sociologist, anthropologist, historian, writer, painter, journalist and congressman." Peter Burke has never served in parliament, does not write much for newspapers, nor so far as we know does he paint; but in their range his writings qualify him as 'sociologist, anthropologist, historian, writer,' and he has spent a good deal of time studying painters too.

See also: Annales School; Cultural History; Freyre, Gilberto (1900–87); Knowledge, Sociology of; Media and History; Prosopography (Collective Biography); Public Sphere and the Media; Renaissance; Social History; Sociological Theory.

Bibliography

Anderson, Perry, 1998. Origins of Postmodernity. Verso, London.
Braudel, Fernand, 1972[1949]. The Mediterranean and the Mediterranean World in the Age of Philip II. Collins, London.
Briggs, Asa, Burke, Peter, 2002. A Social History of the Media: From Gutenberg to the Internet. Polity, Cambridge.
Burckhardt, Jakob, 1944[1860]. The Civilisation of the Renaissance in Italy. Phaidon Press, London.
Burke, Peter, 1969. The Renaissance Sense of the Past. Edward Arnold, London.
Burke, Peter, 1972a. Culture and Society in Renaissance Italy. Batsford, London. Paperback edition, Tradition and Innovation in Renaissaince Italy, Fontana, London, 1974.
Burke, Peter (Ed.), 1972b. Economy and Society in Early Modern Europe: Essays from Annales. Routledge & Kegan Paul, London.
Burke, Peter (Ed.), 1973. A New Kind of History: From the Writings of Febvre. Routledge & Kegan Paul, London.
Burke, Peter, 1974. Venice and Amsterdam: A Study of Seventeenth-Century Elites. Temple Smith, London.
Burke, Peter, 1978. Popular Culture in Early Modern Europe. Temple Smith, London.
Burke, Peter, 1980, 1992, 2005. Sociology and History. Allen & Unwin, London. Rev. ed., History and Social Theory (Polity Press, Cambridge, 1992); third ed., same publisher 2005.
Burke, Peter, 1987. The Renaissance, second ed. 1997. Macmillan, London.
Burke, Peter, 1990. The French Historical Revolution: The Annales School, 1929–89. Polity, Cambridge.
Burke, Peter (Ed.), 1991. New Perspectives on Historical Writing, second ed. 2001. Polity, Cambridge.
Burke, Peter, 1992a. The Fabrication of Louis XIV. Yale University Press, New Haven, CT.
Burke, Peter (Ed.), 1992b. Critical Essays on Michel Foucault. Scolar Press, Aldershot.
Burke, Peter, 1993a. Antwerp: A Metropolis in Comparative Perspective. Martial and Snoeck, Antwerp.
Burke, Peter, 1993b. The Art of Conversation. Polity, Cambridge.
Burke, Peter, 1995. The Fortunes of the Courtier: The European Reception of Castiglione's Cortegiano. Polity, Cambridge.
Burke, Peter, 2000–12. A Social History of Knowledge: [I] From Gutenberg to Diderot (Polity Press, Cambridge, 2000); [II] From the Encyclopédie to Wikipedia (Polity Press, Cambridge, 2012).
Burke, Peter, 2001. Eyewitnessing: The Use of Images as Historical Evidence. Reaktion Books, London.
Burke, Peter, 2004. Languages and Communities in Early Modern Europe. Cambridge University Press, Cambridge.
Burke, Peter, 2009a. Invitation to historians: an intellectual self-portrait, or the history of a historian. Rethinking History: The Journal of Theory and Practice 13 (2), 269–281.
Burke, Peter, 2009b. Cultural Hybridity. Polity Press, Cambridge.
Burke, Peter, Pallares-Burke, Maria Lucia G., 2008. Gilberto Freyre: Social Theory in the Tropics. Peter Lang, Witney.
Burke, Peter, Po-chia Hsia, R. (Eds.), 2007. Cultural Translation in Early Modern Europe. Cambridge University Press, Cambridge.
Burke, Peter, Porter, Roy (Eds.), 1987. The Social History of Language. Cambridge University Press, Cambridge.
Burke, Peter, Porter, Roy (Eds.), 1991. Language, Self and Society. Polity, Cambridge.
Burke, Peter, Porter, Roy (Eds.), 1995. Languages and Jargons. Polity, Cambridge.
Elias, Norbert, 2009. The retreat of sociologists into the present. In: Essays III: On Sociology and the Humanities. Collected Works, vol. 16. UCD Press, Dublin, pp. 107–126.
Elias, Norbert, 2012[1939]. On the Process of Civilisation. In: Collected Works, vol. 3. UCD Press, Dublin [The orginal German title of the book was Über den Prozess der Zivilisation, 2 vols. (Basel: Haus zum Falken, 1939); earlier English editions appeared under the title The Civilizing Process.].
Foucault, Michel, 1977. Discipline and Punish: The Birth of the Prison. Pantheon, New York.
Goudsblom, Johan, 1977. Sociology in the Balance. Basil Blackwell, Oxford.
Haskell, Francis, 1963. Patrons and Painters: A Study in the Relations between Italian Art and Society in the Age of Baroque. Chatto & Windus, London.
Hauser, Arnold, 1951–62. In: The Social History of Art, 4 vols. Routledge & Kegan Paul, London.
Hauser, Arnold, 1982. The Sociology of Art. Routledge & Kegan Paul, London.
Ladurie, Emmanuel Le Roy, 1974. The Peasants of Languedoc. University of Illinois Press, Urbana, IL.
McNeill, William H., 1964. Europe's Steppe Frontier. University of Chicago Press, Chicago.
Mannheim, Karl, 1936. Ideology and Utopia. Routledge & Kegan Paul, London.
Namier, Lewis, 1957[1929]. The Structure of Politics at the Accession of George III, second ed. Macmillan, London.
Sahlins, Marshall, 1985. Islands of History. University of Chicago Press, Chicago.
Sarpi, Paolo, 1619. Istoria del Concilio Tridentino. Giovanni Billio, London.
Stone, Lawrence, 1971. Prosopography. Daedalus 100 (1), 46–71.
Wachtel, Nathan, 1977. Vision of the Vanquished: The Spanish Conquest of Peru through Indian Eyes, 1530–1570. Harvester, Hassocks.
Williams, Raymond, 1958. Culture and Society 1780–1950. Chatto & Windus, London.

Burnout, Psychology of

Christina Maslach, University of California, Berkeley, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C. Maslach, volume 2, pp. 1415–1418, © 2001, Elsevier Ltd.

Abstract

Burnout is a psychological syndrome that develops in response to chronic emotional and interpersonal job stressors. The three defining components of this syndrome are (1) exhaustion, (2) cynicism and detachment from the job, and (3) a sense of inefficacy and reduced accomplishment. Burnout impairs both personal and social functioning. While some people may quit the job as a result of burnout, others will stay on, but will only do the bare minimum rather than their very best. This decline in the quality of work and in both physical and psychological health can be costly – not just for the individual worker, but for everyone affected by that person. Consequently, there has been a growing interest in developing interventions to reduce burnout, from government agencies and organizations in both the public and private sectors.

A Brief History of Burnout Research

Although the birth of the burnout concept, as we know it today, began in the 1970s, that does not mean that the problem itself suddenly burst on to the scene at that time. The relationship that people have with their work, and the difficulties that can arise when that relationship goes awry, have been long recognized as a significant phenomenon in people's lives. Earlier writing, both fictional and nonfictional, described similar phenomena, including extreme fatigue and the loss of idealism and passion for one's job. Perhaps the clearest example is Greene's 1961 novel, *A Burnt-Out Case*, in which a spiritually tormented and disillusioned architect quits his job and withdraws into the African jungle. It is probably fair to say that burnout had been identified as a problem by both social commentators and practitioners long before it had a recognized name and had become a focus of systematic study by researchers.

And it was the identification of that name, 'burnout,' which has such evocative power to capture the psychological reality of people's experiences in the workplace, which occurred in the 1970s. The origins of its usage are somewhat murky, but appear to come from the illicit drug scene, where 'burnout' referred to the physical effects of chronic drug abuse. Counselors and therapists who worked with drug addicts borrowed the term to describe their own psychological deterioration and stress. This use of 'burnout' was reported by a clinical psychologist, Herbert Freudenberger (1974), in a volume of articles about free clinics. At that same time, a social psychologist, Christina Maslach, was conducting exploratory research with workers in health care and human service occupations, studying how they coped with strong emotional arousal on the job. During the course of the interviews, some of the workers described their psychological difficulties as 'burnout,' and Maslach shifted her focus to describe this phenomenon (1976). For both Freudenberger and Maslach, burnout was rooted within caregiving and service occupations, in which the core of the job is the relationship between provider and recipient. This interpersonal context of the job meant that, from the beginning, burnout was studied not simply as an individual stress response, but in terms of an individual's relational transactions in the workplace. Moreover, this interpersonal context focused attention on the individual's emotions, and on the motives and values underlying his or her work with other people.

The clinical and social psychological perspectives of the initial articles influenced the nature of the first phase of burnout research. On the clinical side, the focus was on symptoms of burnout and on issues of mental health. On the social side, the focus was on the relationship between provider and recipient, and on the situational context of service occupations. Most of this initial research was descriptive and nonempirical – it was more qualitative in nature, using such techniques as interviews, case studies, and on-site observations. In addition, this first phase was characterized by a strong applied orientation, which reflected the particular set of social, economic, historical, and cultural factors of the 1970s. Both Farber (1983) and Cherniss (1980) provided analyses of how these factors had influenced the professionalization of the human services in the United States, and had made it more difficult for people to find fulfillment and satisfaction in these careers. Burnout workshops became a primary mode of intervention, as exemplified in the work of Pines et al. (1981).

A second phase of burnout research developed in the 1980s, in which the emphasis shifted to more systematic empirical research. This work was more quantitative in nature, using questionnaire and survey methodology, and studying larger subject populations. A particular focus of this research was the assessment of burnout, and several different measures were developed. The scale that has had the strongest psychometric qualities and has been used in over 90% of the research studies is the Maslach Burnout Inventory (MBI) developed by Maslach and Jackson in 1981.

The shift to greater empiricism was accompanied by theoretical and methodological contributions from the field of industrial–organizational psychology. Burnout was viewed as a form of job stress, with links to such concepts as job satisfaction, organizational commitment, and turnover. The industrial–organizational approach, when combined with the prior work based in clinical and social psychology, generated a richer diversity of perspectives on burnout and strengthened the scholarly base via the use of standardized tools and research designs.

Another aspect of this empirical phase is that burnout extended beyond its original American borders. At first, the phenomenon drew attention in English-speaking countries, such as Canada and the United Kingdom. Soon articles, books, and research measures were translated into numerous languages, and subsequently research on burnout emerged in many countries of Europe, as well as Israel. By the 1990s, the intellectual contributions of European authors in terms of theory, research, and intervention were considerable (most notable is the work of Schaufeli and his colleagues; see Schaufeli et al., 2001; Schaufeli and Enzmann, 1998). Now, in the twenty-first century, burnout research is even more global, and is being done in many countries around the world. As a result of all these new contributions, the concept of burnout has been extended to occupations beyond the human services and education, and burnout research has been enhanced by more sophisticated methodology and statistical tools, as well as new structural models.

Defining and Understanding Burnout

Burnout research was exploratory at first, very bottom-up and anchored in people's experiences. As studies began to show a clearer pattern of the burnout phenomenon, it became possible to establish it as a prolonged response to chronic interpersonal stressors on the job. The three key dimensions of this response are an overwhelming exhaustion; feelings of cynicism and detachment from the job; and a sense of ineffectiveness and lack of accomplishment. The exhaustion component represents the basic individual stress dimension of burnout. It refers to feelings of being overextended and depleted of one's emotional and physical resources. The cynicism (or depersonalization) component represents a motivational, interpersonal distancing dimension of burnout. Depersonalization refers to a negative, callous, or excessively detached response to various aspects of the job; cynicism refers to losing an emotional or cognitive involvement with work. The component of reduced efficacy or accomplishment represents the self-evaluation dimension of burnout. It refers to feelings of incompetence and a lack of achievement and productivity in work.

These three dimensions of burnout are assessed by the MBI (see Maslach et al., 1996, for all versions of the measure, which has been translated and validated in many languages). Recently, an alternative burnout questionnaire has been proposed – the Oldenburg Burnout Inventory – that assesses the two dimensions of exhaustion and disengagement from work (Halbesleben and Demerouti, 2005), and has shown high convergent validity with the MBI. Other conceptualizations of burnout have considered exhaustion to be the sole defining criterion (Freudenberger, 1983; Pines et al., 1981). Although more recent theories make distinctions between various aspects of exhaustion – e.g., physical fatigue, emotional exhaustion, and cognitive weariness (Shirom and Melamed, 2005) or physical and psychological exhaustion (Kristensen et al., 2005) – their measures inevitably produce a single overriding factor of exhaustion. However, the fact that exhaustion is a necessary criterion for burnout does not mean it is sufficient. Exhaustion is not something that is simply experienced – rather, it prompts actions to distance oneself emotionally and cognitively from one's work, presumably as a way to cope with the work overload. Distancing is such an immediate reaction to exhaustion that a strong relationship from exhaustion to cynicism is found consistently in burnout research, across a wide range of organizational and occupational settings.

Three decades of burnout research conducted at the individual level of analysis has resulted in a deep understanding of how individual employees react to their jobs. The exhaustion component of burnout is more predictive of stress-related health outcomes than the other two components. These physiological correlates mirror those found with other indices of prolonged stress. In terms of mental health, the link with burnout is more complex. It has been assumed that burnout may result in subsequent mental disabilities, and there is some evidence to link burnout with greater anxiety, irritability, and depression. However, an alternative argument is that burnout is itself a form of mental illness, rather than a cause of it. Some of this research has focused on the distinction between burnout and depression: burnout is job-related and situation-specific, as opposed to depression, which is general and context free. Other research has established clinical burnout as the equivalent of work-related neurasthenia (Schaufeli et al., 2001).

Some of the burnout research has considered a range of personal qualities. Burnout scores tend to be higher for people who have a less 'hardy' personality or a more external locus of control, or who score as 'neurotic' on the Five-Factor Model of personality. People who exhibit Type A behavior tend to be more prone to the exhaustion dimension of burnout. There are few consistent relationships of burnout with demographic characteristics. Although higher age seems to be associated with lower burnout, it is confounded with both years of experience and with survival bias. The only consistent gender difference is a tendency for men to score slightly higher on cynicism.

Burnout has been associated with various forms of job withdrawal – absenteeism, intention to leave the job, and actual turnover. However, for people who stay on the job, burnout leads to lower productivity and effectiveness at work. Consequently, it is associated with decreased job satisfaction and a reduced commitment to the job or the organization. People who are experiencing burnout can have a negative impact on their colleagues, both by causing greater personal conflict and by disrupting job tasks. Thus, burnout can be 'contagious' and perpetuate itself through informal interactions on the job. There is also some evidence that burnout has a negative 'spillover' effect on people's home life.

Much of the burnout research has found that social and organizational conditions are the primary correlates of burnout across many occupations (see reviews by Maslach et al., 2001; Schaufeli and Enzmann, 1998). These factors can be summarized within six key domains of the workplace environment: workload, control, reward, community, fairness, and values. The first two areas are reflected in the Demand–Control model of job stress (Karasek and Theorell, 1990). Work overload occurs when job demands exceed human limits. People have to do too much in too little time with too few resources. Increased workload has a consistent relationship with burnout, especially with the exhaustion dimension. Control includes employees' perceived capacity to influence decisions that affect their work,

to exercise professional autonomy, and to gain access to the resources necessary to do an effective job. A major control problem occurs when people experience role conflict. Reward refers to the power of reinforcements (monetary, social, and intrinsic) to shape behavior. Insufficient reward and lack of recognition increases people's vulnerability to burnout by devaluing their work and increasing their feelings of inefficacy. Community is the overall quality of social interaction at work, including issues of conflict, mutual support, closeness, and the capacity to work as a team. When these workplace relationships lack support and trust, or are characterized by incivility and unresolved conflict, then the risk of burnout is higher. Fairness emerges from the literature on equity and social justice, and refers to whether people feel they are being treated fairly and with respect. Burnout is more likely when people perceive the workplace to be characterized by inequity, lack of reciprocity, and an effort–reward imbalance. Finally, values are the ideals and motivations that originally attracted people to their job, and thus they are the motivating connection between the worker and the workplace, which goes beyond the utilitarian exchange of time for money or advancement. When there is a value conflict on the job, and thus a gap between individual and organizational values, workers will find themselves making a trade-off between work they want to do and work they have to do. Research has found that a conflict in values is related to all three dimensions of burnout.

A consistent theme throughout this research literature is the problematic relationship between the individual and the situation, which is often described in terms of imbalance or misalignment or misfit – for example, the demands of the job exceed the capacity of the individual to cope effectively, or the person's efforts are not reciprocated with equitable rewards. Some of the earliest models of organizational stress focused on this notion of job–person fit, which was assumed to predict less strain and better adjustment, and subsequent theorizing continued to highlight the importance of both individual and contextual factors (see Kahn and Byosiere, 1992). A mediation model of burnout focuses on the degree of perceived match, or congruency, between the individual and the six key areas of the workplace environment listed above. When there is greater mismatch between the person and the job, there is a greater likelihood of burnout, which in turn predicts work-related outcomes (Leiter and Maslach, 2005). In other words, burnout appears to be the mediating state between work stressors and work outcomes (see **Figure 1**). It is not simply that burnout is an important psychological outcome in its own right, but that it is related to people's commitment to their work and their motivation to perform well on the job.

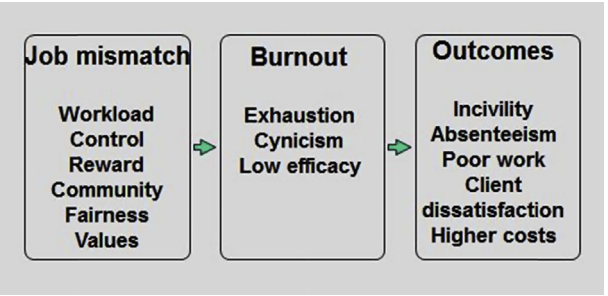

Figure 1 The mediation role of burnout.

Research on Burnout Interventions

This analysis of the burnout phenomenon points to an important bottom line in terms of the fit between people and their job. Employees need to have some choice and control over how they handle their workload, some sense of recognition and reward, support from the people they work with, and a sense of being treated with respect and fairness. There must be effective ways to resolve conflicts, and there should be clear organizational values that can mesh with employees' personal motives. All workers want to know that they are doing meaningful work that makes them proud. Burnout can be prevented and reduced when employers design workplaces that engage employees, and employees have the opportunity to make job choices that match their personal attributes. These guidelines may seem fairly simple and straightforward, but putting them into practice can be a complex process that requires continuous time and effort. However, the outcomes are well worth the commitment – healthier workers and more productive organizations.

Unfortunately, the work that has been done to test and evaluate the efficacy of burnout interventions has been fairly sparse over the years. Despite the public interest in finding solutions to the burnout problem, there are many obstacles and difficulties in actually doing the kind of work that would provide some clear answers to the question of "what can we do to deal with burnout?" Most burnout interventions have focused on changing or treating individual people, and have not dealt with the situational factors that have been identified in the research. Thus, there have been debates about whether to help people cope with stress, or to put them on disability leave and treat them for severe depression. Such debates reopen the issue about how to define burnout, almost as if there has not been over 30 years' of research on this topic. The clearest metamessage that comes from reviewing the literature on burnout interventions is that the significance of the work environment has been greatly underestimated, relative to the significance of the individual. Intervention efforts must underscore the necessity of understanding individual behavior in its social context. If future work does not address directly this issue of person–situation interaction, then true progress in the prevention of burnout is likely to be negligible.

A promising development in burnout research has been to expand the framework to the positive opposite of burnout, rather than just focus on the negative state that it represents. This positive experience has been called work engagement, and many studies are being done to further understand the personal and situational factors that promote greater engagement in the workplace (Bakker and Leiter, 2010). Such a framework leads people to consider what factors are likely to enhance employees' energy, vigor, and resilience; to promote their involvement and absorption with the work tasks; and to ensure their dedication and sense of efficacy and success on the job. The question for interventions, then, has been changed to "how do we build/enhance engagement with work" rather than "how do we reduce the occurrence of burnout." Focusing on job

engagement represents a new way of thinking about solutions to burnout.

In addition, burnout researchers have been developing new tools and paradigms that can be adapted to applied research on interventions, and have been developing better partnerships with practitioners and organizations in the search for effective solutions (Maslach et al., 2012). Moreover, a new international e-journal, *Burnout Research*, will provide a much-needed, dedicated venue for the publication and dissemination of cutting-edge research, critical reviews, and new translational research on promising interventions. Given the ongoing importance of the burnout phenomenon, and the rising interest in making real progress in alleviating it, there is a need for a primary place to capture the many contributions being made.

See also: Academic Engagement; Affect-Regulation Motivation; Coping with Job Stress; Flow in Motivational Psychology; Occupational Health; Passion and Motivation; Personality and Health; School Burnout and Engagement: Lessons from a Longitudinal Study in Finland; Stress at Work; Teacher Motivation; Workplace Safety and Health.

Bibliography

Bakker, A.B., Leiter, M.P. (Eds.), 2010. Work Engagement: A Handbook of Essential Theory and Research. Psychology Press, Hove, UK.

Cherniss, C., 1980. Professional Burnout in Human Service Organizations. Praeger, New York.

Farber, B.A. (Ed.), 1983. Stress and Burnout in the Human Service Professions. Pergamon, New York.

Freudenberger, H.J., 1974. Staff burn-out. Journal of Social Issues 30, 159–165.

Freudenberger, H.J., 1983. Burnout: contemporary issues, trends and concerns. In: Farber, B.A. (Ed.), Stress and Burnout in the Human Service Professions. Pergamon, New York, pp. 23–28.

Greene, G., 1961. A Burnt-out Case. Viking Press, New York.

Halbesleben, J.B.R., Demerouti, E., 2005. The construct validity of an alternative measure of burnout: Investigation of the English translation of the Oldenburg Burnout Inventory. Work & Stress 19, 208–220.

Kahn, R.L., Byosiere, P., 1992. Stress in organizations. In: Dunnette, M.D., Hough, L.M. (Eds.), Handbook of Industrial and Organizational Psychology, vol. 3. Consulting Psychologists Press, Palo Alto, CA, pp. 571–650.

Karasek, R., Theorell, T., 1990. Stress, Productivity, and the Reconstruction of Working Life. Basic Books, New York.

Kristensen, T.S., Borritz, M., Villadsen, E., Christensen, K.B., 2005. The Copenhagen Burnout Inventory: a new tool for the assessment of burnout. Work & Stress 19, 192–207.

Leiter, M.P., Maslach, C., 2005. A mediation model of job burnout. In: Antoniou, A.S.G., Cooper, C.L. (Eds.), Research Companion to Organizational Health Psychology. Edward Elgar, Cheltenham, UK, pp. 544–564.

Maslach, C., 1976. September. Burned-out. Human Behavior 9 (5), 16–22.

Maslach, C., Jackson, S.E., 1981. The measurement of experienced burnout. Journal of Occupational Behavior 2, 99–113 [All versions of the MBI are now published online by Mind Garden, mindgarden.com].

Maslach, C., Jackson, S.E., Leiter, M.P., 1996. Maslach Burnout Inventory Manual, third ed. Consulting Psychologists Press, Palo Alto, California [All versions of the MBI are now published online by Mind Garden, mindgarden.com].

Maslach, C., Leiter, M.P., Jackson, S.E., 2012. Making a significant difference with burnout interventions: researcher and practitioner collaboration. Journal of Organizational Behavior 33, 296–300.

Maslach, C., Schaufeli, W.B., Leiter, M.P., 2001. Job burnout. In: Fiske, S.T., Schacter, D.L., Zahn-Waxler, C. (Eds.), Annual Review of Psychology, 52, pp. 397–422.

Pines, A., Aronson, E., Kafry, D., 1981. Burnout: From Tedium to Personal Growth. Free Press, New York.

Schaufeli, W.B., Bakker, A.B., Hoogduin, K., Schaap, C., Kladler, A., 2001. The clinical validity of the Maslach Burnout Inventory and the Burnout Measure. Psychology and Health 16, 565–582.

Schaufeli, W.B., Enzmann, D., 1998. The Burnout Companion to Study and Practice: A Critical Analysis. Taylor & Francis, London.

Shirom, A., Melamed, S., 2005. Does burnout affect physical health? A review of the evidence. In: Antoniou, A.S.G., Cooper, C.L. (Eds.), Research Companion to Organizational Health Psychology. Edward Elgar, Cheltenham, UK, pp. 599–622.

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 3

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser—Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 3

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 3 *xliii*

VOLUME 3

Business and Society: Social Accounting 1
Edmund I Marshall

Business Anthropology 4
Marietta L Baba

Business Education 10
Bob Garvey

Business Ethics 15
Christoph Luetge

Business History 21
Hartmut Berghoff

Business Law 27
Constance E Bagley

Business Models 33
Raphael Amit and Christoph Zott

Business Platforms 37
Annabelle Gawer and Michael A Cusumano

Business Process: Outsourcing 43
Rafiq Dossani and Samina Dossani

Busyness 49
Charles N Darrah, Nicole M Conand, and Alicia Dornadic

Butler, Judith (1956–) 56
Birgit Schippers

Cambridge School of Intellectual History 61
Richard Whatmore

Campaigning: Political 66
L Sandy Maisel

Campbell, Donald Thomas (1916–96) 69
Andrew C Gordon

Cancer: Psychosocial Aspects 73
Tanja Zimmermann

Cancer Screening 78
Arja R Aro

Cannibalism 83
Shirley Lindenbaum

Capital: History of the Concept 90
Heinz D Kurz

Capitalism 94
Ivan T Berend

Capitalism: Global 99
Leslie Sklair

Capitalism: The History of the Concept 105
Jürgen Kocka

Cardiac Disease, Coping with 111
Heike Spaderna and Susan Hellwig

Cardiovascular Reactivity and Recovery 117
Joseph P Gyekis and William Gerin

Cardoso, Henrique (1931–) 122
Diego Pereyra

Career Development: Multilevel Perspective 128
Yehuda Baruch

Caregiver Burden 135
Christine Fekete

Cargo Cults 139
Lamont Lindstrom

Carnival 145
Philip W Scher

Caro Baroja, Julio (1914–95) 150
Andreas Hess and Mayte Martín

Carrying Capacity of the Environment 155
Cang Hui

Cartographic Visualization 161
Menno-Jan Kraak

Cartography 167
Judy M Olson

Case Method and Case Study Collections 176
Kirsten Lundberg

Case, Morphological: Linguistics 183
Barry J Blake

Case-Oriented Research 187
Charles C Ragin

Case Studies 194
Robert K Yin

Case Study: Logic 202
Anthony M Orum

Case Study: Methods and Analysis 208
Andrew Bennett

Cash Transfers 214
Nicola Jones and Elizabeth Presler-Marshall

Caste 223
Veena Das

Castells, Manuel (1942–) 228
Chaime Marcuello-Servós and Brenda Reed

Catholicism 233
Adrian Hastings

Catholicism and Gender 237
Rosemary Radford Ruether

Cattell, Raymond Bernard (1905–98) 243
John R Nesselroade

Causal Analysis, History of 247
Tim F Liao and Inna F Deviatko

Causal Counterfactuals in Social Science Research 251
Paul W Holland

Causal Inference: Overview 255
Jennifer Hill and Elizabeth A Stuart

Causation: Physical, Mental, and Social 261
Daniel Stoljar

Causes and Laws: Philosophical Aspects 266
Helen Beebee

Celebrity 274
Joshua Gamson

Censorship in Research and Scholarship 279
Nicholas G Evans and Michael J Selgelid

Census Microdata 284
Steven Ruggles

Censuses, Population: Comparative International Aspects 290
Tahu Kukutai and Victor R Thompson

Censuses: Current Approaches and Methods 296
Paolo Valente

Censuses: History and Methods 302
Margo Anderson

Center–Periphery Relationships 308
Ulf Hannerz

Central Africa: Sociocultural Aspects 312
René Devisch

Central Asia, Anthropology of 316
Catherine Alexander

Ceramics in Archaeology 323
Patrick V Kirch

Cerebellum: Cognitive Functions 327
Marco Molinari and Maria Leggio

Cerebral Cortex 332
Phillip Grant

Change Blindness 344
Robert W Kentridge

Charisma and Charismatic 350
Bryan S Turner

Charisma, Social Aspects of 353
Christopher Adair-Toteff

Charity and Philanthropy, Economics of 358
James Andreoni

Charity and Philanthropy: Overview 364
Dennis W Cheek, Michal Kramarek, and Patrick Rooney

Chemical Science: History and Sociology 371
David M Knight

Chiefdoms, Archaeology of 376
Laura L Junker

Child and Adolescent Psychiatry, Principles of 383
Nisha Dogra

Child Care and Child Development 391
AC Huston

Child Care and Development across Cultures 395
Birgit Leyendecker

Child-Directed Speech: Influence on Language Development 399
Rupa G Gordon and Linda R Watson

Child Labor 405
Jackline Wahba

Child Neglect 410
Beth Casey

Child Protection 417
Pam G Lister

Child Protection: International Issues 423
Simon Hackett

Child Soldiers 430
Brandon A Kohrt, Sauharda Rai, and Sujen M Maharjan

Child Soldiers as Victims and Actors: Lessons from Ethiopia 439
Tassé Abye

Childhood and Adolescence: Developmental Assets 445
Richard M Lerner, Lisette M DeSouza, Mary H Buckingham, Daniel JA Warren, Robey B Champine, and Kathleen N Greenman

Childhood: Anthropological Aspects 451
Thomas S Weisner

Childhood Cancer: Psychosocial Aspects 459
Meinolf Noeker and Franz Petermann

Childhood Memories 465
Harlene Hayne, Damian Scarf, and Kana Imuta

Children and Families in Social Work 471
Nick Frost

Children and Law 477
Laura A Rosenbury

Children, Rights of: Cultural Concerns 485
Michael S Wald

Children's Geographies 491
Claire Freeman and Paul Tranter

Children, Value of 498
Elizabeth Thomson

China as a Rising Aid Donor 502
Shino Watanabe

Chinese Literacy Acquisition: A Multidimensional Puzzle 508
Xiuli Tong, Silvia Siu-Yin Lam, and Catherine McBride-Chang

Chinese Revolutions: Twentieth Century 515
Jeffrey N Wasserstrom

Choice, Empowerment, and Social Work 520
Graeme Simpson

Chomsky, Noam (1928–) 526
Robert Barsky

Christianity in Asia 531
Michael Northcott

Christianity in Latin America 534
Gustavo Benavides

Christianity in the English-Speaking West 540
Wade Clark Roof and Steven Barrie-Anthony

Christianity, Origins of 545
Michael L Humphries

Chronic Illness, Psychosocial Coping with 549
Linda D Cameron and Christopher M Wally

Chronic Pain: Models and Treatment Approaches 555
Dennis C Turk

Chronology, Stratigraphy, and Dating Methods in Archaeology 562
Steven Shackley

Chronotype 568
Ana Adan

Church and State 574
Gerhard Robbers

Cities and the Financial Crisis 579
Manuel B Aalbers

Cities: Capital, Global, and World 585
Saskia Sassen

Cities: Internal Structure 593
Elisabeth Lichtenberger

Cities Under Postsocialism 605
Luděk Sýkora

Citizen Journalism 612
James F Hamilton

Citizen Mapping 619
Wen Lin

Citizen Participation 625
Howell S Baum

Citizen Science 631
Hauke Riesch

Citizenship, Anthropology of 637
Ayse Caglar

Citizenship, Historical Development of 643
Richard Bellamy

Citizenship: Political 650
Mark Hoipkemier

Citizenship, Sociological Aspects of 655
Thomas Janoski and Sara Compion

City Marketing 662
John Rennie Short

Civil Law 669
Peter Hay

Civil Liberties and Human Rights 674
Donald A Downs

Civil Religion 680
Todd Nicholas Fuist and Rhys H Williams

Civil Rights 686
David Kairys

Civil Rights Movements 691
James M Vardaman

Civil Service 695
Itzhak Galnoor and Jennifer Oser

Civil Society and the Public Sphere: History of the Concept 701
Craig Calhoun

Civil Society, Concept and History of 707
H Islamoglu

Civil Wars 713
Bill Kissane

Civilization, Concept and History of 719
A Al-Azmeh

Civilizations 725
Samuel N Eisenstadt

Clan 730
Philip Burnham

Class Actions, Legal 733
Deborah R Hensler

Class Consciousness 739
Jesus de Felipe-Redondo

Class: History of the Concept 745
Jan Pakulski

Class Size 751
Diane W Schanzenbach

Class: Social 755
David B Grusky

Classical Archaeology 760
Alain Schnapp

Classical Conditioning Methods in Psychotherapy 764
William C Follette and Georgia Dalto

Classical Legacies 771
Norman Vance

Classical Mechanics and Motor Control 778
Pietro G Morasso

Classical (Psychometric) Test Theory 785
Rolf Steyer

Classical Tests for Speech and Language Disorders 792
Yves Turgeon, Joël Macoir, and Anne Lafay

Classifications in Popular Music 799
Alex van Venrooij and Vaughn Schmutz

Classification, Understandings of in the Social Sciences 805
Hubert Feger

Classifiers, Linguistics of 811
Colette Grinevald

Classroom Assessment 819
James H McMillan

Classroom Climate 825
Barry J Fraser

Classroom Level in Educational Effectiveness Research 833
Christopher A Hafen and Erik A Ruzek

Climate Change 839
Ian Bailey and Piers Revell

Climate Change and Global Poverty 848
Abigail Jones and Joshua Hermias

Climate Change and the Media 853
Mike S Schäfer

Climate Change: Economics 860
Gary W Yohe

Climate, History of 865
Christian Pfister

Clinical Psychology in Europe, History of 873
I Lunt

Clinical Psychology in North America, History of 877
Donald K Routh

Clinical Psychology, Psychiatry and Homosexuality 884
Frank Muscarella

Closed and Open Systems: Organizational 893
Thomas G Cummings

Cluster Randomized Trials, Ethical Implications of 897
Monica Taljaard, Charles Weijer, and Jeremy M Grimshaw

Clustering and Cohesion in Networks: Concepts and Measures 906
James Moody and Jonathan Coleman

Co-constructivism in Educational Theory and Practice 913
Kurt Reusser and Christine Pauli

Code Switching: Linguistic 918
Shana Poplack

Coeducation and Single-Sex Schooling 926
Sum Kwing Cheung

CONTRIBUTORS TO VOLUME 3

Manuel B. Aalbers
University of Leuven, Heverlee, BE, Belgium

Tassé Abye
University of Nouakchott, Ksar, Nouakchott, Mauritania

Christopher Adair-Toteff
University of South Florida, FL, USA; and University of Kent, Kent, UK

Ana Adan
Department of Psychiatry and Clinical Psychobiology, School of Psychology, University of Barcelona, Barcelona, Spain; and Institute for Brain, Cognition and Behavior (IR3C), University of Barcelona, Barcelona, Spain

A. Al-Azmeh
Yale University, New Haven, CT, USA

Catherine Alexander
Durham University, GBR, Durham, UK

Raphael Amit
University of Pennsylvania, Philadelphia, PA, USA

Margo Anderson
University of Wisconsin – Milwaukee, Milwaukee, WI, USA

James Andreoni
University of California, La Jolla, CA, USA

Arja R. Aro
University of Southern Denmark, Esbjerg, Denmark

Marietta L. Baba
Michigan State University, East Lansing, MI, USA

Constance E. Bagley
Yale University, New Haven, CT, USA

Ian Bailey
School of Geography, Earth and Environmental Sciences, Plymouth University, Plymouth, UK

Steven Barrie-Anthony
University of California, Santa Barbara, CA, USA

Robert Barsky
Vanderbilt University, Nashville, TN, USA

Yehuda Baruch
University of Southampton, Southampton, UK

Howell S. Baum
University of Maryland, College Park, MD, USA

Helen Beebee
University of Manchester, Manchester, UK

Richard Bellamy
London University, London, UK; and European University Institute, Florence, Italy

Gustavo Benavides
Villanova University, Villanova, PA, USA

Andrew Bennett
Georgetown University, Washington, DC, USA

Ivan T. Berend
UCLA, CA, USA

Hartmut Berghoff
German Historical Institute, Washington, DC, USA

Barry J. Blake
La Trobe University, Bundoora, VIC, Australia

Mary H. Buckingham
Institute for Applied Research in Youth Development, Tufts University, Medford, MA, USA

Philip Burnham
University College of London, London, UK

Ayse Caglar
University of Vienna, Vienna, Austria

Craig Calhoun
New York University, NY, USA

Linda D. Cameron
Psychological Sciences, University of California-Merced, Merced, CA, USA

Beth Casey
Durham University, Durham, UK

Robey B. Champine
Institute for Applied Research in Youth Development, Tufts University, Medford, MA, USA

Dennis W. Cheek
National Creativity Network, USA; and IESEG School of Management, France

Sum Kwing Cheung
Early Childhood and Elementary Education Division, School of Continuing Education, Hong Kong Baptist University, HKSAR, China

Jonathan Coleman
Duke University, Durham, NC, USA

Sara Compion
University of Kentucky, Lexington, KY, USA

Nicole M. Conand
Design Practices Collaborative, CA, USA

Thomas G. Cummings
University of Southern California, Los Angeles, CA, USA

Michael A. Cusumano
MIT Sloan School of Management, Cambridge, MA, USA

Georgia Dalto
University of Nevada Reno, Reno, NV, USA

Charles N. Darrah
San Jose State University, San Jose, CA, USA

Veena Das
Johns Hopkins University, Baltimore, MD, USA

Jesus de Felipe-Redondo
University of Michigan, Ann Arbor, MI, USA

Lisette M. DeSouza
Institute for Applied Research in Youth Development, Tufts University, Medford, MA, USA

Inna F. Deviatko
Higher School of Economics — National Research University, Moscow, Russia

René Devisch
University of Leuven, Leuven, Belgium

Nisha Dogra
Greenwood Institute of Child Health, University of Leicester, Leicester, UK

Alicia Dornadic
Design Practices Collaborative, CA, USA

Rafiq Dossani
Center for Asia Pacific Policy, RAND Corporation, Santa Monica, CA, USA

Samina Dossani
School of Oriental and African Studies (SOAS), La Jolla, CA, USA

Donald A. Downs
University of Wisconsin, Madison, WI, USA

Samuel N. Eisenstadt
Van Leer Jerusalem Institute, Jerusalem, Israel

Nicholas G. Evans
Centre for Human Bioethics, Monash University, VIC, Australia

Hubert Feger
Free University of Berlin, Berlin, Germany

Christine Fekete
Swiss Paraplegic Research, Nottwil, Switzerland

William C. Follette
University of Nevada Reno, Reno, NV, USA

Barry J. Fraser
Science and Mathematics Education Centre, Curtin University, Perth, WA, Australia

Claire Freeman
Geography Department, University of Otago, Dunedin, Aotearoa, New Zealand

Nick Frost
Leeds Metropolitan University, Leeds, UK

Todd Nicholas Fuist
Loyola University Chicago, Chicago, IL, USA

Itzhak Galnoor
Hebrew University, Jerusalem, Israel; and The Van Leer Institute, Jerusalem, Israel

Joshua Gamson
University of San Francisco, San Francisco, CA, USA

Bob Garvey
York St John University, Business School, York, UK

Annabelle Gawer
Imperial College Business School, London, UK

William Gerin
Pennsylvania State University, PA, USA

Andrew C. Gordon
University of Washington, Seattle, WA, USA

Rupa G. Gordon
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Phillip Grant
Justus-Liebig-University, Giessen, Hesse, Germany

Kathleen N. Greenman
Institute for Applied Research in Youth Development, Tufts University, Medford, MA, USA

Jeremy M. Grimshaw
Ottawa Hospital Research Institute, Centre for Practice-Changing Research (CPCR), The Ottawa Hospital, Ottawa, ON, Canada; and University of Ottawa, Ottawa, ON, Canada

Colette Grinevald
University of Lyon, Lyon, France

David B. Grusky
Stanford University, Stanford, CA, USA

Joseph P. Gyekis
Pennsylvania State University, PA, USA

Simon Hackett
Durham University, Durham, UK

Christopher A. Hafen
University of Virginia, Charlottesville, VA, USA

James F. Hamilton
Grady College of Journalism and Mass Communication, The University of Georgia, Athens, GA, USA

Ulf Hannerz
Stockholm University, Stockholm, Sweden

Adrian Hastings[†]
University of Leeds, Leeds, UK

Peter Hay
Emory University School of Law, Atlanta, GA, USA; and Technische Universitaet Dresden, Dresden, Germany

Harlene Hayne
University of Otago, Dunedin, New Zealand

Susan Hellwig
Bergische Universität Wuppertal, Wuppertal, Germany

Deborah R. Hensler
Stanford Law School, Stanford, CA, USA

Joshua Hermias
Washington, DC, USA

Andreas Hess
University College Dublin, Dublin, Ireland

Jennifer Hill
Steinhardt School of Culture, Education, and Human Development, New York University, New York, NY, USA

Mark Hoipkemier
University of Notre Dame, Notre Dame, IN, USA

Paul W Holland
University of California, Berkeley, CA, USA; and Educational Testing Service, Princeton, NJ, USA

Cang Hui
Centre for Invasion Biology, Department of Mathematical Sciences, Stellenbosch University, Matieland, South Africa; and African Institute for Mathematical Sciences, Muizenburg, South Africa

Michael L. Humphries
Southern Illinois University, Carbondale, IL, USA

A.C. Huston
University of Texas, Austin, TX, USA

Kana Imuta
University of Queensland, St Lucia, QLD, Australia

H. Islamoglu
Bogazici University, Istanbul, Turkey

Thomas Janoski
University of Kentucky, Lexington, KY, USA

Abigail Jones
Washington, DC, USA

Nicola Jones
Overseas Development Institute, London, UK

Laura L. Junker
University of Illinois at Chicago, Chicago, IL, USA

David Kairys
Temple University, Philadelphia, PA, USA

Robert W. Kentridge
Durham University, Durham, UK

Patrick V. Kirch
University of California, Santa Barbara, CA, USA

Bill Kissane
London School of Economics, Houghton St., London, UK

David M. Knight
University of Durham, Durham, UK

Jürgen Kocka
Free University of Berlin, Berlin, Germany

Brandon A. Kohrt
Duke Global Health Institute, Duke University, Durham, NC, USA

Menno-Jan Kraak
University of Twente, Enschede, The Netherlands

[†] Deceased.

Michal Kramarek
Indiana University Lilly Family School of Philanthropy, Indianapolis, IN, USA

Tahu Kukutai
The University of Waikato, Hamilton, New Zealand

Heinz D. Kurz
University of Graz, Graz, Austria

Anne Lafay
Centre de recherche de l'Institut universitaire en santé mentale de Québec, Québec, QC, Canada

Maria Leggio
Sapienza University of Rome, Rome, Italy; and Santa Lucia Foundation IRCCS, Rome, Italy

Richard M. Lerner
Institute for Applied Research in Youth Development, Tufts University, Medford, MA, USA

Birgit Leyendecker
Ruhr University Bochum, Bochum, Germany

Tim F. Liao
University of Illinois at Urbana–Champaign, Urbana, IL, USA

Elisabeth Lichtenberger
University of Vienna, Vienna, Austria

Wen Lin
Newcastle University, Newcastle upon Tyne, Tyne and Wear, UK

Shirley Lindenbaum
City University of New York, New York, NY, USA

Lamont Lindstrom
University of Tulsa, Tulsa, OK, USA

Pam G Lister
Glasgow, UK

Christoph Luetge
Technische Universität München, Peter Löscher Chair of Business Ethics, München, Germany

Kirsten Lundberg
Columbia University, New York, NY, USA

I. Lunt
University of London, London, UK

Joël Macoir
Université Laval, Québec, QC, Canada

Sujen M. Maharjan
Tribhuvan University, Kirtipur, Nepal

L. Sandy Maisel
Colby College, Waterville, ME, USA

Chaime Marcuello-Servós
Facultad de Ciencias Sociales y del Trabajo, Universidad de Zaragoza, Zaragoza, Spain

Edmund I. Marshall
University of Bradford, Bradford, UK

Mayte Martín
University College Dublin, Dublin, Ireland

Catherine McBride-Chang
The Chinese University of Hong Kong, Hong Kong, China

James H. McMillan
Virginia Commonwealth University, Richmond, VA, USA

Marco Molinari
Santa Lucia Foundation IRCCS, Rome, Italy

James Moody
Duke University, Durham, NC, USA

Pietro G. Morasso
Istituto Italiano di Tecnologia, Genoa, Italy

Frank Muscarella
Barry University, Miami Shores, FL, USA

John R. Nesselroade
University of Virginia, Charlottesville, VA, USA

Meinolf Noeker
University of Bremen, Bremen, Germany

Michael Northcott
Edinburgh University, Edinburgh, UK

Judy M. Olson
Michigan State University, East Lansing, MI, USA

Anthony M. Orum
University of Illinois at Chicago, Chicago, IL, USA; and Loyola University, Chicago, IL, USA

Jennifer Oser
University of Pennsylvania, Philadelphia, PA, USA

Jan Pakulski
University of Tasmania, Hobart, TAS, Australia

Christine Pauli
Department of Education, University of Fribourg, Fribourg, Switzerland

Diego Pereyra
Instituto de Investigaciones Gino Germani, Universidad de Buenos Aires – CONICET, Argentina

Franz Petermann
University of Bremen, Bremen, Germany

Christian Pfister
University of Bern, Bern, Switzerland

Shana Poplack
Linguistics Department, University of Ottawa, Ottawa, ON, Canada

Elizabeth Presler-Marshall
Chapel Hill, NC, USA

Charles C. Ragin
University of California, Irvine, CA, USA; and University of Southern Denmark, Denmark

Sauharda Rai
Transcultural Psychosocial Organization (TPO) Nepal, Kathmandu, Nepal

Brenda Reed
Facultad de Ciencias Sociales y del Trabajo, Universidad de Zaragoza, Zaragoza, Spain

Kurt Reusser
Institute of Education, University of Zurich, Zurich, Switzerland

Piers Revell
School of Government, Plymouth University, Plymouth, UK

Hauke Riesch
Media & Communications, Brunel University London, Uxbridge, Middlesex, UK

Gerhard Robbers
Institute for European Constitutional Law, University of Trier, Trier, Germany

Wade Clark Roof
University of California, Santa Barbara, CA, USA

Patrick Rooney
Indiana University Lilly Family School of Philanthropy, Indianapolis, IN, USA

Laura A. Rosenbury
Washington University Law School, St. Louis, MO, USA

Donald K. Routh
University of Miami, Coral Gables, FL, USA

Rosemary Radford Ruether
Claremont Graduate University, Claremont, CA, USA

Steven Ruggles
University of Minnesota, MN, USA

Erik A. Ruzek
University of Virginia, Charlottesville, VA, USA

Saskia Sassen
University of Chicago, Chicago, IL, USA

Damian Scarf
University of Otago, Dunedin, New Zealand

Mike S. Schäfer
Universität Zürich, Institut für Publizistikwissenschaft und Medienforschung, Zürich, Switzerland

Diane W. Schanzenbach
Northwestern University, Evanston, IL, USA

Philip W. Scher
University of Oregon, Eugene, OR, USA

Birgit Schippers
St Mary's University College Belfast, Northern Ireland, UK

Vaughn Schmutz
UNC Charlotte, Charlotte, NC, USA

Alain Schnapp
Institute of Art and Archeaology, Paris, France

Michael J. Selgelid
Centre for Human Bioethics, Monash University, VIC, Australia

Steven Shackley
University of California, Berkeley, CA, USA

John Rennie Short
University of Maryland Baltimore County, Baltimore, MD, USA

Graeme Simpson
University of Wolverhampton, Wolverhampton, UK

Silvia Siu-Yin Lam
Communication Sciences and Disorders Northwestern University, Chicago, IL, USA

Leslie Sklair
London School of Economics, London, UK

Heike Spaderna
Bergische Universität Wuppertal, Wuppertal, Germany

Rolf Steyer
Friedrich-Schiller-Universität, Jena, Germany

Daniel Stoljar
Australian National University, Canberra, ACT, Australia

Elizabeth A. Stuart
Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

Luděk Sýkora
Charles University in Prague, Prague, Czech Republic

Monica Taljaard
Ottawa Hospital Research Institute, Ottawa Hospital, Ottawa, ON, Canada; and University of Ottawa, Ottawa, ON, Canada

Victor R. Thompson
Rider University, Lawrence, NJ, USA

Elizabeth Thomson
Stockholm University, Stockholm, Sweden; and University of Wisconsin—Madison, Madison, WI, USA

Xiuli Tong
Faculty of Education, The University of Hong Kong, Hong Kong, China

Paul Tranter
Geography School of Physical, Environmental and Mathematical Sciences, UNSW Canberra, ACT, Australia

Yves Turgeon
Clinique de mémoire, Centre de santé et de services sociaux du Rocher-Percé, QC, Canada

Dennis C. Turk
University of Washington, Seattle, WA, USA

Bryan S. Turner
The Graduate Center, CUNY, New York, NY, USA

Paolo Valente
United Nations Economic Commission for Europe, Geneva, Switzerland

Norman Vance
University of Sussex, Falmer, Brighton, UK

Alex van Venrooij
University of Amsterdam, Amsterdam, The Netherlands

James M. Vardaman
School of Culture, Media and Society, Waseda University, Shinjuku-ku, Tokyo, Japan

Jackline Wahba
University of Southampton, Southampton, UK

Michael S. Wald
Stanford University, Stanford, CA, USA

Christopher M. Wally
Psychological Sciences, University of California-Merced, Merced, CA, USA

Daniel J.A. Warren
Institute for Applied Research in Youth Development, Tufts University, Medford, MA, USA

Jeffrey N. Wasserstrom
University of California, Irvine, CA, USA

Shino Watanabe
Sophia University, Tokyo, Japan

Linda R. Watson
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Charles Weijer
Rotman Institute of Philosophy, Western University, London, ON, Canada

Thomas S. Weisner
University of California Los Angeles, Los Angeles, CA, USA

Richard Whatmore
School of History, The Scores, University of St Andrews, St Andrews, Fife, UK

Rhys H. Williams
Loyola University Chicago, Chicago, IL, USA

Robert K. Yin
COSMOS Corporation, Bethesda, MD, USA

Gary W. Yohe
Wesleyan University, Middletown, CT, USA

Tanja Zimmermann
Institute of Psychology, Technische Universität Braunschweig, Braunschweig, Germany

Christoph Zott
IESE Business School, Barcelona, Spain

Business and Society: Social Accounting

Edmund I Marshall, University of Bradford, Bradford, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 2, pp. 1418–1421, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article describes all the interfaces between a modern business organization and its surrounding community, for business success depends on meeting the requirements and demands of that community. Advertising market research, and public relations are all means of influencing and matching the requirements of the market. The application of business ethics, including concern for the physical environment, is a means of enhancing the social and economic standing of a business. Equally important is the use of up-to-date technological developments within the operation of a business, with sensitive awareness of their impact on industrial relations. At the same time, all businesses are affected by political processes introducing new laws and taxes, and by the changing structure of the population in age, balance between the sexes, and employment requirements. Each of these interfaces is examined in turn. Social accounting is more concerned with macroeconomics investigating how whole industries and institutions relate to the wider economy and society, and how the production of goods and services income formation, and income distribution relate to each other. Social accounting techniques have been widely applied in the fast developing economies of the third world.

To succeed, any modern business manager needs to understand all the interfaces between the business and wider society no less than all the internal operations of the business itself. For no business can survive, let alone flourish, unless it provides goods or services for which customers outside the business will be prepared to pay. Business success depends on meeting market demand. Awareness of market conditions and of how the market may be influenced is a key part of business management. In this article the various interfaces between businesses and their social context are briefly examined, using a form of analysis which has come to be known by the acronym SLEPT, in which the different letters stand respectively for the Social, Legal, Environmental, Political, and Technological factors influencing business organizations.

Social Factors

Market demand depends on the needs, desires, and interests of people in society, and knowledge of these is essential if a business is successfully to meet that demand. Such factors may be determined by the physical climate, by the language and culture, or by the demographic structure of a society. There is likely to be little demand, for example, for heavy overcoats in tropical countries, or for umbrellas in arid lands. In some cultural settings it is still rare for women to seek paid employment. In others the increase in life expectancy may have resulted in increasing demand for leisure activities for pensioners who have retired from employment. Changes to birth rates cause fluctuations in demand for goods and services required by children. These are examples of the ways in which knowledge of the market place will help a business entrepreneur to decide what goods to produce and offer for sale to meet demand.

Part of the cultural context of a business is the prevailing ethos or climate of moral opinion about what are right standards for business activity and trading. During the last two decades, such questions of business ethics have figured increasingly in public concern and on the agendas of business organizations. These highlight the need for complete probity and accountability in business, and for more openness in the operations of boards of directors (Carroll and Buchholtz, 2012). Ethical considerations may involve how a firm exploits its markets, how it obtains raw materials or cheap labour for its operations in different parts of the world, or how its activities affect the physical environment (which will be explored more fully in Section Environmental Factors).

The ethical standing of a business will affect its public relations, which in turn may impact upon the demand for its products, one way or the other. Modern businesses, especially larger ones, pay increasing attention to their public relations and how their images – and the brand images of their products – appear to the general public. Part of that public image is determined by how a business is presented by the news media, part by the reputation which products have among customers, and part by the business' own public advertising of its products. No successful business can ignore its need for a continual campaign to advertise its products, for in this way it can significantly influence market demand and maintain its share of markets with strong competition. The outlets for advertising are constantly spreading, using newspapers, magazines and journals, individual mail-shots, public hoardings, television and radio, and more recently, the Internet. This leads to the whole field of electronic commerce to be considered more fully in Section Technological Factors. At the same time, businesses need to be conducting market research, assessing public opinion about their existing products and exploring what other demands there are in the marketplace for potential new products. Market research enables a business to maintain direct links with any community where it seeks to trade.

The social factors relating to a business company have been recognized in many circles by introducing the concept of stakeholders of the company, who are all the groups of people with a legitimate interest in the operations of the company, not

just the shareholders. Stakeholders equally include the employees of a business, the customers, local residents and communities, and representatives of environmental interests (Carroll and Buchholtz, 2012). For many years there have been schemes of co-ownership and copartnership which have involved employees in the actual ownership of companies. More recently, and particularly in the United States, some corporations have introduced special constitutional features so that there are boards with representation from all stakeholder groups. A growing number of large firms are now publishing regular social or environmental reports as a means of accounting to all their stakeholders; and independent professional bodies are now issuing reviews of whole industries in this way, such as the Turnbull Report of the Association of Chartered Accountants of England and Wales. These arrangements vary in their levels of formality, but they are all attempts to involve wider social interests in the decision-making processes of the company. The growth of consumerism and of specific organizations established to protect and further consumer interests has inevitably brought increasing pressures to bear on the decisions of many businesses.

Legal Factors

The framework for business activity in any society may be constrained by legal requirements in that society. As long ago as 1570, the English Parliament enacted a law requiring every member of the 'lower classes' over the age of 6 years to wear a woollen cap on Sundays and holidays. Such legislation would today be regarded as a gross infringement of personal freedom, but it was introduced to further the interests of the politically powerful sheep farmers and wool traders at the time.

In later centuries in Britain there was significant dismantling of such legal constraints on business, with the ending of the mercantile system and the growth of '*laissez faire*.' Free market economists like Adam Smith, Thomas Malthus, David Ricardo, and John Stuart Mill advocated the freeing of business markets from legal restraint and control, so that the 'invisible hand' of free market forces could determine the price levels at which supply would equal demand. It was maintained that the interests of both producers and consumers were optimised by the free play of market forces, and that this required the absence of legal constraints, to enable businesses best to serve society by pursuing their own interests as vigorously as possible.

During the last 180 years, however, the political pendulum has swung very much the other way. Acts of Parliament in Britain ended the slave trade, banned the employment of women and children in coal mines, and introduced legislation for the health and safety of employees in factories and other places of work. Other statutes provided for the registration of business companies, giving them legal entity as corporate bodies rather than partnerships of all their various owners, and limiting the financial liability of shareholders to what they paid for their shares. Increasingly under governments of differing political shades, more and more legislation has come into force in virtually all countries of the world which sets the standards and requirements to be observed in business trading. Business managers need a very full awareness of the legal framework in which they are operating, both of the statutes established by political processes and of the decisions of courts in applying those statutes and developing case law.

Environmental Factors

Concern for the health and safety of employees is one example of a limitation on business operations to curb the potentially harmful effects of those operations. Just as people within the workplace may be injured and exposed to health risk, so too the activities of a business, in both production and distribution, may be detrimental to people living in the vicinity, along transport routes, and in the community generally. The discharge of waste materials from industrial processes, be they solid, liquid, or gaseous, on to waste land, into rivers, or into the atmosphere can have harmful effects reaching far into other countries, and into the ozone layer above the atmosphere. There is global concern that raw materials are being depleted at rates which will restrict or even exhaust their availability for future generations. For instance, the rain forests of Brazil are being decimated by the worldwide demand for timber and paper (Parker, 2011). Similarly, known reserves of fossil fuels in the world are finite. Public opinion is increasingly aware of the environmental effects of business activities in the widest contexts and longest timescales, and there is growing demand that all economic activity should be environmentally sustainable and without permanent damage to the planet.

Most goods supplied to the market today are packaged in some form, and difficulties may arise in the disposal of the packaging. Cardboard, paper, bottles, and some metallic containers can now be recycled for reuse, but containers made of some plastics have no future use. Manufacturers therefore need to be more aware of the environmental impact of the way they package their products. And in this age of the throw-away society, there are immense problems arising from the disposal of goods which are no longer of use to their owners. It is not easy for society to get rid of motor cars and household applicances which have outlived their useful lives, and there are limits to how the world can dispose of its growing mountains of rubbish. All such considerations challenge business operators to develop products which are easier to recycle or get rid of, or which will last much longer. Political pressures on governments will almost certainly lead to more legal constraints on business to meet these challenges.

Political Factors

Section Legal Factors highlighted the ways in which business activities can be constrained by laws in force in society, and in most cases those laws are made, amended, or repealed through the established political processes of the country in question. Political structures vary between countries, and any business manager needs to understand the working of those structures in any country where the business has operations, in securing raw materials, in production, or in marketing and distribution. In countries which have democratic structures, a business needs to be constantly alert to what changes may come about in public policies. The programs of potential alternative governments need to be scrutinized and possibly influenced through the use

of open democratic processes, so that when new laws are enacted they do not take the business by surprise and may be more helpful for the business than would otherwise have been the case.

Equally important is an understanding of the administrative processes of government through which laws and statutes are implemented. A business must be especially alert to the taxation implications of all its activities, and be ready to rearrange and retime those activities to reduce its tax liabilities. Similarly, the day-to-day administration of other departments and agencies of government may bear on the activities of a business, and it is essential for the manager to have an easy means of communication with appropriate public servants in order to sort out any problems which may crop up for the business. The best means of cooperation in this respect will vary from country to country, but one general principle is that all representations to officials must be open and never underhand nor illegal.

Technological Factors

In Section Social Factors it was noted that technological changes provide another important influence on how a business is run, introducing new products to be manufactured, new methods of manufacturing them, new means of packaging and distribution, new methods of advertising and marketing, and new means of communication to be employed in all operations. The pace of technological change appears to be accelerating all the time, and the contemporary electronic revolution in communicating information is introducing a new way of business life for all large organizations. The mobile phone, the fax, and email enable decisions to be taken and implemented instantly.

Technological changes can bring vast benefits to producers and consumers alike, but they may also pose difficult problems, particularly in the management of human resources. As old machinery becomes obsolete, staff need to be retrained to use new technology or redeployed to new activities because the new technology is not so labor-intensive. Many new machines take the place of people, so it is not unusual for established workforces to resist the introduction of new technology, and industrial relations may become fraught. Nowhere was this shown more dramatically than in the English newspaper industry in the 1980s, when decisions to introduce new technology for print-setting led to mass picketing and some violence at Wapping in east London.

At the same time, the modern revolution in electronic communications enables many more people to work at home in more conducive surroundings and avoiding the traffic congestion associated with commuting to work. Word processors, fax machines, and emails enable the workforces of many businesses to be geographically scattered, with many individuals working on a contract or consultancy basis on tasks which had previously to be done in central offices. Add to all this new technology the far-reaching potential of the Internet, and it becomes possible for more and more business to be conducted electronically, with consumers ordering, and paying for, goods on the Internet; the effectiveness of such commerce depends partly on the strength and security of computer systems, and partly on the reliability of networks used for the prompt distribution of goods.

Social Accounting

Whereas the foregoing description of the interfaces between business and society was in terms of the links between an individual business and its social context, social accounting is more concerned with macroeconomics, involving a study of how whole industries and institutions relate to the wider economy and society, and of how the production of goods and services, income formation, and income distribution relate to each other. Social accounting techniques have been widely applied in the fast developing economies of the third world, such as Iran, Malaysia, Sri Lanka, Pakistan, Indonesia, Ecuador, and Swaziland.

In the application of these techniques, Social Accounting Models, or SAMs, can be constructed for specific countries and used as integrated databases for understanding the effects of national economic policies. While the British Treasury has used Economic Models for over a quarter of a century, in most of the less developed countries similar SAMs have now been built, varying greatly according to the economic characteristics of local economic conditions. Such models may be diagrammatically represented as shown in the literature, and are increasingly used in the formation of public economic policy.

Social accounting concerns microeconomics as well, since a growing number of corporations are now active in social accounting, auditing and reporting. They publish social and environmental reports on a yearly basis, which provide detailed information about the relations with their stakeholders, such as investors, employees, consumers, suppliers, central and local governments, and environmental associations. Financial rating agencies evaluate firms not only in terms of their economic and financial performance, but also in terms of their social, ecological and ethical performance.

See also: Advertising and Advertisements; Advertising: General; Business Law; Environmental Risk; Marketing Strategies; Technology and Social Control.

Bibliography

Alarcon, J., van Heemst, J., Keuning, S., de Ruijter, W., Vos, R., 1990. The Social Accounting Framework for Development. Avebury, Aldershot, UK.

Carroll, A., Buchholtz, A., 2012. Business and Society: Ethics, Sustainability, and Stakeholder Management, ninth ed. Cencage Learning, Stamford, CT.

Jefkins, F., 1987. Public Relations for your Business. Mercury Books, London.

Marshall, E., 1993. Business and Society. Routledge, London.

Pearce, D., 1991. Corporate Responsibility and the Environment. British Gas plc.

Parker, L.D., 2011. Twenty-one years of social and environmental accountability research: A coming of age. Accounting Forum 35 (1), 1–10.

Business Anthropology

Marietta L Baba, Michigan State University, East Lansing, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article is a historical overview of some of the influences of business upon the development of sociocultural anthropology and the consequences thereof, and a review of other major developments that shaped business anthropology over the past century. The article also discusses the most prominent areas of inquiry and practice in business anthropology that have emerged recently and how these have contributed to the current status of the subfield.

Introduction

Business anthropology is inquiry or practice in the business domain that is grounded in anthropological epistemology, methodology, or substantive knowledge. Business anthropology also may be viewed as a term of reference for the overarching field in which anthropology has forged much of its experience in the private sector. In its most general sense, business refers to the buying and selling of goods and services in the marketplace, and the organized economic activities attendant to those practices. Anthropologists have been engaged in the business domain since the early twentieth century as scholars, practitioners, and strategic consultants. From these encounters, numerous designations have emerged in anthropology: corporate, enterprise, organizational, design, finance (anthropology), and so on. The proliferation of alternative terms for the construct 'business anthropology' is a reflection of several historical currents, including tensions in the relationship between anthropology and business in the discipline, difficulties surrounding the early uses of anthropological knowledge in applied anthropology, and a lack of consensus surrounding efforts to institutionalize the study of business-related phenomena in anthropology.

The twentieth century was a turbulent period for the relationship of anthropology and business, encompassing both theoretically serious engagement through the Human Relations School and sharp criticism through Marxist and neo-Marxist studies of industry. The latter decades of the century saw more anthropologists in Europe, North America, and Asia entering into increasingly diversified relationships with businesses, including the absorption of anthropology and ethnography into the business realm and the commodification of ethnographic practice into the private sector. These developments have called for self-reflection with respect to anthropologists' engagement in business, although, for the most part, the long-term consequences have not been confronted. The present situation is one of gradual institutionalization of business anthropology in society, with hundreds of anthropologists around the globe studying about and practicing in business, and the emergence of dedicated journals and conferences (e.g., *Journal of Business Anthropology*, *International Journal of Business Anthropology*, Ethnographic Praxis in Industry Conference).

It is important not only to examine specific instances of anthropological engagement in business activity but also to explore the ways in which business has influenced anthropology. Not only is business influential in and formative of society, but as a field of study business is juxtaposed to other developments in social science that could be important to anthropology.

Anthropology and Business: Early Influences

A received view of anthropology's relationship with business often places its origins in the Hawthorne Project of the Western Electric Corporation. While these events were unprecedented in shaping a dominant paradigm for social science in US industry, it is also important to examine the role of American business in financing anthropological research in the British colonies to grasp the dynamics underlying the relationship of anthropology and business.

The idea of social anthropology (i.e., empirically grounded, theoretically relevant studies of small-scale society) initially was not provided with much support by colonial governments in Europe. To obtain financing for field research, aspiring anthropologists found it necessary to 'sell' social anthropology to colonial administrators based on its practical relevance. Such approaches were not particularly effective: colonial administrators and anthropologists competed for the same professional niches; anthropologists were disinclined to undertake practical problem-solving tasks; and there was conflict between colonial policies and anthropologists' views. It is not surprising that there were few funded chairs of social anthropology in the colonial empire, and there were no other significant sources of funding for anthropological research at the time.

Social anthropology may not have emerged in Europe at all had the Laura Spellman Rockefeller Memorial (LSRM, a branch of the Rockefeller Foundation) in the United States not set out to establish the social sciences as empirical fields that could strengthen the practical basis for social welfare. At the time, the rise of the natural sciences created pressure for a positivist approach for social science, which required funding for equipment and personnel. The financing strategy for this was devised by Beardsley Ruml, a psychology PhD from the University of Chicago who reconceptualized the social work agenda of the LSRM and distributed $21 million in grants between 1922 and 1929.

The 1920s was an era of 'welfare capitalism' in the United States, when corporations in the private sector provided

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12231-7

benefits to employees to persuade them to draw closer to a management agenda and away from unionization. This approach halted the rise of unions but was expensive to firms and was implemented on an ad hoc basis. An empirical social science could provide a more systematic basis for social welfare policies, both public and private.

Ruml's mission was politically sensitive. John Rockefeller Jr. (who oversaw the LSRM) and the parent Rockefeller Foundation (RF) had to avoid the perception of private interests directing any of the Foundation's subsidiaries. The political climate of the Progressive Era had stoked sharp criticism of the Rockefeller family by politicians and the media, particularly when the RF was used to support a study of industrial relations as a response to labor-management conflict at one of the companies controlled by the Rockefellers, after which the RF discontinued all social science research.

Thus, Ruml was careful to direct LSRM funding toward basic empirical social science research, with arms-length relationships between Rockefeller funds and the grantees. He also sought to ensure that such research led toward a positivist vision for the social sciences and social welfare aims. Ruml traveled the world, visiting institutions that met his criteria of excellence, and he met with prominent intellectuals and promising individuals. He provided block grants to outstanding institutions and allowed them to make awards to individuals. He also guided chosen individuals to institutions where funding already had been arranged.

Four major names in the history of anthropology were guided and funded by Ruml, the LSRM, and/or the RF: Bronislaw Malinowski, Elton Mayo, Alfred R. Radcliffe-Brown, and W. Lloyd Warner. In turn, these individuals influenced the relationship between anthropology and business through the Hawthorne Project, the Human Relations School, the Yankee City Series, and the intellectual ties of these programs with British social anthropology.

Perspectives on the Hawthorne Project

Ruml met the Australian psychologist Mayo early in his search for funding opportunities, and he was attracted to Mayo's ideas about industrial management and a potential role for social science. Mayo believed that labor-management relations reproduced society-wide class relationships that generated hostility within industry and a form of class warfare. He blamed management for failing to consult employees in decision making, and he saw a role for the social sciences in bringing knowledge-based guidance to the relationship. Ruml placed Mayo at the University of Pennsylvania, using Rockefeller's personal funds to support Mayo's research in industry. At Penn, Mayo began to read anthropology, and he began to send Malinowski manuscripts for comment.

Progress on Mayo's research at Penn and the strictures of that placement led Ruml to move Mayo to the Harvard Business School, where he was funded directly by the LSRM. At Harvard, Mayo began to meet corporate executives through Rockefeller's networks, including the personnel director of Western Electric Corporation, who told him about their experiments on worker fatigue at the Hawthorne Plant in Cicero, Illinois (Gillespie, 1991).

Mayo had come to realize that the Hawthorne workers were not willing to be open with him regarding their private thoughts or to share details of their lives outside the plant, which were emerging as necessary requirements in Mayo's empirical framework (Gillespie, 1991). From his reading of the literature, Mayo believed that an anthropologist might be able to access such information. The anthropologist W. Lloyd Warner came on to the scene when he was appointed as a tutor and then an assistant professor of anthropology at Harvard, and was scouting for someone to help him set up a community study in the United States where he could experiment with anthropological methods. At the same time, Mayo was interested in identifying an anthropologist who could initiate a community study in the neighborhood around Cicero, Illinois. Warner famously spent time at Hawthorne where he was involved in designing the final phase of the experiments, the Bank Wiring Observation Room (BWOR), which Warner did not carry out himself, but was conducted by Hawthorne researchers. The BWOR confirmed that workers and supervisors systematically restricted output, based upon an informal organization of the workforce (i.e., workers' spontaneous social relationships). Though Warner declined to establish his own community study in Cicero, in designing the BWOR he developed a methodology with characteristics that approximate ethnography, including the synchronized collection and analysis of observational and interview data that create an empirical record of naturalistic behavior in the workplace, thereby establishing a standard for future ethnographic studies of work (Roethlisberger and Dickson, 1939; Schwartzman, 1993).

Mayo and Warner differed sharply in their interpretations of data emerging from the BWOR. Warner and Hawthorne researchers interpreted the findings on output restriction as evidence of distrust between workers and managers based upon a dysfunctional industrial relations regime, where workers believed that if productivity increased management would cut the piece rate, so they had a disincentive to produce more per hour. Mayo rejected this view and insisted on a 'psychopathological' interpretation. According to Mayo, the 'informal organization' of the workforce could support or oppose management, and in this case, workers had developed an illogical, maladjusted perception of management that required intervention. Mayo's close relationship with Western Electric's top leadership enabled him to gain control over the Hawthorne data (which he moved to Harvard), and ultimately, his analysis of the data prevailed (Gillespie, 1991).

Human Relations School

With his interpretation of the data secure, Mayo launched a major counseling program at Hawthorne, to address industrial relations problems. This approach became the basis for the Human Relations School, which incorporated the leading works of anthropologists and psychologists to adjust labor-management relations in the name of 'scientific knowledge.' Mayo envisioned an 'administrative elite' that would carefully apply such knowledge to issues of social organization and control, echoing the interests of the RF and LSRM.

A group of anthropologists (e.g., Conrad Arensberg, Eliot Chapple, Burleigh Gardner, F.L.W. Richardson, and W. Lloyd

Warner) was sufficiently interested in the Human Relations School to warrant a launch of their own brand, with an emphasis on social structure, systems relationships, and human interactions rather than psychology (Schwartzman, 1993). When Warner moved from Harvard to the Department of Anthropology at Chicago and others joined him, this group was enabled to conduct research and consult in businesses from an anthropological perspective. The distinctive methodology employed by this group – the systematic sociology of small-scale groups, recording of behavioral interactions, and efforts to balance their focus upon workers and managers – became hallmarks of applied anthropology in industry in the United States (Partridge and Eddy, 1978) and could be traced to the intellectual tradition of British social anthropology and Warner's methodological innovations in the Yankee City Series (see below).

Adherents of the Human Relations School embraced a functionalist theory of society, in which an equilibrium state (i.e., all parts interact to support the whole) was considered normal and conflict pathological. Functionalist theory was well known and accepted in Rockefeller circles, and it was supported by industrialists such as John Rockefeller Jr., who promoted it widely. Interventions recommended to corporate managers by members of the Human Relations School were perceived by their practitioners as advancing a 'science of society' while also contributing to human welfare (i.e., promoting industrial peace and prosperity), in keeping with the goals of the LSRM. These goals were of special relevance in the years leading up to and during World War II, when industrial productivity was important to the nation's war effort.

Efforts to establish business anthropology in the United States were an integral aspect of a larger project to institutionalize applied anthropology in the United States led by pioneers of the discipline (e.g., Conrad Arensberg, Gregory Bateson, Ruth Benedict, Margaret Mead, George Murdoch, Julian Steward, and others; Partridge and Eddy, 1978). The American Anthropological Association (AAA) declined to allow this group to establish its own section within the AAA, more than hinting at the marginalization of practical anthropology on both sides of the Atlantic, and helping to explain why the American group formed the Society for Applied Anthropology (SfAA) at Harvard in 1941. The SfAA's mission was to promote 'scientific investigation' of principles underlying human relationships and to encourage the 'wide application of these principles to practical problems,' a theme that resonated with the goals of the LSRM (see for discussion Baba and Hill, 2006).

British Social Anthropology

British social anthropology was a source of theoretical and methodological rigor and innovation whose fusion of theory and ethnographic method was financed by the LSRM. Ruml was the benefactor of the London School of Economics (LSE) in the early 1920s, providing $1.25M between 1924 and 1928 (Bulmer and Bulmer, 1981; Stocking, 1995). While some of these funds benefited Malinowski and his students through the International Institute of African Languages and Cultures, the entrepreneurial Malinowski also sought funds directly from the RF. He wrote a proposal for systematic fieldwork on the tribal context of modern economic activities such as native mining, which was attuned to the RF's interests in practical applications of knowledge and addressed interventionist strategies for the colonies (Mills, 2002). Malinowski was viewed with favor by the RF, which was growing impatient with the 'lack of cooperation' from anthropologists in the United States (Goody, 1995). When the RF voted to fund Malinowski directly, his protégés at the LSE were placed in a strategic position to receive continuing support from the Colonial Social Science Research Council, which took over funding of anthropological research in the colonies after the RF wound up its support at the end of the 1930s due to the financial stress of the Great Depression (Mills, 2002).

This development provided an opportunity for Malinowski and his students to attempt to effect policy-driven economic change within the colonial context and to recognize the complexities and fraught nature of such efforts, an irony that was not lost on historians of anthropology (Stocking, 1995). After these experiences, Malinowski's enthusiasm for practical anthropology as an approach toward improvement of the colonial apparatus appears to have diminished, reinforcing the impression among some American anthropologists that practical applications in the policy realm were a risk-prone and ill-advised endeavor.

Meanwhile, Radcliffe-Brown was establishing an empirical basis for the study of kinship among the indigenous people of Australia, based on funding received from the RF through the Australian National Research Council. Radcliffe-Brown viewed society as a 'natural system,' with social structure a primary explanatory variable in relation to similarities and differences across cultures. Although the RF approved of his approach, the Australian establishment became irritated by his ersatz aristocratic and outré lifestyle and chafed at his criticism of their colonial policies. When Radcliffe-Brown tried to bypass the Australian authorities to receive funding directly from the RF near the end of his 5-year term as chair, the resulting crisis in the colony was resolved by his decision to accept an offer of employment from the University of Chicago. While at Chicago (1931–37), Radcliffe-Brown outlined his theory of structural-functionalism, and he was influential in the debate between the scientific and historical forces within anthropology, swaying a new generation away from history and toward the social sciences, the direction promoted by the RF. Significantly, however, Radcliffe-Brown did not agree with Malinowski regarding the application of anthropology to policy questions. He was a firm advocate of maintaining the primacy of theoretical questions in anthropology and leaving policy to administrators. Thus, his views were aligned with those of many academic anthropologists in the United States during the middle and late twentieth century.

The 'scientific' direction of American sociocultural anthropology strengthened by Radcliffe-Brown's structural-functionalism became a driving force within the nascent movement of applied anthropology in the United States as it gathered momentum during World War II. Adherence to science and policy as orienting frames of reference have been features that distinguish practical anthropology in America, and they suggest the intellectual foundations of its eventual association with business anthropology.

Yankee City Study

W. Lloyd Warner, Radcliffe-Brown's protégé in Australia and a former student of Malinowski at Berkeley, modernized the foundational theory and methods of British social anthropology in his own version of small-scale society research in the Yankee City Series. In this 5-year project supported by the RF, Warner and his research group developed an innovative methodology for assigning individual subjects to social classes, ranks, and statuses, and examining interactions among members of social classes within various associations (e.g., families, cliques) through interviews and ethnographic observations. This project defined the standard American social class hierarchy of lower, middle, and upper, each with a lower and upper tier (Warner and Lunt, 1941). From his complex matrix of social interaction data, Warner was able to make observations about relationships among social class and other phenomena, such as the relocation of capital outside of small towns, shifts in alliances across social classes and resentment against the wealthy, unprecedented strikes in local factories and the formation of industrial unions, and the development of support for strikers among town folk (Warner and Low, 1947).

The conceptual and methodological advances of the Yankee City Series, whose first volume was published the same year the SfAA was established (1941), grounded and legitimized applied anthropology in the United States. Ironically, the Human Relations School with which Warner was associated met its demise due to the rise of organized labor and its practices of ritualized conflict, leading to the fall of functional equilibrium theory. Critics of the Human Relations School claimed that its theory and methods failed to acknowledge unequal power relationships within corporations and upheld these relationships through manipulative interventions (Burawoy, 1979). Whether such criticism is fair is a matter of debate. Nevertheless, it is likely that this critique was a factor in the lag between the fall of this School in the 1950s and the time when business anthropology was accepted as a domain of application much later in the twentieth century.

Business Anthropology Following World War II

The academic discipline of anthropology expanded in the United States following World War II, tracing the expansion of American business into international markets. American anthropology already had shown tendencies to resist business influences, and its members' encounters with multinational corporations abroad were not always framed in mutually beneficial terms; the companies' actions sometimes threatened or harmed the well-being of peoples that anthropologists cared about (Sherry, 1983). It may have been overseas encounters that accelerated a process through which businesses were categorized as 'harm industries' even beyond the technical definition of that term (Benson and Kirsch, 2010).

Business management also tended to be cast in critical terms during mid-twentieth century, sometimes represented as proponents of schemes that would bring dire consequences to anthropological subjects (e.g., deskilling; Lamphere, 1979). A strong inflection toward Marxian theory in American anthropology after the 1960s, together with a Vietnam-era ethics code in the AAA that forbade proprietary research, created a chilly atmosphere for research or practice in/of/for businesses.

Nevertheless, several developments in the last quarter of the twentieth century brought anthropology together with business subjects in unexpected ways. A prominent impetus for change emerged from the discipline itself. One potent force was the 'crisis of representation' emerging from the critical and postmodernist movements, and the resulting diaspora of anthropologists into new venues, seeking experiences that responded to the crisis (e.g., avoidance of former colonies, entering nontraditional venues, experiments in methodology, and 'studying up'). Anthropologists learned from other fields and became hybridized, creating the institutional anthropologies (e.g., medical, legal, and educational). Some of these hybrids discovered career paths beyond the academy, finding that globally integrated forms of capitalism readily incorporate anthropological and ethnographic knowledge(s) and techniques (see Cefkin, 2009). At the same time, the discipline of anthropology became more inclusive, more involved in public interest and urgent anthropology, and less alienating of applied and practicing anthropology. Lines between academic and applied anthropology blurred as academia reached out toward partnerships with other types of institutions.

These changes have given rise to a suite of new subfields of inquiry and practice that are affiliated with the business domain, each situated somewhat differently with respect to institutions and literature, but all contributing to the knowledge of business from an anthropological perspective.

Major Areas of Inquiry and Practice

Ethnographically Informed Design

One of the first alternative field sites to emerge in the late 1970s and early 1980s as a destination for anthropologists seeking to 'study up' and experiment with new research subjects and methods was the high technology geographic center of Silicon Valley. Large corporate research venues such as Xerox's Palo Alto Research Center (PARC) provided locations where enterprising anthropologists could explore intriguing ideas such as the 'paperless office' while also observing the 'up' of management and corporate life in situ.

An early pioneer was Lucy Suchman, a graduate student in anthropology at Berkeley who drew upon Garfinkel's ethnomethodology to inspire a new approach to the study of human–computer interaction. She introduced videotaping and frame-by-frame analytic comparisons of human–human and human–machine interactions to reveal inadequacies of machine design at Xerox (Suchman, 1987). Xerox was sufficiently impressed with Suchman's work that the company changed the design of its copier machines and ultimately modified its product development process.

Suchman formed her own research group with Jeanette Blomberg, Julian Orr, and others at the Institute for Research on Learning. They became known around the world for their innovative integration of anthropology and ethnography into the product development stream of major corporations, opening up cognate areas for anthropology and ethnography in business (e.g., communities of practice, knowledge management).

In the late 1980s and 1990s, Suchman and her team collaborated with the Doblin Group of Chicago in a long-term project focused on the workplace of the future for Steelcase, known as the Workplace Project. One member of the team was Rick Robinson of Doblin, who went on to found the entrepreneurial consulting firm E-Lab (E-Lab was purchased by Sapient in 1999), a design-oriented boutique whose ethos integrated design and ethnographic methods (Squires and Byrne, 2002). The goal of the practice was to acquire nuanced, visually based, contextualized knowledge of the user's or consumer's behaviors, and to know both what the user is doing and why he or she is doing it. E-Lab introduced ethnography to the fields of product design and development, and competitors followed suit, initiating a process of hybridization between ethnography and design that has created the ethnographically informed design that may be found in corporate venues across North America, Europe, and Asia. This hybrid field grounds the Ethnographic Praxis in Industry Conference, which now incorporates many other disciplines.

Suchman's group at Xerox PARC, which initially enjoyed significant autonomy, was eventually pressured to focus on the company's priorities. When tensions between the research group and management could not be resolved to mutual satisfaction, the group elected to disband. Suchman's group at Xerox PARC was perhaps one of the most significant influences on business since the Human Relations School of the early twentieth century. On the other hand, an emphasis on ethnomethodology in design initiated at Xerox could have inadvertently exposed a vulnerability to the representation of ethnography as methodology per se, which now may be manifesting itself in the commodification of ethnography in the market.

Organizational Cultures

Management consultants did not invent the idea that organizations may resemble the cultures of societies, but they did discover that this concept was marketable if presented as a variable that their clients could influence, and they found ready markets in the West where Asian firms had unleashed competition in the last quarter of the twentieth century. Particularly appealing was the notion that organizational outcomes could be enhanced by a 'strong culture' that was consistent or integrated in its mission, goals, members' beliefs, values, and behavioral practices, and that such a culture could be achieved by an entrepreneur's intervention (Schein, 1985/1992). For example, an entrepreneur could imprint an organization with his or her values, and these might come to shape the learning experiences of a founding group and persist over time.

Academic anthropologists were challenged to respond to the idea of organizational culture in the 1980s, and it was in the technology-based firms of Silicon Valley, Route 128, and the heart of the manufacturing zone that anthropologists found an opportunity to establish authority on the subject. What emerged was a methodological and theoretical sophistication contributing to the literature on occupations and professions in complex institutions (e.g., Gregory, 1983; Dubinskas, 1988), correcting the early notion of 'strong culture' by demonstrating that organizational culture(s) in business firms may differentiate as well as integrate.

The critical and postmodern movements of the late twentieth century brought many different individual voices and points of view into anthropologists' studies of organizations (e.g., Kondo, 1990; Hamada, 1995). Fragmentation is one way to describe the perspective on organizational cultures that emerged from this period. However, it is also possible to understand this work as part of a transition toward a more nuanced view of organizations as institutional actors within larger social, economic, and technological contexts, with this phase acknowledging the agency of individuals and organizational subgroups or quasi-groups and the potency of their interests in organizations. A fuller transition to neoinstitutionalism requires not only multivocalism but also historical and regional/global contextualization of organizations, so they are no longer viewed as isolates. For example, if a founder can imprint an organization (i.e., infuse it with values beyond its technical purpose, as Selznick theorized), then one might expect to find such organizations in regions where founders' influences continue over long periods of time (e.g., in Europe, where businesses often are family-run). And it is in Europe where ethnographic studies of organizational culture remain of interest today.

The Anthropology of Consumers, Marketing, and Advertising

Intensifying economic competition among global corporations in the late twentieth century was accompanied by a need to understand diverse consumers across the regions of the world. One consequence was that Cold War-era views of consumers as rational, information-driven, and utility-maximizing beings were modified to accommodate notions of the customer as contextually embedded, emotional subjects whose lives require interpretation. With this shift, the dominant positivist standards in consumer research were adjusted to assimilate naturalistic methods such as ethnography and ethnomethodology.

An important development in anthropology that supported these changes was the acknowledgment of consumption (Miller, 1995) as the local idiom through which cultural forms express their creativity and diversity. This opened the door to more creative exchanges between anthropology and marketing. An interdisciplinary theory of consumer culture (CCT) emerged, which is a family of theoretical perspectives that define the relationships among consumer behavior, cultural meaning, and the market. Anthropologists have been central to the development of CCT. From this perspective, material goods and services are involved in the definition of the self and a coherent sense of identity, even if fragmented. Consumption is seen as integral to identity in the West, where individualism is prevalent, and much about the individual is ambiguous at birth. Consumers are not passive adopters, but culturally constituted actors who express sociocultural meanings, and act as creative agents, resisting, mutilating, and reconfiguring what they find in the market as patterns for action and reinterpretation.

These ideas were prevalent in business schools and marketing through the turn of the twenty-first century, as an alternative to the dominant scientific norm. Even so, some anthropologists were uneasy about marketing and advertising practices that might bring harm to consumers and/or harmful consequences to society over the long term. Marketing and advertising may not always be comfortable places for anthropologists, so close to the mechanisms by which consumption is produced in a capitalist society (Malefyt and Moeran, 2003).

Within the first decade of the new century, doubts began to emerge about the role of anthropology in consumer research

and marketing. Ethnography has remained popular in these fields, but the meaning of the method is contested. Ever more practitioners claim to offer 'ethnography' to businesses as a service, with ever widening conceptions of the craft. Some boutique firms offer consumer research services but do not believe anthropology is necessary or even desirable; perhaps anthropology is only a 'middleman' introducing extraneous interpretations, while the corporate clients desire the unvarnished consumer (Malefyt, 2009). Ethnography may mean no more than videotaping consumer behavior and/or using various brands of digital technology to facilitate customer self-documentation and transmission of recordings direct to corporate clients. Such 'technoethnography' could be faster and less expensive than anthropological ethnography or rapid ethnographic methods. As in an earlier era of industrialization, the skilled work of ethnography could be in jeopardy unless anthropologists are able to explain and/or enact its strategic significance with clients.

The Anthropology of Finance

Twenty-first-century society has fostered a new subfield of business anthropology, focused on the professional cultures and gendered landscapes of finance and their intersection with the larger society. The modern financial crisis was first reported in the *Financial Times* by anthropologist Gillian Tett (2009) who understood the role of exotic banking practices (e.g., derivatives). Recent ethnographies of financial professionals reunite anthropology with a century-long tradition of inquiry focused on social institutions – the shared rules, norms, and cognitive constructs that enable both continuity and resilience in complex societies. Anthropologists have been creative in approaches to understanding financial institutions, apprenticing to future traders as a means to discover the rules of market discipline in a technology-mediated environment (Caitlin Zaloom, 2006); learning how disbanded members of a Japanese securities firm apply the metaphor of arbitrage trading (which rests upon a very specific construction of time) to areas of life that fall outside the market (Hirokazu Miyazaki, 2003); accessing investment bankers' experiences with job loss to learn how they reproduce corporate downsizing as a result of their own embodied experiences (Karen Ho, 2009); and developing innovative methodologies to study the first generation of women on Wall Street and their navigation of the masculine world of finance through networking organizations dedicated to women's advancement (Melissa Fisher, 2012). Such studies have placed the anthropology of finance at the forefront of understanding the relationships among professional knowledge and expertise, technological change, and social institutions, and continue to suggest the potency of anthropology as a contributor to research on business and society.

See also: Anthropology at Home; Brokers and Brokerage, Anthropology of; Development: Social-Anthropological Aspects; Economic Anthropology; Exchange in Anthropology; Expertise, Anthropological Approaches To; Finance, Anthropology of; Globalization, Anthropology of; Industrial Societies; Money: Anthropological Aspects; Organizational Culture, Anthropology of; Political Economy in Anthropology; Property: Anthropological Aspects; Work, Anthropology of.

Bibliography

Baba, M.L., Hill, C.E., 2006. What's in the name "applied anthropology"? An encounter with global practice. In: Baba, M.L., Hill, C.E. (Eds.), The Globalization of Anthropology. National Association for the Practice of Anthropology #3. American Anthropological Association, Washington, DC, pp. 176–207.

Benson, P., Kirsch, S., 2010. Capitalism and the politics of resignation. Current Anthropology 51 (4), 459–486.

Bulmer, M., Bulmer, J., 1981. Philanthropy and social science in the 1920s: Beardsley Ruml and the Laura Spelman Rockefeller Memorial, 1922–29. Minerva 19, 347–407.

Burawoy, M., 1979. The anthropology of industrial work. Annual Review of Anthropology 8, 231–266.

Cefkin, M. (Ed.), 2009. Ethnography and the Corporate Encounter: Reflections on Research in and of Corporations. Studies in Applied Anthropology, vol. 5. Bergham Books, New York.

Dubinskas, F. (Ed.), 1988. Making Time: Ethnographies of High-Technology Organizations. Temple University Press, Philadelphia, PA.

Fisher, M., 2012. Wall Street Women. Duke University Press, Durham, NC.

Gillespie, R., 1991. Manufacturing Knowledge: A History of the Hawthorne Experiments. Cambridge University Press, New York.

Goody, J., 1995. The Expansive Moment: Anthropology in Britain and Africa: 1918–1970. Cambridge University Press, Cambridge.

Gregory, K., 1983. Native view paradigms: multiple cultures and culture conflicts in organizations. Administrative Science Quarterly 28, 359–376.

Hamada, T., 1995. Inventing cultural others in organizations. Journal of Applied Behavioral Science 31 (2), 162–185.

Ho, K., 2009. Liquidated: An Ethnography of Wall Street. Duke University Press, Durham, NC.

Kondo, D., 1990. Crafting Selves: Power, Gender and Discourses of Identity in a Japanese Workplace. University of Chicago Press, Chicago, IL.

Lamphere, L., 1979. Fighting the piece rate system: new dimensions of an old struggle in the apparel industry. In: Zimbalist, A.S. (Ed.), Case Studies on the Labor Process. Monthly Review Press, New York, pp. 257–276.

Malefyt, T. de W., 2009. Understanding the rise of consumer ethnography: branding techno-methodologies in the new economy. American Anthropologist 111 (2), 201–210.

Malefyt, T. de W., Moeran, B. (Eds.), 2003. Advertising Cultures. Berg, New York.

Miller, D., 1995. Acknowledging Consumption: A Review of New Studies. Routledge, London.

Mills, D., 2002. British anthropology at the end of empire: the rise and fall of the Colonial Social Science Research Council, 1944–1962. Revue d'Histoire des Sciences Humaines 6, 161–188.

Miyazaki, H., 2003. The temporalities of the market. American Anthropologist 105 (2), 255–265.

Partridge, W.L., Eddy, E.M., 1978. The development of applied anthropology in America. In: Eddy, E.M., Partridge, W.L. (Eds.), Applied Anthropology in America. Columbia University Press, New York, pp. 3–45.

Roethlisberger, F.J., Dickson, J.W., 1939. Management and the Worker. Harvard University Press, Cambridge, MA.

Schwartzman, H., 1993. Ethnography in Organizations. Sage Publications, Newbury Park, CA.

Schein, E., 1985/1992. Organizational Culture and Leadership. Jossey-Bass, San Francisco, CA.

Sherry, J.F. Jr., 1983. Business in anthropological perspectives. Florida Journal of Anthropology 8 (2), 15–36.

Squires, S., Byrne, B. (Eds.), 2002. Creating Breakthrough Ideas: The Collaboration of Anthropologists and Designers in the Product Development Industry. Bergin and Garvey, Westport, CT.

Stocking, G., 1995. After Tylor: British Social Anthropology, 1888–1951. University of Wisconsin Press, Madison, WI.

Suchman, L., 1987. Plans and Situated Actions: The Problem of Human-machine Communication. Cambridge University Press, Cambridge.

Tett, G., 2009. Fool's Gold: How the Bold Dream of a Small Tribe at J.P. Morgan Was Corrupted by Wall Street Greed and Unleashed a Catastrophe. Free Press, New York.

Warner, W.L., Low, J.O., 1947. The Social System of the Modern Factory. The Strike: A Social Analysis. Yankee City Series, vol. IV. Yale University Press, New Haven, CT.

Warner, W.L., Lunt, P.S., 1941. The Social Life of a Modern Community. In: Yankee City Series, vol. I. Yale University Press, New Haven, CT.

Zaloom, C., 2006. Out of the Pits: Traders and Technology from Chicago to London. The University of Chicago Press, Chicago, IL.

Business Education

Bob Garvey, York St John University, Business School, York, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores business education from its rather primitive beginnings to the complex topic it is today. It offers critical perspectives on business education and aims to not only inform but provoke thought and discussion. It starts with a brief historical overview of business education and moves to explore the links between business strategy and education. It considers the relevance and importance of technical versus personal development and the role of innovation and creativity in modern business settings. It then considers the questions: 'What type of people does modern business require?' and 'How are they going to be developed?' These are discussed in the context of leadership and leadership development. Finally, it raises issues about the future by considering the behaviors and practices of business and offers a perspective on the largely underdeveloped concept of 'knowledge productivity' within the construct of a corporate curriculum.

Introduction

This article considers the development of business education from its rather primitive beginnings to the relatively complex topic it is today. It offers critical perspectives on business education as it aims to not only inform but provoke thought and discussion – there are few right answers here!

Beginning with a brief historical overview of business education, it moves on to explore the links between business strategy and education. Then it considers the relevance and importance of technical versus personal development and then the role of innovation and creativity in modern business settings. It considers the questions: 'What type of people does modern business require?' and 'How are they going to be developed?' These are discussed in the context of leadership and leadership development.

Finally, it raises issues about the future in relation to the changing social circumstances brought about by the global economic crisis. It discusses the behaviors and practices of business and offers a perspective on the largely underdeveloped concept of 'knowledge productivity' within the construct of a corporate curriculum (Kessels, 2002).

Brief History of Business Education

> It is groups, not individuals, that adapt to their environments and this is the basis for the peculiar genius of the human species. We are a successful species because we cheat; we tell each other the answers
> Emler and Heather, 1980.

The above quotation, in a nutshell, says all there is to say about education! If we do not support each other to learn, we do not sustain (*see* Sustainable Lifestyles; Sustainable Practices; Sustainable Production and Consumption) or progress. From ancient Greek times, where the epic poem, The Odyssey, was written with a story line about the education and development of the young prince Telemachus through to medieval guilds whose specific purpose was for the development of skills (*see* Education: Skills Training), to modern times where education has become even more important as a driver to economic progress, business education has always had a function.

Business Education and Strategy

Why would any business be interested in educating its workforce?

One response may be found in the imperatives of business strategy (*see* Education, Economics of). Most businesses within capitalist economies are driven by profit. This may translate into developing market share or seeking that elusive 'competitive advantage.'

In recent history, competitive advantage seems to be sought through a range of business drivers. The late-1980s and early-1990s organizations sought competitive advantage through the superior quality of their products or services. This strategy led to great improvements in productivity, efficiency, and technological innovation and a demand for a highly skilled technical workforce able to be flexible and adaptable, to be innovative and self-sufficient.

A business faced the choice of developing its own people or buying in the skills needed. The second option, while attractive, was more difficult to achieve as the competition for skilled workers increased. This in turn put pressure on education systems. Paradoxically, some skilled workers started to become deskilled as a result of technology, as their traditional skills were replaced by automation.

In the UK during this period, the public sector also saw many changes. A restructuring of the National Health Service saw the notion of competition becoming a driver for efficiencies and cost savings, and the long tradition of cooperation within the Service was put under strain. The attitudes, skills, and knowledge needed to manage health care started to change, and this meant that large numbers of doctors sought business education.

Competitive advantage based on quality alone became difficult to sustain. The 'quality' race led to organizations discovering that much less divided them, and high-quality products with high-quality service at reduced cost became the entry point for business (Hamel and Prahalad, 1989, 1991).

There followed a search for new differentiating strategies. Businesses developed new customer-led strategies that required not only quality products and service but value for money and after-sales care. Slogans began to appear in organizational documents such as 'people mean business,' 'we are in it for

the long haul,' 'people are our most important asset,' and the concepts of strategic human resource development and team working evolved. This created 'people-led' strategies with organizations seeking competitive advantage through their people.

In business education terms, perhaps the most significant publication to appear during this period was Senge's *The Fifth Discipline: The Art and Practice of the Learning Organization* (1992). This was reviewed in 1997 in the Harvard Business Review as one of the seminal business works in the last 75 years – business education was on the map!

Research in this period, for example, Huselid (1995) and Patterson et al. (1998), explored the relationship between people development and business performance. These studies suggested that the relationship was positive – developed people contribute to business progress.

Some saw these events as managerial attempts to 'gloss' over the flaws in 'disorganized capitalism' (Lash and Urry, 1987). Later, others took a more generous view, as Kessels (1996) put it:

> …a far more cogent argument would be that organizations have a direct stake in the personal enrichment of employees because excellence on the job requires employees who are comfortable with their work and who have strong and stable personalities. Personal enrichment is thus less an employee privilege than a condition for good performance
>
> p. 4; *see* Metacognitive Development: Educational Implications

As we moved into the new millennium, businesses still searched for competitive advantage. The drivers of innovation and deregulation came on the agenda. Business was using its lobbying power to achieve changes in legislation on both sides of the Atlantic and it became more global and less interested in national boundaries. New markets were sought; new, cheaper centers for manufacturing were established. New distribution channels were developed and computerization dominated all aspects of business. These searches raised moral questions for business and notions such as 'corporate social responsibility' and 'ethical' developed. The British economist Will Hutton's views on deregulation in 1997 were almost prophetic:

> […] the operation of the unchecked market has an inherent tendency to produce unreasonable inequality, economic instability and immense concentrations of private, unaccountable power. To protect itself, society has to have countervailing powers built into the operation of the market, otherwise it cannot deliver its promise. Instead it collapses into licence masquerading as liberty, spivvery dressed up as risk-taking and exploitation in the guise of efficiency and flexibility.
>
> Hutton, 1997, p. 4

Deregulation happened and, arguably at least, this led to the global financial crisis of 2008. Here, 'innovations' in traditional business practices and changes in fundamental economic assumptions created unsustainable rapid growth and considerable individual wealth. In wider society, deregulation and uncontrolled globalization brought us corporatized 'neofeudalism' (Shearing, 2001) with localized 'fiefdoms' where rules are enforced and power bases established rather like the Baron of medieval times. The controllers of the 'fiefdoms' called for deregulation and freedom at the same time as imposing greater regulation and surveillance of those they controlled (Saul, 1997). The sad result is 'paternal authority' where the rights of individuals are sidelined "under the guise of business ethics…" (Schwartz, 2000, p. 175). This is both a leadership and ethical development issue.

So where are we now?

Tyson (2012) points out that in the "US jobs data reveals economy is bouncing back strongly from recession," but there is a serious mismatch between the needs of employers for technologically based skills and those available in the labor market. She argues that the "technological change is accelerating, fuelling demand for more skills at a time when the workforce's educational attainment levels have plateaued."

In the UK, employers face similar technical skill shortages. While UK organizations are recruiting more young people, the recruitment of more specialist employees is becoming more difficult.

Managerial- and professional-level staff are the most difficult to retain, and the methods most commonly practiced to encourage retention are providing business educational opportunities (CIPD, 2012).

Looking globally, "in the fifth year after the outbreak of the global financial crisis, global growth has decelerated and unemployment has started to increase again, leaving an accumulated total of some 197 million people without a job in 2012" (International Labour Organization (ILO), 2013, p. 2). Echoing the comments made by Tyson (2012), the findings of the ILO (2013) and the CIPD (2012) reports are clear that the "new jobs that become available often require competences that the unemployed do not possess" (ILO, 2013, p. 2).

There is now a highly competitive international market in high-level leadership and technical skills. Those who fail to keep ahead of these developments will lose their competitive advantage.

There are long-term cultural shifts in attitude and expectation. The global labor market has meant that people take a more strategic view of their career development and invest personal capital in their own business education. It is not surprising then that more and more employees negotiate opportunities for further training and professional development.

Organizations who respond to these expectations are likely to attract highly qualified staff and those that do not will disadvantage themselves.

It is vital that managers in organizations think strategically about the longer-term development of their human capital. Successful businesses will be those that take business education seriously.

To summarize, attitudes, skills, and knowledge change over time in relation to political, economic, and social changes. The role and purpose of business education changes with these shifts in thinking. Education remains an essential element in the political, social, and economic development – to quote a poster seen in a business recently: "If you think education is expensive, try ignorance!"

The Technical versus the Personal

Garvey (1997) noted that 'technical' skills and knowledge development dominate the training agenda. He also noted

that 'personal' skills development is largely underrepresented. However, he argues that 'personal' skills are essential management abilities that enable managers to lead teams effectively. It is here also that issues of power, status, and ethical problems can appear. So, while the technical development of people in business is essential, it is also important to develop managerial and leadership abilities. These may include communication skills, skills of influencing and persuading, time and project management, and in recent times, ethical awareness, coaching and mentoring, and critical and creative thinking.

Innovation and Creativity

The expectation of change is pervasive. The future is wide open, and new attitudes, skills, and knowledge are required. The new challenges on people require a new paradigm for business education.

As Garvey and Williamson (2002, p. 23) state, "It is rarely possible to predict from where the latest innovation will come. This means that the work organization needs to build potential or strategic capability within its people so that they are ready to adapt, react, respond and innovate (*see* Innovation and Employment). Chance favors the prepared mind and the preparation comes through people being open to rich learning opportunities so readily available in the work place…"

So, what type of person is needed in today's business world?

The Type of Person

A simple Web search for jobs anywhere in the world generates the same type of descriptions. Below is a composite list of the key attributes cited by potential employers in large, small, private, and public sector organizations around the world:

- Creative, analytical, problem solving, and technical skills (*see* Creativity Training)
- Excellent people skills, self-awareness
- Able to produce results to meet both customer and business needs
- Initiative and ability to learn and innovate
- Leadership abilities
- Personal values
- Honesty, positive self-esteem, and good personal presentation
- Capabilities to create, develop, and support a harmonious workplace
- Able to work with change
- Strong and stable personalities (*see* Personality Development and Temperament)
- Leadership

A challenging list indeed! How might these attributes be acquired, developed, and sustained?

One answer may be found in the quality of leadership within a business. Technical abilities remain essential for business progress, but the quality of leadership within a business determines its strategic progress.

Leadership development (*see* Leadership Education: Theory and Practice) is an important preoccupation of modern business. Many organizations develop competency frameworks against which they can assess readiness for leadership positions, design development interventions, develop talent, and plan succession routes. However, the issue of competencies within management development is controversial.

The argument against competencies is that they can lead to a 'training gap' orientation focusing on what people cannot do rather than on what they can.

Bolden and Gosling (2006, p. 148) suggest that competencies, being reductionist in nature, cannot prepare people for the complexities of a job role; they are not sensitive to specific situations, tasks, or individuals; they represent a view of past performance rather than acting as a predictor of future behavior; they tend to exclude subtle qualities, interactions, and situational factors, and they create a limited and mechanistic approach to learning. Garvey et al. (2009, p. 191) note "…when professionals are highly anxious – perhaps because of detailed surveillance and under strong resource pressure, then the delivery of competencies can degrade."

Taken together, these arguments suggest that a competency framework cannot prepare learners for either an uncertain future or changing circumstances – the very condition for which the above list of personal attributes seems to be designed! The implications here again for business education are considerable.

This seeming mismatch between the need for certain attitudes, skills, and behaviors, and the approach to acquiring them may be explained in the concept of 'mental models' (Senge, 1990).

Arguably, an overarching mental model in management is a positivistic orientation (Johnson and Duberley, 2000, p. 12). This mental model is based on the notion that 'good practice' in management is about cause-and-effect decisions, taking action, establishing objectives, and measuring their outcome. This discourse translates into organizational policies as the rational, pragmatic manager (Garvey and Williamson, 2002) attempts to control the system.

However, Barnett (1994, p. 37) states: "Society is more rational, but it is a rationality of a limited kind" and "genuinely interactive and collaborative forms of reasoning" are in danger of being driven out by this technical reasoning. When applied to business education, this approach tends to favor the prespecification of 'learning outcomes' or goals.

Prespecification for leadership development, for example, is common practice. While this does offer the possibility of accountability, measurement, and quality control to determine value for money, it also leads to reductionist thinking, creates unquestioning compliance, and stifles innovation and flexibility (Bernstein, 1971; Buckingham, 2001) – it limits 'collaborative reasoning.'

This mode of rationality cannot be adequate to guide learners if they are to be capable of flexibility, innovation, creativity, and improvisation (Habermas, 1974; Jessup, 1991; Barnett, 1994; Buckingham, 2001) – the widely agreed qualities required of the leaders of the future (Sosik and Megerian, 1999; Harris, 2004; Bolden and Gosling, 2006).

According to many writers (Polanyi, 1958; Argyris and Schon, 1981; Vygotsky, 1985; Bruner, 1990; Lave and Wenger, 1991) (*see* Vygotsky's Theory of Human Development and New Approaches to Education), learning is fundamentally

a social activity, which happens in a social context through dialogue. Therefore, leaders create the social context in which learning can take place – or not! As Daloz (1986, p. 68) states: "To understand human development, we must understand the environment's part, how it confirms us, contradicts us, and provides continuity."

The environment within a business makes it more or less possible for people to learn, develop, and perform. Traditionally, learning in business took place in classrooms with the education provided by an external agent. While there is still a place for this type of educational model and, indeed, many advantages, such as the introduction of new ideas and approaches, the potential for external critical comment, there is also a need for a more 'situated' (Lave and Wenger, 1991) or 'community of discovery' (Coffield and Williamson, 2011) approach.

'Situated learning' (*see* Situated Learning (Learning *In situ*)) is where "…learners inevitably participate in communities of practitioners and that the mastery of knowledge and skill requires newcomers to move toward full participation in the sociocultural processes of a community…" (Lave and Wenger, 1991, p. 29). It is an inevitable consequence of being at work, of doing the job!

The notion of 'situated learning' is not without its problems. It assumes that an individual learns through participating in a 'community of practice,' but groups can create power bases with those who are 'in' and those who are 'out.' This behavior is facilitated through the cultural rituals, language, assumptions, symbolic displays, and specific narratives at play within a particular community. With the idea of a 'community of practice' comes the risk of restricted openness or the limited introduction of new ideas and thus constrained innovation.

Community of Discovery – An Alternative View

Sir Isaac Newton observed, "I have seen further because I stood on the shoulders of giants." Most human discoveries have been collective social achievements, and here is the basis for the concept of a 'community of discovery.' Through inter- and multidisciplinary working, there is the potential for new thoughts to emerge as different perspectives engage with each other. If there is a sense among members of a group that new ideas are valued, acknowledged, and debated critically and openly, a 'community of discovery' becomes possible and new solutions to old problems may emerge.

Beyond the Organization

At a local level, organizations create environments to facilitate learning, but governments play their part by supporting business education. In recent times, the European Union memorandum on lifelong learning (2000), the G8 Group Cologne memorandum on lifelong learning (2000), many of Europe's national governments, Japan, Singapore, Australasia, and Taiwan are all committed to a public policy agenda that promotes lifelong learning in the workplace (see Field, 2000) for global competitiveness.

Business education is a global business!

Knowledge Productivity and the Corporate Curriculum

According to Garvey and Williamson (2002), all economies are knowledge economies and they always have been. Economic, social, and political progress has always been as a result of the application of new ideas and discoveries. In recent years, the notion of a 'knowledge economy' has acquired greater significance. Here, organizations need to develop their strategic capability and add value to their goods and services by the application of knowledge (*see* Intelligence, Prior Knowledge, and Learning). Kessels (2002) refers to this as 'knowledge productivity.'

We have Kessels to thank for developing the concept of the 'corporate curriculum.'

A curriculum is a program or course of study. All organizations have a curriculum whether planned, emergent, or chaotic! Kessels (1996) suggests that the idea of 'curriculum' is central to business education and is also is central to any credible account of organizational learning culture.

Bernstein (1971) suggested that to design any curriculum, it was important to consider the following questions:

1. What is valid knowledge in a specific context?
2. What is a valid way to deliver this knowledge?
3. How might this learning be evaluated?
4. How might this knowledge be realized in practice?

He suggested that there are two main elements to any curriculum – the 'formal' and the 'informal.'

In a 'formal' curriculum, the content is specific and prespecified. Knowledge is categorized by subject matter boundaries, and evaluation is dominant and therefore easier to deliver and assess. The content is largely controlled by the educator or teacher. The 'formal' curriculum tends to result in lower initiative, lower innovation, and lower creativity for the 'learner.'

In contrast, the 'informal' curriculum has less specific and more holistic content with loose subject matter boundaries. Rather than prespecifying the objectives of the learning, the 'informal' curriculum is driven by the prevailing circumstances or needs of the context – it is about relevance. Evaluation and assessment are related to the need that was being addressed and linked to utility and usefulness. This aligns it to Knowles' (1984) notion of andragogic adult learning, where the curriculum is

- based on the adult's need to know;
- experience-based;
- involving, shared, and participative;
- relevant and applied;
- problem-centered rather than content-oriented; and
- driven by the individual's internal motivations rather than external motivators (after Knowles, 1984).

The 'informal' curriculum fosters greater initiative and creativity in the learner.

The notion of curricula is deeply significant and has implications for organizations who take business education seriously. The 'formal' approach is advocated in many textbooks on the subject of training and development. The 'formal' position is often accompanied by the statement that business education needs to be strategically aligned, despite the fact that it is unlikely to deliver the innovation, flexibility, and

creativity needed for the knowledge economy (Broad and Newstrom, 1992).

However, the knowledge economy is dynamic – it is highly changeable and organizations need to be quick to anticipate and adapt. Therefore, working with a balance between the 'formal' and the 'informal' is more likely to build that all-important strategic capability for a capacity and potential for rapid change. A key question then is, 'What is best delivered formally and what is best delivered informally?'

The challenge is for managers to recognize that the workplace itself is a rich learning environment, which can provide the necessary stimulus and support for learning and knowledge generation. There is much evidence that this 'situated approach' (Lave and Wenger, 1991) offers a greater depth, understanding, and participation in learning for knowledge productivity.

Kessels (1996) embeds the metaphor of the 'rich landscape' of learning in his 'corporate curriculum' framework in the form of seven elements or 'learning functions':

- Expertise related to the core competence
- Problem solving
- Reflection and generation of new knowledge
- Communication
- Self-knowledge, metacognitions (*see* Teaching for Metacognition)
- Peace and stability
- Creative turmoil

The seven elements in the 'rich landscape' contain both 'formal' and 'informal' elements, and much depends on the type and purpose of a business as to the nature of the balance of the elements. In the 'rich landscape,' learning occurs through an active engagement and participation that encourages creativity, lateral thinking, and flexibility – the much-needed abilities of an employee of the future and the potential answer to the business education challenge in the twenty-first century.

See also: Creativity Training; Education, Economics of; Education: Skills Training; Innovation and Employment; Intelligence, Prior Knowledge, and Learning; Leadership Education: Theory and Practice; Metacognitive Development: Educational Implications; Personality Development and Temperament; Situated Learning (Learning In Situ); Sustainable Lifestyles; Sustainable Practices; Sustainable Production and Consumption; Teaching for Metacognition; Vygotsky's Theory of Human Development and New Approaches to Education.

Bibliography

Argyris, C., Schon, D.A., 1981. Organisational Learning. Addison-Wesley, Reading, MA.
Barnett, R., 1994. The Limits of Competence: Knowledge, Higher Education, and Society. SRHE/Open University Press, Buckingham.
Bernstein, B., 1971. On the classification and framing of educational knowledge. In: Young, M.F.D. (Ed.), Knowledge and Control: New Directions for the Sociology of Education. Open University, Collier–MacMillan, London.
Bolden, R., Gosling, J., 2006. Leadership competencies: time to change the tune. Leadership 2 (2), 147–163.
Broad, M.L., Newstrom, J.W., 1992. Transfer of Training. Action-Packed Strategies to Ensure High Payoff from Training Investments Reading. Mass: Addison-Wesley Publishing Company, Inc.
Bruner, J., 1990. Acts of Meaning. Harvard University Press, USA.
Buckingham, M., December 3, 2001. 'Don't waste time and money. Gallup Management Journal. Available online at: http://gmj.gallup.com/content/default/asp?ci=259&pg=1.
Coffield, F., Williamson, B., 2011. From Exam Factories to Communities of Discovery: The Democratic Route. The Institute of Education, London, UK.
CIPD, 2012. A Barometer of HR Trends and Prospects 2013. http://www.cipd.co.uk/hr-resources/survey-reports/cipd-surveys-overview-hr-trends-prospects-2012.aspx (accessed 02.01.13.).
Daloz, L.A., 1986. Effective Teaching and Mentoring. Jossey Bass, San Francisco.
Emler, N., Heather, N., 1980. Intelligence: an ideological bias of conventional psychology. In: Salmon, P. (Ed.), Coming to Know. Routledge and Kegan Paul, London.
Field, 2000. Lifelong Learning and the New Educational Order. Trentham Books. London.
Garvey, B., 1997. What's in It for Me?. In: The Learning Organization, vol. 4 MCB, Leeds, pp. 3–10.
Garvey, B., Williamson, B., 2002. Beyond Knowledge Management: Dialogue, Creativity and the Corporate Curriculum. Pearson Education, Harlow, UK.
Garvey, B., Stokes, P., Megginson, D., 2009. Coaching and Mentoring Theory and Practice. Sage, UK.
Habermas, J., 1974. Theory and Practice, Heinemann, UK (first published in 1971 as Theorie und Praxis).
Hamel, G., Prahalad, C.K., May–June 1989. Strategic intent. Harvard Business Review 1–14.
Hamel, G., Prahalad, C.K., July–August 1991. Corporate imagination and expeditionary marketing. Harvard Business Review 69 (4), 81–92.
Harris, P.R., 2004. European leadership in cultural synergy. European Business Review 16 (4), 358–380.
Huselid, M., 1995. The impact of human resource management practices on turnover, productivity, and corporate financial performance. Academy of Management Journal 38, 645.
Hutton, W., 1997. The State to Come. Vintage, London.
International Labour Organization, 2013. Global Employment Trends: Recovering from a Second Jobs Dip. ILO, Geneva, Switzerland.
Jessup, G., 1991. Outcomes: NVQs and the Emerging Model of Education and Training. Falmer, Oxford.
Johnson, P., Duberley, J., 2000. Understanding Management Research. Sage, London.
Kessels, J.W.M., 1996. Corporate Education: The Ambivalent Perspective of Knowledge Productivity. Centre for Education and Instruction, Leiden University.
Kessels, J.W.M., 2002. You cannot be smart against your will. In: Garvey, B., Williamson, B. (Eds.), Beyond Knowledge Management: Dialogue, Creativity and the Corporate Curriculum. Pearson Education, Harlow.
Knowles, M., 1984. Andragogy in Action. Jossey-Bass, San Francisco.
Lash, S., Urry, J., 1987. The End of Organised Capitalism Cambridge. Polity Press, UK.
Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, Cambridge.
Patterson, G., West, M.A., Lawthorn, R., Nickells, S., 1998. Impact of People Practices on Business Performance. IPD, London, UK.
Polanyi, M., 1958. Personal Knowledge: Towards a Post-Critical Philosophy. Routledge and Kegan Paul, London.
Saul, J.R., 1997. The Unconscious Civilization. Penguin, Ringwood.
Schwartz, M., 2000. Why ethical codes constitute an unconscious regression. Journal of Business Ethics 23, 173–184.
Senge, P.M., 1992. The Fifth Discipline, Century Business, London, UK.
Shearing, C., 2001. Punishment and the changing face of the governance. Punishment & Society 3 (2), 203–220.
Sosik, J.J., Megerian, L.E., 1999. Understanding leader emotional intelligence and performance: the role of self-other agreement on transformational leadership perceptions. Group & Organization Management 24 (3), 367–390.
Tyson, L., 2012. US Jobs Data Reveals Economy Is Bouncing Back Strongly from Recession. The Guardian Online: 17.10.12. http://www.guardian.co.uk/business/2012/oct/17/us-employment-data-recession-recovery (accessed 30.12.12.).
Vygotsky, L.S., 1985. In: Wertsch, J.V. (Ed.), Culture, Communication and Cognition: Vygotskian Perspectives. Cambridge University Press, Cambridge, London and NY.

Business Ethics

Christoph Luetge, Technische Universität München, Peter Löscher Chair of Business Ethics, München, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article gives a brief overview of the history of business ethics starting with Aristotle. Different contemporary approaches to business ethics are surveyed, with an emphasis on continental European discussions. The article distinguishes between and focuses on the deliberative, the contractarian, and the contractualist approach to business ethics. The second part presents a short list of current topics and questions in business ethics.

Introduction

Business ethics is a discipline of ethics that is concerned with moral and ethical problems, norms, and values that arise in the area of business. As such, it requires interdisciplinary cooperation with other disciplines, in particular economics and the social sciences.

Business Ethics and Economic Ethics

In some parts of the non-English speaking world (e.g., Germany), business ethics is distinguished from economic ethics, in the same way that the discipline of economics is distinguished from business administration and management (cf. Luetge, 2012a). Business ethics, in this sense, would focus on ethical problems within and concerning mainly corporations and related organizations, whereas economic ethics would concentrate on larger scale ethical problems of entire economies. While this distinction can be useful in a number of contexts, it is, however, not common in most international debates. Therefore, this article uses the term 'business ethics' for both ethical enterprises.

History of Business Ethics

While the term business ethics only started to be used widely in the 1970s, the occupation with and the analysis of ethical problems in the area of business reaches back to Antiquity. The Greek philosopher Aristotle (384–322 BC) did not distinguish between ethics and economics as separate disciplines, but regarded them both as parts of 'praxis,' i.e., oriented toward action as their final goal (as opposed to theoretical disciplines concerned with truth, and 'poiesis' concerned with producing things) (cf. Aristotle, 2003). This mirrored the situation in the Greek city state (polis) in which ethical, economic, and political aspects were closely intertwined. Citizens could (to a much greater degree than in modern societies) directly observe and monitor the behavior of others. This situation was typical of premodern societies that can be characterized by face-to-face structures and interactions. Ethics, in this social context, was mostly concerned with recommending guidelines for individual behavior: the virtues. These virtues, in the premodern world, served directly both the interest of their bearers and the interest of the society as a whole.

In the premodern situation, the ethical challenge arose in particular due to the problem of justly sharing the fruits of a society that were mostly fixed in total: Premodern societies did not experience substantial and stable growth over longer periods of time (the Roman Empire being a possible exception). Therefore, individuals could only gain significantly at the expense of others. The result is a 'zero-sum society' (cf. McCloskey, 2006), which had significant implications for ethics. The successful fifteenth-century Florentine merchant Giovanni Rucellai stated it in this way: "by being rich, I make others (which I might not even know) poor" (Rucellai, 1772, written about 1450).

In order to remain stable, a zero-sum society needed an appropriate kind of ethics, i.e., an ethics that curbed excessive profits, greed – what Aristotle (2003) called 'pleonexia.' An ethics of temperance was the result, which called for the citizens to remain moderate, to be content with moderate gains and a 'just' price, or – in former times – to refrain from lending money at interest: Usury was condemned by a long tradition of theologians and philosophers ranging from Aristotle, the Bible, the Qur'an, Thomas Aquinas, and Martin Luther up to modern representatives (like S. Gesell).

The ethical view on the economy only started to change when the beginnings of significant economic growth became visible in the eighteenth century. Authors like B. Mandeville (1670–1733) began to reinterpret the role of self-interest, in particular by stating that actions motivated by self-interest have unintended consequences that may be beneficial for other citizens (cf. Mandeville, 1714/1924). The first systematic approach to the role of self-interest in modern economies was consequentially undertaken by Adam Smith (1723–90) in his two books *The Theory of Moral Sentiments* (1759/1976) and the *Wealth of Nations* (1776/1976). Both works are sometimes seen as being in opposition with one another, as the *Theory* recognizes the workings of sympathetic feelings in communities, while the *Wealth* focuses on the workings of self-interest in economies. However, both works do not necessarily contradict each other. Rather, they exhibit an important background problem of business ethics: the premodern vs. the modern society. Smith's *Theory* is concerned with how morality (sympathy) works in the premodern society, in small group settings. By contrast, the *Wealth* is concerned with the modern economy and society, where feelings of sympathy from premodern contexts cannot play a direct role in governing economies and societies. Rather, the

feelings can only become indirectly involved, via the structures, rules, and institutions of economics: This is the meaning of Adam Smith's doctrine, according to which mutually beneficent results of economic transactions cannot be derived directly from (in particular, 'moral') motivations of actors, but from their self-interest, which gets channeled in beneficent ways by the mechanisms of the economy: "It is not from the benevolence of the butcher the brewer, or the baker that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity, but to their self-love, and never talk to them of our own necessities, but of their advantages" (Smith, 1776/1976). This doctrine is highly relevant to business ethics in modern societies.

The opposite reaction to modern societies can be seen in the works of J.-J. Rousseau (1712–78), who espoused the idea of a static economy (Rousseau, 1755/1964), and K. Marx (1818–83). Marx actually welcomed the benefits capitalism brought to the workers, but he regarded it as a mere transitional phase on the way to the ultimate end of a communist society (cf. Marx and Engels, 1959).

The next important step was Max Weber's work on the supposed Protestant origins of capitalism (Weber, 2005, Original 1904/1905 and Weber, 2001, Original 1915–19), which contributed descriptively to business ethics. Weber's work focuses on the relation between religious convictions and the economic performance of a society.

Another historical development that is connected to Weber's ideas is the ideal of the 'honest businessman,' an ideal that has its European origins in both the Italian Renaissance as well as in the heyday of the Hanseatic League. This ideal consisted in ethical ideas about being honest and trustworthy, keeping promises, contributing to the community, and promoting not only cultural affairs but also practical abilities like speaking several languages and having organizational talent. The honest businessman is an ideal that originates from the practice of business, not from philosophical or economic theory. As such, it has been – and sometimes still is – present in Europe but evolved (in somewhat different ways) in other parts of the world as well, for example, in Japan (cf. Shibusawa, 1916).

Before dealing with the contemporary discussion of business ethics, its theological origins must be mentioned as well. In particular, both the Catholic and Protestant churches developed thoughts on business ethics (for other religious schools, see Luetge, 2013, Parts 11, 13, and 14). Several papal encyclicals had an important influence on ethical reflection in business ethics, such as 'Rerum Novarum' (1891, by Pope Leo XIII). Theologians like Heinrich Pesch (1854–1926) developed the doctrines of Catholic social teaching, which aimed at finding a third way between capitalism and socialism (Pesch, 1905ff.), and highlighted the importance of the principle of subsidiarity, according to which individuals were to resolve many social problems on their own, with the state stepping in only if individuals (or private law organizations) failed to resolve the problem. These ideas were developed further by others like Johannes Messner (1927) and Oswald von Nell-Breuning (1928).

The Protestant social teaching was developed by thinkers like Otto Baumgarten (1921) and Georg Wünsch (1927). Baumgarten aimed at combining capitalism with altruist motives in personal life. Wünsch, however, promoted a Christian socialism, which regarded capitalism only as a historical episode.

The contemporary debate on business ethics started in the United States during the 1970s, in particular as a reaction to scandals in business and the changed perception of the public. Before that, corporations had already used the term 'social responsibility' for different activities intended to be beneficial for society, albeit not in a systematic way. When business ethics entered the US scene, it made use of more systematic theoretical resources, namely, ethical theory and empirical case studies (cf. De George, 2006).

The Continental European discussion of business ethics, however, developed in quite a different way. It started, in the 1980s, with much more theoretically focused discussions between different schools, before it turned into a more empirically oriented debate as well. Although this had the disadvantage of being slightly more removed from the world of business originally, in the end a theoretical sophistication was achieved that took (and still takes) into account many theories from other disciplines, especially economics.

The following section discusses the main contemporary approaches to business ethics, both from the Anglo-American and the continental tradition.

Approaches to Business Ethics

While there is disagreement in a number of matters among the different approaches to business ethics, there is also consensus at least in some points: Nearly all approaches now accept (1) that ethics and business, morality and profit, or solidarity and efficiency hold no contradictions in principle and (2) that the institutions of the society have an important bearing on its ethics (although the exact extent of this is controversial).

Among the most prominent approaches to business ethics are the following: deliberative or discourse approaches, contractarianism, and the stakeholder approach. These are discussed in turn.

The Deliberative or Discourse Approach

Several approaches in business ethics rely on a foundation in the discourse ethics of Jürgen Habermas and Karl-Otto Apel. Some of them are Horst Steinmann, Peter Ulrich, and Andreas Scherer (cf. Steinmann and Scherer, 1998; Ulrich, 2008 (Original 1987); Scherer et al., 2006; and Palazzo and Scherer, 2006). Its theoretical core is the idea of an ideal discourse situation in which economic actors deliberate rationally on the foundations and rules of their society. This ideal discourse situation relies on a number of presuppositions, in particular, that coercion and force are not used in the discourse, as they belong to the domain of strategic rationality, which Habermas and Apel keep strictly separate from communicative rationality. Communicative rationality is regarded as having higher dignity, as (so Habermas insists) only the rational argument is to rule, against economic and administrative power.

Similarly, Peter Ulrich's 'integrative economic ethics' (Ulrich, 2008) tries to integrate economic rationality into a much broader framework of communicative rationality. In practice, this would mean that corporations are to include aims other than profit making into their goals – not as a better means to

profit making itself, but rather as an additional, second way of doing business. The individual actor is to reflect the rationality of his or her economic activities – and, when in doubt, give primacy to ethical considerations. Ulrich does not regard economic constraints as crucial, let alone decisive (that would be 'economism' in his eyes), but argues that it is in the hands of citizens to conduct a discourse on the foundations of their societal and economic basis – which will be constitutive to the economy and its processes.

Discourse ethics has spawned other approaches to business ethics as well, in particular the deliberative approach by Scherer et al. (2006). Deliberative democracy (cf. for example, Elster, 1998; Cohen, 1989; Aaken et al., 2004) is a theoretical approach that regards all actors (including corporations) as being subject to legitimation in societal deliberation. Deliberation is to prevent the dominance of economic rationality – and in that is directed against economics in general. However, economic actors like corporations are seen as engaging in public discourse for their own long-term success. Scherer regards it as the moral and political duty of corporations to develop a moral integrity based on their public deliberation, and thus actively engage in societal discourse on issues like human rights, working conditions, or environmental protection.

The dualist approach inherent here persists in other approaches to business ethics: ethics and economics, morality and business, or competition and solidarity are ultimately seen as opposing concepts.

Contractarian Approaches

Theories of the social contract have been among the most influential in political and social philosophy (cf. for example, Hobbes, 1651/1991; Hume, 1739–40/1978; and Rawls, 1971). Their theoretical core is the normative criterion of the actors' consent, which is usually only given if it serves the actor's self-interest.

In business ethics, this tradition has generated two different schools: contractarian and contractualist approaches, which can be distinguished in the following way (cf. Wempe, 2004): Contractualism assumes that contracting is subject to an internal morality, according to which contracts only work if the contracting parties have certain moral capabilities and can commit themselves to certain norms that are not ruled by self-interest alone. Contractarianism, on the other hand, does not regard an internal morality as a necessary condition of the contracting process or of its stability. The actors involved are seen as being self-interested in some way.

One approach to business ethics that makes use of social contract theory is the integrative social contracts theory (ISCT) of Donaldson and Dunfee (1995, 1999). Donaldson and Dunfee, whose theory can be classified as broadly contractualist, aim at developing a theory of business that starts from concrete problems of business. These concrete problems lie mainly in the nature of contracting, which, in the globalized world, is subject to many different standards and norms. ISCT is meant to explicitly allow for such a variety of norms – within limits, however. The contractarian idea is used to conceptualize both these intentions by distinguishing between macrosocial and microsocial contracts.

All (boundedly rational) human beings would consent to the macrosocial contract, which then regulates the process of norm finding on lower levels. This macro level, however, leaves a great degree of 'moral free space' to be filled by specific communities, corporations, or individuals: They agree upon microsocial contracts, within the macrosocial framework, which may comprise quite different norms and standards. To deal with the question whether any such norms and standards can be ethically acceptable, Donaldson and Dunfee introduce the concept of hypernorms. Hypernorms are meant to be universal, overarching ethical principles that "are sufficiently fundamental that they serve as a source of evaluation and criticism of community-generated norms, and may include not only rules specifying minimum behaviour, such as the rule against the killing of innocents, but imperfect duties such as virtue, beneficence and decency" (Donaldson and Dunfee, 1995, p. 96). Hypernorms express principles "so fundamental to human existence that one would expect them to be reflected in a convergence of religious, political and philosophical thought" (Donaldson and Dunfee, 1995, p. 96).

Having the contract paradigm in mind, it must be mentioned that ISCT's hypernorms are not subject to a contract themselves. Rather, they are seen as boundaries of the moral free space, together with the consent required from the members of the community and certain priority rules as rules of thumb for dealing with the inevitable norm conflicts between communities. To sum up, ISCT mainly tries to achieve a reconstruction of the (necessary) internal norms of contracting (cf. Wempe, 2009). It is therefore contractualist rather than contractarian.

The contractarian approach of order ethics ("Ordnungsethik"), by contrast, focuses on the order framework of a society as a means for implementing ethics within the economic world. It can be regarded as the complement of the German conception of 'Ordnungspolitik,' which stresses the importance of a regulatory framework ('Ordnung') for the economy. Order ethics is based on contractarianism: The members of a society agree on its rules, implicitly or explicitly. This agreement will only be given if the rules improve one's own position, at least in the long run, and if they improve it to a larger extent than the relevant alternatives.

Thus, the order framework of a society is regarded as a means for implementing ethics. Therefore, the discipline of economics takes a large space within this approach, as it has a long history of dealing with well-being, normativity, and norms under conditions of modern societies and economies. In the 'pure' tradition of ethics, implementation is most often conceptualized in terms of appealing to actors or their conscience.

By contrast, economics stresses the – not only economic but also ethical – importance of rules and incentives: Economics can teach ethically relevant lessons, one of which is the following.

The Role of Competition in Modern Societies

Order ethics emphasizes the special role competition plays in a modern society with market economy. Certainly, competition has many positive aspects (cf. Hayek, 1978). It fosters innovation and the spreading of new ideas, and it tends to erode positions of power (for example, those of former

monopolists). The negative aspects are the nonregard to justice, but (in a direct sense) only to efficiency, the role of luck (not only of merit), and the constant danger that morality gets crowded out when the incentives work against it.

Dilemma situations from game theory, in particular the prisoners' dilemma or the stag hunt game, model such situations in which morality gets crowded out if the incentives set by the rules, by the order framework, thwart the ethical aim (cf. Axelrod, 1984, for the systematic use of the stag hunt game cf. Binmore, 1994, 1998, 2005). In the classic model of the prisoners' dilemma, the prisoners cannot be expected to cooperate, because the conditions of the situation are set in a way so that cooperation is punished by defection on the part of the other player. The actors are permanently faced with the possibility of being 'exploited' by others if behaving cooperatively, and therefore they stop cooperating themselves preemptively. This leads to a situation where rational, self-interested actors end up with a result that leaves everybody worse and no one better off: Morality gets crowded out. However, rules are designed to prevent this from happening.

For example, corruption is seen as unethical, but some kinds of order frameworks (for example in Germany until 1998) allow for bribes to be deducted from tax, with the consequences of fostering unethical behavior. Order ethics aims at changing the order framework of a society rather than at appealing to moral behavior, by insisting that ethical behavior should be made possible by the rules and not get punished by (counterproductive) incentives.

The ethical consequence of this theory is to make markets more moral – not managers. This is the way Boatright (1999) put it, when he distinguished between a moral manager model and a moral market model. The first one focuses on appealing to individuals (managers, officials, trade unionists, etc.); the second one aims primarily at other aspects; in particular, this encompasses changing the institutional framework (thus indirectly inducing ethical behavior) and preventing rules and individuals' actions from getting into opposition with one another. Rules themselves are not sacrosanct, of course, but subject to meta-rules, in particular, constitutional rules. For these meta-rules, higher degrees of consent must be demanded (cf. Brennan and Buchanan, 1985).

Regarding the nature of rules, it should be stressed that order ethics does not deal only with those rules incorporated in the law, but with rules on other levels as well. This includes, in particular, agreements at branch level and also self-constraining actions of individual corporations. The underlying economic idea of mutual advantages, however, stays the same: to aim for a win–win situation.

Order ethics, as it becomes clear, does not equate altruism with moral behavior nor egoism with immoral behavior. The demarcation lies elsewhere, between unilaterally and mutually beneficial action: In order to act in an ethical way, actors should be pursuing their advantage in such a way that others benefit as well. This implies that there usually are different ways of pursuing one's own advantage, and the ethical challenge lies in finding new and better ways. What is needed here is some kind of 'moral imagination' (Werhane, 1999), which does not have to be detrimental to actors themselves, but rather give them advice on how to make profit in a mutually beneficial way.

Topics of Business Ethics

Apart from different theoretical approaches, the field of business ethics can be differentiated into a number of subtopics and subfields. The following are some of the most discussed:

1. Ethics of globalization: Business ethics in times of globalization presents specific ethical challenges. These are some of the central questions: How can the entire process of globalization be understood and analyzed from ethical points of view? How can different cultural, religious, or ideological differences be represented adequately? What responsibilities do multinational corporations (MNCs) have (cf., e.g., Homann et al., 2007)?
2. Criticism of capitalism: Already, but especially after the financial crisis of 2008–2009, fundamental criticism of the market economy has grown significantly. These are the important questions in this field of research: (How) Can capitalist societies be justified ethically or do we need to develop fundamental alternatives? Does capitalism have inherent ethical problems (cf., e.g., Chomsky, 1999)?
3. Sustainability: The concept of sustainability has become especially widely used, not only in ethics but also in other disciplines (e.g., terms like corporate sustainability, corporate sustainable management, and so on). From an ethical point of view, questions can be posed like: What counts as sustainable from an ethical perspective? What are ethical criteria for a sustainable economy? Do we need to develop sustainable lifestyles (cf., e.g., Marrewijk, 2003)?
4. Poverty: Poverty is still among the greatest both economic and ethical challenges in the world. What are the ethical consequences of poverty? Does the gap between the 'rich' and the 'poor' get larger or narrower? (How) Can inequalities be justified? Is extraordinarily high compensation justified (cf., e.g., Luetge, 2012b)? (How) Can instruments like microcredits be useful for ethical purposes (cf. Yunus, 2003)?
5. Unemployment: Unemployment, both an economic and an ethical problem, is abundant in many countries. Can a right to labor be justified? Is the concept of a basic (unconditional) income ethically desirable?
6. Ethics of financial markets: This kind of subfield has gained much ground since the financial crisis. These are some questions asked: Do financial markets need a special kind of ethics? Have bankers been too greedy and ruined the economy? Has the banking sector disentangled itself too far from the 'real' economy (cf. Boatright, 2007)?
7. Human rights and business ethics: The greatly increased role and power of MNCs have also increased the number of human rights problems for MNCs (cf. Bilchitz, 2013), such as: How should MNCs conduct business in countries with human resource (HR) deficits? How far can MNCs be required ethically to go to improve the local HR situation not only within their factories but also in society in general?
8. Management ethics: Many studies, articles, and concepts in business ethics deal with particular management instruments and how to make ethical use of them. These include (but are not limited to) the triple bottom line, ethics

committees, codes of ethics, ethics audits, whistle-blowing, and others (cf. Hosmer, 2008).

9. Corruption: The problem of corruption can be counted among the internationally most well-studied phenomena in business ethics. Some of the most pressing ethical questions posed are the following: How should corporations do business in highly corrupt countries? Can corporations be held responsible for corrupt practices that occur despite, and/or contrary to, company policy (cf. Lambsdorff, 2007)?
10. Diversity and gender: Many of the originally gender-focused discussions have now turned to the concept of diversity as an umbrella term encompassing many different aspects, including gender, ethnical background, religions and convictions, age diversity, sexual orientation, disabilities, and more. Most discussions center on the question of how MNUs can recognize and include all these aspects in their business (cf. Thompson, 2013; Luetge, 2013, Part 16).
11. Marketing ethics: What kinds of marketing can be regarded as ethically acceptable? Are there ethical limits to marketing strategies (cf. Murphy et al., 2004)?
12. Ethics of consumption: What responsibilities can be attributed to consumers (if any)? How can the concept of 'sustainable consumption' (cf. Seyfang and Elliott, 2009) be spelled out in detail?
13. Sponsoring: What kinds of sponsoring are ethically acceptable? What can sponsored organizations demand from their sponsors from an ethical point of view?
14. Entrepreneurship and ethics: This is a topic that has been neglected, which is, however, very important to the discussion of business ethics. What ethical responsibilities can be attributed to entrepreneurs? Do they (and if so, in what way) differ from the responsibilities of established corporations? How can social entrepreneurship be classified from an ethical standpoint?

Many of these topics lead into areas that have, since the 1990s, differentiated themselves from business ethics: corporate social responsibility, corporate citizenship, corporate social accounting, corporate governance, corporate sustainability management, and others.

In all these cases, the central question remains: How can ethics and profit and morality and self-interest be realized simultaneously, without one of the two getting suppressed by the other? This is the key challenge of business ethics.

See also: Contractarianism; Conventions and Norms: Philosophical Aspects; Democracy, History of; Economic History; Economics and Ethics; Economics, Philosophy of; Liberalism: Historical Aspects.

Bibliography

Aaken, A.v., List, C., Luetge, C., 2004. Deliberation and Decision: Economics, Constitutional Theory, and Deliberative Democracy. Ashgate, Aldershot.
Aristotle, 2003. Nicomachean Ethics. Penguin Classics, London.
Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York.
Bilchitz, D., 2013. Corporations and Fundamental Rights: What is the Nature of Their Obligations, If Any? In: Luetge, C. (Ed.), Handbook of the Philosophical Foundations of Business Ethics. Springer, Heidelberg/New York, pp. 1053–1076.
Binmore, K., 1994. Game theory and the social contract. In: Playing Fair, vol. 1. MIT Press, Cambridge, MA, London.
Binmore, K., 1998. Game theory and the social contract. In: Just Playing, vol. 2. MIT Press, Cambridge, MA, London.
Binmore, K., 2005. Natural Justice. Oxford University Press, Oxford.
Boatright, J., 1999. Presidential address: does business ethics rest on a mistake? Business Ethics Quarterly 9, 583–591.
Boatright, J., 2007. Ethics in Finance. Wiley-Blackwell, Hoboken.
Brennan, G., Buchanan, J., 1985. The Reason of Rules. Constitutional Political Economy. Cambridge University Press, Cambridge.
Chomsky, N., 1999. Profit over People: Neoliberalism and Global Order. Seven Stories Press, New York.
Cohen, J., 1989. Deliberative democracy and democratic legitimacy. In: Hamlin, A., Pettit, P. (Eds.), The Good Polity. Blackwell, Oxford.
De George, R., 2006. The history of business ethics. In: Epstein, M.J., Hanson, K.O. (Eds.), The Accountable Corporation. Business Ethics, vol. 2. Praeger, Westpoint.
Donaldson, T., Dunfee, T., 1995. Integrative social contracts theory. A communitarian conception of economic ethics. Economics and Philosophy 11 (1), 85–112.
Donaldson, T., Dunfee, T., 1999. Ties That Bind. A Social Contracts Approach to Business Ethics. Harvard Business School Press, Cambridge, MA.
Elster, J. (Ed.), 1998. Deliberative Democracy. Cambridge University Press, Cambridge.
Hayek, F.A.v., 1978. Competition as a discovery procedure. In: Hayek, F.A.v. (Ed.), New Studies in Philosophy, Politics and Economics. University of Chicago Press, Chicago.
Hobbes, T., 1651/1991. In: Tuck, R. (Ed.), Leviathan. Cambridge University Press, Cambridge.
Homann, K., Koslowski, P., Luetge, C. (Eds.), 2007. Globalisation and Business Ethics. Ashgate, Aldershot.
Hosmer, L.T., 2008. The Ethics of Management, sixth ed. McGraw-Hill, New York.
Hume, D., 1739–40/1978. A Treatise of Human Nature, second ed. Clarendon, Oxford.
Lambsdorff, J.G., 2007. The New Institutional Economics of Corruption and Reform – Theory, Policy and Evidence. Cambridge University Press, Cambridge.
Luetge, C., 2012a. Economic ethics, overview. In: Chadwick, R. (Ed.), Encyclopedia of Applied Ethics, second ed., vol. 2. Academic Press, San Diego, pp. 13–19.
Luetge, C., 2012b. Executive compensation. In: Chadwick, R. (Ed.), Encyclopedia of Applied Ethics, second ed., vol. 4. Academic Press, San Diego, pp. 243–249.
Luetge, C. (Ed.), 2013. Handbook of the Philosophical Foundations of Business Ethics. Springer, Heidelberg/New York.
Mandeville, B., 1714/1924. In: Kaye, F. (Ed.), The Fable of the Bees, or, Private Vices, Publick Benefits. Clarendon, Oxford.
Marrewijk, M.v., 2003. Concepts and definitions of CSR and corporate sustainability: between agency and communion. Journal of Business Ethics 44 (2–3), 95–105.
Marx, K., Engels, F., 1959. Marx-Engels-Werke. Dietz, Berlin.
McCloskey, D., 2006. The Bourgeois Virtues. Ethics for an Age of Commerce. University of Chicago Press, Chicago.
Murphy, P.E., et al., 2004. Ethical Marketing. Prentice Hall, Upper Saddle River, NJ.
Palazzo, G., Scherer, A., 2006. Corporate legitimacy as deliberation: a communicative framework. Journal of Business Ethics 66, 71–88.
Pesch, H., 1905ff.. Lehrbuch der Nationalökonomie. Herder, Freiburg.
Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge, MA.
Rousseau, J.-J., 1755/1964. Discours sur l'origine et les fondements de l'inégalité parmi les hommes, Oeuvres compl, vols. 3. Bibliothèque de la Pléiade, Paris.
Rucellai, G., 1772. Ricordanze. Padua.
Scherer, A., Palazzo, G., Baumann, D., 2006. Global rules and private actors: toward a new role of the transnational corporation in global governance. Business Ethics Quarterly 16 (4), 505–532.
Seyfang, G., Elliott, D., 2009. The New Economics of Sustainable Consumption: Seeds of Change. Palgrave MacMillan, Basingstoke.
Shibusawa, E., 1916. Rongo to Soroban. Toado Shobo ("The Analects and the Abacus"), Tokio.
Smith, A., 1759/1976. In: Raphael, D., Macfie, A. (Eds.), The Theory of Moral Sentiments. Oxford University Press, Oxford.
Smith, A., 1776/1976. In: Campbell, R., Skinner, A. (Eds.), An Inquiry into the Nature and Causes of the Wealth of Nations. Oxford University Press, Oxford.

Steinmann, H., Scherer, A., 1998. Corporate ethics and global business: philosophical considerations on intercultural management. In: Kumar, B.N., Steinmann, H. (Eds.), Ethics in International Business. De Gruyter, Berlin/New York.

Thompson, L., 2013. Framing Global Gender Issues: Cross-Cultural Theory and Analysis. In: Luetge, C. (Ed.), Handbook of the Philosophical Foundations of Business Ethics. Springer, Heidelberg/New York, pp. 1113–1127.

Ulrich, P., 2008. Integrative Economic Ethics: Foundations of a Civilized Market Economy. Cambridge University Press, Cambridge (Original 1987).

Weber, M., 2001 (Original 1915–1919). In: Schmidt-Glintzer, H. (Ed.), Die Wirtschaftsethik der Weltreligionen, Gesamtausgabe, vol. 20. Tübingen: Mohr Siebeck (in English: The Social Psychology of the World Religions).

Weber, M., 2005. In: Parsons, T., Giddens, A. (Eds.), The Protestant Ethics and the Spirit of Capitalism. Routledge Classics, London (Original: Weber, M. Die protestantische Ethik und der Geist des Kapitalismus. Archiv für Sozialwissenschaften und Sozialpolitik 20 (1905) 1–54; 21(1905) 1–110).

Wempe, B., 2004. On the use of the social contract model in business ethics. Business Ethics: A European Review 13 (4), 332–341.

Wempe, B., 2009. Extant social contracts and the question of business ethics. Journal of Business Ethics 88 (4), 741–750.

Werhane, P., 1999. Moral Imagination and Management Decision Making. Oxford University Press, New York.

Yunus, M., 2003. Banker to the Poor: Micro-Lending and the Battle against World Poverty. Public Affairs, New York.

Business History

Hartmut Berghoff, German Historical Institute, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes business history as an academic discipline before it sketches the main methods and approaches of the field. It also introduces the most important literature, the best journals, and the leading institutions of business history.

Business History

Strictly speaking, business history as the history of private and/or public corporations with specific legal structures must be distinguished from entrepreneurial history, the history of entrepreneurially active individuals. In the first instance, the concern is with the development of organizations in which questions of economic success loom paramount. In the second case, the object of study is a certain social type and/or personality as well as behavior, both inside and outside of the enterprise, including its social, cultural, or political implications. In practice, however, institutional and biographical approaches tend to overlap.

History of the Scholarly Discipline

Modern business and entrepreneurial history have two sources. First is the need of corporations and businesspeople for image cultivation and self-affirmation, which has led to the publication of numerous commissioned biographies and company histories as well as autobiographical accounts. The second major impetus behind the development of business history was the work of the German historical school of economics. This school questioned the explanatory power of classic liberalism and attempted to replace it with alternative concepts in which long-term perspectives as well as the entrepreneur played key roles in explaining economic growth. Of primary significance was the work of the scholars Gustav Schmoller, Werner Sombart, and Lujo Brentano.

Against this theoretical background, scholars around 1900 began to write academic business histories, thus establishing a counterweight to the celebratory accounts that banks and, later, large industrial firms produced. Both genres were reacting to the mounting criticism of the conduct and profits of big business. In this way, business history served, among other functions, to legitimize a rising social class that was rapidly acquiring both affluence and influence. In 1912, Richard Ehrenberg, whose history of the Siemens Brothers (1906) was one of the first company histories based on archival documents, collaborated with Hugo Racine on a study of the Krupp workforce and their families, which defended the company policies of the Ruhr's heavy-industry sector. This period also saw the founding of the first systematically organized company archives. In Great Britain, George Unwin (1904) published his pioneering work on the origins of the modern enterprise in 1904, which traced the decline of the guilds through to the emergence of functionally differentiated management structures in the seventeenth century.

In the twentieth century, scholarly approaches to business history received a further impulse from technology and from management colleges. Here historical methods were combined with the self-promotional interests of new academic disciplines. College-trained engineers and managers not only wanted to secure their academic standing, but also harbored naive hopes of being able to derive general laws of historical development.

In the nineteenth-century United States, numerous nonacademic business histories on banks, canal companies, and railroads were published. Business history first established itself academically in the 1920s at the Graduate School of Business Administration of Harvard University, which sought to use business history for teaching purposes. Future managers were to be trained through an analysis of historical case studies and through a comparison of past and current methods of executive leadership. With the appointment of Norman Gras to the first-ever chair for business history in 1927, Harvard laid the groundwork for the institutionalization of American business history. In his magnum opus *Business and Capitalism* (1939), inspired by the work of Sombart, Ehrenberg, and Unwin, Gras divided the economic history of America into stages defined by each period's distinctive business type: petty, mercantile, industrial, financial, and national capitalism. The *Casebook in American Business History*, which Gras coauthored with his Harvard colleague Henrietta M. Larson, also appeared in 1939. In 1948, Larson published an extensive bibliography in her *Guide to Business History*. The year before, Gras and Larson had gone a step further down the road to firmly establishing their discipline with the creation of the Business History Foundation. In 1954, the Business History Conference was founded, which has remained the discipline's central American organization to the present day. In this period, periodicals appeared such as the *Journal of Economic and Business History* (1928–32), *Bulletin of the Business Historical Society* (1938–53), and *Business History Review* (since 1954).

Business history received an additional stimulus from the favorable reception of the work of Joseph Schumpeter and his stress on entrepreneurial innovation as the driving force of economic development. Following his programmatic address to the Economic History Association, Schumpeter himself helped to establish the Research Center in Entrepreneurial History (1948–58), funded by the Rockefeller Foundation and the Carnegie Corporation. The scholars associated with the center – Arthur H. Cole, its first director, as well as Thomas C. Cochran, Leland H. Jenks, and the German economist Fritz Redlich – centered their work on the concept of entrepreneurship and sought to anchor it in a broader social context. The Center also

published its own journal, *Explorations in Entrepreneurial History*. In 1959, after 10 years, Cole summarized the Center's most significant findings, which went far beyond those of mere biographical studies. The main emphasis was on the development of entrepreneurial typologies and on the relation of the entrepreneur to the corporation and its environment.

Beginning in the 1940s, more and more American universities, colleges, and business schools introduced courses and chairs in business history. In Germany, business history is primarily taught under the rubric of economic and social history in departments of history or economics. In most other European countries, it is mainly taught at business schools.

The growing interest in business history during the 1950s was reflected in the founding of numerous new scholarly journals, including *Business History* (founded in 1956, United Kingdom), *Tradition* (1956, West Germany; since 1978, *Zeitschrift für Unternehmensgeschichte*), *Technology and Culture* (1958, United States), and *Histoire des Entreprises* (1958, France). Ironically, business history remained a force even when its object of study was undergoing severe periods of crisis. In the wake of the student movements of the 1960s, entrepreneurs and corporations were subjected to harsh criticism. Then, in the 1970s, the pioneer countries of the industrial age lapsed into a structural crisis that, within a few years, had shrunk the long-dominant sectors of mining, iron, steel, and textiles. These upheavals led to the founding of numerous museums and business archives that have since become lively research sites. Added to this were newly established academic organizations such as Germany's *Gesellschaft für Unternehmensgeschichte* (Society for Business History, established 1976).

The maturing of business history as a discipline has been reflected in the emergence of subdisciplines. The 1990s saw the founding of specialized business history journals such as *Accounting: Business and Financial History* (1991), *Financial History Review* (1994), and *Journal of Industrial History* (1998). By contrast, the successor of *Business and Economic History Review* (1962–99), *Enterprise and Society* (2000), began to pose questions inspired and informed by the broader fields of social and cultural history.

The regional embeddedness of much business activity has not infrequently led to the founding of business history museums and archives. Prominent examples are the Center for Business History in Glasgow, Scotland; the Hagley Museum and its affiliated Center for the History of Business, Technology, and Society in Wilmington, Delaware, United States; and the Society for Westphalian Economic History in Germany's Ruhr district. In the 1990s, a number of international organizations were founded, such as the European Association for Banking History (1990) and the European Business History Association (1994). It is a sign of the consolidation and maturation of the field that, in recent years, a growing number of handbooks (Jones and Zeitlin, 2008), textbooks with source collections (Blaszczyk and Scranton, 2006), introductions for students (Berghoff, 2004), and accounts of long-term developments (Landes et al., 2010; McCraw, 2000) have been published.

Methods and Approaches of Business History

From the moment it entered academia, business history began to borrow methodologies from neighboring disciplines, particularly economics and management studies but also from history, sociology, political science, and anthropology. Equally variegated are the research areas and methodologies of modern business history. Roughly speaking, one can distinguish 10 areas of study, which are discussed in the remainder of this section.

Finance Capital and Industrial Production

Stimulated chiefly by Rudolf Hilferding's thesis of the dominance of big banks over industry, the relationship between the financial and the industrial sectors is a focus of analysis. For European and American industrialization, it has been argued that banks and money markets at first played only a subordinate role in the financing of business. However, when these firms reached certain growth thresholds – for example, merging with or investing in capital-intensive technologies – external funding was required, which would lead them into a close relationship with the banking sector.

Of Great Britain, it is said that between 1880 and 1930, London financial institutions neglected industrial interests, thus hindering the modernization and concentration of the country's industrial base. This was in contrast to the big banks of Paris and Berlin. On the other hand, Collins sees the aloofness of the money market to industry as a consequence rather than a cause of slow corporate growth. Davis argues that the American money market *c.*1900 was less efficiently organized than its British counterpart, and the limited availability of credit explains the concentration of capital in the hands of multimillionaires like Rockefeller and Carnegie (Grossman, 2010). By contrast, Chandler emphasizes the dominance of internally generated cash flow as the result of successful integration. Case studies on the close relationship between big banks and industry in Germany show that one cannot speak of a general ascendancy of financial institutions (Fohlin, 2007).

Businesses as Innovators and Users of Technology

Technological innovations, either self-generated or initiated from outside the firm, are central to the development of many enterprises. Thus, business history undertook the description of technical processes; however, this often resulted in uncritical lists of glorious inventions and paeans to the genius of particular entrepreneurs. The new business history discusses questions regarding the timely application of technical innovations or their adequacy for specific conditions of production and sales. In the debates on the competitiveness of different nations, comparative investigations have been particularly enlightening. The mass production technologies that had been suitable for the large and relatively homogeneous American market and a predominantly unskilled labor force were nontransferable – at least with the same level of success – to the completely different economic terrain of Europe. It is also not always an advantage to be in the forefront of technological innovation. Wait and see – then imitate. Later investments in more mature technologies have sometimes proven a superior strategy, especially for countries lagging behind the pioneers of industrial development. Further significant themes in this field are mechanisms of technology transfer, research and development, the relationship of companies and academic research, as

well as the social, economic, and mounting ecological consequences of technical progress.

Organizational Capability, Corporate Governance, and Comparative Advantage

Among the central themes of business history is the quest for an optimal organizational structure with respect to strategy, technology, and the market. Alfred D. Chandler of the Harvard Business School put forward theses on this question that were to preoccupy entire generations of scholars. In his first major work, *Strategy and Structure* (1962), Chandler asserted that changes in the "market – cum – technological environment" modify the "basic long-term goals and objectives of an enterprise,' which themselves determine the shape of the company's organization. Chandler's *The Visible Hand* (1977) stressed the superior performance of managers (i.e., salaried entrepreneurs) in comparison to owner-entrepreneurs. His third notable work, *Scale and Scope* (Chandler, 1990), extended the study to Great Britain and Germany, concluding that differences in national competitiveness were the result of divergent utilization of economies of scale. In employing this approach, Chandler elevated the American model of managerial capitalism and its large-scale "three-pronged investment in production, management, and marketing facilities" to a universal prescription for success.

Chandler's main achievement was that he moved beyond the stage of individual case studies and proposed a general paradigm for the analysis of industrial capitalism. At the same time, his claim to general validity provoked vehement protest. The ensuing debate was very productive but more or less refuted Chandler's idealization of the American model of big industry. His critics complained that Chandler had ignored labor relations, culture, regulation, politics, and smaller companies below the top 200.

Michaele Piore and Charles Sabel (1984) pointed to the viability of historical alternatives to mass production. According to their 'theory of dualism,' mass production is not the only paradigm of industrial capitalism. Small-scale enterprises, always representing the majority of companies, were also able to operate successfully. The most efficient were those specializing in small-market segments and using flexible production technologies along with highly skilled labor. Fluctuations in demand and design could be easily coped with. This 'modern craft production' was a 'necessary complement to mass production' and explains the persistence of the small firm. Multiple paths did coexist, and successful specialty production hinged on quality, differentiation, human skills, and networks rather than on high-throughput technologies and size (Scranton, 1998).

Associated with this debate are questions concerning the optimal size of firms, the reasons for and success of mergers and acquisitions, the relation of transaction costs and corporate structure, and control and governance. Research by Mary O'Sullivan (2000) has shown that, in the 1980s and 1990s, corporate control became more controversial. Previously silent stockholders were empowered by institutional investors and pressed for the maximization of 'shareholder value.' As a consequence, many top executives in the United States and, soon after, Europe subscribed to this maxim, rearranged their incentive structures, and by doing so, fundamentally changed the way their corporations were run. This provoked criticism from other stakeholders, who rightly claimed that their interests also mattered, socially and economically.

Multinationals in the Process of Internationalization and Globalization

Closely connected to the question of appropriate organizational structures is the internationalization process within firms (Jones, 1996). Why do companies prefer foreign direct investments with direct control as opposed to portfolio investments or cooperation with independent foreign partners? Firms like Nobel and Vickers invested abroad in order to exploit competitive advantages acquired at home. By contrast, Shell and Lever secured access to vital raw materials. Others were desirous of skirting customs barriers, operating more closely to the market, or securing foreign know-how, locational advantages, and subsidies. Casson maintains that transaction costs economies caused cross-border market exchange to be substituted by multinational hierarchical organization. Internalization is a superior method of coordination across national frontiers. Related hybrid structures such as joint ventures, cartels, and strategic alliances are also analyzed with the use of similar questions (Casson, 2000).

Firms as a Venue of Social Interaction

The history of work relations was given particular impetus by the classic school of social history. It discovered that in the process of industrialization, many entrepreneurs rose from related occupations like trading or craftsmanship from which they derived important social and financial capital, as well as technological and/or organizational know-how. Their degree of social mobility has long been overestimated. Social history further asks how owners of capital, managers, employees and workers, and men and women organize their social relations within firms. In what ways are individual interests bound together organizationally, and how do they factor into the bargaining process? Thus the focal points of research are the history of unions, of worker participation, and of company social policy, and methods of in-house company relations (Cassis and Minoglou, 2006).

Another area of study is the mechanisms with which top management uses its power and control in the implementation of directives. How is the principal – agent problem solved? That is, in what way is the self-interest of managers and employees curbed with regard to the goals of the entire firm and the capital owners? Chandler sees the rise of better educated managers bereft of family interests as one of the positive and decisive competitive advantages of the American model. He sees the main reason for the relative decline of British industry in its insistence on 'personal capitalism.' Numerous critics have warned against such generalized statements. There is little doubt, however, that the emergence of large company structures frequently coincided with the rise of managers who had no familial connection to the capital owners.

The Survival of the Family Business

In recent decades, family businesses have been reevaluated and acquitted of Chandler's verdict of being generally inferior to

managerial firms and incapable of reaping economies of scale and scope (Colli, 2003). In fact, they are the dominant type of business in most countries. There are many examples of family-controlled and even family-run giant firms and 'global champions'. Moreover, scholars, noting the persistence of family firms, have noted their diversity and stressed their comparative advantages, such as flexibility and lower transaction costs. Family members are often driven by motives beyond short-term financial gain, and they mobilize qualities like loyalty, intrinsic motivation, emotional reward, and long-term commitment. These findings were indirectly given credence by the corporate scandals of the 1990s and 2000s, which dealt a severe blow to the belief in the potential and integrity of managers.

The debate generally assumes a rigid dichotomy between family and managerial businesses, but there in fact exists a broad array of hybrid forms of governance. The recent literature, stressing variety rather than general developmental paths, leaves the impression that family firms, once they have reached a certain size, complexity, and age, are likely to include outside managerial expertise, and, after a phase of hybrid governance, the family's retreat begins.

The small and medium-sized firms of Germany's *Mittelstand* – many of them family owned as well as global champions in specialized niche markets – are no longer seen as relics but rather as models. However, the classic *Mittelstand* model began to be transformed in the 1970s (Berghoff, 2006). The new *Mittelstand* is less dependent on individual families, has more access to external capital, and displays a more international orientation.

Corporations as Cultural Institutions

Firms are bearers of specific corporate cultures that play a key role in internal and external communication. As concerns the internal structure, it is a question of the 'software of organization,' that is, the organization's shared rules and values. While Chandler suggests that the formal structures – the 'hard facts of organization' – are decisive, proponents of cultural business history assert that internal integration and coordination are essentially dependent on the strength of the given corporate culture. This assures loyalty, team spirit, and communication skills while also lowering transaction costs. It consolidates the results of long-term learning processes, especially firm-specific routines and tacit knowledge.

Corporate culture is frequently influenced by strong individuals, chiefly company founders, but also by external socio-cultural factors. Organizational models like those of the bureaucracy, the military, or large agrarian enterprises were especially effective in the industrialization process (Cole, 1959). Following mergers and takeovers, there is often a clash of cultures and incompatibility between differing value systems. Furthermore, it is precisely strong corporate cultures that are subject to inertia that can impede adaptation to external change. Research has shown that organically evolved corporate cultures determine the inner quality of a company in essential ways. However, social engineering – the arbitrary, short-term modification of corporate cultures – is impossible.

Corporate culture not only is effective as an internal link of identity and cohesion but also serves as an exogenous label and a seal of quality. It determines the firm's image in the minds of customers and the broader public, and defines the self-projection of the enterprise and its advertised products. The image-forming power of corporate culture extends to brands that rationalize the communication process with customers, while also implying an adherence to quality standards. The strong trust of such arrangements is central to all forms of marketing but can also be transformed into its diametrical opposite when that trust is abused.

Marketing History

Markets have to be made by marketers, advertisers, market researchers, and other intermediaries between production and consumption. Markets are not perfect mechanisms of self-coordination, and products don't find their way to customers all by themselves. The historical development of markets over the last 200 years or so has done little to cure market imperfections and lessen the need for intermediation. On the contrary, as industrialization and globalization broke down regional and national barriers, markets became much more complex. Rapid technological and cultural change made them ever-faster moving targets for those who sought to understand them. Marketing tries to process information on markets and translate it into a language that corporations can use to steer their operations. Consumers' varying preferences generate insecurity among marketers, who have to develop methods to overcome information asymmetries. This task calls for professional, even scientific, expertise as well as special instruments and institutions.

A number of studies, inspired by cultural history, have examined diverse aspects of corporations' marketing histories, such as the contributions of sales forces (Friedman, 2004) and advertising (Laird, 1998), the marketing efforts of individual firms, the targeting of certain ethnic groups (Walker, 2007), and market research. Other authors have taken a broader perspective and looked at the emergence and transformation of mass markets (Strasser) or the consumer society in general, or at the interaction of these factors (Berghoff et al., 2012).

Companies and Entrepreneurs as Political Actors

Modern companies operate within regulated frameworks. The state establishes legal forms that superintend ownership laws, property rights, responsibilities, and tax and disclosure duties. In addition to company law, labor law, environmental law, trade law, and consumer protection law are also of prime importance. Therefore, it is in the vested interest of the firm to influence the development of regulatory frameworks and to involve itself in the political process. The range of its dealings extends from simple forms of lobbying to participation in *coups d'états*.

Central topics of research are the effects of organized interests; the activity of companies in political parties, parliaments, administration, and governments; the thin line between sponsoring, donations, and corruption; private – public partnerships; the sale of products to the state; the nationalization or privatization of enterprises; and the correlation between business goals and politics in general.

The view of the political role and ethical responsibility of the private sector has gained in significance through research on

the behavior of businessmen in relation to the National Socialist (Nazi) regime. In reaction to the decades-long taboo placed on studies of the Nazi past, numerous works have emerged rejecting ideologically motivated generalizations and accusations, while at the same time confirming the deep involvement of many businessmen in National Socialism. More recent research has demonstrated that this was also true of businessmen in occupied, neutral, and allied countries (Turner, 2005). What is still debated hotly is the extent to which they still enjoyed some autonomy and followed 'normal' commercial considerations or whether they primarily acted under compulsion by the brutal dictatorship (Buchheim and Scherner, 2009; Hayes, 2009).

The Biographical Approach (Entrepreneurial History)

Entrepreneurial history is concerned with the people behind the strategic decisions. What sociocultural variables determined their worldview and thereby their business personality? Such questions were at first chiefly posed by biographical investigations that frequently sought to portray their subjects in heroic terms. Simultaneously, there was the countertradition of critical and even defamatory entrepreneurial biography.

Since the 1970s, social history has introduced quantifying methods into entrepreneurial biography. The great advantage of collective-biographical studies consists in the ability to distinguish between exceptional facts and general structures. Drawing upon selective and representative samples, they investigate heritage, education and upbringing, religion, marital and social status, financial circumstances, the chances of ennoblement, and political influence. One of the most heatedly discussed questions is whether European businessmen became gentrified by imitating the aristocracy and abandoning their own identity, a debate that has led to an overwhelming rejection of this thesis.

Emerging parallel to the wave of collective-biographical works have been several excellent individual biographies, family histories, and large collections of biographical articles (Daumas and Chatriot, 2010). Increasingly, online biographical compilations that include images and other source material are gaining in importance as reference and teaching tools. With the current wave of historiography addressing everyday life and mentalities, biographical methods are currently on the rise. Vivid and atmospheric narratives, including the experiences of spouses, children, and other relatives, now complement more representative statements on general structures that are analyzed through statistical methods.

Conclusion

On the whole, despite its varying status in different countries, business history has firmly established itself as an academic discipline and continues to evolve and expand. It is a field with many cross-discipline connections and a steady exchange with researchers from history, economics, management studies, sociology, political science, and anthropology. As before, there is still the danger of narrow accounts of individual case studies composed with an antiquarian mindset or commissioned histories in the interest of image building, legitimation, or adulation. Although this kind of business history has long been overshadowed by research that fulfills the highest scholarly criteria, publications that present a wealth of material but have little regard for historical context or try to construct heritage stories in the interest of corporate public relations are still published in large numbers. For serious scholars, however, the commitment to macrohistorical and macroeconomic contexts and the inclusion of theoretical approaches from related disciplines have become standard practice.

See also: Bourgeoisie and Middle Classes, History of; Business Ethics; Capitalism; Culture and Economic Development; Economic History; Economics and Ethics; Family Business; Family and Kinship, History of; Industrial Relations, History of; Industrialization, Typologies and History of; Labor History; Prosopography (Collective Biography); Social History.

Bibliography

Berghoff, Hartmut, 2004. Moderne Unternehmensgeschichte: Eine themen- und theorieorientierte Einführung. Schöningh, Paderborn.

Berghoff, Hartmut, 2006. The end of family business? The *Mittelstand* and German capitalism in transition, 1949–2000. Business History Review 80 (2), 263–295.

Berghoff, Hartmut, Scranton, Philip, Spiekermann, Uwe (Eds.), 2012. The Rise of Marketing and Market Research: Information, Institutions and Markets. Palgrave Macmillan, Houndmills/New York.

Blaszczyk, Regina Lee, Scranton, Philip B. (Eds.), 2006. Major Problems in American Business History: Documents and Essays. Houghton Mifflin, Boston.

Buchheim, Christoph, Scherner, Jonas, 2009. Corporate freedom of action in Nazi Germany: a response to Peter Hayes. Bulletin of the German Historical Institute 45, 43–50.

Cassis, Youssef, Minoglou, Pepelasis Ioanna (Eds.), 2006. Entrepreneurship in Theory and History. Palgrave Macmillan, Basingstoke/New York.

Casson, Mark, 2000. Economics of International Business; A New Research Agenda. Edward Elgar, Cheltenham.

Chandler, Alfred D., 1990. Scale and Scope: The Dynamics of Industrial Capitalism. Belknap Press, Cambridge, MA.

Cole, Arthur Harrison, 1959. Business Enterprise in Its Social Setting. Harvard University Press, Cambridge, MA.

Colli, Andrea, 2003. The History of Family Business, 1850–2000. Cambridge University Press, Cambridge, MA.

Daumas, Jean-Claude, Chatriot, Alain, 2010. Dictionnaire historique des patrons Français. Flammarion, Paris.

Fohlin, Caroline, 2007. Finance Capitalism and Germany's Rise to Industrial Power. Cambridge University Press, Cambridge (UK)/New York.

Friedman, Walter A., 2004. Birth of a Salesman: The Transformation of Selling in America. Harvard University Press, Cambridge, MA.

Grossman, Richard S., 2010. Unsettled Account: The Evolution of Banking in the Industrialized World since 1800. Princeton University Press, Princeton, NJ.

Hayes, Peter, 2009. Corporate freedom of action in Nazi Germany. Bulletin of the German Historical Institute 45, 29–42.

Jones, Geoffrey, 1996. The Evolution of International Business: An Introduction. Routledge, London/New York.

Jones, Geoffrey, Zeitlin, Jonathan, 2008. The Oxford Handbook of Business History. Oxford University Press, Oxford/New York.

Laird, Pamela Walker, 1998. Advertising Progress: American Business and the Rise of Consumer Marketing. Johns Hopkins University Press, Baltimore.

Landes, David S., Mokyr, Joel, Baumol, William J. (Eds.), 2010. The Invention of Enterprise: Entrepreneurship from Ancient Mesopotamia to Modern Times. Princeton University Press, Princeton, NJ.

McCraw, Thomas K., 2000. American Business, 1920–2000: How It Worked. Harlan Davidson, Inc., Wheeling.

O'Sullivan, Mary, 2000. Contests for Corporate Control: Corporate Governance and Economic Performance in the United States and Germany. Oxford University Press, Oxford.
Piore, Michaele, Sabel, Charles F., 1984. The Second Industrial Divide: Possibilities for Prosperity. Basic Books, New York.
Scranton, Philip, 1998. Endless Novelty: Specialty Production and American Industrialization, 1865–1925. Princeton University Press, Princeton, NJ.
Turner, Henry Ashby, 2005. General Motors and the Nazis: The Struggle for Control of Opel, Europe's Biggest Carmaker. Yale University Press, New Haven.
Unwin, George, 1904. Industrial Organization in the Sixteenth and Seventeenth Centuries. Oxford Clarendon.
Walker, Susannah, 2007. Style and Status: Selling Beauty to African American Women 1920–1975. University Press of Kentucky, Lexington.

Relevant Websites

http://www.thebhc.org
http://www2.h-net.msu.edu/~business/index.html
http://www.ebha.org
http://www.unternehmensgeschichte.de/englisch/index-eng.html
http://www.hbs.edu/leadership/database/index.html
http://www.immigrantentrepreneurship.org

Business Law

Constance E Bagley, Yale University, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the body of statutes, regulations, and case law governing transactions for economic gain among individuals, corporations, partnerships, and other forms of business entities. The laws governing business can be broken down into several basic categories: contracts; agency; employment law; forms of business entities; commercial laws governing the sale of goods, negotiable instruments, and secured lending transactions; product liability; and intellectual property. Other laws affecting the legal environment in which businesses operate are addressed as well, including antitrust and competition law, environmental law, consumer protection laws, regulations governing banking and the offer and sale of securities, and bankruptcy law.

Business law comprises the body of statutes, regulations, and case law governing transactions for economic gain among individuals, corporations, partnerships, and other forms of business entities. The laws governing business can be broken down into several basic categories: contracts; agency; employment laws; forms of business entities; commercial laws governing the sale of goods, negotiable instruments, and secured lending transactions; product liability; and intellectual property. Other laws affecting the legal environment in which businesses operate include antitrust and competition law, environmental law, consumer protection laws, regulations governing banking and the offer and sale of securities, and bankruptcy law.

Modern firms are awash 'in a sea of law' (Edelman and Suchman, 1997). Public law provides the 'rules of the game' (North, 1990) that govern a firm's relationship with its owners, employees, customers, creditors, and competitors. Business law does not just regulate, however; it also enables and facilitates (Suchman et al., 2001). For example, documented property rights, such as recorded deeds to real property, make it possible for entrepreneurs to borrow money collateralized by liens on their assets, thereby converting 'dead capital' into the financial capital necessary to start new firms and expand existing ones (De Soto, 2000). Contract law provides the mechanism whereby individuals and firms can create 'private law' to govern their relationships enforceable by the power of the state.

Law promotes economic growth by protecting property rights, enforcing contracts, allocating risks, facilitating capital markets, creating incentives to innovate, promoting liquid and skilled labor markets, providing economic incentives and infrastructure, and promoting free trade in the global markets (Bagley, 2013). It protects workers by regulating terms and conditions of employment, requiring employers to provide certain benefits, and protecting civil rights in the workplace by banning discrimination or the creation of a hostile environment. Law furthers consumer welfare by promoting safe, low-cost and innovative products and services and by preventing deceptive practices. It promotes public welfare by ensuring national security, promoting the fair and effective administration of justice, collecting taxes and spending money, guaranteeing fundamental rights, and protecting the environment (Bagley, 2013). It thus should come as no surprise that the World Economic Forum found a statistically significant [and positive] relationship between a country's gross domestic product per capita and each of the following: (1) the adequacy of legal recourse, (2) demanding product standards, (3) the extent of intellectual property protection, (4) the effectiveness of the antitrust laws, and (5) stringent environmental regulations (Porter, 2002).

Firms increasingly take an active role seeking to shape the legal environment in which they do business (Bagley, 2010). Legally astute top management teams can use the law and legal tools to increase both the total value created and the share of that value captured by the firm (Bagley, 2008). Conversely, failure to integrate legal considerations into firm strategy can put the firm at a competitive disadvantage and imperil its economic viability.

The remainder of this article focuses on the particular legal doctrines that affect the business environment.

Contracts

A contract is a legally enforceable promise. According to Douglass North and Barry Weingast, "one necessary condition for the creation of modern economies dependent on specialization and division of labor (and hence impersonal exchange) is the ability to engage in secure contracting across time and space" (North and Weingast, 1989, p. 831).

Elements

Formation of a valid contract generally requires four basic elements. First, there must be an agreement between the parties formed by an offer and acceptance of that offer. Second, the parties' promises must be supported by something of value, known as consideration. Third, both parties must have the capacity to enter into a contract. Fourth, the contract must have a purpose that is legal.

Intent to be Bound

As business transactions have grown more complex, courts have wrestled with the issue of when preliminary negotiations have ripened into a contract (Farnsworth, 1987). The leading US case arose out of Pennzoil Company's failed attempt to merge with Getty Oil Company (*Texaco, Inc. v. Pennzoil Co.*, 1987). The court held that a five-page memorandum of agreement, signed by the majority shareholders of Getty Oil and approved (but never signed) by the board, constituted a binding contract under New York law even though a definitive agreement for the $5 billion transaction was never signed.

In Japan, businesspersons are more likely to rely on oral contracts or on short contracts with a broad catch-all provision

stating that the parties will consult in good faith to resolve any questions of interpretation or matters not covered by the contract (Yanagida et al., 1994). As a result, Japanese courts are more likely than their US counterparts to conclude that negotiations have ripened into a binding contract. In addition, as the Japanese High Court explained in a case decided in 1987:

> In the event that preparation between two parties progresses toward conclusion of an agreement and the first party comes to expect that the agreement will surely be concluded, the second party becomes obligated under the principles of good faith and trust to try to conclude the agreement, in order not to injure the expectation of the first party. Therefore, if the second party, in violation of its obligation, concludes that the agreement is undesirable [absent certain circumstances], it is liable for damages incurred by the first party as the result of its illegal acts (Yanagida et al., 1994; pp. 228–229).

The US law imposes more limited precontractual liability under the doctrine of promissory estoppel for the out-of-pocket reliance damage suffered when one party reasonably and foreseeably relies on a promise made during negotiations.

Consideration

In most jurisdictions, each side must provide something of value to form a valid contract. The thing of value, known as consideration, can be money, property, a promise, or a service. Consideration is present if a party agrees to do something that he or she is not otherwise legally required to do or if the party agrees to refrain from doing something that he or she was otherwise legally entitled to do. Courts usually will not scrutinize the value of the consideration, thus giving rise to the adage that even a peppercorn can be adequate consideration. Indeed, under Japanese law, there is no requirement for consideration at all.

Unconscionability

A court will refuse to enforce a contract if it is so oppressive or fundamentally unfair as to shock the conscience of the court. This doctrine of unconscionability dates back to Aristotelian theory and Roman law (Gordley, 1981). The doctrine of *laesio enormis* provided a remedy when the contract price for the sale of goods deviated by at least half from the just price, defined as the market price for similar goods under similar circumstances. German and French laws that relieve a party from its obligations under a contract have their roots in this doctrine. Similarly, in England and the United States, the general principle of freedom of contract has always been limited by the doctrine of unconscionability.

Genuineness of Assent

Even if a contract meets all the requirements of validity (agreement, consideration, capacity, and legality), a court will not enforce it if there has been no true 'meeting of the minds' between the parties. Courts have found no meeting of the minds when there is fraud, duress, or a mistake of fact.

Fraud and duress

There is no meeting of the minds if a contract is tainted with fraud or if one party was forced to enter into it under duress. Thus, a contract is voidable – that is, subject to being undone by one or more parties – if one party induced another to enter into the contract by making false statements, by failing to disclose information when under a legal duty to do so (e.g., when the parties are partners), or by engaging in blackmail or extortion.

Mutual Mistake of Fact

A contract is also voidable if it was predicated on a substantial mutual mistake of fact. For example, in the classic case *Raffles v. Wichelhaus* (1864), two parties had signed a contract in which Wichelhaus agreed to buy 125 bales of cotton to be brought by Raffles from India on a ship named *Peerless*. Unbeknownst to both parties, there were two ships named *Peerless* both sailing out of Bombay during the same year. Raffles meant the *Peerless* sailing in December, and Wichelhaus meant the *Peerless* sailing in October. When the cotton arrived in the later ship, Wichelhaus refused to complete the purchase, and Raffles sued for breach of contract. The court concluded that the mutual mistake of fact had resulted in there being no meeting of the minds and, therefore, no contract.

Mistake of Judgment and Commercial Impracticability

Although a mutual mistake of fact makes a contract voidable, a mere mistake of judgment does not. Therefore, even if a party is mistaken about some aspect of what is bargained for, the contract is still enforceable. For example, the Wisconsin Supreme Court held that a contract to sell a stone for $1 was enforceable when neither party knew at the time that the stone was in fact a diamond (*Wood v. Boynton*, 1885).

Some jurisdictions will relieve a party of its contractual obligations if performance has become commercially impracticable, that is, if performance is made financially impracticable by the occurrence of an event unforeseen by the contract (Eagan, 1980). The added cost of performance must be very substantial before a court will relieve a seller or shipper of its contractual obligations. For example, the House of Lords in the United Kingdom refused to relieve a shipping company of its promise to deliver goods for the agreed-upon price after the closure of the Suez Canal caused the freight costs to double (*Tsakiroglou & Co. Ltd. v. Noblee Thorl G.m.b.H.*, 1960).

Statutes of Fraud

United States

Although most oral contracts are enforceable, almost every state in the United States has statutes requiring that certain types of contracts be evidenced by some form of writing. These Statutes of Fraud require that any (1) contract for the transfer of any interest in real property (such as a deed, lease, or option to buy), (2) promise to pay the debt of another, (3) agreement that by its terms cannot be performed within a year, (4) prenuptial agreement (i.e., an agreement entered into before marriage that sets forth the manner in which the parties' assets will be distributed and the support to which each party will be entitled in the event of divorce), and (5) contract for the sale of goods priced at $500 or more must be evidenced by some writing signed by the party against whom enforcement is being sought.

England, France, and Other Jurisdictions

Although the original Statute of Frauds was adopted in England in 1677, England has repealed all of it except the provisions

relating to transfers of interests in land and promises to pay the debt of others. Many other countries that once had a Statute of Frauds have repealed all or parts of it. France, Japan, and other civil law countries never had a writing requirement.

Remedies

In the event that a valid contract is breached, the other party is entitled to monetary damages or to a court order requiring performance. The nonbreaching party is generally entitled to expectation damages, that is, such amount of money as is necessary to put the plaintiff in the cash position it would have been if the contract had been fulfilled.

Consequential damages, that is, compensation for losses that occur as a foreseeable result of the breach, are also often available. However, if the party suing for consequential damages has suffered lost profits that were not reasonably foreseeable or a natural consequence of the breach, then there will be no award of consequential damages (*Hadley v. Baxendale*, 1854).

Agency

Agency law specifies under what circumstances the contracts entered into by one party (the agent) will bind another (the principal) on whose behalf the agent is acting. Agency law also addresses the question of when the principal will be liable for the torts or civil wrongs of the agent or for crimes committed by the agent. The employer–employee relationship (historically described as the master–servant relationship) is perhaps the most common type of agency in business settings.

Authority to Bind the Principal

A contract entered into by an agent on behalf of the principal is binding on the principal only if the agent had authority to enter into the contract or if the contract is later ratified by the principal. Authority to enter into a contract can be either actual or apparent. Actual authority can be given expressly or can be implied from the nature of the responsibilities the principal has given the agent to fulfill. Apparent authority exists when the principal, either directly or indirectly, leads a third party to believe that the agent has power to act on the principal's behalf.

Respondeat Superior

Under the doctrine of *respondeat superior*, the principal is liable for any torts (and certain crimes) committed by the agent while acting within the scope of employment. A minor detour, whereby the agent deviates from the assigned task, is still considered to have been within the scope of employment unless the deviation is so great as to constitute a frolic, that is, an excursion that in no way furthers the business of the principal.

A principal is more likely to be responsible for activities that otherwise might be deemed frolics when the harm done by the agent was made possible by an instrumentality provided by the principal. For example, if the principal provides the agent a truck to drive, the principal is more likely to be responsible for automobile accidents occurring while the agent is driving the truck even if at the time of the accident the agent was not directly furthering the business of the principal.

In Japan, the employer is not liable for the misdeeds of an employee if the employer exercised due care selecting the employee and supervising the employee's work or if the damages would have arisen even if due care had been exercised.

Employment Laws

Most jurisdictions have statutes and case law setting the minimum wage, imposing limits on the employment of children, giving workers the right to form unions and engage in collective bargaining, and specifying whether an employee without a written employment contract is an employee at-will who can be terminated at any time for any or no reason, or an employee who can be fired only for just cause. In addition, in many jurisdictions, the employment relationship is subject to statutes banning discrimination based on race, color, national origin, religion, sex, or age. For example, in 2007, Japan prohibited discrimination based on age in recruiting and hiring. The United States, the European Union, and India also prohibit discrimination based on disability.

Forms of Business Entities

A fourth body of law fundamental to the conduct of business prescribes the types of entities by which business can be conducted. A business may be conducted as a sole proprietorship owned by one person, a general partnership, a limited partnership, a limited liability company, or a corporation.

Corporations

The most prevalent entity used for large businesses is the corporation. A corporation is a fictitious person that has capacity to enter into contracts and to hold property. The equity of a corporation, evidenced by shares of stock, is owned by the shareholders. The shareholders elect the board of directors, which has the responsibility for managing the corporation. The board elects the officers, such as the president, treasurer, and secretary, who report to the board and serve at its pleasure.

Corporate governance

The respective rights and obligations of shareholders, directors, and officers of a corporation are dictated by the statutes of the jurisdiction in which the corporation is formed. For example, in the state of Delaware, where more than half of the Fortune 500 companies are domiciled, the board of directors has broad latitude to determine when it is in the best interests of the corporation to rebuff a hostile takeover offer (Bagley, 2013).

Jurisdictions vary significantly in their treatment of corporate stakeholders, such as employees, customers, suppliers, and communities. In large German corporations, codetermination statutes require that half of the supervisory board be elected by the employees and labor unions and councils. In Japan,

boards traditionally included no nonexecutive directors and put the welfare of the employees before shareholder return. In the United States, shareholder primacy tends to be the rule, at least once a decision has been made to sell control of the corporation or to break it up (Bagley and Page, 1999).

Limited Liability

A key advantage of a corporation is the limited liability it provides to its owners. Absent facts justifying piercing the corporate veil, such as the commingling of personal and corporate assets or the shareholders' disregard for the corporate form, the shareholders are not liable for the debts of the corporation. In contrast, in a general partnership, each of the general partners is personally liable for the debts and other obligations of the partnership. Similarly, the owner of a sole proprietorship is personally liable for all debts of the business. In a limited partnership, the managing partner has the same unlimited liability as a general partner in a general partnership, but the limited partners have no liability beyond their capital contribution unless they actively participate in management. In the case of a limited liability company, the owners (called members) have limited liability even if they are active in the management of the business.

Commercial Laws Governing the Sale of Goods, Negotiable Instruments, and Secured Lending Transactions

Because of the key role played by purchases and sales of goods in virtually every type of business, many jurisdictions have specific statutes designed to facilitate their sale. These statutes modify certain of the general contract rules discussed earlier in this article. Commercial laws also govern the creation and transfer of negotiable instruments and the lending of money secured by property.

Uniform Commercial Code

For example, in the United States, every state other than Louisiana has adopted a version of the Uniform Commercial Code (UCC).

Sale of Goods

The provisions in UCC Article 2 concerning the sale of goods eliminate many of the formalities required under the common law. For example, the UCC abolishes the 'mirror image rule,' which requires that an acceptance be identical in terms to the offer. Instead, the UCC acknowledges that sales transactions are often represented by an exchange of preprinted forms with additional or conflicting terms (the 'battle of the forms'), and it provides guidance as to which terms will be deemed to be part of the contract.

Negotiable instruments

Article 3 of the UCC governs negotiable instruments, including drafts, checks, notes, and certificates of deposit. The most common type of draft is a check: the writer of the check is the drawer; the bank on which the check is drawn is the drawee; and the person to whom the check is payable is the payee. Article 3 sets forth the requirements for negotiability: a writing signed by the maker or drawer containing a definite and unconditional promise or order to pay a fixed amount of money on demand or at a definite time. It also specifies what endorsements are required to negotiate an instrument payable to order, such as a check.

The UCC prescribes the rights given a holder in due course, that is, the holder of a negotiable instrument who took it for value, in good faith, and without notice that it had been dishonored or that any person has a defense against or claim to it. Unlike an ordinary holder who has only those rights that the transferor had, a holder in due course takes the instrument free of most of the defenses and claims to which the transferor was subject.

Secured Lending Transactions

Article 9 of the UCC governs transactions in which payment of a debt is guaranteed, or secured, by property owned by the debtor. It permits the secured party to perfect its security interest by filing a financing statement or, in some cases, by possession of the collateral. In general, the first to perfect has priority over other third parties who may wish to have their debts satisfied out of the same collateral.

Convention on Contracts for the International Sale of Goods

Most sales of goods involving parties in more than one country are governed by the Convention on Contracts for the International Sale of Goods (CISG), unless the parties expressly opt out of CISG. Unlike the UCC, CISG requires that the terms of the acceptance mirror those of the offer. It provides that an offer is irrevocable if the offeror orally states that it is irrevocable, or if the offeree reasonably relies on the offer as being irrevocable. In contrast, the UCC requires that an irrevocable offer not supported by consideration be in writing. Under CISG, the parties must specify the price or provisions for its specification, whereas the UCC requires that only the quantity be specified.

CISG imposes a duty for merchants to conduct themselves in good faith (Duncan, 2000). Thus, for example, a buyer may not arbitrarily deny a seller's request for additional time to deliver goods. This contrasts sharply with the UCC's perfect tender rule, which permits a buyer to reject goods not conforming perfectly to the contract.

Product Liability

Most jurisdictions impose strict liability on the sellers of goods for defective products, that is, liability without regard to fault. For example, the Council of Ministers of the European Union adopted a directive in July 1985, which made manufacturers and producers strictly liable for injuries caused by defects in their products. In certain jurisdictions, there is no strict liability in defective design cases. Instead, the plaintiff has the burden of proving that there was a defect in the product that could reasonably have been avoided by the use of an alternative design.

Intellectual Property

Intellectual property is any product or result of a mental process that is given legal protection against unauthorized use. It includes patents, copyrights, trademarks, and trade secrets.

Patents

A patent is a government-granted right to exclude others from making, using, or selling an invention for 20 years after the date the patent application is filed. A utility patent, the most common type of patent, protects any novel, useful, and nonobvious process, machine, manufacture, or composition of matter. In certain jurisdictions, human-made living organisms can be patented, as can computer software and business processes.

Copyrights

Unlike a patent, which can protect an idea, a copyright protects original expressions fixed in a tangible medium, such as books, records, musical scores, and films. Although registration of a copyright gives the author certain additional rights, the copyright is created as soon as the work is fixed in a tangible medium of expression. Under the Berne Convention, most industrialized countries give persons with a copyright under the laws of another jurisdiction the same rights that a national would have under local law.

Trademarks

A trademark is a word, name, symbol, or device used by a manufacturer to identify and distinguish its goods from those manufactured by others and to indicate the source of the goods. Examples include Coca-Cola, Marlboro, and Nescafe. A service mark, such as Barclays, is similar but is used in connection with services.

Trademarks are granted for a particular field of use, for example, nonalcoholic beverages or automobiles. Thus, more than one company can register the same mark. The importance of a company's domain name, or electronic address, for Internet branding purposes, and the fact that domain names contain no field-of-use restriction, have required courts to consider when use of a trademark within a domain name constitutes trademark infringement. Courts have ruled against 'cybersquatters,' who register a trademarked name as a domain name to force the trademark owner to buy the domain name. They also have found infringement when the company using the domain name offers the same or similar products as the trademark owner or uses the name in a way that dilutes or tarnishes a famous mark. For example, Hasbro, owner of the trademark Candyland for a children's game, was able to force the owner of Candyland.com to stop using the domain name for a pornography site.

Trade Secrets

A trade secret is information, including a customer list, formula, compilation, or process, that derives independent economic value from not being generally known and is the subject of reasonable efforts to maintain its secrecy. Perhaps the most famous trade secret is the formula for Coca-Cola. In response to perceived problems facing US businesses from foreign theft of trade secrets, the United States adopted the Economic Espionage Act in 1996, which imposes criminal liability (including fines and prison sentences) on any person who knowingly steals a trade secret or knowingly receives a wrongfully obtained trade secret.

In 1997, Volkswagen AG (VW) paid General Motors (GM) Corporation $100 million to settle allegations that VW's purchasing chief stole trade secrets, including plans for future GM models and car-building techniques, when he left GM to join VW. Under the emerging doctrine of inevitable disclosure, an employer can challenge a former employee's decision to work for a competitor if the new position would result in the inevitable disclosure or use of the former employee's trade secrets.

Other Laws Affecting the Legal Environment of Business

Other laws affecting the legal environment in which businesses operate include antitrust and competition law, environmental laws, consumer protection laws, regulations governing the offer and sale of securities, and bankruptcy law.

Antitrust and Competition Law

Cases brought by the European Commission and the US Justice Department against Microsoft Corporation and Google Inc. have required regulators and courts to wrestle with the question of how to reconcile the exclusionary rights arising out of intellectual property law with the bans on anticompetitive behavior and abuse of dominant position contained in such laws as the Sherman Act in the United States, Articles 85(1) and 85(2) of the Treaty of Rome in the European Union, and the Act of Prohibition of Private Monopolization and Maintenance of Fair Trade in Japan. Regulators in the European Union, the United States, and China can block mergers that would unduly limit competition even if neither merger partner is domiciled in that jurisdiction.

Environmental Laws

In response to high-cost hazardous waste spills and clean-ups, a number of jurisdictions have adopted laws governing the storage, transport, and disposal of hazardous waste. The European Union and Taiwan have adopted the policy that the polluter pays, with the objectives of protecting the quality of the environment and human health and promoting the prudent and rational utilization of natural resources. The United States imposes liability not only on the person who stored, transported, or disposed of the waste, but also on the owner or operator of the property at the time the waste is discovered. Increasing concerns about climate change have led to stricter limits on the release of greenhouse gases, such as carbon dioxide.

Consumer Protection

Jurisdictions continue to adopt laws and regulations to protect consumers from deceptive sales practices and advertising. The rise of the Internet and social networks, as well as the ease

with which electronic databases can be used to collect personal data, have sparked new concerns about consumer privacy and more stringent government oversight.

Financial Regulation

The United States has the most restrictive rules regarding the offer and sale of securities, such as common and preferred stock and bonds. Generally, any public offering of securities must be made pursuant to a registration statement declared effective by the Securities and Exchange Commission (SEC). The most important part of the registration statement is the prospectus, which contains detailed information about the company issuing the securities and its management, principal shareholders, business, and financial condition. The Jump-start our Business Startups Act of 2012 (the JOBS Act) created new exemptions for public offerings using the Internet and reduced the disclosure requirements for "emerging growth companies," defined as firms with less than $1 billion in annual revenues. Manchester United relied on these new provisions when it went public on the New York Stock Exchange in 2012. The US laws provide civil and criminal penalties for any offer or sale of a security, regardless of whether it is registered, involving a misstatement of a material fact or an omission that makes the facts stated materially misleading. SEC Regulation S facilitates the offering of securities outside of the United States by US firms by exempting many such offerings from the burdensome US registration requirements that would be applicable if the securities were sold within the United States.

The subprime mortgage crisis of 2008, which contributed to a global financial crisis, led to more stringent regulation of banks and the other financial institutions. They include heightened net capital requirements and greater regulation of financial derivatives, such as credit-default swaps.

Bankruptcy Law

The principle that a debtor should be permitted to discharge certain debts has ancient roots. As early as 1400 BC, Mosaic Law required unconditional forgiveness of debts every seven years. The British Parliament under King Henry VIII passed a Bankruptcy Act in AD 1542 to regulate failing English merchants.

A key objective of bankruptcy law is to ensure that creditors and equity holders of financially distressed businesses are treated fairly in a judicial setting designed to ensure the efficient handling of claims. In general, creditors must be paid in full before equity holders receive anything.

The United States provides perhaps the most prodebtor bankruptcy relief available. An insolvent firm in the United States generally can elect to liquidate its assets or to reorganize the business, often with the debtor still in control. In contrast, in Australia, creditors can force the liquidation of a company that cannot pay its debts even if the debtor argues that the business would be worth more as an operating whole. Some countries, such as Germany, do not provide for the complete discharge of debts in bankruptcy, only arrangements for their payment over time.

See also: Area and International Studies: Political Economy; Corporate Finance: Financial Control; Corporate Governance; Corporate Law: Publicly Traded Corporations; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; Law and Economics: Empirical Dimensions; Law and Economics; Monetary Policy Since the 2007–2009 Financial Crisis; Regulation, Economic Theory of; Venture Capital; Working Conditions.

Bibliography

Bagley, C.E., 2008. Winning legally: the value of legal astuteness. Academy of Management Review 33, 378–390.

Bagley, C.E., 2010. What's law got to do with it?: Integrating law and strategy. American Business Law Journal 47, 587–639.

Bagley, C.E., 2013. Managers and the Legal Environment: Strategies for the 21st Century, seventh ed. South-Western, Cengage Learning, Mason, OH.

Bagley, C.E., Page, K.L., 1999. The devil made me do it: replacing corporate directors' veil of secrecy with the mantle of stewardship. San Diego Law Review 36, 897–945.

De Soto, H., 2000. The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else. Basic Books, New York.

Duncan Jr., J.C., 2000. Nachfrist was ist? Thinking globally and acting locally: considering the extension principles of the U.N. convention on contracts for the international sale of goods in revising the uniform commercial code. Brigham Young University Law Review 2000, 1363–1411.

Eagan, W., 1980. The Westinghouse uranium contracts: commercial impracticability and related matters. American Business Law Journal 18, 281–302.

Edelman, L.B., Suchman, M.C., 1997. The legal environments of organizations. Annual Review of Sociology 23, 479–515.

Farnsworth, E.A., 1987. Precontractual liability and preliminary agreements: fair dealing and failed negotiations. Columbia Law Review 87, 217–294.

Gordley, J., 1981. Equality in exchange. California Law Review 69, 1587–1656.

Hadley v. Baxendale, 1854. 156 Eng. Rep. 145 (Exch.).

North, D.C., 1990. Institutions, Institutional Change and Economic Performance. Cambridge University Press, Cambridge, MA, pp. 3–4.

North, D.C., Weingast, B.R., 1989. Constitutions and commitments: the evolution of institutions governing public choice in seventeenth-century England. The Journal of Economic History 49, 803, 831.

Porter, M.E., 2002. Enhancing the microeconomic foundations of prosperity: the current competitiveness index in World Economic Forum. In: Porter, M.E. (Ed.), The Competitiveness Report 2001–2002.

Raffles v. Wichelhaus, 1864. 159 Eng. Rep. 375 (Exch.).

Suchman, M.C., Steward, D.J., Westfall, C.A., 2001. The legal aspects of entrepreneurship: observations on the legitimization of venture finance in Silicon Valley. In: Schoonhoven, C.B., Romanell, E. (Eds.), The Entrepreneurship Dynamic: Origins of Entrepreneurship and the Evolution of Industries. Stanford University Press, Stanford, CA.

Texaco, Inc. v. Pennzoil Co., 1987. 729 S.W. 2d 768 (Tex. Ct. App.).

Tsakiroglou & Co. Ltd. v. Noblee Thorl G.m.b.H., 1960. 2 Q.B. 318, 348, *aff'd*, [1962] A.C., 93(1961).

Wood v. Boynton, 1885. 25 N.W. 42 (Wis.).

Yanagida, Y., Foote, D.H., Johnson Jr., E.S., Ramseyer, J.M., Scogin Jr., H.T., 1994. Law and Investment in Japan: Cases and Materials. Harvard University Press, Cambridge, MA.

Business Models

Raphael Amit, University of Pennsylvania, Philadelphia, PA, USA
Christoph Zott, IESE Business School, Barcelona, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A firm's business model is a template that depicts the way the firm conducts its business. It describes the system of interdependent activities that are performed by the firm and by its partners and the ways that these activities are linked to each other through transactions in factor and product markets. In this article, we focus on the evolution of business model research during the last 10 years and explain the various ways in which the business model has been conceptualized. We review scholarly work on business models within the fields of strategic management, entrepreneurship, and innovation.

Introduction

A business model is crafted by a focal firm's managers in order to best meet the perceived needs of their customers. The business model is a template that depicts the way the firm conducts its business. It describes the system of interdependent activities that are performed by the firm and by its partners and the mechanisms that link these activities to each other.

An activity in a focal firm's business model can be viewed as the engagement of human, physical, and/or capital resources of any party to the business model (the focal firm, end customers, vendors, etc.) to serve a specific purpose toward the fulfillment of the overall objective. An activity system is a set of interdependent and interconnected activities that are centered on a focal firm; it encompasses activities that are conducted either by the focal firm or by partners, customers, or vendors. To fully address the market opportunity, the firm's activity system may transcend the focal firm and span across the firm and its industry boundaries, but it remains firm-centric, enabling the focal firm not only to enhance total value with its partners, but also to appropriate a share of the value created.

A focal firm's business model thus has two objectives: (1) to create value for all business model stakeholders (commonly referred to as the 'value proposition'); that is, to fulfill customers' needs and create customer surplus while generating a profit for the focal firm and all its partners (Amit and Zott, 2001), and (2) to appropriate value for the focal firm.

Later in the article, we briefly describe various approaches toward conceptualizing business models. We then review some key contributions to business model research in the areas of strategic management, entrepreneurship, and innovation. We conclude with a future research agenda.

Business Model Conceptualizations: A Brief Overview

At this stage in the evolution of the business model concept, a unified view has not yet emerged, as different conceptualizations of the business model emphasize different aspects of wealth creation (Zott et al., 2011). Broadly speaking, some concepts center on value creation, while others focus more on value appropriation through, for example, revenue generation. The value creation perspective of the business model has been advanced in a series of articles by Amit and Zott (2001), Zott and Amit (2007, 2008) to capture the essence of 'how firms do business.' These studies have begun to describe the business model as a source of innovation, for example, when it connects previously unconnected parties, links transaction participants in new ways, or introduces new transaction mechanisms. McGrath (2010) concurs with this value creation view of the business model. According to her definition, business models include two core managerial choices: units of business (i.e., what a firm is selling that someone is prepared to pay for) and the set of activities employed to sell those units.

Other authors have proposed a more encompassing definition of the term, with greater emphasis on value appropriation (e.g., Casadesus-Masanell and Ricart, 2010). Chesbrough and Rosenbloom (2002: 529), for example, link the business model to the technology management literature and define it as the "heuristic logic that connects technical potential with the realization of economic value," emphasizing its role in linking technology to market outcomes. Consistent with this perspective, Casadesus-Masanell and Ricart (2010) posit that one important component of business models is the set of choices made by management regarding how the organization operates, such as compensation practices, procurement contracts, location of facilities, or assets employed. Another component of business models, according to this view, is reflected in the consequences of these choices, such as low cost or a culture of frugality.

Despite the absence of a clear consensus in the literature about the concept definition, researchers have begun to converge on some common themes characterizing business models (Zott et al., 2011): (1) business models center on the logic of how value is created for all stakeholders, not just how it is captured by the focal firm; (2) activities performed by the focal firm as well as by partners, suppliers, and even customers play an important role in the various conceptualizations of business models that have been proposed; (3) business models emphasize a system-level, holistic approach toward explaining how firms 'do business'; and (4) the business model is emerging as a new level and unit of analysis. Together, these emerging themes point toward the business model as a concept that extends that of the value chain (Porter, 1985) by (1) emphasizing value creation and delivery dynamics, (2) spanning firm and industry boundaries, and (3) allowing for a nonlinear sequencing of interdependent activities. Although the notion of the business model draws on arguments that are central to the value chain framework, in

particular on the ideas that activities and multiple sources of value matter, it extends those arguments in important ways.

The business model concept became prevalent with the development of the Internet in the mid-1990s. Since that time, the concept has resonated first with practitioners and then with academics in such fields as strategic management, entrepreneurship, and innovation.

Business Models and Strategic Management

Scholars have emphasized that the business model can play an important role for a firm's strategy. Firm strategy could indeed be viewed as having two broad components: the firm's competitive positioning and its underlying business model, with the latter complementing the former. In support of this argument, Zott and Amit (2008) examine the possible contingent effect of business model design in mediating between product market strategy and firm performance. They ask: How do the firm's business model and product market strategy interact to impact the firm performance? They find that business models that emphasize novelty and are coupled with either differentiation or cost leadership product market strategies can have a positive impact on the firm's performance; and novelty-centered business models together with early entry into a market have a positive effect on performance. Thus, business model design and product market strategy are complements, not substitutes (Zott and Amit, 2008).

According to Teece (2007, 2010), the business model reflects a "hypothesis about what customers want, and how an enterprise can best meet those needs, and get paid for doing so" (2007: 1329); it "embodies nothing less than the organizational and financial 'architecture' of the business" (2010: 173). However, is the business model part of a firm's strategic repertoire (i.e., a strategic tool in and of itself) or is it an expression of the firm's strategy? According to Richardson (2008), the business model explains how the activities of the firm work together to execute its strategy, thus bridging strategy formulation and implementation. In a similar vein, Casadesus-Masanell and Ricart (2010) view the business model as a reflection of a firm's realized strategy.

An important factor of interest to strategy scholars is the focus of the business model concept on the value proposition to its stakeholders and a generalized emphasis on the role of the customer, which appears to be less pronounced elsewhere in the strategy literature. A common view seems to have emerged that the business model revolves around customer-focused value creation (Chesbrough and Rosenbloom, 2002; Teece, 2007, 2010; Zott et al., 2011). Viewed from this perspective, the business model embodies the essential details of a firm's value proposition for its various stakeholders, as well as the activity system the firm uses to create and deliver value to its customers (Zott and Amit, 2010).

Increasingly scholars also acknowledge that firms do not execute their business models in a competitive vacuum, and that they can indeed compete through their business models (Casadesus-Masanell and Ricart, 2010). The business model, then, represents a potential source of competitive advantage, one that is distinct from the firm's product market position (Zott and Amit, 2008).

Business models can thus play a central role in explaining firm performance. They complement a firm's product market strategy (Zott and Amit, 2007), embody its value proposition to customers and other stakeholders, and describe how that value is created and delivered through the firm's (possibly boundary spanning) activity system. Correspondingly, Afuah and Tucci (2001: 3) propose the business model as a unifying construct for explaining competitive advantage and firm performance and define it as "the method by which a firm builds and uses its resources to offer its customer better value and to make money in doing so."

Business Models and Entrepreneurship

In addition to importance in the field of strategic management, business models are also relevant for entrepreneurship. The entrepreneur could be portrayed as the architect of the activity system of a new venture, as a creative designer (and, potentially, innovator) of economic exchanges, a description which complements received notions of entrepreneurs as product and process innovators. The business model concept is thus central to entrepreneurship.

Consistent with the theoretical development by Amit and Zott (2001), entrepreneurs' business models can create value through efficiency (emphasized by transaction costs economics), novelty (Schumpeterian innovation), complementarities (advocated by resource-based theory), and lock-in (inherent in strategic networks). The emphasis of the business model perspective on value creation calls for integrating important entrepreneurship and strategy theories and potentially heightens our understanding of the complex processes and mechanics that drive wealth creation.

Zott and Amit (2007, 2008) show empirically that the design of the business model is indeed central to wealth creation through entrepreneurship. In particular, Zott and Amit (2007) have analyzed the performance implications for entrepreneurial firms of business model designs by considering two distinct effects: the total value creation potential of the business model and the focal firm's ability to appropriate that value. They hypothesize that the greater the total value created, and the greater the entrepreneurial firm's bargaining power, the greater the amount of value that the focal firm can appropriate. How much of the possible total value is actually captured, however, depends on the revenue model chosen by the entrepreneur. The researchers find through large sample, cross-sectional empirical analyses of start-ups that introducing novel business models positively influences entrepreneurial firm performance, even when the environment switches from resource-rich to resource-poor.

In line with these findings, Hargadon and Douglas (2001) point to the crucial role of the design of the business system for entrepreneurial innovation. Such design is centered on "a particular set of understandings and interactions" (2001: 494) between different stakeholders, such as the firm and its customers. As Romme (2003) notes, the purpose of design is to produce systems that do not yet exist – a key task for entrepreneurs. Indeed, in (re-)designing their business models, entrepreneurs throughout the world, in new as well as existing companies, have been leveraging recent technological

developments. Among other things, this has contributed to a shift in the locus of competitive advantage from the venture and its internal stakeholders (e.g., management, shareholders, and employees) to its activity system, which encompasses external stakeholders such as partners, vendors, and customers. Thus, both new and established firms have become increasingly interested in innovating their business models.

Business Models and Innovation

Rapid advances in information and communication technologies have facilitated new types of technology-mediated interactions between economic agents and enabled firms to fundamentally change the ways they do business (Bock et al., 2012; Chesbrough, 2010; Dewald and Bowen, 2010; Sako, 2012). Because of that, senior managers of focal firms have more possibilities in how they pursue innovation. For example, besides innovating products, processes, and management practices, they can now also innovatively design new business models along various dimensions, for example, construct, link, and/or sequence boundary-spanning exchanges and activities (Amit and Zott, 2012). For instance, Chesbrough (2010) argues that business model innovation provides a tool for value creation, complementing new technologies. He proposes ways to overcome business model innovation barriers through experimentation and successful leadership of organizational change. Bock et al. (2012) analyze the perceptions of 107 CEOs and find that business model innovation effort in companies positively moderates the relationship between activity reconfiguration and strategic flexibility. Dewald and Bowen (2010) analyze the reactions of incumbents to the introduction of new business models and find that incumbent managers are less likely to adopt the new business model when innovation is perceived as a threat than when it is perceived as an opportunity. Sanchez and Ricart (2010) explore business model innovation in low-income markets and distinguish between isolated and interactive new business models introduced by firms in those markets.

Business model innovation seems to be of growing importance. A study from the Economist Intelligence Unit (2005: 9) found that a majority of the more than 4000 senior managers who were surveyed worldwide favored new business models over new products and services as a source of future competitive advantage. The study stated that, "the overall message is clear: how companies do business will often be as, or more, important than what they do." In a similar study conducted by IBM (Pohle and Chapman, 2006) in which over 750 corporate and public sector leaders from around the world were interviewed on the subject of innovation, "one key finding was that competitive pressures have pushed business model innovation much higher than expected on CEOs' priority lists" (2006: 35). Indeed, the study found that outperforming companies whose operating margins had grown faster than their competitors' over the previous 5 years were twice as likely as their lower performing peers to emphasize business model innovation, as opposed to product or process innovation.

Much of the prior research on business model innovation has considered the extent to which business models are *de novo*, that is, new to the state-of-the-art, and not just new to the firm (Birkinshaw et al., 2008). Nidumolu et al. (2009) view the development of new business models as a key step in their five-stage model of corporate transformation to become environmentally sustainable. They declare as the central challenge "to find novel ways of delivering and capturing value, which will change the basis of competition" (2009: 60). According to the authors, opportunities for business model innovation lie in developing new delivery technologies that change the value chain, in combining digital and physical infrastructures, or in turning products into services. Similarly, Johnson et al. (2008) focus on *de novo* business models, based on the belief that there is "no point in instituting a new business model unless it is not only new to the company, but in some way game-changing to the industry or market" (2008: 58). Changes to business model design, however, can also be subtle and new to the firm; they may not have the potential to disrupt an industry, but could still yield important benefits to the business model innovator (i.e., the focal firm).

The novelty presented by new business models can result in superior value creation and replace the old way of doing things to become the standard for the next generation of entrepreneurs to beat. Some scholars have suggested very broad domains for business model innovation, in line with their corresponding definitions of the business model concept. Mitchell and Coles (2003), for example, propose that business model innovation involves modifications in the 'who,' 'what,' 'when,' 'why,' 'where,' 'how,' or 'how much' involved in providing products and services to customers. Similarly, Johnson et al. (2008) describe business model innovation as involving the firm's value proposition, target customers, product and service offering, resources (such as people, technology, or equipment), revenue model, cost structure, processes, rules, and norms.

The business model can thus represent a new subject of innovation, which complements the traditional subjects of process, product, and organizational innovation, and involves new forms of cooperation and collaboration. It can, however, also simply help firms commercialize innovative ideas and technologies by unlocking the value potential embedded in new technologies and converting it into market outcomes. Chesbrough and Rosenbloom (2002) detail an extensive case study, in which they show how the Xerox Corporation grew in part by employing an effective business model to commercialize a technology rejected by other leading companies. The study also compares successful and unsuccessful technology spin-offs with comparable market potential and finds that in successful ventures the search and learning for an effective business model was significantly higher than in failed ventures. Björkdahl (2009) argues that the integration of new technologies into the technology base of a product (i.e., technology cross fertilization) can open up new subspaces in the existing technical performance and functionality space, which in turn requires a new business model if the economic value potential of the new technology is to be captured.

Business models can not only entail consequences for technological innovations, they are also shaped by them. Technological innovation can trigger changes in the company's operational and commercial activities, and hence in the business model.

Opportunities for Future Research

Research on business models needs to be advanced through both theory development and empirical analysis. Increasing consensus on the theoretical foundations, the definition, and the fundamental properties of business models should lead to the emergence of broadly accepted concepts and typologies, which are currently lacking. Such research will help deepen our understanding of the linkages between value creation and value appropriation. Empirical research on the measurement of business model innovation, structured to capture all lines of a firm's business that have revenue potential, holds great promise to enhance our understanding of business models. Examining the dynamics of business model evolution, how they emerge, and how they are shaped and adapted by entrepreneurial actors over time – as well as how business models coevolve with product market strategy and organization design – reflects an important cross-disciplinary research program that will substantially solidify the business model as a pivotal unit of analysis in our understanding of value creation and capture.

Conclusion

We are still in the early stages of developing the business model as a new unit and level of analysis for strategy, entrepreneurship, and innovation research. Theoretical, empirical, and field research on the foundations and evolutions of business models promises to broaden our understanding of this important concept.

Acknowledgments

We are grateful to Jay Barney and to Marvin Lieberman for their insightful suggestions. We also acknowledge the excellent research assistance of Yuliya Snihur.

See also: Business Platforms; Entrepreneurship and Economic Growth; Hierarchies and Markets; Strategic Management; Strategizing.

Bibliography

Afuah, A., Tucci, C., 2001. Internet Business Models and Strategies: Text and Cases. Irwin/McGraw-Hill, New York.

Amit, R., Zott, C., 2001. Value creation in e-business. Strategic Management Journal 22, 493–520.

Amit, R., Zott, C., 2012. Creating value through business model innovation. Sloan Management Review 53, 41–49.

Birkinshaw, J., Hamel, G., Mol, M., 2008. Management innovation. Academy of Management Review 33, 825–845.

Björkdahl, J., 2009. Technology cross fertilization and the business model: the case of integrating ICTs in mechanical engineering products. Research Policy 38, 1468–1477.

Bock, A., Opsahl, T., George, G., Gann, D.M., 2012. The effects of culture and structure on strategic flexibility during business model innovation. Journal of Management Studies 49 (2), 279–305.

Casadesus-Masanell, R., Ricart, J.E., 2010. From strategy to business models and to tactics. Long Range Planning 43, 195–215.

Chesbrough, H., 2010. Business model innovation: opportunities and barriers. Long Range Planning 43 (2), 354–363.

Chesbrough, H., Rosenbloom, R., 2002. The role of the business model in capturing value from innovation: evidence from Xerox Corporation's technology spinoff companies. Industrial and Corporate Change 11, 529–555.

Dewald, J., Bowen, F., 2010. Storm clouds and silver linings: responding to disruptive innovations through cognitive resilience. Entrepreneurship: Theory & Practice 34 (1), 197–218.

Economist Intelligence Unit Report, 2005. Business 2010: Embracing the Challenge of Change.

Hargadon, A.B., Douglas, Y., 2001. When innovations meet institutions: Edison and the design of the electric light. Administrative Science Quarterly 46, 476–501.

Johnson, M.W., Christensen, C.M., Kagermann, H., 2008. Reinventing your business model. Harvard Business Review 86, 58–68.

McGrath, R., 2010. Business models: a discovery driven approach. Long Range Planning, Special Issue on Business Models 43 (2–3), 247–261.

Mitchell, D., Coles, C., 2003. The ultimate competitive advantage of continuing business model innovation. Journal of Business Strategy 24 (5), 15–21.

Nidumolu, R., Prahalad, C.K., Rangaswami, M.R., 2009. Why sustainability is now the key driver of innovation. Harvard Business Review 87 (9), 57–64.

Pohle, G., Chapman, M., 2006. IBM's global CEO report 2006: business model innovation matters. Strategy & Leadership 34 (5), 34–40.

Porter, M.E., 1985. Competitive Advantage: Creating and Sustaining Superior Performance. Free Press, New York.

Richardson, J., 2008. The business model: an integrative framework for strategy execution. Strategic Change 17 (5–6), 133–144.

Romme, G.A., 2003. Making a difference: organization as design. Organization Science 14, 558–573.

Sako, M., 2012. Business models for strategy and innovation. Communications of the ACM 55 (7), 22–24.

Sanchez, P., Ricart, J.E., 2010. Business model innovation and sources of value creation in low-income markets. European Management Review 7, 138–154.

Teece, D.J., 2007. Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance. Strategic Management Journal 28, 1319–1350.

Teece, D.J., 2010. Business models, business strategy and innovation. Long Range Planning 43, 172–194.

Zott, C., Amit, R., 2007. Business model design and the performance of entrepreneurial firms. Organization Science 18, 181–199.

Zott, C., Amit, R., 2008. The fit between product market strategy and business model: implications for firm performance. Strategic Management Journal 29, 1–26.

Zott, C., Amit, R., 2010. Designing your future business model: an activity system perspective. Long Range Planning 43, 216–226.

Zott, C., Amit, R., Massa, L., 2011. The business model: recent developments and future research. Journal of Management 37, 1019–1042.

Business Platforms

Annabelle Gawer, Imperial College Business School, London, UK
Michael A Cusumano, MIT Sloan School of Management, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the literature on platforms and sheds light on the design principles, economics fundamentals, and business strategies associated with platforms. We divide the discussion into two main types: product platforms designed by individual companies or assemblers of closed supply chains to develop new products or services on the basis of common and reusable components and architectures, and industry platforms which act as a foundation for an ecosystem of firms to develop a set of interrelated products and services. Present in a variety of organizational contexts, platforms create value for innovative ecosystem participants by structuring the innovation process around core and complementary elements and by creating the network effects that accelerate the adoption and use of platforms. With increased globalization of innovation capabilities, we expect the phenomenon of platforms to become an important new form of industrial organization. We conclude with highlighting concerns associated with platform dominance.

Business platforms, both internal to the firm and at the industry level, have become increasingly important in a large number of markets, especially with high-technology products that require complementary innovations from various firms to be truly valuable to end users. There is considerable confusion, however, as to what the term 'platform' refers to in practice and in management-related research. This article reviews different types of platforms commonly discussed in the management literature and attempts to define more precisely the characteristics of a specific types of platforms. We also discuss managerial practices associated with effective platform leadership as well as gaps in our current knowledge and avenues for future research. (This article draws extensively on prior writings by the authors, especially Gawer, 2009b and Gawer and Cusumano, 2013, 2014.)

Main Types of Platforms

What managers and researchers refer to as platforms exist in a variety of industries, especially in high-technology businesses that involve multiple products and services. For example, Microsoft, Apple, Google, Intel, Cisco, ARM, Qualcomm, EMC, and many other firms design or build hardware and software products and components for computers, cell phones, and consumer electronics devices. Many of these products have been referred to in the business press and in scholarly literature as 'platforms' or key platform elements (Gawer and Cusumano, 2002; Eisenmann et al., 2006; Gawer, 2009a). In addition, these firms and their partners as well as customers are also said to participate in platform-based 'ecosystems' (Moore, 1996). While the term platform is used in many ways, most scholars seem to agree that platforms are distinct from conventional products in that they have the potential to generate positive feedback loops, also called 'network effects,' between the primary product (the platform) and its users as well as with complementary products and services. In other words, the more users who adopt the platform, and the more third-parties that create complementary products and services, the more valuable the platform becomes to the owner and to the users. The positive feedback loops create increasing incentives for more users and more complementors to adopt a particular platform and join a particular ecosystem.

Platforms can also operate in relation to other platforms. For example, the Intel x86 line of microprocessors have served as essential elements of the personal computer platform, with Windows 'sitting on top' to control the hardware as well as create a foundation for launching applications and other functions such as accessing the Internet through a browser. Other microprocessor designs and operating systems form key elements of smartphone platforms dominated by Google and Apple to access the Internet and the phone network or run a variety of applications. Search engines such as from Google and Microsoft as well as social media networks such as from Facebook, Skype, and LinkedIn exist on top of the Internet.

A number of researchers have studied platforms from various perspectives. For example, we have research on new product development from the innovation and operations management fields (Meyer and Lehnerd, 1997); research on platform-based competition in the strategy field (Gawer and Cusumano, 2002, 2008; Eisenmann et al., 2006); and analyses of the economic implications of platform competition and dynamics in the economics field (Rochet and Tirole, 2003; Armstrong, 2006). But our review of the literature and numerous industry examples suggests that these and other scholars have largely been describing two predominant forms of platforms: product platforms, used mainly for product development, and industry platforms, used to create value and innovations at the ecosystem level. We can consider both types as 'business platforms' in that the primary actors, apart from individual users, are firms or networks of firms functioning as suppliers and partners.

In short, we define product platforms as a set of assets organized in a common structure from which a firm can efficiently develop and produce a stream of derivative products (Muffatto and Roveda, 2002). We define industry platforms as products, services, or technologies that are similar in some ways

to the former but provide the foundation upon which an entire industry ecosystem, including both the platform owner and third-party firms as well as users can develop their own complementary products, technologies, or services (Gawer and Cusumano, 2002; Gawer, 2009b). The following sections elaborate on these distinctions as well as the importance of network effects to the second type of platform.

Product Platforms

The first popular usage of the term platform seems to have been in the context of new product development and incremental innovation around reusable components or technologies. We refer to these as product platforms in that a firm, either working by itself or with suppliers, can build a family of related products or sets of new features by deploying these components. For example, Wheelwright and Clark (1992) describe how various companies developed 'product platforms' to meet the needs of different customers simply by modifying, adding, or subtracting different features. Other research in this vein tends to define these kinds of platforms as subsystems and interfaces that form a common structure from which a company can efficiently develop and produce a family of products, such as new automobiles or consumer electronics devices. Robertson and Ulrich (1998) propose an even broader definition, viewing platforms as the collection of assets (i.e., components, processes, knowledge, people, and relationships) that a set of products share. In the marketing literature, Sawhney (1998) even suggests that managers should move from 'portfolio thinking' to 'platform thinking,' which he defines as understanding the common strands that tie the firm's offerings, markets, and processes together, and exploit these commonalities to create leveraged growth and variety.

These and other researchers have identified, with a large degree of consensus, several potential benefits of product platforms: savings in fixed costs; efficiency gains in product development through the reuse of common parts and 'modular' designs, in particular, the ability to produce a large number of derivative products with limited resources; and flexibility in product feature design. One key objective of platform-based new product development seems to be the ability to increase product variety and meet diverse customer requirements, business needs, and technical advancements while maintaining economies of scale and scope within manufacturing processes.

The empirical evidence indicates that, in practice, companies have successfully used product platforms to increase product variety, control high production and inventory costs, and reduce time to market. Most of the early research is about durable goods, whose production processes involve manufacturing, such as in the automotive, aircraft, equipment manufacturing, and consumer electronics sectors (Sanderson and Uzumeri, 1997; Cusumano and Nobeoka, 1998). Researchers have also identified a few fundamental design principles or 'design rules' that appear to operate in product platforms, in particular the stability of the system architecture, and the systematic or planned reuse of modular components (Baldwin and Woodard, 2009). We can see as well a fundamental trade-off couched in terms of functionality and performance: the optimization of any particular subsystem may result in the suboptimization of the overall system (Meyer and Lehnerd, 1997). In this sense, product platforms may promote only incremental innovation or constrain some types of innovation – a theme that we will return to later in this article.

We also view supply chains as a special kind of product platform. In a supply chain, a set of firms follow specific guidelines to supply intermediate products or components to the product owner or the final product assembler (Doran, 2004). The objective is also to improve efficiency and reduce cost such as by the systematic reuse of modular components. Major potential benefits are that a firm with access to a platform supply chain can tap into external capabilities to find more innovative or less expensive components and technologies. At the same time, the product or platform owner may have less control over the components and technology, which can have its own negative consequences. Supply chain platforms are commonly discussed with regard to assembly industries such as consumer electronics, computers, and automobiles. We can also link this literature to other research on industry architecture (Jacobides et al., 2006).

The research suggests that there are distinctions between supply chains and industry platforms. One is that, in the latter case, the firms developing the complementary innovations – such as applications for Windows or the Apple App Store – do not necessarily buy from or sell to each other. Nor are they usually part of the same supply chain or share patterns of cross-ownership, such as Toyota does with its major component suppliers. There also are no obvious network effects between the main product or platform and components in a supply chain platform, though when this does occur, we are probably viewing an embedded industry platform. For example, an auto parts supplier might develop a new fuel-injection system and provide it to a particular automaker. Other automakers then decide to adopt this technology and redesign their engines to work with this new fuel-injection system. Independent software companies start providing new programing applications that optimize performance of the fuel-injection system under different driving conditions. Other parts suppliers redesign their components to work with the new fuel-injection system. The fuel-injection system, once simply part of one firm's supply chain, has gradually become an industry-wide platform within the automobile industry, with its value considerably enhanced by the efforts of multiple firms.

Industry Platforms

We have defined industry platforms as products, services, or technologies developed by one or more firms, and which serve as foundations upon which a larger number of firms, organized in an ecosystem, can build further complementary innovations and potentially generate network effects. There is a similarity to product platforms in that industry platforms provide a foundation of reusable common components or technologies, but they differ in that this foundation is 'open' to outside firms. The degree of openness can also vary on a number of dimensions – such as level of access to information on interfaces to link to the

platform or utilize its capabilities, the type of rules governing use of the platform, or cost of access (as in patent or licensing fees).

There are some similarities between the concept of an industry platform and that of a dominant design. A 'dominant design' (Abernathy and Utterback, 1978), when it emerges, sets the standard for what form and features users expect a particular product to take in the future. The two concepts differ, however, in that the purpose of an industry platform is to bring multiple parties within the industry together – primarily users and complementors. Moreover, like product company platforms, industry platforms do not simply emerge without deliberate, firm-driven agency, or deliberate managerial decisions and actions. And in platform markets – those industries characterized by a foundation technology around which third-party firms create complementary innovations, with adoption driven by positive feedback loops and network effects – the most likely winner is not necessarily the originator of the dominant design or the owner of the most elegant product. The most likely winner is the owner of the 'best' platform, the characteristics of which we discuss later in this section.

Despite different degrees of openness to outside complementors, various products and technologies have served as industry platforms: the Microsoft Windows and Linux operating systems; Intel and ARM microprocessor designs; Apple's iPod, iPhone, and iPad designs along with the iOS operating system; Apple's iTunes and App Store; Google's Internet search engine and Android operating system for smartphones; social networking sites such as Facebook, LinkedIn, and Twitter; video game consoles; and the Internet itself. We can even view payment technologies, ranging from credit and debit cards to micropayment schemes, as platforms that enable different types of financial transactions.

Early research on industry platforms and their innovation ecosystems generally focused on computing, telecommunications, and other information-technology intensive industries. For example, Bresnahan and Greenstein (1999), in their study of the computer industry, analyzed platforms as a bundle of standard components around which buyers and sellers coordinated their efforts. West (2003) defined a computer platform as an architecture of related standards that allowed modular substitution of complementary assets such as software and peripheral hardware. Gawer and Henderson (2007) described a product as a platform when it is one element or subsystem of an evolving technological system, when it is strongly functionally interdependent with most of the other components of this system, and when end-user demand is for the overall system, so that there is no demand for components when they are isolated from the overall system.

Taken together, these studies suggest several generalizations with regard to what makes for the best industry platform and how this can affect competitive dynamics as well as innovation at the ecosystem level. Positions of industrial leadership are often contested and lost when industry platforms emerge, as the balance of power between assemblers and component-makers changes. At the same time, industry platforms tend to facilitate and increase the degree of innovation on complementary products and services. The more innovation there is on complements, the more value it creates for the platform and its users via network effects, creating a cumulative advantage for existing platforms: As they grow in adoption, they become harder to dislodge by rivals or new entrants, with the growing number of complements acting like a barrier to entry. The rise of industry platforms may also raise complex social welfare questions regarding trade-offs between the social benefits of platform-compatible innovation versus the potentially negative effects of preventing competition on overall systems.

It follows that the design principles or 'design rules' for industry platforms overlap somewhat with those for product and supply chain platforms, but go beyond them to serve a larger purpose. For example, the stability of the platform architecture is still essential, but there are important differences. In contrast to what happens for product and supply chain platforms, the logic of design for industry platforms is inverted. Instead of a firm being a 'master designer' or assembler, here we start with a core component that is part of an encompassing modular structure, and the final result of the assembly is either unknown *ex ante* or incomplete. In fact, for truly successful industry platforms, the end-use of the end product or service does not seem to be fully predetermined by the platform owner. This creates unprecedented scope for innovation on complementary products, services, and technologies. The situation simultaneously evokes the fundamental question of how incentives (for third-parties) to innovate can be embedded in the design and governance of the platform.

This leads to another apparent design rule for effective industry platforms: The interfaces around the platform should be sufficiently 'open' to allow outside firms to 'plug in' complements as well as innovate on these complements and make money from their investments. Although this resonates well with research on open innovation, recent research on platforms highlights the complex trade-offs between 'open' and 'closed' (Greenstein, 2009). This research suggests that, while opening up interfaces should increase complementors' incentives to innovate, it is important to preserve as proprietary some source of revenue and profit. It therefore adds a more subtle take on the literature on open innovation that had extolled the simple benefits of opening interfaces.

Specific strategic questions also arise in the context of industry platforms. For example, Gawer and Cusumano (2008) argue that not all products, services, or technologies can become industry platforms. To perform this industry-wide role and convince other firms to adopt the platform as their own, the platform must (1) perform a *function* that is essential to a broader technological system and (2) solve a *business* problem for many firms and users in the industry. While necessary, these conditions alone are still not sufficient to help firms transform their products, technologies, or services into industry platforms, nor indicate how platform leaders can stimulate complementary innovations by other firms, including some competitors, while simultaneously taking advantage of owning the platform.

One particular managerial challenge is that platform leaders and competitors must navigate a complex strategic landscape where both competition and collaboration occur, sometimes among the same actors. For example, as a technology evolves, platform owners often face the opportunity to extend the scope of their platform and integrate into complementary markets. This creates disincentives for complementors to invest in

innovation in these complementary markets. Farrell and Katz (2000) identified the difficulty for platform owners to commit, not to squeeze, the profit margins of their complementors. Gawer and Henderson (2007) show how Intel's careful selection of which complementary markets to enter (the connectors) while giving away corresponding intellectual property allowed the firm to push forward the platform/applications interface. Intel thereby retained control of the architecture at the same time it renewed incentives for complementors to innovate 'on top of' the newly extended platform.

Another challenge is that, as technology is constantly evolving, platform leaders need to make business decisions and technology or design decisions in a coherent manner. For example, consider a firm that designs open interfaces to its platform: this will stimulate innovation on complements, and the firms that will act as complementors by designing these complements need therefore to be treated by the focal firms as allies, not potential competitors. This will mean that the focal firm should probably refrain from entering as a competitor in complementors' markets if it wishes to sustain the complementors' incentives to innovate. This need for coherence across business decisions and technological design decisions can be difficult to achieve since these decisions are often made by different teams within the organization. The coherence imperative requires top management's awareness of the interdependencies between these decisions, and the right product process in place to allow ongoing coordination across functional silos. Hence, to make the whole greater than the sum of the parts, we can see the need in many complex systems industries for one firm or a small group of firms to act as a 'platform leader' (Gawer and Cusumano, 2002).

Network Effects and Multisided Markets

Perhaps the most critical distinguishing feature of an industry platform compared to a product company platform or supply chain is the potential creation of network effects. As mentioned earlier, these are positive feedback loops that can grow at exponentially increasing rates as adoption of the platform and the number of complements rise. The network effects can be very powerful, especially when they are 'direct' (sometimes called 'same-side') between the platform and the user of the complementary innovation and reinforced by a technical compatibility or interface standard that makes using multiple platforms ('multihoming') or switching from one platform to another difficult or costly. For example, VHS tapes and Windows or Apple iOS applications only work on compatible devices. Facebook and LinkedIn users can only view profiles or send messages within their groups. The network effects can also be 'indirect' or 'cross-side,' and sometimes these are equally or even more powerful. Indirect effects occur when, for example, advertisers become attracted to the Google search engine or to Facebook because of the large number of users. Companies can also innovate in business models and find ways of charging different sides of the market to make money from their platform or from complements and different kinds of transactions or advertising (Eisenmann et al., 2006).

In parallel with the strategy literature, some researchers in industrial organization economics have begun using the term platform to denote markets with two or more sides, and potentially with network effects that cross different sides. Such a 'multisided market' provides goods or services to several distinct groups of customers, all of whom need each other in some way and rely on the platform to mediate their transactions (Evans, 2003; Rochet and Tirole, 2003, 2006; Parker and Van Alstyne, 2005). While the concept of a multisided market can sometimes apply to supply chain platforms as well as industry platforms, it does not entirely conform to either category. Nonetheless, there are important similarities between industry platforms and multisided markets. Among the similarities are the existence of indirect network effects that arise between two different sides of a market when customer groups must be affiliated with the platform in order to be able to interact or transact with one another (Armstrong, 2006; Rochet and Tirole, 2003, 2006).

At the same time, not all multisided markets are industry platforms as we describe them in this article. Double-sided markets where the role of the platform is purely to facilitate exchange or trade, without the possibility for other players to innovate on complementary markets, seems to belong to the supply chain category. A multisided market that stimulates external innovation could be regarded as an industry platform. However, while all industry platforms function in this way, not all multisided markets do. For example, dating bars and Web sites, a common example used in the literature, are double-sided markets in that they facilitate transactions between two distinct groups of customers. But there need not be a market for complementary innovations facilitated by the existence of the platform.

The literature on double-sided markets is particularly useful to understand the 'chicken-and-egg problem' of how to encourage access to a platform for distinct groups of buyers or sellers. Nonetheless, the literature has limitations as platform research. For example, it takes for granted the existence of the markets that transact through the platform. With the notable exceptions of Parker and Van Alstyne (2005) and Hagiu (2007), this literature has delivered only limited insight into *why* such platforms come into existence in the first place: the drivers of platform emergence and evolution. Most research focuses on pricing as the key to encouraging access and adoption. However, Evans (2009) focuses on start-up platform strategies, while Hagiu (2007), Eisenmann et al. (2009), and Boudreau and Hagiu (2009) focus on the importance of nonprice mechanisms for the governance of platform ecosystems. They suggest, in accordance with Gawer and Cusumano (2002), that pricing alone cannot be the answer to the inevitable strategic questions of platform dynamics, such as how to share risks among members of an ecosystem. These papers take double-sided (or multisided) research to the next level and bridge the strategy and product design literature as well as the industrial organization economics literature.

Managerial Implications and Further Research

There is considerable research to date on how firms can manage the process of platform leadership. In particular, companies that wish to tap into the innovative capabilities of

an external ecosystem need to develop a vision for their platform and build a sufficiently open and modular platform architecture to facilitate third-party innovation. They need to build a vibrant coalition and carefully manage ecosystem relationships that are mutually beneficial for participants. They need to continue evolving the platform and the ecosystem to remain competitive over time. Overall, managers must make a series of technological decisions on the one hand and strategic business decisions on the other, in a coherent manner.

Yet there are still several aspects of platform strategy and industry dynamics that we do not well understand. For example, in this article, we implied that some industry platforms may emerge from product platforms, but this is only a suggestion from a few cases and needs much more research. Economists tend to assume that platforms already exist and generally do not address this issue, whereas scholars of technological change, competitive dynamics, and organizational processes are more interested in change over time and could help explore more in depth the emergence platforms and associated ecosystems (Adner and Kapoor, 2010). In addition, empirical researchers such as Le Masson et al. (2009), whose focus is on collective design processes for developing industry platforms, might offer other useful theoretical and methodological perspectives.

A related and important area of further study is that business ecosystems. The networks approach from the organizational literature, by bringing its insights on network dynamics and field evolution as well as strategic networks is well positioned to make significant contributions to the question of how business networks appear and evolve. For example, recent work by Nambisan and Sawhney (2011) explicitly examines the link between platform leadership and orchestration processes in network-centric innovation. The new institutional literature rooted in sociology, which offers concepts such as legitimacy, collective identity, and institutional work, can be useful as well to examine whether and how platform leaders can successfully establish themselves as trustworthy brokers (Gawer and Phillips, 2013).

Finally, we need to expand our understanding of the impact of platforms on innovation and competition. In the economics, innovation, operations management, and strategy literature, scholars associate technological platforms with a positive impact on innovation. The positive effect stems from the fact that, by offering unified and easy ways to connect to common components and foundational technologies, platform leaders should help reduce the cost of entry in complementary markets, and provide demand for complements, often fueled by network effects. Platforms therefore offer a setting where it is in the interest of private firms to elicit and encourage innovation by others. However, concern over the dominant positions that platform leaders such as IBM, Microsoft, Google, Apple, Cisco, and Qualcomm can achieve has raised awareness that platforms may have a potentially negative effect on competition and possibly on innovation. Specific platform design approaches might also limit innovators to particular approaches which may be platform-enhancing, rather than platform-destroying. Platform leaders have the incentives and the power to encourage ecosystem innovation that is platform-enhancing, but can also act in ways that discourage or increase the cost of innovation that may be either competing with the platform, or create possible substitutes to the platform. We need further research to evaluate the impact of platform leadership, such as maintained through platform interfaces and architecture control, on different kinds of innovation. This type of research would likely highlight the potential trade-offs between innovation on modules and discrete products versus innovation on systems and platforms at the industry level.

See also: Boundaries and New Organization Forms; Business Models; Dominant Designs; Ecology: Organizations; Hierarchies and Markets; Innovation; Modularity and Organizations; Network Analysis; Network Paradigm: Applications in Organizational Science; Strategic Management; Supply Chain Management.

Bibliography

Abernathy, W.J., Utterback, J.M., 1978. Patterns of innovation in technology. Technology Review 80 (7), 40–47.

Adner, R., Kapoor, R., 2010. Value creation in innovation ecosystems: how the structure of technological interdependence affects firm performance in new technology generation. Strategic Management Journal 31, 306–333.

Armstrong, M., 2006. Competition in two-sided Markets. RAND Journal of Economics 37, 668–691.

Baldwin, C.Y., Wooward, J., 2009. The architecture of platforms: a unified view. In: Gawer, A. (Ed.), Platforms, Markets and Innovation. Edward Elgar, Cheltenham, UK and Northampton, MA, pp. 19–44.

Boudreau, K., Hagiu, A., 2009. Platforms rules: multi-sided platforms as regulators. In: Gawer, A. (Ed.), Platforms, Markets and Innovation. Edward Elgar, Cheltenham, UK and Northampton, MA, pp. 163–191.

Bresnahan, T., Greenstein, S., 1999. Technological competition and the structure of the computer industry. Journal of Industrial Economics 47, 1–40.

Cusumano, M.A., Nobeoka, K., 1998. Thinking beyond Lean. Free Press, New York.

Doran, D., 2004. Rethinking the supply chain: an automotive perspective. Supply Chain Management: An International Journal 9 (1), 102–109.

Eisenmann, T., Parker, G., Van Alstyne, M., 2006. Strategies for two-sided markets. Harvard Business Review 84 (10), 92–101.

Eisenmann, T., Parker, G., Van Alstyne, M., 2009. Opening platforms: how, when, and why? In: Gawer, A. (Ed.), Platforms, Markets and Innovation. Edward Elgar, Cheltenham, UK and Northampton, MA, pp. 131–162.

Evans, D.S., 2003. Some empirical aspects of multi-sided platform industries. Review of Network Economics 2 (3), 1–19.

Evans, D.S., 2009. How catalysts ignite: the economics of platform-based start-ups. In: Platforms, Markets and Innovation. Edward Elgar, Cheltenham, UK and Northampton, MA, pp. 99–128.

Farrell, J., Katz, M.L., 2000. Innovation, rent extraction, and integration in systems markets. Journal of Industrial Economics 97 (4), 413–432.

Gawer, A. (Ed.), 2009a. Platforms, Markets and Innovation. Edward Elgar, Cheltenham, UK and Northampton, MA.

Gawer, A., 2009b. Platforms dynamics and strategies: from products to services. In: Gawer, A. (Ed.), Platforms, Markets and Innovation. Edward Elgar, Cheltenham, UK and Northampton, MA, pp. 45–76.

Gawer, A., Cusumano, M.A., 2002. Platform Leadership: How Intel, Microsoft, and Cisco Drive Industry Innovation. Harvard Business School Press, Boston, MA.

Gawer, A., Cusumano, M.A., 2008. How companies become platform leaders. MIT Sloan Management Review 49 (2), 28–35.

Gawer, A., Cusumano, M.A., 2013. Industry platforms and ecosystem innovation. Journal of Product Innovation Management 31 (3), 417–433.

Gawer, A., Cusumano, M.A., 2014. Platforms and innovation. In: Dodgson, M., Gann, D., Phillips, N. (Eds.), The Oxford Handbook of Innovation Management. Oxford University Press, Oxford.

Gawer, A., Henderson, R., 2007. Platform owner entry and innovation in complementary markets: evidence from Intel. Journal of Economics and Management Strategy 16 (1), 1–34.

Gawer, A., Phillips, N., 2013. Institutional work as logics shift: the case of Intel's transformation to platform leader. Organization Studies 34 (8), 1035–1071.

Greenstein, S., 2009. Open platform development and the commercial Internet. In: Gawer, A. (Ed.), Platforms, Markets and Innovation. Edward Elgar, Cheltenham, UK and Northampton, MA, pp. 219–248.

Hagiu, A., 2007. Merchant or two-sided platform? Review of Network Economics 6 (2), 115–133.

Jacobides, M.G., Knudsen, T., Augier, M., 2006. Benefiting from innovation: value creation, value appropriation and the role of industry architectures. Research Policy 35 (6), 1200–1221.

Le Masson, P., Weil, B., Hatchuel, A., 2009. Platforms for the design of platforms: collaborating in the unknown. In: Gawer, A. (Ed.), Platforms, Markets and Innovation. Edward Elgar, Cheltenham, UK and Northampton, MA, pp. 219–248.

Meyer, M.H., Lehnerd, A.P., 1997. The Power of Product Platforms: Building Value and Cost Leadership. Free Press, New York.

Muffatto, M., Roveda, M., 2002. Product architecture and platforms: a conceptual framework. International Journal of Technology Management 24 (1).

Moore, J.F., 1996. The Death of Competition: Leadership and Strategy in the Age of Business Ecosystems. HarperBusiness, New York.

Nambisan, S., Sawhney, M., 2011. Orchestration processes in network-centric innovation: evidence from the field. Academy of Management Perspectives 25 (3), 40–57.

Parker, G., Van Alstyne, M., 2005. Two-sided network effects: a theory of information product design. Management Science 51, 1494–1504.

Robertson, D., Ulrich, K., 1998. Planning for product platforms. MIT Sloan Management Review 39 (4), 19–31.

Rochet, J.-C., Tirole, J., 2003. Platform competition in two-sided markets. Journal of the European Economic Association 1 (4), 990–1029.

Rochet, J.-C., Tirole, J., 2006. Two-sided markets: a progress report. RAND Journal of Economics 35, 645–667.

Sanderson, S.W., Uzumeri, M., 1997. Managing Product Families. Irwin, Chicago.

Sawhney, M.S., 1998. Leveraged high-variety strategies: from portfolio thinking to platform thinking. Journal of the Academy of Marketing Science 26 (1), 54–61.

Wheelwright, S.C., Clark, K.B., 1992. Creating project plans to focus product development. Harvard Business Review 70 (2), 67–83.

West, J., 2003. How open is open enough? Melding proprietary and open source platform strategies. Research Policy 32, 1259–1285.

Business Process: Outsourcing

Rafiq Dossani, Center for Asia Pacific Policy, RAND Corporation, Santa Monica, CA, USA
Samina Dossani, School of Oriental and African Studies (SOAS), La Jolla, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Business Process Outsourcing (BPO) is a relatively new global industry. In 2011, worldwide spending on BPO services totaled $153 billion (USD) and it is one of the largest and fastest growing outsourcing activities. The business is driven by firms in developed countries looking to save costs by outsourcing work to lower-wage developing countries. A number of developing countries with highly skilled labor force have benefited from the rise of the outsourcing industry. This article focuses on the case of India, the most successful outsourcing destination for BPO services. We look at the development of the BPO market and firm-level outsourcing decisions. Then, we turn our attention to India's BPO industry, considering both the drivers of growth and the organizational features.

Introduction

Business Process Outsourcing (BPO) is the outsourcing of a firm's internal back-office business functions. Outsourcing is the general term used when a firm contracts out specific operational tasks to another firm or person; this comprises of a wide range of activities such as IT software outsourcing, engineering research and development, and outsourced software product development. In the case of BPO, noncore business operations that are necessary for a firm's functioning are outsourced, such as human resource (HR), finance and accounts (F&A), and data entry. Further, when firms (usually in developed countries) decide to relocate processes to another country (usually to developing countries), this is known as offshoring. Traditionally, the primary motivation for outsourcing has been to take advantage of labor cost arbitrage. However, in more recent times, outsourcing is driven by sophisticated service offerings, higher labor productivity, and better service quality.

Over the past few decades, outsourcing and offshoring of services increased dramatically. As of 2011, worldwide spending on software products, IT, and BPO services totaled greater than $1.7 trillion (USD). Global spending on BPO services alone was $153 billion (USD) and is one of the fastest growing segments of the overall offshoring market (NASSCOM, 2012, p. 26). A number of developing countries with highly skilled labor forces have benefitted from the rise of the outsourcing industry, the most well-known case being India. While other countries like China, Philippines, and Czech Republic made headway into this market, India remains the leading destination for global outsourcing, capturing 58% of global market share (in 2011) (NASSCOM, 2012: p. 40). For this reason, we shall mainly focus on India's BPO industry.

We start off by providing a general overview of the BPO sector, looking at its historical development and considering the outsourcing decisions from the client firm's perspective. We then turn our attention specifically to India's BPO sector, looking at both the drivers of growth and understanding the organizational features of this industry.

Development of BPO Sector

The recent phenomenon of services outsourcing has similar features to the manufacturing outsourcing of the 1970s and 1980s. In the case of manufacturing sector, as labor costs increased in the West, firms sought to reduce costs by offshoring low-skilled work to low-wage countries. What emerged was a disaggregated, often modularized, value chain, whereby product development and R&D took place in different locations from where goods were produced.

The outsourcing of the manufacturing sector was successful and accelerated throughout 1970s and 1980s. At the time, services outsourcing was considered to be less feasible as they were localized, nontradable, and nonstorable. Despite offshoring of software services starting as early as the 1970s by US firms to India, Ireland, and Israel (Dossani, 2004), firms still viewed business processes as a fixed cost. These were, therefore, never scrutinized in cost-saving exercises. However, all these changed with technological advances and the arrival of the 'digital age.' Digitization allowed information to be stored and easily shared between different entities (Dossani and Kenney, 2007). Coupled with lower data-transfer costs as a result of improvements in IT and communications, firms turned to reorganizing processes to cut costs. Furthermore, processes that were digitized allowed these to be removed of location constraints. Consequently, as in the case of manufacturing, processes were separated and modularized, allowing firms to find the most cost-effective way of performing individual tasks, even if they were separated spatially.

Originally, Western firms relocated their noncore business processes from higher-cost urban areas to lower-cost small towns and rural locations. Here firms found labor and land cost savings to total 20–30% less than if they set up in urban areas (Dossani and Kenney, 2007). With low costs and higher productivity, firms continued to engage in intranational outsourcing. Although initially the pool of adequately skilled labor was enough to generate these savings, labor shortages in these areas began to emerge over the 1980s and 1990s and firms sought out alternatives sources of low-cost skilled labor.

During this period, a number of low-wage English-speaking countries started undertaking simple processing work. For example, some Caribbean nations started processing airline tickets and credit cards for multinational companies (MNCs), while in the 1980s, airlines began to fly ticket stubs to Ireland and India for processing (Dossani and Kenney, 2009). Although the scale and number of employees in these operations were small, it gave rise to MNCs and firms experimenting and relocating all kinds of business process aboard.

As improvements in technology and further digitization occurred, the pace of outsourcing business services accelerated throughout the 1990s. Firms were no longer constrained by geographical factors, and instead made decisions based on labor skills and cost. Initially, BPO to other countries started with simple, routine work such as call centers, F&A, and other work that could be done remotely. As time went on, firms began to experiment with a range of other back (and front) office services to English-speaking countries characterized by a well-educated but low-cost labor force. Countries such as India, Malaysia, China, Ireland, and Philippines all emerged as contenders for this outsourcing business. As the outsourced country becomes more familiar with the nature of the work, not only did firms outsource more processes but they also began to transfer higher value-added work.

Hence the advancement of technology and digitation of information created a huge potential for a wide range of business process services to be offered remotely. Firms moved from intranational outsourcing to offshoring to lower-waged countries. As firms and outsourced countries become more familiar with the nature of the work, BPO provisions continue to expand and accelerate.

BPO and Firm-Level Decisions

In this section, we look at some of the decisions the outsourcing firm has to consider when thinking about offshoring business processes. As the main motivation for offshoring rests on cost cutting and higher productivity, firms have to make the right choices about the kinds of work to be outsourced; otherwise it could potentially end up being a costly exercise. Factors such as what processes to outsource, where to relocate, whether processes can be separated and modularized are important considerations for successful outsourcing. We now discuss some of these factors in further detail.

The most important decision for firms is to deliberate on what processes to outsource. The simplest tasks get outsourced first as these are the easiest to manage and have lower risk in experimentation and failure. Furthermore, simple tasks allow the outsourced country to catch up and dynamic learning to take place. The rate of learning in the outsourced country depends on its knowledge base, education and skill levels, entrepreneurial spirit, and the local managers (Dossani and Kenney, 2007). The faster the rate of learning, the more potential the outsourced country has for higher value-added work. For example in insurance claims processing, Indian operations initially started off by simply processing claims; as learning took place, these operations took over the role of valuation and investigations of claims, thereby moving up the value chain of work.

Closely related to a firm's decision on what processes to outsource is the knowledge component of the process. The higher the knowledge requirement of the process the more worried the firms, may be about whether the quality of the services will change or even decline as work is transferred to the outsourced country. Any disruption to the business process should be minimized or could prove to be a risky undertaking. This is often the reason why simple tasks are outsourced initially, as both the outsourcing client and vendor firm get comfortable with the overall process.

While lower labor cost is an important factor for relocation of business processes, firms must make decisions on where to locate and the scale of BPO offered. For example in relocating to India, firms benefit from a large pool of college-educated workforce, making recruitment relatively quick. If firms want to expand, Indian operations have the capacity to increase BPO provisions without labor shortages. Thus, the scale of BPO offering in the outsourced country is significant for firms. In addition to this, firms may wish to set up within clusters where work of similar nature is being conducted. The advantage for outsourcing firms is that learning and knowledge spillovers may take place and therefore business process can be offshored faster. Outsourcing firms entering such an environment also benefit from a supportive infrastructure such as good telecommunications facilities or large pools of labor all working on similar operations.

Fourth, firms have to separate and modularize processes such that they are ready for outsourcing. In some cases, it is relatively easy to offshore the process as a whole as activities may be fairly standardized and routinized. For example, in customer care, the process as a whole can be outsourced and offshored as it does not require proximity to the firm. However, not all business processes are straightforward; usually, the ability to outsource work depends on the ability to separate each individual stage of the process. If the business process is not straightforward, firms may need to experiment or reengineer current processes to allow outsourcing to take place. The potential here is great: discarding or reformulating legacy practices and incorporating new methods mean that firms are likely to experience higher productivity brought about by experimentation and new business models.

Finally, there may be some processes that are time sensitive with specific deadlines, for example overnight medical transcriptions that need to be processed for the next day. In this case, firms may choose to offshore to locations that accommodate these time deadlines. For example, as India is 12 h ahead of United States, work that needs to be done overnight in the United States can be completed during the day in India hence shortening the turnaround time.

Drivers of the Indian BPO Sector

So far, we focused on outsourcing decisions and motivation from the individual firm's perspective. We now turn our attention to the BPO sector from the perspective of the outsourced countries, i.e., countries that are receiving outsourcing work. While this industry has taken off in a number of countries, India has emerged as the leading recipient of outsourcing. For this reason, we shall be concentrating on India's BPO

industry for the remainder of this article. In 2011, India captured 58% of market share of global sourcing work, while BPO export earnings totaled over 15 billion (USD) in 2012 (NASSCOM, 2012: p. 102). India has a number of key advantages for firms looking to relocate and outsource business processes. In this section, we shall outline some of the key drivers of the Indian BPO industry.

First and arguably the most important input for BPO industry is a well-educated labor force. Unlike outsourcing of low-end manufacturing jobs, service relocation requires much higher skill levels. In the case of low-end manufacturing outsourcing, labor force can be relatively low skilled since the skills can be embedded in sophisticated machinery. Therefore, a developing country has access to a large portion of its population with the ability to engage in the manufacturing process. However for the production of business services, skill levels must be relatively high and a country must have a ready pool of labor to perform tasks established by the outsourcing firm. In the case of India, the BPO sector has benefited from a large well-educated English-speaking population, allowing a continuous pool of skilled workers for this sector. In 2012, India's BPO industry alone directly employed 880 000 workers and indirectly employed an additional 3.5 million people (NASSCOM, 2012: p. 117). Furthermore, India's BPO operations initially developed in large metropolises like Bangalore, Mumbai, and Delhi, attracting a large number of potential employees and ability to expand operations. Hence, unlike the early relocation of services to rural areas in the West, where skilled labor shortages was the main constraint to expansion, locating in large Indian cities ensured continuous labor flows.

Currently, India is experiencing an interesting shift in the location of its operations. Although service provision continues to grow in traditional locations (like Bangalore, Mumbai, Delhi), we observe the rise of secondary and tertiary cities as alternative service delivery locations. As pressure on labor costs as well as infrastructure cost (real estate, transportation) mount, Indian operations look to these locations as a way of keeping cost low. These secondary and tertiary cities, like Bhubaneswar, Jaipur, and Chandigarh, have access to a large skilled workforce with lower infrastructure costs and state government support, delivering significant cost advantage for firms (NASSCOM, 2011, 2012). Furthermore, the growth of these secondary/tertiary cities indicates that India has significant capacity for expansion, if BPO business continues to develop. For India, the emergence of these alternative service delivery centers mean that firms can keep costs low and fight off competition from other emerging centers in China or Philippines.

In the past 30 years, advances in technology and communications have brought about revolutions in many industries, overhauling processes like never before. It is no doubt that these technological advanced have contributed greatly to the development of the BPO sector. Digitization allowed information to be shared across borders, while the declining costs of data transfer and the 'proximately' to places like India, allows companies to benefit from labor cost arbitrage. Yet, Indian BPO expansion has not relied solely on technological change. As in the case of software outsourcing, it was important to demonstrate that BPO was a legitimate and normal business decisions for firms. In Indian software outsourcing, firms obtained international certification (such as International Standards Organization (ISO) certification; Carnegie Mellon software certification) (Dossani and Kenney, 2007). However in the BPO industry, legitimacy was provided by the huge success of early adopters of service outsourcing. Entry by large MNCs in India started in 1993 by American Express, followed by British Airways in 1996, and General Electric in 1998. It was not until the early 2000s that a whole host of fortune 500 firms, such as HSBC, Hewlett–Packard, Citicorp, all relocated their business processes to India (Dossani and Kenney, 2007). By this time, firms looking to cut costs saw this as a part of a viable business decision.

Traditionally BPO work in India started off by firms undertaking voice-based services such as call centers or data entry work. This was followed by more generic back-office functions such as HR and F&A. As Indian service providers established a strong offshore business, they gradually moved up the value chain. More recently, as these traditional markets matured and cost savings on the margin start to diminish, Indian firms began to reengineer processes to move into higher knowledge-based work. This process of moving up the value chain has been vital for the survival of the BPO business in India, generating a huge growth potential in this sector. Indian firms have shifted their focus from simply providing generic support services to providing higher value-added activity in specific industries.

This shift toward knowledge process outsourcing is seen as a natural progression for BPO firms. Driven by better analytical tools, huge volumes of data, and better understanding of specific industry processes, Indian operations are now looking to provide high-end services such as legal process outsourcing, knowledge management services, or research and analytics to both existing and new clients. This is evident from the fact that data analytics is the fastest growing sector in India after financial services (NASSCOM, 2012: p. 117). As BPO firms move from generic to more specific outsourcing work, firms need to reengineer processes toward industry specific operations. As Indian BPOs offer higher-end processing services, their success will depend upon whether they are able to develop capability across the entire value chain. One strategy for BPOs to quickly obtain industry specific expertise is to acquire smaller niche, specialist companies to fill in the knowledge gap. By engaging in mergers and acquisitions (M&A) activities, BPO can offer far more varied service offerings and extend market reach. This new development in the BPO market suggests Indian firms are far more likely driven by value creation services than simply relying on cost arbitrage alone.

Furthermore, India's goal to be the leading provider of services is not restricted to its own borders. As India moves up the value chain, some lower end work is being outsourced to cheaper countries, prompting the rise of Indian MNCs. The emphasis on a 'Global Delivery Model' has become an important feature for the larger service providers. The idea is to provide services using a team that is globally distributed. In doing so, clients are offered services that are seamlessly delivered from global locations to the local markets, providing continuous 'round the clock' services. Furthermore, clients can enter local markets easily and take advantage of local knowledge, language, and culture in a particular country (PWC,

2011). The major advantage for the Indian BPO firms is the ability to access resources of varying costs across different operations. Firms can 'pick and choose' where individual process should be done, mixing costlier 'on-site' resources with cheaper 'offshore' components. This allows firms to deliver services to its customers at optimal costs and the ability to remain competitive. Finally, by going global, Indian operations are 'risk-proofing' from major disasters, either natural (flooding, earthquakes) or man-made (political instability). In the event of a disaster, service delivery for clients can be easily shifted to another location hence not disrupting business. The 'Global Delivery Model' has been successfully implemented by some large software and service providing firms such as TCS and Infosys. In 2008, the delivery footprint of the Indian BPO industry extended to over 25 countries outside of India, increasing their global presence (NASSCOM-Everest, 2008).

While India's labor pool, higher value-added work and its eyes on global markets are the more contemporary drivers of the BPO industry, there are of course a whole host of other contributing reasons for its success. Long-term factors such as low geographical, political, and data security risk; good government support; and excellent infrastructure provisions all contribute to the Indian BPO success. The reforms of 1991 and subsequent liberalization of India brought about two important changes to business in India. Firstly, the state-run telecommunication sector was reformed by allowing competition from large numbers of private providers into the market. These telecommunication firms could bid for a wide range of contracts from niche specialization to fully integrated services. As a result, India ended up with telecommunication networks with quality and cost levels on par with developed countries. Secondly, as foreign venture capital rules were changed, this led to the availability of financial support for Indian start-ups including those emerging in the provision of business services (Dossani and Kenney, 2007).

India currently remains the world's leading outsourcing base (in both IT and BPO); however, other countries such as China, Philippines, and Czech Republic are also emerging as major outsourcing destinations. The survival of Indian BPO firms depends on the ability to expand offerings and capture more of the value chain. This will require further experimentations, new business models, and focusing on reengineering processes. Indian firms are also looking to expand into new markets. Traditionally, with India's strength in English, English-speaking regions form the bulk of export, the UK and US market accounting for over 80% of BPO exports. However, Indian firms are venturing into newer markets through acquisitions, enhancing language capability and setting up global delivery centers in Eastern European, Asia Pacific, and BRIC (Brazil, Russia, India, and China) regions. As Western markets are currently economically depressed, it is no surprise that Asia Pacific and BRICs (including India itself – see below) have seen the fastest growth rate for Indian BPO organizations in recent years (NASSCOM, 2012).

Finally, as India is one of the largest economies in the world and experiencing high economic growth despite global uncertainties, Indian BPO firms can look to its own domestic market to provide services. As BPO operations move into niche, specific markets and the types of services expand, BPO firms can target domestic firms to provide servicing solutions. The domestic BPO segment is expected to reach 149 billion (rupees) driven mainly by voice-based services (particularly local language provisions) and other niche services. Hence domestic demand may be one future driver of the Indian BPO business.

Organizational Structure of the BPO Industry in India

Although the growth of the BPO industry in India was initially triggered by MNC subsidiaries, over time the market has evolved and with it the organizational structure has changed. In India, there are over 500 BPO organizations across various industries, all differing in size and diversity in service offering (NASSCOM, 2012). In addition to this, a wide range of ownership and organizational structures exist. Dossani and Kenney (2007) identify seven organizational structures featured in the Indian BPO market: MNC subsidiaries, MNC outsourcing service providers, MNC specialist, US high-technology start-up, Indian specialist, Indian independents, and Indian IT industry subsidiaries, which we will discuss next.

MNC Subsidiaries

As mentioned earlier, MNC subsidiaries were the first movers in the Indian BPO market. American Express established the first back-office BPO operation in 1993, followed by British Airways in 1996, and General Electric in 1998. Since 2000, a whole host of global firms entered into this market, across different industries. Even MNCs such as Dell computers, AOL, and SAP with no previous Indian operations were successful in establishing BPO operations. Furthermore, BPO provision was a natural progression for MNC subsidiaries that already had software development facilities in India.

By setting up an internal operation, MNCs benefit from greater controls over key business decisions such as labor processes, business organization, and ability to keep sensitive information in-house. As the MNC subsidiary gains substantial experience with BPO work and organizational process become easier with time, MNCs can outsource higher value-added work with lower risk, allowing the subsidiary to handle increasingly sophisticated work on par with their Western counterpart. Today there are more than 100–120 MNCs in India and about 30% of these businesses are focused on delivering high-end services such as research, analytics, consulting, and legal services (NASSCOM, 2012).

Higher value-added work is not the only indication of maturity of the Indian subsidiaries. MNCs handing over global managerial responsibilities to their Indian subsidiaries show confidence in the quality of Indian operations, as they are trusted with enforcing standard operating practices in other overseas facilities. Finally, in some cases, the Indian subsidiary is successful enough that the MNC feel its subsidiary can become an independent service provider to other third parties. For example, Genpact, which initially started off as a business process subsidiary of General Electrics is now an independent company offering BPO solutions (Dossani and Kenney, 2007).

MNCs benefit from operating a subsidiary as it is able to control the business operations internally, reduce risk

associated with knowledge transfer, and expand service offerings relatively easily. As more MNC subsidiaries have entered the Indian market, the BPO sector continues to evolve. Indian subsidiaries are carrying out more sophisticated work, enjoy greater management responsibility, and in some cases the Indian operations can spin off as an independent company. It is expected that MNC subsidiaries will continue to set up and expand in India.

Multinational Outsourcing Service Providers

MNC outsourcing service providers are no strangers to outsourcing globally. Many companies like IBM, HP, and Keane (data systems firms) or ADP, EDS (payroll and accounting) offered service solutions internationally, prior to the rush of outsourcing to India in 2000s. After 2000, many of these firms entered the Indian market due to competition from the Indian independents and the demands for lower cost from customers. By locating in India, MNC outsourcing service providers were able to deliver nearshore and offshore services, at a lower cost base. These companies were able to enter the Indian market quickly and scale up operations rapidly. While MNC outsourcing providers had long established customer relationship and experience of offshoring, entering India was a response to cost and delivery pressures.

MNC Specialist

It is not only the large sized MNC companies that can benefit from outsourcing to India. Smaller MNC companies that offer niche and varied specialization may also move into the Indian market to access high skilled, cheap labor. MNC specialists undertake a wide range of processes from medical transcription, tax preparation, map digitization, engineering drawing, and so on. Often these MNC specialists may have little experience in operating abroad, yet it makes sense to outsource labor intensive, highly specialized work as this is often more expensive in a developed country.

US High-Technology Start-Up

The traditional model of start-up companies is one where a firm begins in a single location and only offshores operations as it matures. However, there are advantages for start-up companies to consider offshoring in the early stages of its development. Firstly, start-ups are often constrained by limited resources; therefore, offshoring can supplement onshore operations, at a much lower cost. For example, after the dot.com bust in the early 2000s, high-tech companies were forced to operate at lower cost, as financing was a major issue (Dossani and Kenney, 2007). Secondly, by offshoring at an earlier stage in the business model, firms can experiment at a much lower cost. Experimentation is important for businesses, particularly at early stages of operations. Hence by offshoring early on firms can conduct more experiments then they would otherwise. Alternatively, start-up companies can locate their headquarters, marketing, and sales in the United States and all other operations in India (including R&D). For Indian firms, not only does this create further employment opportunities, but they also benefit from knowledge spillovers and experience in high-technology entrepreneurship.

Indian Specialist

Unlike the software industry in India, the impact of the Indian specialist firms in nonsoftware service sector was less significant. Initially, Indian firms suffered from a lack of legitimacy, customer contact, and domain knowledge. However, as Indian firms gain experience with service outsourcing, the number of Indian specialist firms grew. Labor cost arbitrage is not the only reason companies' relocated work; Indian specialist quickly gained recognition for high skilled, high value-added work. As a result, the Indian BPO sector consists of large numbers of niche firms engaged in a wide range of service provision, across a variety of industries.

Indian Independents

Indian independents are like the Indian specialist but focus on providing general services like call-center operations or claims processing. These firms tend to be much larger than the Indian specialist, some employing more then 6000 employees and may have large MNCs as customers. The growth of these large Indian independent firms accelerated after the dot.com bust as US firms looked to reduce costs. For example, Tracmail was established to provide back-office services to US firms such as Amazon and Yahoo! (Dossani and Kenney, 2007). From 1999 to 2004, the rapid expansion of low-end service work, such as call centers, brought about many benefits to the Indian market; however there were a number of problems with this kind of work. High attrition rates, lack of career path for employees, cut-throat pricing, training expense, and problems of working through the night (Ramesh, 2004) encouraged Indian Independent to undertake higher value-added BPO work. As the nature of the BPO work means that business activities are part of a larger process, firms have the opportunity to move up the value chain as they become familiar with process. As a result by 2004, Indian independents were diversifying away from call centers into other business areas such as F&A.

As the market in India consolidates, M&A activity in recent years has increased for both the Indian independents and Indian specialist (NASSCOM, 2012). Indian independents are often brought over by MNCs looking to set up in Indian or large capture-all firms looking to provide 'bundle offerings.' The BPO market is likely to be dominated by these larger, diversified firms with multiple service offerings, as smaller firms are often overlooked for large contracts both domestically and internationally.

Indian IT Industry Subsidiaries

As early as 1970s, Indian software service subsidiaries were the initial trigger for the relocation of services to India, even before BPO took off. These companies initially grew by using low-cost Indian labor to provide high-quality software provisions globally. Over the last few decades, Indian software service vendors developed into large globally competitive firms, with significant global market share; companies such as TCS,

Infosys, Wipro, Satyam Computers, and HCL are leaders in software outsourcing. These companies are all experienced with operating on the global level, managing foreign customers, with excellent executive and managerial capabilities, and the ability to get overseas contracts.

The move to provide BPO was therefore a strategic opportunity for Indian Software firms as they acquired significant advantage over new entrants in the BPO market. Once they decided to provide business-processing services, these software firms took different routes to enter the market. Some grew organically; others established in-house subsidiaries or through M&A activities. While Indian software firms were leaders in the provision of software services, many found the nature of BPO provision somewhat different from software provision. Particularly, the skills and requirements of the workforce in BPO differed from software workers who were mainly from science and engineering backgrounds. BPO workers often possess qualities such as good interpersonal skills, the ability to directly deal with customer interactions, and were often with a social science or commerce background. Hence, while it is natural to assume Indian software subsidiaries would have been the most successful in providing BPO service, in reality these firms had to act strategically and reorganize processes to enter the BPO industry. For example, rather then moving into simple call-center service provisions (and recruiting a large number of call-center staff), Infosys established an in-house subsidiary, Progeon (now known as Infosys BPO Limited). Progeon focuses on only providing high-end back-office work, utilizing the already existing software skills of Infosys. Likewise, HCL established its in-house subsidiary, HCL BPO. The subsidiary provided call-center work by acquiring UK call centers rather then expanding in India. By locating in UK, HCL solved the language problems of Indian call-center operations and left Indian operations to concentrate on back-office work.

To summarize, we see that the BPO sector in India has a number of different organizational and ownership structures. As the economic model behind India's BPO industry is constantly changing, so too will the organizational structures of the industry. While the MNC is a large part of the BPO market and will likely to dominate this sector, Indian firms are also growing dynamically. Looking ahead, the Indian BPO market is likely to remain a high-growth industry as firms constantly evolve its service offerings and diversified into many different activities. Although India is currently still the number-one destination for sourcing, its dominance is threatened by other emerging centers like China and Philippines. Survival of Indian firms' depends on their ability to capture more of the value chain and expansion into new markets.

See also: Boundaries and New Organization Forms; Multinational Corporations; Supply Chain Management; Technology and Organization.

Bibliography

Dossani, Rafiq, 2004. Origins and Growth of the IT Industry in India. Working paper. Asia-Pacific Research Center, Stanford University.

Dossani, Rafiq, Kenney, Martin, 2007. The next wave of globalization: relocating service provision to Indian. World Development 35 (5), 772–791.

Dossani, Rafiq, Kenney, Martin, 2009. Service provision for the global economy: the evolving Indian experience. Review of Policy Research 26 (1–2), 77–104.

NASSCOM, 2011. The IT-BPO Sector in India. NASSCOM, New Delhi.

NASSCOM, 2012. The IT-BPO Sector in India. NASSCOM, New Delhi.

NASSCOM-Everest, 2008. India BPO Study: Roadmap 2012-Capitalizing on the Expanding BPO Landscape. NASSCOM, New Delhi.

PWC-ORN, 2011. The Ever – Changing Global Service Provider Industry: Key Findings for 2012. The Duke Offshoring Research Network (ORN), Fuqua School of Business, Duke University, Durham: North Carolina.

Ramesh, P. Babu, 2004. 'Cyber coolies' in BPO: insecurities and vulnerabilities of non-standard work. Economic and Political Weekly 39 (5), 492–497.

Busyness

Charles N Darrah, San Jose State University, San Jose, CA, USA
Nicole M Conand and Alicia Dornadic, Design Practices Collaborative, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Busyness (or being busy) is a phenomenon experienced by many Americans and it is a folk concept used to describe everyday life; the concept is also used by researchers describing and analyzing the pace of those lives. Busyness includes assumptions about time and its use, everyday activities, individual and family choices, and constraints on decision making. It assumes values such as performing activities that are consistent with commitments to other people, the necessity to create individual identities by adopting lifestyles, and demonstrating responsibility by being productive and efficient. Busyness was initially associated with increases or decreases in paid work, but it has increased with increases in leisure time, wages, consumption of goods and services, and unpaid work. It is a complex phenomenon that is at once individual and social, and behavioral and experiential. Facets of the concept are addressed by different disciplines and fields, and it is often associated with individual or social costs, such as stress. Information technology has been used by individuals and organizations to control and coordinate activities by providing connectivity to other people and information, although doing so may simultaneously intensify the experience of busyness.

Defining Busyness

Being busy, or busyness, can be broadly defined as having too much to do in the time available. Time is experienced as being a scarce resource and busyness may be expressed as a struggle to balance or juggle commitments, of time being 'squeezed' from the day, or of a speeding up of daily life. The concept is used both in scholarly discourse and as a folk category in everyday conversations. As a scholarly concept it includes considerations of time, activities or tasks, individuals, and context or constraint. Despite its apparent simplicity and familiarity, busyness is a complex construction that is embedded in characteristics of modernity.

Being busy is closely related to time. First, busyness is situated in historical time and it is different than working hard. Busyness seems to reflect a distinctly contemporary worldview in which activity is both ubiquitous and urgent, but also where the stakes might not be as high as sheer survival. Second, busyness is associated with clock time in which days, minutes, and seconds provide a metric for assessing and comparing use or consumption of time (Snyder, 2013). Third, claims of busyness go beyond describing how discrete activities fit into periods of time, but they also describe a state of being that defines a way of life, either that of an individual or a society. Busyness implies a time horizon or duration of time in which one is busy and, perhaps, never not busy.

Being busy also differentiates time by the activities it contains. Activities can be observed and recorded and doing so simultaneously acknowledges that the time could be used to do something else. Yet busyness goes beyond the description of activities. It has a subjective and non-observable quality; it is as private and internal as it is a matter of display. This is partially because people learn to produce and recognize distinct activities and the scripts that define how they are appropriately performed. These distinctions increase the number of activities as well as their heterogeneity. There is an important granularity to busyness, for the more distinctions made among activities and their components, the more there is to be managed (tracked, categorized, reassembled, etc.). Likewise, there is a private work of anticipating, planning, and prioritizing that is necessitated by being busy but that cannot be observed by others. Life in general or specific days do not just accelerate familiar activities, but when activities proliferate, new coordination or management tasks are created, ironically adding to busyness. This is tacit work that is both individual (individuals may define and organize tasks differently) and social, since scripts, routines, and values are learned in society.

Third, busyness implies a context or constraints in which people are acting. These constraints permit comparisons that allow people to identify themselves as busy. Constraint is important because it shifts the experiential locus of busyness from the individual to the environment in which they act. Constraints allow people to experience busyness as something that happens to them apart from the actions they take. There is a sense of fullness or of pressure to initiate and complete activities against a backdrop of fear that everything might not be accomplished. This defines the day as a set of accomplishments that should or must be completed, although what will happen if they are not is often unclear. In fact, being busy means experiencing life as both a set of accomplishments and of activities that once seemed essential, but which were not completed.

The ultimate time constraint is death. Being busy can push back thoughts of mortality by filling days with activities and it may seemingly extend life by allowing people to do and experience more in the same amount of time. Despite the seeming simplicity of being busy it is richly complex. Individuals may distinguish activities performed from the private, internal work of managing time. Evidence that they are busy may be provided by remembering tasks they desired or expected to perform but did not.

Busyness affects individual capabilities and their interactions with others. Activities and tracking them are related but distinct, and it is an individual responsibility to do both.

There is a solitary aspect of being busy, for the focus is with how efficiently individuals allocate time. The contemporary discourse of being connected identifies an emergent category of activities that requires effort and cannot be taken for granted. The busyness of others affects lives, even those of people who eschew a life packed with activities and commitments. The busyness of others is salient when daily schedules are laid and plans to undertake activities are contingent on the availability of other people.

Being busy is therefore not a mere descriptor of individual lives but also of the society in which they live. Busyness is social because what drives so many of the activities is situated beyond individual agency – government regulation, deregulation, consumerism, knowledge work, technology, etc. The sociality of busyness is deeper, however, for a modern person does not simply fit into his or her society, but from it they choose how to live: they construct a lifestyle in consort with others who are doing the same. In modern societies, there is a sense that things could be different than they are and that it is the individual's responsibility to be the author of their own life. Under such conditions, busyness assumes a moral weight (Greenfeld, 2005).

To be busy implies interpretation and judgment, as well as enumerating activities. Interpretation enters because we cannot simply watch people in order to exhaust the possibilities of their busyness, for everyone is always performing activities, even if it is sleeping or 'doing nothing.' Researchers can watch people rapidly performing diverse activities, but they might not judge themselves to be busy. Either an analyst or a cultural native proclaims that what is occurring is evidence of being busy, and there is no reason to assume that insider and outsider assessments will coincide.

Moreover, busyness itself has changed. First, busyness has historically been addressed as an issue of the relationship between work and leisure, so that it is difficult to reconcile survey results in which work paid hours are generally not increasing but individuals claim to be busier. As familiar work-life boundaries have dissolved, it is difficult to locate the causes of busyness in work hours. Linder (1970) spoke of a harried leisure class to identify precisely this disconnection of work and leisure. Accordingly, it is difficult to differentiate real and perceived busyness, just as activities and experience can be difficult to categorize as that of work and family.

Second, claims that busyness is exclusively a middle or upper class affliction are challenged when lower-income, under-employed, and even unemployed workers experience busyness. Cutbacks to social services and increasingly flexible, customizable work hours shift responsibilities to individuals and families, and results in fragmented daily schedules found across income levels.

Third, there is increased sensitivity that busyness is not a uniform, homogeneous transformation of society to one that occurs unevenly and is structured by gender. Wajcman (2008) observed that men and women enjoy different kinds of leisure time. Men's occurs in longer blocks of time and is relatively uninterrupted, while women's leisure is tied to the nature of household work and the activities of children, and it is more likely than men's leisure to be interrupted. Women's leisure may be less restorative and, in fact, simply another form of work. Although Wajcman was describing the United Kingdom, her analysis suggests that fine-grained differences in activities patterned by the domestic division of labor can produce qualitatively different experiences of time and busyness. Differential participation in labor markets by men and women, and the characteristics of particular jobs, should have these fine-grained consequences in households and in the experience of busyness by their members.

Fourth, the experiential component of busyness is essential and it cannot be inferred from quantified measures of time allocated to activities. No amount of observation can fully capture busyness since it is partially a function of subjective experience that comes from comparison to normative standards that vary culturally. Thus, busyness is descriptive and prescriptive: it is about what should or must occur during a day and tracking progress in fulfilling plans is just an additional activity.

Fifth, members of modern societies assemble their own lifestyles from resources provided by society. This work is largely individualized and further enabled by information technologies (ITs) that provide seemingly unlimited information and capacities to build networks and to 'be connected'. While this lifestyle work is liberating, it simultaneously increases individual responsibilities to forge identities, typically through activities of production and consumption. For example, advertisements that proclaim opportunities to 'have it your way' place pressure on people to think about their lives and to participate in activities that will enact them. Collectively, this lifestyle-building both adds to the public and private activities of daily life, and its individualized nature intensifies the fragmentation of daily schedules that then must be coordinated among family members, friends, coworkers, and acquaintances.

Busyness in the United States, as in any country, must be understood as one manifestation of an international phenomenon. It is grounded in modernity and explanations draw upon general theories of production and consumption (Vercruyssen et al., 2011). Accordingly, both busyness and its scholarship transcend national boundaries; it is not situated exclusively in the United States, despite the influence of pioneering contributions by Schor (1993) and Hochschild (1997).

Intellectual Backgrounds, Schools of Thought, and Advocates

Conditions that affect time use are ubiquitous and are addressed in many different fields and disciplines. Scholarship about time allocation is well developed and fundamental to comparative studies of busyness, but it does not capture all that is salient to the latter since being busy is about more than time and activities. A central challenge to the study of being busy is that much is relevant to the topic and little can be excluded, yet a core of 'busyness studies' is absent. The study of busyness is therefore more an intersection of disciplines and fields than a discrete field.

Academic Approaches

Sociology, economics, psychology, and anthropology are especially germane to the scholarship of busyness. Sociologists

contribute time allocation studies that permit cross-national comparisons and that capture trends in how Americans allocate time. Several approaches are used. In 24-h recall time diaries, individuals complete a diary about the previous day's activities. Increases in one activity must be offset by reductions in others. Analysis of activities may reflect categories (e.g., work, recreation) defined by researchers that facilitate comparison or by diarists that correspond to the distinctions they use to organize their everyday lives. Broader categories of activities may facilitate comparative research, but may not capture subtle shifts within them. The approach relies on recall of time use (not observation of behavior) and it may miss infrequent or episodic activities. The focus is on behavior rather than experience. The American Heritage Time Use Study database, for example, combines four surveys conducted by the University of Michigan, University of Maryland, and the American Time Use Survey (ATUS) conducted since 2003 by the US Bureau of Labor Statistics (http://www.timeuse.org/ahtus/).

A second approach is that of estimated activity data sets in which individuals are asked to estimate the time they spend on a sample of total activities. Advantages include a typically high response rate, precisely replicated questions that allow comparison, potentially large sample sizes, and the ability to capture infrequent activities. Limitations include informant recall bias, survey questions and categorizations of data that are not salient to local activities or discourse, attention to behavior and not experience, and data that does not capture subtle shifts in usage of time. Examples include the General Social Survey from the University of Chicago which has covered the broadest range of activities since 1972; the Survey of Public Participation in the Arts conducted by the US Census Bureau for the National Endowment for the Arts since 1982; and the Fitness Surveys by the National Center for Health Statistics since 1985.

Despite limitations for understanding busyness, time allocation studies provide the best tracking in usage of time by Americans, and the findings provide much of the data that is salient to public policy and civic discourse about the nature of contemporary society.

Economists look at time as a scarce resource to be rationally allocated by individuals and households. Time allocation data is relevant to how people choose to participate in domestic production versus working additional hours and purchasing goods and services, with time as the ultimate limited resource. The decline in hours spent cooking, for example, is correlated with women increasingly entering the labor market, thereby increasing income but not time available to prepare meals. Time allocation at least partially reflects trends in wages, and the costs of goods and services. Hamermesh (2009) suggests that economists should use time allocation data collection to explore topics such as household production and bargaining, joint time use by spouses, peer effects and social norms (including timing activities), and the relationships of macroeconomics to individual and household activities.

Psychologists view time use as a behavioral manifestation of preferences and values, specifically as it correlates with mental processes in daily life (Csikszentmihalyi and Larson, 1987). The authors' Experience Sampling Method (ESM) takes various forms, but it attempts to describe self-reported mental processes as they occur in context. Protocols may require individuals to respond at regular intervals, to random signals, or after specific events. ESM can obtain empirical data on daily activities and psychological states and patterning of thoughts, thereby providing data relevant to the experiential component of busyness. Multitasking, for example, has become a basic strategy used by many couples to perform more activities in the same unit of time. Offer and Schneider (2011) use ESM to demonstrate that men and women differ in kinds of multitasking they perform, the contexts in which they perform it, and how they assess their experiences. Krueger et al. (2008) propose a National Time Accounting that links time diaries with respondents' recalled emotional experiences of a day's activities. They approach time accounting with a focus on psychological well-being as measured multidimensionally.

Finally, anthropologists conduct cross-cultural comparisons of time use, and they situate discrete activities as components of social practices that occur in time and space. They focus on meaning and the material accoutrements of social interaction (Graesch, 2009), as well as emic (insider) perspectives on activities and time use (Darrah et al., 2007). Ethnographic fieldwork attempts to link household-level data with macro-level societal conditions. The methodology gets at fine-grained discriminations among activities as they are categorized by participants. This approach allows observed activities to be connected to how they are discussed by actors, thereby linking agency with structure, experience, and meaning. The limitations of in-depth fieldwork are the small sample sizes and lack of random sampling, as well as analytical categories that may limit comparisons.

Interest in busyness is grounded in scholarly interest, but because it addresses the fundamental question of how people spend time, it is also associated with broader societal issues and civic discourse about them. One interpretation of busyness is that of a time crunch or time famine. Based on national survey data collected 1948–87, Schor (1993) argued that Americans were overworked because, despite productivity gains, hours in paid work had not declined. The picture painted was one of trends external to individuals that forced them to undertake new tasks while simultaneously spending more time in paid work. A counter argument is that hours of paid work have not increased and any time famine has been created by individuals taking on additional work hours, in effect trading time for income that is spent on goods and services, yet another activity. Hochschild (1997) examined the time pressures felt by families as more women entered labor markets. From research she conducted in a small midwest town during summers between 1990 and 1993 on a company she calls 'Amerco', Hochschild found that while people articulated values in which family commitments were primary, they were often drawn into workplaces where they often felt more rewarded. A result was that previously distinct cultural domains of work and family became blurred as family became work and work became family.

Organizational Implications

Increased levels of worker productivity and work hours, as well as employer requirements for mandatory overtime, flexible hours, or availability to fit into corporate optimal scheduling practices can make it difficult for workers to plan other

activities. Perceived stress, increased absenteeism, and attrition may result, thereby threatening productivity (Perlow, 1997). Computer-mediated communication, such as e-mail, originally promised to enable employees to communicate quickly across time and distance. Mobile technologies allow many workers to perform at least some job-related tasks virtually anywhere, while simultaneously being constantly engaged in work and accessible. This can amplify the blurring of domains of work, leisure, and family, and it presents new responsibilities for managing time use. For example, because e-mail is asynchronous, when coupled with norms of being accessible and responsive, its stored messages can intensify feelings of work overload (Barley et al., 2011).

Current Knowledge

There are no comprehensive measures of busyness since the concept includes both behavior and experience, and because the concept is primarily of interest to separate fields and disciplines with diverse interests. Current knowledge is accordingly about the conceptualization and utility of the concept as it is about empirical findings about its extent.

Acceleration of Social Change

Busyness as a social phenomenon is distinct from hard work, and it is grounded in specific cultural milieus and historical moments (Greenfeld, 2005). Certain cultural milieus value busyness and make it a component of meaningful individual lives, as when individuals prize activity and efficiency. Historically, busyness is associated with three forms of social acceleration (Rosa, 2003). Technological acceleration results from speeding up processes of transport, communication, and production through technological innovation. IT, in particular, affects acceleration by allowing more activities to be allocated to the same period of time. Such efficiency is in turn linked to increased leisure. IT also underlies activities that are valued in themselves; gaming and many uses of social media are exemplary. These new activities in principle require time that must be taken from other activities.

The acceleration of social change is manifested in rapid changes in social organization and structure, patterns of action, and scripts for action. A result is that past experience may be less reliable as a source of lessons for the present and future. Work and family exemplify the effects of this acceleration, where families are less often multigenerational enterprises and work is no longer stable throughout an individual's life, much less across generations. Third, the pace of life itself accelerates, which is paradoxical since technology may have been adopted to provide more leisure. This pace is reflected in personal assessments of scarce time and feelings of being harried (Darrah, 2007). The acceleration is reflected objectively in contraction of time spent in activities, such as eating or walking faster, or the compression of activities and experiences within a given time period by simultaneously performing activities, and reducing pauses or intervals.

Social acceleration is driven by capitalism in which time equates to money, innovators need to be first to market, and material consumption is intense. It reflects competitive pressures on individuals to work longer hours in order to make themselves more attractive to employers and on employers to intensify work processes in order to maintain smaller headcounts in order to reduce training costs, wages, and benefits. Flexible work hours and locations, too, lead to scattered working hours and more consumption. Flexibility and a reduction in planning and formal patterns of interaction weaken sociotechnical structures and create additional tasks of coordination (Southerton and Tomlinson, 2005).

The cultural driver is created by an image of the good life as one fulfilled during the human lifespan. Social acceleration promises to produce more and diverse experiences, hopefully resulting in a fulfilling life. Such promises are especially important during historical conditions when paths to meaningful lives are uncertain.

Time Allocation

Time allocation studies sensitize us to the fact that people are always doing something; in this sense, everyone is always busy. These studies divide time into periods that contain distinguishable activities. Individuals and households then make decisions about allocating time to perform the activities. The power of the approach lies in the zero-sum nature of time: by tracking activities over the 24 h of the day, increases in any activities must be offset by reductions in others and vice versa (Gershuny, 2000). Changes in allocations reveal how individuals use time that is categorized in socially meaningful ways. The categories themselves are not immutable and their definitions tell us about how time allocation is meaningful within specific cultures. Although time allocation studies permit cross-national comparisons, the categories must be created either by observers or the people whose activities are being described. The former distinctions facilitate comparison, but at the risk of distorting how people experience and make decisions about their daily lives. The latter perspective is sensitive to how people experience and organize their daily lives, but it inhibits comparative studies.

Robinson and Martin (2009) use ATUS to describe trends in paid and unpaid work in the United States, providing the broad-brush portrait of national time allocation that underlies most discussions of busyness. First, men's time devoted to paid work (including education, training, and searching for jobs) declined from 1965 to 1985 and then remained flat through 2003. In contrast, women's hours in paid work remained flat from 1965 to 1985, increased in 1993, and then remained flat through 2003. Aguiar and Hurst (2009), too, find that the average man in 2003–05 worked 40 h per week (including commuting time), compared to 43.5 h in 1985 and 51 h in 1965 (in demographically adjusted hours). Women's participation in the labor force between 1985 and 1995 contributed to the increase in their work time. In summary, overall time devoted by men to paid work has declined; for women, it has increased.

Second, men's unpaid work increased between 1965 and 1985, and then did not change through 2003. Men performed on average two episodes of unpaid work per day, with 85–90% of men performing unpaid work on any day. Nearly 100% of women performed unpaid work each day although their total

time in it declined between 1965 and 2003, offsetting their increase in paid work. Most of the decline is attributed to women's reduced time in food preparation and the average number of episodes of unpaid work for women dropped from five in 1965 to three by 1990, where it has remained. Strikingly, total work time declined by 15% between 1965 and 1985, while leisure time increased.

Aguiar and Hurst (2009) note that leisure has increased for men and women since 1965 by 5 and 3.5 h per week, respectively, with most of the gains coming before 1985. In 1965, the average American spent 31 h in leisure activities compared to 35 h in both 1985 and 2003–05. They define leisure as including time spent watching television, socializing with friends and family members, sports, reading, hobbies, entertainment, and all similar activities. For both men and women, leisure inequality has increased since 1985. Less educated men and women have experienced substantial increases in leisure, while more educated individuals have experienced few leisure gains. Men with at least 16 h schooling spend less than 100 h per week in leisure activities, while men with 12 years and those with less than that spend, respectively, 108 and 113 h. The increase in leisure inequality corresponds to documented increases in wage and consumption inequality (Aguiar and Hurst, 2009: pp. 57, 61).

A contribution of time allocation studies then is that being busy is not simply a function of time at paid work. Americans perceive that they rush more and are more stressed, despite the decline in hours at work, although it is also true that the mean minutes on break at paid work declined significantly for both men and women between 1965 and 2003 (Robinson and Martin, 2009). While differences in reporting activities may be a factor, the results suggest that time pressure during paid work may have intensified.

Hamermesh and Lee (2007) compare data from Australia, Germany, the United States, and South Korea. Just over 42% of US men (42.6%) and 54.5% of women report they are 'almost always' or 'often' under time pressure. Their analysis suggests that the celebrated time crunch better characterizes higher-income families both because their members choose to work more hours and because they have higher incomes to spend during the same amount of nonwork time. This result is consistent with Linder's argument (1970) that increasingly specialized work leads to productivity gains that result in products and services that have to be consumed, thereby intensifying the pace of consumption, the demand for higher quality goods and services, and their simultaneous consumption.

The time crunch debate frequently assumes that being busy is harmful by equating it with being harried or stressed. Yet busyness may simply provide a neutral descriptor of contemporary life. In fact, being busy or experiencing time pressure may be partially explained as a culturally acceptable way to account for everyday experiences in societies dominated by modern industrial values. Even when complemented by assessments of quality of life, it is difficult for time allocation studies to capture the complex contexts in which seemingly discrete activities occur and the coordinative work that permeates daily life and binds together presumably separate domains of paid work, unpaid work, and leisure. This is not to discount the significant contributions of time allocation research, only to point out that allocating activities to time captures an important, but not determinative component of busyness. Likewise, efforts to directly link activities with perceived stress captures only part of the experience of being busy.

Household Ethnographies

Microlevel studies complement macrolevel time allocation studies and allow researchers to address additional questions. Ethnographic interviewing permits researchers to ask questions that are meaningful to people about their lives and how they organize and process information about them, regardless of researchers' analytical categories. This insider perspective is arguably closer to the experiential dimension of busyness. Observational studies, especially when accompanied by informant commentaries, allow researchers to appreciate the subtle shifting between scripts, values, and meaning associated with activities that is an essential but tacit part of busyness. Talking through sequences of e-mails or text messages to family members, strangers, coworkers, and other acquaintances, for example, is accompanied by shifts between intimacy and distance, risk, and safety that are difficult to elicit in interviews, much less surveys.

Microlevel studies also allow researchers to study the moral order that underlies decisions about allocating time. Busyness implies knowledge about what people can, should, or must do. Activities are not simply performed, but they also entail different kinds of commitments to people and to how individuals define themselves. Individuals navigate days marked by activities and commitments, and they keep track of where and who they are; these, too, are time-consuming internal activities that accompany the behaviors of preparing meals, transporting children, and reading books or looking at screens. Above all, being busy is about what is not done and therefore not seen, as much as it is about what we do.

Graesch (2009) describes the procedures used by researchers conducting week-long studies of 32 working families in Los Angeles. The methodology included videography of parents and children in their homes, and an ethnoarchaeological approach to mapping and documenting family assets. Activities were captured using a scan sampling method in which behavior is sampled at regular predetermined times. Kitchens emerged as 'hubs of busyness', with clocks, calendars, telephones, and message-covered refrigerators providing the infrastructure for time management. These were homes where families had increased leisure, but also segmented, hectic schedules.

Darrah et al. (2007) followed 14 dual-career families throughout a sampling of days, journeying with them to work and school, as well as on numerous errands. Fieldwork revealed intelligence gathering, planning, prioritizing, and purchases of technology to manage or cope with busyness rather than a simple speeding up or acceleration of time. It provided details of how time, tasks, location, and meaning came together in specific situations, with the household as a setting for experimenting with ways to manage busyness. The older contrast between work and leisure was subordinated to new activities, including many created by deregulation and consumerism in the neoliberal state.

New Directions in Theory and Research

Research and theory germane to busyness continues to be generated in different disciplines, even without a comprehensive research agenda. The primary challenge of studying busyness remains that of integrating theory and findings from different disciplines and fields in order to provide a more holistic approach to how it is conceptualized and investigated. Time allocation studies provide continuity within the study of busyness, although the categorizations and data do not specifically address the concept of busyness. Several social science disciplines bring additional concepts, data, and questions to the field.

Distinguishing the mechanisms that lead to different facets of busyness is essential. Southerton and Tomlinson (2005), for example, discuss three elements of being harried that are consistent with busyness. They identify the volume of time to complete the tasks that people judge to be necessary so insufficient time to complete them results in expressions of being harried. Coordination becomes more difficult when daily lives occur in the context of casual, flexible, or ad hoc structures that necessitate individualized scheduling, often in real time. Consumption occurs within networks and 'omnivores' with wide-ranging interests have more networks that increase the cost of coordinating schedules and feelings of being rushed. Daily tasks, too, may be situated in temporal rhythms that become intensified, resulting in expressions of juggling and multitasking. Busyness is accordingly not unidimensional and it involves often tacit coordination activities that largely elude time allocation studies, with their summative approach to being busy.

Interest in busyness is partially motivated by concern that it has unintended, harmful consequences for both children and adults, such as stress or limited attention. The costs of busyness may include effects on learning and sociability, as well as productivity and job turnover. Explicating these costs and linking them to potential remedies is important to individuals, families, and workplaces, and also to public policy. Shifting focus from individual adjustments and adaptations to institutional responses, such as workplaces that are flexible in responding to the idiosyncratic commitments of workers is increasingly linked to worker recruitment, retention, and productivity.

Changing the constraints upon daily lives complements enhancing individuals in their ability to cope with busyness. One approach is to provide workplace, instead of worker flexibility through practices such as offering employees more control over scheduling work hours, places of work (e.g., telecommuting), and the number of hours worked, and even results-only work environments in which workers control work scheduling as long as productivity goals are met. Such flexible arrangements offer benefits as well as costs and they are not appropriate for all conditions. Although data regarding their impact is limited, they do provide opportunities to experiment with environmental conditions in order reduce the effects of busyness.

Although busyness sometimes seems to permeate society, its manifestations can be distributed. It may also provoke resistance such as through various slow movements that are the antithesis of efficiency and productivity. Globalized capitalism binds together many people around the world through networks of rapid communication, access to information, and consumption, but there is also a slow world in which the pace of life is relaxed through deliberate public policies. Miele (2008) describes CittàSlow, an international movement aimed at diffusing the philosophy of 'slow food' throughout urban design and planning to improve quality of life. Such movements as slow food and slow cities provide opportunities for people to create alternatives to social structures in which busyness is inevitable and difficult to avoid.

Finally, ITs, and personal mobile devices in particular, are deeply entangled with busyness. IT's material affordances allow the efficiency and mobility that provides people with the flexibility of working under conditions they choose, even if those efficiencies are often translated into taking on additional activities that increase aggregate busyness. But IT also provides products that are embedded in leisure practices that people enjoy, just as they add more to do each day. Researchers and designers continue to explore the potential of IT to enhance communication, collaboration, and coordination among multiple individuals, support nomadic work lives, and integrate our physical and digital worlds, but also to reduce busyness (Snyder et al., 2007). New technologies that allow communication through video chat or coordinating e-books for children to read with family members at a distance, and even calendar sharing applications, highlight the importance of staying connected and organized, and including extended family in those routines (Cao et al., 2010). Researchers are also focusing on potentially negative consequences of increased technology use that may affect busyness. Turkle (2011), for example, argues that IT use may not only influence time allocation by minimizing boredom and loneliness, but it may also reduce opportunities and incentives for self-reflection and building deeper connections with other people.

As theory and concepts about busyness have changed, so have methodological approaches. A first methodological issue is how to integrate data that describe behavior with those which capture the experiential component of busyness. Direct observations, surveys, and interviews provide data about behavior or people's recollections of it. There is, however, a gap between such description and evidence of busyness. Claims of being busy or not can be based on the analysis of time allocations by a community of researchers or by individuals who are describing their own lives. However, busyness is difficult to establish without talking to people about how they experience and categorize their daily activities. The private, internal work of busyness – anticipating, planning, adjusting, prioritizing, etc. – is only accessible through open-ended interviewing conducted in the flow of activities. Eliciting and integrating data about these two very different components of busyness is a central methodological challenge.

A second challenge is to integrate quantitative and qualitative perspectives on busyness. Considered quantitatively, busyness can be measured by an amount or level of activity calculated by summing minutes or hours or assessment of stress. Alternatively, busyness may be considered qualitatively by focusing on the types of activities – their nature and variety, duration, differentiation by scripts or rules, etc. – that create distinctions among activities that are critical to being busy. Here it is not the amount of activity that matters, but the

differentiation and complexity of activities, as well as their coordination and management.

Third, busyness can be examined at the level of individual lives or at that of a society that generate busyness. The former is the more familiar approach and it sensitizes researchers to how individuals make choices about their use of time. But it takes for granted those activities among which choice is being exercised. The latter approach draws attention to the social processes that entangle individuals in activities regardless of their choices. In effect, it treats as problematical the very activities that individuals choose among and it explicates the processes and structures that characterize a society and which produce individuals who are willing and able to embrace busyness.

A final challenge is to further develop international, cross-cultural approaches to busyness as a phenomenon that transcends national boundaries. Explanations that rely on national conditions can be complemented and strengthened by a comparative approach and international time allocation studies provide a foundation (Gershuny, 2000). However, surveys alone may be inadequate to provide the requisite data (Sullivan, 2008). The challenge is conceptual as well since busyness requires analysts to understand activities in detail, as well as how they are coordinated and experienced, a profound exercise in cross-cultural comparison.

See also: Computer Mediated Communication; Coping across the Lifespan; Coping with Job Stress; Emotions and Work; Everyday Life, Anthropology of; Family Policy; Family and Work; Future Time Perspective and Motivation; Household Economics; Life Course and Event History Analysis; Mixed Methods; Modernity: Anthropological Aspects; North America: Sociocultural Aspects; Subjective Wellbeing, Psychology of; Sustainable Lifestyles; Time Use Research Methods; Time, Anthropology of.

Bibliography

Aguiar, M., Hurst, E., 2009. A summary of trends in American time allocation: 1965–2005. Social Indicators Research 93, 57–64.

American Heritage Time Use Study. Retrieved 10.02.14 from http://www.timeuse.org/ahtus/.

Barley, S.R., Meyerson, D.E., Grodal, S., 2011. E-mail as a source and symbol of stress. Organization Science 22 (4), 887–906.

Cao, X., Sellen, A., Bernheim Brush, A.J., Kirk, D., Edge, D., Ding, Xianghua, 6–10 February 2010. Understanding family communication across time zones. In: Paper presented at the 2010 ACM conference on Computer Supported Cooperative Work, Savannah, Georgia. ACM, New York, NY, pp. 115–158. http://dx.doi.org/10.1145/1718918.1718947.

Csikszentmihalyi, M., Larson, R., 1987. Validity and reliability of the experience-sampling method. Journal of Nervous and Mental Disease 175, 526–536.

Darrah, C.N., 2007. The anthropology of busyness. Human Organization 66 (3), 262–269.

Darrah, C.N., Freeman, J.M., English-Lueck, J., 2007. Busier than Ever!: Why American Families Can't Slow Down. Stanford University Press, Stanford.

Graesch, A.P., 2009. Material indicators of family busyness. Social Indicators Research 93, 85–94.

Gershuny, J., 2000. Changing Times: Work and Leisure in Postindustrial Society. Oxford University Press, Oxford.

Greenfeld, L., 2005. When the sky is the limit: busyness in contemporary American society. Social Research 72 (2), 315–338.

Hamermesh, D.S., 2009. It's time to "do economics" with time-use data. Social Indicators Research 93, 65–68.

Hamermesh, D.S., Lee, J., 2007. Stressed out on four continents: time crunch or yuppie kvetch? Review of Economics and Statistics 89, 374–383.

Hochschild, A., 1997. The Time Bind: When Work Becomes Home and Home Becomes Work. Macmillan, New York.

Krueger, A.B., Kahneman, D., Schkade, D., Schwarz, N., Stone, A.A., 2008. National Time Accounting: The Currency of Life. Working Paper #523. Princeton University Industrial Relations Section.

Linder, S., 1970. The Harried Leisure Class. Columbia University Press, New York.

Miele, M., 2008. CittàSlow: producing slowness against the fast life. Space and Polity 12 (1), 135–156.

Offer, S., Schneider, B., 2011. Revisiting the gender gap in time-use patterns: multitasking and well-being among mothers and fathers in dual-earner families. American Sociological Review 76 (6), 809–833.

Perlow, L.A., 1997. Finding Time: How Corporations, Individuals, and Families Can Benefit from New Work Practices. Cornell University Press, Ithaca, NY.

Robinson, J., Godbey, G., 1997/1999. Time for Life: The Surprising Ways American Use Their Time. University Park, Pennsylvania University Press.

Robinson, J., Martin, S., 2009. Changes in American daily life: 1965–2005. Social Indicators Research 93, 47–56.

Rosa, H., 2003. Social acceleration: ethical and political consequences of a desynchronized high-speed society. Constellations 10, 3–33.

Schor, J., 1993. The Overworked American: The Unexpected Decline of Leisure. Basic Books, New York.

Snyder, B.H., 2013. From vigilance to busyness: a neo-Weberian approach to clock time. Sociological Theory 31 (3), 243–266.

Snyder, M., Zimmerman, J., Forlizzi, J., 22–25 August 2007. Your dinner's calling: supporting family dinnertime activities. In: Paper presented at Designing Pleasurable Products and Interfaces, Helsinki, Finland. ACM, New York, pp. 485–489. http://dx.doi.org/10.1145/1314161.1314209.

Southerton, D., Tomlinson, M., 2005. 'Pressed for time': the differential impacts of a 'time squeeze'. Sociological Review 53, 215–239.

Sullivan, O., 2008. Busyness, status distinction and consumption strategies of the income rich, time poor. Time & Society 17, 5–26.

Turkle, S., 2011. Alone Together: Why We Expect More from Technology and Less from Each Other. Basic Books, New York.

Vercruyssen, A., Roose, H., Van de Putte, B., 2011. Underestimating busyness: indications of nonresponse bias due to work-family conflict and time pressure. Social Science Research 40, 1691–1701.

Wajcman, J., 2008. Life in the fast lane? Towards a sociology of technology and time. British Journal of Sociology 59, 59–77.

Relevant Websites

http://svcp.org/index.html.
http://www.timeuse.org/ahtus/.

Butler, Judith (1956–)

Birgit Schippers, St Mary's University College Belfast, Northern Ireland, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The American philosopher Judith Butler is one of the most influential scholars working in the field of contemporary critical theory. Best known for her contributions to gender theory and feminism, specifically for her account of gender as performative, her recent writings on international theory and global ethics, her critique of U.S. foreign policy in the wake of the wars in Iraq and Afghanistan, and her interest in the conflict in Israel–Palestine, have broadened her oeuvre and have made her work of interest to a wider public, beyond academia.

Biography

Questions of recognition, of who counts or matters, constitute a leitmotif of the writings of the American philosopher Judith Butler (1956–present). That her own work counts in the social sciences is beyond dispute, and no serious effort to survey the contemporary scene of social and political theory should omit her contribution. In fact, it is with considerable justification that Butler could be described as one of the most important contemporary critical theorists of the late twentieth and early twenty-first centuries. Her interventions into the feminist theory debates of the late 1980s and 1990s had a profound influence on the direction of contemporary feminist theorizing, and continue to shape the research agendas of many scholars working within this field. Terms such as performativity, vulnerability, and precariousness, introduced by Butler into wider contemporary critical discourse (see Butler, 1990, 2004a, 2009), firmly belong to the repertoire of social and political theorizing. Considered as one of the founders of queer theory, her unsettling of the category of gender and her deconstruction of gender's relationship to (biological) sex and desire remain hugely influential. Her more recent writings (2004a, 2009, 2012) on international theory and global ethics, her critique of U.S. foreign policy in the wake of the wars in Iraq and Afghanistan, and her interest in the conflict in Israel–Palestine have broadened her oeuvre beyond the boundaries of feminist scholarship and gender studies; they have also extended her public appeal beyond academia.

At present, Butler is Maxine Elliot Professor in the Department of Rhetoric and Comparative Literature and the Co-Director of the Program of Critical Theory at the University of California, Berkeley. She also holds the Hannah Arendt Chair of Philosophy at the European Graduate School in Switzerland, and she is Wun Tsun Tam Mellon Visiting Professor of the Humanities at Columbia University, New York. Previous academic appointments include two short periods at Wesleyan and John Hopkins universities. Butler is the recipient of many prizes and honors, including the Adorno Prize, awarded by the City of Frankfurt, Germany, for outstanding performances in the fields of philosophy, music, theater, and film. She is also a regular contributor to public events, such as those associated with the boycott, divestment, and sanctions (BDS) movement against Israeli policies in the occupied territories; she was a speaker at Occupy Wall Street and she served on the International Gay and Lesbian Human Rights Commission.

Butler's scholarly background lies in continental philosophy, and much of her writing is informed by the interplay between German and French ideas, specifically the works of Hegel and Foucault. We learn about her philosophical journey in *Subjects of Desire: Hegelian Reflections in Twentieth-Century France* (1987/1999), a book based on her doctoral thesis. In the preface to the book's second edition, Butler narrates the history of her scholarly training, including her graduate studies in German philosophy, specifically German idealism and, above all, the work of Hegel. Her interest in phenomenology led to a period of study in Heidelberg on a Fulbright scholarship, where she took classes with Hans-Georg Gadamer (1900–2002). Butler's focus on German philosophy swiftly broadened to encompass twentieth-century French thought, such as the writings of Simone de Beauvoir (1908–86), whose ideas underpin some of her early publications (see Butler, 1986, 1987). Hegel's philosophy continues to shape Butler's thinking to date, but following her completion of her doctorate at Yale University in 1984, she broadened the range of her philosophical interlocutors to include the ideas of French post-structuralism, in particular the works of Michel Foucault and Jacques Derrida. More recent texts engage with the writings of Theodor W. Adorno, Hannah Arendt, Walter Benjamin, Emmanuel Levinas, as well as Edward Said. Notwithstanding her initial reluctance to embrace psychoanalytic thought, Butler has also been profoundly influenced by the writings of Freud (see Butler, 1997a), and she pursues a critical engagement with the ideas of Jacques Lacan, Melanie Klein, and Jean Laplanche.

Butler's scholarly formation is closely intertwined with the context of her upbringing, which she articulates through the trope of trouble. To fully appreciate the development of her ideas, it is important to acknowledge two formative influences on her thinking: these are her upbringing as a Jew and her exposure to Judaic ideas as part of her education in her synagogue, and the lived experience of gender and sexuality in a heteronormative culture. In a recent interview (2011), Butler recounts how, as punishment for being troublesome, she was sent to classes with the rabbi. However, rather than experiencing these classes as penalty, she saw these lessons as an opportunity to engage with a set of ethical and philosophical questions that continue to occupy her thinking in the present: these are questions to do with ethical responsibility, the role of Jewishness, and wider questions of identity formation. The classes also introduced Butler to the works of Spinoza, whose conception of the *conatus*, interpreted by her as the desire to

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61319-3

persist in one's being, has exercised a significant influence on her thought, articulated in her recent focus on the notion of livability (Butler, 2004b; see also Lloyd, 2008; Schippers, 2015). Equally significant, and perhaps better known, is her engagement with feminism and gender politics in the context of a heterosexist culture. As she tells her readers in her most famous book to date, *Gender Trouble: Feminism and the Subversion of Identity* (1990), the fate of two gender-troubled uncles and her own sexual identity formation have contributed to an early interest in questions of sexuality and its relation to gender. Central to the later development of this awareness is Butler's insistence on the power exerted by hegemonic regimes of gender regulation. It is this latter aspect that has informed her work on gender, identity, and feminism. These key themes weave themselves through Butler's oeuvre, beginning with early writings in philosophy, right up to her most recent texts on binationalism and global protest. Notwithstanding developments and additions in her thinking, her theoretical apparatus remains anchored in a set of key categories and approaches that stem from her early work on Hegel and Foucault, and the wider framework of continental philosophy.

Major Works

Desire, Power, and the Performativity of Gender

Given the wide range of influences on the development of Butler's ideas, it is impossible to align her work with one particular school of thought. However, three frameworks continue to inform her scholarly endeavors: post-structuralism, existential phenomenology, and psychoanalysis. These approaches also configure the persistent themes in Butler's work, which, although addressed in different contexts and at different stages of her academic career, evolve around questions of political subjectivity and conditions of alterity. As already mentioned, two thinkers feature prominently in the development of Butler's work: the French philosopher Michel Foucault and the towering figure in modern German philosophy, G.F.W. Hegel. While Foucault's conception of the subject as imbued with power has been hugely influential to Butler's writings, she borrows from Hegel her continued attention to questions of alterity and to recognition, specifically, how to engage with the other under conditions of unknowingness, and how to relate to the other upon whom the subject is dependent for its very physical survival. These themes feature prominently in Butler's recent writings on international theory (see Butler, 2004a, 2007, 2009), but they surface early on in her work. Of particular importance to the development of this early work on gender is her deployment of Hegel's concept of desire. In *Subjects of Desire: Hegelian Reflections in Twentieth-Century France* (1987/1999), Butler pursues the notion of desire in its encounter with the historicizing impetus and imperative articulated in the French reception of Hegel, specifically in the works of Alexandre Kojève, Jean Hyppolite, and Jean-Paul Sartre. Although Butler later belittles her book in the preface to the second edition as her juvenilia, *Subjects of Desire* flags up important themes that continue to inform her writings; these include her attention to conditions of alterity, recognition, and identity formation; they also encompass the historicizing impetus of power; and, central to the further development of her work, the acknowledgment of corporeality and affect, and their implication in power.

Subjects of Desire's concluding chapter, which covers the treatment of desire in the works of contemporary French theorists, including Lacan, Foucault, Derrida, and Kristeva, connects her early phenomenological studies with her later Foucauldian and Derridean writings, for which she is best known. A key figure in this transition is Simone de Beauvoir, whose relevance to Butler's thinking is twofold: first, de Beauvoir provides the link to Butler's work on gender and feminism; second, de Beauvoir's Hegelian readings, specifically the idea of becoming, prepare Butler's own reading of gender as performative (see below). The engagement with structuralist and post-structuralist influences, which surface toward the end of *Subjects of Desire*, together with her interest in existential phenomenology, shape the future direction of Butler's work, and they culminate in the publication of *Gender Trouble: Feminism and the Subversion of Identity* (1990). This book attends to three issues, considered as deeply problematic by Butler: (1) the representational claims of feminism; (2) the dominance of expressive theories of gender; and (3) structuralist accounts of sex, gender, and the body.

What underpins the discussion in *Gender Trouble* is Butler's criticism of feminism as a representational political project. In other words, she challenges those who claim that feminism can speak straightforwardly on behalf of women. Her ambition is both epistemological and normative: for her, political projects cannot and should not presume the homogeneity of the subjects they seek to represent. However, rather than abandoning feminism as a worthwhile political project, her assertion regarding the heterogeneity of the subjects of feminism, as well as the decentered nature of the category 'woman,' constitute the point of departure for her reformulation of feminism and gender. What drives the argument of *Gender Trouble* is Butler's discontent with expressive accounts of gender. She challenges in particular the sex-gender analogy, that is, the idea that gender, understood as the historical, social, or cultural representation of a biological body, relates in any direct way to the body. Instead of emanating from, or expressing, the 'truth' of a biological body, gender is said to be performative: it is a never-ending process of *doing* one's gender (rather than *being* a gender), conducted under constraint, restricted and policed by the norms of gender, and operating in what she terms a 'heterosexual matrix.' This matrix is a generative framework that produces heterosexual gender as normal, and that marginalizes nonnormative genders and sexual minorities. Further, the desire that is said to emanate from this sex-gender analogy is also deconstructed, opening up possibilities for different articulations between (sexed) body, gender identity, and sexual desire.

Gender Trouble's influence on the direction of feminist theory in the 1990s, and the controversy surrounding some of its key assertions, cannot be underestimated. Sympathetic readings have highlighted the value of Butler's ideas for a politics of difference that takes seriously the internal differentiation and heterogeneity of the subject of feminist politics (see Lloyd, 2007). Further, her Foucauldian-inspired analysis

of the body has contributed a great deal to scholarship that seeks to trace and illuminate the historical discursive constitution of sexed corporeality. Much of the criticism of Butler's ideas derives from her reference to drag, which she deploys as an allegory for practices of gender performativity. Drag, according to Butler, dramatizes and illustrates the illusion of expressive accounts of gender; that is, the claim that gender expresses the truth of the sexed body. Despite Butler's insistence that her use of drag served merely illustrative purposes, it was (mis-)read by some of her critics as a normative assertion of gender as a voluntaristic expression (see Benhabib, 1995; Nussbaum, 1999). Such a voluntarist account, according to her critics, constitutes a serious oversight of the deeply entrenched injustices that emerge from women's corporeal vulnerability and that limit the potential for women's agency. Butler's Foucauldian-inspired critical genealogy of the subject proved to be equally controversial, interpreted by some critics as an account of the death of the subject and a denial of female agency (see Benhabib, 1995). A more challenging debate ensued around Butler's account of the body and the status of its materiality, or 'fleshiness.' These questions are taken up in *Bodies That Matter: On the Discursive Limits of* "Sex" (1993), where Butler revisits and clarifies some of her ideas presented in *Gender Trouble*; whereas *Gender Trouble* maps and critiques the conceptual terrain on which feminist theorizing of gender has operated, *Bodies That Matter* attends to the historicity of the body's materiality, and of the body's implication in (heteronormative) discourses of power.

Toward the end of the 1990s and into the 2000s, Butler's writings open up new areas of critical inquiry, focusing on international theory and (global) ethics (see Butler, 1996, 2004a, 2005, 2009). However, she keeps returning to gender theory and feminism, and to questions of sexuality and desire. These themes are brought together in *Antigone's Claim: Kinship between Life and Death* (2000a). Profitably read as a treatise on the gendered nature of speech, and as a development of Butler's previous work on speech act theory and the performativity of language (see 1997b), *Antigone's Claim* also constitutes a significant intervention into public debates on marriage equality. The book is underpinned by a philosophically informed discussion on the vagaries of kinship structures in the context of a heterosexist culture, and it deconstructs the contours of these heterosexual kinship arrangements. In *Undoing Gender* (2004b), a collection of essays on what she terms the New Gender Politics, an alliance of feminist, gay, lesbian, and trans and intersex movements focused on the pursuit of progressive sexual politics, Butler tracks the radical challenge, emanating from trans and intersex movements to the categories of gender and feminist theory, specifically to the idea of womanhood, of binary conceptions of gender, and, yet again, of feminism's representational claims.

Power and the Psyche

The trans and intersex movements also expose the deeply entrenched violence at work in the operation of gender norms, and in the policing, regulation, and forced correction of bodies. To map such deep-rooted subjection to norms, Butler draws on the ideas of Sigmund Freud, which she reads through the lens of Hegelian and Foucauldian philosophy. In *Gender Trouble*, she presents a critical reading of Freud's narrative of the acquisition gender identity that challenges its alleged heterosexist presuppositions. She returns to Freud's ideas in *The Psychic Life of Power: Theories in Subjection* (1997a), where she ponders the persistent attachment of subjects to their own subjection. Supplementing her reading of Freud with insights from Hegelian and Foucauldian philosophy, she builds on Freud's concept of melancholia, arguing that the melancholic attachment to a same-sex love object becomes incorporated as the barred prohibition of homosexuality. This internalized prohibition of homosexuality contributes to the development of (heterosexual) conscience, and it explains the often violent abjection of homosexual identification.

This development in Butler's theory of gender is significant because it puts to rest any assertions relating to an alleged voluntarism in her theory of the subject. Equally significant to Butler's discussion in *The Psychic Life of Power* is her continued reading of Hegel and Foucault. Of key importance to Butler's Hegel reception is the *Phenomenology of Spirit* (1977), in particular Hegel's narrative of the master–slave dialectic and the ensuing struggle for recognition. *The Psychic Life of Power* develops this reading of the *Phenomenology* by drawing on the Foucauldian notion of *assujetissement*. Thus, reading Hegel through Foucault, Butler articulates a concept of modern subjectivity as paradoxical. Becoming a subject entails subjection to norms and to an other, which the subject experiences as deeply agonizing. Yet at the same time, subjection brings the subject into existence as intelligible and recognizable, enabling being and persistence (hence the link to Spinoza). Developing this insight, Butler concludes that the subject is passionately attached to its own subjection: for Butler, there is no subject without subjection. The idea of a passionate attachment to subjection is also instrumental in the development of her thought after 9/11 and her turn to questions of ethics.

International Theory, Universality, and the Turn to Ethics

One of the more heavily debated topics in the recent Butler scholarship relates to her so-called turn to ethics (see Dean, 2008; Lloyd, 2008; Rushing, 2010; Schippers, 2015). This aspect of Butler's work has puzzled many of her readers because it seems to run counter to a concern, expressed in her earlier work, that the alleged universalist and formalist assumptions of ethics may be at odds with the kind of Foucauldian investigations to which she remains committed; ethics, she worried, may replace or displace a concern with power and politics. However, building on her reformulation of the critique of universalism (see 1996, 1997b, 2000b), her recent work espouses the importance of ethics and its relevance for understanding politics in the post-9/11 era. This interest in the political fallout of 9/11 should be considered alongside her foregrounding of a new set of political concepts, such as her pursuit of the question of responsibility, which she articulates in relation to post-9/11 politics, and in her discussion of Zionism and the conflict in Israel–Palestine. It also informs her ongoing interest in conceptions of the subject: although Butler remains attuned

to subjectivity's gendered dimension, she has added to her repertoire by considering subjecthood in the context of global interdependence, and by pondering its ethnic and racial framing under conditions of war.

The source for this interest can yet again be traced back to her early work on Hegel and existential phenomenology; there, Butler articulates a narrative of the subject as relational and fundamentally dependent upon others. Its political context lies in the events surrounding the attacks on September 11, and the ensuing military invasions of Afghanistan and Iraq. It is in the encounter with these political events that Butler reformulates her questions and interests, and reworks her conceptual apparatus, which lead her to foreground the notions of precariousness and precarity, vulnerability, violence and grievability, and livability. In her two most significant works of the last decade, *Precarious Life: The Powers of Mourning and Violence* (2004a) and *Frames of War: When Is Life Grievable?* (2009), Butler challenges those assertions which insist on the unconditional sovereignty of the state and its right to engage unilaterally in acts of war. Against such assertions of sovereignty, she champions instead an account of global interdependence, which would reconfigure dominant patterns of recognizability, and which could sustain the livability of populations across diverse ethnic and racial frames. Building on her critical reading of the ideas of Emmanuel Levinas, Butler suggests that human life is subject to a fundamental vulnerability and precariousness, dependent upon the care provided by others. Such care, however, is distributed unequally, leaving some populations more exposed to violence than others.

Jewish Philosophy and the Politics of Cohabitation

Much of Butler's criticism of U.S. foreign policy since 9/11 turns on her effort to articulate a critique of violence that acknowledges violence's paradoxical role as simultaneously destructive and constitutive. This critique of violence, conducted in the context of the war on terror, the military invasions of Iraq and Afghanistan, and the treatment of detainees at Abu Ghraib and Guantánamo Bay, informs another area of scholarly inquiry and public engagement: this is Butler's intervention into the conflict in Israel–Palestine. Two issues occupy her writings in this area: first, she challenges those positions that frame criticism of Israel as inherently anti-Semitic (see Butler, 2004a, 2012). Developing her discussion of hate speech (see Butler, 1997b), she aims to articulate a specifically Jewish position from which to criticize Israeli state violence. Such a position would not only challenge the claim that criticism of Israel is, by definition, anti-Semitic; it would also put forward a distinctively Jewish critique of violence from which to judge the violent actions of the (Israeli) state.

A further facet of Butler's engagement with Jewishness pertains to practices of state formation and citizenship in the context of heterogeneous populations. In brief, she seeks to formulate conceptions of a polity beyond nationalism that would facilitate heterogeneous populations, such as those inhabiting the territory of Israel–Palestine, and to construct forms of citizenship that accommodate plurality. These themes are outlined in one of Butler's most recent books, *Parting Ways: Jewishness and the Critique of Zionism* (2012). Drawing on work by Hannah Arendt, Edward Said, and Emmanuel Levinas, Butler thinks through prospects for cohabitation in multiple inhabited spaces, across boundaries, and within contested political spaces. Such considerations of citizenship also inform her reflections on the fate of illegal migrants and their precarious status. This latter point features in *Who Sings the Nation-State? Language, Politics, Belonging* (2007), a conversation with Gayatri Spivak on the role of the nation-state under conditions of migration, multiculturalism, and global interdependence. Both considerations, on conditions of contested state borders and on migration, are underpinned by her critique of the nation-state, which Butler, following the work of Arendt, regards as structurally prone to producing and expelling minorities. As yet, this aspect of Butler's writings is not as developed as her ideas on gender and feminism, or her discussion of vulnerability and precariousness, but it opens up a new field which explores the prospects for citizenship under conditions of heterogeneity and plurality, and which connects her recent writings with current debates in international theory and in political studies and sociology more widely.

Impact

It would be imprudent to identify the lasting impact of a living writer who produces scholarly output at the staggering rate that Butler does. Judging by Butler's most recent work, on Jewishness, Zionism, and the critique of Israel (2012), and on the politics of the performative in the era of global protest (2013), it is fair to claim that her writings continue to bear out her attentiveness to significant developments in the politics of the left. Of particular interest to her readers has been Butler's explicit connection between pressing political problems and the sophisticated philosophical considerations that underpin her analyses. Her pronouncements on global ethics, her advocacy of binationalism, and her criticism of state violence, as well as her work on Jewishness, are underpinned by her articulation of a relational account of the subject and the assertion of the subject as vulnerable, and by her reflections on what she terms a politics of dispossession, out of which ethical obligations toward (potentially distant) others arise. Her writings on ethics and violence have already made a substantial impact on the Butler scholarship (Chambers and Carver, 2008; Lloyd, 2008; Rushing, 2010; Schippers, 2015.). But while her recent work has received some attention in the field of international theory (Schippers, 2014), it has, as yet, not attained the same level of attention as her earlier work on gender. Given the relative novelty of her recent writings on Jewishness and binationalism, and her criticism of Israel, it is too early to establish their significance beyond the immediate Butler scholarship.

To give a fair assessment of Butler's wider impact, it is important to acknowledge her influence beyond the boundaries of the academy. Butler is a sought-after speaker at scholarly and nonscholarly events alike, and her recent works on international politics, such as her criticism of U.S. foreign policy and her support for the BDS campaign, have

opened her work and ideas to new audiences. The importance accorded to interventions into public policy is also reflected in the outlets of Butler's contributions: she is a regular interviewee for scholarly and nonscholarly publications, and many of her interviews and speeches can now be accessed online (many of these are available on YouTube). However, Butler's public statements, specifically in relation to Israel and her support for the BDS movement, have also left their mark and led to a series of hostile encounters and indeed threats.

Undoubtedly, it is in the fields of gender studies, feminism, and queer theory that her impact is most widely felt. Her writings have led to numerous and still growing publications, both critical and sympathetic. Her deployment of the trope of 'trouble,' which opens her discussion in *Gender Trouble*, has inspired those readers who, like Butler, aim to unsettle hegemonic conceptions of gender, gender relations, and the role of feminism. Despite the, at times stinging, criticism of Butler, notoriously described by Martha C. Nussbaum as 'Professor of Parody' (1999), it is with the notion of performativity that she has made a lasting impact on the theorizing of gender for more than two decades. Further, Butler's work on the new gender politics (2004b) has broadened the appeal of her ideas beyond the initial concern for the alleged heterosexist presumptions of feminist theory. This focus on her enormous contribution to gender theory should not cloud her equally significant role in the development of contemporary critical thought beyond gender. While her scholarly contributions alone, specifically her articulation of a relational conception of the subject, her theory of performativity, or her reworking of universalism are of considerable merit, they gain additional force through the way that they relate to the political and social problems of the early twenty-first century. One can only hope that Butler will continue to stimulate her readers with more work.

See also: Arendt, Hannah (1906–75); Beauvoir, Simone de (1908–86); Femininities; Feminist Political Theory and Political Science; Foucault, Michel (1926–84); Freud, Sigmund (1856–1939); Gender and Women's Studies, Applied Research on; Gender: Gendering of Categories; Hegel, Georg Wilhelm Friedrich (1770–1831); Heterosexuality; Identity and Identification; Judaism; Power; Responsibility: Philosophical Aspects.

Bibliography

Benhabib, S., 1995. Feminism and postmodernism. In: Benhabib, S., Butler, J., Cornell, D., Fraser, N. (Eds.), Feminist Contentions: A Philosophical Exchange. Routledge, New York and London, pp. 17–34.

Butler, J., 1986. Sex and gender in Simone de Beauvoir's second sex. Yale French Studies 72, 35–49.

Butler, J., 1987. Variations on sex and gender: beauvoir, wittig and foucault. In: Benhabib, S., Cornell, D. (Eds.), Feminism as Critique: Essays on the Politics of Gender in Late-Capitalist Societies. University of Minnesota Press, Minneapolis, pp. 128–142.

Butler, J., 1990. Gender Trouble: Feminism and the Subversion of Identity. Routledge, New York and London.

Butler, J., 1987/1999. Subjects of Desire: Hegelian Reflections in Twentieth-Century France. Columbia University Press, New York.

Butler, J., 1993. Bodies that Matter: On the Discursive Limits of "Sex". Routledge, New York and London.

Butler, J., 1996. Universality in culture. In: Cohen, J. (Ed.), For Love of Country: Debating the Limits of Patriotism. Martha C. Nussbaum with Respondents. Beacon Press, Boston, pp. 45–52.

Butler, J., 1997a. The Psychic Life of Power: Theories in Subjection. Stanford University Press, Stanford.

Butler, J., 1997b. Excitable Speech: A Politics of the Performative. Routledge, New York and London.

Butler, J., 2000a. Antigone's Claim: Kinship between Life and Death. Columbia University Press, New York.

Butler, J., 2000b. Contingency, Hegemony, Universality: Contemporary Dialogues on the Left. With Ernesto Laclau and Slavoj Žižek. Verso, London and New York.

Butler, J., 2004a. Precarious Life: The Powers of Mourning and Violence. Verso, London and New York.

Butler, J., 2004b. Undoing Gender. Routledge, London and New York.

Butler, J., 2005. Giving an Account of Oneself. Fordham University Press, New York.

Butler, J., 2007. Who Sings the Nation-State? Language, Politics, Belonging. With Gayatri Spivak. Seagull Books, Oxford and New York.

Butler, J., 2009. Frames of War: When Is Life Grievable? Verso, London and New York.

Butler, J., 2011. There are some muffins there if you want….A conversation on queerness, precariousness, binationalism, and BDS. In: Aloni, U. (Ed.), What Does a Jew Want? On Binationalism and Other Specters. Columbia University Press, New York, pp. 204–227.

Butler, J., 2012. Parting Ways: Jewishness and the Critique of Zionism. Columbia University Press, New York.

Butler, J., 2013. Dispossession: The Performative in the Political. With Athena Athanasiou. Polity Press, Cambridge.

Carver, T., Chambers, S.A. (Eds.), 2008. Judith Butler's Precarious Politics: Critical Encounters. Routledge, London and New York.

Chambers, S.A., Carver, T., 2008. Judith Butler and Political Theory: Troubling Politics. Routledge, London and New York.

Dean, J., 2008. Change of address: Butler's ethics at sovereignty's deadlock. In: Carver, T., Chambers, S.A. (Eds.), Judith Butler's Precarious Politics: Critical Encounters. Routledge, London and New York, pp. 109–126.

Hegel, G.W.F., 1977. Hegel's Phenomenology of Spirit. Oxford University Press, Oxford.

Lloyd, M., 2007. Judith Butler: From Norms to Politics. Polity Press, Cambridge.

Lloyd, M., 2008. Towards a cultural politics of vulnerability: precarious lives and ungrievable deaths. In: Carver, T., Chambers, S.A. (Eds.), Judith Butler's Precarious Politics: Critical Encounters. Routledge, London and New York, pp. 92–105.

Nussbaum, M.C., 1999. The professor of Parody: the hip defeatism of Judith Butler. The New Republic 22, 37–45.

Rushing, S., 2010. Preparing for politics: Judith Butler's ethical dispositions. Contemporary Political Theory 9 (3), 284–303.

Schippers, B., 2014. The Political Philosophy of Judith Butler. Routledge, London and New York.

Schippers, B., 2015. Violence, affect and ethics. In: M. Lloyd, M. (Ed.), Butler and Ethics. Edinburgh University Press, Edinburgh.

C

Cambridge School of Intellectual History

Richard Whatmore, School of History, The Scores, University of St Andrews, St Andrews, Fife, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Cambridge School of Intellectual History is most associated with the work of Quentin Skinner and John Pocock, but encompasses the work of a number of scholars of political thought and political philosophy who have had connections with the University of Cambridge from the 1950s to the present. Members of the Cambridge School, who now teach at institutions of higher education across the globe, share a commitment to the historical study of texts, and reject approaches characterized by presentism and teleology (labeled prolepsis). They also reject canons of dominant authors and histories of philosophy that evaluate past ideas. Members of the Cambridge School have accordingly been in the vanguard of a historical turn in the study of past politics and ideas more generally, alongside the similarly historically minded approach, albeit distinctive methodologically, of Begriffsgeschichte or Conceptual History pioneered by Otto Brunner, Werner Conze, and Reinhart Koselleck. Beyond a belief in the need to establish a context for the understanding of historical texts, and distaste for grand or theory-inspired narratives, the members of the Cambridge School have conflicting views of the origins and nature of modern political thought. This is illustrated through a comparison of the work of Skinner and Pocock.

Introduction

The philosopher R.G. Collingwood, in his posthumously published *The Idea of History* (1946) made the argument that the past could not be understood by reference only to what he called the 'outside' of events, the facts about bodies and their movements. It was additionally necessary to describe the 'inside' of actions, or the relationship between an act and what was later termed its ideological context. The famous example Collingwood gave was of Julius Caesar's assassination. The murder could be described by reference to the assault on Caesar near Theatre of Pompey on the Ides of March, the descent of the knives into his flesh, and the resulting flow of blood. But this did not explain why it happened. Understanding Caesar's death entailed knowledge of Caesar's mind and those of his enemies Gaius Cassius Longinus and Marcus Junius Brutus. Caesar was killed because of his recent ascent to the position of long-standing dictator (dictator perpetuo). He was seen as a tyrant and a proponent of political ideas that clashed with the assassins' perspective on the nature of the Roman Republic and of constitutional law. In short, the reconstruction of ideas expressed by historical actors was necessary to any act of historical understanding.

Only 3 years after the appearance of Collingwood's book Peter Laslett published an edition of the political writings of Robert Filmer, the seventeenth-century advocate of the divine right of monarchs. He spent the following years working on something more monumental; a critical edition of John Locke's *Two Treatises of Government*, which appeared in 1960. Laslett was seeking to date the authorship of texts, and used evidence drawn from contemporary correspondence and commentaries to reveal that Filmer's *Patriarcha* had been composed prior to his other writings, but was only published posthumously between 1679 and 1680. More surprisingly still, Locke's classic work of political thought, rather than being a defense of the 'Glorious Revolution,' as it had traditionally been described, having appeared in print from 1690, was shown to have been written around 1681, at a time when Whigs of Locke's stamp were contemplating violence against the Stuart court. Laslett's editions raised questions for historians concerning the relationship between an author's intentions in writing a text and the intentions behind its publication, causing in the process a wholesale reevaluation of late seventeenth-century political thought. Collingwood's demand that historians think about the ideas of historical actors in their own time, and Laslett's that an author's intentions could be reconstructed from textual evidence, led to the establishment of the Cambridge School.

In treating Caesar as a historical individual whose ideas were accessible through his writings, Collingwood was arguing against approaches to history that emphasized the special qualities of great men. In arguing that Locke was not the author whose writings had justified the Glorious Revolution of 1688–89, but rather an activist in politics whose ideas for rebellion had come to nothing a decade earlier, Laslett was directly questioning traditional Whig historiography, in which liberty was readily identifiable, acquired and defended in England, from the time of the Magna Carta, an approach associated with Henry Hallam's *Constitutional History of England* (1827) or Thomas Babbington Macaulay's *The History of England from the Accession of James II* (1848). Whig approaches assuming historical progress and linear development were

lauded by readers as justifications of the glorious national present, and as such sustained the role of the historian as a public intellectual and sometimes as a public moralist. The point the critics of the Whigs were making was that this was all well and good, but the result was not history. True historians had to be skeptical of grand narratives, of presumed straightforward connections between the present and the past, and of teleological accounts founded on the assumption of gradual improvement. Herbert Butterfield made all of these arguments in *The Whig Interpretation of History* (1931), attacking the presumed causal linkage between progress, Protestantism and liberty that he saw as underlying nineteenth-century Whig historiography. The following generation of historians in Cambridge became skeptics, seeking to challenge what they held to be the formulaic and inaccurate philosophies of history of the overconfident Victorians and their successors.

The Cambridge Method

What exactly an anti-Whig historiography would look like remained uncertain, especially because Butterfield himself was doubtful that it could be written in practice, as evinced by his own *The Englishman and His History* (1944). But the following generation of historians took the challenge seriously. In the 1950s and 1960s, substantive reinterpretations of early modern intellectual history were authored by Felix Gilbert, Caroline Robbins, Hans Baron, John Pocock, and Bernard Bailyn. The work of these historians emphasized discontinuity, unintended consequences, tragic failures, and lost traditions of political argument; investigated 'minor' figures and used them to illuminate neglected elements of the thought of figures traditionally deemed canonical; and opposed overarching narratives or those inspired by philosophers' perspectives upon history, and especially that of Karl Marx. Attacks on the methods employed in writing such Whig histories, the North American equivalent of which was the notion of evolving 'unit ideas' associated with the work of A.O. Lovejoy, were written by John Pocock ('The History of Political Thought: A Methodological Enquiry'), John Dunn ('The Identity of the History of Ideas'), and Quentin Skinner ('Meaning and Understanding in the History of Ideas') between 1962 and 1969, all then associated with the University of Cambridge. They have come to be seen as classic statements about the practice of intellectual history in the area of political thought. Skinner's 'Meaning and Understanding' proved especially influential, and has been described as the manifesto of the 'Cambridge School,' in part because it identified opponents most clearly, and condemned with particular verve descriptions of classic texts as the 'sole object of inquiry,' and as contributing to a 'dateless wisdom.' Skinner attacked the presumed connections between canonical authors by reference to presumed historical continuities and political philosophies. He argued that historians of political thought were attributing to past authors concepts that could not have been available to them (the sin of anachronism) and sought in the work of historic authors anticipations of arguments, which could not actually be found in the texts (the sin of prolepsis). Skinner demanded that the arguments of texts be treated as acts performed in history. Such arguments needed to be related to the discourses of the time, identified as the context, being related or opposed to arguments articulated by people who were living at the time of the author and whose work would have been known to the author. Skinner said that the goal of the historian was to reveal what the author of a particular text 'was doing,' encompassing what the author had intended to do and had succeeded in doing as interpreted by the responses of other authors. One of Skinner's most fascinating points was that the language or discourse in which an author was working, meaning the set of assumptions that he was adopting and employing in the articulation of his arguments, set limits to the argument itself. Authors could innovate and alter existing languages, but the point was to search for exactly this because existing languages were employed in articulating what was happening in the ideological and material present.

In each of their essays, Pocock, Dunn, and Skinner proposed that texts be treated as products of specific historical contexts. In working out the meaning of the texts, Dunn and Skinner identified the intentions of an author as the principal guide to the nature of the text, although neither unproblematic as an intellectual objective, nor sufficient in terms of understanding an author's work. The three authors opposed approaches founded on the presumption of fixed concepts of historical analysis, of meta-theoretical assumptions about human nature, and of opaque or a historical theoretical vocabularies. Within the historical fraternity, one goal of the attempt to analyze utterances in political texts as 'speech acts' was to liken them to the 'acts' analyzed by fellow historians, thereby countering prevalent Namierite opposition in Anglophone circles to the study of ideas (Lewis Namier had called intellectual history 'flapdoodle' in his England in the Age of the American Revolution published in 1930). This aspect of the enterprise has been the most obviously successful, as a generation of scholars has confidently described themselves as intellectual historians, formed professional associations, and acquired academic positions in this field. At the same time the methodology of the Cambridge School has been attacked by Straussians, Marxists, empirical political scientists (behaviorists), and post-structuralists. Critics have questioned the capacity of any scholar to identify an author's intentions, or to get 'inside the mind' or a historical actor, and at the same time have stated that a method that limits itself to the recovery of past ideas, and rejects any evaluation of such ideas as anachronistic, amounts to the worst kind of antiquarianism.

Pocock's and Skinner's Historical Writing

Among the major questions for historians in the 1970s was the extent to which use of the 'Cambridge' method would alter contemporary perspectives upon the history of political thought? Amidst a range of revisionist interpretations, Pocock's *The Machiavellian Moment* (1975) and Skinner's *Foundations of Modern Political Thought* (1978) have been seen to be seminal, and more especially interesting because of their earlier methodological essays. Pocock's *The Machiavellian Moment* followed his earlier *The Ancient Constitution and the Feudal Law* (1957) in leading historians to broaden their definition of early modern political theory, which was shown to encompass civil law, theology, and political economy, by contrast with many narrower twentieth-century conceptions, which had inspired

altogether different readings of history. *The Machiavellian Moment* dealt with changing concepts of time from Aristotle to the American Revolution and the 'Machiavellian' reflection upon the foundation and crisis of republican states. Skinner's *Foundations* commenced with a volume on Renaissance political theories and a second volume on 'The Age of the Reformation,' describing the gradual identification of the state with an artificial person standing above the unruly elements of political society, being a potential solution to endemic religious conflict. Skinner's work ended with scrutiny of the resistance theories of the early seventeenth century, by which time modern politics were in his view discernable. Each work generated a great deal of scholarly comment, much of which has linked them together as the most important studies to date of early modern political thought in Europe and the Atlantic World. They deserve particular attention as test cases of whether the practice followed the method, and successfully changed the way intellectual historians worked if they adopted the 'Cambridge' standard.

Skinner and Pocock agreed on the importance in early modern history of the articulation of a view of human flourishing developed in the Italian city republics from the thirteenth century onwards, in which a life lived in an independent republic was praised as natural to man, and in which civic activity to sustain this state was described as being the highest secular aspiration. They agreed on little else. Neither Pocock nor Skinner had ever anticipated agreement about intellectual inquiry. But the extent of the disagreement was perhaps surprising, given the extent to which each shared, broadly, the same approach to the past in terms of method. Many have seen such disagreement as indicating that the Cambridge method would not lead to more obviously improved understanding of the past, and have questioned whether the Cambridge School exists, beyond a rejection of teleology and prolepsis.

Pocock's history of civic humanism extended chronologically and geographically, describing James Harrington's adaptation in the mid-seventeenth century of Machiavelli's perspective on the history of Rome, according to which the practice of arms and inheritable landed property were the precondition for the enjoyment of liberty and the exercise of civic virtue. Pocock went on to detail the controversy over this perspective in the eighteenth century, in the very different context of ideas about liberty and autonomy that accompanied the end of the wars of religion and the rise of large and competing commercial monarchies in Europe. Maintaining these new forms of state required the development of standing armies and public credit; these were defended by association with modern forms of politeness, and praised by such men as Daniel Defoe alongside consumption and financial independence, in societies organized in accordance with the division of labor. 'Neo-Harringtonians,' as Pocock termed them, such as Andrew Fletcher, favored ancient virtue, protected and sustained by militias and an elite of landowners, whose interest in the state ensured their political wisdom and the moderation of their laws. The neo-Harringtonians despised modern politeness, believing it to entail the corrosion of masculinity, the growth of forms of corruption accompanying the rise of parties and the specialist politician, and far greater uncertainty in civil society and in politics, exemplified by the reliance of the state upon the expertise of the stockjobber.

The clash between 'ancient' and 'modern' ideas about liberty has been characterized by Pocock as a debate in which the ancients saw liberty as a "direct act of the personality…the liberty of the hedgehog who knows himself but may know nothing else" and in which the moderns supposed liberty to entail the mediation of the personality, "through all the multifarious activities which may relate humans to one another in society…the liberty of the fox, who knows so many things that he may have no self left to know." In the Atlantic world, the controversy was characterized by concerns about the likely stability of real and of moveable property, and was in turn complicated both by the domestically uncertain identity of composite states comprised of different 'nations,' and by discussions in Protestant and Catholic states of the likely consequences of commercial society for religious belief, some of which questioned the compatibility of different forms of Christian worship with a polity competing for survival by commerce and by war. This state of affairs in turn generated an influential literature of jeremiad, which was particularly marked in the writings of historians, politicians, and philosophers in the late eighteenth century. Pocock has always contended that such issues remain identifiable in the political cultures of modern states, and particularly the North American Republic, and this has been one of the reasons for the controversial reception of his work within and beyond the academy.

In the *Foundations*, Skinner narrated a distinctive story, which challenged Pocock's account by emphasizing the differences between the Greek and the Roman political heritage, arguing specifically that Pocock had overestimated the importance of Aristotle's *Politics* as a source of Renaissance understanding of what it meant for citizens to live as equally sovereign agents in a law-governed republic. For Skinner, civic humanism was more heavily indebted to Cicero, who had extolled the life of civic virtue and republican citizenship that was possible under a free government. From Pocock's perspective, the fundamental difference between his view of civic humanism and Skinner's lies in the importance of justice as a political value. While Cicero encouraged the enjoyment of all the goods of human society and their just distribution, entailing the development of a moral code capable of defining and sustaining the life fully lived, Machiavelli was less concerned with justice and morals than with *virtù*, and placed personal autonomy and public participation at the center of his political philosophy. For Pocock, historical Ciceronianism was more a moral philosophy than a doctrine for political action, and it was Machiavelli's distinctive vision, which defined republican thought and the practice in Europe's small states, enjoying another history when it was adapted to the circumstances of Europe's larger monarchies and their colonial empires. The 'Afterword' to *The Machiavellian Moment* is revealing of Pocock's response to Skinner's criticism. It states that Skinner's approach should be placed within the history of jurisprudence, rather than being directly related to the republicanism of Machiavelli. Pocock has argued that Skinner's study of the emergence of the state should been seen as a contribution to 'the ideology of liberal empire,' founded on acceptance of the transformation of a republic into an empire, because the Ciceronian virtues Skinner described were capable of being realized under popular rule or the rule of a prince, such as

Augustus, Trajan, or Justinian, or indeed the 'sacred' monarch described by Eusebius, because of the existence of law. Pocock has criticized the notion that civic humanist ideas that paralleled the Florentine condemnation of tyrannical princes could be found in Otto of Freising's writings in the twelfth century. Pocock's argument is that none of Skinner's cast take seriously the idea of decline and fall, which became so powerful in European thought, questioning whether liberty depended upon law, and holding that virtues could easily be lost and states distempered and destroyed in consequence. In other words, Skinner has neglected to place texts in a context that is sufficiently broad. This criticism fits with Pocock's own description of the intellectual journey he has undertaken since *The Machiavellian Moment*. Pocock now states that his own notion of the politics employed in 1975 was insufficiently encompassing, especially of the theological ideas that played such a prominent role in seventeenth- and eighteenth-century life. A more theologically embedded perspective on politics has characterized his subsequent historical writing, and most directly that concerning Edward Gibbon's *Decline and Fall*.

The Origins of Modern Political Thought

Skinner's work since the *Foundations* has included major reinterpretation of Thomas Hobbes' perspective upon rhetoric and engagement with Aristotle, which for many exemplifies the Cambridge method. At the same time Skinner has retained an emphasis on the impersonal character of the modern state and an interest in the origins of modern ideas about liberty. It is in this area that his work has been most controversial. In *Liberty before Liberalism* (1998) and in a number of essays, Skinner has taken steps to refute the accusation of antiquarianism, by underscoring the continuing importance of a neglected tradition of political argument, evident after the regicide of 1649, when John Milton, Marchamont Nedham, and others were commissioned to defend the new state. Milton and Nedham were convinced that the actions of a political body had to be determined by the will of its members. They advocated representation rather than democracy, defining democracy as a system in which the people acted as their own government. Representation, by contrast, was one means of promoting government by the most virtuous and most wise, although Milton and Nedham recognized that it would also be necessary to have laws that encouraged *virtù* in the people, raising the possibility of laws to force citizens to be free. Arguing against the exercise of prerogative power by governments, Milton and Nedham held that the existence of such executive powers rendered subjects or citizens unfree, because of the threat to life and to property inherent in the existence of such powers. As such, they countered Hobbes's definition of a free man as someone 'not hindered to do what he has a will to' by praising civil liberty and political action with reference to Sallust, Seneca, and Tacitus. In consequence, Skinner has termed the doctrine of Milton and Nedham "the neo-Roman theory of free states and free citizens."

The essence of the neo-Roman theory was nicely captured in Harrington's response to Hobbes' famous claim that the liberty proclaimed on the turrets of the city of Lucca was illusory, because the extent of the citizens' freedom within the law was no greater than that enjoyed under the Sultan. Harrington stated that "even the greatest bashaw in Constantinople is merely a tenant of his head," enabling Skinner to conclude that an obsession with certain civil rights to be asserted against governmental power characterized by the thought of the neo-Romans, and constituted a significant classical legacy to modern politics. That someone is not free unless they are free from direct coercion, or indeed from the threat or likelihood of coercion, is a perspective that Skinner wants to reassert today, especially in cases where we may presume ourselves to be free, but where we are not because of the coercive potential of state action, or that of other corporate bodies.

Skinner's restated "foundations of modern political thought," with the lost tradition of Milton and Nedham added, have been criticized on numerous grounds. The concern with a fixed 'concept' of liberty defining a developing historical school of thought identical with Skinner's personal political philosophy, and deemed to be of direct relevance to modern problems in politics, has been called whiggish, and has been likened to the approaches founded on prolepsis that Skinner attacked in 'Meaning and Understanding.' The fact that none of Skinner's cast called themselves neo-Romans and advocates of a particular approach to liberty, or identified themselves as a separate group with the other authors listed as neo-Romans or opponents of slavery, has led to the accusation that Skinner's categories are ahistorical, or at best misleading, because they were not used by the historical actors concerned. Despite being the person who has done the most in recent years to bring philosophers and intellectual historians into dialogue, Skinner has been accused of being too much of a historian and too much of a philosopher by members of each tribe. His definition of politics as comprising precise but narrow concepts, explicable by means of a critique of Isaiah Berlin's discussion of liberty in the mid-twentieth century, has also been described as too blunt an instrument of early modern historical analysis, in part because of Pocock's success, and that of Jonathan Clark, Colin Kidd and others, in studying early modern politics together with theology and political economy. Skinner's claim that the neo-Romans' idea of 'civil liberty' was 'strictly political' is odd in that it implies that the neo-Romans were uninterested in the nature of the Christian polity, and in the forms of international competition considered vital to maintain security.

Some of these criticisms can be seen to have missed their target once it is accepted that Skinner's main aim has been to make sense of a complicated controversy over the meaning of liberty in seventeenth-century England, while valuably reminding his audience that liberalism is itself a historical construct. If the rise of the impersonal state remains central to Skinner's interpretation of modern political thought, a genuine question is the nature of the neo-Roman response, from the late seventeenth century, to the great change in attitudes that accompanied the rise of commercial monarchies that were able to use public credit to fight wars. This difficulty was identified sometime ago by John Dunn, Skinner's sometime Cambridge colleague, who has argued that Skinner's definition of the political is too narrow, and largely irrelevant to the modern world because it does not take seriously the economic limits to modern politics, in the form of debts and their consequences for national security, which have dominated political discussion since the early eighteenth century. In Dunn's view, the more accurate foundations of modern political thought have been delineated by Istvan

Hont in his *Jealousy of Trade* (2006), describing the battle about the relationship between the political control of the economy and the economic control of politics since David Hume sought to refute Bernard Mandeville's view of commercial society. Hont's work has inspired an international turn in intellectual historical research, which is leading to the reinterpretation of political thought from the seventeenth century to the present.

What all of this means is that while the term 'Cambridge School' is increasingly seen to be a misnomer, because those associated with it disagree about almost every aspect of the history of political thought, the Cambridge historians have at the same time produced a remarkable body of work that has contributed to the success of intellectual history in disciplines across the humanities. Facile assumptions of intellectual progress have been abandoned, and the use of modern categories in charting alien intellectual landscapes is increasingly questioned.

See also: Classical Legacies; Cultural History; Historical Thought and Historiography: Classical Period (Especially Greece and Rome); Historical Thought and Historiography: Current Trends; Historicism; History and Literature; Intellectual History; Isms and History; Linguistic Turn and Discourse Analysis in History.

Bibliography

Armitage, D., 2000. The Ideological Origins of the British Empire. Cambridge University Press, Cambridge.
Armitage, D., 2013. Foundations of Modern International Thought. Cambridge University Press, Cambridge.
Armitage, D., Subrahmanyam, S., 2010. The Age of Revolutions in Global Context, C. 1760–1840. Palgrave Macmillan, Houndmills, Basingstoke, Hampshire.
Bevir, M., 2001. The Logic of the History of Ideas. Cambridge University Press, Cambridge.
Clark, J.C.D., 2000. English Society, 1688–1832: Ideology, Social Structure and Political Practice during the Ancien Regime. Cambridge University Press, Cambridge.
Coleman, J., 2000. The history of political thought in a modern university. History of Political Thought 21, 152–171.
Collini, S., 2000. General introduction. In: Collini, S., Whatmore, R., Young, B. (Eds.), Economy, Polity, Society: British Intellectual History, 1750–1950. Cambridge University Press, Cambridge.
Collini, S., 2001. Postscript. Disciplines, canons and publics: the history of 'the history of political thought in comparative perspective'. In: Castiglione, D., Hampsher-Monk, I. (Eds.), The History of Political Thought in National Context. Cambridge University Press, Cambridge, pp. 280–302.
Condren, C., 1985. The Status and Appraisal of Classic Texts. Princeton, NJ.
Dunn, J. The identity of the Bourgeois Liberal Republic. In: Fontana, B. (Ed.), The Invention of the Modern Republic. Cambridge University Press, Cambridge, 209–210.
Dunn, J., 1968. The identity of the history of ideas. Philosophy 43, 85–104.
Gelderen, M., Skinner, Q., 2002. Republicanism: A Shared European Heritage. Cambridge University Press, Cambridge, UK.
Gunnell, J., 1998. Time and interpretation: understanding concepts and conceptual change. History of Political Thought 19, 641–658.
Gunnell, J., 1979. Political Theory: Tradition and Interpretation. Harvard University Press, Cambridge, MA.
Haakonssen, K., 2006. The history of eighteenth-century philosophy: history or philosophy? In: Haakonssen, K. (Ed.), The Cambridge History of Eighteenth-Century Philosophy. Cambridge University Press, Cambridge.
Hampsher-Monk, I., 1984. Political languages in time: the work of J. G. A. Pocock. The British Journal of Political Science 14, 159–174.
Hont, I., 2006. The Jealousy of Trade: International Competition and the Nation State in Historical Perspective. Harvard University Press, Cambridge, MA.
Janssen, P.L., 1985. Political thought as traditionary action: the critical response to Skinner and Pocock. History and Theory 24, 115–146.
Kidd, C., 1999. British Identities before Nationalism: Ethnicity and Nationhood in the Atlantic World, 1600–1800. Cambridge University Press, Cambridge, England.
Kidd, C., 2006. The Forging of Races: Race and Scripture in the Protestant Atlantic World, 1600–2000. Cambridge University Press, Cambridge, UK.
Palonen, K., 2003. Quentin Skinner: History, Politics, Rhetoric. Cambridge University Press, Cambridge.
Pocock, J., 1962. The history of political thought: a methodological enquiry. In: Laslett, P., Runciman, W.G. (Eds.), Philosophy, Politics, and Society, 2nd ser. Oxford University Press, Oxford, pp. 183–202.
Pocock, J., 1975. The Machiavellian Moment: Florentine Political Thought and the Atlantic Republican Tradition. Princeton University Press, Princeton, NJ.
Pocock, J., 1981. The Machiavellian moment revisited: a study in history and ideology. Journal of Modern History 53, 49–72.
Pocock, J., 1987. Between Gog and Magog: the republican thesis and the *Ideologia Americana*. Journal of the History of Ideas 48, 325–346.
Pocock, J., 1994. Machiavelli and the rethinking of history. Il Pensiero Politico 27, 215–230.
Pocock, J., 1996. Political thought in the English-speaking Atlantic, 1760–1790. In: Pocock, J.G.A., Schochet, G., Schwoerer, L.G. (Eds.), Varieties of British Political Thought, 1500–1800. Cambridge University Press, Cambridge, pp. 246–320.
Pocock, J., 2003. Barbarism and Religion. In: The First Decline and Fall Cambridge, vol. III. Cambridge University Press, Cambridge.
Pocock, J., 2004. Quentin skinner. The history of politics and the politics of history. Common Knowledge 10, 532–550, 537–538.
Pocock, J.G.A., 1971. Politics, Language, and Time: Essays on Political Thought and History. Atheneum, New York.
Pocock, J.G.A., 1985. Virtue, Commerce, and History: Essays on Political Thought and History, Chiefly in the Eighteenth Century. Cambridge University Press, Cambridge, Cambridgeshire.
Pocock, J.G.A., 1987. The Ancient Constitution and the Feudal Law: A Study of English Historical Thought in the 17th Century. Cambridge University Press, Cambridge [u.a.].
Pocock, J.G.A., 1999a. Barbarism and Religion, vol. 2. Cambridge University Press, Cambridge.
Pocock, J.G.A., 1999b. Barbarism and Religion, vol. I. Cambridge University Press, Cambridge, UK.
Pocock, J.G.A., 2000. Barbarism and Religion, vol. 2. Cambridge University Press, Cambridge [u.a.].
Pocock, J.G.A., 2009. Political Thought and History: Essays on Theory and Method. Cambridge University Press, Cambridge, UK.
Pocock, J.G.A., Schochet, G.J., Schwoerer, L.G., Folger Institute, 1993. The Varieties of British Political Thought, 1500–1800. Cambridge University Press, Cambridge, England.
Richter, M., 1990. Reconstructing the history of political languages: Pocock, Skinner and the Geschichtliche Grundbegriffe. History and Theory 29, 38–70.
Robertson, J., 1995. A Union for Empire: Political Thought and the British Union of 1707. Cambridge University Press, Cambridge, England.
Runciman, D., 2001. The history of political thought: the state of the discipline. British Journal of Politics and International Relations 3, 84–104.
Skinner, Q., 1969. Meaning and understanding in the history of ideas. History and Theory 8, 3–53.
Skinner, Q., 1981. Machiavelli. Oxford University Press, Oxford.
Skinner, Q., 1998. Liberty before Liberalism. Cambridge University Press, Cambridge.
Skinner, Q., 2002a. Visions of Politics, vol. 1. Cambridge University Press, Cambridge.
Skinner, Q., 2002b. Visions of Politics, vol. 3. Cambridge University Press, Cambridge.
Skinner, Q., 2004. Visions of Politics, vol. 2. Cambridge University Press, Cambridge.
Skinner, Q., 2008. Hobbes and Republican Liberty. Cambridge University Press, Cambridge, UK.
Skinner, Q., 1983. Machiavelli on the maintenance of liberty. Politics 18, 3–15.
Skinner, Q., 1984. The idea of negative liberty: philosophical and historical perspectives. In: Rorty, R., Schneewind, J.B., Skinner, Q. (Eds.), Philosophy in History. Cambridge University Press, Cambridge, pp. 193–221.
Skinner, Q., 2002c. A third concept of liberty. Proceedings of the British Academy 117, 237–268.
Skinner, Q., 2003. States and the freedom of citizens. In: Skinner, Q., Strath, B. (Eds.), States and Citizens. History, Theory, Prospects. Cambridge University Press, Cambridge, pp. 11–27.
Skinner, Q., Gelderen, M., 2013a. Freedom and the Construction of Europe. Cambridge University Press, Cambridge.
Skinner, Q., Gelderen, M., 2013b. Freedom and the Construction of Europe, vol. II. Cambridge University Press, Cambridge.
Whatmore, R., 2012. Against War and Empire: Geneva, Britain and France in the Eighteenth Century. Yale University Press, New Haven, CT.

Campaigning: Political

L Sandy Maisel, Colby College, Waterville, ME, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by L.S. Maisel, volume 3, pp. 1433–1435, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Political campaigns can be viewed at a micro- or a macro-level. At the micro-level, campaigning involves an individual candidate attempting to win political office. At the macro-level, political parties seek to gain seats in government for their partisans and ultimately to gain power or at least to share power in the relevant governmental arena. At each level the same steps in the process must be followed – decisions must be made about candidacies, organizations must be built, messages must be refined, strategies must be set, and the public must be convinced to support those waging a particular campaign. Only certain aspects of political campaigns are viewed by the public. Most frequently, the actions taken away from public view are more determinative of success or failure.

At the microlevel, political campaigning involves an individual candidate attempting to win political office. At the macrolevel, political parties seek to gain seats in government for their partisans and ultimately to gain power or at least to share power in the relevant governmental arena. At each level decisions must be made about candidacies, organizations must be built, messages must be refined, strategies must be set, and the public must be convinced to support those waging a particular campaign. Only certain aspects of political campaigns are viewed by the public. Frequently, the actions taken away from public view are more determinative of success and failure.

The Choice of Candidates

Decisions about who will run for office involve decisions by potential candidates as to whether or not they will run and decisions by political parties, acting under a variety of rules, concerning who their candidates will be.

The Decision to Run

Individuals who run for political office must exhibit political ambition beyond the desire to simply serve their community. Campaigning for and serving in political office involves personal costs and public exposure that deter many from seeking office. However, even those ambitious for elective office must make decisions about what offices to seek and when to seek those offices (Jacobson and Kernell, 1981). These decisions are based on a complex, often-implicit cost-benefit analysis. Potential candidates weigh the costs and benefits of both winning and losing. Their decision calculus can be summarized:

$$p_r = p_w(B_w - C_w) + p_l(B_l - C_l)$$

where p_r = probability of running; p_w = probability of winning; p_l = probability of losing; B_w = benefit of winning; C_w = cost of winning; B_l = benefit of losing; C_l = cost of losing.

The difficulty that potential candidates face in making these decisions relates to the uncertainty regarding virtually all of the relevant variables (Maisel, 1986; Kazee, 1994). Generally the rather obvious benefits far outweigh the costs of winning for those with the requisite political ambition; but costs must be weighed – loss of privacy, loss of time with family, frequently loss of income. Similarly, the evident costs of losing are almost always greater than the benefits; but some benefits do accrue to losers – making a name for oneself, building up good will with party officials. Beyond these broad generalizations, little is known about how specific candidates make individual decisions given the dazzling array of difficult-to-estimate variables (Maisel and Stone, 1997).

Parties' Choices of Candidates

The ways in which political parties choose their candidates for office varies significantly from political system to political system and from party to party. The key variable in this aspect of the process relates to the extent to which party organization controls places on the ballot contrasted with voters acting independently. In the former, the process is often described as candidate recruitment, in the latter, candidate emergence.

At one extreme are those systems in which political parties slate their candidates for office, ranking candidates or placing them in safe constituencies. Most parties operating in electoral systems with proportional representation follow this procedure. It is also followed in systems with strong party systems, in which citizens vote more for parties than they do for individual candidates. Many European systems fit this description (Hix and Lord, 1997).

At the other extreme are systems, of which the United States is prototypical, in which voters choose party candidates in primary elections. The shape of the electorate in primary elections varies considerably from polity to polity. In theory, primary elections are open only to members of one political party or another, but in actual practice how membership is defined varies according to local rules and can be quite restrictive or totally open. In either case, voters in the primary elections determine the party nominee, even if that nominee would not have been the choice of party officials (Maisel, 1999).

Many systems are hybrids, with party officials and the voters sharing power. The US presidential nominating system stands

as perhaps the most complex means of choosing party candidates. Delegates elected to national party conventions make the actual nominations. Some delegates are selected because of their positions within the party or as elected officials with party affiliations; others are chosen at party meetings (or caucuses) held in local communities; still others are chosen by the voters in primary elections, based on lists of delegate candidates approved by the presidential campaigns they seek to support (Polsby and Wildavsky, 2000; Wayne, 2000).

The Campaign Organization

In polities with strong party systems, campaign organizations tend to be party-centered. In these cases, the voters cast their ballots for candidates as representatives of a party. The party defines the campaign message and communicates it to the voters. The party raises the money to fund the campaign, does the polling necessary to understand what the electorate is thinking, and structures the campaign strategy and tactics. Typically full-time, year-round party workers perform all these functions or they are contracted out to professionals.

Where strong party systems do not exist, individual candidates must build their own organizations. In these cases, candidates are most concerned with their own elections, not that of those who share this party label. Candidates choose campaign managers whose sole goal is victory for that one candidate. They build personal organizations to handle the important functions of campaigning – doing research to develop issue positions and to counter opponents' proposals; polling the public to ascertain perceptions of the candidate and reactions to strategies; writing speeches; communicating with the press; developing and placing advertisements; sending out direct mail appeals, telephoning and leafleting the district; scheduling and preparing for candidate appearances; and perhaps most important, raising money. For larger and more expensive campaigns, all of these functions are performed by paid workers or volunteers in a candidate's own organization, or bought with campaign contributions.

Recent campaign organizations have been less able to handle all of these functions, and two means of coping have emerged. For campaigns in smaller constituencies, the candidate often performs all of these functions him or herself. They are done more or less well depending on the skill of the candidate, the time available, the level of competition, and the sophistication that is expected in campaigning for office (Maisel, 1999).

Even those with relatively high budget campaigns often find that they cannot hire staff to perform all of the necessary campaign activities. These candidates frequently turn to paid political consultants, experts from outside the campaign who specialize in certain aspects of the campaign process and sell their services to many campaigns during any election cycle. Firms specialize in media advertising, polling, direct mail, telephone banks, fundraising, and virtually every other aspect of campaigning; other firms are structured to take over all or many aspects of a campaign. Some work in particular geographic areas. Others are ready to sell their services anywhere throughout an entire nation or even internationally.

Defining Messages and Setting Strategy

A political campaign requires getting a message to voters. To do that, a candidate must define what the message is and must devise a strategy to reach the voters who will be swayed by that message. Again, how this is to be done varies with whether the campaign in question is party-centered or candidate-centered.

For party-centered campaigns, the message is the party platform, and the voters to be reached at those in the party's core constituency and others to whom the message might appeal in a particular year. Neither the message nor the strategy vary much from year to year, though marginal changes are made as the context changes – and these marginal changes might well spell the difference between victory and defeat.

For candidate-centered campaigns these decisions are among the most crucial. Candidates must decide whether to stick with the party message or to devise one of their own; they must decide what particular issues will play best to their voters; and they must figure out who precisely those voters are. Campaign strategists divide the electorate in a variety of ways – by demographic characteristics, by economic interests, by geographic locations. The goal is to use research and polling data to determine what appeals will work with which audiences and to have the candidate address those audiences appropriately. For incumbents seeking re-election, the strategic consideration often involves relying upon strengths demonstrated earlier. For challengers, the strategy must be to find an opponent's weakness and to emphasize one's own strengths.

Effective campaign strategies remain more art than science. Professional political campaigners know what has worked in the past. But many strategists learn those lessons. The successful strategists are those who can see how to change the approach as the context changes, as is often said in the military, 'not to refight the last war but to plan for the next one.'

Communicating with the Voters

To be effective, a political campaign must be able to send the message it has devised to the audience it has targeted. Campaign messages are communicated in two ways, through free media and paid media.

Free Media Coverage

Political campaigns seek free media coverage in whatever form they can, whenever they can. Candidates want to be reported on in newspapers and newsmagazines; they want to grant interviews on radio; they want their events covered on television. Free media coverage is advantageous not only because there is no cost but also because the viewer sees news coverage of a candidate as giving that candidate a sense of legitimacy; the candidate is seen as part of the day's news, not as part of a paid advertisement. However, free media is uncontrolled exposure. Campaign strategists can try to structure what topics will be covered – and campaigns are exploring increasingly sophisticated means to do this – but too frequently the message delivered is very different from the message the candidate seeks to convey.

Paid Media Advertising

The opposites apply to paid advertising. Paid media – whether in the form of television, radio, or print advertisements, of Internet sites, or of direct mail – is costly and is often viewed by the prospective voter as the slanted message it in fact is. However, these paid media also have the distinct advantage of allowing a candidate to emphasize exactly the message desired. These messages can be narrowcast, that is designed for and directed to specific target audiences. These ads can create images, discuss positions, or – increasingly in recent years – attack an opponent, all in precisely the way strategists feel will be most effective for a specified audience.

The critical concerns about political campaigns today relate to paid media. Many fear that effective political messages cannot be conveyed in 30-second advertisements, the medium preferred by advertising executives. Others complain that the emphasis on negative, attack advertising has poisoned the atmosphere that surrounds political life, keeping some of the best candidates from running (Ansolabehere and Iyengar, 1995). Virtually everyone bemoans the exorbitant costs of campaigns, costs spurred by reliance on paid media. However, political campaigns will continue to rely on paid media so long as they are effective in communicating a candidate's message to the target audience; for that is the means to electoral success – the ultimate message of a campaign's effectiveness.

Editor's Note

Recent research on political campaigning has focused on two major themes: (1) uses of social media and Internet communicaitons in political campaigns; and (2) comparative international studies of campaigns, strategies and effects. One interesting and recent study combining these two interests is Lilleker and Jackson, 2011. Also of interest: Vergeer et al., 2013; Vergeer, 2012.

See also: Advertising: Effects; Electoral Systems; Mass Media, Political Economy of; Media Effects; Media and Social Movements; Party Identification and Nonpartisanship; Political Money and Party Finance; Political Science and Polling; Voting, Sociology of.

Bibliography

Ansolabehere, S.A., Iyengar, S., 1995. Going Negative: How Attack Ads Shrink and Polarize the Electorate. Free Press, New York.

Hix, S., Lord, C., 1997. Political Parties in the European Union. St. Martin's Press, New York.

Jacobson, G.C., Kernell, S., 1981. Strategy and Choice in Congressional Elections. Yale University Press, New Haven, CT.

Kazee, T.A. (Ed.), 1994. Who Runs for Congress? Ambition, Context, and Candidate Emergence. Congressional Quarterly, Washington, DC.

Lilleker, Darren G., Jackson, Nigel A., 2011. Political Campaigning, Elections and the Internet: Comparing the US, UK, France and Germany. Routledge, Oxford.

Maisel, L.S., Stone, W.J., 1997. Determinants of candidate emergence in U.S. House elections: an exploratory study. Legislative Studies Quarterly 22, 79–96. Current work from the Candidate Emergence Project can be found at http://socsci.colorado.cdu/CES/home.html.

Maisel, L.S., 1986. From Obscurity to Oblivion: Running in the Congressional Primary, revised ed. University of Tennessee, Knoxville, TN.

Maisel, L.S., 1999. Parties and Elections in America: The Electoral Process, third ed. Rowman & Littlefield, Lanham, MD.

Polsby, N.W., Wildavsky, A., 2000. Presidential Elections: Strategies and Structures of American Politics, tenth ed. Chatham House Publishers, New York.

Vergeer, Maurice, Hermans, Liesbeth, Sams, Steven, 2013. Online social networks and micro-blogging in politicla campaigning: the exploration of a new campaign tool and a new campaign style. Party Politics 19 (3), 477–501.

Vergeer, Maurice, 30 September 2012. Politics, elections and online campaigning: past, present ... and a peek into the future. New Media and Society.

Wayne, S.J., 2000. The Road to the White House 2000. St. Martin's Press, New York.

Campbell, Donald Thomas (1916–96)

Andrew C Gordon, University of Washington, Seattle, WA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 1435–1439, © 2001, Elsevier Ltd.

Abstract

Donald T. Campbell (1916–96), American social scientist. Campbell received his B.A. and his Ph.D. at the University of California (Berkeley). Though he taught initially at the University of Chicago and later at several other institutions (Syracuse, LeHigh), Campbell spent the vast majority of his scholarly years (1953–79) in the Psychology Department at Northwestern University with which Campbell is 'permanently and inextricably co-identified.' Social psychology was his home base, but Campbell also influenced a vast array of other disciplines (philosophy, anthropology, sociology, statistics, education, and political science among them), scholars, and students through his seminal methodological and theoretical writings. He also insisted on exploring the connections among and interstices between disciplines as legitimate and often ignored areas for scholarship and debate. His emphasis was always on the acquisition of knowledge, and particularly on the fallibility of any single method or observation as a representation of reality. From those basic points, Campbell developed an enormous volume of rich concepts ('triangulation,' 'convergent vs discriminant validity,' 'plausible rival hypotheses,' 'evolutionary epistemology,' 'heterogeneity of irrelevancies') and practical strategies ('quasi-experiments,' 'unobtrusive measures') which have influenced dramatically both individual scholars as well as the current direction of the social sciences. Working simultaneously on basic epistemological and practical measurement and political issues confronting many disciplines, Campbell's contributions were fundamental and multirfaceted. With a twinkle in his eye and a gift for precise but colorful language, Campbell collaborated generously with generations of students and colleagues to produce some of the most oft-cited and influential work in twentieth-century social science.

Donald Thomas Campbell, born 20 November 1916 in Grass Lake, Michigan, died 6 May 1996 in Bethlehem, Pennsylvania, thus ending a career marked by an array of superlatives. Prolific scholar and author, original thinker, ebullient teacher with a notable twinkle in his eye, and generous colleague, Campbell was widely regarded as the most important social science methodologist of the twentieth Century. Several of Campbell's articles vie with one another as among the most widely cited pieces of social science scholarship. Repeatedly he coined phrases any one of which most scholars would be proud to have conceived, for example, 'quasi-experiments,' 'unobtrusive measures,' 'internal and external validity,' and 'plausible rival hypotheses.' His concepts are incorporated as fundamental in several fields – psychology, sociology, anthropology, organization and management sciences, public policy, evaluation, education, and philosophy – and in common use by scholars unaware of their inventor. Moreover, Campbell's welcoming and self-critical personal style, modeled on his scholarship, led him to embrace and examine all 'plausible rival hypotheses' and endeared him to students and colleagues alike.

Campbell received his B.A. (1939) and Ph.D. (1947), both in Psychology, from the University of California (Berkeley), and taught at several institutions, including Ohio State University (1947–50), the University of Chicago (1950–53), Syracuse University (as N.Y. State Board of Regents Albert Schweitzer Professor, 1979–82) and Lehigh University (as Distinguished University Professor of Sociology-Anthropology, Psychology and Education, 1983–96), but the majority of his scholarly years were spent as professor of psychology at Northwestern University (1953–79) with which Campbell's name and reputation are 'permanently and inextricably co-identified' (Campbell, 1981).

Campbell's honors include a Fulbright Visiting Professorship in Social Psychology at Oxford (1968), the Distinguished Scientific Contribution Award from the American Psychological Association (1969), election to the National Academy of Sciences and the American Academy of Arts & Sciences (1973), the Kurt Lewin Memorial Award from the Society for the Psychological Study of Social Issues (1974), presidency of the American Psychological Association (1975), the Williams James Lectureship at Harvard (1977), the Myrdal Prize in Science (1977), and the Distinguished Contribution Award from the American Educational Research Association (1981). In addition, two annual awards are given in his name: The Donald Campbell Award for Significant Research in Social Psychology (1982–) and the Donald Campbell Award from the Policy Studies Organization, to an 'outstanding methodological innovator in public policy studies' (1983–). Campbell received honorary degrees from numerous universities, including Oslo, Michigan, Chicago, and Southern California.

But Campbell felt most honored by the vast array of books dedicated to him, a list that grew steadily in several disciplines before and after his death.

Major Contributions

When he died, Campbell's resumé listed more than 230 published books, monographs, and articles. Any brief discussion of his contributions must necessarily focus on a few. Best known as a methodologist, with considerable scholarly work in the areas of experimental design, measurement, and social experimentation, Campbell is perhaps remembered especially for his explorations of the concept of validity. He was also a well-regarded epistemologist with a keen interest in the sociology of science. To each domain, Campbell brought his unique mind to bear on the problem of knowledge production. In a remarkable array of works, Campbell explored from several

perspectives the inevitable fallibilities inherent in both observers and methods in accurately portraying the world.

In his earliest publication, 'The Indirect Assessment of Social Attitudes' (1950), Campbell evidenced an aspect of the interests that would become the overall focus of his scholarly life: the imperfections introduced by human observers and methods, including the scientific method, in the search for veridical description of the physical world and mechanisms for compensating for these errors. (Campbell later declared profound 'ambivalence' toward the use of disguised measures; see Kidder and Campbell, 1970.) Campbell explored the sources and loci of bias and developed and refined methods for illustrating and for minimizing these biases. For example in Jacobs and Campbell (1961) he demonstrated how consensus in the 'reality' of an arbitrarily invented but shared group norm could persist over generations of experimental subjects. In a series of crosscultural explorations Campbell and co-workers (e.g., Segall et al., 1966) explored sources of validity and invalidity in perceptions of 'in-groups' and 'out-groups,' and the differential susceptibility to perceptual illusions between European and non-European cultures.

Campbell and Fiske (1959) published 'Convergent and Discriminant Validation by the Multitrait, Multimethod Matrix,' a complex and frequently cited analysis of the need for multiple measures of underlying constructs (traits) and of the need for the demonstrated capacity of methods to distinguish among traits if one is to minimize irrelevant measurement artifacts. Since every measure is partially invalid, multiple and distinctive methods are called for to yield a 'heterogeneity of irrelevancies.' Campbell and Fiske explored the value of simultaneously deploying maximally different measures of an underlying trait (in pursuit of convergent validation) and of different traits measured the same way (in pursuit of discriminant validation). Only by demonstrating that an underlying trait can be measured in various distinct ways and that the chosen measurement strategies distinguish among measured traits, can one minimize the impact of misleading artifacts.

By the early 1960s Campbell had established an enviable reputation for the vigor and reach of his campaign for advancing the methods and theory associated with issues of validity. His reputation was further advanced with the publication with (Campbell and Stanley, 1963) of 'Experimental and Quasi-experimental Designs for Research,'a widely circulated work (elaborated as Cook and Campbell, 1979) in which they popularized the terms 'internal' and 'external' validity:

> Internal validity is the basic minimum without which any experiment is uninterpretable: did in fact the experimental treatment make a difference in this specific experimental instance? External validity asks the question of generalizability: To what populations, setting, treatment variables, and measurement variables can this effect be generalized? (Campbell and Stanley, 1963).

Another widely cited treatise, *Unobtrusive Measures: Nonreactive Research in the Social Sciences* (Webb et al., 1966) evolved during affable lunchtime conversations with colleagues from different departments. Over several years these colleagues sought to identify relatively unbiased, nonreactive ways of measuring behavior. A frequently cited example describes ways a Chicago museum tried to identify the popularity of its exhibits. 'Obtrusive measures' included questioning visitors leaving the museum; an 'unobtrusive measure' was the frequency with which museum staff had to replace worn tiles in front of exhibits. By that index, the exhibit of baby chicks hatching live and wet from quivering eggs was by far the most popular.

Campbell's participation in an interdisciplinary conference spawned his brilliant foray into the sociology of science entitled 'Ethnocentrism of Disciplines: A Fishscale Model of Omniscience' (1969a). On reading a paper 'in an area of high relevance' to his own work by esteemed fellow social psychologist William McGuire, Campbell realized and then acknowledged that he had not 'read or read-at even half of McGuire's citations, and was not at all aware of the existence of another sizeable proportion' (Campbell, 1969a). While many scholars might react with chagrin, Campbell instead reflected on how the 'highly arbitrary' organization of universities and academic fields skewed knowledge production:

> Thus anthropology is a hodgepodge of all novelties venturing into exotic lands – a hodgepodge of skin color, physical stature, agricultural practices, weapons, religious beliefs, kinship systems, history, archeology, and paleontology ... Thus psychology is a hodgepodge of sensitive subjective biography, or brain operations, or school achievement testing, of factor analysis, of Markov process mathematics, of schizophrenic families, of laboratory experiments on group structure in which persons are anonymous ... Thus economics is a hodgepodge of mathematics without data, of history of economic institutions without mathematics or theory, of an ideal model of psychological man... (Campbell, 1969a)

Campbell took a breathtaking leap when he conceived the 'myth of unidisciplinary competence,' demonstrating how social, administrative, structural, and political supports which accrue over time around arbitrary disciplines ultimately create incentives that reinforce these arrangements and punish others of equal epistemological worth. Campbell detailed the inevitable consequence–biased knowledge production – and offered a model of 'interdisciplinary narrowness' for mitigating this 'ethnocentrism of disciplines.'

By the late 1960s, following Vietnam War protests, campus unrest, and a national focus on urban inequities, Campbell began trying conscientiously to influence the design, implementation, and evaluation of social policies. In 'Reforms as Experiments' (1969b) Campbell introduced a range of real-world quasiexperiments (attempts to approach the standards of experimentation even where random assignment of subjects to conditions is unachievable) – each tested against standards of 'internal and external validity' – to the practical world of social reform. That article frequently is cited as the most important single work in the field now known as 'program evaluation.'

In a brilliant tongue-in-cheek section of this article Campbell advised 'trapped administrators whose political predicament will not allow the risk of failure' how to capitalize on 'threats to validity' to assure positive results (Campbell, 1969b). For example, were they to accept the occasional 'grateful testimonial' as if it were a representative outcome, to

reserve interventions for carefully selected subpopulations most likely to succeed, and to eliminate from their analyses all those who prematurely quit the program, 'successful' program outcomes could be virtually assured.

Campbell continued in various fora to insist that social scientists have both obligations and opportunities to test the relevance of their theories and methods in the service of public good. In this, Campbell's unswerving concern was with the biases affecting the generation of knowledge. For example, in his too-infrequently cited 'Assessing the Impact of Planned Social Change' (1975) Campbell again displayed remarkable methodological and sociological insight in generalizing the problem of data validity in the policy arena in the following way:

> The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures, and the more apt it will be to distort and corrupt the social processes it is intended to monitor (Campbell, 1975).

Corrupted indicators remain a persistent and only vaguely explored source of real-world invalidity (e.g., Cochran et al., 1980).

Another highly influential, widely circulated piece went unpublished for nearly 20 years: in 'Methods for the Experimenting Society' Campbell attempted specifically to design a strategy for a society willing to test innovations and their intended effects. He argues for a society that would

> vigorously try out possible solutions to recurrent problems and would make hard-headed, multidimensional evaluations of outcomes, and when the evaluation of one reform showed it to have been ineffective or harmful, would move on to try other alternatives (Campbell, in Overman, 1988).

Surveying the multitude of societal reforms underway at the time, many of which he endorsed (e.g., Alexander Dubcek in Czechoslovakia, Salvador Allende, in Chile), Campbell argued:

> There is no such (experimenting society, committed to reality testing, self-criticism, to avoiding self-deception) anywhere today. While all nations are engaged in trying out innovative reforms, none of them are yet organized to adequately evaluate the outcomes of these reforms (Overman, 1988).

After specifying requirements of such an 'experimenting society' Campbell attempted in various ways to determine conditions under which these results might be achievable. For example, after Allende was assassinated in Chile, Campbell invited Chilean social scientist Ricardo Zuñiga to Northwestern University specifically to reflect on the conditions that might allow vigorous, large-scale reforms to be tested.

Campbell made other major contributions in his application of evolutionary theory to the realm of ideas. Campbell, the eternal epistemologist, focused in particular on the joint operation of the 'blind-variation-and-selective-retention process fundamental to all inductive achievements, to all genuine increases in knowledge.' In his 'evolutionary epistemology,' which perhaps received its fullest flowering in his essay honoring Karl Popper (in Overman, 1988), Campbell endorsed the natural selection paradigm for explaining the production of knowledge. Through the selective winnowing of ideas, Campbell argued, what survive over time are those ideas that are most meritorious, particularly as each is subjected to 'critical realism.' In his William James lectures, Campbell (1977) summarized his perspective in this way:

> I am a fallibilist and antifoundationalist. No part of our system of knowledge is immune to correction. There are no firm building blocks, either as indubitable and therefore valid axioms, or in any set of posits that are unequivocal once made. Nor are there any unequivocal experiences or explicit operations on which to found a certainty of communication in lieu of a certainty of knowing.

> I am some kind of a realist, some kind of a critical, hypothetical, corrigible, scientific realist. But I am against direct realism, naive realism, and epistemological complacency (Campbell, in Overman, 1988).

Legacy

Campbell's legacy is vast and pronounced. It affects a wide range of social scientists as well as public policy scholars and practitioners, and is reflected in the work of hundreds of his students and more distant heirs. He introduced fundamentally new ways of thinking that are taken for granted today about how we perceive, discover, and assess accumulated knowledge. Many of his concepts are now fundamental throughout the social and policy sciences. Campbell is the most important intellectual figure in the history of evaluation research. The strategies he employed in 'Reforms as Experiments' for unearthing and illustrating threats to interpretations of policy implementation are an often-cited model for exploring complex data (see Tufte, 1983).

In remarkably vivid, fetching but precise language, Campbell repeatedly captured the central importance of fallibilities inherent in human observers and their methods as they accumulate knowledge about the world, and thus of the importance of multiple perspectives, the triangulation of methods, and an open process whereby ideas are tested against one another to promote the winnowing effects of competition. Thus, in the James lectures, Campbell stated that

> We are cousins to the amoeba, and have received no direct revelation not shared with it. How then, indeed, could we know for certain? ... Our only hope as competent knowers is that we be the heirs of a substantial set of well-winnowed propositions (Campbell, in Overman, 1988).

In the heady intellectual mix he created and sustained, Campbell was known for changing his mind when persuaded by carefully tested evidence, and for challenging others to do the same. For example, over the years Campbell was persuaded

by sociologist Howard S. Becker and others to appreciate the falsifiability of carefully wrought case studies, in spite of their failure to live up to his long-cherished methodological criteria (Campbell, in Overman, 1988). In return, he prompted qualitatively-oriented colleagues to persist in reconsidering the validity claims inherent in their own work (see, for example, Becker, 1986). For this open-mindedness, Campbell is revered by quantitative and qualitative methodologists alike (see Patton, 1999).

Campbell also influenced legions of colleagues to reflect honestly on their own work. He talked and wrote openly about his own insecurities and frustrations, including how late in his career he began to write, droughts in his own productivity, his discomfort with the production pressure and relentless expectation of genius when he was a young assistant professor at the University of Chicago, his frequent bouts of depression, and the times when he believed his students and colleagues 'carried him.'

He explained his messy desk cheerfully by saying that he 'filed archeologically,' and as long as no one moved things around he could find whatever he wanted. Amazed colleagues would watch him reach into a pile and pull out the precise papers relevant to their discussions.

Although its publishers have sold thousands of copies of *Experimental and Quasi-Experimental Designs for Research* and students return to it year after year, Campbell never received any royalties; he asked only for unlimited copies to give freely to his colleagues.

Consideration of Campbell's legacy would be incomplete without a discussion of his personal style, for it, too, promoted the production of knowledge while providing his colleagues with a vital and successful paradigm. It would understate Campbell's colossal influence if there were no mention of his towering warmth, kindness, and generosity, and his deeply honest and self-critical perspective on his own work (Campbell, 1981). Campbell was remarkably able to reflect on his own fallibilities and to create conditions that would model his epistemology, and improve both his work and the work of others. Examples are diverse and legendary: he would often send a requested reprint accompanied by a colleague's critical rejoinder. He was always in pursuit of contrary evidence and counter-intuitive examples, particularly if they confronted his own work. He encouraged students to write minority sections to project reports if they disagreed with him or with an emerging majority. Campbell was resolutely encouraging of others and in fact was thought by some to 'suffer fools gladly.' He gave three grades: A, B, and incomplete. If the data collected in dissertations failed to support anticipated hypotheses, or failed to replicate published findings in the literature, Campbell urged students to write up the apparent reasons for the obtained results.

Over time, Campbell connected his early deep philosophical explorations to a program of practical work that will continue to guide generations of scholars and practitioners. Campbell lives on through his voluminous work in many fields, through his students, and through their adoption of his style.

See also: Evolutionary Epistemology; Experimental Design: Large-Scale Social Experimentation; Experimentation in Psychology, History of; External Validity; Foucault, Michel (1926–84); Internal Validity; Panel Surveys: Uses and Applications; Pearson, Karl (1857–1936); Quasi-Experimental Designs; Unobtrusive Methods, History of.

Bibliography

Brewer, M., Collins, B., 1981. Scientific Inquiry and the Social Sciences. Jossey-Bass, San Francisco.

Campbell, D.T., 1950. The indirect assessment of social attitudes. Psychological Bulletin 47 (1), 15–38.

Campbell, D.T., 1969a. Ethnocentrism of disciplines and the fish-scale model of omniscience. In: Sherif, M., Sherif, C. (Eds.), Interdisciplinary Relationships in the Social Sciences. Aldine, Chicago, pp. 328–348.

Campbell, D.T., 1969b. Reforms as experiments. American Psychologist 24, 409–429.

Campbell, D.T., 1975. Assessing the impact of planned social change. In: Lyons, G. (Ed.), Social Research and Public Policies. Public Affairs Center. Dartmouth College, Hanover, NH, pp. 3–45.

Campbell, D.T., 1977. Descriptive Epistemology: Psychological, Sociological, and Evolutionary. William James Lectures. Harvard University, Cambridge, MA.

Campbell, D.T., 1981. Another Perspective on a Scholarly Career. In: Brewer, M.B., Collins, B.E. (Eds.), Scientific Inquiry and the Social Sciences. Jossey-Bass, San Francisco, pp. 454–486.

Campbell, D.T., Fiske, D.A., 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin 56, 81–105.

Campbell, D.T., Stanley, J.C., 1963. Experimental and quasi-experimental designs for research on teaching. In: Gage, N. (Ed.), Handbook of Research on Teaching. Rand McNally, Chicago, pp. 171–246.

Cochran, N., Gordon, A.C., Krause, M., 1980. Proactive records: Reflections on the village watchman. Knowledge 2 (1), 5–18.

Cook, T.D., Campbell, D.T., 1979. Quasi-experimentation Design and Analysis Issues for Field Settings. Rand McNally, Chicago.

Jacobs, R.C., Campbell, D.T., 1961. The Perpetuation of an arbitrary tradition through several generations of a laboratory microculture. Journal of Abnormal & Social Psychology 62, 649–658.

Kidder, L., Campbell, D.T., 1970. The indirect testing of social attitudes. In: Summers, G. (Ed.), Attitude Measurement. Rand McNally, Chicago.

Overman, E.S. (Ed.), 1988. Methodology and Epistemology for Social Science: Selected Papers of Donald T. Campbell, Chicago.

Patton, M.Q., 1997. Utilization–Focused Evaluation, third ed. Sage, Beverly Hills, CA.

Segall, M.H., Campbell, D.T., Herkovits, M.J., 1966. The Influence of Culture on Visual Perception. Bobbs-Merrill, Indianapolis, IN.

Tufte, E.R., 1983. The Visual Display of Quantitative Information. Graphics Press, Cheshire, CT.

Webb, E.J., Campbell, D.T., Schwartz, R.D., Sechrest, L., 1966. Unobtrusive Measures: Nonreactive Research in the Social Sciences. Rand McNally, Chicago.

Cancer: Psychosocial Aspects

Tanja Zimmermann, Institute of Psychology, Technische Universität Braunschweig, Braunschweig, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.I. Hasenbring, volume 3, pp. 1443–1449, © 2001, Elsevier Ltd.

Abstract

With advances in treatment efficacy and improvements in survival rates, cancer is increasingly viewed and treated as a chronic disease that can be effectively managed for many years. From the onset of symptoms until the completion of therapy and into survivorship, patients face different physical, psychological, and social challenges. In addition, not only patients but also their partners have to deal with the consequences of cancer. Significant advances have been made in the understanding of psychosocial adjustment to cancer. Routine screening and management of psychosocial distress with psychosocial interventions is recommended as an integral part of cancer care. Moreover, the interactions between patients and family members are of great importance to further the field by providing more insight into the family dynamics of dealing with cancer.

Psychological Distress

An aging and growing population is associated with more elderly patients being treated for cancer. Remarkable advances have been made in innovative medical technologies and therapies for cancer patients. Almost 60% of adult cancer patients are expected to survive at least 5 years from diagnosis or to be cured, with 2/3 of them over 65 years of age (Valdivieso et al., 2012). The number of people treated or followed up for cancer or who have survived the disease, therefore, is growing. Nevertheless, cancer remains a life-threatening diagnosis with profound psychosocial impact even years after medical treatment has ended. Cancer patients are known to endure psychosocial, emotional, and physical distress. Psychosocial distress describes an unpleasant psychological (behavioral, emotional, cognitive), social, and/or spiritual experience, which can negatively impact cancer patients' ability to cope effectively with cancer, its physical symptoms, and its treatment (Pirl et al., 2014). Moreover, prolonged exposure to stress can have a negative effect on quality of life (QoL) and results in a host of physical symptoms. In the light of such substantial and sustained disease burden, the management of survivorship issues in cancer assumes paramount importance. Such issues revolve around the effective maintenance of individual functioning (e.g., psychological distress, QoL) throughout the course of cancer. The research base evaluating the prevalence of psychological burden (such as depression and anxiety) is growing steadily, and there is now also a sizable body of clinically relevant research.

Psychological Distress in Cancer Patients

Psychological distress covers a wide spectrum, ranging from normal feelings of vulnerability, sadness, and fears to problems that can become disabling, such as depression, anxiety, extensive worries, negative thoughts, or social isolation. The prevalence of distress, depression, anxiety, and other psychiatric disorders has been studied for many types and stages of cancer. A recent review shows a prevalence of 38.2% of mood disorders in patients with various types of cancer (Mitchell et al., 2011). Significant distress ranges from 20% to 40% in newly diagnosed patients, as well as patients experiencing a recurrence (Holland et al., 2010). Distress is associated with higher rates of nonadherence to medical treatment, QoL, or even survival (Shimizu, 2013). In cancer care, patients' distress levels have been positioned as the sixth vital sign in addition to heart rate, respiration, blood pressure, temperature, and pain (Bultz and Carlson, 2006). Psychological distress may include physical and cognitive dysfunctions resulting from cancer treatment, anxiety and depression, disruptive social and dyadic relationships as well as fatigue and sleep disturbances. Patients experience significant disruptions to normal daily activities, which challenge a person's assumptions about their current and future life trajectory and sense of self. Patients have to cope with role changes and distress provoked by the increase in functional disability and physical side effects associated with cancer and its medical treatment. Some cancer patients blame themselves for the cancer diagnosis (e.g., head and neck cancer patients due to the fact that tobacco and alcohol represent the major risk factors for these cancers (Lang et al., 2013)). Moreover, long-term cancer survivors worry about cancer recurrence or development of new cancers, as well as physical, cognitive, socioeconomic, sexual, and legal issues (Valdivieso et al., 2012).

Depression and Anxiety

Distress in cancer patients can also precede clinical depression and anxiety, with prevalence rates ranging between 5% and 50% (Artherholt and Fann, 2012). A recent meta-analysis found mood symptoms in all patients (men and women) and at all stages of treatment (palliative and nonpalliative; Mitchell et al., 2011). In identifying depression, some symptoms such as low mood, loss of interest in usual activities, feelings of hopelessness or guilt, or suicidal ideations are particularly important to consider whereas physical symptoms such as appetite and weight loss, fatigue, and sleep disturbances may overlap with disease or treatment effects. Depression should carefully be distinguished from sadness or grief. Numerous patients that receive a cancer diagnosis assume that they are close to death and experience anxiety, depression, fear, and anger as the disease progresses.

Fear of Recurrence

A widespread problem among cancer patients in addition to psychosocial burden and physical symptoms is the fear of

recurrence (FoR). FoR is defined as the fear of the disease recurring or progressing in the same organ or a different area of the body. FoR is negatively associated with QoL and psychosocial well-being, inducing depression, and anxiety (Herschbach et al., 2005). Specific situations and/or symptoms (e.g., aftercare appointments, anniversaries of the diagnosis, or environmental triggers) can activate the intensity of FoR. Hence, estimates of prevalence rates depend on the time of assessment. A recent systematic review shows that even years after the initial cancer diagnosis, cancer survivors still experience FoR with modest intensity (Koch et al., 2013). This implies that FoR can persist over an extensive period of time after the initial cancer diagnosis in unchanged intensity. Moreover, FoR was also found in partners of cancer patients (Zimmermann et al., 2011).

Cognitive Dysfunction

Problems with memory, attention span, or concentration in cancer patients who received chemotherapy are known as chemotherapy-related cognitive dysfunction, also known as "chemobrain" or "chemofog." Between 4% and 75% of patients are affected by these problems, and for 20–35% of cancer survivors, it may persist for months or years after treatment has been completed (Hines et al., 2014). Although the causes for these cognitive deficits remain unknown, they can have a considerable negative impact on QoL and functional ability (e.g., work-related difficulties, social withdrawal for fear of deficits being noticed).

Sexual Impairment

Almost any cancer diagnosis can lead to sexual impairment, those directly affecting areas related to sexual function are particularly concerned (e.g., breast or genitals). Among cancer survivors, sexual dysfunctions are common, but often not addressed. Side effects of the medical treatment (e.g., chemotherapy, radiation, and hormone therapy) can contribute to lower self-esteem, depression, less desire for sexual activity, and lower libido. Sexual disturbances are frequent among cancer patients regardless of gender, and need attention by health care providers (Valdivieso et al., 2012). It is also important to recognize that psychosocial aspects as depression or anxiety can contribute to sexual dysfunction (specifically in terms of desire, arousal, and orgasm). In women with cancer (i.e., breast cancer) the iatrogenic menopause can cause significant detriment to sexual functioning (e.g., lowered libido, decreased vaginal lubrication, dyspareunia, and numbness in previously sensitive breasts; Morreale, 2011). Comparable results were found for the iatrogenic effects of hormonal ablation in men with prostate cancer, which include mood disturbance, cognitive impairment, fatigue, and sexual dysfunction (Chambers et al., 2011). Cancer patients with a negative view of their body after cancer treatment have been found to have lower rates of sexual satisfaction. For example, patients with colorectal cancer and newly formed enterostoma experience disturbances in body image and may face complications such as leakage, skin and stoma problems, and odor, which may lead to low self-esteem, poor self-concept, and sexual dysfunction (Hoon et al., 2013). In addition, the relationship between the cancer patient and the partner is an important factor for sexuality. For example, for breast cancer patients, the quality of the relationship emerged as a strong predictor of sexual functioning (Gilbert et al., 2010).

Body Image

Cancer patients often have to cope with significant changes to appearance and functioning. Body image is, therefore, a critical psychosocial issue for these patients. Cancer and its medical treatment can profoundly affect patients' body image during various stages of treatment and have strong implications for psychosocial adjustment and social-functioning. For example, patients with serious body image concerns experience negative emotions (i.e., distress), avoid social situations, and become isolated. Independent from cancer or its site, a negative body image adversely impacts QoL and psychosocial-functioning (e.g., anxiety and depression, and sexual functioning). Cancer and its medical treatment can lead to appearance changes (i.e., scarring, hair loss, disfiguration), sensory alterations (e.g., numbness, pain), and functional impairment (e.g., dysphagia, impotence, dysarthria). Problems with body image occur in patients with different cancer sites, which are most prevalent in the immediate postoperative and treatment period (Fingeret et al., 2014), and involve perceptions, thoughts, feelings, and behaviors related to the whole body and its functioning. Only few studies have investigated the prevalence rates of body image concerns in cancer patients; however, there is some evidence that patients with head and neck cancer show the highest prevalence rates of up to 75% (Fingeret et al., 2014), primarily due to surgical disfigurement and obvious changes in their appearance, or changes in how they speak and eat (Lang et al., 2013; Semple et al., 2013). In addition, patients with colorectal cancer experience body image disturbance after surgery (enterostoma), which may be associated with greater distress and poorer QoL. Nevertheless, further investigations are necessary to confirm these associations (Sales et al., 2014). Other studies have focused on the prevalence of body image concerns in women with breast cancer and found prevalence rates between 17% and 33%. Also, long-term breast cancer survivors' rates range between 15% and 30%. In summary, body image concerns affect a substantial number of patients with cancer, which may persist into long-term survivorship. Further research is necessary to identify demographic, disease, and treatment-related factors that increase the risk for body image concerns.

Positive Aspects

Despite the negative consequences of a cancer diagnosis, a growing literature also reports positive aspects of the cancer experience such as greater appreciation of life, changed priorities, closer relationships, and enhanced spirituality. In a study by Dorval et al. (2005), 42% of the 282 couples with women newly diagnosed with nonmetastatic breast cancer, reported getting closer together after the cancer experience. Cancer can be a growth experience for patients and couples (Tedeschi and Calhoun, 1995), in terms of (1) benefit finding (e.g., positive change in relationships, greater appreciation of life, and change in priorities), i.e., the acquisition of benefit from adversity, and (2) posttraumatic growth (e.g., increased sense of personal strength, changed priorities, and richer existential and spiritual life), i.e., success in coping with the aftermath of a trauma,

reconstructing or strengthening perceptions of oneself, others and the meaning of events (Tedeschi and Calhoun, 2004).

Psychological Distress in Partners of Cancer Patients and the Couple

There is growing awareness that cancer not only affects patients, but also the entire family, friends, and caregivers. Usually the patient's intimate partner is the primary caregiver. The diagnosis and treatment of cancer can affect both members of the dyad, and induce serious distress and negatively impact successful coping throughout the cancer trajectory (McClure et al., 2012). Social relationships are among the most-studied contributing factors to QoL in cancer patients (Stanton et al., 2007). Among these relationships, the patient's relationship with the spouse is considered paramount because cancer patients identify their partners as their most important source of emotional and instrumental support (Baanders and Heijmans, 2007). The diagnosis and treatment of cancer, however, can affect every aspect of both patients' and partners' psychological functioning. Partners and patients can show higher levels of distress compared to the general population (Hinnen et al., 2008; Nakaya et al., 2010). For a long time, cancer was considered an individual stress experience demanding individual coping strategies. Since both partners of the dyad interact with and mutually influence each other, it is important to consider the individual and dyadic functioning of the patient in relation to the partner's physical and mental states, and vice versa. The intercorrelation between the psychological distress of patients and spouses support the interdependence between patient and partner (Hagedoorn et al., 2008). Therefore cancer should be considered as a stressor concerning both partners simultaneously and should be studied as a couple's disease if a partner is present (Kayser et al., 2007).

Cancer not only impacts the patients' life in demanding restoring physical, psychological, and social homeostasis, but also threatens the spouse (Traa et al., 2014). Both the patients' and the partners' reactions to a cancer diagnosis include fear, uncertainty, anxiety, and depression with patients' partners' reactions mutually determining how they cope with and adjust to the illness. The partner has to adapt to the patients' disease, has worries, has to confront the potential loss of a life partner, is burdened by taking over the patients' tasks, and by providing support and caring. Moreover, spouses can experience important restrictions in their social life, and interruption to daily life; they might suffer from economic or financial consequences, have to manage household and child care responsibilities, and are threatened by an insecure future (Traa et al., 2014). Sleep disturbances, fatigue, loss of physical strength, pain, loss of appetite, and weight loss are the most prevalent physical problems in caregivers (Stenberg et al., 2010). Prevalence of depression and anxiety among partners of cancer patients varies between 10–53% and 16–56%, respectively (Girgis et al., 2013). Several studies demonstrate disruptions in social involvement and activities of the partners due to their caregiving responsibilities (Girgis et al., 2013).

In addition, cancer-related stress contributes to general life stress and daily hassles. Partner experiences a 'double role': on the one hand they undergo their own psychosocial distress, and on the other hand they have to provide support during a time when they themselves are under extreme stress. Therefore, coping with cancer can challenge communication patterns, roles, and responsibilities of a couple. While the majority of partners dealing with cancer experience a variety of psychosocial burden, many caregivers report also positive emotions (e.g., satisfaction, happiness about quality time with the spouse, feelings of value and self-worth) from their caring role (Girgis et al., 2013). Not surprisingly, some couples report that the cancer experience brought them closer together (cohesion), whereas others experience adjustment and communication problems that may result in decreased intimacy and lower relationship satisfaction.

There is emerging evidence that dyadic processes, specifically communication and dyadic coping, may be involved in the adjustment process for both partners when facing a cancer diagnosis. Under the stressful circumstances of the diagnosis, couples may tend to be kind to each other (increase positivity in the relationship), but this may occur within the context of avoidance (e.g., not talking about the stressor, mutual withdrawal to give the other person time to adapt, and not upsetting the other person). Thus, despite an increase in positive interaction between partners at present, such avoidance patterns potentially come at the cost of not addressing the issues surrounding the illness, which places long-term adjustment to cancer and relationship satisfaction at risk (Manne et al., 2010). This phenomenon, known as 'protective buffering' (i.e., hiding concerns, denying worries, and avoiding discussions), is associated with higher levels of patient and partner distress. Investigating specific dyadic behaviors in couples coping with cancer may, therefore, be more fruitful than relating cancer adjustment to more general marital quality. Lafaye et al. (2014) conclude that coping strategies used by cancer patients or spouses not only influence their own well-being but also the well-being of the partner. Especially, dyadic coping as a stress management strategy entails that both partners are mutually involved in the stress coping process in providing and receiving support from each other and engaging in joint problem-solving activities, and shared emotion regulation to behave as a unit rather than as individuals (Bodenmann, 1995; Badr et al., 2010).

In summary, both members of a couple facing a cancer diagnosis experience psychosocial distress, which might also impact their individual and relationship functioning. Coping with cancer from a couple-based perspective as a dyadic situation demonstrates potentially beneficial effects on psychosocial adjustment and relationship functioning of patients and spouses (Badr and Krebs, 2013; Martire et al., 2010; Regan et al., 2012).

Psychosocial Interventions

The field of psycho-oncology has progressed over the last 40 years from mapping the extent and severity of psychological distress in cancer patients, through the development and evaluation of psychosocial intervention to the current phase of disseminating psychological skills to health care professionals (Moorey, 2013). Over the past few decades, a variety of psychosocial interventions has been developed to address the

psychological and social (i.e., psychosocial) aspects of disease for cancer patients. These interventions involve a wide range of components, such as psychoeducation, enhancing coping skills, and providing supportive care to cancer patients to address the psychosocial needs of cancer patients. Research has shown that treatment of distress may improve health-related QoL in cancer patients (Artherholt and Fann, 2012).

Interventions for Patients

While unmet psychosocial needs exist across cancer types and the cancer care continuum, most studies of psychosocial interventions have been conducted in patients with breast or prostate cancer, and only few studies included patients with late-stage cancer and cancer survivors (Subnis et al., 2014). A large number of patients continue to live with cancer following their surgical and other adjuvant treatments. Consequently, it is necessary to provide also interventions that focus on long-term benefits in patients' adjustment and coping (Tamagawa et al., 2012).

Longitudinal studies are, therefore, needed that include patients from the time of diagnosis until the end of survivorship in a variety of cancer populations, which will provide valuable information about psychosocial aspects and effects of psychosocial interventions across the cancer care continuum. In addition, distress is often underrecognized by medical staff in clinical oncology setting. Inadequate recognition and treatment for distress can interfere with medical decision making and treatment adherence. In 2015, cancer centers in the US will be required to implement screening programs for psychological distress as a new criterion for accreditation (Pirl et al., 2014). Screening is an optimal strategy for detecting problems such as distress in cancer patients. In addition, the National Comprehensive Cancer Network guidelines recommend regular screening for distress with brief screening instruments (e.g., the Distress Thermometer) because of the substantial consequences of disregarded and untreated distress in cancer patients (Holland et al., 2011).

Interventions for Couples

A cancer diagnoses poses many challenges for both patients and their partners. Several reviews have addressed the potential of couple-based interventions to improve individual and dyadic functioning of couples coping with cancer, and the beneficial effects on psychosocial adjustment and relationship satisfaction (Badr and Krebs, 2013; Martire et al., 2010; Regan et al., 2012). Therefore, both patients and their partners should be included in psychosocial interventions, with the aim to support and strengthen their ability to cope with cancer. To date, several couple-based intervention studies aimed at improving dyadic coping and relationship satisfaction of couples coping with cancer. A recent review illustrates the importance of stress communication (e.g., expressing worries, feelings, and needs regarding cancer- or relationship-related topics), supportive behavior, and effective dyadic coping for relationship satisfaction in couples coping with cancer (Traa et al., 2014). Moreover, interventions with multiple targets, e.g., reduction of emotional distress, improved communication and sexual intimacy, and strategies to enable partners to express their distress show the most promising results in enhancing couples' well-being (Wootten et al., 2014). In summary, couple-based interventions have yielded promising results for couples in which one partner has a cancer diagnosis in improving communication, dyadic coping, QoL, psychosocial distress, sexual functioning, and marital satisfaction (Li and Loke, 2014; Badr and Krebs, 2013).

Conclusions

The increase in survival rates of cancer patients suggests that aspects of illness adjustment and QoL are becoming increasingly important. Psychosocial aspects have been examined in patients with different cancer diagnoses and at varying disease stages. Significant impairments in physical-, role-, emotional-, cognitive-, and social-functioning negatively affect QoL and psychosocial adjustment in cancer patients. Cancer patients' psychological distress follows distinct trajectory patterns during the course of the disease. A proportion of patients maintain low psychological distress levels and adequate QoL during the entire duration of the illness, while others show a U-shaped pattern, i.e., initially high stress levels, followed by a gradual reduction as cancer progresses with an increase again toward the terminal stages of the disease. Maintaining and enhancing the well-being of cancer patients is of great importance in cancer care. Further research and improvements are necessary (1) in identifying patients' distress (e.g., with effective screening instruments), (2) in identifying partners' distress, (3) in improving psychosocial interventions (for patients, partners, couples, or families), (4) in offering cancer patients (and their partners) appropriate psychosocial therapies, which suit their psychosocial needs (indicated care), and (5) in integrating psychosocial therapies and services into the standard of quality cancer care. More research is needed to develop more effective methods of identification and targeted interventions for psychosocial aspects in the context of cancer treatment. Clinicians should be aware of the impact of psychosocial aspects on cancer patients, even years after the initial diagnosis, e.g., FoR, anxiety, depression, body image concerns, sexual dysfunction, etc. Therefore, the identification and management of major psychosocial aspects should be a clear clinical priority. Because of the potential of psychosocial aspects to reduce QoL and to induce psychosocial morbidity, even years after the initial diagnosis, the lack of research in this area and an increasing number of long-term cancer survivors, future studies in the field of psychosocial aspects of cancer should focus on these factors, and on effective ways to improve psychosocial adjustment and QoL for cancer patients and their partners.

See also: Behavioral Medicine; Body Image; Cancer Screening; Caregiver Burden; Childhood Cancer: Psychosocial Aspects; Co-morbidity of Mental and Physical Conditions; Gynecological Health: Psychosocial Aspects; Illness and Dyadic Coping; Personalized Medicine; Social Support and Recovery from Disease and Medical Procedures.

Bibliography

Artherholt, S.B., Fann, J.R., 2012. Psychosocial care in cancer. Current Psychiatry Reports 14, 23–29.

Baanders, A.N., Heijmans, M.J., 2007. The impact of chronic diseases: the partner's perspective. Family & Community Health 30, 305–317.

Badr, H., Carmack, C.L., Kashy, D.A., Cristofanilli, M., Revenson, T.A., 2010. Dyadic coping in metastatic breast cancer. Health Psychology 29, 169–180.

Badr, H., Krebs, P., 2013. A systematic review and meta-analysis of psychosocial interventions for couples coping with cancer. Psychooncology 22, 1688–1704.

Bodenmann, G., 1995. A systematic-transactional view of stress and coping in couples. Swiss Journal of Psychology 54, 34–49.

Bultz, B.D., Carlson, L.E., 2006. Emotional distress: the sixth vital sign–future directions in cancer care. Psychooncology 15, 93–95.

Chambers, S.K., Pinnock, C., Lepore, S.J., Hughes, S., O'connell, D.L., 2011. A systematic review of psychosocial interventions for men with prostate cancer and their partners. Patient Education and Counseling 85, e75–88.

Dorval, M., Guay, S., Mondor, M., Masse, B., Falardeau, M., Robidoux, A., Deschenes, L., Maunsell, E., 2005. Couples who get closer after breast cancer: frequency and predictors in a prospective investigation. Journal of Clinical Oncology 23, 3588–3596.

Fingeret, M.C., Teo, I., Epner, D.E., 2014. Managing body image difficulties of adult cancer patients: lessons from available research. Cancer 120, 633–641.

Gilbert, E., Ussher, J.M., Perz, J., 2010. Sexuality after breast cancer: a review. Maturitas 66, 397–407.

Girgis, A., Lambert, S., Johnson, C., Waller, A., Currow, D., 2013. Physical, psychosocial, relationship, and economic burden of caring for people with cancer: a review. Journal of Oncology Practice 9, 197–202.

Hagedoorn, M., Sanderman, R., Bolks, H.N., Tuinstra, J., Coyne, J.C., 2008. Distress in couples coping with cancer: a meta-analysis and critical review of role and gender effects. Psychological Bulletin 134, 1–30.

Herschbach, P., Berg, P., Dankert, A., Duran, G., Engst-Hastreiter, U., Waadt, S., Keller, M., Ukat, R., Henrich, G., 2005. Fear of progression in chronic diseases: psychometric properties of the Fear of Progression Questionnaire. Journal of Psychosomatic Research 58, 505–511.

Hines, S., Ramis, M.A., Pike, S., Chang, A.M., 2014. The effectiveness of psychosocial interventions for cognitive dysfunction in cancer patients who have received chemotherapy: a systematic review. Worldviews on Evidence-Based Nursing 11, 187–193.

Hinnen, C., Hagedoorn, M., Ranchor, A.V., Sanderman, R., 2008. Relationship satisfaction in women: a longitudinal case-control study about the role of breast cancer, personal assertiveness, and partners' relationship-focused coping. British Journal of Health Psychology 13, 737–754.

Holland, J.C., Andersen, B., Breitbart, W.S., et al., 2010. Distress management. Journal of National Comprehensive Cancer Network 8, 448–485.

Hoon, L.S., Chi Sally, C.W., Hong-Gu, H., 2013. Effect of psychosocial interventions on outcomes of patients with colorectal cancer: a review of the literature. European Journal of Oncology Nursing 17, 883–891.

Kayser, K., Watson, L.E., Andrade, J.T., 2007. Cancer as a "we-disease": examining the process of coping from a relational perspective. Families, Systems, & Health 25, 404–418.

Koch, L., Jansen, L., Brenner, H., Arndt, V., 2013. Fear of recurrence and disease progression in long-term (>/=5 years) cancer survivors–a systematic review of quantitative studies. Psychooncology 22, 1–11.

Lafaye, A., Petit, S., Richaud, P., Houede, N., Baguet, F., Cousson-Gelie, F., 2014. Dyadic effects of coping strategies on emotional state and quality of life in prostate cancer patients and their spouses. Psychooncology 23, 797–803.

Lang, H., France, E., Williams, B., Humphris, G., Wells, M., 2013. The psychological experience of living with head and neck cancer: a systematic review and meta-synthesis. Psychooncology 22, 2648–2663.

Li, Q., Loke, A.Y., 2014. A systematic review of spousal couple-based intervention studies for couples coping with cancer: direction for the development of interventions. Psychooncology 23, 731–739.

Manne, S., Badr, H., Zaider, T., Nelson, C., Kissane, D., 2010. Cancer-related communication, relationship intimacy, and psychological distress among couples coping with localized prostate cancer. Journal of Cancer Survivorship 4, 74–85.

Martire, L.M., Schulz, R., Helgeson, V.S., Small, B.J., Saghafi, E.M., 2010. Review and meta-analysis of couple-oriented interventions for chronic illness. Annals of Behavioral Medicine 40, 325–342.

McClure, K.S., Nezu, A.M., Nezu, C.M., O'hea, E.L., Mcmahon, C., 2012. Social problem solving and depression in couples coping with cancer. Psychooncology 21, 11–19.

Mitchell, A.J., Chan, M., Bhatti, H., Halton, M., Grassi, L., Johansen, C., Meader, N., 2011. Prevalence of depression, anxiety, and adjustment disorder in oncological, haematological, and palliative-care settings: a meta-analysis of 94 interview-based studies. Lancet Oncology 12, 160–174.

Moorey, S., 2013. I know they are distressed. What do I do now? Psychooncology 22, 1946–1952.

Morreale, M.K., 2011. The impact of cancer on sexual function. Advances in Psychosomatic Medicine 31, 72–82.

Nakaya, N., Saito-Nakaya, K., Bidstrup, P.E., Dalton, S.O., Frederiksen, K., Steding-Jessen, M., Uchitomi, Y., Johansen, C., 2010. Increased risk of severe depression in male partners of women with breast cancer. Cancer 116, 5527–5534.

Pirl, W.F., Fann, J.R., Greer, J.A., Braun, I., Deshields, T., Fulcher, C., Harvey, E., Holland, J., Kennedy, V., Lazenby, M., Wagner, L., Underhill, M., Walker, D.K., Zabora, J., Zebrack, B., Bardwell, W.A., 2014. Recommendations for the implementation of distress screening programs in cancer centers: report from the American Psychosocial Oncology Society (APOS), Association of Oncology Social Work (AOSW), and Oncology Nursing Society (ONS) joint task force. Cancer 120, 2946–2954.

Regan, T.W., Lambert, S., Girgis, A., Kelly, B., Kayser, K., Turner, J., 2012. Do couple-based interventions make a difference for couples affected by cancer? A systematic review. BioMed Central. Cancer 12, 279.

Sales, P.M., Carvalho, A.F., Mcintyre, R.S., Pavlidis, N., Hyphantis, T.N., 2014. Psychosocial predictors of health outcomes in colorectal cancer: a comprehensive review. Cancer Treatment Reviews 40, 800–809.

Semple, C., Parahoo, K., Norman, A., Mccaughan, E., Humphris, G., Mills, M., 2013. Psychosocial interventions for patients with head and neck cancer. Cochrane Database of Systematic Reviews 7, Cd009441.

Shimizu, K., 2013. Effects of integrated psychosocial care for distress in cancer patients. Japanese Journal of Clinical Oncology 43, 451–457.

Stanton, A.L., Revenson, T.A., Tennen, H., 2007. Health psychology: psychological adjustment to chronic disease. Annual Review of Psychology 58, 564–592.

Stenberg, U., Ruland, C.M., Miaskowski, C., 2010. Review of the literature on the effects of caring for a patient with cancer. Psychooncology 19, 1013–1025.

Subnis, U.B., Starkweather, A.R., Mccain, N.L., Brown, R.F., 2014. Psychosocial therapies for patients with cancer: a current review of interventions using psychoneuroimmunology-based outcome measures. Integrative Cancer Therapies 13, 85–104.

Tamagawa, R., Garland, S., Vaska, M., Carlson, L.E., 2012. Who benefits from psychosocial interventions in oncology? A systematic review of psychological moderators of treatment outcome. Journal of Behavioral Medicine 35, 658–673.

Tedeschi, R.G., Calhoun, L.G., 1995. Trauma and Transformation: Growing in the Aftermath of Suffering. Sage, Thousand Oaks, CA.

Tedeschi, R.G., Calhoun, L.G., 2004. Posttraumatic growth: conceptual foundations and empirical evidence. Psychological Inquiry 15, 1–18.

Traa, M.J., De Vries, J., Bodenmann, G., Den Oudsten, B.L., 2014. Dyadic coping and relationship functioning in couples coping with cancer: a systematic review. British Journal of Health Psychology. http://dx.doi.org/10.1111/bjhp.12094.

Valdivieso, M., Kujawa, A.M., Jones, T., Baker, L.H., 2012. Cancer survivors in the United States: a review of the literature and a call to action. International Journal of Medical Sciences 9, 163–173.

Wootten, A.C., Abbott, J.M., Farrell, A., Austin, D.W., Klein, B., 2014. Psychosocial interventions to support partners of men with prostate cancer: a systematic and critical review of the literature. Journal of Cancer Survivorship 8, 472–484.

Zimmermann, T., Herschbach, P., Wessarges, M., Heinrichs, N., 2011. Fear of progression in partners of chronically ill patients. Behavioral Medicine 37, 95–104.

Relevant Websites

www.cancer.org – American Cancer Society.
www.cancer.gov – National Cancer Institute.
www.nccn.org – National Comprehensive Cancer Network.

Cancer Screening

Arja R Aro, University of Southern Denmark, Esbjerg, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cancer is a growing health problem worldwide, mainly due to the expanding elderly population. Screening for cancer aims to reduce cancer mortality and morbidity by targeting asymptomatic people to identify those at increased risk of having either disease or precancerous changes. Screening is an intervention on groups of people who have not demanded the service. This means that benefits and harms of screening programs need to be carefully weighed against each other; these can be studied at societal, service provider, and individual levels. Rapidly developing biomedical knowledge on cancer and screening technology as well as increasingly informed citizens pose challenges for cancer screening programs.

Cancers

Cancer is a generic term for a large group of diseases that can affect any part of the body. Other terms used are malignant tumors and neoplasms. One defining feature of cancer is the rapid creation of abnormal cells that grow beyond their usual boundaries, and which can then invade adjoining parts of the body and spread to other organs. This process is referred to as metastasis. Metastases are the major cause of death from cancer. Typically cancer develops slowly and has a long preclinical or prepathological phase. Cancer screening is based on the assumption that it improves prognosis that the required treatment would be less severe thereby contributing to improved quality of life and savings for the society.

Difference between Screening and Testing

Screening for cancer refers to any kind of test performed for the systematic detection or exclusion of cancer, precancer, or susceptibility to cancer assuming that the *initiative comes from outside the individual*, i.e., from the health care provider or screening organizer. Screening distinguishes apparently healthy people who probably have the target condition from those who do not. The former can then be offered information, further tests and appropriate treatment to reduce their risk and/or any complications arising from the disease or condition. Screening is different from *testing* where the *initiative comes from the individual*, or from a health professional, *in the course of a diagnostic process*. People screened are less prepared for the screening result, less aware about the condition they are screened for, and normally have less contact with medical staff than those being tested. Since screening is aimed at unaware, asymptomatic people, it has heavier public health, social, psychological, and ethical implications than testing.

Rationale for Screening

Using the classical screening criteria by Wilson and Jungner (Wilson and Jungner, 1968) as a basis, several later screening criteria have been synthesized into a general version (Andermann et al., 2008), which accommodates developments in research on biomedicine and technology, especially genetics and genomics but also the growing evidence-base and societal changes reflecting more citizens' influence on policies.

The screening criteria are the following (Andermann et al., 2008):

- The screening program should respond to a recognized need.
- The objectives of screening should be defined at the outset.
- There should be scientific evidence of screening program effectiveness.
- The program should integrate education, testing, clinical services, and program management.
- There should be quality assurance with mechanisms to minimize potential risks of screening.
- The program should ensure informed choice, confidentiality, and respect for autonomy.
- The program should promote equity and access to screening for the entire target population.
- Program evaluation should be planned from the outset.
- The overall benefits of screening should outweigh the harms.

Specific screening guidelines for different cancer types have been developed and updated to be in line with the scientific advancements, growing evidence-base of the screening effectiveness and contextual priorities such as national screening policies. The public health rationale behind cancer screening is to reduce mortality and morbidity of the target condition. According to the present knowledge, for some cancer forms (e.g., cervical cancer, colorectal cancer) precancerous, treatable lesions or changes can be found via screening and thus screening can be preventive. For some other cancers (e.g., breast cancer) the idea is to find cancer in a preclinical asymptomatic phase before it would be found without screening. The assumption is that earlier detection of cancer is beneficial in terms of less invasive (and less costly) treatment, improved survival from cancer and less disruption of quality of life of those screened. Screening for biological markers, such as prostate-specific antigen (PSA) for prostate cancer risk and human papillomavirus (HPV) types for cervical cancer risk, identifies those who have a heightened risk of cancer and thus might benefit from follow-up or further investigations. Recent developments in genomics have provided the technology, which enables screening for susceptibility or genetic defects indicating increased risk of cancer. In general, however,

according to current knowledge, only 5–10% of the cancers seem to be inherited – resulting directly from gene defects (called *mutations*) inherited from a parent. Thus due to most cancers being multifactorial in their origin, genetic screening is rarely cost-effective except in certain subpopulations (e.g., families) with increased risk for certain forms of cancer, for example, certain familial breast cancer mutations (e.g., BRCA1 & 2). This kind of genetic screening is based on the assumption that knowledge of risk enables either primary prevention via lifestyle, environmental changes, or via targeted medical interventions.

Screening Programs as Policies

The fact that even countries with rather similar cancer prevalence rates and the same available international scientific research evidence on the effectiveness of cancer screening, have adopted very different screening programs, shows how much screening is a policy decision. Policy decisions are dependent on the best available research evidence but also on contextual issues such as resources, characteristics of the population, know-how of the practitioners and policymakers as well as priorities and values of citizens (Satterfield et al., 2009). Further, the cost of large-scale or high-technology screening exceeds the resources of most societies. For example, in developing countries, low-technology options such as visual inspection for cervical cancer have been found feasible and also effective.

Screening programs are optimal for frequent cancer types for which a cost-effective, affordable, acceptable, and accessible screening test is available to the majority of the population at risk. Estimates of premature deaths avoided by screening vary from 3 to 35%. Different screening programs target different populations such as age groups depending on the prevalence of cancer and other factors. Performance of the screening tests can be measured by sensitivity, specificity, positive and negative predictive values. Sensitivity refers to the probability that a person who has cancer gets a positive test result, whereas specificity is the likelihood that a person not having cancer receives a negative test result. The positive and negative predictive values are the proportions of positive and negative results that are true positive (i.e., cancer) and true negative results (i.e., no cancer). The positive predictive value, and also to some extent negative predictive value, are affected by the prevalence of the disease in the screened population. For a given sensitivity and specificity, the higher the prevalence, the higher the positive predictive value of screenings tests.

Examples of Cancer Screening Programs

An example of a cancer amenable to screening is breast cancer (the most common female cancer) (by mammography screening), and also cervical cancer (by Pap test screening). Screening programs in several countries are based on the evidence from randomized controlled studies showing that disease-specific mortality among women over the age of 50 years was reduced by up to one-third due to mammography screening every 2–3 years (Kerlikowske et al., 1995). Nevertheless, more recently, benefits and harms of mammography screening have been debated. The decline in mortality from breast cancer has been explained in terms of improved treatment approaches rather than the effects of screening; also, the effects of potential overdiagnosis and the psychosocial consequences of the screening process have been emphasized (Gøtzsche and Jørgensen, 2013). An independent reevaluation of the evidence on mammography screening concluded that the breast cancer mortality reduction is probably around 15% instead of close to 30% as previously estimated, and that the claimed overdiagnosis with its negative consequences happens to some extent (Marmot et al., 2013).

For cervical cancer screening effectiveness there has been less strictly controlled research but also less debate; this is probably since the available screening programs can identify precancerous changes, which can be treated and thus prevent cancer from occurring. Screening programs in different countries have targeted a range of age groups and used different screening intervals, but in general, it has been shown that Pap test screening has been able to significantly decrease cervical cancer mortality (Peirson et al., 2013). Research has also brought new screening methods in addition to Pap test screening. One example of these is screening for HPV, some forms of which seem to be related to cervical cancer. Developments in this area have led to launching vaccination programs with the aim of preventing cervical cancer. In developing countries with a lower level of available technology and resources, studies have shown feasibility and usefulness of visual inspection using acetic acid as a cervical cancer screening method (e.g., Ibrahim et al., 2012).

Prostate cancer among males is an example of cancer screening, which has been studied in well-controlled trials. The most common screening method has been the PSA blood test indicating an increased risk for prostate cancer. Nevertheless, there is no conclusive evidence that PSA screening results in a significant reduction in either disease specific or overall mortality; further, potential harms (e.g., overdiagnosis and overtreatment) may outweigh the (uncertain) benefits of screening (Hoffman, 2010; Lee et al., 2013). Prostate cancer occurs mostly in old age and invasive treatment can cause a significant decrease in quality of life and functioning.

An example for cancer screening programs for both genders, which is carried out in several countries is colorectal cancer screening. For colorectal cancer screening several clinical tests exist, but also a self-administered stool sample test (fecal occult blood test) is available. It is currently estimated that these screening programs have led to a reduction in colorectal cancer mortality by 23% (Hewitson et al., 2008).

In the future instead of cancer-specific screening programs, it might be possible to do whole body screening for multiple sites using imaging technologies and serum marker tests.

Social and Psychological Aspects of Cancer Screening

Determinants of Screening Provision and Uptake

In society, screening can be studied from the perspectives of society, health services offered, and of people taking part in screening, although in real life these levels are interdependent

and imbedded. Different cultural, health system, and resource factors and values at society and service level *determine the screening provided* (Table 1). *Uptake of screening* is affected by the social and cultural context, the way screening is organized, and by individual factors. The more initiative is left to the individual, the more important is the person's health motivation in determining the use of screening programs. Interventions using postal invitation, telephone reminders, general practitioners' invitation, and appointment instead of open invitation and self-sampling, have been shown to increase cancer screening participation (Camilloni et al., 2013; Ferroni et al., 2012). Examples of individual-level determinants are socioeconomic background, health behavior in general, beliefs, perceived risk, and health literacy, knowledge as well as experiences related to illnesses in general, and to cancer and screening (e.g., Aro et al., 1999).

Theoretical models used to study determinants of cancer screening uptake have mostly been based on cognitive expectancy value models predicting self-initiated participation. These models have only been partly successful at predicting screening behavior, which is probably due to the fact that in screening, the social and service contexts are essential, which are not included in many individual and cognitive models (e.g., concept of intention, excluding concern and emotions). Also, these models are often too general for different types of cancer screening. Theories on treatment adherence or coping with threat can offer alternatives, but also stages of change theories can be applied to adherence to repeated screening rounds, especially if the context factors are included in the model. Though not easily administered in large-scale screening programs, personalized risk communication for informed decision-making about taking screening tests seems to increase informed decision-making compared to more general information. In addition, knowledge and accuracy of risk perception might increase using personal risk communication; anxiety about screening, however, does not seem to be affected (Edwards et al., 2013). A step forward to a more multilevel framework to understand cancer screening uptake and the role of socioeconomic factors has been presented by von Wagner et al. (2011). In this article, the authors describe possible pathways by which life experiences related to socioeconomic status (stressors, illness experience, resources, education) might influence psychosocial factors (e.g., beliefs, confidence) playing a role in cancer screening participation.

Implications of Screening Provision and Uptake

Screening Provision

Screening provision can have both positive and negative implications. Positive implications at society level include increased awareness of cancer and health in general (Table 1). At service level, positive implications can include increased competences among the screening and health care staff. At the individual level, screening provision can increase health motivation. On the potential negative side, at society-level screening provision increases costs and can also increase medicalization of the health conditions targeted by screening programs. The service level faces increased diagnostic work and treatment, demand on staff capacity and resources in general. People *invited* to, or otherwise made aware of screening, might get worried about the screened condition.

Screening Uptake

Screening programs aim at high uptake rates, which would increase lives saved, a decrease in morbidity and improvement in quality of life of those taking part, preferably in a cost-effective way (Table 1). This would bring benefits to the society. However, since screening also detects cancers, which otherwise would never become clinical, or which would appear so late that they would not lead to death or increased morbidity (overdiagnosis and overtreatment), screening

Table 1 Screening in a society: examples of determinants and implications of screening provision and uptake on different levels of society

	Screening provision		*Screening uptake*	
	Determinants	*Implications*	*Determinants*	*Implications*
Society level	Culture Health and insurance policy Health care Resources Values	Awareness Costs Medicalization	Cultural and social context Health care and insurance	Lives saved Decrease of morbidity Awareness Costs In case of overdiagnosis and overtreatment: increased morbidity
Service level	Health care Insurance Tradition Values Resources	Staff competences Workload from screening program Need for diagnostics, treatment, follow-up, education, staff capacity Costs	Organization of screening, e.g., invitation	Increased staff competences Workload in diagnostics, treatment, follow-up Costs
Individual level	NA[a]	Health motivation Awareness Worry	Socioeconomic background Health behavior, beliefs, knowledge, and literacy Experience	Relief Improved quality of life inconvenience Worry False assurance Decreased quality of life in case, no survival or morbidity benefit

[a]Not applicable.

uptake can also bring (unnecessary) workload and costs for service providers as well as harms to the participants, although these can be estimated only at group level. Most of the individual-level implications are related to the screening process, which may include several rounds of assessments. Already invitation to screening may raise worry and a screening test may be inconvenient (e.g., painful or embarrassing to some) or bring other hazards. A *normal screening finding* normally brings relief, but the screening procedure with additional assessments can cause worry (Aro et al., 2000; Metsälä et al., 2012). For those with a *false positive* screening finding, worry, especially cancer-specific distress, and its behavioral indicators such as preoccupation with symptoms, decreased screening uptake but also increased use of health services, might last long past the reassuring diagnosis (Aro et al., 2000; Bond et al., 2013). Further, distress seems to increase with the level of invasiveness of the screening procedure (Bond et al., 2013). *A true positive diagnosis* of cancer due to screening may both prolong life and improve its quality. However, this impact cannot be evaluated individually but only at a group level. Finding and treating cancer, which would never have become clinical in a lifetime, is clearly a major adverse effect of screening. *A false negative* finding misleadingly reassures and delays diagnosis. *A true negative* finding is reassuring, assuming that screening has no adverse effects.

Theories guiding research on implications of the screening process and its findings should cover cognitive issues such as perceptions and illness beliefs, coping with health threat, health behaviors, and quality of life. Persons needing to go through further assessments in the screening process especially need targeted attention and accessible material to alleviate their concern. It is important to take into account health literacy in general and especially cancer literacy of the target group of screening and of all screening participants; in delivering cancer risk information new genomic knowledge and its implications for cancer prevention and monitoring need to be taken into account.

Challenges in Cancer Screening and Research on It

Cancer – A Growing Problem

Screening, however, is not the solution for cancer control. It is cost-effective only for certain age or risk groups, and mostly not globally feasible. Prevention would be better than screening for an already existing disease, but only part of the causal factors of cancer are understood. It is estimated that about one-third of cancer deaths are due to the five leading behavioral and dietary risks: high body mass index, low fruit and vegetable intake, lack of physical activity, tobacco use, and alcohol use. Examples of these lifestyle-related cancers linked to economic development are lung, breast, and colorectal cancers. Due to the aging population the number of cancer cases will increase in the future. Even if fewer people will die from cancer in the future due to improved treatment in the high resource countries, the situation might not be the same in low resource settings. By 2030, the global cancer burden is expected to nearly double, growing to 21.4 million cases and 13.2 million deaths; most of these cancer deaths will probably occur in low- and middle-income countries.

Cancer Screening

Health literacy and especially cancer literacy, risk perception and risk communication as well as informed consent of participants in large-scale screening programs pose major challenges since screening programs should deliver messages to very diverse target groups with little face-to-face contact. Growing evidence on screening effectiveness and rapid developments both of biomedicine and of screening technology, require regular updating of screening-related material, decisions, and guidelines. Development of decision aids for public health decision-making (e.g., Tso et al., 2011) might prove useful – as long as these respect the inclusive and ethical principles of public health decision-making (Satterfield et al., 2009; Tannahill, 2008). From the public health perspective, instead of individually based *informed consent*, one could talk about *informed contract* between rapidly developing science and society in developing and deciding of health policies (Nuffield Council on Bioethics, 2007). To realize this kind of informed contract, public debate and the empowerment of citizens to take part in developing screening policies can be supported by public discussion on values, ethics, and priorities in communities and societies. Modeling software provides one way of forecasting effects and impacts of screening where as experimental studies are not acceptable or feasible. People need knowledge and skills to control the cancer problem in their roles of societal level decision-makers, service providers, and as individual citizens.

Research

Comprehensive research on social and psychological aspects of cancer screening is best done using a multilevel approach so that both determinants and implications of screening provision and uptake, including the levels of society, service, and the individuals, are taken into account. A prospective design with nonscreened referents – most easily built in organized screening programs with planned schedules – enables more reliable measurement of predictive determinants of both uptake and consequences of screening. Theory-based research using context-specific measures, which are specific to the health condition screened in addition to standard measures or those measuring clinical morbidity levels, might provide more useful information in terms of developing tailored support and interventions for screening participants in different phases of the screening process.

Conclusion

Cancer screening programs as health policies belong to multidisciplinary and multisector actions, which need careful updating, planning, monitoring, and evaluation. Determinants and consequences of both screening provision and uptake can best be studied at multiple levels of society, service, and individual level.

See also: Cancer: Psychosocial Aspects; Health Behavior, Psychosocial Theories of; Health Behaviors; Health Care,

Rationing of; Health Literacy; Illness Behavior and Care-Seeking; Patient Adherence to Health Care Regimens; Personalized Medicine; Public Health as Social Science.

Bibliography

Andermann, A., Blancquaert, I., Beauchamp, S., Déry, V., 2008. Revisiting Wilson and Jungner in the genomic age: a review of screening criteria over the past 40 years. Bulletin of the World Health Organization 86 (4), 317–319.

Aro, A.R., Absetz, P.S., van Elderen, T.M., van der Ploeg, E., van der Kamp, L.J.T., 2000. False-positive findings in mammography screening induce short-term distress—breast cancer-specific concern prevails longer. European Journal of Cancer 36, 1089–1097.

Aro, A.R., de Koning, H.J., Absetz, P.S., Schreck, M., 1999. Psychosocial predictors of first attendance for organised mammography screening. Journal of Medical Screening 6, 82–88.

Bond, M., Paney, T., Welch, K., Cooper, C., Garside, R.-, Dean, S., Hyde, C., 2013. Systematic review of the psychological consequences of false-positive screening mammograms. Health Technology Assessment 17 (13), 1–170.

Camilloni, L., Ferroni, E., Jimenez, C.B., Pezzarossi, A., Furnari, G., Borgia, P., Guasticchi, G., Giorgi Rossi, P., Methods to Increase Participation Working Group, 2013. Methods to increase participation in organised screening programs: a systematic review. BMC Public Health 13, 464.

Edwards, A.G., Evans, R., Dundon, J., Haigh, S., Hood, K., Elwyn, G.J., 2013. Personalised risk communication for informed decision making about taking screening tests. Cochrane Database Systematic Reviews 18 (4), CD001865. Review. Update in: Cochrane Database System Reviews 2, CD001865.

Ferroni, E., Camilloni, L., Jimenez, B., Furnari, G., Borgia, P., Guasticchi, G., Giorgi Rossi, P., Methods to Increase Participation Working Group, 2012. How to increase uptake in oncologic screening: a systematic review of studies comparing population-based screening programs and spontaneous access. Preventive Medicine 55 (6), 587–596.

Gøtzsche, P.C., Jørgensen, K.J., 2013. Screening for breast cancer with mammography. Cochrane Database Systematic Reviews 4 (6), CD001877 [Epub ahead of print].

Hewitson, P., Galsziou, P., Watson, E., Towler, B., Irwig, L., 2008. Cochrane systematic review of colorectal cancer screening using the fecal occult blood test (hemoccult): an update. American Journal of Gastroenterology 103 (6), 1541–1549.

Hoffman, R.M., 2010. Randomized trial results did not resolve controversies surrounding prostate cancer screening. Current Opinion in Urology 20 (3), 189–193.

Ibrahim, A., Aro, A.R., Rasch, V., Pukkala, E., 2012. Cervical cancer screening in primary health care setting in Sudan: a comparative study of visual inspection with acetic acid and Pap smear. International Journal of Women's Health 4, 67–73.

Ilic, D., Neuberger, M.M., Djulbegovic, M., Dahm, P., 2013. Screening for prostate cancer. Cochrane Database Systematic Reviews 31 (1), CD004720.

Kerlikowske, K., Grady, D., Rubin, S.M., Sandrock, C., Ernster, V.L., 1995. Efficacy of screening mammography. A meta-analysis. Review Journal of the American Medical Association 273, 149–154.

Lee, Y.J., Park, J.E., Jeon, B.R., Lee, S.M., Kim, S.Y., Lee, Y.K., 2013. Is prostate-specific antigen effective for population screening of prostate cancer? A systematic review. Annals of Laboratory Medicine 33 (4), 233–241.

Marmot, M.G., Altman, D.G., Cameron, D.A., Dewar, J.A., Thompson, S.G., Wilcox, M., 2013. The benefits and harms of breast cancer screening: an independent review. British Journal of Cancer 108 (11), 2205–2240.

Metsälä, E., Pajukari, A., Aro, A.R., 2012. Breast cancer worry in further examination of mammography screening – a systematic review. Scandinavian Journal of Caring Sciences 26, 773–786.

Nuffield Council on Bioethics, 2007. Public Health: Ethical Issues. Nuffield Council. http://www.nuffieldbioethics.org/sites/default/files/Public%20Health%20-%20short%20guide.pdf (accessed 05.07.13.).

Peirson, L., Fitzpatrick-Lewis, D., Ciliska, D., Warren, R., 2013. Screening for cervical cancer: a systematic review and meta-analysis. Systematic Reviews 2, 35.

Satterfield, J.M., Spring, B., Brownson, R.C., Mullen, E.J., Newhouse, R.B., Walker, B.B., Whitlock, E.P., 2009. Toward a transdisciplinary model of evidence-based practice. The Milbank Quarterly 87 (2), 368–390.

Tannahill, A., 2008. Beyond evidence—to ethics: a decision-making framework for health promotion, public health and health improvement. Health Promotion International 23 (4), 380–390.

Tso, P., Culyer, A.J., Brouwers, M., Dobrow, M.J., 2011. Developing a decision aid to guide public sector health policy decisions: a study protocol. Implementation Science 6, 46.

von Wagner, C., Good, A., Whitaker, K.L., Wardle, J., 2011. Psychosocial determinants of socioeconomic inequalities in cancer screening participation: a conceptual framework. Epidemiologic Reviews 33, 135–147.

Wilson, J.M.C., Jungner, G., 1968. The Principles and Practice of Screening for Disease. WHO, Geneva. http://www.int/bulletin/volumes/86/4/07-05011BP.pdf (accessed 01.07.13.).

Relevant Websites

http://www.who.int – Definition of cancer.
http://www.screening.nhs.uk/screening – Description of the screening process.
http://www.cancer.org/cancer/cancercauses – Inherited cancers.
www.cancer.gov – Performance of screening tests.

Cannibalism

Shirley Lindenbaum, City University of New York, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D. Tuzin, volume 3, pp. 1452–1454, © 2001, Elsevier Ltd.

Abstract

Cannibalism has a long-established place in the chronicles of human history. Shifting representations of the cannibal tell a story of changing Western sensibilities. Difference and novelty, once appreciated, were displaced during the eighteenth century by the ranking of societies on a single evolutionary continuum. Cannibalism became the prime signifier of barbarism, an attitude that would harden during the nineteenth century in biology-based racism. Archaeological and genetic evidence of cannibalism in ancient European population, as well as more recent historical reports, provide an opportunity for critical reflection on the meaning the practice held for ourselves and for others.

The practice of cannibalism, the eating of one's own kind, has a long-established place in the chronicles of human history. Archaeologists find evidence of human cannibalism before the invention of metals, the building of the Egyptian pyramids, and the origins of agriculture. The ancient literature from Homer to Strabo and the Roman historians mentions anthropophagi living on the borders of the known world. In the fifth century, Herodotus said man-eating people were found among the Scythians in the environs of the Black Sea. The Crusades provided images of savages who preferred human flesh to that of nightingales stuffed with spices. The word cannibal, rather than the older word anthropophagus, is said to be a legacy of Columbus' encounter with man-eating Caribs of the Antilles in 1492.

The shifting representation of the cannibal provides a history of changing sensibilities. Ancient, medieval, and early modern sources described savage nations of anthropophagi in semihuman form, hybrids of animal and man. For medieval travel writers, the typical cannibal monster was the cynocephalus, the dog headed being who inspired the imagination of Marco Polo. Idolatry and cannibalism were not condoned, but the behaviors of distant and strange races in Asia and Africa presented no challenge to European values. Thirteenth- and fourteenth-century authors portrayed difference and novelty as positive qualities, a source of pleasure and delight, although cannibal behavior served also to satirize courtly and aristocratic culture. Montaigne's denunciation in 1577 of the cruelty displayed by Christians against one another during the religious wars in France is viewed as an early statement of cultural relativism. There was more barbarity, he thought, in eating a man alive than in eating him after he was dead, as was customary in other nations, and in the tortures of the rack for bodies full of feeling, or roasting a man bit by bit, as had occurred among his fellow citizens on the pretexts of piety and religion (Montaigne, 1581). The Renaissance image of the heroic cannibal adopted by Montaigne and the first missionaries in the New World reflected the aristocratic values of magnanimity and chivalry. Renaissance views also held cannibalism to be 'natural' because the four humors make human flesh tasty to humans who are made of the same stuff.

The earliest images of cannibalism for European eyes were provided by the illustrators of the de Bry's *Great Voyages* (1590–1634). For these seventeenth-century Protestant engravers cannibalism, although not considered acceptable, was viewed as a theological sin, a form of immoderate eating; the Tupinambas were thus judged to be pagans and savages for committing one or another of four kinds of gluttony (Bucher, 1981). This view would soon be challenged by the greed of the conquistadors and the excessive zeal of missionaries. Political and economic conditions in Europe provided the context for dwelling on images of gourmandizing excess. Poverty and unrest allowed Protestant and Republican authors to focus on the sins of appetite associated with unreformed religion, and morality plays depict monks and priests as gluttonous lechers. The Spanish were viewed as tyrants consuming other nations to satisfy their Babylonian, Roman Catholic, and absolutist ambitions (Healey, 2002). Debate over the Eucharist, fraught with cannibal associations since the earliest days of the Christian church, became a major point of polemical contest in Reformation Europe. Carried to the New World, the analogy between the sacrament and New World anthropophagy allowed Calvinists to compare Catholics to the Outeca, cannibals who ate their victims raw, an allusion to sanguinary sacrifice in the Catholic mass (Rawson, 1997).

In the context of increasing overseas trade, religious proselytization, and imperial expansion, some eighteenth-century authors began to explore the perverse attractiveness of the image of the cannibal to reflect on certain features of their own society (Cottom, 2001). The trope of the good and the bad cannibal (the noble and ignoble savage) reappeared and became transformed with French and British adventures in the Pacific, which settled on new images of seduction and transgression. In 1719, Defoe portrayed the cannibal as a nexus of ambiguous desire; Robinson Crusoe is attracted both to the idea of eating human flesh and to being eaten. The ambivalence of cannibal transgression reached its extreme in de Sade's 1797 novel, *Juliette*, whose characters gorge on wines, liqueurs, shit, semen, and bits of human flesh, his critique of colonial exploitation perhaps undone by his exploitation of cannibal eroticism.

During the eighteenth century, a significant change took place in the way the inhabitants of the new worlds came to be understood. Renaissance perceptions of difference had been

religious, rather than racial or national. The intruders were not described as Spaniards, Europeans, or whites, but as Christians, and the 'Indian' was not a savage of any particular type, but a naked pagan, to be converted. The shift from religiously framed difference entailed new models for constructing otherness, couched in a narrative of natural history. Humanitarian assessments of dissimilarity (as well as some negative judgments) were displaced by a more uniform, intolerant denigration of physical inferiority and moral fault. This decisive move in Enlightenment thinking, away from mythological, religious causes of social evolution to material causes, recast the history of mankind in a single continuum, with societies ranked higher and lower on the same scale. A discourse about social progress depended on the figures of the noble and ignoble savage, formulated in terms of the West and the Rest, and provided a language in which Western modernity first came to be defined (Hall, 1992). Cannibalism was to become the prime signifier of barbarism for a language of natural typification and essential difference that would harden in the biologically based negative racism of the nineteenth century.

The image of the noble savage had attained sociological status with Rousseau's account of simple, unsophisticated people living in a state of nature, unfettered by government, property, or social divisions. This state of paradise was discovered anew with French and British explorations in the Pacific. Bougainville said of Tahiti that he thought himself transported to the Garden of Eden. When Cook arrived in Tahiti in 1769, the same idyll of sexual paradise and ideas of innocence were repeated, and when he brought the Tahitian Omai to Europe in 1774, his tattoos fascinated London society, sparking a tattooing vogue among the English aristocracy.

By the end of the eighteenth century, however, as maritime exploration turned its focus to the scientific classification of animal and human species, and with the death and disappearance of the navigator Defresne, killed and eaten by the Maori, as well as the death of Cook in Hawaii in 1779, the image of the noble savage was no longer tenable. The stereotype of the noble savage always contained its mirror opposite, the ignoble savage. In the South Pacific, the dualism became gendered in the form of the beautiful, enticing siren, and the malicious, bestial cannibal, a distinction that would be extended to civilized, white, Polynesians and savage, black Melanesians (Boulay, 2001).

Savage cannibals were reported to be present in nineteenth-century New Zealand. Maori warriors were said to eat fallen enemies and to have massacred and consumed the passengers of the crew of the *Boyd* in 1806. Perhaps the best-known account of the hazards of seafaring was that of the crew of the *Essex*, the whaling ship rammed by a sperm whale in the Pacific in 1820, the event that inspired Melville's *Moby Dick*. The crew of 20 took to three small boats and were blown around the Pacific for 90 days, enduring thirst and hunger, succumbing finally to cannibalism. Only eight of them survived.

The construction of the stereotype of the ignoble savage took place on many fronts. Beginning in London in 1851 and ending in 1931 in Paris, international exhibitions proclaimed the superiority of Western civilization. Visitors to the Paris exhibition were invited to view the skulls of primitive men organized by the exhibit's director of anthropology, illustrating the ladder of races with whites at the top. One hundred 'kanaks' were imported from New Caledonia for the 1931 exhibition in the Bois de Boulogne. Newspapers ran stories about two sensational attractions: crocodiles and wild cannibals. The kanaks also were exhibited as cannibals and polygynists in German towns, although most of these Polynesians by then were practicing Catholics.

Nineteenth-century missionary literature, called on to explain the existence of savagery in divine creation, described a migration of the sons of Ham, starting from the banks of the Euphrates, passing through India, and coming to rest in Melanesia. Thus, one found among the blacks in Africa, Asia, and Oceania "loathsome vices … and cannibalism in all its hideous subtleties" (Boulay, 2001: pp. 104–109). Nineteenth-century romantic images of the heroic missionary becoming a meal for cannibals, perhaps to impress mission donors, continued well into the twentieth century, a chestnut still found in cartoons and the repertoire of comedians. The discovery of gold in Papua New Guinea extended the life of similar tales of personal encounters with savage cannibals.

Airbrushed out of these mythologies of heroic adventure are the complexities of interethnic encounter and sexual confrontation, as well as the more gruesome stories of conflict, violence, and dispossession. To justify political intervention at the end of the nineteenth century, the adventure literature had settled on the Fijian King Thakombau (for the English) and Chief Bourarate de Hienghene, the 'king' of New Caledonia (for the French) to represent the abominations of anthropophagy (Boulay, 2001). Famine-based cannibalism was reported during the Siege of Leningrad in the winter of 1941–42 and in southern China during the Cultural Revolution. Japanese soldiers testified to eating their own troops, enemy dead, and prisoners of war in New Guinea during World War II when their daily rations were insufficient.

Salacious accounts of cannibalism are now a staple for the popular press, providing a frisson of pleasure for audiences reading about fellow citizens said to be suffering from severe personality disorder or psychosis. Interest in cannibalism appears to have increased in recent years, even as the practice becomes less frequent.

This brief walk through the history of human cannibalism illustrates the ubiquity, geographic diversity, and changing perceptions of the practice. The topic has long provided a vehicle for historians, geographers, philosophers, theologians, and others to speak about the realities of their time. The eater of human flesh once provided a vehicle for theoretical debates among the philosophers of natural law, an encompassing form of scholarship that had influence in the worlds of politics and science from the sixteenth to the eighteenth centuries, before philosophy, history, science, and anthropology became separate disciplines. The sciences of natural law may have disappeared, but dimensions of the debates are still present in anthropological inquiry (Lindenbaum, 2004).

Natural Law

The cannibal is said to be one of the great forgotten figures in the history of philosophy (Avramescu, 2011: p. 2). Avramescu's cannibal is depicted as a scholarly creature, viewed from the

perspective of natural law, the content of which is set by nature, and thus having universal validity. This pre-Enlightenment cannibal enters the collective Western imagination as a central figure in a theater of lawless figures existing in a condition of natural anarchy, a generalized war of all against all, as a shaper of European moral, legal, and political thought. The discovery of man-eaters in the Caribbean and travelers' accounts of cannibalism, real and imaginary, raised new questions for natural law.

Did these people believe in God and what was their conception of divine nature? Did they have a form of government, and if so, was it monarchical? Did they have private property? And did they eat people? Finding no description of such savages in the literature of antiquity or in the Bible, natural philosophers began to question the relevance of these texts. Were cannibals in a state of nature merely representatives of that state, and thus essentially good, or lawless savages whose human nature must be restrained by a strong sovereign? The cannibal entered the history of philosophy on both sides, as a positive representative, governed by the laws of nature, acting from absolute necessity, and as a deviant from its laws, committing crimes against nature.

The implications of anthropophagy raised a serious challenge also to the Christian doctrine of resurrection and the question of identity. When a starving man ate another, in what body would the man eaten be resurrected? Had his flesh been converted into the man who ate him? Would his flesh be restored to him, or remain part of the cannibal? Some Christian apologists solved this problem by asserting that the personal identity of a rational being consisted of its interior sentiment and that spiritual bodies differed from their earthly counterparts. The resurrected body thus would be a body that differed from the natural one.

The New World practice of cannibalism, considered to be a monstrous crime, presented a problem for evangelization. Are these anthropophagi men or subhuman beings, incapable of receiving the Gospels? Even if their quality as human beings is admitted, cannibalism is incompatible with any religious belief. The opinions of the learned were divided. Some said that knowledge of God was diffused among all nations, including the most barbarous, who might be enlightened by knowledge of duties toward God, without which they would lack the strength to respect God's commandments.

Difficulties arose also over the concept of salvation in the case of shipwrecks and war, when the bodies of men fed fish or animals that were later eaten by humans, a consumption chain that might result in men feeding other men. The seventeenth-century sciences redefined the problem of resurrection as a scientific project. Aided by the microscope, the original solid particles and vessels were said to be the entire body, the rest but extraneous matter, the nucleus remaining unaffected, and no confusion of bodies would be possible.

The seventeenth-century decline in the doctrine of demonism (a belief that fallen angels had, of their own free will, chosen Satan as their ruler), as well as increasing doubts about travelers' tales, and questions about some aspects of natural law, coincides with the disappearance of the cannibal from learned philosophical discussion, along with a particular style of argument, values, and sensibilities. The marvels that once commanded the attention of philosophers and medical men during the sixteenth and early seventeenth centuries also fell out of favor. Protestant and Catholic clergy, aided by philosophers and naturalists, banished marvels and miracles from intellectual discussion. The passion and unprecedented diffusion of marvels seemed to have fallen into the wrong hands and were now considered potentially dangerous, threatening the monopoly of the power of wonder by princes and churches (Daston and Park, 1998: p. 362). The exit of cannibals and demons from philosophical thought accompanied a change in elite sensibilities. The anthropophagus (the older term) was once a central figure in an inverted world in which the absence of supreme power led to a state of chaos, characterized by the practice of cannibalism. Once also a protagonist in a political and moral philosophy that had addressed the nature of the social order, the anthropophagus is said to have been eclipsed by the State, the new agent of absolute cruelty (Avramescu, 2011: p. 262).

Anthropology

By the end of the nineteenth century, anthropology had emerged institutionally and conceptually from natural philosophy, retaining the observational and taxonomic practices of the earlier sciences for the study of human diversity. The anthropological study of human cannibalism now distinguishes between *endocannibalism*, the volitional eating of a person from within the group, and *exocannibalism*, eating someone from outside the group.

Endocannibalism, often in the form of funerary consumption, has been reported to be present in Papua New Guinea (Gillison, 1983; Glasse, 1967; Knauft, 2002; Lindenbaum, 2013; Whitfield et al., 2008) and Amazonia (Conklin, 2001). *Exocannibalism*, as an act of aggression, frequently in the context of warfare, has been documented for the Americas, the South Pacific, and Africa. The Maka of southeast Cameroon reported eating people from outside their kinship network before their subjugation by the Germans in 1905 (Geschiere, 1998), and more recently by soldiers in the Congo. In some cases in Papua New Guinea and Amazonia both in-group and out-group consumption took place.

A taxonomy of cannibalism now includes behaviors that shade into Western medical and dietary regimes. *Medicinal cannibalism*, the ingestion of human tissue, has been reported since the first century AD. Pliny the Elder said that drinking human blood was a cure for epilepsy. From the sixteenth to the eighteenth century, Europeans ingested human blood and tissue as a medicine or tonic. Human flesh, blood, heart, skull, and bone marrow was obtained from mummy shops, where the embalmed and dried bodies of people who ideally had met with sudden or violent death was considered to be a universal panacea by the Paracelsians, who in contrast to the Galenists, promoted the use of a variety of medicinal body substances. The present-day traffic in human organs, tissue grafts, and human cloning now provides an avenue for *technological cannibalism*. The use of cadaver-derived drugs from pituitary glands for the production of human growth hormones from the late 1960s to early 1980s, led to several hundred iatrogenic cases of Creutzfeldt–Jakob disease in the United States, Europe, and Japan.

Autocannibalism (*autophagy*) refers to the act of eating parts of oneself: hair, nail clippings, mucous, and excrement, but also instances in which individuals under torture and duress partake of their own living flesh, raw or cooked.

Placentophagy, in which the mother eats her newborn baby's placenta, became popular in the United States with the spread of the home-birth movement in the 1970s. Some home-birth vegetarians eat the placenta, which they view as unkilled meat. Just as the home-birth movement provided a niche for *placentophagy*, consuming the ashes of a loved one became feasible in the United States as professional services for the cremated body were no longer felt to be obligatory. *Bone ash cannibalism* has been reported also in South America among the Amahuaca and the Yanomami.

Sacrificial cannibalism, in which the victim is treated with solicitude and honor as a prelude to sacrifice to a god, followed by cannibalism, occurred in fifteenth century among the Aztecs in Mexico, and in nineteenth-century Fiji.

Innocent cannibalism is said to occur when a person is unaware that he or she is eating human flesh. In Greek legend Atreus punishes his brother Thyestes for seducing his wife by tricking him into eating his own children. The unwitting ingestion of human flesh is also a theme in popular entertainment. 'Sweeney Todd' is based on the case of a murderous barber and a baker, partners in crime, who dispose of bodies by baking them into pies. A farmer in 'Motel Hell' makes smoked meats from individuals he considers to be contributing undesirable elements to the gene pool.

Invisible cannibalism is reported in Papua New Guinea among the Ankave-Anga (Lemonnier, 2006). Cannibal beings in immaterial form are said to travel through the air or underground, transforming into their semihuman or animal appearance when they arrive at their destination.

As Tuzin (2001) observed, the inventory of cannibal practices varies not only according to motive and circumstance, but also the diversity is so great that it overwhelms the common feature of ingestion, and confounds efforts to understand the practice as a unitary phenomenon.

Cannibalism and Anthropological Theory

Recent anthropological accounts of cannibalism take us through a twentieth-century history of theory in the social and biological sciences. Cannibalism, like incest, aggression, and other phenomena of supposed universal import, is a productive concept for exploring theoretical propositions. In the late 1950s, the structuralist concerns of British anthropology led scholars to examine cannibalism in the context of relationships among social structure, sorcery, politics, and warfare. At the same time, the utilitarian aspects of functionalism seemed to suggest that human flesh was of some nutritional value (an observation made by Evans-Pritchard in 1937), and an ecological explanation might be applied to tropical peoples living in low-medium population densities where they exploited a diverse range of animal foods. A materialist explanation also was proposed for the ritual slaughter and consumption of war captives by the Aztecs in fifteenth-century Mexico during times of famine and shortage of animal protein. Human flesh was said to provide incentives for warriors and also to sustain the power of elites. The Aztec case provoked a heated debate about population levels and ecological conditions in the Valley of Mexico, as well as the bourgeois grounding of Western-oriented materialist explanations.

Anthropological interest in the practice during the 1980s followed, in part, from Arens (1979) provocative suggestion that institutionalized cannibalism never existed because centuries of reports were not based on reliable eyewitness evidence. In response, many anthropologists reevaluated their data, and more nuanced studies became available from Papua New Guinea, South America, Africa, and Fiji. Anthropologists now acknowledge that institutionalized cannibalism occurred in some places at some times. Noting the ethnocentricity of the criticism, some also called for reflection on our own horror of the practice and for the topic to be discussed in ways that avoided resort to denial or euphemism. Like their scholarly predecessors, twentieth-century anthropologists were challenged by the problem of the 'thinkability' of cannibalism. Levi-Strauss (1964) asserted that superficial reactions of dismay to reports of the 'barbarous' customs of native communities should be resisted, first by separating off its purely alimentary forms, because famine can force men to eat anything, and what remains can be termed positive, since they stem from a mystic, magic, or religious cause. In the context of a theoretical shift in the 1970s and 1980s toward matters of cultural belief, cosmology, and ritual, Sahlins insisted that cannibalism is also 'symbolic' even when it is 'real' and that the practical function of institutions is not adequate to explain their cultural structure. The search for cultural meaning resulted in rich descriptions of cannibal cosmologies, dramas, and rituals.

Refinements of natural selection theory in the 1960s and 1970s had encouraged evolutionary biologists to tackle what they viewed as the vile and unsavory behaviors in animals, which they depict as rape, slavery, infanticide, mate-desertion, and cannibalism (Mock, 1992). Shifting away from classical group selection (the idea that traits evolve for the good of the population or species), most biologists proposed that phenotypic traits, including behaviors, evolved because of the net benefit to the individual's inclusive fitness. One consequence was that sporadic reports of such 'unsavory' behaviors in animals could no longer be assumed to be pathological or aberrant. Even these behaviors might have evolved if they typically conferred reproductive advantages on their practitioners. Attention turned to such topics as the proximate cues underlying cannibalistic acts (in rodents), the genetic basis for cannibalism (in flour bugs), and how population and community structure might be affected by these dynamics. For evolutionary biology, as in the other sciences, cannibalism had become a more-or-less respectable topic of investigation.

Archaeologists, too, had begun to examine what they viewed as the darker side of humanity. During the 1980s, studies of violence, slavery, and cannibalism displaced an earlier focus on nationalism and gender. For some decades, archaeologists had examined human and animal bones in the American Southwest, debating whether these remains were the result of starvation cannibalism, war, or the execution of

witches. In the late 1990s, sensationalist reports of cannibalism among Ancestral Pueblo people were met with dismay by Native American communities, and the findings were discounted by some scholars who acknowledged violence but not cannibalism. In 2000, however, the discovery of a human coprolite among the disarticulated bones of seven individuals in the hearth of a pithouse at Cowboy Wash in southwestern Colorado provided confirmation of violence and cannibalism (Marlar et al., 2000). Analysis of the coprolite, as well as the residue on a cooking pot, revealed the presence of human myoglobin that could have resulted only from the cooking and consumption of human flesh. Scholars from a variety of disciplines later reexamined the contentious topic of social violence in the pre-Hispanic American Southwest and its possible association with cannibalism. A multidisciplinary group representing archaeology, ethnohistory, biological, and cultural anthropology concluded that violence had been associated with warfare and probably witchcraft and that anthropophagy had occurred in specific instances, associated with ritual, and not related to hunger (Nichols and Crown, 2008).

The turn toward representation in cultural anthropology in the late 1980s gave new life also to accounts of cannibalism as a colonizing trope and stratagem used by colonizers to justify their predatory behavior. Postcolonial studies proposed that the figure of the cannibal was created to facilitate the cultural cannibalism of colonialism, the projection of Western appetites onto the cultures they consumed. A common feature in the history of cannibal allegations was said to be the combination of its denial among ourselves and its attribution to those we wished to defame, conquer, and civilize. In an atmosphere of postcolonial guilt, denial of cannibalism was extended to denial on behalf of those we now wished to rehabilitate and acknowledge as our equals (Rawson, 1997). Cannibal denial and the debate it engendered diverted attention from a deeper analysis of collective ideas of prejudice. The assumption about the cannibal nature of others is one instance of broader attempts to discredit political rivals, unfriendly neighbors, and ethnic groups to which the accuser does not belong.

As Clifford Geertz (1995) observed, cultural anthropology draws the greater part of its vitality from the controversies that animate it. In the 1990s, the topic of cannibalism lured Obeyesekere and Sahlins into a high-profile debate concerning how we should understand non-Western cultures. Obeyesekere's (1992) reading of the logs and journals of Cook's Pacific voyages had led him to consider the way 'cannibalism' had been defined. The dialogues tapped dark fantasies, psychic structures shared by both the British and Hawaiians – "an anxiety that the Other is going to roast and eat us" (Obeyesekere, 1992: p. 635). The rhetoric about cannibalism, he said, revealed more about the relations between Europeans and Savages during early and late contact than as ethnographic statements about the nature of Savage anthropophagy (see also Sahlins, 2003). The debate settled on details of Captain Cook's death in the Hawaiian Islands in 1779. Obeyesekere (an Enlightenment universalist) proposed that the Hawaiians were pragmatic, calculating, and strategic rationalists, rather like ourselves, and their lives should be understood along psychological lines. Sahlins (a relativizing historicist) took the position that cultures exist within distinct schema, discontinuous with modern, bourgeois rationality, to be understood along French structuralist lines. In many ways, the debate seemed to revive earlier discussions about primitive mentality (Levy-Bruhl, 1923).

The echo of the structuralism of Levi-Strauss is present also in French and Spanish essays in which cannibalism was 'good to think.' Lemonnier's (1992) account of the homogeneity of forms of bodily aggression among different Anga groups led him to suggest historical and structural transformations of an older cultural system common to them all. Similarly, Vivieros de Castro (1992) proposed that the theme of divine cannibalism held by the Arawete of eastern Amazonia, central to their definition of the human condition, was part of a broader complex of Tupi-Guarani ritual anthropophagy. Tupinamba cannibalism had no privileged explication, but it was part of an interrelation of several explanatory levels.

The idea of the cannibal in the postmodern world has been emptied of the urbane irony present in Swift's 1729 satire, *A Modest Proposal*, in which he argued that the sale of small Irish children for human consumption would rescue the economy, prevent the children of the poor from being a burden to parents, put a stop to abortions, and make men as fond of their pregnant wives as they were of their mares in foal. Contemporary forms of 'immoderate eating' for profit and entertainment, in which people eat as much of a given food as possible in a set time, or consume beetles and other questionable substances, also have lost the moral and political resonance associated with earlier forms of transgressive consumption. No longer a sin of gluttony, or a platform for ridiculing bourgeois behavior, popular enactments of transgressive eating merely play with the borders of modern sensibilities. Verrips (1991) suggested that cannibal themes in recent movies, books, and music, and in foods named after body parts, provide substitutes and satisfaction for cannibalistic tendencies in societies where the Eucharist has lost its sway, substituting new forms of sublimation that harness the deep-seated desire of humans to eat their own kind.

The topic of cannibalism continues to lend itself to explanatory trends in the human sciences, as anthropologists acknowledge the dangers of framing cultural difference in historical terms, ignoring the extent to which social and cultural practices encountered in the present are themselves products of modern, social, political, cultural processes, and sensibilities. Knauft's 1985 ethnography of Gebusi social life in Papua New Guinea presented an array of all-night ceremonies, dramatic sorcery inquests, and divinations that had led to the execution of sorcery suspects and an extraordinarily high rate of homicide. Most of those executed were sorcerers, considered to be inhuman, and routinely were cooked and eaten, treated in the same manner as game animals of the forest. When Knauft returned in 1998 he found that fundamentalist Christianity, schools for children, and the marketing of garden produce had effected changes so great that indigenous rituals mostly were limited to remote villages. Sorcery beliefs still were present, but sorcery inquests were rare and cannibalism no longer occurred. The Gebusi had internalized aspects of Western subjectivity, Christian morality and conduct, emulating what they perceived to be a more modern style of life and in the

process, incorporating a modern sense of self and personhood (Knauft, 2002).

The Cannibal We Know Today

All the major positions on cannibalism discussed by theorists today were well represented by the eighteenth century (Cottom, 2001: p. 145): Expressions of doubt, the idea that aggressive cannibalism was an act intended to evoke terror, and the idea that cannibalism was a result of famine. The universality of cannibalism is also a recurrent theme. The philosophers' discussion of the cannibal as a subject of universal history preceded the notion of the cannibal within every human being, the judgment of a contributor to the *Encyclopedie* (1751–65) edited by Diderot. In the early eighteenth century, the distinction between exocannibalism and endocannibalism arose in the context of arguments assembled around the right to conquer; if anthropophagi eat only the flesh of people who died from natural causes, or who were killed by others in their midst, this was not a justification for attack. The notion that the conqueror was as savage as the savages, a proposition associated with cultural relativism, appears in Book 1 of the *Iliad*, in Plato's *Republic*, in Aristotle's *Politics*, and Montaigne's *Selected Essays*.

In the 1980s, following 50 years of theoretical engagement with the topic, and global theorizing about cannibal practices, however, anthropologists confessed to being notably uninformed about ideas and practices in specific cultural settings, eliciting a call (Brady, 1982) for philosophical house cleaning around the complexities of getting to know cannibals. The taxonomic impulse of nineteenth-century science, which produced a list of behaviors based on the common feature of ingestion, and categories emerging from our own culture and institutions, was not well suited to this task.

We began to know cannibals somewhat better when Western-educated Papua New Guineans visited their own communities and wrote about the practice. John Waiko (1970) interviewed elderly Binandere men in Northern Papua who recalled with pride their own participation in 'cannibalism, raiding, and the like.' Martin Buluna's discussion (1995) of cannibalism in the Milne Bay area provided detailed descriptions of the protocol associated with the treatment and consumption of victims, elaborate feasting and dancing.

In recent years, anthropologists have begun to speak with women and men who consumed loved ones, and in some cases the enemy dead, and to hear what they thought about the practice. Former cannibals in Amazonia and Papua New Guinea describe the legal rules about who may and may not eat the deceased person, the moral responsibilities of the living to one another and to the dead, as well as the fate of the body and the spirit of the deceased person. The myths and rituals enacted in mortuary ceremonies are said to convey sentiments of love, compassion, and grief. We are presented with a vast theater of bodily and spiritual practices that define the political and moral boundaries of social life, conjoining ideas about cannibalism, human identity, the destiny of the human spirit, the relations among people and communities, as well as between humans and animals, embedded in concepts, philosophies, and moral landscapes that once engaged the natural philosophers (Gillison, 1983; Glasse, 1963, 1967; Lemonnier, 1992; Lindenbaum, 2013; Whitfield et al., 2008). As the Wari told Beth Conklin (2001), cannibalism was not all about the eating.

Recent studies suggest another shift in contemporary sensibilities that may rescue the image of the cannibal consigned to the pathological margins of culture and to the behavior of others. Tim White (2001) has asserted that osteological and archaeological evidence provides convincing proof of prehistoric cannibalism among many different peoples, including many of our ancestors. Molecular geneticists working with an anthropologist and the Fore people in Papua New Guinea have come to the same conclusion. Evidence of selection at the human prion gene locus suggests that kurulike epidemics (the epidemic transmitted when the Fore consumed deceased kin who had died of kuru) is said to have occurred during the evolution of modern humans when cannibalism was thought to be widespread (Mead et al., 2003). Elderly Fore women known to have consumed deceased kin, but who had not developed the disease, were shown to have a distinct type of prion protein gene that had provided relative protection when they were exposed to the infectious prion. Global patterns of diversity in the same gene suggest that European populations show similar, but older, evidence of selection, resulting in survival advantage during episodes of cannibalism and prion-related epidemics in ancient human populations. White's proposition that we had all once been cannibals now seems less provocative than it did in 2001.

The alliance between contemporary anthropology and medicine could be said to occupy some of the space vacated by the departure of natural philosophy. The message that we too were cannibals can be used to reinvigorate the moral imagination, said to have been censored by the conventions and discourse of modern science. The figure of the cannibal, long used to establish difference and construct racial boundaries, can now be called on in projects to deconstruct them.

See also: Ancient Greece and Rome, History of; Natural Law; Ritual.

Bibliography

Arens, W., 1979. The Man-Eating Myth: Anthropology and Anthropophagy. Oxford University Press, New York.

Avramescu, C., 2011. An Intellectual History of Cannibalism. Princeton University Press, Princeton.

Boulay, R., 2001. Kannibals and Vahinés. Imagerie des mers du Sud. Editions de la Réunion des musées nationaux des Arts d'Afrique et d'Océanie, Paris.

Brady, I., 1982. The myth-eating man. American Anthropologist 84, 595–610.

Bucher, B., 1981. Icon and Conquest. A Structural Analysis of the Illustrations of de Bry's Great Voyages. Chicago University Press, Chicago.

Buluna, M., 1 May 1995. Having the neighbours for dinner. Eastern Star, 12–13.

Conklin, B.A., 2001. Consuming Grief. Compassionate Cannibalism in an Amazonian Society. University of Texas Press, Austin.

Cottom, D., 2001. Cannibals and Philosophers. Bodies of Enlightenment. Johns Hopkins University Press, Baltimore, Maryland.

Daston, L., Park, K., 1998. Wonders and the Order of Nature. Zone Books, New York, 1150–1750.

Geschiere, P., 1998. Rubber and cannibalism. In: Meyer, B., van Velzen, T. (Eds.), The Power of Images in a Globalizing World. Amsterdam.

Geertz, C., 1995. Culture war. New York Review of Books 42 (19), 4–6.
Gillison, G., 1983. Cannibalism among women in the Eastern Highlands of Papua New Guinea. In: Brown, P., Tuzin, D. (Eds.), The Ethnography of Cannibalism. Society of Psychological Anthropology, Washington, DC, pp. 33–50.
Glasse, R.M., 1967. Cannibalism in the kuru region of New Guinea. Transactions of the New York Academy of Science 29 (Series II), 748–754.
Hall, S., 1992. Formations of Modernity. Polity Press, Cambridge.
Healey, M., 2002. Monstrous tyrannical appetites: "& what wonderfull monsters have there now lately ben borne in Englande". In: McAvoy, L.H., Walters, T. (Eds.), Consuming Narratives. University of Wales Press, Cardiff.
Knauft, B.M., 2002. Exchanging the Past. A Rainforest World of Before and After. University of Chicago Press, Chicago.
Lemonnier, P., 1992. Attaques corporelles et cannibalisme ches Les Anga de Nouvelle-Guinée. Terrain 18, 87–94.
Lemonnier, P., 2006. Le Sabbat Des Lucioles. Sorcellerie, chamanisme et imaginaire cannibale en Nouvelle-Guinée. Editions Stock, Paris.
Levy-Bruhl, L., 1923. Primitive Mentality. London: George Allen & Unwin Ltd.
Levi-Strauss, C., 1964. Tristes Tropiques. An Anthropological Study of Primitive Societies in Brazil. Atheneum, New York.
Lindenbaum, S., 2004. Thinking about cannibalism. Annual Review of Anthropology 33, 475–498.
Lindenbaum, S., 2013. Kuru Sorcery. Disease and Danger in the New Guinea Highlands. Paradigm Publishers, Boulder, Colorado.
Marlar, R., Leonard, B., Billman, B., Lambert, P., Marlar, J., 2000. Biochemical evidence of cannibalism at a prehistoric Puebloan site in southwestern Colorado. Nature 407, 74–78.
Mead, Simon, Stumpf, M.P.H., Whitfield, J., Beck, J.A., Poulter, M., Ampbell, T., Uphill, J.B., Goldstein, D., Alpers, M., Fisher, E.M.C., Collinge, John, 2003. Balancing selection at the prion protein gene consistent with prehistoric kurulike epidemics. Science 300, 640–643.
Mock, D., 1992. Dining respectably. Science 258, 1969–1970.
Montaigne, M., 1964 (1581). Selected Essays. Houghton Mifflin, Boston.
Nichols, D.L., Crown, P.L., 2008. Social Violence in the Prehispanic American Southwest. University of Arizona Press, Tucson.
Obeyesekere, G., 1992. "British Cannibals." Contemplation of an event in the death and resurrection of James Cook, explorer. Critical Inquiry 18, 630–654.
Rawson, C., 31 October 1997. The horror, the holy horror. Times Literary Supplement, 3–5.
Sahlins, M., 2003. Artificially maintained controversies. Anthropology Today 19 (3), 3–5.
Tuzin, D., 2001. Cannibalism. International Encyclopedia of the Social & Behavioral Sciences 3, 1452–1454.
Verrips, J., 1991. 'Ik kan je wel opvreten.' En(i)ge notities over het thema in westerse samenlevingen. Etnofoor 1V (1), 19–50.
Vivieros de Castro, E., 1992. From the Enemy's Point of View. University of Chicago Press, Chicago.
Waiko, J.D., 1970. A payback murder. The green bloodbath. Journal of the Papua New Guinea Society 4 (2), 27–35.
White, T., 2001. Once were cannibals. Scientific American 285 (2), 58–65.
Whitfield, J.T., Pako, W.H., Collinge, J., Alpers, M.P., 2008. Mortuary rites of the South Fore and Kuru. Philosophical Transactions of the Royal Society. Series B, Biological Sciences 363 (1510), 3721–3724.

Capital: History of the Concept

Heinz D Kurz, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article discusses various types of capital and their roles in the social sciences. The emphasis is first on the concept of capital in economics and especially in the form of produced means of production. When one speaks of capital in this way, the problem of explaining profits is close at hand. The surplus approach of the classical economists is compared with Marx's theory of exploitation, marginal productivity theory, and Schumpeter's to trace profits back to innovations. The article then turns briefly to other types of capital, in particular 'social', 'cultural', and 'symbolic' capital, as they were investigated in sociology, especially by Bourdieu.

The term 'capital' was used in Medieval Latin as an adjective, *capitalis* (from *caput*, head), modifying the word *pars*, to designate the principal sum of a money loan. The principal sum was contrasted with the 'usury', the payment made to the lender in addition to the return of the sum lent, later called interest. This usage had become common by the thirteenth century and possibly can be traced back to the eleventh century in the first chartered towns of Europe (see Fetter, 1937). Later capital also represented, for example, a set of produced means of production, a fund invested in liquid assets, or human skills, so-called human capital. Here we focus attention first on capital as a factor of production, that is, instruments (plant and equipment, tools, machinery) and objects (raw materials, semifinished products) of production and skilled labor. Capital is a factor that can be accumulated, whereas other factors such as land and simple labor cannot. Earlier authors reckoned the means of subsistence ('bread') enabling workers to perform their tasks among the necessary inputs and thus capital. To the extent to which natural resources, such as land, have first to be brought into a form that can be used productively, the investment in these resources and the resources themselves get confounded; Karl Marx (1818–83) spoke of 'la terre-capital'. Productive capital yields its proprietor profits, a money loan yields interest, and land yields a rent. The explanation of the origins of the different types of property incomes and the determination of their levels was and still is one of the most controversial areas in economics.

In authors such as David Hume (1711–76) durable instruments of production were typically called 'stocks', a word also used to describe stores of tradable commodities and sums of money. With the inception of systematic economic analysis at the time of the physiocrats in France and the classical economists in Britain in the eighteenth century the term 'capital' gradually gained wider acceptance and was used alongside with such terms as 'stocks', 'funds', or 'advances'. François Quesnay (1694–1774) distinguished between three kinds of advances (*avances*) or capital goods, depending on their durability: yearly (raw materials, means of subsistence); long lived, i.e., surviving more then one yearly cycle of production (animals, tools); and very durable (melioration of land, investments in infrastructure). In *The Wealth of Nations* (1776) of Adam Smith (1723–90), we encounter the distinction between *circulating* capital and *fixed* capital. The former refers to capital goods advanced at the beginning of the period of production that contribute exclusively to the period's output: they 'disappear' from the scene at the same time as their value is transferred to the product. Fixed capital instead refers to capital goods that are long lived and cannot be traced on a given unit of output. In their case the idea of a material-cum-value transmigration into the product loses much of its appeal. Smith calls the former circulating, because it changes hands from producer to buyer, whereas the aged fixed capital item typically (but not necessarily) remains with the producer. David Ricardo (1772–1823), a most successful stock jobber at the London Exchange and thus familiar with annuities, in his *Principles of Political Economy and Taxation* (1817) put forward a theory of value that also covered the problem of the wear and tear of durable instruments of production and their depreciation over time. Interestingly, he and other classical authors put forward the idea that a coherent treatment of fixed capital presupposes a joint-products framework: a fixed capital item that enters production at the beginning of the production period is considered as a different commodity from the item that exits at its end as a by-product of the main product (think, e.g., of a tractor used in the production of wheat). In this way fixed capital can be reduced to circulating capital. John von Neumann (1903–57) and Piero Sraffa (1898–1983) analyzed the problem within such a framework.

In Marx's *Das Kapital* (1867) we find the distinction between *constant* and *variable capital*: the former refers to raw materials, tools, and instruments of production and represents 'dead' or 'saved-up' labor incorporated in the things mentioned, the latter refers to wage goods spent in employing living labor, which, according to Marx, is the sole creator of value. Profits in this view reflect 'unpaid labor' and thus the 'exploitation' of workers. This fact is concealed by the seeming equality of people in markets, where both sides agree to enter a contract. However, propertyless workers, who essentially own only their capacity to work, are forced to do so in order to make a living for themselves and their families. Their bargaining power is weak and therefore they have to accept long working days, low wages, and bad working conditions. Capital is to Marx essentially an asymmetric power relationship between the class of workers and that of capitalists. Labor saving technical progress, engendered by the capitalist competitive process, generates and continually refills an 'industrial reserve army of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03068-3

the unemployed', which perpetuates the weak position of workers in the struggle over the distribution of income. However, by progressively replacing living labor by dead labor (machines, etc.), self-seeking capitalists unconsciously destroy the source from which profits flow: the general rate of profits is bound to fall and with it capitalism. Marx puts forward a particular version of an invisible hand argument, as it had been developed at the time of the Scottish Enlightenment, with Smith as one of its towering figures: the unintended consequences of human actions, which may run counter to the interests of the agents.

The view that profits are the result of exploitation was challenged toward the end of the nineteenth century by the so-called marginalist authors (the concept of 'marginal productivity' had, however, already been developed by the German economist Johann Heinrich von Thünen (1783–1850)). The American economist John Bates Clark (1847–1938) and the Austrian economist Eugen von Böhm-Bawerk (1851–1914) were among the severest critics of Marx. Clark distinguished between *capital* and *capital goods*, the former being a fund of value earning its owner a return, which equals the 'marginal productivity of capital'. By this is meant the increase in output due to an infinitesimally small increase in the amount of capital employed, given the employment of all other factors of production. Profits thus reflect the relative scarcity of capital in production; they are the higher the more scarce capital is.

Böhm-Bawerk did not agree with Clark's explanation, but argued that profits (or rather interest, as he preferred to call them) are rooted essentially in two 'grounds': (1) Man exhibits a positive time preference in consumption, that is, prefers a given bundle of goods today to the same bundle tomorrow, etc. (2) More 'roundabout ways of production' are superior. To provide a well-known example of more roundabout processes: catching fish with bare hands is not very productive; jousting them with a spear is more productive; using a net and a boat is even more productive, etc. Through producing intermediate products (spear, boat, net) and then equipping workers with them labor productivity is increased. To produce first such intermediate products, that is capital goods, and not immediately consumption goods, presupposes some saving and investment. The first ground however says that people are not inclined to wait to consume, that is to save. They do so only, if they are offered an interest rate that exceeds their rate of time preference. The superiority of more roundabout processes of production allows for such an *agio*, which is interest. Since workers equipped with (more) capital goods exhibit a higher labor productivity than workers without, they have to share the increase in output with capitalists, who saved and invested and thus were able to endow them with capital goods.

As these examples already show, a main issue in capital theory is whether capital is 'productive' in the sense that it explains the existence of *profits*. Critical reviews of early profit theories are contained in Marx (1905–1910[1954]) and Böhm-Bawerk (1884[1890]). At the cost of severe simplification, the various traditions in the theory of capital and profits may be divided in two principal groups, one rooted in the surplus approach of the classical economists, the other in the demand and supply approach of the marginalist authors. Both traditions developed their arguments essentially within an analytical framework that presupposes competitive conditions and focuses attention on fully adjusted positions of the economic system, characterized by a uniform rate of profits and the corresponding set of normal prices.

The classical authors from Smith to Ricardo explained profits in terms of the *surplus product* left after the means of production used up in the course of producing given levels of output and the means of subsistence in the support of workers had been deducted from these outputs. Taking real wages as given is thus a characteristic feature of the early classical economists' approach (the level of wages was then discussed in another part of the theory, typically by taking into account, in particular, whether the society was 'improving', stagnant, or declining and whether the demand for 'hands' tended to grow more quickly than the workforce). Production was conceived of as a *circular flow* involving a strong degree of interconnectivity of the different industries of the economy. The rate of profits, expressed in material terms, is the ratio between the social surplus and social capital, i.e., two aggregates of heterogeneous commodities. A comparison of these two bundles of commodities necessitated the development of a theory of value. The classical economists tried to tackle this problem typically by first identifying an 'ultimate measure of value', which was designed to render heterogeneous commodities homogeneous. Several authors, including Ricardo and Marx, then reached the conclusion that 'labor' was the sought standard and therefore advocated some version of the labor theory of value. According to it the values of commodities are regulated by the amounts of labor needed directly and indirectly in their production. By means of this theory Marx in particular, in a first step, determined the rate of profits by assessing the labor values of the surplus product and the social capital. In a second step he then applied this rate to the capitals employed in the different lines of production in order to determine normal competitive prices, or prices of production. Ladislaus von Bortkiewicz (1868–1938) aptly called this approach 'successivist'.

Yet, as Sraffa (1898–1983) showed (Sraffa, 1960: p. 6), the successivist approach cannot generally be sustained: "the distribution of the surplus must be determined through the same mechanism and at the same time as are the prices of commodities", that is, simultaneously. The classical authors did not have at their disposal the instrument of simultaneous equations and the mathematics needed to solve them. This helps to explain why they had recourse to the labor theory of value. This landed them, in the case of Marx, in the (in)famous problem of the 'transformation' of labor values in prices of production. Commodities were produced by means of commodities and there was no way to circumnavigate simultaneous equations. Sraffa showed that a coherent formulation of the classical approach that was quite independent of the labor theory of value was possible: The rate of profits and competitive prices could be determined consistently in terms of the givens of the problem under consideration: (1) the system of production in use, characterized by the dominant methods of production employed to produce given gross outputs; and (2) the ruling real wage rate(s) (or the share of wages).

We may illustrate this in terms of the following numerical example with only two commodities, a and b. Assume that 30 units of a are produced by means of 10 units of itself and 12 units of commodity b, where the input quantities cover both

means of production and means of subsistence in the support of workers. We may write this in the following way: $(10a, 12b) \rightarrow 30a$, where $\rightarrow$ means 'produce'. Similarly, 20 units of b are produced by means of 8 units of a and 4 units of itself: $(8a, 4b) \rightarrow 20b$. The economic system as a whole apparently produces more than it uses up (or 'destroys') productively: $30a > 10a + 8a = 18a$ and $20b > 12b + 4b = 16b$. The system produces a *surplus* or *net product* of $12a$ and $4b$. In conditions of free competition this surplus product is distributed in terms of a uniform *general* rate of profits, r, on the *value* of the capitals employed in the two industries. This gives rise to the following normal price equations, where p_a designates the price of one unit of commodity a and p_b the price of one unit of commodity b:

$$(1 + r)\,(10p_a + 12p_b) \;=\; 30p_a$$

$$(1 + r)\,(8p_a + 4p_b) \;=\; 20p_b$$

If we measure all value magnitudes in terms of, say, commodity a, then this is equivalent to setting the price of commodity a equal to unity: $p_a = 1$. The two equations above thus have only two unknowns, r and p_b (the price of commodity b expressed in terms of commodity a) and can be solved. We get:

$$r \approx 0.4876 \;\text{ and }\; p_b \approx 0.847$$

In the numerical example given the rate of profits is just under 49% and one unit of b is worth about 0.85 units of commodity a.

The alternative marginalist explanation, which we have already encountered above, traces profits back to the productivity enhancing effect of the use of capital goods. It consists essentially of a generalization of the principle of intensive diminishing returns in agriculture indiscriminately to all industries and all factors of production alike. The older marginalist authors, with the exception of Léon Walras (1834–1910), were aware of the fact that in order to be consistent with the concept of a long-period equilibrium, the capital equipment of the economy could not be conceived as a set of arbitrarily given physical quantities of heterogeneous capital goods, but had to be expressed as a *value* magnitude: its commodity composition was seen to be a part and parcel of the equilibrium solution, determined by (1) preferences, (2) the technical alternatives from which cost minimizing producers can choose, and (3) initial endowments of the economy of labor, land, and *value capital*. The formidable problem for the marginalist approach consisted in the necessity of establishing the concept of a 'quantity of capital', which could be expressed independently of the 'price of its service', the rate of profits, and whose relative scarcity then determined that rate. If such a concept could be shown to exist, profits could be explained analogously to intensive rent on homogeneous land and a theoretical edifice could be erected on the universal applicability of the principle of the relative scarcity of a factor of production, which is the marginal productivity principle.

However, doubts as to the sustainability of this concept had already surfaced at an early time, especially in Knut Wicksell (1851–1926). Wicksell understood that a collection of heterogeneous capital goods can only be expressed as a single magnitude by using prices to render the various capital goods commensurable. However, as the equations above show, prices cannot be ascertained independently of the general rate of profits. The concept of a given 'quantity of capital', whose relative scarcity is taken to hold the key to an explanation and determination of the rate of profits, therefore presupposes already the knowledge of this rate. This implies arguing in a circle. This finding had prompted some authors such as John R. Hicks (1904–89) in the late 1920s and early 1930s to abandon the long-period method and adopt instead temporary and intertemporal equilibrium methods (Garegnani, 1976). Yet it was only during the so-called Cambridge controversies in the theory of capital in the 1960s and 1970s that the marginalist concept of capital was conclusively shown to be untenable in general (see Garegnani, 1970; Kurz and Salvadori, 1995: Chapter 14). It is based, among other things, on the idea that the methods of production can be ordered according to the amount of capital per unit of labor employed. Cost minimizing producers will then adopt from the set of alternative methods that one, which at a given wage rate maximizes the rate of return on capital. The higher (lower) the wage rate the more (less) capital intensive the method chosen, because producers have an incentive to substitute more expensive factors of production by less expensive ones: labor (capital) will be replaced by capital (labor). Yet Sraffa (1960) showed that this view cannot generally be sustained: methods of production cannot be ordered monotonically according to capital intensity or roundaboutness (Böhm-Bawerk).

The marginalist view applies only in exceedingly special cases, and it is ironic to see that these are precisely those in which the labor theory of value holds. Despite these findings, the concept is still widely employed, in much of macroeconomic theory, for example, with its reliance on 'aggregate production functions'. In more recent times also temporary and intertemporal models have been criticized from the point of view of capital theory; see the contributions by Garegnani and Bertram Schefold in Kurz (2000).

Joseph Alois Schumpeter broke with basically all previous analyses of capital and capitalism, but praised Marx for his deep insights into the *endogenous* dynamism of the capitalist system. He conceived of economic development as processes of transition between states of 'circular flows', that is, stationary economic conditions, triggered by 'new combinations', that is, innovations. He insisted that in the circular flow there are neither entrepreneurs nor capital nor profits nor interest. There are only static firms and the only incomes paid are wages of labor and rents of land. It is the entrepreneur who awakens the system from its slumber in terms of new goods, new methods of production, new forms of organization, new markets for his products and raw materials, etc. The innovator typically enters the scene without the means to realize his projects. He needs credit, which he gets from the banker, who creates additional money. Credit is the lever of a withdrawal of resources on behalf of the innovator. In an economy in which all productive resources are fully employed, an increase in the overall demand in monetary terms leads to an increase in the prices of productive resources. The inflation induced by credit expansion can be compared to a tax on static firms and channels into the hands of the entrepreneur the resources he needs. It is only with credit that 'capital', 'profits', and 'interest' enter the stage. According to Schumpeter capital consists exclusively of the

purchasing power handed over to innovators. It is not itself productive, but a precondition to be met in order for productivity increases to obtain. Neither produced means of production nor accumulated savings are capital, as received theory contends, capital is rather credit given to new producers. Schumpeter therefore speaks of his "purchasing power theory of capital" (1912: p. 255). Capital, profits, and interest exist only in a dynamic economy, not in the circular flow – a view with which he challenged basically all existing theories, classical, Marxist, marginalist (neoclassical), or Austrian. Both Böhm-Bawerk and decades later Schumpeter's student Paul Anthony Samuelson rejected this view.

With the process of globalization going on, there is a tendency toward an internationalization of capital and the worldwide equalization of the rate of profits on capital. An analysis of the factors affecting this rate is an important task in contemporary economic analysis.

From an early time onward, economists and social scientists in general drew the attention to the role of institutions, norms, customs, and habits for the quality of life and the performance of socioeconomic systems. Smith, for example, insisted: "Commerce and manufactures … can seldom flourish in any state in which there is not a certain degree of confidence in the justice of government" (WN V.iii.7). Justice, fairness, and the absence of corruption were seen to engender trust among individuals, thereby reducing transaction costs and increasing productivity. Early soundings of the more recent concept of 'social capital' referring to social cohesion and interconnectedness can be found in writers such as James Madison, a coauthor of *The Federalist Papers*, and Alexis de Toqueville in his study of the American democracy.

In more recent times the concept of social capital was developed to encompass all the benefits, economic and other, derived from the quality of relationships and forms of cooperation between individuals and groups. The concept became particularly prominent in sociology with the contributions of Pierre Bourdieu (1930–2002), especially Bourdieu (1984), who also introduced other categories of capital into the analysis, especially 'cultural capital' and 'symbolic capital'. Bourdieu defines social capital as the amount of resources controlled by an individual or a group as a consequence of it being a part of a network of more or less institutionalized relationships of mutual acquaintance and recognition. The social capital of an individual (or a group) reflected in its social positioning gives it access to resources above and beyond its economic capital. Each individual is seen to be describable in terms of the amounts of the various kinds of capital at its disposal and the networks it belongs to; a description in terms of social class membership is not good enough. The distribution of the various kinds of capital among individuals (and groups) is the key to power relations in social life shaping the latter's dynamics. Bourdieu shows how the complex of social relations and the activation of each individual's array of capitals in everyday practice are able to engender or reproduce inequality. In Bourdieu's view 'symbolic capital' is a particularly important source of power, because it is typically not perceived as such. By activating their symbolic capital (prestige, honor, etc.) – in the extreme as 'symbolic violence' – its holders exert an influence on the thinking, perception, and action of dominated agents who are inculcated with a belief in the legitimacy of the going social order. Bourdieu succeeds in showing how the ruling classes manage to preserve their social position across time, a fact that stands in striking contrast to the view that postindustrial societies are characterized by a high degree of social mobility, equal opportunities of its members, and the like.

The concept of social capital was further developed and popularized by Robert Putnam's *Bowling Alone* (2000). In economics it implies parting company with the analytical workhorse of much of modern economic analysis: the *homo oeconomicus*. The World Bank in the late 1990s initiated a research project designed to assess the value added gained by taking the concept of social capital into account in economic analysis and policy.

See also: Bourdieu, Pierre (1930–2002); Cultural Capital and Education; Marx, Karl (1818–83); Power in Society; Schumpeter, Joseph A. (1883–1950); Smith, Adam (1723–90); Social Capital.

Bibliography

Böhm-Bawerk, E. von, 1884[1890]. Kapital und Kapitalzins. Erste Abteilung: Geschichte und Kritik der Kapitalzins-Theorien. Wagner, Innsbruck. Translated as Capital and Interest, Macmillan, London.

Bourdieu, P., 1984. Distinction: A Social Critique of the Judgment of Taste. English translation of the French original. Harvard University Press, Harvard.

Fetter, F.A., 1937/1977. Reformulation of the concepts of capital and income in economics and accounting. Reprinted in Capital, Interest, & Rent.

Garegnani, P., 1970. Heterogeneous capital, the production function and the theory of distribution. Review of Economic Studies 37, 291–325.

Garegnani, P., 1976. On a change in the notion of equilibrium in recent work on value and distribution. In: Brown, M., Sato, K., Zarembka, P. (Eds.), Essays in Modern Capital Theory. Amsterdam, North-Holland.

Kurz, H.D. (Ed.), 2000. Critical Essays on Piero Sraffa's Legacy in Economics. Cambridge University Press, Cambridge.

Kurz, H.D., Salvadori, N., 1995. Theory of Production. A Long-Period Analysis. Cambridge University Press, Cambridge, Melbourne and New York.

Marx, K., 1867/1954. Das Kapital. Kritik der Politischen Ökonomie, Bd. I. In: Capital, English edition, vol. I. Progress Publishers, Moscow.

Marx, K., 1905–1910[1954]. Theorien über den Mehrwert. In: Kautsky, Karl (Ed.), Theories of Surplus Value, English edition, 3 vols. Progress Publishers, Moscow.

Putnam, R., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon and Schuster, New York.

Schumpeter, J.A., 1912. Theorie der wirtschaftlichen Entwicklung. Duncker & Humblot, Leipzig. Translation of an abridged version of the second German edition as Theory of Economic Development, Harvard University Press, Cambridge, MA.

Smith, A., 1776/1976. An inquiry into the nature and causes of the wealth of nations. In: Campbell, R.H., Skinner, A.S. (Eds.), The Glasgow Edition of the Works and Correspondence of Adam Smith, vols. I and II. Oxford University Press, Glasgow.

Sraffa, P., 1960. Production of Commodities by Means of Commodities. Prelude to a Critique of Economic Theory. Cambridge University Press, Cambridge.

Capitalism

Ivan T Berend, UCLA, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstarct

The term, capitalism appeared in the mid-nineteenth century. It has various interpretations, Marxist and post-Marxist, as well as anti-Marxist concepts. Chronologically one may differentiate among merchant, industrial, and financialized globalized capitalism. In the interwar decades, alternative regimes – Fascist and Communist economic systems – emerged. After World War II consumer and welfare capitalism were introduced. In the history of capitalism regulated and laissez-faire system rotated. In the late twentieth century, an Asian model and state capitalism became especially successful in emerging market systems.

Capitalism is an economic system based on free markets, credit creation, private enterprise, and ownership. The term and concept of capitalism as an exploitative socioeconomic system was initially introduced by Marxist social sciences. The spread of the term, however, was strengthened and its concept changed by the Western critiques of Marxism and socialism. Although rarely used, sociology and economics accepted the term and historiography broadly discusses its half-a-millennium development and various stages from commercial to globalized capitalism. The term was rarely used in the Western world during the Cold War ideological confrontation of East and West, but reappeared again, especially after the 2008 financial crisis.

Concepts and Theories

The term, derived from *capital* (root in Latin *caput*, 'head') and related to *capitalist*, appeared in the mid-late nineteenth century in the English, German, French, and other European languages. The word *capital* had been used from the seventeenth century onward in the economic sense, meaning 'fund of money,' and later 'principal' (as opposed to *interest*). The term *capitalist* appeared in mid-late eighteenth century French, English, and American texts. The first recorded occurrence of the form *capitalism* in English is 'The sense of capitalism sobered and dignified Paul de Florac' in Thackeray's *The Newcomes* II (1854). The spread of the term *capitalism* in common language use paralleled its use in Marxist socioeconomic theory, the appearance of s*ocialist–communist* economies in various countries, and the expansion of Cold War rhetoric based on the opposition of the two socioeconomic systems. The term and concept of *capitalism* as a socioeconomic system was basically introduced by the first theorists of Marxist social sciences. Although rarely used, sociology and economics accepted the term and historiography broadly discussed the transition from feudalism to *capitalism*, and the various stages of capitalist development from its early commercial phase via industrial and financial *capitalism*, to its late twentieth-century characteristics. The capitalist world system is often discussed according to the pattern of advanced core and dependent, backward periphery. Development studies and economics analyze late twentieth-century globalization and the increased role of transnational companies in a more interrelated, but at the same time even more polarized capitalist world economic system.

Karl Marx, who often used the terms capital, capitalist, and capitalistic, did not use the term capitalism as a noun in such basic writings as the *Communist Manifesto* and *Das Kapital*. The term appeared only later in his correspondence on the development of Russian capitalism during the late 1870s. The concept finally gained scholarly interpretation and became widely used by Marxist theory and the socialist movement in the last third of the nineteenth century. The concept was broadly used in historiography and sociology, but much less frequently in non-Marxist economics, to describe a phase of history or a set of ideas and mentality. During the twentieth century, capitalism became a popular term, especially to contrast the private market system with socialism. Marxist theorists interpreted the term capitalism as a socioeconomic system, a 'mode of production,' in which 'capital is not a thing but a social production relation belonging to a definite historical formation of society' (Marx, 1952; Chap. 48, English trans.; first published in 1867). Capitalism is a class society basically consisting of two antagonistic classes: bourgeoisie and proletariat. The system historically came about through 'primitive accumulation,' the separation of peasantry from the means of production, and the creation of a free proletariat deprived of property, whose only saleable commodity was its labor power. The ruling class, the bourgeoisie, owns the means of production, buys the labor power, and controls all economic transactions and production processes. In the capitalistic reproduction process, capitalists who own the capital buy various types of commodities (raw material, machinery, labor power, etc.) for producing new commodities for sale. In this reproduction process money is transformed into commodities, then again into money (M–C–M). The capitalist, however, is motivated by making profit and wants to get more money back at the end of the reproduction process (M–C–M′). The source of profit is one of the commodities – the labor power, which creates a 'surplus value.' The value of the goods produced by the workers is higher than the value of the labor power bought for a wage, which covers the expenses of the reproduction of labor power (including the reproduction of the class itself). Capitalists expropriate this 'surplus,' thus exploiting the proletariat. In capitalism, therefore, social production is in sharp contrast with individual expropriation. As a consequence, wealth is

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62003-2

accumulated in the hands of capitalists, while the proletariat is kept in poverty. Capitalism, however, creates its own 'grave diggers' – the ever-increasing proletariat who, uniting together in the advanced West, rises up to destroy the capitalist system through a proletarian revolution. Bourgeois expropriators become expropriated and a just-communist society without classes and private ownership is established.

Post-Marxian Marxism revised this theory. Eduard Bernstein, the leading theorist of the powerful German Social Democratic Party, rejected the concept of proletarian revolution in his landmark article series, *Probleme des Socialismus* (1896). He argued that capitalism has developed in a direction different from the one indicated by Marx. Poverty has not grown but diminished, and welfare reforms have been accepted peacefully. Bernstein and the Western Social Democratic movement maintained that the proletariat could realize its interests and goals by organized mass movement and parliamentary reforms. Another leading German socialist, Rudolf Hilferding, in his *Das Finanzkapital* (1909), added that modern economy created 'consciously regulated social relations.' Later he concluded in the theory of 'organized capitalism,' thus, a state-controlled welfare system.

Revolutionary socialism emerging during the first decades of twentieth century redefined contemporary capitalism as the higher or the 'last' stage of capitalism, otherwise called *imperialism*. In her *The Accumulation of Capital*, Rosa Luxemburg (1951, English trans.; first published in 1913) maintained that C would not be able to expand its markets, recruit new labor force, and mobilize new resources without a 'third person,' i.e., noncapitalist (independent farmer) sectors of the advanced countries, and, most of all, the agricultural economy of the colonies and peripheries. The Russian Vladimir I. Lenin defined imperialism as a new, expansionist stage of capitalism, which was also characterized by the domination of 'monopolies,' where the competition of small individual capitalists is replaced by the deadly competition of monopolies, not only locally but also in the international arena. According to this view, capital exports became a new instrument for exploiting backward peripheral countries. Imperialism has led to the fight for the redistribution of the colonies and spheres of interest, and this unavoidably generates wars, and leads to the proletarian revolution of the backward world. This version of Marxism was the leading ideology of the Bolshevik Revolution in Russia, a revolutionary wave in Central and Eastern Europe after World War I, and the communist takeover in a series of Central and Eastern European, Asian, and African countries after World War II.

Non-Marxist scholarship did not build on the Marxist concept and interpretation of capitalism, but the term itself became commonly used in scholarship. Many twentieth-century sociologists and economists used the term capitalism to mean a society and economy based on free market and entrepreneurial interest, and accepted these characteristics as the essence of capitalism, in contrast to socialism. Critiques of Marxism, therefore, contributed just as much to the spread of the term and concept of capitalism. Because of his work on the role of Protestant ethic in the development of capitalism, Max Weber played an instrumental role in making capitalism a commonly used term in scholarship. Ludwig von Mises and his disciple Friedrich von Hayek frequently used the term capitalism when launching the most radical attacks against Marxist theory and socialist economy and defending an undisturbed free market. Joseph Schumpeter's book *Capitalism, Socialism, and Democracy* (1943) emphasized the changing evolutionary character of capitalism because of its entrepreneurial interest that 'incessantly revolutionarizes the economic structure from within' and with a 'creative destruction' destroys the old, outdated and creates the new. Joyce Appleby also characterizes capitalism as a 'relentless revolution' (2010), with a built-in ability of renewal.

The broadly accepted non-Marxist meaning of capitalism is rather different from the Marxist interpretation. According to the most widespread definition, capitalism is a complex socioeconomic system where the bulk of the means of production is in private hands guaranteed by modern property rights and laws. An 'invisible hand' (Adam Smith), the operation of market, regulates the entire economy, production, distribution, and labor. The market operates according to the play of supply and demand, which influence price fluctuations and automatically regulate investments and output. As the Say-law phrased it: each supply creates its demand. The market mechanism, therefore, is a self-regulating system. Capitalist entrepreneurs are acting individually and freely in a rational organization and control themselves with double entry bookkeeping, motivated by gain and increase of profit. Capitalism, in this sense, is in sharp contrast with both its rigidly regulated feudal predecessor and twentieth-century rival socialism. Hereditary, almost unchangeable social status, basically non-commercialized production for local consumption, strict communal and guild regulations, and the lack of free private property, labor, and market make economic relations in feudalism rigid opposed to capitalism; and nonmarket economy based on state ownership and central planning contrast socialism to capitalism. While capitalist society in its early stage has been, and in contemporary backward countries is characterized by extreme social polarization and income inequity, capitalism granted equal rights and, in principle, freedom of choice later in the advanced countries. From the 1870s on, but mostly from the 1950s, the industrial working class began shrinking, and the 'white-collar' layers, including the middle class, gradually expanded and became the most dominant layer of society. From the last third of the twentieth century, a two-thirds to three-quarters white-collar and middle class majority dominated advanced societies.

Early Stages of Development

Capitalism, broadly accepted as a phase of history, is dominating the last half-a-millennium. It has also become conventional that capitalism itself had various phases and stages. Its antecedents go back to ancient history. Barter and exchange of one thing for another, as Adam Smith, the main advocate and ideologue of market capitalism argued, are a part of human nature. The pockets of private market economy and its institutions, especially in urban settlements, flourished in medieval Europe. Joseph Schumpeter maintains that history of capitalism that goes back to the period when credit was created. Moneylenders and urban merchants, mostly in North Italy were its main representatives. Its continuous development,

however, began mostly from the sixteenth century. Historiography produced a vast literature on the transition from feudalism to capitalism. These writings maintain that declining productivity, a demographic crisis, and the scarcity of peasant labor generated the lowering of rents, a decrease of labor services, and an increase in peasant mobility and freedom in Western Europe. In most of these countries, consequently, serfdom was gradually loosened and disappeared by the sixteenth century. Commercialization of agriculture played a central role in the development of capitalism. The flourishing textile industry in Britain with an excess production over consumption led to the enlargement of productive capacities in an unheard scale and made England the 'workshop of the world.' This stage between the sixteenth and eighteenth centuries is often called 'commercial capitalism' in history. The system was based on private ownership of warehouses and ships and buying and selling goods all over the world. The merchants also lent out money at interest and established connections with production. They bought and distributed raw materials for peasants who worked at home in the traditional way (putting-out system), and gradually subordinated the rural cottage industry. It led to a proto-industrialization, the development of a decentralized putting-out system coordinated by merchants and later the establishment of (nonmechanized) factories, which introduced division of labor in the production procedure.

The emergence of capitalism, as various historians have discussed it, also gained incentive from the discovery of the 'New World.' The unlimited inflow of gold and silver from Latin America led to the accumulation of wealth. Although Spain was the main beneficiary at the beginning, most of the wealth got into the hands of Dutch and English merchants. "It is not the importation of gold and silver – stated Adam Smith – that the discovery of America has enriched Europe … [It] made a most essential [change] … by opening a new inexhaustible market to all the commodities of Europe, it gave occasion to a new division of labor." The new markets and emerging Atlantic trade enriched the commercial powers of Northwest Europe and made them the core of the emerging modern world system from the sixteenth century on. This period saw the rise of modern world trade, characterized by the trade of mass consumption goods of food and textile, instead of the medieval trade of luxury goods from the Near East and India. While the core countries of Western Europe sold processed industrial products, they bought unprocessed raw materials and agricultural produces from the peripheral countries of the world system: Latin America and Eastern Europe.

The rise of capitalism was strongly influenced by cultural-ideological and institutional factors. The Protestant reformation of the sixteenth century created a new work ethic and lifestyle, which was a prime mover of capitalist development. Hard work and thrift became virtues, which was in sharp contrast to parasitic noble values and lifestyle, the so-called *Hidalgo* (Spanish) and *Szlachta* (Polish) attitude. Meanwhile, the new moral code legitimized capitalist inequity: character was idealized, richness considered to be the triumph of will and work, poverty became a moral failing. In the absolute state between the sixteenth and eighteenth centuries, centralizing policy creating uniform monetary systems, legal codes, unified and defended home markets without internal tariffs, also played an important role. The isolation of the domestic market by limiting imports was accompanied by a tremendous effort to build up a large colonial empire and a system of various privileges and stimuli for exports. The *mercantilist* state itself became a promoter of the development of market, industry, trade, and transportation by the creation of road systems, canals, navies, and even state-owned industries. In the early modern centuries, important institutions such as modern property right were introduced that created a stable environment for rising capitalism. Meanwhile, in the most advanced countries in Northern Europe new consumption attitude emerged, the 'industrious revolution,' and a new scientific method, experiment and applied sciences, an 'industrial Enlightenment' gained ground. All these developments historically paved the way for dramatic capital accumulation, enlargement of markets, and technological experiments, new trends that led to the rise of modern capitalism.

Industrial Capitalism and Beyond

In the last quarter of the eighteenth century, Adam Smith published his powerful theory and ideology of the free market system, the *Wealth of Nations*, which became instrumental in the development of laissez-faire capitalism. Smith argued that from the selfish activity of the entrepreneurs, driven by profit motivation, the entire society profited, and that the individual actions, because of the self-regulating impact of the market, created an economic harmony and well-being for the nation. Smith established the economic thought and popular thinking on capitalist market economy for two centuries. Political thinkers and economist of latecomer countries, however, attacked British *laissez-faire* ideology. The American Hamilton and Germans Johann-Gottlieb Fichte and Friedrich List argued against *laissez-faire* in favor of protective state interventionism.

The end of the eighteenth century is considered to be a turning point in the development of capitalism. As a consequence of the British Industrial Revolution, anew stage of more mature, as social scientists and historians called it, 'industrial capitalism' emerged and dominated the nineteenth century. Starting in Britain, the industrialization process transformed the core countries of the capitalist world system into industrialized economies with a high, sustained economic growth. Industrial capitalism, in the first period of its existence, institutionalized long working days, and exploited female and child labor ruthlessly in the factory system. Living conditions of the workers in crowded and polluted industrial cities, as described by Charles Dickens and Friedrich Engels, were extremely poor. Capitalism, however, reached its dynamic period of development. A banking revolution in Belgium and France, then, most of all, in Germany, introduced the Credit Mobilér type, then the German type of investment banks with huge industrial portfolios. These banks played an instrumental role in industrialization. The relatively small, family-managed factories enlarged, and most of them became joint stock companies with professional management. From the mid-nineteenth century on, a European railroad system created a united European transportation network and an integrated marketplace. In this period virtually the entire globe became part of the capitalist world system. The core countries offered an unlimited market for food and raw materials. Britain and a handful of Western

European countries absorbed nearly 70% of the world's food and raw material exports. New institutions strengthened international capitalism, such as a free trade system and the gold standard – a fiscal underpinning of the free trade system making the national currencies convertible, which were gradually introduced from the 1860s and 1870s.

The expansion of capitalism around the turn of the nineteenth–twentieth centuries led to considerable changes in the system. Werner Sombart introduced the term of 'late capitalism,' and Rudolf Hilferding named this new period 'finance capitalism.' The concept of 'Imperialism' (Hobson) also appeared and became a leading concept of revolutionary Marxism. Some of the socialist theorists, however, such as Karl Kautsky, considered imperialism as a 'policy' of capitalism, similar to long workdays and child labor in its early phase. He maintained that capitalism could exist without colonies by introducing new policies. Twentieth-century capitalism was characterized by the challenge of major wars, worldwide economic depression, and a significant slowing down of economic growth. During World War I, the German war economy produced a 'war socialist' system that offered a more efficient regime even in peacetime. (Keynes) Alternative regimes such as fascist economic dirigisme emerged and challenged traditional capitalism. More importantly, capitalism had to face the challenge of socialism, first as a movement, then as a rival socioeconomic system. A great part of the twentieth century was characterized by the struggle of the two socioeconomic systems. Capitalism reacted by a highly flexible adjustment to the new challenges.

During the wars and depression, the development of state capitalism with a strong and effective state interventionism, a return to protectionism, inflationary financing, and even state investments and planning, helped the system to respond to the new challenge. Large state-owned sectors emerged in Italy, Spain, and Poland. National planning was introduced in Germany and Hungary to cope with economic decline and unemployment, and also preparing for war. Self-regulating, *laissez-faire* market capitalism was replaced everywhere by a regulated market system. Adam Smith's theory was replaced by John Maynard Keynes' concept of state interventionism to create additional demand.

The Welfare State, Mixed Economy, and Back to Globalized Laissez-Faire

The socialist challenge of capitalism required an effective response. The first chapter of this rivalry was opened in the late nineteenth century. Chancellor Otto von Bismarck, in his struggle with the strong German socialist movement, turned toward the introduction of welfare institutions. Pension system, health care, and other welfare institutions were introduced for workers to take the wind out from the sail of the emerging socialist movement. For these developments, the Great Depression of the early 1930s brought about a turning point. The Swedish Social Democratic government, elected in 1932, started building a welfare state. President Franklin D. Roosevelt also initiated major social legislation and the introduction of the social security system. The comprehensive system of the welfare state, however, emerged only after World War II in Western Europe. Capitalism, in contrast with its nineteenth-century features, initiated widespread social legislation. Paid vacation, short workweek, nationwide pension systems, free health care, education, maternity leave, cheap housing, and many other institutions were established. Citizen rights were reevaluated and the right to work, i.e., a full employment policy, became dominant. Inequity, at least in the middle layers of the society, markedly diminished in the advanced capitalist countries of Europe. The period after World War II also saw the spread of mixed economies. Almost all of the advanced Western European countries, led by France and Britain, nationalized entire sectors of their economies and established new state-owned firms in various industries. In some countries, such as France or Japan, even government planning was used to guide modernization of the economy. Capitalism underwent a spectacular transformation.

In his article on *The Instability of Capitalism* (1928), the leading Austrian-American economist, Joseph Schumpeter, prophesied the transformation of the system and the fading of the original meaning of capitalism. He maintained that capitalism is inherently stable and able to recover from great crises, but generates socio-intellectual effects discordant with its spirit and institutions. As a consequence, 'Capitalism is … in so obvious process of transformation into something else … [that] it will be merely matter of taste and terminology to call it Socialism or not.' After World War II, and also toward the end of the twentieth century, capitalism adjusted to the new historical situation and markedly transformed but preserved entrepreneurial interest, market flexibility, efficiency, competitiveness, and reached its highest prosperity and fastest growth rate ever. The Western core countries of the world system successfully adjusted to a new technological communication revolution and became the leader of new technologies and the rise of postindustrial society.

During the long period of postwar prosperity, a part of the former Asian and South European peripheries of the capitalist world system reached a higher than average growth and became an integrated and equal part of the core countries. After a long period of rivalry, capitalism emerged victorious while the parallel world system of centrally planned state socialism, which conquered one-third of the world and influenced an even greater part by its policy after World War II, collapsed in 1989–91. The former Soviet Bloc countries introduced radical market reforms, thus reestablishing capitalism. During the last third of the twentieth century, *laissez-faire* capitalism became the leading Western model again. Neo-liberal economics triumphed and challenged mixed economy and welfare state. Capitalism in late twentieth century reached a new stage in its history. The main trend of this new age is globalization. National boundaries and national economies rapidly lost their importance. Multinational or transnational companies penetrated previously independent economies. Foreign direct investment became instrumental all over the world, including the advanced core countries. Toward the end of the 1990s nearly one-third of American exports and two-thirds of imports were intrafirm deliveries. About one-third of French, Dutch, and British industrial output was produced and roughly 25–40% of their research and development expenditure was financed by affiliates of transnational companies. In Ireland, and former state-socialist Hungary, foreign affiliates produced

two-thirds of industrial output and financed roughly 70% of research and development expenditures. In several former Soviet Bloc countries the entire banking system was in the hands of Western banks.

From the late twentieth century, the advanced capitalist countries were deindustrialized and financialized. The financial and real estate sectors became larger than manufacturing. Capitalism became deregulated and high-risk 'casino' type gambling led to ballooning real estate bubbles in several countries. That period also became the rise of consumer-, or later even hyperconsumer societies. Government and household spending surpassed available incomes and led to extreme indebtedness of households and governments in several advanced countries, and former peripheral ones that virtually did catch-up and reached the advanced level. This trend led to the explosion of the 2007–08 liquidity crisis that generated a financial panic and recession in several countries. In Europe, the common currency of the European Union, the euro was also endangered. The Great Recession of the early twenty-first century led to a new wave of regulations in the United States and Europe, mostly coordinated by international organizations such as the G-20 organization of the economic powers.

While globalization, in some cases, contributed to a successful catching-up process with the core, it also preserved and even strengthened the core–periphery inequity. The gap between advanced core and the peripheries was growing considerably: intercountry income spread was 10:1 in 1913; 26:1 in 1950, but it increased to 40:1 at the end of the twentieth century. The so-called *Gini index* of inequality reflects broadening income disparities in the United States, and also in Europe. A newly globalized but even more polarized capitalist world system, dominated by an expanding Western core, was emerging at the turn of the millennium.

Globalization, however, had also positive influence for several Asian countries and an Asian model of modernization and capitalism emerged. After Japan and the small 'Asian Tigers,' China, India, Vietnam introduced an effective state-capitalist system and a strong export-led industrialization. Their unparalleled economic growth started transforming world capitalism and the global power relation.

See also: Capitalism: Global; Economic History; Field Theory; Industrialization, Typologies and History of; Polanyi, Karl (1886–1964); Social History; Weberian Social Thought, History of.

Bibliography

Appleby, J., 2010. The Relentless Revolution. A History of Capitalism. W.W. Norton, New York.
Aston, T.H., Philpin, C.H.E., 1985. The Brenner Debate. Cambridge University Press, Cambridge, UK.
Baldwin, P., 1990. The Politics of Solidarity. Class Bases of the European Welfare State 1875–1975. Cambridge University Press, Cambridge, UK.
Berend, I.T., 2006. An Economic History of Twentieth Century Europe. Economic Regimes from Laissez-Faire to Globalization. Cambridge University Press, Cambridge, UK.
Berend, I.T., 2012. An Economic History of Nineteenth Century Europe. Diversity and Industrialization. Cambridge University Press, Cambridge, UK.
De Vries, J., 2008. The Industrious Revolution: Consumer Behavior and the Household Economy, 1650 to the Present. Cambridge University Press, New York.
Hacker, L.M., 1940. The Triumph of American Capitalism. Simon & Schuster, New York.
Heimann, E., 1964. History of Economic Doctrines. Oxford University Press, New York.
Hobson, J.A., 1902. Imperialism. A Study. Nisbet & Co, London.
Kautsky, K., 1914. Der Imperialismus. Neue Zeit, Berlin.
List, F., 1904. The National System of Political Economy. Green and Co., London.
Luxemburg, R., 1951. The Accumulation of Capital. Routledge, London.
Marx, K., 1952. Capital. Encyclopedia Britannica, Chicago.
Mokyr, J., 2008. The Gift of Athena. Historical Origins of the Knowledge Economy. Princeton University Press, Princeton.
Schumpeter, J., 1943. Capitalism, Socialism, and Democracy. Allen & Unwin, London.
Smith, A., 1976. An Inquiry into the Nature and Causes of the Wealth of Nations. University of Chicago Press, Chicago.
Sombart, W., 1921–8. Der moderne Kapitalismus, vols. 1–3. Duncker and Humbolt, Munich, Germany.
Sweezy, P.M., 1942. The Theory of Capitalist Development. Principles of Marxian Political Economy. Oxford University Press, New York.
Vries de J.
Wallerstein, I., 1974. The Modern World System: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century. Academic Press, New York.
Weber, M., 1992. The Protestant Ethic and the Spirit of Capitalism. Routledge, London.

Capitalism: Global

Leslie Sklair, London School of Economics, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Capitalist globalization is defined as the phase of capitalism in which the transnational capitalist class (TCC) gradually takes a leading role in the global system. The TCC consists of four overlapping fractions, dominated by the corporate (those who own and/or control transnational corporations and their local affiliates). The analysis highlights economic, political, and culture–ideology issues. The crises of class polarization and ecological unsustainability, endemic to capitalist globalization, are discussed in the light of resistance to global capitalism and alternative forms of globalization, based on noncapitalist principles of social justice.

Introduction

While the idea of international capitalism has been part of political economy and related disciplines for more than a century, the concept of global capitalism (or capitalist globalization) is of more recent origin. Despite the fact that many scholars have used the terms interchangeably, one consequence of the popularity of the concept of globalization in the 1990s and the first decade of the new millennium was the increasing force of the argument for a firm distinction between the international and the global. International capitalism has generally been conceptualized in state-centrist terms, focused mainly on how national capitalists based in competing national economies and working through national companies operated across borders. The distinctive concept of global capitalism takes its departure from the idea of a global economy dominated by those who own and control globalizing corporations, and those who serve their interests. In the 1960s, a distinctive concept of global capitalism began to emerge from the transitional idea of capitalist world economy through several attempts to establish a genuinely global conception of capitalism not grounded in national economies and societies. In the years between then and now, periodic booms and recessions in various parts of the world have increasingly been analyzed in terms of the global capitalist system. The global financial crisis that came to a head in 2008 and spread throughout the world, with a few notable exceptions, is the latest in a long history of capitalist crises, recessions, and depressions. While global capitalism appears to have weathered the 2008 storm and has (so far) averted collapse at the systemic level, the severity of the crisis has been compared by many commentators to the great world depression of the 1930s. As in the 1930s, different regions of the world, different parts of countries, and even different neighborhoods in cities and towns have experienced the events of the last few years in very different ways, further confirmation of an increasing class polarization crisis in the era of capitalist globalization.

Researchers on global capitalism have focused on three interrelated phenomena, increasingly significant since the 1960s. These are, first, the ways in which transnational corporations (TNCs) have facilitated the globalization of capital and the production of goods and services (Sklair, 2002; Dicken, 2007); second, the rise of new global forms of organization of the capitalist class (Sklair, 2001; Carroll, 2010); and third, transformations in the global scope of TNCs that own and control the mass media, notably television channels, transnational advertising agencies, and the digital economy (Herman and McChesney, 1997; Fuchs & Mosco, 2012). Attention has been especially focused on the role of TNCs, large and not so large, in promoting global brand consumer goods and the emergence of a global culture and ideology of capitalist consumerism. Theory and research on each of these three phenomena roughly coincide with attention to the economic, political, and culture–ideology spheres of global capitalism. Resistance to capitalist globalization in theory and practice also revolves around these three areas of contention.

Global Capitalism as an Economic System

It is no coincidence that interest in a global capitalist system in contrast to competing national capitalisms has increased perceptibly since the 1960s. The context of theoretical and empirical interest in competing national capitalisms was (and for many still is) the history of colonialism and imperialism. This is overlaid with several versions of the theory that capitalist states could more or less successfully plan their own economic futures (see, for example, Keynesianism, regulation theory, and the developmental state). The concept of international capitalism, therefore, refers to a system of interacting and competing national capitalist economies, in which national elites of various types use 'their' big business (and businesses) to further 'their' national interests around the world. As direct imperialism and colonialism came to an end and as increasing numbers of very large TNCs began to emerge in the 1960s, attention began to shift decisively from national to global capitalism.

The dependency approach to development and underdevelopment of Gunder Frank and the related world-systems approach of Wallerstein both highlighted the systemic nature of capitalism as a worldwide phenomenon over several centuries. While these theoretical innovations can be said to have prepared the ground for it, neither entirely succeeded in establishing a coherent concept of global capitalism for what came to be termed the 'age of development' from the 1950s onward. This was due to ambivalence over their units of analysis and insufficient focus on the role of the major corporations in development in general. For example, the analysis of core, semiperiphery, and periphery in the world-systems approach is

based on national economies, as is the related concept of commodity chains. While both these approaches have generated large numbers of valuable studies on specific sets of economic exchanges and their productive and exploitative consequences (Babones and Chase-Dunn, 2012), the simple assertion that these take place within a world economy or global economy does not suffice for an analysis of global capitalism.

Theories of global capitalism, in the sense used here, take off from the proposition that capitalism entered a new, global phase in the second half of the twentieth century (Ross and Trachte, 1990; Robinson, 2004). By the end of the century, the largest TNCs had assets and annual sales far in excess of the gross national products of most of the countries in the world and this trend has persisted into the new millennium. In 2010, according to the World Bank, only 100 countries out of a total of around 200 for which there were even approximate numbers had gross domestic products (GDPs) of more than US$20 billion. By contrast, the Fortune Global 500 (FG500) list of the biggest corporations by turnover reported in 2011 that around 484 of them had annual sales greater than US$20 billion. This comparison, however, underestimates the economic scale of major corporations compared with sovereign states, as most TNC revenues are counted as part of the GDP of some countries. The largest economic entities in the world in terms of GDP are the United States, China, Japan, Germany, France, United Kingdom, Brazil, Italy, India, Canada, and the Russian Federation, all in excess of a trillion US dollars (the United States in 2010 was still the leader with a GDP of US$14.6 trillion, over twice the size of its two main (Asian) rivals). However, the top 50 corporations in the FG500, led by Walmart, the largest company in the world with US$420 billion, all had revenues in excess of US$100 billion. Therefore, in an important sense, such well-known names as Walmart, Shell, Exxon Mobil, BP, Toyota, Sony, Pepsico, Coca Cola, Apple, the huge Japanese trading houses, and many other corporations most people have never heard of like the Chinese state giants Sinopec and State Grid, have more economic power at their disposal than the majority of the countries in the world. These figures indicate the gigantism of TNCs relative to the economies of most countries.

Not only have TNCs grown enormously in size since the 1950s, but their global reach has expanded dramatically. Many TNCs regularly earn more than half their revenues outside the countries in which they are legally domiciled. This is the case not only for TNCs from countries with relatively small domestic markets (for example, Switzerland, Sweden, Canada, and Australia), but also for those legally domiciled in the United States and Japan. While most of the biggest corporations are still headquartered in what used to be termed the First World, around 100 companies originating in what was called the Third World or that part of it known as the newly industrializing countries have been numbered among the FG500 over the last few decades. This group has included the state-owned oil companies of Brazil, China, India, Mexico, Taiwan, and Venezuela (owned by the state but increasingly run like private corporations); banks in Brazil and China; and the Korean manufacturing and trading conglomerates (chaebol), some of which have attained global brand-name status (for example, Hyundai and Samsung).

Writers who are skeptical that capitalism is a global system argue that because most major TNCs are legally domiciled in the United States, Japan, and Europe, and because they trade and invest mainly between themselves, capitalism is still best analyzed in terms of national corporations. For such writers, the global economy is a myth and, consequently, there is no global capitalist system as such. Against this conclusion, theorists of the global capitalist system argue that an increasing number of corporations operating outside their countries of origin see themselves as developing global strategies of various types, as is obvious from the contents of their annual reports, other corporate publications and how they represent themselves in the mass media. While all parts of all economies are clearly not equally globalizing, an increasing volume of empirical research indicates that the production and marketing processes of most major industries are being de-territorialized from their countries of origin and that these processes are being driven by the TNCs. The central issue for economic globalization is the extent to which TNCs domiciled in the United States, Japan, and European and other countries can be more fruitfully conceptualized as expressing the national interest of their countries of origin (the globoskeptic argument) or expressing what can be conceptualized as the private interests of those who own and control them, aggregated as the interests of global capitalism. Even if historical patterns of TNC development have differed from country to country and region to region, it does not logically follow that TNCs and those who own and control them express any type of 'national' interest or national character. Industrial sector, for example, is often more important for explaining patterns of corporate operation than national origins.

The formal ownership of capital and the corporations has been transformed since the 1960s. The distribution of ownership of share capital has increased throughout the world by means of greater participation of the general population in stock markets and the indirect investments that hundreds of millions of people have through their pension funds and other forms of savings in the richer countries and for wealthy elites in less rich countries (although these people still constitute a small minority in most communities). The increase in the numbers of those actively involved in speculating in stock markets around the world has led some to argue that economic globalization has created a popular capitalism, although others argue the more elitist thesis that the real drivers of the global capitalist system are the managers of unit trusts and pension funds. Formal ownership, of course, does not necessarily mean effective control over capital and the resources of TNCs.

The focus of attention of those interested in generating a model of the future of global capitalist growth in the last quarter of the twentieth century was on the so-called newly industrializing countries (notably export-processing enclaves like Singapore, Hong Kong, South Korea, and Taiwan). However, in the opening decade of the twenty-first century, the focus switched to former member countries of the Third and Second (postcommunist) Worlds, heralding the rise of the BRICS (Brazil, Russia, India, China, and more recently South Africa) – all posting impressive economic growth performances while the traditional First World powerhouses of capitalism (North America, Western Europe, and Japan) languished in the throes of enormous government and private citizen debt, austerity policies, deep cuts in public investment, rising unemployment, low or negative economic growth, and financial crisis. It is too early to say, definitively, how robust

the BRICS model is, but the boosterism through which it is represented in the business media may be read as one indication of the increasing desperation of those charged with the management of global capitalism.

It is now commonplace to trace the present crisis of global capitalism to the financial crisis that broke in 2008. The globalization of the international financial and trading system can be fruitfully analyzed in terms of the progressive weakening of the nation-state, widespread deregulation, especially of financial markets, and the growing recognition that the major institutions of global capitalism, notably TNCs and globalizing financial and trading organizations, are setting the agenda for these weakened nation-states, not excluding the most powerful. The latest crisis of capitalism triggered by the global financial upheavals that saw the demise of some major institutions all over the world and the bailout by governments with taxpayers' money of others to avoid system meltdown has been vividly portrayed in the Oscar-winning documentary film 'Inside Job' (Ferguson, 2010). This can be understood within the context of what has been termed the culture–ideology of consumerism (C–I of C). The C–I of C is the consequence, on the cultural and ideological levels, of a crisis created by overproduction in the global capitalist system, made possible by the efficiency gains of the electronic revolution, and a chronic lack of real spending power – as opposed to borrowing via credit cards and bank loans – among consumers globally (Sklair, 2002). It is no coincidence that this latest and probably most severe crisis since the 1930s seems to have come to light due to problems with subprime mortgage lending in the United States, Europe, and elsewhere. Home ownership or its equivalent in secure leases is a central institution of capitalist globalization insofar as it stimulates ever-expanding demand for all the goods and services that keeps a capitalist economy growing. A bad situation – mortgage holders defaulting on their debts – was made significantly worse by the gradual disclosures that financial institutions all over the world had been speculating on these junk assets (derivatives). Thus, when the revenue streams from mortgage lending began to dry up due to foreclosures and home 'owners' walking away from their negative equity burdens, such financial instruments became worthless and once-revered financial institutions collapsed and some had to be bailed out by governments with tax-payers' money. The effects that this had on the 'real economy' (a term that the ideologues of capitalist globalization had successfully kept largely under wraps until now) are well known – banks are less willing to lend, businesses vulnerable to cash flow problems suffer, and jobs are lost. The only solutions on offer are platitudes that the economy needs to be stimulated, and governments become desperate to put cash in people's pockets to get them spending again while introducing ever more drastic austerity measures, debt increases, and so on. The general public is thrown into confusion and while criticisms of the irresponsibility of bankers and unregulated markets become a flood, especially in the context of a culture of huge bankers' bonuses, no alternatives to the so-called free market system are taken seriously by those charged with managing the global economy. Theory and research on this issue has, not surprisingly, led to an increased interest in the politics of global capitalism.

Global Capitalism as a Political System

The politics of global capitalism is debated intensely inside and outside the social sciences. Since the disintegration of the Soviet empire from the late 1980s, the struggle between capitalism and communism has been largely replaced by the struggle between the advocates of the neoliberal form of capitalist globalization and its opponents. Many theorists have discussed these issues within the triadic framework of states, TNCs, and international economic institutions. From this perspective, the global capitalist system is dominated by the relations between the major states and state systems (United States, North American Free Trade Agreement, the European Union, and Japan); the major corporations (FG500); and the World Bank, the International Monetary Fund, and the World Trade Organization (WTO) (supplemented in some versions by other international bodies, major regional institutions, and so on).

The idea of a transnational ruling class has been suggested by several theorists of the emergence of a global class structure in what was identified as the Gramscian turn in International Relations; for example, in the work of Gill (1990) on the Trilateral Commission, where he identifies a "developing transnational capitalist class fraction." Sklair (2001) proposes a more explicit concept of a transnational capitalist class, and it plays a central role in his theory of capitalist globalization. Here, the transnational capitalist class is the characteristic institutional form of transnational political practices in the global capitalist system (paralleling the role of TNCs in the economic sphere and consumerism in the culture–ideology sphere). In this formulation, the transnational capitalist class is analytically divided into four main fractions: (1) those who own and/or control major TNCs and their local affiliates (corporate fraction); (2) globalizing politicians and officials (political fraction); (3) globalizing professionals (technical fraction); and (4) merchants, media, and advertising (consumerist fraction).

A transnational capitalist class might be transnational in at least three senses. First, its members would tend to have outward-oriented global rather than inward-oriented national perspectives on a variety of issues; for example, support for free trade and neoliberal economic and social policies. Second, they would tend to be people from many countries, more and more of whom begin to consider themselves as citizens of the world as well as of their places of birth and residence (these might differ). And third, they would tend to share similar lifestyles, particularly patterns of higher education (in cosmopolitan business schools) and consumption of luxury goods and services. Theory and research to support this argument are as yet in a very embryonic phase. Nevertheless, and despite real geographical and business sector conflicts, the whole of the transnational capitalist class shares a fundamental interest in the continued accumulation of private profit on a global scale.

There exists a very wide variety of approaches to the politics of global capitalism. While Marxist, anti-Marxist, Marx-inspired scholars and those who have little interest in Marx mostly see capitalism as a global system, others tend to conceptualize globalization in much wider terms and minimize the importance of global capitalism as an explanatory variable (see the

collection edited by Lechner and Boli, 2011). Consequently, it is important to distinguish those approaches that assume that globalization simply means capitalist globalization from those that make distinctions between generic globalization, capitalist globalization, and alternative (usually noncapitalist) forms of globalization.

Global Capitalism as a Culture–ideology System

The third distinctive aspect of capitalism as a global system is the worldwide diffusion and increasingly concentrated ownership and control of the electronic mass media, particularly television (TV) and the Internet. In the last decades of the twentieth century, the number of TV sets per capita grew rapidly in developing countries (from fewer than 10 per 1000 population in 1970 to 145 per 1000 in 1995 (according to the United Nations Educational, Scientific, and Cultural Organization)) and growth has accelerated since then (it is estimated that there are now around 500 million TV sets in China). Ownership of personal computers also increased rapidly into the new millennium, ranging from over 800 per 1000 population in Switzerland and the United States to less than 2 per 10 000 in Rwanda. The average today appears to be around 60 per 1000 of the population, with the top 71 out of 168 countries having at least 100+ computers per 1000 of the population. Internet access has also penetrated deeply all over the world, even in towns and cities in very poor countries. By 2007, Greenland had almost total coverage (922 per thousand), the top 43 countries had 500 per thousand, and the top 134 countries had 100 per thousand (all data from www.nationmaster.com via Central Intelligence Agency Factbooks). Many researchers use such data to argue that a globalizing effect due to the mass media is taking place all over the world. Ownership and control of television, including satellite and cable systems, and associated media, including newspaper, magazine and book publishing, films, music distribution, and a wide variety of other, notably Internet-based, media are concentrated in relatively few very large TNCs. While the predominance of US-based corporations is being challenged by Japan-, Europe-, and Australia-based groups globally, and even by corporations like the Brazil-based media empire of TV Globo, there is wide consensus that – with few exceptions, notably Internet-based activity challenging the hegemony of global capitalism – the culture–ideology of consumerism drives all these media in the pursuit of private profit.

The culture–ideology of consumerism prioritizes the centrality of consumption and consumerism in contemporary capitalism – a system in which consumption expectations and aspirations are inexorably increased without necessarily ensuring the income to buy. The extent to which economic and environmental constraints on the private accumulation of capital undermine the global capitalist project in general and its culture–ideology of consumerism in particular is a central issue for theory and research on the capitalist global system. Nevertheless, most scholars studying these issues of global culture do so not in terms of capitalism per se but in terms of the potential impact of global culture on national and local cultures and identities, notably in the context of gender, ethnic, and lifestyle differences.

Crises of Global Capitalism

While Marxist and Marx-inspired theories of the inevitability of a fatal economic crisis of capitalism appear to have lost most of their adherents, at least two related but logically distinct contradictions of global capitalism have been identified. The first is the simultaneous creation of increasing poverty and increasing wealth within and between societies for rapidly expanding minorities all over the world, and increasing economic insecurity for the squeezed middle class (class polarization crisis). While Marx's emiseration thesis did not predict the enormous increases in the numbers of the very wealthy that have occurred in the last few decades even in some nominally poor societies, the relative position of the very poor all over the world appears to be worsening. The second contradiction of global capitalism is the unsustainability of the system in terms of environment and natural resources (ecological crisis). Globalizing corporations (in some cases rather more clearly than national governments) recognize the class polarization crisis, but largely in marketing terms. In most communities around the world the absolute numbers of people who are becoming global consumers have been increasing rapidly over recent decades. However, it is also true that in some communities the absolute numbers of the destitute and near-destitute are also increasing, sometimes alongside the new rich consumers. The best available empirical evidence comes from the United Nations Development Program annual Human Development Reports (undp.org/en/) demonstrating that the gaps between the rich and the poor have widened since the 1980s in many parts of the world. The very poor cannot usually buy the goods and services that global capitalists sell. While there is a long way to go before consumer demand inside the rich First World is entirely satisfied, the gap between the rich and the poor all over the world is not a welcome news for TNCs. In addition to the profits lost when poor people who want to buy goods and services do not have the money or even the credit to do so, the increasing visibility of the new rich and the new poor in an age of constant global media exposure directly challenges claims that everyone eventually benefits from capitalist globalization.

The ecological crisis is also directly connected with consumerism, encapsulated in the political and ideological struggles over the concept of sustainable development. Since the 1980s, most major TNCs have developed detailed policies in response to the challenges of environmental harm and declining stocks of resources essential for the maintenance of global capitalist consumerism. However, the persistence of problems of pollution, health risks, environmental degradation, and waste management intrinsic to the system suggests that it will be difficult to avoid fatal ecological crisis in the long term under the conditions of capitalist globalization. In this context, the attempt by TNCs and their supporters in international bureaucracies, the professions, government (globalizing politicians and officials at the national, international, and transnational levels), and the mass media to capture the idea of sustainable development and to reconcile it with capitalist globalization is worth further study (Sklair, 2001: Chapter 7).

This issue is inextricably bound up with the struggle between those who campaign for free access to information against those who deal in suppression and the dissemination of

disinformation, as recently exposed in the Wikileaks affair. While it has been on the agenda for at least the past 100 years, the issue of corporate spin and corporate lobbying has been receiving increased critical public and scholarly attention in the last decade in this context (Dinan and Miller, 2007).

The magazine, *New Internationalist*, reported in 2009 that 12 488 unique registered lobbyists actively lobbied the US Congress and Senate in the period January–July 2010; 15 000 Brussels-based lobbyists sought to influence European legislation, and 2600 special interest groups have a permanent office in Brussels. The lobbying industry employs around 14 000 people in Britain. In the United States, where lobbyists are more strictly monitored (although not necessarily regulated), between 1998 and 2010, over two billion dollars were spent by the pharmaceutical/health products industry, and over one billion dollars were spent on lobbying on behalf each of insurance, electrical utilities, business associations, computers/Internet, and oil and gas interests. These are large sums of money, even by US corporate standards, and it is significant that none of the biggest corporations in the lobbying industry has much of a public profile. While the industry leader, Burson Marsteller, has had some exposure due to its lobbying on behalf of repressive regimes (for example, Nigeria and Argentina in the 1970s) and corporate polluters (representing Union Carbide after Bhopal), the other major players have very low public profiles. For example, Hill & Knowlton (part of the global communications giant WPP), Fishbourne Hedges, Weber Shandwick, APCO, and Fleishmann Hillard, despite the fact that they all have substantial transnational portfolios, lobbying governments and the public on behalf of their corporate clients, are hardly known at all outside the industry. Nevertheless, it is probable that most of what we see and hear in the corporate mass media about global capitalism originates from lobbying companies and the public relations industry. Lobbyists also channel funds from corporations to politicians and political parties and other types of political entities in support of the election campaigns of friendly candidates, those who were identified above as the globalizing politicians and officials of the political fraction of the transnational capitalist class. Unfriendly candidates and officials are also targeted with promises of campaign funds and encouraged to become more corporate friendly.

Resistance to Global Capitalism

Global capitalism is often seen in terms of impersonal forces (notably market forces and free trade) wreaking havoc on the lives of ordinary and defenseless people and communities. It is not coincidental that interest in economic globalization has been accompanied by an upsurge in what has come to be known as New Social Movements (NSM) research. NSM theorists, despite their substantial differences, argue that the traditional response of the labor movement to global capitalism, based on class politics, has failed. In its place, a new analysis based on identity politics (notably of gender, sexuality, ethnicity, age, community, belief systems, and lifestyles) has been developed, directed toward resistance to sexism, racism, environmental damage, war-mongering, capitalist exploitation, and other forms of human rights abuses.

The globalization of identity politics involves the establishment of global networks of people with similar identities and interests outside the control of international, state, and local authorities. There is a substantial volume of research and documentation on such developments in the women's, peace, indigenous peoples, and environmental movements, some of it in direct response to perceived TNC malpractices. This provides a series of research-rich connections for scholars influenced by postmodernist and global capitalist theories.

Serious challenges to global capitalism in the economic sphere have also come from those who 'think global and act local.' This normally involves disrupting the capacity of TNCs and global financial institutions to accumulate what are widely perceived to be excessive private profits at the expense of their workforces, their consumers, and the communities that are affected by their activities. An important aspect of global capitalism is the increasing dispersal of the manufacturing process into many discrete phases carried out in many different places, popularized by Dicken (2007) as 'global shift.' Being no longer so dependent on the production of one factory and one workforce gives capital a distinct advantage, particularly against the strike weapon that once gave tremendous negative power to the working class. Global production chains can be disrupted by strategically planned stoppages, but this is generally more of an inconvenience than a real weapon of labor against capital. The global division of labor builds flexibility into the system so that not only can capital migrate anywhere in the world to find the cheapest reliable productive sources of labor but also few groups of workers can any longer decisively 'hold capital to ransom' (in the common slogan of the capitalist-controlled media) by withdrawing their labor. At the level of the production process, globalizing capital has a substantial advantage over labor. In this respect, the global organization of the TNCs and allied institutions like the World Bank and the WTO have generally proved to be too powerful for local labor and community organizations. Nevertheless, global capitalists, if one is to believe their own propaganda, are continuously beset by opposition, boycott, legal challenge, and moral outrage from the consumers of their products, concerned citizens, and by disruptions from their workers. There are also many ways to be ambivalent or hostile about the C-I of C, some of which the Green Movement has successfully exploited.

The issue of democracy is central to the prospects for global capitalism and the struggle against it. The rule of law, freedom of association and expression, and freely contested elections, as minimum conditions and however imperfectly sustained, are as necessary in the long run for mass market-based global consumerist capitalism as they are for alternative social systems to thrive. Most research on the democratic issues raised by capitalism, naturally, continues to be state-centrist, focusing largely on how it works within specific countries. Nevertheless, the growing influence of globalization in the social sciences appears to be encouraging more scholars to consider capitalism in the global as well as the local and/or national context and the prospects, however fragile at present, to effect changes in the system on a transnational and/or global basis as well as at the local and national levels. This is evidenced by the significant growth in research on transnational social movements, notably those that seek alternatives to the global

capitalist system (see Smith and Wiest, 2012). Resistance at the level of ideology is also flourishing, in particular contemporary Marxist analyses and critiques of the capitalist control of the media of mass communications (Fuchs and Mosco, 2012).

Alternatives to Global Capitalism

The crisis of global capitalism that began with the financial turmoil of 2008 unsurprisingly led to intensified debates around alternatives to capitalist globalization, not only in groups on the political extremes who have always been looking for alternatives to capitalism, but also to some extent by those in the more mainstream areas of political life all over the world. The antiglobalization movement that filled the media at the so-called battle of Seattle in 1999 and continued to gather pace during the first decade of the twenty-first century is slowly giving way to a rather more mature understanding that capitalist globalization is not the only possible form of globalization. Counterhegemonic globalization has been conceptualized in many ways, most usually in terms of a partial or complete rejection of capitalist globalization (Evans, 2012). Seeking alternatives to capitalist globalization, thus, does not mean rejecting generic globalization per se. It has been argued that capitalist globalization has undermined the emancipatory potential of generic globalization in terms of the electronic and postcolonial revolutions, the creation of transnational social spaces, and new forms of cosmopolitanism – and there are even signs that, despite its often problematic past history, ideas around socialist globalization are being revived (Sklair, 2009). Despite the daily barrage of propaganda from the corporate-controlled mass media and the message from most political parties in capitalist democracies that there is no alternative to global capitalism, serious research and practical programmes on realistic alternatives to it are underway. These range from radical rejections of global capitalism to progressive reforms of the existing system, from the traditional Marxist Left to the multifaceted World Social Forums, now with an organization hub in the United States.

The widespread appeal of the 'Occupy' movements became evident in September 2011 when, under the slogan of 'We are the 99%' the Occupy Wall Street campaigners appeared in Zuccotti Park in Manhattan. By October 2011, according to media reports, Occupy protests had taken place or were ongoing in almost 100 cities in more than 80 countries. In the United States alone, Occupy groups were reported to be active in over 600 communities. Unlike many previous anti-globalization protests, the Occupy movements have clearly targeted global capitalism as the problem. While actual occupations (notably in New York and in London) were forcibly ended, lively debates on the Internet and social networking media suggest that the process of movement building against capitalist globalization has sunk viable roots in many communities around the world. While there is little evidence that processes such as the 'Arab Spring' that emerged in 2011 contained any significant anticapitalist element, the extent to which they might eventually become and/or contribute to anticapitalist movements remains an open question. Certainly, phenomena like the mass mobilization of protesters in Tahrir Square in Cairo and in other places in North Africa and the Middle East inspired Occupy protestors everywhere. Other, less dramatic, initiatives, like the movement to create Transition Towns all over the world based on the principle of zero carbon economy and society, and various types of producer–consumer cooperatives, some under the aegis of the transnational cooperative movement and some not, have been given a considerable boost by extreme economic hardship brought on by the global financial crisis. The most interesting feature of these movements is that many of the participants appreciate not only the traditional values of localism, for example, sourcing food supplies as locally as possible, but also the possibilities of wider, even transnational linkages with other like-minded groups all over the world. The peasant-based movement, Via Campesina, provides some useful clues as to how such linkages might create nonexploitative mutually beneficial opportunities for sustainable standards of living, for example, in anticorporate struggles over food sovereignty. For those who participate in such movements and millions like them all over the world, alternatives to global capitalism are already emerging.

See also: Capitalism; Economic Growth; Global History: Universal and World; Globalization and World Culture; Globalization: Geographical Aspects; International Business: Geographic Aspects; International Business; International Trade: Economic Integration; Multinational Corporations; World Systems Theory.

Bibliography

Babones, S., Chase-Dunn, C. (Eds.), 2012. Routledge Handbook of World-systems Analysis. Routledge, London and New York.

Carroll, W., 2010. The Making of a Transnational Capitalist Class: Corporate Power in the 21st Century. Zed Books, London.

Dicken, P., 2007. Global Shift: Mapping the Changing Contours of the World Economy, fifth ed. Sage, London.

Dinan, W., Miller, D. (Eds.), 2007. Thinker, Faker, Spinner, Spy: Corporate PR and the Assault on Democracy. Pluto Press, London.

Evans, P., 2012. Counter-hegemonic Globalization. In Ritzer, G. (Ed.), Wiley-Blackwell Encyclopedia of Sociology. Wiley-Blackwell, Oxford and Malden, MA, pp. 309–315.

Ferguson, C., 2010. Inside Job (Documentary Film). Representational Pictures, Sony Pictures Classic.

Fuchs, C. and Mosco, V. (Eds.) 2012. Marx is back: the importance of marxist theory and research for critical communication studies today. TripleC–Open Access Journal for a Global Sustainable Information Society (2), 127–632. <http://www.triple-c.at/index.php/tripleC/issue/view/25>.

Herman, E., McChesney, R., 1997. The Global Media: The New Missionaries of Corporate Capitalism. Cassell, London.

Lechner, F., Boli, J. (Eds.), 2011. The Globalization Reader, fourth ed. Blackwell, Malden, MA.

Robinson, W., 2004. A Theory of Global Capitalism. The Johns Hokins University Press, Baltimore, MD.

Ross, R., Trachte, K., 1990. Global Capitalism: The New Leviathan. State University of New York Press, Albany, NY.

Sklair, L., 2001. The Transnational Capitalist Class. Blackwell, Malden, MA.

Sklair, L., 2002. Globalization: Capitalism and Its Alternatives, third ed. Oxford University Press, Oxford.

Sklair, L., 2009. The emancipatory potential of generic globalization. Globalizations, 523–537. 6/4 (December).

Smith, J., Wiest, D., 2012. Social Movements in the World-system: The Politics of Crisis and Transformation. Russell Sage Foundation, New York.

Capitalism: The History of the Concept

Jürgen Kocka, Free University of Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Capitalism' is a controversial concept. It emerged as a key concept of social criticism and as a concept for socioeconomic and historical analysis, in Western languages in the second half of the nineteenth century. The contributions by Marx and Hilferding, Sombart and Max Weber, Veblen and Keynes, Schumpeter and Braudel were particularly important. The concept played a role in controversial politico-ideological debates. Since the 1990s, a limited revival of the concept can be observed, but it is used more by social scientists and in the humanities than by economists. In conclusion, a working definition of 'capitalism' is suggested.

Origins: A Concept of Difference

'Capitalism' is a controversial concept. Many scholars avoid it, because they find it too ideological, value-loaded, vague, or broad. On the other hand, 'capitalism' has been a central concept for many important analyses and debates in academia and the general public. In the present time, it is increasingly used again, although, more by social scientists and in the humanities than by economists. If one wants to use it for scholarly purposes, one must define it and should know its history.

The concept 'capitalism' is a neologism, which emerged in Western languages in the second half of the nineteenth century. It is much younger than the historical reality it denotes and also younger than its major theoretical justifications. Neither Adam Smith and his contemporaries nor John Stuart Mill and the classical political economists used the concept. Certainly, 'capital' and 'capitalist' (as a noun and an adjective) were older and frequently used in European languages at least since the sixteenth and the eighteenth century, respectively. But the noun 'capitalism' appeared only after the European revolutions of 1848. The French socialist Louis Blanc (1811–82) used it in 1850 and defined it critically as "appropriation of capital by some to the exclusion of others." In 1851, Pierre Joseph Proudhon (1809–65) castigated landed property on the Parisian house market as a 'fortress of capitalism' when demanding public regulation against exorbitant rents and speculation. The same year, the liberal economist Frédéric Bastiat (1801–50) also quoted the term. In 1867, a French encyclopedia, the *Grand dictionnaire universel du XIXe siècle*, defined 'capitalism' as 'power of capital or capitalists.'

In German the concept was not pioneered by Karl Marx (1818–83) and Friedrich Engels (1820–95) who in the 1850s and 1860s wrote much about the 'capitalist mode of production' and 'capitalist accumulation,' but used the noun 'capitalism' only later and rather marginally. It was an economist with state socialist sympathies, Johann Karl Rodbertus (1805–75) who in 1869 wrote: "Capitalism has become a social system." In 1870, a liberal-conservative professor of economics, Albert Eberhard Friedrich Schäffle (1831–1903), published a voluminous monograph *Capitalism and Socialism*. He extensively dealt with the conflicts between capital and wage labor, advocated state-sponsored reforms in order to mitigate those conflicts, and defined 'capitalism' as a "national and international organism of production under the leadership of 'entrepreneurial' capitalists competing for highest profit." "The Socialists are correct," he added, "when they declare the present economy to be characterized by the 'capitalist mode of production,' i.e., by the hegemony of 'capitalism'." In 1872, the socialist Wilhelm Liebknecht (1826–1900) thundered against capitalism as a 'juggernaut on the battlefields of industry.' In 1876, a much read encyclopedia (Meyers Universal-Lexikon) mentioned 'capitalism' for the first time, while its 1896 edition carried a whole entry on the term defining it as the "capitalist mode of production in contrast to the socialist and collectivist one" (Hilger, 1982: 443f.; Kocka, 2013: 6–9; Bastiat, 1851: 407, 415; Rodbertus-Jagetzow, 1869: XVf.; Schäffle, 1870: 116; Silberner and Silberner, 1940).

In Great Britain, the concept was not unknown already in 1855, but then reluctantly introduced since the 1880s, especially in Fabian circles. John A. Hobson (1858–1940) published a book on *The Evolution of Modern Capitalism* in 1894 in which he concentrated on the rise of the factory system. The Encyclopedia Britannica mentioned the concept first in its 1910–11 edition (still in Capitalism) and carried a whole entry on the term in 1922 defining capitalism as "a system in which the means of production were owned by private proprietors who employed managers and workers for production." The history of the concept in the United States paralleled that of Great Britain, though there is evidence that the term was known to radical working-class circles before journalists and scholars adopted it. Here, Thorstein Veblen (1857–1929) was one of the first who used it in his *The Instinct of Workmanship* of 1914. He followed European authors when stressing that capitalism was much older than industrialization, growing out of the handicraft system, of trade and of financial businesses in the fifteenth and sixteenth century. But he emphasized that "its highest development comes with the advanced stages of the machine technology and is manifestly conditioned by the latter" (Pendennis, 1855: 75; Brick, 2006: 23–33; Veblen, 1914: 282–83, 302; Williams, 1976: 42–44; Merrill 1990, 470–473).

Individualized property rights; commodification on markets for goods, labor, land, and capital; the price mechanism and competition; investment, capital, and profit; the distinction between power holding proprietors and dependent property-less wage workers, tensions between capital and labor, rising inequality; the factory system; and industrialized

production – these were, in varying combinations, major characteristics of the concept 'capitalism' as it emerged until World War I. It was mostly used to denote an economic practice or an economic system, frequently with special attention to its social and cultural consequences.

To judge from German examples, the concept 'capitalism' emerged as a *concept of difference*. It was used to highlight, usually in a critical way, certain features of the then dominant or emerging socioeconomic reality *in contrast* to what this reality was thought to have been in previous ('precapitalist') times, and to what it might become under socialism in the future. It is remarkable how closely discourses about capitalism and socialism were interrelated. It is equally striking how much the capitalist elements of the contemporary reality – or the capitalist period in which one lived – were seen as something modern. In other words, many observers did not only discuss capitalism as they observed it, but they also placed it on an imaginary time line. The contrast with a selectively commemorated and frequently idealized past, as well as the contrast with an imagined and sometimes utopian future was constitutive for the emergence of the concept.

Critique and Analysis

While all authors saw capitalism in its developed form as a modern phenomenon, the concept was also used for historical research. It was applied to medieval and ancient Europe, i.e., to historical constellations in which capitalism was much less developed than in modern times or in the present. Werner Sombart (1902), Max Weber (1893–1908), Giuseppe Salvioli (1906), Henri Pirenne (1914), William Cunningham (1916), Richard Henry Tawney (1922), Henri Sée (1928), and many others participated in the frequently controversial exploration of distant pasts by using the modern category 'capitalism.' The findings of the Belgian economic historian, Henri Pirenne (1862–1935), have stood the test of the time best, while others like Sombart (1863–1941) exaggerated the static, noncapitalist, even anticapitalist spirit of the medieval period, in which, as recent research has shown, capitalist forms of behavior were not altogether alien to both long-distance merchants and market-related producers, in different regions of Europe – even before the spread of capitalist principles accelerated between the sixteenth and the eighteenth century (in long-distance trade, banking and finance, agriculture, and proto-industrialization) and industrialization got under way (since the eighteenth century).

It is more than obvious that 'capitalism' originated as a *polemical*, *critical* concept. Of course, there were those who defended capitalism on the basis of its economic strength, as a creator of wealth and welfare as well as a motor of modernization and a force of liberation from traditional bonds and hierarchies. However, among conservatives as well as on the political left, capitalism was seen as an irresistible force of erosion: custom was replaced by contract, *Gemeinschaft* by *Gesellschaft*, the traditional by the modern, social ties by the market. On the right, anticapitalism frequently went hand in hand with antiliberalism and antisemitism, particularly since the Great Depression of the 1870s. But the socialist critique of capitalism was the most powerful one. On the one hand, it attacked the exploitation of labor by capital, increasing social inequality, suppression at the work place, and the lack of a fair deal in the distribution of product and wealth. On the other hand, it predicted the decline of capitalism due to its internal contradictions and its eventual replacement by something new, namely socialism.

Because the concept 'capitalism' was frequently used for critical and polemical purposes, and because its definition varied between single authors a lot, many scholars stayed in distance to it or avoided it altogether. The economist Richard Passow (1880–1949) documented the lively discussion about the different meanings of the concept in the scholarly literature from the late-nineteenth century to the 1920s (1918/1927). But as it has happened so often with key concepts in the social sciences – 'bureaucracy,' 'democracy,' and 'modernization' can serve as examples – the value-loaded, mostly critical and even polemical use of 'capitalism' did not prevent, but sometimes even helped to turn it into a fruitful descriptive and analytical tool for scholarly purposes. Marx and Weber are two cases in point.

Marx did not only criticize the 'capitalist mode of production' and the 'bourgeois society' based on it. Together with Engels, he developed the building blocks of a critical theory of capitalism. Impressed by the victorious industrial capitalism of his time, he emphasized: (1) the market and its increasingly 'compulsive' laws; (2) the basic role of work and labor as well as the asymmetric relations and built-in tensions between capitalist employers and property-less wage workers to be transformed into the class conflict between bourgeoisie and proletariat; (3) investment, profit, accumulation, and endless expansion of the capitalist economy, which would spread over the world and transform social relations and dominant cultures, before being destroyed by its internal contradictions and being replaced by something new – socialism – in a revolutionary way. Marx' theories and visions have been sharply criticized. At the same time, they have influenced not only social movements, political parties, and historical developments, but also scholarly discourses and research more than the work by anybody else, ever since. Among his intellectual followers, Rudolf Hilferding (1877–1941) deserves particular attention who published *Finance Capital* in 1910. In his book, Hilferding transformed the Marxist analysis of capitalism in a way that it could deal with the rise of the modern corporation and the partial separation between ownership and control at its top, with the increasing tendencies toward concentration and centralization of capital as well as with the frequently close relation between industrial enterprises and banks. Hilferding developed the building blocks of a critical theory of 'Organized Capitalism' (his concept), which would not abandon the market as the basic regulatory mechanism, but would supplement and modify it by adding elements of collective coordination, structured negotiation, and conscious organization (Marx, 1967; Muller, 2003: 166–207; Hilferding, 1910; Winkler, 1974; Bottomore, 1985).

Classical Contributions

Max Weber (1864–1920) combined analysis and critique in another way. He conceptualized capitalism as a world-historical

phenomenon with many variations – merchant capitalism, political capitalism, spoils capitalism, among others – but he was primarily interested in *modern capitalism*, which he saw defined by the capitalist enterprise (*Unternehmung*) with its specific 'rationality,' its discipline, contractual wage work, division and coordination of labor as well as a clear separation from the households of the economic actors, including the entrepreneurs. Weber found *modern capitalism* in this sense only in the West, since about the sixteenth century. He analyzed its interdependence with important social, cultural, and political developments, particularly with legal change and the rise of the modern states in a long history of formation and coercion, negotiations, and wars. Weber emphasized that *modern capitalism* had indispensable economic presupposition, especially markets for goods, labor, and land as well as for money, credit, and capital. It also needed specific economic sentiments (*Wirtschaftsgesinnung*), which would motivate main actors to relentlessly strive and intensively work for innovation and profit, used not for consumption but for reinvestment with a perspective of long-term growth per se. Weber saw this 'capitalist spirit,' in its early phases, grounded in the ascetic religiosity of early modern Calvinists and Puritans. This 'Weber thesis' has been much criticized on empirical grounds. While Weber appreciated capitalism as a central pillar of Western modernity, he saw its built-in irrationalities, too. He doubted whether capitalism would lead to general welfare, and he feared that in the long run it might threaten human freedom (Weber, 1988, 2012; Parsons, 1928–29; Mommsen, 1997; Schluchter, 2009: 63–87).

By World War I historical and social science-based thought on capitalism had produced results, which have remained influential until the present. But over the next decades, economists, social scientists, and historians contributed major new thoughts to the discussion, which remained controversial, fluid, and stimulating.

For John Maynard Keynes (1883–1946), the concept did not play a central role, but he did not avoid it, either. In 1927, he saw "the essential characteristic of Capitalism" in its "dependence upon an intense appeal to the money-making and money-loving instincts of individuals as the main motive force of the economic machine." In some parts of Europe, but not in the United States, he observed a "latent reaction, somewhat widespread, against basing Society to the extent that we do upon fostering, encouraging, and protecting the money-motives of individuals." In contrast to Weber and Sombart who had emphasized the calculability and the rational instrumentality of capitalist behavior, he stressed the great importance of tempers, emotions, and 'animal spirits' in capitalism.

In the work of Joseph A. Schumpeter (1883–1950) who always enjoyed being outside the mainstream, 'capitalism' was a central category. He defined "that word which good economists always try to avoid: capitalism is that form of private property economy in which innovations are carried out by means of borrowed money, which in general … implies credit creation." On this basis, he assumed that capitalism began in Southern Europe around 1300. As the core of capitalism, Schumpeter famously discovered 'creative destruction.' Innovation on the basis of new combinations would lead to new methods of production and distribution, new forms of organization, new demand, and new economic opportunities, but also to the displacement or the destruction of older methods, patterns, and opportunities. There were always winners and losers. Schumpeter argued that booms and busts, ups and downs of the capitalist economy were related to the rhythms of innovation; he developed his theory of the business cycles on this basis. In the 1940s, Schumpeter was highly pessimistic as to the future of capitalism. He did not expect its failure due to internal contradictions and crises like Marx, but he observed capitalism's propensity to undermine itself: Its success would transform and weaken social and cultural forces – e.g., the middle class family, precapitalist elites, and the protection by a powerful state – which capitalism had needed and would need for its ascent and survival (Keynes, 1927: 50f.; Schumpeter, 1939: 223f.; 1976).

In his monumental *Civilisation matérielle, économie et capitalisme* (1967–1979), Fernand Braudel offered an alternative definition by sharply distinguishing 'capitalism' from 'market economy.' He saw local markets, fairs and stores, the everyday practice of exchange by producers, traders, and consumers as constitutive for market economies throughout the ages and nearly everywhere in the world, while he tried to reserve the term 'capitalism' to the businesses of a relatively narrow and exclusive superstructure of wealthy and influential merchants, bankers, ship owners, proprietors, entrepreneurs, and financial capitalists. Among them, competition played minimal, close contacts to the holders of political power – princes, bishops, magistrates, and governments – an important role. While Braudel's contraposition of 'market economy' and 'capitalism' has met many insurmountable problems, it had the merit of highlighting that oligopolistic and monopolistic practices have always been frequent in capitalism, and that a close symbiosis between capitalists and the politically powerful has been much more the rule than their clean separation. Braudel also further developed the tradition of seeing capitalism as a world-historical development. He influenced authors like Immanuel Wallerstein and Giovanni Arrighi, who in the last quarter of the twentieth century conceptualized and investigated capitalism as a global phenomenon. Wallerstein built on older theories of imperialism and *dependencia* when placing capitalism into the center of his 'world-system approach.' Arrighi stressed finance capitalism and proposed a scheme according to which capitalist *lead regions* have changed over time and moved from Northern Italy in the late-middle ages via the Netherlands in the early modern period and England since the eighteenth century to the United States (twentieth century), and perhaps China in the future (Braudel, 1967–1979; Wallerstein, 1974; Arrighi, 1994, 2007; Vries, 2012).

Political Contexts and Recent Revival

No doubt, 'capitalism' has always been more than a political catchword or instrumentalized slogan. Scholarly definitions, clarifications, and application have made the term a penetrating and far-reaching analytical instrument, ever since the nineteenth century. But the term has preserved part of its affinity to social criticism. Over the decades it has frequently,

though not at all exclusively, been used by authors who saw themselves somewhere on the left and combined analysis with criticism in varying combinations when writing about economy, society, politics, and culture. For illustration, consider economic interpretations of imperialism (e.g., Luxemburg, 1913), studies on monopoly capitalism (Baran and Sweezy, 1966), treatises on the relation between the crisis of capitalism and the rise of fascism (e.g., Poulantzas, 1974), critical comments about the internal contradictions of capitalism (Habermas, 1973; Bell, 1976; Boltanski and Chiapello, 2000), the need for reembedding capitalism in order to make it more compatible with human needs (Polanyi, 1944; Streeck, 2009), or reflections on how capitalism will cope with present-day challenges like the hypotrophy of finance capitalism, increasing social inequality, or climate change (Wallerstein et al., 2013). Analysis and criticism can reinforce each other.

In addition, the term 'capitalism' has continued to play a role in controversial political debates in the public arena, reaching far beyond academic discourses. 'Capitalism' has been a central combat term within socialist and communist rhetoric and propaganda, sometimes ambitious and pointed, frequently simplified, aggressive, and totalizing. As a consequence, the term got involved into and damaged by the political and ideological struggles of the past century. 'Capitalism' was a key word in deep-reaching conflicts between different positions on the left and various shades of the right as to the correct or desirable economic, social, and political shape of the future. It played a role in controversies between democratic speakers and the advocates of both left- and right-wing dictatorships. It has been shown how the deep tension between the United States and the Soviet Union after World War I and later on between West and East in the Cold War influenced and flavored the public use of the term 'capitalism' in the United States (Marks, 2012). The Cold War impacted politics, culture, and language in most European countries, too. Again, 'capitalism' served as a concept of difference. It was defined and applied in contrast to the Communist alternative and challenge. This sometimes led to dichotomic simplifications and distortions. The more 'capitalism' was used as a polemical catchword in politico-ideological conflicts, the less useful it appeared to many scholars. Consequently they criticized or avoided it altogether (Sombart, 1935: 195; Hayek, 1954; Hartwell, 1983; Grassby, 1999).

Since the 1990s, something like a limited revival of the concept can be observed, within general discussions as well as in the social and historical sciences (Sklansky, 2012). On the one hand, the Cold War is over. Nowadays 'capitalism' is much less a combat term in a systemic conflict of global scale than it used to be during most of the twentieth century. The concept has been freed from this burden. On the other hand, a resurgence of market liberalism if not of laissez-faire capitalism has taken place since the late 1970s, primarily on the level of neoliberal discourses, but partly also in economic and social policy (Harvey, 2007). Between World War I (at the latest) and the 1970s, powerful trends toward more political coordination of markets, more planned organization of competition, and more welfare state intervention had been on the way, at least in parts of Europe and North America. Some observers spoke of a move toward 'organized capitalism,' others argued that a transition from nineteenth century industrial capitalism to a 'mixed system' or a postcapitalist era was on the way (Brick, 2006: 54–85). In the increasingly neoliberal atmosphere and the waves of deregulation since the 1970s – at least in the United States and in some parts of Europe (including Eastern Europe since the 1990s) – this perspective has lost much credibility. In some parts of the intellectual spectrum, 'capitalism' had always been used in an unconditionally positive sense, at least in the United States (e.g., Friedman, 1962). Now, the affirmative (e.g., Mackey, 2013) or strictly neutral usage (e.g., Hall and Soskice, 2001) of the term gained additional ground. The Great Recession of 2008 had – *thirdly* – an opposite effect. It led to a renewed questioning of the neoliberal belief in the self-regulating capacities of capitalist markets and to new proposals in favor of reregulation. But it also contributed to rising interest in the term 'capitalism.' *Fourthly*, the accelerated globalization of capitalism in the last decades has demonstrated that capitalism can flourish in very different social contexts and under very different political regimes, even in authoritarian and dictatorial systems, at least for a while. A well-defined concept of 'capitalism' appears to be well suited for guiding comparative research on a global scale, now on the agenda. Global perspectives enlarge the typology of observable forms of capitalism. They question received theses, e.g., with respect to the relation between bonded labor, wage labor on a contractual basis, and capitalism. They reopen the debate on the relations between capitalism, freedom, and democracy (Pryor, 2010: 74–111; van der Linden, 2008; Sanyal, 2007).

A Lean Definition

The definitions of capitalism will vary with different topics and different research priorities. In the light of the previous survey, and if a broad systematic and historical analysis of capitalism is intended, a working definition is useful, which emphasizes decentralization, commodification, and accumulation as basic characteristics. On the one hand, it is essential that individual and collective actors dispose of rights, which enable them to make economic decisions in a relatively autonomous and decentralized way. On the other hand, markets serve as the main mechanisms of allocation and coordination; commodification permeates capitalism in many forms, including the commodification of labor. Thirdly, capital is central, that means utilizing resources for investment in the present with the expectation of higher gains in the future, accepting credit besides using savings and returns, dealing with uncertainty and risk as well as having profits, and accumulation as aims. Change, growth, and expansion are inscribed (Kocka, 2010, 2013).

Such a lean definition allows to distinguish between very different types of capitalism and to study the ways in which they have changed and mixed. It also allows to trace capitalist phenomena in epochs and regions in which they were or are just minority phenomena within dominantly noncapitalist environments while acknowledging that to speak of capitalism as an economic *system* its mechanisms must be fully developed and dominant. While such a definition concentrates on economic criteria, it also directs attention to the numerous,

changing, and changeable social, political, legal, and cultural conditions and consequences of capitalism. While the term has been used and can be used for the analysis of very specific problems – like the change from entrepreneurial to managerial capitalism in the history of modern firms (Chandler, 1977; John, 2008) – it has been and can be particularly useful as a frame within which to study the changing interconnections between economic, social, legal, political, and cultural dimensions (e.g., Zakim and Kornblith, 2012). The term can serve as a tool for bridging boundaries between disciplines.

Throughout its history, the concept has been used for relating specific objects of study to grand questions of the past, present, and future, questions in which normative orientations play a role. In this respect, the term 'capitalism' resembles concepts like 'modernity' or 'democracy.' The term 'capitalism' emerged as a critical *concept of difference*, contrasting observations of the present with constructions of different pasts and imaginations of a future, which would be different again. Criticism and analysis have been closely intertwined. While this accounts for some of the problems and weaknesses of the concept, it also explains some of its particular strengths and its continuous attraction. Today, it is not at all clear how an alternative to capitalism could and should look like. People who use the concept surely diverge as to their personal and political visions of the future. But some, though vague, idea or protovision about how things could and perhaps should be, differently from the present, seem to be in the back of the head of some of those who use the term 'capitalism' to talk about the past, present, and future. In this sense, the critical potential of the concept has not fully disappeared, while not exacerbating empirical description and theoretical analysis.

See also: Capital: History of the Concept; Capitalism: Global; Capitalism; Communism, History of; Marx, Karl (1818–83); Marxism in Contemporary Sociology; Schumpeter, Joseph A. (1883–1950); Socialism: Historical Aspects; Socialism; Veblen, Thorstein (1857–1929); Weber, Max (1864–1920); Weberian Social Thought, History of.

Bibliography

Arrighi, G., 1994. The Long Twentieth Century. Money, Power and the Origins of Our Time. Verso, London.
Arrighi, G., 2007. Adam Smith in Beijing. Lineages of the Twenty-First Century. Verso, London.
Baran, P.A., Sweezy, P., 1966. Monopoly Capital. An Essay on the American Economic and Social Order. Verso, New York.
Bastiat, F., 1851. Harmonies économiques, second ed. Guillaumin et Cie, Paris.
Bell, D., 1976. The Cultural Contradictions of Capitalism. Basic Books, New York.
Boltanski, L., Chiapello, E., 2000. Le nouvel esprit du capitalisme. Gallimard, Paris.
Bottomore, T., 1985. The Theories of Modern Capitalism. G. Allen & Unwin, London.
Braudel, F., 1967–79. Civilisation matérielle, économie et capitalisme, 3 vols. A. Colin, Paris.
Brick, H., 2006. Transcending Capitalism. Visions of a New Society in Modern American Thought. Cornell University Press, Ithaca, NY.
Chandler Jr., A.D., 1977. The Visible Hand. The Managerial Revolution in American Business. Harvard University Press, Cambridge, MA.
Cunningham, W., 1916. The Progress of Capitalism in England. Cambridge University Press, Cambridge.
Friedman, M., 1962. Capitalism and Freedom. The University of Chicago Press, Chicago.
Grassby, R., 1999. The Idea of Capitalism before the Industrial Revolution. Rowman & Littlefield, Lanham.
Habermas, J., 1973. Legitimationsprobleme des Spätkapitalismus. Suhrkamp, Frankfurt.
Hall, P.A., Soskice, D. (Eds.), 2001. Varieties of Capitalism. The Institutional Foundations of Comparative Advantage. Oxford University Press, Oxford.
Hartwell, R.M., 1983. The origins of capitalism: a methodological essay. In: Pejovich, S. (Ed.), Philosophical and Economic Foundations of Capitalism. D.T. Heath & Co., Lexington, MA, pp. 11–23.
Harvey, D., 2007. A Brief History of Neoliberalism. Oxford University Press, Oxford.
Hayek, F.A. (Ed.), 1954. Capitalism and the Historians. University of Chicago Press, Chicago.
Hilferding, R., 1910. Das Finanzkapital. Eine Studie zur jüngsten Entwicklung des Kapitalismus. Verlag der Wiener Volksbuchhandlung Ignaz Brand & Co., Wien.
Hobson, J.A., 1894. The Evolution of Modern Capitalism: A Study of Machine Production. W. Scott, London.
Hilger, M.-E., 1982. Kapital, Kapitalist, Kapitalismus. In: Brunner, O., et al. (Eds.), Geschichtliche Grundbegriffe, vol. 3. Klett, Stuttgart, pp. 339–454.
John, R.R., 2008. Turner, Beard, Chandler: progressive historians. In: Business History Review, vol. 82, pp. 227–240.
Keynes, J.M., 1927. The End of Laissez-Faire. The Hogarth Press, London.
Kocka, J., 2010. Writing the history of capitalism. In: Bulletin of the German Historical Institute, vol. 47. German Historical Institute, Washington, pp. 7–24.
Kocka, J., 2013. Geschichte des Kapitalismus. Beck, München (Engl. transl. forthcoming 2015 with Princeton University Press).
van der Linden, M., 2008. Workers of the World. Essays toward a Global Labor History. Brill, Leiden.
Luxemburg, R., 1913. Die Akkumulation des Kapitals. Ein Beitrag zur ökonomischen Erklärung des Kapitalismus. Buchhandlung Vorwärts Paul Singer, Berlin.
Mackey, J., 2013. Conscious Capitalism. Liberating the Heroic Spirit of Business. Harvard Business Review Press, Cambridge, MA.
Marks, S.G., 2012. The word "capitalism": the Soviet Union's gift to America. In: Society, vol. 9, pp. 155–163.
Marx, K., 1967. Capital: A Critique of Political Economy, vol. I. International Publishers, New York.
Merrill, M., 1990. The Anticapitalist Origins of the United States. In: Review: The Journal of the Fernand Braudel Center 13 (4), 465–497.
Merrill, M., 2014. How Capitalism Got Its Name. In: Dissent Fall, pp. 87–92.
Mommsen, W., 1997. Die Wirtschaftsgesinnung des modernen marktorientierten Kapitalismus. In: L'èthique protestante de Max Weber et l'esprit de la modernité. Maison des Sciences de l'Homme, Paris, pp. 208–224.
Muller, J.Z., 2003. The Mind and the Market. Capitalism in Western Thought. Anchor Books, New York.
Parsons, T., 1928–29. "Capitalism" in recent German literature: Sombart and Weber. In: Journal of Political Economy, vol. 36, pp. 641–644, 37, pp. 31–51.
Passow, R., 1918/1927. Kapitalismus". Eine begrifflich-terminologische Studie, second ed. G. Fischer, Jena.
Pendennis, A., Thackeray, W.M. (Eds.), 1855. The Newcomes. Memoirs of a Most Respectable Family, vol. 2. Bradbury and Evans, London.
Pirenne, H., 1914. The stages in the history of capitalism. In: American Historical Review, vol. 19, pp. 494–515.
Polanyi, K., 1944. The Great Transformation. Farrar & Rinehart, Inc., New York.
Poulantzas, N., 1974. Fascism and Dictatorship. The Third International and the Problem of Fascism. NLB, London.
Pryor, F.L., 2010. Capitalism Reassessed. Cambridge University Press, Cambridge.
Rodbertus-Jagetzow, K., 1869. Zur Erklärung und Abhülfe Der heutigen Cerditnoth des Grundbesitzes, vol. 2. Friedrich Mauke, Jena.
Salvioli, J., 1906. Le capitalisme dans le monde antique. V. Giard & E. Brière, Paris.
Sanyal, K., 2007. Rethinking Capitalist Development. Primitive Accumulation, Governmentality and Post-Colonial Capitalism. Routledge, London.
Schäffle, A.E.F., 1870. Kapitalismus und Socialismus mit besonderer Rücksicht auf Geschäfts- und Vermögensformen. H. Laupp, Tübingen.
Schluchter, W., 2009. Die Entzauberung der Welt. Sechs Studien zu Max Weber. J.C.B. Mohr, Tübingen.
Schumpeter, J.A., 1939. Business Cycles. A Theoretical, Historical and Statistical Analysis of the Capitalist Process, vol. 1. Porcubine Press, Philadelphia.
Schumpeter, J.A., 1976. Capitalism, Socialism, and Democracy (1942). Harper & Row, New York.
Sée, H., 1928. Les origines du capitalisme modern. Adelphi Co., New York.
Silberner, L.F., Silberner, E., 1940. Le mot capitalisme. Annales d'histoire sociale 2 (2), 133–134.
Sklansky, J., 2012. The elusive sovereign: new intellectual and social histories of capitalism. In: Modern Intellectual History, vol. 9, pp. 32–248.

Sombart, W., 1919–27. Der Moderne Kapitalismus. Historisch-systematische Darstellung des gesamteuropäischen Wirtschaftslebens von seinen Anfängen bis zur Gegenwart, third, second and first ed., 3 vols. Duncker & Humblot, Leipzig (first ed. 1902).

Sombart, W., 1935. Capitalism. In: Seligman, R.A., Johnson, A. (Eds.), Encyclopaedia of the Social Sciences, vol. 3. MacMillan, New York, pp. I95–208.

Streeck, W., 2009. Reforming Capitalism: Institutional Change in the German Political Economy. Oxford University Press, Oxford.

Streeck, W., May/June 2014. How will capitalism end? New Left Review 87, 35–64.

Tawney, R.H., 1962. Religion and the Rise of Capitalism. A Historical Study (1922, 1926). P. Smith, Gloucester, MA.

Veblen, T., 1914. The Instinct of Workmanship. The Macmillan Company, New York.

Vries, P., 2012. Europe and the rest. Braudel on capitalism. In: Garner, G., Middell, M. (Eds.), Aufbruch in die Weltwirtschaft. Braudel wiedergelesen. Leipziger Univ.-Verl., Leipzig, pp. 81–144.

Wallerstein, I., 1974. The Modern World System, vol. 1. Academic Press, New York.

Wallerstein, I., et al., 2013. Does Capitalism Has a Future? Oxford University Press.

Weber, M., 1988. Die protestantische Ethik und der Geist des Kapitalismus (1904/05). In: Gesammelte Aufsätze zur Religionssoziologie, vol. 1. J.C.B. Mohr, Tübingen, pp. 17–206.

Weber, M., 2006. Zur Sozial- und Wirtschaftsgeschichte des Altertums. Schriften und Reden 1893–1908. J.C.B. Mohr, Tübingen.

Weber, M., 2012. Abriß der universalen Sozial- und Wirtschaftsgeschichte. Mit- und Nachschriften 1919/20. J.C.B. Mohr, Tübingen.

Williams, R., 1976. Keywords. A Vocabulary of Culture and Society. Oxford University Press, New York.

Winkler, H.A. (Ed.), 1974. Organisierter Kapitalismus. Voraussetzungen und Anfänge. Vandenhoeck & Ruprecht, Göttingen.

Wood, E.M., 2002. The Origin of Capitalism. A Longer View. Verso, London.

Zakim, M., Kornblith, G.J. (Eds.), 2012. Capitalism Takes Command. The Second Transformation of 19th-Century America. University of Chicago Press, Chicago.

Cardiac Disease, Coping with

Heike Spaderna and Susan Hellwig, Bergische Universität Wuppertal, Wuppertal, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Because of rising prevalence rates of coronary heart disease and chronic heart failure, an increasing number of persons have to cope with cardiac disease. This article gives an overview of stress and coping concepts applied in the domain of adult heart disease. Approach-oriented coping strategies and pursuing goals in life appear to protect cardiac patients from negative emotions. Rumination, however, appears to contribute to the development of depressive symptoms. Interventions should consider patients' dispositional coping styles. Methodological limitations and suggestions for future research including prospective designs and objectively assessed outcomes in the area of cardiac diseases are outlined.

Introduction

Cardiac diseases constitute a major health problem worldwide. Coronary heart disease (CHD), myocardial infarction (MI), and chronic heart failure (chronic HF) are the main contributors to this disease burden. Chronic HF denotes a clinical syndrome that can result from any structural or functional cardiac disorder that impairs the ventricles' ability to fill with or to eject blood, such as CHD and MI (Hunt et al., 2009). This article focuses on coping processes in adults with the most common cardiac conditions, CHD and chronic HF. Stress and coping are defined, and research regarding associations of coping styles and strategies with outcomes investigated cross-sectionally and longitudinally will be reviewed.

Persons who cope effectively with stressful encounters are expected to report good well-being, few psychological problems, a good quality of life, and, importantly, also good physical health (Lazarus, 1993). Thus, we give an overview about associations of coping styles and strategies with emotional outcomes including depression and post-traumatic stress disorder (PTSD), behavioral outcomes such as resuming walking, and medical outcomes including mortality. In the concluding section, open questions and directions for future research are outlined.

Definitions of Stress and Coping in Cardiac Disease

Stress and coping in the context of cardiac disease are typically conceptualized according to the transactional stress and coping theory by Lazarus and Folkman (1984). Psychological stress is defined as "a particular relationship between the person and the environment that is appraised by the person as taxing or exceeding his or her resources and endangering his or her well-being" (Lazarus and Folkman, 1984: p. 19). According to Lazarus and Folkman (1984), the way a person appraises situations determines both stress reactions and coping efforts. Coping in this situational approach is defined as "constantly changing cognitive and behavioral efforts to manage specific external and/or internal demands that are appraised as taxing or exceeding the resources of the person." (Lazarus and Folkman, 1984: p. 141). In order to evaluate which type of coping is beneficial for cardiac patients' adjustment, it is important to note that the term 'coping' refers to all efforts to deal with a stressful encounter, independent of their effectiveness (Lazarus and Folkman, 1984).

Lazarus (1993) emphasizes that when studying how patients cope with their illness, it is important to specify the particular threats the patient is experiencing at that time, rather than focusing on the illness in general. In line with this reasoning, specific stressors experienced by patients with CHD and chronic HF have been identified. Survivors of an MI report distress due to the emotional impact of the disease, life style changes, dealing with health professionals, reactions of the partner, changes in roles of social life, return to work, financial difficulties, and dependency (e.g., Stewart et al., 2000). Similar stressors have been reported by patients with chronic HF (Bosworth et al., 2004; Spaderna et al., 2012). Additional stressors comprise the disease's interferences with personal goals (Garnefski and Kraaji, 2010) or medication side effects (Bosworth et al., 2004). In addition to disease-related stressors, patients also encounter common stressors such as death of a loved one or children's problems (Bunyamin et al., 2013). Considering that perceived stress contributes to future adverse events and reduced long-term survival in patients with cardiac disease (e.g., Aboa-Éboulé et al., 2007; Arnold et al., 2012), identification of dispositional coping styles and actual coping strategies that promote successful adjustment is crucial.

Coping styles refer to stable personality dispositions that determine a person's typical way of responding to stressful encounters. These approaches often are based on theoretical distinctions of approach versus avoidance of threats. Trait approaches of coping that have been applied to cardiac patients include sensitization/repression (Weinberger et al., 1979) and vigilance/cognitive avoidance (Krohne, 1993). Others postulated a broader set of possible coping dimensions. For example, Carver et al. (1989) developed the Coping Orientation to Problem Experience (COPE) to assess 14 different dispositional coping strategies persons usually adopt in stressful situations.

According to the situational coping approach of Lazarus and Folkman, the COPE can also be used to assess which of various coping strategies were actually applied in a specific situation (Carver et al., 1989). This instrument has been widely used with cardiac patients, both in its dispositional and actual version. Other self-report instruments to assess coping strategies exist. Importantly, these instruments cover a different range of actual coping behaviors, and yield different strategies.

Although the term 'strategy' implies that persons pursue specific goals when dealing with a stressor, these goals generally are not part of the assessment.

Dispositional Coping Styles in Cardiac Disease

Coping Styles and Emotional Well-Being

It has been argued that a repressive coping style might be beneficial in the short run, but might undermine necessary cognitive and behavioral adaptations to the illness, thereby leading to adverse outcomes in the long run (Suls and Fletcher, 1985). According to these expectations, cross-sectional empirical evidence indicates that cardiac patients with an avoidant coping style reported less negative affect, more positive affect (Bedi and Brown, 2005), and lower distress and depressive symptoms than patients with a nonavoidant coping style (Denollet et al., 2008). Similarly, cognitive avoidance was related to better emotional quality of life (but unrelated to physical quality of life) in patients newly listed for cardiac transplantation, whereas vigilance, i.e., directing attention toward threatening information, was accompanied with worse quality of life (Bunyamin et al., 2013).

A longitudinal study by Ginzburg et al. (2002) contradicts the assumption that a repressive coping style leads to negative emotional long-term consequences. Repressors not only showed less-severe acute stress disorder than high-anxious and defensive patients 1 week after their MI, but were also less likely than nonrepressors to be classified as having clinical PTSD (7% vs 17%, 19%, and 20% in the other three coping style groups) 7 months later. Repressors also reported lower symptom levels independent of education, prior traumatic events, and acute stress disorder (Ginzburg et al., 2002). Interestingly, repressors and nonrepressors did not differ in their primary appraisals, i.e., in their threat perception of the MI or their medical status (Ginzburg et al., 2002). Replications and data from studies with a longer follow-up are warranted in this area.

Assessing denial via the dispositional version of the COPE, on the other hand, was unrelated to emotional well-being as indicated by depressive symptoms after 2 years in 83 patients with chronic HF (Murberg et al., 2002). However, behavioral disengagement (i.e., giving up attempts to attain goals) predicted higher depression scores, even after controlling for sex, age, neuroticism, extraversion, and New York Heart Association (NYHA) class, and indicator for disease severity, which is based on patients' self-reports (Murberg et al., 2002). Thus, diverting attention from threats should be distinguished from behavioral disengagement.

Coping Styles, Medical Outcomes, and Mortality

An avoidant coping style appears to help patients to maintain positive affect and to prevent negative affect and stress. However, one study indicates that avoidance (repression) might be associated with impaired survival in the long run. After a mean follow-up of 6.6 years, repressors had a two-fold risk of death or nonfatal MI compared to nonrepressors, even after statistically controlling for covariates including disease severity at baseline (Denollet et al., 2008).

There is some evidence that not an avoidant coping style itself, but the incompatibility of information or interventions with a patient's coping style, are detrimental to patients' physical health. Early evidence regarding the adverse effect of risk information on cardiac patients' health status comes from a small study of 97 patients who had received percutaneous coronary transluminal angioplasty (PCTA) (Shaw et al., 1986). Sensitizers (see Weinberger et al., 1979) were at higher risk for restenosis 6 months later when they had little knowledge about procedural and risk factors, compared to sensitizers who had a high level of information. Regarding additional medical complications 6 months after PCTA (e.g., hospitalization for chest pain, MI, arrhythmias), repressors with a high level of information were more likely to have complications than repressors with low-information level. However, this finding applied only to patients with no prior history of heart attack (Shaw et al., 1986).

A mismatch of patients' coping style with psychosocial treatment can even challenge survival. In the Montreal Heart Attack Readjustment Trial, 1.376 patients were randomized to control or treatment, with the latter consisting of monthly telephone calls to monitor patients' distress (Frasure-Smith et al., 2002). If a patient reported elevated distress or was rehospitalized, he or she received a home-based nursing intervention. In the treatment group, both men and women with a repressive coping style had a two-fold mortality risk compared to repressors of the control group, even after 5 years of follow-up (Frasure-Smith et al., 2002). By the end of the 1-year intervention program, repressors of the treatment group were more likely to be prescribed benzodiazepines and to have had emergency room visits than those in the control group. They also tended to report more somatic depression symptoms. On the contrary, for men who qualified as truly anxious (i.e., high anxiety regardless of social desirability), the intervention was beneficial, which was mediated by reductions in somatic depression symptoms (Frasure-Smith et al., 2002). The authors conclude that their intervention interfered with repressors' coping style, undermining their need to avoid perceptions of distress. Thus, interventions aiming at distress regulation should be tailored to a patient's coping style (Frasure-Smith et al., 2002).

Summary

In sum, the few studies available to date indicate that among patients with CHD and advanced chronic HF, those with an avoidant coping style, i.e., a disposition to avert attention from threats, do better in terms of emotional well-being and quality of life than nonrepressors or persons who score high on vigilance or sensitization. Repressors appear also to be at lower risk to develop PTSD. However, they might be at increased risk for medical complications, additional MIs, and premature death in the long run, particularly if they cannot cope according to their favored style, i.e., if they are forced to focus on distress and risk factors. Sensitizers, on the contrary, appear to benefit from information that helps to reduce their anxiety.

Actual Coping Strategies in Cardiac Disease

Coping Strategies and Depression

During the last decade, it has been consistently observed that 20–40% of patients with CHD and chronic HF are clinically

depressed or experienceelevated depressive symptoms (Celano and Huffman, 2011). As depression is independently related to an impaired prognosis in CHD and chronic HF (Celano and Huffman, 2011), even among patients with advanced disease who are listed for cardiac transplantation (Spaderna et al., 2014), researchers started to investigate whether specific appraisals and coping strategies are related to the development of clinical depression or depressive symptoms in these patients.

Threat Appraisal

Threat appraisals, i.e., appraising disease-related stressors as harmful to one's physical and emotional well-being, have been cross-sectionally linked to a diagnosis and to symptoms of depression in patients hospitalized because of an MI (e.g., Vögele et al., 2012). Longitudinally, threat appraisal contributed to the development of depressive symptoms after 6 months in patients with chronic HF, whereas challenge appraisal was unrelated to depression (Park et al., 2006).

Approach versus Avoidance Coping Strategies

Contrary to findings regarding coping styles, avoidance coping strategies and depressive symptoms occur together (e.g., Vollman et al., 2007). In contrast, approach-oriented coping strategies such as planful problem solving and seeking social support were negatively related to depressive symptoms (Vollman et al., 2007).

Regarding the development of depressive symptoms, prospective studies indicate that avoidance strategies are irrelevant (e.g., Schröder et al., 1998). van Elderen et al. (1999), who measured *cognitive* approach and avoidance coping in 278 patients with CHD, observed neither any cross-sectional nor longitudinal relationships between avoidance (assessed by items such as "If I'm worried about my heart disease … I try to keep my mind off it"; "I think that worrying about it doesn't make much sense") and depressive symptoms. Approach coping, however, was cross-sectionally related to higher depressive symptoms (and state anxiety), and longitudinally (3 and 12 months later) to lower depression (van Elderen et al., 1999). The latter result is surprising, considering that approach coping was operationalized via items such as "I allow myself to be controlled by it (heart disease)" or "I think how easy it is for healthy people", self-blaming and viewing the world pessimistically, indicating a rather passive approach to the stressor. A composite labeled 'active coping' also helped to counter depressive symptoms assessed after 6 months in patients with chronic HF (Park et al., 2006). This might be particularly relevant for those patients who perceive their health condition as controllable (Park et al., 2012).

Garnefski and Kraaji (2010) emphasized that approaching the stressor 'cardiac disease' by finding new goals in one's life (goal reengagement) contributes to a stable low level of depressive symptoms. In the study by Vögele et al. (2012), the coping strategies 'threat minimization' and 'search for affiliation' were associated with lower depression scores immediately after the MI and up to 6 months later (Vögele et al., 2012). Interestingly, patients who perceived higher threat reported to make less use of these strategies (Vögele et al., 2012). Social coping was also related to fewer depressive symptoms and more energy after surgery in the study by Schröder et al. (1998). However, these studies cannot explain by which means (avoiding or approaching the problem) threat is minimized and, similarly, if patients seek social contacts in order to avoid threat or to approach it together with the support of others.

In sum, cardiac patients who are also depressed, report to make less use of active approach coping strategies such as problem solving or reframing, and, at least in some studies, report to adopt more avoidant strategies and self-blaming. The finding that depression is accompanied by the adoption of passive coping strategies, less reframing, more self-blaming, and viewing the world pessimistically, is not surprising and could be conceived as part of the depressive disorder's symptomatology. Based on the few longitudinal studies, it appears that avoidance strategies are irrelevant for the development of depression, whereas approach coping, but also seeking social support helps to diminish depressive symptoms. The discrepancy of findings regarding avoidant strategies (often accompanied by higher depression) and an avoidant coping style (associated with lower reported depression) deserves more attention.

Rumination

In a different approach to explain the development of depression in cardiac patients, researchers have focused on specific coping strategies, particularly rumination. This construct has been defined as "behaviours and thoughts that focus one's attention on one's depressive symptoms and on the implications of those symptoms" (Nolen-Hoeksema, 1991, p. 569). Recent studies did find support for the relevance of rumination both for self-reported cardiac symptoms such as fatigue or sleep problems (León et al., 2010) and for depressive symptoms (e.g., Garnefski and Kraaji, 2010), although not consistently. Schröder and colleagues found that those patients who had often used rumination as a coping strategy during the month before their cardiac surgery reported more depressive symptoms, reduced energy, and more worries about their infarct, functioning, and symptoms in the week after cardiac surgery than those patients who had adopted rumination infrequently (Schröder et al., 1998). However, Vögele et al. (2012) did not observe any associations of rumination with future depression (self-reported and clinically diagnosed according DSM-IV) in patients after an MI.

It is conceivable that the nature of ruminative thoughts and additional factors such as existing depressive symptoms and vulnerabilities (Denton et al., 2012) need to be taken into account to explain the development of depressive symptoms. In 387 patients who showed no signs of depression after an acute coronary syndrome, not brooding rumination (e.g., "I think, what am I doing to deserve this?") per se, but it's interaction with depression vulnerabilities predicted depressive symptoms 3 months later. That is, only when vulnerabilities such as dysfunctional beliefs, few pleasant events, or poor dyadic adjustment were also present at baseline, rumination did predict an increase in depressive symptoms. In patients who already showed elevated depression scores at baseline, rumination directly predicted depression severity 3 months later (Denton et al., 2012). Reflective rumination (e.g., "I analyze recent events to understand why I am depressed") was generally unrelated to depression (Denton et al., 2012). Thus, rumination contributes to the development of depression in cardiac patients, yet mechanisms might be more complex than originally expected.

Coping Strategies and Anxiety

Anxiety was found to be a common emotional response to heart failure due to the unpredictable nature of the disease (Bosworth et al., 2004) and it has also been linked to an increased risk of cardiac events (Frasure-Smith and Lespérance, 2008). Yet, the few studies available to date yield inconsistent findings regarding the association of coping strategies with anxiety.

Cross-sectionally, a frequent use of avoidant coping strategies was linked to higher anxiety and to a stronger relationship of anxiety with poor physical functioning in the study by Eisenberg et al. (2012), whereas van Elderen et al. (1999) reported associations of avoidant coping with lower anxiety and higher well-being at baseline. In this study, avoidant coping did not predict negative emotions later on, although patients, on average, reported to use more avoidant than approach coping.

With regard to approach coping strategies, results are also inconsistent. Approach coping strategies were unrelated with anxiety (Eisenberg et al., 2012), while anxious CHD patients reported less problem-focused coping in the study by Tung et al. (2008). Approach coping, on the other hand, was cross-sectionally associated with higher anxiety, and longitudinally with lower anxiety (and depression) (van Elderen et al., 1999). Thus, it is unclear whether approach coping is accompanied by negative affect in the short run, but pays off with reduced anxiety in the long run. More prospective studies are needed to establish such mechanisms.

In sum, the few studies examining the relationship of anxiety and coping activities yield inconsistent results, which might in part be explained by use of different operationalizations of avoidant and approach coping strategies. In addition, depression and anxiety are often highly correlated and it remains unclear if these emotional states or disorders are differentially related with coping.

Coping Strategies and Behavioral Outcomes

Resuming an active life after MI is an important goal of rehabilitation. However, the role of coping for achieving functional outcomes, e.g., return to work, in patients with cardiac disease has not received much attention. Schröder et al. (1998) investigated associations of coping with behavioral indicators of resuming an active life after cardiac surgery. Patients' self-reports regarding activities such as resuming reading, ambulating in the ward, and efforts to regain physical fitness were assessed in the week after surgery. Higher scores in rumination before surgery were associated with a longer stay in the intensive care unit, and a delayed resumption of reading, walking, and visits. Patients who often sought for information started walking later. Social coping, i.e., seeking social contacts and support, was associated with faster recovery regarding reading and ambulation (Schröder et al., 1998).

Coping Strategies and Medical Outcomes

Regarding associations of coping with indicators of disease severity, results of cross-sectional studies consistently show that coping strategies are unrelated to objectively assessed indicators of disease severity, such as degree of obstruction of the coronary arteries (León et al., 2010), number of bypasses, infarcts, medication (Schröder et al., 1998), or left-ventricular ejection farction (Park et al., 2006, 2012). It appears that coping efforts are not restricted by disease severity. Whether the adoption of certain coping strategies in specific stressful encounters is related to medical outcomes remains elusive.

Summary

In sum, it appears that appraising situations as threat instead of challenge contributes to future depression in patients with cardiac disease. In addition, ruminating about stressors and the disease may further increase the risk for depression, particularly if rumination is of the brooding kind and additional vulnerabilities are present. Avoidant coping strategies are accompanied with psychological distress and negatively associated with subjective well-being in patients with CHD and chronic HF, but this type of strategies appears to be of less relevance regarding future depression. It is conceivable, however, that once patients are depressed they tend to rely on avoidant coping strategies more often than nondepressed patients. Active coping on the other hand, an approach strategy, appears to counteract depressive symptoms. Social support and other resources might assist patients to apply coping strategies that are more effective with regard to impairing depression. However, results are heterogeneous due to different conceptualizations of coping styles and strategies and suboptimal study designs. Larger longitudinal studies are needed to elucidate the relationships between appraisal, adoption of coping strategies, and the development of depression and anxiety after MI and in patients with chronic HF. Moreover, studies investigating the impact of coping strategies on behavioral outcomes are warranted.

Future Directions

Coping styles and strategies are worth to be considered in cardiac disease. Because there is neither a generally accepted definition of coping styles and strategies nor agreement regarding their measurement, instruments yield assessments of coping constructs with different structures and varying numbers and types of strategies. Some authors use the terms 'style' and 'strategy' interchangeably, or they conceptualize coping in terms of a disposition while assessing it by an instrument designed to measure strategies. In addition, some researchers rely on idiosyncratic repertories of coping concepts, often derived from factor analysis of very small samples. Also, in some studies, different coping subscales are merged to a composite measure, either based on correlational patterns or theoretical presumptions. In consequence, despite using the same instruments and labels, the underlying subscales might differ considerably. Thus, inconsistencies may simply result from the heterogeneous use of coping terms and inadequate operationalizations of the coping constructs under investigation.

Inconsistencies in findings may also result from different operationalizations of stressors. In some studies, the instructions are adapted to the domain of cardiac disease, yet some mention only the disease itself, while others specify distinct stressors such as angina symptoms. In other studies, it remains unclear if patients had to refer to a specific stressor at all. As has

been shown, cardiac patients report a variety of disease-related stressors and, additionally, are exposed to stressors that are not directly related to their disease. Therefore, types of stressor and situational characteristics such as controllability of the stressor need to be taken into account when evaluating coping efforts. However, this goodness-of-fit-hypothesis was rarely considered (e.g., Park et al., 2012). Furthermore, to better understand possible causal mechanisms, more studies applying longitudinal or experimental designs are recommended.

Moreover, future studies could benefit from addressing a broader range of outcomes. Particularly medical outcomes such as cardiac events, rehospitalizations, and survival as consequences of coping need to be examined in prospective studies. In relation to disease-related stressors, behavioral indicators such as resuming daily activities or returning to work deserve more attention. In addition, positive outcomes in cardiac disease such as a greater appreciation of life, greater sense of personal strength, or spiritual development are understudied in these patients. Future studies should also explicitly explore potential gender differences in coping with cardiac disease.

With regard to psychosocial interventions to improve patients' coping efforts, not much is known except for the importance to consider patients' coping styles. Stress management is often part of psychological interventions for cardiac patients (Linden et al., 2007), but its components vary from teaching of specific cognitive coping strategies to relaxation techniques, yoga, and support groups. Although comprehensive psychosocial interventions including dietary changes and regular exercise have been shown to benefit cardiac patients (Linden et al., 2007), no conclusion with regard to specific coping strategies can be drawn. Psychological interventions to decrease depression in cardiac patients mostly rely on problem solving and cognitive behavior therapy and are able to reduce depression, although effect sizes are small (Dickens et al., 2013).

Conclusion

Patients with cardiac diseases experience illness-related and common stressors, and, in addition, prolonged stress experience appears to predict additional adverse events and mortality. However, current knowledge regarding effective ways of coping for cardiac patients is limited by a variety of different coping concepts, operationalizations, and limited research designs. Patients with an avoidant coping style divert their attention from threats and therefore report less depressive symptoms than nonrepressors and are at reduced risk for developing PTSD. However, prognosis might be impaired if patients' coping styles are undermined by highlighting potential problems and risks, and providing information for repressors, and withholding important information for sensitizers, respectively. Approaching stressors, adopting active coping strategies, and developing new goals appear to counteract depressive symptoms in cardiac patients. Brooding rumination, on the contrary, might promote the development of depressive symptoms, particularly if additional risk factors for depression are present. In sum, more research is needed to address relevant issues of coping in cardiac disease. Studies should be based on theoretical approaches and apply a sound assessment of coping and related variables. Prospective research designs and randomized controlled trials examining medical and behavioral outcomes with varying time frames are warranted. This can help to elucidate the impact of appraisal processes, coping strategies, and styles on adjustment in men and women with cardiac diseases.

See also: Chronic Illness, Psychosocial Coping with; Coping, Personality and; Coronary Heart Disease: Psychosocial Aspects; Depression, Pessimism, and Health; Depression; Personality and Health.

Bibliography

Aboa-Éboulé, C., Brisson, C., Maunsell, E., Masse, B., Bourbonnais, R., Vézina, M., Milot, A., Théroux, P., Dagenais, G.R., 2007. Job strain and risk of acute recurrent coronary heart disease events. Journal of the American Medical Association 298, 1652–1660.

Arnold, S.V., Smolderen, K.G., Buchanan, D.M., Li, Y., Spertus, J.A., 2012. Perceived stress in myocardial infarction. Long-term mortality and health status outcomes. Journal of the American College of Cardiology 60, 1756–1763.

Bedi, G., Brown, S.L., 2005. Optimism, coping style and emotional well-being in cardiac patients. British Journal of Health Psychology 10, 57–70.

Bosworth, H.B., Steinhauser, K.E., Orr, M., Lindquist, J.H., Grambow, S.C., Oddone, E.Z., 2004. Congestive heart failure patients' perceptions of quality of life: the integration of physical and psychosocial factors. Aging & Mental Health 8, 83–91.

Bunyamin, V., Spaderna, H., Weidner, G., For the Waiting for a New Heart Study Group, 2013. Health behaviors contribute to quality of life in patients with advanced heart failure independent of psychological and medical patient characteristics. Quality of Life Research 22, 1603–1611.

Carver, C.S., Scheier, M.F., Weintraub, J.K., 1989. Assessing coping strategies: a theoretically based approach. Journal of Personality and Social Psychology 56, 267–283.

Celano, C.M., Huffman, J.C., 2011. Depression and cardiac disease: a review. Cardiology Review 19, 130–142.

Denollet, J., Martens, E.J., Nyklíček, I., Conraads, V.M., de Gelder, B., 2008. Clinical events in coronary patients who report low distress: adverse effect of repressive coping. Health Psychology 27, 302–308.

Denton, E.-g. D., Rieckmann, N., Davidson, K.W., Chaplin, W.F., 2012. Psychosocial vulnerabilities to depression after acute coronary syndrome: the pivotal role of rumination in predicting and maintaining depression. Frontiers in Psychology 3, 1–9.

Dickens, C., Cherrington, A., Adeyemi, I., Roughley, K., Bower, P., Garret, C., Bundy, C., Coventry, P., 2013. Characteristics of psychological interventions that improve depression in people with coronary heart disease: a systematic review and meta-regression. Psychosomatic Medicine 75, 211–221.

Eisenberg, S.A., Shen, B.-J., Schwarz, E.R., Mallon, S., 2012. Avoidant coping moderates the association between anxiety and patient-rated physical functioning in heart failure patients. Journal of Behavioral Medicine 35, 253–261.

van Elderen, T., Maes, S., Dusseldorp, E., 1999. Coping with coronary heart disease: a longitudinal study. Journal of Psychosomatic Research 47, 175–183.

Frasure-Smith, N., Lespérance, F., 2008. Depression and anxiety as predictors of 2-year cardiac events in patients with stable coronary artery disease. Archives of General Psychiatry 65, 62–71.

Frasure-Smith, N., Lespérance, F., Gravel, G., Masson, A., Juneau, M., Bourassa, M.G., 2002. Long-term survival differences among low-anxious, high-anxious and repressive copers enrolled in the Montreal Heart Attack Readjustment Trial. Psychosomatic Medicine 64, 571–579.

Garnefski, N., Kraaji, V., 2010. Do cognitive coping and goal adjustment strategies used shortly after myocardial infarction predict depressive outcomes 1 year later? Journal of Cardiovascular Nursing 25, 383–389.

Ginzburg, K., Solomon, Z., Bleich, A., 2002. Repressive coping style, acute stress disorder, and posttraumatic stress disorder after myocardial infarction. Psychosomatic Medicine 64, 748–757.

Hunt, S.A., Abraham, W.T., Chin, M.H., Feldman, A.M., Francis, G.S., Ganiats, T.G., Yancy, C.W., 2009. 2009 focused update incorporated into the ACC/AHA 2005 guidelines for the diagnosis and management of heart failure in adults. Journal of the American College of Cardiology 53, e1–e90.

Krohne, H.W., 1993. Vigilance and cognitive avoidance as concepts in coping research. In: Krohne, H.W. (Ed.), Attention and Avoidance: Strategies in Coping with Aversiveness. Hogrefe & Huber, Seattle, pp. 19–50.

Lazarus, R.S., 1993. Coping theory and research: past, present, and future. Psychosomatic Medicine 55, 234–247.

Lazarus, R.S., Folkman, S., 1984. Stress, Appraisal, and Coping. Springer Publishing Company, New York.

León, T.C., Nouwen, A., Sheffield, D., Jaumdally, R., Lip, G.Y.H., 2010. Anger rumination, social support, and cardiac symptoms in patients undergoing angiography. British Journal of Health Psychology 15, 841–857.

Linden, W., Philips, M.J., Leclerc, J., 2007. Psychological treatment of cardiac patients: a meta-analysis. European Heart Journal 28, 2972–2984.

Murberg, T.A., Bru, E., Stephens, P., 2002. Personality and coping among congestive heart failure patients. Personality and Individual Differences 32, 775–784.

Nolen-Hoeksema, S., 1991. Responses to depression and their effects on the duration of depressive episodes. Journal of Abnormal Psychology 100, 569–582.

Park, C.L., Fenster, J.R., Suresh, D.P., Bliss, D.E., 2006. Social support, appraisals, and coping as predictors of depression in congestive heart failure patients. Psychology & Health 21, 773–789.

Park, C.L., Sacco, S.J., Edmondson, D., 2012. Expanding coping goodness-of-fit: religious coping, health locus of control, and depressed affect in heart failure patients. Anxiety Stress Coping 25, 137–153.

Schröder, K.E.E., Schwarzer, R., Konertz, W., 1998. Coping as a mediator in recovery from cardiac surgery. Psychology & Health 13, 83–97.

Shaw, R.E., Cohen, F., Fishman-Rosen, J., Murphy, M.C., Stertzer, S.H., Clark, D.A., Myler, R.K., 1986. Psychologic predictors of psychosocial and medical outcomes in patients undergoing coronary angioplasty. Psychosomatic Medicine 48, 582–597.

Spaderna, H., Vögele, C., Barten, M.J., Smits, J.M.A., Bunyamin, V., Weidner, G., 10 February 2014. Physical activity and depression predict event-free survival in heart transplant candidates. Health Psychology. Advance online publication. http://dx.doi.org/10.1037/hea0000033.

Spaderna, H., Weidner, G., Koch, K.C., Kaczmarek, I., Wagner, F.M., Smits, J.M.A., 2012. Medical and psychosocial predictors of mechanical circulatory support device implantation and competing outcomes in the Waiting for a New Heart Study. Journal of Heart and Lung Transplantation 31, 16–26.

Stewart, M., Davidson, K., Meade, D., Hirth, A., Makrides, L., 2000. Myocardial infarction: survivors' and spouses' stress, coping, and support. Journal of Advanced Nursing 31, 1351–1360.

Suls, J., Fletcher, B., 1985. The relative efficacy of avoidant and nonavoidant coping strategies: a meta-analysis. Health Psychology 4, 249–288.

Tung, H.-H., Hunter, A., Wei, J., 2008. Coping, anxiety and quality of life after coronary artery bypass graft surgery. Journal of Advanced Nursing 61, 651–663.

Vögele, C., Christ, O., Spaderna, H., 2012. Cardiac threat appraisal and depression after first myocardial infarction. Frontiers in Psychology 3, 1–7.

Vollman, M.W., LaMontagne, L.L., Hepworth, J.T., 2007. Coping with depressive symptoms in adults living with heart failure. Journal of Cardiovascular Nursing 22, 125–130.

Weinberger, D.A., Schwartz, G.E., Davidson, R.J., 1979. Low-anxious, high-anxious, and repressive coping styles: psychometric patterns and behavioral and physiological responses to stress. Journal of Abnormal Psychology 88, 369–380.

Relevant Website

http://www.who.int/gho/mortality_burden_disease/causes_death/2000_2011/en/index.html – WHO 2013.

Cardiovascular Reactivity and Recovery

Joseph P Gyekis and William Gerin, Pennsylvania State University, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Acute mental stress leads to transient blood pressure (BP) and heart rate elevations, and these are risk factors for the development of hypertension and heart disease. A corollary is that individuals particularly prone to strong cardiovascular reactivity (CVR) in the lab may experience these responses frequently enough in daily life for these elevations in BP to directly contribute to tissue damage and the occurrence of cardiovascular disease events. In this article, we provide a guide to the methodology of CVR studies. Following a brief review of the empirical literature, we discuss issues regarding the design, implementation, and analysis of reactivity studies.

Introduction

The cardiovascular system is tightly regulated to control the flow of gases and nutrients throughout the body. The circulatory functions of the heart and blood vessels change in response to chemical signals within the blood stream as well as innervation to their tissues from the nervous system. Within the realm of psychology, cardiovascular reactivity (CVR) specifically refers to the degree to which an individual experiences changes in cardiovascular function in response to engaging, challenging, or aversive behavioral situations, especially those that are mental in origin (Manuck, 1994).

Nearly a century of research on CVR has shown that acute mental stress leads to transient changes in a spectrum of cardiovascular responses – prominent among them a characteristic elevation of heart rate and blood pressure (BP) in most individuals experiencing psychosocial stress in the laboratory. Researchers hypothesize that individuals particularly prone to strong CVR in the lab may also experience these responses frequently enough in daily life for these mental challenge-related elevations in BP to directly contribute to tissue damage and the occurrence of cardiovascular disease events (Manuck, 1994). Thus, researchers have proposed that CVR might be a mechanism linking behavioral and psychosocial factors such as anger, hostility, Type A personality, lack of social support, and exposure to cardiovascular diseases.

The purpose of this article is to update a prior article (Goyal et al., 2008) and provide a guide for researchers interested in conducting studies of CVR. Following a brief review of the empirical literature, we discuss issues regarding the design, implementation, and analysis of reactivity studies.

CVR as a Possible Cause of Cardiovascular Disease

Unlike profound risk factors that consistently show large effects on disease outcomes, studies on the role of CVR on future disease outcomes are mixed and study-specific. Chida and Steptoe (2010) recently conducted a meta-analysis on prospective studies of the relationship between CVR and subclinical and clinical endpoints. They found that although studies varied substantially, overall there were significant effects of CVR on various outcomes, especially for future development of hypertension. Although only about 1% of the variability in outcomes was explained by laboratory tests of CVR, this 1% is deemed to be clinically significant because even incremental increases in hypertension risk lead directly to increases in risk of cardiovascular events.

Since findings regarding the generalizability of laboratory CVR to the natural environment have been inconsistent (Gerin et al., 1994; Harshfield et al., 1988; Ironson et al., 1989; Kamarck et al., 2000, 2003, 2005; Turner et al., 1994), it seems possible that the true effect of real-world CVR on cardiovascular disease events may be stronger than those derived from laboratory measures.

In recent years, improvements in technology are making it possible to investigate whether CVR to daily stressors better predict future hypertension and cardiovascular events. However, studies employing ambulatory BP monitoring of CVR to natural stressors recorded by ecological momentary assessment are still in the proof of concept stage, and have not been followed up long term to evaluate their prospective utility. These studies are urgently needed to evaluate whether real-life CVR is indeed linked to subsequent cardiovascular disease (Zanstra and Johnson, 2011).

Still, even if such studies eventually prove compelling, the concern about the possible role of unmeasured confounders ever looms over the temptation to claim that CVR itself is directly responsible for the associated events (Carroll, 2011). The cardiovascular diseases that are more likely to occur in individuals with high CVR might be caused by other complex factors that are difficult to fully control for such as diet, drug intake, physical activity, sleep quality, or as yet unknown factors influencing disease progression. Because effect sizes are small, these issues cannot be addressed by incremental improvements in the ability to control for these confounders. Thus, the hypothesis that CVR itself is damaging to the cardiovascular system will await confirmation until intervention methods are devised to specifically reduce CVR to stress. Then clinical trials will show whether CVR-targeted therapies do indeed influence the development of cardiovascular disease. For example, it will soon be possible to test whether patients with heart rate and BP under artificial control (perhaps recovering heart surgery patients with pacemakers and implanted variable rate drug pumps containing BP reducing medication) fare better if their CVR is effectively

dampened whenever the patient self-reports psychosocial stress.

Laboratory Research on CVR

Typical Procedural Timeline

The typical features of common CVR studies include baseline measures, stressor presentation, and recovery periods.

First, participants are instrumented with a device that measures BP and/or heart rate and then sit through baseline period usually between 5 and 30 min. Cardiovascular measurements are taken throughout, and the mean of these represents the resting measure.

Next, a stressor is then presented, typically for 2–5 min, although longer tasks have been used. More will be said about how stressors are chosen in a later section. Cardiovascular measurements are usually taken at the end of the stressor period, and the average represents the 'task' level. Reactivity is usually defined as the difference between the mean baseline value and the mean task value. More sophisticated representations are also used, in which the baseline is controlled statistically, to account for individual differences in resting levels.

Finally, recovery data may be collected. A post-stress rest period provides a way of evaluating return of the physiological parameter from the stress-induced levels to prestress levels. Cardiovascular measurements are taken during this period. Recovery is more difficult to measure than reactivity; area under the curve is often used as the measure. See Linden et al. (1997) for a review of the various reactivity measures.

The Laboratory Milieu

The laboratory does not necessarily represent a neutral test venue. Subjects come in with a range of expectation, hostility, fear, anxiety, and other characteristics that may influence their baseline and possibly task levels. Carefully controlling as much of this otherwise unexplained error will result in cleaner data, reducing the likelihood of Type II error. It is important that the experimenter be aware of these issues, because it is crucial that his or her demeanor, tone, and other verbal and nonverbal behaviors be carefully controlled between subjects. Because this level of detail may begin to slip over time, quality control checks may be necessary to ensure consistency throughout the study period.

Selection of a Stressful Challenge

Several factors must be taken into account when deciding on a task.

Conceptuality

Does the task conceptually represent the situation being studied, in terms of both psychological and physiological aspects?

Psychometric Properties

Is the task well supported in terms of its reliability, and validity? The results depend on the selection of the task. Make sure there is evidence that the task manipulates what it is supposed to manipulate (e.g., Is the commonly used serial subtraction task manipulating anxiety or active coping?).

Individual Tasks/Challenges

The question of what precisely is being manipulated by many of the common tasks remains highly controversial. They are meant to arouse stress, and undoubtedly they do, but stress is a complex construct, and what specific aspect of 'stress' is aroused by, e.g., public speaking versus the Stroop task, is unclear. Alternatively, it is important to remember that although two tasks may have very different 'phenotypes,' they may still share a conceptual basis. Thus, the mirror star-tracing task does not share any obvious characteristics with, e.g., the Stroop task; however, to the extent that the two tasks involve a manipulation of frustration, they may be said to represent two measures of a single construct.

The Cold Pressor

Historically, the cold pressor was the first task used in a published study of CVR (Hines and Brown, 1932), and was then used in virtually every such study until the late 1950s. The justification for its selection was simply the authors' need for a simple, noninvasive way to elevate BP that could be used in the physician's office (they were trying to develop a diagnostic test for predicting hypertension). It is not clear, however, precisely which dimension of stress is being represented.

Mental Arithmetic and Serial Subtraction

Although the procedures vary from study to study, mental arithmetic is the most commonly used mental stressor in studies of cardiovascular psychophysiology. Typically, the subject is told that he or she will be given a large number and asked to subtract backward from that number aloud, often by 13 s. The task may be used to represent harassment, when the experimenter interrupts to tell the subject that she has made an error and must begin again.

The Stroop Color-Word Test and the Mirror Star-Tracing Procedure

These are two examples of brief cognitive tasks that are commonly used in reactivity studies. In the *Stroop Color-Word Test*, subjects are presented with a list of words that are the names of colors; the words are printed in ink of a different color from the color name. Subjects are asked to name the color of the ink for each word, rather than read the color that is spelled out. The discrepancy between the ink color and the color name creates interference, which is stressful for the subject. In the *mirror star-tracing procedure*, subjects are asked to trace the outline of a star while looking at its reversed image o in a mirror rather than looking directly.

Social Interaction Tasks

A relatively simple social interaction that can be modeled in the laboratory is public speaking, which is stressful for many people. The speech should last at least 5 min. The participant then proceeds to deliver the speech.

The Trier Social Stress Test (TSST) was developed (Kirschbaum et al., 1993) and is believed to induce stress responses by inducing perception of a social-evaluative threat and uncontrollability (Dickerson and Kemeny, 2004). In the first task, subjects are given 10 min to prepare a speech about why they should be hired for a job; they then deliver the 5-minute speech, which they are told is being recorded and judged based on speech and body language. If subjects pause for more than 10 s during the speech, they are told "you still have time, please continue"; after a total of 20 s panel members will begin asking questions. Immediately following this task, subjects are instructed to perform oral serial subtraction (counting backward from 1687 in 13 s) in front of the panel for 5 min, starting over each time they make an error. At the end of the 20-minute test, subjects are debriefed and informed that no taping or evaluation occurred.

Distress/Anger Recall

The methods described so far have used some means of provoking feelings of stress, anger, anxiety, frustration, and so on. A different strategy, however, is to take an idiopathic approach in which the subject describes events and feelings concerning a past stressful situation. The first report of an 'anger-recall' task was provided in a study by Ironson et al. (1992), in which she had subjects recall life events that had greatly angered them, as a means of recreating a state of anger in the laboratory. An advantage to this stressor is that it does not depend on the subject's reaction to an artificial task, but allows the experimenter to measure physiological changes in response to a stressor that is real and important for each subject.

Poststress Recovery Phase

Researchers have tended to focus on the cardiovascular changes that occur during exposure to a stressor, and therefore many studies of cardiovascular psychophysiology end after the task is presented. However, it is often useful to have the subject sit for a second rest period after the task phase, as recovery of the prestress baseline may yield valuable data. The CVR paradigm has been extended in recent years because data suggest the return to the prestress resting level may also have a role in predicting cardiovascular disease. In a meta-analysis, Hocking-Schuler and O'Brien (1997) reported that persons at high risk for hypertension showed poorer BP recovery than low-risk persons, and several studies have shown that recovery of BP following exposure to a pressor stimulus such as the cold pressor test is slower in hypertensives than in normotensives (Hines and Brown, 1936; Hocking-Schuler and O'Brien, 1997). In addition, we and others have shown that children of hypertensive parents have poorer BP recovery than children of normotensives (Gerin and Pickering, 1995). To date, four studies have prospectively investigated the relationship between delayed BP recovery from psychological stressors in the laboratory and the development of hypertension (Borghi et al., 1986; Steptoe and Marmot, 2005; Stewart and France, 2001; Treiber et al., 2001). Each of these studies found that poor BP recovery following mental stress predicted future BP levels; two found that recovery had greater predictive power than CVR (Borghi et al., 1986; Steptoe and Marmot, 2005).

General Methodological Issues

Frequency of Measurements

The frequency of measurement varies with the specific measurement. Thus, heart rate can be measured at smaller intervals than BP (when using an arm cuff), as each BP measurement takes at least a minute and perhaps more. The general concern is the reliability of the measure. Thus, one reason for the importance of the interval size is that it determines the number of measurements that can be taken within a given phase of the study. In a 1993 paper, Gerin et al. (1993) discussed this topic in depth, using BP measurements as an example, providing a useful discussion in the context of psychophysiological studies. Briefly, basing a measure on the average of several measurements reduces the random error associated with the measure, which increases the reliability of the measure.

Duration of Stressor Exposure

Because BP regulation is almost immediately responsive to one's reactions to the environment, the task duration can be quite brief when BP is the outcome. Three measurements can reasonably be taken during a 5-minute stressor; one at the end of minute 1, at the end of minute 3, and at the end of minute 5. However, there may be advantages of a longer duration task, say, one that goes on for 15 min. In terms of measurement reliability, discussed earlier, a longer duration allows more intermittent measurements to be taken. The TSST is an example of a stressor of longer duration than most laboratory stressors.

Probe Measures

Throughout the session, it may be useful to take measures that will help to interpret the physiological outcomes. For example, a simple measure involves asking the subject a question, such as, "How angry do you feel at this moment?" and having the subject use an interval-level scale (for example, 1 equals 'not at all,' 7 equals 'extremely').

Type I Error

Psychophysiological experiments provide many opportunities to take advantage of chance. Every additional outcome measure included, for which there is no strong, a priori hypothesis based on the literature and relevant pilot data, adds to the experimentwise error rate. For example, BP is often measured as both systolic and diastolic pressure, with each outcome treated in a separate analysis; however, rarely are hypotheses provided for when each measure, on its own, should be affected by the stressor; thus, the Type I error rate in that instance is almost double what is reported.

Individual Difference Characteristics

A number of demographic, constitutional, and personality factors that are associated with hypertension and coronary heart disease have also been shown to influence CVR, and should be considered for inclusion as covariates when

conducting CVR studies. An exhaustive list of these is beyond the scope of this article; however, it is worth mentioning that family history of hypertension affects reactivity although both positive (Lawler et al., 1998) and negative (Muldoon et al., 1993) findings have been reported. A meta-analysis of 48 studies (Pierce et al., 2005) found that individuals with a positive family history showed increased CVR to a wider range of stressors compared with those without a family history, but not a larger magnitude of reactivity overall.

Ethnic background has also been studied as a determinant of CVR. The majority of studies that have examined black–white differences to standard laboratory stressors such as mental arithmetic have shown that both black children and black adults exhibit larger cardiovascular responses to these stressors (Anderson, 1989; Brondolo et al., 2003). Based on the hypothesis that exposure to racism may partially explain differences in hypertension (e.g., Krieger, 1990), a number of CVR studies have used a stressor that involves racial discrimination. The results have been inconclusive, with some studies showing blacks having greater reactivity to the stressor, and other studies showing null results (Brondolo et al., 2003; Harrell et al., 2003; Lepore et al., 2006).

Gender. Studies reporting gender comparisons have tended to show that men exhibit greater systolic BP reactivity to stress, while women tend to have greater heart rate responses (Allen et al., 1993; Kudielka et al., 2004; Stoney et al., 1987). Although few studies of menopausal status have been conduced, the data that do exist suggest that postmenopausal women tend to have greater heart rate and systolic BP reactivity than premenopausal women (Bairey Merz et al., 1998; Saab et al., 1989; see Rose et al., 2004; for an exception).

Socioeconomic Status (SES). There is evidence indicating that higher SES, reflected by higher educational attainment or occupational grade, is associated with increased CVR to mental stress (Carroll et al., 1997; Musante et al., 2000; Rose et al., 2004). This relationship is complex, however, with a number of mediators and moderators of the SES–CVR association having been reported (e.g., Armstead et al., 1994; Chen and Matthews, 2001; Merritt et al., 2004).

Summary and Conclusion

CVR continues to be a dominant biopsychosocial approach to understanding the effects of environmental stressors as they affect cardiovascular disease. The greater challenge, however, is to design studies that us to understand the roles that reactivity and recovery models may play in the pathogenic chain leading to hypertension and heart disease.

The reactivity model has evolved greatly over the past three decades. Researchers are investigating processes never envisioned by the original theorists, with a focus on those that appear to be implicated in the pathogenicity that causes cardiac events, including platelet aggregation and endothelial dysfunction. In addition, the changes in these many parameters, as a function of mental stress, are also being examined with indicators of dysregulation in other biological systems, including immune function and the hypothalamic pituitary axis. The behavioral genetics of acute cardiovascular responses is being elaborated, and sophisticated techniques employed to identify the genes implicated in the observed individual differences in cardiovascular reactivity, and the subsequent cardiac outcomes.

Even with its 80-year history of study, the role of CVR as a cause of cardiovascular disease remains controversial, and more study is needed before strong conclusions can be drawn. The ultimate question, of course, is whether or not intervening to attenuate CVR will reduce the likelihood of developing hypertension and coronary heart disease.

See also: Coronary Heart Disease: Psychosocial Aspects; Environmental Stress and Health; Hypertension: Psychosocial Aspects; Race: Ethnicity and Health; Socioeconomic Status and Health; Startle Reflex and Health; Stress and Cardiac Response.

Bibliography

Allen, M.T., Stoney, C.M., Owens, J.F., Matthews, K.A., 1993. Hemodynamic adjustments to laboratory stress: the influence of gender and personality. Psychosomatic Medicine 55, 505–517.

Anderson, N.B., 1989. Racial differences in stress-induced cardiovascular reactivity and hypertension: current status and substantive issues. Psychological Bulletin 105, 89–105.

Armstead, C.A., Anderson, N.B., Lawler, K.A., 1994. The Interaction of Socioeconomic Status, Ethnicity and Cardiovascular Reactivity Among Women. Presented at the Annual Meeting of the American Psychosomatic Society, Boston.

Bairey Merz, C.N., Kop, W., Krantz, D.S., Helmers, K.F., Berman, D.S., Rozanski, A., 1998. Cardiovascular stress response and coronary artery disease: evidence of an adverse postmenopausal effect in women. American Heart Journal 135, 881–887.

Borghi, C., Costa, F., Boschi, S., Mussi, A., Ambrosioni, E., 1986. Predictors of stable hypertension in young borderline subjects: a five-year follow-up study. Journal of Cardiovascular Pharmacology 8 (Suppl. 5), S138–S141.

Brondolo, E., Rieppi, R., Kelly, K.P., Gerin, W., 2003. Perceived racism and blood pressure: a review of the literature and conceptual and methodological critique. Annals of Behavioral Medicine 25, 55–65.

Carroll, D., Smith, G.D., Sheffield, D., Shipley, M.J., Marmot, M.G., 1997. The relationship between socioeconomic status, hostility, and blood pressure reactions to mental stress in men: data from the Whitehall II study. Health Psychology 16, 131–136.

Carroll, D., 2011. A brief commentary on cardiovascular reactivity at a crossroads. Biological Psychiatry 86, 149–151.

Chen, E., Matthews, K.A., 2001. Cognitive appraisal biases: an approach to understanding the relation between socioeconomic status and cardiovascular reactivity in children. Annals of Behavioral Medicine 23, 101–111.

Chida, Y., Steptoe, A., 2010. Greater cardiovascular responses to laboratory mental stress are associated with poor subsequent cardiovascular risk status: a meta-analysis of prospective evidence. Hypertension 55, 1026–1032.

Dickerson, S.S., Kemeny, M.E., 2004. Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. Psychological Bulletin 130, 335–391.

Gerin, W., Pickering, T.G., 1995. Association between delayed recovery of blood pressure following acute mental stress and parental history of hypertension. Journal of Hypertension 13, 603–610.

Gerin, W., Pieper, C., Marchese, L., Pickering, T.G., 1993. The reliability of cardiovascular change scores: a comparison of intermittent vs. continuous methods. Journal of Psychosomatic Research 37, 1–9.

Gerin, W., Rosofsky, M., Pieper, C., Pickering, T.G., 1994. A test of generalizability of cardiovascular reactivity using a controlled ambulatory procedure. Psychosomatic Medicine 56, 360–368.

Goyal, T.M., Shimbo, D., Mostofsky, E., Gerin, W., 2008. Cardiovascular stress reactivity. In: Luecken, L.J., Gallo, L.G. (Eds.), Handbook of Physiological Research Methods in Health Psychology. Sage, Thousand Oaks, pp. 133–158.

Harrell, J.P., Hall, S., Taliaferro, J., 2003. Physiological responses to racism and discrimination: an assessment of the evidence. American Journal of Public Health 93, 243–248.

Harshfield, G.A., James, G.D., Schlussel, Y., Yee, L.S., Blank, S.G., Pickering, T.G., 1988. Do laboratory tests of blood pressure reactivity predict blood pressure variability in real life? American Journal of Hypertension 1, 168–174.

Hines, E.A., Brown, G.E., 1932. A standard stimulus for measuring vasomotor reactions: its application in the study of hypertension. Proceedings of the Staff Meeting of the Mayo Clinic 7, 332–335.

Hines, E.A., Brown, G.E., 1936. The cold pressor test for measuring the reactibility of the blood pressure. Data concerning 571 normal and hypertensive subjects. American Heart Journal 11, 1–9.

Hocking-Schuler, J.L., O'Brien, W.H., 1997. Cardiovascular recovery from stress and hypertension risk factors: a meta-analytical review. Psychophysiology 34, 649–659.

Ironson, G.H., Gellman, M.D., Spitzer, S.B., Llabre, M.M., Pasin, R.D., Weidler, D.J., Schneiderman, N., 1989. Predicting home and work blood pressure measurements from resting baselines and laboratory reactivity in black and white Americans. Psychophysiology 26, 174–184.

Ironson, G., Taylor, C.B., Boltwood, M., Bartzokis, T., Dennis, C., Chesny, M., et al., 1992. Effects of anger on left ventricular ejection fraction in coronary artery disease. American Journal of Cardiology 70, 281–285.

Kamarck, T.W., Debski, T.T., Manuck, S.B., 2000. Enhancing the laboratory-to-life generalizability of cardiovascular reactivity using multiple occasions of measurement. Psychophysiology 37, 533–542.

Kamarck, T.W., Schwartz, J.E., Janicki, D.L., Shiffman, S., Raynor, D.A., 2003. Correspondence between laboratory and ambulatory measures of cardiovascular reactivity: a multilevel modeling approach. Psychophysiology 40, 675–683.

Kamarck, T.W., Schwartz, J.E., Shiffman, S., Muldoon, M.F., Sutton-Tyrrell, K., Janicki, D.L., 2005. Psychosocial stress and cardiovascular risk: what is the role of daily experience? Journal of Personality 73, 1749–1774.

Kirschbaum, C., Pirke, K.M., Hellhammer, D.H., 1993. The trier social stress test – a tool for investigating psychobiological stress responses in a laboratory setting. Neuropsychobiology 28, 76–81.

Krieger, N., 1990. Racial and gender discrimination: risk factors for high blood pressure? Social Science & Medicine 30, 1273–1281.

Kudielka, B.M., Buske-Kirschbaum, A., Hellhammer, D.H., Kirschbaum, C., 2004. Differential heart rate reactivity and recovery after psychosocial stress (TSST) in healthy children, younger adults, and elderly adults: the impact of age and gender. International Journal of Behavioral Medicine 11, 116–121.

Lawler, K.A., Kline, K., Seabrook, E., Krishnamoorthy, J., Anderson, S.F., Wilcox, Z.C., et al., 1998. Family history of hypertension: a psychophysiological analysis. International Journal of Psychophysiology 28, 207–222.

Lepore, S.J., Revenson, T.A., Weinberger, S.L., Weston, P., Frisina, P.G., Robertson, R., et al., 2006. Effects of social stressors on cardiovascular reactivity in black and white women. Annals of Behavioral Medicine 31, 120–127.

Linden, W., Earle, T.L., Gerin, W., Christenfeld, N., 1997. Physiological stress reactivity and recovery: conceptual siblings separated at birth? Journal of Psychosomatic Research 42, 117–135.

Manuck, S.B., 1994. Cardiovascular reactivity in cardiovascular disease: once more unto the breach. International Journal of Behavioral Medicine 1, 4–31.

Merritt, M.M., Bennett, G.G., Williams, R.B., Sollers, J.J., Thayer, J.F., 2004. Low educational attainment, John Henryism, and cardiovascular reactivity to and recovery from personally relevant stress. Psychosomatic Medicine 66, 49–55.

Muldoon, M.F., Terrell, D.F., Bunker, C.H., Manuck, S.B., 1993. Family history studies in hypertension research: review of the literature. American Journal of Hypertension 6, 76–88.

Musante, L., Treiber, F.A., Kapuku, G., Moore, D., Davis, H., Strong, W.B., 2000. The effects of life events on cardiovascular reactivity to behavioral stressors as a function of socioeconomic status, ethnicity, and sex. Psychosomatic Medicine 62, 760–767.

Pierce, T.W., Grim, R.D., King, J.S., 2005. Cardiovascular reactivity and family history of hypertension: a meta-analysis. Psychophysiology 42, 125–131.

Rose, K.M., North, K., Arnett, D.K., Ellison, R.C., Hunt, S.C., Lewis, C.E., Tyroler, H.A., 2004. Blood pressure and pulse responses to three stressors: associations with sociodemographic characteristics and cardiovascular risk factors. Journal of Human Hypertension 18, 333–341.

Saab, P.G., Matthews, K.A., Stoney, C.M., McDonald, R., 1989. Premenopausal and postmenopausal women differ in their cardiovascular and neuroendocrine responses to behavioral stressors. Psychophysiology 26, 270–280.

Steptoe, A., Marmot, M., 2005. Impaired cardiovascular recovery following stress predicts 3-year increases in blood pressure. Journal of Hypertension 23, 529–536.

Stewart, J.C., France, C.R., 2001. Cardiovascular recovery from stress predicts longitudinal changes in blood pressure. Biological Psychology 58, 105–120.

Stoney, C.M., Davis, M.C., Matthews, K.A., 1987. Sex differences in physiological responses to stress and in coronary heart disease: a causal link? Psychophysiology 24, 127–131.

Treiber, F.A., Musante, L., Kapuku, G., Davis, C., Litaker, M., Davis, H., 2001. Cardiovascular (CV) responsivity and recovery to acute stress and future CV functioning in youth with family histories of CV disease: a 4-year longitudinal study. International Journal of Psychophysiology 41, 65–74.

Turner, J.R., Ward, M.M., Gellman, G.A., Harshfield, G.A., Johnston, D.W., Light, K.C., van Doornen, L.J.P., 1994. The relationship between laboratory and ambulatory cardiovascular reactivity: current evidence and suggested future directions. Annals of Behavioral Medicine 16, 12–23.

Zanstra, Y.J., Johnson, D.W., 2011. Cardiovascular reactivity in real life settings: measurement, mechanisms and meaning. Biological Psychiatry 86, 98–105.

Cardoso, Henrique (1931–)

Diego Pereyra, Instituto de Investigaciones Gino Germani, Universidad de Buenos Aires – CONICET, Argentina

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Fernando Henrique Cardoso is one of the most prominent Latin American intellectuals, since both his sociological contributions in topics such as social change, democracy, and modernization and his recognition as political leader, reaching the presidency of Brazil. His ideas on dependency theory gave him nevertheless an international recognition. This article, firstly, presents a reconstruction of his lifetime in terms of his academic and political actions. It firstly examines his education and main sociological ideas, revising his main works. Secondly, it offers a summarized picture of theory of dependency, its authors and different perspectives, indicating some themes of disagreement. Finally, the article discusses some reasons that could explain the success of Cardoso's ideas and specific points of recent and likely future debate on Cardoso's thinking; specially wondering about what he changed his vision on social change and some controversies in use of classical Marxism.

Introduction

Fernando Henrique Cardoso is one of the most renowned Latin American intellectuals, since both his sociological contributions and his recognition as political leader. Cardoso reached top positions in academic and political fields. He founded and also chaired the *Centro Brasileiro de análise e planejamento* (CEBRAP); he was the president of the International Sociological Association (ISA) as well and developed an extensive career in Brazilian politics as senator and national minister, a calling that included at the end the presidency of the country. Cardoso's influence and legacy is very important, not only in Latin America but also worldwide, in topics such as social change, democracy, and modernization. Though, his ideas on dependency theory gave him an international recognition.

Cardoso was born in Rio de Janeiro in 1931. He was the eldest of three children in a family with military tradition. His paternal relatives had strong political links and had participated in previous revolts and revolutions. However, his father became a lawyer when he retired from the army and was elected a federal deputy defending combinatory leftist, liberal, and nationalist viewpoints. When the son entered Secondary School, the family moved from the cosmopolitan city of Rio to the then more bucolic atmosphere of São Paulo. This city was at that moment changing from a rural alike place to a developed urban area, a process driven by the modernization of local elites. It would become very rapidly the center of industrialization of Brazil and one of the biggest metropolises in the world, which is today. This was the location where the teenager Cardoso watched profound social transformations and processed the eclectic ideas he absorbed everyday at home and school (Kahl, 1976: 130–131).

Brazil had not at the 1940s a long-term academic tradition; universities were founded lately. This country was a case in which sociology institutionalized very late but professionalization of the discipline happened relatively promptly, in contrast with Argentina, for instance, in which it was the opposite (Pereyra, 2010). Teaching of sociology at the universities started in the 1930s at the level of degree studies and in 1950s at the postgraduate level (Porto and Dwyer, 2010). Nevertheless, sociology was not absolutely a novel field in Brazil before 1930. It was included in secondary teaching from the 1920s to 1942 (certainly from a strong positivist standpoint), and particularly São Paulo's schools openly welcomed that disciplinary knowledge.

One key institution was the University of São Paulo (USP). It was established in 1934 on the basis of the *Escola Livre de Sociologia e Política de São Paulo* (Free School of Sociology and Political Science), created during the previous year by donors from local emerging business class. It was clearly a step in the path of modernizing the state and educational system of Brazil. The USP rapidly combined several independent research and higher educational institutions and hired professors coming from France, Italy, Spain, Germany, and other European countries. The *Escola Livre* remained attached to the university but kept some intellectual and administrative autonomy. This led to a successful institutional development. The USP became a central academic actor in the country. It created an intellectual revolution, especially in social sciences field; meanwhile, the *Escola* organized postgraduate programs and strongly contributed in the institutionalization of sociology in the region (Limongi, 1995). Cardoso's education was plainly marked up by the experience of his studies at the USP.

Education, Traditions, and Early Training

Fernando H. Cardoso registered in 1949 as a university student at the Social Sciences Department. Curiously, he first applied for entering into Legal Studies Area (probably following a family mandatory) but he failed a test in Latin. This examination's adverse result changed thus all his vocational career and future. After changing option, he witnessed in the university an interesting debate about the nature of sociology and the need of renovating teaching and research. While the sociology courses at the *Escola Livre* followed the empirical model (mainly tutored by the American sociologist, Donald Pierson), the teaching of sociology at the USP was more inspired by French intellectual tradition. Cardoso was immersed in the discussion and learned from both sides (Kahl, 1976; Limongi, 1995).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61318-1

Two decisive names who taught at the USP influenced Cardoso's student ideas: Roger Bastide and Florestán Fernándes. The first was a French sociologist (1898–1974) who specialized in literature, social history, and anthropology, giving special attention to religion and conscience. The second was a local professor (1920–95) who was renewing sociological tradition in Brazil. Although Cardoso was just a decade younger than Fernándes, the former found an inspirational mentor in the latter. Cardoso became an obedient disciple, learning from the Fernándes' lessons different authors from German sociological tradition (Weber, Marx, Mannheim), reflecting upon the relationship between economy and culture, and discussing the impact of social structure in social and individual actions. But these continental and classical theories were jointly studied with an empirical emphasis inspired by aforementioned Pierson's ideas (Arruda, 1995). Cardoso always remarked that Fernándes was very influential in modeling his own sociological framework (Kahl, 1976: 130). Their careers separated (not openly, certainly) much later, nevertheless, since they followed diverse political ways in terms of political preferences and party participation. Probably, diverse social backgrounds (a deprived home in Fernándes' case and a prosperous middle-class childhood in Cardoso's one) could partially explain the path but this is not a sufficient explicative.

While still a student Cardoso took part in an intellectual debate, which was probably foundational to the establishment of modern sociology in Brazil. It was the topic coined as the 'Negro question,' in which local intellectuals from diverse disciplines tried to understand the social integration of Brazilian people with African origin and the possible emerging of a unique national identity. That idea became a well-funded UNESCO project on the position of Negroes in Brazil. The aim of the sponsors was to prove that Brazilian society was an unproblematic melting pot. Fernándes was interested in participating in the research because he wanted to show the wrongness of that idea. So, he invited Cardoso to join himself, Bastide and all the investigative staff, so the advanced students collaborated in fieldwork and data analysis (Kahl, 1976: 131; Platt, 2010a). The voluminous collected material was the basis for the writings from professors and a whole cohort of young sociology students at the USP, including Cardoso as likely the most active of them.

Cardoso began lecturing both economics and sociology at the USP even before graduating. He received a degree in sociology from the same university in 1952 and a master degree one year later when he was appointed as teaching assistant, position held until 1961 (Kahl, 1976: 131). Cardoso's main research interests at that time were to deal with the issue of slavery and the situation of *Negros* in society. That question connected empirical studies with the national issue. Consequently, he realized gradually that their status was based on their class position within a capitalist system.

His first book on this research was published jointly with Octavio Ianni (1960) *Cór e mobilidade social em Florianópolis* (Race and social mobility in Florianopolis), in which the authors specially studied social obstacles that stopped and closured social change and upright mobility of Negros in certain regions of Brazil, scrutinizing slavery and salary regimes in economies oriented both for internal and external markets. This first book clearly displayed the legacy he acknowledged from Bastide and Fernándes, but also from Antonio Cándido, another key figure of sociology at the USP (Jackson, 2008). The second publication was Cardoso's doctoral dissertation (1962), *Capitalismo e escravidão no Brasil meridional* (Capitalism and slavery in Southern Brazil); which was supervised by Fernándes himself. Both works showed that his research method was definitely established: a combination of ethnographic field strategies (interviews and direct observation) and historical data analysis, all oriented by a vision on society's plenty of structuralism. At that point, Cardoso was convinced that his sociology could connect the singularity of Brazilian society, Marxist debate, and national politics.

In 1961, he was promoted to full professor and organized a research unit at the USP aimed to study problems on industrial sociology, taking ideas from the Laboratory of Industrial Sociology founded by Alain Touraine in Paris. Cardoso visited this city for some months, changing ideas and experiences with Touraine, Raymond Aron, and Michael Crozier (Platt, 2010a). After returning home, the political situation of Brazil troubled. In 1964, there was a military coup. The new administration blacklisted him from teaching and he had to hide initially and later, when family connections were insufficient to keep him safe, he fled to Argentina. After few weeks, he moved once again to Santiago de Chile, having been called for a new and promising intellectual project.

Cardoso was hired by the *Instituto Latinoamericano de Planificación Económica y Social* (ILPES, Latin American Institute for Economic and Social Planning), which was a research body recently created by the UNESCO in association with the Economic Commission for Latin America (CEPAL, as usually known by its Spanish acronym). When established, the ILPES was conceived as an organization oriented to train bureaucracy all over Latin America, strengthening its capacity for both the formulation and application of finance plans and the development of investment policies. That was the context of the so-called Alliance for Progress pushed up by Kennedy administration. But the agenda changed since a new and different intellectual project emerged (Gabay, 2010).

Chile was a crucial pathway in the production of social knowledge at that time. Not only CEPAL had headquarters in Santiago, also the Latin American School of Social Sciences (FLACSO) had been founded in that city by UNESCO. Those institutions contributed all together to both the reinforcement of Latin American identity and the institutionalization of local social sciences (Beigel, 2013). Cardoso found in that country a brainstorming atmosphere: local minds, Latin American scholars, enduring contacts with Paris (that meant keeping links with Touraine), and a group of Brazilian intellectuals who also were pushed into exile (Celso Furtado, Francisco Weffort, Theotonio Dos Santos, Vania Bambirra, among others), with which would redefine the question of national development and social change in the region.

CEPAL was one of the five regional commissions of the United Nations. It was founded in 1948 for the purposes of contributing to the economic and social development of Latin America and coordinating international actions. Under the strong leadership of Raúl Prebish, it profoundly influenced Latin American thought, applying an original approach in the analysis of development: the center–peripheral model

(Dosman, 2008; Bielschowsky, 2009). Cardoso realized very promptly that the theory of development coined by CEPAL should be questioned since the explanation was based excessively in economic factors. After studying the case of entrepreneurs in Argentina and Brazil, Cardoso started understanding the role of elites in stopping or blocking development in peripheral areas, giving special attention to culture and ideological factors (Cardoso, 1971). With his Chilean colleague, Enzo Faletto, he opened a reflection upon a new idea on social change. It would be the internationally famous theory of dependency, which responded to the then existing sociological and economic theories that were unsatisfactory explanations about development in Latin America.

Theory of Dependency and International Actions

After circulating for some time in manuscript and discussed massively in classes and seminaries in Santiago, Rio, and Buenos Aires, their work was finally published as *Dependencia y Desarrollo en America Latina* (1969a, Dependency and Development in Latin America), having a countless impact. The same year, Cardoso sent to press many drafts of his works during his period in Santiago, in which the empirical and hypothetical framework of the new theory could be found (1969b). This theoretical project was not merely an isolated effort since almost the entire research staff at ILPES had been working in related fields. The academic audience understood and was ready to discuss the issue; Cardoso and Faletto were not original but they put the debate in plain words, offering an easy interpretative scheme. The theory was a boom since the field was disenchanted by preceding developmental theories. They were shifting to leftish positions. This intellectual corpus was read in a time that academic agenda was redefining the role, challenges, and commitments of Latin American intellectuals and both anti-colonial and anti-imperialist discourses were emerging (Gilman, 2003).

Dependency theory offered therefore a new sociological definition of underdevelopment combining the analysis of society with economy and politics. It advocated an inward looking approach to development, in contrast with previous CEPAL ideas. Cardoso and Faletto claimed for an integrative and global analysis of social change, focusing both on the singularity of national historical conditions and the goals and interests that oriented social conflict among social groups. Thus, they demanded for a perspective based on the idea of 'historic and social process,' a way of understanding social domination in combinatory terms of politics, social structure, and culture (Cardoso and Faletto, 1969).

After *Dependency and Development* was printed, the debate on this issue expanded. It became the core of intellectual arena and hundreds of papers, chapters, and books were published. Gradually, discussion moved from official agencies to universities and academic meetings, mainly sponsored by the Latin American Sociological Association. In 1977, Cardoso and Faletto needed to rewrite some ideas in response to misunderstandings and confusions and the way that, according to the authors, their colleagues established wrongly a discussion on the idea of dependency. That resulted in a kind of brief sequel of the original book, the *postscriptum of Dependency and development*. Basically, diverse authors have discussed on dependency in two different ways. Some (Vania Bambirra and Ruy Mauro Marini, among others) affirmed that it was entirely a theoretical framework and dependency was a key category of capitalism in underdevelopment societies since it was thought in terms of mode of production. Others, led by Cardoso, stated that the so-called theory of dependency was not exactly an articulated system of ideas but an intellectual approach to a distinctive problem; in other words, a perspective of analysis. Thus, it aimed to design an analytical study on a particular articulation between social classes, economy, and state.

After exiling in Chile, Cardoso returned to Brazil in 1968 and passed an examination that allowed him to teach at the USP. However, the military government put his political rights off and his university contract was canceled. In contrast with his earlier similar experience, he did not move. Although things were not quite calm, he stayed in São Paulo, with the conviction of improving life in Brazil. In conjunction with other lecturers and researchers in the same conditions, Cardoso founded in 1969 a new body which would become his home during the following decade. It was the CEBRAP (Brazilian Centre for Analysis and Planning). It was a private organization sharing a common history of similar institutional experiences in the region during 1970s and early 1980s. Such bodies expanded in Argentina, Chile, and Uruguay; they aimed to gain autonomy from dictatorial political intervention in those countries. Soon, the CEBRAP reached a significant academic prestige. It was well funded by Ford Foundation, had consultancy contracts, and even received resources from national and federal official agencies with which could build up a steady interaction. In some way, it was a model to follow by regional bodies alike: how to get independency and collect money for doing research; its institutional visibility helped its actions to have better impact in local culture and politics (Brunner and Barrios, 1987). Researching at the CEBRAP, Cardoso reinforced his reputation as a critic of the military regime and an advocate of a peaceful transition to democracy, but also established an important network of political and academic links both nationally and globally.

These connections implied an active participation at international bodies. So, Cardoso occupied important positions in the ISA. After holding a vice-presidency in the period 1978–82, he was elected president for the following four-year term. At the moment of the election he was not only a key man within the field but also a well-known name to the general public in Brazil and abroad. His job at the ISA was regarded by his contributions to the development of sociology in peripheral areas; it should remember that 1982 and 1986 ISA congresses were held in Mexico City and New Delhi, respectively; which was a fact of the ISA decision to first time move sociological discussion away from Europe and the US (Platt, 2010a).

It was clear in his agenda that sociology had to follow scientific tradition but it was not a reason for abandoning commitment and ethical values (like justice and fraternity) that orient the actions of sociologists (Cardoso, 1986). His political skills were very helpful to be a good institutional leader, although he often exercised his functions at a distance. Cardoso's successful career at the ISA could be understood tracing back his earlier contacts with UNESCO, which also refreshed the productive French–Brazilian academic

connection. There could have been a Touraine-related trail in some way, but Platt (2010b) suggests some methodological caution in following such trails; the study of those connections should be developed, though.

Political Career

Cardoso was active in local politics since the late 1970s. He started his political career when he was put forward as a candidate for the Senate in 1978. He failed this first attempt. He was elected twice as senator from Sao Paulo, nevertheless, from 1983 to 1992. During his job in the national congress, he was one of cofounders of the centre-left Brazilian Social Democratic Party. In 1992, after President Fernando Collor de Mello was impeached for corruption and was replaced by Itamar Franco, Cardoso was called for duty to become successively Foreign Affairs and Finance minister. Therefore, he had to face two definitive challenges: inflation and social disturbances. After establishing some consensus with the opposition and labor leader Luiz Inácio da Silva (nicknamed Lula), Cardoso coordinated with firm determination the application of a sensible and simple financial plan, named *Plano Real*. The plan resulted at last, stabilizing prices and ordering the economic situation. It was widely credited for bringing Brazil out of economic chaos (Cardoso, 2006).

To take advantage of his political prestige and recent fame of 'storm tamer,' Cardoso entered later the race for and won the presidency in 1994. It was clear that the support of the national business community was crucial for the decisive victory. During his administration, he emphasized further economic reform, including privatization and increased foreign investment, as well as funding for education and social services. In 1997 voters approved an amendment to the constitution that allowed the president to hold consecutive terms and one year later Cardoso became the first Brazilian president to be reelected for a second four-year period. By that time, however, Brazil was facing severe financial problems, and his administration was forced to pursue an austerity plan, which included spending cuts and tax increases, to reduce the budget deficit and to secure international loans.

Probably, one of Fernando Henrique Cardoso's greatest triumphs at presidency was his successful policy to reduce AIDS rates in Brazil. His time in government is also acknowledged by his challenge to pharmaceutical sector and the decision to promote generic cheap drugs in order to combat illness in his country. Barred constitutionally from seeking a third consecutive term, Cardoso left office in 2003, leaving the post to the *Partido dos Trabalhadores* (PT, Labour Party) under the strong leadership of Lula, which is still in power. When delivering the presidential scepter or when hanging his sash up, Cardoso could have recalled that Florestán Fernándes had been one of the main and well-harked officials of the PT.

In addition to the University of São Paulo, of which he is emeritus professor, Cardoso has taught at a vast number of universities: Santiago, Chile, California (Stanford and Berkeley), Cambridge, Paris-Nanterre, the École des Hautes Études en Sciences Sociales, and the Collège de France. From 2003 to 2007, he was professor at the Thomas J. Watson Jr. Institute for International Studies, at Brown University in the United States. Along his career, Cardoso has as well received honorary doctorates from more than 20 universities; he is also an honorary overseas member of the American Academy of Arts and Sciences. He was elected twice (2005 and 2008) as one of the hundred major public intellectuals of the world in a survey conducted by the renowned scholarly periodicals *Prospect* and *Foreign Policy*. Nowadays, he has an active academic and political agenda. He enthusiastically lectures at conferences on topics from economics to AIDS. He also participates at different forums related with international affairs and developmental policies and takes part in different boards of cultural or academic organizations. His intellectual production is really wide. Besides numerous journal articles published in diverse countries, a complete set of his pronouncements and speeches as president of Brazil was published in 16 volumes by the Presidency of the Republic (2002). In *The Accidental President of Brazil* (2006), he presented a good summary of both the political events and a personal insight into the sociology of Brazil.

Debate and Controversies

What remain unclear are the precise reasons of Cardoso's intellectual success. Kahl (1976), Goertzel (1999), and Platt (2010a) suggest that he could define a key question and was always able to keep it at the center of his intellectual worries, offering a sociological explanation historically situated and giving up any pretention to build up a general theory. Kahl (1976: 186–187) additionally advocates that the sociological impact of Cardoso's work is so shocking since his thinking did give out an answer that could fill the gap and bias of existing theories such as functionalism, Marxism, and modernization theory in the explanation of social change in Latin America. A further point to be well-thought out is that dependency theory emerged in a time when an enthusiastic audience was waiting for revelation: the singularity of Latin America in terms of culture and economy should be explained by a singular not imported theory. Formulae from center had evidently failed so it was welcoming a new recipe with local ingredients and flavors. Cardoso was the major chef at that occasion.

Moreover, a debate on Cardoso's ideas was emerging, which could lead to a promising research agenda. Some academic voices questioned Cardoso's policies when he was at ministerial or presidential power. Many accused him of betrayal since his governmental policies seemed to be far away from his previous ideas on Marxism and dependency. This could be considered only a local dispute between Labour and Social Democracy in Brazil. But in fact it is more complex since the discussion could be traced back to the 1970s when the polemics on the nature of dependency peaked. For example, one key author of dependency theory, Theotonio dos Santos, questioned Cardoso in 2010 via a public letter in which the former president was blamed to have applied neoliberal policies and to have blocked both a real industrialization process and the empowerment of working class; such actions are said to be distant from left legacy. It is important to say that dos Santos affirmed Cardoso's shift had already begun 30 years before when they argued each other on the interpretation on social change in the region.

The key point to take into account is whether Cardoso changed or not his sociological mind along his academic and political career. In the first scenario, Cardoso dramatically changed his vision on society. Cardoso had a strong compromise with revolutionary changes in the past and later became a neoliberal politician. Nevertheless, the second picture showed a more invariable intellectual outlook in Cardoso's actions. One should reflect upon the nature of Cardoso's dependency approach, though.

One may reason that his ideas made a rupture with the classical theory of modernization. In that sense, the destiny of Latin America should follow a different and alternative path that the more advanced societies, which meant that the entire idea of development is false and wrong. In that reasoning, the only way to change dependence situation was a social revolution. On the contrary, one might also think that the theory of dependency is a critique of developmental theories but also a continuation; since dependency could be assumed as one singular obstacle to the progress of peripheral societies. Thus, when countries could solve that problem they can establish modern economic and political institutions.

Cardoso seemed to have not renewed much of his ideas since one can observe, he has never abandoned his ideas on democracy and political participation. *Dependency and development* did not question democratic institutions. In a competition between liberal and populist traditions on the idea of democracy, Cardoso seems to belong to the first one and opposed to the second. One could see as well that Cardoso had a liberal standpoint from his first writings; he believed all nations follow the same civilizatory process, although found different obstacles and blockages. Dependency was one that distinguished Latin American countries from other national experiences. Hence, Cardoso was more familiar to Gino Germani's vision (1979) on modernization than it was considered before.

The issue of the use of Marx seems to be crucial. Here, Cardoso's ideas are likely related with the interaction between liberalism and Marxism, an intellectual project raised at the same time by names such as Anthony Giddens and Jürgen Habermas. Cardoso's thinking looks like to admit an international use of Marx's ideas in contrast with the other side of dependency theorists who aimed to find a Latin American Marxist approach. Cardoso's thinking could be as well placed in the context of a new interpretation on Marxism and the redefinition of democracy in Latin America during the early 1980s, just the time Cardoso himself was beginning his political career. The research agenda shifted then from development and dependency to an examination of political, economic, and cultural conditions that promoted democracy in the region (Lechner, 1988; Pereyra, 2010: 219).

But the main theme tune of debate is the relationship between social classes and social change. While classical Marxism emphasizes the role of proletariat in the revolutionary process, Cardoso's ideas seems to run down such actions from the masses, which was the base of dos Santos' critique. Contrarily, Cardoso's theory of dependency appears to be appropriate to study social elites since, according to him, would be the group which promotes social change by moving dependency conditions and redefining state autonomy. That stress on both the autonomy of upper class and the importance of cultural and ideological factors was an oddity in Classical Marxism but was not so strange in the region where Gramscian ideas became popular. It is necessary to recall the importance of Gramsci in Brazil, as well in Argentina. But this vision could also help in understanding why Brazilian bourgeoisie have enthusiastically read Cardoso's ideas and backed his leadership.

The question is open and debate will continue. Although, it complicates even more if you add to the scene the political parties' competition within Brazilian context and the disputants' social capitals. To sum up, Cardoso' ideas are fecund and creative, giving an original vision of social change in Latin America and the conditions that could block development and democracy in those countries, rethinking the idea of social structure and modernity, served with liberalism, spiced with Marxism, and well seasoned with classical sociology, though.

See also: Dependency Theory; Development and the State; Development: Social-Anthropological Aspects; Elites: Sociological Aspects; Marxism in Contemporary Sociology; Modernization and Modernity in History; Modernization; Planning and Development; Social Changes: Models; Social Democracy.

Bibliography

Arruda, María Arminda, 1995. A sociologia no Brasil. Florestán Fernándes e a "escola paulista", in Sergio Miceli (org) Historia das Ciencias Sociais no Brasil, Sumare, IDESP. São Paulo II, 107–231.

Beigel, Fernanda, 2010. Dependency analysis: the creation of new social theory in Latin America. In: Patel, Sujata (Ed.), The ISA Handbook of Diverse Sociological Traditions. Sage, London, pp. 189–200.

Beigel, Fernanda (Ed.), 2013. The Politics of Academic Autonomy in Latin America. Ashgate, London.

Bielschowsky, Ricardo, 2009. Sixty years of ECLAC: structuralism and neo-structuralism. CEPAL Review 97, 171–192.

Brunner, Joaquín, Barrios, Alicia, 1987. Inquisición, mercado y filantropía, Ciencias sociales y autoritarismo en Argentina, Brasil, Chile y Uruguay. FLACSO, Santiago de Chile.

Cardoso, Fernando H., 1962. Capitalismo e escravidão no Brasil meridional. Capitalism and slavery in Southern Brazil. Difusão Européia do Livro, São Paulo.

Cardoso, Fernando H., 1969. Mudancas sociais na America Latina. Social transformations in Latin America. Difusão Européia do Livro, São Paulo.

Cardoso, Fernando H., 1971. Ideologías de la burguesía industrial en sociedades dependientes (Argentina y Brasil). Ideologies of industrial bourgeoisie in dependent societies…. Siglo XXI, Mexico.

Cardoso, Fernando H., August 1986. Opening speech at the XIth World Congress of Sociology. New Delhi. http://www.isa-sociology.org/about/presidents/welcome_address_f_cardoso.pdf.

Cardoso, Fernando H., 2006. The Accidental President of Brazil. A Memoir with Brian Winter. Public Affairs, New York.

Cardoso, Fernando H., Ianni, Octavio, 1960. Cór e mobilidade social em Florianópolis. Race and social mobility in Florianopolis. Companhia Editora Nacional, São Paulo.

Cardoso, Fernando, Faletto, Enzo, 1969. Dependencia y Desarrollo en América Latina. Siglo XXI, México, corrected edition, 1978. [Dependency and development in Latin América]. University of California Press, Berkeley.

Dos Santos, Theotonio, 2010. Carta abierta a Fernando Henrique Cardoso. Ciencias Sociales hoy.

Dosman, Edgar J., 2008. The Life and Times of Raúl Prebisch, 1901–1986. McGill-Queen's University Press, Montreal.

Gabay, E., 2010. El "fantasma" de Prebisch: el ILPES entre 1963 y 1969. In: Pereyra, Diego (Ed.), El desarrollo de las ciencias sociales. Tradiciones, actores e

instituciones en Argentina, Chile, México y Centroamérica. FLACSO, San José de Costa Rica, pp. 73–97.
Germani, Gino, 1979. The Sociology of Modernization: Studies on Its Historical and Theoretical Aspects with Special Regard to the Latin American Case. Transaction, New Brunswick.
Gilman, Claudia, 2003. La pluma y el fusil. Siglo XXI, Buenos Aires.
Goertzel, Ted George, 1999. Fernando Henrique Cardoso: Reinventing Democracy in Brazil. Lynne Rienner Publishers, Boulder.
Jackson, Luiz Carlos, 2008. Pioneering generations in são paulo sociology (1934–1969). Tempo Social 4 selected edition, São Paulo 2008.
Kahl, Joseph A., 1976. Modernization, Exploitation and Dependency in Latin America. Germani, González Casanova and Cardoso. Transaction, New Bruswick.
Lechner, Norbert, 1988. Los patios interiores de la democracia. FLACSO, Santiago.
Limongi, Fernando, 1995. A escola libre de sociología e política en São Paulo. In: Historia das Ciencias Sociais no Brasil, op. cit, I, pp. 257–274.
Platt, Jennifer, 2010a. Fernando Henrique Cardoso. published at the ISA Web page.
Platt, Jennifer, 2010b. ISA Presidents and Their Presidential Addresses: Introduction. published at the ISA Web page.
Pereyra, Diego, 2010. Dilemmas, challenges and uncertain boundaries of Argentinian Sociology. In: Patel, S. (Ed.), The ISA Handbook of Diverse Sociological Traditions, op. cit, pp. 212–222.
Porto, Maria Stela Grossi, Dwyer, Tom, 2010. Development, dictatorship and re-democratization- trajectories of Brazilian sociology. In: Patel, S. (Ed.), The ISA Handbook of Diverse Sociological Traditions, op. cit, pp. 201–211.

Relevant Websites

http://cebrap.org.br/v2/ – Cebrap.
www.ifhc.org.br – Fundación Fernando Henrique Cardoso.

Career Development: Multilevel Perspective

Yehuda Baruch, University of Southampton, Southampton, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A.B. Weinert, volume 3, pp. 1471–1476, © 2001, Elsevier Ltd.

Abstract

In this article, I discuss careers from various perspectives, in particular, as an individual phenomenon, organizational system, and society in general. I discuss the changing nature of careers and how they became more dynamic, in line with the wider business environment. Exploring labor markets and careers as an ecosystem I explore factors influencing careers and the nature of career planning and management. I then move to career outcomes, again, at the three level of analysis: individual outcomes, organizational outcomes, and national outcomes. In terms of practical management of careers, I list the various practice organizations that may employ to deal with career planning and management and how they can be integrated. I refer to issues that associate careers with other elements of human resource management, such as global careers, labor markets, work–life balance, and the issue of diversity. Finally, I present a view of the shape of the things to come in the future of careers.

Careers: Individual and Organizational Perspectives

Career is "an evolving sequence of a person's work experience over time" (Arthur et al., 1989) or "a process of development by employee along a path of experience and roles in one or more organizations" (Baruch and Rosenstein, 1992). Career is a salient element in peoples' working life, hence life in general, as work life and private life are interwoven to a high degree. Careers are managed simultaneously by individuals and their employers, and take place in increasingly dynamic (Hall, 2002) and global labor markets (Tams and Arthur, 2007). Already early seminal works highlighted the micro- and macroframes of references for studying careers. One example is by Schein (1978), who combined both individual and organizational perspectives. Career studies originating from the individual perspective tend to deal with career choice (most notably Holland, 1959), and career stages or life cycles (Super, 1957). From organizational career system perspective, career systems deal with identifying, planning, developing, and managing people throughout their employment within the organization.

A distinction typically made within career studies is between individual-focused and organizational/cultural-focused perspectives. I argue that a better approach would realize the complementariness between the two perspectives. Rather than distinction, there is certain convergence in the literature, pointing out that to reach comprehensive understanding of careers, it is necessary to place it within a wider context. Balancing individual and organizational needs is a critical element needed in career scholarship (Baruch, 2006; Lips-Wiersma and Hall, 2007). Earlier, Herriot and Pemberton (1996) offered a career model that places in its core the prerequisite for a match between individual and organizational needs, wants, and provisions. A system that neglects the fulfillment of such a match will see star performers leaving, lack of ability to attract talent, and overall diminishing commitment and investment. It can be the case for organizations (Kets de Vries and Balazs, 1997), nation (Brain Drain phenomenon – Dickmann and Baruch, 2011), and the wider society (Baruch, 1998).

A broader perspective places career systems within the wider environment – professional, organizational, and cultural realms. These provide the values, norms, and framework for individuals to generate and follow their aspiration, and to benchmark against others (Baruch, 2004a: p. 4).

Yet, there is more to it than individual and organizational relationships. Other actors within the wider systems are significant stakeholders. Governments, unions, professional associations, legal systems, and regulations, all play part in shaping career and career development within and outside organizations. The wider labor market, which is becoming global, can be described as an Ecosystem (Baruch, 2013; Higgins, 2005). Ecosystem is "A system that contains a large number of loosely coupled (interconnected) actors who depend on each other to ensure the overall effectiveness of the system" (Iansiti and Levien, 2004). The career ecosystem operates in a constellation that is not merely economical but also part of social-political phenomenon in a global environment (Gunz et al., 2011). There is a significant 'flow of talent' across organizations, sectors, and nations. People chose to have new, second careers (Baruch and Quick, 2007), to move to a different location or country, either for a limited period or permanently (Carr et al., 2005; Mahroum, 2000). More broadly, people are more open to, or accept the necessity of changes. This goes hand in hand with the individualization of societies, both in western and in eastern societies.

The Changing Nature of Career Systems and Emerging Career Models

New career models emerge, some are revolutionary, and others are evolutionary. These new models developed alongside the traditional model, not replacing it. **Figure 1**, titled traditional alongside multidirectional model, presents what Peiperl and Baruch (1997) termed the postcorporate career, where organizations are changing their philosophy and strategy regarding people management. As a result, their structure and career systems are changing too.

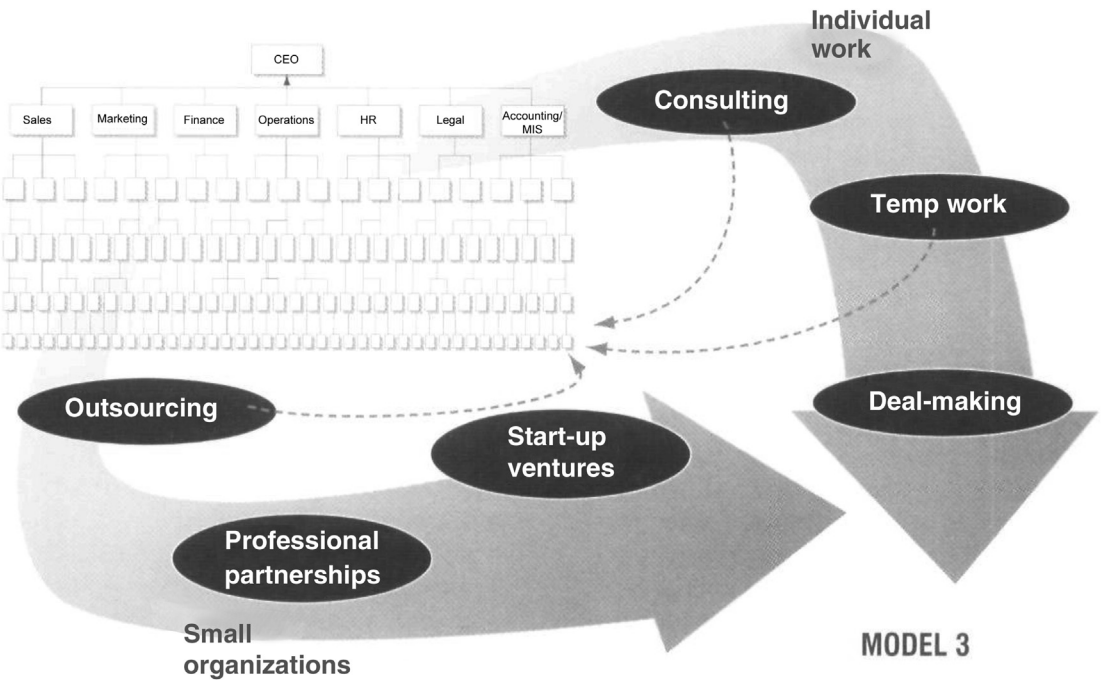

Figure 1 Traditional alongside multidirectional model.

The change in the career's realm is not a singular phenomenon. It should be taken in the perspective of other changes in life generally and management particularly.

Traditional career systems assumed that such treatment will last through the entire career cycle of the employee, whereas contemporary career systems recognize the dynamic and transitional nature of the system (Sullivan, 1999; Baruch, 2004a). Until the 1980s, career structures traditionally focused on advancing in the hierarchy, typically climbing vertical ladders. Even though it was not a reality for many, as the majority hit a career plateau at certain stage, it was still perceived as the ideal type of career – secured, within single organization, with upward mobility. It was in line with conventional perception that a successful career is built out of a set of successive linear moves up the organizational structure, accompanied by increments in formal authority, prestige, and rewards (De Vos and Soens, 2008). The shape of career systems has changed into a boundaryless (Arthur and Rousseau, 1996) and multidirectional career (Baruch, 2004b), where people may hold different career attitudes, such as the protean career (Hall, 1996, 2004). There is certain conceptual ambiguity associated with the new terminology. For example, the protean and boundaryless career concepts overlap significantly, but the protean career concept predominantly focuses on an individual's motives to follow a particular career path, and to take their destination in their own hands, whereas the boundaryless career concept mainly concerns different forms of mobility (Gubler et al., 2013).

Below I list some of the concepts that gained attention in the development of the so-called 'new careers' (Arthur et al., 1999). One caveat to mention is that it is not 'all change,' and as Inkson et al. (2012) argue, the new careers do not necessarily reflect true change, as many, if not most of careers are still more traditional rather than, for example, boundaryless or protean. Current empirical studies (Rodrigues and Guest, 2010; Kattenbach et al., 2013) indicate that there is indeed more rhetoric than actual dynamic in the labor markets, possibly due to the changing economic conditions. In terms of academic perspective, there is a need for a balance in describing new trends and formerly held systems (Baruch, 2006; Lips-Wiersma and Hall, 2007). Another criticism on new career concept is that they tend to reflect individual agency and missing on the context in which careers take place (Grote and Hall, 2013). While there is a higher level of overcoming organizational structural constraints, occupational boundaries still prevent many from going through a significant job change, unless it is a move to managerial role within existing industry, or that the person took a significant retraining and qualifications. Even so, reentering the labor market at a later career stage can prove challenging. In particular, employment for the over 50 years old is fragile, and although people at this stage can perform and contribute, leaving the labor market proved too many times a permanent one-way move.

The Boundaryless Career

A boundaryless career was defined as "a sequence of job opportunities that goes beyond the boundaries of any single employment setting" (Arthur and Rousseau, 1996; DeFillippi and Arthur, 1994). It is possible to capture the boundaryless concept by involving various underlying meanings (Sullivan and Arthur, 2006). Sullivan and Arthur revisited the boundaryless theory by defining varying levels of physical and psychological career mobility between successive employment situations. The posit both physical movement and psychological movement along a continuum, suggesting the concept be

viewed and measured by the degree of boundarylessness as displayed by specific career actors.

The Protean Career

The protean idea (Hall, 1976) suggests that careers can be driven by the values of career actors and directed through personal choice, in contrast to the traditional way, where external agents such as the organization would plan, direct, and monitor careers (Hall, 2004; Hall and Mirvis, 1996). Therefore, in a similar vein to the boundaryless career, the protean career can be conceived along two dimensions: the extent to which the career is driven by values and the extent to which it is self-directed (Briscoe et al., 2006). In a complementary manner, being able to conduct careers according to values and in the direction individuals wish for, the protean career can also be viewed as an unidimensional orientation that reflects or lead to subjective career success (Baruch and Quick, 2007; Baruch et al., 2005).

There are similarities and differences between the protean and boundaryless careers (Briscoe et al., 2006). The protean model is still in a shape of 'becoming' due to the variety of views for the concept and its measurement (Gubler et al., 2013; Baruch, 2014). Protean career is an attitude or orientation of an individual, whereas boundaryless is a depiction of a system where traditional barriers are shattered or reduced, though never cease to exist. People with protean career orientation would find it easier to plan and manage their careers within a boundaryless career environment.

The Postcorporate Career

The concept of the postcorporate career (Peiperl and Baruch 1997) depicts careers that take place outside large organizations. Individuals who leave, voluntarily or involuntarily, individuals unable or unwilling to pursue corporate careers, and can lead career external to the corporate, are the actors and stakeholders in this era. The postcorporate followed from a mass redundancies and 'rightsizing,' where, alongside the large corporation a need emerged for smaller and agile organizations to complement the large corporations.

Rather than looking for merely hierarchical promotion, individuals identify or create opportunities to gain intrinsic and financial rewards. The postcorporate career is still highly contemporary, as major waves of redundancy reoccur, but the postcorporate careers have not become the norm, and fit mostly for professionals, consultants, and managers. The traditional, corporate-bound, career has proved quite robust and has survived through the pressures of competition and rationalization.

The Kaleidoscope Career

The Kaleidoscope Career Model (KCM) (Mainiero and Sullivan, 2006) was developed independently from the protean or boundaryless concepts. Mainiero and Sullivan suggest that like a kaleidoscope mechanism, the KCM depicts how individuals change the pattern of their career by rotating the varied aspects of their lives in order to arrange their relationships and roles through different practices. Changes may be internal or external such as a development in the labor market. People look to reach best fit among work demands, constraints, and opportunities as well as relationships and personal values and norms of behaviors.

The metaphorical mirrors of the career kaleidoscope are the following three parameters or motivators: (1) *authenticity*, in which the individual makes choices that permit being true to oneself; (2) *balance*, whereby the individual strives to reach an equilibrium between work and other facets of life; and (3) *challenge*, a personal need for stimulating work and career advancement. As life span, people search for the best fit that matches the character and context of their life, and the kaleidoscope's parameters shift in response. There is already empirical support for the KCM (August, 2011; Smith-Ruig, 2009).

The Multidirectional Career

The multidirectional career (Baruch, 2004b) is a concept, which manifests a trend in career systems and their meaning and implications for individuals, organizations, and society. The traditional one direction, upward climb in the hierarchy is not a sustainable model any longer. Individuals and organizations are becoming open to wide directions of possibilities to reach meaningful and successful careers. Otherwise they might have a painful 'wake-up call' when the next round of redundancy reaches the individual or the organization faces a shortage of talent (Bedeian and Armenakis, 1998). This is particularly true for people in knowledge and creative industries (Greenhalgh and Rosenblatt, 1984, 2010). Job insecurity can lead not only to negative personal outcomes (Jordan et al., 2002), but also to 'keep one on their toes' and make sure people reinvent and reshape their career, keep updated and competent, ready to move to various directions on the career frontiers.

To sum up the above conceptual developments in the study of careers, it is fair to say that the introduction of new career models and concepts reflects a true change and paints a new landscape of careers where former constraints are dissolved or shattered. In the past, people had the inclination to follow the social norm of commitment, loyalty, and stability, tended to keep with a few workplaces, and preferred to stay with the same organization. Conversely, today's careers frequently go beyond the boundaries of any single organization, and people can change roles and employers, even occupations, without acquiring a negative stigma of job-hopping or unstable professionals (Sullivan and Baruch, 2009). A new psychological contract has been established (Rousseau, 1996; Herriot and Pemberton, 1995), offering altered implications for the wider career landscape. For a sober examination of new and old realities in the area of careers, the reader is directed to Inkson and King (2011).

Factors Influencing Careers and Career Planning and Management

The components forming the *Career Active System Triad* (CAST) are set at three levels of analysis (see **Table 1**): values, approaches, and behaviors (Baruch, 2004a). The basic underlying level of values – the principles, morals, and culture – form

Table 1 The career active system triad – levels presentation

Level	*Individual*	*Organization*
Values	Aspirations	Philosophy (strategy)
Approaches	Attitudes	Policies
Behaviors	Actions	Practices

the roots from which the other levels emerge. The second level comprises of approaches and assumptions that translate those values into the third level, that of action, behavior, and practice. The values convey the aspiration (for individuals) and strategy (for organizations) into the attitudes (for individuals) and policies (for organizations). The third level of actual behaviors is where people will act and apply career-related actions (make a move, retrain, start a new venture), and organizations will utilize managerial practices. The results of such actions are the career outcomes. This is an active system, always in a perpetual motion, responding to both external and internal forces and requirements.

At the organizational system level, organizations develop, explicitly or implicitly, a system wherein the career strategies and practices follow their business strategy (Sonnenfeld and Peiperl, 1988). Organizations realize the need for a clear aligning of HRM strategy and practice with the strategy and practice of the whole business (Holbeche, 1999). The link is relevant not merely for individuals' attitudes and feelings, but also for organizational financial performance (Huselid, 1995; Combs et al., 2006).

The CAST model builds on the theory of planned action of Fishbein and Ajzen (1975) – see **Figure 2**. People's career-related actions are shaped by internal and external factors that influence their work attitudes and work performance, and their deriving intentions lead to behaviors.

Career Outcomes

Individuals thrive to gain career success. The meaning of it is not set in stone. Career success, in terms of outcomes, are typically evaluated or explored in two basic dimensions – objective and subjective. It is never simple to identify what leads to reaching career success. At the individual level, there is a distinction between merit-based factors, such as competence, skills, qualifications, that should be instrumental in people's career progress, and nonrelated issues such as gender, race, age, and other factors that can cause distortion for system that if it is not based on procedural and distributive justice.

A comprehensive meta-analysis by Ng et al. (2005) focuses on the two main dimensions of career outcomes, typically examined in terms of success. The success can be internal (intrinsic) and external (extrinsic). Heslin (2005) has identified avenues for improving the conceptualization and measurement of both objective and subjective career success, including the need for greater sensitivity to the criteria that participants, in different contexts, use to construe and judge their career success. Heslin illustrates that contextual and individual factors are likely to be associated with the relative salience of objective and subjective criteria of career success.

Career outcomes can be not only a source of great life satisfaction and fulfillment, but also a source of stress and frustration. The changing nature of careers means a significant change is the way career is shaped and managed in partnership (or otherwise) between individuals and organizations, within a wider system, regional, national, and global. Arthur et al. (2005) studied career success in a boundaryless career world and argue that typically career research makes inconsistent use of career theory: in particular, it neglects the interdependence of the objective and subjective careers, and 'boundaryless career' issues of inter-organizational mobility and extra-organizational support.

Career success is associated with significant positive outcomes, including happiness, psychological health, and longevity (Kern et al., 2009; Pan and Zhou, 2013). It is typically associated with the social capital theory (Seibert et al., 2001), which is conceptualized in terms of network structure and social resources.

Hall and Chandler (2005) expand our understanding of the relationship between subjective and objective careers, describing a condition under which the subjective career takes on particular salience: when the person feels a sense of calling, a sense of purpose, in his or her career. Indeed there is a clear difference between external success, as can be measured by hierarchy position and income, and internal success, which is about the individual feeling about their achievements.

Career outcomes can be negative, of course: A grim view about the way careers are developing in a growingly individualized society is depicted by Kalleberg (2009). He realized that the landscape of Western careers in a society comprised of individuals who live and suffer from isolation (see also Turkle, 2011).

At the organizational level, successful career planning and management means that they are able to attract, recruit, and retain talent they need. The organization needs to manage a 'talent flow' where new people enter the system, where existing members of staff continuously keep competitive and satisfied,

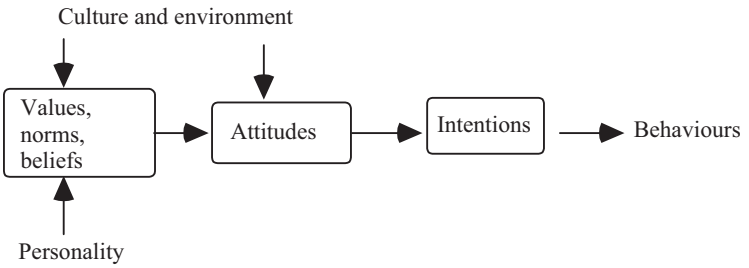

Figure 2 Fishbein and Ajzen model.

and where individuals' and organizational needs are mutually met. Internal labor markets is one strategy where the organization, at the extreme case, only recruit at the low level, and promotes from within. External labor markets means that the organization is open to recruit at any level from external sources.

Career Outcomes – National Level

At a national level of analysis, career systems can be directed and influenced by strategies and actions of governments and other political and economic bodies. First and foremost, the education system shapes the competence portfolio of a nation. It also helps to shape values of the population, though these are more influenced by the culture, religion, and other relevant factors. Investment in certain areas can generate support for certain industries (for example, high-tech industrial parks). Attractive life condition would form pull factors to attract global talent (Baruch, 1995) whereas countries, which cannot offer such conditions might suffer from the Brain Drain phenomenon, which is typical to global labor markets.

Career Practices and Systems

Sonnenfeld and Peiperl (1988) is a model of career strategies, where organizations adjust their career system to fit their organizational business strategies. In particular by exploring supply flow – do they develop internal talent via internal labor markets, or look outside for talent pool. The other dimension in their model is the assignment flow, where people are assigned with roles, and are promoted according to their individual contribution or their fit for the team/organization. Such strategies are reflected in specific practices that the organizations utilize for the planning and management of careers. Baruch and Peiperl (2000) present an empirical examination of career practices in organizations. Some of these practices are enabling organizational monitoring and control; other practices focus on organizational support and career self-management (Sturges et al., 2010).

Below I list the main career practices available for organizations (based on Baruch, 1999; Baruch and Peiperl, 2000).

Posting (advertising) internal job openings;
Formal education as part of career development;
Performance appraisal as a basis for career planning;
Career counseling by direct manager;
Career counseling by the HRM department;
Lateral moves;
Retirement preparation programs;
Succession planning;
Mentoring;
Common career paths;
Dual ladder (careers offering upward mobility in career tracks other than supervisory or managerial);
Booklets and/or pamphlets on career issues;
Written personal career planning for employees;
Assessment centers;
Development centers;
360° Performance appraisal systems (an evaluation system where employees receive confidential and anonymous feedback from all the people who work with and around them, e.g., peers, managers, etc.);
Career workshops;
Induction/orientation;
Special programs:
for high flyers (a system to identify future leaders based on their capacity to learn and adapt, rather than on the basis of seniority), dual-career couples, and
for ethnic minorities, age, gender, disabled, etc. (EEO);
Creating psychological contracts;
Secondments (temporary transfers of employees from their regular organizations to new assignments).

Integrating Practices into a Coherent System

The above list of practices vary in their applicability and relevance for different kinds of organizations: Factors such as size, sector of operation, location, dependence on human capital, local vs multinational, to name a few, are instrumental for any organization in making a learned choice about which practices to apply. Further, these practices should not be considered in isolation, but rather organizations should engage in developing an integrated system. The system should be designed to answer the needs and requirements of both the individual and the organization. The fit between organizational and individual needs is at the core of Herriot and Pemberton (1996) careers' model. It is also necessary to apply a twofold level of integration: 'internal,' among the variety of practices, and 'external' – integration between the career system and the organizational culture and strategy (Baruch, 1999).

In deciding which practices to use and how to apply them, HRM should consider the following factors or dimensions (Baruch, 2003):

- *Involvement*: from a very low to a very high level of organizational involvement needed while dealing with the specific career practice.
- *Sophistication and complexity*: from very simplistic to highly sophisticated and complex.
- *Strategic orientation*: from very practical, 'tactical,' to very strategic.
- *Developmental focused*: from low to high relevance for developing individuals.
- *Organizational decision making focused*: from low to high relevance for organizational decision-making processes.
- *Innovative*: from very traditional or conventional, to innovative and unorthodox.

Career-Related Issues

Other career-related issues worthy of mentioning here, but mostly covered elsewhere are diversity, global/international HRM, work–life balance, to name a few.

Career and Diversity

The issue of diversity, at the more general level of diversity management, discrimination, and inclusion, is dealt in depth in the article dedicated to these issues – see Gonzalez and Zamanian in this volume. Yet, there are a number of specific diversity-related issues of direct relevance for careers and career management. Looking for talent means that merit and competence should be the leading factors in selecting people for recruitment and later,

for promotion. Yet, in practice, people who do not belong to the majority group tend to suffer from direct and indirect discrimination. For a comprehensive study on the literature on global careers of women see Altman and Shortland (2008). Gender is just one issue. For example, relating to age, Dries et al. (2008) examine whether four different generations (Silent Generation, baby boomers, Generation X, and Generation Y) hold different beliefs about career. The majority of their participants still had rather 'traditional' careers, although younger generations seemed to exhibit larger discrepancies between career preferences and actual career situation.

Careers and Work–Life Balance

Slipovers from work to other facets of life (in particular family) gain high attention within academics (e.g., Jones et al., 2013; Ollier-Malaterre, 2010). Yet, in practice, pressing pressures for highly competitive business environment means that individuals committed to work are being expected to compromise on their work–life balance. Greenhaus and Powell (2003, 2006) discuss the pros and cons of work–life boundaries and propose a theoretical model of work–family enrichment.

Global Careers

Many careers are unfolded across national and other geographical boundaries. There are many types of global careers, and factors such as culture differences play a crucial role in managing careers across borders (Dickmann and Baruch, 2011). In particular, expatriation and repatriation form a growing niche of both career studies and International Human Resource Management (IHRM) literature (Bolino, 2007; Cerdin and Le Pargneux, 2009).

Expatriate assignments can help or hinder advancement of employees who have international experience (Bolino, 2007). Repatriate career success influences an organization's ability to retain its current repatriates and recruit future expatriates. The fit between career and international assignment is likely to affect the success of the assignment at its various stages (Cerdin and Le Pargneux, 2009). Cerdin and Le Pargneux offer a three-stage model: preexpatriation, expatriation, and repatriation to explicate how individuals' career characteristics influence international assignment success during expatriation and repatriation. The literature about global career goes beyond expatriation/repatriation, and the interested reader can find more on it in a number of books (e.g., Dowling et al., 2008; Harzing and Pinnington, 2010) and specific related journals.

The Future of Careers

In the past, wealth was based on physical capabilities, then on capital, primarily land. In the industrial revolution the capital that matter have changed to plant – equipment etc., later to be replaced by financial capital, which was the core key to competitive advantage. These days, the creation of wealth moves to rely on human asset, so individuals look to improve their human capital. In the information age knowledge matters, and, as was always the case, technology – all influence careers and their progress. The shift from agriculture to industry and production generated different career systems. The move to services and now to knowledge management and high reliance on IT generates new professions, and new landscape for careers, in particular virtual careers. People can work from many places – home, office, car; people are connected via fast-pace mechanisms, and updating competencies is a prerequisite for a long-term career success.

One cannot predict the shape of the things to come, but surely future careers will be quite different for the coming generations. The psychological contract will continue to evolve, technology will play a crucial role, and globalization will continue to shape careers.

Readers interested in comprehensive reviews of career studies would find the following useful: Sullivan and Baruch (2009) and Baruch and Bozionelos (2011). For attempts to cover career theory, see the two career handbooks: Arthur et al. (1989) and Gunz and Peiperl (2007).

See also: Aging and Work; Burnout, Psychology of; Diversity in Organizations; Global Development: Markets and Wellbeing; Identity and Identification, Social Psychology of; Immigration, Citizenship, and Integration: Social Work Connections; Job Analysis and Work Roles, Psychology of; Labor Market Policy; Marriage and the Dual-Career Family; Migration: Motivations; Organizational Climate in the Work Setting; Organizational Commitment; Organizations, Sociology of; Personality and Values at Work; Psychological Contracts; Social Networks; Values, Social Psychology of; Vocational Interests, Values, and Preferences, Psychology of; Wellbeing and Burnout in the Workplace, Psychology of; Work Motivation; Work, Anthropology of.

Bibliography

Altman, Y., Shortland, S., 2008. Women and international assignments: taking stock—A 25-year review. Human Resource Management 47, 199–216.

Arthur, M.B., Hall, D.T., Lawrence, B.S., 1989. Generating new directions in career theory: the case for a transdisciplinary approach. In: Arthur, M.B., Hall, D.T., Lawrence, B.S. (Eds.), Handbook of Career Theory. Cambridge University Press, Cambridge, pp. 7–25.

Arthur, M.B., Inkson, K., Pringle, J.K., 1999. The New Careers: Individual Action and Economic Change. Sage, London.

Arthur, M.B., Khapova, S.N., Wilderom, C.P.M., 2005. Career success in a boundaryless career world. Journal of Organizational Behavior 26, 177–202.

Arthur, M.B., Rousseau, D.M., 1996. The boundaryless career as a new employment principle. In: Arthur, M.G., Rousseau, D.M. (Eds.), The Boundaryless Career. Oxford University Press, New York, pp. 3–20.

August, R.A., 2011. Women's later life career development: looking through the lens of the kaleidoscope career model. Journal of Career Development 38, 208–236.

Baruch, Y., 1995. Business globalization – the human resource management aspect. Human Systems Management 14, 313–326.

Baruch, Y., 1998. The rise and fall of organizational commitment. Human Systems Management 17, 135–143.

Baruch, Y., 1999. Integrated career systems for the 2000s. International Journal of Manpower 20, 432–457.

Baruch, Y., 2003. Career systems in transition: a normative model for career practices. Personnel Review 32, 231–251.

Baruch, Y., 2004a. Managing Careers: Theory and Practice. FT-Prentice Hall/Pearson, Harlow.

Baruch, Y., 2004b. Transforming careers – from linear to multidirectional career paths: organizational and individual perspective. Career Development International 9, 58–73.

Baruch, Y., 2006. Career development in organizations and beyond: balancing traditional and contemporary viewpoints. Human Resource Management Review 16, 125–138.

Baruch, Y., 2013. Careers in academe: the academic labour market as an eco-system. Career Development International 18, 196–210.

Baruch, Y., 2014. Measuring the protean career – new construct validation. International Journal of Human Resource Management. http://dx.doi.org/10.1080/09585192.2014.896389.

Baruch, Y., Bozionelos, N., 2011. Career issues. In: Zedeck, S. (Ed.), APA Handbook of Industrial & Organizational Psychology. APA Publication, Washington DC, pp. 67–113.

Baruch, Y., Peiperl, M.A., 2000. Career management practices: an empirical survey and theoretical implications. Human Resource Management 39, 347–366.

Baruch, Y., Quick, J.C., 2007. Understanding second careers: lessons from a study of U.S. navy admirals. Human Resource Management 46 (4), 471–491.

Baruch, Y., Rosenstein, E., 1992. Career planning and managing in high tech organizations. International Journal of Human Resource Management 3, 477–496.

Bedeian, A.G., Armenakis, A.A., 1998. The cesspool syndrome: how dreck floats to the top of declining organizations. The Academy of Management Executive 12, 58–67.

Bolino, M.C., 2007. Expatriate assignments and intra-organizational career success: implications for individuals and organizations. Journal of International Business Studies 38, 819–835.

Briscoe, J.P., Hall, D.T., Frautschy DeMuth, R.L., 2006. Protean and boundaryless careers: an empirical exploration. Journal of Vocational Behavior 69, 30–47.

Carr, S.C., Inkson, K., Thorn, K., 2005. From global careers to talent flow: reinterpreting 'brain drain'. Journal of World Business 40, 386–398.

Cerdin, J.L., Le Pargneux, M., 2009. Career and international assignment fit: toward an integrative model of success. Human Resource Management 48, 5–25.

Combs, J., Liu, Y., Hall, A., Ketchen, D., 2006. How much do high-performance work practices matter? A meta-analysis of their effects on organizational performance. Personnel Psychology 59, 501–528.

De Vos, A., Buyens, D., 2008. The professional career on the right track: a study on the interaction between career self-management and organizational career management in explaining employee outcomes. European Journal of Work and Organizational Psychology 18, 55–80.

De Vos, A., Soens, N., 2008. Protean attitude and career success: the mediating role of self-management. Journal of Vocational Behavior 73, 449–456.

DeFillippi, R.J., Arthur, M.B., 1994. The boundaryless career: a competency-based perspective. Journal of Organizational Behavior 15, 307–324.

Dickmann, M., Baruch, Y., 2011. Global Careers. Routledge, NY.

Dries, N., Pepermans, R., De Kerpel, E., 2008. Exploring four generations' beliefs about career. Journal of Managerial Psychology 23, 907–928.

Dowling, P.J., Festing, M., Engle Sr, A.D., 2008. International Human Resource Management: Managing People in a Multinational Context. CengageBrain. com.

Fishbein, M., Ajzen, I., 1975. Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research Reading. Addison-Wesley, MA.

Greenhalgh, L., Rosenblatt, Z., 1984. Job insecurity: toward conceptual clarity. Academy of Management Review 9, 438–448.

Greenhalgh, L., Rosenblatt, Z., 2010. Evolution of research on job insecurity. International Studies of Management & Organization 40, 6–19.

Greenhaus, J.H., Powell, G.N., 2003. When work and family collide: deciding between competing role demands. Organizational Behavior and Human Decision Processes 90, 291–303.

Greenhaus, J.H., Powell, G.N., 2006. When work and family are allies: a theory of work-family enrichment. Academy of Management Journal 31, 72–92.

Grote, G., Hall, D.T., 2013. Reference groups: a missing link in career studies. Journal of Vocational Behavior 83, 265–279.

Gubler, M., Arnold, J., Coombs, C., 2013. Reassessing the protean career concept: empirical findings, conceptual components, and measurement. Journal of Organizational Behavior. http://dx.doi.org/10.1002/job.1908.

Gunz, H., Mayrhofer, W., Tolbert, P., 2011. Careers as a social political phenomenon in the global economy. Organization Studies 32, 1613–1620.

Gunz, H.P., Peiperl, M.A. (Eds.), 2007. Handbook of Career Studies. SAGE, Thousand Oaks, CA.

Gutteridge, T.G., Leibowitz, Z.B., Shore, J.E., 1993. Organizational Career Development. Jossey-Bass Pub, San Francisco, CA.

Hall, D.T., 1976. Careers in organizations. Pacific Palisades, Goodyear, CA.

Hall, D.T., 1996. The Career is Dead – Long Live the Career. Jossey-Bass, San Francisco, CA.

Hall, D.T., Mirvis, P.H., 1996. The new protean career: psychological success and the path with a heart. In: Hall, D.T. (Ed.), The Career Is Dead – Long Live the Career. Jossey-Bass, San Francisco, CA, pp. 15–45.

Hall, D.T., 2002. Careers in and Out of Organizations. Sage Publications, Thousand Oaks, CA.

Hall, D.T., 2004. The protean career: a quarter-century journey. Journal of Vocational Behavior 65, 1–13.

Hall, D.T., Chandler, D.E., 2005. Psychological success: when the career is a calling. Journal of Organizational Behavior 26, 155–176.

Harzing, A.W., Pinnington, A. (Eds.), 2010. International Human Resource Management. Sage Publications, London.

Herriot, P., Pemberton, C., 1995. New Deals: The Revolution in Managerial Careers. John Wiley, Chichester.

Herriot, P., Pemberton, C., 1996. Contracting careers. Human Relations 49, 757–790.

Heslin, P.A., 2005. Conceptualizing and evaluating career success. Journal of Organizational Behavior 26, 113–136.

Holbeche, L., 1999. Aligning Human Resources and Business Strategy. Butterworth Heineman, Oxford.

Higgins, M., 2005. Career Imprints: Creating Leaders across an Industry. Jossey-Bass, San Francisco, CA.

Holland, J.L., 1959. A theory of occupational choice. Journal of Counseling Psychology 6, 35–45.

Huselid, M.A., 1995. The impact of human resource management practices on turnover, productivity, and corporate financial performance. Academy of Management Journal 38, 635–672.

Iansiti, M., Levien, R., 2004. Strategy as ecology. Harvard Business Review 82 (3), 68–81.

Inkson, K., Gunz, H., Ganesh, S., Roper, J., 2012. Boundaryless careers: bringing back boundaries. Organization Studies 33, 323–340.

Inkson, K., King, Z., 2011. Contested terrain in careers: a psychological contract model. Human Relations 64, 37–57.

Jones, F., Burke, R.J., Westman, M. (Eds.), 2013. Work-life Balance: A Psychological Perspective. Psychology Press, New York, pp. 235–260.

Jordan, P.J., Ashkanasy, N.M., Hartel, C.E., 2002. Emotional intelligence as a moderator of emotional and behavioral reactions to job insecurity. Academy of Management Review 27, 361–372.

Kalleberg, A.L., 2009. Precarious work, insecure workers: employment relations in transition. American Sociological Review 74, 1–22.

Kattenbach, R., Schneidhofer, T.M., Lücke, J., Latzke, M., Loacker, B., Schramm, F., Mayrhofer, W., 2013. A quarter of a century of job transitions in Germany. Journal of Vocational Behavior. http://dx.doi.org/10.1016/j.jvb.2013.11.001.

Kern, M.L., Friedman, H.S., Martin, L.R., Reynolds, C.A., Luong, G., 2009. Conscientiousness, career success, and longevity: a lifespan analysis. Annals of Behavioral Medicine 37, 154–163.

Kets de Vries, M.F.R., Balazs, K., 1997. The downsize of downsizing. Human Relations 50, 11–50.

Lips-Wiersma, Hall, D.T., 2007. Organizational career development is *not* dead: a case study on managing the new career during organizational change. Journal of Organizational Behavior 28, 771–792.

Mahroum, S., 2000. Highly skilled globetrotters: mapping the international migration of human capital. R & D Management 30, 23–31.

Mainiero, L.A., Sullivan, S.E., 2006. The Opt-out Revolt: Why People Are Leaving Companies to Create Kaleidoscope Careers. Davis-Black, Palo Alto, CA.

Ng, T.W.H., Eby, L.T., Sorensen, K.L., Feldman, D.C., 2005. Predictors of objective and subjective career success: a meta-analysis. Personnel Psychology 58, 367–408.

Ollier-Malaterre, A., 2010. Contributions of work-life and resilience initiatives to the individual/organization relationship. Human Relations 63, 41–62.

Pan, Zhou, W., 2013. Can success lead to happiness? the moderators between career success and happiness. Asia Pacific Journal of Human Resources 51, 63–80.

Peiperl, M.A., Baruch, Y., 1997. Back to square zero: the post-corporate career. Organizational Dynamics 25 (4), 7–22.

Rousseau, D.M., 1996. Changing the deal while keeping the people. Academy of Management Executive 10 (1), 50–59.

Rodrigues, R.A., Guest, D., 2010. Have careers become boundaryless? Human Relations 63, 1157–1177.

Schein, E.H., 1978. Career Dynamics: Matching Individual and Organizational Needs. Addison-Wesley, Reading, MA.

Seibert, S.E., Kraimer, M.L., Liden, R.C., 2001. A social capital theory of career success. Academy of Management Journal 44, 219–237.

Smith-Ruig, T., 2009. Mapping the career journey of accountants in Australia. In: Baugh, S.G., Sullivan, S.E. (Eds.), Maintaining Energy, Focus and Options over the Career: Research in Careers, vol. 1. Information Age, Charlotte, NC, pp. 163–196.

Sonnenfeld, J.A., Peiperl, M.A., 1988. Staffing policy as a strategic response: a typology of career systems. Academy of Management Review 13, 568–600.

Sturges, J., Conway, N., Liefooghe, A., 2010. Organizational support, individual attributes, and the practice of career self-management behavior. Group & Organization Management 35, 108–141.

Sullivan, S.E., 1999. The changing nature of careers: a review and research agenda. Journal of Management 25 (3), 457–484.

Sullivan, S.E., Arthur, M.B., 2006. The evolution of the boundaryless career concept: examining physical and psychological mobility. Journal of Vocational Behavior 69, 19–29.

Sullivan, S.E., Baruch, Y., 2009. Advances in career theory and research: critical review and agenda for future exploration. Journal of Management 35, 1452–1571.

Super, D.E., 1957. The Psychology of Careers. Harper & Row, New York.

Tams, S., Arthur, M.B., 2007. Studying careers across cultures: distinguishing international, cross-cultural, and globalization perspectives. Career Development International 12, 86–98.

Turkle, S., 2011. Alone Together: Why We Expect More from Technology and Less from Each Other. Basic Books, NY.

Caregiver Burden

Christine Fekete, Swiss Paraplegic Research, Nottwil, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Schulz, volume 3, pp. 1476–1479, © 2001, Elsevier Ltd.

Abstract

In this article, the terms caregiving and caregiver burden are defined, basic epidemiological data on the prevalence of caregiving is presented, and the typical caregivers are described. Based on the stress-theoretical approach, caregiving and its potential consequences on different life domains are discussed in more detail. Finally, some future challenges of caregiving are discussed.

Defining Caregiving and Caregiver Burden

Although no general definition on caregiving is available, there is broad agreement that informal caregiving can be described as 'extraordinary care, exceeding the bounds of what is normative or usual in family relationships' (Schulz and Martire, 2004: p. 240). Typically, caregiving is characterized by a substantial expenditure of time, financial resources, and energy, in an extent that exceeds the 'normal' as the care receiver somehow depends on the effort of the caregiver. Close relatives or spouses are often involved in caregiving over longer periods, and they perform it voluntarily, due to a lack of available alternatives or due to the strong wish of the care receiver to stay in the home setting. Caregiving might include displeasing activities, such as helping an incontinent person with bladder management, and can be physically exhausting, e.g., helping a mobility-impaired person getting up. Given unstable health conditions of care recipients, caregiving is often associated with lack of control and unpredictability (Schulz and Martire, 2004).

In contrast to formal caregiving provided by health professionals and being paid by the recipient or any responsible body (e.g., health insurances or the social security system), informal caregiving is provided by nonprofessionals, without financial compensation. To emphasize this lack of remuneration, the term 'informal caregiving' is often used synonymously with 'family caregiving' (Bastawrous, 2013). Services performed by informal caregivers are considered a substantial health care resource, saving the public health care systems hundreds of billions of dollars each year (Arno et al., 1999).

As caregiving may be physically and emotionally strenuous, lasting over a longer period without a predictable development, caregiving is described as a chronic stressor with relevant impact on diverse life domains, such as physical and mental health (Pinquart and Sorensen, 2003, 2007), labor market integration, or the financial situation (Family Caregiver Alliance, 2013). As the label of 'caregiver burden' suggests, caregiving can be seen as strenuous or burdensome. Yet, the term 'caregiver burden' remains vague, as there is no general agreement on its definition, neither at the conceptual level nor at the operational level (Bastawrous, 2013). Generally, caregiver burden describes a variety of different experiences related to caregiving, thus requiring a complex definition. One such definition maintains that caregiver burden is "a multidimensional response to physical, psychological, emotional, social, and financial stressors associated with the caregiving experience" (Kasuya et al., 2000: p. 119). The caregiver burden as a result of potentially distressing caregiving activities relates to 'objective' and 'subjective' aspects of caregiving. The objective burden refers to the objectively observable physical or instrumental care provision in terms of number of hours or type of care provided. In contrast, subjective measures of the caregiver burden mostly relate to emotional or psychological consequences of caregiving; for example, depressive mood or distress (Bastawrous, 2013).

To date, a variety of standardized measurement instruments have been designed to assess the caregivers' feelings of burden, mostly including both objective and subjective indicators of burden (for an overview see Whalen and Buchholz, 2009). However, as a consequence of the multidimensional concept, many instruments include a multitude of constructs, such as impact of caregiving, quality of life, risk, strain, hassles, cost of care, experiences, and domain-specific areas of stress (physical, psychological, emotional, social, or financial). These instruments often assess very detailed aspects of caregiver burden, e.g., worry, isolation, guilt, blame, or fatigue. Conversely, some instruments include positive outcomes of caregiving, such as the satisfaction that results from the caregiving experience (Whalen and Buchholz, 2009).

The Prevalence of Caregiving

As informal or family caregiving is not defined uniformly, prevalence estimates vary according to the underlying definition. However, there is a general consensus that its prevalence is high. Different estimates from the United States are used to illustrate the dimension of caregiving in Western countries. For example, estimates from the National Alliance for Caregiving suggest that 65.7 million persons, and therefore approximately 29% of the US adult population, provide care to an ill, disabled, or aging person (The National Alliance of Caregiving and the American Association of Retired Persons, 2009). Other sources estimate that 52 million persons in the United States provide care to an adult aged over 18 years with disability or illness (Coughlin, 2010). Data from the Alzheimer's Association estimate that around 43.5 million adult caregivers care for someone above age 50, and 14.9 million caregivers care for someone with Alzheimer disease or another type of dementia (Alzheimer's Association, 2011). Additionally, an estimated

5.9 million children in the United States live with severe disabilities, most of them receiving care at home by their parents and families (Perrin, 2002).

The Typical Caregiver

As populations age across most Western countries (European Commission, 2007; U.S. Department of Health and Human Services, 2012), the group of those caring for elderly persons has steadily increased over the past decades. Many old people suffer from one or more chronic conditions where their care cannot be given by professional care providers. In these instances, spouses or children are most likely to care for their elderly family members. Around 70% of 50- to 64-year-old caregivers provide care for an elderly parent. Among caregivers aged over 75, caring for a spouse is the most commonly reported care situation. In the group of spousal caregivers, the proportion of male caregivers is higher than in any other care situation (Wagner and Takagi, 2010).

Based on theoretical concepts of gender role socialization, gender role expectations, and labor market segregation indicating that women more often engage in household activities, it has been suggested that females are more often involved in caregiving than males. However, evidence concerning gender differences in caregiving is inconclusive. In a review from Yee and Schulz (2000), four studies found that women spend more time in caregiving than men, while two studies reported no gender differences, and one study observed that men spend more time in caregiving than women. Despite these inconsistent results, there is some consensus that the proportion of male caregivers has steadily increased over the past years, and that it is especially high in older cohorts (Wagner and Takagi, 2010; Family Caregiver Alliance, 2013).

The average age of caregivers has slightly increased during the past decades and is now around 50 years. The average duration of caregiving is about 4 years with widely varying intensity (Wagner and Takagi, 2010; The National Alliance for Caregiving, 2005, 2009). According to a US study, the average caregiver spends 21 h with care provision per week, where about half of the respondents indicated a caregiving provision of less than 8 h, but where almost 20% reported to spend more than 40 h weekly (The National Alliance for Caregiving, 2005). Despite the wide range of care intensity, it is generally assumed that the number of hours caring for someone increases with the age of the caregiver (Wagner and Takagi, 2010; The National Alliance for Caregiving, 2005, 2009).

Around three of four caregivers are employed. Still, over 50% of middle-aged employed caregivers indicate that their caregiving task affects their work life such that they report leaving work early or come in late as a consequence of their caregiving role (Wagner and Takagi, 2010). According to a review from 2007, caregivers engaged with low amounts of time in caregiving had the same chance of labor market participation as noncaregivers. In contrast, caregivers who were heavily involved in caregiving were less likely to work than noncaregivers. Likewise, researchers found an inverse relationship between time spent on caregiving and time spent on participation in the labor market (Lilly et al., 2007).

Caregiving and Its Multifaceted Consequences

Although caregiving is a highly important resource for society and individuals in need of care, there is a consensus on the fact that caregiving is a chronic stressor affecting different life domains of caregivers. Caregiving has therefore been analyzed within a stress-theoretical framework that considers caregiving as the proliferation or spread of stress from primary stressors to increased risk of physical or psychological morbidity of the caregiver (Schulz and Martire, 2004; Pearlin et al., 1997).

Within this analytical framework, objective primary stressors are composed of characteristics of the care receivers' health (e.g., level of functioning, severity of disability, cognitive impairment) and the type and intensity of caring activities. Further objective stressors are care receivers' behavioral problems including verbal and physical aggression, agitation, psychotic symptoms (e.g., hallucinations and delusions), sleep disturbances, depression, and obstructive behavior (Ornstein and Gaugler, 2012). Interestingly, higher levels of care recipient behavior problems were more consistently related to the caregiver burden in terms of poor health than were care receivers' impairment and intensity of caregiving (Pinquart and Sorensen, 2007), possibly due to their unpredictability.

As caring for a person with cognitive impairment seems particularly stressful, dementia caregiving provides an instructive example of the relation between objective stressors and its consequences for caregivers. Dementia caregiving is therefore the predominant health condition studied in the caregiving literature (Schulz and Martire, 2004). For example, dementia caregivers spend significantly more time providing care and supervision than nondementia caregivers and are more often confronted with behavioral problems of care recipients (Ornstein and Gaugler, 2012). As a consequence of these primary stressors, persons caring for someone with cognitive problems report higher caregiver burden and suffer from more mental and physical health problems than nondementia caregivers (Schulz and Martire, 2004; Pinquart and Sorensen, 2003, 2007).

Besides primary stressors related to care receiver characteristics, stress-coping models also include the examination of secondary stressors, mainly related to the role conflicts due to caregiving demands (Schulz and Martire, 2004; Pearlin et al., 1997). For example, long-term care for a family member may increase the risk of family conflicts in case of disagreement on caregiving roles, division of responsibilities, or the way of carrying out caregiving tasks. Additionally, work-related problems such as being tired or absent from work, missing job opportunities, or missing chances for job promotion are potential problems that occur as a consequence of primary stressors of caregiving. Further consequences of caregiving including financial difficulties or social isolation were described (Ranmuthugala et al., 2009). Again, dementia caregivers seem to be particularly vulnerable to secondary stressors as they more often face complications with their involvement in the labor market, and as they report more limitations in their leisure time and their engagement in social relationships (Etters et al., 2008).

Based on the stress-health model, primary (care receivers' characteristics, type of care provided) and secondary stressors (e.g., role conflicts, financial hardship) result in the experience

of stress, in case caregivers evaluate the demands as threatening and their coping resources as inadequate. Research has documented the physiological responses related to caregiving in terms of stress hormones, immune competence, and medication use (Vitaliano et al., 2003). It is assumed that chronically stressful experience elicits negative emotional, physiological, and behavioral responses that increase the risk of physical and psychiatric morbidity (Schulz and Martire, 2004). According to a comprehensive meta-analysis, caregivers report worse physical and psychological health than noncaregivers. More specifically, they suffer from more stress, more depressive symptoms, lower levels of subjective well-being, poorer physical health, and lower self-efficacy than noncaregivers. Again, the health differences were highest in dementia caregivers (Pinquart and Sorensen, 2003). Ultimately, caregivers reporting a burden related to their caregiving activity had a 63% increased mortality risk compared to that of persons who were not involved in caregiving. Conversely, caregivers who did not report caregiver burden were not likely to show an increased mortality risk compared to noncaregivers (Schulz and Beach, 1999).

Since most of the research has concentrated on negative effects of caregiving, criticism concerning the focus on detrimental caregiving consequences has been raised, and it has been claimed that positive aspects of caregiving need to be included in research as well. In fact, several studies observed that persons providing care feel enriched by their caregiving role due to the fact that caregiving contributes to personal growth, a stronger purpose in life, a feeling of mastery, or a sense of usefulness (Kramer, 1997). Yet, it remains unclear how the positive and negative experiences related to caregiving interact.

Whether caregiving is perceived as stressful experience or not may also depend on a variety of factors moderating and/or mediating the effects on caregiver outcomes. Socioeconomic position, psychosocial resources, sociodemographic characteristics, or psychological traits play an important role in predicting the caregiver burden. For example, caregivers in higher socioeconomic position are at lower risk of poor health than caregivers in disadvantaged social positions (Pinquart and Sorensen, 2007). It is supposed that caregivers in privileged social positions have better access to the health care system and are less often exposed to additional stressors unrelated to caregiving. Social support is a further resource potentially important to reduce the caregiver burden as social support may buffer the stressors related to the caregiver role, strengthen coping strategies, and enhance health-promoting behaviors. Evidence from an extensive meta-analysis supports this notion by showing that informal support of caregivers is associated with better physical health (Pinquart and Sorensen, 2007). Also, psychological resources such as self-efficacy or self-esteem of caregivers have been identified as protective factors in the experience of caregiver burden. Furthermore, age and gender may have impact on the development of caregiver strain and ill-health of caregivers. Not surprisingly, higher age was related to poorer physical health of caregivers (Pinquart and Sorensen, 2007), and it has been suggested that the health effects of caregiving are more serious in female than in male caregivers. Caregiving women reported somehow higher burden, lower subjective well-being, more depressive symptoms, and lower physical health than caregiving men (Pinquart and Sorensen, 2006). Finally, the quality of the relationship between the caregiver and the care receiver might be an important factor in the perception of burden and subsequent health outcomes. A recent review indicated that caregiver–care receiver mutuality is an important correlate of mental health and emotional well-being of the caregiver (Park and Schumacher, 2013).

Future Challenges

Demographic changes and further progress in medical treatment of chronic disorders may lead to an increased importance of informal caregiving in the near future. For example, recent estimates claim that by the year 2040, around 21% of North Americans will be aged over 65 years (U.S. Department of Health and Human Services, 2012), and estimates for Europe predict that the old age dependency ratio, expressed by the number of people aged over 65 per number of people aged 15–64, will increase from 25% in 2007 to 53% in 2050 (European Commission, 2007). As it is highly unlikely that the aging population will be adequately supplied by professional care, these developments result in a high need of informal care. In addition, progress in medical treatment will extend caregiving episodes, e.g., through drug therapies, improving a patient's functional status and extending his or her home care.

Although the overall need for informal caregiving is likely to increase, changes in family structures and social roles may compromise the potential of informal caregiving or may at least increase the caregiver burden. Family structures are likely to change such that future cohorts of elderly will have smaller families and, in consequence, fewer children available for engagement in caregiving. Additionally, as female labor market participation will increase in the near future, female resources to provide care will be reduced.

These societal changes in family composition and in the distribution of social norms within families will result in a growing concern about opportunities and threats of caregiving in the future. Obviously, caregiving is closely related to demographic changes and level of functioning among aging generations. These developments call for action at two levels. First, at the individual level, interventions targeted to reduce the burden on caregivers need to be established. Experiences with support groups have shown promising results in enhancing caregivers' subjective well-being (Chien et al., 2011). Second, at the macro-social level, the challenge of caregiver burden has to be addressed by appropriate initiatives from social and health policy as well as from supportive private sector programs.

See also: Alzheimer's Disease: Behavioral and Social Aspects; Chronic Illness, Psychosocial Coping With; Disability: Social and Psychological Aspects; Gender and Health Care; Health-Related Support and Self-help Mutual Aid Groups.

Bibliography

Alzheimer's Association, 2011. 2011 Alzheimer's disease facts and figures. Alzheimer's & Dementia 7, 208–244.

Arno, P.S., Levine, C., Memmott, M.M., 1999. The economic value of informal caregiving. Health Affairs (Millwood) 18, 182–188.

Bastawrous, M., 2013. Caregiver burden–a critical discussion. International Journal of Nursing Studies 50, 431–441.
Chien, L.Y., Chu, H., Guo, J.L., Liao, Y.M., Chang, L.I., Chen, C.H., Chou, K.R., 2011. Caregiver support groups in patients with dementia: a meta-analysis. International Journal of Geriatric Psychiatry 26, 1089–1098.
Coughlin, J., 2010. Estimating the impact of caregiving and employment of well-being. Outcomes & Insights in Health Management 2, 1–7.
Etters, L., Goodall, D., Harrison, B.E., 2008. Caregiver burden among dementia patient caregivers: a review of the literature. Journal of the American Association of Nurse Parctitioners 20, 423–428.
European Commission, 2007. European's Demographic Future: Facts and Figures on Challenges and Opportunities. European Communities, Directorate-General for Employment, Social Affairs and Equal Opportunities, Luxembourg.
Family Caregiver Alliance, 2013. Fact Sheet: Selected Caregiver Statistics. www.caregiver.org/caregiver/jsp/content_node.jsp?nodeid=439 (accessed 13.09.13.).
Kasuya, R.T., Polgar-Bailey, P., Takeuchi, R., 2000. Caregiver burden and burnout. A guide for primary care physicians. Postgraduate Medicine 108, 119–123.
Kramer, B.J., 1997. Gain in the caregiving experience: where are we? what next? Gerontologist 37, 218–232.
Lilly, M.B., Laporte, A., Coyte, P.C., 2007. Labor market work and home care's unpaid caregivers: a systematic review of labor force participation rates, predictors of labor market withdrawal, and hours of work. Milbank Quarterly 85, 641–690.
The National Alliance of Caregiving, 2005. Caregiving in the U.S. The National Alliance of Caregiving, Washington, DC.
The National Alliance of Caregiving and the American Association of Retired Persons, 2009. Caregiving in the U.S. National Alliance of Caregiving, Washington, DC.
Ornstein, K., Gaugler, J.E., 2012. The problem with "problem behaviors": a systematic review of the association between individual patient behavioral and psychological symptoms and caregiver depression and burden within the dementia patient-caregiver dyad. International Psychogeriatrics 24, 1536–1552.
Park, E.O., Schumacher, K.L., 2013. The state of the science of family caregiver-care receiver mutuality: a systematic review. Nursing Inquiry. http://dx.doi.org/10.1111/nin.12032.
Pearlin, L.I., Aneshensel, C.S., Leblanc, A.J., 1997. The forms and mechanisms of stress proliferation: the case of AIDS caregivers. Journal of Health and Social Behavior 38, 223–236.
Perrin, J.M., 2002. Health services research for children with disabilities. Milbank Quarterly 80, 303–324.
Pinquart, M., Sorensen, S., 2003. Differences between caregivers and noncaregivers in psychological health and physical health: a meta-analysis. Psychology and Aging 18, 250–267.
Pinquart, M., Sorensen, S., 2006. Gender differences in caregiver stressors, social resources, and health: an updated meta-analysis. The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences 61, 33–45.
Pinquart, M., Sorensen, S., 2007. Correlates of physical health of informal caregivers: a meta-analysis. The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences 62, 126–137.
Ranmuthugala, G., Nepal, B., Brown, L., Percival, R., 2009. Impact of home based long term care on informal carers. Australian Family Physician 38, 618–620.
Schulz, R., Beach, S.R., 1999. Caregiving as a risk factor for mortality: the Caregiver Health Effects Study. Journal of the American Medical Association 282, 2215–2219.
Schulz, R., Martire, L.M., 2004. Family caregiving of persons with dementia: prevalence, health effects, and support strategies. The American Journal of Geriatric Psychiatry 12, 240–249.
U.S. Department of Health and Human Services, 2012. Administration on Aging. A Profile on Older Americans. http://www.aoa.gov/AoARoot/Aging_Statistics/Profile/2012/4.aspx.
Vitaliano, P.P., Zhang, J., Scanlan, J.M., 2003. Is caregiving hazardous to one's physical health? A meta-analysis. Psychological Bulletin 129, 946–972.
Wagner, D., Takagi, E., 2010. Informal Caregiving by and for Older Adults. http://healthaffairs.org/blog/2010/02/16/informal-caregiving-by-and-for-older-adults/ (accessed 17.09.13.).
Whalen, K., Buchholz, S.W., 2009. The reliability, validity and feasibility of tools used to screen for caregiver burden: a systematic review. JBI Library of Systematic Reviews 7, 1373–1430.
Yee, J.L., Schulz, R., 2000. Gender differences in psychiatric morbidity among family caregivers: a review and analysis. The Gerontologist 40, 147–164.

Relevant Websites

http://www.alz.org/care/.
http://www.caregiver.org.
http://caregiveraction.org/.
http://caregivertoolkit.ca.
http://www.caring.com/support-groups.

Cargo Cults

Lamont Lindstrom, University of Tulsa, Tulsa, OK, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by K. Burridge, volume 3, pp. 1481–1484, © 2001, Elsevier Ltd.

Abstract

Cargo cult spread in the 1950s as a new cover term for Melanesian social movements, although it has also labeled ritual political–economic endeavors elsewhere. In these movements, islanders turned to religious ritual to obtain 'cargo' (or *kago* in Melanesian Pidgin English). The term is rich in meaning. It has signified manufactured goods or money, a new political order, local sovereignty and the withdrawal of colonial rulers, or even human salvation. Cargo cult prophets devised or revived ritual practices, creating new communities waiting for cargo's arrival. Although anthropologists have largely abandoned the term, cargo cult survives in popular media, journalistic, and literary circles to castigate or celebrate all sorts of untoward and wild desire.

The Birth of Cargo Cult

Melanesian cargo cults, so they came to be called, boomed in the 1950s. These southwest Pacific social movements attracted the attention of colonial administrators, ethnographers, journalists, and filmmakers. Although anthropologists borrowed and then refined the term in comparative research and analysis, the label quickly spread beyond ethnographic purviews into sometimes unexpected corners of popular culture. In the final scenes, for example, of the 1962 film *Mondo Cane*, Gualtiero Jacopetti's original 'shockumentary,' we watch yearning Papua New Guinea islanders clustered around a huge, roughly made wooden model airplane. They are gathered high up in the mountains, sitting on a new homemade airstrip carved out of the forest. Their eyes eagerly search the skies, so the narrator tells us, for airplanes full of wonderful cargo that they expect will soon arrive. But they will be disappointed. No cargo planes will land. These islanders are the misguided followers of a cargo cult.

Anthropologists and others, during the first half of the twentieth century, noted and described a variety of island movements and miscellaneous social disturbance throughout Melanesia (that part of Oceania which today comprises the Indonesian province of West Papua, and the southwest Pacific island nations Papua New Guinea, Solomon Islands, Vanuatu, Fiji, and New Caledonia). Australian colonial anthropologist F.E. Williams' (1923) account of the 'Vailala Madness' in Papua was one of the more comprehensive descriptions of a movement that erupted in the earlier years of European domination of these islands. It was not until after the Pacific War, however, that Melanesian social movements would come to be bundled together under the new rubric of cargo cult. Peter Worsley, who compiled an important initial overview and survey, characterized cargo cults as:

> strange religious movements in the South Pacific during the last few decades. In these movements, a prophet announces the imminence of the end of the world in a cataclysm which will destroy everything. Then the ancestors will return, or God, or some other liberating power, will appear, bringing all the goods the people desire, and ushering in a reign of eternal bliss (1957: 11).

Following the Second World War, anthropological concern was gradually shifting from describing static social orders through the lens of structural functionalism, which presumed cultural continuity, to issues of development and upheaval. In the 1950s, social change and revolution became difficult to ignore. Two celebrated accounts of the history of protest and rebellion appeared in the same decade as the first cargo cult monographs: Norman Cohn's *Pursuit of the Millennium* (1957) and Eric Hobsbawm's *Primitive Rebels* (1959). Issues of protest, insurrection, insurgence, and upheaval were then in the air and these questions percolated even in the far Melanesian fringes of the colonial world. The war had mortally undermined the colonial system, although the final effects of this would not play out in the Pacific until the 1970s. After 1945, new problems of social change, syncretism, governance, and new challenges to colonial authority attracted increased anthropological attention, and the cargo cult emerged as a central problem which confronted both colonial power and academic social theory.

The cargo cult story depicted South Pacific islanders as desperate, confused, or sometimes misled who turned to religious ritual (which was sometimes traditional and sometimes innovative) in order to obtain cargo. They expected and worked for a sudden, miraculous transformation in their lives and villages. Steinbaurer, in 1979, published a catalog of 186 cargo cults and others have counted up to 400–500 of these, although this has required some stretching and diversifying the label, both temporally and spatially. Especially notable or long-lived cargo cults (or Melanesian social movements that attracted the label) include the John Frum and Nagriamel movements of Vanuatu, the Christian Fellowship Church and the Moro Movement of the Solomon Islands, and the Paliau and Yali movements, Hahalis Welfare Society, Pomio Kivung, and Peli Association of Papua New Guinea.

Leading these movements, cult prophets commonly drew on Christian millenarianism, sometimes conflating the arrival of cargo with Christ's second coming and Judgment Day (locally often called 'Last Day'). The word 'cargo' (or *kago* in Melanesian Pidgin English) is rich in meaning. It has signified various sorts of manufactured goods (vehicles, packaged foods, refrigerators, guns, tools, and the like) and European money. Or cargo could represent the struggle for a new political order,

the assertion of local sovereignty, and the withdrawal of colonial rulers (see Burridge, 1960). Or, again, cargo more metaphorically could signify human salvation and a final, successful transcendence of everyday reality (Strelen, 1977).

The first academic accounts of cargo cults featured in anthropological journals beginning in the early 1950s. The first ethnographic monographs appeared in 1956. The period between 1956 and 1964 was cargo cult's golden age (or literary high point, at least). During these years, six important anthropological analyses were published. Five of these dealt with particular cargo movements: Jean Guiart (1956) on Tanna Island's John Frum Movement in what was then the colonial New Hebrides; Margaret Mead (1956) and Theodore Schwartz (1962) on the Paliau Movement of Manus, Admiralty Islands; Kenelm Burridge (1960) on northern Madang movements in Papua New Guinea; and, not too far away, Peter Lawrence's account of the Yali Movement (1964). The sixth, now classic volume was Peter Worsley's comparative compendium of Pacific cargo cults, *The Trumpet Shall Sound* (1957). One might also include here Robert Maher's *New Men of Papua: A Study in Culture Change* (1961) that described the Tommy Kabu Movement of the Purari River delta area of Papua, although Maher did not use cargo cult idiom to frame his analysis. The term only appeared as an afterthought on the book's final page where Maher warned that Purari people, although up to then pragmatic, could turn instead to cargo cult should their growing desire for social change be thwarted. After the mid-1960s, cargo literature and accounts hugely expanded as ethnographers described more and more Pacific social movements under the cargo cult rubric. Cargo cults, too, demanded increased attention as one of the main theoretical puzzles of Pacific ethnography (why, when, where, and how does cargoism erupt). Cargo cult analyses also informed grander anthropological theories of cultural organization and social change (see Jarvie, 1964; Jebens, 2004).

The Cargo Cult Label

The term cargo cult first appeared in the November 1945 issue of the colonial news magazine *Pacific Islands Monthly*. That month, Norris Mervyn Bird, whom the editors identified as an old territories' resident, wrote to warn of postwar flare-ups among islanders misled by ill-digested Christian teaching and an increasingly liberal and dangerous colonial policy. Bird was particularly perturbed that the new Labor Party government in Canberra might liberalize its colonial native affairs policies, including maintaining islander militias into the postwar era. Several anthropologists including Lucy Mair and H. Ian Hogbin, who then lectured at the Australian Army School of Civil Affairs in Canberra, adopted the term soon afterward. This quickly spread through Australian academia and then into anthropological circles overseas as anthropologists and others borrowed 'cargo cult' to label almost any sort of organized village-based social movement with religious and political aspirations – movements which were increasingly on colonialist and academic radar throughout Melanesia.

Although cargo cult, the label, dates only back to 1945, it spread rapidly as a popular and soon ubiquitous cover term for Pacific social movements. Some extended it back in time to label recorded political disturbances as old as 1857 (Steinbauer, 1979). The term also spilled beyond Melanesia and was also applied to movements in the Polynesian and Micronesian cultural regions of the Pacific. Although an improvement over 'Vailala Madness' and other prewar terms for island social movements, the label cargo cult proved problematic in several ways. Islanders involved in such movements always aspired to achieve many things beyond simple cargo, or material goods themselves. And movement organization was often misunderstood in cultic terms. More than one observer has had to protest that some social movement in question was 'neither cargo nor cult' (e.g., Kaplan, 1995). Moreover, critics within the Pacific and beyond also quickly adopted the term as a form of political abuse. Some politicians today continue to belittle the plans and aspirations of their rivals by deprecating these as 'cargo cultist.' Despite cargo cult's popularity, from the beginning anthropologists juggled a long list of alternative labels. These have included nativistic movements, vitalization or revitalization movements, accommodation, acculturation, adaptation, adjustment, dynamistic, or enthusiastic movements, messianic movements, millenarian movements, enthusiastic movements, crisis cults, Holy Spirit movements, protonationalist movements, culture-contact movements, and more. These alternative labels appreciated cargo cult's affinities with social movements elsewhere also sparked by the global spread of the colonialist and capitalist systems. Melanesian cargo cults were in significant ways similar to the North American Ghost Dance and Native American Church, China's Boxer Rebellion, and Mau Mau of East Africa, all local reactions to globalizing political and economic forces.

Cargo Belief

The defining aspect of cargo ideology was, of course, unrequited desire for cargo itself. Cultists sought the arrival of ships full of cargo: manufactured goods and tinned foods, vehicles, weapons, and money. After observing military transport during the Pacific War, cultists also began scanning their horizons for cargo submarines and planes. The particulars of desired cargo, as reported, reflected not only local aspirations but also stereotypic presumptions of what islanders should want. Following the war, refrigerators for example, occupied a suspiciously prominent place in many such reported cargo lists and one might expect the refrigerator to be principally an object of western, not island, desire (Lindstrom, 1996).

Details of cargo belief varied from movement to movement. A prevalent theme was the assertion that island ancestors were somehow involved in the production of manufactured goods. During some cults, islanders spun stories of a technologically wise ancestor who long ago had sailed away to America, Europe, or Australia and there taught the secrets of cargo, sometimes leaving behind his less accomplished brother, who was ancestral to less technologically advanced island communities. In others, cargo myth presumed that wily Europeans had stolen industrial knowledge from Pacific ancestors, or were stealing the cargo itself that ancestors were beneficiently shipping back home to the islands. In either case, people invented new rituals to induce the dead to reroute cargo flows and, sometimes, to reanimate and return home aboard cargo-filled ships and planes and reunite with their descendants.

The cultic task was to penetrate the secret knowledge people presumed to support European technological and political prowess and to acquire this knowledge and goods alike. The human task was to restore local dignity and realize egalitarian relations with troublesome white newcomers (Lattas, 1998).

After the Pacific War, the American military occasionally assumed the role of cargo provider. Many Melanesians, particularly those recruited to work at allied bases on Efate and Espiritu Santo in the New Hebrides, Vanuatu, Guadalcanal in the Solomon Islands, and Manus, Hollandia, and elsewhere in New Guinea, received better pay and gained access to a welcome variety of new wartime goods and services. After the militaries withdrew from most of the southwest Pacific, money and goods became scarcer. John Frum movement supporters active on Tanna began predicting the return of US military units along with the wartime cargo they had enjoyed as labor corps recruits. John Frum leaders also incorporated experience of military routines and symbols into their postwar cult ritual and liturgy, including drill team marching, bamboo rifles, red crosses (from Army ambulances), numbered dog (identification) tags, khaki uniforms, and American flags (Figure 1).

Many movements, in addition to simple material goods, pursued various sorts of world transformation. This, too, partly reflected political conditions at the end of the Pacific War. The Japanese advance had dislodged the Dutch and the Australians from much of New Guinea as well as the British from the Solomon Islands. Large American occupation forces similarly destabilized colonial authority in the New Hebrides and, to a degree, in New Caledonia and Fiji. At war's end, when colonial powers moved to reestablish their authority, many islanders who had largely governed themselves during the war years resisted this reassertion of European control. Some postwar cult prophets in response began predicting that ancestors, or returning Americans, would drive the colonial powers from the region. In these areas, cargo cults were as Jean Guiart (1951) argued 'forerunners of Melanesian nationalism.'

Figure 1 John Frum followers raise American flag at Sulphur Bay movement headquarters (copyright Lamont Lindstrom).

In addition to articulating desire for freedom, dignity, and independence from European domination, cult prophets often predicted more millenarian sorts of change. Mountains would flatten and valleys would rise up. Land would become sea, and the seas dry up. Cultists expected the coming of a new world and transformed personhood, and many cult rituals offered motifs and elements of rebirth, or baptism, to mark the advent of a new order. Two common cargo prophesies were that the dead would come back to life, and that the skins of the faithful would whiten. The first of these prophecies reflected the importance of ancestral spirits in traditional island religions and their connections with fertility and production. The second responded to the stark racial inequalities of colonialist regimes where Europeans controlled access to money, goods, and education. Some read these desires in Christian, or at least syncretic, terms as signifying a universal human quest for salvation and ultimate transcendence of everyday existence and sourced cult ideologies to the more millennially inclined Christian messages offered by missions, such as the Seventh-day Adventists, that spread throughout the region before the war. Whether or not particular cultists embraced this strand of millennialist Christian theology, much cult ideology did expect world transformations of various sorts. Burridge (1960, 1969) summarized all cargo ideology as a new 'myth-dream': the fusion of shared hopes, desires, and imagined possibilities that a prophet or visionary brought into communal focus.

Cult Organization

Orality rather than literacy continues to characterize much of Melanesia. People communicate by talking. All sorts of rumor, news, and speculation flow from village to village – some of this still about mysterious sightings, dreams of the future, or prophetic visions. Only some of these stories attract much public attention, however, let alone a following. Anthropologists attempted to figure why cargo movements occurred in one village while bypassing another. Two important factors proposed to explain the distribution of cult outbreaks were the degree of people's sense of 'relative deprivation' – how unhappy they were with their current lives vis-à-vis the past, their future aspirations, or their neighbors – and the absence of a strong, local power structure. Where people did not feel relatively deprived, they were less likely to follow cargo prophets. And where village leaders remained powerful (whether these leaders were chiefs or what, in Melanesian Pidgin English, are called 'big-men'), they usually could deflect upstart cargo cult enthusiasm, chase away prophets, and stifle the local spread of a movement.

Cargo movements typically were sparked by local prophets who foretold the return of ancestors or of ancestral favor and prescribed what people must do in order to obtain cargo, instigating, for example, novel sorts of dance, exchange, and other ritual. Cult leaders claimed special knowledge of supernatural instruction and usually also privileged strategies to acquire such. Given island orality, prophecy spread largely by word of mouth. Prophets or their representatives sometimes itinerated throughout a region to spread the message or, emulating Christianity, they established regular holy days and headquarters where followers periodically convened. John Frum leaders, for example, established a new ritual calendar which included weekly Friday night meetings at the Sulphur Bay movement headquarters on the east Tanna coast, and an annual celebration every February 15th that celebrated cult founders. During these events, they enjoyed the opportunity to plan, circulate information, and reinforce their message among the followers.

Sometimes cargo prophets urged the revival of traditional ceremony and practice that European missionaries and officials had devalued, including dance, kava drinking, male initiation ritual, polygyny, and so forth. Others focused on ritually miming European practices and styles, including dinner tables, dress, and letter writing. Cargo prophets often instructed people to drill and march. The faithful cleared new airfields, as in Jacopetti's *Mondo Cane*, and built makeshift cargo warehouses to store wealth expected soon to arrive. Prophets also often advised followers that the ancestors required novel offerings of food, flowers, or money to be left in graveyards in order for cargo to materialize. They sometimes instructed followers to dig up crops, kill their pigs, or as did John Frum leaders on Tanna, discard all European money in order to open the gate to the cargo road. Some cults engineered the abolition of marriage exchange and incest prohibition, and prophets and followers alike engaged in unrestrained sexuality (Rimoldi and Rimoldi, 1992). Elsewhere, prophets outlawed sex entirely as ritually necessary to ensure cargo's arrival. In regions where movements became politically dominant, as occasionally happened in Tanna's John Frum areas, leaders instituted police forces and cult courts to enforce their edicts and protect movement orthodoxy.

In much of Melanesia, knowledge remains a politically valued resource and inspiration, not personal creativity, the dominant mechanism of socially legitimate knowledge production. Men achieve personal reputation and also political status with a comprehensive grasp of family genealogy, history, personal and place names, ritual procedures, and curing and divination skills. Inspired knowledge of cargo has had similar political weight. Prophetic messages of all sorts, in fact, are not unusual in Melanesia. Many people, even though today longtime Christians, continue to sense the presence of ancestral ghosts and other spirits. It is not uncommon for men and women to acquire knowledge and information in their dreams. Dreamers whose cargo messages were accepted as useful and true by a community became leaders of the movements that formed around them, although sometimes others stepped in to organize and distribute the prophetic messages that dreamers provided. Women, for instance, have also been cargo prophets, although men typically appropriated and broadcast the messages that less mobile, less discursive women received in their dreams or otherwise.

Cargo prophesies have united people – at least temporarily – into areal organizations that conjoined villages and kin groups from across a region (Worsley, 1957). Participation in many of these movements was much larger than in village and kin-based Melanesian social groupings typically of modest size. Broader social unity, in fact, was an explicit message and goal in many movements, and cult ideology focused demand for social cooperation and harmony. Prophets and leaders worked to get everyone involved in cult ritual; for example, mass dances and marches to invite ancestral arrival, or ritual procedures to wash and bless money to promote its reproduction. Novel movement rituals of communal dancing, drilling, and marching, in fact, represented and enacted new cultic social unity. Prophets also often preached against socially divisive practices of sorcery, in particular, along with other threats to group unity. Attempts to reengineer marital and sexual practice within movements typically aimed to reduce intergroup conflict and to encourage a unified community. Collapse of movement solidarity served sometimes to excuse the failure of prophecy. Cargo did not arrive because followers did not fully obey prophetic command to create a stable, harmonious community.

The history of most cargo cults was short. Followers abandoned a prophet and his movement when no cargo arrived and the world did not transform. Active culting melted back down into background, ordinary cargo desire (see Jebens, 2010). Some cargo cults, however, have been successfully institutionalized and continue to be active in different form. John Frum on Tanna, for example, which began in the late 1930s, 70 years later is managed by third and fourth generation leaders, and has elected members to Vanuatu's national parliament (Tabani, 2008). Other cargo cults have similarly been institutionalized as political parties, new religions, or both. The Peli Association, Pomio Kivung, and New Britain's Tutukuval Isukal in Papua New Guinea are successful political organizations at the local level (Lattas, 1998; Whitehouse, 1995; Billings, 2002). The Christian Fellowship Church thrives today on New Georgia, Solomon Islands as a syncretic church and notably successful business enterprise engaged in forestry, logging, and fisheries.

Cargo Cults and Melanesian Culture

Anthropologists of Melanesia used to argue that the region suffered an enduring, bedrock cargo thinking, or cargo mentality, or cargoism, or even cargo culture. This made it difficult to discern exactly when a particular cargo cult began or ended. Everyday cultic frisson kept the atmosphere charged even when no cargo prophets were actively stirring things up. One can certainly contextualize cargo belief and cult structure within enduring, fundamental patterns in Melanesian cultures. Everywhere, the exchange of goods and wealth objects is an important aspect of creating and maintaining social relationships. People give one another garden produce, pigs, mats and baskets, traditional shell and contemporary money, and other valuables to celebrate births and marriages, and to mourn deaths. Moreover, men and women earn social reputation and political influence through generous giving. Cultic focus on wealth, thus, elaborated ubiquitous concern with the political

management of economic production and exchange. And, because people believe that ancestral supervision ensures the fertility of people, gardens, and pigs, it made sense to turn to ancestors and spirits also to acquire new kinds of goods such as money or shotguns or tinned peaches. Cargo cult ceremony paralleled traditional ritual practice that ensured ancestral benevolence.

Many islanders also continue to believe that ancestors speak to them in dreams, providing important knowledge, hints of the future, and instruction for proper living. Messages about the arrival of cargo ships and planes were similar to other sorts of spiritual communication. Furthermore, some have suggested that Melanesian conceptions of time and social transformation are 'episodic' rather than developmental (McDowell, 1988). People presume that sudden transformations are normal; that one cosmic order at any moment may replace another. Prophecies of cargo's arrival, the return of the dead, and the emergence of a new world are more compelling where people do not believe that the future must develop incrementally over time from the present. Finally, cultic organization – a community of believers who follow cult prophets and leaders – resembled ordinary social organization in much of Melanesia where big-men attract followers by managing the exchange of goods and information (Cochrane, 1970). Cargo prophets, along these lines, functioned as just another sort of traditional island leader.

If cargo cults are a Pacific version of millenarian movements that erupt everywhere in times of uncertainty and change, then these underlying aspects of Melanesian culture help explain the regional particulars of cult organization and the details of cargo ideology (Trompf, 1990). More than this, some have suggested that cargo culting is an indigenous Melanesian form of political organization and change that predates colonial interference in the region. If this is the case, nativistic, revivalistic, and utopian social movements in Melanesia may not quiet down and ultimately disappear, as many have predicted, despite the end of colonialism and gradual economic development throughout the region.

Cargo Cult Futures

Cargo cults may continue to erupt, or they may prove to have been a twentieth century reaction to globalization, colonial inequalities, and the disruptions of world war. The most successful movements, however, have survived into the twenty-first century, now institutionalized as political parties, churches, and business enterprises in Vanuatu, Solomon Islands, and Papua New Guinea. Some, like John Frum on Tanna, also serve as minor tourist attractions celebrated and promoted by touristic entrepreneurs. Much of the desire and energy that once animated these social movements flow today into new Evangelical and Pentecostal Christian mission enterprises. Beginning in the 1980s, fundamentalist Christian missions based in Australia, New Zealand, and the United States notably strengthened their presence in the Pacific. Influenced by this renewed Christian millenarianism, many islanders have become involved in Holy Spirit movements. Rather than waiting for cargo ships, followers seek instead to be possessed by the Holy Ghost to bring about sudden transformations of self, society, and world. Typically, local Christian leaders often also undertake healing and antisorcery campaigns, cleansing villages of hidden sorcery paraphernalia believed to be causing illness, death, and disorder. New Christian prophets predict the Last Day – the return not so much of cargo but of Christ – and the impending establishment of a new cosmos. What could be called neo-cargoism has also sometimes been sparked by failed or successful development projects such as Lihir Island's gold mine (Bainton, 2010). As Pacific islanders are increasingly absorbed into a global, capitalist order where economic inequalities persist, and even deepen, social movements (Christian, cargoist, and otherwise) certainly will persist.

Whatever happens to cargo cults themselves in the Pacific, the label cargo cult is now widely applied – and no longer just in Melanesia (Lindstrom, 1993). Cargo culting today is anything that some critic depreciates. Any fervid desire for wealth or goods that people pursue with apparently irrational means can be condemned as cargo cultic, and the Internet abounds with examples of creative uses of the cargo cult label far almost everywhere. Almost anything might be called cargo cult if it: demonstrates fervid desire (for wealth, or whatever); involves some group or collectivity of the desirous; and if these desires, or purported means to achieve them, can be discredited as irrational. Or, "the distilled essence of cargo cult, in this most simplified delineation of the term, is a tragic relationship between rational ends and irrational means – between genuine desire and ineffective practice" (Lindstrom, 1993: 185–186). Cargo desire, though, can also sometimes be an honorable sentiment as it was, occasionally, in stories of Melanesian cultists. Good cargo culting focuses on personal development, creativity, and individual freedom while bad and tragic cargo desire merely seeks tawdry material goods or dubious political programs, quick but ineffective fixes to life's problems.

Cargo cult is a just-so story about a peculiarly modern mode of desiring that is always ultimately unrequited, unquenchable, and never ending. These underlying cultural currents of, and about, desire sustain cargo cult's continuing popularity and use, despite the particularities of its deployment online and elsewhere. As cultural moderns we take desire itself, no matter its good or bad objects, to be naturally unending. We rightly desire human perfectibility, freedom, and wealth but we would be astonished should we one day actually achieve this. Self-development is, and should be, an infinite task. Cargo cultists, both good and bad, presume desire naturally to be an unquenchable, unrequited human condition. Tales of the cargo cult have served to phrase our understandings of human desire in constant, universal form. We continue to appreciate a universal cargo cult story, now relocated from Melanesian islanders to desiring humans everywhere.

See also: Ancestors, Anthropology of; Big Man, Anthropology of; Colonialism, Anthropology of; Creolization: Sociocultural Aspects; Development: Social-Anthropological Aspects; Globalization, Anthropology of; Oceania, Sociocultural Overviews: Melanesia; Ritual; Spirit Possession, Anthropology of; Syncretism.

Bibliography

Bainton, Nicholas A., 2010. The Lihir Destiny: Cultural Responses to Mining in Melanesia. Australian National University E Press, Canberra.

Billings, Dorothy, 2002. Cargo Cult as Theater: Political Performance in the Pacific. Lexington Books, Lanham, MD.

Burridge, Kenelm, 1960. Mambu: A Study of Melanesian Cargo Movements and Their Social and Ideological Background. Methuen, London.

Burridge, Kenelm, 1969. New Heaven, New Earth: A Study of Millenarian Activities. Schocken Books, New York.

Cochrane, Glynn, 1970. Big Men and Cargo Cults. Clarendon Press, Oxford.

Guiart, Jean, 1951. Forerunners of Melanesian nationalism. Oceania 22, 81–90.

Guiart, Jean, 1956. Un siècle et demi de contacts culturels à Tanna, Nouvelles-Hébrides. Musée de l'Homme, Paris.

Jarvie, I.C., 1964. The Revolution in Anthropology. Routledge & Kegan Paul, London.

Jebens, Holger (Ed.), 2004. Cargo, Cult and Culture Critique. University of Hawaii Press, Honolulu.

Jebens, Holger, 2010. After the Cult: Perceptions of Other and Self in West New Britain (Papua New Guinea). Berghahn Books, New York.

Kaplan, Matha, 1995. Neither Cargo Nor Cult: Ritual Politics and the Colonial Imagination in Fiji. Duke University Press, Durham.

Lattas, Andrew, 1998. Cultures of Secrecy: Reinventing Race in Bush Kaliai Cargo Cults. University of Wisconsin Press, Madison.

Lawrence, Peter, 1964. Road Belong Cargo: A Study of the Cargo Movement in the Southern Madang District, New Guinea. Manchester University Press, Manchester.

Lindstrom, Lamont, 1993. Cargo Cult: Strange Stories of Desire from Melanesia and beyond. University of Hawaii Press, Honolulu.

Lindstrom, Lamont, 1996. Cargo inventories, shopping lists, and desire. In: Haviland, William, Gordon, Robert (Eds.), Talking about People: Readings in Contemporary Cultural Anthropology, second ed. Mayfield Publishing, Mountain View, CA, pp. 35–39.

Maher, Robert, 1961. New Men of Papua: A Study of Culture Change. University of Wisconsin Press, Madison.

McDowell, Nancy, 1988. A note on cargo cults and cultural constructions of change. Pacific Studies 11, 121–134.

Mead, Margaret, 1956. New Lives for Old: Cultural Transformation – Manus, 1928–1953. Morrow, New York.

Rimoldi, Max, Rimoldi, Eleanor, 1992. Hahalis and the Labour of Love: A Social Movement on Buka Island. Berg, Oxford.

Steinbauer, Friedrich, 1979. Melanesian Cargo Cults: New Salvation Movements in the South Pacific. University of Queensland Press, St. Lucia.

Strelen, John G., 1977. Search for Salvation: Studies in the History and Theology of Cargo Cults. Lutheran Publishing House, Adelaide.

Tabani, Marc, 2008. Une pirogue pour le paradis: Le culte de John Frum á Tanna (Vanuatu). Éditions de la Maison des Sciences de l'Homme, Paris.

Trompf, Garry W., 1990. Cargo Cults and Millenarian Movements: Transoceanic Comparisons of New Religious Movements. Mouton de Gruyter, Berlin.

Whitehouse, Harvey, 1995. Inside the Cult: Religious Innovation and Transmission in Papua New Guinea. Oxford University Press, New York.

Williams, Frances E., 1923. The Vailala Madness and the destruction of native ceremonies in the Gulf Division. Anthropology. Report no. 4. Port Moresby: Territory of Papua.

Worsley, Peter, 1957. The Trumpet Shall Sound: A Study of "cargo" Cults in Melanesia. Macgibbon & Kee, London.

Relevant Website

http://vanuatu.travel/discover-the-islands.html – Vanuatu Tourism Office and Cargo Cult Tours.

Carnival

Philip W Scher, University of Oregon, Eugene, OR, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Within the European Christian tradition, Carnival is the festive period that either came some days before Lent or signaled the start of Lent itself. It is a time when people abstained from eating meat, but was primarily marked by large public celebrations that included costumes, masks, music, dances and street theater. With the spread of European colonialism, Carnival became a global phenomenon. Carnivals that have developed in other regions around the world have reflected the local, indigenous or immigrant cultures that were present in those societies.

Carnival is not simply a single event celebrated in many places. Everywhere that Carnivals have taken root and grown they have become unique events unto themselves. Whatever soil they germinate from tends to influence their final flowering. In the United States the term Carnival is often applied to a variety of fairs or temporary amusement parks that have little or nothing to do with Carnival or Mardi Gras as it is known in New Orleans, Louisiana, and Mobile, Alabama. The latter are examples of New World versions of an old world festival whose origins can be traced in Europe to about the AD tenth century.

From the end of the 900s to the middle of the 1100s some reference is made in Latin documents to *carnelevare*, a word interpreted to mean literally 'to lift up the flesh, or remove the meat.' The term is used in these documents generally to refer to a regular time at which people were meant to settle debts. However, the origin of the name 'Carnival' is disputed between those who argue for a connection to the Italian word 'carne' (meat), and those that argue a link with the word 'carrus' (car). The argument here really concerns the putative origins of the modern festival. The 'carne' camp would suggest an origin within Christianity, linking the festive traditions to the Lenten season, while the 'carro' camp would allow the origins of the festival to be much earlier with pre-Christian cultural traditions. By far the most common Folk etymologies that exist agree with the connection to the later Latin expression 'carne vale,' which means 'farewell to meat.' The word carne may also be translated as flesh, and thus some interpret the expression carne vale as farewell to one's corporeal self and therefore as an invitation to free oneself of one's previous identity. The alternate origin story stems from the Roman name for the festival of the Navigium Isidis (ship of Isis). Indeed there were such processions marked by costuming, music and the carrying of what we might consider 'floats' today. However these forms of celebration, marked by these general characteristics, may be found in many contexts geographically and historically.

A document dating back to the year AD 1140 alludes to the slaughter of animals and a parade through Rome in anticipation of Lent. Lent in the Christian calendar lasts 40 days and culminates in Easter Sunday. It represents the 40 days that Jesus retreated into the wilderness to meditate, and to undertake fast. Lent is preceded by Ash Wednesday, at which time ashes taken from the burned palm fronds of the previous year's Palm Sunday are applied to the forehead of the believer in the shape of a cross. Carnival was incorporated into the Christian calendar as a final festive moment at which meat could be consumed and frivolity exercised in anticipation of the seriousness of the Lenten season.

In the fifteenth and sixteenth centuries writers began to associate the rites of Carnival with ancient Roman and Greek antecedents such as the Saturnalia, Bacchanalia, and Lupercalia. There does not appear to be any direct connection with these festivals, however, and there is a large gap, historically speaking, between the last mention of these ancient rites and the first references to medieval Carnival. What appears to be the most likely scenario for the origins of medieval Carnival is that it served as a sort of ritual magnet to surviving pagan activities. That is, the ritual of a celebration before Lent attracted to it many non-Christian practices and rituals that had been enacted at approximately the same time as Carnival. These included a host of fertility rites; agricultural and hunting rituals; and forms of sun, river, or mountain worship. Many of these rituals were undoubtedly quite ancient themselves, but had clearly gone through enormous changes through the ages before they became incorporated as features of Carnival. The costuming traditions of many of today's European Carnivals bear traces of these ancient rites just as many new world carnivals bear traces of ancient African, South Asian, and Native American/Amerindian practices.

Within the medieval European context, some of the most elaborate and well-known traditions, including Carnival parades, masquerade balls and costumed characters, were first recorded in medieval Italy with Venice emerging as the acme of the Carnival arts. From Italy, Carnival traditions spread to the Catholic nations of Spain, Portugal, and France. From France, they spread to the Rhineland of Germany, to Greece and to the French colonies of North America and the Caribbean. From Spain and Portugal, Carnival accompanied the Catholic colonization to the Caribbean and Latin America. The next section will deal with the proliferation of Carnivals in Europe followed by a discussion of Carnival's spread through colonization, conquest and empire and the addition of cultural elements from Africa, South Asia, and Native America. Carnival from the fourteenth to the sixteenth century developed into one of the most important celebrations in Europe; so much so that by the seventeenth and eighteenth centuries the Church attempted to radically reform what was perceived to be an unmitigated adventure in sin and the pursuit of pleasure. But by the early

nineteenth century, especially in the urban centers of Western Europe, Carnival had made a comeback, especially in Paris.

Northern Europe

The two major Carnivals of France are the Nice Carnival and the Paris Carnival. The Nice Carnival was held as far back as 1294, and is still held annually, attracting over a million visitors yearly during the two weeks preceding Lent. The Paris Carnival occurs after the Feast of Fools and dates back to the sixteenth century or earlier. To the north, in Belgium there are several notable Carnivals. The Carnival of Binche in Belgium has a history dating back at least to the fourteenth century. Although, like most Carnivals, the season itself begins many weeks prior, major parades are held over the three days before Lent; the most important participants are the Gilles, who go out in traditional costumes on Shrove Tuesday and throw blood oranges to the crowd. In 2003, the Carnival of Binche was listed as one of the Masterpieces of the Oral and Intangible Heritage of Humanity by UNESCO. The Carnival at Aalst, in Belgium, which is celebrated for the full week preceding Ash Wednesday, itself was inscribed in 2010.

While the most celebrated and by far the largest Carnival in England is Notting Hill Carnival (see below) historically the season immediately before Lent was called Shrovetide and was a time for confessing sins (also known as 'shriving'). This period was generally met with fewer festivities than the mainland European Carnivals and many of these did not survive the sixteenth-century English Reformation, although England has a host of other seasonal festivals. In German-speaking countries, two distinct varieties of Carnivals are observed. The Rheinish Carnival is held in the west of Germany. The Cologne Carnival is the largest and most famous of these. Unique to this celebration is 'Weiberfastnacht,' held in commemoration of an 1824 revolt by washer-women. On this day women traditionally cut off men's neck ties and are permitted to kiss any man they like. The Carnival of southwestern Germany is known as Fastnacht. This festival takes place in Baden and Swabia, Switzerland, Alsace and Vorarlberg in Western Austria. It is linked to pre-Christian rites dedicated to the expulsion of winter and the renewal of spring and growth. Carnival in the Netherlands is also called 'Vastenavond' or 'Vastelaovend(j),' and is mostly celebrated in Catholic regions of the country, mainly in the south. Dutch Carnival is officially celebrated on the Sunday through Tuesday preceding Ash Wednesday. Although there are many traditions that are region-specific, some common characteristics of Dutch Carnival include a parade, a 'prince' plus cortège ('Council of 11'), a farmer's wedding (boerenbruiloft), and eating herring on Ash Wednesday. Dutch Carnival is perhaps best remembered in the modern era through the celebrated paintings of fifteenth-century painter Hieronymus Bosch.

Southern Europe

The most famous Carnival in Italy is the Carnival in Venice which may have begun as early as the twelfth century, with masking traditions first recorded in the thirteenth century. That the festival was often seen as a threat to order by local authorities is reflected in Italy's many laws over the past several centuries attempting to restrict celebrations and the wearing of masks which are, of course, a central feature of the Carnival. Carnival celebrations in Venice ceased for many years after the city came under Austrian control in 1798, but were revived in the late 1970s.

Carnival in Portugal is celebrated throughout the country, most famously in Ovar, and Madeira. It was also introduced to the Cape Verde Islands by the Portuguese who settled there bringing along Catholic festivities and traditions to the uninhabited islands. The Portuguese also brought Carnival to their colony of Goa in India. Here the Carnival is known as 'Intruz' (from the Portuguese word Entrudo) and its most extensive observance takes place in the city of Panaji. Ironically, although Portugal introduced Christianity and the customs related to Catholic practice to Brazil, the country has begun to adopt some aspects of Brazilian-style Carnival celebrations, in particular those of Rio de Janeiro with their brilliant parades, samba music and other Brazilian performance forms.

Spain has a long and much varied tradition of Carnival celebrations. The most famous of these are in Cádiz, and Catalonia. In Cádiz, as in many Carnival traditions, the costumes often reflect recent events in the community or, increasingly, in the world. For example, during the bird flu epidemic in 2006, many revelers masqueraded as chickens. The Cádiz Carnival, like certain West Indian aspects of Carnival, is famous for the social commentary of its masqueraders such as the chirigotas, choirs, and comparsas. The chirigotas are well-known witty, satiric popular groups who sing about politics, new times and household topics, wearing the same costume, which they train for the whole year.

Catalonia has also developed a host of specific traditions within its Carnivals. Solsona's Carnival is one of the longest Carnivals in Catalonia; free events in the streets, and concerts every night, run for more than a week. The Carnival is known for a legend in which well-meaning townspeople attempting to help a donkey graze on some grasses growing atop a bell tower hoisted the animal up by its neck. To commemorate this story every year, people in Solsona hang a stuffed donkey from the tower. It then 'pisses' above the excited crowd by virtue of a water pump.

Carnival in the New World

The popularity and fame of the Parisian Carnival spread not only to other locales on the continent, but to the growing and increasingly wealthy colonies in the New World. In the 1770s the French colonies of Saint Domingue (now Haiti), Martinique, Guadeloupe, and Saint Lucia were on the rise, rapidly becoming sources of enormous wealth for France. This wealth, generated by slave labor on sugar plantations, lead to a colonial high society with plenty of disposable income and few of the pleasures of the Parisian metropole. Carnival emerged as an important pastime for the wealthy plantocracy who invested great sums in holding Carnival balls and masquerades. Participation by slaves was certain prior to emancipation, and they would have been present not only as servants but also often as musicians or performers. By 1783 the French and English had

become dominant forces in the Caribbean and, of course, in North America as well. As their empires expanded the Spanish empire diminished in the Caribbean archipelago.

Most of the islands in the Caribbean celebrate Carnival including the Dominican Republic, Antigua, Aruba, Cayman Islands, Curaçao, Barbados, Dominica, Haiti, Grenada, Jamaica, Saint Maarten, Saint Lucia, Saint Kitts, Saint Thomas, U.S. Virgin Islands, Saint Vincent, and the Grenadines which are all known for their Carnival seasons and unique celebrations. Probably the most famous and widely copied Carnival in the Caribbean is the Carnival in Trinidad. Originally a colony of the Spanish, the influx of French planters in the eighteenth century introduced Carnival to the island where it was primarily a white colonial affair. After the emancipation of the enslaved African population in 1838, however, the tone of the festival changed dramatically as more and more freed African-Trinidadians began to participate. A wide variety of Carnival characters and masqueraders evolved and bore evidence of not only African influences but also Indian influences after the arrival of indentured laborers from South Asia during the nineteenth and twentieth centuries. These characters developed slowly, often beginning as satirical reflections on people or issues in society, creatures from folklore or history or creative elaborations of French and other European Carnival figures. Other characters, like the 'Neg Jadin' or 'garden slave' were originally played by artistocrats pretending to be their own servants during Carnival. These portrayals were later taken on by emancipated Africans themselves in one of those role reversal performance traditions, like minstrelsy, so common in slave societies of the new world where the oppressed imitates the oppressors impression of him/herself. Along with the Neg Jadin there were the Dames Lorraine, the Burrokeets, the Clowns, Bats, Bears, Red and Blue Devils, Jab Jabs, Jabs Molassi, the Wild Indian bands, the Moko Jumbies (stilt walkers), the Midnight Robbers and later the multitudes of sailors and soldiers commandeering the streets and sidewalks from would-be onlookers. Each character had a unique costume and attendant performance style.

One of the masquerading traditions that horrified the upper classes the most was called the *pissenlit*. Literally the 'bed-wetter,' this piece of performance entailed a grown man sporting bedsheets or linens stained red in imitation of bloody menstrual rags. By the late 1930s and right through the 1950s and early 1960s, the sailor bands emerged as the most popular form of masquerade for the average (male) Trinidadian. Trinidadians had been imitating sailors since at least the nineteenth century when the British sailors began making regular calls at Port-of-Spain. But in the 1940s and 1950s, when Americans were stationed on Trinidad during the World War II, the sailor costume achieved the height of its popularity. Thousands of people wearing the basic white sailor outfit, some carrying talcum powder to spray on cringing spectators, would parade behind their favorite steelbands as they moved across the city. The Steel Pan (sometimes erroneously called the Steel Drum by non-Trinidadians) was invented in Trinidad and steelbands were deeply rooted in their neighborhoods and their followers were fiercely loyal to the music, the costume, and the home turf. When two steelbands, coming from rival neighborhoods met, break-out of pitched street battles, known as steelband clashes, were not uncommon. The steelbands, with glorious names like Desperadoes, Renegades, Red Army, and Invaders taken from American movies, advanced on the downtown in snaking columns of warriors holding banners that read "To Hell and Back" or "D-Day." Today much of that tradition exists on the margins of Carnival, with center stage being taken by large masquerade bands with hundreds, if not thousands of revelers in purchased costumes and dancing to prerecorded or live-performing soca bands. Increasingly, however, there has been a return to the smaller art form of masquerade with parody and social commentary at its heart. These endeavors have been pioneered mostly by local artists and musicians.

Carnival is an important cultural event on the Dutch Antilles islands of Aruba, Curaçao, Sint Maarten, Saba, Sint Eustatius (Statia), and Bonaire. Festivities on these islands in many respects parallel those of Trinidad with the exception that the J'Ouvert in the Dutch Antilles usually ends at sunrise with the burning of a straw King Momo, cleansing the island of sin and ill fortune. Humorously enough, on the island of St. Eustatius he is called Prince Stupid.

Carnival on the small island nation of Dominica has a variety of unique masquerade traditions, most strikingly in what is perhaps the oldest form of costume called the sensay. The name itself is most likely Ghanaian in origin and outfit itself is made of frayed rope and other fibrous material such as pounded leaves of the agave, and 'langue beff.' The material is tied around the body in layers so that it gives the effect of strips of monstrous plumage, which is, in effect, the West African source of this particular masquerade. In addition, a mask is worn on the face and cow horns form the headpiece. Sensay costumes are also made of strips of paper, cloth, frayed plastic sacks and dry banana leaves.

Carnival has also been celebrated in Cuba since the eighteenth century. After Fidel Castro's Communist Revolution, Carnival's religious overture was suppressed, however many of the events remained, existing with basic tolerance from the state. Carnival celebrations were in decline throughout Cuba from the 1960s, but have revived in recent years with an emphasis on the African cultural antecedents to the modern event.

In Haiti, Carnival or Kanaval exists alongside another tradition called Rara which brings in elements of the Haitian religion of Vodoun, itself deriving from belief systems in what is now Benin in West Africa. The Kanaval most likely derives from celebrations brought to what was then the sugar colony of Saint Domingue but, as in many Caribbean Carnivals, was radically altered and reshaped by Africans after the successful rebellion of 1804 which gave Haiti its independence.

In the modern era a sort of reverse colonization has developed as a result of the decolonization of the Caribbean in the 1960s.West Indian migrants to the United Kingdom, Canada, and New York, for example, have brought with them the traditions of the Carnivals of Trinidad, Haiti, and the smaller islands. The Carnivals now celebrated at Notting Hill, London, and Brooklyn, New York, however are not tied to the religious calendar of Catholicism. Taking place mostly in the late summer, such events as Caribana are what might properly be called 'ethnic display events' in which Caribbean culture is the featured element.

Carnival is celebrated in New York City in the borough of Brooklyn. As in the UK, the timing of Carnival has been

separated from the Christian calendar and is celebrated on Labor Day Monday, in September. It is called the West Indian American Day Parade or West Indian Day Carnival, and was founded by immigrants from Trinidad. The West Indian Day Carnival is one of the largest parades and street festivals in North America with approximately one million people participating or attending. The parade, which consists of steel bands, floats, elaborate Carnival costumes and sound trucks proceeds down Brooklyn's Eastern Parkway in the Crown Heights neighborhood.

The best-known, most elaborate, and most popular Carnival events in the United States are in New Orleans, where it is commonly referred to as Mardi Gras (Fat Tuesday). The New Orleans Carnival season begins after Twelfth Night, on Epiphany (January 6). It is a festive time of parades, balls (some of them masquerade balls), and 'king cake' parties. It has also been part of the winter social season for elite families at which time young women were 'presented' at débutante balls. The parades in New Orleans are organized by Carnival 'krewes.' Krewes select a Carnival king, known as Rex, and build elaborate floats. Riders atop the floats toss 'throws' such as strings of plastic colorful beads, doubloons (aluminum or wooden dollar-sized coins usually impressed with a krewe logo), decorated plastic throw cups, and small inexpensive toys.

Latin America

In Mexico, Carnival is celebrated in many cities and towns, most notably in Mazatlán, Sinaloa, Mérida, Yucatán, and in the city of Veracruz, where Carnival is celebrated with traditional music, folklore, arts, and dances. People dress in bright, feathered costumes resembling the indigenous traditions, and create a series of performances on the streets as well as on stages. In most cases there is a large set-up of fair games and roller coasters. In Bolivia, Chile and Peru La Diablada has been named a Masterpiece of the Oral and Intangible Heritage of Humanity by UNESCO. It is an essential aspect of Carnival in the Bolivian highlands and is marked by performers wearing intricate and colorful devil masks and costumes. It is the result of cultural traditions deriving both from the Spanish and indigenous Andean peoples.

The Carnival in Brazil is perhaps the most famous Carnival in Latin America, if not the world. It has its roots in the seventeenth century and is celebrated throughout the country. The most iconic expression of the Brazilian Carnival takes place in the Carnival in Rio de Janeiro, with samba schools parading in the Sambadrome (an vast stadium built in 1984 to host Carnival processions and allow viewing by mass numbers of spectators). It's the largest Carnival event in the country, and one of the largest in the world. Blocos organize community members or neighborhoods into themed samba groups accompanied by music (Bandas). There are more than 440 blocos in Rio de Janeiro.

Approaches to Carnival

Carnival is generally experienced as a chaotic, confused, and disorienting time. All the comforts of time and place tend to evaporate in a gross assault on the senses. Reason flies away. This goes as much for the scholarly interpretations of the event as for the event itself.

One often reads of Carnival as a form of resistance or, conversely, as a 'steam valve' manifesting as an allowable expression of discontent permitted by the powerful within rigidly defined social hierarchies. One may also find interpretations of Carnival as a key feature of some primordial identity suddenly threatened by commercialism, political patronage, and corruption. It is both full of profound meaning and profoundly meaningless. What accounts for this diversity of views and the denial of this diversity by many scholars?

Historically the focus on Carnival by scholars has been guided by political agendas and a woeful lack of theoretical reflection that explains the too frequent dismissal of ambiguity and contradiction in the Carnivals. The Carnival is subject to numerous forms of reductionist analysis. That is to say that many scholars are eager to demonstrate just how Carnival is a form of resistance, or hopelessly distorted by political manipulation and commercialism, or an assertion of identity that should be good for development, or the exemplar of resistance through the reversal of social strata, etc. Social scientists, historians, and writers have analyzed Carnival from a variety of perspectives often examining how the social and political contexts affect the ways in which Carnival occurs as well as the dynamics of race, gender, and class that characterize Carnival. Carnival performance has been placed at the center of a 'national culture' and given the burden of fulfilling hopes generated by an ardent nationalist rhetoric.

Judging by the literature, Carnival can be both a 'Bread and Circuses' kind of mass distraction and an occasion for political action and defiance of authority. The most illuminating analyses of Carnival take the greatest care to identify and trace the implications of political manipulations and social transformations. They socially ground and locate the controversies and the stakes involved in the debate over the festival's form and meaning. For instance *contra* the idea that Carnivals may be inherently oppositional, Stallybrass and White (1986: p. 19) suggest that nostalgia, uncritical populism, and licensed complicity are common errors in cultural nationalist or biased analyses of Carnivals.

One common way of understanding is to see it as the "collective expression of the perceptions, meanings, aspirations, and struggles engendered by the material conditions of social life and informed by the cultural traditions of the group" (Lee, 1991: 417). Another common approach to the analysis of Carnivals is to examine controversies over the nature of a group's cultural heritage that arise during the Carnival season (Cohen, 1993; Gallaugher, 1995; Green, 2007). The most important issues raised consider exactly whose culture is presented as either the national or ethnic culture. In these studies Carnival is an important platform in which cultural conflicts are expressed. Controversies about culture during Carnival tell various ethnic groups about themselves and their 'cultural citizenship.'

Finally, many contemporary studies of Carnivals place them within the broader context of the political economy of identity. Key to such a project is an understanding of the ways in which local and global histories are interwoven. For example, the Carnival in Trinidad and those in England, the United States,

Canada, and other parts of the Caribbean, are aspects of the processes of cultural change in the 'metropole' and in the 'periphery.' Yet, to prioritize either of these locations is to lose sight of the active and on-going relationships those so-called core locations and peripheral ones have been involved with for centuries. The histories of all of the Carnivals under consideration here occur as a result of similar, sometimes the same processes. The particular configuration that any Carnival may take and the role it may play in the lives of those who participate in it and those who do not depend upon the particular conjunctions of class, gender, ethnicity, and generation as well as broader structures of community and nation found in each locale.

This approach might be characterized as a kind of political economy of identity that compels one to consider the place of 'class, capitalism, and power' in the constitution of localized subjects, yet can also account for subjects who may occupy multiple localities as can be argued in the case of the transnational communities. In that regard scholarship cannot separate the analysis of cultural identity, cultural practices, and cultural change from an analysis of work, power, and politics. The key factors informing these analyses include shifting labor markets and immigration policies, struggles for political recognition, economic autonomy, and cultural legitimacy. Ideally, scholarly approaches to these festivals will locate the social and cultural phenomena of the Carnivals within an analysis of the conditions of existence of those who make it, watch it, experience it, and talk about it within particular historically constituted structures of power and representation that influence those activities (Roseberry, 1988).

There is increasing recognition of the futility of reducing Carnival to any singular kind of entity with universally realized qualities and attributes. What development bureaucrats, corporate sponsors, political functionaries, culture ministers, folks on the street, and scholars may speak of when they use the term Carnival, may be quite different from one another. What any one of these figures may argue is the 'true essence' of the Carnival must be situated within any number of political, economic, and even cultural projects. The invocation of 'the Carnival' calls forth idealized images of resistance, the assertion of identity, of national pride and struggle that resonate in myriad ways for different peoples each engaged in located struggles for economic autonomy, cultural integrity and respect, and political power. The work of anthropologists, ethnomusicologists, art historians etc. demonstrates the polysemic nature of the Carnival as a socio-, cultural, economic phenomenon. The event is composed of numerous events and activities none of which may be said to be of greater importance than another. Whatever these Carnivals may be they are the products of histories, a manifestation of social tensions, a barometer of cultural change, and a crucible for creating, discovering, and asserting identities. The Carnivals encompass contradictory ideas and practices. People in position to assert their visions of what the Carnivals are seek to extinguish or at best minimize alternative interpretations and activities by rewarding those representations and practices that conform with their vision and by punishing or condemning those that do not. That in itself is part of the masquerade.

See also: Americas, Sociocultural Overviews: Caribbean; Catholicism; Dance, Anthropology of; Folklore; Music: Anthropological Aspects; Performance: Anthropological Aspects; Ritual.

Bibliography

Abrahams, Roger D., 1983. The Man-of-words in the West Indies: Performance and the Emergence of Creole Culture. Johns Hopkins University Press, Baltimore.

Aching, Gerard, 2002. Masking and Power: Carnival and Popular Culture in the Caribbean. University of Minnesota Press.

Bettelheim, Judith, 1990. Carnaval in cuba: another chapter in the nationalization of culture. Caribbean Quarterly 36 (3 and 4), 29–42.

Cohen, Abner, 1993. Masquerade Politics. University of California Press, Berkeley.

Cowley, John, 1999. Carnival, Canboulay and Calypso: Traditions in the Making. Cambridge University Press, pp. 309.

Danow, David K., 1995. The Spirit of Carnival: Magical Realism and the Grotesque. University Press of Kentucky, Lexington.

DaMatta, Roberto, 1991. Carnivals, Rogues and Heroes. University of Notre Dame Press, Notre Dame.

Gallaugher, Annemarie, 1995. Constructing Caribbean culture in Toronto: the representation of Caribana. In: Ruprecht, A., Taiana, C. (Eds.), The Reordering of Culture in the Hood: Latin America, the Caribbean and Canada. Carelton University Press, Ottowa.

Gilmore, David D., 1998. Carnival and Culture: Sex, Symbol, and Status in Spain. Yale University Press, New Haven.

Green, Garth L., Scher, Philip W. (Eds.), 2007. Trinidad Carnival: The Cultural Politics of a Transnational Festival. Indiana University Press, Bloomington.

Guillermoprieto, Alma, 1991. Samba. Vintage Books, New York.

Harris, Max, 2003. Carnival and Other Christian Festivals: Folk Theology and Folk Performance. University of Texas Press, Austin.

Hill, Errol, 1997. The Trinidad Carnival: Mandate for a National Theatre. New Beacon Books.

Humphrey, Chris, 2001. The Politics of Carnival: Festive Misrule in Medieval England. Manchester University Press, Manchester.

Kendrick, Christopher, 2004. Utopia, Carnival, and Commonwealth in Renaissance England. University of Toronto Press, Toronto.

Kinser, Samuel, 1990. Carnival American Style, Mardi Gras at New Orleans and Mobile. University of Chicago Press, Chicago.

Lee, Ann, 1991. Class, race, color and the trinidad carnival. In: Ryan, S. (Ed.), Social and Occupational Stratification in Contemporary Trinidad. St Augustine: Institute of Social and Economic research. University of the West Indies.

Mitchell, Reid, 1999. All on a Mardi Gras Day: Episodes in the History of New Orleans Carnival. Harvard University Press, Cambridge.

Stallybrass, Peter, White, Allon, 1986. The Politics and Poetics of Trangression. Cornell University Press, New York.

Caro Baroja, Julio (1914–95)

Andreas Hess and Mayte Martín, University College Dublin, Dublin, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Julio Caro Baroja was one of the founding fathers and most influential figures in Spanish social anthropology and cultural history during the Franco years and the *transición*. Veering between Madrid and the Basque Country and Navarre and defending a liberal agenda he always kept a distance to the Franco regime and its academic supporters. Caro Baroja's academic career opportunities were severely limited because of his opposition to the Franco regime and because he spoke out against its isolationist policies. This was also the main reason why he opted out of academia to become an independent scholar. However, the distance from official institutions would not diminish the role and the impact his writings had on shaping modern Spanish social anthropology and cultural history.

Introduction

Julio Caro Baroja (1914–95) was one of the founding fathers of Spanish and Basque anthropology. His work comprises over 30 books and numerous book chapters, contributions to academic journals, newspaper, and magazine articles, not to speak of the many essays he wrote for exhibition and museum catalogs. Marginalized as an academic and intellectual during the Franco dictatorship, his liberal-pluralist views – which also marked his scholarly work – served as a much valued counterweight against the powerful Spanish nationalist cause and its 'academic' use of folklore and traditions. Although working as an independent scholar for most of his life, his writings became an important point of reference in anthropological discussions particularly during the transitional years that followed the death of Franco. Today his impact is still detectable. His work is cited with regular frequency. Some of his books have even achieved the status of modern classics. One of Caro Baroja's greatest lasting legacies as an anthropologist is perhaps the fact that the various cultures of the Iberian peninsula are now studied without resorting to the rhetoric of race, biological explanations of ethnicity, or narrow nationalist interpretations of culture.

Caro Baroja's writings do not fit easily traditional academic labels or descriptions. It veers between a cultural anthropology that is reminiscent of Franz Boas's work but also contains detailed ethnographic descriptions and biographical accounts. On occasion, Caro Baroja's work pays homage to intellectual traditions and paradigms that seem nowadays somewhat dated, such as the German–Austrian dominated *Kulturkreislehre*. However, it can be argued that Caro Baroja always managed to stay away from unthinking applications or anything that smelled of theoretical dogma. While some of his oeuvre has been translated into other languages, his work continues to be studied mainly by Spanish-speaking anthropologists or cultural historians who have an interest in the pluralistic cultural landscape of the Iberian peninsula. Impact and influence there certainly was, but Caro Baroja's writings deserve to be mined and to be critically evaluated beyond those interested in Spanish-speaking anthropology. In particular, scholars interested in cultural sociology and cultural history will find in Caro Baroja's work much to appreciate.

Education and Training

Julio Caro Baroja was born in 1914 in Madrid. His father was a well-known publisher and his mother played an active role in Madrid's cultural scene. His uncle, the writer Pio Baroja, would also exercise a considerable intellectual influence over him. For most of his life, Caro Baroja commuted between the two family homes in Madrid and Vera de Bidasoa, in the north-eastern part of Navarre that borders on the Basque province of Gipuzkoa. The fact that Caro Baroja was brought up in Madrid and in the Basque Country and Navarre would shape and influence his academic career and his later life choices considerably. He first studied ancient history at Complutense in Madrid but his studies were interrupted by the Spanish Civil War. After a longer spell in Itzea, the residence of the Baroja family in Vera, he eventually returned to Madrid to resume his studies in 1939 and 3 years later was awarded a PhD in history for his thesis 'Traditional cults and rituals in Spanish folklore.' Caro Baroja had been extremely lucky in having been considered unfit for military service. Not being drafted meant first and foremost avoiding potentially challenging situations. As an intellectual, bookworm, or 'egghead' – in Francoist Spain these things were interchangeable – Caro Baroja would have aroused all kinds of comments from both superiors and 'comrades.'

Even before achieving academic distinction, Caro Baroja had already two longer book-length studies in anthropological journals to his credit. He was seen as a promising young student and over the course of two summers in 1930 and 1931 he became part of the pioneering excavating effort of a group of Basque archaeologists and social anthropologists under the leadership of José-Miguel Barandiaran and Telesforo de Aranzadi, the founders of Basque anthropology. From Barandiaran he learned about Malinowksi, Durkheim, and the Durkheimians. At the university he was in close contact with Ortega y Gasset who introduced him to the work of Wilhelm Schmidt, one of the exponents of the then fashionable theory of cultural cycles and cultural areas, and to the biological and early environmental models of Jakob von Uexküll. There were also other academic visitors from Germany, such as Hugo Obermaier and Herman Trimborn, who spent considerable time in Madrid and from whom Caro Baroja learned 'the tricks of the trade.' After having graduated, he worked first as an assistant in the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61172-8

Department of Ancient Spanish History and Dialectology and Popular Traditions at the University of Madrid. Later this activity partly overlapped with his serving as the Director of the Museum of the Spanish People, a position which he held from 1944 until 1954.

As a young academic and subsequently in his administrative role as director, Caro Baroja was part of the intellectual scene in Madrid, or at least what was left of it after the Civil War. Being a liberal-minded young man who had sympathized with the republic but who also tended to avoid the extreme choices that many in the crisis-ridden Spanish Republic seemed to prefer, he tried to hang on if not to a publicly declared liberal agenda (something that was impossible at the time) then at least to some sense of free academic thinking. For example, in his role as academic expert and director of the anthropological museum, he led various efforts against the isolationist policies of the Franco regime and its devastating effects on academic life. He played host to visitors and became an important informant to those who had an interest in the cultures of the peninsula: among them were George M. Foster from the Smithonian, Walter Starkie, the Director of the British Institute in Madrid, and Julian Pitt-Rivers from the University of Chicago who conducted field work in Andalucia and who greatly valued Caro Baroja's advice and support.

Caro Baroja's growing interest in French, English, and American anthropology was against the grain of much that Madrid's social and political elite were thinking at the time. Particularly during the Second World War his growing international network aroused some suspicion. His contacts with the Smithsonian and the anthropologist Pitt Rivers, the British Council, and Walter Starkie were subject to the authorities' scrutiny. Later, during the Cold War the new relationship between the Western Allies and Franco's regime made it possible for him to visit the United Kingdom (in 1951) and the United States (in 1952) where he met Evans Pritchard and Alfred Kroeber, respectively – a rare visit for a Spanish scholar in those times. Despite his international leanings, Caro Baroja remained skeptical about any particular schools of thought or anthropological paradigm. His interest and training in history made him more receptive toward cultural history and what could be described as a cultural anthropology of the Boas type.

The Study of Culture Applied to the Iberian Peninsula

In 1949a, Caro Baroja published *Análisis de la cultura: Etnología, historia, folklore* (*Cultural Analysis: Ethnology, History, Folklore*). The text was based on lectures given in Barcelona in 1946. In the book the author paid reference to the German and Austrian cycle-of-culture approach, British functionalism, and American cultural anthropology – but also Durkheim and the anthropological agenda that followed the Durkheimian approach; he even discussed linguistic approaches like that of de Saussure. However, we can also see where Caro Baroja differed from all these paradigms by insisting on the materiality of culture(s) and its historical dimensions. The final part of the book discussed a number of examples ranging from folklore to ethnography, which threw light on what the author meant by 'materiality of culture.'

In contrast to American anthropologists like Boas or Kroeber, Caro Baroja also maintained a strong interest in the study of individuals and distinguished characters. It was not so much that he admired charismatic personalities but rather that the study of some individuals almost functioned like a prism. He was not so much interested in individuals or individualism per se but in the plasticity and variety of cultural and social life. Sometimes, he argued, the particularities of historical and cultural change were better studied by looking through the lens of an individual's life and work than studying larger social groups or subpopulations; in any case, the two perspectives should not be regarded as being mutually exclusive but should help to enrich our understanding of social and cultural life.

The same year *Análisis de la cultura* was published saw also the publication of Caro Baroja's most important work, *Los Vascos* (*The Basques*). A few years earlier he had already made some important inroads into the study of Basque culture. His first publication had been *Tres estudios etnograficos relativos al País Vasco* (*Three Ethnographic Studies Related to the Basque Country*) (1934), a presentation of the findings of his work with Barandiaran and Aranzadi. This was followed by an impressive output over the next few years: *Los pueblos del Norte de la Peninsula Ibérica* (*The Northern People of the Iberian Pensinsula*) (1943), *La vida rural en Vera de Bidesoa* (*Rural Life in Vera de Bidesoa*) (1944), *Los pueblos de España* (*The Peoples of Spain*) (1946), and *Materiales para una historia de la lengua vasca en su relación con la latina* (*Documents for a History of the Basque Language's Relation to Latin*) (1946).

While *Los pueblos del Norte* and *Los pueblos de España* were more comparative in outlook, the other titles obviously addressed the social and cultural conditions of the Basque country and Navarre, including the rural community in Vera, which he increasingly considered to be his home. As he confessed in an autobiographical account many years later, he liked Madrid as a city that would always remain associated with his childhood and youth but he was not a 'castillanista.' Particularly during the bleak years in which the Franco regime consolidated its power, Madrid seemed often rather monoglot and one dimensional. In contrast, small-town life in Vera, Navarre, and the nearby cities along the Basque coast appeared to be more varied and full of life and stories.

Many of Caro Baroja's Basque writings would become testimonies to the rich small-town and rural life in the Basque Country and Navarre – an argumentation that was directed against the all-too-common perception that it was only cities which provided excitement and a colorful life, while the countryside by definition had to be dull and monotonous. Spending long periods in Vera held also the major attraction of conversations and travels with his uncle Pio Baroja and an exposure to a different kind of learning that included the reading of the humanist classics. His uncle's library provided Caro Baroja with a second, humanist education, which was very much in contrast to his 'official' student years. This second education had the advantage of the student being able to choose his own curriculum and to read at his own pace, and according to his own humanist preferences. Some of the titles Caro Baroja read were in the language in which they were originally written. His Latin was certainly above average. He also mastered Italian, Portuguese, French, and some German. His English was passable but, by his own admission, far from

perfect – so that he would always try to improve it. His knowledge of Basque was certainly good enough for reading purposes and for basic communication, an astonishing achievement for somebody who had not been raised in or constantly exposed to the Basque-speaking environment.

Like the other studies which preceded it, *Los Vascos* appears to be remarkably restrained in terms of sympathizing with any particular anthropological approach or paradigm. In the introduction to the book, the author makes allusions to environmental and functionalist approaches of the cultural cycles type; however, in the same introduction, the author also states that society needs to be studied as a totality and that it would be wrong to single out one particular function or cultural cycle over any other. To compartmentalize would limit our understanding. Caro Baroja's notion of functions seems particularly interesting. His unique historical method does not see some primordial function as being specific to a particular past period or an earlier cycle but sees functions rather as having been part of a cultural inheritance that shaped and continue to shape the present in particular ways, yet without making that society totally path dependent on such functions.

It is the openness to change and modification which is most interesting in Caro Baroja's observation of cultures. In particular it is the unique combination of structure/function *and* openness that marks Caro Baroja's 'liberal' anthropology of the Basques and which also contributes to a rejection of the false alternative of universalist versus nationalist-particularist values and perspectives. From there it was but a short step to argue that culture as a concept should be given clear preference over concepts such as 'race,' 'ethnic belonging,' 'homogeneous people,' or 'national communities.' Thus, *Los Vascos* could be considered as a rejection of, and a rupture with, essentialist concepts and theories. Instead, Caro Baroja's study offered thick descriptions of cultural configurations, each of which alluded to some particular cultural and human geographical context and function, such as the countryside, the sea, and different types of human settlements like the *auzola* (community neighborhoods), village, or town. Retrospectively it appears almost as if *Análisis de la cultura* provided the theoretical scaffolding, including discussions of examples of what and what not to do, while *Los Vascos* became the free-standing edifice that remained after the theoretical scaffolding had been removed.

In 1954 Caro Baroja experienced a personal crisis after a string of deaths in the family. The Baroja family had always consisted of distinguished characters who contributed to the arts and to public life in general, even under the difficult circumstances after the Spanish Civil War. The death of his mother in 1950 followed by his two uncles, Ricardo in 1953 and Pio in 1956, was experienced by Caro Baroja as a huge loss. He felt increasingly isolated and without intellectual and social support. Never an admirer of academic life and even less a fighter interested in achieving academic distinction for its own sake, but also tired of the increasing isolationist positions of the Franco regime which made his international contacts look increasingly suspicious and a one-man affair, he resigned from all academic and administrative positions. His financial independence, secured by the royalties of his uncle's literary legacy and his parents' inheritance, allowed him to continue to work as an independent scholar without the burden of academic duties and responsibilities and without running the risk of becoming politically suspect.

From Spain to the Sahara and Back

It was anthropological curiosity but perhaps also the chance to escape from the pain caused by the death of his beloved mother that made Caro Baroja embark on something totally different. Known to the local colonial authorities and partly supported by them, Caro Baroja spent some periods in 1952 and 1953 in the Maghreb and in particular what is known today as Western Sahara. To be sure, Caro Baroja was not the first to study Sahauri culture, and like its French or British counterpart the Spanish colonial power (or what was left of it after the loss of Cuba and the Philippines in 1898) showed some interest in the anthropologist's field work, particularly when it came to information concerning language, local customs, and traditions. Conducting fieldwork, Caro Baroja used a translator, which naturally limited his scope of understanding. Still, the field work and investigation allowed for some fascinating insights into Sahauri nomad culture, as the two publications *Estudios saharianos* (*Sahauri Studies*) (1955) and *Estudios mogrebies* (*Maghreb Studies*) (1957) clearly demonstrate. For the author, these studies constituted something new because for the first time he had left familiar terrain and the intellectual comfort zone to study a culture which was clearly distinct from anything he had observed before.

It would be mere speculation to say that there is a straightforward link between studying Sahauri and Maghreb nomads and returning to Spain to study outsider cultures of the Iberian peninsula. However, one might wonder about the connection between the two preoccupations because over the next two decades Caro Baroja dedicated significant time writing about those who did not fit easily into what was considered to be a homogeneous Spanish culture: witches, Jews, and non-Christian practices and practitioners. *Los Brujas y su mundo* (*Witches and Their World*) (1961) became, after *Los Vascos*, his most famous study. It was followed by *Vidas magicas e inquisición* (*The Lives of Magicians and the Inquisition*) (1967) and *Inquisición, brujería y criptojudaismo* (*Inquisition, Witchcraft and Crypto-Judaism*) (1970). Earlier, in *Razas, pueblos y linajes* (*Races, Peoples and Lineages*) (1957) Caro Baroja had already argued that the concept of race was of little or no value. He also argued against the officially supported view that Jews played no role in Spanish culture in his three-volume study *Los judios en la España moderna y contemporanea* (*Jews in Modern and Contemporary Spain*) (1962).

That the cultures of the Iberian peninsula were not as unitarian as Francoist ideology reaffirmed was a constant theme in Caro Baroja's work and culminated in a critical study which openly challenged the official ideology, *El mito del carácter nacional* (*The Myth of National Character*) (1970). Some of Caro Baroja's work about customs and folk traditions looked at the rich pagan traditions that prevailed in the midst of Christian Spain, another proof that Spain was not one homogenous cultural entity as Francoist cultural policies claimed. *El carnaval* (*The Carnival*) (1965), *La estación de amor* (*The Season of Love*), *El laberinto vasco* (2003), and *El estío festivo* (*The Festive Summer*) (1984) are all thick descriptions of local or

regional festive days and celebrations, which challenged the hegemony of Catholic dogmatism and its political-cultural usages. Obviously Caro Baroja continued to follow in the tracks of his atheistic uncle Pio Baroja. However, in contrast to his uncle, the agnostic social anthropologist nephew continued to show a deep interest in the cultural-anthropological dimensions of both pagan and Christian celebrations as is evident from studies like *Las formas complejas de la vida religiosa* (*Complex Forms of Religious Life*) (1978), *De la superstición al ateismo* (*From Superstition to Atheism*) (1974), and *Ritos y mitos equivocos* (*Rites and False Myths*) (1974).

The Study of Basque and Navarrese Culture

Despite his move into other themes and areas, there remained for Caro Baroja an intellectual homeland to which he would return time and time again – the study of culture of the Basque Country and Navarre. Particularly from the early 1970s onward, his output amassed to what looks now like an impressive number of books, the multivolume sets *Etnografía historica de Navarra* (*Ethnografic History of Navarre*) (3 volumes) (1972) and *La casa en Navarra* (*Homes in Navarre*) (4 volumes) (1982), and the series *Estudios Vascos* (*Basque Studies*) (18 volumes) (1973–1988) being the most visible signs. What was new, however, but which did not come as a total surprise if one thinks about Caro Baroja as a committed liberal, was what could be called an 'individual' or 'subjective' turn in his work. This is most visible in his family portrait *Los Baroja* (*The Barojas*) (1972) and the two essay collections *Biografías y vidas humanas* (*Biographies and Human Lives*) (1986) and *Los hombres y sus pensiamientos* (*Men and Their Thoughts*) (1989).

The most distinct study of this later phase in Caro Baroja is his impressive *Los Vascos y su historia a través de Garibay* (*The Basque and Their History according to Garibay*) (1972), a study of the Mondragon chronicler and educator Esteban de Garibay (1533–99), who spent most of his life between the Basque Country and Navarre and Madrid. In the final chapter of the book it is evident that this veering and intellectual traveling between the Basque lands and the capital constituted a kind of Barojaian hermeneutic cycle, which allowed the social anthropologist and cultural historian to gain a unique perspective and observation point, which in turn allowed him to gain distance from both Spanish and Basque nationalisms.

It was such critical distance and perspective that not only informed his late writings on the profession itself, like *La aurora del pensamiento antropológico* (*The Golden Age of Anthropological thinking*) (1983), *Los fundamentos del pensamiento antropológico moderno* (*The Foundations of Modern Anthropological Thought*) (1985), and *Las falsificaciones de la historia en relación con la de España* (*Historical Falsifications in Relation to Spain*) (1991) but also allowed for some genuine insights into what it meant to be Basque, Spanish, or both. It was in thought-provoking essay collections such as *Terror y terrorismo* (*Terror and Terrorism*) (1989), *Ser o no ser vasco* (*To Be Nor Not To Be Basque*) (1998), and *El laberinto vasco* (*The Basque Labyrinth*) (2003) in which Caro Baroja's liberal thinking truly found its mature expression and through which he became known to a wider audience, almost to such an extent that he turned into a public intellectual (which is perhaps a bit of a misnomer since the label does not honor his wish for a quiet and private life, preferably in his retreat in Itzea). The essay was the most ingenious, explorative, and experimental form which allowed Caro Baroja to think about the complications of what it meant to be or not to be Basque.

Caro Baroja remained active and continued to publish until his death in 1995. Late acknowledgment came not only in the form of awards, prizes, and Festschriften but also in the form of increased academic praise and acknowledgment. The historian Juan Pablo Fusi called him 'the most important personality of contemporary post-war Spanish historiography,' and the renowned American anthropologist William A. Douglass considers the works of 'Don Julio' and particularly *Los Vascos* to be classics in the field. Such praise seems justified. Particularly Caro Baroja's Basque writings constitute an invaluable body of work indeed. Not only anthropologists but also cultural sociologists and cultural historians have come to rely heavily on Caro Baroja's findings and will, predictably, continue to do so in the future. His most important contribution, however, still consists in having considerably shaped the modern anthropological discourse in Spain, particularly in the difficult times of the Franco regime and then later during the Spanish *transición*. That Spanish and Basque social anthropology today avoid the use of concepts such as 'race,' or making dubious references to blood types or genetic differences, is not at least due to Caro Baroja's subtle influence. He not only argued for another, more culturally oriented anthropology but also led by example, as can be seen in his most important work such *The Basques* or his book about Garibay.

See also: Anthropology: Overview; Collective Memory, Anthropology of; Cultural Critique, Anthropological; Cultural History; Cultural Sociology, History of; Dictatorship in History; Ethnography; Europe: Sociocultural Overview; Nation, Sociological Aspects of; National Character.

Bibliography

Caro Baroja, J., 1934. Tres estudios etnograficos relativos al Pais Vasco. Caro Raggio, Madrid.
Caro Baroja, J., 1943. Los pueblos del Norte de la Peninsula Ibérica. CSIC, Madrid.
Caro Baroja, J., 1944. La vida rural en Vera de Bidesoa. CISC, Madrid.
Caro Baroja, J., 1946. Los pueblos de España. Barna, Barcelona.
Caro Baroja, J., 1946. Materiales para una historia de la lengua vasca en su relación con la latina. Acta Salamanticensia, Salamanca.
Caro Baroja, J., 1949a. Análisis de la cultura: Etnología, historia, folklore. CISC, Barcelona.
Caro Baroja, J., 1949b. Los Vascos. Biblioteca Vascongada de Amigos del País, San Sebastián.
Caro Baroja, J., 1955. Estudios saharianos. CSIC, Madrid.
Caro Baroja, J., 1957. Estudios mogrebies. CISC, Madrid.
Caro Baroja, J., 1957. Razas, pueblos y linajes. Revista de Occidente, Madrid.
Caro Baroja, J., 1961. Los Brujas y su mundo. Revista de Occidente, Madrid.
Caro Baroja, J., 1962. Los judios en la España moderna y contemporánea. Arión, Madrid.
Caro Baroja, J., 1965. El carnaval. Taurus, Madrid.
Caro Baroja, J., 1967. Vidas Magicas e Inquisición. Taurus, Madrid.
Caro Baroja, J., 1970. Inquisición, brujería y criptojudaismo. Ariel, Barcelona.
Caro Baroja, J., 1970. El mito del carácter nacional. Seminarios y Ediciones, Madrid.
Caro Baroja, J., 1972. Los Baroja. Taurus, Madrid.

Caro Baroja, J., 1972. Los Vascos y su historia a través de Garibay. Txertoa, San Sebastián.
Caro Baroja, J., 1972. Etnografía historica de Navarra, 3 vols. Aranzadi, Pamplona.
Caro Baroja, J., 1974. De la superstición al ateísmo. Taurus, Madrid.
Caro Baroja, J., 1974. Ritos y mitos equivocos. Istmo, Madrid.
Caro Baroja, J., 1978. Las formas complejas de la vida religiosa. Akal, Madrid.
Caro Baroja, J., 1979. La estación de amor. Taurus, Madrid.
Caro Baroja, J., 1982. La casa en Navarra, 4 vols. Caja de Ahorros, Pamplona.
Caro Baroja, J., 1983. La aurora del pensamiento antropológico. CSIC, Madrid.
Caro Baroja, J., 1984. El estío festivo. Taurus, Madrid.
Caro Baroja, J., 1985. Los fundamentos del pensamiento antropológico moderno. CSIC, Madrid.
Caro Baroja, J., 1986. Biografías y vidas humanas. Txertoa, San Sebastián.
Caro Baroja, J., 1989. Los hombres y sus pensiamientos. Txertoa, San Sebastián.
Caro Baroja, J., 1991. Las falsificaciones de la historia en relación con la de España. Circulo de Lectores, Barcelona.

Essay Compilations

Caro Baroja, J., 1973–1988. Estudios Vascos, 18 vols. Txertoa, San Sebastián.
Caro Baroja, J., 1989b. Terror y terrorismo. Plaza & Janés-Cambio 16, Barcelona.
Caro Baroja, J., 1998. Ser o no ser vasco. Espasa, Irun.
Caro Baroja, J., 2003. El laberinto vasco. Caro Raggio, Madrid.

English Translations

Caro Baroja, J., 1964. The World of the Witches. Weidenfeld and Nicholson, London.
Caro Baroja, J., 2009. The Basques (With an Introduction by William A. Douglass). Center for Basque Studies, Reno, NV.
Caro Baroja, J., 2011. The Selected Essays of Julio Caro Baroja (Compiled and with an Introduction by Jesús Azcona). Center for Basque Studies, Reno, NV.

Carrying Capacity of the Environment

Cang Hui, Centre for Invasion Biology, Department of Mathematical Sciences, Stellenbosch University, Matieland, South Africa; and African Institute for Mathematical Sciences, Muizenburg, South Africa

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.E. Geores, volume 10, pp. 7038–7039, © 2001, Elsevier Ltd.

Abstract

Catton defines carrying capacity as the maximum persistently supportable load of a focal population in a bounded environment. Here, definitions of carrying capacity for natural populations, ecosystems and humans, as well as biocapacity, are presented, with issues fully discussed pertaining to their proper interpretation, models and estimation, debate and derived paradoxes. With the explosion of the human population, the threat of overstepping the earth's carrying capacity has become a sword of Damocles, hanging over humanity and urging it to choose a sustainable path of development. Carrying capacity has become arguably the most important concept in the era of Anthropocene.

Humans are fascinated by growth dynamics, from the weight of an infant to the size of an economy. Two intrinsic questions on any growth dynamics are how fast and how long the system can grow. That is, the rate and the bound of growth. An unbounded growth is infeasible due to limited resources, and carrying capacity is a measure of this limitation. In 1798, Thomas Malthus proposed an exponential growth curve in his *Essay on the Principle of Population*, stating that a population with constant fecundity greater than its mortality will grow exponentially till it causes inevitable conflicts with the limited supply of resources. Deminishing resources and the expanding population will inevitably lead to scramble competition. Malthus's idea has further inspired Charles Darwin to adopt resource competition as one prerequisite of evolution via natural selection. Darwin said in his *Autobiography* that, "I happened to read for amusement Malthus on *Population* and being well prepared to appreciate the struggle for existence which everywhere goes on from long-continued observation of the habits of animals and plants, it at once struck me that under these circumstances favourable variations would tend to be preserved, and unfavourable ones to be destroyed. The result of this would be the formation of new species." Such struggle of life will ensure only the survival of the fittest, or equivalently as Russel Wallace put, the elimination of the unfit. Such a density-dependent mechanism of population regulation that intense competition happens when resource demand approaches supply was formally formulated by Pierre Verhulst in 1838, by adding a parameter K into the Malthusian equation to curb the unbounded growth. This equation is now well-known as the logistic equation, and the parameter K the carrying capacity.

Population Carrying Capacity

Catton (1986), defined an environment's carrying capacity as its maximum persistently supportable load. In ecology, carrying capacity is normally defined as the maximum population of a given species that can be supported indefinitely in a defined habitat without permanently impairing the productivity of that habitat (Rees, 1992). The concept of carrying capacity is rooted in the logistic equation depicting population dynamics under simple density-dependent regulation and has been developed across different hierarchical levels of living systems and in many sectors of human society (Seidl and Tisdell, 1999; Monte-Luna et al., 2004). In the Malthusian equation, the probability that an organism will reproduce or die is assumed to remain constant and is independent of the density of the population. Obviously, this can only be true when there is no competition among the individuals, normally when the population density is low. The growth of any population in a limited environment will eventually halt due to the shortage of resources. When the demand of the existing population on the resource (i.e., the population size times the basic per capita resource consumption rate required for maintaining life) is equal to the rate of resource supply, the population will reach its saturation level. This saturation level is decided by both the resource supply and the per capita resource consumption rate, called the carrying capacity of the environment for the focal population. Using the language of differential equations in mathematics, the Verhulst logistic equation reads as follows:

$$\frac{dN}{dt} = rN\left(\frac{K-N}{K}\right)$$

where N is the population size; dN/dt the derivative of population size, indicating the change rate of the population size; r the intrinsic population growth rate, indicating how fast the population can grow at small size; K the carrying capacity. The Malthusian equation only includes the first part of the right term, $dN/dt = rN$, meaning that the population change rate is proportional to its current population size. It also means that the per capita rate of increase $(dN/dt)/N$, often an indicator of population fitness, is a constant and thus density independent. In the logistic equation, the per capita rate of increase, $r(K-N)/K$, is a declining function of current population size N, thus becoming density dependent. In **Figure 1**, we can clearly see the difference between Malthusian's exponential growth curve and the saturation form of the logistic equation. Clearly, when K approaches infinity, the logistic equation becomes the

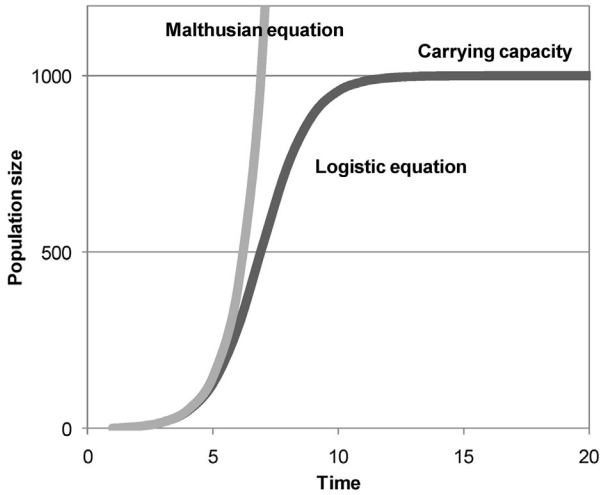

Figure 1 An illustration of population dynamics governed by the Malthusian equation and the logistic equation. Parameter values: $r = N_0 = 1$ and $K = 1000$.

Malthusian equation. The exact solution to the logistic equation exists,

$$N(t) = \frac{K \cdot N_0 \cdot \exp(r \cdot t)}{K + N_0(\exp(r \cdot t) - 1)}$$

where N_0 is the population size when time $t = 0$. When time t increases, the population size will converge to the carrying capacity K (**Figure 1**). The logistic equation is the simplest model, which contains a carrying capacity for the population. In fact, carrying capacity exists in many models that implement the negative density-dependence term, i.e., $(dN/dt)/N$ is a declining function of N.

Because the carrying capacity is also the asymptote, or the equilibrium of the population size in the logistic equation, the concept of carrying capacity has often been confused with population equilibrium, leading to unnecessary confusion and so-called paradoxes. In practice, many studies have considered the maximum population equilibrium or maximum population size as the carrying capacity, making the distinction even muddier. In 1969, Richard Levins hypothesized that if r in the logistic equation is negative, then we will have a positive dN/dt when $N > K$ and thus the population size will increase indefinitely. This is clearly counterintuitive. Although constraining r to be only positive can make Levins' paradox disappear, many declining populations are indeed experiencing the situation of $r < 0$, thus such a constraint could exclude many biologically meaningful scenarios. Although constraining r and K to be the same sign could resolve Levins' paradox, such a constraint nonetheless incurs a new problem on the meaning of negative carrying capacity ($K < 0$). The negative carrying capacity could be a measure of just how unfavorable the environment is (Hui, 2006). An important finding from metapopulation theory is that the source–sink dynamics are such that where a local population is in a 'sink habitat' a positive equilibrium can still be maintained due to contributions from migrants from source populations. These sink patches have a negative carrying capacity but support a positive population equilibrium. Clearly, the carrying capacity of a sink habitat is not equal to the maximum population equilibrium that can survive through the inflow of propagules, known as the rescue effect. This highlights the difference between carrying capacity and the maximum population equilibrium. Indeed, many preindustrial cities often benefited from being the hub of trade routes and supported a large population beyond their capacity.

In 1992, Lev R. Ginzburg considered incorporating additional mortality into the logistic equation without changing any other aspects of the environment. One of his proposals is: $dN/dt = rN(1 - N/K) - \mu N$, where $\mu N(\mu > 0)$ is the additional mortality term. As the resources do not change in this hypothetical scenario, intuitively the population with higher mortality might attain the same K, but only slower. However, the new population equilibrium is $K(r - \mu)/r(\mu < r)$, lower than the carrying capacity K. Once again, the Ginzburg paradox is due to the confusion between the concept of carrying capacity and population equilibrium. To illustrate, let us consider a thought experiment (Hui, 2006). There are three eggs in a six-position egg box (**Figure 2**). If we eat one and then put a new one in it everyday, the size of the egg population will be maintained at three. Now what is the carrying capacity of this egg population, three or six? Now, if we eat two eggs per day, the population equilibrium will become zero but the carrying capacity remains six. According to our knowledge, the above equation does not have any mechanistic problem but has been widely used as a theoretical foundation in spatial and metapopulation ecology. This new stable equilibrium is called the local carrying capacity of a metapopulation, whereas other scientists call this equilibrium local density. Once again, the concept of equilibrium was confused with carrying capacity. This confusion also appears in some textbooks, yet many others do differentiate between these two concepts. The carrying capacity of the egg box example should be the 'environment's maximal load,' i.e., the six positions in the egg box, and the equilibrium of egg population is three if we eat one and add one per day. If we can clarify these two concepts: carrying capacity and population equilibrium (Hui, 2006), there is no contradiction between intuition and there should be no Ginzburg's paradox.

Ecological carrying capacity is a measure of the amount of renewable resources in the environment in units of the number of organisms these resources can support. Normally, K is a function of both the species and the environment, and is expected to change only in evolutionary time (Daily and Ehrlich, 1992).

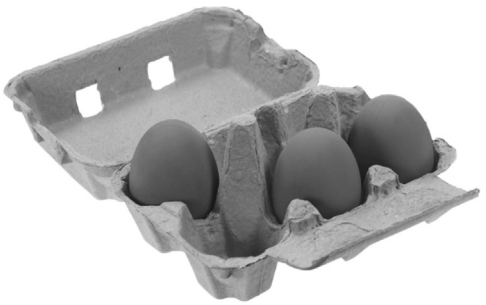

Figure 2 A population of three eggs at the equilibrium in an egg box with the six-holder carrying capacity.

Rees (1992) suggested that carrying capacity is the number of individuals of a given species that a given habitat can support without being permanently damaged. If the population of a given species exceeds the carrying capacity of a given habitat, then either the resources required to meet the needs of that species will be depleted, or the wastes produced by that species will build to the point of poisoning members of the species, or both, and the population will crash. In livestock management, carrying capacity is affected by many factors, such as habitat, climate, vegetation, soil profile, food quantity and quality and accessibility, inter- and intracompetition, foraging, parasitism, diseases, population density, social behavior, and anthropogenic disturbance. In terms of nutrition intake, carrying capacity has been defined as the accessible and utilizable energy in the habitat divided by per capita energy consumption (Ayllon et al., 2012). In terms of space, carrying capacity is also affected by aggregation or social group behavior, group size, dispersal distance, and habitat suitability, and can be defined as the amount of habitat available divided by the expected individual territory area for a given life stage (Ayllon et al., 2012). It is closely related to subsistence density, tolerance density, security density, maximum harvest density, minimum impact density in livestock and wildlife management (Hobbs and Hanley, 1990).

Carrying capacity is not a static number but is affected by the abundance and distribution of limited resources and by how individuals compete for these limiting resources (Ayllon et al., 2012). This notion is especially relevant in organisms that compete via both exploitation and interference because of behavioral responses, such as competition avoidance, induced by aggressive interactions typically result in a much reduced exploitation of the limited resource than could be accounted for by resource depletion alone (Ayllon et al., 2012). In territorial species, the behavioral adjustment of the size and shape of territories has profound consequences for their population regulation, demography, and spatial ecology. In practice, carrying capacity is determined by the resource in the least supply, 'the weakest link' as it were. For production purposes, accurate estimation of carrying capacity will affect the maximum sustainable yield, which is thought to be obtained when the population is maintained at half of the carrying capacity (Ayllon et al., 2012). By contrast, in conservation efforts, the maximum carrying capacity is the desired target.

Carrying Capacity of Ecosystems

The directionality of community succession is a powerful concept for conservation biology, analogous to the irreversibility of time in physics, and has revolutionized the understanding of complex adaptive systems. By definition, succession is 'an orderly process of community change' after disturbance. Knowing the directionality of succession is necessary for (1) distinguishing new from mature communities (i.e., defining the age of a community), (2) understanding how communities evolve and respond to disturbance (e.g., habitat loss and climate change), and (3) designing more efficient conservation and restoration plans. Two important concepts of an ecosystem are productivity (often measured by the total biomass) and biodiversity (often measured by species richness or the Shannon index that considers the relative abundance of species). The directionality of succession can potentially be indicated by the increase of net primary production. Per capita community productivity can be very high during the initial phases of succession, and decreases progressively as an upper limit to biomass is reached, and this phenomenon may be expressed as a sigmoid curve roughly akin to the logistic growth curve of a population. The upper limit of net primary production, or the maximum potential biomass, has been suggested to indicate the carrying capacity of biomass in a community (Monte-Luna et al., 2004).

If the number of species in a habitat is relatively low, immigration and diversification through disruptive selection will take place, which progresses until the region attains a maximum supportable load of species (Monte-Luna et al., 2004). This can also define the directionality of succession and, thus, the carrying capacity of biodiversity, depicting the maximum number of species or the 'biodiversity ceiling' that an environment can support. In fact, the form of species diversification on evolutionary timescales resembles the logistic population growth. By analogy, the number of species resembles the number of individuals or biomass, and the difference between rate of speciation and extinction the difference between fecundity and mortality (i.e., the intrinsic growth rate). The carrying capacity of biodiversity may be regulated by both physical factors such as climate and habitat heterogeneity and biotic interactions between species. Once the biodiversity reaches its ceiling, interspecific interactions can lead to competitive exclusion of some taxa, and adaptive niche partitioning (Monte-Luna et al., 2004). This can be further related to the r/K selection theory in ecology. It specifies a life-history tradeoff between the values of r and K that a species can possess. Species with high r and low K are opportunistic and often occur at the early stage of succession, whereas species with low r and high K are good competitors for limiting resources and often occur at the late stage of succession. An intermediate level of disturbance in a landscape, arguably, can create patches at different levels of succession, thus promoting the coexistence of these two types of species at the regional scale.

Human Carrying Capacity

Our planet is practically a closed system in terms of physical resources, and all human activity depends on these limited resources. Due to the continue growing of the world population and the rapid diminishing of pristine ecosystems from our exploration for consumption, the concept of carrying capacity is essential and should be considered paramount in our future planning. The Club of Rome has warned us the possible consequences of world population growth and industrialization for food production and resource exhaustion. They projected that within 50 years from now, the planet would reach its limit of growth (**Figure 3**), and thus advocated an urgent mutually beneficial integration of economics and ecology. Approaching carrying capacity means the deterioration of ecosystem services, the loss of biodiversity and habitat heterogeneity and the break down

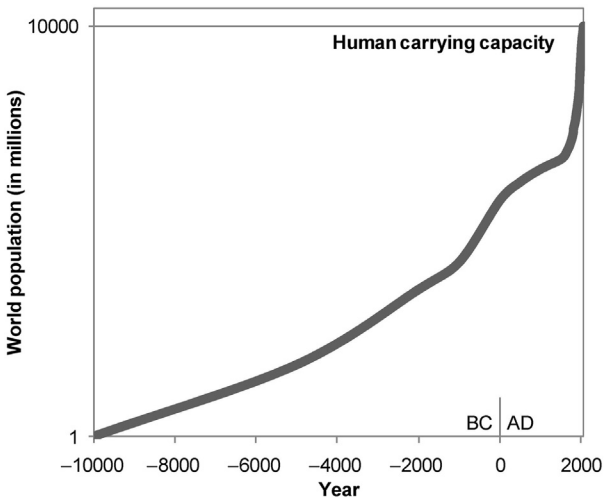

Figure 3 World human population from 10 000 BC to AD 2050 when the carrying capacity of 10 billion is projected to be reached.

of ecosystem resilience (Arrow et al., 1995). This could lead to irreversible change, enlarged uncertainty to the biosphere, and even a regime shift of the world's living and climate systems. This concern has greatly influenced the popularity of environmentalism and the creation of the United Nations Environment Program.

Carrying capacity is commonly assumed to be constant in population dynamics models used for resource management. Human carrying capacity, however, is not a fixed, static, or a simple relationship, but is dynamic and adaptive. Humans can have more than one *K* value for a given resource at a given time, and the carrying capacity can vary markedly with culture and level of economic development (Daily and Enrlick, 1992). Earth's human carrying capacity cannot be characterized by a simple monotonic decline of the natural capital as the number of humans increase. For instance, Cohen (1995) studied a dynamic *K* model, with the change of carrying capacity affected by the population change rate.

Human carrying capacity describes the number of human beings that can be supported on a sustainable basis in a given area (or on the whole Earth) within natural resource limits and by human choices concerning social, cultural, and economic conditions (Franck et al., 2011). This concept is twofold: Biophysical carrying capacity is the maximum population that can be supported by the resources of planet Earth at a given level of technology, whereas social carrying capacity is the sustainable biophysical carrying capacity within a given social organization (Franck et al., 2011). In the case of very efficient agriculture the carrying capacity is only determined by the ratio of the total productive area and the nonagricultural used area (per person). For 2005 land cover, it can support 11.4 billion (10^9) (Franck et al., 2011), ranging from 7.7 to 12 billion in literature (Cohen, 1995). It has been predicted to be much higher than this range under certain scenarios, ranging from 30 to 100 billion (Franck et al., 2011). It is the limiting resource that constrains human carrying capacity. Although there exists the technology and economic willingness to develop intensive sustainable agriculture and other forms of food supply compatible with the aims of world development programs.

In a Malthusian framework, economic growth generates population growth; the population growth increases pressure on the natural resource base; the deterioration of the natural resource base decreases per capita food output; and the reduced per capita food limits population growth which, in turn, limits exploitation of the environment (Berck et al., 2012). This moderating feedback loop (density dependence) can be diverged by technological improvements. Population pressure on the resource base stimulates technological improvements in food production and hence facilitates continued population growth, which intensifies exploitation of the resource base (Berck et al., 2012). Alternatively, as income goes up there is increasing environmental degradation up to a point, after which environmental quality improves, causing an inverted-U shape (Arrow et al., 1995). Moreover, the detrimental impact from human society to ecosystems is often discounted over time and distance in decision-making, i.e., discounting future and faraway activities than can reduce carrying capacity. This encourages behavior that may reduce carrying capacity for future generations (Daily and Enrlick, 1992). Globalization further creates the illusion that all regions could simultaneously sustain populations that sum to more than global *K*, a typical Netherlands fallacy.

Biocapacity

Since the concept of sustainable development was put forward, it has become an ideal development mode and a common policy goal. To date, many indicators have been developed to assess the status of sustainable development, such as the life-cycle assessment, human development index by the United Nations Development Programme, barometer of sustainability, index of sustainable economic welfare, environmental pressure indicator, genuine progress indicator, sustainable technology development, environmental sustainability index, and ecological footprint (Rees, 1992). The latter has gained popularity due to its compatibility with the data format commonly derived from economic and social surveys. The ecological footprint for a particular population is defined as the total area of productive land and water ecosystems required to produce sufficient resources and assimilate wastes (Rees and Wackernagel, 1994). Biocapacity can be defined as the locally available carrying capacity of the ecosystem for generating resources and absorbing wastes (Rees and Wackernagel, 1994) and is constrained by the carbon sequestration rate of the ecosystem. According to the global average rate of carbon sequestration, we need to have about 17 ha of land and ocean area to absorb 1 metric ton of carbon emission. Ecological footprint and biocapacity, thus, represent the demand on and the supply from a regional ecosystem, respectively. As both ecological footprint and biocapacity are measured in the same unit (the global hectare: gha), it is straightforward to calculate regional ecological budget as surplus and deficit, after discounting the influence of trading. This specifies whether a regional population is potentially self-sufficient or is at least partially reliant on imported biocapacity. An ecological surplus has been proposed as a minimum criterion for sustainability.

Due to it being highly operable and easy to understand for public and policy makers, with the data required accessible from government yearbooks, to date they have been applied at a variety of spatial scales, from municipality/provincial level to national/global extents, covering all aspects of socioeconomic sectors, such as industry, education, agriculture, tourism, and waste management.

As a continuously developing field, the ecological footprint methodology has been widely criticized and amended. For instance, it has been considered a static indicator of weak sustainability as no dynamics and bounds are imposed on the level of ecosystem services and their demands. This has been partially solved by time series analysis and extrapolation (Yue et al., 2013). The introduction of spatial features, with the help of the geographic information system (GIS), has largely released the methodology from this constraint. To calculate the biocapacity of a region, one first needs to estimate the available areas of biologically productive land and water. Specifically, this biologically productive area can normally be divided into six or more main categories (cropland, grazing land, fishing land, forest, built-up area, and barren land), and the sizes of these six land covers can then be either extracted from government agencies or calculated using remote sensing images with the aid of GIS (Yue et al., 2013):

$$K = \sum_i A_i \cdot Y_i \cdot E_i$$

where A_i is the biologically productive area of land cover category i; γ_i is the yield factor of land category i and is calculated annually as the ratio of the local yield of a generic product to the global average yield of the same product. The yield factor converts local biologically productive land into units of global average productivity and thus facilitates comparisons across regions; E_i represents the equivalence factor of land cover category i and is a scaling factor needed for converting a specific land use type into a universal unit of biologically productive area (gha). Equivalence factor is calculated each year as the ratio of the global average productivity of a specific land type to the average productivity of all biologically productive land on the earth.

An important issue that is associated with area-based information is the scale dependency of spatial features. Evidently, the shape and size of different land covers are sensitive to the spatial scale as most landscape features are scale dependent and have self-similar, fractal structures. Put simply, results for biocapacity can be influenced by the resolution of the map (Yue et al., 2013). Since area-based information has been widely implemented for estimating the sizes of different land covers and therefore the biocapacity, it is important to at least be aware of the scale dependency of biocapacity and its impact on regional sustainability assessment (Figure 4). Moreover, there is often a power–law relationship between the biocapacity and population density (Figure 5), suggesting that biocapacity serves not only as a support for social development and human well-being but also sets an ecological limit for human activities (Yue et al., 2011). The biocapacity pressure index is the regional ecofootprint to biocapacity ratio and reflects the pressure posed by the human population on the local ecosystem. If the biocapacity pressure index is greater than one, the population is in ecological overshoot, indicating that regional development is not ecologically self-sustainable. If the biocapacity pressure index is less than one, the region is potentially self-sufficient (Yue et al., 2011).

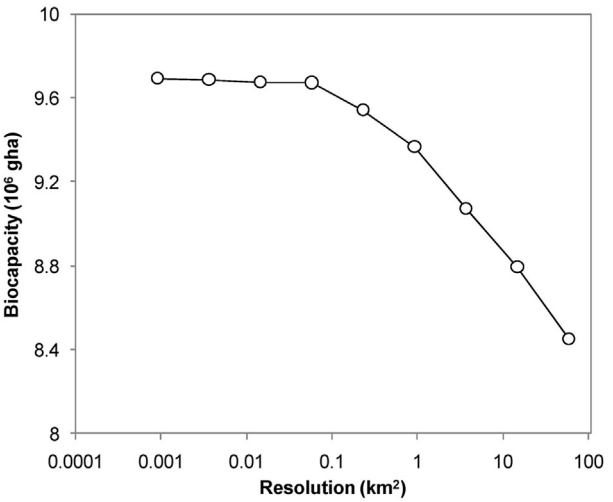

Figure 4 The scale dependency of the biocapacity in Jinghe River Watershed, Northwest China.

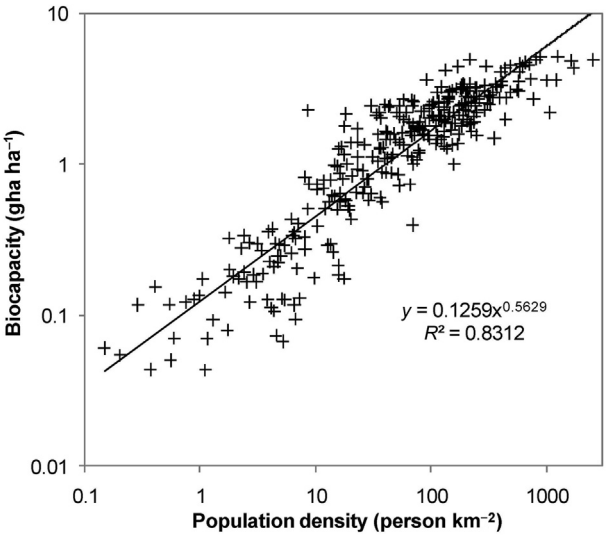

Figure 5 The relationship between population density and biocapacity per hectare for 320 counties in Northwest China.

Relevance in Anthropocene

The growing number of studies on carrying capacity and its further development into different research fields have made this concept the most important issue in this century, for obvious reasons. First, the world has entered a rather peaceful era after the World War II, allowing both the living condition to proliferate and the economy to boom at a staggering pace. Although it took the modern human population of *Homo*

sapiens 200 000 years to reach its first billion in the early nineteenth century, the second billion took only another 100 years. Standing at 7 billion currently, the human population is projected to reach 10 billion in the middle of this century. The sheer number and the increasing per capita consumption rate in this era of Anthropocene place an enormous pressure on the planet's resource supply, both renewable and unrenewable. Humanity as a single species consumes more than one-fifth of the planet's net primary production each year. Will this great era of modern humanity continue to grow indefinitely; fluctuate around a sustainable level after a soft landing; or follow the boom–bust pattern of many earlier civilizations? The concept of carrying capacity has become a great reminder, a type of 'sword of Damocles,' hanging over humanity and driving it onto a sustainable path.

Second, as human population continues to expand, the planet's pristine ecosystems are degrading at a rapid rate due to man-made global environmental changes – climate change, habitat loss, biological invasion, soil erosion, and pollution, to name a few. Ongoing agricultural intensification, overfishing, industrialization, and urbanization further transfer how semi-natural ecosystems function and behave, creating many novel ecosystems that face high disturbance and gene flow. The planet's greatest asset – biodiversity, 100 million species strong – is quickly losing its grip. The rapid loss of species is estimated to be between 1000 and 10 000 times higher than the natural extinction rate, reaching 10 000 species extinction per year many of which disappear before they have even been formally described. Human society relies heavily on the health of ecosystem functions, the provision of ecosystem services and the benefit from biological diversity for cleaning air and water, stabilizing weather, maintaining soil fertility, dissipating waste, controlling pests, pollinating crops, generating power and discovering new antibodies, and providing food, timber, cloth, medicine, industrial material (coal, oil, gas, rubber, plastics, chemicals), together with minerals. We are not making anything new but utilizing products and services of the planet's ecosystem and physical system. To sustain humanity, we need to manage the planet's biosphere within its bearable margin to avoid disruptive regime shift and massive extinction, although some argue that we have already crossed the point of no return. On the positive side, the concept of carrying capacity is a powerful tool for assessing and conserving natural capital.

See also: Environment and Development; Environmental Movements; Food Security and 'Green Revolution'; Limits to Growth; Sustainable Development: An Economic Perspective.

Bibliography

Arrow, K., Bolin, B., Constanza, R., Dasgupta, P., Folke, C., Holling, C.S., Jansson, B.O., Levin, S., Maler, K.G., Perrings, C., Pimentel, D., 1995. Economic growth, carrying capacity, and the environment. Science 286, 520–521.

Ayllon, D., Almodovar, A., Nicola, G.G., Parra, I., Elvira, B., 2012. Modelling carrying capacity dynamics for the conservation and management of territorial salmonids. Fisheries Research 134–136, 95–103.

Berck, P., Levy, A., Chowdhury, K., 2012. An analysis of the world's environment and population dynamics with varying carrying capacity, concerns and scepticism. Ecological Economics 73, 103–112.

Catton, W., 1986. Carrying Capacity and the Limits to Freedom. Paper Prepared for Social Ecology Session 1. XI World Congress of Sociology, New Delhi, India.

Cohen, J.E., 1995. Population growth and Earth's human carrying capacity. Science 269, 341–346.

Daily, G.C., Ehrlich, P.R., 1992. Population, Sustainability, and Earth's Carrying Capacity. Paper No. 0046. Department of Biological Sciences, Stanford University, Stanford.

Franck, S., von Bloh, W., Muller, C., Bondeau, A., Sakschewski, B., 2011. Harvesting the Sun: new estimations of the maximum population of planet Earth. Ecological Modelling 222, 2019–2026.

Hobbs, N.T., Hanley, T.A., 1990. Habitat evaluation – do use availability data reflect carrying capacity. Journal of Wildlife Management 54, 515–522.

Hui, C., 2006. Carrying capacity, population equilibrium, and environment's maximal load. Ecological Modelling 192, 317–320.

Monte-Luna, P.D., Brook, B.W., Zetina-Rejon, M.J., Cruz-Escalona, V.H., 2004. The carrying capacity of ecosystems. Global Ecology and Biogeography 13, 485–495.

Rees, W.E., 1992. Ecological footprint and appropriated carrying capacity: what urban economics leaves out. Environmental Urbanization 4, 121–130.

Rees, W.E., Wackernagel, M., 1994. Ecological footprints and appropriated carrying capacity: measuring the natural capital requirements of the human economy. In: Jansson, A.M., Hammer, M., Folke, C., Constanza, R. (Eds.), Investing in Natural Capital: The Ecological Economics Approach to Sustainability. Island Press, Washington, DC, pp. 362–390.

Seidl, I., Tisdell, C.A., 1999. Carrying capacity reconsidered: from Mathus' population theory to cultural carrying capacity. Ecological Economics 31, 395–408.

Yue, D., Xu, X., Hui, C., Xiong, Y., Han, X., Ma, J., 2011. Biocapacity supply and demand in Northwestern China: a spatial appraisal. Ecological Economics 70, 988–994.

Yue, D., Guo, J., Hui, C., 2013. Scale dependency of biocapacity and the fallacy of unsustainable development. Journal of Environment Management 126, 13–19.

Relevant Websites

http://www.carryingcapacity.com.au/
http://www.carryingcapacity.org/
http://www.facingthefuture.org/
http://www.sustainablescale.org/

Cartographic Visualization

Menno-Jan Kraak, University of Twente, Enschede, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The cartographic visualization process, based on the statement "How do I say what to whom, and is it effective?" is explained for both presenting and exploring spatiotemporal data. The visualization environment and required functionality are described.

Setting

Assume that one is interested in an overview of the population distribution of a particular area. An option would be to give a written description such as: 'in the northeast only a few small towns are found, while in the southeast …' Before one gets a clear picture, several pages could have been filled, especially if a description of the relations of the towns with other geographic phenomena, such as the rivers or railroad network, is also incorporated. A better option is to tell this story using a map. Maps are the most efficient and effective means to transfer geospatial information. The map user can locate geographic objects, while the shape and color of signs and symbols representing the objects inform about their characteristics, such as the town's location and population size. Maps reveal geospatial relations and patterns and offer the user insight into, and an overview of, the distribution of particular phenomena, such as an area's population distribution.

The map easily allows one to answer questions of the nature 'Where …?' 'What …?' and 'When …?' These questions deal with the basic components of geospatial data: location, characteristics, and time, or their combination. The map discussed herein and found in Figure 1 will quickly result in answers to questions such as 'Where are the large towns located?' or 'What town has the highest number of inhabitants?' Questions regarding 'When?' depend on the map contents. Often the map is just a snapshot in time; for instance, the map in Figure 1 depicts the status as of 1998. In answering questions, another quality of maps will be revealed – the ability to offer an abstraction of reality. It simplifies by selection, but at the same time it puts, when well designed, the remaining information in a clear perspective. The map of the area only needs the boundaries of municipalities and a symbol for the number of people living in each town. In this particular case there is no need for roads, mountains, or other physical features. Maps like these are often static and come as they are, especially when they are printed on paper. On-screen maps often allow for interaction, such as getting data from the database behind the map.

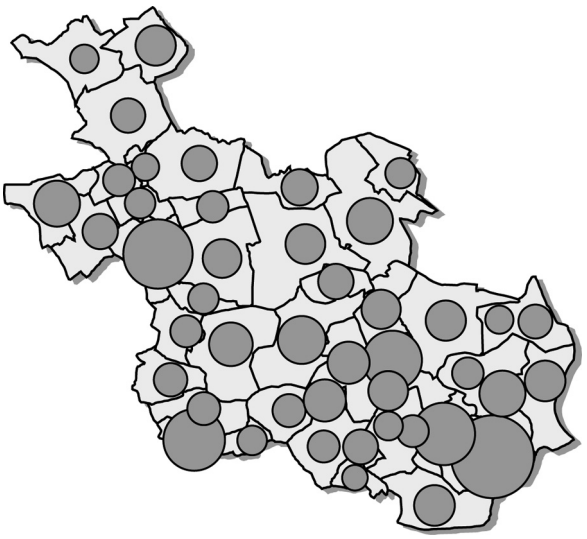

Figure 1 Population distribution.

For more complex questions one needs more than just a static map. For instance, to study the impact of industrial development on the growth of towns, one would need the possibility to generate maps depending on the nature of the question. This could be to switch layers with information on or off. This would also include the possibility to view the data in alternative visualizations. Questions are no longer the simple 'Where?' 'What?' and 'When?' especially if, for instance, specific demographic models are included that can predict alternative future population trends based on census and socioeconomic data. These trends could be visualized in an animation. The mapping environment requires interaction and dynamics. This is what today's visualization is all about, as will become clear in the remainder of this article. However, according to the dictionary, visualization means 'to make visible.' From this perspective map making has always been visualization. Here the term visualization is used in the context that allows the map user to interact with the (digital on-screen) map (for a detailed discussion see Hearnshaw and Unwin, 1994; MacEachren and Taylor, 1994).

Visualization Process

Any map, static or dynamic, on screen or on paper, complex or simple, is created during what is called the cartographic visualization process. This process is considered to be the translation or conversion of geospatial data from a database into graphics. Predominantly these are maplike products as is schematically explained in Figure 2. This process should be seen in the context of geospatial data handling, which is the acquisition, storage, manipulation, visualization, and use of geospatial data in the context of particular applications.

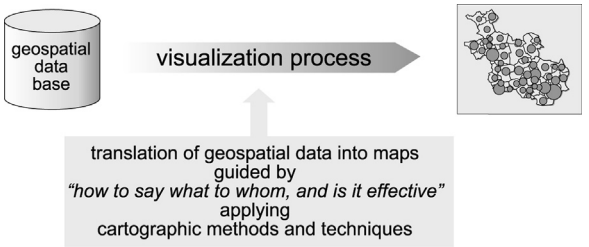

Figure 2 The cartographic visualization process.

During the visualization process, cartographic methods and techniques are applied. These can be considered as a kind of grammar that allows for the optimal design, production, and use of maps, depending on the application. The process is guided by the phrase 'How do I say what to whom, and is it effective?' The phrase holds four key words: *How* refers to cartographic methods and techniques (in the case of **Figure 1** a proportional point symbol map has been chosen); *What* refers to the geospatial data (**Figure 1** deals with quantitative population data); *Whom* refers to the map audience and the purpose of the map – (the map in **Figure 1** is rather basic and could function in a newspaper or school atlas); and *Effective* reflects the usefulness of the map (do the map readers understand the message the map intends to bring, an overview of the area's population distribution).

The producer of maps could be a professional cartographer, but could also be an expert who is mapping, for instance, vegetation stands using remote sensing images or health statistics in the slums of a city. With today's availability of Web mapping tools, the mapmaker could be anyone – with or without any notion of cartographic design.

The visualization process can vary greatly depending on where the visualization takes place and the purpose for which it is needed. Visualizations can be, and are, created during any phase of the geospatial data-handling process. They can be simple or complex, while the production time can be short or long. Some examples are the creation of a full, traditional topographic map sheet, a newspaper map, a sketch map showing a route, a map from an electronic atlas, an animation showing the growth of a city, a three-dimensional (3D) view of a building or a mountain, or a real-time display of traffic conditions on a mobile device. Other examples include 'quick and dirty' views of part of the database, the map used during the updating process, or during a geospatial analysis. The environment in which the visualization process is executed can vary considerably. It can be done on paper, on a stand-alone personal computer, or on a networked tablet computer.

Many tools are available to visualize the data. These tools consist of functions, rules, and habits. Algorithms to classify the data or smooth a coastline are samples of functions. Rules tell us, for instance, to use proportional symbols to display quantities or position an artificial light source in the northwest to create shaded relief maps. Habits, or conventions as some would call them, tell us to color the sea in blue, lowlands in green, and mountains in brown (Robinson et al., 1995; Kraak and Ormeling, 2011).

Visualization and Cartography

In the past, even dealing with incomplete and uncertain data, the visualization process nearly always resulted in an authoritative map. The maps created by a cartographer were good enough for the user. This shows that cartography, for a long time, has been very much driven by supply rather than demand. Somehow, this is still the case. However, nowadays, it is also accepted that just making a map is not the only purpose of cartography. Especially since 1980, many people have become involved in making maps. The widespread use of Geographical Information Systems (GIS) has significantly increased the number of maps being created (Longley et al., 2010). Even the spreadsheets used by most office workers today have mapping capabilities, although many people are probably not aware of this. The opportunities offered by the World Wide Web lead to an incredible increase in maps produced. Some websites, such as Google Maps, Bing Maps, or Open Street Map produce over a million maps a day! Many of these maps are not produced as final products, but rather as intermediate products to support the user in his or her work dealing with geospatial data. The map, as such, has started to play a completely new role: it is not just a communication tool but also a tool to aid the user's (visual) thinking process.

This process is being accelerated by the opportunities offered by hardware and software developments. These have changed the scientific and societal needs for geo-referenced data and, as such, for maps. New media not only allow for dynamic presentation but also for user interaction. Users do expect immediate and real-time access to the data, and geospatial data has become abundant. In what is known as a Web2.0 environment, people collect and map their own location-based and annotated data. They do so using mobile phones and make their maps available via social media such as Facebook or Twitter. These massive amounts of data can also be mined for other purposes.

These developments have given the word *visualization* enhanced meaning since progress in other disciplines has linked the word to more specific ways in which modern computer technology can facilitate the process of 'making visible' in real time. Specific software toolboxes have been developed, whose functionality is based on two key words: *interaction* and *dynamics*. Separate disciplines called *scientific visualization* (McCormick et al., 1987) and *information visualization* (Ware, 2012) have developed and have had a major impact on cartography. If applied in cartography, dynamic interaction offers the user the possibility of instantaneously changing the appearance of the map. Interacting with the map will stimulate the user's thinking and add a new function to the map. As well as communication, it will prompt thinking and decision making.

These developments have stimulated a model for map-based visualization (DiBiase, 1990). As such, it is also known as 'geovisualization' (MacEachren, 1995; Dykes et al., 2005), which covers both the communication and thinking functions of the map. Communication is described as 'public visual communication' since it concerns maps aimed at a wide audience. Thinking is defined as 'private visual thinking' because it is often an individual playing with the geospatial

data to determine its significance (see Figure 3). On a lower level, different visualization stages can be recognized; each stage requires a different strategy from the perspective of map use, based on audience, data relations, and the need for interaction. These stages are exploration, analysis, synthesis, and presentation.

From Figure 3 it is obvious that presentation fits into the traditional realm of cartography, where the cartographer works on known geospatial data and creates communicative maps. These maps are often created for multiple uses. However, exploration often involves a discipline expert creating maps while dealing with unknown data. These maps are generally single purpose and are related to the expert's attempt to solve a problem. While dealing with the data, the expert should be able to rely on cartographic expertise provided by the software or some other means.

This process describes the 'democratization of cartography' (Morrison, 1997). Morrison explains that 'using electronic technology [means that] no longer does the map user depend on what the cartographer decides to put on a map. Today the user is the cartographer'. And 'users are now able to produce analyses and visualizations at will to any accuracy standard that satisfies them (Morrison, 1997).'

Visualization and Exploration

What is exploratory cartography? The environment has just been described: a person is trying to solve a particular geoproblem and is exploring various geospatial databases. Exploration also means working with unknown data. However, what is unknown for one is not necessarily unknown to others. For instance, browsing Google Earth is an exploration for most of us because of its wealth of information. With products like these, such exploration takes place within boundaries set by the producers. Cartographic knowledge is incorporated in program wizards resulting in predesigned maps. Some users feel this to be a constraint, but those same users will no longer feel constrained as soon as they follow the Web links attached. This example shows that the environment, the data, and the type of users influence one's view of what exploration entails.

Returning to the sentence driving the visualization process, 'How do I say what to whom and is it effective?' – some similarities as well as differences exist between modes of visualization that entail presentation and those that involve exploration. 'How?' still represents the cartographic methods and techniques. However, new technology is emerging that offers challenges and opportunities, such as animation, the application of the third dimension and virtual reality, multimedia, and so forth. 'I' is no longer just the cartographer, but an expert geoscientist. In the very near future, it can probably be just anyone having access to the World Wide Web. 'What?' no longer represents a relatively well-defined and known data set – at least, certainly not from the user's perspective. 'To whom' seems to be simpler than before; it is not a relatively well-defined user group but the same person represented by 'I,' the expert geoscientist in the role of cartographer. 'Effective' raises some interesting questions. When a map is used, the information to be transferred is known and, for all problems involved, can somehow be measured. But how can the visual thinking process be measured? If it is considered positively, is it because of the efficient graphics or because of the geoscientist's clever thinking? These questions become more complex if we realize that we do not even know the initial aim of the visualization in these circumstances.

The most prominent change is the shift from supply-driven cartography to a demand-driven approach. Although having many people making maps without any cartographic knowledge might seem wrong, those noncartographers will also introduce fresh views, for instance as described by Keller and Keller (1993). They distinguish three steps in the visualization process: the first identifies the visualization goal; the second removes mental roadblocks; and the third designs the display in detail. In the second step, Keller and Keller suggest taking some distance from the discipline in order to reduce the effects of traditional constraints. Why not choose an alternative mapping method? For instance, show a video of the landscape next to a topographic map. New, fresh, creative graphics could

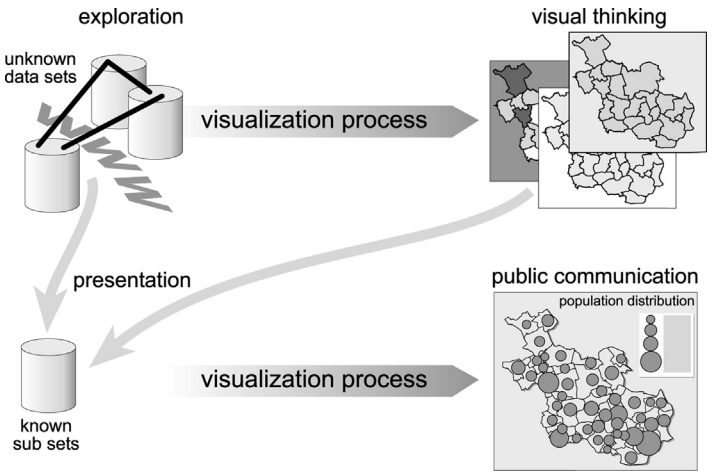

Figure 3 The visualization process: visual thinking and visual communication.

be the result; such graphics might also offer different insights and would probably have more impact than traditional mapping methods. During the third step, which is especially applicable in an exploratory environment, one has to decide between mapping data and visualizing phenomena.

An exploratory visualization environment offers the tools to act in the ways previously suggested. Such an environment should allow the user to look at geospatial and other geo-referenced data in any combination, at any scale, with the aim of seeing or finding geospatial patterns (which may be hidden). Geospatial patterns can be defined as variations in location, attributes or time, or a combination of any of the three geospatial components within an area of interest. One of the first concepts of visual geospatial data exploration was introduced by Monmonier (1989) when he described the term *brushing* (see Figure 4). This is when the selection of an object in a map automatically highlights the corresponding elements in the other graphics. Depending on the view in which one selects the objects, there is geographical brushing (clicking in the map), attribute brushing (clicking in the diagram or table), and temporal brushing (clicking on the time line). As such, the user gets an overview of the relation among geographic objects based on location, characteristics, and time.

Visualization Functions

What are the basic requirements of an exploratory visualization environment? The necessary functions described in the following sections need high interactivity options.

Basic Display

Map displays need tools to allow the user to pan, zoom, scale, transform, and rotate the image contents. These geometric tools should be available and independent of the dimensionality of the displayed geospatial data. Figure 5(a) illustrates panning and zooming facilities. Another option could be rotating the map. This last function is of particular need when displaying 3D maps, since it also allows objects that otherwise might be hidden by other objects to be seen.

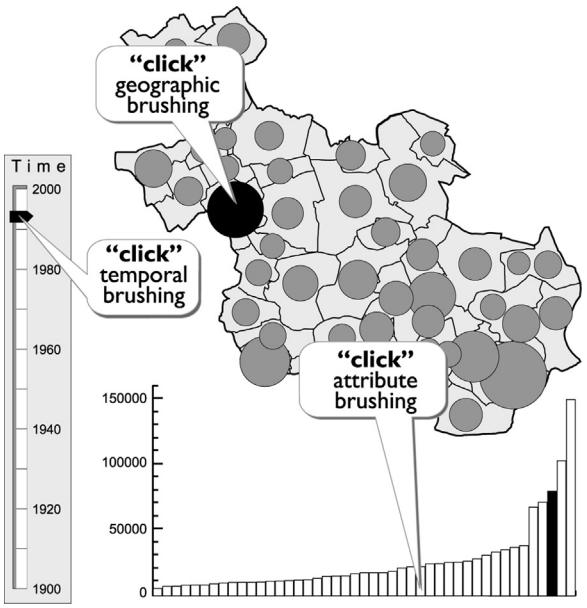

Figure 4 Brushing.

Navigation and Orientation

This involves the keys to the map. At any time, the user should be able to know where the view is located and what the symbols mean. To illustrate this function, Figure 5(b) shows a map with its marginal information as well as the coordinates at the cursor location in the map.

Query Data

During any phase of the visualization process, the user should have access to the geospatial database to query the data. The questions should not necessarily be limited to a simple 'What?' 'Where?' or 'When?' As Figure 5(c) shows, clicking a geographic object in the map reveals the information available in the database, as well as the hyperlinks that are attached to the object.

Multi-Scale

Combining different data sets is a common operation in an exploratory environment. The chance that these sets will have the same data density or the same level of abstraction is unlikely. Generalization operators to solve these multi-scale problems remain necessary, if not just to make sure zoom-in and zoom-out operators result in sensible maps. In Figure 5(d) zooming-in results in a more detailed map.

Reexpression

To stimulate visual thinking, an unorthodox approach to visualization was recommended. This requires options to manipulate data behind the map or offer different mapping methods for displaying the data. An example of data manipulation is the application of several classification systems, while the use of different advanced map types to view the same data represents display differences. Figure 5(e) illustrates how the population data shown in the figure's main map can be displayed alternatively. The upper inset shows a prism map (a 3D representation of the data in which the height of the area corresponds to the attribute value; the big towns have high columns) and a cartogram (a map in which the area of a geographic unit is equal to the attribute value and not, as usually, to its geographic area; the highly populated towns take a lot of map space).

Multiple Dynamically Linked Views

These tools represent a combination of multimedia and the brushing technique already mentioned. The user will be able to view and interact with the data in different windows, all representing related aspects of the data. These views do not necessarily contain maps; video, sound, text, and so forth can all be included. Clicking an object in a particular view will

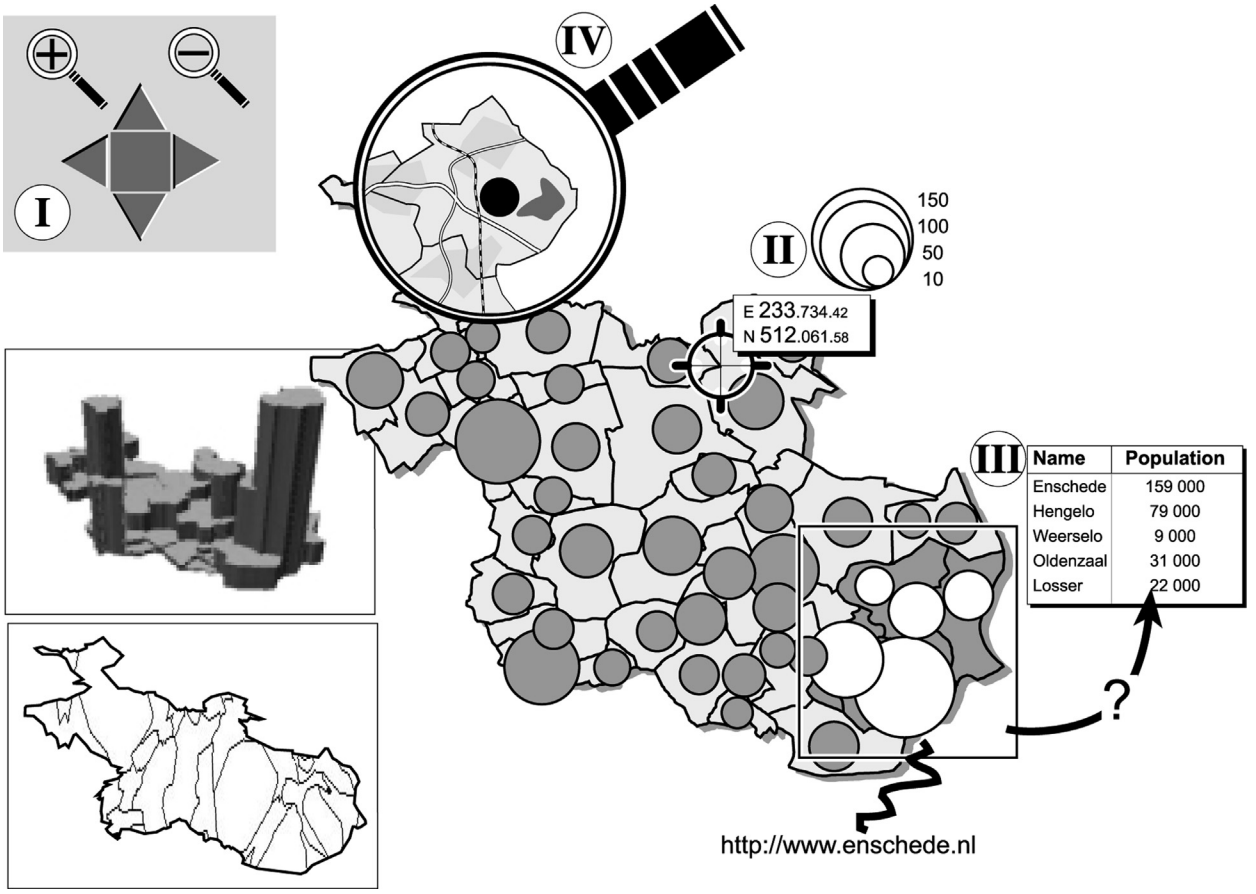

Figure 5 Exploratory cartographic functions: (a) basic display; (b) navigation and orientation; (c) query; (d) multi-scale; and (e) reexpression.

show its geospatial relations to other objects or representations in all the other views (see Figure 6).

Animation

Maps often represent complex processes that can be well expressed by animation. Animation can be used for temporal as well as nontemporal changes in geospatial data. Aspects to be solved are related to the interface (user interaction – navigation) and the legend. Since the exploratory user will also be the creator of the animation, he or she should be able to influence the flow of the animation.

Visualization Environments

Most cartographic components of geographical information systems can only handle one or two of the functions described in the previous section. Scientific and information visualization software might have the most functions available, but it lacks the typical geo-referencing options needed to deal with geospatial data, options that are found in the typical Geographic Information Systems software. However, operational experimental exploratory visualization environments do exist in academic environments. An example of an early experimental environment is the Cartographic Data Visualizer (CDV) (Dykes, 1997). The program requires a base map (administrative boundaries) and statistical data. From the menu on the left the user can select active variables and visualizations. Options are choropleth (or polygon) maps, proportional circle maps, cartograms, and noncartographic visualization such as dotplots and scatter diagrams. When creating these different views on the data, they will all be linked together. Clicking an area in a map or a symbol in a diagram will highlight the corresponding area/symbol in the other views. The user has several options available to choose the layout of each of the views. One can also execute basic calculations to create new variables or classify the data according to different methods. CDV is, by the nature of its functionality, capable of showing geospatial patterns that would not be apparent when single maps or diagrams are viewed.

Conclusions

Packages with the functionality similar to CDV offer an advanced cartographic visualization environment. However, remembering 'How do I say what to whom, and is it effective?' leaves the question open if it indeed works. From a technical point of view it certainly does, but can the user handle all those linked views? It is argued that if users can change their perspective on the data through selection and transformation, as well as alternative visualizations, meaningful relations among data variables are more likely to be revealed. For some

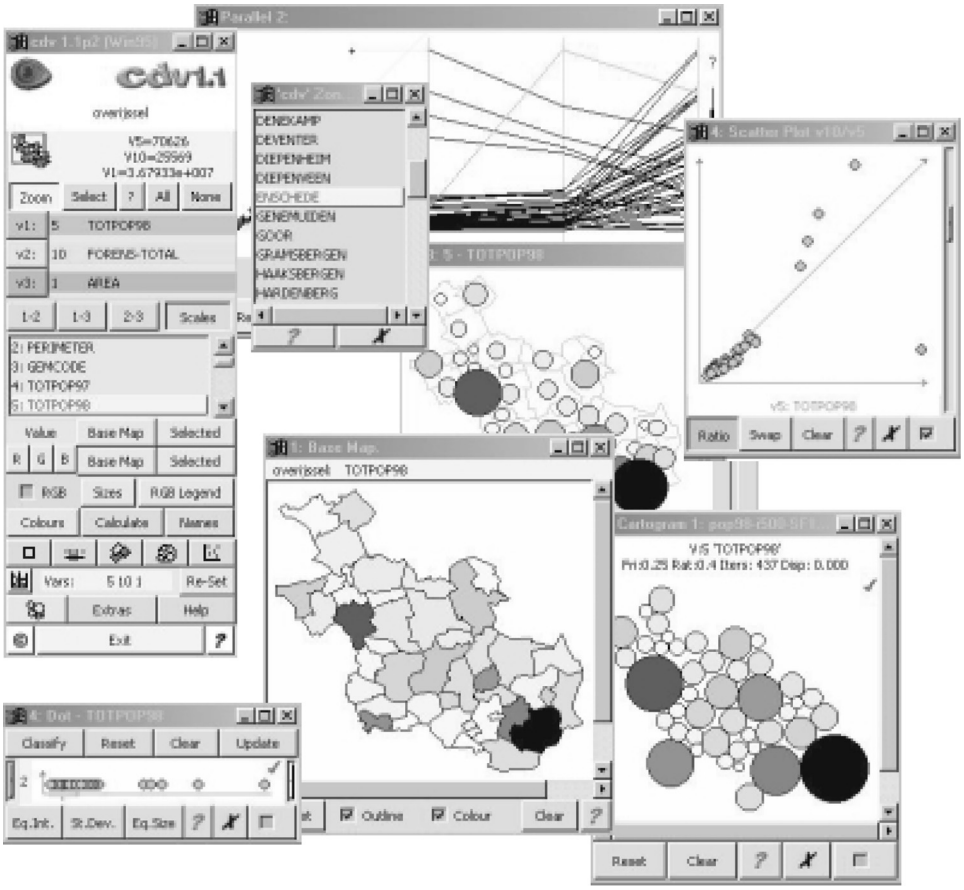

Figure 6 A typical view of a CDV session.

products, usability is easy to determine; for others it is quite a job. To understand how maps work is not easy (Wood and Fels, 1992), especially as we change into a demand-driven mapping environment. In the supply-driven environment, it was known what the map should tell and who the customers were. Today cartographers deal partly with providing facilitating tools to visualize geospatial data. How the tools are used remains, for the moment at least, unknown. These trends will only increase as more and more maps are produced via the World Wide Web by both cartographers and mapmakers. One has to find out why someone wants to make a map in order to be able to judge the future tools that need to be developed.

See also: Cartography; Cognitive Maps; Dynamic Mapping in Geography; Ethnic Conflict, Geography of; Geographic Information Systems; Tactile Maps in Geography; Thematic Maps in Geography.

Bibliography

DiBiase, D., 1990. Visualization in earth sciences. Earth & Mineral Sciences, Bulletin of the College of Earth and Mineral Sciences 59, 13–18.

Dykes, J., 1997. Exploring spatial data representation with dynamic graphics. Computers and Geosciences 23, 345–370.

Dykes, J., MacEachren, A.M., Kraak, M.J. (Eds.), 2005. Exploring Geovisualization. Elsevier, Amsterdam.

Hearnshaw, H.M., Unwin, D.J. (Eds.), 1994. Visualization in Geographical Information System. Wiley, Chichester, UK.

Keller, P.R., Keller, M.M., 1993. Visual Cues, Practical Data Visualization. IEEE Press, Piscataway, NJ.

Kraak, M.J., Ormeling, F.J., 2011. Cartography Visualization of Spatial Data. Guilford Press, New York.

Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W., 2010. Geographic Information Systems and Science. Wiley, Chichester.

MacEachren, A.M., 1995. How Maps Work: Representation, Visualization and Design. Guilford Press, New York.

MacEachren, A.M., Taylor, D.R.F. (Eds.), 1994. Visualization in Modern Cartography. Pergamon Press, London.

McCormick, B.H., DeFanti, T.A., Brown, M.D., 1987. Visualization in scientific computing. IEEE Computer Graphics 7, 69.

Monmonier, M., 1989. Geographic brushing: enhancing exploratory analysis of the scatterplot matrix. Geographical Analysis 21, 81–84.

Morrison, J.L., 1997. Topographic mapping for the twenty first century. In: Rhind, D. (Ed.), Framework of the World. Geoinformation International, Cambridge, UK, pp. 14–27.

Robinson, A.H., Morrison, J.L., Muehrcke, P.C., Kimerling, A.J., Guptill, S.C., 1995. Elements of Cartography. Wiley, New York.

Ware, C., 2012. Information Visualization: Perception for Design. Morgan Kaufmann Publishers.

Wood, D., Fels, J., 1992. The Power of Maps. Routledge, London.

Cartography

Judy M Olson, Michigan State University, East Lansing, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cartography is "the discipline dealing with the conception, production, dissemination and study of maps" (International Cartographic Association, 1996). This article covers the basic concepts of cartography as well as the role of maps in society.

Introduction

Cartography is "the discipline dealing with the conception, production, dissemination and study of maps" (International Cartographic Association, 1996). It is one of several mapping sciences, including geodesy, surveying, aerial photogrammetry, and satellite remote sensing. It is closely allied with geographic information systems, geographic information science, and geomatics; the term cartography now coexists with many other terms that have developed as modern electronics have come into use in the production, dissemination, and study of maps. Responding to changing social, intellectual, and technological innovations, cartography has risen from its roots in gestures and marks on the ground to a highly sophisticated and varied endeavor that uses data from aerial photographs, satellite images, global positioning systems, crowd sourcing, and the harvesting of social media. Traditional concepts in cartography, including scale, projection, spatial relationship, generalization, and symbolization and data modeling, remain central. Critiques of cartography have presented views about how maps function beyond their role as scientific documents, focusing attention on the underlying culture and motivations for mapping and their socially constructed nature (Harley, 1989). In the last half of the twentieth century, the concept of maps as communication devices brought the user to the center of concern for the mapmaker (MacEachren, 1995). That development distinguished the period from a tradition more limited to a focus on visual representation, i.e., the logical encoding of data.

Traditionally, maps have been classified as general reference (showing a variety of individual features), thematic (showing a distribution), or somewhere between these ends of the spectrum. Even though the classification is inexact and the explosion of mapmaking using new tools and data sources has further taxed the system, the terms are still used. They are often used now to distinguish different features on a map: A population map has thematic data (population), but it probably also has reference data such as political boundaries.

Cartography as an academic discipline has most often been found in geography departments in colleges and universities in the United States and Canada. It has commonly been an independent department in Europe and other parts of the world. Labels for some of those departments have changed in recent decades. Commercial enterprises, nonprofit entities, and government agencies that produce maps generally have large cartography departments that are central to their operations, though they, too, may bear some other label.

The term 'research' within the field of cartography has at least two distinct meanings. One is the systematic gathering of information from a variety of sources for compilation into a coherent map. The other is the quest for discovery of knowledge about maps, the processes associated with them, and the underlying meaning of maps within the culture in which they exist. The former definition is closely associated with government and commercial map production and the latter with academic cartography. The subject matter of academic cartographic research ranges from historical studies of the cultural, technical, and political context of maps to the changing processes of production to the ways in which maps are used in wayfinding and in the production of knowledge.

The relationship between people and maps, including symbol perception, the development of map-reading abilities, and cognitive processes in map use, was of special interest in cartography in the late twentieth century, as were the processes by which maps could be produced by computer. Maps delivered electronically and on demand have had a profound effect on how maps are obtained and used, and modern tools have made possible a wide range of innovative and dynamic products. The common paper street and road maps of previous decades, for example, have largely given way to maps on the screens of computers, global positioning system devices, and smartphones. The choice of scale for these displays affects what and how features are symbolized (and vice versa) and the ability to see less information at a time, but far more information as the user scrolls seamlessly over space. Users can also choose which features to display (e.g., restaurants, parks, or hotels) and whether to turn on voice directions when driving or walking or otherwise navigating from place to place.

Although maps are most closely associated with geography, practitioners in many disciplines and professions use the map as a device for recording and preserving information, as a tool in research, and as pedagogical illustration. Given the use of computers in mapping, their production is no longer strictly the domain of professional cartographers, or even others in the mapping sciences, as it once was. Anyone with a computer and appropriate software can produce maps, albeit not always the ones they would like to see because data may not be available or the mapping software with which the person is familiar may not produce the desired product. This democratizing of mapping has had both negative and positive effects. The products of nonspecialists range from misleading to highly insightful and creative. On the other hand, some types of maps are now far more widespread and are available in a more timely

way to fit immediate needs than in the past, and they are becoming less constrained by convention. Even more important, highly motivated mapmakers from many disciplines and often with high levels of technical expertise are producing maps that are truly innovative (e.g., Weldon Cooper Center for Public Service, 2013).

Cartography and its products have a profound effect on human thinking and behavior. The map is a metaphor for the real world and is often the model that shapes it. Existing hills and valleys, rivers, and political boundaries are recorded on maps, but decisions about surface excavation, river redirection, or political redefinitions are marked on maps and become reality. The terms 'map' and 'mapping' have become common metaphors in everyday language; 'on the map' signifies importance, and 'mapping a strategy' implies attentiveness, organization, and purpose in the planning of actions.

Cartography has responded to changing social, intellectual, and technological conditions throughout history, and the permeation of computers into all aspects of life in the late twentieth century brought on profound changes that go well beyond who can make a map. Cartography transformed from a data-poor to a data-rich field, as information is now gathered with aerial and satellite imaging, global positioning systems, and computerized field systems. In just 11 days in February 2000, the Shuttle Radar Topography Mission collected high-resolution elevation data for nearly the entire globe (Jet Propulsion Laboratory, 2009). Inexpensive computer storage and the Internet have made it possible to share and receive massive amounts of data. Professional cartographers are now heavily involved in information management and in creating mechanisms for shared use of data. Scientific visualization in cartography, which implies the use of maps for discovery of knowledge and generally involves user interaction with the map, has gained considerable attention in recent years as a result of the ubiquity of computer usage in spatial representation.

Major Cartographic Concepts

Mapping the Earth's surface involves the concepts of scale, resolution, projection, spatial relationships, generalization, symbolization and data modeling, and categories of maps. Some level of understanding of each of these concepts is inherent in every instance of the employment of a map.

Scale

The map is a scaled model of the Earth's surface (or of the surface of some other planetary body), and the reduced size is one of its most useful features. The Earth itself is too large for ready comprehension.

There are three types of scale expression commonly used on conventional maps: verbal, representative fraction, and graphic (**Figure 1**) (Robinson et al., 1995). Although each can refer to either distance (linear scale) or area (areal scale), linear scale is by far the more commonly used. Areal scale is generally indicated only if area is consistently scaled over the entire map. Since distance cannot be consistently scaled, no such constraint is associated with the appearance of a linear scale indicator.

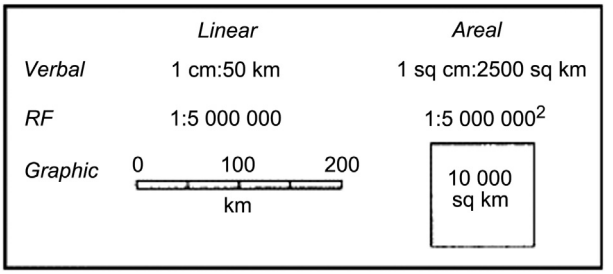

Figure 1 The three types of scale expression, illustrated in both linear and the rarely used areal form. (RF = representative fraction.)

The terms large and small, or the comparative terms larger and smaller, are often used in connection with scale. Larger scale refers to a map that is a larger representation than a smaller-scale map of the same area. A map at 1:100 000 of a given area is larger than a map at 1:200 000, and, accordingly, 1:100 000 is a larger scale than 1:200 000.

Explicit scale expression is not the only indication of scale on maps (Eastman, 1981; Brewer, 1990). Map content and design give scale clues as well. Double-line streets, for example, immediately convey that the map is large for the size of area represented and is therefore a relatively large-scale map; a shaded relief background and sparse line work almost always indicate that the map is small relative to the area represented and is therefore a relatively small-scale map. Scale clues are becoming increasingly important because the user of an animated map often cannot use an explicit expression of scale while observing the changing map. Because users seldom measure distances or areas on thematic maps, especially those of familiar areas displayed on the Web, it is becoming increasingly common to omit scale expressions on them.

Resolution

Resolution is related to scale. For images captured electronically (say, by satellite) and for maps displayed electronically, resolution is the distance on the Earth's surface that is covered by each pixel. It is an indication of how much detail has been captured and how much detail can be represented in a visual image made from the data. For example, a satellite image with a resolution measured in kilometers cannot capture individual buildings, whereas another image with a resolution measured in centimeters could capture not only buildings but much smaller features. The resolution of the image does not dictate the scale of display, although an image with coarse resolution will not usefully be displayed at large scale and the image with fine resolution is inefficient for small-scale display. If many kilometers are represented by each pixel, the map could not show sufficient detail to navigate in a city, whereas another with a resolution in centimeters would likely contain too much detail for that purpose. Detail on maps, however, is also a function of design choice, whether by the mapmaker or the user. A map of population by county is of much finer data resolution than one showing the data by state, even if the two maps are shown at exactly the same display scale. A map showing only major roads and their names might also be capable of showing more minor roads, or additional

information of some other kind, if the person using the map display chooses to turn on the additional features. There is a lower limit scale at which data at a given resolution can be represented, and there is an upper limit scale at which the map is noticeably wasteful of space and inefficient for conveying information (Tobler, 1988).

Projection

The term 'projection' refers to the system by which points on a spherical or spheroidal surface are assigned to points on a flat surface. Every point on the sphere or spheroid has a corresponding point on the flat map. Perhaps the most fundamental observation about map projections is that linear scale cannot be the same everywhere. In other words, map projections are geometrically distorted representations in contrast to globes, which maintain a shape very close to that of the planet and a scale that is virtually the same everywhere. There are many map projections in use and the five represented in **Figure 2** illustrate a variety of shapes and relative sizes of landmasses that can result when different systems of projection are applied.

Although it is necessary to tolerate distortion when projecting the Earth onto a flat surface, the distortion in map projections is systematic and in some cases highly useful. Every rectangular projection of the world, for example, has one or two straight lines on which there is no distortion, and distortion increases at right angles away from the line(s) of no distortion. Although small-scale maps of large areas have major visible distortion, maps of small areas, say a city or small region, have so little distortion when the projection is centered on the represented area that the distortion is of interest only for the most exacting uses.

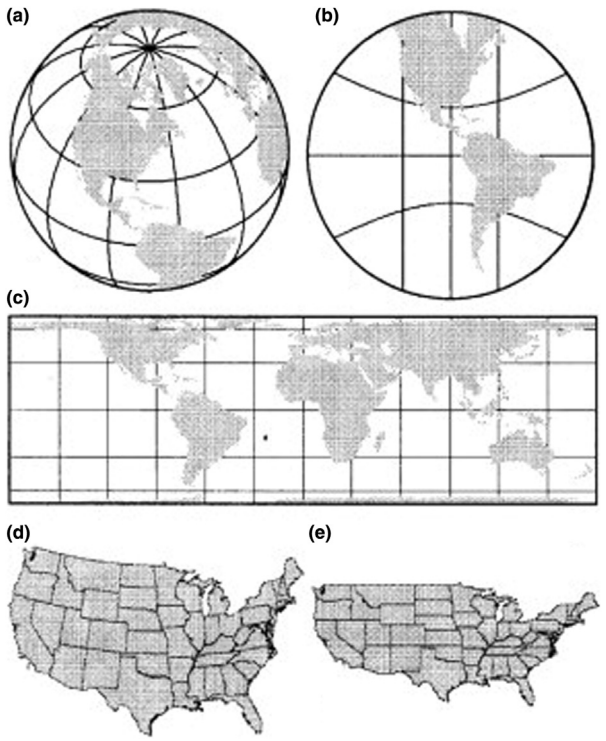

Figure 2 Five map projections showing a variety of shapes and sizes of landmasses when different projection systems are applied: (a) orthographic, the projection that looks like a view of a globe; (b) gnomonic, a very distorted projection, but the only one on which all great circles are straight; (c) cylindrical equal area, which preserves correct sizes of all areas but has extreme changes in linear scale; (d) Alber's, which is equal area and is commonly used for the USA; and (e) plane chart, which represents all degrees of latitude and longitude at the same length regardless of their length on the Earth's surface.

The development of geographic information systems resulted in renewed interest in map projections in the later decades of the last century because many maps were transformed into digital files and data had to be 'unprojected' (transformed from flat map Cartesian coordinates to longitude and latitude) or 'reprojected' (transformed to another projection) to be compatible with data from other maps or other sources. Data are now generally gathered directly in digital form, but because coordinate systems (e.g., Universal Transverse Mercator) by which data are gathered are tied to specific map projections, the user of the data must be keenly aware of that underlying projection. Even the limited amounts by which map projections affect location within small areas will affect the compatibility of the multiple data sets that are brought together to produce a map. Digital files are also commonly tied to a world projection, not one centered on the user's particular small area of interest. If the underlying world projection is one on which each degree of latitude and longitude is plotted at the same length, even a small patch of land such as a property parcel can be greatly distorted. The projection used for a small-area map is of little concern only when the projection is centered on that small area.

Despite the impossibility of maintaining consistent linear scale on a map projection, there are two other properties that may be maintained, though not both on the same projection. When, at every point on the map, maximum linear scale distortion is compensated by opposite distortion in the opposite direction, area scale is maintained. On the other hand, if the linear scale is the same in every direction at each point, angles at all points will be maintained (not angles between distant points, but angles right at the individual points). The terms for these two categories of projection are equal area (or equivalent) and conformal (or orthomorphic), respectively. Many projections have neither compensating nor equal linear scale in opposite directions at all points; these projections are neither equal area nor conformal.

Some projections have special properties that make them useful for specific purposes. The Mercator projection, when centered on the equator, shows all loxodromes (or rhumb lines, i.e., lines of constant geographic direction) as straight lines. This property has been important to navigators. It was a profound development because most loxodromes on the Earth (and on globes) are spirals and the Mercator projection simplified route planning. The projection suffers from extreme deformation of area, and, like many other projections, has been used inappropriately (Monmonier, 2004). The Mercator maps hanging on classroom walls have led many students to think of South America and Africa as a relatively small land bodies relative to the greatly exaggerated high-latitude landmasses of Greenland and Russia, whereas South America is over eight times the size of Greenland and Africa is about 1.75 times the size of Russia.

There are more than 100 named projections (see Snyder, 1993), and cartographers try to choose a projection (named or not) that is appropriate for the mapping purpose at hand. The discrete naming of projections, while providing convenient reference, obscures the fact that there are an infinite number of ways to represent the Earth on a flat surface. Computer construction of projections is resulting in increasingly tailored projections (Laskowski, 1997).

Spatial Relationships

Maps show spatial relationships in a readily comprehended form. Metric relationships include distance, direction (angle), and area; topological relationships include such properties as connected to, inside, outside, left of, and right of. The concept of spatial relationship also includes spatial (or geometric) form; an entity may be considered to have a footprint of a point, line, or area, and if quantities are associated, the entity may be thought of as a pole, ribbon, or volume. Spatial form also includes shape (e.g., Long Island vs Martha's Vineyard) and distribution ('more of something here than there'). The relevant spatial relationships vary with the type of map and with how it is used. Street map users want to know the location of a feature to which they are navigating and the turns and distances to reach it. Users of a world population map want to know where population is concentrated and where it is sparse.

Like the general concept of scale, spatial relationship is often taken for granted by the practiced map user. Sometimes the reason for using a map is not the spatial relationship itself, but what it means. A 'hot spot' on a map showing cancer rates suggests the need for further study and possible action; the spatial relation on the map is simply the trigger and is suggestive, at best, of any factors of causation. Whether a street, population, or cancer map, it is the spatial relationships that make it a useful device. Without them, there would be no need for a map.

Generalization

As scaled representations of the Earth or of portions thereof, every map is generalized. Lines have fewer details in them than present in the feature on the Earth's surface, and only selected features are shown. When quantities are depicted, they are usually shown by category rather than by exact value. Generalization is highly useful in focusing the map on important information instead of cluttering it with overabundant detail. It also means that some important features may not be there, not because they do not exist but because they have been generalized out of the map.

Despite the obvious necessity for generalization in making representations that are literally thousands and millions of times smaller than the original territory, it is not always understood. The small street maps of major cities that appear on paper road maps may be assumed by users to include every street, and a surface feature that falls between the contours of an elevation map may surprise and confuse the hiker. The failure of a map due to generalization is one of the reasons that there are many different maps of the same area at different scales and with different selections of features. The larger the scale (i.e., the larger the representation of a given area), the less generalization is necessary and the more useful it is likely to be for finding small features. The smaller the scale (the smaller the representation of a given area), the more likely it is that broad patterns will be visible because of the lack of obfuscating detail.

The relationship between scale and generalization is not an exact one. Some maps have much more detail than others of the same physical scale. The capacity of a map for depicting detail and the clarity of a detailed map for the user are affected by design choices. A traffic map, for example, with color of road segment indicating vehicle density is a clearer representation than one using road width, which is likely to have overlapping symbols or considerably fewer categories.

Symbolization and Data Modeling

The very act of representing a feature with a symbol is a form of generalization. The symbol is not the original and every map is a set of symbols. Mapmakers have been highly creative in employing symbols.

Geometrically, mapmakers can employ only points, lines, and areas as symbols on flat maps, though they may be used so as to simulate the illusion of a third dimension. The use of three-dimensional media, such as molded plastic or the product of a 3D printer, allows the explicit use of volume as a symbol. Dynamic mapping can include an additional dimension, usually time or changing point of view. The development of virtual reality allows the illusion not only of a third dimension, but of the intersection of surfaces and of movement through perceived three-dimensional space or over a period of time. Regardless of medium, the geometry of symbols generally reflects the way in which humans think of the object relative to the scale of the map. For example, a city is a large area on a street map but a point on a world map. A river on a local, large-scale engineering map is an area, but the same river appears as a line on a smaller-scale regional map. On thematic maps, on the other hand, symbols often represent conceptual volume, not specific features. A dot map of population, for example, uses numerosity of dots to represent quantities of persons.

There are a limited number of characteristics of symbols, called 'visual variables,' which can be used to represent the characteristics of the feature represented (Bertin, 1983). Darkness–lightness of the symbol and size are the most commonly used characteristics to represent quantity. Shape, color hue, and orientation most often represent differences in kind. Patterning in area symbols may also be used to distinguish characteristics of features.

The modeling of geographic space in digital files introduces a different form of representation from that on paper or similar physical medium (Peuquet, 1991). Geometric concepts of point, line, area, and volume are still valid, but there is no scaled physical representation or visual symbols associated with features. Rather, numbers represent location, and codes or descriptive terms represent what exists at a location. These digital files can be thought of as data models as opposed to a physical model. They can almost always be converted into a map, but they are much more flexible. A digital elevation model, for example, i.e., a file of elevation values at points, can be converted into any of the maps in **Figure 3** and is often portrayed with a combination of these techniques.

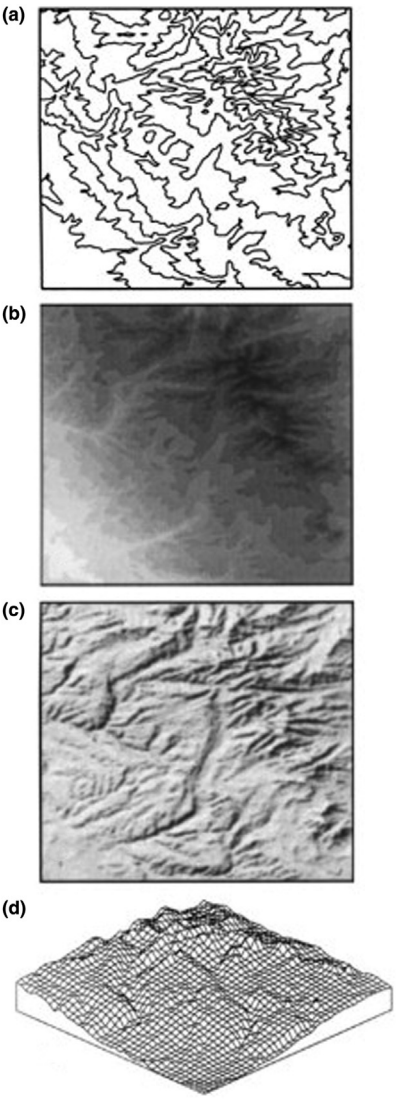

Figure 3 Four ways of representing a set of elevation data; lines are used in the contour map (a) and three-dimensional diagram (d), whereas area symbols are used in the layer-shaded map (b) and in the shaded relief map (c).

Categories of Maps

In the latter half of the twentieth century, maps were often broadly categorized as general reference or thematic. Although the distinction is highly artificial, often difficult to apply, and increasingly controversial, it is still in use. A general reference map (Figure 4) shows a variety of individual features, such as cities, rivers, regions, lakes, and oceans, and is generally used to find specific features. The general reference map is selective of features in any category: only 'important' cities, rivers, etc., are identified. Although lettering on the map often stands out visually, there is limited visual hierarchy among the features of the map, as all the features must be distinguishable from one another. A thematic map (Figure 5) shows a distribution of one or more phenomena and is used to visualize spatial form. The thematic map is generally based on all the known instances of the phenomenon even though none is shown individually. The thematic map generally has clear visual hierarchy, with the symbols representing the featured distribution (e.g., the colorful symbols representing mussels in Figure 5) standing out more than other features on the map (e.g., lines representing state boundaries).

Sometimes a third category, such as navigation maps, is added. Others consider the general reference and thematic categories to be opposite ends of a scale, with individual maps fitting somewhere between. Navigation maps, such as street guides, are then placed near the general reference end because many different specific items are represented, but not necessarily every item in a category. They have certain thematic qualities, however, because of the focus on the features relevant to finding one's way. Inventory maps, such as a map showing the residence of every patient with a certain disease, are closer to the thematic end of the scale, but they share the general reference characteristic in that individual items are shown.

There is good argument that the general reference–thematic classification refers more to map use than to maps themselves. One can look at a thematic map to see how large one political unit is relative to another, which is a function generally associated with the term 'general reference.' Likewise, one can look at a general reference map and observe overall pattern of drainage or evidence of population, uses generally associated with the term 'thematic' (Robinson and Petchenik, 1976).

There are valid arguments as well that these categories are so artificial that they are not useful (Wood et al., 2010). There is no question that many, if not most, maps are difficult to place into these categories, especially with the flexibility afforded by data availability and production possibilities of electronic mapping. A recent textbook avoids any use of such categorization (Krygier and Wood, 2011).

Design

The design of a map is the whole set of decisions on how underlying information is symbolized, and it also refers to the resulting look of the map. It has been affected by virtually every technological development in the history of mapping, and the ability of the observer to distinguish differences between maps printed using intaglio as opposed to lithograph exemplifies the profound influence that technological change has on the way a map looks. In the late twentieth century and early twenty-first century, electronic production and dissemination are having profound influences on design. Traditional rules of thumb, such as 'the darker the more,' when depicting quantitative information on the traditional monocolor paper map have given way to rules of thumb about using the full range of color characteristics effectively. In digital display, using a bright color on a dark background may be far more effective for emphasizing a high category of quantitative data than dark on a light background. We can also use motion for emphasis; a blinking item on the map will surely catch attention. Even more important, maps need not be made for everyone but can change to meet user needs. Color choices, for example, can accommodate users with red-green color deficiencies, but the rare person with a yellow-blue deficiency need not be left out. An alternate version of the map, identical in all respects except the choice of colors, can be an option.

Figure 4 Madagascar physiography, an example of a general reference map, Central Intelligence Agency, 2014. https://www.cia.gov (original is in color).

Tools are also available on the World Wide Web to assist in design decisions. ColorBrewer helps a map author to find combinations of colors that will be effective for various media and determine which combinations are suitable for those with red-green deficiencies (Brewer et al., n.d.). TypeBrewer gives assistance in the selection of type (Sheesley, 2006).

Most maps today are viewed on electronic devices, not in books or printed paper sheets. A new set of maps on the Web, if

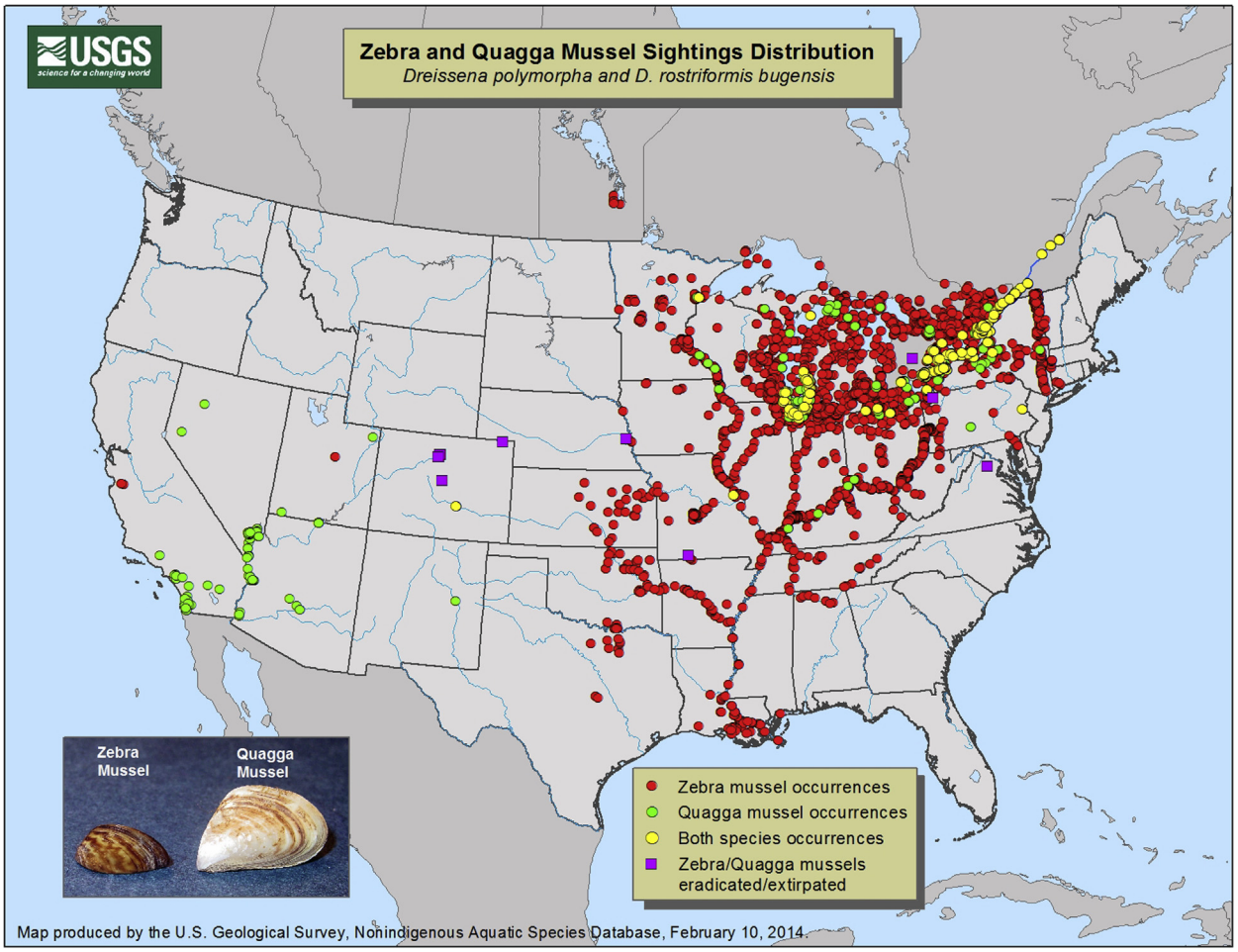

Figure 5 Zebra and quagga mussels map, an example of a thematic map, US Geological Survey. http://nas.er.usgs.gov/taxgroup/mollusks/zebramussel/ (original is in color).

content or design is sufficiently attractive, can draw thousands of views within a short time of posting because the link is shared on social media. Having been created by a wide variety of authors, online displays are broadening the way in which maps are designed, with a mixture of quality in the results. Some of the old common design features, such as a neat line (line delineating the area of the map display), are used far less because they are less often relevant. On the other hand, there is little doubt that some maps are designed better than others and that some of the traditional as well as new design guidelines are useful.

Technology is not the only influence on design. Intellectual changes are reflected in map textbooks (Krygier and Wood, 2011), graphic critiques more broadly (Tufte, 2001), and publications of exemplars in cartography (Wallace and Huffman, 2012).

The History of Maps and Cartography

The history of cartography begins earlier than recorded time and in all likelihood had its origins in gestures and ephemeral marks in the soil (Godlewska, 1997). Maps have undergone profound development in the millenia since then, and the following summary is based on Thrower's (1996) readable account.

The development of civilizations and of cartographic practice were probably highly intertwined, and at least some cartographic artifacts remain from ancient Mesopotamia and Egypt, including a clay tablet from roughly 3800 BC with mountains and a river. Extant land ownership maps date to 2000 BC. The development of grid systems, associated with less subjective and more scientific approaches to mapping, was reflected in written materials in China in third century AD, and the earlier development of paper in China provided a flexible and highly portable physical medium for drawing maps.

The concept of the world as a sphere was basic to the development of logical representation of large areas of the world. Measurements make sense only when the curvature of the planet is taken into account. The idea seems to have had its origins with the Pythagoreans and was supported in work by Plato and Aristotle. Eratosthenes (276–196 BC), who served as the librarian at Alexandria, used the concept of the spherical

Earth and the relationship between noon Sun angles at different locations to measure the size of the planet. Because the exact length of a stade, the unit of measure in use at the time, is not known, it is difficult to know exactly how accurate his estimate was, but it seems to have been remarkably close and a better estimate than several that followed, as scholars attempted to correct the figure. A successor in his position in the second century AD, Claudius Ptolemy, compiled the seminal tome entitled *Geographia* that contained a guide to making maps, including instructions for several map projections and the latitude–longitude location of about 8000 places.

Mapmaking in China and South Asia is characterized by remarkable developments in roughly the same time period as the accomplishments of the ancient Greeks. There is reference to a map on silk in the third century BC, and maps were used by the military and administrators in the Han Dynasty. In later centuries, the first printed map appeared in China well before the development of printing in the West (approximately 1153 AD vs 1472 AD).

European cartography during the Middle Ages was characterized by maps that are highly schematic by modern standards but were important cultural documents of their time, having been associated with religious institutions and beliefs. One schema was a circular map with east at the top showing Asia as the upper section of the circle, separated from Europe in the lower left and Africa in the lower right by the Don River and the Nile. Europe and Africa were separated by the Mediterranean Sea. These maps are known as T–O maps – the rivers and Mediterranean forming a rough T shape and the oceans surrounding the landmasses forming the O. The second schema was a circle with north at the top and, within the circle, temperature zones that included the torrid one along the equator, bordered by the temperate zones, which in turn were bordered by frigid zones at the poles. Route maps of the time carried far more detail than these schematic world maps, but the most remarkable detail and the most scientific approach was found in the portolan charts. These maps were used in navigation, primarily in the Mediterranean Sea, and showed intricate coastlines with hundreds of identified coastal locations. At several locations within the water are compass roses with direction lines extending from them.

Cartography in the Islamic world during the European Middle Ages, while varied, included continuation of the Greek scientific tradition. The scientific instrument called the astrolabe, a device for measuring angles to stars, was important in finding the way to Mecca. In the early 1400s, Ptolemy's *Geographia*, which had been preserved in Byzantium, reached Italy via refugees fleeing the invading Turks. Translated into Latin, the work had an extraordinary influence on subsequent development of cartography. Systematic projection of the spherical Earth to the flat map followed Ptolemaic instructions for well over a century.

Accurate mapping depends on the ability to measure both latitude and longitude accurately. Latitude, which could be measured by angles to the Sun and other heavenly bodies at specific times of the day, was far easier to measure than longitude, which depends on differences in time from one location to another. In a competition sponsored by the British government, John Harrison in the latter half of the eighteenth century produced a chronometer that could retain accurate time at sea (Sobel, 1995). It was a turning point in the accuracy of maps covering large regions of the world.

The rise of detailed scientific mapping of the Earth's surface in the form of topographic mapping owes much to the Cassini family, whose influence on cartography began with the appointment of astronomer Giovanni Domenico Cassini to the Académie Royale in Paris, which along with the Royal Society of London was concerned with mapping and other scientific problems. Under his direction, maps recorded only locations that were determined by astronomical methods, and detailed measurement of a degree of latitude was undertaken at various locations to test the theory of the Earth as a prolate spheroid, which affects the placement of locations on flat maps. The detailed mapping of France was continued under three more generations of the Cassinis (Jacque, César-François, and Jean) and was completed in 1793.

In the eighteenth and early nineteenth centuries, one of the most important developments in cartography was the mapping of thematic data (Robinson, 1982). With the practice of accurate base maps established, it was feasible to focus attention on the spatial arrangement of various phenomena of interest such as elevation, population, and winds. Symbol systems such as isolines, dots, and the choropleth method came into being and are commonplace in modern mapping.

Changes in cartography in the twentieth century included the use of aerial photographs in topographic mapping and changes in production technologies. The latter half of the century saw the development of satellite sensing for gathering data for maps, the deployment of global positioning systems (constellations of satellites sending signals from which latitude and longitude are determined), and the infusion of computers into mapping, including the development of geographic information systems and Internet mapping. The variety of thematic mapping methods has recently undergone rapid change as attention has focused on the use of computers to overlay multiple phenomena, produce animated maps, and develop sophisticated visualizations of spatial phenomena (Thrower, 1996).

The richness of cartographic history throughout the world and over many centuries is reflected in *The History of Cartography*, a monumental multivolume series, which continues production as this encyclopedia is being written (Harley and Woodward, 1987; History of Cartography, 2014). The project itself is an important intellectual event in cartographic history.

See also: Cartographic Visualization; Cognitive Maps; Dynamic Mapping in Geography; Ethnic Conflict, Geography of; Geographic Information Systems; Tactile Maps in Geography; Thematic Maps in Geography.

Bibliography

Bertin, J., 1983. The Semiology of Graphics (W.J. Berg, English Trans.). University of Wisconsin Press, Madison, WI.

Brewer, C., Harrower, M., The Pennsylvania State University. n.d. ColorBrewer 2.0: Color Advice for Cartography. www.colorbrewer.org.

Brewer, C.A., 1990. The effect of color on the perception of map scale. In: Everett, W. (Ed.), Student Honors Competition Winning Papers, Association of American Geographers. Cartography Specialty Group, Toronto, Canada.

Eastman, J.R., 1981. The perception of scale change in small-scale map series. The American Cartographer 8 (1), 5–21.
Godlewska, A., 1997. The idea of the map. In: Hanson, S. (Ed.), Ten Geographic Ideas that Changed the World. Rutgers University Press, New Brunswick, NJ.
Harley, J.B., 1989. Deconstructing the map. Cartographica 26 (2), 1–20.
Harley, J.B., Woodward, D. (Eds.), 1987. The History of Cartography. University of Chicago Press, Chicago, IL.
History of Cartography, The. 2014. http://www.geography.wisc.edu/histcart/.
International Cartographic Association, Association Cartographique Internationale, 1996. New definition of cartography adopted as resolution 1 by the [10th] General Assembly [Barcelona 1995]. ICA News 26, 1. Reprinted in Cartography, 1997, 26 (1), 83.
Jet Propulsion Laboratory, 2009. The Shuttle Radar Topography Mission: The Mission to Map the World. http://www2.jpl.nasa.gov/srtm/.
Krygier, J., Wood, D., 2011. Making Maps: A Visual Guide to Map Design for GIS, second ed. Guilford Press, New York.
Laskowski, P., 1997. The distortion spectrum, monograph 50. Cartographica 34 (3), 67–95.
MacEachren, A.M., 1995. How Maps Work: Representation, Visualization, and Design. Guilford Press, New York.
Monmonier, M.S., 2004. Rhumb Lines and Map Wars: A Social History of the Mercator Projection. University of Chicago Press, Chicago.
Peuquet, D., 1991. Methods for structuring digital cartographic data in a personal computer environment. In: Taylor, D.R.F. (Ed.), Geographic Information Systems: The Microcomputer and Modern Cartography. Pergamon Press, Oxford, UK.
Robinson, A.H., 1982. Early Thematic Mapping in the History of Cartography. University of Chicago Press, Chicago, IL.
Robinson, A.H., Petchenik, B.B., 1976. The Nature of Maps: Essays toward Understanding Maps and Mapping. University of Chicago Press, Chicago.
Robinson, A.H., Morrison, J.L., Muehrcke, P.C., Kimerling, A.J., Guptill, S.C., 1995. Elements of Cartography, sixth ed. Wiley, New York.
Snyder, J.P., 1993. Flattening the Earth: Two Thousand Years of Map Projections. University of Chicago Press, Chicago, IL.
Sheesley, B., 2006. TypeBrewer: A Map Design Help Tool for Selecting Typography. www.typebrewer.org.
Sobel, D., 1995. Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time. Walker, New York.
Thrower, N.J.W., 1996. Maps and Civilization: Cartography in Culture and Society. University of Chicago Press, Chicago, IL.
Tobler, W., 1988. Resolution, resampling, and all that. In: Mounsey, H., Tomlinson, R. (Eds.), Building Databases for Global Science. Taylor & Francis, London, pp. 129–137.
Tufte, E.R., 2001. The Visual Display of Quantitative Information. Graphics Press, Cheshire, CT.
Wallace, T.R., Huffman, D.P., 2012. Atlas of Design, vol. 1. North American Cartographic Information Society, Madison, WI.
Weldon Cooper Center for Public Service, University of Virginia, 2013. The Racial Dot Map, Dustin A. Cable, author. http://www.coopercenter.org/.
Wood, D., Fels, J., Krygier, J., 2010. Rethinking the Power of Maps. Guilford Press, New York.

Case Method and Case Study Collections

Kirsten Lundberg, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

An introduction to the case method (CM) of teaching, its history, and a representative list of global case collections focused on the disciplines of public policy, public health, journalism, environment, and other social sciences. The CM, also known as discussion-based learning, is a powerful, interactive, and student-centered teaching tool in the physical or online classroom. 'Teaching' case studies are ready-made curricula.

What Is Case Method?

The case method (CM) of instruction has existed, in one form or another, since ancient Greece. Socrates was renowned for his approach to teaching: instead of lecturing, he engaged his listeners in a give-and-take dialogue. (Example is often given of the dialogue called Meno, in which Socrates through questions leads Meno, a self-declared expert on virtue, to the uncomfortable acknowledgment that he is not sure what virtue is. See: http://classics.mit.edu/Plato/meno.html.) The so-called Socratic method became the backbone of the CM of teaching, which emerged in the nineteenth century at Harvard Law School and was adapted in the 1920s by Harvard Business School (HBS). (For a précis of the history of CM in teaching law, business, and medicine see Garvin (2003).) While law school cases concerned legal case law, HBS applied the teaching approach to professional challenges. Its cases were not hypothetical but based on practitioners' efforts to resolve some of the thorniest dilemmas of the profession.

By the 1960s, the CM had become the predominant pedagogical approach at business schools around the globe. As INSEAD, a leading global business school, put it: "The case method challenges students to use problem-solving, decision-making skills just as they would in the workplace. The approach fosters teamwork, individual reflection and active participation." (INSEAD Web site. See: http://www.insead.edu/facultyresearch/research/insead_cases.cfm.)

As word of its appeal spread, the teaching approach – alternatively called active learning, participant-centered learning, or discussion-based learning – moved slowly into other disciplines. By the early twenty-first century, it was in use to varying degrees at schools of public policy, public health, education, military training, medicine, religion, physical sciences, engineering, and journalism. T. Grandon Gill, a professor of information systems and decision sciences at the University of South Florida, wrote that CM is especially relevant when students are or will soon become professional practitioners. He said:

> The "real world" tends to produce complexity as the needs of stakeholders collide in a way that makes "fit", rather than "optimization" the most relevant goal. Where fit is the goal, judgment is nearly always required. And the case method is well suited to presenting situations where judgment is needed.
>
> Gill, 2011

CM in the Classroom

The Aim of CM

The CM won its growing legion of adherents, thanks to its powerful effect in the classroom. But what is it? It is perhaps easier to define what it is not: CM is not a lecture, not a problem set, and not a small-group discussion. Rather, it features a faculty-guided discussion based on a close, pre-class reading of a case. Students are encouraged to speak more than the instructor, to debate with one another rather than respond directly to the teacher, and to provide substantive explanations for the arguments they advanced.

Under the CM model, faculty serves not as the source of correct answers, but as discussion guides. Students, for their part, are no longer empty vessels into which learned professors pour knowledge (the lecture model) but are instead active participants in their own learning. As David Garvin of HBS puts it in an essay on case teaching: "Only if teachers and students work as partners will the true ends of education—the ability to use knowledge, to think creatively, and to continue learning on one's own—be achieved" (Christensen et al., 1991: p. 5). Students are asked to think for themselves and be on their feet to analyze, hypothesize, prescribe, and defend. Garvin's colleague C. Roland Christensen wrote:

> The discussion process itself requires students to become profoundly and actively involved in their own learning, to discover for themselves rather than accept verbal or written pronouncements. They must explore the intellectual terrain without maps, step by step, blazing trails, struggling past obstacles, dealing with disappointments.
>
> Christensen et al., 1991: pp. 16–17

CM asks students to learn at three levels: cognitive, behavioral, and personal. It strengthens both their subject-specific and transferable skills. The Australia and New Zealand School of Government (ANZSOG) says that "… the power of case teaching comes from the fact that it gives practical shape and illustration to concepts. It focuses on people as decision makers, so class participants can 'stand in the shoes' of the manager whose circumstances or dilemma are described in the case." (See the ANZSOG Web site: http://casestudies.anzsog.edu.au/.) Case use can be flexible, according to the preference of the instructor: in some courses,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.41003-2

cases feature in every class; in others, students might discuss only one.

Effectiveness of CM in Classroom Discussions

During case discussions, students come to understand that the world is not black and white, but gray, and that throughout history, most decisions are far from preordained. Rather, history is made by ordinary individuals making the best choice they can in spite of significant deficits. Typically, decision makers don't have enough money, personnel, time, and – most crucially – information. Nonetheless, they must act. Case discussion provides students with a vicarious workplace experience. While the effectiveness of CM has been under researched, one study found that it increases students' tolerance for ambiguity (Banning, 2003). In a world characterized by rapid change and a need for flexibility, this was a powerful outcome.

CM has proved effective in multiple fields. The National Center for Case Study Teaching in Science, for example, found it invaluable in teaching the hard sciences. As the Center says, "… cases can be used not only to teach scientific concepts and content, but also process skills and critical thinking." Moreover,

> Since many of the best cases are based on contemporary, and often contentious, science problems that students encounter in the news, the use of cases in the classroom makes science relevant. We have also found the method to be extraordinarily flexible. We have seen it used as the core of entire courses or for single experiences in otherwise traditional lecture and lab courses. Moreover, cases can be presented in a variety of formats and taught in a variety of ways, ranging from the classical discussion method used in business and law schools to Problem-Based Learning and Team Learning, with their emphasis on small-group, cooperative learning strategies.
>
> See the Center's Web site: http://sciencecases.lib.buffalo.edu/cs/about/

'Teaching' Cases

Definition of 'Teaching' Case

An essential component of CM is the case study. Multiple definitions of the term 'case study' have long muddied public understanding of how a case is used in teaching. For example, 'case study' in popular parlance usually means an example of best practices; it can also refer to an academic analysis.

But what has became known as a 'teaching case' (to differentiate it from the others) is not an analysis of a situation, nor a description of best practices. It is a real-life story intended as a vehicle for classroom discussion. A teaching case describes a situation; it does not prescribe a solution. It provides a snapshot of a specific incident which, through discussion, illuminates larger principles. A teaching case engages students so they can practice how to take action. It introduces students to the realities of their chosen profession. It also tests students' ability to apply theory to a real-world situation.

Even with that clarification, practitioners differ on what constitutes a teaching case. Consensus has settled around three broad categories: decision forcing, retrospective, and point of view. Decision-forcing cases are by far the most popular kind of teaching case and, many would argue, the most effective in the classroom. A decision-forcing case puts students in the shoes of a protagonist confronting a thorny dilemma to which there was (1) no single correct answer and (2) a range of equally viable solutions. A fully researched, interview-based case describes the wide range of constraints – political, economic, and personal – and stops at the point where a decision is imminent. Students are asked to discuss what they would do in his or her place.

In contrast, a retrospective or research case (or 'cadaver case,' as a former case program director termed it) looks at a case incident in its entirety, from inception to resolution. Instead of placing students in a decision-making role, it asks them to look back at a concluded episode, identify the turning points, speculate on what might have been done differently, and try to rethink the outcome. Finally, a point-of-view case tells a story from a single point of view. It is useful when a more comprehensive case is impossible for logistical or political reasons – for example, a case based on a political dissident in an authoritarian regime. All three types of cases can, if necessary, be complimented by secondary sources.

'Teaching' Case Design

There is considerable controversy over whether a teaching case should be hypothetical, disguised (names and other details fictionalized), a simulation, or a hybrid. Purists insist that the best cases are real stories about real people, that fact trumps fiction in nearly every instance, and that students cannot easily dismiss a reality-based case as 'implausible' because it did, in fact, happen. One introduction to a collection of teaching cases argues:

> (Cases) … based on actual events tend to be more effective than the ones that are made up. This is because fictitious cases often tend to be too neat and too comprehensive… Unfortunately, life … is typically messy and the information available to decision makers at any given time is neither comprehensive nor logically cohesive.
>
> Brynard and Erasmus, 1995: p. 7

Moreover, real cases based on primary sources (interviews) serve a second function as 'microhistory,' or the history of how decisions are made. Cases can be an important resource for future historians because they chronicle the minutiae of what leaders considered in coming to decisions about crucial aspects of the field.

Many faculties, however, routinely use and express professional satisfaction with hypotheticals, disguised cases, simulations, and hybrids. These are especially useful in dealing with incidents considered too sensitive to discuss on record by the participants as it is. Yet another category sometimes termed as 'case study' is an assembly of raw materials such as newspaper articles, which students read in preparation for a class discussion. (For an interesting review of case classroom challenges from the faculty point of view, see the collection of more than 150 short case studies written by economics teaching staff hosted by The Economics Network at the University of Bristol (UK): http://www.economicsnetwork.ac.uk/showcase.)

Case Examples

There is no single case that can serve as an example of all the possible approaches to writing a case. But a good teaching case tells a story, raises discussions of a range of topics, and stops at a decision or a turning point. Here are two examples of decision-forcing cases. One is a journalism case; the other about public health. Each could be used in other disciplines as well, such as public policy, international relations, conflict studies, or crisis management.

Case Example 1

In January 2006, assailants kidnapped Jill Carroll, a stringer in Iraq for the *Christian Science Monitor*. The case tracks the *Monitor's* editorial leadership over the next 2 months as they struggle to manage their multiple constituents: the State Department, the Baghdad bureau, the Arab press, the US Central Intelligence Agency (CIA), the Federal Bureau of Investigation (FBI), Carroll's family members, and so forth. It ends as the *Monitor* prepares a public service announcement (PSA) to run on Al Jazeera pleading for information about Carroll. The FBI has advised the *Monitor* to stay under the radar to secure Carroll's release, while the CIA has advised it to publicize her plight widely. The decision point: should the *Monitor* run the PSA?

The case discussion will cover far more than what to do about the PSA. Under the skillful guidance of an instructor, students can talk about how a news organization, which usually covers the problems of others, manages an internal crisis; how to negotiate with a wide range of stakeholders or constituents; how to report in conflict situations; and what obligations a news organization might have to a stringer in a war situation.

Case Example 2

In October 2012, as Superstorm Sandy approached the East Coast, the North Shore-Long Island Jewish Hospitals (North Shore-LIJ) network braced for landfall. The network owned three hospitals in low-lying areas; just a year earlier, during Hurricane Irene, it had evacuated those hospitals. But Irene veered off and the expensive evacuations were for naught. The decision point: should North Shore-LIJ evacuate this time?

The case discussion addresses this very question. But the case document also allows students to explore the legal and ethical trade-offs between evacuating patients and sheltering in place, the role of hospitals during a natural disaster, the relationship between a private hospital network and its government overseers, the characteristics of emergency management, and what kind of decision-making structure works best in a disaster?

Case Example Reuse

These are only two examples. There exist thousands of cases about thousands of topics. Some are narrowly conceived to serve in a specific class; others can be used in a variety of courses and classes. Good cases can have a shelf life of many years. With good writing and a strong conceptual approach, the case 'incident' becomes a snapshot which captures reality for a moment, and raises a discussion on a range of higher level issues grounded in the reality of the original event.

Many cases have epilogues; some have teaching notes as well. What actually happens in a case should not matter to its pedagogical usefulness. If students are asked to put themselves in the shoes of the decision maker *at that time*, they will engage with the dilemma regardless of the eventual outcome. What happened was simply the choice made on the spot; the case does not speak to if that choice was the only, the best, or the most beneficial to all parties.

Levels of Learning

In the course of preparing for and discussing a case in the classroom, students typically learn at three levels simultaneously: factual material, behavioral/analytical, and personal. The first level is cognitive: students absorb information about the subject treated in the case, be it the history of government/media relations in a given country or the details of an aid project in Haiti or the factors an engineering firm considered in determining where to site a bridge. Such knowledge is most easily acquired and, often, most quickly forgotten.

The related analytical skills, however, last a lifetime. Preparing for and discussing a case require students to go through five steps. The first is analysis: what is going on here? Whose problem is it, and are there shared interests? The second is diagnosis: what are the available options for a solution? Third, students are asked to anticipate consequences of options and weigh the respective pros/cons. Fourth, they must prescribe a solution to the dilemma and fifth, defend it (or allow themselves to be persuaded by others). Taken together, these steps help students exercise and strengthen their skills of judgment, and of critical thinking.

In his classic article on why CM outperformed lectures, HBS professor Charles I. Gragg memorably explained in the title: 'Because wisdom can't be told' (Gragg, 1954: p. 6). Subsequent research reinforces the view that students retain the longest what they teach themselves. According to Professor Clyde Freeman Herreid of the University at Buffalo (NY), who has written extensively on the use of CM in science teaching:

> If reading, arguing, and challenging are hallmarks of critical thinking, then case studies are the poster children for the process. Most of them are discipline specific, certainly. But they all grapple with the essence of critical thinking—asking for evidence—developing a habit of mind that should permeate everyday life.
>
> Herreid, 2004

Students also acquire lifelong behavioral skills in a case classroom. In the course of discussing a case, they practice negotiation and public speaking. They often have the chance to work in teams. They learn how to persuade others, and also the value of listening – of being willing to change one's mind in the face of a convincing argument. These skills will serve them well in whatever field they choose, and they are unlikely to forget them.

At the deepest level, case discussions are able to help students confront their own biases and prejudices. Not necessarily the big issues such as racism or sexism, but smaller ones: that growing up in a suburb might lead one to think differently than someone who grew up on a farm or that socioeconomic differences can influence one's world view. Students in case classrooms are routinely surprised that a problem whose solution they deemed 'obvious' inspired a very different solution from fellow students. No matter the field, a better understanding of one's personal history and social context helps guard against an emotional, knee–jerk reaction to professional situations. A skillfully guided case discussion could lead to valuable self-discovery.

Case Collections

Below is a list of case collections useful to instructors in the behavioral and social sciences. A few, such as science and engineering, may seem outside the disciplinary limits. But they are included because the cases by and large focus on aspects of those fields that overlap with behavioral and social science concerns.

While every effort has been made to discover case programs around the world, this list does not pretend to be comprehensive. Some of the collections make their cases available to the public for free, others charge, and a third group may require a personal request to acquire a case. See each individual Web site for detailed information.

Case Consortium @ Columbia, Columbia University, New York, NY (https://casestudies.jrn.columbia.edu)

This collection of multidisciplinary, online, multimedia cases started in 2007, with a focus on journalism. By 2013, it included upwards of 60 cases for use in a variety of disciplines, including public health, public policy, journalism, and sustainable development. The Case Consortium's mission is to produce excellent, case-based curricula for disciplines outside business, based on the view that students in all professional fields should have the opportunity to benefit from CM. The cases are free to registered faculty and are available for a fee to other users. The Web site offers guidelines on how to use cases in the classroom. The consortium also offers custom training to case teachers and writers.

Case Method Project, US Marine Corps University (MCU), Quantico, VA (http://guides.grc.usmcu.edu/content.php?pid=273794&sid=2339824)

The Case Method Project at MCU started as a one-person project in December 2007. By 2013, its 100+ cases were being used in a variety of programs throughout the Marine Corps. Instructors developed and taught their own cases, both in the US and globally. Some cases are based on original research but most came from secondary sources. The decision-forcing cases give the US Marines practice in solving a wide variety of problems, most of which defied conventional categories. They also encourage Marines to engage in the serious study of history. Besides decision-forcing cases, the project produced 'conference-style' cases, so-called because participants were not expected to prepare in advance but to learn about the case dilemma in the classroom. Finally, hybrid cases provided some background prior to class, and some in class.

Case Program, ANZSOG, Melbourne, Australia (www.casestudies.anzsog.edu.au)

The ANZSOG case collection, established in 2004, celebrated the publication of its 150th case study in August 2013. Its public sector cases focus on a range of issues in Australia and New Zealand, from accountability to policy development and disaster management. ANZSOG was established in 2002 as a unique initiative of governments, universities, and business schools to educate future public sector leaders in public policy and management. Case teaching is an integral part of its interactive pedagogical style. The program also offers workshops on case teaching and writing.

Case Program, Harvard Kennedy School, Cambridge, MA (www.ksgcase.harvard.edu)

The Case Program's collection of public policy cases numbered over 2000 by 2013. The cases cover a wide range of public policy topics, from health to national security, nonprofits, crisis/risk management, economics, public/private partnerships, intelligence, international relations, and public management. The online collection dates back to 1974, and includes principally decision-forcing and retrospective cases. The cases are free to registered faculty and are available for a fee to other users. The Web site includes guides to how to learn and teach using the CM.

Case Study Program, The Institute of Public Administration of Canada (IPAC), Toronto, Canada (http://www.ipac.ca/CaseStudyProgram-About)

The IPAC Case Study Program is Canada's premier portal to case study material for both teachers and public organizations. Founded in the late 1970s, the case inventory, as of 2013, included over 200 cases. The program offers a variety of services designed to foster the use and creation of case studies either through academics who share cases they use in class or via new cases based on real-life experiences at public organizations. The cases are public-sector oriented, and include decision-forcing cases, simulations, and knowledge transfer (sharing lessons learned). They are available electronically for a fee; free to IPAC members.

Case Study Unit, Lee Kuan Yew School of Public Policy, University of Singapore (http://lkyspp.nus.edu.sg/research-centres/publications/case-studies/)

This small collection of 20 cases covers a wide range of topics about public policy and public administration in Asia: foreign policy, health care and education, environment and climate change, leadership and public management, and others. Some are based on primary, others on secondary, sources of information. The cases are all based on real events, but in some instances details are disguised for teaching purposes. The cases

have been used in one or more of the master's degree programs at the school, as well as in executive education programs. Teaching notes exist for most of the cases. They are freely available on the Web site.

Centre for Social Investment, Heidelberg University, Germany (http://www.thecasecentre.org/educators/ordering/whatsavailable/collections/csi)

The Centre for Social Investment (https://www.csi.uni-heidelberg.de/) was founded in 2006 as an academic institute of Heidelberg University (economics, social science, law, and theological faculties). By 2013, its case collection held about 40 cases (plus teaching notes) about nonprofit management. The cases are based on two research projects: Strategies for Impact in Philanthropy (https://www.csi.uni-heidelberg.de/projekte_sip_e.htm) and Strategies for Impact in Education (https://www.csi.uni-heidelberg.de/projekte_sie_e.htm). The Center aims to improve understanding of social investment, civil society, social economy, and philanthropic institutions, and to inform about the policymaking that affected them.

China Case Center for Public Policy & Management (CCCPPM), Tsinghua University, Beijing (http://www.sppm.tsinghua.edu.cn/english/research/case/)

The China Case Center was founded in July 2004. Its goal is to produce cases for the School of Public Policy & Management in order to improve the quality of teaching and to train professional public managers. The Center also holds seminars on case teaching and writing to promote use of the CM among its faculty. The Center makes its cases available for a fee to other master of public administration programs in China. It also hosts an online forum for instructors to share case teaching experiences.

Electronic Hallway, University of Washington (http://hallway.evans.washington.edu/)

The Electronic Hallway at the Evans School of Public Affairs, University of Washington, is an online library of teaching cases and other curricular materials for faculties who teach public administration, public policy, and related subjects. The Hallway is a membership organization; cases are submitted by member faculties. Cases are grouped according to their organizational setting or policy area, such as arts and culture, nonprofit and NGOs, or law enforcement and corrections. The cases are sold around the world to graduate, undergraduate, and government training programs. The Hallway also runs short courses on CM.

Enduring Legacies Native Cases, Evergreen State College, Olympia, WA (http://nativecases.evergreen.edu/index.html)

This collection of free, decision-forcing cases examines Native American issues, including biology, economics, chemistry, environmental studies, ethics, health, women's studies, and natural resources, among others. The collection, which numbered over 80 by 2013, began in 2006 as a partnership between Evergreen State College, Northwest Indian College, Salish Kootenai College, and Grays Harbor College. Its goal is to develop and disseminate curricula and teaching resources about key issues in Indian Country. It has support from the Lumina Foundation for Education, the Bill and Melinda Gates Foundation, and the National Science Foundation. An Indian Advisory Board helps identify key issues.

Engineering Case Program, Rose-Hulman Institute of Technology, Terre Haute, IN (http://www.civeng.carleton.ca/ECL/5index.html)

The Engineering Case Program, which features accounts of real engineering work, originated at the Stanford University in 1964 with the support of the National Science Foundation. It continued under the sponsorship of the American Society for Engineering Education (ASEE). The Rose-Hulman Institute of Technology houses a collection of some 250 cases, and is in charge of distribution. ASEE jointly administers the Case Program with Rose-Hulman.

EnviroDx-Clinical Cases in Environmental Medicine, University of Utah (http://library.med.utah.edu/envirodx/index.html)

This innovative and small collection of four cases addresses environmentally related diseases. In the computer-based cases, the user is put in the role of a physician treating a patient with an unknown disease or condition possibly caused by exposure to environmental factors. The cases were developed by a team of content experts, instructional designers, graphic artists, multimedia producers, and programmers. The cases are designed to help students develop diagnostic and treatment skills, as well as obtaining relevant and reliable information quickly.

E-PARCC, Collaborative Governance Initiative, Program for the Advancement of Research on Conflict and Collaboration, Maxwell School, Syracuse University, New York (http://www.maxwell.syr.edu/parcc_eparcc.aspx)

E-PARCC is an open (Creative Commons license) online resource for those teaching the skills and concepts of collaborative governance: public management, networks and public management, collaborative governance, and collaborative problem solving. The Web site includes case studies about real-world occurrences, and simulations that engage students in collaborative problem-solving. Materials in the collection are selected through an annual competition, and are double-blind peer reviewed.

Epidemiologic Case Studies, Centers for Disease Control, Atlanta, GA (http://www.cdc.gov/epicasestudies/classroom_noro.html)

This small collection of eight interactive case studies was created to teach epidemiologic principles and practices. Based on real-life outbreaks and public health problems, they were developed in collaboration with the original

investigators and experts from the Centers for Disease Control and Prevention. The case studies require students to apply their knowledge and skills to the kinds of problems confronted daily by public health practitioners at the local, state, and national level.

Global Health Delivery (GHD) Project, Harvard University, Cambridge, MA (http://www.thecasecentre.org/educators/casemethod/resources/freecases/ghd)

The GHD Project (http://www.globalhealthdelivery.org/) at Harvard University offers a series of some 30 free teaching cases that examine principles of health-care delivery in areas with few resources. Teaching notes accompany the cases. The cases allow students to examine the real-world challenges faced by global health programs and organizations that seek to deliver health care in low-resource settings. GHD is working to establish a network across disciplines to help health-care stakeholders and practitioners benefit from and build on each other's successes.

Hubert Project, University of Minnesota, Minneapolis, MN (http://www.hubertproject.org)

Publicly launched in August 2012, the Hubert Project is both a platform for public affairs educators to connect globally in transforming public affairs education and a portfolio of open-access, web-based multimedia learning materials created and shared by this community. It is governed by an international group of public affairs schools. The portfolio contains three types of multimedia learning materials: *video briefs* featuring short introductions to a public affairs issue, *curated cases* that package publicly available materials into timely and relevant mini-cases, and in-depth *e-cases* describing a critical quandary that a leader or team must resolve.

Innovations for Successful Societies (ISS), Woodrow Wilson School of Public & International Affairs and the Mamdouha S. Bobst Center for Peace and Justice, Princeton University, Princeton, NJ (http://www.princeton.edu/successfulsocieties/)

ISS cases are used by a wide variety of universities, research institutes, public policy organizations, and NGOs around the world. Its cases, usually examples of best practices, support public servants, policy makers, and scholars seeking to build effective and accountable government in challenging contexts. The cases focus on institution building and government implementation. Started on a pilot basis in 2006, as of 2013 its 100 plus+ cases deal with topics from agency transformation to citizens monitoring public services, and police services.

INSEAD Humanitarian Research Group, Fontainebleau, France (http://www.insead.edu/facultyresearch/centres/isic/humanitarian/CasesHumanitarianResearch.cfm)

Part of a business school, this small collection of 20+ cases from the Humanitarian Research Group deals with the logistics of disaster planning and research. Some cases look at disasters that made global headlines, while others examine issues of planning and response, such as transport, funding, or distributing information online. They cover subjects from organizational behavior to supply chain management, political science, ethics, negotiations, and marketing.

National Center for Case Study Teaching in Science, University at Buffalo, Buffalo, NY (http://sciencecases.lib.buffalo.edu/cs/)

The National Center for Case Study Teaching in Science at the University at Buffalo started in the late 1990s to promote the development and dissemination of materials and practices for case teaching in the sciences. In 2013, nearly 500 cases were available on their Web site for use at the graduate, undergraduate and high-school level. In addition, the center offers a 5-day summer workshop and a 2-day fall conference to train faculty in the CM of teaching science. Support for the center comes from the National Science Foundation, the Pew Charitable Trusts, and the US Department of Education. Science faculties write the cases, which are published after peer review and revision. All have teaching notes.

Pew Case Study Center, Institute for the Study of Diplomacy, Georgetown University, Washington, DC (http://faculty.guisd.org/index.cfm)

This collection of nearly 280 decision-forcing and retrospective cases at the GUISD focuses on international affairs and diplomacy. The cases are used in professional, graduate, and undergraduate courses. Twenty-one of the cases were produced under the auspices of the Carnegie Council on Ethics and International Affairs between 1990 and 2001. Cases can be ordered online; the price of a pdf version is cheaper than that of a hard copy.

School of Public Affairs and Administration Case Collection (SPAA), Rutgers University–Newark, NJ (http://spaa.newark.rutgers.edu/casesimportal/contact/295.html)

SPAA at Rutgers-Newark published cases on public and nonprofit management. Case topic areas include public/nonprofit administration, organizational theory, human resources and diversity, decision making, politics/policy-making, multi-level governance, performance measurement, budgeting and finance, leadership, and ethics. The collection includes fictionalized as well as real-world cases. Many of them are linked to multimedia sources – news, video, social networks, and others.

The Case Centre, UK (formerly ECCH: European Case Clearing House), Cranfield University, UK (www.thecasecentre.org)

The Case Centre is an independent, nonprofit, membership-based organization dedicated to promoting the CM of learning. It maintains a collection of some 49 000 cases on management. While its focus is business cases, it also holds case collections on humanitarian, charitable, political, ethical, and not-for-profit issues. It distributes many large case collections, including those at the Stanford University and Harvard's Kennedy School. It provides multiple resources on CM for instructors and runs training programs.

See also: Associations, Scholarly; Case-Oriented Research; Problem Solving and Reasoning: Case-based; Publishing: Academic and University Presses; Social Justice and Social Inequalities; Social Science Professions and Professionalization; Social Science and Civil Society; Social Science and Universities; Social Science and the Media.

Bibliography

Banning, Kevin C., October 2003. The effect of the case method on tolerance for ambiguity. Journal of Management Education 27 (5), 556–567.

Brynard, P., Erasmus, D. (Eds.), 1995. Public Management and Administration, Case-Study Resource Book. J.L. van Schaik Publishers, Pretoria.

Christensen, C. Roland, Garvin, David A., Sweet, Ann (Eds.), 1991. Education for Judgment. Harvard Business School Press, Boston.

Garvin, D., September–October 2003. Making the case. Harvard Magazine 106 (1). See: http://harvardmagazine.com/2003/09/making-the-case-html.

Gill, T. Grandon, 2011. Informing with the Case Method: A Guide to Case Method Research, Writing and Facilitation. Informing Science Press, Santa Rosa, CA. See: http://www.grandon.com/publications/CaseMethod.pdf.

Gragg, Charles I., 1954. Because wisdom can't be told. In: McNair, Malcolm P. (Ed.), The Case Method at the Harvard Business School. McGraw Hill Book Company, New York, p. 11.

Herreid, Clyde Freeman, May 2004. Can case studies be used to teach critical thinking? Journal of College Science Teaching 33 (6). http://sciencecases.lib.buffalo.edu/cs/pdfs/Can%20Case%20Studies%20Be%20Used%20to%20Teach%20Critical%20Thinking-XXX-6.pdf.

Case, Morphological: Linguistics

Barry J Blake, La Trobe University, Bundoora, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Case, as traditionally understood, refers to a system of inflection, usually in the form of suffixes, which marks the relation of nouns in a clause. In the Latin sentence *Calpurnia Clodiam vīdit* 'Calpurnia saw Clodia' the fact that Clodia rather than Calpurnia is the one who is seen is signaled by the -m suffix on Clodia, rather than by word order as in English. This -m is a case suffix. Cases can be classified as grammatical (signaling a grammatical relation) or semantic (signaling a semantic role). Grammatical cases include the nominative marking the subject, the accusative marking the direct object, the dative marking the indirect object and the genitive marking a dependent of a noun. In some languages the subject of a transitive verb is distinguished by an ergative case and the subject of an intransitive verb and the direct object are treated alike, usually left unmarked in what some call the absolutive case. Semantic cases include the locative indicating location or place where (e.g., 'in,' 'at'), the allative indicating movement toward (e.g., 'to') and the ablative indicating direction from (e.g., 'from'). Case may be marked on the head of a noun phrase and on its dependents via concord, or it may be marked on only one constituent of the noun phrase, usually the last, whether the last member is the head or not.

Case, Case Form, and Case Marker

A morphological case system is a system of inflection that marks dependents for the type of relation they bear to their heads. Typically, case marks the relationship of a noun to a verb at the clause level or of a noun to a preposition, postposition, or another noun at the phrase level. Consider the following Turkish sentence:

(1)	Ali	adam-a	elma-lar-ı	ver-di
	Ali.NOM	man-DAT	apple-PL-ACC	give-PAST.3SG

'Ali gave the apples to the man.'

The suffix -ı indicates that elmalar is the direct object of the verb vermek 'to give' and the suffix -a indicates that Adam is the indirect object. The suffix -ı is said to be an accusative (or objective) case marker and the word form elmaları is said to be in the accusative case. The suffix -ı also indicates that elmalar is specific, since in Turkish only specific direct objects are marked as accusative. The suffix -a is a dative case marker. Ali contrasts with elmaları and adama in that it bears no overt suffix. It is said to be in the nominative case, which in this sentence indicates the subject.

Turkish has a system of six cases as in Table 1. The locative marks location as in Istanbul-da 'in Istanbul,' and the ablative indicates 'from' or 'out of' as in Ankara-dan 'from Ankara.' The genitive is used in phrases like adam-ın ev-i 'the man's house' where ın corresponds to 's in English. The suffix -ı on ev 'house' is a third person possessive form translatable as 'his,' 'her,' or 'its.' In Turkish 'the man's house' is literally 'the man's, his house.'

Table 1 Turkish case system

Nominative	adam
Accusative	adamı
Genitive	adamın
Dative	adama
Locative	adamda
Ablative	adamdan

The word forms displayed in Table 1 make up a paradigm, that is, they constitute the set of case forms in which the lexeme *adam* can appear. In Turkish one could say that there is only one paradigm in that a single set of endings is found for all nouns, although these suffixes do undergo some phonologically conditioned changes such as undergoing vowel harmony and the like. The locative, for instance, has the form-da following stems with back vowels and -de following stems with front vowels. A distinction needs to be made between cases (of which there are six in a system of oppositions), and the case markers or case forms through which the cases are realized. A case marker is an affix and a case form is a complete word. In Turkish the case affixes can be separated from the stem, so it is possible to talk about case markers. In some languages, however, it is not possible to isolate a case suffix, so it is necessary to talk in terms of the various word forms that express the cases of the stem. These are case forms.

Case Marking Fused with Number and Gender

The traditional notion of cases and the associated terminology were developed on the basis of Ancient Greek and Latin where it is not possible to separate number marking from case marking. This means separate paradigms for the two number categories, singular and plural. Moreover, there are different case/number markers for different stem classes. In Latin five such classes are recognized, or five declensions as they are called. The three major declensions are illustrated in Table 2.

In Latin there is also a three-way gender distinction: masculine, feminine, and neuter. With a few exceptions male creatures are masculine and females feminine, but inanimates are scattered over all three genders (though almost all neuter nouns are inanimate). There is a partial association of form and gender in that ā-stems are almost all feminine and *o*-stems mostly masculine (except for a subclass of neuters represented by bellum in Table 2). This means that there can be fusion of

Table 2 Latin case paradigms

	1	*2*		*3a*	*3b*
	ā-stems	o-stems		cons. stems	i-stems
	Feminine	Masculine	Neuter		
	domina	dominus	bellum	cōnsul	cīvis
	'Mistress'	'Master'	'War'	'Consul'	'Citizen'
Singular					
Nominative	domina	dominus	bellum	cōnsul	cīvis
Vocative	domina	domine	bellum	cōnsul	cīvis
Accusative	dominam	dominum	bellum	cōnsulem	cīvem
Genitive	dominae	dominī	bellī	cōnsulis	cīvīs
Dative	dominae	dominō	bellō	cōnsulī	cīvī
Ablative	dominā	dominō	bellō	cōnsule	cīvī, cīve
Plural					
Nominative	dominae	dominī	bella	cōnsulēs	cīvēs
Vocative	dominae	dominī	bella	cōnsulēs	cīvēs
Accusative	dominās	dominōs	bella	cōnsulēs	cīvīs, cīvēs
Genitive	dominārum	dominōrum	bellōrum	cōnsulum	cīvium
Dative	dominīs	dominīs	bellīs	cōnsulibus	cīvibus
Ablative	dominīs	dominīs	bellīs	cōsulibus	cīvibus

gender, number, and case. The point is illustrated in Table 2 where we have *domina* 'mistress (of a household)' illustrating feminine ā-stems and *dominus* 'master (of a household),' which is based on the same root, representing masculine o-stems. As can be seen from Table 2 the word form *domina* simultaneously represents nominative case, feminine gender, and singular number, *dominum* represents accusative case, masculine gender, and singular number, and similarly with other word forms.

Where there is more than one paradigm, it frequently happens that distinctions made in one paradigm are syncretized or neutralized in others. In the traditional descriptions a case is established wherever there is a distinction for any single class of nominals. Informally one might consider that a case is a row in a set of paradigms like those in Table 2. Dealing in cases ('rows') facilitates the description of the relation of form and function. A particular case can be said to have one or more functions without reference to the details of the forms.

The functions of the cases in Latin, and in Ancient Greek, are important in that they have served as a model for applying case labels in other languages such as Turkish. In Latin the nominative encodes the subject and the accusative the direct object. The dative marks the indirect object of a few three-place verbs such as dare 'to give' and the complement of a few two-place verbs including crēdere 'to believe,' nocēre 'to be harmful to,' and subvenīre 'to help.' The genitive is used mainly to mark noun phrases as dependents of nouns, in particular to mark the possessor: cōnsulis equus 'the consul's horse.'

The ablative in Latin represents the syncretism or merger of three once-distinct cases: the ablative, the locative, and the instrumental so it is not surprising then to find that it expresses source, location, and instrument.

The vocative is anomalous in that it does not mark the relationship of dependents, but is used as a form of address as in Quō vādis, domine? (whither go.2SG lord. VOC) 'Where are you going, master?' However, the vocative belongs with the other cases since its marking belongs to the same system as the unequivocal cases.

Latin uses a system of prepositions in addition to the case system. Some prepositions, including all those expressing destination, such as ad 'to' govern the accusative (ad urbem 'to the city'), while others such as ex 'out of, from' govern the ablative (ex urbe 'from the city').

The Range of Case Systems

Case systems differ in terms of the number of distinctions they make and range from a minimum of two as in some Iranian languages, which have a nominative and an oblique, to figures like 15 in Finnish (see Table 3).

Case systems also vary in terms of the type of distinction they make in a given domain. While many languages are like Latin in having one case for the subject, namely the nominative, and another case for the direct object (the accusative), some languages including Basque, northeast Caucasian languages, and many Australian languages have one case for the subject of a transitive verb, which is called the ergative, and another case to cover the subject of an intransitive predicate and the direct object. This case is generally morphologically unmarked. Some writers call this case the absolutive; others call it the nominative. The following illustration is from the Australian language Kalkutungu where the ergative is realized by the suffix -yu and the absolutive/nominative is unmarked. Note, however, that with the cross-referencing clitics that are attached to the purposive auxiliary a common from represents the subject and a distinct form the object.

(2) Nhutu a=nhurr ingka?
you.NOM PURP=2PL.SUBJ go
'Are you mob going?'

(3) Nhutu-yu a=nhurr nuwa puyu?
you-ERG PURP=2PL.SUBJ see 2DU.NOM
'Are you mob going to see them two?'

(4) Nhutu a=kutu nuwa puyu-yu
you.NOM PURP=2PL.OBJ see 2DU-ERG
'They two want to see you mob.'

In a number of languages, particularly in Australia, accusative and ergative marking are found in the same language. As pointed out by Silverstein (1976), there tends to be some complementarity with ergative marking confined to nouns

Table 3 Finnish local cases

	Location	*Source*	*Destination*
Ø	-na	-ta	-ksi
	essive	partitive	translative
Interior	-ssa	-sta	-(h)Vn, -sVVn
	inessive	elative	illative
	'In'	'From (inside)'	'Into'
Exterior	-lla	-lta	-lle
	adessive	ablative	allative
	'At'	'From (outside)'	'To(wards)'

and accusative to pronouns or to pronouns plus certain classes of noun such as kinship nouns, personal names or in some instances all human nouns. The following paradigms from another Australian language, Yuwaalaray, show ergative marking with nouns and accusative with pronouns (see Table 4).

Every language has to provide for the expression of a large number of local notions. This can be done via adpositions as with English prepositions such as on, to, over, and under. Where a number of local notions are expressed via the case system the number of cases naturally rises. In Finnish, for instance, the notions of location, source, and destination can combine with the notions of interior and exterior to produce a set of nine local cases as illustrated in Table 3. Originally there were two systems (location, destination, and source) and (interior, exterior, and neither), but the markers have partially fused. The functions are no longer entirely local either. The partitive is used for the patient if it represents part of a whole or an indefinite quantity, or if the action is incomplete, or if the polarity of the clause is negative. The translative refers to the endpoint of a transformation as in 'She turned the frog into a prince.'

Other cases commonly encountered include the comitative, expressing accompaniment (I went with them) and instrumental expressing the instrument used to carry out an action (She wiped the screen with a cloth). In the Uralic languages a case called abessive is found. It means 'without,' 'not having.' In Finnish, for instance, rahta-tta is (money-AB) 'moneyless.'

Classifying Cases

The Greeks made a primary distinction between the nominative and the other cases, collectively called the oblique cases. This makes sense when one considers that the nominative was the form used in isolation, whereas the other forms generally required a particular syntactic context. In some languages there is a formal distinction between the nominative and the oblique cases inasmuch as there is a special stem for the obliques. This is the situation in many Dravidian and northeast Caucasian languages, for instance. In most languages the nominative bears no marking as in Turkish (Table 1), but consists of the bare stem. Indo-European case languages are unusual in having a marked nominative in most paradigms (Table 2).

Another distinction that is often made is between grammatical (or syntactic) cases and semantic (or concrete) cases. The grammatical cases are traditionally taken to include the nominative and accusative and often the genitive, and should include the dative and the ergative. The basis for the distinction is often not made clear, but if the criterion for classifying a case as grammatical or syntactic is that it expresses a purely syntactic as opposed to semantic relation, then the dative and ergative should certainly be included since the dative encodes the indirect object and the ergative encodes the 'transitive subject.'

If the distinction between grammatical and semantic cases were to be clear-cut, the grammatical cases would encode only purely syntactic relations and the semantic cases would encode only homogeneous semantic relations such as location or source. However, it is common for a syntactic case to encode a semantic relation or role that lies outside of whatever syntactic relation it expresses. In Latin, for instance, the accusative not only expresses the direct object, it also expresses the semantic role of destination as in Vādō Rōmam 'I am going to Rome.' On the other hand, there are situations where the so-called semantic cases encode a purely syntactic relation. This commonly occurs with the passive where the demoted subject is often expressed via a semantic case. In Latin the demoted subject is encoded in the ablative as in vīsus ā cōnsule 'seen by the consul.'

It is also convenient for the purposes of making certain generalizations to distinguish the cases that encode the complements (including the subject) of typical one-place and two-place transitive verbs. These are the nominative, ergative and accusative, and they are known collectively as the core cases and the others as peripheral cases. In Hindi-Urdu and other Indo-Aryan languages a single case is used for subjects and objects and it is called the direct case as opposed to the oblique case. In some treatments the term local cases is used to cover cases designating notions of place such as location, source, destination or path (Table 3).

Table 4 Yuwaalaray case marking

Nominative	dhayn 'man'	ngali 'we two'
Ergative/instrumental	dhayndu	ngali
Accusative	dhayn	ngalinya
Genitive	dhayn.gu	ngalingu
Dative	dhaynda	ngalingunda
Ablative	dhayndi	ngalingundi

Exponence

The relation between the verb and its dependents, whether complements or adjuncts, may be expressed by an adposition, either a preposition as in English or a postposition as in Turkish. In languages with morphological case systems there are usually adpositions as well as cases. In Latin, as we have seen, there are prepositions, and in Turkish there are postpositions such as sonra 'after,' which governs the ablative: tiyatrodan sonra (theatre-ABL from) 'after the theater.' In such languages there are essentially two systems: an inner, restricted system of case and an outer, more elaborate system of adpositions. In some languages such as Japanese there is just one system of postpositions, which includes markers for subject, direct object, indirect object, and topic, as well as for local notions such as 'from.' In such a language the distinction between case and adposition is essentially neutralized. In the following Japanese example ga marks the subject, ni marks the indirect object, and o marks the direct object.

(5) Sensei ga Tasaku ni hon o yat-ta
teacher SUBJ Tasaku IO book DO give-PAST
'The teacher gave Tasaku a book.'

Morphological case is almost always realized as a set of suffixes, though there are instances where there is some

mutation of the stem as with Old English fēt, the dative singular (as well as nominative and accusative plural) of fōt 'foot,' tone, as in some western Bantu languages, or complete suppletion as with the English pronouns: I/me, she/her, we/us, and also he/him and they/'em/them where there is an -m suffix (historically dative) plus some stem alternation.

Where categories such as number, gender or possessor are marked as well as case and where these categories are not fused, case marking normally follows these other categories as in Turkish: adam-lar-ım-la (man-PL-1SG.POSS-LOC) 'with my men' where the number marking and possessor precede the case marking.

Within the noun phrase there are two common distributions for case marking. Either it occurs on all types of nominal or it is occurs on only one word. In the former type the distribution is traditionally described in terms of dependents showing case concord or agreement with a noun head. In the following example from Ancient Greek bios is a nominative singular form of a second declension (o-stem) masculine noun, the nominative indicating that bios is the subject of the predicate. The definite article and the adjective are in the nominative singular masculine form, their concord in case, number and gender indicating that they are dependents of bios.

(6) Ho aneksetastos bios ou biōtos anthrōpō
the.NOM.SG unexamined.NOM.SG life.NOM.SG not livable.NOM.SG man.DAT.SG
'The unexamined life is not livable for man.'

This example also illustrates concord between a predicative adjective (biōtos) and the subject (bios). While concord of case, number, and gender of predicative adjectives is the norm in Latin and Ancient Greek [GJ1], as well as in some modern languages such as Slavic languages (some of which have 7 cases), in many other languages predicative adjectives display no concord. This is the situation in German where the predicative adjective consists of the bare stem and is distinct from the various case forms.

Full case concord within the noun phrase is found in the older Indo-European languages and in some modern branches including Romance and Slavic. In Polish, for instance, *ta piękna biała róża*, 'this beautiful white rose' NOM versus *tej pięknej białej róży* 'of this beautiful white rose' GEN. It is also found in Balto-Finnic, Slavic, Semitic [GJ2], in some Australian languages, and elsewhere, but it is a minority system. In some languages such as Hungarian the concord does not extend to the adjective and in the Indo-Aryan languages adjectives exhibit fewer case distinctions than nouns. In the more conservative Germanic languages, including modern German, the adjective shows fewer case marking distinctions where a determiner is present. In most languages case concord does not extend to dependents in the genitive case. However, in Georgian a postposed genitive exhibits concord, which results in double case marking: compare davit-is mama-s (David-GEN father-DAT) and mama-s davit-isa-s (father-DAT David-GEN-DAT 'to David's father'). Where case marking occurs on only one word in the noun phrase, it is usually on the final word. In many languages, including a vast number in northern Asia and South Asia, the final word is the head noun in the noun phrase. The following example is from the Dravidian language Kannada:

(7) Naanu ellaa maanava janaangavannu priitisutteene
I.NOM all human community.acc love.1SG
'I love all mankind.'

In some languages the final word in the noun phrase is not always the head. This is the situation in various Australian and Amazonian languages, and in Basque: etxe zaharr-etan (house old-PL.LOC) 'in old houses.'

History

Since bound forms in general derive from free forms, it is not surprising that there is evidence of case markers deriving from nouns and verbs, often via postpositions. For instance, the Hungarian illative ('into') case marker -ba as in világ-ba 'into the world' derives from a noun bél 'innards' and the allative case marker-tharri in the Australian language Kunggari derives from a verb tharri 'to become.'

Case marking may also be lost. This is quite striking in Europe, particularly among the Romance languages. The Latin system of noun inflection illustrated in Table 2 has been lost from the Romance languages except for a reduced system retained in Rumanian. Modern Greek and the Germanic languages too have lost case distinctions over the last two millennia, with English and Afrikaans having lost all case marking on nouns, though German still retains a four-case system: nominative, accusative, genitive, and dative and Modern Greek a three-way system (nominative, accusative and genitive, dative and genitive having merged in the language). In languages like English and the Romance languages case distinctions have been retained on pronouns. Case marking can fall prey to phonetic attrition and the availability of prepositions and word order as alternative means of making the relevant distinctions must also be a factor in the loss of case.

See also: Gender: Grammatical; Morphology in Linguistics; Number, Grammatical.

Bibliography

Blake, Barry J., 1994, 2001. Case. Cambridge University Press, Cambridge, UK.
Blake, B.J., 2007a. Playing with Words: Humour in the English Language. Equinox, London.
Blake, B.J., 2007b. All about Language. OUP, Oxford.
Blake, B.J., 2010. Secret Language. OUP, Oxford.
Breen, J.G., Blake, B.J., 2006. A Grammar of Yalarnnga. Pacific Linguistics, Canberra.
Brecht, R.D., Levine, J.S. (Eds.), 1986. Case in Slavic. Slavica Publishers, Columbus, OH.
Butt, Miriam, 2006. Theories of Case. Cambridge University Press, Cambridge, UK.
Comrie, Bernard, 1986. On delimiting cases. In: Brecht, Richard D., Levine, James S. (Eds.), Case in Slavic. Slavica Publishers, Columbus, OH, pp. 86–105.
König, Christa, 2008. Case in Africa. Oxford University Press, Oxford and New York.
Mel'čuk, Igor A., 1986. Toward a definition of case. In: Brecht, R.D., Levine, J.S. (Eds.), Case in Slavic. Slavica Publishers, Columbus, OH, pp. 35–85.
Malchukov, Andrej, Spencer, Andrew (Eds.), 2009. The Oxford Handbook of Case. Oxford University Press, Oxford.
Silverstein, Michael, 1976. Hierarchy of features and ergativity. In: Robert, M., Dixon, W. (Eds.), Grammatical Categories in Australian Languages. Australian Institute of Aboriginal Studies, Canberra, Humanities Press, New Jersey, pp. 112–171.

Case-Oriented Research

Charles C Ragin, University of California, Irvine, CA, USA; and University of Southern Denmark, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Case-oriented research focuses on interconnections among parts and aspects within single cases. In the case-oriented approach, the researcher attempts to make sense of each case as a singular, interpretable entity. In-depth knowledge of the cases included in a study is usually considered a prerequisite for the examination of patterns observed across cases. Case-oriented researchers often study one case at a time, but they also may study multiple instances of a given phenomenon (e.g., comparable instances of ethnic conflict). The distinctiveness of case-oriented research is apparent when this approach is contrasted with the variable-oriented approach, where researchers focus almost exclusively on cross-case patterns, without first gaining case-based knowledge.

The Goals of Case-Oriented Research

Today, social scientists tend to identify case-oriented research with specific techniques of data collection linked to the observation and analysis of singular cases (e.g., direct observation of individuals at the micro level and archival research on nation-states at the macro level). While generally useful, the identity of case-oriented research with specific techniques of data collection is unfortunate, for it obscures basic differences between case-oriented research and conventional variable-oriented research. More fundamental than differences in methods of data collection is the contrast between goals (Ragin, 1987). Case-oriented strategies are distinctive in that they are centrally concerned with making sense of a relatively small number of cases, selected because they are substantively or theoretically significant in some way (e.g., Eckstein, 1975; Gerring, 2006; Mahoney and Goertz, 2004; George and Bennett, 2005). Conventional variable-oriented strategies, by contrast, are centrally concerned with the problem of assessing the relationship between aspects of cases across a sample of generic 'observations', often with the goal of inferring general patterns that hold for the larger population.

For example, a researcher might use a case-oriented approach in order to study a small number of firms in an in-depth manner. Suppose these firms are all thought to be unusually successful in retaining their best employees while at the same time investing in them and thus enhancing their potential value to competing firms. How do these firms do it? Do they all do it in the same way? Clearly, to answer these questions a researcher would need to conduct an in-depth study of the firms in question, focusing on how they accomplish this outcome. By contrast, a variable-oriented researcher might study the predictors of variation in rates of 'employee retention' across a large sample of firms. Is it more a matter of firm or industry characteristics? Do these two sets of factors interact? Useful answers to these questions would be based on careful analysis of relationships between variables, using data drawn from a survey of a large number of firms – the more firms (and the more varied they are), the better.

As these two examples illustrate, what matters most is the researcher's starting point, and here the issue is very simple: Does the researcher seek to understand specific cases or to document general patterns characterizing a population? This contrast follows a long-standing division in all of science, not just social science. von Wright (1971) argues in *Explanation and Understanding* that there are two main traditions in the history of ideas regarding the conditions an explanation must satisfy in order to be considered scientifically respectable. One tradition he calls 'finalistic' and is anchored in the problem of making facts understandable. The other is called 'causal-mechanistic' and is anchored in the problem of prediction. The contrast between case-oriented and variable-oriented research closely parallels this fundamental division. In the two examples just described, the first researcher uses the case-oriented approach in order to make certain facts understandable – the spectacular success of a handful firms in retaining their most valuable employees; and the second researcher uses the variable-oriented approach in order to derive an equation predicting levels of retention, based on a large sample of firms, and to draw inferences from this analysis to an entire population.

Researchers may select specific cases for in-depth study – seeking to make them 'understandable' – for a variety of different reasons (Stake, 1995; Gerring, 2006; George and Bennett, 2005). For example, cases might be chosen because they are extreme in some way and thus present a process or phenomenon in relatively pure form (e.g., Dumont, 1980). Cases are sometimes selected for in-depth study for the opposite reason – because they are typical or run-of-the-mill (e.g., Becker et al., 1961). Researchers select 'typical' cases to learn more about conventional arrangements and practices. Another justification for studying specific cases in an in-depth manner is because they are historically or culturally significant in some way (e.g., Lipset, 1963). For example, some cases may be pioneers, establishing standards and routines imitated by others (Meyer and Strang, 1993). Cases also may be selected for explicitly theoretical reasons – because they challenge a widely held theory or because they support a widely questioned theory (Eckstein, 1975). Very often case-oriented methods are used because the phenomena that researchers wish to study are too complex, context bound, or context sensitive to be studied in any other way (Yin, 1994). Such phenomena require a more encompassing approach, and the researcher must triangulate different types of evidences using different methods. It is also important to note that to make sense of cases researchers

sometimes must collect evidence relevant to the motives and intentions of actors, which, in turn, mandates the use of intensive methods. In short, whenever there is a compelling reason to make the facts of a case or a small number of related cases 'understandable,' a case-oriented research strategy is essential.

Once the distinction between case-oriented and variable-oriented research is established and their contrasting goals acknowledged, it is clear that the importance of techniques of data collection as bearers of the 'case-oriented versus variable-oriented' distinction fades. For example, it is clear that a researcher using case-oriented methods to study a handful of firms might benefit from conducting surveys of their employees and performing a conventional variable-oriented analysis of these data. The results of the survey would contribute to this researcher's depth of knowledge about the firms in question, just as interviewing their top executives or studying their archives would elicit useful information. Likewise, it is clear that the researcher using variable-oriented methods to predict rates of retention could benefit from interviews of top executives or personnel officers to help interpret results found in the analysis of the predictors of retention. Still, the first researcher would remain focused on the problem of understanding the handful of firms in question, while the second would remain focused on the problem of explaining variation in retention rates across a large number of firms and making inferences from this sample to a population. The important point here is that data collection techniques per se are relatively neutral; what matters most are the researcher's goal and the contrasting research strategies that follow from different goals.

The Logic of Case-Oriented Research

The logic of case-oriented research is fundamentally configurational. Different parts of each case are understood in relation to one another and in terms of the total picture or package that they form together. The organizing idea in such research is that the parts of a case may constitute a coherent whole – that they have an integrity and coherence considered together. For example, researchers who study families often find that patterns of interpersonal accommodation are so enmeshed that a 'dysfunction' cannot be remedied without addressing many different aspects of a family system all at once. Likewise, researchers who study cultures often observe that cultural traits seem to come in packages that defy easy disassembly. The 'parts' that case-oriented researchers examine can be quite varied. In macro-level research, for example, the 'parts' of a case might include institutions, path dependencies, social structures, historical patterns and trends, routine practices, cultural beliefs, singular events, event sequences, connections to other cases, the case's larger environment, and so on. What matters most is that the investigator makes sense of multiple aspects of the case in an encompassing manner, using theory as a guide.

Campbell (1975) offers a rough formalization of this approach in his examination of the logic of case study research. Campbell argues that at first glance the in-depth study of a single case appears to be lacking in scientific merit because there is only one case to explain and many possible explanations to choose from. He notes, however, that case study researchers routinely 'reject' theories because they do not explain the facts of their cases. Furthermore, despite having only one case, they often must struggle to find theories that work. Why is it so difficult? The key to this puzzle is the simple fact that every theory has many implications, relevant both to features of the case in question and to causal processes and sequences operating within the case. Thus, researchers evaluate many theoretical implications relevant to their cases to see if their cases conform to expectations. Not all features of a case may be compatible with an initial theory, and the researcher must either find an alternate theory that works better, revise an existing theory, or propose an entirely new one (Walton, 1992).

Campbell (1975) suggests that each separate theoretical implication can be seen as a different 'observation' for 'testing' a theory. Thus, a single case becomes many observations – some contradicting and some supporting competing theories. Collectively, competing theories and their different implications define all theoretically relevant aspects of the case in question. The researcher's task is to see which theory does the best job of explaining aspects of the case relevant not only to its own implications, but also to the implications of competing theories. Thus, after defining and selecting relevant aspects of the case, the researcher assesses the explanatory power of each theory. The theory (or combination of compatible theories) that best covers both its own implications and those of competing theories prevails and provides the basis for the investigator's representation of the case.

In the end, the researcher crafts an explanation that satisfies as many theoretical implications as possible in a coherent manner. The success of a case study hinges on (1) the number of relevant aspects of the case the researcher can encompass with his or her explanation, (2) the success of the researcher in showing that his or her portrait of the case actually makes sense of all theoretically relevant aspects, and (3) the agreement of other scholars that all relevant aspects of the case in question have in fact been addressed by the researcher in a convincing manner. Of course, the most successful case studies accomplish much more than theoretical and substantive coherence. They also may advance theory or establish important lessons for policy makers (White, 1992). Still, theoretical and substantive coherence should be considered preconditions for these more ambitious goals (Yin, 1994).

Campbell's (1975) argument that case-oriented researchers test multiple theoretical implications for each case underscores the configurational nature of this type of research. Because theoretical implications direct the researcher's attention to 'observations' (i.e., different aspects of a single case) that cannot be independent from one another, the researcher must make sense of them all at once, as a package. Furthermore, because different theories typically have implications of different aspects of the case, the researcher's attention is directed to a broad range of aspects, all of which may be interconnected in some way. Thus, case-oriented researchers must examine overlapping configurations of aspects as they weigh the explanatory power of competing theories.

Case-Oriented Research on Multiple Instances

Case-oriented researchers sometimes conduct a series of case studies, selecting cases strategically so that the knowledge that

these cases generate may accumulate. For example, a specific type of evidence that is not available in one case may prompt the investigator to select for his or her next study a case that offers this evidence. In this manner, case-oriented researchers may develop theoretical arguments in much the same way that psychoanalysts build theory from the analysis of a series of patients, admitting to their practice patients who promise to enhance their thinking and their theories in some way.

This case-by-case, grounded approach can be contrasted with a research strategy that takes *multiple instances* of a phenomenon as its starting point. For example, a researcher might be interested in the formation of green parties in Western Europe and focus on a handful of parallel cases. The key contrast between the study of multiple instances and studying a series of cases is the *timing* of the selection of cases. In serial case studies, findings from one case determine the selection of the next case; in the study of multiple instances, by contrast, the researcher identifies multiple instances of a phenomenon at the outset of the investigation. Of course, the researcher's definition of the set of relevant cases may shift in the course of the research as the investigator learns more about the phenomenon (see Ragin, 1997, 2000). The key point is that the study of multiple instances begins with a preliminary specification of the set of relevant cases.

Generally, when researchers study multiple instances, they focus on a qualitative outcome of some sort, a change that occurred over a specific period of time, signaling some sort of shift or transformation relevant to the case as a whole (e.g., Skocpol, 1992; Morgan and Prasad, 2009; Stryker, 1990; Bloemraad, 2006; Chen and Weir, 2009; Fourcade-Gourinchas and Babb, 2002; Riley, 2005; Steinmetz, 2008; Mahoney, 2010; Swenson, 2002). Accounts and explanations of these qualitative changes typically invoke convergent causal conditions – the specific combination of causally relevant conditions that prompted the change. In many respects, these combinations can be seen as causal recipes (Ragin, 2008) because all the relevant ingredients for change must be in place for it to occur. Specifying a causal recipe entails not only listing the relevant ingredients, but also identifying the processes and mechanisms involved in the production of the outcome (Steel, 2004; Bennett, 2008; George and Bennett, 2005). Even as lists of combined conditions, causal recipes are pregnant with implications about processes and mechanisms (see, e.g., Boswell and Brown, 1999).

Depending on the goals of the study, the specification of a causal recipe for a chosen case may be considered a hypothesis relevant to other cases. In this way, the endpoint of a case study may be a testable hypothesis about other cases, especially when this endpoint is a more or less 'portable' causal recipe linked to a general outcome – a qualitative change experienced by other cases as well. The case is viewed not just as an empirical event but also as a token of a type of empirical phenomenon which is analyzed in order to develop a general hypothesis that explains events of that type.

Consider, for example, a researcher conducting a study of protest against the International Monetary Fund (IMF) in Peru. Through analysis of a variety of case materials – newspaper articles, interviews, secondary sources, published surveys, and so on – the researcher might conclude that these protests occurred because of (1) the severity of austerity measures imposed by the IMF as a condition for the restructuring of Peru's debt repayment, (2) the high concentration of the poor in urban slums, (3) the perception of widespread corruption on the part of government leaders, and (4) the substantial prior level of political mobilization and contention. This explanation of IMF protest cites a specific combination of conditions and has the character of a recipe: all four conditions were met simultaneously in the case of Peru, and together they explain the dramatic turn to protest.

Using the analysis of Peru as a springboard, a researcher could move in either of two directions. The first direction would be to find other instances of protest against IMF-imposed austerity and examine the extent to which these cases agree in displaying the same recipe, the same combination of four causal ingredients found in Peru. Do all (or virtually all) instances of IMF protest display these four antecedent conditions? This strategy employs the common qualitative research strategy of 'selecting on the dependent variable,' an approach that is almost universally, but mistakenly, condemned by quantitative researchers (see, e.g., King et al., 1994; Geddes, 1990; for a defense, see Dion, 1998; Ragin, 2000, 2008). The second direction would be to try to find other instances of Peru's recipe and examine whether these countries also experienced IMF protest. In essence, the researcher would select cases on the basis of their scores on the independent variable. In this example, however, the 'independent variable' is a causal recipe with its four main conditions all satisfied. The goal of the second strategy would be to assess the sufficiency of the recipe: Does it invariably (or at least with substantial consistency) lead to protest?

Both of these strategies are set theoretic in nature and conform to the two general set theoretic approaches described in Ragin and Rihoux (2004) and Ragin (2008). The first is an examination of whether instances of the outcome (protest against IMF-mandated austerity) constitute a subset of instances of a combination of causal conditions (i.e., Peru's recipe). The researcher examines other instances of IMF protest to see if the same combination of four ingredients existed in these other cases. This demonstration would establish that the four causal conditions in question are consistent with an argument of necessity. The second is an examination of whether instances of a specific combination of causal conditions (Peru's recipe) constitute a subset of instances of an outcome (IMF protest). The researcher would identify countries that experienced the same intersection of these four conditions, to see if they exhibited IMF protest as a consequence. This demonstration would establish that the combination of causal conditions is consistent with an argument of sufficiency. Of course, both strategies could be used, and if both subset relations were confirmed, then the two sets (the set of cases with Peru's causal recipe and the set of cases with IMF protest) would coincide. While it might appear that the two strategies together constitute a correlational analysis, it is important to recognize that correlations are strong only if there are many 'null–null' instances – cases that lack both the causal recipe and the outcome. Neither of the two research strategies just described depends on or uses such cases in any direct way. It is also important to understand that from a case-oriented viewpoint, these two tasks – the assessment of sufficiency and the assessment of necessity – are distinct and should not be

Table 1 Case-oriented research: assessing necessity versus sufficiency

	Causal recipe absent	*Causal recipe present*
Outcome present	Cell 1: cases in this cell challenge the argument of necessity but are not directly relevant to the assessment of sufficiency.	Cell 2: cases in this cell may confirm either necessity or sufficiency, depending on the researcher's goals.
Outcome absent	Cell 3: cases in this cell are not directly relevant to the assessment of either necessity or sufficiency.	Cell 4: cases in this cell challenge the argument of sufficiency but are not directly relevant to the assessment of necessity.

conflated, as they are in a conventional variable-oriented analysis based on symmetric measures of association.

Table 1 summarizes the case-oriented approach to assessing the connections between conditions and outcomes. It is important to understand that different ideas about the causal recipe in question motivate different research strategies, especially with regard to case selection. If the researcher suspects that the conditions identified in the recipe are necessary but not sufficient, then the next analytic step is to examine other instances of the outcome (i.e., select on the dependent variable). Specifically, the researcher seeks to establish that instances of the outcome reside in cell 2 of Table 1, with no cases in cell 1. Cases in cell 4 do not challenge the argument of necessity. By contrast, if the researcher suspects that the recipe is sufficient but not necessary, then the next analytic step is to examine other instances of the causal combination. Specifically, the researcher seeks to establish that instances of the causal combination reside in cell 2 of Table 1, with no cases in cell 4. Cases in cell 1 do not challenge the argument of sufficiency.

Of course, it is common for a hypothesis based on one case to be disconfirmed by other cases. For example, suppose that the 'true' recipe for IMF protest has five ingredients, not four, indicating that the researcher overlooked an important ingredient in the case of Peru. The results of the sufficiency analysis of the four-condition recipe would reveal cases in cell 4, which in turn would challenge the argument of sufficiency. The cell 4 cases would have only four of five causal conditions in the 'true' recipe and thus would not exhibit the outcome. Alternatively, suppose that the recipe based on the Peruvian case is needlessly complex and should have only three causal ingredients, not four. The results of a necessity analysis of the four-condition recipe would uncover cases in cell 1.

The key point is that in case-oriented research, the standard 2×2 table of outcome cross-tabulated against cause (which in case-oriented research is usually a causal recipe, not a single independent variable) is disaggregated and partitioned because the different cells of this table have different interpretations and different analytic uses. This approach contrasts sharply with conventional variable-oriented research, where the usual endpoint of the analysis of a 2×2 table (or its interval-level equivalent, a scatterplot) is a global, symmetric measure of association embracing all cases at once. It is also important to emphasize that these symmetric statistics reward researchers for having as many cases as possible in cell 3 (the null–null cell). By contrast, cell 3 plays no direct role in the case-oriented assessment of either the sufficiency or necessity of a causal condition or a combination of causal conditions.

Incongruities Between Case-Oriented and Variable-Oriented Research

Case-oriented research is distinct from variable-oriented research in many ways. The distinctiveness of the case-oriented approach is often overlooked, and researchers using different strategies talk past each other when discussing their methods. While some scholars emphasize philosophical differences between case-oriented and variable-oriented strategies, there are striking and profound differences between the two approaches at the practical level. The practical level refers to the relatively mundane procedures researchers use when working with evidence to produce social scientific representations of what they have learned – that is, summaries of findings and patterns.

The importance of practical differences as a basis for miscommunication between case-oriented and variable-oriented researchers can be seen clearly in the contrasts between two very simple procedures: the search for causal commonalities across multiple instances of an outcome in case-oriented research and the examination of the correlation between two variables across a sample of observation in variable-oriented research. The search for commonalities is usually part of an effort to derive limited generalizations from the in-depth study of a modest number of instances. A researcher might attempt to identify common antecedent conditions for an outcome these cases share. This researcher might argue further that the shared antecedents are necessary conditions for the outcome – that is, if theoretical and substantive knowledge support this interpretation.

Consider the following contrast: a researcher conducting an in-depth study of a select group of firms – those that are able to both invest in and retain their best employees – might identify causally relevant features shared by these firms. Shared features, in turn, might point to a possible formula for success and offer lessons for other firms. The search for commonalities across a select group of cases is the most basic case-oriented strategy applied to multiple instances. By contrast, the examination of a correlation between two variables across a sample of observations in variable-oriented research is usually conducted with the goal of establishing that two conditions or aspects covary. For example, high rates of retention might be correlated with high levels of employee compensation across a sample of firms. This correlation, in turn, might be interpreted causally: the researcher would argue that one reason firms differ in their employee retention rates is that they differ in how well they compensate employees.

While both procedures seem deceptively simple and straightforward, these two ways of producing results from evidence involve sharply contrasting practical orientations toward cases, outcomes, and causal conditions.

Cases

When a researcher using variable-oriented methods computes a correlation between two variables, the relevant cases become more or less invisible and variables take center stage. Furthermore, the set of cases included in the computation must be fixed before the researcher can compute the correlation. Once this set is fixed, usually at the outset of an investigation, it is rarely altered. What matters most is that the cases (which are seen as 'observations') belong to the same general 'population' and that they are drawn from this population with an eye toward randomness or representativeness or some combination of these criteria.

In a case-oriented study of commonalities, by contrast, cases have clear identities and are usually chosen specifically because of their substantive significance or theoretical relevance. Furthermore, the set of relevant cases may shift during the investigation because the researcher may decide that one or more cases do not 'fit' with the others. For example, a researcher studying 'firms that invest in and retain their top employees' might decide that several firms originally thought to belong to this group really do not belong, and that maybe one or more that were thought to be outside this group actually belong in it. This flexibility is maintained throughout the investigation because the core concepts (e.g., 'investing in employees') may be revised as the researcher learns more about relevant instances.

Outcomes

In correlational studies, researchers usually identify a dependent variable – an outcome that varies across cases. Typically, such outcomes are aspects of cases that vary by level, for example, level of satisfaction, level of bureaucratization, and so on. Sometimes the outcome variable is categorical, indicating whether or not some event has occurred (e.g., filing a complaint), and sometimes it is a frequency or a rate (e.g., rate of employee retention). The important point is that the outcome must vary across observations. The goal of the research typically is to explain, if possible, why cases have different values on the dependent variable. Such research is centrally concerned with the question of 'why.' For example, a researcher might seek to explain why some employees are more satisfied than average and others less so, why some firms have higher retention rates than average and others less than average, and so on.

In a case-oriented study of commonalities, by contrast, the outcome is often something that does not vary substantially across the chosen cases. In a study of firms, for example, cases might be chosen precisely because they all display the same outcome – a specific pattern of successful retention. Recall that the goal of a case-oriented study of commonalities is to identify common causal conditions linked to a specific outcome across a relatively small number of purposefully selected cases. Thus, the focus is on cases with a specific outcome, not cases that vary widely in how much they display this outcome. While the outcomes in a study of this type will not be exactly identical across cases, the researcher must demonstrate that the outcomes in the cases selected are in fact enough alike to be treated as instances of the same thing. Finally, unlike correlational studies, which are centrally concerned with the question of 'why' (as in: Why some more than others?), case-oriented studies are centrally concerned with the question of 'how' (as in: How does it happen?). How do firms succeed in retaining their most valuable employees?

Causal Conditions

In a correlational study, causation may be inferred from a pattern of covariation. If a variable thought to represent a cause or to be an indicator of a key causal condition is strongly correlated with the outcome variable, then the researcher may make a causal inference. Usually, the researcher will assess the relative strength of several causal variables at the same time. The typical goal is either to find out which one explains the most variation in the outcome variable or simply to assess the relative importance of the different independent variables. In effect, variables compete with each other to explain variation in the dependent variable. In most investigations, each causal variable is considered sufficient, by itself, for the outcome or an increment in the outcome. That is, each one is considered an 'independent' variable capable of affecting the outcome variable, regardless of the values of other causal variables.

In a case-oriented study of commonalities, by contrast, causation is typically understood conjuncturally. The goal of this type of analysis is to identify the main causal conditions shared by relevant cases. Causal conditions do not compete with each other, as they do in correlational research; often, they combine. How they combine or 'fit together' is something that researchers try to discern using their in-depth knowledge of cases. Because all the cases have more or less the same outcome, the usual reasoning is that the causally relevant conditions shared by cases provide important clues regarding which factors must be present to produce the outcome in question. When constructing this argument, the researcher is especially sensitive to the possibility that a given causal requirement (i.e., a necessary condition) may be met in a variety of different ways.

The study of the commonalities shared by a set of similar instances is not always explicitly concerned with causal conditions. Instead, the commonalities identified by researchers help them deepen their knowledge of the research subject and thus lay the groundwork for the construction of a 'composite portrait' (see Ragin and Amoroso, 2011). The researcher works back and forth between ideas and evidence to refine the portrait (Katz, 1982). While not explicitly causal, procedurally the method is the same: researchers make sense of the 'facts' of their cases by identifying and interpreting commonalities. Case-oriented researchers often shy away from causal statements because they have been warned repeatedly that case studies are not good tools for making causal inference (e.g., Campbell and Stanley, 1966). However, even the construction of composite portraits based on commonalities results in representations of cases that are full of causal implications.

These and other practical differences in how researchers using case-oriented versus variable-oriented methods work with evidence to produce results provide many opportunities

for disjunctures in findings. For example, from the perspective of variable-oriented work, the study of commonalities across a small number of instances is fraught with analytic sins and errors. (1) The number of cases is too small and too nonrandom to warrant any kind of inference (King et al., 1994). (2) The procedure 'selects on the dependent variable' (i.e., focuses on cases with more or less the same value on the outcome variable). This practice may deflate otherwise robust correlations (Collier, 1995; Collier and Mahoney, 1996; King et al., 1994). (3) Researchers may drop cases that 'do not fit' at various stages of the analysis, which seriously undermines any effort to generalize beyond the cases that remain (King et al., 1994). (4) The most important causal factors (i.e., causally relevant commonalities) do not vary and thus are impossible to assess, and so on (Lieberson, 1991, 1994; Goldthorpe, 1991, 1997).

Likewise, from the perspective of case-oriented work, the examination of the correlation between a causal and outcome variable across many cases is fundamentally flawed. (1) Typically, there are so many cases that there is no way for the researcher to know if they are all really comparable and thus belong together in the same analysis. (2) Fixing a population or sample boundary also fixes the assumption of homogeneity, which usually is not warranted. Investigators should be able to redefine the set of relevant cases as they learn more about them (Ragin, 1997). (3) It is difficult to determine how something 'comes about' by comparing cases with different levels of the outcome. The partial instances (i.e., those with lower scores on the outcome variable) are likely to provide many false leads (Lijphart, 1971, 1975; Ragin, 1997). (4) It is pointless to try to isolate the 'independent' effect of any causal condition when several factors usually must combine for a particular outcome to occur (George, 1979), and so on. In short, at a practical level the two approaches seem almost antithetical. It is no wonder that findings diverge and researchers often talk past each other (Rueschemeyer, 1991; Ragin, 1997).

Bridging Case-Oriented and Variable-Oriented Research

While the gulf between case-oriented and variable-oriented research seems great, it is possible to span it. Consider the following scenario: a researcher using case-oriented methods studies several instances of an outcome (e.g., a small number of firms that are very successful in investing in and retaining top employees), documents causally relevant commonalities, and then constructs a general, composite argument about how these firms do it. This argument leads to four specific recommendations (which might be labeled X_1 to X_4) based on the observed commonalities. A second researcher reads the report of this study and decides to evaluate it with a large sample using variable-oriented methods. This researcher collects information on a random sample of firms and finds that independent variables X_1 to X_4 do not distinguish more successful from less successful firms, using various measures of employee retention. In short, the second researcher shows that there is no statistically significant difference in the retention rates for firms with and without these four aspects, considering these aspects one at a time or in an additive, multivariate equation.

What went wrong? Usually, the researcher using variable-oriented methods will claim that the first researcher's 'sample' was 'too small' and 'unrepresentative.' Thus, the identification of X_1 to X_4 took advantage of specific aspects of the selected cases. The first researcher might counterattack by arguing that causally relevant commonalities identified through in-depth study are very difficult to represent as 'variables,' and that the second researcher's crude attempt to operationalize them fell far short. Indeed, the first researcher might argue that it would take in-depth knowledge of each firm included in the variable-oriented study to capture these conditions appropriately and contextually.

These criticisms and countercriticisms are quite common. However, the incongruity between the two hypothetical studies can be resolved without resorting to derogation. The first researcher in this example – the one using case-oriented methods – selected on instances of the outcome and identified four causally relevant conditions shared by the firms in question. In essence, this researcher worked backward from the outcome to causes and thus identified potential necessary conditions for the outcome (Ragin, 2000). Are these conditions truly necessary? In part, this is an empirical question. To gain confidence, the researcher should examine as many instances of the outcome as possible, to see if they agree in displaying these four causally relevant conditions (or their causal equivalents). But it is also a question about existing knowledge. Is the argument that these four conditions are necessary consistent with theoretical and substantive knowledge? Do they make sense as necessary conditions? If the researcher's finding is consistent with existing substantive and theoretical knowledge, then the argument that these four conditions are necessary is strengthened.

How should the second researcher – the one using variable-oriented methods – respond to the argument that these four factors are necessary conditions? At a more abstract level, the specification of necessary conditions is relevant primarily to the identification of cases which are candidates for an outcome. Cases cannot be candidates for an outcome if they do not meet the necessary conditions. But many cases may meet the necessary conditions for an outcome and still not exhibit the outcome because they lack additional conditions which, when combined with the necessary conditions, establish sufficiency for the outcome. In fact, the cases displaying the outcome may be only a small minority of those that meet the necessary conditions. Thus, while there are clear gains from identifying necessary conditions, as in the hypothetical case-oriented study, the identification of causally relevant commonalities shared by instances of an outcome does not establish the conditions that are sufficient for an outcome. Thus, the variable-oriented researcher's finding that these four conditions do not distinguish low-retention firms from high-retention firms across a large sample of firms does not directly challenge the argument that these conditions are necessary.

The variable-oriented analysis of these four conditions across a large sample of firms is much more directly relevant to their sufficiency. To argue that these conditions are jointly sufficient for the outcome, it would be important to show that when these four conditions are combined the outcome follows. In other words, the variable-oriented researcher could evaluate sufficiency by examining the correspondence between the combination of the four causes, on the one hand, and the outcome, on

the other, across a large sample of firms. Still, this analysis would be an evaluation of sufficiency, not necessity, and the results of this analysis would not bear in a direct manner on the claim of necessity made by the case-oriented researcher.

Notice that the resolution of the disjunction described here provides a framework for integrating the two strategies. The results of case-oriented research on shared antecedent conditions may establish potential necessary conditions. Assuming that theoretical and substantive knowledge support the idea that these conditions are necessary, the next task is to verify this, moving from a small number of positive instances of the outcome to many. If the extension of the initial case-oriented results to many cases holds, then the next step is to address the sufficiency of these conditions. Researchers should select instances that display these conditions to see if they agree in displaying the outcome. If they do not, then in-depth study of disconfirming cases should help researchers identify relevant sufficient conditions.

This sketch identifies only one of several ways to integrate case-oriented and variable-oriented inquiries. The general and most important point is that it is possible to integrate these two approaches if researchers are careful to distinguish between necessity and sufficiency and separate the analysis of these two aspects. More generally, this discussion underscores the distinctiveness of case-oriented research, especially the case-oriented study of multiple instances of an outcome.

See also: Case Study: Logic; Case Study: Methods and Analysis; Comparative Method in Anthropology; Comparative Studies: Method and Design; Ethnography; Explanation: Conceptions in the Social Sciences; Human–Environment Interactions: Case Studies; Triangulation: Methodology.

Bibliography

Becker, H.S., Greer, B., Hughes, E., Strauss, A., 1961. Boys in White: Student Culture in Medical School. University of Chicago, Chicago, IL.

Bennett, A., 2008. Process tracing: a Bayesian perspective. In: Brady, H., Collier, D. (Eds.), The Oxford Handbook of Political Methodology. Oxford, New York, NY.

Bloemraad, I., 2006. Becoming a citizen in the United States and Canada: structured mobilization and immigrant political incorporation. Social Forces 85, 667–695.

Boswell, Terry, Brown, Cliff, 1999. The scope of general theory: methods for linking deductive and inductive comparative history. Sociological Methods and Research 28, 154–185.

Campbell, D.T., 1975. Degrees of freedom and the case study. Comparative Political Studies 9, 178–193.

Campbell, D.T., Stanley, J.C., 1966. Experimental and Quasi-experimental Designs for Research. Rand McNally, Chicago, IL.

Chen, Anthony S., Weir, Margaret, 2009. The long shadow of the past: risk pooling and the political development of health care reform in the states. Journal of Health Politics, Policy, and Law 34, 679–716.

Collier, D., 1995. Translating quantitative methods for qualitative researchers: the case of selection bias. American Political Science Review 89, 461–466.

Collier, D., Mahoney, J., 1996. Insights and pitfalls: selection bias in qualitative research. World Politics 49, 56–91.

Dion, Douglas, 1998. Evidence and inference in the comparative case study. Comparative Politics 30, 127–145.

Dumont, L., 1980. Homo Hierarchicus: The Caste System and Its Implications. University of Chicago, Chicago, IL.

Eckstein, H., 1975. Case study and theory in political science. In: Greenstein, F.I., Polsby, N.W. (Eds.), Handbook of Political Science. Strategies of Inquiry, vol. 7. Addison-Wesley, Reading, MA.

Fourcade-Gourinchas, M., Babb, S.L., 2002. The rebirth of the liberal creed: paths to neoliberalism in four countries. American Journal of Sociology 108, 533–579.

Geddes, B., 1990. How the cases you choose affect the answers you get: selection bias in comparative politics. In: Stimson, James A. (Ed.), Political Analysis. University of Michigan Press, Ann Arbor, MI, pp. 131–150.

George, A., 1979. Case studies and theory development: the method of structured, focused comparison. In: Lauren, P.G. (Ed.), Diplomacy: New Approaches in History, Theory and Policy. Free Press, New York, NY.

George, A., Bennett, A., 2005. Case Studies and Theory Development in the Social Sciences. MIT, Cambridge, MA.

Gerring, J., 2006. Case Study Research: Principles and Practices. Cambridge University, New York, NY.

Goldthorpe, J., 1991. The uses of history in sociology: reflections on some recent tendencies. British Journal of Sociology 42, 211–230.

Goldthorpe, J., 1997. Current issues in comparative macrosociology. Comparative Social Research 16, 1–26.

Katz, J., 1982. Poor People's Lawyers in Transition. Rutgers University, New Brunswick, NJ.

King, G., Keohane, R.O., Verba, S., 1994. Designing Social Inquiry: Scientific Inference in Qualitative Research. Princeton University, Princeton, NJ.

Lieberson, S., 1991. Small N's and big conclusions: an examination of the reasoning based on a small number of cases. Social Forces 70, 307–320.

Lieberson, S., 1994. More on the uneasy case for using Mill-type methods in small-N comparative studies. Social Forces 72, 1225–1237.

Lijphart, A., 1971. Comparative politics and the comparative method. American Political Science Review 65, 682–693.

Lijphart, A., 1975. The comparable cases strategy in comparative research. Comparative Political Studies 8, 157–175.

Lipset, S.M., 1963. The First New Nation: The United States in Historical and Comparative Perspective. Basic Books, New York, NY.

Mahoney, J., 2010. Colonialism and Postcolonial Development: Spanish America in Comparative Perspective. Cambridge University Press, Cambridge, MA.

Mahoney, J., Goertz, G., 2004. The possibility principle: choosing negative cases in comparative research. American Political Science Review 98, 653–689.

Meyer, J., Strang, D., 1993. Institutional conditions for diffusion. Theory and Society 22, 487–511.

Morgan, K.J., Prasad, Monica, 2009. The origins of tax systems: a French– American comparison. American Journal of Sociology 114, 1350–1394.

Riley, D., 2005. Civic associations and authoritarian regimes in interwar Europe: Italy and Spain in comparative perspective. American Sociological Review 70, 288–310.

Ragin, C.C., 1987. The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies. University of California, Berkeley, CA.

Ragin, C.C., 1997. Turning the tables: how case-oriented research challenges variable-oriented research. Comparative Social Research 16, 27–42.

Ragin, C.C., 2000. Fuzzy-Set Social Science. University of Chicago, Chicago, IL.

Ragin, C.C., 2008. Redesigning Social Inquiry: Fuzzy Sets and Beyond. University of Chicago, Chicago, IL.

Ragin, C.C., Amoroso, L., 2011. Constructing Social Research: The Unity and Diversity of Method, second ed. Pine Forge, Thousand Oaks, CA.

Ragin, C.C., Rihoux, B., 2004. Qualitative comparative analysis (QCA): state of the art and prospects. Qualitative Methods 2, 3–13.

Rueschemeyer, D., 1991. Different methods – contradictory results? Research on development and democracy. In: Ragin, C.C. (Ed.), Issues and Alternatives in Comparative Social Research. E.J. Brill, Leiden, The Netherlands.

Skocpol, T., 1992. Protecting Soldiers and Mothers: The Political Origins of Social Policy in the United States. Harvard University Press, Cambridge, MA.

Stake, R., 1995. The Art of Case Study Research. Sage, Thousand Oaks, CA.

Steel, D., 2004. Social mechanisms and causal inference. Philosophy of the Social Sciences 34, 55–78.

Steinmetz, G., 2008. The colonial state as social field: ethnographic capital and native policy in the German overseas empire before 1914. American Sociological Review 73, 589–612.

Stryker, R., 1990. Science, class, and the welfare state: a class-centered functional account. American Journal of Sociology 96, 684–726.

Swenson, P.A., 2002. Capitalists against Markets: The Making of Labor Markets and Welfare States in the United States and Sweden. Oxford University Press, Oxford.

von Wright, G.H., 1971. Explanation and Understanding. Cornell University, Ithaca, NY.

Walton, J., 1992. Making the theoretical case. In: Ragin, C.C., Becker, H.S. (Eds.), What Is a Case? Exploring the Foundations of Social Inquiry Cambridge, New York, NY.

White, H., 1992. Cases are for identity, for explanation, or for control. In: Ragin, C.C., Becker, H.S. (Eds.), What Is a Case? Exploring the Foundations of Social Inquiry Cambridge, New York, NY.

Yin, R.K., 1994. Case Study Research: Design and Methods, second ed. Sage, Thousand Oaks, CA.

Case Studies

Robert K Yin, COSMOS Corporation, Bethesda, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the practice of doing case study research at a high level. Rather than attempting a comprehensive overview, the article focuses attention on three critical issues, by: (1) offering a working definition of case studies to determine whether case study research is to be the method of choice in the first place; (2) identifying exemplary case studies that can serve as a way of setting high standards; and (3) suggesting procedures for seeking to make a case study credible and generalizable – two of the most common challenges in strengthening case study research.

Case Study Research: Practicing at a High Level

The present article discusses the practice of doing case study research, including its use in evaluation research. The encyclopedia has other articles on case studies, covering the history, logic, and methods and analysis of case studies, with less directly related articles on "case-oriented research" and "single-case designs" (see cross-references to other articles at end of this text).

From a practice perspective, the article assumes that you may be doing or planning to do case study research, not just wanting to learn about its history or analytic foundations. Moreover, appearance in an encyclopedia implies that an article should have more lasting value than, say, a chapter in an edited book. As a result, this article strives to cover bolder ground – not only emphasizing the practice of doing case study research, but also practicing it at a *high level*.

To practice at a high level and complementing comprehensive guidance already appearing elsewhere (Yin, 2014), the most critical issues include knowing: whether a case study should be the method of choice in the first place; how to set standards for doing high-quality case studies; and what procedures to consider in anticipating the most likely challenges when doing case study research. The article deals with the three issues, by:

1. offering a working definition of case studies and discussing the implications of this definition in knowing whether a case study is to be the method of choice;
2. identifying exemplary if not classic case studies – deliberately representing different academic disciplines and professional fields – as a way of setting high standards for case study research; and
3. suggesting procedures for addressing two of the most common challenges doing case study research – seeking to make a case study credible and to make it generalizable.

Overall, the purpose of the article is to review the critical issues that you – who may be a seasoned or a budding scholar – may encounter in actually doing a case study.

A Working Definition and Its Implications

An initial concern deals with whether you have appropriately identified your method of choice. To stay on track for doing a case study (as opposed to pursuing some other method), you might ask how closely your planned study matches a working definition of case studies – as well as the implications of abiding by such a definition.

The working definition considers a case study to be:

- An empirical inquiry that closely examines a contemporary phenomenon (the *case*) within its *real-world context*.

The resulting case study will be an up-close and in-depth inquiry into a specific, complex, and real-world phenomenon (the *case*).

This working definition befits case study research but also distinguishes it from other social science research methods, such as surveys, experiments, histories, and economic studies. The definition's main implications therefore deserve a bit more of your attention.

First, by following the working definition, have no doubt that the 'up-close' and 'in-depth' aspects imply an intensive data collection effort. The effort can involve *multiple sources of evidence*, because your inquiry is likely to be based on a combination of: direct observations in the field; lengthy interviews with different kinds of informants; and, where relevant, potentially far-reaching access to documents and archives (e.g., media and data files). You even might carry out a quantitative subanalysis about some aspects of the case. In addition, the interviews and documentary sources will not only add depth to your case study but also can cover actions over an extended period of time. With such a range, your completed case study can be more than a mere 'snapshot' of events.

Second, as the main subject of your case study, the 'case' will preferably be a concrete entity (e.g., a person, organization, community, program, policy, practice, decision, or event), existing in a specific time and place. In contrast, when the term 'case' is used in an abstract sense (e.g., an argument, claim, or proposition), its study does not readily point to using the case study method. Rather, most if not all social science methods, and not just case study research, are capable of dealing with arguments, claims, or propositions.

Third, your case study may consist of a single case (a *single-case study*) or contain two or more cases (a *multiple-case study*). Both situations satisfy the working definition. Your decision on whether to include a single or multiple cases will depend on the nature of your case study. For instance, if the case being studied involves a part of the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10507-0

social world rarely accessible to social scientists (e.g., the illicit activities of drug dealers; or the inside relationships among the key executives in a company), a single revelatory case study will be the likely result. Alternatively, if your case study starts with a theoretical proposition that will benefit from a comparison of deliberately contrasting or complementary cases, a multiple-case study will be the likely result.

Fourth, accounting for the *real-world context* assures that a case study will embed a case within an actual setting and circumstance. In this sense, case studies do not isolate their case(s) in some artificial environment (as in conducting an experiment in a laboratory or holding survey interviews over some electronic medium like a phone or a Web site). Covering the contextual conditions adds to a fuller and hopefully more accurate understanding of the case.

At the same time, accounting for the context creates an extra burden on a case study: The case and its context may not necessarily have clear boundaries. For example, a case study of a contemporary organization can lead to a possibly difficult choice: whether a predecessor version of the organization should be considered to be part of the case or part of the context (see **Example 1** for another illustration).

As a final observation, the working definition can accommodate different epistemological orientations in doing your research–typically, a contrast between an *interpretivist* (or *relativist*) orientation and a *realist* orientation (see Schwandt, 2007, for succinct definitions of these terms). The differences in orientations represent deeply felt worldviews and not merely terminological differences (e.g., Yin, forthcoming). To give but an inkling of the worldview differences, from an interpretivist perspective, you might want your case study to acknowledge a world with multiple realities and multiple meanings, with findings that are observer dependent. In doing such a *qualitative case study*, attending to multiple sources of evidence (such as field observations and interviews) can help you to get at these multiple meanings. Alternatively, from a realist perspective, your case study may be examining a reality that is assumed to be observer *in*dependent. An important motive for collecting evidence from multiple sources would then assume a cross-checking objective – to determine the convergence of the evidence.

Exemplary (if not Classic) Case Studies

Besides testing your planned study against the working definition and its implications, you also can check whether you are on the right track in doing a case study by closely examining previously published case studies – preferably ones that have been frequently cited in the research literature. More important, you can use such case studies to set a high standard for your work, hoping that some aspect of your planned case study – its design, data collection, or analysis strategies – resembles the desirable features of one or more of these published case studies.

Making such comparisons is not only possible but also fair, because the published case studies exist in nearly every academic discipline and professional field. Such diversity means that you should readily be able to find case studies in your chosen discipline or field. The brief review that follows will give you a flavor of the kinds of case studies available for your inspection and review.

Early Roots in Sociology, Psychology, and Anthropology

For instance, scholars readily trace some of the earliest case studies to the department of sociology at the University of Chicago (the *Chicago School*). Its members long ago produced case studies with an enduring heritage in the social sciences (e.g., Platt, 1992). To mention but a few of these case studies, they typically focus on different social roles, ranging from *The Polish Peasant in Europe and America* (Thomas and Znaniecki, 5 vols., 1918–20), to *The Hobo* (Anderson, 1923), to juvenile delinquents (Shaw, 1930), and to medical students (Becker et al., 1961).

In addition to the case studies from the Chicago School, other early case studies also examined local communities, with *Middletown* and *Middletown in Transition* perhaps being two of the most well-known and highly regarded (Lynd and Lynd, 1929, 1937). Social groups, such as members of local clubs (e.g., Whyte, 1943/1993) and unemployed men (Liebow, 1967) also have been frequent subjects of case studies. The works by Whyte and Liebow in particular have attained a classic status, coming in part as a result of scholars giving their works an inordinate amount of attention, including inquiries about their methods and leading to an array of supplemental writings (e.g., Whyte, 1984, 1989, 1992). Besides providing a revealing understanding of underappreciated segments of society, both works demonstrate how someone outside of the social group being studied can befriend the group and use participant–observation to conduct a case study. Liebow (1993) later produced yet another variant of case study research, which takes the form of life histories as cases.

At the same time, a different strand of case studies, with equally deep roots, also has existed in the field of psychology. The case studies there are likely to be about clinical patients or about rare occurrences involving individual human behavior (e.g., Rolls, 2005). These kinds of case studies still appear with frequency in clinical psychology. The case studies are likely to appear in separate journal articles, with the most famous neurological case – a person who had lost the ability to record new memories (Corkin, 2013) – having been the topic of over 30 such articles published between 1957 and 1968 alone.

Example 1 Defining the Case in a Case Study

The case being studied is a small group of friends and their interactions and relationships. Whether the friends' parents and siblings are to be considered part of the case being studied or part of the context for the case will affect the nature of the case study's questions and findings. The initial definition of the case should therefore explicitly identify the people considered to be inside or outside of the group, with the consequences explained and justified as part of the case study design – even though such a preliminary bounding of the case can later be revisited.

Finally, a longstanding case study tradition also has existed in the field of cultural anthropology. The anthropological field methods continue to provide a model for doing case study fieldwork, with participant – observation along with questionnaires administered to field participants being a frequent combination of data collection (e.g., Emerson, 2001, pp. 4–7). The earliest anthropological studies tend to emphasize the customs, traditions, and daily life in specific societies or social groups, such as Margaret Mead's classic *Coming of Age in Samoa* (1928), which studied individuals and families in three villages. As with many other older and classic works, although Mead's conclusions have come under later scrutiny, her methods and approach have nevertheless been accorded their due recognition (e.g., Reichardt and Rallis, 1994, p. 7). Although the early anthropological studies may not have dwelled on any particular 'case,' later studies have certainly focused on specific cases, such as the six high schools in Lawrence-Lightfoot's (1983) contrast between public and private schooling.

Contemporary Case Studies

Continuing to the more contemporary scene, one classic case study examined the confrontation between the United States and the then-Soviet Union during the Cuban missile crisis (Allison, 1971; Allison and Zelikow, 1999). The case study has long been a best-selling book in the field of political science. In this case study, and in contrast to those typically based on extensive field observations and interactions, the bulk of the case study's evidence comes from an extensive set of post hoc interviews as well as access to an exhaustive array of documents, some only made available years after the crisis. Another case study, marked by a similar combination of post hoc interviews and access to an exhaustive array of documents, analyzed the decision-making and other actions leading to the tragic loss of the space shuttle *Challenger* (Vaughan, 1996). The findings led to a differing interpretation of the disaster's causes – 'sociology of mistake' – than that offered by the Presidential Commission created to investigate the incident.

Community studies continue to be a frequent topic of other contemporary case studies. A case study can deal with a single community (e.g., Small, 2004), as in the *Middletown* studies. Other case studies can focus on a local community organization, rather than on the community as a whole, as the 'case' to be studied. When such organizations are the subject of study, the typical contemporary case study may include two or more organizations as part of the same multiple-case study. A benefit of this approach is the ability to develop insights into different types of organizations, as in a study of four different types of organizations in each of two urban neighborhoods (Marwell, 2007).

In like manner, the contemporary literature contains other exemplary examples of both single- and multiple-case studies. How to choose between these two variants depends on your research questions as well as on the practical constraints of finding one or more cases. In evaluation research, a common design involves a single-case study with wholistic evaluation questions, such as the effects of reforming a single-school district; however, an integral part of the ensuing case study relies on the results from a quantitative analysis, such as regressions that examine the achievement trends among the district's students (e.g., Supovitz and Taylor, 2005).

In contrast, an evaluation in the field of public health followed a multiple-case study design. The case study examined how and why immunization programs can be successful, with the case study covering 12 cases – four in each of three countries in sub-Saharan Africa (Mookherji and LaFond, 2013). Of the four programs in each country, three were considered positive examples because they had shown improved outcomes, and one was considered a negative example because it had shown no such improvement. A further distinctive feature of this study was that, while the main methods and results were reported in a journal article (Mookherji and LaFond, 2013), the individual case studies were made available through a separate Web site (JSI Research & Training Institute, Inc., http://arise.jsi.com). Such a dual arrangement has become increasingly common and gives readers access to both the main work and the supporting evidence.

A third evaluation example chose the interactions between a teacher and a student with developmental disabilities as the 'case,' studying eight such pairs as cases (Miyahara and Wafer, 2004). The case study systematically varied the teaching strategies used in each of the pairs and applied a replication logic to determine the efficacy of the strategies by assessing the students' performance on a variety of psychometric measures over time.

Finally, the preceding study as well as one by Rivera (2008) shows how good case studies have not been limited to book formats but have appeared as articles in contemporary journals. Rivera's case study (discussed later in this article) comes from the field of foreign affairs and appears in one of the top journals in sociology.

Overall, many of the cited case studies have been among the leaders in their field. Others that have been cited have distinctive features in addition to having been carefully designed and conducted. Because debates over the quality of case studies may be unending, a practical strategy for setting high standards for your work will be to acknowledge if not emulate these previously published case studies.

Two Common Challenges in Doing Case Study Research: Credibility and Generalizability

Two common challenges to the quality of case study research derive from caveats over a case study's credibility and generalizability. A high-level case study needs to address these challenges. Although the caveats cannot be entirely dispelled, they can be reduced by practicing certain methodological procedures. Successfully following these procedures in doing a case study will increase the confidence in a case study's findings as well as in the utility of its findings.

Many people may have developed a strong skepticism about case study research because of an early exposure to a poorly executed case study. Alternatively, inappropriate impressions about case study research may have come from the use of case studies as classroom materials in an academic course. Such *teaching case studies* have been specially prepared to

illustrate a key issue in the course and to stimulate classroom discussion. The teaching case studies exist in virtually every field – both the academic disciplines as well as the professions, such as law, business, public administration, urban planning, social work, public health, and medicine. Unfortunately, the teaching case studies, though appearing under the seemingly same 'case study' moniker as *research case studies*, differ sharply from the research case study that you are trying to conduct.

Teaching case studies have a pedagogical motive. Their texts can be manipulated to create a teachable moment, without regard to the need to present evidence completely or fairly. In contrast, research case studies should satisfy research, not pedagogical motives. Therefore, your case study will need to show explicitly how it has collected and presented its evidence, both completely and fairly. Meeting such a challenge becomes one way of producing a *credible* research case study.

Teaching case studies also may not have to be concerned with developing generalizations – possibly other than pedagogical ones or those related to the course being taught. In contrast, your research case study will be of increasing value the more that its findings have implications beyond the study itself. Meeting this challenge means making generalizations from your case study. On exceptional occasions, a single-case study may be so unique or important that no generalization is desired – for instance, Stake's (2005) *intrinsic* case studies and Lawrence-Lightfoot and Davis's (1997) *portraits*. However, barring an extremely compelling subject matter, you will then have to argue compellingly the value of doing such a case study.

These two topics – producing a credible case study whose findings are generalizable – have become the main challenges in doing case study research at a high level.

Building Credibility

How you deal with credibility may depend on the differences, as discussed at the outset of this article, between the *interpretivist* and *realist* orientations. The relevant evidence may assume different characteristics, and the strategies for dealing with different kinds of evidence may differ.

As a general guide, you may think of three ways of strengthening the credibility of a case study: creating an aura of *trustworthiness*, dealing with concerns over *validity*, and striving for *reliability*. You may attend to these three ways in isolation from each other or in any combination.

Creating Trustworthiness

Especially if you have adopted an *interpretivist* orientation, creating a strong sense of trustworthiness may become the main way of building credibility. The trustworthiness begins by convincing readers that you actually did the fieldwork reported in your case study – that is, being "really and fully present – physically, cognitively, and emotionally – in the scenes of action under study" (Eisenhart, 2006, p.574). Even within the pluralistic, multiple-reality assumptions embedded in an interpretivist perspective, a case study needs to demonstrate that the data and interpretations are accurate *from some point of view* [emphasis added], which also implies reporting about the reflexivity or two-way interplay between the researcher and the events and participants in the field (Eisenhart, 2006, pp. 575–579). Helpful procedures in this process of building trustworthiness include keeping a careful field diary and having well-organized field notes. If nothing else, such procedures will convey an important message – which you are trying to be transparent about your research methods.

In pursuing this transparency, a particular ethical challenge arises in assuring readers that your case study has reported and fairly analyzed *all* the collected data. Such a challenge also exists with other research methods, as in doing experiments, where a constant temptation is to find some procedural excuse to exclude data that might in fact have been contrary to initial expectations. In doing case study research, the conversational nature of field interviews and serendipity of field observations can create a gray area whereby contrary but seemingly casual or even incredulous remarks or observations may be ignored because they were not considered to be part of the 'formal' data collection (Yin, 2011, p. 39). As a partial remedy, you might develop a tool, ahead of time, containing a brief but written decision-making framework, explicitly acknowledging your own rules for reporting or ignoring field information (Newman and Brown, 1996, pp. 101–113). You will then need to know when you might be tempted to "make an exception" and to "counter the temptation with an even stronger admonition regarding the dire consequences of breaking your own rules" (Yin, 2011, p. 39).

Shifting to yet another way of dealing with trustworthiness, engaging in *thick description* has been a hallmark of ethnographic research (Ryle, 1949; Geertz, 1973, 1983). Typically, the highly-detailed accounts enable a reader to appreciate and ultimately derive a deep understanding of the social conditions being studied. However, thick description also can play a role in strengthening the trustworthiness of case study research and therefore deserves your consideration as yet another procedure. By engaging in thick description, you can increase or at least surface your awareness of the filter regarding your potential selectivity biases and your use of preconceived categories (Becker, 1998, p. 85). In other words, such description, portraying the people, events, and actions within their locally meaningful contexts, can help to reveal implicitly-held researcher-centric perspectives. These revelations will add to the transparency and in turn the trustworthiness of a case study.

Dealing with Validity

Worrying about the validity of a study offers a second way of addressing the credibility challenge. The validity concept largely derives from a realist orientation. Any interest, much less concern, over validity may not arise with case studies done with an interpretivist orientation (Yin, forthcoming).

Validity issues especially arise when doing an explanatory (compared to a descriptive) case study. In such a case study, the main explanation tries to cover how and why the studied happenings transpired as they did. The quest in effect implies some kind of causal relationships, while fully acknowledging the complexity of such relationships. In setting forth its main explanation, a case study needs to amass credible data in support of these claimed relationships (see **Example 2**). At the same time, some scholars differ in accepting the relevance of causality as a relevant concept in the social sciences (Yin, forthcoming). For instance, Lincoln and Guba (1985, pp. 150–157) consider the social world to be one of *mutual*

Example 2 Using field observations to explain how and why real-world events occur

Case studies often call for the direct study of social processes in local settings – e.g., the operations inside an organization or school, or the workings of community or neighborhood groups. Doing such fieldwork can include direct field observations about causal processes, with the observations leading to explanations about how and why the real-world events have occurred (e.g., Betts, 2013, p. 255; Erickson, 2012, p. 688; Maxwell, 2004; 2012; Miles and Huberman, 1994, p. 132). The observations even can help to appreciate the complexity of causal relationships as they occur in the real world, such as the potential nonlinear and recursive flows of events and the potential multiplicity of causes, both proximal and distal (e.g., Betts, 2013).

Example 3 Illustrating an explanation and a rival explanation

A case study aims to explain how and why an interorganizational partnership came to operate successfully. The main explanation involves the leadership by the lead partner. The case study also entertains and collects data about a rival explanation – that the prepartnership relationships among several of the organizations, but not the lead partner, established the partnership's central working patterns. The case study then thoroughly compares the evidence supporting the main explanation with that supporting the rival explanation.

simultaneous shaping, where 'everything influences everything else' with 'no directionality,' even though outside observers may label some of the happenings as 'outcomes or effects.' In setting forth its main explanation, a case study needs to amass credible data in support of these claimed relationships (see **Example 2**).

To boost the validity of the findings, a case study needs to examine closely not only its main explanation but also one or more *plausible rival explanations*. These rival explanations are similar to those that sit at the core of the experimental method (Campbell, 2014). In doing experiments, much attention has been given over the years to the identification of methodological rivals posing threats to a study's 'internal validity' (Cook and Campbell, 1979). However, little attention has been given to the types of substantive (or real-world) rivals that represent rival substantive claims and not just methodological rivals. Articulating these real-world rivals (see **Example 3**), and then collecting data to check them out, often can be one of the more stimulating aspects of doing case study research, while simultaneously strengthening a case study's validity. **Exhibit 1** presents a tentative typology of such real-world rivals, also listing a well-known group of methodological rivals.

To address rivals, the analytic procedure calls for comparing the evidence in support of a case study's main explanation with the evidence in support of the plausible rival explanation(s). To make the comparison as fair as possible, you should try aggressively and exhaustively to collect data about the rival(s) – as if trying to prove the potency of the rival(s); then, if finding a lack of such support, you can reject the rival(s) (Patton, 2002, p. 553; Rosenbaum, 2002, pp. 8–10). Assuming that the evidence also strongly supported the main explanation, successful comparisons with plausible rivals will increase confidence in the validity of a case study's findings. As noted in the field of education research (but also applicable to many other related fields), "the use of qualitative methods...can be particularly helpful in ruling out alternative explanations…[and] can enable stronger causal inferences" (Shavelson and Towne, 2002, p. 109).

At the same time, a continuing challenge in doing case study research is to understand how plausible rival explanations are best identified, and how best to judge whether a rival is truly plausible and strong or merely a red herring. Similarly, no benchmarks yet exist for deciding how and whether a rival is to be rejected or accepted, or to remain in an equivocal status. Future research will need to fill these gaps.

Striving for Reliability

A third way of dealing with the credibility challenge consists of strengthening the reliability of a case study. The challenge may arise in doing any kind of case study but in particular, in doing descriptive as well as explanatory case studies. The case study needs to convince readers that data have been collected and analyzed consistently and fairly, and that the case study descriptions represent the claimed reality or the perspective of field informants – but not the analyst's imagination. To be avoided is any reference to a case study as a 'story-telling' activity, or to case study analysts as case study 'writers.' Both references connote a freedom of expression that can go beyond the collected evidence. Even when a case study adopts an interpretivist orientation – for instance, deliberately acknowledging the existence of multiple perspectives or fully appreciating the 'meaning' of what is being studied – these perspectives and meanings also need to be accurately rendered.

To strengthen reliability, the desired case study method will first make the data collection procedure as transparent as possible. An important procedure, using a *case study protocol*, can help to promote the needed transparency. The protocol differs from the conventional survey questionnaire that is directed at an interviewee. Rather, the protocol has a series of questions or topics directed at the case study analyst, along with clues and probes pointing the analyst to the potential measures and sources of evidence. A good protocol covers the entire line of inquiry to be followed in a case study. The preferred line of inquiry has questions and topics that nevertheless leave room for discovery and unexpected field occurrences, as well as for the multiple realities likely to emerge qualitatively from different interviewees.

During actual data collection, the desired method will then include a continual striving to triangulate findings, a procedure that also will boost the reliability of a case study. *Triangulation* occurs when the evidence from several different sources converges on the same finding, increasing the reliability of the finding. The triangulating process also can help to operationalize the concepts in a case study, partially offsetting the common concern over the lack of formal measures in much

Methodological Rivals:

1. The Null Hypothesis (the occurrence is the result of chance circumstances only)
2. Threats to Validity (e.g., history, selection, testing, instability, maturation, instrumentation, regression, experimental mortality, and selection-maturation interaction)
3. Investigator Bias (reactivity in field research; also "experimenter effect")

Real-World Rivals:

1. Direct Rival (alternative events explain the results better)
2. Commingled Rival (other events also help to explain the results)
3. Implementation Rival (the implementation process, not a substantive rationale, explains the results)
4. Rival Theory (an alternative theory explains the results better)
5. Super Rival (a larger force, including the ones being studied, explains the results)
6. Societal Rival (secular trends, not any particular actions, explain the results)

Exhibit 1 Types of rivals.

case study research–and thereby boosting a case study's *construct validity* (Yin, 2014, pp. 46, 47).

The opportunity to triangulate derives from the fact that the typical case study fieldwork should actively embrace multiple sources of evidence. In other words, the fieldwork should deliberately seek to corroborate key findings by searching for multiple sources of evidence. In interpretivist situations, as in accurately rendering an interviewee's unique perspective over or 'meaning' about some matter, no other source of evidence may be relevant. Under this circumstance, you would use a triangulating procedure to prompt the interviewee to express the perspective or meaning on more than one occasion or in more than one way, thereby establishing the interviewee's consistency about the matter.

Still, and as with the situation in examining plausible rival explanations, little attention has been given to the development of a formal process for carrying out the triangulation procedure. For instance, no benchmarks exist to define when triangulation might be considered 'strong' or 'weak' or 'complete' or 'incomplete,' and thus whether an analyst needs to continue pursuing new sources of evidence to strengthen an existing triangulation. Future research will again need to fill this gap.

Generalizing from a Case Study

The supposed difficulty in generalizing from a single-case study – or from a small set of case studies – has long been considered a major shortcoming of case study research. How to go about seeking to make generalizations from case studies will probably remain a longstanding quest.

Making Analytic Generalizations

The desired generalization takes the form of an *analytic generalization*, which consists of a substantive proposition or other principle or set of principles, derived from a case study's findings and then claimed to be applicable to other situations (Yin, 2013, 2014).

The generalizing procedure involves extracting a more abstract level of ideas from a case study's findings and asserting, mainly through careful argument rather than any numeric or statistical calculation, how these ideas might pertain to newer situations other than those in the original study. (Generalizing in this manner may resemble what others have called 'transferability,' with the extent of transferability depending upon "the degree of similarity of the sending and receiving contexts" (Lincoln and Guba, 1985, p. 297).) Thus, the preferred analytic generalization is posed at a conceptual level higher than that of the specific case(s) in a case study (**Exhibit 2** highlights the two levels). Presumably, this higher level was needed to justify the importance of studying the chosen case(s) in the first place. In other words, the substantive proposition may be similar to the one that might have started a case study and that has now, upon completion of the case study, gained empirical support (Small, 2009).

In addition, an analytic generalization should go beyond serving only as a 'working hypothesis' (e.g., Cronbach, 1975) – that is, one in need of further study rather than being ready to be generalized or applied to a new situation. This shortcoming is not easily overcome. However, thoughtfully linking an analytic generalization to the related research literature by identifying conceptual overlaps and gaps will help. Replicating the same findings by conducting a second or third case study – analogous to conducting a second or third experiment to corroborate the findings from an initial experiment – also will help (e.g., Yin, 2014, pp. 57–59). Eventually the ideal analytic generalization may not only extend to situations having other 'like' cases but also "apply to many *different types* of cases" (Bennett, 2004, p. 50) [emphasis added].

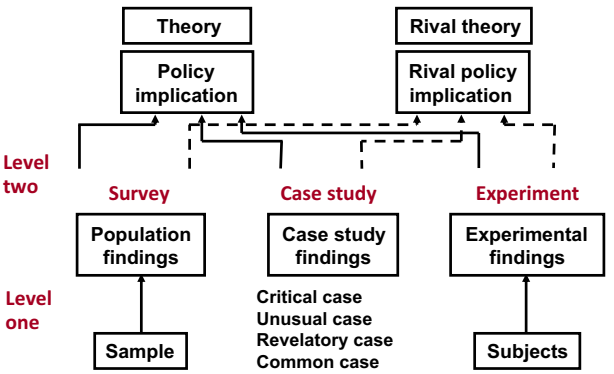

Exhibit 2 Making generalization: Two levels.

Example 4 Making analytic generalizations from a single-case study

A classic example of the use of analytic generalizations comes from Graham Allison's well-known single-case study of the Cuban missile crisis, in which the United States and the then-Soviet Union confronted each other in a warlike manner that threatened to spread into a global conflict (Allison, 1971; Allison and Zelikow, 1999). The case study posits three theories to explain the crisis, alternatingly matching the actions predicted by each theory with the actions that actually occurred. The most potent combination of theories then becomes the basis for the analytic generalization and its implications for a broad array of international relationships described by the author – certainly was not limited to the features of the particular 'case' that was studied (i.e., limited to missile crises or even to relationships involving the United States and the Soviet Union).

Example 5 Making analytic generalizations from another single-case study

Another illustration of an analytic generalization appears in a journal article (Rivera, 2008). The case study examines how the Croatian government came to portray the country's past, present, and future in the aftermath of the wars of Yugoslav secession. The wars had left a reputation-damaging effect, threatening Croatia's highly-valued tourist industry. The case study shows how, in response, the government reframed the country's past by omitting the war in its representations of national history, repositioning the country as more closely sharing a history and culture with its Western European neighbors. The explanation for these findings then draws from a theoretical framework, in which the author innovatively extends Erving Goffman's well-regarded sociological work on "stigma and the management of spoiled identity" (Goffman, 1963) from the individual to the institutional realm. The author concludes by claiming that the analytic generalization has applicability to other situations of collective memory and cultural sociology.

'Case Study,' not 'Case'

One major barrier in trying to generalize from case studies may have arisen as a result of confusion between generalizing from a 'case study' and generalizing from a 'case.' Thus, when people question how to generalize from a 'case study,' they may inadvertently think in terms of how to generalize about the 'case(s)' in the case study – typically invoking a sample to population logic, as in making a *statistical generalization*. Such thinking results in an immediate pressure to expand the number of cases. However, you need to remember that the classic case study research may be limited to examining a single case in-depth, and the relevant issue is whether and how to claim a generalization under such a circumstance.

Overcoming the confusion requires understanding that the goal is to generalize from a case study, regardless of the number of cases in the case study. Two other scholars also make the same distinction between a case and a case study. While the case being studied may (or may not) be considered "an instance of a larger class of events" (e.g., George, 1979; George and Bennett, 2004, p. 17), the case study approach involves "the detailed examination of an aspect of a historical episode to develop or test historical explanations that may be generalizable to other events" (George and Bennett, 2004, p. 5). In their terms, the *episode* might be considered the 'case,' and the *historical explanations* the analytic generalizations derived from the case study.

Though not using the same terminology, other prominent works have devoted attention to analytic generalization, also distinguishing it from statistical generalization: (1) Mitchell's (1983) discussion of *logical inference* and *statistical inference*; (2) Bromley's (1986) *case inference* compared with *statistical inference* (pp. 290–291); and (3) Donmoyer's (1990) *schema*. A fourth work, by Burawoy (1991), covers the *extended case method* – his way of describing how a generalization 'extends' a case study to some broader significance (pp. 271–280).

At the same time, no formal procedures yet exist that can guide you in seeking to make analytic generalizations, unlike the ample and almost mechanical guidance available for making statistical generalizations. As a result, you should again appeal to the existing literature for the best examples, while also setting a high standard. To get you started, consider the following two examples. Both deliberately illustrate analytic generalizations on the basis of *single*-case studies (see **Examples 4 and 5**).

Where You Go from Here

This article has defined and discussed the practice of case study research emphasizing the focal topics that will enable you to practice case study research *well*. A major assumption has been that you are in the midst of doing or planning to do your own case study. See if you can: work with an explicit definition of case studies, as presented in the working definition at the outset; cite, use, and relate your case study to some exemplary if not classic case studies; and adopt and use the procedures and tools suggested here and in the literature to meet the challenges of credibility and generalizability. You may not excel the first time. But over a series of case study projects, you should eventually be practicing case study research at a high level.

See also: Case Method and Case Study Collections; Case Study: Logic; Case Study: Methods and Analysis; Case-Oriented Research; Human–Environment Interactions: Case Studies; Single-Case Designs: Methodology.

Bibliography

Allison, G.T., 1971. Essence of Decision: Explaining the Cuban Missile Crisis. Little, Brown, Boston.

Allison, G.T., Zelikow, P., 1999. Essence of Decision: Explaining the Cuban Missile Crisis, second ed. Addison Wesley Longman, Boston.

Anderson, N., 1923. The Hobo. University of Chicago Press, Chicago.

Becker, H., 1998. Tricks of the Trade: How to Think about Your Research While You're Doing It. University of Chicago Press, Chicago.

Becker, H., Geer, B., Hughes, E.C., Strauss, A.L., 1961. Boys in White: Student Cultdure in Medical School. University of Chicago Press, Chicago.

Bennett, A., 2004. Testing theories and explaining cases. In: Ragin, C.C., Nagel, J., White, P. (Eds.), Workshop on Scientific Foundations of Qualitative Research. National Science Foundation, Arlington, VA, pp. 49–51.
Betts, J., 2013. Aid effectiveness and governance reforms: applying realist principles to a complex synthesis across varied cases. Evaluation 19, 249–268.
Bromley, D.B., 1986. The Case-Study Method in Psychology and Related Disciplines. Wiley, Chichester, England.
Burawoy, M., 1991. The extended case method. In: Burawoy, M., et al. (Eds.), Ethnography Unbound: Power and Resistance in the Modern Metropolis. University of California Press, Berkeley, pp. 271–287.
Campbell, D.T., 2014. Foreword. In: Yin, R.K. (Ed.), Case Study Research: Design and Methods, fifth ed. Sage, Thousand Oaks, CA, pp. xvii–xviii.
Cook, T.D., Campbell, D.T., 1979. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Rand McNally, Chicago.
Corkin, S., 2013. Permanent Present Tense: The Unforgettable Life of the Amnesic Patient, H.M. Basic Books, New York.
Cressey, P., 1932. The Taxi-Dance Hall. University of Chicago Press, Chicago.
Cronbach, L.J., 1975. Beyond the two disciplines of scientific psychology. American Psychologist 30, 116–127.
Donmoyer, R., 1990. Generalizability and the single-case study. In: Eisner, E.W., Peshkin, A. (Eds.), Qualitative Inquiry in Education: The Continuing Debate. Teachers College Press, New York, pp. 175–200.
Eisenhart, M., 2006. Representing qualitative data. In: Green, J.L., Camilli, G., Elmore, P.B. (Eds.), Handbook of Complementary Methods in Education Research, third ed. American Educational Research Association, Washington, DC, pp. 567–581.
Emerson, R.M. (Ed.), 2001. Contemporary Field Research: Perspectives and Formulations, second ed. Waveland Press, Prospect Heights, IL.
Erickson, F., 2012. Comments on causality in qualitative inquiry. Qualitative Inquiry 18, 686–688.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Geertz, C., 1983. Local Knowledge: Further Essays on Interpretive Anthropology. Basic Books, New York.
George, A.L., 1979. Case studies and theory development: the method of structured, focused comparison. In: Lauren, P.G. (Ed.), Diplomacy: New Approaches in History, Theory, and Policy. Free Press, New York, pp. 43–68.
George, A.L., Bennett, A., 2004. Case Studies and Theory Development in the Social Sciences. MIT Press, Cambridge, MA.
Goffman, E., 1963. Stigma: Notes on the Management of Spoiled Identity. Prentice--Hall, New York.
JSI Research & Training Institute, Inc. http://arise.jsi.com).
Lawrence-Lightfoot, S., 1983. The Good High School: Portraits of Character and Culture. Basic Books, New York.
Lawrence-Lightfoot, S., Davis, J.H., 1997. The Art and Science of Portraiture. Jossey-Bass, San Francisco.
Liebow, E., 1967. Tally's Corner. Little, Brown, Boston.
Liebow, E., 1993. Tell Them Who I Am: The Lives of Homeless Women. Penguin Books, New York.
Lincoln, Y.S., Guba, E.G., 1985. Naturalistic Inquiry. Sage, Thousand Oaks, CA.
Lynd, R.S., Lynd, H.M., 1929. Middletown: A Study in Modern American Culture. Harcourt Brace Jovanovich, New York.
Lynd, R.S., Lynd, H.M., 1937. Middletown in Transition: A Study in Cultural Conflicts. Harcourt Brace Jovanovich, New York.
Marwell, N.P., 2007. Bargaining for Brooklyn: Community Organizations in the Entrepreneurial City. University of Chicago Press, Chicago.
Maxwell, J.A., 2004. Using qualitative methods for causal explanation. Field Methods 16, 243–264.
Maxwell, J.A., 2012. The importance of qualitative research for causal explanation in education. Qualitative Inquiry 18, 655–661.
Mead, M., 1928. Coming of Age in Samoa: A Psychological Study of Primitive Youth for Western Civilisation. Perennial Classic Edition published in 2001, New York.
Miles, M., Huberman, M., 1994. Qualitative Data Analysis: A Sourcebook for New Methods. Sage, Thousand Oaks, CA.
Mitchell, J.C., 1983. Case and situation analysis. Sociological Review 31, 187–211.
Miyahara, M., Wafer, A., 2004. Clinical intervention for children with developmental coordination disorder: a multiple case study. Adapted Physical Activity Quarterly 21, 281–300.
Mookherji, S., LaFond, A., 2013. Strategies to maximize generalization from multiple case studies: lessons from the Africa routine immunization system essentials (ARISE) project. Evaluation 19, 283–303.
Newman, D.L., Brown, R.D., 1996. Applied Ethics for Program Evaluation. Sage, Thousand Oaks, CA.
Patton, M.Q., 2002. Qualitative Research and Evaluation Methods, third ed. Sage, Thousand Oaks, CA.
Platt, J., 1992. "Case study" in American methodological thought. Current Sociology 40, 17–48.
Reichardt, C.S., Rallis, S.F., 1994. The relationship between the qualitative and quantitative traditions. New Directions for Program Evaluation 61, 5–11.
Rivera, L.A., 2008. Managing 'spoiled' national identity: war, tourism, and memory in Croatia. American Sociological Review 73, 613–634.
Rolls, G., 2005. Classic Case Studies in Psychology. Hodder Education, Abingdon, England.
Rosenbaum, P.R., 2002. Observational Studies, second ed. Springer, New York.
Ryle, G., 1949. The Concept of Mind. Basic Books, New York.
Schwandt, T.A., 2007. The Sage Dictionary of Qualitative Inquiry, third ed. Sage, Thousand Oaks, CA.
Shavelson, R., Towne, L. (Eds.), 2002. Scientific Research in Education. National Academies Press, Washington, DC.
Shaw, C., 1930. The Jack Roller. University of Chicago Press, Chicago.
Small, M.L., 2004. Villa Victoria: The Transformation of Social Capital in a Boston Barrio. University of Chicago Press, Chicago.
Small, M.L., 2009. "How many cases do I need?" on science and the logic of case selection in field-based research. Ethnography 10, 5–38.
Stake, R.E., 2005. Qualitative case studies. In: Denzin, N.K., Lincoln, Y.S. (Eds.), The Sage Handbook of Qualitative Research, third ed. Sage, Thousand Oaks, CA, pp. 443–466.
Supovitz, J.A., Taylor, B.S., 2005. Systemic education evaluation: evaluating the impact of systemwide reform in education. American Journal of Evaluation 26, 204–230.
Thomas, W.I., Znaniecki, F., 1918–1920. The Polish Peasant in Europe and America. Dover, New York (Reprinted in 1958).
Vaughan, D., 1996. The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA. University of Chicago Press, Chicago.
Whyte, W.F., 1943/1993. Street Corner Society: The Social Structure of an Italian Slum, fourth ed. University of Chicago Press, Chicago.
Whyte, W.F., 1984. Learning from the Field: A Guide from Experience. Sage, Thousand Oaks, CA.
Whyte, W.F., 1989. Introduction to action research for the twenty-first century: participation, reflection, and practice. American Behavioral Scientist 32, 502–512.
Whyte, W.F., 1992. In defense of street corner society. Journal of Contemporary Ethnography 21, 52–68.
Yin, R.K., 2011. Qualitative Research from Start to Finish. Guilford Press, New York.
Yin, R.K., 2013. Validity and generalization in future case study evaluations. Evaluation 19, 321–332.
Yin, R.K., 2014. Case Study Research: Design and Methods, fifth ed. Sage, Thousand Oaks, CA.
Yin, R.K. Causation, generalizability, and the future of mixed methods research. In: Hesse-Biber, S.N., Johnson, R.B. (Eds.), The Oxford Handbook of Mixed Methods Research. Oxford, New York, in press.

Case Study: Logic

Anthony M Orum, University of Illinois at Chicago, Chicago, IL, USA; and Loyola University, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of single cases, or case study research, has become one of the most important and popular methodological tools at the disposal of social science researchers. This article discusses the general logic of case study research. It provides a basic definition of the case study and then suggests three kinds of case studies – the typical, the prototypical, and the deviant. In addition, it also speaks to the strengths of the case study, as a way of studying history and the evolution of the case, and identifies pitfalls and challenges. Illustrations and examples are furnished throughout.

Introduction

Today there is a renewed interest in qualitative research and in case study research, in particular. Scholars can thank a number of people for sparking this interest, among them, Skocpol (1979) who through a careful and brilliant analysis of social revolutions revealed how researchers can employ a small sample of cases to study broad and significant issues. Case studies have a long and distinguished pedigree in the social sciences, dating from writings like those of Lynd and Lynd (1929) on Middletown, as well as Gans (1962) and Whyte (1943) on neighborhoods. The resurgent interest is not merely based on impressionistic evidence. In just the past several years, new and important writings, such as those by Gerring (2004, 2007) and George and Bennett (2005), have helped to elucidate the logic of case studies, both single cases as well as comparative cases. Moreover, a quick review of the social science citation index will reveal the growing, indeed exploding, number of citations to the works of Gerring (2004, 2007), Ragin (1992), Ragin and Becker, (1992) as well as the present writer and his colleagues (Feagin et al., 1991).

A Definition of the Case Study

While people use case studies for a host of different projects, as, for example, the study of social revolutions, cities, democracies, among other objects, very little careful attention has been given to the definition of a case study itself. This is a major problem. Without a clear definition of a case study, it is hard to know how, if at all, one can generalize to some larger set of cases. Is a particular study a case of a city, or a case of a neighborhood? Is it a case of a person, or of a social role? These and other questions must be answered at the very outset of the project in order for other researchers to know how to properly employ the results of a particular empirical investigation, and this becomes all the more critical as case study research is done more often and in a more diverse set of fields in the social sciences.

Here we shall define a case study as an in-depth study of a single phenomenon whose boundaries and content can be made conceptually and empirically clear. There are two key elements to this definition. One is that the phenomenon has a clear content, and the other is that it also has boundaries within which it can be examined. Thus, one could be interested in social revolutions, as the contents, and the boundaries for such social revolutions can be those of states, or even regions, within which they occur. Very often, in fact, the contents may shade into the definition of the boundaries, as social revolutions begin within states but then spread into regions, e.g., the recent Arab Spring in the Middle East. Equally often the contents themselves may be defined and redefined over time as the particular phenomenon evolves. It may begin with a small network of people but then evolve into a variety of organizations that become, in effect, social movement organizations. Social movements provide perhaps the clearest example of how the contents and boundaries may shift and become vague. Other case studies, like those of neighborhoods, or even of social roles, can furnish instances where the clarity is greater – but conceptual clarity will always require considerable and careful attention in the study of cases.

The Selection of Cases: Types and the Reasons/ Criteria for Selection

Once one identifies and defines the contents and boundaries of a particular case study, the next issue is to think about and decide on the larger population of cases to which one wants to generalize. Will the case, for example, be one that sheds light on the nature of American political parties, or will it be all political parties in general? Does one wish to generalize to all cities, or to all neighborhoods, or only to low-income neighborhoods?

In thinking more deeply about the choice of a particular case and of the larger population to which one wants to generalize, one must carefully decide about the specific nature and character of the case itself. In the past, social science researchers sometimes neglected this issue, but the more that scholars turn to case studies today the more important it becomes. Here we shall suggest three particular types that have found their way into the literature recently. For a more refined definition of cases, the reader is advised to turn the illuminating article by Brenner (2003).

The Typical Case

Generally speaking, most students of case studies, if they are confined to only one or two cases, in particular, choose to think of their case as typical of the larger population to which they

 http://dx.doi.org/10.1016/B978-0-08-097086-8.44002-X

hope to generalize. There are a number of such instances, but perhaps the most famous is found in the pioneering research of Lynd and Lynd (1929). Working in the 1920s and 1930s, well before the methods and costs of social science research had advanced, the Lynds wanted to identify a city whose population was typical of that of the United States. Such a city, they hoped, would allow them to study the nature of work and family life in a setting that represented the average, or standard, setting in which Americans at the time lived. The city they chose was Muncie, Indiana, and the title they gave to their now-famous books was that of *Middletown*. They spent years exploring the city and its past, did interviews with local residents, and created a portrait of everyday Americans and America. They believed that their findings, uncovered during the Depression years of the 1930s, represented the way that urban life in America was experienced. Their work and that of other researchers who following their landmark research, also uncovered in Muncie, Indiana, uncovered such things as the dominance of a single upper-class family and the precarious nature of economic life at the time (see, for example, Caplow et al. 1985).

Muncie, the Lynds argued, provided a typical urban setting, and because it was typical, the Lynds believed they could generalize their research to the nature of American life in general. Note that unlike scholars today, who would draw a random sample of the American population, based on such parameters as age, gender, class, and race, the Lynds contextualized the life experiences they studied. Work was defined by the nature of work in Muncie, religious life was defined by the nature of religious institutions in Muncie, and civic life was defined by the nature of civic life in Muncie. The urban and historical context was essential, they believed, to their research: people were embedded in the daily comings and goings of the place rather than being disconnected as would happen with a typical social survey.

The Prototypical Case

A second type of case that one might choose to select is what can be called a prototypical case. The prototype represents a case that is not merely the average or typical instance of a phenomenon but is one that, in the eyes of the researcher, furnishes a model of how particular phenomenon might develop in the future in a range of cases whose character is similar to that of the prototypical case. The prototype represents a single case, but its full dimensions can be illuminated through comparisons to other instances of the same general population of cases.

One of the most compelling instances of the prototype occurs in de Tocqueville's (1945) analysis of democracy in America. Tocqueville traveled in the United States in the 1830s and succeeded in writing the most masterful portrait ever of the origins of American culture and politics. He cast his analysis as an effort to understand both the causes and consequences of American democracy. He believed that in understanding these conditions his analysis would foreshadow how democracy would eventually emerge in European nations.

Thus he argued that it was the general equality of conditions among Americans, as compared to their European counterparts that helped to promote and sustain the democratic institutions of the United States. America, he argued, did not have a long history of social and political inequality as existed during the course of European feudalism, and thus it was the very cultural foundations here that helped to sustain the effort to promote political equality. In addition, Americans were accustomed to creating and populating various types of voluntary associations, such as clubs and similar organizations, and this too, by comparison with European countries, further helped to sustain the democratic institutions.

By laying out these several and varied preconditions for the emergence of democracy in America, then, Tocqueville hoped to show systematically how the United States differed from such countries as France and England, and therefore, by implication, what the special causes of democracy were. None of this is to suggest, by the way, that he was unaware of the flaws and failures of America in the nineteenth century. He was well aware of the real inequalities that existed, among them in terms of racial differences as well as in terms of the development of capitalism. It was the American ideals, embedded both in the culture and in the institutional makeup that was so especially unique and powerful to the character of the country.

While Tocqueville of course developed his analysis through a continuing, if often implicit, comparison between the United States and European countries, his work was notable for the depth of his analysis and profound insights in the study of America, itself. Like other students of case studies, particularly single cases, he was intent on digging more deeply into the nature of American democracy and comparing it, often implicitly, with other nations he knew well.

The case study as a prototype is a very important method and strategy for case study researchers, and one that should be employed more often as case study research becomes more popular. However, the researcher must remember that this type of case is notable not for its typicality but for its exceptional nature, one that, in the eyes of the researcher, could serve as a model that other similar cases could follow in the future.

The Deviant Case

A third type of case is that of the deviant case study. Here, as in the instance above, the particular case is taken as special, and its analysis proceeds on the assumption that its unique qualities will help to shed light on a broader set of cases and instances in the population. However, unlike the prototypical case, the use of the deviant case is pursued to reveal how the exceptional instance illuminates the broader set of typical or unexceptional cases. One of the most notable examples of the deviant case in the social sciences is the study undertaken of the International Typographers Union (ITU) by Lipset et al. (1956) in the 1950s.

This particular case study chose a question very different from studies of typical cases. Instead of asking what the characteristics of this typical instance are, they wondered how and why the ITU was able to sustain democratic institutions, such as elections and a rotation of leadership, when, in fact, at the time no other unions were able to do so. In other words, how and why was the ITU so special? They presumed, in their research (as Tocqueville did in his), that their findings and insights would enable other unions, if they chose to do so, to introduce democratic institutions and practices as well.

Through an examination of the archives of the union, personal interviews, and election results, Lipset and his colleagues uncovered some important results. Among other things, they found that the members of the ITU tended to be more highly skilled and educated as well as more engaged in union activities than the members of other unions. They reasoned then that these higher levels of skills and active engagement in the workings/politics of the union promoted a greater democracy within it, allowing alternative points of view and working against the continuing domination by a single group of leaders and their cohorts. Although these results only pertained to the case of unions, and their internal politics, it was obvious that they also had a bearing on the nature of politics in other organizations, perhaps even those of nation-states. (Indeed, Lipset et al. suggested that higher levels of skills and greater engagement among the citizens of nation-states also might work to promote more democratic politics in those settings as well.)

Although it has been many decades since this work was completed, this particular case study stands as an important example of how a deviant case – one that displays an atypical pattern – can be extremely useful in highlighting the usual, or typical, conditions for a particular set of institutions. Like Tocqueville's example, that of union democracy takes a case that is an outlier, in effect, and attempts to show it as a model that might be followed by the larger set of cases to which it generalizes.

Case Study Analysis: An Emphasis on Process and Time over Structure

Case study research is celebrated for the depth of its analysis, and, in successful instances, for the range and quality of its insights. While it can be used to confirm, or disconfirm, specific hypotheses, as in the case of the research by Lipset and his colleagues on union democracy, it is often used as a means of generating and developing new ideas. What exactly is it about case study research that allows this to happen? If the difference does not lie in the quality of the researcher – and we assume it does not – what is it about the very nature of the research itself that promotes new insights and ideas?

Research on case studies permits the researcher to study more deeply and extensively the nature of the phenomenon itself, whether a city, a social role, or a political movement. In a word, it enables the researcher to study the process, or the development of the particular phenomenon, over time (George and Bennett, 2005). Let us take an example to show how this can take place. Suppose that a researcher is interested in the consequences of living in a segregated neighborhood as compared to an integrated neighborhood, and chooses to focus only on African–American and white American households, and their relative proportions. Now a standard quantitative study, using a wide number of cases, typically would use some kind of statistical measure of segregation, or integration, as the celebrated studies using the index of dissimilarity do. Here the advantage of the quantitative approach lies in the range of cases that one can bring to bear on the problem.

A researcher who uses case study research has different, but equally significant, options. He can explore the matter of segregation versus integration in another way. Suppose the researcher chooses to study a single neighborhood, one that has changed over time in terms of its racial composition. Here the advantage of case study research lies in the ability of the research to explore the process, or actual history, of segregation and key moments in how it unfolds over time. The researcher may choose to study a neighborhood that was previously populated primarily by blacks – hence was segregated – and that has now become more integrated, populated both by blacks and whites. (Alternatively, as is evident in a great deal of research on segregation, the researcher could choose a neighborhood that once was primarily white, and has now become home to more and more black residents.)

The recent research of Pattillo (2007), a sociologist, examines just such a phenomenon. Pattillo was curious about the process that unfolded as a neighborhood changed from a poor black to mixed income neighborhood that included black families of greater wealth. Sociologists had argued that a mixture of family households of different incomes ultimately would entail many positive consequences for the remaining poor black families themselves, furnishing a wide range of amenities and offering new and important role models for adults and children alike. Pattillo wished to discover whether this expectation was correct. In her 10-year study of the development of the Oakland–Kenwood neighborhood of Chicago she found that the positive scenario envisioned by many sociologists missed many unintended consequences of gentrification. In this particular setting, she uncovered many conflicts and resentments. The poor black families, in particular, came to resent many of the privileges and attitudes of the wealthier black families.

Case study research, moreover, not only permits the researcher to study the process of how a single phenomenon unfolds over time but also has the potential for permitting the researcher to identify the 'tipping point,' – when, for example, a neighborhood changes from a segregated to an integrated one, and the consequences of this transformation. It is not clear if a simple measure of statistical segregation can tap into this kind of shift in the consequences of the shifting composition of the neighborhood, even though one can establish a range of segregation based upon the relative percentages of different racial groups living in an area. However, a case study, with its capacity to furnish a detailed and exacting focus on the changing character of a neighborhood over time, would have precisely this advantage over quantitative studies. It can permit the researcher to focus on the typical outcomes associated with integration, for example, the crime rate among young black males, and then reveal when in a neighborhood's history the 'tipping point' that leads to a lower crime rate is reached – and why. (Conversely, of course, if the researcher is interested in how an integrated neighborhood changes into a segregated black neighborhood, and the consequences that follow, one could study the 'tipping point' when the composition of the neighborhood promotes a higher rate of crime among young black males.)

Finally, let us note that while case studies allow the researcher to examine and explore the dimensions of time and process, the issue of structure always remains available for scholars to study with cases. In the two above examples concerning democracy, the one by Tocqueville emphasizes the role

of history and of time in fashioning American democracy. By contrast, Lipset and his associates focus far more on the ways in which the structural elements of unions, as, for example, the skills and educational level of workers, encourage the development of 'union democracy.'

The Selection of Cases: Some Challenges and Pitfalls

Critics of case study research who typically favor quantitative research in the social sciences almost always raise the issue of the representativeness of case studies, an issue with which one is always concerned in large-scale quantitative studies. Students of case studies can address this criticism in two ways. First, they can do so by noting the special research purposes of case studies, which is to examine a particular phenomenon in depth as compared to the breadth and range of large-scale samples. Second, they can, and must, provide a clear and emphatic rationale for the particular cases they select. Is there something special about a particular neighborhood, for example, which will illustrate some larger concern about low-income neighborhoods? Is there something about a particular state, or set of states, i.e., regions, whose study will further illuminate the nature of social revolutions? The whys and wherefores of case selection, in other words, are of utmost importance to define, and thereby to enable other researchers to effectively use the evidence produced by specific case studies. If there is a frequent flaw in case study research it is that the researcher fails to spell out the whys and wherefores of the choice of cases, making it appear at times as though the choice were simply an arbitrary one. Indeed, if this case is only arbitrary then it is impossible to know whether, or how, the results can be generalized to other cases – even what the population of other cases might represent.

Beyond these two concerns, there is another possible pitfall involved in the selection of cases, and that concerns the way that the researcher chooses the specific cases to examine. Let us consider Skocpol's pathbreaking analysis of social revolutions. She chose to focus on a small handful of successful revolutions in order to illustrate the preconditions of revolution. She examined those revolutions in depth, and to support her claims about why revolutions occurred in one country but not another, she compared the countries with successful revolutions to those where such revolutions were unsuccessful. It was an important and useful strategy. Nevertheless later analyses of social revolutions suggest that had she chosen a larger set of cases, her results might have been different – and not nearly so robust as she claimed (Geddes, 1990).

Skocpol's work also illustrates another pitfall. She chose her cases, of social revolutions in countries, by focusing on the outcomes, on the revolutions, and then making comparisons. Recent work by Collier and Mahoney (1996) suggests that making choices about cases on the outcome, or dependent variable, can be problematic, resulting in what they refer to as 'selection bias.' The following is an example to illustrate how this might work.

Suppose a researcher is interested in whether gentrification is taking place in urban neighborhoods and what the causes of such gentrification might be. The researcher defines the gentrification of a neighborhood in terms of the changes that take place in the material wealth of the neighborhood and, in turn, defines wealth by the average level of family income, the value of residential properties, and the amount of new small commercial business, such as small high-end retail stores, businesses like Starbucks, and the like.

Because the researcher wants to study gentrification and its causes in depth he/she chooses one neighborhood that has gentrified and compares it with one that has not. Now the problem arises because of the level, or degree, of gentrification that can occur in a neighborhood. As in much of social science, outcomes rarely have one single cause, but often multiple ones. If in this instance, the researcher has chosen, let us say, a neighborhood that has become heavily gentrified over time (in terms of the indices noted above), and found the role of developers to be prominent in promoting gentrification, he/she might miss other possible causes, such as the role of real estate agents, or changes in zoning ordinances initiated by the local government, whose role often appears much earlier in the process but furnishes the initial, and critical, push that allows gentrification to occur.

The general point here is this: the selection of cases to study based on the dependent, or outcome, variable can miss possible causes – and also possible consequences – by virtue of its choice of cases. If there is a range of values in the outcome – as in this instance from highly to minimally gentrified – then a selection of only one or even two cases will obviously miss some of this critical variation. The result will be that the analysis will miss other important causes, and even consequences, of the phenomenon in question.

Issues of Theoretical Generalization and Refinement: The Use of Parallel and Different Cases

Because the analysis of a single case can only take the analyst so far, the best rule to follow, if possible, is to examine at least three or four pertinent cases in depth. Multiple cases permit the researcher to refine and develop theoretical arguments, and they do so precisely because case studies permit the researcher to examine cases deeply and, ultimately, to make careful comparisons among those cases.

One example will illustrate the advantages of using several cases and examining them in depth. Orum (1995) was interested in examining various theories about the origins of urban growth. Popular theories had suggested that such growth came about because of special alliances among the leading figures in different major institutional spheres, particularly local government, business, and the media. Orum suggested that such theories were limited only to contemporary circumstances and that they did not help much to explain the sources of growth in the past. Using several different cases and studying them in great depth permitted him to come up with several important discoveries bearing upon issues of urban growth and expansion.

The first case examined was that of Milwaukee, Wisconsin. Tracing its origins back to the early nineteenth century, Orum discovered that the city seemed to go through certain stages in its growth, and that one could not simply rely on explanations of alliances among leaders of different institutional spheres. Thus, based upon intensive study, he argued that, in the earliest

years, the growth of the city relied heavily on local entrepreneurs, people who promoted the city and who, among other things, sold land for it. However, over time this motive force for growth changed, and the main forces turned to new economic institutions, ones that brought in growth by attracting new residents. A third stage happened when the impetus for growth shifted from the local economic enterprises, which had grown somewhat stagnant, to local government. Government sought to expand the community by adding new territory. Among other things, it also became clear in this stage that a single conception of urban growth was inappropriate; growth could take the form of not only new population but also new territory. Such an insight would have been impossible had Orum relied only on the study of this single case in depth. The fourth and final phase happened as the city began to decline. In large part, the decline took place when many of its original industries left for greener pastures in other cities. Under these circumstances, a heavy burden was placed upon local government to devise ways to promote growth, either by expanding its borders or by attracting new industry through various incentives.

In effect, then, Orum's explanation for Milwaukee's growth suggested that such expansion happened by stages, and that at each stage there was a different configuration of local business and political forces responsible for such growth, and that the principal form of growth itself could vary between stages. Next, he asked whether this pattern held only for Milwaukee, or whether it could also describe the trajectory of growth and decline in other parallel industrial cities. Thus, he made comparisons in the pattern and stages of growth in Milwaukee with those in Cleveland, Ohio, a city which, on the face of it, seemed similar in many respects. In fact, the stages and forces of growth at each stage were virtually identical in Cleveland and Milwaukee. He thus argued that among American industrial cities, the nature of growth, and the forces responsible for it, showed important parallels, underscoring his argument on behalf of a stage theory of urban growth.

Finally, Orum pushed the argument one step further, wondering whether the pattern found was only characteristic of industrial growth, and cities, in America. Could the pattern of stages also apply to cities that grew up in the postindustrial era and were themselves postindustrial cities? Here he made a comparison between the patterns in Milwaukee and those in Austin, Texas, a postindustrial city he also had come to know intimately. This comparison showed that the first three stages were very similar in terms of the rapidity of growth, and in the role of entrepreneurs, local business, and local government in promoting such growth in Milwaukee and Austin alike. In the end, then, he refined his overall argument to suggest that the stage view of urban growth in America fit not only industrial-era cities but also cities in the postindustrial era. This suggested that there appeared to be an underlying structural pattern to the growth of American cities, regardless of era. This last comparison also enabled him to draw out differences between Milwaukee and Austin, allowing for further understanding of the nature of urban growth, and difference, in the United States.

In brief, then, as the example of the various cities is intended to illustrate, the careful and systematic study of several cases can permit the researcher both to substantiate important theoretical generalizations and to refine and extend them. Again, this can be done because of the deep acquaintance the researcher gains from the study of cases, an acquaintance unlike that to be obtained with methods like surveys or experiments.

Conclusion: Making the Case for Case Studies

A number of years ago, the author and his colleagues (Feagin et al., 1991) put together a book of essays on how case study research could be done and the advantages of such research. Authors in a variety of sociological fields provided entries on cases in the life cycle, cases of cities, cases of religious institutions, and cases of people who were homeless. At the time, we hoped to help revive a significant methodological tool for the social sciences, one whose use had declined over the course of the twentieth century.

"The times, they are changing," to quote a famous lyric from Bob Dylan. Today case study research is to be found everywhere in the social sciences, although it still faces many of the same standard criticisms it faced years ago. In this article I hope to have made clear that the study of cases is a very powerful and important technique for social scientists. However, one must be ever vigilant about how and why one chooses particular cases, and how to proceed to analyze them. Examples like those mentioned in this article should furnish a start to the nature and logic of case study research. Other recent writings, like those of Gerring (2007), George and Bennett (2005), Chen et al. (2012), and Ragin and Becker (1992) should help to explain and explore these important issues even further.

See also: Case Study: Methods and Analysis; Case-Oriented Research; Generalizability Theory; Generalization: Conceptions in the Social Sciences; Qualitative Methods, History of; Selection Bias, Statistics of.

Bibliography

Brenner, Neil, September 2003. Stereotypes, archetypes, and prototypes: the uses of superlatives in contemporary urban studies. City & Community, 205–216.

Caplow, Theodore, Chadwick, Bruce A., Bahr, Howard M., Hill, Reuben, 1985. Middletown Families: Fifty Years of Change and Continuity. University of Minnesota Press, Minneapolis, MN.

Chen, Xiangming, Orum, Anthony M., Paulsen, Krista, 2012. Introduction to Cities: How Space and Place Shape Human Experience. Wiley-Blackwell, Malden, MA (Chapter 4).

Collier, David, Mahoney, James, October 1996. Insights and pitfalls: selection bias in qualitative research. World Politics 49, 56–91.

De Tocqueville, Alexis, 1945. Democracy in America. In: The Henry Reeve Text as Revised by Francis Bowen, with Notes and an Interpretive Essay by Phillips Bradley, vols. I–II. Vintage Books, New York, NY.

Feagin, Joe, Orum, Anthony M., Sjoberg, Gideon (Eds.), 1991. A Case for the Case Study. University of North Carolina Press, Chapel Hill, NC.

Gans, Herbert, 1962. Urban Villagers: Group and Class in the Life of Italian–Americans. Free Press, New York, NY.

Geddes, Barbara, 1990. How the cases you choose affect the answers you get: selection bias in comparative politics. In: Stimson, James A. (Ed.), Political Analysis, vol. 2. University of Michigan Press, Ann Arbor, MI, pp. 131–150.

George, Alexander L., Bennett, Andrew, 2005. Case Studies and Theory Development in the Social Sciences. MIT Press, Cambridge, MA.
Gerring, John, 2004. What is a case study and what is it good for? American Political Science Review 98 (2), 341–354.
Gerring, John, 2007. Case Study Research: Principles and Practices. Cambridge University Press, New York, NY.
Lipset, Seymour Martin, Trow, Martin, Coleman, James S, 1956. Union Democracy: The Internal Politics of the International Typographical Union. Free Press, Glencoe, IL.
Lynd, Robert, Lynd, Helen Merrell, 1929. Middletown: A Study in American Culture. Harcourt Brace, New York, NY.
Orum, Anthony M, 1995. City-Building in America. Westview Press, Boulder, CO.
Pattillo, Mary, 2007. Black on the Block: The Politics of Race and Class in the City. University of Chicago Press, Chicago, IL.
Ragin, Charles C, 1992. Cases of 'what is a case?' In: Ragin, Charles C., Becker, Howard S. (Eds.), What is a Case? Exploring the Foundations of Social Inquiry. Cambridge University Press, Cambridge, MA, pp. 1–18.
Ragin, Charles C., Becker, Howard S (Eds.), 1992. What Is a Case? Exploring the Foundations of Social Inquiry. Cambridge University Press, Cambridge, UK.
Skocpol, Theda, 1979. States and Social Revolutions: A Comparative Analysis of France, Russia and China. Cambridge University Press, Cambridge, MA.
Whyte, William Foote, 1943. Street Corner Society: The Social Structure of an Italian Slum. University of Chicago Press, Chicago, IL.

Case Study: Methods and Analysis

Andrew Bennett, Georgetown University, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp. 1513–1519, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The article reviews case study methods from a historical perspective, starting with the initiative to formalize these methods in the 1970s and examining the distinct contributions to different types of case studies, such as atheoretical or configurative-ideographic cases, heuristic cases, or within-case analysis, among others. It also covers the new developments in case study methods built on the first attempts, particularly its relation to the Philosophy of Science, the elaboration of the concept of typological theory, and the emerging consensus on their comparative advantages and limitations. It concludes by encouraging research collaboration between scholars using diverse methodological approaches.

This article reviews case study methods, also called 'small *n*,' 'comparative,' or 'qualitative' methods. It outlines within-case methods of analysis, including process tracing and congruence testing, and cross-case methods of comparison, including John Stuart Mill's methods of agreement and difference, Alexander George's method of 'structured, focused comparison,' and Charles Ragin's method of 'qualitative comparative analysis.' It argues that case studies can help develop typological theories, or contingent generalizations on the different causal patterns leading to a particular outcome (related notions include configurational analysis and fuzzy logic). This article notes philosophies of science relevant to case study methods, including scientific realism, Bayesian confirmation theory, and Imre Lakatos' methodology of scientific research programs. Finally, it summarizes the comparative advantages of case study methods, including conceptual validity, identification of new variables, and modeling of complex interaction effects, as well as their comparative limitations, including the potential for indeterminacy and the dangers of selection bias. The central conclusion is that the comparative advantages of case study methods, statistical methods, and formal modeling are different and complementary, and that research proceeds more effectively through the combination of these methods than through the use of any one method alone.

Case study methods have been around as long as recorded history, and they presently account for a large proportion of the books and articles in anthropology, biology, economics, history, political science, psychology, sociology, and even the medical sciences. The logic of case study methods, much like that of any historian's or detective's efforts to make inferences from patterns within cases and comparisons between them, is more intuitive than the logic of statistical inference. Until relatively recently, however, the lack of formalization of the logic of case study methods inhibited them achieving their full potential for contributing to the progressive and cumulative development of theories. It is only in the last three decades that scholars have formalized case study methods and linked them to underlying arguments in the philosophy of science. Ironically, statistical methods, though less intuitive, were standardized earlier, so that attempts to formalize case study methods often misappropriated terms and concepts from statistics (McKeown, 1999). More recently, case study methods have evolved from being defined 'negatively,' via contrasts to statistical methods, to being defined 'positively,' by their distinctive logic, techniques, and comparative advantages. This continuing evolution remains a contested process, but there is growing consensus on the proper procedures for carrying out case studies and the strengths and limits of such studies. It is becoming increasingly clear through this process that the comparative advantages of case study and statistical methods are largely complementary and that the two methods can thus achieve far more scientific progress together than either could alone.

The Definition of 'Case' and 'Case Study'

Early efforts to define 'case studies' relied on distinctions between the study of a small versus a large number of instances of a phenomenon. Case studies thus became characterized as 'small *n*' studies, in contrast to 'large *N*' statistical studies. Related to this, one early definition stated that a 'case' is a "phenomenon for which we report and interpret only a single measure on any pertinent variable" (Eckstein, 1975). Case study researchers have increasingly rejected this definition, however, because it wrongly implies, in the language of statistics, that there is an inherent 'degrees of freedom' problem in case studies. In other words, this definition of case studies suggests that with a greater number of variables than observations on the dependent variable, case studies provided no basis for causal inference. In the view of case study researchers, however, each case includes a potentially large number of observations on intervening variables and qualitative measures of different aspects of the dependent variable, so there is not just a 'single measure' of the variables or an inherent degrees of freedom problem. This point is increasingly recognized by researchers from the statistical tradition (King et al., 1994; Bennett and Elman, 2004, 2006).

In addition, the 'small *n*/large *N*' distinction implies that large *N* methods are always preferable whenever sufficient data are available. As argued below, however, case studies can serve useful theory-building purposes, such as the inductive generation of new hypotheses, even when instances of a phenomenon are sufficiently numerous to allow the application of statistical methods.

For present purposes, then, *a case is defined as an instance of a class of events* (George, 1979a,b). The term 'class of events' refers here to a phenomenon of scientific interest, such as revolutions, types of governmental regime, kinds of economic system, or personality types. A case study is thus a well-defined aspect of a historical happening that the investigator selects for analysis, rather than a historical happening itself. The Cuban Missile Crisis, for example, is a historical instance of many different classes of events: cases of deterrence, coercive diplomacy, crisis management, and so on. In deciding which class of events to study and which theories to use, the researcher decides what data from the Cuban Missile Crisis are relevant to their case study of it. Of course, even if one accepts the present definition of a case, there is still room in the context of particular studies to debate such questions as: 'what is this event a case of' and "given this phenomenon, is this event a case of it?" (Ragin and Becker, 1992).

There is potential for confusion among the terms 'comparative methods,' 'case study methods,' and 'qualitative methods.' In one view the comparative method, or the use of comparisons among a small number of cases, is distinct from the case study method, which in this view involves the internal examination of single cases (*see* Comparative Studies: Method and Design). For the present purposes, however, case study methods are defined to include both within-case analysis of single cases and comparisons between or among a small number of cases. This is not an effort to claim wider meaning for the term 'case studies,' but an outgrowth of the growing consensus that the strongest means of drawing inferences from case studies is the use of a *combination* of within-case analysis and cross-case comparisons within a single study or research program, although single case studies can also play a role in theory development. As for the term 'qualitative methods,' this is sometimes used to encompass both case studies carried out with a positivist view of the philosophy of science and those implemented with a postmodern or interpretive view. This present article hews to the traditional terminology in focusing on 'case studies' as that subset of qualitative methods that has adopted a largely positivist framework.

The Historical Development of Case Study Methods

Case study methods have developed through several phases over the last three decades. Prior to the 1970s, 'case studies' consisted primarily of historical studies of particular events, countries, or phenomena, with little effort to cumulate results or progressively develop theories (Verba, 1967). Throughout the 1970s, however, scholars who were dissatisfied with the state of case study methods, and encouraged by the example of the formalization of statistical methods, began to formalize case study methods.

First, Adam Przeworski and Henry Teune (1970) clarified the logic of 'most similar' and 'least similar' case comparisons. In the former comparison, which draws on the logic of John Stuart Mill's method of difference and mimics the experimental method, the researcher compares two cases that are similar in all but one independent variable and that differ in the outcome variable. Such a comparison may be consistent with the inference that the difference in the single independent variable that varies between the cases accounts for the difference in the dependent variable (although for a variety of reasons discussed below, this inference may be spurious). In a comparison of least similar cases, which draws on Mill's method of agreement, the researcher compares two cases that differ in all but one independent variable that have the same value on the dependent variable. If, for example, we find that teenagers are 'difficult' in both tribal societies and industrialized societies, we might be tempted to infer that it is the nature of teenagers rather than the nature of society that accounts for the difficulty of teenagers.

Arend Lijphart (1971) and Harry Eckstein (1975) contributed further to the formalization of case study methods by clarifying the differences among various types of case study research designs and theory-building goals. These authors identified similar types, although their terminology differs, and Lijphart adds an important type, the 'deviant case,' for which Eckstein does not make explicit provision. Their types of case studies correspond as shown in **Table 1**.

The atheoretical or configurative-ideographic case study takes the form of a detailed narrative or 'story' presented in the form of a chronicle that purports to illuminate how an event came about. Such a narrative is highly specific and makes no explicit use of theory or theory-related variables. Most case studies, however, do have an explanatory purpose. These studies generally fall into the category of 'disciplined-configurative' or 'interpretive' case studies, in which general propositions are used, often implicitly, to explain specific historical cases. Another variant of such case studies is the use of cases as examples that illustrate a theory.

Heuristic case studies seek to generate new hypotheses inductively from the study of particular cases. Notably, statistical methods lack this capacity for inductively generating hypotheses, and they typically rely instead on hypotheses derived deductively or borrowed from case study research. An especially important type of case study for developing new hypotheses is the 'deviant' case study. This is the study of a case whose outcome is not predicted or explained adequately by existing theories. Unless the outcome of a deviant case turns out to be a consequence of measurement error, the case is likely to be useful for identifying variables that have been left out of existing theories. Finally, researchers can use case studies to test whether the outcomes and processes that theories predict in particular cases are in fact evident.

Eckstein's and Lijphart's contributions demonstrated that there was not just a single type of case study, but many kinds of case study research designs and many different theory-building purposes that they could serve. Their treatments differed,

Table 1 Types of case studies

Lijphart	*Eckstein*
Atheoretical	Configurative-ideographic
Interpretative	Disciplined configurative
Hypothesis generating	Heuristic
Deviant	?
Theory-confirming/infirming	Crucial, most likely, least likely

however, in that Lijphart relied greatly on statistical concepts and language. He was thus skeptical of the value of single case studies for building social science theories, and, consistent with the widespread preference at the time for 'large *N*' over 'small *n*' methods, he urged researchers to consider several means of either decreasing the number of variables in their models or increasing the number of cases to be studied in order to make use of statistical rather than case study methods. This advice, however, raised the risk "conceptual stretching" (Sartori, 1970), or of lumping together dissimilar cases under the same definitions. Possibly for this reason, Lijphart later placed greater emphasis instead on the controlled comparison of most similar to cases as a basis for causal inference (Lijphart, 1975; Collier, 1993).

Eckstein, in contrast, focused on the use of case studies for theory testing and argued that even single case studies could provide tests that might strongly support or impugn theories. In so doing, Eckstein developed the idea of a 'crucial case,' or a case that "*must closely fit* a theory if one is to have confidence in the theory's validity, or, conversely, *must not fit* equally well any rule contrary to that proposed" (Eckstein, 1975; his emphasis). Eckstein argued that true crucial cases are rare, so he pointed to the alternative of 'most likely' and 'least likely' cases. A most likely case is one that is almost certain to fit a theory if the theory is true for any cases at all. The failure of a theory to explain a most likely case greatly undermines our confidence in the theory. A least likely case, conversely, is a tough test for a theory because it is a case in which the theory makes only a weak prediction. A theory's ability to explain a least likely case is strong evidence in favor of the theory. In this way, Eckstein argued, even single case studies could greatly increase or decrease our confidence in a theory or require that we alter its scope conditions.

Alexander George (1979a,b) further developed case study methods by refining 'within-case' analysis and cross-case comparisons in ways that help each method compensate for the limits of the other. George argued, as Mill himself had, that the 'method of difference' and the corresponding practice of comparison of most similar cases could lead to spurious inferences. One reason for this is that no two nonexperimental cases achieve the ideal of being similar in *all* respects but one independent variable and the outcome. Thus, there is always the danger that left-out variables or residual differences in the values of the independent variables account for the difference in the outcomes of similar cases of (*see* Human-Environment Interactions: Case Studies). In addition, as Mill recognized, phenomena might be characterized by what general systems theorists have termed 'equifinality,' or the condition in which the same outcome can arise through different causal pathways or combinations of variables. Thus, there might be no single necessary or sufficient variable for a phenomenon: it might be that either ABC or DEF causes Y, and that none of the variables A–F is itself sufficient to cause Y. In such circumstances, pair-wise comparisons of cases might wrongly reject variables that contribute to the outcome of interest in conjunction with some contexts but not with others, and might also accept as causal variables that are in fact spurious.

To compensate for these limits of controlled comparison, George developed the 'within-case' methods of 'congruence testing' and 'process tracing' as means of checking on whether inferences arrived at through case comparisons were spurious. In congruence testing, the researcher checks whether the prediction a theory makes in a case, in view of the values of the case's independent variables, is congruent with the actual outcome in the case. In process tracing, the researcher examines whether the causal process a theory hypothesizes in a case is in fact evident in the sequence and values of the intervening variables in that case. Thus, process tracing might be used to test whether the residual differences between two similar cases were causal or spurious in producing a difference in these cases' outcomes. Process tracing can perform a heuristic function as well, generating new variables or hypotheses on the basis of sequences of events observed inductively in cases.

George (1979a,b) also systematized case study procedures by developing what he called the method of 'structured focused comparison.' In this method, the researcher systematically: (1) specifies the research problem and the class of events to be studied; (2) defines the independent, dependent, and intervening variables of the relevant theories; (3) selects the cases to be studied and compared; (4) decides how best to characterize variance in the independent and dependent variables; and (5) formulates a detailed set of standard questions to be applied to each case. In addition, consistent with his emphasis on equifinality, George argued that case studies could be especially useful in developing what he called 'typological theories,' or contingent generalizations on "*the variety of different causal patterns* that can occur for the phenomena in question … [and] *the conditions under which each distinctive type of causal patterns occurs*" (George, 1979a; his emphasis). He thus advocated a kind of 'building block' approach to the development of theories in which each case, while rendered in terms of theoretical variables, might prove to be a distinctive causal pathway to the outcome of interest.

In the 1980s and 1990s, thousands of books and articles made use of these improvements in case study methods in a wide variety of social science research programs. Meanwhile, scholars continued to elaborate case study methods and articulate the ways in which they differed from statistical methods. David Collier, reviewing the development of case study and comparative methods, argued that these methods have advantages in defining and measuring qualitative variables in conceptually valid ways and forestalling the problem of conceptual stretching (Collier, 1993). Charles Ragin argued that qualitative methods were also better than statistical methods at accounting for equifinality and complex interaction effects. Although statistical methods can model several kinds of interaction effects, Ragin noted, they can do so only at the cost of requiring a larger sample size, and models of nonlinear interactions rapidly become complex and difficult to interpret. Ragin also introduced the method of Qualitative Comparative Analysis (QCA), which uses Boolean algebra to reduce a series of comparisons of cases to the minimum number of logical statements or hypotheses that entail the results of all the cases compared (Ragin, 1987). This method, he argues, makes comparisons among cases in ways that treat them inherently as configurations of variables and that allow for the possibility of equifinality and complex interactions. Both Collier and Ragin also noted the limitations of case study methods, including the potential for indeterminacy when attempting to sort out rival explanations in a small number of cases, the difficulty of

attaining a detailed understanding of more than a few cases, and the inability to make broad generalizations on the basis of small numbers of cases.

New Developments in Case Study Methods

Thousands of applications of case study methods in the last two decades have provided fertile ground for further methodological refinements. Three recent key developments include the strengthening of linkages between case study methods and the philosophy of science, the elaboration of the concept of typological theories, and the emergence of elements of consensus on the comparative advantages and limitations of case study methods.

Case Studies and the Philosophy of Science

With regard to the philosophy of science, the 'scientific realist' school of thought has emphasized that causal mechanisms, or independent stable factors that link causes to effects under certain conditions, are important to causal explanation (Little, 1998). This has resonated with case study researchers' use of process tracing to uncover evidence of causal mechanisms at work. It has also provided a philosophical counterpoint to attempts by researchers from the statistical tradition to place 'causal effects,' or the expected difference in outcomes brought about by the change in a single independent variable, at the center of causal explanation (King et al., 1994). Case study researchers have argued that both causal mechanisms, which are more easily addressed by case studies, and causal effects, which are best assessed through statistical means, are essential to the development of causal theories and causal explanations (George and Bennett, 2001).

Another relevant development in the philosophy of science has been the resurgence of interest in Bayesian logic, or the logic of using new data to update prior confidence levels assigned to hypotheses. Bayesian logic differs from that of most statistics, which eschew reliance on prior probabilities. Eckstein's crucial, most likely, and least likely case study designs implicitly use a Bayesian logic, assigning prior probabilities to the likelihood of particular outcomes (McKeown, 1999). One new development here is the refinement of Eckstein's approach, taking into consideration the likelihood of an outcome not just in view of one theory, but in the presence of alternative hypotheses. If a case is 'most likely' for a theory, and if the alternative hypotheses make the same prediction, then the theory will be strongly impugned if the prediction does not prove true. The failure of the theory cannot be blamed on the influence of the variables highlighted by the alternative hypotheses. Conversely, if a theory makes only a weak prediction in a 'least likely' case, the alternative hypotheses make a different prediction, but if the first theory's prediction proves true, this is the strongest possible evidence in favor of the theory (Van Evera, 1997). This helps address the central problem of a Bayesian approach – that of assigning and justifying prior probabilities – even if it does not fully resolve it.

The continuing development of the logic of hypothesis testing has also been relevant to case study methods (*see* Hypothesis Testing: Methodology and Limitations). On this topic, Imre Lakatos argued that a theory can be considered progressive only if it predicts and later corroborates 'new facts,' or novel empirical content not anticipated by other theories (Lakatos, 1976). This criterion helps provide a standard for judging whether process tracing, the designation of new subtypes, and the proposal of new theories from heuristic case studies are being done in a progressive or regressive way. It also provides a philosophical basis for arguing that a hypothesis can be derived from one set of observations within a case and then to some extent tested against the 'new facts' or previously unexamined or unexpected data that it predicts within that same case, although independent corroboration in other cases is usually advisable as well (Collier, 1993).

Typological Theories and 'Fuzzy Logic'

A second recent development in case study methods has been the elaboration of the concept of typological theory. Typological theories occupy a middle ground between covering laws, or highly general abstract propositions, and causal mechanisms. Typological theories identify recurring conjunctions of mechanisms and provide hypotheses on the pathways through which they produce effects. Thus, like QCA, typological theories treat cases as configurations. Unlike QCA, they do not attempt to reduce the number of theoretical statements about the variables, but retain a diverse and admittedly complex set of contingent generalizations, with potentially one generalization per type. Consequently, typological theories are well suited to modeling equifinality.

To construct typological theories, researchers first specify the variables and use them to define the typological space, or the set of all mathematically possible combinations of the variables (this is sometimes termed a truth table in the philosophy of science). At first this may seem to produce an unmanageably large number of combinations: a model with five dichotomous variables, for example, would have 32 possible types. However, once the researcher begins to categorize extant cases in a preliminary way into particular types, it often becomes possible to narrow the range of cases of interest for study. Many types may remain empty, with no extant cases. Some types may be overdetermined for the outcome of interest, and hence not worthy of study unless they have an unexpected outcome. From among the cases and types that remain, the researcher can use the preliminary categorization of cases within the typological space to help identify most likely, least likely, most similar, least similar, and crucial cases for study. Cases in the typological space with unexpected outcomes, or deviant cases, can help identify new causal pathways that can be added to the existing theory in a kind of "building block" approach (George and Bennett, 2001). A related development concerns the concept of "fuzzy logic" (Ragin, 2000). Fuzzy logic treats cases as configurations but rather than using dichotomous or trichotomous variables and categorizations of cases, it allows the use of scaling to give a score on the extent to which a case fits into a certain type. In other respects, the use of fuzzy logic proceeds in ways much like those of typological theories.

The Emerging Consensus on the Strengths and Limits of Case Study Methods

A third development is that while several debates on case study methods continue, others have moved toward synthesis or even

closure, and the overall picture is of an emerging consensus on the advantages limitations of case study methods. As noted above, researchers from a variety of methodological traditions have recognized that because case studies can include many observations, they do not suffer from an inherent degrees of freedom problem. At the same time, it is also widely agreed that particular case studies may suffer from indeterminacy, or an inability to exclude all but one explanation on the basis of available process tracing evidence (Njolstad, 1990). When this occurs, it may still be possible to narrow the number of plausible explanations, and it is also important to indicate as clearly as possible the extent to which the remaining hypotheses appear to be complementary, competing, and incommensurate in explaining the case.

Second, most case study researchers have readily acknowledged the limits of Mill's methods. Ragin's alternative of qualitative comparative analysis makes less restrictive assumptions, but its results are highly sensitive to changes in the measurement or coding of a single case (Goldthorpe, 1997). There has thus been a movement toward typological theories and fuzzy logic, which make still less restrictive assumptions than QCA and are not so sensitive to the results of a single case. In addition, there is growing consensus that the use of within-case methods of analysis helps provide a check on the potential spuriousness of cross-case comparisons (Collier, 1993; Mahoney, 1999; George and Bennett, 2001). Case study researchers consequently seldom if ever rely on case comparisons alone.

Third, there is growing recognition that the case selection criteria necessary for statistical studies are in some respects inappropriate for case studies. Random selection in a case study research design, for example, can result in worse biases than intentional selection (King et al., 1994). There is also increasing understanding that, consistent with the reliance of some case study designs on a Bayesian logic, case studies are sometimes intentionally selected not to be representative of some wide population but to provide the strongest possible inferences on particular theories (McKeown, 1999). There is still disagreement between those who warn against any selection on the dependent variable (King et al., 1994) and who argue that selection on the dependent variable is appropriate for some research objectives (Collier and Mahoney, 1996; Ragin, 2000; George and Bennett, 2001). Related to this is a continuing disagreement over whether single case studies can make only limited contributions to theory building (King et al., 1994), or whether single case studies have indeed reshaped entire research programs (Rogowski, 1995). There is wider agreement, however, that selection bias is potentially more severe in case studies than statistical studies because biased selection of case studies can overstate as well as understate the relationship between the independent and dependent variables (Collier and Mahoney, 1996).

On the whole discussions of these issues have moved toward an emerging consensus on the comparative advantages and limitations of case study methods (Bennett, 2004). These methods' advantages include the conceptualization, operationalization, and measurement of qualitative variables (conceptual validity), the avoidance of conceptual stretching, the heuristic identification of new variables and hypotheses (often through study of deviant cases), the assessment of whether statistical generalizations offer plausible or spurious explanations of individual cases, the incorporation of equifinality and complex interactions effects, and the inferences made possible by combining within-case and cross-case analyses (Collier, 1993; Munck, 1998; George and Bennett, 2001). It is possible that new statistical methods may be able to improve upon the statistical treatment of equifinality and interaction effects, and at least narrow the gap in the treatment of this issue, but the other comparative advantages of case study methods appear to be inherent in their differences from statistical methods.

The limits of case study methods include their inappropriateness for judging the relative frequency or representativeness of cases, their weakness at performing partial correlations and establishing causal effects or causal weight, the necessarily narrow and contingent nature of their generalizations, and the danger that selection bias can be more catastrophic than in statistical studies (Collier, 1993; King et al., 1994; Munck, 1998; George and Bennett, 2001). Fortunately, these are precisely the strengths of statistical studies.

Future Directions in Case Study Methods

Just as the formalization of case study methods in the 1970s inspired a generation of more sophisticated research, the recent further refinements in these methods are likely to lead to still greater sophistication in their use in the social, biological, and even physical sciences. The increasingly evident complementarity of case study and statistical methods is likely to lead toward more collaborative work by scholars using various methods. The recent interest among rational choice theorists in using case studies to test their theories, for example, is an important step in this direction (Bates et al., 1998). Because case studies, statistical methods, and formal modeling are all increasingly sophisticated, however, it is becoming less likely that a single researcher can be adept at more than one set of methods while also attaining a cutting-edge theoretical and empirical knowledge of their field. Collaboration might therefore take the form of several researchers working together using different methods, or of researchers more self-consciously building on the findings generated by those using different methods. In either form, effective collaboration requires that even as they become expert in one methodological approach, scholars must also become conversant with alternative approaches, aware of their strengths and limits, and capable of an informed reading of their substantive results.

See also: Case Studies; Case Study: Logic; Case-Oriented Research; Contextual Studies: Methodology; Ethnography in Applied Social Research; Qualitative Methods, History of; Qualitative Research; Single-Case Designs: Methodology.

Bibliography

Bates, R.H., Greif, A., Rosenthal, J., Weingast, B., Levi, M., 1998. Analytic Narratives. Princeton University Press, Princeton, NJ.

Bennett, A., 2004. Case study methods: design, use, and comparative advantages. In: Sprintz, D.F., Wolinsky-Jahmias, Y. (Eds.), Models, Numbers, and Cases: Methods for Studying International Relations. University of Michigan Press, Ann Arbor, MI.

Bennett, A., Elman, C., 2004. Complex causal relations and case study methods: the example of path dependence. Political Analysis 14, 250–267.

Bennett, A., Elman, C., 2006. Qualitative research: recent developments in case study methods. Annual Review of Political Science 9, 455–476.
Collier, D., 1993. The comparative method. In: Finifter, A.W. (Ed.), Political Science: The State of the Discipline II. American Political Science Association, Washington, DC.
Collier, D., Mahoney, J., 1996. Insights and pitfalls: selection bias in qualitative research. World Politics 1, 56–91.
Eckstein, H., 1975. Case studies and theory in political science. In: Greenstein, F.I., Polsby, N.W. (Eds.), Handbook of Political Science, vol. 7. Addison-Wesley, Reading, MA.
George, A.L., 1979a. Case studies and theory development. In: Lauren, P.G. (Ed.), Diplomacy: New Approaches in Theory, History, and Policy. Free Press, New York.
George, A.L., 1979b. The causal nexus between cognitive beliefs and decision-making behavior: the 'operational code.'. In: Falkowski, L.S. (Ed.), Psychological Models in International Politics. Westview, Boulder, CO.
George, A.L., Bennett, A.O., 2001. Case Studies and Theory Development. MIT Press, Cambridge, MA.
Goldstone, J., 1997. Methodological issues in comparative macrosociology. Comparative Social Research 16, 107–120.
Goldthorpe, J., 1997. Current issues in comparative macrosociology. Comparative Social Research 16, 1–26.
King, G., Keohane, R., Verba, S., 1994. Designing Social Inquiry. Princeton University Press, Princeton, NJ.
Lakatos, I., 1976. Falsification and the growth of scientific research programs. In: Lakatos, I., Musgrave, A. (Eds.), Criticism and the Growth of Knowledge. Cambridge University Press, Cambridge, UK.
Lijphart, A., 1971. Comparative politics and the comparative method. American Political Science Review 65, 682–693.
Lijphart, A., 1975. The comparable cases strategy in comparative research. Comparative Political Studies 8, 158–177.
Little, D., 1998. Microfoundations, Method, and Causation. Transaction, New Brunswick, NJ.
Mahoney, J., 1999. Nominal, ordinal, and narrative appraisal in macro-causal analysis. American Journal of Sociology 104, 1154–1196.
McKeown, T., 1999. Case studies and the statistical world view. International Organization 1, 161–190.
Munck, G.L., 1998. Canons of research design in qualitative analysis. Studies in Comparative International Development 33, 18–45.
Njolstad, O., 1990. Learning from history? Case studies and the limits to theory-building. In: Gleditsch, N.P., Njolstad, O. (Eds.), Arms Races: Technological and Political Dynamics. Sage, Newbury Park, CA.
Przeworski, A., Teune, H., 1970. The Logic of Comparative Social Inquiry. Wiley-Interscience, New York.
Ragin, C.C., 1987. The Comparative Method: Moving beyond Qualitative and Quantitative Strategies. University of California Press, Berkeley, CA.
Ragin, C.C., 2000. Fuzzy-Set Social Science. University of Chicago Press, Chicago, IL.
Ragin, C.C., Becker, H.S., 1992. Introduction. In: Ragin, C.C., Becker, H.S. (Eds.), What Is a Case? Exploring the Foundations of Social Inquiry. Cambridge University Press, Cambridge, UK.
Rogowski, R., 1995. The role of theory and anomaly in social-scientific inference. American Political Science Review 89, 467–470.
Sartori, G., 1970. Concept misformation in comparative politics. American Political Science Review 64, 1033–1053.
Van Evera, S., 1997. Guide to Methods for Students of Political Science. Cornell University Press, Ithaca, NY.
Verba, S., 1967. Some dilemmas in comparative research. World Politics 20, 111–127.

Cash Transfers

Nicola Jones, Overseas Development Institute, London, UK
Elizabeth Presler-Marshall, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social protection programming in general, and cash transfers in particular, have been one of the successes of international development's 'war on poverty' since the 1990s. This article discusses the evidence base on the contribution of cash transfer programming to human capital development, economic growth, humanitarian protection, and social norm change. It also considers program governance modalities and challenges. The final section discusses the future direction of cash transfer programming.

Introduction

Social protection programming in general, and cash transfers (CTs) in particular, have been claimed as one of the development successes since the 1990s. Originally pioneered in Brazil and Mexico in the mid-1990s, CTs, which give money to target populations either conditionally (CCTs) or unconditionally (UCTs), have been shown to have significant effects in reducing household income poverty, improving human capital outcomes, and contributing to declining levels of inequality. Buttressed by robust evaluation evidence, CTs have now spread throughout the developing world and are reaching more than 750 million people in middle- and low-income countries (Arnold et al., 2011).

Distinct from safety nets, social protection can take a variety of forms, ranging from public works programs to social transfers and social insurance; but they all aim to act as a buffer against transitory and/or chronic poverty. CTs are one of the most widely favored social protection instruments, as they have proved to be effective in enhancing and smoothing household consumption and increasing uptake of basic services at scale across a wide range of developing countries. More recently, in the context of global financial crises and in light of discussions around international development goals post-2015, social protection is increasingly seen as essential – not only to tackle rising levels of risk and vulnerability, but also to promote social justice and social inclusion (**Figure 1**).

The growing popularity of CTs is partly due to their fungible nature, empowering beneficiaries to spend according to their priorities. However, CCTs, in particular, have gained political acceptability as they are usually contingent on beneficiary families accessing basic services – for example, sending their children to school and attending health clinics. Evidence is still mixed, however, as to how important the conditional aspect of such programs is, and whether positive effects are more related to general household economic strengthening (Baird et al., 2011). Because of this, and informed by

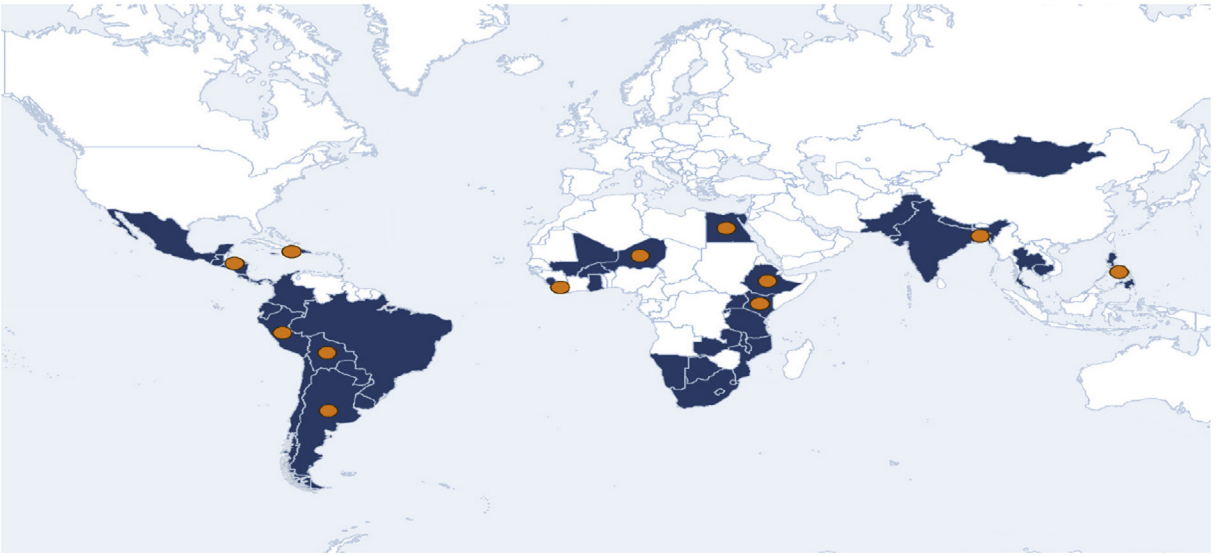

Figure 1 Cash transfer programs in Latin America, Africa, and Asia. Countries shaded in blue were implementing CTs as of 2012; countries with an orange circle began programs between 2009 and 2012. Only programs reaching more than 5000 people were included. Source: Global Savings and Social Protection Database, 2012 as cited in Sridharan, V., 2012. Enhancing the Impact of Cash Transfers: The Latest Trends toward Savings-Linked Social Protection. http://newamerica.net/sites/newamerica.net/files/policydocs/Trends0709.pdf.

commitments to a rights-based approach, a growing number of programs are unconditional. However, as there have been very few methodologically robust comparative evaluations, the research community remains undecided as to the relative merits of CCTs versus UCTs. (One of the few exceptions is the World Bank–funded CT experiment in Malawi in 2010–11, which included a conditional and an unconditional strand. The program aimed to improve girls' school attendance and sexual and reproductive health (see, for example, Baird et al., 2011).)

This article first discusses the evidence base on the contribution of CTs to human capital development (especially nutrition, health, and education), economic empowerment and growth, humanitarian protection, and social norm change. After looking at program governance modalities and challenges, considerations for the future direction of CT programming are offered.

Human Capital Development

Improving the uptake of basic services, especially among children, is one of the core features of most CTs, as the aim is to reduce child poverty and, by investing in the development of children's human capital, to reduce the risk of life course and intergenerational poverty transfers. Here, we briefly review the evidence on how CTs have contributed to improved access and outcomes in three key areas: nutrition, health, and education.

Nutrition

CTs have been shown to improve food consumption, particularly in the poorest countries. In Kenya, for example, participant households were found to consume greater quantities and more diverse foods – with the government reporting a 15% increase in the frequency of consumption of five food groups (meat, fish, milk, sugar, and fats) (Aston and Jones, 2012). Similarly, in Malawi, dietary diversity increased by nearly 25% for households receiving a CT (Audsley et al., 2010); in Ethiopia, beneficiaries of the government's Productive Safety Net Programme (PSNP) were significantly more likely to consume the required 1800 calories per day than non-beneficiaries. However, despite impacts on consumption, the final impact of CTs on measurable outcomes such as stunting and anemia is somewhat mixed – due to the difficulty in establishing the interplay between cash and other inputs such as clean water, sanitation, and health care (Fiszbein and Schady, 2009; **Figures 2** and **3**).

Dietary improvements are partly due to the fact that CTs allow households to purchase food directly. In Malawi, for example, 75% of transfers were spent on groceries (Arnold et al., 2011). While this is higher than the average of 50% found for most CTs, one synthesis study across sub-Saharan Africa found that in six of the seven programs reviewed, the cash was primarily spent on food (Adato and Bassett, 2008). CTs also contribute to dietary improvements by enabling landholding rural households to purchase livestock and other inputs required to boost agricultural productivity and better feed themselves. In Zambia, for example, participants in a pilot CT spent over one-fifth of their CT on livestock (Garcia and Moore, 2012), and in Ethiopia, PSNP participation has been found to increase livestock ownership by 7% (Gilligan et al., 2009).

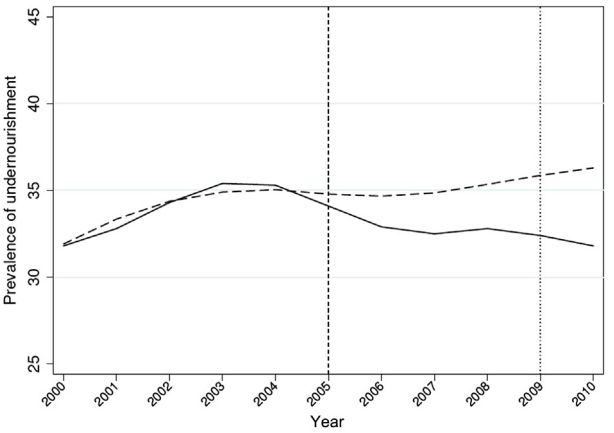

Figure 2 Prevalence of undernourishment, Kenya (solid line), and control (dashed line) – with vertical line representing introduction of CT. Source: d'Agostino, G., Pieroni, L., Scarlato, M., 2013. Social Protection and Food Security in Sub-Saharan Africa: An Evaluation of Cash Transfer Programmes. MPRA Paper No. 55265. Munich Personal RePEc Archive. http://mpra.ub.uni-muenchen.de/55265/ (accessed 09.19.14.).

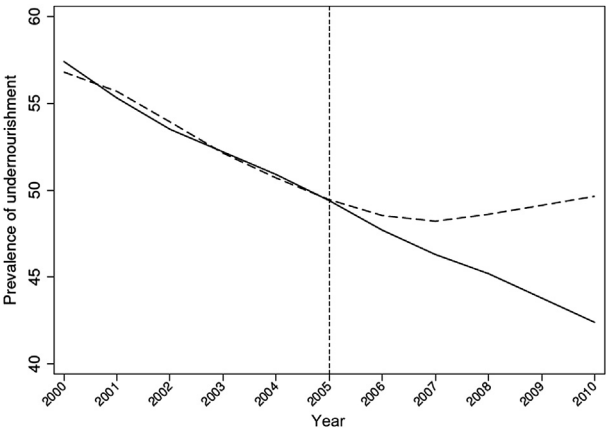

Figure 3 Prevalence of undernourishment, Ethiopia (solid line), and control (dashed line) – with vertical line representing introduction of CT. Source: d'Agostino, G., Pieroni, L., Scarlato, M., 2013. Social Protection and Food Security in Sub-Saharan Africa: An Evaluation of Cash Transfer Programmes. MPRA Paper No. 55265. Munich Personal RePEc Archive. http://mpra.ub.uni-muenchen.de/55265/ (accessed 09.19.14.).

While CTs increase household food consumption, better nutrition-related outcomes do not always follow. Indeed, Manley et al. (2013), in their analysis of 17 CT programs, found no significant impacts on children's height for age – a key indicator of child malnutrition. Similarly, in Kenya, despite better diets, Ward et al. (2010) found no impact on anthropometric indicators. Other studies have found positive impacts though. For example, in Nicaragua, Hoddinott and Bassett (2009) concluded that the Red de Protección Social CCT had extremely robust impacts on preschooler height, with children

in beneficiary communities 5.3% less likely to be stunted. Similarly, CTs have been associated with significant declines in global acute malnutrition in countries ranging from Myanmar to Niger (Bailey and Hedlund, 2012).

While the specific program components that lead to better outcomes are as yet unclear, evidence suggests that the better diets facilitated by CTs are more likely to translate into better nutritional outcomes when the cash is paired with complementary programming that targets other known deficits (Bailey and Hedlund, 2012; Hoddinott and Bassett, 2009). For example, given that the first 2 years of a child's life are critical to healthy growth and development, messaging directed at infant feeding practices is often key. Instructional components directed at hygiene and sanitation are also important to combat parasites and diarrheal illness. Finally, given that infant nutrition is intimately linked to maternal nutrition, the provision of nutritional supplements to pregnant and breast-feeding women can also have long-term impacts on health; supplements have reduced intrauterine growth retardation by nearly one-third (Hoddinott and Bassett, 2009).

Health

Not surprisingly, since many programs condition benefits on the uptake of certain health care services, CTs have improved uptake rates, particularly among the most disadvantaged communities, who are least likely to utilize such services (Fiszbein and Schady, 2009). However, impacts on health outcomes have been inconsistent (Fiszbein and Schady, 2009). Not only do services vary significantly in their quality, with obvious impacts on outcomes, but also the constellation of influences on health outcomes makes it challenging to isolate impacts – or even impact channels (Fiszbein and Schady, 2009). That said, given some strong successes, CTs have recently been expanded beyond their initial child health mandate and are now targeting diverse outcomes, including safe motherhood and HIV prevention.

An analysis of 13 CT programs, most of them in Latin America, found clear impacts on the use of preventative services such as immunization (Ranganathan and Lagarde, 2012). Long-term impacts are less clear, however. Rasella et al. (2013) found that Brazil's Bolsa Familia has reduced the under-5 mortality rate – in part, through its impacts on poverty-related causes such as malnutrition and diarrhea. Similarly, Ghana's Livelihood Empowerment Against Poverty program has been found to reduce morbidity among older children (Handa et al., 2013), while a small-scale, community-implemented CCT in Tanzania reduced illness rates among preschoolers by more than 10% (Evans et al., 2014).

A key finding on the impact of CTs on health outcomes is the importance of targeting – with the poorest groups experiencing the strongest impacts. UNICEF (2010), for example, found that the highest returns on investment in child health are in rural areas, where fertility rates tend to be higher, preventable child deaths are more common, and health care coverage is low. The importance of targeting was echoed by Manley et al. (2013) in their analysis of child nutrition outcomes, and by Barnham (2013), who found that Mexico's Oportunidades program achieved the greatest declines in infant mortality among groups with high rates of illiteracy and mortality.

The focus of some CCTs has been extended to safe motherhood, given poor progress in reaching the Millennium Development Goal 5 (reducing maternal mortality). Mexico's Oportunidades, and the Philippines' 4Ps (Pantawid Pamilyang Pilipino Program), now have elements promoting safe motherhood. Others, such as Nepal's Safe Delivery Incentive Programme and India's Janani Suraksha Yojana, focus entirely on safe motherhood, paying pregnant women if they attend prenatal clinics and are delivered by a skilled attendant. In their recent review of the effects of demand-side financing on maternity care, Murray et al. (2014) conclude that while the evidence is still slim, CTs can help improve uptake of maternal health services. In El Salvador, for example, Comunidades Solidarias Rurales has had robust effects on rates of institutional delivery (de Brauw and Peterman, 2011).

However, a cautionary note is required. While CTs can improve access to health care, better outcomes also depend on the availability of good-quality health services (Fiszbein and Schady, 2009). Murray et al. (2014) observe that CTs and other demand-side financing schemes do not necessarily lead to higher quality services, and can in fact reduce quality by increasing utilization rates, putting additional strain on already stretched systems. This means that supply-side investments are crucial in order to capitalize on the effects of CTs (Ranganathan and Lagarde, 2012).

The newest application of CTs to health-related outcomes – and one, which highlights the complex decision-making process that produces good health programming – is HIV prevention. An analysis of program impacts identified two pathways for CTs to reduce HIV risk: upstream, by reducing poverty and improving education; and downstream, by directly rewarding healthy behaviors (Pettifor et al., 2013). Though few studies have tracked HIV incidence due to concerns about ethics, the evidence overall suggests that both uptake of education services and sexual behaviors – particularly for girls and women – are malleable.

Upstream impacts are clearly potent. For example, Kenya's CT for orphans and other vulnerable children – which is unconditional and pays caregivers, not adolescents, $20 a month – found that beneficiaries had significantly reduced odds of sexual debut (35% for girls and 18% for boys) compared with their nonbeneficiary peers (Handa et al., 2014). Similarly, Malawi's Zomba CCT, which targeted school attendance, found that beneficiary girls were 60% less likely to become infected with HIV – primarily because participation reduced the odds of transactional sex (Baird et al., 2012). Girls who were enrolled in school had fewer older partners and less sex.

While downstream effects are more rarely measured, there is some evidence that direct conditioning may also work. For example, in Tanzania, a CT that offered payments of either $10 or $20 every quarter if tests for sexually transmitted infections came back negative, reduced the odds of disease after 1 year at the $20 level (de Walque et al., 2012). However, the lack of negative results at 4 and 8 months may mean that it takes time for participants to trust the conditioning – or to extricate themselves from complicated sexual relationships.

The complicated nature of personal decision-making is highlighted by Kohler and Thornton (2012) in their evaluation of an HIV-targeted CCT in Malawi. The program randomly

offered participants financial incentives at the end of 1 year, if they maintained their HIV status. Ironically, while the reward changed neither self-reported behavior nor serology results, men who received a reward were then more likely to engage in risky sex than men who had not received a reward. Women, on the other hand, were less likely to do so.

Education

As with health, CTs have proved to be critical for stimulating demand for basic education services. CT programs targeting education may be either conditional (i.e., children must attend school in order to receive the CT) or quasiconditional (whereby attendance is not necessarily monitored, but communities and households are strongly encouraged to use the cash to invest in children's schooling, as in Ghana and Kenya) (Jones et al., 2013). Global findings suggest that the impact on educational attendance and reduced absenteeism, often stemming from domestic and paid work pressures, is consistently significant (Fiszbein and Schady, 2009; **Figure 4**).

Although schooling is free in many of the low- and middle-income country contexts where CT programs have been rolled out, a range of other costs (e.g., uniforms, books, snacks) make continued access to school challenging, with inability to meet these costs a major cause of dropout for children from poor households. By contrast, the ability to purchase such items due to the CT is associated with confidence, self-esteem, and a sense of belonging to a peer group.

Evidence also suggests that dropping out of school can lead to involvement in more exploitative and dangerous activities, such as transactional sex or sex work, child labor, stealing, or begging (Jones et al., 2013). Similarly, in some contexts, CTs may prevent girls being withdrawn from school and forced into early marriage (Jones et al., 2013).

There is also evidence of some indirect spill-over effects, which reinforce the more direct and positive inputs from schools. Because of a combination of increased awareness about the importance of schooling and the availability of more cash (which, among other things, reduced men's need to migrate for economic reasons), greater parental involvement in children's education has been reported. For example, Jones et al. (2008) found that fathers enrolled in Peru's Juntos Program were becoming more involved in their children's education (benefiting from awareness-raising and parenting classes) and were supporting children's learning at home.

Despite these positive effects, however, there is also a risk that poverty and social class may undermine social relations and play a potentially divisive role in educational settings. Children from beneficiary families in Gaza and the West Bank, for example, reported feeling excluded, because being identified as poor led to discrimination from their peers and, in some cases, teachers (Pereznieto et al., 2014).

Furthermore, while increasing household income can be critical in terms of promoting enrollment and helping to prevent dropouts, the CT amount is often inadequate to improve educational attainment (Fiszbein and Schady, 2009). Part of the problem lies with the resource and capacity limitations of the home environment, but supply-side constraints, i.e., the quality of education provided, are also key. Concerns about the quality of teaching, teacher absenteeism, school-based violence, and limited or poor-quality school psychosocial counseling services are all issues that have emerged in studies on child-sensitive social protection, and which potentially undermine the positive dividends of CTs in education (e.g., Jones et al., 2008; Pereznieto et al., 2014).

Economic Growth

Propoor CTs are also linked to broader economic growth (FAO, 2014; Barrientos, 2012). While it has traditionally been assumed that CTs are a long-term investment, with benefits expected to accrue to the broader economy through improved health and education outcomes for today's children, evidence suggests that CTs also have immediate positive effects on economic growth. These are primarily through increases in

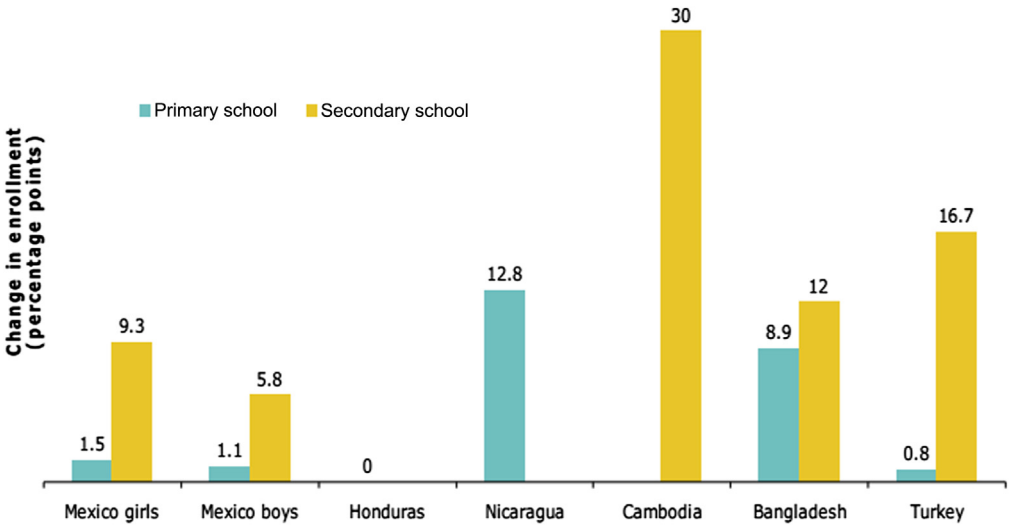

Figure 4 Impact of CCTs on primary and secondary enrollment. Source: Arnold, C., Conway, T., Greenslade, M., 2011. Cash Transfers Literature Review. Department for International Development, Policy Division, London.

beneficiary household consumption, which have follow-on impacts on local economies, and strengthening of recipient asset portfolios, which facilitates self-employment. However, CTs may also affect growth at the highest levels – insofar as they help build a more secure state, they encourage growth, since security is essential to local, national, and international economic investment (Samson, 2009).

Evidence suggests that even small CT programs can stimulate local demand, as poor households gain more purchasing power, which can, in turn, stabilize prices by incentivizing farmers, making trade more competitive. For instance, Malawi's Mchinji CT pilot, targeted at the labor-constrained ultra-poor, averaged only $14 a month but increased spending significantly on food and nonfood items. In an evaluation of the program's impact on economic growth, Miller (2011) found that beneficiaries were important customers for local businesses and were especially important for smoothing sales during the rainy season. The Food and Agriculture Organization of the United Nations (2014), in its evaluation of the impacts of numerous CT programs on growth in sub-Saharan Africa, notes that this consumption helps spread – and multiply – impact, as local businesses are primarily owned by noneligible households. Indeed, models suggest that multiplier effects can be as high as 2.5, though complementary interventions may be required to maximize these (Miller, 2011).

CTs also stimulate economic growth by helping poor households acquire productive assets that enable them to build more secure futures. As the FAO (2014) notes, self-employment is often the exit route from poverty. Even small transfers, if regular, can help poor households manage risk and invest in production. For example, Miller (2011) reports that in Malawi, $14 a month allowed households to increase livestock holdings and agricultural production. Similarly, the FAO (2014) notes that CTs in Zambia, Malawi, and Kenya increased household investment in agricultural assets and inputs; qualitative research in Ghana, Kenya, and Zimbabwe found that CTs increased investment in economic activities – with female-headed households benefiting most in some cases.

Larger transfers may offer households even greater opportunities for investment. In Kenya, for example, Haushofer and Shapiro (2013) report that an unconditional CT, experimental in its design, randomly offered households a one-off grant of either $300 or $1100; beneficiaries increased their livestock holdings by over 50% and their income from self-employment by nearly 40%. Similarly, in Uganda, where the government distributed unconditional CTs of $380 (on average) to more than 5000 young people for enterprise development – worth about a year's average income – Blattman et al. (2013) found that recipients invested most of the cash in training and business assets. Four years later, recipients were 65% more likely to have a skilled trade and had real earnings nearly 50% higher than nonbeneficiaries. Interestingly, impacts were highly gendered; women's incomes raised approximately 2.5 times as much as men's (84% vs 31%) (Blattman et al., 2013).

Barrientos (2012) notes that in order to maximize growth impacts, the design of CT programs is crucial. Transfers must be regular and reliable, and last long enough to rebuild livelihoods. They should also be large enough, as noted above, to encourage investment rather than merely smoothing daily consumption. Eligibility conditions should not encourage asset depletion, which ultimately makes it far more challenging for households to exit poverty. Instead, CTs should be paired with complementary programs that facilitate asset accumulation.

Humanitarian Protection

CTs have a broad set of applications in humanitarian contexts. They have been successfully used in short-term situations ranging from environmental disasters, such as the 2010 floods in Pakistan and the 2011 drought in Somalia, to the reintegration of refugees and former combatants in Burundi and Liberia. They are also addressing longer-term fragility in places such as the occupied Palestinian territories (oPt). While in-kind support (in the form of food, medical supplies, etc.) remains central in some cases and important in many others, experience increasingly demonstrates the appropriateness and efficacy of CTs – even in disaster and fragile contexts (Bailey and Hedlund, 2012).

CTs, whether conditional or unconditional, offer numerous advantages over more traditional forms of humanitarian aid. They can, for example, be planned and implemented far more quickly than other forms of support (Global Humanitarian Assistance, 2012; Carpenter et al., 2012). Metz et al. (2012) report that in South Sudan, while food distributions were only punctual half of the time and were often delayed for over a month, cash payments were typically on time or very close to being so. Similarly, after the floods in Pakistan, the government was able to promptly issue prepaid Visa cards, worth over $200 each, to millions of households (European Commission, 2013). Given the severe damage to infrastructure, in-kind aid would undoubtedly have taken far longer to reach recipients.

Cash can be a more efficient form of distribution. In Jordan and Syria, for example, refugees are able to access their cash benefits via local bank ATMs, which saves the time and expense of creating new distribution systems (Harvey and Bailey, 2011). In Somalia, cash was transferred through local traders and money transfer companies; aid agencies worked closely with local vendors to ensure that food supplies would be available in normal markets and to prevent inflation (Hedlund et al., 2013).

Given gains in efficiency, CTs can be a cost-effective response to emergency situations. In Ecuador, where the World Food Programme offered assistance to Colombian refugees in 2011, CTs were found to be the least expensive way of delivering aid. The marginal cost of direct food distribution was nearly 4 times higher ($11.50 vs $3.03, with food vouchers costing $3.30) (Hidrobo et al., 2012). Metz et al. (2012) concluded that direct food aid was less cost-effective than CTs in South Sudan as well.

CTs allow for a greater level of flexibility than most other forms of aid. That flexibility can be vital for programs and households alike, allowing a more targeted response that protects local livelihoods while also providing a safety net. For instance, after the 2009 earthquake in Indonesia, Oxfam found that while people's livelihoods had not been disrupted, the need to find shelter was precluding adequate participation in income-generating activities. Cash was seen as the best way to help families meet their immediate needs so that they could

resume productive activities to ensure their longer-term survival (Harvey and Bailey, 2011). Similarly, in Kenya, a voucher program not only improved refugee diets, but also strengthened and expanded local markets, positively impacting other people's livelihoods (Harvey and Bailey, 2011).

CTs offer greater flexibility for households too, allowing each to prioritize spending based on their specific needs. While food and shelter are clear priorities, education is also important for families with school-age children. In Southern and Central Somalia, UNICEF's CT reunified families – improving children's school performance as a result – and saved money by reducing demand for expensive institutional care. The flexibility to pay school and medical fees, as well as pay down debt, was also judged crucial by Somali recipients (Hedlund et al., 2013).

While CTs have immediate protective benefits in humanitarian contexts, there is also evidence that they can segue into long-term preventive and promotive programs for vulnerable populations. In Ethiopia, for example, the PSNP, which provides cash (and food) to food-insecure households in exchange for labor on community projects, grew out of decades of emergency food aid that was having no lasting impact on household resilience. Similarly, programs in Kenya that were started by nongovernment organizations (NGOs) in response to high food prices are seeking to influence larger social protection debates (Mohanty, 2010).

This is not to argue that CTs are a panacea for humanitarian protection. Indeed, they share many of the weaknesses of more traditional aid programs – weaknesses that are often driven by the need for urgency in humanitarian contexts. For example, after the 2010 election violence in Liberia, a program offering cash for work or an unconditional grant to adolescents to keep them out of trouble was implemented without respect to national law. Work was offered to children aged 15 and above, despite the law prohibiting employment of children under 16 (Thompson, 2012). Similarly, postdisaster responses all too often suffer from limited coordination. With many agencies working in multiple ways in limited areas, it can be difficult to monitor impacts and draw conclusions. Where the need is great and coordination is limited, such as after the 2010 floods in Pakistan and the 2004 tsunami, establishing what worked and what did not can remain unclear for many years.

Effective targeting and communication can be especially urgent but fraught with difficulty in fragile situations, in which most people are vulnerable and aid is limited. In these contexts, careful targeting is vital, as is clear communication, so that communities understand how benefits are assigned – something that seems particularly rare in humanitarian situations. In Yemen and the oPT, for example, CT recipients often report that they have no idea where their cash comes from, or why they receive it and others do not (Jones et al., 2013). Similarly, in Sierra Leone, the social safety net, far from improving relationships between communities and the state, engendered considerable mistrust (Osofisan, 2011).

CTs, like other aid programming, need to pay careful attention to exclusion and bias on the basis of gender or ethnicity. Oxfam GB and Concern Worldwide (2011), reviewing the gendered impacts of three crisis-oriented CTs (in Kenya, Zimbabwe, and Indonesia), found that while programs can improve intrahousehold relations where women are the targeted beneficiary, overall, CTs reinforced women's traditional roles. Beneficiaries reported that not only did staff emphasize women's responsibility for children – but that they also highlighted men's irresponsibility, marginalizing men's role as fathers.

Social Norm Change

Over time, CTs may also facilitate a shift in constraining social norms. For example, they may empower women to make decisions and handle money, encourage men to spend more time caring for their children, and discourage early marriage. While it takes time to change underlying beliefs about what people ought to do – i.e. injunctive norms – there is increasing evidence that CTs are already changing the descriptive norms that govern people's actions.

In the case of Peru's Juntos program, for example, Jones et al. (2008) found that the program has had some impacts on gender norms. It is enabling women's decision-making and changing local understanding about women's mobility and community participation. There is also evidence that the program is reducing domestic violence and encouraging men to assume greater household responsibility (Jones et al., 2008; see also Hidrobo and Fernald, 2013). Even small steps are likely to have multiplier effects over time, particularly as children witness their mothers assuming productive as well as reproductive roles, and becoming more confident in their ability to contribute to the household economy. Similar results have been reported in Brazil, where women report greater latitude in decision-making, including on contraception.

While evidence is mixed, CTs may also be playing a role in ending harmful practices such as child marriage. In Ethiopia, for example, a program that offered families livestock in exchange for delaying their daughters' marriage was found to prevent marriages of the youngest girls; rates for girls aged 10–14 fell by 90%, though the rate for older girls increased (Erulkar and Muthengi, 2009). There is also evidence that programs in India designed to encourage girls' education and discourage early marriage have been successful at keeping girls in school. Nanda et al. (2014) report that beneficiary girls are nearly 25% more likely to be enrolled in school after the age of 15. Though final evaluations are still in progress, given that education offers girls some protection from early marriage, it seems likely that the overall effects are positive.

However, observed changes in social norms have tended to be small and confined to specific subgroups of CT participants, for several reasons. First, because CTs typically target women as beneficiaries, they often unintentionally reinforce traditional gender roles by relying on women's unpaid work and time to ensure fulfillment of conditions or to attend program-related awareness-raising sessions (Holmes and Jones, 2013; Oxfam GB and Concern Worldwide, 2011). Second, there is some evidence that as programming expands women's horizons, men respond by holding on more tightly to traditional roles. For example, Camacho and Rodriguez (2012) found that not only did Colombia's Familias en Acción program fail to empower women on five measured dimensions, but also that after enrollment, women beneficiaries had a lower probability of being the sole decision-maker about children's education.

They speculate that when school attendance became a condition for receiving the transfer, men took control. Similarly, there is also evidence that (at least in the short term) CTs can worsen domestic violence. Hidrobo and Fernald (2013), for example, found that for Ecuadorian women with less than sixth grade education, CT participation tended to increase emotional violence if their level of education was higher than their partner's.

Cash alone is rarely sufficient to alter deeply held beliefs or change age-old traditions. In India, even where CT participation has increased girls' schooling, there is little evidence that participation – or education – has fundamentally altered their self-efficacy (Nanda et al., 2014). Similarly, in Bangladesh, where the female secondary school stipend has contributed to rising enrollment rates for girls (who are now more likely than boys to attend school), two-thirds of girls still marry before their 18th birthday (UNICEF, 2011).

Program Governance Modalities and Challenges

For the chronically poor and most vulnerable, who are least likely to benefit from economic growth, politics and political change are often the only channels through which they can challenge poverty. National governments' institutional capacity to implement large-scale, poverty-targeted CT programs is therefore key, as are program accountability mechanisms and existing state–citizen relations.

Unfortunately, social protection schemes, including CTs, are often located within comparatively weak ministries, with limited financial resources and high levels of donor dependence. This is especially the case in low-income contexts in sub-Saharan Africa. In a study on CTs in Kenya, Mozambique, Palestine, Uganda, and Yemen, Jones et al. (2013) also found an important disconnect between central and local levels in terms of the rationale for program reforms and procedures. This is often due to weak information flows and human resource capacity deficits at subnational level, with too few (and inadequately trained) program staff to deal with the multiple vulnerabilities facing beneficiaries. For instance, in Gaza, social workers were responsible for managing up to 800 beneficiary cases; in Kenya, community-level work was done entirely by volunteers, and in Mozambique, by poorly trained and remunerated facilitators. Moreover, staffs are not always competent to deal effectively and sensitively with specific vulnerabilities faced by people with disabilities, children, or older people, for example.

Because of the cross-sectoral nature of social protection, it is also difficult to establish effective coordination mechanisms across government, development partners, and NGOs. Single registry databases are beginning to tackle this, recording all information about registered beneficiaries and their program entitlements in order to reduce duplication of effort and promote synergies. In Palestine, for example, following the Brazilian and Chilean models, a computerized single registry database is gradually being made available to all relevant ministries and is expected to significantly strengthen program coordination.

Another critical approach aimed at improving coordination among program designers, implementers, and beneficiaries is to embed accountability mechanisms within CT programs, creating spaces for interaction and information exchange between staff and beneficiaries. Current CTs vary considerably in this respect. In Mozambique, permanentes (local volunteers who take on a program liaison role) and neighborhood committees are important links between beneficiaries and program staff; however, respondents also highlighted that they often serve as gatekeepers. In Uganda, program implementers are trying to engage with local leaders who organize regular community meetings, as well as interacting with beneficiaries on payment days (Jones et al., 2013). Program design has gone a step further in Kenya, where since 2011, community-level committees have been channeling citizen inputs into program decision-making and monitoring (Jones et al., 2013).

Many programs have grievance mechanisms for beneficiaries and nonbeneficiaries, but Jones et al. (2013) found that implementation appears to be weak, partly because in sub-Saharan Africa, CTs are largely seen as a gift – either from political leaders or God – leading to a general reluctance to complain. In other instances, the grievance process was simply too tardy to inspire confidence, or people were concerned about reprisals (Jones et al., 2013).

Finally, delivering regular and predictable social protection support can buttress a 'social contract' – a reciprocal relationship between citizens and their government. In countries where there is relatively good information provision, and where community leaders and local program implementers have not been excessively politicized (e.g., Brazil and Uganda), CTs appear to have consolidated citizen–state ties. However, where programs have been politicized at the local level and are plagued by irregular payments or corruption (e.g., Mozambique and Yemen), confidence in the state has generally not improved, and has even been further undermined. In the oPT, while the Palestinian National Cash Transfer Programme is recognized as a critical coping strategy, there is little awareness of the Palestinian Authority's contribution, even though it funds approximately half of the budget. Many beneficiaries regard the support as being from 'Europeans,' effectively undermining the political mileage that could be gained to help strengthen state–citizen relations.

Future Directions

CTs have undoubtedly contributed to poverty and vulnerability reduction at scale in middle- and low-income countries alike. Long-established programs such as Brazil's Bolsa Familia, Mexico's Oportunidades, and South Africa's Child Grant have achieved meaningful and sustained improvements in the lives of millions of program beneficiaries and continue to protect families from on-going risks and stresses. However, there is still much to do to ensure that programs facilitate a more sustainable exit from poverty and multidimensional vulnerability. Priorities include the need to pay greater attention during program design and implementation to the intersection between social exclusion and poverty (e.g., through developing more gender- or disability-sensitive programs); the importance of developing more embedded monitoring and evaluation systems that draw on qualitative and participatory approaches alongside quantitative methods so as to more effectively tailor programs in the context of profound resource shortages; and the imperative of forging

stronger linkages to complementary services and programs (e.g., vocational training, psychosocial counseling, and microcredit) that can help tackle the multidimensional nature of poverty and vulnerability. Here programs such as Chile's Chile Solidario, which provides families not only with regular cash supplements but equally importantly also with tailored support overseen by a dedicated social worker to help households tackle challenges they jointly identify (e.g., domestic violence, drug addiction, lack of marketable skills, inadequate housing) offers an innovative approach toward which more narrow CT program models could aspire.

See also: Children, Rights of: Cultural Concerns; Conditional Cash Transfers: Influence on Marriage and Fertility; Gender and Health Care; Gender, Economics of; Health in Low-Income Countries; Inequality: Comparative Aspects; Justice and Health; Social Inequality and Schooling; Social Inequality in History (Stratification and Classes); Time Use and Gender; Water, Health and Social Inequality; Welfare Economics; Welfare Policy Comparisons; Welfare State, History of.

Bibliography

Adato, M., Bassett, L., 2008. What Is the Potential of Cash Transfers to Strengthen Families Affected by HIV and AIDS? A Review of the Evidence on Impacts and Key Policy Debates. Submitted to the Joint Learning Initiative on Children and AIDS (JLICA). International Food Policy Research Institute (IFPRI), Washington, DC.

d'Agostino, G., Pieroni, L., Scarlato, M., 2013. Social Protection and Food Security in Sub-Saharan Africa: An Evaluation of Cash Transfer Programmes. MPRA Paper No. 55265. Munich Personal RePEc Archive. http://mpra.ub.uni-muenchen.de/55265/ (accessed 09.19.14.).

Arnold, C., Conway, T., Greenslade, M., 2011. Cash Transfers Literature Review. Department for International Development, Policy Division, London.

Aston, T., Jones, N., 2012. Social Protection and Child Malnutrition: Kenya Briefing. Save the Children, London.

Audsley, B., Halme, R., Balzer, N., 2010. Comparing cash and food transfers: a cost-benefit analysis from rural Malawi. In: Omamo, S.W., Gentilini, U., Sandström, S. (Eds.), Revolution: From Food Aid to Food Assistance – innovations in Overcoming Hunger. World Food Programme, Rome.

Bailey, S., Hedlund, K., 2012. The Impact of Cash Transfers on Nutrition in Emergency and Transitional Settings: A Review of Evidence. Humanitarian Policy Group, Overseas Development Institute, London.

Baird, S., McIntosh, C., Özler, B., 2011. Cash or condition? evidence from a cash transfer experiment. Quarterly Journal of Economics 126, 1709–1753.

Baird, S., Garfein, R., McIntosh, C., Ozler, B., 2012. Effect of a cash transfer programme for schooling on prevalence of HIV and herpes simplex type 2 in Malawi: a cluster randomised trial. Lancet 379, 1320–1329.

Barnham, T., 2013. A healthier state: the effect of conditional cash transfers on neonatal and infant mortality in rural Mexico. Journal of Development Economics 94, 74–85.

Barrientos, A., 2012. Social transfers and growth: what do we know? what do we need to find out? World Development 40 (1), 11–20.

Blattman, C., Fiala, N., Martinez, S., 2013. The Economic and Social Returns to Cash Transfers: Evidence from a Ugandan Aid Program. http://cega.berkeley.edu/assets/cega_events/53/WGAPE_Sp2013_Blattman.pdf (accessed 05.14.14.).

de Brauw, A., Peterman, A., 2011. Can Conditional Cash Transfers Improve Maternal Health and Birth Outcomes? Evidence from El Salvador's Comunidades Solidarias Rurales. IFPRI Discussion Paper 01080. International Food Policy Research Institute, Washington, DC.

Camacho, A., Rodriguez, C., 2012. Who's the Boss at Home after Receiving Conditional Cash Transfers? Department of Economics, Universidad de Los Andes.

Carpenter, S., Slater, R., Mallett, R., 2012. Social Protection and Basic Services in Fragile and Conflict-Affected Situations. ODI Working Paper 8. Overseas Development Institute, London.

Erulkar, A., Muthengi, E., 2009. Evaluation of Berhane Hewan: a program to delay child marriage in rural Ethiopia. International Perspectives on Sexual and Reproductive Health 35 (1), 6–14.

European Commission, 2013. The Use of Cash and Vouchers in Humanitarian Crises. DG ECHO Funding Guidelines. http://ec.europa.eu/echo/files/policies/sectoral/ECHO_Cash_Vouchers_Guidelines.pdf (accessed 05.14.14.).

Evans, D., Hausladen, S., Kosec, K., Reese, N., 2014. Community Based Conditional Cash Transfers in Tanzania. The World Bank, Washington, DC.

FAO, 2014. From Protection to Production: The Role of Cash Transfer Programmes in Fostering Broad-Based Economic Development in Sub-Saharan Africa. FAO, Rome. http://www.fao.org/economic/ptop/home/en/ (accessed 05.14.14.).

Fiszbein, A., Schady, N., 2009. Conditional Cash Transfers: Reducing Present and Future Poverty. The International Bank for Reconstruction and Development/The World Bank, Washington, DC.

Garcia, M., Moore, C., 2012. The Cash Dividend: The Rise of Cash Transfer Programs in Sub-Saharan Africa. The World Bank, Washington, DC.

Gilligan, D.O., Hoddinott, J., Kumar, N.R., Taffesse, A.S., 2009. An Impact Evaluation of Ethiopia's Productive Safety Nets Program. International Food Policy Research Institute, Washington, DC.

Global Humanitarian Assistance, 2012. Tracking Spending on Cash Transfer Programming in a Humanitarian Context. Briefing. Development Initiatives, Somerset. http://www.globalhumanitarianassistance.org/wp-content/uploads/2012/03/cash-transfer-financing-final.pdf (accessed 05.14.14.).

Handa, S., Park, M., Darko, R., Osei-Akoto, I., Davis, B., Diadone, S., 2013. Livelihood Empowerment against Poverty Program Impact Evaluation. University of North Carolina, Chapel Hill.

Handa, S., Halpern, C.T., Pettifor, A., Thirumurthy, H., 2014. The government of Kenya's cash transfer program reduces the risk of sexual debut among young people age 15–25. PLoS ONE 9 (1), e85473. http://dx.doi.org/10.1371/journal.pone.0085473.

Harvey, P., Bailey, S., 2011. Cash Transfer Programming in Emergencies. Good Practice Review 11. Humanitarian Practice Network, Overseas Development Institute, London.

Haushofer, J., Shapiro, J., 2013. Policy Brief: Impacts of Unconditional Cash Transfers. http://web.mit.edu/joha/www/publications/Haushofer_Shapiro_Policy_Brief_UCT_2013.10.22.pdf (accessed 05.14.14.).

Hedlund, K., Majid, N., Maxwell, D., Nicholson, N., 2013. Final Evaluation of the Unconditional Cash and Voucher Response to the 2011–12 Crisis in Southern and Central Somalia. UNICEF Somalia, Nairobi.

Hidrobo, M., Fernald, L., 2013. Cash transfers and domestic violence. Journal of Health Economics 32 (1), 304–319.

Hidrobo, M., Hoddinott, J., Margolies, A., Moreira, V., Peterman, A., 2012. Impact Evaluation of Cash, Food Vouchers, and Food Transfers among Colombian Refugees and Poor Ecuadorians in Carchi and Sucumbíos, Final Report. International Food Policy Research Institute, Washington, DC.

Hoddinott, J., Bassett, L., 2009. Conditional Cash Transfer Programs and Nutrition in Latin America: Assessment of Impacts and Strategies for Improvement. Working Paper #09. Food and Agriculture Organization of the United Nations (FAO)/Hunger-Free Latin America and the Caribbean Initiative.

Holmes, R., Jones, N., 2013. Gender and Social Protection in the Developing World: Beyond Mothers and Safety Nets. Zed Books, London.

ILO, 2013. Cash Transfer Programmes, Poverty Reduction and Empowerment of Women: A Comparative Analysis. Experiences from Brazil, Chile, India, Mexico and South Africa. Working Paper 4/2013.

Jones, N., Samuels, F., Malachawska, A., 2013. Holding Cash Transfers to Account: Beneficiary and Community Perspectives. Overseas Development Institute, London.

Jones, N., Vargas, R., Villar, E., 2008. Cash transfers for children in Peru: tackling the multi-dimensionality of poverty and vulnerability. Environment and Urbanization 20 (1), 255–273.

Kohler, H.P., Thornton, R., 2012. Conditional cash transfers and HIV/AIDS prevention: unconditionally promising? World Bank Economic Review 26, 165–190.

Manley, J., Gitter, S., Slavchevska, V., 2013. How effective are cash transfers at improving nutritional status? World Development 48, 133–155.

Metz, M., Biel, M., Kenyi, H., 2012. Comparing the Efficiency, Effectiveness and Impact of Food and Cash for Work Interventions: Lessons Learned from South Sudan. Deutsche Gesellschaft fur Internationale Zusammenarbeit (GIZ).

Miller, C., 2011. Cash transfers and economic growth: a mixed methods analysis of transfer recipients and business owners in Malawi. Poverty and Public Policy 3 (3), 1–36.

Mohanty, S., 2010. Nairobi Urban Social Protection Programme. Oxfam.

Murray, S., Hunter, B., Bisht, R., Ensor, T., Bick, D., 2014. Effects of demand-side financing on utilisation, experiences and outcomes of maternity care in low- and middle-income countries: a systematic review. BMC Pregnancy and Childbirth 14, 30.

Nanda, P., Datta, N., Das, P., 2014. Impact of Conditional Cash.

Osofisan, W., 13–15 April 2011. Towards Strengthening State-Citizen Relationship in Fragile States and Environments: The Role of Cash Transfer Programmes. International Conference: Social Protection for Social Justice. Institute of Development Studies, UK.

Oxfam GB and Concern Worldwide, 2011. Walking the Talk: Cash Transfers and Gender Dynamics. https://www.concern.net/sites/default/files/resource/2011/05/cash_transfers_in_gender_relations.pdf (accessed 05.14.14.).

Pereznieto, P., Jones, N., Hamad, B., Shaheen, M., 2014. Effects of the Palestinian National Cash Transfer Programme on Children: A Mixed Methods Analysis. Overseas Development Institute, London.

Pettifor, A., MacPahil, C., Nguyen, N., Rosenberg, M., 2013. Can money prevent the spread of HIV? a review of cash payments for HIV prevention. AIDS and Behavior 16, 1729–1738.

Ranganathan, M., Lagarde, M., 2012. Promoting healthy behaviors and improving health outcomes in low and middle income countries: a review of the impact of conditional cash transfer programmes. Preventive Medicine 55, S95–S105.

Rasella, D., Aquino, R., Santos, C., Paes-Sousa, R., Barreto, M., 2013. Effect of a conditional cash transfer programme on childhood mortality: a nationwide analysis of Brazilian municipalities. Lancet 382, 57–64.

Samson, M., 2009. Social cash transfers and pro-poor growth. In: Promoting Pro-Poor Growth: Social Protection. OECD. http://www.oecd.org/development/povertyreduction/43280571.pdf (accessed 05.14.14.).

Sridharan, V., 2012. Enhancing the Impact of Cash Transfers: The Latest Trends toward Savings-Linked Social Protection. http://newamerica.net/sites/newamerica.net/files/policydocs/Trends0709.pdf.

Thompson, H., 2012. Cash and Child Protection: How Cash Transfer Programming Can Protect Children from Abuse, Neglect, Exploitation and Violence. Save the Children UK, London.

UNICEF, 2011. The State of the World's Children 2011: Adolescence: An Age of Opportunity. United Nations Children's Fund, New York.

UNICEF, 2010. Narrowing the Gaps to Meet the Goals. United Nations Children's Fund, New York. http://www.unicef.org/media/files/Narrowing_the_Gaps_to_Meet_the_Goals_090310_2a.pdf (accessed on 05.14.14.).

de Walque, D., Row, W., Nathan, R., et al., 2012. Incentivising safe sex: a randomised trial of conditional cash transfers for HIV and sexually transmitted infection prevention in rural Tanzania. BMJ Open 2, e000747.

Ward, P., Hurrell, A., Visram, A., Riemenschneider, N., Pellerano, L., O'Brien, C., MacAuslan, I., Willis, J., 2010. Cash Transfer Programme for Orphans and Vulnerable Children (CT-OVC), Kenya Operational and Impact Evaluation 2007–2009. Final Report. Oxford Policy Management, Oxford. http://www.unicef.org/evaluation/files/OPM_CT-OVC_evaluation_report_july2010-final_Kenya_2010-019.pdf (accessed 05.14.14.).

Relevant Websites

https://www.gov.uk/government/organisations/department-for-international-development – The Department for International Development (DFID) leads the United Kingdom's work to end extreme poverty, including through social protection.

http://www.ids.ac.uk/idsresearch/centre-for-social-protection – The Institute of Development Studies (IDS) Centre for Social Protection supports the development of social protection systems that are comprehensive, long-term, sustainable, and pro-poor.

http://www.odi.org.uk/programmes/social-protection – The Overseas Development Institute (ODI)'s Social Protection programme aims to support the design and implementation of effective social protection policy and programming in poor countries.

http://www.unicef.org/socialpolicy/index_socialprotection.html – The United Nations Children's Fund (UNICEF) works to improve children's lives in 191 countries.

http://www.worldbank.org/en/topic/socialprotectionlabor – The World Bank provides technical and financial assistance to developing countries in order to reduce extreme poverty, including through social protection.

Caste

Veena Das, Johns Hopkins University, Baltimore, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The career of caste as an anthropological concept provides a fascinating terrain on which to examine how anthropological objects come to be invented and stabilized, and the relation they bear to patterns of sociality, needs of government, and production of knowledge. Many textbooks in sociology and social anthropology represent caste as a fundamental institution of Indian society. In this view, castes are ranked in a hierarchical order and are governed by rules of endogamy, commensality, and purity–pollution that put strict limits on exchange of women and food between castes. The hierarchical order, however, permits the flow of women from lower to higher castes through marriage and the flow of food from higher to lower castes. The caste system is seen to correspond to a rough division of labor, although it is recognized that certain occupations such as agriculture cut across different castes. This textbook picture of the caste system, although influential, has come under serious scrutiny in recent years as being not only ahistorical but also ignoring the power–knowledge axis in the production of social science concepts.

Caste as the Ideology of Indian Society

Louis Dumont's (1980) *Homo Hierarchicus* has long been regarded as an outstanding contribution to the understanding of caste. Dumont argued that principles of hierarchy and holism were central for explaining the caste system. The principle of hierarchy in India, he proposed, was based upon the religious opposition between pure and impure – pollution incurred in the biological processes of life and death was removed in India not through processes of reciprocity (I bury your dead – you bury mine) but through principles of hierarchy. The task of removal of pollution was assigned to the lower castes who became permanently imbued with it. Thus, the separation between castes as well as their hierarchical ordering could be derived from the opposition between pure and impure. The scheme had the simplicity and elegance to make the bewildering diversity of Indian civilization immediately knowable, especially to Western readers.

Dumont's characterization of Indian society has been challenged on the ground that what he saw as a timeless ideology was itself a result of certain practices of classification and enumeration instituted in the context of colonial administration that gave a dominant place to Brahmanical texts as representatives of Indian society. An important intervention made by McKim Marriott (1990) needs mention here. Marriott provided a significant alternative to Dumont's formulation, arguing that different transactional strategies defined the position of different castes – it was not a simple case of hierarchy versus equality but rather of a universe governed by a complex set of rules and strategies regarding matching, mixing, and marking through which different regional and local configurations of castes were generated. What was at stake for both Dumont and Marriott, despite their differences, was the representation of India as the 'other' of the modern West. They were much less interested in either the concrete historical processes through which institutions were formed or the contemporary changes in the caste system. It is instructive to compare this with the way that caste was rendered in the work of the Indian anthropologist, M.N. Srinivas. Srinivas's stake in the local and his deep concern with the way caste was shaping Indian democratic politics distinguish him from these authors.

The Stake in the Local

The decade of the 1950s saw the emergence of 'village studies' as an important genre of anthropological writing in India. For Srinivas, the village was an indispensable site for fieldwork because of the importance he attached to what he called the 'field view' as distinct from the overly textual view of the caste system. The two important concepts that Srinivas formulated with far-reaching consequences for the study of caste were 'sanskritization' and 'dominant caste.'

Srinivas contended that a preoccupation with textual sources had led many to assume that the caste system was a rigid hierarchical system. Attention to historical and ethnographic processes on the ground, on the other hand, showed that while individual mobility was severely restricted in the caste system, group mobility had always been possible for castes in the middle rungs through the process of sanskritization. This process referred to the ability of a caste group, which had achieved economic or political mobility, to stake a claim to a higher status by recasting its customs and rituals to correspond more closely to Brahmanical rituals and practices. The process of change in ritual practices and social customs was not open to everyone. However, when a caste was successful in acquiring a measure of economic or political power through physical mobility, opening up of markets, or changed political alignments, it could translate this into claims to a higher caste status. With the decline in traditional kingship brought on by the colonial rule, such functions were shifted to colonial authorities. It is one of the exquisite ironies of the colonial rule that when caste began to be recorded systematically in the census, it became a major source for claiming of higher status by various caste groups who adduced their sanskritic practices or their myths of origin as evidence of their higher status.

Unlike the concept of *sanskritization*, which assumed an interaction between local and regional levels, the concept of

the dominant caste emerged out of an understanding of local and especially agricultural society. Srinivas argued that the ritual hierarchy of caste notwithstanding, the dominant role in village life was played by landowning peasant-proprietor castes who were rarely Brahmins. The relation between these rich peasant castes and the other castes in the village was characterized as a form of patron–client relations. Thus, Srinivas argued that the horizontal solidarity of caste expressed by endogamy and commensality was counterbalanced by the vertical solidarity observable at the village level through patron–client relations. Many institutions such as fictive kinship within the village, ritual processes for creation of boundaries, and the division of labor and economic exchange, created the village as an economic, political, and ritual unit that established solidarities cutting across caste. As the general understanding of locality in the field of anthropological theory deepened, interest shifted to the mapping of specific histories through which localities were created. The search for overarching structural principles seemed less fruitful than understanding how the momentous changes that were taking place in Indian society impacted upon caste, at the local, regional, and national levels.

Colonial Constructions

Once the search for enduring structural principles gave way to historically grounded work, the colonial archive became an important source for understanding the nature of colonial rule and how it transformed the institution of caste. The processes of enumeration and classification played a particularly important role in this process. Although the predecessor states of the British in India did have apparatuses for counting, classifying, and controlling populations, these were tied to specific needs of the state such as revenue collection, or rising of temporary armies. The British colonial state instituted a new way of collecting information in the form of maps, settlement reports, revenue records, statistical information, censuses, folklore, and narratives, to name a few.

This new form of governance, or 'rule-by-records and rule-by-reports' in the felicitous phrasing of Richard S. Smith had a decisive influence on the shaping of caste identities in the twentieth century. In the processes of classifying and enumerating the population, the British did not start with caste as if it were a natural category found on the ground. In the early phases of colonial rule, the emphasis was on cadastral control. Statistics on land ownership, tenancy, crop production, and instruments of agricultural production were geared toward standardizing methods of revenue collection and it was by no means obvious that caste rather than village would be the obvious unit for the organizing of data. According to Smith, it was only around 1850 that the census in the case of Punjab was transformed from an instrument of tax to an instrument of knowledge. Earlier census reports were more pragmatic and local in orientation. There was considerable tension between the concerns of centrist census officials collating data in an encyclopedic manner and the local officials who were concerned with recording the nature of social groups and categories for more practical purposes such as collection for revenue or settlement of disputes (see Appadurai, 1996). In order to understand the concerns of the centrist census officials who recorded information on caste for which there could have been no practical use, it is necessary to turn to the notions of race in the scientific discourses of this period.

In shifting to castes as the most natural groups around which information was to be organized on Indian society, British officials relied on their notions of race and physical types. Herbert Risley (1969[1908]) was the most vocal proponent of using anthropometrical measures in conducting the ethnographic survey of India because, according to him, anthropometry yielded particularly good results in India by reason of the caste system, which allowed marriage only within a limited circle. Relying on such symbolism as that of dark versus light skin color and shape of nose and jaw, the category of caste was thus collapsed with that of race. Ghurye (1932) was the first Indian sociologist to criticize this view of caste and to challenge its political implications.

It is not anyone's case that the process of recording caste created this institution *ex nihilo*. What it did was to freeze the ongoing processes by which caste came to be solidified in the official imagination, and to generate the conception of community as enumerative, which had a strong influence on processes of political representation (Cohn, 1984). It is important to underscore the fact that the census, gazetteers, reports, and other such forms of knowledge came to represent the power of official discourse to name and fix the status of caste groups in local imaginaries. Many caste groups began to see the census as the source for claiming higher status. Thus census commissioners were besieged with petitions challenging a particular status ascribed to a caste.

It is also interesting to see that the theories of the racial origin of castes had a serious impact on social and political mobilization, as in the anti-Brahmin movements in the South of India and in the mobilization of the untouchable castes under the leadership of Ambedkar in Western India – both of which used the idea of original settlers versus Aryan invaders. The discursive formations around caste quotas in educational institutions, government jobs, and reserved constituencies, as well as the politics of mobilization around these issues, show the interaction and circulation of categories between popular imagination and official discourse (Dirks, 1992).

The Contemporary Context

An important paradox about caste in public life in India is that while the Constitution of India recognizes only individuals as bearers of rights and duties, and bans any discrimination based on caste or religion, the processes of democratic politics with the imperatives of grass root mobilization have created new arenas for the evolution of caste-based politics (Bèteille, 1996; Kothari, 1970). Interestingly, the legal prohibitions on caste and the political mobilization around caste reflect the contrary pulls of formal legal justice system and the imperatives of representational politics in contemporary India. This can be illustrated with the example of 'untouchability' and its contradictory careers in law, politics, and everyday life (Galanter, 1972). It is not that a similar exercise cannot be done on other castes, but the example of untouchability has a particular resonance because of its unique association with the caste system.

While it could conceivably be argued that some form of caste-based discrimination is to be found in all the legal sanskritic texts, recent examination of the genealogy of the terms through which untouchability makes an appearance in discourse shows the importance of political and social processes in the negotiation of group identities in democratic societies (Charsley, 1996). The term 'untouchability' is ascribed to Sir Herbert Risley (1969[1908]) and was part of his effort to classify and rank castes in the subcontinent as a whole. While the category *shudra* occurs in the Sanskrit texts and is sometimes taken to be equivalent to 'untouchables,' its referents are varied – ranging from kings, powerful land-owning castes, to castes with extreme disabilities. Prior to Risley, compilers of district gazetteers and state census reports had experimented with other terms, such as depressed classes, depressed castes, *panchamas*, *pariahs*, etc.

The use of the term 'untouchable,' in public life, owed much to the reformist politics of early twentieth century, especially to Gandhi's politics of reform and the agenda for the abolition of untouchability in the nationalist movement (Gandhi, 1954). In 1931, Gandhi adopted the term *Harijan* (people of god) and increasingly substituted it for other terms in his writings. While the prototypical Harijan for Gandhi was a member of the *bhangi* caste who cleaned lavatories and was thus rendered 'unclean,' his major political opponent, Ambedkar, who himself came from the Mahar caste of Bombay Presidency, was much more interested in forging political alliances between the major agricultural-dependent castes, whose low status came from their dependency rather than their polluting occupations. Yet he preferred to retain the label 'untouchable' in his politics (Charsley, 1996). In the embroilment of nationalist politics with caste mobilization, Gandhi's strategy was to retain untouchables within the fold of Hinduism while Ambedkar's attempt was to forge a new identity for them, but within the Indian civilizational context – hence the choice of neo-Buddhism as the new religion for the erstwhile untouchables (Ambedkar, 1946). The term 'untouchable,' like other such generic terms, masks considerable local heterogeneity. Its importance lay in the binary division between untouchable and nonuntouchable castes created through the bureaucratic imperative of creating a single category for purposes of reservations. Under the Government of India Act of 1935, the term 'scheduled castes' replaced the earlier category of depressed classes. Although untouchability was to be the criterion for inclusion of castes into the list of scheduled castes, the names of the castes that were finally included in the state lists formed a kind of unity only through a "common relationship their members have with government" (Dushkin, 1972: p. 166). The fact that the category of untouchability was basically a twentieth-century creation does not detract from either the experiences of oppression of castes grouped under this category, or the importance of the grounding of struggles for greater equality under identity politics. In fact, the creativity of the new social movements can be seen in the emergence of the term Dalit (literally 'downtrodden') among the neo-Buddhists and scheduled castes in Maharashtra, which has become now the most commonly accepted term of self-reference by such groups. The most interesting feature of this phase of the movement of Dalits is the emphasis not only on political action but also on representation of the experiences of untouchable castes through literature, life histories, and collection of oral literature as a critique of caste society (Murugkar, 1991; Omvedt, 1995).

An interesting question arises as to whether politics is the only domain within which the legacy of oppression and humiliation is articulated in contemporary India. How are caste identities performed in everyday life? Khare (1984) had suggested that instead of playing the impure foil to the Brahmin, the untouchables conduct themselves even at the community level as civilizational critics, showing a complex relationship with caste ideology. Other important ethnographies of caste show how caste identity is performed in relation to everyday activities of work and worship, how elaborate strategies for management of caste identities are evolved, and how these performances come to be experienced as forms of embodiment. Debates on caste have shown the intricate relations between forms of aesthetic, political, and social representations of caste. Simultaneously, much of the scholarly agenda of research in India in recent years has been influenced by the public controversies on such issues as caste-based reservations or reintroduction of caste-based enumeration for the next decennial census. New patterns of sociality are evident in ways in which caste identity is performed in the context of family, community, economy, and polity, which reconceptualize caste not as a sign of the *exotic* but of the *contemporary* repositioning of the subjects and objects of anthropology in new ways.

Dalits, History, and Modernity

New directions in the understanding of the Dalits and caste oppression, as well as an independent Dalit perspective on modernity emerged as Dalit intellectuals have succeeded in reorienting the scholarship on caste. Simultaneously, there are important changes in the political landscape since the 1990s, signaled by the growing importance of Bahujan Samaj Party and its iconic Dalit leader, Mayawati, which have created the conditions under which new and controversial questions about Dalit identity can be raised, including by Dalit intellectuals and literary figures themselves.

There seems to be complete agreement among scholars that the prominence of Dalit literature has been crucial for bringing the experience of everyday humiliation as well as the brutal violence against Dalits into the public sphere by the literary device of testimonios (Ganguly, 2009; Rege, 2006; Bhagwan and Feldhaus, 2008a,b). Careful historical scholarship has been brought to bear on the manner in which classification and representation managed to eclipse the role of particular untouchable castes in historical struggles. Thus, Rawat (2011) shows the historical processes through which the Chamars, who worked as agricultural labor, were assimilated into the category of leather workers in the colonial classifications of caste and viewed as a criminal caste. In nationalist discourses too, their contributions to peasant movements were not recognized because they were not seen as a peasant caste. Rawat shows how some contemporary scholars have been quick to assume a stable composite caste identity of lower castes (Dalit-Bahujan) that is projected back to immemorial times. Such formulation obscure the intensification of

conflicts between members of previously untouchable castes and the middle-level agrarian castes that are now subsumed under the category of other backward castes. Narayan (2011) too makes the point that Dalit identity is not necessarily a stable identity – which is not to say that identities carved through experiences of making collective political subjectivities are inauthentic, but to signal the future openness of these constructions. Rao's (2009) careful historical reconstruction of the way that Dalit political subjectivity emerged through the nineteenth- and twentieth-century struggles shows how suffering might be transformed from a politics of resentment to a politics that claims agency through contestation.

The second important set of questions relates to the impact of various legislative and administrative measures to prevent caste atrocities on intercaste relations. Here, the folding of gender into caste is significant. Rao (2011) analyzes the cases of atrocity including infamous cases of rape and murder and concludes that scheduled caste women were seen as the property of upper caste men and that women were specially targeted for spectacular and public punishment for defying upper caste men and their sexual control over their bodies. In a later work, Rao (2009) adds more nuance to this argument as she argues that despite the good intentions behind laws to prevent upper caste violence against lower castes, especially Dalits, they can have the effect of revitalizing archaic practices. As laws that single out violence against scheduled castes become broader in scope, they constitute the Dalits as vulnerable subjects already subject to injury and thus might perpetuate caste stigmatization.

While on the one hand, it is clear that caste violence has become more intensified in areas in which the socioeconomic difference between upper and lower castes has narrowed (Sharma, 2012), it is also the case that studies of everyday life in areas reveal interesting contradictory trends. One survey designed by Dalit intellectuals and implemented by Dalit surveyors in two blocks in Uttar Pradesh found that Dalit households expressed greater satisfaction with their lives as compared to that of their parents. This was attributed in large part to greater accessibility of consumer goods and participation in markets (Kapur et al., 2010). In another case in a predominantly scheduled caste neighborhood in Delhi, Das (2011) found that connivance between local big men belonging to one's own caste and appointed officials surfaced in accusations of corruption and manipulation of government schemes to benefit residents so that the residents saw their own leaders as engaged in massive deceptions and exploitation. The Dalit perspective on modernity, it is argued, is not informed by valorization of tradition (Guru, 2000) as is evident in the desire for new kinds of goods or the desire for learning English (Dasgupta, 2000). Yet Dalit intellectuals as well as other scholars might have been too quick in dismissing what it is to be attached to tradition because of a rather simple view of tradition.

A more complex view of everyday life that incorporates heterogeneous elements from tradition is evident in the work of anthropologists and literary scholars who have been more attentive to the multiplicity of subject formations. Consider the work of Singh (2011) who found that the relation between Sahariyas in Rajasthan (classified as a scheduled tribe and some regions and scheduled caste in others) and the local landowning castes over the bonded labor issues might be best interpreted in terms of agonistic intimacy, in that same people who were involved in intense conflicts in the arena of politics were drawn to each other in ritual and worship through the mobility of their gods. In a bold move, Ganguly (2005) argued that the normative modernity that informs the conceptual legislation of the social sciences does not allow other conceptual positions through which attachment to caste might be viewed as a mode of dwelling and belonging. She proposes a shift from the ideological (caste as oppressive) to the phenomenological that can account for both pain and attachment to one's mode of being (however defined) within the everyday. Such a view would neither obscure the injustices and oppression to which members of lower castes are subject, nor assume that the practices of one's caste are all treated as stigmatized. Betlem (unpublished) offers a very interesting and thoughtful example from her study of the custom of dedicating young girls to the goddess among the Madigas (a Dalit caste) in Andhra Pradesh – a custom that is considered to be 'repugnant' by the reformist state because of the fact that the girls then become sexually available to higher caste patrons. Betlem shows that members of the Madiga caste display a complex subjectivity in which they simultaneously participate in the reform projects of the State and in the ritual performances through which the goddess is placated, negotiating multiple transactions within the self. Finally, we should note that such complex constructions of the self are found not only in the relation that Dalit castes have to Hinduism but also to Christianity. The fine study on caste and Christianity by Mosse (2012) demonstrates how Dalit Christians contributed to a distinctive culture of protest in the Tamil region drawing the Church into caste politics of the region. Simultaneously, however, Dalit festivals have been incorporated in the ritual calendar of many local churches and the appeal by Dalit priests to incorporate an Indo Dalit Christian rite within the Roman Catholic Church distinct from the Latin rite points to complex forms of attachment that put the entire question of tradition and modernity in a different light than a notion of modernity embedded exclusively within consumption would suggest.

See also: Ethnic Intermarriage; Religions of India.

Bibliography

Ambedkar, B.R., 1946. What Congress and Gandhi Have Done to the Untouchables. Thacker, Bombay, India.

Appadurai, A., 1996. Number in the colonial imagination. In: Appadurai, A. (Ed.), Modernity at Large: Cultural Dimensions of Globalization. University of Minnesota Press, Minneapolis, MN, pp. 114–139.

Bèteille, A., 1996. Caste in contemporary India. In: Fuller, C.J. (Ed.), Caste Today. Oxford University Press, Delhi, India.

Betlem, H., unpublished. When Custom is a Crime: Law, Life and the Goddess in Rural South India, PhD Dissertation. Johns Hopkins University, Baltimore.

Bhagwan, M., Feldhaus, A. (Eds.), 2008a. Speaking Truth to Power: Religion, Caste and the Subaltern Question in India. Oxford University Press, Delhi.

Bhagwan, M., Feldhaus, A. (Eds.), 2008b. Claiming Power from Below: Dalits and the Subaltern Question in India. Oxford University Press, Delhi.

Charsley, S., 1996. 'Untouchable': what is in the name? Journal of Royal Anthropological Institute 2, 1–23.

Cohn, B.S., 1984. The census, social structure and objectification in South Asia. Folk 26, 25–49.
Das, V., 2011. State, citizenship and the urban poor. Citizenship Studies 13 (3–4), 319–333.
Dasgupta, P., 2000. Sanskrit, English and Dalits. Economic and Political Weekly 5 (16), 1407–1412.
Dirks, N.B., 1992. Castes of mind. Representations 37, 56–78.
Dumont, L., 1980. Homo Hierarchicus: The Caste System and Its Implications. University of Chicago Press, Chicago.
Dushkin, L., 1972. Scheduled caste politics. In: Mahar, J.M. (Ed.), The Untouchables in Contemporary India. University of Arizona Press, Tucson, AZ.
Galanter, M., 1972. The abolition of disabilities: untouchability and the law. In: Mahar, J.M. (Ed.), The Untouchables in Contemporary India. University of Arizona Press, Tucson, AZ.
Gandhi, M.K., 1954. The Removal of Untouchability. Navajivan Press, Ahmadabad, India.
Ganguly, D., 2005. Caste, Colonialism, and Counter-Modernity: Notes on Postcolonial Hermeneutics. Routledge, London.
Ganguly, D., 2009. Pain, personhood, and the collective: Dalit life narratives. Asian Studies Review 37 (4), 429–442.
Ghurye, G.S., 1932. Caste and Race in India. Kegan Paul, London.
Guru, G., 2000. Dalits in Pursuit of Modernity. In: Thapar, R. (Ed.), India: Another Millennium? Penguin, India, Delhi, pp. 123–137.
Kapur, D., Prasad, C.B., Pritchett, L., Babu, S.D., 2010. Rethinking inequality: Dalits in Uttar Pradesh in the market reform era. Economic and Political Weekly XLV (30), 39–49.
Khare, R.S., 1984. The Untouchable as Himself: Ideology, Identity and Pragmatism among the Lucknow Chamars. Cambridge University Press, Cambridge, UK.
Kothari, R., 1970. Caste in Indian Politics. Orient Longmans, Delhi, India.
Marriott, M. (Ed.), 1990. India through Hindu Categories. Sage, Delhi, India.
Mosse, D., 2012. The Saint in the Banyan Tree: Christianity and Caste Society in India. University of California Press, Berkeley.
Murugkar, L., 1991. Dalit Panther Movement in Maharashtra: A Sociological Appraisal. Popular Prakashan, Bombay, India.
Narayan, B., 2011. The Making of Dalit Publics in North India: Uttar Pradesh, 1950-Present. Oxford University Press, New Delhi, India.
Omvedt, G., 1995. Dalit Visions: The Anti-Caste Movement and the Construction of an Indian Identity. Orient Longman, New Delhi, India.
Rege, S., 2006. Writing Caste, Writing Gender: Reading Dalit Women's Testimonios. Zuban, New Delhi.
Rao, A., 2009. The Caste Question: Dalits and the Politics of Modern India. University of California Press, Berkeley.
Rao, A., 2011. Violence and humanity: or, vulnerability as political subjectivity. Social Research: An International Quarterly 72 (2), 607–632.
Rawat, R., 2011. Reconsidering Untouchability: Chamars and Dalit History in North India. Indiana University Press, Bloomington.
Risley, H.R., 1969[1908]. The People of India. Oriental Books, New Delhi, India.
Sharma, S., 2012. Hate Crimes in India: An Economic Analysis of Violence and Atrocities.
Singh, B., 2011. Agonistic intimacy and moral aspiration in popular Hinduism: a study in the political theology of the neighbor. American Ethnologist 38 (3), 430–450.

Castells, Manuel (1942–)

Chaime Marcuello-Servós and Brenda Reed, Facultad de Ciencias Sociales y del Trabajo, Universidad de Zaragoza, Zaragoza, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Manuel Castells Oliván is a highly cited sociologist born in 1942 in Hellín (Albacete), Spain. He began his academic training in Spain, but left for political reasons in the 1960s and continued his very active life and higher education in Paris. In 1979 he began to develop his academic career in the United States at Berkeley and other universities. His scholarly contributions are in the field of urban sociology and, from the beginning of the 1990s, in globalization and information and communication technologies. His work can be characterized as situated within the context of praxis, downplaying disciplinary scholarly development and focusing more on the identification of social and cultural inequities, and solutions to those inequities.

An Active Lifestyle

Manuel Castells Oliván was born in Hellin, Albacete Province of Spain on 2 February 1942. His mother, Josefina Oliván Escartín, was Aragonese from Biescas (Huesca), and his father, Fernando de Castells Adriaenses, was from Valencia. In a sense, it was a coincidence that he was born in the La Mancha region of southern Spain, where his father had been stationed. Later, Manuel Castells lived in several other places as the family moved because of his father's job. In his adolescence and youth he moved between the provinces of Valencia and Barcelona This article has been drawn from three types of sources. The first source is the two 'official' Web sites of Professor Castells, his professional web at the University of Southern California, Annenberg School for Communication and Journalism, http://annenberg.usc.edu/Faculty/Communication%20and%20Journalism/CastellsM.aspx and his 'scientific web,' a project financed by the University of Catalonia and the Fundación Telefónica, available at http://www.manuelcastells.info/es/index.htm. In the latter, Castells has updated and detailed the whole of his academic and professional activity. The second source of information consists of Castells' own texts as well as articles, monographs, anthologies, and published popular works from various stages of his life. Finally, the third source is a set of research materials available on various Web sites such as videos, interviews, conferences, etc., along with this author's own observations and interactions with Professor Castells on two occasions, first his appointment as an honorary member of the Spanish Federation of Sociology in Madrid in 2010 and second, the World Congress of Sociology, 2010 in Gothenburg.

Although Manuel Castells does not usually talk about his family or childhood, on several occasions he has said that he came from a very conservative and Francoist (supporters of the Spanish dictator Francisco Franco) family, which was common in Spain at that time. This background, undoubtedly, conditioned his subsequent development, his selection of topics for analysis, and his construction of theories about the world and society. It is possible that he inherited a strong sense of realism, with both feet firmly on the ground, and the critical spirit that characterizes him, from his 'Altoaragonese' mother (Altoaragonese are people from the North of Aragon. This includes the different valleys of the Pyrenees). His concern about power and the search for freedom probably have much to do with his sanguine, active, and sensual temperament and enjoyment of life, which put him up against the authoritarian ways of Francoist society of that time. He himself has said in interviews that he was an active, dynamic, and lively student, who found himself constrained by an authoritarian society and political regime. This made him leave Spain and look outside for what he could not find in Spanish society in the 1960.

At the age of 16 he began his university studies in Barcelona, in the field of social sciences and law. However, he could not study Sociology because it was not institutionalized until the 1970s, although several Sociology Chairs had existed in Spain from 1898. (In 1898, Prof. Manuel Sales y Ferré (1843–1910) was awarded the first chair of Professor of Sociology at the Spanish University Complutense of Madrid.) His initial training focused on law and economics from 1958 to 1962. At the age of 20, without concluding his studies, he was exiled to Paris. He resumed his studies in 1964 obtaining a License (Masters) in Public Law and Political Economy at the University of Paris.

He continued his education and in 1965 received a Diploma in Sociology of Work from the Institute of Social Sciences of Work at the University of Paris. In 1966, he obtained the Diplome d'Etudes Approfondies (Masters) in Sociology, Ecole Pratique des Hautes Etudes, University of Paris. This was followed by a Doctorat 3eme Cycle (PhD) in Sociology from the University of Paris in 1967, a Doctorate (PhD) in Sociology from the Universidad Complutense of Madrid, and a Doctorat d' Etat in Sciences Humaines from the Universite de Paris-V, 'Rene Descartes – Sorbonne.'

His capacity for work and sound training achieved in the field of sociology, under the tutelage of his teacher Alain Touraine, enabled him to be a researcher and assistant lecturer in France, firstly from 1965 to 1967 as Researcher at the Laboratory of Industrial Sociology at the Ecole Pratique des Hautes Etudes in Paris, then from 1967 to 1969 as Assistant Lecturer in Sociology at the University of Paris (Nanterre Campus). It is said that his classes were a key place for the French students who mobilized in May 1968. This, however, meant that he had to leave France for a while from 1969 to 1970, taking up the post of Assistant Lecturer in Sociology at the University of Montreal (Canada). He returned once again

to the French university as Associated Professor in Sociology (with tenure), and director of the Seminar for Urban Sociology, *Ecole des Hautes Etudes en Sciences Sociales* in Paris, from 1970 to 1979.

He was Professor of City and Regional Planning and Professor of Sociology at the University of California at Berkeley from 1979 to 2003, engaging in various tasks such as teaching, research, and management, which led him to affiliation with a wide range of centers and institutes at Berkeley that focused on Western European Studies, Slavic and Eastern European Studies, Environmental Design, and Southeast Asian Studies. In addition, he combined these duties at Berkeley with posts of Professor of Sociology and Director of the Institute of Sociology of New Technologies at the Universidad Autónoma of Madrid (1988–93) and with his position of Research Professor, Consejo Superior de Investigaciones Cientificas, Barcelona (1997).

As of 2013, he held positions as Director of Research, Professor and Professor Emeritus in four different institutions: Cambridge, University of Southern California, Open University of Catalonia, and University of California at Berkeley.

Social Scientist

Castells has said, "I am against disciplines including sociology, economics, anthropology, I believe that the world is complex and can only be understood from an interdisciplinary approach (…). I don't believe in disciplines." (Taken from the Radio Televisión Española (RTVE; Spanish State Radio Television) documentary 2013, http://www.rtve.es/television/20130612/sociologo-manuel-castells/686369.shtml.) His words, while they contain a contradiction because to speak of interdisciplinarity, the distinctions that apply to each discipline are necessary, allow us to understand, to some extent, the changes and the evolution of his career as a researcher and teacher in the field of social science. His initial training, as shown in the previous section, started in the field of economics and law, and then passed through sociology, political science, and humanities, and is currently in the field of communication. Manuel Castells himself has pointed out, "When something matters to me and I don't understand it, I become a professor of this in order to learn." Perhaps this is the key that has resulted in him becoming one of the most cited academics of the early twenty-first century.

Internationally, despite his statements of interdisciplinarity, Castells is recognized first, as an urban sociologist, second, from the late 1990s, as a referent in the sociology of globalization and information and communication technologies (ICTs). In fact, the field of sociology, it could be said is his primary field, however his success has been to go beyond the limits of the discipline(s) to address social reality itself as an object that requires customized empirical attention and comprehensive effort in its completeness.

For Castells, his approach is not to 'jibarize' (minimize) the object of study and to compress it into preconceived models, but to study the concrete by going beyond abstract models: from empirical to theory, and not the reverse. His methodological position is that any explanation and analysis, if it wishes to be scientific, must be based on data and observation of phenomena to explain and analyze, but not just models. It is a task, a work of intellectual construction that starts from observation, collects data empirically, submits data to an analysis methodology, and then proceeds to an intellectual reconstruction preparing a statement that makes sense and is not just a statement, but gives sense to an observed reality.

Castells advocates the empiricism of praxis in touch with the world, in the face of the sophist temptation of empty words grounded in previously established abstractions. As he has explained publicly, his doctoral thesis at the University of Paris dealt with 'the determinants of the location of industrial enterprises in the metropolitan area of Paris, a statistical model of analysis of 1500 companies' and he points out "there is nothing more concrete and statistical" (RTVE documentary). Castells says he has always been empirical.

Manuel Castells, therefore, emphasizes the need for empirical observation that focuses analysis and makes rigor possible, because you cannot say anything, "what you say has to relate to what you observe" (RTVE documentary). He goes a step further in his approach than mere empirical verification to opt for an 'intellectual framework' because it has a criterion of coherence and relevance in the process of explaining phenomena.

Urban Sociologist

Castells' current rejection of preconceived models that simplify social reality is an evolution from his early works. In his youth he shared a Marxist approach where the solution is already known due to observation being limited by the Marxist-defined coordinates that see conflict and class struggles such as the role of social movements in urban transformation (Castells, 1976). His contributions to urban sociology brought him significant international importance. He is credited with introducing the concept of 'collective consumption' (Castells, 1974a; Castells 1975, 1976), as a way to study changes in the management of social struggles, moving from the realm of the economic to the political because of state intervention. This focus on the political was another way of studying the faces of power; something that continues to occupy him and which he regards as central in his academic and research career.

Castells considers himself as an advocate and member of the 1970s 'new urban sociology' (Castells, 2012, Holberg Prize Ceremony) (Elliot, 1980), where power – conflicts, politics – and not just spatial forms were the central issues. Included in this group were Lefevre, Harvey, and a few others, who left behind the Chicago School's concern for 'urban integration.' In this early stage as an urban sociologist he produced an important list of books and articles. One of the most provocative 'Y a-t- il une sociologie urbaine?' was translated into English in 1976 although published in French in 1968 in the newly created magazine *Sociologie du Travail*. Other works that stand out include *Problemas de investigación en sociología urbana* (1971), *La Question Urbaine* (1972), the collection of essays City, Class, and Power (1978), *Urban Crisis and Social Change* (1981), and *The City and the Grassroots: A Cross- Cultural Theory of Urban Social Movements* (1983).

His research on the subjects of urbanism, urban planning, structure of cities, industrial areas, organizations, and issues

that revolve around the city, social movements, and power structures also led him to be concerned about social and economic development in Latin America and Asia. From this it is also possible to trace his gradual transition to studies on globalization and its attempt to respond to the new challenges that have emerged, in that the city and society have been transformed at the hand of the so-called ICTs.

Thus, in 1989, he published *The Informational City. Information Technology, Economic Restructuring and the Urban – Regional Process* and an article entitled *Social Movements and the Informational City* that can be considered to be key references in the evolution of his work. Castells puts forward the concept of *space of flows*, which is described as "the material organization of time-sharing social practices that work through flows" (Castells, 1989a, p. 147). From this conceptualization, an almost natural transition is articulated from his studies in the field of urban sociology to the new dimension of a globalized world where ICTs become the backbone of its new research and publications.

Informational Society

For Castells, in the 'information age' it is not intellectually correct to speak of the information and knowledge society. All human societies are based on both pillars, processing them in particular ways according to the time and circumstances. He prefers to say 'informational,' that is, "the use of a new technological system based on ICTs with microelectronic support and of communication through digital networks" (RTVE documentary, 2013).

From his viewpoint, this technological framework changes the mental and communication processes of human beings. ICTs cause a material change in relation to how we process information and communicate. Castells is among those, including Niklas Luhmann and Gordon Pask, who emphasize that communication is the essential element of human interaction and that meaningful communication is the key to our social relations. Castells, however, appears to discover a novel way that had not been explored that focuses on that we are in an era where microelectronics have changed the forms, rhythms, and communication mechanisms so that the human mind has become radically transformed. This means that we are faced with a new scenario: a great transformation that generates a new digital culture, where the convergence of cultural change and technological change revolutionizes society.

Castells posits that his arrival in Berkeley allowed him to see firsthand the consolidation of the digital information industry in the Silicon Valley, and witness the emergence of this process of transformation that is now on another level. In a sense, he was in the right place at the right time. He was present at the electronics and information explosion, seeing emerge what now seems obvious: a new technological system affecting all and producing the electronic revolution.

Manuel Castells argues that these structural phenomena such as technological change, digital culture, and globalization require various viewpoints. Different approaches from different cultural positions are needed to see their interdependence. It can only be explained in this way, knowing that there has been a radical change that transforms the present and the future of humanity. Castells argues that 97% of information is digitalized, therefore accessible by electronic means and the network. This leads him to pass from the diagnostic and the empirical description of the world around us to the proposal and the prescription of what one should do. It could be said that he moves from science to consciousness and then to political praxis, in the best sense of the word. He considers that, faced with this overwhelming circumstance produced by the excess of available information, where a person can become paralyzed without knowing what to do, one has to pass from passive consumption in the network ocean, to navigation, with direction and orientation. He proposes that everyone becomes active creators of information and communication.

Castells finds an example for this issue in the social (wiki) movements of occupation and rebellion produced in different places and countries. Here he observes how the use of available ICTs has made other forms of organization possible. The people involved in these social movements deploy deliberate and nonhierarchical forms precisely because technology allows these forms of communication. This is the underlying current in his work, *Networks of Outrage and Hope: Social Movements in the Internet Age* (2012).

For him, the Internet arose as a social media, with the intention of connecting with others. Castells argues that the Internet is not an information system, it is a communication system (RTVE documentary, 2013). He attributes elements of passion, innovation, and altruism to the network and downplays its military and power control origins. However, apart from this detail, he has also called his 'informational society' a 'network society.' The essential work for understanding his position is the trilogy *The Information Age: Economy, Society and Culture*, consisting of the volumes *The Rise of the Network Society*, *The Power of Identity*, *End of Millennium*. The first edition was published between 1996 and 1998 and the second edition, extensively modified, between 2000 and 2004, this work is considered by many to be one of the key references in global sociology and has placed him among the most important authors. It is, in fact, one of his most cited works. However, his critics say it is only a more or less successful compilation of other authors, where the only novelty is the actual summary made. Castells has continued his research and publications in this field and of his works written in 2006 the following are of note, *The Network Society: From Knowledge to Policy* and *Mobile Communication and Society: A Global Perspective*.

Power and Change

Manuel Castells describes his work as coming from the same vertebral axis, the same line of argument, namely, the question of power. For him, it has been a central theme in his personal development and career. As he stated in an interview in 2012: "My analysis of power. This has been the overarching theme I have studied in my works, in many contexts and in different domains of society. Power in the city. Power in and by information technology. Power in globalization. Communication power, that is the power and counterpower built in the media and in the Internet networks. In terms of specific themes I think that what will remain from my work is to have been able to map out economies, societies and culture at the dawn of the information age, coining the concept of the network society as

a key thread to understand the transformations brought about by the digital revolution in interaction with the culture of autonomy, and the globalization of the economy" (from the Holberg Prize interview with Manuel Castells, http://www.holbergprisen.no/en/manuel-castells/interview.html).

His analysis of the phenomenon of power is posed as an empirical question on which he elaborates theory, but from an instrumental view of theory. For him, "theory is useful to produce knowledge through research, different kinds of research. But it is not a goal in itself" (Lecture by Manuel Castells at USC Annenberg about Network Theories of Power). This is key to his overall research strategy and, in particular, in the study of power. Castells says "that power is based on the control of communication and information, be it the macro-power of the state and media corporations or the micro-power of organizations of all sorts" (2009: p. 4).

He is interested in the question of power because those who have power model and control society. He states that power is the matrix of rules that constitute society. He defines power as "the relational capacity that enables a social actor to influence asymmetrically the decisions of other social actor(s) in ways that favour the empowered actor's will, interest and values" (2009: p. 10). Additionally, Castells argues that resistance as a counterpower always appears where there is power and domination. An important component in his study of power is his extension of existing theories of power to four forms of power in networks: (it is possible to listen (directly) to Manuel Castells explaining this at http://www.youtube.com/watch?v=skcUYhRaEas) (1) networking power, (2) network power, (3) networked power, and (4) network-making power.

In his research on the question of power there is an implicit proposal of utopia. On its particular horizon is freedom as a supreme value. It is what allows everything else. Equality yes, but without freedom one cannot fight for equity. He calls for real democracy, now, which overcomes inertia and uses current tools provided by ICTs. For instance, ICTs can help achieve new forms of democratic representation through increased communicative autonomy; the social sciences contribute to this new form of social change.

For Castells, power has no abstract or theoretical nature; it is in the specific hands of individuals, organizations, and networks. He asserts that you cannot make general theories, only specific analysis of particular cases. For him, the most important issue is the creation of meaning; those in power have the ability to program the socially available meanings. Therein lies the true power. This reveals the road of change, and as a result of this he takes the position that because everyone develops their critical spirit, and from here seeks cooperation and the network to be together, get value, lose fear, and take the step to being able to act and change, with the fundamental principle of nonviolence. It could be said that he has discovered the French saying: *Tout se passe dans la tête*. What Castells seeks is a utopian freedom from below and from the Internet that changes the world as we know it for the better.

Conclusion

Manuel Castells has received over 40 awards and honors ranging from his first award, the John H. Simon Guggenheim Fellowship in 1982–83, to his most outstanding and most recent, the Holberg International Memorial Prize from the Parliament of Norway, in 2012, and the 2013 Balzan Prize for Sociology. His contributions, first, in the area of urban sociology and, second, in the field of ICTs have positioned him throughout the world as an author of reference. His work has been innovative and visionary in many ways. He could be considered thus a key figure in the construction of international sociology, and continues to be active in producing and generating significant contributions for a better understanding of the current world. He has generated a substantial theoretical corpus with his publications.

His work does not stop with publications, Castells has sought to take up a position and excel in all processes of 'revolution' and change that have taken place close by. These have included the demonstrations and sit-ins arising from 15-M in Spain. (Watch him and listen to his own words at http://www.youtube.com/watch?v=iUsNhz9Vxkk.) His activities emphasize his multilocal character; "after being expelled from several countries, whenever there was a dictatorship I was going to fight it and they end up expelling me" (RTVE documentary, 2013). This inclination to fight against power and the powerful is admirable, though it also provides a ready lead into questions that arise regarding his contributions to a solid and cohesive body of scholarly knowledge about power and organizational processes.

While many followers of Castells see in his work and person a figure equivalent in contribution to the social sciences as Karl Marx, Max Weber, or (more recently), Bourdieu, Luhmann, Habermas; others raise concerns about the impact on knowledge creation of Castells' prioritization of empiricism and flexibility over discipline-based theory. (It is possible to listen to him on http://www.rtve.es/television/20130612/sociologo-manuel-castells/686369.shtml. Some examples of a type of 'adoration' and idolization can be found at 7′57″, 15′, and 17′ 31″. In the last example Imma Tobella said [*sic*]: "Castells demonstrated that he is the Max Weber of XXI century.")

Regarding empiricism and flexibility, one could say that Castells advocates the empiricism of praxis in touch with the world and implicitly characterizes a focus on discipline-based theory as using empty words grounded in previously established abstractions. As he has explained publicly, his doctoral thesis at the University of Paris dealt with "the determinants of the location of industrial enterprises in the metropolitan area of Paris, a statistical model of analysis of 1500 companies" and as he emphasizes "there is nothing more concrete and statistical."

However, it could be said that his epistemological and methodological perspective has remained in a first order observation, where there is not a review process and critical reflection on the way in which observer and observed produce the observing process. The world is given, you only have to look (outside). This approach, from a Sociocybernetical perspective (Marcuello, 2006) or from the Subject Oriented Approach of Arne Kjellman (2003), lacks sufficient epistemological sophistication. As described by Van Dijk (2010) Castells approach is fuzzy and unclear, where meanings and psychology appear in his particular repertoire of concepts without giving a coherent and solid theory. Castells analysis lacks a more comprehensive

review of other approaches that add more complexity to complicated and complex issues. This lack of complexity is critically important in his way of studying power, or rather, power relations, which continues to retain a dichotomous perspective where it is clear that there are some who command because they have power and others who obey and resist because they do not. Old patterns are repeated under an undeclared Marxist approach, but the elements of complexity that the modern social system produces are obviated.

It is true that the work of Manuel Castells is vast and significant. However, if his work is analyzed and compared with other great social scientists, questions arise as to whether he will be considered to be of the same standing as either classical or contemporary figures of sociology. Only the passage of time will tell if his work has enough density to establish itself in the History of Ideas and be a landmark reference.

See also: Bourdieu, Pierre (1930–2002); Habermas, Jürgen (1929–); Luhmann, Niklas (1927–98); Oversharing: The Eclipse of Privacy in the Internet Age.

Bibliography

Boulding, K.E., 1989. Three Faces Of Power Sage-London, *ISBN:* 9780803938625.

Castells, Manuel, 1968. Y a-t-il une sociologie urbaine? Sociologie du Travail 1, 72–90.

Castells, Manuel, 1971. Problemas de investigacion en sociologia urbana. Siglo XXI, Madrid-Mexico.

Castells, Manuel, 1972. La Question Urbaine. Francois Maspero, Paris.

Castells, Manuel, 1974a. Consommation collective, interets de classe et processus politique dans le capitalisme avance. Papers: Revista de Sociologia (Special Issue), 63–90.

Castells, Manuel, 1974b. Monopolville. L'entreprise, l'etat, l'urbain. Mouton, Paris.

Castells, Manuel, 1975. Advanced capitalism, collective consumption and urban contradictions. In: Lindberg, Leo, et al. (Eds.), Stress and Contradiction in Modern Capitalism. Heath, Lexington, MA, pp. 175–198.

Castells, Manuel, 1976. Crise de l'Etat, consommation collective et contradictions urbaines. In: Poulantzas, Nicos (Ed.), La Crise de l'Etat. Presses Universitaire de France, Paris, pp. 179–208.

Castells, Manuel, 1978. City, Class, and Power. Mac Millan, London and New York: St Martin's Press.

Castells, Manuel, 1981. Crisis Urbana Y Cambio Social. Siglo XXI, Madrid-Mexico.

Castells, Manuel, 1983. The City and the Grassroots. A Cross-Cultural Theory of Urban Social Movements. University of California Press, Berkeley.

Castells, Manuel, 1989a. The Informational City. Information Technology, Economic Restructuring and the Urban-Regional Process. Blackwell Publishers, Oxford, England, and Cambridge, Massachussets.

Castells, Manuel, 1989b. Social movements and the informational city. Hitotsubashi Journal of Social Studies 21, 197–206.

Castells, Manuel, 1996. The Information Age: Economy, Society, and Culture. In: The Rise of the Network Society, second ed., vol. 1. Blackwell Publishers, Oxford, England, and Cambridge, Massachusetts. 2000.

Castells, Manuel, 1997. The Information Age: Economy, Society, and Culture. In: The Power of Identity, second ed., vol. 2. Blackwell Publishers, Oxford, England, and Cambridge, Massachusetts. 2004.

Castells, Manuel, 1998. The Information Age: Economy, Society, and Culture. In: End of Millennium, second ed., vol. 3. Blackwell Publishers, Oxford, England, and Cambridge, Massachusetts. 2000.

Castells, Manuel, 2006a. The Network Society: From Knowledge to Policy. Center for Transatlantic Relations, The Johns Hopkins University, Washington DC. Co-edited with Gustavo Cardoso. Co-published in Portuguese. Lisboa: Gulbenkian Foundation.

Castells, Manuel, 2006b. Mobile Communication and Society: A Global Perspective. The MIT Press, Cambridge, Massachussets.

Castells, Manuel, 2012. Networks of Outrage and Hope: Social Movements in the Internet Age. Polity Press, Cambridge.

Elliot, B., 1980. Manuel Castells and the new urban sociology. British Journal of Sociology 31 (1), 151–154.

Kjellman, A., 2003. Doctoral Thesis, Constructive systems science - the only remaining alternative? A contribution to science and human epistemology, defended at The Royal Institute of Technology (Sweden), ISBN: 91-7283-531-1, and available at http://people.dsv.su.se/~kjellman/e-index.html.

Marcuello Servós, C.H., 2006. Sociocibernética. Lineamientos de un paradigma. Inst. Fernando el Católico- CSIC.

Van Dijk, J.A.G.M., 2010. Review of Manuel Castells, communication power [Bespreking van het boek Communication Power]. European Journal of Communication Research 35, 485–489.

Relevant Websites

http://annenberg.usc.edu/Faculty/Communication%20and%20Journalism/CastellsM.aspx – University of Southern California, Annenberg School for Communication and Journalism.

http://www.manuelcastells.info/es/index.htm – UOC y la Fundación Telefónica.

Multimedia

Book presentation Comunicación y Poder, 2009 at Universidad Complutense de Madrid. http://www.youtube.com/watch?v=BfSmA6HHVIE.

Ceremony of Holberg International Memorial Prize, 2012: Manuel Castells. http://www.holbergprisen.no/en/manuel-castells.html; http://www.holbergprisen.no/en/manuel-castells/interview.html.

Documental RTVE: El mundo según el sociólogo Manuel Castells. Pienso, luego existo, La 2 de TVE, 16.06.2013, (29:56 min). http://www.rtve.es/television/20130612/sociologo-manuel-castells/686369.shtml.

Gabilondo entrevista al sociólogo Manuel Castells (Extracto de la entrevista). Presentador: Iñaki Gabilondo. CNN+, 10.06.10, (05:27 min). http://www.cuatro.com/noticias/Gabilondo-entrevista-sociologo-Manuel-Castells_2_1042680038.html.

Manuel Castells – Poder y Comunicación – Acampada Barcelona Mayo, 2011 (Publicado el 06.05.2012). http://www.youtube.com/watch?v=iUsNhz9Vxkk.

Manuel Castells, at Occupy London, (Published 26.11.2011). http://www.youtube.com/watch?v=BS82Do3bU-w#t=499.

Manuel Castells, Network Theories of Power – USCAnnenberg. http://www.youtube.com/watch?v=skcUYhRaEas.

Catholicism

Adrian Hastings,[†] University of Leeds, Leeds, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 1538–1542, © 2001, Elsevier Ltd.

Abstract

Catholicism, meaning universality, signifies the central form that Christianity has taken in history from the second century onward, a form shared by all the main early traditions – Greek, Latin, Syriac, Armenian – including hierarchy, sacraments, religious orders, the cult of saints. From about the eleventh century, however, the 'Catholicism' of the Western church became increasingly clearly 'Roman Catholicism,' dependent upon the exercise of papal authority and the extension of Roman Canon Law, an ordering of ecclesiastical life different to the 'orthodoxy' of the eastern churches. Challenged by the Reformation, the medieval form of this Catholicism was only strengthened by both the Counter-Reformation and the papal supremacy (Ultramontanism) of Vatican I in the nineteenth century. It was, however, profoundly altered by the reforms of Vatican II and is now in a state of contestation between those who appeal above all to the latter, reincorporating Protestant Reformation insights, and those who, on the contrary, believe that Vatican I still provides the decisive model.

Catholicism signifies the central form that Christianity has taken in history from, at latest, the second century, but it is a form that has continued steadily to evolve. In the second millennium Catholicism has come to designate most particularly the character Christianity has assumed within the Roman Communion under the authority of the papacy.

'Catholic' is not a term to be found in the New Testament but it is already being used to characterize and designate the church in its early postapostolic period. Its primary meaning is 'universal,' signifying both Christianity's claim to be a faith unbounded by language, ethnic, or class boundaries and the insistence that local church communities must be in communion with one another: the church in Corinth or Ephesus is 'Catholic' because part of an association of communities sharing the same faith and rituals present up and down the Roman Empire and, potentially, everywhere. As such the term quickly became the church's most regularly used name, incorporated into the creeds. All Christians using the creeds, whether Roman Catholic, Orthodox, or Protestant, claim to believe in, and belong to, the Catholic Church.

The character assumed by Christianity in the second century and generally termed 'Early Catholicism' is one in which each local church is ruled by a bishop, assisted by priests (presbyters) and deacons, claiming 'succession' from the apostles. Entry is via the sacrament of baptism while regular worship is centered upon the weekly celebration of the Eucharist, renewing Jesus's last supper with his disciples. The Hebrew scriptures were retained as inspired, generally in their Greek Septuagintal form, to which was added a collection of the earliest Christian writings, the New Testament, to which apostolic authority was attributed. The shape of Catholicism had, however, been established well before the full canon of the New Testament was agreed upon, certain books remaining for long in dispute. While there was at first no formal machinery for linking local churches together in the one church, local councils of bishops from neighboring towns soon began to be called, while the churches in major cities, particularly those claiming an apostolic foundation, assumed a leading position. Among these the Church of Rome, where both Peter and Paul were martyred, came to be accorded an unquestionable preeminence from an early date.

While all the books of the New Testament are centered unequivocally upon the figure of Jesus Christ, they already suggest, particularly in the account of the first Christian martyr, Stephen, but also in concentration upon certain of the apostles, the importance for early Christianity of the personal 'witness' to Christ (the strict meaning of martyr). A cult of the martyrs, focused upon their tombs, began early and developed into a wider cult of the saints that remained a permanent characteristic of Catholicism. Furthermore virginity/celibacy was stressed from early times as an alternative road to exceptional holiness, a way later institutionalized in monasticism.

This 'Early Catholicism,' constituted by hierarchical ministry, sacraments, the canon of scripture, and the cult of the saints, was common to east and west, the Greek and Latin churches, as also to Syrian, Coptic, and Armenian churches. Despite diversities of language, liturgy, and theological tradition, Catholic Christianity was held together by the fact and theology of 'communion,' a sacramental bond of fellowship from which individuals or whole churches could be excluded on account of heretical teaching or moral deviation.

After Christianity became a privileged religion with the conversion of the emperor Constantine in the early fourth century, it grew enormously in numbers. What had hitherto been a community of a voluntary sort whose members knew that they might well have to face persecution on account of their beliefs became instead quite quickly a church of the masses, many of whose members had received very limited Christian instruction. The myths and customs of the rural population were carried on, often covered over with a thin Christian veneer. Henceforth Catholicism would be characterized by the co-existence of a clerical and monastic religious upper class with a popular religion which it would preach to generation after generation but never fully convert.

Post-Constantinian Catholicism grew too in the development of an institutional framework designed to resolve conflicts and maintain unity, although the result was often to precipitate still worse division. The most striking development of this was what came to be recognized as the 'ecumenical council' beginning with that of Nicaea (325). At the same time the authority of

[†] Deceased

the major churches – Rome, Alexandria, Antioch, and then Constantinople and Jerusalem – was further accentuated. By the fifth century this had been stabilized in the theory of the 'Five Patriarchates' but, among them, Rome and Constantinople held a quite privileged position. That of Rome, with its unquestionable apostolic origins, had been there from the start, but was reinforced greatly from the late fourth century. Constantinople, the new imperial capital, claimed to be for that reason also the church's second see in dignity, though Rome denied that its own primacy was due to its civic, rather than its apostolic, status. In practice, the two co-existed for centuries, at times somewhat tensely, but also co-operatively. Rome enjoyed an increasingly effective primacy of jurisdiction throughout the Latin west but of little more than honor in the east. At the same time the advance of monasticism profoundly affected the ethos of Catholicism in both east and west.

It was only after the eleventh century that the 'Catholicism' of the western church could be distinguished clearly from the 'Orthodoxy' of the eastern. The Gregorian reform (named after Gregory VII, pope from 1073–85) profoundly affected the western church but through the accentuation of tendencies already present. Three are crucial: the first was a large extension in the claims but, still more, the effective jurisdictional power of the papacy. Its ability to impose a defined model of ecclesiastical life was turning 'Catholicism' into 'Roman Catholicism': a hitherto largely decentralized church was being made to conform ever more closely to rules laid down in Rome, rules which included insistence on higher moral and educational standards in the clergy and greater separation between clergy and laity. The second was a vast extension of canon law and its implementation through a hierarchy of church courts with appeal always possible to the highest courts in Rome. Canon law became the effective tool for implementing the new Gregorian ideals. The third was the law of celibacy for all the clergy. Hitherto most ordinary priests were, or had been, married, just as they had been and continued to be in the east. Henceforth in the west the married priest would be an offender, though in practice he would long continue to exist. The law of celibacy was a necessary tool for ensuring the segregation of clergy from laity, making it possible for the former to be controlled effectively by higher church authority. The 'church' became in practice more and more a clerical reality in which the laity were expected ideally to remain passive and obedient. The identification of church and clergy was accentuated further by language. The early church had allowed great linguistic flexibility but in the medieval west Latin came to have a wholly privileged position as the language of the church and its public services as well as of theology. As a range of vernaculars developed vigorously their own literatures, the ecclesiastical use of Latin became a further characteristic of Catholicism. It expressed a certain clerical universalism at odds with the national vernaculars used by the laity.

Two further developments, linked particularly with the thirteenth century, were also lastingly influential for the shaping of Catholicism and its differentiation from eastern Orthodoxy. One was the founding, beginning with the Franciscan and Dominican friars, of religious orders which were not purely monastic but, instead, active, mobile, and involved in the world around them, especially the urban world. Further societies of this sort, most notably the Jesuits in the sixteenth century, would multiply, providing a powerful, well-organized, and educated religious army at the service of the church, available not only internally but also for missionary service abroad, as the Franciscan mission to China in the fourteenth century already demonstrated. The second was scholasticism, formal systems of theology, heavily influenced by the rediscovery of Aristotle's works, which grew out of the teaching system of the new universities. Here formal logic was learned and applied in a systematization of theology very different from that of the patristic period or the continued traditions of the Greek east.

Thus, by the later Middle Ages the type of Christianity which we have come to call Catholicism was almost fully evolved – a mix of subapostolic, post-Constantinian, and medieval developments. The Reformation brought an attack on many of these developments, almost anything in fact which could not be defended with firm scriptural authority. Against this assault the papacy yielded almost nothing. It admitted moral and educational, but not doctrinal or institutional, failings. The Council of Trent (1545–63) formulated the lines of the Counter-Reformation position, emphatically reaffirming almost everything that Luther, Zwingli, and Calvin had impugned. The one major new element that it added to the character of Catholicism as a social reality was the seminary. Hitherto there had been no institutions formally established for the training of diocesan priests, who had learned what they needed to know either through the general education of a university or through a measure of apprenticeship, a son often enough learning the job from his father. The Tridentine seminary would improve greatly the effectiveness and education of the clergy but also further distance them socially from the laity, particularly on account of the 'minor seminary,' effectively a secondary school in which candidates for the priesthood would be segregated from their early teens. This could lead easily on to a further gap between the clergy and the university world, and therefore also between the clergy and modern ways of thought. While with time medieval scholasticism was ousted from the university, it would in a much degraded form continue to dominate the characteristic mindset of the clergy until the twentieth century.

Tridentine Catholicism was, nevertheless, far more creative than the term 'Counter-Reformation' might suggest. While classical Protestantism with its commitment to '*sola scriptura*' was averse to religious involvement in secular culture, the Catholicism of the sixteenth and seventeenth centuries produced its last great cultural era, the Baroque, as rich artistically as the earlier Catholic art cultures, the Romanesque and the Gothic. In art and music, as in theology and spirituality, Catholicism is by nature open to the secular and developmental, even luxuriantly. This was always true, though only in the nineteenth century did John Henry Newman justify the development of doctrine as a human and Catholic necessity, to be contrasted with Protestant insistence upon not going beyond the explicit content of scripture.

Sixteenth- and seventeenth-century Catholicism was also particularly fruitful in the literature of prayer, from the mystical writings of the Spanish Carmelites, Teresa of Avila and John of the Cross, to the Jesuit *Spiritual Exercises*, Francis de Sales' *Introduction to the Devout Life*, and countless other minor classics, many of them popular also among Protestants. Catholicism then has traditionally combined institutional rigidity with

spiritual and artistic creativity. In the nineteenth century with the triumph of Ultramontanism (a name used to describe a form of Catholicism in Europe north of the Alps anxious to enhance papal authority and uniformity with Rome in contrast with the more old-fashioned Gallicanism which accepted papal supremacy but limited to a minimum papal intervention in non-Italian ecclesiastical affairs) institutional rigidity grew while creativity declined. Faced with the often anticlerical, and even antireligious, ethos of the French Revolution and later liberal movements, the Catholic Church drew in on itself as a fortress against the modern world, defining the authority of the pope at the first Vatican Council of 1870 under Pius IX (1846–78) while a little later Pius X (1903–14) condemned every form of 'modernism.' In the first decades of the twentieth century Catholicism could then appear as a consistently reactionary force, intellectually, culturally and socially. Yet it was far from being a community in decline. Modern technology made Roman centralization work in a way never previously possible. The railway brought countless pilgrims to Rome and a cult of the reigning pope among ordinary Catholics quickly developed. Pius IX, antimodern as he was, became in consequence the first 'modern' pope, almost idolized by millions. At the same time a series of apparitions of the Blessed Virgin Mary mostly to children, of which the most famous were at Lourdes in 1858 and at Fatima in 1917, much reinforced a principal focus for Catholic popular devotion, both backed by and supporting papal definitions of Mary's Immaculate Conception (1854) and Assumption into Heaven (1950). The missionary orders, mostly nineteenth-century foundations, were extending Catholicism throughout the southern hemisphere; there was a move to encourage forms of lay apostolate across Europe and North America, while for the first time there had developed a vast army of nuns at work almost everywhere. This last may have been the most significant shift. In the Middle Ages nuns were few in comparison with monks and included no active orders. Only in the seventeenth century, led by heroic women like Mary Ward and Jane Frances de Chantal, did unenclosed orders of women, committed to apostolic work as well as prayer, begin to exist despite huge clerical opposition. By the close of the nineteenth century there were more nuns in the church than priests. Silent in many ways they were forced to remain, Roman Catholicism being traditionally an intensely male-dominated and clerical form of Christianity, but a changing balance between the sexes was evident even here.

Catholicism continued to present publicly an almost unchanged face until the end of Pius XII's pontificate in 1958; highly clerical, intellectually repressed, well organized, expansive, more tightly controlled from Rome than ever, insistent upon the use of Latin, retaining apparently untouched its medieval legacy, institutional, and liturgical. Yet processes of modernization had already gone far, intellectually especially in western Europe, institutionally elsewhere with the appointment of Asian and African bishops, and the creation of numerous cardinals to represent every continent. For the first time since the fourteenth century, the majority of cardinals were not Italian. It needed only the gentle touch of John XXIII and the calling of the Second Vatican Council (1962–65) for a truly revolutionary process to set in, only part of which was formally approved by the Council. The intellectual debate unleashed by 4 years of conciliar discussions set a movement in train which went well beyond the letter, though often not far beyond the spirit, of the Council's decrees. Within a few years Latin was replaced by the local vernacular in the liturgy and theological education; scholastic theology almost disappeared from the seminaries; hitherto frigid relationships with the Orthodox and Protestant churches gave way to a near-frenzied spate of official ecumenism.

In strictly theological terms the changes of Vatican II may seem less than revolutionary, but in the wider terms of the sociology of Catholicism, its sense of identity and shared ethos, they undoubtedly were. Symbolic ramparts, such as the denial of the cup in communion to the laity, which had been fiercely maintained against Protestantism for centuries, were dismantled quickly. It was not only that the barriers which made Catholics feel they were different – from Latin to Friday abstinence from meat – had almost disappeared, and with them the shared certainty that the pope was always right. It was also that Catholics now found themselves divided in a way that had never previously been the case in modern times. If, for Pope Paul VI (1963–78) and the official church, Vatican II had gone just far enough this was challenged from both right and left. For the former it had gone much too far, betraying authentic Catholic values in the pursuit of ecumenism, while for the latter it had failed through evading the most awkward issues, notably that of Roman centralization. The conflict hardened particularly over two matters that refused to go away: contraception and the law of celibacy for priests. Pope Paul had prevented the Council from discussing either. Contraception was passed to a papal commission that actually recommended a change in teaching. Paul rejected the recommendation and in 1968, in the encyclical *Humanae Vitae*, reaffirmed the use of contraception to be always wrong. He was quite unprepared for the degree of open opposition this aroused and for the consequent decline in confidence in papal teaching, particularly among the more thoughtful laity, who disregarded widely the papal ban. Again, despite numerous appeals, at least to allow married men to be ordained, both he and John Paul II (1978–) repeatedly renewed insistence on the law of celibacy in face of the near-catastrophic decline in the number of priests in the western world in the 30 years following the council, due both to the withdrawal of priests from the ministry, usually to marry, and a linked fall in recruitment. A further factor was that minor seminaries were now seen as obsolete and closed almost everywhere in Europe and North America in the early post-conciliar years. The credibility of the law had been internally undermined by several things: Vatican II's praise for the married clergy of Catholic Uniate churches in the east, its permission for married men to be ordained to the diaconate, and, finally, the acceptance of married priests, especially in Britain, if they were converts who had been ministers in another church previously.

Opposition to contraception set Vatican policy against almost all international attempts to curb the rapid rise in global population. This, together with papal rejection of the ordination of women and refusal to open the door to a married priesthood, did much to restore the nineteenth-century image of official Catholicism as incurably reactionary and anti-modern. At no time since the Middle Ages has the papacy appeared to matter so much within public policy as in the age

of John Paul II but its image has reverted to that of Pius XII. In papal history John XXIII could seem an aberration.

Shortage of priests resulted in the closing or merging into one of many parishes, especially in France, and the transfer of duties from priests to deacons or to the laity in general. Women, who in the past had never been allowed even to serve at mass, were now enrolled in thousands as 'eucharistic ministers.' Moreover, the religious orders, both male and female, whose numbers had grown enormously between 1850 and 1960, went into rapid decline in Europe and North America. In practice a rapid declericalization of the Church's ministry was taking place. While this could be seen as advantageous, its value was partially undermined by the haphazard way in which it happened, lack of a positive pastoral strategy and continued insistence by Rome that the traditional clerical pattern remained normative.

Behind the conflicts over practice and the loss of confidence in authority there were now huge differences in theological perception. If Vatican II seemed almost unbelievably radical to Catholics content with the church of Pius XII, the policies of John Paul II seemed almost unbearably reactionary to Catholics who had found the Council a harbinger of religious liberation. Many 'progressive' Catholics, while deeply loyal to the unity of a worldwide communion and to the idea of a central 'see of unity,' not only rejected much current papal teaching, particularly that of John Paul, whose personal theology often appeared fundamentally preconciliar, but had come to share so full a sense of fellowship with Christians of other churches that intercommunion had come to be seen as obviously right and was practiced frequently with the encouragement of many priests, despite condemnation by authority. 'Conservatives' accused 'progressives' of practicing an '*à la carte*' Catholicism. 'Progressives' charged 'Conservatives' with 'integralism,' confusing the substance of Catholicism with clericalism and Ultramontanism. While 'Progressives' tended to define Catholicism in terms inclusive of the principal insights of the Reformation as constituting a necessary critique of the one-sidedness of the late medieval church, 'Conservatives' continued to define it as inherently contradictory to Protestantism. Vatican II had, in fact, led to a profound questioning of the model of Catholicism as it developed across a 1000 years from the Gregorian reform through the Counter-Reformation to the decrees of Vatican I and Ultramontanism. How far that model will in consequence alter in the long run remains unclear.

It may well be claimed that the areas where this contestation is most in evidence are also areas of church decline. As against that decline, most obvious in western Europe, is to be set a huge growth of Catholicism in Africa, Latin America, and several parts of Asia. Brazil is now the country with by far the largest number of Catholics. Here too, however, there are great differences. Latin America has suffered an acute shortage of clergy for far longer than Europe and most of the issues dividing the church in Europe are evident here too, as also in the Philippines. The advance of Pentecostal Protestantism is also altering the hitherto almost exclusively Catholic character of Latin American Christianity. Elsewhere in Asia and in most of Africa the number of priests and religious has escalated. Religious orders whose recruitment plummeted after Vatican II in their traditional areas have grown remarkably in Africa, India, and Indonesia. For the Jesuits in particular India became the most flourishing of provinces, but the growth of houses of contemplative nuns in Latin America also is noteworthy. Nevertheless throughout the southern hemisphere there are profound tensions between Roman control and local processes of inculturation and political orientation. The future of Catholicism as the principal constituent within Christianity may depend upon how far a lasting *modus vivendi* is established between pope and Curia on the one hand and the young churches of the south on the other. John Paul's understandable preoccupation with Poland and eastern Europe is unlikely to be continued by his successors. Brazil, Argentina, India, the Philippines, Nigeria, and the Congo may matter more for the shaping of Catholicism in the twenty-first century than traditional pillars such as France, Spain, Ireland, and Poland, though the mediating position of the vast church of the United States may still prove to matter most of all. While the decline in church practice or missionary enthusiasm in France or Ireland is undeniable, there is as yet little evidence to suggest that world wide the strength of Catholicism is in decline and, if the reforms of Vatican II have led to a shaking of the foundations unprecedented for many centuries, they have also contributed to a massive renewal evident above all within the third world.

See also: Catholicism and Gender; Charisma and Charismatic; Christianity, Origins of; Historical Thought and Historiography: Christian Tradition; Historical Thought and Historiography: Classical Period (Especially Greece and Rome); Middle Ages, The; Religion, History of; Religion, Sociology of; Religion: Evolution and Development; Renaissance; Revolutions of 1989–90 in Eastern Central Europe.

Bibliography

Alberigo, G. (Ed.), 1995. History of Vatican II. Orbis, Maryknoll, NY.
Aubert, R., 1978. The Church in a Secularised Society. Paulist Press, New York.
Cleary, E.L., Stewart-Gambino, H. (Eds.), 1992. Conflict and Competition: The Latin-American Church in a Changing Environment. Lynne Rienner Publishers, Boulder, CO.
Duffy, E., 1997. Saints and Sinners: A History of the Popes. Yale University Press, New Haven, CT.
Ellis, J.T., 1969. American Catholicism. University of Chicago Press, Chicago.
Flannery, A., 1992. Vatican Council II: The Conciliar and Post-conciliar Documents. Dominican Publications, Dublin, Eire.
Gannon, T.M. (Ed.), 1988. World Catholicism in Transition. Macmillan, New York.
Gifford, P., 1998. African Christianity: Its Public Role. Indiana University Press, Bloomington, IN.
Greeley, A.M., 1977. The American Catholic: A Social Portrait. Basic Books, New York.
Hastings, A., 1989. African Catholicism. SCM Press, London.
Hastings, A. (Ed.), 1991. Modern Catholicism. SPCK, London.
Hastings, A. (Ed.), 1999. A World History of Christianity. Eerdmans, Grand Rapids, MI.
Hebblethwaite, P., 1984. John XXIII: Pope of the Council. Chapman, London.
Hebblethwaite, P., 1993. Paul VI: The First Modern Pope. Paulist Press, New York.
Hennesey, J.J., 1981. American Catholics: A History of the Roman Catholic Community in the United States. Oxford University Press, New York.
Hornsby-Smith, M.P., 1987. Roman Catholics in England: Studies in Social Structure Since the Second World War. Cambridge University Press, Cambridge, UK.
Keough, D., 1990. Church and Politics in Latin America. St Martin's Press, New York.
McBrien, R.P., 1994. Catholicism. Harper, San Francisco.
McBrien, R.P. (Ed.), 1995. The Harper Collins Encyclopedia of Catholicism. Harper-Collins, New York.
O'Farrell, P., 1977. The Catholic Church and Community in Australia: An History. Nelson, West Melbourne, Australia.

Catholicism and Gender

Rosemary Radford Ruether, Claremont Graduate University, Claremont, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article surveys the developing views of gender in Roman Catholic thought and practice. It includes New Testament thought, the era of the Fathers of the Church (AD 200–600), Medieval Latin Christianity, Counter-Reformation Catholicism, and Modern Catholicism through the Second Vatican Council to the pontificate of Benedict XVI.

New Testament Era

Christianity in its New Testament beginnings appeared to offer a gender-inclusive promise of redemption. Both women and men entered the baptismal waters equally, to be freed from the corrupting power of the old Adam and to rise to newness of life in Christ. This vision of an inclusive community of redemption was rooted in the ministry of Jesus and the early Jesus community. Here it was suggested that women and other despised people who suffered from various forms of impurity and subordination are not just included as equals in a new community of redemption, but this new community represented a radical dismantling of the old system of social hierarchy. This vision was expressed in the baptismal formula cited in St. Paul's *Letter of the Galatians* (3:28):

> For as many of you as were baptized into Christ have put on Christ. There is neither Jew nor Greek, there is neither slave nor free, neither male nor female; for you are all one in Christ Jesus

This baptismal theology probably preexisted Paul himself and was taken over by Paul from the Christian community he joined. In its roots in Jewish Christianity, this theology may have suggested a redemptive return to the pregendered humanity before the 'fall' into male and female, an idea found in Philonic Judaism (Hellenistic Judaism as written by Philo Judaeus) expressed in celibate communities of spiritual life, such as the Egyptian Therapeutae (a Jewish contemplative community of the first century). Earlier versions of this theology focused only on the transformation of the gendered pair in redemption. This view is found in several early Christian gospels, such as the *Gospel of the Egyptians*. Here Jesus replies to Salome's question about when redemption will happen by saying, "when you tread on the garment of shame and when the two become one, the male with the female, neither male nor female." Similarly, in the *Gospel of Thomas*, "Jesus said to them, 'when you make the two one and you make the inside as the outside and the outside as the inside, and the above as the below, and when you make the male with the female into a single one, so the male will not be male and the female not be female, then you shall enter the Kingdom'" (22:4-7).

In the context of the early Christian mission where slaves and free persons, Jews and gentiles, mingled in assembled communities, some leaders extended this theology to a three-paired formula of Jew and Greek, slave and free, male and female. It was this three-paired formula that Paul took over in the baptismal theology he cites in Galatians. But Paul did not actually focus on the slave/free or male/female pair. His main interest was in the Jew/Greek dichotomy, or as he put it, "For in Christ Jesus neither circumcision nor uncircumcision is of any avail, but only faith working through love" (Gal. 5:6). Only when a conflict arose in the church in Corinth, which had a constituency that strongly supported gender equality and emancipation from slavery, did Paul take heed of these other social implications and begin to formulate a theology of redemption to curb them.

In his letter to the Corinthians, Paul addresses a number of issues relating to the status of women in the Church. One of these is celibacy. Some women who exercised prophetic roles in the liturgy had embraced celibacy, which they understood as freeing themselves from subordination in the patriarchal household. Paul responds to this by cautioning against celibacy as a temptation to fall into immorality. Despite his own celibacy, Paul advises that almost everyone else should marry and maintain sexual relations, although they might abstain briefly for prayer. Underlying this advice is his belief that redemptive transformation has not yet happened and so the Corinthians cannot act as if they were already in a state of freedom from sin through baptism (I Cor. 7).

Paul then moves to the issue of women covering their heads when they pray and prophesy. He opens this discussion with the assertion of a hierarchical theology of 'orders of creation.' God is the head of Christ, Christ is the head of the male, and the male is head of the female. This hierarchy has not changed for the baptized. So women have no right to remove their head coverings, which symbolize their subordinate status under the male. Moreover woman in herself is not the image of God. The male should not cover his head because he is the image and glory of God, but the women possesses this image only secondarily under the male (I Cor. 11: 7).

Paul then turns to the relation of speaking in tongues to orderly intelligible speech. Because women were not trained in didactic speech and were freed to speak in public only when ecstatic speech was allowed, this argument had the effect of subordinating women's speech to that of male leaders, a consequence that was enforced by the (later added) commandment that "the women should keep silence in the churches for they are not permitted to speak but should be subordinated, even as the law says" (I Cor 14: 34).

Paul's successors in treatises such as Ephesians and Colossians sought to systematize his efforts to shore up social hierarchy in the church by spiritualizing equality in Christ and separating this inward transformation from social relationships. For them the hierarchical relations of the patriarchal family remains intact. This is expressed in commandments, such as "wives, be subject to your husbands," "children, obey your parents in everything, for this is your acceptable duty in the Lord," and "slaves obey your earthly masters in everything, not only while being watched and in order to please them, but wholeheartedly, fearing the Lord" (Col. 3: 18, 20, 22).

But more egalitarian visions of the church did not die out. The conflict between these two views continued into the second century. The deutero-Pauline (secondary Pauline writers, followers of Paul) letter of I Timothy sought to counteract women's claims to public ministry in the church by a patriarchal reading of Genesis I-3. The author asserts that women were created second in the order of creation but were first in sin. Their subjugation, reflecting their secondary place in creation, is redoubled as punishment for their primacy in sin. They may not preach but must keep silence. They will be saved only by accepting this subjugation in marriage, child bearing, and modest, submissive behavior (I Tim. 2:11-15).

But this letter also reveals the existence of alternative views among women in this church who are claiming an active ministry that the author seeks to curb. This alternative is reflected in writings such as the *Acts of Paul and Thecla* in which a converted woman takes the lead in itinerant ministry, resists imperial repression while baptizing herself, and claims Paul's authority as the one who commissions her to 'go and teach the word of God.'

The Patristics Era

Second-century Christianity was diverse and syncretistic. A number of movements, which were later marginalized by Orthodox leaders, advocated withdrawal from marriage and gender equality. Among these was Montanism, led by three prophets, two female, and one male. They saw themselves as possessed by the Paraclete sent by Christ and speaking in the name of the risen Lord. They believed they were living in the last days of world history, prepared for martyrdom, and awaiting the imminent arrival of the redemptive New Jerusalem.

Another expression of radical Christianity was Marcionism. Its founder, Marcion, was a dualist who saw the creator of the world and the giver of the law as the author of an oppressive universe of social hierarchies and violence. The true God of Love has sent his son Jesus into the world to liberate us from the laws of the evil Creator. Redemption gathers a community of equals that overcomes these divisions.

Valentinianism was another variant of radical egalitarian Christianity of the second century. The Valentinian school interpreted the fall of Adam and Eve and redemption in Christ through a story of cosmic devolution from and reintegration back into a higher divine world of divine beings. A female divinity, Sophia, is the mediator between the two worlds, bringing about its fall, but also calling it back into communion with the divine Pleroma. Valentinian communities were led by men and women who were seen as rising above gender division into androgynous unity.

In an early third-century writing influenced by Valentinianism, *The Gospel of Mary*, the risen Christ chooses Mary as his beloved disciple to interpret his message to the other disciples. Andrew and Peter then dispute her authority, but Levi defends her and even suggests that attaining redemptive humanity depends on being able to hear spiritual insight from women. Rather than contending against women, the true Christian should "put on the perfect humanity and preach the gospel, not laying down any rule or law beyond what the Savior said." For these Christians the Savior did not lay down any law subordinating women.

Other theological leaders, Tertullian in North Africa, Irenaeus in Gaul, and Clement and Origen in Alexandria, among others, criticize Marcionite and Valentinian views and establish an orthodoxy that rejects them. Women are eliminated from public ministry in these churches. Although women were recognized as having prophetic inspiration, they were to communicate their visions to the male leaders, not testify in the community assemblies.

By the fourth century the more radical versions of redemptive equality had been mostly defeated. The women who flocked into celibate Christian monastic life in the fourth and fifth centuries mostly accepted their segregation from public ministry and from preaching in the church. Yet the idea continued to linger in monasticism that the female ascetic, by virtue of renouncing marriage and sexuality, was no longer under male domination. A number of women from leading Roman families, such as Marcella, Melania the Elder and her daughter Melania the Younger, Paula and her daughter Eustochium, pioneered prayer and study groups in Rome and later built monastic communities in Jerusalem and Bethlehem.

The leading monastic circle of Cappadocia, led by Basil the Great and his brother Gregory Nyssa, was pioneered by their sister Macrina who turned the family estates into a religious community. Gregory Nyssa saw his sister as his spiritual mentor and wrote his major treatise, *On the Soul and the Resurrection*, as a dialogue with her. She is seen as the embodiment of divine wisdom who leads him through a process of increasing understanding of the journey of the soul from its fall into sin and death into restored glory in communion with God.

In the Latin Church, Augustine emerged as the major theological voice. He sought to reconcile these contradictions between creational patriarchy and redemptive equality in Christ. He recognized that to be redeemed, women must be affirmed as created in the image of God. But he also felt bound by the Pauline dictum of I Corinthians that the male has priority as the image and likeness of God. Augustine's solution was to split the image of God as a gender-neutral soul from woman's nature as female. Women possess the image of God and hence a redeemable soul in their spiritual humanness, but as females they are not the image of God, but rather the image of the body, the subjugated lower creation.

For Augustine, the woman was made subject to the male in the original creation. This subjugation has to do with her role as wife and sexual partner in the production of children. Furthermore, due to Eve's primacy in leading the male to

disobedience to God in Eden, women's original subordination has been worsened into coercive servitude. This subjugation has not been overcome by baptism nor is it lessened by celibacy. Monastic women are still subject to male church authorities. The Christian woman should accept her subordination as both an expression of her original place in creation and punishment for her primacy in sin. This subjugation will be overcome only in heaven when the created order of sex, procreation, mortality, and sin will be swallowed up in redemptive immortality. Only then will women become fully equal with men according to her redemptive merits.

Medieval Christianity

The Latin Middle Ages also saw further denigration of women. This affected the status of women as ordained ministers. The New Testament and Patristic periods understood women as ordained to the deaconate by the same ceremonies by which men were ordained. There is some debate in the twenty-first century whether women were ordained priests and bishops in some communities. This seems to have been the case for groups such as the Montanists and Marcionites. Orthodox Christians sought to deny women the priesthood, but women as ordained deacons continued into the Middle Ages. Some abbesses were ordained deacons. Ordination was understood functionally. A minister was ordained to specific services within particular communities.

In the twelfth to thirteenth centuries, this view was radically changed. Ordination was defined as conferring the specific power to 'confect' the Eucharist and was limited to males only who could exercise this power wherever their superiors assigned them. A renewed misogyny demonized women as 'the cause of our ruin.' Roman law was cited to subjugate all women, and Aristotle's anthropology used to prove that women were inherently defective as human beings.

The Dominican theologian, Thomas Aquinas (1225–74), developed his view of women based on Aristotle. Aristotle believed that the male seed possessed the entire formative power of procreation, while women provided only the 'matter' shaped by the male seed. Every seed could shape a male embryo. Only when the female matter was shaped insufficiently by this seed was a female produced who lacked full humanity, in intellectual capacity, moral will, and physical strength. For Aquinas, following Aristotle, women could not take any leadership role but was by nature under male subjugation.

Aquinas' view of women had implications for his Christology. God's Word had to be incarnated as a male to manifest 'perfect' humanity. It then followed that only males could be ordained priests to represent Christ. Women are not only barred from ordination juridically but are incapable by nature of validly receiving this sacrament. Women also cannot exercise any form of sovereignty in society, such as a ruling monarch, even though women as queens did have such roles in the Medieval world. Aquinas did accept that God might bestow prophetic gifts on women, although he believed they should exercise such powers only privately through male superiors. Yet he did not deny that women might possess personal holiness, at times superior to many men. Like Augustine, he saw women coming into equality with men according to their spiritual merits in heaven.

Yet many women did flourish as spiritual leaders in Medieval Christianity. Although barred from being priests, the Medieval Church continued the early Christian view that women could be prophets. This opened a significant space for women's spiritual gifts and teaching roles in religious communities. Unlike the Patristic era during which few women's actual writings are preserved, the Middle Ages provides us with many scores of women's religious works. A key resource for this existence of women's writings was women's monastic communities that cultivated women's literacy in Latin or the vernacular, possessed scriptoria where manuscripts were copied and libraries where they were preserved.

One of the leading medieval woman writers is abbess Hildegard of Bingen (1098–1179) who was accepted as a prophetess by the highest church authorities, including the Pope.

Hildegard wrote theological, botanical, and medical texts, as well as almost 400 letters, many liturgical songs and musical compositions and poetry, as well as supervising the artistic illumination of her visions. She wrote three major theological works, *Scivias* or 'Know the Way,' the *Book of Life's Merits*, and the *Book of Divine Works*, which lay out her visions of Creation, Fall, and Redemption and its ethical mandates for Christian life.

Hildegard sees the female divine powers of 'Wisdom and Love' (*Sapientia* and *Caritas*) as God's primary agents in shaping the creation of the Earth and guiding humanity after the Fall into redemptive history and its eschatological culmination. Although Hildegard accepts the gender and class hierarchies of her society, she sees those gifted with the prophetic Word, such as herself, as playing key redemptive roles for humanity and the Church.

An important movement for medieval Christian women was the Beguines. These were urban communities of women, mostly middle and working class, who took no permanent vows, were not enclosed in monastic walls, but supported themselves through manual labor, especially in textiles, cared for the poor and the sick, and taught children and women. The Beguines produced a number of leading mystical theologians. Among them are Mechtild of Magdeburg (1210–83) and Hadewijch (thirteenth century).

Mechtild, in her writing, *The Flowing Light of Godhead*, describes her vision of the journey of the soul in its love relation with God the Father and his Son. She joined a Beguine community in her early 20s where she lived for 40 years. In her old age, Mechtild moved to the learned women's monastery of Helfta where she was supported in writing the seventh and last book of her manuscript. Hadewijch was a Flemish Beguine who was a contemporary of Mechtild. In her *Poems in Stanzas*, she writes as a male knight errant pursuing the love of God as her Lady Love; while in her *Visions*, she pursues the love of God as a female soul longing for union with her Beloved.

Another important female mystic is the fourteenth-century English visionary Julian of Norwich (1343-after 1413). When she was 30, Julian had a series of visions of Christ dying on the cross and God's love for us. Retiring to an anchorhold (enclosed cell connected to a parish church), Julian lived the rest of her life meditating and writing on these visions. Notable in Julian's thought is her exploration of the Trinitarian nature

of God as both mother and father. She writes, "in our making God almighty is our kindly father and God all-wisdom is our kindly mother, with the love and goodness of the Holy Spirit."

Although Julian and other Medieval theologians, such as Bernard of Clairvaux, cultivated the feminine image of God as Wisdom, the most important female symbol for the divine is Mariology. Theologically, Mary is not divine, but the representative of the ideal human, but for Catholics she functioned as a divinity, installed in heaven next to her divine Son and mediating between God and humanity. Mariology grew over the two millennia of Catholicism. The New Testament Church made no direct Marian claims, affirming Jesus' virgin birth primarily as a Christological doctrine. Patristic Christianity elaborated the idea of Mary as the New Eve, parallel to Christ as New Adam, Mary's perpetual virginity, her title as Theotokos or Mother of God and the Assumption or belief that she was carried to heaven at death.

Medieval Christianity added the idea of the Immaculate Conception, the belief that Mary was conceived in the womb cleansed of original sin. This was disputed by some theologians, such as Thomas Aquinas, who preferred the view that she was sanctified in the womb later in gestation, but the Franciscans advocated the dating of this to her conception. This doctrine, along with the Assumption, were declared doctrines only in modern times (1854, 1950).

But the Middle Ages saw a great flowering of devotion to Mary. Her feast days proliferated, hundreds of churches were dedicated to her, and the Mary altar became a standard part of every church. Her relics multiplied in numerous shrines, and private devotions, such as the rosary, directed worshippers to pray to her daily. Theologically she was exalted as the embodiment of the unfallen human in its original and perfect state, the appropriate vehicle for God's incarnation, and the Queen of Heaven extending a forgiving hand to unworthy humans. As manifest in modern appearances, such as the Virgin of Guadalupe in Mexico, Mary becomes the central mediator of God's grace.

Counter-Reformation Catholicism

The Reformation era of the sixteenth and seventeenth centuries was in some ways a setback for women. Laws that allowed women membership in guilds and control of property were narrowing. The Protestant Reformation abolished celibacy, while championing the patriarchal family. Women could not be ordained and should not preach. Their sole vocation was that of dutiful wife and mother. Women lost many vocations and roles that had been available through celibate female communities.

In Catholic areas, monastic life for women continued, but the narrowing of women's legal and economic opportunities also happened there. The Catholic Counter-Reformation stressed the strict cloistering of women and resisted the efforts of many creative women of the period, such as Mary Ward in England, who sought to found new noncloistered activist orders with teaching and social roles for women. Women were either to be wives obedient to their husbands or cloistered nuns chanting prayers under the supervision of male church authorities.

A few Catholic humanists, such as Erasmus, championed women's expanded education. The maverick Catholic magician Agrippa von Nettesheim even suggested that women originally were not only spiritually equal to men but also superior to them as representatives of the divine Wisdom of God. He claimed that male domination was recent and expressed, not God's will or female inferiority, but simply male tyranny. But this view conflicted with the trends of the times and would not be heard again for two centuries, when the late-eighteenth-century feminist, Mary Wollstonecraft, championed women's rights in her treatise, *The Vindication of the Rights of Women* (1792). Feminism would not appear among Catholic women until after the Second Vatican Council in the 1960s.

Modern Catholicism

There was virtually no organized Catholic participation in the first wave of American feminism that ran from the Women's Rights Convention in Seneca Falls, New York, in 1848 to the winning of women's suffrage in 1920. The reasons for this lie in several factors; first, the Catholic bishops were vehemently opposed to feminism, associating it with secular modern apostasy to Christian and 'family' values, represented by child labor laws, birth control, the Equal Rights Amendment, divorce, liberalism, and socialism. Many bishops explicitly rejected women's suffrage itself, claiming it would take women out of their proper sphere in the family.

The view of gender that came to prevail in nineteenth- and twentieth-century Catholicism was no longer one that stressed women's inferiority, but rather their complementarity to men. Women were different and even in some cases better than men, but their specific nature lay in their passive auxiliary being. Women were to be 'Marylike' rather than 'Christlike.' They are to represent the church in its subordination to Christ and to masculine leadership. They should not take any such leadership themselves. In these views, Catholicism echoed and put its particular stamp on the romantic views of gender complementarity common to the period.

The conflict between official Catholicism and feminist advocacy for women has developed in the Catholic community in the post–Vatican II era. Vatican II itself largely ignored women. Women, even women religious, were absent from the delegates to the Council. Protest about women's absence by Cardinal Leon-Joseph Suenens in 1962 brought a slight change. A few women religious and lay women were allowed as 'auditors' without vote in the later sessions of the Council. Three even served on commissions that drafted the final Council statements, specifically those on 'The Church in the Modern World' and on the 'Apostolate of the Laity.'

The final documents of the Council have a few concessions to women's new role in work and legal equality in modern society. It is said that women are claiming a new equity for themselves before the law and in fact, and the participation of women in cultural life should be acknowledged and supported by the Church. These statements are echoed in Pope John XXIII's encyclical *Pacem in Terris* (1963). Here it is said that "women are becoming ever more conscious of their human dignity. They will not tolerate being used as mere material

instruments, but demand rights befitting a human person in both domestic and in public life" (Sec. 41).

These concessions to women's role in public society did not include changes in the life of the Church itself, in terms of women's ordination or in teachings on sexuality and reproduction. Both of these issues were being challenged in the 1960s. Mainline Protestant churches in Western Europe and North America had begun to ordain women in increasing numbers in the 1960s. In 1976, the American Episcopal Church endorsed women's ordination, a change that brought women into the priesthood in a tradition closely related to Catholicism.

The Roman Magisterium responded to these developments with an encyclical that same year insisting that women could not be ordained because women lack the capacity to 'image Christ.' Maleness and the representation of Christ as a male were thus identified in a way that many Catholics came to see as questionable. Some theologians, such as the eminent Karl Rahner, even suggested that such literalness about Christ's maleness as necessary for his sacramental representation is heretical. Far from ending the discussion, support groups for women's ordination have grown among Catholics since 1975 in Europe and North America.

In 1975, U.S. Catholics convened the first Women's Ordination Convention in Detroit. From this conference emerged the Women's Ordination Conference that works on this issue in Roman Catholicism nationwide. In 1996, the Women's Ordination Conference Worldwide was founded at the First Women's Synod in Gmuden, Austria, to promote women's ordination, particularly in the Catholic church, worldwide. This group is closely connected with the U.S. Women's Ordination Conference and has held several global meetings.

Another stage of this challenge developed in 2002 when seven women from Germany, Austria, and the United States were ordained priests on a boat in the Danube River in Austria by several male Catholic bishops. Several of these women were subsequently ordained bishops by these Catholic bishops. They, in turn, have continued to ordain women deacons, priests, and bishops, creating a new movement called Roman Catholic WomenPriests. By 2011, there were five women bishops, 65 women priests, and 13 women deacons in the United States and one bishop, seven priests, and one deacon in Canada.

The movement is organized into a number of administrative districts in the two countries that oversee theological education, preparation for ministry, and ordination of women seeking priesthood. These women priests in turn have created a variety of ministries, worshipping communities, retreat centers, and chaplaincies in prisons, hospitals, and schools. This movement has been strongly condemned and declared excommunicated several times by the Vatican. But this has not deterred their expansion and their continued claim that they are validly ordained as priests in the Catholic tradition.

Another area of contention among Catholics has been the traditional teaching that sex (allowable within marriage alone) is primarily for procreation and cannot be separated from its procreational potential. Birth control (other than the rhythm method) is forbidden. As moral theologians and ordinary lay Catholics began increasingly to question these teachings, and called for this issue to be discussed in the Second Vatican Council, Pope Paul VI decided to steer this discussion away from the Council by convening a special commission to study it. The commission meet from 1964 to 1967. Its membership expanded, and it came to include three married couples from the Catholic Family Movement. One of these couples, Pat and Patty Crowley from the United States, collected documentation from members of the Family Movement, which testified to the anxiety caused by the rhythm method and its unreliability as a method of family planning.

Members of the Commission were deeply moved by these findings, and they voted to change the teachings of the Church to one that accepted all medically approved methods of birth control within the context of committed, child-raising married families. But this official report, which was supported by the vast majority of the Commission, was rejected by Pope Paul VI after he had been advised by a few dissenters that such a change would threaten the Church's claim to infallible teaching. The reaffirmation of the anti-birth control teachings in the papal encyclical *Humanae Vitae* (1968) created a storm of protest throughout the Church. Many priests and moral theologians openly dissented. Many lay Catholics simply have chosen to ignore Church teaching on this subject. Yet the Vatican under Popes John Paul II and Benedict XVI have continued to insist on the inerrancy of this teaching, deepening the conflict between the laity and the Magisterium of the Catholic Church on the question of contraception.

Realizing the deeply conflicted state of relations between the official church and women, the American Catholic bishops voted in the early 1980s to engage in an official dialogue between the bishops and the Women's Ordination Conference. After several years of meetings, the bishops who were members of the dialogue persuaded the other bishops that they should undertake a pastoral letter on women in the Church. They conducted a series of listening sessions with Catholic women across the United States to hear their issues. They then drafted an epistle that declared that men and women should be partners in the family, society, and the church. This epistle condemned sexism as a 'sin.'

But the Vatican intervened in the process of drafting this letter, demanding the condemnation of birth control and women's ordination as a part of the letter, and that the model of relationship between men and women should be one of complementarity, not partnership. The result of these interventions was a decision by the American bishops to table the letter, rather than issue a document that was likely to worsen, rather than improve, relations with women in the church.

These conflicts between teaching and practices of the Catholic Church and the growing consciousness of their rights to equality among Catholic women (and their male supporters) is by no means limited to North America. Catholic feminist theological networks have grown in Western Europe, as well as in Latin America, Asia, and Africa. With the growing lack of adequate ministry by celibate male priests throughout the world, more and more of the ministerial work in Catholic parishes and chaplaincies is being done by theologically trained laity, especially by women.

Thus, the conflict between the Catholic hierarchy and women, exacerbated by the reactionary policies of Popes John Paul II and Benedict XVI that undermined the reforming hopes of the Second Vatican Council, shows no signs of improving in the second decade of the twenty-first century. The issues go far beyond the inclusion of a few token women in clerical office.

They call for a fundamental rethinking of ancient traditions rooted in male ecclesial patriarchy and a view of femaleness and sexuality as expressions of inferiority and sin to be distanced from sacramental office.

A thoroughgoing revision of these patterns any time soon in official teachings is unlikely. Since women who claim more egalitarian views about themselves and their rights to decide about their own bodies show no signs of disappearing in large numbers from the Catholic Church worldwide, this conflict will continue to be present for the foreseeable future as a deep schism concerning the meaning of gender identity in Catholicism.

See also: Catholicism; Feminism in Anthropology; Gender History; Historical Thought and Historiography: Christian Tradition.

Bibliography

Borresen, K.E., 1981. Subordination and Equivalence: The Nature and Role of Women in Augustine and Thomas Aquinas. University Press of America, Washington, DC.

Macy, G., 2008. The Hidden History of Women's Ordination: Female Clergy in the Medieval West. Oxford University Press, Oxford.

McClory, R., 1995. Turning Point: The Inside Story of the Birth Control Commission. Crossroads, New York.

Raming, I., 1976. The Exclusion of Women from the Priesthood: Divine Law or Sex Discrimination? Trans. N Adams. Scarecrow Press, Metuchen, NJ.

Ruether, R.R., 2012. Women and Redemption: A Theological History. Fortress Press, Minneapolis, MN.

Swidler, L., Swidler, A., 1977. Women Priests: A Catholic Commentary on the Vatican Declaration. Paulist Press, New York.

Weaver, M.J., 1985. New Catholic Women: A Contemporary Challenge to Traditional Religious Authority, first ed. Harper & Row, San Francisco, CA.

Cattell, Raymond Bernard (1905–98)

John R Nesselroade, University of Virginia, Charlottesville, VA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 1545–1550, © 2001, Elsevier Ltd.

Abstract

Raymond B. Cattell was born on 20 March 1905 in Staffordshire, England. He was educated in England, where he began his academic career in chemistry but eventually moved into psychology. He received his Ph.D. in 1929. Over a long and very productive career, Cattell managed to keep one foot in the theory of behavior and the other well-grounded in empirical research to test and extend his theories. His approach to the study of behavior was highly quantitative and dominated by the model of factor analysis. Cattell applied his methods to the domain of personality, broadly defined, and made salient contributions to the measurement and modeling of human abilities, temperament, motivation, learning, and the nature of developmental change. He pioneered the study of intraindividual variability in personality, both methodologically and substantively. Cattell was the central force in organizing the Society of Multivariate Experimental Psychology, in founding its journal *Multivariate Behavioral Research*, and in producing the first edition of *The Handbook of Multivariate Experimental Psychology*. He died on 2 February 1998 at the age of 92.

In describing the psychologist Raymond Bernard Cattell's comprehensive research program, Child, in a single sentence, gave an amazingly complete and accurate description: 'His major concern was to map out an integrated theory of human intellectual, temperamental and motivational characteristics within the context of hereditary and environmental influences using multivariate methods of analysis' (Child, 1998, p. 356). In trying to convey the sweep of Cattell's theoretical and empirical work Goldberg (1968) referred to him as 'psychology's master strategist' (p. 617). Indeed, the sheer audacity of venturing into so many major research areas during one career is unlikely to be seen again for some time.

Raymond B. Cattell was born on 20 March 1905 in Staffordshire, England. He was raised in Devon, near the town of Torquay. Cattell began his academic career in chemistry at University College, London and received his Bachelor of Science degree with first class honors in 1924. Drawn to the study of psychology, Cattell began work on a doctorate at King's College, also in 1924. After completing his Ph.D. in 1929, Cattell did some teaching at University College, Exeter before becoming Director of the School Psychological Services and Clinic at Leicester in 1932.

In 1937, E. L. Thorndike invited Cattell to come for a year to Columbia University as a research associate. Subsequently, he also spent brief periods at Clark, Harvard, and Duke universities. During his time at Harvard, Cattell began seriously to develop the methodology for intensively studying the single case (P-technique factor analysis) in which he remained interested throughout his career. Cattell acknowledges that with his ' ... office ... having been next door to Gordon Allport's, the latter may also have played an initiating part' but goes on to lament that although Allport, too, was ' ... in pursuit of uniqueness of personality ... ' he had ' ... some unreadiness to come to grips with statistical models' (Cattell, 1984, p. 141). Exhibiting a certain disdain for those who refused to use 'proper' methods, no matter how celebrated their status, was one of Cattell's stable characteristics.

Cattell worked for a time during the latter years of World War II in the Adjutant General's Office. In 1945 he became a research professor at the University of Illinois in Urbana-Champaign and remained there until 1973 when, in part due to age restrictions then in force, he retired. He moved on to the University of Hawaii where he maintained a special appointment for a few years. When that appointment ended, Cattell stayed in Honolulu and continued to work informally with several younger colleagues and to publish an occasional paper. One of his last publications (McArdle and Cattell, 1994) brought to a close his 50-year affair (Cattell, 1944) with the beguiling concepts of 'parallel proportional profiles' and 'confactor rotation,' more of which will be said later.

By far the longest and most productive portion of Cattell's scientific career occurred while he was Director of the Laboratory of Personality Assessment and Group Behavior (later the Laboratory for Personality Analysis) at the University of Illinois. It was from this setting that his influence on the development of the science of psychology hit its maximum. The sheer number of big ideas and level of productivity in following them up during those years are remarkable. His success in claiming his 'fair' share of CPU time on the Illiac, the university's automatic computer, was legendary. Although by today's standards, its capacity was meager, at the time the Illiac allowed Cattell to do large factor analyses in weeks and months that would otherwise have taken years.

Contributions to the Methods and the Substance of Psychology

Raymond B. Cattell is an exemplar of a class of psychological researchers who were heavily invested in the application of quantitative models to the study of behavior. Also included were the likes of Cyril Burt, Hans J. Eysenck, J. P. Guilford, Charles Spearman, Godfrey Thompson, and L. L. Thurstone. Let it be said openly that this is not a list of like-minded individuals who necessarily had more in common than not. Rather, it is a list of important contributors to the development of psychology as a science who relied substantially on the models and methods of factor analysis in their empirical research pursuits. In both strategy and tactics, they differed from each other in the use of factor analytic methods.

Because of Cattell's heavy reliance on factor analysis, it is not possible to separate the substance from the methods. In fact, in an earlier review of Cattell's research program in *Theories of Personality* (Hall and Lindzey, 1978), the chapter heading is 'Cattell's Factor Theory.' Therefore, a discussion of his work can be usefully begun with a methodological introduction.

At the heart of Cattell's research program lay the basic equation of the common factor model. Cattell referred to it as the factor specification equation and wrote it as:

$$a_{ji} = b_{j1} \cdot f_{1i} + b_{j2} \cdot f_{2i} + \cdots + b_{jk} \cdot f_{ki} + u_{ji}$$

where a_{ji} is a score for person i on behavior (or act) a_j, and fqi ($q = 1, 2, \ldots, k$) represent the ith person's endowments or scores on the k common factors, the b_{jq} ($q = 1, 2, \ldots, k$) are weights (factor loadings) that specify the amount to which any individual's score on the qth factor contributes to (in Cattell's theoretical framework, *determines*) his score on variable a_j, and u_{ji} is a contribution to a_{ji} that is unique to both the behavior and the individual. The u_{ji} terms include both errors of measurement and portions of the observed score that are reliably measured, but fall outside the purview of the common factors defining the interrelations of a given set of variables.

The b_{jq} values, which Cattell referred to as situational indices (later, behavioral indices), brought the situation and the individual's traits into the same equation so that both individual and situational information could be used to predict behavior. Suppose, for example, a given individual had a very high endowment on the factor Cattell labeled Exvia, which was similar in nature to the individual differences dimension called Extraversion in many personality research and theory contexts. If, on the one hand, the b_j for Exvia with respect to performance of a given a_j was small, then that high level of Exvia had little or no bearing on that person's a_j score. On the other hand, however, if b_j for Exvia on performance of a given a_j was large, then that high level of Exvia had major influence in determining that person's a_j. By contrast, for the person with a near zero level of Exvia, little would be contributed to their performance on a_j by the Exvia factor even if the b_j for Exvia was large, because the product $b_{j,exvia} \cdot f_{exvia,i}$ would be close to zero.

This factor specification equation is thus a 'working model' of the individual differences orientation in that the differing amounts of the latent attributes possessed by different people (variation in the f_q) are 'converted' into differences in their scores on the observed variables (variation in the a_j). By knowing the amounts of a person's trait endowments and the magnitude of the contribution the traits made to manifest variables, one could predict an individual's scores on the manifest variables, that is, behavior.

Within that framework, the scientific agenda is rather clear: define the latent factors as clearly as possible, construct tests to measure them accurately, and develop empirically based estimates of the b values for the manifest variables, the a_j, of interest. Then one has made the prediction objective of science reachable and can also elaborate an explanatory system as well by detailing more and more regarding the nature of the factors (e.g., see how the factors behave across different media of observation and how stable they are across time, examine their genetic and environmental variance contributions, plot their age changes, etc.). Attacking these several objectives was a major part of Cattell's research program.

Thus, despite knowing the adaptations in the methods of the physical sciences had to be made in studying behavior, Cattell's methodological approach showed a deep reliance on fundamental principles of physical science. He believed that factors were causal influences on observable variables and that, in analytical work, when the factors had been properly rotated to their 'causal influence' identities, the factor loadings (the b_j values in the specification equation) would be proportional for a given set of behaviors (the a_j values in the specification equation) across different samples. This represented an explicit commitment to a fundamental objective of scientific research: the identification of relationships that remain *invariant* under different transformations. In his context, Cattell formalized this notion in the terms 'parallel proportional profiles' (the results) and 'confactor rotation' (the method) mentioned earlier. Along the way, he worked on the development of methods for assessing the similarity of factor loading patterns so that one could render some judgment about factorial invariance even without reaching the ideal parallel proportional profiles.

Cattell was also firmly committed to the principles of rigorous, quantitative measurement and scaling and once presented, with characteristic lack of modesty, a proposed more basic set of concepts (Cattell, 1964) to replace the prevailing notions of reliability and validity. In so doing, he again challenged the status quo regarding the role of item homogeneity in evaluating the quality of measuring instruments.

Cattell placed a strict reliance on mathematical representations of phenomena of interest although, in his case, this amounted generally to using some variant of the common factor model. His insistence on the introduction of rigor into the definition of psychological concepts left its mark in a number of ways. He argued that ambiguity and confusion about concepts could be reduced by discarding as many of those in vogue as possible, discovering and defining new ones, and giving them new names to mark their unique status. In the process, he delivered such labels as 'cortertia' (cortical alertness) and 'threctia' (threat reactivity) to identify major individual differences in dimensions of personality.

A monograph on real-base, true zero scaling (Cattell, 1972) illustrates the concern he held for fundamental issues of scaling, the commitment and energy with which he would attack a problem, provide a solution, and advance it for discussion and debate. Debate, however, was not always the reaction evoked by his proposals.

In addition to this strong analysis and modeling orientation, Cattell's research program was buttressed by several other key attributes. Where others might carve up the study of behavior to identify more clearly a piece for themselves in the broader scheme, Cattell did so as away of more systematically approaching the whole. Much of his work was organized around the tripartite division of abilities, temperament, and motivation. But he enriched this classical triumvirate by distinguishing between surface and source traits, estimating heritability coefficients, and incorporating concepts of psychological states, learning theory, and situational influences. He introduced the multiple abstract variance analysis (MAVA) design to incorporate different family configurations into the estimation of heritability coefficients and in the design of correlational studies he saw not one, but six important ways to interrelate variables, persons, and occasions of

measurement. He systematized the latter in a heuristic often referred to as the 'data box' or covariation chart (Cattell, 1952) which is still frequently encountered in the literature.

As an example of his capacity for orderly abstraction and the willingness and ease with which he would think 'big,' Cattell was not satisfied to name and report the results of his many factor analyses. Instead, he proposed the development of a Universal Index (U. I.) of factors, the first 15 of which (U. I. 1 through U. I. 15) were to be reserved for what he considered to be the well-established human ability factors. He then proceeded to assign the next 17 numbers (U. I. 16 through U. I. 32) for factors based on objective tests on which he had published. He invited further contributions to this indexing system but, as one might expect from even a passing acquaintance with human nature, one person's proposal for a Universal Index is hardly likely to be adopted readily by others.

In addition to a frequently observed trait of successful research scientists – a desire to be first and to receive credit for it – Cattell was gifted with a high level of intelligence and capacity for abstract thought. He made up many of the items for the Culture Fair Intelligence tests (Cattell, 1954) himself; took delight in the rapidity with which he could solve items from 'would be' intelligence test item writers and, in return, relished 'stumping' them with a recently written item or two of his own.

The levels of abstraction with which Cattell felt quite comfortable were in many cases a far remove from data. For example, second-order factor analyses (factor analyzing the intercorrelations of factors) were routine in his laboratory. Third-order factor analyses, while not routine, were completed on several occasions to which published papers attest. Along the way, at least one fourth-order factor analysis was conducted.

Cattell was not inclined to be tolerant of others' lack of knowledge of his system. Those who worked in his laboratory were apt to be rebuffed when they would occasionally suggest ways by which he might strengthen the bonds of communication with his readers. 'Let them go and read my 1957 book' (Cattell, 1957), was the reply heard on more than one occasion to the suggestion that a little more detail might be helpful to the reader. Nor, as was already mentioned, was he kindly diposed toward those he felt were not using the multivariate tools of the trade wisely.

Cattell was an avid reader (and writer) of poetry and also enjoyed biographies from which he enriched his own writing. His enormous legacy of books, chapters, and articles, numbering well over 500 publications, exhibit many examples of his caring about the craft of writing as well as the ideas being written about. He would search for just the right word or turn of phrase – and did so amazingly quickly! When he was well into his eighties the author once asked him in what areas of his intellectual functioning he could detect losses. He responded that he didn't act on visual input as rapidly as he used to when, for instance, driving his automobile. (This was true!) But, he very quickly added, he could still retrieve adjectives, etc., quite handily when he was writing and he was very pleased about it. Some of his 'classic' remarks, many of which are buried in footnotes, will likely go unheeded by the future generations, and that will be a stylistic as well as a substantive loss. For example, his characterization of the hyperplane as 'the footprint of a causal influence' or his accusation that by developing and making widely available the by then nearly ubiquitous orthogonal analytic rotation program, Varimax, Henry Kaiser had 'put opium on the market,' are illustrative.

Cattell's deep reliance on factor analysis was symptomatic of his belief that the workings of variables should be examined with as little interference as possible from the observer. The methods of factor analysis could then be used to tease out the key patterns of relationships. As any craftsman, Cattell felt strongly about having access to proper tools and worked to evolve factor analytic methods during the length of his career. He did this both by formalizing design with the presentation of the covariation chart or 'data box' and by restlessly pushing technological developments. The ever present need for faster, more efficient methods for conducting factor analysis if his research goals were to be realized, directly resulted in software applications such as Maxplane, Procrustes, and Rotoplot to facilitate rotation of factors to a final solution. The Rotoplot program, for instance, displayed the pairwise factor (actually, reference vector) plots on an oscilloscope and filmed them by 35 mm camera mounted on the oscilloscope. As soon as the film strip could be processed, Cattell would project the plots on his office wall and write down the tangents of the angles of rotation for another run, urgently trying to achieve 'maximum simple structure' so that factors could be interpreted and the article written and submitted. At the time that Rotoplot was invented, others who were doing graphical rotation of factors were having to have each pairwise plot rendered by triangle and T-square.

Bridging Psychology's Past and its Future

Cattell and his research program represent an important bridge between psychology's past and its future. While he was alive, Cattell was a personal, as well as scientific link to the psychology's past. On occasion, he would tell of conversations (sometimes, encounters) with the likes of Burt, Fisher, Pearson, and Spearman. Although he was not a student of Thurstone, he seized on the tool of multiple factor analysis which Thurstone was so instrumental in bringing to psychologists and future refined it to his own ends. At the same time, his zeal in promoting, and passing along to future generations of researchers not only his substantive findings but also the methods that he thought were appropriate and necessary for psychological research, forged important links to the future. He was a prime mover in establishing the Society of Multivariate Experimental Psychology (SMEP) and the founding of the society's journal, *Multivariate Behavioral Research* (MBR). He helped to launch MBR in 1966 with a provocative lead article entitled 'Multivariate Behavioral Research and the Integrative Challenge,' in which he summarized his own version of the two general traditions of psychological research – the experimental and the individual differences traditions – and pointed out the directions for future research that would elevate the study of behavior to new levels of competence. In that same year the first edition of the *Handbook of Multivariate Experimental Psychology*, for which Cattell was both the editor and the major contributor, was published. A second edition followed 22 years later (Nesselroade and Cattell, 1988).

McArdle (1984) credited Cattell with having a profound effect on the development of what is now referred to generally as 'structural modeling' the fitting of quantitative models to covariance matrices (and occasionally the associated arithmetic means, as well), which has grown to enormous proportion in current day psychological research. This connection reinforces the amazingly productive service Cattell extracted from the factor analytic model. Cattell's empirical and theoretical work, as no one else's, has helped to clarify the distinction between latent and manifest variables and the key properties of both.

The contributions that Cattell made both substantively and methodologically were many. Here, we have only touched on some ofthe major ones. He developed and presented one of the most systematic and encompassing theories of personality that psychology has ever seen. Along the way, he introduced a large number of methodological innovations, some of which have made their way into today's behavioral science argot with an ease that today gives no hint of Cattell's pioneering contribution. The 'scree test,' 'Procrustes' rotation, and 'P-technique' come readily to mind.

The theory of fluid and crystallized intelligence, on which Cattell collaborated closely with Horn (Horn and Cattell, 1966) remains very much in currency today as far as so-called psychometrically oriented models of cognition are concerned. The apparently different age gradients of these two kinds of abilities has helped to stimulate thinking not only about the complexity of the organism and the nature of its development across the lifespan (e.g., Baltesm, 1987), but also about the need to consider what level of aggregation of variables is optimal for a given purpose.

Despite Cattell's strong investment in the study of human abilities (Cattell, 1971), he did not neglect the temperament and motivation components of his research program. For example, the fruits of his efforts to measure temperament by questionnaire, especially, are evident today in the continuing usage of the High School Personality Questionnaire (HSPQ) and the 16 Personality Factor questionnaire (16PF). Some of the work of which Cattell was most proud and for which he had the highest hopes sprang from his contributions to the study of motivation. His concept of the dynamic calculus, which was an attempt to write a systematic prediction equation for behavior that included trait, state, and dynamic aspects was refined into what he came to call structured learning theory. In the author's view, this was a key contribution in Cattell's own evaluation of his scientific legacy.

Many of the individuals who received their scientific training in Cattell's laboratory are still active and productive today. They, in turn, are passing on to their students the concepts of the data box, fluid and crystallized intelligence, covariation designs, factorial invariance, states versus traits, etc.

Despite the comprehensive perspective, the numerous insights and innovations, and the sheer bulk of his scholarly output, which spanned parts of eight different decades, Raymond B. Cattell's place in the history of psychology will not be identified without controversy for the foreseeable future. A lifetime achievement award for his contributions to psychology to be given by the American Psychological Foundation was announced in early 1997. Although Cattell had never taken kindly to the science 'establishment,' he had reached a point in his journey where such a high level of recognition for his scientific work meant something to him. But it was not to be. Some of his more philosophical, provocative writings having to do with the roles in our society of genetics, evolution, and public policy had generated bitter and sufficient resentment that, after some 'negotiating,' the award was withdrawn prior to the 1997 annual meeting of the American Psychological Association in Chicago, where it was to have been given. Cattell died the following 2 February at the age of 92.

See also: Darwin, Charles Robert (1809–82); Factor Analysis and Latent Structure Analysis: Overview; Galton, Sir Francis (1822–1911); Lifespan Development, Theory of; Neo-Piagetian Theories of Cognitive Development; Parsons, Talcott (1902–79); Personality Theory and Psychopathology; Quetelet, Adolphe (1796–1874); Thorndike, Edward Lee (1874–1949).

Bibliography

Baltes, P.B., 1987. Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. Developmental Psychology 23 (5), 611–626.

Cattell, R.B., 1944. 'Parallel proportional profiles' and other principles for determining the choice of factors by rotation. Psychometrika 9, 267–283.

Cattell, R.B., 1952. Factor Analysis. Harper, New York.

Cattell, R.B., 1954. Culture Fair Intelligence Tests, Scales 1, 2, and 3, Forms A and B, rev. ed. IPAT, Champaign, IL.

Cattell, R.B., 1957. Personality and Motivation Structure and Measurement. World Book Co., New York.

Cattell, R.B., 1964. Validity and reliability: A proposed more basic set of concepts. Journal of Educational Psychology 55, 1–22.

Cattell, R.B., 1971. Abilities: Their Structure, Growth, and Action. Houghton Mifflin, Boston.

Cattell, R.B., 1972. Real base, true zero factor analysis. Multi-variate Behavioral Research Monographs 72 (1), 1–162.

Cattell, R.B., 1984. The voyage of a laboratory, 1928–1984. Multivariate Behavioral Research 19, 121–174.

Child, D., 1998. Raymond Bernard Cattell (1905–1998). British Journal of Mathematical and Statistical Psychology 51, 353–357.

Goldberg, L.R., 1968. Objective Personality and Motivation Tests – theoretical introduction and practical compendium – Cattel R B and Warburton F W. Contemporary Psychology 13, 617–619.

Hall, C.S., Lindzey, G., 1978. Theories of Personality, third ed. Wiley, New York.

Horn, J.L., Cattell, R.B., 1966. Refinement and test of the theory of fluid and crystallized general intelligences. Journal of Educational Psychology 57, 253–270.

McArdle, J.J., 1984. On the madness in his method: R. B. Cattell's contributions to structural equation modeling. Multivariate Behavioral Research 19, 246–267.

McArdle, J.J., Cattell, R.B., 1994. Structural equation models of factorial invariance in parallel proportional profiles and oblique confactor problems. Multivariate Behavioral Research 29 (1), 63–113.

Nesselroade, J.R., Cattell, R.B. (Eds.), 1988. Handbook of Multivariate Experimental Psychology, second ed. Plenum, New York.

Causal Analysis, History of

Tim F Liao, University of Illinois at Urbana–Champaign, Urbana, IL, USA
Inna F Deviatko, Higher School of Economics – National Research University, Moscow, Russia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the history of causal research traditions in the social sciences. We identify two major bases for the methods and logic of causal analysis in the social sciences – experimental designs and statistical methods – and discuss the developments in these two correlated research traditions, especially the implications of these developments for the social sciences. While the focus of our discussion is on the developments in western societies, we also briefly review prominent features of causal analysis in the social sciences in non-western societies.

History of Causal Analysis in the Social Sciences

The emergence of causal analysis in the social sciences coincided with the emergence of social science disciplines in the West in the late nineteenth and early twentieth centuries. These disciplines established themselves as university departments and sought to distinguish themselves from natural sciences and humanities that had enjoyed a much longer history in academic research. The distinguishing feature of these social sciences, in particular sociology, has been to understand the causal role of society. Put differently, social scientists want to understand the effect of environment on human behavior. Following centuries-old nature-nurture debate, social scientists intend to distinguish the role of environment from heredity. Since the emergence of social sciences, causal analysis has matured to a recent stage where researchers moved beyond the mere consideration of observables to the analysis of counterfactuals (see Morgan and Winship, 2007).

For the American sociologist Albion Small (1898; 11 May 1854–24 March 1926), the first two of the three phases of social science research, embodied in what he called the 'essential methodological processes,' are elementary analysis, which is descriptive of individual parts or variables of a phenomenon, and causal analysis, which is what we know today as analysis of causal mechanisms. In quantitative social sciences, we still practice these two stages of research today in that order. We further identify two major sources for the methods and logic of causal analysis in the social sciences – experimental designs and statistical methods. The evolutions of these foci have been correlated processes. Note that because of space constraints, we cannot go into the details of other causal research traditions, notably John Stuart Mill's (1883; 20 May 1806–8 May 1873) configurational type of analysis of causality and its contemporary refinements and advances, especially in Charles Ragin's (1987) and his later works in the twenty-first century. In the following sections, we briefly review the two influential sources of causal analysis in the social sciences, namely experiments and statistical analysis.

Randomization, Experimentation, and Social Science Research

Two books on social science research methods that have been considered as modern classics and that have been the standard texts in many graduate-level social science research methods course are Donald Campbell and Julian Stanley's (1963) *Experimental and Quasi-Experimental Designs for Research* and Thomas Cook and Donald Campbell's (1979) *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. A chief difficulty for causal analysis is spuriousness or the effect of confounding variables. Luminaries in both Europe and America in the likes of Small, Emile Durkheim, and Karl Pearson already knew long ago what every social science graduate student is taught today that the specter of spuriousness could only be truly eliminated either by randomization of observations or by a method that somehow resembles or simulates the randomization process. Campbell and Stanley (1963) discuss various types of true and quasi-experimental designs, some of which originated in the early twentieth century, distinguish between internal and external validity of these designs, and warn researchers against various threats to internal validity including history, maturation, testing, instrumentation, statistical regression, selection of subjects, experimental mortality, selection–maturation interaction, and John Henry effect, and to external validity including reactive or interaction effect of testing, interaction effects of selection biases and the experimental variable, reaction effects of experimental arrangements, and multiple treatment interference.

What separates true experimental designs from quasi-experimental designs is the application of randomization, the process that randomly assigns research subjects to treatment versus control groups to allow the understanding as well as the analysis of true causal effects. However, as is well known, randomization in the social sciences is rarely practical or even possible. Samuel Stouffer (1950; 6 June 1900–24 August 1960) emphasized quasi-experiments – experiments that do not apply random assignments for comparison to make inferences about treatment-induced changes but otherwise resemble true experiments. Cook and Campbell's (1979) book reviews some preexperimental and experimental designs but nevertheless focuses on quasi-experimental designs that control many of the validity threats, often drawing on examples from psychology and educational research.

There are other variations of true experiments. One type that was developed in the twentieth century but has received more recent attention is field experiments, which recruit subjects in the field instead of in classrooms or into laboratory settings

although still allowing randomization. Arguing the 'controlled' environment of a laboratory can be artificial, Glenn Harrison and John List (2004) review six criteria that can serve to determine the field context of an experiment: the nature of the subject pool, the nature of information brought to the task by the subjects, the nature of the community, the nature of the task, the nature of stakes, and the environment the subjects operate in.

A special type of quasi-experiments has received a lot of academic and applied attention recently is natural experiment even though the most famous early example is the 1854 Broad Street cholera outbreak in London. Obviously, natural experiments do not have random assignments of subjects. However, they are useful in that there is often a clearly defined and large change in the treatment or exposure occurring to a clearly defined population or subpopulation, with no change in the treatment or exposure for another population or subpopulation. Many good examples of natural experiments come from economic or socioeconomic research. This is an increasingly popular method in the social sciences (for a complete review, see Dunning, 2008).

Statistical Analysis of Causality and Social Science Research

Early statisticians were deeply interested in the analysis of experimental data. Noted statisticians who made fundamental contributions, directly or indirectly, to statistical causal analysis include Ronald Fisher, Jerzy Neyman, Karl Pearson, and George Udny Yule, among others. Fisher's contribution to the design and statistical analysis of experiments is particularly fundamental, proposing the principles of design of experiments and analysis of variance of experimental data. His research culminated in two books (Fisher, 1925, 1935; 17 February 1890–29 July 1962), both of which were standard references for decades to come. Separately, Neyman (1923; 16 April 1894–5 August 1981) was concerned with analyzing experiments with statistical and probability theory as well. Together with Egon Pearson, Neyman also contributed to hypothesis testing in statistics, resulting in what we know as Neyman–Pearson hypothesis testing today and the Neyman–Pearson lemma paved the foundation for hypothesis testing practices up to this day.

The developments have been swift in the past few decades for statistical causal analytic methods in the social sciences. Major developments are represented in the research on structural equation modeling, matching analysis, and more generally, counterfactual causal modeling. Drawing on Sewall Wright's (1920) early work on path analysis in biology, simultaneous equation models in economics (Haavelmo, 1943), and factor analysis in psychology (Spearman, 1904), Karl Jöreskog (1973), James Keesling (1972), and David Wiley (1973) developed a comprehensive analysis system for causal analysis structural equation models with latent variables. The analytic framework became popular with the advances of computer programs, initially LISREL, and later on EQS, Amos, and MPlus. Today, virtually all major statistical software packages contain a module for structural equation modeling, in particular R, SAS, and Stata. The structural equation modeling framework allows causal analysis among not just observed variables but also unobserved, latent variables. The framework may also have some effects on uniting social scientists from different disciplines with a common analytic interest. This analytic approach allows not just a formal way of understanding causal chains but reciprocal causation as well. More importantly, it recognizes that our measurement is never perfect and that it contains measurement error, which can be properly modeled in structural equation modeling. Kenneth Bollen's (1989) book, together with other treatments, has helped make the method more accessible for social science researchers. Also see Judea Pearl's (2009, 2010) treatment of causality in this and related approaches.

A serious challenge to making causal inference is the presence of potential confounding factors when random assignment is absent. Observational studies are situations that lack randomization. Matching, or matched sampling, is a method for reducing bias and for more properly estimating treatment effects (Rubin, 1974; Rosenbaum and Rubin, 1983). Propensity score matching is a prominent matching method that uses observed variables to predict the probability of experiencing the event to create a 'counterfactual' group for making a more proper causal inference. The propensity score matching method has become very popular in the social sciences in recent years, and is implemented in major statistical software packages including R, SAS, Stata, and SPSS. For a notable recent assessment of the method applied in medical research, see Peter Austin (2009).

The essence of propensity score matching is to create a sample of cases that did not experience treatment but that are statistically equivalent, in all other aspects, to members of the treatment group. The procedure obtains a factual–counterfactual matched sample with which treatment effects can be estimated. One reason for the popularity enjoyed by propensity score matching today is because matching on the score is as good as matching on the configuration of all the observed variables involved in the matching (Rosenbaum and Rubin, 1983). Fundamental to this method is the assumption that the nonrandom assignment process is exogenous to the outcome distribution process, an assumption that Donald Rubin (1973) uses. Put differently, the assumption can be viewed as whether assignment to treatment is a self-selection process related to the expected outcome. James Heckman (2005) questions this assumption and proposes a control function approach that explicitly models the relation between the unobserved variable(s) in the outcome and selection equations by modeling the correlated errors in the two equations. Heckman's treatment relaxes the assumption and can be useful when the assumption is violated, thus enabling the researcher to estimate treatment effects when self-selectivity exists. For a comprehensive discussion of counterfactuals and causal modeling, see Stephen Morgan and Christopher Winship (2007). Morgan and Winship have a second edition of the book in press.

Causal Research in Non-western Societies

From the discussion above it is obvious that causal analysis in the social sciences has been entirely developed in the West. The development of the social sciences in non-western societies typically followed the western model. In East Asia, social sciences first developed in Japan, with their roots formed in the Meiji period when the country was open to new ideas from the outside. For example, sociology was introduced to Japan by

Ernest Francisco Fenollosa who gave lectures on the topic at Tokyo Imperial University in 1878 (Shoji, 1996). Accordingly, theoretical and methodological concerns in the social sciences there followed the western model. In China, the development of social sciences was also heavily influenced by the West even though drawing conclusions from social observations can be traced to early thinkers such as Confucius and Lao Tzu. Serious concerns with causal analysis primarily follow the western tradition (see Peng, 2011).

Russian social sciences of the late nineteenth and early twentieth century in prerevolutionary Russia developed closely with the West European tradition of social–economic surveys (*zemskaya statistika*) conducted by elective district councils (*zemstvo*). Further development of this tradition was halted after early 1930s during Stalin's rule, when sociology was claimed to be ''bourgeois pseudo-science.' The revival of sociology in the Soviet Union began during Khrushchev's era from the late 1950s to the early 1960s. The renewal of sociological research led to noticeable advances in methodology of social sciences with influences from both the western tradition and relevant developments in East European social sciences. In particular, the Polish sociologist Stefan Nowak's (1960, 1977) idea of statistical interpretation of types of causal relations in social sciences is worth mentioning. He introduced a typology of causal relations for dichotomic attributes (e.g., general vs conditional, positive vs negative). The typology enabled one to derive statistical relationships corresponding to each type of causal relations for elementary cases.

A version of information-based approach to causal analysis was developed in 1970s by Igor N. Taganov (1977). This is an early attempt to use Markov-chain-based causal reasoning in Soviet social sciences although it has never been widely applied. Soviet/Russian logician Victor K. Finn (1983) developed a formal JSM method of plausible reasoning (named after J.S. Mill) based on multivalent logic and intended for automatic generation and testing of hypotheses. With its foundation in mathematical logic, the method synthesizes the procedures of induction, constructive abduction (in the Piercian sense), and analogical reasoning and can be applied to qualitative social science data for identifying underlying causal structures (Finn, 1983). In his 1997 book coauthored with K.V. Khvostova, *Problems of Historical Knowledge in the Light of Contemporary Multidisciplinary Research* and other publications, Finn describes the procedures of causal hypotheses generation from data consisting of qualitative empirical information on objects and their properties in order to produce a knowledge base comprising explanatory and causal propositions inferred from propositions about the properties of objects. A software implementation for causal analysis (based mostly on 'direct method of agreement' by J.S. Mill) has been successfully applied in different disciplines including sociology, paleography, toxicology, and others. When compared with comparative methods such as Charles Ragin's QCA (Ragin, 1987), this JSM method has the benefit of implementing a rule of 'ban on counterexamples' (Kuchenkova and Tatarova, 2013).

Readers interested in an up-to-date treatment of methods for causal research in the social sciences are referred to the handbook edited by Stephen Morgan (2013), which also includes a chapter on the history of causal analysis in the social sciences.

See also: Experimental Design: Overview; Experimental Design: Randomization and Social Experiments; Experimental Economics; Experimental Methods in Survey Research in Demography; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Organizations and Culture; Organizations and Culture.

Bibliography

Austin, Peter C., 2009. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. Biomedical Journal 51, 171–184.

Bollen, Kenneth A., 1989. Structural Equations with Latent Variables. Wiley, New York.

Campbell, Donald T., Stanley, Julian, 1963. Experimental and Quasi-experimental Designs for Research. Cengage Learning, Stamford, CT.

Cook, Thomas D., Campbell, Donald T., 1979. Quasi-experimentation: Design and Analysis Issues for Field Settings. Houghton Mifflin, Boston.

Dunning, Thad, 2008. Natural Experiments in the Social Sciences: A Design-Based Ahpproac. Cambridge University Press, Cambridge.

Finn, Victor K., 1983. О машинно-ориентированной формализации правдоподобных рассуждений в стиле Ф. Бэкона-Д.С. Милля (On computer-oriented formalization of plausible reasoning in F. Bacon – J.S. Mill's style). *Семиотика и информатика* (Semiotics and Information Science) 20, 35–101.

Fisher, Ronald A., 1925. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.

Fisher, Ronald A., 1935. The Design of Experiments. McMillan, London.

Haavelmo, Trygve, 1943. The statistical implications of a system of simultaneous equations. Econometrica 11, 1–12.

Harrison, Glenn W., List, John A., 2004. Field experiments. Journal of Economic Literature 42, 1009–1055.

Heckman, James J., 2005. The scientific model of causality. In: Stolzenberg, Ross M. (Ed.), Sociological Methodology, vol. 35. Wiley-Blackwell, Hoboken, NJ, pp. 1–97.

Jöreskog, Karl, 1973. A general method for estimating a linear structural equation system. In: Goldberger, Arthur S., Duncan, Otis Dudley (Eds.), Structural Equation Models in the Social Sciences. Academic, New York, pp. 85–112.

Keesling, James Ward, 1972. Maximum Likelihood Approaches to Causal Flow Analysis. Department of Education, University of Chicago, Chicago.

Khvostova, Ksenia V., Finn, Victor K., 1997. *Проблемы исторического познания в Свете Совретенных Междисциплинарных исследований* (Problems of Historical Knowledge in the Light of Contemporary Multidisciplinary Research). Russian State University of Humanities Press, Moscow.

Kuchenkova, Anna V., Tatarova, Galina G., 2013. Стратегия применения логико-комбинаторных методов в процедурах типологического анализа (A strategy for using logical-combinatorial methods in procedures of typological analysis). *Социология: Методология, методы, математическое моделирование* (Sociology: Methodology, Methods, Mathematical Modeling) 36, 7–35.

Mill, John Stuart, 1882. A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation. Harper & Brothers, New York.

Morgan, Stephen L., Winship, Christopher, 2007. Counterfactuals and Causal Inference: Methods and Principles for Social Research. Cambridge University Press, Cambridge.

Morgan, Stephen L. (Ed.), 2013. Handbook of Causal Analysis for Social Research. Springer, New York.

Neyman, Jerzy, 1923. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Roczniki Nauk Rolniczych X, 465–472 (Annals of Agricultural Sciences 1–51), translated in Statistical Science 5 (1990).

Nowak, Stefan, January, 1960. Some problems of causal interpretation of statistical relationships. Philosophy of Science 27 (1), 23–38.

Nowak, Stefan, 1977. Причинные интерпретации статистических связей в социальном исследовании (Causal interpretations of statistical relationships). In: Аганбегян, А., Блейлок, Х., Бородкин, Ф., Будон, Р., Капекки, В. (Eds.), *Математика в социологии: Моделирование и Обработка инфортации* (Mathematics in Sociology: Modeling and Data Analysis). Мир, Москва, pp. 76–123 (English Edition: Blalock H., Aganbegian, A., Borodkin, F., Capecchi V. (Eds), 1975. Quantitative Sociology. Academic Press Inc., New York).

Pearl, Judea, 2009. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge.

Pearl, Judea, 2010. The foundations of causal inference. In: Futing Liao, Tim (Ed.), Sociological Methodology, vol. 40. Wiley-Blackwell, Hoboken, NJ, pp. 75–149.

Peng, Yusheng, 2011. 社会科学中的因果分析 (Causal analysis in the social sciences). 社会学研究 (Sociological Studies) 3, 1–31.

Rosenbaum, Paul R., Rubin, Donald B., 1983. The Central role of the propensity score in observational studies for causal effects. Biometrica 70, 41–55.

Ragin, Charles C., 1987. The Comparative Method: Moving beyond Qualitative and Quantitative Strategies. Univ. of California Press, Berkeley.

Rubin, Donald B., 1973. Matching to remove bias in observational studies. Biometrics 29, 159–183.

Rubin, Donald B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688–701.

Shoji, Kokichi, 1996. Institutionalization of sociology in Japan. In: Lee, Su-Hoon (Ed.), Sociological in East Asia and Its Struggle for Creativity, pp. 33–44. Proceedings of the ISA Regional Conference for Eastern Asia, Seoul, Korea, November 22–23, 1996, the International Sociological Association.

Small, Albion W., 1898. Seminar notes: the methodology of the social problem. Division I. The sources and uses of material. American Journal of Sociology 4, 113–144.

Spearman, Charles, 1904. 'General Intelligence', objectively determined and measured. American Journal of Psychology 15 (2), 201–292.

Stouffer, Samuel A., 1950. Some observations on study design. American Journal of Sociology 55, 355–361.

Taganov, Igor N., 1977. Информационные меры причинного влияния: IH-анализ (Information measures of causal influence: IH-analysis). In: Аганбегян, А., Блейлок, Х., Бородкин, Ф., Будон, Р., Капекки, В. (Eds.), *Математика в социологии: Моделирование и Обработка инфортации* (Mathematics in Sociology: Modeling and Data Analysis). Мир, Москва, pp. 124–134 (English Edition: Blalock H., Aganbegian, A., Borodkin, F., Capecchi V. (Eds), 1975. Quantitative Sociology. Academic Press Inc., New York).

Wiley, David E., 1973. The identification problem for structural equation models with unmeasured variables. In: Goldberger, Arthur S., Duncan, Otis Dudley (Eds.), Structural Equation Models in the Social Sciences. Academic, New York, pp. 69–84.

Wright, Sewall, 1920. The relative importance of heredity and environment in determining the piebald pattern of guinea pigs. Proceedings of the National Academy of Sciences 6, 320–332.

Causal Counterfactuals in Social Science Research

Paul W Holland, University of California, Berkeley, CA, USA; and Educational Testing Service, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp. 1550–1554, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article explains the connection between counterfactual reasoning and the Fundamental Problem of Causal Inference. The Neyman–Rubin causal model is introduced, as the potential outcome framework on which it is based makes explicit the counterfactual nature of causal inference. The definition of average causal effects and prima facie average causal effects reveals the sources of bias for causal estimates in observational studies Discussion is also given to 'empty' counterfactuals that arise when unalterable attributes of units are treated as if they were treatments in experiments.

The term 'counterfactual conditional' is used in logical analyses to refer to any expression of the general form: "If A were the case then B would be the case." In this usage, A is usually false or untrue in the world so that A is 'contrary to fact' or counterfactual. Examples abound. "If kangaroos had no tails, they would topple over" (Lewis, 1973b). "If an hour ago I had taken two aspirins instead of just a glass of water, my headache would now be gone" (Rubin, 1978). Perhaps the most obnoxious counterfactuals in any language are those of the form: "If I were you, I would …."

Lewis (1973a) observed the connection between counterfactual conditionals and references to causation. He finds these logical constructions in the language used by Hume in his famous discussion of causation. Hume defined causation twice over. He wrote, "we may define a cause to be *an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second*. Or, in other words, *where, if the first object had not been, the second never had existed*" (Lewis, 1973a: italics are Lewis's).

Lewis draws attention to the comparison between the factual first definition where one object *is* followed by another and the counterfactual second definition where, counterfactually, it is supposed that if the first object 'had not been' the second object would not have been either.

It is the connection between counterfactuals and causation that makes them relevant to social science research. From the point of view of some authors, it is difficult, if not impossible, to make any sense of causal statements without using counterfactual language (Lewis, 1973a; Holland, 1986; Rubin, 1978; Robins, 1985, 1986). Other authors are concerned that using such language gives an emphasis to unobservable entities that is inappropriate in the analysis of empirical data (Dawid, 1997; Shafer, 1996). The discussion here accepts counterfactuals in discussions of causation and will explain their role in the estimation of causal effects based on the work of Neyman (1923, 1935), Rubin (1974, 1978), and others.

We begin with a simple observation. Suppose that we find that a student's test performance changes from a score of X to a score of Y after some educational intervention. We might then be tempted to attribute the pretest–posttest change, $Y - X$ to the intervening educational experience, that is, to use the gain score as a measure of the improvement due to the intervention. However, this is social science and not the tightly controlled 'before–after' measurements made in a physics laboratory. There are many other possible explanations of the gain, $Y - X$. Some of the more obvious are simple maturation, other educational experiences occurring during the relevant time period, and differences in either the tests or the testing conditions at pre- and posttests. Cook and Campbell (1979) provide a classic list of 'threats to internal validity' that address many of the types of alternative explanations for apparent causal effects of interventions. For this reason, it is important to think about the real meaning of the attribution of cause. In this regard, Lewis's discussion of Hume serves us well. From it we see that what is important is what the value of Y *would have been* had the student *not had* the educational experiences that the intervention entailed. Call this score value, Y^*. Thus enter counterfactuals. Y^* is not directly observed for the student, that is, they *did have* the educational intervention of interest, so asking for what their posttest score *would have been* had they *not had it* is asking for information collected under conditions that are contrary to fact. Hence, it is not the difference $Y - X$ that is of interest, but the difference $Y - Y^*$, and the gain score has causal significance relative to the effect of the educational experience only if X can serve as a substitute for the counterfactual Y^*. In physical science laboratory experiments such a substitution is often easy to make, but it is rarely believable in many social science applications of any consequence.

A formal model or language for discussing the problem of estimating causal effects (of the form $Y - Y^*$ rather than $Y - X$) was developed by Neyman (for randomized experiments) and Rubin (for a wide variety of causal studies) and will be called the Neyman–Rubin model for causal effects here.

Prospective Causal Studies

The Neyman–Rubin model is most easily understood in the context of a *prospective causal study* that has the general structure specified by this sequence of events:

1. Subjects or experimental units of study are identified.
2. Baseline or pretest information about these units is recorded.

3. The units are either assigned to (in a controlled study) or select themselves to (in studies without the control of assignment) exposure to one of the treatment conditions or interventions of the study.
4. These units are then subsequently exposed to their assigned or self-selected treatment condition (and each unit is affected by this exposure in a manner that is unrelated to the exposure conditions of the other units).
5. At an appropriate later time an outcome, end point, or posttest measure is recorded for each unit in the study.

The type of study that this five-part schema is intended to cover includes most randomized comparative experiments as well as many types of pretest–posttest quasi-experiments or observational studies. It should be emphasized that, properly interpreted, the Neyman–Rubin model has application to other types of causal studies (see Holland, 1988; Holland and Rubin, 1988; Robins, 1997), but, for our purposes here, prospective causal studies are already sufficiently complicated and inclusive.

The condition mentioned parenthetically in (4), that the exposure conditions of the other units do not affect the outcomes associated with a given unit, is very important, and is clearly an assumption that would not be true in general (Hong and Raudenbush, 2013). For example, in the study of infectious diseases your vaccination will affect my likelihood of contracting polio. Rubin explicitly identifies this assumption calling it the stable unit treatment value assumption, or SUTVA. SUTVA will be assumed throughout this article.

The Neyman–Rubin Model

The version of the Neyman–Rubin model used here is adapted from that of Holland (1986), and differs from the original versions by Neyman (1935) and Rubin (1978) mainly in its emphasis on population quantities. One of the main benefits of this model is that it identifies certain 'counterfactual conditional expectations' as the location of key assumptions about the inferential structure of any causal study and its resulting data.

The prospective causal study begins with the 'units,' 'subjects,' or 'cases' of the study, and the ith unit is denoted by the subscript 'i.' It may help the reader to imagine that this is a discussion about a very large sample of units. Denote the population of units under study by P. For the most part P will lie quietly in the background without being noticed. The baseline information that is collected or recorded for unit i will be denoted as a *vector* of numerical information, z_i.

There is a 'causal' variable denoting a set of possible 'treatments' or 'exposure' conditions to which each unit in the study could be exposed. For simplicity, we assume that there are only two treatment conditions denoted by $x = 1$ (treatment) and $x = 0$ (control). A more complicated version of this would let x be a number representing the 'strength' of the treatment level, but we will just use the dichotomous case here.

An important aspect of causal variables designating such treatments levels or intervention conditions is the assumption that the level of exposure for any unit *could have been different* from what it actually was. This condition excludes 'attributes' of units (such as race, gender, age, or pretest score) as causal variables in the sense that such attributes *cannot* have 'unit-level causal effects' in the sense that we will define below. This idea is discussed more extensively in Holland (1986, 1988), and is mentioned again in another context at the end of the present article. Each unit is exposed to one treatment level and the value of x to which i is exposed is denoted by x_i.

Finally, we come to the outcomes or dependent variables in the study, and here is where a special notation is needed. We let $Y_i(x)$ denote the (numerical) response that would be recorded for unit i if unit i were exposed to treatment level x. For each i, $Y_i(x)$ is a function of x. It should be emphasized that $Y_i(x)$ is not directly observed unless $x_i = x$. This is an important point because it is crucial to realize that the $\{Y_i(x)\}$ do not denote *observed data* like z_i and x_i do, but rather the $\{Y_i(x)\}$ are 'potential outcomes' that lie behind the observed values of the outcome variable (Rubin, 2005). For this reason, we denote the potential observations by capital letters to distinguish them from quantities that are directly observable, which we denote by lowercase letters.

The connection between the potential outcomes and the actually observed outcomes is then given by the equation

$$y_i = Y_i(x_i) \qquad [1]$$

where y_i is the observed outcome or value of the dependent variable for unit i. The idea behind eqn [1] is that to get from the potential outcomes $\{Y_i(x)\}$ to an observed outcome, we must select the value of x in $Y_i(x)$ to be the value to which i is actually exposed, that is, $x = x_i$, and then we obtain y_i from the potential observations, $\{Y_i(x)\}$, via eqn [1].

The *observed data* for each unit i are the vector (z_i, x_i, y_i). The potential outcomes, $\{Y_i(x)\}$, are never observed for *all* values of x for a fixed unit, i, but only for the specific x-value to which i is actually exposed, x_i. It is sometimes said that the $\{Y_i(x)\}$ are 'counterfactual' because they are not actual observations. Here they are called *potential observations* because they could have been observed had x_i been different than it was.

It is helpful to use a notation, such as $E(y)$, to mean the average value of y_i across the (large number of) units in P. Furthermore, an expression such as

$$E(y|x = a) \qquad [2]$$

will mean the average value of y_i across all of the (large number of) units in P for which $x_i = a$. The use of the expectation notation is to make certain quantities clearer in their meaning, and may be justified from either a frequentist or Bayesian point of view. It should be noted that within the expectation notation, the subscript i, denoting the unit, is suppressed because, within the scope of the $E(\ )$ operator, i is averaged over.

Using the Neyman–Rubin Model

An important fact about the Neyman–Rubin model is the Fundamental Problem of Causal Inference (Holland, 1986): It is impossible in principle to *observe* $Y_i(x)$ for more than one value of x for any one unit, i. Any procedure that claims to have avoided the Fundamental Problem of Causal Inference can always be shown to be based on untestable assumptions.

Sometimes such assumptions are plausible, and sometimes they are not.

A basic definition that we are now in a position to make is that of a 'unit-level causal effect.' Because we are restricting attention to the simple case of two treatment levels, $x = 0/1$ we may restrict attention to these differences

$$Y_i(1) - Y_i(0) = \text{the casual effect of } x = 1 \text{ relative to } x = 0 \text{ for unit } i \quad [3]$$

In the Neyman–Rubin notation, the unit-level causal effects are the *basic* quantities of interest in causal inference. However, the Fundamental Problem of Causal Inference is now immediately seen to *be fundamental* because it implies that unit-level causal effects are, themselves, *never directly observable*. Thus, we are always reduced to making assumptions that allow us to make some sort of conclusion or inference about these causal effects. This is the place where the Neyman–Rubin model might appear to be impractical for applied research, that is, because its most basic parameters, the unit-level causal effects, are not directly observable. Furthermore, this is exactly the place where the potential exposability of all of the levels of x to any unit is seen to be crucial to the foundations of the theory. The definition of causal effect requires this assumption so that the difference, $Y_i(1) - Y_i(0)$, is meaningful. It is the Fundamental Problem of Causal Inference and this definition of causal effect that make causal inference both more interesting and more difficult than the simple computation of correlational and associational measures. The unit-level causal effects mimic the comparison between a 'factual' and a 'counterfactual' identified in the quote mentioned earlier from Hume by Lewis.

It is now time to introduce the notion of an average causal effect (ACE). In general, an ACE is any average of unit-level causal effects. The most general ACE has the form

$$\text{ACE} = E[Y(1) - Y(0)|A] \quad [4]$$

where A denotes some collection of units defined in terms of either z_i or x_i or both. In eqn [4] we again suppress the subscript i because it is being averaged over. As an example of an A in eqn [4] we might use $A =$ 'all the units in the study,' in which case the ACE is the average causal effect over all of P. But other cases might be of interest, for example, $A =$ 'all units where i is male and for whom $x_i = 1$.' In this case the ACE is for the males in treatment group 1. Here we restrict our attention to the ACE that is called the 'effect of the treatment on the treated' in which case A denotes all of the units for which $x_i = 1$, that is

$$\begin{aligned}\text{ACE} &= E[Y(1) - Y(0)|x = 1] \\ &= E[Y(1)|x = 1] - E[Y(0)|x = 1]\end{aligned} \quad [5]$$

Up to now, we have simply defined the basic structure of the data collection design as well as the causal connection between the potential outcomes and the causal variable x, that is, eqn [3]. We have not yet identified the connection between any quantities that could be estimated with data and the causal parameters given by either the unit-level causal effects in eqn [3] or the ACEs in eqn [5]. This leads us to the 'prima facie average causal effects,' (FACEs). The FACEs are what can be estimated from the data. To parallel the ACE in eqn [5] we examine the FACE, which is simply the difference between the mean of the outcome variable observed in each treatment group, that is

$$\text{FACE} = E[y|x = 1] - E[y|x = 0] \quad [6]$$

If we substitute the definition of y in terms of the potential observations, $\{Y_i(x)\}$, that is given in eqn [1] into eqn [6] we obtain

$$\text{FACE} = E[Y(1)|x = 1] - E[Y(0)|x = 0] \quad [7]$$

Finally, if we combine eqn [5] and eqn [7] we obtain the following basic formula that relates the ACE to the FACE

$$\text{FACE} = \text{ACE} + \text{BIAS} \quad [8]$$

where

$$\text{BIAS} = E[Y(0)|x = 1] - E[Y(0)|x = 0] \quad [9]$$

The BIAS term contains two parts: one *factual*, that is, $E[Y(0)|x = 0] = E[y|x = 0]$, and the other *counterfactual*, that is, $E[Y(0)|x = 1]$. The factual part is just the mean of y_i for those units with $x_i = 0$. The counterfactual part is the mean of $Y_i(0)$ for those units for whom $x_i = 1$. $E[Y(0)|x = 1]$ is a quantity for which there can never be any data because the conditioning event makes the quantity being averaged over a counterfactual. Thus, it is a counterfactual conditional expectation. The value of such counterfactual parameters is that they pinpoint exactly where assumptions must be made that allows causal inference to take place using empirical data. When BIAS = 0, we have FACE = ACE and the empirical FACE equals the causal ACE.

An important condition that insures that BIAS = 0 is the construction of x_i by random assignment, which forces x_i and $Y_i(0)$ to be statistically independent of each other as functions of i over P (Holland, 1986). When this independence holds, $E[Y(0)|x = 1] = E[Y(0)|x = 0]$ and BIAS = 0.

Empty Counterfactuals

There is an unsatisfactory and rather misleading use of counterfactuals that sometimes arises in social science research. It occurs when the counterfactual condition, that is, the 'if A were the case' part could never occur in any real sense. Such empty counterfactuals arise when a nonmanipulable factor in a causal study is described as having an 'effect' on some outcome. Examples easily come about in casual causal talk. The effect of gender on salary suggests considering "what her salary would have been had she been a man." The effect of test performance on future employment suggests "the job he would have had had he scored higher on the test." The effect of English language proficiency on a math test in English suggests "the mathematics score a non-English speaker would have received had he or she been an English speaker." These empty counterfactuals arise when the value of a variable for a factor in a study could not have been other than the value that it was. The interesting and useful counterfactuals arise in those cases when the variable could have had a different value than it did for the individuals in a study, at least in principle. Judgment as to when a counterfactual is empty or not is not always easy and may require careful thought in many cases (King and Zeng,

2006). Consider the examples "if I were you, I would …," Lewis's kangaroo, and Rubin's aspirin in the opening paragraph. These represent very different kinds of counterfactuals from this perspective. The first is as empty as they get, the emptiness of the second depends on how the kangaroo might not have a tail (our imagination vs an axe), while Rubin's aspirin could be taken or not.

See also: Causal Inference: Overview; Experimental Design: Bayesian Designs; Experimental Design: Overview; Experimental Design: Randomization and Social Experiments; Factor Analysis and Latent Structure Analysis: Overview; Mechanism-Based Causal Analysis; Optimal Experimental Design; Probability and Chance: Philosophical Aspects; Quasi-Experimental Designs.

Bibliography

Cook, T.D., Campbell, D.T., 1979. Quasi-experimentation: Design and Analysis Issues for Field Settings. Houghton Mifflin, Boston.

Dawid, A.P., 1997. Causal Inference without Counterfactuals. Research Report No. 188. Department of Statistical Science, University College, London.

Holland, P.W., 1986. Statistics and causal inference. Journal of the American Statistical Association 81, 945–970.

Holland, P.W., 1988. Causal inference, path analysis and recursive structural equations models. In: Clogg, C. (Ed.), Sociological Methodology. American Sociological Association, Washington, DC, pp. 449–484.

Holland, P.W., 2008. Causation and race. In: Zuberi, T., Bonilla-Silva, E. (Eds.), White Logic, White Methods: Racism and Methodology. Rowman and Littlefield, Lanham, pp. 93–109.

Holland, P.W., Rubin, D.B., 1988. Causal inference in retrospective studies. Evaluation Review 12, 203–231.

Hong, G., Raudenbush, S.W., 2013. Heterogeneous agents, social interactions, and causal inference. In: Morgan, S.L. (Ed.), Handbook of Causal Analysis for Social Research. Springer, New York, pp. 331–352.

King, G., Zeng, L., 2006. The dangers of extreme counterfactuals. Political Analysis 14, 131–159.

Lewis, D.K., 1973a. Causation. Journal of Philosophy 70, 556–567.

Lewis, D.K., 1973b. Counterfactuals. Harvard University Press, Cambridge, MA.

Neyman, J., 1923. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. Roczniki Nauk Rolniczki 10, 1–51 (In: Polish, English Trans., Dabrowska, D., Speed, T., 1991. Statistical Science 5, 463–480).

Neyman, J., 1935. Statistical problems in agricultural experimentation. Supplement of the Journal of the Royal Statistical Society 2, 107–180.

Robins, J.M., 1985. A new theory of causality in observational survival studies – application of the healthy worker effect. Biometrics 41, 311.

Robins, J.M., 1986. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. Mathematical Modelling 7, 1393–1512.

Robins, J.M., 1997. Causal inference from complex longitudinal data. In: Berkane, M. (Ed.), Latent Variable Modeling with Applications to Causality. Springer-Verlag, New York, pp. 69–117.

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and non-randomized studies. Journal of Educational Psychology 66, 688–701.

Rubin, D.B., 1978. Bayesian inference for causal effects: the role of randomization. Annals of Statistics 6, 34–58.

Rubin, D.B., 2005. Causal inference using potential outcomes. Journal of the American Statistical Association 100, 322–331.

Shafer, G., 1996. The Art of Causal Conjecture. MIT Press, Cambridge, MA.

Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and Quasi-experimental Designs for Generalized Causal Inference. Wadsworth Cengage Learning.

Causal Inference: Overview

Jennifer Hill, Steinhardt School of Culture, Education, and Human Development, New York University, New York, NY, USA
Elizabeth A Stuart, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses causal inference in statistics. It describes the theoretical framework and notation needed to formally define causal effects and the assumptions required to identify them nonparametrically. This involves definition of potential outcomes that represent the potential value of the outcome across different treatment exposures. Designs that allow researchers to satisfy or weaken these assumptions are briefly described. Then common parametric assumptions used to model effects and more current approaches that require weaker assumptions are discussed.

Introduction: Causal Inference as a Comparison of Potential Outcomes

Causal inference refers to an intellectual discipline that considers the assumptions, study designs, and estimation strategies that allow researchers to draw causal conclusions based on data. As detailed below, the term 'causal conclusion' used here refers to a conclusion regarding the effect of a causal variable (often referred to as the 'treatment' under a broad conception of the word) on some outcome(s) of interest. The dominant perspective on causal inference in statistics has philosophical underpinnings that rely on consideration of counterfactual states. In particular, it considers the outcomes that could manifest given exposure to each of a set of treatment conditions. Causal effects are defined as comparisons between these 'potential outcomes.' For instance, the causal effect of a drug on systolic blood pressure 1 month after the drug regime has begun (vs no exposure to the drug) would be defined as a comparison of systolic blood pressure that would be measured at this time given exposure to the drug with the systolic blood pressure that would be measured at the same point in time in the absence of exposure to the drug. The challenge for causal inference is that we are not generally able to observe both of these states: at the point in time when we are measuring the outcomes, each individual either has had drug exposure or has not.

Notation for Potential Outcomes and Causal Effects

There is a long history of philosophers, statisticians, and economists thinking about causality and how to establish causal relationships between variables. This philosophical discussion continues to this day, with many different perspectives still being expressed (see, e.g., Pearl, 2009; Dowd, 2011). Statisticians have generally focused on the investigation of the extent to which a given 'cause' leads to a change in some outcome(s) of interest. This is the framework detailed in this article. More general discussions of causal relationships, such as those investigated through structural equation models or path analysis, or those that attempt to determine a (or the) cause of an effect, typically require much stronger assumptions than do approaches that focus on estimating specific causal effects (Holland, 1986; VanderWeele, 2012).

To formalize assumptions and models in this framework, traditional statistical notation has been extended to explicitly define the 'potential outcomes' described above (Rubin, 1978). For instance, for an individual, i, and a binary treatment, Z_i, two outcomes can be defined that have the potential to be observed at a given time point, $Y_i(Z_i = 0) = Y_i(0)$ and $Y_i(Z_i = 1) = Y_i(1)$. In common language, these can be described as the outcome that would be observed if individual i had not received the treatment ($Z_i = 0$) and the outcome that would be observed if individual i had received the treatment ($Z_i = 1$), respectively. In the context of the example of the effect of a drug on blood pressure, $Y_i(0)$ would reflect the systolic blood pressure at a particular point in time for individual i observed in the absence of the drug and $Y_i(1)$ would reflect the systolic blood pressure observed *at that same point in time* for individual i after treatment with the drug.

One benefit of the potential outcome notation is that it allows for a simple and clear definition of the individual-level causal effect for observation i as a comparison between these two potential outcomes, for instance, $Y_i(1) - Y_i(0)$. (Nonlinear comparisons of potential outcomes, such as those used to define an odds ratio can also be considered; for simplicity this article focuses on the simple difference, in part because of complications that arise with noncollapsible functions of the potential outcomes; see, e.g., Greenland et al., 1999; Austin, 2007.) This formalization makes explicit the difficulty in inferring causality at the individual level given that we can never observe both $Y_i(1)$ and $Y_i(0)$ for the same individual, at any given time point; Holland (1986) refers to this as the 'fundamental problem of causal inference.' Using this potential outcomes framework to define causal effects also allows us to define those effects nonparametrically, and creates a useful distinction between structural assumptions and modeling assumptions (discussed below).

While it is nearly impossible to empirically identify individual-level causal effects without very strong assumptions or very specialized settings (for instance, in a laboratory where conditions can be precisely controlled), it is possible to make progress with respect to estimating average causal effects. The most common such estimand is the average of individual-level causal effects over a population or sample of interest, expressed generally as $E[Y(1) - Y(0)] = (1/N)\Sigma_i(Y_i(1) - Y_i(0))$. When this average is taken over the full population of interest, it is

referred to as the population average treatment effect. When this average is taken over the analysis sample it is referred to as the sample average treatments effect (SATE).

There is also sometimes interest in causal effects for particular subgroups. As a simple example, researchers may be interested in the SATE for males as distinct from the SATE for females. A slightly more complex example is sometimes referred to as the 'average treatment effect on the treated' (ATT). This is the average treatment effect for those in the treatment group, and is defined as $E[Y(1) - Y(0)|Z = 1] = (1/N_t)\Sigma_\iota(Y_i(1) - Y_i(0)|Z_i = 1)$. (An analogous quantity called the sample average treatment effect on the control (SATC) can also be defined.) Researchers are sometimes interested in the ATT in settings where there is no thought of giving the treatment to the full population, and rather interest is in the effect of the treatment on the people who actually received it. This might be the relevant quantity, for instance, in studies of the effects of adolescent drug use, where there is no thought of 'imposing' drug use on all adolescents. The sample average treatment effect on the treated (SATT) in such a study would compare the average outcomes for drug users with the potential outcomes for the same group had they not used drugs. This could inform the potential effect of programs that help to prevent or terminate drug use in teenagers. The SATE, on the other hand, would represent the difference between the average potential outcomes if all adolescents in the study were heavy drug users versus the potential outcomes that would manifest if none of the adolescents in the study were heavy drug users. This would not be the relevant scientific quantity given that some portion of that sample would likely never choose to use drugs so the effect for these people is rather meaningless. See Imai et al. (2008) and Harder et al. (2010) for more discussion of possible estimands.

To illustrate some of these concepts, consider the hypothetical data in Table 1. These data reflect 14 participants in a study that aimed to estimate the effect of a job training program on hourly wages. The table presents an omniscient view of the data that the researcher could never see in entirety because the Y(0) cells for those in the treatment group and the Y(1) cells for those in the control group are complete. The column for Y denotes the outcome that is observed for each person: $Y = \text{Treat}*Y(1) + (1 - \text{Treat})*Y(0)$. Seeing this complete data set with all of the potential outcomes observed is of course not possible in the real world, but, for illustration, it allows us to see the causal effects for each individual and, by extension, to calculate several different sample estimands. The treatment effects vary in this example depending on whether the participant had received a high school education (HS). The subgroups without high school education (HS = 0) are all defined to have treatment effects of 4. On the other hand, those with a high school education would experience no gain from the job training (for each, $Y(0) = Y(1)$). Since the distribution of those with high school education varies across the treatment groups, the average treatment effects vary as well. The SATT is equal to 2.67 and the SATC is equal to 1.

Table 1 Displays data from a hypothetical job training study. Data on high school graduation status (HS), age, and hourly wage (Y) could be observed by the researcher. The full set of potential outcomes could not be observed. They are displayed here to aid in conceptualization of individual-level causal effects as well as several sample average estimands

Person	*Treat (Z)*	*HS*	*Age*	*Y(0)*	*Y(1)*	*Y*
1	1	0	25	12	16	16
2	1	0	22	11	15	15
3	1	0	21	10	14	14
4	1	0	23	11	15	15
5	1	1	22	17	17	17
6	1	1	24	19	19	19
7	0	0	20	10	14	10
8	0	0	24	12	16	12
9	0	1	22	17	17	17
10	0	1	22	16	16	16
11	0	1	26	19	19	19
12	0	1	20	18	18	18
13	0	1	28	19	19	19
14	0	1	21	17	17	17

Structural Assumptions Required for Nonparametric Identification

If we could observe all of the potential outcomes for our sample, then causal inference would be easy. However, as discussed above, the fundamental problem of causal inference reflects the fact that at best we can observe only half of the potential outcomes, and thus need to rely on assumptions about those missing potential outcomes. In some sense this is what distinguishes causal inference from standard statistical inference, in that it (generally) requires untestable assumptions regarding the unobserved potential outcomes. This section discusses the structural assumptions needed to estimate causal effects, followed by a brief discussion of possible study designs and additional parametric assumptions.

Ignorability

Unconditional Ignorability

What assumption would allow us to identify average causal effects? To make the issues concrete, consider each piece of the SATE. Estimating $E[Y(1)]$ for the sample is challenging because we only observe Y(1) for those observations that were assigned to receive $Z = 1$. If those observations assigned to $Z = 0$ have a different mean of Y(1) than those who were assigned to $Z = 1$, then an estimate of the mean of the treated outcomes, $E[Y(1)|Z = 1]$, will be a biased estimate of $E[Y(1)]$. Similarly, estimating $E[Y(0)]$ is challenging because we only observe Y(0) for those observations that were assigned to receive $Z = 0$. If those observations assigned to $Z = 1$ have a different mean of Y(0) than those who were assigned to $Z = 0$, then an estimate of the mean of the control outcomes, $E[Y(0)|Z = 0]$, will be a biased estimate of $E[Y(0)]$. The most straightforward set of assumptions that ensures that a simple difference in means can provide an unbiased estimate of the average treatment effect relies on mean independence:

$$E[Y(0)] = E[Y(0)|Z = 0]$$

$$E[Y(1)] = E[Y(1)|Z = 1]$$

Designs and estimation procedures discussed in a later section highlight the distinction between assumptions and estimation. However, it is worth mentioning at this point that a completely randomized experiment would automatically satisfy this mean independence assumption. In fact, it will satisfy an even stronger assumption, that the distributions of the potential outcomes are independent of the binary outcome Z: $Y(0), Y(1) \perp Z$. Because of the connection to this classic design in causal inference, this assumption, sometimes referred to as the ignorability assumption, is also referred to as the randomization assumption.

The other estimands discussed above require slightly weaker assumptions. For instance, the effect of the treatment on the treated can be decomposed into two pieces, $E[Y(1)|Z = 1]$ and $E[Y(0)|Z = 1]$. Since we observe $Y(1)$ for the treated ($E[Y(1)|Z = 1] = E[Y|Z = 1]$), it is straightforward to estimate $E[Y(1)|Z = 1]$. However, it is difficult to estimate $E[Y(0)|Z = 1]$, since we do not observe $Y(0)$ for any treated individuals. It only becomes straightforward if $E[Y(0)|Z = 1] = E[Y(0)|Z = 0]$. This would be a consequence of randomization of the treatment Z as well since that design ensures that $Y(0) \perp Z$. Similarly, $Y(1) \perp Z$ would allow us to estimate the effect of the treatment on the controls. While these are weaker assumptions than are necessary for identification of the average treatment effect, it is difficult to find situations in which one assumption is clearly more plausible than the other.

Ignorability

Unless a randomized experiment has been performed, the above version of the ignorability assumption is typically not plausible because it implies that there are no systematic differences between the treatment and control groups with respect to the potential outcomes. Therefore, many causal inference methods rely on the more general version of the assumption above that conditions on observed covariates, X, formalized as $Y(0), Y(1) \perp Z|X$. In words, this assumption requires that the treatment and control group members that have the same values of X are not systematically different with regard to their potential outcomes. Some authors refer to this formulation as 'conditional ignorability,' however, this nomenclature is redundant since the standard formulation of the assumption allows for conditioning on covariates.

A simple example in which we would know that the ignorability assumption holds would be a randomized block experiment in which the blocks are defined by covariates, X. Observational studies differ in a crucial way from randomized block experiments, however, because no explicit randomization occurs; thus, we must take it on faith that the ignorability assumption holds.

How does ignorability help? When the ignorability assumption holds it implies the following properties:

$$E[E[Y|Z = 0, X]] = E[E[Y(0)|Z = 1, X]] = E[E[Y(0)|X]] = E[Y(0)]$$

and

$$E[E[Y|Z = 1, X]] = E[E[Y(1)|Z = 1, X]] = E[E[Y(1)|X]] = E[Y(1)]$$

This means that if ignorability holds (conditional on X) and we can appropriately condition on X, then all we have to do is average these conditional means over the distribution of X and we can recover the marginal means of the potential outcomes. If X is not low-dimensional, however, this conditioning can prove difficult in practice, which has led to use of some of parametric, semiparametric, and sophisticated nonparametric approaches discussed below or in other articles.

Common Support

Satisfying ignorability ensures comparability (with regard to potential outcomes) across treatment and control groups for observations with the same values of the covariates. However, this is only helpful if there is *common support* across treatment groups with respect to the confounders. That is, there must be observations from both treatment groups for each neighborhood of the covariate space where we want to make inferences. If, for example, no control units exist for a given neighborhood of the covariate space where treated units exist, then, in essence, no 'empirical counterfactuals' exist for those treatment units and we cannot make causal inferences about them without further assumptions. For this reason, causal researchers often require an assumption of overlap or 'common support' when attempting to identify causal effects, typically formalized as $0 < \Pr(Z|X) < 1$. In a randomized experiment, this is operationalized by the fact that all individuals in the experiment have a positive probability of receiving either treatment condition; individuals who could not receive one of the treatments would be deemed ineligible for the study.

In practice, if researchers determine that this assumption is not satisfied, they might decide to perform inference only for observations in portions of the covariate space where both treated and control units exist. This may lead to a different causal estimand than originally intended and results must be interpreted accordingly (see, e.g., Crump et al., 2006). This trade-off in what is being estimated and how well it can be estimated is often of particular relevance in studies that use propensity score methods, discussed further in other articles on observational studies.

The determination of common support depends on the inferential goal. For instance, if the goal is to estimate the effect of the treatment on the treated, this would require that control observations exist in every neighborhood of the covariate space where treatment observations exist, but not vice versa. If there are parts of the covariate space where there are controls but no treated subjects, it does not matter for nonparametric identification of the ATT since the goal is not to make inferences about the controls anyway. This lack of overlap could influence parametric estimation, however.

Common support is the only standard causal inference assumption that is empirically verifiable. However, the usefulness of this property is hampered by two issues. The first is that researchers may condition on a large number of covariates in an attempt to satisfy ignorability and it becomes increasingly difficult to verify overlap as the dimensionality of the covariate space grows. A common approach to this problem is simply to examine overlap with regard to the propensity score (*see* Observational Studies: Overview). However, this requires correct specification of the propensity score model. The second issue is that even if the included covariates satisfy ignorability, they may represent a superset of the covariates

actually needed to satisfy ignorability. Researchers forcing common support with respect to all included covariates may make overly conservative judgments about which observations need to be discarded from an analysis. Alternatives exist that aim to satisfy *common causal support*, $0 < \Pr(Z|W) < 1$, which requires overlap only with respect to the subset of covariates needed to satisfy ignorability, W (Hill and Yu-Sung, 2013).

Stable Unit Treatment Value Assumption

The potential outcome notation introduced above has an assumption embedded in it. The potential outcomes for individual i under treatment and control conditions, respectively, were defined as a function of the treatment exposure for individual i alone, $Y_i(Z_i = 1) = Y_i(1)$ and $Y_i(Z_i = 0) = Y_i(0)$. This ignores the possibility that this outcome could be influenced by the treatments received by others. A more general formulation, $Y_i(Z)$, could define potential outcomes based on the full vector of treatments received by everyone in the study, Z. Even with a modest sample size, such a formulation would increase the complexity of the research questions and subsequent analysis beyond a reasonable limit, not to mention the impossibility of ever achieving common support. For instance, in a study with a binary treatment and 10 people this formulation would lead to $2^{10} = 1024$ different potential outcomes for each person. The assumption that each person's potential outcomes are only defined based on his or her own treatment assignment can be formalized by the following expression: $Y_i(Z) = Y_i(Z')$, if and only if $Z_i = Z'_i$. This assumption is often discussed colloquially as the 'no interference assumption' to reflect the requirement that one individual's potential outcomes cannot be changed by another person's treatment. However, implicit in the stable unit treatment value assumption (SUTVA) is also the assumption that the treatment and control conditions mean the same thing for all units. In other words, there is only one 'version' of the treatment condition and one 'version' of the control condition (see Hernan and VanderWeele, 2011 for more thorough discussion of this assumption and its implications).

The most common strategy for avoiding concerns about SUTVA violations is through study designs that assign treatments at the group level. For example, many studies of educational interventions randomize whole schools to treatment conditions rather than randomizing individuals within schools, due to concerns that the treatment received by one individual may affect the outcomes of other individuals in the same school. This interference is less of a concern when considering schools as the units.

Temporality

Implicit in our understanding of causality is the notion that our causal variable (or treatment) occurs temporally prior to the outcome. Moreover, when conditioning on covariates it is crucial that they represent events that occur or characteristics measured prior to the causal (treatment) variable, otherwise they are implicitly outcomes and need to be conceived as such. Rosenbaum (1984) presents a discussion of this issue as do other authors (such as Gelman and Hill, 2007). Conditioning on variables that may be affected by the treatment (e.g., mediators) brings in additional complexities because they require consideration of the potential values of those post-treatment variables. A framework for thinking about post-treatment variables, called principal stratification, is discussed further in the article with that name. Other approaches to exploring causal pathways posttreatment are discussed in the article on Mediation.

Design Considerations

Causal researchers have carefully considered the types of research designs that facilitate satisfying the required assumptions. Generally, randomized experiments are seen as the strongest design for estimating causal effects since ignorability is assured through the randomization. Randomization also obviates the need to fit models that condition on pretreatment covariates and makes inferences from such models (when used, for instance, to increase efficiency) more robust to departures from the parametric assumptions. However, randomized experiments are not always possible, and can have their own complications that make inferences challenging, such as noncompliance and missing data (Barnard et al., 2003). A final concern with randomized experiments is that they often enroll subjects that are not representative of the population of interest, potentially making them less useful for estimating population treatment effects (Imai et al., 2008).

A number of strong nonexperimental study designs exist for settings when randomized experiments are infeasible or do not answer the question of interest. A common theme across nonexperimental studies is the need to design them with as much (or more) care as is used for randomized experiments (Rosenbaum, 1999; Rubin, 2008). Some nonexperimental designs are motivated by the desire to avoid or make more plausible the assumption of ignorability of the treatment assignment. These designs include regression discontinuity, interrupted time series, fixed effects, and instrumental variables approaches. They are discussed in more detail in other articles on Observational Studies.

Other nonexperimental study designs are motivated by the desire to avoid or weaken parametric assumptions by creating as much balance and overlap across treatment and control groups as possible. Matching and stratification (with or without propensity scores) are the most prominent designs in this genre. This type of 'preprocessing' (see Ho et al., 2007) can allow either for nonparametric inference on the newly constructed treatment and control groups (for instance, by comparing mean outcomes across matched treatment and control groups) or for more robust inference using standard parametric models (for instance, by fitting a regression on matched treatment and control groups). Treatment effect estimation is more robust to deviations from the parametric models when the treatment and control groups have common support and are well balanced (that is, the covariate distributions are similar across groups).

Nonexperimental studies of all types are ideally designed in ways that limit their sensitivity to unobserved confounding (violations of ignorability). Rosenbaum (2004) formalizes this

in a concept called 'design sensitivity' and Zubizarreta et al. (2013) provide one example of implementing these ideas in practice, such as comparing extreme values of a continuous treatment variable, matching on a large set of observed characteristics, and tying the statistical test used to the hypothesized pattern of effects. Gelman and Hill (2007) and Shadish et al. (2002) provide more information on nonexperimental study designs, and cite relevant articles on observational studies.

Parametric Assumptions

The focus of this article thus far has been on identification of causal effects through nonparametric approaches (comparisons of means or subgroup-specific means). It can be difficult to maintain a completely nonparametric approach in situations where it is necessary or beneficial to condition on covariates. For instance, if ignorability requires conditioning on covariates, then we must find a way to model the requisite conditional expectations. A straightforward approach to such conditioning is by fitting a linear regression model to the observed data. In particular, if one assumes that $Y(0) = X\beta + \varepsilon_0$ and $Y(1) = X\beta + \tau + \varepsilon_1$, then regressing observed Y on X and Z would be a reasonable approach for estimating τ. However, we rarely believe that such simple additive and linear models hold in practice, and it can be difficult to diagnose deviations from these models when X is high-dimensional or when there is lack of common support across treatment and control groups.

As discussed above, some designs (notably matching and stratification) are motivated by the desire to avoid these types of stringent parametric assumptions. However, an alternative (or complementary) approach is to simply use a semiparametric or nonparametric strategy when estimating causal effects. Examples of semiparametric estimation including regression models that incorporate weights based on the propensity score could also be considered to be a part of this paradigm (Hahn, 1998; Rosenbaum 1987; Hernan et al., 2001; Kurth et al., 2006). A more liberal interpretation might also frame weighting strategies as observational 'designs' because the goal is to create a 'pseudo-population' within which treatment and control groups are balanced. Nonparametric regression methods that can accommodate large numbers of covariates are typically Bayesian or computationally driven, or both (examples can be found in Hill, 2011 and Karabatsos and Walker, 2012).

Some causal inference approaches target the structural assumptions without considering the parametric assumptions. For instance, classic fixed effects models are often used in causal inference settings as a way of capturing group-level confounders that are not directly measured (*see* Hierarchical Models: Random and Fixed Effects). However, these models are traditionally embedded in a classical linear regression framework with all the parametric assumptions inherent in that framework. Moreover there is no guarantee that such strategies will reduce rather than amplify bias.

Conclusions

This article has discussed the potential outcomes framework for defining and estimating causal effects, in particular determining the effects of treatments on outcomes. It has highlighted the key structural and parametric assumptions commonly made, and the implications of those assumptions in experimental and nonexperimental settings. For more information on the details of estimating causal effects please see the other associated articles in this volume.

See also: Experimental Design: Bayesian Designs; Experimental Design: Compliance; Experimental Design: Large-Scale Social Experimentation; Experimental Design: Overview; Hierarchical Models: Random and Fixed Effects; Instrumental Variables in Statistics and Econometrics; Latent Structure and Causal Variables; Longitudinal Causal Inference; Mediation, Statistical; Observational Studies: Overview; Semiparametric Models; Time Series: General.

Bibliography

Austin, Peter C., 2007. The performance of different propensity score methods for estimating marginal odds ratios. Statistical Medicine 26, 3078–3094.

Barnard, J., Frangakis, C.E., Hill, J.L., Rubin, D.B., 2003. Principal stratification approach to broken randomized experiments: a case study of school choice vouchers in New York city. Journal of the American Statistical Association 98 (462), 299–323.

Crump, R.K., Hotz, J.V., Imbens, G.W., Mitnik, O.A., 2006. Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand. NBER Technical Working Paper No. 330.

Dowd, Bryan E., 2011. Separated at birth: statisticians, social scientists, and causality in health services research. Health Services Research 46 (2), 397–420.

Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge.

Greenland, S., Robins, J.M., Pearl, J., 1999. Confounding and collapsibility in causal inference. Statistical Science 14 (1), 29–46.

Hahn, J., 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66, 315–322.

Harder, V.S., Stuart, E.A., Anthony, J., 2010. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychological Methods 15 (3), 234–249.

Hernan, M.A., VanderWeele, T.J., 2011. Compound treatments and transportability of causal inference. Epidemiology 22 (3), 368–377.

Hernan, M., Brumback, B., Robins, J., 2001. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. Journal of the American Statistical Association 96 (454), 440–448.

Hill, Jennifer L., 2011. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics 20, 217–240.

Holland, P.W., 1986. Statistics and causal inference. Journal of the American Statistical Association 81 (396), 945–960.

Imai, K., King, G., Stuart, E.A., 2008. Misunderstandings between experimentalists and observationalists about causal inference. Journal of the Royal Statistical Society, Series A 171, 481–502.

Karabatsos, G., Walker, S.G., 2012. A Bayesian nonparametric causal model. Journal of Statistical Planning and Inference 142, 925–993.

Kurth, T., Walker, A.M., Glynn, R.J., Chan, K.A., Gaziano, J.M., Berger, K., Robins, J.M., 2006. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of non-uniform effect. American Journal of Epidemiology 163 (3), 262–270.

Pearl, J., 2009. Causality: Models, Reasoning, and Inference, second ed. Cambridge University Press, New York.

Rosenbaum, P.R., 1999. Choice as an alternative to control in observational studies. Statistical Science 14 (3), 259–304.

Rosenbaum, Paul R., 1987. Model-based direct adjustment. Journal of the American Statistical Association 82, 387–394.

Rosenbaum, Paul R., 2004. Design sensitivity in observational studies. Biometrika 91 (1), 153–164.
Rubin, Donald, 1978. Bayesian inference for causal effects: the role of randomization. The Annals of Statistics 6, 34–58.
Rubin, Donald B., 2008. For objective causal inference, design trumps analysis. The Annals of Applied Statistics 2 (3), 808–840.
Shadish, William R., 2002. Revisiting field experimentation: field notes for the future. Psychological Methods 7, 3–18.
VanderWeele, Tyler J., 2012. Invited commentary: structural equation models and epidemiologic analysis. American Journal of Epidemiology 176 (7), 608–612.
Zubizarreta, J.R., Cerda, M., Rosenbaum, P.R., 2013. Effect of the 2010 Chilean earthquake on posttraumatic stress: reducing sensitivity to unmeasured bias through study design. Epidemiology 24 (1), 79–87.

Causation: Physical, Mental, and Social

Daniel Stoljar, Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Many philosophers hold (what might be called) the causal hierarchy picture of the world. The picture is hierarchical because it views social entities (institutions, groups, or corporations) as being composed of psychological entities (persons or individuals), which are themselves composed of biological entities (cells), which are themselves composed of fundamental physical entities (atoms and subatomic particles). The picture is causal because it assumes that causal statements are potentially true at every level in the hierarchy. While attractive, the picture raises an important question about how any upper level causal statements can be true. On the one hand, it seems part of the picture that upper level causal statements are true only if corresponding lower level statements are, but on the other, if the corresponding lower level statements are true, there seems no reason for supposing that the upper level statements are true in the first place. This article discusses the formulation of the problem in contemporary philosophy and its relation to a different but superficially similar problem that was historically raised against Cartesian dualism. The article also suggests that the problem emerges in the causal hierarchy picture only from a mistaken interpretation of a common causal principle, usually called the exclusion principle.

We invoke the notion of causation to describe many different aspects of our experience and of the world. For example, in describing the physical domain, we might say the earthquake caused the building to collapse; in describing the psychological or mental domain, we might say paranoia was the cause of Oswald shooting Kennedy; and in describing the social domain, we might say excessive military spending caused instability in the Soviet Union. In addition, we often use causal notions to link these various domains together. For example, we might say her arm rose because she wanted to signal, linking the mental domain (wanting to signal) and physical domain (her arm rising); or social policy causes the degradation of the environment, linking the social domain (social policy) and the physical domain (the environment). In short, causation is a very general notion: it applies in and across the physical, mental, and social domains, and perhaps others too. According to many philosophers, however, there is an interesting and difficult class of problems that emerges when we focus on how causal statements in one domain – for example, the physical domain – interact with causal statements in others – for example, the social or mental domains. Problems in this class – which we will call problems of causation in multiple domains – is what this article is about (for causation in general, *see* Causes and Laws: Philosophical Aspects).

Cartesian Dualism and Gassendi's Objection

The seventeenth century philosopher and mathematician Rene Descartes famously held that, while the nature and behavior of nonhuman animals could be explained completely by the physical science of his day, various features of human beings – including in particular the fact that human thought and action are not under direct environmental control – seemed to place them beyond the reach of physical science. Descartes therefore endorsed dualism, the idea that every human being is a complex of a physical body located in space and subject to physical laws and an immaterial mind not located in space and not subject to physical laws. Descartes' contemporary Gassendi famously objected that dualism made a mystery of how bodily actions and movements could be caused by immaterial psychological events: 'How can there be effort directed against anything, or motion set up in it, unless there is mutual contact between what moves and is moved? And how can there be contact without a body?' (Descartes et al., 1984).

Gassendi's objection is an example of a problem of causation in multiple domains. It turns on the fact that there is causal contact between the mental and physical domains, something that is difficult to understand if Cartesian dualism is true. However, problems of this sort are by no means unique to the relationship between the mental and the physical, nor to Descartes' specific account of that relationship, and nor do they arise only from older ideas in science and philosophy. On the contrary, they arise from a general picture of the world, which is very common in contemporary intellectual culture: the causal hierarchy picture.

The Causal Hierarchy Picture

What sort of picture of the world does contemporary science presents us with? According to one common view, science presents us with a twofold picture. One part of the picture is that the world is organized into a series of levels (cf. Oppenheim and Putnam, 1958; Schaffer, 2003). The most basic level (assuming there is a most basic level) is the physical level, which is described by the most basic science, physics. Arranged in a hierarchy on top of this level are other levels such as the chemical level, the biological level, the psychological level, and the social level, and each of these levels are described by nonbasic sciences (sometimes called special sciences), such as chemistry, biology, psychology, and sociology. To put things the other way around, the picture views social entities (institutions, groups, or corporations) as being composed of psychological entities (persons or individuals), which are themselves composed of biological entities (cells), which are themselves composed of chemical entities (molecules), and which are themselves composed of fundamental physical entities (atoms and subatomic particles).

Another part of the picture adds causation to the hierarchy, in two distinct ways. First, it is part of the picture that, since

causation is a very general notion, causal statements are true at every level. Thus, causal claims about the behavior of corporations or persons are, from the point of view of the picture, perfectly genuine and potentially true. Second, it is also part of the picture that every time a causal statement is true at some nonbasic level, there is a corresponding causal statement that is true at the basic level.

The intuitive idea behind this second point is that causal statements at higher levels require mechanisms at lower levels, and, ultimately, at the physical level. Thus, consider the statement, she raised her arm because she wanted to signal. According to the picture, this statement (if true) is made true by, and requires for its truth, a further statement drawn from a lower level, presumably a statement about neural activity and how that activity causes movement of the limbs. Or again, consider the statement, urbanization in Australia in the twentieth century resulted in a decline of religious practice. This statement (if true) is made true by, and requires for its truth, a further statement drawn from a lower level, presumably a statement about the thoughts and goals of particular people that made up the relevant populations. In both cases, causal statements at upper levels require mechanisms at lower levels.

Many aspects of this causal hierarchy picture have been discussed and examined in recent philosophy of (natural and social) science (*see* Social Properties (Facts and Entities): Philosophical Aspects). But for our purposes, the important point is that it raises a problem of causation in multiple domains, and in fact a whole class of such problems. Here is a simple way of seeing the issue. Consider some causal statement *S*, which is putatively true at some nonbasic level *N*. Given the causal hierarchy picture, there will be another causal statement *S**, which is true at the basic level *B*. However, if *S** is true, and we know it is true given our picture of science, why should we suppose in addition that *S* is true? Intuitively, after all, all the causal work – all the pushing and shoving – is done at the basic level. But then it is mysterious why *S* – or any nonbasic causal claim – should be true.

While the issues are sometimes assimilated, the problem generated by the causal hierarchy picture is different from the problem that Gassendi raised for Descartes, in two respects. First, Gassendi's objection flows from the fact that, for Descartes, the physical and the mental domains are so different in nature: one is in space, the other not. But the problem raised by the causal hierarchy picture does not have its source in the fact that the mental level is so different in nature from the physical level. Its source is rather the simple fact that upper level causal claims are distinct from (i.e., not identical with) lower level claims, together with the idea that every time a causal claim is true at some nonbasic level, a corresponding causal claim will be true at the basic level. Second, the problem generated by the causal hierarchy picture is not limited to the mental and physical domains, as was Gassendi's objection. As we have seen, it arises for any nonbasic causal claim, whether in biology, psychology, or social theory. So long as one supposes that there are causal claims at these levels, a problem of causation of multiple domains can be raised.

We can now see the significance of the problem of causation in multiple domains. On the one hand, we have a picture of the world, the causal hierarchy picture, which is presented to us by science and therefore has a reasonable claim to be our best and most sophisticated picture of the world. On the other hand, the problem of causation in multiple domains apparently tells us that picture cannot be right, because the picture itself makes it hard to see how any nonbasic causal statements can be true.

The Exclusion Argument

We have so far been concerned with presenting the problem generated by the causal hierarchy picture in an informal way. However, it is desirable to be a bit more explicit about the reasoning that leads to the problem. The crucial fact about this reasoning is that it exploits a principle about causation often called the exclusion principle (Kim, 1993, 1998; Yablo, 1992; Bennett, 2003). As we shall shortly see, the precise formulation of this principle is a matter of controversy, but a reasonable initial formulation is (E1):

(E1) If e_1 causes e_2, then there is no event e_3 such that: (a) e_3 is not identical to e_1 and (b) e_3 causes e_2.

Intuitively, the idea here is that the discovery of a cause for some phenomenon usually means that other causes have been ruled out or excluded. For example, if a doctor tells you that the pain in your foot is caused by uric acid crystals in the joint of your big toe, you do not regard it as an open question whether the pain has some other cause – calcium crystals, for example.

With the exclusion principle in place, we can now be more explicit about the problem generated by the causal hierarchy picture. It is easiest to state the argument if we focus on a particular kind of causal statement, one in which a certain nonbasic event n causes a physical event p. (Later we will consider whether this assumption is misleading.) Thus, suppose that (1) is true

(1) n causes p.

Given the causal hierarchy picture, it would seem that there must be a physical event p*, which also causes p; hence (2) is true:

(2) p^* causes p.

Now, given that n and p* are from different domains, i.e., are drawn from different levels, we may further assume that

(3) n is distinct from (i.e., not identical to) p^*.

However, the principle of exclusion – in the form of (E1) – says that, in general, no two distinct events can cause a single event. Hence,

(4) abc

(4) If n is distinct from p^*, then it is not the case that both n and p^* causes p.

But (1)–(4) are jointly contradictory: while any three of them might be true, all four cannot be true together. It follows (barring some subtle ambiguity) that at least one of (1)–(4) must be false. In other words, what the reasoning tells us is that the causal hierarchy picture together with the exclusion principle presents a contradiction. To solve the problem, one must

give up or modify either the exclusion principle or the causal hierarchy picture or both.

Possible Responses

The most common response to the exclusion argument is to modify the principle of exclusion. Before turning to that response, however, it is important to briefly consider some alternative possibilities.

Rejecting the Hierarchy Part of the Picture

One obvious response is to suggest that the causal hierarchy picture is a distorted picture of the world because it takes the metaphor of levels too literally. This objection might go as follows. While it is true that there are apparently lots of different sciences, this is only a methodological or pragmatic fact, which represents our limited epistemological access to the world. In fact, there is only a single science – physics – which genuinely explains the world. The other sciences are simply convenient descriptions, which, apart from their pragmatic usefulness, will or could be dispensed within the future.

The main problem with this position is a feature of the causal hierarchy picture that so far has not been discussed: multiple realizability. Earlier we saw that according to the causal hierarchy picture, entities from upper levels are composed of entities from lower levels. But it seems an obvious empirical fact that the same entities from an upper level might be composed of many different entities at lower levels. Thus, the same corporation over time might be composed of different individuals. The same mental processes in different people might be made up of different neural processes – indeed, given science fiction possibilities, they might not even be made of neural processes at all. The lesson usually drawn from multiple realizability (as these sort of facts are called) is that entities from upper levels are genuinely distinct from lower level entities and, as a corollary, that the sciences dealing with upper level entities are genuinely distinct from lower level sciences (Fodor, 1974/1981); see Kim (1993) for an opposite view. However, if this is the right lesson, one should take seriously the hierarchy part of the causal hierarchy picture.

It is important to note that drawing this lesson from multiple realizability does not represent a return to Cartesian dualism. One can see this if one introduces another feature of the causal hierarchy picture: supervenience. Philosophers often say that, in the causal hierarchy picture, the upper levels supervene on the basic level, where this means that, while the upper levels in the hierarchy are distinct from the basic level, they are nevertheless wholly determined by, or are entailed by, the basic level (cf. Kim, 1993). A simple way to think of supervenience is by switching metaphors from levels to patterns. To say that the psychological or the social supervene on the physical is to say that these are patterns in the physical – hence are wholly physical – it is simply that they are not patterns that physicists are interested in. To put the point a little more formally, supervenience tells us that if you imagine a possible world that completely matches the actual world in all physical respects, then you have *ipso facto* imagined a possible world that also matches the actual world in all psychological and social respects. But this idea is something that Descartes would have denied. According to Descartes, the psychological (and presumably the social also) is only contingently related to the physical, and thus a possible world that completely matches the actual world in physical respects is not necessarily a world that completely matches it in these other respects.

Rejecting the Causal Part of the Picture

If one cannot reject the hierarchy part of the causal hierarchy picture, can one reject the causal part? One way to develop this objection starts with the observation that it is not at all obvious that particular sciences make explicit use of causal notions. Bertrand Russell (1917/1963) famously claimed that physics had no use for causation. More recently, many have argued convincingly that causation has no place, or at least no central place, in the explanations offered by particular special sciences such as psychology, linguistics (Chomsky, 1959; Cummins, 1983), and anthropology (Sperber, 1996). So perhaps the proper account of the picture of the world that science presents is that it is hierarchical but not causal.

However, while it might very well be true that particular sciences do not provide explicitly causal explanations, this does not mean that various causal claims will not be true at the various levels. We can see this if we examine briefly the dispute between Chomsky and Skinner, surely one of the central moments of twentieth-century science (Chomsky, 1959). Skinner had proposed, among other things, that the basic aim of scientific psychology and linguistics should be to provide causal explanations of human behavior – functional analysis, Skinner called it. Part of Chomsky's response to Skinner was that it is premature to suppose that scientific psychology could provide such explanations. He suggested instead that a more reasonable aim for scientific psychology would be the decomposition of certain psychological capacities and abilities, and in particular the ability on the part of a speaker to speak a language. I think it is fair to say that, in this regard, Chomsky's suggestion has been enormously influential and that most cognitive psychologists follow him in not aiming at causal explanations of behavior. However, even if this is true, it also remains true both that there are truths about how behavior is caused, and that these truths are in all probability multiply realized. But then we have the problem of causation of multiple domains back again.

One can imagine a more radical development of this point, where the claim at issue is not that causation plays no role in the explanations offered by special sciences, but rather that no nonbasic causal claims are true. In fact, this idea has been historically popular in philosophy of mind and psychology. In part because of the influence of Wittgenstein, it was at one time very common to hear that causal talk in the psychological domain either should not be taken seriously or should be interpreted very differently from similar talk deployed in the physical domain (Ryle, 1949; for criticism, see Davidson, 1963/1981). But suggestions of this sort consistently underestimate the sense in which causation is an extremely abstract and general notion. W.V.O. Quine famously argued that the difference between the statements 'cows exist' and 'numbers exist' is not a difference in existence, but rather a difference between cows and numbers. Similarly, one might

say that the difference between causal claims in the physical domain and, e.g., the psychological domain is not a difference in causation; it is simply a difference between physics and psychology.

Rejecting the Metaphysical Presuppositions of the Picture

Finally, a number of philosophers have attempted to avoid the exclusion argument by suggesting that it only arises from mistaken metaphysical views lying behind the causal hierarchy picture. For example, in stating the exclusion argument above, we assumed that an upper level event might cause a lower level event – (1) summarizes just that possibility. According to some philosophers, however, this sort of 'downward causation' is deeply mysterious, and causation takes place only within levels, not across levels (cf. Kim, 1993). However, if one rejects downward causation, one might reject the exclusion argument right from the start.

However, the problem with this suggestion is twofold. First, it is not clear that downward causation is so mysterious. As we have noted a number of times, causation is an extremely general notion. There seems no reason in principle why we should not suppose that upper level events might not cause lower level ones. Second, it seems possible to formulate the exclusion argument even if one does not start by assuming that downward causation is possible. For consider: given the causal hierarchy picture, every nonbasic event will supervene on some basic event. But then, one could causally explain the presence of any nonbasic event by causally explaining the basic event on which it supervenes. This brings the problem back again: given the causal hierarchy picture, there seems no rationale for supposing that any nonbasic causal statement is true.

Problems with Exclusion

If we cannot answer the exclusion argument by rejecting the causal hierarchy picture or some metaphysical presupposition of the argument, it would seem that the only option left is to give up or modify the exclusion principle. Indeed, objections to this principle are easy enough to find. Consider the spy shot by the firing squad: the first soldier on the firing squad caused his death, but so did the second. Hence, the spy's death is overdetermined. But a little reflection shows that the exclusion principle as stated so far rules out cases of overdetermination a priori.

This objection to the exclusion principle is suggestive but limited. It is true that there are possible cases of overdetermination. But it still seems unlikely that every case of nonbasic causation should be analogous to the case of the firing squad. To accommodate the possibility of overdetermination, we can reformulate the exclusion principle to read

E(2) If e_1 causes e_2, then (probably, in general) there is no event e_3 such that: (a) e_3 not identical to e_1 and (b) e_3 causes e_2

(E2) allows, where (E1) does not, the possibility of firing squad cases, but it also says that we can expect such cases to be the exception rather than the rule. On the other hand, it seems clear that we could formulate a probabilistic version of the exclusion argument based on (E2), which would be almost as bad as the nonprobabilistic version based on (E1).

A second objection to the exclusion principle cuts deeper, however. Consider the possibility of composition (or multiple realizability) all the way down: whenever we arrive at a potential basic level, it turns out that this level is itself composed at a lower level (for discussion of this idea, see Schaffer, 2003). Combining this idea with the exclusion principle yields some absurd results. For if there is composition all the way down, it is easy to see that the exclusion argument will tell us that any candidate causal statement is false. But this suggests that there really is something wrong with the exclusion principle. Surely a principle of causation should not have the consequence that no candidate causal claim is true.

In response to this objection, a proponent of the exclusion principle might take one of two options. First, one might deny the possibility that there is composition all the way down. However, this is a fairly desperate maneuver. It is surely an empirical question, not a question that can be decided a priori, whether the levels of the world simply go on forever. Second, and much more plausibly, one might revise the exclusion principle to accommodate the objection. The obvious way to do this is to replace 'identical to' in (E2) with 'supervenient on,' invoking the relation we discussed in the context of multiple realizability. That would result in the following principle:

E(3) If e_1 causes e_2, then (probably, in general) there is no event e_3 such that: (a) e_3 not supervenient to e_1 and (b) e_3 causes e_2

This version of the exclusion principle does not have the result that, if the world is composed all the way down, none of our paradigm causal claims are true. For it is consistent with (E3) that causal claims can be true at levels that supervene on the basic level. Thus, even if we assume that there is composition at every level, it is still consistent with (E3) that some causal claims at nonbasic levels are true.

On the other hand, if it is (E3), and not (E1) or (E2), which provides the proper articulation of the exclusion principle, then we have a response to the exclusion argument. The exclusion argument mistakes (E1) for (E3). Combining the causal hierarchy picture and (E1) yields a contradiction, but there is no problem with combining the causal hierarchy picture and (E3), for (E3) does not exclude the possibility of nonbasic causal claims.

Further Questions

Problems of causation of multiple domains threaten to undermine our picture of the world. But, as we have seen, these problems arise only from a mistaken conception of the exclusion principle, i.e., only if we formulate that principle as (E1) or (E2) and not (E3). However, it would be a mistake to conclude that this is the end of the matter. For (E3) generates some further puzzling questions of its own. I will close by briefly mentioning two of these.

The first question arises when we consider the difference between (5) and (6):

(5) For some events e_1, e_2, and e_3, if e_1 causes e_2, and if e_3 supervenes on e_1 then e_3 causes e_2

(6) For all events e_1, e_2, and e_3, if e_1 causes e_2, and if e_3 supervenes on e_1 then e_3 causes e_2

(E3) allows the possibility that (5) is true. But it had better not allow the possibility that (6) is. For (6) is subject to counterexamples such as the following (Jackson and Pettit, 1990). Suppose living in a particular neighborhood on the north side of town (e_1) makes you happy (e_2); living in a particular neighborhood on the north side of town entails living on the north side of town – as we might put it, living on the north side of town (e_3) supervenes on living on a particular neighborhood on the north side (e_1). Nevertheless, it might not be true that living on the north side (e_3) makes you happy (e_2) – after all, with the exception of your own particular neighborhood, you may dislike the north side, thinking of yourself as a 'south side person.' But now we are faced with a problem: if (6) is not true and (5) is, how are we to draw the line?

In order to answer this problem, Jackson and Pettit appeal to a principle of causation they call invariance of effect under variance of realization. They go on to develop this principle, embedding it in a framework for thinking about the relation between causation, causal explanation, and the causal hierarchy picture (especially as that applies to the social and psychological realms), which they call program explanation (cf. Jackson and Pettit, 1990, 1992; Pettit, 1993; for some similar work, see List and Menzies, 2009). Unfortunately, reviewing these ideas is beyond the scope of the present discussion. But the important point is that the idea that (E3) articulates the exclusion principle incurs an obligation: to explain what distinguishes the event triples $\langle e_1, e_2, e_3 \rangle$ that satisfy (5) from the event triples that do not.

The second question returns us to Gassendi's objection to Descartes. In response to Gassendi's objection, many Cartesian dualists are tempted to endorse epiphenomenalism. Epiphenomenalists accept that there is a correlation between mental events and behavioral events, but insist that the correlation in question is not causal. Rather, the correlation is the product of two other relations: a nomological (or lawful) relation between brain events and psychological events, and a causal relation between brain events and behavioral events. Since these latter relations can obtain without a causal relation obtaining between mental events and behavioral events, epiphenomenalism is a possible position, no matter how counterintuitive it seems to suppose that mental events do not cause behavioral events.

The problem that epiphenomenalism presents for (E3) and our discussion of it is the following. If we restrict attention to the relation between the psychological level and the neurological level, we too have articulated a correlation between mental and behavioral events, and for us too the correlation is a product of two further relations: first, there is a supervenience relation between the mental event and the brain event; second, there is a causal relation between the brain event and the behavioral event. However, why should we suppose that this correlation is causal? As the case of epiphenomenalism makes clear, the mere fact that a correlation is a product of a causal relation and some other relation does not make the correlation causal. Perhaps then, the answer we have given to the exclusion argument is no better than the epiphenomenalist answer to Gassendi?

Part of the answer to this question is that there are correlations and there are correlations. True, the fact that a correlation is the product of a causal relation and some other relation does not make it causal. But it had better not make it noncausal either – after all, every causal relation is a product of a causal relation and identity. So the structural similarity between the account we have offered and epiphenomenalism does not entail that account is no better than epiphenomenalism.

But there is also a somewhat deeper issue here. The world presents us with a myriad of correlations. Our concept of causation is in part a tool to divide these correlations into the causal and the noncausal. Obviously, there are some correlations that are clearly causal and some that are not. But it needs to be admitted that there are hard cases, cases in which it is difficult to say whether we have a causal correlation here or not. The causal hierarchy picture that we have been discussing presents us with such hard cases. Adopting a concept of causation that employs (E1), these cases will turn out not to be causal; adopting a concept of causation that employs (E3), they will. As we have seen, the latter concept does seem a reasonable concept of causation. But defending that choice in detail requires more investigation of the concept of causation that we can enter into here.

See also: Causes and Laws: Philosophical Aspects; Individualism versus Collectivism: Philosophical Aspects; Intentionality and Rationality.

Bibliography

Bennett, K., 2003. Why the exclusion problems seem intractable and how, just maybe, to tract it. Nous 37 (3), 471–497.
Chomsky, N., 1959. Verbal behavior Skinner B F. Language 35 (1), 26–58.
Cummins, R., 1983. The Nature of Psychological Explanation. MIT Press, Cambridge, MA.
Davidson, D., 1963/1981. Actions, reasons and causes. In: Davidson, D. (Ed.), Essays on Actions and Events. Oxford University Press, Oxford, UK.
Descartes, R., Cottingham, J., et al. (Eds.), 1984. The Philosophical Writings of Descartes, vol. 22. Cambridge University Press, Cambridge, UK.
Fodor, J.A., 1974/1981. Special sciences: or, the disunity of science as a working hypothesis. In: Fodor, J.A. (Ed.), Representations, first ed. MIT Press, Cambridge, MA.
Jackson, F., Pettit, P., 1990. Causation and the philosophy of mind. Philosophy and Phenomenological Research 50, 195–214.
Jackson, F., Pettit, P., 1992. In defense of explanatory ecumenism. Economics and Philosophy 8, 1–21.
Kim, J., 1993. Mind and Supervenience. Cambridge University Press, Cambridge, UK.
Kim, J., 1998. Mind in a Physical World. Cambridge University Press, Cambridge, UK.
List, C., Menzies, P., 2009. Nonreductive physicalism and the limits of the exclusion principle. Journal of Philosophy 106 (9), 475–502.
Oppenheim, P., Putnam, H., 1958. The unity of science as a working hypothesis. Minnesota Studies in Philosophy of Science 2, 3–36.
Pettit, P., 1993. The Common Mind. Oxford, University Press, New York.
Russell, B., 1917/1963. On the notion of cause. In: Russell, B. (Ed.), Mysticism and Logic. Penguin, New York.
Ryle, G., 1949. The Concept of Mind. Hutchinson's University Library, london.
Schaffer, J., 2003. Is there a fundamental level? Nous 37 (3), 498–517.
Sperber, D., 1996. Explaining Culture. Oxford, UK: Blackwell.
Yablo, S., 1992. Mental causation. Philosophical Review 101, 245–280.

Causes and Laws: Philosophical Aspects

Helen Beebee, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The questions 'what makes it the case that one event causes another?' and 'what makes it the case that something is a law of nature?' are highly controversial ones for which, among contemporary philosophers, there is no answer that can claim orthodoxy. This article sketches some of the most influential theories of causation and lawhood.

Preliminaries

'Cause' and 'law' are perhaps two of the most important and fundamental concepts that human beings deploy in their attempts to understand and intervene in their environment, both in everyday life and in their scientific endeavors. When we want to explain why an event occurred, we seek out its causes. When we act, we do so because we believe the action will have certain effects. When we want to know why our environment and our fellow human beings behave in regular, predictable ways, we look for the laws that govern that behavior.

It is part of the scientist's job to discover what causes what, and to investigate what the laws of nature are. The philosopher's job, however, is a more abstract one; the philosopher wants to know what makes it the case that something – anything, be it the decay of a subatomic particle or the assassination of Archduke Ferdinand – causes something else, or that something is a law of nature – whether it is a physical, biological, psychological, economic, or any other kind of law. We know that when one strikes a match and it lights, the first event – the striking – caused the second. But this causal fact depends for its truth on more than the mere fact that the two events occurred one after the other; there must be some extra feature of the world that binds these two events together as cause to effect. Similarly, it seems that laws of nature must be more than mere regularities; it is doubtless true that all lumps of gold are smaller than a mile in diameter, but it is not a law that this is so. So genuine laws must have some extra feature, apart from regularity, which marks them off from mere 'accidental' regularities. The philosopher's job, or at least part of it, is to say what, if anything, this 'extra feature' might be.

Determinism vs Indeterminism

A question that is relevant to any discussion of the nature of causes and laws is that of whether the universe is fundamentally deterministic or indeterministic. One way of putting the thesis of determinism is this: the complete state of the universe at any given time together with the laws of nature determines precisely what the complete state of the universe will be at all future times. Indeterminism is simply the denial of determinism.

The view that the universe is fundamentally indeterministic has been widely (although by no means universally) held within science since the advent of quantum mechanics; however, most philosophers only began to take the possibility of indeterminism seriously in the 1960s and 1970s. One consequence of this reluctance to take indeterminism on board is that some current theories of causes and laws were originally formulated under the assumption of determinism, and were then modified to take into account the possibility that at least some laws and causal processes are fundamentally indeterministic. The theories of laws and causation discussed below are mostly presented in their simpler, deterministic form. While the indeterministic versions of the theories present additional complications and problems – for example, because they employ the notion of probability – the fundamental philosophical issues addressed, and problems faced, by the indeterministic and deterministic versions, are mostly the same or similar.

The Particular and the General: Events, Facts, and Properties

Many of our causal claims are claims about individual occurrences: the assassination of Archduke Ferdinand, the Great Recession, the birth of a particular child, a specific meteor shower, or whatever. Such occurrences are thus the 'relata' of particular or 'token' causal relations. There is an ongoing debate about whether such individual occurrences should be conceived as events (the assassination of Archduke Ferdinand) or facts (the fact that Archduke Ferdinand was assassinated) – see, for example, Mellor, 1995 and Lewis, 1986 – but for simplicity we shall assume that they are events.

Sometimes, however, we make general causal claims about whole populations rather than their individual members. Thus, it is doubtless true that smoking causes cancer, but it does not follow that every individual who smokes will be caused to get cancer as a result; and it can be true that a particular person dies as the result of a bee sting without its being true that bee stings in general cause death. General or 'population-level' causation is generally taken to apply to properties or repeatable features that different individuals can share, such as smoking, getting cancer, being red, or weighing 3 kg.

Laws of nature are likewise concerned with general relationships between properties – whether it is the relationship between force, mass, and acceleration enshrined in $f = ma$ or the economic 'law' of diminishing marginal utility. In what follows, general features or properties will be denoted 'F,' 'G,' and so on, while individual events will be denoted 'c,' 'e,' etc.

Laws of Nature

Laws vs Accidents

What is the difference between laws of nature and merely 'accidental' regularities? (Suppose that it is a law that all metals expand when heated and that it is merely accidentally true that all (past, present, and future) Queens of England are less than 2 m tall.) A crucial difference is that laws 'support counterfactuals,' whereas accidental regularities do not. It is true of any unheated piece of metal that if it were to be heated it would expand. But it need not be true of any non-Queen of England that if she were to be Queen she would be less than 2 m tall. If someone is currently, say, 2.05 m tall, becoming Queen of England would not result in her getting any shorter.

Another difference is that violations of law are said to be physically impossible, whereas violations of accidental regularities are physically possible. It is physically impossible for a heated metal to fail to expand when heated; but it is perfectly possible for a Queen of England to be taller than 2 m. Finally, it is sometimes claimed as a third difference that laws, unlike accidental regularities, 'govern' what goes on in the universe. We tend to suppose that statements of laws of nature are rather like pieces of divine legislation: decrees that the universe must obey rather than mere general descriptions of what in fact happens.

Realist vs Regularity Accounts of Laws

Can an analysis of lawhood be given that does justice to these intuitions about the nature of laws? A popular view (Armstrong, 1983) is that laws are relations of necessitation between universals. To say that it is a law that all *F*s are *G*s is to say that there is a necessary relation, *N*, that holds between the universals (properties) *F* and *G*. (Following Armstrong, we can write this '*N*(*F*,*G*)'.) When an object instantiates *F*, the instantiation of *F* guarantees, via *N*, that *G* will also be instantiated. Thus, *N*(*F*,*G*) – its being a law that all *F*s are G*s* – guarantees that all *F*s will in fact be *G*s.

The basic point of this 'realist' theory of laws is that it provides an 'ontological ground' for the difference between laws and accidents; accidental regularities just happen, whereas lawful regularities are grounded in and explained by a necessary relation holding between the co-occurring properties. And, on this view, sense can be made of the idea that laws govern what happens; if *N*(*F*,*G*) holds, then *G* must be instantiated whenever *F* is instantiated.

The main rival to this conception of laws is the 'Humean' or 'regularity' conception, according to which laws are really no more than regularities. The classic (although not the only) motivation for the regularity view is an 'empiricist' epistemology according to which we ought not to believe in entities that we cannot perceive. Since our evidence that there are any laws of nature comes solely from the observation of regularities – when we see a law (as opposed to an accidental regularity) being instantiated we do not see any extra ontological feature – we ought not to believe that there is any extra ontological feature that laws have but accidental regularities lack.

A 'naïve' regularity theory, however – according to which statements of law are merely true universal generalizations – is untenable, for it simply collapses the distinction between laws and accidental regularities. If statements of law are merely true universal generalizations, then accidental regularities are themselves laws; it turns out to be a law of nature that no Queen of England is taller than 2 m after all.

A more sophisticated version of the regularity theory is provided by the 'Ramsey–Lewis view,' after F.P. Ramsey and David Lewis (see Lewis, 1973b: pp. 73–75; J.S. Mill also held a similar view). The basic idea of the Ramsey–Lewis view is as follows. When we seek, as scientists do, to find out how the universe behaves, we are not satisfied with isolated facts about what happened in the laboratory on Tuesday afternoon, what happened when a particular patient took a particular drug, and so on. Rather, what we seek are wide-ranging generalizations that abstract as far as possible from the particularities of the laboratory or the patient. Generalizations about the heights of Queens of England or the diameters of lumps of gold apply to an extremely restricted number and range of phenomena. On the other hand, generalizations about subatomic particles, chemical reactions, the relationship between force, mass, and acceleration, and so on, apply to an enormous number of diverse phenomena, and are hence much more useful and explanatorily powerful. According to the Ramsey–Lewis view, the difference between laws of nature and merely 'accidental' regularities amounts to roughly this difference between generalizations that are wide ranging and powerful and those that are not. It is not that there is any ontological distinction between laws and accidents – laws do not have an extra feature (*N*, say) that accidental regularities lack; rather, it just so happens that some generalizations (the laws) are interesting and useful ways of codifying what happens in the universe, and others (the accidents) are not.

The Ramsey–Lewis view (and the Humean view of lawhood in general) diverges from a conception of laws of nature according to which laws 'govern' what happens in the Universe, since on the Humean view the laws just are regularities, rather than features of the world that explain why the world is regular (Beebee, 2000). Some philosophers argue that Humeanism about laws makes it impossible for us to know any laws of nature (Armstrong, 1983: pp. 52–59; Bird, 2008; Beebee, 2011). On the other hand, some philosophers claim that Humeanism about laws takes the sting out of the alleged conflict between determinism and free will (see Swartz, 1985: Chapters 10 and 11; Beebee and Mele, 2002; Berofsky, 2012) (*see* Free Will and Action).

Ceteris Paribus Laws

The accounts of laws canvassed above are perhaps best suited to the laws we find in, say, physics and chemistry – the places where we are most likely to find truly exceptionless generalizations that hold across the whole of space and time (although it is often argued that even in the 'hard' sciences, truly exceptionless generalizations are extremely rare). There are plenty of generalizations in, say, economics, sociology, and psychology that, while not exceptionless, are clearly lawlike in some sense. The law of diminishing marginal utility, for example, does not hold in all cases, so if lawhood requires an exceptionless regularity, there will turn out to be no laws in economics.

On the other hand, it is surely no accident that, by and large, the utility of something eventually decreases as one obtains a greater quantity of it. Similarly, there are plenty of generalizations in biology and psychology that are not exceptionless but nonetheless provide a good foundation for prediction and control.

Such cases are generally thought of as '*ceteris paribus*' laws ('cp-laws' for short): generalizations that hold not in all cases *simpliciter*, but only 'other things being equal,' that is, in cases where there is no additional, interfering factor. We might try to subsume cp-laws under our existing accounts of laws simply by building in the absence of such interfering factors, so that our candidate law has the form 'all *F*s are *G*s, except in circumstances *X*, *Y*, and *Z*.' This approach is problematic, however, since in many, if not most, cases, we either have no idea what the relevant *X*, *Y*, and *Z* are, and so – even if there is a law in the offing – we will not be in a position to formulate it, or else there will be indefinitely many ways in which the generalization in question could in principle fail, so there is no way in principle of specifying what the relevant law is.

We might try to avoid these problems by claiming that cp-laws are only supposed to hold in 'ideal' or 'normal' conditions, thus removing the need to specify exactly what those conditions are within the content of the law itself. However, unless we can independently specify which conditions count as normal or ideal, the threat of circularity looms; we cannot allow 'normal' or 'ideal' conditions to simply be those in which the regularity obtains, or cp-laws will turn out to be trivially true (all *F*s are *G*s – except when they are not). (See Hausman, 1992: Chapter 8; Pietroski and Rey, 1995; Earman and Roberts, 1999; Schrenk, 2007a.)

Stability and Invariance

Some contemporary accounts of lawhood take a different starting point to the idea that a law is an exceptionless regularity of a certain kind (the Ramsey–Lewis view) or else is the obtaining of some metaphysical relation that entails the existence of an exceptionless regularity (Armstrong's view). Instead, they start from the idea that the distinctive feature of laws is that they are stable or invariant under a range of different background conditions. Suppose that the generalization 'all the pens currently on Obama's desk are blue' is true. Despite being true, it is highly unstable; there are very many possible situations in which it would be false. Obama might, for example, have left the red pen he was using earlier today on his desk rather than putting it in his pocket, or one of his aides might have left their own black pen on the desk. By contrast, 'nothing travels at the speed of light' is highly stable; there are (so we currently believe) no physically possible situations at all that would render it false. The notion of stability is thus a 'modal' notion; the stability or otherwise of a generalization is a matter of whether it would still be true in various different, nonactual situations. And, since stability is not all-or-nothing but rather comes in degrees – one generalization can be more or less stable than another – we can think of lawhood itself as coming in degrees, so that some generalizations are more lawlike than others, rather than thinking that a generalization either expresses a law or it does not.

The notion of stability (or invariance) may be a good place to start if we want to make sense of cp-laws, because it seems to capture what is 'lawlike' about, say, the law of diminishing marginal utility. While the generalization in question only holds *ceteris paribus*, it arguably still exhibits a good deal more stability than, say, the generalization concerning the pens on Obama's desk; it would remain true (*ceteris paribus*) given a wide range of changes in prices, consumer preferences, and so on. Of course, we still need to find a way to spell out the restrictions imposed by the *ceteris paribus* clause; it has been suggested that we think of cp-laws as generalizations that are stable across circumstances that are of particular interest to a given science, such as psychology or economics (see Lange, 2002; Woodward and Hitchcock, 2003a,b).

Laws and Dispositions

A final kind of account of the nature of laws – again, one that is perhaps well suited to dealing with cp-laws – is a 'dispositionalist' account. According to a naïve version of the so-called counterfactual account of dispositions (such as fragility, which, when applied to panes of glass and china cups at least, we can think of as the disposition to break when dropped), an object *o* (a china cup, say) has disposition to manifest feature *M* (being broken) in 'stimulus conditions' *S* (being dropped) just if, if *o* were to be in situation *S*, it would manifest *M*. As it stands, however, this account cannot be right. For example, if I were to drop a china cup that is wrapped in a lot of cotton wool, it would not break – but the cup is nonetheless fragile. The presence of the cotton wool 'masks' the disposition – it prevents the disposition from being manifested. Antidotes provide another counterexample to the naïve counterfactual account; the redback spider is deadly (its bite has the disposition to kill people) despite the fact that, thanks to the existence of an antivenom, nobody has died from a redback bite for many years.

Whether the counterfactual account can be amended to deal with problem cases such as these is a controversial question (see e.g., Lewis, 1997; Bird, 1998). What is relevant to the discussion of laws, however, is that we can think of laws of nature as describing not how things actually behave, but how they are disposed to behave. So a candidate law of nature might have the form 'all *F*s are *D*s,' where property *D* is the disposition or capacity to be *G*, rather than 'all *F*s are *G*s' (Cartwright, 1989; Ellis, 2002; Bird, 2005).

A dispositionalist account of laws may be well suited to deal with cp-laws because – as we have just seen – something's having a disposition is compatible with its failing to manifest that disposition even though the stimulus condition is present. If we think of laws as generalizations about dispositions, then we can think of cp-laws as those laws that describe dispositions that are susceptible to masks, antidotes, and the like. Thus, for example, 'all china cups break when dropped' is false; I can easily falsify it by dropping a china cup wrapped in cotton wool. But this does not falsify 'all china cups are disposed to break when dropped,' since, as we have seen, a china cup is still disposed to break when dropped – it is still fragile – even if, thanks to the cotton wool, it is not in a position to manifest that disposition when I drop it. (See Schrenk, 2007b.)

Causation

Theories of causation have proliferated in the last 50 years or so, with different theories speaking to different underlying metaphysical commitments and to different views about how the ontology of causation (i.e., how the world has to be arranged in order for causal claims to be true) ought to relate to its epistemology (i.e., how we can come to know, or justifiably believe, that a causal claim is true). This section surveys some of the leading theoretical approaches.

The Naïve Regularity Theory

The agenda for much of the contemporary discussion about the nature of causation was set by Hume (1748/1975: Sections IV–VII). Hume was an empiricist, believing that there could be no 'ideas' in the mind that do not somehow come from our senses. How, then, do we come to have an idea of causation? What features of the world could furnish us with that idea? Famously, Hume maintained that we do not perceive any intrinsic connection or relation between two events we judge to be causally related; we see the match being struck, and we see it light, but we do not see any causation between those two events. Causation, then, cannot be a mysterious intrinsic relation between events – for if it were, since we cannot perceive any such relation, we could have no idea of causation. (See Menzies, 1998; Beebee, 2009 for discussion of the question of the observability of causation.)

The exact nature of Hume's positive theory of causation is a matter for interpretative dispute (see Beebee, 2006a). However, the standard interpretation – and therefore the one that has had the most impact on contemporary discussion – characterizes Hume as a 'naïve regularity theorist.' On this view, two events are related causally in virtue of contiguity (causes and their immediate effects are right next to each other in space and time), temporal priority (causes precede their effects), and constant conjunction (events similar to the cause are always followed by events similar to the effect). So what makes it true that the striking caused the lighting is that the two events are contiguous (or, if there is some spatiotemporal gap between them, the events are mediated by further events so that there is a chain of contiguous events starting with the striking and ending with the lighting); the striking occurs before the lighting; and similar strikings are always followed by similar lightings. The constant conjunction requirement ensures that for Hume (and for most modern day Humeans) causation is an extrinsic relation: a causal relation obtains between two events not purely in virtue of how those events themselves are, but in virtue of features of other events; the striking only gets to count as a cause of the lighting because other strikings are also followed by lightings.

The other notable feature of the naïve regularity theory is that it is a reductionist view: causal claims, like 'the striking of the match caused the lighting,' are made true, at bottom, not by any primitive causal feature of the world, but by noncausal features (since the fact that two events are contiguous, the fact that one occurs before the other, and the fact that they are similar to other events are not themselves causal facts).

The naïve regularity theory is now universally regarded as untenable, although, as we shall see, elements of it are clearly discernable in many contemporary theories. It is untenable for a variety of reasons. For example, the contiguity requirement rules action at a distance (one event's causing another, spatially or temporally distant event without there being a chain of events 'hooking up' the first to the last) to be impossible, while many philosophers think that is at least conceptually possible. And the constant conjunction requirement is obviously flawed: there can be constant conjunction without causation – for example, joint effects of a common cause are constantly conjoined but neither causes the other – and, if determinism is false, it seems there can be causation without constant conjunction. My betting on number 13 on the roulette wheel is a cause of my winning some money, even if in exactly the same circumstances the following day I bet on number 13 and lose.

Nevertheless, the general issues that Hume raised are still among the main foci of dispute in contemporary philosophy of causation. In particular, the question of whether causal facts reduce, somehow or other, to facts about regularities is the driving force behind many of the theories of causation currently on offer.

Counterfactual Analyses

Arguably, the most influential analysis of causation in recent years has been Lewis's counterfactual analysis (Lewis, 1973a). According to Lewis's analysis, an event e causally depends on event c if e counterfactually depends on c – which is to say, if, had c not occurred, e would not have occurred either (*see* Counterfactual Reasoning, Quantitative: Philosophical Aspects and Counterfactual Reasoning, Qualitative: Philosophical Aspects). A 'chain of causal dependence' is a series of events $\langle a, b, c, \ldots, n\rangle$ such that b causally depends on a, c causally depends on b, and so on. Finally, c causes e if there is a chain of causal dependence (perhaps involving only c and e themselves, but perhaps involving many hundreds of intermediate events) from c to e.

The need for chains of dependence arises when there is no counterfactual dependence between cause and effect because of a backup mechanism or 'preempted alternative.' Suppose that the bus stop is next to the taxi rank. Susan takes the bus to work (c); but if the bus had been late she would have taken a taxi. She attends a 9.30 meeting (e). e does not counterfactually depend on c, since if Susan had missed the bus she would have taken a taxi and attended the meeting just the same. But there is some intermediate event d – Susan's getting off the bus at the stop outside her office, say – which counterfactually depends on c (if she had not got on the bus in the first place, she would not have got off it either) and upon which e counterfactually depends (if she had not got off at that stop – given that she was already on the bus, and it was already nearly 9.30 – she would have had to walk back to her office from the next stop, and she would have missed the meeting).

Lewis's counterfactual analysis of causation has been found to be subject to a number of counterexamples (e.g., Schaffer, 2000a). Much recent work on causation has been devoted to devising alternative analyses which, while remaining true to Lewis's basic claim that causation is to be analyzed in terms of counterfactuals, add further complexities in an attempt both to circumvent the problem cases and to deal with

indeterministic causation (see, e.g., Menzies, 1989; Noordhof, 1999; Lewis, 2000).

Both Lewis's original analysis and its more complex successors fall squarely within the Humean tradition, since the counterfactual analysis is a reductionist project: it seeks to show how causal facts are made true by noncausal features of the world. Indeed, since Lewis's account of counterfactuals make the truth conditions of counterfactuals depend primarily on the laws of nature, and the laws of nature themselves, on Lewis's view, are merely a subset of the regularities, one can think of counterfactual analyses of causation generally as more sophisticated descendants of the naïve regularity theory of causation.

Not everyone, however, is a Humean. Many philosophers deny Hume's central motivating claim that causation cannot be perceived (see Anscombe, 1971; Armstrong, 1997: Chapter 14; Menzies, 1998; Beebee, 2009). Some argue that the features of the world to which reductionist analyses seek to reduce causal facts are themselves really causal features; hence reductionist analyses are circular and therefore unsuccessful (see Carroll, 1994). Others argue that if a broadly Humean analysis of causation were true it would render the pervasive regularities that the universe exhibits a sort of cosmic fluke (see Strawson, 1989: Chapter 5; Beebee, 2006b). And many simply think that the persistent failure of reductionist accounts to deal with counterexamples, despite ingenious amendments, is good reason to think that the whole reductionist project is doomed to failure.

Process and Mechanistic Theories

We generally take it for granted that causal influence is propagated through mechanisms or processes of various kinds. When the impact of the white snooker ball causes the black ball to roll into the pocket, we can grant that the latter event counterfactually depends on the former, but may yet be inclined to say that this dependence is symptomatic rather than constitutive of the causal relation between the two. This is because there is an identifiable process linking the two, involving the transfer of energy-momentum from the white to the black ball and subsequent movement of the black ball toward the pocket. Some philosophers have attempted to analyze causation by focusing on the distinction between a genuine causal process and a mere 'pseudoprocess.' Salmon (1984), for example, takes the distinguishing feature of causal processes to be their ability to 'transmit a mark.' A mouse walking across the floor is a process that can transmit a mark; for example, if I paint a blue patch on the mouse's fur, the blue patch will 'transmit' from earlier to later stages of the process (since it will go wherever the mouse goes). By contrast, if I make a spot of light move across the floor by means of a torch, the movement of the light across the floor is a pseudoprocess; if I paint the first illuminated patch blue, that mark will not transmit to later stages of the 'process'; it will stay just where I put it.

Salmon's 'mark transmission' theory is problematic for various reasons (see Dowe, 2008: Section 6). Dowe's suggested replacement, the 'Conserved Quantity Theory' (2000), identifies the causal relation with the transfer of whatever physical quantities are conserved according to the conservation laws of physics, for example, energy-momentum. A problem with this account is that it only tells us what causation actually consists in. It seems that we can imagine possible situations where there is causation in the absence of any transfer of a conserved quantity (e.g., by the casting of spells). Other analyses of causation – such as the counterfactual analysis – tell us why both the movement of snooker balls does, and the effective casting of spells would, count as cases of causation, since they both involve counterfactual dependence. Dowe's account fails to do this.

One solution to the problem is to attempt to spell out the 'folk theory' of causation: those features that, according to our ordinary concept of causation, are essential for a relationship to be causal, and hence those features that will apply to snooker balls and spells alike. And we can then claim that the transfer of conserved quantities counts as causation precisely because, as things actually are, it is such transfer that satisfies our folk theory or definition of causation – while at other possible worlds, such as one in which magic spells are effective, other kinds of relation count as causal (see Dowe, 2000: Chapter 1; Menzies, 1996; Lewis, 2004).

Like process theories, 'mechanistic' accounts of causation also seek to identify causation with the process by which the cause produces the effect (Glennan, 1996; Machamer et al., 2000; see Williamson, 2011 for a survey). However, mechanistic accounts are more pluralist in spirit, and do not try to locate a single feature of the world (such as the transfer of a conserved quantity) that can be identified with the causal relation. Machamer et al. note that "terms like 'cause' and 'interact' are abstract terms that need to be specified with a type of activity and are often so specified in typical scientific discourse. Anscombe (1971) … noted that the word 'cause' itself is highly general and only becomes meaningful when filled out by other, more specific, causal verbs, for example, scrape, push, dry, eat, burn, knock over. An entity acts as a cause when it engages in a productive activity" (Machamer et al., 2000: p. 6; see also Cartwright, 2004).

Dowe's account is explicitly reductionist in spirit, in two senses. First, it attempts to identify causation with a feature of the world that can itself be understood in noncausal terms; the 'transfer' of conserved quantities is not supposed to be conceived as, for example, the presence of an energy-momentum in the white ball causing its subsequence presence in the black ball. It is also reductionist in the sense that all (actual) causation is, at bottom, causation at the level of fundamental physics; the presence of a causal process is determined by features of the world that are visible, as it were, only from the perspective of physics. Machamer, Darden, and Craver's account is antireductionist in both senses. 'Productive' activities are not themselves to be analyzed in noncausal terms (e.g., the instantiation of regularities or the transfer of conserved quantities); nor are the mechanisms that underpin causal relations to be found only at the level of fundamental physics. Indeed, part of the purpose of their view is to make sense of causal explanation in biology.

Probabilistic Theories

Frequently, it is claimed on cigarette packets that smoking causes heart disease; this is a claim about 'general' or 'type-level'

or 'population-level' causation. What makes (or would make) such a claim true? We might try to regard it as some kind of generalization over individual instances of particular smokers whose heart disease has been caused by their smoking. But it is clearly not a universal generalization; the fact that Jane – a lifelong smoker – had no heart disease when she died from other causes (falling off a cliff, say) does not serve to refute the claim that smoking causes heart disease. Nor is our general causal claim merely an existential claim to the effect that at least one smoker has been caused by their smoking to develop heart disease. If that were so, then it would take just one unlucky person to die because of choking on a Mars Bar to make it true that eating Mars Bars causes imminent death.

This being so, we should perhaps regard general and 'token' or 'individual' causation as two different relations, the former relating types or properties, and the latter relating individual events or instantiations of properties – something that is suggested in any case by the grammatical form of general causal claims ('smoking causes heart disease,' and not 'instances of smoking cause instances of heart disease'). Indeed, some authors have argued that we have not only two different relations, but two relations that are independent of one another (e.g., Eells, 1991: Chapter 1). Others argue that the two 'levels' of causation are in fact closely related (Hitchcock, 1995), so that, once we have the right account of causation on the table, we will see that we do not really have two distinct species of causation at all.

What is generally accepted, however, is that general causation should be understood in terms of probabilistic or statistical relationships, so that, in essence, a claim of the form 'C causes E' is made true by C's raising the probability of E. (Similarly, 'C prevents E' is made true by C's lowering the probability of E.) According to one account (Eells, 1991; see also Suppes, 1970), we should understand C's raising the probability of E as the obtaining of $\Pr(E|C \,\&\, B_i) > \Pr(E|\sim C \,\&\, B_i)$, where B_1, B_2, … are all the different combinations of the presence and absence of other relevant factors within a given population. So, for example, suppose we divide our population (for example, UK citizens) into 'homogeneous reference classes,' so that within each reference class everyone is exactly alike with respect to all other risk factors for heart disease (alcohol consumption, cholesterol levels, relevant genetic factors, etc.). Then 'smoking causes heart disease' is true if and only if in each such reference class, Pr(heart disease|smoker) > Pr(heart disease|nonsmoker).

This condition is extremely hard to satisfy, however. For example, imagine that there is a very rare genetic trait that interacts with smoking in such a way as to decrease the risk of heart disease. Then on the above account it will be false that smoking causes heart disease for any population that includes possessors of this genetic trait as a subpopulation. A rival proposal (Dupré, 1984, 1990) is that C causes E if and only if C raises the probability of E in a fair sample of the population (i.e., in a sample of the population in which the relative frequency of other relevant factors is the same as it is for the population as a whole). This analysis yields the desired result that smoking causes heart disease in the above example, since in any fair sample those with the genetic trait will be vastly outnumbered by those who lack it; so C will still raise the probability of E in a fair sample. An alleged benefit of Dupré's analysis is that it makes the truth conditions for general causal claims match up in an obvious way with a standard way in which general causal claims are in fact established, namely, randomized control trials. The point of randomization is precisely to make it likely that other relevant factors are present with equal frequency in both the trial group and the control group, and an increase in frequency of the effect under investigation in the trial group relative to the control group is supposed to be good evidence that the drug, or whatever, is a cause of that effect.

The accounts just discussed are reductionist – indeed, Humean – in spirit, in that they seek to analyze (general) causation in terms of regularities (although the regularities in question are statistical rather than exceptionless). A more recent and similar approach – the 'manipulability' or 'interventionist' approach – shuns reductionism, but nonetheless aims to shed light on how causal structure can be inferred from statistical evidence (Woodward, 2003, 2008).

The need for the appeal to either homogenous reference classes or to fair samples stems in part from the need to rule out 'spurious causation;' cases of probability increase in the absence of causation. One factor F can easily increase the probability of another, G, in the population as a whole without causing G, as when, for example, F and G are effects of a common cause H. For example, the barometer needle pointing to 'rain' (F) is positively correlated with rain (G), but both are effects of a common cause, namely, low atmospheric pressure (H). (We can think of the population in this case as comprising 'weather episodes.') There will be no such spurious correlation in a homogeneous reference class, however, since by hypothesis all members of any such reference class will be alike with respect to whether or not they possess H. 'Holding fixed' factor H screens off the statistical relationship between F and G. Similarly, if we compare a fair sample of Fs with a fair sample of $\sim F$s, by definition H will have the same relative frequency within each of the samples, and so, again, there will be no correlation between F and G.

Interventionist accounts provide a different way of 'breaking' spurious probabilistic relationships. The basic idea is very simple. Take the barometer case again. F and G are positively correlated within the population as a whole. But if I intervene on F, say by manually moving the pointer so it either points to 'rain' (f_1) or to 'sunny' (f_2), the correlation disappears: there is no positive correlation between F and G if we consider just those cases where I have directly brought about either f_1 or f_2 (so long as my manually moving the pointer is not itself correlated with the atmospheric pressure). In effect, my intervening breaks the causal (and hence probabilistic) connection between the putative cause we are interested in – here, the position of the barometer needle – and its usual causes (atmospheric pressure).

A great advantage of the interventionist account is that, with the notion of an intervention in place, we can define causation in terms of increase in probability without the need to make any stipulations about other (relevant) background factors – unlike the probabilistic accounts considered previously. Interventionist accounts are not, however, reductionist accounts, since the notion of an intervention is itself defined in causal terms: an intervention is, by definition, the 'setting'

of a variable X (e.g., needle position) at a particular value x (e.g., 'rain') in a way that breaks the connection between (or 'screens X off' from) its usual causes.

One Concept of Causation or Two?

We can think of the various accounts of causation surveyed above as falling into two distinct categories. On the one hand, we have process and mechanistic accounts (Section Process and Mechanistic Theories), which take causation to be a matter of what we might call 'production'; on the other, we have counterfactual and probabilistic accounts, which take causation to be a matter of 'difference making.' Other analyses of causation not considered here are generally pretty easy to categorize as either 'production' or 'difference-making' accounts.

To see how these two notions can in principle come apart, consider a 'possible world' where much of what goes on is a result of spells cast by witches and wizards. Let us stipulate that the spells can operate at large spatial and temporal distances, and that they do so directly: there is no process or mechanism that connects the casting of a spell to its implementation. It seems right to say that when Merlin, at midday, casts a spell to turn the prince into a frog at midnight – which he duly does – the former caused the latter. Merlin's spell made a difference to what happened, but there was no process connecting the two; the former did not 'produce' the latter, in just this sense of 'produce' (Schaffer, 2000a). Or consider a less far-fetched scenario: Linda the lifeguard sees a swimmer in trouble and rushes to help, intent on preventing him from drowning. Tanya deliberately trips Linda up as she runs toward the sea, thus preventing her from preventing the swimmer from drowning. The swimmer duly drowns. (Such cases are known as cases of 'double prevention.') It seems plausible to say that Tanya's action is a cause of the swimmer's death; after all, had Tanya not tripped Linda up, the swimmer would have been saved. But again, there is no process connecting the two; we have the situation on the beach involving Linda and Tanya, and the swimmer way out at sea, with nothing linking them together. So again we seem to have difference making without production (Schaffer, 2000b).

It looks as though we can also have production without difference making. There are two routes of the same length to the shop – route A and route B. Which route I take makes no difference to the time and manner in which I arrive at the shop, but I am intent on going to the shop and will definitely take one or other of these routes; in fact, today I take route A. It looks as though my taking route A is not a difference maker: if I had not taken route A, I would have taken the route B, and I would have got to the shops just the same. However, there is clearly a causal process linking my setting out on route A and my subsequent arrival; I put one foot in front of the other, repeated a lot of times, and as a result reached the shop. So we have production without difference making.

Ned Hall (2004) argues that we really have two distinct concepts of causation: the production concepts and the difference-making concept. He diagnoses the repeated failure of various analyses of causation to cover the full range of standard problem cases by noting that analyses fall into either the production category or the difference-making category; so analyses of the former kind will inevitably fail to deal with cases of difference making without production, and analyses of the latter kind will inevitably fail to deal with cases of production without difference making. And, he claims, it is easy to fail to distinguish the two kinds of causation because cases of one without the other are pretty rare; in our ordinary dealings with the world, causes are both difference makers and producers of their effects. (When I sink the black ball in snooker, my taking the shot causes the black ball to drop into the pocket in both senses of 'cause.')

Many philosophers, however, have been reluctant to accept that there is no univocal concept of causation (Lewis, 2004). For example, Dowe (2001), whose Conserved Quantity Theory is clearly a 'production' account, argues that cases of difference making without production are not really cases of causation at all. Meanwhile, defenders of difference-making accounts argue that alleged cases of production without difference making can in fact be dealt with by their accounts. For example, we might respond to the shop case just described by appealing to the idea that causation is context-sensitive, in the sense that causal claims are true or false relative to contrast class (which is normally only implicit). Thus, my taking route A rather than route B really is not a cause of my subsequent arrival, since my taking route A rather than route B made no difference to the time and manner of my arrival. My taking route A to the shops rather than taking no route at all, however, is a cause of my arrival. So we can save the intuitive judgment that my taking route A is a cause of my arrival by finding a context within which the former does make a difference to the latter; hence we do not have a clear counterexample to a broadly difference-making account after all (Menzies, 2007).

Causes and Laws

The above discussion of causation and laws of nature has treated them as separate phenomena, to be given distinct analyses; yet intuitively there is some relationship between, for instance, the law that all metals expand when heated and the fact that a particular piece of metal's expansion is caused by its being heated. How, then, are causation and laws related?

Different theories of causation connect with laws of nature in different ways. Hume's constant conjunction requirement on causation, for example, can be seen as a requirement that causes and effects are lawfully correlated (or 'fall under' a covering law). Laws of nature enter into Lewis's counterfactual analysis of causation indirectly, via his analysis of counterfactuals (see Lewis, 1973b). For Lewis, the truth of 'if A had been the case, B would have been the case' requires that B is true at the 'closest possible world(s)' in which A is true, and closeness of possible worlds depends in part on the extent to which they share the same laws of nature. Dowe's Conserved Quantity Theory of causation (Section Process and Mechanistic Theories above) identifies the causal relation with the transfer of any quantity that is subject to conservation laws, so that, again, facts about causation depend on facts about the laws of nature – although in a rather different way. Heathcote and Armstrong (1991)

argue that the causal relation is identical with *N*, the relation of nomic necessity in virtue of which laws of nature obtain (see Section Realist vs Regularity Accounts of Laws above).

On the other hand, it has been argued that lawful correlation is not necessary for causation. Anscombe (1971), for example, holds that *c* may cause *e* even though *c* and *e* do not fall under any regularity. (Causation of this kind is sometimes referred to as 'singular causation'; but note that 'singular causation' can also be used to refer to particular causation, in contrast to general causation.) Heathcote and Armstrong (1991) claim that even though the causal relation is in fact identical with nomic necessity, this identity does not itself hold as a matter of necessity. Heathcote, Armstong, and Anscombe all agree, then, that 'higgledy-piggledy' worlds are conceivable – that is, possible worlds where there is plenty of causation but where causally related events are not subsumable under any laws (and so under any regularities) at all.

See also: Causal Inference: Overview; Causation: Physical, Mental, and Social; Counterfactual Reasoning, Qualitative: Philosophical Aspects; Counterfactual Reasoning, Quantitative: Philosophical Aspects; Determinism: Social and Economic; Economics, Philosophy of; Empiricism, History of; Free Will and Action; Idealization, Abstraction, and Ideal Types; Joint Action; Logical Positivism and Logical Empiricism; Mental Causation; Natural Law; Probability and Chance: Philosophical Aspects; Scientific Explanation.

Bibliography

Anscombe, G.E.M., 1971. Causality and Determination. Cambridge University Press, London.

Armstrong, D.M., 1983. What Is a Law of Nature? Cambridge University Press, Cambridge.

Armstrong, D.M., 1997. A World of States of Affairs. Cambridge University Press, Cambridge.

Beebee, H., 2000. The non-governing conception of laws of nature. Philosophy and Phenomenological Research 56, 571–594.

Beebee, H., 2006a. Hume on Causation. Routledge, Abingdon.

Beebee, H., 2006b. Does anything hold the universe together? Synthese 149, 509–533.

Beebee, H., 2009. Causation and observation. In: Beebee, H., Hitchcock, C., Menzies, P. (Eds.), The Oxford Handbook of Causation. Oxford University Press, Oxford, pp. 471–497.

Beebee, H., 2011. Necessary connections and the problem of induction. Nous 45, 504–527.

Beebee, H., Mele, A.R., 2002. Humean compatibilism. Mind 111, 201–223.

Berofsky, B., 2012. Nature's Challenge to Free Will. Oxford University Press, New York.

Bird, A., 1998. Dispositions and antidotes. The Philosophical Quarterly 48, 227–234.

Bird, A., 2005. Laws and essences. Ratio 18, 437–461.

Bird, A., 2008. The epistemological argument against Lewis's regularity view of laws. Philosophical Studies 138, 73–89.

Carroll, J.W., 1994. Laws of Nature. Cambridge University Press, Cambridge.

Cartwright, N., 1989. Nature's Capacities and Their Measurement. Cambridge University Press, Cambridge.

Cartwright, N., 2004. Causation: one word, many things. Philosophy of Science 71, 805–819.

Dowe, P., 2000. Physical Causation. Cambridge University Press, Cambridge.

Dowe, P., 2001. A counterfactual theory of prevention and 'causation' by omission. Australasian Journal of Philosophy 79, 216–226.

Dowe, P., 2008. Causal processes. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2008 Edition). http://plato.stanford.edu/archives/fall2008/entries/causation-process (accessed 31.07.13.).

Dupré, J., 1984. Probabilistic causality emancipated. Midwest Studies in Philosophy 9, 169–175.

Dupré, J., 1990. Probabilistic causality: a rejoinder to Ellery Eells. Philosophy of Science 57, 690–698.

Earman, J., Roberts, J., 1999. Ceteris paribus, there is no problem of provisos. Synthese 118, 439–478.

Eells, E., 1991. Probabilistic Causality. Cambridge University Press, Cambridge.

Ellis, B., 2002. The Philosophy of Nature: A Guide to the New Essentialism. Acumen, Chesham.

Glennan, S., 1996. Mechanisms and the nature of causation. Erkenntnis 44, 49–71.

Hall, E.J., 2004. Two concepts of causation. In: Collins, J., Hall, E.J., Paul, L.A. (Eds.), Causation and Counterfactuals. MIT Press, Cambridge, MA.

Hausman, D., 1992. The Separate and Inexact Science of Economics. Cambridge University Press, Cambridge.

Heathcote, A., Armstrong, D.M., 1991. Causes and laws. Noûs 25, 63–73.

Hitchcock, C., 1995. The mishap at Reichenbach fall: singular vs. general causation. Philosophical Studies 78, 257–291.

Hume, D., 1748/1975. Enquiry concerning human understanding. In: Selby-Bigge, L.A., Nidditch, P. (Eds.), Enquiries Concerning Human Understanding and Concerning the Principles of Morals, third ed. Clarendon Press, Oxford.

Lange, M., 2002. Who's afraid of ceteris paribus laws? Or: how I learned to stop worrying and love them. Erkenntnis 52, 407–423.

Lewis, D.K., 1973a. Causation. Journal of Philosophy 70, 556–567.

Lewis, D.K., 1973b. Counterfactuals. Blackwell, Oxford.

Lewis, D.K., 1986. Events, in His Philosophical Papers, vol. II. Blackwell, Oxford pp. 241–269.

Lewis, D.K., 1997. Finkish dispositions. The Philosophical Quarterly 47, 143–158.

Lewis, D.K., 2000. Causation as influence. The Journal of Philosophy 97, 182–197.

Lewis, D.K., 2004. Void and object. In: Collins, J., Hall, E.J., Paul, L.A. (Eds.), Causation and Counterfactuals. MIT Press, Cambridge, MA.

Machamer, P., Darden, L., Craver, C., 2000. Thinking about mechanisms. Philosophy of Science 67, 1–25.

Mellor, D.H., 1995. The Facts of Causation. Routledge, London.

Menzies, P., 1989. Probabilistic causation and causal processes: a critique of Lewis. Philosophy of Science 56, 642–663.

Menzies, P., 1996. Probabilistic causation and the preemption problem. Mind 105, 85–117.

Menzies, P., 1998. How justified are Humean doubts about intrinsic causal links? Communication and Cognition 31, 339–364.

Menzies, P., 2007. Causation in context. In: Corry, R., Price, H. (Eds.), Causation, Physics and the Constitution of Reality: Russell's Republic Revisited. Oxford University Press, Oxford.

Noordhof, P., 1999. Probabilistic causation, preemption, and counterfactuals. Mind 108, 95–125.

Pietroski, P., Rey, R., 1995. When other things aren't equal: saving ceteris paribus laws from vacuity. British Journal for the Philosophy of Science 46, 81–110.

Salmon, W., 1984. Scientific Explanation and the Causal Structure of the World. Princeton University Press, Princeton, NJ.

Schaffer, J., 2000a. Trumping preemption. The Journal of Philosophy 97, 165–181.

Schaffer, J., 2000b. Causation by disconnection. Philosophy of Science 67, 285–300.

Schrenk, M., 2007a. The Metaphysics of Ceteris Paribus Laws. Ontos, Frankfurt.

Schrenk, M., 2007b. Can capacities rescue us from ceteris paribus laws? In: Kistler, M., Gnassounou, B. (Eds.), Dispositions and Causal Powers. Ashgate, Aldershot, pp. 221–247.

Strawson, G., 1989. The Secret Connexion. Clarendon Press, Oxford.

Suppes, P., 1970. A Probabilistic Theory of Causality. North-Holland, Amsterdam.

Swartz, N., 1985. The Concept of Physical Law. Cambridge University Press, New York.

Williamson, J., 2011. Mechanistic theories of causality. Philosophy Compass 6, 421–447.

Woodward, J., 2003. Making Things Happen. Oxford University Press, Oxford.

Woodward, J., 2008. Causation and manipulability. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy (Winter 2008 Edition). http://plato.stanford.edu/archives/win2008/entries/causation-mani (accessed 31.07.13.).

Woodward, J., Hitchcock, C., 2003a. Explanatory generalizations, part I: a counterfactual account. Noûs 37, 1–24.

Woodward, J., Hitchcock, C., 2003b. Explanatory generalizations, part II: plumbing explanatory depth. Noûs 37, 181–199.

Celebrity

Joshua Gamson, University of San Francisco, San Francisco, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Modern celebrity culture emerged as a focus of inquiry and critique in the mid-twentieth century, and has blossomed in recent years into a vibrant field of study. Critics have pointed to the role of celebrities as mass-produced commodities and celebrity as ideological support for consumer capitalism, and to the disconnection between notoriety and merit. Scholars have approached celebrities as symbolic entities, mediating a tension between egalitarian and aristocratic cultural tropes, and embodying cultural anxieties about the relationship between media images and lived realities; documented the economic, social, and political structures in which celebrity has been produced, noting the significance of the Hollywood star system, and the recent transformations brought by reality television and new media technologies; and studied celebrity fans and fandom, shedding light on forms of fan engagement with celebrity culture, fan–celebrity encounters, and the uses of celebrity images by marginalized groups.

In popular usage, 'celebrity' typically refers simply to the famous or notorious person, whose existence and activities are known to an audience that is unknown to him or her personally, or to the condition of such notoriety. Theory and research on the topic, while often focusing also on individual celebrities, have tended to work from a broader definition of 'celebrity' (or, occasionally, 'celebrityhood' or 'celebritydom'): as a social and cultural phenomenon, in which large-scale knownness and public visibility become important sources, measures, and currencies of social, economic, and sometimes political standing. Celebrities are generally defined as distinctly contemporary versions of famous people, and celebrity a distinctly contemporary form of fame. Their distinction is found in their relationship to modern, industrialized, commercial communications media, which makes it possible to achieve notoriety without benefit of extraordinary action or achievement. The celebrity is thus famous for the media-disseminated image of himself or herself; in Daniel Boorstin's (1961) well-known phrasing, he or she is "well known for his [or her] well knownness." Celebrities are, moreover, a specific kind of social elite or status group, characterized by both extraordinary privilege and uncertain boundaries, and whose members occupy positions that are both socially central and easily lost (Kurzman et al., 2007; see also Milner, 2005). They are, as Italian sociologist Francesco Alberoni (1972) put it, a 'powerless elite' – or more accurately, an elite with high status and visibility but limited institutional power – commanding a high level of interest unrelated to the consequences of their activity. The study of celebrity, which has emerged within both humanities and social science disciplines, has run along several overlapping tracks. Beginning with critics of mass culture, analysis has focused on the distinctive characteristics and symbolic functions of celebrity in contemporary Western societies; more structurally focused studies have considered the political, economic, and social organization of celebrity, especially as it has become industrialized in the entertainment business, spread across a variety of institutional settings, and undergone changes due to new media genres and technologies; and, more recently, researchers have investigated the interpretations, uses, pleasures, and sources of engagement for audiences or fans of celebrities and celebrity.

Celebrity and Critiques of Mass Culture

The term 'celebrity' in its modern meanings began to be used in the nineteenth century, but the study of the phenomenon began in earnest with the rise of mass-produced culture, and in particular with the elaboration of an industrialized Hollywood film 'star system' in the early decades of the twentieth century. (For an excellent history of fame and fame discourse from ancient times to the near-present, see Braudy, 1986.) It first emerged as a sustained focus of inquiry through mid-twentieth-century criticism of mass culture, from both the left and the right. Although these works were rarely based on empirical research, and were filled with unsupported assertions about those creating and receiving celebrity images, they called attention to the mass production and management of celebrities, and to the question of the social impact of industrialized celebrity culture.

The Celebrity and Capitalist Ideology

Early perspectives on celebrity were largely theoretical, emerging as part of Marxist-influenced Frankfurt School cultural criticism of 'mass culture,' 'mass society,' and the 'culture industry.' Celebrities were seen as mass-produced, standardized commodities posing as unique human individuals, and celebrity discourse as a major ideological support beam for consumer capitalism. In the 1940s, Theodor Adorno and Max Horkheimer (1977), for instance, saw celebrities as the products of 'the culture industry,' the cultural apparatus of mass society; Hollywood stars serve as distractions from the dissatisfactions created by industrial capitalism, and to manipulate 'the masses' into capitalism's false promises of both choice (standardized, mass-produced celebrities appear to be different individuals) and universal success (celebrities appear to demonstrate the rewards available to all).

Leo Lowenthal (1968), also writing in the 1940s, researched changes in 'mass idols' in popular magazines, charting the move from 'idols of production' (business and politics) to 'idols of consumption' (entertainment and sports); he too suggested that these popular culture heroes perpetuated the myth of an open social system, such that the existing social

system is celebrated along with the star. C. Wright Mills (1956: p. 71) wrote in the 1950s of the professional celebrity as a summary of American capitalist society's promotion of competition and winning; as "the crowning result of the star system in a society that makes a fetish of competition," the celebrity shows that rewards go to those who win, regardless of the content of the competition.

The definition of celebrities as mass-produced distractions, and their ideological role in promoting consumption, competition, individualism, and the myth of open opportunity, has continued in much contemporary cultural criticism and analysis. For example, P. David Marshall (1997: p. x), has argued that the celebrity, as a "public individual who participates openly as a marketable commodity," embodies and legitimates the linked ideologies of market capitalism and individualism; Chris Rojek (2001: p. 90) has located celebrities within 'the culture of distraction today,' as human commodities who fill the void left by 'the death of God and the decline of the church'; Ellis Cashmore (2006: p. 269) has concluded that "celebrity culture's most basic imperative is material: it encourages consumption at every level of society"; and Karen Sternheimer (2011: p. 24) has suggested that celebrity culture "reflects and reinforces the ever changing notion of what it means to achieve the American Dream."

Celebrity vs Heroism

For more conservative cultural critics from the 1950s onward, the role of the celebrity system as an ideological support for capitalism was less important than its reflection of a large-scale disconnection between notoriety and merit. Such a view crystallized in the 1960s with the publication of Daniel Boorstin's *The Image* (1961), which distinguished celebrity from heroism. In an argument that presaged more recent postmodernist theory on 'simulation and simulacra' and the implosion of artifice and reality (see Baudrillard, 1988), Boorstin argued that, with the growth of mass media, public relations, and electronic communication, it was possible to produce fame without any necessary relationship to outstanding action or achievement. Thus, the hero, whose fame is the result of distinctive action or exceptional, meritorious character, has been superseded by the celebrity, whose notoriety is manufactured by mass media without regard for character or achievement; the signs of greatness are mistaken for its presence. In Boorstin's definition, the celebrity is a 'human pseudoevent.' The phenomenon of celebrity is a symptom of a media-driven culture in which artifice has displaced reality, and in which merit and attention have become uncoupled. Although as Leo Braudy (1986) has shown, the oppositions between pure, 'real' fame and inauthentic 'artificial' celebrity do not fully hold up – historically, fame and merit have never been firmly and exclusively coupled – the conservative critical approach to celebrities as false, vulgarized heroes has pointed toward historically new features. Modern media, through the increasingly sophisticated creation, management, and reproduction of images, have an unprecedented capacity to place a person on the cultural radar screen, quickly and with no necessary reliance on the person's publicly celebrated actions or character. Thus the modern celebrity system does not require of celebrities that they be extraordinary in terms of achievement, talent, or character – some celebrity is 'acquired' and some 'achieved' (Rojek, 2001: pp. 16–18) – only that they provide engaging narratives, starring in what Neal Gabler (2001: p. 5) calls the 'life movie.'

Celebrities as Symbolic Entities

The interest shown by cultural critics in what the workings of contemporary celebrity tell us about the culture that makes it so central has been taken up in many humanities-based approaches to stars and stardom, which tend to consider the symbolic activity that takes place in and through celebrity discourse. Who gets attention, the logic goes, tells us much about the core values, or ideological contradictions, of the society giving the attention. In addition to the theme, noted above, that celebrities serve as ideological supports for capitalism and consumerism, two other themes have been particularly pervasive: the tension between egalitarian and aristocratic cultural strands; and the pursuit of the authentic self.

Celebrities as a Democratic Aristocracy

One striking feature of contemporary Western celebrity discourse is the way celebrities are treated to a cultural status that is simultaneously 'above' the rest of the populace and 'of' that populace. Celebrities are culturally constructed as a sort of elected aristocracy, both elevated and brought down by the watching crowds; the celebrity has become one symbolic means through which the population of the unfamous declares its own power to shape the public sphere (Marshall, 1997). Moreover, while celebrity culture certifies some people as more deserving of attention and rewards because of their difference from the rest of the population, it also continually demonstrates that such people are ordinary, just like everyone else (Braudy, 1986). Thus, popular celebrity discourse embodies ambivalence about hierarchy in Western democracies: celebrities are celebrated for being better than, but no better than, those who watch them.

The Search for the 'Real' Self

A second outstanding feature of contemporary celebrity discourse is the thematic emphasis on getting 'behind' celebrity images to the 'true' or 'real' self. Celebrity discourse, as Richard Dyer (1991: p. 135) has demonstrated, involves a 'rhetoric of authenticity': the question of what a celebrity is 'really like,' what kind of self actually resides behind the celebrity image, is a constant, whether in the form of tabloid exposés, behind-the-scenes reporting, celebrity profiles, or fan activities such as autograph seeking. In part, this is because celebrities have the unique characteristic of appearing to audiences only in media texts, while also living in the world as actual human beings – they are images, but are "carried in the person of people who do go on living away from their appearances in the media" (Dyer, 1991: p. 135). In part, the theme of realness is the result of the increasing visibility over time of celebrity production mechanisms, raising the question of whether the celebrity image has been manufactured to attract an audience, or

whether it reflects a true, deserving self (Gamson, 1994). Celebrity discourse, with its heavy rhetorical emphasis on authenticity, thus manifests a larger cultural anxiety about the relationship between media images and lived realities. This dynamic has been reworked and advanced in recent years through the genre of 'reality television,' in which celebrities purportedly show that they are real through the filming of their 'everyday lives,' and noncelebrities become famous for living their 'real lives' in front of the cameras (Grindstaff, 2010). Here, the exposure of the ordinary self – offered as a demonstration of authenticity – itself becomes a means to celebrity.

The Political Economy of Celebrity

While many of the attempts to grapple with the unique symbolic or ideological features of contemporary celebrity have been either entirely speculative or based exclusively on textual analysis, much of the empirical research on the topic has focused on celebrity as an economic and social system. Influenced by the strategies of political economists and organizational sociologists, this research has investigated not so much the cultural meaning of celebrity as the internal organization and economic logic of the celebrity system. In contrast to approaches that assume that film stars are popularly selected for attention, for instance, such analysts tend to see celebrity as the result of "the exigencies of controlling the production and marketing of films" (King, 1986: p. 155). Although celebrities increasingly emerge in other social domains (politics, academia, etc.), most attention has been given to the major celebrity production center, the entertainment industry (Currid-Halkett, 2010).

The Hollywood Star System and Beyond

The pursuit of celebrity, especially in the entertainment business, became highly routinized, rationalized, and industrialized over the course of the twentieth century, with the development of industries, such as public relations, specifically devoted to the generation and management of public visibility. Celebrities are, in this context, marketing tools. In the notoriously risky entertainment business, which requires high capital investment for most of its products, a star is an insurance policy against audience disinterest, used primarily to minimize the risk of financial loss. Thus, star images are typically managed in accordance with the needs of the financiers of the vehicle with which a celebrity is associated – with film stars, for instance, a movie studio. The key nexus is not so much the celebrity and his or her audience, but the celebrity's backers, who pursue publicity, and journalists, editors, and producers, who provide it (Gamson, 1994).

The structure of the star system has changed significantly over its brief lifetime. The early studio star system involved tight control of the production, exhibition, and distribution of films and their associated film-star images by several major studios; stars were under studio contracts, and studio publicity operations were responsible for producing and disseminating celebrity stories and images. When the studio oligopoly was broken up by a US Supreme Court decision in the 1950s, many more parties with a financial interest in celebrities' careers became involved in the management of celebrity images – personal publicists, managers, agents, in addition to the celebrity himself or herself, joined studio publicists in battling for control of the process. With the growth of television since the 1950s, the explosion of celebrity-driven media outlets since the 1970s, and the development and expansion of reality television since the 1980s and the Internet since the 1990s, moreover, it has become both easier to build celebrity and more difficult to retain it – hence Andy Warhol's famous declaration that eventually everyone will be world famous for 15 minutes. While Hollywood movie studios still generate a large proportion of American and international celebrities, celebrity has become less centralized, and the logic of celebrity has taken hold within a wider range of social spheres, including the worlds of literature (Moran, 2000), sports (Smart, 2005), politics (West and Orman, 2002), art, business, and academia, each with its own somewhat distinction celebrity production practices (Turner, 2004).

Celebrity Production in the Twenty-First Century

Recent history has seen some major changes in celebrity production apparatuses. Reality television, which developed in the 1980s in response to changing economic conditions in the television industry as a cheaper, quicker alternative to scripted programming (Collins, 2008), transformed celebrity production: no specialized training or prior experience was needed to enter the celebrity field, and celebrity became 'an outcome of a programming strategy' (Turner, 2004: p. 53). In addition, although in many ways the Internet simply extended the reach of existing entertainment industry organizations, it has also rapidly changed the dynamics of celebrity production, especially through Web 2.0 phenomena such as YouTube, Myspace, Twitter, and Facebook (Gamson, 2011; see also Rojek, 2012). As "the tools of self-publicity are increasingly available to ordinary people" (Bennett and Holmes, 2010: p. 76), barriers to entry to celebrity are reduced, and access to potential audiences does not require industry gatekeepers, celebrity production has become partly autonomous from the centralized, tightly controlled celebrity industry. The result of these changes was a large influx of 'civilians' into the celebrity field; forms of celebrity that are more fleeting, dispensable, and difficult to sustain than earlier ones (Collins, 2008; Currid-Halkett, 2010; Rojek, 2012), and the emergence of new celebrity types, such as the anticelebrity viral star (celebrated for being unlike conventional celebrities), the do-it-yourself celebrity (who has pursued fame outside the established celebrity system), and the microcelebrity (famous to a small community of fans) (Gamson, 2011).

Fandom and the Reception of Celebrity

Analyses of the social and economic organization of celebrity tend to bracket questions of its cultural meanings, and textual analyses of celebrity tend to operate with untested assumptions or assertions about the meaning and impact of celebrity for audiences. Methodologically, both have tended to exclude empirical research into the meaning of celebrities and celebrity in the everyday lives of the fans or audiences encountering

them. As the study of culture in general began, in the 1980s, to take more of a methodological turn toward audience research, audience-related aspects of celebrity have also come more sharply into focus. This territory remains, however, under-investigated, in part because the study of audiences is both methodologically cumbersome and costly.

Considerable thought, and a small but growing body of research, has been devoted to the question of the fans' or audiences' relationship to celebrities. One early theory, for instance, proposed that mass media such as television facilitate a 'parasocial relationship' between performers and audience members, in which the spectator comes to relate to the celebrity as if they were in a face-to-face relationship, with the 'illusion of intimacy' (Horton and Wohl, 1956). Since then, largely through psychoanalytic film theory, discussions have focused on the processes by which audience members identify with celebrities, especially film stars (Tudor, 1974; Stacey, 1991). Various typologies of identification have been set forth, emphasizing quite a range of activities, uses, and types of attachment developed by audiences in their encounters with celebrities (Marshall, 1997), with particular attention to celebrity gossip activities as 'fantasies of belonging' to a 'moral community' (Hermes, 1995: p. 132).

This has been especially important in challenging the assumption that 'the audience' for celebrity is a homogenous mass, and that celebrities mean the same thing for all of its members. Parasocial identification and celebrity hero worship, for instance, both appear to be common stances; but ironic, playful, and irreverent interpretations of celebrity images also appear to be prevalent, especially as the pursuit of celebrity itself has become a more common focus of public discussion (Gamson, 1994). Particular attention has been paid to the ways groups marginalized on the basis of race, class, gender, sexuality, ethnicity, nationality, or age make use of celebrity images for their own purposes: building group solidarity or expressing rebelliousness and alienation, for example, through the celebration of particular types of stars (strong women, for example, or stars from their own ethnic group) (Stacey, 1991). Finally, recent empirical work has detailed the meanings made by fans of particular celebrities (Vannini, 2004), celebrity fandom online (Soukup, 2006), teenagers' attitudes toward fame (Halpern, 2008), and interactional dynamics of encounters between celebrities and fans (Ferris and Harris, 2011).

The Future of Celebrity

While there is considerable consensus on certain distinctive features of contemporary celebrity, including its recent transformations, quite a bit of room remains for empirical investigations of the contours of celebrity as a socially organized system of meaning, status, identification, and pleasure; even with the growth of the 'celebrity studies' field in recent years, theory and speculation on the phenomenon remains more developed than empirical research. Considerable room remains, in fact, for a synthesis of the theoretical conceptualizations and explanations of celebrity culture, and subsequently for theory-driven empirical research that tests and refines those theories. Several areas in particular are likely to prove fruitful for future mixed-methods research. First, audience research can continue not only to document the range and types of audience positions vis-à-vis celebrities, but also to examine various possible explanations for the variance – audience characteristics, for instance, or qualities of different celebrity domains, genres, and types. Second, the process by which the logic of celebrity can and does spread to spheres other than entertainment, and how it may operate differently in those realms, is still not well understood. More systematic panel-data allowing comparisons across institutional domains are needed. Third, it remains the case that nearly all of the literature on celebrity has emerged from, and focused on, the United States and the United Kingdom. Cross-cultural comparative research is a very promising, and almost entirely untapped, source of insight into the cultural, economic, political, and social logic of celebrity. Finally, while it is clear that celebrity production has been transformed by digital technologies and new media – and presumably along with it the experience of celebrity for those encountering it – the contours and impact of these changes have yet to be fully documented, studied, and understood.

See also: Audiences, Media; Entertainment; Film and Video Industry; Hegemony and Cultural Resistance; Journalism; Mass Media and Cultural Identity; Mass Media, Political Economy of; Mass Media, Representations In; Media Effects; Public Sphere and the Media; Reputation; Social Media; Television: General; Television: History.

Bibliography

Adorno, T., Horkheimer, M., 1977. The culture industry: enlightenment as mass deception. In: Curran, J., Gurevitch, M., Woolacott, J. (Eds.), Mass Communication and Society. Sage Publications, Beverly Hills, CA, pp. 349–383.

Alberoni, F., 1972. The powerless 'elite': theory and sociological research on the phenomenon of the stars. In: McQuail, D. (Ed.), Sociology of Mass Communications. Penguin Books, Harmondsworth, UK, pp. 75–98.

Baudrillard, J., 1988. Selected Writings. Stanford University Press, Stanford, CA.

Bennett, J., Holmes, S., 2010. The 'place' of television in celebrity studies. Celebrity Studies 1, 65–80.

Boorstin, D.J., 1961. The Image: A Guide to Pseudo-events in America. Harper & Row, New York.

Braudy, L., 1986. The Frenzy of Renown: Fame and Its History. Oxford University Press, New York.

Cashmore, E., 2006. Celebrity Culture. Routledge, London.

Collins, S., 2008. Making the most out of 15 minutes: reality TV's dispensable celebrity. Television and New Media 9, 87–110.

Currid-Halkett, E., 2010. Starstruck: The Business of Celebrity. Faber & Faber, New York.

Dyer, R., 1991. A star is born and the construction of authenticity. In: Gledhill, C. (Ed.), Stardom: Industry of Desire. Routledge, London, pp. 132–140.

Ferris, K., Harris, S., 2011. Stargazing: Celebrity, Fame, and Social Interaction. Routledge, New York.

Gabler, N., 2001. Toward a New Definition of Celebrity. Norman Lear Center, University of Southern California Annenberg School of Communication, Los Angeles.

Gamson, J., 1994. Claims to Fame: Celebrity in Contemporary America. University of California Press, Berkeley, CA.

Gamson, J., 2011. The unwatched life is not worth living: the elevation of the ordinary in celebrity culture. Publications of the Modern Language Association 126, 1061–1069.

Grindstaff, L., 2010. Just be yourself—only more so: ordinary celebrity in the era of self-service television. In: Kraidy, M., Sender, K. (Eds.), The Politics of Reality Television: Global Perspectives. Routledge, New York, pp. 44–58.

Halpern, J., 2008. Fame Junkies: The Hidden Truths behind America's Favorite Addiction. Mariner Books, Boston, MA.

Hermes, J., 1995. Reading Women's Magazines: An Analysis of Everyday Media Use. Polity, Cambridge.
Horton, D., Wohl, R.R., 1956. Mass communication and para-social interaction: observations on intimacy at a distance. Psychiatry 19, 215–229.
King, B., 1986. Stardom as an occupation. In: Kerr, P. (Ed.), The Hollywood Film Industry. Routledge and Kegan Paul, London, pp. 154–184.
Kurzman, C., Anderson, C., Key, C., Lee, Y., Moloney, M., Silver, A., Van Ryn, M., 2007. Celebrity status. Sociological Theory 25, 347–367.
Lowenthal, L., 1968. The triumph of mass idols. In: Literature, Popular Culture, and Society. Pacific Books, Palo Alto, CA, pp. 109–136.
Marshall, P.D., 1997. Celebrity and Power. University of Minnesota Press, Minneapolis, MN.
Milner, M., 2005. Celebrity culture as a status system. Hedgehog Review 7, 66–77.
Mills, C.W., 1956. The Power Elite. Oxford University Press, London.
Moran, J., 2000. Star Authors: Literary Celebrity in America. Pluto Press, London.
Rojek, C., 2001. Celebrity. Reaktion, London.
Rojek, C., 2012. Fame Attack: The Inflation of Celebrity and Its Consequences. Bloomsbury USA, New York.
Smart, B., 2005. The Sport Star. Sage, London.
Soukup, C., 2006. Hitching a ride on a star: celebrity, fandom, and identification on the world wide web. Southern Communication Journal 71, 319–337.
Stacey, J., 1991. Feminine fascinations: forms of identification in star-audience relations. In: Gledhill, C. (Ed.), Stardom: Industry of Desire. Routledge, London, pp. 141–166.
Sternheimer, K., 2011. Celebrity Culture and the American Dream. Routledge, New York.
Tudor, A., 1974. Image and Influence: Studies in the Sociology of Film. St. Martin's Press, New York.
Turner, G., 2004. Understanding Celebrity. Sage, London.
Vannini, P., 2004. The meanings of a star: interpreting pop music fans' reviews. Symbolic Interaction 27, 47–69.
West, D., Orman, J., 2002. Celebrity Politics. Prentice Hall, Upper Saddle River, NJ.

Censorship in Research and Scholarship

Nicholas G Evans and Michael J Selgelid, Centre for Human Bioethics, Monash University, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Censorship involves the suppression of communication that may be valuable or harmful to the censor, the individual or group censored, or the intended recipient of the communication. Researchers and scholars have been censored throughout history by religious, political, military, and corporate powers. During the twenty-first century, much debate has surrounded questions about the potential censorship of 'dual-use' biological research that could enable production and use of biological weapons. Whether or not censorship would be ethically appropriate in this contemporary context depends on how trade-offs should be made between the values of free communication and harm prevention.

Censorship involves the intentional suppression of communication in order to prevent dissemination of information that may be valuable or harmful to the censor, the individual or group censored, or the intended recipient of the communication. Within debates about censorship one finds a range of definitions, which may be so expansive as to include any act of restricting information flow, or so narrow that only the suppression of expression by the state or one of its proxies is encompassed. Broad definitions will include – and narrow definitions will exclude – what should arguably be considered gray-area cases of censorship, such as legislation not aiming at suppression of speech that nonetheless causes individuals to suppress themselves through fear of legal action (Schauer, 1978). The diversity of definitions has lead some to suggest that censorship falls on a spectrum with other acts that restrict speech or are otherwise more or less *like* censorship (Alexander, 2005).

What binds these definitions together is a set of exemplar cases that inform a well-understood paradigm of censorship. Paradigmatic censors include government authorities, militaries, and religious institutions; more recently, corporations have engaged in censorship and self-censorship (that is, suppressing their own communication) to protect their interests. Researchers and scholars are often vulnerable to censorship by these powerful actors, whose power depends on a particular orthodoxy. The history of research and scholarship is thus replete with examples involving the abuse of power in the suppression of work that challenges established knowledge or practice.

Cases of Censorship: History, Methods, and Motivations

The censorship of research and scholarship has a long history, but its early roots lie in the suppression of the dissemination of knowledge by religious authorities. Charges of heresy – the crime of speaking contrary to church orthodoxy – have been used to justify censorship of priests and laypersons for hundreds of years. A prominent example of this is that of Peter Abelard, a fourteenth-century Catholic scholar accused of heresy for attempting to interpret the Catholic Trinitarian doctrine along rationalistic lines. Abelard, found guilty of this alleged heresy, was required to burn his works and forbidden to preach (Godman, 2000). Paradigmatic of religious censorship, Abelard's trial also demonstrates some of the iconic means of achieving censorship: legal sanction, destruction of works, and formal silencing under threat of further punishment.

The Catholic Church is best known for its censorship of works listed on the *Index Prohibitum Librorum*, or 'Index of Prohibited Books,' including those of Denis Diderot, John Stuart Mill, and George Berkeley. The works of Nicolaus Copernicus and Galileo Galilei, whose research spurred the scientific revolution, also once occupied the list. Copernicus' *De Revolutionibus Orbium Coelestium* (On the Revolutions of the Heavenly Spheres) forwarded the claim that the earth orbited the sun, contesting the Aristotelian cosmology of the day, that situated the earth as the immobile center of the universe. *De Revolutionibus* was initially published openly. The controversy surrounding *De Revolutionibus* was initially not forthcoming, and the work was published openly. The reason for the lack of controversy surrounding Copernicus' *De Revolutionibus* was partly due to the disclaimer provided in the preface by Osiander, Copernicus' student, added after his teacher's death. While Osiander's motives remain unknown, his claim that *De Revolutionibus* was "not put forward to convince anyone that [Copernicus' theories] are true, but merely to provide a reliable basis for calculation," effectively silenced the intent of Copernicus' work. By subverting the intent of Copernicus' work–the preface was not signed, so it appeared to be Copernicus' own thoughts on his work–Osiander accomplished something that might well be considered equivalent to censorship in outcome (Henry, 2012).

De Revolutionibus was finally censored by the Church in 1616, following the release of Galileo's *Sidereus Nuncius* (The Sidereal Messenger). *Sidereus Nuncius* provided… a further challenge to Aristotelian cosmology through Galileo's telescopic observations of a changing, imperfect cosmos, including a cratered moon and impermanent, moving stars (Tolbert, 2003). While on its own, *De Revolutionibus* was initially deemed harmless; but when placed alongside *Sidereus Nuncius* it became a threat to Church doctrine. Works subjected to censorship may only demonstrate the qualities that make them targets of censorship years – even decades – after the fact.

Galileo's *Dialogo sopra i due massimi sistemi del mondo* (Dialogue Concerning the Two Chief World Systems), however, caused controversy almost immediately. In *Dialogo*,

Galileo defended his heliocentric conception of the solar system, and presented a strong challenge to the orthodox view that placed the earth – and thus man – at the center of the cosmos. *Dialogo* was initially well received by Pope Urban, with one caveat: Urban requested that Galileo include a disclaimer on the subject of the reciprocal motion of the tides, which Galileo (wrongly) believed was due to the movement of the earth around the sun. Urban noted that although Galileo's argument was convincing, it was still possible that God had designed the system of the tides to appear that way, while still being controlled by some other, divine means. Galileo did incorporate these remarks; but by presenting the Pope's views through the voice of the dullard Simplicius, he insulted the Church, bringing about his censorship and house arrest (Miller, 2008).

Orthodoxy may be central to censorship motivated by religious conviction, but the political process can play a strong part in the practice of censorship. The Butler Act of 1925, for example, is an example of censorship in which the State Government of Tennessee passed a bill that rendered it unlawful to teach evolution, or any other material that denied the Biblical account of the origin of humans and sought to account for human existence as ancestors of 'lower' animals (House Bill 185, 1925). The Butler Act was motivated by religious belief, and a desire to deny evolution in favor of creationist arguments, but relied on legal means to deter teachers from disseminating information. As with the trial of Abelard, legal injunction, or its threat, is often a powerful method of censorship. In case of the Butler Act, it is also important to note that preventing dissemination of scholarship by censoring teachers is a powerful way to prevent research or scholarship from achieving recognition in the public eye.

The persecution of Soviet scientists by Trofim Lysenko, director of the Institute of Genetics within the Soviet Academy of Sciences from the 1940s through to the 1960s, is one of the most horrific examples of modern censorship of science. Lysenko advocated a theory of heredity that posited that organisms inherited traits from the surrounding environment, in contrast with the Mendel–Morgan gene theory that supported the idea that the transmission of genetic information from parents to offspring influenced the traits of an organism. As director of the Academy of Sciences, and friend of Joseph Stalin, Lysenko embarked on a campaign to ensure the supremacy of his theory by imprisoning those who held competing theories. Lysenko's reasoning, at the time, was that other theories weakened the Soviet State, and threatened Soviet agriculture. During his reign, Lysenko imprisoned and killed thousands of scientists who held theories of genetics that conflicted with his own. Lysenko's actions are believed to have set back Soviet biology by decades, demonstrating the extent of the harms that can be inflicted through the abuse of power by a censor (Joravsky, 1986).

Militaries have often sought to censor research that might provide an adversary with some advantage: as early as 1792, the French military restricted access to technology and research on improving artillery (Gillispie, 1992). In World War II, science and technology became deciding factors in the war effort, and the censorship of research became part of both Allied and Axis strategy. The Manhattan Project, which led to the development of the first two atomic weapons, was perhaps the ultimate example of this practice: an entire community of scientists and their families were veiled behind a wall of secrecy and censorship in the name of a wartime scientific endeavor.

Following World War II, the control of science and technology to achieve nuclear supremacy led to the *Atomic Energy Act* in the United States. The Act presumptively restricted access to "(1) [the] design, manufacture, or utilization of atomic weapons; (2) the production of special nuclear material [uranium-233, uranium-235 or plutonium in their natural or enriched states]; or (3) the use of special nuclear material in the production of energy" (AEA, 1946: pp. 1–15). This policy, known as the 'born classified' doctrine, is almost unique in liberal Western democracies, and is certainly the only one of its kind in the censorship of science (Quist, 1994).

Science is now big business, and companies with vested interests may engage in censorship and self-censorship in order to maintain their advantage in the market or prevent damage to their reputations. The tobacco industry, for example, has engaged in a sustained campaign to obfuscate, avoid, and outright suppress research on the health risks of smoking. Phillip Morris and British American Tobacco, for example, engaged in the practice of self-censorship of research associated with the health risks of smoking, and of second-hand smoke. Lawyers employed by these companies would review both in-house research and externally commissioned research for its potential to cast smoking in a negative light. Results and publications that posed such risks were rewritten to provide weaker or alternate conclusions, and at times were completely suppressed. By placing themselves in a position where they could censor research on the risks of smoking, tobacco companies were able to control and suppress information that might show smoking to be harmful, thereby manipulating scientific consensus and policy-making (Glantz et al., 1996).

Current Issues: The Dual-Use Dilemma

In the twenty-first century, the censorship of research has been suggested as a possible regulatory strategy in addressing the *dual-use dilemma*, which arises when one and the same piece of research can be used to benefit or harm humanity. A paradigmatic case of dual-use research occurred in 2001, when Australian virologists modified the mousepox virus genome. The (unexpected) outcome of this research was the creation of a virus that killed mice that were genetically resistant to, and mice that had been vaccinated against, ordinary mousepox. Mousepox is genetically similar to smallpox, a disease that killed hundreds of millions of people in the early twentieth century; and, supposedly, more people than any other infectious disease in recorded history. Critics fear that publication of the mousepox study might enable bioterrorists to create and unleash a strain of vaccine resistant smallpox (Miller and Selgelid, 2008).

A 2004 report by the National Academies of Science, *Biological Research in an Age of Terrorism*, considered the possibility of censoring dual-use research posing serious risk of misuse. The report (known as the 'Fink report' after its chairperson, Gerald Fink) concluded that any suppression of potentially

dangerous research results should be left to voluntary self-governance of scientists (NRC, 2004). On the advice of the Fink report, the National Science Advisory Board for Biosecurity (NSABB) was created with the aim of providing guidance on the governance of dual-use research. In June 2007, the NSABB released its initial guidance, including a 'responsible communication' plan for federally funded research with dual-use potential. This plan, in line with the Fink report, relied on self-governance of scientists and, in particular, the discretion of principal investigators (NSABB, 2007).

The model of self-governance has, however, been challenged. In 2005 two researchers at Stanford University modeled a terrorist attack distributing botulinum toxin through a cow-to-consumer milk supply, using parameters taken from the United "States" milk supply. The study demonstrates how 10 g of botulinum toxin, distributed prepasteurization, could result in the death of 200 000 people. These deaths would be concentrated, the researchers argued, among child populations, as milk is distributed regularly outside of normal channels to schools in the United States, and the youth and low body mass of children makes them more susceptible to botulinum toxin poisoning (Wein, 2009: p. 803).

The dual-use potential of this research prompted the Department of Health and Human Services to request the *Proceedings of the National Academies of Science* (PNAS) to withhold publication of the study (Simonson, 2005; Wein and Liu, 2005: pp. 9987–9988). The PNAS embargoed the article for a week, before publishing it with an accompanying editorial by Dr Bruce Alberts, editor in chief of the PNAS, claiming that the scientific process required openness to improve human health, and that the research should be published to better assist in this aim (Alberts, 2005).

In 2011, Ron Fouchier, a virologist from Erasmus University in the Netherlands, announced that he had successfully created a strain of H5N1 avian influenza that was transmissible between ferrets inhabiting the same cage. Ferrets are a well-known analogue for human immune responses to influenza viruses, and their proximity showed that this variant was airborne transmissible in mammals – something not seen in wild H5N1. Fouchier – together with Yoshihiro Kawaoka of the University Wisconsin, who completed similar research that same year – claimed that this research showed that bird flu, a virus with a reported fatality rate of up to 60% in humans, could become transmissible between mammals – a recipe for a potentially devastating pandemic.

The purported benefits of this H5N1 influenza research include raising awareness about the risks of emerging influenza viruses, assisting in disease surveillance, and advancing influenza research (e.g., vaccine development). Yet concerns about the potential danger of this research, either through accidental or intentional release of the engineered virus, prompted the NSABB to recommend the partial censorship of the research (i.e., via omission of detailed description of materials and methods). However, the NSABB later recanted in response to revisions (providing further explanation of the purported public health benefits of the studies) made to the submitted papers by the researchers (Evans, 2013).

The NSABB's policies were reevaluated in 2012 in response to the H5N1 avian influenza studies. The United States Government released new policy regarding the regulation of state-sponsored dual-use research of concern, outlining an assessment procedure for use in assessing the dual-use character of government-sponsored research and determining the capacity research possesses for misuse. In cases where the risks posed by the research cannot adequately be mitigated, federal departments and agencies were instructed to determine whether it was appropriate to:

1. Request voluntary redaction of the research publications or communications.
2. Classify the research:
   a. In accordance with National Security Decision Directive/NSDD-189, departments and agencies will make classification determinations within the scope of their classification authorities and appropriate classification guidelines or may consult with other departments and agencies to make these determinations.
   b. Departments and agencies may consider whether to refer classified research to another department or agency for funding.
3. Not provide or terminate research funding (Office of Biotechnology Activities, 2012).

These regulations represent the first known policy that actively makes possible the censorship of dual-use life sciences research not already restricted by virtue of being classified research, as in the case of nuclear research.

The Ethics of Censoring Dual-Use Research

Dual-use research provides a backdrop from which to understand the ethics of censoring research and scholarship. Censorship is, on the face of it, a violation of an individual's right to freedom of speech (which tends to incorporate publishing, among other forms of communication), widely recognized and justified through appeal to a plural set of values. Protecting freedom of speech allows us to communicate, negotiate, and satisfy our preferences; and speech can be desirable in itself (i.e., free speech promotes utility). Speech is an exercise in individual self-determination or autonomy. Autonomy is valuable to an individual's quality of life, and a self-governing citizenry is important to a well-functioning liberal democratic society. Protecting speech assists in securing equality between individuals by ensuring everyone has sufficient opportunity to participate within a society. Finally, preventing interference in the activities of individuals necessarily promotes freedom, which may be seen as a good in itself (Alexander, 2005).

Yet few commentators hold that all speech promotes these values (and is thus deserving of protection as a matter of right); moreover, at times the right to freedom of speech may be outweighed by competing considerations even if speech is still valuable (other things being equal) (Waldron, 2010). Even the most extreme of libertarian thinkers, such as Robert Nozick, acknowledge that there are some (exceptional) cases where rights might be outweighed by competing considerations (Nozick: p. 30). The right to freedom of speech is valuable, but it is not always the only or most valuable thing to promote. In cases of conflict, a balance should be struck between free speech and other values (Selgelid, 2007).

Censorship is commonly considered permissible in cases where the dissemination of information may lead to immediate and/or catastrophic harm. Dual-use research, in particular, raises this possibility. The concern that the information disseminated in scientific publication will enable acts of bioterrorism lies at the heart of the dual-use dilemma in the biological sciences. In the mousepox example above, the concern was the possibility of similar modification of the smallpox virus, which in its wild form killed hundreds of millions of people during the twentieth century. The botulinum toxin studies modeled an attack that could kill approximately 200 000 people, using relatively small amounts of the toxin. In the case of the H5N1 avian influenza studies, wild H5N1 has a case-fatality rate of 60% in humans. Modifying the virus so that it was transmissible between humans could cause a catastrophic loss of life – the 1918 'Spanish flu' pandemic killed somewhere between 20 and 100 million people, but had an average case-fatality rate of approximately 2.5% (Taubenberger and Morens, 2006).

Moreover, nuclear science arguably provides a precedent for censoring research with immensely destructive potential. Research on the first atomic bombs – which, combined, killed approximately 300 000 people – was censored vigorously. Many atomic scientists accepted this practice (though not without complaint) (Rhodes, 1986). Given sufficient evidence of the likelihood that a devastating biological attack could occur, censorship of contemporary dual-use life science research might arguably be appropriate (at least in exceptional circumstances). A consensus in debates about dual-use life science research is that an appropriate balance should be struck between the goals to promote scientific progress, on the one hand, and security, on the other. What an appropriate balance would be, however, remains an open question (Selgelid, 2007; Evans, 2013).

Beyond the right to freedom of speech, there are other reasons that have been given for protecting the publications of researchers and scientific publications in particular. None of these reasons gives an absolute justification for open communication of science, but they can be incorporated into the considerations of value described above. First, scientific freedom and openness are essential to the structure of science by providing a forum within which ideas can be examined, reviewed, and critiqued. If there is little prior knowledge of where innovative and beneficial research will arise, maximizing openness maximizes the chance of reaping the benefits of new research (Resnik, 2008).

Not all agree, however, that this is sufficient reason to protect scientific speech. Significant effort is put into funding allocation to decide which projects and scientists will produce the best results. Though subject to critique, these systems show that research is planned and funded with certain expectations about where beneficial research will occur. If such judgments are appropriate in decision-making about what science should be done, the argument goes, why should they not also play a role in decision-making about what research will be published? There is a clear trade-off between security and scientific progress, but neither is typically held to be absolutely important (Evans, 2013).

Some claim that the institutional function of a university is to produce and disseminate knowledge. Research and scholarship conducted at universities, therefore, should perhaps be protected in virtue of the moral purpose of the institution of the university. Interfering with this function would not only violate the rights of researchers, but arguably also pervert the purpose of the university (Miller, 2010).

Others have noted, however, that the interference of private funding bodies, government agencies, and companies into the research affairs of universities can pervert their course *even if knowledge results* (Elliott, 2010; Evans, 2013). Thus, the function of a university is not sufficient reason to argue against interference, but merely to argue against certain kinds of interference – e.g., such as that of Lysenko and tobacco companies. Whether censorship of dual-use research would be an unacceptable kind of interference, however, is debatable.

Finally, there are pragmatic considerations about the practice of censorship, and what they mean for research and scholarship. As the history of censorship demonstrates, censorship is a practice that is either indicative of, or prone to, egregious abuses of power. Even if there are good reasons to censor science in some cases, there is still some risk that the censorship process would be perverted by powerful interests. The Soviet scientists persecuted under Lysenko are the most extreme example of the horrors that can be inflicted against scientists; less violent, but still troubling, is the suppression of public health research by tobacco companies.

Even the *Atomic Energy Act*, in which research was censored out of concern about the security risks of nuclear science and technology, has been used inappropriately. Empowered through the *Atomic Energy Act*, the Atomic Energy Commission wielded extensive power that it abused to promote its own interests as well as those of the people it served. Famously, this power was wielded against J. Robert Oppenheimer, the former director of the Manhattan Project, in depriving him of his security clearance and exiling him from the classified nuclear community.

Numerous aspects of Oppenheimer's trial arguably involved abuses of power. There was a well-known vendetta between Lewis Strauss, chairman of the Atomic Energy Commission, and Oppenheimer. There was also the background hysteria about communism, and Oppenheimer's past – a past that he lied about during security proceedings – included association, though not formal affiliation, with the Communist movement. The US Air Force, moreover, desired the construction of large, strategic nuclear weapons, a program that Oppenheimer publicly opposed. Oppenheimer's trial serves as a reminder that personal grudges, ideological disputes, and political maneuvering can all be used to pervert a process designed to ensure the security of a people (Bernstein, 1982).

Conclusion

Any further debate about the ethics of censoring research and scholarship will ultimately turn on the weight of the factors discussed above: the balance of values in particular acts of speech; the design, function, and operation of social institutions; and the practical difficulties associated with censorship, which has a murky history. Censorship of research and scholarship is, like the censorship of any speech, ethically fraught. It is, however, a source of contemporary debate as the technologies that stem from scientific research become the potential tools of terrorists. Though considerable uncertainty remains

about how best to weigh the values implicated in the dissemination of research findings against the harms of the misuse of biotechnology, there is wide agreement that the right to freely disseminate research is neither absolute, nor always top priority.

Acknowledgment

Significant sections of this work were completed at the Center for Human Bioethics, Monash University, as a Visiting Fellow with support from the Wellcome Trust grant "Building a Sustainable Capacity in Dual-Use Bioethics."

See also: Civil Liberties and Human Rights; Freedom of the Press; Intellectual Property Rights: Ethical Aspects; Research Funding: Ethical Aspects; Secrecy, Anthropology of.

Bibliography

Alberts, B., 2005. Modeling attacks on the food supply. Proceedings of the National Academy of Sciences 102 (28), 9737–9738.

Alexander, L., 2005. Is There a Right of Freedom of Expression? Cambridge University Press, New York.

Atomic Energy Act of 1946, 60 Stat. 755, United States Congress.

Bernstein, B.J., 1982. In the matter of J. Robert Oppenheimer. Historical Studies in the Physical Sciences 12 (2), 195–252.

Elliott, C., 2010. White Coat, Black Hat: Adventures on the Dark Side of Medicine. Beacon Press, Boston.

Evans, N.G., 2013. Great expectations—ethics, avian flu and the value of progress. Journal of Medical Ethics 39 (4), 209–213.

Gillispie, C.C., 1992. Science and secret weapons development in revolutionary France, 1792–1804: a documentary history. Historical Studies in the Physical and Biological Sciences 23 (1), 35–152.

Glantz, Stanton A., Slade, John, Bero, Lisa A., Hanauer, Peter, Barnes, Deborah E., 1996. The Cigarette Papers. University of California Press, Berkley.

Godman, P., 2000. The Silent Masters: Latin Literature and Its Censors in the High Middle Ages. Princeton University Press, Princeton.

Healy, R., Marsh, J.G., Woodbridge, D.B., Mackey, J.B., 1965. Panel—industry looks at classification management. Journal of the National Classification Management Society 1 (2–4), 33–38.

Henry, J., 2012. A Short History of Scientific Thought. Palgrave MacMillan, London.

House Bill 185, State of Tennessee, 1925.

Joravsky, D., 1986. The Lysenko Affair. Harvard University Press, Cambridge.

Miller, D.M., 2008. The thirty years war and the Galileo affair. History of Science 46, 49–74.

Miller, S., Selgelid, M.J., 2008. Ethical and Philosophical Considerations of the Dual-Use Dilemma in the Biological Sciences. Springer, Dordecht, NE.

Miller, S., 2010. The Moral Foundations of Social Institutions. Cambridge University Press, Cambridge.

National Research Council (NRC), 2004. Biotechnology Research in an Age of Terrorism. National Academies Press, Washington, DC.

National Science Advisory Board for Biosecurity (NSABB), 2007. Proposed Framework for the Oversight of Dual Use Life Science Research: Strategies for Minimizing the Potential Misuse of Research Information. Working Paper for the NSABB.

Office of Biotechnology Activities, 2012. U.S. government policy for oversight of life sciences dual use research of concern. Office of Biotechnology Activities. National Institutes of Health, Bethesda, MD. http://www.phe.gov/s3/dualuse/Documents/us-policy-durc-032812.pdf (accessed on 06.08.14.).

Quist, A.S., 1994. Security Classification of Information: Principles of Classification of Information, vol. 2. United States Department of Energy, Oak Ridge Gaseous Diffusion Plant.

Resnik, D.B., 2008. Scientific autonomy and public oversight. Episteme 5 (2), 220–238.

Rhodes, R., 1986. The Making of the Atomic Bomb. Simon & Schuster, New York.

Schauer, F., 1978. Fear, risk and the first amendment: unraveling the chilling effect. Boston University Law Review 58, 685–732.

Simonson, S., 2005. Letter to Dr. Bruce Alberts. Department of Health and Human Services, Washington, DC. http://www.fas.org/sgp/bush/hhs052705.pdf (accessed on 06.08.14.).

Selgelid, M.J., 2007. A tale of two studies: ethics, bioterrorism, and the censorship of science. The Hastings Center Report 37 (3), 35–48.

Suolahti, J., 1963. The Roman Censors. A Study on Social Structure. Annales Academiae Scientiarum Fennicae, Swedenm.

Sweeney, M.S., 2000. Secrets of Victor: The Office of Censorship and the American Press and Radio in World War II. University of North Carolina Press, Chapel Hill.

Taubenberger, J.K., Morens, D.M., 2006. 1918 Influenza: the mother of all pandemics. Revista Biomédica 17, 69–79.

Tolbert, Jane T., 2003. Peiresc and censorship: the inquisition and the new science, 1610–1637. The Catholic Historical Review 89 (1), 24–38.

Waldron, J., 2010. Dignity and Defamation: The Visibility of Hate. Harvard Law Review. JSTOR, 1596–1657.

Wein, L.M., 2009. OR Forum–homeland security: from mathematical models to policy implementation: the 2008 Philip McCord Morse lecture. Operations Research 57 (4), 801–811.

Wein, L.M., Liu, Y., 2005. Analyzing a bioterror attack on the food supply: the case of Botulinum Toxin in milk. Proceedings of the National Academy of Sciences 102 (28), National Academy of Sciences. 9984–9989.

Relevant Websites

http://www.article19.org/ – Article 19: Defending Freedom of Expression and Information.

http://www.fas.org/biosecurity/resource/ – Federation of American Scientists: Biosecurity and Biodefense Resources.

http://www.fas.org/biosecurity/education/dualuse/ – Federation of American Scientists: Case Studies in Dual Use Biological Research.

http://www.indexoncensorship.org/ – Index on Censorship.

http://osp.od.nih.gov/office-biotechnology-activities/biosecurity/nsabb – National Science Advisory Board for Biosecurity.

http://www.brad.ac.uk/bioethics/ – University of Bradford School of Social and International Studies, "Dual-Use Bioethics."

http://www.virtualbiosecuritycenter.org/ – Virtual Biosecurity Center.

Census Microdata

Steven Ruggles, University of Minnesota, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Census microdata provide individual-level information describing the characteristics of persons, families, and households. Compared with conventional aggregate-level census data, microdata offer several advantages, including the ability to customize tabulations, construct new variables based on combining characteristics of family members, conduct individual-level multivariate analysis, and make data comparable across time and space. Limitations of census microdata vary depending on the source, but include suppression of detailed information to ensure confidentiality, limited sample size, and the need for specialized statistical software. Census microdata have existed since the 1960s, but vast new data sets are becoming available for social science research.

Definition of Census Microdata

Census microdata consist of individual responses to census questionnaires, coded into a form suitable for statistical analysis. The term census microdata refers to responses from both complete censuses and from large-scale household surveys conducted by statistical agencies, such as the American Community Survey for the United States or the *Mikrozensen* (Microcensus) for Germany. Census microdata are among the most widely used sources in the social sciences.

Publicly accessible census microdata are usually samples of both individuals and households. With a few exceptions, records describing individuals are nested within families and households, and the data provide information about the interrelationships of the members of each residential group. The samples are typically large; most include at least 1% of the population and the most common sample densities are 5 or 10% of the population. Complete microdata including all enumerated individuals in a population are sometimes available for historical populations or through restricted-access data enclaves.

Most census microdata include information on economic activities, ethnicity, educational attainment, fertility, migration, marital status, and consensual unions. In developing countries, most censuses provide information about mortality and disabilities, and there are often extensive housing characteristics, usually including water supply, sewage, and physical characteristics of the dwelling such as materials used for the floor and roof.

The availability of census microdata is growing rapidly. In 2000, publicly accessible microdata were available for just five countries, and included data on fewer than 100 million individuals. By 2013, census microdata were freely available for 700 censuses and surveys of 78 countries, which collectively provide data on 750 million persons residing in 200 million households. As shown in **Figure 1**, it is estimated that the total will grow to approximately one thousand million records from 100 countries by 2018 (Ruggles, 2014).

Advantages and Limitations of Census Microdata

Advantages

Until recently, most census statistics were available only in aggregated form. Whether disseminated in books of statistical tables or in machine readable form, these data were designed to describe the characteristics of administrative units (such as states, counties, or municipalities) or statistical units defined by the census agency (such as census tracts or block groups). Unlike aggregated census data that provide characteristics of *places*, microdata provide the characteristics of *people*. This allows investigators to pursue powerful research strategies.

- Tabulations customized to particular research questions. For example, suppose one wants to study the relationship of teen marriage to high school dropout rates. The US census never published a table on marital status by school enrollment by age, so that topic cannot be investigated using the aggregated census data. With microdata, however, investigators can analyze any combination of characteristics.
- Customized variables based on multiple records. Because of the hierarchical structure of the microdata – individuals are organized into families and households, and the interrelationships of the individuals are known – microdata make it possible to develop new measures that combine characteristics of more than one person. For example, by comparing a husband's ethnicity with that of his wife, one can measure ethnic intermarriage. Investigators can develop a virtually limitless number of customized measures that combine characteristics from multiple persons, such as number of own children attending school, age differences of siblings, or coresidence of kin.
- Multilevel multivariate analysis. Investigators can move beyond the simple three-way or four-way tables provided by aggregate census data and control for many individual, family, and community-level variables simultaneously using multilevel multivariate techniques.
- Comparability across time and space. Aggregated tables are often difficult or impossible to compare across time or between countries. If an investigator is interested in studying trends on high school dropout rates in the United States during the 1980s, the aggregate census tables are of little use. The 1980 census provides information on school enrollment for the age groups of 14–15 and 16–17, whereas in 1990 the age groups were 10–14 and 15–17; the incompatibility of the age groups means we cannot compare dropout rates over time. Microdata make the comparison simple, since school enrollment at every age can be easily calculated.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.41082-2

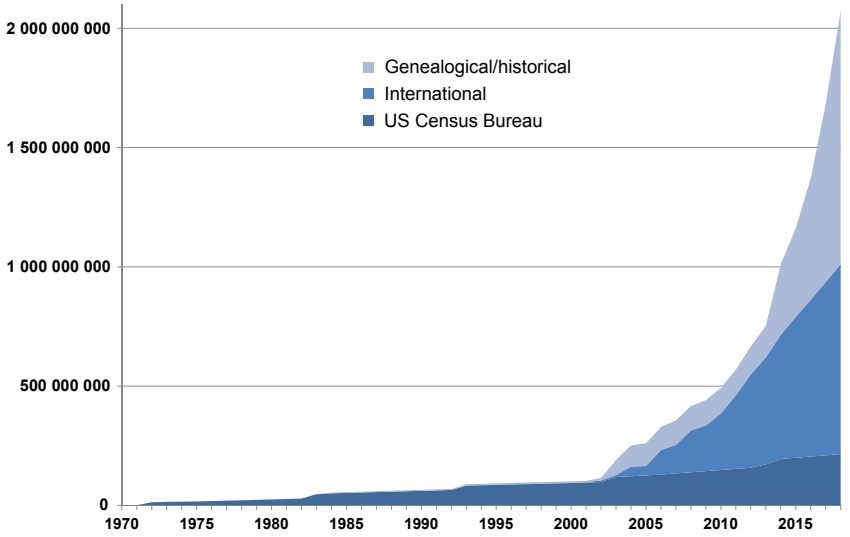

Figure 1 Number of individual cases of census microdata readily accessible for research. Source: Ruggles, S. Big microdata for population research. Demography, 51, 287–297.

Limitations

Despite the power of microdata, there are three limitations that make preaggregated data preferable for some applications.

- Limited detail. In some countries – particularly in Europe and North America – privacy concerns limit the geographic detail available in the publicly accessible census microdata. In the United States, for example, only places of 100 000 population or greater can be identified. This means that analyses of small geographic areas, such as neighborhoods, the only potential sources are aggregated data or restricted-access microdata. Beyond geography, a few countries suppress other key information commonly needed for demographic analysis. Several countries, group age into 5-year groups. In a few countries, such as the Netherlands, the individual records are not grouped into families and households, greatly compromising the utility of the data. These limitations tend to be substantially less severe outside of the most-developed countries, and older historical microdata generally provide full detail on all variables.
- Limited sample size. Census microdata provide many more cases than most other forms of social and economic microdata; most census microdata samples include hundreds of thousands or millions of observations. Nevertheless, the sample sizes are sometimes inadequate for studying small geographic areas or small subpopulations; in these cases, aggregated data or narrowly focused surveys may be the only alternatives.
- Need for specialized analysis software. Unlike published tabular statistics, use of microdata traditionally has required specialized statistical software packages. These packages – such as SPSS, SAS, and Stata – are complicated, expensive, and demand substantial computing power. Online data analysis tools are beginning to mitigate this constraint.

Census microdata may also be compared with small-scale demographic survey data, such as the Demographic and Health Surveys widely used for studying conditions on the developing world. Such surveys generally have more comprehensive subject coverage than censuses, and data quality is usually high. Sample sizes are comparatively small, however, and they are usually clustered, limiting the potential for spatial analysis. Census microdata samples are usually an order of magnitude larger than demographic surveys. Although some census data are unreliable, overall nonresponse rates compare favorably with sample surveys. Census microdata can be used fruitfully in conjunction with surveys to provide information about local context that cannot be generated from the surveys themselves.

Origins and Development of US Census Microdata

1960s and 1970s

The first census microdata was created as a byproduct of the 1960 US census (U.S. Bureau of the Census, 1964). The expansion of social science research in the late 1950s and early 1960s had led to growing demand for special tabulations designed to answer specific research questions. The Census Bureau prepared these special tables on a reimbursable basis, but the costs were high; the Census Bureau typically charged $5000 for a basic tabulation project at a time when the average annual salary of a US social scientist was just a little over $7000 (Kraus, 2011; Ruggles et al., 2012).

To meet the demand for customized tabulations, in 1962 the Census Bureau drew a 1-in-1000 extract of the basic data tapes they had used to create tabulations for the published census volumes. To preserve confidentiality, the Census Bureau removed names, addresses, and other potentially identifying information. The Census Bureau made the data available to the research community on 7 IBM tapes or 11 UNIVAC tapes for a charge of $1500 (Brunsman, 1963; Duncan and Shelton 1978; Hauser, 1960). Only a few research centers had access to computers, so the Census Bureau also made a 1-in-10 000

version available on 18 000 punch cards that investigators could analyze using older unit record machines (Mason et al., 1977; U.S. Census Bureau, 1964).

The 1960 public use sample revolutionized analysis of the American population and led to an explosion of new census-based research. It allowed researchers to make tabulations tailored to their specific research questions. As Otis Dudley Duncan put it,

> The importance of this innovation can hardly be overestimated. We have known for a long time that certain essential social indicators are available in principle from the Federal statistical system. Yet all too often efforts to put information into an appropriate form are frustrated by the inadequacy of the published summary tables for the purpose at hand. With access to the unit records, the social scientist may specify in detail how variables are to be manipulated so as to produce an optimal estimate of the magnitude desired.
> Duncan, 1974: 5097

Mason et al. (1977) concurred, explaining that the 1960 census microdata sample was a 'development of profound significance to social research' because it gave the research community 'freedom to retabulate or manipulate without the constraints imposed by a fixed set of printed volumes.' Not only did the new sample allow researchers to make tabulations tailored to their specific research questions, it also allowed them to apply new methods – such as individual-level multivariate analysis – to the analysis of census data.

Despite its high impact, the initial 1960 sample had two serious limitations. First, the sample density was small. The one-in-1000 sample density yielded about 180 000 person records. Given the modest capacity of computers in 1964, this was a substantial number of cases. Nevertheless, as researchers began to use the sample for detailed analyses of small population subgroups, however, its limitations became apparent. Second, the 1960 public use sample provided little geographic information. To ensure confidentiality, the Census Bureau stripped off all information on places below the state level. This meant, for example, that it was impossible to extract a subsample of the New York City population.

Both of these problems were addressed by the 1970 public use samples. The Census Bureau expanded the size of the 1970 samples more than 60-fold compared with the 1960 sample. The Census Bureau provided six independent public use samples for 1970, each of which had a one-in-100 density. Users who required an exceptionally large number of cases could combine the samples to obtain six-percent sample density, or about 12 million person records. In addition, the 1970 samples provided a variety of alternate geographic codes, although the Census Bureau still did not identify places of less than 250 000 population (U. S. Census Bureau, 1972).

In 1972, the Center for Research Libraries obtained funding from the National Science Foundation to make a new version of the 1960 sample. The new version enlarged the sample density from one-in-1000 to one in-100, and at the same time reorganized the coding schemes and record layouts to be compatible with the samples from 1970 (U.S. Census Bureau, 1973). This compatibility made it relatively easy for investigators to pool data from 1960 to 1970 and thus incorporate change over time into their analyses, and this became a widespread research strategy. The public use samples for 1960 and 1970 quickly became essential tools of American social scientists. In the mid-1970s, demographers Samuel Preston at the University of Washington and Halliman Winsborough at the University of Wisconsin independently developed projects to extend the series backward by creating new historical samples of earlier censuses.

1990s

By 1991, there were national census microdata samples available for eight US censuses: 1880, 1900, 1910, 1940, 1950, 1960, 1970, and 1980. A ninth census year – 1990 – was planned. For the first time, the potential existed for individual-level national studies of long-run social and economic change, but in practice, such analysis was cumbersome. The nine samples resulted from separate projects headed by four different investigators and six different performance sites. Even in the few cases where the same investigator working at the same location created samples for more than one census year, there were often incompatibilities, even if there was a recognizable family resemblance. For example, the two Winsborough samples (1940 and 1950) use completely different record layouts and have different coding schemes for many variables, including occupation, institution type, citizenship, and employment status. As noted, the Census Bureau samples for 1960 and 1970 were largely compatible with each other. The 1980 sample, however, was completely different, and offered substantially greater detail than did 1960 or 1970 (U.S. Census Bureau, 1982).

The Integrated Public Use Microdata Series (IPUMS) was launched in 1992 to reconcile the differences among the eight US census microdata samples. Using structured data-integration metadata, IPUMS implemented a system of consistent codes across census years without losing information, and provided comprehensive integrated documentation highlighting comparability issues. IPUMS began disseminating integrated microdata through the Internet in 1993. The database now includes microdata from every US decennial census from 1850 forward, as well as Census Bureau microdata from the Current Population Survey and the American Community Survey. In 1995, IPUMS introduced Web-based data access technology that allowed investigators to design customized data sets incorporating microdata from multiple sources, greatly simplifying access to large-scale microdata (Ruggles, 2005).

International Census Microdata

Despite the success of large-scale US microdata, statistical agencies in other countries were slow to create similar resources. A few national statistical offices – including those of Australia, Brazil, China, Colombia, Mexico, Norway, and South Africa – made internal microdata available to selected academic researchers on a case-by-case basis. Several organizations developed substantial archives of census microdata from multiple countries for comparative research, but they were unable to provide access to off-site researchers.

Statistics Canada began producing Public Use Microdata Files (PUMF) in 1974, and the United Kingdom created Samples of Anonymized Records (SARs) in 1993. In both these cases, the samples were very small, confidentiality edits compromised usability for many applications, and access was difficult for researchers from other countries. Notwithstanding these limitations, the PUMF and SARs have been successful, and they rank among the most important sources for demographic research in Canada and the United Kingdom.

Most countries have had no systematic program for preservation or reuse of census microdata once the statistical agency had published summary tables. As a result, most machine-readable census microdata from the 1960s and 1970s had already disappeared by the mid-1990s, and large portions of the surviving microdata was at immediate risk of destruction.

Robert McCaa, a historical demographer at the University of Minnesota, led an effort to liberate and preserve the world's census microdata. From 1998 to 2013, McCaa recruited 100 national statistical agencies around the world to collaborate with the IPUMS project (**Figure 2**); perpetual agreements with each country ensure long-run preservation and free access for the academic community (Minnesota Population Center, 2013).

By 2013, IPUMS had released anonymized integrated microdata samples for 238 censuses of 74 countries taken between 1960 and 2011. These 74 countries have more than three-fourths of the world's population. Most of the countries outside of Western Europe and North America provide good geographic precision, identifying places with as few as 20 000 residents. For 63 countries, IPUMS provides microdata from multiple census years (3.6 census years per country on average). The samples are usually large: almost two-thirds of the samples include 10% or more of the population, and 85% of the samples include at least 5%. Taken together, the 238 samples currently available include information on 544 million individuals; by 2018 the IPUMS project expects to have released data for about 800 million persons drawn from 300 censuses of about 100 countries.

Census Microdata from Genealogical Sources

Since 2000, the availability of historical census microdata has expanded rapidly through the adaptation of genealogical data to meet scientific purposes. Starting in 1982, the genealogical arm of the Church of Jesus Christ of Latter-Day Saints – known as FamilySearch – organized an army of 25 000 volunteers to transcribe information stored on microfilm images of the 1880 US census and the 1881 censuses of Britain and Canada. When the work was complete in the late 1990s, FamilySearch provided copies of digital census transcriptions describing 84 million persons for academic use. These data became the centerpiece of a new international collaboration to develop integrated complete enumerations of the North Atlantic region suitable for scientific research. The North Atlantic Population Project (NAPP) began in 1999 with participants from Britain, Canada, Iceland, Norway, and the US. The scope of the collaboration has grown over the past decade to include universities and national archives in Denmark, Egypt, Ireland, Mexico, and Sweden. Each year, more censuses are added, and NAPP presently disseminates data on 130 million persons drawn from 33 censuses of nine countries enumerated between 1703 and 1930 (Ruggles et al., 2011).

Over the coming 5 years, the quantity of accessible historical microdata will increase by an order of magnitude, to 1.1 billion individual records. Most of this results from collaborations of academic researchers with large genealogical organizations, including the two leading commercial vendors of genealogical

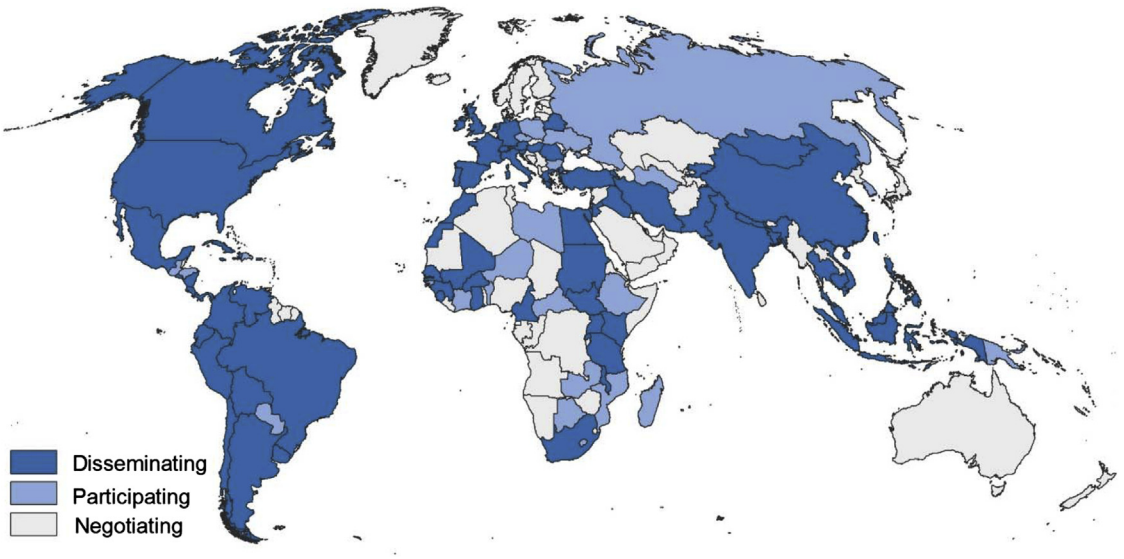

Figure 2 Countries participating in Integrated Public Use Microdata Series-International. Disseminating means that microdata were being disseminated as of June 2013. Participating means that the statistical agency has joined the partnership, and microdata will be available soon. Source: Minnesota Population Center, 2013. Integrated Public Use Microdata Series, International: Version 6.2 [Machine-readable Database]. University of Minnesota, Minneapolis.

data, Ancestry.com and findmypast.com. New data products will include the entire censuses of Great Britain from 1851 to 1911 (a net addition of 175 million person records) and complete census microdata covering all enumerated persons in the United States from 1790 to 1940 (a net addition of 730 million person records).

With a few comparatively small exceptions, the historical data are transcribed from public sources and there are no confidentiality restrictions. This means that the historical data sets ordinarily provide full information on geographic location. The historical data sets also include names, although to protect commercial interests of the commercial partners, there are restrictions on the dissemination of names for the data sets donated by genealogical organizations. Nine countries have multiple complete enumerations with full identification, allowing investigators to trace individuals and households over the life course and across several generations. The data will be freely disseminated through IPUMS.

Restricted Census Microdata

The Census Research Data Centers (RDCs) operated by the US Census Bureau's Center for Economic Studies house complete decennial census microdata for 1970 through 2000, including both short-form and long-form records. Thanks to a recently completed project to restore the 1960 census, complete long-form data covering 25% of the 1960 population are now available through the RDCs (Ruggles et al., 2011). The RDCs also house complete American Community Survey data covering 47 million persons, and are adding about 5.4 million persons per year. In all, the microdata housed in the RDCs currently describe 1.05 billion persons with full geographic identification down to the block level. The Minnesota Population Center is currently converting these data to IPUMS format.

The Minnesota Population Center is developing an international restricted data enclave modeled on the Census RDCs to provide access to about 130 complete census enumerations with full geographic detail. These data provide information on approximately 1.6 billion persons. Like the Census RDC, no restricted data will ever leave the secure servers; instead, the enclave will provide remote access through virtualization. All results will be assessed for potential violations of confidentiality, and the data enclave staff will email them to the researchers only after they are cleared (Ruggles, 2014).

New Applications of Census Microdata

Researchers will soon have access to data spanning two centuries for some North Atlantic countries and the last two to five decades for much of the rest of the world, allowing direct observation of demographic and economic transformations that are reshaping society. Key research topics include residential segregation; migration and migrant settlement patterns; urban sprawl; the economic and ideational context of declines in fertility, mortality, and intergenerational coresidence; rural depopulation and agricultural consolidation; the identification of concentrated poverty; causes and levels of change in ecosystems as a function of human–environment interactions; comparative cross-national policy analysis; and multilevel analysis of the impact of community characteristics on individual behavior.

Variation in economic development and demographic behavior creates opportunities to assess the impact of local characteristics on individual behavior, thereby offering the potential for understanding the consequences of both structural and ideational change. For example, Esteve et al. (2012) examined the rise of cohabitation for 350 regions of 13 Latin American countries over four decades, and assessed the impact of national and regional conditions on the pace of change.

The large scale of census microdata enables study of particular communities and small dispersed populations, but it also allows studies that span many countries and large time spans. Restricted census microdata with full geographic information will allow analysis of the effects of local context on individual behavior at multiple scales, focusing on immediate neighbors, enumeration districts, and labor market areas.

Linked census microdata have shifted our perception of life-course change in the past. They have revealed that occupational mobility was far higher in the nineteenth century than it is today, migration was far more frequent, and the formation of intergenerational families was most common among the rich, rather than the poor (Ferrie, 2005; Ruggles, 2011). The next generation of linked microdata will be far more powerful, with 100 times the number of records, more reliable links, and coverage across entire lives and across multiple generations, allowing multilevel analysis of the demographic and economic context of mobility.

See also: Censuses, Population: Comparative International Aspects; Censuses: Current Approaches and Methods; Censuses: History and Methods; Data Bases and Statistical Systems: Demography; Data Bases and Statistical Systems: Economics (General); Data Bases and Statistical Systems: Sociology.

Bibliography

Brunsman, H.G., 1963. Letter to Dr. Joshua Lederberg, 22 April 1963. Joshua Lederburg Papers, National Library of Medicine. http://profiles.nlm.nih.gov/BB/G/K/F/C/_/bbgkfc.pdf.

Duncan, O.D., 1974. Developing social indicators. Proceedings of the National Academy of Sciences 71, 5096–5102.

Duncan, J.W., Shelton, W.C., 1978. Revolution in United States government statistics: 1926–1976. U.S. Department of Commerce Office of Statistical Policy and Standards. U.S. Government Printing Office, Washington, DC.

Esteve, A., Lesthaeghe, R., López, A., 2012. The latin american cohabitation boom, 1970–2007. Population and Development Review 181, 55–82.

Ferrie, J.P., 2005. History lessons: the end of American exceptionalism? Mobility in the United States since 1850. Journal of Economic Perspectives 19, 199–215.

Hauser, P.M., 1960. The 1960 census as an instrument for demographic research. Population Index 26, 199–211.

Kraus, R.S., 2011. Statistical Déjà Vu: The National Data Center Proposal of 1965 and its Descendants. Census Bureau Working Paper. http://www.census.gov/history/www/reference/publications/working_papers.html.

Mason, W.M., Taeuber, K.E., Winsborough, H., 1977. Old Data for New Research. Report of a Workshop on Research Opportunities and Issues in the Design and Construction of Public Use Samples from the 1940 and 1950 Censuses and from Current Population Surveys from 1960 Forward. Madison, WI June 283-0, 1976. Center for Demography and Ecology Working Paper, 77–83.

Minnesota Population Center, 2013. Integrated Public Use Microdata Series, International: Version 6.2 [Machine-readable Database]. University of Minnesota, Minneapolis.

Ruggles, S., 2005. The Minnesota population center data integration projects: challenges of harmonizing census microdata across time and place. In: 2005 Proceedings of the American Statistical Association, Government Statistics Section. American Statistical Association, Alexandria, VA, pp. 1405–1415.

Ruggles, S., 2011. Intergenerational coresidence and family transitions in the United States, 1850–1880. Journal of Marriage and the Family 73, 136–148.

Ruggles, S., 2014. Big microdata for population research. Demography 51, 287–297.

Ruggles, S., Alexander, J.T., Genadek, K., Goeken, R., Schroeder, M., Sobek, M., 2012. Integrated Public Use Microdata Series: Version 5.0 [Machine-readable Database]. University of Minnesota, Minneapolis.

Ruggles, S., Roberts, E., Sarkar, S., Sobek, M., 2011. The North Atlantic population project: progress and prospects. Historical Methods 44, 1–6.

U.S. Census Bureau, 1964. Censuses of Population and Housing, 1960: Two National Samples of the Population of the United States, 1/1,000, 1/10,000: Description and Technical Documentation. GPO, Washington, DC.

U.S. Census Bureau, 1972. Public Use Microdata Samples of Basic Records from the 1970 Census: Description and Technical Documentation. U.S. Government Printing Office, Washington, DC.

U.S. Census Bureau, 1973. Technical Documentation for the 1960 Public Use Microdata Sample. U.S. Government Printing Office, Washington, DC.

U.S. Census Bureau, 1982. Public Use Microdata Samples of Basic Records from the 1980 Census: Description and Technical Documentation. U.S. Government Printing Office, Washington, DC.

Relevant Website

Integrated Public-Use Microdata Series https: www.ipums.org/ (access to census microdata from all sources with associated metadata).

Censuses, Population: Comparative International Aspects

Tahu Kukutai, The University of Waikato, Hamilton, New Zealand
Victor R Thompson, Rider University, Lawrence, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by B.A. Bailar, volume 3, pp. 1595–1599, © 2001, Elsevier Ltd.

Abstract

In most countries around the world, the population census is a crucial tool of governance, providing vital information about a country's size, spatial distribution, and socio-demographic characteristics. This article identifies some of the defining characteristics of the population census and explores variations in census practices in an international context. The expansion in the conceptualization of the census, combined with the introduction of technological innovations in data capture, coding, and dissemination, has both complicated and enriched the possibilities for comparative census analysis in the regional and global contexts.

Introduction

In most countries around the world, the population census is a crucial tool of governance, providing vital information about a country's size, spatial distribution, and socio-demographic characteristics. Often the census is the only source of data from which it is possible to derive reliable estimates and social indicators for small areas and subpopulations. Although the systematic enumeration of a country's entire population is a quintessentially modern development, the counting of populations for military or taxation purposes dates back to ancient times. The United States census of 1790 is often credited as the first national population census, although it excluded Native Americans and counted Black individuals as only three-fifths of a person. The census spread throughout Western Europe and the 'New World' in the nineteenth century, and was ubiquitous by the middle of the twentieth century. In the current 2010 'census round' (2005–14), 227 out of a total of 233 countries had either undertaken or were planning to undertake some form of population census (United Nations, 2010).

With respect to its core features, the census has changed very little since the United Nations instigated its 'World Program on Population and Housing Censuses' 60 years ago (United Nations, 2010). Concepts such as universality, simultaneity, and periodicity (see below) are as important now as they were then. However, the ways in which countries compile census data has undergone significant change. The last decade, in particular, has seen a shift away from the traditional census approach, involving a total population enumeration at a single time point, to alternative census models which make greater use of register, administrative, and survey data (Coleman, 2013; Dugmore et al., 2011; United Nations, 2010; Valente, 2010). This expansion in the conceptualization of the census, combined with the introduction of technological innovations in data capture, coding, and dissemination, has both complicated and enriched the possibilities for comparative census analysis in regional and global contexts. These developments and their implications are discussed throughout this article.

Defining Features of National Censuses

Bodies Responsible for Population Censuses

Population and housing censuses are always undertaken by the national government, often under the auspices of a National Statistical Office (NSO) working within a set of statutory limitations relating to content and data confidentiality. Where data are used for the purpose of apportionment, separate legislation may also apply. Beyond political functions, census data are widely used to inform policy and planning at regional, national, and local levels, as well as for research, commercial, and other uses. Confidentiality is a core concept and influences how individuals' data are stored, coded, aggregated, and accessed. In some countries public access to individual census records is permitted after a certain period of time. In the US, access to individual census records is restricted to 72 years by law, while in Ireland it is 100 years.

Geographical Specificity

Each country's population census conforms to a set of core characteristics. The most distinguishing feature of a census is its geographic specificity, with the population to be counted clearly contained within a defined territorial boundary. Census geography is multilevel and ranges in granularity from national boundaries to small areas such as 'census blocks' (US), 'meshblocks' (New Zealand and Australia), and 'output areas' (United Kingdom). While some change to census geography over time is inevitable, this needs to be carefully planned. Sometimes data classification schemes have to be adjusted in order to preserve temporal comparability. There is no standard approach for the enumeration of nonsovereign territories and dependencies. In some cases, such as in Puerto Rico, the territory is included within the national census of the mainland; in other cases, such as the French overseas departments or territories (DOM-TOMs), a separate enumeration is conducted.

Household Data

Although the individual is the key unit of analysis for the population census, most countries that undertake a population census also enumerate dwellings so that household structure and characteristics can be analyzed alongside data on individuals. The dual population and housing census approach typically involves two separate census questionnaires: one is designed to elicit information on an individual's characteristics and circumstances; the other is concerned with the relationships of individuals within a housing unit and specific features of the dwelling. The distribution and planning of the housing census must also take into account individuals who do not live in regular housing, such as those who are homeless, refugees, or incarcerated. This is sometimes a controversial issue. In the US, for example, political power is based on the number of people residing in a legislative district. Consequently, decisions on the legislative districts, to which people are assigned, are highly politicized. For example, the incarcerated population in the US is concentrated in rural areas far from their home legislative districts. As a result, census counts are biased toward rural areas with prisons over urban minority areas, from where a disproportionate number of prisoners originate. In developing countries, group enumeration is sometimes used to enumerate nomadic groups, where only the number of persons for broad age-sex groups is recorded (Bryan and Heuser, 2004). Seasonal workers and individuals engaged in daily labor-related migration which takes them across national borders are another example of special populations for which different procedures might be used. Differences in how countries treat special populations can have an important effect on the estimated size of the world's population (Bailar, 2001). While the census aims to cover the entire population, some groups, such as military personnel serving overseas, are necessarily omitted. This is distinct from the exclusion of subpopulations on the basis of citizenship distinctions such as race, ethnicity, gender, and property ownership. The inclusion of absentee residents and overseas visitors depends on whether the basis of enumeration is *de jure* (place of usual residence) or *de facto* (location on census day). Making comparisons between countries that use different counting rules can be problematic, as is undertaking time-series analysis when there has been a shift from one method to another. In many countries census participation is compulsory by law, but responses to specific questions such as religious affiliation may be voluntary.

Simultaneity and Periodicity

Simultaneity enables the census to provide an entire population snapshot at a discrete point in time. For that reason the census is designated to take place on a specific date, even though the data collection process may stretch over weeks or even months. The census is usually undertaken at a time of the year when potential disruptions, such as the occurrence of extreme weather, are minimal. The duration of enumeration varies tremendously between countries and is influenced by a number of factors including resources, landmass, population size and distribution, and the availability of technology.

Periodicity refers to the frequency and regularity of the census cycle. The UN recommends that all countries conduct a census at least once in every 10 years, preferably closer to years ending in 0. Most countries, including the US, India, and England, conduct a decennial census, although some countries belonging to the former British Empire (e.g., Australia) undertake a census every 5 years. A spacing of exactly 5 or 10 years is desirable because it enables greater flexibility for cohort analysis and the presentation of results in standard demographic terms. Factors such as political unrest and natural disasters can lead to the delay or postponement of the census. The Christchurch earthquakes, for example, delayed the planned 2011 New Zealand census by 2 years. The delay or postponement of a single census can have longer-term implications if it leads to a permanent change in the frequency and regularity of the census cycle.

Census Content

Core Topics

In census-based analysis, household and individual level data are always clearly distinguished. Household data typically includes information on the structural characteristics of dwellings, occupancy information, and relationship characteristics of household members. Households are not the same as families although the terms are often used interchangeably by the public. As households may contain families or unrelated individuals living together, data on relationships within households is crucial for identifying the status of families. NSOs routinely employ socio-economic data on households and families to assess the relative well-being of their populations.

All censuses contain a large number of questions on key demographic variables for both individuals and households. Questions on age, sex, location of residence during census, number of occupants, and place or country of birth are nearly universal in their adoption across countries. Even small countries with limited resources contain some version of these questions. While the wording of core demographic questions varies across countries, there is generally a consensus on how to convert them into data. The classic demographic variable of age, for example, is always collected as an exact age, whereby respondents identify their age in exact years rather than a range of years. Likewise, sex is universally treated as a binary concept, representing either males or females. Other key variables like place or location of residence typically reflect exact locations based on recognized municipal designations and officially designated geographies. These core topics constitute a baseline set of variables around which countries build their census. In 2011 the European Union member states agreed to adopt a set of core topics to be included in each member state's census (Eurostat, 2011). These include individual level variables such as age, marital status, educational attainment, and occupation, in addition to household level variables such as housing arrangements, number of occupants, and type of ownership.

Core topics are particularly useful for demographic analysis of the population; consequently many NSO's allocate a sizable share of their time and resources to ensure the accuracy of these data (see Stukel, 2008 for a breakdown of expenditures for different countries). Age and sex are by far the most important of all information collected in a census. Nevertheless, despite

special attention given to ensure that errors of coverage are limited, errors do inevitably occur (Ewbank, 1981). Beyond age, sex, and location, the inclusion of other core topics such as marital status, residence, and place of birth are influenced by factors that include the census budget, the resources available to NSOs, and the availability of population register data.

Noncore Topics

The inclusion of noncore topics in the population census varies considerably across countries. NSOs with large budgets or those lacking population registers tend to include more questions than NSOs with limited resources or access to register-based alternatives. Ireland, for example, sent out a household survey in 2011 that was 24 pages long and included dozens of questions. Kenya's 2011 census questionnaire, on the other hand, was only two pages long, with far fewer questions. The extent to which a particular noncore topic (e.g., migration, literacy, or method of commuting) is deemed relevant can also influence its inclusion in the census. Countries with a high level of immigration, such as Canada and Australia, are more likely to include questions on migration than countries that receive less immigrants (e.g., China), although there are always exceptions.

Some noncore topics, including race, ethnicity, nationality, language, and religion, are not as universally adopted as the aforementioned topics. Indeed, some countries are hesitant about including these types of questions because of privacy issues or concerns about data being used for discriminatory purposes (Simon, 2007). While the United Nations has no formal recommendation regarding a universally accepted set of questions for race and ethnicity or related items (United Nations Statistical Division, 1998), the United Nations Committee on the Elimination of Racial Discrimination regularly advises states to collect information about these items (United Nations, 2012). From a conceptual perspective, identity-based questions are challenging because of the variable meanings that can be attributed to them, both across and within countries (Morning, 2008). Unlike demographic characteristics such as age and sex, socially ascribed identities such as ethnicity and language group are neither mutually exclusive nor consistent. The 2000 US Census, for example, allowed respondents to choose more than one race, which resulted in 2.4% of the population choosing more than one race. Native Hawaiian and Pacific Islanders reported a much higher mixed-race component at 54.4% (Jones et al., 2001). Quality analysis found that 60% of all respondents who chose more than one race in the 2000 census only reported a single race in follow-up surveys (Bentley et al., 2003).

The census is often used to evaluate the overall state of affairs in a country, making it both an important tool of scientific evidence and a flashpoint for political debate. Both core and noncore topics can be highly contested and the ultimate decision to include some and not others is as much political as it is practical. Likewise, socio-economic variables vary tremendously over time and place with respect to nomenclature, level of detail, and their conceptualization. Religious, ethnic, and national identities, for example, differ from country to country and can change dramatically even within individual countries over time. Similarly, as countries develop, the types of socio-economic variables also change. Unpaid work and time traveled to work are illustrative examples.

Recent Shifts from Traditional Censuses to Alternative Models

The substantial cost of undertaking a census, combined with the long spell between enumerations, has raised questions about the relevance and sustainability of the traditional census model. In countries where accelerated migration has led to rapid population turnover, the limited capacity of the census to capture these shifts has diminished its perceived usefulness. Growing objections to state-sponsored data collections, as well as the ballooning costs of censuses, have provided additional challenges (Coleman, 2013).

For many decades, only the Scandinavian countries used alternative methods to derive population data, but in the last decade the census has assumed a wider concept, particularly in Europe (UNECE, 2006). In the 2000 round, 18 of the 27 European Union countries (67%) responding to a United Nations Statistical Division (UNSD) survey had conducted traditional censuses (United Nations, 2010). By the 2010 round (until 2009), this had reduced to 11 countries. In most cases, the shift was to a combined approach utilizing a range of data sources including field enumeration, registers, and surveys. Of the four purely register-based countries, three were Scandinavian (Finland, Sweden, Denmark). The remaining country, Austria, shifted from a traditional census to a purely register-based model in 2011 (Lenk, 2008).

North America is interesting as the census approach in the US and Canada has undergone significant changes since 2000. In previous census rounds, a short census form was distributed to the entire population in both countries while a long form, asking detailed socio-demographic questions, was distributed to a random sample. For the most recent censuses (2010 and 2011, respectively) only short forms were used, with the long forms being replaced by sample surveys. In the case of Canada, the sample survey was entirely voluntary. At the time of writing, a number of countries were actively looking at alternative census models, among them the United Kingdom, Australia, and New Zealand (Coleman, 2013; Statistics New Zealand, 2012). France remains the only country in the world that uses a 'rolling census' model in which annual sample surveys are combined to cover 70% of the French population over a 5-year cycle (Valente, 2010).

The shift toward a more expansive concept of census has been aided by technological innovations enabling easier and more comprehensive linkage of administrative data. Although the infrastructural costs associated with establishing robust alternative models are significant, so are the long-term efficiencies and savings. In the US, for example, the expected average cost per head for the 2010 census was US$ 48.90. In many of the register-based or mixed census models of European countries such as Austria, Belgium, Denmark, and Norway, the cost per head was less than US$ 1.30 per head (UNECE Secretariat, 2010, Table 4; expected costs on a purchasing power parity basis). In Denmark the cost was a mere US$ 0.03 per head. Nevertheless, as has been demonstrated by countries that have recently transitioned from

a traditional census approach to an alternative model, there are significant challenges associated with census transformation, including population compliance and interagency collaboration. These challenges suggest that it will be many decades yet before the population census reaches its 'twilight' years (Coleman, 2013). The significant capital investment and high degree of public conformance required of registration systems suggests that register-based alternatives are likely to be restricted to developed countries with relatively well-established administrative data infrastructures.

Methodologies of Censuses

Methods of Enumeration

The decision to undertake a census raises a number of methodological issues that NSOs must address. The mode of administration is particularly important as it can have a direct impact on the accuracy of the data and the overall cost. The most common mode of enumeration is the face-to-face interview but changes in technology and concerns over budgets have made other types of administration more popular. A self-report sample of 124 countries reveals that 85% of these countries conducted face-to-face interviews in the 2010 round (United Nations, 2010). Of all the world's regions, Europe was the less likely to employ the face-to-face mode, followed by North America and Oceania. Other modes of enumeration include the paper-based self-enumeration mode, which is popular in North America and Oceania, and population registers, which are found primarily in Europe. A growing number of countries also appear to be utilizing web-based technologies. The 2011 Australian census recorded an online response rate of 33%, up from 9% in 2006 (Australian Bureau of Statistics, 2012). The 2011 rate for Australia was comparable with the target online response rate for the 2013 New Zealand census (35%), but somewhat higher than for the Czech Republic in 2011 (Coleman, 2013). The Internet is only used in combination with other modes of administration. However, given its significant cost-saving benefits, it is likely to be more widely adopted by NSOs in coming years.

Accuracy Assessment of Data

Each time a census is undertaken NSOs must also deal with coverage errors associated with miscounting the population, and content errors arising from the collection of incorrect information. Each country must adopt some method for dealing with these errors in a robust and efficient manner that corresponds to their resources and expertise. While there is no universally recognized agreement about the best methods for dealing with these errors, there have been some attempts to assist countries with data capture (United Nations, 2009a), editing procedures (United Nations, 2009b), coding practices (United Nations, 2009b), and post-enumeration surveys (PES) (United Nations, 2009c). Because the census is supposed to be an accurate count of a country's population, coverage errors assume high priority for NSOs. Net coverage (a report of the total difference between who was counted and who was not) and gross coverage (separate reporting for overcounts and undercounts) are the primary means of reporting coverage errors. Regardless of which way it is reported, the fundamental goal of evaluating census coverage errors is to determine who was counted more than once, and who was left out. In the US, for example, demographic analysis suggests that African Americans were undercounted by as much as 6.5% in 1970, and by about 2.5% in the 2010 census. By comparison, non-African Americans were only undercounted by about 2.2% in 1970, and were actually overcounted by .5% in 2010 (Mather et al., 2011).

Some countries provide detailed analyses of their coverage errors, while others do not. Developed countries typically engage in some form of post-enumeration analysis of content errors and coverage errors. Canada, for example, has been reporting on their coverage errors in their quinquennial census since 1961. Comparisons across countries show significant variation in coverage. In the 2000 census round, the net undercount was about 5.7% in the UK; 2.8% in Australia; 3.2% in Suriname, and 2.4% in Canada. Net coverage is dependent on a variety of factors and censuses do not always become more accurate. In Canada, for example, the net undercount estimate was about 2.4% in 1996 and 2.7% in 2006.

One of the primary methods of evaluating coverage and content errors is through a PES. In its most basic form, the PES replicates the census for a subset of the population and the results are compared to estimate coverage and content errors. A PES can be very complicated and resource intensive. However, at its core are the postenumeration sample (P sample) and the enumeration sample (E sample). The P sample contains a sample of the population independent of whether or not they show up in the census. The E sample is a sample based on census records (i.e., they were chosen based on their involvement in the census). The presence of the P sample allows NSOs to evaluate coverage errors and the inclusion of the E sample allows NSOs to evaluate content errors. Though not all countries use a PES to evaluate their censuses, a large number do use them on a regular basis. Demographic analysis is also commonly used, sometimes in conjunction with the PES, to evaluate the accuracy of a census. Demographic analysis is most often conducted by analyzing existing data from other sources, such as vital registrations or other registers that include basic demographic information. The US, for example, uses its Medicare enrollment database as one source of data in its demographic analysis of the total population (U.S. Census, 2010). Suriname, however, uses a subset of its census and compares it to expected distributions of the population, based on various characteristics. For example, it compares the age distribution of its census sample to the UN age-sex accuracy index. Other strategies for conducting demographic analysis include errors of coverage analysis, comparisons of age and sex distributions across censuses, and a variety of other demographic comparisons to known distributions of the population.

There are other means of assessing the accuracy and validity of censuses, such as the use of record checks to estimate coverage. Canada has used a sophisticated version of the record check, known as the Reverse Record Check (RRC), to evaluate its coverage since 1961. The RRC is a procedure that takes a random sample of the population using other sources of data and then address matches it to census data to

compare the accuracy of the census. The RRC is a very effective way of estimating net coverage but is reliant on accurate sources of alternative data that can be linked to census data.

Finally, analysis of census evaluation procedures across countries reveals that not all countries use them for the same reasons, or as extensively (United Nations, 2010). In a sample of 124 countries which reported on their usage of PES, 78% reported using a PES, either alone or in conjunction with demographic methods; 16% reported only using demographic methods, and 6% reported that they only used other methods to evaluate the accuracy of their census. While most countries (97%) reported using the PES to evaluate coverage errors, a smaller number of countries (74%) used them for evaluating content errors.

Data Processing and Dissemination

Data Processing

The processing of census information includes all the tasks associated with taking the original information collected in the field and making it available to end users. Historically, this process includes data editing, coding, tabulation, verification, review, and the preparation and presentation of the census results (Shryock et al., 1971). Data capture techniques have changed considerably from the days when all data was captured through manual entries on paper forms. Optical mark reading (OMR), where scanners automatically detect tick boxes or other information; intelligent character recognition software that automatically recognizes handwriting; personal digital assistants (PDA) handheld devices for recording information; and telephone and internet technologies are all being used to assist in data capture (United Nations, 2009a). While some developing countries, like Mauritius in 2000, continue to use manual entry to capture data, others have adopted various data capture strategies that reduce costs and increase efficiency. In Ethiopia for example, OMR has effectively cut the time taken to process census data from 18 months in the 1994 census to around 4.5 months in the 2007 census (United Nations, 2009a).

Once captured, census data are typically edited (correction of errors) and coded (assignment of values) before final verification and presentation of the data. In the past, much of the editing and coding was performed by hand in the field office or in conjunction with machine-readable cards that changed or added missing information, based on a set of predetermined rules (Shryock et al., 1971). Much of this work has been replaced by computer-assisted methods for coding and editing data. Few countries solely use manual techniques for editing and those that do, tend to be concentrated in Latin America and the Caribbean (United Nations, 2010). Most use some combination of automated and manual editing techniques (United Nations, 2010). Many countries outsource various tasks during the processing and dissemination stages of the census. Confidentiality becomes an important issue with the outsourcing of data processing and many countries have begun to establish guidelines and policies for dealing with the confidentiality issues involved in censuses.

Dissemination of Data

The processing of data plays a critical role in connecting end users to the census. Generally, the dissemination of data is in the form of a print publication or some form of electronic media. The electronic media can take several forms, including devices such as DVD/CD, websites, online sources, and/or Geographic Information Services (GIS). A variety of software sources and electronic data tools also exist to provide users with one on one interaction with census data. Instead of simply providing printed publications, electronic interfaces with census data have emerged around the world. 'American Factfinder' in the US and 'NZ. Stat' in New Zealand, for example, give users an opportunity to interact with the data in ways that were not possible even as recently as the 1990s.

Censuses in Comparative Contexts

As the flow of people, goods, and ideas across national borders has rapidly expanded, governments and international organizations have become more aware of the potential benefits of using census data for comparative purposes across countries and regions. The UNSD has been a leading advocate for the collection and dissemination of population and housing census data in accordance with the UN's recommended principles (United Nations, 2008). Data harmonization is an important issue in the European Union where comparable data is crucial to regional policymaking and governance. In the 2010 census round, the European Parliament passed a number of regulations aimed at improving comparability of census data across member states. According to Eurostat, the agency which coordinates statistical activities across the EU, the legal and technical framework set up for the 2011 census far exceeds previous attempts to integrate data regionally (Eurostat, 2011: p. 10). Because of diverse approaches to census collection in Europe, the framework has focused on achieving harmonization of census output rather than input. One of the preconditions for comparability of census output is that every member state follows the same definitions and technical specifications, and publishes the same breakdowns. Each member state also has to provide cross-classified census data in the form of hypercubes to Eurostat, which allows for regional comparisons across a range of standard socio-demographic variables (Annex I, Regulation (EU) No 519/2010).

Another relatively recent innovation in the dissemination of census data is the systematic sharing of integrated individual-level micro data. IPUMS-International (Integrated Public Use Microdata Series-International) has the world's largest archive of publicly available census samples, holding more than 200 census samples from 68 countries, and dating back to the 1960 census round. The anonymous datasets are freely available to qualified researchers through a web dissemination system. Output harmonization is a key feature of the project, with micro data coded and documented consistently across countries and over time, to enable comparisons. Notwithstanding the significant advancements that initiatives like the EU framework and IPUMS-International have made with respect to facilitating census-based comparative research, some variables,

such as migration status, ethnicity, and race, continue to pose major challenges (Bell and Muhidin, 2009).

Future Prospects for Censuses

The national census has a long and rich history and continues to be widely viewed as an integral tool for national and regional governance and planning. In the twenty-first century, however, political and fiscal concerns, along with the demand for more timely data, have stimulated NSOs to explore census alternatives. There is little doubt that ongoing technological advances in data capture, management, and integration systems will further expand the possibilities open to governments, albeit within the constraints of legal requirements and confidentiality concerns (Coleman, 2013). Furthermore, as new methods in geospatial, post-enumeration, and statistical sampling techniques become more widely incorporated into national censuses, the need for new technologies and the demand for NSO staff skilled in these practices will also increase, both to process and interpret the results of these censuses (Coleman, 2013).

Growing pressures to harmonize data and standardize collection practices, along with concerns about the costs of administering a census and the need for more frequent data, means that countries will continue to adopt new strategies for counting their populations. However, the extent to which a country engages with new technologies or incorporates the most efficient and productive enumeration practices into the census, will continue to be heavily influenced by socio-political factors. The US is as an illustrative example of how politics may outweigh resources in the quest to develop new and better ways to enumerate the population. Despite having one of the most advanced NSOs in the world, with significant economic resources (the US spent roughly US$ 13 billion on the 2010 census alone) and access to advanced technologies, the adoption of new technologies has occurred much later in the US than in other industrialized, census-taking countries. The apparent reluctance to use the internet to supplement paper forms and to incorporate statistical adjustments into its census (which has been outlawed), means censuses in the US are in some ways anachronistic. Its geographic neighbor to the north, Canada, recently conducted over 50% of its census online and is a strong advocate of statistical adjustments for a variety of purposes. In 2011, Brazil completed its first entirely digital census through the use of PDAs. In the end, the politics of counting the population (Skerry, 2000; Anderson and Fienberg, 2001; Prewitt, 2003) may well trump the desire to count accurately.

See also: Censuses: Current Approaches and Methods; Censuses: History and Methods.

Bibliography

Anderson, M.J., Fienberg, S.E., 2001. Who Counts?: The Politics of Census-Taking in Contemporary America. Russell Sage Foundation Publications, New York.

Australian Bureau of Statistics, 2012. Fact Sheet: eCensus Take-up Rates. http://www.abs.gov.au/websitedbs/censushome.nsf/home/factsheetsecensus?opendocument&navpos=450 (accessed 13.04.13.).

Bailar, Barbara A., 2001. Censuses: comparative international aspects. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences, vol. 2. Elsevier, Amsterdam, pp. 1595–1599.

Bell, M., Muhidin, S., 2009. Cross-National Comparisons of Internal Migration. Human Development Research Paper 2009/30. United Nations Development Programme, New York.

Bentley, M., Mattingly, T., Hough, C., Bennett, C., 2003. Census Quality Survey to Evaluate Responses to the Census 2000 Question on Race: An Introduction to the Data. Census 2000 Evaluation.

Bryan, T., Heuser, R., 2004. Collection and processing of demographic data. In: Siegel, J., Swanson, D. (Eds.), The Methods and Materials of Demography, second ed. Elsevier Academic Press, Amsterdam, The Netherlands, pp. 43–63.

Coleman, D., 2013. The twilight of the census. Population and Development Review 38, 334–351.

Dugmore, K., Furness, P., Leventhal, B., Moy, C., 2011. Beyond the 2011 census in the United Kingdom. International Journal of Market Research 53 (5), 619–650.

Eurostat, 2011. Methodologies and Working Papers. EU Legislation on the 2011 Population and Housing Censuses: Explanatory Notes.

Ewbank, Douglas C., 1981. Age Misreporting and Age-Selective Underenumeration: Sources, Patterns, and Consequences for Demographic Analysis. National Academy Press, Washington, DC.

Jones, N.A., Smith, A.S., Black, A.A., Hawaiian, N., Indian, A., 2001. The Two or More Races Population, 2000, vol. 8(2). US Department of Commerce, Economics and Statistics Administration, US Census Bureau, Washington, DC.

Lenk, M., 2008. Methods of Register-based Census in Austria. Statistik Austria, Wien.

Mather, M., Pollard, K.M., Jacobsen, L., 2011. First Results from the 2010 Census. Population Reference Bureau, Washington, DC.

Morning, A., 2008. Ethnic classification in global perspective: a cross-national survey of the 2000 census round. Population Research and Policy Review 27 (2), 239–272.

Prewitt, Kenneth, 2003. Politics and Science in Census Taking. Russell Sage Foundation, New York.

Shryock, H., Siegel, J., Larmon, E., 1971. The Methods and Materials of Demography. US Department of Commerce, Bureau of Census, Washington, DC.

Simon, P., 2007. "Ethnic" Statistics and Data Protection in the Council of Europe Countries. European Commission against Racism and Intolerance, Strasbourg.

Skerry, P., 2000. Counting the Census?: Race, Group Identity, and the Evasion of Politics. No. 56. Brookings Institution Press.

Statistics New Zealand, 2012. Transforming the New Zealand Census of Population and Dwellings: Issues, Options and Strategies. Statistics New Zealand, Wellington, NZ.

Stukel, D., 2008. Projected Census Dates, Funding Requirements and Sources, and Technical Assistance Needs for the 2010 Round of Population and Housing Censuses. United Nations Statistics Division, New York.

United Nations, 1998. Principles and Recommendations for Population and Housing Censuses. Department of Economic and Social Affairs, Statistics Division, New York.

United Nations, 2008. Principles and Recommendations for Population and Housing Censuses, Revision 2. Series M No. 67/Rev.2. United Nations, Department of Economic and Social Affairs, Statistics Division, New York.

United Nations, 2009a. Census Data Capture Methodology Technical Report. United Nations, Department of Economic and Social Affairs, Statistics Division, New York.

United Nations, 2009b. Handbook on Population and Housing Census Editing, Revision 1. United Nations, Department of Economic and Social Affairs Statistics Division, New York.

United Nations, 2009c. Post Enumeration Surveys Operational Guidelines Technical Report. United Nations, Department of Economic and Social Affairs Statistics Division, New York.

United Nations, 2010. Report on the Results of a Survey on Census Methods Used by Countries in the 2010 Census Round. United Nations Statistics Division, New York.

United Nations, 2012. Report of the Committee on the Elimination of Racial Discrimination, Eightieth Session, 13 February–9 March 2012. General Assembly Official Records Sixty-seventh Session Supplement No. 18 (A/67/18).

United Nations Economic Commission for Europe, 2006. Conference of European statisticians recommendations for the 2010 censuses of population and housing. United Nations, New York and Geneva.

UNECE Secretariat, 2010. Main results of the UNECE-UNSD survey on the 2010–2011 round of censuses in the UNECE region. United Nations Economic and Social Council, Geneva.

U.S. Census, 2010. The Development and Sensitivity Analysis of the 2010 Demographic Analysis Estimates by the Demographic Analysis Research Team. U.S. Census Bureau, Washington, DC.

Valente, P., 2010. Census taking in Europe: how are populations counted in 2010? Population & Societies 467, 1–4.

Censuses: Current Approaches and Methods

Paolo Valente, United Nations Economic Commission for Europe, Geneva, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

The views expressed in this publication are those of the contributor, and do not necessarily reflect the views of the United Nations.

Abstract

The population census is the official counting of a population in the territory of a country and the collection of information on selected demographic and socioeconomic characteristics of the population. Almost all countries conduct a population census at least once every 10 years, often together with a housing census, which is the official counting of the housing stock and collection of housing characteristics. The traditional approach to the population census consists in the registration of all individuals and their details using paper questionnaires during a field operation that normally lasts a few days or weeks. Some countries conduct censuses using data from population registers and other administrative sources, without collecting data in the field. Recently, various alternative census approaches have been developed, making use of register data combined with field data collection and other sources, as well as partial annual field data collection and other sampling methodologies.

Definition/Overview

The population census is the operation that produces the official counting of the population within the territory of a country and its smallest geographical subterritories, as well as the registration of selected demographic and socioeconomic characteristics of the population. Almost all countries conduct a population census at least once every 10 years, frequently together with a housing census, which is the official counting of the housing stock and collection of selected housing characteristics.

The population census plays a key role in the administration of a country as its results are generally used for many important purposes, e.g., allocating funds to regional governments and municipalities or defining electoral districts. Census results also influence key policy decisions at the national and local level with regard to economic development, infrastructures, provision of education, health, and other services to the population.

Population censuses are also important for national statistical systems. In fact, the census not only provides the benchmark for the population count at national and local levels, but also represents the best or sometimes the only source of data on small population groups. Moreover, censuses allow the creation of a solid framework that can be used to draw samples from all other surveys conducted in the country. The role of the population census is even more important in countries where other vital statistics are not complete or accurate: in these countries, the census is the only statistical source providing correct and detailed estimates of the population size and structure.

The Traditional Census

The methodology used to conduct a census has been basically the same for centuries, starting in the ancient world. Evidence suggests that censuses were carried out already around 4000 BC by the Babilonians. At that time censuses were taken every 6–7 years to take record – on clay tiles – of the number of people and livestock, and to know how much food was needed for the population. In Egypt census were taken from around 2500 BC and the results were used – among other purposes – to plan how to share out the land after the annual flooding of the Nile and to calculate the labor force available to build the pyramids (ONS United Kingdom, 2013; CSO Ireland, 2013). Censuses were conducted regularly also in China and in ancient Rome (the word census comes from the Latin word 'censere' which means 'to estimate'), by recording all individuals and their details through the completion of population lists. More recently, census forms replaced the lists, but the information is still collected in the field across the whole country in a relatively short period of time, normally within a few weeks.

Traditional censuses in modern times used two main alternative methods of enumeration. In the first method, census enumerators (temporary staff recruited for the census field data collection) are responsible for collecting the information from the households with the help of an interview and the filling out of census forms. This approach is usually adopted in countries where a relatively high proportion of the population has minimal standards of education or is illiterate. This method is particularly expensive because enumerators have to conduct interviews in addition to delivering and collecting the forms.

In the second method, called 'self-enumeration,' forms are filled in by the members of the household. The enumerators deliver the forms to the households and collect them some days later. In some countries where this method is adopted, the postal system is used instead of enumerators for the delivery and/or the collection of the forms. More recently, some countries have offered the option to fill in the census questionnaires on the Internet, as an alternative to filling the paper forms.

Some countries use two types of questionnaires to reduce the amount of information to be processed. They supply only a sample of the population with an extensive long form, while a short form is used for the rest of the population. This method has the advantage of providing extensive details about the population, yet reducing the amount of time and labor to process the collected information. This also limits the complexity and costs of the census operations. However, since the information collected with long forms only represents a sample of the population, census results on detailed

 http://dx.doi.org/10.1016/B978-0-08-097086-8.41006-8

characteristics (covered in the long form) are not available for small areas or small population groups. This approach has been used in the United States and Canada since the 1970s. For the 2010 US census, the long form was replaced by a large household sample survey (the American Community Survey or ACS) that is conducted every year and provides detailed demographic, social, and economic data about households. As a result, the US census model for the 2010 census is based on a decennial traditional enumeration – conducted in 2010 using only a short form – with yearly updates of the population characteristics on a sample basis provided by the ACS.

Although the concept of the traditional census is relatively simple, its implementation is a huge and very complex operation that requires significant financial resources, the participation of various administrations at central and local levels, as well as the recruitment and training of a large temporary work force hired as census field staff (enumerators, supervisors, etc.).

From the point of view of the census management, there are also a number of problems and issues when it comes to the traditional approach. First of all, recruiting and training a very large temporary work force to collect data implies huge costs. But there are also other activities required for a traditional census that are very expensive, like printing, distributing, collecting a huge number of census forms, entering the data (manually or using scanners), and processing them. Moreover, in a traditional census, the costs are concentrated around the year when data are collected and are much lower in the other years of the census cycle. This cost distribution over time can create management and budgeting problems.

From the point of view of the organization of the fieldwork operation, a census conducted with the traditional approach normally requires the cooperation of various administrations at the national and local level (including government departments, cartographic agencies, postal administration, municipalities), which may pose problems in some countries, especially if the budget does not fully cover the census expenses, or if the respective tasks and responsibilities of the various administrations involved are not clearly specified.

An emerging problem reported by countries using the traditional census model is the increasing difficulty to enumerate certain population groups, particularly those characterized by high mobility and multiple residences (including young professionals, students, workers, retired people, or other categories that commute regularly between two or more places). In general, it can be difficult to find these persons at home in order to deliver or collect the census forms. Moreover, identifying the place of usual residence for these people is often complicated.

Several countries have reported an increasing reluctance of the population (in particular among specific subgroups) to participate in the census. This can be due to various reasons, including reluctance to open the door for security reasons, in particular by old people or in areas with security problems; distrust toward statistical institutes or the public authorities in general; fear that the information collected could be used for purposes other than statistical uses; and reluctance to provide information that is already available in registers or other administrative sources.

Finally, there are also some problems with the outputs produced by traditional censuses. Publishing census results often takes a long time because of the large amount of information that is being processed. The frequency of the results of a traditional census is often not considered to be acceptable. In fact, the traditional census is normally conducted every 10 years due to its costs and complexity. However, there is a certain necessity from various users to obtain more frequently updated data than every 10 years.

The Register-Based Census

The most common alternative to the traditional census approach is the so-called register-based census. With this method, traditional field enumeration is replaced by the use of administrative data coming from various registers and administrative sources (population registers, building registers, social security registers, business registers, student/education registers, etc.) through a matching process. The matching is normally done making use of a personal identification number (unique for each individual) that should be included in the various registers. By this method, it is possible to match demographic, education, economic, and social data on individuals existing in a large number of registers (UNECE, 2007).

This approach, developed by several Nordic countries in the 1970s and adopted for the first time in Denmark in 1981, permits the production of census data at a limited cost and with relatively limited work, once a good quality system of statistical registers has been set up. This approach has the advantage of placing no burden on individuals and updated data are potentially available every year. Moreover, there is no cyclic distribution in the costs and census staff, as they are distributed relatively evenly across time.

It should be noted that setting up and maintaining a statistical system based on registers requires significant initial investments and a very long development time. Moreover, this approach requires good cooperation between the authorities responsible for the registers and national statistical institutes, because data have to be transmitted from the former to the latter following specific formats and deadlines. Legislation that allows using administrative data for statistical purposes and matching records across registers is also required, as well as acceptance by the public of such a system.

Another disadvantage of this approach is that the details to be collected are limited to those available in registers, and the quality of the data depends on the coverage and quality of the registers themselves. However, statistical agencies can combine data from different registers to assess and increase quality and derive new variables. Another way to provide data on variables that may not be available from registers is to make use of results from existing household sample surveys. An example is the so-called Virtual census conducted in the Netherlands in 2001 and 2011, where register data were integrated with results from the labor force survey (LFS) on some of the economic characteristics that were not adequately covered in the registers. This approach allows producing census data without field data collection, but the detail of the information derived from sample surveys is clearly limited.

The Combined Census, Based on Data from Registers and Field Data Collection

Many countries have registers with population and other data that could potentially be used for a census, but the coverage and quality of the data are not sufficient for producing reliable census data. Sometimes key census variables (for instance, on occupation, educational attainment, or the composition of households) are not available. Some of these countries have used data from registers in combination with data from exhaustive or partial (sample) field data collection in order to produce census results.

Data from the registers can be combined with data from sample surveys conducted ad-hoc to fill in the gaps or confirm already existing data. Hence, the purpose of additional surveys is usually to evaluate the accuracy of population registers and/or to collect information on topics that are not covered in registers, or for which the coverage and quality of registers is not sufficient. The method has the advantage that costs and burden on individuals for providing information are limited, because surveys are just submitted to a sample of the population. Moreover, by comparing the data of the sample surveys with those from the registers in the same areas it is possible to assess the accuracy of the registers, and possibly make adjustments to the population counts derived from the register. This method was adopted in 2008 by Israel for the first time and in 2011 by various other countries in Europe.

Another approach to combined censuses is based on using data from registers in combination with full enumerations in the field. The field enumeration provides information about the coverage and accuracy of the data available in the registers, and allows collecting information on variables not available in registers. This approach is more expensive than the previous one because the whole population is counted, but it is in general less expensive than a traditional census, because of the efficient use of data available in registers. Compared to a register-based census, this method is clearly much more expensive and poses response burden on the public, but on the other hand it provides improved precision of the results and may help to increase the coverage and quality of registers. For this reason, this method is sometimes used to transition from a traditional to a fully register-based census.

Often, when a combined census approach is used – with either samples or full field data collection – the respondents have the opportunity to fill in web-based census questionnaires. The Internet option is generally considered to be more beneficial, especially in terms of improving/maintaining the census participation among population groups that would be difficult to enumerate in a traditional way; responding to social/public expectations; improving data quality; and reducing costs if a significant proportion of the respondents fill the questionnaires on the Internet.

The Rolling Census

With this method, the census is conducted by a cumulative and continuous (or 'rolling') field data collection over long periods of time, rather than during a relatively short time period as in the traditional census. This is the only alternative census methodology that does not require using data from population registers or other administrative data. This method was developed in theory in the 1970s, and first implemented in France starting in 2004 (see later). Under this method, data are collected in the field every year in the various municipalities on a rolling basis. In France, for instance, a 5-year cycle is used. The census results are based on rolling averages calculated over the 5-year cycle and are updated yearly. This method provides for improved frequency of data and spreads out across time the financial and human burden associated with censuses. However, this method is complex and hard to implement. Complications may arise from movements of persons across municipalities over the various years. These movements could potentially lead to double counting or missing certain individuals, although specific mechanisms have been put in place to deal with these cases.

Global and Regional Trends about Census Methodology

As mentioned earlier, historically and until the 1970s, population censuses were conducted in all countries across the world exclusively using the traditional approach. During the 1970s, some countries in Northern Europe started developing the register-based census. The first register-based census was carried out for the first time in 1981 in Denmark, followed by other Nordic countries in the 1990s. Outside Europe, Singapore carried out in 1990 a combined census, based on data from registers and field data collection. However, basically all other countries in the world were still using the traditional census approach until the 1990s.

For the 2000 census round (covering the period from 1995 to 2004, according to the United Nations), the number of countries using alternative census methods increased significantly in Europe: three countries conducted a register-based census (without field data collection) and five countries conducted a combined census. In total, 8 European countries out of 38 that conducted a census (21%) adopted an alternative method. Outside Europe, in the 2000 round only few countries including Bahrain and Singapore conducted a combined census, and the traditional approach remained the dominant method.

In the 2010 census round (covering the period from 2005 to 2014), the trend toward the adoption of alternative census methods became even stronger in Europe. The number of European countries conducting register-based censuses passed from three in the 2000 round to nine in the 2010 round, and those conducting a combined census from five to eight. If we also include France with its new rolling census, then the total number of countries that adopted an alternative census method more than doubled from 8 in the 2000 round (21%) to 18 in the 2010 round (42%). Outside Europe, Israel and Turkey joined Bahrain and Singapore among the few countries conducting a combined census in the 2010 round, but the traditional approach continues to be the method adopted by almost all countries, in some cases with Internet response or other innovations.

Selected National Census Methodological Practices

A selection of census methods adopted by countries in the 2010 census round, including traditional, combined, register-based, and rolling censuses, is presented. Several of the examples presented refer to European countries since the proportion of countries adopting alternative census methods is particularly high in this continent. In this regard, it should be noted that the European Union member states were free to decide the methodology, sources, and technology to be used for their census. In fact, the EU census legislation was concerned with the census dates (the reference date had to be in 2011 for all EU countries) and the harmonization of the census output rather than the input, including methodology and sources (EU, 2011).

United States: Traditional Census with Yearly Updates (2010)

Since the first census taken in 1790, the United States has been conducting censuses regularly, and often introduced important innovations. In 1890, punch cards and machine tabulation of responses were introduced. In 1940, sampling was introduced: long forms were used to collected detailed information from about 15% of the population, while the rest of the population filled a short form with only limited information. In 1950, tabulation was made using a dedicated computer. In 1970, the census became mainly a mail enumeration: since then the large majority of the population receive and return the forms by mail, and enumerators are used only for nonresponse follow-up (US Census Bureau, 2013a).

For the 2010 census, the enumeration was conducted using only a short form asking 10 questions (name, gender, age, race, ethnicity, relationship, and whether the home was owned or rented). Approximately 74% of the households returned their census forms by mail; the remaining households were counted by census workers throughout the United States. Detailed information on education, housing, jobs, etc. that in previous censuses was collected with long-form questionnaires, in 2010 was collected by the ACS, a large annual household survey launched in 2005, covering approximately 3 million addresses annually. Participation in the ACS, as in the 2010 Census, is mandatory by law (US Census Bureau, 2013b).

Canada: Traditional Census with Internet as Primary Response Channel (2011)

In Canada the first modern census was conducted in 1871, and then censuses were carried out regularly every 10 years. In 1941, sampling was used for the first time: one-tenth of the households were asked additional questions on housing. This use of sampling was considered effective, yielding high-quality data while reducing costs and response burden. In 1951 and in subsequent censuses, the sample was expanded to one household in five in order to obtain greater geographic detail. Starting in 1956, mid-decade censuses were conducted regularly. In 1971, self-enumeration (i.e., respondents complete their own questionnaire) replaced interview as the main enumeration method, and sampling was extended to include population data, using 'short form' and 'long form' (see earlier text). In the 2006 census, forms were delivered by Canada Post to about 70% of households, and for the first time households could complete their questionnaire online (with a 18.5% take up rate) (Statistics Canada, 2013).

For the 2011 census, major innovations were introduced. The questions normally asked on the census long form were asked in the National Household Survey. Questionnaires for this voluntary sample survey were distributed approximately 4 weeks after the census. All households were asked to complete 10 questions on basic topics (the 'short form'), and Internet response was promoted as primary enumeration method, with the objective to ensure better quality data and reduce costs. The majority of households received by post a letter replacing the traditional paper questionnaire, to encourage respondents to complete the questionnaire online. The paper questionnaire was still available on request. The Internet response target was set to 40%, but the final take up rate was even higher: 54% of households responded by Internet (Statistics Canada, 2012).

Israel: Combined Census (2008)

Israel conducted six censuses since its establishment. The first one was carried out in November 1948 and was used as the base for the Central Population Register (CPR). The four censuses that followed were conducted as classic censuses in which the whole country's area was canvassed, all possible dwelling facilities were enumerated, and a demographic short questionnaire was completed by the interviewer for 100% of the residents. A short socioeconomic questionnaire was completed for one-fifth of the households. All of the five first censuses were conducted with face-to-face interviews, using paper questionnaires (CBS Israel, 2012).

In 2008, Israel carried out an innovative combined census (called 'Integrated census'), consisting in a register-based census complemented by two sample surveys aimed to evaluate the quality of the register and to supply parameters for calculating corrections to the data from the CPR. The approach was intended to combine the advantages of both traditional and register-based censuses. The advantages of this approach consisted in reducing the burden on respondents (as only part of the population had to answer the questionnaire) and improving the quality of information available in registers (CBS Israel, 2013).

Turkey: Combined Census (2011)

The first census of Turkey was carried out in 1927. The following population censuses were carried out regularly every 5 years between 1935 and 1990, and then in 2000. The traditional method was used, with the particularity that data were collected in 1 day by application of the curfew. The enumerators visited all the households in the country and filled in the questionnaires by face-to-face interview. The population was enumerated on the basis of the place where each person was physically present on census day (*de facto* population).

The 2011 census was based on a combined method, using data from population registers and conducting a register-based large-scale sample survey. The backbone of the census was the Address-Based Population Registration System, established in 2007 in order to record and update information on the

population of localities, and to register population movements. The sample survey provided information that could not be derived from registers. The survey covered 13.5% of the total households, was conducted with interviews, using paper questionnaires and netbooks (Computer Assisted Personal Interview, or CAPI), and the duration of the data collection was 2 months (Ural, 2012).

Estonia: Combined Census with Internet as Primary Response Channel (2011)

The first population census of the Republic of Estonia took place in 1922, with a very exhaustive questionnaire. The data collection lasted 2 days but data processing took one and a half years, even though two special tabulating machines, designed specifically for population censuses, were used. Other censuses were carried out in 1934, 1941 (during the German occupation), 1959, 1970, 1979, and 1989 (as part of the Soviet Union censuses, with data processed centrally in Moscow). After the restoration of independence, the first census was carried out in 2000, with a traditional method using a household and dwelling questionnaire and a personal questionnaire. Processing of the results was a major challenge for Statistics Estonia, as there was no prior experience of processing census datasets in the country (Statistics Estonia, 2013).

The population and housing census of 2011 was based on a combined methodology and the intensive adoption of modern technologies. The method consisted of self-enumeration via Internet (e-enumeration) and traditional face-by-face interview using laptops. Paper questionnaires were considered only as a backup for emergency situations. Selected variables were taken from registers (when coverage and accuracy were considered sufficient) and used for prefilling the questionnaires, to be then verified by the respondents, and corrected if necessary. In the period before the census, an intensive publicity campaign was launched, motivating the population to use the Internet response option. The response of the population was exceptional, and the participation in the Internet census exceeded the projections, reaching almost 67%. The nonresponse follow-up was carried out by enumerators using laptops instead of paper questionnaires. The goal of the new solutions adopted was to ensure a bigger coverage for the census, better quality, and faster publication of the results (Beltadze, 2013).

Poland: Combined Paperless Census with Extensive Use of Technology (2011)

The first census in Poland was carried out in 1789, and the results were used to collect a tax covering the costs of maintaining a permanent army. Other censuses were carried out in the following decades more or less at regular intervals. In 1918, the Central Statistical Office was established, and censuses were then carried out at regular intervals (except for the period of World War II) between 1921 and 2002, with a traditional method. From the 1970 census, data have been captured, processed, and stored electronically (Pappelbon, 2011).

For the 2011 census, a new very innovative model was adopted, based on the combined use of data from about 30 administrative registers and data obtained from respondents (via Internet or with direct or telephone interviews), exclusively with the use of electronic questionnaires prefilled with data from registers. As a result, paper questionnaires were eliminated altogether, and replaced by IT solutions. GIS technology was used for the census preparatory work and for on-going census process monitoring. The new census model, that was considered very successful, was adopted to minimize census costs and burden on respondents (considering increasing nonresponse in statistical surveys) and to take advantage of the intensive development of IT systems of public administration, based on advanced technologies (Dygaszewicz, 2013).

Netherlands: Register-Based Census Using also Data from Existing Surveys (2001 and 2011)

In the Netherlands, the first census was held in 1795 for the purpose of establishing voting constituencies. After the establishment of Statistics Netherlands in 1899, six traditional censuses were carried out between 1909 and 1971. At the time of the 1971 census, the willingness of the population to participate had fallen sharply due to privacy concerns, and the decision was taken not to conduct traditional censuses anymore. Starting in 1981, the censuses were based on registers and surveys available at Statistics Netherlands. The data that have been compiled in 1981 and 1991 were of limited character, while in 2001 Statistics Netherlands developed for the first time the so-called virtual census, that was then repeated in 2011.

The virtual census uses a combination of available registers and surveys. The backbone of the census is the CPR, which is the combination of all municipal population registers. Other registers and databases are also used, including the Social Statistical Database with data on employees and self-employed. Different variables that are not available from registers, such as occupation and level of education, are obtained from the LFS or other surveys. To ensure numerical consistency in the set of census tables between data from registers and sample surveys, a new estimation method called 'repeated weighting' was developed, based on the repeated application of a regression method. According to Statistics Netherlands, the virtual census is cheaper (costing only about 1% of the traditional census in the Netherlands, but excluding the costs of the registers), comparable with earlier Dutch censuses, and more socially acceptable. However, for the variables derived from sample surveys the level of detail is clearly lower compared to a traditional census due to the use of sampling (Schulte Nordholt et al., 2004).

Austria: Direct Transition from Traditional to Register-Based Census (2011)

In Austria, censuses were carried out already since the eighteenth century, mainly to serve military and fiscal purposes. Modern censuses, aimed at serving the needs of the administration of the state, were carried out from 1869 at regular intervals. After World War II, censuses were conducted from 1951 every 10 years, covering population and housing. In the 1990s census taking became increasingly controversial, due to high costs, low frequency, and increasing reluctance in filling in forms. As a result, it was decided that the 2001 census would be the last traditional census.

In 2011, Austria carried out its first register-based census (after a test census in 2006), using information taken directly from existing administrative registers. The CPR represents the backbone of the register-based census. Other registers used include the Housing Register of Buildings and Dwellings, the Business Register of Enterprises and their Local Units, and the Register of Educational Attainment – all of which are maintained by Statistics Austria – as well as the Central Social Security Register, the Unemployment Register, and the Tax Register. The new register-based census of Austria, compared to the traditional census, is expected to be less expensive and to provide data sooner and with higher frequency (tentatively every 5 years) (Statistics Austria, 2013).

France: Rolling Census (Implemented Since 2004)

In France, the first census of modern times was carried out in 1801, under Napoleon Bonaparte. Subsequent censuses took place every 5 years until World War II, and then at irregular intervals (every 6–9 years) between 1946 and 1999, using the traditional method. Since 2004, the new rolling census was introduced for the first time, with the objectives to meet the expectations of census users (particularly in terms of frequent updates of the results), and facilitate census management in terms of constant budget over time.

As mentioned earlier, the rolling census is based on a cumulative and continuous (or 'rolling') field data collection over long periods of time, rather than during a relatively short time period as in the traditional census. In France, a 5-year cycle is used, with two different strategies to collect data of small municipalities (population under 10 000) and large municipalities. Small municipalities are divided into five groups, and a full enumeration is conducted each year in one of the groups. In large municipalities, a sample survey covering 8% of dwellings is conducted each year. At the end of the 5-year cycle, the population of small municipalities (which in France amounts up to about half of the total population) is fully enumerated, in addition to about 40% of the population of large municipalities. In total, about 70% of the country's population is enumerated. This is considered to be sufficient to guarantee robust information at the municipal level.

The census results are based on rolling averages calculated over the 5-year cycle and are updated yearly. Since the data collection for the French rolling census started in 2004, the first results for the population at the national level were based on data collected in the 5-year period 2004–08, with 2006 used as reference year being the central year of the period.

See also: Census Microdata; Censuses, Population: Comparative International Aspects; Censuses: History and Methods; Data Collection: Interviewing.

Bibliography

Beltadze, D., 2013. Enumeration via Internet - Estonian experience. Paper submitted to the UNECE-Eurostat Meeting on Population and Housing Censuses, Geneva, 30 September–3 October 2013. http://www.unece.org/stats/documents/2013.10.census1.html (accessed 22.11.13.).

CBS Israel, 2012. The integrated rolling census in Israel. Paper submitted to the Sixtieth plenary session of the Conference of European Statisticians, Paris, 6–8 June 2012. http://www.unece.org/stats/documents/2012.06.census.html (accessed 22.11.13.).

CBS Israel, 2013. 2008 Census – the Integrated Census. http://www.cbs.gov.il/census/census/pnimi_sub_page_e.html?id_topic=2&id_subtopic=1 (accessed 22.11.13.).

CSO Ireland, 2013. The Census through History. http://www.census.ie/-in-history/the-census-through-history.150.1.aspx (accessed 22.11.13.).

Dygaszewicz, J., 2013. New Technologies Used in 2010 Census Round – Polish Case Study. http://www.unece.org/stats/documents/2013.10.census1.html (accessed 22.11.13.).

EU, 2011. EU Legislation on the 2011 Population and Housing Censuses - Explanatory Notes. Publications Office of the European Union, Luxembourg.

ONS United Kingdom, 2013. Census-Taking in the Ancient World. http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/about-censuses/census-history/census-taking-in-the-ancient-world/index.html (accessed 22.11.13.).

Pappelbon, J., 2011. IPUMS Europe - country Report: POLAND, paper submitted to the 11th IPUMS International Workshop "Integrating Global Census Microdata", Dublin Ireland, 20–21 August 2011; http://www.hist.umn.edu/~rmccaa/ipums-global/poland_ipums_dublin_workshop.pdf (accessed 22.11.13.).

Schulte Nordholt, E., et al., 2004. The Dutch Virtual Census of 2001-Analysis and Methodology. Statistics Netherlands, Voorburg/Heerlen. http://www.cbs.nl/nr/rdonlyres/d1716a60-0d13-4281-bed6-3607514888ad/0/b572001.pdf (accessed 22.11.13.).

Statistics Austria, 2013. Register-Based Census. http://www.statistik.at/web_en/surveys/register_based_census/index.html (accessed 22.11.13.).

Statistics Canada, 2012. Changes in the Canadian census of population, paper submitted to the Sixtieth plenary session of the Conference of European Statisticians, Paris, 6–8 June 2012. http://www.unece.org/stats/documents/2012.06.census.html (accessed 22.11.13.).

Statistics Canada, 2013. History of the Census of Canada. http://www12.statcan.gc.ca/census-recensement/2011/ref/about-apropos/history-histoire-eng.cfm (accessed 22.11.13.).

Statistics Estonia, 2013. History of Population Censuses in Estonia. http://www.stat.ee/62931 (accessed 22.11.13.).

UN-ECOSOC, 1995. Resolution 1995/97 of the United Nations Economic and Social Council – Ecosoc.

UNECE, 2007. Register-Based Statistics in the Nordic Countries – Review of Best Practices with Focus on Population and Social Statistics. United Nations Publications, Geneva and New York. Sales No. E.07.II.E.11. http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf (accessed 22.11.13.).

Ural, E., 2012. The 2011 Population and Housing Census of Turkey. Paper submitted to the UNECE-Eurostat Expert Group Meeting on Censuses Using Registers, Geneva, 22–23 May 2012; http://www.unece.org/stats/documents/2012.05.census1.html (accessed 22.11.13.).

US Census Bureau, 2013a. History of US Censuses through the Decades. http://www.census.gov/history/www/through_the_decades/overview/ (accessed 22.11.13.).

US Census Bureau, 2013b. How We Count America. http://www.census.gov/2010census/about/how-we-count.php (accessed 22.11.13.).

Relevant Websites

http://unstats.un.org/UNSD/Demographic/sources/census/2010_PHC/default.htm – 2010 World Population and Housing Census Programme. Website maintained by the United Nations Statistics Division.

http://www1.unece.org/stat/platform/display/censuses/2010+Population+Census+Round – 2010 Population Census Round – Information and Material for UNECE Countries. Web page maintained by the United Nations Economic Commission for Europe.

http://www.unece.org/stats/census.html – Population and Housing Censuses. Web page maintained by the United Nations Economic Commission for Europe.

http://unstats.un.org/UNSD/Demographic/sources/census/censusdates.htm – Census Dates for all Countries. Web page maintained by the United Nations Statistics Division.

Censuses: History and Methods

Margo Anderson, University of Wisconsin – Milwaukee, Milwaukee, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A census is a complete count of a population, generally a nation, conducted by government, as of a fixed date. The periodic, national population census became a common function of Western governments around 1800 and is now routine throughout the world. Government agencies contact households to provide information for each member on a small number of questions such as age, sex, relationship to head. The agency tabulates responses and publishes the results. Data serve a variety of functions, including political apportionment, program planning, funds allocation, marketing and forecasting analyses. Censuses are expensive, logistically and technologically complex public activities, and serve to focus public attention on demographic patterns and change.

A census is a complete count of a population of a geographic area, generally a nation or country, conducted by the government. It is the most common method used for population measurement. Other methods include population registers and sample surveys. While governments have conducted censuses since ancient times, the regular periodic national population census only became a common function of Western governments in the late eighteenth and nineteenth centuries. The practice is now routine throughout the world. The basic historical continuity in census practice over time is straightforward. Government statistical agencies conduct censuses by contacting the household reference person in every residential household in person, or by postal mail or electronic means. The respondent provides information for each household member on a small number of questions such as age, sex, relationship to head, and other demographic markers. Individual-level data are generally considered confidential. The government statistical agency tabulates the responses and publishes the results. Data are used for a wide variety of public functions, including political apportionment, program planning, and funds allocation. The private sector uses census data for planning and marketing. Social science researchers and the media use the data to monitor demographic trends and illustrate social change. Within this broad outline, improvements in method and use result in more accurate, detailed, and timely data over time, and in shifting political and scientific demands on the particular census.

Definition/Overview

A census is a count of the population of a country as of a fixed date. National governments conduct censuses to determine how many people live in different areas of the country, whether the population is growing, stable, or declining in the country as a whole or in particular parts of the country. They also determine what the characteristics of the population are in terms of age, sex, ethnic background, marital status, or income. The traditional method is for governments to collect the information by sending a questionnaire in the mail or an interviewer to every household or residential address in the country. The questionnaire asks the head of the household, or a responsible adult in the household (the respondent), to list all the people who live at the address as of a particular date, and answer a series of questions about each of them. The respondent or the interviewer is then responsible for sending the answers back to the government agency, which in turn adds up the results, or tabulates or aggregates the answers from the country overall and for political subdivisions such as states or provinces, cities, counties, or other civil divisions. The agency usually reports the results to the public a few months or years after the census; the results are considered 'news' and are reported in the media. Since censuses aim to count the entire population of a country, they are very expensive and elaborate administrative operations, and are conducted relatively infrequently, generally at 5 to 10-year intervals.

Between censuses, governments estimate the size and characteristics of the population, either by extrapolating the trends in the census into the future, estimating the population from other data systems such as tax or drivers license records, or by conducting periodic sample surveys to collect information about the population. Representative probability samples collect information from a small portion of the population, and thus can be conducted frequently, even monthly. In the USA, the Current Population Survey of around 60 000 households is conducted monthly. National governments also conduct other types of censuses, particularly of economic activity, such as an agriculture, manufacturing, or business census. Such censuses collect information on the number and characteristics of farms, businesses, or manufacturing firms. In the nineteenth century, such censuses were conducted at the same time as the population census. Today, the economic censuses are generally conducted on a different schedule from the population census.

History

Censuses have been taken since ancient times by emperors and kings trying to assess the strength of their realms. These early censuses were conducted sporadically, and generally served to measure the tax or military capacity of a particular area. They tended to count adult men, men liable for military service, or people liable to tithe (pay taxes). For census taking of an entire

 http://dx.doi.org/10.1016/B978-0-08-097086-8.41007-X

population to become feasible, a uniform unit of analysis was required. Hence census taking in the West had to await the emergence of the concept of commensurate households, generally seen as a development of the Medieval European West (Herlihy, 1985). The household or family served as the unit or analysis or the locus for counting the members within it.

Generally speaking the modern periodic census of all persons is an invention of the early modern period in the European West and, particularly from a New World perspective, was associated with efforts by the home country to determine the success of the overseas colonies. Thus the British Crown and the British Board of Trade ordered repeated counts of the colonial American population in the seventeenth and eighteenth centuries, starting in the 1620s in Virginia (Wells, 1975; Cassedy, 1969). In Canada, French efforts to count the population began in 1665–1666 when Jean Talon came to the New World on behalf of Louis XIV and took a census of what became Quebec. Censuses were continued at irregular intervals after Canada became a British colony in 1763 (Worton, 1997). By the early nineteenth century, census taking began to be a regular feature of government in Western Europe and North America (Alterman, 1969; Nissel, 1987; Glass, 1973; Desrosieres, 1993; Patriarca, 1996). The International Statistical Congress (established 1853) and the International Statistical Institute (established 1885) proposed and promoted the uniform regular censuses for all national states. In the twentieth century, census taking spread throughout the world. Currently, the 2010 World Population and Housing Census Programme of the United Nations Statistical Office provides guidance and support for census taking worldwide.

Functions and Techniques

Censuses serve a variety of purposes in different countries. At a minimum a census provides a measure of the size of the population of a country, which can be compared with the population in the past, the population of other countries, and to make estimates of the likely population in the future. Governments use census information in almost all aspects of public policy, from determining how many children an educational system must serve, to determining where to put new roads. Censuses are also used to provide the denominators of other measures, e.g., measures of per capita income for a state or local area, or for such measures as crime rates or birth or death rates. Private businesses use census data for marketing analyses to determine where to locate new businesses, or to decide where to advertise particular products.

Other government agencies and private researchers use the census to provide the 'sampling frame' for other survey research. This is the address list that researchers use to determine where to conduct a sample or poll. In the USA, the US Census Bureau conducts the census during the 10 year of each decade. Each decade, the population count provides the data for reapportioning seats among the states in the House of Representatives and Electoral College, and for redrawing district boundaries for seats in the House, in state legislatures, and in local legislative districts. In Canada, a full census is taken during the first year of every decade and an abridged census is taken during the sixth year of the decade. The population data are also used to apportion seats among the provinces in the House of Commons and to draw electoral districts.

Most nations create a permanent national statistical agency to take the census, such as the US Census Bureau or Statistics Canada. The agency in charge generally undertakes a public review process to determine the questions to be asked. The questions vary from nation to nation depending on the particular political and social history and conditions of the country. Most censuses include basic demographic information such as the age, sex, educational background, occupation, and marital status of the individual. Race, ethnic or national origin, and religious affiliation are important questions in many nations. Further questions can include the person's place of birth, relationship to the household head; the individual's or the family's income; the type of house the household occupies; and whether the person is a citizen, has moved in the past 5 years, or speaks a particular language. Questions that are quite routine in one nation may be seen as quite controversial in another, depending on the history of the country. Americans do not ask questions on religious affiliation on the census since it is considered a violation of the First Amendment right to freedom of religion. Other nations, such as India, do collect such information. Questions on the number of children born to a woman were quite controversial in China in recent years because of the emphasis on the one-child policy of population limitation. In the USA, asking a question on income was considered controversial in 1940 when it was first asked. It is no longer as problematic. Questions change in response to public debate about the state of society. Americans wanted to know which households had radios in 1930, and introduced questions on housing quality in 1940. Canadians ask census questions on the unpaid work done in the home.

The standard nineteenth and twentieth century method for taking a census encompassed several phases. The first involves designing the procedures for the census, planning the questions to be asked, and created the field operation to staff the census. The statistical agency first divided the country into geographical subdivisions to be counted, and made maps and address lists, and prepared instructions for the local census takers. In most countries, this phase requires hiring large numbers of temporary workers (enumerators) to do the counting, or calling upon other government employees at a local level, such as schoolteachers, to conduct the count. The central statistical agency prepares and prints the questionnaires, and distributes them to households either through the mail, or by delivery by the enumerators.

During the second phase of the count a responsible adult or household head in every household, family, or equivalent institution is asked to fill out the form or respond to the enumerator and supply the required information about each member of the household. Questions for all people usually include a brief set, on a 'short' form, for example, name, age, sex, racial or ethnic status, marital status, and relationship to the household head. Interest in collecting more information on the characteristics of the population tended to burden the census form with additional questions over time, and thus increase the cost and logistics of the census. The development of probability sampling methods in the mid-twentieth century alleviated some of these issues and resulted in a shift in questionnaire design, which permitted census officials to sample

a portion of households to receive a more complicated or 'long' form questionnaire. A sample census questionnaire generally has many detailed questions on the individual's work status, income, housing, educational background, citizenship, and recent moves.

During this phase, either the person who fills out the form or an enumerator is responsible for making sure the completed questionnaire is delivered to the statistical agency for processing. The respondent may deliver the form by mail, or to the enumerator in person or by phone interview. In the 2010 census round, some countries have permitted respondents to answer the census electronically on the web.

In the third phase, the statistical agency enters the data onto a computer and adds up, or tabulates the responses for the nation, states, or provinces, and cities, towns, and other local jurisdictions. The agency also cross-tabulates the answers, reporting not merely on the number of people in a local area, but on the number of people, for example, in 5-year age cohorts, for each sex, for local areas. The agency only publishes the tabulated results of the count, and keeps the individual responses confidential. In the USA, individual census responses are stored at the National Archives. After 72 years, the original forms are opened to the public for use. These original responses are frequently used by people researching the history of their families.

Through most of the nineteenth century, the census tabulations were tallied by hand. Hand tabulation gave way to punch card and mechanical counter/sorter tabulations of census data from 1890 through the 1940s and electronic computer tabulation thereafter. Until the 1980s, statistical agencies published census results in large volumes of numeric tables – sometimes numbering in the hundreds of volumes. Since then, census results have become available electronically, on disc, magnetic tape, CD-ROM, or the Internet. The choice of census technique for a particular country is the result of its social and political traditions and technological capacities. The US census is highly automated, and is primarily conducted by mail. Canada sends enumerators to deliver the census form to each household; the household head fills it out and sends it in. Other nations use even more labor-intensive techniques of collecting and tabulating the data. In October 2010, for example, Argentina declared a national holiday and required that each household have a person available to answer questions from the legions of enumerators deployed that day.

National Traditions

The United States

The US census was mandated in the 1787 Constitution. At the time of the American Revolution, the framers faced many problems of uniting the 13 separate colonies into a national government, including how to allocate political representation among the states, and how to levy taxes.

The initial government structure under the Articles of Confederation gave each state one vote in Congress and required the states to collect taxes for the national government. This proved unsatisfactory since the states were of widely different sizes, tax capacity, and cultures. The 1787 Constitution derived the sovereignty of the state from the 'people of the US' and created a bicameral legislature with representation in the Senate based on the states; and representation in the House of Representatives based upon the population of each state. The Electoral College was composed of electors determined by summing the House and Senate members for each state. The decennial census was designed to provide the population figures for apportioning the seats in the House according to population. 'Direct taxes' levied on the states were also to be apportioned on the basis of population (Anderson, 1988).

At the time, a racially based slave labor system also existed, and almost 20% of the American population were enslaved African-Americans. The framers debated whether slaves were 'persons' or 'property' and thus whether states should receive representation for their slave populations. The southern states where slavery dominated did not consider the slave population for purposes of apportioning their state legislatures, but they did consider slaves property for tax purposes. The framers could not find an easy solution to this dilemma, and developed what came to be called the Three-Fifths Compromise, which 'discounted' the size of slave population as equivalent to 60% of the free population when determining the apportionment of the House. The Three-Fifths Compromise thus required a separate count of slave and the free, mainly white, population. The Constitution also specified that 'Indians not taxed,' that is, those Indians who were not considered part of the civil society were not to be counted in the census. The Three-Fifths Compromise was abolished with the abolition of slavery after the Civil War; the tradition of accounting for the various racial groups in the population continued after abolition.

The first census was taken in 1790. Assistant US marshals were instructed to travel the country and ask six questions at each household: the name of the family head; and for each household, the number of free white males aged 16 and older; the number of free white males younger than 16; the number of white females; the number of other free people (the free colored); and the number of slaves. They totaled the figures for their local jurisdiction and sent them to the US marshal for the state that in turn totaled the figures for the state and sent them to the President. The first American census counted 3.9 million people.

In later years, the census became more elaborate, with more questions asked, and more data published (Cohen, 1982). In 1850, Congress mandated a census schedule (form) with a line of questions for each person, including their name. A temporary Census Office, as it was then called, was set up in Washington to add up the tables of responses and publish the results in large volumes. By 1880, when the American population topped 50 000 000, the census was still being compiled by hand, using a primitive tally system. In 1890, the Census Office introduced machine tabulation of the responses, and each person's answers were converted to codes punched in Hollerith cards, a precursor to the IBM punch card. The cards were then run through counting machines. This innovation was the beginning of modern data processing and led to further innovations in tabulating large amounts of data. By the 1940s, the Census Bureau commissioned the construction of the first nondefense computer, UNIVAC, to tabulate the 1950 census. By the late 1950s, the Census Bureau dispensed with the punch cards and developed an electronic scanning system, called

FOSDIC, or Film Optical Scanning Device for Input to Computers, to read the answers on the census form and input the data to computer (Eckler, 1972).

In 1940, the US began to collect some census information from a sample of the population, and slowly shifted the detailed questions on the census to the long form asked of only about 15–25% of households. In 1970, the American census became primarily a mail enumeration, as the Census Bureau developed automated address files for the country. As of the year 2000, over 90% of the roughly 100 000 000 residential addresses in the US received the census form in the mail. If the Census Bureau did not receive a response, it sent an enumerator to determine if the address is correct and get the information from the household at the address.

In the late 1990s, the Census Bureau began to experiment with a new method of collecting long form information through 'continuous measurement' or 'rolling sample survey.' The American Community Survey was tested in the 2000 census and it replaced the census long form in the 2000s. Annual national-level household data became available in 2005. Each month, a sample of 250 000 households and group quarters receive an extensive mail survey and return the questionnaire by mail. If the respondents do not return the survey, the Census Bureau follows up with a personal or phone interview to complete the survey. The results are tabulated and results for large geographic areas are reported annually. Smaller jurisdictions receive results rolled up from the results of 3 or 5 years of data, depending on the size of the reporting area. The official population of the United States at the 2010 census was 308.7 million.

Canada

Regular decennial censuses began in 1851 in Canada. The British North America Act of 1867 required that the census provide population counts for the apportionment of representation in the House of Commons among the four provinces of Ontario, Quebec, Nova Scotia, and New Brunswick, and for the periodic readjustment of the boundaries of electoral districts. The first census of the Dominion was taken in 1871, and counted 3.7 million people.

The first Canadian census after Confederation was also a very elaborate affair, collecting information on agriculture, livestock, animal products, industrial establishments, forest products, shipping and fisheries, mining, and public institutions as well as on population. Canada was primarily an agricultural nation at the time, and the Department of Agriculture conducted the census. In 1905, the census bureau was made a permanent government agency in the Agriculture Department. In 1918, an independent Dominion Bureau of Statistics was created with oversight for taking the census and collecting other statistical information, headed by a Chief Statistician. In 1971, the bureau was renamed Statistics Canada. Canada has a centralized statistical system, with all major statistical work conducted in the same agency (Worton, 1997).

Canada began to conduct sample surveys along with the full census in 1941, introduced a quinquennial census (a census on the sixth year of the decade) in 1956, and self-enumeration in 1971. The population of Canada in 2011 was 33.5 million.

Great Britain

The first Parliamentary proposals for regular censuses of population in Britain date to the mid-eighteenth century, but they did not lead to the institutionalization of census taking until the end of the century. In late eighteenth century Britain, the publication of Thomas Malthus' *Essay on the Principle of Population* and concerns that the population was declining led Parliament to pass a bill "for taking an Account of the Population of Great Britain and of the Increase or Diminution thereof." The first count was taken in March 1801 by the Overseers of the Poor under the direction of former clerk of the House of Commons, John Rickman, and counted 9 million people. Census taking continued every decade afterward except for the cancellation of the 1941 Census during the Second World War. In 1837, the General Register Office was organized. The decennial census was made a responsibility of the GRO and used the administrative structure of the registration districts to organize the field portion of the count (Wrigley, 2011).

The 1841 census introduced individual-level enumeration. The mid-nineteenth century census was particularly influenced by the ongoing medical and public health work of the GRO leaders such as William Farr. It therefore developed detailed data on occupation, age, and locale for the denominator data required for the publication of mortality and other statistics (Glass, 1973; Nissel, 1987; Higgs, 2005). Machine tabulation of the census was introduced in 1911. The 1920 Census Act established a permanent legislative mandate for the census. Sampling was introduced in 1951 by way of the analysis and publication of results from a 10% sample of the complete 1951 forms. In 1961, the census used a short form for the entire population and an additional long form for a 10% sample. In 1966, a 1 in 10 sample count was taken. The results of the 1961 census were computer tabulated. By 1966, the Census Office had its own computer facilities. The population of the United Kingdom in 2011 was 63.2 million.

France

France established a population registration system in the late eighteenth century, and began a tradition of quinquennial census taking under Napoleon in 1801. The first census counted a population of 33 million. For much of the nineteenth century, the administration of the count was local with a small national office (*Statistique Generale de la France*) publishing the results tabulated locally. Individual forms were used in Paris starting in 1817; by 1876 the entire country used three standardized forms: one for the individual, one for the family or household, and a third for the building enumerated. Tabulation was done locally until 1896 when Lucien March introduced machine tabulation. The French pioneered in the analysis of occupational statistics derived from the census as a means of understanding the evolution of the economy from an agricultural and artisan basis to one characterized by professional qualifications and employees with credentials. In 1946, after World War II, the French National Statistical Agency was titled *Institut National de la Statistique et des Etudes Economiques* (INSEE) (Desrosieres, 1991).

In the 2000s, France restructured its population measurement system, and replaced the periodic census with a continuous measurement sample survey. Since 2004, local officials count the population on a rolling basis. In municipalities under 10 000 population, a 20% sample is collected annually; the entire population is counted every 5 years. In municipalities larger than 10 000, an 8% sample is surveyed annually, achieving 40% coverage over 5 years. INSEE tabulates and publishes annual reports from the samples. The population of France in 2011 was 65 million (Clanché, 2011).

Issues

Censuses provide important information about the population of a country, and can become embroiled in political or social controversy simply by reporting information relevant to ongoing issues in the society. Complaints about the census generally involve concerns about the accuracy of the count, the propriety of particular questions, and the uses to which the data are put.

Censuses require public understanding, support, and cooperation to be successful. Concerns about government interference with private life can prevent people from cooperating with what is an essentially voluntary counting process. People may be suspicious of giving information to a government agency, or may object that particular census questions are invasions of privacy.

When such public trust is lacking, people may fail to participate. In recent times, individuals doubled up in illegal housing units, undocumented immigrants who do not reside in the country legally, or individuals who do not wish to reveal their economic or social situation to a government agency, can be reluctant to respond to a census. In the most serious challenges, people claim the results will not be held in confidence and that census should not be conducted at all. During World War II, census records and population registration records from countries occupied by the Nazis, for example in the Netherlands, were used to identify Jews for detention, removal, and extermination. The effect of such use was to undermine the legitimacy of the census after World War II. In the Netherlands, which took its last regular census in 1971 and collects population information through other mechanisms, the legacy of the Nazi era was one of the major justifications for ending census taking (Seltzer, 1998).

At the other side of the spectrum, political challenges can be made to the census if the census does not count the population well enough. All censuses contain errors of various kinds. Some people and addresses are missed. People may misunderstand a question, or fail to answer all the questions. Census officials have developed elaborate procedures to catch and correct errors as the data are collected, but some errors tend to remain. Since census results are often used to allocate seats in legislative bodies and government funds, such errors, particularly undercounts, undermine the credibility of the census as an allocation mechanism.

In recent years, developments in statistical analysis have made it possible to measure the accuracy of censuses (Choldin, 1994; Anderson and Fienberg, 2001). Census results may be compared with population information from other sources, such as the records of births, deaths, and marriages in vital statistics. Census officials can also determine the level of accuracy of the count by conducting a second, sample count (a postenumeration survey) shortly after the complete census, and then matching the records of the sample and the census. Census officials estimate who is missed, and who is counted twice or in the wrong geographic location, and use these estimates to evaluate the overall accuracy of the count. The United States Statistics Division (2011: 22) reports that 26% of the countries responding to their survey reported that they would use their evaluation results to adjust the published census figures for the 2010 census round.

In the USA, people who live in cities, the poor, and minorities tend to be undercounted relative to the rest of the country. From 1970 to 2000 officials representing such undercounted jurisdictions claimed that these jurisdictions suffered loss of political representation and government funding since the apportionment and funding formulas are based on incorrect data. Mayors and leaders of civil rights organizations pressed for adjustment of the census results, and filed lawsuits to compel adjustment of the enumeration. The courts have permitted adjustment for some census uses but to date the Census Bureau has not produced sufficiently robust adjusted census results to justify supplementing or replacing the results of the door to door count.

Calls for Future Innovation

The emergence of digital technologies for collecting and reporting information is posing both new challenges and new opportunities for the census as an institution. On the one hand, increasingly wired and mobile populations question why governments need to spend substantial amounts of scarce public funding on a population count, when existing digital records might be able to provide the relevant population information. Several nations, including Canada's 2011 census and the US American Community Survey, are experimenting with electronic response methods for data collection. Planning for the 2020 census round includes proposals for integration of digital administrative records with sample survey methods to produce a cheaper and more efficient census operation. These proposals are being tested to determine if they produce data tabulations of sufficient quality to replace traditional methods.

Perhaps paradoxically, the same digital revolution is also giving new life to historically collected census data, both tabulated published results and the microdata if they have been preserved and archived. Large-scale research projects such as IPUMS USA and IPUMS International promise to retrieve and harmonize microdata from centuries of censuses from around the world for researcher use. Geographic information system technology is permitting historical analysis of tabulated census data from the past in projects such as Social Explorer.

See also: Census Microdata; Censuses, Population: Comparative International Aspects; Censuses: Current Approaches and Methods.

Bibliography

Alterman, H., 1969. Counting People: The Census in History. Harcourt, Brace & World, New York.

Anderson, M., 1988. The American Census: A Social History. Yale University Press, New Haven, CT.

Anderson, M., Fienberg, S.E., 2001. Who Counts? The Politics of Census Taking in Contemporary America, revised ed. Russell Sage Foundation, New York.

Cassedy, J., 1969. Demography in Early America: The Beginnings of the Statistical Mind. Harvard University Press, Cambridge, MA.

Choldin, H., 1994. Looking for the Last Percent: The Controversy over Census Undercounts. Rutgers University Press, New Brunswick, NJ.

Clanché, F., 2011. France's Redesigned Census: Lessons and Prospects. Paper presented at the Conference of the Italian Statistical Society Statistics in the 150 years from the Unit of Italy, University of Bologna Bologna, June 2011. http://unstats.un.org/unsd/censuskb20/KnowledgebaseArticle10695.aspx.

Cohen, P.C., 1982. A Calculating People: The Spread of Numeracy in Early America. University of Chicago Press, Chicago.

Desrosieres, A., 1991. Official statistics and medicine in nineteenth-century France: the SGF as a case study. Social History of Medicine 4, 515–537.

Desrosieres, A., 1993. La Politique des Grands Nombres: Histoire de la Raison Statistique. Edition La Découverte, Paris [1998. The Politics of Large Numbers: A History of Statistical Reasoning. Harvard University Press, Cambridge, MA].

Eckler, A.R., 1972. The Bureau of the Census. Praeger, New York.

Glass, D.V., 1973. Numbering the People. Saxon House, Farnborough, UK.

Herlihy, D., 1985. Medieval Households. Harvard University Press, Cambridge, MA.

Higgs, E., 2005. Making Sense of the Census Revisited. Census Records for England and Wales, 1801–1901 – A Handbook for Historical Researchers. The National Archives and Institute of Historical Research, 2005, London.

Nissel, M., 1987. People Count: A History of the General Register Office. Her Majesty's Stationery Office, London.

Patriarca, S., 1996. Numbers and Nationhood: Writing Statistics in Nineteenth-Century Italy. Cambridge University Press, New York.

Seltzer, W., 1998. Population statistics, the Holocaust, and the Nuremberg trials. Population and Development Review 24, 511–552.

United Nations Statistics Division, Department of Economic and Social Affairs, 2011. Report on the Results of a Survey on Census Methods used by Countries in the 2010 Census Round. Working Paper: UNSD/DSSB/1. http://unstats.un.org/unsd/census2010.htm (accessed 14.01.13.).

Wells, R., 1975. The Population of the British Colonies in America before 1776: A Survey of Census Data. Princeton University Press, Princeton, NJ.

Worton, D.A., 1997. The Dominion Bureau of Statistics: A History of Canada's Central Statistics Office and its Antecedents: 1841–1972. McGill-Queens University Press, Kingston, ON.

Wrigley, E.A., 2011. The Early English Censuses. Oxford University Press, New York.

Relevant Websites

http://unstats.un.org/unsd/demographic/sources/census/2010_PHC/default.htm – 2010 World Population and Housing Census Programme, United Nations Statistical Division, Demographic and Social Statistics.

http://www.ipums.org/ – Integrated Public Use Microdata Series (IPUMS).

http://www.nappdata.org – North Atlantic Population Project.

http://www.socialexplorer.com/pub/home/home.aspx – Social Explorer.

http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml – U.S. Census Bureau, American FactFinder.

http://www.census.gov/population/international/links/stat_int.html – U.S. Census Bureau, International Statistical Agencies.

Center–Periphery Relationships

Ulf Hannerz, Stockholm University, Stockholm, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Center and periphery, and related conceptual pairs, have been prominent in the social sciences since the mid-twentieth century. Some such pairs suggest a more consensual relationship, others portray conflict. The emphasis may be more on political economy or on cultural influence. Centers may rise and fall; peripheries may be involved with one or more centers. With increasing global interconnectedness, there is a greater differentiation of transnational center–periphery relationships.

Concepts of center and periphery became increasingly prominent in the social sciences during the second half of the twentieth century. They were of diverse origins and entailed varied emphases, but these differences have not always been obvious in continued use. There has actually been a cluster of related conceptual pairs: center and periphery, but also core and periphery, metropolis and satellite, and metropolis and province. But they have not been mere synonyms. Moreover, in obscure and perhaps treacherous ways, ideas of center and periphery may occasionally have overlapped in social thought and discourse with other classical contrasts, such as modernity–tradition or urban–folk.

The sociologist Edward Shils had a major part in introducing the center–periphery pair of concepts into the vocabulary of academic social science (cf Greenfeld and Martin, 1988). His paper 'Center and Periphery' originally appeared in 1961, but there was a close affinity between it and several of his other publications in the same period. The overall conception of society and culture in Shils' perspective was consensualist; but not, he asserted in a later comment, as merely a facile expression of the mood of the era. In his view, the importance of consensus to social life had been underestimated (Shils, 1975: p. xi). Thus he concentrated on notions such as 'tradition,' 'ritual,' 'deference,' and 'charisma.'

Notably, in his 'Center and Periphery' paper, these two terms were used in a metaphorical sense; centrality had 'nothing to do with geometry and little with geography.' The center was identified with what was ultimate, irreducible, and sacred in the realm of symbols, values, and beliefs; it was also identified in the realm of action with those roles and institutions which embodied such cultural understandings and were most actively engaged in propagating them. Yet in 'Metropolis and Province in Intellectual Life,' published in the same year, the facts of uneven spatial organization which are now usually linked to concepts of center and periphery become clear, and are set in what we would now describe as a transnational context. Drawing on his study of the situation of intellectuals in India, Shils argued that in their minds, people have varyingly extensive maps of the world significant to them, and that a major feature of such maps would be their portrayal of one's qualitative proximity to or distance from the metropolis (Shils, 1972: p. 356).

In the former of these two papers, then, the center is a locus of excellence, of grace; the periphery defers to it. In the latter, the metropolis is a center of vitality, a seat of creativity. The province, on the other hand, Shils notes, is frequently taken to be in itself "rude, unimaginative, awkward, unpolished, rough, petty, and narrow" (1972: p. 357). Cultural salvation lies in an involvement with the metropolis. The choice seems to be one between impoverished autonomy and enriching dependence.

If Shils was a theorist of consensus and cultural authority, the metropolis/satellite and core/periphery pairs, as inserted into the debates of the 1960s and 1970s, pointed in an entirely different direction, resonating with another political mood. 'Metropolis' and 'satellite' were the terms used by the economist Frank (e.g., 1967), who had spent much of the 1960s in Latin America, studying what he described as the development of underdevelopment, and engaging with the growth of 'dependency theory.' Then in the 1970s, introducing 'world-system analysis,' with its emphasis on changes in a global system since the late Middle Ages until the present, the sociologist Wallerstein (e.g., 1974) contrasted 'core' with 'periphery' (but inserting a 'semiperiphery' category as well). Both Frank and Wallerstein were primarily concerned with political economy, and with the expanding control and exploitation of the material resources of the 'periphery' or 'satellite' on the part of the 'core' or 'metropolis.' In both instances, the emphasis was on conflict and domination within a global order.

The points of view represented by Shils, on the one hand, and by Frank and Wallerstein on the other, seem not often to be brought into direct confrontation, although over the years, they may have come to interact and blend with one another. By now, when center–periphery concepts are more passingly referred to, it is not always entirely clear which of the varieties is involved; if not some combination of them. Yet there is a field of tension here between studies focusing on distributions of material assets and the exercise of power on the one hand and studies of culture on the other; between views suggesting consensus and views emphasizing conflict; and between views treating the social organization of meaning and meaningful forms somewhat in isolation and other views which insist on setting it in the context of political and economic structures.

Probably all would accept, however, that centers and peripheries imply each other. We are dealing with relational phenomena: there is no center without periphery, no periphery unless there is a center. The particular way the center is a center is also reflected in the way the periphery is a periphery.

Levels and Varieties

Our conception of center–periphery relationships here is that they are relationships of inequality existing in geographical space. Such relationships, however, have been defined at very

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12031-8

different levels of specificity. During the latter half of the twentieth century, especially in the Cold War era, the world was understood to be divided into three main segments – one Western and capitalist, one Socialist, and one 'developing' and postcolonial. Here, the First and the Second Worlds had their internal center–periphery structures. The Third World, on the other hand, may have had these, but also tended to be seen rather more as a periphery in its entirety, to one First or Second World center or other. At present, the extreme macroview may be that of a center Global North and a peripheral Global South. Much of the cultural and political debate over center–periphery relationships has remained mostly at such levels of identification. Notions such as Orientalism, Occidentalism, and Eurocentrism refer to understandings and representations of self and other in the context of these large-scale center–periphery structures.

Alternatively, identifications may indeed be made at the national level: the United States or France is seen rather more as centers than Sweden, Romania, or Burkina Fasso. In other instances, centers are yet more specifically placed. Nations and regions may have their own centers in particular localities. In the study of historical civilizations, there is frequently an emphasis on centers as sites combining political power with complex ritual life and elaborated knowledge systems in the hands of specialists. A well-known paper by the anthropologists Bernard Cohn and McKim Marriott (first published in 1958), for example, delineates the complementary roles of centers and networks in the integration of Indian civilization (Cohn, 1987: p. 78ff.). In related mid-twentieth century work, as Redfield and Singer (1954), also anthropologists, discussed the cultural role of cities, they described the dynamic relationship between the 'great tradition' of early urban centers and the 'little traditions' of the surrounding peasant societies. The emphasis on the symbolic authority of the center here is clearly reminiscent of Shils' view.

Yet Redfield and Singer also noted that later urban centers, rather than refining particular local traditions, are often at the crossroads of the world, bringing together diverse traditions and serving as communication nodes for wider areas. In this the authors adumbrated recent interest in the role of particular cities in contemporary transnational center–periphery relationships. 'World cities,' such as New York, London, or Paris, may be seen as generalized, multipurpose centers, combining various kinds of power and drawing the attention of the periphery for many different reasons (Hannerz, 1993; Knox and Taylor, 1995). The contemporary global structure of center–periphery relationships, however, can be understood as more internally varied. Cities may be centers in particular ways to particular people, dispersed in a transnational periphery – Rome to the Catholic world; San Francisco to gay people not only from elsewhere in the United States, but also from other continents; Memphis, Tennessee, to friends of country music. As we see, attachments of very diverse kinds are involved. This means that the same people in the periphery may well be involved with several, noncompeting centers for different purposes. Yet centers of the same type can also compete in the periphery.

How do centers become centers, even drawing attention across national boundaries? Some are centers for mainly historical or mythical reasons, because something occurred there once in the past, or occurred there first; or because the memory of a charismatic figure is preserved there. They belong on the map of the significant past, or of eternal truths. Often they are intensified as centers at particular points in a ritual calendar. Many 'centers of pilgrimage' have such characteristics. Other centers are very much of the present. They draw attention and stimulate imagination from afar because they are 'where the action is' now, with regard to one or more lines of human preoccupation. The contemporary proliferation of center–periphery relationships of transnational reach is undoubtedly related to a greater ease of transportation and communication, making centers easy to get to, and easy to stay in touch with. As the participants, practitioners, adherents, activists, and employees of more subcultures, life styles, ideologies, occupations, or corporations use new means to extend their circles out of local or national habitats, new far-away centers are discovered, or even made. As more diasporas are generated by more migrants, old centers may also be reaffirmed, from further away.

A significant part of what has happened in world center–periphery relationships in recent times also seems to involve the ways in which the periphery helps make the center what it is. As a center comes to be recognized, it may draw to itself the most committed of the periphery, as immigrants, tourists, participants, contributors, observers, and commentators. Thus the periphery not only adds to the outside reputation of the center, but may also contribute directly to the richness of its internal life. As the center is in the periphery, so to a degree the periphery may establish its presence in the center.

Center–Periphery Diffusion

In part, it may be in the passage of culture from center to periphery that the two parties to the relationship manifest themselves. In the clearest case, the center would be always the donor, the periphery always the recipient – the center is active, the periphery is passive. Yet even if the imbalance is not quite so great, some measure of net cultural export would appear to be one conception of center–periphery influence. What is prominently involved here, consequently, is diffusion.

In a wider arena of debate about the effects of global interconnectedness, since mid-twentieth century, the preoccupation with cultural diffusion from center to periphery has often been expressed in one or the other of two ways. Post–World War II modernization theory offered a forecast of where the world was heading – where the West actually already was, more or less, and where the rest would follow, with some assistance from the center. When not drily 'value-free,' modernization theory tended to take a positive view of the changes in question. The other major way of describing cultural diffusion on a global scale vas less favorably inclined. The term 'cultural imperialism,' as Tomlinson (1991: pp. 3ff.), has pointed out in a review, is a combination of two complicated concepts into one, with a considerable ideological charge. It has tended to refer to a growing dominance of western, and especially American, media and consumer goods in other parts of the world. Coca-Cola, McDonald's, and Barbie dolls have become the predictable pieces of evidence in a genre of cultural critique. Both the early form of modernization theory

and the critique of cultural imperialism thus tend to offer global homogenization scenarios; implying, forecasting, or warning of the end of cultural diversity.

Such scenarios, however, have been increasingly often contested. There has been a growing attention to the multicentricity of culture, to crisscrossing cultural flows, and not least to instances of cultural counterflow, from periphery to center. While some net asymmetry of flow may yet be undeniable, many commentators have also become less inclined to accept the assumption of passive reception at the periphery. They are much more likely to see an active periphery engaged in managing diffusion from the center: the periphery accepts this, rejects that, modifies one thing, and synthesizes something else with items from its own local cultural inventory. Metaphors of creolization and hybridity summarize this combination of cultural diffusion with creativity at the periphery.

Moreover, it has been argued that a periphery engaging with more centers can be in a quite different situation from one linked to only a single center – in a way it may even be in an advantaged position compared to these centers. Instead of being complacently self-sufficient and self-centered, as these often are, the multicentering periphery can at best go beyond eclecticism to work out its own cosmopolitan synthesis of its selections from the centers. Here, then, creativity draws on the peripheral condition.

Soft Power

A recently popular concept which may be placed in the context of center–periphery relationships is that of 'soft power,' originally proposed by the American political scientist Nye (2004). Here is a notion which links politics with culture. Soft power is understood to stand in contrast to that which is hard power: command power, not least military power. It co-opts rather than coerces people; "if I can get you to *want* to do what I want, then I do not have to force you to do what you do *not* want to do." Nye, concerned largely with the role of American power and its varieties in the world order, points to the different ways American values can be projected to support the country's global leadership role, but gives some special emphasis to the diffusion of American popular culture. This line of thought, then, portrays Washington and Hollywood working in tandem to build a global center–periphery structure.

One reason why the idea of soft power has had a wider appeal, however, is that it can attract commentators, and political actors, also in countries which have less of a real opportunity to choose between hard power and soft power. On a global or regional basis, countries or other entities can hope to be able to convert cultural penetration or persuasive power into political influence. This view, consequently, goes with a more multicentric understanding of relationships between centers and peripheries.

One factor in the global cultural order which may be linked to a notion of soft power is that of language, and language diversity. The Dutch sociologist De Swaan (2001) has delineated the characteristics of the contemporary 'world system of language.' It is a system with three major tiers; and what is important is less how many people use a language as their first language, their 'mother tongue,' but how many can use it as a second language. This is what at present makes English more important than any other language, the only language in the first tier. In the second tier, there are a number of languages with many native speakers, but also sizable numbers of other speakers: French, Spanish, German, Russian, Arabic, Chinese, and some others. Then in the third tier, there are innumerable languages, including national languages with few speakers outside the nation.

Languages have certainly not reached their places in this hierarchy as a matter of chance, or through some intrinsic merit. The world language system is an accumulated, although changing, product of history, politics, and commerce. Yet as it is, it may have some influence of its own. The present situation is one where a very considerable proportion of the people of the world can understand spoken or written English, and even though at least some of those who use it as a second language use it actively enough to make their voices heard, this means that those parts of the world who have it as a first language gain a certain advantage. It is easy to pay attention to them. What is said and written in other places, in other languages, becomes less widely accessible. Somewhat paradoxically, this center–periphery structure of language use is also strengthened by the distribution of translations. While it would seem that the widespread English-language skills would make translations of written work from this language into other languages somewhat superfluous, and translations into English would be more desirable, the fact is that translations from English into other languages are much more common.

Mixed Feelings

It is in the nature of their asymmetrical relationship that center and periphery take different views of it, and may even be differently aware of it. A periphery has to attend to its center. The center for its part may at times be preoccupied with its internal affairs, and may not have a clear grasp of the more widely dispersed consequences of its actions. The periphery tends to have a more developed, more detailed, sometimes more imaginative and sometimes more knowledgeable idea of the center than vice versa. (This is obviously a spatialized aspect of the general tendency that the subordinate knows more about the dominant than the other way around.)

Emotions may be attached to such linkages. Center–periphery relationships often do not leave people cold, passionless, and neutral. In the consensualist view, centers generate warm, deferential attitudes at the peripheries. In a conflict perspective, in contrast, center and periphery are primarily defined by political and economic structures – and we may then see their cultural concomitants at the periphery in the responses to control, coercion, and constraint. Along such lines, the many forms of cultural resistance have been a major theme in the portrayal of center–periphery relationships (also becoming more prominent in Wallerstein's world-systems analysis over time). The periphery may defend itself symbolically against intrusions through a celebration of the self and its local and traditional roots, but there may also be a denigration of the other, an uncomplimentary representation of the center.

Yet the responses of the periphery to the center are not always clearcut. For one thing, claims to center–periphery

consensus now increasingly meet with skepticism. With Gramsci as a source of theoretical inspiration, observers may interpret deference in the more complicated terms of hegemony. Nye's notion of soft power certainly has hegemonical implications. It is likewise possible, however, that people at the periphery may be genuinely of two minds about a center. If ambivalence is a prevalent quality of social life, as Smelser (1998) has argued, perhaps this is even as common a response to a center as any more one-sidedly favorable or unfavorable stance. It may be a place one loves to hate and hates to love. One may feel that good things come from there, and at the same time resent its influence, or what may seem to be its narcissism. Especially those multipurpose centers understood to be 'where the action is,' moreover, may inspire sentiments which are positive, but instrumental or playful rather than deferential. They are among the liminal spaces of contemporary life.

All views of the center, however, and all messy feelings toward it, are not necessarily to be found in any single inhabitant of the periphery. They may rather be complicatedly distributed among its population, and such variations can be the foci of intense debates and conflicts structuring local life at the periphery.

The End of Center–Periphery Relationships?

With some frequency, one now hears voices criticizing center–periphery concepts, or proposing the decline of center–periphery structures. It is said that the world economy is increasingly decentered (see e.g., Lash and Urry, 1994: p. 4). Human life, it is also argued, is increasingly deterritorialized. People move quickly between places, being rooted in neither of them, and have relationships which belong in no place in particular. 'Critical masses' for cultural elaboration may be built up in cyberspace, and do not require local face-to-face contacts. As would-be centers and would-be peripheries are closely in touch, by way of electronic media or jumbo jets, the 'cultural lag' which divided them in the past is no longer there.

Even if they are often closer to futuristics than to the really existing world, there may be something to such arguments, At times, however, both the concepts criticized and the alternatives to them would need to be elaborated more precisely. Indeed, we must be sensitive to the ambiguity of our terms, and to the actual range of variations out there. Asymmetries may indeed only be relative, culture flows are not entirely one way, the world is crisscrossed by more symmetrical relationships as well, and there are many centers, and many kinds of centers, rather than only one. Centers also rise and decline. In the early twenty-first century, some of the large countries of that Global South – China, India, Brazil – are turning into major economic powers. It remains to be seen what global cultural influence will go with that, but one could note that Japan, rising earlier in the world economy, fairly soon thereafter did gain more global attention for its developing popular culture (Iwabuchi, 2002). In some Japanese circles, the hope was expressed that this would be accompanied by more soft power.

Center–periphery conceptualizations, it seems, deserve to be continuously scrutinized and developed, rather than merely rejected. It is notable, too, that the end of centers and peripheries appears often proclaimed by commentators whose own vantage point is at the center; in this case, not a privileged position.

See also: Creolization: Sociocultural Aspects; Diffusion: Anthropological Aspects; Global South; Globalization, Anthropology of; Modernity: Anthropological Aspects; Third World; World Systems Theory.

Bibliography

Cohn, B.S., 1987. An Anthropologist Among the Historians and Other Essays. Oxford University Press, Delhi.
De Swaan, A., 2001. Words of the World. Polity, Cambridge.
Frank, A.G., 1967. Capitalism and Underdevelopment in Latin America. Monthly Review Press, New York.
Greenfeld, L., Martin, M. (Eds.), 1988. Center: Ideas and Institutions. University of Chicago Press, Chicago.
Hannerz, U., 1993. The cultural role of world cities. In: Cohen, A., Fukui, K. (Eds.), Humanising the City? Edinburgh University Press, Edinburgh.
Iwabuchi, K., 2002. Recentering Globalization. Duke University Press, Durham, NC.
Knox, P.L., Taylor, P.J. (Eds.), 1995. World Cities in a World-System. Cambridge University Press, Cambridge.
Lash, S., Urry, J., 1994. Economies of Signs and Space. Sage, London.
Nye, J.S., 2004. Soft Power. Public Affairs, New York.
Redfield, R., Singer, M., 1954. The cultural role of cities. Economic Development and Cultural Change 3, 53–73.
Shils, E., 1972. The Intellectuals and the Powers. University of Chicago Press, Chicago.
Shils, E., 1975. Center and Periphery. University of Chicago Press, Chicago.
Smelser, N.J., 1998. The Rational and the Ambivalent in the social sciences. American Sociological Review 63, 1–16.
Tomlinson, J., 1991. Cultural Imperialism. Pinter, London.
Wallerstein, I., 1974. The Modern World-System. Academic Press, New York.

Central Africa: Sociocultural Aspects

René Devisch, University of Leuven, Leuven, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Most authors define Central Africa as the vast area comprising speakers of the different western branches of Bantu. The great diversity of western Bantu populations and traditions has gradually extended, beginning 5000 years ago, from southern Cameroon, into Equatorial Guinea, Gabon, People's Republic of Congo, the Democratic Republic of Congo, and Angola. Side by side with Ubangi and Nilotic speakers, they have moreover spread across the Central African Republic, South Sudan, western Uganda, Rwanda, and Burundi, to eastern Congo, Tanzania, and northwestern Zambia. Whereas the eastern Bantu have specialized in grain growing, the western Bantu adopted cassava and yams for staple foods. The region has been the scene of complex interactions between the local and outer worlds.

In the equatorial forest in the Congo basin and its northern fringe, for centuries decentralized agriculturalists have tended to privilege their patrilocal households as an economic and political stronghold. By the sixteenth century, long-distance exchange economies of textiles and firearms for ivory and local captives may have connected the equatorial forest even with Egypt via eastern Sudan. The savannah regions in the southern and western half of Central Africa saw the emergence of important political leadership roles, including kingship. Their history became tragically bound up with the growth of the Atlantic slave trade. Following Livingstone's and Stanley's trans-African 'explorations,' the 1885 Conference of Berlin legitimized imperial Europe's colonial scramble for Central Africa. From the 1950s on, frustration on the side of those having assimilated the European civilizational capital grew with the glaring contradictions between their 'Whitened' skills and aspirations, on the one hand, and their second-rank position in the colonial institutions, on the other. In the 1970s, Authenticity Movements as well as parody in popular culture cropped up in various newly independent African states and prompted in people a radical shift away from passive assimilation of the White civilizational models; now they urged them to take possession in a militant way of their own history and a dignified self-representation on the world scene.

Most authors define Central Africa as the vast area comprising speakers of the different western branches of Bantu. The great diversity of western Bantu populations and traditions has gradually extended, beginning 5000 years ago, from southern Cameroon, into Equatorial Guinea, Gabon, People's Republic of Congo, the Democratic Republic of Congo, and Angola. Side by side with Ubangi and Nilotic speakers, they have moreover spread across the Central African Republic, South Sudan, western Uganda, Rwanda, and Burundi, to eastern Congo, Tanzania, and northwestern Zambia. Whereas the eastern Bantu have specialized in grain growing, the western Bantu adopted cassava and yams for staple foods. The region has been the scene of complex interactions between the local and outer worlds.

AD 1000–1880

Probably due to its sparse population, Central Africa did not develop the type of domestic markets and specialized local industries as found in West Africa. Until the late fifteenth century, it stayed out of the lines of long-distance communication such as those that had developed from coastal East Africa or along the caravan routes of West Africa. The history of precolonial Central Africa from AD 1000 can be divided broadly into two main regional developments.

With regard to the northern half of Central Africa, comprising the great equatorial forest with its northern woodland fringe and the Congo basin, linguistic and ethnographic data – common 'words and things' (Vansina, 1990) – bear witness to the interaction between the farming and pastoral peoples in the forest and savanna areas north and south of the Equator. The most important hunter-gatherer groups of northern Central Africa were the matrilineal Pygmies, who interacted with their agricultural neighbors and long-distance trading groups (like the salt-making Vili of Loango, and the neighboring copper-trading Tio). The Zaire River and its tributaries have comprised an important corridor for the transmission of cultural influences and systems of government, as well as for the trading economies of the Atlantic zone. The farming in the thinly scattered patrilineal societies of the northern and eastern corner of the rain forest remained for long untouched by the food plants originating from South America and already imported into other parts of Central Africa.

In the central part of the equatorial forest, the Mongo patrilineal societies were strongly influenced by the neighboring agriculturalists to the northwest along the Ubangi and higher up in the Sudan. Mongo Big Men, often combining agriculture and trade, tended to privilege their patrilocal household and patriclan rather than the village community as an economic stronghold and political unit, also by way of matrimonial transactions and cultic and initiatory specializations. By the sixteenth century, long-distance exchange economies of textiles and firearms for ivory and local captives may have connected the equatorial forest even with Egypt.

The savanna regions in the southern and western half of Central Africa saw the emergence of important political leadership roles, including kingship. Depending on the rainfall, the wooded savanna was moderately suitable for agriculture and/or pastoralism. Priests were the guardians of all reproductive resources. The extended family was the focus of physical and spiritual well-being of both living and deceased members of the patrilineage. Lineage chiefs gained control over the

rain-making cults, the territorial custody of hunting traditions, and the collection of tribute. From the fourteenth century, the chiefs gained regional control over iron and smithy work and, particularly from the seventeenth century, over firearms and market-oriented commerce. This as well as their indirect involvement in the Atlantic slave trade made their power paramount over large areas in central Angola and western Congo.

Challenges from the Atlantic and Colonizing Europe

Unlike West and East Africa, Central Africa had not entered the maritime trade prior to the arrival of the Portuguese caravels seeking trading bases, from the 1480s onward. Reports by Portuguese missionaries and merchants as well as occasional correspondence by Western-educated Kongo show how the foreign influence brought radical changes to Kongo kingdom and culture. The kingdom then comprised prosperous farming communities linked to a regional trade of fish and craft goods. From this period, the *nzimbu* shells in the hands of Europeans took on the functions of coinage for the previously unfamiliar purposes of wage payment and marketing.

Portuguese traders carried Mediterranean manufactures to Kongo in exchange for raffia cloth, ivory, dye wood, and copper. The redistribution of these manufactures (such as North African textiles, iron knives, glass mirrors, glass schnapps bottles, glazed bowls, Venetian beads, and glazed china) was controlled carefully by the royal court. Moreover, Portuguese teachers, artisans, lawyers, and priests enhanced the authority of the king and his closest supporters. A number of Kongo were taken to Europe for further education. From the sixteenth century, the history of the southern savanna became tragically bound up with the growth of the Atlantic slave trade. Manpower, rather than landed property as in early modern Europe, was the key to value in the Central African communities. Slave trading of servile subjects and captives gained via long-distance marketing at the Malebo Pool (where Kinshasa would later develop) became the grim solution to local needs for foreign goods. Like any slave-trading society, the Kongo kingdoms and other neighboring equatorial kingdoms became oppressive and fractious, and doomed to collapse.

The Atlantic trade gradually came under the control of the semicolonial Luso – an African community of the Loanda hinterland, Africa's first White colony. The long-distance savanna trade routes probably became avenues for the dissemination of European goods and influence, as well as of the South American crops: maize, cassava, tobacco, and tomatoes. Soon, frontier states such as Matamba and Kasanje as well as a new class of trading entrepreneurs took control over trade relations with the European powers. Following the industrial revolution in Europe, new goods were being shipped to southern Cameroon, Gabon, and the lower Congo. Toward the end of the eighteenth century, with the involvement of French, English, and Dutch merchants, the Atlantic slave trade reached its peak.

Swahili-Arab penetration in northeastern Zambia and southeastern Congo in the early nineteenth century started as trade of guns and powder for ivory and slaves. It would soon overthrow the local chiefs and gain independent political power for itself.

Following Livingstone's and Stanley's 'explorations,' the 1885 Berlin Conference legitimized Europe's division of Central Africa into colonies. The conquest of Equatorial Africa would last until the 1920s before military power, forced labor, harsh rubber exploitation, diseases, and hunger would have decimated half of the population and broken local resistance. Colonial rule gradually deprived local societies of the institutional ability to confront the exploitative forces and exogenous 'civilizing programs' imposed upon them. In the views of the colonizing masters, the bringing about of philanthropic medical provision and widespread school education in the colonies was an essential feature in the international justification of political and economic colonialism. The colony's extractive economy, directed at the 'so far unutilized resources' and relying on mandatory labor in lieu of tax payment, as well as its transport and communication technology were to open up the remote areas to the larger world scene. They were to 'uplift' and integrate the local populations in the new era of 'universal civilization and progress.'

Bantu traditional rule was deeply alien to the colony's display of order. Colonial state power was imposed through regulations, enticements, and direct interference, as both text and texture, absorbing and domesticating people and events by the very acts of *writing* and administrative *ordering*. The textual economy disassembled and inserted local realities into panoptical recording and regulation. The management of people's civil identity and geographic confinement was achieved through the identity cards and mandatory 'passbooks.' Records of marriage, descent, settlement, and land use were meant to bind the populations geographically, and were authoritative for the succession to chiefly titles and for settling matrimonial and family disputes. On the other hand, such regulatory documents were intended to modernize society by freeing the so-called *évolué* from customary collectivism. Population censuses, together with geographic and linguistic-ethnographic mapping and recording of data, allowed for a compartmentalization of languages and ethnic groups.

Unlike the nineteenth century bureaucratic nation-state in the West, the Bantu political traditions of Central Africa do not draw their inspiration from orders of visual representation and architectonic spatial models. They are molded by organic, hydraulic, and/or animal-totemic metaphors informing political networks and strategies as an order of events, forces, sources, and relations. Membership of and alliances between particular social groups are not primarily tied to a geographic partitioning but to blood ties and to the mythical or primal space-time order in line with the constant cosmogenetic reenactment of the reproductive and hierarchical weave between the founding ancestors, their foundational and migratory exploits, and their multiple descendants.

The functions of traditional political title holders were, and are, being thought of as prior to, and the source of, all life forms as well as the guarantee of their order. A chief represents and surpasses his subordinates by his twofold function. First, like his totemic animals (leopard, eagle, and crocodile), he is a conqueror. Through enthronement, the ruler embodies the

founding ancestors and represents the primal space-time order initiated by the immigration of his ancestral people and their conquest of the land. Through his body, in particular, his keen perceptiveness and the nightly forces that he shares with his totemic animals, the ruler impersonates the founding ancestor. He imposes the qualities of a perennial hierarchical social organization, territorial unity, and moral order on his society. Second, the chief acts as the androgynous lifegiver or mediator of the (re)generative processes between the land, society, humans, and the 'the primal womb' in the earth. His rule is thus one of cosmic and physical regeneration, guaranteeing human, agricultural, and social reproduction, and instituting commensality and sharing in his territory. In this diurnal ordering role, the chief protects his people from the nocturnal antirule of envy, theft, sexual abuse, and sorcery. The elders, through councils, ceremonial exchange, and authoritative speech, extend the regenerative capacity into the daily weave of events and relations uniting kin groups under common rule.

In the Christian mission stations or in the plantation, mining, and industrial enterprises, ever larger numbers of individuals and families joined the workers' camps and 'indigenous townships,' set up outside customary rule, on the fringes of White settlements. They thus created new ideological and physical spaces of identity and collective imagination. Indeed, from the last decades of the nineteenth century, Christian congregations set out to work side by side with the concessionary companies and the colonial administration. Their aim was to 'bring civilization and universal salvation to the Dark Continent,' as well as to 'save the individual's soul' through the removal of 'pagan African customs.' While the first decades of the missionary endeavor were directed at the adaptation of Christianity to local society and culture, from the 1930s on, missions became increasingly involved in the social engineering of the townships, echoing the dominant mood of the plantation or mining trusts and the colonial administration. From then on, church authorities designed paternalist programs to educate converts toward assimilating modern skills and dispositions. By the 1950s, the term 'modernization' had become a new banner for the assimilationist option.

In the Belgian Congo, the colonial government had secured the agreement of the missions to manage the schools in exchange for land concessions. Nearly all mission stations developed boarding schools with enough fields, cattle, poultry, and workshops to be self-sufficient. The colonial school presented Western progress as the source of a more dignified identity on the world scene. Vinck (1995) convincingly shows the extent to which, from the 1920s in the Belgian Congo, the school books conveyed the primary symbols and values of the West. By encouraging the pupils to develop a new identity and self-image, these manuals influenced the entire political élite-to-be in postindependence Congo. The concept of state, a key notion in the school books, was the comprehensive, abstract, depersonalized expression of the new power and authority that came to dominate the various spheres of life. According to the manuals, all Whites shared in the power of the state and thus were to be considered as authorities. Colonial school books depicted traditional society and religion as dominated by sorcery and the machinations of the devil.

Endogenization

From the 1950s, among the second generation of those having assimilated the European civilizational capital, the frustration grew with the glaring contradiction between the promises that the colonial master had put before them and their second-rank position in the colonial institutions. A few Christian priests, pastors, and élite got imbued with the Third World emancipation movement enhanced by the rapid decolonization in Asia, the 1955 Afro-Asian Conference in Bandung, the independence struggle in Algeria, the *Négritude* call for the rehabilitation of African cultures, and the Pan-Africanism radiating from Ghana and Guinea.

The militant élite used a triple banner in the battle for decolonization (Devisch and Nyamnjoh, 2011). First, in the terms of the Christian, humanist, and/or socialist discourses of the colonizers, they claimed the right to dignity, social emancipation, and equality. Second, their political aspirations were largely drawing on European ideological discourse associating modern economic development with gradual political emancipation. They thereby became complicit with the modernizing and authoritarian endeavor of the White welfare state and nation-state, however alien it was to people at the grassroots level. Third, in the terms of western Bantu traditions, they favored a palaver model of negotiation, ceaselessly co-opting 'brothers' and 'doing things with words.'

In the 1970s, authenticity movements in various states prompted a radical shift away from assimilation of the White civilizational models, now urging people to take possession in a militant way of their own history and a dignified self-representation on the world scene. Throughout the 1970s and 1980s, artists and customers in the discos celebrated liberation from paternalist and moralizing colonization. Today, people in both the rural and urban milieux have become aware how much their lifestyles, modes of production, and environments are both in rupture with their parental models, and are simultaneously excluded from the current economic and informational globalization of Western consumerist lifestyles. The masses in the poverty-stricken suburbs have been haunted, through television and downtown scenes, by the imageries of ease and extravagance that the transcontinental mass media as well as the few very rich nationals exhibit in fine clothing, expensive cars, and luxury goods. Many are increasingly bitter about their exclusion. However, they do not fail to develop their own proud visions of society and the way things are. In the early 1990s, through the waves of demonstrations and Dead-Town protest manifestations or even uprisings, suburban people counteracted the governmental views on modernization and the neocolonial politics. Irony and parody allowed the populace to deconstruct imperialist twentieth-century modernity.

Suburbanites straddle worlds through hybridity or the imaginary transgression of codes, in particular in the utopian fields of humor, daydream, and language. These are gradually furthering a cultural critique of the postcolonial situation: rather than fighting one another as ill-fortuned or discontented consumers of modern cash goods (increasingly brought in secondhand from the North), people in suburban neighborhoods and rural communities are reexploring their

genuine sense of communality, and their collective memory stored in body techniques and sensuous culture. Hybridity in many cultural expressions blurs the tradition–modernity, Bantu–Western, precapitalist–capitalist oppositions that have been created by the Europeanizing tropes. Songs on the radio today in vernacular languages and along old rhythms, recalling both the collective frenzy and euphoria in the bars of the 1970s and 1980s and village festivities or rituals, are now unsettling the Reformist voices of the (post) colony which had connected city life with French or English speech and etiquette, with school education and the petit bourgeois lifestyle, and with well-equipped biomedical services. The pidginization or creolization of French or English colloquial language, and of narrative styles in songs and newspapers, vitiate modernist master tales about the literate African city dweller and the retrograde and illiterate villager. Thousands of independent prophetic healing churches of the holy spirit, through exorcizing materialist greed and Westernization of public mores, or even state politics identified with the work of *sataani* (the local term for the Christian-imported notion of Satan, referring to those engaged in illegal practices and sorcery), develop a forceful critique of the catastrophic collusion between economic modernization and people's sociocultural dispossession. At the same time, through collective trances and 'Christian' forms of telepathy and clairvoyance in the name of the holy spirit, the prophetic churches explore ways toward domesticating modernizing forces. During the ceremonial offering of money to the prophet and assistant assembly leaders, adepts may circle more than a full hour around the congregation, dancing and singing as if to reenchant one of modernity's most striking secularizations, anonymous cash trade. The offering aims to transform the subaltern's poverty and needs into what is defined as divine grace or healing. The prophetical churches thereby speak back to global capitalism and celebrate their participation in a new economy. They subject the monetary economy to the utopian ideals of equality and solidarity proper to brotherhood and sisterhood.

See also: Africa's Environment: A Stressed Biogeographical and Cultural Landscape; African Studies: Culture; African Studies: Health; African Studies: History; Colonization and Colonialism, History of; Near Middle East/North Africa Studies: Society and History.

Bibliography

Andersson, E., 1958. Messianic Popular Movements in the Lower Congo. Almqvist and Wiksells, Uppsala, Sweden.

Birmingham, D., 1981. Central Africa to 1870: Zambezia, Zaire and the South Atlantic. Cambridge University Press, Cambridge, UK.

Birmingham, D., Martin, P.M. (Eds.), 1983. History of Central Africa, vols. 2. Longman, London.

Boahen, A. (Ed.), 1990. General History of Africa: VII. Africa under Colonial Domination 1880–1935, abridged ed. UNESCO/Currey, Paris/London.

Cairns, H., 1965. Prelude to Imperialism: British Reactions to Central African Society 1840–1890. Routledge, London.

De Boeck, F., 1996. Postcolonialism, power and identity: local and global perspectives in Zaire. In: Werbner, R., Ranger, T. (Eds.), Postcolonial Identities in Africa. Zed, London, pp. 75–106.

Devisch, R., 1995. Frenzy, violence, and ethical renewal in Kinshasa. Public Culture 7, 593–629.

Devisch, R., 1996. Pillaging Jesus: healing churches and the villagisation of Kinshasa. Africa 66, 555–586.

Devisch, R., Nyamnjoh, F. (Eds.), 2011. The Postcolonial Turn: Re-Imagining Anthropology and Africa. Langaa & Leiden: African Studies Centre, Bamenda.

Harms, R., 1981. River of Wealth and Sorrow: The Central Zaire Basin in the Era of Slave and Ivory Trade, 1500–1891. Yale University Press, New Haven, CT.

Hunt, N.R., 1999. A Colonial Lexicon of Birth Ritual, Medicalization and Mobility in the Congo. Duke University Press, Durham, NC.

Mazrui, A.A., Wondji, C. (Eds.), 1993. General History of Africa: VIII Africa since 1935. Heinemann, London.

Mbembe, A., 1992. Provisional notes on the postcolony. Africa 62, 3–37.

Mudimbe, V.Y., 1988. The Invention of Africa: Gnosis, Philosophy, and the Order of Knowledge. Indiana University Press, Bloomington, IN.

Vansina, J., 1966. Kingdoms of the Savanna: A History of the Central African States until European Occupation. University of Wisconsin Press, Madison, WI.

Vansina, J., 1990. Paths in the Rainforests. Currey, London.

Vaughan, M., 1991. Curing Their Ills: Colonial Power and African Illness. Polity Press, Cambridge, UK.

Vinck, H., 1995. The influence of colonial ideology on schoolbooks in the Belgian Congo. Paedagogica Historica 31, 355–405.

Central Asia, Anthropology of

Catherine Alexander, Durham University, GBR, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since the end of the Soviet Union and with improved access to the region, the anthropology (in the Western tradition) of Central Asia has taken off. Recently, there has been greater acknowledgment of local sources, archives, and scholars. Moving from a sense of Central Asia as an isolated area, a variety of approaches have begun to theorize the region as partly integrated with other regions and traditions: Islam, Turkic and nomadic peoples, Southwest Asia, and postsocialist and postcolonial areas. An emphasis on past and current migrations and trading networks to, from, and within the area further stretch the idea of Central Asia and reveal the complex lived experience of ethnic groups – often at odds with nation-state ethnoterritories. Debates continue as to the extent to which clan organization informs current elite politics and pre-Soviet traditions are reemerging intact, in hybrid forms, or were wiped out. The main focus of studies has been on religion, politics, and ethnicity. Further research is needed on new economic and other inequalities, industry, labor, and the environment.

Central Asia is arguably one of the oldest ingénues on the world stage: repeatedly discovered by various scholarly communities, commercial or state powers bent on opening up new ethnographic regions, seeking new resources and reconfiguring geopolitical alliances. Newton (2001) suggested the image of the loose fish to evoke the scramble to discover and make various claims in, and about, Central Asia after the collapse of the Soviet Union as though the area were a blank spot on the map. Arguably, this sense prevails in conceptualizing the anthropology of the region as catch-up (Liu, 2011) for Anglophone or Western style scholarship. The recognition that a wealth of other work exists, however, by pre-Soviet, Soviet, and post-Soviet local scholars has begun to reshape the contours of the anthropology of Central Asia through new translations and critical engagements with historical accounts (e.g., Levi and Sela, 2009), Soviet-era ethnography (Mühlfried and Sokolovskiy, 2012) scholars from the region (Masanov, 1996; Naumova, 1991) and archives (Jacquesson, 2010), following what Stoler (2009) christened the archival turn.

Elastic Area Definitions

While geographical definitions of Central Asia are notoriously variable, overlapping with other categorizations such as Russian Turkestan, which covered the Turkic lands, the author adopts the current convention of Kazakhstan, Turkmenistan, Kyrgyzstan, Uzbekistan, Tajikistan, and Xinjiang (although an argument could be made for including the post-Soviet Caucasus region). It is worth being reminded of the differences of each country plus the specificity of Central Asia and elements it shares with other regions and histories. Indeed, as discussed below, the region often seems to appear and disappear in the same stroke. A few characteristics will serve to frame the remainder of the article. Unlike the steppe, nomadic past of the other states, Uzbekistan has a long history of urban settlement; indeed Tashkent was the main city of Soviet Central Asia in terms of population size and cultural purchase; a point that is, perhaps obviously, contested by other cities of the region. Despite the aridity of Tajikistan and Uzbekistan, much of this land was turned over to cotton monoculture in the last century, which in turn necessitated huge irrigation projects, in part responsible for the Aral Sea's desiccation and environmental catastrophe from which many fled. The mountainous terrain of Tajikistan and Xinjiang, sometimes known as Chinese Turkestan, means only a small area is available for habitation and cultivation. Xinjiang is home to many ethnic groups from the rest of Central Asia, particularly the Uyghurs, a Turkic people with no titular state (Bellér-Hann et al., 2007).

As Roy (2000) observed, Central Asian polities, particularly the Ferghana valley borders between Kyrgyzstan, Uzbekistan, and Tajikistan, appear to have been devised by Stalin with an eye to the necessity for an overarching political entity to keep the peace. Sure enough, with the end of the Soviet Union, civil war broke out in Tajikistan and there have been conflicts in the region since. Border violence following independence has echoes in many other postcolonial settings. Again as Kandiyoti (2002a) noted, Central Asia was distinct economically and politically from other Soviet republics. In eastern European states, such as Hungary and Czechoslovakia, economy and political control were roughly coterminous with the polity and managed locally, whereas the Central Asian states were largely extractive economies or agricultural monocultures (cotton and wheat) effectively controlled by Moscow, via all-Union ministries and local cadres who also directed resources across Soviet space. Technical and professional work was typically dominated by Slavs, presenting an acute shortage of such expertise when many migrated West after 1991.

One further point is the Soviet-era energy network that weaves the Central Asian states together into a curious cat's cradle of interdependency.

One indication that Central Asia is now recognized globally to be as vital an ethnographic region as any other, not just a curious relic shuffling back into life, is the recent overviews in mainstream anthropology journals and handbooks (Rasuly-Paleczak, 2005; Liu, 2011; Marsden, 2013). Ironically then, even as Central Asia has emerged as an ethnographic region, it has largely been through revealing the essential plurality and fluidity of the area, challenging macrolevel accounts, both emic and etic, of nation-state building enterprises on four counts. First, the region is often theorized as an extension of other regions and traditions – often being treated as though

remarkably plastic to the impress of theories cultivated elsewhere. Second, past and present migrations for economic, political, and environmental reasons extend the reach of the region's peoples globally – and certainly challenge anthropological accounts to take such diasporas into account. Third, the coexistence of stateless groups such as the Uyghur and other ethnic groups within nation-state boundaries query the monocultural productions of elites and, fourth, films and the built environment are increasingly being analyzed as a way of engaging with an imagined global audience.

Thus, Liu (2011) suggests following Chari and Verdery's (2009) linking of postsocialist (typically the Soviet Union and eastern bloc) and postcolonial areas, although Kandiyoti points to the limitations, as well as the potential, of extending postcolonial (2002b) and postsocialist (2002a) analyses to Central Asia. Rasuly-Paleczek (2005) shows the advantages of drawing on other Turkic analyses of relationships between tribe and state; Marsden (2012) suggests comparison and integration with southwest Asia and broader imagined Muslim communities. An increasing emphasis on tracking historic and contemporary migrant flows, the changing positions of different ethnic groups within and beyond each country, and the palimpsests of ethnic groups and social practices within cities present sharply different views of ethnoterritories from that presented by nation-state elites, more akin to Appadurai's (1990) notion of ethnoscapes. Finally, a recent interest in political performativity (Cummings, 2010; Beyer et al., 2013) and spectacle, whether events (Adams, 2010), the built environment (Laszczkowski, 2011; Alexander et al., 2007), or films made by or about (Yessenova, 2011; Norris, 2012; *Slavic Review*, 2008) Central Asian states, suggests another avenue for extending the region as government elites seek to brand the nation through monopolizing notions of national culture and tradition speaking both to their own populations and to an imagined global stage. Lane's (1981) classic account of the rites of rulers appears to be holding its own, albeit in a nominally different context.

In the remainder of this article, the author touches on and extends the main themes (covered by these excellent overviews, economics, politics, and religion) ending with topics largely omitted, for want of space, by these essays: migration, gender, industry, labor, the current economic crisis, and environmental challenges. However, the author starts with a brief overview of the historiography and institutional framing of the region.

Non-Western Accounts of Central Asia

Levi and Sela's (2009) anthology of historical sources on Central Asia from the seventh century on shows a sequence of Arabic, Muslim, Turkic, and Mongol encounters via trade and conquest and accounts of the area. The Uzbek dynasties, of the sixteenth and seventeenth centuries, most famously chronicled by Babur, were also described by English merchants, explorers sponsored by Catherine the Great and, later, by players of the first 'great game.' Paksoy suggests that these early forays into linking Turkic Central Asia with Turcology, such as the sixteenth-century Dede Korkut epics translated by Barthold (1894–1936) in the 1930s were suppressed by Stalin in the 1950s as indicative of a potential pan-Turkic polity. The phenomenal resource wealth of Central Asia has long attracted careful mappings of resources: lands, people, flora, and fauna. Shoqan Walikhanov (1835–65) is one of the best known Kazakh scholars of the region, although as a fully signed up member of the Russian imperial army his adoption as one of Kazakhstan's national heroes is perhaps a little strained. Access to Soviet space by Western scholars was restricted, with Central Asian encounters particularly sparse. Lubin's (1985) account of labor relations in a factory in Andijan in 1979 is almost unique as both an ethnography and going beyond religion, politics, and kinship, three topics that continue to dominate studies. Piirainen's (1996) ethnography is a fairly unique and excellent description of a one-company industrial town in Kazakhstan breaking down after 1991. Otherwise, studies of Central Asia by Western scholars were largely limited to macrolevel analyses based on secondary data such as official statistics and archives, as it was in the earlier days of the 1990s, when state formation was of central interest (Smith et al., 1998).

Institutional Framing for Areas and Topics

For anthropologists, institutional and national support concentrated fieldwork in some countries and on certain topics, although not exclusively. Thus, research programs at the Max-Planck Institute for Anthropology have covered property changes (Hann, 2003) and civil religion (Hann, 2006) in Central Asia. The establishment in 1992 of the French Institute for Central Asian Studies in Tashkent, enabled several French scholars to undertake research in Uzbekistan. Elsewhere access continues to be difficult. Tajikistan's civil war 1992–97 and continued devastation has attracted few anthropologists (pace Harris, 2004), although several PhD dissertations are based on the country and Marsden (2012) provides wonderful accounts of Afghan traders in Tajikistan; Turkmenistan is more forbidding yet (see, however, Peyrouse, 2012). So, however much one wants to quibble over the geographical extent of Central Asia, most of the anthropology, in the Western tradition, whether from the East Pacific, United States, or Europe, has concentrated on Kazakhstan, Kygryzstan, and Uzbekistan. Xinjiang, interestingly, has variously attracted scholars of China, Islam, postsocialism, and Turkic peoples and languages.

Property Regime Change

Most accounts of economic change in the 1990s can be aligned with broader postsocialist narratives across the former Soviet region, which focused on immediate economic survival in the aftermath of the abrupt contraction of the state and, with it, the disappearance of support for health, education, basic welfare, and state enterprises. Alongside this, mass unemployment appeared as industry, collective farms, and services were privatized, restructured, or simply ceased to function. Although property regime change is sometimes glossed as the privatization of state assets and services, state disinvestment did not necessarily result in complete transfer of property objects to private control since assets, such as the social infrastructure

belonging to industrial and rural enterprises, were often seen simply as capital draining liabilities. The landscape was therefore one of unclear ownership of property objects just left to rot alongside privatized state enterprises, land, and housing (Alexander, 2004b). To an extent, these concerns were shared with studies elsewhere of other postsocialist states. Unemployment took many forms from 'unpaid leave' to below-subsistence wage levels, delayed wages, payment in kind, and, finally, complete loss of jobs. When not sold off in toto, the nominal transfer of urban and rural enterprises to their workers via shares, often resulted in the rapid consolidation of control by a few, usually previous directors, as shares were exchanged for immediate cash needs or the small patches of land allocated to individuals or families were unable to support a household, especially where seeds, fertilizer, fuel, and markets were no longer provided. Trevisani (2010) also shows how control was kept by the same people, despite reforms, in Uzbekistan's cotton farms. Corruption and embezzlement were widespread (Khazanov, 2012). Nazpary (2001) proposed a theory of systematically created chaos to obfuscate the deliberate dispossession of people who ended up losing out in Almaty in the 1990s. Although casting emic confusions about what was going on as an account of systematic change, it is nevertheless a powerful narrative of the experiences of those who ended up with nothing. Explanations of changing economic conditions often spun on elites; Rasanyagam (2011) proposed the idea of the state continuing to be the fount of resources to be manipulated by emerging elites as part of the informal economy.

The effects of the abrupt cessation of much industry and associated farms and infrastructure were manifold. More highly valued male jobs were also often those that collapsed first, often seeing a shift in women's roles to primary breadwinner as they went to work in bazaars (Yessenova, 2006; Spector, 2008). Kaiser (2005) tracked the *chelnoki* (suitcase traders), initially often women (Werner, 2004) who spread across the newly fixed and newly permeable borders from Uzbekistan and Kazakhstan, or from neighboring countries to trade in Central Asia (Marsden, 2012). The move to previously devalued business and trade represented a profound moral shift; a once peripheral, often illegal, activity, suddenly appeared to be everywhere and championed by elites. Whether or not the recent economic crisis has caused another bout of moral doubt, or has had other effects, has not yet been fully documented; Petric's (2013) recent volume on the longer term effects of destroying livestock in Kyrgyzstan, Yessenova (2010, 2011) on the solidification of property rights, and Bissenova's (unpub) account of Astana's housing boom and bust are exceptions.

Okada (unpub) provides an imaginative analysis of how the labor theory of value was aligned with a gendered Soviet aesthetics in Tashkent. Thus, the most highly valued paintings were on heroic scales and subjects, in oil, hung in public places and created by men; the slightness of watercolors was reserved for women and domestic spaces. With the collapse of state commissioning and the introduction of alternative systems of valuation via foreign markets, watercolors and sketches increased in value and female artists' economic position improved rapidly.

Rural poverty as a result of decollectivization, privatization, and unemployment has been relatively less studied, particularly recently. Shreeves (2002) and Kandiyoti (2003, 2004) consider gender and rural household economies in Kazakhstan and Uzbekistan, respectively. Just as cities were dotted with discarded buildings, many farms were not so much privatized as abandoned to subsistence; men feeling emasculated by the inability to provide and reliance on women (Shreeves, 2002). In a sense, both urban and rural privatizations showed the continued use of standard neoliberal development paradigms and definitions – and their failure to understand how things worked on the ground and therefore to be efficacious in alleviating new forms of poverty. Understanding rural households and economic units in terms of nuclear families, for example, did not relate to collective, multigenerational labor and property regimes (Kandiyoti, 1999). Werner (1998, 1999) also showed that exchanging lavish gifts and feasts continued to be both a major expense and a source of social/financial stability for rural Kazakh families. Since Khazanov's (1984) foundational work on steppe nomadism, Kerven's (2003) and particularly Khazanov's (2012) studies have provided a trenchant account of changing pastoral property regimes in Kazakhstan and Turkmenistan. Now it is rarely possible to survive through pastoralism alone after the rapid disinvestment of state support and continued policy neglect of rural areas, which often led to stock being sold off for immediate needs. Cartels of distributors fix prices and make survival problematic for small-scale wool, dairy, and meat producers. Khazanov states that this way of life has been irreparably broken: economic and social differentiation is growing in rural areas and pastoralism even less mobile than in the Soviet period.

As the initial shock of unfettered market processes faded or became more familiar, anthropological studies have tended to shift away from a focus on property and value to questions of representation, power inequalities, and identity, often expressed through contested notions of ethnicity, religious diversity, and their entailments. In this sense, trends may be seen as mirroring a more general turn in Western left politics from a politics of allocation to a politics of recognition and identity.

Performing Politics: Identity, Clans, and Fluid Ethnicities

The creation and securitization of new nation-states on the basis of Soviet borders is a continuing concern particularly shared by geographers and anthropologists. What ethnography adds is accounts of how ethnic fluidity in the region, arrested by the Soviet emphasis on ethnoterritoriality, resulted in multiethnic areas, or pockets of ethnic concentrations, that sit uneasily with a renewed emphasis on the nation-state as the political container of legitimate power and identity.

At a macrolevel of analysis, Roy (2000) and Smith et al. (1998) provided early accounts of the solidification of Soviet boundaries noting the difficulties with assuming that a nation-state model had any practical salience for understanding either the ethnic composition of each state or the distribution of each titular group. Kazakhstan's titular group, for example, was less than 50% in 1991 with Russian population concentrations in the north. Many of Uzbekistan's

cities were predominantly Tajik speaking, but with different ethnic concentrations within cities. The plurality and coexistence of ethnic groups within countries and cities in particular (Liu, 2012) as well as border regions (Megoran, 2007) emphasize the need for a historical context to understand the different causes and effects of successive migrations to and from the region before, during, and after the Soviet Union.

Thus, the descendants of mass deportations of Koreans, Meshketian Turks, Germans, Tatars, Germans, and Chechyns, to name but some (Oka et al., 2002), to Central Asia in the 1930s and 1940s have taken different paths after 1991. Kazakhs, for example, who left as a result of environmental devastation or Stalinist repressions have been nominally welcomed back as repatriates, but have faced huge economic and bureaucratic obstacles to reintegration, alongside a sense that the motherland is no longer a place and people with whom they can easily identify (Diener, 2009; Werner and Barcus, 2009). The fact that many of Kazakhstan's cities are primarily Russophone (Davé, 2007) adds to the sense of estrangement. While ethnic consolidation has certainly taken place, the notion of 'return' is far from straightforward, equally so for Russians who have either gone to their own alien motherland (Flynn, 2004) or remained in Central Asia as a newly disempowered minority (Laruelle and Peyrouse, 2004; Kaiser, 1998). In contradistinction to many official proclamations that titular ethnic groups have once again come into their kingdom, the prevalence of so many different groups, together with various generational affiliations within such groups, complicates the reception of such pronouncements and how they are enacted. One result is new ethnically based economic and power inequalities alongside the sheer creativity with which people manipulate and move in and out of classifications (Allès, 2005; Finke, 2006; Tishkov, 1995; Finke and Sancak, 2012).

Studies of power relations in the region are largely analyses of how and to whom power was transferred. Alexander (2007a) and Jones Luong's edited volume (2003) discuss the variety of mechanisms used from central to regional and city-level administrations; there are few such microstudies of the institutional and bureaucratic reforms that enabled the control of resources to shift hands (or stay in the same hands) so fast, simply because access to this kind of material is hard to obtain. Rather debates have centered on the extent to which kinship, in the form of clans, and state power is imbricated. In brief, Rasuly-Paleczak (2005) notes that the historical oscillation, and contemporary imbrication, of power between tribes and state forms is characteristic of Turkic polities and therefore a useful heuristic for analyzing Central Asian politics; Schatz (2004) provides the most detailed account of how this has been operationalized in Kazakhstan, suggesting that clan solidarity simply went underground during the Soviet period only to resurface in sufficiently ambiguous form also to be used as an accusation of corruption. Gullette (2010) draws on Kyrgyzstan's 2005 Tulip Revolution to note not only how frequently 'clan politics' is used as an emic term of disapprobation, but also how this obscures alternative political modalities.

Schatz's argument also traces another continuing debate: the extent to which current social forms are new; a return to pre-Soviet structures as a kind of retraditionalization (e.g., Bazin, 2008); a continuation of Soviet organization, which effectively eradicated earlier systems (Trevisani, 2010); and/or a range of hybrid forms depending on the actors and events (Kandiyoti, 1999) – returning once again to the complex fluidity characterizing Central Asia. Three examples illustrate this debate well: national cultural production; the *mahalle* (neighborhood organization); and Soviet territorial borders, which the newly independent states embraced and sought to make more rigid.

It is worth noting that first, cities are often the stage where these affirmations and contestations of power and belonging play out, continuing the Soviet emphasis on cities as the engines of progress (Alexander and Buchli, 2007) and, second, that such debates range from elite policies to everyday encounters. Typically, national legitimation by elites has been enacted in spectacular form through the mass orchestration of celebrations, recalling similar Soviet extravaganzas, such as Adams (2010) describes in Uzbekistan and through recasting the built environment to signal and enact particular political ideologies (Laszczkowski, 2011). Both Uzbekistan's and Kazakhstan's governments encourage a carefully curated version of Uzbek and Kazakh histories, respectively, using a familiar folkloric lexicon (Alexander, 2004a), effectively performing its rights to the land through blood, where the Soviet period was merely an aberrant blip (see Grant, 2001 for a similar discussion of Moscow's new monuments). Kazakhstan's elites endeavor to keep one foot in a romanticized nomadic past, the other in a hi-tec, global/Eurasian future (Laszczkowski, 2011). Liu's (2011) point is important that the Uzbek government has often practiced a foreign policy sympathetic to the interests of its allies but injurious to the Uzbek people, even as it leads celebrations of Uzbek national culture.

The *mahalle* is both a neighborhood and a form of community organization typical of Turkic cities; in Central Asia it is particularly associated with Uzbekistan. Its reemergence in urban politics has been interpreted as a traditional structure that has been co-opted by the new municipal administrations (Trevisani and Massicard, 2003; Rasanyagam, 2009), and an instance of how concentrations of different forms of sociality continue to coexist within one city, Osh, home to Uzbeks excluded from both Kyrgyzstan and Uzbekistan (Liu, 2012).

The rush of the newly independent states to secure and cement territorial borders (Reeves, 2011; Megoran, 2007) points not just to the state work of manning and creating such borders as material artifacts but the everyday work of living with and imagining such borders, particular in the Ferghana valley region where Uzbekistan and Kyrgyzstan meet. The effects and affects of such borders are manifold.

Religion

As Liu (2011: p. 120) notes, Islam has been privileged in studies of the region. In part, this is down to grand narratives from the Soviet period onward according it a monolithic role that more nuanced ethnographic accounts have been able to tease apart. Thus, in Massell's (1974) persuasive account of the early Soviet period, veiled women took the place of the missing proletariat, as symbolic objects of all that was oppressive,

patriarchal, and backward about Central Asia, even though veiling was limited to Uzbekistan and Tajikistan. There are, at least, three reasons for academic interest in Islam rising again from the 1990s onward: (1) there was a huge burst of building mosques (alongside Protestant missionary activity), many of which were funded by external sources; (2) the notion that allowing Islamic practice would allow the resurgence of related suppressed traditions; and (3) the apparent growth of Islam together with links to Muslim 'states of concern' via funding neatly replaced north American fears of communism with fears of militant Islam. On this point, Hetherington and Heathershaw and Megoran (2011) highlight the recurrent orientalist discourse of danger and violence in Central Asia that continues to affect and distort both policy and much scholarship.

For all the Soviets rejected Freudianism as linked to false bourgeois consciousness, there is a curious consistency in how Islamic practice is first seen as repressing gender equality and then in turn as being both repressed and a container of all other traditions similarly disparaged or outlawed during the Soviet Union. The problem with this archaeological, or layered, approach to a national consciousness, even if not so termed, is the temptation to act as though one need only scratch the surface for a particular social organization to be restored. Once again, there are several views on the relationship between contemporary and earlier forms of Islam in Central Asia, as well to other lost practices and indeed, to widen the frame, to Islamic states elsewhere.

Khalid (2007) suggests that the Soviet evisceration of religion was so effective, that after 1991 there was little sense of what Islam was, beyond a loose recognition that certain ethnic affiliations and religious confessions once had an affinity: to be Kazakh, was thus *ipso facto* to be Muslim, irrespective of practice, knowledge, or belief. The efflorescence of different ways of practicing Islam created new anxieties over what was or was not 'correct' or might be seen as extremist (McBrien, 2009) and who had authority to make such pronouncements (Waite, 2006). Arguably (McBrien and Pelkmans, 2008), there are parallels between the ideology of a competitive neoliberal market and the range of new religions on offer, echoing Ruthven (1991). Again, where some shrines had once been ignored, and thus covertly allowed a variety of practices, a new openness also brought out attempts to control and circumscribe observances (Alexander, 2004a; Kehl-Bodrogi, 2006). Similarly perhaps, since the domestic realm was not immediately in the public eye, it became an important conduit for continuing and passing on religious and other rituals and traditions (Kandiyoti and Azimova, 2004; Fathi, 2006). Clark (2009) notes that homes and domestic networks of families and friends are typically the means through which Protestant Christianity is also taking a strong hold in Almaty.

State responses to religion throw up the contradictions of nominally embracing liberalism while also trying to control the national ideology. Kazakhstan's president may be seen on television almost running from Eid celebrations one minute to attend Easter Mass the next. The lack of open religious debate in Uzbekistan on the other hand, suggests that, once again, it is in the less observed everyday arenas that a variety of religious and healing practices are to be found.

Conclusion

Many topics remain unstudied or scarcely covered. Energy and water politics seem to be largely the province of politics and international relations, with little sense of how energy and water shortages locally or giant pipelines using productive land are affecting people's everyday lives. Just as Central Asia received waves of settlers hungry for land after the emancipation of the serfs in 1861, political prisoners, volunteers for new enterprises, and deportations of entire populations so too was it subject to extreme Soviet agricultural, military, and other experiments, often with long-lasting, devastating environmental consequences. Three examples give an indication. Cotton monoculture in Tajikistan and Uzbekistan required high irrigation that drained the Aral Sea; the central steppe's thin soil was plowed up; northeast Kazakhstan was turned over to nuclear tests. The legacy of these interventions is extreme; so far, there have been few anthropological engagements with what they mean (Werner and Purvis-Roberts, 2006, 2007 are the exceptions here). Labor migration continues apace, between farm and city, between impoverished countryside and the West. Marat (2009) aside, there have been few studies of the effects of the economic crisis. Next to nothing has been written on industrial relations in these resource-rich countries (where strikes appear to be increasing) and on emerging forms of inequality within, as well as between, ethnic groups. It would be good to see the concern with the politics of performance and representation matched by the politics of resource allocation as the growing gap between rich and poor that marks neoliberal regimes also takes hold. Certainly, the increasing distance between rich and poor, urban and rural areas often outweighs ethnic affiliation.

In returning to the movement, fluidity and far-reaching imaginations that have always characterized this region, ethnography is a vital means of countering stereotypes encountered both externally and in self-exegeses – whether demonstrating transformative, traditional or hybrid, syncretic patterns of social life. If early accounts emphasized top-down nation-state building, ethnography shows the inadequacy of the nation-state concept in suggesting the lived plurality of lives in Central Asia. Fully capturing the far-flung networks of ethnic groups that variously live (often choosing to live) in diasporic conditions, in mobile trading relationships, and in their titular state has yet to happen. Equally, a move to consider Central Asia as part of global processes of accumulation by dispossession in the late twentieth and twenty-first centuries would enable a stronger sense of its commonality and specificity juxtaposed with global economies of which it is now irrevocably part.

See also: Ethnic Identity and Ethnicity in Archaeology; Ethnicity and Migration in Europe; Ethnicity: Anthropological Aspects; International Migration by Ethnic South and Southeast Asians; Islam: Asia; Migration History; Migration, Theory of; Migration: Anthropological Perspectives; Migration: Cultural Aspects; Migrations, Colonizations, and Diasporas in Archaeology; Nations and Nation-States in History; Tradition, Anthropology of.

Bibliography

Adams, L., 2010. The Spectacular State: Culture and National Identity in Uzbekistan. Duke University Press, Durham.

Alexander, C., 2004a. The cultures and properties of decaying buildings. Focaal: European Journal of Anthropology 44, 48–60.

Alexander, C., 2004b. Values, relations and changing bodies: industrial privatisation in Kazakhstan. In: Humphrey, C., Verdery, K. (Eds.), Property in Question: Appropriation, Recognition and Value Transformation in the Global Economy. Berghahn Books, Oxford.

Alexander, C., 2007a. Soviet and post-Soviet city planning: a case study from Almaty. Critique of Anthropology 27 (2), 165–182.

Alexander, C., 2007b. Almaty: rethinking the public sector. In: Alexander, C., Buchli, V., Humphrey, C. (Eds.), Urban Life in Post-Soviet Asia. UCL Press, Routledge, pp. 70–101.

Alexander, C., Buchli, V., 2007. Introduction. In: Alexander, C., Buchli, V., Humphrey, C. (Eds.), Urban Life in Post-Soviet Asia. UCL Press, London.

Alexander, C., Buchli, V., Humphrey, C., 2007. Urban Life in Post-Soviet Asia. UCL Press, Routledge, London.

Allès, E., 2005. The Chinese-speaking Muslims (Dugans) of Central Asia: a case of multiple identities in a changing context. Asian Ethnicity 6 (2), 121–134.

Appadurai, A., 1990. Disjuncture and difference in the global cultural economy. In: Featherstone, M. (Ed.), Global Culture. Sage, London.

Bazin, L., 2008. Some responses to globalisation in Uzbekistan: state authoritarianism, migrant labour and neo-traditionalism. Anthropology of the Middle East 3 (1), 10–27.

Bellér-Hann, I., Cesàro, M., Harris, R., Smith Finley, J. (Eds.), 2007. Situating the Uyghars between China and Central Asia. Ashgate, Aldershot.

Beyer, J., Rasanayagam, J., Reeves, M. (Eds.), 2013. Performing Politics: Ethnographies of the State in Central Asia. Indiana University Press, Bloomington.

Bissenova, A. Post-socialist Dreamworlds: Housing Boom and Urban Development in Kazakhstan. (Ph.D. dissertation), University of Cornell, unpub.

Chari, S., Verdery, K., 2009. Thinking between the posts: postcolonialism, post-socialism, and ethnography after the Cold War. Comparative Studies in Society and History 51 (1), 6–34.

Clark, W., 2009. Networks of faith in Kazakhstan. In: Pelkmans, M. (Ed.), Conversion after Socialism: Disruptions, Modernisms and Technologies of Faith in the Former Soviet Union. Berghahn, Oxford.

Cummings, S. (Ed.), 2010. Symbolism and Power in Central Asia: Politics of the Spectacular. Routledge, London.

Davé, B., 2007. Kazakhstan. Ethnicity, Language and Power. Routledge, London.

Diener, A., 2009. One Homeland or Two? The Nationalization and Transnationalization of Mongolia's Kazakhs. Stanford University Press, Washington, DC.

Fathi, H., 2006. Gender, Islam, and social change in Uzbekistan. Nationalities Papers 25 (3), 303–317.

Finke, P., 2006. Variations on Uzbek Identity: Concepts, Constraints and Local Configurations. Berghahn Books, Oxford, New York.

Finke, P., Sancak, M., 2012. To be an Uzbek or not to be a Tajik. Ethnicity and locality in the Bukhara Oasis. Zeitschrift für Ethnologie 13 (1), 47–70.

Flynn, M., 2004. Migrant Resettlement in the Russian Federation: Reconstructing 'Homes' and 'Homelands'. Anthem Press, London.

Grant, B., 2001. New Moscow monuments, or states of innocence. American Ethnologist 28 (2), 332–362.

Gullette, D., 2010. The problems of the 'clan' politics model of Central Asian statehood: a call for alternative pathways for research. In: Kavalski, E. (Ed.), Stable Outside, Fragile Inside?: Post-Soviet Statehood in Central Asia. Farnham, Surrey, England. Ashgate Publishers, Burlington, VT.

Hann, C., 2003. The Postsocialist Agrarian Question: Property Relations and the Rural Condition. LIT (with the "Property Relations" Group), Münster.

Hann, C. (Ed.), 2006. The Postsocialist Religious Question: Faith and Power in Central Asia and East Central Europe. Lit Verlag, Munich.

Harris, C., 2004. Control and Subversion: Gender Relations in Tajikistan (Anthropology, Culture and Society). Pluto Press, London.

Heathershaw, J., Megoran, N., 2011. Contesting danger: a new agenda for policy and scholarship on Central Asia. International Affairs 87 (3), 589–612.

Jacquesson, S., 2010. Pastoréalisms – Anthropologie Historique des Processus d'intégration chez les Kirghiz du Tian Shan Intérieur. In: Nomaden und Sesshafte, 14. Dr. Ludwig Reichert Verlag, Wiesbaden.

Jones Luong, P. (Ed.), 2003. The Transformation of Central Asia. Cornell University Press, Ithaca.

Kaiser, M., 1998. Russians as minority in Central Asia. In: Migration. A European Journal of International Migration and Ethnic Relations. Special Issue: The Former Soviet Union, 31, pp. 83–117.

Kaiser, M., 2005. Cross-border traders as transformers. In: Kollmorgen, R. (Ed.), Transformation als Typ Sozialen Wandel. Postsozialistische Lektionen, Historische und Interkulturelle Vergleiche. LIT Verlag, Münster, Hamburg, London.

Kandiyoti, D., 1999. Poverty in transition: an ethnographic critique of household surveys in post-Soviet Central Asia. Development and Change 30, 499–524.

Kandiyoti, D., 2002a. Post-colonialism compared: potentials and limitations in the Middle East and Central Asia. International Journal of Middle East Studies 34, 279–297.

Kandiyoti, D., 2002b. How far do analyses of postsocialism travel? The case of Central Asia. In: Hann, C.M. (Ed.), Postsocialism. Ideals, Ideologies and Practices in Eurasia. Routledge, London, pp. 238–257.

Kandiyoti, D., 2003. The cry for land: agrarian reform, gender and land rights in Uzbekistan. Journal of Agrarian Change 3 (1–2), 225–256.

Kandiyoti, D., 2004. Post Soviet Institutional Design, NGO's and Rural Livelihoods in Uzbekistan, Civil Society and Social Movements Programme Paper 11. UNRISD.

Kandiyoti, K., Azimova, N., 2004. The communal and the sacred: women's worlds of ritual in Uzbekistan. Journal of the Royal Anthropological Institute 10 (2), 327–349.

Kehl-Bodrogi, K., 2006. Islam contested: nation, religion, and tradition in post-Soviet Turkmenistan. In: Hann, C., "Civil Religion" Group (Eds.), The Postsocialist Religious Question: Faith and Power in Central Asia and East-Central Europe. LIT, Münster, pp. 125–146.

Kerven, C., 2003. Prospects for Pastoralism in Kazakhstan and Turkmenistan: From State Farms to Private Flocks. Routledge, London.

Khalid, A., 2007. The Politics of Muslim Cultural Reform: Religion and Politics in Central Asia. University California Press, Berkeley.

Khazanov, A., 1984. Nomads and the Outside World. Cambridge University Press, Cambridge.

Khazanov, A., 2012. Pastoralism and property relations in contemporary Kazakhstan. In: Khazanov, A., Schlee, G. (Eds.), Who Owns the Stock? Collective and Multiple Property Rights in Animals. Berghahn Books, Oxford.

Lane, C., 1981. Rites of Rulers: Ritual in Industrial Society – The Soviet Case. Cambridge University Press, Cambridge.

Laruelle, M., Peyrouse, S., 2004. Les Russes du Kazakhstan. Identités Nationales et Nouveaux Etats dans l'espace Post-Soviétique (Russians in Kazakhstan. National Identities and New States in the Post-Soviet Space). Maisonneuve & Larose/IFEAC, Paris.

Laszczkowski, M., 2011. Building the future: construction, temporality, and politics in Astana. Focaal 60, 77–92.

Levi, S., Sela, R., 2009. Islamic Central Asia: An Anthology of Historical Sources. Indiana University Press, Bloomington.

Liu, M., 2011. Central Asia in the post–Cold War world. Annual Review of Anthropology 40, 115–131.

Liu, M., 2012. Under Solomon's Throne: Uzbek Visions of Renewal in Osh. University of Pittsburgh Press, Pittsburgh.

Lubin, N., 1985. Labour and Nationality in Soviet Central Asia. Princeton University Press, Princeton.

Marat, E., 2009. Labor migration in Central Asia: implications of the global economic crisis. Anthropological Quarterly 79 (3), 431–461.

Marsden, M., 2012. Fatal embrace: trading and hospitality on the frontiers of South and Central Asia. Journal of the Royal Anthropological Institute 18 (s1), 117–130.

Marsden, M., 2013. Southwest and Central Asia: comparison, integration or beyond. In: Fardon, R. (Ed.), The Sage Handbook of Social Anthropology. Sage, Los Angeles.

Masanov, N.E., 1996. Kazakhstanskaya politicheskaya i intllektualnaya elita: klanovaya prinadlezhnost' i vnutrietnichskoe sopernichestvo (The Kazakhstan political and intellectual Elite: Clan belonging and interethnic rivalry). Vestnik Evrazii 1, 46–61.

Massell, G., 1974. The Surrogate Proletariat: Moslem Women and Revolutionary Strategies in Soviet Central Asia, 1919–1929. Princeton University Press, Princeton.

McBrien, J., 2009. Mukadas's dilemma: veils and modernity in Kyrgyzstan. Journal of the Royal Anthropology Institute 15 (1), S127–S144.

McBrien, J., Pelkmans, M., 2008. Turning Marx on his head: missionaries, 'extremists' and archaic secularists in post-Soviet Kyrgyzstan. Critique of Anthropology 28, 87–103.

Megoran, N., 2007. On researching ethnic conflict: epistemology, politics and a Central Asian boundary dispute. Europe-Asia Studies 59 (2), 253–277.

Mühlfried, F., Sokolovskiy, S. (Eds.), 2012. Exploring the Edge of Empire. Soviet Era Anthropology in the Caucasus and Central Asia. Halle Studies in the Anthropology of Eurasia, vol. 25. Lit-Verlag, Berlin.

Naumova, O.B., 1991. Sovremennye etnokul'turnye protsessy kazakhov v mnogonatsional'nykh rayonakh Kazakhstana (Contemporary Ethnic and Cultural Processes in Multicultural Regions of Kazakhstan) (Ph.D. thesis), Institute of Ethnology and Anthropology of the Russian Academy of Sciences, Moscow.

Nazpary, J., 2001. Post-Soviet Chaos: Violence and Dispossession in Kazakhstan. Pluto Press, London.

Newton, S., 2001. Conflicting Interests: The "Great Game" Replayed? Conference Presentation, Royal Society of Asian Affairs Conference, SOAS.

Norris, S., 2012. Nomadic nationhood: cinema, nationhood, and remembrance in post-Soviet Kazakhstan. Ab Imperio 2, 378–402.

Oka, N., Masanov, N., Karin, E., Chebotarev, A., 2002. The Nationalities Question in Post-Soviet Kazakhstan. Institute of Developing Economies, Chiba.

Okada, K. Re-legitimising Painting in Tashkent Uzbekistan after Fall of the Soviet Union (1991–2004) (Ph.D. dissertation), University of London, Goldsmiths, unpub.

Petric, B., 2013. On a Mange nos Moutons: le Kighizstan, du Berger au Biznesman. Belin, Paris.

Peyrouse, S., 2012. Turkmenistan: Strategies of Power, Dilemmas of Development. M.E. Sharpe, Armonk, NY.

Piirainen, T., 1996. The Death of the Iron Mother: Transition to Market Economy in an Industrial Town in Kazakhstan, EU-TACIS, Social Impacts of Privatisation and Economic Restructuring I Kazakhstan with a Focus on Aktyubinsk Oblast. TACIS Publications, Almaty. European Commission: Brussels.

Rasanayagam, J., 2009. Morality, self and power: the idea of the *mahalla* in Uzbekistan. In: Heintz, M. (Ed.), The Anthropology of Moralities. Berghahn Books, Oxford, pp. 102–117.

Rasanyagam, J., 2011. Informal economy, informal state: the case of Uzbekistan. International Journal of Sociology and Social Policy 31 (11–12), 681–696.

Rasuly-Paleczak, G., 2005. Comparative perspectives on Central Asia and the Middle East in social anthropology and the social sciences. Central Eurasian Studies Review Publication of the Central Eurasian Studies Society 4 (2), 2–17.

Reeves, M., 2011. Fixing the border: on the affective life of the state in southern Kyrgyzstan. Environment and Planning D: Society and Space 29 (5), 905–923.

Roy, O., 2000. The New Central Asia: The Creation of Nations. New York University Press, New York.

Ruthven, M., 1991. The Divine Supermarket: Shopping for God in America. Vintage, London.

Schatz, E., 2004. Modern Clan Politics: The Power of 'Blood' in Kazakhstan and Beyond. University of Washington Press, Seattle, WA.

Shreeves, R., 2002. Gender and development in rural Kazakhstan. In: Mandel, R., Humphrey, C. (Eds.), Markets and Moralities: Ethnographies of Postsocialism. Berg, Oxford.

Smith, G., Law, V., Wilson, A., Bohr, A., Allworth, E., 1998. Nation-Building in the Post-Soviet Borderlands. Cambridge University Press, Cambridge.

Spector, R., 2008. Bazaar politics: the fate of the market place in Kazakhstan. Problems of Post-communism 55 (6), 42–53.

Stoler, A., 2009. Along the Archival Grain: Epistemic Anxieties and Colonial Common Sense. Princeton University Press, Princeton.

Tishkov, V., 1995. Don't kill me, I'm a Kyrgyz! An anthropological analysis of violence in the Osh ethnic conflict. Journal of Peace Research 32 (2), 133–149.

Trevisani, T., 2010. Land and Power in Khorezm: Farmers, Communities and the State in Uzbekistan's Decollectivisation, Halle Studies in the Anthropology of Eurasia, vol. 23. LIT Verlag, Berlin.

Trevisani, T., Massicard, E., 2003. The Uzbek Mahalla: between state and society. In: Everett-Heath, T. (Ed.), Central Asia. Aspects of Transition. Routledge/Curzon, London, New York.

Waite, E., 2006. The impact of the State on Islam amongst the Uyghurs: religious knowledge and authority in the Kashgar Oasis. Central Asian Survey 25 (3), 251–265.

Werner, C., 1998. Household networks and the security of mutual indebtedness in rural Kazakhstan. Central Asian Survey 17 (4), 597–612.

Werner, C., 1999. The dynamics of feasting and gift exchange in rural Kazakhstan. In: Svanberg, I. (Ed.), Contemporary Kazaks: Social and Cultural Perspectives. Curzon Press, London, pp. 47–72.

Werner, C., 2004. Feminizing the new Silk Road: women traders in rural Kazakhstan. In: Nechemias, C., Kuehnast, K. (Eds.), Post-Soviet Women Encountering Transition: Nation-Building, Economic Survival, and Civic Activism. Johns Hopkins University Press, Baltimore, pp. 105–126.

Werner, C.A., Purvis-Roberts, K.L., 2006. After the cold war: international politics, domestic policy, and the nuclear legacy in Kazakhstan. Central Asian Survey 25, 461.

Werner, C., Purvis-Roberts, K., 2007. Unravelling the secrets of the past: contested versions of nuclear testing in the Soviet Republic of Kazakhstan. In: Johnson, B. (Ed.), Half-lives and Half Truths: Confronting the Radioactive Legacy of the Cold War. School for Advanced Research Press, Santa Fe, New Mexico.

Yessenova, S., 2006. Hawkers and containers in Zarya Vostoka: how 'bizarre' is the post-Soviet bazaar? In: Dannhaeuser, N., Werner, C. (Eds.), Markets and Market Liberalisation: Ethnographic Reflections, Research in Economic Anthropology. Texas A&M University, pp. 37–59.

Yessenova, S., 2010. Borrowed places: eviction wars and property rights formalization in Kazakhstan. Research in Economic Anthropology 30, 11–45.

Yessenova, S., 2011. Nomad for export, not for domestic consumption: Kazakhstan's arrested endeavour to 'put the country on the map'. Studies in Russian & Soviet Cinema 5 (2), 181–203.

Ceramics in Archaeology

Patrick V Kirch, University of California, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 1623–1627, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Ceramics, among the oldest and most significant technological innovations in the history of humankind, are ubiquitous in many archaeological contexts, and provide archaeologists with one of their main categories of empirical data. Studies of variation in ceramic production, style, and use have assisted archaeologists both in the construction of chronologies and in the interpretation of ancient societies. Four main areas of archaeological study of ceramics are reviewed: (a) the classification of pottery; (b) the use of ceramics in chronology construction and dating; (c) the characterization and sourcing of pottery; and (d) ethnoarchaeological studies of traditional pottery-making industries.

Ceramics, or pottery, are among the oldest and most significant technological innovations in the history of humankind, the first truly synthetic material. Being highly plastic and thus virtually infinite in the range of shapes and forms possible, offering ready surfaces for decoration, and being ubiquitous in many archaeological contexts, ceramic containers and artifacts have provided archaeologists with one of their main categories of empirical data. Studies of variation in ceramic production, style, and use have assisted archaeologists both in the construction of chronologies and in the interpretation of ancient societies.

Ceramics denotes objects made from clay and fired to achieve hardness, which range from low-fired terra- cottas and earthenwares (firing range 900–1200 °C), through stonewares (1200–1350 °C), to porcelains (1300–1450 °C) (Gosselain, 1992). Most archaeological analyses of ceramics are focused on earthenwares, which dominated pottery industries throughout most of prehistory. Aside from vessels and containers of all kinds, ceramic artifacts also include such important items as bricks and tiles, figurines and models, and tableware.

History of Ceramics

The earliest human experimentation with clay – including the discovery that moist clay is plastic, can be shaped, and when heated or fired will retain its shaped form – is not well documented archaeologically. Fired clay figurines are known from the Dolní Vëstonice site in Czechoslovakia, dated to 30 000 BC (Vandiver et al., 1991). In the Old World, the first ceramic vessels appear in the early Jomon culture of Japan at about 10 000 BC, and in the Near East (Anatolia) at around 8500–8000 BC. Ceramics do not appear until considerably later in the New World, where it appears that the technology was independently developed in several localities between 5500–2500 BC. In all cases, the origins of ceramic vessels appear to be closely associated with sedentary modes of life, and with storage either of agricultural products or of quantities of gathered plant foods (as in the Jomon case). There is a high degree of correlation between sedentism and pottery making, as evident in a world ethnographic sample where only two out of 46 pottery-making societies are nonsedentary (Eerkens et al., 2002).

The oldest known pottery in the world is the 'cord-marked' Jomon pottery industry of Japan, dating to 10 000 BC and associated with sedentary or semi-sedentary hunter-gatherers (Zhushchikhovskaya, 2012). Jomon pottery is highly distinctive, both with its cord or string-impressed surface decoration, and its elaborate modeled rims; vessel forms are typically high jars or beakers. Elsewhere in the Far East, pottery appears in parts of China by around 7000 BC, if not earlier. The Yangshao-Culture, centered in the Yellow River valley of northern China and dating between 4800–4200 BC, is noted for its beautifully formed earthenware jars and dishes, decorated with red-and-black painted geometric designs. In later time periods, the Chinese perfected many technical aspects of ceramic production, such as the horizontal and vertical updraft kilns, the potter's wheel, and high-fired stonewares and glazes (Kuzmin, 2006). Porcelain ceramics were innovated in China sometime around the beginning of the first millennium AD, and were widely traded throughout the Old World, and later to Europe and the Americas.

In the Near East, architectural uses of clay (e.g., for walls, floors, and roofs) are widespread by around 7500 BC, in association with the domestication of plants and animals, and the origins of sedentary village life. Pottery containers appear by 8500–8000 BC in Anatolia, and slightly later in other parts of the Near East. A variety of other clay implements and artifacts are also characteristic of early Near Eastern sites, such as clay spindle whorls and loom weights, clay sickles (with obsidian or flint blades), clay toys and models, and clay stamps and cylinder seals (Byrd, 2005). Ceramic technology was not independently developed in Europe, but rather was borrowed or diffused from the Near East, in association with the spread of agriculture and a sedentary mode of existence. In the classical world, of course, Greek black- and red-figured pottery (600–400 BC) represents a high level of technical and artistic achievement, as does slightly later Roman Arretine ware (100 BC–AD 400).

Ceramic technology in the New World arose considerably later in time, evidently as an independent development; moreover, the potter's wheel and the use of glazes never developed in the New World. The earliest known pottery in

the New World is from the Amazon basin and dates to approximately 5500 BC (Roosevelt et al., 1991; *see* South America, Archaeology of). Fiber-tempered pottery has been found in the Southeastern United States from 2500 BC (Saunders and Hays, 2006) while pottery in Pacific coastal Mesoamerica dates to approximately 2000 BC (Clark and Gosser, 1995). Among the outstanding pottery traditions of the New World, one must count the coastal Peruvian Nazca and Moche traditions (200 BC to AD 700) with their elaborated modeled anthropomorphic and zoomorphic vessels (Chatfield, 2007); the Late Classic Maya pottery which included exquisite polychrome funerary vessels; and the Pueblo pottery traditions of the southwestern United States.

Archaeological Approaches to Ceramics

Ceramics have played a central role in archaeological method and theory, for several reasons: (a) pottery has a long history, and is virtually ubiquitous in most sedentary societies; (b) pottery is nonperishable, and is often recovered in very large quantities from archaeological excavations; (c) pottery functioned both as utilitarian cooking, storage, and serving vessels for all strata of a society, as well as special purpose functions for elite ceremonial or funerary use; and, (d) as a highly plastic material, pottery displays seemingly infinite variation in composition, manufacturing method, shape, and decoration (Orton and Hughes, 2013) . These variations, moreover, were culturally conditioned, resulting in particular ceramic styles or traditions, which can be traced through time and space.

Early in the development of modern archaeology, scholars such as Flinders Petrie in Egypt, and somewhat later James Ford, Irving Rouse, and James Griffin in North America, recognized the value of ceramic studies in developing relative chronologies based on changes in ceramic style over time. Prior to the development of such absolute chronometric dating methods as radiocarbon or dendrochronology (*see* Chronology, Stratigraphy, and Dating Methods in Archaeology), the construction of ceramic chronologies were essential to developing local and regional time frames for prehistoric cultures.

With the rise of absolute dating methods in the second half of the twentieth century, archaeological studies of ceramics have shifted somewhat, with less emphasis on classification, seriation, and chronology construction, in favor of new approaches. These include detailed physical and chemical studies of ceramic composition and of production techniques and sequences, which inform archaeologists about the technology of ceramic production and use, and about the distribution and movement of ceramic vessels during their life spans (Sharer and Ashmore, 2003). Such information in turn is useful to archaeologists who are attempting to understand prehistoric economic and sociopolitical organization.

Classification of Ceramics

As perhaps the most plastic of all material culture media known in antiquity, ceramics exhibit enormous variation – in fabric, methods of production, shape, decoration, and function – offering tremendous advantages as well as challenges to the archaeologist. In order to bring some order to this variability, archaeologists must classify pottery into sets of like objects. The aims and methods of ceramic classification, however, depend greatly upon the particular archaeological approach. For example, a classification which is aimed at the discovery of historical types, those which show meaningful variation over time, will most likely emphasize different aspects of variation from a classification designed to exhibit key differences in manufacturing process. The literature on archaeological classification of ceramics is vast, but useful overviews may be found in Shepard (1965), Rice (1987), and Sinopoli (1991).

Ceramic classifications may be based on a number of different kinds of variables, the most common being technological dimensions, vessel shape, and decoration (Orton and Hughes, 2013). Classifications based on technological dimensions are especially useful when the aim is to understand production processes, although such dimensions may also be useful in the definition of historical types. Major technological dimensions include raw materials (especially clay and nonplastic inclusions such as mineral temper), methods of vessel forming (such as coiling or slab building, and the use of the potter's wheel), secondary treatments (such as the application of slips or glazes), and methods of firing (open air firing, use of kilns). Vessel shape may be classified formally according to various systems, such as that developed by Shepard (1965), which distinguishes between restricted and unrestricted vessels, and further between composite and inflected shapes. Vessel shape, naturally, is closely linked with vessel function. The third major category of variation used in ceramic classification is that of decoration and surface treatment. Possible surface treatments include slipped, glazed, burnished, polished, paddle-impressed, and smoothed. Decorations may be applied by painting, stamping, incising, carving, or other methods, such as three-dimension appliqué or relief. Designs used in decorations typically consist of individual design elements, which are systematically organized into motifs and larger decorative panels, and generally follow culturally-determined rules. The classification and analysis of such design systems may be based on individual motif catalogs and set of design rules (e.g., Mead et al., 1975), or on analysis of underlying principles of symmetry (e.g., Washburn, 1977).

Methods of ceramic classification also vary. Pottery producers have their own indigenous or folk classification systems, which have been the subject of considerable ethnoarchaeological study (see Section Ethnoarchaeology and Ceramics below). Formal archaeological procedures for classifying pottery include the well-developed 'type-variety' system, first applied in North America and later extended to Mesoamerica. This system uses a 'binomial' nomenclature, in which a geographic name is combined with some specific technological descriptor (e.g., 'Barton incised'). Classification schemes range from simple paradigmatic classifications, to complex, hierarchical taxonomies. An alternative approach, much favored in the 1960s and 1970s, is quantitative or phenetic classification, in which 'types' are generated by a computer, following certain mathematical algorithms operating upon a number of qualitative and/or quantitative parameters. The prime example of such a numerical taxonomy of ceramics is that of Clarke (1970) for the Bell Beaker pottery

complex of Great Britain. Scholars in have revisited the use of quantitative methods in the 21st century as computers have become increasingly powerful and capable of analyzing very large quantities of data (Smith et al., 2014).

Ceramics and Chronology

The emphasis accorded ceramic studies by archaeologists reflects the importance of using ceramic change as a means for constructing cultural chronologies. In the late nineteenth century, pioneering Egyptologist Flinders Petrie recognized that pottery vessels which had been placed in tomb groups as funerary objects showed subtle but continuous stylistic changes over time (Stevenson, 2014). Petrie arranged representative vessels in an inferred chronological sequence, resulting in the first seriation of pottery.

The methods of seriation were greatly refined by American archaeologists, such as Irving Rouse and James A. Ford (1962), working with ceramic assemblages from various New World localities, including the Mississippi Valley region, and the Viru Valley of Peru (O'Brien et al., 2000). Their work generated much debate concerning the methods of ceramic classification and the reality of ceramic 'types.' The fundamental principle of seriation was to define a set of historical types which displayed gradual temporal changes, with a particular type arising at some point in time, gradually increasing in popularity (hence reflected by increased frequency in archaeological assemblages), and later decreasing until the type disappeared from the archaeological record (Peeples and Schachner, 2014). When the frequencies of such historical types were plotted as percentage diagrams, they displayed a characteristic frequency distribution resembling the plan of a battleship, and hence were known as 'battleship curves.'

The great advantage of seriation was that it permitted the construction of cultural chronologies independent of any 'absolute' method of direct dating. Surface collections of potsherds from sites of unknown age could be tabulated according to the frequency of key ceramic types, and then chronologically ordered by arranging the frequency distributions to form 'battleship curves.' Moreover, local ceramic chronologies could be linked together using trade pottery which occurred in more than one region, or by tracing the diffusion of particular stylistic traits. With the invention of radiocarbon dating and other methods of 'absolute' dating in the later half of the twentieth century, the importance of seriation has declined, although it is still important as a cross-check on radiocarbon-based chronologies (Peeples and Schachner, 2014).

Pottery may also be directly dated, either by radiocarbon dating of organic inclusions within the ceramic fabric (i.e., dung or chaff included as temper), or by thermoluminescence (TL) dating. TL dating is based on the fact that when clay and other geological inclusions in pottery are fired at temperatures of 500 °C or higher, electrons which had been 'trapped' in the crystal lattice structure are freed, emitting light or thermoluminescence (Raju et al., 2014). Following the firing process, new trapped electrons gradually accumulate in the crystalline imperfections of the pottery, as a natural consequence of radioactive decay. If ancient pottery is then reheated in the laboratory up to 500 °C, and the emitted light is measured and plotted as a 'glow curve' of intensity vs. temperature, the age of the specimen can be calculated, since the intensity of light emitted will be proportional to age. There are, of course, many possible complications deriving from the geological composition of the fabric, and from post-depositional conditions affecting the pottery. As a result, TL dating is less widely used than the radiocarbon method.

Compositional Studies

Within many ancient societies, pottery was produced by specialists who then traded, exchanged, or sold their wares to other nonpottery producing sectors of society, and/or to other villages or geographic localities. Moreover, pottery was frequently traded or exchanged over considerable distances. Tracing the production, distribution, and specialized use of pottery in ancient societies requires that the archaeologist be able to characterize the unique composition of a particular ceramic product, typically a mix of clay and other nonplastic inclusions (including purposefully added 'temper.') A key phase in archaeological pottery analysis is thus ceramic characterization (Bishop et al., 1982). Characterization may also include efforts at sourcing, in which the materials that make up a particular ceramic ware are traced to their geographic points of origin, such as local clay quarries or sources of sand used as temper.

A wide range of mineralogical and geo-chemical techniques have been applied to ceramic compositional analysis, whether for characterization or sourcing. Petrographic analysis of the nonplastic inclusions within a ceramic fabric, in which the specific mineral grains are identified by examining a thin section of the pottery under a polarizing microscope, is a widely used technique (Baxter et al., 2008). X-ray diffraction, which identifies minerals by their crystalline structures, is another frequently applied method for the characterization of pottery on the basis on the temper or non-plastic inclusions. Other techniques which have been applied more recently include optical emission spectroscopy, X-ray fluorescence spectroscopy, atomic absorption spectroscopy, neutron activation analysis, proton-induced X-ray emission, Mössbauer spectroscopy, electronic microprobe analysis, and inductively-coupled plasma analysis (see Rice, 1987 for a review of these and other methods; Pillay et al., 2000).

Ethnoarchaeology and Ceramics

Given the importance of ceramics in archaeology, it is not surprising that archaeologists have turned to traditional pottery-making societies to learn more about potential variability in ceramic production, distribution, use, and discard. The study of contemporary peoples, using ethnographic methods of participant-observation, in order to gain knowledge of material culture variability which is potentially applicable to the interpretation of archaeological assemblages, is called ethnoarchaeology (*see* Ethnoarchaeology). Ceramic ethnoarchaeology (Kramer, 1985; Stark, 2003) is perhaps one of them most important subfields within this topic (Longacre, 1991).

Because archaeologists had long been concerned with the classification of ceramics, a number of ethnoarchaeological studies have focused on the ways in which traditional potters

classified or categorized their own products. Indigenous potters typically pay little attention to the technological attributes often accorded emphasis by archaeologists (such as details of temper, paste, surface treatment, or decoration), but rather emphasize general function in their folk classifications. Thus, among the Kalinga of the Philippines, pottery vessels are distinguished by whether they are intended for cooking rice or for cooking vegetables and meat, and by their sizes (Beck, 2006; Longacre, 1981). The Fulani of Cameroon (David and Hennig, 1972) lexically discriminate among five size classes of jars, with secondary classification based on their intended contents.

Other ethnoarchaeological studies of pottery have focused on aspects of ceramic production, including the social role or status of potters in their societies, on ceramic distribution, on the use of vessels, their life spans, and on their breakage and discard rates (see Kramer, 1985 for a general review). Such studies have aided archaeologists in their interpretations of ancient ceramics by revealing the complex linkages between behavior and material culture, demonstrating that strictly utilitarian explanations for archaeological phenomena are not always preferable, and by showing that multiple lines of evidence may help to discriminate between alternative explanations (Stark, 2003).

See also: Art History; Art, Anthropological Aspects of; Chronology, Stratigraphy, and Dating Methods in Archaeology; Ethnoarchaeology; Intensification and Specialization, Archaeology of; Trade and Exchange, Archaeology of.

Bibliography

Arnold, D.E., 1985. Ceramic Theory and Culture Process. Cambridge University Press, Cambridge, UK.

Baxter, M.J., Beardah, C.C., Papageorgiou, I.I., Cau, M.A., Day, P.M., Kilikoglou, V.V., 2008. On statistical approaches to the study of ceramic artefacts using geochemical and petrographic data. Archaeometry 50 (1), 142–157.

Beck, Margaret E., 2006. Midden ceramic assemblage formation: a case study from Kalinga, Philippines. American Antiquity 71 (1), 27–51.

Bishop, R.L., Rands, R.L., Holley, G.R., 1982. Ceramic compositional analysis in archaeological perspective. In: Schiffer, M. (Ed.), In Advances in Archaeological Method and Theory, Vol. 5. Academic Press, New York.

Byrd, B.F., 2005. Reassessing the emergence of village life in the near East. Journal of Archaeological Research (3), 231.

Chatfield, M., & University of California, S. (2007). From Inca to Spanish Colonial: Transitions in ceramic technology. Dissertation Abstracts International.

Clark, J.E., Gosser, D., 1995. Reinventing Mesoamerica's first pottery. In: Barnett, W., Hoopes, J.W. (Eds.), The Emergence of Pottery: Technology and Innovation in Ancient Societies. Smithsonian Institution Press, pp. 209–221.

Clarke, D.L., 1970. Beaker Pottery of Great Britain and Ireland. Cambridge University Press, Cambridge, UK.

David, N., Hennig, H., 1972. The Ethnography of Pottery: A Fulani Case Seen in Archaeological Perspective. Addison-Wesley, Reading, MA.

Deetz, J., 1965. The Dynamics of Stylistic Change in Arikara Ceramics. University of Illinois Press, Chicago.

Eerkens, Jelmer W., Neff, Hector, Glascock, Michael D., 2002. Ceramic production among small-scale and mobile hunter and gatherers: a case study from the southwestern great basin. Journal of Anthropological Archaeology 21 (2), 200–229.

Ford, J.A., 1962. A Quantitative Method of Deriving Cultural Chronology. Pan American Union Technical Manual No. 1, Washington, DC.

Gosselain, O., 1992. Bonfire of the enquiries. Pottery firing temperatures in archaeology: What for ? (English). Journal of Archaeological Science 19 (3), 243–259.

Kramer, C., 1985. Ceramic ethnoarchaeology. Annual Review of Anthropology 14, 77–120.

Kuzmin, Y.V., 2006. Chronology of the earliest pottery in East Asia: progress and pitfalls. ANTIQUITY-OXFORD 80 (308), 362.

Longacre, W.A., 1981. Kalinga pottery, an ethnoarchaeological study. In: Hodder, I., Issac, G., Hammond, N. (Eds.), Pattern of the Past. Cambridge University Press, Cambridge, UK.

Longacre, W.A., 1991. In: Longacre, William A. (Ed.), Ceramic Ethnoarchaeology. University of Arizona Press, Tucson.

van der Leeuw S E, Pritchard A C (eds.) The Many Dimensions of Pottery: Ceramics in Archaeology and Anthropology. University of Amsterdam, Amsterdam.

Mead, S.M., Birks, L., Birks, H., Shaw, E., 1975. The Lapita Pottery Style of Fiji and its Associations. Polynesian Society Memoir No. 38, Wellington (New Zealand).

O'Brien, M.J., Lyman, R., Darwent, J., 2000. Time, space, and marker types: James A. Ford's 1936 chronology for the lower Mississippi valley. Southeastern Archaeology (1), 46. http://dx.doi.org/10.2307/40713178.

Orton, Clive, Hughes, Michael, 2013. The Potential of Pottery as Archaeological Evidence. In: Pottery in Archaeology. Cambridge University Press, New York.

Peeples, M., Schachner, G., 2014. Refining correspondence analysis-based ceramic seriation of regional data sets (English). Journal of Archaeological Science 39 (8), 2818–2827.

Pillay, A., Punyadeera, C., Jacobson, L., Eriksen, J., 2000. Analysis of ancient pottery and ceramic objects using x-ray fluorescence spectrometry. X-Ray Spectrometry 29 (1), 53–62.

Raju, S., Sudhakar, B., Reddy, B., Dhoble, S.J., Thyagarajan, K.K., Raju, C., 2014. Synthesis, photoluminescence and thermoluminescence properties of Sm and Dy ions doped barium gadolinium titanate ceramics. Ferroelectrics Letters Section 41 (1-3), 9–19.

Rice, P.M., 1987. Pottery Analysis: A Sourcebook. University of Chicago Press, Chicago.

Roosevelt, A.C., Housley, R.A., Da Silveira, M.I., Maranca, S., Johnson, R., 1991. Eighth millennium pottery from a prehistoric shell midden in the Brazilian Amazon. Science 254 (5038), 1621–1624.

Rye, O.S., 1981. Pottery Technology: Principles and Reconstruction. Taraxacum, Inc., Washington, DC.

Saunders, R., Hays, C.T., 2006. Introduction: Themes in Early Pottery Research. In: Saunders, R., Hays, C.T. (Eds.), Early Pottery: Technology, Function, Style, and Interaction in the Lower Southeast. University of Alabama Press, Tuscaloosa, pp. 1–22.

Sharer, R., Ashmore, W., 2003. Analysis of artifacts. In: Archaeology: Discovering our Past, third ed. McGraw-Hill Higher Education, New York.

Shepard, A.O., 1965. Ceramics for the Archaeologist. Carnegie Institution of Washington Publication 609, Washington, DC.

Sinopoli, C.M., 1991. Approaches to Archaeological Ceramics. Plenum Press, New York.

Smith, N.G., Karasik, A., Narayanan, T., Olson, E.S., Smilansky, U., Levy, T.E., 2014. The pottery informatics query database: a new method for mathematic and quantitative analyses of large Regional ceramic datasets. Journal of Archaeological Method and Theory 21 (1), 212–250.

Stark, M.T., 2003. Current issues in ceramic ethnoarchaeology. Journal of Archaeological Research 11 (3), 193–242.

Stevenson, A., 2014. Artefacts of excavation. Journal of the History of Collections 26 (1), 89–102.

Vandiver, P., Soffer, O., Klima, B., 1991. The origins of ceramics: Figurine Manufacture at Dolni Vestonice, circa 26,000 B.P. Studio Potter 20 (1), 4.

Washburn, D.K., 1977. A Symmetry Analysis of Upper Gila Area Ceramic Design. Peabody Museum Papers of the Peabody Museum of Anthropology and Archaeology, No. 68, Cambridge, UK.

Zhushchikhovskaya, I., 2012. Early Jomon pottery (based on material from Ookubo site) (English transl of Russian title). Rossijskaâ Arheologiâ 1, 5–15.

Cerebellum: Cognitive Functions

Marco Molinari, Santa Lucia Foundation IRCCS, Rome, Italy
Maria Leggio, Sapienza University of Rome, Rome, Italy; and Santa Lucia Foundation IRCCS, Rome, Italy

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by B. Suchan, I. Daum, volume 3, pp. 1631–1634, © 2001, Elsevier Ltd.

Abstract

In the nineteenth century, researchers reached a consensus on the basis of animal ablation experiments that damage to the cerebellum leads to motor disorders but does not affect its sensory or cognitive functions. This model has been so influential that for over 100 years, the cerebellum has been considered to be strictly involved in motor control, whereas cerebellar function in cognition has been hotly debated. The current consensus is centered around the belief that the cerebellum intervenes in many domains, from motor to emotion and from sensory processing to learning. This article discusses the cognitive functions of the cerebellum. Examples of cerebellar involvement in various higher-cognition functions will be provided, and theories on the specificity of the cerebellar contribution to cognition will be advanced.

History

In the early twentieth century, Stewart and Holmes (Holmes, 1917) reported that cerebellar lesions due to tumors or gunshots effect motor deficits in humans, such as decreased muscle tone, impaired coordination of movements, and deficits in the regulation of gait and posture. Holmes's studies were influential and introduced the cerebellum as a purely motor structure. Nevertheless, even in the 1800s, intellectual, psychiatric, and social–emotional dysfunctions were observed in patients with cerebellar degeneration and agenesis (see reviews by Macchi and Bentivoglio, 1987; Schmahmann, 1991).

The concept of cerebellar function outside of the motor domain, focusing primarily on autonomic and limbic functions, was renewed by Snider in the 1950s (Snider, 1950). But only in the 1980s has research on cerebellar function in cognition gained momentum (Leiner et al., 1986; Schmahmann, 1991).

General Principles of the Anatomophysiological Cerebro-Cerebellar Interactions

The organization of the cerebello-thalamo-cortico-ponto-cerebellar loop has been examined extensively in monkeys (Schmahmann and Pandya, 1997; Middleton and Strick, 2000). According to these findings, cerebello-cortical circuits are organized in close parallel loops. Each cortical area that sends elements to the cerebellum via the pontine nuclei is also the recipient of a returning loop by the cerebello-thalamic bundle. Cerebellar efferents originate from deep cerebellar nuclei and are organized into segregated channels that reach specific cortical zones by a disynaptic pathway through the thalamus. Similarly, the cortico-ponto-cerebellar stream presents segregated channels (Ramnani, 2006).

The introduction of electrical and magnetic stimulation of brain structures through the scalp has made the neurophysiological study of neuronal circuitries in humans possible on a large scale; this technique has also been used to address the nature of cerebro-cerebellar interactions neurophysiologically. Thus, by modifying the function of cerebellar circuits through noninvasive brain stimulation techniques, such as transcranial magnetic stimulation and transcranial direct current stimulation, it is possible to analyze the effects of cerebellar modulation over the cerebral cortex in humans directly (Bijsterbosch et al., 2012).

It is generally accepted that cerebellar input has a facilitating effect on the contralateral cerebral cortex – chronic cerebellar damage reduces the excitability of the contralateral motor cortex (Di Lazzaro et al., 1995). Conversely, disparate effects on the cerebral cortex after cerebellar stimulation have been reported (Ferrucci et al., 2012; Lu et al., 2012). By varying the parameters of cerebellar stimulation, it is possible to modify the ultimate effect on the cerebral cortex (Hamada et al., 2012).

Early neurophysiological studies on cerebro-cerebellar interactions focused on sensory motor processing (Bower and Parsons, 2003), and recent reports have determined the function of the cerebellum in allowing sensory information to modulate motor cortex excitability (Taib and Manto, 2009). In the sensorimotor domain, clinical and experimental data converge, indicating that the cerebellum links the crosstalk between the somatosensory and motor cortices. How does this mechanism work? Are somatosensory influences on the motor cortex mediated by cerebello-M1 circuits, or does the cerebellar output affect the somatosensory cortex directly?

Hints to the answers can be found in observations in humans with focal lesions of the cerebellum. Isolated cerebellar lesions affect S1 intracortical processing of simple somatosensory stimuli, as shown in subjects with unilateral focal cerebellar lesions, in which early somatosensory evoked potentials (SEPs) of the primari somatosensory (SI) cortex that is contralateral to the lesioned cerebellum are altered (Restuccia et al., 2001). In particular, the amplitude of the frontal N24 and parietal P24 SEP components and the strength of the corresponding dipolar sources are significantly lower in the hemisphere that lacks the cerebellar input compared with the contralateral side.

The finding that cerebellar damage influences the early components of SI processing supports the existence of a direct influence of the cerebellum over S1- rather than an undirected one through M1. This datum is consistent with findings in monkeys of different segregated cerebro-cerebellar loops targeting, beside the motor areas, also the parietal lobe

(Clower et al., 2001). The function of such an early influence can be understood, considering the predictive properties of cerebellar processing (Bastian, 2006). In this framework, the early influence might be to gauge cerebellar capacity to alert the cortex before the arrival of a predicted incoming volley (Tesche and Karhu, 2000; Ivry, 2000).

How can the cerebellum predict the next event in the cerebral cortex? The cerebellar capacity to process sequences of events and, in particular, detect predictable ones (Molinari et al., 2009) by sequencing has been observed in the motor (Manto et al., 2012), somatosensory (Restuccia et al., 2007), temporal (Moberget et al., 2008), and cognitive domains (Leggio et al., 2008). These studies reported sequencing and event prediction impairments that were associated with cerebellar damage. This significance of sequencing in defining cerebellar function in various domains is a novel support for the hypothesis of sequencing as the basic cerebellar mode of operation (Braitenberg et al., 1997; Molinari et al., 2008).

Cerebellar Cognitive Functions

In 1998, Schmahmann and Sherman (Schmahmann and Sherman, 1998) described 'cerebellar cognitive affective syndrome,' reporting impairments in executive functions, disturbances in spatial cognition, language deficits, and personality changes in subjects with cerebellar pathologies. This early report was confirmed and expanded by several groups. The current list of cognitive functions that are impaired in association with cerebellar malfunctioning includes executive functions (Kim et al., 1994; Gottwald et al., 2004; Bellebaum and Daum, 2007), visuospatial functions (Neau et al., 2000; Molinari et al., 2004), working memory (Silveri et al., 1998; Chiricozzi et al., 2008; Marvel and Desmond, 2010), verbal memory (Grasby et al., 1993), linguistic processing (Petersen et al., 1988), verbal fluency (Leggio et al., 2000; Peterburs et al., 2010), procedural learning (Molinari et al., 1997; Shin and Ivry, 2003), attention (Allen et al., 1997; Gottwald et al., 2004), sequencing (Leggio et al., 2008; Moberget et al., 2008), and emotion (Schmahmann and Sherman, 1998; Turner et al., 2007).

Consistent with lesion studies, there are functional magnetic resonance imaging (fMRI) data on healthy subjects on cerebellar activation in several cognitive functions, from which a consensus has been forming on the topographical distribution of various cognitive domains in the posterior lobe of the cerebellum, independent of sensorimotor somatotopy (Stoodley et al., 2012). More recent brain stimulation studies that have targeted cerebellar structures are providing additional evidence of cerebellar cognitive processing (Ferrucci et al., 2008, 2012; Pope and Miall, 2012). An exhaustive list of reports that implicate cerebellar involvement in various domains is beyond our objective; thus, we will focus on verbal working memory (vWM) and mental functions with regard to increasing our understanding of cerebellar function in cognition.

Cerebellum and vWM

The contribution of the cerebellum to vWM has been examined by several groups, which have generated notable results that have enhanced our comprehension of the cerebellar contribution to cognition. After early fMRI findings of cerebellar activation during vWM tasks (Desmond and Fiez, 1998; Thomas et al., 1999), Silveri et al., first demonstrated vWM dysfunction in a cerebellar patient (Silveri et al., 1998).

Since then, many groups have noted cerebellum-related vWM impairments that generally refer to the highly influential model that was proposed by Baddeley (Baddeley, 2003) to support their findings. According to this model, the phonological loop is the system that processes language-related information. This loop comprises two distinct components: The articulatory rehearsal, located in Broca's, the premotor, and supplementary motor cortical areas, and the phonological store, located in the parietal lobe (Keren et al., 2012).

The cerebellum has only recently been added to the list of neural structures that are involved in vWM, but its contribution remains unknown. Due to the clinical characteristics of the deficit, the first report on vWM impairment from cerebellar damage concluded that the cerebellar contribution lay in the articulatory control process rather than the phonological storage component of vWM (Silveri et al., 1998). Later, in various subjects with cerebellar damage, Justus and colleagues reported vWM impairments to the phonological output buffer or phonological analyses (Ravizza et al., 2006). A recent single-case study reported cerebellum-related damage to the phonological short-term store (Chiricozzi et al., 2008). Thus, data from lesion studies indicate that the cerebellum acts on many levels of vWM processing.

The complexity of the cerebellar contribution to vWM has been confirmed by neuroimaging studies, which have noted that various cerebellar lobules are active when different vWM components are called into action (Cooper et al., 2012). In particular, Desmond and colleagues differentiated at least two segregated cerebro-cerebellar loops that were active during vWM tasks in a series of fMRI studies. According to their findings, the superior cerebellum is a part of a cerebro-cerebellar motor circuit that includes Broca's area and the supplementary motor area, which creates motor-related articulatory trajectory paths (orthographic to phonological translation). Conversely, the inferior cerebellum is part of a circuit that involves BA40 and supports the maintenance of information in the phonological store (Marvel and Desmond, 2010).

The presence of various function-specific cerebro-cerebellar loops that support vWM has also been confirmed by morphometric studies in subjects with spinocerebellar ataxia type 6 (SCA6) cerebellar degeneration. In a comparison of vWM performance and cerebellar cortex volume, a critical relationship between vWM and gray matter density was demonstrated in the superior (bilateral lobules VI and crus I of lobule VII) and inferior (bilateral lobules VIIIa and VIIIb, and right lobule IX) sections of the cerebellum. Notably, in these areas, distinct cerebellar regions subserve various vWM functions. Specifically, the areas that are devoted to the phonological loop can be differentiated from those that support the central executive components of Baddeley's model of vWM (Cooper et al., 2012).

Considering that Baddeley's cognitive modules lie in disparate areas of the cerebral cortex (Smith and Jonides, 1998) and based on the modular organization of cerebello-cortical

connections, the clinical manifestations of cerebellar damage on vWM might depend on the region of the cerebellum and the associated cerebro-cerebellar connection that is damaged (Marvel and Desmond, 2010).

Thus, on examination of a specific cognitive function, such as vWM, the function of the cerebellum appears to be difficult to identify. As in the early approaches of studying cerebellar function in cognition, nearly all abilities require cerebellar contribution. Nevertheless, due to the unitary nature of cerebellar anatomical modules, the basic operational mode in various domains should be the same.

The Cerebellum and Mental Function

The study of mental functions with regard to the cerebellum is an area of significant interest. As discussed, since the early definition of Cerebellar Cognitive Affective Syndrome (Schmahmann and Sherman, 1998), mood disorders, personality changes, and inappropriate behaviors have been linked to cerebellar damage. Conversely, the cerebellum has been considered to mediate the pathogenesis of several neuropsychiatric disorders – schizophrenia, depression, autism, and dyslexia. Further, among degenerative cerebellar disease, the rate of psychiatric diseases is high (Leroi et al., 2002). How cerebellar processing intervenes in affecting mental health is unknown, but the answer might be obtained by analyzing cerebellar involvement in the pathogenesis of schizophrenia and autism.

Schizophrenia is defined as a mental disorder that is characterized by a dissociation between internal representations and external reality. Considering the classical cerebellar function as a comparator between functional states (Ito, 2005), it is conceivable that internal vs external reality representations involve the cerebellum. Moreover, it is notable that in reporting cognitive defects in subjects with cerebellar damage, terms that are characteristic of psychosis, such as cognitive dysmetria and dysmetria of thought, have been used (Schmahmann, 2010). Several years ago, a cerebellar theory of schizophrenia was even advanced (Picard et al., 2008).

Further, as cerebellar patients, schizophrenic patients present with deficits in perceptual and motor processing, attentional skills, vigilance, vWM, and verbal fluency (Synofzik et al., 2010; Tedesco et al., 2011). Notably, impairments in predictive abilities have been reported, at least for the sensorimotor domain, in schizophrenia (Synofzik et al., 2010) and after cerebellar damage (Bastian, 2006).

Neurophysiologically, it is possible to evaluate the capacity to recognize unexpected changes in incoming volleys of sensory data using the MisMatch Negativity (MMN) paradigm. MMN is a brain event-related potential that reflects the capacity of the human brain to detect deviance in the environment preattentively. MMN is considered as an effective measure of the processing of preattentive information by change detection. MMN is thus recorded as a negative event-related potential (ERP) component that is elicited by deviant stimulus – i.e., changes in the frequency, duration, intensity, or location of a standard stimulus. The MMN is believed to be the outcome of a comparison process that is commonly elicited when an incoming stimulus differs from the memory of repetitive ones (standards) that occurred in the recent past. This comparison is altered for acoustic stimuli in schizophrenia (Zhou et al., 2013) and in cerebellar patients for acoustic (Moberget et al., 2008) and somatosensory stimuli (Restuccia et al., 2007).

For both pathologies, the core alteration in associated behavioral disturbances has been linked to the inability to detect changes in sensory patterns. The inability to predict sensory patterns correlates with the difficulties of schizophrenia in recognizing an action on the basis of subject intent (D'Angelo and Casali, 2013). Somehow, in parallel, cerebellar patients encounter difficulties in recognizing actions and behaviors (Leggio et al., 2008).

Autism is a developmental disorder in which cerebellar damage or malfunction is often considered among possible pathogenic mechanisms (Fatemi et al., 2012). The strongest link between the cerebellum and autism derives from the coexistence of cerebellar-like motor symptoms that affect basic motor control, postural stability, and skilled gestures and abnormal patterns of motor learning in most persons with autism (Nebel et al., 2012), as well as the presence of abnormalities in cerebellar structures (Abell et al., 1999).

Similarities between autism and cerebellar symptoms are present with regard to cognitive deficits. Individuals with autism have impairments in executive function, memory, and language – all domains that are altered in subjects with cerebellar damage (Tedesco et al., 2011). Notably, the core neurobehavioral features of autism – impaired social interaction, isolated areas of interest, and an insistence on sameness – can be related to impairments in cerebellar processing. Based on neurophysiological findings, it has been suggested that the cerebellum is central to the analysis of sequential events in the behavioral domain (Molinari et al., 2008).

The capacity to adapt to a given context behaviorally depends on the ability to recognize the perceived pattern. Only if the perceived context is recognized as pertinent to previous knowledge it is possible to develop the correct behavior. For example, if one is presented to the Japanese emperor for the first time, it is highly likely that he will not know the proper decorum. Yet, an experienced courtier of the Japanese Imperial palace will immediately adapt his behavior, even to small changes in the attitudes of others in the room.

Cerebellar damage will reduce one's capacity to detect a pattern, thus impeding the recognition of a previously encountered context and preventing the appropriate behavior from being assumed. Thus, it is understandable how cerebellum-related difficulties in recognizing a behavioral pattern will impair social interactions and the capacity to adapt to new areas, and favoring the tendency to repeat a fixed behavior.

The Basic Operational Mode of the Cerebellum in Cognition

Despite the increase in data on the cerebellum in cognitive and behavioral functions, the precise definition of cerebellar processing has not been determined. It is generally acknowledged that due the regularity and substantial uniformity of cerebellar modules, the same basic functions should be exerted by the cerebellum in the motor, cognitive, and affection-related domains (Ito, 2008). In particular, based on studies in the

motor domain, the hypothesis that cerebellar processing allows event prediction of subsequent states (Bo et al., 2008) has been expanded to the cognitive domain.

Cerebellar sequencing has been suggested to be at the core of cerebellar capacities in predicting future states in the sensorimotor, cognitive and behavioral domains (Braitenberg et al., 1997; Molinari et al., 2009; D'Angelo and Casali, 2013). In a recent study of more than 200 cerebellar patients, over a wide range of cognitive domains, sequencing was the most affected domain, independent of the nature (degenerative or focal) or anatomical location of the lesion (Tedesco et al., 2011).

Evidence of the significance of cerebellar processing in behavioral predictions and in pathogenesis neuropsychiatric disorders, such as autism and schizophrenia, has been reported in the studies above. Neuroimaging, neurophysiological, and clinical data are consistent in suggesting that the cerebellum, in parallel to proposals in the somatomotor domain (Bo et al., 2008), is a predictive device in cognitive processes, including vWM (Ben Yehudah et al., 2007).

Notably, disparate lines of research suggest that the detection and coordination of spatiotemporal patterns – such as sequencing – constitute the common basic principle of cerebellar neuronal operation in all functional domains.

The model of the cerebellum as the key element that allows smooth, organized spatiotemporal muscular patterns dates back to Holmes (Holmes, 1917). Based solely on anatomical evidence of the cerebellar circuits, sequencing has been suggested to be the chief result of cerebellar neuronal processing. Specifically, the cerebellar cortex allows to detect sequences of incoming activity and generate precise output sequences (Braitenberg et al., 1997). As regards cognition, the same basic processing of sequencing and the complexity of the cortico-cerebellar loops allows the cerebellum to support cortical function by detecting and coordinating spatiotemporal cortical activity patterns at the level of spike activity (Fatemi et al., 2012) or functional modules (Molinari et al., 2008).

See also: Associative Learning in the Cerebellum; Autism Spectrum Disorders; Human Cognition, Evolution of; Schizophrenia: Metabolic Consequences; schizophrenia; Working Memory, Neural Basis of; Working Memory, Psychology of.

Bibliography

Abell, F., Krams, M., Ashburner, J., et al., 1999. The neuroanatomy of autism: a voxel-based whole brain analysis of structural scans. Neuroreport 10, 1647–1651.

Allen, G., Buxton, R.B., Wong, E.C., Courchesne, E., 1997. Attentional activation of the cerebellum independent of motor involvement. Science 275, 1940–1943.

Baddeley, A., 2003. Working memory: looking back and looking forward. Nature Reviews Neuroscience 4, 829–839.

Bastian, A.J., 2006. Learning to predict the future: the cerebellum adapts feedforward movement control. Current Opinion in Neurobiology 16, 645–649.

Bellebaum, C., Daum, I., 2007. Cerebellar involvement in executive control. The Cerebellum 6, 184–192.

Ben Yehudah, G., Guediche, S., Fiez, J.A., 2007. Cerebellar contributions to verbal working memory: beyond cognitive theory. The Cerebellum 6, 193–201.

Bijsterbosch, J.D., Barker, A.T., Lee, K.H., Woodruff, P.W.R., 2012. Where does transcranial magnetic stimulation (TMS) stimulate? Modelling of induced field maps for some common cortical and cerebellar targets. Medical & Biological Engineering & Computing 50, 671–681.

Bo, J., Block, H.J., Clark, J.E., Bastian, A.J., 2008. A cerebellar deficit in sensorimotor prediction explains movement timing variability. Journal of Neurophysiology 100, 2825–2832.

Bower, J.M., Parsons, L.M., 2003. Rethinking the "lesser brain". Scientific American 289, 50–57.

Braitenberg, V., Heck, D., Sultan, F., 1997. The detection and generation of sequences as a key to cerebellar function: experiments and theory. Behavioral and Brain Sciences 20, 229–277.

Chiricozzi, F.R., Clausi, S., Molinari, M., Leggio, M.G., 2008. Phonological short-term store impairment after cerebellar lesion: a single case study. Neuropsychologia 46, 1940–1953.

Clower, D.M., West, R.A., Lynch, J.C., Strick, P.L., 2001. The inferior parietal lobule is the target of output from the superior colliculus, hippocampus, and cerebellum. Journal of Neuroscience 21, 6283–6291.

Cooper, F.E., Grube, M., Von Kriegstein, K., et al., 2012. Distinct critical cerebellar subregions for components of verbal working memory. Neuropsychologia 50, 189–197.

D'Angelo, E., Casali, S., 2013. Seeking a unified framework for cerebellar function and dysfunction: from circuit operations to cognition. Frontiers in Neural Circuits 6.

Desmond, J.E., Fiez, J.A., 1998. Neuroimaging studies of the cerebellum: language, learning and memory. Trends in Cognitive Sciences 2, 355–362.

Di Lazzaro, V., Restuccia, D., Nardone, R., et al., 1995. Motor cortex changes in a patient with hemicerebellectomy. Electroencephalography and Clinical Neurophysiology 97, 259–263.

Fatemi, S.H., Aldinger, K.A., Ashwood, P., et al., 2012. Consensus paper: pathological role of the cerebellum in autism. The Cerebellum 11, 777–807.

Ferrucci, R., Giannicola, G., Rosa, M., et al., 2012. Cerebellum and processing of negative facial emotions: cerebellar transcranial DC stimulation specifically enhances the emotional recognition of facial anger and sadness. Cognition & Emotion 26, 786–799.

Ferrucci, R., Marceglia, S., Vergari, M., et al., 2008. Cerebellar transcranial direct current stimulation impairs the practice-dependent proficiency increase in working memory. Journal of Cognitive Neuroscience 20, 1687–1697.

Gottwald, B., Wilde, B., Mihajlovic, Z., Mehdorn, H.M., 2004. Evidence for distinct cognitive deficits after focal cerebellar lesions. Journal of Neurology, Neurosurgery, and Psychiatry 75, 1524–1531.

Grasby, P.M., Frith, C.D., Friston, K.J., Bench, C., Frackowiak, R.S.J., Dolan, R.J., 1993. Functional mapping of brain-areas implicated in auditory – verbal memory function. Brain 116, 1–20.

Hamada, M., Strigaro, G., Murase, N., et al., 2012. Cerebellar modulation of human associative plasticity. Journal of Physiology-London 590, 2365–2374.

Holmes, G., 1917. The symptoms of acute cerebellar injuries due to gunshot injuries. Brain 40, 461–535.

Ito, M., 2005. Bases and implications of learning in the cerebellum – adaptive control and internal model mechanism. Progress in Brain Research 148, 95–109.

Ito, M., 2008. Control of mental activities by internal models in the cerebellum. Nature Reviews Neuroscience 9, 304–313.

Ivry, R., 2000. Exploring the role of the cerebellum in sensory anticipation and timing: commentary on Tesche and Karhu [comment]. Human Brain Mapping 9, 115–118.

Keren, H., Chen, S.H.A., Ho, M.H.R., Desmond, J.E., 2012. A meta-analysis of cerebellar contributions to higher cognition from PET and fMRI studies. Human Brain Mapping. Epub ahead of print: n/a.

Kim, S.G., Ugurbil, K., Strick, P.L., 1994. Activation of a cerebellar output nucleus during cognitive processing. Science 265, 949–951.

Leggio, M.G., Silveri, M.C., Petrosini, L., Molinari, M., 2000. Phonological grouping is specifically affected in cerebellar patients: a verbal fluency study. Journal of Neurology, Neurosurgery, and Psychiatry 69, 102–106.

Leggio, M.G., Tedesco, A.M., Chiricozzi, F.R., Clausi, S., Orsini, A., Molinari, M., 2008. Cognitive sequencing impairment in patients with focal or atrophic cerebellar damage. Brain 131, 1332–1343.

Leiner, H.C., Leiner, A.L., Dow, R.S., 1986. Does the cerebellum contribute to mental skills? Behavioral Neuroscience 100, 443–454.

Leroi, I., O'Hearn, E., Marsh, L., et al., 2002. Psychopathology in patients with degenerative cerebellar diseases: a comparison to Huntington's disease. American Journal of Psychiatry 159, 1306–1314.

Lu, M.K., Tsai, C.H., Ziemann, U., 2012. Cerebellum to motor cortex paired associative stimulation induces bidirectional STDP-like plasticity in human motor cortex. Frontiers in Human Neuroscience 6.

Macchi, G., Bentivoglio, M., 1987. Agenesis or hypoplasia of cerebellar structures. In: Vinken, P.J., Bruyn, G.W. (Eds.), Handbook of Clinical Neurology, pp. 367–393.
Manto, M., Bower, J.M., Conforto, A.B., et al., 2012. Consensus paper: roles of the cerebellum in motor control – the diversity of ideas on cerebellar involvement in movement. The Cerebellum 11, 457–487.
Marvel, C., Desmond, J., 2010. Functional topography of the cerebellum in verbal working memory. Neuropsychology Review 20, 271.
Middleton, F.A., Strick, P.L., 2000. Basal ganglia and cerebellar loops: motor and cognitive circuits. Brain Research. Brain Research Reviews 31, 236–250.
Moberget, T., Karns, C.M., Deouell, L.Y., Lindgren, M., Knight, R.T., Ivry, R.B., 2008. Detecting violations of sensory expectancies following cerebellar degeneration: a mismatch negativity study. Neuropsychologia 46, 2569–2579.
Molinari, M., Leggio, M.G., Solida, A., et al., 1997. Cerebellum and procedural learning: evidence from focal cerebellar lesions. Brain 120, 1753–1762.
Molinari, M., Petrosini, L., Misciagna, S., Leggio, M.G., 2004. Visuospatial abilities in cerebellar disorders. Journal of Neurology, Neurosurgery, and Psychiatry 75, 235–240.
Molinari, M., Restuccia, D., Leggio, M.G., 2009. State estimation, response prediction, and cerebellar sensory processing for behavioral control. The Cerebellum 8, 399–402.
Molinari, M., Chiricozzi, F., Clausi, S., Tedesco, A., De Lisa, M., Leggio, M., 2008. Cerebellum and detection of sequences, from perception to cognition. The Cerebellum 7, 611–615.
Neau, J.P., Arroyo-Anllo, E., Bonnaud, V., Ingrand, P., Gil, R., 2000. Neuropsychological disturbances in cerebellar infarcts. Acta Neurologica Scandinavica 102, 363–370.
Nebel, M.B., Joel, S.E., Muschelli, J., et al., 2012. Disruption of functional organization within the primary motor cortex in children with autism. Human Brain Mapping n/a.
Peterburs, J., Bellebaum, C., Koch, B., Schwarz, M., Daum, I., 2010. Working memory and verbal fluency deficits following cerebellar lesions: relation to interindividual differences in patient variables. The Cerebellum 9, 375–383.
Petersen, S.E., Fox, P.T., Posner, M.I., Mintun, M., Raichle, M.E., 1988. Positron emission tomographic studies of the cortical anatomy of single-word processing. Nature 331, 585–589.
Picard, H., Amado, I., Mouchet-Mages, S., Olié, J.-P., Krebs, M.O., 2008. The role of the cerebellum in schizophrenia: an update of clinical, cognitive, and functional evidences. Schizophrenia Bulletin 34, 155–172.
Pope, P.A., Miall, R.C., 2012. Task-specific facilitation of cognition by cathodal transcranial direct current stimulation of the cerebellum. Brain Stimulation 5, 84–94.
Ramnani, N., 2006. The primate cortico-cerebellar system: anatomy and function. Nature Reviews Neuroscience 7, 511–522.
Ravizza, S.M., McCormick, C.A., Schlerf, J.E., Justus, T., Ivry, R.B., Fiez, J.A., 2006. Cerebellar damage produces selective deficits in verbal working memory. Brain 129, 306–320.
Restuccia, D., Della, M.G., Valeriani, M., Leggio, M.G., Molinari, M., 2007. Cerebellar damage impairs detection of somatosensory input changes. A somatosensory mismatch-negativity study. Brain 130, 276–287.
Restuccia, D., Valeriani, M., Barba, C., et al., 2001. Functional changes of the primary somatosensory cortex in patients with unilateral cerebellar lesions. Brain 124, 757–768.
Schmahmann, J.D., 1991. An emerging concept. The cerebellar contribution to higher function. Archives of Neurology 48, 1178–1187.
Schmahmann, J.D., 2010. The role of the cerebellum in cognition and emotion: personal reflections since 1982 on the dysmetria of thought hypothesis, and its historical evolution from theory to therapy. Neuropsychology Review 20, 236–260.
Schmahmann, J.D., Pandya, D.N., 1997. The cerebrocerebellar system. International Review of Neurobiology 41, 31–60.
Schmahmann, J.D., Sherman, J.C., 1998. The cerebellar cognitive affective syndrome. Brain 121 (Pt 4), 561–579.
Shin, J.C., Ivry, R.B., 2003. Spatial and temporal sequence learning in patients with Parkinson's disease or cerebellar lesions. Journal of Cognitive Neuroscience 15, 1232–1243.
Silveri, M.C., Di Betta, A.M., Filippini, V., Leggio, M.G., Molinari, M., 1998. Verbal short-term store-rehearsal system and the cerebellum. Evidence from a patient with a right cerebellar lesion. Brain 121 (Pt 11), 2175–2187.
Smith, E.E., Jonides, J., 1998. Neuroimaging analyses of human working memory. Proceedings of the National Academy of Sciences of the United States of America 95, 12061–12068.
Snider, R.S., 1950. Recent contributions to the anatomy and physiology of the cerebellum. Archives of Neurology and Psychiatry 64, 196–219.
Stoodley, C.J., Valera, E.M., Schmahmann, J.D., 2012. Functional topography of the cerebellum for motor and cognitive tasks: an fMRI study. Neuroimage 59, 1560–1570.
Synofzik, M., Thier, P., Leube, D.T., Schlotterbeck, P., Lindner, A., 2010. Misattributions of agency in schizophrenia are based on imprecise predictions about the sensory consequences of one's actions. Brain 133, 262–271.
Taib, N.O.B., Manto, M., 2009. Trains of transcranial direct current stimulation antagonize motor cortex hypoexcitability induced by acute hemicerebellectomy laboratory investigation. Journal of Neurosurgery 111, 796–806.
Tedesco, A.M., Chiricozzi, F.R., Clausi, S., Lupo, M., Molinari, M., Leggio, M.G., 2011. The cerebellar cognitive profile. Brain 134, 3672–3678.
Tesche, C.D., Karhu, J.J., 2000. Anticipatory cerebellar responses during somatosensory omission in man [see comments]. Human Brain Mapping 9, 119–142.
Thomas, K.M., King, S.W., Franzen, P.L., et al., 1999. A developmental functional MRI study of spatial working memory. Neuroimage 10, 327–338.
Turner, B.M., Paradiso, S., Marvel, C.L., et al., 2007. The cerebellum and emotional experience. Neuropsychologia 45, 1331–1341.
Zhou, Z., Zhu, H., Chen, L., 2013. Effect of Aripiprazole on mismatch negativity (MMN) in schizophrenia. PLoS One 8, e52186.

Cerebral Cortex

Phillip Grant, Justus-Liebig-University, Giessen, Hesse, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Schuz, volume 3, pp. 1634–1640, © 2001, Elsevier Ltd.

Abstract

Forming the outer layer of the forebrain, the cerebral cortex is the largest body of gray matter in the brain. It is made up of three distinct parts, namely the isocortex, which is phylogenetically the youngest part of the brain and is therefore often referred to as the neocortex, and the ancestrally older archi- and paleocortices, of which the hippocampus is the most prominent example. The cerebral cortices are highly versatile in their function and exhibit great synaptic plasticity and change and therefore have the capacity, inter alia, for storage of vast amounts of information.

The cerebral cortex is the outer mantle (pallium) of the forebrain (telencephalon or cerebrum). It is not only the largest body of gray matter in the primate brain but also the most versatile regarding its function. Although the amount of research into cortical function has drastically increased in humans with the invention of newer neuroimaging techniques (like functional magnetic resonance imaging, fMRI), it can still be considered the greatest enigma of the nervous system. On the one hand, the immense network of fibers connecting different cortical areas, both directly and via the dorsal thalamus (see below) combined with the participation of the cortex in almost all higher cognitive functions suggest to some that it may be considered the higher-order driving force behind most brain functions. On the other hand, lesion studies, case-reports from, e.g., stroke patients or findings regarding the tremendous capacity for neural plasticity and change lead others to view upon the cerebral cortex (especially the neocortical areas; q.v., next paragraph) as the equivalent of the hard drive in a computer, rather than its central processing unit.

In general, the participation of specific cortical regions in certain tasks is interpreted differently by members of the neuroscientific world: While locationists propose that abstract mental or brain functions are attributable to discrete cortical areas, connectionists assume that the interplay of various brain regions actually represents the neural basis of brain function.

Although, to use a common idiom, the jury is still out on this question, evidence exists supporting all of the aforementioned hypotheses. It is therefore helpful to draw a distinction between those areas of the cerebral cortex that are phylogenetically older – the so-called (greater) limbic system – and those relatively young in primate evolution and, in part, unique and emergent in *Homo sapiens*.

The cerebral cortex is commonly divided according to its phylogenetic age into the younger neocortex or neopallium and the ancestrally older archi- and paleocortices. Another distinction – based on cytoarchitectonic lamination – differentiates a six-layered isocortex from a three-layered allocortex, whereby the isocortex is mostly equivalent to the neocortex and the allocortex to the archi- and paleocortices (Figure 1).

Apically, the outer layers of the cerebral cortex are surrounded by a fine glial membrane, the superficial limiting glial membrane, which in turn borders onto the microscopic subpial space. Inwards, the cerebral cortex transits into the telencephalic white matter, which encompasses the cerebral nuclei (i.e., the caudate nucleus, putamen, claustrum, amygdala, and the innominate substance of the basal forebrain including the nucleus basalis of Meynert).

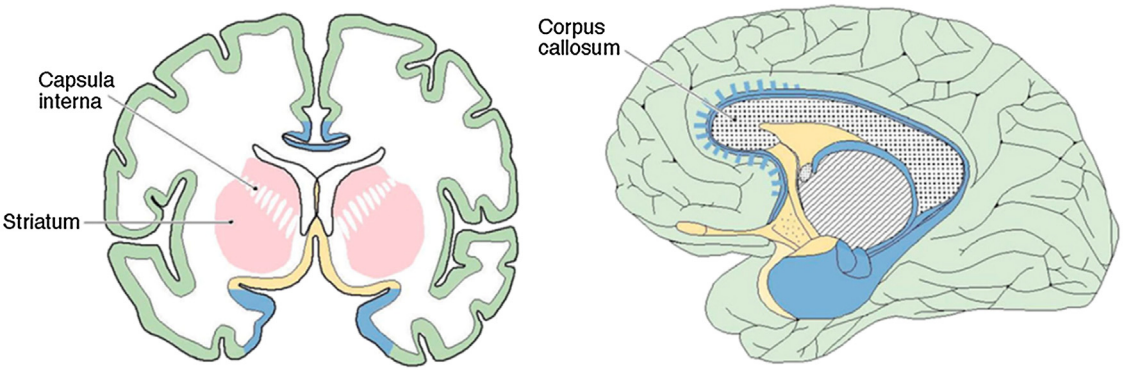

Figure 1 Overview of the cerebral cortex and the cerebral nuclei (pink).
Green = neocortex; blue = archicortex; yellow = paleocortex. Modified from Benninghoff and Drenckhahn. Anatomie, vol. 2, sixteenth ed. ISBN-10: 3437423517 ISBN-13: 978-3437423512; image 12_8-01.

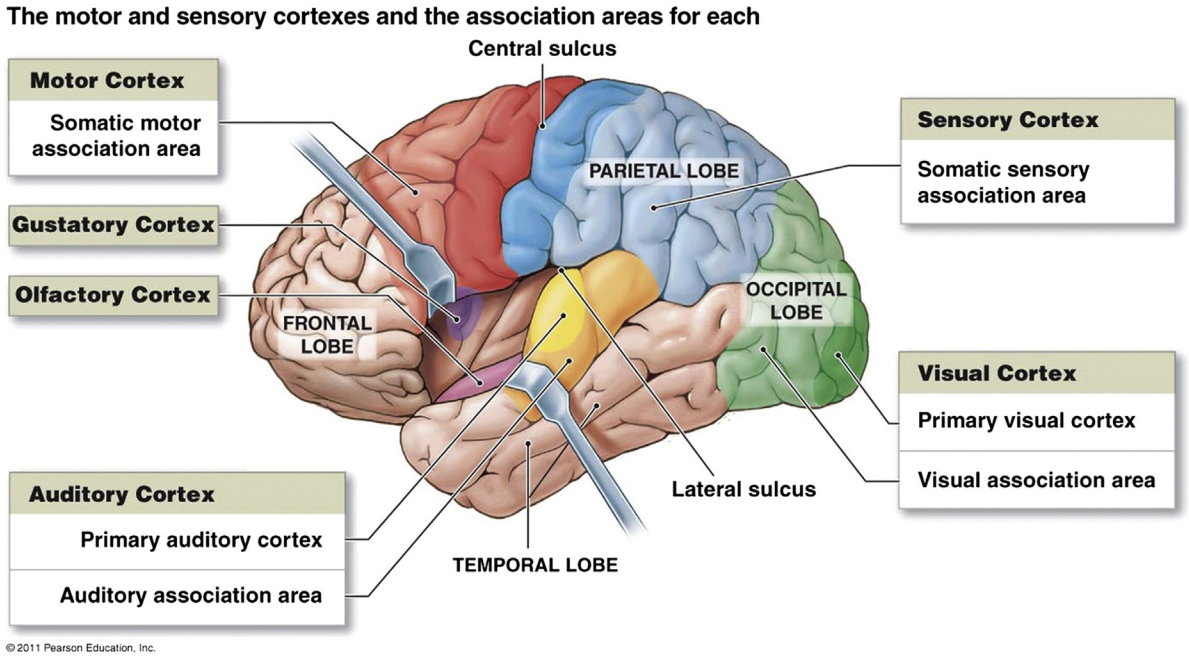

Figure 2 Cerebral cortical lobes. Paulsen and Waschke. Sobotta Atlas der Anatomie des Menschen, 23rd ed. ISBN-10: 3437440705 ISBN-13: 978-3437440700; image 12.34.

The cerebral cortex is further divided into lobes, whereby the frontal, parietal, occipital and temporal lobes as well as the insula are mostly isocortical, while the – somewhat inconsistently defined – limbic lobe is made up primarily of allocortical areas. Due to the large growth of the brain in higher-mammalian phylogeny, the cerebral cortex is gyrated in many species, meaning that its surface is heavily folded into gyri, which are separated by sulci. The largest of the sulci are the lateral sulcus, separating the temporal from the frontal and parietal lobes, the central sulcus between the frontal and parietal lobes and the parietooccipital sulcus. Both cerebral hemispheres are separated by the large longitudinal fissure and connected through several commissures, the largest thereof being the corpus callosum (Figure 2).

Ontogeny and Phylogeny of the Human Cerebral Cortex

The entire forebrain develops from the most rostral of the primary brain vesicles, the prosencephalic vesicle. It forms from the neural tube around the fourth week of gestation. Around the fifth week of gestation it has further divided into the diencephalic and telecencephalic vesicles, whereby the latter gives rise to two distinct hemispheral anlages the subventricular zone, combined by a median mass known as the telencephalon impar, or the median telencephalon (Figure 3).

Initially, the cerebral hemispheres mainly increase in size, while at the end of the fifth month usually gyration starts. It is important to notice that gyration (and neural ontogeny as a whole) does not end with but continues after birth and, to a certain extent, throughout any individual's life (Figure 4).

Neuronal progenitor cells of the cerebral cortex develop, divide, and differentiate near the center of the hemispheral anlages (the subventricular zone) and then migrate in a centrifugal fashion, thereby forming the distinct layers of the cerebral cortex.

Although the human brain is by far the largest in all primate species, size and gyration are not the ultimate determinants of higher-order functions. Comparing the brain of an average human being with that of, e.g., a cetacean, one finds the brain of the latter both larger and more gyrated. The major difference, however, lies in the thickness of the cerebral cortex and the density of neurons therein (Figure 5). Primate brains have approximately double the number of nerve cells in comparison with the brains of equally sized insectivores (Rapoport, 1990). There actually appears to be a trend in hominin evolution toward a reduction in total brain size (as indicated by cranial capacities) with a growing density of neurons; i.e., Neanderthals probably had similarly sized if not slightly larger brains than modern humans; with, however, less higher-order cognitive capacities.

Of all primates, humans may have the largest brains in total, but this increase in brain volume is by no means proportional but rather specific to the emergence of higher-order cognitive and motor functions. Comparing regions of the brain belonging to, e.g., the visual or olfactory systems, humans actually have considerably less neurons than other non-hominid primates (Rapoport, 1990). Those regions having undergone the most drastic growth in human evolution can be grouped into two major systems, namely the motor system as well as large parts of the cerebral cortex, the amygdala and hippocampal formation, which may be viewed upon as the cognitive or association system (Rapoport, 1990) (Table 1).

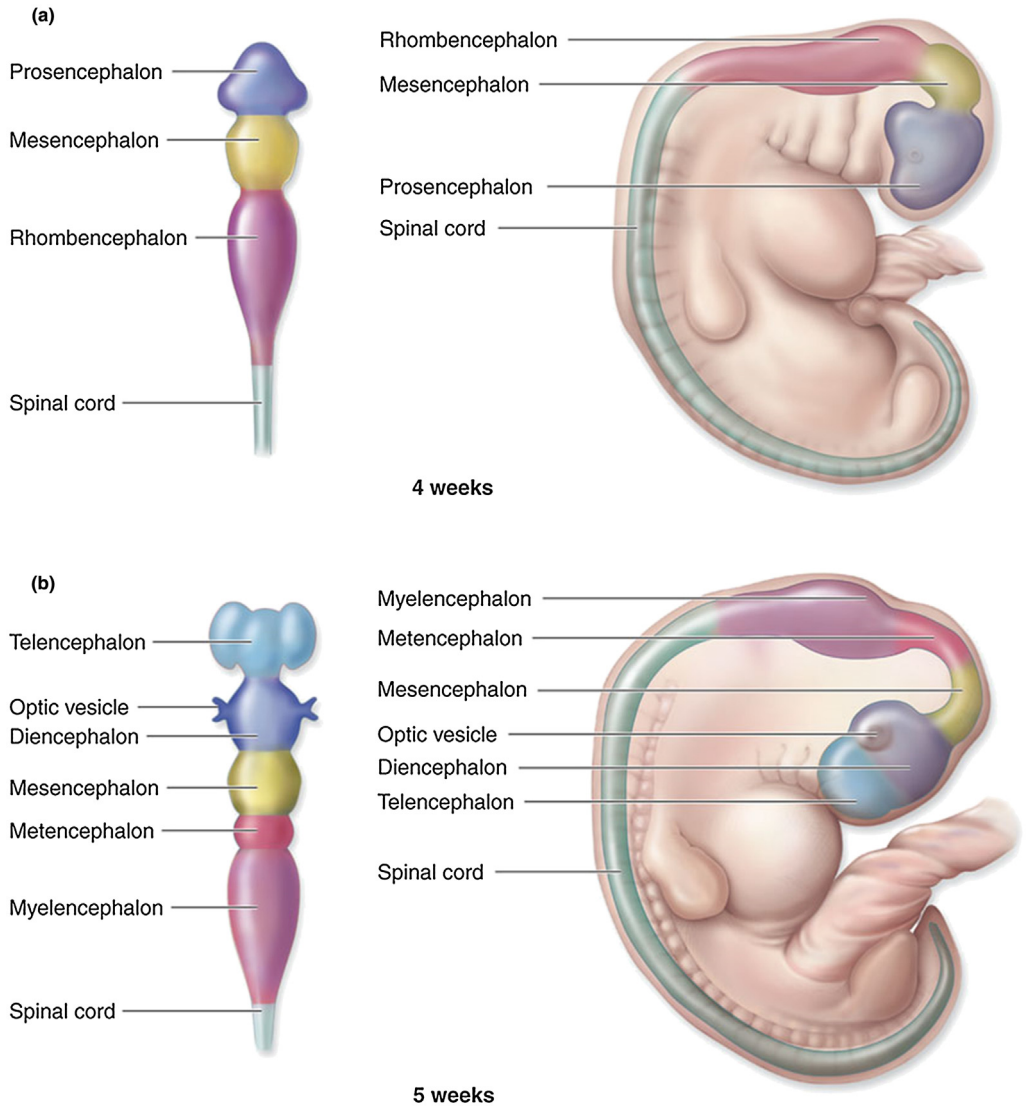

Figure 3 Early ontogeny of the human brain. McKinley and O'Loughlin. Human Anatomy, second ed. ISBN-10: 0077361369 ISBN-13: 978-0077361365; figure 15.2.

While the rapid growth of the brain's motor system may initially appear surprising, it has to be stressed that *H. sapiens* is the only extant primate-species to have developed exclusively bipedal locomotion, thereby necessitating a tremendous increase in neuronal control of the locomotor system. Of the cognitive or association system, two important areas need to be highlighted: the first is the cerebral cortex in general and the frontal lobe in particular. Humans are the only species in which the frontal lobe is larger – and considerably so at that – than the other neocortical lobes and extends above the eye-socket (i.e., the orbitofrontal cortex). This is due to the fact that the frontal lobe is involved in motor control and movement initiation, which actually forms the basis of all higher-order human functions, including speech, thinking, imagination and – as suggested by some authors – even empathy and theory-of-mind (e.g., Guillot et al., 2012).

The second mentionable region of the human forebrain to have undergone rapid and unique evolution in humans is the nucleus basalis of Meynert. It innervates almost the entire neocortex – using the neurotransmitter acetylcholine – in a manner only found in humans and is believed to represent the basis of drive and unspecific motivation: *The activity of the newly evolved nucleus basalis stimulates humans, starting at birth, to continuously explorative activity in the social and material world. This special functional construction of the human brain is responsible for all playing activities and trials of children, leading up to social, cultural and technical creativity of adults.* (Duncker, 2005: 20) (**Figure 6**).

Cortical Lobes and Hemispheres – Interconnectivity and Lateralization

As mentioned in the introductory paragraph, the cerebral cortex is divided into two hemispheres, which are each comprised of five lobes. Each of these lobes is related to a specific function (or modality) and possesses a cortical field, which directly

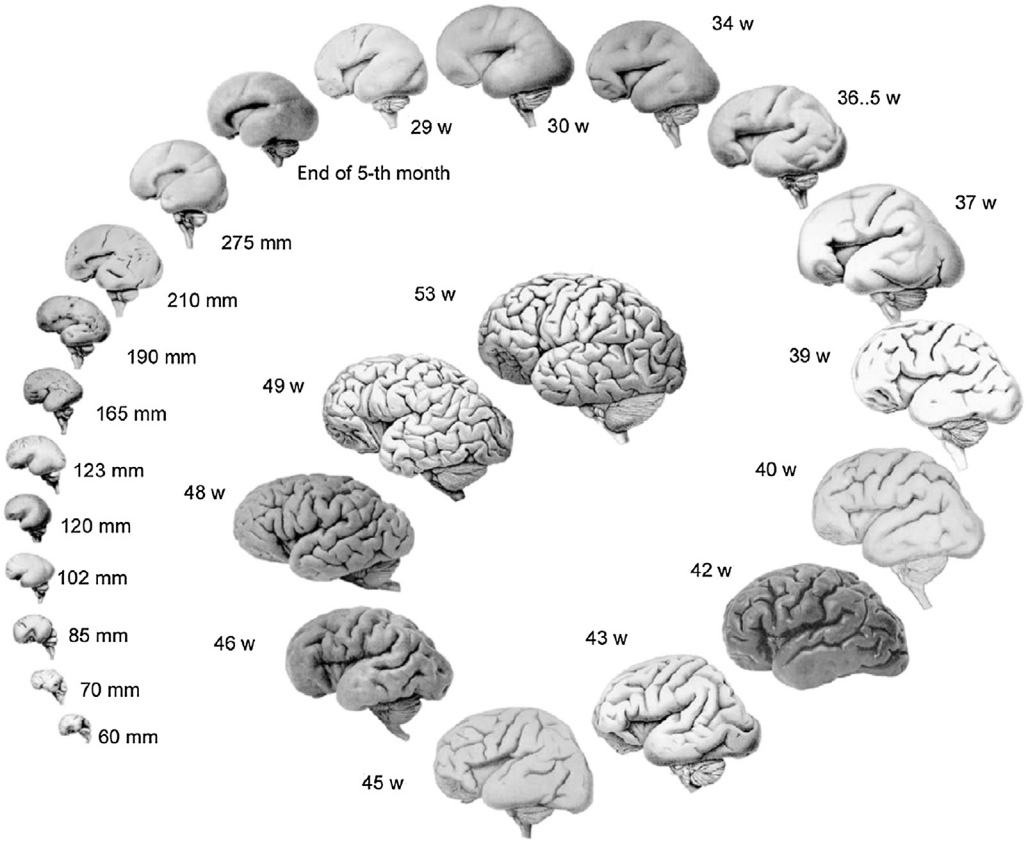

Figure 4 Development of the human cerebral cortex. Paxinos and Mai. The Human Nervous System, second ed. ISBN: 978-0-12-547626-3. http://www.sciencedirect.com/science/book/9780125476263; figure 3.1.

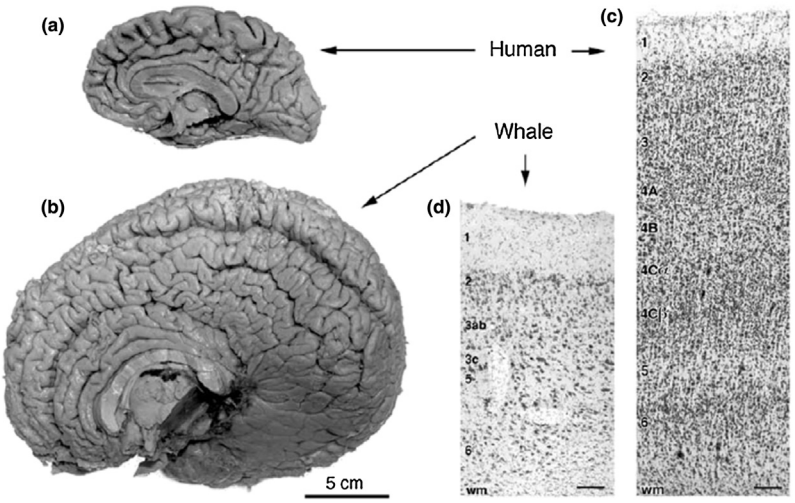

Figure 5 Comparison of neuronal density in the human compared to the whale cerebral cortex. Paxinos and Mai. The Human Nervous System, second ed. ISBN: 978-0-12-547626-3. http://www.sciencedirect.com/science/book/9780125476263; figure 1.1.

interacts with subcortical areas. These fields are called primary fields and border on secondary and tertiary fields, which connect with the primary and other cortical fields but not directly with other brain regions. The modalities associated with the frontal, parietal, occipital and temporal lobe as well as the insula are motor control, somatosensibility, vision, hearing, and viscerosensibility, respectively. Cortical fields not clearly attributable to any specific modality but involved in the integration of different modalities or the storage of memory are referred to as association areas.

Table 1 Overview of the brain areas having undergone massive evolution in human phylogeny

Brain region	*PI*[a] *humans*	*Division with maximal progression during primate evolution*
System I: Telencephalic regions, including amygdala and hippocampal formation		
Neocortex	156	Association areas, including: – prefrontal cortex – visual association – Broca's speech area (44 and 45, Brodmann) – inferior parietal lobule (39 and 40, Brodmann) – Brodmann area 37 (part of Wernicke's speech area)
Corpus callosum	↑	
Amygdaloid complex	3.9	Corticobasolateral group subiculum, CA1, dentate gyrus
Hippocampal formation	4.2	
Entorhinal cortex	5.5	
Septum	4.0	
Nucleus basalis of meynert	↑	
System II: Segregated circuits involving basal ganglia and thalamus		
Basal ganglia	16.4	Internal segment, globus pallidus
Thalamus	↑	pulvinar, dorsomedial nucleus
Substantia nigra	16	pars compacta
Cerebellar and mesencephalic-midbrain catecholaminergic nuclei		
Cerebellar nuclei	?	Dentate nucleus
Red nucleus	↑	Parvocellular division
Raphe nucleus	?	
Locus coeruleus	?	
Ventral tegmental area	?	
Retrorubral nucleus	?	

↑: PI > 1 but value not known in humans; ?: PI not reported.
[a]Progression index in *H. sapiens* (ratio of weight of human brain to weight of brain in primitive notional insectivore of equivalent body weight).
Rapoport, S.I., 1990. Integrated phylogeny of the primate brain, with special reference to humans and their diseases. Brain Research Reviews 15, 267–294.

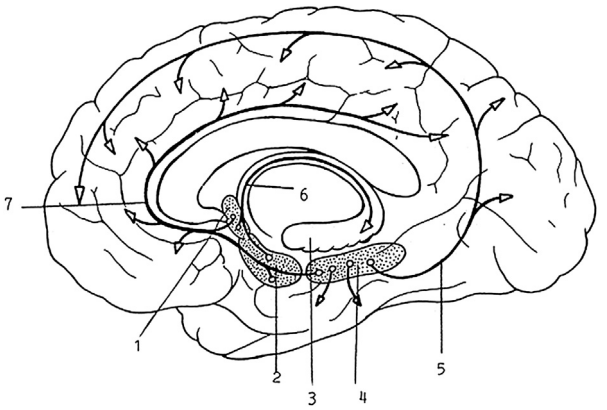

Figure 6 Schematic representation of the proposed organization of cholinergic projections of the basal forebrain in the human brain. 1. Medial septal nucleus; 2. nucleus of the diagonal band of Broca; 3. hippocampus; 4. nucleus basalis of Meynert; 5. external capsule; 6. fornix; and 7. cingulate bundle. Fadda, F., Rossetti, Z.L., 1998. Chronic Ethanol Consumption: From Neuroadaptation to Neurodegeneration. Progress in Neurobiology 56, 385–431.

The primary field of the frontal lobe is the precentral gyrus and is separated from the primary field of the parietal lobe – the postcentral gyrus – by the large central sulcus. The precentral gyrus projects directly to interneurons of the anterior horn of the spinal cord (also called the ventral horn), thereby giving collaterals to the parvocellular and magnocellular parts of the red nucleus in the midbrain. The parvocellular red nucleus sends these motor collaterals via the central tegmental tract to the superior olivary nucleus for integration in the cerebellum (Guillain–Mollaret triangle), while the magnocellular red nucleus comprises the starting point of the rubrospinal motor tract. A small number of exceptionally large pyramidal neurons of the precentral gyrus (Betz giant cells) project directly to those alpha-motoneurons in the spinal cord which innervate the small muscles of the hand.

The postcentral gyrus receives input from somatosensory thalamic nuclei via the corona radiata. These projections form the extensions of the ascending spinal and the trigeminal lemniscal tracts and convey protopathic (e.g., coarse touch, temperature and pain) and epicritic (gentle touch) information from the somatic areas of the body. Proprioceptive information from the locomotor apparatus project to the somatosensory thalamus indirectly via the superior peduncle of the cerebellum (brachium conjunctivum).

The primary field of the occipital lobe – the striate area – represents the termination field of the optic tract. Photoceptors in the retina relay information via the optic chiasm to the lateral geniculate body of the thalamus, which in turn projects to the striate area via an extensive fiber bundle within the corona radiata, referred to as the optical radiation.

The inner aspect of the temporal lobe, directly opposite the insular cortex, is comprised of three large gyri, the transverse temporal gyri of Heschl. These form the primary auditory field and receive their input from the medial geniculate body of the thalamus via the acoustic radiation within the corona radiata, thereby comprising the termination area of the auditory tracts.

The insular cortex receives viscerosensory information from various regions of the brainstem, primarily the nuclei of the solitary tract within the reticular formation. The only visceroafferent system not to project directly to the insula is the olfactory system. The primary cortical field representing olfaction is not located within the neocortex but in a part of the allocortex called the rhinencepahlon. It is located near the basal forebrain and is especially distinguished by its direct inputs from the olfactory bulb, which are not relayed within the thalamus (Figure 7).

Although the cerebral hemispheres appear identical to the naked eye, they are actually involved in different processing strategies. The so-called dominant hemisphere (left hemisphere in most humans) processes information in a sequential fashion and is therefore related to analytical functions like speech-processing, logical and mathematical thinking. The other (right in most humans) hemisphere processes in a more

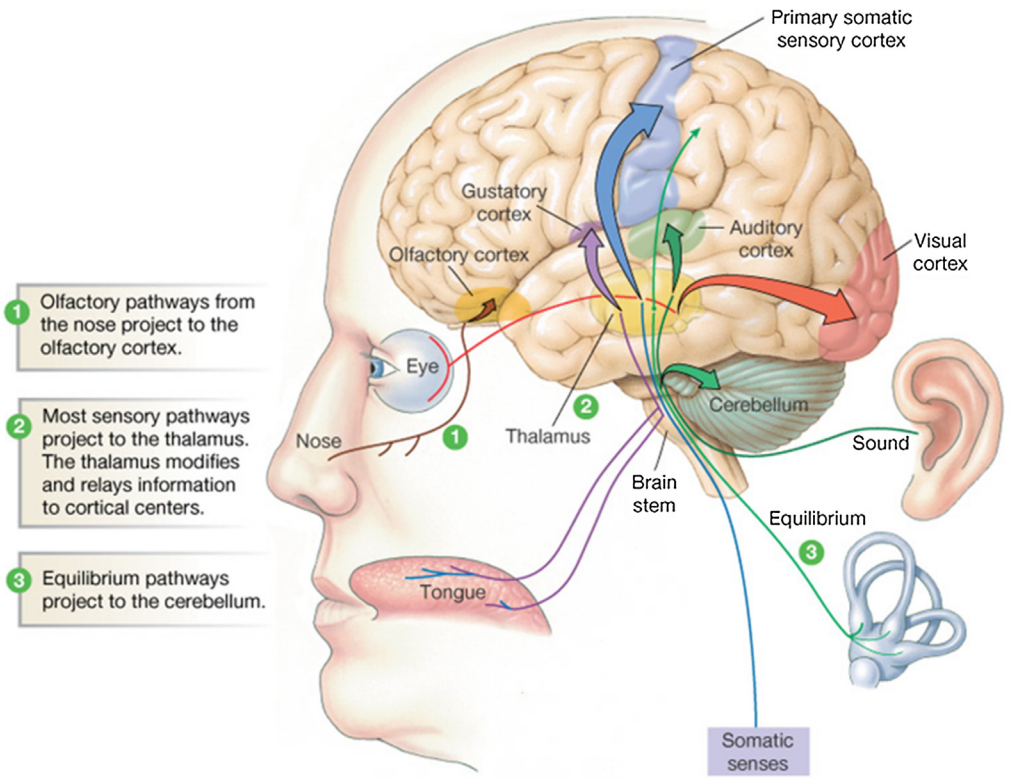

Figure 7 Overview of the cortical projection of the sensory pathways. http://faculty.pasadena.edu/dkwon/chap10_A/chap%2010_A%20accessible_files/textmostly/slide6.html.

holistic fashion and is therefore relevant for creativity, music- and art-processing, spatial orientation, personality, and facial recognition. The extent to which the cerebral hemispheres function differently from each other is called lateralization and is found to the greatest degree in humans, within whom it has also been suggested to be correlated positively with cognitive functions (Lebel and Beaulieu, 2009).

In order to function as a whole, especially with regard to higher-order processing, various areas within the cerebral cortex need to be interconnected, and the cortex additionally needs to communicate with noncortical and noncerebral brain areas. Those fiber tracts connecting similar areas of opposite hemispheres as called commissural fibers. The largest commissure in the brain is the corpus callosum, connecting major parts of both cerebral hemispheres. Although other cerebral commissures exist, the most important ones to mention are the anterior cerebral commissure, which connects both temporal lobes (including the amygdalae), and the fornical commissure, which connects both hippocampal formations.

Fibers connecting different areas within the same cerebral hemisphere are called association-fibers and can be divided into long and short association-fibers. The first span great lengths by connecting, e.g., the frontal with the occipital (superior longitudinal fasciculus) or temporal (arcuate and uncinate fasciculi) or the occipital with the temporal lobe (inferior longitudinal fasciculus). Another important long association tract is the cingulum, which lies within the eponymous cingulate gyrus and connects it with the hippocampal formation. Short association-fibers, on the other hand, are also called cerebral arcuate fibers and only connect adjacent gyri. The major body of long cerebral association-fibers is macroscopically referred to as the external capsule, whereby the arcuate fibers of the insula are analogously also called the extreme capsule.

Finally, fibers connecting the cerebral cortex with subcortical areas of the CNS, whereof the internal capsule is the largest and most prominent, are called projection-fibers. The internal capsule begins or terminates (depending on fiber direction) in the cerebral cortices and comprises many fiber tracts; inter alia, the corona radiata (between cortex and thalamus), the pyramidal tract (between the precentral gyrus and the anterior horn of the spinal cord), the spinothalamic tracts, and the medial as well as trigeminal lemnisci (connecting the epicritic nuclei of the brainstem with the thalamus). During its passage of the basal cerebral nuclei (aka basal ganglia), it lies between the caudate and lentiform nuclei. During ontogeny it displaces the caudate from the putamen, wherefore the resulting structure is called the striatum, due to small gray matter connections still remaining between the two nuclei. The part of the internal capsule near the apex of the lentiform nucleus exhibits a twisting, wherefore the various axons are in extremely close proximity to each other. This is of great clinical relevance, since lesions within this so-called genu of the internal capsules (e.g., through infarction) have dire consequences for the patient. After its passage of the basal cerebral nuclei and the diencephalon, the internal capsule lies at the anterior aspect of the midbrain and is here called the cerebral crura or (together with

parts of the midbrain tegmentum) the cerebral peduncles. The corticospinal or pyramidal tract can also be seen on the anterior aspect of the medulla oblongata where it forms the pyramids.

A second important projection-fiber is the fornix. It connects the hippocampal formation in the temporal lobe with the septal area of the basal forebrain and the mammillary bodies of the hypothalamus and is involved in memory consolidation as part of the extended neuronal circuit of Papez as formulated by Paul MacLean.

Cortical Plasticity and Change

Although the nervous system is generally known for its enormous capacity for synaptic plasticity and change, both cortical regions of the brain (i.e., the cerebellar and cerebral cortices) are exceptional in this respect. While often attributed to the works of Donald Hebb and the group surrounding Per Andersen (including Terje Lomo, who was the first to show long-term potentiation in the rabbit hippocampus during his doctoral work in the Andersen lab), the notion of synaptic plasticity as the basis of learning and memory was actually introduced by Santiago Ramón y Cajal during his Croonian lecture in 1894.

Two major factors are of relevance for the change in synaptic transmission and morphology, namely afferent input and exercise or repetition. The first factor has been shown repeatedly by, e.g., Witt-Lajeunesse (2001) and Comeau et al. (2010), who showed that rodents presented with a diversified and rich array of stimuli (so-called complex housing) show significantly higher levels of synaptic plasticity as well as resilience to neural lesions as compared with animals housed in a less stimulating environment. Especially the findings of the latter group are of special noteworthiness, since they were the first to show that stimulating sensory inputs not only lead to plasticity in primary sensory areas of the cortex but also in areas upstream regarding higher cortical functions; i.e., the prefrontal cortex. The adverse effect was shown, e.g., by Plank et al. (2011) in patients suffering from retinal dystrophy, who exhibited specific atrophy of the primary visual cortex.

The effects of exercise on synaptic plasticity have also been shown by numerous researchers. The aforementioned study by Plank et al. (2011) found a linear relationship between the fixation-stability of the eyes on a target and the gray matter volume of the frontal eye-field. Similarly, Draganski et al. (2004) recruited 24 healthy volunteers and instructed 12 of these in juggling, while the others served as controls. After 3 months, brain scans of the juggling-group showed significant increases in gray matter volume both compared with baseline and to controls. These changes began to normalize after another 3 months during which participants were instructed to no longer practice juggling, whereby gray matter volumes were still higher than at baseline, as the capability for juggling was not completely lost.

When comparing the volume of the posterior hippocampus, Woollett and Maguire (2011) found significant increases in participants who had successfully acquired the knowledge of the layout of Central London as part of their training to become a taxi-driver. With no initial differences between aspiring taxi-drivers and controls, several years later the authors found significant increases in gray matter volume in those participants that had successfully passed their examination; both compared with controls and to participants who had not acquired a license as a taxi-driver. The latter group had – on average – spent only half the time studying compared with the successful participants and did not differ significantly from controls regarding their change in the posterior hippocampal volume.

Finally, a number of studies both in humans and nonhuman primates show that either loss or the intense sensory stimulation of body parts lead to redistribution of corresponding areas of the cerebral cortex, a process referred to as recruitment or cortical remapping. Those areas that no longer receive sensory input are taken over by adjacent cortical fields, while vice versa cortex areas receiving increased sensory input enlarge and take over regions irrespective of whether they initially were related to other areas of the body.

Differential Lamination of the Cerebral Iso- and Allocortices

Isocortex

As mentioned in the introductory paragraph, the isocortex (corresponding to the neocortex) is generally comprised of six layers, while the allocortex (corresponding to the archi- and paleocortices) is mostly three-layered. The borders between one cortex type and the other are usually not sharp but made up of transition zones that can be subsumed under the name mesocortices.

The basic layout of the isocortex begins with an apical glial layer, the external glial limiting membrane, which borders on the outer isocortical layer, named the molecular layer due to its sparse neuronal population. The molecular layer entails a fiber plexus (plexus tangentialis or Exner-band) and is heavily involved in cortical ontogeny. As the first layer to develop from the subventricular zone, it encompasses a number of neurons (Cajal-Retzius cells) that attract subsequently developing neurons while simultaneously producing proteins that inhibit neuronal migration. As a consequence, the following layers of the isocortex develop inversely; i.e., sixth-layer neurons migrate until they reach the border of the molecular layer, followed by fifth layer neurons, that migrate through the anlage of layer VI, again until reaching the border of the molecular layer. This process subsequently is repeated by neurons forming layers IV, III, and finally II. After the termination of the migration of second-layer neurons, Cajal-Retzius cells go into apoptosis (programmed cell death), leading to the final low neuronal density in the fully developed isocortical molecular layer (**Figure 8**).

The innermost layer of the isocortex, the multiform layer, is made up of various different neuronal types and borders on the cortical white matter. Between the molecular and multiform layers (I and VI, respectively) two neurons typical to the cerebral cortices, namely pyramidal and spiny stellate or granule cells, form distinctive layers that are eponymously named. From the molecular layer downward these are the external granular (II), external pyramidal (III), internal granular (IV), and internal pyramidal (V) layers. Similar to the molecular layer, the layers III, IV, and V also encompass tangentially oriented fiber plexus, whereof the outer band of

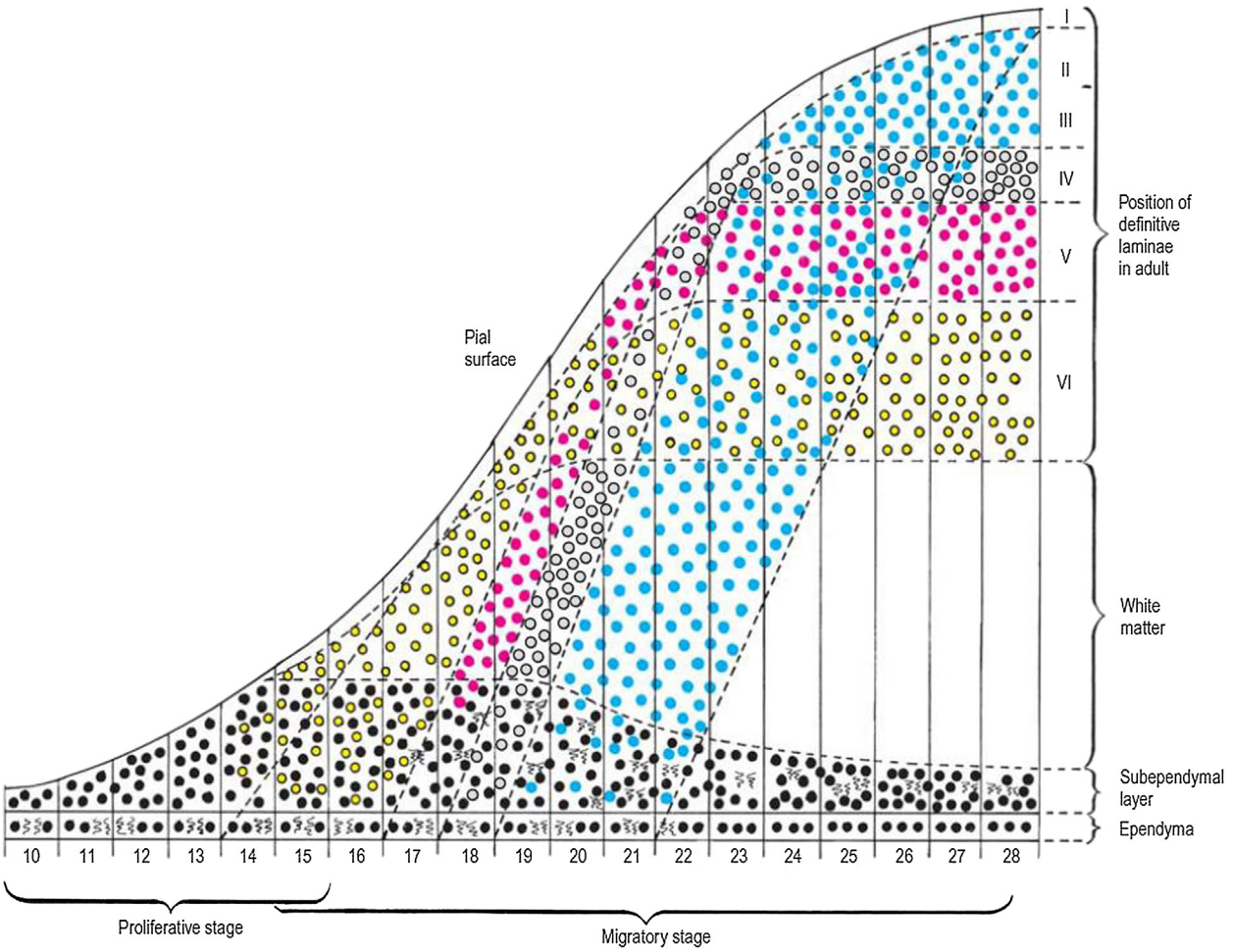

Figure 8 Neuronal migration during ontogeny of the human neocortex.

Baillarger in layer IV is especially noteworthy due to its historical connotations. Within the primary visual cortex it is of such prominence that it can be seen with the naked eye and was so seen by the medical student Gennari, wherefore it is also called the stria of Gennari within the analogously named striated area of the occipital cortex (in some texts, the stria of Gennari is also referred to as the stria of Vicq-d'Azyr) (Figure 9).

The functions of pyramidal cells and spiny stellate cells within the isocortex are discrete, whereby pyramidal cells give rise to efferent axons and spiny stellate cells represent the termination points of afferent fibers. The external layers (II and III) are mainly involved in corticocortical connections, whereas the internal layers (IV and V) receive or respectively give rise to projection fibers. Due to these functional differences, primary sensory cortical fields lack a clearly distinguishable layer V, but exhibit a very prominent layer VI, and are therefore called granular cortices. Primarily efferent cortical fields (most importantly the precentral gyrus) have a large layer V but not clearly distinguishable layer VI and are therefore called agranular cortices. It therefore becomes clear that, although histologically named isocortex, the lamination within the neocortex is not equal in all cortical fields. This differential lamination lead Korbinian Brodmann to propose over 50 individual cortical fields, many of which have since been corroborated by more modern methods (Figure 10).

Allocortex

Although phylogenetically older and three-layered, the same basic principles as mentioned for the isocortex apply to the allocortex. Since a detailed coverage of the entire allocortex would exceed the scope of this article, the lamination of the allocortex shall only be illustrated exemplarily for the two gyri comprising the retrocommissural hippocampus (or proper hippocampus). These gyri are the dentate gyrus and Ammon's horn. Unlike the isocortex, both of these gyri encompass only one fiber directional modality, making the basic connectivity of the proper hippocampus a one-way street. The dentate gyrus receives input (primarily from the entorhinal cortices) and therefore exhibits a prominent granular layer, surrounded by a molecular and a multiform or plexiform layer, but no pyramidal layer. Ammon's horn, on the other hand, is the efferent gyrus of the proper hippocampus and therefore has a pyramidal but no granular layer. Due to ontogeny, both gyri of the proper hippocampus are inversely positioned, wherefore their respective white matter tracts lay apically rather than

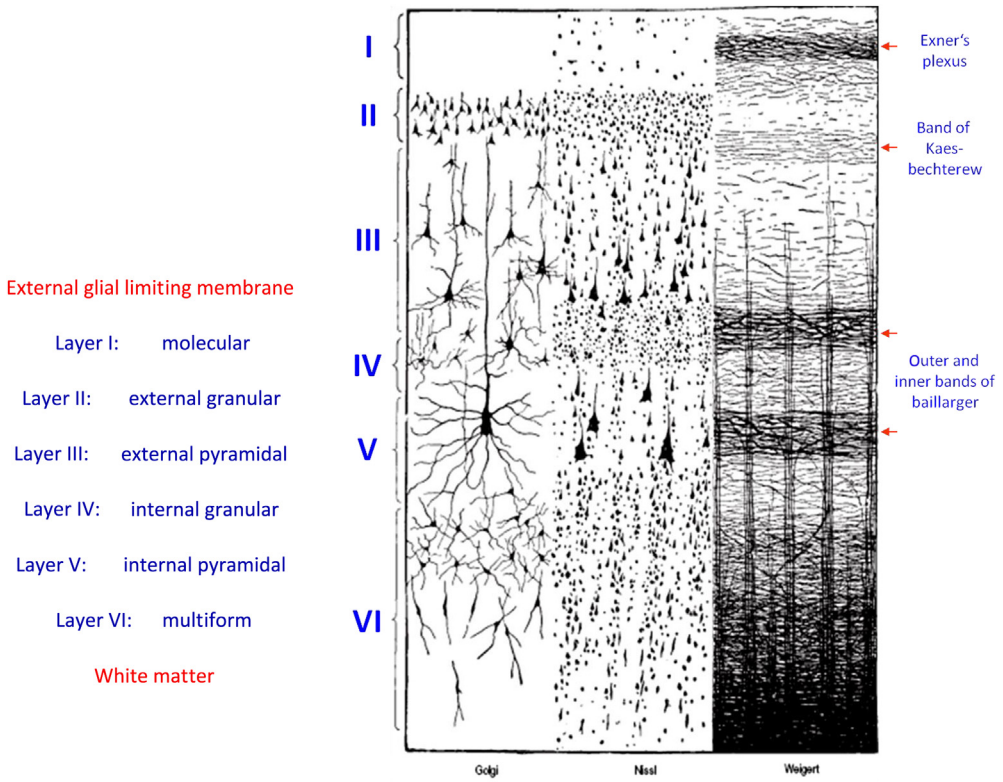

Figure 9 Cytoarchitecture of the neocortex (in three staining-methods). Modified by Grant from Strandring, Gray's Anatomy, 39th ed.; figure 22.10.

away from the surface. The white matter tract conveying inputs from the entorhinal cortices to the dentate gyrus is the perforant tract and lies within the molecular layer of Ammon's horn, the white tract of which is the alveus bordering on the cerebral ventricle. The molecular layer of Ammon's horn can be divided further, wherefore the layer encompassing the perforant tract is referred to as the lacunose molecular layer. It is separated from the pyramidal layer by the radiatum layer, which holds axon collaterals – the Schaffer-collaterals – connecting pyramidal cells from different sectors of Ammon's horn (namely CA3 to CA1; CA standing for Cornu ammonis, which is Latin for Ammon's horn). The final layer of the proper hippocampus, between the pyramidal layer and the alveus, is the oriens layer. Within the sector CA3, which lies near the dentate gyrus, a third layer can be distinguished with the molecular layer of Ammon's horn (also called Ammon's horn). This layer carries unmyelinated axons from the dentate gyrus' granule cells, that contact the CA3-pyramidal cells. It is therefore only present in CA3 and not stainable with classical techniques (e.g., Kluver-Barrera or Weigert), wherefore it is called the lucid layer (Figure 11).

Interactions of the Cerebral Cortex with the Thalamus

A number of misconceptions exist regarding thalamocortical and corticothalamic interactions. First, it is important to distinguish between two major portions of the thalamus, namely the dorsal (i.e., the thalamus proper) and the ventral thalamus. While the dorsal thalamus is usually referred to only as 'thalamus,' the ventral thalamus (not to be confused with the ventral nuclear group of the (dorsal) thalamus) is often subsumed under mainly the basal cerebral nuclei. It consists, inter alia, of the subthalamic nucleus, the external and internal pallidum, and the reticular thalamic nucleus, whereby only the latter shall by explained functionally here.

The main function of the ventral thalamus is to regulate the activity of the dorsal thalamus. This becomes inherently clear when taking into account that connections between the dorsal thalamus and the cerebral cortex are generally reciprocal; i.e., excitation or inhibition of the cortex by the thalamus in turn leads to respective excitation or inhibition of the thalamus by the cortex and vice versa. The existence of a controlling structure is therefore necessary and is found especially in the reticular nucleus of the ventral thalamus. This structure is thalamotopically organized, surrounds the dorsal thalamus like a shell, and receives collaterals of all axons between the dorsal thalamus and the cerebral cortex. While these axons are usually excitatory in nature, the ensuing projections of the reticular nucleus have an inhibiting effect of the respective nuclei of the dorsal thalamus (Figure 12).

Another misconception stems from the existence of functional subdivisions of the dorsal thalamus. It was long believed that the dorsal thalamus consisted of such regions involved in the relaying of information to the cerebral cortex and as such functioned as regulators of general cerebral activity. These regions were named the specific or corticothalamus and the unspecific or truncothalamus, respectively. The unspecific thalamus was considered an extension of the brain stem's ascending reticular activation systems and was, inter alia,

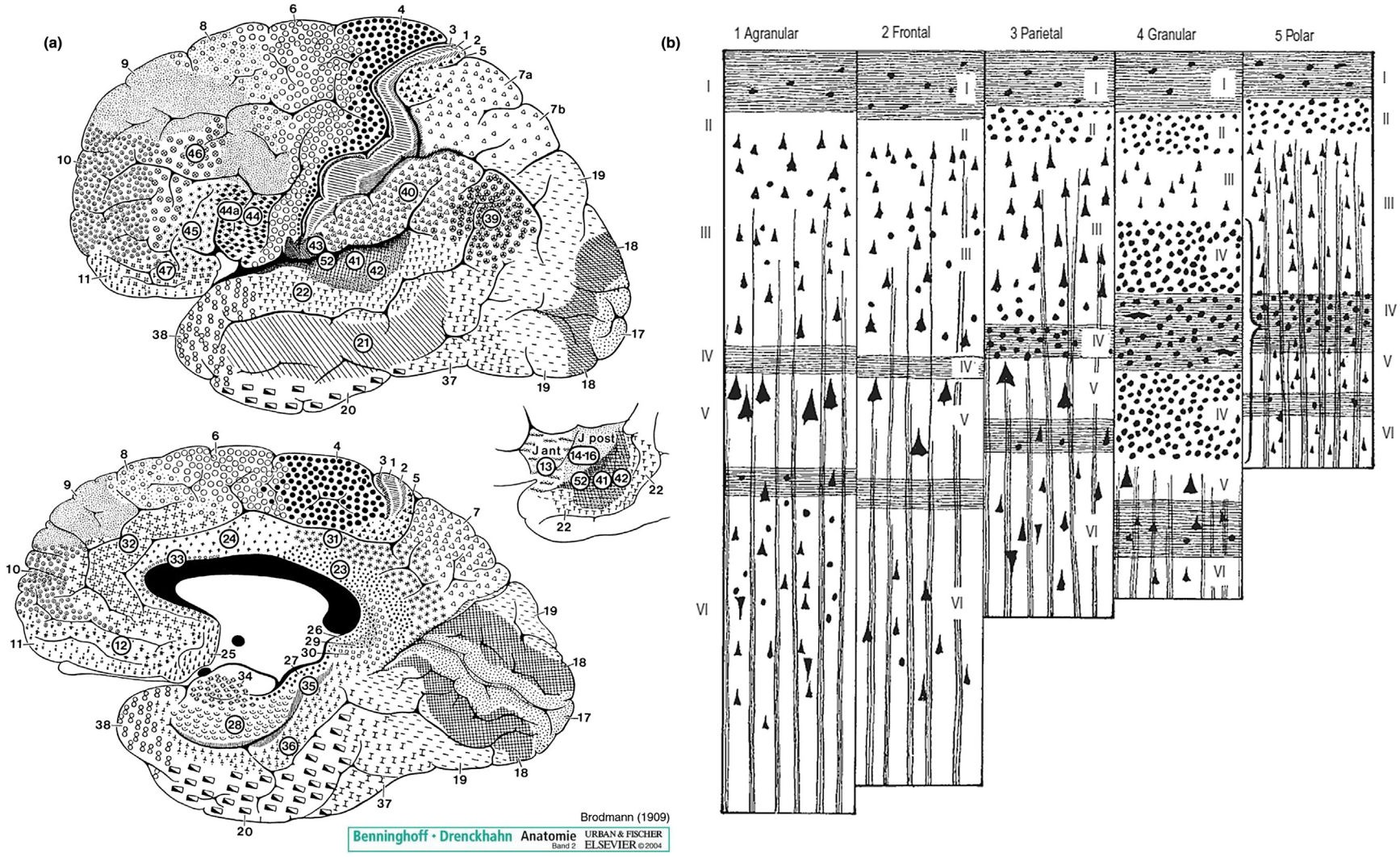

Figure 10 Differential lamination of the human neocortex. (a) Brodman's areas, (b) cytoarchitecture of major lamination types.

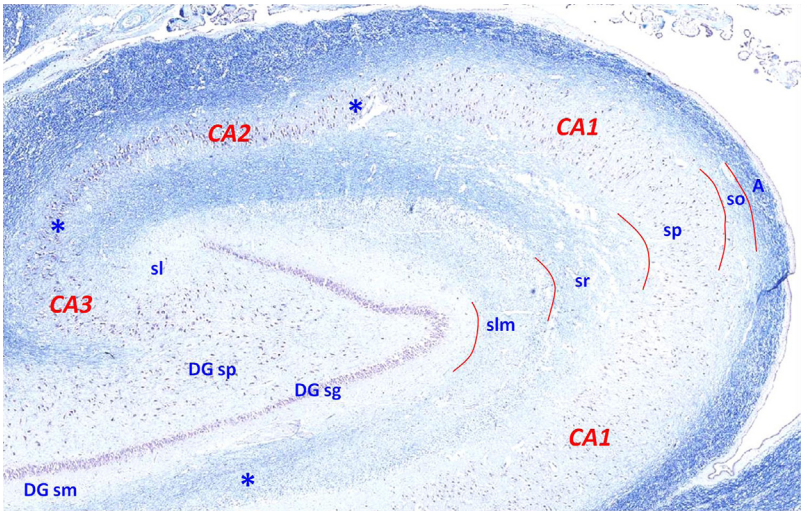

Figure 11 Microscopic anatomy of the human hippocampal formation.
A: alveus, CA: cornu ammonis, asterisks show the approximate borders between the CA fields, DG sg: granule cell layer of the dentate gyrus, DG sm: molecular layer of the dentate gyrus, DG sp: plexiform layer of the dentate gyrus, sl: stratum lucidum of the CA, slm: stratum lacunosum-moleculare of the CA; so: stratum oriens of the CA, sp: stratum pyramidale of the CA, sr: stratum radiatum of the CA. Source: Grant, P., 2011. Dopamine neurotoxicity, oxidative stress and schizophrenia – *in vitro* and *in vivo* studies of peroxisomal reactions to increased dopamine.

proposed by Eysenck as the neurobiological foundation of extraversion.

A number of studies from the early twenty-first century have, however, conclusively disproven this notion, wherefore neuroscientists now propose the distinction between isothalamus and allothalamus: While the allothalamus communicates primarily with the basal cerebral nuclei, the isothalamus interacts with the cerebral cortex. It has been shown that each and every isothalamic nucleus has both neurons that relay information to the thalamus (by projecting to the internal granular layer) as well as neurons that have slightly broader regions of cortical innervation, projecting into the molecular and multiform layers and being therefore believed to influence the way in which the cortex processes information relayed by the thalamic neurons projecting into layer IV.

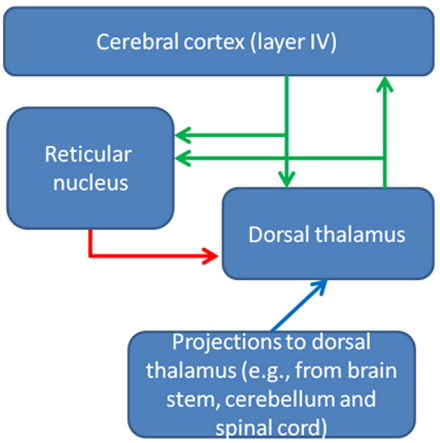

Figure 12 Schematic illustration of the neuronal interaction between the dorsal thalamus and the cerebral cortex; role of the reticular nucleus; green indicates excitatory and red inhibitory projections.

There also exist isothalamic interneurons, wherefore it is believed that a number of corticocortical connections may actually not be direct but relayed through the dorsal thalamus. This mechanism additionally allows for control over thalamic activity by the cerebral cortex itself. This is also of major importance to thalamocortical function, as a major job of the thalamus is the regulation of cortical processing (see above). Here also misconceptions exist, as it is suggested that the thalamus 'decide' on the filtering of information. This is, however, not true, as the thalamus itself does not actually filter information out and also does not possess the necessary 'background information' to achieve this task (Figure 13).

This can be illustrated for selective information processing by example of the so-called 'Cocktail-Party effect': An individual following a specific conversation during a party will filter out all other conversations going on within earshot. Hearing his or her name (or any other personally salient word) used within any of these other conversations, however, will lead to an immediate relocation of attention to the previously ignored conversation. This example clearly illustrates both the function of corticothalamic connectivity as well as the impossibility of filtering solely on the thalamic level. First, the thalamus does not store information of this nature, wherefore it cannot 'interpret' its personal relevance as only the forebrain can. Second, the 'dropping' of one's own name is not anticipated and therefore occurs during a time when the conversation is supposedly still 'filtered out,' wherefore it is clear that it is in actuality not filtered out, but reaches the limbic areas of the cerebral cortex, if only processed differently there compared with the conversation that had been followed to this point. Third and finally, the immediate relocation of attention – in other words, the nature of cortical processing of received information – shows that connections exist between different cortical regions via the thalamus, in this example from (probably) the prefrontal and association cortices to the primary and secondary auditory cortices.

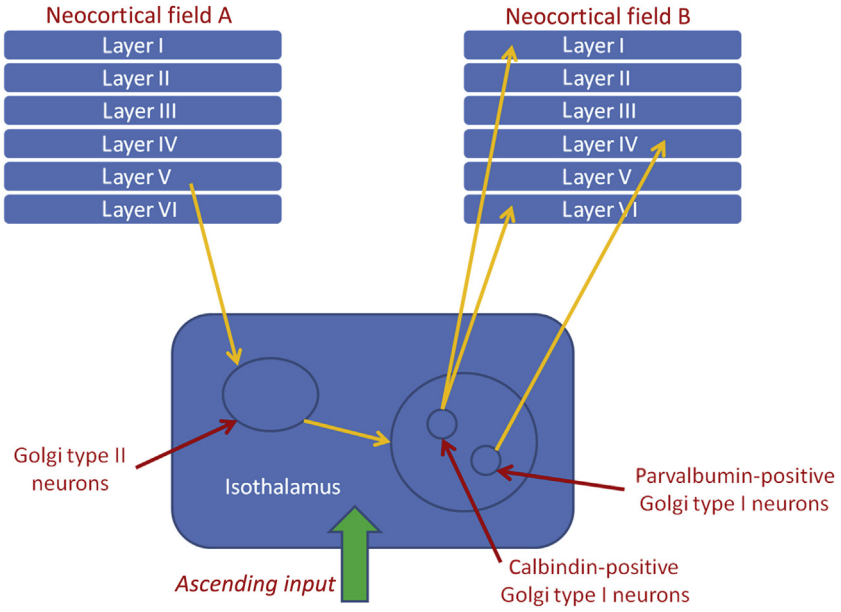

Figure 13 Schematic illustration of thalamo-cortical and cortico-thalamo-cortical projections.

In conclusion, the thalamus is fundamentally involved in the function of the cerebral cortex as it not only relays information to layer IV spiny stellate cells; it additionally protects the cerebral cortex from overload (a concept that actually exists on the neuronal level in so-called neuronal excitotoxicity) through the ventral thalamus, especially the reticular nucleus. Finally, it is inherently relevant for the cortex' own control of information processing through the isothalamus as outlined above.

See also: Amnesia: General; Attention, Neural Basis of; Brain Asymmetry; Brain Development: Ontogenetic Neurobiology of; Brain, Evolution of; Brain: Response to Enrichment; Cognition, Evolution of; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Consciousness, Neural Basis of; Cortical Areas Engaged in Movement: Neuroimaging Methods; Cortical Development: Transplantation and Rewiring Studies; Declarative Memory, Neural Basis of; Electroencephalography: Basic Principles and Applications; Epigenetics of Brain Plasticity and Behavior; Extraversion; Hippocampal Amnesia; Hippocampus and Related Structures; Human Cognition, Evolution of; Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of; Intelligence: Central Conceptions and Psychometric Models; Intelligence: Historical and Conceptual Perspectives; Learning and Memory, Neural Basis of; Limbic System; Long-Term Potentiation and Long-Term Depression; Motor Cortex; Neural Plasticity; Neurons and Dendrites, Integration of Information in; Neurotransmitters; Occipital Lobe (Visual Cortex): Functional Aspects; Olfactory System; Parietal Lobe; Personality, Biological Models of; Personality, Trait Models of; Prefrontal Cortex Development and Development of Cognitive Function; Prefrontal Cortex; Premotor Cortex; Primate Behavior: Significance for Understanding Humans; Primates, Evolution of; Semantic Memory; Sleep: Neural Systems; Temporal Lobe; Topographic Maps; Visual Perception, Neural Basis of.

Bibliography

Comeau, W.L., McDonald, R.J., Kolb, B.E., 2010. Learning-induced alterations in prefrontal cortical dendritic morphology. Behavioural Brain Research 214, 91–101.

Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U., May, A., 2004. Changes in grey matter induced by training – newly honed juggling skills show up as a transient feature on a brain-imaging scan. Nature 427, 311–312.

Drenckhahn, D. (Ed.), 2003. Benninghoff: Anatomie, sixteenth ed. Urban & Fischer, München Jena.

Duncker, H.-R., 2005. Human as a biological and cultural being. In: Duncker, H.-R., Prieß, K. (Eds.), On the Uniqueness of Humankind. Springer, Berlin, Heidelberg, New York.

Fadda, F., Rossetti, Z.L., 1998. Chronic ethanol consumption: from neuroadaptation to neurodegeneration. Progress in Neurobiology 56, 385–431.

Grant, P., 2011. Dopamine Neurotoxicity, Oxidative Stress and Schizophrenia – In Vitro and In Vivo Studies of Peroxisomal Reactions to Increased Dopamine. Giessen, Univ., Diss urn:nbn:de:hebis:26-opus-84954.

Guillot, A., Di Rienzo, F., MacIntyre, T., Moran, A., Collet, C., 2012. Imagining is not doing but involves specific motor commands: a review of experimental data related to motor inhibition. Frontiers in Human Neuroscience 6. http://dx.doi.org/10.3389/fnhum.2012.00247.

Lebel, C., Beaulieu, C., 2009. Lateralization of the arcuate fasciculus from childhood to adulthood and its relation to cognitive abilities in children. Human Brain Mapping 30 (11), 3563–3573.

Nieuwenhuys, R., Voogd, J., Van Huijzen, C., 2008. The Human Central Nervous System, fourth ed. Steinkopff.

Paulsen, F., Waschke, J. (Eds.), 2010. Sobotta: Atlas der Anatomie des Menschen – Kopf, Hals und Neuroanatomie. Elsevier, Munich.

Plank, T., Frolo, J., Brandl-Ruehe, S., Renner, A.B., Hufendiek, K., Helbig, H., Greenlee, M.W., 2011. Gray matter alterations in visual cortex of patients with loss of central vision due to hereditary retinal dystrophies. NeuroImage 56, 1556–1565.

Rapoport, S.I., 1990. Integrated phylogeny of the primate brain, with special reference to humans and their diseases. Brain Reserach Reviews 15 (3), 267–294.

Standring, S. (Ed.), 2008. Gray's Anatomy: The Anatomical Basis of Clinical Practice, 40th ed. Churchill Livingstone Elsevier.

Witt-Lajeunesse, A., 2001. Effects of behavioral therapies and pharmacological intervention in brain damage (Master's Thesis). Lethbridge College, Alberta, Canada.

Woollett, K., Maguire, E.A., 2011. Acquiring "the knowledge" of London's layout drives structural brain changes. Current Biology 21, 2109–2114.

Change Blindness

Robert W Kentridge, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Change blindness is a phenomenon in which major changes to a visual scene go unnoticed. There are many methods of inducing change blindness, for example, by presenting a blank image between presentation of the original and changed pictures. Change blindness is thought to occur when visual attention is prevented from being drawn to the change. Detecting the changes requires a comparison between the changed state of the picture and a visual memory of its original state. Without visual attention the memory may not be retrieved at all or the available memory may lack sufficient visual detail for a change to be registered. Change blindness is employed as a tool for studying visual attention and has obvious real-world implications for tasks such as driving.

Overview

If we are looking at a picture and one thing in it changes, someone's hat turns from black to white, we would normally expect to see the change easily. In fact, we would expect the change to jump out at us. There are, however, circumstances where such changes are almost impossible to detect. This is the phenomenon of change blindness. The circumstances in which we fail to detect changes easily all seem to affect visual attention. If, for example, some irrelevant bright spots appear at the same time as the picture changes then we usually fail to detect the change in the picture. Understanding change blindness can help us understand the role of visual attention in perception, and in particular, conscious perception. The mechanisms underlying change blindness are also likely to depend upon the properties of visual short-term memory and the nature of visual representations. Explaining change blindness is therefore of considerable theoretical interest to psychologists. The ability to detect changes in a scene is also of great practical importance. Drivers, for example, need to see a child stepping into the road or a traffic light changing. Understanding how factors in their environment might interfere with these abilities is an example of especially valuable applied psychology.

History and Methods

The discovery of change blindness arose out of research aimed at understanding how the visual system integrates information across eye movements. Although the visual world appears stable to us, we actually move our gaze rapidly from place to place several times per second. These rapid shifts of gaze are known as saccadic eye movements or saccades. The intervening periods during which we keep our gaze fixed on one part of the visual scene are known as fixations. We acquire very little, perhaps no, information during saccades. At each fixation most of the information we do acquire is from a small region of space centered on the direction of fixation. We acquire much less information from the visual periphery as the resolution of visual system decreases markedly away from the direction of gaze. Our sensation of an apparently stable continuous world must be constructed by integrating disjointed patches of vision acquired one after another. One key issue in understanding the integration of information across fixations is the extent to which the relatively poor information acquired from locations in more peripheral vision during one fixation is used to interpret information from subsequent fixations that are made to those locations. Information from the periphery, although not as precise as that obtained from central vision, is certainly used by the visual system in selecting destinations for subsequent saccades. But how much peripheral information is actually used for purposes beyond saccade guidance? Does this peripheral information allow us to match parts of transitory visual representations between fixations in the creation of a global representation of the world?

Early work on integration of information across saccades focused on reading. In this work subjects' direction of gaze was continuously measured with an eye-tracker so that the visual stimulus they were presented with could be manipulated between fixations. Creating such 'gaze-contingent' displays is technically demanding today and was all the more so in the 1970s. At that time it was possible to manipulate a display of text (but not yet a realistic image) so that it changed as a subject moved his or her eyes. This allows one to ask exactly what information is acquired from the periphery on a given fixation. If we can, for example, detect the identity of a letter in the periphery then when we subsequently saccade to it we will notice if the letter has changed. If we are unable to detect letter identity in the periphery then if the letter changes as we saccade to it we will be unaware and will continue reading as normal. One early, and very surprising, result found by McConkie and Zola (1979) was that the visual details of letters could be changed across saccades without being noticed even when changes in the identity of those letters would be detected. If a letter changed from an 'a' to a 'b' the change would be detected, but a change from a lowercase 'a' to an uppercase 'A' would go unseen (in the experiment the text alternated in case, e.g., 'ThE qUiCk BrOwN fOx,' but changed in the order of case alternation across saccades so that a change to 'tHe QuIcK bRoWn FoX,' was undetected). This was all the more remarkable because subjects claimed that they were seeing text in their peripheral vision and that no changes were occurring. This suggests that, in reading at least, our global representation of the world is maintained at quite an abstract level and our

impression of continuous physical detail is 'filled in' without truly being represented or at least explicitly available to us.

A few years later John Grimes (described in Grimes, 1996), working in McConkie's laboratory, began using a system that allowed full color pictures to be switched as a subject made a saccade. Grimes manipulated pictures using Photoshop so that one or two objects in the picture changed. For example, one original picture shows two men, one wearing a black hat and one a gray hat. In the manipulated picture the hats switch colors. If the change between the two images occurs during a saccade subjects very rarely detected the change even though it was glaringly obvious when not synchronized with an eye movement. This failure to detect large changes in a visual scene is an example of what we know known as change blindness.

As the eye-tracker in Grimes' experiments records eye-position as well as controlling the times at which images change it is possible to determine where a subject has been looking during the experiment. Grimes discovered that even when subjects had looked directly at the changing object before it changed, and looked at it again at some time after it changed, they still did not notice the change. Even with information acquired directly from fixation the representation stored in visual short-term memory must either be so abstract that changes in physical detail are unseen or, if physical detail is represented, changes must be integrated without reference to the prior state of the representation.

Grimes' results prompted a number of investigators, most prominently Ronald Rensink and Daniel Simons, to ask why a saccade prevented people from detecting changes in visual scenes. When we make a saccade the visual system partially suppresses sensitivity to visual signals (at least in the part of the visual system sensitive to rapid motion, Burr et al., 1994), a process known as saccadic suppression. What signals that remain are blurred by the motion of the eye. Rensink et al. (1997) asked whether the process of saccadic suppression itself was essential to change blindness or whether just the brief period of motion blur between steadily fixated images was sufficient to prevent changes from being detected. Rensink addressed the question using a very simple experimental design. Instead of waiting for subjects to move their eyes before switching images he interleaved a very brief presentation of a plain gray image in between presentation of original and manipulated pictures (see **Figure 1**). The brief gray image is, effectively, substituting for motion blur. Rensink was able to alternate back and forth between the two versions of an image repeatedly and, provided there was a blank gray screen between each change, subjects would fail to notice large changes to an image over large numbers of alternations. There was no need to eye-track and subjects were free to move their eyes while trying to detect what was changing in the image.

In Rensink's flicker paradigm the change between images is never directly visible to the subject. It is possible that subjects' failure to detect the change might be due to some form of masking occurring between the gray frame and the changed part of the image. To test whether masking effects might account for change blindness O'Regan et al. (1999) developed a variant of the flicker paradigm that did away with the interleaved gray frames. Instead, every time the image changed a number of small, high-contrast, checkered patched appeared briefly at random locations in the image. O'Regan and his colleagues refer to these patches as 'mudsplashes' in analogy to the effect of mud splashing on a car windscreen. The mudsplash patches did not cover the part of the image that changed but, nevertheless, subjects were still unable to detect the

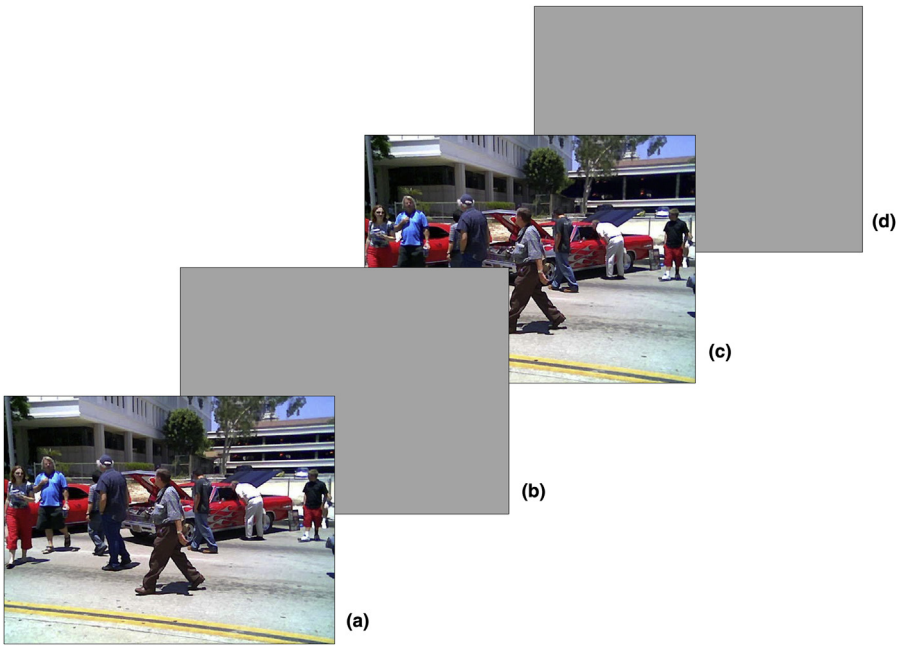

Figure 1 A schematic illustration of the change blindness flicker paradigm. The observer is presented with the images (a), (b), (c), and (d) in quick succession, for example for half a second each. The presentation repeats in the same sequence until the observer detects that change made between images (a) and (c). Notice that the middle storey of the car-park in the upper right background has been removed in image (c). Observers typically find such changes extremely difficult to detect and may take hundred of repetitions of the sequence before they spot the change.

change. Masking cannot, therefore, account for change blindness.

One explanation for the effectiveness of both the flicker and mudsplash paradigms is that they both prevent the change in the image from attracting visual attention. If one simply alternated between two versions of an image without an interleaved frame then the elements of the image that change produce a spatially localized visual transient that can attract attention to the location of the changing part of the image. In the flicker paradigm the entire image changes to gray when the interleaved frame is presented and so there is no localized transient that can specifically guide attention to the location of the change. In the mudsplash paradigm the changing part of the image is not obscured and so does generate a localized transient, but at the same time as this transient occurs the mudsplashes are also producing competing highly salient transients and so reducing the chances that the changing part of the image itself will attract attention. Changes in a scene only attract our attention if they are abrupt. When changes occur gradually then even in the absence of any other manipulation they are likely to be missed (see e.g., Simons et al., 2000).

The mudsplash paradigm allows explanations of change blindness beyond disruption of attentional capture to be tested. In a picture of a real-world scene some aspects of the image will be more important, in terms of their meaning, than others. In a picture of two people having dinner al fresco the people are more important to our understanding of the image than the position of clouds in the sky, even though the clouds may form a larger part of the image in terms of area. O'Regan et al. used a variant of the mudsplash paradigm to intentionally draw subjects' attention to the location of a change within an image. They positioned a single mudsplash at the location of the change. The change was therefore obscured, but the single mudsplash drew subject's attention to its location. When the change occurred in a part of the scene important to its meaning subjects could identify the nature of the change immediately. This, again, implies that masking cannot explain change blindness. Interestingly, however, if the change (and accompanying mudsplash) occurred in a part of the image with little significance for its meaning subjects were often unable to identify the nature of the change. Even though subjects were attending to the change location they were not encoding that part of the environment with sufficient detail to know what had changed there. This implies that the richness of our visual representation varies according to the significance we attribute to different parts of the scene.

All of the paradigms discussed above point to the importance of attention, or specifically the lack of it, in inducing change blindness. Lack of attention cannot, however, be the whole story. Even when the changing object is attended subjects can fail to notice that an object has changed if the moment of change is not visible. One of the most dramatic demonstration of this is from a study by Simons and Levin (1998) in which subjects fail to notice that the (real) person to whom they are talking changes from one individual to another when the switch is obscured by a visual interruption (a door is carried between them). Similar failures to detect change occur when films cut from one view of an actor to another. Detecting a change in an object involves a comparison between the state of the object before and after the change. There are a number of ways in which this comparison could fail. Memory is clearly involved in comparing the current state of an object and its state in the past.

We have seen that the psychological significance of objects affect the likelihood that changes to them are detected. Change detection involves making a comparison between the current and previous states of an object so perhaps both attention and significance affect memory for the initial state of objects. It is certainly the case that factors known to influence memory affect change blindness. For example, we become less likely to detect changes in items, the longer the interval between the time they were fixated and the time at which the change occurs (Hollingworth, 2004) – there is a 'recency effect' in change blindness. Of course, the fact that a change can go undetected for considerable time and yet eventually be recognized demonstrates that change blindness is not solely a failure to encode memories. Changes in a scene may interfere with memories that have already been encoded or encoded memories of objects may only be retrieved when those objects are attended. Informing subjects which object in a scene may have changed using a visual cue presented after the change can have no effect on encoding but subjects' ability to indicate whether the cued object changed was markedly better than their ability to report whether *any* object in the scene had changed in the uncued condition (Hollingworth, 2003). Of course, some changes may remain undetectable even when a subject is cued to report changes in a specific object. As we have seen in the early experiments of McConkie and Zola (1979) in which a change in the case of letters was undetected, change is only detectable if the property that changed has been encoded and retrieved. There is, however, evidence that changes to visual properties that have been encoded may not always be enough to permit change detection even when retrieval is cued. Fernandez-Duque and Thornton, 2000 showed that when subjects report that a specific object has not changed they take longer to respond if the object did in fact change than they do if the object remained the same. Laloyaux et al. (2008) presented subjects with photographs each showing three people. Unbeknown to the subjects the emotional expression on one of the faces changed gradually over the period that the photograph was being viewed. At the end of the stimulus presentation the subjects are shown one of the three faces from the photograph and asked to report whether it has changed. They are also asked to report which of five versions of the face they saw. Subjects fail to detect the change in the majority of trials, however, even when the change is undetected they choose the face from the five examples that most closely matched an expression seen during the course of the undetected change rather than the expression seen either at the start or at the end of the presentation. Information about a changing object may therefore be encoded in memory and remain to some extent intact despite changes in the stimulus and still not drive change detection.

Theories

It is clear from the variety of paradigms in which change blindness can be demonstrated that both attention and visual short-term memory play a part in determining whether or not

a change will be detected. Explanations couched solely in terms of attention or solely in terms of memory (e.g., the proposal that the representation of a scene after a change completely overwrites prior representations) are not, however, able to explain all aspects of change blindness. A number of theories propose that combinations of attention, memory, and the significance of objects within the scene determine whether or not changes in these objects will be detected.

Change blindness can be explained at a number of different levels. Hollingworth (2003) and Mitroff et al. (2004) specifically address the issue of *why* changes are missed. Their explanations are couched in terms of the importance of changes and their relevance to the task of interpreting a scene. Changes may be missed because they do not contribute to the meaning of a scene and so remain largely unattended. In many examples of change blindness the undetected changes, although often large, do not affect the meaning of the scene as a whole. The precise location of the horizon in scene showing a sailing boat cutting through the waves has no impact on our interpretation of the scene, although moving the horizon may change a large area of the image. One can then go on to consider why attention may be necessary in order to detect change.

Many authors start from the position of assuming that the initial preattentive visual representation of objects is transitory. It is something akin to iconic memory that decays rapidly. This representation may already be parsed into objects by some scene segmentation process, but if objects are segmented they still have no meaning, their contents are just visual details. Perception involves taking this raw representation of objects' components of and transforming it into a stable representation in which the meaning of the scene and the object in it are represented. The notion that attention serves to bind features into object representations has a long history and many theories of change blindness and of scene perception in general emphasize this dissociation between a rich preattentive experience and a sparse cognition of the gist of scene. We only notice changes in the sparse semantically meaningful representation and only access visual detail within the scene as we need it – we do not store a representation of visual detail but obtain from the world.

Rensink (2000), for example, suggests that before attention is deployed our representation of the portion of a scene that we are currently looking at takes the form of 'proto-objects' containing the features of objects. This representation is rich in that all of the information about the features of the object are present; however, these features are not bound together into coherent representation encompassing the identity of the specific object. These mutable representations only reflect the current content of the world. As things change in the world, so do their proto-object representations and so, in the normal course of events, changes will be undetectable as there is no stable memory for these proto-objects. If, however, attention is focused on an object then its representation is stabilized – the component features of the proto-object are used to identify it. In this model it is the representation of the proto-objects that is rich containing all the details of their visual properties. The representations that are formed when a proto-object is attended are sparse – it is as if they are merely labels identifying the objects and a few of their key properties. Once attention is withdrawn from an object its representation reverts to that of a proto-object. Attention has only limited capacity and so we can only maintain stable representations of very few objects at any one time.

A key aspect of these models is the fragility of visual memory. The basis of change blindness in these models is that object representations dissolve back into unbound collections of features as soon as attention is withdrawn from them. It is not therefore possible to make comparison with previous states of objects after attention has been withdrawn from them. Experiments in which change blindness is eliminated when location of a change is cued after it has occurred, such as those of Hollingworth described previously, seem to pose a problem for models based on the fragility of memory. If we are told which object has changed after the change occurred we are still able to recover memory of its prior state. It appears, however, that in normal circumstance we simply do not recover these memories.

Change blindness was discovered in the course of work on the integration of visual information across saccades and it is work still focused on this problem that may explain our failure to use what visual memories we have in change detection. One remarkable feature of the visual system is the efficiency with which we correct inaccurate saccades. If, when we move our eyes to fixate a target, our gaze misses the intended location we make a corrective saccade almost immediately. Moreover, we do not notice that we have done so. The mechanisms of corrective saccades have been studied by intentionally moving visual targets as a subject is making a saccade toward them. Again, these movements typically remain undetected. It has been suggested that one of the assumptions built into the visual system, a Bayseian prior, as it were, applied in interpreting the visual stimulus, is that objects remain stable in identity and position across saccades. It is far more likely that a saccade is slightly inaccurate than it is that an object moved or changed during the very brief time it takes to make the saccade to it. So, although we may have visual memory for currently unattended objects, the visual system does not access these memories unless there is strong evidence against the assumption of object stability. This was first demonstrated by showing that movements in the position of a saccade target that were normally unnoticed became detectable if the entire visual display blanked briefly just after the competition of saccades (Deubel et al., 1996). This blanking is a strong cue that stimuli are not stable. During normal vision, however, there is a very strong bias in favor of object stability and so although information about currently unattended objects may potentially be available in visual memory it will not normally be accessed and hence normally changes in the scene that do not capture attention may be missed.

Related Work

Change blindness is often used as a tool in studies aimed at assessing exactly which properties of a stimulus capture attention and are hence immune to change blindness (e.g., Cole et al., 2004). Understanding our abilities to detect or miss changes has obviously application in the real world. Studies using change blindness paradigms have shown that we are poor at detecting changes in real-world scenes. Importantly,

however, we are more likely to detect changes in portions of a scene that are important for a given task. This means that when driving, for example, we may be poor at detecting changes in advertising billboards (which we hope have relatively low importance) but we are better at detecting task-relevant changes such as the appearance of a cyclist on the road (Velichkovsky et al., 2002). We are, moreover, capable of evaluating the complexity of scenes and modifying the level to which we intentionally monitor a scene for changes (Smilek et al., 2008).

Change blindness is one of a group of phenomena in which the effects of attention influence our ability to detect events. In 'inattention blindness,' subjects voluntarily focus their attention on a demanding task. Large changes in the scene can go unnoticed even when those changes occur at the point of fixation (Mack and Rock, 1998). One of the most celebrated demonstration of this is reported in Simons and Chabris' (1999) paper "Gorillas in our midst." Observers are asked to keep track of the number of times a basketball is passed between one group of players, dressed in white, and to ignore passes between player dressed in black. In the middle of the movie a man dressed in a gorilla suit walks across the scene in full (and beats his chest). Remarkably a large proportion of observers fail to notice the unattended gorilla. Variants of this task (Drew et al., 2013) have shown that radiologists when searching a CT-scan for one particular abnormality could miss other unexpected features in the scan (in this case, an image of a gorilla).

Our ability to detect stimuli is also, somewhat paradoxically, reduced just after we have attended to a stimulus. In a paradigm known as the 'attentional blink' subjects are shown a sequence of stimuli, typically letters, presented rapidly one after another and asked to respond when a specific target appears (Raymond et al., 1992). They are then asked whether a second type of stimulus appeared in the sequence at any point following the appearance of the target. Subjects' ability to detect this second stimulus is greatly reduced if it appears in the half-second or so immediately following presentation of the target letter. It is suggested that once attention has been attracted to the target there is a period immediately following during which it cannot be reallocated to the remaining of the stimulus stream.

The common theme to all of these phenomena is that unattended stimuli are poorly detected. This may lead one to conclude that attention is necessary for the conscious detection of a stimulus. This is an especially difficult hypothesis to test as to falsify it one must show that stimuli can be consciously perceived in the absence of attention. Measures of attention are always relative. We might, for example, show than a cued stimulus is processed more effectively that an uncued one. We cannot, however, conclude that the uncued stimulus is completely unattended. Li et al. (2002) conducted an experiment in which subjects focused on a primary task but occasionally had to report properties of probe stimuli presented in peripheral vision. They titrated the difficulty of the central task until performance in the probe task was at a minimum. Their rationale was that attention must now be maximally focused on the central task. They then showed that subject could consciously make complex categorizations of images shown in the periphery (e.g., report whether a scene contains a vehicle). The results are interpreted as showing that attention is not strictly necessary for conscious detection of stimulus properties. Note that, in contrast to most change blindness experiments, the peripheral stimuli in Li's study are important to the subject even though they are unattended.

In conclusion, change blindness and related phenomena demonstrate the importance of attention in recovering a rich representation of objects within a scene. Information from objects that are no longer attended may remain in visual memory but is not routinely accessed. Attention gates access to the information required to detect changes in those objects.

See also: Attention, Neural Basis of; Consciousness, Neural Basis of; Environmental Cognition, Perception, and Attitudes; Eye Movement, Control of: Oculomotor Control; Eyewitness Testimony; Facial Perception; Implicit Memory; Perception of Action Space: Using Multiple Frames of Reference; Person Perception, Accuracy of; Recognition; Retrieval; Short Term Memories, Theories of; Split Brains and Human Consciousness; Time Perception Models; Visual Imagery; Visual Memory, Psychology of; Visual Perception, Neural Basis of.

Bibliography

Burr, D.C., Morrone, M.C., Ross, J., 1994. Selective suppression of the magnocellular visual pathway during saccadic eye movements. Nature 371, 511–513.

Cole, G.G., Kentridge, R.W., Heywood, C.A., 2004. Visual salience in the change detection paradigm: the special role of object onset. Journal of Experimental Psychology: Human Perception & Performance 30, 464–477.

Deubel, H., Schneider, W.X., Bridgeman, B., 1996. Postsaccadic target blanking prevents saccadic suppression of image displacement. Vision Research 36, 985–996.

Drew, T., Võ, M.L.-H., Wolfe, J.M., 2013. The invisible gorilla strikes again: sustained inattentional blindness in expert observers. Psychological Science 24, 1848–1853.

Fernandez-Duque, D., Thornton, I.M., 2000. Change detection without awareness: do explicit reports underestimate the representation of change in the visual system? Visual Cognition 7, 323–344.

Grimes, J., 1996. On the failure to detect changes in scenes across saccades. In: Akins, K. (Ed.), Perception (Vancouver Studies in Cognitive Science), vol. 2. Oxford University Press, New York, pp. 89–110.

Hollingworth, A., 2003. Failures of retrieval and comparison constrain change detection in natural scenes. Journal of Experimental Psychology: Human Perception and Performance 29, 388–403.

Hollingworth, A., 2004. Constructing visual representations of natural scenes: the roles of short- and long-term visual memory. Journal of Experimental Psychology: Human Perception and Performance 30, 519–537.

Hollingworth, A., 2006. Visual memory for natural scenes: evidence from change detection and visual search. Visual Cognition 14, 781–807.

Laloyaux, C., Devue, C., Doyen, S., David, E., Cleeremans, A., 2008. Undetected changes in visible stimuli influence subsequent decisions. Consciousness and Cognition 17, 646–656.

Li, F.F., VanRullen, R., Koch, C., Perona, P., 2002. Natural scene categorization in the near absence of attention. Proceeding of the National Academy of Science of the United States of America 99 (14), 9596–9601.

McConkie, G.W., Zola, D., 1979. Is visual information integrated across successive fixations in reading? Perception & Psychophysics 25, 221–224.

Mitroff, S.R., Simons, D.J., Levin, D.T., 2004. Nothing compares two views: change blindness can occur despite preserved access to the changed information. Perception & Psychophysics 66, 1268–1281.

Mack, A., Rock, I., 1998. Inattentional Blindness. MIT Press, Cambridge, MA.

O'Regan, J.K., Rensink, R.A., Clark, J.J., 1999. Blindness to scene changes caused by "mudsplashes". Nature 398, 34.

Rensink, R.A., O'Regan, J.K., Clark, J.J., 1997. To see or not to see: the need for attention to perceive changes in scenes. Psychological Science 8, 368–373.

Rensink, R.A., 2000. Seeing, sensing, and scrutinizing. Vision Research 40, 1469–1487.

Raymond, J.E., Shapiro, K.L., Arnell, K.M., 1992. Temporary suppression of visual processing in an RSVP task: an attentional blink? Journal of Experimental Psychology: Human Perception & Performance 18, 849–860.

Simons, D.J., Franconeri, S.L., Reimer, R.L., 2000. Change blindness in the absence of a visual disruption. Perception 29, 1143–1154.

Simons, D.J., Levin, D.T., 1998. Failure to detect changes to people during a real-world interaction. Psychonomic Bulletin & Review 5, 644–649.

Smilek, D., Eastwood, J.D., Reynolds, M.G., Kingstone, A., 2008. Metacognition and change detection: do lab and life really converge? Consciousness and Cognition 17, 1056–1061.

Simons, D.J., Chabris, C.F., 1999. Gorillas in our midst: sustained inattentional blindness for dynamic events. Perception 28, 1059–1074.

Velichkovsky, B.M., Dornhoefer, S.M., Kopf, M., Helmert, J., Joos, M., 2002. Change detection and occlusion modes in road-traffic scenarios. Transportation Research Part F: Traffic Psychology and Behaviour 5, 99–109.

Charisma and Charismatic

Bryan S Turner, The Graduate Center, CUNY, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Charisma is associated with the sociology of Max Weber, especially his typology of authority and legitimacy. Weber developed his famous contrast between tradition, charisma, and legal-rational authority. Charisma was seen to be a revolutionary form of authority, and charismatic religious leaders claimed authority on the basis of an extraordinary dispensation. With the death of the charismatic leader, followers either abandoned the movement or it became more traditional and eventually even legal-rational. Weber called this process the 'routinization of charisma.' There is much debate as to whether charisma still operates in modern society, and whether the theory can be generalized to Asian religions.

'Charisma' or 'gift of grace' is a theological notion that has been widely used in the social and religious sciences to either describe the hierarchical organization of religious roles or explain the growth and development of social movements based on religious inspiration or the basis of authority and leadership in society generally. In its strictly religious context, it means a divinely conferred power, being derived from the Greek *kharisma* (*kharis*, favor or grace). Charismatic power is associated with the idea of the sacred as a force in human affairs. A person in possession of charisma is thought to have a talent, for example, in terms of healing or prophecy. In anthropological research, there has been considerable interest in 'shamanism' as a form of charismatic authority that depends on a capacity to have visions and to perform healing (see Eliade, 1964). In sociology, it is conceptually part of an analytical framework that is concerned with understanding large-scale changes in religious institutions and the foundations of authority.

The Sociology of Charisma

In the sociology of religion, the study of charisma has been closely associated with Max Weber (1864–1920), who adopted the idea from the historical and theological research of Rudolf Sohm and Karl Holl, who in turn had developed the concept in their analysis of Canon law. Weber wrote that "'the concept of charisma' (gift of grace) is taken from the vocabulary of early Christianity. For the Christian hierocracy Rudolf Sohm, in his *Kirchenrecht*, was the first to clarify the substance of the concept, even though he did not use the same terminology. Others (for instance, Holl in *Enthusiasmus und Bussgewalt*) have clarified certain important consequences of it. It is thus nothing new" (Weber, 1978, pp. 1, 216). Despite Weber's modesty about the adoption of the idea, it became a fundamental dimension of his analysis of authority and had far-reaching consequences for the development of the sociology of religious institutions. Weber generalized the idea and grasped its radical implications for the study of social and political change in human societies.

In the rise of religion, certain individuals are recognized as having a capacity to experience ecstatic states that are regarded as the precondition for healing, telepathy, and divination. It is primarily through "these extraordinary powers that have been designated by such special terms as 'mana,' 'orenda,' and the Iranian 'Maga' (the term from which our word 'magic' is derived). We shall henceforth employ the term 'charisma' for such extraordinary powers" (Weber, 1963, p. 2). Such charismatic power either is inherited as a natural endowment or is acquired by extraordinary means. The religious talent in both cases is a substance that may remain dormant in a person until it is aroused by asceticism or trance.

Weber did not dwell on these features of charisma in folk religion because he wanted to employ the term to understand the secular dynamic of authority and leadership in social institutions. His main intention was to compare and contrast three types of authority: charismatic, traditional, and legal-rational. As we have seen, charismatic authority rests on the ability of a leader to inspire disciples in the belief of the authenticity of a calling. In practical terms, an authentic claim is validated by a talent such as healing, but this alone cannot be the basis of authority. In the case of genuine charisma, a follower has a duty to accept the authority of a leader. In *Economy and Society* (Weber, 1978, pp. 1, 241), the term charisma is applied to a certain quality of an individual personality by virtue of which he is considered extraordinary and treated as endowed with supernatural, superhuman, or at least exceptional powers or qualities. Traditional authority involves the acceptance of a rule that expresses a custom, namely, an established pattern of belief or practice. Finally, legal-rational authority is typical of bureaucracies in which formal rules of conduct are underpinned by procedural norms. These forms of authority are in turn forms of compliance. Tradition depends on compliance through empathy, legal-rational authority rests on rational argument, and charismatic authority and leadership require inspiration. Weber provided many diverse illustrations of charismatic leaders, including Jesus Christ, Mohammed, Napoleon, Stefan Georg, and the Chinese emperor. Although the cases were heterogeneous, he argued that charismatic authority is confronted by a common problem of succession with the death of the leader. Charismatic authority is thus unstable. With the demise of the charismatic leader, the disciples typically disband, but occasionally a solution for continuity will be developed. In the case of the Christian Church, Weber argued that the charismatic authority of Christ was invested in the Church itself (as the body of Christ), and thus the bishops who control the 'keys of grace'

 http://dx.doi.org/10.1016/B978-0-08-097086-8.84003-9

enjoy a vicarious authority. This 'institutionalization of charisma' becomes over time increasingly formal, bureaucratic, and impersonal. As the charismatic power of Christ becomes transformed into a set of formal procedures and bureaucratic rules, Weber spoke of the 'routinization of charisma.'

Forms of Religious Association

Weber's account of charisma was also an important component of his sociology of religion within which he attempted to identify different religious roles and patterns of organization. For example, in *Ancient Judaism* he distinguished between the prophet who, as a charismatic figure, has a personal calling, and the priest who has authority by virtue of his service in a sacred tradition (Weber, 1952). The prophets, who often emerge from the ranks of the priesthood, are unremunerated and depend on gifts from followers. Weber also distinguished two forms of prophecy as represented, on the one hand, by Buddha and, on the other, by Zoroaster and Muhammad. The latter are involved in 'ethical prophecy' and are conceived as instruments of God. These prophets receive a commission from God to preach a revelation and demand obedience from their disciples as an ethical duty. By contrast, exemplary prophets demonstrate to their followers a salvational path through the example provided by their own lives. Exemplary prophecy was, according to Weber, characteristic of Asia; ethical prophecy, of the Abrahamic religions of the Middle East. Weber's analysis of charisma with respect to ethical prophecy in the Old Testament has been subject to considerable criticism (see Zeitlin, 1984), but his conceptual framework continues to influence both sociology (see Lindholm, 1990) and anthropology (see Werbner and Basu, 1998).

This discussion of charisma with respect to different social roles should be seen as part of a larger sociological debate about the forms of association that characterize the social organization of religious belief and practice. Weber wanted to argue that any group that is subject to charismatic authority forms a charismatic community (*Gemeinde*) and that such a community is inherently unstable. With the death of the leader, the group either dissolves or charisma undergoes a process of routinization. The disciples have no career, formal hierarchy, offices, or qualifications. The church that provides the organizational context of the priesthood is very different. Ecclesiastical organizations require a hierarchical administration of the 'charisma of office' in which there are definite stages in clerical and administrative careers. It is clearly the case that Weber's sociology of charisma should be understood as somewhat parallel to Ernest Troeltsch's 'church-sect typology' in which there is a historical oscillation between the evangelical sects and the bureaucratic churches (see Troeltsch, 1931).

Charismatic Movements

Because Weber believed that in modern societies legal-rational authority would become dominant, tradition and charisma were regarded as early forms of authority and thus as characteristic of premodern societies. The decline of charisma is thus associated with the idea of secularization. The notion that charismatic authority was not a resilient aspect of modern society was in turn a function of Weber's pessimistic understanding of social change in terms of secular rationalism and the erosion of religious meaning. Against Weber's assessment, the late nineteenth and twentieth centuries were profoundly influenced by charismatic movements. Charismatic renewal has been a common theme of diverse religious movements in 'primal societies' and in the industrial societies of Europe and North America. The collapse of aboriginal or tribal societies under colonial settlement saw the spread of charismatic movements against the supremacy of White-settler societies such as the Ghost Dance among the Cheyenne and Sioux tribes of the American Plains. A Paiute prophet called Wovoka had received a vision in which through ritual dance the dead would return to restore the pristine culture of native societies. This anti-White charismatic movement subsided after the murder of Sitting Bull and the subsequent massacre of his followers at Wounded Knee in 1890. Charismatic leadership has also played a significant role in those new religious movements that have been a response to the social and economic disruptions associated with the decolonization of the Third World (see P. Worsley, The Trumpet Shall Sound, 1970).

In contexts of rapid social change, charismatic leadership is an important component of so-called revitalization movements that function to create an effective transition from tribal to urban society. However, charisma is also associated with religious forms that are a response to personal alienation, isolation, and meaninglessness in the developed, industrial world. One illustration is the Family, whose members were disciples of Charles Manson and operated in southern California in the late 1960s. Manson, who was a social dropout with a criminal record, provided a message of personal liberation based on mind-expanding drugs and found a receptive set of disciples in the cultic milieux of the American counterculture. The Manson Family offered disoriented disciples the experience of an absolute community that depended on a powerful ideology and indoctrination (see Schreck, 1988). Another American illustration would be Jim Jones and the People's Temple (see Weightman, 1983). These manifestations of charisma in modern society are closely associated with the dislocations of youth culture, the anomie of the modern city, and the spread of personal alienation. Charismatic figures can also enjoy a global following with modern communication and international travel. For example, Sathya Sai Baba (1926–2011), who was born in Puttaparthi, a remote Indian village, gained a global following from a range of different religious traditions and built numerous religious centers around the world. He claimed to be an incarnation of the Indian saint and Muslim faqir Shirdi Sai Baba, announced that he was the 'Cosmic Christ,' offered healing to the sick, and showered his disciples with gifts (see Srinivas, 2010).

Charisma is not exclusively of interest to sociologists of religion. It plays an important role in political sociology. Commenting on Weber's discussion of charisma in the famous 'Politics as a vocation' (Weber, 2009) – a lecture from Munich in 1919 – Peter Baehr (2008: 90) suggests that Weber implied a tripartite division between religious, military, and political dimensions. What Weber called the city-state demagogue and the parliamentary party leader were unique to the West. British Prime Minister William Gladstone (1809–98) was in Weber's

eyes a grand demagogue whose charismatic powers wooed the masses, allowing him to become a plebiscitary dictator.

Conclusion: Charisma in the Information Age

Weber's analysis of charisma has produced a rich legacy of sociological research, but there are serious conceptual problems associated with its application in contemporary society. The concept of the charismatic movement is used to describe a variety of revivalist, nativist, messianic, and healing movements. There is considerable debate in sociology and religious studies about whether charisma in modern society is degraded and inauthentic (see Wilson, 1975). The term is frequently employed to describe the fleeting popularity of political leaders whose social appeal is constructed by campaigns in the media. Whereas charisma in traditional societies arises spontaneously in the collective enthusiasm for divinely inspired leaders, in modern politics it is inevitably the product of deliberate orchestration of the media. The notoriety of the Manson Family and the People's Temple was a conscious product of media attention. However, this argument indicates a persistent issue in charismatic movements, namely, how to distinguish between genuine and false claims to religious authority. Biblical warnings against 'false prophets' recognized the fact that disciples could be misguided. While the term has been implausibly stretched to describe a diverse and heterogeneous range of leaders in the twentieth century from Adolf Hitler to John F. Kennedy, it remains a basic concept in sociology and in religious studies.

See also: Charisma, Social Aspects of; Millennialism; New Religious Movements; Prophetism; Religion, Sociology of; Weber, Max (1864–1920).

Bibliography

Baehr, P., 2008. Caesarism, Charisma and Fate. Historical Sources and Resonances in the Work of Max Weber. Transaction Publishers, New Brunswick and London.
Eliade, M., 1964. Shamanism. Archaic Techniques of Ecstasy. Ballingen Foundation, New York.
Lindholm, C., 1990. Charisma. Blackwell, Oxford, UK.
Schreck, N. (Ed.), 1988. The Manson File. Amok Press, New York.
Srinivas, T., 2010. Winged Faith. Rethinking Globalization and Religious Pluralism through the Sathya Sai Movement. Columbia University Press, New York.
Troeltsch, E., 1931. The Social Teaching of the Christian Churches, vols. 2. Allen and Unwin, London.
Weber, M., 1952. Ancient Judaism. Free Press, Glencoe.
Weber, M., 1963. The Sociology of Religion. Methuen, London.
Weber, M., 1978. Economy and Society. In: An Outline of Interpretive Sociology, vols. 2. University of California Press, Berkeley, CA.
Weber, M., 2009. Politics as a vocation. From Max Weber. Routledge, London. 77–128.
Weightman, J.M., 1983. Making Sense of the Jonestown Suicides. A Sociological History of the People's Temple. Mellen Press, New York.
Werbner, P., Basu, H. (Eds.), 1998. Embodying Charisma Modernity Locality and the Performance of Emotion in Sufi Cults. Routledge, London and New York.
Wilson, B.R., 1975. The Noble Savages. The Primitive Origins of Charisma and Its Contemporary Survival. University of California Press, Berkeley, CA.
Worsley, P., 1970. The Trumpet Shall Sound. A Study of Cargo Cults in Melanesia. Shocken Books, New York.
Zeitlin, I.M., 1984. Ancient Judaism: Biblical Criticism from Max Weber to the Present. Polity Press, Cambridge, UK.

Charisma, Social Aspects of

Christopher Adair-Toteff, University of South Florida, FL, USA; and University of Kent, Kent, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by S. Abbruzzese, volume 3, pp 1653–1659, © 2001, Elsevier Ltd.

Abstract

Charisma is the extraordinary and revolutionary power that is given to the one chosen to lead. Charismatic leadership may be religious, political, artistic, or social, and it is often temporary: the person's charismatic power lasts only as long as he or she is able to perform extraordinary tasks. When the person is regarded as no longer being successful, the charismatic power is considered to have been withdrawn. The power of charisma has long been recognized, but its study is relatively recent. The concept of charisma has been broadened extensively in the last decade or two, with the result that it has lost some of its traditional characteristics.

Charisma is the extraordinary power that is supernaturally given to a specifically chosen person. The charismatic individual was a magician, a prophet, a shaman, or a particular kind of religious leader. The charismatic person was recognized as being special by virtue of the special powers that were given to him or her. These powers were highly unusual, and were often transitory. Charisma seemed to have been endowed in an individual for a short time and then withdrawn rapidly. Charisma was an unstable power and seemed resistant to permanence and transference. Its instability, its intensity, and its unusualness contributed to its allure.

Much of this article will focus on Max Weber because, as Stephen Turner has written, "Max Weber is the original and sole source of the social science concept of charisma (as distinct from the religious concept)… ." (Turner, 2011: p. 82).

Charisma, Origins of

Max Weber was the one who introduced the notion of charisma into sociological discourse, and he paved the way for expanding its use into the general public, but he was not the originator of the term. Weber was not always concerned with identifying the sources for his ideas, and when he did, it is important to examine them for they frequently shed considerable light on the issue that he addresses. This is particularly true regarding the notion of charisma, because his theological sources do not simply deal with charisma as a religious concept, but indicate qualities that become important for Weber's own social and political notion of charismatic authority.

As a religious concept, charisma does not have a long history. It does not have an entry in either the first (1909) or the second edition (1927) of *Religion in Geschichte und Gegenwart*, nor in the third edition (1899) of the *Realencyklopädie für protestantische Theologie und Kirche*. Instead, one is directed to the entry 'Geistesgaben.' or 'spiritual gifts.' The reader is referred to I Corinthians 12, where 'spiritual gifts' are briefly mentioned. What is important is that these are miraculous gifts, and they come directly from God. They can be utilized for the benefit of the receiver or, more generally, for the receiver's community (Cremer, 1899: pp. 460–463). Weber's two acknowledged sources represent each of these two points of view; Karl Holl's focus on charisma is in relation to the soul of the receiver, and Rudolf Sohm's is on its importance for the community.

In the first part of *Wirtschaft und Gesellschaft*, Weber explicitly mentions Karl Holl and Rudolf Sohm in connection with the concept of "'Charisma' ('Gnadengabe')" (Weber, 1922: p. 124). Weber's reference to Holl occurs only once, and it is to his early work *Enthusiasmus und Bussgewalt*. As the subtitle makes clear, Holl's book is devoted to Greek monks and specifically to Symeon, a Greek monk living about 1000 AD. Symeon was an ascetic, and he considered ascetic suffering to be a type of cleansing that then resulted in a closer relation to God. For Symeon, the world of the senses was unimportant; the real world was invisible and was filled with devils and spirits. Symeon believed that charisma was a gift from God and that its spiritual effects helped him in his struggle with the demons (Holl, 1898: pp. 143, 148–150). It was a 'spiritual gift' that allowed the monk to concentrate on what was most important to him – his own spiritual salvation (Holl, 1898: p. 153).

The second, and more important, source for Weber's conception of charisma was the Church historian and legal scholar Rudolf Sohm (Kroll, 2001: p. 51). Sohm is more important because Weber cites him often. Weber makes reference to his notion of charisma in the early part of *Wirtschaft und Gesellschaft*, but also in the main section on charisma (Weber, 2005: p. 462). Furthermore, he specifically cites Sohm's notion of charisma as indicated in the account of his Vienna lecture "Probleme der Staatssoziologie" (Weber, 2005: p. 755). Finally, he explicitly mentions Sohm and the treatment of charisma in his lecture course "Allgemeine Staatslehre und Politik (Staatssoziologie)" as well as in the posthumously published "Die drei reinen Typen der legitimen Herrschaft" (Weber, 2009: pp. 77–79 and Weber, 2005: p. 735). Second, Weber spells out his debt to Sohm: Sohm was the first scholar to investigate the notion of charisma from a "scientific view point," and Sohm's treatment of it was "brilliant" (Weber, 2005: p. 735). Third, Sohm not only treated the concept from a "purely historical" perspective but also indicated its "sociological specialty of this category of power structures" (Weber, 2005: p. 462). Sohm intended to indicate the importance the concept played in the early Church and thereby show the fundamental difference between the early believers and the later Catholic Church. Holl's

discussion of eleventh-century Greek monks and Sohm's investigation of charisma in the early Church may not seem to have much in common; however, both noted that the revolutionary power of charisma was used against institutions. That is why Holl writes of the opposition between "office and spirit" and why Sohm describes the contradiction between the free world of the spirit and the dominating Church law (Holl, 1898: p. III and Sohm, 1892: p. X). Sohm explains that 'ecclesia' does not mean Church, but it does mean "God's people" and he distinguishes between the church's structure and the real notion of church. The real church is a gathering of believers – where "two or three come together" (Sohm, 1892: pp. 19–20). Charisma is a God-given gift, and it requires recognition. However, there can be no talk of 'legal obligation'; rather, there is only the 'the love of the highest.' Charismatic power requires 'free recognition' and 'free obedience'; thus, it stands in direct opposition to the 'demands of the law' (Sohm, 1892: p. 27). Sohm's pro-charisma and antilegal argument about the early Church provoked a lengthy and heated discussion with the noted Church historian Adolf Harnack. Thomas Kroll maintains that this debate prompted Weber to think more fully about constructing his typology of domination (Kroll, 2001; Weber, 2005: pp. 37–41). Whether this is true is not important here; what is important is that Karl Holl and Rudolf Sohm both regarded the notion of charisma as being fundamentally important with its opposition between spirit and organization, not just for theology but also for society. In his introduction to the English translation of *Wirtschaft und Gesellschaft*, Guenther Roth maintained, "He [Weber] secularized Rudolf Sohm's notion of the charisma of the Christian Church" (Weber, 1978: p. XCVI). It was Weber who introduced it to a wider audience.

Weber's Concept of Charisma

Weber frequently discussed concepts by setting out their opposites, and his discussion of charisma follows this pattern. If anything, Weber believed that charisma could be understood only by contrasting it with very different but related concepts (see Adair-Toteff, 2005). There are three such concepts, all of which fall under the category of Herrschaft. The term 'Herrschaft' is resistant to translation, but it incorporates 'domination,' 'rule,' 'authority,' and 'leadership.' Weber defines Herrschaft as the likelihood that one will obey a specific order given by a specific person (Weber, 1922: pp. 28, 122; Weber, 2005: pp. 135, 139, 727; Weber, 2009: pp. 74, 75). Weber acknowledges that people will obey an order for a variety of reasons: out of self-interest, out of simple convention, or even out of sheer habit (Weber, 1922: p. 122; Weber, 2005: p. 726). However, Weber's concern is with those who obey because they acknowledge the validity of that authority, hence his concern is with 'legitimate' Herrschaft. In his view, there are only three 'pure' forms of legitimate Herrschaft, and he readily acknowledges that they are rarely found in reality. Because they are rarely found in reality, Weber is convinced that it is even more imperative to offer "clear concepts" of these types (Weber, 1922: pp. 123–124; Weber, 1992: pp. 160–161; Weber, 2005: p. 726).

The first type of legitimate Herrschaft is 'legal,' and it has a significant number of features. It is based upon rational rules, and the person who interprets and implements these rules is fully trained; the person has had a rigorous education and has proven his or her expertise by passing exams. These are people who then hold an office and are part of a hierarchy. The officials who make the determinations must follow objective rules, so these determinations are never arbitrary. The determinations are totally independent of the person and are neither made for or against that individual. As Weber repeatedly writes, these decisions are made "without hate or love" ("sine ira et studio") (Weber, 1922: p. 129; Weber, 2001: pp. 400–401, 429; Weber, 2005: pp. 187, 727). The rules are always in continuous effect; as a result, legal Herrschaft is both stable and permanent (Weber, 1922: pp. 125, 127; Weber, 2005: pp. 726–727). These are all important components of legal Herrschaft, but the most defining one is the notion of competency. Weber emphasizes this in virtually every single account of legal Herrschaft (Weber, 1922: p. 125; Weber, 1992: pp. 160–161; Weber, 2005: pp. 136–137, 726; Weber, 2009: pp. 78–80). The second type is 'traditional' (or even 'conventional'), and it shares a number of features with legal Herrschaft (Weber, 1922: p. 123). It is also stable and permanent; Weber speaks of the "eternal laws" (Weber, 1992: p. 160). Weber grants that the traditional Herrschaft is personal, but he also stresses that he is mostly restricted by tradition (Weber, 2009: pp. 76–77, 82). The patriarch is able to introduce new laws as long as they do not conflict with the 'holiness' of the long-established order and he also has enough respect from the others (Weber, 1922: p. 133; Weber, 2005: pp. 149, 729). While not as strictly formal as legal authority, traditional authority tends not to be arbitrary, and the decisions are based on more than simply who the person is. But Weber stresses that traditional Herrschaft lacks the objective and dispassionate competency of the trained official, and it lacks clear and rational regulations (Weber, 1922: pp. 131–132).

Weber spends a considerable amount of effort explaining what charisma is not: it has no official, it has no training, and it has no hierarchy. It has no rules, it has no competency, and it has no legal principle. It is not impersonal, it is not permanent, and it is not stable. These are all important differences, but the single most important difference between charismatic and legal and traditional Herrschaft is that it is extraordinarily unusual. Weber insists that both traditional Herrschaft and legal Herrschaft are "specifically *ordinary* (*everyday*) (*Alltags-*) forms of 'Herrschaft.'" Traditional Herrschaft has existed forever, and while legal Herrschaft is a relatively modern phenomenon, it is found everyday in many parts of the Western world. In marked contrast, charismatic Herrschaft is extremely rare, or, as Weber says – it is always "unusually extraordinary" (*Außeralltaglich*). Weber makes this point or some variation of it repeatedly (Weber, 1922: pp. 140–1411; Weber, 1989: pp. 99, 175; Weber, 2001: pp. 315, 355; Weber, 2005: pp. 739, 740; Weber, 2009: pp. 90, 91; see Shils, 1965: p. 199; Bendix, 1977: p. 299; and Adair-Toteff, 2002: p. 194). However, the phrase 'unusually extraordinary' can be applied to a number of things: a person's height or weight, an animal's speed or strength, or a storm's range and force. In contrast, charisma also has the power to awe. That is why Weber believes that charisma is such a revolutionary power; he called it the "specifically 'creative' revolutionary power in history" (Weber, 2005: pp. 468, 482; Riesebrodt, 2001: p. 151). Late in his

life, Weber insisted that there were only two "revolutionary" powers that were historically important: rationality, which "comes from above" and charisma, which "comes from within" (Weber, 2009: pp. 92–95).

The charismatic person convinces others of his/her unique power by performing magical feats, by offering revelations, or by doing heroic acts (Weber, 1922: pp. 140, 141; Weber, 2005: p. 737). The individual is regarded as charismatic only as long as he/she is successful; as soon as there is a failure, there is a risk that the others will withdraw their recognition and stop obeying instructions and orders. The warrior who begins to lose battles, the magician who can no longer conjure tricks, and the political leader whose words have lost power are some examples of the person who has lost his/her charismatic power. Charisma is predicated upon the recognition of the person's charismatic powers and that this recognition is subject to change; it is highly personal and very temporary.

Weber contrasts the uniqueness and instability of charisma with the ordinariness and regularity of everyday life. Charisma is opposed to both traditional and bureaucratic rationalism (Weber, 2005: p. 542). Economic, familial, and societal concerns are alien to the charismatic leader; money, family, and order are not important to him/her (Weber, 2005: p. 484; Weber, 2009: p. 93). The demands of everyday living are antithetical to the extraordinariness of charisma, and they serve mostly to compromise its power. Every compromise with the everyday order alters the character of charisma in a very negative manner (Weber, 2009: pp. 94, 95). In a passage remarkable for its intensity, Weber describes how charisma's interaction with the material needs of the everyday world diminishes and dooms it: "Each charisma finds itself on this way from a stormy-emotional economic-alien life to a slowly suffocating death under the weight of material interests in each hour of its life and indeed with each growing hour in increasing measure" (Weber, 2005: pp. 488, 489).

Charisma – Transference and Institutions

The instability and lack of permanence in charisma pose the great question about transference: can it be passed on to another individual, or can it be institutionalized? Weber recognized that charisma was broken in its conflict with the demands of everyday life, and it was transformed into "churches, sects, academies, schools" (Weber, 2005: p. 489). Weber cites the example of Jesus appointing Peter as his legitimate 'successor' and directing him to found his 'church' (Weber, 2005: p. 495). Weber notes that the most common way of ensuring that the power of charisma can be passed on is by legitimizing it through the "bond of blood" (Weber, 2005: pp. 517, 740). Weber sometimes refers to this as "Erbcharisma der Familie" (Weber, 2005: pp. 522, 525). Another way of transferring charismatic power is through "Amtscharisma," which Weber defines as the specific conveying of the power through a social institution. While this is predominantly a church, it need not be restricted to just that (Weber, 2005: p. 527). However, Weber points to the Catholic Church as the clearest example of "Amtscharisma" and notes how it is based upon the "authority willed by God" (Weber, 2005: p. 529). Weber contrasts the spontaneity of charisma with discipline. He points to military discipline but also to the economic discipline needed to run a large business (Weber, 2005: p. 556). There is a fundamental tension between the revolutionary power of charisma and everyday routine living (Roth and Schluchter, 1979: p. 37).

There are some scholars who argue that charisma and order and institutions are not diametrically opposed. In the introduction to his *Max Weber on Charisma and Institution Building*, S.N. Eisenstadt admitted that Weber himself stressed the "dichotomy" between charisma and social organization, but he emphasized that Weber actually believed that the two were fundamentally interrelated (Eisenstadt, 1968: p. ix). Eisenstadt points to two types of creativity – one in which an idea is nurtured for a long time and finally comes to fruition, and the other, which seems to be close to being spontaneous. However, he insists that both can come about only within a permanent social framework (Eisenstadt, 1968: pp. xvi, xvii). Eisenstadt notes that the charismatic leader creates in freedom but leaves an impact on the institution through routinization. He suggests that this process is cyclical – charismatic power rises up against the ordinary, becomes routine, and then rises again (Eisenstadt, 1968: p. xxi). In "Charisma, Order and Status" Edward Shils also points to the social need for order and looks to Weber's own examples for transferring charisma (Shils, 1965: pp. 202, 203). However, Eisenstadt appears to underscore the human need for order found in institutions and to overlook the uniqueness and awe-inducing power that are the marks of charisma.

Religious and Political Charismatic Leaders

Charismatic leaders are both very important and rather rare, so it is instructive to note those who Weber specifically identifies. Sometimes Weber simply names them, while other times he spends some time discussing them. He orders them into three types: artistic charisma, religious charisma, and political charisma.

Weber does not devote much time to discussing artistic charisma, but he identifies Stefan George as a poet who possessed enormous charisma (Weber, 1922: p. 142; Weber, 2005: pp. 40, 465; Weber, 2009: p. 93). Stefan George was a firm believer in aesthetics and promoted an "arts for art's sake" philosophy. Before 1900, George had come to be regarded as a modern-day prophet, and this group of followers was referred to as the "Stefan George Circle." Weber appreciated George's separation of art from money, and he respected George enough as a translator to use his rendition of Shakespeare; however, he was highly critical of George's cult-like following. George was charismatic in that he had a forceful personality and a magical way with words. In Weber's opinion, George was too much of a proponent of irrational romanticism, which led his followers to reject sober counsel and instead to seek out extreme emotional experiences (Weber, 1992: p. 34; Weber, 2005: p. 465).

Weber devotes more time to religious charisma, and he identifies several figures from Christianity. Weber draws attention to Jesus and his charismatic power (Weber, 2001: p. 355). Jesus is the one who personifies the rejection of all traditional rules and laws – "It is written but I say unto you." Instead, he introduces the "eternally new," and he gives "new commands"

(Weber, 1922: p. 141; Weber, 2005: pp. 468, 734; Weber, 2009: pp. 92, 93). He may be sent by God, but he is also forsaken by God (Weber, 2005: pp. 466, 734). Weber also singles out St Francis, but he does not explain why he thinks that St Francis is a charismatic figure. What he does suggest is that St Francis' belief in unconditional goodness was a special doctrine that drew people (Weber, 2001: pp. 318, 319). He also notes that St Francis' songs were extraordinarily powerful and thus had an almost mystical effect (Weber, 2001: p. 413). Christians are not the only religious charismatic leaders; Weber also identifies the Buddha and even the Dalai Lama (Weber, 1989: p. 521; Weber, 2005: p. 494).

Prophets, and especially Old Testament Prophets, are repeatedly mentioned by Weber as being charismatic (Weber, 1922: pp. 140–143; Weber, 1992: p. 160; Weber, 2001: pp. 177–180; Weber, 2005: pp. 483, 515, 594, 691, 734–737; Weber, 2009: pp. 90–93). Prophets are difficult to classify; while it seems intuitive to place them in the religious charisma category, Weber often accentuates their political involvement; hence, I place them in the political charisma category (Weber, 1992: p. 161). Like many of the charismatic leaders, the Old Testament Prophets believed that they had been 'called' and that they had been given a mission. Weber refers to this as a "prophetic mission" or a "Godly mission" (Weber, 2001: p. 203; Weber, 2005: pp. 462–464). Weber differentiates the prophet from the priest; the former has been personally called, while the latter has been professionally trained (Weber, 2001: p. 178). Weber does not deny that some priests have charismatic traits, but he emphasizes that these are personal and are not derived from the priestly office. Weber also differentiates the prophet from the magician; the latter's charisma comes through performing magic, while the former's is derived from God-given ethical commands (Weber, 2001: p. 178). Accordingly, the prophet demands obedience out of an ethical duty rather than from any type of law or tradition (Weber, 2001: p. 189). The Old Testament Prophets revealed divine instructions; thus, they were not involved in any political tradition.

Weber does identify several genuine political leaders as possessing charisma: Gladstone and Napoleon (Weber, 2009: pp. 90–93). But it was Pericles who seemed to be the most interesting of the charismatic political leaders, and that was because he combined the power of words with the power of deeds (Adair-Toteff, 2008: pp. 147, 155, 160). Weber reminded us that Pericles was a demagogue, but one who embodied the characteristics of a responsible political leader (Weber, 1992: p. 191; Weber, 2005: p. 736; Weber, 2009: pp. 102, 103).

Evil Charismatic Leaders

For the most part, all of Weber's examples of charismatic leaders are revolutionary, but they also appear to be forces for good. However, there are examples of charismatic leaders who were evil. Charles Lindholm is one scholar who spent considerable effort in discussing the charisma of evil leaders. He emphasized three, all of whom were from the twentieth century.

The name Charles Manson is not normally connected to the notion of charisma, but Lindholm makes this a compelling case. While Manson did not commit any murders himself, his charismatic power was so great that his 'family' members were more than willing to commit any crimes, including murder, that he ordered. There are several things that separate Manson from many other charismatic leaders, but one of the most pronounced was the way in which he shared the power with his disciples. Rather than maintaining distance in order to preserve a sense of awe, Manson seemed to relish drawing his disciples to him and sharing the secrets of his power (Lindholm, 1990: p. 133).

In contrast to Manson, Jim Jones, the charismatic founder of the "Peoples Temple," maintained his authority by being aloof and delegating responsibility to a very close inner circle. The inner circle carried out many of Jones' commands, but it also served to insulate him when there were mistakes. In this manner, Jones was able to maintain his sense of always being right, even when the facts seemed to show the opposite. Like Manson's followers, Jones' followers also gave up much in order to belong, but unlike the Manson 'family,' Jones' disciples were a much larger and more diverse group. They gave their possessions to Jones and followed him from place to place, ultimately to Guyana. While there were the usual middle-class white radicals, many of Jones' followers were impoverished black people. Moreover, the followers were not just young people, but older people were involved as well. As Lindholm points out, the followers came from three generations (Lindholm, 1990: p. 139). As Weber indicated, charismatic leaders care little about money; Jim Jones was totally indifferent to wealth. What he thrived on was power.

The most famous evil charismatic leader that Lindholm discusses is Adolf Hitler. Lindholm stresses that it was Hitler's great power as a propagandist and especially as an orator that brought him so much success (Lindholm, 1990: p. 99). Lindholm utilizes some of Weber's classifications: the focus on Hitler as a person, his personal force, and the religious fervor of his followers (Lindholm, 1990: pp. 100–107). Lindholm is not the only person to emphasize Hitler's charismatic power. There have been a number of recent books devoted specifically to Hitler's charisma: three examples are *Hitler's Charisma: Leading Millions into the Abyss*, *The Third Reich: Charisma and Community*, and *The Dark Charisma of Adolf Hitler*. The point in each is to help explain how and why so many people were not only willing but also even eager to believe in the need for mass killings and genocide.

To Lindholm's list, we can also add David Koresh and Osama bin Laden. The Branch Davidians were ready to die for their convictions, and bin Laden's followers were ready to kill for their beliefs. Both groups were convinced of the power of their charismatic leader and as a result were willing to follow whatever commands their leader gave. Whether good or bad, these charismatic leaders were regarded as possessing a highly unusual power; they were rarities in history. Today, however, charisma has lost that specialness and now is applied to the most wide and varied objects.

'Charisma' Today

The notion of charisma has now gained widespread currency; it is found in hundreds of recently published books (Turner, 2003). Most of those that have appeared between 2010 and 2013 can be divided into three categories. There are some

that are primarily religious in orientation and deal with the notion of charisma in the eternal fight between good and evil; however, the vast majority can be divided into two, overlapping categories. They are in the categories of leadership skills or self-improvement. Examples of the first include *Charisma-Based Leadership: How to Be the Leader That Everyone Wants to Follow*, *How to Be Charismatic: Learn from the World's Most Charismatic Leaders*, and *The Psychology of Human Leadership: How to Develop Charisma and Authority*. Examples of self-improvement are even more plentiful; they include *The Charisma Myth: How Anyone Can Master the Art and Science of Personal Magnetism*; *How to Light Up a Room: 55 Techniques to Help You Increase Your Charisma, Build Rapport, and Make People Like You*; *Art of Conversation: Enhance Your Charisma for Success*; *The Laws of Charisma: How to Captivate, Inspire, and Influence for Maximum Success*; and *The Charismatic Edge: The Science of Confidence, Captivating and Compelling Communication*. These titles indicate the many ways in which the notion of charisma has been altered and expanded since Weber's introduction of it. First and foremost, it is no longer considered a gift; instead, it is now regarded as something that can be acquired, something that can simply be learned. Furthermore, it is no longer that unusual; anyone can acquire charisma – as long as the individual is willing to buy the book that promises it. Third, it is no longer mysterious; instead, it is a science, an art, or even an amalgam of each. Fourth, the fact that charisma was regarded as a fleeting and transitory power has been replaced with the sense that once one has gotten it, one will always have it. This connects to the fifth point; originally, the person was regarded as possessing charisma only so long as that person was successful – performing miracles, winning battles, and the like. As soon as the person failed, the adherents believed that God or the gods had withdrawn that special power. Finally, charismatic power is no longer regarded as extraordinary and highly unusual; instead, the notion has been generalized to mean almost any type of attractiveness, popularity, and even charm (Wikipedia). Thus, the notion of charisma has been generalized; while it can still be applied to various Western and Eastern types of 'charismatic awakening' churches, it is now applied to almost anything. It has now been applied to cars, not just to special ones like the iconic Porsche 911 but also to the Mitsubishi Carisma, which was a play on 'car' and 'charisma.' Charisma is also a brand name for bed sheets, pillow cases, and even towels. It is also a brand name for dolls, toys, collectibles, and jewelry. It has been used as a name of a hair-styling salon. Finally, it is also a woman's name, which strangely enough seems to have been adopted by authors of semipornographic novels (Charisma Knight and Kelsey Charisma) and even an actress and Playmate (Charisma Carpenter). The unfortunate result of all of this is that, as Marta B. Calás (1993) has suggested, we have been "making 'charisma' as *ordinary* as possible"; thus, the idea of it has been so expanded that it is now far removed from Weber's (and Sohm's) original notion of charisma as an *extraordinary* and awesome power. Nonetheless, the concept of charisma continues to play an outsized role in modern society.

See also: Bureaucracy, Sociology of; Political Sociology; Secular Religions; Social Psychology, Sociological; Status and Role, Social Psychology of.

Bibliography

Adair-Toteff, Christopher, 2005. Max Weber's charisma. Journal of Classical Sociology. 5 (2), 189–204.

Adair-Toteff, Christopher, 2008. Max Weber's Pericles: the political demagogue. Max Weber Studies. 7 (2), 147–162.

Bendix, Reinhard, 1977. Max Weber. An Intellectual Portrait. University of California Press, Berkeley.

Calás, Marta B., 1993. Deconstructing charismatic leadership: re-reading Weber from the darker side. Leadership Quarterly. 4 (3/4), 305–328.

Cremer, Hermann, 1899. Geistesgaben, Charismata. Sechster Band. In: Hauck, Albert (Ed.), Realencyklopädie für protestantische Theologie und Kirche. J.C. Hinrichs'sche Buchhandlung, Leipzig, pp. 460–463.

Eisenstadt, S.N., 1968. Max Weber. On Charisma and institution building. In: Selected Papers, Edited and with an Introduction by S.N. Eisenstadt. University of Chicago Press, Chicago and London.

Holl, Karl, 1898. Enthusiasmus und Bussgewalt beim Griechischen Mönchtum. J.C. Hinrichs'sche Buchhandlung, Leipzig.

Kroll, Thomas, 2001. Max Webers Idealtypus der charismatischen Herrschaft und die zeitgenössenische Charisma-Debatte. In: Hanke, Edith, Mommsen, Wolfgang J. (Eds.), Max Webers Herrschaftssoziologie. J.C.B. Mohr (Paul Siebeck), Tübingen, pp. 47–72.

Lindholm, Charles, 1990. Charisma. Basil Blackwell, Oxford.

Riesebrodt, Martin, 2001. Charisma. In: Kippenberg, Hans G., Riesebrodt, Martin (Eds.), Max Webers "Religionssystematik". J.C.B. Mohr (Paul Siebeck), Tübingen, pp. 101–118.

Roth, Guenther, Schluchter, Wolfgang, 1979. Max Weber's Vision of History. University of California Press, Berkeley.

Shils, Edward, 1965. Charisma, order, and status. American Sociological Review 30 (1), 199–213.

Sohm, Rudolf, 1892. Kirchenrecht. Erster Band. Die Geschichtlichen Grundlagen. Verlag von Duncker & Humblot, Leipzig.

Turner, Stephen P., 2003. Charisma reconsidered. Journal of Classical Sociology. 3 (1), 5–26.

Turner, Stephen P., 2011. Classic sociology: Weber as an analyst of charisma. In: Harvey, Michael, Riggio, Ronald E. (Eds.), Leadership Studies. The Dialogue of Disciplines, pp. 82–88. Cheltenham, UK and Northampton, MA.

Weber, Max, 1922. Grundriss der Sozialökonomik. III. Abteilung. Wirtschaft und Gesellschaft. Verlag von J.C.B. Mohr (Paul Siebeck), Tübingen.

Weber, Max, 1978. In: Roth, Guenther, Wittich., Claus (Eds.), Economy and Society. University of California Press, Berkeley.

Weber, Max, 1989. In: Schmidt-Glintzer, Helwig, Kolonko, Petra (Eds.), Die Wirtschaftsethik der Weltreligionen. Konfuzianismus und Taoismus. J.C.B. Mohr (Paul Siebeck), Tübingen.

Weber, Max, 1992. In: Mommsen, Wolfgang J., Schluchter, Wolfgang, Morgenbrod, Birgitt (Eds.), Wissenschaft als Beruf/Politik als Beruf. J.C.B. Mohr (Paul Siebeck), Tübingen.

Weber, Max, 2001. Wirtschaft und Gesellschaft. In: Kippenberg, Hans G., Schilm, Petra, Niemeier, Jutta (Eds.), Religiöse Gemeinschaften, Teilband 2. J.C.B. Mohr (Paul Siebeck), Tübingen.

Weber, Max, 2005. Wirtschaft und Gesellschaft. In: Hanke, Edith, Kroll, Thomas (Eds.), Herrschaft, Teilband 4. J.C.B. Mohr (Paul Siebeck), Tübingen.

Weber, Max, 2009. In: Hübinger, Gangolf, Terwey, Andreas (Eds.), Allgemeine Staatslehre und Politik (Staatssoziologie). Unvollendet. Mit- und Nachschriften 1920. J.C.B. Mohr (Paul Siebeck), Tübingen.

Charity and Philanthropy, Economics of

James Andreoni, University of California, La Jolla, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Philanthropy is defined as a benevolent behavior toward others in the society, usually in the form of charitable gifts. Charitable giving accounts for a significant fraction of income in the US and other nations and is often directly encouraged through government tax policy. This article discusses the motives, causes, and influences of philanthropic behavior. Special emphasis is placed on how government policy – through the charitable deduction on income taxes and through direct grants to charities – can affect the amount of philanthropy in the economy.

Introduction

In 2006, the second richest man in America, Warren Buffett, pledged that over the remainder of his life he would give away 85% of his wealth – at the time, a sum of over $40 billion – to a foundation established by the richest man in the world, Bill Gates. By 2012, the two men have given away nearly $26 billion, and held another $36 billion in assets waiting to be handed out.

These are examples of philanthropic, or charitable, behavior – individuals freely giving money to help others. Most people are, in fact, philanthropists, although not on the scale of the two mentioned above. Through the last few decades, over two-thirds of Americans donated money to charity in any given year, with households giving over 2% of income on average. In 2011, average cash donations to charity were $2227 per married household, and $1303 for single people, making charitable giving a major category of household expenditures.

In addition, the government is an active partner with private donors in funding the charitable sector of the economy. It does this in two ways. The first way is direct grants to charities. In 2009, for instance, charitable organizations reported receiving $140 billion in government grants, or about 10% of total revenue. The second way the government spends money on charity is by excluding income spent on charitable donations from the income tax. The US tax code was modified in 1917 to allow many taxpayers to deduct their charitable donations from their taxable incomes, and by 2011 individuals were claiming nearly $175 billion in charitable deductions.

Here we will describe how economists think about and do research on philanthropy, what economists know about charitable giving, and how public policy influences it (see also Andreoni and Payne, 2013). The focus will be primarily on gifts of money and assets by individuals, although corporations and other institutions also make charitable gifts (see Williams, 2003), and people also give by volunteering time (see Brown and Martin, 2012). In addition, the discussion will center on philanthropy in the United States, since a large majority of the academic research has centered on the American experience (for information on charitable organizations internationally, see Anheier and Kendall, 2001). Finally, we will highlight what challenges remain ahead for the study of philanthropy.

An Economist's View of Charitable Behavior

Economics is founded on the view that people are self-interested. Economists assume that since individuals have a choice in how they behave, they must always make the choices that they think are the best at the time. This assumption, called the 'axiom of rationality,' provides the bedrock from which economic models of behavior are formed.

Applying this framework to philanthropic behavior, we can ask why would someone who works hard for his or her money simply turn around and give it away? Does not this contradict the assumption that individuals are self-interested? Maybe. But by the axiom of rationality, we should first look for ways in which an individual believes that giving to charity is the best thing to do with that money at that time.

One possibility is that people desire more of the service provided by the charity. National Public Radio (NPR), for instance, is a charitable organization that survives largely on donations. Over a million people a year give to NPR with most giving under $100. Are these givers getting their money's worth? Certainly not. Anyone can listen to NPR broadcasts, even without donating, and anyone who gives $100 is unlikely to notice any increase in broadcast quality as a result. It follows that a self-interested person is better off giving nothing and taking a 'free ride' on the donations of others. Hence, this alone cannot be an adequate explanation for why people give to charity (Andreoni, 1988).

A second reason is that individuals may be getting something directly from the charity in exchange for their contributions. For instance, big donors may get better seats to the opera, or donors to a university may get buildings named for them. While this must surely matter for large donors, most individuals only receive tokens (a coffee mug?), if anything, in exchange for donations.

A third reason could be that individuals get some internal satisfaction – a 'warm-glow' – from giving to their favored charity, and the more they give, the better they feel (Andreoni, 1989). This would mean that giving to charity is like buying any other good, such as ice cream. We easily accept that people have a natural taste for ice cream, so why not a natural taste for warm-glow? If we accept this, then we can analyze charitable giving just like any other consumer good – when income goes up people should want more, and when the price goes up people should want less. Scholars have examined preferences

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71001-4

for giving from a number of different approaches, and it seems that warm-glow is in fact a core economic motivation for giving (Andreoni, 1993; Ribar and Wilhelm, 2002; Harbaugh et al., 2007). But what is it that produces the private benefit from giving? Warm-glow could stem from religious duty, from pressure at work or from friends, as a signal of social status (Rose-Ackerman, 1996), or to project a positive image to others (Andreoni and Bernheim, 2009). At the heart of all of these processes, however, seems to be a basic human interest in helping others or doing one's share. So for this discussion, we will take as our launching point the idea that people get joy from giving and apply our economist's tools for analyzing consumer behavior to the study of charitable giving.

The Facts about Philanthropy in the United States

How do we learn about charitable giving? One source is surveys of consumers. The Center on Philanthropy, for instance, conducts regular surveys of over 8000 households. Surveys are valuable since they can obtain information on age, education levels, and other personal characteristics of respondents. A disadvantage is that individuals must rely on imprecise memories when answering questions, or may be reluctant to give true information about their incomes or donations.

A second important source is samples of tax returns. Since individuals who itemize their tax returns in the US can take a charitable deduction, we can learn about donations for this sector of the economy. The advantage to tax returns is that the information on them is precise, as long as people do not cheat on their taxes. (Slemrod (1989) explored this potential problem and found that, while there is some evidence of cheating by overstating charitable deductions, the effects are small and do not appreciably affect the analysis.) The disadvantage to this data is that tax returns contain very little information about the personal characteristics of the filers that would be helpful in explaining giving, such as education levels or religious affiliation, nor can we learn about the giving habits of those who do not itemize their tax returns. Since no data source is perfect, economists must conduct many studies on many data sources in order to uncover the 'facts' on charitable giving.

Sources, Totals, and Trends in Giving

Charitable donations can come from individuals, charitable foundations, corporations, or through bequests. While all are significant, by far the dominant source of giving is from individuals. Table 1 shows that in 2011 individuals gave over $217 billion to charity, or 73% of the total dollars donated. The second biggest source, foundations, was responsible for 14% of all donations.

Table 1 Sources of private philanthropy, 2011

Source of gifts	*Billions of dollars*	*Percent of total*
Individuals	217.8	73
Foundations	41.7	14
Bequests	24.4	8
Corporations	14.6	5
Total for all sources	298.5	100

Source: Giving USA, annual report 2012.

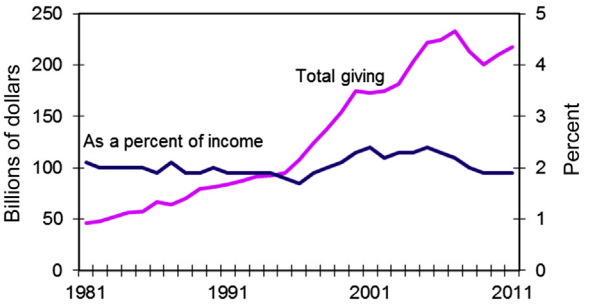

Figure 1 Giving by individuals, 1981–2011. Dollars are inflation-adjusted to 2011 values. Source: Giving USA, annual report 2012.

The trends in giving over the last 30 years can be seen in Figure 1. Total giving has been on a steady rise, with especially big jumps coming in 1996, 1997, and 1998. When measured as a percent of income, however, giving seems much more stable. Since 1968, giving has varied from 1.6 to 2.1% of the income. This figure has been remarkably stable for decades, with the occasional dips due to recessions or increases for special times of need.

Who Gives?

The Center on Philanthropy survey indicates that 65.6% of households gave to charity in 2009 and that the average gifts among those giving was $2313, exactly 2% of income. Table 2 shows that as household incomes rise, the household is more likely to give, and will make larger donations. This table also reveals an interesting pattern typically found in charitable statistics. Those with the lowest incomes tend to give the highest share of income to charity, in this survey it is 3.6% of income, while middle-income people give the least. For very high incomes, which are not captured by this data, giving again tends to swing up on average, although often with quite large differences in giving among the very rich (Auten et al., 2000).

What could cause this 'u-shaped' giving pattern? One explanation is that those with low incomes may be young people who know their wages will be rising, hence they feel they can afford more giving now. It may also be due to the types of charities people give to, since lower income people tend to give significantly more to religious causes. As we see, it will be important to account for all the factors that may explain giving before offering explanations for the averages seen in these tables.

Table 2 also illustrates that giving varies significantly with the age and educational attainment of the givers. As people get older, they are typically more likely to give to charity and to give a greater fraction of their incomes. Likewise, those with more education give more often, give more dollars, and generally give a higher fraction of income. Again, age, education, and income all vary with each grouping in the table and will have to be considered jointly.

Table 2 Private philanthropy by income, age, and education of the giver, 2009

	Percent of households who give	*Average amount given by those who give (in dollars)*	*Percent of household income*
All contributing households	65.6	2313	2.0
Household income			
Under $15,000	32.5	899	3.6
$15,000–$29,999	50.1	1219	2.7
$30,000–$44,999	53.5	1322	1.9
$45,000–$59,999	68.8	1608	2.1
$60,000–$74,999	73.3	1699	1.9
$75,000–$99,999	77.7	2298	2.1
$100,000–$124,999	87.9	2850	2.3
$125,000–$149,999	91.1	3277	2.2
$150,000–$200,000	94.4	3428	1.9
$200,000 and above	91.4	6989	1.8
Age of giver			
18–24 years	33.0	709	0.8
25–34 years	52.3	1385	1.2
35–44 years	61.4	1701	1.2
45–54 years	64.7	2692	1.9
55–64 years	76.3	2762	2.4
65–74 years	80.2	2785	3.3
75 years and above	80.3	2618	4.8
Highest education of giver			
Not a high school graduate	40.5	1209	1.2
High school graduate	57.0	1770	1.9
Some college	70.6	2122	1.2
College graduate or more	84.1	3145	2.3

Source: Author's calculations, data from the 2009 Center on Philanthropy Panel Study.

What Do They Give To?

In 2009, over 320 000 charitable, religious, and other nonprofit organizations filed with the US government. Table 3 attempts to categorize these charities by the types of services they provide. This reveals that, among all types, households are most likely to give to religious organizations and to give them the most money – 41.9% of all households give to religion and 57.5% of all charitable dollars go to religion.

Table 3 Private philanthropy by type of charitable organization, 2009

Type of charity	*Percent of households who give*	*Average amount given by those who give (in dollars)*	*Percent of total household contributions*
Arts, culture, humanities	8.0	360	1.9
Combined	25.3	613	10.2
Education	14.5	544	5.2
Environment	9.2	215	1.3
Health	23.6	305	4.8
International	5.4	402	1.4
Need	31.5	549	11.4
Neighborhood	4.7	288	0.9
Religious	41.9	2086	57.5
Youth	12.0	247	2.0
Other	7.8	662	3.4

Source: Author's calculations, data from the 2009 Center on Philanthropy Panel Study.

Taxes and Giving

How can taxes encourage philanthropic behavior? Taxpayers in the US who itemize their deductions can deduct their charitable giving from their taxable income. If someone in the 15% marginal tax bracket gives $100 to charity and takes a charitable deduction, this person will save $15 in taxes, making the net cost of the gift $85. If the person was in the 31% tax bracket then the $100 gift would have a net cost of $69. Thus, the higher one's tax rate, the lower the net cost of giving. In this way the tax rate acts as a subsidy rate to giving, and those in higher tax brackets get bigger subsidies. (Note that state income taxes often add additional subsidies. See Feenberg, 1987.)

Since the tax deduction subsidizes giving, we should expect that those with higher subsidies will make larger contributions. It is also natural to expect that those with higher incomes will make larger contributions. However, those with higher incomes will typically also have higher subsidies, so when we see richer people giving more how can we know whether their income or the subsidy is causing the change? If we can compare people with similar subsidy rates but different incomes, and similar incomes but different subsidy rates, we can apply statistical techniques to separate the influences of income and the subsidy.

Note that no one can give to charity solely to 'get a tax benefit.' Even a person in the highest federal tax bracket, 35%, will still pay 65 cents to give away another dollar. But this can skew how we view the statistics when we report giving as a percent of income. Earlier we noted that someone with income of $35 000 would give 1.9% of income, as would someone with income of $175 000. Since the marginal tax rate at the low income is 15% and at the high income is 35%, the higher income person is getting a bigger tax benefit. After the tax savings, the low earner's net contribution is 1.6% of income, while for the high earner it is 1.3%.

An important thing for policy makers to determine is whether subsidizing giving through tax deductions is worth the cost to the government. If, for instance, the subsidy does not increase giving at all, then the policy is costing the government tax dollars without benefiting the charities. If, on the other hand, the subsidy creates a dramatic response, tax deductions may be an effective tool to increase social well-being. Generally, policy makers look for the increase in contributions due to the subsidy (the benefit) to be larger than the loss in tax dollars (the cost). The next section discusses how economists explore this issue.

The Effect of the Subsidy on Giving

Economists began conducting detailed empirical studies of giving in the 1970s in an attempt to isolate the effect of the subsidy on giving (see Clotfelter, 1985, for a review). You can imagine asking the counterfactual policy question this way: Consider a proposal that would allow everyone to claim 110% of their current deduction on their income taxes – make a $100

contribution and take a deduction for $110. This means that the net cost of a $100 contribution would fall by 10%. For someone in the 28% marginal tax bracket, for instance, the cost of giving $100 would fall from $72 to $62.80. We would expect that this lower cost might encourage more giving, but by how much?

The first generation of studies to explore this counterfactual relied on cross-sectional data, primarily from tax returns. Observing income and deductions for thousands of people in a given tax year, the researchers use the variations in income and cost to estimate how, on average, individuals would respond to such a change in policy. There were several limiting features of the data used in these studies. Of particular importance is that the samples, for reasons of confidentiality, typically did not include extremely wealthy people.

Depending on the source of the data, the years studied, the size of the sample, and the statistical approach, the results naturally varied. It has been generally agreed, however, that the prediction that best characterizes the results is that a policy that would reduce the cost to donors by 10% – without affecting the other income taxes paid by the individual – would increase their contributions by about 13% (Clotfelter, 1990). This suggests that behavior is quite responsive to the incentives offered by the tax deduction, and that the response is large enough to satisfy the policy criterion noted above. In particular, if the government pays an additional 10% of the cost of giving and if people respond by giving 13% more, then the policy generates more new charity than it costs the government in tax revenue.

These studies also confirmed that giving increases with income. If after-tax incomes were to rise by 10%, estimates are that giving would rise by about 8% (Clotfelter, 1990). Notice that this quantifies the left-hand part of the u-shaped giving curve discussed in Section An Economist's View of Charitable Behavior – since giving rises by a lesser percent than income, giving as a percent of income (all else equal) must be declining.

Later studies were able to supplement these findings using data from surveys rather than tax returns (see Clotfelter, 1990; Steinberg, 1990). Although the survey results varied, they generally confirmed the policy predictions stated above. In addition, they revealed the importance of other individual characteristics in explaining giving. Most strikingly, they consistently found that giving increases with the age and the education of the giver, as suggested in the tables above. It could be that those who are educated have more interest in charity or that interest in charity changes with age. A more plausible explanation, however, is that researchers have no information on individual wealth, and since wealth is correlated with both age and education, it is likely that the influence of wealth is being filtered through these other variables.

A second generation of studies, conducted mostly in the 1990s, has begun to shift the consensus view noted above. Unlike the earlier studies that relied on cross-sectional data, these newer studies used panel data, that is, data with observations on the same set of individuals over several years. Panel data is generally seen as superior to cross-sectional data. The reason is that by observing the same people at different points in time, the researcher is able to get a more precise measure of how behavior responds to changes in the environment.

Randolph (1995) used tax returns on about 12 000 people from 1979 to 1988, and statistical techniques designed explicitly for panel data identified a much weaker effect of the subsidy and a much stronger effect of income. His results indicate that a policy that would increase the government subsidy by 10% would increase giving by only about 5%, whereas if income were to rise by 10% giving would rise by 11%. Others using panel data have found similar effects. Hence, at the end of the 1990s, the literature on the effect of the subsidy to giving was in a state of flux.

Bakija and Heim (2011) waded into these waters in an attempt to reconcile these differences. Like Randolph, they assembled a panel of tax returns, but their panel was much larger, with over 550 000 returns on almost 60 000 tax filing units. They also adopted an estimation technique that contained elements of the original analysis from the 1970s and 1980s as well as the more dynamic analysis of the 1990s. In addition, the data had a distinct advantage in having data on very wealthy households.

Their results were a mix of the two prior generations of findings. For households with incomes below $200 000, a 10% increase in the subsidy only raises giving by 7–8%. On these people, the government is getting less new charity than it is losing in revenue. For families reporting incomes of $200 000 on up to millionaires, Bakija and Heim predict that, on average, a 10% increase in the subsidy will raise giving by 11–13%, thus the government gets a good return on its investment from these donors. An important caveat to these results, however, is that while the prediction for low- to upper-middle-income households is fairly precise, estimates for the top 5% of incomes is very imprecise – the true response to a 10% increased subsidy could be anywhere from 5 to 20% increase in giving by the wealthy. While this article is a step forward, the literature is still inconclusive about the giving behavior of the very wealthy.

Do Government Grants Displace Private Philanthropy?

There are several reasons to suspect that government grants to charities might depress private donations. First, if givers are aware of the grants, then they may perceive the charity as less in need of their donations, leading them to keep the money or give it to a different cause. This is known as the 'crowding out' hypothesis (Warr, 1982; Roberts, 1984; Bergstrom et al.,1986). A second reason giving may decline is that the charity may be less aggressive in fundraising after receiving a government grant. Both effects would lead to government grants to offset private philanthropy.

There are also reasons to think that government grants could have the opposite effect. Citizens who are unsure about the quality of a charity could see a grant as a 'stamp of approval' (Andreoni et al., 2014). Or the grant could be used as 'seed money' that will allow the charity to expand into a bigger operation by, for instance, constructing new buildings or opening branches in different cities (Andreoni, 1998). In both of these cases, government grants could actually encourage private philanthropy.

An important difficulty in unraveling these effects is that the government and the voters are likely to have similar tastes. For instance, a tsunami may cause both individuals and the government to give more to that cause, but it would be incorrect to conclude that the government giving 'caused' the private giving – the tsunami caused them both simultaneously.

Failure to account for this 'simultaneity bias' would make it appear that crowding out is smaller than it appears.

This was pointed out in important study by Payne (1998). She used a 10-year panel of 430 US charities, and applied special statistical techniques to account for the fact that government grants to charities may be caused by the same needs and interests as private giving. While prior research that did not account for this found no crowding out, her analysis predicts about 50% crowding out.

Several years later, Andreoni and Payne (2003, 2011) discovered something important was missing from this research – charitable fundraising. Most charities find fundraising as a necessary distraction from their main mission, and a government grant gives them a reason to redirect their efforts back to the charity itself. That is, a grant may crowd out giving indirectly by causing the charities to cut back their efforts at fundraising.

Andreoni and Payne explore this possibility by analyzing tax filings of over 8000 charitable organizations in the US for up to 12 years. Looking at the congressional district the charity is located in, and including measures for how powerful their representative may be in helping secure grants in the district, they are able to map the simultaneity of government and private giving to the charity. They find that overall, a government grant of $10 000 reduces giving about $7500 – a substantial crowding out. However, a large fraction of this is due to a reduction in fundraising by the charity. Breaking it down, of the $7500 reduction in giving, from $5250 to all $7500 is due to the charity reducing fundraising after a grant. This points to the importance of understanding fundraising for making wise policy decisions.

Fundraising and Matching Grants

The last section highlighted how important fundraising is in understanding the effects of government policy. Often times, major donors to charity can affect the same kinds of policies used by the government. For instance, a single wealthy donor will sometimes give a large sum, which they then promise as 'matching gifts' to other donors. The wealthy philanthropist may promise to spend up to $1 million by giving $1 for every $1 raised from other donors. This is very similar to a government subsidy. What is the effect? Does it work as an effective subsidy to others' gifts?

Two papers looked at this issue, Karlan and List (2007) and Huck and Rasul (2011). There are several factors that make inference about matching gifts difficult. First, is it really a match? If the charity usually raises $5 million in a fund drive, then the donor expects to spend the entire $1 million, meaning that no one's donation really had an effect on whether $1 would be collected. In that case the matching gift is really more like a 'leadership gift' (Andreoni, 1998). Leadership gifts have importance often because they can signal the quality of the charity, or they can provide assurance that the charity's goals will be met. In both cases, the leadership gift can inspire others to give more, that is, they can 'crowd in' rather than crowd out smaller donors.

These two research studies conducted experiments with actual charities to try to determine which story fit the data. What they found was that individuals tended to respond to the fact of a matching gift, but not to the match level; a $1-for-$1 match raised as much as a $3-for-$1 match. But giving responded to a simple announcement of an equivalent leadership donation about the same as it did to an announcement of a matching gift. Thus, matching gifts work, but it is likely because they work like leadership gifts rather than that they act like subsidies to giving.

Conclusions and Future Research

Philanthropy is a significant factor in the US economy, accounting for about 2% of income almost every year for the past 40 years. Government and private givers are in a partnership to fund the services of almost 200 000 charitable organizations. This article has discussed the economic influences on private philanthropy and how it interacts with government policy.

The most important finding is that individuals are indeed sensitive to the charitable deduction in the US tax system. By subsidizing giving, the tax deduction has clearly increased giving by the private sector, although the effect clearly differs across income groups. The tax deduction is only available to those who itemize deductions on their tax returns, thus low-income individuals get no benefit. For middle-income people who itemize, the deduction returns 10–28% of their donation in lower taxes, and increases their donations by 7–20%. High-income households, making $200 000 per year and more get 33–35% of their donation back in lower taxes, and this raises their donations by about 35–40%.

The literature on the crowding out of private giving by government grants indicates that an important factor in the analysis is the role of fundraising. When charities get grants, many will redirect their efforts away from fundraising. As a result, government grants only result in about 30 cents on the dollar of new charitable services.

What can we look forward to as the important questions for future research? A surprising new development in the world of charitable giving is the explosion in private foundations. A foundation is a tax-exempt organization with a charitable or not-for-profit mission. Individuals, especially very wealthy individuals, have been setting up foundations at increasing rates. Part of this is to gain tax savings during life rather than to allow money to pass into an estate, but part is also a sincere desire to do good during one's life. Recent studies have indicated that foundations are increasingly important and powerful participants in the market for charity and deserve greater attention by researchers and policy makers.

See also: Altruism and Prosocial Behavior, Sociology of; Charity and Philanthropy: Overview; Income Taxes; Wealth Distribution.

Bibliography

Andreoni, J., 1988. Privately provided public-goods in a large economy- the limits of altruism. Journal of Public Economics 83, 57–73.

Andreoni, J., 1989. Giving with impure altruism: applications to charity and Ricardian equivalence. Journal of Political Economy 97, 1447–1458.

Andreoni, J., 1993. An experimental test of the public-goods crowding-out hypothesis. American Economic Review 83, 1317–1327.
Andreoni, J., 1998. Toward a theory of charitable fund-raising. Journal of Political Economy 106, 1186–1213.
Andreoni, J., Bernheim, B.D., 2009. Social image and the 50-50 norm: a theoretical and experimental analysis of audience effects. Econometrica 77 (5), 1607–1636.
Andreoni, J., Scholz, J.K., 1998. An econometric analysis of charitable giving with interdependent preferences. Economic Inquiry 36, 410–428.
Andreoni, J., Payne, A.A., 2003. Do government grants to private charities crowd out giving or fundraising? American Economic Review 93, 792–812.
Andreoni, J., Payne, A.A., 2011. Is crowding out due entirely to fundraising? Evidence from a panel of charities. Journal of Public Economics 95 (5–6), 334–343.
Andreoni, J., Payne, A.A., 2013. Charitable giving. In: Auerbach, Alan J., Chetty, Raj, Feldstein, Martin, Saez, Emmanuel (Eds.), Handbook of Public Economics, vol. 5. Elsevier, pp. 1–50.
Andreoni, J., Payne, A.A., Smith, S., 2014. Do grants to charities crowd out other income? Evidence from the UK. Journal of Public Economics 114, 75–86.
Anheier, H.K., Kendall, J., 2001. Third Sector Policy at the Crossroads: An International Non-profit Analysis, vol. 10. Routledge, London and New York.
Auten, G.E., Clotfelter, C.T., Schmalbeck, R.L., 2000. Taxes and philanthropy among the wealthy. In: Slemrod, J. (Ed.), Does Atlas Shrug? The Economic Consequences of Taxing the Rich. Harvard University Press, Cambridge, MA.
Bakija, J., Heim, B.T., 2011. How does charitable giving respond to incentives and income? New estimates from panel data. NationalTax Journal 64 (2), 615–650.
Bergstrom, T., Blume, L., Varian, H., 1986. On the private provision of public goods. Journal of Public Economics 29, 25–49.
Brown, E., Martin, D., 2012. Individual giving and volunteering. chapter 13. In: Salamon, Lester (Ed.), The State of Nonprofit America, second ed. Brookings Institution Press.
Clotfelter, C.T., 1985. Federal Tax Policy and Charitable Giving. The University of Chicago Press, Chicago, IL.
Clotfelter, C.T., 1990. The impact of tax reform on charitable giving: a 1989 perspective. In: Slemrod, J. (Ed.), Do Taxes Matter? The Impact of the Tax Reform Act of 1986. The MIT Press, Cambridge, MA, pp. 203–242.
Feenberg, D., 1987. Are tax price models really identified—the case of charitable giving. National Tax Journal 40, 629–633.
Harbaugh, W.T., Mayr, U., Burghart, D.R., 2007. Neural responses to taxation and voluntary giving reveal motives for charitable donations. Science 316 (5831), 1622–1625.
Huck, S., Rasul, I., 2011. Matched fundraising: evidence from a natural field experiment. Journal of Public Economics 95 (5), 351–362.
Karlan, D., List, J.A., 2007. Does price matter in charitable giving? Evidence from a large-scale natural field experiment. The American Economic Review 97 (5), 1774–1793.
Payne, A.A., 1998. Does the government crowd-out private donations? New evidence from a sample of non-profit firms. Journal of Public Economics 63, 323–345.
Randolph, W.C., 1995. Dynamic income, progressive taxes, and the timing of charitable contributions. Journal of Political Economy 103, 709–738.
Ribar, D.C., Wilhelm, M.O., 2002. Altruistic and joy-of-giving motivations in charitable behavior. Journal of Political Economy 110, 425–457.
Roberts, R.D., 1984. A positive model of private charity and public transfers. Journal of Political Economy 92, 136–148.
Rose-Ackerman, S., 1996. Altruism, nonprofits, and economic theory. Journal of Economic Literature 34, 701–728.
Slemrod, J., 1989. Are estimated tax elasticities really just tax evasion elasticities? The case of charitable contributions. Review of Economics and Statistics 71, 517–522.
Steinberg, R., 1990. Taxes and giving: new findings. Voluntas 1, 61–79.
Warr, P.G., 1982. Pareto optimal redistribution and private charity. Journal of Public Economics 19, 131–138.
Williams, R.J., 2003. Women on corporate boards of directors and their influence on corporate philanthropy. Journal of Business Ethics 42 (1), 1–10.

Charity and Philanthropy: Overview

Dennis W Cheek, National Creativity Network, USA; and IESEG School of Management, France
Michal Kramarek and Patrick Rooney, Indiana University Lilly Family School of Philanthropy, Indianapolis, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Charity and philanthropy are overlapping concepts whose habitual, periodic, and diverse expressions throughout recorded history across cultures and time have resulted in mainly positive benefits for humankind and societies. Formal research regarding the nature, history, extent, impact, and changing outlook concerning these twin concepts is a fairly recent development and has spawned a growing and ever more robust literature with many underexplored and still poorly understood dimensions.

Charity and philanthropy have been defined in different ways through the ages and in diverse cultures. Both concepts involve, in general, a basic desire to make a positive difference in the lives of others. Philanthropy literally conveys the idea of goodwill demonstrated toward the human race, while charity is an empathetic and altruistic response, frequently to human misfortune or suffering ('love in working clothes'). Both concepts involve a direct or indirect relationship between a personal giver (donor) and others within the realm of civic life, rather than relying solely on governmental or strictly institutional actions.

Philanthropy, according to a broad Western definition, is voluntary action for the public good, where 'voluntary' means unforced and free, where 'action' refers to giving, serving, and associating, and where 'public good' refers to creation of goods that are indivisible and nonexcludable or improve the quality of life for other(s) for whom one is not formally responsible. This improvement in the quality of life may include the alleviation of suffering by means of satisfying basic needs (e.g., food and shelter) and/or enhancing life by satisfying nonbasic needs (e.g., recreation, education, and self-actualization) (Payton and Moody, 2008). Traditionally, both the subject and object of philanthropy are humans or groups of humans, although in more recent centuries other creatures and the environment itself increasingly are objects of interest. Many other definitions of philanthropy and related terms exist, including diverse non-Western perspectives (Smith et al., 2006; Ilchman et al., 1998). This article focuses primarily on a Western perspective since the overwhelming bulk of research conducted to date has been undertaken in Western nations by researchers working within this broad cultural tradition.

Historical Meaning of Philanthropy

Philanthropy is an umbrella word used to describe a range of concepts from past to present and from academic to practical. Philanthropy originated in ancient Greece between the fifth and fourth centuries BC as a specialized philosophical and theological term to describe a facilitator of human advancement and was applied to an exceedingly wide range of human activities (Sulek, 2009a). One of the oldest documented philanthropic acts, more in keeping with modern usage, was the endowment of Plato's Academy where Plato picked his successor and bequeathed his wealth upon his death (347 BC). It lasted until Philo of Larissa's death in 83 BC, followed by a subsequent revival in AD 410, and final abolition of the Academy by order of Justinian I in AD 529 for a total duration of 383 years. Preliminary evidence from around the globe regarding the ancient world highlights charitable and philanthropic actions by rulers, their subjects, an various social organizations with religious texts from various traditions encouraging such behaviors and public and more private solicitations, formal petitions (from commoners and nobility alike), and commendations accompanying its practice (e.g., Veyne, 1992).

The meaning of philanthropy by the nineteenth century still referred to one of seven broad categories: literal, archaic, ideal, ontological, volitional, actual, and social (Sulek, 2009b). During the twentieth century, the term philanthropy became narrower in both scholarly discourse and public use, primarily referring to the act of giving and actions that follow from it.

Present-Day Paradigms of Giving

In current academic discourse, three paradigms of giving can be distinguished (Gunderman, 2008). The first one, egoistic giving, is described as a function of a donor's egoism. For example, in the fourth century BC, Spurius Maelius was slain for his philanthropic deeds, because he was suspected of trying to obtain political power. Since the Industrial Revolution, egoistic giving was viewed as a means of controlling the working class (Davis, 1999; Adam, 2009). During the Great Depression in the 1930s, egoistic giving served as 'riot insurance' for businessmen and bankers (Bremner, 1988). In the twentieth century, egoistic giving was pilloried as a means to socially repent for acquired wealth, build a legacy (Rosenwald, 2008), or feed one's pride (Menninger, 2008). Despite this critique, egoistic giving may be perceived as beneficial to society overall, if there is no such thing as 'bad' giving (Kelley, 2008) or a giver embodies some notion of the public interest (Kristol, 2008).

Compassionate giving or charity, the second paradigm of giving, is described as a model of philanthropy historically aiming to satisfy immediate human needs. In the United States, for example, formal expression of compassionate giving can be traced to Governor John Winthrop's 'A Model of Christian Charity,' composed in 1630. Within Judaic traditions, compassionate giving relates to the concepts of *Tzedakah* (works of charity, justice) and *Tikkun Olam* (repairing the world). In Christian traditions, compassionate giving relates to concepts of *caritas* (unlimited loving kindness to all others), *tithe* (literally a 'tenth' of one's income, drawing upon the Old Testament), and *offerings* given as part of Christian worship. In Muslim traditions, compassionate giving relates to concepts of *takaful* (social solidarity), *zakat* (donating a fraction of wealth as part of one's worship), *sadaqa* (benevolence), and *waqf* (endowments). Historical studies of these and other traditions as they relate to charity clearly demonstrate both strategic and far-sighted perspectives that informed giving and efforts to improve both communities and societies (e.g., Loewenberg, 2001 (ancient Judaism); Hamel, 1990 (Christianity and Judaism); Singer, 2008 (Islam); Constantelos, 1991, 1992 (Orthodoxy); Smith, 2009 (medieval China); and Ilchman et al., 1998 and Barnett and Stein, 2012 (current world religions)).

Scientific giving, the third paradigm of giving, aspires to improve systematically particular aspects of the human condition. For example, instead of focusing on feeding the hungry, scientific giving aims to make the hungry self-sufficient, so they can feed themselves. To achieve such far-reaching goals, scientific giving takes a long and wide perspective. It insists on rationalized giving and often aspires to objective measurable outcomes. Scientific giving is frequently presented as a novel approach and superior to compassionate giving, which, in both cases, is not correct (Schambra, 2008; Ruswick, 2013). While scientific giving has proved highly successful in some areas, most notably medicine, its overall effectiveness is not as obvious. In Western culture, the roots of scientific giving can be traced to the Enlightenment era, when philanthropy began to refer increasingly to rational consciousness. Recent examples of these approaches include strategic, performance, and outcomes philanthropy (Brest, 2012; CSR Asia, 2010), catalytic philanthropy (Kramer, 2009), venture philanthropy and social impact investing (where philanthropy focuses on maximizing socially valuable outcomes rather than strictly economic ones) (John, 2012; John et al., 2013), philanthrocapitalism (Bishop and Green, 2008), and similar conceptualizations. There are a number of hybrid concepts, which contain elements from the above paradigms or which relate to paradigms not classified as philanthropy (e.g., cause-related marketing).

While all three paradigms of giving can be found in nations around the globe, the mix and extent of them will vary depending on historical, sociopolitical, cultural, geographic, economic, and other factors. Western European nations, for example, considerably expanded the depth and breadth of the welfare state as working class movements, political parties, and formal organizations worked out varied approaches to provide social supports and safety nets as a matter of fundamental rights across the human lifespan (Kaufmann, 2012; Castles et al., 2010). A corollary to these fundamental changes in societal provision for housing, education, health care, and social services as entitlements was a correspondingly substantial rise in tax rates for many wage earners and a decrease in both the need for and provision of private charity and philanthropy. Yet, within these strong social welfare states there is considerable variation in private charitable giving and philanthropy with Swedes, for example, living in what is usually regarded as one of the most extensive social welfare regimes, giving at a rate considerably above the European average (Vamstad and von Essen, 2013). Despite the sometimes immense differences in both social welfare provision and charitable giving, the influence of tax incentives and price and income elasticities appears to differ very little, if several comparative studies of Germany and the United States are a reliable guide (e.g., Bönke et al., 2013).

Social Capital

Social capital is a concept closely linked to philanthropy. Social capital, in general, refers to the value of social relations that positively affect individual or collective productivity. There is no one single, universally accepted definition of social capital. Philanthropy both promotes social capital and relies on it. The term 'social capital' can be traced to a 1916 article on the rural school community center by Lyda Judson Hanifan. Wider academic use of the term began in the 1950s.

Current research identifies a number of positive and negative consequences of social capital. The positive consequences of social capital include better education and child welfare, safe and productive neighborhoods, better health, happiness, economic prosperity, and better democracy, social order, and social support. The negative consequences of social capital are related to excessive costs of membership (e.g., the costs of belonging to a social group may impose demanding norms of behavior), limited individual freedom (solidarity of a social group may restrict individual freedoms), downward leveling of norms (successful outliers may be exploited by their social group), exclusiveness (those outside social relations may be excluded from the benefits), and antagonism (those outside the group may resent the solidarity of the group) (Portes, 1998; Putnam, 2000).

Nonprofit Sector

The nonprofit sector, in general, refers to a sphere of private, formal economic activity without paying dividends or profits to owners or investors. Philanthropy is an important source of support for the nonprofit sector (e.g., Zunz, 2011; Bradley, 2009; Tennant, 2007). Philanthropy in a broader sense also includes activities carried out in the sector, for example, community building as a means of creating social capital (Salamon, 2010). A multitude of substitute terms exist to define nonprofit organizations across different geographical locations, legal systems, cultural traditions, and disciplinary perspectives (Lohmann, 1992). Four distinguishable concepts for the nonprofit sector can be delineated (Salamon and Anheier, 2006).

Charitable, nonprofit, and voluntary sector are terms that can be traced back to the Elizabethan Poor Law of 1601 in England. The US federal tax code has used the term 'nonprofit' since the 1950s to define these incorporations bound by a nondistribution constraint and also employs the term 'public

charity.' The concept also emphasizes a type of service that is defined by law as 'charitable' and includes services to the general public or to those in need, an orientation seen in the Statute of Charitable Uses in the United Kingdom and organizations that have proliferated in the United Kingdom (Hilton et al., 2012). Nonprofit is frequently critiqued for defining a sector by what it is not (Salamon and Anheier, 1996).

Social economy is a second concept that emphasizes the ethos of solidarity or ideology governing organizations within the nonprofit sector. The French term 'social economy' (économie sociale) can be traced to the beginning of the twentieth century. It is used to describe a sector comprising cooperatives, mutuals, associations (voluntary organizations), and foundations of public utility, which place service ahead of profit, place people and work ahead of capital, have autonomous management, and make decisions in a democratic manner. The term 'social enterprise' and the more recent term 'social business' carries this tradition around the globe, contextualized from place to place (Kerlin, 2009; Yunus and Weber, 2011).

Nongovernmental sector, independent sector, and third sector are terms referring to a third concept of the nonprofit sector that emphasizes sovereignty and specifically autonomy from the government. (The 'third sector' is everything that is not part of business or government. An older and less widely used term is the 'voluntary sector.') It focuses on the sector's function to empower the general public or those in need. As such, this concept is narrower than the other three uses and tends to be employed more frequently in emerging economies or post-socialistic countries (Salamon and Anheier, 1996).

Civil society is the term referring to a fourth concept of the nonprofit sector. This concept can be traced back to the 1980s when Central European states began their transition from socialism to democracy. This concept emphasizes the sovereignty of the sector in general with a focus on the noninstitutional, grassroots activity it conducts. This concept may incorporate suspicion toward organized institutions of the sector, which are sometimes believed to serve hidden agendas (Salamon and Anheier, 1996).

Given the conceptual diversity of the nonprofit sector, the Johns Hopkins Comparative Nonprofit Sector Project developed what was intended to be an internationally universal, structural/operational definition encompassing different legal and cultural traditions based on five key features (Salamon, 2010). The fundamental feature of the sector is that it consists of organizations. These organizations do not have to be formally constituted or legally registered but are characterized by structure and consistency in operations. The organizations are not part of government but may rely on public funding and may be private, not public. The primary purpose of the organizations is not commercial activity and they do not distribute profits to owners or managers. The organizations are self-governed, managed independently, control their own activities, and can independently decide to dissolve. The sector is voluntary and organizations do not impose membership or participation.

Theories of the Nonprofit Sector

The three-failures theory is the most widely employed idea to explain the existence of the nonprofit sector and the relationship among institutions of the public, private, and nonprofit sectors (Steinberg, 2006). Three-failures theory argues that the competitive market fails for three reasons. First, the market fails to supply pure public goods (i.e., goods consumed collectively and goods that do not diminish with additional consumption, e.g., national security). Second, the market fails to supply excludable goods to all who need them (i.e., some cannot afford even low-cost ones). Third, the market fails to supply the right quantity of certain private goods (i.e., goods that can be owned exclusively and customers can compete for them), a result of asymmetric information (Hansmann, 1980). Government could address these failures, because it does not operate to maximize profits, but it fails in two ways. First, government fails to provide goods that would satisfy the needs of minorities. Second, government fails to supply a sufficient quantity of collective goods to those with high levels of demand (Weisbrod, 1991). The nonprofit sector can address both market and government failures because it does not operate to maximize profits and it is not restricted to render a homogeneous response to heterogeneous demand as required by democratic principles. However, the nonprofit sector fails to offer an adequate supply of public goods, due to particularism, paternalism, philanthropic insufficiency, and amateurism (Salamon, 1987). This failure, in turn, is addressed by the private for-profit market and government sectors.

Giving in the United States of America

The most complete data sources on charitable giving can be found in the United States, partly because of long-standing federal tax authority rules that require annual disclosure of financial and operational details by nonprofits including lists of grants made and to whom; income received and sources of major income; salaries of the five highest paid staff, chief executive officer, and board members; contracts to third parties; investments; and subsidiary organizations. Individual tax payers as well as corporations may also file federal tax forms to claim various qualifying deductions, including those for charitable contributions. These two annual, comprehensive, and generally reliable sources of data can be used to understand giving across the nation.

Giving USA is the most comprehensive annual report about charitable giving. It has been published since 1956 by the Giving USA Foundation and researched and written by the Indiana University Lilly Family School of Philanthropy. *Giving USA* estimates are mostly based on econometric methods using tax data, government estimates for economic indicators, and information from other research institutions. The estimates are revised 2 years after their first publication when tax return information about itemized deductions by individuals and corporations and tax statements from nonprofit organizations (including foundations) become available. *Giving USA 2013* on giving for the year 2012 reports charitable giving amounting to $316.23 billion or 1.9% of the gross domestic product (GDP) (Giving USA Foundation, 2013). In comparison, over the last 40 years, total giving ranged from 1.7% of GDP to 2.3%. Total giving increased in current dollars every year but three (1987, 2008, and 2009) since recording began in 1956. In 2012, the largest source of contributions was individual giving

at $228.93 billion or 72% of total contributions. The other sources included giving by foundations ($45.74 billion or 15%), giving by bequests ($23.41 billion or 7%), and giving by corporations ($18.15 billion or 6%). Since the beginning of the study, the religious subsector has received the largest share of contributions. In 2012, giving to religion was estimated at $101.54 billion or 32% of total giving. However, giving to religious organizations has been decreasing as a share of the total giving since the 1986–90 period. The other major uses of giving include education ($41.33 billion or 13%); human services – response for those affected by the economic crisis, especially by providing emergency basic needs services ($40.40 billion or 13%); gifts to grant making foundations ($30.58 billion or 10%); health ($28.12 billion or 9%); gifts to public-society benefit funds, such as United Ways, Jewish Federations, and freestanding donor-advised funds ($21.63 billion or 7%); international affairs ($19.11 billion or 6%); arts, culture, and the humanities ($13.28 billion or 5%); and environment/animals ($8.30 billion or 3%). Besides monetary and in-kind gifts, *Giving USA* incorporates estimates on volunteering and the number of nonprofit organizations. According to the Corporation for National and Community Service, 64.3 million adults volunteered in 2011. At an estimated average value of volunteer work of $22.14 per volunteer hour, this amounts to $171 billion worth of volunteering in 2011. The number of tax-exempt organizations was estimated at 1 081 891.

The 2012 Bank of America Study of High Net Worth (HNW) Philanthropy is a leading resource on HNW individuals. It is a biannual study conducted by the Indiana University Lilly Family School of Philanthropy. The study is based on a random sample of US residents with a household income greater than $200 000 and/or net worth (excluding the value of their primary residence) of at least $1 000 000. The importance of research focusing specifically on HNW households derives from the reliance for nonprofit sector growth on the income volatility of the top income group (Parker and Vissing-Jorgensen, 2009), which raises concerns about rich households' ability to consistently finance the nonprofit sector. According to the fourth study, published in 2012 on giving in 2010, HNW households remained committed to supporting nonprofits despite the economic downturn, as 95.4% of those households donated to charity, as compared to 65.4% of the general population (Indiana University Lilly Family School of Philanthropy, 2012). Nearly 69% of wealthy households gave consistently year after year to the same organizations or causes. However, average charitable giving decreased by 6.8% from 2009 to 2011. The structure of giving by wealthy households differs from the structure of giving by the general population. In 2011, the most popular causes among HNW households were education (79.6%), basic needs (79.3%), art (68.8%), health (65.3%), and religion (65.2%). In 2011, 61% of HNW households gave their largest gift to fund the general operations of an organization, an increase of about 5% points from 2009 and likely in response to the increased difficulties organizations faced raising core operating expenses in a downturned economy. The study also found that HNW households tend to give strategically and are less willing to take risks in philanthropy than they are with their personal investment portfolio. The study also estimated volunteering. In 2011, 88.5% of HNW households volunteered. More than one-third of these volunteers gave 200 h or more of their time in 2011.

Giving in Global Perspective

A growing body of research informs the understanding of philanthropy, civil society, and the nonprofit sector in global perspective (e.g., Salamon and Anheier, 2006). This research is difficult for a number of reasons, including differences in the concept of philanthropy on cultural and legal levels, differences in the amount and quality of data available, risk of double counting, differences in purchasing power parity, and the scale and expense of a global study.

The Johns Hopkins Comparative Nonprofit Sector Project is a large-scale collaborative research project undertaken in some 40 countries beginning in the 1990s that generated for the first time a systematic body of global comparative data on the nonprofit sector. It is based on evidence from more than 600 nonprofits and other entities as well as secondary data that formed reasonably representative samples within the nations studied. In regards to the overall size of the nonprofit sector, the study estimates that in 35 countries, the nonprofit sector accounted for a total of 39.5 million full-time workers or an average of 4.4% of the economically active population in these countries. Such size would make it the seventh largest economy in the world, if all nonprofit sectors globally were a single country. The size of the sector varies considerably between countries, from 0.4% of the economically active population of Mexico to 14% in the Netherlands. The United States has the fourth largest nonprofit sector at 9.8% of the economically active population. There is also variation between developed countries and developing ones, but, on average, the nonprofit workforce in developed countries is over four times larger than in developing nations. In regards to the composition of the sector, the study found that 64% of nonprofit activities are carried out in service fields, comprising mainly education (23% of the workforce), social services (19%), health (14%), and development (8%). The bulk of remaining activities are carried out in expressive fields comprising mainly culture and recreation (19%), professional and business representation (7%), as well as civic advocacy and environmental protection (6%). With regards to volunteer inputs, the study found that 16.8 million or over 40% of full-time equivalent nonprofit workers in the analyzed 35 countries were volunteers. However, countries vary considerably in their reliance on volunteering as opposed to paid labor. For example, in Sweden, over 5% of the economically active population volunteered and less than 3% received pay for their work in the sector. On the other end (Belgium), less than 3% volunteered and over 8% received pay. These differences among countries were explained by cultural variations in the concept of volunteering. The study found that in Anglo-Saxon countries the concept of volunteering is rooted in the democratic concept of a self-organizing society, separate from business and the state. In other Western countries, such as Germany, the concept of volunteering is rooted in the autocratic concept of communal service for the public good. In some non-Western countries,

there is no tradition of volunteering and, therefore, its conceptualization is not well developed (for a recent view from Asia see Brassard et al., 2010). Regarding the revenue structure, the study found that in 32 analyzed nations, on average, 53% of all revenue came from fees, 35% came in the form of government funding, and 12% came from philanthropy. When the value of volunteer work is considered as a form of philanthropic support, philanthropy constituted 30% of total revenue. As in other cases, the structure of revenues in the nonprofit sector varied significantly between countries. Traditionally, government support is the most important source of funding for the nonprofit sector in Western Europe, while fees play a dominant role in the nonprofit sectors of countries in Africa, Central and Eastern Europe, Latin America, as well as in Australia, the United States, and Japan.

The Civil Society Organization Index is a composite index of civil society development used in the Johns Hopkins Comparative Nonprofit Sector. The index comprises three dimensions (capacity, sustainability, and impact) where each dimension is constructed on a number of different measures. The capacity dimension measures the employment level, the diversification of employment, mobilization of volunteers, and stimulation of charitable resources in the nonprofit sector. The sustainability dimension measures the financial base, membership, volunteering, and the supportiveness of the legal environment in which the nonprofit sector operates. The impact dimension measures the service function and expressive function of the nonprofit sector. The country scoring the highest in the index was the Netherlands with the highest scores in all three dimensions. The United States was third on the list and Mexico received the lowest score.

The World Giving Index (WGI) is based on data from the *Gallup WorldView World Poll*, an ongoing research project carried out in up to 153 countries involving over 150 000 people in its now annual stratified samples. The WGI is based upon three charitable behaviors: giving money to an organization, volunteering time in an organization, and helping a stranger. According to data collected in 2011 from 146 nations, societies give in different ways (Charities Aid Foundation, 2012). The frequency of giving money ranges from as low as 3% in Georgia and Senegal to 79% in Ireland. The frequency of volunteering ranges from 3% in Greece and Serbia to 58% in Turkmenistan. Helping strangers is the main way that the world gives as 45% of the world's population helped a stranger at least once in 2011. Based on the WGI aggregate value, around half of the 20 most charitable countries are wealthy nations, but around half are not (e.g., Liberia, Philippines, Sri Lanka, and Turkmenistan). The top 10 nations for overall giving (all three forms combined) in descending order were Australia, Ireland, Canada, New Zealand, the United States, the Netherlands, Indonesia, the United Kingdom, Paraguay, and Denmark. According to a 5-year trend analysis (Charities Aid Foundation, 2012), giving money is the highest among the oldest age groups, volunteering attracts the 35- to 49-year and 15- to 24-year age groups the most, while helping strangers is least common among those 50+.

The Index of Global Philanthropy and Remittances is an annual metastudy, which combines findings from multiple sources in different countries (Adelman et al., 2012). The data on remittances are derived from the World Bank's bilateral matrix, which is considered the only comprehensive and comparable source of all bilateral remittance flows between the Organization for Economic Cooperation and Development (OECD) and recipient countries. According to the 2012 publication on results from 2010, global philanthropy, remittances, and private capital investment worldwide totaled $575 billion or 82% of the developed world's economic dealings with developing countries. In this amount, government aid accounted for only 18% of total financial flows. The remittances from all Development Assistance Committee (DAC) donors within the OECD to developing nations were $190 billion. Among these donor countries, the United States was the biggest donor in terms of total assistance to developing countries at $165.2 billion (remittances, philanthropy, and private capital investments), but ranked 12th in terms of assistance as a percentage of gross national income (GNI) at 1.13%. On the recipient side, the largest receivers were India, China, and Mexico. The development of new funding mechanisms and technologies leads to increasingly blurred lines among remittances, philanthropy, and private capital investment, since varied percentages of these types of financing provide at least some direct or indirect charitable or philanthropic support to both individuals and households. Government support from the now 24 OECD nations that comprise the DAC generally fail to meet the United Nations Official Development Assistance Target of 0.7% of GNI per participating nation as only Sweden, Norway, Luxembourg, Denmark, and the Netherlands meet or slightly exceed this figure. The United States despite being the largest giver at $30.46 billion of direct government aid is sixth from the bottom in terms of percentage of GNI (Organization for Economic Cooperation and Development, 2013).

Emerging Trends in Philanthropy Research

Interest in research and education on charity and philanthropy is growing globally with active centers for research throughout Europe, Australia, New Zealand, and North America as well as emerging sites in Asia, Africa, Central and South America, and the Pacific. The Nonprofit Academic Centers Council serves as a voluntary federation of many of these centers. Research in charity and philanthropy is supported by private banks, corporations, government agencies, international agencies, private and public foundations, nonprofits, and individuals (e.g., Barclays Wealth, 2010 (global study); Budani, et al., 2010 (Latin America); Morris et al., 2011 and Mahmood and Santos, 2011 (Asia); Sheth and Singhal, 2011 (India); Fauzia, 2013 (Indonesia); Harrow et al., 2014 (global perspective from principally European researchers)). Large repositories of information regarding charity and philanthropy can be found in organizations such as the Foundation Center (New York City); Indiana University Lilly Family School of Philanthropy; the Council on Foundations (USA); the World Initiatives for Grantmaker Support, Inc. (WINGS, Brazil); and the European Foundation Centre (EFC, Brussels). Organizations like these serve as key connectors among researchers, research networks, data sources, practitioners, corporations, governments, foundations, and other organizations. The expansion of these

networks over the past decade in particular has prompted large-scale comparative work and many smaller comparative research projects.

Undergraduate and especially graduate degree programs up through the level of PhD specializing in charity and philanthropy can be found in many universities, often in different departments and with variable rigor. Annual academic conferences specializing in this arena include the meetings of the International Society for Third Sector Research, Association for Research on Nonprofit Organizations and Voluntary Action, Voluntary Action History Society, and the EMES European Research Network. A study of academic publishing trends in philanthropic studies concluded that the field is both growing and maturing in terms of the depth and breadth of the questions being asked and the amount and complexity of research undertaken (Graddy et al., 2011; cf Bekkers and Wiepking, 2011; Oppenheimer and Olivola, 2010). Due to the inherent multidisciplinary nature of charity and philanthropic research, literature is published by scholars around the globe in a huge variety of disciplines and specialties, particularly in business (chiefly management and organizational behavior), social work, sociology, psychology, history, economics, law, political science, religious studies, public policy studies, evaluation, and development studies. The lack of a comprehensive system of nomenclature can make locating all relevant research regarding the sector quite challenging.

See also: Afrocentric Approaches to Social Work; Charity and Philanthropy, Economics of; Volunteerism, Psychology of.

Bibliography

Adam, T., 2009. Buying Respectability: Philanthropy and Urban Society in Transnational Perspective, 1840s to 1930s. Indiana University Press, Bloomington, IN.

Adelman, C., Miller, P., Norris, J., Marano, K., Spantchak, Y., 2012. Index of Global Philanthropy and Remittances 2012. Center for Global Prosperity, Hudson Institute, Washington, DC.

Barclays Wealth, November 2010. Global Giving: The Culture of Philanthropy. Barclays Bank PLC in Co-operation with Ledbury Research, London.

Barnett, M., Stein, J.G., 2012. Sacred Aid: Faith and Humanitarianism. Oxford University Press, New York.

Bekkers, R., Wiepking, P., 2011. A literature review of empirical studies of philanthropy: eight mechanisms that drive charitable giving. Nonprofit and Voluntary Sector Quarterly 40 (5), 924–973.

Bishop, M., Green, M., 2008. Philanthrocapitalism: How the Rich Can Save the World. Bloomsbury Press, New York.

Bönke, T., Massarrat-Mashhadi, N., Sielaff, G., 2013. Charitable giving in the German welfare state: fiscal incentives and crowding out. Public Choice 154 (1), 39–58.

Bradley, K., 2009. Poverty, Philanthropy and the State: Charities and the Working Classes in London. Manchester University Press, Manchester, UK.

Brassard, C., Sherraden, M.S., Lough, B.J., 2010. Emerging Perspectives on International Volunteerism in Asia. Research Paper for IVCO 2010, "Innovation and Challenges in International Volunteerism & Development – An Asian Perspective" October 3–6, 2010. International FORUM on Development Service and Singapore International Foundation, Singapore.

Bremner, R., 1988. American Philanthropy. University of Chicago Press, Chicago, IL.

Brest, P., 2012. A decade of outcome-oriented philanthropy. Stanford Social Innovation Review 10 (2), 42–47.

Budani, A., Luft, M., López, C., 2010. Main Philanthropy Trends in Latin America. Inter-American Development Bank and AVINA Foundation, Washington, DC.

Castles, F.C., Leibfried, S., Lewis, J., Obinger, H., Pierson, C. (Eds.), 2010. The Oxford Handbook of the Welfare State. Oxford University Press, New York.

Charities Aid Foundation, 2012. The World Giving Index 2012: A Global View of Giving Trends. Author, West Malling, Kent, UK.

Constantelos, D.J., 1991. Byzantine Philanthropy and Social Welfare, second ed. Aristide D. Caratzas Publisher, Athens, Greece.

Constantelos, D.J., 1992. Poverty, Society and Philanthropy in the Late Medieval Greek World. Aristide D. Caratzas Publisher, Athens, Greece.

CSR Asia, 2010. Strategic Philanthropy: Unlocking Entrepreneurial Potential. Credit Suisse AG, Zurich, Switzerland. Philanthropy White Paper 01.

Davis, D., 1999. The Problem of Slavery in the Age of Revolution, 1770–1823. Oxford University Press, New York.

Fauzia, A., 2013. Faith and the State: A History of Islamic Philanthropy in Indonesia. E. J. Brill, New York.

Giving USA Foundation, 2013. Giving USA: The Annual Report on Philanthropy for the Year 2012. Author, Chicago, IL.

Graddy, E.A., Ferris, J.M., Sohn, Y.J., 2011. The Evolution of Research on Philanthropy and Nonprofit Organizations as an Intellectual Field: An Analysis of Leading Journals. The Center on Philanthropy & Public Policy, University of Southern California, Los Angeles, CA. Research Paper – 35, August.

Gunderman, R., 2008. We Make a Life by What We Give. Indiana University Press, Bloomington, IN.

Hamel, G.H., 1990. Poverty and Charity in Roman Palestine First Three Centuries CE. University of California Press, Berkeley, CA.

Hanifan, L.J., 1916. The rural school community center. Annals of the American Academy of Political and Social Science 67, 130–138.

Hansmann, H., 1980. The role of nonprofit enterprise. The Yale Law Journal 89 (5), 835–901.

Harrow, J., Jung, T., Phillips, S. (Eds.), 2014. The Routledge Companion to Philanthropy. Routledge, New York.

Hilton, M., Crowson, N., Mouhot, J.-F., McKay, J., 2012. A Historical Guide to NGOs in Britain: Charities, Civil Society and the Voluntary Sector since 1945. Palgrave Macmillan, New York.

Ilchman, W.F., Katz, S.N., Queen II, E.L. (Eds.), 1998. Philanthropy in the World's Traditions. Indiana University Press, Bloomington, IN.

Indiana University Lilly Family School of Philanthropy, 2012. The 2012 Bank of America Study of High Net Worth Philanthropy. Author, Indianapolis, IN.

John, R., 2012. The Emerging Ecosystem of Entrepreneurial Social Finance in Asia. Asia Centre for Social Entrepreneurship and Philanthropy, National University of Singapore Business School, Singapore. Working Paper No. 1, April.

John, R., Tan, P., Ito, K., 2013. Innovation in Asian Philanthropy: Entrepreneurial Social Finance in Asia. Asia Centre for Social Entrepreneurship and Philanthropy, National University of Singapore Business School, Singapore. Working Paper No. 2, April.

Kaufmann, Franz-Xaver, 2012. European Foundations of the Welfare State. Berghahn Books, New York.

Kelley, T., 2008. There's no such thing as 'bad' charity. In: Kass, A. (Ed.), Giving Well, Doing Good: Readings for Thoughtful Philanthropists. Indiana University Press, Bloomington, IN, pp. 245–247.

Kerlin, J.A. (Ed.), 2009. Social Enterprise: A Global Comparison. University Press of New England for Tufts University Press, Lebanon, NH.

Kramer, M.R., 2009. Catalytic philanthropy. Stanford Social Innovation Review 7 (4), 30–35.

Kristol, I., 2008. Foundations and the sin of pride: the myth of the third sector. In: Kass, A. (Ed.), Giving Well, Doing Good: Readings for Thoughtful Philanthropists. Indiana University Press, Bloomington, IN, pp. 260–267.

Loewenberg, F.M., 2001. From Charity to Social Justice: The Emergence of Communal Institutions for the Support of the Poor in Ancient Judaism. Transaction Publishers, New Brunswick, NJ.

Lohmann, R.A., 1992. The commons: a multidisciplinary approach to nonprofit organization, voluntary action, and philanthropy. Nonprofit and Voluntary Sector Quarterly 21 (3), 309–324.

Mahmood, M., Santos, F., 2011. UBS-INSEAD Study on Family Philanthropy in Asia. UBS AG Philanthropy Services, Singapore.

Menninger, R., 2008. Foundation work may be hazardous to your mental health: some occupational dangers of grantmaking (and grant receiving). In: Kass, A. (Ed.), Giving Well, Doing Good: Readings for Thoughtful Philanthropists. Indiana University Press, Bloomington, IN, pp. 127–135.

Morris, A., Line, D., Vadaketh, S., May 2011. Something's Gotta give – the state of philanthropy in Asia. A report from *The Economist* Intelligence Unit.

Oppenheimer, D.M., Olivola, C.Y. (Eds.), 2010. The Science of Giving: Experimental Approaches to the Study of Charity. Psychology Press, New York.

Organization for Economic Cooperation and Development, 2013. Official Development Assistance. April 2013 Update www.oecd.org (accessed 03.07.13.).

Parker, J., Vissing-Jorgensen, A., 2009. Who bears aggregate fluctuations and how? American Economic Review 99 (2), 399–405.

Payton, R., Moody, M., 2008. Understanding Philanthropy in Meaning and Mission. Indiana University Press, Bloomington, IN.
Portes, A., 1998. Social capital: its origins and applications in modern sociology. Annual Review of Sociology 24, 1–24.
Putnam, R., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon & Schuster, New York.
Rosenwald, J., 2008. Principles of public giving. In: Kass, A. (Ed.), Giving Well, Doing Good: Readings for Thoughtful Philanthropists. Indiana University Press, Bloomington, IN, pp. 185–193.
Ruswick, B., 2013. Almost Worthy: The Poor, Paupers, and the Science of Charity in America, 1877–1917. Indiana University Press, Bloomington, IN.
Salamon, L., 1987. Of market failure, voluntary failure, and third-party government. toward a theory of government-nonprofit relations in the modern welfare state. Nonprofit and Voluntary Sector Quarterly 16 (1–2), 29–49.
Salamon, L., 2010. Putting the civil society sector on the economic map of the world. Annals of Public and Cooperative Economics 81 (2), 167–210.
Salamon, L., Anheier, H.K., 1996. Defining the Nonprofit Sector: The United States. Johns Hopkins University Institute for Policy Studies, Baltimore, MD.
Salamon, L., Anheier, H.K., 2006. The nonprofit sector in comparative perspective. In: Powell, W., Steinberg, R. (Eds.), The Nonprofit Sector: A Research Handbook, second ed. Yale University Press, New Haven, CT, pp. 89–116.
Schambra, W., 2008. The evaluation wars. In: Kass, A. (Ed.), Giving Well, Doing Good: Readings for Thoughtful Philanthropists. Indiana University Press, Bloomington, IN, pp. 301–306.
Sheth, A., Singhal, M., 2011. India Philanthropy Report 2011. Bain & Company, Mumbai, India.
Singer, A., 2008. Charity in Islamic Societies. Cambridge University Press, New York.
Smith, D.R., Stebbins, R., Dover, M., 2006. A Dictionary of Nonprofit Terms and Concepts. Indiana University Press, Bloomington, IN.
Smith, J.H., 2009. The Art of Doing Good: Charity in Late Ming China. University of California Press, Berkeley, CA.
Steinberg, R., 2006. Economic theories of nonprofit organizations. In: Powell, W., Steinberg, R. (Eds.), The Nonprofit Sector: A Research Handbook, second ed. Yale University Press, New Haven, CT, pp. 117–139.
Sulek, M., 2009a. On the classical meaning of philanthropia. Nonprofit and Voluntary Sector Quarterly 39 (3), 385–408.
Sulek, M., 2009b. On the modern meaning of philanthropy. Nonprofit and Voluntary Sector Quarterly 39 (2), 193–212.
Tennant, M., 2007. Fabric of Welfare: Voluntary Organisations, Government, and Welfare in New Zealand, 1840–2005. Bridget Williams Books, Wellington, NZ.
Vamstad, J., von Essen, J., 2013. A part of the problem or the solution? Charitable giving in the Swedish welfare state. Paper Presented at the Annual Meeting of the ISTR 10th International Conference, Universita Depli Studi Di Siena, Italy. http://citation.allacademic.com (accessed 03.07.13.).
Veyne, P., 1992. Bread and Circuses: Historical Sociology and Political Pluralism. Penguin Books, New York.
Weisbrod, B.A., 1991. The Nonprofit Economy. Harvard University Press, Cambridge, MA.
Yunus, M., Weber, K., 2011. Building Social Businesses: The New Kind of Capitalism that Serves Humanity's Most Pressing Needs. Public Affairs, New York.
Zunz, O., 2011. Philanthropy in America: A History. Princeton University Press, Princeton, NJ.

Relevant Websites

www.philanthropy.iupui.edu – Indiana University Lilly Family School of Philanthropy is the largest research entity globally focused on research and education on charity and philanthropy. The trustees of Indiana University in June 2012 voted to further expand the Center on Philanthropy into a new School of Philanthropy due to the depth of research and education activity occurring and the potential for even greater national and global impact (accessed 03.07.13.).

www.arnova.org – The Association for Research on Nonprofit Organizations and Voluntary Action (ARNOVA) is an outgrowth of a membership organization of researchers interested in research on nonprofit organizations, voluntary action, philanthropy, and civil society that formed in 1971. It sponsors an annual conference and publishes the Nonprofit and Voluntary Sector Quarterly (accessed 03.07.13.).

www.cof.org – The Council on Foundations in Arlington, VA, USA, is the largest membership organization globally for foundations, corporations, or philanthropic entities that provide charitable support. It has a large number of interest groups, specialized resources, and events (accessed 03.07.13.).

www.emes.net – EMES – The European Research Network, headquartered at the University of Liége, Belgium, is the leading European membership organization for researchers on matters related to the Third Sector. It welcomes collaboration with non-member researchers or centers (accessed 03.07.13.).

www.efc.be – The European Foundation Centre, established in Brussels in 1989, is an organization formed by European foundations that facilitates the work of foundations and corporate partners across Europe with governments, civil society, and other global partners and serves as a creator, advocate, and repository of important information for the sector throughout Europe (accessed 03.07.13.).

www.foundationcenter.org – The Foundation Center in New York City, established in 1956, is the leading source for philanthropy statistics for the United States and increasingly the world dealing with foundation activities, expenditures, and impacts among other items. The Center operates a large publishing house and provides research, education, and training programs to advance the field of philanthropy (accessed 03.07.13.).

www.icnl.org – The International Center for Not-for-Profit Law, headquartered in Washington, DC, with ten other offices, is a global center of resources related to the intersection of law and the nonprofit sector. It publishes the International Journal of Not-for-Profit Law (accessed 03.07.13.).

www.istr.org – The International Society for Third Sector Research, established in 1992, includes attention to charity and philanthropy within its purview with a particular focus on its impact on civil society and the nonprofit sector. It publishes Voluntas – International Journal of Voluntary and Nonprofit Organizations (accessed 03.07.13.).

www.wingsweb.org – The World Initiatives for Grantmaker Support Inc. (WINGS) is a federation of nearly 150 grantmaker associations and philanthropic support organizations around the world. WINGS seeks to coordinate global information and activities regarding charity and philanthropy and influence international law and agencies regarding involvement of the nonprofit sector in global affairs (accessed 03.07.13.).

Chemical Science: History and Sociology

David M Knight, University of Durham, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 1659–1663, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Metallurgy, ceramics, and pharmacy are ancient chemical technologies, but chemical science developed in the West from the seventeenth century. Antoine Lavoisier in the late eighteenth century, describing his work as revolutionary, introduced new language and new theory; and its exciting links to electricity made chemistry the fundamental science of the early nineteenth century. By mid-century, graduate schools like Justus Liebig's journals, institutes, and connections with industry had inaugurated professional chemistry. Its first major international conference, settling formulae, was held in 1860; by then crude vitalism had been abandoned, and chemistry was separating into specialisms (inorganic, organic, and physical) and subspecialisms. From about 1900, chemistry has been perceived as reduced to physics, a service science rather than a fundamental one; and its achievements in fertilizers, pesticides, explosives, and drugs are looked at askance. But chemists are ubiquitous, essential in scientific teamwork; and the creation of new molecules unknown in nature has gone on apace. The history of chemistry used to be written by practitioners interested in finding the roots of modern ideas, but has been largely taken over by historians more interested in contexts.

Introduction

The chemical sciences are concerned with specific kinds of matter and their transformations. Chemistry is very ancient, going back into remote prehistory with cookery, the preparation of drugs and dyes, the baking of clay into ceramics, and metalworking. Its evolution into a science, where theory guides practice, and into a profession, with formal courses and qualifications, happened in the eighteenth and nineteenth centuries (Brock, 1992). The boundaries of chemistry, notably with physics and biology, are social constructions that have varied in different times and places, especially in recent decades.

Ancient Technologies

From very remote times, people have been using techniques and processes that we would call chemical, involving careful control as part of a craft or art passed from father to son, mother to daughter, or master to apprentice. Indeed, the word 'chemistry' is supposed to have come from an ancient Egyptian word *chem* meaning earthy: the Arabic definite article *Al* was added to yield our 'alchemy,' and dropped to give *chymistry* and then by 1700, 'chemistry.' Early technologies, culminating in triumphs such as the making of porcelain and Japanese swords, include features we would regard as magical; but since the course of chemical reactions depends upon the purity of components, those involving natural products are hard to predict and to repeat. Recourse to prayers, incantations, and curious additives should not amaze us.

Metallurgy, Alchemy, and Pharmacy

Alchemy, with its objective of converting base metals into gold, which chemist–historians portrayed as absurd or dishonest, was not unreasonable. Nature was believed to be perfecting metals within her womb, and the alchemist was simply speeding up the process. If everything was composed, as Aristotle believed, of the four elements earth, water, air, and fire in different proportions, then changing these ratios would transform one substance into another, so that lead might become gold. If, alternatively, Democritus and Epicurus were correct in believing that in appearance there were colors, smells, and tastes, but in reality atoms and void; then because lead, gold, and everything is made up of different arrangements of these ultimately similar atoms (or 'corpuscles'), again conversions should be possible.

Alchemy began in Egypt and Babylonia, and also in China: with emphasis both upon making gold (or maybe something resembling it) using an elixir to expedite the process, and of prolonging and enhancing life by giving humans the noble and permanent qualities of gold. Pharmacy grew out of trial and error, but in the West the maverick Swiss doctor calling himself Paracelsus (1493–1541) brought alchemy into it. He introduced metallic compounds into previously herbal medicine, notably for the treatment of the new disease, syphilis, which was ravaging Europe. He publicly burnt the books of the great Greek physician, Galen, and saw chemical study as essential for medicine. His career outraged the medical establishment, but the powerful and dangerous new remedies proved irresistible to doctors and patients, and medical schools became centers for chemistry.

The First Chemical Theories

Until the mid-twentieth century, it was believed that alchemy was abandoned by the rational thinkers of the Scientific Revolution; especially Robert Boyle (1627–91) and Isaac Newton (1642–1727). Close examination of their manuscripts (Principe, 1998) shows that both of them were in fact adepts, copying out and trying alchemical recipes, and believing that they were well on the way to a transmutation. But they were also adherents to the atomic view of matter, seeing hard and indestructible corpuscles or particles as fundamental. These

formed very stable primary mixts, such as iron, gold, or sulfur, which in turn combined with each other. Unlike gravity, which was universal, chemical affinity was elective: some substances reacted together, others did not. J.W. Goethe wrote a novel, *Elective Affinities* (1809) exploring chemical and human bonding; chemists in eighteenth-century Germany and Sweden (the center of chemical activity) drew up tables of affinity in attempting to predict the outcome of reactions.

From Germany also came the first chemical paradigm. G.H. Stahl (1660–1734) proposed that everything, which would burn contained *phlogiston* (Greek, flammable): this idea brought order into chemical understanding, whereas atomic ideas were vague and untestable. Moreover, in Germany, Lorenz Croll in 1778 began the first chemical journal, *Chemische Annalen*, bringing into being a chemical community there (Hufbauer, 1982). His example was followed in France and England by Antoine Lavoisier and William Nicholson.

Lavoisier's Revolution

Lavoisier (1743–94) was a wealthy man, prominent in the privatized tax system of France; his spare time he devoted to chemistry, in a splendidly equipped laboratory. Becoming a member of the Royal Academy of Sciences, the small salaried body charged with scientific research, he resolved to reform the language and theory of chemistry. As in Carl Linnaeus' botany, names should be international, clear, and free from changeable theory: while *phlogiston* should be replaced as incoherent. Stahl saw *phlogiston* emitted in burning; Lavoisier by contrast (in a classic paradigm shift) saw something absorbed from the air, leading to an increase in weight. He drew upon the work of Joseph Priestley (1733–1804), who had isolated 'vital' or 'eminently respirable' air in a British tradition of work on gases. Lavoisier christened this substance 'oxygen' (Greek, sour) because he believed that it was also responsible for acidity (generalizing from analyses of nitric and sulfuric acids). Water was a compound of oxygen with another gas, hydrogen: such elements were the basis of chemistry, rather than the hypothetical corpuscles which might concern physicists, or the earth, water, air, and fire with which Priestley's friend Thomas Jefferson (1743–1826) structured his book on Virginia. In 1794, during Robespierre's Reign of Terror, Lavoisier was executed as a tax profiteer, while the left-wing views of Priestley (who continued to disagree with him over *phlogiston*) led to his exile in Pennsylvania. But their new and exciting chemistry survived and prospered (Bensaude-Vincent and Abbri, 1995; Knight and Kragh, 1998).

Electricity and Chemistry

In 1799, Alessandro Volta (1745–1827) showed that electricity was generated when two metals were dipped into water; there was no need for any animal tissue, as Luigi Galvani (1737–98) had supposed. His paper was an alarm bell, as Humphry Davy (1778–1829) put it, and chemists everywhere repeated and extended the experiments. But results were confusing until in 1806 Davy did the careful experiments confirming his intuition that pure water is decomposed electrically into oxygen and hydrogen only. Just as Newton had found that gravity was the force behind planetary motions, so Davy inferred that electricity and chemical affinity were manifestations of one power. In 1807, he used this insight in isolating the light and reactive metals potassium and sodium, and (putting Britain back on the chemical map) went on to demonstrate, with chlorine, that Lavoisier had been wrong about acidity.

Davy had been appointed to the newly founded Royal Institution in London's fashionable West End, where he proved himself to be an enormously popular lecturer, making professing a performance art (Golinski, 1992; Knight, 1998). The fees that men and women paid to join the Royal Institution, and hear Davy lecture, supported a research laboratory in the basement. Davy became one of the first people in Britain to make a living out of chemical research, which had previously perforce been a hobby for an aristocrat like Boyle, a minister of religion like Priestley, or a doctor like Galvani. At the Royal Institution, Davy trained (in a kind of informal apprenticeship) his successor, Michael Faraday (1791–1867), and the pursuit of Davy's insight that chemical affinity was electrical continued there.

With Lavoisier, chemistry had acquired an exact language, closer to algebra than to the evocative terms of the alchemists; and it included testable theories, for example, of acidity. It was the science of the secondary qualities – of colors, smells, and tastes – and it promised to be useful (chlorine for disinfecting and bleaching, for example, and oxygen for chest complaints). It also is proved popular everywhere. With its connection to electricity, it became the dynamic fundamental science, concerned not just with matter but also with force; there was as yet no unified science of physics. Mechanical explanations seemed shallow, while chemistry's connections with heat, light, and electricity went deep.

A Mature Science

J.J. Berzelius (1779–1848) in Sweden used the unsystematic Davy's insight to create a structure for chemistry: 'dualism' based on the idea that every compound had a positive and a negative part. He also picked up John Dalton's idea that each element was composed of identical atoms that differed from those of other elements: Berzelius arranged these in an electrochemical series from oxygen, the most negative, to potassium. The number of elements known steadily grew through the century with improvements in chemical analysis.

Berzelius trained a number of chemists by having them to stay in his house, where Anna the housekeeper washed up dishes and flasks. But in the 1820s, Justus Liebig (1803–73) at the University of Giessen launched the first graduate school for turning out a stream of chemists with PhD degrees (Brock, 1997; Morrell, 1997). Liebig's success depended upon his having perfected apparatus for analyzing organic compounds; his students usually did their research on some natural product, and published it in the journal that became called *Liebigs Annalen* after its editor. They found jobs, particularly in the dye industry (Fox and Nieto-Galan, 1999) and in pharmacy, both of which were becoming based in science rather than craft skills; many went to England, a rich country with a poor educational system. With the collapse of Napoleon's empire in

1815, the University of Berlin had emerged dedicated to research and teaching (*Wissenschaft und Bildung*), and the various German states began to compete in their opera houses and universities. They followed Giessen, building better laboratories and bidding for star chemists. Schools began teaching chemistry, textbooks were needed (Lundgren and Bensaude-Vincent, 2000), and academic careers opened up in a field now largely separated from medicine. Universities in Britain and the USA followed the German model, usually demanding German research experience from professorial candidates.

In the 1850s, chemists could agree about what things were made of, but not about formulae. Dalton had supposed that water must have the simplest possible formula, HO; Davy and others, notably Amadeo Avogadro (1776–1856), opted for the H_2O formula because two volumes of hydrogen combine with one of oxygen. An atom of oxygen thus weighed either 8 or 16 times as much as one of hydrogen, and such uncertainties ran through the whole list of elements. In 1860, August Kekule (1829–96), a pioneer in working out chemical structures such as that of benzene, called for an end to this confusion through an international conference, which met in Karlsruhe. It was poorly organized, but afterward chemists came to accept the reformulation of Avogadro's arguments by Stanislao Cannizzaro (1826–1910). With agreed atomic weights, tabular arrangement of the elements became possible; and the most successful was the periodic table of Dmitri Mendeleev (1834–1907).

Mendeleev's predictions of the properties of some hitherto undiscovered elements were startlingly accurate; and with the table (as he hoped) the student had to remember fewer brute facts. From its position in the table an element's properties could be known. It is striking that so many bright ideas, from Dalton via Cannizzaro to Mendeleev, came from people on the periphery rather than in the great scientific centers.

The Fragmentation of Chemistry

In death, we rot: for we (and animals and plants) are then subject to chemical reactions that go differently than while we are alive. Most people believed in a vital force that maintained life. It is claimed that Friedrich Woehler (1800–82), pupil of Berzelius and friend of Liebig, destroyed this vitalism when in 1828 he synthesized urea. In fact, the chief interest in this reaction was that ammonium cyanate and urea turned out to have the same atomic constitution: their different properties were the result of different arrangements (Brooke, 1995). So the story has more to do with understanding molecular structure; but the synthesis, and the work of Liebig and his students in analysis, showed that no gulf separated organic and inorganic worlds. Nevertheless, by 1848, when Berzelius died, it was clear that dualism did not fit organic compounds well, and as the chemical community grew it was convenient to separate organic chemistry, based upon carbon, from the inorganic branch. The expansion of universities led to new professorships and laboratories devoted to the specialism of organic chemistry, from which in the twentieth century emerged biochemistry.

Chemists had relied upon balances, test tubes, condensers, blowpipes, and other items of apparatus that were difficult to manipulate. The chemist had to think with his (or occasionally her) fingers and was proud of skills in glassblowing. Chemistry was essentially experimental, exciting, attractive, and often dangerous. Then in 1860, came collaboration between Robert Bunsen, inventor of the controllable gas burner, and the physicist G.R. Kirchhoff, who found that elements heated at a high temperature have characteristic spectra. Analysis could be done by physical methods, and this optical spectroscopy was the first of what is now an armory of such techniques that has transformed the appearance of chemical laboratories (Morris and Travis, 1997: pp. 715–740).

At about the same time, the new science of thermodynamics, based on energy and its transformations, brought together into classical physics sciences which had been separate or had been part of the empire of chemistry. Davy and Faraday had been pioneers in what became a new specialism, physical chemistry, investigating energy changes in reactions, and the mechanisms, rates, and reversibility of processes. The leaders here were Wilhelm Ostwald (1853–1932) and J.H. Van't Hoff (1852–1911) who launched a journal and promoted academic positions and laboratories. The new profession of chemical engineering was closely linked to the rise of physical chemistry. Whereas, early in the nineteenth century, chemists had been called in only as consultants or troubleshooters when something went wrong, by the end of the century they were employed full time (Bud and Roberts, 1984). In industry, intellectual property belongs generally to the company and not the individual, and is secured by patents (Travis et al., 1998).

The Reduction of Chemistry

The nineteenth century was the heyday of chemistry, the golden age in which it came to maturity and seemed fundamental. The chemist and spectroscopist William Crookes (1832–1919) followed Faraday in studying cathode rays, but J.J. Thomson in 1897 identified them as composed of subatomic corpuscles, soon named 'electrons.' The subsequent nuclear atomic model of Ernest Rutherford (1871–1937) – for whom all science was physics or stamp collecting – and Niels Bohr (1885–1962) accounted not only for spectra, but also for the periodic table. Chemistry became a branch of physics (Nye, 1996); the properties of gold could in principle be calculated from data about protons, neutrons, and electrons, though in practice the chemistry laboratory is essential. This meant that chemistry lost its glamor; the chemist was as ubiquitous as ever, an essential member of the teams or groups so characteristic of twentieth-century science, but playing a service role (Knight, 1995).

The number of chemists has continued to grow, as has the number of new substances they have synthesized that had been unknown in nature. Davy wrote of the chemist being a godlike creator, and this creativity is nowadays celebrated by chemists such as Roald Hoffmann. The engineer or architect must remember the law of gravity, but to mourn that architecture has been reduced to physics would be absurd: like the poet or the painter, the chemist has to work within constraints, but that is a feature of life – indeed making creativity possible (Hoffmann and Torrance, 1993).

Hoffmann, born in Poland, surviving World War II, escaping to the USA, learning chemistry there, and doing

research that brought him a Nobel Prize in 1981, exemplifies another trend. Chemistry reached the West via Islam. In the eighteenth and nineteenth centuries, it was a European science, with Germany the most important center by 1900. Papers in German journals, and research experience in Germany, counted high in any pecking order; but already the USA was becoming a major power in science. There in 1916, G.N. Lewis proposed the electronic theory of chemical combination, much developed by his pupil Linus Pauling. Since 1945, the USA has been the center of things, making the English language and publication in US journals the key to prestige in research. Two World Wars, and Hitler's coming to power between them, are part of the reason for this; equally important has been American prosperity, itself dependent on science. Chemistry has steadily gone West.

The Status of Chemistry

Davy and Liebig (Rossiter, 1975) wrote famous books on agricultural chemistry, and in the nineteenth century, chemical fertilizers and pesticides were unequivocally welcomed in a Europe of food shortages. Lavoisier improved French gunpowder, and later chemists produced high explosives making possible engineering achievements and also formidable weapons. All these things were seen as benefits. National chemical societies, their academic and professional aspects sometimes in tension, were formed and enjoyed prestige (Russell et al., 1977). Although pollution from new chemical industries (as from older ones like tanning) was palpable and led to legislation, the expectation was that the chemists would be able to cure it. It did not happen: Rachel Carson's book *The Silent Spring* (1962) alerted the world to the dangers. So in the late twentieth century, despite successes such as plastics, and the array of new drugs available for medicine, chemistry is seen as boring and its applications as threatening. Chemists feel misunderstood and underappreciated.

Twentieth-century chemistry is dominated not only by universities in the Giessen tradition, but also by big-spending international companies with research laboratories, now turning toward biotechnology (Galambos and Sturchio, 1997; Kevles, 1997) and by the military. Research is carried on no longer by a Woehler or a Crookes, on their own or with assistants, but by teams of people possessing various skills. Chemistry has been taught in an impersonal way, with less hands-on experiment in a world more conscious of health and safety.

This is a strange eventful history, which until the mid-twentieth century was mainly written by participants who looked for progress. They had the advantage of being familiar with chemicals and apparatus, but professional historians of science have come to look more closely at contexts and careers. The history that emerges deserves to be known beyond the world of chemists.

Editor's Note

The history of the chemical sciences continues to converge with the histories of other fields, and not only through links with other traditional disciplines, such as physics and biology. Chemistry also is being folded into broad international initiatives that involve large public expenditures as well as efforts to attract commercial interest and capital. These initiatives include biotechnology and genomics, synthetic biology, nanotechnology and molecular electronics. Science, technology, and medicine are more closely integrated than ever, and major chemical companies such as Dupont and Monsanto have expanded into agricultural and medical biotechnology, and not without courting plenty of controversy. Recent research in science and technology reflects this trend. Historians of science continue to write about the chemical revolution, but sociological, anthropological and other studies of contemporary chemistry are pursuing chemistry through a thicket of initiatives involving multiple fields and corporate-university relations.

Recent studies on the history of chemistry include Ursula Klein and Wolfgang Lefèvre (2007), Lissa Roberts and Rina Knoeff (2007). For studies on the recent era of multidisciplinary initiatives and entrepreneurial research see: Steven Shapin (2008), Stephen Hilgartner (2012).

See also: Archaeometry; Biomedical Sciences, History and Sociology of; Biotechnology; Ceramics in Archaeology; Nanotechnology: Social and Cultural Aspects; Physical Sciences: History and Sociology; Science, New Forms of; Scientific Disciplines, History of; Universities and Science and Technology: United States.

Bibliography

Bensaude-Vincent, B., Abbri, F., 1995. Lavoisier in European Context. Science History, Canton, MA.
Brock, W.H., 1992. The Fontana History of Chemistry. Fontana, London.
Brock, W.H., 1997. Justus von Liebig: The Chemical Gatekeeper. Cambridge University Press, Cambridge, UK.
Brooke, J.H., 1995. Thinking about Matter. Ashgate Variorum, Aldershot, UK.
Bud, R., Roberts, G.K., 1984. Science versus Practice. Manchester University Press, Manchester, UK.
Fox, R., Nieto-Galan, A., 1999. Natural Dyestuffs. Science History, Canton, MA.
Galambos, L., Sturchio, F., 1997. The transformation of the pharmaceutical industry in the twentieth century. In: Krige, J., Pestre, D. (Eds.), Science in the Twentieth Century. Harwood, Amsterdam, pp. 227–252.
Golinski, J., 1992. Science as Public Culture: Chemistry and Enlightenment in Britain, 1760–1820. Cambridge University Press, Cambridge, UK.
Hoffmann, R., Torrance, V., 1993. Chemistry Imagined. Smithsonian, Washington, DC.
Hufbauer, K., 1982. The Formation of the German Chemical Community, 1720–1795. University of California Press, Berkeley, CA.
Hilgartner, S., 2012. Novel constitutions? New regimes of openness in synthetic biology. BioSocieties 7, 188–207.
Kevles, D., 1997. From eugenics to genetic manipulation. In: Krige, J., Pestre, D. (Eds.), Science in the Twentieth Century. Harwood, Amsterdam, pp. 301–318.
Knight, D.M., 1995. Ideas in Chemistry. Athlone, London.
Knight, D.M., 1998. Humphry Davy: Science and Power. Cambridge University Press, Cambridge, UK.
Knight, D.M., Kragh, H., 1998. The Making of the Chemist. Cambridge University Press, Cambridge, UK.
Krige, J., Pestre, D., 1997. Science in the Twentieth Century. Harwood, Amsterdam.
Klein, U., Lefèvre, W., 2007. Materials in Seventeenth-Century Science: A Historical Ontology. MIT Press, Cambridge, MA.
Lundgren, A., Bensaude-Vincent, B., 2000. Communicating Chemistry: Textbooks and Their Audiences, 1789–1939. Science History, Canton, MA.
Morrell, J., 1997. Science, Culture and Politics in Britain, 1750–1870. Ashgate Variorum, Aldershot, UK.

Morris, P.J.T., Travis, A.S., 1997. The role of physical instrumentation in structural organic chemistry. In: Krige, J., Pestre, D. (Eds.), Science in the Twentieth Century. Harwood, Amsterdam.

Nye, M.J., 1996. Before Big Science. Twayne, New York.

Principe, L., 1998. The Aspiring Adept. Princeton University Press, Princeton, NJ.

Rossiter, M., 1975. The Emergence of Agricultural Science: Justus Liebig and the Americans, 1840–1880. Yale University Press, New Haven, CT.

Russell, C.A., Coley, N.G., Roberts, G.K., 1977. Chemists by Profession. Open University Press, Milton Keynes, UK.

Roberts, L., Knoeff, R., 2007. The places of chemistry in eighteenth century England and The Netherlands. Special Issue of Ambix, Journal of the Society for the History of Alchemy and Chemistry.

Shapin, S., 2008. The Scientific Life: A Moral History of a Late Modern Vocation. University of Chicago Press.

Travis, A.S., Schröter, H.G., Homburg, E., Morris, P.J.T., 1998. Determinants in the Evolution of the European Chemical Industry, 1900–1939. Kluwer, Dordrecht.

Chiefdoms, Archaeology of

Laura L Junker, University of Illinois at Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Developed as a cross-culturally comparative concept within early to mid-twentieth century cultural evolutionary anthropology and processual archaeology, chiefdoms are defined as smaller scale complex societies that are sometimes intermediate to state development, with centralized decision-making hierarchies, some degree of hereditary social ranking, and economic centralization. In historically and ethnographically known chiefdoms, chiefs generally construct and maintain political power bases through various economic means (tribute collection, control over production and exchange, foreign trade monopolies), ideological means (ritual, myth, sacred landscapes), and military coercion. Archaeologists have recognized that these power strategies are materialized archaeologically in a variety of highly visible ways (such as monumental construction, settlement hierarchies, specialized prestige goods production and exchange, differential household wealth, and indicators of large-scale militarism), and a significant amount of archaeological research has been carried out on the origins, evolution, and organization of chiefdoms in premodern societies of the Near East, Africa, East Asia, Europe, Mesoamerica, South America, North America, and Oceania. However, there is considerable recent debate about the utility of 'chiefdoms' as a cross-culturally applicable evolutionary concept, with challenges to this archetype ranging from decoupling elements of 'chiefdom' structure, examining these societies at more diverse scales of analysis, and rejecting the 'chiefdom' model in favor of a variety of postprocessual approaches in archaeology. This article reviews this controversy and synthesizes current archaeological approaches to sociopolitical structure, political economy, ideology, warfare, and evolutionary transformations in chiefdoms.

The use of the term 'chiefdom' to refer to prestate complex societies is a relatively recent phenomenon, beginning with Oberg's (1953) historical classification of South and Central American societies and Sahlins' (1963) ethnographic work on Polynesian societies in the 1950s. Service (1962) included 'chiefdoms' a decade later in his neo-evolutionary categories of *band*, *tribe*, *chiefdom*, and *state*, and the concept become closely associated with the cultural evolutionary paradigm developed earlier in the century by such anthropologists as Julian Steward and Leslie White. Almost immediately, archaeologists began to apply the chiefdom concept to investigations of early societies such as the Late Neolithic European builders of Stonehenge and the Mississippian towns of the Southeastern US, which manifested some complexity but generally lacked the long-held archaeological indicators of 'civilization' (e.g., large urban centers, massive kingly tombs, written dynastic histories and legal codes, markets and monetary systems) (Flannery, 1972; Peebles and Kus, 1977; Renfrew, 1973; Wright, 1984). Chiefdoms are included in what archaeologists have increasingly referred to as 'middle range' societies, encompassing the traditional cultural evolutionary categories of both *tribes* and *chiefdoms* and emphasizing the great diversity of social and political structures in societies of intermediate scale and structural complexity between *bands* and *states* (Beck, 2003; Earle, 1991).

Chiefdoms and Cultural Evolutionary Theory

Chiefdoms have been defined by some archaeologists in primarily political terms, as smaller-scale complex societies with centralized decision-making hierarchies of one or two levels above the individual village (compared to three plus levels in states) (Sahlins, 1963; Wright, 1984). Chiefly leaders obtain their political authority through both ascription and performance (compared to largely hereditary kings), they have very generalized administrative roles (judicial, economic, ritual, military), they have few bureaucratic specialists (unlike the specialized bureaucracies of states), and they generally rely on kin-based alliances to structure political relations with subordinates (rather than the nonkin-based political institutions and formal legal codes of states) (Earle, 1991). Most archaeologists would expand the definition to include social ranking that has at least some hereditary component (unlike the achieved ranking of tribal societies and the almost wholly hereditary social classes of early states) and some degree of economic centralization (control over staple production, building and managing irrigation systems, tribute mobilization, control over the production and exchange of prestige goods, and/or foreign trade monopolies) (Brumfiel and Earle, 1987; Renfrew and Shennan, 1982). Chiefs generally maintain their political power not only through varying strategies for accumulating and disbursing material resources, but also through ideological means (manipulating cosmologies, ritual, and myth, and creating power-imbuing sacred landscapes) and military coercion (using strategies of threatened or actual force supported through controlled warrior cadres and military technologies) (Earle, 1997). Chiefdoms rarely exist or evolve in isolation, but instead are found as clusters of interacting *peer polities* that often share aspects of elite culture, have similar structures, and compete for regional supremacy (Renfrew and Cherry, 1986). Many chiefdoms are also part of larger *world systems* that link them to more developed states and empires (Peregrine, 1992)

Chiefdom-level societies have been identified by cultural anthropologists, archaeologists, and historians in many parts of the world and in many time periods (**Figure 1**), often as precursors to state development but sometimes as long-term, stable structures that do not transform into more complex states (*see* States and Civilizations, Archaeology of).

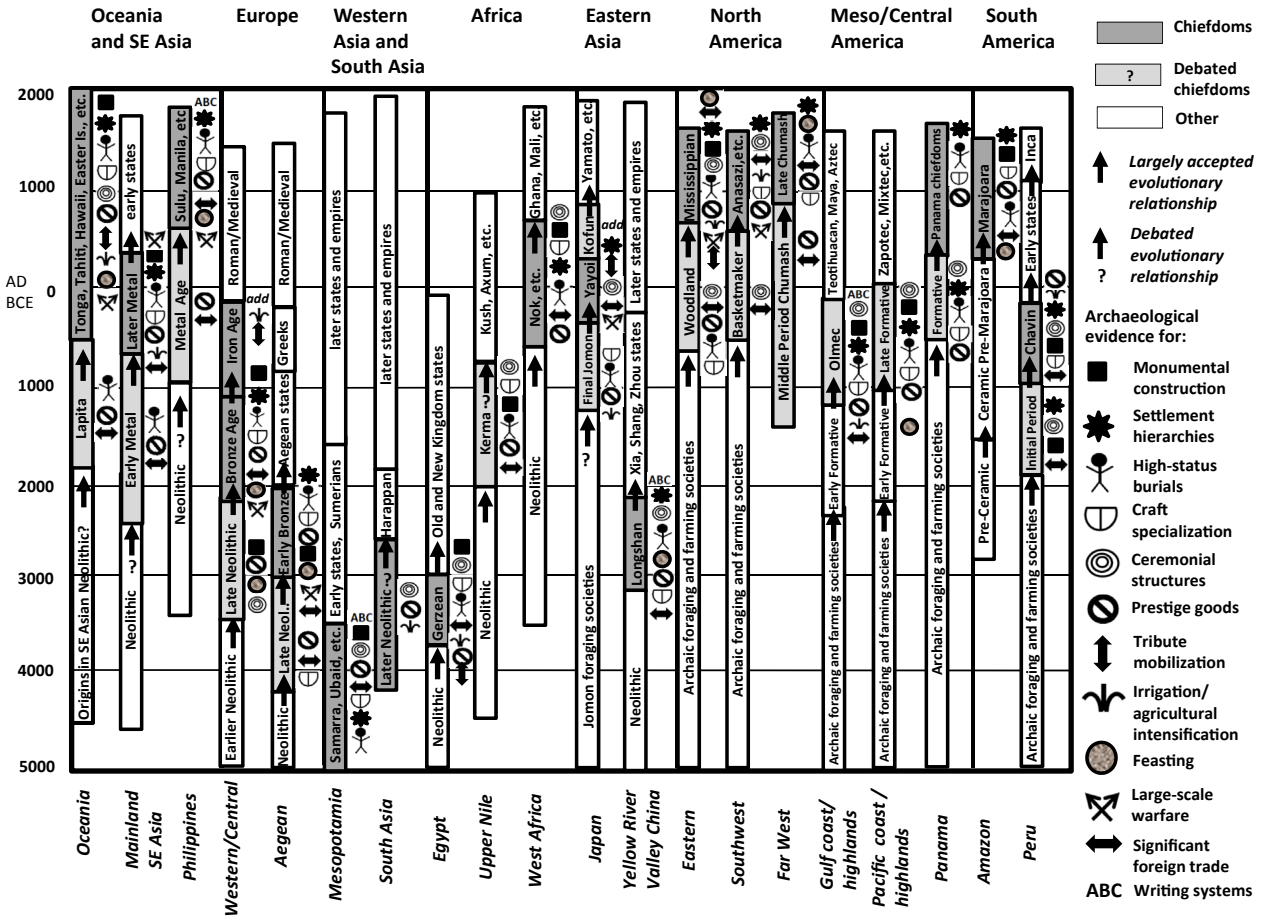

Figure 1 The chronological and geographic distribution of some archaeologically known chiefdoms with varying archaeological indicators. The 'debated chiefdoms' have been variously described as 'tribal' or 'ranked' societies, 'chiefdoms' or 'early states.' Evolutionary transitions marked with '?' indicate cases of disputed cultural/population continuity.

The Polynesian chiefdoms (*see* Pacific Islands, Archaeology of), the pre-Roman (Late Neolithic, Bronze Age, and Iron Age) societies of Europe (*see* Europe, Archaeology of), and eastern North American complex societies (*see* North America, Archaeology of) received the most intense archaeological study in the mid part of the twentieth century. However, archaeologists have more recently expanded their use of the chiefdom model to societies in the Near East, Africa, East Asia, South Asia, Mesoamerica, South America, and western North America (e.g., Feinman et al., 2000; Junker, 1999; McIntosh, 1999; Redmond, 1998; Rothman and Stein, 1994), although the 'evolutionary status' of many as chiefdoms is controversial (e.g., the Woodland Period societies of the Eastern US, the Late Neolithic societies of Europe, the Halaf and Samarran societies of the Near East, and many prestate African societies). Early in the development of the chiefdom concept, archaeologists identified a number of common material correlates of these types of societies (e.g., two to three level settlement hierarchies, monumental architecture, burials with hereditary status indicators, specialist-manufactured prestige goods, regional and interhousehold differences in wealth; see **Figure 1**). However, as more of these complex societies are intensively studied by archaeologists, it has become clear that material patterns, and presumably the ideological, political, social, and economic structures underlying them, vary substantially (Earle, 1997; Drennan and Peterson, 2006; McIntosh, 1999).

Critiques and Refinements of the Chiefdom Concept

A number of recent critiques of the chiefdom concept, and cultural evolutionary models in general, have emphasized the organizational diversity and historical uniqueness that are ignored when anthropologists attempt to fit societies in broadly defined classifications and assume uniform trajectories of development. As noted above, many scholars have pointed out the great deal of variability in societies designated as 'chiefdoms.' One response to this recognized diversity has been to subdivide prestate complex societies into developmentally distinct types, such as Carneiro's (1981) 'minimal,' 'typical,' and 'maximal' chiefdoms (based on complexity of political hierarchies), Johnson and Earle's (1987) 'simple' and 'complex' chiefdoms (based on scales of chiefly political, economic, and ritual control), and Renfrew's (1973) 'group-oriented' and 'individualizing' chiefdoms (based on differing strategies of lineage vs individual aggrandizement and social display).

Some scholars have emphasized that transformations in social, political, economic, and ideological structures that have previously defined chiefdoms do not always occur in tandem and should be studied independently as trajectories of change (Earle, 1997; Feinman and Neitzal, 1984). Other archaeologists have criticized the simplistic emphasis on hierarchical socio-political structures in chiefdoms (McIntosh, 1999), emphasizing *heterarchy* (Ehrenreich et al., 1995), or nonhierarchical or horizontally binding relations that operate along multiple dimensions (social, economic, political, ritual) to structure these societies.

While many archaeologists favor continual refinement of the chiefdom concept and its continued utility in comparative archaeological research, others have largely advocated its abandonment, citing the misplaced archaeological search for the roots of 'chiefdom' emergence and structure in systemic evolutionary processes, elite-centered viewpoints, and largely economic-focused notions of power (Pauketat, 2007; Yoffee, 1993). New approaches in the last several decades to the archaeological study of emerging complex societies include 'actor-based' methods of analysis focused on the agency of individuals and groups in shaping power relations, Marxist theories of political control that emphasize the ideological and symbolic bases for structure, and 'interpretive' approaches that recognize Western and capitalist bias in modeling ancient societies (Dobres and Robb, 2000; Hodder, 2012). Shared among these postprocessual critiques is the recognition that both structure and change in complex societies are contingent on highly variable social and environmental landscapes, as well as unique historical circumstances. In addition, there is an increasing emphasis on studying these societies at multiple scales, ranging from individual actors and interest groups within the society to expansive networks of interacting polities, using new spatial and quantitative tools such as *social network analysis*. Despite these criticisms, revisions, and outright rejections of the chiefdom concept, the term 'chiefdom' as a designate for these societies and key elements of political, social, economic, ideological, and military structure defining it, remain highly visible in archaeological work, cross-cultural comparisons, and theoretical dialogue. Therefore, subsequent sections examining these prominent aspects of chiefdom organization focus on the work of scholars who continue to find this a useful analytic concept.

Political and Social Features

Chiefdoms are characterized by partially inherited and partially achieved political leadership, reinforced through ideological manipulation, control of economic resources, and militarism (Earle, 1991). The political power of chiefs is often reflected materially in symbols of office, such as the feather capes of Polynesian chiefs (Kirch, 1984) and the stone maceheads of Gerzean Period predynastic polities in Egypt (Bard, 1994), and in monumental architecture which demonstrates a chief's ability to control and mobilize large labor forces (e.g., the henge monuments of Late Neolithic Europe and the monumental statuary of Formative Period Mesoamerica and Easter Islanders (Kolb, 1994; Renfrew, 1973)). Chiefly centers, as the focal point of political, economic, and religious activities of chiefs, are generally archaeologically distinguishable from subordinate villages by their greater size, centrality, and presence of monumental architecture, wealth objects, and specialized production locales. Multiple levels of political authority (*paramount chiefs* and local chiefs) typical of larger-scale chiefdoms are often archaeologically visible in regional settlement hierarchies of up to three levels (Cobb, 2003; Drennan and Peterson, 2006; Peebles and Kus, 1977; Wright, 1984).

A number of archaeologists have offered new conceptual frameworks for analyzing differing political power strategies in chiefdoms. One model offered by Blanton et al. (1996), contrasts *network* vs *corporate* strategies of political dominance, elaborating Renfrew's earlier distinction between individualizing and group-oriented chiefdoms. These differing strategies are viewed as part of political dynamics in all complex societies, but political actors in a particular society may emphasize one mode of control more than the other. In a network strategy, political actors try to create personal networks of dominance through the strategic distribution of portable wealth and symbolic capital (e.g., ritual potency and religious knowledge). The emphasis on individual aggrandizement is archaeologically manifested in lavish individual burials, elaborate household wealth display, and competitive presentational events such as ceremonial feasting. The highly conflictive nature of leadership and long-term instability of political configurations are evidenced in rapidly changing settlement hierarchies. In contrast, a society emphasizing a corporate political strategy disperses power across different groups and sectors of society through bureaucratic institutions promoting consensus, solidarity, and collective action, often reinforced through archaeologically visible public architecture, collective tombs, and unifying emblems (e.g., the henge monuments, copper axes, and collective mound tombs of Late Neolithic chiefdoms of Europe; the massive earthen burial mounds topped by wooden buildings and shell and copper valuables of Mississippian chiefdoms of North America) (Renfrew, 1973; Peebles and Kus, 1977). Beck's (2003) work on Mississippian chiefdoms in the American Southeast has framed these varying strategies similarly, but around issues of relative scale and mobility of economic resources and labor, with *coercive expansion* and *persuasive aggregation* working alternatively or even simultaneously to form *apical* (i.e., top-down) and *constituent* (bottom-up) forms of political authority.

In terms of social structure, chiefdoms invariably have some form of ascribed social ranking (Sahlins, 1963). While social status differences are characteristic of all human societies, social ranks in chiefdoms are at least partially inherited intergenerationally, they are given structural rigidity through behavioral taboos and symbolic expression, and in most cases, there is distinct economic advantage (differential access to resources) conferred on an elite stratum. Social stratification in chiefdoms is often manifested in archaeologically visible rank insignia (e.g., the shell gorgets of Mississippian chiefs in the American Southeast and the jade ornaments of China's Longshan chiefdoms) (Barnes, 1999; Cobb, 2003), differential wealth in households (e.g., imported Chinese porcelain fragments in fourteenth to sixteenth centuries Philippine chiefs' pile-house remains) (Junker, 1999); and varying size and complexity of residential architecture at

settlements (e.g., village house-clusters in Hawaiian chiefdoms; Kirch and Jones, 2003).

Ascribed rather than strictly achieved social ranking is most archaeologically evident in burial practices. Inherited status is generally expressed through variation in body positioning, body treatment, grave forms, and burial accompaniments, which cross cuts age and sex categories, making it archaeologically distinguishable from mortuary variability attributable to age differences, gender roles, and achieved statuses. Mortuary analyses based on this premise have been carried out most successfully at Mississippian, Chinese Longshan, Egyptian Gerzean, and Bronze Age and Iron Age European cemeteries (Bard, 1994; Barnes, 1999; Cobb, 2003; Gilman, 1981), where relatively large samples of well-preserved burials have allowed the identification of distinct social hierarchies. In cases where larger burial populations are lacking, archaeologists have frequently made an argument for the presence of elites based on finds of elaborate, often monumental chiefly burials (e.g., jade-yielding Formative Period Mesoamerican tombs, the Nok and other mounded tombs of West Africa, the Yayoi stone-lined crypts of Japan) (Barnes, 1999; McIntosh, 1999). Status-related dietary and health differences between elites and non-elites have also been studied through zooarchaeological and paleobotanical analyses of food remains at settlements, and through osteological assessments of nutritional stress and disease in burial populations (e.g., Blanton et al., 1996; Kirch and Jones, 2003).

Political Economy

Just as chiefdoms studied by archaeologists vary in political structure, they differ in the ways in which chiefs attempt to assert economic control in support of their governing institutions. Brumfiel and Earle (1987; Earle, 1997) have made a useful distinction between strategies of *staple finance* and *wealth finance* in chiefdoms. Staple finance involves the systematic levying of tribute payments (in the form of staple goods and/or labor) from subjugated commoners, with the revenue used to directly finance activities such as monument-building, warfare, trade, and ceremonial feasting that enhance the authority of the chief. Since control of land and its production is essential to staple finance, chiefdoms emphasizing this economic strategy generally have strongly developed institutions for land tenure (archaeologically manifested in boundary features such as walls and corporate monumental structures) and significant chiefly investment in agricultural intensification (archaeologically manifested in large-scale irrigation works, water control, and artificial terracing) (Kirch, 1984). Archaeological investigations suggest that the Hawaiian chiefdoms, pre-Inca South American chiefdoms, and many prestate Mesopotamian complex societies had political economies dominated by staple finance (Blanton et al., 1996; Brumfiel and Earle, 1987; Earle, 1997; Rothman and Stein, 1994).

Wealth finance involves the use of prestige goods or valuables as political currencies by chiefs and other elites, to cement alliances with other elites and to reward subordinates for loyalty and service (particularly associated with a network political strategy). Prestige goods vary according to socially defined standards of value and can be any rare or easily controlled objects, such as: jades in Longshan China (Barnes, 1999), gold in Central American chiefdoms (Helms, 1993), bronze in European chiefdoms (Earle, 1997), and shell and copper ornaments in eastern North American chiefdoms (Cobb, 2003). Politically charged wealth can be locally manufactured or obtained through foreign trade. Emerging elites control access to this political currency through support of *attached specialists*, highly skilled artisans who produce these goods wholly for elite consumption; by monopolizing valuable raw materials and technologies; by dominating trade routes; and by defining limited social contexts for their circulation, such as elite-controlled ceremonial feasting events (Brumfiel and Earle, 1987; Schortman and Urban, 2004). While specific chiefdoms tend to emphasize one means of economic control over another, chiefs tend to use elements of both staple finance and wealth finance in varying combinations, resulting in unique forms of political economy (Earle, 1997).

In addition to the emergence of attached specialists, the rise of chiefdoms is sometimes accompanied by the growth of *independent specialists*. Independent specialists are full-time craft production specialists who concentrate at chiefly centers due to economies of scale and produce more mundane household goods for an unrestricted set of consumers. While the archaeological recovery of large-scale craft workshops containing mass-production equipment (e.g., large kilns or smelting furnaces, molds, potter's wheels) is the most direct line of evidence for specialist production, such finds are rare. More often archaeologists have assessed craft production modes through analysis of the products themselves and their regional distribution. Specialized production by a limited number of concentrated, recurrently interacting, full-time craftsperson (either attached specialists or independent specialists) generally results in a more standardized product, as measured through object form, raw material, and/or decoration. In addition, specialist-produced prestige goods are largely restricted to elite centers, while household-produced domestic goods are widely dispersed throughout a region, and the socially unrestricted products of independent specialists are concentrated at, but not restricted to, regional centers (Brumfiel and Earle, 1987; Earle, 1997; Schortman and Urban, 2004).

Particularly in chiefdoms relying heavily on wealth finance, political relationships, and hierarchies of authority are reinforced through the continual circulation of prestige goods, most often in the context of bridewealth exchanges (e.g., gifts of porcelain in Southeast Asia and gold in Panama; Junker, 1999), political investitures (e.g., the presentation of feather capes in Polynesia; Earle, 1997), ritualized feasting, and other politically charged events. Of these exchange contexts, ritual feasting is particularly evident in the archaeological record. Archaeologists have identified feasts that were likely aimed at political integration and competitive status display using evidence, including: specialized feasting paraphernalia like bronze drinking vessels in the European Iron Age or large ceramic cooking vessels in Mississippian and Formative Mesoamerican chiefdoms; unusual food remains such as large pig concentrations in Longshan China; and the association of similar artifacts and ecofacts with ritual architecture like Mesoamerican ballcourts (Dietler and Hayden, 2001; Kirch and Jones, 2003).

The Ideology of Rulership

Ideologies are collective representations of the social and political order in particular societies, often encoded in myths, ceremonies, and various public performances, but also frequently materialized in archaeologically visible monumental architecture, iconography, and portable objects. In both state-level societies and chiefdoms, the creation of a dominant ideology and its imposition on the populace are an important basis for power (Earle, 1991; Pauketat, 2001). In many chiefdoms, elites developed their own language, dialect, lexicon, and/or notational systems to control the flow of esoteric knowledge and to restrict the performance of religious rites. The archaeological contexts in which written texts are found and their decipherable content (e.g., inscriptions found with monumental earthen works in Formative Period Mesoamerican centers, imported Sanskrit-based writing or Arabic scripts on bamboo in historic chiefdoms of island Southeast Asia) (Blanton et al., 1996; Junker, 1999) suggest that literacy was an elite prerogative which served to institutionalize cosmological notions and to legitimate the political and social domination of elites.

In the absence of written texts, archaeologists often infer shared cosmic orders and ruling ideologies from iconography on portable objects like were-jaguar figurines of Formative Period Mesoamerica, 'sun' motifs and other symbols on Mississippian pottery (Cobb, 2003), 'patron-deity' symbols on Egyptian Gerzean stone palettes and maces; from monumental constructions (e.g., the platform temple complexes of Polynesia, the henge monuments of Late Neolithic Europe (Earle, 1997)); from material symbols in burials (e.g., painted scenes in Japanese Yayoi mounded tombs (Barnes, 1999)); and from the spatial organization of chiefly centers (e.g., the layout of Cahokia and other Mississippian centers (Cobb, 2003; Pauketat, 2001)). An influx of exotic symbols representing foreign cosmologies or religious beliefs (e.g., Buddhist and Hindu art and architecture in Southeast Asia, widespread Olmec styles in Formative Period Mesoamerica, Egyptian themes in the early Nubian chiefly tombs of Kerma) has often been interpreted by archaeologists as a particularly effective strategy for chiefs to add to their monopolistic store of power-enhancing knowledge.

While past archaeological research has largely focused on ideologies of domination in chiefdoms, many researchers now recognize that the many individuals and social factions comprising these complex societies have distinct, and sometimes conflicting, world views based on personal experiences and interests. Increasingly, archaeologists are attempting to tease out diverse values and perspectives related to gender, class, ethnicity, occupation, and individuality (Dobres and Robb, 2000).

Warfare and Militarism

Anthropological studies have suggested that warfare in chiefdom-level societies differs from that in tribal societies by its expansionistic focus on acquisition of territory, resources, and captives and by its concentration of military power in the hands of chiefly leaders who use it to expand their political sway. In an early study of chiefdoms, Carneiro (1981) suggested that both the rise and expansion of chiefdoms were related to warfare. Under conditions of land shortages and population stress, militarily powerful leaders systematically assault settlements along polity boundaries and conquered enemies were incorporated into the expanding chiefdom, resulting in the coalescence of a polity of greater complexity as well as scale. However, ethno-historic work on chiefdoms with 'low-density' populations and abundant resources and land (e.g., in island Southeast Asia, the Caribbean, and parts of Africa), shows that large-scale and highly organized warfare and seemingly perpetual raiding is endemic but not generally related to resource stress and territorial conquest aims (Junker, 1999). In fact, in early historic island Southeast Asian chiefdoms, slave-raiding was the most prominent form of militarism, a practice that can be connected to labor shortages rather than population pressure on resources. Warfare seems to be an almost continuous occurrence in many chiefdoms, one of a number of forms of peer polity interaction that are politically transformative in a variety of ways. Political loyalties may be shifted to military victors who control the ideology of warrior prestige and who increase their political currency through captured labor and valuables, resulting in larger alliance networks (Earle, 1997; Kirch, 1984; Kolb and Dixon, 2002).

Archaeologists have focused on a variety of material evidence to document changes in the scale, intensity, and behavioral and ideological aspects of warfare in chiefdom-level societies (*see* Conflict and War, Archaeology of: Behavior and Social Organization). These include the development of fortifications or other defensive works, changes in regional settlement patterns (e.g., concentration of population in large centers, depopulation along polity boundaries, relocation of settlements to defensible positions), systematic destruction of power-symboling architecture (e.g., dismantling of monumental architecture, toppling of statuary, burning of chiefly centers)(Cobb, 2003; Kirch, 1984). Examples are the alternating rebuilding and burning of fortifications surrounding Iron Age European towns such as Heuneberg and the creation of 'no man's lands' between Mississippian centers (Cobb, 2003), and the defacement of sacred monumental statuary on Easter Island (Kirch, 1984). Osteological evidence from burials allows archaeologists to evaluate rates of violence in populations (e.g., scalping, skeletal trauma, embedded projectiles), the ritualized use of body parts in prewar and postwar ceremonialism (e.g., decapitation for trophy head-taking), and debilitating physical conditions and malnutrition related to prolonged exposure to military siege (Kolb and Dixon, 2002). Archaeological studies of burials also allow archaeologists to identify 'warrior insignia' associated with the development of specialized warrior classes and strong ideologies of warrior prestige, exemplified in the elaborate bronze horse fittings of Bronze Age European warriors and the gold-pegged teeth of island Southeast Asians warriors (Carman and Harding, 1999; Junker, 1999). The development of new warfare technologies associated with escalating military competition can be traced archaeologically in intensified metals mining, 'hoards' of weaponry stored for battle (e.g., sling stones at Polynesian sites), labor-saving equipment (e.g., molds for mass-

producing bronze weapons and arrowheads in Bronze Age Europe and Metal Age Southeast Asia) or large-scale facilities for mass-production of weapons (e.g., the sometimes massive iron smelting areas in Iron Age European fortresses), and the adoption of horses for mounted cavalry (e.g., the horse-riding warrior images in Japanese Yayoi and Kofun tombs) (Barnes, 1999; Carman and Harding, 1999; Earle, 1997; Junker, 1999; Kolb and Dixon, 2002).

The Evolution of Chiefdoms

Many archaeologists view the transition to inherited, institutionalized power and status, associated with chiefdoms but further developed in states, as the key transformation in human societies. Since the organizational dynamics of chiefdoms are viewed as similar to that of states, theories of state development are commonly applied at the chiefdom level (Flannery, 1972; Wright, 1984). Early theories emphasized the managerial benefits of chieftainship as societies expanded in scale. Chiefs arose to administer irrigation works (Wittfogel, 1957), coordinate localized production and exchange (Service, 1962), mediate land conflicts and warfare (Carneiro, 1981), ameliorate economic risk through ritual intervention (Drennan, 1976), and to meet other organizational challenges.

More recently, archaeologists have criticized the functionalist nature of these theories and their simplistic emphasis on single factors, instead focusing on the varied acquisitive and power-seeking strategies of political actors as they vie for symbolic capital and control over the labor and resources of others (Earle, 1997). In this view, the origins of permanent forms of social inequality and political authority are to be found in the competitive and aggrandizing behavior of 'big men' in tribal societies. Archaeologists now examine this transformation at varying scales of analysis, ranging from the transfiguring actions and motivations of individual political actors (Dobres and Robb, 2000), to multiple polities developing in tandem through forms of peer polity interaction (Renfrew and Cherry, 1986), to change-stimulating contacts with more complex states and empires (world systems theory).

In addition, many archaeologists now recognize that not all chiefdom-level societies have unilinear evolutionary trajectories leading inevitably toward greater complexity. Many chiefdoms fail to develop state-level institutions over the long-term, instead perpetually cycling between complex and simple forms, 'devolving' into tribal societies (e.g., New Zealand's South Island Maori, possibly Amazonian societies), mysteriously collapsing (e.g., Mississippian chiefdoms, Easter Island), or otherwise maintaining a chiefdom-level organization until eventual absorption by colonizing states and empires (e.g., many chiefdoms of the Central American Isthmus, North America, the Caribbean, island Southeast Asia, Europe and Polynesia). As noted in the critiques of cultural evolutionary theory, other scholars reject the notion of 'chiefdoms' as a structural type in human societies altogether (as argued forcefully by Yoffee, 1993) and the explanation for their rise and fall is as varied as the 'chiefdoms' themselves.

See also: Archaeology, Theory In; Center–Periphery Relationships; Conflict and War, Archaeology of: Behavior and Social Organization; Conflict and War: Anthropological Aspects; Exchange in Anthropology; Intensification and Specialization, Archaeology of; Marxist Archaeology; North America, Archaeology of; Political Anthropology; Political Economy in Anthropology; Power in Society; Ritual and Symbolism, Archaeology of; States and Civilizations, Archaeology of; Trade and Exchange, Archaeology of.

Bibliography

Bard, K., 1994. The Egyptian predynastic: a review of the evidence. Journal of Field Archaeology 21, 265–288.
Barnes, G., 1999. The Rise of Civilization in East Asia: The Archaeology of China, Korea and Japan. Thames and Hudson, New York.
Beck, R., 2003. Consolidation and hierarchy: chiefdom variability in the Mississippian Southeast. American Antiquity 68, 641–661.
Blanton, R., Feinman, G., Kowalewski, S., Peregrine, P., 1996. A dual-processual theory for the evolution of Mesoamerican civilization. Current Anthropology 37, 1–14.
Brumfiel, E., Earle, T.K. (Eds.), 1987. Specialization, Exchange and Complex Societies. Cambridge University Press, Cambridge.
Carman, J., Harding, A. (Eds.), 1999. Ancient Warfare: Archaeological Perspectives. Sutton, Stroud.
Carneiro, R., 1981. The chiefdom as precursor of the state. In: Jones, G., Krautz, R. (Eds.), The Transition to Statehood in the New World. Cambridge University Press, Cambridge, pp. 37–79.
Cobb, C., 2003. Mississippian chiefdoms: how complex? Annual Review of Anthropology 32, 63–84.
Dietler, M., Hayden, B. (Eds.), 2001. Feasts: Archaeological and Ethnographic Perspectives on Food, Politics, and Power. University of Alabama Press, Tuscaloosa.
Dobres, M., Robb, J. (Eds.), 2000. Agency in Archaeology. Routledge, London.
Drennan, R., 1976. Religion and social evolution in Formative Mesoamerica. In: Flannery, K. (Ed.), The Early Mesoamerican Village. Academic Press, New York, pp. 245–268.
Drennan, R., Uribe, C. (Eds.), 1991. Chiefdoms in the Americas. University Press of America, Lanham.
Drennan, R., Peterson, C., 2006. Patterned variation in prehistoric chiefdoms. Proceedings of the National Academy of Sciences of the United States of America 103 (11), 3960–3967.
Earle, T. (Ed.), 1991. Chiefdoms: Power, Economy and Ideology. Cambridge University Press, Cambridge.
Earle, T., 1997. How Chiefs Come to Power: The Political Economy in Prehistory. Stanford University Press, Stanford.
Ehrenreich, R., Crumley, C., Levy, J. (Eds.), 1995. Heterarchy and the Analysis of Complex Societies. Archeological Papers No. 6. American Anthropological Association, Washington, DC.
Feinman, G., Lightfoot, K., Upham, S., 2000. Political hierarchies and organizational strategies in the Puebloan Southwest. American Antiquity 65, 449–470.
Feinman, G., Neitzel, J., 1984. Too many types: an overview of prestate societies in the Americas. In: Schiffer, M. (Ed.), Advances in Archaeological Method and Theory, vol. 7. Academic Press, New York, pp. 39–102.
Flannery, K., 1972. The cultural evolution of civilizations. Annual Review of Ecology and Systematics 3, 399–426.
Gilman, A., 1981. The development of social stratification in bronze age Europe. Current Anthropology 22, 1–23.
Helms, M., 1993. Craft and the Kingly Ideal. University of Texas Press, Austin.
Hodder, I. (Ed.), 2012. Archaeology Theory Today. Polity Press, Cambridge.
Johnson, A., Earle, T.K., 1987. The Evolution of Human Societies: From Foraging Group to Agrarian State. Stanford University Press, Stanford.
Junker, L., 1999. Raiding, Trading and Feasting: The Political Economy of Philippine Chiefdoms. University of Hawaii Press, Honolulu.
Kirch, P., 1984. The Evolution of Polynesian Chiefdoms. Cambridge University Press, Cambridge.
Kirch, P., Jones, S., 2003. New archaeological insights into food and status: a case study from pre-contact Hawaii. World Archaeology 34, 484–497.

Kolb, M., 1994. Monumentality and the rise of religious authority in precontact Hawai'i. Current Anthropology 35, 521–547.
Kolb, M., Dixon, B., 2002. Landscapes of war: rules and conventions of conflict in ancient Hawai'i (and elsewhere). American Antiquity 67, 514–534.
McIntosh, S. (Ed.), 1999. Beyond Chiefdoms. Pathways to Complexity in Africa. Cambridge University Press, Cambridge.
Oberg, K., 1953. Indian Tribes of Northern Mato Grosso, Brazil. Smithsonian Institution Press, Washington.
Pauketat, T., 2001. The Archaeology of Traditions: Agency and History before and after Columbus. University of Florida Press, Tallahassee.
Pauketat, T., 2007. Chiefdoms and Other Archaeological Delusions. Altamira Press, Lanham.
Peebles, C., Kus, S., 1977. Some archaeological correlates of ranked societies. American Antiquity 42, 421–448.
Peregrine, P., 1992. Mississippian Evolution: A World-system Perspective. Prehistory Press, Madison.
Redmond, E., 1998. Chiefdoms and Chieftaincy in the Americas. University of Florida Press, Gainesville.
Renfrew, C., 1973. Monuments, mobilization and social organization in neolithic wessex. In: Renfrew, C. (Ed.), The Explanation of Culture Change: Models in Prehistory. Duckworth Press, London, pp. 539–558.
Renfrew, C., Cherry, J. (Eds.), 1986. Peer Polity Interaction and Socio-political Evolution. Cambridge University Press, Cambridge.
Renfrew, C., Shennan, S. (Eds.), 1982. Ranking, Resource, and Exchange. Cambridge University Press, Cambridge.
Rothman, M., Stein, G. (Eds.), 1994. Chiefdoms and Early States in the Near East: The Organizational Dynamics of Complexity. Prehistory Press, Madison.
Sahlins, M., 1963. Poor man, rich man, big man, chief: political types in melanesia and polynesia. Comparative Studies in Society and History 5, 285–303.
Schortman, E., Urban, P., 2004. Modeling the roles of craft production in ancient political economies. Journal of Archaeological Research 12, 185–226.
Service, E., 1962. Primitive Social Organization: An Evolutionary Perspective. Random House, New York.
Wittfogel, K., 1957. Chinese society: an historical survey. The Journal of Asian Studies 16, 343–364.
Wright, H., 1984. Prestate political formations. In: Earle, T.K. (Ed.), On the Evolution of Complex Societies. Undena Publications, Malibu, pp. 41–77.
Yoffee, N., 1993. Too many chiefs? In: Yoffee, N., Sherratt, A. (Eds.), Archaeological Theory: Who Sets the Agenda? Cambridge University Press, Cambridge, pp. 60–78.

Child and Adolescent Psychiatry, Principles of

Nisha Dogra, Greenwood Institute of Child Health, University of Leicester, Leicester, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.H. Schmidt, A. Maras, volume 3, pp. 1676–1682, © 2001, Elsevier Ltd.

Abstract

This article begins with a brief review of child and adolescent psychiatry (CAP) as a distinct medical discipline and also the training required to practice it. We also briefly consider the options when there is a shortage of or no child psychiatrists available to deliver care. There is a short discussion on the features that distinguish CAP from other specialties. The influence of culture, development, family, and school are discussed. This is followed by outlining the principles of an effective approach to delivering high-quality care for children and young people. This article will not consider the presentation or treatment of specific disorders.

Introduction

This article begins with a brief review of child and adolescent psychiatry (CAP) as a distinct medical discipline and also the training required to practice it. There is a short discussion on the features that distinguish CAP from other specialties. The influence of culture, development, family, and school are discussed. This is followed by outlining the principles of an effective approach to delivering high-quality care for children and young people. This article will not consider the presentation or treatment of specific disorders.

Terminology

For the ease of the reader, the word child is used to denote all those under the age of 18 years. The word adolescent is used specifically to refer to those aged 12 and above.

CAP as a Distinct Medical Specialty

CAP is a relatively young specialty and in many ways continues to be a developing one especially in middle- and low-income countries. In some countries, CAP has well-established roots as a separate specialty, whereas in others, the work is carried out by general psychiatrists or pediatricians. Training in CAP varies widely across international contexts and there is a large disparity of standards and practice in CAP including European countries where it has been formally compared (Simmons et al., 2012). This may be due to the wide variation in historical origins of the specialty. In the United States, CAP has its origins in the post second war guidance clinics. In the United Kingdom, it was promoted through some of the work undertaken by psychotherapists and further cemented by epidemiological studies indicating the level of unmet need. The changing views of children and the notion that children had specific rights may also have played a part (O'Reilly et al., 2013).

Of the 34 European countries for whom data was collected about child psychiatry (Simmons et al., 2012), six (Belarus, Bosnia, Ukraine, Malta, Russia, and Spain) reported that there is no official CAP training program. The remaining 28 participants, who were Austria, Belgium, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Israel, Italy, Latvia, Lithuania, The Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, Sweden, Switzerland, Turkey, and the United Kingdom, reported official CAP training programs. In 7/28 countries (25%), there is a core common trunk in general psychiatry before trainees specialize in CAP; 19 countries (68%) do CAP from the start of psychiatry training, and 2 countries (7%) start CAP training straight from medical school. All the respondent countries, except Greece, have national training standards for CAP, although only 11 reported their rigorous implementation. Countries use a number of assessment strategies ranging from oral and written exams to submission of a curriculum vitae, research thesis, or analysis of video recordings. Canada, Australia, New Zealand, and the United States all have recognized pathways for specializing in psychiatry. In common with some European countries including the United Kingdom, these countries all require some general medical training after qualification from medical school and some training in general psychiatry. Training varied in terms of time from very little time up to 5 years and more. Other countries that also have formal CAP programs include Morocco (since 2007) and Paraguay has started such training in 2013. Recruitment into CAP is a problem in a significant number of countries, and appropriate training opportunities are issues for developing countries (e.g., Pakistan; personal communication).

There are difficulties in ensuring that there are enough appropriately trained clinical staff to meet the mental health needs of children given the increasing prevalence rates.

In situations where there is no direct access to a child psychiatrist or a severe shortage, provision of services requires other staff such as general practitioners, pediatricians, or adult psychiatrists to develop special interests to ensure the mental health needs of children are met. Another option may be to train up nursing staff or other staff (for example, teachers) who have access to families for other reasons such as delivering maternity or educational services, respectively. Training should allow staff to deliver care for the problems most encountered in their context. This requires applying the principles described

below. Clearly, it is important for staff to be supported and this may be done through links with centers that have child psychiatrists who are able and willing to provide consultation and support. There are examples of modern technology being used to provide support to those delivering frontline services and the potential of this in child mental health has yet to be fully realized. There are also a number of training opportunities specifically aimed at non-mental health specialists to develop their understanding of child mental health. Eapen et al. (2012) wrote to specifically address this issue by suggesting that staff working in primary contexts such as primary health care or education could make some basic provisions. It is probably best to be open with families about the potential limitations so that they are fully aware of what may or may not be available. In these situations CAP may have a greater role in consultation and supporting staff rather than providing direct care.

The models to deliver services will need to take local contexts into account. The World Health Organisation states that "Mental health promotion requires multi-sectoral action, involving a number of government sectors and non-governmental or community-based organizations. The focus should be on promoting mental health throughout the lifespan to ensure a healthy start in life for children and to prevent mental disorders in adulthood and old age."

The challenge may be in thinking about how we can promote children's mental health with whatever limited resources are available. The reality is that if we continue to wait for the perfect time and adequate resources, nothing may materialize.

What Distinguishes Child Psychiatry?

The key features that distinguish the practice of CAP from general adult psychiatry are the following:

- Nowhere is development more marked than in childhood and adolescence.
- Children have to be viewed in the context of their families on whom they are generally dependent and definitions and compositions of the families are changing all the time.
- The cultural context of the child is even more significant as both the first two features are related to this aspect.
- Most often, nominated patients do not present themselves but are presented.
- Cultural mismatches between young people and their carers that may affect presentations to mental health services and how they are managed.

Social Contexts and Experiences

Social factors are, to varying degrees, of undoubted importance in the prevention, development, and maintenance of mental health problems including those of childhood. As Jenkins (2008) highlights, disentangling genetic from environmental influences is not easy.

Bronfenbrenner's (1979) model of layered contexts is a fairly widely used model to try and understand the various factors that need to be considered when thinking about CAP. The layers are the microsystem, mesosystem, exosystem, and macrosystem. The first layer (microsystem) is at the level of the relationships in which the child is actively involved, for example, with parents, siblings, and peers. These interactions inevitably depend on context and the child's characteristics (or personality). The mesosystem describes how the different components of the microsystem come together. The exosystem is the wider local community such as the neighborhood. The macrosystem is usually more remote, but may still be a major social influence, for example, through socioeconomic policy on child rearing, education and health policies, and wider social and cultural contexts. It is worth noting that different factors may be important at various stages of the child's life. Some adverse family factors may be mitigated by external factors that come into play as children get older (e.g., peer support). Those working in CAP need to be aware of the dynamics and the interplay between the different layers.

Cultural Considerations

Culture itself influences the way that children and their perspectives are viewed (Dogra, 2010; Dogra et al., 2007). Although most countries have signed up to international agreements, there is still huge variation in the value that children are accorded (O'Reilly et al., 2013). Among many issues, culture in its broadest sense (see Dogra, 2010) influences

- the meaning of childhood and expectations of children
- the way developmental milestones and family development are viewed
- parenting styles
- the interpretation and meanings of the symptoms that may be present
- religion
- views of gender.

An understanding of the cultural context may help clinicians make sense of the presentations and formulate the issues in ways that are acceptable to the family and thereby amenable to intervention. There are very few culturally specific problems. However, some disorders are more likely to occur in some contexts than others (Dogra et al., 2007). For example, the eating disorder anorexia nervosa is more common in Western contexts than others. Rousseau et al. (2009) surveyed the literature on culture and Diagnostic and Statistical Manual of Mental Disorders (DSM)-IV in child psychiatry. They concluded that, although the DSM-IV diagnostic categories may be found cross culturally, clinicians need to be aware of how culture may influence the diagnostic process.

Family Environments

One cannot think about child psychiatry without considering the family and the part it plays in the child's life. While the precise makeup of families may be culturally variable, the concept of the family is universal.

Parenting

The role of parenting is to adequately nurture the child and prepare them for adulthood (whether that is independent living or living in an adult role within an extended family). Parenting styles are strongly influenced by broader cultural and political factors. There are widely accepted to be four different types of parenting styles that bring with them different advantages and disadvantages. Baumrind (1966) identified authoritative, authoritarian, and permissive parenting styles. Macoby and Martin (1983) divided permissive parenting into two subtypes – neglectful and indulgent. Awareness of parenting styles and how these impact on the child are of relevance in child psychiatry.

Authoritative Parenting

Authoritative parents make demands that fit with their children's ability to take responsibility for their own behavior. Children, it is suggested, subsequently learn that they are competent individuals who can do things successfully for themselves, which then fosters high self-esteem, cognitive development, and emotional maturity. Characteristics of authoritative parenting include emotional support, clear bi-directional communication, firm limit setting, reasoning, and responsiveness (Baumrind, 1966). Warm supportive parenting aids child development and helps children to regulate their own emotions and behaviors. It may also be protective in adverse circumstances. Children are more likely, according to this framework, to learn to express their perspective in a constructive socially accepted way. This is because there is appropriate modeling and the parents are responsive to their child's needs. This type of parenting requires a secure attachment with the primary caregiver, a positive relationship between the parent and child, and age-appropriate and consistent disciplinary practices, e.g., reasoning with a younger child may be inappropriate in the context of the child's development.

Authoritarian Parenting

In contrast, authoritarian parenting is described as punitive and directive. Again stereotypically, parents who use this style are unlikely to discuss rules and will emphasize discipline and obedience (Baumrind, 1966). The views of young people are deemed to be of little importance irrespective of age. It is implied that this style is not responsive to young people's emotional needs; however, in practice, this may not be the case.

Permissive Parenting

Both permissive styles of parenting – neglectful and indulgent – make few demands on children, with limited boundaries. In the case of neglectful parenting, there is reduced care or supervision (Macoby and Martin, 1983) and, in extreme form, there is a lack of responsiveness to the needs of children, emotional or otherwise. Indulgent styles may also be problematic as again there are limited boundaries – the child is just given everything (Baumrind, 1966). There is evidence that 'overindulged' children show poor emotion regulation (underregulated) and rebel when they cannot have their own way. 'Overindulgent' parenting may be used to avoid having to deal with the conflicts that may arise or to manage parental separation and hostility.

Development

There is no doubt that culture influences child rearing and the way children are encouraged to grow up. Culture influences the nature of their relationships with peers and siblings and the roles and responsibilities that accompany their development. The key areas of development are physical, emotional, social, cognitive, and moral development. Culture can influence all aspects of development but it can be difficult to identify exactly how it influences any particular child as there will be other complex dynamic features at work. The influences on social and emotional development are most evident.

As already discussed above, parenting styles may be influenced by cultures. However, families are in many ways unique entities of their own. It is often difficult to define what a family is because the 'nuclear' family that so many of our ideas are based on is increasingly changing in shape and content. CAP is frequently influenced by the notion of the family life cycle outlined by Carter and McGoldrick (1989). Despite its limitations, the Carter–McGoldrick model highlights that life and development inherently bring with them changes in social context and practice. The way that families as a whole or individuals within that unit manage that change has implications for all members of the family. While the stages of the family life cycle may be variable between and within cultures, the concept of different stages of development is likely to hold true. Practitioners can compare their own experiences and those of the families they have worked with using the model as a benchmark. It becomes unnecessary to debate whether the model is right or wrong but it offers families and staff stages that can be considered whatever the context.

School Environments

After home, school is where children spend most of their time, so it is unsurprising that school environments and experiences are an important factor when considering the mental health of children. Schools can have a very positive impact on children and mitigate some of the negative impacts of other social factors, including aversive home environments (Mortimore, 1995). However, for some young people they can be a considerable source of stress, worry, and unhappiness, particularly if bullying is experienced. Elias et al. (1997) argued that schools are most successful in their educational mission when they promote academic, social, and emotional learning. This approach recognizes that emotions can facilitate or hamper young people's learning, which in turn influences their mental health.

There are many factors such as poverty, war, and trauma that are beyond the scope of interventions that can be applied by CAP on a case by case basis. However, the

influences of these on the child's mental state need to be borne in mind especially by those who work with children with mental health problems but are not specifically trained in the area.

Whose Perspective Is Most Valued?

In child psychiatry, there is the additional complication that the clinician has to deal with different viewpoints within the family, which will impact on the assessment and treatment that can be agreed. Cultural and family factors clearly influence whose perspective in the family is afforded the greatest value but the clinician working within CAP has to balance all perspectives. Garland et al. (2004) interviewed 170 adolescents, their caregivers, and their therapists to identify three desired outcomes for each stakeholder. Almost two-thirds of the triads did not agree on even one of the desired outcomes for the adolescent's treatment when desired outcomes for individual cases were compared. Essentially, there was a lack of agreement of the desired outcome among key stakeholders or desired outcome priorities for adolescent services. Williams et al. (2011) found low parent–adolescent concordance on perceived need for help, with a greater proportion of adolescents reporting a need compared to their parents. Concordant dyads kept a higher number of appointments than discordant dyads. A study by Roberts et al. (2005) comparing concordance between parent and child in different ethnic groups found that ethnicity is a significant factor in parental labeling and awareness of adolescent mental health problems. This lack of consensus may limit engagement in treatment and the effectiveness of care and this highlights a fundamental issue for child and adolescent mental health service (CAMHS) in identifying who the client is and whose needs are prioritized. Parents may attend and/or support their child's attendance depending more on how they feel than whether the child requires help or not.

It is clear that these factors strongly influence emotional and social development and a sense of self. How the child's sense of self, parental, and wider expectations interplay in the cultural context with each other are of great relevance in the practice of child psychiatry.

Principles of an Effective Approach To Service Delivery

Context of Child Psychiatry Services

Most child psychiatry care is delivered in the outpatient community setting, with some being delivered in day centers and some in inpatient units. Generally, inpatient care is reserved for those at high risk or those who are so seriously ill that they need constant nursing and/or medical attention. Outpatient teams are dependent on local resources but multidisciplinary teams are generally viewed as an effective way to deliver services, although the evidence for any approach is relatively limited.

There is also the issue of whether child psychiatry services should be part of mental health provision or pediatric/child services. This is best answered by considering the conceptual, theoretical, and practice issues in turn. While child psychiatrists and paediatricians at first sight may appear to have much in common, there may be little commonality in their approaches. The etiological and treatments of child and adult psychiatry share more in common than pediatrics and child psychiatry. In developing countries, the pressures to deal with physical health problems may mean that the emotional health of children is always relegated to second place if it is even on the agenda. However, there is evidence that children experience mental illness whatever their background and sadly we also know that the stigma surrounding children with mental illness is a universal phenomenon (Leighton and Dogra, 2009).

High-Quality Services

There is increasing debate about what constitutes high-quality child psychiatry care. The Institute of Medicine (2001) quality framework identified that quality care is timely, efficient, effective, patient centered, equitable, and safe.

Assessment in CAP

Most Western contexts would argue for services to be child centered but the meaning of child centered will depend on local contexts. Whatever the approach, good practice in CAP does need the psychiatrist to keep the child's interests as paramount and pay attention to them and their needs whatever other pressures there may be. This section will outline the structure of a CAP assessment before dealing with interviewing children as summarized in **Box 1**.

The structure of the assessment is fairly standard, although surprisingly there is little evidence for why we use this format. Renou et al. (2004) identify that the structured interviews used for CAP research are not practical for clinical practice, although work to improve the reliability of the assessment is ongoing. Hirsch (2013), in describing the practice at a New York facility, states that they send out a packet of questionnaires and rating scales ahead of the first interview. This is probably not appropriate or possible except for some research facilities. Questionnaires cannot replace good clinical practice and interviews remain the best way of undertaking a CAP assessment. There is a need to be flexible about the order and the way in which information is collected. It can be helpful to start by talking about areas

Box 1 Key components of a comprehensive CAMHS assessment

- Presenting concerns and history of these
- Family history
- Social history
- Developmental history
- Risk assessment
- Mental state examination of child
- Observation of young people with their carers
- Formulation and care plan

such as school, family, or hobbies before raising the 'problem' to help the child become comfortable. Good practice means that often children do not even realize they are participating in an assessment as the style is conversational and not interrogative.

Setting the Context

It is important to clarify what the child and/or their carers are expecting and what they know of the service. It is also important to establish their expectations and hopes for the meeting. Consent and confidentiality issues are best addressed early on and it can be helpful for the child to know that there are clear boundaries and that their perspective is as important as those of any adult.

Presenting Concerns

It is important to establish what the child and their family think are the main worries. These may be different from what the referrer has told you, and the relative importance given to different things may vary between different members of the family too. An accurate history of how long the symptoms or worries have been present may also give a clue to precipitating factors and the need to examine how these have affected people differently.

Family History

This involves identifying who is in the family and what constitutes the family from the child's perspective. A genogram (as shown in Figure 1) can be a useful way to demonstrate who is in the family and family relationships. Parental experiences of parenting may be relevant and dependent on the history; this area will be touched upon or explored in more detail.

Social History

This usually means asking the young people about their peer relationships and their interests. As children get older, they begin to have more contacts outside of the family. This is a good opportunity to find out what is of interest to the child and can be a good place to begin interviews as it helps engage with the child. As children become older, asking them about peer activities becomes more important. Young people may be reluctant to share all aspects of this part of the history with their carers. Drug and alcohol history is an important part of the social history, though again sensitivity may be required in getting honest answers from young people about their use of substances and alcohol.

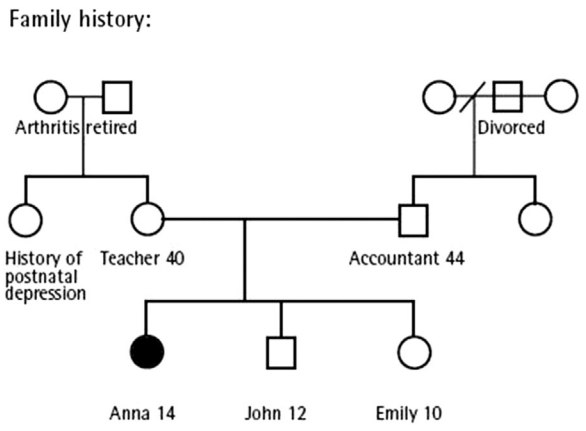

Figure 1 An example of a genogram.

Developmental Assessment

Given that children and young people are developing toward maturation for adulthood, a developmental assessment is a key component of a comprehensive child and adolescent mental health assessment. A developmental assessment is a structured evaluation of the child's development and a comprehensive assessment needs to cover the key domains – physical, social, emotional, moral, and intellectual. Such an assessment is likely to involve several members of the multidisciplinary team (e.g., formal intellectual assessment is best undertaken by a psychologist who has specific training in this; an assessment of motor development, a component of physical development may be best undertaken by an occupational therapist). It will also usually involve seeing the child in different contexts and environments. While this may not be within the exact remit of CAP, the reality is that many referrals to CAMHS have developmental components as potential etiological factors and even if a developmental assessment has been undertaken it may be insufficient.

Why Is Development Important To Consider?

Consider the symptom of 'inattention and poor concentration.' In a 3 year old, this is not unusual, so a diagnosis of attention deficit disorder is unlikely to be appropriate. However, a child who is 7 years of age but with global developmental delay may still have features of inattention and poor concentration. Only a detailed assessment could clarify the situation and an initial assessment establishes whether the child's development is appropriate or not. Unless there is a clear understanding of a child's level of functioning and understanding, it is difficult not only to make an assessment of their presentation but also to identify the appropriate intervention.

Risk Assessment

Risk management is an integral component of a CAP assessment and is especially important with an increasingly litigious society. The nature, source, and assessment of the risks assessed should be clearly recorded. There is the risk children may present to themselves, the risk to others, and also the risk they may be at from others. In CAP, almost all of the children seen in specialist services will be seen with an adult who has parental responsibility for the child and therefore effectively carry some of the responsibility for managing the risk associated with that child. Any risk management plan associated with preventing further occurrences of self-harm, for example, is likely to require their input. There may well be a number of complicating factors within the relationship to take into account, but

the adult carer also has some responsibility to protect and safeguard the child.

Mental State Assessment

In children, this will usually include the form of an observation of behavior, particularly in relation to how they interact with their parents (or carers), siblings (if present), and the clinician. This may take the form of looking for normal developmental markers as well as social skills, appropriate eye contact, and looking for unusual or odd behavior. In older children and young people, this may be a much more formal process, looking at issues more familiar to a mental state examination as in adults. In some older children seen in contexts such as specific 'drop-in' services or those that do not wish for their parents or carers involvement, the interaction between family members may not be possible to observe. Observations need to be done in context of what is appropriate for the child or young person's developmental stage, as outlined above. What is appropriate at a very young age may be interpreted as 'clinginess' for a slightly older child. Attachments to family and carers become looser in adolescents where part of the task is to develop appropriate levels of independence.

The Assessment Process

Assessment of most children will involve spending some time with the child, some time with the carers, and some time with them together and perhaps even some time with the wider family. All children should be afforded some time alone and the extent of this will be influenced by their age, level of cognitive functioning, and the nature of their presentation. Some clinicians recommend starting with the family together and then giving each party some time alone. My own experience has led to a different approach. With older adolescents (age 14 years upward), I usually offer the young person the opportunity to start with me alone to emphasize that I value their perspective and second I have found that parental presence can limit the child's input to the assessment process.

It is likely that most children and families will have limited information about what child psychiatry services do; so it is probably helpful to provide an information leaflet such as the one shown in Table 1. This has clearly been written from a 'Western UK perspective' and the information may not be the same across different contexts but the need to be clear about what the service does and how it functions is relevant whatever the context.

As in other services, professionals need to be aware that the experiences children have had may make them suspicious of mental health services. Children may also have been given misinformation about who they are going to see by parents who do not want to upset the child. It is widely acknowledged that nonspecialist professionals often have a poor understanding of mental health and may not have discussed the referral in any degree of detail with the family. After introductions have been made, a useful opening question is to ask the young person (and anyone else present) what they anticipate will happen. It is important to explore where they have obtained these ideas from, as their sources may be unreliable. Young people need to know that they are not being negatively judged and that they are valued.

The principles of engagement, respect, and such are as important and relevant in child psychiatry as in all other areas where human communication is a key part of the process. There is, however, an issue that in different contexts respect means different things to different people. It is best not to rely on stereotypes to consider what might be important to any particular child or family, but to discuss this at an individual level. Children who come from families where parents are authoritarian (and this can be from any cultural background) or backgrounds in which children are seen but not heard may be less forthcoming. Clinicians need to be open to the idea that children may need different levels of coaxing. Just because a child does not say much, does not mean he or she does not have anything to say. It is worth paying attention to the contexts. Exploring the meanings of mental health for the child and the family is important, as this will reveal so much about their understanding and the work that lies ahead for everyone. At the same time, it is important to avoid making assumptions or trying to explain things in a way that supports one's own world view.

Diagnosis and Management

Diagnosis should use either the International Classification of Diseases or the DSM. The treatment approach is guided by the assessment and should always be discussed with the child and family, as their knowledge, consent, and motivation are essential if the treatment is to be successful. In planning treatment, the setting, the content, and the method of the treatment should again be considered. Additionally, the mode of treatment should be considered. The setting may again be in the family home, general practitioner surgery, local child psychiatry clinic, or pediatric liaison clinic.

The content of treatment can again be divided into psychological, environmental, biological, or developmental. Each of these should be taken into account when deciding on the method of treatment, which may be individual therapy, work with parents/carers, group therapy, family therapy, drug treatment, or a combination of these. The mode of treatment in child psychiatry is commonly behavioral, psychodynamic, cognitive, educational, supportive, or a combination of these. Each mode of treatment can be applied in at least one of the methods of treatment. However, it is usually best to consider each treatment method in turn: individual therapy, parent work, group therapy, family therapy, and drug treatment. Most therapies in psychiatry can be used in children and adolescents but will require adaptation for use with particular groups or may not be suitable for them (e.g., very young children and individuals with severe learning disabilities). Basic parenting work can usually be delivered by other agencies whilst family therapy and family work is more likely to take place within CAP although not exclusively limited to CAP services.

Summary

In summary, although a relatively young specialty, CAP has made significant advances and there are clear principles of good practice. While there have been advances in understanding

Table 1 Child and Adolescent Mental Health Service information leaflet for first appointment (to be sent with appointment letter)

What is a Child and Adolescent Mental Health Service?

A Child and Adolescent Mental Health Service (CAMHS) is a specialist service that offers assessment and treatment for child mental health problems. We generally see children and young people up to the age of 18 years.

What is a mental health problem?

Children and young people can experience a range of emotional, psychological, behavioral, and psychiatric problems for which specialist help is required. A mental health problem can be seen as a problem that affects the child or young person or their family and prevents them from achieving their potential. Not all child mental health problems need specialist services but those that are complex, severe, or persistent benefit from such involvement.

Examples you may have heard of include depression, anxiety, attention deficit hyperactivity disorder, autism, or family problems.

Child mental health problems are rarely caused by a single factor and it is usually a combination of individual, family, social, and environmental factors that lead to problems.

Who will you see?

There are many different types of professionals who work in CAMHS. They include

Doctors (specialized in child psychiatry)

Nurses

Psychologists

Therapists trained in a variety of treatment strategies such as psychotherapy, art, cognitive behavior, and family

Our staff are trained in working with children, young people, and their families to ensure that the service we provide is tailored to your needs using the best evidence there is. We are committed to working together with you so that any interventions offered are appropriate and acceptable to you and your family.

Who needs to come?

For the first appointment, we suggest the carers come along with the young person who has been referred. You are also welcome to bring other family members, a friend or an advocate if you feel that would be helpful to you.

If we need to make any special arrangements such as access, please let us know before coming as we want to make your visit as comfortable as we can.

What will happen when you come to see us?

When you attend your first appointment, we will ask you to complete some forms. It seems like a lot of work but it is really important to us that we understand your point of view. You will then be seen by one or two clinicians and the appointment lasts between one and one and a half hours.

Depending on the young person's age, some of the assessment may be done with the young person and the carers together and part of it with them separately. This is so we can understand everyone's point of view to better work with you in planning any further care. The appointment is quite long because we need to collect quite a lot of information to be able to make a thorough assessment.

After the first meeting, the clinicians will feedback their assessment to you and there may be one of several outcomes:

- No further assessment or treatment required and the person is discharged from the service.
- Further assessment is required; so more appointments are made.
- A diagnosis is made and suggestions for treatment options are discussed. Treatment may be offered immediately or the young person placed on a treatment waiting list dependent on the issues.

Whatever the outcome, the clinicians will discuss this with you to ensure that your views have been taken into account. The clinician will write to the person who asked you to come and see us and you will get a copy of the assessment letter.

You should know that we are happy to be asked about what we do as it is important for us that you are as fully informed as is right for your development about your own care.

Confidentiality

We offer a confidential service but there are times we need to share information with partner agencies to ensure we deliver the best care that we can. If we plan to share information about you we will ask for your or your carers' permission unless there are concerns for your safety in which case we are legally obliged to share information to ensure your protection. For young people, it is important to know that they can share information with clinical staff that is just between them. Again, information will be shared with parents only in cases where there are concerns for your safety.

How to find us?

We are based at Westcotes House on Westcotes Drive, Leicester.

the etiology and treatments available, there is still much to be done in identifying the relationship between treatment and outcomes.

See also: Academic Performance, Effects of Socio-Economic Status on; Behavioral Problems, Effects of Socio-Economic Status on; Bereavement and Complicated Grief across the Lifespan; Counseling and Psychotherapy: Ethnic and Cultural Differences; Cultural Views of Life Phases; Defiant Behavior During Adolescence across Cultures; Defiant Behavior During Infancy and Early Childhood; Diagnostic Processes in Clinical Psychology and Psychotherapy; Early Emotional Development and Cultural Variability; Emotional Development, Effects of Parenting and Family Structure on; Evidence-Based Psychotherapy; Intellectual Functioning, Assessment of; Life Management, Developmental Psychology of; Mass Trauma: Psychopathological Effects across the Life Span; Motivational Development, Systems Theory of; Narrative Exposure Therapy; Neonatal Temperament: From Phenomenology to Mechanism; Obsessive-Compulsive Disorder across the Life Span; Panic Disorder and Agoraphobia Across the Lifespan; Parenting Attitudes and Beliefs across Cultures; Personality Changes During Adolescence Across Cultures; Phobias Across the Lifespan; Psychodynamic Therapy: Short-Term; Psychological Treatment, Effectiveness of; Psychological Treatments: Randomized Controlled Trials; Self and Identity Development During Adolescence across Cultures; Self in Culture: Early Development; Self-Development in Childhood and Adolescence; Self-Regulation During Adolescence: Variations

Associated with Individual–Context Relations; Sleep Disorders across the Life Span; Social Phobia across the Lifespan; Social and Emotional Development in the Context of the Family; Socioeconomic Status and Social-Emotional Development of Children from Birth to 36 Months of Age: A Systematic Narrative Review; Stress in Adolescence: Effects on Development; Stress in Childhood: Effects on Development; Stress in Infancy and Early Childhood: Effects on Development; Stress: Effects on Neonatal and Child Development; Temperament Development, Theories of; Temperament and Human Development; Therapist–Patient Relationship.

Bibliography

Baumrind, D., 1966. Effects of authoritative parental control on child behavior. Child Development 37, 887–907.

Bronfenbrenner, U., 1979. Contexts of child rearing: problems and prospects. American Psychologist 34, 644–850.

Carter, E., McGoldrick, M., 1989. The Changing Family Life Cycle: A Framework for Family Therapy, second ed. Gardner, New York.

Dogra, N., 2010. Culture and child psychiatry. In: Bhattacharya, R., Cross, S., Bhugra, D. (Eds.), Clinical Topics in Cultural Psychiatry. Royal College of Psychiatrists Press, London, pp. 209–221.

Dogra, N., Vostanis, P., Karnik, N., 2007. Child and adolescent psychiatric disorders. In: Bhugra, D., Bhui, K. (Eds.), Textbook of Cultural Psychiatry. Cambridge University Press, Cambridge, pp. 301–313.

Elias, M.J., Zins, J.E., Weissberg, R.P., Frey, K.S., Greenberg, M.T., Haynes, N.M., Kessler, R., Schwab-Stone, M.E., Shriver, T.P., 1997. Promoting Social and Emotional Learning: Guidelines for Educator. Association for Supervision and Curriculum Development, Alexandria, VA.

Garland, A.F., Lewczyk-Boxmeyer, C., Gabayan, E.N., Hawley, K.M., 2004. Multiple stakeholder agreement on desired outcomes for adolescents' mental health services. Psychiatric Services 55, 671–676.

Hirsch, G.M., 2013. Psychiatric Evaluation and Diagnosis: What Is It and What to Expect. http://www.aboutourkids.org/articles/psychiatric_evaluation_diagnosis_what_it_what_expect (accessed 09.08.13.).

Institute of Medicine, 2001. Crossing the Quality Chasm: A New Health System for the 21st Century. National Academy Press, Washington, DC.

Jenkins, J., 2008. Psychosocial adversity and resilience. In: Rutter, M., Bishop, D., Pine, D., Scott, S., Stevenson, J., Taylor, E., Thapar, A. (Eds.), Rutter's Child and Adolescent Psychiatry, fifth ed. Blackwell Publishing, Oxford, pp. 377–391.

Karnik, N., Dogra, N., Vostanis, P., 2007. The management of child and adolescent psychiatric disorders. In: Bhugra, D., Bhui, K. (Eds.), Textbook of Cultural Psychiatry. Cambridge University Press, Cambridge, pp. 471–483.

Leighton, S., Dogra, N., 2009. Defining mental health and mental illness. In: Dogra, N., Leighton, S. (Eds.), Nursing in Child and Adolescent Mental Health. McGraw-Hill, Maidenhead.

Macoby, E.E., Martin, J.A., 1983. Socialisation in the context of the family: parent-child interaction. In: Mussen, P.H., Hetherington, E.M. (Eds.), Handbook of Child Psychology, Socialisation, Personality and Social Development, fourth ed., vol. 4. Wiley, New York.

Mortimore, P., 1995. The positive effects of schooling. In: Rutter, M. (Ed.), Psychosocial Disturbances in Young People: Challenges for Prevention. Cambridge University Press, Cambridge, pp. 333–365.

O'Reilly, M.J., Ronzoni, P.D., Dogra, N., 2013. Research with Children. Sage, London, pp. 2–8.

Personal communication of paper with Rahman, R. Challenges to Psychiatry in Pakistan & Possibilities for Regional Cooperation Delivered at the Asia Pacific Mental Health Forum on Tuesday, 28th April, 2013, during Royal Australia & New Zealand College of Psychiatrists Annual Congress.

Renou, S., Hergueta, T., Flament, M., Mouren-Simeoni, M.C., Lecrubier, Y., 2004. Diagnostic structured interviews in child and adolescent's psychiatry. Encephale 30, 122–134.

Roberts, R.E., Alegria, M., Roberts, C.R., Chen, I.G., 2005. Concordance of reports of mental health functioning by adolescents and their caregivers: a comparison of European, African and Latino Americans. Journal of Nervous and Mental Disease 193 (8), 528–534.

Rousseau, C., Measham, T., Bathide-Suidan, M., 2009. DSM-IV, culture and child psychiatry. Journal Canadian Academy of Child Adolescent Psychiatry 17, 69–75.

Simmons, M., Barrett, E., Wilkinson, P., Pacherova, L., 2012. Trainee experiences of child and adolescent psychiatry (CAP) training in Europe: 2010–2011 survey of the European Federation of psychiatric trainees (EFPT) CAP working group. European Child Adolescent Psychiatry 21, 433–442.

Williams, C.D., Lindsey, M., Joe, S., 2011. Parent-adolescent concordance on perceived need for mental health services and its impact on service use. Children and Youth Services Review 33, 2253–2260.

World Health Orgnaisation, 2014. What Is Metnal Health? http://www.who.int/features/qa/62/en/index.html (accessed 16.06.14.).

Child Care and Child Development

AC Huston, University of Texas, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Most children in modern industrialized societies will spend a significant portion of their early lives in child care, and many will require such care during the early school years as well. Initial fears that child care per se would be harmful have not been supported, but there is some evidence that extensive care, beginning in early infancy, can make it more difficult for mothers to relate sensitively to their infants. Whether child care has positive or negative effects on children's development depends primarily on its quality. Quality is defined by sensitive, responsive, and stimulating caregivers and age-appropriate activities and curriculum. High-quality care promotes cognitive, academic, and social development, and its effects last beyond the preschool years. Children attending child care centers as opposed to home-based child care perform better on language and academic skills and have skills with peers, but are also more aggressive and noncompliant.

Most children in modern industrialized societies will spend a significant portion of their early lives in child care, and many will require such care during the early school years as well. Initial fears that child care per se would be harmful have not been supported, but there is some evidence that extensive care, beginning in early infancy, can make it more difficult for mothers to relate sensitively to their infants. Whether child care has positive or negative effects on children's development depends primarily on its quality. Quality is defined by sensitive, responsive, and stimulating caregivers and age-appropriate activities and curriculum. High-quality care promotes cognitive, academic, and social development, and its effects last beyond the preschool years. Children attending child care centers as opposed to home-based child care perform better on language and academic skills and have skills with peers, but are also more aggressive and noncompliant.

Child care is regular care provided by someone other than a child's parents. Throughout human history, grandparents, siblings, and relatives have cared for young children, but since about the 1950s, nonparental child care has become an increasingly visible and prevalent part of the lives of young children in most industrialized societies, largely because of increasing rates of maternal employment and of single-mother families. In the United States, by 1997, the majority of infants and preschool children spent some time in child care; many of them were in full-time care (35-plus hours per week) (Capizzano and Adams, 2000).

This rapid and dramatic change in the ecology of children's experience has raised a host of questions about the benefits and risks associated with nonparental child care. Strong value commitments and assumptions have often influenced the questions asked. In the first wave of research in the 1970s, for example, researchers asked whether 'day care' was harmful to children; they rarely considered potential benefits. At the same time, others were investigating the benefits of 'early childhood intervention' without considering possible harmful effects. Both groups were examining systems of nonparental care for young children, but the ways they framed questions and the labels they used led to different conclusions.

Is Extensive Child Care Harmful?

Why might we expect nonparental child care to be harmful for children? According to the 'maternal deprivation' view, having one primary caregiver (usually the mother) with whom to develop an early attachment relationship is critical to the socioemotional development of young children. This view implies that even very high-quality nonmaternal care could impair the development of a secure attachment to the mother, if it seriously reduced the amount of time that a child spends with its mother. The most commonly used measure of attachment is the Strange Situation, a laboratory procedure in which children are observed during and after brief separations from their mothers. Secure attachment is indicated when the child goes to the mother and is comforted by her. Insecure attachment may be manifested by avoiding or ignoring the mother or by clinging to her without being comforted.

A large number of early studies concluded that child care in the preschool years did not have harmful effects on attachment security or later socioemotional development, but the evidence is more mixed with respect to extensive nonmaternal care in the first year of life. Although the majority of children who receive full-time child care in infancy have secure attachments to their mothers, some studies find elevated rates of insecure attachment when comparing these children to those in exclusive maternal care or part-time child care (Bystrova et al., 2009). A major study of 1153 infants across the United States indicated that children who spent extensive time in child care during the first 3 years of life had mothers who were slightly less sensitive in interactions with them, and that extensive care was associated with elevated rates of insecure attachment only for children whose mothers were insensitive (NICHD Early Child Care Research Network, 1997, 1999a,b). As one reviewer concluded, "Adverse effects on infant-mother attachment appear to occur only when infant day-care co-occurs with other risky conditions, …" (Lamb, 1997; Lamb and Ahnert, 2006).

It does not appear that time in child care, in and of itself, has long-term effects on most other aspects of socioemotional development, but children with extensive care from infancy on tend to be more aggressive and assertive and to be less compliant to adults in some settings than are those with less

child care experience. Many of these differences are not apparent when children have experienced high-quality care (Lamb and Ahnert, 2006).

Quality of Care

A second theory predicts that the effects of child care depend on the physical and social environment provided. Harmful effects on both social and intellectual development might be expected if children receive less attention, affection, interaction, and stimulation from nonparental adults than they would from their parents. This view implies that the effects of care depend on its quality and on the quality of the home environment. Quality can be defined by the processes that occur in the setting: (1) sensitive, responsive, positive interactions of adults with children; (2) intellectually stimulating adult actions and activities, including appropriate language, reading, and play; and (3) toys, materials, and curricula that provide age-appropriate opportunities for learning. Quality can also be defined by structural characteristics, some of which can be mandated by regulatory agencies: (1) small ratios of children to adults; (2) small group sizes; (3) caregiver education and training in child development; (4) sufficient space per child; (5) a safe and clean physical environment; and (6) continuity and wages of the staff. Structural and process measures are correlated moderately with one another.

Whether process or structural indicators of quality are used, children in high-quality care have higher levels of language and intellectual skills, and perform better on academic tasks than do children in low-quality care. Although this finding is consistent across studies, there is some disagreement about whether quality produces a large enough difference to be socially significant (Lamb and Ahnert, 2006; Vandell and Wolfe, 2000). In one analysis, the quality of the child care environment and characteristics of the home environment during the first 3 years of life were compared as predictors of school readiness and language skill at age of 3 years. The size of the child care quality effect was about half that of the family environment, suggesting that child care makes a substantial contribution to language and academic development (NICHD, 1999a,b). Critics have also argued that child care effects do not last into later childhood, but evidence for long-term effects is beginning to accumulate. In a large study following children from age of 4 years through grade 3, children who attended higher quality child care centers performed better on measures of cognitive skills (e.g., math and language abilities) while they were in child care and, in many cases, through the end of second grade (Currie, 2001).

High-quality child care also predicts social skills with peers and social competence with adults, but the relations between quality and social behavior are weaker than those between quality and cognitive/academic performance. Some studies suggest that children's relationships with their teachers in child care forecast social skills and relations with teachers in elementary school (Ahnert et al., 2006, Currie, 2001).

One problem in evaluating child care effects is that families with more resources, better education, and more sensitive parenting styles (to name only a few attributes) place their children in higher quality child care settings than do less advantaged families. Most recent studies have measured both family and child care characteristics so that the independent contributions of each could be evaluated, but there is still the possibility that other unmeasured family attributes could affect child care choices and children's cognitive and intellectual development.

Experimental studies, in which children are assigned randomly to enriched early care experiences or to control groups, demonstrate clearly that high-quality early care has lasting effects on cognitive and academic skills, at least for children from economically disadvantaged families. Experiments avoid possible confounds between family and child care attributes. The Abecedarian project enrolled children from low-income families in educational child care from infancy until they reached school age. Children in the treatment group performed better than children in a control group (who received social and nutritional services) on tests of intelligence and school achievement throughout childhood and adolescence (Ramey et al., 2000). Less-intensive early intervention programs for children aged 3–5 years also produced lasting effects on children's school progress; participating children were less likely to be retained in a grade or to need special education than were children in control groups (Reynolds et al., 2001).

The importance of child care does not end when children enter school. Participation in formal after-school care programs that provide cognitive stimulation and positive adult interactions is associated with high academic achievement and low levels of behavior problems, particularly among low-income children (Mahoney et al., 2005). Children without adult supervision in the out-of-school hours are at risk for behavior problems and poor adjustment, particularly if they live in low-income families or unsafe neighborhoods (Pettit et al., 2001).

Type of Care

A great deal of child care occurs in the child's home or someone else's home, especially for infants and toddlers. Caregivers can be grandparents, other relatives, or nonrelatives. In the United States, a small percentage of infants receive center-based care, but the percentage increases as children reach ages 3–5 years. Many western European countries have extensive center-based care for children aged 3 years and over, but care settings for younger children vary considerably (Waldfogel, 2001). The available data suggest both advantages and disadvantages of center-based care relative to home-based care for very young children. Infants and young children in center care show better language and cognitive development by ages 2 and 3 years than those in home-based settings of comparable quality, but they also have more communicable illnesses (e.g., colds and other respiratory illnesses) (NICHD, 2000).

Children attending a center or a child care home with several other children have more experience with peers than do children in other types of home-based care. Children with extensive group experience develop better skills interacting with peers, but their caregivers also rate them higher on negative behavior (e.g., aggression, disobedience) than

children in other forms of care (Lamb and Ahnert, 2006). For young children, social skills and sociability with peers are often accompanied by a certain amount of aggression; perhaps early contact with peers leads to increases in a range of social behaviors.

Do Effects Differ for Children from Different Family Backgrounds?

In the United States, where most child care is funded privately, children from affluent families receive higher quality care than children from families with low and moderate incomes do. In some cases, children from very poor families receive slightly higher quality care than children from families with modest incomes, largely because of publicly funded programs for children in poverty (Johnson et al., 2003). In countries with publicly funded child care for children from all income levels, average quality appears to be considerably better than it is in the United States, and there are not wide discrepancies in quality associated with family income (Lamb and Ahnert, 2006).

Some researchers have proposed a compensatory hypothesis: children from disadvantaged homes might profit from child care of reasonable or high quality because it provides more opportunities for learning and development than they receive at home. The complementary lost resources hypothesis suggests that children from highly advantaged homes may be harmed by child care because it provides fewer opportunities than their home environments do. The evidence for these hypotheses is mixed. Almost all of the experiments exposing children to high-quality care have included only economically disadvantaged children, and these children clearly profit from such care. Some investigations have found support for the lost resources notion, but most studies support the idea that high-quality care provides benefits for children from a wide range of backgrounds (Lamb and Ahnert, 2006).

Public Policy Issues

Public versus Private Funding

Child care is expensive. In some industrialized nations, public funds pay the great majority of the costs. Sweden, France, and some other countries provide large-scale publicly funded, full-day, preschool programs. In the United States, at the other end of the spectrum, limited amounts of public funds are available to low-income families, and tax credits cover a small portion of child care costs for people earning enough to owe taxes. Overall, parents pay the vast majority of the costs of care. At the same time, child care workers and staff receive very low wages. In 1997, teachers in child care centers, many of whom had college degrees, earned from approximately $13 000 to $19 000 per year; salaries of teaching assistants ranged from $10 500 to $12 250 per year (Whitebrook et al., 1998). As a result, there is high turnover among child care workers.

Family Leave versus Infant Care

Care for infants is more expensive than care for preschoolers because they require more concentrated adult attention and they fare best in small groups. Many industrialized nations provide parents with paid family leave as an alternative to supplying center-based infant care (Waldfogel, 2001). In the United States, some workers receive paid family leave, but federal law requires only that parents receive up to 12 weeks of unpaid leave with guaranteed job security and medical benefits. Even this policy applies only to full-time, full-year workers in large organizations. As a result, many infants enter nonmaternal care between ages 2 and 3 months. There is some evidence that mothers who return to work very early and whose babies have extensive nonmaternal care are less sensitive during interactions with their infants than are mothers who have more time at home (Bystrova et al., 2009). It may be more difficult to learn your infant's signals and respond appropriately when you spend a great deal of time at work starting very early in the child's life.

Conclusion

Most children in modern industrialized societies will spend a significant portion of their early lives in child care, and many will require such care during the early school years as well. Initial fears that child care per se would be harmful have not been supported, but there is some evidence that extensive care, beginning in early infancy, can make it more difficult for mothers to relate sensitively to their infants. Whether child care has positive or negative effects on children's development depends primarily on its quality. Quality is defined by sensitive, responsive, and stimulating caregivers and age-appropriate activities and curriculum. High-quality care promotes cognitive, academic, and social development, and its effects last beyond the preschool years. High-quality care for children from all economic levels is provided more successfully in countries that make large public investments in early care than in countries that require parents and child care staff to pay the costs of care.

Future Directions

Nonmaternal care is here to stay. Policy debates about what is 'good enough' care abound, and research is needed to identify thresholds of quality that make large differences, with particular attention to individual child and family characteristics. We understand how early child care can contribute to cognitive and academic performance better than we understand how it can influence positive social behavior with both peers and adults. New designs, including experiments with random assignment, could help to answer these important questions.

See also: Parents and Teachers as Partners in Education.

Bibliography

Ahnert, L., Pinquart, M., Lamb, Michael E., 2006. Security of children's relationships with nonparental care providers: a meta-analysis. Child development 77 (3), 664–679.

Bystrova, K., et al., 2009. Early contact versus separation: effects on mother–infant interaction one year later. Birth 36 (2), 97–109.

Capizzano, J., Adams, G., 2000. The Hours that Children under Five Spend in Child Care: Variations across States. A Report of Assessing the New Federalism. The Urban Institute, Washington, DC.

Currie, J., 2001. Early childhood education programs. Journal of Economic Perspectives 213–238.

Johnson, D.J., et al., 2003. Studying the effects of early child care experiences on the development of children of color in the United States: toward a more inclusive research agenda. Child Development 74 (5), 1227–1244.

Lamb, M.E., Ahnert, Lieselotte, 2006. Nonparental child care: context, concepts, correlates, and consequences. Handbook of Child Psychology.

Lamb, M.E., 1997. Nonparental child care: context, quality, correlates, and consequences. In: Damon, W., Sigel, I., Renninger, K.A. (Eds.), Handbook of Child Psychology, Child Psychology in Practice, fifth ed., vol. 4. Wiley, New York, pp. 73–134.

Mahoney, J.L., Lord, H., Carryl, E., 2005. An ecological analysis of after-school program participation and the development of academic performance and motivational attributes for disadvantaged children. Child development 76 (4), 811–825.

NICHD Early Child Care Research Network, 1997. The effects of infant child care on infant-mother attachment security: results of the NICHD study of early child care. Child Development 68, 860–879.

NICHD Early Child Care Research Network, 1999a. Child care and mother-child interaction in the first 3 years of life. Developmental Psychology 35, 1399–1413.

NICHD Early Child Care Network, 1999b. Child outcomes when child-care classrooms meet recommended standards for quality. American Journal of Public Health 89, 1072–1077.

NICHD Early Child Care Research Network, 2000. The relation of child care to cognitive and language development. Child Development 71, 960–980.

Pettit, G.S., et al., 2001. Antecedents and behavior-problem outcomes of parental monitoring and psychological control in early adolescence. Child development 72 (2), 583–598.

Ramey, C.T., Campbell, F.A., Burchinal, M.R., Skinner, M.L., Gardner, D.M., Ramey, S.L., 2000. Persistent effects of early childhood education on high-risk children and their mothers. Applied Developmental Science 4, 2–14.

Reynolds, A.J., et al., 2001. Long-term effects of an early childhood intervention on educational achievement and juvenile arrest: A 15-year follow-up of low-income children in public schools. Jama 285 (18), 2339–2346.

Vandell, D., Wolfe, B., 2000. Child Care Quality: Does It Matter and Does It Need to Be Improved?, vol. 78. University of Wisconsin-Madison, WI. Institute for Research on Poverty.

Waldfogel, J., 2001. International policies toward parental leave and child care. The Future of Children 99–111.

Whitebrook, M., Howes, C., Phillips, D., 1998. Worthy Work, Unlivable Wages. The National Child Care Staffing Study, 1988–1997. Center for Child Care Workforce, Washington, DC.

Child Care and Development across Cultures

Birgit Leyendecker, Ruhr University Bochum, Bochum, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Every culture has found unique solutions for the universal task of raising healthy children. These solutions may be adaptive and adequate in one cultural context yet they may appear inadequate in another context. Thus, there is no one single right way but rather multiple cultural pathways to children's well-being and healthy development both within and outside of the family. While mothers are the primary caregivers of infants in basically all cultures, the social context in which these relationships are embedded as well as the role of professional caregivers varies greatly.

Tom Weisner (1996) described what happens when he asks students and colleagues about what they consider to be the single most important influence on children's development. The answers are predictable, for example, to provide a safe socioemotional environment, nutrition, medical care, education, etc. As Weisner points out, almost no one comes up spontaneously with what he considers to be the single most important influence: the culture in which a child develops and matures. Whether a child grows up in rural Arkansas or in downtown San Francisco, on a tiny island in the North Sea or in the Amazon rainforest, in Singapore or in a township in South Africa is going to greatly influence his development, the abilities and skills he is going to acquire, and his perception of the world. Culture includes the shared knowledge and shared meaning (cultural interpretation) as well as shared practices that are transmitted and transformed over the generations.

Variations of childrearing practices within a culture are influenced by features such as education, social and economic resources, size of the family, and urban or rural living environment (Keller, 2013). Thus, childrearing practices and the parental ethnotheories are adapted to the resources of a family as well as to the cultural habitat. This has implications for the study of caretaking arrangements and children's development in diverse cultural niches. Childcare practices that may appear meaningful and may make sense in one cultural habitat may appear inadequate and senseless in another habitat (Keller, 2013). Therefore, research on early childcare and parenting practices are presented with the challenge of disentangling cultural effects on children's development from the environmental features and the resources available to families.

The acknowledgment of "the startling, disturbing, wonderful variety of what it means to be a child or parent and the variety of forms development can take" (Weisner, 1996: p. 306) has challenged the conception of the universal child and has shifted the focus to understanding the "child in context" (Bronfenbrenner, 1979). Super and Harkness (1986) have developed the concept of the developmental niche as a framework to understand how culture influences the organization of the developmental environment and thereby affects the course of development. A developmental niche is conceptualized by three interacting subsystems: the physical and social settings of a child's everyday life, the customs of childcare and childrearing, and the psychology of the caretaker. These subsystems account for similarities across developmental niches due to similar influences. Childcare and childrearing practices are, for example, associated with other customs in the child's developmental niche that have proven to be successful and adaptive. Each generation of parents is likely to modify childrearing practices according to their needs, yet instead of reinventing the wheel anew in each generation they are likely to build their childcare practices on the cultural practices of previous generations. These subsystems also account for differences and variability in relationships between children's development and their environment. For example, variability may result due to the personality of the parents and their individual resources or the children's gender and temperament (Super and Harkness, 2002).

Culture Matters in Human Development

Despite the enormous variety of developmental niches and the customs of childcare and childrearing within these niches, there is at least one overarching similarity: Regardless of their cultural background, all parents want to foster their children's healthy development. They want them to survive and to thrive, to grow up to function well in their everyday life, to become well adjusted, responsible members of the society who get along with others, can comply to rules and authority, and who are able to provide for themselves economically. What constitutes a responsible member of the society, what is meant by providing for oneself, and how these goals can be achieved may differ greatly both within and across cultures (Keller, 2013; Leyendecker et al., 2002). Accordingly, Keller (2013: p. xiii) describes the "developmental pathways as cultural solutions for universal developmental tasks." These developmental pathways and associated childrearing practices are not static but are likely to change gradually (or rapidly) over time as sociodemographic ecologies change and cultural values are adapted accordingly (Greenfield, 2009).

In her research, Rogoff and her colleagues (Rogoff et al., 1993; Rogoff, 2003) have addressed the question of how culture matters in human development, and how children acquire knowledge relevant to their sociocultural niche. Rogoff (2003) differentiates between child-focused activities and children's participation in mature activities in their community. Children's participation in adult activities requires children to observe the ongoing activities. Adults or older children

may facilitate and support their participation. In contrast, structured child-focused activities require that an adult caregiver provides the learning opportunities and structured teaching situations, e.g., teaching a child how to insert wooden blocks of different shapes into a box with openings that match these shapes. Parents in Western cultures who are more likely to provide child-focused activities are also more likely to provide a daily schedule for their infants and young children with fixed times for sleeping, napping, and eating, and more opportunities for dyadic one-on-one interactions (Leyendecker et al., 1995). In contrast, children in non-Western cultures are more likely to participate in shared activities with the members of their community, their daily schedule is likely to be centered around an individualized schedule, and they are likely to have more opportunities for interactions with multiple members of their family and other people who make up their daily social environment (Leyendecker et al., 1995; Rogoff, 2003). All developmental niches include participation in adult activities, guided participation, and specialized child-focused activities, yet the likelihood that children are presented with these different learning opportunities is likely to vary across developmental niches.

Cultural Variability of Caretaking Arrangements: Mothers and Others

From a psychological perspective, the quality of children's relationships with their parents and nonparental caregivers, as well as the availability of adequate economic, social, and physical resources are likely to promote children's healthy development and adjustment (Lamb, 2012). Once the enormous variety of cultures is taken into account, it becomes evident that there is no single best pathway to foster children's well-being, healthy development, and adjustment. Instead, there are as many pathways as there are different cultures in which children are raised (Greenfield et al., 2003; Keller, 2013). Especially in early childhood, the caretaking arrangements have been at the center of numerous debates on what is considered appropriate for children's healthy development. Keller (2013) criticizes that Western developmental psychology has adopted the norm that infants form an attachment bond primarily to the person they spend most of their time with – their mothers. She points out that from an evolutionary point of view as expressed by Hrdy (1999) "humankind would not have survived if solely mothers had been infants' caretaker" (Keller, 2013: p. 182). Instead, mothers and children benefit from alloparental help in childcare. Therefore, in line with Hrdy's (1999) work, she strongly advocates including mothers as well as others as potentially important caretakers and attachment figures in cross-cultural research on early child development.

Regardless of the cultural context they are born into, infants are likely to have their mothers as their primary caretaker in the earliest stages of life. Keller (2013) assumes that besides breastfeeding and nursing, other biological roots may foster this exclusive bond over the first 6 months of life. Aside from these similarities, the social environment the mother–infant pair is embedded in is likely to differ greatly. While differences between nuclear Western families and extended large non-Western families are most likely suitable to demonstrate the vast differences in social environments, variability within similar ecologies can occur as well. A study comparing the social environment of 10-week-old infants in diverse cultural settings in Europe, North and South America, and Africa found that with one exception, mothers were present or responsible for the child between 63 and 81% of the time the infants were awake (Schölmerich et al., 2002). The one exception was mothers of the Ngandu people in Central Africa. Interestingly, the two African samples from the Central African Republic – Aka and Ngandu – live in close proximity to each other and they can be described as traditional, small-scale, preindustrial, and non-Western (Berry et al., 1986; Hewlett et al., 2000). These two culturally distinct groups live in the same tropical forest yet their patterns of childcare vary greatly. The Aka are foragers (hunter-gatherers) who move their camps several times a year, have a high gender egalitarianism with men and women contributing equally to subsistence. In contrast, the Ngandu are sedentary farmers, have marked gender inequality with women receiving less although they are the primary providers and in charge of working the fields. When mothers were working in the fields, they were within earshot of their infants; however, older siblings were likely to be in charge. While both Aka and Ngandu mothers held their infants frequently, the Aka mothers were twice as likely to carry their infants and to be in constant body contact, and fathers were more likely to be involved in childcare (Hewlett, 2004; Schölmerich et al., 2002). While the social environment of infants in these two groups is markedly different, they still have more opportunities for interactions with multiple caregivers in a group setting when compared to infants growing up in a typical Western urban environment.

In her book *Cultures of Infancy*, Keller (2013) describes that while in basically all cultural environments mothers are the ones who are most likely to be in charge and responsible for the infant, differences lie in the social context – the amount of time siblings, kin, and others are in proximity and potentially involved. Keller (2012) argues that this contextual variability is largely ignored in attachment research and that the so-called Strange Situation developed by Ainsworth and her colleagues (Ainsworth et al., 1978) is suitable only for children growing up in an urban Western environment. While Keller recognizes the importance of the child's mother, she argues that a culturally sensitive attachment theory has to take the role of the social context into account as well. In a social context with many other caring adults or siblings present and involved, the mother–child relationship may be less exclusive as the child is used to be taken care of by others. Thus, these children are less likely to be distressed when their mothers are not present and more likely to allow other adults to pick them up and take care of them (Keller, 2012, 2013).

Conversely, Lamb (2012) criticizes the idea that the role of the biological mother is considered to be essential to children's well-being while all other potentially important attachment figures and caretakers are largely ignored. He states that relatives and nonrelatives alike, regardless of their gender, are able to take care of both younger and older children. In Western societies, the traditional family with two biological parents is more often replaced by a large variety of nontraditional families. Lamb (2012) argues that parents' capacity for good

parenting does not depend on their gender and that single-headed households as well as nonbiological parents do not put children at risk. Instead, availability of competent, caring, and nurturing adults, in conjunction with the availability of sufficient social and financial resources, contributes to the outcome of children in traditional and nontraditional families alike.

Early Childhood Education and Care by Professional Caretakers

In the first part of this article, the focus was on the enormous cultural variability of early childcare arrangements in diverse cultural niches. While mothers are the primary caretakers in almost all cultures, the social context in which these relationships are embedded may differ greatly. In addition, relatives and nonrelatives are likely to play an increasingly important part as the infants grow older. In the final part of this article, the focus shifts to the organized nonrelative caretaking arrangements, which are available, for example, in Europe and North America. Formal childcare arrangements differ greatly from the informal childcare arrangements with relatives or others who share either the same household or live in close proximity. Aside from live-in caretakers, most childcare arrangements with professional caretakers take place outside of the family in a separate building with little or no contact between parents and children during this period. This has raised much concern among parents and psychologists. Here, the question will be addressed whether nonparental formal childcare is beneficial for young children and whether qualitative nonparental childcare is available and accessible for families from all socioeconomic strata.

Despite great advances in research on early childcare, such as the NICHD Study of Early Child Care (e.g., NICHD Early Child Care Research Network, 2002), there are still many open questions. Is nonrelative childcare of toddlers a source of enrichment that promotes cognitive and linguistic development or is childcare at this young age associated with developmental risks? Can high-quality childcare counterbalance low-quality care in the home environment? Should we encourage families to enroll their children at a young age in a nursery school? In the United States, Canada, and some parts of Europe, many small as well as several large-scale studies have addressed these questions. While findings based on research on the relationship between quality and especially quantity of early childcare on social skills and externalizing behavior problems are mixed (e.g., Loeb et al., 2007; Love et al., 2003; Vandell et al., 2010), there is general agreement about the association between high quality of childcare and children's advanced cognitive and linguistic functioning (Belsky, 2006; Love et al., 2005; NICHD ECCRN, 2002; Ruzek et al., 2014). These effects were still visible during middle childhood and adolescence. Children attending high-quality childcare were found to receive higher scores on tests of math, reading achievement, and memory in elementary school (NICHD Early Child Care Research Network, 2005), and higher scores on cognitive academic achievement later in adolescence (Vandell et al., 2010). An analysis of the Early Head Start program in the United States found significant gains in cognitive and language functioning in 2- and 3-year-old children from low-income families enrolled in high-quality programs (Love et al., 2003). Burchinal et al. (2000) who analyzed data from three large childcare studies found support that the quality of childcare is related to developmental outcomes of all children regardless of their social risk status. For minority children, however, the advantages of high-quality childcare appear to be most pertinent to their language development (Burchinal et al., 2000; Buysse et al., 2013).

Accessibility of Early childcare

As Belsky (2009) points out, children of all ages benefit from high-quality care relative to those experiencing poor-quality care. Ideally, nonparental care should be at least as stimulating as parental care (Ahnert et al., 2006). Nonparental caretakers, who are attentive to children's verbal and nonverbal signals, respond adequately, and provide a stimulating learning environment can foster children's development. Aside from the issue of quality of care, there is the issue relating to what extent children have access to nonparental care. In some countries, such as the Netherlands, Sweden, France, and Switzerland, early childhood education and care is mostly publicly funded for children from the ages of 3 or 4 years. In other countries, such as Germany, Norway, and Austria, early childhood education and care is publicly available but not free of charge. In Germany, for example, some municipalities offer childcare free of charge. Most families, however, have to pay for childcare before children start school. Even though these fees are quite low for low-income families, children from these families are less likely to attend nursery schools (for the ages 1–3) and are likely to enroll in the so-called Kindergarten (ages 3–6) at a later age (Leyendecker et al., 2014; Tietze et al., 2013) About one-third of all children in Germany have at least one immigrant parent. While immigrant families are from all social strata, they are most likely to have less financial and educational resources when compared to the German majority. They are less likely to be familiar with strategies on how to enroll a child into public childcare at an early age. In addition, relatives such as grandparents who may support parents are less likely to be available and involved when compared to nonimmigrant German families (Tietze et al., 2013).

In sum, children all over the world are likely to be cared for primarily by their mothers in the first 6 months of their life. Aside from this parallel, their social environment, however, is likely to differ greatly and influence their early experiences. Despite the great variety of developmental niches, there is general agreement that all children "thrive psychologically, when they are nurtured by caring competent parents who are themselves embedded in warm supportive relationships" (Lamb, 2012: p. 101). The parental role can be filled by the child's biological or nonbiological parents as well as competent older children or caring adults. In their first years of life, some children especially in Western cultures may be cared for primarily by their parents. Most children around the world, however, are likely to be cared for by parental and nonparental caregivers. These early experiences and the quality of care these caregivers provide are part of their diverse cultural pathways, and will influence children's development.

See also: Cross-Cultural Psychology; Cultural Psychology; Culture and the Self: Implications for Psychological Theory; Ethnicity and Migration in Europe; Human Development, Theories of; Immigration, Citizenship, and Integration: Social Work Connections; Parenting Attitudes and Beliefs across Cultures; Prenatal and Infant Development: Overview, Current Trends, Future Directions; Residential Child Care.

Bibliography

Ainsworth, M.D.S., Blehar, M.C., Waters, E., Wall, S., 1978. Patterns of Attachment: A Psychological Study of the Strange Situation. Erlbaum, Hillsdale, NJ.

Ahnert, L., Pinquart, M., Lamb, M.E., 2006. Security of children's relationship with nonparental care providers: a meta-analysis. Child Development 74, 664–679.

Belsky, J., 2006. Early childcare and early child development: major findings of the NICHD study of early child care. European Journal of Developmental Psychology 3 (1), 95–110.

Belsky, J., March 2009. Effects of Child Care on Child Development: Give Parents a Real Choice. Birbeck University, London. http://www.noviny-mpsv.cz/files/clanky/6640/9_Jay_Belsky_EN.pdf (accessed 11.10.13.).

Berry, J.W., van de Koppel, J.M.H., Senechal, C., Annis, R.C., Bahuchet, S., Cavalli-Storze, L.L., Witkin, H.A., 1986. On the Edge of the Forest: Cultural Adaptation and Cognitive Development in Central Africa. Swets & Zeitlinger, Lisse.

Burchinal, M.R., Peisner-Feinberg, E., Byrant, D.M., Clifford, R., 2000. Child-care quality: testing for differential associations related to poverty, gender, or ethnicity. Applied Developmental Science 4 (3), 149–165.

Bronfenbrenner, U., 1979. The Ecology of Human Development. Harvard University Press, Cambridge, MA.

Buysse, V., Peisner-Feinberg, E., Paez, M., Haymmer, C., Knowles, M., 2013. Effects of early education programs and practices on the development of learning of dual language learners: a review of the literature. Early Childhood Research Quarterly. http://dx.doi.org/10.1016/j.ecresq.2013.08.004.

Greenfield, P.M., Keller, H., Fuligni, A., Maynard, A., 2003. Cultural pathways through universal development. Annual Review of Psychology 54, 461–490.

Greenfield, P.M., 2009. Linking social change and developmental change: shifting pathways of human development. Developmental Psychology 45 (2), 401–418.

Hewlett, B.S., Lamb, M.E., Leyendecker, B., Schölmerich, A., 2000. Internal working model, trust and sharing among foragers. Current Anthropology 41 (2), 287–297.

Hewlett, B.S., 2004. Fathers in forager, farmer, and pastoral cultures. In: Lamb, M.E. (Ed.), The Role of the Father in Child Development. Wiley, Hoboken, NJ, pp. 182–195.

Hrdy, B.S., 1999. Mothers and Others. The Evolutionary Origins of Mutual Understanding. Harvard University Press, Cambridge.

Keller, H., 2013. Cultures of Infancy, second. ed. Erlbaum, Mahwah, NJ.

Keller, H., 2012. Attachment and culture. Journal of Cross-Cultural Psychology 44 (2), 175–194.

Lamb, M.E., 2012. Mothers, fathers, families, and circumstances: factors affecting children's adjustment. Applied Developmental Science 16 (2), 98–111.

Leyendecker, B., Lamb, M.E., Schölmerich, A., Fracasso, M.P., 1995. The social worlds of 8- and 12 months-old infants: early experiences in two subcultural contexts. Social Development 4, 194–208.

Leyendecker, B., Harwood, R.L., Lamb, M.E., Schölmerich, A., 2002. Mothers' socialisation goals and evaluations of desirable and undesirable everyday situations in two diverse cultural groups. International Journal of Behavioral Development 26 (3), 238–258.

Leyendecker, B., Citlak, B., Schräpler, J., Schölmerich, A., 2014. Bildungserwartungen beim Übergang in die Grundschule – Ein Vergleich von deutschen und zugewanderten Eltern. Zeitschrift für Familienforschung/Journal of Family Research 26 (1), 70–93.

Loeb, S., Bridges, M., Bassok, D., Fuller, B., Rumberger, R.W., 2007. How much is too much? The influence of preschool centers on children's social and cognitive development. Economics of Education Review 26 (1), 52–66.

Love, J.M., Kisker, E.E., Ross, C., Raikes, H., Constantine, J., et al., 2005. The effectiveness of early head start for 3-year old childen and their parents: lessons for policy and programs. Developmenta Psychology 41, 885–901.

Love, J.M., Harrison, L., Sagi-Schwartz, A., van IJzendoorn, M.H., Ross, C., Ungerer, J.A., Raikes, H., Brady-Smith, C., Boller, K., Brooks-Gunn, J., Contantine, J., Kisker, E.E., Paulsell, D., Chazan-Cohen, R., 2003. Child care quality matters: how conclusions may vary with context. Child Development 74 (4), 1021–1033.

Magnuson, K.A., Waldfogel, J., 2005. Early childhood care and education: effects of ethnic and racial gaps in school readiness. The Future of Children 15 (1), 169–196.

NICHD ECCRN, 2002. Early child care and children's development prior to school entry: results from the NICHD study of early child care. American Educational Research Journal 39 (1), 133–164.

NICHD ECCRN, 2005. Early child care and children's development in the primary grades: follow-up results from the NICHD study of early child care. American Educational Research Journal 43, 537–570.

Rogoff, B., Mistry, J., Göncü, A., Mosier, C., 1993. Guided participation in cultural activity by toddlers and caregivers. Monographs of the Society for Research in Child Development 58.

Rogoff, B., 2003. The Cultural Nature of Human Development. Oxford University Press, Oxford.

Ruzek, E., Burchinal, M., Farkas, G., Duncan, G., 2014. The quality of toddler child care and cognitive skils ar 24 months: propensity score analysis results from the ECLS-B. Early Childhood Research Quarterly 29, 12–21.

Schölmerich, A., Leyendecker, B., Lamb, M.E., Hewlett, B.S., Tessier, R., 2002. Alltagserfahrungen von 3 Monate alten Säuglingen in verschiedenen Kulturen. In: Alt, K., Kemkes-Grottenthaler (Hg.) Kinderwelten, A. (Eds.), Anthropologie – Geschichte – Kulturvergleich. Böhlau Verlag, pp. S386–S399.

Super, C., Harkness, S., 1986. The developmental niche: a conceptualization at the interface of child and culture. International Journal of Behavioral Development 9, 545–569.

Super, C., Harkness, S., 2002. Culture structures the environment for development. Human Development 45 (4), 270–274.

Tietze, W., Becker-Stoll, F., Bensel, J., Eckhardt, A.G., Haug-Schnabel, G., Kalicki, B., Keller, H., Leyendecker, B. (Eds.), 2013. Nationale Untersuchung zur Bildung, Betreuung und Erziehung in der frühen Kindheit (NUBBEK). Verlag das netz.

Vandell, D.L., Belsky, J., Burchinal, M., Steinberg, L., Vandergrift, N., 2010. NICHD early child care research network. Do effects of early child care extend to age 15 years? Results from the NICHD Study of Early Child Care and Youth Development Child Development 81 (3), 737–756.

Weisner, T., 1996. Why enthnography should be the most important method in the study of human development. In: Jessor, R., Colby, A., Shweder, R. (Eds.), Ethnography and Human Development. Chicago University Press, Chicago, pp. 305–325.

Child-Directed Speech: Influence on Language Development

Rupa G Gordon and Linda R Watson, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Child-directed speech (CDS), or speech adjustments made by adults when interacting with children, is characterized by higher pitch, more repetition, simplified vocabulary, and exaggerated intonation patterns. Here, we review the role of CDS in enhancing language acquisition. Not only have strong global relationships have been found between maternal linguistic input and language development, but also specific features of CDS (e.g., higher pitch and exaggerated prosody) have been discerned that seem to assist in language learning. Finally, we review evidence from developmental disorders further support the relationship between CDS and language acquisition.

Introduction

Child-directed speech (CDS) refers to speech adjustments made by adults when speaking to children and infants. CDS differs from adult-directed speech (ADS) as it is characterized by higher pitch, greater pitch range, exaggerated intonation patterns, elongated vowels, shorter utterances, more repetition, more questions, simplified vocabulary, and less use of past tenses (Pine, 1994; Snow, 1977). In general, the characteristics of CDS are believed to serve many important functions, including (1) enhancing language acquisition, (2) gaining and holding the infant's attention, (3) developing the affective bond between infant and mother, and (4) facilitating social interaction and early communication (Sachs, 1977). Beyond speech, adults also make other modifications to their communicative behaviors during interactions with children, including the use of simplified gestures and exaggerated facial expressions and movements (Brand and Shallcross, 2008; Chong et al., 2003; Iverson et al., 1999; O'Neill et al., 2005).

Infants display strong preferences for CDS compared to ADS as early as 1 month of age (Cooper and Aslin, 1990), regardless of the speaker's gender (Pegg et al., 1992), and this preference continues through toddler (Kuhl et al., 2004; Paul et al., 2007) and preschool years (Klin, 1991). They also show increased neural responses to CDS compared to ADS as measured by Event-Related Potentials (ERP) (Zangl and Mills, 2007) and Near-Infrared Spsectroscopy (Saito et al., 2007). Furthermore, this difference in brain activity is greatest when the CDS is produced by the infant's own mother (Naoi et al., 2012). Infants also display strong social preferences for speakers who use CDS, suggesting that CDS plays an important role in the development of social bonds. Schachner and Hannon (2011) found that 5-month-old infants preferred to look at an individual who previously used CDS compared to a novel individual, and this preference is not found for individuals who used ADS.

The characteristics of CDS change as a function of the child's age and developmental abilities. Stern et al. (1983) found that CDS to neonates (2–6 days old) is characterized by long pauses, short utterances, and steady rhythms, as neonates require low levels of stimulation and social interaction. In contrast, at 4 months old, the child is much more active and the mother uses the most exaggerated features of CDS, including pitch and repetition. At 12 and 24 months, when the child is exploring independently and playing with objects, the mother facilitates the child's explorations by using speech that is more object focused, and CDS is characterized by longer durations of vocalizations and more complex utterances (Adamson and Bakeman, 1984; Stern et al., 1983).

The aim of this article is to review evidence that the modifications associated with CDS impact language acquisition. Note that although many studies have focused on maternal use of CDS and this article will reference maternal behaviors frequently, many of these findings are also applicable to fathers and other adults (e.g., Fernald et al., 1989; Jacobson et al., 1983; Papoušek et al., 1987).

Associations between Maternal Linguistic Input and Language Acquisition

Some early studies, notably the work of Newport et al. (1977) discounted the role of CDS in language acquisition because their findings suggested no relationship between the simplicity of maternal language and children's language acquisition. They argued that CDS might not be beneficial for language learning as its apparent simplifications often involve deletions that may make the content more difficult for the child to understand. Despite these early findings, current research has found strong associations between the use of CDS and language acquisition, and has also discerned specific characteristics of CDS that benefit language learners.

Multiple studies have examined the global relationship between maternal linguistic input and language acquisition using longitudinal correlational designs. Several aspects of maternal speech at 18 months, including utterance length and use of verbs and pronouns predicted characteristics of children's speech at 27 months, such as utterance length and verbs use (Furrow et al., 1979). Numerous studies have found that maternal linguistic input and vocabulary diversity are positively associated with vocabulary development (Hart and Risley, 1995; Hoff and Naigles, 2002; Huttenlocher et al., 1991; Weizman and Snow, 2001).

Furthermore, certain interactive features of early caregiver–child communication appear to promote language development. For example, children whose mothers engaged with them in more joint interactions, rather than directive

interactions, developed larger vocabularies (Tomasello and Todd, 1983), and maternal responsiveness to vocalizations at both 9 and 13 months predicted the child's achievement of certain milestones of expressive language, including first words, first 50 words, and first use of language about the past (Tamis-LeMonda et al., 2001). Although directive maternal utterances have been associated with poorer child language outcomes in numerous studies of typically developing children (e.g., Hampson and Nelson, 1993; Tomasello and Farrar, 1986), other research points to the need for more nuanced analyses of caregiver directiveness. For example, caregiver directives given when the caregiver and child share an attentional focus are associated with more positive child language outcomes in typically developing children (Akhtar et al., 1991; Masur et al., 2005). These findings collectively point to the potential benefits of providing language input that is responsive to the child's focus of attention rather than redirecting the child's focus. But this generalization also may be too broad. A recent study of low-income African-American mother–child dyads in the United States found that the number of directive object labels (i.e., labels for objects outside of the child's focus of attention) that mothers used was positively associated with the children's later vocabulary, whereas the number of responsive object labels (labels for objects within the child's focus of attention) was not (Shimpi et al., 2012). Thus, there remain considerable gaps in understanding the complexities of the relationships between CDS and child language outcomes, and the mechanisms accounting for these relationships.

It is critical to remember when interpreting the findings of these correlational studies that social interaction involves mutual and reciprocal influences from both interactants. Indeed, an important influence on the mother's behaviors and responsiveness are the infant's own behaviors. Behaviors such as smiling, crying, and infant vocalizations influence the mother's responsiveness and speech to the infant (Henning et al., 2005), as do more explicit uses of gestures (e.g., pointing) and infant sign language (Vallotton, 2009). Experimental manipulations of the feedback mothers receive from the infant also provide strong support for this idea. For example, when mothers interacted with their child via video, but the video feed of the infant was surreptitiously delayed, mothers tended to use more interrogatives than when it was truly live, suggesting that mothers are sensitive to important behaviorally contingent feedback cues from the infant (Murray and Trevarthen, 1986). Mothers also tend to raise their pitch when they receive positive feedback from the child (Smith and Trainor, 2008).

Identifying Features of CDS That Impact Language Acquisition

Correlational studies describing the relationship between CDS and language acquisition suggest a strong relationship, but it is difficult to parse out the specific influences of CDS. More recently studies have experimentally manipulated aspects of CDS in order to identify specific characteristics of CDS that can benefit language development. Research has shown that the impact of different features of CDS on language acquisition varies with the age of the child (Rowe, 2012). Indeed, CDS seems to be advantageous for language acquisition, particularly for infants and young children. In infants (7–8 months of age), long-term learning (delay of 24 h) only occurred for words presented in CDS compared to ADS (Singh et al., 2009). Similarly, 21-month-old children learned novel words better when presented using CDS rather than ADS, but by 27 months children learned words equally well in both conditions (Ma et al., 2011), highlighting the relevance of considering the child's age and language level as a factor in evaluating the impact of CDS on language acquisition.

Pitch and Prosody

The vast majority of research on the specific differences between CDS and ADS has focused on pitch and prosody. The higher pitch and exaggerated pitch contours of CDS serve multiple functions, including gaining the infant's attention, communicating intent and emotion, and supporting language acquisition.

The higher pitch of CDS is ideal for gaining the infant's attention because the fundamental frequency of CDS is typically within 400–600 Hz, falling in the optimal orienting range for infants (500–2000 Hz) (Sachs, 1977). ADS, on the other hand, typically falls within a lower, less salient range (e.g., 220 Hz for a woman; Sachs, 1977; Trainor and Desjardins, 2002). The prosodic patterns of CDS seem to convey communicative intent better than ADS, as adults are able to identify the speaker's intent with higher accuracy in CDS compared to ADS (Fernald, 1989). In fact, even non-English speaking adults from a nonindustralized, nonliterate culture are able to infer intentions from English CDS better than ADS, suggesting that there are universally recognized prosodic cues in CDS (Bryant and Barrett, 2007). The pitch height and exaggerated pitch contours of CDS also enhance the vocal expression of emotion, and therefore may support the development of affective bonds between mother and child (Trainor et al., 2000).

The pitch and prosody characteristics of CDS also serve other functions that may assist in language acquisition. Prosody in CDS may facilitate speech processing by highlighting and emphasizing target words, whereas prosody in ADS is more variable and less likely to highlight target words (Fernald and Mazzie, 1991). The exaggerated pitch contours of CDS are beneficial for discriminating and learning vowel categories (Trainor and Desjardins, 2002; Vallabha et al., 2007), as well as syntax, as it contains a greater number and more reliable markers of major syntactic units than ADS (Nelson et al., 1989). Prosody in CDS facilitates word segmentation from fluent speech, evidenced by infants' ability to distinguish words from nonsense sentences when the sentences exhibit the intonation patterns and structure of CDS rather than ADS (Thiessen et al., 2005). Finally, a prospective study found that 20 month children who heard CDS that was characterized by reduced prosodic stress on target words and less varied prosodic contours had reduced vocabulary development at 24 months, suggesting that the lack of prosodic stress in CDS potentially impaired the child's ability to segment and highlight target words (D'Odorico and Jacob, 2006).

Other Features of CDS

There are multiple characteristics of CDS beyond prosody and pitch that contribute to language acquisition. The simplicity and repetitive nature of CDS is believed to serve the purpose of reducing confusion and enhancing gains in language acquisition (e.g., Snow, 1977 but see Newport et al., 1977). Specifically, the frequency with which mothers use words impacts the acquisition of vocabulary (Goodman et al., 2008) and the frequency with which words are spoken in isolation (i.e., not as part of a fluent speech stream, a common occurrence in CDS) predicts acquisition better than the total frequency of exposure in a naturalistic longitudinal study (Brent and Siskind, 2001). Similarly, exposure to words in isolation as well as in fluent speech enhanced infants' learning of statistical properties of words in an experimental study, whereas exposure to the words only in fluent speech alone did not (Lew-Williams et al., 2011). Furthermore, the variability with which a word is presented (i.e., used by multiple speakers in multiple contexts) also positively influences word recognition and generalization (Singh, 2008). The phonetic units of CDS are altered so that the vowel space is expanded compared to ADS, such that vowels are clearer, longer, and more distinct from each other (Kuhl et al., 1997). Moreover, the manner in which the phonetic cues to distinguish vowels are altered in CDS is language specific. Japanese mothers alter the phonetic cues that are more relevant for vowel discrimination in Japanese, while English mothers alter those cues that are more relevant for vowel discrimination in English (Werker et al., 2007). Word segmentation in multiple languages is also facilitated by the use of diminutives (e.g., doggie, Stevie) by increasing invariance in word endings and stress patterns (Kempe et al., 2007).

Extralinguistic Features of Child-Directed Communication

Recent research suggests that adults make other communicative alterations when talking to children beyond changing the characteristics of their speech (Iverson et al., 1999; O'Neill et al., 2005). Gestures are more concrete and simplified when used with children. Gestures such as pointing and showing were used more often to support speech and disambiguate and emphasize verbal references (Iverson et al., 1999; O'Neill et al., 2005). Furthermore, emphatic or beat gestures that are common in ADS, such as making circular motions with a flat hand while speaking, are not typically found in CDS (Iverson et al., 1999; O'Neill et al., 2005).

Other aspects of the interaction tend to be exaggerated, such as movements and facial expressions (Green et al., 2010; Stern, 1974). Distinct facial expressions have been found in infant interactions of both English-speaking and Chinese-speaking mothers that are not typically found in adult interactions (e.g., mock surprise, pursed lips) (Chong et al., 2003). Mothers also modify their movements, termed motionese, in order to enhance attention and to highlight aspects of the movement and concurrent speech (Brand and Shallcross, 2008; Rutherford and Przednowek, 2012). Specifically, Brand et al. (2002) found that when mothers demonstrated novel objects to their children vs to an adult, they used more exaggerated movements, displayed more enthusiasm, and used movements that were simpler, larger, and more repetitive. Furthermore, infants preferred watching motionese compared to movements produced in adult-directed interactions (Brand and Shallcross, 2008; Koterba and Iverson, 2009) and infants spent more time engaged with those objects for which the adults used more repetition in their movements (Koterba and Iverson, 2009).

Mothers also integrate the use of multiple modalities when talking to children, which positively influences language learning. The presentation of visual information (e.g., moving or manipulating an object) closely synchronized with auditory information from speech increases the saliency of the target. This type of temporal synchrony is used to highlight novel word-referent relationships, and children spend more time attending to targets and show better learning of new words (Gogate et al., 2000, 2006).

Necessity of Social Interaction

Exposure to speech spoken using the characteristics of CDS may not be enough to influence language learning; rather, at least for infants, the ability to learn language also requires social interaction. Nine-month-old American infants were exposed via either live sessions or video recordings to Mandarin spoken using CDS over 12 sessions in which the speaker read books and played with toys. Only those children who had been exposed to the live sessions were able to learn a phonetic contrast that only occurs in Mandarin (Kuhl et al., 2003). The authors suggest that social and referential cues found only in live interactions, such as shared gaze and joint attention, may be critical for language learning.

Furthermore, social interaction also impacts the development of the production of speech and babbling in infants. Adult behaviors in response to infant vocalizations such as speech, smiling, and touching increases the production of infant vocalizations (e.g., Bloom et al., 1987; Ramey et al., 1972). Goldstein et al. (2003) found that adult contingent responses to 8-month-old infant vocalizations not only increased the frequency of vocalization, but also increased the proportions of vocalizations that had more developed voicing, syllable structure, and faster consonant–vowel transitions compared to baseline. Furthermore, when adult contingent behavioral responses were vocalizations, the infants altered their babbling according to the phonological patterns of the adult's speech (Goldstein and Schwade, 2008).

Universal Nature of CDS

If CDS is important for language acquisition, then these modifications should be found across languages and cultures. Indeed, modifications of speech for children are highly universal. Many of the characteristics of CDS, including shorter utterances, longer pauses, and expanded pitch contours, occur in cultural linguistic contexts that include native speakers of American English, British English, French, Italian, German, Japanese, and Mandarin (Fernald et al., 1989; Grieser and Kuhl, 1988). Although there are differences in the exact features of CDS across languages (e.g., Pye, 1986), and the degree to which

modifications of certain features occur (e.g., alterations in prosody are more pronounced in American English than other languages to which it has been compared (Fernald et al., 1989)), caregivers across virtually all cultures display modifications when speaking to children.

Child-directed alterations also occur in sign language. Masataka (1992, 1996) found that compared to when signing with adults, deaf mothers signed slower and used more repetitions and exaggerated movements when signing to their deaf infants. The infants also spent more time attending to child-directed signs compared to adult-directed signs.

CDS and Language Acquisition in Developmental Disabilities

Alterations in CDS to children with developmental delays in language acquisition lend further support for the relationship between CDS and language development. Compared to mothers of typically developing children of the same chronological age, mothers of children with Down syndrome produced more simplified speech, including fewer nouns (Cardosomartins et al., 1985), reduced syntactic complexity (Buium et al., 1974), exaggerated prosody (Fidler, 2003), and less complex gestures and combinations of gesture and utterances (Iverson et al., 2006). Mothers with children with autism spectrum disorder (ASD) modified their speech based on the level of functioning of their child, such that mothers of children with higher functioning verbal skills asked more questions, answered more questions, and reinforced the use of language, while mothers of children with lower functioning verbal skills used more directives, shorter utterances and reinforced the child's motor actions (Konstantareas et al., 1988).

Some argue that these modifications signify that mothers are sensitive to their child's developmental state (e.g., Konstantareas et al., 1988), while others argue that this lack of complexity in language may be detrimental to the child's ability to develop language (e.g., DePaulo and Bonvillian, 1978). In particular, the increase in maternal directiveness with children with developmental disabilities has been argued to negatively impact the interaction. Mothers of these children tend to exhibit more verbal and nonverbal behaviors that seek to control or direct the child's behaviors and actions (Marfo, 1990). Furthermore, correlational studies have found that mothers of infants with language impairments tend to spend less time engaged in conversational interaction, and tend to be less responsive compared to mothers of typically developing children (Alston and St James-Roberts, 2005; Hammer et al., 2001; Wulbert et al., 1975). Parents of children with ASDs who make more nondemanding comments that follow in to their child's focus of attention have children with better long-range language outcomes (Siller and Sigman, 2002). Similar to research on typically developing children, however, caregiver directives given when the caregiver and child share an attentional focus are associated with more positive child language outcomes in children with ASDs (Haebig et al., 2013a,b).

Importantly, many of these studies are correlational and child's behaviors likely influence the amount and quality of maternal language input at the same time that CDS impacts child language acquisition. For example, compared to typically developing children, children with ASD show a reduced preference for CDS over other stimuli (Klin, 1991; Kuhl et al., 2004; Paul et al., 2007), and attention to CDS is positively associated with concurrent language skills (Kuhl et al., 2004; Paul et al., 2007; Watson et al., 2010). Children with ASD also do not show the typical neural response to CDS as measured by EEG (Kuhl et al., 2004). These differences in child responsiveness to CDS may have a transactional impact on the quantity and quality of CDS that caregivers use with children with ASD that may in turn further restrict optimal language learning conditions for the children (e.g., Wan et al., 2013).

To address some of these issues, many interventions have focused on teaching parents how to become more responsive and facilitate mutual interactions. A recent meta-analysis found that parent-implemented language interventions, which often include a focus on increasing parent responsivity to their children, do have a positive impact on the expressive and receptive language skills of children with language impairments (Roberts and Kaiser, 2011).

Effects of Maternal Depression

Providing further support for the idea that CDS impacts language acquisition is the concomitant alteration in CDS and learning that occur related to maternal depression. Maternal depression has numerous effects on infant–mother interactions. The social interactions of depressed mothers tend to be reduced in quality, such that the less the mothers facilitate interactions, the less the children tend to be sociable, even with strangers (Stein et al., 1991). In terms of CDS, depressed mothers produce less exaggerated prosody and have delayed responses to infant vocalizations and more variable pauses (Bettes, 1988). These differences in speech behaviors have implications for infant learning, as infants of depressed mothers fail to learn that their mother's voice signals an upcoming face, although they are able to successfully learn this association from the CDS of unfamiliar nondepressed female speaker (Kaplan et al., 2002). However, infants of mothers with postpartum depression show less learning from the CDS of a nondepressed female speaker than infants of nondepressed mothers, and this difference continues even after the infants' mothers are no longer clinically depressed, suggesting exposure to early features of CDS may serve a 'social gating' function that supports the infants attention and responsiveness to others in relatively long-lasting ways (Kaplan et al., 2012)

Conclusions

The vast majority of the literature suggests that the adjustments in speech and nonverbal behaviors made by adults when interacting with children serve numerous purposes and provide benefits for young language learners. Not only have global relationships between maternal linguistic input and child language development been found, but research has also begun to elucidate specific features of CDS that support language learning. Furthermore, other aspects of communication beyond speech are also modified in child-directed interactions, including gestures and movements, which appear to benefit

language learning. Finally, evidence from developmental disabilities and maternal depression also support a strong role for CDS in language learning. Specifically, increasing maternal linguistic input with children with developmental disabilities can enhance language development, while the negative effects of maternal depression on CDS impairs language learning in children. In sum, there is increasing evidence that CDS is highly universal and beneficial for language acquisition.

See also: Biological Preconditions for Language Development; Language Development, Theories of; Language Development: Influence of Socio-Economic Status; Neural Basis of Language Development; Neurophysiological Markers of Language Development.

Bibliography

Adamson, L.B., Bakeman, R., 1984. Mothers' communicative acts: changes during infancy. Infant Behavior and Development 7 (4), 467–478.

Akhtar, N., Dunham, F., Dunham, P.J., 1991. Directive interactions and early vocabulary development: the role of joint attentional focus. Journal of Child Language 18, 41–49.

Alston, E., St James-Roberts, I., 2005. Home environments of 10-month-old infants selected by the WILSTAAR screen for pre-language difficulties. International Journal of Language & Communication Disorders 40 (2), 123–136.

Bettes, B.A., 1988. Maternal depression and motherese: temporal and intonational features. Child Development 59 (4), 1089–1096.

Bloom, K., Russell, A., Wassenberg, K., 1987. Turn-taking affects the quality of infant vocalizations. Journal of Child Language 14 (2), 211–227.

Brand, R.J., Baldwin, D.A., Ashburn, L.A., 2002. Evidence for 'motionese': modifications in mothers' infant-directed action. Developmental Science 5 (1), 72–83.

Brand, R.J., Shallcross, W.L., 2008. Infants prefer motionese to adult-directed action. Developmental Science 11 (6), 853–861.

Brent, M.R., Siskind, J.M., 2001. The role of exposure to isolated words in early vocabulary development. Cognition 81 (2), B33–B44.

Bryant, G.A., Barrett, H.C., 2007. Recognizing intentions in infant-directed speech: evidence for universals. Psychological Science 18 (8), 746–751.

Buium, N., Rynders, J., Turnure, J., 1974. Early maternal linguistic environment of normal and downs syndrome language-learning children. American Journal of Mental Deficiency 79 (1), 52–58.

Cardosomartins, C., Mervis, C.B., Mervis, C.A., 1985. Early vocabulary acquisition by children with down syndrome. American Journal of Mental Deficiency 90 (2), 177–184.

Chong, S.C.F., Werker, J.F., Russell, J.A., Carroll, J.M., 2003. Three facial expressions mothers direct to their infants. Infant and Child Development 12 (3), 211–232.

Cooper, R.P., Aslin, R.N., 1990. Preference for infant-directed speech in the first month after birth. Child Development 61 (5), 1584–1595.

D'Odorico, L., Jacob, V., 2006. Prosodic and lexical aspects of maternal linguistic input to late-talking toddlers. International Journal of Language & Communication Disorders 41 (3), 293–311.

DePaulo, B.M., Bonvillian, J.D., 1978. Effect on language development of special characteristics of speech addressed to children. Journal of Psycholinguistic Research 7 (3), 189–211.

Fernald, A., 1989. Intonation and communicative intent in mothers' speech to infants: is the melody the message? Child Development, 1497–1510.

Fernald, A., Mazzie, C., 1991. Prosody and focus in speech to infants and adults. Developmental Psychology 27 (2), 209–221.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., Fukui, I., 1989. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. Journal of Child Language 16 (3), 477–501.

Fidler, D.J., 2003. Parental vocalizations and perceived immaturity in down syndrome. American Journal on Mental Retardation 108 (6), 425–434.

Furrow, D., Nelson, K., Benedict, H., 1979. Mothers' speech to children and syntactic development: some simple relationships. Journal of Child Language 6 (3), 1–21.

Gogate, L.J., Bahrick, L.E., Watson, J.D., 2000. A study of multimodal motherese: the role of temporal synchrony between verbal labels and gestures. Child Development 71 (4), 878–894.

Gogate, L.J., Bolzani, L.H., Betancourt, E.A., 2006. Attention to maternal multimodal naming by 6- to 8-month-old infants and learning of word-object relations. Infancy 9 (3), 259–288.

Goldstein, M.H., King, A.P., West, M.J., 2003. Social interaction shapes babbling: testing parallels between birdsong and speech. Proceedings of the National Academy of Sciences of the United States of America 100 (13), 8030–8035.

Goldstein, M.H., Schwade, J.A., 2008. Social feedback to infants' babbling facilitates rapid phonological learning. Psychological Science 19 (5), 515–523.

Goodman, J.C., Dale, P.S., Li, P., 2008. Does frequency count? Parental input and the acquisition of vocabulary. Journal of Child Language 35 (3), 515–531.

Green, J.R., Nip, I.S.B., Wilson, E.M., Mefferd, A.S., Yunusova, Y., 2010. Lip movement exaggerations during infant-directed speech. Journal of Speech Language and Hearing Research 53 (6), 1529–1542.

Grieser, D.L., Kuhl, P.K., 1988. Maternal speech to infants in a tonal language – support for universal prosodic features in motherese. Developmental Psychology 24 (1), 14–20.

Haebig, E., McDuffie, A., Weismer, S.E., 2013a. The contribution of two categories of parent verbal responsiveness to later language for toddlers and preschoolers on the autism spectrum. American Journal of Speech-Language Pathology 22, 57–70.

Haebig, E., McDuffie, A., Weismer, S.E., 2013b. Brief report: parent verbal responsiveness and language development in toddlers on the autism spectrum. Journal of Autism and Developmental Disorders 43 (9), 2218–2227.

Hammer, C.S., Tomblin, J.B., Zhang, X., Weiss, A.L., 2001. Relationship between parenting behaviours and specific language impairment in children. International Journal of Language and Communication Disorders 36 (2), 185–205.

Hampson, J., Nelson, K., 1993. The relation of maternal language to variation in rate and style of language acquisition. Journal of Child Language 20, 313–342.

Hart, B., Risley, R.T., 1995. Meaningful Differences in the Everyday Experience of Young American Children. Paul H. Brookes, Baltimore.

Henning, A., Striano, T., Lieven, E.V.M., 2005. Maternal speech to infants at 1 and 3 months of age. Infant Behavior & Development 28 (4), 519–536.

Hoff, E., Naigles, L., 2002. How children use input to acquire a lexicon. Child Development 73 (2), 418–433.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., Lyons, T., 1991. Early vocabulary growth – relation to language input and gender. Developmental Psychology 27 (2), 236–248.

Iverson, J.M., Capirci, O., Longobardi, E., Caselli, M.C., 1999. Gesturing in mother-child interactions. Cognitive Development 14 (1), 57–75.

Iverson, J.M., Longobardi, E., Spampinato, K., Caselli, M.C., 2006. Gesture and speech in maternal input to children with down's syndrome. International Journal of Language and Communication Disorders 41 (3), 235–251.

Jacobson, J.L., Boersma, D.C., Fields, R.B., Olson, K.L., 1983. Paralinguistic features of adult speech to infants and small children. Child Development 54, 436–442.

Kaplan, P.S., Bachorowski, J.-A., Smoski, M.J., Hudenko, W.J., 2002. Infants of depressed mothers, although competent learners, fail to learn in response to their own mothers' infant-directed speech. Psychological Science 13 (3), 268–271.

Kaplan, P.S., Danko, C.M., Kalinka, C.J., Cejka, A.M., 2012. A developmental decline in the learning-promoting effects of infant-directed speech for infants of mothers with chronically elevated symptoms of depression. Infant Behavior and Development 35, 369–379.

Kempe, V., Brooks, P.J., Gillis, S., Samson, G., 2007. Diminutives facilitate word segmentation in natural speech: cross-linguistic evidence. Memory & Cognition 35 (4), 762–773.

Klin, A., 1991. Young autistic children's listening preferences in regard to speech: a possible characterization of the symptom of social withdrawal. Journal of Autism and Developmental Disorders 21 (1), 29–42.

Konstantareas, M.M., Zajdeman, H., Homatidis, S., Mccabe, A., 1988. Maternal speech to verbal and higher functioning versus nonverbal and lower functioning autistic children. Journal of Autism and Developmental Disorders 18 (4), 647–656.

Koterba, E.A., Iverson, J.M., 2009. Investigating motionese: the effect of infant-directed action on infants attention and object exploration. Infant Behavior and Development, 32, 437–444.

Kuhl, P.K., Andruski, J.E., Chistovich, I.A., Chistovich, L.A., Kozhevnikova, E.V., Ryskina, V.L., Lacerda, F., 1997. Cross-language analysis of phonetic units in language addressed to infants. Science 277 (5326), 684–686.

Kuhl, P.K., Coffey-Corina, S., Padden, D., Dawson, G., 2004. Links between social and linguistic processing of speech in preschool children with autism: behavioral and electrophysiological measures. Developmental Science 8 (1), F1–F12.

Kuhl, P.K., Tsao, F.M., Liu, H.M., 2003. Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. Proceedings of the National Academy of Sciences of the United States of America 100 (15), 9096–9101.

Lew-Williams, C., Pelucchi, B., Saffran, J.R., 2011. Isolated words enhance statistical learning in infancy. Developmental Science 14, 1323–1329.

Ma, W., Golinkoff, R.M., Houston, D.M., Hirsh-Pasek, K., 2011. Word learning in infant- and adult-directed speech. Language Learning and Development 7 (3), 185–201.

Marfo, K., 1990. Maternal directiveness in interactions with mentally-handicapped children – an analytical commentary. Journal of Child Psychology and Psychiatry and Allied Disciplines 31 (4), 531–549.

Masataka, N., 1992. Motherese in a signed language. Infant Behavior & Development 15 (4), 453–460.

Masataka, N., 1996. Perception of motherese in a signed language by 6-month-old deaf infants. Developmental Psychology 32 (5), 874–879.

Masur, E.F., Flynn, V., Eichorst, D.L., 2005. Maternal responsive and directive behaviors and utterances as predictors of children's lexical development. Journal of Child Language 32, 63–91.

Murray, L., Trevarthen, C., 1986. The infants' role in mother-infant communications. Journal of Child Language 13 (1), 15–29.

Naoi, N., Minagawa-Kawai, Y., Kobayashi, A., Takeuchi, K., Nakamura, K., Yamamoto, J., 2012. Cerebral responses to infant-directed speech and the effect of talker familiarity. NeuroImage 59 (2), 1735–1744.

Nelson, D.G.K., Hirsh-Pasek, K., Jusczyk, P.W., Cassidy, K.W., 1989. How the prosodic cues in motherese might assist language learning. Journal of Child Language 16 (1), 55.

Newport, E., Gleitman, H., Gleitman, L., 1977. Mother, I'd rather do it myself: some effects and noneffects of maternal speech style. In: Snow, C.E., Ferguson, C.A. (Eds.), Talking to Children. Cambridge University Press, Cambridge, England, pp. 109–149.

O'Neill, M., Bard, K.A., Linnell, M., Fluck, M., 2005. Maternal gestures with 20-month-old infants in two contexts. Developmental Science 8 (4), 352–359.

Papoušek, M., Papoušek, H., Haekel, M., 1987. Didactic adjustments in fathers' and mothers' speech to their 3-month-old infants. Journal of Psycholinguistic Research 16 (5), 491–516.

Paul, R., Chawarska, K., Fowler, C., Cicchetti, D., Volkmar, F., 2007. "Listen my children and you shall hear": auditory preferences in toddlers with autism spectrum disorders. Journal of Speech, Language and Hearing Research 50 (5), 1350–1364.

Pegg, J.E., Werker, J.F., Mcleod, P.J., 1992. Preference for infant-directed over adult-directed speech – evidence from 7-week-old infants. Infant Behavior & Development 15 (3), 325–345.

Pine, J.M., 1994. The language of primary caregivers. In: Gallaway, C., Richards, B.J. (Eds.), Input and Interaction in Language Acquisition. Cambridge University Press, New York, pp. 15–37.

Pye, C., 1986. Quiche Mayan speech to children. Journal of Child Language 13 (1), 85–100.

Ramey, C.T., Klisz, D., Hieger, L., 1972. Synchronous reinforcement of vocal responses in failure-to-thrive infants. Child Development 43 (4), 1449.

Roberts, M.Y., Kaiser, A.P., 2011. The effectiveness of parent-implemented language interventions: a meta-analysis. American Journal of Speech-language Pathology 20 (3), 180.

Rowe, M., 2012. A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. Child Development 83 (5), 1762–1774.

Rutherford, M.D., Przednowek, M., 2012. Fathers show modifications of infant-directed action similar to that of mothers. Journal of Experimental Child Psychology 111 (3), 367–378.

Sachs, J., 1977. The adaptive significance of linguistic input to prelinguistic infants. In: Snow, C.E., Ferguson, C.A. (Eds.), Talking to Children. Cambridge University Press, Cambridge, England.

Saito, Y., Aoyama, S., Kondo, T., Fukumoto, R., Konishi, N., Nakamura, K., Toshima, T., 2007. Frontal cerebral blood flow change associated with infant-directed speech. Archives of Disease in Childhood-Fetal and Neonatal Edition 92 (2), F113–F116.

Schachner, A., Hannon, E.E., 2011. Infant-directed speech drives social preferences in 5-month-old infants. Developmental Psychology 47 (1), 19–25.

Shimpi, P.M., Fedewa, A., Hans, S., 2012. Social and linguistic input in low-income African American mother-child dyads from 1 months through 2 years: relations to vocabulary development. Applied Psycholinguistics 33, 781–798.

Siller, M., Sigman, M., 2002. The behaviors of parents of children with autism predict the subsequent development of their children's communication. Journal of Autism and Developmental Disorders 32, 77–89.

Singh, L., 2008. Influences of high and low variability on infant word recognition. Cognition 106 (2), 833–870.

Singh, L., Nestor, S., Parikh, C., Yull, A., 2009. Influences of infant-directed speech on early word recognition. Infancy 14 (6), 654–666.

Smith, N.A., Trainor, L.J., 2008. Infant-directed speech is modulated by infant feedback. Infancy 13 (4), 410–420.

Snow, C.E., 1977. Mothers' speech research: from input to interaction. In: Snow, C.E., Ferguson, C.A. (Eds.), Talking to Children. Cambridge University Press, Cambridge, England, pp. 31–49.

Stein, A., Gath, D.H., Bucher, J., Bond, A., Day, A., Cooper, P.J., 1991. The relationship between post-natal depression and mother-child interaction. British Journal of Psychiatry 158, 46–52.

Stern, D.N., 1974. Mother and infant at play: the dyadic interaction involving facial, vocal and gaze behaviors. In: Lewis, M., Rosenblum, L.A. (Eds.), The Effect of the Infant on its Caregiver. Wiley, New York, pp. 187–213.

Stern, D.N., Spieker, S., Barnett, R., MacKain, K., 1983. The prosody of maternal speech: infant age and context related changes. Journal of Child Language 10 (1), 1–15.

Tamis-LeMonda, C.S., Bornstein, M.H., Baumwell, L., 2001. Maternal responsiveness and children's achievement of language milestones. Child Development 72 (3), 748–767.

Thiessen, E.D., Hill, E.A., Saffran, J.R., 2005. Infant-directed speech facilitates word segmentation. Infancy 7 (1), 53–71.

Tomasello, M., Farrar, M.J., 1986. Joint attention and early language. Child Development 57 (6), 1454–1463.

Tomasello, M., Todd, J., 1983. Joint attention and lexical acquisition style. First Language 4, 197–212.

Trainor, L.J., Austin, C.M., Desjardins, R.N., 2000. Is infant-directed speech prosody a result of the vocal expression of emotion? Psychological Science 11 (3), 188–195.

Trainor, L.J., Desjardins, R.N., 2002. Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels. Psychonomic Bulletin & Review 9 (2), 335–340.

Vallabha, G.K., McClelland, J.L., Pons, F., Werker, J.F., Amano, S., 2007. Unsupervised learning of vowel categories from infant-directed speech. Proceedings of the National Academy of Sciences of the United States of America 104 (33), 13273–13278.

Vallotton, C.D., 2009. Do infants influence their quality of care? Infants' communicative gestures predict caregivers' responsiveness. Infant Behavior & Development 32 (4), 351–365.

Wan, M.W., Green, J., Elsabbagh, M., Johnson, M., Charman, T., Plummer, F., 2013. The BASIS Team. Quality of interaction between at-risk infants and caregiver at 12–15 months is associated with 3-year autism outcomes. Journal of Child Psychology and Psychiatry 54 (7), 763–771.

Watson, L.R., Baranek, G.T., Roberts, J.E., David, F.J., Perryman, T.Y., 2010. Behavioral and physiological responses to child-directed speech as predictors of communication outcomes in children with autism spectrum disorders. Journal of Speech, Language and Hearing Research 53 (4), 1052.

Watson, L.R., Roberts, J.E., Baranek, G.T., Mandulak, K.C., Dalton, J.C., 2012. Behavioral and physiological responses to child-directed speech of children with autism spectrum disorders or typical development. Journal of Autism and Developmental Disorders 42, 1616–1629.

Weizman, Z.O., Snow, C.E., 2001. Lexical input as related to children's vocabulary acquisition: effects of sophisticated exposure and support for meaning. Developmental Psychology 37 (2), 265–279.

Werker, J.F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., Amano, S., 2007. Infant-directed speech supports phonetic category learning in English and Japanese. Cognition 103 (1), 147–162.

Wulbert, M., Inglis, S., KriegsmaE, Mills, B., 1975. Language delay and associated mother-child interactions. Developmental Psychology 11 (1), 61–70.

Zangl, R., Mills, D.L., 2007. Increased brain activity to infant-directed speech in 6-and 13-month-old infants. Infancy 11 (1), 31–62.

Child Labor

Jackline Wahba, University of Southampton, Southampton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Child labor is still a cause of public concern. This article focuses on the relationship between poverty and child labor. Many children work to ensure the survival of their families and themselves. Furthermore, child labor today leads to poverty tomorrow: child labor perpetuates poverty and results in intergenerational poverty traps. Yet, policies banning child labor, even if they can be implemented, might harm the welfare of children and their families rather than help. This is why devising effective policies to address child labor is a challenging task and needs to tackle the root cause of the problem.

Child labor is still a cause of public concern in the twenty-first century. Recent estimates by the International Labor Office (ILO) indicate that there were about 168 million child laborers, aged 5–17 years, worldwide in 2012. Almost half of these children, around 85 million, work long hours under hazardous conditions. About 4 in 10 child laborers were younger than 12 years old (73 million). However, not all children work is seen as problematic, only when children employment interferes with the children's education and welfare that it is referred to as child labor. So, if child labor is a concern, what could be done about it?

Although this might seem to be a straightforward question, answering it is not so straightforward. In order to address the problem of child labor, it is necessary to understand the circumstances that lead children to work. Many economists believe that poverty is the main reason behind child labor. Many children work to ensure the survival of their families and themselves. However, child labor today, leads to poverty tomorrow. Child labor perpetuates poverty and results in intergenerational poverty traps. Yet, policies banning child labor, even if they can be implemented, might harm the welfare of children and their families rather than help. This is why devising effective policies to address child labor is a challenging task and cannot be done in isolation but has to tackle the root cause of the problem.

This article will examine the intricate relationship between poverty and child labor, and assess alternative policies to address child labor when children are driven to work by poverty. Section What Is Child Labor? will provide an overview of the linkage between poverty and child labor. Section Poverty Trap will examine the impact of child labor on poverty persistence, while Section What Are the Main Policy Implications? will evaluate the effectiveness of various policy recommendations to address child labor. Section Conclusion concludes.

What Is Child Labor?

Not all work done by children is classified as child labor. According to the ILO's Statistical Information and Monitoring Program on Child Labor (SIMPOC) 'child labor' is work that deprives children of their childhood, their potential and their dignity, and that is harmful to physical and mental development. Labor that jeopardizes the physical, mental or moral well-being of a child, either because of its nature or because of the conditions in which it is carried out, is known as 'hazardous work.' The ILO stipulates a minimum age for employment of 15 and 18 years for hazardous occupations and defines child labor by the ILO Minimum Age Convention, 1973 (No. 138) and the ILO Worst Forms of Child Labor Convention, 1999 (No. 182).

Estimating the number of child laborers is a difficult task. The ILO produced new child labor estimates for the year 2012 based on 60 datasets from national household surveys. The new global and regional estimates on child labor benefited from refinements embodied in the Resolution concerning Statistics of Child Labor, adopted by the 18th International Conference of Labor Statisticians (ICLS) in 2008. Based on those estimates for 5–17 years old in 2012, the incidence of child labor was 10.6% worldwide, and has fallen slightly from 14.2% in 2004. As **Table 1** shows, based on ILO (2013), that the largest number of child laborers was in Asia and the Pacific (77.7 million), but in relative terms Sub-Saharan Africa had by far the highest incidence rate around 21.4 versus 9.3% in Asia and the Pacific. One in 10 children was child laborer in Latin America and the Caribbean. The ILO estimates show that more boys than girls were engaged in child labor, in absolute and in relative terms.

The relative level of the child labor by economic activity can vary widely from one country to another. The ILO estimates suggest that child laborers were engaged in all the three broad groupings of economic activity (agriculture, industry, and services). However more than half (60%) were engaged in agriculture, 7% were employed in industry and 26% in services.

Table 1 Regional estimates of child labor in 2012, 5–17 years old

Region	*Total children (thousand)*	*Child labor (thousand)*	*Incidence rate (%)*
World	1585566	167956	10.6
Asia and the Pacific	835334	77723	9.3
Latin America and the Caribbean	142693	12505	8.8
Sub-Saharan Africa	275397	59031	21.4
Other regions	332143	18697	5.6

Source: ILO, 2013. Global Child Labor Trends 2008 to 2012, International Programme on the Elimination of Child Labor (IPEC). ILO, Geneva, p.10.

Two thirds (64 for boys vs 73% for girls) of child laborers were unpaid family workers making any form of intervention very difficult to implement. Paid employment and self-employment accounted respectively for 21 and 5% of all child laborers: only one in five child laborers were in paid employment. The overwhelming majority are unpaid family workers. The vast majority of child labor is found in rural settings and informal urban economy. Participation rates of children in economic activity are on average twice as high in rural communities than in urban centers. See ILO (2013).

Why Is Child Labor a Cause for Public Concern?

Child labor is a cause for concern, not only because of the number of children affected, which is still very high as seen above, but also and more importantly, because of the negative repercussions that starting working life too young has on the personal development of children. There is no disagreement that any type of child labor that involves abusive commercial exploitation of children or hazardous work is harmful and should be eliminated. Yet, when it comes to child labor in nonhazardous work the main concern is that child labor prevents children from receiving education. Low human capital base is the most serious developmental constraint in developing countries. Investments in child schooling add to the stock of human capital, and therefore to future income and living standards for society as a whole. Child labor constrains the children's abilities to benefit from education and schooling, thereby condemning them to perpetual poverty and low-wage employment.

Poverty and Child Labor

One of the most influential studies on child labor is the article by the Basu and Van (1998) that has spurred a large number of articles. The seminal theoretical article by Basu and Van (1998) argues that household decisions to send their children into employment are driven by poverty. Their model builds on two fundamental assumptions: first that households send children to work only if the adult wage is very low that the household subsistence requirements cannot be met without the income generated by the children- which they referred to as the luxury axiom. In addition, they assumed that adult labor and child labor are substitutes in production- known as the substitution axiom. They concluded that if a labor market is characterized by more than one equilibrium- one in which adult wages are low and children work, and another where adult wages are high and children do not work- banning child labor is a benign policy. But, in very poor countries, where there is only one labor market equilibrium at which adult wages are low, banning child labor can worsen the households' welfare.

This article has prompted a substantial body of literature and ignited the interest of many economists in child labor over the last decade aided by availability of large-scale, nationally representative household surveys from developing countries that has enabled more rigorous studies of this complex issue. Many believe that poverty is the main cause of child labor. However, the empirical evidence on the causal impact of poverty on child labor fails to consistently find a negative relationship between household income and child labor. For example, Ray (2000) fails to find positive relationship between child labor and poverty in Pakistan. As discussed by Edmonds (2008), there are two fundamental problems in studying the impact of poverty on child labor which are responsible for the lack of consistent relationship. First, poor households differ from rich households and those differences in many cases are due to reasons associated with child labor. Hence, disentangling these omitted factors from the underlying causal relationship is empirically difficult. Secondly, the relationship between household income and child labor is a two-way one where household income affects child labor but child labor affects household income as well. Thus the second challenge is to control for the endogeneity between household income and child labor.

Studies that use panel data and track children in the same household over time consistently find that large increases in child labor are associated with substantive declines in household incomes. For example, Beegle et al. (2006) track children over a 3-year period in rural Tanzania and find that children tend to work when households experience unexpectedly poor harvest, and that children stop working when households recover from the bad harvest. Also, Duryea et al. (2007) find that child labor increases during adult unemployment spells in urban Brazil. Edmonds (2005) using a panel dataset of Vietnamese households from 1993 to 1997 finds that improvements in per capita expenditure explain 80% of the decline in child labor.

Some studies have taken a different approach by focusing on market wages as opposed to parents' wages or income. As discussed above, Basu and Van (1998) stress the role of adult market wage; they assume that households send children to work only if the adult market wage is very low and as the wage rise they withdraw their children from the labor force. Thus, the elasticity of child labor supply to adult market wages is important in particular for policy formulation. Wahba (2006) finds support for this hypothesis, examining the relationship between adult market wages and child labor in Egypt and finds that a 10% increase in the illiterate male market wage lowers the probability that a child engages in wage work by 22 for boys and 13% for girls.

Another contentious issue in the empirical evidence on the relationship between household income and child labor is what Bhalotra and Heady (2003) refer to as the wealth paradox where they find a positive correlation between market work participation and household assets. Since land is usually strongly correlated with a household's income, this seems to challenge the presumption that child labor involves the poorest households. Basu et al. (2010) argue that the missing ingredient may be an explicit modeling of the labor market. They develop a simple model which suggests the possibility of an inverted-U relationship between land holdings and child labor. Using a unique data set from India, that has child labor-hours they find that, controlling for child, household and village characteristics, the turning point beyond which more land leads to a decline in child labor occurs around 4 acre of land per household. Thus taking into account the nonlinear relationship between household's income and child labor explains this wealth paradox.

Households also may send children to work not only in order to augment household income but also to manage better the income risk they face. Child labor can be part of a strategy to minimize the risk of interruption of a household's income stream and hence to reduce the potential impact of job loss by a family member or of failed harvest for example. More importantly, because of capital market imperfection in developing countries, poor households cannot borrow to manage during the bad times and in many cases they have to rely on their children's earnings for survival, see e.g., Baland and Robinson (2000). For example, Dehejia and Gatti (2005) find a negative relationship between child labor and access to credit, where access to credit is proxied by the extent of financial development.

Meanwhile at the macrolevel, there is a debate on whether international trade and globalization reduces or increases child labor. A common argument in the child labor literature is that foreign trade, and globalization in general, increases child labor by increasing the demand for goods produced by children. See Edmonds (2008) for an extensive summary of this literature. For instance, Edmonds and Pavcnik (2007) study the relationship between trade liberalization (measured as the change in the relative price of an exported commodity) and child labor using household level data from Vietnam. They find that higher rice prices are associated with declines in child labor, suggesting that in this case greater market integration is associated with less child labor, not more. Fafchamps and Wahba (2006) show that children living in or around cities work more for the market in the form of wage work or small business employment. The increased intensity of market exchange associated with cities thus appears to be empirically correlated with an increased involvement of children in market-related forms of work. At the same time, work by children on the family farm and on household chores falls dramatically with proximity to urban centers. These results suggest that urban proximity in Nepal is associated with an improvement in child welfare – at least as far as total child labor and education are concerned.

There are potentially many other determinants for child labor besides poverty. One strand of the child labor literature focuses on the lack of availability of good schools. The overall condition of the education system can be a powerful influence on the supply of labor. The failure of the education system may lead many parents to view work as the preferred option for their children because education is no longer a road toward obtaining a degree and a civil service job. In addition, traditional factors such as rigid cultural and social roles further limit educational attainment and increase child labor. Another strand of the literature pushes sociodemographics, family characteristics and social norms as important determinants of child labor. Parents' education play an important role: educated parents are less likely to send their children to work. The child's age, gender, and birth order also affect child labor. Furthermore, cultural factors matter, so there is a huge variation among patterns and incidence of child labor in different countries that stem from cultural and social norms and traditions.

Are Poor Households Acting Rationally?

One might ask, are poor households acting rationally by sending their children out to work? Well, to be able to answer that question, let's consider a very simple model of household time allocation. Let's assume that a child can allocate his/her time between three activities: schooling, work, and leisure. The family allocates the time of its children to wherever the perceived private return is highest. The decision by any family to send his or her child to school or to work depends on the private benefits and costs. First, let's consider the private returns and costs of schooling. The benefits obtained from education clearly depend on the quality of the education the child is likely to receive and how that education is likely to increase the child's productivity in the future and therefore his/her future earnings. The overall condition of the education system can be a powerful influence on the supply of labor. Low quality of the education, for example too crowded classes or bad teaching, would reduce the benefits of schooling. As a result, parents may find no use in sending their children to school when they could be home learning a skill (for example, agriculture) and supplementing the family income. Schooling is costly even if it is free (no fees) because there are other direct costs involved, such as books, uniform, and travel. Also, there are indirect costs of schooling; the opportunity cost of child schooling (the loss incurred by a family whose child is in school rather than working) is wage rate earned in the labor market by that child. The higher the opportunity cost of school attendance in relation to a household's income, the greater the perceived need for the child to work. Schooling problems, such as schools being too far away, would also increase the cost of education and make it unattractive for parents to invest in children schooling.

Families which send their children to school give up short term consumption in favor of investment in human capital of the child, which is likely to pay off in old age. Families which choose to engage their children in paid market work, sacrifice the future of their children, and hence that of their own, in an attempt to raise their current standard of living of the household. Families' decisions to take a child out of school and send him or her to work is typically a rational decision based on a comparison of the discounted future stream of benefits accruing to the child, and/or household from the child's education relative to the opportunity cost of the child's time which is his/her productivity at work, plus the direct costs of education. Poor households are likely to be present-oriented as compared with more well-off households and would be likely to discount the future more than well-off households and are therefore more likely to take their children out of school.

Poverty Trap

The main concern about child labor is its detrimental effects on schooling. Few studies have examined the dynamics of child labor. For example, Basu (1999) show that child labor, when is undertaken at the expense of human capital formation, would lead to child labor trap: if children work, they will not go to school and will have low human capital. As adults, they will also need to send their children to work thereby perpetuating child labor in the future. Households that are very poor are much more likely to send their children to work, and child labor contributes to poverty in the next generation by reducing schooling attainment. For example, Ray (2003) finds that an additional hour of wage work in Ghana is associated with more

than a year's less completed educational attainment. This circular pattern of positive feedback between poverty and child labor may lead to a vicious cycle of poverty, in which the descendants of the poor remain poor because they were poorly educated. This cycle can be the foundation of a classical 'poverty trap.'

There is little empirical evidence on the intergenerational transmission of poverty in developing countries, in particular, on how child labor perpetuates poverty from one generation to another, or on how parents who were child laborers are more likely to have their children work as well. Parents who worked as children are more likely to have underinvested in schooling and become poverty-trapped and hence would expect their children to work as well. The relationship between family background, such as race, ethnic origin, religion and in particular education, and child labor is fairly established in the empirical literature. Studies show that a low level of parents' educational attainment is an important factor in increasing the likelihood of children working.

A few studies have examined the effects of child labor on the children once they become adults. Ilhai et al. (2000) examine the effects of child labor on lifetime earnings and poverty status using retrospective data on adult earnings and information on whether adults worked as children in Brazil. They find that entry in the workforce before age 13 results in a reduction in adult wages of 13–17% and an increase in the probability of being in the lowest two income quintiles of 7–8%. They estimate the effect of child labor both independent of its effect on education and through its effects on education. They find both effects to be significant. They argue that early entry in the labor force reduces years of education and lowers the returns per year of schooling. Similarly, Emerson and Souza (2003) explore the evidence of intergenerational persistence of child labor. They build an overlapping generation model of the household child labor decision to capture a child labor-trap story and then examine the empirical evidence using household survey data from Brazil. They find that adults who worked as children have lower earnings even when controlling for their educational levels and that the likelihood of being child laborer increases the younger the parents entered the labor market. They explain the persistence of child labor above that of parental education and household income, given their theoretical model, to be the result of some unobservable human capital characteristic i.e., human capital accumulation is not determined only by the amount of education but also by maybe the quality of education, the level of education of siblings or the household environment. Wahba (2006) also provides direct evidence on the intergenerational persistence of child labor where she finds that parents who themselves worked as child laborers are more likely to send their children to work in Egypt.

Thus child labor if driven by severe poverty is highly likely to reinforce itself and result in an awful equilibrium of persistent poverty, excessive child labor and low education over generations.

What Are the Main Policy Implications?

Tackling child labor is not straightforward. Various policy options such as banning child labor or trade sanctions, even if they can be implemented, are unlikely to be effective in eradicating child labor. Indeed such policies may even be counterproductive because they would harm the same people they want to protect namely, the poor. Basu and Zarghamee (2009) argue that boycott and trade sanctions may increase child labor. Their argument is that sanctions tend to depress child wages, and that the response of poor families depending on child labor to avert extreme poverty may be characterized by strong income deficits. Doepke and Zilibotti (2009) use a political-economy perspective to show that these types of international action on child-labor tend to lower domestic political support within developing countries for banning child-labor. Hence, international labor standards and product boycotts may delay the ultimate eradication of child-labor.

An alternative policy option that is often advocated to address child labor is drawing children into school. To the extent that schooling displaces child labor, policies that affect the costs of or the returns to school might prove effective in reducing child labor. However, Ravallion and Wodon's (2000) point that increased school enrollment does not necessarily lead to an equal fall in child labor as they find that Bangladesh's Food for Education program increased school attendance, but reduced the incidence of child labor by only a small proportion. Kondylis and Manacorda (2012) go one step further, by arguing that increases in enrollment induced by improved school accessibility might, as an unwanted consequence, lead to a rise in child labor. Similarly de Hoop and Rosati (2012) investigate the impact of school subsidies and increased access to education on child work using data from 'Burkinabé Response to Improve Girls' Chances to Succeed' (BRIGHT), an integrated program that aims to improve school participation in rural communities in Burkina Faso. They find that BRIGHT substantially improved school participation, yet it did not reduce – in fact may have increased – children's participation in economic activities and household chores. They argue that if education programs are implemented to achieve a combination of increased school participation and a reduction in child work they may either have to be combined with different interventions that effectively reduce child work or they may have to be tuned more carefully to the incentives and constraints the child laborer faces.

A number of countries have adopted policies designed to discourage child labor and increase schooling through conditioning cash transfers to poor households on children attending school. The PROGRESA program is particularly important as one of the first conditional cash transfers that incentivized families to send their children to school. The cash transfer increases with the age of the child in order to compensate the household for the older child's greater opportunity cost of schooling. In addition, at secondary school ages, girls receive larger cash payments for attending school than do boys. Several studies find the program to have had a significant reduction in child labor e.g., Schultz (2004) and Skoufias and Parker (2001).

A recent study by Edmonds and Schady (2012) considers how child time allocation in Ecuador responds to receipt of the *Bono de Desarrollo Humano* (BDH) cash transfer. The BDH cash transfer is randomly assigned to some eligible poor households and not to others. The BDH transfer is $15 per household per month, 7% of monthly expenditures for recipient households,

and does not vary across eligible families. The transfer is paid to mothers, and does not have any conditions attached. They find that $15 a month given as the BDH reduces child labor. These results are consistent with a growing literature that documents that cash and in-kind transfers can reduce child work. Their findings highlight the central role that poverty plays in the child labor decision, and suggest the possibility of affecting large changes in child labor with relatively modest investments in poverty relief.

Conclusion

Parents do not send their children into paid employment out of a selfish preference. In the vast majority of cases, children are driven into work by poverty and the inability of poor households to manage external shocks. Households that are very poor are much more likely to send their children to work, resulting in the poverty of their children when they become adults. Hence child labor leads to a vicious cycle of poverty. Yet, any measures aiming to ban child labor might hurt poor children by reducing their welfare.

As a result addressing child labor needs to be mindful of its main driver and deal with the root of the problem: poverty. Cash Transfer programs have proved success in mitigating the poverty effects that drive children out of school and into labor markets. For example, in Brazil, the 'Bolsa Familia programme,' which provides cash transfers to some 820 000 children aged 7–14, is estimated to have increased school attendance by 4% nationally, but 9 and 12% respectively in rural regions and the north-east (the poorest part of the country). There is now a growing body of evidence that both conditional and unconditional cash transfers, in the range 0.5 to 2% of GDP, can play a role in increasing child schooling and reducing child labor.

Any intervention to control child labor should recognize the powerful market forces that give rise to child labor in the first place as there are many pitfalls and risks of backlash that occur in this area of complex interaction between household economics and market structures. As long as the main cause that pushes children to work, namely poverty, is not tackled, child labor will never be a phenomenon of the past. However, the recent evidence on the success of cash transfers in reducing child labor and poverty is promising.

See also: Children and Families in Social Work; Children, Rights of: Cultural Concerns; Globalization and Human Trafficking; Street Children: Cultural Concerns.

Bibliography

Baland, J.-M., Robinson, J., 2000. Is child labor inefficient? Journal of Political Economy 108, 663–679.

Basu, K., Van, P.H., 1998. The economics of child labor. The American Economic Review 88, 412–427.

Basu, K., 1999. Child labor: cause, consequence, and cure, with remarks on international standards. Journal of Economic Literature 37, 1083–1119.

Basu, K., Zarghamee, H., 2009. Is product boycott a good idea for controlling child labor? A theoretical investigation. Journal of Development Economics 88, 217–220.

Basu, K., Das, S., Dutta, B., 2010. Child labor and household wealth: theory and empirical evidence of an inverted-U. Journal of Development Economics 91, 8–14.

Bhalotra, S., Heady, C., 2003. Child farm labor: the wealth paradox. World Bank Economic Review 17, 197–227.

Beegle, K., Dehejia, R., Gatti, R., 2006. Child labor and agricultural shocks. Journal of Development Economics 81, 80–96.

Dehejia, R., Gatti, R., 2005. Child labor the role of financial development and income variability across countries. Economic Development and Cultural Change 53, 913–931.

Doepke, M., Zilibotti, F., 2009. International labor standards and the political economy of child-labor Regulation. Journal of the European Economic Association 7, 508–518.

Duryea, S., Lam, D., Levison, D., 2007. Effects of economic shocks on children's employment and schooling in Brazil. Journal of Development Economics 84, 188–214.

Edmonds, E.V., 2008. Child labor. In: Strauss, J., Schultz, T.P. (Eds.), Handbook of Development Economics, vol. 4, pp. 3607–3709.

Edmonds, E., 2005. Does child labor decline with improving economic status? The Journal of Human Resources 40, 77–99.

Edmonds, E., Pavcnik, N., 2007. International trade and child labor: cross-country evidence. Journal of International Economics 68, 115–140.

Edmonds, E., Schady, N., 2012. Poverty alleviation and child labor. American Economic Journal: Economic Policy 4, 100–124.

Emerson, P.M., Souza, A.P., 2003. Is there a child labor trap? Intergenerational persistence of child labor in Brazil. Economic Development and Cultural Change 51, 375–398.

Fafchamps, M., Wahba, J., 2006. Child labor, urban proximity, and household composition. Journal of Development Economics 79, 374–397.

de Hoop, J., Rosati, F.C., 2012. Does promoting School Attendance Reduce Child Labor? Evidence from Burkina Faso's Bright Project, Understanding Children's Work (UCW) Programme. Working Paper.

Ilhai, N., Orazem, P., Sedlacek, G., 2000. The Implications of Child Labor for Adult Wages, Income and Poverty: Retrospective Evidence from Brazil. The World Bank, Washington, DC.

ILO, 2013. Global Child Labor Trends 2008 to 2012, International Programme on the Elimination of Child Labor (IPEC). ILO, Geneva.

Kondylis, F., Manacorda, M., 2012. School proximity and child labor evidence from rural Tanzania. Journal of Human Resources 47, 32–63.

Ravallion, M., Wodon, Q., 2000. Does child labor displace schooling? Evidence on behavioral responses to an enrolment subsidy? The Economic Journal 110, c158–c175.

Ray, R., 2000. Analysis of child labor in Peru and Pakistan: a comparative study. Journal of Population Economics 13, 3–19.

Ray, R., 2003. The determinants of child labor and child schooling in Ghana. Journal of African Economics 11, 561–590.

Schultz, T., 2004. School subsidies for the poor: evaluating the Mexican PROGRESA poverty program. Journal of Development Economics 74, 199–250.

Skoufias, E., Parker, S., 2001. Conditional cash transfers and their impact on child work and schooling: evidence from the PROGRESA program in Mexico. Economica 2, 45–96.

Wahba, J., 2006. The influence of market wages and parental history on child labor and schooling in Egypt. Journal of Population Economics 19, 823–852.

Child Neglect

Beth Casey, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In recent years there have been important academic and theoretical developments around child neglect, including the nature of the phenomenon, causes and consequences, and the professional approaches most appropriate in neglect cases. This article begins with an analysis of how awareness of neglect has grown, before examining the complexity of definitional issues. It goes on to explore the ways in which knowledge in relation to neglect has been constructed, examining the major constructions and contextual issues that have emerged. Key contextual factors including the relationship between neglect and poverty and the gendered nature of neglect and policy responses are discussed.

Introduction

Child neglect has become an increasingly important issue in recent years with academic and theoretical developments, covering the nature of the phenomenon, causes and consequences, and the professional responses in neglect cases. This article begins with a discussion of the 'neglect of neglect' and then analyzes how awareness of neglect has grown, before examining the complexity of definitional issues. It then explores how knowledge in relation to neglect has been constructed, examining the major constructions and contextual issues that have emerged. Key contextual factors, including the relationship between neglect and poverty and gendered nature of neglect and policy responses are discussed.

The Neglect and Reemergence of Neglect

The 'neglect of neglect' in the child maltreatment field has been noted by researchers over a significant period (Dubowitz, 2007; Gabarino and Collins, 1999) with Coohey (1995: p. 885) concluding that "the overwhelming focus of research and theory building has been on sexual abuse, physical abuse or child maltreatment." Stone (1998) suggests that neglect is poorly understood by practitioners theoretically and that training for practitioners in this area is lacking. The reasons underlying the failure of some parents to provide adequate care for children remain significantly underresearched (Hildyard and Wolfe, 2007).

Coupled with these gaps in research and theory is a growing awareness since the 1970s about the extent to which class-based assumptions can influence practice, leading, according to Scourfield (2003: p. 109), to "a reluctance to bring families into the child protection system" on grounds of being "poor, shabby and dirty." Stone (1998: p. 88) suggests that the relatively low profile of neglect might be "due to the fact that neglect is a long-term developmental issue rather than a crisis." Further, physical abuse may be more tangible and visible, commanding more attention from professionals than neglect, which is viewed as an act of omission (Garbarino and Collins, 1999).

Several studies have highlighted the 'neglect of neglect' at a practice level. Stone (1998) found that neglect cases would be given low priority and were filtered out of the system at various thresholds until a specific incident of sexual or physical abuse arose in the same case. Similarly, Swift (1995: p. 78) suggests that "because neglect is usually characterized as being of long duration, such a case may easily slip to the bottom of a worker's list of action priorities until a child is clearly put at risk."

The 'neglect of neglect' is increasingly being challenged through recognition of its increased incidence and evidence of its adverse effects on children. Neglect has become the largest category of child protection registration in the UK (Scourfield, 2003; Turney, 2000) and accounts for the highest percentage of maltreatment in the US and Canada (Sedlak et al., 2010; Ruiz-Casares et al., 2012). In the Swedish context, Kaunitz et al.'s study revealed that for children aged 0–12, 62% of referrals concerned neglect (Cocozza and Hort, 2011).

Research evidence strongly supports the view that child neglect can lead to poor outcomes for children in both the short term and long term (Daniel and Taylor, 2006). Neglect can impact upon the child's emotional, physical and cognitive, psychological and behavioral development (Glaser and Prior, 2002). Deleterious effects include anxiety and low self-esteem, problematic behavior, educational underachievement, and adverse impacts upon peer and social relationships (Glaser and Prior, 2002). The long-term impact on children may be influenced by the severity, chronicity, and developmental stage at which the neglect occurs (Hildyard and Wolfe, 2002).

Defining Neglect

Definitions of neglect are varied and contested, making investigation of professional responses to neglect complex. Daniel et al. (2011: p. 13) argue that neglect can be defined both broadly and narrowly, highlighting a distinction between "neglect as a concept denoting the experience of a child whose developmental needs are not being met and 'neglect' as an operational, legislative or policy label." Hence, definitions vary depending on purpose, for example, criminal proceedings, determining eligibility for services, and investigating allegations and research (Daniel et al., 2011). When defining neglect then as a "category for compulsory action," definitions tend to be narrower (Daniel et al., 2011: p. 15). In

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28011-2

the UK, national guidance in the shape of *Working Together* defines neglect as:

> the persistent failure to meet a child's basic physical and/or psychological needs, likely to result in the serious impairment of the child's health or development.
> (HM Government, 2013)

In the UK, legally neglect is embedded within The Children Act's (1989: section 31) definition of significant harm: "the child is suffering, or is likely to suffer significant harm; and that the harm or likelihood of harm is attributable to a lack of adequate parental care or control." Common to most definitions of neglect is the notion of parental omission of care, failure to provide for (Garbarino and Collins, 1999; Swift, 1995) or protect the child (CIS, 2008). Most definitions address the failure to meet a child's basic needs (Garbarino and Collins, 1999). In Australia there is no uniform legal definition of neglect, although eight aspects of neglect are covered in different states and territories (Lawrence and Irving, 2004). In other countries, for example, China, there has been no attempt to develop a legal definition of maltreatment (Liao et al., 2011).

In the UK, applying the *Working Together* definition of neglect in practice becomes problematic, particularly given a lack of guidance on the application of the broad statement and explanation of key terms (Turney, 2000). Legally informed definitions assert a dichotomous categorization of neglect, in which the child is not neglected (Polansky et al., 1981). Although the legal definition of neglect appears 'concrete' (Daniel et al., 2011), in practice it is recognized that neglect is a "complex and multifaceted phenomenon" (Stone, 1998: p. 90) and "notoriously difficult to define" (Turney, 2000: p. 249). Moreover, in China, where no legal definition exists there is no consensus among Chinese researchers regarding how neglect should be defined and conceptualized (Liao et al., 2011).

Critiques of narrow definitions of neglect have highlighted the focus on parental omission rather than on children's unmet needs (Dubowitz et al., 1993; 2007). Dubowitz et al. (1993) argue that legal definitions of neglect tend to be narrow and carry the implicit assumption of parental responsibility, while failing to recognize the complexity of the phenomenon. They argue that the focus of concern should 'be on children and their unmet needs, rather than on the presence or absence of parental (or caregiver) behaviors.' Narrow definitions may fail to take into account the dynamic nature of children's age and developmental levels, which influences the level and type of care they require (Zuravin, 2001). Crucially, by defining neglect more broadly, Dubowitz et al. (1993) argue that neglect should be conceptualized within an ecological framework, recognizing the wider social context in which children are neglected. At policy level in the UK, the *Framework for the Assessment in Need and their Families* (DoH, 2000) promoted a broader understanding of neglect taking an ecological approach that positions the child within the family, wider community, and social factors. Although in England the government's revised statutory guidance *Working Together* (HM Government, 2013) has superseded that which originally accompanied the Assessment Framework in 2000, the model itself is retained in the new guidance. A systems approach allows the practitioner to understand the interplay of biological, psychological, social, and cultural factors in child neglect. Further, it is child centered and rooted in child development specifically through the incorporation of the dimension of the 'child's developmental needs.' In countries with a 'family service' model, including Denmark, Finland, Sweden, and Norway (Skivenes, 2011), child maltreatment is understood in terms of the wider context and circumstances in which a child is cared for including psychological, familial, and socioeconomic problems (Cocozza and Hort, 2011).

Neglect can be defined variously in terms of parental characteristics or behavior or the impact of both of these factors upon the child (Daniel et al., 2011). Parental care, and lack of it, clearly exists on a continuum, ranging from excellent to grossly inadequate (Dubowitz, 2007). Given that neglect exists on a continuum, Dubowitz (2007: p. 604) suggests that it is difficult to identify when "exactly the inadequacy of care becomes problematic" with practitioners relying on measures such as 'faltering growth' or the notion of developmental milestones.

Research highlights the heterogeneity of neglect as a phenomenon. For example, Zuravin (2001) identifies several subtypes of neglect with different operational definitions. Determining neglect "rests on an assessment of frequency of the omission as well as the likelihood and severity of harm" (Dubowitz et al., 1993: p. 18). Some omissions in care are not likely to be harmful unless they are more frequent (Dubowitz et al., 1993). Zuravin (2001: p. 50) suggests that certain types of neglect require chronicity, for example, in discussing educational neglect she suggests that "a child must be absent from school for a minimum of twenty days without legitimate reason". A further complexity concerns whether definitions should include intentionality, whether neglect is the result of a desire to harm, of 'ignorance' or of circumstances such as poverty (Zuravin, 2001). Liao et al. (2011: p. 1710) suggest that in Chinese culture neglect is understood as being "unintentional and is generally attributed to ignorance." A broader definition of neglect widens the responsibility from parental blame to include wider social and institutional responsibility.

From this brief review of the existing literature on defining neglect, it is evident that wide-ranging categorical and conceptual differences exist within and between definitions. Such differences are likely to both reflect and influence how neglect is dealt with by social workers in practice and the way in which service users understand the basis of professional involvement in their lives.

Constructing Child Neglect

Several authors, mainly within the UK and US, have explored the constructions and discourses given precedence within social work practice relating to neglect.

A Preoccupation with Dirt and Disorder

In his study of child neglect responses in a social work team, Scourfield (2003) found that professional discourses drew upon selected messages from research. He cites the influence of findings from the *Bridge Child Care Consultancy* (1995) and *Child Protection: Messages from the Research* (DoH, 1995). The former emphasizes physical neglect and the servicing of the

child's body, "their neglected child is dirty and smelly" (Scourfield, 2003: p. 111), while the latter highlights the emotional impact of parenting styles characterized as "low warmth, high criticism" (DoH, 1995: p 19). Interestingly, the practical construction of neglect was "more influenced by the discourse of the servicing of the child's body" (Scourfield, 2003: p. 112). Swift (1995) found that physical neglect predominated both in relation to casework records and in social workers' talk about neglect. Similarly, in considering factors of concern in a chronic case scenario, Horwath (2007) found that 135 professionals focused on lack of supervision and physical harm with only six considering lack of stimulation and nine poor nurturing.

Swift (1995) and Scourfield (2003) indicate that there exists a professional preoccupation with dirt and disorder, both in relation to the child's appearance and home conditions. In Scourfield's (2003) study deciding whether a home was tidy and clean was "judged against tacit standards," involving levels of order and cleanliness (Scourfield, 2003: p. 113). Additionally, adequate feeding was "another abiding concern about the servicing of children's bodies" (Scourfield, 2003: p. 116). The emphasis in practice on physical neglect is noteworthy, given that there is also some evidence that professionals recognize the damaging effects of emotional neglect. For example, Horwath (2007) found that 75% of respondents agreed that a high criticism, low-warmth environment is damaging for a child.

Measuring and Proving Neglect

The dominant focus on physical neglect in practice may be explained through its being observable and visible: "you can see and smell dirt, check a child's weight" (Scourfield, 2003: p. 126). In contrast, emotional neglect is considered less easy to evidence, viewed as less tangible, more subjective, not as visible, and more difficult to observe developmental delays (Casey, 2013). Several studies have indicated that physical neglect dominates as it is easier to prove in court and in legal terms. The visibility of physical neglect is well placed in a legal context embedded in objectivity, 'facts,' and rational evidence (Casey, 2013). At the same time, "harm to children through neglect is difficult to establish in legal terms because it is incremental, often invisible and not directly observable as an outcome of parental behaviour" (Swift, 1995: p. 84). In a context in which professionals are under considerable pressure to 'get it right,' it is not surprising that "dirty, hungry, ill-clad children are bodies of evidence…concrete proof of parental failure to clean, feed and clothe" (Scourfield, 2003: p. 123). Moreover, this focus on physical needs is embedded within a climate preoccupied with risk (Scourfield, 2003) and "may reflect the incident driven system which has dominated child welfare practice for over twenty years" (Horwath, 2007: p. 1291). This focus on physical neglect and needs marginalizes the child's other needs (Horwath, 2007) and presents a partial view of children's welfare (Scourfield, 2003).

'Good Enough' and 'Bad Enough' Parenting

Given the complexities of defining neglect, the point at which care becomes inadequate is difficult to ascertain. In Rodriguez's (2002) study, investigating abuse-reporting decisions of professionals on abuse scenarios, accuracy was lowest for child neglect. Identifying neglect involves professional judgment of whether parenting is of an adequate standard (Lawrence and Irving, 2004). In Horwath's study, 25% of professionals believed staff in social work departments accept lower standards of parenting than other professionals (Horwath, 2007). Alongside differences between professional groups, the notion of 'good enough' parenting is also contested between social workers. In the Finnish context, the criteria for child maltreatment, varies between areas, teams, and officers as a result of "different organizational cultures, communities and…norms and expectations about appropriate childhood and parenthood" (Pösö, 2011: p. 116). In Casey's (2013) study, standards were viewed as dependent on experience, with newly qualified workers adopting higher standards of care to warrant action. Accepting lower standards of parenting may result when social workers work to the rule of optimism (Daniel, 2000). While social workers may accept relatively low standards of care, thresholds into services may be high. In this context, professionals make decisions as to whether care is bad enough to warrant services. Very high thresholds can cause cases to be closed or deregistered from child protection services too rapidly, resulting in situations of repeated registration, deregistration, and reregistration (Casey, 2013). Defining neglect becomes part of service rationing strategies, with levels of 'good enough' care determined by available resources (Casey, 2013).

Contextual Factors

Gender and Neglect

Mothers are "overwhelmingly identified as perpetrators of neglect" (Daniel and Taylor, 2001: p. 24). The question of responsibility for the maltreatment of children is "a key gender controversy" (Scourfield, 2003: p. 23) and studies have identified that gendered constructions of women are mainstream organizing principles (Scourfield, 2003). As Turney (2000: p. 50) argues, care and nurturing are identified with the feminine, in the West: "the association between caring and mothering…has particular salience for a consideration of child neglect."

Social workers' holding of women responsible (Strega et al., 2008) in cases of neglect builds on and reproduces traditional psychoanalytical positions and attachment theory, which are explored below, specifically around the 'good enough' mother. This association between femininity and caring has implications for understanding neglect with any breakdown in care being attributed to women (Daniel and Taylor, 2006). As Daniel and Taylor (2006: p. 427) elaborate, the theoretical underpinning of the majority of research on neglect is that: "neglect is an indication of a lack of nurturing; nurturing is carried out by mothers; therefore, when nurturing is absent the problem must be a problem of mothering."

Research and theory on child neglect is influenced by the assumption of women as primary caregivers (Daniel and Taylor, 2006). When fathers are included, research "tends to focus on the father's relationship with the mother and his presence or absence in the home" (Daniel and Taylor, 2006: p. 428). In Gaudin et al. (1996) study, fathers had limited involvement in child rearing and supporting the mother.

The mother then becomes a key variable in writing about child neglect, in terms of role rejection, lack of nurturing, knowledge, immaturity, and poor nurturing (Swift, 1995). Swift (1995: p. 89) suggests that "many researchers are concerned with establishing the main causal variables of neglect, they also contribute to the definition of the problem by framing it in personal and intra-familial terms."

Characteristics of Poor Parenting

A range of studies have investigated the characteristics of neglectful parents, with the focus mainly on mothers' personalities, caring roles, attitudes, relationships, and attachment histories (Coohey, 1995; Hildyard and Wolfe, 2007; Polansky et al., 1981). Egeland and colleagues' longitudinal study of Minnesotan mother–child interaction among first-time mothers considered to be at risk of maltreatment concluded that, for children aged six, "psychologically unavailable mothers were more tense, angry, depressed and confused" and less intelligent (MacDonald, 2001: p. 50). Similarly, Hildyard and Wolfe's (2007: pp. 898–899) comparative study of neglectful and nonneglectful mothers' responses to child emotions and behaviors found that "neglectful mothers had significantly more children; had experienced a greater degree and severity of maltreatment in their own childhoods; were more likely to be unemployed…and reported significantly more depressive symptomatology than the comparison sample."

Connell-Carrick and Scannapieco (2006) found that neglectful caregivers were characterized by a history of depression and attempted suicides. Substance abuse, mental illness, or learning difficulties were common in case files in Stone's (1998: p. 92) study that concluded that "most caregivers were significantly damaged individuals who were ill-equipped, emotionally or practically, to care for children." In Casey's (2013) study there was a preoccupation with the balance between child and parental 'needs' within case files and professional dialog: neglectful mothers were considered a risk when they put their own needs first. Gaudin et al. (1996: p. 367) examined casework assessments and found that neglectful families were viewed as "less healthy; less able to resolve conflicts; less cohesive…less well led and less verbally expressive." Similarly, Connell-Carrick and Scannapieco (2006) indicate that neglectful parents have fewer parenting skills, including effective communication and coping with stress. They were also viewed as lacking empathy and expressing dislike toward their child.

Crittenden (1999) identifies three 'types' of neglect within which the characteristics of parents are seen to differ:

- *Disorganized neglect*: families are characterized as multi-problem, disorganized, and crisis prone. Parents are described as failing to attend to children's needs of all kinds until crisis point;
- *Emotional neglect*: parents may provide materially for their children but do not give adequate attention to their emotional needs; and
- *Depressed neglect*: parents are withdrawn and do not perceive their children's needs. Children become silent and depressed.

Drug and Alcohol Misuse and Other 'Risk' Factors

Studies have also focused on specific risk factors for child neglect, including alcohol and substance misuse, mental health issues, learning disability, and domestic violence. Cleaver et al. (2007) highlight considerable evidence showing that children who grow up in families where there is domestic violence and/or parental alcohol or drug misuse are at increased risk of significant harm. Further, Forrester (2000: p. 241) found that "substance-using families were very significantly over-represented in neglect cases" with 9 out of 10 children whose parents used heroin registered as neglected. Alcohol misuse was also cited in 20 out of the 59 children registered for neglect. Similarly, Christoffersen and Soothill (2003) found that a mother's alcohol abuse contributed significantly to children's neglect. Cleaver et al. (2007) found that in social work case files, three-quarters of children living with domestic violence and/or substance misuse had unmet needs in at least one area of development. Cleaver et al. (2007) also revealed how many families experienced a combination of domestic violence, parental alcohol misuse, drug misuse, mental illness, and learning disability.

Research indicates that fathers may exacerbate the likelihood of neglect, e.g., if they are using household income to finance substance misuse. Scourfield (2000: pp. 376–377) argues that "the abusive man has a complementary relationship to the neglectful woman where his violence, abuse of money, avoidance of domestic work and frequent absence from the home are seen to hinder her care of the children." In Casey's (2013) discussion of the professional construction of 'responsible neglectful mother/dangerous or absent men', women were held responsible for being unable to meet the child's needs as a consequence of the emotional impact of men's violence.

Family 'Dysfunction'

Coohey (1995: p. 885) argues that "neglectful mothers exchange fewer resources with both their partners and mothers." Stone (1998) found that 7 out of 20 neglectful families lived in reconstituted families: mostly a mother plus a new partner. Relationships and family dysfunction were central concepts guiding practitioners' understandings of how children became neglected (Stone, 1998). Gaudin et al. (1996) compared family functioning in nonneglectful and neglectful families and found that neglectful mothers had more unresolved family conflict than those in the comparison group. Neglectful families also rated themselves as having "less closeness and less clear internal family boundaries; poorer negotiating skills; …less responsiveness to other family members' statements; less warmth…and less empathy toward one another" (Gaudin et al., 1996: pp. 368–369). Coohey (1995: p. 885) found that neglectful mothers "were more likely to state their mothers had fewer positive attributes…their relationship was less positive and they were less interested in receiving resources from their mothers." Significantly, explicit discussion of fathers is absent from many studies of neglectful families and, where they are included, are addressed primarily regarding their relationship with the mother (Daniel and Taylor, 2006) and being a resource for the mother's caring

role. Casey (2013) found that within professional constructions, mothers were blamed for the impermanency, chaotic, and 'inappropriate' nature of their relationships with men contributing to or exacerbating neglect.

Intergenerational Neglect

The intergenerational nature of neglect and abuse has been a consistently reported, but somewhat controversial finding of child maltreatment research (Connell-Carrick and Scannapieco, 2006; MacDonald, 2001). Stone (1998: p. 92) found repeated references in case file data of mothers' own experiences of neglect and abuse: "mum has never experienced effective nurturing parenting herself." In this study, 18 out of 20 cases scored positively for poor parenting of caregivers, with many having a history of neglect or abuse and experiences of the care system. Moreover, Hildyard and Wolfe (2007) found that mothers with an unresolved attachment status had the most severe difficulties in interpreting their own child's emotions. Hildyard and Wolfe (2007: p. 904) argue that research has indicated that experiences of childhood maltreatment, produces "negative representational models…based on…self-blame and shame," adversely affecting caregiving abilities.

Theories about the intergeneration transmission of child maltreatment are highly gendered, with the overriding focus on the mother–daughter relationship. For example, Coohey states (1995: p. 893) that "the mechanisms that perpetuated child neglect may lie in personality and behavioural traits of the mother…that are then reproduced in the adult daughter through the mother-daughter relationship."

Attachment Theory and Mother-Blaming

Swift (1995) suggests that 'bonding theories' are highly influential within social work practice in cases of neglect, and establish the *mother* as primarily responsible for meeting both material and emotional needs of their children. An implicit conceptualization of the 'good enough' mother is central to all of these theories. Although an understanding of attachment theory is considered crucial in identifying neglect, attachment theories have been critiqued for upholding mothers as solely and individually responsible for inadequate care and for blaming and pathologizing women.

Furthermore, physical descriptions of mothers "directs us to see poor care as closely connected to 'defects' in the mother and simultaneously directs attention away from the experienced problems of the mother" (Swift, 1995: p. 110). Although, as Swift (1995: p. 89) suggests Polansky's influential studies deemed poverty and social deprivation as important factors when investigating neglect, the personality features of mothers were seen as the overriding and primary cause.

Specifically, the 'cycle of neglect' discourse "render[s] the social and economic context not only invisible but irrelevant" (Swift, 1995: p. 113). The neglect schema, argues Swift (1995: p. 99), validates preexisting individualized services, which are 'directed at changing people rather than addressing social ills': services that provide 'reparenting' to break the cycle. Hence, mothers are policed to care, to reach a state where she is less needy, a state, which she needs to arrive at herself in the face of limited resources (Swift, 1995). Such blaming ignores the reality of women's lives, including material factors of limited financial resources or housing (Kufeldt et al., 2003).

Poverty and Neglect

Swift (1995) identifies two main historical discourses about the attribution of cause surrounding child neglect: personality and poverty. The relationship between poverty and child neglect is now commonly recognized. Slack et al. (2004) highlight an inverse relationship between income levels and child neglect. Similarly, Stone (1998: p. 93) highlights that "neglected children evidently suffer from poverty both in their material and their emotional environments." Professionals in Casey's (2013) study recognized that living in poverty can adversely impact the 'ability to parent' and increase the risk of neglect, while being aware of the danger in assuming that poverty was the sole cause or factor in cases of child neglect.

Beyond Simple Causation

Despite the above findings, there has, however, been a move away from notions of poverty as a simplistic *causal* factor of child neglect, toward an acknowledgment of the complex *relationship* and correlations between the two (Dubowitz et al., 1993; Slack et al., 2004). Turney (2000: p. 54) argues that practitioners need to consider the "importance of intrapersonal dynamics but…also highlight the nature of relationships between individuals, their families and wider communities." Garbarino and Collins (1999) emphasize the interplay of biological, psychological, social, and cultural factors.

At the same time, there is evidence that high unemployment and poverty hinder parental ability to provide care, increasing the likelihood of neglect. Garbarino and Collins (1999) suggest that neglect is a function of influences on caregivers that lead to lowered morale, deteriorated functioning, and depression. Cleaver et al. (2007) have highlighted that domestic violence and parental substance misuse can impact on housing, income, and social integration. Studies have indicated that parents' personal issues interact with contributory factors including poor housing, an absence of reliable child care, inadequate income, and institutional constraints (Casey, 2013; Dubowitz et al., 1993; Garbarino and Collins, 1999).

Connell-Carrick and Scannapieco (2006) found that families who were rated as having a poorer social climate (including social support and isolation) were more likely to be substantiated for neglect. Dubowitz et al. (1993: p. 9) argue that "the interactions between children and parents are influenced by community and societal factors such as the availability of child care and poverty." Swift (1995) argues that living in poverty impedes parents' capacity to improve their parenting through an absence of access to resources.

Social Neglect and Poverty

By positioning poverty as a contributory factor, social, community, and institutional forms of neglect are highlighted (Garbarino and Collins, 1999; Lawrence and Irving, 2004). As Garbarino and Collins (1999: p. 17) argue "we already know that rates of neglect are higher among poor people. So when

society ignores them, it contributes to child neglect." Institutional policies that ignore the needs of children become a form of institutional neglect (Garbarino and Collins, 1999). Dubowitz (2007) argues that poverty in itself is a form of social neglect. Hence, the neglect of children "becomes not simply an indictment of those families which fail to meet the basic needs of their children, but also of a society which fails to construct and maintain an infrastructure which facilitates parenting and values children" (Stone, 1998: p. 95).

An ecological conceptualization of neglect (Dubowitz et al., 1993: p. 10) recognize that "although parents are responsible for nurturing and protecting their children, social factors influence their ability to provide that care." An ecological approach would allow for the "*shared responsibility* among individuals, families, communities and society. Replacing individual blame by a shared responsibility enables a more constructive approach, and suggests interventions be targeted at multiple levels" (Dubowitz et al., 1993: p. 15). Similarly, Turney (2000: p. 54) advocates that social workers rethink the concept of care, taking into account the political and social contexts in which it takes place: "a more holistic account of the process and relationships of caring reduces the likelihood of a simple mother-blaming response to neglect and promotes the chance of a response based on support." For example, Swift (1995) found that reference to social and economic issues did not appear on case files pertaining to neglect, with child welfare rooted in the private family and obligation of parents to provide for children within the family structure: "the organization of records provides a format for noticing and collecting information that simultaneously reflects and supports the personalised stream of discourse about neglect and not the discourse to do with poverty" (Swift, 1995: p. 95).

Conclusion

This article has highlighted how the 'neglect of neglect' has been challenged through recognition of its increased incidence and evidence of its adverse effects on children. Definitions of neglect are varied and contested and can be defined narrowly and more broadly. It is recognized that neglect is complex, multifaceted, and difficult to define. However, common to most definitions is the persistent parental omission of care, a failure to provide, and to meet a child's basic needs. There is an emphasis in practice on physical neglect, viewed as observable and visible and easier to prove in court and in legal terms. The literature also highlights that given the complexities of defining neglect, the point at which care becomes inadequate is difficult to ascertain. The notion of 'good enough' parenting is contested. The range of studies that explore the characteristics of neglectful parents, including mothers' personalities, caring attitudes, relationship and attachment history, and risk factors for neglect have also been discussed. This article has explored the major discourses and contextual issues that have emerged, including neglect and poverty and the gendered nature of child neglect and policy responses.

See also: Child Protection.

Bibliography

Casey, B., 2013. Performing Child Neglect in Social Work Practice (unpublished Ph.D. thesis). School of Applied Social Sciences, Durham.

Christoffersen, M.N., Soothill, K., 2003. The long-term consequences of parental alcohol abuse: a cohort study of children in Denmark. Journal of Substance Abuse Treatment 25 (2), 107–116.

CIS, 2008. Canadian Incidence of Reported Child Abuse and Neglect Guidebook. http://cwrp.ca/sites/default/files/publications/en/CIS-2008_Guidebook.pdf (accessed 01.09.2013).

Cleaver, H., Nicholson, D., Tarr, S., Cleaver, D., 2007. Child Protection, Domestic Violence and Parental Substance Misuse: Family Experiences and Effective Practice. Jessica Kingsley Publishers, London.

Cocozza, M., Hort, S., 2011. The dark side of the universal welfare state? Child abuse and protection in Sweden. In: Gilbert, N., Parton, N., Skivenes, M. (Eds.), Child Protection Systems: International Trends and Orientations. Oxford University Press, New York, pp. 89–111.

Connell-Carrick, K., Scannapieco, M., 2006. Ecological correlates of neglect in infants and toddlers. Journal of Interpersonal Violence 21 (3), 299–316.

Coohey, C., 1995. Neglectful mothers, their mothers, and partners – the significance of mutual aid. Child Abuse and Neglect 19 (8), 885–895.

Crittenden, P.M., 1999. Child neglect: causes and contributors. In: Dubowitz, H. (Ed.), Neglected Children: Research, Practice and Policy. Sage Publications, Thousand Oaks, pp. 47–68.

Daniel, B., 2000. Judgements about parenting: what do social workers think they are doing? Child Abuse Review 9, 91–107.

Daniel, B.M., Taylor, J., 2001. Engaging with Fathers: Practice Issues for Health and Social Care. Jessica Kingsley, London.

Daniel, B.M., Taylor, J., 2006. Gender and child neglect: theory, research and policy. Critical Social Policy 26 (2), 426–439.

Daniel, B., Taylor, T., Scott, J., with Derbyshire, D., Neilson, D., 2011. Recognising and Helping the Neglected Child: Evidence-based Practice for Assessment and Intervention. Jessica Kingsley Publishers, London.

Department of Health, 1995. Child Protection: Messages from the Research. HMSO, London.

Department of Health, 2000. Framework for the Assessment of Children in Need and Their Families. HMSO, London.

Dubowitz, H., 2007. Understanding and addressing the 'neglect of neglect': digging into the molehill. Child Abuse & Neglect 31 (6), 603–606.

Dubowitz, H., Black, M., Starr, R.H., Zuravin, S., 1993. A conceptual definition of child neglect. Criminal Justice and Behavior 20 (1), 8–26.

Forrester, D., 2000. Parental substance misuse and child protection in a British sample. A survey of children on the child protection register in an inner London district office. Child Abuse Review 9 (4), 235–246.

Garbarino, J., Collins, C., 1999. Child neglect: the family with a hole in the middle. In: Dubowitz, H. (Ed.), Neglected Children: Research, Practice and Policy. Sage, Thousand Oaks, pp. 1–23.

Gaudin, J.M., Polansky, N.A., Kilpatrick, A.C., Shilton, P., 1996. Family functioning in neglectful families. Child Abuse and Neglect 20 (4), 363–377.

Glaser, D., Prior, V., 2002. Predicting emotional abuse and neglect. In: Browne, K., Hanks, H., Stratton, P., Hamilton, C. (Eds.), Early Prediction and Prevention of Child Abuse: A Handbook. John Wiley and Sons, Chichester.

HM Government, 2013. Working Together to Safeguard Children: A Guide to Inter-agency Working to Safeguard and Promote the Welfare of Children. Department for Education, London.

Hildyard, K.L., Wolfe, D.A., 2002. Child neglect: developmental issues and outcomes. Child Abuse and Neglect 26 (6–7), 679–695.

Hildyard, K.L., Wolfe, D.A., 2007. Cognitive processes associated with child neglect. Child Abuse & Neglect 31 (8), 895–907.

Horwath, J., 2007. The missing assessment domain: personal, professional and organizational factors influencing professional judgements when identifying and referring child neglect. British Journal of Social Work 37 (8), 1285–1303.

Kufeldt, K., et al., 2003. Critical issues in child welfare. In: Kufeldt, K., McKenzie, B. (Eds.), Child Welfare: Connecting Research, Policy and Practice. Wilfrid Laurier University Press, Ontario, Canada.

Lawrence, R., Irving, P., 2004. Redefining fatal child neglect. Child Abuse Prevention Issues 21, 1–23.

Liao, M., Lee, A.S., Roberts-Lewis, A.C., Hong, J.S., Jiao, K., 2011. Child maltreatment in China: an ecological review of the literature. Children and Youth Services Review 33 (9), 1709–1719.

MacDonald, G., 2001. Effective Interventions for Child Abuse and Neglect: An Evidence-Based Approach to Planning and Evaluating Interventions. John Wiley and Sons, Chichester.

Polansky, N.A., Chalmers, M.A., Williams, D.P., Buttenwieser, E.W., 1981. Damaged Parents: An Anatomy of Child Neglect. The University of Chicago Press, Chicago.

Pösö, T., 2011. Combating child abuse in Finland: from family to child-centered orientation. In: Gilbert, N., Parton, N., Skivenes, M. (Eds.), Child Protection Systems: International Trends and Orientations. Oxford University Press, New York, pp. 112–130.

Rodriguez, C.M., 2002. Professionals' attitudes and accuracy on child abuse reporting decisions in New Zealand. Journal of Interpersonal Violence 17 (3), 320–342.

Ruiz-Casares, M., et al., 2012. Supervisory neglect and risk of harm: evidence from the Canadian child welfare system. Child Abuse and Neglect 36, 471–480.

Scourfield, J., 2000. The rediscovery of child neglect. The Sociological Review 48 (3), 365–382.

Scourfield, J., 2003. Gender and Child Protection. Palgrave Macmillan, Houndsmills.

Sedlak, A.J., et al., 2010. Fourth National Incidence Study of Child Abuse and Neglect (NIS-4) Report to Congress. US Department of Health and Human Services, Administration for Children and Families, Washington, DC.

Skivenes, M., 2011. Norway: toward a child-centric perspective. In: Gilbert, N., Parton, N., Skivenes, M. (Eds.), Child Protection Systems: International Trends and Orientations. Oxford University Press, New York, pp. 154–179.

Slack, K.S., Holl, J.L., McDaniel, M., Yoo, J., Bolger, K., 2004. Understanding the risks of child neglect: an exploration of poverty and parenting characteristics. Child Maltreatment 9 (4), 395–408.

Stone, B., 1998. Child neglect: practitioners' perspectives. Child Abuse Review 7 (2), 87–96.

Strega, S., Fleet, C., Brown, L., Dominelli, L., Callahan, M., Walmsley, C., 2008. Connecting father absence and mother blame in child welfare policies and practice. Children and Youth Services Review 30, 705–716.

Swift, K., 1995. Manufacturing 'Bad Mothers': A Critical Perspective on Child Neglect. University of Toronto Press, Toronto; Buffalo, NY.

The Children Act, 1989. Her Majesty's Stationary Office (HMSO), London.

Turney, D., 2000. The feminizing of neglect. Child and Family Social Work 5 (1), 47–56.

Zuravin, S., 2001. Issues pertinent to defining child neglect. In: Morton, T.D., Salovitz, B. (Eds.), The CPS Response to Child Neglect: An Administrator's Guide to Theory, Policy, Program Design and Case Practice. National Resource Center on Child Maltreatment, Duluth, pp. 37–59.

Child Protection

Pam G Lister, Glasgow, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article will provide an overview of the international policy context of child protection. It will identify current trends in a range of countries. The key tension between promoting child and family welfare and ensuring child protection is discussed in these contexts. The article then discusses the practice issues of risk assessment and the skills required in child protection work. It will conclude by examining the child protection issues faced by black and minority ethnic children and disabled children.

Introduction

This article describes the international policy context of child protection before discussing key elements of risk assessment and skills required in practice and will conclude by highlighting key debates in the area.

The International Policy Context

The United Nations Convention on the Rights of the Child (UNCRC, 1989) established internationally agreed standards for children's rights. All but two countries, Somalia and the United States, have ratified the convention. While all 54 articles have implications for child protection, Article 19 'Protection from all forms of violence' is the most relevant article to explore in this paper. This article states:

1. States Parties shall take all appropriate legislative, administrative, social, and educational measures to protect the child from all forms of physical or mental violence, injury or abuse, neglect or negligent treatment, maltreatment or exploitation, including sexual abuse, while in the care of parent(s), legal guardian(s), or any other person who has the care of the child.
2. Such protective measures should, as appropriate, include effective procedures for the establishment of social programs to provide necessary support for the child and for those who have the care of the child, as well as for other forms of prevention and for identification, reporting, referral, investigation, treatment and follow-up of instances of child maltreatment described heretofore, and, as appropriate, for judicial involvement (UN General Assembly, 1989 Article 19: p. 7).

In 2011, the United Nations Convention on the Rights of the Child (UNCRC) established CRC General Comment 13 (GC13). The aim of GC13 was to provide clearer interpretation of Article 19 by establishing measurable indicators regarding outcomes for children, thereby strengthening child protection at all levels in every country.

The relationship between the UNCRC and national legislation varies across countries. For example, at a European level, the UNCRC is not formally incorporated into the European Convention on Human Rights (ECHR 1950), although some elements are included. Nations who ratify the treaty are bound to it under international law and are known as State Parties. However, full integration of the principles is not always evident in national legislation. In the UK, for instance, all policies must be compatible with children's rights as outlined in the UNCRC. However, although there have been attempts to harmonize the UNCRC with UK national legislation, it has not been fully incorporated. Since 2004, a Children's Rights Index has been established for over 190 countries to monitor the implementation of children's rights at national level (see Gran, 2010 for an overview). The UNICEF commissioned report (Lundy et al., 2012) provides an overview of how the UNCRC has been incorporated into domestic legislation in 12 countries with an in-depth analysis of implementation in 6 of these countries. The report also examined nonlegislative means by which nations may fulfill their duties under the convention including independent human rights organizations, national strategies, visible cross-sectoral coordination, child impact assessment, budgeting, data collection, training and awareness, and participation.

There are significant differences in how governments interpret their statutory duties to protect children. Gilbert et al. (2011) argue that there are typically two orientations toward child protection practice, one which focuses on child protection and is typified by concentration on investigation and the establishment of whether or not abuse has taken place, and a second which takes a family service approach based on assessment of need and typically leading to a therapeutic response. They argue that Anglo-American countries tend to take a child protection response, while European and Nordic countries broadly favor a family service response. However, there are significant variations within these broad approaches.

Within the child protection-based jurisdictions, there are different approaches to mandatory reporting (see Gilbert et al., 2011). There is mandatory reporting in Canada, Australia, and in some American states. However, there is no mandatory reporting in the United Kingdom and New Zealand. The countries, and states within countries, differ with regard to the existence of national guidance and procedures, and the use of specific assessment tools. Furthermore, there are different trends developing in each country. In England, following a major review of child protection (Munro, 2011) there has been a shift from a child protection orientation, toward one of

prevention. In the US, differential responses to child protection concerns are developing and in some states where the family group decision-making model from New Zealand has been introduced (Duerr Berrick, 2011), while in Canada there has been an intensification of the child protection model, although, parallel to this development, Canadians have also adopted family group conferencing (Swift, 2011).

Similar differences exist between Nordic and continental countries, which use a broader family service approach. In Sweden and Finland, while child protection work is located within child and family welfare services, there is mandatory reporting. Additionally, there is a growing concern that the Swedish state intervenes too quickly (Cocozza and Hort, 2011). Similar debates are taking place in other parts of continental Europe. For example, Wolff et al. (2011) argue that the German system fluctuates between strengthening the child welfare approach and adopting an investigatory risk assessment approach. The Belgian response to child and family welfare concern has historically been based on principles of social responsibility but this approach could be under threat in the future, following cases such as the Dutroux Case in 1996 after which public demonstrations took place to demand that more attention be paid to the child and for more substantial professional and judicial responses (Desair and Adriaenssens, 2011).

A different pattern emerges in the developing world where "sophisticated child abuse services are a relatively recent phenomenon" (Lachman and Poblet, 2002: p. 588). Moreover, the focus of child protection research and practice has been on issues such as poverty (Egibo, 2002), children affected by HIV/Aids (Nyandiya-Bundy and Bundy, 2002), and children affected by war and displacement (Killian, 2002).

Lachman and Poblet (2002) argue that

> Intrafamilial abuse must be addressed by child protection services that have a comprehensive and holistic approach and therefore services should also tackle extra-familial abuse … The poverty of the South should not negate the development of child protection services. Rather it should encourage the development of a different model, where child protection services exist despite war and death, where they are enmeshed within the services of a community, and where the participation of the community itself and children is seen as paramount.
> (p. 591)

In a similar vein, following a detailed statistical analysis of the consequences of child maltreatment in high-income countries, Gilbert et al. (2011) argue that international comparative studies outside of North America and northern Europe are needed "to help learn lessons from different settings about how to prevent child maltreatment and its consequences" (p. 10). They suggest that the changing patterns in child maltreatment between countries require further investigation. For example, one significant difference between high-income countries and developing countries is what appears to be a decrease in high-income countries of physical and sexual abuse. What lessons could be learned from this trend, which could inform child welfare and protection services in developing countries? They suggest further research is also needed into the characteristics of families, and communities and services, which promote health development in children (see this study for detailed statistical analyses of incidence and prevalence).

Some of this comparative research is being undertaken by the International Society for the Prevention of Child Abuse and Neglect, which produces a text on world perspectives on child abuse every 2 years (Dubowitz, 2012). This is informed by findings from a 2012 World Perspectives Survey completed by 69 countries, representing 77% of countries where the researchers were able to find a potential respondent. As the author acknowledges, the report offers a "somewhat representative view of the state of child protection-across different regions of the world and according to countries' income levels" (Dubowitz, 2012: p. 5), making the findings an important source of cross-country comparisons. It offers information on how child maltreatment is defined in different countries, some insight into incidence and prevalence rates, the surveillance of child maltreatment and current responses to child maltreatment worldwide. The survey found that the most common forms of maltreatment considered to be child abuse and neglect were physical and sexual abuse (incest, sexual abuse, or pornography). Eighty percent of respondents also cited other issues like failure to provide adequate food and clothing, sexual exploitation, and forced labor. One of the behaviors mentioned least by respondents was that of physical discipline. While this was of concern to countries in northern Europe, it did not feature significantly in responses from Oceania and Africa. Child abandonment was also not uniformly considered to be a form of child maltreatment.

With regard to surveillance and responses to child protection, the survey found that the formal documentation of cases of child maltreatment was becoming widespread, with 54% of respondents reporting that their country maintained official statistics on reported cases and 90% of countries reporting that their country had an official policy regarding child maltreatment. Surveys were completed by 69 countries with 10 respondents from Africa, 10 from the Americas, 27 from Asia, 20 from Europe, and 2 from Oceania. Mandatory reporting by professionals had been established in the countries of 92% of respondents, specific criminal penalties in 90% and provisions for removing children from their parents and carers in 87%. However, while the existence of policies is encouraging, only one-third of respondents considered the policies to be comprehensively implemented in practice. Furthermore, while the researchers found that services such as home visiting and therapy were developing across all countries, very few services were deemed adequate with low income and African countries reporting the least resources.

Critical Issues in Child Protection Practice

Risk Assessment

At the centre of debates across countries is that of the nature of risk assessment in child protection. Child protection guidance differs according to country, depending on how risk is defined, if at all. Risk is not a quantifiable entity and professional judgment is core in the assessment of risk. Parton (1997) has long argued that the social work response to child protection

concerns have been overly risk averse, and intervention and responses to intervention have been characterized by a focus on blame and culpability.

However, analyses of child protection interventions, which were not successful and where a child has been severely harmed or killed, have identified some common failures with regard to risk assessment. Although there have been individual errors where social work services are involved, a common feature of cases of child death or severe injury has been related to errors of reasoning when social workers are faced with large quantities of information and have to make decisions in the context of confusion and chaos (Munro, 2011; Brandon et al., 2008, 2009). Research has found that professionals are prone to make certain errors when undertaking child protection assessments. They are more likely to take account of verbal information that is easily available to them, both at the beginning and of the assessment process (Munro, 1999). Practitioners are also more likely to take account of information, which is striking; so they are more likely to focus on visible information. Therefore, less-specific referrals, for example, involving neglect or emotional abuse, are given lower priority although such referrals are likely to result in the child suffering more harm. In a significant proportion of cases, it has been found that social workers have failed to assess the past history of families and so have missed significant patterns of behavior.

In the light of these concerns, one response has been the development of risk-assessment tools, procedures, protocols, and guidance for practice. These tools and guidance allow for sharing of information on what had been found to be a successful intervention and provide a structure for social work assessments (Munro, 2010). However, there are concerns that overdependency on such tools may result in a rigid responses in which the child does not remain at the centre of good child protection practice. Overemphasis on performance indicators may result in attention being concentrated on the most easily accessible and understood quantitative data, which then runs the risk of workers avoiding the analysis of complex qualitative information. These tools may also give a false sense of objectivity and accuracy, when the evidence of the accuracy of the predictive capacity of these tools is not conclusive. They may undermine the professional skills, autonomy, and judgment of those engaged in child protection work (Munro, 2011).

Skills in Child Protection Practice

Research and inquiries into child protection cases have identified the core cognitive and practice skills required by child protection workers. Attention has been paid to the importance of workers gathering historical information to build a family chronology through which patterns of behavior can be viewed over time. From whom workers gather information is crucial, and research has found that too often workers direct their attention to working with parents who are leading chaotic lives, often linked to domestic abuse, mental health and addiction issues, at the expense of seeking the views of children (Brandon et al., 2008, 2009).

A key issue in a number of inquiries has been parental deceit and feigned compliance (Brandon et al., 2008, 2009). Munro (2005) gives a frank statement on the crux of the matter:

> Professionals dealing with a concern of child abuse need to operate with a higher level of suspicion than usual: parents are anxious to hide abusive behaviour [and] are likely to be dishonest.
> (2005: p. 379)

Fauth et al.'s (2010) scoping study of research between 2000 and 2009 into effective practice to protect children living in families where parents are highly unwilling to engage with support services identifies that professional empathy and a focus on relationship building with parents needed to be balanced with an 'eyes wide open' approach in which professionals retain an authoritative stance. This involves social workers dealing openly with issues of power in the relationship while at the same time involving the family in identifying appropriate methods of support. Brandon et al.'s research (2008) identified a pattern of workers being overwhelmed by the complexity of such cases and developing a strategy of concentrating on the present, ignoring past history and being over optimistic at small changes in the parental situation or attitude, resulting in a 'start again syndrome' (p. 324).

However, although overattention to parental issues has been found to result in poorer outcomes for children, developing appropriate partnerships with parents where there are child protection concerns is an essential skill in child protection work, and the difficulties of developing effective partnership working with parents have been highlighted in a number of studies (Brandon et al., 2008; Fauth et al., 2010). Balancing building a relationship and developing critical analysis of the situation makes great demands on social workers' communication skills. Forrester et al. (2008) found that social workers had a tendency to engage in high levels of confrontation and display poor levels of listening when trying to discuss child protection concerns with parents. The authors found little evidence of workers empathizing with the parental viewpoints and argue that those who did form empathic relationships were able to raise, concerns, which could then be addressed in partnership.

There is some evidence that social workers can effectively build good working relationships with parents when there are child protection concerns. Spratt and Callan's (2004) study found that parents were very positive about the ability of workers to empathize and communicate with them. They found that a key feature of effective practice is the ability of the worker to be clear to parents about initial concerns they have in relation to issues of inadequate parenting in order that effective work on addressing the concerns could take place quickly and in partnership.

There has been a longstanding concern about the gendered nature of child welfare and child protection processes (Daniel and Taylor, 1999; Dominelli, 2002; Milner, 1993). With regard to working with mothers in the child protection process, several authors have emphasized that the majority of mothers in the child protection process are single mothers living in poverty and deprivation. They argue that child welfare services have taken a punitive approach to marginalized and disadvantaged young mothers who are subject to

excessive policing and surveillance (Dominelli et al., 2005; Rutman et al., 2002; Strega et al., 2008).

A number of reasons have been suggested to explain the absence of fathers from the child protection process. A fundamental issue is the predominant discourse of fathering, which it is suggested is viewed as having a higher status than mothering (Milner, 1993) and characterized as providing financial support rather than undertaking caring responsibilities with children. Fathers are perceived as being absent or dangerous, and work with fathers is not seen as core business of social work (Dominelli, 2002; Scourfield, 2006).

Although there is some research evidence to suggest that fathers want active roles in child welfare processes (Dominelli et al., 2011), there is little concrete evidence on what promotes engagement with fathers. However two literature reviews in this area conclude that early engagement, persistence, direct work with fathers, the provision of relevant and flexible services, and worker training are essential (Maxwell et al., 2012; Zgoni et al., 2013).

The failure to engage with children effectively in the child protection process has also been identified as a feature in a number of inquiries, which suggests that workers usually only take account of children's views where they supported the workers assessment of the situation. It has been argued that taking children's views seriously is a fundamental protective mechanism (Munro, 1999; Laming Report, 2003). Children are generally unclear about the child protection process in which they find themselves, and workers tend to underestimate the ability of children to express their view and be involved in decisions about them. Core skills of communicating with children are, therefore, the foundation of child protection practice and this requires workers being able to build relationships over a period of time. This is very difficult given the urgency of some child protection work and the imperatives of time scales for assessment and investigations in many child protection procedures.

Developing Understandings

Black and Minority Ethnic Children

There are two areas that are striking in the research in this field. The first is the concern raised about the treatment and status of child protection work with the indigenous population of countries of the Australia, Aotearoa/New Zealand, Canada, the United States, and Europe. Historical displacement and discrimination has led to current major concern with regard to child welfare and child protection for these diverse populations (Haig-Brown, 1988; Libesman, 2012). A report in Australia (2007) revealed the widespread violence and sexual abuse of women and children in the Northern Territory (Wild and Anderson, 2007). Oliver (2012) argues that intergenerational trauma and dysfunction, disadvantaged living conditions, alcohol and substance misuse, and the use of pornography are key elements in the maltreatment of Aboriginal children. She suggests that victims may be unable or reluctant to come forward to give evidence for fear of ostracism from the community.

With regard to indigenous communities in New Zealand, Taaouni (2012) argues that "Maori are experiencing a four – decade cultural renaissance marred by several horrific Maori child homicides" and that "Debate has raged around causes and solutions" (p. 161). She states that some commentators argue that these events are a result of inadequate Maori parenting and a 'warlike' society. Others argue that the abuse of violence is exaggerated and that abuse arises from the colonization of Maori culture. Therefore, social, economic, political, and cultural processes have established a context for violence (see Tait-Rolleston and Pehi-Barlow, 2001 for discussion on Maori social work).

In respect of the North American continent, the over-representation of Aboriginal children in child protection statistics, and out-of-home placement in Canada has been attributed to the multiple disadvantages experience by Aboriginal families such as socioeconomic factors (Trocme et al., 2004). For example, the poverty rates of Aboriginal children in Canada in 2000 was 41%, whereas the total poverty rate for children was 19.1% (Swift, 2011). Similarly, in the United States concern had been expressed about the major policy issues, which have not been addressed in the field of child welfare, leading to Native American children not being afforded the same rights of protection and welfare that is given to other children in the United States (Cross et al., 2000).

The UNICEF (2012) report on the implementation of children's rights in 12 countries provides a succinct summary of the issues for these groups of children:

> In all countries in the study, the most vulnerable groups of children (separated children, asylum seekers and indigenous children and asylum seekers and indigenous children and children in conflict with the law) continued to fare less well compared to peers. This was linked to higher levels of poverty and social exclusion, concerns that have increased significantly in recent years. In several countries interviewees suggested that separated children and asylum seekers were not seen as rights holders in the same way as other children, and this was linked to the weakness of the CRC in these areas.
>
> (Lundy et al., 2012)

In the Europe and North American context similar concerns emerge with regard to child protection services to black and minority ethnic children and families. In the UK, there is evidence to show that black and minority ethnic children are over-represented on child protection registers, placement in local authority care and referrals to child protection teams (Barn, 1993). There is a debate about why black and minority children are overly represented in child protection statistics, with some commentators citing parenting traditions and others the impact of institutional racism on black and minority ethnic children and families (Ahmed, 2005; Chand and Thoburn, 2006). Black and minority ethnic families are more likely to live in poverty, which is a significant indicator of the possibility of child maltreatment, particularly in the case of neglect (Dominelli, 1997; Stokes and Scmidt, 2011). The failure to acknowledge issues of race, language, culture, and religion are noted in a number of the analyses of child deaths (Brandon et al., 2008). Concern has been expressed about the overpathologizing of black and minority ethnic parents by professionals (Chand, 2000). However, more attention is paid in the literature to the concept of 'cultural

relativism' among professionals working with black and minority ethnic families (Barn, 2007; Laming Report, 2003). Cultural relativism is the process by which professionals make assumptions about cultural practices and accept different standards of care, as it is assumed that these practices are viewed as culturally acceptable in the family's community. Furthermore, cultures may be viewed as having a homogenous identity and the different cultures in which the child is embedded can be ignored, for example, the childhood culture, family culture, geographical culture, and religious and linguistic cultures (Connolly et al., 2005). Finally, systemic issues in safeguarding or child protection processes, where policy and procedures fail to address the needs of black and minority ethnic families are viewed as an important factor in the overrepresentation of black children in the child protection system (Chand, 2008).

There is a further issue regarding asylum-seeking children in Europe (Cardol, 2012), Canada (Ayoke, 2001), Australia (Bhahba, 2007), and North America (Molina, 2012; Somers et al., 2010). Child protection systems need to take into account the additional stresses faced by refugees and asylum-seeking parents, who may have experienced violence and persecution in their home country and then endure persecution, suspicion, and racism in their host countries (Cardol, 2012). Social work responses to the particular challenges of working with unaccompanied asylum-seeking children are a further area for practice development (Hill and Hopkins, 2009).

Disabled Children

The needs and rights of disabled children in need of child protection is an underresearched area. In an international literature review of disabled children and child protection, Stalker and McArthur (2010) identified 38 relevant studies. Overwhelmingly, the studies found that disabled children are more likely to be abused than nondisabled children. Evidence was found of underreporting of abuse of disabled children (Cooke and Standen, 2002; Hershkovitz et al., 2007). The correlation between the types of abuse and types of impairment was inconclusive, but the authors found that the evidence suggests that children with communication impairments, behavioral disorders, learning difficulties, and sensory impairments were more vulnerable to being abused. Evidence on the direction of causality, the extent to which maltreatment contributes to impairment or whether impairment predisposes children to abuse is also inconclusive.

Stalker and McCarthur (2011) found a 'complicated and inconsistent' (p. 9) pattern with regard to the gender of disabled children who had been abused. A similar inconsistent pattern was found with regard to the role of cultural and social factors, but with very little research in this area the authors considered it unwise to draw conclusions. However, Kapitanoff et al. (2000) argue that definitions of abuse vary across cultures, and prevalence in some black and minority groups may be confused by the fact that black and minority ethnic children are more likely to be in lower socioeconomic groups, which as shown above, is positively correlated with child maltreatment. Few studies had asked disabled children about their experiences of being abused or being involved in the child protection system. Similarly, they found a dearth of research, which examined professional responses to the abuse of disabled children or the effectiveness of child protection services for these children.

Conclusion

This article has given an overview of the key concerns facing policy makers and practitioners in the field of child protection in an international context. A key issue for policy makers is to ensure that the UNCRC is fully integrated into domestic law and policy. Evidence shows that where this is the case, children's rights are more fully upheld. Participation of children in decision making is a key element in ensuring children's rights. While progress has been made with regard to listening to children and ascertaining their views, participation of children is an underdeveloped area. At a practice level, social workers can contribute to upholding children's rights to be protected and to participate in decisions by ensuring children are listened to and encouraged to voice views and enabled to participate in decision-making processes. The UNCRC is an international mandate, which nations and countries can utilize to ensure children's rights and views are upheld.

See also: Child Labor; Child Neglect; Child Protection: International Issues; Children and Families in Social Work; Children and Law; Disability and Disabling Practices; Disability: Sociological Aspects; Indigenous Child Welfare Practice: The Canadian Case; Intelligence: Assessments of; International Issues in Child Protection; Law and People with Disabilities; Probation, offenders and Risk; Risk Assessment in Social Work; Social Work Theory; Street Children: Cultural Concerns.

Bibliography

Ayoke, J., 2001. Separated Children Seeking Asylum in Canada. UNHCR, Ottowa, Ontario.

Barn, R., 1993. Black Children in the Public Care System. Batsford, London.

Barn, R., 2007. Race, Ethnicity and Child Welfare: A fine balancing act. British Journal of Social Work, 37 (8) 1425–1434.

Bhahba, J., 2007. Seeking Asylum alone, a Comparative Study: A Comparative Study of Laws, Policy and Practice in Australia, the United Kingdom and the United States. Themis Press, Annadale, New South Wales.

Brandon, M., Belderson, P., Warren, C.R., Howe, A., Sidebotham, P., Dodsworth, J., Black, J., 2008. Analysing Child Death and Injury through Abuse and Neglect: What Can We Learn? A Biennial Analysis of SCRs 2003–5. University of East Anglia, Norwich.

Brandon, M., Bailey, S., Warren, C., Gardner, C.R., Sidebotham, P., Dodsworth, J., Warren, C., Black, J., 2009. Understanding Serious Case Reviews and Their Impact, a Biennial Analysis of SCRs 2003–5. University of East Anglia, Norwich.

Brown, L., Callahan, M., Strega, S., Walsmley, C., Dominelli, L., 2009. Manufacturing ghost fathers: the paradox of father presence and absence in child welfare. Children and Family Social Work 14 (1), 229–234.

Cardol, G., 2012. Non-national children and vulnerability: the child protection context. In: Sheehan, R., Rhoades, H., Stanley, N. (Eds.), Vulnerable Children and the Law. Jessica Kingsley, London and Philadelphia, pp. 145–160.

Chand, A., 2000. The over-representation of black children in the child protection system: possible causes, consequences and solutions. Child and Family Social Work 5 (1), 67–77.

Chand, A., 2008. Every child matters?: A critical review of child welfare reforms in the context of minority children and families. Child Abuse Review 17 (1), 6–22.

Chand, A., Thoburn, J., 2006. Research review: child protection referrals and minority ethnic children and families. Child and Family Social Work 11 (4), 368–377.

Cocozza, M., Hort, S.E.O., 2011. The dark side of the universal welfare state? Child abuse and protection in Sweden. In: Gilbert, N., Parton, N., Skivenes, M. (Eds.), Child Protection Systems. International Trends and Orientations. Oxford University Press, New York, pp. 89–111.

Connolly, M., Crichton Hill, Y., Ward, T., 2005. Culture and Child Protection: Reflexive Responses. Jessica Kingsley, London.

Cooke, P., Standen, P.J., 2002. Abuse and disabled children: hidden needs…? Child Abuse Review 11 (1), 1–18.

Cross, T.A., Earle, K.A., Simmons, D., 2000. Child abuse and neglect in Indian country: policy issues families in society. The Journal of Contemporary Human Services 81 (1), 49–58.

Daniel, B., Taylor, J., 1999. The rhetoric versus the reality: a critical perspective on practice with fathers in child care and child protection work. Child and Family Social Work 4 (3), 209–220.

Desair, K., Adriaenssens, P., 2011. Policy towards child abuse and neglect in Belgium: shared responsibility, differentiated response. In: Gilbert, N., Parton, N., Skivenes, M. (Eds.), 2011. Child Protection Systems. International Trends and Orientations. Oxford University Press, New York, pp. 204–222.

Dominelli, L., 1997. Anti-racist Social Work, second ed. Macmillan, Basingstoke.

Dominelli, L., 2002. Feminist Social Work Theory. Palgrave Macmillan, Basingstoke.

Dominelli, L., Strega, S., Callahan, M., Putman, D., 2005. Endangered children: the state as failed parent and grandparent. British Journal of Social Work 35 (5), 1123–1144.

Dominelli, L., Strega, S., Walsley, C., Callahan, M., Brown, L., 2011. 'Here's my story' fathers of 'looked after' children recount their experiences of the Canadian welfare system. British Journal of Social Work 41 (2), 351–367.

Dubowitz, H. (Ed.), 2012. World Perspectives on Child Abuse, tenth ed. International Society for the Prevention of Child Abuse and Neglect, Aurora, CO.

Duerr Berrick, J., 2011. Trends and issues in the US child welfare system. In: Gilbert, N., Parton, N., Skivenes, M. (Eds.), Child Protection Systems. International Trends and Orientations. Oxford University Press, New York, pp. 17–35.

Egibo, P.O., 2002. The effect of poverty on child protection in Africa, in: challenges facing child protection. Child Abuse and Neglect 26 (6–7), 587–617.

Fauth, R., Jelici, H., Hart, D., Burton, S., Shemmings, D., 2010. Effective Practice to Protect Children Living in Highly Resistant Families. National Children's Bureau, University of Kent.

Forrester, D., Kershaw, S., Moss, H., Hughes, L., 2008. Communication skills in child protection: how do social workers talk to parents? Child and Family Social Work 13 (1), 41–51.

Gilbert, R., Widom, C., Browne, K., Fergusson, D., Webb, E., Janson, S., 2008. Burden and consequences of child maltreatment in high-income countries. The Lancet. http://dx.doi.org/10.1016/S0104-6736(08) 61706–7. www.thelancet.com.

Gilbert, N., Parton, N., Skivenes, M., 2011. Child Protection Systems. International Trends and Orientations. Oxford University Press, New York.

Gran, B.R., 2010. Introducing the children's rights index. International Journal of Children's Rights 18, 1–17.

Haig-Brown, C., 1988. Resistance and Renewal. Surviving the Indian Residential School. The Secwepemc Cultural Education Society, Vancouver.

Hershkovitz, I., Lamb, M.E., Horowitz, D., 2007. Victimisation of children with disabilities. American Journal of Orthopsychiatry 77 (4), 629–635.

Hill, M., Hopkins, P., 2009. Safeguarding children who are refugees or asylum seekers: managing multiple levels of legislation and policy. In: Broadhurst, K., Glover, C., Jamieson, J. (Eds.), Critical Perspectives on Safeguarding Children. Wiley–Blackwell, Chichester, pp. 229–246.

Jemta, L., Fugli-Meyer, K.S., Oberg, K., 2008. On intimacy, sexual activities and exposure to sexual abuse among children and adolescents with mobility impairments. Acta Paediatrica 97 (5), 641–646.

Kapitanoff, S.H., Lutzker, J.R., Bigelow, K., 2000. Cultural issues in the relation between child disabilities and child abuse. Aggression and Violent Behaviour 5 (3), 227–244.

Killian, B., 2002. Is child protection possible in areas of war and violence? In: Lachman, P., et al. (Eds.), challenges facing child protection. Child Abuse and Neglect 26 (6–7), pp. 587–617.

Lachman, P., Poblet, X., 2002. Overview – lessons from the south, In: challenges facing child protection. Child Abuse and Neglect 26 (6–7), 587–617.

Laming Report, 2003. The Victoria Climbie Inquiry: Report of an Inquiry by Lord Laming Cm5730. The Stationery Office, London.

Libersman, T., 2012. International human rights law and the needs of indigenous children. In: Sheehan, R., Rhoades, H., Stanley, N. (Eds.), Vulnerable Children and the Law. Jessica Kingsley, London and Philadelphia, pp. 181–200.

Lundy, L., Kilkelly, V., Byrne, B., Kang, J., 2012. The UN Convention on the Rights of the Child. A Study of Legal Implementation in 12 Countries. UNICEF, London.

Maxwell, N., Scourfield, J., Featherstone, B., Holland, S., Thomas, R., 2012. Engaging fathers in child welfare services: a narrative review of recent research evidence. Child and Family Social Work 17 (2), 160–169.

Milner, J., 1993. A disappearing act: the differing career paths of fathers and mothers in child protection investigations. Critical Social Policy 13 (38), 48–63.

Molina, G.E., 2012. Florence immigration and refugee rights project (FIRRP) USA. In: Sheehan, R., Rhoades, H., Stanley, N. (Eds.), Vulnerable Children and the Law. Jessica Kingsley, London and Philadaelphia, pp. 93–108.

Molina, G.E., 2012. Unaccompanied children as illegal immigrants in the United States. In: Sheehan, R., Rhoades, H., Stanley, N. (Eds.), Vulnerable Children and the Law. Jessica Kingsley, London and Philadelphia, pp. 93–108.

Munro, E., 1999. Common errors of reasoning in child protection. Child Abuse and Neglect 23 (8), 745–758.

Munro, E., 2005. What tools do we need to improve identification of child abuse? Child Abuse Review 14 (6), 374–388.

Munro, E., 2011. The Munro Review of Child Protection: Final Report. The Stationery Office, London.

Nyandiya-Bundy, S., Bundy, R., 2002. The influence of HIV and AIDS on child protection in: challenges facing child protection. Child Abuse and Neglect 26 (6–7), 587–617.

Oliver, S., 2012. The victimisation of indigenous children. In: Sheehan, R., Rhoades, H., Stanley, N. (Eds.), Vulnerable Children and the Law. Jessica Kingsley, London and Philadelphia, pp. 127–144.

Parton, N., 1997. Child Protection and Family Support: Tensions, Contradictions and Possibilities. Routledge, London.

Rutman, D., Strega, S., Callahan, M., Dominelli, L., 2002. 'Undeserving mothers'? Practitioner experiences of working with young mothers in/from care. Child and Family Social Work 7 (3), 149–160.

Scourfield, J., 2006. The challenge of engaging fathers in the child protection process. Critical Social Policy 6 (2), 440–449.

Sobsey, D., Randall, W., Parrila, R.K., 1997. Gender differences in abused children with and without disabilities. Child Abuse and Neglect 21 (8), 701–720.

Somers, M., Herrera, P., Rodriguez, L., 2010. Constructions of childhood and unaccompanied children in the immigration system in the United States. U.C. Davies Journal of Juvenile Law and Policy 14 (2), 311–380.

Spratt, T., Callan, J., 2004. Parent's views on social work interventions in child welfare cases. British Journal of Social Work 34 (2), 199–224.

Stalker, K., McArthur, K., 2010. Child protection and disabled children: a review of recent research. Child Abuse Review 21 (1), 24–40.

Strega, S., Fleet, C., Brown, L., Dominelli, L., 2008. Connecting father absence and mother blame in child welfare policies and practice. Child and Youth Services Review 30 (7), 705–716.

Stokes, J., Scmidt, G., 2011. Race, poverty and child protection decision making. British Journal of Social Work 41 (6), 1105–1121.

Sullivan, P.M., Knutson, J.F., 2000. Maltreatment and disability: a population based epidemiological study. Child Abuse and Neglect 24 (10), 1257–1273.

Swift, K., 2011. Canadian child welfare: child protection and status quo. In: Gilbert, N., Parton, N., Skivenes, M. (Eds.), Child Protection Systems. International Trends and Orientations. Oxford University Press, New York, pp. 36–55.

Taaouni, R., 2012. Mana tamariki, takahi tamariki: Maori child pride, Maori child abuse. In: Sheehan, R., Rhoades, H., Stanley, N. (Eds.), Vulnerable Children and the Law. Jessica Kingsley, London and Philadelphia, pp. 127–144.

Tait-Rolleston, W., Pehi-Barlow, S., 2001. A Maori Social Work Construct. In: Dominelli, L., Lorenz, W., Soydon, H. (Eds.), Beyond Racial Divides: Ethnicity in Social Work. Ashgate, Aldershot.

Trocme, N., Knoke, D., Blackstock, C., 2004. Pathways to the over-representation of aboriginal children in Canada's welfare system. Social Services Review 78 (4), 577–600.

United Nations Convention on the Rights of the Child, 1989. http://www.ohchr.org/EN/ProfessionalInterest/Pages/CRC.aspx (accessed 19.08.13.).

Woolf, R., Biesel, K., Heinitz, S., 2011. Child protection in an age of uncertainty: Germany's response. In: Gilbert, N., Parton, N., Skivenes, M. (Eds.), Child Protection Systems. International Trends and Orientations. Oxford University Press, New York, pp. 183–203.

Wild, R., Anderson, P., 2007. Little Children are Sacred. Board of Inquiry into the Protection of Aboriginal Children. Northern Territory Government, Australia.

Zgnoni, L., Warbuton, W., Bussey, K., McMaugh, A., 2013. Fathers as 'core business' in child welfare practice and research: an interdisciplinary review. Child and Youth Services Review 35 (7), 1055–1070.

Child Protection: International Issues

Simon Hackett, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article considers the controversies, debates, and challenges associated with promoting children's welfare and ensuring their protection from abuse and harm. An overview of child protection as a concept and a practice is offered. Key concepts and definitions are examined emphasizing the contested nature of child abuse. Current issues in child protection theory, research, and practice at the international level are explored including the need to protect children in situations of 'natural' disasters and emergencies, as well as in political conflict and war. The global challenge of protecting children from sexual abuse and exploitation in the context of new technologies is also considered.

Introduction

Examples of the ill-treatment of children exist throughout history and across cultures, including infanticide, abandonment, forced marriage, child labor, and sexual exploitation. While the standards of care afforded to many children in previous centuries may fall short of those considered to be necessary in the present day, there is, at the same time, considerable evidence that children have been valued, nurtured, and cared for in every era:

> Every society has to deal with the issue of the care of its young and has devised some means of intervention into family life to ensure this. These efforts have been influenced by the cultures, material circumstances, technologies and politics impinging on those societies. (Corby, 2000: p. 20)

This article explores the controversies and debates associated with such efforts in order to offer an overview of child protection as both a concept and a practice. It begins by exploring and clarifying definitions before examining some of the key changes in focus in protecting children, as well as current challenges and emerging directions in child protection theory, research, and practice. While most attention previously has been given to child protection at the level of the individual state, the processes of globalization have brought about, now more than ever before, cross national issues which threaten the welfare and development of children. For example, the Pinheiro's (2006) *World Report on Violence against Children* for the United Nations found that physical violence against children in the home is widespread in all regions of the world and estimated that each year worldwide 150 000 000 girls and 73 000 000 boys experience sexual violence; 1 800 000 children are involved in commercial sexual exploitation and child abuse images; and 1 200 000 children are victims of trafficking. This means that child protection is no longer merely an issue that individual societies have to address, but a global concern. Therefore, this article focuses on child protection internationally as a complementary chapter to a further article in the Encyclopedia which primarily addresses the child protection system in the United Kingdom.

Definitions

Defining what is meant by 'child protection' and associated concepts such as 'child abuse,' 'welfare,' 'harm,' and even 'childhood' is a complex and contested activity. This means that conceptualizing child protection issues is fraught with definitional dilemmas and complexities. It is, therefore, vitally important to debate and clarify what is meant by the various terms associated with this field of practice to understand child protection at the international level.

Defining Childhood

Childhood itself is a social construct rather than a fixed biological period, with the rights and protections afforded to children and young people of particular ages and stages of development at times highly variable and contradictory (Ariès, 1962). For instance, the age at which children are held criminally responsible, can have sexual intercourse, are entitled to marry or vote has shifted throughout history and still varies considerably across, and even sometimes within, nation states. The United Nations (UN) (1989) *Convention on the Rights of the Child* defines a child as a person below the age of 18, unless the laws of a particular country set the legal age for adulthood younger. The UN Committee on the Rights of the Child, the monitoring body for the Convention, recommends that states where the age of majority is set below 18 should increase the level of protection for all children under 18.

Defining Children's Rights

Built on varied legal and cultural traditions, the UN *Convention on the Rights of the Child* provides a universally agreed set of non-negotiable standards and obligations for those nations who are signatories. The Convention outlines the fundamental rights of children, including the right to be protected from economic exploitation and harmful work, from all forms of sexual exploitation and abuse, and from physical or mental violence, as well as establishing that children should not be separated from their family against their will. Specifically, Article 19 of the Convention sets out

children's rights to protection from all forms of violence in the following terms:

> Children have the right to be protected from being hurt and mistreated, physically or mentally. Governments should ensure that children are properly cared for and protect them from violence, abuse and neglect by their parents, or anyone else who looks after them…. In most countries, laws already define what sorts of punishments are considered excessive or abusive. It is up to each government to review these laws in light of the Convention.

Additionally, Article 34 states that:

> Governments should protect children from all forms of sexual exploitation and abuse.

The UN Convention is widely acknowledged as the most complete statement of children's rights ever produced and is the most ratified international human rights treaty with all UN member states except for the United States, South Sudan, and Somalia having approved it. As such, it provides a useful foundation for the building of a theory and a practice of international child protection. At the same time, however, the Convention has generated continuing debate, with opponents fearing that ratification could undermine individual nations' sovereignty and interfere in the private lives of citizens (see Gran, 2010). Others have argued that simply adopting the UN *Convention on the Rights of the Child* does not guarantee the protection of children's rights in real-world situations (Blanchfield, 2013). For example, it has been noted that some countries widely criticized for abuses of children's rights, such as the Democratic Republic of Congo, China, and Sudan, are signatories (Blanchfield, 2013). Even a country such as the United Kingdom, where the notion of children's rights is well established socially and embedded within social policy and social work practice, has an ambiguous track record on ensuring the rights of *all* children. The UK government has been criticized for breaching some of its human and children's rights obligations in the way in which it approaches children involved in the asylum system (Clements, 2006) or in the way that young offenders are transferred from youth custody to adult prisons (Puffett, 2013). Additionally, Blanchfield points out that a number of Islamic countries have attached reservations to their adoption of the UN Convention which state that ratification would not apply to provisions that they deem incompatible with Islamic Shari'a law or values. She adds that this raises concern that such reservations could allow for broad interpretations of the Convention's provisions, particularly in the area of child marriage and education for girls (Blanchfield, 2013: p. 16). Ultimately, the most significant limitation of the Convention is not the rights it enshrines, nor its ability to influence legislation in those nations endorsing it, but the variability of actions which flow from such legislation in protecting children and their rights.

Defining Child Abuse

While the UN *Convention on the Rights of the Child* sets out the rights of children to be protected from all forms of violence and abuse, as outlined earlier, it does not offer definitions of specific types of abuse or the range of behaviors that might constitute them. This too is a contested and difficult area. Child abuse is as much as a social construct as is the concept of childhood (Corby, 2000). What are deemed to be optimal, acceptable, or abusive parenting practices vary considerably over time and across and within cultural contexts. For example, in some jurisdictions, the use of parental corporal punishment as a means of disciplining children is viewed as harmful and is outlawed, while in other cultures, it is seen as a necessary and responsible way of raising children. According to the World Health Organisation (WHO, 2013):

> Child maltreatment, sometimes referred to as child abuse and neglect, includes all forms of physical and emotional ill-treatment, sexual abuse, neglect, and exploitation that results in actual or potential harm to the child's health, development or dignity. Within this broad definition, five subtypes can be distinguished – physical abuse; sexual abuse; neglect and negligent treatment; emotional abuse; and exploitation.

The above-mentioned definition is one of thousands on child abuse that have been proposed internationally. The discrepancies and inconsistencies between them reflect real and, at times, deeply conflicting social values and standards relating to the state's role in intervening in the lives of its citizens. Nonetheless, there is now considerable support for the five subtypes of child abuse included in the above WHO definition. The development of these core categories of child abuse is, in large part, the consequence of the long history of social workers and other professionals being confronted with ever more complex and diverse cases where children have been harmed. This has forced appraisal and reappraisal of the nature and range of abuses faced by children. For example, it is widely stated that awareness of child abuse in the modern era stemmed from the ground-breaking work in the 1960s of American pediatrician Henry Kempe and colleagues who outlined the term 'battered child syndrome' thereby bringing the issue of child physical abuse into the professional and public consciousness. However, it was not until a decade and a half later until survivors of child sexual abuse began to raise awareness of this 'new' form of abuse (Amstrong, 1978). Awareness of emotional abuse and neglect followed yet later (Garbarino et al., 1986). The nature of sexual abuse perpetrated by children and young people (Hackett, 2004) and the abuse of children in institutions, such as the well-documented series of cases of abuse within the global Catholic Church (Keenen, 2012), are even more recent developments.

It is also necessary to operationalize each of the five subtypes of child abuse within national policy, legislation, and guidance. In the United Kingdom, this guidance comes in the form of the *Working Together* document which was first published in 1988 (DHSS, 1988) but which has been updated at regular intervals until the most recent iteration in 2013 (HM Government, 2013). While the categories or subtypes of abuse included in these documents have remained largely static, the accompanying guidance has been expanded considerably and the definitions of each abuse category has shifted considerably

as new knowledge and better research evidence about the nature of the form of abuse has been generated. In addition, there is a further complexity faced by social workers who need to work with general categories of abuse in their practice and make decisions about whether the circumstances of individual cases fall within these categories and, importantly, what action is therefore warranted. Defining and redefining child abuse, it appears, is a continual process.

Defining Child Protection

Finally, in terms of definitions, increasing awareness and a body of practice knowledge and empirical research about the various subtypes of child abuse has led to the development of a particular strand of social work practice which has become known as 'child protection.' 'Child protection' is a term that describes the philosophies, policies, standards, guidelines, and practices that guide when to, and how to, intervene in the lives of children in order to ensure their protection from these various types of abuse and maltreatment. While social workers do not have a monopoly on child protection practice, with professionals such as doctors, teachers, psychologists, and the police integral to the child protection task, child protection represents a field of social work practice that is recognized, and recognizable, internationally. For example, a review of published research within the last 2 years in the journal *International Social Work* reveals papers addressing child protection issues in the following diverse countries: Nigeria, Uganda, China, India, Brazil, El Salvador, Guatemala, Nepal, Sierra Leone, Turkey, the United Kingdom, and the United States. While all countries have some sort of child protection system, it is however clear that some are more developed and better resourced than others. It is important that each country develops its own approach taking into account tradition, culture, and the specific challenges it faces (Ortiz et al., 2012). Each country should identify the way in which social workers should respond to these challenges, while at the same time clarifying the child protection responsibilities of other professional groups. It is increasingly acknowledged that it is vital to involve those who are on the receiving end of child protection services, both children and parents and carers in this process, not least because their rights and their experiences have been overlooked in the past.

UNICEF (2006) defines eight components necessary for the building of protective environments internationally that will prevent and respond to child abuse, including actions across societal, governmental, and professional service levels:

- Strengthening government commitment and capacity to fulfill children's right to protection;
- Promoting the establishment and enforcement of adequate legislation;
- Addressing harmful attitudes, customs, and practices;
- Encouraging open discussion of child protection issues that includes media and civil society partners;
- Developing children's life skills, knowledge, and participation;
- Building capacity of families and communities;
- Providing essential services for prevention, recovery, and reintegration, including basic health, education, and protection; and
- Establishing and implementing on-going and effective monitoring, reporting, and oversight (UNICEF, 2006: p. 1).

Social workers' specific responsibilities in relation to these tasks vary considerably according to how their professional roles are conceived in different countries, but often they are central to the identification of cases of abuse and in the provision of services to help children recover from their experiences.

Current Challenges in Protecting Children at the International Level

Such is the diversity of factors associated with the abuse of children internationally that it is impossible here to offer a comprehensive list of the challenges currently facing the international community. Instead, a number of key indicative themes are presented which illustrate the challenges that social workers face internationally in their attempts to improve the protection of children in the communities in which they work.

Child Protection in the Context of Disasters and Emergencies

The United Nations estimates that 66 million children around the world are affected every year by 'natural' disasters such as floods, earthquakes, and drought. Dominelli (2012) points out that the term 'natural disaster' is problematic as the behavior of human beings strongly influences the occurrence of what are seen as naturally occurring crises and disasters. Child protection is a particular concern in situations of emergency and humanitarian crisis. Many of the specific elements that underlie such crises, such as displacement and poverty, bring with them specific risks to children. The likelihood that children are vulnerable to the range of forms of child abuse discussed earlier is heightened in situations of natural disasters. Disasters can undermine community and family structures and large numbers of children can become orphaned or separated from their families without warning. Even where children are not separated from their parents as a consequence of disasters, families may be projected into situations of serious unpredictability and can lack access to their most basic needs, such as food and shelter. In such conditions, children may become vulnerable to neglect or exploitation and experience acute stress responses and profound loss. Their education can become disrupted and their development delayed. Girls may be particularly vulnerable to sexual abuse during emergencies. For example, research in Bangladesh and Ethiopia has indicated that risk of sexual violence and forced marriages increases for girls during and after disasters (Swarup et al., 2011).

Social workers play an important role in stabilizing the conditions in children's lives in the aftermath of disasters and restoring the conditions necessary for survival, recovery, and growth. For example, Sim et al. (2013) outline the critical role played by social workers in the aftermath of the 2008 Wenchuan earthquake in China. Here, it is estimated that over 1000 social workers were involved in immediate rescue efforts and the subsequent reconstruction of damaged communities. Social workers worked to develop community

cohesion, support parenting, and offered direct counseling to children.

Nonetheless, despite some good examples of practice, there is a general lack of connection between the orthodox practices associated with humanitarian aid and those involved in protecting children. As Ridsdel and McCormick suggest:

> Despite the heightened vulnerability of girls and boys during and after conflict and disasters, currently humanitarian action does not give adequate priority to child protection and care, a situation that reflects the broader lack of attention given to this important issue in fragile and non-fragile states. (Ridsdel and McCormick, 2013: p. 3)

This is further compounded by many states in the global South that are particularly vulnerable to natural disasters having underdeveloped child protection systems. As well as calling for social work educators to develop more systematic programs of training and education in disaster social work (Sim et al., 2013), it is important to develop awareness among social workers of the child protection implications of disasters so that addressing such needs becomes a core element of the practice of social work interventions in disasters. This is particularly the case given predictions about the likely future increase in the frequency, severity, and unpredictability of disasters. For instance, Save the Children (2009) predicts that every year over the next decade 175 million children will be affected by disasters brought about by climate change across the world. Perhaps then, climate change poses the biggest challenge to child protection at the global level. In this way, Dominelli's (2012) call for the social work profession to move to a new model of 'green social work' is echoed in the need for a 'green child protection practice' that sees protection of the environment as a fundamental element in preventing and responding to the abuse of children.

Safeguarding Children Abused in Armed Conflicts

Political conflict, civil unrest, and war violates children's rights and blights their lives in many areas of the world. Political conflict and war is a human, rather than a 'natural,' disaster, created and waged in large part by adult men but with both a direct and an indirect impact on children's development, well-being, and safety (Santa Barbara, 2006). Children are frequently abducted during the course of armed conflicts and displaced from their families and communities, with some being forced into the military against the provision of Article 38 of the UN *Convention on the Rights of the Child* which limits the involvement of children under the age of 15 in any branch of the armed forces. The organization Human Rights Watch estimates that between 200 000 and 300 000 children are serving as soldiers in armed conflicts around the world, with approximately 100 000 of these children fighting in Africa alone. Children are harmed in war as a result of their direct involvement in armed conflicts as soldiers, where they are vulnerable to death and injury, disablement, and emotional trauma. Child soldiers are brutalized through being forced to kill and inflict violence on others, leading them to be both victims and perpetrators at the same time. Kimmel and Roby (2007) describe the use of child soldiers as a form of institutionalized child abuse. Other children abducted in war are sold into physical slavery or prostitution where they are subjected to rape and sexual abuse. Children who are not recruited directly into fighting are often the victims of war and political conflict. It is estimated that half of all civilians killed in wars across the world are children. Additionally, living in communities bound by war and political conflict is an emotionally disastrous condition for children, breeding insecure attachments and leading to post-traumatic stress, and grief and loss responses. Furthermore, political conflict often exposes children to religious and social intolerance and models the inappropriate use of violence. War is therefore a core and distinct form of abuse of children at the international level.

Social workers are active across the world working with communities affected by war and political conflict. For example, Bhadra (2012) describes support offered to children by professionals and community volunteers affected by communal violence in Gujarat, India. Interventions included working with children in refugee camps using activity-based and narrative approaches that allowed children to tell their own stories of the conflict and its impact, bolstering caregivers' capacities to providing nurturing environments for children, through to the development of community-oriented peace-keeping measures and school-based interventions that promote the values of tolerance and acceptance. Denov (2010) highlights the complexity of responses needed by children involved as soldiers in the decade-long civil war in Sierra Leone. She points to the important role social work should play in protecting children involved in war given the profession's commitment to the empowerment of marginalized communities, but laments that:

> Social work has been largely silent on the issue of children affected by war, and has been suspiciously absent in international debates and discussions concerning war and political violence and related global trends and processes. (Denov, 2010: p. 804)

Protecting Children from Sexual Abuse, Exploitation, and Sexual Trafficking

The sexual abuse of children was traditionally viewed as a domestic crime, perpetrated by adults on children, most often fathers on their daughters (Corby, 2000). However, increasing awareness of differing forms of sexual abuse over the last 30 years has seen social workers' understandings of the dynamics and manifestations of child sexual abuse develop radically. This includes the awareness that a significant proportion of all cases of child sexual abuse are perpetrated not by adults but by children and juveniles (Hackett et al., 2013), knowledge that a significant proportion of victims are boys (Hunter, 1990), as well as the realization that child sexual abuse is not confined to individual homes, but extends to all areas of society and to many institutions (Finkelhor et al., 1988). However, one of the most disturbing features of an increasingly globalized society is the spread of child sexual abuse and exploitation as an issue which now

transcends national boundaries. Not only does child sexual abuse occur in all societies, but the ways in which children are abused and exploited sexually often have international dimensions. These global dimensions contain three major strands: the involvement of children in prostitution; the production of abusive images of children (sometimes referred to as 'child pornography'); and the trafficking of children for the purposes of sexual abuse (US Department of Justice, 2010). The need to protect children from more traditional forms of child sexual abuse continues, but such protective efforts are compounded by, and made more complex as a result of, these globalized features.

Throughout Asia, Latin America, and the countries of the North, there have been increasing reports of growth in the number of children involved in prostitution. There are many factors underpinning this growth, and these factors differ across the world, as does the scale of the problem. In some countries, child prostitution as a phenomenon has existed for centuries, embedded in historical and cultural practices. In most countries, it is brought about as a combination of problems of poverty and socioeconomic inequality, the subordination of women and children, and the perpetuation of attitudes and values that view children as economic commodities (Heiberg, 2001). Approximately, one million children around the world are forced into prostitution every year and the total number of prostituted children could exceed 10 million (Willis and Levy, 2002). Children are frequently trafficked for the purpose of prostitution, either through well-organized criminal networks or through the actions of family members or acquaintances. In some cases, parents or children themselves are lured through offers of a better life or good employment, only to be trafficked into sexual or physical slavery in unfamiliar and alien cultures. Once involved in prostitution, children are also highly vulnerable to rape, physical violence, and a wide range of physical and mental health problems.

Child protection social workers have a role to play in preventing child trafficking through providing increased social welfare and educational opportunities for families at risk, as well as in supporting the rehabilitation of children from prostitution into safe and protective communities. A review of support for trafficked children in the United Kingdom (Franklin and Doyle, 2013) highlights how responses should be primarily dealt with as a child protection issue, with a need for some specialism to ensure that trafficked children's particular needs and circumstances are understood. However, their research into the reality of responses revealed that a lack of understanding among social workers of the immigration system and pressures relating to the immigration process acted as a barrier to supporting trafficked children, as did lack of appropriate placement provision.

Internet Abuse and Online Protection

A final illustrative example of the challenges facing child protection social workers across the world is that of the online abuse of children. The rapid development of new online technologies is a truly transnational phenomenon, even though the availability of such technologies is unevenly spread among children in industrialized as opposed to low-income societies. New information technologies are, however, transforming children's lives in many ways across the world. In terms of child protection, the Internet has significantly increased the availability of information for children about abuse and has provided routes for them to communicate their experiences and gain support. As such, new technologies can challenge the secrecy and isolation brought about by abusive experiences and exert considerable protective influence on children. The widespread use of social media platforms, such as Facebook and Twitter, has networked children in ways not possible previously. At the same time, these new technologies have also brought with them considerable additional risks to children. While new technologies have not created sex offences against children, their use has enabled a shift in the nature and scale of such abuse (Buschman et al., 2010).

Online abuse can include both virtual and actual elements. Livingstone and Haddon (2009) have described the following set of categories to understand risk and harm related to children's online activities:

online harm from content: where the child is a passive recipient of pornographic or harmful sexual content;

harm from contact: where the child is targeted as a participant by an adult or another child in activities such as sexual abuse that is photographed and then disseminated, for online grooming for sexual abuse or for bullying; and

harm from conduct: where the child actively initiates risky or abusive behaviour, for example, by creating or uploading pornographic material, physically meeting an adult encountered online, placing images of herself or himself or another young person online, downloading abusive images of children or bullying.

The range of adult behaviors that constitutes online child sexual abuse includes adults who sexually exploit their own or other children for the production of child abuse images, as well as those who download images for their own personal use. Sex offenders also use online technologies as a way of targeting children for subsequent abuse offline. This includes 'grooming' children online with the intention of abusing them sexually offline. Such approaches may involve the assumption by the adult of a false identity, in particular with the offender pretending to be a child (Whittle et al., 2013). Other forms of online child abuse include children being engaged in sexually explicit talk and being encouraged to perform sexually explicit acts on themselves or their friends.

Given these newly emerging types of sexual abuse of children, online safety is now a major child protection concern. The protection task requires a careful balance between shielding children from harm online on the one hand and their rights for information and for freedom of expression on the other. The law enforcement response in such cases is often very challenging as the behaviors involved are not contained within national borders. An initial sexual offense against a child perpetrated and filmed in the USA and then made available online may be viewed by a network of offenders worldwide. In this example, one initial contact offense may lead to thousands of online offences with the comparative anonymity of the Internet allowing the original offense to be compounded many times over. There are frequently difficulties in the identification of children abused through child abuse images of this kind and

in seeking prosecution of offenders cross nationally (Wortley and Smallbone, 2012).

The development of these new types of sexual abuse of children challenges child protection practices to develop new ways of responding, placing an emphasis on the child protection community to work together with internet service providers to improve safety measures for children and online reporting mechanisms, as well as educating children and parents about the potential dangers of the online environment. The example of the UK-based Child Exploitation and Online Protection Center, which is a world leader in this area, demonstrates the effectiveness of social workers working alongside law enforcement professionals to identify cases where children have been abused online. To play their part in the protection of children from online abuse, social workers need to engage with children in the online environment. Social work has traditionally worked at the heart of service users' social worlds to improve the quality of people's lives in the context of their environment. As children move seamlessly between online and offline social worlds, social workers need to embrace both of these dimensions of children's lives in their protective efforts. Social workers also have a core remit to work with parents to improve their parenting skills and capacities where parenting has been compromised. As such, social workers can educate parents about online risks to children, just as they have done traditionally about, for example, physical risks to children in the home and the external community.

Conclusions

Child protection is not the sole preserve of the social work profession, even though social workers play an important role in safeguarding children from harm across the world. Effective child protection practice requires the sharing of information across professional boundaries and interagency collaboration. Global issues in child protection demonstrate that this collaboration needs to extend beyond the boundaries of individual states. This is a particular challenge to social workers who mainly work within local environments. The variable development of child protection systems across countries, as well as variable levels of awareness about the diversity of issues associated with child abuse, further compounds the challenges and issues associated with international child protection.

See also: Armed Conflict and Social Work: Significance and Implications; Child Neglect; Child Protection; Child Soldiers as Victims and Actors: Lessons from Ethiopia; Disasters in Social, Cultural and Political Context; Green Social Work; Human Rights and Social Work; Indigenous Child Welfare Practice: The Canadian Case; International Social Problems: A Systems Perspective; Motherhood; Social Work Theory; Street Children: Cultural Concerns; Trafficking of Women and Children in Latin America; United Nations and its Agencies.

Bibliography

Amstrong, L., 1978. Kiss Daddy Goodnight. Dell, New York.
Ariès, P., 1962. Centuries of Childhood. Jonathan Cape, London.
Bhadra, 2012. Psychosocial support for the children affected by communal violence in Gujarat, India. International Journal of Applied Psychoanalytic Studies 9 (3), 212–232.
Blanchfield, L., 2013. The United Nations Convention on the Rights of the Child. Congressional Research Service. Available at: http://www.fas.org/sgp/crs/misc/R40484.pdf (accessed 08.10.13.).
Buschman, J., Wilcox, D., Krapohl, D., Oelrich, M., Hackett, S., 2010. Cybersex offender risk assessment: an explorative study. Journal of Sexual Aggression 16 (2), 197–209.
Clements, L., 2006. The treatment of children under the UK asylum system – Children first and foremost? Web Journal of Current Legal Issues 5.
Corby, B., 2000. Child Abuse. Towards a Knowledge Base, second ed. Open University Press, Buckingham.
Denov, M., 2010. Coping with the trauma of war: Former child soldiers in post-conflict Sierra Leone. International Social Work 53 (6), 791–806.
DHSS, 1988. Working Together: A Guide to Inter-agency Co-operation for the Protection of Children from Abuse. HMSO, London.
Dominelli, L., 2012. Green Social Work. From Environmental Crises to Environmental Justice. Polity Press, Cambridge.
Finkelhor, D., Meyers, M.W., Burns, N., 1988. Nursery Crimes: Sexual Abuse in Day Care. Sage, Beverly Hills, CA.
Franklin, A., Doyle, L., 2013. Still at Risk. A Review of Support for Trafficked Children. The Children's Society.
Garbarino, J., Guttman, E., Seeley, J.W., 1986. The Psychologically Battered Child. Jossey-Bass Publishers, California.
Gran, B., 2010. Comparing children's rights: introducing the children's rights index. International Journal of Children's Rights 18, 1–17.
Hackett, S., 2004. What Works for Children and Young People with Harmful Sexual Behaviours? Barnardo's, Barkingside.
Hackett, S., Masson, H., Balfe, M., Phillips, J., 2013. Individual, family and abuse characteristics of 700 British adolescent sexual offenders. Child Abuse Review 22 (4), 232–245.
Heiberg, T., 2001. Commodities in Stigma and Shame. An International Overview of Save the Children's Work against Child Sexual Abuse and Exploitation. Save the Children, Stockholm.
HM Government, 2013. Working Together to Safeguard Children: A Guide to Inter-agency Working to Safeguard and Promote the Welfare of Children. Department for Education, London.
Hunter, M. (Ed.), 1990. The Sexually Abused Male: Prevalence, Impact, and Treatment, vol. 1. Lexington Books, Lexington, MA.
Keenen, M., 2012. Child Sexual Abuse and the Catholic Church: Gender, Power, and Organizational Culture. Oxford University Press, New York.
Kimmel, C., Roby, J., 2007. Institutionalized child abuse. The use of child soldiers. International Social Work 50 (6), 740–754.
Livingstone, S., Haddon, L., 2009. EU Kids Online: Final Report 2009. EU Kids Online Network, London, UK.
Ortiz, I., Moreira Daniels, L., Engilbertsdóttir, S. (Eds.), 2012. Child Poverty and Inequality. New Perspectives. UNICEF, New York.
Pinheiro, P., 2006. World Report on Violence against Children. United Nations, Geneva.
Puffett, N., Wednesday 16 January 2013. Government Admits Placing Young Offenders in Adult Jails. Children and Young People Now. http://www.cypnow.co.uk/cyp/news/1075953/government-admits-placing-offenders-adult-jails (accessed 08.10.13.).
Ridsdel, J., McCormick, C., 2013. Protect My Future. The Links between Child Protection and Disasters, Conflict and Fragility. Available at: http://www.terredeshommes.org/wp-content/uploads/2013/06/20130315_child_protection_and_disaster_conflict_and_fragility.pdf (accessed 08.10.13.).
Santa Barbara, J., 2006. Impact of war on children and imperative to end war. Croatian Medical Journal 47 (6), 891–894.
Save the Children, 2009. Children and Climate Change. Save the Children, London.
Sim, T., Yuen-Tsang, A., Quen, C., Dong, Q., 2013. Rising to the occasion: disaster social work in China. International Social Work (published online first May 2013).
Swarup, A., Dankelman, I., Ahluwalia, K., Hawrylshyn, K., 2011. Weathering the Storm: Adolescent Girls and Climate Change. Plan International, Woking, UK.
United Nations, 1989. Convention on the Rights of the Child. United Nations, New York.

UNICEF, 2006. What Is Child Protection? The United Nations Children's Fund. Available at: www.unicef.org/protection/files/What_is_Child_Protection.pdf (accessed 08.10.13.).

US Department of Justice, July 2010. The National Strategy for Child Exploitation Prevention and Interdiction. A Report to Congress. Available at: http://www.justice.gov/psc/docs/natstrategyreport.pdf (accessed 08.10.13.).

Whittle, H.C., Hamilton-Giachritsis, C., Beech, A., Collings, G., 2013. A review of young people's vulnerabilities to online grooming. Aggression and Violent Behavior 18, 62–70.

Willis, B., Levy, B., 2002. Child prostitution: global health burden, research needs, and interventions. The Lancet 359 (9315), 1417–1422.

World Health Organisation, 2013. Child Maltreatment. Available at: http://www.who.int/topics/child_abuse/en/ (accessed 08.10.13.).

Wortley, R., Smallbone, S., 2012. Child Pornography on the Internet. US Department of Justice/Center for Problem Oriented Policing. Available at: http://www.cops.usdoj.gov/Publications/e04062000.pdf (accessed 08.10.13.).

Relevant Websites

ceop.police.uk – Child Exploitation and Online Protection Centre.
www.nspcc.org.uk/inform – NSPCC Child Protection Resources.
www.unicef.org/crc – UN Convention on the Rights of the Child.
www.unicef.org/protection – UNICEF Child Protection Resources.

Child Soldiers

Brandon A Kohrt, Duke Global Health Institute, Duke University, Durham, NC, USA
Sauharda Rai, Transcultural Psychosocial Organization (TPO) Nepal, Kathmandu, Nepal
Sujen M Maharjan, Tribhuvan University, Kirtipur, Nepal

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Child soldiers can be any person under 18 associated with an armed group and fulfilling many roles that may or may not involve combat. An estimated 300 000 children across the globe are members of state militaries and other armed groups. Although armed groups used children throughout history, the widespread availability of small arms has made it physically possible for children to participate in a more lethal combat, thereby increasing their participation value. Children become involved in armed groups through abduction, threats, pressure, and manipulation. Voluntary association also occurs in the context of poverty, gender and ethnic discrimination, and state-sponsored violence. Among published studies of child soldiers, high levels of exposure to violence are reported. Moreover, post-traumatic stress disorder may affect more than half of all child soldiers at some point after association with armed groups. Postconflict factors such as educational and economic deprivation, stigmatization, and discrimination have been associated with adverse outcomes. Reintegration and rehabilitation is particularly difficult for girl soldiers. Potential protective factors for successful reintegration include community sensitization, cleansing rituals, transitional periods in interim care centers, religious support, psychosocial counseling, family mediation, and skills-based vocational training. Policy and advocacy is needed to provide educational and employment alternatives to children in place of joining armed groups; and, regulation and control of small arms trade is needed to reduce recruitment of children by armed groups.

One of the most pressing issues in global conflict is the exploitation of children by armed groups. An estimated 300 000 children across the globe are members of state militaries and other armed groups (UNICEF, 2003). The military recruitment of children has taken place in 86 countries and territories worldwide with approximately half of all child soldiers in African countries (Coalition to Stop the Use of Child Soldiers, 2008). Reflecting this pattern, the vast majority of published research on child soldiers comes from studies in Sierra Leone, Uganda, and Mozambique. Child soldiers are involved in a range of political conflicts such as formally recognized wars, terrorist activities, and ethnic conflict. Although armed groups have used children throughout history, the widespread availability of small arms has made it physically possible for children to participate in a more lethal manner and thus be of greater value to armed groups (Wessells, 2006).

The popular image, often used by organizations to raise funds to support intervention programs for former child soldiers, depicts rogue armed groups violently tearing children away from idyllic family settings and those children as faultless passive victim, psychologically devastated, and irreversibly damaged as a result of conflict. The story continues that these haphazard militias shatter children's lives through the dehumanizing experiences of war, drugs, and crime. These children are also exposed to traumatic experiences and are often forced to commit atrocities themselves. Child soldiers are thus imagined as severely disabled because of the effects of psychological trauma. The final goal for recovery and rehabilitation is the return to and reunification with the child soldier's family.

Although these portrayals do reflect actual cases, the situation of child soldiers is far more varied and complex (Betancourt et al., 2008a; Wessells, 2006). UNICEF and other groups working with child soldiers recognize that children associated with armed groups may assume a range of roles and are not always combatants: the 2007 Paris Principles refer to child soldiers as "children associated with armed forces or armed groups" (CAAFAG) meeting the following criteria:

> any person below 18 years of age who is or who has been recruited or used by an armed force or armed group in any capacity, including but not limited to children, boys, and girls used as fighters, cooks, porters, messengers, spies, or for sexual purposes. It does not only refer to a child who is taking or has taken a direct part in hostilities.
>
> UNICEF (2007): p. 7

Moreover, the reasons that children become associated with armed groups vary considerably. Contexts that promote recruiting children into armed groups are the result of larger national and international processes that produce local vulnerability. Poverty, gender and ethnic discrimination, and legacies of state-sponsored violence create circumstances in which children voluntarily join armed groups. Reintegration is also a complex process that varies widely across different conflicts and can be very different for children in the same conflict. While it might be assumed that removing a child from an armed group is the best approach, it is not a panacea to assure psychosocial well-being. The very focus of *re*integration programs operates with an assumption that former child soldiers were previously *integrated* into their communities. In reality, some child soldiers join armed groups because they feel excluded from society. Reintegration not only means returning back home but also building and sustaining positive relationship, regaining identity, and managing their intrusive thoughts and reveries amidst the stigma they face as former child soldiers (Boothby and Thomson, 2013). Moreover, the tremendous variation in exposure to traumatic events during involvement with armed groups and in resilience among child soldiers throws into question whether psychological trauma is a universal response (Wessells, 2006). Lastly, local definitions

of 'children' vary within and between cultural groups, such that individuals identified as 'child soldiers' by humanitarian groups may locally be considered adults (Kohrt and Maharjan, 2009). This diversity must be considered when understanding the experiences of child soldiers to best inform interventions to prevent recruitment of children into armed groups and to facilitate positive reintegration after leaving armed groups.

Ethical Conduct of Research and Interventions with Former Child Soldiers

Before discussing research and interventions, it is crucial to consider the ethical conduct when working with current and former child soldiers. Child soldiers are victims of exploitation through exposure to military association. Research and intervention should not be another form of exploitation. In conflict settings, research and intervention has the potential to violate the 'do no harm' ethic (Betancourt, 2011; Kohrt et al., 2010a; Wessells, 2009).

Harm can manifest as threats to safety and discrimination of former child soldiers, their families, and their communities. Research and intervention may expose child soldiers who have tried to hide their association with an armed group, thus placing them and their families in danger of revenge by other soldiers or aggrieved civilians. To address this potential harm, the best practices of research and service provision have advocated for working with the broader community, not exclusively focusing on child soldiers. Additionally, participation in research and interventions should not automatically identify a child as a former combatant publicly. A solution to this is the development of research and intervention programs that target a range of vulnerable children, rather than only child soldiers. For example, programs can also work with children with family members who have been targets of enforced disappearance, children whose family members have been tortured, and children in families who have been forcibly displaced. Research also risks stigmatization of former child soldiers (Annan et al., 2011; Betancourt et al., 2010a; Kohrt et al., 2008). Mental health problems carry a strong stigma in most parts of the world (Sartorius et al., 2005). Focusing on normative language related to idioms of distress can give children a therapeutic opportunity to share feelings and emotionally support other children in a nonstigmatizing atmosphere (Karki et al., 2009).

Participatory research is one approach that can help address some of the potential harms when research and intervention programs are dictated by outside actors. Participatory research provides a strong connection between research and intervention salient to the local community. Child soldiers identified psychosocial problems affecting them and processes to address these needs (Karki et al., 2009). The children themselves developed indicators to evaluate the effectiveness of interventions.

Child Soldier Experience in Life-History Context

Understanding the experience of child soldiers requires that the soldier experience be considered in the broader context of life history. Former child soldiers have a range of potential risk and resilience factors related to (1) *presoldier and noncombatant experiences*, (2) *experiences during association with armed groups*, and (3) *postcombatant reintegration experience* (see **Figure 1**).

(1) Presoldier and noncombatant experience

Exposure to trauma and other stressors as a child soldier cannot be considered the sole determinant of psychosocial well-being for former child soldiers. If all children's problems are attributed exclusively to war, interventionists and policy makers risk misapplication of care in resource poor settings (Derluyn et al., 2013). More detailed knowledge of the impact of political violence versus other chronic social problems can lead to better-informed interventions in postconflict settings with scarce resources (IASC, 2007). Children's experiences before becoming a child soldier and the experience of stressors outside of their combatant roles need to be taken into account. One of the important issues is to understand that the threats to child well-being are structural conditions in the presoldier context. Contexts that promote recruiting children into armed groups are the result of larger national and international processes that produce local vulnerability. Poverty, gender and ethnic discrimination, and legacies of state-sponsored violence create circumstances in which children voluntarily join armed groups (Kohrt et al., 2010b). Poverty, especially, is an impediment to well-being of children. Poverty creates risks for violence and exploitation of children.

Countries where children are exploited through involvement in armed group are also likely to be countries where children are exploited for labor in other domains (United Nations, 2005). Children are relied upon heavily for labor in most subsistence agricultural settings. Children are also trafficked to urban centers and other countries for manual labor and commercial sex work. Children have been used for their labor in garment factories, brick manufacturing plants, and mines where they are vulnerable to both mechanical and respiratory injuries.

Another preconscription issue is the lack of education in many settings vulnerable to conflict. The opportunity for education is tied strongly to economic status and potential to avoid conscription (Betancourt et al., 2008b). The quality of education is tied strongly to economic status and potential to avoid conscription. In Mozambique, school drop-out students were recruited into armed forces with promises of better future (Boothby and Thomson, 2013).

Another major problem for children worldwide is abuse and neglect. Due to high rates of alcoholism, caregivers expose children to domestic violence and abuse. In addition to abuse, children suffer from neglect: because of endemic poverty, children are often unable to have their basic needs met. Moreover, in conflict-affected countries in Africa and Asia, there are few laws to protect children from child abuse and even fewer avenues for enforcement in most settings where children have been conscripted into armed groups (United Nations, 2005). Abuse and neglect creates vulnerability to recruitment. Child marriage is also child abuse. Child marriage was a common historical practice and still occurs in many settings where children have been used as combatants. Legacies of child marriage make girls vulnerable to being abducted as 'bush wives' for

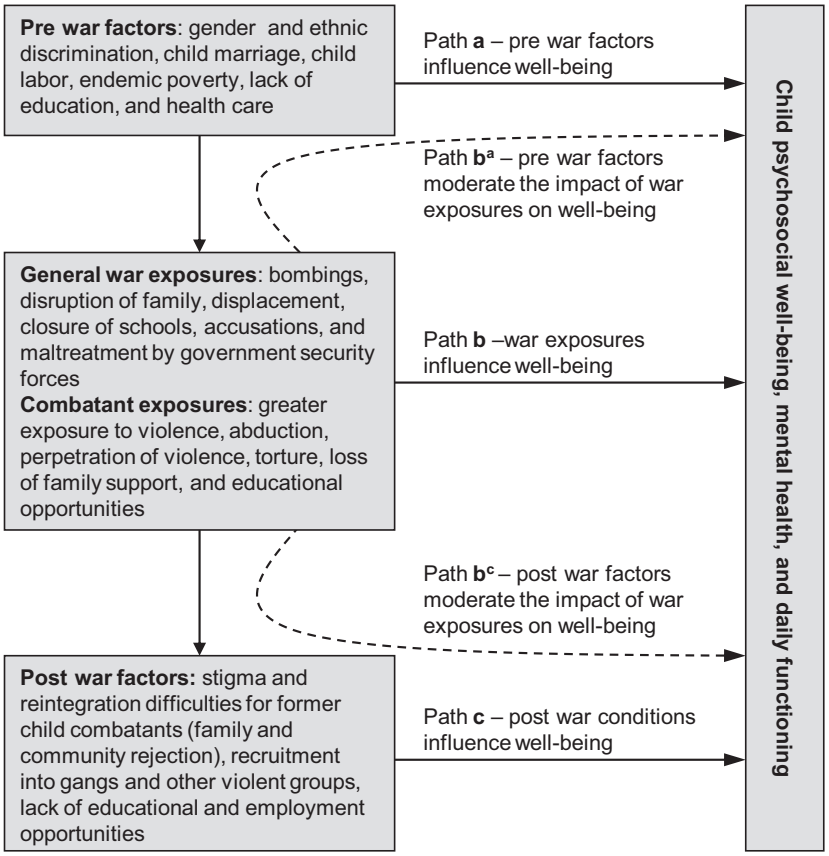

Figure 1 Framework for life-history approach to risk factors for psychosocial problems among former child soldiers.

older men in military groups (McKay and Mazurana, 2004). In Nepal, fleeing from child marriage was a reason girls cited for joining armed groups (Kohrt et al., 2010b).

Gender discrimination is a profound factor that increases vulnerability to a range of psychosocial and mental health problems. Countries with a legacy of exploiting children as combatants score on the lower end of the Gender Development Index and Gender Empowerment Index. This pattern reflects the difference between women and men in the burden of labor; women are more poorly compensated than men. Lack of empowerment of girls may make them vulnerable to recruitment into armed groups as sexual slaves (Stark, 2006). In some settings, lack of opportunities for women may lead girls to join armed groups as a form of empowerment (Kohrt et al., 2010b).

(2) Experience of child soldiers during association with armed groups

The first issue to consider regarding association with armed groups is how a child was conscripted, as this may be a major determinant of subsequent well-being. Children become involved in armed groups through abduction, threats, pressure, manipulation and voluntary association. It is helpful to consider a division of risk factors for conscription into 'push factors' and 'pull factors' (Somasundaram, 2002). 'Push factors' are those factors that place an individual at risk of recruitment. Poor economic conditions, unemployment, and inability to meet basic needs in the community were reasons for joining. Children may join because of lack of opportunities in their communities and family problems. Girls may join to escape abusive marriages or may join prior to entering forced marriages. 'Pull factors' are aspects of the armed group that include both methods of recruitment, enticement, and promises made to children.

The most extreme 'pull factor' is abduction. Across the studies included in a recent systematic review (Betancourt et al., 2013a), abduction was the main mechanism by which children reported becoming involved with an armed group. Children in Mozambique were forcibly be made to join armed groups through kidnapping and abduction where they go through a series of cohesive and inhuman training and indoctrination before being integrated into armed forces; the children were kidnapped, separated gender-wise, and boys forcibly sent to armed forces while girls were made to serve the soldiers (Boothby and Thomson, 2013). In a study in Uganda among 330 sample child soldiers 99.7% were forcefully abducted (Klasen et al., 2013). However, in a Nepal study, nearly half of study participants joined without abduction (Kohrt et al., 2008), as did 18 out of 19 former child soldiers in the study from Sri Lanka (De Silva et al., 2001). In a study from El Salvador of 293 former child soldiers, 73.7% of study participants reported joining voluntarily and 3.4% reported

forced conscription (Santacruz and Arana, 2002). In these studies, ideological reasons for joining voluntarily included revenge, martyrdom, and 'desire to defend the motherland.' Other subjects reported joining out of fear of abduction.

Across the published studies of child soldiers, high levels of exposure to violence during association were reported. For example, several studies found that over 70% of child soldiers were severely beaten by armed forces, with similar rates reported by boys and girls (Amone-P'Olak, 2005; Amone-P'Olak et al., 2007; Bayer et al., 2007; Betancourt et al., 2011; Klasen et al., 2010; Kohrt et al., 2008). In an Ugandan study, more than half of the 330 sample child soldiers reported being directly involved in fighting, looting property, and injuring and killing other people (Klasen et al., 2013).

High levels of sexual abuse among girls were observed in samples from Sierra Leone (44%; Betancourt et al., 2011) and Uganda (30%; Klasen et al., 2010), as well as a mixed sample from Uganda and the DRC (57%; Bayer et al., 2007). Within the armed groups in Mozambique, girls were made to serve the commanders and soldiers as wives, servants and cargo carriers (Boothby and Thomson, 2013). Many girl child soldiers from African countries have reported sexual violence. This contrasts with Asian countries including Sri Lanka, Nepal, and the Philippines where intimate sexual or personal relationships without the female's consent and commander's approval were forbidden and would be punished. However, girls in the Philippines and Colombia have agreed to sexual relationships within groups if they knew it brought them some kind of benefits (Stevens, 2014). Boys were also victims of sexual abuse in these settings, although only Amone-P'Olak (2005) found higher rates of sexual abuse among boys (81 vs 72%) as compared to girls.

(3) Reintegration and postcombatant experience

While it might be assumed that removing a child from an armed group is the best approach, it does not assure psychosocial well-being. Increasingly, researchers have shown that the postcombatant reintegration experience is one of the main determinants of mental health, psychosocial well-being, educational outcomes, and livelihood (Annan et al., 2013).

Former child soldiers struggle with community stigma upon return home (Annan et al., 2009; Burman and McKay, 2007; Denov and Maclure, 2007; Denov, 2010; Shakya, 2010; Stark, 2006). For many adults, child soldiers are seen as threatening and morally corrupted (Boyden, 2003). In Sierra Leone stigma and perceived discrimination for being a former child soldier was significantly associated with worsening of posttraumatic stress symptoms over time (Betancourt et al., 2013b). In El Salvador, participants who experienced stigma scored higher on a psychological impact scale, as did youth who experienced feelings of disappointment upon return home (Santacruz and Arana, 2002). Longitudinal data from Sierra Leone indicated that stigma predicted higher levels of hostility/externalizing problems (such as behavioral problems characteristic of attention deficit disorder, oppositional defiant disorder, and conduct disorder) as well as deficits in prosocial behaviors over time, even after adjusting for war exposures (Betancourt et al., 2010a,c). In Uganda, abductees and non-abductees experienced similar levels of family and community rejection, while non-abductees were more likely to engage in prosocial behaviors such as participation in politics and community activities (Blattman, 2009).

Stigma may be connected with local idioms of distress associated with perpetration of violence. A study in Uganda found that the failure of reintegration was associated with widespread community resistance due to perceived unaccountability for the crimes committed and persecution from revenging spirits (*cen*) of those who have been killed (Akello et al., 2006). Community members perceived that minors committed war crimes and were free from punishment for their acts. Regarding *cen*, community members and child soldiers themselves perceived that child soldiers were being punished by spirits of persons they had killed. Therefore, the community avoided these child soldiers because of fear of their violent pasts and their affliction with spiritual afflictions related to their misdeeds.

Across settings, girls appear to suffer greater stigma and negative experiences during reintegration. There are estimated 120 000 female child soldiers worldwide and reintegration is particularly difficult for them (Stevens, 2014). Studies that stratified community acceptance by gender indicated differences in the ways families and communities treated returning boys compared to girls. In Sierra Leone, female former child soldiers experienced lower rates of acceptance as compared to their male counterparts (Betancourt et al., 2010b) and were commonly described as being akin to 'prostitutes' (Betancourt et al., 2008b). Girls in one study from Uganda also reported experiencing less community support (Annan et al., 2011). The trend continues even in Nepal where a study reported girl child soldiers experiencing gender-based community stigma related to local perceptions that their Hindu purity had been violated (Kohrt et al., 2010b). However, in a longitudinal study among child soldiers of Sierra Leone, girls demonstrated higher rates of improvement across the 4-year follow-up period (Betancourt et al., 2013b).

In addition to stigma and discrimination, postconflict factors such as educational and economic deprivation are associated with adverse outcomes (Betancourt et al., 2010a; Blattman and Annan, 2010; Wessells, 2009). Most studies have found that former child soldiers felt ill-equipped to contribute to their family economy upon return to civilian life (Annan et al., 2006; Boothby et al., 2006). In Uganda, access to resources have been identified as an important factor to help the child soldiers break away from their child soldier experiences, cope with bad feelings and mental health problems including nightmare, suicidal thoughts, and hallucination (Vindevogel et al., 2013a).

Reintegration means building positive relationships and coping with their intrusive feelings and stigma they face as former child soldiers (Boothby and Thomson, 2013). Family, community and peer support have been identified as important factors facilitating successful reintegration and psychosocial well-being. Family and community acceptance along with religious and spiritual cleansing ceremonies are important factors that facilitated successful social

reintegration of former child soldiers of Mozambique (Boothby and Thomson, 2013). In Sierra Leone, higher levels of family acceptance were linked to lower average levels of emotional distress among participants (Betancourt et al., 2010a,c). In this setting, caregiver–child open communication about traumatic experiences, and perpetration of violence in particular, might have a beneficial effect on positive adjustment outcomes (Borisova et al., 2013). Caregiver's adequate knowledge of direct violence was predictive of reported levels of depression among adolescents and of positive adjustment outcomes like prosocial attitudes, confidence, and self-agency. Supportive parenting was associated with better psychosocial adjustment in one study of Ugandan child soldiers (Derluyn et al., 2004). In Nepal, peer-related support was a strong predictor of psychosocial well-being: former child soldiers having strong peer support experienced increased hope and fewer post-traumatic stress disorder (PTSD) symptoms as well as reported better performance in daily functioning (Morley and Kohrt, 2013).

In Uganda, access to resources was linked with the child soldiers overcoming their current challenges, belonging to their returned environment, and a sense of being whom they wanted to be (Vindevogel et al., 2013a). Returning to and staying in school was associated with higher levels of confidence and prosocial behaviors among youth in a Sierra Leone study (Betancourt et al., 2010b).

Reintegration Programs, Traditional Healing, and Other Interventions

Facilitating successful reintegration of former child soldiers is one of the goals of the United Nations disarmament, demobilization, and reintegration (DDR) framework. Reintegration programs for child soldiers typically comprise economic and education support packages such as vocational training, apprenticeships, micro-finance loans, and formal or informal schooling. The intention of these programs is to prevent or mitigate the social, economic, and psychological sequelae of participation in an armed group. DDR programs should also involve caregivers and families in tangible ways to promote psychosocial well-being of reintegrated child soldiers (Borisova et al., 2013).

Potential protective factors for successful reintegration include community sensitization, cleansing rituals, transitional periods in interim care centers, religious support, psychosocial counseling, family mediation, and skills-based vocational training (Boothby, 2006; Kryger and Lindgren, 2011; Williamson, 2006). Attending school and training programs is considered critical in helping war-affected youth attaining a sense of normalcy and safety in their everyday lives while also increasing their future employment opportunities (Betancourt and Khan, 2008). In a study in Uganda, former child soldiers who were given the opportunity to continue their education scored lower on assessments of depression, compared to those who directly entered vocational training (Ovuga et al., 2008). Involvement of family is also crucial for effective interventions. In Sierra Leone inclusion of family-based elements resulted into greater impact of an intervention (Betancourt et al., 2013b).

A study of multisectorial reintegration assistance in Burundi found that former child soldiers who were socially integrated within communities had high work/employment rates, literacy rates above national average, and no differences in present functioning and mental health compared to never-recruited peers (Jordans et al., 2012). While coaching, on-the-job and vocational training and provision of material kits were associated with improved socioeconomic reintegration trajectories, a sense of satisfaction with received services is most strongly and recurrently associated with better outcomes. Community connection and a sense of future were also identified as protective factors among Colombian former child soldiers (Cortes and Buchanan, 2007). In the study, six Colombian child soldiers identified the factors that helped them overcome the effects of war. Feeling connected to the community and hope for the future were key factors. This was reflected in Nepal as well were having a sense of foreshortened future (i.e., thinking that one will not live as long as one's peers) was one of the strongest predictors of poor functioning in school, work, and social activities among children exposed to political violence (Kohrt et al., 2011).

Given the high burden of stigma and psychosocial problems among former child soldiers, using traditional healing rituals is an important trend in humanitarian interventions, particularly as combined with Western mental health services. Psychosocial practitioners, as well as anthropologists and other social scientists, have increasingly questioned the appropriateness and effectiveness of Western psychiatric therapies for survivors of mass violence, especially for children (Bracken and Petty, 1998; Psychosocial Working Group, 2003; Summerfield, 1998). In a Mozambique study, these traditional ceremonies played a vital role in rebuilding community trust and cohesion, as these ceremonies were believed to clean the child soldiers from their 'contaminated' atrocities of war (Boothby and Thomson, 2013).

In a study of former male child soldiers in Mozambique, traditional cleansing rituals were seen as promoting community acceptance (Boothby, 2006). In Sierra Leone, communities and families employed 'rituals of welcome' to incorporate returned girl soldiers into acceptable social positions (McKay and Mazurana, 2004). Honwana (2006) describes traditional healers in Angola conducting communal cleansing rituals for child soldiers to purify them of their exposure to killing. In Mozambique, traditional healers facilitated reintegrating ex-combatants by 'taking the war out of the people,' (Nordstrom, 1997: p. 146).

Reunification rituals or rituals of reintegration, however, may come at a cost to former child soldiers. For example, child soldiers are threatening to adults because they do not fall into socially expected roles of submission to adult authority (Boyden, 2003). Rituals may restore the expected socially submissive role of children, but this may represent a lost opportunity for children to gain a greater voice in social processes. The issue of girl soldiers exemplifies this challenge. Traditional rituals often reinforce gender discrimination by promoting the status of '(older) males' and "threaten the human security and well-being of women and girls" (Denov,

2007). Furthermore, Denov adds, "when assessing whether 'culture is always right', one cannot discount … the reality and implications of gendered exclusionary practices." Girl soldiers are particularly threatening for patriarchal societies. Therefore, adults use ritual to disempower girls and return them to socially acceptable subjugated roles. This raises the question of whether restorative rituals are the best practice for reintegration of child soldiers, particularly for girl soldiers. In Nepal, girls soldiers refused to participate in rituals during reintegration because they were perceived as a form of subjugation and antithetical to the reason they originally joined armed groups (Kohrt, 2015).

The peer-reviewed literature on former child soldiers includes few systematic studies of interventions to inform major policy decisions. In Côte d'Ivoire, 345 former child soldiers were enrolled in trauma-focused counseling supplemented by therapeutic workshops, literacy workshops, and resocialization exercises (Bissouma et al., 2010); in a pilot study in Liberia, 130 former child soldiers participated in a brief trauma-focused intervention (Gregory and Embrey, 2009). Two intervention trials were conducted in northern Uganda: Bolton and colleagues conducted a randomized controlled trial of Interpersonal Therapy with 314 war-affected youth (Betancourt et al., 2012; Bolton et al., 2007). Ertl et al. (2011) used three arms to compare Narrative Exposure Therapy, an academic catch-up and supportive counseling group, and a waitlist control condition among 85 former child soldiers, with significantly greater reduction of PTSD symptoms in the narrative group. In a randomized controlled trial in the Eastern DRC, Narrative Exposure Therapy for Forensic Offender Rehabilitation (FORNET) was offered to 15 ex-combatants with the goal of reducing traumatic stress and appetitive aggression compared to 'treatment as usual.' The results showed that FORNET helped reducing the severity of PTSD symptoms but not the levels of appetitive aggression (Hermenau et al., 2013). More research is needed to identify effective common elements in child soldier interventions as well as context specific elements for successful promotion of well-being.

Mental Health of Former Child Soldiers

It is difficult to make generalized conclusions about prevalence of mental health problems among former child soldiers, and it is especially difficult to isolate the effect of soldiering trauma on mental health among the range of other stressors that child soldiers experience such as poverty, lack of education, and stigma and discrimination. Variations in prevalence rates of mental health problems are in part attributable to problems in methodology, sampling, and instrumentation. There is notable heterogeneity in PTSD prevalence among studies with similar research designs. For example, in studies of formerly abducted youth in rehabilitation centers in northern Uganda, documented PTSD rates ranged from 99% (Amone-P'Olak, 2005; Amone-P'Olak et al., 2007) to 27% (Okello et al., 2007). Also in northern Uganda, one study investigating PTSD symptoms among 71 former child soldiers determined that PTSD symptoms were present in 97% of the sample (Derluyn et al., 2004).

It is now being suggested that research should extend beyond PTSD to tackle what is currently known about perpetration-induced trauma (Klasen et al., 2013). Another measure called Development Trauma Disorder (DTD) has been developed. In an Uganda study 80% of the sample former child soldiers met criteria for DTD while only 33% met criteria for PTSD, which supports that classic PTSD symptoms do not capture the overall trauma affect (Klasen et al., 2013).

Children may continue to suffer from psychological problems years after reintegration. In Uganda the effect of war-related trauma was seen even 6 years after the end of the war reflecting the adverse and long-term mental health effect on former child soldiers (Amone-P'Olak et al., 2014). Postwar environment are a key determinant of continued mental health problems in former child soldiers (Amone-P'Olak et al., 2014).

One question is whether child soldiering has a more negative impact on girls compared to boys. Female gender was a predictor of poorer mental health in studies in Sierra Leone (Betancourt et al., 2011) and Nepal (Kohrt et al., 2008). In the Sierra Leone research, more females scored within the clinical range for anxiety (80% girls vs 52% of boys) as well as depression (72% girls vs 55% of boys). When controlling for war exposures, female gender was significantly associated with lower levels of confidence and prosocial behaviors over time (Betancourt et al., 2010b, 2011). In the Nepal sample, there was a significant interaction between gender and child soldier status: girl soldiers had six times greater odds of having PTSD compared to those never conscripted, whereas boy soldiers had nearly three times greater odds of having PTSD compared with never-conscripted boys (Kohrt et al., 2008).

Comparing samples of children in the same conflict who were not conscripted in armed groups to child soldiers is an approach to identify the unique contribution of child soldiering on mental health. In studies with comparison groups, PTSD prevalence was greater among former child soldiers compared with never-conscripted children (Kohrt et al., 2008; MacMullin and Loughry, 2004; Okello et al., 2007; Vindevogel et al., 2013b). In Nepal, this distinction was maintained even after controlling for exposure to violence (Kohrt et al., 2008). By contrast, studies from northern Uganda found little difference in psychosocial distress levels between former child soldiers and comparison groups (Betancourt et al., 2013b).

Future Directions for Research and Intervention

With the growing knowledge about child soldiers, there are new directions in prevention and reintegration activities. Jordans and colleagues (Jordans et al., 2012) argue that successful reintegration is about the *equality* in opportunities, participation, well-being, and social functioning of former child soldiers compared to those of their never-conscripted peers. Reintegration also means building positive relationship in face of stigma they face as former child soldiers (Boothby and Thomson, 2013). Whether or not a former child soldier has found employment, is married, received education, or experiences few mental health problems are indicators of

long-term reintegration primarily vis-à-vis the average of the population at large or their peers.

There is a need to consider long-term development of mental health care services rather than services that attend to acute relief efforts. World regions affected by war and armed conflict often lack a standing and effective mental health care system, especially for children. There is a need to consider how the energy, human resources, and finances invested during the postconflict period can be extended to services that will last beyond the acute bolus of support. Amone-P'Olak et al. (2014) argues that policies and interventions to reduce long-term mental ill-health should address both postwar environment and mental health care as complementary. Strengthening the health systems in war-affected regions to respond to the mental health situation through both psychosocial interventions and mental health treatment should be prioritized (Amone-P'Olak et al., 2014). Moreover, in postconflict settings, mental health and psychosocial problems after war may be as related to chronic structural violence factors as to war-related exposures. Interventions should strongly consider chronic social problems as well as war trauma.

To prevent recruitment of child soldiers, interventions at the national level are needed to promote the inclusion of children's issues and voices. Through dialogue with policy makers, advocacy is needed for increased child protection to tackle abuses such as forced child marriage and child conscription in armed groups, as well as further capacity building of local institutions concerned with human rights protection, which currently often take place in a culture of impunity. Further initiatives are required whereby children's voices could be incorporated in planning, designing and monitoring, and evaluation of child protection programs. At the macro-social level, there also needs to be interventions to address the issues of poverty, ethnic discrimination, gender discrimination, protection of human rights, and decentralization of power.

At a more fundamental level, structural violence including poverty, inequality, and militarization are key issues. Policy and advocacy is needed to reduce the small arms trade that makes child soldiers more appealing, and to reduce the provision of large arms to militaries. Moreover, there should be continued international political pressure to encourage more representative inclusion of women, and ethnic minorities in governments in conflict-affected countries to assure that their needs are met during the rebuilding of society. International political pressure is needed to ensure a transitional justice process that allows for the involvement and participation of children and youth in political processes in postconflict societies. A clear mechanism for children to express their political rights, such as in the Truth and Reconciliation Bill, and citizenship is needed to supplant the system of child exploitation by political groups. International pressure and monitoring bodies, such as the 1612 U.N. Task Force (United Nations, 2005), can foster enactment of child-centered legislation and assure that political and legal processes meet international standards.

See also: Armed Conflict and Social Work: Significance and Implications; Cross-Cultural Training in Industrial and Organizational Psychology; Culturally Appropriate Interventions in Social Work; Dehumanization; Depression; Gangs, Sociology of; Indigenous Child Welfare Practice: The Canadian Case; Mental Health; Military and Disaster Psychiatry; Post Traumatic Amnesia; Post-Traumatic Stress Disorder; Post-Traumatic Stress in Social Work; Risk Assessment in Social Work; Stigma, Social Psychology of; Stress in Adolescence: Effects on Development; Violence as a Public Health Problem; Violence in Anthropology; Youth Gangs.

Bibliography

Akello, G., Richters, A., Reis, R., 2006. Reintegration of former child soldiers in northern Uganda: coming to terms with children's agency and accountability. Intervention 4 (3), 229–243.

Amone-P'Olak, K., 2005. Psychological impact of war and sexual abuse on adolescent girls in northern Uganda. Intervention: International Journal of Mental Health, Psychosocial Work & Counselling in Areas of Armed Conflict 3 (1), 33–45.

Amone-P'Olak, K., Stochl, J., Ovuga, E., Abbott, R., Meiser-Stedman, R., Croudace, T.J., Jones, P.B., 2014. Postwar environment and long-term mental health problems in former child soldiers in northern Uganda: the WAYS study. Journal of Epidemiology and Community Health. http://dx.doi.org/10.1136/jech-2013-203042.

Amone-P'Olak, K., Garnefski, N., Kraaij, V., 2007. The impact of war experiences and physical abuse on formerly abducted boys in northern Uganda. South African Psychiatry Review 10, 76–82.

Annan, J., Blattman, C., Horton, R., 2006. The State of Youth and Youth Protection in Northern Uganda: Findings from the Survey of War Affected Youth. UNICEF, New York.

Annan, J., Brier, M., Aryemo, F., 2009. From "Rebel" to "Returnee": daily life and reintegration for young soldiers in northern Uganda. Journal of Adolescent Research 24 (6), 639–667.

Annan, J., Blattman, C., Mazurana, D., Carlson, K., 2011. Civil war, reintegration, and gender in northern Uganda. Journal of Conflict Resolution 55 (6), 877–908.

Annan, J., Green, E.P., Brier, M., 2013. Promoting recovery after war in northern Uganda: reducing daily stressors by alleviating poverty. Journal of Aggression, Maltreatment & Trauma 22 (8), 849–868.

Bayer, C.P., Klasen, F., Adam, H., 2007. Association of trauma and PTSD symptoms with openness to reconciliation and feelings of revenge among former Ugandan and Congolese child soldiers. Journal of the American Medical Association 298 (5), 555–559.

Betancourt, T.S., Borisova, I., Rubin-Smith, J., Gingerich, T., Williams, T., Agnew-Blais, J., 2008a. Psychosocial Adjustment and Social Reintegration of Children Associated with Armed Forces and Armed Groups: The State of the Field and Future Directions. Psychology Beyond Borders and the Francois-Xavier Bagnoud Center for Health and Human Rights/Harvard School of Public Health, Cambridge, Massachusetts, p. 97.

Betancourt, T.S., Khan, K.T., 2008. The mental health of children affected by armed conflict: protective processes and pathways to resilience. International Review of Psychiatry 20 (3), 317–328.

Betancourt, T.S., Simmons, S., Borisova, I., Brewer, S.E., Iweala, U., de la Soudiere, M., 2008b. High hopes, grim reality: reintegration and the education of former child soldiers in Sierra Leone. Comparative Education Review 52 (4), 565–587.

Betancourt, T.S., Agnew-Blais, J., Gilman, S.E., Williams, D.R., Ellis, B.H., 2010a. Past horrors, present struggles: the role of stigma in the association between war experiences and psychosocial adjustment among former child soldiers in Sierra Leone. Social Science & Medicine 70 (1), 17–26.

Betancourt, T.S., Borisova, I.I., Brennan, R.B., Williams, T.P., Whitfield, T.H., de la Soudiere, M., Gilman, S.E., 2010b. Sierra Leone's former child soldiers: a follow-up study of psychosocial adjustment and community reintegration. Child Development 81 (4), 1077–1095.

Betancourt, T.S., Brennan, R.T., Rubin-Smith, J., Fitzmaurice, G.M., Gilman, S.E., 2010c. Sierra Leone's former child soldiers: a longitudinal study of risk, protective factors, and mental health. Journal of the American Academy of Child & Adolescent Psychiatry 49 (6), 606–615.

Betancourt, T.S., 2011. Attending to the mental health of war-affected children: the need for longitudinal and developmental research perspectives. Journal of the American Academy of Child & Adolescent Psychiatry 50 (4), 323–325.

Betancourt, T.S., Borisova, I., de la Soudière, M., Williamson, J., 2011. Sierra Leone's child soldiers: war exposures and mental health problems by gender. Journal of Adolescent Health 49 (1), 21–28.

Betancourt, T.S., Newnham, E.A., Brennan, R.T., Verdeli, H., Borisova, I., Neugebauer, R., Bolton, P., 2012. Moderators of treatment effectiveness for war-affected youth with depression in northern Uganda. Journal of Adolescent Health 51 (6), 544–550.

Betancourt, T.S., Borisova, I., Williams, T.P., Meyers-Ohki, S.E., Rubin-Smith, J.E., Annan, J., Kohrt, B.A., 2013a. Research review: psychosocial adjustment and mental health in former child soldiers – a systematic review of the literature and recommendations for future research. Journal of Child Psychology and Psychiatry 54 (1), 17–36.

Betancourt, T.S., Newnham, E.A., McBain, R., Brennan, R.T., 2013b. Post-traumatic stress symptoms among former child soldiers in Sierra Leone: follow-up study. The British Journal of Psychiatry 203, 196–202.

Bissouma, A.C., Te Bonle, D.M., Yeo-Tenena, J.M.Y., Moke, B.L., Kipre-Koiho, A., 2010. Profil psychopathologique des enfants associés au combat à l'ouest de la Côte d'Ivoire. Neuropsychiatrie de l'Enfance et de l'Adolescence 58 (6–7), 410–415.

Blattman, C., 2009. From violence to voting: war and political participation in Uganda. American Political Science Review 103 (02), 231–247.

Blattman, C., Annan, J., 2010. The consequences of child soldiering. Review of Economics and Statistics 92 (4), 882–898.

Bolton, P., Bass, J., Betancourt, T., Speelman, L., Onyango, G., Clougherty, K.F., Verdeli, H., 2007. Interventions for depression symptoms among adolescent survivors of war and displacement in northern Uganda: a randomized controlled trial. Journal of the American Medical Association 298 (5), 519–527.

Boothby, N., 2006. What happens when child soldiers grow up? The Mozambique case study. Intervention: International Journal of Mental Health, Psychosocial Work & Counselling in Areas of Armed Conflict 4 (3), 244–259.

Boothby, N., Crawford, J., Halperin, J., 2006. Mozambique child soldier life outcome study: lessons learned in rehabilitation and reintegration efforts. Global Public Health 1 (1), 87–107.

Boothby, N., Thomson, B., 2013. Child soldiers as adults: the Mozambique case study. Journal of Aggression, Maltreatment & Trauma 22 (7), 735–756.

Borisova, I.I., Betancourt, T.S., Willett, J.B., 2013. Reintegration of former child soldiers in Sierra Leone: the role of caregivers and their awareness of the violence adolescents experienced during the war. Journal of Aggression, Maltreatment & Trauma 22 (8), 803–828.

Boyden, J., 2003. The moral development of child soldiers: what do adults have to fear? Peace and Conflict: Journal of Peace Psychology 9 (4), 343–362.

Bracken, P., Petty, C., 1998. Rethinking the Trauma of War. Free Association Books, New York.

Burman, M.E., McKay, S., 2007. Marginalization of girl mothers during reintegration from armed groups in Sierra Leone. International Nursing Review 54 (4), 316–323.

Coalition to Stop the Use of Child Soldiers, 2008. Child Soldiers: Global Report 2008. London.

Cortes, L., Buchanan, M.J., 2007. The experience of Columbian child soldiers from a resilience perspective. International Journal for the Advancement of Counselling 29, 43–55.

De Silva, H., Hobbs, C., Hanks, H., 2001. Conscription of children in armed conflict – a form of child abuse. A study of 19 former child soldiers. Child Abuse Review 10, 125–134.

Denov, M., 2007. Is culture always right? The dangers of reproducing gender stereotypes and inequalities in psychosocial interventions for war-affected children. In: Dowdney, L. (Ed.), Psychosocial Web Pge. Coalition to Stop the Use of Child Soldiers, London.

Denov, M., Maclure, R., 2007. Turnings and epiphanies: militarization, life histories, and the making and unmaking of two child soldiers in Sierra Leone. Journal of Youth Studies 10 (2), 243–261.

Denov, M., 2010. Coping with the trauma of war: former child soldiers in post-conflict Sierra Leone. International Social Work 53 (6), 791–806.

Derluyn, I., Broekaert, E., Schuyten, G., De Temmerman, E., 2004. Post-traumatic stress in former Ugandan child soldiers. Lancet 363 (9412), 861–863.

Derluyn, I., Vindevogel, S., De Haene, L., 2013. Toward a relational understanding of the reintegration and rehabilitation processes of former child soldiers. Journal of Aggression, Maltreatment & Trauma 22 (8), 869–886.

Ertl, V., Pfeiffer, A., Schauer, E., Elbert, T., Neuner, F., 2011. Community-implemented trauma therapy for former child soldiers in northern Uganda. JAMA: The Journal of the American Medical Association 306 (5), 503–512.

Gregory, J., Embrey, D.G., 2009. Reducing the effects of profound catastrophic trauma for former child soldiers: companion recovery model. Traumatology 15 (1), 52–62.

Hermenau, K., Hecker, T., Schaal, S., Maedl, A., Elbert, T., 2013. Addressing post-traumatic stress and aggression by means of narrative exposure: a randomized controlled trial with ex-combatants in the eastern DRC. Journal of Aggression, Maltreatment & Trauma 22 (8), 916–934.

Honwana, A., 2006. Child Soldiers in Africa. University of Pennsylvania Press, Baltimore, MD.

IASC, 2007. Guidelines on Mental Health and Psychosocial Support in Emergency Settings. Inter-Agency Standing Committee, Geneva.

Jordans, M.J.D., Komproe, I.H., Tol, W.A., Ndayisaba, A., Nisabwe, T., Kohrt, B.A., 2012. Reintegration of child soldiers in Burundi: a tracer study. BMC Public Health 12.

Karki, R., Kohrt, B.A., Jordans, M.J.D., 2009. Child led indicators: pilot testing a child participation tool for psychosocial support programmes for former child soldiers in Nepal. Intervention: International Journal of Mental Health, Psychosocial Work & Counselling in Areas of Armed Conflict 7 (2), 92–109.

Klasen, F., Oettingen, G., Daniels, J., Adam, H., 2010. Multiple trauma and mental health in former Ugandan child soldiers. Journal of Traumatic Stress 23, 573–581.

Klasen, F., Gehrke, J., Metzner, F., Blotevogel, M., Okello, J., 2013. Complex trauma symptoms in former Ugandan child soldiers. Journal of Aggression, Maltreatment & Trauma 22 (7), 698–713.

Kohrt, B.A., Jordans, M.J., Tol, W.A., Speckman, R.A., Maharjan, S.M., Worthman, C.M., Komproe, I.H., 2008. Comparison of mental health between former child soldiers and children never conscripted by armed groups in Nepal. The Journal of the American Medical Association 300 (6), 691–702.

Kohrt, B.A., Maharjan, S.M., 2009. When a child is no longer a child: Nepali ethno-psychology of child development and violence. Studies in Nepali History and Society 14 (1), 107–142.

Kohrt, B.A., Jordans, M.J., Morley, C.A., 2010a. Four principles of mental health research and psychosocial intervention for child soldiers: lessions learned from Nepal. International Psychiatry 7 (3), 58–60.

Kohrt, B.A., Tol, W.A., Pettigrew, J., Karki, R., 2010b. Children and revolution: the mental health and psychosocial wellbeing of child soldiers in Nepal's Maoist Army. In: Singer, M., Hodge, G.D. (Eds.), The War Machine and Global Health. Altamira Press: Rowan & Littlefield Publishers, Inc., Lanham, MD, pp. 89–116.

Kohrt, B.A., Jordans, M.J., Tol, W.A., Luitel, N.P., Maharjan, S.M., Upadhaya, N., 2011. Validation of cross-cultural child mental health and psychosocial research instruments: adapting the depression self-rating scale and child PTSD symptom scale in Nepal. BMC Psychiatry 11 (1), e127.

Kohrt, B.A., 2015. The role of traditional rituals for reintegration and psychosocial wellbeing of child soldiers in Nepal. In: Hinton, A.L., Hinton, D.E. (Eds.), Legacies of Mass Violence. Cambridge University Press.

Kryger, L.S., Lindgren, C.L., 2011. Fighting for a future: the potential for posttraumatic growth among youths formerly associated with armed forces in northern Uganda. Intervention 9 (1), 6–20.

MacMullin, C., Loughry, M., 2004. Investigating psychosocial adjustment of former child soldiers in Sierra Leone and Uganda. Journal of Refugee Studies 17 (4), 460–472.

McKay, S., Mazurana, D., 2004. Where Are the Girls? Girls in Fighting Forces in Northern Uganda, Sierra Leone, and Mozambique. Rights and Democracy, Montreal.

Morley, C.A., Kohrt, B.A., 2013. Impact of peer support on PTSD, hope, and functional impairment: a mixed-methods study of child soldiers in Nepal. Journal of Aggression, Maltreatment & Trauma 22 (7), 714–734.

Nordstrom, C., 1997. A Different Kind of War Story. University of Pennsylvania Press, Philadelphia.

Okello, J., Onen, T., Musisi, S., 2007. Psychiatric disorders among war-abducted and non-abducted adolescents in Gulu district, Uganda: a comparative study. African Journal of Psychiatry 20, 225–231.

Ovuga, E., Oyok, T.O., Moro, E.B., 2008. Post traumatic stress disorder among former child soldiers attending a rehabilitative service and primary school education in northern Uganda. African Health Sciences 8 (3), 136–141.

Psychosocial Working Group, 2003. Psychosocial intervention in complex emergencies: a framework for practice. In: T.P.W. Group (Ed.), Working Paper. Queen Margaret University College, Edinburgh, p. 2.

Santacruz, M.L., Arana, R.E., 2002. Experiences and psychosocial impact of the El Salvador civil war on child soldiers. Biomedica 22 (Suppl. 2), 283–397.

Sartorius, N., Schulze, H., World Psychiatric Association, 2005. Reducing the Stigma of Mental Illness: A Report from a Global Programme of the World Psychiatric Association. Cambridge University Press, Cambridge, UK; New York.

Shakya, A., 2010. Experiences of children in armed conflict in Nepal. Children and Youth Services Review 33 (4), 557–563.

Somasundaram, D., 2002. Child soldiers: understanding the context. British Medical Journal 324 (7348), 1268–1271.

Stark, L., 2006. Cleansing the wounds of war: an examination of traditional healing, psychosocial health and reintegration in Sierra Leone. Intervention: International Journal of Mental Health, Psychosocial Work & Counselling in Areas of Armed Conflict 4 (3), 206–218.

Stevens, A.J., 2014. The invisible soldiers: understanding how the life experiences of girl child soldiers impacts upon their health and rehabilitation needs. Archives of Disease in Childhood. http://dx.doi.org/10.1136/archdischild-2013-305240.

Summerfield, D., 1998. Protecting children from armed conflict. Children affected by war must not be stigmatised as permanently damaged. British Medical Journal 317 (7167), 1249.

UNICEF, 2003. Fact Sheet: Child Soldiers. United Nations Children's Fund, New York, p. 4.

UNICEF, 2007. Paris Principles.

United Nations, 2005. Security Council Resolution, 1612 (S/RES/1612-2005). United Nations, New York, p. 6.

Vindevogel, S., Broekaert, E., Derluyn, I., 2013a. "It helps me transform in my life from the past to the new": the meaning of resources for former child soldiers. Journal of Interpersonal Violence 28 (12), 2413–2436.

Vindevogel, S., Schryver, M., Broekaert, E., Derluyn, I., 2013b. War-related experiences of former child soldiers in northern Uganda: comparison with non-recruited youths. Paediatrics & International Child Health 33 (4), 281–291.

Wessells, M.G., 2006. Child Soldiers: From Violence to Protection. Harvard University Press, Cambridge, MA.

Wessells, M.G., 2009. Do no harm: toward contextually appropriate psychosocial support in international emergencies. American Psychologist 64 (8), 842–854.

Williamson, J., 2006. The disarmament, demobilization and reintegration of child soldiers: social and psychological transformation in Sierra Leone. Intervention 4 (3), 185–203.

Child Soldiers as Victims and Actors: Lessons from Ethiopia

Tassé Abye, University of Nouakchott, Ksar, Nouakchott, Mauritania

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The international consensus achieved through a large campaign organized by humanitarian groups depicts children involved in armed conflict only as victims, an approach challenged by research-based and empirical studies. Consequently, divergent views on the sociological and/or political meaning of child soldiering have emerged recently. Humanitarian organizations have also tended to focus on child soldiers on the African continent. This article uses a child-centered perspective to demonstrate that child soldering may also result from political engagement.

Introduction

The Geneva Convention of 1949 defined the age of child soldiers as 15 years for child protection purposes; the 1977 protocols raised it to 18. The Cape Town principles accepted the definition in the 1977 Protocols and extended their scope. These principles suggest that a 'child soldier' is any child (boy or girl) less than 18 years who is part of any regular or irregular armed force or armed group involved in any capacity including cooks, porters, messengers, and anyone accompanying such groups, other than family members. It includes girls and boys recruited for forced sexual purposes and/or forced marriage. The definition does not refer only to a child who is carrying, or has carried, weapons. This definition is widely accepted and was endorsed by the United Nations General Assembly in 1989 through the Convention on the Rights of the Child (CRC). However, in 2002, the Optional Protocol of the CRC on Children's Involvement in Armed Conflict provided legal instruments that outlawed child soldiering. The international consensus gained through the humanitarian groups' campaign depicted children only as victims. This has been challenged by subsequent research. Consequently, divergent views on the sociological and/or political meaning of child soldiering have emerged during the past 20 years (Krijn and Richards, 1998; Hoffman, 2010; Brett and Specht, 2004). During this time, humanitarian organizations have also focused on child soldiers in the African continent. I examine these issues in this article, using original research based on Ethiopian experiences.

The Paris Principles of 2007 further developed and gave prominence to the 'victim approach' to child soldiering. Article 11 proposed that "children under 18 years of age who are or who have been unlawfully recruited or used by armed forces or groups and are accused of crimes against international law are considered primarily as victims of violations against international law and as alleged perpetrators."

That over 300 000 children are engaged in different conflicts across the globe is a dramatic international concern. Jezequel (2006: p. 1) identified the humanitarian focus on those living in the African continent. He states, "seven out of nine reports … by Human Rights Watch on the use of child soldiers in the last ten years, concern sub-Saharan African countries." The atrocities committed during the recent 'civil' wars in Sierra Leone, Somalia, Uganda, and different parts of Africa, contributed to shaping and framing the image of child soldiers as an African phenomenon. These child soldiers are described simultaneously as nonhuman, victims, perpetrators, killing machines, and pirates. Beah's (2007) personal experiences, other publications, and the international media have contributed to this image.

The devastating consequences of these different wars and the important roles children played in these armed conflicts, have contributed to this significant 'consensus' around the victim approach among humanitarian and governmental organizations. Academics approaching child soldiering from another perspective have found it a highly intimidating exercise. Shedding light upon or analyzing these issues from different angles has been interpreted as denial and insensitivity at best; at worst, a betrayal of the suffering of thousands of children from the consequences of inhuman wars. Moreover, when academics of African origin attempt to articulate child soldering in a fashion differing from the mainstream approach, it may be understood as merely defensive attitudes about their continent – compelling 'African' authors like me to avow that this is not our intention. Such research may be limited in scope to one country and there is no pretense of drawing a general conclusion from it.

Methodological Considerations

The argument I present in this article comes from a small research project aiming to understand the journey of those engaged in armed conflict in Ethiopia during the military regime (1974–91). Those study participants engaged in armed conflict during this specific period were part of an armed movement against a dictatorial government. Those siding with the military regime are not covered by this research.

This article is part of a larger research project of 8 years' duration on Ethiopian guerrilla movements. Among my sample for this publication are individuals who are 'happy' winners of the war against the military regime. Some hold political power in Ethiopia now; others have abandoned political life for other activities becoming professors, consultants, computer scientists, writers. I also interviewed former guerrilla fighters who have ended their struggle against the

military regime and live in other countries as migrants/refugees. I chose qualitative rather than quantitative research to explore the trajectories of former guerrilla fighters, and used snowball sampling to select the participants. Recently, I also had the opportunity to live in Ethiopia for 6 years. During this period, I observed former guerrilla fighters in various situations and enriched my understanding.

I also attended several gatherings of former guerrilla fighters who remained at home and comrades returning to Ethiopia from their migration journeys. These gatherings would often occur in the house of a former combatant and provided important opportunities for ethnographic observations. These events became 'celebrations' that began during late afternoon on weekends and lasted until late night. The learning provided was amazing. The former fighters turned migrants and those who never left would scrutinize, discuss, celebrate, mock, and tease each other, thereby rehearsing their previous collective past. There were moments where one of their roles during a specific operation was recalled and I observed how these discussions clarified questions that had been obscure for some participants. Combatants who had lost their lives during the war were remembered. I used these moments to learn about individual roles and positions before and during guerrilla life, and to contact individuals for further in-depth interviews. The informality of the discussions and their settings gave me a tremendous opportunity to clarify many questions that would have been difficult to grasp in formal individual interviews.

As researcher, I also observed two congresses of the Amhara National Democratic Movement (ANDM), attended by over a 1000 former guerrilla fighters. The congresses provided a huge opportunity to observe discussions among former guerrilla fighters currently living in different socio-economic conditions. During my time at Addis Ababa University, I could also work closely with former guerrilla fighters who had become members of academia.

I conducted 54 in-depth interviews with former guerrilla fighters (37 men and 17 women). However, the findings for this article relate chiefly to fighters aged less than 18 when they joined the armed struggle. This category encompassed 30 individuals (21 men and 9 women); each was interviewed for 2 h. They varied in age from 11 to 17 years. With five ex-rebels, I held lengthy, multiple interviews that lasted around 3 consecutive days. These usually took the form of a 'retreat.' The names (fictitious) I gave to these former fighters are Tarikuwa, Tesfayé, Kumeta, Tesfa, and Tadios. I thank them for the substantial time they gave to share their knowledge.

Women guerrilla fighters constituted around 30% of interviewees. Although limited in number, my research gives them voice. A major limitation of this study is the exclusion of former guerrilla fighters who today live as peasants. They represent the largest and most important group among Ethiopia's former guerrilla fighters. Their inclusion could bring another level of understanding to modes of engagement and their lives within the guerrilla movement and afterwards.

Thus, this research provides a limited and tentative understanding of the trajectories of a small number of Ethiopian former child/youth guerrilla fighters, without pretension of having a total understanding of the entire domain. However, it contributes to enlarging an area of research that has been insufficiently studied by social scientists in Ethiopia and beyond.

Context for Engagement

A number of guerrilla movements were formed during Ethiopia's Imperial reign. However, the February 1974 revolution became a turning point in contemporary Ethiopian history because this led to former combatants becoming part of the country's political elite and impacting daily life. This revolution 'surprised' political observers and activists. Started with a political slogan, *Yale Meneme Deme* (without a drop of blood), it became one of the bloodiest revolutions during the late twentieth century. In the absence of a strong, identifiable, and organized political party, a military group known as the *DERG*, who had neither a clear political agenda nor an articulated strategy, seized power.

The Ethiopian Left was recruited disproportionally among university and high school students. While the student movement's level of engagement prior to the revolution had been relatively high, its intensity was not comparable to the revolutionary momentum. The rapidity and magnitude of political involvement had no precedence in the country's history. While *a priori*, nothing had prepared them for a high commitment to politics, they acted as if they were ready and waiting for the revolution to come. The notion of *Zeitgeist* (spirit of the time) used by Gérard Chaliand (2008: p. 35) could explain their rush to engagement?

The relative ease with which the Imperial regime was ousted could have created in the perception and analyses of emergent political groups and their supporters, an illusion that a popular force could bring down any government if the will existed. Kiflu Tadesse, a leading Central Committee member of EPRP, ensured that 5 months after the beginning of the revolution, "a provisional youth league committee was formed with an assignment to set up the structure of the League" (Tadesse, 1998: p. xxix).

The politicization of the country's youth and its engagement in political organizations was reinforced through the *Zemecha* (Campaign). On 24 December 1974, the military government formulated a policy demanding that all university and high school students and teachers participate in a campaign that contributed to the betterment of peasant life. The military regime sent young people, aged 15 to 16 years and beyond, to the countryside to fragment and limit the emergence of an opposition, because it feared social upheaval. The closure of high schools and institutions of tertiary education had exactly the opposite effect. The engagement of young people in political activism can largely be attributed to ERP mobilization work during the *Zemecha*. About two-thirds of the interviewees trace their political engagement to their involvement in the *Zemecha*.

During this period of Ethiopian history, membership and allegiance to, or sympathy with, political movements had become a significant practice even for those aged 13 to 14 years. Organizing dance parties and listening to American music on the radio program 'Listeners Choice' had been a 'must' for large groups of young people living in cities, but they became unfashionable after the revolution. Young urban youths chose to participate in political meetings, workshops, read the works of Marx, Engels, Lenin, Stalin, and Mao, and actively engaged in

clandestine political Party Cells. Compared to young people, the number of 'adults' participating in underground political movements was smaller. Spending the night close to a library to be among the first to access Marxist books that arrived from abroad was comparable to today's reaction to the launch of a new Harry Potter book or new version of a playstation in major European or American cities.

The transformation of Ethiopian youth was radical. Young people in cities, especially the capital but also in the different provinces, discovered new role models through TV, radio programs, films, books, and pamphlets distributed by political groups such as *Democracia*, the EPRP official publication. Watching Chinese films like 'From Victory to Victory' had become an extremely popular activity.

Activities such as massive participation in demonstrations against the regime, delivering Party and Youth League pamphlets clandestinely, tagging cities with revolutionary and antigovernment slogans under the cover of night, placed the regime under high levels of stress. In response to the large-scale mobilization of young people, exactly 2 years after assuming power, the military government framed a new policy and announced the *Red Terror* on 11 September 1976. The military regime acted with extreme and widespread brutality. The military junta authorized government supporters, the army, police, and kebele militia to kill, torture, and imprison whomever they suspected. The government published multiple decrees limiting if not suppressing the freedom to write, demonstrate, organize, and even to think. Granted the right to kill, government supporters locked the cities and implemented the 'search and destroy' campaign.

During the May Day demonstration of 1977, according to Ottaway and Ottaway (1977: p. 147), "Hundreds were arrested, taken to three different sites on the outskirts of Addis Ababa, and executed *en masse*. Scores of others were gunned down in the streets by the Derg's 'permanent secretaries', the jeeps mounted with machine guns constantly patrolling the streets of Addis Ababa. The death toll may have been as high as 1000 for that single day." Christopher Andrew and Vzsili Mitrokhin (2005: p. 457) citing Amnesty International estimates, inform us that during the first days of the *Red Terror*, "the death toll could be as high as 500 000." The army, not satisfied with liquidating the country's youth, demanded that parents paid for the bullets that had killed their children if they wished to bury them.

That the EPRP quickly gained substantial support not only from young people but also from the few industrial workers and some members of the National Defense Forces (including members of the DERG) may have created over-confidence among the EPRP's leadership. They may have believed that they could take political power from the military junta without a long guerrilla war. A number of EPRP publications of that time suggested that the military government was very weak, without any social base to sustain its power. The military junta '*stands on pins and needles*' was the EPRP's credo.

The overall situation country-wide did not provide other alternatives for Ethiopia's young people. The slightest difference with the regime was immediately sanctioned by imprisonment, torture, or killing. While leaving the country and seeking refuge in neighboring countries had become the solution for many, others decided to organize in armed struggle against the regime.

"Almost unnoticed in these chaotic conditions, in 1975 a small group of university students calling themselves the Tigray People's Liberation Front (TPLF) launched an armed struggle for national autonomy for the northern province of Tigray" (Young, 1998: p. 36). Unlike the EPRP, the TPLF started rural armed struggle as the main form of engagement. Alongside collective struggle, some Tigrayan students and activists had been organized around their regional identities. Within the Left movement in Ethiopia, the issue of ethnicity has been an important organizing factor and remains central even today. Left activist Wallelign Mekonnen's (1969) famous article, *On the Question of Nationalities in Ethiopia*, became a landmark.

The Ethiopian People's Democratic Movement (EPDM) (later renamed Amhara National Democratic Movement) emerged in 1980, as a result of the 'partial' disintegration of EPRP/EPRA, which did not last long as the major political party attracting and fascinating Ethiopian youth. Five years after the official launching of the party and the formation of its armed wing, the EPRP/EPRA was dismantled.

Political Activism – A Prelude to Armed Struggle

Contrary to the image the international media broadcast of young people captured and forced to fight in wars that they did not understand, those interviewed for this research had made a conscious decision to participate in armed struggle. Among these were two types of participants: those who had been activists prior to becoming guerrilla fighters (24 out of 30); and those with little experience of militancy before joining the armed struggle (6 out of 30). Additionally, age when joining the armed struggle, place of residence, and level of education distinguish these two groups. Most of those who joined the guerrilla movement with earlier activist experience came from the cities, particularly Addis Ababa, Gondar, Mekele, and Axum; their average age was 15 years; only five were aged 17 years. Among the six respondents who joined the armed struggle without prior political experience, all but one were from rural areas and lived close to military bases of guerrilla organizations. The average age for this group was 15 years.

Many activists, particularly those who joined the armed wing of EPRP/EPRA, decided to engage in the rural armed struggle as a means of escaping imprisonment by the military regime. Tedros, who occupied ministerial positions for close to 20 years after the victory of Ethiopian People Revolutionary Democratic Forces (EPRDF), was a young EPRP activist studying at high school in Addis Ababa. He explained his decision to join the guerrilla movement as:

> I knew from other comrades that I was identified by the regime security. I had to stay underground in Addis for several weeks before the party arranged my escape to *Assimba*, Tigray and join the EPRA-…It was not safe to stay in Addis either for me or the party. Tedros said he was identified as *Tebeleché neber* (translated as 'I was eaten'). Moreover, *tebeleché, tebelah*' in some regions referred to the evil eye (Buda in Amharique), bringing an association between the military regime's security services' evil eye (*buda*) with another translation 'I was cheated, or I lost'.

Tadios, minister in the current government, joined the armed struggle when he was 17 years old. A member of the EPRP at 15, he worked against the military regime's objective of mass conscription of peasants from his region. His decision to join the armed struggle resulted from difficulties in remaining in the city. Exposed to some form of political life while studying at high school, he was recruited into EPRP during the *Zemecha*:

> While we were in the *Zemecha*, we heard that we were going to be killed by the security [forces] of the government. We left the *Zemecha* camp and two months later, I was asked to leave for the EPRA through a Party channel. At that time, *Belessa* was identified as the main area for the armed struggle in Gondar; because of my knowledge of that region, I and two others were sent by the Party to the EPRA.

Hailu, born into a family with a long tradition in the priesthood, a traditional career for the country's educated elite, became engaged in political activities during junior high school. Here, Hailu's major activity besides formal education was political action. The level of political activities in junior- and senior-high school in Ethiopia's cities during that period is difficult for outsiders to imagine. The kids at the school were divided into two major groups: those considered Left-wing and those supporting Haile Selassié's regime. Hailu recalled, "We had our red hats to distinguish us from those reactionaries … at one time there were no red hats left in town, they had been all sold out." A Left-winger at age 14 years, he was elected as member of the student association in high school. "It is through a university student coming back for vacation and others arriving for university service that I started to become involved in larger political activities prior to the revolution." Arrested by the Imperial security forces for being a political activist before reaching 18 years, he joined the EPRP a few days after Tadios.

The notorious *Shaleka* (Commander) Melaku governed the region where Tadios and Hailu lived. As Governor of the Gondar region, *Shaleka* Melaku was known for implementing the most monstrous program of mass killing and torture. Under *Shaleka* Melaku's leadership, thousands of young men, children, and women were killed, tortured, and jailed for belonging to the target age group: being young. Hailu, as many others who lived in the city, recited a popular lamentation of a woman whose son had been sent to Gondar prison: 'Shaleka Melaku, Ye Igzeir Tanash Wondem, Ye Zarein Marelign Ke ingdih Alwoldem' (translation: 'Commander Melaku, God's youngest brother, forgive my son for today. I promise never to give birth').

Given such oppression, Tariku recalled that sending activists to join the guerrilla movement was a last resort. Many had to stay and work underground in different Party structures for months and only when it became unbearable were they sent to join the rural armed struggle. Living underground in cities like Addis Ababa or Gondar was made extremely difficult after the government started house-by-house searches for EPRP members or anyone considered a potential activist. The numbers of those caught, imprisoned, and killed was countless. According to Kiflu Tadesse (1998), after 3 years of existence (1974–77), the EPRP/EPRA in Tigray, for example, counted only 200 fighters. Starting in 1976/1977, the *Derg* had launched a massive offensive against the EPRP in cities and killed many activists. However, by the end of 1977, the number of fighters in the ERPA in Tigray rose to over 1000 (Tadesse, 1998: pp. 365–373).

Also, the TNO/TPLF was preparing for rural armed struggle. Tito, a former top-ranking military officer, now a consultant on security and peace, recalls, "My engagement into political activities dated from my high school days. Having Zeru Keshen as a teacher (Zeru Keshen, was a leading University student activist and later a member of the Central Committee of the EPRP) was instrumental for me. His discussions pushed us to reflect on the situation of the country and made us think about and analyze it. But I think that it was during the *Zemecha* that I totally embraced the revolutionary struggle."

Berhane had the same experience as Tito: "After hesitating whether to join the TPLF or EPRP, I joined the TPLF directly from the *Zemecha* at 17 and never went back to my parent's home." Tesfayé, born to a poor peasant family, was in high school when he became interested in politics. He recalls that his political engagement was highly influenced by the teachers he met, a professor of Sri Lankan origin, and university service students like Estiphanos Seyoum.

Seifu, a 10th grade student, discovered politics during the *Zemecha*. He performed a wide range of political, administrative, and economic activities on behalf of the current ruling party until 2003 when he resigned from government and political activities to join the higher education system. He holds a Masters degree from a major American university and is currently enrolled for a PhD there. Explaining why he joined the armed struggle, he said:

> It was during the *Zemecha* that I discovered politics as did many students at that time, I was about 16 years old when I decided to join the armed struggle. The government was then engaged in a systematic harassment of the young generation in the entire country but more importantly in Tigray.

Like many others, Tafesse, was in high school in 1974 and traces his political engagement to the *Zemecha*.

> We were, 120 students who had been sent to the Benichangol region. The *Zemecha* students were organized in sub-groups of seven or eight, I was placed in a group of four strong university students who were articulate about the revolution and inspired us. Back in Addis, I joined the EPRP Youth League. That is how I started politics.

The different interviews and discussions held during this research revealed that those who joined guerrilla movements in the early stages of the struggle had some exposure to the Left-wing literature available then. Four of the nine women interviewed in this study, had become politically active through the *Zemecha*. Most were junior or senior high school students.

Although there is evidence of young men and women joining the armed struggle through some form of collective 'obligation,' (some villagers were persuaded to contribute by encouraging children to join the armed struggle in the combatant cadres), those interviewed stated clearly that they joined the armed struggle on a voluntary basis, thus

contradicting the commonly accepted assumption that they ended up in armies or armed opposition groups because they were either forced to join or because they lacked other options. Despite not being representative, respondents under the age of 18 indicated their determination to join the armed struggle. In Ethiopian society, *Jegnenet* (bravery) has structured the collective psyche. Ethiopian revolutionaries have used (and sometimes abused) traditional songs that celebrated individual engagement in performing acts of bravery, for mobilization and recruitment: 'Ye feri inate atechenki, lejish koura new nefsun tebaki, kezaf lay hono were teyaki. Ye Jegna inate tateki begemed, lejishin amora inji aykebrewem Balchaede.' (Translation: 'Mother of a chicken-hearted son, have no fear, your son is as wise as a raven, and will only gather information sheltered on the tree (sits idle while others sacrifice is implicit). Mother of a hero be brave and fasten your belt, unlike others, your son will die fighting and not be buried by his relatives (as a coward is implicit).')

Not a single interviewee mentioned that he or she had been forced or abducted, as observed by scholars of fighting forces in other conflicts. That the individuals I interviewed had been 'politically active' prior to their guerrilla engagement partly explains this situation. Moreover, for those living in neighborhoods near guerrilla military bases, their proximity to and contact with guerrilla fighters who originated from the large cities of Ethiopia, represented some form of access to 'modernity' and sophistication to which they aspired. Some may have considered joining the guerrilla movement as a 'rite of passage' to modern life. Toruwa, a member of the national parliament, said when interviewed, that he was the youngest in a family of five children and the only one to access education. Toruwa was in grade five and aged 13 years when she joined the armed movement. She said:

> I decided to join the armed struggle. The military base of the EPRP was not far from my village. These young men and women's commitment and discipline that I observed each time they passed our village impressed me. When I arrived in their military base, they did not know what to do with me. They discussed this for hours among themselves and were upset with those who allowed me, a kid, to join. Finally, they accepted me in the armed structures.

During several discussions held with members of the EPDM about their different wars, most praised Toruwa's courage. For many peasants in the countryside, Toruwa's unit was known to have been among the best fighters and was feared by many. That she was one of the few women occupying the position of commander played an important role in building her reputation.

While writing this article, Kumeta has just finished his PhD in a European university and returned to Ethiopia to teach and do research. Joining the guerilla movement at age 11, he was the youngest interviewee to do so:

> My uncle was a political activist and lived with us at my parents' home. He often explained the struggle to me. I had finished 6th grade…and really wanted to join the guerrilla fighters. I had tried twice to join the TPLF. Both times, the Front refused to enroll me arguing that I was too young. At my third approach, they finally accepted me. This time they were convinced that I would not go back. Somehow, I think that my insistence at remaining amused them.

Kumeta performed tasks ranging from radio operator to message conveyor for over 2 years before being engaged in the combatant forces.

Lungo, today in charge of an important national institution, recently gained a Masters degree from a well-known university in the United States. He joined the EPDM at age 16 years. Born into a small shopkeeper's family, he excelled at this trade when very young, "I use to go from one village to another to sell fabrics. Our area was a frequent passageway for the guerrillas. We really liked the way they were disciplined, the way they carried weapons, the way they walked.… At that time the Military regime's MIG fighter planes had already bombarded my village three times. When passing through our village, the guerrilla fighters had created structures for young people to have discussions with them … That is how some friends and I decided to join the guerrilla movement."

Conclusion

My interviews demonstrate that joining the armed struggle often resulted from prior political engagement. Beyond the national context of high repression, the revolutionary model for initiating change was highly valued by young generations in Ethiopia and other countries during the 1970s.

For Ethiopia, the *foco* (Mobile Strategic Force), a nucleus of a 'people's army,' to use Régis Debray's (1967: p. 24) term, was begun by university and high school students. Ethiopian armed struggle leaders, moved rapidly and deeply toward guerrilla organization and tactics, inspired by Maoism as well as Chinese and Vietnamese Irregular War theorists. The majority of combatants in Ethiopian guerrilla groups were drawn from the Ethiopian peasantry even though the instigators had been recruited primarily from the educated elite.

Most of those interviewed for this research possess important academic qualifications in their respective domains of study. They often mention their excellent academic standing in high school and at university. Statements about their academic performances are often followed by implicit comparisons to others in their age group. In evidencing academic achievements, former guerrilla fighters systematically indicated their young age (16–17 years) when joining university. It is as if these respondents wanted to ensure that the researcher understood that they made a conscious decision to become politically active and to sacrifice the possibilities for individual professional or social success that lay wide open to them.

Convinced that their socialist concepts could build a democratic system, these young university and high school students waged armed struggle to oppose the military regime that had usurped political power in their country and to liberate their people. These militants' characteristics do not correspond to those Ibrahim Abdullah and Patrick Muana (1998: p. 179) identified for the majority of those in the Revolutionary United Front of Sierra Leone, namely 'Lumpens, school drop-outs, young unemployed, illicit diamond miners

locally referred as *san–san boys*, drifters, hooligans, drug addicts and thieves.' The modalities of engagement, their social aspirations for Ethiopia and its population, their ambitions for the future, and indeed the lives of the young Ethiopian guerrilla fighters that I interviewed, had different motives and objectives from those in Sierra Leone.

See also: Armed Conflict and Social Work: Significance and Implications; Child Protection; Marxist Approaches to Social Work; Oppression; Social Work Theory.

Bibliography

Abdullah, I., Muana, P., 1998. The revolutionary united front of sierra Leon: a revolt of the Lumpenproletariat. In: Clapham, Christopher (Ed.), African Guerrillas. James Currey Publisher, Oxford, pp. 172–193.

Andrew, C., Mitrokhin, V., 2005. The World Was Going Our Way: The KGB and the Battle for the Third World. Basic Books, New York.

Beah, I., 2007. The True Story of a Child Soldier: A Long Way Gone. Harper Perennial, London.

Brett, R., Specht, I., 2004. Young Soldiers: Why They Choose to Fight. International Labour Organization, Geneva.

Chaliand, J.G., 2008. Le nouvel art de la guerre. Edition Archipel, Paris.

Debray, R., 1967. Révolution dans la révolution? Lutte armée et lutte politique en Amérique latine. Maspero, Paris.

Hoffman, K., 2010. The ethics of child-soldiering in the Congo. Young – Nordic Journal of Youth Research 18 (3), 339–358.

Jezequel, J.H., 2006. Child Soldiers in Africa: A Singular Phenomenon? on the Necessity of a Historical Perspective. Available on: www.diplomatie.gov.fr/frIMG/pdf/0605/jezquel-UK-2.pdf (accessed 19.01.14.).

Krijn, P., Richards, P., 1998. Youths in Sierra Leone: why we fight, africa. Journal of the International African Institute 68 (2), 183–210.

Mekonnen, W., 1969. Retrieved on the 14 March 2014. http://walilegnfordemocracia.com/onationalqu.pdf.

Ottaway, M., Ottaway, D., 1977. Ethiopia: Empire in Revolution. Africana, New York.

Tadesse, K., 1998. The Generation, Part II. The History of the Ethiopian People's Revolutionary Party. University Press of America, Inc, New York.

Young, J., 1998. The Tigray People's Liberation Front. In: Clapham, C. (Ed.), African Guerrillas. Indiana University Press, Bloomington.

Childhood and Adolescence: Developmental Assets

Richard M Lerner, Lisette M DeSouza, Mary H Buckingham, Daniel JA Warren, Robey B Champine and Kathleen N Greenman, Institute for Applied Research in Youth Development, Tufts University, Medford, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P.L. Benson, volume 3, pp. 1690–1697, © 2001, Elsevier Ltd.

Abstract

Research pioneered by Peter L. Benson, of Search Institute, has enhanced the study of developmental assets and thriving among youth. Thriving occurs when the strengths of youth are aligned with ecological developmental assets. Findings from the 4-H Study of Positive Youth Development (PYD) indicate that when individual and ecological assets are integrated, thriving occurs. We discuss contributions of Search Institute to the study of thriving, as well as findings from tests of the model of PYD by Lerner and Lerner.

In Peter Benson's last published work on the concept of developmental assets, an idea that he introduced into the developmental literature in 1990, Benson and his colleagues (Benson et al., 2011, p. 198) noted that:

> The framework of developmental assets, first posited in 1990 (Benson, 1990) and refined in 1995 (Benson, 1997, 2006), was explicitly designed to provide greater attention to the positive developmental nutrients young people need for successful development, not simply to avoid high-risk behaviors, and to accent the role that community plays in adolescent well-being…[t]he framework establishes a set of developmental experiences and supports hypothesized to have importance for all young people during the second decade of life. Recent work is taking a broader life span perspective, positing that developmental assets reflect developmental processes that have age-related parallels in infancy, childhood, and young adulthood.

This growing use of the concept of developmental assets occurred as a consequence of the importance of this concept within a broader model of youth development termed the positive youth development (PYD) perspective. We review the PYD perspective, and place special emphasis on the envisioned role of developmental assets in thriving among diverse youth. In turn, we summarize research examining the role of developmental assets in the promotion of PYD. Finally, we discuss implications of this research for evidence-based programs aimed at enhancing developmental assets for youth.

The PYD Model

The goal of developmental science is to describe, explain, and optimize intraindividual change and interindividual differences in intraindividual change across the life span (Baltes et al., 1977). Scholarship within this field is framed by relational developmental systems theoretical models (Overton, 2013; Overton and Mueller, 2012). According to these models, human development is shaped by the mutually influential relations between the developing individual and the multiple levels of his or her dynamic context. These bidirectional relations may be represented as individual ↔ context relations. These relations regulate the course of development. When these 'developmental regulations' are mutually beneficial to both the person and his or her ecology, they may be termed adaptive developmental regulations (Brandtstädter, 2006).

Examples of relational developmental systems theoretical models include Bronfenbrenner's bioecological theory (e.g., Bronfenbrenner and Morris, 2006), action theory models of intentional, goal-directed behaviors (e.g., Brandtstädter, 2006), Elder's (1998) life course theory, Magnusson's (e.g., Magnusson and Stattin, 2006) holistic person–context interaction theory, and the Ford and Lerner (1992) and the Gottlieb (1998) developmental systems formulations.

History, or temporality, is part of the ecology of human development that is integrated with the individual through developmental regulations. As such, there is always change and, as well, at least some potential for systematic change (i.e., for plasticity) across the life span (Baltes et al., 2006; Lerner, 1984). This potential for change represents a fundamental strength of human development. Of course, plasticity is never absolute; it is relative in the sense that development involves an integration of change-promoting and change-constraining processes and, as well, the integration of gains and losses in structure and function across the life span (Baltes et al., 2006). In addition, change for the better or worse can characterize any individual's developmental trajectory. Nevertheless, a key assumption of relational developmental systems theories – and, as we will note, of the use of these theories to understand both adolescent development in general and to frame the PYD conception of developmental processes more specifically – is that the developmental system is sufficiently diverse and complex such that there are a multitude of ways for researchers and practitioners to facilitate mutually beneficial individual ↔ context relations that may enhance the probability of change for the better and of promoting more positive outcomes of human development (J. Lerner et al., 2013).

During adolescence, there are several individual and contextual changes that are manifested in the developmental system. Changes in the prefrontal cortex, increases in the interconnectivity among brain regions, and increases in dopamine levels provide both vulnerabilities to risk and opportunities for growth in cognitive control and self-regulation (Steinberg, 2010). At the same time, youth may experience contextual

changes, such as changing schools (e.g., Eccles, 2004) and the increased relevance of peer pressure for risk-taking (e.g., Brown and Larson, 2009). Moreover, during adolescence, the individual has the cognitive, behavioral, and social skills to contribute actively, and often effectively, to his or her own development (Lerner, 1982). Accordingly, adolescence is an ideal developmental period for studying the relative plasticity of human development and for exploring how, through matching individual strengths with ecological developmental assets within the developmental system, positive development may be promoted.

The Study of Adolescence within the Relational Developmental System

The adolescent period is characterized by multiple dimensions of change involving levels of organization ranging from the physical and physiological, through the cognitive, emotional, and behavioral, and to the social, relational, and institutional. As previously mentioned, relative plasticity represents a fundamental strength of the adolescent period (Lerner, 2005, 2009) in that it reflects the potential for systematic changes that may result in more positive functioning. If adaptive developmental regulations emerge or can be fostered between the plastic, developing young person and features of his or her context (e.g., ecological assets linked to the structure and function of his or her family, school, peer group, and community), then the potential for thriving in youth may exist across the adolescent decade.

Indeed, predicated on relational developmental systems theories, the links among plasticity, adaptive developmental regulations, and positive development suggest that all young people have strengths that may be capitalized on to promote *thriving* (i.e., exemplary positive development; Benson et al., 2011) when integrated with developmental assets across the adolescent years. One example of the emerging strengths of adolescents is their ability to intentionally influence adaptive developmental regulations with their context (Gestsdóttir and Lerner, 2008). For instance, such intentional self-regulation may involve selecting positive goals that reflect important life purposes, using cognitive and behavioral skills (such as executive functioning or resource [asset] recruitment), to optimize the likelihood of achieving one's goals and, when goals are blocked, possessing the capacity to adjust or compensate effectively (Baltes and Baltes, 1990; Freund and Baltes, 2002).

When viewed from the perspective of relational developmental systems theories, young people represent 'resources to be developed' as opposed to the historical deficit view that young people are 'problems to be managed' (e.g., Roth and Brooks-Gunn, 2003). This strength-based view of adolescents has been used to study PYD within the United States (e.g., J. Lerner et al., 2013) and internationally (e.g., Gestsdóttir and Lerner, 2007; Silbereisen and Lerner, 2007). As previously noted, this research has been framed at a 'meta-level' by the ideas of individual ↔ context relations within relational developmental systems models. In addition, the research has been influenced by interest in the characteristics of PYD that emerge from this relational process, by the individual and ecological bases of these characteristics, and by interest in outcomes of the PYD process (e.g., youth community contribution or active and engaged citizenship) (e.g., Zaff et al., 2010).

There are several instantiations of this theoretical approach. Benson and his colleagues (e.g., 2006, 2011) formulated one such illustration and, in turn, Lerner and Lerner (e.g., J. Lerner et al., 2012) have expanded on Benson's model to propose a related conception of positive development (see J. Lerner et al., 2012 for a fuller description of other models, e.g., those of Damon, 2008; Eccles, 2004; Hamilton and Hamilton, 2005; Larson, 2000; Masten, 2001; Spencer, 2006). We first discuss the Benson developmental assets model, followed by the Lerner and Lerner model.

Peter Benson and Search Institute, and the Study of Developmental Assets

The work of Peter Benson and his colleagues at Search Institute (e.g., Benson, 2008; Benson et al., 2011) has been integral in providing the vocabulary for, and vision of, the strengths of young people and the communities in which they live, through the study of 'developmental assets.' Benson and his colleagues describe 'internal' or individual assets (i.e., 'skills, competencies, and values' of a young person) that are grouped into four categories: (1) commitment to learning; (2) positive values; (3) social competencies; and (4) positive identity (Benson et al., 2011). These individual assets represent the talents, energies, strengths, constructive interests, and 'sparks' that every young person possesses (Benson, 2008). When these individual strengths are aligned with 'external' or ecological assets (i.e., other individuals, community-level networks, institutions), thriving occurs. The ecological assets are conceived of as 'environmental, contextual, and relational features of socializing systems' and are organized into four categories: (1) support; (2) empowerment; (3) boundaries and expectations; and (4) constructive use of time (e.g., Benson et al., 2011).

Benson and his colleagues found these assets to be predictive of seven behavioral indicators of *thriving* including (1) school success; (2) leadership; (3) helping others; (4) maintenance of physical health; (5) delay of gratification; (6) valuing diversity; and (7) overcoming adversity (see Benson et al., 2011 for a review). For example, when internal assets (such as achievement motivation and school engagement) are combined with external assets (such as time spent in youth programs), school success was significantly predicted for six different racial/ethnic groups of 6th–12th grade students. In addition, higher levels of assets have been related to other positive developmental outcomes, including better physical health, lower levels of risk behaviors, and resilience (Scales et al., 2003).

Benson and colleagues have emphasized the importance of exploring potential practical applications of research on developmental assets by highlighting the role of communities in fostering well-being and positive development among young people. During the last two decades, research from Search Institute has helped communities to develop strategic goals and program initiatives for promoting PYD. Today, more than 300 communities across the country have incorporated Search Institute's asset-building framework into their efforts to foster positive and healthy functioning among youth.

In sum, the work of Benson and his colleagues has focused on research and applications aimed at connecting the strengths of youth with the assets of communities and building upon them to promote positive development. Current research by Search Institute seeks to extend the developmental assets approach to diverse youth in American and international contexts.

Richard M. Lerner, Jacqueline V. Lerner, and Colleagues, and the Study of Individual ↔ Context Relational Processes and PYD

Lerner, Lerner, and colleagues (e.g., J. Lerner et al., 2013) conceptualize PYD by drawing on the mutually influential individual ↔ context relations that underlie the PYD process. Their model has been tested longitudinally in the 4-H Study of PYD (e.g., Lerner et al., 2005, 2011), which seeks to identify the individual and ecological relations that may promote thriving and prevent risk or problem behaviors. Within the study, thriving is perceived as the growth of attributes that mark a flourishing, healthy young person (e.g., the characteristics termed the 'Five Cs' of PYD – competence, confidence, character, connection, and caring) (Eccles and Gootman, 2002; Lerner et al., 2005; Roth and Brooks-Gunn, 2003).

A key hypothesis tested in the 4-H Study is that, if: (1) the strengths of youth (e.g., a young person's cognitive, emotional, and behavioral engagement with the school context or possession of intentional self-regulation skills such as selection, optimization, and compensation (SOC)); can (2) be aligned with ecological developmental assets (e.g., other individuals, institutions, social networks); then (3) young people's healthy development may be optimized (J. Lerner et al., 2013). In addition, since positive development is characterized by adaptive and mutually beneficial relations between the developing young person and his or her environment, then thriving youth positively shape and are shaped by their environment and, as such, should be positively engaged with and act to enhance their world. In addition, they should be less likely to engage in risk/problem behaviors. **Figure 1** presents an illustration of the Lerner and Lerner conception of the PYD developmental process.

The figure illustrates that adaptive developmental regulations and their positive and problematic *sequelae* exist within the broader ecology of human development. This ecology includes history and culture and, thus, introduces change at all levels of organization within the developmental system (Bronfenbrenner and Morris, 2006; Elder, 1998).

Empirical Support for the Five Cs Model of PYD

As indicated earlier, in order to test the ideas presented in **Figure 1**, researchers at the Institute for Applied Research in Youth Development at Tufts University launched the 4-H Study of PYD in the fall of 2002. The 4-H Study of PYD is a longitudinal investigation supported by a grant from the National 4-H Council and the Altria Corporation. Data were collected annually from youth in Grades 5 through 12 in order to study youth in their actual environments, where youth and their parents, rather than research investigators, make decisions about how they spend their time. Full details of the methodology of the 4-H Study have been presented in numerous empirical publications (e.g., Lerner et al., 2005, 2011). Findings of the 4-H Study have provided empirical support for several key ideas within the Lerner and Lerner PYD model, including the antecedents of PYD as they may exist in the relations between youth strengths and ecological assets.

From the relational developmental systems PYD perspective, all young people have strengths that may be capitalized on to promote thriving across the adolescent years. One example of the emerging strengths of adolescents is their ability to contribute intentionally and actively to the adaptive developmental regulations with their context; for instance, as indexed through the use of SOC behaviors (Freund and Baltes, 2002; Gestsdóttir and Lerner, 2007, 2008). Other instances of strengths include specific self-regulations in key contexts of adolescents, as demonstrated through school engagement (Li, 2011) and future-oriented beliefs and emotional structures (Schmid et al., 2011a,b). For example, having a hopeful future orientation may help to activate adolescents' self-regulatory skills, even in the face of challenges.

The Lerner and Lerner PYD model specifies that the strengths of youth are linked to PYD in the context of ecological assets. The 4-H Study assessed the role of these assets in the PYD process depicted in **Figure 1**. The relations among observed ecological assets in the families, schools, and neighborhoods of youth with positive and negative developmental outcomes were assessed among 5th-grade youth from the 4-H Study (Theokas and Lerner, 2006). Ecological asset indicators were categorized into four categories: (1) human; (2) physical or institutional; (3) collective activity; and (4) accessibility, and were measured equivalently across the three contexts. Different dimensions of the family, school, and neighborhood settings had the most comprehensive impact on developmental outcomes, specifically on collective activity in the family, accessibility in school, and human resources in the neighborhood. However, in all settings, assets associated with individuals were the most potent predictors of PYD. Family assets were most important in the lives of youth; as such, one of the strongest predictors of PYD was eating dinner together as a family.

Subsequent analyses of the youth from Theokas and Lerner's work (2006) indicated that neighborhood dimensions interact with adolescent extracurricular activity involvement to predict PYD, depressive symptoms, and risk behaviors (Urban et al., 2009). The direction of these relationships differed by gender. When they engaged in extracurricular activities, girls who lived in lower-asset neighborhoods exhibited higher levels of PYD and lower levels of depressive symptoms and risk behaviors. At high levels of activity involvement, girls in high-asset neighborhoods exhibited increased levels of risk behaviors, particularly if they lived in neighborhoods with ample physical resources. The opposite relationships were found in boys. Moderate to high levels of involvement in extracurricular activities predicted lower levels of PYD and higher levels of risk behaviors for boys living in lower-asset neighborhoods. For boys living in high-asset neighborhoods, activity involvement was generally beneficial. Increased activity involvement was associated with increased levels of PYD and decreased levels of risk behaviors. The findings from this study pointed

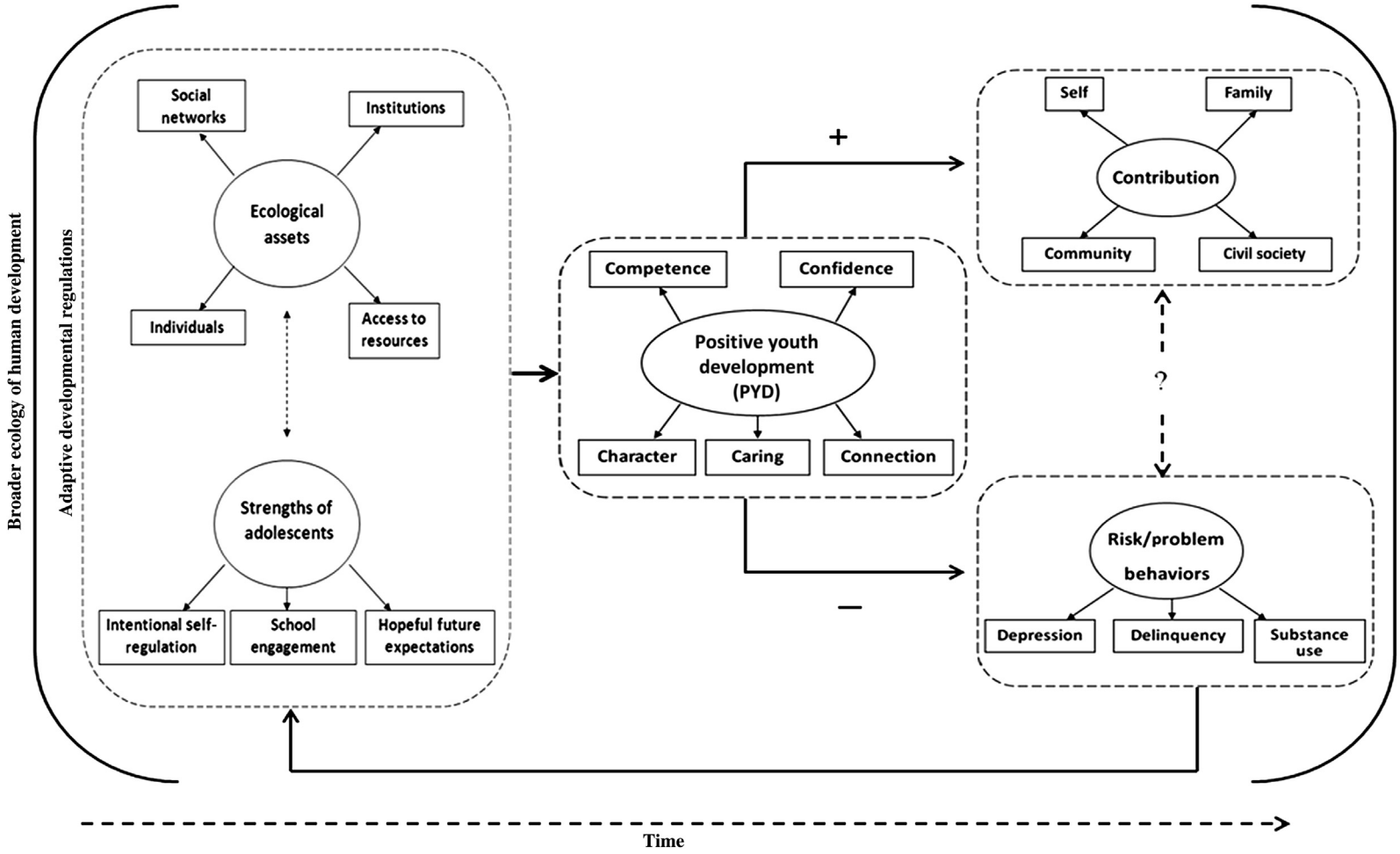

Figure 1 A relational, developmental systems model of the individual ↔ context relations involved in the Lerner and Lerner conception of the PYD developmental process. PYD = positive youth development.

to the need to consider the influence that multiple contextual influences can have on development.

In a follow-up study to work by Theokas and Lerner (2006), analyses by Bowers et al. (2011) assessed the relationships between these ecological assets and trajectories of positive and problematic development. Assets at the family, school, and neighborhood levels differentiated goal-optimization trajectories, while factors at the school level differentiated delinquency trajectories. However, an ecological asset was not found that consistently differentiated both goal-optimization trajectories and delinquency trajectories. The results indicated that collective activity in the family best predicted membership for the five goal-optimization trajectories that were identified, whereas school-based assets (physical resources and accessibility) differentiated the four delinquency trajectories that were identified.

The 4-H Study data have also been used to examine more specifically the ecological assets of parenting and youth programs in relation to the youth intentional self-regulation. For example, using data from Grades 5 to 8 from the 4-H Study, Lewin-Bizan et al. (2010) found a developmental cascade wherein positive parenting (as indexed by warmth and monitoring) was an important contextual asset predicting subsequent intentional self-regulation, and intentional self-regulation predicted subsequent scores for PYD, which, in turn, positively predicted later youth Contribution scores.

As already mentioned, several studies have also used the 4-H Study data set to examine possible interactions between self-regulatory processes and out-of-school-time (OST) activity participation. For example, Urban et al. (2010) found that both the strengths of youth and contextual resources are involved in thriving. Using data collected from youth in Grades 5 to 7, Urban et al. (2010) explored whether youth intentional self-regulation abilities moderated the effect of participation in OST activities on PYD among adolescents living in neighborhoods with relatively low levels of ecological assets. Overall, Urban et al. (2010) found that youth in these settings who had the greatest capacity to self-regulate (i.e., youth with the highest SOC scores) benefitted the most from involvement in OST activities, as demonstrated through their scores on measures of PYD, depressive symptoms, and risk behaviors. These relations were particularly strong for girls.

Mueller et al. (2011) examined whether the relationship between adolescents' SOC abilities and their participation in youth development (YD) programs in Grades 8 and 9 predicted their PYD and Contribution scores in Grade 10. Results indicated that while self-regulation skills alone predicted PYD, self-regulation and YD program participation both predicted Contribution. In addition, Grade 8 YD participation positively predicted Grade 9 SOC, which, in turn, predicted Grade 10 PYD and Contribution.

Conclusions

The results of the 4-H Study of PYD (as well as other research relevant to the PYD model; J. Lerner et al., 2012) provide important insights into what constitutes PYD and which individual strengths and ecological developmental assets help to promote adolescent thriving. We believe that the relational developmental systems approach used across all instances of the PYD model has been useful in understanding, first, the plasticity of human development and, second, the importance of dynamic relations between adolescents and their real-world ecological assets as promoting thriving among youth.

It would be useful if future research exploring the promotion of PYD adopted a common language and system of measurement for individual strengths and ecological assets. In addition, in order to enhance the understanding of ecological assets, the characteristics of contexts other than YD programs merit further study. For instance, more research should investigate the role of families and schools as ecological assets in promoting PYD. Similarly, contexts such as workplaces or faith institutions should be given more attention (Greenberger and Steinberg, 1986; King et al., 2011). Professionals working with families and schools, as well as employers who hire youth, would benefit from being educated about possible applications of how the ecological assets in these settings can enhance PYD among diverse youth.

Acknowledgments

The preparation of this article was supported in part by grants from the John Templeton Foundation, the National 4-H Council, and the Thrive Foundation for Youth.

See also: Life Course and Event History Analysis; Personality Changes During Adolescence Across Cultures; Self-Regulation During Adolescence: Variations Associated with Individual–Context Relations; Transition to Adulthood.

Bibliography

Baltes, P.B., Baltes, M.M., 1990. Psychological perspectives on successful aging: the model of selective optimization with compensation. In: Baltes, P.B., Baltes, M.M. (Eds.), Successful Aging: Perspectives from the Behavioral Sciences. Cambridge University Press, New York, pp. 1–34.

Baltes, P.B., Lindenberger, U., Staudinger, U.M., 2006. Life span theory in developmental psychology. In: Lerner, R.M. (Ed.), Theoretical Models of Human Development. In: Damon, W., Lerner, R.M. (Editors-in-chief), Handbook of Child Psychology, vol. 1, sixth ed. Wiley, Hoboken, NJ, pp. 569–664.

Baltes, P.B., Reese, H.W., Nesselroade, J.R., 1977. Life-Span Developmental Psychology: Introduction to Research Methods. Brooks/Cole, Oxford, UK.

Benson, P.L., 1990. The Troubled Journey: A Portrait of 6th–12th Grade Youth. Search Institute, Minneapolis, MN.

Benson, P.L., 1997. All Kids Are Our Kids: What Communities Must Do to Raise Caring and Responsible Children and Adolescents. Jossey-Bass, San Francisco.

Benson, P.L., 2006. All Kids Are Our Kids: What Communities Must Do to Raise Caring and Responsible Children and Adolescents, second ed. Jossey-Bass, San Francisco.

Benson, P.L., 2008. Sparks: How Parents Can Help Ignite the Hidden Strengths of Teenagers. Jossey-Bass, San Francisco, CA.

Benson, P.L., Scales, P.C., Hamilton, S.F., Semsa Jr., A., 2006. Positive youth development: theory, research, and applications. In: Lerner, R.M. (Ed.), Theoretical Models of Human Development. In: Damon, W., Lerner, R.M. (Editors-in-chief), Handbook of Child Psychology, vol. 1, sixth ed. Wiley, Hoboken, NJ, pp. 894–941.

Benson, P.L., Scales, P.C., Syvertsen, A.K., 2011. The contribution of the developmental assets framework to positive youth development theory and practice. In: Lerner, R.M., Lerner, J.V., Benson, J.B. (Eds.), Advances in Child Development and Behavior, vol. 41, pp. 195–228.

Bowers, E.P., von Eye, A., Lerner, J.V., Arbeit, M.R., Weiner, M.B., Chase, P., Agans, J.P., 2011. The role of ecological assets in positive and problematic developmental trajectories. Journal of Adolescence 34 (6), 1151–1165.

Brandtstädter, J., 2006. Action perspectives on human development. In: Lerner, R.M. (Ed.), Theoretical Models of Human Development. In: Damon, W., Lerner, R.M. (Editors-in-chief), Handbook of Child Psychology, vol. 1, sixth ed. Wiley, Hoboken, NJ. pp. 516–568.

Bronfenbrenner, U., Morris, P.A., 2006. The bioecological model of human development. In: Lerner, R.M. (Ed.), Theoretical Models of Human Development. In: Damon, W., Lerner, R.M. (Editors-in-chief), Handbook of Child Psychology, vol. 1, sixth ed. Wiley, Hoboken, NJ.

Brown, B.B., Larson, J., 2009. Peer relationships in adolescence. In: Lerner, R.M., Steinberg, L. (Editors-in-chief), Contextual Influences on Adolescent Development. Handbook of Adolescent Psychology, vol. 2, third ed. Wiley, Hoboken, NJ, pp. 74–103.

Damon, W., 2008. The Path to Purpose: Helping Our Children Find Their Calling in Life. Simon and Schuster, New York.

Eccles, J.S., 2004. Schools, academic motivation, and stage-environment fit. In: Lerner, R.M., Steinberg, L. (Eds.), Handbook of Adolescent Psychology, second ed. John Wiley & Sons, Inc, Hoboken, NJ, pp. 125–153.

Eccles, J., Gootman, J.A. (Eds.), 2002. Community Programs to Promote Youth Development/Committee on Community-Level Programs for Youth. National Academy Press, Washington, DC.

Elder Jr., G.H., 1998. The life course and human development. In: Damon, W. (Series Ed.), Handbook of Child Psychology. In: Lerner, R.M. (Vol. Ed.), Theoretical Models of Human Development, vol. 1, fifth ed., Wiley, New York, pp. 939–991.

Ford, D.H., Lerner, R.M., 1992. Developmental Systems Theory: An Integrative Approach. Sage Publications, Newbury Park, CA.

Freund, A.M., Baltes, P.B., 2002. Life-management strategies of selection, optimization and compensation: measurement by self-report and construct validity. Journal of Personality and Social Psychology 82 (4), 642–662.

Gestsdóttir, S., Lerner, R.M., 2007. Intentional self-regulation and positive youth development in early adolescence: findings from the 4-H study of positive youth development. Developmental Psychology 43 (2), 508–521.

Gestsdóttir, S., Lerner, R.M., 2008. Positive development in adolescence: the development and role of intentional self-regulation. Human Development 51 (3), 202–224.

Gottlieb, G., 1998. Normally occurring environmental and behavioral influences on gene activity: from central dogma to probabilistic epigenesis. Psychological Review 105 (4), 792–802.

Greenberger, E., Steinberg, L.D., 1986. When Teenagers Work: The Psychological and Social Costs of Adolescent Development. Basic Books, New York.

Hamilton, M.A., Hamilton, S.F., 2005. Work and service-learning. In: Dubois, D.L., Karcher, M.K. (Eds.), Handbook of Youth Mentoring. Sage, Thousand Oaks, CA, pp. 348–363.

King, P.E., Carr, D., Boitor, C., 2011. Religion, spirituality, positive youth development, and thriving. In: Lerner, R.M., Lerner, J.V., Benson, J.B. (Eds.), Advances in Child Development and Behavior, vol. 41, pp. 159–193.

Larson, R.W., 2000. Toward a psychology of positive youth development. American Psychologist 55 (1), 170–183.

Lerner, J.V., Bowers, E.P., Minor, K., Lewin-Bizan, S., Boyd, M.J., Mueller, M.K., Lerner, R.M., 2013. Positive youth development: processes, philosophies, and programs. In: Lerner, R.M., Easterbrooks, M.A., Mistry, J. (Eds.), Handbook of Psychology. In: Weiner, I.B. (Editor-in-chief), Developmental Psychology, vol. 6, second ed. Wiley, Hoboken, NJ, pp. 365–392.

Lerner, R.M., 1982. Children and adolescents as producers of their own development. Developmental Review 2 (4), 342–370.

Lerner, R.M., 1984. On the Nature of Human Plasticity. Cambridge University Press, New York.

Lerner, R.M., September 2005. Promoting positive youth development: theoretical and empirical bases. In: White Paper Prepared for the Workshop on the Science of Adolescent Health and Development, National Research Council/Institute of Medicine. National Academies of Science, Washington, DC.

Lerner, R.M., 2009. The positive youth development perspective: theoretical and empirical bases of a strength-based approach to adolescent development. In: Snyder, C.R., Lopez, S.J. (Eds.), Oxford Handbook of Positive Psychology, second ed. Oxford University Press, Oxford, England, pp. 149–163.

Lerner, R.M., Lerner, J.V., Almerigi, J., Theokas, C., Phelps, E., Gestsdottir, S., von Eye, A., 2005. Positive youth development, participation in community youth development programs, and community contributions of fifth grade adolescents: findings from the first wave of the 4-H study of positive youth development. Journal of Early Adolescence 25 (1), 17–71.

Lerner, R.M., Lerner, J.V., von Eye, A., Bowers, E.P., Lewin-Bizan, S., 2011. Individual and contextual bases of thriving in adolescence: a view of the issues. Journal of Adolescence 34 (6), 1107–1114.

Lewin-Bizan, S., Bowers, E.P., Lerner, R.M., 2010. One good thing leads to another: cascades of positive youth development among American adolescents. Development and Psychopathology 22, 759–770.

Li, Y., 2011. School engagement: what it is and why it is important for positive youth development. In: Lerner, R.M., Lerner, J.V., Benson, J.B. (Eds.), Advances in Child Development and Behavior, vol. 41, pp. 129–158.

Magnusson, D., Stattin, H., 2006. The person in context: a holistic-interactionistic approach. In: Damon, W. (Series Ed.), Lerner, R.M. (Vol. Ed.), Handbook of Child Psychology, sixth ed. John Wiley & Sons Inc., Hoboken, NJ, pp. 400–464.

Masten, A.S., 2001. Ordinary magic: resilience processes in development. American Psychologist 56 (3), 227–238.

Mueller, M.K., Phelps, E., Bowers, E.P., Agans, J., Urban, J.B., Lerner, R.M., 2011. Youth development program participation and intentional self-regulation skills: contextual and individual bases of pathways to positive youth development. Journal of Adolescence 34 (6), 1115–1125.

Overton, W.F., 2013. Relationism and relational-developmental-systems: a paradigm for developmental science in the post-Cartesian era. In: Lerner, R.M. (Ed.), Benson, J. (Series Ed.), Embodiment and Epigenesis: Theoretical and Methodological Issues in Understanding the Role of Biology within the Relational Developmental System. Advances in Child Development and Behavior, vol. 44. Elsevier, Amsterdam.

Overton, W.F., Mueller, U., 2012. Meta-theories, theories, and concepts in the study of development. In: Lerner, R.M., Easterbrooks, M.A., Mistry, J. (Eds.), Irving B. Weiner (Editor-in-Chief), Comprehensive Handbook of Psychology: Developmental Psychology, vol. 6. Wiley, New York.

Roth, J.L., Brooks-Gunn, J., 2003. What exactly is a youth development program? Answers from research and practice. Applied Developmental Science 7 (2), 94–111.

Scales, P.C., Leffert, N., Vraa, R., 2003. The relation of community developmental attentiveness to adolescent health. American Journal of Health Behavior 27 (1), S22–S34.

Schmid, K.L., Phelps, E.M., Kiely, M.K., Napolitano, C.M., Boyd, M.J., Lerner, R.M., 2011a. The relationship between intra-individual change in positive youth development and adolescents' hopeful futures: findings from the 4-H study of positive youth development. Journal of Positive Psychology 6 (1), 45–56.

Schmid, K., Phelps, E., Lerner, R.M., 2011b. Constructing positive futures: modeling the relationship between adolescents' hopeful future expectations and intentional self regulation in predicting positive youth development. Journal of Adolescence 34 (6), 1127–1135.

Silbereisen, R.K., Lerner, R.M., 2007. Approaches to positive youth development: a view of the issues. In: Silbereisen, R.K., Lerner, R.M. (Eds.), Approaches to Positive Youth Development. Sage Publications, London, pp. 3–30.

Spencer, M.B., 2006. Phenomenology and ecological systems theory. Development of diverse groups. In: Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology, sixth ed., vol. 1. Wiley, New York, pp. 829–893.

Steinberg, L., 2010. A behavioral scientist looks at the science of adolescent brain development. Brain and Cognition 72, 160–164.

Theokas, C., Lerner, R.M., 2006. Observed ecological assets in families, schools, and neighborhoods: conceptualization, measurement and relations with positive and negative developmental outcomes. Applied Developmental Science 10 (2), 61–74.

Urban, J.B., Lewin-Bizan, S., Lerner, R.M., 2009. The role of neighborhood ecological assets and activity involvement in youth developmental outcomes: differential impacts of asset poor and asset rich neighborhoods. Journal of Applied Developmental Psychology 30 (5), 601–614.

Urban, J.B., Lewin-Bizan, S., Lerner, R.M., 2010. The role of intentional self regulation, lower neighborhood ecological assets, and activity involvement in youth developmental outcomes. Journal of Youth and Adolescence 39 (7), 783–800.

Zaff, J., Boyd, M., Li, Y., Lerner, J.V., Lerner, R.M., 2010. Active and engaged citizenship: multi-group and longitudinal factorial analysis of an integrated construct of civic engagement. Journal of Youth and Adolescence 39 (7), 736–750.

Childhood: Anthropological Aspects

Thomas S Weisner, University of California Los Angeles, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Anthropologists believe that the most important influence in human development is the ecological and cultural setting within which a child will grow up. The anthropological study of childhood documents and accounts for the variety of childhoods found around the world, using comparative ethnographic evidence to test hypotheses about human development. It also studies the mechanisms in child, family, and community life for the acquisition, internal transformations, sharing, and intergenerational transmission of culture. Most importantly, it does this with close attention to the everyday contexts and routines of life, experience, meanings, intentions, and beliefs and goals of the communities, parents, families, and children themselves.

Introduction

The anthropological study of childhood documents and accounts for the variety of childhoods found around the world, using comparative ethnographic evidence to test hypotheses about human development. It also includes the study of the mechanisms in child, family, and community life for the acquisition, internal transformations, sharing, and intergenerational transmission of culture. Most importantly, it does this with close attention to the everyday contexts and routines of life, experience, meanings, intentions, and beliefs and goals of the communities, parents, families, and children themselves.

A thought experiment will illustrate the anthropological point of view about childhood. Imagine a newborn, healthy infant. What is the most important thing that you could do to influence the life of that infant? Most respond by mentioning dyadic interaction with the baby: hold and touch the infant a lot; provide good nutrition and health care; provide stimulation to achieve school success; love the baby; give it wealth and social capital; and so forth. Anthropologists believe that the most important influence in human development is the cultural setting within which the infant will grow up. Where on earth is the child growing up – where did the child end up in the 'birth lottery'? It is how, why, and by whom children are held, loved, fed, stimulated, punished, provided resources, and so forth, and how that varies so widely across human communities, that is, the focus of inquiry. Shaping a whole person engaged in family and cultural community life is the purpose of childhood development from an anthropological perspective. Childhood is a cultural project with goals, meanings, constant adaptation, and struggle, and anthropology provides the evidence for the startling and remarkable varieties of childhoods found around the world. Biological, psychological, and cultural anthropologists collaborate in the study of childhood, since biology, mind, and the cultural context are all required to understand childhood (Weisner, 2011).

The study of childhood and the process of children acquiring culture was almost entirely neglected by anthropologists until after 1925 (Whiting, 1968). Although a great deal of progress has been made, anthropology does not yet provide a single unified theory of why and how childhoods vary around the world or of childhood acquisition of culture. Rather the field offers rich, multivariate hypotheses and data on childhood (Super and Harkness, 1997).

The Stages of Childhood

Five stages of human growth and development are common to *Homo sapiens*: infancy, childhood, juvenility, adolescence, and adulthood (Bogin, 1999). Margaret Mead described lap children (infants, aged 0–1), knee children (toddlers, 2–3), yard children (preschool, 4–5), community children (juveniles in middle childhood, 6–12), and adolescence (Schlegel and Barry, 1991). These are maturational periods of development, alongside which cultural groups provide their own varied categories and marked stages of life. A new stage perhaps has emerged for study in the United States and some other Western, developed countries: emerging adulthood, the period from about age 18 to 25 or so, during which marriage may be delayed, education extended, and a search for work and identity continues (Arnett, 2011). Anthropologists analyze the cultural meaning of the very idea of stages, since they are used to account for children's behavior ('he's crying, but it is OK, because he's still a toddler'), as well as to assure and define normal and appropriate development ('she is eight and, so, old enough to start helping run our household'). Human cultures weave wonderful variations, meanings, and stories around panhuman maturational stages of childhood. The Beng of Ivory Coast, for example, believe that young children are still partly in yet another stage, a cultural world called *wrugbe*, where ancestors share life with prebirth children who are ambivalent about leaving that world. This helps explain for Beng why infants cry or are sickly: they want to return to *wrugbe* (Gottlieb, 2004).

Along with these maturational and social stages, evolution has prepared children and their caregivers to seek out information from the environment in all its many forms, cultural and noncultural. There is no question that despite the remarkable variation in parenting and children's learning environments around the world, children and adults alike are

prepared by evolution to be what Melvin Konner (2010) has called 'culture acquisition devices,' or CADs. Thinking of children as uniquely prepared by our evolutionary past to respond to and learn from the environment is a useful way to think about the continual interactions between our genetic inheritance and the importance of the cultural learning environment (CLE). Konner (2010) outlines 20 of these CADs or putative mechanisms that evolved to acquire cultural knowledge. These are divided into four broad categories of these learning mechanisms that are involved in the acquisition of culture: reactive processes in the cultural surround (such as classical conditioning, or social facilitation due to reduced inhibition, or instrumental or intentional conditioning); social learning (such as scaffolding, mimicry, imitation, and direct instruction); emotional/affective learning processes (attachment processes, along with positive or negative identification, or emotional management through rituals and scripts); and symbolic processes (cognitive modeling, schema learning, narrative, and thematic meaning systems).

Conceptions of Childhood in Anthropology

There is a variety of perspectives on childhood in anthropology. In one view, children are socialized into a set of norms and customs that they learn and then perpetuate. In this view, children are small adults in the making, ready receptors of traditions, shaped by parents and community adults to insure continuity in cultural and moral education, provide competence for survival in the ecology of the community, respect for tradition and appropriate behavior, and respect for elders in demeanor and gender roles.

In a second view, children's personalities and minds are understood as reflections of the cultural themes as well as the anxieties children grow up with (such as in the work in Bali of Bateson and Mead, 1942). The focus also is on the semiotics and communication of cultural meanings to children and on how these cultural patterns are absorbed and internalized, which in turn reproduce the meanings as well as neurotic obsessions and conflicts of their parents' cultures.

Third, the psychocultural, or personality integration model (Whiting and Whiting, 1975) begins with the climate, history, and ecology of a community, which shapes child-care practices, which in turn produce psychological effects on children. These contextual influences on children are produced by direct social learning as well as by psychodynamic processes shaping personality and defenses in children. These children become adults who then project into myths, rituals, art, and other forms (including in turn their own practices as parents) of the learned patterns, as well as intrapsychic conflicts produced in childhood and shared by others in their community.

The psychocultural model does not depend on a psychodynamic theory of learning. Children and adults alike have universal needs of the self, including the hunger for recognition, reward, and material and bodily satisfaction, to which communities respond through the cultural careers made available to children in a community. This core, psychocultural motivational force has been described as *affect hunger* by Walter Goldschmidt, who argues that this underlying need drives children to voraciously acquire cultural knowledge and also drives their parents and others to want them to do so (Goldschmidt, 1992). The desire for social recognition and inclusion more generally insures that most individuals link to and engage with society; however, culture channels these desires and, so, inevitably thwarts these needs, leading to intrapsychic and cultural conflicts. Melford Spiro, for example, used both psychodynamic and sociocultural approaches to understand the ideological, political, and ecological reasons for and consequences of the care of children by designated community caretakers, or *metapelets*, in socialist-inspired agricultural collective groups in Israel (Spiro, 1975).

Fourth, anthropologists study the 'developmental niche' of childhood: everyday physical/social settings, cultural customs of care, and the psychology of the caretakers of children. The cultural models of parenthood that are part of the customs and psychology of care direct behavior in their local, everyday settings (which include the goals, meanings, and rationales for parenting and being a child) (Harkness and Super, 1996). Parenting of children also is shaped by the organic hardware given by our common mammalian heritage and by socioeconomic conditions in the community. Children experience culture as it is practiced within their family's daily routine of cultural life. Cultural routines consist of activities children engage in (mealtimes, bedtimes, family visits, chores, going to church, school, play, etc.). Activities are the primary mechanisms bringing culture to and into the mind of the child, since cultural activities are what a child directly experiences day in and day out, repeated thousands of times. Activities consist of goals and values; tasks of an activity; the scripts for how to engage in that activity; the people present and participating in the activity and their relationships with the child and with one another; the motives and feelings and affective and emotional experiences of those involved; and the stability and persistence in the lives of the child and family (Weisner, 1996).

A related concept to the developmental or ecocultural niche is the CLE (Edwards and Bloch, 2010). The CLE consists of a cultural activity in progress, occurring in a social setting in some ecological and historical context, with norms of behavior for that activity and with characteristic people and relationships in the settings. Many features of the CLE have been shown to influence children's social behavior and mind. For example, gender, status in the community, resources, age, family structure (nuclear or extended), formal education, and other features all influence the amounts and ways of expression of children's prosocial behaviors, aggression, nurturance, and other outcomes.

Cultural and ethnic group identification, and the child's and family's own strength of social identity, is an important influence both of a child's own self and, therefore, of the child's and family's connections to the cultural community. Families often are the context in which this sociocultural identity is acquired and experienced due to co-residence, shared surnames and language, and similarity of physical appearance, among other features. The more this social identity is tied to provision of resources and support, and the more salient it is, whether due to external threats or opportunities or a positive sense of a shared heritage, the greater the importance the cultural community has in a child's development.

Finally, some anthropologists view childhood itself as a cultural construction shaped by forces within as well as outside a single cultural community. The very idea of what a child or parent is, in this view, is more the outcome of processes of power in an increasingly global political economy, in which children as well as parents are *constructed* or *positioned* by these agents of power (the state, market economy, political and other elites, and other economic and political forces) (Stephens, 1995). The emphasis of this approach also is on the child as active agent or engaged actor in defining its own socialization. This view counters the idea that socialization and enculturation is done primarily to the child. Culture acquisition also depends on the child's active, often thoughtful participation as well. Children can and do resist what family and society want for them and want of them, and this process of resistance and transformation is an important perspective in anthropology.

Anthropology has contributed to our understanding of formal education through work on classrooms and schools, showing the cultural influences on how teachers teach and manage classroom behavior. Anthropology broadens the study of 'education' to include all forms of social and cultural learning outside of schools, however, and how this much broader world of everyday learning connects with formal classroom learning (Lancy et al., 2010). Anthropological studies also include nonformal and informal education as well as formal schooling, such as out-of-school activities (clubs, music, church groups for children and youth, sports, gaming), various forms of private and commercial schooling (other than formal education), and the intersections between these various settings for learning that are common everywhere.

Some Cultural Influences on Children's Development

Cultural Scale and Complexity

Children in more complex societies (with occupational specialization, an extensive market economy, a nucleated settlement pattern, centralized and hierarchical political and legal system, and a centralized religious priesthood) are more likely to seek help and assistance from others, to try and dominate or control others, and to be more egoistic. Children in less complex societies are more likely to show nurturance toward other children (to offer assistance and respond to their requests), be more responsible, make more responsible suggestions to others, and do more tasks required for family and community survival. Mothers who have heavy subsistence workloads are more likely to expect responsible work from children and use stricter discipline. Children living in extended, joint, or expanded households and family systems are more often involved in directive, aggressive interactions, while children living in smaller nuclear families are more often engaged in sociable and intimate interactions with parents and others. Additionally, fathers in smaller conjugal households and family systems are, on average, more involved with children compared to fathers in joint/extended families.

Of course, children and adults everywhere seek help, show responsibility, or are aggressive and so forth. These patterns only reflect the modal tendencies of communities, not a rigid uniformity within them (Whiting and Edwards, 1988).

Gender Differences

Gender differences in children's development are recognized and shaped by all cultures (Ember, 1981). Of five kinds of interpersonal behavior in children aged from 3 to 11 (nurturance, dependency, prosocial dominance and affiliations, egoistic dominance, and sociability), girls on average were more likely than boys to be nurturing toward others, while boys were more likely to be egoistically dominant and aggressive than girls. Play styles and types of play vary by gender (girls are more likely to do joint work-play and to do so nearer their homes, for instance). Women and girls do most caretaking of children during the juvenile period in most cultures, so girls experience care by their own sex, while boys do not, leading to differences in early gender identification and psychosocial and self-development. Peer groups have a tendency to segregate by gender, and children in middle childhood tend to prefer same-sex children to interact with where there is a choice. Cultures with more mixed-age and mixed-gender groups in settings around children are likely to have less sex-segregated roles for children. Individual differences among boys and girls are usually substantial, even within communities where there are strong overall gender differences in development.

Father roles are recognized in all societies (Shwalb et al., 2013). Fathers seldom are involved in direct care of infants and young children in most societies, but they do have complementary nurturing and affiliative roles. Fathers also are more likely to be involved in economic, protective, and didactic child training. In a study of 80 preindustrial societies, fathers were more often nearer to and involved with young children in monogamous, nuclear family, and non-patrilocal residential situations and in sociocultural circumstances where mothers make relatively large contributions to family subsistence. Father involvement is related to sociocultural evolution: foraging societies report more father participation in childcare, while horticultural, agrarian/peasant, and pastoral subsistence-based societies have tended to have less. There is an upswing in contemporary societies in encouraging paternal care. The cultural beliefs about gender (how women's as well as men's roles are defined by parenting), as well as the ability of fathers to provide consistently for their children, influence father involvements. Poverty and uncertain economic life, or migration and dislocation, can drastically change father as well as mother involvement in patterns of childcare.

The study of gender continues to broaden in scope to include anthropological studies of the trafficking of girls for sexual exploitation – a growing world problem. There is selective abortion practiced around the world, which favors male births over female. Anthropology also is focused on the education of girls and the consequences of that education for a broad range of outcomes. For example, LeVine et al. (2012) ask why there is a correlation between increasing education for girls and women and gradually declining fertility rates and improved health for those women with more formal education and the communities they are in. Although the association has been known for a long time, the mechanisms

were not. Girls' and women's education and literacy provides tools and mechanisms for increased interactions with health and other bureaucracies regarding managing family size, birth control, and family health. These women also are more likely to recognize and practice birth control and wider birth spacing and to socialize their own children using more pedagogical child-rearing practices.

Emotional Development

Emotional development in childhood is influenced by cultural expectations at each developmental stage about the demeanor expected by boys or girls of that stage in particular cultural settings. A child should show that he or she is a certain kind of cultural person with an appropriate self and identity in his/her emotional comportment. Cultural management of emotion relies on what Robert Levy (1973), in his study of Tahitians, called *redundant cultural control*. Tahitians, for example, as well as many other Pacific Island communities, expect children to be calm, gentle, and quiet in demeanor (except for an extended period of adolescence and youth called *taure'are'a* in Tahiti, in which adventures, autonomy, rebellion, and aggressiveness are culturally expected and common). Redundant community management of 'gentleness' includes many beliefs and practices: children are somewhat distanced from their mothers and fathers after infancy and spend much of their time with peers; socialization networks are diffuse, meaning that affect toward others also is diffused; severe anger is strongly discouraged, while mild transient episodes are tolerated; threats are common, while actual aggression toward children is not; accidents are reinterpreted as punishment by spirits for aggression; there can be magical retaliation for serious anger; and, it is generally shameful to show lack of control. A culture complex of many interrelated beliefs and practices of this kind, such as the Tahitian example, including lexical labeling of the expected behavior patters, recognized cultural scripts, and stories adding to its emphasis, all are strong signs that some emotional pattern or competence in children is of adaptive and moral importance to a society.

Ochs and Izquierdo (2009) have tied emotional development to processes of learning responsibility and a sense of obligation and respect. There are clear differences in how communities socialize their children for responsibility and task performance around the world. The US middle class is unusually low in accomplishing such training in responsible tasks for their children compared to many regions of the world.

Basic Trust and Attachment

Another example of cultural variation in development comes from studies of basic trust and attachment. The universal socialization task for cultures regarding attachment concerns the learning of trust, rather than insuring the *secure* attachment (as defined by particular psychological scales) of an individual child to a single caretaker in a dyadic relationship. The question that is important for many if not most parents and communities is not, 'Is (this individual) child securely attached?', but rather, 'How can I insure that my child knows whom to trust and how to show appropriate social connections to others?' 'How can I be sure my child is with others and in situations where he or she will be safe?' Parents are concerned that the child learns culturally appropriate social behaviors that display proper social and emotional comportment and also show trust toward others.

Successful attachment does not depend on only one kind of maternal care in nuclear families, nor a specific kind of infant and toddler behavioral style. Anthropological studies show that a wide range of family and parenting practices found around the world can produce close affiliation and trusting attachments in children (Otto and Keller, 2013; Quinn and Mageo, 2013). Although the individual child is named and recognized everywhere (this is the basic process of individuation), individualism and egoistic autonomy as goals clearly are not universal. Sociocentric and interdependent self-development is more common ideals in much of the world compared to autonomous and independent development (Shweder and Bourne, 1991).

Chisholm (1999) proposes an evolutionary developmental hypothesis regarding trust and attachment. Considering the long course of evolution, children and families faced highly changing kinds of environments. These environments varied in how risky, uncertain, and variable they were. Less threatening, more favorable material and social conditions led to greater investment in fewer children and, so, encouraged closer attachments to one or a few caregivers. Unfavorable conditions encouraged what are labeled (inappropriately so from an anthropological perspective) *insecure* or avoidant/ambivalent infant and child behaviors observed during standardized assessments or rated through interviews or questionnaires. In threatening, insecure conditions, those often were the more adaptive, successful parental and child responses that were likely to increase the chances of children reaching reproductive age.

Multiple caretaking is found widely across communities around the world (Weisner and Gallimore, 1977). Most cultures provide multiple caretakers to children, rather than a single person. Care is socially distributed across many in the family and community in most children's lives, not dependent on a single person. Indeed, a child living an entire childhood exclusively in his or her natal home may well be the exception around the world. Older siblings and cousins are widely used as caretakers as well as other kin, hired care, and group care. Extended families in village and agrarian-based societies have high levels of multiple care of children. In India and elsewhere in Southeast Asia, for instance, there is an intense childhood experience with several maternal figures (Seymour, 1999). This means that a focus on a single dyadic connection as the key indicator of 'secure' emotional attachment does not fit the experience of most children and caregivers, nor evolutionary evidence (Hrdy, 2009).

Many communities (particularly more hierarchical, agrarian/horticultural, and respect-oriented) want infants and children to exhibit calm, respectful, quiet demeanor in the presence of strangers and community members generally, and there is evidence from the Nso of Cameroon, Gusii of Kenya, and many other societies that even very young children learn

and display this relationship style. This is not a sign of insecure attachment (Keller, 2012; LeVine and Norman, 2001). Learning about trust and security in many communities includes socialization for respecting hierarchy, showing deference, and training children to not expect parental responsiveness and warmth, as Jeannette Mageo has shown for Samoan family life (Mageo, 1998). Mageo suggests that achievement of group trust and attachment is not based only on a positive feeling of security. Trust and relational security may well include experiences of ambivalence and even avoidance of some family and kin, as well as acceptance within the group, including kin and community support. Similarly, managing anger and expecting and accepting only intermittent maternal attention are believed to be the appropriate training for learning how to anticipate and deal with security and separation among the Murik of Papua New Guinea (Barlow, 2004). The Murik do provide early mother-specific attachment socialization, which is closely tied to nursing and the provision of food. They also emphasize the diffusion of emotional relations outward to other caretakers (siblings and kin), who generally are harsher and who discipline the young child most often (rather than the mother). The Murik cultural intention is to complement the one with the other kind of care. The cultural message is that both of these kinds of caretaking experiences are needed for true, realistic emotional security in the Murik world. Both patterns certainly are what Murik will face in life. Hence, both patterns are emphasized and defined as good mothering and secure attachment. To be pluralistic and situation-centered, not only person-centered or dyadic relationship-centered, is an important component of the socialization for trust and a sense of security in most communities.

Developmental Goals

Anthropology does not assume that competencies valued in Western communities (verbal skills, cognitive abilities, or signs of egocentric autonomy of the self, for instance) are necessarily desired or even meaningful child developmental outcomes elsewhere (LeVine, 1988). However, nearly all cultures are concerned over some version of good communication, mental ability, self-construal, and personhood. LeVine et al. (1994) contrast *pedagogical* goals (cognitive and social stimulation to prepare children for literacy and schools, as well as for an individualistic and autonomous self-construal) and *pediatric* goals (concern for survival, health, and physical growth of infants, and subsequent responsible engagement in family subsistence and family continuity). Of course, in the contemporary world, many parents and communities want a hybrid combination of these goals for their children (Edwards and Whiting, 2004).

Anthropology has a unique point of view regarding the goals for a good childhood: the production of cultural well-being in children. Well-being is more than physical health, the attainment of skills and competence, the absence of illness or misfortune, or of successful subsequent reproduction, important as these are. Well-being is the ability of a child to engage and participate in the activities deemed desirable by a cultural community and the psychological experiences that go along with that participation.

The Acquisition of Culture

The roles and settings in which children learn of course influence when and how children will acquire cultural knowledge. Children are apprentices to more experienced community members in doing important tasks, and this apprenticeship situation is a powerful learning experience for children. Play and work blend in childhood learning. Imaginative, fantasy, toy, physical, and motoric kinds of play (including organized sports with rules) vary according to whether adults encourage it or whether it is considered beneficial for children by adults because they believe that it enhances desired competencies or furthers their societies' developmental goals (such as cognitive and school-like activities in many contemporary cultures). Of course, play is universal among children, as soon as their sheer inventiveness, creativity, and exuberance take over.

Adults are rarely involved in play with children in most societies around the world, in contrast to the emphasis on adult-child play from early childhood on in the United States and other Western, developed countries (Lancy, 2007). For example, the Kpelle of Liberia in the 1970s had children playing on 'the mother-ground,' or open public spaces where children can observe, lurk nearby, and imitate adults going about their activities in their agrarian village community. Formal schooling contrasts sharply with this mother-ground of childhood and everyday learning in context. Anthropological studies of schooling find striking differences in the culture of classrooms around the world, including different teaching practices and student expectations. A moral and cultural curriculum accompanies the formal literacy, numeracy, and sciences training in classrooms. Preschool classrooms around the world reflect implicit cultural models regarding the importance of collective, group activities (in contrast to more individually chosen activities preferred in many US preschools). For example, US teachers intervene and use words to negotiate disputes, while Chinese and Japanese classrooms monitor and encourage children to settle their own disputes. There are many other such national/cultural differences (Tobin et al., 2009). Participation in ceremonies and rituals at times of baptism, birthdays, naming ceremonies, puberty, and marriage also has powerful influence on children's acquisition of cultural knowledge. Such ceremonies crystallize cultural beliefs and practices. They intensify emotionally, politically, and socially salient key concerns that parents and communities have about childhood and elevate the goals the community shares for children and parents (Turner, 1967).

Multiple cultural and mental processes are involved in culture acquisition; however, the relative importance of different mental and cultural mechanisms for emotional, social, or cognitive learning are currently not well understood in anthropology. Evolved tendencies of the mind prepare children to understand the world in certain ways. For example, children in widely disparate cultural communities seem to share understandings about what living things are like and how they behave and think (Hirschfeld and Gelman, 1994). Psychodynamic processes transform emotionally salient cultural information. Stories and narratives embed cultural knowledge, shape recall, and organize cultural knowledge into scripts with shared local meaning. Sociolinguistic studies

of child language acquisition show wide variations in how and when parents talk to their children. Language learning is embedded in interactional routines shaped by cultural practices, with children as active learners (Schieffelin and Ochs, 1986). Children's play, games, and language use show strong evidence of class and race/ethnic influences as well as cultural knowledge. Children display remarkable creativity and agency and use the richness of language in shaping social relationships, beliefs, and practices (Goodwin, 1990). Neither children nor parents generally have explicit, conscious awareness of all the complex cultural knowledge that they know. Most cultural scripts are seldom in conscious awareness, even as they drive actions in accord with their community's beliefs. Cultural beliefs and practices have powerful *directive force* in guiding child behavior and child socialization in part because of this shared, implicit, everyday understanding put into action (D'Andrade, 1995; Lancy et al., 2010).

Anthropological Methods and the Study of Children

Anthropological methods for the study of children include qualitative interviews and conversations, ethnography, and participant observation (Weisner, 1996). These methods are important for operationalizing the anthropological concept of childhoods lived in cultural pathways in naturalistic settings. Systematic observational procedures, field guides for comparative studies, and special procedures for sampling children's activities and time use enhance ethnography (Bernard, 1995; Munroe and Munroe, 1994). Anthropologists also use assessments standard in child development for comparing physical growth and the cognitive and socioemotional life of children, often revising these standardized measures to insure that culturally appropriate procedures and meaningful outcomes are being measured. The processes of language-mediated culture acquisition, using sociolinguistic methods, are essential to understand the process of 'becoming a speaker of culture' (Ochs, 2002) and are central in anthropology. Film and video records of childhood are invaluable for comparative studies of cultural activities, emotional expression, daily routines and activities, the physical environment of the home, and patterns of gaze and attention, for example (Ochs and Kremer-Sadlik, 2013).

The cumulative cross-cultural ethnographic record of the lives of children and parents around the world is a unique, invaluable, and remarkable contribution to the scientific study of childhood and human development (Lancy, 2008; LeVine, 2007). The use of ethnographic and qualitative evidence in the contemporary study of children's lives is also increasing throughout the social sciences (Yoshikawa et al., 2008; Weisner, 2011). Using these methods for studying children's lives insures that the social, cultural, and ecological context of children's lives is fully included in the study of child development, and that the experiences, meaning systems, and intentions, goals, and values that influence the developmental pathways available to children are never excluded from study. For analytic reasons, it is, of course, important to bracket context out, and use a wide range of methods for study, but anthropologists do not forget that the world is not linear, additive, and decontextualized. Anthropological methods therefore will always be essential (Weisner and Duncan, 2013).

Anthropology and the Study of Childhood in the Twenty-First Century

Anthropology has always been concerned with the experiences and the cultural worlds of minorities, of the poor and nonliterate, and of those, including children, who are so often unable to give voice to and represent their own world. Life histories and autobiographical accounts of childhood have provided rich data, as in the classic *Nisa: The Life and Words of a !Kung Woman* (Shostak, 1981). Scheper-Hughes (1992) describes infants and young children in deeply impoverished political and economic circumstances in northeastern Brazil, circumstances leading to high infant and child mortality, and anger and despair among parents (and anthropologists). Anthropological studies of African-American families and economically downwardly mobile families in the United States demonstrate how some (but not all) can rely on extended kin in their struggles with poverty. Anthropological studies of childhood disability and deviance find greater acceptance and social integration of children with physical and cognitive disabilities in many communities than in the United States and other Western societies, as long as children are able to live as sufficiently cultural persons in their communities and are not violent or dangerous to others (Ingstad and Whyte, 1995). Disability and mental health usually are considered by anthropology in the context of other sources of poverty and social exclusion along with the disability or mental illness itself.

Anthropologists are concerned with children at risk around the globe, including children under stress from academic examinations in Japan and Korea, immigrant children in Europe and elsewhere, including street children, or children facing war, forced migration, and economic exploitation due to globalization and other change in Africa and elsewhere. Child sexual and physical abuse around the world is now a recognized concern for anthropologists. Cultural beliefs and practices regarding appropriate discipline and treatment of children clearly do vary widely, and Western notions of abuse are not universal; however, repeated and unchecked physical aggression or intrafamilial sexual relations between close kin and children are nowhere defined as normative and acceptable (Korbin, 1981). Anthropologists are concerned with children's rights, recognizing their vulnerable status and the lack of provision of basic protections for children (*Cultural Survival Quarterly*). World youth cultures are growing in importance due to the influence of the Internet and mass communications around the world. The recognition of and attention to children who are deprived, vulnerable, and excluded for a wide variety of reasons continues to be a central concern for anthropology. These are all topics for the anthropology of childhood in the twenty-first century (Weisner and Lowe, 2005).

Globalization is an important force in the world today. Nonetheless, regional, cultural, and family system differences remain very powerful. Rogoff (2011) illustrates this remarkable

blending of tradition and change in a Mayan village in Guatemala across four generations, as does Greenfield's multigenerational study of weaving, child development, and economic change in Southern Mexico (2004). Ngecha, a Kikuyu community in periurban Nairobi, Kenya, shows three generations of continuity as well as dramatic changes in education, work, and state-level change. The roles of women in this community are a key to understanding these changes and their consequences for children and family life (Edwards and Whiting, 2004).

More generally, Therborn (2009) characterizes seven broad cross-cultural family patterns that clearly persist and influence anthropological studies of children and families today and will continue to do so for the foreseeable future: Christian European; Islamic West Asian/North African; South Asian Hindu; Confucian East Asian; sub-Saharan African; Southeast Asian; and Creole (U.S. South, Caribbean, Brazil, parts of South America). Each of these seven broad family patterns differs in social dimensions that have profound influences on the pathways of social development for children growing up within those family systems. These highly culturally variable dimensions of family systems include norms concerning inheritance (e.g., whether all children inherit equally, or only males, or only firstborn males) or descent rules (e.g., bilateral as in the United States or patrilineal) and marriage customs that are preferred or permitted in different communities (e.g., whether there is an ideal norm of lifetime monogamy; whether divorce and serial monogamy are allowed or plural marriage is permitted). Beliefs about sexuality and gender and patterns of household formation also vary systematically across these family regions (i.e., whether couples form independent households, live with parents, or form joint or extended households; whether children typically remain in one household or move between multiple households during childhood). Influences of religious practices on families and many other norms, laws, and customs that shape family life show similar patterned variation, and so they deeply affect the developmental pathways of children and youths. As Therborn (2009) comments about the widely varying families around the world, and the lives of children in them, "The boys and girls of the world enter many different childhoods and depart them through many different doors" (p. 338). Along with globalization, the anthropological comparative study of powerful local and regional cultural differences in parenting, childhood, and family life across populations around the world will continue to not only be important, but also to provide enduring scientific questions.

See also: Cross-Cultural Research Methods in Sociology; Early Emotional Development and Cultural Variability; Moral Development, Cultural Differences In; Parenting Attitudes and Beliefs across Cultures; Play, Anthropology of; Pretend Play and Cognitive Development; Prosocial Behavior, Effects of Parenting and Family Structure On; Self and Identity Development During Adolescence across Cultures; Sex Selective Abortion; Sociolinguistics; Street Children: Cultural Concerns; Youth Culture, Anthropology of.

Bibliography

Arnett, J.J., 2011. Emerging adulthood(s). The cultural psychology of a new life stage. In: Jensen, L.A. (Ed.), Bridging Cultural and Developmental Approaches to Psychology. New Syntheses in Theory, Research and Policy. Oxford University Press, New York, pp. 255–274.

Barlow, K., 2004. Critiquing the good enough mother: a perspective based on the Murik of Papua New Guinea. Ethos 32 (4), 514–537 (Special issue: Contributions to a feminist psychological anthropology).

Bateson, G., Mead, M., 1942. Special Publications of the New York Academy of Sciences. Balinese Character: A Photographic Analysis, vol. 2. New York Academy of Sciences, New York.

Bernard, H.R., 1995. Research Methods in Anthropology: Qualitative and Quantitative Approaches. Alta Mira Press, Walnut Creek, CA.

Bogin, B., 1999. Patterns of Human Growth. Cambridge University Press, Cambridge.

Chisholm, J.S., 1999. Death, Hope, and Sex: Steps to an Evolutionary Ecology of Mind and Morality. Cambridge University Press, New York.

Cultural Survival Quarterly. World Report on the Rights of Indigenous Peoples and Ethnic Minorities, Cambridge, MA (http://www.cs.org).

D'Andrade, R., 1995. The Development of Cognitive Anthropology. Cambridge University Press, Cambridge.

Edwards, C.P., Bloch, M., 2010. The Whitings' concepts of culture and how they have fared in contemporary psychology and anthropology. Journal of Cross-Cultural Psychology 41 (4), 485–498.

Edwards, C.P., Whiting, B.B. (Eds.), 2004. Ngecha: A Kenyan Village in a Time of Rapid Social Change. University of Nebraska Press, Lincoln.

Ember, C.M., 1981. A cross-cultural perspective on sex differences. In: Munroe, R.H., Munroe, R.L., Whiting, B.B. (Eds.), Handbook of Cross-Cultural Human Development. New York, Garland, pp. 531–580.

Goldschmidt, W., 1992. The Human Career. The Self in the Symbolic World. Blackwell, Cambridge, MA.

Goodwin, M.H., 1990. He-Said-She-Said. Talk as Social Organization among Black Children. Indiana University Press, Bloomington.

Gottlieb, A., 2004. The Afterlife is Where We Come from: The Culture of Infancy in West Africa. University of Chicago Press, Chicago.

Greenfield, P.M., 2004. Weaving Generations Together: Evolving Creativity in the Maya of Chiapas. School of American Research, Santa Fe.

Harkness, S., Super, C.M. (Eds.), 1996. Parents' Cultural Belief Systems. Guilford Press, New York.

Hirschfeld, L.A., Gelman, S.A. (Eds.), 1994. Mapping the Mind: Domain Specificity in Cognition and Culture. Cambridge University Press, Cambridge.

Hrdy, S.B., 2009. Mothers and Others: The Evolutionary Origins of Mutual Understanding. Harvard Press, Cambridge, MA.

Ingstad, B., Whyte, S.R. (Eds.), 1995. Disability and Culture. University of California Press, Berkeley.

Keller, H., 2012. Attachment and culture. Journal of Cross-Cultural Psychology 44 (2), 175–194.

Konner, M., 2010. The Evolution of Childhood: Relationships, Emotion and Mind. Belknap Press of Harvard University Press, Cambridge, MA.

Korbin, J.E. (Ed.), 1981. Child Abuse and Neglect. Cross-Cultural Perspectives. University of California Press, Berkeley.

Lancy, D.F., 2007. Accounting for variability in mother-child play. American Anthropologist 109 (2), 273–284.

Lancy, D.F., 2008. The Anthropology of Childhood: Cherubs, Chattel, Changelings. University Press, Cambridge, MA.

Lancy, D.F., Bock, J., Gaskins, S. (Eds.), 2010. The Anthropology of Learning in Childhood. AltaMira Press, Walnut Creek, CA.

LeVine, R.A., 1988. Human parental care: universal goals, cultural strategies, individual behavior. In: LeVine, R.A., Miller, P.M., West, M.M. (Eds.), Parental Behavior in Diverse Societies. Jossey-Bass, San Francisco, pp. 3–12.

LeVine, R.A., 2007. Ethnographic studies of childhood: a historical overview. American Anthropologist 109, 247–260.

LeVine, R.A., Levine, S., Dixon, S., Richman, A., Leiderman, P.H., Keefer, C., 1994. Child Care and Culture: Lessons from Africa. Cambridge University Press, Cambridge.

LeVine, R.A., LeVine, S., Schnell-Anzola, B., Rowe, M.L., Dexter, E. (Eds.), 2012. Literacy and Mothering: How Women's Schooling Changes the Lives of the World's Children. Oxford, New York.

LeVine, R.A., Norman, K., 2001. The infant's acquisition of culture: early attachment reexamined in anthropological perspective. In: Moore, C.C., Mathews, H.F. (Eds.), The Psychology of Cultural Experience. Cambridge University Press, New York, pp. 83–104.

Levy, R., 1973. The Tahitians. Mind and Experience in the Society Islands. University of Chicago Press, Chicago.
Mageo, J., 1998. Theorizing Self in Samoa: Emotions, Genders, and Sexualities. University of Michigan Press, Ann Arbor.
Munroe, R.L., Munroe, R.H., 1994. Behavior across cultures: results from observational studies. In: Lonner, W.J., Malpass, R. (Eds.), Psychology and Culture. Allyn and Bacon, Boston, MA, pp. 107–111.
Ochs, E., 2002. Becoming a speaker of culture. In: Kramsch, C. (Ed.), Language Acquisition and Language Socialization. Ecological Perspectives. Continuum Press, London, pp. 99–120.
Ochs, E., Izquerdo, C., 2009. Responsibility in childhood: three developmental trajectories. ETHOS 37 (4), 391–413.
Ochs, E., Kremer-Sadlik, T. (Eds.), 2013. Fast Forward Family: Home, Work and Relationships in Middle Class America. University of California Press, Berkeley.
Otto, H., Keller, H. (Eds.), 2013. Different Faces of Attachment. Cambridge University Press, New York.
Quinn, N., Mageo, J. (Eds.), 2013. Cross-Cultural Challenges to Attachment Theory. Palgrave, New York.
Rogoff, B., 2011. Developing Destinies: A Mayan Midwife and Town. Child Development in Cultural Context. Oxford, New York.
Scheper-Hughes, N., 1992. Death without Weeping. The Violence of Everyday Life in Brazil. University of California Press, Berkeley, CA.
Schieffelin, B.B., Ochs, E. (Eds.), 1986. Language Socialization across Cultures. Cambridge University Press, Cambridge.
Schlegel, A., Barry, H., 1991. Adolescence: An Anthropological Inquiry. The Free Press, New York.
Seymour, S.C., 1999. Women, Family, and Child Care in India: A World in Transition. Cambridge University Press, Cambridge.
Shwalb, D.W., Shwalb, B.J., Lamb, M.E. (Eds.), 2013. Fathers in a Cultural Context. Routledge, New York.
Shostak, M., 1981. Nisa: The Life and Words of a !Kung Woman. Harvard University Press, Cambridge, MA.
Shweder, R.A., Bourne, E.J., 1991. Does the Concept of the Person Vary Cross-Culturally? In: Shweder, R.A. (Ed.), Thinking through Cultures: Expeditions in Cultural Psychology. Harvard University Press, Cambridge, MA, pp. 113–155.
Spiro, M.E., 1975. Children of the Kibbutz. Harvard University Press, Cambridge.
Stephens, S. (Ed.), 1995. Children and the Politics of Culture. Princeton University Press, Princeton.
Super, C.M., Harkness, S., 1997. The cultural structuring of child development. In: Berry, J., Dasen, P.R., Saraswathi, T.S. (Eds.), Handbook of Cross-Cultural Psychology, second ed. Allyn & Bacon, Boston, pp. 1–39.
Therborn, G., 2009. Family. In: Shweder, R.A., Bidell, T.R., Dailey, A.C., Dixon, S., Miller, P.J., Modell, J. (Eds.), The Child: An Encyclopedic Companion. University of Chicago Press, Chicago, pp. 333–338.
Tobin, J., Hsueh, Y., Karasawa, M., 2009. Preschool in Three Cultures Revisited: China, Japan, and the United States. University of Chicago Press, Chicago.
Turner, V.W., 1967. The Forest of Symbols. Cornell University Press, Ithaca, NY.
Weisner, T.S., 1996. Why ethnography should be the most important method in the study of human development. In: Jessor, R., Colby, A., Shweder, R.A. (Eds.), Ethnography and Human Development. Context and Meaning in Social Inquiry. University of Chicago Press, Chicago, pp. 305–324.
Weisner, T.S., 2011. Culture. In: Underwood, M.K., Rosen, L.H. (Eds.), Social Development: Relationships in Infancy, Childhood and Adolescence. Guilford, New York, pp. 372–399.
Weisner, T.S., Duncan, G., 2013. The world isn't linear or additive or decontextualized: pluralism and mixed methods in understanding the effects of antipoverty programs on children and parenting. Ch. 8, pp. #. In: Gershoff, E.T., Mistry, R.S., Crosby, D.A. (Eds.), Societal Contexts of Child Development: Pathways of Influence and Implications for Practice and Policy. Oxford University Press, New York.
Weisner, T.S., Gallimore, R., 1977. My brother's keeper: child and sibling caretaking. Current Anthropology 18 (2), 169–190.
Weisner, T.S., Lowe, E., 2005. Globalization and the psychological anthropology of childhood and adolescence. In: Casey, C., Edgerton, R. (Eds.), A Companion to Psychological Anthropology: Modernity and Psychocultural Change. Blackwell Publishers, Oxford, pp. 315–336.
Whiting, B., Edwards, C.P., 1988. Children of Different Worlds: The Formation of Social Behavior. Harvard University Press, Cambridge, MA.
Whiting, J., 1968. Socialization: anthropological aspects. International Encyclopedia of the Social Sciences 14, 545–551.
Whiting, J., Whiting, B., 1975. Children of Six Cultures: A Psychocultural Analysis. Harvard University Press, Cambridge, MA.
Yoshikawa, H., Weisner, T.S., Kalil, A., Way, N., 2008. Mixing qualitative and quantitative research in developmental science: uses and methodological choices. Developmental Psychology 44 (2), 344–354.

Childhood Cancer: Psychosocial Aspects

Meinolf Noeker and Franz Petermann, University of Bremen, Bremen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Medical progress has been able to make many childhood cancers nonfatal. Therapies are now available that turn cancer into a chronic condition or even cure it completely, even if lasting impairments may remain. In Europe and North America, about 80% of all patients with cancer in childhood and adolescence survive their leukemia, lymphomas. or tumors. The improved treatment options and changed ethical approaches have led to a consensus that the child should be told the life-threatening nature of the disease in an honest and age-appropriate way. Psychological research does not focus on suspected psychosomatic causes of childhood cancer manifestation but on the stress impact imposed by disease and treatment and the transactional coping and adaptation process in the child and family. The patient's and the family's adaptation to the challenges imposed by disease and treatment depends on the balance of stress impact, on the one side, and available coping competencies and resources, on the other side. A psychological intervention program is introduced that focuses on counseling and supporting the family across different treatment phases (initial phase of diagnosis; intensive oncological treatment; long-term remission of the disease and reintegration of the child versus, in case of a negative disease course, recurrent relapses, and final death). Key interventions refer to the enhancement of open and clear family communication including mutual exchange of personal needs and concerns and the practical organization of day-to-day routines as well as an effective communication with the healthcare team or with the social network. During the last decade, psychological research has shifted toward the long-term psychological outcome in survivors of pediatric cancer patients. The risk of neuropsychological impairments, psychopathological comorbidity, and potential impairments of health-related quality of life have become complementary parameters for conceptualization and evaluation of modern oncological treatment protocols. A conceptual framework is introduced that delineates mediators of adaptation process and adjustment outcome covering a continuum of emergence of psychopathology (e.g., posttraumatic stress disorder (PTSD)), on the one side versus occurrence of developmental benefit and resilience, on the other side.

Substantial progress in cancer treatment (chemotherapy, radiotherapy, surgery, bone-marrow transplantation) has dramatically improved the long-term survival rates of children and adolescents with the diagnosis of cancer (malignant tumors, lymphoma, and leukemia). Some decades ago, the only probable outcome without any alternative for the family was to face the death and loss of their child. As nowadays more than two-thirds of the young patients survive, cancer in children and adolescents no longer is regarded as fatal but as a chronic disease, which can result in complete cure, death, or survival with potential physical, neurological, neuropsychological, psychosocial, and psychopathological impairments.

To date, a broad consensus is established on telling the child honestly the life-threatening nature of the disease. The statistical survival rates, however, do not determine the particular prognosis in the individual child, turning the former threat of certainty of death into the threat of uncertainty of the very personal fate. The term 'Damocles syndrome' characterizes the psychological situation of the family that has completed oncological therapy but cannot be sure to have definitely defeated the disease.

Psychological Adaptation to Childhood Cancer as a Balance of Stress and Resources

The manifestation of childhood cancer is not associated with any known predisposing social or psychological feature in the child or in the family. Its presence can be regarded a randomly occurring highly stressful life event for 'normal' families. As a consequence, psychological research and intervention in pediatric oncology is not concerned with the detection of suspected 'psychosomatic' causes of cancer but with an understanding of the 'biopsychological' stress impact imposed by cancer diagnosis on the child's and family's psychological functioning and well-being.

The risk for the child or the family to master or to fail in the turmoil of the cancer experience may be compared to a balance with one scale pan containing all the stressors resulting from the disease and the other scale pan containing all the resources, competencies, and sources of social support to handle the stress. In some vulnerable families, strong stressors and demands clearly outweigh only weak resources, thus increasing the probability of emotional suffering and behavioral disturbance in the child as well as the risk of marital discord or parental depression. Other families are characterized by resilience, i.e., the available resources are strong enough to outweigh the threatening disease impact, thus giving rise to good adjustment and a personal sense of mastery. The large majority of patients succeed in mastering the plethora of disease- and treatment-related stressors and challenges (Long and Marsland, 2011).

Enhancement of Effective Coping with Childhood Cancer by Family Counseling

Understanding adaptation to childhood cancer as a psychological balance between burden and resources can serve as a basis for effective psychological counseling strategies. Therapeutic enhancement of adaptation includes two major approaches: decrease stress impact and increase coping

competencies and resource availability. In order to administer individually tailored counseling strategies, a broad knowledge is required on the most common and threatening source of stress, on the one hand, and of the most powerful resource that can be mobilized, on the other hand. Typical stressors imposed on the family and child are presented in Table 1 covering practical, emotional, social, and existential domains. As a counterpart, Table 2 shows resources that can buffer and diminish cancer stress impact, improve adaptation, stabilize quality of life, and protect from mental health disturbances for both individual family members and the entire family system. Research has highlighted the key role of family communication behavior for adjustment outcome at the individual and system level. Intensity and quality of communication between the parents about disease issues is closely correlated with an improved parent–child contact as well as with a high disease-related information level in the child and the siblings. A fair and respectful sharing of the cancer-related burden and tasks is correlated with a high family coherence and stability suggesting mutual pathways (Noeker et al., 1990).

Table 1 Typical stressors, demands, and challenges frequently occurring during the course of childhood cancer

Organization of treatment (meetings with physicians, insurances, etc.)
Financial expenditures
Gathering information on cancer and its treatment
Repeated hospitalizations including the alternative to stay at home with the siblings or be in the hospital with the ill child
Treatment-related distress according to painful and frightening medical procedures (e.g., bone marrow aspirations) and treatment side-effects (e.g., nausea and vomiting due to chemotherapy)
Reduced time for recreation and holiday
Role conflict (e.g., father between ongoing job demands and family need for support)
Reexamination of future perspectives (job perspectives of the mother; schooling of the child)
Sharing responsibility for treatment decisions (e.g., amputation in the case of bone sarcoma)
Subconscious violation, disappointment, frustration of the hope for a normal and healthy child
Religious doubts
Uncertainty of final outcome; fear of relapse, fear of loosing the child

Source: Noeker, M., Petermann, F., 1990. Treatment-related anxieties in children and adolescents with cancer. Anxiety Research 3, 101–111; Noeker, M., Petermann, F., Bode, U., 1990. Family counseling in childhood cancer: conceptualization and empirical results. In: Schmidt, L.R., Schwenkmezger, P., Weinman, J., Maes, S. (Eds.), Theoretical and Applied Aspects of Health Psychology. Harwood Academic Publishers, Chur, pp. 241–253.

Table 2 Protective family resources that enhance an effective adaptation to childhood cancer

Communication competencies
Attributing positive meanings to the situation
Commitment to the family as a unit
Engaging in active, flexible coping efforts
Balancing illness demands with other family needs
Maintaining social integration
Developing collaborative relationships with professionals
Ability to ask for and accept support
Clear religious or philosophical belief systems
Parenting style demonstrating consistent rules and expectations

Sources: Kazak, A.E., Brier, M., Alderfer, M.A., Reilly, A., Fooks, P.S., Rogerwick, S., Ditaranto, S., Barakat, L.P., 2012. Screening for psychosocial risk in pediatric cancer. Pediatric Blood Cancer 59, 822–827; Koocher, J.E., O'Malley, G.P., 1981. The Damocles Syndrome: Psychosocial Consequences of Surviving Childhood Cancer. McGraw-Hill, New York, London. Kupst, M.J., Natta, M.B., Richardson, C.C., Schulman, J.L., Lavigne, J.V., Das, L., 1995. Family coping with pediatric leukemia: 10 years after treatment. Journal of Pediatric Psychology 20, 601–617; Long, K.A., Marsland, A.L., 2011. Family adjustment to childhood cancer: a systematic review. Clinical Child and Family Psychology Review 14, 57–88; Noeker, M., Petermann, F., 1990. Treatment-related anxieties in children and adolescents with cancer. Anxiety Research 3, 101–111; Patterson, J.M., 1991. Family resilience to the challenge of a child's disability. Pediatric Annals 20, 491–499.

Psychological Counseling Strategies and Interventions across Different Phases of Childhood Cancer Treatment

Psychological counseling programs have been developed based on the knowledge of typical stress and resource factors in families with childhood cancer (Kazak et al., 2007, 2012; Noeker et al., 1990). The particular counseling interventions applied are specifically tailored taking into account the specific stressors that characterize certain phases of the treatment course. Counseling sessions may take place at the clinic or at home. While all family members should join them, there may also be an indication for sessions with family subsystems, e.g., when mother and father want to find out an agreement on consistent parenting behavior. In the following, selected topics and strategies of a psychological counseling program are delineated following the major treatment phases.

Supporting the Initial Coping with Cancer Diagnosis

After being told the diagnosis of childhood cancer, most parents show strong affective reactions like shock, sadness, anger, and despair. The primary task of the counselor here is to support awareness and acceptance of these emotional reactions via verbalization and clarification. Idiosyncratic and biased lay theories of the parents on the reasons why exactly one's own child has cancer have to be resolved as they may interfere with the medical treatment plan in the long run. While exploring subjective disease and treatment concepts, special attention has to be paid to frequent parental beliefs of being personally responsible for and feelings of guilt about their child's cancer (e.g., via genetics, bad food, psychological stress). Cognitive-emotional restructuring of inappropriate disease concepts enhances the parents' cognitive capacity for understanding and accepting the oncological treatment concept.

The next major counseling task is to initiate an active, mutually respectful, and supportive coping attitude and behavior within the family system including:

1. clarifying the kind of support every person in the family needs most from other family members for stabilizing outer and inner equilibrium;
2. encouraging open communication and frank expression of personal feelings, concerns, sorrows, and anxieties among the family members; and
3. promoting an attitude and commitment to approach the new burden and tasks as a shared challenge for the entire family.

From a 'Protective' to an Open Approach

Some decades ago, a so-called 'protective' approach was preferred based on the notion not to disclose the life-threatening nature of the disease to the child. In the meantime, not only ethical considerations but also progress in oncological treatment strategies has led to a broad consensus that there is no responsible alternative to telling the truth even to preschool children. Today, medical progress makes it possible to counterbalance the bad news of cancer diagnosis with the good news of profound hope for cure. Therefore, telling children the truth includes two realistic key messages: first, that death has to be expected if no treatment is performed, and second, that even the treatment does not necessarily guarantee survival. Information given to the child should be appropriate to their cognitive and emotional age and maturity. In small children, showing concrete treatment procedures and instruments and offering plausible explanations for sensory experiences like vomiting during chemotherapy is more effective than abstract explanations about malignant cells. An approach that avoids telling the truth implies serious medical and psychological 'side-effects' to be considered:

1. A child will not accept the enormously distressing treatment procedures without the awareness of the serious consequences in case of treatment refusal. Thus, true information is necessary for compliance with treatment, and compliance with treatment is necessary for medical treatment success.
2. Children who lack plausible information on their condition tend to develop other explanations for the occurrence of their disease (Eiser, 1993). Such a frequently reported belief is that the illness represents a punishment for one's own misbehavior.
3. Trust in the parents is one of the most important psychological resources for the child to protect from feelings of loneliness, separation, and depression. A child uncovering the parents' well-meant lies will not only have to struggle with the adversities of disease and treatment but, in addition, with parental alienation.

Phase of Intensive Oncological Treatment

After the family has overcome the initial shock of cancer diagnosis, psychological counseling should assess vulnerability and resource factors in the child and in the family. Premorbid factors of vulnerability that were already operating during the time before the diagnosis (e.g., behavioral or developmental disorders in the child, intense sibling rivalry, marital discord, inconsistent parenting styles, socioeconomic disadvantages) may interfere with a competent adaptation to demands of disease and treatment. Assessment of premorbid risk factors, therefore, improves the chances for early, preventive and individualized intervention. At least as important as the assessment of risks and deficits, however, is the exploration of strengths, skills, competencies, and resources. If a counselor turns the family's attention to positive experiences of crises that were resolved successfully during the past, problem-solving strategies may be reactivated that have proven specific adaptability in this particular family. Past experience of successful mastery strengthens self-efficacy also toward cancer challenges. Psychological counseling during intensive oncological treatment includes:

1. encouraging active participation in the management of the illness;
2. defining clear and practical responsibilities for the day-to-day management of additional chores and tasks; making sure that agreements are perceived as a fair sharing of the burden by every family member;
3. parenting issues (e.g., ways to show affection without spoiling and pity);
4. mobilization and acceptance of concrete social support from neighborhood or classmates; and
5. facilitating honest communication between family and healthcare team.

Negative Disease Course: Recurrent Relapses, End-Stage of Disease, Death, and Dying

A new diagnosis of recurrent cancer after a time of suspected remission requires similar accepting and verbalizing counseling strategies concerning the intense emotional reactions that have been experienced during the encounter of being informed of the diagnosis. Now the focus of counseling may shift toward the ambivalent tendencies toward continuation versus discontinuation of oncological treatment as the parents may wish to maintain any chance for cure, on the one hand, and may not want to see the child suffer from a stressful but potentially futile treatment, on the other. With the worsening of the child's condition, the threat of loss of the child is gradually intensified leading to the following counseling topics:

1. Preparing parents for communication with the child on death and dying. Finding individually appropriate images integrating truth and hope.
2. Searching for phantasies on a good way to spend the remaining time together; development of other important goals in addition to survival like freedom from pain.
3. Enhancing imagery on a suitable way of parting in dignity.

If the child actually dies, it is important to say that the treatment failed and not the child or the family. The family's engagement shown for the child should be acknowledged. Individually suitable mourning rituals may be conceived that combine keeping related to the deceased child and, at the same time, letting them go. Family members may be encouraged to take their time for mourning and grief according to their very personal needs not according to expectations of their social network (*see* Bereavement).

In summary, many counseling interventions focus on the enhancement of open and clear family communication. The mobilization of this key resource for effective coping with the burden of childhood cancer may refer to the mutual exchange of personal needs and concerns, the practical organization of day-to-day routines, or the subtle communication between parents and child about their chances of survival. Effective extrafamilial communication may refer to establishing cooperation and agreements with

the staff or to mobilizing and accepting support from the extended social network.

Positive Course: Long-Term Survival and Reintegration

After termination of intensive oncological therapy, anxieties remain in patients and their families concerning a return of the disease, which would require a completely new course of therapy with, however, decreased chances for cure. Families suffer from intrusive thoughts and chronic sorrow about potential cancer relapse comparable to symptoms in Generalized Anxiety Disorder. Intervention strategies based on Cognitive-behavioral Therapy (CBT) and on Acceptance- and Commitment-based Therapy (ACT) can support the patient and family members to observe these recurrent thoughts consciously, to accept them as a natural reaction to an abnormal situation, and at the same time, not to get too much involved every time anxious cognitions emerge. Counselor and family may cautiously anticipate and go through possible responses and options for the case of an actually occurring relapse.

Termination of therapy also requires encouragement of the child to give up their sick role and corresponding privileges and to promote social reintegration in previous roles, groups, and obligations.

Long-Term Adjustment Outcome in Childhood Cancer Survivors

During the last decade, a number of patient registries, multicenter studies and *metaanalyses* have investigated the interaction of disease- and treatment-associated risk factors with sociodemographic and psychosocial vulnerability factors in long term survivors of childhood cancer. The Childhood Cancer Survivor Study (CCSS) funded by the National Cancer Institute represents the largest and most comprehensively characterized group of childhood cancer survivors in North America (Leisenring et al., 2009; Zeltzer et al., 2009). The sample includes 7147 (77%) survivors and 388 (78%) siblings. Whereas adult survivors of childhood cancer from the CCSS report more symptoms of psychological distress compared with siblings, both groups report better psychological adjustment than that found in population norms. Survivors report poorer health-related quality of life (HRQOL) in physical and social but not in mental health domains compared with siblings. Compared with U.S. population norms, survivors have increased risk of poor HRQOL in physical and social domains, with the largest effect size in vitality. Previous treatment with cranial radiation and surgery are associated with poor HRQOL, particularly in physical domains. Participants diagnosed <10 years of age or treated in the past 25 years report better HRQOL than those diagnosed >10 years of age or treated >30 years ago. Several studies have indicated that brain tumors and treatment strategies including central nervous system (CNS) surgery, cranial radiotherapy, and intrathecal chemotherapy carry an increased risk for neurological and neuropsychological long-term outcomes, which in turn also threaten the patients' psychosocial and vocational participation.

Developmental Benefit

Threatening and distressing experiences such as the life-threatening condition of cancer do not inevitably exert a detrimental impact on psychological functioning on the contrary, may also enhance personal mental growth and development gains, which predispose to an increased resilience to future crises and challenges. Currier et al. (2009) have developed a questionnaire that reflects both the stressful and beneficial side of pediatric cancer (Revised Benefit/Burden Scale for Children, BBSC). The instrument separates psychometrically two very distinct factors, the perceived stress versus the perceived benefits. Both 10-item scales show excellent internal consistency. In some patients, both scales are highly increased indicating that the very dramatic, vital experience is not simply processed as only positive or negative. Rather, the impressive challenges of disease and treatment include both an intense stress impact and perception of a meaningful life experience. Maybe in some patients the very existential confrontation with the risk of death and dying acts as a precursor for a final strengthening of basic assumptions concerning ultimate personal safety, protection, self-efficacy, and confidence. Studies highlighting positive long-term outcome, however, have been questioned as survivors of cancers may be biased reporters and tend to deny difficulties and overestimate their positive health and satisfaction.

Adjustment between Dysfunction and Resilience

Despite the methodological progress achieved by large multicenter trials and *metaanalysis*, psychological research continues to face the task to disentangle the mediators and moderators that explain the enormous interindividual variation in psychological outcome in childhood cancer survivors, even in the case of high similarity in oncological conditions. **Figure 1** provides a heuristic developmental model that depicts factors resulting in a spectrum of unfavorable versus favorable developmental pathways and adjustment outcomes. As the most unfavorable outcome mental disorders may emerge, often in the form of PTSD, as well as depression or anxiety disorders. The next outcome category refers to patients with physical, cognitive, emotional, functional, or social impairments concerning HRQL, but who do not fulfill the criteria for any specific mental disorder. The next category refers to the large group of patients that do not show differences in psychological outcome measures compared to healthy controls. At the most favorable end of the spectrum, there is the subgroup of patients that show developmental benefits and increased resilience.

As illustrated in **Figure 1** the adjustment outcome in the individual case depends on the interaction of disease- and treatment-related factors, on the one hand, and psychological factors of premorbid vulnerability and patterns of the transactional process of coping with disease- and treatment-related stressors, on the other hand. The individual pattern of premorbid cognitive-emotional schemata and behavioral skills predisposes to a higher or lower coping capacity and,

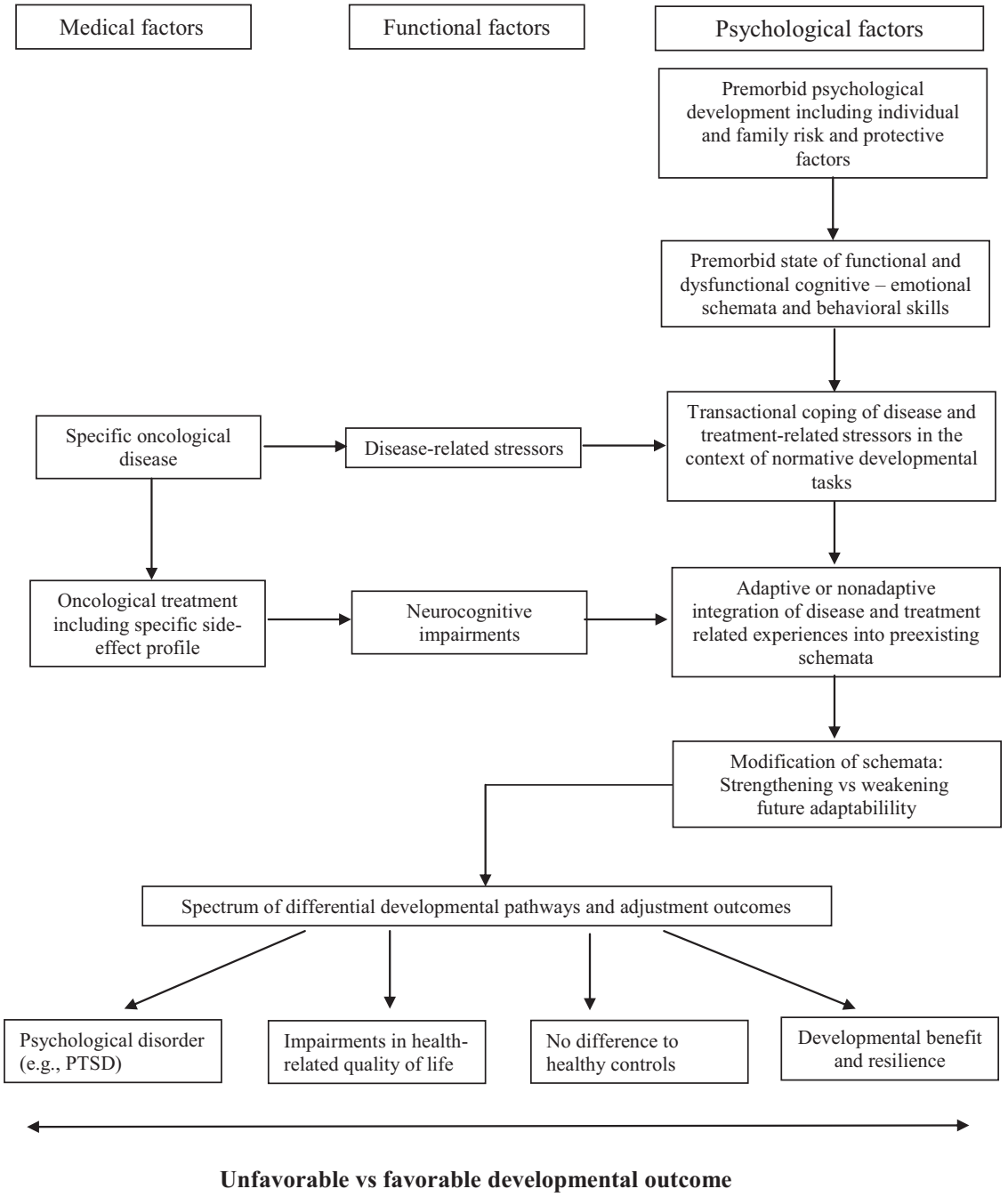

Figure 1 Differential developmental paths and outcomes in survivors of childhood cancer.

finally, coping success. Cognitive appraisals and coping behavior may be altered by neuropsychological impairments acquired by brain tumor or CNS surgery, radiation, or chemotherapy.

A key assumption underlying this model is that of a continuity or even enhancement of premorbid adjustment levels by the impact of living with the cancer experience. Patients with a high level of premorbid resources and coping skills are likely to take a developmental path toward uncompromised mental health or even developmental gain. Patients with a premorbid vulnerability, however, carry an increased risk of coping failure and negative (traumatic, depressive, anxious) reappraisals and conclusions drawn from the cancer experience. Negative inferences promote dysfunctional belief systems increasing the risk of adjustment failure and final psychopathology.

See also: Cancer Screening; Cancer: Psychosocial Aspects; Chronic Illness, Psychosocial Coping with; Health Psychology

in Childhood; Illness and Dyadic Coping; Pain, Health Psychology of; Social Support and Recovery from Disease and Medical Procedures.

Bibliography

Chesler, M.A., Barbarin, O., 1987. Childhood Cancer and the Family. Meeting the Challenge of Stress and Support. Brunner/Mazel, New York.

Currier, J.M., Hermes, S., Phipps, S., 2009. Brief report: children's response to serious illness: perceptions of benefit and burden in a pediatric cancer population. Journal of Pediatric Psychology 34, 1129–1134.

Eiser, C., 1993. Growing up with a Chronic Disease: The Impact on Children and Their Families. Kingsley, London.

Kazak, A.E., Brier, M., Alderfer, M.A., Reilly, A., Fooks, P.S., Rogerwick, S., Ditaranto, S., Barakat, L.P., 2012. Screening for psychosocial risk in pediatric cancer. Pediatric Blood Cancer 59, 822–827.

Kazak, A.E., Rourke, M.T., Alderfer, M.A., Pai, A., Reilly, A.F., Meadows, A.T., 2007. Evidence-based assessment, intervention and psychosocial care in pediatric oncology: a blueprint for comprehensive services across treatment. Journal of Pediatric Psychology 32, 1099–1110.

Koocher, J.E., O'Malley, G.P., 1981. The Damocles Syndrome: Psychosocial Consequences of Surviving Childhood Cancer. McGraw-Hill, New York, London.

Kupst, M.J., Natta, M.B., Richardson, C.C., Schulman, J.L., Lavigne, J.V., Das, L., 1995. Family coping with pediatric leukemia: 10 years after treatment. Journal of Pediatric Psychology 20, 601–617.

Kurtz, B.P., Abrams, A.N., 2010. Psychiatric aspects of pediatric cancer. Child and Adolescent Psychiatric Clinics of North America 19, 401–421.

Leisenring, W.M., Mertens, A.C., Armstrong, G.T., Stovall, M.A., Neglia, J.P., Lanctot, J.Q., Boice Jr., J.D., Whitton, J.A., Yasui, Y., 2009. Pediatric cancer survivorship research: experience of the Childhood Cancer Survivor Study. Journal of Clinical Oncology 27, 2319–2327.

Long, K.A., Marsland, A.L., 2011. Family adjustment to childhood cancer: a systematic review. Clinical Child and Family Psycholy Review 14, 57–88.

Noeker, M., Petermann, F., 1990. Treatment-related anxieties in children and adolescents with cancer. Anxiety Research 3, 101–111.

Noeker, M., Petermann, F., Bode, U., 1990. Family counselling in childhood cancer: conceptualization and empirical results. In: Schmidt, L.R., Schwenkmezger, P., Weinman, J., .Maes, S. (Eds.), Theoretical and Applied Aspects of Health Psychology. Harwood Academic Publishers, Chur, pp. 241–253.

Noll, R.B., Gartstein, M.A., Vanatta, K., Correl, J., Bukowski, W.M., Davies, W.H., 1999. Social, emotional, and behavioral functioning of children with cancer. Pediatrics 103, 71–78.

Patterson, J.M., 1991. Family resilience to the challenge of a child's disability. Pediatric Annals 20, 491–499.

Zeltzer, L.K., Recklitis, C., Buchbinder, D., Zebrack, B., Casillas, J., Tsao, J.C., Lu, Q., Krull, K., 2009. Psychological status in childhood cancer survivors: a report from the Childhood Cancer Survivor Study. Journal of Clinical Oncology 27, 2396–2404.

Childhood Memories

Harlene Hayne and Damian Scarf, University of Otago, Dunedin, New Zealand
Kana Imuta, University of Queensland, St Lucia, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

During infancy and early childhood, what we encode and remember is stored, retained, and retrieved, largely without the benefit of language, but during the rest of our lives, our memories are established, maintained, and expressed using our extensive language networks. In this article, we will introduce a few experimental procedures that can be used to study the transition between nonverbal and verbal memory processing during early childhood and discuss the long-term fate of early childhood memories.

Humans are the only animals who, over the course of their lifetime, make the transition from encoding their memories in a nonverbal format to a format that is primarily based on language. During infancy and early childhood, what we encode and remember is stored, retained, and retrieved, largely without the benefit of language, but during the rest of our lives, our memories are established, maintained, and expressed using our extensive language networks. What happens during this important transition from one representational format to the other? Are we eventually able to translate some of our early preverbal memories into words as we acquire our native language? And, if we can, under what conditions might this translation occur?

Historically, research on memory development has typically been conducted in two distinct ways – researchers who study infants and young children have relied exclusively on nonverbal learning and memory techniques, while researchers who study older children, adolescents, and adults have relied primarily on verbal procedures instead. The goal of this article is to explore the transition between nonverbal and verbal memory processing during early childhood. Our specific aims are twofold. First, we will compare children's nonverbal and verbal memories of the same event. Second, we will review what we know about the fate of memories that were originally encoded largely without language.

Nonverbal Memory Procedures

Given human infants' preverbal nature, the study of memory processing by them has relied exclusively on nonverbal procedures. In the past, researchers have used the visual paired-comparison (VPC) paradigm (e.g., Fagan, 1974; Fantz, 1964; Rose, 1981), operant conditioning techniques (e.g., Hartshorn and Rovee-Collier, 1997; Hartshorn et al., 1998; Rovee-Collier et al., 1980), and imitation tasks (e.g., Barr et al., 1996; Bauer and Shore, 1987; Meltzoff, 1985) to study memory development beginning from birth to the age of 3 (and sometimes even older). In these tasks, nonverbal instructions are built into the procedure such that participants know what to do in the absence of verbal instructions, and participants can 'show' rather than 'tell' us what they remember. The results obtained using these three nonverbal tasks are highly consistent. Across all three, retention increases as a function of age, the range of effective retrieval cues increases as a function of age, and memories that appear to be forgotten can be reactivated and expressed given the appropriate retrieval cues (for a review, see Hayne, 2004).

Verbal Memory Procedures

As children begin to acquire language, researchers typically employ verbal memory procedures similar to those used with older children and adults. In these studies, children are asked to 'tell' rather than to 'show' what they remember. Katherine Nelson was one of the first researchers to ask children about familiar, past experiences (for a review, see Nelson, 2006). Since her pioneering studies, verbal interview procedures have been used extensively with preschool-aged children who have been asked to recall both repeated (e.g., Fivush, 1984; Hudson and Nelson, 1986) and unique events (e.g., Farrant and Reese, 2000; Fivush et al., 1984, 1987; Gross and Hayne, 1998). Collectively, these studies have shown that children begin to talk about the past beginning at about 18–24 months of age, the length and organization of their past-event narratives increase as a function of age, and, as with measures of nonverbal memory, the duration of retention increases as a function of age (for a review, see Nelson and Fivush, 2004).

Despite the wealth of knowledge that we have accumulated about nonverbal and verbal memory development during infancy and early childhood, what we lack is a clear understanding of whether memory processes change when children make the transition from one encoding format to another and, if change does occur, what the nature of that change might be. To address this issue, researchers have recently developed a number of 'blended' memory techniques that incorporate both verbal and nonverbal tests. In this article, we will review some of the studies that have used these blended memory procedures and present findings that provide key insights about the continuities and discontinuities in memory development during the transition from infancy to childhood.

Blended Memory Procedures

Hide-and-Seek Task

The hide-and-seek task was originally developed by DeLoache and Brown (1979) as a measure of nonverbal memory in very

young children. At the time the task was developed, work by Rovee-Collier and her colleagues had already shown that infants as young as 3 months of age exhibited retention after delays as long as several days when tested using the mobile-conjugate reinforcement paradigm (e.g., Rovee-Collier et al., 1980, 1981). Ironically perhaps, comparable evidence for long-term memory of toddlers and young children was extremely limited (e.g., Daehler et al., 1976; Horn and Myers, 1978). Importantly, the hide-and-seek task provided some of the first evidence for memory over periods longer than just a few seconds in very young children.

In the basic hide-and-seek task, children are told that a stuffed toy (e.g., Big Bird) is going to hide, and they are encouraged to remember where it is hidden so that they can find the toy later. The child then watches the experimenter hide the toy in a specific location in their house (e.g., under the bed). After a delay of approximately 5 min, during which the child and the experimenter engage in another activity, the child is then asked to retrieve the hidden toy. Using this procedure, DeLoache and her colleagues found that children as young as 18 months of age have robust memories for the location of hidden objects (DeLoache and Brown, 1979, 1983, 1984; DeLoache et al., 1985).

Although the original hide-and-seek task was used as a purely nonverbal measure of young children's memory, more recently, we have extended the task to include both nonverbal and verbal measures. In addition, with the aim of using the hide-and-seek task as a measure of children's episodic memory, we modified the task such that it would allow us to test children's recall of the three W's of episodic memory (Tulving, 2002): the what, when, and where of the hiding event. In our modified version of the hide-and-seek task, the experimenter first familiarizes the child with seven stuffed toys (Donald Duck, Ronald McDonald, Mickey Mouse, Eeyore, Bert, Bob the Builder, and The Count). The child then chooses three toys and accompanies the experimenter to hide each of the three toys in a specific location in separate rooms around his or her house. When entering each room, the experimenter asks the child to name the room and to watch carefully as she hides the toy. Once all three toys are hidden, the child and the experimenter return to the living room and read two books during a 5-min delay to prevent children from rehearsing the hiding locations. Immediately following the delay interval, the verbal recall test is administered. In this test, the child is asked, "Which room did we go into first (or next)?" then "Who was hiding there?" and "Where did we hide him?" The answers to these questions allow us to assess children's verbal memory of the what (e.g., Eeyore), where (e.g., behind the curtain in the bedroom), and when (e.g., the last room entered) of each hiding event. Once the child has answered all of the questions on the verbal recall test or indicates that he or she can recall no more, the nonverbal recall test is administered. In the nonverbal recall test, the experimenter asks the child to show the experimenter the room that they went to first (or next) and to retrieve the toy in each room.

In this modified version of the hide-and-seek task, 4-year-old children verbally recalled significantly more information about the what (i.e., toy), where (i.e., room and specific location), and when (i.e., order of room visit) of the hiding event than did 3-year-old children. In contrast, on the nonverbal test, the 3-year-olds recalled as much information as the 4-year-olds regarding where the toys were hidden. In fact, their nonverbal memory for information about both the room and the specific location within that room (e.g., under the bed) was almost at ceiling and exceeded their verbal memory for the same information by almost a factor of 2. In contrast, 3-year-olds still struggled, even on the nonverbal test, to accurately indicate when they visited each room; 4-year-olds were significantly more accurate at recalling the when information than the 3-year-olds. (We did not test children's recall of the *what* information on the nonverbal test because children are not able to express their memory for the toy behaviorally.)

Taken together, while the findings from the verbal test, on its own, suggest that 3-year-olds experience great difficulty in remembering the what, where, and when of the hiding event, the findings from the nonverbal test indicate that 3-year-olds do have a robust memory of the where information. Relying on verbal tests alone underestimated the amount of information they retained. Furthermore, our findings suggest that the key difference between 3- and 4-year-olds' memory of this particular event may, more specifically, be their ability to remember the temporal sequence of when each object was hidden (Hayne and Imuta, 2011). This finding supports the notion that, of the three W's of event memory, the temporal aspect of the event may be the most difficult to recall (e.g., Clayton et al., 2003; Suddendorf and Corballis, 2007; Tulving, 2005). By blending verbal and nonverbal tests, the hide-and-seek task allows us to gain a better insight into the similarities and differences in children's memory of events (defined by the three W's) between 3 and 4 years of age and underscores the fact that, even once they begin to talk, children may continue to rely heavily on nonverbal forms of representation.

Magic Shrinking Machine

The magic shrinking machine is a task that can be mastered by 2-year-olds, yet it is still fascinating and challenging for 4-year-olds (Simcock and Hayne, 2002, 2003). In this task, children insert large objects into the machine and, by performing a series of actions, create smaller but otherwise identical versions of the same object. In order to make the machine 'work,' children turn on the machine by pulling a lever to activate the lights on the front panel. They then place an object (e.g., a teddy bear) into an opening at the top and turn a handle on the side. Turning the handle produces a series of unique sounds. Once the sounds stop, the child can retrieve an identical but 'shrunken' toy from a door on the front of the machine.

In one of the original experiments with the magic shrinking machine, 2- to 4-year-old children were visited in their homes and were shown how to operate the machine (Simcock and Hayne, 2003). During this visit, the child's caregiver completed the MacArthur Communicative Development Inventory: Words and Sentences (CDI) and a checklist of words that were specifically associated with the magic shrinking machine event (e.g., 'handle' and 'ball'). The following day, the experimenters returned to the child's home to test the child's memory of the magic shrinking machine event. During the test, children's memory was assessed using verbal free recall, nonverbal photograph identification, and behavioral reenactment tests.

Although children's verbal recall increased as a function of age, it lagged considerably behind both their nonverbal memory and their general verbal skills. During the verbal interview, children failed to report aspects of the event that they clearly remembered on the basis of their nonverbal memory performance, even though they had the relevant words in their vocabulary. For example, 24-, 27-, and 30-month-olds reported almost nothing during the verbal interview, yet they could, depending on their age, accurately identify 2–6 of the target photos (maximum = 8) and reproduced 2–4 of the target actions (maximum = 5). Children older than 33 months of age generally had most of the 22 target words as part of their vocabulary, and they approached ceiling on the nonverbal memory measures; however, their verbal reports were still extremely lean, and they reported less than half of what they could have reported based on their productive vocabulary alone (Simcock and Hayne, 2003). Again, these findings underscore the fact that, even as fluent language speakers, young children still rely extensively on nonverbal forms of representation.

The Long-Term Fate of Preverbal Memories

The research reviewed thus far has clearly shown that children's ability to use language in the service of memory lags considerably behind their ability to use language to discuss the here and now. Thus, if we rely exclusively on what children can tell us about past events, we are likely to underestimate what they actually remember. In contrast, children's nonverbal memory skills emerge very early in development (for a review, see Hayne, 2004) and continue to exceed their verbal memory skills at least through early childhood (and possibly beyond). What happens to these early nonverbal representations over time? Do they disappear as children make the final switch to a primarily verbal coding format, or are some of these memories retained and translated into language over time?

In an initial attempt to answer these questions, Simcock and Hayne (2002), tested independent groups of 27-, 33-, and 39-month-old children for their memories of the magic shrinking machine event after a 6-month or 1-year delay. In that study, children took part in two demonstration sessions that took place on two consecutive days. The procedure for each demonstration session and for the test session was identical to that used by Simcock and Hayne (2003). Consistent with the data obtained when children were tested after a 24-h delay (Simcock and Hayne, 2003), children's performance on the verbal recall test lagged behind their performance on the two nonverbal tests when they were tested after a 6-month or 1-year delay. Even after the long delays, children, on average, demonstrated robust nonverbal memory of the magic shrinking machine event, recalling over 60% of the information on the behavioral reenactment test and 50% on the photograph identification test. In contrast, they reported only about 20% of the information during the verbal recall test. Most importantly, even though all of the children had acquired the majority of the target words as part of their productive vocabulary by the time of the 6-month and 1-year interviews, there was not a single instance in which their verbal reports of the event contained a target word that had not been part of their productive vocabulary at the time of encoding. That is, the children in Simcock and Hayne's (2002) study failed to translate their preverbal memories into language – instead preserving those memories in their original preverbal format.

Evidence of Verbal Translation of Preverbal Memories

Although Simcock and Hayne (2002) found no evidence for the translation of memories into language, there is at least some anecdotal evidence that such translation occurs. For example, when older children and adults are asked to recall their earliest personal memory, a handful report events that occurred around or before their second birthday (e.g., Eacott and Crawley, 1998; Jack and Hayne, 2007; Jack et al., 2009; MacDonald et al., 2000; Tustin and Hayne, 2010). In light of Simcock and Hayne's (2002) findings, what are we to make of these reports? On the one hand, given their retrospective nature, it is entirely possible that these very early accounts are not reliable – the events may have not taken place, or the participants may be relying on what they have been told about the event rather than what they actually remember on the basis of firsthand experience. On the other hand, given the frequency with which these early accounts occur in the literature, it is possible that a small handful of preverbal experiences have been retained over time and have been translated into language in such a way that they can be reported during an interview years, or even decades, later. If this is the case, how might it occur?

In addition to the magic shrinking machine event, the magic bubble machine task is another blended memory procedure that was developed by Morris and Baker-Ward (2007) to explore the translation of preverbal memories into language over a relatively brief developmental period. In the Morris and Baker-Ward (2007) study, 2-year-old children were first tested for their receptive and productive vocabulary for six basic colors; children's color knowledge was also assessed using teacher and parent reports. The following day, children took part in the magic bubble machine event during which they learned that one of the six colors of bubble soap (i.e., the target color) activated the machine, producing bubbles. In the 8 weeks following this event, an experimenter visited the children's classrooms for approximately an hour each week and conducted color-learning activities. After the 8-week color training, children were tested again for their knowledge of the six color words. At this time, children were also assessed for their memory of the magic bubble machine event using verbal recall, nonverbal color identification of the target color on a color chart, and behavioral reenactment in which children demonstrated how to activate the bubble machine.

Using this procedure, Morris and Baker-Ward (2007) found that 9 of the 31 children (i.e., ~30%) who did not know the target color word prior to the magic bubble machine event, but who subsequently acquired the word during the 8-week training period, used the target color word to describe the event during the verbal recall test. In other words, these children presumably encoded the target color without the benefit of language, maintained the memory over the 8-week

delay, and then mapped the newly learned word onto their earlier, preverbal representation. In contrast to the earlier findings reported by Simcock and Hayne (2002) and Morris and Baker-Ward's (2007) findings suggest that children might be able to translate a preverbal memory into words.

It is important to note that there were a number of procedural differences between the Simcock and Hayne (2002) and Morris and Baker-Ward (2007) studies, including the complexity of the events, the context in which the events took place, the length of the retention interval, and the addition of specific language training. One or all of these factors may have contributed to the difference in results. For example, the timing of the language training and the fact that it occurred in the same context as the original event may have facilitated children's ability to map a newly learned word onto a previously established representation. Whether similar translation would occur under more real-world learning conditions is not known, but it is a fruitful area for further research.

In addition to the high level of contextual support during the language-training portion of the study, the 2-year-olds in the Morris and Baker-Ward (2007) study were tested for verbal recall in the presence of the bubble machine (i.e., a highly salient retrieval cue), whereas in Simcock and Hayne's (2002) study, the magic shrinking machine was not present during verbal recall. The critical role of reminder object cues for young children's verbal recall has been demonstrated in previous studies (e.g., Bauer et al., 1995, 2004) and, therefore, suggests that the level of contextual support at the time of test may be another key determinant of whether children can access their preverbal memories via language (Morris and Baker-Ward, 2007).

Although Simcock and Hayne (2002) found no evidence of children using target words that were not part of their vocabulary at the time of the magic shrinking machine event to describe the event after a 6-month or 1-year delay, Jack et al. (2012) recently found some evidence in support of verbal translation of preverbal memories after a 6-year delay. In their study, 27- to 51-month-old children experienced the magic shrinking machine event and then were tested after 24-h and 6-year delays. Following the 6-year delay, 9 of the 46 children who took part in this study (i.e., 20%) were able to verbally report information about the event. Importantly, of these nine children, two described the event using a target word that was not part of their vocabulary at the time of original encoding.

So what set apart the children who were able to provide verbal report after a 6-year delay and those who did not? The nine children who were able to verbally recall the magic shrinking machine event had not performed any better during the 24-h interview than did the children who did not verbally recall the event at the 6-year interview. Furthermore, children's vocabulary skills and general memory skills at the times of encoding and test were not related to their long-term verbal recall performance. Jack et al. (2012) did find, however, that the few children who were able to provide verbal reports of the magic shrinking machine event after a 6-year delay were the ones who had actively participated in conversations about the event with their parents during the retention interval.

The importance of mother–child conversations on children's early memory development is now well established. A large body of research has shown that mother–child conversations about the past influence children's ability to develop well-organized and elaborate memory narratives (Dunn et al., 1987; Eisenberg, 1985; Farrant and Reese, 2000; Fivush, 1991; Fivush and Fromhoff, 1988; Haden et al., 2009; Harley and Reese, 1999; Hudson, 1990; Newcombe and Reese, 2004; Peterson and McCabe, 1992, 1994; Reese, 2002). Jack et al.'s (2012) finding extends this literature by suggesting that the opportunity for children to verbally rehearse their memories may play an essential role in their subsequent ability to translate their preverbal memories into language. Similarly, it is also possible not only that the language training in the Morris and Baker-Ward (2007) study with the magic bubble machine may have helped children to learn the relevant color word, but also that the training per se may have provided a rehearsal opportunity for the target event, keeping the memory active over the retention interval and providing the opportunity for the translation to occur. This possibility remains to be tested.

Verbal Reminders

There is a well-established literature with human infants showing that even after a nonverbal memory has been forgotten, the memory can be reactivated and expressed given the appropriate retrieval cues (e.g., Butler and Rovee-Collier, 1989; Herbert and Hayne, 2000; Rovee-Collier et al., 1980, 1985). Although the nature of an appropriate retrieval cue is initially quite specific to the conditions of original encoding, as a function of both age and experience, the nature of these cues gradually broadens (Hayne, 2006). Over the course of development, the nature of effective retrieval cues gradually broadens to such a degree that language cues also become effective retrieval aids – potentially even for memories that were originally preverbal or nonverbal in nature. We have recently found evidence that suggests that verbal reminders of past events may be a key to the maintenance of early memories, even before children have the full-fledged language skills to verbally rehearse and retrieve memories on their own. In particular, for preverbal and early-verbal children, exposure to simple questions such as "Do you remember when…?" may help maintain a preverbal representation long enough that they can use their emerging language skills to translate the memory into words.

As an initial attempt to investigate the age at which children begin to exploit simple verbal reminder questions, we tested 2-, 3-, and 4-year-olds on the VPC paradigm and measured their retention of a visual stimulus with or without a verbal reminder (Imuta et al., 2012). In this study, we first familiarized children with a three-dimensional cartoon face, and, following a predetermined delay, we tested children's memory of the visual stimulus by presenting them with the familiar face and a novel face, side by side, and coded for children's looking behavior. To investigate the effect of simple verbal reminder questions on children's memory retrieval, immediately prior to the testing session, children in the verbal reminder condition were asked, "Can you remember coming

here a little while ago?" followed by "Can you remember what face you saw the last time you came?" Children in the control conditions were either simply complimented on their clothing or asked an unrelated question about their favorite toy. Using this procedure, we found that children in the control conditions did not exhibit retention of the visual stimulus after a delay. In contrast, we found that 2-, 3-, and 4-year-olds who were provided the verbal reminder questions exhibited excellent retention (Imuta et al., 2012). The finding that children as young as 2 years of age can exploit simple verbal reminder questions to retrieve a nonverbal memory suggests that this kind of reminding may play a crucial role in allowing these preverbal memories to be translated into language over time.

Beyond the Preverbal-to-Verbal Transition

Although we have largely focused on the fate of preverbal memories, it is important to note that even our early *verbal* memories are rarely recalled in adulthood (Tustin and Hayne, 2010). Indeed, the general absence of memories from early childhood was noted by Freud (1963), who coined the term 'childhood amnesia' (or 'infantile amnesia') to describe the phenomenon. Since Freud (1963), a number of theories have emerged, and, with few exceptions, these theories have largely focused on *human* developmental transitions such as the preverbal-to-verbal transition that we have discussed here and elsewhere (Simcock and Hayne, 2002), the emergence of a sense of self (Howe and Courage, 1993), and the emergence of episodic memory (Wheeler et al., 1997), to name a few. Interestingly, work with nonhuman animals has demonstrated that they too experience 'childhood' amnesia (Callaghan and Richardson, 2012), suggesting that theories that focus solely on *human* developmental transitions cannot provide a complete explanation of the phenomenon.

Importantly, work with nonhuman animals underscores what Hayne and Jack (2011) referred to as the 'forgotten explanation' of childhood amnesia, that is, forgetting (Frankland et al., 2013; Josselyn and Frankland, 2012). The role of forgetting in childhood amnesia is often ignored because normal forgetting functions cannot explain the specificity and degree of amnesia that adults display (Wetzler and Sweeney, 1986). With respect to specificity, Tustin and Hayne (2010) have demonstrated that forgetting occurs across a protracted, rather than specific, developmental time frame. They found that as the age of participants increased, so did the age of their earliest memories, such that younger participants had significantly earlier memories than older participants (Tustin and Hayne, 2010). With respect to the degree of amnesia, recent nonhuman animal work has demonstrated that childhood amnesia coincides with a period of high neurogenesis in the hippocampus (Frankland et al., 2013; Josselyn and Frankland, 2012), providing an explicit neural mechanism for the steeper forgetting functions that we commonly see early in development in humans and other animals. That is, the high neuronal turnover in the hippocampus that occurs early in development impairs the ability to recreate the specific patterns of activation that are required to recall specific memories (Guskjolen et al., 2013).

Conclusions

In conclusion, the data reviewed here show that children continue to rely on their nonverbal memory skills over a protracted period of development. If we only ask them questions about what they remember, we are likely to underestimate the content of their memory. Although most of these early, nonverbal memories will ultimately be forgotten, there is some evidence that at least some of them are retained over significant periods of development. Although rare, it may also be possible that at least a small handful of these memories might be translated into language. We conclude that the probability that translation will occur may be dependent, at least in part, on the availability of verbal reminders during the retention interval. The fate of these early preverbal memories may provide yet another key to the intriguing phenomenon of childhood amnesia.

See also: Autobiographical Memory; Child-Directed Speech: Influence on Language Development; Episodic Memory; Hippocampal Amnesia; Intentionality in Language and Communication, Emergence of; Language Acquisition; Memory Binding; Memory Development During Infancy and Early Childhood across Cultures.

Bibliography

Barr, R., Dowden, A., Hayne, H., 1996. Developmental changes in deferred imitation by 6- to 24-month-old infants. Infant Behavior and Development 19, 159–170. http://dx.doi.org/10.1016/s0163-6383(96)90015-6.

Bauer, P.J., Hertsgaard, L.A., Wewerka, S.S., 1995. Effects of experience and reminding on long-term recall in infancy: remembering not to forget. Journal of Experimental Child Psychology 59, 260–298. http://dx.doi.org/10.1006/jecp.1995.1012.

Bauer, P.J., Shore, C.M., 1987. Making a memorable event: effects of familiarity and organization on young children's recall of action sequences. Cognitive Development 2, 327–338. http://dx.doi.org/10.1016/S0885-2014(87)80011-4.

Bauer, P.J., van Abbema, D.L., Wiebe, S.A., Cary, M.S., Phill, C., Burch, M.M., 2004. Props, not pictures, are worth a thousand words: verbal accessibility of early memories under different conditions of contextual support. Applied Cognitive Psychology 18, 373–392. http://dx.doi.org/10.1002/acp.1006.

Butler, J., Rovee-Collier, C.K., 1989. Contextual gating of memory retrieval. Developmental Psychobiology 22, 533–552. http://dx.doi.org/10.1002/dev.420220602.

Callaghan, B., Richardson, R., 2012. The effect of adverse rearing environments on persistent memories in young rats: removing the brakes on infant fear memories. Translational Psychiatry 2, e138.

Clayton, N.S., Bussey, T.J., Dickinson, A., 2003. Can animals recall the past and plan for the future? Nature Reviews Neuroscience 4, 685–691.

Daehler, M., Bukatko, D., Benson, K., Myers, N., 1976. The effects of size and color cues on the delayed response of very young children. Bulletin of the Psychonomic Society 7, 65–68.

DeLoache, J.S., Brown, A.L., 1979. Looking for Big Bird: studies of memory in very young children. Quarterly Newsletter of the Laboratory of Comparative Human Cognition 1, 53–57.

DeLoache, J.S., Brown, A.L., 1983. Very young children's memory for the location of objects in a large-scale environment. Child Development 54, 888–897. http://dx.doi.org/10.2307/1129893.

DeLoache, J.S., Brown, A.L., 1984. Where do I go next? Intelligent searching by very young children. Developmental Psychology 20, 37–44. http://dx.doi.org/10.1037/0012-1649.20.1.37.

DeLoache, J.S., Cassidy, D.J., Brown, A.L., 1985. Precursors of mnemonic strategies in very young children's memory. Child Development 56, 125–137. http://dx.doi.org/10.2307/1130180.

Dunn, J., Bretherton, I., Munn, P., 1987. Conversations about feeling states between mothers and their young children. Developmental Psychology 23, 132–139.

Eacott, M.J., Crawley, R.A., 1998. The offset of childhood amnesia: memory for events that occurred before age 3. Journal of Experimental Psychology: General 127, 22–33. http://dx.doi.org/10.1037/0096-3445.127.1.22.

Eisenberg, A.R., 1985. Learning to describe past experiences in conversation*. Discourse Processes 8, 177–204. http://dx.doi.org/10.1080/01638538509544613.
Fagan, J.F., 1974. Infant recognition memory: the effects of length of familiarization and type of discrimination task. Child Development 45, 351–356.
Fantz, R.L., 1964. Visual experience in infants: decreased attention to familiar patterns relative to novel ones. Science 146, 668–670. http://dx.doi.org/10.1126/science.146.3644.668.
Farrant, K., Reese, E., 2000. Maternal style and children's participation in reminiscing: stepping stones in children's autobiographical memory development. Journal of Cognition and Development 1, 193–225. http://dx.doi.org/10.1207/s15327647jcd010203.
Fivush, R., 1984. Learning about school: the development of kindergartners' school scripts. Child Development 55, 1697–1709. http://dx.doi.org/10.2307/1129917.
Fivush, R., 1991. The social construction of personal narratives. Merrill-Palmer Quarterly 37, 59–81.
Fivush, R., Fromhoff, F.A., 1988. Style and structure in mother–child conversations about the past. Discourse Processes 11, 337–355. http://dx.doi.org/10.1080/01638538809544707.
Fivush, R., Gray, J.T., Fromhoff, F.A., 1987. Two-year-old talk about the past. Cognitive Development 2, 393–409. http://dx.doi.org/10.1016/s0885-2014(87)80015-1.
Fivush, R., Hudson, J.A., Nelson, K., 1984. Children's long-term memory for a novel event: an exploratory study. Merrill-Palmer Quarterly 30, 303–316.
Frankland, P.W., Köhler, S., Josselyn, S.A., 2013. Hippocampal neurogenesis and forgetting. Trends in Neurosciences 36, 497–503.
Freud, S., 1963. New introductory lectures on psycho-analysis. In: Strachey, J. (Ed.), The Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 15. Hogarth Press, London, pp. 199–201.
Gross, J., Hayne, H., 1998. Drawing facilitates children's verbal reports of emotionally laden events. Journal of Experimental Psychology: Applied 4, 163.
Guskjolen, A., Yiu, A.P., Akers, K., Martinez-Canabal, A., Restivo, L., Josselyn, S.A., et al., 2013. Inhibiting hippocampal neurogenesis promotes memory persistence in infant mice. Paper Presented at the Canadian Neuroscience Meeting, University of Toronto, The Hospital for Sick Children, ON, Canada.
Haden, C., Ornstein, P.A., Rudek, D.J., Cameron, D., 2009. Reminiscing in the early years: patterns of maternal elaborativeness and children's remembering. International Journal of Behavioral Development 33, 118–130. http://dx.doi.org/10.1177/0165025408098038.
Harley, K., Reese, E., 1999. Origins of autobiographical memory. Developmental Psychology 35, 1338–1348.
Hartshorn, K., Rovee-Collier, C.K., 1997. Infant learning and long-term memory at 6 months: a confirming analysis. Developmental Psychobiology 30, 71–85. http://dx.doi.org/10.1002/(sici)1098-2302(199701)30:1<71::aid-dev7>3.0.co;2-s.
Hartshorn, K., Rovee-Collier, C.K., Gerhardstein, P., Bhatt, R.S., Wondoloski, T.L., Klein, P.J., et al., 1998. The ontogeny of long-term memory over the first year-and-a-half of life. Developmental Psychobiology 32, 69–89. http://dx.doi.org/10.1002/(SICI)1098-2302(199803)32:2<69::AID-DEV1>3.0.CO;2-Q.
Hayne, H., 2004. Infant memory development: implications for childhood amnesia. Developmental Review 24, 33–73. http://dx.doi.org/10.1016/j.dr.2003.09.007.
Hayne, H., 2006. Age-related changes in infant memory retrieval: implications for knowledge acquisition. In: Munakata, Y., Johnson, M.H. (Eds.), Processes of Change in Brain and Cognitive Development: Attention and Performance. Oxford University Press, New York, pp. 209–231.
Hayne, H., Imuta, K., 2011. Episodic memory in 3- and 4-year-old children. Developmental Psychobiology 53, 317–322. http://dx.doi.org/10.1002/dev.20527.
Hayne, H., Jack, F., 2011. Childhood amnesia. Wiley Interdisciplinary Reviews: Cognitive Science 2, 136–145.
Herbert, J., Hayne, H., 2000. Memory retrieval by 18–30-month-olds: age-related changes in representational flexibility. Developmental Psychology 36, 473–484.
Horn, H.A., Myers, N.A., 1978. Memory for location and picture cues at ages two and three. Child Development 49, 845–856. http://dx.doi.org/10.2307/1128255.
Howe, M.L., Courage, M.L., 1993. On resolving the enigma of infantile amnesia. Psychological Bulletin 113, 305–326.
Hudson, J., Nelson, K., 1986. Repeated encounters of a similar kind: effects of familiarity on children's autobiographic memory. Cognitive Development 1, 253–271. http://dx.doi.org/10.1016/S0885-2014(86)80004-1.
Hudson, J.A., 1990. The emergence of autobiographical memory in mother–child conversation. In: Fivush, R., Hudson, J. (Eds.), Knowing and Remembering in Young Children. Cambridge University Press, Cambridge, UK, pp. 166–196.
Imuta, K., Scarf, D., Hayne, H., 2012. The effect of verbal reminders on memory reactivation in 2-, 3-, and 4-year-old children. Developmental Psychology 49, 1058–1065. http://dx.doi.org/10.1037/a0029432.
Jack, F., Hayne, H., 2007. Eliciting adults' earliest memories: does it matter how we ask the question? Memory 15, 647–663. http://dx.doi.org/10.1080/09658210701467087.
Jack, F., MacDonald, S., Reese, E., Hayne, H., 2009. Maternal reminiscing style during early childhood predicts the age of adolescents' earliest memories. Child Development 80, 496–505. http://dx.doi.org/10.1111/j.1467-8624.2009.01274.x.
Jack, F., Simcock, G., Hayne, H., 2012. Magic memories: young children's verbal recall after a 6-year delay. Child Development 83, 159–172. http://dx.doi.org/10.1111/j.1467-8624.2011.01699.x.
Josselyn, S.A., Frankland, P.W., 2012. Infantile amnesia: a neurogenic hypothesis. Learning & Memory 19, 423–433.
MacDonald, S., Uesiliana, K., Hayne, H., 2000. Cross-cultural and gender differences in childhood amnesia. Memory 8, 365–376.
Meltzoff, A.N., 1985. Immediate and deferred imitation in fourteen- and twenty-four-month-old infants. Child Development 56, 62–72.
Morris, G., Baker-Ward, L., 2007. Fragile but real: children's capacity to use newly acquired words to convey preverbal memories. Child Development 78, 448–458. http://dx.doi.org/10.1111/j.1467-8624.2007.01008.x.
Nelson, K., 2006. Narratives from the Crib. Harvard University Press, Cambridge, MA.
Nelson, K., Fivush, R., 2004. The emergence of autobiographical memory: a social cultural developmental theory. Psychological Review 111, 486–511. http://dx.doi.org/10.1037/0033-295X.111.2.486.
Newcombe, R., Reese, E., 2004. Evaluations and orientations in mother–child narratives as a function of attachment security: a longitudinal investigation. International Journal of Behavioral Development 28, 230–245. http://dx.doi.org/10.1080/01650250344000460.
Peterson, C., McCabe, A., 1992. Parental styles of narrative elicitation: effect on children's narrative structure and content. First Language 12, 299–321. http://dx.doi.org/10.1177/014272379201203606.
Peterson, C., McCabe, A., 1994. A social interactionist account of developing decontextualized narrative skill. Developmental Psychology 30, 937–948.
Reese, E., 2002. A model of the origins of autobiographical memory. In: Fagen, J.W., Hayne, H. (Eds.), Progress in Infancy Research, vol. 2. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 215–260.
Rose, S.A., 1981. Developmental changes in infants' retention of visual stimuli. Child Development 52, 227–233.
Rovee-Collier, C.K., Enright, M., Lucas, D., Fagen, J.W., Gekoski, M.J., 1981. The forgetting of newly acquired and reactivated memories of 3-month-old infants. Infant Behavior and Development 4, 317–331. http://dx.doi.org/10.1016/s0163-6383(81)80033-1.
Rovee-Collier, C.K., Patterson, J., Hayne, H., 1985. Specificity in the reactivation of infant memory. Developmental Psychobiology 18, 559–574. http://dx.doi.org/10.1002/dev.420180611.
Rovee-Collier, C.K., Sullivan, M.W., Enright, M., Lucas, D., Fagen, J.W., 1980. Reactivation of infant memory. Science 208, 1159–1161. http://dx.doi.org/10.1126/science.7375924.
Simcock, G., Hayne, H., 2002. Breaking the barrier? Children fail to translate their preverbal memories into language. Psychological Science 13, 225–231.
Simcock, G., Hayne, H., 2003. Age-related changes in verbal and nonverbal memory during early childhood. Developmental Psychology 39, 805–814. http://dx.doi.org/10.1037/0012-1649.39.5.805.
Suddendorf, T., Corballis, M.C., 2007. Mental time travel across the disciplines: the future looks bright. Behavioral and Brain Sciences 30, 335–345. http://dx.doi.org/10.1017/S0140525X0700221X.
Tulving, E., 2002. Episodic memory: from mind to brain. Annual Review of Psychology 53, 1–25. http://dx.doi.org/10.1146/annurev.psych.53.100901.135114.
Tulving, E., 2005. Episodic memory and autonoesis: uniquely human? In: Terrace, H.S., Metcalfe, J. (Eds.), The Missing Link in Cognition: Origins of Self-Reflective Consciousness. Oxford University Press, New York, pp. 3–56.
Tustin, K., Hayne, H., 2010. Defining the boundary: age-related changes in childhood amnesia. Developmental Psychology 46, 1049.
Wetzler, S.E., Sweeney, J.A., 1986. Childhood amnesia: an empirical demonstration. In: Rubin, D.C. (Ed.), Autobiographical Memory. Cambridge University Press, New York, pp. 191–201.
Wheeler, M.A., Stuss, D.T., Tulving, E., 1997. Toward a theory of episodic memory: the frontal lobes and autonoetic consciousness. Psychological Bulletin 121, 331–354.

Relevant Websites

http://www.otago.ac.nz/psychology/staff/harlenehayne.html.
http://www.otago.ac.nz/psychology/staff/otago025837.html.

Children and Families in Social Work

Nick Frost, Leeds Metropolitan University, Leeds, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores social work with children in families. Different levels and forms of social work practice are explored. It is argued that such practice exists at the complex interface of the state and the family, and that this makes such work inherently complex and contested. The article concludes by considering the importance of professional judgment and what the future of such work may look like.

Introduction

Social workers across the world work with children and their families to promote the perceived best interests of the child. Social workers operate at the interface of the state and the family: an interface that is often contested and controversial. Their work, particularly when concerned with child maltreatment, is often subject to high levels of public, political, and media scrutiny (Nelson, 1986; Parton, 2006).

This article will explore how social work with children across the world has many common features: the core of the work shares approaches such as developing a positive relationship with the child, and where possible their parent(s), while empowering and protecting children and young people. But there are also differences in social work practice dependent on the social and political context in which it is situated.

This article will argue that social work can be understood at three key levels: the social structural level, the community level, and the family/individual level. It will also be argued that social work with children and families consists of a number of core practices: family support, safeguarding and the provision of alternative care, and permanent placements for children. This article will conclude with a discussion of professional judgment and development before speculating how social work practice with children may look in the future.

The objectives of what follows are to

- analyze different levels of social work practice with children and families;
- explore differing forms of practice with children and their families;
- reflect upon professional judgment and development; and
- suggest future policy and practice framework for working with children.

Underpinning Levels and Approaches to Social Work with Children and Families

When social workers work with a child it is usually because the child is facing some sort of problem or challenge. A child living with a consistent and loving parent, in a materially and emotionally comfortable environment, is unlikely to ever come across a social worker. Rather social workers practice with children who may be abused, neglected, abandoned, or deprived. Within social work there are many ways of understanding the many challenges that a child might face: how professional social workers conceptualize the challenges facing the child will have a profound impact on how they choose to intervene. We can understand children's problems in a number of ways as demonstrated in Table 1.

While Table 1 is useful in helping us understand different styles and approaches to social work, in reality many social workers will utilize combinations of different models and interventions, depending on the organizational framework they operate within and the specific child they are working with. Often workers will combine styles of working – moving from being a therapist, to an advocate, to a campaigner, for example. In this sense, social work occupies a particular flexible space – in contrast to the more defined role of a lawyer or doctor, for example.

Social work is a responsive role which has to be fluid and flexible to achieve its goal – that of promoting the best interests of the child. As we will see later this gives social workers a particular role in the multidisciplinary professional 'team around the child' (Anning et al., 2010).

Levels of Intervention

Structural and Political Interventions

The most global and overarching forms of addressing the challenges of childhood exist at the political and social structural level (Frost, 2011). Social workers all over the

Table 1 Understanding childhood challenges

Theory/level	*Explanatory framework*	*Interventions*
Social structural/ political	Childhood problems are situated in deep-seated social injustice and inequalities	Political/campaigning initiatives
Community/ ecological approaches	Community, family, and individual factors combine to produce challenges for children	Combination of methods ranging from political and community initiatives
Family systems/ Individual approaches	Childhood problems are rooted in household interactions	Therapy Family casework Family support

world will be engaged in such styles of work: it may involve campaigning for change; gathering data that demonstrates the nature of childhood disadvantage; and/or mobilizing adults or children themselves to bring about change. Social work as a profession needs to reflect upon how effective it is in tackling major social problems such as domestic violence, rape, and genital mutilation. This form of action has long historical antecedents going back to the origins of social work and Victorian figures such as Octavia Hill and the founders of the large childhood charities such as the Societies for the Prevention of Cruelty to Children (Ferguson, 2004). Often the Victorian philanthropists would combine the levels explored in **Table 1** – they would campaign for social change, work in communities, and practice with individual children and their families.

Today, social workers are working for such change through global organizations such as UNICEF or in NGOs such as Save the Children. While the Victorian pioneers often worked as if they somehow knew 'the best interests of the child,' most organizations aim to give a voice to children by enhancing their participation in both the policy and practice areas in a manner consistent with the United Nations Convention on the Rights of Children (www.unicef.org.uk accessed 05.06.13).

The campaigning and collective role of social work sometimes positions it in a vulnerable position. Social workers are often described as being overly 'politically correct,' 'revolutionaries,' or alternatively 'wimps' and 'bullies' (Franklin and Parton, 2001). These accusations reflect, to return to one of the earlier themes, the complex relationship between the state and the family. The social worker can be accused of 'intervening too much,' when removing children from their birth parents, or 'intervening too little,' when a child dies at the hands of a parent. Particular children, such as the child known as 'Baby Peter' in England, often attract high levels of media and political attention (Laming, 2009). This complex tension forces social workers to walk a 'tight wire,' which is sure to be both controversial and contested.

This 'tight wire' leads to social work existing in a complex relationship with the media, politicians, and public opinion. In the United Kingdom, for example, ever since the late 1960s, social work has been subject to sustained public, media, and political criticism. During the same period, social work has been invested in by government, grown both in terms of numbers and professional status, and been granted an increasing number of statutory responsibilities. This we might identify as something of a paradox: the more social work is seen as failing and controversial, then the more it has to be supported and invested in. Alongside this investment, particularly in the Western world comes increased monitoring and inspection by the state and its agencies (Frost and Parton, 2009).

The political explanations of this paradox are complex and at times contradictory: the explanations exist somewhere along the 'welfare'/'social control' continuum. Modern liberal governments are obliged to invest in childhood: to utilize an overused cliché, 'children are our future,' and their welfare must be sustained. Thus, any social system that wishes to reproduce itself (in both the political and biological senses) has to invest in childhood. In liberal and social democratic societies the family will play the major role in nurturing childhood: with varying degrees of state support. The social work role emerges when the family is perceived as failing in this nurturing role (Donzelot, 1979). Social work will therefore intervene with the abused child, the neglected child, the child not attending education, and so on. The social worker is therefore simultaneously both a supporter of the vulnerable child and the enforcer of the authority of the state. Different commentators vary in their weighting of these roles. Some see the social worker as a champion of children's welfare, some perceive them as the enforcer of the authoritarian state and yet others as a complex combination of the two. Few ask how the child might view the social worker.

Social workers will vary in their approach to the family–state interface. Some will perceive themselves as 'radical' (Bailey and Brake, 1975; Bolger et al., 1979) or 'feminist' (Dominelli and McLeod, 1989) social workers pursuing overtly political approaches to their work and toward social change, others will focus on a strongly individualized agenda of therapy (Payne, 2005).

Community-Based Interventions

The global and political initiatives outlined above often take a concrete form of locally based and delivered initiatives. These range from major health or nutrition-based projects in countries facing shortages and absolute poverty, to community play schemes or parenting support in industrial and postindustrial countries. The underlying theory is the same as those working at the political level and shares a political element that is we cannot address childhood inequalities solely through work with individuals, and therefore broader-based interventions are required (Lavelette, 2011).

Community-based schemes vary considerably. They may include area-based antipoverty schemes, community work schemes encouraging collective campaigning and organizing, or campaigns against domestic violence or substance abuse. Social workers will normally be working in these contexts alongside community development workers, education workers, health professionals, and many others. The social worker may bring their specific skills in relationship building, professional liaison, and child-centered welfare work (Brennan et al., 2013).

Such initiatives are often seen as 'preventive.' They are a form of social intervention that is often universal in intent, that is, available without stigma to all sections of the population. This idea of prevention is central to social work. The initial ambition of many social workers will have been to work in prevention, which is sometimes conceptualized as 'family support' or 'early intervention.' The theory is that by offering support before social problems have become embedded the social worker can contribute to the prevention of later problems such as youth crime, child maltreatment, or family breakdown. Family support, often related to younger children, can also occur in the form of early help with older young people, early in the emergence of a particular problem or challenge. Contrary to popular perceptions, the social worker value-base is strong in relation to family support, and usually social workers regard the removal of the child to care as a last

resort. Prevention sits at the heart of child and family social work (Frost and Dolan, 2012).

Family-Based Initiatives

Family support may exist at both the levels explored so far: the political level and the community level. Fundamentally, however, it is an activity that exists at the level of the household (Georgas et al., 2006). The social worker may offer a 'personalized' combination of counseling, therapeutic interventions, practical help, and services. Usually this will involve cooperation with other professionals – in health, education, and related services. The role of the social worker may be involved in the assessment of a social situation – often using an ecological approach which takes into account the individual, the family, and community settings and their interplay (Jack, 1997). Having undertaken an assessment, social workers may play the role of 'broker' or 'case manager' – planning a coherent package of support in partnership with the family. They may provide direct support services – building relationships with the child and their parent(s) (Folgheraiter, 2003). Many social workers in modern industrialized societies report that they are no longer able to play this role – and that their work has become dominated by the protection of maltreated children. This will often involve rapid assessments and finely judged interventions – the practice safeguarding of children is examined below (Ferguson, 2011).

Forms of Social Work Practice

Social work with children and their families can take many forms. It is argued here that social work can be understood as comprising of four major practices: family support, safeguarding, provision of alternative care, and the provision of permanent alternatives.

Supporting Families

Family support work involves assisting families to raise their own children and is often linked to reducing the need to bring children into state care. Within family support practice there are a range of approaches often theorized as being at primary, secondary, and tertiary levels (Frost and Dolan, 2012). Primary prevention can operate at the social structural level, which we have explored earlier. It may also involve initiatives involving universal provisions such as play schemes, drop-in centers, and welfare advice. At the secondary level, family support practice is more focused on families that face specific challenges – for example, parents might be lonely, isolated, or depressed. The international organization Home Start provides an example of intervention at this level by providing professionally supported volunteer home visiting to families with identified needs. At tertiary level, interventions are focused at more severe social problems and a smaller number of families: perhaps involving domestic violence, substance abuse, or parenting problems.

Family support is a 'slippery concept,' which is difficult to define (Frost and Dolan, 2012). While social workers are involved in family support, it involves other professionals – such as day care, health, and play workers – and covers five core ideas.

- First, family support is based in the idea of identifying problems early, and assumes that it is better to offer support early in the life of a problem, before it becomes entrenched and difficult to resolve. For example, if a parent has problems communicating with a child it is better to provide advice and support when this emerges as an issue, before conflict arises between the child and parent.
- Second, family support is proactive. It attempts to work with children and their parents, and aims to bring about positive change for the child and the family.
- Third, family support implies partnership. It cannot be imposed on a family. This demands particular skills and approaches from the professional involved.
- Fourth, family support is based in a theory of change. It assumes that positive change can be brought about through family support initiatives and that social work has a key role to play within this change.
- Finally, family support is seen as bringing about wider social benefits through the prevention of social problems such as youth crime and child abuse.

Protecting Children

It is the protection of children that is the highest profile activity that social workers are involved in. In many parts of the globe this has been associated with 'scandals' where social work practice has been blamed for intervening too little or too late, or sometimes for intervening too much. Safeguarding children is very challenging for social workers – basically it involves attempting to assess whether the child's best interests are served by remaining with their birth families or being removed to substitute carers either on a short-term or long-term basis.

Our understanding of child abuse is constantly evolving and changing. For example, since the end of the Second World War we have seen the emergence of 'the battered baby syndrome,' child sexual abuse, institutional abuse, Internet abuse, and most recently we have a growing understanding of the organized sexual exploitation of young people through group and gang grooming. We can see that our understanding of child abuse is 'socially constructed': ever changing and changing across time and space. These understandings are a shared professional process but social work has played a key role in our understanding of the changing nature of child abuse (Parton, 2006).

The complex nature of the social work task is hinted at in the previous paragraph: how can we respond to ever-changing phenomena? The uncovering of then newly articulated forms of abuse could be accompanied by denial: 'battered babies,' (Parton, 2006) incest and ritual forms of abuse have all been accompanied by strenuous denial (Nelson, 1986). Awareness of the damage of more long-term subtle abuse has also been vital to the timing of interventions, for example, in cases of emotional abuse and long-term neglect. New forms of proactive practice have to evolve to detect abuse and to protect children. Social workers, and associated academics, are often in the forefront of campaigning and practice developments in this field.

Good practice is focused on listening to children, providing healing for any trauma they have suffered and in providing safe and sustainable care for them. This takes us to a discussion of the provision of alternative care for separated children.

Alternative Care

In most social systems where children cannot live with their birth parent(s), social workers are involved in finding alternative forms of care for them: this may be in the form of institutional care, kinship care, foster care with strangers, or adoption. The deployment and legal base of these alternatives varies across the world. In some countries, kinship care predominates (e.g., Spain), in others, large residential provision exists (e.g., India), and in yet others, foster care with strangers dominates (e.g., United Kingdom) (see Lerch and Stein, 2010).

The provision of alternative care for separated children is a core social work activity where the social work authority is 'in loco parentis' (Hannon et al., 2010) and has clear and unequivocal responsibility for the child. In many ways this is an 'acid test' for the social work profession: can it provide positive outcomes for children in care?

There is some positive and suggestive evidence in relation to children in alternative care. Popular discourse is often that outcomes for children in care are poor. This discourse tends to underplay the poorer starting points for children who enter the care system: it is more valuable to measure the 'value-added' by the care system. More realistically there are probably three groups: children who have stable care careers who are often successful when they move on; a middle group who can be successful if adequately supported; and finally a group, often with troubled pasts and disrupted care careers who may well struggle once having left care. Stein refers to these groups as 'moving on,' 'survivors,' and 'strugglers' (Stein, 2012). Judgment on the care system, therefore, has to be more nuanced than that it is 'failing children' and social workers need to recognize the differing needs of the three groups. All three will require some support with a mentoring role more useful for the 'moving on' group, more targeted support for the 'survivors' particularly when they face transitions, and intensive, outreach, and crisis support provided for the 'strugglers' group.

Of all areas of social work, the research evidence in field of children in care and leaving care is probably the strongest. The evidence strongly suggests that good quality care, provided in stable placements, with a gradual and supported transition to adulthood can provide successful outcomes for children in care (Stein, 2012). It also needs to identify and strengthen resilience factors in children – those factors that enable them to survive childhood adversity (Rutter, 1990).

Permanence

Most countries have a form of adoption: the permanent transfer of parental rights and responsibilities to someone other than the birth parent. Adoption is therefore a form of 'permanence': aiming to provide stable and ongoing care for children separated from their birth parents. Adoption acts as a route out of the care system or as an alternative to the system where children are adopted from birth. In most jurisdictions, the social worker has a key role in the adoption process: it remains an important social work role (Javier et al., 2007).

As with other aspects of social work adoption is a publicly high profile and controversial form of child welfare practice. Each aspect of the adoption process is complex and involves subtle elements of professional judgment. Let us examine some of these complexities in turn and assess how they can be addressed.

When should children be placed for adoption? The removal of children from their birth parent(s) remains one of the most demanding arenas of social work practice. Sometimes the birth parent(s) requests adoption for their child, and in some cultures this may be arranged informally. In more legally regulated states, frequently the social worker removes the child utilizing the powers available to the family court. In the context of international adoption, issues of poverty and inequality are strongly at play (Repo and Yrola, 2011). The birth parent will usually be angry and uncooperative, certainly at the time of removal and during court proceedings. Sometimes a child will be removed at birth – especially where the mother has been abusing substances during pregnancy or when she has had previous children removed: this is clearly a highly emotionally charged and demanding process (Javier et al., 2007). In other situations, the child may be removed after a period of abuse and neglect. Here, the social worker, usually with the assistance of the court and other professionals, is exercising complex professional judgment. The parent(s) who have had children removed will normally have been offered a package of family support before any decision to remove the child is made. This will vary considerably depending on the social and political contexts that the social worker is operating within. Decisions about when the support package should be withdrawn and the child removed, and how many packages of support should be offered, remain among the most challenging decisions within the social work lexicon. It is unsurprising that social work practice is often open to challenge – for intervening either 'too early' or 'too late.'

Having explored when children should be placed for adoption, who should be approved as adopters is explored next. The assessment of adopters by social workers is a further area of complexity. Should prospective adopters be rejected for smoking too much, eating too much, being too old, or too culturally insensitive as the popular press often reports? Further complications exist: should gay and lesbian people be enabled to adopt, for example? (Gates et al., 2007). Should people of 'extreme' political views be able to adopt? In the case of well-publicized 'celebrity' adoption wealth and fame also seem able to influence the adoption process (Repo and Yrola, 2011). All these examples have been subject to recent controversies and are likely to continue into the future as these issues lay at the heart of belief systems about the family, parenting, and childhood.

There are perhaps two helpful frameworks for addressing these dilemmas: but of course there are no easy answers. First of all, it is helpful to see adoption as a service for children, not for adults. Thus, the measure should not be what is good for adults but what is in the 'best interests of the child.' The concept of 'best interests' is contested and raises many issues about the child's active participation and to what degree adults can decide what is in the child's 'best interests.' Secondly, potentially

successful adopters should be carefully and professionally assessed for their ability to address the needs of a child. Thus, it is important that adopters are assessed for their caring skills and that they are not discriminated against on grounds of gender, class, and sexuality, for example (Javier et al., 2007).

How should children and adopters be matched? Here we find further controversial decisions that social workers have to address through their practice. Perhaps the fundamental question is around transracial adoption: can adopters effectively care for a child from a different ethnicity? In practice, this issue usually revolves around whether white carers can care for children of color (Lee, 2003). This is perhaps the most controversial and contested issue in social work practice and has generated a considerable literature and campaigns, both for and against transracial adoption (Kirton, 2000). The research evidence here is contradictory. Some researchers argue that transracial placements can provide positive placements with good outcomes (Harris, 2006), while other research points to a lack of self-esteem and confused racial identities resulting from transracial placements (Gaber and Aldridge, 1994). The issue has been subject to political campaigns and strong opinions on both side of the debate particularly when taking children out of their country of origin into more Western environments (Repo and Yrola, 2011). Professionally social workers have been inclined to emphasize the need for sensitive ethnic matching – arguing that this encourages children to grow up with positive self-esteem and the ability to handle racism and prejudice later in life (Kirton, 2000).

What support should be offered postadoption? If adoption is permanent and provides a 'normalized' placement for children, how should social workers view postadoption support? What permanence means to adopters and traumatized children often has a very different meaning to social workers. The term 'forever mummy or daddy' means little to a child whose life has no concept of forever and to an adopter when marriages or partnerships break up or difficulties remain unsupported. Here, we have seen a revolution in social work practice. In the post–Second World War environment adoption was a 'closed' process – a service for orphans or the stigmatized children born 'out of wedlock' (Luckock, 2008). Many of the children were healthy babies and were brought up as if they were the birth children of their adoptive parents. The logic of this approach to permanence was that little or no support was offered to adoptive parents. But in recent decades, adoption has changed. There are fewer 'healthy white babies' to adopt and the service is now provided for older children, who may have had traumatic childhoods or who may be disabled or from a minority ethnic group. Good adoption practice is now 'open,' which means that children and young people know that they are adopted and have life story books utilized to help them understand their past and why they were adopted. Recognizing the challenges and openness of contemporary adoption, it has followed that postadoption support can be provided and this can help reduce the tragedy of adoption placement breakdown. However support in this area remains underdeveloped in many parts of the world. The idea that children have been 'rescued' from previous 'damaged' families and placed in a 'better' family has fed into ideas of them then having to be independent. In recent years, however, adoption support has been gaining support (Luckock, 2008).

Professional Judgment

Social work with children and families involves taking delicate and nuanced decisions about how the state and its agents should relate to children and families. For this reason, social work is often controversial. Good social work practice needs to make sure that practice is grounded in opportunities for critical reflection on both the process and outcomes of social work. In order to promote effective social work, the profession requires:

- An educated and trained workforce, including high level continuing professional development.
- A workforce that is supported and encouraged in the complex roles that it undertakes.
- A workforce that has opportunities for high quality professional supervision that supports critical reflection.
- A workforce that can access a range of high quality services to support the work it undertakes.
- Political and social systems that recognize and support the development of the social work role.

We also need to recognize the challenging nature of social work decision making and the risks involved in such complex work. Harry Ferguson in his body of work reflects well on the complexities of the social work task and the number of complex microdecisions involved in the home visit, for example (Ferguson, 2011).

Social work also needs to evolve as a profession. There are tensions between the idea of professionalism – which traditionally may have implied forms of elitism and ensuring high material rewards – to a new form of professionalism that strives toward genuine participation by children and young people. Developing from the UN Convention on the Rights of the Child, social work since the latter quarter of the twentieth century has been increasingly embedded in the participation of children and young people. All over the world there are youth parliaments, forums, and 'children in care councils' that enable young people to have a real impact on services. The exact forms differ across the world: in India, working children have been organized (Reddy and Ratna, 2002); in England, a day where young people 'takeover' has been organized (http://www.childrenscommissioner.gov.uk/takeover_day, accessed 05.06.13). This participation is an unfinished process and attempts to develop genuine partnership and participation need to be at the center of contemporary forms of professionalism (Cree and Davis, 2007).

This article has focused on social work, but in an increasingly complex and networked society, intractable social problems are not amenable to 'single profession' solutions (Anning et al., 2010). Many social workers are now embedded in multiprofessional settings – such as hospitals, children's centers, or youth crime teams. This enables professionals to share information, skills, and ideas in order to build more flexible, holistic, and responsive approaches, by working through 'the team around the child.'

A Future of Social Work with Children and Families

Social work with children and families has changed and evolved over time, varying according to the social context in which it operates. While it is difficult to predict future trends

and patterns, I can speculate on some future challenges that social work is likely to face. Multidisciplinary approaches seem to be one future direction that is becoming increasingly embedded in the profession and that social work needs to promote more deeply rooted forms of participation.

Technological advances have led to new challenges for safeguarding. The Internet allows perpetrators of abuse new forms of unprecedented global access to children. The mobile phone has equally underpinned new forms of organized and networked abuse that would have been more difficult for abusers before such technology existed. New safeguarding challenges are likely to emerge as technology evolves.

Many of the social problems discussed in this article have their roots in social inequalities and poverty. As much of the Western world suffers from 'austerity' and the economic crisis, it is probable that problems such as family breakdown and youth crime will become more widespread. Global developments such as conflict and displacement also tend to have an impact on children and young people – sometimes leaving them vulnerable to abuse or abandonment. Unaccompanied asylum seeking children are among the most vulnerable children and social work needs to be alert because such issues arise in a globalized and sometimes destabilized world. Social workers have an important, although not exclusive, role to play in relation to these social problems.

Conclusion

This article has outlined the complex, diverse, and challenging roles facing social workers in their professional practice with children and families. Specifically the following issues have been explored:

- the different levels of social work practice with children and families,
- differing forms of practice with children and their families,
- the role of professional judgment and development, and
- what future policy and practice may look like.

Fundamentally, it has been argued that social work exists at the interface between the state and the family and that this relationship is contested and controversial. For this reason, the social work profession requires both space for reflection and robust political support if it is to address its demanding and changing roles successfully in the best interest of children and young people.

See also: Adoption, Demography of; Adoption: Domestic, International and Global Perspectives; Child Care and Child Development; Child Neglect; Child Protection; Childhood and Adolescence: Developmental Assets; Children, Rights of: Cultural Concerns; Defiant Behavior During Infancy and Early Childhood; Human Development, Theories of; Identity in Childhood and Adolescence; Infancy and Childhood: Emotional Development; Social Work Theory; Street Children: Cultural Concerns; Stress in Childhood: Effects on Development.

Bibliography

Anning, A., Cottrell, D., Frost, N., Green, J., Robinson, M., 2010. Developing Multiprofessional Teamwork for Integrated Children's Services. Open University Press, Maidenhead.
Bailey, R., Brake, M. (Eds.), 1975. Radical Social Work. Edward Arnold, London.
Bolger, S., Docking, J., Corrigan, P., Frost, N., 1979. Towards Socialist Welfare Work. Macmillan, London.
Brennan, M., Bridger, J., Alter, T., 2013. Theory, Practice and Community Development. Routledge, London.
Cree, V., Davis, A., 2007. Social Work: Voices from the Inside. Routledge, London.
Dominelli, L., McLeod, E., 1989. Feminist Social Work. Macmillan, London.
Donzelot, J., 1979. The Policing of Families. Pantheon, New York.
Ferguson, H., 2004. Protecting Children in Time. Palgrave Macmillan, Basingstoke.
Ferguson, H., 2011. Child Protection Practice. Palgrave Macmillan, Basingstoke.
Folgheraiter, F., 2003. Relational Social Work. Jessica Kingsley Press, London.
Franklin, B., Parton, N., 2001. Press-ganged! Media reporting of social work and child abuse. In: May, M., Payne, R., Brunsdon, E. (Eds.), Understanding Social Problems: Issues in Social Policy. Blackwell, Oxford, pp. 223–237.
Frost, N., 2011. Rethinking Children and Families. Continuum, London.
Frost, N., Dolan, P., 2012. The theoretical foundations of family support. In: Davies, M. (Ed.), Social Work with Children and Families. Palgrave Macmillan, Basingstoke, pp. 40–49.
Frost, N., Parton, N., 2009. Understanding Children's Social Care. Sage, London.
Gaber, I., Aldridge, J. (Eds.), 1994. In the Best Interests of the Child: Culture, Identity & Trans-Racial Adoption. Free Association Books, London.
Gates, G., Badgett, M., Macombe, J., Chambers, K., 2007. Adoption and Foster Care by Gay and Lesbian Parents in the United States. The Williams Institute, Los Angeles, CA. http://escholarship.org/uc/item/2v4528cx (accessed 05.06.13.).
Georgas, J., Berry, J.W., van de Vijver, F., Kagitcibasi, C., Poortinga, Y.H. (Eds.), 2006. Family Structure and Function: A 30 Nation Psychological Study. Cambridge University Press, Cambridge.
Hannon, C., Wood, C., Bazalgette, L., 2010. In Loco Parentis. Demos, London.
Harris, P. (Ed.), 2006. In Search of Belonging. BAAF, London.
Jack, G., 1997. An ecological approach to social work with children and families. Child and Family Social Work 2 (2), 109–120.
Javier, R., Baden, A., Biafora, F., Camacho-Gingerich, A. (Eds.), 2007. Handbook of Adoption: Implications for Researchers, Practitioners, and Families. Sage, London.
Kirton, D., 2000. 'Race', Ethnicity and Adoption. Open University Press, Buckingham.
Laming, H., 2009. The Protection of Children in England: A Progress Report. TSO, London.
Lavelette, M., 2011. Radical Social Work Re-Visited. Policy, Bristol.
Lee, R.M., 2003. The transracial adoption paradox. Counselling Psychology 31 (6), 711–744.
Lerch, V., Stein, M. (Eds.), 2010. Aging Out of Care: From Care to Adulthood in European and Central Asian Societies. SOS Children's Villages, California.
Luckock, B., 2008. Adoption support and the negotiation of ambivalence in family policy and children's services. Journal of Law and Society 35 (1), 3–27.
Nelson, B., 1986. Making an Issue of Child Abuse. University of Chicago, Chicago.
Parton, N., 2006. Safeguarding Childhood: Early Intervention and Surveillance in a Late Modern Society. Palgrave Macmillan, Basingstoke.
Payne, M., 2005. Modern Social Work Theory. Routledge, London.
Reddy, N., Ratna, K., 2002. Journey in Children's Participation. The Concerned for Working Children, Bangalore.
Repo, J., Yrola, R., 2011. The gender politics of celebrity humanitarianism in Africa. International Feminist Journal of Politics 13 (1), 44–62.
Rutter, M., 1990. Psychosocial resilience and protective mechanisms. In: Rolf, J., Masten, A., Cichetti, D., Nuechterlein, K., Weintraub, S. (Eds.), Risk and Protective Factors in the Development of Psychopathology. Cambridge University Press, Cambridge.
Stein, M., 2012. Young People Leaving Care: Supporting Pathways to Adulthood. Jessica Kingsley Press, London.

Relevant Websites

http://www.childrenscommissioner.gov.uk/takeover_day – Office of the Children's Commissioner (England) (accessed 05.06.13.).
http://www.unicef.org.uk/UNICEFs-Work/Our-mission/UN-Convention/ – UNICEF – Convention on the Rights of the Child (accessed 05.06.13.).

Children and Law

Laura A Rosenbury, Washington University Law School, St. Louis, MO, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J. Roche, volume 3, pp. 1716–1720, © 2001, Elsevier Ltd.

Abstract

This article offers a broad survey of the multiple ways law governs children's lives. The legal regulation of children is neither a subset of family law nor a limited set of exceptions to the legal regulation of persons generally. Law instead creates a legal category of child and governs that category differently than it governs adults across many spheres, including the home, schools, child welfare systems, juvenile justice institutions, and the market. The article analyzes the justifications and effects of this legal approach, highlighting the ways law determine whether parents or state actors may exercise authority over children.

Law governs children's lives in multiple ways. Although parents generally make decisions on behalf of their children, law grants parents that authority. Law does so by recognizing parent–child relationships, granting and protecting parental rights, and specifying that children lack the legal capacity to make most decisions on their own. In some situations, however, law carves out exceptions to parental authority, substituting the state's judgment for that of parents. Law thereby more directly governs children's lives through compulsory education regimes, child protection systems, child labor restrictions, and laws responding to children's misconduct, among other means. In each of these ways, law specifies whether parents or state actors may exercise authority over children (Buss, 2004; Guggenheim, 2005). Debates continue about where to draw the line between parental and state authority (Rosenbury, 2007), but children remain subject to the decisions of either their parents or the state.

This legal framework responds to the fact that children are born completely dependent on adults and only over time develop the mental and physical capabilities needed to pursue more independent courses of action. Parents and the state therefore share the duty of caring for children as vulnerable individuals in need of protection, support, and guidance (Dailey, 2006). Yet law does more than reflect periods of reduced capacity due to chronological age. Law itself participates in the construction of childhood, most noticeably by specifying different ages for when children become legal adults for purposes of marriage, sexual activity, employment, driving, drinking, voting, and criminal prosecution, among other activities (Cunningham, 2006). Individuals therefore may be legal adults for some purposes and legal children for others.

These diverse legal definitions of childhood and adulthood at times reflect the relative nature of maturity, as individuals of a certain age may be mature enough to make decisions in some contexts but not others (Todres, 2012). At other times, however, law's shifting age cutoffs reflect political considerations other than, or in addition to, understandings of capacity and maturity. Moreover, some forms of children's incapacity are explicit consequences of law. Because law generally does not permit children to engage in wage labor, enter into enforceable contracts, or consent to medical care, for example, children are dependent on adults to perform those functions for them. In this respect, the age and maturity of children often does not matter; instead, the relevant factor is the legal status of 'child' attaching to the individual in question (Appell, 2013).

This article surveys the ways this legal treatment of children affects multiple spheres of children's lives. Specific legal regulation of children varies from jurisdiction to jurisdiction, but the general legal framework set forth above produces a baseline level of treatment that is remarkably consistent across jurisdictions. Such consistency remains even in jurisdictions that have embraced robust conceptions of children's rights, including those set forth in the United Nations Convention on the Rights of the Child, which has been adopted by every jurisdiction except the United States and Somalia (Bartholet, 2011; Guggenheim, 2006). The article concludes by briefly highlighting some of the strengths and weaknesses of law's general approach to children.

Law across Children's Lives

Law treats children differently from adults across a wide spectrum of substantive areas. The legal regulation of children is therefore neither a subset of family law nor a limited set of exceptions to the legal regulation of persons generally. Law instead affects children in many spheres, both inside and outside the family, including schools, child protection systems, juvenile justice institutions, and the market. This article will survey these diverse spheres by discussing children's relationships with caregivers, including the establishment of parental ties; children's protection from maltreatment; children's educational experiences, including access to quality education; children's experience in the juvenile justice and criminal law systems; children's interactions with the market and civic life, as workers, consumers, and citizens; and children's peer relationships.

Children's Relationships with Caregivers

Law recognizes the parent–child relationship and then bestows rights and obligations upon legal parents and, to a lesser extent, children. The parent–child relationship arises most commonly through (assumed) genetic ties, with birth certificates listing as legal parents the woman giving birth to a child and the man she identifies as the father. Legal parentage may also arise through

adoption, marital presumptions, gestational agreements, and, in limited circumstances, functional or *de facto* caregiving (Dwyer, 2006; Meyer, 2006). Generally, no more than two adults are recognized as the legal parents of a child. Accordingly, children may not have access to genetic parents who are not their legal parents, including sperm and egg donors, although open adoption is increasingly common (Blauwhof, 2009; Cahn, 2012).

In these ways, adults opt into the legal status of parent, through conduct or agreement. Children, in contrast, are born into the legal status of child or moved into it. Despite children's lack of agency in this respect, the United Nations Convention on the Rights of the Child declares that children have a right to be raised by their parents, as opposed to the state, and to have a relationship with both parents, even if the parents are separated. The United States has not ratified, accepted, or acceded to the Convention, in large part because of concerns that children's rights will threaten traditional lines of authority within the family. Children in the United States therefore do not possess an affirmative right to a parental relationship, although robust notions of family privacy generally dictate that children are raised by legal parents instead of the state (McClain, 2006; Meyer, 2000).

Once the state recognizes a parent–child relationship, it generally does not specify the content of that relationship beyond dictating that parents must provide their children with life necessities, including food, clothing, shelter, and medical care. In exchange for this financial support, legal parents are given broad authority to direct the upbringing of their children, imparting whatever values they please (Harris et al., 1990). For example, the United Nations Convention on the Rights of the Child specifies that parents have a right to control their children's religious, educational, and cultural heritage. The United States Supreme Court has also interpreted the Fourteenth Amendment of the United States Constitution to protect the fundamental right of parents to make decisions concerning the care, custody, and control of their children.

Parents therefore enjoy a right to direct the upbringing of their children largely free from state intervention. The state may intervene to determine custody arrangements and specify the details of financial support when children are born to unmarried legal parents or to legal parents who subsequently divorce. In such situations, judges enter custody and child support orders based on their views of the best interests of the child (Mnookin, 1975; Scott and Emery, 2013). If parents are not subject to such custody or child support orders, parents alone may determine where their children will reside, how they will interact with others, and how they will be financially supported. If children object to their parents' choices, they generally have no recourse in court unless the state believes their parents are defaulting in their duties, as discussed in more detail in the next section.

States may intervene in parent–child relationships, often against the wishes of parents and children, if parents require direct financial support from the state or otherwise are not performing their duties. In the United States, for example, parents receiving public assistance often must use the funds to support their children in certain ways and open their homes to social workers for inspection (Cahn, 1999). Some jurisdictions also affirmatively monitor children being raised by nonlegal parents, such as foster parents, at times disrupting relationships between children and foster parents in order to preserve the primacy of children's relationships with their legal parents until such relationships are terminated, as discussed in the next section.

Children thereby remain under the authority of their legal parents, subject to limited exceptions, until they reach the age of majority, usually 18 years of age, or are otherwise emancipated. Children under the age of majority may become emancipated by operation of law, most commonly when they marry or enlist in the military, or by petition. In order for emancipation petitions to be successful, children often must provide evidence that they are financially independent, able to manage their affairs and, in some jurisdictions, that emancipation would be in their best interests. These requirements make emancipation unavailable for most children. Moreover, some states in the United States rescind emancipation if an emancipated child subsequently becomes indigent.

Some children have attempted to 'divorce' their parents outside of the emancipation context, but they have generally failed. Even if not emancipated, children may direct their own medical care in some situations. As discussed in more detail below, many jurisdictions permit children to access reproductive health care and treatment for sexually transmitted diseases or drug abuse and alcoholism without parental notification or consent. Judges in most jurisdictions may also declare a child to be a 'mature minor' capable of directing medical care for certain purposes (Wadlington, 1994).

Children's Protection from Maltreatment

States most intensely intervene in parent–child relationships when parents are suspected of neglecting or abusing their children. Such intervention may take various forms, including mandated counseling and parenting classes, removal of children from the home, criminal prosecution, and the termination of parental rights. Race and class play a large role in this area of law. Poor families and families of racial and ethnic minorities may more likely experience state intervention and removal of children, often because of bias and the imposition of norms derived from majoritarian childrearing practices (Roberts, 2002). This differential treatment persists despite facially neutral laws and robust antidiscrimination norms and rights.

Jurisdictions define child maltreatment and govern the varied responses to it. When children are harmed by adults who are not their parents, jurisdictions generally invoke the criminal process. When children are harmed by their parents, jurisdictions often do not invoke the criminal process. Instead, most jurisdictions have developed separate, noncriminal child protection systems to respond to parents who are harming their children or otherwise endangering them. This noncriminal approach is thought to better protect children's welfare, as it directs state services to parents and children and often permits children to remain with their parents or to reunite with their parents more quickly than the criminal process would allow (Huntington, 2006).

Child protection systems often employ broader definitions of child maltreatment than those found in criminal law. In the United States, for example, federal law specifies that child abuse must include "any recent act or failure to act on the part

of a parent or caretaker, which results in death, serious physical or emotional harm, sexual abuse, or exploitation, or an act or failure to act which presents an imminent risk of serious harm" (42 U.S.C. §5106(g)), and most of the individual states employ even broader definitions of maltreatment. For instance, many states specify that making or selling drugs in the presence of one's children constitutes abuse, even if that exposure does not result in serious harm or imminent risk of serious harm to the child. In addition, many states have broad definitions of neglect, hinging on parental failure to exercise a minimal degree of care in providing children with food, clothing, shelter, or medical care or failure to protect children from harms within the home. In these states, judges have found children to be neglected when they are living in dirty homes or exposed to domestic violence, for example, on the theory that children living under such circumstances will not develop properly. Over half of the children found to be maltreated in the United States are in fact neglected pursuant to such definitions (Huntington, 2006).

Given broad definitions of abuse and neglect, disputes arise about whether certain parenting decisions constitute maltreatment or instead are legitimate exercises of parental prerogatives. Corporal punishment is a classic example. Although over 30 countries ban all corporal punishment of children, parents in the United States may generally discipline their children physically so long as the force used is reasonable and moderate, both of which are subject to differing interpretations (Freeman, 2010). Other disputes arise in the context of medical decisions, including with respect to elective surgery, the administration of growth hormones, 'corrective' surgery for intersex children, experimental treatments, and decisions to forego both these and other treatment options. Health decisions at birth are often particularly fraught, with doctors or hospitals appealing to the state when parents refuse to follow treatment recommendations (Wadlington, 1994).

Jurisdictions generally require school officials and health care workers to report suspected child maltreatment to the police or child welfare agencies. In a minority of jurisdictions, this duty extends to anyone who knows about or suspects maltreatment. Child protection agencies then exercise discretion in determining which reports to investigate and how to respond to information discovered during such investigations. In the United States, federal law requires agencies to make reasonable efforts to avoid removing children from parents' homes if the child will be safe. Agencies therefore often work with parents on an informal basis, achieving voluntary changes in parenting practices (Huntington, 2006). At other times, state agencies remove children on an emergency basis, although the United States Supreme Court has held that agencies may not be held liable under federal law for failure to remove children from dangerous homes. Conversely, state agencies may be liable for emergency removals in the United States if the removals are not justified by sufficient evidence of risk of immediate harm to the children removed or if adequate process is not provided to parents. Notably, children's wishes are not part of the calculus in the United States; the state's view of child safety is dispositive at every stage.

When state agencies believe that ongoing judicial intervention is warranted, most jurisdictions oversee such cases in special family courts. These special courts subsequently conduct adjudicatory hearings, overseen by judges or referees without juries, to determine if the state has proven, generally by a preponderance of the evidence, that parents have in fact committed conduct falling within the relevant definition of child maltreatment. In most jurisdictions, legal parents have a statutory right to counsel during these proceedings; stepparents and functional parents who are not legal parents may be considered parents for purposes of such hearings in some jurisdictions. Children have a right to participate in such proceedings in many jurisdictions, at times through legal counsel. In the United States, children are not granted that right but federal law requires that children in abuse and neglect proceedings be represented by a 'guardian ad litem,' who may be a lawyer or a lay advocate, and state law often goes beyond federal law to mandate that a lawyer be appointed instead of or in addition to a lay advocate (Atwood, 2005).

When courts find that the evidence is sufficient to support a finding of child maltreatment, they generally have broad discretion to make a disposition in the best interests of the child. Dispositional options include leaving legal and physical custody with parents, provided they submit to state oversight, the state taking legal custody over children but leaving physical custody with parents, and the state taking both legal and physical custody by removing the child from the home and placing the child with a relative or foster parents or in an institution. With respect to the last option, jurisdictions often permit removal only if the state made reasonable efforts to permit the child to remain in the home, generally by offering services to parents. After removal, jurisdictions also often require states to continue to make reasonable efforts to reunite removed children with their parents.

If children cannot be reunited with their parents, jurisdictions generally terminate parents' rights, so that children may enter stable relationships with new caregivers. In the United States, federal law requires states to file termination petitions, and propose potential adoptive parents, when children have been in foster care for 15 of the most recent 22 months. This requirement is motivated by the view that children generally do not thrive while in 'foster care limbo' and that their best interests are served by a relationship with new legal parents when reunification efforts consistently prove unfruitful (Bartholet, 1999). Pursuant to this goal of establishing new legal parent–child relationships, most of the individual states in the United States require that adoption first be ruled out before proceeding to legal guardianship or other planned permanent living arrangements. Other jurisdictions often embrace a wider range of caregiving alternatives for children, including kinship care, when parental rights must be terminated.

To terminate parental rights absent parental consent, courts generally must find a substantive basis for the termination (severe or chronic abuse or neglect, for example) and that termination is in the child's best interests. Some jurisdictions also permit termination in the absence of ongoing parental fault when it would nevertheless be in the best interests of the child, for example, when neglectful parents have been rehabilitated but the child in the meantime has developed strong bonds with a foster family. As in earlier stages of the proceedings, all children subjected to termination proceedings must

either be represented by guardians ad litem or otherwise participate.

Children's Educational Experiences

The United Nations Convention on the Rights of the Child specifies that all children have a right to an education. In the United States, the Supreme Court has held that access to education is not a fundamental right or liberty protected by the US Constitution, but all state constitutions provide that children have a right to free public education and require state legislatures to establish public school systems. All of these rights to education are also obligations, in that children must attend school regardless of their own desires or the desires of their parents.

Beyond compulsory education laws, laws in many jurisdictions seek to ensure access to quality education for all students. The United Nations Convention prohibits all forms of discrimination against children, including in the context of education. In the United States, the Supreme Court has held that all children, including the children of undocumented immigrants, must be able to access state public school systems that are not explicitly segregated by race. Federal law further prohibits discrimination on the basis of sex and language ability in all schools receiving federal funds and requires all states to provide a 'free appropriate public education' to students with disabilities. Public schools in the United States must also provide 'fundamentally fair procedures' to all students facing expulsion.

Within these parameters, jurisdictions remain largely in control of curricula and policies in state-run schools, even in the face of parental opposition. Parents objecting to school decisions generally lose, based on the theory that local school officials represent the majority views of the relevant population and objecting parents may send their children to privately-run schools (Rosenbury, 2007). The robust conception of parental rights discussed above protects parents' decisions to send their children to schools run by nonstate actors, including religious institutions (Gilles, 1996). Parents generally have no right to state subsidies for private school tuition, however.

This general deference to state educational authority often limits children's autonomy as well as the autonomy of their parents. Students, like their parents, largely have not succeeded in challenging curricula or policies in state-run schools. In the United States, for example, students in public schools have failed in challenges to the removal of books from school libraries and in challenges to school dress codes unless those codes discriminate on the basis of religion, race, or (in some states) gender identity or infringe on political speech that is not disruptive. Public school students in the United States also may not engage in public prayer at school events, or otherwise express religious viewpoints during school-sponsored activities. Students participating in extracurricular activities sponsored by local school districts in the United States must submit to drug testing if the school district so mandates.

Students have succeeded in challenging the practices of state-run schools, however, when those practices infringe on students' individual rights and do not promote other legitimate state goals (Dailey, 2011). The United States Supreme Court, for example, has held that public school students must be given the opportunity to opt out of reciting the Pledge of Allegiance at the beginning of the school day. The Supreme Court has also held that students may express opinions in public schools so long as they do not disrupt or otherwise materially interfere with school activities.

Students in schools run by private actors are not guaranteed these protections, as such schools have wide latitude to develop their own curricula and policies. Students also are not entitled to due process upon expulsion from such schools. Private schools generally must meet minimal educational and health requirements imposed by the state, however. Private schools are also subject to jurisdictions' antidiscrimination laws.

Children's Relationships to the Juvenile Justice and Criminal Law Systems

Most jurisdictions maintain special courts designed to respond to misconduct by children. These juvenile courts, operating apart from criminal courts, reflect the view that children accused of crime should not be punished like adults because children are still developing and thus may be more amenable to rehabilitation than are adults. This focus on development and rehabilitation means that juvenile courts often exercise jurisdiction over a wide range of misconduct by juveniles, including conduct that would not be considered criminal if performed by adults (Scott and Steinberg, 2008).

At the same time, many jurisdictions increasingly view children accused of serious crime as responsible and culpable offenders, just like adults. These jurisdictions therefore try some children accused of crime in criminal courts, or they impose adultlike punishments on children in juvenile courts (Feld, 1993). Protections remain for children even in these situations, however. Most jurisdictions have held that children may not be subject to the death penalty and may not be sentenced to life imprisonment (Feld, 2013a).

Children in juvenile courts receive many of the procedural protections extended to adult criminal defendants, but the protections enjoyed by children are not identical to those enjoyed by adults (Birckhead, 2009). Many of these differences are thought to accrue to children's benefit. Juvenile court hearings are often closed and court records kept confidential. Judges are often permitted to impose indeterminate sentences because it may be difficult to predict in advance how much treatment will be needed to rehabilitate the child offender. The United States Supreme Court has also held that judges in both juvenile and criminal courts must take a child's age into account when determining whether a child was subject to custodial interrogation or validly waived their rights to a lawyer or to remain silent. Other jurisdictions, including some states in the United States, presume that children under a specified age cannot waive their rights or cannot do so without consulting a parent or guardian. Most jurisdictions require that parents be notified before state officials interrogate children (Feld, 2013b).

In other ways children accused of misconduct may receive lesser protections than adults receive. Children in juvenile court proceedings may be detained pretrial on broader preventative grounds than those that apply to adults, and children so detained are often not permitted to post bail. Many jurisdictions permit parents to waive counsel on their children's behalf,

although some jurisdictions require advice from counsel or a judge before parents may do so. In addition, although attorneys representing children in juvenile court proceedings are generally instructed to represent their clients' 'expressed interests,' just as they do for adult clients, parents, prosecutors, and judges often push counsel to instead act in their clients' 'best interests.' Finally, in the United States, children in juvenile court do not have a right to a jury trial, and courts have consistently held that parents may validly consent to searches of their children's living quarters even when similar consent between adults sharing a residence, but maintaining separate bedrooms, would not be valid under the Fourth Amendment (Birckhead, 2009).

Children's Interactions with the Market and Civic Life

All jurisdictions restrict the ability of many children to work for wages. These child labor laws are based on the theory that paid work harms children and society at large by preventing children from obtaining the education needed to develop into productive adults. Children's paid work is therefore regulated very differently than adult labor market participation (Humbert, 2009). Employers who violate these laws are generally subject to civil, and at times criminal, liability. Parents may also be subject to criminal prosecution in many jurisdictions, although most child labor laws contain exceptions permitting children to work under parental supervision in many otherwise prohibited jobs. When children legally engage in paid work, their parents are generally entitled to their wages until they reach the age of majority or are otherwise emancipated, on the theory that such wages offset parents' support obligations.

Many jurisdictions deviate from this norm of differential regulation of child and adult labor in one employment context, prostitution. Although the United Nations Convention on the Rights of the Child requires states to protect children from sexual exploitation, and the United States similarly has robust laws criminalizing child sexual abuse, jurisdictions that criminalize prostitution tend to treat children engaging in sex work the same as adults. This equal treatment has been criticized on multiple grounds (Birckhead, 2011). Some argue that the practice is illogical, as some children are prosecuted for prostitution even when they are under the minimum age for consenting to sexual intercourse. Others argue that children, unlike adults, are incapable of consenting to prostitution. Defenders of the practice emphasize, however, that the prosecution of child prostitutes keeps vulnerable children off the streets and induces children to cooperate with the prosecution of their pimps.

Outside of the employment context, children may enter into contracts and enforce them against other parties, but many jurisdictions also permit children to disavow most contracts at any time during their minority, or within a reasonable time after becoming adults, by invoking the infancy defense. To disavow the contract on this basis, children must return the purchased goods that remain in their possession, if any; if the contract was for necessaries, children must also return whatever value they received from the goods. Some jurisdictions go farther, requiring children to make restitution for the benefits received from any disavowed contract or to pay quasi-contractual damages in some contexts beyond the context of necessaries. Children are thought to be protected by this general approach, but the availability of the infancy defense also makes many parties reluctant to enter into contracts with children (Preston and Crowther, 2012).

At the same time, children increasingly engage in commerce, particularly electronic commerce, and companies often aggressively advertise to children as consumers. Some jurisdictions impose additional regulation on advertising targeted to children, given that children may be more subject to manipulation than adults. Some jurisdictions also impose additional regulation designed to protect the privacy of child consumers. In the United States, for example, federal law requires web sites and other online services that knowingly collect personal information to obtain parental consent before collecting such information from children under 13 years of age.

In other market contexts, children's decisions are accorded the same respect extended to adults. In many jurisdictions, children on their own may seek contraceptive, prenatal, and abortion services, as discussed in more detail in the next section, and may seek treatment for sexually transmitted diseases or drug and alcohol abuse (Hill, 2012). In the United States, the Supreme Court also recently held that states may not limit the sale or rental of violent video games to children. In reaching that holding, the Court emphasized that although "a State possesses legitimate power to protect children from harm," that power "does not include a free-floating power to restrict the ideas to which children may be exposed" (Brown v. Entertainment Merchants Ass'n, 131 S. Ct. 2729 (2011)). In a similar vein, courts have held that schools may not limit student speech on social networking sites, so long as students did not intend for the speech to reach the school and there is no reasonable forecast that the speech would cause substantial disruption at school (Waldman, 2011).

In addition to engaging with the market, children may inherit property or receive property as gifts, and parents do not automatically enjoy legal power over such property in many jurisdictions. Children holding sole title to such property often encounter problems effecting transfers or sale, however, given that the contracts of children are voidable, as discussed above. In addition, children generally may not execute valid wills or file lawsuits until they reach the age of majority or are otherwise emancipated (Cunningham, 2006). Given such issues, courts may empower parents or other guardians or custodians to manage the property subject to fiduciary duties or court supervision. In order to achieve more flexibility, some donors or decedents convey property to children through trusts, wherein the trustee holds title to the property and manages it for the benefit of the child beneficiary.

Children's participation in civic life is more limited than in the market. Although all jurisdictions recognize children as citizens, if they or their parents meet certain requirements, that citizenship does not enable children to vote. More than a dozen nations have lowered the voting age to 16 years for purposes of at least some elections, but no jurisdiction has permitted children less than 16 years of age to vote (Hamilton, 2012). Some believe that children's political interests are represented by their parents, but parents neither receive more votes than

other adults, nor are parents permitted to vote by proxy on behalf of their children.

Children's Peer Relationships

Law focuses most of its attention on children's relationships with adults, whether they are parents, other adults serving as parents at the direction of the state, teachers, or juvenile justice personnel. Law generally does not recognize or otherwise respond to children's relationships with peers, except in two, relatively narrow circumstances: sibling relationships and harmful peer relationships.

Jurisdictions confer legal status on sibling relationships, and some states in some circumstances recognize stepsibling relationships. The effect of that status is often confined to the context of inheritance law, however. Laws in many jurisdictions direct judges to take sibling relationships into account, by keeping siblings together or arranging for meaningful visitation, when making custody or maltreatment decisions, but attempts to keep siblings together often fail, particularly in maltreatment proceedings. In the face of such failure, most jurisdictions do not guarantee that siblings may live with one another or even visit one another in different households (Hasday, 2012).

Accordingly, if the parent–child relationship falls apart, the sibling relationship is not always protected. Moreover, the value of sibling relationships is often assumed to flow primarily from the common parent–child relationship. For example, some states in the United States prohibit judges from terminating parental rights when doing so would substantially interfere with sibling relationships. Yet to preserve sibling ties, children in this context are moved into guardianship instead of termination of parental rights and adoption. Although some children may benefit from guardianship, other children may benefit more from adoption. By protecting sibling ties only by preventing the termination of parental rights, parents may benefit more than children.

Organ donation between siblings also raises complex legal and ethical issues. Siblings often provide the best match as donors, and parents often want one sibling to donate to another. In disputed cases, courts generally look at the best interests of the donor child, sometimes applying the substituted judgment rule under which the court determines whether the child would agree to be a donor if the child were old enough to make the decision (Hill, 2012). In reaching this conclusion, courts tend to consider the psychological relationship between the siblings, whether the donation will help keep the relationship intact, and the physical risk to the child donor.

Outside of the context of sibling ties, jurisdictions generally respond to children's peer relationships only when those relationships are perceived to be threatening to the children themselves, to other children, or to larger communities. In such situations, law often authorizes state intervention designed to end the peer relationship and alleviate the harm. For example, the US Supreme Court has interpreted federal law to impose tort liability on schools that fail to protect students from harassment by other students on school grounds. Many jurisdictions have also adopted antibullying statutes and policies and passed various forms of antigang legislation (Silbaugh, 2013).

Other forms of state intervention are less directly coercive, but they still attempt to alleviate the harm presumed to flow from children's peer relationships. By passing curfew ordinances that limit the hours in which children may congregate outside of the home, for example, jurisdictions implicitly recognize that children seek to interact with one another but seek to limit that interaction in order to prevent potential harm. In the United States, children are generally permitted in public after curfew hours only when accompanied by a parent, engaging in an organized activity approved by a parent, or running an errand for a parent, thereby reinforcing the primacy of the parent–child relationship while simultaneously positing children's unsupervised peer relationships as threatening and unsafe (Federle, 1995).

Sexual relationships between children are often viewed as particularly threatening. Some jurisdictions criminalize teen sexual behavior in the hope that potential sanctions will deter some children (generally boys) from pursuing sex and encourage other children (generally girls) to resist. Children often are not prosecuted under such statutes, however, and the few prosecutions that occur often appear motivated by concerns about particular sexual practices, including those between same-sex partners and partners of different races and those that result in pregnancy (Goodwin, 2013). In the United States, some children have been charged with violations of child pornography laws when they sext one another, even when they are sending their own images to peers (Sweeny, 2013–14).

Some jurisdictions, in particular the United States, also do not affirmatively provide children with the means to manage the reproductive and health consequences of sexual relationships. Public schools in the United States often offer limited sex education, and none of the states mandate that schools provide children with condoms or other contraception (McClain, 2006). Some public schools in the United States have chosen to distribute condoms on their own, however, and courts have rejected parental challenges to such distribution so long as schools permit parents to veto their children's access.

In other ways, jurisdictions address the consequences of children's sexual relationships as part of public health initiatives (Hill, 2012). Most jurisdictions, including the United States, allow children to consent to treatment for sexually transmitted diseases or drug and alcohol abuse without requiring them to notify their parents. Many of these jurisdictions also mandate that girls be vaccinated against the HPV virus, which is transmitted sexually. Many jurisdictions similarly allow children of certain ages, generally 12 and older, to seek contraceptive services without parental involvement. Children's access to abortion is more tightly restricted, particularly in the United States, where only two states explicitly allow all pregnant children to consent to abortion on their own, without parental involvement.

Assessing Law's Role in Children's Lives

This general approach to allocating legal authority over children serves multiple goals. Law treats children as a distinct legal category because children are legal subjects with vulnerabilities, developmental needs, capacities, interests, and experiences often different from those of adults. Law defers to parental

prerogatives in most contexts in order to encourage parents to invest in and provide for their children, both in terms of responding to children's incapacities and dependencies and in developing the close relationships necessary for nuanced decisions about each child's unique circumstances (Buss, 2004; Dailey, 2006). Robust norms of family privacy also foster diversity and pluralism (Cooper Davis, 1994; Galston, 1998). At the same time, the state retains an interest in protecting and educating children in order to ensure they are properly prepared for the obligations of citizenship.

Debate remains about whether law could better achieve each of these goals and whether law's current focus frustrates other goals. In particular, law's focus on the appropriate allocation of childrearing authority between parents and the state may create a perception of children as objects to be possessed (Woodhouse, 1992) which, in turn, may detrimentally affect children's perceptions of their own autonomy. Law's emphasis on children's dependency, and adults' responsibility for children's welfare, may similarly overlook the ways that children gradually attain the ability to act autonomously. Accordingly, some lawmakers and scholars increasingly have recognized children's developing capacities, positing a continuum between infancy and adulthood, with children acquiring greater decision-making abilities as they grow older (Dailey, 2011; Zimring, 1982). As discussed above, this approach currently comes into play most often with respect to state regulation of adolescent decision making outside the home, particularly in the market and when children are accused of serious crimes, but the approach could also be invoked to support children's agency even within the family.

Law has yet to develop a systemic approach for deciding whether children's dependencies or capacities should govern particular contexts or a mechanism for combining the strengths of each approach. Doctrinal inconsistencies therefore abound (Appleton, 2014). For example, many children accused of committing crimes are adjudicated in juvenile courts on the theory that they lack the capacity and maturity to be held fully responsible for their crimes. This treatment resembles the legal treatment of children in other areas where doubts about children's decision-making abilities exist, such as voting, driving, and sexual relationships. At the same time, doctrines in other areas, such as free speech and reproductive rights, reflect greater deference to children's decisions, and many children are tried in criminal courts based on the severity of their crimes alone without regard to their decision-making abilities. Approaches privileging children's dependency over their capacities, or their capacities over their dependency, thus have failed to produce a coherent body of law governing children's lives.

In addition, these approaches are limited in their aspirations, emphasizing children's dependencies and capacities while minimizing other important interests. Dependency and autonomy do not capture the full range and complexity of children's experiences. Children also live active lives in the here and now, experiencing a full range of emotions and interactions as children, separate from or in addition to their interests in becoming adults, even as they are dependent on adults for many aspects of their lives. By focusing almost entirely on children's relationships with adults or the contexts in which children might be treated like adults themselves, lawmakers and scholars often overlook the ways that children are more than passive recipients of adult childrearing until they become adults themselves. Law thus minimizes children's intrinsic experiences.

This article offers a broad survey of the diverse ways that law governs children's lives as a way of understanding existing frameworks of children and law and envisioning new ones. Law does more than regulate children within the family or in state institutions. Instead, law mediates children's interactions with multiple actors in multiple spheres across public and private divides. Law's traditional approach of allocating decision-making authority between parents and the state serves many purposes, but law does and can do more to respond to and facilitate the full range of children's needs, desires, and experiences.

See also: Civil Liberties and Human Rights; Family Law in the United States; Family and Schooling; Juvenile Justice: International Law and Children's Rights; Law and Everyday Life; Prostitution and Sex Work; Religion and Youth; Religion, Family, and Kinship; Rights: Legal Aspects; Special Education in the United States: Legal History.

Bibliography

Appell, Annette R., 2013. Accommodating childhood. Cardozo Journal of Law & Gender 19 (3), 715–779.

Appleton, Susan, 2014. Restating childhood. Brooklyn Law Review 79 (2), 525–549.

Atwood, Barbara A., 2005. Representing children: the search for clear and workable standards. Journal of the American Academy of Matrimonial Lawyers 19 (2), 183–222.

Bartholet, Elizabeth, 1999. Nobody's Children: Abuse and Neglect, Foster Drift, and the Adoption Alternative. Beacon Press, Boston, MA.

Bartholet, Elizabeth, 2011. Ratification by the United States of the convention on the rights of the child: pros and cons from a child's rights perspective. Annals of American Academy of Political and Social Science 633 (1), 80–101.

Birckhead, Tamar R., 2009. Toward a theory of procedural justice for juveniles. Buffalo Law Review 57 (5), 1447–1513.

Birckhead, Tamar R., 2011. The 'youngest profession': consent, autonomy, and prostituted children. Washington University Law Review 88 (5), 1055–1115.

Blauwhof, Richard J., 2009. Foundational Facts, Relative Truths: A Comparative Law Study on Children's Right to Know Their Genetic Origins. Intersentia, Antwerp.

Buss, Emily, 2004. Allocating development control among parent, child and the state. University of Chicago Legal Forum 2004 (1), 27–55.

Cahn, Naomi R., 1999. Models of family privacy. George Washington Law Review 67 (5), 1225–1246.

Cahn, Naomi R., 2012. The new kinship. 100 Georgetown Law Journal 100 (2), 367–429.

Cooper, Davis P., 1994. Contested images of family values: the role of the state. Harvard Law Review 107 (6), 1348–1373.

Cunningham, Larry, 2006. A question of capacity: towards a comprehensive and consistent vision of children and their status under law. UC Davis Journal of Juvenile Law & Policy 10 (2), 275–378.

Dailey, Anne C., 2006. Developing citizens. Iowa Law Review 91 (2), 431–503.

Dailey, Anne C., 2011. Children's constitutional rights. Minnesota Law Review 95 (6), 2099–2179.

Dwyer, James G., 2006. The Relationship Rights of Children. Cambridge University Press, New York.

Federle, Katherine H., 1995. Children, curfews, and the constitution. Washington University Law Quarterly 73 (3), 1315–1638.

Feld, Barry C., 1993. Criminalizing the American juvenile court. Crime & Justice 17 (1), 197–280.

Feld, Barry C., 2013a. The youth discount: old enough to do the crime, too young to do the time. Ohio State Journal of Criminal Law 11 (1), 107–148.

Feld, Barry C., 2013b. Real interrogation: what actually happens when cops question kids. Law & Society Review 47 (1), 1–36.

Freeman, Michael D.A., 2010. Upholding the dignity and best interest of children: international law and the corporal punishment of children. Law & Contemporary Problems 73 (2), 211–251.

Galston, William A., 1998. The legal and political implications of moral pluralism. Maryland Law Review 57 (1), 236–250.

Gilles, Stephen G., 1996. On educating children: a parentalist manifesto. University of Chicago Law Review 63 (3), 937–1034.

Goodwin, Michele, 2013. Law's limits: regulating statutory rape. Wisconsin Law Review 2013 (2), 481–540.

Guggenheim, Martin, 2005. What's Wrong with Children's Rights. Harvard University Press, Cambridge.

Guggenheim, Martin, 2006. Ratify the U.N. Convention on the Rights of the Child, but don't expect any miracles. Emory International Law Review 20 (1), 43–68.

Hamilton, Vivian E., 2012. Democratic inclusion, cognitive development, and the age of electoral majority. Brooklyn Law Review 77 (4), 1447–1513.

Harris, Leslie J., Waldrop, Dennis, Waldrop, Lori R., 1990. Making and breaking connections between parents' duty to support and right to control their children. Oregon Law Review 69 (3), 689–739.

Hasday, Jill E., 2012. Siblings in law. Vanderbilt Law Review 65 (3), 897–931.

Hill, Jessie B., 2012. Medical decision making by and on behalf of adolescents: reconsidering first principles. Journal of Health Care Law & Policy 15 (1), 37–73.

Humbert, Franziska, 2009. The Challenge of Child Labour in International Law. Cambridge University Press, New York.

Huntington, Clare, 2006. Rights myopia in child welfare. 53 UCLA Law Review 53 (3), 637–699.

McClain, Linda C., 2006. The Place of Families: Fostering Capacity, Equality, and Responsibility. Harvard University Press, Cambridge.

Meyer, David D., 2000. The paradox of family privacy. Vanderbilt Law Review 53 (2), 527–595.

Meyer, David D., 2006. Parenthood in a time of transition: tensions between legal, biological, and social conceptions of parenthood. American Journal of Comparative Law 54 (1), 125–144.

Mnookin, Robert H., 1975. Child custody adjudication: judicial functions in the face of indeterminacy. Law & Contemporary Problems 39 (3), 226–293.

Preston, Cheryl B., Crowther, Brandon T., 2012. Infancy defense inquiries. Santa Clara Law Review 52 (1), 47–80.

Roberts, Dorothy, 2002. Shattered Bonds: The Color of Child Welfare. Basic Books, New York.

Rosenbury, Laura A., 2007. Between home and school. University of Pennsylvania Law Review 155 (4), 833–898.

Scott, Elizabeth S., Steinberg, Laurence, 2008. Rethinking Juvenile Justice. Harvard University Press, Cambridge.

Scott, Elizabeth S., Emery, Robert E., 2013. Gender politics and child custody: the puzzling persistence of the best interest standard. Law & Contemporary Problems 76 (3), forthcoming.

Silbaugh, Katherine, 2013. Bullying prevention & boyhood. Boston University Law Review 93 (3), 1029–1047.

Sweeny, JoAnne, 2013–14. Sexting and freedom of expression: a comparative approach. Kentucky Law Review 102 (1), 103–146.

Todres, Jonathan, 2012. Maturity. 48 Houston Law Review 48 (5), 1107–1165.

Wadlington, Walter, 1994. Medical decision making for and by children: tensions between parent, state, and child. University of Illinois Law Review 1994 (2), 311–336.

Waldman, Gold E., 2011. Badmouthing authority: hostile speech about school officials and the limits of school restrictions. William & Mary Bill of Rights Journal 19 (3), 591–659.

Woodhouse, Barbara B., 1992. 'Who owns the child?': Meyer and Pierce and the child as property. William & Mary Law Review 33 (4), 995–1122.

Zimring, Franklin E., 1982. The Changing Legal World of Adolescence. The Free Press, New York.

Children, Rights of: Cultural Concerns

Michael S Wald, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the historical development of children's rights worldwide. It then describes the different categories of claims made under the rubric of children's rights and discusses value issues that need to be resolved as governments, courts, and citizens seek to define children's status in each society. It concludes by examining the challenges countries face implementing any such rights.

The last 100 years have witnessed enormous progress in the recognition of individual human rights throughout the world. In addition, various identity groups, including women, racial, ethnic, linguistic, cultural and religious minorities, the disabled, and gay males and lesbians all have claimed, and received to varying degrees, rights to equal treatment under law, nondiscrimination in employment in both the public and private sectors, and fuller integration into the economic and political life of their countries.

It is perhaps not surprising, therefore, that a worldwide movement to recognize children as a distinct subgroup with claims of rights also occurred. In 1924, the League of Nations adopted the first Declaration of the Rights of Children. In 1989, the United Nations General Assembly adopted the UN Convention on the Rights of the Child (UNCRC) (UN Convention, 1989), which has been called the 'most comprehensive and detailed of all … international human rights instruments' (Alston, 1994: p.1). The Convention has been ratified by 192 of the 194 UN members; the United States and Somalia have not ratified. In the United States, conservative political groups, concerned that the CRC (Convention on the Rights of the Child) undermines parental authority and gives government too much power over families, have contributed to blocking ratification. In Somalia, there is no internationally recognized government, which precludes ratification.

Despite the seemingly widespread acceptance of the concept of children's rights, the idea that children, individually or as a class, should have *rights* (claims enforceable against others) is actually revolutionary and problematic both in philosophical and practical terms. Determining what rights to provide children requires that societies confront fundamental value issues quite different from those entailed in recognizing other groups' claims.

First, children generally are not autonomous; they are dependent; moreover, the period of dependency has lengthened with modernity. Ideally, they are part of families; their life chances are greatly affected by the care they receive from their families. Parents need authority to fulfill these responsibilities. Some highly respected child development specialists argue that giving children rights may be incompatible with sound child development (Goldstein et al., 1973). In addition, giving children rights may conflict with parental rights and privacy. However, parents do not always act in their children's best interests and the community as a whole has an interest in the development of future generations. Children may have their own views and desires. Therefore, it must be decided how responsibility for the well-being and upbringing of children should be allocated between parents, families, the child, and the community as a whole.

Second, intellectual and emotional capacities develop gradually, for both biological and social reasons. These limits in capacity may conflict with claims that children should have the same liberties and privileges as adults (Purdy, 1992). A theory of children's rights must address the question, what entitles persons to be full participants in a country's civil and political life?

The term children's rights encompasses five different claims. Most important is the right to be recognized as persons and therefore entitled to basic human rights, including the right not to be 'owned' by another. The worldwide acceptance of this premise reflects a fundamental shift in the conception of children. However, this right may conflict with parents' right to form their children's values and life plans. Second is the right to have government guarantee their basic well-being, by assuring adequate food, shelter, and health care. While children clearly need these goods, there is disagreement whether such support should be considered a right or a moral obligation of citizens to each other. Third is the right to protection from inadequate care from caretakers. While this right is accepted everywhere, the definition of inadequate and the degree of justifiable state intrusion to protect children remains highly debated. The fourth category focuses on liberty and participation rights, such as the right to freedom of expression and the right to vote. Supporters of liberty rights must develop a theory of why children should be granted some rights and not others. The final category involves children's right to act independently of parental control. There is no easy resolution to the tension between parental rights and children's rights. Children cannot have total freedom from parental control, both because this would often be harmful and the state cannot regulate most intimate family relations. There may be areas, however, where the consequences of not giving children rights outweigh the harms of limiting parental oversight of their children, for example, the right to receive certain medical care or access to contraceptives without parental permission.

The Historical Context

For much of history children were viewed largely as nonpersons, the property and responsibility of their parents, who had a right

to control their upbringing, even their very existence (Eekelaar, 1986). Children (usually defined as persons under 21 or 18 years of age) also could not participate in the political or civic life of their countries.

In the mid-1800s, some countries began assuming state responsibility for promoting and protecting children's well-being. Initially, this entailed providing free public education and passage of laws protecting children from severe physical abuse by their parents. Over the next 100 years, Western industrialized countries adopted more laws affecting the status of children, including laws limiting child labor in settings outside the family, passage of compulsory school attendance laws, and greater state efforts to protect children from parental maltreatment. The idea that childhood was a period for gradual maturation and lesser responsibility also was reflected in the creation of the juvenile court, which was directed to treat, not punish, children who violated the law (Hawes, 1991).

While these policies were advocated in the name of children's rights, and certainly contributed to their well-being, these were not the types of civil rights that other groups were demanding and receiving. The focus was on the protection of children, not their autonomy or fuller integration into the economic and political life of their countries. Moreover, except where their behavior was considered inimical to the social order, parents continued to have virtually total authority over their children's upbringing. In the United States, where the movement toward child protection was most evident, the US Supreme Court ruled early on that parents had a constitutional right to control their children's upbringing (Meyer v. Nebraska, 1923).

In fact, many 'children's rights' laws often were supported primarily as a means of protecting adult interests. For example, child labor laws and compulsory schooling were championed by newly emerging labor unions, which viewed the availability of child labor as a barrier to higher wages for adults (Woodhouse, 1992). Proponents of the juvenile court promoted it as a needed intervention to protect society from children who were not being adequately supervised by their parents, often poor, immigrant parents (Schlossman, 1977).

Since the 1960s, most countries have continued expanding the protection of children's well-being. The establishment of a right to education for virtually all children throughout the world increased state efforts to guarantee children a minimal level of material well-being, and greatly expanding state intervention into family privacy through child abuse laws that have all benefited children. Moreover, during this period, in many countries children were recognized as rights holders with respect to freedom of speech and other civil liberties, were given a greater voice in decisions affecting their lives, such as custody decisions when parents divorce, and in some situations have been given a degree of legal autonomy over behaviors once controlled by their parents, for example, access to contraception.

Yet the scope of applicable rights remains highly contested. At the extreme, some commentators have argued for 'children's liberation,' with the complete elimination of distinctions between children and adults with respect to all political, social, and familial rights (children, as reflecting of their status, usually have not been the demanders of these rights, in contrast to other groups). For example, they argued that children should have the right to vote, to decide whether to go to school, and to a share of the family income to spend as they choose (Farson, 1974).

Few proponents of children's rights go this far. Most advocates propose limiting, but not eliminating, parental autonomy in controlling their children's lives and giving children increasing autonomy as they mature, in recognition that capacities develop over time. Yet even the more limited types of claims of rights raise profoundly difficult issues.

What Is Meant by Children's Rights?

As noted, the term children's rights is used to encompass a variety of different claims (Campbell, 1992). Rights for adults generally involve claims to be autonomous. While this also is true of some children's rights, often the term is used to support the argument that adults should act in a manner that protects children's best interests. Distinguishing these categories reveals the tensions in the general concept and helps identify the issues related to the implementation of various types of claims. Often this entire debate revolves around who is in the best position to determine the child's 'best interests' – the state, the family, or the child?

Basic Human Rights

The first claim is that children are entitled to basic human rights, including the right to life itself, the right not to be 'owned' by another, the right to freedom from torture or other cruel punishment, and the right to be free from discrimination based on race, color, gender, religion, or national, ethnic, or social origin. Such rights are widely accepted for adults, despite continuing problems with implementation, especially with respect to nondiscrimination based on gender, race, and ethnicity. Since these rights are not tied to competence or capacity, there is no reason for denying them to children on this basis. The critical step is accepting that children are *persons*, not chattels or incomplete human beings. By adopting the 1989 CRC virtually all countries accepted this fundamental shift in the conception of children, although the ongoing differential treatment of girls, including the practice of infanticide and limiting schooling, remains a significant problem in many countries.

Moreover, application of these rights to children is not entirely unproblematic, at least with respect to the relationship between parents and children. Few countries still consider children to be the property of their parents. However, many people agree with Fried that "the right to form one's child's values, one's child's life plan … are extensions of the basic right not to be interfered with in doing these things for oneself" (1978). Yet, such a right constitutes a form of 'ownership' and may conflict with the child's right to an 'open future' and with society's interest in how future adults are socialized (Feinberg, 1980). The increase in home schooling in the United States beginning in the 1990s and the denial of education to girls in many countries are reflections of the tension between a liberal view that children should be given the opportunity to become autonomous adults able to choose their own ends and the view, often derived from religious values, that children should

be raised in a manner that maximizes the likelihood that they will accept specific religious or cultural norms. (I am not sure where the sentences I have deleted came from: they do not seem to fit here.)

There is no easy resolution to this conflict about the location of rights and responsibilities. Unless the state were to intervene in family life in ways that would be deemed unacceptable in most countries, parents will continue to make critical decisions for children, regardless of the child's wishes. As noted, respected child development experts assert that parents need autonomy to successfully raise their children. Moreover, deference to parental or family authority may be demanded by a society's political, cultural, or religious traditions. In some countries, including the United States, supporting parental autonomy is seen as critical to supporting political and cultural diversity. In others, the proper role of parents may be seen as teaching children to accept national cultural values and traditions, not as helping children develop into 'autonomous' adults (see generally Alston, 1994). Many countries have reserved the right to apply the CRC in light of their own cultural or religious laws, although there is evidence that the CRC is influencing these values (*International Journal of Children's Rights*, 1995). Whether children's rights are universal or culturally specific likely will remain one of the major debates in implementing the CRC. For recent studies, see the chapters in the edited volume by Invernizzi and Williams (2011).

In fact, given that children's values and perspectives are malleable and constructed, someone will shape their values and personhood. Parents, families, neighborhoods, peers, government institutions such as schools, and general cultural norms all play a part. Thus, the issue is what are the appropriate roles of the family and the state in making critical decisions, such as those regarding schooling, religion, where the child shall live, not whether children should have the right to control their own upbringing (Coons et al., 1991). This question goes to the heart of how a society defines the relationship between the state and the individual, adults, and children alike.

Children also are denied some basic civil rights available to adults because societies have an interest in how they are socialized. At a minimum, children are compelled to go to school, thus restricting their liberty rights. Finally, there is still substantial debate over the status of the fetus with respect to abortion and maternal behavior during pregnancy. Because no consensus could be reached on this issue, the CRC leaves the question to each country.

Despite these problems, the recognition that children are persons, with rights, along with the acceptance of the basic premise of the CRC that in all decisions affecting children their 'best interests' shall be a primary consideration, constitutes a great advance that clearly benefits the vast majority of children.

Welfare Rights

A second category of claimed rights is that government should guarantee children a basic level of well-being; these are often called welfare rights. The CRC provides that children should have a right to "a standard of living adequate for the child's physical, mental, spiritual, moral, and social development" (Article 27), "the enjoyment of the highest attainable standard of health…" (Article 24), and to education (Article 28). The CRC goes beyond these basic goods, and seeks to afford children a broader set of welfare rights, including the right "to rest and leisure, to engage in play and recreational activities appropriate to the age of the child and to participate freely in cultural life and the arts" (Article 31).

Since the 1950s, governments have assumed increasing responsibility for guaranteeing that children have adequate food, nutrition, shelter, and health care. In many countries, children benefit from general social welfare policies, which provide all citizens with universal health insurance and various forms of income support. Most industrialized countries also have partially socialized the cost of caring for children through children's allowances, paid parental leave after the birth of a child, and public provision of day care. Many less economically developed countries are moving in the same direction, despite limited financial resources.

The laggard in this regard, given its capacity, is the United States. Despite its wealth, approximately 20% of children in the United States live in poverty. Resistance to greater income support for children in the United States reflects, in part, the belief of many political figures that childhood poverty is caused by 'irresponsible' adult behavior, such as having children out of wedlock, and that it is impossible to provide goods to children without also giving them to adults, thereby 'rewarding' the adults' actions. In addition, many politicians support the premise of libertarian theorists like Robert Nozick that there is no moral justification for the redistribution of wealth (1974).

Even in countries that accept a welfare state, the level of support that a child should receive remains largely a political issue, not a legal right enforceable by courts. No matter how wealthy a country, trade-offs between levels of socially provided health care, housing, education, and other welfare goods are now seen as inevitable. Courts, appropriately, are reluctant to get too deeply into these decisions. Since the scope of the claimed right is so vague, courts lack the legitimacy and authority to order elected governments to provide people with specific levels of well-being. In addition, some theorists contend that, in the political process, framing political debates in terms of moral obligations that the citizens of a country pledge to provide to each other may prove more productive than trying to establish rights (O'Neill, 1988; cf Freeman, 1992).

Another critical question is whether welfare rights should encompass the right to 'equal opportunity.' Opportunity focuses on each child's options in adulthood. Social scientists have struggled to identify the goods that must be provided to children during childhood in order to equalize their opportunities as adults. Much of the focus is on education. But opportunity is also highly influenced by the quality of parenting a child receives and the social, economic, and cultural environment in which the child is raised. A right to equal opportunity is far more difficult to guarantee than a right to basic necessities and may require restricting parental rights (Fishkin, 1983).

Protection Rights

A third category of claimed rights for children is protection from inadequate care from their caretakers, especially their

parents. Since giving children more protection changes their relationship with their parents and the state's relationship with parents and families, this right raises fundamental issues quite different from welfare rights.

The degree of justifiable state intrusion against a parent's will remains highly debated. Should the state treat parents as 'stewards' or 'trustees' of their children, and seek to ensure that parents provide children a high level of care, or should the state be the protector 'of last resort,' intervening only if parental care is clearly harmful?

Most countries have decided that coercive intervention should be limited to situations where the actual or threatened harm to the child's physical or emotional well-being is substantial. This approach reflects the judgment that unwanted intervention is a drastic step from the child's, as well as the parent's, perspective. It also is based on evidence that the care provided to children removed from their families sometimes is as bad as or worse than the parental care (Wald, 1982). Even when some form of state intervention is needed to protect children, many countries are following the practice, first established in New Zealand, of maximizing the involvement of parents and extended family in deciding how best to protect the child in the future. In the United States, there is a preference for placement with kin when a child must be removed from the parent's home.

A few countries have adopted laws that seek to provide more extensive protections for children. Five countries have banned any corporal punishment of children; while these laws do not carry any significant sanctions for violation, they reflect a different conception of the rights of children and the role of the state. In recent years, evidence that early child rearing can substantially affect a child's brain development and capacity for school performance and social regulation has led to increased calls for more extensive monitoring of how each child is being reared (Shonkoff et al., 2011).

In the future, governments will continue to struggle with the appropriate grounds for protective intervention. The critical task, however, will be to develop and connect children to social programs that enhance all children's well-being and future opportunities without resorting to coercive interventions. By the time protection rights must be invoked, the child has suffered substantial harm and the capacity of the state to rectify the situation is often limited (Lonne et al., 2009).

Liberty and Participation Rights

The fourth category of rights does focus on autonomy. In virtually all countries, persons under 18 lack the right to vote, they are restricted in making decisions about whether to work (but compare the widespread use of child labor in poorer countries), and in their freedoms of expression and other civil rights. In addition, in many countries children accused of criminal behavior generally are denied some due process rights available to adults in that country, for example, the right to a jury trial; these limitations have been justified as necessary to promote the rehabilitative ideals of the juvenile justice system.

In 1967, the US Supreme Court determined that children are entitled to at least some of the rights guaranteed by the free speech provision of the US Constitution (Tinker v. Des Moines Independent Community School District, 1969). Since then the liberty rights of minors have been gradually increased in the United States and many other countries. The CRC provides that: "(the) child shall have the right to freedom of expression; this right shall include the freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of the child's choice" (Article 13); "(s)tates parties shall respect the right of the child to freedom of thought, conscience, and religion" (Article 14); "(s)tates parties recognize the rights of the child to freedom of association and to freedom of peaceful assembly" (Article 15). Notably, the CRC does not provide adult status in other areas, such as the right to vote, marry, work, or to be free from constraints imposed only on children, such as the obligation to attend school, or to be at home by a certain hour (curfew laws).

Courts, legislatures, and the drafters of the CRC have not arrived at any consistent theory justifying granting children some of these rights and not others. Nor have they resolved the tension between claims for special protection and for full participation. If adolescents are seen as having the maturity and capacity to make decisions about their health care, to choose what they read or see, or to vote, does it not then follow that they should be held fully responsible for their antisocial behavior? While few child advocates would go so far, this claim was invoked in the United States to justify prosecuting juveniles who commit serious offense in adults courts.

The critical issues relate both to capacities and to questions about developmental needs. For some rights, such as voting or marrying, it is generally accepted that some minimum level of competence is necessary to the proper exercise of the right. Children often are denied other rights, such as unlimited access to books or movies, on the assumption that early exposure to certain materials or ideas may be harmful to them.

An overriding question is whether it makes sense to assume that the same age line should govern access to all political and civil rights. Zimring has proposed treating children as possessing a 'learner's permit,' whereby they are given increasing autonomy and responsibility over time (1982). Different rights might be available at different ages, based on best guesses about average capacity. Age lines are inherently arbitrary; not all people attain capacities at the same age. But individualized decisions about whether someone is old enough to vote, marry, or leave school are likely, in practice, to be even more arbitrary, as is having a single age trigger all rights.

Autonomy within the Family

The issues regarding autonomy rights are even more difficult with respect to the right of children to act independently of their parents' wishes. The extreme view, that children of any age should be free to make their own decisions, has not received acceptance. In many countries giving children any autonomy within the family would be incomprehensible. In other countries, there are debates about whether children should have a right to make certain major decisions, such as whether to have an abortion, to have access to contraceptives, or to receive treatment for substance abuse, without their parents' permission or knowledge. Advocates of giving children these rights contend that many children will be reluctant to seek out needed services if their parents are informed. Children have

also been given some rights to determine important aspects of their life, such as which parent to live with in the case of divorce-or at least to have their voice heard in legal proceedings affecting their interests. The case for giving children rights is strongest with respect to decisions where parents may not be counted upon to generally act in their children's best interests.

There is no easy resolution to the tension between parental rights and children's rights. It is unrealistic to assume that children can have total liberty rights, enforceable against their parent's opposition. Asking courts to enforce most such rights would be a waste of court resources and the orders would be unenforceable as a practical matter. But giving children some rights is feasible, for example, access to drug treatment or abortion without parental permission. The question is whether the benefits of giving children autonomy with respect to the decision outweigh the harms of limiting parental oversight of their children. These rights will have to reflect the legal and social structure of each country and should not be thought of as universal children's rights.

The Future

The situation of children throughout much of the world improved substantially during the twentieth century. Concepts of rights likely contributed to these advances. The recognition that children are independent human beings with claims to special attention and resources is a major advance with respect to universal recognition of human rights (Freeman, 1992).

However, further development of children's rights will require resolving the value issues often masked by rights talk. Despite the unprecedented rapidity of the ratification of the CRC, the fact that the contours of many of the proposed rights are not very clearly defined will make enforcement and monitoring by the international bodies problematic. In fact, the very vagueness of the obligations may account for the rapid ratification and almost total support of the CRC, despite the fact that many of the proposed rights seemingly conflict with deeply held cultural and religious values in many countries. Countries expect to interpret the provisions in light of their customs and values. But progress is definitely being made; for example, many countries have children's rights ombudspeople or children's commissioners that seek to educate the public and enforce the Convention.

With respect to welfare rights, those concerned with helping children must recognize that realization of most of the proposed rights requires policies that help families, not just children (Luttrell-Rowland, 2012). Children's well-being generally is inseparable from that of the parents. Greater provision of welfare rights should lessen the occasions that protection rights will need to be invoked. However, in establishing welfare rights, countries need to ensure that when adults seek to act on behalf of children, their own needs do not take precedence.

With respect to all general social policies, countries must think through the implications of the fact that children, who may constitute a third to half of the population, cannot vote. The poorest children are further disenfranchised by the general powerlessness of their parents. Some countries are making significant efforts to include the voices of children in discussions of issues related to them and to more general public policy issues. For example, in the United States, youths are often members of commissions created to assess public policies and a number of cities have youth commissions. Children were involved in Nepal when it considered adoption of the CRC. The recognition of children's right to speak for themselves could significantly alter the place of children in the state (Melton and Limber, 1992).

Many countries continue to deal with problems of street children, child prostitution, child labor, and infanticide. The advancement of children's rights in these countries will be intimately linked to women's rights, since it is unlikely that children will have economic security and greater liberty if their mothers lack these rights. It will also require wealthier countries to demonstrate concern for children beyond national borders.

Children's rights advocates sought to make the 1900s, the century of the child. They were successful in many respects. Future advances will require finding adequate balances for the tensions that have been identified, as well as eliminating the negative impacts on children of discrimination based on race, class, and gender. For most children, these latter factors may be greater barriers to their rights and well-being than age.

See also: Children and Law; Children, Value of; Human Rights and Social Work; Human Rights, Anthropology of; Human Rights, History of; Street Children: Cultural Concerns; War, Political Violence, and Effects on Children.

Bibliography

Alston, P., 1994. The best interests principle: towards a reconciliation of culture and human rights. In: Alston, P. (Ed.), The Best Interests of the Child. Clarendon Press, Oxford, UK.

Armstrong, M., Chuulu, M.S., Himonga, C., Letuka, P., Mokobi, K., Ncube, W., Nhlapo, T., Rwezania, B., Vilakazi, P., 1995. Towards a cultural understanding of the interplay between children's and women's rights: an Eastern and Southern African perspective. International Journal of Children's Rights 3, 333–368.

Campbell, T., 1992. The rights of the minor: as person, as child, as juvenile, as future adult. In: Alston, P., Parker, S., Seymour, J. (Eds.), Children, Rights and the Law. Clarendon Press, Oxford, UK.

Convention of the Rights of the Child, 1989. G.A. Res 44/25 U.N., GAOR, 44th sess., Annex, Supp. No 49 at 167 U.N. Doc A/44/49.

Coons, J., Mnookin, R., Sugarman, S., 1991. Puzzling over children's rights. Brigham Young University Law Review, 307–350.

Eekelaar, J., 1986. The emergence of children's rights. Oxford Journal of Legal Studies 6, 161–182.

Farson, R., 1974. Birthrights. Macmillan, New York.

Feinberg, J., 1980. The children's right to an open future. In: Aiken, W., LaFolette, H. (Eds.), Whose Child? Children's Rights, Parental Authority, and State Power. Rowman and Littlefield, Tottowa, NJ.

Fishkin, J.S., 1983. Justice, Equal Opportunity, and the Family. Yale University Press, New Haven, CT.

Freeman, M., 1992. Taking children's rights more seriously. In: Alston, P., Parker, S., Seymour, J. (Eds.), Children, Rights and the Law. Clarendon Press, Oxford, UK.

Fried, C., 1978. Right and Wrong. Harvard University Press, Cambridge, MA.

Goldstein, J., Freud, A., Solnit, A.J., 1973. Beyond the Best Interests of the Child. Free Press, New York.

Hawes, J., 1991. The Children's Rights Movement: A History of Advocacy and Protection. Twayne Publishers, Boston, MA.

International Journal of Children's Rights, 1995. Special Issue: Multiculturalism and The Rights of the Child 3, 1–144.

Invernizzi, A., Williams, J. (Eds.), 2011. The Human Rights of Children: From Visions to Implementation. Ashgate Publishing Company, Burlington, VT.

Lonne, B., Parton, N., Thompson, J., Harries, M., 2009. Reforming Child Protection. Routledge, London.
Luttrell-Rowland, M., 2012. Ambivalence, conflation, and invisibility: a feminist analysis of state enactment of children's rights in Peru. Signs: Journal of Women in Culture and Society 38 (1), 179–202.
Melton, G., Limber, S., 1992. What children's rights mean to children: children's own views. In: Freeman, M., Vreeman, R. (Eds.), The Ideologies of Children's Rights. Kluwer Academic Publishers, Dordrecht, The Netherlands.
Meyer v. Nebraska, 1923. United States Supreme Court Reports. 262, 390–403.
Nozick, R., 1974. Anarchy, State and Utopia. Basil Blackwell, Oxford, UK.
O'Neill, O., 1988. Children's rights and children's lives. Ethics 98, 445–463.
Purdy, L., 1992. In Their Best Interest? Cornell University Press, Ithaca, NY.
Schlossman, S.L., 1977. Love and the American Delinquent. University of Chicago Press, Chicago.
Shonkoff, J.P., Garner, A.S., The Committee on Psychosocial Aspects of Child and Family Health, Committee on Early Childhood, Adoption, and Dependent Care, Section on Developmental and Behavioral Pediatrics, 2011. The lifelong effects of early childhood adversity and toxic stress. Pediatrics 129 (1), e232–e246.
Tinker v. Des Moines independent community school district, 1969. United States Supreme Court Reports 393, 503–526.
UN Convention on the Rights of the Child, 1990. United Nations. http://www.ohchr.org/en/professionalinterest/pages/crc.aspx.
United States Advisory Board on Child Abuse and Neglect, 1991. Creating Caring Communities. US Department of Health and Human Services, Washington, DC.
Wald, M., 1982. State intervention on behalf of endangered children. International Journal of Child Abuse and Neglect 6, 3–45.
Woodhouse, B., 1992. 'Who owns the child?': Meyer and Pierce and the child as property. William and Mary Law Review 33, 995–1122.
Zimring, F., 1982. The Changing Legal World of Adolescence. Free Press, New York/London.

Children's Geographies

Claire Freeman, Geography Department, University of Otago, Dunedin, Aotearoa, New Zealand
Paul Tranter, Geography School of Physical, Environmental and Mathematical Sciences, UNSW Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Children's geographies, or more correctly the geographies of children and young people, explore the relationship between children, young people, and space. Space matters as it impacts directly on the ways children interact with society and on their physical and mental well-being. It is also an important determinant of their life chances. Children's geographies is increasingly recognized as a significant focus of geographical research, especially in relation to the lives of children in the majority world. Children's lives are being directly affected by the changing global environment. Children's geographies has much to offer in understanding and responding to these changes.

What Is Children's Geographies?

Geography involves the study of space and place, its formation, character, and, significantly, the interaction between place, society, and individuals within that society. For children, space is critical; it determines their well-being and life chances, it molds their identity, influences behavior, and provides the stage on which they find their way in society. Children are shaped by and in turn shape the physical worlds in which they live, and in this regard space matters enormously. Geography with its disciplinary focus on space and society is ideally placed to contribute to better understanding children's lives. Children's geographies, therefore, is the study of the relationship between children and space. However, the reality is far more complex than this simple relational definition suggests.

Childhood itself is a highly contested notion. Definitions and concepts of childhood differ. Childhood can be defined chronologically but clarity over what constitutes the age of childhood is contested. The United Nations Convention on the Rights of the Child (UNCROC) defines a child as being up to age 18; however, few young people appreciate the appellation of child. Many organizations concerned with children's well-being use the phrase children and young people. Yet, the term young person is itself problematic as in some contexts it refers to someone up to age 25. Childhood can also be defined with reference to access to rights. Again these are age-variable by country and context. For example, the age of criminal responsibility in India is age 7.

Children's geographies generally follows the UNCROC definition of a child being aged under 18 but uses the term young people for children of secondary school age or beyond. The term children's geographies is in some ways a misnomer, as its focus is not just on children but on young people, so really should be called geographies of children and young people. Neither is 'children's geographies' purely the prerogative of geographers, rather it is a coalescence of research around children's geographies undertaken by researchers from a range of disciplines. It is characteristically and increasingly multidisciplinary in both participation and content, and has application to a range of professions including education, psychology, sociology, anthropology, urban design, architecture, planning, transportation, and health studies.

Development of Children's Geographies as a Discipline

Children were conspicuously absent from geography for most of the discipline's early development. The 1970s and early 1980s saw a number of early studies on children from a geographic perspective. Bunge is usually credited with providing the first focused geographical study of children for his work on spatial oppression and children in Detroit and Toronto (Bunge and Bordessa, 1975). Other significant early works include those by Colin Ward (1978), Roger Hart (1979), and Roger Moore (1985). The works of Hart and Moore broke new ground in studying children's use of space as they took their lead from the children in their research. Child-led research is now seen as good practice, but at that time it was groundbreaking in its novelty. While these were significant works, children's geographies as a discipline usually takes its starting point from the article by James (1990) which asked "Is there a 'place' for children in geography?" The answer to James' question is a clear 'yes,' as children's geographies has grown markedly since then and is a clearly recognized subdiscipline in geography that has its own journal, holds international conferences, and is supported by a wealth of scholarly work. As well as research being published in the journal *Children's Geographies*, children-focused studies are now being published in a wide range of geographical journals. However, despite this growth, within geography, children's geographies, like children themselves, is seen as a worthy endeavor but has a limited voice and is generally consigned to the margins of mainstream geographic thought. Children's geographies has received considerably less attention from geographers than the geographies of adults. This is unfortunate, particularly in terms of the potential for children's geographies to offer insights into wider geographical research and current areas of concern, notably research on creating more resilient environments (Skelton and Gough, 2013).

A Schema for Children's Geographies

The field of children's geographies, like the subdiscipline itself, has broadened and matured. In 1999, Matthews and Limb proposed an agenda for the geography of children, putting

forward seven propositions focusing on children's geographic relationships. Their agenda identified that children differ from adults in their 'ways of seeing,' place use, free range, environmental fears and sense of danger, place feelings, relationship to environmental decision making, and democratic responsibility. This agenda has stood children's geographies in good stead, but in recent years increasing attention is being paid to wider concerns, notably developing and challenging the ways that children are conceptualized and positioned in society.

Children's geographies has been instrumental in challenging children's absence from and marginalization in society in respect to their knowledge and voice, and in repositioning children as a focus of study in their own right, rather than as adjuncts to adult-based studies. This repositioning is important as it has given recognition to children as social actors who are knowledgeable, competent, and capable of coherently articulating their views. Thus, children are seen as dynamic agents capable of not just being acted on by the environment and society, but also through their own actions, themselves shaping the environment and society. They are, therefore, not just adults in waiting but capable social actors in their own right. This maturing of how children are represented and the emerging agenda is most apparent in the contemporary development of what has been termed 'new social studies of childhood'; its foundational principles are:

1. to explore the social construction of childhood and youth;
2. to critique the assumption that childhood is universally experienced and understood;
3. to foreground children and young people's own issues, experiences, practices, and perspectives; and
4. to recognize and explore children's and young people's agency in their lives and in broader social structures (Kraftl et al., 2012).

While these foundational principles are well intentioned and have the potential to promote both children's well-being and broader social agendas, the reality is that challenges persist in promoting change in policy and practice. Despite the upsurge of interest in the lives and well-being of children, recognition of children's social and environmental rights remains limited (Freeman, 2007). Enabling children and young people to take up a full role in society requires adults to take a more child-focused and supportive role and to devolve to children and young people some of the power they currently hold. Whether societies and institutions are ready for this remains uncertain as the values that at one time suggested that 'children should be seen but not heard' are still evident (Freeman and Tranter, 2011).

A Global Discipline

Understanding of and engaging with children's geographies is increasingly important for a number of reasons. Children are part of an increasingly spatially extensive and complex world. They reflect highly varied social and geographical relationships that occur at a diversity of scales from local home neighborhood to the global. Children are increasingly mobile as they move within and between homes, schools, and various service spaces. This mobility can comprise microscale movements as in daily or weekly movements between parental homes, or macroscale movements as they migrate between countries and cross international borders with their families or as independent travelers, child refugees, or workers. Children's geographies includes studies from across the globe but the Minority/Western world remains dominant in research, focus, and academic participation. This is reflected, for example, in the adult–child dichotomy which has been central to childhood studies but which is less central in situations where children have extended families, play a central role in the family economy, or take on parental responsibility due to parental HIV/AIDS, situations more commonly found in the global south (Punch and Tisdall, 2012). The Eurocentric dominant views of children, family, and space with their western notions of childhood innocence and dependency are being and need to be increasingly challenged if children's geographies is to be truly 'global' in orientation. In reality, as Skelton (2009: 1434) points out, "most of the world's children live in the economically poor countries of the Majority World." Although existing studies are dominated by western researchers, many of these researchers themselves are calling for a wider global remit for children's geographies.

Children's Rights

The UNCROC requires governments to respect and uphold children's rights in a range of areas from basic rights such as the right to a name, identity, life, and religion to more complex rights including the right to have a say in decisions that impact on their lives (article 12). In children's geographies, articles 12 and 13 (the right to access information) are especially important, as they emphasize the child's rights in matters that affect their lives, including matters relating to housing, environment, transport, leisure, access to schools, sports grounds, and health facilities. The progress of many countries toward fully implementing UNCROC has been slow, as illustrated by the fact that in 2008 one in three children in the UK lived in poverty and "disabled children continue to face barriers to the enjoyment of their rights, in particular access to appropriate health services, leisure and play" (UK Children's Commissioners, 2011). Recognition of children's rights across the world remains variable especially in countries experiencing political instability, war, and economic and environmental crises. Globally the position of children remains poor (Smith and Ansell, 2009). It is imperative that geographers recognize and prioritize giving voice to children's perspectives and views in their work.

The 'Universal' Child

There has been a tendency in traditional childhood studies and to some extent in children's geographies to talk of the child in a sense that implies a universal childhood, where identifying a 'child' as distinct to adult is seen as sufficient. More recently though recognition is being given to the fact that children lead highly diverse lives in often quite different social, political, geographic, and cultural contexts, where norms around behavior, age-appropriate activities, expectations, and so on can vary enormously (Hörschelmann and van Blerk, 2012). For

instance, in countries especially in the global south, children contribute to the family economy at a young age whereas in parts of the developed world young people commonly seek employment to fund their lifestyle choices. In other places it is seen as inappropriate for children to work. These types of differences can occur in different geographic locations, such as different countries, but can also be apparent in geographically similar but socially or economically different portions of society as in the same town. Within geographically close communities, children's lives can be markedly different depending on factors such as gender, family size, neighborhood, race, school attended, parental occupation or lack of, ability or disability, and so on. Some geographers have also investigated how differences in children's lives in terms of work/play balance are affected by global forces, and how the notion of a childhood based on play and school is not the norm for many of the world's children (Aitken et al., 2008). Growing international trade inequities and expansion of development in developing countries have led to increased pressure to engage in employment for children in many poor countries (Abebe, 2007).

Children reflect the diversity of the wider population in their experiences. To talk of children as a single entity masks the variability and indeed richness of their lives. Children can also act in multiple roles in different parts of and at different times in their lives. They may move between different cultural worlds that may have different languages. As they age, their physical abilities develop which enhances their ability to access places and to take up different social, economic, and cultural roles. Children's geographies needs to look beyond generic social and geographic boundaries, to the particular circumstances contributing to the child's position and identity. Disaggregation of children is particularly important if policy development is to recognize and accommodate children's diverse positions in society and their varied well-being. Recognition of difference is especially important in research.

Research

If research is to reflect accurately the reality of children's lives and perspectives, then the research must provide ways for the child's views and perspectives to emerge. There are a number of challenges that researchers face in this regard. As identified earlier, children are a highly heterogeneous and dynamic group whose lives are continually evolving as they age and develop. To enable children's knowledge to emerge, different research methods will be required to those used with adults, and special skills and techniques may be required in working with children. Some innovative work has been done in this regard using methods such as researcher observation, mental mapping, computer software programs such as softGIS, photo voice, walk along/go-along methods, child-focused interviews, and activities such as box city creation. The need for child- and age-appropriate methods is especially pertinent for working with very young children who are sorely underrepresented in research (see Hancock and Gillen, 2007 for an example of geographical research with preschool children).

Ethical considerations are important and more challenging than most adult-oriented research and relate to the need to ensure all children's best interests are fully protected. Ethical issues can and do arise during the research process as in disclosures by the child of matters of a harmful nature (e.g., neglect, exploitation) occurring in the child's own life. Such disclosures can cause the researcher to have to weigh up confidentiality against the need to act in a child's best interest. The researcher's own position, beliefs, and experiences also impact on the research process in which the researcher needs to be sensitive to the power differential present between adult and child and researcher and the child participant. Although research with children can be challenging, it is enlightening, fascinating, and often great fun for all involved. It is also potentially valuable for providing insights for researchers concerned with responses to major global challenges. The more active and engaging and fun the research process is, the better the data derived is likely to be.

Dominant Research Themes

An appreciation of the substance of children's geographies can be gleaned by looking at the broad range of themes it explores, including applications using different lenses (from positivist to postmodern and poststructuralist). The following list, by no means a complete one, indicates the breadth of coverage of research within children's geographies, including how children engage with the various spaces in their lives; health and hazards (traffic, pollution, inactivity, bullying); designing environments for children (including the concept of children as planners); children as citizens and social agents; children's cognitions/imaginings of their environment; exploitation of children, including child trafficking and slavery; the role of migration in children's lives; children and information technologies; and a lively and wide debate around the ethical issues involved in researching children. Many of these themes strongly interrelate with each other.

Rather than attempting to cover all the themes examined in children's geographies research, the focus here is on two broad themes that have attracted the attention of researchers within the subdiscipline:

1. studies of how children engage with the major spaces that they use in their daily lives (e.g., designated play space, schools, and the city center), particularly the way in which these spaces provide opportunities for children's play; and
2. research on the theme of children's mobilities (including active travel and independent mobility).

Children's Spatial Engagements

An important focus of children's geographies research involves children's perceptions of and engagement with key spaces in their daily lives. Children are portrayed as having an intimate connection with their local environments in ways that are qualitatively different from adult's connections. Largely through playful exploration, they have the ability to live in the present and be fascinated by their immediate surroundings, rather than being focused on a destination, and seeing the next task as more important than the present, as adults often do. Key spaces for children that have been

studied in children's geographies include playgrounds (Karsten, 2003), their own homes (Wood and Beck, 1994), streets (Fotel, 2009; Tranter and Doyle, 1996), school grounds (Dyment et al., 2009; Thomson, 2005), neighborhoods (Mee, 2010; Freeman, 2010), city centers (Woolley et al., 1999), and various service spaces other than playgrounds (ranging from community gardens, museums, hospitals to public markets) (Schmelzkopf, 1995; Watson, 2009).

The focus on these spaces occupied by children is partly a consequence of the progressive confinement of children to child-specific spaces, a trend beginning more than a century ago. The dominant view of children's place in western societies is centered on the home, family, private space and in the public domain to child-designated public spaces, predominantly playgrounds and skate parks. Children's access to public space has diminished, not least due to the dominance of a neoliberal policy agenda which has led to the winding back of public investment in cities, particularly in the UK, the US, Australia, and New Zealand. The erosion of investment in public space for children, as well as the commercialization of play spaces (McKendrick et al., 2000), has reinforced the view that children are illegitimate occupiers, or even troublemakers when occupying some 'adult' public spaces (Valentine, 2004).

A growing body of empirical studies has identified the need to consider the views of children themselves, and how they relate to their worlds in varied geographical settings. Children may be influenced by space, but they are not passive reactors; children can and do shape and influence both the social and physical nature of space. They create their own spaces in neighborhoods, school grounds, and public spaces in the city. They are in a state of continual negotiation with adults over space: where they can go, how they get there, whom they go with, and what they can do in those spaces. Children need to be mobile and to access space independently to develop. They have a strong need to explore, to encounter, to grow, and to engage with the world. They will attempt to pursue these needs passionately, vigorously, and autonomously, using their knowledge, understanding, and their social relationships to seek to promote their right to access the spaces that they need.

Playful Spaces

Play is often seen as the central business of childhood, and is the way that children learn about their world. Opportunities for play have been marginalized by a range of issues in society, from parental concerns about fast-tracking children's learning, to the loss of spaces for play and the exclusion of children from spaces regarded as 'adult' spaces. The contested notion of the play itself has attracted attention in children's geographies, warning against taking play for granted, and arguing that geographical studies are important in helping to define play as much as identifying the way in which children's play can contribute to the meanings of place (Harker, 2005).

Perhaps unsurprisingly, a great deal of children's geographies research has focused on playgrounds, as these are seen as a service space specifically designed for children. Geographic research has examined ways in which playgrounds are used by children, including children of different abilities, ages, and gender (Karsten, 2003; Moore et al., 1997). While this research is important, other geographic research has called into question the concept of special spaces for children, identifying the possibility that restricting children's play to playgrounds and school grounds may lead to 'childhood ghettoization' (Matthews, 1995: 223).

School

A research focus on schools (and school grounds) as a children's space is justified in terms of the centrality of schools to children's everyday life in many societies, their ubiquitous nature, the length of time spent in the school, and the way in which the various spaces and practices at the school, including a hidden curriculum, are critical in children's experience of both the inclusionary and exclusionary aspects of society (Collins and Coleman, 2008). Schools are not only spaces where children learn about an educational curriculum, but also spaces where they learn about societal structures, as reflected in the different quality of environments in rich versus poor neighborhoods. School grounds are particularly important for geographical research, as they allow the observation and analysis of key aspects of children's lives in a contained environment (Hemming, 2007). Key topics within school ground research include environmental learning in school grounds (Tranter and Malone, 2004), green school grounds (Dyment and Bell, 2007), physical activity, emotional geographies of pleasure, enjoyment, and playfulness (Bundy et al., 2009; Hemming, 2007), the loss of play opportunities as a consequence of a culture of risk aversion (Thomson, 2007), and the demonstration of social agency by children (Hemming, 2007).

City Center

Despite the city center being a potentially valuable site for children to engage with other people (including children) in public space, there has until recently been surprisingly little attention by geographers to this topic. This may reflect an underlying perception that children are out of place in the city center. Indeed, an Australian geographer, Fincher (2004: 337) became concerned by developers' attitudes to children in Melbourne: "… presentation of children … as burdensome – people from whom to 'escape' and 'be freed' … the construction of a myth that the perfect lifestyle is to be lived without children or other dependants – one's own or other people's." This argument has also been used to promote child-free gated communities, in some condominiums in the US, and adult tourism developments, and is implicit in elderly people's residential complexes even though children frequently visit relatives there. With the growing movement of families back into central city areas, interest in the lives of city children and the opportunities and constraints this environment offers has been the focus of renewed research in cities such as Auckland, New Zealand (Ergler et al., 2012) and Melbourne, Australia (Mizrachi and Whitzman, 2009). The reality of children's presence in the city centers confronts the widespread view that children's lives are focused on those spaces specially designated for children. While the city and the other spaces that provide important constituents of children's daily lives are important, so too is the ability of children to access these spaces, and hence mobility is a key theme in children's geographies.

Children's Mobilities

The term 'children's mobilities' has emerged only recently in children's geographies. It represents a combination of children's geographies with what has become known as the mobility 'turn' in the social sciences that emerged in the 1990s. The mobilities paradigm applies new theoretical approaches to understand the way that mobilities are "at the center of constellations of power, the creation of identities and the micro-geographies of everyday life" (Cresswell, 2011: 551). Children's geographies is now finding its way into journals about mobility, including the journal *Mobilities* (Barker et al., 2009), and special issues of the *Journal of Transport Geography* (Buliung et al., 2011). However, children's mobility is still developing its status in the transport/mobility field in geography. As Barker et al. (2009) point out only one paper in the first 3 years of the journal *Mobilities* focused specifically on children and young people. This relative neglect of children echoes the fact that children are a "user group that is poorly served by transport systems" in many cities (McMillan, 2005: 441), particularly in the US, Australia, Britain, and Canada.

Geographers were studying children's mobilities long before the 'new mobilities' paradigm (Sheller and Urry, 2006) emerged in the social sciences, though not with the same theoretical underpinnings. Colin Ward's (1978) *Child in the City* raised important questions about how children move around the city, and the stimulating experiences provided for children who were able to explore their city. A landmark quantitative study was the study by Hillman, Adams, and Whitelegg – *One false move: a study of children's independent mobility* (Hillman et al., 1990). This study of children's freedom of movement in cities was based on the various 'licences' parents give to their children, e.g., to cross roads alone, come home from school alone. The main findings of the research were that children's freedoms in England in 1990 were much lower than in 1971 (using data from the same schools in both years) and children in England in 1990 were much more restricted than children in similar cities in Germany. This pioneering study was replicated in cities in several countries, indicating considerable variability between nations, within cities, and between rural and urban areas (Collins and Kearns, 2001; Fotel and Thomsen, 2004; Tranter and Pawson, 2001). New data from a 2010 follow-up study of the Hillman et al. (1990) study showed that in 2010, German primary school children were still granted all the licences in greater proportions than their English counterparts, despite car travel to school in Germany rising to English levels (Shaw et al., 2013). As well as examining 'licences,' several different methodologies to study children's independent mobility have been employed by geographers, including qualitative research (e.g., using photography, guided walks) and quantitative approaches (e.g., travel diaries and spatial range analysis through the application of geographic information systems) (Tillberg Mattsson, 2002).

Geographers are also involved in research on resistance to the trend toward reduced independent mobility in western cities, examining the loss of children's freedoms in the context of changing expectations by parents, the growing dominance of private transport, and the possibilities for new arrangements (e.g., walking school buses) to regain some of the freedom for children (Kearns et al., 2003). Mobilities research has also incorporated the topic of play, focusing on both 'affect' and the concept of playful mobility (Tranter and Sharpe, 2012), suggesting that if children's mobilities involve playful activity, children will enjoy their mobility and adults will spend less time facilitating and supervising their children's physical activity.

As well as independent mobility, a related concept that has attracted the attention of geographers is that of children's 'active travel' – travel that involves some physical activity (walking, cycling, scootering, and public transport). While children's independent mobility and children's active travel are sometimes conflated, they are two separate concepts, and recent research has suggested that they should be viewed as distinct. Indeed, some of the health benefits believed to be associated with independent mobility can also be achieved with active travel supervised by adults (Mammen et al., 2012). The health and developmental benefits of active travel and independent mobility are now regarded as potentially more beneficial for children's health and development than organized sport and supervised exercise classes. Even the journey to school, if it can be made independently by children, allows them to develop a deeper understanding of the geography of their own neighborhoods and communities, as well as opportunities for free-play or day-dreaming – important activities for children (Ross, 2007).

New Directions for Children's Geographies

The search for new directions for children's geographies might well be guided by consideration of the changes in children's lives over recent decades, particularly for children in developed nations. These include a growing ethos of individualism; closure of local shops and schools; loss of wild spaces (natural spaces); increased population densities; increased pressure on children to compete in a consumerist world; replacement of children's independent mobility with adult-dependent mobility; increased use of technology for surveillance of children; more play indoors rather than outdoors; increased fear or risk aversion; and increasingly scheduled lives and an increased pace of life and time pressure. Many of these changes have impacts on adults as well as children, yet children have little choice but to accept and adapt to the changes. Many of the changes have led to a loss of opportunity for children to playfully explore their own neighborhoods and cities, and a concomitant increase in the pressures on parents to drive, accompany, or supervise their children's activities.

Importantly, many of these changes are driven by global forces. Children's lives are influenced not only by the global financial crisis, but also by other global trends and challenges. A growing global human population places increasing pressure on cities. A probable crisis in access to energy, particularly cheap oil (Brecha, 2013), is likely to have significant (if unpredictable) impacts on children's lives (Tranter and Sharpe, 2007). As Brecha (2013) argues, "being societally unprepared for declining oil production will have serious consequences." Climate change impacts on children are rarely forecast in a positive light (Bartlett, 2011). Many of these challenges are interdependent. For example, if a global response to a shortage of oil is to switch to liquid fuel from coal, this will very likely accelerate global climate change.

With few exceptions, children's geographies has been largely silent regarding these major global challenges and their potential impact on children. Here lies a valuable opportunity for new directions in the subdiscipline. Children's geographies needs to look not just at specific children in specific places (the microspaces discussed earlier), but at broader, global perspectives, considering the macroscale issues as they are influenced by "broad-brush political-economic and social-cultural transformations" Philo (2000: 253). Ansell (2009) reinforces this, arguing that "it is not *only* locally occurring processes, practices and events that impinge on children's lives. Economic globalization and neoliberal policies are transforming life for many young people, but they remain untouched by the small-scale participatory projects that are the focus of much research" (Ansell, 2009: 191). Geographers need to research and challenge the national and even global forces and processes that affect children. These include a neoliberal agenda in planning, the striving for economic growth that is clearly unsustainable, and a disregard for the likely impacts of energy stress and global warming.

If children's geographies are to be relevant, they need to be considered in spheres beyond those limited areas in which children's views are currently seen to be important and accepted (e.g., in the design of parks and playgrounds and leisure centers for children and young people). Virtually every aspect of city management and planning impacts directly or indirectly on children, whether it is housing, transport, education, health, or energy generation. Researchers within all these areas should be encouraged to consider children's interests. Article 3 of the UNCROC states: "The best interests of children must be the primary concern in making decisions that may affect them." Every decision made about any aspect of our economy, environment, and society ultimately affects children. An urgent need for a new direction in children's geographies is for a link to be provided between research on children's use of space and place and the *sustainability agenda* (Percy-Smith and Burns, ahead-of-print). Not only are children shaped by and in turn shape their worlds, but also their worlds are influenced by factors that operate from local to global scales. Future children's geographies may demonstrate that the views of children are vital in our responses to global challenges.

Acknowledgments

We would like to acknowledge the excellent advice on this article provided by Dr Christina Ergler.

See also: Mobility and Travel Activity Patterns; Place in Geography; Social Geography; Youth, Geography of.

Bibliography

Abebe, T., 2007. Changing livelihoods, changing childhoods: patterns of children's work in rural southern Ethiopia. Children's Geographies 5 (1–2), 78–93.

Aitken, S., Lund, R., Kjørholt, A.T. (Eds.), 2008. Global Childhoods: Globalization, Development and Young People. Routledge, London.

Ansell, N., 2009. Childhood and the politics of scale: descaling children's geographies? Progress in Human Geography 33 (2), 190–209.

Barker, J., Kraftl, P., Horton, J., Tucker, F., 2009. The road less traveled–new directions in children's and young people's mobility. Mobilities 4 (1), 1–10.

Bartlett, S., 2011. Children and the culture of climate change. The Journal of the History of Childhood and Youth 4 (3), 497–505.

Brecha, R.J., 2013. Ten reasons to take peak oil seriously. Sustainability 5 (2), 664–694.

Buliung, R., Sultana, S., Faulkner, G., 2011. Guest editorial: special section on child and youth mobility – current research and nascent themes. Journal of Transport Geography 20 (1), 31–33.

Bundy, A., Luckett, T., Tranter, P., Naughton, G., Wyver, S., Ragen, J., Spies, G., 2009. The risk is that there is 'no risk': a simple, innovative intervention to increase children's activity levels. International Journal of Early Years Education 17 (1), 33–45.

Bunge, W., Bordessa, R., 1975. The Canadian Alternative – Survival. Expeditions and Urban Change. York University, Toronto.

Collins, D., Coleman, T., 2008. Social geographies of education: looking within, and beyond, school boundaries. Geography Compass 2 (1), 281–299.

Collins, D., Kearns, R., 2001. The safe journeys of an enterprising school: negotiating landscapes of opportunity and risk. Health and Place 7 (4), 293–306.

Cresswell, T., 2011. Mobilities I: catching up. Progress in Human Geography 35 (4), 550–558.

Dyment, J., Bell, A.C., 2007. Active by design: promoting physical activity through school ground greening. Children's Geographies 5 (4), 463–477.

Dyment, J., Bell, A., Lucas, A., 2009. The relationship between school ground design and intensity of physical activity. Children's Geographies 7 (3), 261–276.

Ergler, C.R., Kearns, R., Witten, K., 2012. Seasonal and locational variations in children's play: implications for wellbeing. Social Science & Medicine 9, 178–185. pii: S0277-9536(12)00796-4.

Fincher, R., 2004. Gender and life course in the narratives of Melbourne's high rise developers. Australian Geographical Studies 42 (3), 325–338.

Fotel, T., 2009. Marginalized or empowered? Street reclaiming strategies and the situated politics of children's mobilities. Geography Compass 3 (3), 1267–1280.

Fotel, T., Thomsen, U., 2004. The surveillance of children's mobility. Surveillance and Society 1 (4), 535–554.

Freeman, C., 2010. Children's neighbourhoods, social centres to 'terra incognita.' Children's Geographies 8 (2), 157–176.

Freeman, C., Tranter, P., 2011. Children and Their Urban Environment. Earthscan, London.

Freeman, M., 2007. Why it remains important to take children's rights seriously. International Journal of Children's Rights 15 (1), 5–23.

Hancock, R., Gillen, J., 2007. Safe places in domestic spaces: two-year-olds at play in their homes. Children's Geographies 5 (4), 337–351.

Harker, C., 2005. Playing and affective time-spaces. Children's Geographies 3 (1), 47–62.

Hart, R., 1979. Children's Experience of Place: A Developmental Study. Irvington Publishers, New York.

Hemming, P.J., 2007. Renegotiating the primary school: children's emotional geographies of sport, exercise and active play. Children's Geographies 5 (4), 353–371.

Hillman, M., Adams, J., Whitelegg, J., 1990. One False Move: A Study of Children's Independent Mobility. Policy Studies Institute, London.

Hörschelmann, K., van Blerk, L., 2012. Children, Youth and the City. Routledge, Oxon.

James, S., 1990. Is there a 'place' for children in geography? Area 22 (3), 278–283.

Karsten, L., 2003. Children's use of public space: the gendered world of the playground. Childhood 10 (4), 457–473.

Kearns, R.A., Collins, D.C.A., Neuwelt, P.M., 2003. The walking school bus: extending children's geographies? Area 35 (3), 285–292.

Kraftl, P., Horton, J., Tucker, F., 2012. Critical Geographies of Childhood and Youth: Contemporary Policy and Practice. Policy Press, Bristol.

Mammen, G., Faulkner, G., Buliung, R., Lay, J., 2012. Understanding the drive to escort: a cross-sectional analysis examining parental attitudes towards children's school travel and independent mobility. BMC Public Health 12 (1), 862.

Matthews, H., 1995. Living on the edge: children as outsiders. Tijdschrift voor Economische en Sociale Geografie (Journal of Economic and Social Geography) 86 (5), 456–466.

Matthews, H., Limb, M., 1999. Defining an agenda for the geography of children: review and prospect. Progress in Human Geography 23 (1), 61–90.

McKendrick, J.H., Bradford, M.G., Fielder, A.V., 2000. Kid customer? Commercialization of playspace and the commodification of childhood. Childhood 7 (3), 295–314.

McMillan, T.E., 2005. Urban form and a child's trip to school: the current literature and a framework for future research. Journal of Planning Literature 19 (4), 440–456.

Mee, K.J., 2010. 'Any place to raise children is a good place': children, housing and neighbourhoods in inner Newcastle, Australia. Children's Geographies 8 (20), 193–211.

Mizrachi, D., Whitzman, C., 2009. Vertical living kids: creating supportive environments for children in Melbourne central city high rises. 4th National Conference on the State of Australian Cities, Perth, November 24–27, 2009.

Moore, R., 1985. Childhood's Domain: Play and Place in Child Development. Croom Helm Ltd, Kent.

Moore, R., Goltsman, S., Iacofano, D., 1997. Play for All Guidelines: Planning, Design and Management of Outdoor Play Settings for All Children. MIG Communications, Berkeley, California.

Percy-Smith, B., Burns, D., 2013. Exploring the role of children and young people as agents of change in sustainable community development. Local Environment 18 (3), 323–339. ISSN 1354-9839.

Philo, C., 2000. "The corner-stones of my world": editorial introduction to special issue on spaces of childhood. Childhood: A Global Journal of Child Research 7 (3), 243–256.

Punch, S., Tisdall, E.K.M., 2012. Exploring children and young people's relationships across majority and minority worlds. Children's Geographies 10 (3), 241–248.

Ross, N.J., 2007. 'My journey to school…': foregrounding the meaning of school journeys and children's engagements and interactions in their everyday localities. Children's Geographies 5 (4), 373–391.

Schmelzkopf, K., 1995. Urban community gardens as contested space. Geographical Review 85 (3), 364–381.

Shaw, B., Watson, B., Frauendienst, B., Redecker, A., Jones, T., Hillman, M., 2013. Children's Independent Mobility: A Comparative Study in England and Germany (1971–2010). Policy Studies Institute, London.

Sheller, M., Urry, J., 2006. The new mobilities paradigm. Environment and Planning A 38 (2), 207–226.

Skelton, T., Gough, K.V., February 2013. Introduction: young people's im/mobile urban geographies: special issue: young people's im/mobile urban geographies. Urban Studies 50, 467–483.

Skelton, T., 2009. Children's geographies/geographies of children: play, work, mobilities and migration. Geography Compass 3 (4), 1430–1448.

Smith, F., Ansell, N., 2009. Children/Childhood, International Encyclopedia of Human Geography. Elsevier-online, pp. 58–64.

Thomson, S., 2005. Territorialising the primary school playground: deconstructing the geography of playtime. Children's Geographies 3 (1), 63–78.

Thomson, S., 2007. Do's and don'ts: children's experiences of the primary school playground. Environmental Education Research 13 (4), 487–500.

Tillberg Mattsson, K., 2002. Children's (in) dependent mobility and parents' chauffeuring in the town and the countryside. Tijdschrift voor economische en sociale geografie 93 (4), 443–453.

Tranter, P., Doyle, J., 1996. Reclaiming the residential street as playspace. International Play Journal 4, 81–97.

Tranter, P., Malone, K., 2004. Geographies of environmental learning: an exploration of children's use of school grounds. Children's Geographies 2 (1), 131–155.

Tranter, P., Pawson, E., 2001. Children's access to local environments: a case-study of Christchurch, New Zealand. Local Environment 6 (1), 27–48.

Tranter, P., Sharpe, S., 2007. Children and peak oil: an opportunity in crisis. International Journal of Children's Rights 15 (1), 181–197.

Tranter, P., Sharpe, S., 2012. Disney-Pixar to the rescue: harnessing positive affect for enhancing children's active mobility. Journal of Transport Geography 20 (1), 34–40.

UK Children's Commissioners, November 2011. UK Commissioners Midterm Report to the UK State Party on the UN Convention on the Rights of the Child. Office of the Children's Commissioner, London.

Valentine, G., 2004. Public Space and the Culture of Childhood. Ashgate, Surrey.

Ward, C., 1978. Child in the City. The Architectural Press, London.

Watson, S., 2009. The magic of the marketplace: sociality in a neglected public space. Urban Studies 46 (8), 1577.

Wood, D., Beck, R.J., 1994. Home Rules. Johns Hopkins University Press, Baltimore, MD.

Woolley, H., Dunn, J., Spencer, C., Short, T., Rowley, G., 1999. Children describe their experiences of the city centre: a qualitative study of the fears and concerns which may limit their full participation. Landscape Research 24 (3), 287–301.

Children, Value of

Elizabeth Thomson, Stockholm University, Stockholm, Sweden; and University of Wisconsin–Madison, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In demography, the value of children refers most often to the benefits parents receive from having and rearing children. Benefits may accrue from the children themselves, from the experience of rearing them, or from the responses of kin, community, and society at large. Children also entail costs for parents and the value of children sometimes refers to their net value (benefits less costs). Benefits and costs of children are shaped by the economic conditions of life, forms of social organization, and cultural beliefs and practices. Studies using surveys to measure the value of children report variation in values across societies; by socioeconomic status; between women and men; for first, second, and higher order births; and for daughters versus sons. Cultural foundations for the value of children have also been explored. The net value of children underlies parents' desires for children which, in combination with their ability to achieve those desires, influences their decisions to have children and how many children they have.

Theoretical Development

Theories of fertility decline focus primarily on the economic value of children. In agricultural economies and during early periods of industrialization, parents and kin make the heavy investment of time and money in young children in order to reap the rewards of children's labor from adolescence onward (Schultz, 1973; Caldwell, 1976). The old-age security value of children is particularly important in contexts where no public provisions exist for elder care (Cain, 1985). For large numbers of people throughout the world, the economic value of children in the intermediate and long term continues to be a primary benefit of parenthood (Nauck and Klaus, 2007). Even in wealthy societies where children's labor is no longer central to their parents' subsistence, adult children may remain important to the economic well-being of elderly parents (Rendall and Bahchieva, 1998; Buehler, 2008).

Scholars have long recognized that the value of children extends beyond their direct economic benefits. The term 'value of children' is most closely associated with Hoffman and Hoffman (1973), who identified nine psychological needs fulfilled by parenthood and the experience of child-rearing: (1) adult status and social identity; (2) expansion of the self; (3) morality; (4) primary group ties, affiliation; (5) stimulation, novelty, fun; (6) creativity, accomplishment, competence; (7) power, influence, effectance; and (8) social comparison, competition. Most of these psychological needs have an inherently social character, an issue that becomes clear in subsequent theory and research.

Several scholars have attempted to develop a unifying theory of the value of children. Friedman et al. (1994) claim that the primary value of children is the reduction of uncertainty. Those with limited access to other means for uncertainty reduction, such as stable careers and marriages, will want to become parents. Few of the hypotheses derived from this theory stand up to empirical scrutiny, however, and those that do can be explained by the economic and social opportunity costs of children (Lehrer et al., 1996).

Building on the work of cultural anthropologists and sociologists, Schoen et al. (1997) propose that children produce social capital, i.e., social relationships and social resources. At birth, children strengthen parents' ties to kin. Through schooling and other activities, they link parents with community resources. As adolescents and young adults, children bring new information, ideas, and social relationships to parental households. And most parents eventually obtain in-laws and grandchildren as a consequence of parenthood. Although the social capital argument was developed in the context where children have low or no economic value, it may have more general applicability. Where family, kin, and community ties provide economic opportunities and security, the social capital produced by children may be as economically valuable as their labor or accumulated wealth. Buehler (2008) makes a similar argument that parenthood is motivated by what he terms the 'structural value' of children, i.e., the expansion of personal networks and exchange relationships arising from the birth of a child. These networks and relationships produce in turn psychological rewards (Hoffman and Hoffman, 1973) and social capital (Schoen et al., 1997). But the extent to which the personal networks and exchange relationships are activated for the parents' well-being depends on the quality of the parent–child relationship.

Nauck (2007) derives the value of children from a general theory of social production functions in which individuals seek to maximize physical well-being and social esteem. Physical well-being requires comfort and stimulation; social esteem requires status, affect, and behavioral confirmation. The value of children is based on the extent to which children enable parents to achieve these five intermediate goals. The nature of the parent–child relationship means that stimulation, affect, and behavioral confirmation are confounded. The theoretically derived value of children is thus organized along lines identified by others – economic (comfort), social (status), and psychological (stimulation, affect, behavioral confirmation).

Children are also quite costly to their parents during the child-rearing years. Fertility declines in the twentieth century have been attributed in large part to declines in the economic value and increases in the economic costs of children to their parents (e.g., Caldwell, 1976). Most theorists distinguish

between the direct costs of children (food, clothing, shelter, education, etc.) and the opportunity costs of childcare for the parents' alternative activities (Fawcett, 1978). Parents also incur social and psychological costs from their responsibilities for and interactions with children (Hoffman and Hoffman, 1973; Fawcett, 1978; Schoen et al., 1997; Buehler, 2008).

Measuring the Value of Children

The Value of Children (VOC) surveys conducted in the mid-1970s (Fawcett, 1983) were based on the social psychological values developed by Hoffman and Hoffman (1973), but also contained information on the perceived economic benefits as well as the psychological, social, and economic costs of children. Respondents were asked open-ended questions about reasons for having or not having children and were asked to rate the importance of several lists of child benefits and costs as reasons to have or not have children or their next child. Most subsequent empirical research was based on closed-format questions. A partial replication of the VOC in the early 2000s (Nauck and Klaus, 2007) extended the importance scale from three to five points, improving the measures' reliability and validity. Seven-point importance ratings were used in the U.S. National Survey of Families and Households (Seccombe, 1991).

Subjective-expected-utility or expectancy-value models (e.g., Davidson and Jaccard, 1975) specify child benefits and costs as the product of the likelihood that having a child will produce a given outcome and the value (positive or negative) of the outcome. Each perceived likelihood and value is measured on a scale with several response options. More recent research in this tradition attempts to measure the combination of value and likelihood with a single response, e.g., agreement with a statement about positive or negative outcomes (Billari et al., 2009; but see Liefbroer, 2005).

Measurement analyses of child benefits and costs generally confirm the theoretical clusters of economic, social, and psychological well-being (Schoen et al., 1997; Nauck and Klaus, 2007), or at least separate the economic from the social and psychological (Mayer and Trommsdorff, 2010). Perceived costs constitute a separate dimension (Nauck and Klaus, 2007; Billari et al., 2009). Benefits and costs of children may also be distinguished in relation to general theories of fertility and fertility change (Liefbroer, 2005).

The real as opposed to perceived economic benefits and costs of children can also be measured indirectly based on the value of children's labor and transfers to parents over the life course, expenditures related to child-rearing, and the value of parental time that might otherwise be spent on income-producing activities (e.g., Rosenzweig, 1978). DiPrete et al. (2003) used a combination of time-budget data, attitude data, and estimates of the effects of a birth on labor earnings and standard of living changes to compare the economic costs – direct and opportunity – of children in different countries. In the aggregate, costs were strongly associated with national fertility rates. Zelizer (1985) argued that in a context where children's net economic value is zero, monetary compensation for a child's death represents children's net socioemotional value.

Variations in the Value of Children

The Value of First- and Later-Born Children

The original Value of Children studies found that first-, second-, and higher order births are associated with distinct benefits and costs (Bulatao, 1981; Fawcett, 1983). The first child confers the status of parenthood, so that benefits associated with parenthood per se can be acquired by having only one child. Adult status, relationship stability, parent–child interaction, and kin connections are all cited as primary reasons for becoming a parent. The first child is also associated with the greatest increase in opportunity costs, that is, constraints on parental time and energy.

The most important value of a second child is to provide a sibling for the first (Bulatao, 1981), a value that could be incorporated into the social dimension of child value discussed above. The value of higher order births may be predominantly economic – each child contributes additional labor or economic security for parents – but could also further strengthen partnership and kin ties and provide additional opportunities for rewarding parent–child interactions.

Griffith et al. (1985) pointed out that some of the unique values of first and second births may apply to the parity of a couple, independent of each partner's parity. She argued that the value of a child as a representation of the couple's commitment to one another would be relevant in stepfamilies, even when one or the other partner already had a child. The couple-specific value is argued to underlie the larger combined family sizes of step-family couples in comparison to couples in stable partnerships (Thomson, 2004).

Each child also adds to the parents' child-rearing costs, including direct financial expenditures and limitations on alternative uses of time. It appears, however, that cost considerations remain relatively stable across parities due to a trade off between quantity and quality (Bulatao, 1981; Nauck, 2007). That is, parents with small numbers of children may invest more time and money in each child than parents with larger numbers of children.

Gender and the Value of Children

Overall, men and women perceive the values and costs associated with children in much the same way. The few differences that are observed are consistent with traditional gender roles (Fawcett, 1983; Seccombe, 1991; Buehler, 2008). Liefbroer (2005) found that gender differences arose primarily from differential likelihood of outcomes for men and women, not the relative value of those outcomes. Men are, on average, more concerned than women about the financial costs of child-rearing and in patrilineal societies about having sons to continue the family name. Women place greater importance than do men on the work and strain of raising children, and in some cases also the opportunity costs of children for other activities, and the benefits of children for the marital relationship (Fawcett, 1983).

Differences in the perceived benefits of daughters and sons are also related to differences in the roles and behaviors of men and women (Fawcett, 1983). In patrilineal societies, sons are valued more than daughters for kinship ties, that is, to continue the family name. Where men have greater access than women

to economic opportunities, sons may also be valued for future financial assistance. Daughters are valued more than sons for household and childcare help and companionship. In patrilineal societies, the values associated with sons are produced for the most part in adulthood, those associated with daughters during childhood, consistent with the practice of daughters' leaving the parental home upon marriage. Many of the benefits of children, particularly social and psychological benefits, do not appear to differ for sons and daughters. Thus, where children provide few economic benefits, few parents express preferences for sons versus daughters, but rather prefer at least one of each (Hank, 2007).

Socioeconomic Variation in the Value of Children

As noted above, the economic value of children is associated with agricultural and household economic production and the absence of social insurance for elderly parents. Although children also entail high economic costs in such contexts, they are necessary for survival. Industrialization and urbanization reduce the economic value of child labor and increase the costs of rearing children to be economically independent. Economic development eventually leads to the development of social insurance that reduces further the economic value of children for support in old age. As a result, children's net economic value is perceived to be lower in industrialized wealthy countries than in poorer countries with a greater dependence on agricultural and household production (Fawcett, 1983; Nauck and Klaus, 2007), except for those at the bottom of the economic ladder in wealthy countries (Rendall and Bahchieva, 1998).

Psychological benefits and costs of children were reported in the original VOC surveys to be more important in industrial and postindustrial settings than in agricultural settings (Fawcett, 1983). This difference could arise because of the priority of economic survival over psychological well-being (Zelizer, 1985). Children may provide psychological benefits (and costs) for parents in all settings, but these components of child value become salient only when children become irrelevant to economic survival. The more recent VOC surveys find no difference across societies at different levels of economic development in the emotional value of children (Nauck and Klaus, 2007; Mayer and Trommsdorff, 2010).

Because the social capital or social esteem value of children has only recently been introduced into theoretical discussions of the value of children, it is difficult to know how such values might vary under different economic conditions. In agricultural settings, there may be a stronger association between economic and social ties so that the latter are not distinguishable from the former. Only in industrial and urban societies may social capital be sufficiently separable from financial capital to identify it as a separate source of child value.

Socioeconomic variation in the value of children is also evidenced across families within societies (Fawcett, 1983). A consistent finding from VOC surveys was that urban respondents place a lower economic value and a higher emotional value on children than do respondents living in rural areas. Similarly, education is inversely associated with children's economic value and directly associated with their emotional value as well as with perceived restrictions or opportunity costs of parenthood. Direct financial costs and childcare stresses do not vary substantially across countries or across individuals in different economic circumstances, likely due to the greater investments parents of higher socioeconomic status make in each child. Although one might expect the opportunity costs for mother's employment to increase with socioeconomic status, well-paid mothers can purchase substitute care. Klein and Eckhard (2007) found no differences in the perceived opportunity costs of children or in their effects on childbearing across socioeconomic levels.

Culture and the Value of Children

Broad cultural values may also serve as sources of specific or general values of children. Religious institutions and beliefs may support the value of children for social and psychological benefits. For example, Catholicism is viewed as a support for large-family values in the Philippines, Confucianism for the high value of sons to carry on the family name in some Asian countries. The relative values of daughters and sons are associated with broad cultural values on gender equality (Fawcett, 1983).

Lesthaeghe (1983) argues that ideational change is an independent force underlying current low fertility in Western countries. He identifies the two most salient features of this change as secularization and individuation. Secularization allows more latitude to individual morality; individuation stresses the importance of personal self-actualization. Using national surveys of social and family values, Lesthaeghe and colleagues (e.g., Lesthaeghe and Meekers, 1986) distinguish two dimensions of family values – a nonconformity dimension linked to partner relationships and nonmarital childbearing; and the 'meaning of parenthood,' including beliefs that children are necessary for fulfillment and for marital success. Measures of secularism and individuation were strongly associated with nonconforming family values, but only weakly associated with the meaning of parenthood.

This finding is consistent with the argument of Ehrhardt and Kohli (2011) who distinguish two dimensions of individualization. The first dimension, an increase in status of the individual – particularly children and women – underlies the decreased economic value and increased economic costs of children, including opportunity costs for their mother's employment. The second dimension, an increase in autonomy and freedom of choice, does not imply an increased competition between self-actualization and parenthood. It does underlie, however, dramatic decreases in union stability and the economic safety net for mothers and children, operating not through children's social and psychological benefits but through their economic costs.

The Value of Children and Fertility

Hoffman and Hoffman (1973) theorized that the influence of psychological needs met by parenthood or by particular numbers of children would depend on the intensity of such needs and the availability of alternative mechanisms for fulfilling them. Studies using ratings such as those in the VOC surveys have generally found that high perceived economic

benefits of children are associated with a large family size. The importance of psychological benefits and restrictions on parental activities are associated with a small family size (Fawcett, 1983; Nauck, 2007; Mayer and Trommsdorff, 2010). The relationship between perceived child values and fertility is not, however, as strong as some scholars had hoped, and the financial and time/effort costs of children are not associated with family size (Fawcett, 1983; Nauck, 2007).

Relatively moderate associations between values of children and completed fertility should not be surprising. As noted above, many child values may be parity specific and should be expected to influence only specific parity progressions. Further, child values can influence only desired fertility, so that the relationship between values and outcomes depends on the degree of fertility control. In addition, early measures of the perceived value of children were relatively crude (e.g., three-point response scale). When a particular parity progression is specified, when contraception is pervasive, and when precise measures of the expected value (net of cost) of children are generated, much stronger associations are observed between values and births (e.g., Liefbroer, 2005; Nauck, 2007).

Nauck (2007) argues further that fertility decision-making is likely to be rational – considering carefully the benefits and costs – primarily during a period of fertility transition. Before and after the transition, when most people share an understanding of children's values and costs, decision-making will be more routinized and spontaneous, thus more strongly influenced by prevailing social norms.

See also: Demographic Measurement: General Issues and Measures of Fertility; Family Size Preferences; Family Theory: Economics of Childbearing; Family Theory: Economics of Intergenerational Relations; Fatherhood; Fertility Theory: Theory of Intergenerational Wealth Flows; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Fertility and Social Networks; Second Demographic Transition.

Bibliography

Billari, F.C., Philipov, D., Testa, M.R., 2009. Attitudes, norms and perceived behavioural control: explaining fertility intentions in Bulgaria. European Journal of Population 25, 439–465.

Buehler, C., 2008. On the structural value of children and its implication on intended fertility in Bulgaria. Demographic Research 18, 569–610.

Bulatao, R.A., 1981. Values and disvalues of children in successive childbearing decisions. Demography 18, 1–25.

Cain, M., 1985. Fertility as an adjustment to risk. In: Rossi, A.S. (Ed.), Gender and the Life Course. Aldine, New York, pp. 145–160.

Caldwell, J.C., 1976. Toward a restatement of the demographic transition theory. Population and Development Review 2, 321–366.

Davidson, A., Jaccard, J., 1975. Population psychology: a new look at an old problem. Journal of Personality and Social Psychology 31, 1073–1082.

DiPrete, T.A., Engelhard, H., Morgan, P., Pacalova, H., 2003. Do cross-national differences in the costs of children generate cross-national differences in fertility rates? Population Research and Policy Review 22, 439–477.

Ehrhardt, J., Kohli, M., 2011. Individualisation and fertility. Historical Social Research 36 (2), 35–64.

Fawcett, J.T., 1978. The value and cost of the first child. In: Miller, W.B., Newman, L.F. (Eds.), The First Child and Family Formation. Carolina Population Center, Chapel Hill, NC, pp. 244–265.

Fawcett, J.T., 1983. Perceptions of the value of children: satisfactions and costs. In: Bulatao, R.A., Lee, R.D. (Eds.), Determinants of Fertility in Developing Countries, Supply and Demand for Children, vol. 2. Academic Press, New York.

Friedman, D., Hechter, M., Kanazawa, S., 1994. A theory of the value of children. Demography 31, 375–401.

Griffith, J.D., Koo, H.P., Suchindran, C.M., 1985. Childbearing and family in remarriage. Demography 22, 73–88.

Hank, K., 2007. Parental gender preferences and reproductive behaviour: a review of the recent literature. Journal of Biosocial Science 39, 759–767.

Hoffman, L.W., Hoffman, M.L., 1973. The value of children to parents. In: Fawcett, J.T. (Ed.), Psychological Perspectives on Population. Basic Books, New York, pp. 19–76.

Klein, T., Eckhard, J., 2007. Educational differences, value of children and fertility outcomes in Germany. Current Sociology 55, 505–525.

Lehrer, E.L., Grossbard-Shechtman, S., Leasure, J.W., 1996. A theory of the value of children – comment. Demography 33, 133–139.

Lesthaeghe, R., 1983. A century of demographic and cultural change in Western Europe: an exploration of underlying dimensions. Population and Development Review 9, 411–435.

Lesthaeghe, R., Meekers, D., 1986. Value changes and the dimensions of familism in the European Community. European Journal of Population 2, 225–268.

Liefbroer, A.C., 2005. The impact of perceived costs and rewards of childbearing on entry into parenthood: evidence from a panel study. European Journal of Population 21, 367–391.

Mayer, B., Trommsdorff, G., 2010. Adolescents' value of children and their intentions to have children: a cross-cultural and multilevel analysis. Journal of Cross-Cultural Psychology 41, 671–688.

Nauck, B., 2007. Value of children and the framing of fertility: results from a cross-cultural comparative survey in 10 societies. European Sociological Review 23, 615–629.

Nauck, B., Klaus, D., 2007. The varying value of children: empirical results from eleven societies in Asia, Africa and Europe. Current Sociology 55, 487–503.

Rendall, M.S., Bahchieva, R.A., 1998. An old-age security motive for fertility in the United States? Population and Development Review 24, 293–307.

Rosenzweig, M.R., 1978. The value of children's time, family size and non-household child activities in a developing country: evidence from household data. In: Simon, J.L. (Ed.), Research in Population Economics: An Annual Compilation of Research, vol. 1, pp. 331–347.

Schoen, R., Kim, Y.J., Nathanson, C.A., Fields, J., Astone, N.M., 1997. Why do Americans want children? Population and Development Review 23, 333–358.

Schultz, T., 1973. The value of children: an economic perspective. Journal of Political Economy 81, S2–S13.

Seccombe, K., 1991. Assessing the costs and benefits of children: gender comparisons among childfree husbands and wives. Journal of Marriage and Family 53, 191–202.

Thomson, E., 2004. Step-families and childbearing desires in Europe. Demographic Research, Special Collection 3, 117–134.

Zelizer, V., 1985. Pricing the Priceless Child: The Changing Social Value of Children. Johns Hopkins University Press, Baltimore, MD.

China as a Rising Aid Donor

Shino Watanabe, Sophia University, Tokyo, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

China's foreign aid dates back to 1950. With a history of more than six decades, China has become one of the most important emerging donors. This article provided a comprehensive overview of China's foreign aid from the following perspectives: terms and definitions, aid modalities and volumes, aid projects, distribution and allocation, decision making, management, aid procedures, and major characteristics. It then analyzes China's contributions and challenges to international society.

Introduction

Since the late 1990s, the People's Republic of China (China) has been expanding its presence in many significant ways. Foreign aid is not an exception. China has been treated as one of the most important rising donors, along with other BRICS countries namely, Brazil, Russia, India, and South Africa. However, it is misleading to assume that rising donors are monolithic and regard China as a typical one (Manning, 2006; Sato et al., 2011; Rowlands, 2012). They are heterogeneous with their distinct characteristics, although some commonalities exist.

China's foreign aid dates back to 1950, not even a year after China was established in October 1949, meaning it has a history of more than 60 years as a donor. China's foreign aid has a far longer history than many other rising donors, with the exception of Russia and India. At the same time, China was a recipient of foreign aid at the outset of its establishment in 1949. Therefore, China's foreign aid has been shaped through its unique experience of simultaneously being a recipient as well as a donor (Watanabe, 2012b).

China's foreign aid is not necessarily well known despite such significance. The Chinese government has not disclosed detailed information on its foreign aid. In April 2011, China first published its white paper on foreign aid. Since then we have come to see some concrete and official information on China's foreign aid. Accordingly, the literature on China's foreign aid is slowly emerging not only in English but also in other languages, including Chinese and Japanese.

Building upon such growing literature, this article seeks to provide a brief but comprehensive overview of China's foreign aid. First, it deals with the thorny issue of terminologies and definitions related to China's foreign aid. It examines what is to be considered China's foreign aid. Second, it provides an overview of China's foreign aid based upon the first white paper. Third, it identifies major characteristics of China's foreign aid, in order to highlight the differences between China's foreign aid and those offered by major traditional Development Assistance Committee (DAC) donors. Fourth, it analyzes the major contributions and challenges of China's foreign aid. Finally, conclusions and some observations are provided.

Terms and Definitions

First and foremost, we should note that China's foreign aid is an ambiguous concept. China has been reluctant to call itself a donor or its efforts toward foreign countries as foreign aid. This is partly because China has been sensitive to the hierarchical relationship that would be implied by using these terms. China prefers to describe itself as a partner and use more mutual terms such as 'economic cooperation', 'development cooperation', and 'development assistance' (King, 2006).

China is willing to emphasize its partnership among developing countries. For many developing countries, the term 'donor' is often associated with concepts such as paternalism, hierarchy, and neocolonial interference; hence, they favor the term, 'development partners', indicating a horizontal relationship that deserves mutual benefits, noninterference, and respect for sovereignty (Mawdsley, 2012).

Therefore, it is no surprise that no official definition of foreign aid exists in China in a way that could be compatible with Official Development Assistance (ODA) of DAC within the Organization for Economic Cooperation and Development (OECD). DAC, a committee established within OECD, serves as a platform of traditional Western donors to promote primarily appropriate and desirable aid practices, enhances coordination among member donors, and monitors members' aid policies and practices.

In 1972, DAC defined ODA as those flows to countries and territories on DAC's List of ODA Recipients (www.oecd.org/dac/stats/daclist) and to multilateral institutions that are: (1) provided by official agencies, including state and local governments, or by their executive agencies and (2) each transaction of which (a) is administered with the promotion of the economic development and welfare of developing countries as its main objective and (b) is concessional in character and conveys a grant element of at least 25% (calculated at a rate of discount of 10%).

Meanwhile, rising donors define their foreign aid in their own ways. Obviously China's foreign aid does not conform to the ODA. From the outset, China has offered foreign aid based upon its own criteria and has not considered it under the category of ODA. Thus, it is misleading to call China's foreign aid ODA, although there is no publicly available definition of China's foreign aid (Davies, 2007). Even China's first white paper on foreign aid did not provide any official definition of China's foreign aid in a manner equivalent to DAC's definition (Information Office, 2011). It merely identified three categories of China's foreign aid: grants, interest-free loans, and concessional loans.

As a matter of fact, China's foreign aid is much broader than what is constituted as ODA. Although China's foreign aid does

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64130-2

include some efforts equivalent to ODA, it also embraces other transactions that would be classified under the Official Development Finance, namely, concessional loans that cannot satisfy the concessional terms of ODA (Brautigam, 2010). In other words, ODA equivalency alone is insufficient for understanding China's entire aid activity. The common wisdom is that Chinese government–sponsored investment activities are an integral part of its foreign aid and should be considered as such (Brautigam, 2009, 2010; Wolf et al., 2013).

Overview of China's Foreign Aid

Aid Modalities and Volumes

China's foreign aid falls in three categories: grants, interest-free loans, and concessional loans. Grants and interest-free loans are provided from China's state finances. Concessional loans are offered by the Export-Import Bank of China (ExIm Bank). China had offered 256.29 billion yuan in aid from 1950 to the end of 2009, including 106.2 billion yuan in grants, 76.54 billion yuan in interest-free loans, and 73.55 billion yuan in concessional loans (Information Office, 2011).

Grants are a sum of money given by the Chinese government for specific purposes, such as building social infrastructure (hospitals, schools, low-cost houses, and well and water-supply projects, among others) and implementing social welfare projects. Grants are also spent for small- and medium-size projects for human resource development, technical cooperation, and humanitarian assistance (Information Office, 2011).

Interest-free loans are mainly offered to developing countries that are economically solvent. They are used to build public facilities and implement projects that improve people's livelihoods. According to the white paper, interest-free loans last 20 years: 5 years of use, 5 years of grace, and 10 years of repayment (Information Office, 2011).

Concessional loans are the most prioritized modality in China today. They are offered mainly to finance infrastructure projects. The ExIm Bank uses the money it raises on the market as principal but can extend concessional loans with relatively low rates, as the Chinese government offers financial subsidies to the bank. According to the white paper, China's concessional loans are 15–20 years long, including a 5- to 7-year grace period, and the annual interest rate is between 2 and 3%. By the end of 2009, China had extended concessional loans to 76 countries and assisted 325 projects (Information Office, 2011).

With the exception of the figures disclosed in the first white paper, China's foreign aid still includes many unknowns. According to Wolf et al. (2013), China's aid pledges expanded dramatically – from $1.7 billion in 2001 to $189 billion in 2011 – for two reasons. First, China has accumulated a massive current account surplus and, hence has been plowed back into international society by way of foreign assistance. Second, China needs to secure a stable supply of natural resources to maintain its rapid economic growth. Foreign aid and government-sponsored investment are effective tools for enabling China to satisfy such needs. In contrast, they estimate that the annual average of aid delivered by China was $6.3 billion from 2000 to 2010 and while cumulative deliveries were $50 billion up to 2010, just 9.4% of China's total pledge (Wolf et al., 2013). The gap indicates that China's aid pledge has yet to be accepted at face value.

Aid Projects

China's foreign aid takes the following eight forms: complete projects, goods and materials, technical cooperation, human resource development cooperation, medical teams sent abroad, emergency humanitarian aid, volunteer programs in foreign countries, and debt relief.

Complete projects are conducted in production and people's livelihood in the fields of industry, agriculture, culture and education, health care, communication, power supply, energy, and transportation, either with grants or interest-free loans. They originated in China's foreign aid to Vietnam and the Democratic Republic of Korea (North Korea) for reconstruction and economic development in the 1950s. By the end of 2009, China had assisted with more than 2000 complete projects. Complete projects today are a major form of China's foreign aid and account for 40% of its aid expenditure (Information Office, 2011).

Goods and materials were the initial form of China's foreign aid and were offered for free. China provided a variety of goods and materials to recipient countries: machinery, equipment, medical devices, testing equipment, transport vehicles (including civil airplanes, locomotives, and container-testing equipment), office equipment, food and medicine, and so forth (Information Office, 2011).

Technical cooperation takes the form of sending Chinese experts in a variety of fields to enhance the self-development capacity of recipient countries. The fields include industrial production and management, farming and poultry raising, handicrafts, culture and education, sports and physical training, medical and health care, clean energy development, geological survey and prospecting, and economic planning. Even after complete projects are constructed, China trains people on production, operation, and maintenance of complete projects by means of technical cooperation. Such cooperation generally takes 1–2 years, with extensions possible based on requests from recipients (Information Office, 2011).

Human resource development cooperation takes the form of various training programs for experts and workers in recipient countries. It started in 1953, and by 1979, China had invited many trainees from foreign countries including North Korea, Vietnam, Albania, Cuba, and Egypt. In 1981, China began to offer practical training courses in developing countries in cooperation with the United Nations Development Program. In 1998, China started to offer seminars for government officials, and the programs extended to cover a variety of fields. China had trained more than 120 000 people by the end of 2009, covering more than 20 fields including economics, diplomacy, agriculture, medical and health care, and environmental protection. Today, China offers training for 10 000 people from developing countries annually (Information Office, 2011).

China first sent a team of medical doctors and nurses to Algeria in 1963 and has since sent medical teams to 69

developing countries in Asia, Africa, Europe, Latin America, the Caribbean, and Oceania. The teams not only offered medical services but also provided medical devices and medicines for free. By the end of 2009, China had sent more than 21 000 medical workers abroad, who have treated 260 000 million patients in developing countries (Information Office, 2011).

China offers emergency humanitarian aid to countries suffering from natural disasters in the forms of cash, materials, or dispatched personnel. In September 2004, China officially developed a response mechanism for emergency humanitarian relief and aid. China has provided emergency humanitarian aid to countries in many occasions, including bird flu in South Asian countries; a locust plague and cholera in Guinea-Bissau; dengue fever in Ecuador; H1N1 influenza in Mexico; serious earthquakes in Iran, Pakistan, Haiti, and Chile; a hurricane in Madagascar; tropical storms in Burma and Cuba; and a flood in Pakistan. China has also offered emergency food aid to many of the least developed countries, such as North Korea, Bangladesh, Nepal, Afghanistan, Burundi, Lesotho, Zimbabwe, and Mozambique (Information Office, 2011).

China began overseas volunteer programs by sending volunteers to developing countries in May 2002 and volunteer Chinese language teachers in 2003. The first country to which China dispatched volunteers was Laos. By the end of 2009, China had sent 405 young volunteers to 19 countries including Thailand, Ethiopia, Laos, Myanmar, Seychelles, Liberia, and Guyana. These volunteers served the local people in many ways, such as Chinese language teaching, traditional Chinese medicine treatment, agricultural technology, sports and physical training, computer skills, and international relief. China had dispatched 7590 Chinese language teachers to more than 70 countries by the end of 2009 (Information Office, 2011).

Finally, China has occasionally canceled its mature interest-free government loans for heavily indebted poor countries or least developed countries. China first announced such debt relief at the First Forum on China–Africa Cooperation (FOCAC) Ministerial Conference in 2000; it has since made similar announcements on other occasions, including the UN High-Level Meeting on Financing for Development in 2005, Beijing Summit of the FOCAC in 2006, UN High-Level Meeting on the Millennium Development Goals in 2008, FOCAC Fourth Ministerial Conference in 2009, and UN High-Level Meeting on the Millennium Development Goals in 2010. By the end of 2009, China had canceled 380 mature debts worth 25.58 billion yuan for 50 countries in Africa, Asia, Latin America, the Caribbean, and Oceania (Information Office, 2011).

Distribution and Allocation

The first white paper disclosed some information on the distribution of China's foreign aid. China has offered its foreign aid to 161 countries, including most of the developing countries by the end of 2009. Among them, 123 developing countries have consistently received foreign assistance from China, including 51 countries in Africa, 30 in Asia, 18 in Latin America and the Caribbean, 12 in Oceania, and 12 in Eastern Europe (Information Office, 2011).

The sector allocation of the concessional loans was as follows: 61.0% for economic infrastructure (transportation, communications, and electricity infrastructure), 16.1% for industry, 8.9% for energy and resources development (such as oil and minerals), 4.3% for agriculture, 3.2% for public facilities, and 6.5% for others areas (Information Office, 2011).

China prioritizes agriculture, rural development, and poverty reduction in recipient countries today. As such, its foreign aid covers a variety of fields: agriculture, industry, economic infrastructure, public facilities, education, and medical and health care, among others.

China's foreign assistance in agriculture includes (1) building farms, agrotechnology demonstration centers, and experiment and promotion stations of agrotechnology; (2) providing irrigation and implementing water-conservancy projects; (3) offering agricultural machinery and implements, farm produce processing equipment, and related agricultural materials; (4) sending experts to share agricultural production technologies and advice on rural development; and (5) training agricultural personnel (Information Office, 2011: p. 15). China had assisted 221 agricultural projects more specifically, 35 farms, 47 agrotechnology experiment and promotion stations, 11 animal husbandry projects, 15 fisheries projects, 47 farmland irrigation and water-conservancy projects, and 66 other types of agricultural projects (Information Office, 2011).

Industrial aid has been a major tool for China particularly from the 1950s to 1970s, and it assisted recipient countries to build the foundation for industrial development. In the 1970s, industrial aid was an important part of China's complete project aid. China had implemented 688 industrial production projects by the end of 2009 and assisted building many industries, such as light, textile, machinery, chemical, metallurgical, electronic, construction materials, and energy industries (Information Office, 2011).

Economic infrastructure has been an important sector of China's foreign aid. China had implemented 442 projects for economic infrastructure by the end of 2009. China has contributed to facilitating economic and social development in recipient countries by assisting in building economic infrastructure (Information Office, 2011).

China has also assisted recipient countries by building public facilities, such as civilian buildings, wells for water supply, conference centers, sports venues, culture venues, and facilities for scientific, educational, and medical care purposes (Information Office, 2011). China had built 687 public facilities in recipient countries by the end of 2009.

China has emphasized its foreign aid in education, particularly by building schools, offering teaching equipment and materials, sending teachers, training teachers and interns from other developing countries, and offering government scholarships to students from other developing countries to study in China. By the end of 2009, China had built more than 130 schools and offered funding for 70 627 students from 119 developing countries. China has built hospitals and medical care centers as well as malaria prevention and treatment centers, sent medical teams, trained medical workers, and provided medicines and other medical materials. China had built more than 100 hospitals and medical care centers by the end of 2009 (Information Office, 2011).

China has offered foreign aid in clean energy and coping with climate change for several decades. For instance, China assisted recipient countries in building small- and medium-size

hydropower stations and projects of power transmission and offered training courses on clean energy sources and climate change. China had held 50 workshops on the development and use of renewable resources, forestry management, and desertification treatment and prevention for more than 1400 people from developing countries. China had also offered biogas technologies to recipient countries to diversify their energy sources (Information Office, 2011).

Decision Making, Management, and Aid Procedure

It is unclear how China's foreign aid policy is determined. It seems that the top level determines the policy direction and budget allocation, working with the Chinese Communist Party's Leading Group and the State Council (IDRC, 2007). A State Council subgroup decides all foreign aid projects (Davies, 2007). The administrative department of foreign aid in China is the Ministry of Commerce (MOFCOM), which is responsible for foreign aid funds and is in charge of formulating foreign aid policies, regulations, and plans, examining and approving aid projects, and managing project execution. MOFCOM also engages in the tendering process and actual operations. The Executive Bureau of International Economic Cooperation, China International Center for Economic and Technical Exchanges, and Academy of International Business Officials – all affiliated with MOFCOM – oversee management of the implementation of complete projects and technical cooperation projects. The ExIm Bank is responsible for project assessment, allocation, and recovery of concessional loans (Information Office, 2011).

Meanwhile, MOFCOM communicates with the Ministry of Foreign Affairs (MOFA), Ministry of Finance (MOF), and the ExIm Bank when it drafts foreign aid programs and funds plans for each country. Other relevant government agencies also participate in the management of foreign aid programs based upon their expertise. In 2008, MOFCOM, MOFA, and MOF established a foreign aid interagency liaison mechanism to enhance coordination. It was developed into an interagency coordination mechanism in February 2011 (Information Office, 2011).

China's foreign aid takes the following procedure. First, the recipient government submits a loan application to the ExIm Bank. If the bank approves the application, it is forwarded to MOFCOM, and MOFCOM concludes a framework agreement with the recipient country (Davies, 2007). The bank then concludes a project loan agreement with the recipient. Because the loan is tied, Chinese contractors engage in the construction of the financed infrastructure projects (Kjollesdal and Welle-Strand, 2010).

Foreign Aid with Chinese Characteristics

China has favored the label 'South–South cooperation', and considers its foreign aid under this framework. This means that both China and the recipient countries are South, and are equal partners, which creates several distinctive characteristics of China's foreign aid.

First, China has been highly sensitive to the hierarchical relationship between donors and recipients. This might reflect China's long history of being recipients of foreign aid. China portrays itself as a developing country and thus advocates such principles as equality, mutual benefits, and mutual development between donors and recipients. Put more directly, national interests – not only political but also economic – are the main drive for China's providing foreign aid. In recent years, China has sought to increase its exports to and investment in recipient countries, secure access to raw materials and energy resources, and gain support from and political influence over recipient countries.

Second, South–South cooperation indicates that China is still a developing country, similar to the recipient countries. This means that China shares common problems that have been caused by the process of development in other developing countries. Of course, traditional donors do have experiences that could be applicable to developing countries as well. Because China is at a relatively similar stage of development, it has been dealing with impediments of development. Thus, China is more likely to have recent and firsthand experience in achieving economic development and realistic and effective solutions for many problems that developing countries might be able to apply. For China, foreign aid is one of the useful mechanisms for sharing its direct experience with other developing countries; for recipient countries, it makes the China model all the more appealing and easier to follow than those of advanced Western countries.

Yet, we should note that there is still no consensus as to what exactly constitutes the China model. In 2004, Joshua Cooper Ramo coined the phrase 'Beijing Consensus' as opposed to 'Washington Consensus' to describe China's development model that emphasizes political factors including a state-led initiative for economic development and self-determination (Ramo, 2004). Discussions and debates both inside and outside China have examined what constitutes the China model (Halper, 2010).

Third, China opposes the imposition of any political conditions in providing foreign aid. Based upon its Five Principles of Peaceful Coexistence (mutual respect for sovereignty and territorial integrity, mutual nonaggression, noninterference in each other's internal affairs, equality and mutual benefit, and peaceful coexistence), China perceives that imposing conditions on recipients involves an intervention into their internal affairs and, hence, rejects the idea of conditionality.

Traditional DAC donors tend to impose some kind of conditionality when they offer foreign aid. Aid conditionality basically takes some of the following forms: budget cut, small government, liberalization of foreign exchange and trade, deregulation, democratization, and good governance. They reflect the common belief shared among most of the aid community – namely, the Washington Consensus – based upon neoclassical economics. This belief implies that developing countries should follow certain paths for development – more specifically, those Western donors experienced in the process of their development.

In contrast, China respects the right of recipient countries to manage their own development paths. Based on its own experience of achieving dramatic economic growth over the last three decades by implementing its Open and Reform Policy and learning through trial and error, China appreciates the freedom of choosing its own development path suitable to its actual conditions. Thus, China advocates for the existence of

multiple paths for development, one of which is undoubtedly the China model.

Although China is less enamored with conditionality, it has requested that recipient countries adhere to the One-China policy rule: China offers foreign aid to countries that have diplomatic relations only with China, not Taiwan. For recipient countries, this rule would be a semi-conditionality in receiving China's foreign aid.

Major Impacts of China's Foreign Aid

China's different administrative norms and practices in its aid operation have produced mixed impacts not only for the donor community, but also for recipient countries. China's growing aid is undoubtedly a welcome engagement to international society.

First, as a rising donor, China has contributed to increasing the total amount of foreign aid flowing from donors to developing countries. By the end of 2009, China's foreign aid had reached most of the world, covering 161 countries and regions. In addition, China increased its budget for foreign aid by 29.4% annually between 2004 and 2009 (Information Office, 2011). Major DAC donors are still suffering from the negative repercussions stemming from the Lehman Brothers crisis in 2008, including economic downturn and high unemployment, and hence have difficulties in boosting budgets for foreign aid significantly. Thus, with its growing economy, China is a promising donor.

Second, China's manner of aid giving offers some strengths. China's swift decision making due to the top-down mechanism of the Chinese Communist Party and quick implementation of aid projects has been highly appreciated by recipient countries. Moreover, China has provided new aid modalities and ideas of development. Based upon recipient countries' growing need for capacity development, China continuously enriches its soft aid type cooperation in terms of human resource development, sending medical teams and experts oversees, not just expanding its hardware type assistance. China's resource-backed financing, the so-called Angola mode is another example. It is a mechanism through which China extends loans in return for resources, for instance oil in the case of Angola. Without such an arrangement, it would be impossible to receive foreign loans.

Meanwhile, the negative side of the ledger is also significant, causing several concerns. China's unique style of foreign aid challenges standardized foreign aid practices, which have been shaped through the international donor community, particularly among DAC donors. China's 'no strings attached' policy has been criticized for several reasons, such as postponing the necessary reform (both democratic and administrative), hindering the improvement of transparency and accountability, and undermining Western donors' leverage for enforcing policy reform in recipient countries. In other words, China's foreign aid offers a loophole to some of its recipient countries to circumvent conditionality that they would otherwise have to address (Manning, 2006).

Tying status is another critical problem. China's concessional loans are all tied as China requests that recipients of its concessional loans procure more than 50% of goods and services from China. This idea contradicts the existing norm of traditional donors, however. In 2001, DAC came up with a formal recommendation that aid for the least developed countries should be untied. Tied aid is perceived to undermine the effect of aid projects in recipient countries. A study by the World Bank pointed out that tied aid is costly to recipient countries, degrading the value of aid by about 25% (World Bank, 1998).

Last but not least, China's foreign aid projects have caused a variety of problems in recipient countries. Although we do not naïvely buy all the stories of problems related to China's foreign aid projects overseas that have been widely covered by media reports, it is true that China's foreign aid sometimes causes problems in recipient countries. China's emphasis on implementing economic infrastructure projects is more prone to corruption and political scandals than other types of foreign assistance. The behavior of Chinese companies and workers has caused various concerns. Chinese workers on China's foreign aid projects sometimes create tension in local communities as the projects do not create new employment opportunities for local people. The poor and severe labor conditions of Chinese workers scare the local community. In addition, some of the facilities and infrastructure constructed by Chinese companies are inferior in quality.

Conclusions and Observations

China is still an emerging donor. It is currently unrealistic to assume that China, in cooperation with other rising donors like BRICS countries as a bloc, challenges DAC by posing an alternative to the existing aid architecture. China's foreign aid has yet to become a substitute to the existing aid architecture (Rowlands, 2012).

Meanwhile the presence of emerging donors still has important implications for the current aid architecture. DAC's long-term efforts in promoting untying aid might be dampened with seemingly successful operations of emerging donors (Rowlands, 2012). In fact, the recommendation issued in 2001 has not been fully realized. Some DAC donors like Canada, Portugal, and Greece have continued to offer tied aid.

With the emergence of rising donors that overtly pursue their own interests, traditional donors began to regain some attention to commercial interests through foreign aid. Due to severe economic situations caused by the prolonged recession triggered by the Lehman Brothers crisis started in September 2008, traditional donors cannot help but take into account their economic gains through foreign aid.

The relationship between traditional donors and rising donors has become much more complex today. On the one hand, China's foreign aid norms and practices undermine some of the common efforts traditional donors seek to pursue through their ODA. On the other hand, emerging donors' aid efforts have uncovered some hidden problems and revealed some limitations of the existing aid norms and practices. The definition of ODA is a typical example. The analysis of China as a rising aid donor suggests that the measurement of development assistance based exclusively on ODA limits our understanding of international development assistance and hence needs to be reconsidered to produce better development outcomes.

In essence, China's foreign aid is not substituting for the traditional donors' ODA. They are rather complementary. Traditional donors and rising donors working in tandem can produce better outcomes of international development in this great revolutionary era for development assistance.

See also: Chinese Revolutions: Twentieth Century; Development Studies; Development and the State; East Asian Studies: Politics; Globalization and Transnational Corporations; Poverty Policy; Sustainable Development and Social Work; United Nations and its Agencies.

Bibliography

Brautigam, D., 2009. The Dragon's Gift: The Real Story of China in Africa. Oxford University Press, Oxford.

Brautigam, D., 2010. China, Africa and the International Aid Architecture. African Development Bank Group Working Paper, 107, Tunis.

Davies, P., 2007. China and the End of Poverty in Africa: Towards Mutual Benefit? Diakonia, Sundbyberg.

Halper, S., 2010. The Beijing Consensus: How China's Authoritarian Model Will Dominate the Twenty-First Century. Basic Books, New York.

Information Office of the State Council, 2011. The People's Republic of China. China's Foreign Aid, Beijing.

International Development Research Centre, 2007. Emerging Donors in International Development Assistance: The China Case. International Development Research Centre, Ottawa.

King, K., 2006. China's Partnership Discourse with Africa. http://www.fe.hku.hk/cerc/KK-article/3.China's%20Partnership%20with%20Africa.King.2006c.pdf (accessed 13.12.15.).

Kjollesdal, K., Welle-Strand, A., 2010. Foreign aid strategies: China taking over? Asian Social Science 6 (10), 3–13.

Manning, R., 2006. Will 'emerging donors' change the face of international co-operation? Development Policy Review 24 (4), 371–385.

Mawdsley, E., 2012. The changing geographies of foreign aid and development cooperation: contributions from gift theory. Transactions of the Institute of British Geographers 37, 256–272.

McCormick, D., 2008. China & India as Africa's new donors: the impact of aid on development. Review of African Political Economy 35 (115), 73–92.

Ramo, J.C., 2004. The Beijing Consensus. Foreign Policy Centre, London.

Rowlands, D., 2012. Individual BRICS or a collective bloc? Convergence and divergence amongst 'emerging donor' nations. Cambridge Review of International Affairs 25 (4), 629–649.

Sato, J., Shiga, H., Kobayashi, T., Kondoh, H., 2011. "Emerging donors" from a recipient perspective: an institutional analysis of foreign aid in Cambodia. World Development 39 (12), 2091–2104.

Severino, J.M., Ray, O., 2009. The End of ODA: Death and Rebirth of a Global Public Policy. Center for Global Development Working Paper 167.

Shimomura, Y., Ohashi, H. (Eds.), 2013. A Study of China's Foreign Aid. Palgrave Macmillan, Houndmills.

Watanabe, S., 2012a. China's foreign aid. In: Kim, H.S., Potter, D.M. (Eds.), Foreign Aid Competition in Northeast Asia. Kumarian Press, Sterling, pp. 59–80.

Watanabe, S., 2012b. Donors' impact on China: How have major donors affected China's economic development and foreign aid policy? In: Sato, J., Shimomura, Y. (Eds.), The Rise of Asian Donors: Japan's Impact on the Evolution of Emerging Donors. Routledge, London and New York, pp. 87–113.

Wolf Jr., C., Wang, X., Warner, E., 2013. China's Foreign Aid and Government-sponsored Investment Activities: Scale, Content, Destinations, and Implications. Rand Corporation, Santa Monica, CA.

Woods, N., 2008. Whose aid? Whose influence? China, emerging donors and the silent revolution in development assistance. International Affairs 84 (6), 1205–1221.

World Bank, 1998. Assessing Aid: What Works, What Doesn't, and Why. Research Report. Oxford University Press for the World Bank, Oxford and New York.

Chinese Literacy Acquisition: A Multidimensional Puzzle

Xiuli Tong, Faculty of Education, The University of Hong Kong, Hong Kong, China
Silvia Siu-Yin Lam, Communication Sciences and Disorders Northwestern University, Chicago, IL, USA
Catherine McBride-Chang, The Chinese University of Hong Kong, Hong Kong, China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Literacy is a multilevel construct comprising word reading, spelling, and text reading comprehension. Understanding the key linguistic and cognitive factors that support word reading, spelling, and text reading comprehension is fundamentally important to elucidate the detailed nature of literacy acquisition. While most contemporary models of literacy acquisition are developed on the basis of English and other alphabetic writing systems, little attention has been paid to nonalphabetic writing system, such as Chinese. We focus on Chinese literacy acquisition and emphasize how different properties of the Chinese language (phonology, morphology, and orthography) may impact the development of different-level literacy skills. A systematic review of extant research on literacy acquisition, along with our own work on Chinese word reading, spelling, and reading comprehension, suggests that phonological awareness, morphological awareness, and orthographic skills are all important to Chinese literacy acquisition, but the relative importance of these skills varies across word reading, spelling, and reading comprehension. Moreover, the contribution of each of these linguistic and cognitive skills to a single component of literacy skills also varies in different developmental stages. This suggests that a full model of literacy acquisition should consider the linguistic features that are characteristics of the writing system, as well as developmental context.

Literacy is multidimensional, comprising reading, spelling, reading comprehension, and higher-order writing skills (McBride-Chang et al., 2003). Literacy acquisition is a pressing challenge faced by every child. Thus, it is vital to understand the structures and cognitive processes that underlie the acquisition of each component skill in order to provide the best support for children's literacy acquisition. A long history of research suggests that the process underlying literacy acquisition is partly shaped by the properties of a given language exhibited in its writing system. In the following section, we focus on how specific features of languages influence the acquisition of different literacy skills in reading, spelling, and reading comprehension. Writing composition in Chinese is beyond the scope of discussion here given space limitations. In addition, we will mention both Mandarin Chinese and Cantonese, the most studied languages in Chinese in the literature, in various given contexts so as to have a better illustration of the properties in Chinese.

What Are the Relations of Component Literacy Skills?

There exist intrinsic and overlapping interrelationships among the skills or recognition and spelling of characters and words and reading comprehension. Word reading is a skill of mapping a print unit onto a speech sound unit carrying meaning (Mason et al., 1991). In the early stage of literacy development, many of the processes involved in word reading are sound–sight repetition (McBride-Chang et al., 2003). When children's reading ability becomes more sophisticated, they will rely on explicit knowledge of the morphological system and orthographic system in Chinese (Perfetti et al., 2005). The skills acquired from reading lay the foundation for spelling. Spelling is an encoding process to retrieve or reproduce the print code of the spoken unit (e.g., Perfetti, 1997). It is an integrated process that requires basic skills of orthographic processing, phonological skills, and morphemic analysis; these signify the fundamental principles of a given writing system and its distinctive characteristics relating to phonology, morphology, and orthography (Perfetti, 1997). It is not surprising to find that word reading and spelling are highly associated as the linguistic factors involved in these two literacy skills overlap. While the former requires one to reproduce the print via the spoken channel, the latter is a representation of the spoken unit using print. Nonetheless, in some writing systems, the cognitive skills required for spelling may be more in-depth than for word reading. For example, in Chinese, spelling requires not only the mapping of the sound unit and the print unit, but the precise understanding of the internal structures of a character in order to produce a correct spelling of a word. Findings from a longitudinal study by Tong et al. (2011) investigating the relationship of various reading-related cognitive tasks and Chinese word reading and spelling among kindergartners affirmed the idea that although there were overlapping cognitive processes in these two aspects of literacy, there were some distinctions between them. In this study, homophone awareness, visual skills, and syllable awareness were strongly associated with word reading, but these cognitive abilities were not related to children's spelling in their early stage of literacy development.

Word reading and spelling are fundamental for growth in reading comprehension (Stanovich, 1986). Reading comprehension requires multiple cognitive and metacognitive processing such as working memory, inference, reasoning, and thinking (Crowder and Wagner, 1992). Word reading and reading comprehension are highly intercorrelated, and word reading is highly predictive of children's reading comprehension (e.g., Cunningham and Stanovich, 1997; Yeung et al., 2012; Yin et al., 2007). According to Stanovich (1986), early competence with reading skill is a good predictor of later reading proficiency. Thus, children's good word decoding skills may help them to become good readers later on.

Furthermore, word reading is recognized as the most fundamental skill to facilitate children's self-achieved information from the word (Sticht and James, 1984). The process of recognizing a word reflects the initial understanding of the correspondence between the speech sound unit and the single writing unit, while spelling shares the same processing mechanisms with word reading (Ehri, 1997). It can strengthen the sound–symbol linkage and then enhance children's understanding of the sound system of a given language and of reading skills (e.g., Ehri, 1997).

However, given that the goal of reading is to derive meaning from print, reading comprehension is the end goal of literacy (Snow et al., 1998). Not surprisingly, there is an imperfect connection between early competence in reading skill and reading performance later on (e.g., Gough and Tunmer, 1986). There is accumulated evidence that word reading, spelling, and reading comprehension are correlated, but their correlations are not perfect, indicating that they are connected, but also dissociable. The processes and outcomes of these three core constructs of literacy might be influenced by other constructs relating to linguistic forms and characteristics, as well as the developmental level of children. Here, we focus on Chinese literacy acquisition and specify how different sound and meaning structures of Chinese influence the process of Chinese literacy acquisition. We focus first on phonology, the sound system of Chinese, then on morphology, the meaning system, and finally on orthographic processing of Chinese print itself, in turn, below.

Chinese Phonology: Syllable and Tones

Chinese has a relatively simple syllable structure, consisting exclusively of syllables that are composed of V (vowel) only, as in /o1/阿 (fawn on), CV (consonant-vowel) only, as in /ma1/媽 (mother), CVC (consonant-vowel-consonant) only, as in /shan1/山 (mountain), or VC (vowel-consonant) only, as in /an4/暗 (dark), and CVVC /liang4/亮 (lightening); there are no consonant clusters, as commonly observed in alphabetical languages such as English. The examples used here are Mandarin Chinese as it is most often discussed in research and it represents one general picture of syllable structure in Chinese. CV is the most common structure in Mandarin Chinese. Although Cantonese has similar syllable structures, it is a little bit more complicated than Mandarin Chinese. For example, Cantonese allows more consonants at the end of the syllable. Despite the differences between Mandarin Chinese and Cantonese, syllabic awareness is particularly important for reading aloud in Chinese.

Also, Chinese is a tonal language. In Chinese, the change of pitch of a spoken syllable can result in changes of meanings. When talking about tones in Chinese, it is particularly interesting to use Cantonese as an example. Among Chinese languages/dialects, Cantonese has a rather complicated tonal structure that consists of at least six tones, i.e., high-level (55), high-rising (25), mid-level (33), low-falling (21), low-rising (23), and low-level (22) (Chao, 1947). For example, a single spoken syllable /fu/ can represent six different meanings of /fu55/膚 (skin), /fu25/虎 (tiger), /fu33/褲 (trousers), /fu21/符 (symbol), /fu23/婦 (woman), /fu22/父 (father) when pronouncing it with six different tones in Cantonese. Moreover, unlike other prosodic features such as intonation and stress in English, tone is fixed for a given syllable and it does not vary much across contexts (the sandhi rule, in which tones of words are influenced by adjacent words or syllables, is an exception). The ability to distinguish tone is crucial in order to have a precise understanding of the meaning of words and to be able to pronounce words correctly. There is evidence that a deficit in perceiving tone is associated with difficulties in learning to read aloud (Tong et al., 2013) and to comprehend the meaning of words (Tong et al., submitted).

Chinese Morphology: Homophone and Lexical Compounding

Chinese has a distinctive morphology in that, for the most part, each character is a morpheme, or (smallest) meaning unit. In Chinese, one morpheme maps onto one syllable. However, the number of spoken syllables is far less than the number of morphemes. For example, the number of spoken syllables used in Mandarin is about 1200, whereas regular words consist of over 7000 syllables (Li et al., 2002). Cantonese includes only 753 syllables excluding tone (Leung et al., 2004). That results in a large number of homophones, so that different morphemes (or characters) are pronounced using the same sound. For instance, one monosyllable /*lann4*/ can represent more than five characters 蘭 (orchid), 欄 (column), 瀾 (wave), 攔 (obstruct), 闌 (waning), 斕 (colorful) with different meanings in Cantonese. Therefore, without contextual cues, it is very challenging to identify the meaning when hearing a single syllable. However, by pairing the syllable with another appropriate character, for example, 專欄/zyun1 lann4/ (special column), one is able to derive the meaning from its spoken form.

Lexical compounding is the most frequently used word formation process in Chinese (Packard, 2000). In Chinese, one single morpheme can combine with other morphemes to form a new word. For example, the single morpheme 書 (*book*) can combine with other morphemes in forming into a group of compounds including 書包 (book bag), 書架 (bookshelf), and 書桌 (reading desk).

However, it is worth mentioning that the definition of morpheme in Chinese can be explained at two separate levels. One level was previously mentioned, i.e., that a single character is usually also a morpheme. For example, a single Chinese character 好/hou2/ can indicate a variety of meanings such as (1) with much merit, (2) good, the opposite of bad, (3) good relationship, and (4) accomplishment. A second level lies at the level of a structural relationship within the internal components of a character: A semantic radical offers a semantic marker. For example, 犭 and 忄 contain partial information, the former indicating canine and the latter indicating emotions or feelings. However, semantic radicals provide only rough semantic categorization of a group of characters, rather than giving exact meaning information for a given character. Therefore, it is relatively difficult to understand whether a partial meaning unit, such as a semantic radical, can be a morpheme or not given the current definition of morpheme. Also, semantic radical is part of print unit and it is operationally

challenging to measure it without involving orthographic information.

As is clear from the above description, it remains open how we could adequately conceptualize morphology and morphological awareness. In general, the morpheme could not be simply defined as a single word form. Here, we conceptualize of morpheme as an independent representation of a linguistic unit which has its own sound, meaning, and orthography, such as 車 (vehicle). Each morpheme can contain either meaning or structural features. In particular, each morpheme must have a specific meaning which is different from its unpronounceable radical. As to structure, it involves the word form and sound variation and modes for forming compound words.

Chinese Orthography: Strokes, Radicals, and Characters

Chinese has a relatively complex orthographic structure. Unlike alphabetical letters, Chinese characters are formed using strokes, which are the basic and smallest units of a character. The number of strokes of a character, on average, varies from 1 to 23. A character can be formed by as few as one stroke, such as 一 (one), or it can be formed from many strokes, which in turn comprise different radicals, such as 唱 (sing). The structure of the latter is most commonly found in Chinese characters; about 90% of Chinese characters comprise of two radicals, including a semantic radical and a phonetic radical (Ho et al., 2003). Positional regularity is found in Chinese characters. The semantic radical is usually positioned on the left while the phonetic radical falls on the right. To illustrate, a Chinese character, 泳/yong3/, which means swim, comprises a left-hand side semantic radical 氵 indicating a water-related concept and a right-hand side phonetic radical 永 pronounced as /yong3/. This regularity has been described as part of the 'orthography–phonology correspondence rules' (Ho and Bryant, 1997).

However, the links between the sound, meaning, and form in Chinese are much less systematic in comparison to English and other alphabetic languages. The associations of sound, meaning, and forms are somewhat arbitrary (Zhou and Marslen-Wilson, 1999). Around 15% of characters are irregular characters, the sounds of which differ from their phonetic radicals (Shu et al., 2003). Examples include 初/co1/ (first) (primary character) versus 刀/dou1/ (phonetic radical), 錯/co3/ (mistake) versus 昔/sik1/ (the past), 台/toi4/ (stage) versus 始/ci2/ (start); for all of these, the phonetic cues are not valid for naming these characters. It commonly occurs in Chinese writing systems that the similar visual-configuration or graphic represent very different meanings or sounds. For example, the single Chinese characters 土/tou2/ (soil), 士/si6/ (soldier), 工/gung1/ (worker) are composed of the same number of identical strokes, but the sounds and meanings are markedly different. On the other hand, no systematic grapheme–phoneme correspondence makes the separation of phonological awareness and orthographic measure feasible in Chinese. These typical features relevant to Chinese orthography may suggest that Chinese children need to rely on visual distinction or memorization of visual-orthographic patterns to correctly identify Chinese characters.

Literacy Acquisition: Phonology, Morphology, and Orthography

Reading and writing are the most important literacy skills learned by young children and they are likely to be intrinsically linked with metalinguistic knowledge (Nagy and Anderson, 1999). Several models of early reading have established the connections among these metalinguistic skills across languages (e.g., Perfetti et al., 2005; Seidenberg and McClelland, 1989). A single mechanism model based on English word reading has conceptualized that phonology, morphology or semantics, and orthography are three interlinked constituents of word recognition. Moreover, the representation of each constituent is influenced by the levels of the other constituents (Plaut et al., 1996; Seidenberg and McClelland, 1989). Extending one single mechanism model to Chinese word reading, the lexical constituency model also recognizes that word identity is constrained by phonology, semantics, and orthography. The process of identifying the word involves a multifacet activation of these three constituents. The sound information is simultaneously activated along with the process of retrieving the meaning from a visual-orthographic unit (Perfetti, Liu, and Tai, 2005). In addition, one recent study used a triangle model to simulate Chinese word reading and found that results were quite different from English in that the processing route from orthography to semantic was much faster than the mapping from orthography to phonology, whereas it was the reverse in English (Yang, Shu and McCandliss, 2013).

Across computational models of word reading, phonology, semantics, and orthography are consistently viewed as three core constructs of word reading (Perfetti et al., 2005). Of these three components, phonology is viewed as the sound system relating to a given language. Phonological awareness is the counterpart construct to measure children's sensitivity or awareness of the sound structure of language in reading study. Semantics can be understood as meaning of the words and sentences in terms of a linguistic framework. Morphology is focused on the study of morphemes. Given this, morphology and semantics are inextricably linked and mutually represented at the word level. Particularly, morphological awareness is most often used to assess semantic knowledge or structure in the reading realm. In addition, orthographic processing refers to understanding a set of conventional or legitimate rules in structuring the written unit. As a logographic system, Chinese has no phoneme–grapheme correspondence. However, there exist some prespecified rules in constraining unpronounceable strokes or components in forming a meaningful square-shaped character.

That is, there are a number of rules codified in the Chinese writing system, defining the specific order of strokes in forming proper Chinese characters. There are six basic types of stroke rules, namely horizontal before vertical, left diagonal before right diagonal, top before bottom, left before right, outside before inside, and inside internal stroke before the outside-sealing stroke. Therefore, sensitivity to the graphic structure and sequence of laying out the strokes might be helpful to correctly identify or retrieve the character to some extent. The potential functions of strokes in Chinese have been documented in a few studies. In a priming study, Flores d'Arcais (1994) found that stroke sequence could serve

as a lexical retrieval from memory in reading Chinese characters. Stroke sequence itself will likely activate one's orthographic presentation from the memory; hence, it is easier to retrieve related characters.

As mentioned, computational models of word reading underscore the importance of phonology (Perfetti et al., 2005). To assess children's knowledge of phonology in their given knowledge, we test their phonological awareness. Phonological awareness is a single and unified ability to recognize, discriminate, and manipulate the sounds in one's language and it manifests itself in different skills through preschool and the early elementary school years (Adams, 1990). Moreover, phonological awareness is bidirectionally associated with reading and writing (Perfetti et al., 1987) and it is the most reliable metalinguistic prerequisite for reading in an alphabetic writing system, primarily because reading and writing can strengthen children's phonological awareness development via this interactively reinforced process (Bialystok et al., 2005a; Bialystok et al., 2005b; for a review, see Adams, 1990). Phonological awareness is a strong predictor of children's progress in learning to read in Chinese (Bialystok et al., 2005b; McBride-Chang et al., 2004).

However, the relative importance of phonological awareness for reading and writing may vary across languages. For example, the prediction power of syllable awareness is robust in Chinese while phonemic awareness is predominant in English (McBride-Chang et al., 2004). Moreover, phonological awareness may be language-specific and its primacy level depends on the accessibility of phonological units in a given language system, while reading skills are developed separately and are determined by language proficiency and instruction (Bialystok et al., 2005b). For example, one study that compared the performance of phonological awareness and reading skill among English–Cantonese bilinguals, Cantonese–English bilinguals, and monolingual English speakers demonstrated that the two bilingual groups performed equally well in syllabic awareness, because the syllable is the most salient unit in Chinese. In contrast, both groups of Canadian children outperformed Hong Kong Cantonese–English bilinguals on phoneme-onset deletion phoneme counting tasks (Bialystok et al., 2005b), at least in part because phoneme skills are not emphasized in work with Hong Kong Chinese children. A 2-year longitudinal study (Tong et al., 2011) also showed that syllable awareness uniquely contributed to Chinese word reading and spelling among kindergarteners across time after all other contributions such as age, vocabulary knowledge, and nonverbal IQ were statistically controlled.

There were several other studies (e.g., Chow et al., 2005; McBride-Chang et al., 2004) that also found that syllable awareness was a strong predictor of reading ability among young children in Chinese. However, it is noteworthy that the importance of phonological awareness may diminish with age. As shown in the study by Tong et al. (2011), the contribution of syllable awareness to word reading decreases with age. Most likely, strategies or processes involved in word reading change as one's reading abilities get more sophisticated. Perhaps it is because of the unique writing system in Chinese that phonological structure is not as salient as it is in alphabetic languages. The reader may need to rely on other cognitive skills such as morphological awareness and orthographic awareness to further assess the meaning of words in Chinese. Homophones were a consistently unique predictor for word reading across time in that study. As mentioned previously, homophones constitute a salient feature in the Chinese language system and hence children's sensitivity toward homophones may explain their ability in mapping a single syllable to different meanings and different print configurations, facilitating their abilities in word reading.

Tone awareness is another crucial element in Chinese phonology. As suggested by several studies (e.g., McBride-Chang et al. (2008); So and Sigel, 1997; Yeung et al., 2012), tone detection ability is highly associated with children's reading skills. A study conducted by So and Siegel (1997) to examine the relationship between Chinese reading ability and tone awareness among primary school children showed that the development of tone awareness may facilitate word decoding skills. The study showed that children in the poor reading group particularly have problems in mapping printed characters with sounds as well as in understanding the meanings of characters. That suggests an interaction between phonology and semantics within the orthography in Chinese. The study suggested that the ability to manipulate the tones system in order to pronounce words correctly and to distinguish different homophones with variations in tone may be highly associated with children's word reading (e.g., Leong, 1996). Therefore, it is reasonable to assume that homophones and tone awareness, two salient features in the Chinese phonological system, are important in understanding children's reading acquisition. However, the importance of these cognitive processes varies with children's age, and that fact may explain why the decoding processes involved in word reading change when children's reading becomes more sophisticated.

According to Ku and Anderson (2006), learning to read is not only a task of decoding the orthographic unit, i.e., converting it to a phonological unit, but also of mapping the sound to the meaning unit (morpheme). Morphological awareness is viewed as the ability to be aware of and manipulate the morpheme, employing rules of word formation. In the intervention training conducted by Zhou and Marslen-Wilson, 1999, after a 2-month training on lexical compounding, a good indicator or morphological awareness in Chinese, children's word reading had improved significantly. An emerging consensus in reading is that morphological awareness is one important construct to understand reading acquisition beyond phonological awareness across orthographies (Carlisle, 2000; Li et al., 2002; Packard, 2000). Inflections, derivations, and lexical compounding are the most often tested in alphabetic languages (Anderson, 2003; Casalis and Louis-Alexandre, 2000; Lewis and Windsor, 1996; Nagy et al., 2006). Moreover, children's acquisition of inflectional knowledge comes earlier than derivational and compound knowledge. However, inflectional morphology is not a particularly important or strong feature of Chinese. In contrast, compounding morphology is predominant in Chinese because a great number of Chinese words are compound words. Moreover, children's ability to flexibly manipulate the morpheme might facilitate the process of mapping a speech unit onto a Chinese character (McBride-Chang et al., 2003). However, the extent to which morphological awareness contributes to different components of reading (e.g., single-morpheme and multiple-morpheme word reading and text reading

comprehension) may vary because the demands for meaning information necessary for single-morpheme reading, multiple-morpheme reading, and text reading comprehension differ. This is an unresolved issue that needs further research.

In the previous paragraphs, we mentioned the relationship of morphology and orthography. Orthography plays a particularly important role in Chinese due to its complex visual structures. Orthographic awareness is sensitivity to orthographic regularities (units of print) in the language. There is substantial evidence showing that orthographic knowledge affects Chinese children's performance in word reading (e.g., Ho et al., 2003; Lin et al., 2010; Packard et al., 2006). The association of orthographic knowledge and children's word reading was shown by Ho et al. (2003) in their study of pseudocharacter reading tasks in Hong Kong Chinese kindergartners, first-graders, and third-graders. Children's use of radical knowledge in relation to position and function regularity increases as children's reading competency grows. These findings offer evidence that orthographic knowledge of stroke patterns and positional constraints of radicals play an important part in Chinese word reading, and they lend credence to the hypothesis that understanding the positional regularity of functional components of the Chinese orthography may contribute to Chinese word reading. Also, a mother–child joint study conducted by Lin et al. (2010) reported that by implementing analytical strategies such as the orthographic strategy of teaching kindergarten children functions of semantic and phonetic radicals, children showed better performance in their word reading skills. Additionally, they found that orthographic awareness uniquely explained character reading.

Orthographic processing is viewed as a construct that is entangled with phonological awareness in English word reading (Foorman, 1994) and some evidence from previous studies has suggested that orthographic processing and phonological awareness are reciprocally influenced in reading and spelling across languages (e.g., Foorman, 1994). The route and the magnitude of the contribution of orthographic processing to word reading might vary across languages. One study from Huang and Hanley (1995) compared word reading among children from England, Taiwan, and Hong Kong. In that study, visual paired associates learning skills, involving memorization and linking of color to abstract figures, was uniquely associated with word reading ability in Taiwan and Hong Kong even with IQ and vocabulary skills partialled, but not among British children. Also, visual-orthographic measures contributed substantially to early Chinese word reading whereas phonological awareness predicted later Chinese reading (Ho and Bryant, 1997b). In addition, Siok and Fletcher (2001) adopted diverse levels of visual-orthographic and phonological awareness measures and suggested that learning to read Chinese develops from a logographic to an orthographic-phonological stage (Siok and Fletcher, 2001).

Although the existing literature is mainly focused on the linkage between metalinguistic skills and word reading, there are also several studies that have moved beyond word reading to other literacy skills, such as spelling and reading comprehension. As described before, spelling and word reading are different sides of the same process (Ehri, 1997). Spelling is the process of employing multiple metalinguistic skills, as well as orthographic knowledge, across developmental stages (Ellis, 1998). Moreover, spelling requires the integration of metalinguistic skills related to speech with orthographic knowledge about print (Leong, 1999; Treiman, 1993). Of these, orthographic knowledge is initially used in reading and is further transferred to spelling (Firth, 1985). Also, morphological awareness and orthographic processes influence early spelling (Bryant, Nunes, and Bindman, 1999; Goswami, 1999). There exists evidence suggesting that phonological awareness, morphological awareness, and orthographic processing are all effective predictors of spelling, but the extent to which each linguistic or cognitive skill is important for spelling varies by language and task (Walker and Hauerwas, 2006). For example, Tong et al. (2011) found that syllable awareness and orthographic skills contributed to kindergarteners' word spelling. However, this effect was eliminated once word wording was statistically controlled. A theoretical reading model proposed by Yeung et al. (2012) showed that morphological awareness uniquely explained first-graders' word reading and spelling when other cognitive constructs such as phonological skills, orthographic skills, and rapid naming were statistically controlled. In this model, both orthographic skills and morphological skills are individually important for literacy development in Chinese.

In contrast, there have been few studies on the association between metalinguistic or cognitive constructs and reading comprehension among young children. Reading comprehension is a higher-order reading skill which requires the reader to construct meaning from print. Morphological awareness is crucial to the development of reading comprehension. A growing number of studies suggest that children's performance on morphological awareness tasks are predictive to the growth of reading comprehension even when controlling critical reading-related measures, such as phonological awareness and vocabulary (e.g., Deacon and Kirby, 2004; Kirby et al., 2012; Carlisle, 1995; Kieffer and Lesaux, 2012).

Over the last two decades, research on Chinese literacy acquisition has made significant progress in identifying factors that affect literacy acquisition. Different aspects of literacy skills overlap with each other but meanwhile each of these skills has its distinctive elements. However, how these factors interact in the process of literacy development, and jointly contribute to the growth of literacy skills, remains unclear. New research is needed to solve the multifaceted puzzle. In addition, there are few studies on Chinese reading comprehension, so future work may probe further into the relationship between reading comprehension and other cognitive constructs, as well as the relationship between reading comprehension and other literacy skills. These are important in understanding writing and reading acquisition and may help to solve the multifaceted puzzle of Chinese literacy.

See also: Case Studies; First Language Acquisition, Linguistic Theory of; Language Acquisition; Language Development, Theories of; Multilingual Language Development; Reading Nonalphabetic Scripts; Second Language Learning; Writing Systems.

Bibliography

Adams, M.J., 1990. Beginning to Read: Thinking and Learning about Print. MIT Press, Cambridge, MA.

Anthony, J.L., Lonigan, C.J., 2004. The nature of phonological sensitivity: converging evidence from four studies of preschool and early grade-school children. Journal of Educational Psychology 96, 43–55.
Anderson, R.C., L., W., Ku, Y.-M., Shu, H., Wu, N., 2003. Children's use of partial information in learning to read Chinese. Journal of Educational Psychology 95, 52–57.
Bialystok, E., Luk, G., Kwan, E., 2005a. Bilingualism, biliteracy, and learning to read: interactions among languages and writing systems. Scientific Studies of Reading 9, 43–61.
Bialystok, E., McBride-Chang, C., Luk, G., 2005b. Bilingualism, language proficiency, and learning to read in two writing systems. Journal of Educational Psychology 97, 580–590.
Bowey, J., Miller, R., 2007. Correlates of orthographic learning in third-grader children's silent reading. Journal of Research in Reading 30, 115–128.
Bryant, P., Bradley, L., 1985. Children's Reading Problems. Blackwell, Oxford.
Bryant, P., Nunes, T., Bindman, M., 1999. Morphemes and spelling. In: Nues, T. (Ed.), Learning Toread: An Integrated View from Research and Practices. Kluwer Academic, London, pp. 15–42.
Carlisle, J.F., 1995. Morphological awareness and early reading achievement. In: Feldman, L.B. (Ed.), Morphological Aspects of Language Processing. Erlbaum, Hillsdale, NJ, pp. 189–209.
Carlisle, J.F., 2000. Awareness of the structure and meaning of morphologically complex words: impact on reading. Reading and Writing: An Interdisciplinary Journal 12, 169–190.
Carlisle, J.F., Lauren, A.K., 2006. Effects of word and morpheme familiarity on reading of derived words. Reading and Writing 20, 669–693.
Casalis, S., Louis-Alexandre, M.,-F., 2000. Morphological analysis, phonological analysis and learning to read French: a longitudinal study. Reading and Writing: An Interdisciplinary Journal 12, 303–335.
Chao, Y.R., 1947. Cantonese primer. Greenwood Press, New York.
Clay, M.M., 1998. By Different Paths to Common Outcomes. Stenhouse, York, Maine.
Cunningham, A.E., Perry, K.E., Stanovich, K.E., 2001. Converging evidence for the concept of orthographic processing. Reading and Writing: An Interdisciplinary Journal 14, 549–568.
Cunningham, A.E., Stanovich, K.E., 1997. Early reading acquisition and its relation to reading experience and ability ten years later. Developmental Psychology 33, 934–945.
Crowder, R.G., Wagner, R.K., 1992. The Psychology of Reading, second ed. Oxford University Press, New York.
Chow, B.W.-Y., McBride-Chang, C., Burgess, S., 2005. Phonological processing skills and early reading abilities in Hong Kong Chinese kindergartners learning to read English as a second language. Journal of Educational Psychology 97, 81–87.
Deacon, S.H., Wade-Woolley, L., Kirby, J., 2007. Crossover: the role of morphological awareness in French immersion children's reading. Development Psychology 43 (3), 732–746.
Deacon, S.H., Kirby, J.R., 2004. Morphological awareness: just 'more phonological'? The roles of morphological and phonological awareness in reading development. Applied Psycholinguistics 25 (2), 223–238. http://dx.doi.org/10.1017/S0142716404001110.
Downing, J., 1989. Cognitive clarity: a unifying and cross-cultural theory for language awareness phenomena in reading. In: Yaden, D.B., Templeton, S. (Eds.), Meta-Lingusitic Awareness and Beginning Literacy. Heinemann, Portsmouth, NH.
Ellis, N.C., 1998. Emergentism, connectionism, and language learning. Language Learning 48, 631–664.
Ehri, L.C., 1997. Learning to read and learning to spell are one and the same, almost. In: Perfetti, C.A., Rieben, L., Fayol, M. (Eds.), Learning to Spell: Research, Theory, and Practice across Languages. Erlbaum, Mahwah, NJ, pp. 237–269.
Ehri, L.C., Wilce, L.S., 1987. Does learning to spell help beginners learn to read words? Reading Research Quarterly 22 (1), 37–65.
Ferroli, L., Shanahan, T., 1987. Kindergarten writing: explaining the relationship to first-grade reading. In: Readence, J.E., Baldwin, R.S. (Eds.), Research in Literacy: Merging Perspectives. National Reading Conference, Rochester, NY, pp. 81–97.
Flores d'Arcais, G., 1994. Order of strokes writing as a cue for retrieval in reading Chinese characters. European Journal of Cognitive Psychology 6 (4), 337–355.
Foorman, B.R., 1994. Phonological and orthographic processing: separate but equal? In: Berninger, V.W. (Ed.), The Varieties of Orthographic Knowledge. I: Theoretical and Developmental Issues. Kluwer Academic, Dordrecht, pp. 321–357.
Frith, U., 1985. Beneath the surface of developmental dyslexia. In: Patterson, K.E., Marshall, J.C., Coltheart, M. (Eds.), Surface Dyslexia. Lawrence Erlbaum, London, pp. 301–330.
Gough, P.B., Tunmer, W.E., 1986. Decoding, reading and reading disability. Remedial and Special Education 7, 6–10.
Goswami, U., 1999. The relationship between phonological awareness and orthographic representation in different orthographies. In: Harris, M., Hatano, G. (Eds.), Learning to Read and Write: A Cross-Linguistic Perspective. Cambridge University Press, Cambridge, pp. 51–70.
Hauerwas, L.B., Walker, J., 2003. Spelling of inflected morphology in children with spelling deficits. Learning Disabilities Research and Practice 18, 25–35.
Ho, C.S.-H., Bryant, P., 1997. Learning to read Chinese beyond the logographic phase. Reading Research Quarterly 32, 276–289.
Ho, C.S.H., Bryant, P., 1997a. Development of phonological awareness of Chinese children in Hong Kong. Journal of Experimental Child Psychology 72, 130–153.
Ho, C.S.-H., Bryant, P., 1999. Different visual skills are important in learning to read English and Chinese. Educational and Child Psychology 16, 4–14.
Ho, C.S.-H., Chan, D.W.-O., Tsang, S.M., Lee, L.H., 2000. The Hong Kong Test of Specific Learning Disabilities in Reading and Writing (HKT-SpLD). Chinese University of Hong Kong and Education Department, HKSAR Government, Hong Kong, China.
Ho, C.S.-H., Lai, D.N.-C., 1999. Naming-speed deficits and phonological memory deficits in Chinese developmental dyslexia. Learning and Individual Differences 11, 173–186.
Ho, C.S.H., Yau, P.W.-Y., Au, A., 2003. Development of orthographic knowledge and its relationship with reading and spelling among Chinese kindergarten and primary school children. In: McBride-Chang, C., Chen, H.-C. (Eds.), Reading development in Chinese children Westport. Connecticut Prager Press, London, pp. 51–71.
Ho, C.S., Ng, T.-T., Ng, W.-K., 2003. A radical approach to reading development in Chinese: The role of semantic radicals and phonetic radicals. Journal of Literacy Research 35 (3), 849–878.
Huang, H.S., Hanley, J.R., 1995. Phonological awareness and visual skills in learning to read Chinese and English. Cognition 54 (1), 73–98.
Katzir, T., Kim, Y.S., Wolf, M., Kennedy, B., Lovett, M., Morris, 2006. The relationship of spelling recognition, RAN, and phonological awareness to reading skills in older poor readers and younger reading-matched controls. Reading and Writing 8, 845–872.
Kieffer, M.J., Lesaux, N.K., 2012. Direct and indirect roles of morphological awareness in the English reading comprehension of native English, Spanish, Filipino, and Vietnamese Speakers. Language Learning 62, 1170–1204.
Kirby, J.R., Deacon, S.H., Bowers, P.N., Izenberg, L., Wade-Woolley, L., Parrila, R., 2012. Children's morphological awareness and reading ability. Reading and Writing: An Interdisciplinary Journal 25 (2), 389–410. http://dx.doi.org/10.1007/s11145-010-9276-5.
Koda, K., 2000. Cross-linguistic variations in L2 morphological awareness. Applied Psycholinguistics 21, 279–320.
Kuo, L.-J., Anderson, R.C., 2006. Morphological awareness and learning to read: a cross-language perspective. Educational Psychologist 4 (3), 161–180.
Lau, B.S., 1999. 中華新詞典 (ZhongHua Xin Ci Dian). Zhonghua shu ju (Xianggang) you xian gong si, Hong Kong.
Law, Y.K., 1998. Relationships between word recognition, reading comprehension and writing performance of grade six students in Hong Kong. Curriculum Forum 7, 98–111.
Leong, C.K., 1996. What does accessing a morphemic script tell us about reading and reading disorders in an alphabetic script? Annals of Dyslexia 36, 82–102.
Leong, C.K., 1999. Phonological coding and children's spelling. Annals of Dyslexia 49, 195–220.
Leong, C.K., Hau, K.T., Cheng, P.W., Tan, L.H., 2005a. Exploring two-wave reciprocal structural relations among orthographic knowledge, phonological sensitivity, and reading and spelling of English words by Chinese students. Journal of Educational Psychology 4, 591–600.
Leong, C.K., Tan, L.H., Cheng, P.W., Hau, K.T., 2005b. Learning to read and spell English words by Chinese students. Scientific Studies of Reading 9, 63–84.
Leung, M.-T., Law, S.-P., Fung, S.-Y., 2004. Type and token frequencies of phonological units in Hong Kong Cantonese. Behavior Research Methods, Instruments, and Computers 36, 500–505.
Lewis, D.J., Windsor, J., 1996. Children's analysis of derivational suffix meanings. Journal of Speech and Hearing Research 39, 208–216.
Li, W.L., Anderson, R.C., Nagy, W., Zhang, H., 2002. Facets of metalinguistic awareness that contribute to Chinese literacy. In: Li, W., Gaffney, J.S., Packard, J.L. (Eds.), Chinese Children's Reading Acquisition: Theoretical and Pedagogical Issues. Kluwer Academic, Boston, pp. 87–106.
Lin, D., McBride-Chang, C., Aram, D., Levin, I., 2010. Mother-child joint writing in Chinese kindergarteners: metalinguistic awareness, maternal mediation and literacy acquisition. Journal of Research in Reading 34 (4), 1–17.
Martin, F., Pratt, C., Claydon, E., Morton, A., Binns, S., 2003. The development of orthographic and phonological strategies for the decoding of words in children. Journal of Research in Reading 26, 191–204.

Mattingly, I.G., 1972. Reading, the linguistic process and linguistic awareness. In: Kavanagh, J., Mattingly, I. (Eds.), Language by Ear and by Eye. MIT Press, Cambridge, MA, pp. 133–147.

Mason, J.M., Herman, P.A., Au, K.H., 1991. Children's developing knowledge of words. In: Flood, J., Jense, J.M., Lapp, D., Squire, J.R. (Eds.), Handbook of Research in Teaching the English Language Arts. Macmillan, New York, pp. 721–731.

McBride-Chang, C., Bialystok, E., Chong, K.K.Y., Li, Y.P., 2004. Levels of phonological awareness in three cultures. Journal of Experimental Child Psychology 89, 93–111.

McBride-Chang, C., Chow, B.W.-Y., Zhong, Y.-P., Burgess, S., Hayward, W., 2005. Chinese character acquisition and visual skills in two Chinese scripts. Reading and Writing: An Interdisciplinary Journal 18, 99–101, 134–147.

McBride-Chang, C., Ho, C.S.-H., 2000. Developmental issues in Chinese children's character acquisition. Journal of Educational Psychology 92, 50–55.

McBride-Chang, C., Kail, V.R., 2002. Cross-cultural similarities in the predictors of reading acquisition. Child Development 73 (5), 1392–1407.

McBride-Chang, C., Shu, H., Zhou, A.B., Chun, P.W., Wagner, R.K., 2003. Morphological awareness uniquely predicts young children's Chinese character recognition. Journal of Educational Psychology 95 (4), 743–751.

McBride-Chang, C., Tong, X., Shu, H., Wong, A.M.-Y., Leung, K., Tardif, T., 2008. Syllable, phoneme, and tone: psycholinguistic units in early Chinese and English word recognition. Scientific Studies of Reading 12, 171–194.

Nagy, W.E., Anderson, R.C., 1999. Metalinguistic awareness and literacy acquisition in different languages. In: Wagner, D.A., Venetzky, R.L., Street, B. (Eds.), Literacy: An International Handbook. Westivew Press, Boulder, CO, pp. 155–160.

Nagy, W.E., Anderson, R.C., 1984. The number words in printed school English. Reading Research Quarterly 19, 304–330.

Nagy, W.E., Berninger, V.W., Abbott, R.C., 2006. Contributions of morphology beyond phonology to literacy outcomes of upper elementary and middle-school students. Journal of Educational Psychology 98, 134–147.

Nunes, T., Bryant, P., Bindman, M., 1997. Morphological spelling strategies: developmental stages and processes. Developmental Psychology 33, 637–649.

Packard, J.L., 2000. The Morphology of Chinese: A Linguistic and Cognitive Approach. Cambridge University Press, New York.

Perfetti, C.A., 1997. The psycholinguistics of spelling and reading. In: Perfetti, C.A., Rieben, L., Fayol, M. (Eds.), Learning to Spell: Research, Theory, and Practice across Languages. Erlbaum, Mahwah, NJ, pp. 21–38.

Perfetti, C.A., Beck, I., Bell, L., Hughes, C., 1987. Phonemic knowledge and learning to read are reciprocal: a longitudinal study of first grade children. Merrill-Palmer Quarterly 33 (3), 283–319.

Perfetti, C.A., Liu, Y., Tan, L.H., 2005. The lexical constituency model: some implications of research on Chinese for general theories of reading. Psychological Review 12 (11), 43–59.

Pikulski, J.J., Cooper, J.D., 1997. Issues in Literacy Development. Retrieved 28.08.07. from: http://www.eduplace.com/rdg/res/literacy/.

Plaut, D.C., McCelland, J.L., Seidenberg, M.S., Patterson, K., 1996. Understanding normal and impaired word reading: computational principles in quasi-regular domains. Psychological Review 103 (1), 56–115.

Seidenberg, M.S., McCelland, J.L., 1989. A distributed, developmental model of word recognition and naming. Psychological Review 96 (4), 523–568.

Shu, H., Chen, X., Anderson, R.C., Wu, N., Xuan, Y., 2003. Properties of school Chinese: implications for learning to read. Child Development 74, 24–47.

Shu, H., McBride-Chang, C., Wu, S., Liu, H., 2006. Understanding Chinese developmental dyslexia: morphological awareness as a core cognitive construct. Journal of Educational Psychology 98, 122–133.

Siok, W.T., Fletcher, P., 2001. The role of phonological awareness and visual-orthographic skills in Chinese reading acquisition. Developmental Psychology 37 (6), 886–899.

Snow, C.E., Burns, M.S., Griffin, P., 1998. Preventing Reading Difficulties in Young Children. National Academy Press, Washington, DC.

Stanovich, K.E., 1986. Mathew effects in reading: some consequences of individual differences in the acquisition of literacy. Reading Research Quarterly 21, 360–407.

Sticht, T.G., James, J.H., 1984. Listening and reading. In: Pearson, D. (Ed.), Handbook of Research on Reading, vol. 1, pp. 293–317.

So, D., Siegel, L.S., 1997. Learning to read Chinese: semantic, syntactic, phonological and short-term memory skills in normally achieving and poor Chinese readers. Reading and Writing: An Interdisciplinary Journal 9, 1–21.

Thorndike, R.L., Hagen, E.P., Sattler, J.M., 1986. Stanford-Binet Intelligence Scale, fourth ed. Riverside, Chicago.

Treiman, R., 1993. Beginning to Spell: A Study of First-Grade Children. Oxford University Press, New York.

Tong, X., Tong, X., McBride-Chang, 2013. A tale of two writing systems: double dissociation and metalinguistic transfer between Chinese and English word reading among Hong Kong children. Journal of Learning Disabilities. http://dx.doi.org/10.1177/0022219413492854 first published on June 19, 2013.

Tong, X., McBride, C., Tong, X., Tune in to the tone: Lexical tone identification is associated with vocabulary and word recognition abilities. Language and Speech (in press).

Tong, X., McBride-Chang, C., Shu, H., Wong, A.M.Y., 2009. Morphological awareness, orthographic knowledge, and spelling errors: keys to understanding early Chinese literacy acquisition. Scientific Studies of Reading 13, 426–452.

Tong, X.L., McBride-Chang, C., Wong, A.M.Y., Shu, H., Reitsma, P., Rispens, J., 2011. Longitudinal predictors of very early Chinese literacy acquisition. Journal of Research in Reading 34 (3), 315–332.

Vandervelden, M.C., Siegel, L.S., 1997. Teaching phonological processing skills in early literacy: a developmental approach. Learning Disability Quarterly 20 (2), 63–84.

Walker, J., Hauerwas, L., 2006. Development of phonological, morphological, and orthographic knowledge in young spellers: The case of inflected verbs. Reading and Writing 8, 819–843.

Whitehurst, G.J., Lonigan, C.J., 1998. Child development and emergent literacy. Child Development 69, 848–872.

Wimmer, H., Mayringer, H., Landerl, K., 2000. The double-deficit hypothesis and difficulties in learning to read a regular orthography. Journal of Educational Psychology 92, 668–680.

Yang, J.F., McCandliss, B.D., Shu, H., Zevin, J.D., 2013. Simulating language-specific and language-general effects in a statistical learning model of Chinese reading. Journal of Memory and Language 61, 238–257.

Yeung, P.-S., Ho, C.S.-H., Chan, D.W., Chung, K.K.-H., 2012. Modeling the relationships between cognitive-linguistic skills and writing in Chinese among elementary grades students. Reading and Writing 1–27.

Zhou, X., Marslen-Wilson, W., 1999. The nature of sublexical processing in reading Chinese characters. Journal of Experimental Psychology: Learning, Memory, and Cognition 25, 1–19.

Zhuang, Z.Y., 2000. 香港小學課本用字規範 (XiangGang Xiao Xue Ke Ben Yong Zi Gui Fan). HKSAR: Joint Publishing.

Ziegler, J.C., Goswami, U., 2005. Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. Psychological Bulletin 131 (1), 3–29.

Chinese Revolutions: Twentieth Century

Jeffrey N Wasserstrom, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

China was once seen as having a population that was unusually prone to 'rebellion' but not susceptible to 'revolutions,' that is, upheavals capable of doing more than just changing who was on the throne. The twentieth century, however, saw a series of interconnected revolutions (e.g., one associated with the name Sun Yat-sen and others with Mao Zedong), which between them thoroughly transformed the country. These events fit in well with some standard ideas about revolution promulgated by social theorists primarily interested in and most knowledgeable about the Western and Russian examples, but in other regards call them into question.

The Chinese were once thought to be the least revolutionary of people, yet the period beginning in 1911 was one in which China was altered forever by a series of revolutionary struggles. These struggles need to be understood as both divisible (in that they broke out at different times over the past century and often affected only certain aspects of the social and cultural order) and parts of a single whole (since many Chinese political actors conceived of themselves as participating in a unified Revolution, the goal of which was to radically transform the country's internal working and its place in the world). The tension between thinking in terms of multiple lower-case 'revolutions' and a singular upper-case 'Revolution' is, in China's case, crucial to keep in mind and impossible to resolve without doing a disservice to the country's complicated twentieth-century history. The Chinese experience confirmed some conventional social scientific ideas about revolutions generated from the study of European cases (e.g., the idea that ineffective reforms often pave the way for revolutions), while calling into question others (e.g., the notion that villagers could never be key players in radical movements for change).

In the 1800s, a leading sinologist claimed that the Chinese were the 'most rebellious' but 'least revolutionary' of peoples and Karl Marx, along with other Western social theorists, argued that peasants would never be the driving force in radical movements for change (Meadows, 1856, p. 25; Feuer, 1959, pp. 319, 338). The twentieth century would prove both the sinologist and the theorists wrong – many times over. Nearly every decade of it saw a revolutionary event of one kind or another break out in China. In many of these, residents of the Chinese countryside, where until recently a great majority of the nation's population lived, played central roles.

Key Events

The first quarter of the twentieth century witnessed a series of interrelated insurrections that toppled the Qing Dynasty (1644–1911). Known collectively as the 1911 Revolution, these upheavals paved the way for the establishment of the Republic of China (ROC) on 1 January 1912, which was also when a revolutionary activist, Sun Yat-sen (1866–1925), became this new country's first President. A pathbreaking student-led mass struggle also took place in the century's first quarter. This was the May Fourth Movement of 1919, which began with a rowdy student demonstration in Beijing protesting the terms of the Treaty of Versailles that would cede to Japan parts of China previously under German control. The struggle gained momentum and support from other classes after police arrested some of the Beijing students and ended with the dismissal from office of three high-ranking officials who educated youths claimed were insufficiently patriotic. In addition, the period saw two revolutionary organizations formed: the Nationalist Party or Guomingdang (GMD), which was led initially by Sun and then after his death by Chiang Kai-shek (1887–1975); and the Chinese Communist Party (CCP) whose most important leaders have been Mao Zedong (1893–1976) and Deng Xiaoping (1904–97).

The second quarter of the century began with the GMD and CCP working together in a United Front (1924–27) orchestrated by Sun and his Soviet advisor Borodin. The alliance's goal was to defeat warlordism and imperialism and thus get the *Geming* (a Chinese term for revolution whose shades of meaning are discussed below) back on track. In 1927, however, Chiang turned on his erstwhile allies, who had just helped him reunify the country through the military campaigns and mass movements known as the Northern Expedition. He launched a White Terror against members of the CCP and left-wing rivals within the GMD. After 7 years, after persistent and bloody but never quite conclusive GMD extermination campaigns, most of the remaining members of the CCP, including both Mao and Deng, set out on their epic Long March (1934–35) to safety in isolated northern base areas. The two organizations, each tightly disciplined and Leninist in structure, allied again from 1937 to 1945 in a Second United Front against Japan. The 1925–50 period concluded, however, with a Civil War (1945–49). This ended with the GMD's retreat to Taiwan and Mao's proclamation, on 1 October 1949, of the establishment of a new People's Republic of China (PRC) – something that this son of a relatively well-to-do farmer was only able to do because his party had won so much support in China's countryside. The CCP then launched a series of radical initiatives, including a partially successful effort to gain adherence to a new Marriage Law (1950) that gave women equal rights in family matters and strove to minimize the power of patriarchal kin groups.

The century's third quarter began with Mao taking the lead in further campaigns, all of which were carried out in the name of The Revolution and some of which caused enormous misery. One of most notorious was the 'Anti-Rightist Campaign' (a Red Terror directed largely against intellectuals). The most consequential and disturbing was the 'Great Leap Forward' (a bizarre experiment in irrational utopianism that contributed to a famine of horrendous proportions, causing the deaths of tens of millions of people). Later in this quarter century, Mao, angered by his post-Great Leap marginalization within the CCP hierarchy, rallied young loyalists (Red Guards) to challenge the Party bureaucracy. Thus began the Cultural Revolution (1966–76). Mao's charge, soon echoed by the Gang of Four (a clique that included Jiang Qing, his wife), was that high officials had lost their revolutionary zeal and become corrupt. A revolution within the revolution was needed, they argued, to get the sacred *Geming* back on course.

The Cultural Revolution spiraled into a chaotic war of all against all that continued into the first year of the century's final quarter. This period also witnessed the abortive revolution of 1989, during which giant demonstrations took place in Beijing and scores of other cities. In the ROC (as Taiwan was rebaptized in 1949), a major breakthrough came with the lifting of a decades-old policy of martial law. This laid the groundwork for an East Asian variation of the 'Velvet Revolutions' and 'Handshake Transitions' of Central Europe: the nonviolent transfer of power from the GMD to an opposition party early in the year 2000.

By the end of the 1900s, so many aspects of Chinese politics and society had been transformed or challenged by insurgents that it now seems strange that the first edition of this Encyclopedia included an article on 'The Chinese Problem' but not 'The Chinese Revolution.' The rise of the CCP and the Cultural Revolution still lay in the future when that entry was written, but there were clear signs by then that the twentieth century would be viewed as a revolutionary one for China. One indication of just how revolutionary is that, by the 1990s, many Chinese were living lives that would have been unrecognizable to their great grandparents or even their parents. Moreover, the words 'China' and 'revolution' were by then inextricably linked in the minds of many people. A newly created and little known compound term early in the 1900s (one of the first important tracts to use it appeared in 1903), by the middle part of the century, it was one of the most important keywords of Chinese politics, with figures on both the right and the left denigrating the other as 'counterrevolutionaries,' who had abandoned the cause of the Geming. Its centrality only began to lessen in the 1990s and first years of the twenty-first century, as a focus on economic growth and national strength took priority, but even now it exerts a special force.

The linkage between 'China' and 'revolution' was particularly strong in the first decades of the PRC's existence. Under Mao and even under Deng, who took control in the late 1970s, the term *Geming* was invested with a holy patriotic meaning. And in the wake of the 1989 protests, which were suppressed with brutal force and triggered an ongoing legitimacy crisis, the authorities found it useful to play the revolutionary legacy card, insisting, for example, that the protesters who took to the streets in 1989 had created 'turmoil' that threatened the Revolution. Finally, in 1997, the government moved to ratchet down the potency of the term by claiming that henceforth writing 'counterrevolutionary' tracts or taking 'counterrevolutionary' actions would no longer be considered crimes – instead things previously classified as 'counterrevolutionary' would be denigrated by other words, such as 'subversive,' for example. This adjustment of the legal code, though, did not mean the end of all efforts to play the rhetorical card of *Geming*. In 1999, for example, when American bombs mistakenly killed three Chinese citizens working in the country's Belgrade Embassy, the victims were quickly dubbed 'revolutionary martyrs' and the official media linked these new deaths to those of patriotic participants in hallowed historic events of the 1910s–50s.

In addition, even though Westerners often view the words 'reform' and 'revolution' as contrastive, in China *gaige* (reform) was often presented by Deng, and has sometimes been presented by his successors, as a method for carrying forward the *Geming*. The policy of opening the PRC to international trade and moving away from a command economy that Deng and after him Jiang Zemin (1926–till date) and Hu Jintao (1942–till date) have followed since 1979 is typically presented in English language works as a pragmatic 'Reform Program' and a step away from state socialism. Its PRC proponents, however, describe it as an effort to 'build socialism with Chinese characteristics' and an effort to keep the Revolution from ossifying.

It is not just in the PRC that the words 'China' and 'revolution' have often gone together naturally. Most of the few Chinese names and faces recognized by foreigners are those of insurgents – from Mao to the anonymous Tank Man of 1989. And many ROC residents grew up learning history from schoolbooks that treated the story of the 1900s as an unending revolutionary struggle, which encountered a terrible setback in 1949 but did not die.

Categories and Definitions

This brings us to a question too rarely addressed: did China experience one Revolution or several? Many official PRC histories (as well as some Marxist ones in other languages and standard GMD accounts) refer to a single ongoing Revolution with many stages. Western social theorists (such as Theda Skocpol) and comparative historians (such as Crane Brinton) sometimes prefer to think of a single great event that began in 1911 and ended in 1949. They then fit this almost 40-year-long event into a Great Social Revolution model that is also used to interpret more compact phenomena such as the French upheavals of 1789–99. Many non-Marxist sinologists, meanwhile, speak of three relatively discrete revolutions: a Republican one in 1911, a National one in the 1920s, and a Communist one.

The Term '*Geming*'

Complicating these questions is the term *Geming*, the meaning of which at times overlaps with, and at others diverges from, its Western equivalents. *Geming* literally means 'stripping the mandate,' that is, the successful carrying out of an upheaval that demonstrates, through its ability to overturn a ruling house and establish a new dynasty, that heaven now sides with

a new regime. It was only from the 1890s on, in part because of new senses the characters had acquired in Japan, that *Geming* began to be linked to the creation of a new political system, not just completion of a dynastic cycle. *Geming*'s original linkage to dynastic cycles makes it curiously like its English and French counterparts, in which visions of return and of forward movement alike can be found. In 1789, the Janus-faced aspect of revolution was signaled by the use of classical goddesses to portray new political virtues. In the 1910s, it was signaled by anti-Qing leaders alternating between presenting themselves as avenging the legacy of the Ming (1368–1644), the ethnically Chinese ruling group the Manchus had displaced, and as struggling to create a new system.

If there are overlaps between the terms 'revolution' and *Geming*, however, there is also a difference: it is never clear whether the sense of the Chinese word is singular or plural, lower or upper case and whether a text is referring to The Revolution or one of many. This is more than merely a linguistic issue: to compare events, we need to know when they began and ended. With France, for example, there is a consensus that the revolution began in 1789 and lasted about a decade, although other supplemental revolutions occurred later (1830, 1848, and 1871). With China, there are more options, including the three alluded to above.

Problems with Standard Categorization Schemes

There are, however, limitations to each. The continuous Revolution vision focuses too tightly on the achievements of a particular organization. It distorts the shape of twentieth-century history by downplaying struggles not guided by the particular Leninist party – the CCP or the GMD.

The 1911–49 vision is problematic for use in comparison, even though comparativists champion it. Skocpol, for example, does not explain why, if the 'Great Social Revolutions' are essentially alike in structural terms, the Chinese one took so much longer to play out. Brinton's 'anatomy of revolution' model becomes problematic because 'terrors' occurred at different points in the Chinese *Geming*. In other revolutions, he claims, this 'fever' hit the patient just once, then broke.

A tripartite schema, finally, leaves too little room for radical developments that do not move in lockstep with the rise of governments; 1911, the late 1920s, and 1949 were key turning points in the political story but not necessarily in the social and cultural ones.

Overlapping Revolutions

A more satisfying approach divides China's recent past, for heuristic purposes, into four overlapping revolutions, each following its own timetable: one political, one cultural/intellectual, one diplomatic, and one socioeconomic. Let us consider each briefly.

Politics

The political revolution has the clearest starting date: 10 October 1911. GMD official historians present the events that followed the Wuchang Uprising of that day (really a military mutiny) as adhering to a master plan crafted by Sun Yat-sen and carried out by members of his Revolutionary Alliance, the organization that evolved into the Nationalist Party. But many participants had no contact with the Alliance and knew little of Sun or his ideas. Only some were committed to republican ideals. Others were motivated by distrust of the Qing 'barbarians,' anger at particular officials, or a desire to be on the winning side of a Mandate shift.

Still, the 1911 Revolution extinguished not just a dynasty but a system; hence it deserves to be thought of as politically revolutionary. The National Assembly and other republican institutions quickly proved ineffectual. But the complete failure of attempts by Yuan and later another warlord to found new dynasties demonstrated that the rules of political life had changed.

Cultural and Intellectual Change

Attempts by Chinese revolutionaries to transform traditional belief structures and patterns of behavior began after the Opium War (1839–42). The military defeat suffered then by China led some within the dominant scholar–official class to begin questioning longstanding assumptions about the inferiority of foreign cultures. The first generations of intellectuals to rethink these issues, while interested in specific things the West had to offer (*yong* or techniques), were seldom revolutionary in their approach to questions of fundamentals (*ti* or essences). Their belief in the superiority of traditional moral codes remained unshaken; they merely sought efficacious ways to combine Confucian *ti* with Western gunboats and other sorts of foreign *yong*.

The *ti/yong* distinction remained at the heart of the debate for decades, until the iconoclastic New Culture Movement (1915–23) led by intellectuals who had gone abroad, usually to Japan, or been exposed to foreign ideas at one of China's recently formed Western-style schools. They disagreed about many things but shared three convictions. China's weakness was due to the enduring power of entrenched 'Confucian' and 'feudal' beliefs (such as the veneration of age) and practices (such as arranged marriages). The Chinese should welcome the best the West had to offer: ideas and methods of inquiry included. And intellectuals needed to educate and mobilize others.

Much New Culture Movement energy was directed toward the publication of new periodicals, which were filled with articles introducing and often lauding Social Darwinism, Pragmatism, Anarchism, Marxism, and other imported ideologies. These periodicals also attempted to serve another purpose: their articles, written in a 'plain speech' as opposed to a 'classical' style, were supposed to inform and inspire the masses. This literary move to break down the barriers between intellectuals and nonintellectuals was reinforced by public speaking campaigns on issues ranging from medicine to international affairs.

An important new turn came with the May Fourth Movement, when student activists with ties to the New Culture Movement actively appealed to members of other classes to join them in protesting the terms of the Treaty of Versailles and the corrupt and autocratic domestic authorities who were doing too little, they claimed, to protect the nation. After sympathy strikes by workers and merchants and a widespread boycott of Japanese goods, the students achieved their main

domestic goals: the dismissal of the three officials they hated most and the release of all those arrested for protesting. But their inability to stop the Treaty of Versailles from taking effect illustrated the need for a diplomatic revolution.

Diplomatic Change

No diplomatic revolution could take place until the rise of a strong central government committed to overturning the unequal treaties forced upon the Qing by foreign powers. The Northern Expedition brought this about. After reunifying the nation, however, Chiang claimed that the spread of communism was more dangerous than the continuation of imperialism. Unlike some of his warlord predecessors, Chiang was an outspoken critic of the unequal treaties, but he insisted that China could only regain its place in the world once its house was in order.

World War II precipitated a sea change, as China's position as one of the *allies* gave its leader new leverage in pushing for an end to unequal relations with the West. And, in 1943, the allies renounced all claims to special privileges within Shanghai and other 'treaty port' cities. This decision and Japan's withdrawal in 1945 made China freer of foreign influence than it had been for a century.

CCP historians argue, however, that the diplomatic revolution was still not complete – and they have a point. The GMD regime remained economically and diplomatically dependent on the United States. If credit for starting the diplomatic revolution rightly goes to Chiang, it remains true that China did not fully regain its status as an independent country until after 1949.

Socioeconomic Change

The socioeconomic revolution also began long before the Communists came to power. Most notably, the late Qing and Republican (1912–49) eras saw dramatic shifts in the types of people in control of the government and witnessed the rise within the social hierarchy of merchants. But the development of new alignments near the top, although important, did not have much effect on the lives of the vast majority of Chinese who continued to reside in villages and work in small plots of land. For them, a socioeconomic revolution did not take place until the CCP's land reform campaigns. This socioeconomic revolution (like the others) did not hit all parts of the country at the same time but rather played itself out according to regionally distinctive timetables. By the early 1950s, however, it had affected most of China's villages.

Unfinished Revolutions

If China's twentieth century is envisioned as a time of multiple revolutions, we need to keep in mind that most were incomplete – and that struggles for change in other spheres were pursued sporadically but never fully carried through. One important unfinished *Geming*, linked to but not quite the same as any of those just described, was the revolution for women. Aspects of sexual politics changed throughout the 1900s, but the dream of complete equality promoted by revolutionaries in various decades never became a reality. And forward movement toward male–female equality has sometimes been followed by a period of backsliding; in post-Mao China, women retain many of the rights they were formally accorded by the New Marriage Law of 1950, for example, but many see the recent past as one in which the status of women has declined in some spheres at least. Much the same can be said for the democratic revolution: calls for democratization were heard often during each stage of the *Geming*, but oligarchy and one-party authoritarianism have tended to prevail – and still do prevail in the Chinese mainland.

Ironically, although China's twentieth century was filled with transformations, its final two decades were strikingly similar to its first three. In Taiwan, moves to institute a system of free elections recalled experiments made around 1911. On the mainland, there were May Fourth era parallels for the calls in 1989 for an end to corruption. And in the automatic weapon fire that accompanied the massacres of June 1989 (the famous one in Beijing and a less well-known one in Chengdu), there were echoes of killings of the 1920s and the 1930s that were justified as necessary efforts to maintain order and/or protect the *Geming*.

Lessons from the Chinese Case

China's twentieth-century experience suggests the need for social scientists to pay more attention in the future to several factors. One, already mentioned, is the revolutionary potential of villagers. Another is scale. Revolutions have often been treated as purely 'national' events. Recent work on China shows that local dynamics are often crucial – even in upheavals driven by nationalism. So, too, are transnational flows of people (key participants in 1911 were overseas Chinese), ideas (each twentieth-century revolutionary outburst was inflected by a mixture of domestic and imported theories and notions), and money (funds coming in from Hong Kong, then still a British Crown Colony, for example, helped keep the 1989 protests going). This suggests that, in future, models used for comparing revolutions will need to make more room not only for revolutions of widely varying lengths but also for revolutions that were influenced by local, national, and transnational impulses and actors.

However, China's twentieth-century history reinforces some important insights into the comparative and theoretical literature on revolutions. It shows that Alexis de Tocqueville was right to insist that nothing does more to inspire revolutionary activism than an inefficient authoritarian regime's introduction of a reform program that offers too little and comes too late – an insight whose significance clearly still has twenty-first century relevance, as Arab Spring struggles have shown. The 1911 Revolution (which was preceded by unsuccessful Qing reforms) and even, in a way, the mainland's abortive revolution of 1989 and the much more recent fall of the GMD in Taiwan illustrate the power of Tocqueville's insight. These cases also illustrates the value of stressing, as Marx and Skocpol among others have, the revolutionary potential of groups at the fringe of the governing elite who have rising expectations and are frustrated by being kept away from real power – although on the mainland, in contrast to Taiwan, the Leninist Party in control has so far proved effective at co-opting a rising middle

class that might have been expected to push for movement toward democracy. In addition, Chinese events of the third quarter of the twentieth century in particular underscored the validity of the claim made by both Tocqueville and Skocpol, again among others, that revolutions, while attempting to do other things, often end up strengthening the power of central governments. Finally, the Chinese case shows just how much truth there is in the tragic maxim that revolutions all too frequently and all too regrettably end up devouring their own children.

A Twenty-First-Century Perspective

Looking back at China's twentieth-century revolutions from this point early in the second decade of the new millennium, some final comments are in order. The opening years of the twenty-first century have seen, according to the Chinese government's own estimates, tens of thousands of protests, in which citizens have called attention to abuses of power and sought redress for their grievances. In some cases (usually when the focus is a local issue and only a single social group is involved), protesters have gotten important concessions from the authorities. In other cases, the struggles have been met with harsh repression. What there has not been so far, due to various factors, is any kind of sustained and organized revolutionary movement that poses a real threat to the political status quo.

It is not enough to say that recent events prove that the Chinese people are exceptionally 'rebellious' but not naturally 'revolutionary' as that nineteenth-century bromide was proved unfounded by twentieth-century events. The explanation needs to be sought elsewhere. An exhaustion with revolutions, which proved so costly in the past century, could be a factor. Efforts by the authorities to clamp down quickly on any form of organized opposition is surely also part of the explanation, as is the way in which recent economic growth has made many Chinese from different walks of life feel, more than they did in some previous periods, that they have distinctively different sets of interests and aspirations. Still, the potential for revolutionary activism in the future (albeit perhaps the quite distant future) cannot be discounted, as many of the grievances that have fueled past risings in China and other parts of the world, such as intense disgust with nepotism and corruption and others kinds of abuses of power at the top, continue to plague the People's Republic.

See also: Asia, Sociocultural Aspects: China; Communism, History of; East Asian Studies: Economics; East Asian Studies: Politics; Political Parties: Communist; Revolution, Theories of; Revolutions, Sociology of.

Bibliography

Anderson, Perry, January–February 2010. Two revolutions. New Left Review 61, 59–96.
Bergère, M.C., 1998. Sun Yat-sen (J. Lloyd, English Trans.). Stanford University Press, Stanford, CA.
Bianco, L., 1971. Origins of the Chinese Revolution, 1915–1949 (M. Bell, English Trans.). Stanford University Press, Stanford, CA.
Brinton, C., 1965. Anatomy of Revolution. Vintage, New York.
Dirlik, A., 1989. The Origins of Chinese Communism. Oxford University Press, New York.
Duara, P., 1988. Culture, Power and the State: Rural North China, 1900–1942. Stanford University Press, Stanford, CA.
Fairbank, J.K., Goldman, M., 1998. China: A New History. enlarged ed. Belknap Press of Harvard University Press, Cambridge, MA.
Feuer, L. (Ed.), 1959. Marx & Engels: Basic Writings on Politics and Philosophy. Anchor, New York, pp. 319, 338.
Fitzgerald, J., 1996. Awakening China: Politics, Culture, and Class in the Nationalist Revolution. Stanford University Press, Stanford, CA.
Friedman, E., Pickowicz, P.G., Selden, M., 1991. Chinese Village, Socialist State. Yale University Press, New Haven, CT.
Goldstone, J.A. (Ed.), 1994. Revolutions: Theoretical, Comparative, and Historical Studies, second ed. Harcourt, Fort Worth, CA.
Hartford, K., Goldstein, S. (Eds.), 1987. Single Sparks: China's Rural Revolutions, Sharpe. Armonk, New York.
Meadows, T.T., 1856. The Chinese and Their Rebellions. Smith, Elder & Company, London, p. 25.
Ono, K., 1989. Chinese Women in a Century of Revolution, 1850–1950 (J. Fogel, English Trans. Ed.). Stanford University Press, Stanford, CA.
Perry, E.J., 2012. Anyuan: Mining China's Revolutionary Tradition. University of California Press, Berkeley, CA.
Perry, E.J., Selden, M. (Eds.), 2010. Chinese Society: Change, Conflict and Resistance, second ed. Routledge, London.
Skocpol, T., 1979. States and Social Revolutions. Cambridge University Press, Cambridge, UK.
Spence, J.D., 1982. The Gate of Heavenly Peace. Penguin Books, New York.
Tang, T., 2000. Interpreting the revolution in China: macrohistory and micro-mechanisms. Modern China 26 (2), 205–238.
Wakeman Jr., F., 1975. The Fall of Imperial China. Free Press, New York.
Wang, Z., 1999. Women in the Chinese Enlightenment: Oral and Textual Histories. University of California Press, Berkeley, CA.
Wasserstrom, J.N., 2010. China in the 21st Century: What Everyone Needs to Know. Oxford, New York.
Wong, R.B., 1997. China Transformed: Historical Change and the Limits of the European Experience. Cornell University Press, Ithaca, NY.
Wright, M.C. (Ed.), 1968. China in Revolution: The First Phase, 1900–1913. Yale University Press, New Haven, CT.
Yu, H., 2011. China in Ten Words (A. Barrtrans, Trans.). Pantheon, New York.
Yue, D., Wakeman, C., 1985. To the Storm: The Odyssey of a Chinese Revolutionary Woman. University of California Press, Berkeley, CA.

Relevant Website

www.morningsun.org

Choice, Empowerment, and Social Work

Graeme Simpson, University of Wolverhampton, Wolverhampton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Choice will be examined through five perspectives incorporating an international perspective. First, agency examines the capacity for autonomous actions, in short, self-determination. Second, choice as empowerment examines the concept of an engagement with promoting and working toward change both on a personal and political level. Third, choice as consumption explores the development of the 'citizen consumer.' Fourth, choice as policy examines the extent to which choice has become the acceptable face of neoliberal politics often masking other economic agendas. Finally, choice will be considered as 'collective action' as an antidote to the individualization inherent in much contemporary policy.

Choice

'Choice' has been a feature of social work throughout its history. Choice, within the field of social work and social policy can be understood in many ways and this article will identify five possible ways of examining and understanding 'choice' as a concept. First, choice will be explored through the concept of agency, rooted in the notion that humans have the capacity for autonomous actions, which in given situations will allow for choices to be made – in short, self-determination. This is perhaps the most obvious understanding of choice within social work; however, social work is a complex profession, which draws its underpinning rationale from many sources; so, the second long-standing idea of choice is that of social work being engaged with promoting and working toward change both on a personal and political level; in short, social work seeks a 'better world.' Recently, these strands have found a unified voice under the broad concept of empowerment. Recently, in developed welfare regimes, choice has become increasingly embedded into policy, where those people who use services are being seen as 'consumers of service.' This is particularly evident in developed liberal democracies, and this article will draw mainly (but not exclusively) upon the United Kingdom as an example of this. Thus, the fourth area is that of choice as 'policy,' which itself is closely connected with the final area of examination, which is that choice has become an acceptable face of neoliberal politics in advanced welfare societies, often masking other economic agendas – for example, the 'austerity' agenda from 2009 onward. Finally, choice will be considered as 'collective action' as an antidote to the individualization inherent in much contemporary policy.

Choice as Agency

Those who have engaged in social work have always held that those they work with have a right to self-determination. In essence, it means that social workers do not impose interventions or support upon their service-users or clients, but rather that the service-user has engaged 'voluntarily' with the social worker. Such a principle is embedded within social work's current and also historical codes of ethics and conduct. It was part of Biestek's (1961) ethical principles that stated that client self-determination should be central to all social work practice. The notion of self-determination can be traced to European philosophical traditions, notably the eighteenth-century Enlightenment. In this context, individual self-determination was founded upon the idea that people were capable of organizing their own affairs and was instrumental in challenging (if not rejecting completely) previous ideas of 'God-given' laws. In other words, people have choice (agency) to direct, control, and organize their lives and through choice, people are self-determining 'free' beings.

Even as such a view was being promoted, the French philosopher Rousseau posed a challenge in that people are "born free, and everywhere in chains" (This is a standard English translation of the original French: "L'homme est né libre, et partout il est dans les fers.") (2007/1762: p. 3). Thus, whatever free choices people may have, they are also somehow constrained by their situation. In the contemporary debates around agency, this can be put slightly differently. Social workers understand that the people they work with frequently make choices constrained by their circumstances; likewise, they, as social workers, also have their professional choices constrained. Many social workers quickly realized that as social workers, their 'choices' are limited by a range of factors beyond their control, for example, funding, procedures, and other political decisions. Social workers also understand that many of their service-users lead lives that are at least in part, influenced – if not directly determined – by external forces. This is perhaps put into even sharper focus when considering social work at a global level and the extreme financial poverty of people in the underdeveloped South. External forces, or structures, are often seen as not only limiting choice (or agency) but also making it an almost redundant concept. So, for social work, and for that matter social policy, choice becomes not something that can be taken for granted, but something to be examined and explored, often within the polarization of the agency-structure debate, where agency equates to individual choice, and structure to external forces that people experience and that will shape their lives.

This debate will be familiar to all those who engage in the delivery of any form of social welfare and a potential resolution is to suggest, following the work of Lavin (2005), that people's actions have to be understood within the context of wider social relationships and this would include arguably abstract factors such as the economic and social structures of any given society.

Within social work, and social welfare, there remains, however, a strong belief in choice – or agency, and yet most social workers would not entirely ignore the question of circumstances or environment, even though many procedural and legal directives would ultimately locate the actions with the individual and not the structure. It is arguably not so straightforward, as the well-documented example of child maltreatment highlights. Even though most cases of child maltreatment occur among people who could be regarded as 'income poor,' attempts to address their poverty as part of the social workers' interventions are usually short lived as the abuse becomes defined as 'individual agency' – a conscious 'commission' or 'omission' of an action.

Social workers should then ask how much choice people actually have. If the condition that someone has to live in generates exceptionally high levels of stress, and that the people living with this stress lack the financial resources to alleviate it – do they maltreat their children 'by choice'? In other words, is this a set of actions and behavior that represents a meaningful choice? Or, are the actions, at least to some extent, determined by the conditions people live in? Within a wider study of social policy, individual agency has often been treated with caution. Callinicos (2004) argued that all too often, attempts to impute individual agency to actions on its own seek to explain social phenomena through the lens of individual action, and, by focusing upon the individual action one can explain the social phenomenon. In the example of child maltreatment, which is a social phenomenon, the almost exclusive focus upon the agent of maltreatment allows for the actual social phenomenon and the social conditions that create it to remain unaddressed (Sidebotham et al., 2002).

This is not to suggest that all actions of maltreatment have their origins with structure or external factors, since clearly this is not the case – but the example serves to highlight the 'agency-structure' debate. The other side of this debate and social work's long connection with 'structure' will now be examined.

Choice as 'Empowerment'

If individual choice and self-determination finds it ethical underpinning in the work of Biestek (1961), and is based upon individual action, the exhortation for social workers and other welfare professionals to engage with structural factors equally has its ethical justifications. The International Federation of Social Workers' (IFSW's) ethical principles include the following in Section Choice as a 'Consumer,' entitled 'Social Justice':

> Challenging unjust policies and practices – Social workers have a duty to bring to the attention of their employers, policy makers, politicians and the general public situations where resources are inadequate or where distribution of resources, policies and practices are oppressive, unfair or harmful. (IFSW, 2012, http://ifsw.org/policies/statement-of-ethical-principles/)

This sets the question of empowerment within a political context. The writings of Marx identified the concept of 'consciousness' in relation to class as being a necessary precursor to any form of social change. Accordingly, Marxists seek to engender class consciousness among the working class to build labor solidarity. Thus, people are empowered by becoming aware of the power structures under which they live. This has a wider application, and for social welfare. Paulo Friere's 1970s work, *Pedagogy of the Oppressed*, although based in Latin America, had considerable influence in identifying the role and purpose of social workers 'raising consciousness.' Central to this is the idea that those working with oppressed and marginalized groups have a legitimate role in raising awareness of the structural factors that bring about their powerlessness and through this they (as the oppressed) can develop strategies to empower themselves. In short, this is not about social workers and other welfare professionals 'empowering' people, but rather ensuring that their actions bring about the conditions whereby such groups can engage in collective action. To develop the theme from the earlier section, social work as empowerment is not merely about individual agency, but significantly it is about enabling collective agency. This was a feature of the radical social work movement of the 1970s (see Bailey and Brake, 1975), the antiracist movement in the 1980s (see Dominelli, 1988a,b), and antioppressive practice in the 1990s (see Dalrymple and Burke, 1995). There is an equally strong trend in social work especially that sees 'individual' empowerment as a legitimate goal, often, as indicated earlier, with the aim of empowering service-users to make 'informal choices.' Empowerment within social work seeks to combine the concept of 'personal' empowerment with 'political' empowerment drawing from the radical social work literature of the 1970s and 1980s (Humphries, 1996).

This in turn connects to the development of new social movements (O'Brian and Penna, 1998) and, closely allied to this was the rise of service-user movements, which can be seen as an example of 'collective agency' or collective 'choice.' In such cases, service-users recognize their own individual agency alongside structures such as the disabling nature of society, highlighted by Michael Oliver (1990) and others in the disability rights field. Another important service-user group has been built around people who are, or who have, experienced mental ill-health and developed campaigns to assume control over their own lives. These activities, similar to those run by people with disabilities, focused around the concept that service-users should control their own services, and emphasized self-advocacy and drawing upon the experiences and accounts of 'mental health survivors.' This has often been a reaction against overly medicalized interventions, and has led to social workers regarding service-users as experts by experience, turning many service-users from passive recipients into active participants in service delivery.

Within a global context, collective agency and social work as empowerment is reflected in much of the community action work that has been developed. Furthermore, the global nature of social work draws more heavily on the idea of social work as empowerment than social work in the developed North, where community work, although it exists, often takes second place to an individualized personal agenda. Cox and Pawar (2006), in their study of international social work, also link empowerment to long-standing elements of community development work, and later to 'capacity building.' They emphasize the commonality of people and the desire to

acquire control over aspects of their environment, although they urge caution with the concept, since 'empowerment' clearly means different things in a range of contexts and their conclusion is that empowerment incorporates individual and group strategies, underscored by conscientization and social development processes.

Thus, empowerment is the coming together of individuals united by their social relations. Service-user movements have had a significant impact upon many aspects of social work and have, through collective responses, developed important alternatives to a more overtly individualized approach, and indeed, this was one of the foundations of the radical social work movement (Langan, 2002) and progressive community work (see also Loney, 1983).

Within the United Kingdom, a key movement is anti-oppressive practice, which began as a model of social work practice that was premised upon bringing about structural, as well as individual, change and reflected in the work of writers such as Dominelli (2002) and Dalrymple and Burke (1995), which developed ideas originally formulated by radical social workers during the 1970s. The model, however, was arguably appropriated by the mainstream and has often in practice been reduced to a series of interpersonal skills, such as enabling service-users to make informed (individual) choices, treating people with respect, and informing them of their 'rights,' in short what Dominelli (2000: p. 12) describes as 'bureaucratized forms of empowerment.' All these actions are good in themselves and it could certainly be argued that not to do this would be 'disempowering' and not conducive to 'choice' but that is quite distinct from these actions being part of an empowerment agenda.

Social work as empowerment takes the focus away from individuals and the type of case work codified by Biestek (1961), which emphasized various forms of casework and counseling. It remains, however, firmly rooted in the concept of choice, since collective action and collective work are the exercising of a shared and voluntarily engaged choice and will. It reinforces the concept of individual agency by ensuring that this is connected to structures and lays bare the false dichotomy of 'structure' and 'agency.'

Nonetheless, empowerment can be, and arguably has been, appropriated to maintain the existing order (Dominelli, 2000) and never is this clearer than in the next section, which examines the increasing role of the service-user as the customer and questions of consumer choice, prevalent in the United Kingdom and other developed welfare systems (Simpson and Connor, 2011).

Choice as a 'Consumer'

The beginnings of the twenty-first century saw a significant shift in the provision of services, in that those who received them (clients, service-users) began to be referred as 'consumers' or 'customers.' It is possible, within the United Kingdom at least, to see how this developed out of both the language of empowerment used by social workers and other welfare professionals, and service-user groups and movements (Ferguson, 2008).

This can be seen to connect readily with the idea of choice and empowerment when considering the use of large-scale institutions, particularly for people who were experiencing mental ill-health and those with disabilities. The so-called one-size-fits-all models began to disappear and the intention was that these were replaced with bespoke services; the service-user, therefore, became active in determining their service, rather than the recipient of an existing large-scale service: the latter became known as resources-led and was replaced by the 'needs-led' model. Thus, service-users identify their needs and, in theory at least, are offered choices to meet them. In short, the service-user becomes a consumer of services, although such approaches are not without their critics (Harris, 2003; Clarke, 2005).

This links to another form of choice, that is, the choices made by the consumer or more simply put, the buyer of goods or services. This draws on the economic theory of rational choice and is an attempt to move the consumerist ideology into a welfare setting. The person driven by rational choice as a consumer would fully research the product they wanted to buy (for example, a washing machine), be fully aware of the range of functions they wanted, and then make their decision based upon what the economist would call 'utility' either material or monetary (Seifert, 2011). Often the consumer will forego some functions for a cheaper model; at times, they may buy the best model, irrespective of price, dependent upon their economic position. While choice for the consumer is limited to what is available, the person acting upon their rational choices would maximize their options – they would not buy the first washing machine they saw. The rational choices of consumers will subsequently influence the producers of products because it is unlikely that a product that no one is likely to want to buy will be produced. If the producer continues to ignore the levels of market demand, according to classical economic theory (see Smith, 1976 [1776]) they would soon go out of business.

The extent to which this works in the actual retail market can be debated, but there are certainly difficulties when trying to apply it to welfare services. Those who support this development, for example, Glennerster (2009), point to its benefits. First, they would argue that the self-interest of the consumer, the mantra of 'I know what I want' is a more reliable mechanism for allocation than the goodwill of the provider. Second, they suggest that such a model brings into welfare the rigors of the market and ultimately a 'profit motive.' This will then result in more efficient, which social workers might say creative, solutions. Finally, and central to the whole consumerist ideology, is the notion of 'consumer sovereignty' or more colloquially, 'the customer is always right.' Partnership between those who buy and sell goods and services, or as is the case with a marketized form of welfare, between providers and users of services, is, therefore, important, not for the sake of 'choice,' but so that the market can accurately reflect service-user needs.

From the point of view of the service-user, the consumerist model is built upon 'choice' and services led by the 'consumer' rather than professionals. This will mean that services become more efficient in responding to their needs since continued payments depend upon it. Crucially, however, proponents of the model argue that consumers (service-users) need to be educated about the implications of their choices. This is akin to arguing that buying an energy-efficient washing

machine will be better for the environment in the long-run and this should be a factor in deciding the purchase. Yet, it is not hard to see how this process of 'education,' far from being the 'empowering tool' that in its purest form it might be, enables people to make the 'right' choices. Critics of the market approach embrace a range of positions but the factor that unites them all is that markets fail and choice is constrained (Glyn, 2006).

In the retail world, most people accept this as a social and economic reality, but, as Harris (2003) argues, social welfare is not amenable to market discipline. One knows where one can afford to shop and one also knows that some businesses fail, so one may not be able to continue buying a particular product. It becomes more problematic when this is transferred to the world of welfare, since markets do want to take high levels of risk, so people with high-level needs who are already vulnerable will be increasingly excluded. Within a profit-driven system, the social risks, which state welfare attempts to manage through welfare provision, become private and market-oriented risks. The desire for profit means that those who are at higher risk – generally poor people, people in poor health, and people with a disability – will not be catered for by such provision as the costs of doing so would result in low levels of return, diminishing profit (Esping-Andersen et al., 2002). Furthermore, by focusing upon individual preferences wider social needs may go unmet. Some writers (see Ferguson, 2008) have also suggested that the implementation of the market will result in a potentially alienated workforce, and the last thing service-users want is a workforce that does not engage with either them or their needs.

Perhaps the most significant aspect, however, of the consumer choice debate revolves around resources. It is perhaps a truism to argue that choices are limited by resources: either the resources that the customer has, or the resources available to produce the goods or services. In welfare terms where the market applies, people with greater resources will be able to pay for and receive a better type of service, whereas those who lack resources will be dependent upon the residual services provided – confirming that those already marginalized are likely to have this confirmed through market provision and not addressed by it, since the market is not designed to deliver universal services – for the simple reason that this will not generate an adequate return. Thus, for there to be genuine choice for all 'consumers' of social care, a full range of options would have to be equally available, almost certainly funded by the state either directly or indirectly.

The next section examines the question of choice as social policy – especially the dominant policy trend within neoliberalism.

Choice as Neoliberal Policy

Within the United Kingdom and much of the developed world, there has been a move toward the adoption of neoliberal economic policies. These are characterized by low rates of taxation and low levels of state expenditure with the attendant withdrawal of the state from provision leading to a general reduction in welfare benefits and more crucially state-led welfare provision. To a greater or lesser degree, this trend has affected all developed countries – although the extent to which governments have embraced such changes varies according to the political choices they make. This shift in policy has been accompanied by a language of choice, which is a key concept within the language of neoliberalism and, as was discussed in the last section, seeks to borrow concepts from the world of business and retail and apply or (in some instances) impose them upon service provision and welfare (Le Grand, 2007).

At one level, there can be at least notional gains for service-users as a result of such policies and the case of 'direct payments' within UK social policy is an example. The service-users receive a sum of money with which they can buy and organize the care and services they want, as opposed to having to receive services that exist and may not be tailored for their needs. In this way, it could be argued that the service-user, as a consumer, is 'empowered.' Yet, even this is not a straightforward transaction, shaped by assessment and, crucially, limits upon funding. Any analysis of policy, however, needs to understand the underlying aims and to this end, it is important when considering the 'choice as policy' to analyze the aims and policy beneficiaries.

Recent trends, notably within the United Kingdom, have seen the language of 'choice' being used to introduce a whole series of 'reforms' that have been linked to the economic policies of neoliberals, and that have frequently resulted in policies designed to bring about reductions in services. In a study of people with a moderate to severe learning disability, Simpson and Price (2010) suggested that policies designed to improve social inclusion had often resulted in greater levels of social exclusion for this service-user group, particularly those whose needs were the most demanding of support. Their conclusion, echoing Scull's (1983) in relation to mental health, was that when social liberal policies meet liberal economic policies, the social liberal view is appropriated for a political end. Further, in his review of these developments, Stan Houston (2010) identified the provision of services being driven by the idea of the *homo economicus* – the economic rational actor of rational choice theory, making decisions based upon a distinctly economic calculation. More recently, in his review of policies relating to well-being, Simpson identified 'choice' as a dominant factor in determining services. He located this within a wider discussion of the different factors that contribute to a person's sense of well-being, where choice is but one of several other more important factors identified by Robert Schalock (2004), which gave greater emphasis to friendships and relationships, community engagement, and wider structural factors that impact upon people – yet, increasingly social work policy and practice privileges 'choice.'

Choice, therefore, has moved from being a social work value based upon 'self-determination' to the rationale underpinning policy shifts – shifts that do not promote actual choice but rather to 'semblance of choice' within increasingly limited options. In one sense, 'choice' becomes part of a dominant policy hegemony, the overarching principles that shape the ways in which services, social work, and social policies surrounding social welfare are discussed. To be able to 'choose' between schools, hospitals, or social care services is seen a progressive empowering and part of what Clarke (2005) argued was the making of the 'citizen

consumer.' The combination of the ideas around 'choice' discussed earlier – self-determination, empowerment, and consumer sovereignty – can all be combined in this series of dominant ideas that shape expectations, and by invoking choice as part of new policies, it acts as a convincing cloak to hide realities of reduced services.

'Choice' and Collective Action

In the previous section, the manner in which 'choice' is spoken about focuses upon the individual and thereby 'individualizes' choice. The development of this into a very clear political framework both rests upon and encourages this level of individualization. The final section will examine how individuals can 'choose' to engage in collective action, which offers a different perspective to the individualized hegemony, which is common to most discussions of 'choice.' To begin with, the concept of 'agency' will be reconsidered.

Agency is significant both for groups and individuals. Not only does agency allow for 'choice' but it also allows for the social movements of, for example, people with disabilities, to campaign for greater levels of choice and control over their lives. For people who are engaged in social work practice, or indeed any social welfare practice, they do so often at an individual level – as individuals and of course, with individuals. Simpson and Connor (2011) examined the question of policy and agency in some detail. They argued that individual practitioners needed to understand their own and their service-users' position within the wider remit of social relations and structures. When considering 'choice,' this cannot simply be located within an individualized account, standing outside of political and social factors. Callinicos (2004) has argued that such traditional models of choice result in explanations of social phenomena through the actions and motives of individuals. Thus, agency and, by extension, choice have to be understood within structures.

It has already been argued that 'choice' is limited by social position or economic standing – in short, the more money you have, the more choices you have. This would lead to potential and actual unequal distribution of resources. Within social work, the IFSW's ethical principles include the following:

> Social workers have a duty to bring to the attention of their employers, policy makers, politicians and the general public situations where resources are inadequate or where distribution of resources, policies and practices are oppressive, unfair or harmful. (2012, http://ifsw.org/policies/statement-of-ethical-principles/)

To do this alone is to risk failure and so such challenges need to be part of collective action. To engage in collective action is, in the first instance, an act of individual will – or choice. A clear example here is that of service-user movements. A Marxist approach would argue that action is both an individual and collective choice, and that often there is a desire to hide the inherent social relationship by recasting those who challenge the social order in collective ways as being 'heroes' or people who have made a 'special contribution or achievement.' In such ways, the actual order is not challenged and the individualized nature of choice persists.

Choice then provides not only potential opportunities for people but also specific challenges. Within an unequal structure, choice allows for the 'hope' of greater 'freedom' while actually concealing the realities of diminishing choices and resources. Those who are engaged in social work and social welfare need to understand the appropriation of choice into a particular set of social relationships, which maintain social, economic, and power differentials. If choice is to be fully embraced as 'empowerment' by and for people who lack these resources, then it needs to be within the concept collective choice, will, and action.

This not only has a 'local' application within established welfare regimes. One of the key challenges for social work internationally is to harness the potential of local communities in collective responses to the difficulties they face, including the nature and work of nongovernmental organizations. It could be argued that developed 'northern' regimes have premised much of their provision upon a particular concept of 'choice,' whereas cultural and social organization in the developing 'South' should be one that allows for indigenous people to have the 'choice' to develop their own welfare futures. Much international work is development within a specific 'Northern' tradition. This can involve the imposition of particular methods and expertise that have their origins elsewhere upon those who have a different world view. Such an expression of social work, set out in Cox and Pawar's (2006) text, is a vital and necessary antidote to the seemingly relentless spread of global capital in its search for new markets and resources.

Conclusion – What Type of Social Worker?

This article has explored the concept of choice within social work, welfare, and policy. In doing so, it has introduced different notions of how 'choice' can be understood within social work and policy. It has drawn mostly from developments within 'northern' welfare regimes, which in the decade 2010 are experiencing significant reductions in welfare and services, which impact upon those who have the least resources and therefore often the fewest choices.

There has been a focus upon some of the more traditional aspects of choice, many of which see choice as progressive but there has also been an exploration of challenges to and critiques of these notions. A concluding comment is that those who enter social work, and any other welfare profession, also have to make 'choices' and answer the question of what kind of worker they want to be. Ferguson's work *Reclaiming Social Work* sets out a new agenda for engaged practice that offers a response to this question within a social work context.

Simpson and Connor argue that this is not 'straightforward,' but, drawing upon the work of Ledwith in 2007, two broad 'alternatives' can be identified. One is the technician, who possesses a range of skills and knowledge, brought to bear on a problem in a rational technical manner, and who aims to 'improve' things for people within the existing frameworks. Arguably, technicians attempt to preserve professional autonomy and freedom, but perhaps do not question the nature of the organization (and ultimately the state) they are employed by. The contrasting position is that of the activist, who also believes that the skills and knowledge gained should

be used for improvement but they will not only deal with the symptoms of injustice but also seek to address the underlying structural causes of this. That then remains the final challenge of 'choice' outlined in this article.

See also: Marxist Approaches to Social Work; Social Work Theory.

Bibliography

Bailey, R., Brake, M. (Eds.), 1975. Radical Social Work. Edward Arnold, London.

Biestek, E.P., 1961. The Casework Relationship. Allen and Unwin, London.

Callinicos, A., 2004. Making History: Agency, Structure, and Change in Social Theory, second ed. Brill, London.

Clarke, J., 2005. New labour's citizens: activated, empowered, responsibilized, abandoned? Critical Social Policy 25 (4), 447–463.

Cox, D.R., Pawar, M.S., 2006. International Social Work: Issues, Strategies, and Programs. Sage Publications, London.

Dalrymple, J., Burke, B., 1995. Anti-oppressive Practice. Open University Press, Buckingham.

Dominelli, L., 1988a. Anti-racist Social Work: A Challenge for White Practitioners and Educators. Palgrave Macmillan, Basingstoke.

Dominelli, L., 1988b. Anti-racist Social Work, second ed., 1995; third ed., 2008. Macmillan, London.

Dominelli, L., 2000. Anti-oppressive practice in context. In: Adams, R., Dominelli, L., Payne, M. (Eds.), Social Work: Themes, Issues and Critical Debates. Palgrave, Basingstoke, pp. 3–19.

Dominelli, L., 2002. Anti-oppressive Social Work Theory and Practice. Palgrave Macmlllan, Basingstoke.

Esping-Andersen, G., Gallie, D., Hemerijck, A., Myles, J., 2002. Why We Need a New Welfare State. Oxford University Press, Oxford.

Ferguson, I., 2008. Reclaiming Social Work Challenging Neo-liberalism and Promoting Social Justice. Sage, London.

Freire, P., 1972. Pedagogy of the Oppressed. Penguin, Harmondsworth.

Glennerster, H., 2009. Understanding the Finance of Welfare: What Welfare Costs and How to Pay for It, second ed. Policy Press in Association with the Social Policy Association, Bristol.

Glyn, A., 2006. Capitalism Unleashed: Finance, Globalization, and Welfare. Oxford University Press, Oxford.

Harris, J., 2003. The Social Work Business. Routledge, London.

Houston, S., 2010. Beyond homo economicus: recognition, self-realization and social work. British Journal of Social Work 40, 841–857.

Humphries, B., 1996. Critical Perspectives on Empowerment. Venture Press, Birmingham.

International Federation of Social Workers, 2012. http://ifsw.org/policies/statement-of-ethical-principles/ (accessed 28.08.13.).

Langan, M., 2002. The Legacy of Radical Social Work. Macmillan, London.

Lavin, C., 2005. Postliberal agency in Marx's Brumaire. Rethinking agency in Marx's Brumaire. Rethinking Marxism 17 (3), 439–454.

Le Grand, J., 2007. The Other Invisible Hand: Delivering Public Services through Choice and Competition. Princeton University Press, Princeton, NJ.

Ledwith, M., 2007. Reclaiming the radical agenda: a critical approach to community development. Concept 17 (2), 8–12.

Loney, M., 1983. Community against Government: The British Community Development Project 1968–1978: A Study of Government Incompetence. Heinemann, London.

Oliver, M., 1990. The Politics of Disablement. Macmillan, Basingstoke.

O'Brian, M., Penna, S., 1998. Theorising Welfare: Enlightenment and Modern Society. Sage, London.

Rousseau, J.-J., 2007[1762]. Du Contrat Social ou Principles du Droit Politique. MetaLibri, Amsterdam. Available at: http://www.ibiblio.org/ml/libri/r/RousseauJJ_ContratSocial_p.pdf.

Schalock, R.A., 2004. The concept of quality of life: what we know and do not know. Keynote address. Journal of Intellectual Disability Research 48 (3), 203–216.

Scull, A.T., 1983. Decarceration: Community Treatment and the Individual – A Radical View. Rutgers University Press, New York, NY.

Seifert, R., June 2011. Neoliberalism, Public Choice Theory and the Reform of the NHS. SHAW Conference. Wolverhampton University, Staffordshire.

Sidebotham, P., Golding, J., Heron, J., 2002. Child maltreatment in the "Children of the Nineties:" deprivation, class, and social networks in a UK sample. Child Abuse and Neglect 26 (12), 1243–1259.

Simpson, G., Connor, S., 2011. Social Policy for Social Welfare Professionals: Tools for Understanding, Analysis and Engagement. The Policy Press, Bristol.

Simpson, G., Price, V., 2010. From inclusion to exclusion: some unintended consequences of "valuing people". British Journal of Learning Disabilities 38 (3), 180–187.

Smith, A., 1976 [1776]. An Inquiry into the Nature and Causes of the Wealth of Nations. Idion-Verlag, München.

Chomsky, Noam (1928–)

Robert Barsky, Vanderbilt University, Nashville, TN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article offers an overview of Noam Chomsky's contributions to linguistics, as well as the multiple fields that are impacted by his approach, including psychology, philosophy, computer sciences, communications, political studies, neuroscience, and behavioral sciences. The emphasis is placed upon his early formative years, the milieus from which he emerges and to which he contributes, and the general attitude that drives his work, right up to the current moment.

Avram Noam Chomsky was born in Philadelphia on 7 December 1928, to Dr William (Zev) Chomsky and to Elsie Simonofsky. William Chomsky, who came from Russia to the United States in 1913, was a Hebrew scholar that the *New York Times* referred to, in an article that appeared on 22 July 1977, as 'one of the world's foremost Hebrew grammarians'. Elsie Simonofsky taught Hebrew at Mikveh Israel, for which William Chomsky was the Principal.

Noam Chomsky grew up in Philadelphia with his parents and his younger brother David, who eventually studied medicine and became a doctor in Philadelphia. The whole family was actively involved in Jewish issues, notably the revival of the Hebrew language, Jewish cultural activities, and Zionism. Noam read Hebrew literature with his father, spent a considerable amount of time in Hebrew school, and later became a Hebrew teacher. Although his parents were 'normal Roosevelt Democrats', he had cousins and uncles who were part of a Jewish working class with ties to various strains of communism. The neighborhood in which Chomsky was raised consisted largely of Germans and Irish Catholics, who Noam recalls as being largely anti-Semitic and pro-Nazi.

From just prior to his second birthday, he was sent to the Oak Lane Country Day School, where he remained until the age of 12. The pedagogical approach taken at this Deweyite experimental progressive school remains for him an oft-recalled model, since it brought together students with different levels of talent, from various backgrounds, and fostered their creative faculties without a competitive system of evaluation. This approach placed a premium upon creativity, and since no one activity was put forth as the standard, each student was encouraged to believe in the value of his or her own activity, so that if competing at all, students were competing with themselves. This kind of thinking still underwrites both his political and linguistic writings, as well as his teaching, emphasizing stimulation of potential through noncoercive social structures. While at this school Noam began writing for the school newspaper, and his first piece appeared shortly after his tenth birthday, when he contributed an editorial on the fall of Barcelona. The impact of the Spanish Civil War is still present in many of his political articles, because this revolution makes manifest the idea that people can organize spontaneously, without a vanguard, to combat systems of oppression. The period leading up to this kind of action, though, required organization and a commitment education and militancy, ideas promoted by the New Left, for which Chomsky remains a leading figure.

He was from an early age a voracious reader in many fields. In terms of literature, for example, early influences included Tolstoy, Dostoevsky, Turgenev, Eliot, and Dickens, as well as the Bible (largely read in Hebrew), and Yiddish-Hebrew writers of the late nineteenth and early twentieth centuries. At the age of 12, he read the proofs for his father's book on David Kimche (1160–1236), a Hebrew grammarian working in the thirteenth century, during the so-called Golden Age of Jewish cultural creativity. This experience provided a philological background, although his eventual aim in his linguistic work steered clear of the search for meaning in favor of descriptive explanations, aimed at providing a theory of the synchronic structure of the language which would explain the distribution of phenomena. Insights concerning the value of forgotten learning and the importance of language studies are both reflected in Noam Chomsky's later work, in particular in books such as *Cartesian Linguistics*. Of the influence of his readings in Jewish texts, though, he recalls a special fondness for the Agadah, the legends of stories, rather than the Talmud, which he did study but which played no special role in his development.

At the age of 12, Chomsky entered the Central High School in Philadelphia, a traditional public school, where a heavy premium was placed upon academic success through imposed curriculum, hierarchies of knowledge and students, and a mainstream system of values that he described as impediments to independent and creative thinking. He was an active and well-liked student at Central High and belonged to numerous clubs and read voraciously. Nonetheless, intrigued by libertarian anarchist movements, and knowledgeable of events in Spain, he was by 1941 making regular trips by train by himself to New York to visit relatives and second-hand bookstores on Fourth Avenue. He also frequented the *Freie Arbeiter Stimme* office, the place from which was published an anarchist journal to which notable influences, such as Rudolf Rocker (a crucial interlocutor), contributed. He also spent a lot of time with a relative of his in New York, an uncle (on his mother's side), who ran a newsstand on 72nd Street that served as an informal literary political salon.

In 1945, Chomsky began undergraduate studies at the University of Pennsylvania. He continued to live at home, paying for his education by teaching Hebrew school in the afternoons, on Sundays, and sometimes in the evenings. His enthusiasm for university rapidly waned, however, when he found that the institutional structure he so loathed in high school was largely replicated at Penn. He did meet scholars

who were to be influential for him, including C. West Churchman in philosophy, and Giorgio Levi Della Vida, who taught Arabic, and in 1947, while still an undergraduate, he met Carol Doris Schatz (1930–2008), whom he'd first met at Hebrew School when he was seven. Noam and Carol Schatz were married in 1949, and in 1953 they stayed on an Israeli kibbutz. They had considered moving there, but found a high level of ideological conformity to Stalinist ideology. After 6 weeks in Israel, they returned to the US, and shortly after started a family, and so Carol interrupted her studies in order to raise the children. She later returned to complete her work in linguistics, and went on to a linguistic career of her own, eventually securing a position at Harvard's School of Education, where she specialized in language acquisition. She has also worked on language abilities in deaf-blind subjects in the Sensory Communication Group of the Research Laboratory of Electronics at the Massachusetts Institute of Technology, and has studied the invented spelling of children who began writing before learning how to read (published in 1979). She also became active in Noam's political work, helping to assemble materials and organize clippings for the huge database from which Chomsky draws for his writing.

In addition to meeting his lifelong partner in 1947, Chomsky also met Zellig Harris, who shared many of his interests and had overlapping interests with Noam's father on issues relating to the Hebrew language. Noam had met Zellig Harris in the Harris family home, since Zellig's father regularly held Jewish services in the basement to which he had on occasion invited the Chomsky family. Harris asked Chomsky to read the proofs of a manuscript called *Methods in Structural Linguistics*, but it was not Harris's linguistics that first attracted Chomsky, but rather his politics, and beginning in 1947 they developed a relationship that would be of great consequence to the field of linguistics. Harris also encouraged Chomsky to take graduate courses in philosophy, with Nelson Goodman, Morton White, and mathematics, with Nathan Fine, and he introduced him to psychoanalysis, and to the Sullivan–Horney–Rappaport school of psychoanalysis. The upshot of this connection to Harris was to lead Chomsky back to the unstructured, lively, and creative debate that had been characteristic of his early education and upon which he had thrived in the company of his uncle in New York. Harris shunned course requirements, formal relationships, and scholarly hierarchies in favor of informal gatherings, broad-based discussions, and intellectual exchanges in a range of fields.

Through the course of these meetings and his varied studies, Chomsky ultimately received his BA from the University of Pennsylvania, which turned around his interests in linguistics, philosophy, and logic. His BA honor's thesis, which set the stage for some of his later work and which is taken to be the first example of modern generative grammar, was completed in 1949. He then wrote a Master's thesis, completed in 1951, that was a revision of the BA thesis. It was further revised and then published in 1979 as *Morphophonemics of Modern Hebrew*. Chomsky described Harris's work during this period as a set of procedures for organizing texts, and his own early work in linguistics as trying to make Harris's methods work. His undergraduate thesis also applied some of Harris's ideas, aimed, most notably, at enriching the system of phrase structure rules by introducing complex categories with features that can percolate downward to the categories contained within them, expressing global dependencies not captured in a simple system of phrase structure rules. This was already an approach that deviated from the dominant features of contemporary structuralist linguistics, although Chomsky did receive encouragement from Israeli logician Yehoshua Bar-Hillel, a colleague of Carnap's who became a close friend to Chomsky beginning in 1951. Henry Hoenigswald, a linguistics professor at Penn who focused on historical linguistics and the Indic tradition, also read Chomsky's work during this period.

In 1953, Chomsky broke from the prevailing structuralist methods of the time by suggesting that linguists should abandon their quest for discovery procedures, and pursue the task of finding evaluation procedures, which he himself elaborated in a framework that came to be known as principles and parameters. This led Chomsky to a break from Harris beginning in the early 1950s and definitively when he abandoned the *Methods* program a few years later.

Much of what would later be considered Chomsky's major contribution to the field was undertaken early in his career and recorded in *Morphophonemics*, his 1956 paper with Halle and Lukoff, his contribution (unpublished) to the 1959 Texas conference (in terms of contemporary generative phonology) and the linguistic parts of *Logical Structure of Linguistic Theory* (LSLT) (in terms of syntax-semantics). And the advancements he and those who follow his approach have made since that time owe much to the radically divergent path that he chose to follow – even as an undergraduate (cf *Morphophonemics*). *Syntactic Structures*, published in 1957 (the year in which he was also a visiting professor at Columbia University, invited by the Yiddishist and early sociolinguist Uriel Weinreich). The aim of this work was to develop a grammar which is able to *generate* sentences, just as the speaker of a language is able to produce a virtually infinite number of sentences (*performance*) using the finite number of words and grammatical rules known to him (*competence*). Another crucial work at this time was his review of B.F. Skinner's 1957 book *Verbal Behaviour*. Skinner, the leading proponent of behaviorism, believed that human behavior, and in particular verbal behavior, can be explained and controlled by the same external processes (reinforcement, for example) as those employed to predict and control the behavior of animals. This in Chomsky's view denies a fundamental characteristic of human behavior, creativity, which allows even very young children to comprehend a great variety of utterances *even if they are hearing them for the very first time*. Furthermore, the application to language processes of behaviorist psychology terminology, such as 'stimulus', 'response', 'habit', 'conditioning', and 'reinforcement', was by Chomsky's reading so loose that it could cover virtually anything, and was thus devoid of empirical content.

Chomsky's viewpoint here is ostensibly situated within a *rationalist* perspective, the tenets of which look back to views developed during the seventeenth century, a connection he would make in his second linguistic study, *Cartesian Linguistics*. Therein, Chomsky traces the Cartesian viewpoint through the Enlightenment and the Romantic period, emphasizing its value in terms of understanding creative discourse. He ultimately dwells upon the work of Wilhelm von Humboldt whose work on linguistics and his postulations with regards to appropriate societal makeup correspond very strongly to Chomsky's own

approach. Humboldt's vision, taken up in various ways in other Enlightenment thinkers, is another kind of *leitmotif* in the work of Chomsky that appears, for example, in his various discussions concerning 'language and freedom'. In a lecture, delivered to the University Freedom and the Human Sciences Symposium on 8–9 January 1970, Chomsky reiterated the language–freedom relationship with regards to historical texts, notably works from the Enlightenment period. Citing Rousseau (especially his *Discourse on Inequality*, 1755), Kant, Descartes, Cordemoy, Simon Linguet, and Humboldt, Chomsky describes the Enlightenment vision of a society set up to encourage rather than stifle human potential. Humboldt is particularly important here because he makes the link between characteristic human traits, an appropriate social setting and the language that sets man apart from animals.

In 1958 Chomsky was promoted to Associate Professor at MIT, and in 1958–59, he was a National Science Foundation Fellow at the Institute for Advanced Study at Princeton which gave him the opportunity to revise LSLT, and to participate once again in the Texas Conferences on Problems of Linguistic Analysis in English in 1959. In the spring of 1959, he then began working on a project involving generative phonology, applying to the English language work that he had previously undertaken in Hebrew as *Morphophonemics of Modern Hebrew*. By the early 1960s, his work had gained him recognition in linguistics, philosophy, and psychology. This growing eminence was in part the result of his having begun a graduate program in linguistics and philosophy at MIT with like-minded colleagues (notably Morris Halle and the first Chair, Jay Keyser). The program attracted a number of gifted scholars during this period, including Robert Lees, Jerry Fodor, and Jerry Katz, graduates of the PhD program at Princeton and Paul Postal, who completed his PhD at Yale. All were eventually named to the MIT faculty; Postal in linguistics, Fodor and Katz in philosophy.

The early to mid-1960s was a time of enormous production; in 1962 he gave a paper, at the Ninth International Congress of Linguists, entitled 'The Logical Basis of Linguistic Theory', which outlined an approach to language known as 'transformational generative grammar'. In June of 1964 Chomsky also gave a series of lectures at the Linguistic Institute of the Linguistic Society of America, which were published in 1966 as *Topics in the Theory of Generative Grammar*. He published *Aspects of the Theory of Syntax* (1965), and *Cartesian Linguistics: A Chapter in the History of Rationalist Thought* (1966). He gave a set of lectures to a general audience, in Berkeley, in January of 1967, which were enlarged and published as *Language and Mind* in 1968 (enlarged with several later essays added, in 1972). And he completed *The Sound Pattern of English* with Halle in 1968. He wrote it during a period in which he was a Fellow of the American Council of Learned Societies with the assistance of the National Institutes of Health at Harvard University, the Centre for Cognitive Studies, and a grant from the Social Science Research Council, and he presented his findings prior to publication in the Christian Gauss seminars in Criticism at Princeton, a series of six weekly lectures.

In 1966, Paul Lauter contacted Chomsky (who at that time had been trying to organize national tax resistance with his friend Harold Tovish, a well-known sculptor) and asked him to team up with him to lend support to draft resistance. The result was a group called Resist, which eventually came to be involved in other forms of resistance to 'illegitimate authority', and which came to be associated with such people as Dwight Macdonald, Dr Benjamin M. Spock, and Reverend William Sloane Coffin. So while he was assembling the Beckman lectures for eventual publication as *Language and Mind* (1968), which traces historical developments concerning the study of language and mind with particular reference of rationalist philosophy, he was also working on a series of political articles. 'Responsibility of Intellectuals', his first (of numerous) articles published in the *New York Review of Books* (*NYRB*) created considerable interest because of the large readership of *NYRB* and the type of audience drawn to such a paper at that time. The contributions to NYRB brought to the fore Chomsky's public persona, his power as political renegade, and would, especially with the publication of his book called *American Power and the New Mandarins* set the stage for his being associated with the vocal dissenting left on the American scene.

By the early 1970s Chomsky, now in his early 1940s, continued to lecture, to receive honors, and to hone his linguistic works. He published *Studies on Semantics in Generative Grammar* in 1972, and received honorary Doctor of Humane Letters from five prominent universities. But at least outwardly, he was increasingly active on the political scene, giving the Russell Memorial Lectures in Cambridge in 1971, and publishing, *At War with Asia* in 1970, *Problems of Knowledge and Freedom* in 1971 (the Russell Memorial Lectures), and *For Reasons of State* in 1973. The basis of much of his political work during this period, aside from his interest in the relationship between the intellectual and the state, was the Vietnam War.

Within the field of linguistics, and specifically in Chomsky's department at MIT, the late 1960s marked a period of dissension and tension among faculty and students. An article called 'Former Chomsky Disciples Hurl Harsh Words at the Master', which appeared on 10 September 1972 in the *New York Times*, describes dissident activities within the department. This schism in the field generated significant discussion and 'linguistic wars' that came to be discussed quite publicly, and generated considerable work. At the same time, in the mid- to late 1970s publications such as *Reflections on Language* (1975) and *Essays on Form and Interpretation* (1977) appeared (as did LSLT, although it had been written in 1955–56), and studies that Chomsky had been working on for years entered the public domain.

Chomsky always lectured widely, and during this period alone he gave the Whidden Lectures at McMaster, the Huizinga Memorial Lectures at Leiden, the Woodbridge Lectures at Columbia, and the Kant Lectures at Stanford, and he was named as Member of the British Academy in 1974. But it was the *Reflections on Language* book that probably created the greatest amount of discussion within the field, at least in part because of John Searle's review article 'The Rules of the Language Game', that appeared in the 10 September 1976 *Times Literary Supplement*. In Searle's opinion, *Rules on Language* (*RL*) confirms a retreat from Chomsky's previous positions, in particular the idea that a sentence's meaning is determined by its syntactical deep structure. The book that Searle was discussing was comprised in part of the Whidden Lectures, which addresses the first of three problems that Chomsky has elaborated in various discussions; this problem, which Chomsky has called 'Plato's problem', that can be described in terms of the

question of why it is that humans, whose contact with the world are brief and limited can come to know so much. The second problem, more often taken up with regards to political rather than linguistic discussions, is what Chomsky called Orwell's problem, asking how is it that human beings know so little given the amount of information to which they have access? And the third, which he refers to as Descartes' problem, relating to mysteries for humans, categories of knowledge that seem to lie beyond epistemic bounds.

Chomsky's political work during this period continued to focus upon issues that he had talked about since the late 1960s and early 1970s, including the Middle East, the Spanish Civil War, the backgrounds to World War II, the framework of US global planning, global order, and Indochina. But he was paying special attention to the question of why, after almost a century of destruction and invasion that had been unusually heavily reported (as a result of new technologies), and after recent events which should have shaken American confidence in their institutions, the society did not rise up (an issue that was in his work related to discussions of historical and ideological reconstruction). On the contrary, the spirit of dissent of the late 1960s and early 1970s was, at least on the surface, becoming a spirit of resigned consensus. At this point, Chomsky began to collaborate with the next person in the milieu, Edward S. Herman, who had similar interests and who had also published books on the Vietnam War in 1966 and 1971. The Chomsky–Herman collaboration would ultimately lead to the publication of several articles and books, including *Counter-Revolutionary Violence* (1973), *The Political Economy of Human Rights* (1979), and *Manufacturing Consent* (1988, the title later adopted to the film *Manufacturing Consent: Noam Chomsky and the Media* by Achbar and Wintonick, and the companion book edited by Achbar). Texts published from this time forth, notably *Language and Responsibility* and *Radical Priorities*, also contain wide-ranging discussions covering the many disciplines that have interested Chomsky since his youth and that continue to be impacted by his work. Chomsky added substantially to his growing corpus during this period with the publication of *Human Rights and American Foreign Policy* (1978), *Language and Responsibility* (1979), and, once again with Edward S. Herman, *Political Economy of Human Rights* (1979).

Out of the sheer volume, vociferousness of the critique, and the radical underpinnings of his approach, Chomsky's political talks during this time stirred up considerable controversy from different sectors. His criticism of American policies in Vietnam and Cambodia had alienated him from establishment intellectuals and politicians, his condemnation of Cold War policies led to his continued ostracization by the proponents of America's 'free world' policies, and his ridiculing mainstream American's propaganda-instilled description of the 'free world' has led to significant efforts to question or ignore his work, which continues to this day. But this period was particularly fraught, marked as it was by two controversies, relating to Chomsky's having signed a petition on behalf of Robert Faurisson, that led to the 'Faurisson Affair', and his (along with Herman's) comparison of the relative number of innocent people killed by Pol Pot in Cambodia and by the Indonesians in East Timor, and the reporting thereon, that demonized Pol Pot and ignored East Timor. The East Timor example demonstrated the ways in which Western powers can ignore their own complicity while playing up the aggressions of the enemy, but this was construed as his being an apologist for Pol Pot's atrocities. He also added during this period a series of books including *Reflections on Language* (1975), *Essays on Form and Interpretation* (1977), *Rules and Representations* (1980), and *Modular Approaches to the Study of the Mind* (1980). These works include essays and lectures which expand and clarify previous work, address issues raised by his critics, update grammatical theory based upon recent work (by Chomsky and others), and all the while developing the thesis that language (and other mental abilities) should be seen as 'growing' in the individual's mind, under genetic control, rather than as 'learned' by organisms lacking any initial predisposition to develop the particular kinds of mental faculty which are eventually acquired.

Chomsky's linguistic work made important progress in the early 1980s leading him to what has been described as a 'new programme', the products of which are recorded in *Lectures on Government and Binding* (1981), *Knowledge of Language: Its Nature, Origin, and Use* (1988), *Barriers* (1988), and finally, in a more accessible form, in *Language and Problems of Knowledge: The Managua Lectures* (1988), which also includes some political discussion arising out of questions posed by the audience in Managua. More recently is *The Minimalist Program* (1995), the work that has taken shape around questions that came into focus in 1980 with the Principals and Parameters model. These texts offer significant advances in the field, offering insights into "empty categories and the principles that govern them, and that determine the nature of mental representations and computations in general, as well as the principles of phrase structure, binding theory, and other subsystems of universal grammar". Chomsky described these developments as part of a revolution in our understanding of the human mind, a leap forward that for him was akin to some of the scientific discoveries of the seventeenth century.

The political work continued to evolve on a tack consistent with principles previously described, but broadened to address a larger number of issues, including expanded work on media research (*Manufacturing Consent: The Political Economy of the Mass Media*, 1988, with Edward S. Herman, *Necessary Illusions*, 1989), Cold War, post–Cold War, and terrorist-style politics (*Towards a New Cold War. Essays on the Current Crisis and How We Got There*, 1982; *Pirates and Emperors: International Terrorism and the Real World*, 1986; *The Culture of Terrorism*, 1988; *Terrorizing the Neighbourhood: American Foreign Policy in the Post–Cold War Era*, 1991; *World Orders Old and New*, 1994), Israel (*The Fateful Triangle: The United States, Israel and the Palestinians*, 1983), Latin America (*Turning the Tide: U.S. Intervention in Central America and the Struggle for Peace*, 1985), Vietnam (*Rethinking Camelot; JFK, the Vietnam War and US Political Culture*, 1993), and imperialism (*Deterring Democracy*, 1992; *Year 501: The Conquest Continues*, 1993). Two popular 'readers' were also published during this period, *Language and Politics* (1988) and *The Chomsky Reader* (1987), two introductions to Chomsky's work were written by Carlos Otero, and collections of interviews such as *Chronicles of Dissent* (1992) and *Keeping the Rabble in Line* (1994) and, more recently, *Understanding Power: The Indispensable Chomsky* have given the reader access to interviews on wide-ranging subjects. This (incomplete) list is itself evidence that a synopsis of Chomsky's work over even

a relatively short period could but scratch the surface of a very deep corpus. Chomsky's many accomplishments earned him admiration from his peers, who rewarded him with such honors as: the Distinguished Scientific Contribution Award, American Psychological Association (1984), the Kyoto Prize in Basic Science, Inamori Foundation (1988), the Orwell Award, National Council of Teachers of English (1987 and 1989), he was made an honorary Member, Ges. Für Sprachwissenschaft, Germany (1990), and he became a William James Fellow, American Psychological Association (1990), and so on.

Whole generations of scholars have been trained by Noam Chomsky and as well by Carol Chomsky, and their own three children; their oldest child, Avi, is a Professor of Latin American history; their second, Diane, has for a long time lived in Managua; and Harry Chomsky works in the computer field on the West Coast. Carol Chomsky worked for 25 years at the Harvard Graduate School of Education, where she taught courses on language and linguistics; she eventually transferred to technology and education, with which she had been involved for years. She passed away in 2008. Chomsky's ideas continue to evolve, and to have had a strong influence on researchers in virtually every discipline in the humanities and social sciences, as well as in neuroscience, computer science, and emerging work that study the brain. As such, he has a very public persona, enhanced as well by notable debates over the years, with William F. Buckley, Alan Dershowitz, Michel Foucault, Christopher Hitchens, George Lakoff, Richard Perle, Jean Piaget, Hilary Putnam, Willard Quine, John Maynard Smith, and others, including Ali G. His recent honors include a 2005 honorary fellowship from the Literary and Historical Society; a 2007 Uppsala University Honorary Doctor's degree in commemoration of Carolus Linnaeus; the President's Medal from the Literary and Debating Society of the National University of Ireland; and a 2010 Erich Fromm Prize in Stuttgart, Germany. Also in 2010, Chomsky became the third scholar to receive the University of Wisconsin's A.E. Havens Center's Award for Lifetime Contribution to Critical Scholarship. He was voted the leading living public intellectual in The 2005 Global Intellectuals Poll; in a list compiled by the magazine New Statesman in 2006, he was voted seventh in the list of 'Heroes of our time'; and on 22 January 2010, a special honorary concert for Chomsky was given at Kresge Auditorium at MIT, featuring music by Edward Manukyan as well as honorific speeches from Chomsky's colleagues. In June 2011, he was awarded the Sydney Peace Prize and, also in 2011, he was inducted into IEEE Intelligent Systems' AI's Hall of Fame for his contributions to the fields of Artificial Intelligence and Intelligent Systems. In 2012, he played a deadpan cameo role in 'MIT Gangnam Style', a parody of the 'Gangnam Style' music video that became one of the most popular on YouTube. And most recently, in 2013, a newly described species of bee was named after Noam Chomsky: Megachile chomskyi.

See also: Bloomfield, Leonard (1887–1949); Boas, Franz (1858–1942); Capitalism; Colonization and Colonialism, History of; Communism, History of; Dewey, John (1859–1952); Enlightenment; Human Rights, History of; Imperialism, History of; Intellectual History; Jakobson, Roman (1896–1982); Labor History; Liberalism: Historical Aspects; National Liberation Movements; Nationalism, Historical Aspects of: South Asia; Piaget, Jean (1896–1980); Sapir, Edward (1884–1939); Skinner, Burrhus Frederick (1904–90); Social Inequality in History (Stratification and Classes); Social Movements, History of: General; Social Protest Movements and Social Work Practice; Socialism: Historical Aspects; Working Classes, History of; von Humboldt, Wilhelm (1767–1835).

Bibliography

Barsky, Robert F., 1997. Noam Chomsky: A Life of Dissent. MIT Press, Cambridge; London.
Barsky, Robert F., 2005. The Chomsky Effect: A Radical Works beyond the Ivory Tower. MIT Press, Cambridge; London.
Barsky, Robert F., 2011. Zellig Harris: From American Linguistics to Socialist Zionism. MIT Press, Cambridge; London.
Bloomfield, Leonard, 1928. Menomini Texts. S.E. Stechert, New York [printed in Germany].
Chomsky, William, 1947. Hebrew, the Story of a Living Language. Education Department, Zionist Organization of America, New York.
Chomsky, William, 1957. Hebrew, the Eternal Language. Jewish Publication Society of America, Philadelphia.
Harris, Zellig S., 1986/1951. Structural Linguistics. University of Chicago Press, London; Chicago.
von Humboldt, Wilhelm, 1969. (1767–1836). The Limits of State Action. Int. J.W. Burrow. Cambridge University Press, London.
Matthews, P.H., 1993. Grammatical Theory in the United States from Bloomfield to Chomsky. Cambridge University Press, Cambridge.
Otero, Carlos P. (Ed.), 1994. Noam Chomsky: Critical Assessments, vols. 4. Routledge, London.
Rocker, Rudolf, 1937. Nationalism and Culture. Freedom Press, London.
Skinner, B.F., 1971. Beyond Freedom and Dignity. Knopf, New York. Harmondsworth: Penguin, 1973.

Flimography

Chomsky & Cie Director: Olivier Azam (out in 2008).
Distorted Morality – America's War on Terror?, 2003. Director: John Junkerman.
Is the Man Who Is Tall Happy?: An Animated Conversation with Noam Chomsky, 2013. Michel Gondry.
Last Party 2000, 2001. Director: Rebecca Chaiklin and Donovan Leitch.
Lees, Robert, 1957. Review of Noam Chomsky's Syntactic structures. Language, 33.3.
Lyons, John, 1991. Chomsky. Wm. Collins, London [Fontana Modern Masters] third expanded ed. (1970, 1978).
Manufacturing Consent: Noam Chomsky and the Media, 1992. Director: Mark Achbar and Peter Wintonick.
Noam Chomsky: Rebel without a Pause (TV), 2003. Director: Will Pascoe.
On Power, Dissent and Racism: A Discussion with Noam Chomsky, 2004. Journalist: Nicolas Rossier; Producers: Eli Choukri, Baraka Productions.
Peace, Propaganda & the Promised Land, 2004. Directors: Sut Jhally and Bathsheba Ratzkoff.
Power and Terror: Noam Chomsky in Our Times, 2002. Director: John Junkerman.

Christianity in Asia

Michael Northcott, Edinburgh University, Edinburgh, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 1761–1764, © 2001, Elsevier Ltd.

Abstract

There are approximately 85 million Christians in Asia but more than 40% of them live in just two countries, the Philippines and South Korea. In the postcolonial era, a strong element of Christian social militancy was observable in many Asian countries. Postcolonial Asian Christian social protest has focused on various concerns including the continuing social power of caste in India, workers' rights in South Korea and Southeast Asia, and democratic reform of authoritarian political regimes which characterize government in much of the region. A more recent form of militancy in Asian Christianity is associated with the rise of charismatic/Pentecostal and fundamentalist/evangelical groups. Asian Christians now find their claimed freedoms to worship and evangelize occasions under increased social scrutiny and resistance among majority religious and cultural groups and their political leaders

Christianity in Asia has an ancient history though it remains a minority religion. There are approximately 85 million Christians in Asia but more than 40% of them live in just two countries, the Philippines and South Korea. The minority status of Christianity in Asia gives it a distinctive sociological character in relation to Christianity in other regions of the world. For many Asian theologians this distinctive sociological context has meant that the particular focus of their articulation of the Christ of Asia is that of dialogue with the dominant religious and cultural traditions of Asia. In the postcolonial era a strong element of Christian social militancy was observable in many Asian countries. This militancy was associated with the ecumenical Christian Council of Asia, and especially its sponsorship of Urban Rural Mission in a number of countries. The Catholic Federation of Asian Bishops Conference is also an organization which has confronted frequently social and political controversies in particular Asian countries. Postcolonial Asian Christian social protest has focused on various concerns including the continuing social power of caste in India, workers' rights in South Korea and Southeast Asia, and democratic reform of authoritarian political regimes which characterise government in much of the region. A more recent form of militancy in Asian Christianity is associated with the rise of charismatic/Pentecostal and fundamentalist/evangelical groups which have arisen within and beyond the denominational boundaries and institutions established during the colonial era. Adherents of the new charismatic and evangelical forms of Asian church regard Christian worship, spirituality, and personal evangelism, rather than social transformation in the wider society, as the essential vocation of Asian Christians. This more transcendental and salvific emphasis represents a clear rejection of the social protest of earlier forms of postcolonial Asian Christianity. The new militancy is more focused on claims to spiritual power for the new forms of Asian Christian spiritual and congregational life, and the rights of all Asian Christians to freedom of worship, and to proselytise their neighbors. Just as the earlier ecumenical and socially radical militancy was the occasion for formal State resistance in South Korea in the 1970s, and in Singapore, Malaysia, and Hong Kong in the 1980s, so Asian Christians now find their claimed freedoms to worship and evangelize occasion increased social scrutiny and resistance amongst majority religious and cultural groups and their political leaders.

South Asia

There is evidence of Christian activity in South India and Sri Lanka from the sixth century. Kerala is the region where most activity was concentrated, and Christians are still strong there, comprising 20% of the population. Christians also comprise a significant proportion of the population of Tamil Nadu. Portuguese colonisers saw a connection between trade and religion and, with the aid of the Jesuit order, encouraged the majority of the population of their settlement in Goa to convert to Christianity by the end of the sixteenth century. Under British rule the East India Company resisted the deployment of missionaries, judging that their presence would exacerbate existing religious tensions between Hindus and Muslims. But in the nineteenth century missionaries were allowed into most of British India. Converts were few and most were from marginal tribal groups, or from untouchables who have much to gain by abandoning Hinduism for Christianity. Since the 1980s a vocal Dalit (untouchable) movement has emerged which is radically anticaste and draws on Latin American-style liberation theology for its inspiration. With the rise of religious politics in India since the Premiership of Mrs Gandhi, religious conflict has become endemic in many parts of India and this has had an increasing impact on Christians. In Sri Lanka Christians are most prominent in the West and make up 30% of Colombo's population.

Christianity in China

Christianity in China goes back to the migration of Nestorian Christians from Iraq to China along the trade routes in the seventh century CE. The Jesuits arrived in China in the sixteenth century under the leadership of Matteo Ricci who experimented with a radical indigenisation of Christianity to Chinese culture. His efforts at cultural translation were controversial in Rome and the Pope insisted on a reversion to Roman rituals and

traditions in 1744. After the opium wars and the Nanking treaty of 1842, Catholic and Protestant missionaries settled in many parts of China. The association between Christianity and Imperialism was strong and conversions were few until the failure of the Boxer rebellion at the end of the nineteenth century when many, especially young, Chinese turned to Christianity as a potential source of social and spiritual restoration. With the advent of Communist government in 1949 all foreign missionaries were removed from China and churches closed. Ironically the exclusion of foreign missionaries allowed Chinese Christianity to flourish in a way it had not before and the Christian presence in China today is much greater than it was at the time of the Communist revolution. Most of the officially recognized churches in China, whether Catholic or Protestant in origin, are known as three-self churches, which means that they are not dependent on extra-Chinese ecclesiastical authority, nor do they rely on funds from overseas. There are also a large number of unofficial churches and congregations in China whose leaders try to circumvent controls on officially recognized churches, and in particular controls on proselytism. Evangelistic and worship meetings are held in secret, in apartments, or in open spaces away from city or town centres. The mass migration of many tens of millions of rural migrants into towns and cities in China since the 1980s, the largest contemporary movement of people on the planet, has occasioned conversions to new religions, including Christianity, as migrants seek to find a substitute for village-folk religion. Churches in urban areas, and especially in southern China, are currently growing very fast, fuelled by the spiritual and ritual vacuum occasioned by the Cultural Revolution and the enforced break-up of traditional Chinese religions, and by intellectual and popular dissatisfaction with secular Communist ideology.

Hong Kong, Taiwan, Japan, and Korea

One fifth of the population of Hong Kong are Christian and just under half are Catholics. The principal Protestant and Catholic traditions are represented and there are also a large number of independent and Pentecostal churches which have become established since the 1970s. The majority of Christians in Taiwan are Presbyterian, reflecting a missionary drive in the nineteenth century. Many new Christian groups emerged after the emigration of supporters of the government of Chian Kai-sheck to Taiwan from mainland China. The Jesuit missionary Francis Xavier introduced Christianity to Japan in 1549, after his initial mission to India who converted the Japanese court to Christianity. More than 300 000 Japanese followed their rulers and converted by the end of that century. However, Christianity was later proscribed, though small pockets of Christian activity remained. In the twentieth century Japan has been characterized by remarkable religious innovation, and many New Religious Movements have emerged, particularly since the Second World War. Parallel developments have occurred in Japanese Christianity. New indigenous styles of Christianity have emerged, and Pentecostal groups have also proliferated.

The Catholic Church has roots in Korea as far back as the sixteenth century but Protestantism has taken root in Korea in a manner it has nowhere else in Asia. Forty-one percent of the population are now Christian, and the majority belong to indigenous Korean and Protestant churches. The beginning of the twentieth century saw a particularly rapid influx of Koreans into the Protestant churches, and many Protestant churches have also experienced considerable growth since the Second World War. Seoul has some of the largest church buildings, and the largest congregations, anywhere in the world. Explanations for the strength of Protestantism in Korea are various. One theory is that there is a uniquely individualist strain to Korean Confucianism and that Protestantism individualism, with its more progressive conception of the role of the individual in the fast changing world of the twentieth century, proved particularly attractive to Koreans who were dissatisfied with Confucianism.

Southeast Asia

As a consequence of more than three centuries of settlement by the Spanish, and after the Spanish American war, by the United States, Christianity is the principal religion of the Philippines. Virtually the whole country was converted to Roman Catholicism under Spanish rule, though a Muslim minority remained, primarily in Mindanao. The Roman Catholic Church, however, was organized and run exclusively by foreign clergy and resentment at this resulted in an internal revolution whose outcome was the Philippine Independent Church which broke away from Catholicism after 1860. Protestant missionaries arrived with the Americans, including Episcopalians who set out to convert the mountain peoples in North Luzon as they had not been reached by the Catholics.

Roman Catholic missions were active in Vietnam, Cambodia, and Laos in the seventeenth and eighteenth centuries, and established churches in all three countries. Cambodia now has very few Christians indeed. After Year Zero Christians were totally purged from Cambodia. In Laos the Catholic Church has had most converts amongst tribal peoples on the border with Thailand but remains an insignificant presence elsewhere in the country. In Vietnam Christians, the great majority of whom are Catholic, constitute 7% of the population. The church was indigenized much more effectively in the North, whereas in the South local leadership was discouraged and there are far fewer adherents. Catholics have been close to political leadership in both North and South Vietnam and the Catholic Church in Rome and in Vietnam was a strong supporter of moves for peace during the war, and of subsequent North–South reconciliation efforts after the war.

Uniquely in Southeast Asia, Thailand has never been colonized by a foreign power and Thai Buddhism is particularly resistant to foreign, including Christian, influence. Strong Chinese communities in the large cities, and especially in Bangkok, are the main locus of Christian adherents. Myanmar Buddhists are also highly resistant to Christianity. Christians are most numerous amongst the Karen tribal group in the forests of the Northwest who are not Buddhist, and do not regard themselves as Burmese.

Christianity was introduced into the Indonesian archipelago by the Dutch in the sixteenth century. Christians are most numerous in Northern Sumatra where the non-Muslim Batak people converted to Christianity in large numbers. There are also

churches in most areas of Java, and on some of the other islands. Indonesian Christianity has been subject to the State religious doctrine of Panchasilla which requires that all religious groups express loyalty to the state and tolerance of other religions. Indonesian Christians were active in the struggle for independence from the Dutch after the Second World War and their partnership in the birth of the new independent nation may partly explain why, almost uniquely in the Muslim world, there have in the past been a number of religious conversions. However, a peaceful inter religious climate characterized by neighborliness and dialogue between the religions recently has turned to conflict as the Suharto and Habibi governments have stirred up religious conflict as a way of maintaining a hold on the votes of the predominantly Muslim population.

Christians in Malaysia number around 7% of the population, and 15% in Singapore. Dating back to the Portugese settlement of Malacca in the fifteenth century, Malaysian Christianity has taken up a range of influences including Portugese Catholic, Dutch Presbyterian, American Methodist, Anglican, and Pentecostal. Christians are confined to Chinese and Indian minorities on the whole, though tribal peoples in East Malaysia have converted to Christianity in large numbers. The fastest growing forms of Christianity in both Singapore and Malaysia today are Pentecostal or Charismatic. This modern style of Christianity has spawned many independent churches and has also been very influential in mainstream Catholic and Protestant congregations. Both countries have experienced a surge in religious interest and affiliation in recent years, sparked in part by Islamic resurgence among the Malays. This is reflected in a range of religious innovations both within Islam, Buddhism, and Hinduism, as well as in Christianity.

See also: Catholicism; Evangelical, Revivalist, and Pentecostal Religions; Religion, Sociology of; Religion: Evolution and Development; Religions of East Asia.

Bibliography

Ackerman, S.E., Lee, R., 1988. Heaven in Transition. University of Hawaii Press, Honolulu, HI.

Barrett, D.B. (Ed.), 1982. World Christian Encyclopedia: A Comparative Study of Churches and Religions in the Modern World AD 1900–2000. Oxford University Press, Nairobi, Zimbabwe.

Digan, P., 1984. Churches in Contestation: Asian Christian Social Protest. Orbis Books, Maryknoll, NY.

von der Mehden, F.R., 1986. Religion and Modernization in Southeast Asia. Syracuse University Press, Syracuse, NY.

Neill, S.C., 1985. History of Christianity in India, 1707–1858. Cambridge University Press, Cambridge, UK.

Palmer, S.J., 1967. Korea and Christianity: The Problem of Identification with Tradition. Hollym Corp., Seoul, South Korea.

Christianity in Latin America

Gustavo Benavides, Villanova University, Villanova, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In its Roman Catholic form, Christianity has been present in Latin America for about 500 years. During the colonial period, Christianity fulfilled a double role: that of validating the social order, beginning with the legitimization of the conquest itself, and also providing the justification for judging that same order. Religions' double function could also be seen at work among the indigenous population, for whom Christianity served as a vehicle of accommodation as well as protest. After independence, Latin American governments have generally regarded the Roman Church as an ideological ally, one which played a crucial role in the legitimization of the social order. But the privileged status of Catholicism has been challenged, first by the generalized, male anticlericalism typical of Catholic countries; then by positivism and the theology of liberation; and finally by the spread of Protestantism. After the breakup of Catholic hegemony and the rapid growth of evangelical churches there was growth and fragmentation; individualism but also the constitution of new communities; political quietism side by side with right wing activism. There are even churches that have reversed the traditional direction of missionary activity and have established branches outside Latin America.

Christianity has been present in Latin America for about 500 years. That period may seem brief if one compares it with almost two millennia of European Christianity, but the crusades that led to the Christianization of the Baltic countries did not end till the fourteenth century, less than two centuries before equally bellicose Christians arrived in the Western hemisphere. If one compares the present situation of Latin American and European Christianity, one finds that Latin America is in some ways more solidly Christian than Europe. The hegemony of Roman Catholicism is a thing of the past, and there is a low degree of religiosity in urban areas much as in Western Europe, especially in the Southern part of the hemisphere. Still, the religion that remains in Latin America, both public and private, is overwhelmingly Christian. The degree of religiosity may be weaker in the southern cone, urban groups everywhere may no longer engage in the ritual activities that were common just a few decades ago, and the influence of the Roman Church on the everyday life of Latin Americans may have diminished substantially, especially in matters regarding sexuality and reproduction; but contemporary Latin Americans are not likely to articulate political grievances in sectarian religious terms. It is true that the political power of conservative evangelicals has increased in recent decades, yet despite differences in terms of their social base, the evangelicals' attitudes to traditional sexual morality does not differ from that of conservative Catholic groups such as the Opus Dei. Oppressed, marginal, or disaffected groups may coalesce along social or ethnic lines (which tend to coincide in the region) forming violent gangs, but they are not likely to form gangs of Catholics or Evangelicals, which would parallel the Muslim youth groups in Western Europe; nor is society likely to suffer cleavages based on religious allegiances as is the case in the Middle East and South Asia. On the contrary, Latin Americans of, for example, East Asian descent, are as likely to be Christian or quasi-Christian, or no-longer-Christian as their fellow citizens. In what follows, we will examine the processes that led to the present situation.

Religion before the Conquest

As is generally the case in the preindustrial world – in which ideological claims are advanced in mythological terms and reinforced through ritual, especially sacrificial practices – in the area now known as Latin America, the social order was religiously legitimized. A proper discussion of the various ways in which this religious legitimization took place would require consideration of the range of societies found in Meso and South America, from kinship-based groups found in the Amazonian area, to chiefdoms, to states. It may suffice to say that societies organized as chiefdoms tend to be theocratic, their political life being heavily ritualized, and this form applies to early states, such as the Inca and the Aztec, in which the rulers were sacralized and considered the pivots of their society. Thus, in the Andean world the Inca dynasty, originally based in the area of Cuzco in Southeastern Peru, claimed descent from Inti, the Sun. In the Maya polities on the other hand, rulers engaged in elaborate ritual blood-letting to propitiate their gods.

More generally, religious legitimization meant that social distinctions, obligation to work, and access to resources were regarded as being regulated by an order that encompassed the social and the physical world. In order to preserve this order it was necessary to establish the links between this world and that of the gods through ritual, usually through sacrifice. Religion also structured time and space, as well as the relations between age groups and the sexes. The organization of time was accomplished at several levels: at the individual level this involved rites of passage, while at the macrosocial level the structuring required the use of liturgical calendars which were related to the seasons and, above all, to agricultural work. Similarly, space was structured by establishing mythological distinctions between neighborhoods and, at a more extended level, by dividing a territory by means of shrines along imaginary lines, an example of which is the *ceque* system centered in Cuzco, the Inca capital.

Conquest, Resistance, Accommodation

The fact that the political, religious, and economic spheres were less differentiated in the Aztec and Inca states than in sixteenth-century Western Europe may have contributed to the speed of the victory of the Spanish and Portuguese adventurers. The 'modernity' of the conquerors should not be exaggerated, since

 http://dx.doi.org/10.1016/B978-0-08-097086-8.84056-8

the processes that would lead to the scientific and cultural transformations of the seventeenth century took place mainly in Northern Europe, bypassing the Iberian Peninsula. Second, the Iberian discovery and conquest of America took place immediately after the defeat of the Muslim kingdom of Granada and the expulsion of the Jews who had refused to convert to Christianity, which means that the conquest itself took place as a kind of crusade at a time of Spanish politico-religious triumphalism. Moreover, attention should be paid to the fact that Columbus seems to have endowed his enterprise with a millenarian aura, hoping that a portion of the wealth found in the Indies would be used in the conquest of Jerusalem. This millenarian component can also be found at work in the self-understanding of some of the members of the religious orders involved in the process of conversion of the vanquished.

In any event, while the crusading spirit validated the conquest, the transcendentalization of religion and the concomitant separation between the roles of priests and those of soldiers and administrators, allowed some clerics to denounce the abuses of conquerors without, however, condemning the conquest itself. During the colonization of Latin America, therefore, Christianity fulfilled the double role traditionally accomplished by religion: that of validating or indeed sanctifying a social order, while also providing the transcendental justification for judging and condemning specific instances of abuse.

Religions' double function can also be seen at work among the indigenous population, for whom Christianity, which had been resisted at first – to the extent that many indigenous people killed themselves after witnessing the destruction their societies – eventually functioned as a vehicle of accommodation as well as of protest. The first aspect involved what is generally known as syncretism, that is, the amalgamation of Christianity and indigenous religions. Thus, in the Andes, the Christian god incorporated elements of Andean divinities such as Wirakocha and Pachacamac, while in Mexico Quetzalcoatl was assimilated into the apostle Thomas. The same identifications were made regarding goddesses: in Yucatan, Ixchel, the goddess of childbirth, was identified with the Virgin Mary; while in Mexico, the dark-skinned Virgin of Guadalupe, subsumed the Aztec goddess Tonantzin, 'Our Mother.' The fact that the physical appearance of Mary, – second only to Jesus in the Catholic pantheon – was said to resemble that of the Indian population rather than that of the conquerors, indicates on the one hand the extent to which at least part of the subjugated population had assimilated the new religion, and on the other, the intimate connection between religion and ethnic identification.

In some cases, conquered elites sought the protection of some of the most bellicose inhabitants of the Christian pantheon, expecting to profit from the power these supernatural beings had shown on the Spanish side; thus, the apostle Santiago, patron of Spain, became syncretized with the Andean Illapa (Lightning). And as early as the sixteenth century the Virgin of Copacabana was enlisted in the intradynastic struggles of native elites. Christianity also played a role in the indigenous rebellions against Spanish domination. This can be seen in the sixteenth-century Taki Onqoy movement, and more clearly in the late eighteenth-century Tupac Amaru rebellion, in which the rebels regarded themselves as better Christians than the Spaniards.

The adoption and mobilization of the symbolic resources of Christianity on the part of the conquered populations took place in informal and formal ways. The former involved the recognition that the supernatural resources of the conquerors were as effective as their military ones, leading to the adoption of both. In this context it was the ritual, especially the sacrificial emphasis of indigenous religions, that facilitated the recognition and then the assimilation of the ritual practices and sacrificial doctrines of Roman Catholicism. In fact, in the Mayan arena, in which sacrifice and self-sacrifice were practiced, Christian missionaries had to be careful with the display of images of the crucified Christ, for these encouraged actual crucifixions among the new Christians. The ritual appeal of the new religion was systematically reinforced by the labor of the missionaries, who carried out their tasks in different ways, seeking not just to gain new converts, but to advance the interests of their respective religious orders. In their effort to gain converts the Franciscans tended to pay little attention to doctrinal matters, preferring to baptize a large number of people in massive ceremonies, a practice that was criticized by the Dominicans and Augustinians, who proceeded in a more deliberate manner, emphasizing the doctrinal component of their task.

The systematic indoctrination of the new Christians was carried out differently according to their social position. Ordinary people would be introduced to the catechism and the rudiments of Christian doctrine through the sermons preached and the confessions heard by a small number of itinerant priests. These indoctrination efforts had to contend with the radical theological differences – whether the theologies were explicit or implicit – between Christians and non-Christians. Non-Christians did not have a conception of the personification of pure evil like the Christian devil, and they also lacked of belief in a unified soul. Even more important was the novelty of the Christian concern with belief and – notwithstanding the populous nature of the Christian pantheon – exclusive allegiance to one god. These theological presuppositions contrasted sharply with the ritual character of the indigenous religions as well as with the additive nature of the indigenous pantheons, which absorbed the gods of conquered peoples. To counteract these difficulties, the church concentrated its efforts on the male children of the conquered elites, who were subjected to systematic instruction in special institutions. Under normal circumstances such Christianization would have had a considerable impact on the population at large. The growth of Christianity was further spurred by the demographic collapses that occurred when plagues decimated the population that accompanied the conquest, which meant that the number of Christians increased as the population decreased, a process that parallels the spread of Christianity in the Mediterranean world more than a millennium earlier.

In addition to converting people to Christianity, the missionaries engaged in seeking the abolition of the old practices. This involved the physical destruction of shrines and temples, whose foundations or even walls served in some cases as the foundations of Christian churches (as can be seen to this day in the Church of Santo Domingo, built upon the Coricancha temple in Cuzco). No less important was the systematic destruction of religious documents and of mummies, which played a central ritual and political role in the Andean

world. The "extirpation of idolatries" campaigns, which lasted into the seventeenth century, in many cases achieved their goal by torturing people suspected of celebrating rituals centered around mummies. The most notorious was the burning alive in 1539 of the leader of Texcoco, Don Carlos Chichimecatecuhtli Ometochtzin, ordered by the 'Protector of the Indians,' the Franciscan Juan de Zumárraga. Another atrocity was the inquisitorial process carried out, once again by Franciscans, under the command of Diego de Landa in the Yucatecan town of Maní in 1562, which resulted in the torture and killing of several thousands of people.

Religion during the Colonial Period

The three centuries of Spanish and Portuguese control saw the continuation of the Christianization of the continent, a process that involved a great degree of ambiguity, as much among the indigenous groups as among the population of European origin. Among the former, the ritual system of Catholicism became part and parcel of everyday life, with sacrificial practices such as the offerings to the *pachamana* – the Andean mother earth – becoming integrated into Catholic rituals, and with the Catholic saints playing the role of the old divinities, much as it had happened in Europe centuries earlier. In rural areas, but even more in the cities, daily life was punctuated by the rituals of baroque Catholicism. At the individual and family level, there were the sacraments, which marked transitions in a person's life. At the level of the society at large, the rhythms of daily life were marked by religious ceremonies, whether these were processions, the public torture and executions organized by the Inquisition, or the solemn masses that celebrated the births and marriages of the royal families in the metropolis. For those of European descent the participation in the ritual life of the church was taken for granted, although in some cases such participation involved Jews who, despite prohibitions, had made their way to the New World pretending to be Christian. It must be remembered that regardless of religious allegiance, the relatively small number of people of European origin, the small size of the towns, the prevalence of gossip, and, not least, the vigilant attitude of the Inquisition, made nonparticipation in Catholic rituals risky. The same applied to the practice of blasphemy, which, widespread throughout the Christian world, had reached virtuosic levels in Spain and its colonies.

Ambivalence toward Catholicism was not unknown. That ambivalence gave place to hostility toward members of the clergy, on the rare occasions when clergy condemned the abuses heaped on indigenous peoples and African slaves. But the condemnations themselves were halfhearted insofar as the legitimacy of the colonial enterprise itself remained above critique. Likewise, despite the occasional clerical condemnation of the violence against slaves of African origin, priests had no difficulty finding numerous passages in the Old and New Testament that justified slavery. For instance, the seventeenth-century Jesuit António Vieira resorted to the Exodus story to show that "God does not complain about the Pharaoh, but rather about the overseers," meaning by this that whereas the abuses of the slave masters should be punished, the institution of slavery itself was divinely ordained. Important in terms of the intimate connection between theology and the social order was the fact that violence was regarded by Vieira and other priests as the ultimately negligible price that the slaves had to pay for the privilege of baptism, without which they were assured of eternal damnation.

The fear of damnation, this time on the part of the elites, led to the church becoming the beneficiary of donations, whose purpose it was to pay for the celebration of masses, which according to Catholic belief would shorten the time that the donor would spend in purgatory. These donations made the church the largest landowner in Spanish America, particularly in the cities, in which the churches, towering above the other buildings, punctuated space, just as the rituals choreographed time. As is the case in preindustrial societies, the theology, institutional weight, and the ritual apparatus of the militant Catholicism of the Counter Reformation, served to prop up the colonial system as one of the bastions of Spanish and Portuguese power during the three centuries of colonial rule. To maintain the Catholic monopoly – and to make sure that the colonies traded only with Spain – colonial authorities sought to protect Catholic orthodoxy from external threats such as Protestantism and the intellectual currents of the Enlightenment.

Religion in the New Republics

The role played by religion in the wars of independence of the early nineteenth century was contradictory. The lower clergy tended to side with those who wanted to achieve independence from Spain, whereas the higher clergy, many of whom had been born in the Iberian peninsula, were generally in favor of maintaining the colonial regime. For example, Miguel Hidalgo, the priest of Dolores who played an important role in the struggle for independence in Mexico, was executed in 1811 after the failure of the rebellion he instigated. On the other side, we find bishops such as the Spaniard Manuel Abad, who not only opposed independence for Mexico but excommunicated his former friend Hidalgo.

Once independence was achieved in the nineteenth century, the Creole elites kept in place the church's prerogatives, while at the same time seeking to control it by continuing the exercise of the right of patronage, just as the kings of Spain had done. In the case of legislation approved by the Congress of Colombia in 1824, this involved nominating bishops and archbishops, granting or refusing the publication of papal bulls and briefs, and enacting legislation to ensure the maintenance of the discipline of the church. After the governments of the new republics and the Vatican signed concordats regulating the status of the church, Roman Catholicism generally enjoyed a privileged position enshrined in the republics' constitutions. In the case of Ecuador, for instance, the concordat signed in 1862 recognized that the "Roman Catholic Apostolic Religion will continue to be the only religion of the Republic of Ecuador." Furthermore, Ecuador also recognized, unlike Colombia, the right of bishops to communicate freely with the Vatican and to make public papal documents without any government interference.

The situation was not always as favorable as clerical groups or proclerical governments would have wanted. As in Spain and elsewhere in the Catholic world, the hegemony of the Church contributed to an atmosphere of anticlericalism which

led to attempts to curtail the Church's privileges. This anticlerical attitude was intensified by the influence of liberal and positivist currents, which became widespread among intellectual elites in the nineteenth and early twentieth centuries. Anticlericalism was furthered by the realization by some intellectual and political elites that the Protestant United States had become a powerful country, while the new Catholic republics remained poor and backward. The most radical measures against Roman Catholicism were taken in Mexico, a country that reestablished relations with the Vatican only in 1992. In general, however, most Latin American governments have regarded the church as an ideological ally, one which played a crucial role in the legitimization of the social order.

Religion and Everyday Life

Catholicism has been an omnipresent reality in the life of Latin Americans, yet Catholicism has been lived differently by men and by women, by members of upper and lower classes, and in urban and rural milieus. As in southern Europe, in Latin America everyday religion in its private and public forms has been the domain of women, and, except for special occasions such as Lent and Christmas, women have constituted the majority of the faithful in Sunday services. But as an example of the ambivalence of popular attitudes towards religion in general and priests in particular, women who spent an inordinate amount of time in churches or hanging around priests were usually derided as *beatas*, (sanctimonious women), while men who did the same were mocked as *beatos*, and had their masculinity questioned. It is the mothers who traditionally have been in charge of introducing children to the rituals and the moral injunctions of Catholicism, more technical matters being left to priests. Social class also plays an important role in the process of religious socialization and children of the middle and upper classes are generally subject to religious indoctrination in schools directed by members of religious orders; whereas children of the lower classes, who attend state schools, are subject to minimal religious education.

One of the consequences of this religious indoctrination based on class distinctions is that Catholic revival movements such as Catholic Action, or political parties linked to Catholicism, such as Christian Democracy, involve mainly middle and upper class persons. In large cities, public religious activity is usually restricted to Sunday masses and to occasional processions, often including the participation of government dignitaries. The legitimizing role of Catholicism as the state religion can be seen at work in the solemn liturgies commemorating the independence days of independent South American States, and in the consecration of entire countries to Jesus or the Virgin Mary.

Popular Religion

The counterpart of official Catholicism is represented by what is known as 'popular religion' – those ritual practices that seem to elude and even oppose the control of religious elites. Ambivalence is at the core of popular religion, since in order to express opposition it is necessary to make use of the symbolic resources of the church. Popular religion is characterized by a complex system of assimilation and rejection, through which subordinate groups negotiate and struggle over access to cultural and material goods. Latin American popular religion comprises the syncretic processes referred to above, in which elements of the symbolic world of the Spanish conquerors were incorporated into that of the Andean, Mesoamerican and other conquered groups. It also comprises contemporary ritual practices in urban and rural milieus, such as processions and pilgrimages, which parallel the rituals choreographed by the official church and seek to establish – through religious fraternities (*cofradías*) – a measure of lay control over religious activities. Despite academic infatuation with symbolic resistance, the parodic and carnivalesque rituals of popular religion generally have not constituted serious threats to the hegemony of Catholicism or of the State. Along the same lines, it can be said that the expenditures required by the ritual calendar of indigenous communities may increase economic disparities instead of leveling them.

Millenarian Movements

Popular religious practices may lead to insurrections, particularly when the state attempt to suppress religious activities or charismatic leaders in a violent manner. This happened in Northeastern Brazil from 1893 to 1897, in the Millenarian movement of Canudos, and also in Southern Brazil from 1912 to 1916 in the Contestado rebellion. Both the Canudos and Contestado movements constituted reactions against the secular, positivistic values of the Brazilian republic – and in the case of the Contestado movement it is clear that the rebellion was triggered by the dislocations brought about by capitalism. Like Millenarian and Messianic movements in general, those that have taken place in Latin America have involved marginal regions in transitional periods and have been led by marginal, literate individuals, such as Antonio Conselheiro – the leader of Canudos – and Miguel Lucena Boaventura, known as José Maria, leader of Contestado.

Liberation Theology

At the elite level, the privileged status of Catholicism in Latin America has not gone unchallenged. In the nineteenth century, anticlerical groups influenced by positivism decried the Church's influence; and in the twentieth century, politically progressive groups opposed the alliance between oligarchic governments and the conservative church. During the early 1960s, prompted first by the triumph of the Cuban revolution and then by the Second Vatican Council, the traditional equation between church and conservative politics was broken. These events led some members of the Latin American clergy to question the role the church had played in the maintenance of an unjust social order. That questioning spurred the emergence of a 'prophetic' understanding of Christianity known as the Theology of Liberation. Liberation theologians were influenced by the social sciences as much as by traditional methods of theological exegesis and they removed concepts such as sin and salvation from the mere personal and spiritual realms to the structures of the societies within which sins were committed and salvation was sought. This theological approach – which

continues to be popular in progressive academic circles in the United States – was attacked by conservative Catholics, clerical and lay alike, who accused its proponents of forsaking Christianity and falling prey to Marxism. Despite its apparent threat to Catholicism, Liberation theology constituted an attempt to save Latin America's dominant religion by enlisting it in the task of reforming society. Liberation theology was ultimately attacked by the local churches and neutralized by Rome, and it gave up its distinctive use of sociological categories and retreated into traditional spirituality.

Alternatives to Catholicism

As the conflicts between Rome and the radical theologians were taking place, a far more important development was unfolding, namely the spread of evangelical Protestantism. There is a strong affinity between Liberation theology, which could be considered as a Protestant Catholicism, and the Evangelical churches. In both cases we find a Puritan faith centered upon the sacred texts and suspicious of the rituals of both official Catholicism as well as popular religion. Both Evangelical Protestantism and Liberation theology appear to be attempts to move away from a sacramental, hierarchical, agrarian-based vision of the world in response to the weakening of the agrarian economy and the ascendance of an urban world ruled by the market.

Instead of the 'disenchantment' of the world, to use Weber's term, we find two religiously ambiguous formations – the Evangelical and Catholic – whose task it is to mediate between the two worlds constituted by economic transactions and by the supernatural. In these two versions of Christianity there is still the attempt to preserve the presence of the transcendent, while at the same time deritualizing everyday activities. Deritualization has been explained in terms of the rationalization of economic activities, since it allows individuals and communities to avoid the burden of wasteful ritual expenditures and provides release from the pressures exercised by community and tradition. The result has been a world of ever-widening economic disparities, which must be confronted in an instrumental manner by the individual, rather than by the community. Rather than repressing emotion, individualism and instrumental reason seem to exacerbate it. Thus, as in the tense coexistence of Pietism and Enlightenment found in eighteenth-century Northern Europe, we find that the strongest forms of Latin American, non-Catholic Christianity are Pentecostal, and this has led to the emergence of charismatic forms of Catholicism – one more example of the proliferation of offerings in the religious market that emerged after the breakup of monopolistic Catholicism. Latin America has also seen the growth of New Age religions among the middle and upper classes, as well as of religions of African origin among groups which are not of African descent, even in countries, such as Argentina, which do not have a significant black population.

While the early spread of Protestantism was caused by the activities of North American missionaries, the subsequent growth of Evangelical churches is now largely fueled by local ferment. In some countries, such as Puerto Rico, El Salvador, Guatemala, Chile, and Brazil, Evangelicals constitute a significant proportion of the population; whereas in others, such as Venezuela and Colombia, that presence is much smaller. Considering the continuous growth of non-Catholic forms of Christianity, the equation between Latin American religion and Catholicism is a thing of the past. The Catholic Church can still exercise pressure upon governments and legislatures on matters related to the control of female sexuality, particularly on issues of abortion and access to birth control, yet the birth rate has decreased in the last several decades throughout the region. Despite the strenuous efforts of the Church, in 2012 abortion was made legal in Uruguay, as it is in Cuba, Guyana, Puerto Rico, and Mexico. In addition, while Chile, the Dominican Republic, El Salvador, Haiti, Honduras, Guatemala, and Surinam forbid abortions under all circumstances, the rest of the Latin American countries now have legislation that permits abortion to preserve the life or the health of the mother.

Conservative Trends in Catholicism

A development that is likely contribute to further loss of the popularity of the Catholic Church among the lower classes, while perhaps increasing it among high-income groups, is the growing importance of conservative Catholic groups such as the Opus Dei, a secretive organization supported by the Vatican, particularly during John Paul II's regime. But even in the case of the Opus Dei, one can see its promotion as an attempt on the part of the church to accommodate itself to a changing world. In effect, despite its authoritarianism and conservative views regarding sexual matters, the values promoted by this organization are consonant with the demands of an unfettered capitalist order and with modernity.

The election of the Argentinean Jorge Bergoglio as pope – the first time that a Latin American has been elected to that office – has caused a sudden increase in the popularity of the papacy, particularly in Latin America. Given the new pope's apparent concern for the poor and his critique of the reigning neoliberal economic orthodoxy, it can be expected that groups such as Opus Dei will lose the influence they once had in the Vatican. It remains to be seen if Pope Francis' popularity will result in an increase in the influence of the Roman Church in matters related to government policies or the everyday behavior of Latin Americans, especially in relation to sexual matters. It seems unlikely that the personal popularity of the Argentinean Pope will return Roman Catholicism to the hegemonic position it once had in Latin America.

See also: Catholicism; Evangelical, Revivalist, and Pentecostal Religions; Religion, Sociology of.

Bibliography

Annis, S., 1987. God and Production in a Guatemalan Town. University of Texas Press, Austin, TX.

Bastian, J.-P., 1997. La mutación religiosa de América Latina. Para una sociología del cambio social en la modernidad periférica. Fondo de Cultura Económica, Mexico City.

Bricker, V.R., 1981. The Indian Christ, the Indian King: The Historical Substratum of Maya Myth and Ritual. University of Texas Press, Austin, TX.

Burga, M., 1988. Nacimiento de una utopía. Muerte y resurrección de los incas. Instituto de Apoyo Agrario, Lima, Peru.

Diacon, T.A., 1991. Millenarian Vision, Capitalist Reality: Brazil's Contestado Rebellion, 1912–1916. Duke University Press, Durham, NC.

Duviols, P., 1971. La lutte contre les religions autochtones dans le Pérou colonial (L'extirpation de l'idolatrie entre 1532 et 1660). Institut Francais d'Études andines, Lima [1977 La destrucción de las religiones andinas (durante la conquista y la colonia). UNAM, Mexico City].

Flores Galindo, A., 1988. Buscando un Inca: identidad y utopía en los Andes. Editorial Horizonte, Lima, Peru.

Garrard-Burnett, V., Stoll, D. (Eds.), 1993. Rethinking Protestantism in Latin America. Temple University Press, Philadelphia, PA.

Gose, P., 2008. Invaders and Ancestors: On the Intercultural Making and Unmaking of Spanish Colonialism in the Andes. University of Toronto Press, Toronto, ON.

Graziano, F., 1999. The Millennial New World. Oxford University Press, New York.

Gruzinski, S., 1988. La colonisation de l'imaginaire. Sociétés indigènes et occidentalisation dans les Mexique espagnol XVIe-XVIIe siècle. Gallimard, Paris [1991 La colonización de lo imaginario. Sociedades indígenas y occidentalización en el México español. Siglos XVI-XVII, Fondo de Cultura Económica, México].

Ingham, J.M., 1986. Mary, Michael, and Lucifer: Folk Catholicism in Central Mexico. University of Texas Press, Austin, TX.

Krickeberg, W., Trimborn, H., Müller, W., Zerries, O., 1961. Die Religionen des alten Amerika. Kohlhammer, Stuttgart [1968 Pre-Columbian American Religions. Weidenfeld & Nicolson, London/Holt, Rinehart and Winston, New York].

Lafaye, J., 1984. Mesías, cruzadas, utopías. El judeo-cristianismo en las sociedades iberoamericanas. Fondo de Cultura Económica, Mexico City.

Levine, R., 1992. Vale of Tears: Revisiting the Canudos Massacre in Northeastern Brazil, 1893–1897. University of California Press, Berkeley, CA.

Martin, D., 1990. Tongues of Fire: The Explosion of Protestantism in Latin America. Blackwell, Oxford, Cambridge, UK.

Mecham, J.L., 1966. Church and State in Latin America. University of North Carolina Press, Chapel Hill, NC.

Pardo, O.F., 2006. The Origins of Mexican Catholicism: Nahua Rituals and Christian Sacraments in Sixteenth-Century Mexico. University of Michigan Press, Ann Arbor, MI.

Penyak, L.M., Petry, W.J. (Eds.), 2006. Religion in Latin America: A Documentary History. Orbis Books, Maryknoll, NY.

Penyak, L.M., Petry, W.J. (Eds.), 2009. Religion in Latin America: Interpretive Essays from Conquest to Present. Orbis Books, Maryknoll, NY.

Prien, H.-J., 1978. Die Geschichte des Christentums in Lateinamerika. Vandenhoeck & Ruprecht, Göttingen.

Reff, D.T., 2005. Plagues, Priests, and Demons: Sacred Narratives and the Rise of Christianity in the Old World and the New. Cambridge University Press, Cambridge, UK.

Stoll, D., 1990. Is Latin America Turning Protestant? The Politics of Evangelical Growth. University of California Press, Berkeley, CA.

Christianity in the English-Speaking West

Wade Clark Roof and Steven Barrie-Anthony, University of California, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W.C. Roof, volume 3, pp. 1756–1761, © 2001, Elsevier Ltd.

Abstract

Canada, the United States, Australia, and New Zealand share origins as English colonies and thus are relatively young as nations and bound by Anglo legacies of language and culture. Protestant Christianity has functioned as a cultural establishment in each of these countries; Anglo religious life has been profoundly shaped by Protestant values such as resistance to religious authority, and a common experience and trajectory of religion and modernity in the West. While these countries are relatively young, their Protestant cultural establishments are increasingly viewed as old and faltering. Moreover, in the twentieth and twenty-first centuries, the historically dominant Protestant bodies in all these settings have suffered institutional declines and threats to cultural hegemony. There has been a notable rise in people identifying as having 'no religion,' but for many this entails more a separation from institutional religion than any divorce from the sacred. Spiritual eclecticism, drawing on Christian and non-Christian sources, as well as a focus on personal spiritual quest exist broadly inside and outside of Protestant institutions.

Historically, the major religious traditions for Anglo-America and Australasia are similar, despite the varying social and historical contexts in which they are located. That they all function, more or less, as cultural establishments within their countries provides a substantial basis for looking at them as a unified religious bloc. The countries making up this large bloc extending across continents – that is, Canada, the United States, Australia, and New Zealand – share origins as English colonies and, thus, are relatively young as nations and bound by Anglo legacies of language and culture. While the countries themselves are relatively young, in another sense the Protestant cultural establishments within them are increasingly viewed by many people as old and faltering – yet another reason for looking at them as of one piece. For all these reasons, comparative analysis of religious trends and dynamics is both possible and desirable, recognizing of course differences in size, polity, and demographics for the various countries.

Historical Considerations

From the outset, we should not underestimate the extent to which Anglo religious life continues to be fundamentally shaped by two historical realities: first, a Protestant background and thus a theological heritage of resistance to religious authority; and, second, a common experience and trajectory in the encounter of religion and modernity in the West. Obviously the two are closely related. The Protestant Reformation unleashed religious energies previously contained in medieval Catholicism and loosened the hold of ascriptive loyalties, thereby giving rise to greater autonomy of the individual in matters of faith and morals. Greater religious choice was accompanied by an elaboration of religious forms and styles, or a further working out generally of what is often referred to as the 'voluntary principle' within Protestantism. This principle is most pronounced in the United States, but elsewhere with Anglo influence, even in England itself, religious establishments have functioned in the modern period in an environment of considerable personal freedom. Protestantism has both given shape to, and been shaped by, high levels of individual freedom. In fact, three quite distinct historical waves of Protestant dissent legitimizing greater individual autonomy may be identified: first, the Calvinist movements; second, the Methodist revivals; and, more recently, the Pentecostals. As David Martin (1978) observes, the three religious waves correspond to shifting, ever-widening spheres of social influence. Unlike the Calvinists, who were limited mainly to social elites, the Methodists empowered working-class and middle-income populations, and the Pentecostals reached still lower social strata, and did so – and still do – in places and contexts often beyond the reach of the other two. For all three, the emphasis on the individual, on education and hard work, and on taking charge of one's life and making the most of it often resulted in upward mobility, and with that, a refocusing of belief systems and moral principles giving greater weight to the role of an individual's preference and conscience.

The encounter with modernity in the West led to a wide array of consequences, but most notably to a loosening of the binding character of tradition and memory (Hervieu-Léger, 2000) and an increased awareness of religious pluralism. Both mesh well with the mounting importance of individual autonomy and reliance upon personal choice in matters of belief. With this also came foundational shifts in faith, or what Peter Berger (1979) describes as a 'loss of ontological certainty.' Broadly speaking, the challenges to religious authority and increased pluralism ushered in an era of greater negotiation in matters of faith and commitment in the face of a widening array of choices. As the term 'secularization' is often used, it is much too encompassing and too easily glosses over the nuances of religion's encounter with modernity to adequately describe what happens in this situation, but it suffices to say that the latter as a historical and increasingly global process creates ruptures in shared religious views and forces upon ordinary people a posture of cognitive bargaining. Organized religion's monopoly over religious and spiritual questions is undermined. Religion's presence within the public arena takes on a new face. Philosopher Charles Taylor (2007, p. 3) describes 'a move from a society where belief in the traditional God is

largely unchallenged and unproblematic to one in which such belief is understood to be one option among others, and frequently not the easiest to embrace' – and yet the desire for spiritual 'fullness' and meaning persists. Growing numbers of people across Anglo-America and Australasia report having 'no religion' or being agnostics, yet many also simultaneously report being engaged in spiritual quests. From a Durkheimian perspective, religion and the sacred do not necessarily coincide, and indeed the emergence of the sacred as broader than or liberated from its earlier organized religious forms is a major transformation in our time.

For our purposes here, what is important are not just the parallels between Protestant influence and religion's encounter with the Enlightenment but also the fact that, as Berger and many other commentators point out, Protestantism has confronted the modern situation longer and to a greater extent than any other major faith tradition. That of course is rapidly changing in the contemporary world, where no religious tradition can insulate itself from pluralizing and individualizing trends. Yet for this very reason, the mainline Protestant experience is important and indeed paradigmatic of what other religions are now facing. Indeed, the fate of Protestantism in the modern West is what people of other faith traditions are quick to notice and worry about as perhaps their own. Apparent are deep internal strains within the tradition and disputes between liberal progressives, on the one hand, and conservative traditionalists, on the other, over moral and religious principles, brought about through selective interpretation of the tradition. In the case of the United States at present, we witness a strong and highly visible fusion of religious and political ideologies, an extension in ways yet also different from the earlier 'culture wars' of the 1980s and 1990s (see Hunter, 2010).

Yet it is not necessarily the case that 'Protestant' translates to 'Anglo' and certainly not that Anglo faith traditions can claim any superiority in their responses to modernity. The spiritual malaise that has currently fallen upon the historic, so-called mainline Protestant institutions throughout much of Anglo-America and Australasia augurs against any such claim. Nor is there any necessary presumption that other faith traditions must inevitably follow the Protestant trajectory in its confrontation with modernity; there are multiple passages and courses of accommodation. Even to argue that the Protestant experience, given its strong Anglo connections, is paradigmatic for other religious traditions is to risk the charge that this is still another, albeit subtle, claim to hegemony – alongside a broader history of racism, sexism, and capitalist exploitation. But there are widespread examples of 'Protestantization' within other faith traditions, such as the adoption of congregational forms of religious organization and the appeal to conscience relating to matters of faith and morality. Rather, we should look to the various countries described here as case studies in which to describe what is happening to these old Anglo religious establishments as they accommodate their peculiar social and cultural circumstances.

United States

The United States is the prime example of a country with a strong Protestant legacy, where the religious norms of individualism and adherence to conscience are now widely diffused and where the largely white, old-line Protestant denominations are mired in a deep spiritual malaise. Some would argue that the United States is the society that has undergone the greatest engagement with modernity, that in many respects its beliefs and practices are highly accommodated to its environment – religion in this context is secularized 'from within,' and thus a model of what to expect increasingly in the modernizing world. Yet, the United States remains highly religious among modern nations with regard to high levels of churchgoing and religious membership. This is true despite serious declines in membership within the old-line constituencies (notably the Episcopalians, Presbyterians, United Church of Christ, Disciples of Christ, and United Methodists) over the past 40 years, a time when the evangelical Christian sector has grown and become more mainstream. This overall switch in institutional forms is due to an array of factors: the influx of new-immigrant populations; higher birth rates and innovative programming, especially in the case of evangelicals; a more diverse Catholic population demographically and in terms of styles; generational changes favoring more direct, experiential faith communities; and, beginning with the Reagan presidency and lasting to the present, a reinvigorated conservative political ideology closely aligned with the social teachings of evangelical and fundamentalist Christian movements. Sometimes it is said that these old Anglo institutions have depleted their theological resources, that the problem is one of 'tired blood' in a context where there are so many fluid religious and spiritual styles.

The distinctiveness and vibrancy of religion in the United States are often explained by 'supply-side' and 'rational-choice' thinking, owing to a history of voluntarism and the necessity for religious constituencies to mobilize recruits as well as funds to sustain themselves. There is an emphasis on institutional innovation and competitive religious leadership. Carefully planned strategies in recruitment; the development of new types of religious organizations, such as megachurches; and the framing of religious messages for particular audiences coincide with a culture stressing choice as a principle across many sectors of life. That many Americans 'switch' religious affiliations as well as frequently move in and out of religious organizations underscores a high level of personal autonomy and decision making. With religion as with the mass marketing of goods and services generally, Americans respond to skillful packaging and presentation of the product. 'Selling God' and the working out of a principle of market differentiation in religion are almost as pervasive within the country as is the selling of anything else, although religious leaders resist the analogy.

A major shift, of course, is the increased religious diversity owing mostly to immigration as a result of the change in immigration policy in the mid-1960s. A major Pew Forum survey (2008) describes the breakdown of religious constituencies as follows: Christian, 78.4% (Evangelical Protestant, 26.3%; Catholic, 23.9%; Mainline Protestant, 18.1%; Historic Black Churches, 6.9%; Orthodox, 0.6%; and other Christians, 0.3%); Jewish, 1.7%; Mormon, 1.7%; Buddhist, 0.7%; Jehovah's Witnesses, 0.7%; Muslim, 0.6%; Hindu, 0.4%; other world religions, 0.3%; other faiths, 1.2%; plus unaffiliated, 16.1%; don't know, or refused, 0.8%; and a range of error of 0.6%. Those identifying as white Protestant

hover around or are less than 50% of the total population, a fact that has not gone unnoticed. Writing even some 40 years ago, sociologist Charles H. Anderson (1970, p. 3) observed, "The decline of the white Protestant majority and of white Protestant hegemony in twentieth century America has encouraged the growth of a self-conscious Protestant community." Such self-consciousness as well as anxieties relating to national identity are now considerably greater, especially among disadvantaged whites but also as articulated in some scholarly circles (see Huntington, 2005). Following 11 September 2001, the Muslim (and also Sikh) presence has evoked fears across many US communities of terrorism and of a changing American religious and cultural landscape.

A deep cleavage has emerged between religious traditionalists and religious progressives, one closely overlapping with political ideologies. The cleavage cuts across religious traditions but is particularly evident within the Christian sector. Since the 1980s, evangelicals and fundamentalists have picked up dropouts and dissidents from the more liberal Protestant churches, and the split within Catholicism along ideological lines has intensified. The thesis of a 'culture war' can be overstated since many Americans are caught in the middle on many social issues and are thus pulled one way or the other depending on the issue in question. Still, there is much tension and occasionally outright conflict, especially with regard to abortion and homosexuality. Disputes such as the latter arise out of seriously conflicting and often incommensurate notions of moral and religious truth and conflicting bases of authority from which to make such appeals: metaphorical as opposed to literalistic interpretations of scripture, and/or expansive, humanistic conceptions of the self versus a static view of human nature as in need of restraint.

Two other trends are important as well. One is the rise of a widespread spiritual quest culture that cuts across religious and nonreligious sectors. It flourishes among old-line Protestants who find their own rituals and worship services in need of rejuvenation; among 'seeker-minded' evangelicals who know very little about Christian tradition but are open to learning more; and among those who describe themselves as 'spiritual but not religious' or New Age spiritual, or simply adhere to a religio-spiritual system of belief and practice of their own choosing, often combining elements from various traditions (e.g., almost a quarter of Americans, many claiming to be Christian, report belief in reincarnation). The quest is for a greater sense of inner, experiential truth, to discover a more authentic and internally centered self (Roof, 1999). How this deepening of spirituality plays out in people's lives varies considerably. For many it takes the form of a personal journey, but for others it involves an enlivened sense of self within a supportive community. Words like 'soul' and 'spiritual,' which in an earlier time had all but disappeared in public discourse, have returned with renewed excitement. In this and other ways, the 'inward turn' today appears to be less a sign of protest against religion as such and more of assuming an effort to cultivate a meaningful sense of self and drawing upon of an array of resources to do so.

A second trend is the increase in numbers of people identifying with no religion, the so-called 'Nones'. In the 1960s, roughly 6% of people in the United States identified as having no religion; by 2012, that had climbed to 19 or 20%. The rise has been especially pronounced among young people, and among members of the millennial generation in particular roughly a third identify as nones (Pew Forum, 2012; Putnam and Campbell, 2010). Public intellectuals and organizations aligned with the 'New Atheism' movement, associated with the names of Richard Dawkins, Daniel Dennett, Sam Harris, and the late Christopher Hitchens, have been quick to claim this trend as evidence of increasing nonbelief. However, while atheist identification has expanded at a modest pace, atheists account for only a fraction of the no religion category; most report actually believing in God. Thus, while the upsurge of the nones and the elevation of 'spirituality' over 'religion' are troubling for many Protestant institutions, the persisting need for spiritual sustenance keeps the door open for religious innovation and perhaps even expansion. Christian movements such as the Emerging Church are taking advantage of this situation, yet it would be a mistake to assume the inexorable decline of traditional denominational forms.

Canada

Similar patterns of post-World War II Protestant decline are evident in Canada, despite the fact that the religious situation is very different there compared to the United States. Canada is historically less pluralistic (though that is now rapidly changing) and much more shaped by the presence of two large religious constituencies, Catholics and Protestants, vying with one another for power and influence. Catholics have long been numerous in Lower Canada, and Protestants, especially Anglicans, are sizable in Upper Canada. For the country as a whole, Protestants for a long time numbered more than Roman Catholics. But as of the 2001 census, Roman Catholics, benefiting from a higher birthrate and immigration streams, have become substantially larger with roughly 12.8 million members, concentrated particularly in Quebec. Protestants collectively number about 8.7 million, a 6% decline from 1991. The United Church of Christ (a merger in 1925 of Congregationalists, Methodists, and some Presbyterians) is the next largest at approximately 2.8 million members, concentrated in Ontario and the Prairies; the Anglicans claim roughly 2 million. Baptists were the only major Protestant group to increase significantly between 1991 and 2001, by 10% to about 729 000. Presbyterians, Lutherans, Pentecostals, and assorted conservative Protestant groups remain a significant presence, most below a million members. Presbyterians declined 36% between 1991 and 2001, the most of any Protestant group, and Pentecostals by 15%. Birth rates for Mormons, the Salvation Army, and conservative Protestant groups are generally higher than is the case for the more established Protestant denominations.

But the fate of Protestantism in Canada differs in some important respects compared with the United States. Mainline Protestants have suffered membership declines in Canada, losing both to Roman Catholic parishes and to some of the conservative churches. Proportionately they have lost more to Catholicism than is probably the case in the United States, where just the opposite pattern appears to be more prominent – that is, somewhat greater losses to the religious conservatives. Statistically, this is what one would expect: the larger the competing population, the greater the likelihood of mainline Protestants switching in as

a result of marriage or religious choice. The conservative religious presence in Canada is neither as well institutionalized nor as publicly visible as its neighbor to the south. Sociologist Reginald W. Bibby (2011, pp. 197–198) argues that in fact the alleged recent growth of conservative Protestantism in Canada is largely a misperception, more really a 'circulation of the saints,' or movement of people from one evangelical church to another, than a case of genuinely active and successful recruitment from other churches. Conservative Protestants tend to be converted over and over again – as a result of social and geographic mobility, marriage, divorce, friendship patterns, preferences for religious leaders, and congregational activities. Bibby's research suggests a slight increase in the recruitment of outsiders into the conservative fold over the past several decades, but even so the actual number of converts remains small.

Overall, levels of religious participation are moderate – higher than those of most European countries but lower than that found in the United States. Conservative Protestants have higher levels of weekly worship attendance than Roman Catholics and other Protestants, but because of their relatively small size proportionately and their tendency to circulate among themselves, their impact is not as widely felt as in the United States. In terms of inter-Protestant tensions, a liberal–conservative cleavage is evident, but the 'culture war' infrastructure in Canada is weaker than in the United States, and consequently it is less of a structural feature. The conservative-moralist flavor of evangelical Protestantism never held sway over Canadians in the way historically it did, and to some extent still does, in the United States. Sociologist Sam Reimer (2003) observes that while US and Canadian evangelicals in many ways resemble each other more than they do their fellow compatriots – for instance, in terms of emphasizing strong boundaries, the conversion experience, and similar styles of sexual morality – Canadian evangelicals are less religiously orthodox and less likely to align themselves with conservative politics, instead moving more fluidly across the political spectrum.

Particularly relevant to understanding the Protestant decline in Canada is the shift toward adopting 'no religion,' which parallels shifts in the United States and Australasia. In 1971, less than 1% of Canadians claimed 'no religion.' By 2001, however, such religious nones had increased to 16% of the population or 4.8 million people, usurping the Protestant position as the second largest 'religious' body in the country behind Roman Catholics. Young Canadians are especially likely to identify with no religion. While the ascent of no religion would seem on its surface to support secularization theories – and, as in the United States, Canadian proponents of the New Atheism are quick to claim the gains as their own – the data suggest a more complex picture. Bibby interprets the restructuring of Canadian religion as neither secularization nor religious revitalization, but rather as 'religious polarization' in which people increasingly move away from ambiguous relationships with religion and toward the entrenched embracing or rejecting of it. In 2005, roughly the same percentage of Canadians attended church weekly (25%) as did not attend church at all (23%). And yet in Canada, as elsewhere, even among the nones there does not seem to be a wholesale waning of spiritual need or quest. Research suggests that many of those identifying with 'no religion' not only seek out spiritual fulfillment but also are able to find it outside of conventional religion.

The wide-ranging constellation of individualized spiritual innovation and quest known as the New Spirituality is thus of great importance in Canada, and not just for spiritually inclined nones. The New Spirituality is less its own separate enclave than a general headwind moving through religious (and, it would seem, nonreligious) communities of all kinds, especially more moderate-to-liberal communities. In both the United States and Canada, many people interested in New Age, metaphysical, and mystical thought continue to identify with an inherited faith tradition – as Protestant or Roman Catholic. Religious as well as 'spiritual but not religious' Canadians often adopt pieces of outside philosophical and spiritual systems, combining them into a mélange that makes intuitive sense without feeling the need to shift their religious identification. It is not clear whether this dynamic is more the case in Canada or the United States, but in any case it represents a major convergence for the two countries in the adaptation of religion to modernity. In both of these highly individualized settings, religion is not so much abandoned in the strictest sense as it is privatized and modified to fit into people's life situations at any given time. Even where religion is overtly abandoned, as by the religious nones, relationships to the sacred often strongly persist.

Australasia

The very term 'Australasia' signals a confrontation and mixing of cultures, of East and West, a term quite fitting to both of our remaining countries, Australia and New Zealand. Both countries have a heritage of European settlement and Anglo dominance over aboriginal cultures, with the practices and structures of that heritage reflected in their social institutions. Both countries know the tensions arising out of an expanding religious pluralism, brought on largely by successive waves of new immigrants. Both are often described as increasingly secular societies, especially in political and civic life.

Yet there are significant differences. Australia has seen fewer new religions or even new versions of older religions than has New Zealand. Religious history for the former is more a telling of stories about a European past, unlike for the latter, where such stories are shaped more by local conditions and cultural and religious admixtures. Even Anglican Church history differs between the two countries. In Australia, the Anglican Church began as a convict church served by chaplains for whites, whereas in New Zealand, the Anglican Church was at first, and for a long time afterward, a Māori aboriginal church served by missionaries. It was not until the mid-1800s that the Pākehā, or white Anglicans, began to outnumber the Māoris. The Māori Rātana and Ringatū faiths are two examples of New Zealand's popular mixing of religious and cultural themes, combining aboriginal and Christian elements. This difference between the two countries need not be exaggerated, but it does register in religious and spiritual styles.

Australia's religious profile today is as follows: Roman Catholics, 25%; Anglicans, 17%; other (non-Christian) religions, 7%; and no religion, 22%. The Catholic ascendency over Anglicans emerged about 60 years ago. The old, well-established religious communities – the Anglicans, the Uniting Church, the Presbyterians and Reformed Churches, and the Lutherans – have all suffered declines during the past several

decades. Of these, the Anglicans fell most precipitously, declining 5% between 1996 and 2011. Christian or Christian-derived groups growing slowly or holding their ground include the Pentecostals, and the Jehovah's Witnesses. Non-Christian groups have expanded rapidly in recent decades due in large part to immigration, but remain proportionately small at less than 3% each. Of these, Hinduism saw the largest recent percentile gain, increasing by 86% between 2006 and 2011; Islam increased by 69%; and Buddhism by 48%.

The greatest change in Australian religious identification since World War II, and particularly among those born during this period, is the increase in people having 'no religion.' Religious nones presently stand at 22% of the population, concentrated especially among young people. Likewise, religious attendance has declined in most of the churches – and yet all the while, stated interest in spiritual well-being appears to have increased. Sociologist Gary Bouma (2006) argues that religion and spirituality are in fact flourishing in a 'postsecular' Australia, but in shifting forms often unhinged from traditional religious institutions and authority, and thus often overlooked by sociologists. This includes conservative evangelicalism and Pentecostalism, as well as New Age and nature spiritualties and boundless spiritual bricolage that draw creatively upon the diversifying religious and cultural landscape. Bouma (pp. 32–33) views Australian religion and spirituality as characterized by 'a shy hope in the heart,' as in temperament less brash and less prone to public display than is the case in the United States. There may actually be movement in the United States toward increased spiritual subtlety and quietude – the growing contingent of 'spiritual but not religious' people is often uneasy shoehorning their spirituality into words. But what seems clear is that Australia is host to a more dynamic spiritual and religious scene than is typically observed.

In New Zealand as of the 2006 census, the Anglicans remain the largest single constituency considering that both the Pākehā and the Māoris are included. Roman Catholics are the next largest population, followed by the Presbyterian, Congregational, and Reformed denomination, with a continuing strong Scottish heritage. About 40% of the population identify with one of the major Christian churches. Christian affiliations that increased between 2001 and 2006, while remaining a small percentage of the population, include Orthodox Christianity and evangelical and fundamentalist Christianity. Recent immigrant religious communities are growing rapidly through immigration but are infinitesimal in size compared with all others. There have been substantial declines in religious affiliation and congregational involvement since World War II, especially for major churches, and mostly as a result of trends among younger whites. Slightly more than a third of the population report having 'no religion,' and recent studies indicate rapid growth in this category. However, as throughout Anglo-America and Australasia, 'no religion' here seems to indicate discomfort with organized religious forms more than wholesale movement away from relationships with the sacred. One recent study found that as institutional religious adherence in New Zealand has declined, reports of religious experience have increased; and that 30% of the population agrees with the statement 'I don't follow a religion, but am a spiritual person interested in the sacred/supernatural' (Vaccarino et al., 2011). Spiritual forms on the ascent include New Age varieties and Anglo interest in Māori spirituality. The white and nonwhite populations are on somewhat different religious trajectories, evident most in the Anglo trends toward reduced organizational involvement. For the Māori, with their folk traditions and close spiritual attachments to the environment, organizational participation was never held up as a religious norm in quite the same way it was for Anglos.

Both Australia and New Zealand experienced enormous changes in the past half-century as they evolved from being made up predominately of villages and tribes to becoming more urban and multicultural in character. Religion, at least in its traditional organized forms, has lost much of its power to integrate life experiences and has become one among many life-worlds in which people may or may not participate. Anglo religious traditions have suffered in the process. There, as elsewhere in modern societies, we observe greater eclecticism and pragmatism. However, compared with Canada and even more so with the United States, two things stand out in making the religious responses to modernity in Australasia different: one is that fundamentalism seems to attract fewer people and offers less of a viable alternative, and the second is that there is less movement generally from one faith tradition to another, the more common movement being simply to leave organized religion in favor of 'no religion.'

See also: Catholicism; Christianity Origins: Primitive and 'Western' History; Christianity in Asia; Church and State; Nationalism: General; Religion and Politics in the United States; Religion, Nationalism, and Identity; Secularization.

Bibliography

Anderson, C.H., 1970. White Protestant Americans: From National Origins to Religious Group. Prentice-Hall, Englewood Cliffs, NJ.

Berger, P.L., 1979. The Heretical Imperative: Contemporary Possibilities of Religious Affirmation. Anchor Press, Garden City, NY.

Bibby, R.W., 2011. Beyond the Gods & Back: Religion's Demise and Rise and Why It Matters. Project Canada Books, Lethbridge, Canada.

Bouma, G.D., 2006. Australian Soul: Religion and Spirituality in the Twenty-First Century. Cambridge University Press, Melbourne, Australia.

Hervieu-Léger, D., 2000. Religion as a Chain of Memory. Rutgers University Press, New Brunswick, NJ.

Hunter, J.D., 2010. To Change the World: The Irony, Tragedy, and Possibility of Christianity in the Late Modern World. Oxford University Press, New York.

Huntington, S.P., 2005. Who Are We? The Challenges to American Identity. Simon and Schuster, New York.

Martin, D., 1978. The Dilemmas of Contemporary Religion. St. Martin's Press, New York.

Pew Forum on Religion and Public Life, 2008. U.S. Religious Landscape Survey – Religious Affiliation: Diverse and Dynamic. Pew Research Center, Washington, DC. http://religions.pewforum.org/pdf/report-religious-landscape-study-full.pdf.

Pew Forum on Religion and Public Life, 2012. "Nones" on the Rise: One-in-Five Adults Have No Religious Affiliation. Pew Research Center, Washington, DC. http://www.pewforum.org/uplodedfiles/topics/religious_affiliation/unaffiliated/Nonesontherise.pdf.

Putnam, R., Campbell, D., 2010. American Grace: How Religion Divides and Unites Us. Simon & Schuster, New York.

Reimer, S., 2003. Evangelicals and the Continental Divide: The Conservative Protestant Subculture in Canada and the United States. McGill-Queen's University Press, Kingston, Ontario.

Roof, W.C., 1999. Spiritual Marketplace: Baby Boomers and the Remaking of American Religion. Princeton University Press, Princeton, NJ.

Taylor, C., 2007. A Secular Age. Harvard University Press, Cambridge, MA.

Vaccarino, F., Kavan, H., Gendall, P., 2011. Spirituality and religion in the lives of New Zealanders. The International Journal of Religion and Spirituality in Society 1 (2), 85–96.

Christianity, Origins of

Michael L Humphries, Southern Illinois University, Carbondale, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by L. Woodhead, volume 3, pp. 1770–1773, © 2001, Elsevier Ltd.

Abstract

The present chapter considers the origins of Christianity in the first two centuries of its development, from the earliest Jesus movements to the writing of the apostolic tradition and institutionalization of the Roman Catholic Church. Departing from customary readings in biblical and religious studies whereby the study of Christian origins yields a singular and extraordinary event along the lines of an immaculate conception, whether literally or symbolically, the current line of inquiry presupposes a sociohistorical and cultural-anthropological perspective, noting that early Christianity is marked by the emergence of disparate and competing communities fighting for identity and social position.

Theoretical Considerations

At least since the appearance of Walter Bauer's (1934) publication, *Rechtgläubigkeit und Ketzerei im ältesten Christentum* (*Orthodoxy and Heresy in Earliest Christianity*), scholars have struggled with the fact that the early Christian texts testify not to a movement branching out into diverse and heretical representations of a singular event but to a discontinuity and difference that only later collapse into an emerging orthodoxy. The question is that if the essence of the movement is to be sought in the hour of its genesis, how is the traditional scholarship to respond when faced with a heterogeneous beginning? Unfortunately, rather than allow the disparity of origins to stand, and to refashion its philosophical sensibilities and methods in light of the human sciences, the scholarship insists that the emergence of Christianity must necessarily reflect a singular defining moment. Understood theoretically, whether the beginning is marked by an immaculate conception or simply by the extraordinary teachings of a Galilean Jew, all disparity is erased in favor of the lofty origin. Expressed historically, it follows the gospel narrative by charting a linear development beginning with the historical Jesus as the only point of origin, moving through the crucifixion and resurrection stories to the establishment of a church in Jerusalem, from which then proceed the missions of the apostles, first to the Jews (Peter) and then to the gentiles (Paul). A line of apostolic succession is then established whereby the core gospel is passed from the Apostles to the apostolic fathers, who in turn set in motion the founding of the Roman Catholic Church. Until recently, most constructions of Christian origins have fallen in line with this understanding, but the scholarship has seen a shift in thinking that takes seriously the plurality of its beginning as reflected in the diversity of texts, and makes full use of the theoretical perspectives of cultural anthropology and social theory (Castelli and Taussig, 1996; Cameron and Miller, 2004, 2011).

The Early Jesus Movements

Notwithstanding the overwhelming number of works that have sought to put forward a portrait of the historical Jesus, the most one can say is that he was a Jewish teacher from the region of Galilee who gathered together a small number of followers and whose teachings challenged local cultural conventions (Vaage, 1994). The nature of these early texts does not lend them to reconstructions of the life and personality of Jesus. On the contrary, the authors of these texts were not interested in preserving the historical person but in the appropriation of Jesus as founder and teacher of a school of thought addressing questions of cultural identity and social formation, questions that arise in response to Hellenistic assimilation, Roman occupation, competing Jewish communities, and the eventual destruction of the Jewish Temple in 70 CE.

Although it is clear that the earliest Jesus movements did not persist in the form in which they were conceived – not to mention the possibility that there could have been many other groups about which nothing is known simply because there is no documentation – there is little doubt that experimentation in social formation occurred early on, and that this process (somewhat disparate, to be sure) served as a foundation for the emergence of the later and more systematic attempts at social formation. Of these early Jesus groups, textual documentation for the following three is most in evidence: (1) the community that produced the Q document, (2) the community that produced the pre-Markan pronouncement stories, and (3) the community that produced the Gospel of Thomas. These texts represent a period of time somewhere between 40 CE and 60 CE. They precede the Gospels by several years, and do not reflect what will later become the Gospel story. They are, for the most part, collections of sayings that convey the teachings of Jesus; they contain neither birth narrative, nor passion narrative (trial and crucifixion), nor resurrection narrative.

Q is short for the German word *Quelle*, meaning 'source.' It is one of two source documents shared by the Gospels of Matthew and Luke (the other being the Gospel of Mark) and is the earliest extant record of a Jesus movement. Stratigraphic analyses of Q have demonstrated a community in the process of social formation, identifying the social context in accordance with the compositional history of Q (Kloppenborg, 1987; Mack, 1993). The formative layer consists of instructional material that represents a group of Jesus people associated with the lower administrative sector of cities and villages in the region of Galilee

who find themselves separated from the customary practices of the Temple and thereby respond by offering an alternative ethos. They call this ethos the kingdom of God, and it is characterized by a simple, almost cyniclike lifestyle. Subsequent layers of text incorporate prophetic and apocalyptic sayings that betray ever-increasing situations of conflict with representatives of the Temple system, including the Scribes and Pharisees. Fighting for position and identity, the Q community begins to imagine Jesus not only as its founder and teacher but also as a son of David, and to identify its members as the true sons of Israel.

Form-critical studies demonstrate that the Gospel of Mark appropriated a collection of stories customarily identified as pronouncement stories. These stories are structured formally as follows: Jesus is placed in a certain situation where someone challenges his teaching and/or activity. He then responds with a sharp answer that turns the tables on his challengers. For example, "When asked why they ate with unclean hands, Jesus replied, 'It is not what goes into a person, but what comes out that makes unclean'" (Mark 7:15). This collection of stories suggests a rather small rural community in conflict with local Pharisees regarding legal aspects of the Torah and the interpretation of purity codes. Unlike the Q community, there is no indication that the members of this group regarded Jesus as a son of David, or necessarily themselves as the true sons of Israel. Jesus is simply presented here as an extraordinary teacher who easily outmaneuvers the Pharisees and thus validates the group's daily practices (Mack, 1995).

In 1945, a collection of Coptic-Gnostic texts was discovered near the city of Nag Hammadi in Upper Egypt. Among these texts was a document with the title "These are the hidden sayings that the living Jesus spoke and Judas Thomas the Twin recorded," and signed at the end as "The Gospel of Thomas." As with the Q document and the collection of pronouncement stories, this text shows no interest in a biographical account of the historical Jesus, it consists almost entirely of sayings, and its focus is upon the teachings of Jesus. Unlike these texts, however, the Gospel of Thomas reflects a distinct trajectory of Jesus followers who embrace a Gnostic understanding of their founder and teacher. Jesus' teachings extend beyond the history of Israel; he is neither messiah nor son of David, nor the extraordinary rabbi who confounds the Pharisees. Rather, Jesus is the gnostic redeemer who awakens his followers to the nature of the cosmos and their place within it. As such, the Thomas community essentially rejects the epic of Israel and locates itself within a universe that transcends history and culture. Having received and understood Jesus' teachings correctly, they have become, like Jesus himself, the enlightened ones within whom the kingdom of God resides.

In addition to these three distinct groups of Jesus people, it is possible to identify at least two others, though the textual documentation is minimal. Like the community that produced the pronouncement stories, there appears to have been a community that produced a collection of miracle stories, which was also used as a source by the author of the Gospel of Mark. This group develops a series of miracle stories that play off of the exodus theme, identifying Jesus as a Moses-like figure who effects liberation not only from sickness and demon possession but also from poverty and oppression. Its social construct is not unlike the formation of Israel in Moses' time. Another group of Jesus people appears to have formed a small community in Jerusalem. Paul's letter to the Galatians, written in 55 CE, mentions two visits he made to the 'pillars' in Jerusalem, whom he identifies as Cephas (Peter), James, and John. The visits apparently dealt with a dispute over the teachings of Jesus with respect to the acceptance of gentiles into the community – most likely concerning the practice of circumcision – which suggests that the members of this group sought to maintain its Jewish tradition and thus held more firmly to the Jewish purity codes. Unfortunately, Paul says little else about these pillars and the community to which they belong, and there is no mention elsewhere of such a community. The names of Peter, James, and John do not reappear until much later, when they are cast in a bad light by the author of the Gospel of Mark shortly after 70 CE.

The Christ Cult and the Gospel According to Paul

The following declaration from Paul's letter to the Galatians reflects the emergence of a very different kind of Jesus movement outside the region of Palestine: "There is no longer Jew or Greek, there is no longer slave or free, there is no longer male and female, for all of you are one in Christ Jesus" (Gal. 3:28). Without question, the statement marks an extraordinary shift in thinking about social formation. To imagine a community wherein all traditional cultural boundaries are erased is, indeed, remarkable. And although one may surmise that the teachings of Jesus sparked a movement toward universality, the early Jesus groups did not betray as much. Moreover, the designation 'Christ Jesus' is here distinctive and no less remarkable. The early Jesus people in Galilee or Jerusalem would not have considered the Greek 'Christos' ('the anointed') an appropriate title for their teacher and founder. Although now attributed to Paul, the substance of the statement precedes Paul; he incorporated into his letters not only the ideas but also fragments of text (Rom. 3:24–26, 4:25; 1 Cor. 13:13, 15:3–5; Phil. 2:6–11) from a community ('the Christ Cult' as a descriptive designation) that emerged somewhere in northern Syria, perhaps in or around the city of Antioch, in the late 40s and early 50s (Mack, 1995). It is likely that some form of the Jesus movement(s) spread to the cities of Syria, Asia Minor, and Greece, where it attracted diaspora Jews as well as gentiles, and that the movement assimilated certain Hellenistic beliefs and practices, eventually emerging as a religious association along the lines of a mystery cult. And in keeping with common practice among the religious associations, the members may have gathered together for meals that became an occasion for storytelling and ritual, thereby constructing social identity. Drawing upon the Greek tradition of the 'noble death' and upon the stories of the Maccabean martyrs, the focus turns from Jesus as teacher to Jesus as noble warrior and martyr who, like Socrates, dies on behalf of his people, and like the Jewish martyr reemerges as a divine spiritual presence. Over time, Jesus the teacher is reimagined as Jesus the dying and rising Christ (the anointed son) who presides over the kingdom, the mystery of which is that all are invited to participate whether Jew or gentile, male or female, free or slave.

Though at one time a persecutor of the Jesus movements, Paul must have been convinced that the Christ Cult was on to something. Not only did he join, but also he became a leading figure in the movement. But in the beginning, as Paul's

correspondence with the Galatian and Corinthian groups reveals, not everything was to Paul's liking, nor was the Christ Cult free from the criticism of other Jesus movements (e.g., the pillars in Jerusalem). The dispute, of course, had everything to do with the all-inclusive nature of this community and the problems associated with this mixing of Hellenistic and Jewish sensibilities, especially as regards codes of table fellowship and the ritual mark of circumcision. Moreover, it would not have been easy for some of the Jesus people to make the transition from Jesus as the teacher and founder to Jesus Christ as the cosmic lord and savior of both Jew and gentile. But herein lies the genius of Paul's solution: if he could show that gentiles were also the children of Abraham, and thereby unite the Christ Cult with the history of Israel, then the ritual marks of ethnicity would not matter. All that is required is that one believed in Christ Jesus, who is now the fulfillment of the Torah. This is precisely what Paul did; and it is precisely this resignification of Israel's story that will help set the course for the emergence of Christianity as heir to the Hebrew epic, though it will be a matter of decades before the Pauline Gospel is fully entrenched and sanctioned by the writing of the apostolic tradition.

The Gospel Tradition

With the advent of the Roman–Jewish war (66–70 CE), everything changes. Not only is the city of Jerusalem and its temple destroyed, but also the entire history and significance of the Second Temple are put into question: the temple-state marked the center of Israel's epic, and it was to be the culmination of Israel's history as it came into full possession of the land and kingdom. But all that was now gone; the epic of Israel had to be imagined differently. The Scribes and Pharisees, forerunners to the rabbinic system, stepped in to fill the void: the written and oral traditions of the Torah were preserved and copied, the purity codes were maintained, and places of gathering (synagogues) had spread throughout the land. For the Jesus movements also, it was a matter of reimagining their place in this new world; they too were sons of Israel and, with the destruction of the Second Temple, had to reconsider their place as heirs to the kingdom. The solution came with a rereading of Israel's story wherein the coming of Jesus displaces the temple-state. It begins with the Gospel of Mark.

Written sometime in the early 70s, the Gospel of Mark is the first text to construct a life of Jesus – not a factual construction as imagined in contemporary biography, but a life narrative that enabled this northern Palestinian community of Jesus followers to establish its own identity and break free of the Pharisees and synagogues. Drawing upon the collection of pronouncement stories and miracle stories, as well as the Christ Cult image of a dying and rising god, the author of the Gospel of Mark writes the story of the son of God who is sent into this world to set right what the authorities of the Second Temple system had so clearly ruined. A new kingdom of God is imagined, and Jesus' death and resurrection mark the end of the temple and its sacrificial practices.

Several years later, the Gospel of Matthew (*c.* 88 CE) and the Gospel of Luke (*c.* 120 CE) will incorporate Mark's Gospel and weave it together with the Q document and material of their own making (the synoptic gospels). Jesus' life narrative will remain essentially the same, as will the displacement of the old (the Temple) with the new (the Passion narrative), but slight alterations, elaborations, and new material will reflect the distinct interests of their respective communities as well. The author of Matthew addresses a Jesus movement located in the region of southern Palestine, a movement that is troubled by the influx of gentiles and thus seeks to maintain a link with its Jewish heritage. Matthew's Jesus is tied genealogically to Abraham and David, he assimilates the role of rabbi and messiah, and his teachings (e.g., the Sermon on the Mount) serve to correct the teachings of his (their) Jewish counterparts. The Gospel of Luke, however, does not address the concerns of a specific Jesus group at all, but rather bears the form of a history intended to reach a much larger gentile audience in a much later period of time. Looking back on the last 50 years, the author constructs a history of salvation that begins with Adam, runs through the history of Israel, manifests itself in the passion narrative, and moves out to the nations of the world. The brilliance of this composition lies in a reconceptualization of an apostolic tradition that links the mission to the nations with the Jesus of the Gospel tradition.

The fourth Gospel (later attributed to John the beloved disciple) makes its appearance in the years between the Gospels of Matthew and Luke, sometime around 95 CE, and reflects a Christian community that developed its own distinct view of Jesus and what it meant to be a member of this community. It is difficult to trace the origins and evolution of this social formation. Nevertheless, textual elements suggest a Jesus group that emerged in the mid- to late 40s somewhere near Syria that, as with other Jesus groups, contended with local Jewish authorities regarding the observance of purity codes. Over time, the community developed a Christology (Gnostic in tenor) that presented Jesus as the divine Logos become flesh and constructed a narrative that reflected its severed ties with its Jewish family entirely: members of the Johannine community were encouraged to leave the synagogues and look upon 'the Jews' (no longer simply Pharisees) as those who failed to heed the signs and acknowledge Jesus as the way, truth, and light – the community of darkness.

Writing the Apostolic Tradition

With the exception of the Pauline letters, very little documentation exists that provides information regarding the early history of the apostles – in fact, the term itself must be used carefully when considering its early history. Aside from those limited references noting Paul's conflict with the pillars in Jerusalem, and apparent opposition to Paul's claim of apostleship (both referencing disputes over ethnic boundaries), all that can be said with any certainty is that the turn of the first century CE marked a distinct shift from Jewish Christian communities to predominately gentile Christian communities – which is not to say, however, that the respective communities were without their own internal disputes. When turning to Luke's Acts of the Apostles, the frequently made mistake is to assume that the book is a descriptive text, a factual record of what actually happened. But this is not the case. The apostles have been dead for many years, and the history that Luke writes comes after the fact. In broad strokes, the book charts out the course from Peter's mission to the Jews to Paul's mission to the gentiles, culminating with Paul's arrival in Rome. What proves so

remarkable about the writing of this tradition is that it sets the foundation for an apostolic succession of inheritance that will authorize the teachings and practices of the Church of Rome. Of course, there never actually was a line of succession that preserved an original orthodoxy. No scholar today believes that the disciples of Jesus, the so-called 12 apostles, passed along a core gospel to the apostolic fathers, who then set in motion the founding of the Roman Catholic Church. Nevertheless, Luke's Acts of the Apostles will come to serve as a myth of origins authorizing the apostolic line of succession, a simple yet successfully imagined history that will come to pass: the Jesus of the gospel tradition invests the apostles, the apostles invest the apostolic fathers, and the apostolic fathers will lay the foundation for the formation of the Roman Catholic Church.

The Path to Orthodoxy

Several years of conflict and resolution lie ahead before Luke's vision of the apostolic tradition is realized. Diversity will persist well into the second century. Numerous gospels, acts, epistles, apologies, and treatises will be written during this period of time, conveying diverse beliefs and practices represented by several Christian communities scattered throughout Syria, Palestine, and North Africa: Judaizers, Hellenists, Gnostics, aesthetics, women-specific communities, and more. Later, with the formation of the New Testament canon, many of these writings will comprise the collection of hidden texts (or Apocrypha) and the Gnostic writings of the Nag Hammadi library. In the midst of this diversity, the Christians who settled in Rome began to organize and develop an episcopate. The central players in this process, those who would later become known as the *apostolic fathers* (a seventeenth-century designation of orthodox scholarship), left a collection of writings that signaled movement toward a desired ecclesiastical organization of the episcopate seated in Rome, as well as a consolidation of early Christian writings and the belief that Christianity is the culmination of Israel's epic: letters attributed to Clement of Rome, letters of Ignatius of Antioch, the Didache, a letter attributed to Barnabas, an apocalyptic text entitled the Shepherd of Hermes, the Martyrdom of Polycarp, and the Epistle to Diogentus. Despite their efforts, however, the fathers were unable to secure a united front whereby their understanding of Christianity as the orthodox transmission of the gospel could be firmly set against so-called heretical manifestations. The turning point occurs in the middle of the second century (*c.* 140–150 CE), when the legitimacy of the episcopate in Rome is directly threatened by the growing popularity of two distinct visions of Christianity: the anti-Jewish vision of Marcion of Sinope and the Gnostic Christian vision of Valentinus of Alexandria.

Marcion drew immediate attention with the introduction of the first 'New Testament' canon, the intent of which was to sever all ties with Judaism. He believed that the God of Israel was the creator of a malevolent world, and that such a deity could not possibly be the God who sent Jesus, the merciful son, into this world. Christianity had already freed itself from the Mosaic law, and so it did not make sense to Marcion that Christians should maintain this connection with the God of Israel. Hence, he rejected the Jewish scriptures and all early Christian writings that retained a connection with Judaism, and proceeded to produce a canon of literature consisting of 10 Pauline letters and a truncated version of the Gospel of Luke. Valentinus, who laid claim to apostolic authority, was a leading figure in the Gnostic Christian movement of Alexandria, a movement that likely had its precursor in the early Jesus group that produced the Gospel of Thomas. The vision of Christ is that of the Gnostic redeemer through whom knowledge (gnosis) of the way back to the divine realm is imparted to those destined to return. They are the enlightened ones who move within a universe that transcends history and culture. The epic of Israel becomes irrelevant, and those who would lay claim to it remain ignorant of the knowledge held by Valentinus and the Gnostic Christians.

For the Church in Rome, the link with Judaism was foundational. The history of salvation, the plan set by the God of Israel, was marked by the coming of Christ, the legacy of which was entrusted to the Apostles who passed it along to the fathers and bishops in Rome. The response to Marcion and Valentinus came swiftly, first with the arguments set forth by Justin (Martyr) of Samaria, and then later supported by the writings of Irenaeus, Clement of Alexandria, and Tertullian. These were all highly educated men who presented powerful arguments demonstrating that the God of Israel was the Christian God who created the world, and that the line of succession leading to the Church in Rome was the divine plan from the beginning. Perhaps more importantly, however, these new Christian intellectuals also made the case to the Romans that Christianity as a religion was good for society and the government of Rome (Mack, 1995).

The conflicts would continue for some time to come, but the path to orthodoxy was clearly set, and the fathers in Rome would eventually win the day. Seven decades later, Constantine would convert to Christianity (313 CE) and shortly thereafter become emperor of the Roman Empire in 324 CE. In 325 CE, he called the first council of bishops to meet in Nicea and charged Eusebius, bishop of Caesarea, to collect the sacred scriptures and produce 50 copies for distribution among the new churches in Constantinople. Thus begins the formation of the New Testament canon and the centralization of Christianity in the Roman Catholic Church.

See also: Catholicism and Gender; Catholicism; Judaism; Myth in Religion; Orthodox Churches; Prophetism; Religion, Sociology of; Religion: Evolution and Development; Religion: Mobilization and Power.

Bibliography

Bauer, W., 1934. Rechtgläubigkeit und Ketzerei im ältesten Christentum. Mohr/Siebeck, Tübingen.

Cameron, R., Miller, M.P. (Eds.), 2004. Redescribing Christian Origins. Society of Biblical Literature, Atlanta.

Cameron, R., Miller, M.P. (Eds.), 2011. Redescribing Paul and the Corinthians. Society of Biblical Literature, Atlanta.

Castelli, E.A., Taussig, H. (Eds.), 1996. Reimagining Christian Origins. Trinity Press International, Valley Forge, Pennsylvania.

Kloppenborg, J.S., 1987. The Formation of Q: Trajectories in Ancient Wisdom Collections. Fortress Press, Philadelphia.

Mack, B.L., 1993. The Lost Gospel: The Book of Q & Christian Origins. HarperSanFrancisco, San Francisco.

Mack, B.L., 1995. Who Wrote the New Testament: The Making of the Christian Myth. HarperSanFrancisco, San Francisco.

Vaage, L.E., 1994. Galilean Upstarts: Jesus' First Followers According to Q. Trinity Press International, Valley Forge, Pennsylvania.

Chronic Illness, Psychosocial Coping with

Linda D Cameron and Christopher M Wally, Psychological Sciences, University of California-Merced, Merced, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Chronic illnesses are the leading cause of mortality worldwide. Generally noncurable, they range in severity from mild to life-threatening. Coping with the diagnosis and management of a chronic illness can be challenging for individuals and close others. Coping methods include problem-focused coping, emotion-focused coping, disengagement, accommodative coping, meaning-focused coping, and relationship-focused coping. How one copes may depend on individual differences such as emotional suppression tendencies, monitoring or blunting of information, and illness beliefs. Poor coping may worsen physical health through chronically elevated distress. Coping skills interventions can facilitate adaptive coping for individuals with chronic illness and their partners and families.

Chronic Illness and Coping Behaviors

Chronic illness is a long-lasting and permanent illness that is not curable by treatments and does not resolve over time. Chronic illnesses are the leading cause of mortality, accounting for an estimated 60% of all deaths worldwide (World Health Organization, 2011). They are highly prevalent in most developed countries. In the United States, for example, over 50% of the population has a chronic condition at any given time. Chronic illnesses range in severity from mild or moderately serious disorders, such as mild and well-controlled asthma, to serious and life-threatening conditions such as diabetes, heart disease, and cancer.

Most individuals diagnosed with a chronic illness experience at least some adverse consequences, threats, and limitations that have the potential to reduce their quality of life. Individuals use a variety of coping behaviors or strategies for managing the consequences brought on by their chronic conditions. Coping behaviors are defined as the efforts taken by an individual to manage threats, challenges and harms, or to manage the emotional distress they create. Understanding how individuals cope with chronic illnesses and the conditions under which specific coping strategies can be adaptive or maladaptive is a key area of research in social and behavioral sciences within the health field, including behavioral medicine, health psychology, nursing, public health, medical anthropology, and medical sociology.

Experiences Requiring Coping Efforts

For individuals diagnosed with a chronic illness, a primary coping goal is to manage or control disease progression through use of medications and other medical treatments as well as through diet, physical activity, and other health behaviors. Yet chronic illnesses also require coping with other aspects of the illness experiences. For many individuals, the development and diagnosis of a chronic illness is a disruptive and life-changing experience requiring significant changes in at least some areas of life. Aspects of life affected by chronic illness include somatic experiences, functional abilities, lifestyles, daily activities, social roles, self-concept, expectations for the future, and goals and priorities.

Individuals face challenges in enduring significant pain and noxious symptoms caused by the chronic condition and its treatment. For example, individuals with migraines suffer excruciating headaches that can last for several days; moreover, their unpredictability of occurrence can further undermine work progress, social plans, and psychological well-being. As another example, individuals diagnosed with colorectal cancer undergo painful treatments such as bowel surgery, chemotherapy, and radiation therapy. Chronic illnesses can cause other symptoms such as dizziness, itchiness, numbness, and lack of muscle control, all of which can cause discomfort and interfere with daily activities.

Chronic illnesses and their treatment can also lead to detrimental changes in cognitive and emotional function. Alzheimer's disease, Parkinson's disease, and brain cancer are just a few examples of chronic illnesses involving deterioration in cognitive functions. Individuals undergoing chemotherapy for cancer often experience cognitive deficits, and medications for numerous chronic conditions carry side effects such as depressive mood, confusion, and nervousness. The physical, somatic, and cognitive consequences of chronic illness can substantially limit abilities to engage in daily life activities such as running errands, socializing, working, and caring for others.

Chronic illness typically creates some degree of psychological distress. Anxiety is common, especially when waiting for test results, undergoing invasive medical procedures, and considering prospects of disability or death. For example, individuals treated for cancer report that fears of recurrence and spread are their top concern (Simard et al., 2013). Depression is also common among individuals with chronic illness, with greater illness severity being associated with higher levels of depression.

Key factors contributing to anxiety and depression include challenges to one's self-identity or self-concept, defined as the set of beliefs about one's qualities and characteristics, and self-esteem, defined as one's sense of self-worth. The physical and mental consequences of a chronic illness can lead to detrimental changes in body image, views of one's capacity for achievements and social activities, and ambitions and goals for the future. Chronic illnesses can also lead to real or perceived

stigma from others (Lee et al., 2002; Chapple et al., 2004). For example, individuals who experienced a stroke, smokers diagnosed with lung cancer, and individuals with HIV or AIDS frequently report instances in which they felt that others avoided contact with them, blamed them for their conditions, or acted toward them in a discriminatory manner.

Chronic illnesses significantly affect spouses, partners, and other family members, particularly those who serve as caregivers. These individuals must also engage in coping behaviors to manage the consequences, threats, and distress created by the illness and caregiving experiences. Individuals with a chronic illness and their spouses or partners often engage in dyadic coping, in which they view the illness as a shared experience and engage in joint coping efforts (Bodenmann, 2005). Dyadic coping may be positive when partners coordinate their efforts and express solidarity and togetherness in managing the illness-related challenges and tasks facing them. Alternatively, dyadic coping can be unbalanced or negative when one or both partners are combative, avoid or withdraw from stressful situations, are unresponsive to attempts to provide empathy or assistance, or fail to reciprocate provisions of support.

Distinctions in Coping Behaviors

Coping behaviors have been characterized in a variety of ways in social, behavioral, and health sciences. Different theoretical models of coping identify distinctions that have been used to guide empirical research and the development of coping measures. In this section, we review some of the more common distinctions used in research on coping with chronic illness.

Problem-Focused and Emotion-Focused Coping

A primary distinction in coping behaviors is between problem-focused and emotion-focused coping efforts (Carver et al., 1989; Lazarus and Folkman, 1984; Leventhal et al., 2012). Problem-focused coping behaviors are those aimed at reducing or resolving specific threats or negative consequences, such as controlling symptoms or finding ways to overcome illness-related barriers to engaging in work tasks. Emotion-focused coping involves the management or regulation of emotional distress caused by illness experiences, such as managing fears about treatment through expressing one's feelings to a spouse or seeking reassurance from a doctor. Common types of emotion coping strategies include expression or suppression of one's emotions with others, cognitive reframing or positive reappraisal of situations (e.g., focusing on one's ability to control one's diabetes through one's diet and physical activity), focus on negative thoughts (e.g., rumination about the illness threats and consequences), avoidance of negative thoughts (e.g., blocking out negative thoughts), avoidance of distressing situations (e.g., delaying or canceling routine health care visits for one's diabetes condition), and modulation of emotional states through the use of relaxation, alcohol or other substance use, or immersion in pleasurable activities.

Individuals often engage in both problem-focused and emotion-focused coping in order to meet the dual goals of controlling the objective threats and consequences created by the illness condition and managing emotional distress. Problem-focused and emotion-focused coping can be enacted in parallel or at different times during the course of illness. Moreover, they tend to interact and affect each other; for example, effective problem-focused coping can decrease emotional distress, thereby reducing the need for emotion-focused coping. Both types of coping behaviors can be helpful or detrimental, depending on the situation and how it is used. Problem-focused coping can be adaptive when a problem is solvable, but potentially maladaptive, if the threat or problem is uncontrollable. For example, an individual with advanced illness may find that persisting with a search for treatments that will cure the disease results in fruitless searching and ongoing frustration. Similarly, emotion-focused coping can be adaptive when it reduces distress, but it can be maladaptive if it interferes with active coping with the illness or leads to unhealthy behaviors.

Coping Engagement and Disengagement

Researchers also make a distinction between engagement and disengagement in coping efforts. Engagement refers to efforts to actively change or resolve problems or emotional distress, whereas disengagement refers to avoiding or ignoring the problematic issues or blocking out the distressing emotions. Disengagement generally involves those emotion-focused coping efforts aimed at escaping from emotional distress such as through avoidance, distraction, denial, and use of mind-altering substances. Disengagement from coping with a chronic illness is commonly unsuccessful over the longer term, and it can lead to dangerous outcomes even in the short term. For example, individuals with diabetes who ignore their condition and fail to eat meals when needed put themselves at risk of severe hypoglycemia, which can lead to mental confusion, unconsciousness, and seizures. Similarly, individuals with asthma who are in denial about their chronic health condition may fail to take their asthma prevention medication and thus increase their risk of serious and potentially fatal asthma attacks.

Disengagement from coping with illness-related problems does not always lead to maladaptive outcomes. In some instances, individuals may find that their quality of life improves when they let go of or give up on a goal. For example, individuals with end-stage cancer may decide to discontinue highly aggressive chemotherapy treatments that have little chance of slowing the progress of the cancer so as to preserve the quality of their final phase of life. As another example, athletes who are diagnosed with a life-threatening heart condition may choose to discontinue their sport as a health-protective measure. Importantly, however, these athletes are likely to cope best if they find a way to replace this important goal pursuit with an equally satisfying goal pursuit, such as an alternative profession that is just as rewarding or a role as a coach in the sport (Scheier and Carver, 2003).

Accommodative Coping and Meaning-Focused Coping

Another perspective that has guided research on coping with chronic illness gives emphasis to forms of engagement coping

involving efforts to adjust or adapt to the chronic illness. This form of coping, termed accommodative coping (Skinner et al., 2003), refers to cognitive adjustments in how one views the illness and the self as a means of coming to terms with unavoidable constraints arising as a consequence of the illness. Emphasis is given to coping strategies such as acceptance and changes in life goals and expectations; cognitive restructuring or reframing to minimize the costs and maximize the benefits of illness experiences; and distraction through engagement in positive activities such as hobbies, volunteering, family roles, and work projects.

Considerable research has focused on specific forms of cognitive restructuring known as meaning-focused coping. Meaning-focused coping refers to efforts to identify benefits and positive outcomes arising from the illness experience. For example, individuals with chronic illnesses often report that the experience has brought one's family closer together, made them stronger, served as a catalyst for positive experiences, and enabled them to understand what is truly important in life (Pakenham, 2011). Meaning-making and benefit-finding tend to increase with increasing severity of the illness condition, suggesting that individuals are increasingly motivated to utilize these coping strategies when illness consequences are regarded as relatively uncontrollable and improvement is unlikely. Evidence on outcomes of benefit-finding is mixed: Use of benefit-finding generally predicts lower depression and more positive affect over time, but it has been found to predict deleterious quality-of-life outcomes for some groups, such as those with more severe disease (Tomich and Helgeson, 2004).

Relationship-Focused Coping

In addition to the coping methods used to personally manage a chronic illness, individuals with chronic illness also utilize coping strategies to maintain their intimate relationships. Relationship-focused coping refers to efforts to preserve the quality of interactions and exchanges of social support with their partners. Active engagement coping, a common strategy involving the use of problem-focused efforts to resolve relationship conflicts brought on by the illness, tends to be beneficial in promoting well-being and marital satisfaction for both partners (Hagedoorn et al., 2000). Also beneficial to psychological adjustment and marital satisfaction among partners is communication about the relationship and the challenges brought on by the illness, as well as empathic responding to the partner's concerns (Acitelli and Badr, 2005; Bodenmann and Shantinath, 2004). Another coping strategy, protective buffering, has shown mixed results. Attempting to protect a partner from distress by masking one's own feelings may reduce distress for the partner, but it may increase distress for the partner who is suppressing emotions.

Individual Differences in Styles of Coping

Although many situational factors influence the selection and use of coping strategies over the course of a chronic illness, stable personal factors also shape these coping behaviors. Individual differences in coping include stable preferences for specific styles of coping, such as emotional suppression tendencies and information seeking styles characterized by either monitoring for or blunting out threat-related information. Individual differences in beliefs or cognitive representations of the nature of the illness also lead to differences in coping behaviors.

Emotional Suppression Tendencies

Suppression of negative emotions is a common emotion regulation tendency. Also known as emotional control, it is defined as a stable tendency to consciously inhibit expressive behaviors (including verbal, facial, and other nonverbal behaviors) reflecting negative emotions such as fear, anxiety, anger, frustration, sadness, and depression. Early research on how individuals cope with cancer revealed that patients commonly reported attempts to 'put on a brave face' and hide their negative emotions from others (Watson et al., 1991; Temoshok and Dreher, 1992).

Although suppression can carry short-term benefits in impression management and protecting others from one's personal distress, it tends to sustain negative emotional arousal, increase sympathetic activation, exacerbate symptoms, and interfere with successful coping. For example, a study of women undergoing chemotherapy for breast cancer treatment revealed that emotional suppression tendencies (particularly anger suppression) predicted higher levels of symptoms relating to immune function and cardiovascular arousal, as well as appraisals of poorer coping with the treatment (Schlatter and Cameron, 2010). Emotional suppression over the longer term can lead to detrimental health outcomes, including poor control of heart diseases and cancers, at least in part through the heightened physiological reactivity generated by sustained emotional distress. For example, correlational research links emotional suppression tendencies with poor adjustment to cancer experiences as well as faster cancer progression (Watson et al., 1991; Weihs et al., 2008).

Monitoring and Blunting Coping Styles

Monitoring and blunting coping styles can influence coping efforts related to information seeking, symptom monitoring, and use of medical treatments. Individuals with a monitoring coping style are motivated to scan for, attend to, and amplify information about their illness and potential treatments, whereas individuals with a blunting style are motivated to avoid and block out such information. High monitors benefit from opportunities to receive detailed information about their illness and treatment procedures, and they are likely to engage in careful monitoring of somatic experiences. For example, women in childbirth who were high monitors benefited from viewing a contraction monitor (Shiloh et al., 1998). High monitors tend to be more knowledgeable about their conditions in relation to low monitors, but they tend to be less satisfied with information provided by health sources, hold more negative beliefs about their condition, and experience more emotional distress (Roussi and Miller, 2014). High blunters tend to compartmentalize their illness from other areas of their lives and generally feel better when not attending

to threatening information about their illness. They tend to benefit from activities that enable them to distract their attention from noxious illness experiences, such as relaxation techniques prior to surgery (Miro and Raich, 1999).

Illness Beliefs

Subjective beliefs about the nature of one's illness influence the selection and enactment of coping strategies. According to Leventhal's Common-Sense Model of Illness Self-Regulation (Leventhal et al., 2012), individuals develop cognitive schema or representations of their illnesses based on past illness experiences and information from health care professionals, the media, and social networks. These representations guide efforts to cope with the illness itself and they generate emotional responses that trigger emotion-focused coping efforts. The Common-Sense Model identifies five primary attributes of illness representations: identity (beliefs about the illness label and associated symptoms), timeline (beliefs about its chronicity and episodic patterns), cause, consequences, and controllability/curability. Each of these beliefs can influence coping efforts (Hagger and Orbell, 2003). For example, individuals with hypertension, an asymptomatic disease, often erroneously identify symptoms that they believe arise when their blood pressure is high and they take their hypertension medication only when those symptoms are present (Meyer et al., 1985). Beliefs that an illness will have catastrophic effects can inhibit problem-focused coping and trigger emotion-focused coping efforts. Beliefs that an illness is acute rather than chronic in nature can lead to cessation of medications once symptoms seem under control.

Health Effects of Coping with Chronic Illness

Coping efforts can affect physical health through two routes. First, engagement in problem-focused strategies to control the illness through treatment adherence, healthy behaviors, symptom monitoring, and use of health care services can facilitate disease control, slow rates of disease progression, and promote social and physical functioning. Second, coping efforts that reduce emotional distress can inhibit physiological distress processes that can exacerbate disease progression. Threat-induced emotional distress activates the body's *flight or fight* response system, which includes activation of the hypothalamic-pituitary-adrenal (HPA) system and the sympathetic-adrenal-medulla (SAM) system. These systems prepare the body to respond to threatening situations by taking actions such as engaging in combat or fleeing from danger. Although adaptive in the short term, prolonged activation of the SAM and HPA systems can lead to detrimental physiological states. Prolonged SAM activation can exacerbate elevations in blood pressure, heart arrhythmias, and sleep difficulties; and prolonged HPA activation can promote inflammation and impair immune function (McEwen, 2004). Through these physiological processes, emotional distress could contribute to chronic illness episodes such as flare-ups of eczema, asthma attacks, and cardiac events; and it could accelerate the progression of immune-related diseases such as cancers and inhibit the healing of wounds from surgery and other procedures (Byrne-Davis and Vedhara, 2008; DeSteno et al., 2013). Problem-focused behaviors that reduce or resolve negative consequences of a chronic illness and emotion-focused behaviors that foster adaptive emotion regulation can not only promote mental well-being, but can also promote physical well-being and disease control.

Interventions to Improve Coping with Chronic Illness

Research on how individuals and their families cope with chronic illness has guided the development of interventions designed to improve their coping skills. These interventions typically target multiple coping skills such as problem-solving, communication, health behaviors, emotion regulation, and cognitive restructuring. Interventions can involve either group-based or individual sessions, and internet-based coping skills interventions are becoming increasingly common. Primary goals of these interventions include improvements in adherence to treatments, health behaviors, adjustment, emotional well-being, physical function, quality of life, and disease control.

Coping skills interventions have been found to increase the use of targeted coping behaviors and to result in improvements in both psychosocial and physical health outcomes, although their effectiveness varies considerably across intervention programs, individuals, and settings. A group-based coping skills intervention for skin cancer patients (Fawzy et al., 1993) serves as an example of a coping skills intervention that led to improvements in multiple quality of life and illness outcomes. This 6-week group intervention (relative to standard care) resulted in greater engagement in active behavioral and cognitive efforts, greater reductions in psychological distress, and greater improvements in immune function and survival rates.

Some interventions are designed to specifically target emotion regulation skills, including stress management strategies. Trials of emotion regulation interventions for individuals with cancer have demonstrated that participation in group interventions targeting emotional suppression and promoting adaptive emotion regulation strategies can reduce suppression tendencies and improve adjustment to cancer (Cameron et al., 2007). The cognitive-behavioral stress management (CBSM) program (Antoni et al., 2009) has been developed for use by individuals with a variety of chronic illnesses, including breast cancer, prostate cancer, HIV infection, and chronic fatigue syndrome. CBSM participants learn a variety of emotion regulation skills, including relaxation techniques, cognitive reframing strategies, and skills for improving use of social support. Trials testing the efficacy of CBSM consistently demonstrate that it increases targeted coping strategies and improves psychosocial adjustment, immune function, and quality of life.

Recent coping skills intervention research includes the development of programs designed for both individuals with chronic illness and their partners, other caregivers, and families. Couples-focused interventions typically provide training in communication skills for promoting effective dyadic coping with illness-related experiences and

relationship-focused coping for maintaining intimacy and relationship satisfaction (Manne and Ostroff, 2008). Family-focused coping interventions can be helpful for siblings of children with chronic illness, particularly through providing parents with skills for communicating information about the illness with their children and managing sibling emotions (Lobato and Kao, 2005).

Summary

Chronic illnesses are the leading cause of mortality worldwide. People living with a chronic illness must cope with the challenges of the illness and the distress it creates. Individuals cope with chronic illness using a variety of strategies. Problem-focused coping efforts involve reducing or resolving specific threats or negative consequences of the illness. Emotion-focused coping refers to the regulation of illness-related emotional distress. These coping strategies can co-occur, and they may interfere with one another at times. How engaged one is in coping efforts can impact management of the disease. Although disengagement from coping efforts often leads to maladaptive health outcomes, it can be adaptive when it involves disengaging from unattainable or unrealistic goals.

Accommodative coping is an engagement coping strategy reflecting cognitive adjustments in how one views the illness and the self, and it can facilitate adjustment when facing unavoidable illness consequences. Another cognitive restructuring strategy is meaning-focused coping, which reflects efforts to identify benefits and positive outcomes arising from the illness experience. Relationship-focused coping involves maintaining positive social support from partners through problem-focused efforts and empathic responding to concerns.

Individuals vary in their tendencies to engage in specific coping responses. Some individuals may naturally suppress emotions to mask feelings from others, which can lead to harmful effects on long-term health. Individuals may exhibit either a monitoring or blunting coping style. Monitors readily seek information, whereas blunters actively avoid it. One's illness beliefs also influence coping responses. Beliefs about illness identity, timeline, cause, consequences, and controllability all guide coping and behaviors.

Coping affects health through problem-focused efforts in managing the illness, as well as emotion-focused efforts in reducing distress and attenuating the effects of HPA and SAM system activation. Several interventions that target coping strategies have demonstrated improvements in psychosocial adjustment, immune function, quality of life, survival rates, and other important outcomes. Interventions that include patients and their spouses, children, and other family members have also produced positive results.

Coping with chronic illness has been, and remains, a topic germane to social and behavioral sciences in health, including health psychology and behavioral medicine. With increasing numbers of people living with a chronic illness at any given time, understanding the situational and dispositional factors related to illness management continues to be a pressing issue.

See also: Behavioral Medicine; Cancer: Psychosocial Aspects; Cardiac Disease, Coping with; Caregiver Burden; Co-morbidity of Mental and Physical Conditions; Depression, Pessimism, and Health; Diabetes: Psychosocial Aspects; Ecology and Health; Health Psychology; Health-Related Support and Self-help Mutual Aid Groups; Health: Anthropological Aspects; Illness Behavior and Care-Seeking; Illness and Dyadic Coping; Respiratory Disorders: Psychosocial Aspects; Rheumatoid Arthritis: Psychosocial Aspects; Self-Efficacy and Health; Stress, Coping and Health.

Bibliography

Acitelli, L.K., Badr, H.J., 2005. My illness or our illness? Attending to the relationship when one partner is ill. In: Revenson, T.A., Kayser, K., Bodenmann, G. (Eds.), Couples Coping with Stress: Emerging Perspectives on Dyadic Coping. American Psychological Association, Washington, DC, pp. 121–136.

Antoni, M.H., Lechner, S., Diaz, A., Vargas, S., Holley, H., Phillips, K., McGregor, B.A., Carver, C.S., Blomberg, B., 2009. Cognitive behavioral stress management effects on psychosocial and physiological adaptation in women undergoing treatment for breast cancer. Brain Behavior Immunity 23, 580–591.

Bodenmann, G., 2005. Dyadic coping and its significance for marital functioning. In: Revenson, T.A., Kayser, K., Bodenmann, G. (Eds.), Couples Coping with Stress: Emerging Perspectives on Dyadic Coping. American Psychological Association, Washington, DC, pp. 33–50.

Bodenmann, G., Shantinath, S.D., 2004. The couples coping enhancement training (CCET): a new approach to prevention of marital distress based upon stress and coping. Family Relations 53, 477–484.

Byrne-Davis, L.M.T., Vedhara, K., 2008. Psychoneuroimmunology. Social Personality Psychology Compass 2, 751–764.

Cameron, L.D., Booth, R.J., Schlatter, M.S., Ziginskas, D., Harman, J., 2007. Changes in emotion regulation and psychological adjustment following use of a group psychosocial support program for women recently diagnosed with breast cancer. Psycho-Oncology 16, 171–180.

Carver, C.S., Scheier, M.F., Weintraub, J.K., 1989. Assessing coping strategies: a theoretically based approach. Journal of Personality and Social Psychology 56 (2), 267–283.

Chapple, A., Ziebland, S., McPherson, A., 2004. Stigma, shame, and blame experienced by patients with lung cancer: qualitative study. British Medical Journal 328, 1470.

DeSteno, D., Kubzansky, L.D., Gross, J.J., 2013. Affective science and health: the importance of emotion and emotion regulation. Health Psychology 32, 474–486.

Fawzy, I., Fawzy, N.W., Hyun, C.S., Elashoff, R., Guthrie, D., Fahey, J.L., Morton, D.L., 1993. Malignant melanoma. Effects of an early structured psychiatric intervention, coping, and affective state on recurrence and survival 6 years later. Archives of General Psychiatry 50 (9), 681–689.

Hagedoorn, M., Kuijer, R.G., Buunk, B.P., DeJong, G.M., Wobbes, T., Sanderman, R., 2000. Marital satisfaction in patients with cancer: does support from intimate partners benefit those who need it most? Health Psychology 19, 274–282.

Hagger, M., Orbell, S., 2003. A meta-analytic review of the common-sense model of illness representations. Psychology Health 18, 141–184.

Lee, R.S., Kochman, A., Sikkema, K.J., 2002. Internalized stigma among people living with HIV-AIDS. AIDS and Behavior 6 (4), 309–319.

Lazarus, R.S., Folkman, S., 1984. Psychological Stress and the Coping Process. Springer, New York.

Leventhal, H., Bodnar-Deren, S., Breland, J.Y., Gash-Converse, J., Phillips, L.A., Leventhal, E., Cameron, L.D., 2012. Modeling health and illness behavior: the approach of the common-sense model. In: Baum, A., Revenson, T., Singer, J. (Eds.), Handbook of Health Psychology, second ed. Psychology Press, Hoboken, pp. 3–36.

Lobato, D.J., Kao, B.T., 2005. Brief report: family-based group intervention for young siblings of children with chronic illness and developmental disability. Journal of Pediatric Psychology 30, 678–682.

Manne, S.L., Ostroff, J.S., 2008. Coping with Breast Cancer: A Couples-Focused Group Intervention, Therapist Guide. Oxford University Press, New York.

McEwen, B.S., 2004. Protection and damage from acute and chronic stress: allostasis and allostatic overload and relevance to the pathophysiology of psychiatric disorders. Annals of the New York Academy of Sciences 1032, 1–7.

Meyer, D., Leventhal, H., Gutmann, M., 1985. Common-sense models of illness: the example of hypertension. Health Psychology 4, 115–135.
Miro, J., Raich, R.M., 1999. Preoperative preparation for surgery. Journal of Advanced Nursing 26, 356–363.
Pakenham, K., 2011. Benefit-finding and sense-making in chronic illness. In: Folkman, S. (Ed.), The Oxford Handbook of Stress, Health, and Coping. Oxford University Press, New York, pp. 242–268.
Roussi, P., Miller, S., 2014. Monitoring style of coping with cancer related threats: a review of the literature. Journal of Behavioral Medicine. http://dx.doi.org/10.1007/s10865-014-9553-x.
Scheier, M., Carver, C.S., 2003. Goals and confidence as self-regulatory elements underlying health and illness behavior. In: Cameron, L.D., Leventhal, H. (Eds.), The Self-Regulation of Health and Illness Behaviour. Routledge, London, pp. 17–41.
Schlatter, M.S., Cameron, L.D., 2010. Emotional suppression tendencies as predictors of symptoms, mood, and coping appraisals during AC-chemotherapy for breast cancer treatment. Annals of Behavioral Medicine 40, 15–29.
Shiloh, S., Mahlev, U., Dar, R., Ben-Rafael, Z., 1998. Interactive effects of viewing a contraction monitor and information-seeking style on reported childbirth pain. Cognitive Therapy and Research 22, 501–516.
Simard, S., Thewes, B., Humphris, G., Dixon, M., Hayden, C., Mireskandari, S., Ozakinci, G., 2013. Fear of cancer recurrence in adult cancer survivors: a systematic review of quantitative studies. Journal of Cancer Survivorship 7, 300–322.
Skinner, E.A., Edge, K., Altman, J., Sherwood, H., 2003. Searching for the structure of coping: a review and critique of category systems for classifying ways of coping. Psychological Bulletin 129, 216–269.
Temoshok, L., Dreher, H., 1992. The Type C Connection: The Behavioral Links to Cancer and Your Health. Random House, New York.
Tomich, P.L., Helgeson, V.S., 2004. Is finding something good in the bad always good? Benefit finding among women with breast cancer. Health Psychology 23, 16–23.
Watson, M., Greer, S., Rowden, L., Gorman, C., Robertson, B., Bliss, J.M., Tunmore, R., 1991. Relationships between emotional control, adjustment to cancer and depression and anxiety in breast cancer patients. Psychological Medicine 21, 51–57.
Weihs, K.L., Enright, T.M., Simmens, S.J., 2008. Close relationships and emotional processing predict decreased mortality in women with breast cancer: preliminary evidence. Psychosomatic Medicine 70, 117–124.
World Health Organization, 2011. Global Status Report on Noncommunicable Diseases 2010. WHO Press, World Health Organization, Geneva, Switzerland.

Relevant Websites

http://www.apa.org/helpcenter/chronic.aspx – American Psychological Association.
http://www.apa.org/pi/aids/resources/research/schneiderman.aspx – American Psychological Association: Cognitive-behavioral stress management and secondary prevention in HIV/AIDS.
http://www.cdc.gov/chronicdisease/ – Centers for Disease Control and Prevention.
http://www.cancer.gov/cancertopics/coping/emotionaleffects – National Cancer Institute.
http://www.nlm.nih.gov/medlineplus/copingwithchronicillness.html – U.S. National Library of Medicine.
http://www.who.int/topics/chronic_diseases/en/ – World Health Organization.

Chronic Pain: Models and Treatment Approaches

Dennis C Turk, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides an overview of the current understanding of pain and its management. It reviews historical conceptualizations of chronic pain, namely, the biomedical, psychogenic, and motivational (secondary gain) models. It provides updated perspectives based on the gate control theory, operant conditioning, and cognitive-behavioral models and perspectives. An integrated, multidimensional model that integrates neurobiological, psychosocial, and behavioral features of these models is presented. The most common perspectives and treatments of pain based on the various models are described with an emphasis on the importance of a multidisciplinary approach to understanding and treating patients with complex chronic pain syndromes.

Pain has been of interest to philosophers, scholars, healers, and laypersons since the beginning of recorded history. Although there have been major advances in knowledge of sensory neurophysiology and the development of sophisticated pharmacological, surgical, and neuroaugmentative interventions, no treatments have proven to be consistently effective (Turk et al., 2011). Traditional medical models have been based on the assumption that there should be a close alignment between objective physical pathology and the extent of self-reported pain. Clinical observation and research have identified a number of challenges to this approach. Several alternatives to the pure biomedical conceptualization have been proposed, namely, psychological causation and maintenance models. These alternative models focus on the role of psychopathology and psychological and motivational factors in pain reports in the absence of objective physical pathology, and the role of environmental reinforcement in the maintenance of behavioral indicators of pain. These conceptualizations are, however, incomplete. Current models of pain (i.e., gate control and cognitive-behavioral) underscore the role of cognitive, affective, and behavioral factors, as well as somatic ones (Flor and Turk, 2011). Understanding, assessment, and treatment of patients experiencing pain based on comprehensive views emphasize the need to incorporate psychological and social variables with neurophysiology. The emphasis on treating patients with chronic pain has moved from cure to rehabilitation and self-management.

Introduction

Pain is one of the most common reasons for seeking medical care (Hart et al., 1995; Gureje et al., 1998). For many, pain is an adaptive response, warning of the potential for bodily harm. This signal helps to prevent further injury and tissue damage. However, research as well as common experience suggests that pain is much more than a mere barometer of the amount of tissue damage. Despite the universal experience of pain, it is a difficult phenomenon to define precisely. We have all seen instances where two people report radically different amounts of pain despite what appears to be an equivalent injury. Clinicians have also noted that soldiers report little or no pain after a combat injury until after they are removed from the battlefield. The wide variations in how people communicate their pain further complicate matters. Some may express pain by crying or groaning, whereas others are more stoic and handle their pain by gritting their teeth and suffering silently.

It is important to distinguish 'pain' from 'nociception.' Nociception is a physiological process by which transduction of sensory information is activated in specialized nerve endings that convey information about tissue damage (e.g., location, descriptive features) to the central nervous system. Pain, on the other hand, is an integrated perceptual process that requires a conscious organism. A person who is completely anesthetized does not report pain associated with a surgical incision. The International Association for the Study of Pain (IASP) recognizes the complexity of pain and treats it as a phenomenological experience unique to each person. According to the IASP definition, pain is "an unpleasant sensory and emotional experience normally associated with tissue damage or described in terms of such damage" (Merskey, 1986). In addition to this global definition, many commonly used pain terms and related phenomena have been described (Turk and Okifuji, 2001). For example, pain may be acute and a response to an identifiable trauma or disease, or it may be episodic and occur intermittently as in the case of migraine headache. Pain may also be relatively constant, extending for years and even decades. Unlike those with acute or recurrent pain, people with chronic pain may experience it 24 h a day, 365 days a year. **Table 1** includes a list of four classes of pain based upon timing and clinical implications.

In this article, I provide an overview of the current understanding of pain and its management. First I review the conceptualizations that have dominated views of pain historically. Then I describe current conceptualizations that integrate multifactorial aspects of pain. The article also focuses on psychosocial factors that have been shown to contribute to the subjective experience of pain as well as adaptation and response to treatment. Following the discussion of models of pain, I provide a historical overview of the most common perspectives and treatments of pain. I emphasize the importance of a multidisciplinary approach to understanding and

Table 1 Four types of pain

Transient pain:	Pain elicited by activation of nociceptive process but no significant injury is present and pain stops as soon as the activation ends. This type of pain is ubiquitous in everyday life and rarely is a reason to seek heath care. Examples: pain associated with injection for immunization
Acute pain:	Pain elicited by nociceptive activation at the site of injury or disease. This type of pain is often a reason to seek medical care. Examples: Postsurgical pain, pain associated with muscle sprain
Recurrent pain:	Episodic or intermittent pain. Each pain episode is relatively short lasting but recurs across an extended period of time. Examples: Migraine headaches, tic douloureux (trigeminal neuralgia), sickle cell crisis
Chronic pain:	Usually elicited by an injury but lasting for a long period of time after the original damage has been healed. At the chronic stage, no objective pathology may be present that can explain the presence of pain. Examples: Chronic low back pain, fibromyalgia

Source: Turk, D.C., Okifuji, A., 2001. Pain terms and taxonomies of pain. In: Loeser, J.D., Butler, S.H., Chapman, C.R., Turk, D.C. (Eds.), Management of Pain, third ed. Lippincott Williams & Wilkins, Philadelphia, pp. 17–25.

treating complex chronic pain syndromes. Finally, I provide an overview of the psychosocial strategies that incorporate behavioral and cognitive-behavioral perspectives with the goals of improving pain and functioning of people with persistent pain. It is not meant to be suggested, however, that pain is purely psychological or that physical factors are unimportant; in fact, they are most often necessary, at least initially. Rather, emphasis is put on psychosocial factors, as they have previously been less integrated into conventional thinking about chronic pain and have been shown to play a growing role in the maintenance of pain and disability over time. The psychological approaches described are almost always employed within the multidisciplinary treatment program (Turk et al., 2011).

Historical Conceptualizations of Pain

Over the centuries, many prominent scholars have attempted to uncover the mechanisms underlying pain. Perhaps among the most influential was the seventeenth-century French philosopher, René Descartes. Descartes conceptualized pain as the result of representation of a painful stimulus traveling along the nervous system starting at the periphery (the flame depicted in Figure 1 that will travel along the nerve from the periphery to the spinal cord eventually reaching the brain where it is acknowledged as 'pain'). Descartes's approach formed the basis for the mind–body dualism that plays a prominent role in the understanding and treatment of pain.

Biomedical (Pain = Physical Pathology) Model

The traditional biomedical view that continues to dominate is reductionist. This perspective assumes that every report of pain must be associated with a specific physical cause (e.g., pain caused by the flame in Figure 1). As a consequence, the extent of pain reported should be directly proportional to the amount of tissue damage. The expectation is that once the physical cause has been identified, appropriate treatment will follow

Figure 1 L'homme de René Descartes. Reproduced by permission of René Descartes Paris: Charles Angot, 1664.

and positive outcomes will be attained. Treatment typically focuses on removing the putative cause of the pain, or by chemically or surgically disrupting the pain pathways (cutting or blocking the transmission of signals from the periphery to the brain). According to this model, once the cause of the pain is removed or the pain pathways are blocked, pain should be eliminated. There are, however, several perplexing features of pain that do not fit neatly into the biomedical model with its suggestion of a one-to-one relationship between tissue pathology and symptoms. Most notably, not all pain accompanies observable pathology. For example, in over 85% of the cases of back pain, one of the most common pain conditions in Western countries, the cause of the pain is unknown (Deyo, 1986). Conversely, physical pathology does not necessarily cause pain. Diagnostic imaging studies identify significant pathology in up to 35% of people who are pain-free (Jensen et al., 1994). Thus, a paradox exists. On the one hand there are people without significant pathology who report severe pain and on the other there are people with objectively determined physical pathology who do not report the presence of any pain.

Psychogenic ('It Is All in Your Head') Model

According to the psychogenic model, pain, in the absence of physical pathology, is a result of the patient's inherent personality traits or psychological problems. The psychogenic view is posed as the flip side of the coin from the physical or biomedical model. From this perspective, if the report of pain occurs in the absence of, or is disproportionate to, objective physical pathology, then, *ipso facto*, the pain reports must have a psychological etiology. Thus, a dichotomy is posed, pain is either somatogenic or psychogenic.

Motivational (Secondary Gain) Model

A related conceptualization to the psychogenic model specifically focuses upon motivation. From this perspective, reports of pain in the absence of physical pathology are attributed to the desire of the person complaining to obtain some benefit such as attention, time off from undesirable activities, financial (disability) compensation (i.e., secondary gain), or to obtain prescribed medication with centrally reinforcing properties. Unlike the psychogenic model, in the motivational model, the assumption is that the person is consciously attempting to acquire a desirable outcome based on the complaint of pain. Thus, the report of pain in the absence of pathology is taken to be suspect and fraudulent. Simply put, the person who reports pain in the absence of objective pathology is pretending to experience pain in order to obtain a desired outcome.

Operant (Social Reinforcement) Model

The operant or social reinforcement model is based upon the behavioral learning paradigm in which the probability of exhibiting a specific behavior is observable, serves a communicative function, and depends on the consequences that follow the behavior. Behaviors are most likely to recur when they are reinforced by desirable consequences (e.g., attention, sympathy), by removal of something aversive (e.g., participating in undesirable tasks, work activity), or avoidance of some undesirable consequence. Fordyce (1976) suggested that certain behaviors associated with pain (e.g., limping, grimacing) could be reinforced to the point where the behaviors recur even when the original nociceptive stimuli is no longer present. Pain behaviors that are consistent with acute pain and that may be reflexive responses to noxious stimulation, may become chronic pain behaviors and be maintained even when there is no physiological basis for those behaviors. Thus, in this model, chronic pain is considered a manifestation of learned behaviors that need to be extinguished, whereas well behaviors, such as activity, need to be increased. The laws of operant learning are viewed as playing a central role in the maintenance of chronic pain behaviors. Other factors, such as physiology, emotions, and the persons' attitudes and beliefs, play little role in the operant model.

Multidimensional Conceptualization of Pain

The unidimensional models described fail to explain individual and situational variation in pain experiences. The article will now focus on more contemporary models of pain with multifactorial bases.

Gate Control Theory

Melzack and Wall (1965) developed the gate control theory that integrated physical and psychological factors regarding pain experience. Briefly, the gate control theory proposes that a mechanism in the dorsal horn of the spinal cord acts as a 'gate' that can inhibit or facilitate transmission of nerve impulses from the periphery to the brain. Melzack and Wall postulated that not only injury sites at the periphery influence the gating mechanism but also people's emotional states. They suggested that the reticular formation in the brain functions as a central biasing mechanism inhibiting the transmission of pain signals. Psychological factors that affect the reticular formation may modulate the pain experience.

Melzack and Wall emphasized the modulation of inputs in the dorsal horn of the spinal cord and on the dynamic role of the brain in pain processes and perception. As a result, the gate control theory integrates psychological variables such as past experience, attention, and other cognitive activities into current research on and therapy for pain. Prior to this formulation, psychological processes largely were dismissed as merely reactions to pain. This new model suggested that severing nerves, the putative pain pathways, was inadequate as a host of other factors modulated the input. Perhaps the major contribution of the gate control theory was that it highlighted the central nervous system as playing an essential role in the perception and interpretation of nociceptive stimuli associated with the experience of pain.

The physiological details of the gate control model have been challenged and it has been suggested that the model is incomplete. As additional knowledge has been gathered since the original formulation of the gate control model, specific points have been disputed and physiological details of the model have been refined. The conceptual aspects of the gate control theory have, however, proved remarkably resilient and flexible in the face of accumulating scientific evidence. The gate control model still provides a powerful summary of the phenomena observed in the spinal cord and brain, and has the capacity to explain many of the most puzzling problems encountered in the clinic. The gate control theory has had enormous heuristic value in stimulating further research in the basic science of pain mechanisms and in spurring new approaches to treatment.

Cognitive-Behavioral Model

The focus of the cognitive-behavioral model is on the person's own subjective perspectives (e.g., attitudes, beliefs, expectations) and feelings about their plight (Turk et al., 1983; Flor and Turk, 2011). The model assumes that, although nociceptive stimuli precede pain, how the person perceives the nociceptive event forms a total pain experience by interacting with the sensory event. For example, negative and pessimistic views by people about their pain condition (e.g., 'It is hopeless and it will never get better.') and their capabilities for managing pain and stress (e.g., 'There is absolutely nothing I can do about my pain.') are likely to exacerbate their emotional distress and sense of disability. Similarly, if one views pain as inevitable (e.g., 'I was injured, that is why I hurt.'), attention to sensory events may become pronounced, and as a consequence, relatively subtle sensory information may be interpreted as being painful.

The cognitive-behavioral model also acknowledges the effects of physiological factors and the environment on behaviors and of behaviors on thoughts and feelings. Reporting of symptoms to family and health care providers is influenced by how people view their pain problem and, of course, physical pathology.

The basic assumption of the cognitive-behavioral model is that people are not passive entities who simply react reflexively to nociception or social reinforcement, but actively engage themselves in defining the experience. Based upon past learning and medical history, people develop subjective representations of illness and symptoms – 'schemas.' These schemas become the filters through which people process new sensory stimulus (Cioffi, 1991). Beliefs about the meaning of pain and one's ability to function despite pain are important aspects of pain schemas. When confronted with pain, people draw causal, covariational, and consequential inferences about their symptoms based upon their own schematic references. For example, if the schema includes a strong belief that all physical activities must be ceased when experiencing pain and that pain is an acceptable justification for neglecting domestic and occupational responsibilities, poor adaptation and coping are likely to result (Williams and Thorn, 1989).

Treatments

As has been discussed previously, pain is a complex phenomenon and diverse in its nature and degrees of associated impairment. It is therefore important to know the objectives of treatments for pain. Many pain states are expected to remit over time with little intervention required. Some types of acute pain may require surgical or pharmacological therapy to treat underlying pathology or to block noxious sensory information. A more comprehensive plan is required for treating people with chronic pain. Many people with chronic pain live with their pain for years, and the adverse effects of pain may be generalized across all aspects of their lives, familial, social, occupational, as well as physical. Given the multilayered problems associated with pain, treatment goals become more global. Rather than treating pain, it becomes necessary to treat people with chronic pain.

Historical Approaches to Treatment

The earliest approach to treating pain was aligned closely with unimodal views, particularly those based on the biomedical model. Pain was thought of as reflecting a physical problem. Early treatment approaches focused on the idea that something had to be removed or signals had to be interrupted to bring about relief from pain. The earliest reference to medication is referred to in an ancient Egyptian text dating back to 1550 BC where the goddes Isis recommended the use of opium to relieve Ra's headache (Bonica, 1990). Ancient pain treatments have included crocodile dung; oils derived from ants, earthworms, and spiders; spermatic fluid from frogs; and moss scraped from the skull of a victim of a violent death. A quick surf of the Internet will reveal that many equally esoteric preparations are endorsed readily even today.

The most common contemporary treatment approaches continue to be based primarily on the biomedical model of pain involving pharmacotherapy (with new drugs touted on an almost weekly basis) or surgery performed along nearly every site of the nervous system including the periphery (sympathectomy), the spinal cord (percutaneous cordotomy, destructive neural blockade), and the brain (thalamatomy, prefrontal lobotomy). Although current methods are more sophisticated than the ancient ones, the basic principles remain the same – alteration of an alleged physical cause should result in symptomatic improvement. Some modalities are quite helpful and a great many people suffering from acute and cancer pain have benefited from those advances in pain medicine and surgery. However, surgery and pharmacotherapy are not panaceas and do not result in adequate control of pain in all cases (Turk et al., 2011).

In contrast, some primitive treatment methods focused on the cause of pain being purely psychogenic. The assumption was that if no physical cause could be detected, then the pain must be caused by emotional disturbance that required treatment and once psychological problems were resolved, the pain would resolve. Such a psychogenic bias could be seen in the fourth version of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV; American Psychiatric Association, 1994). Two types of pain disorders were listed as mental disorders: "Pain Disorder Associated with Psychological Factors" and "Pain Disorder Associated with Both Psychological Factors and a general medical condition." In both cases, pain is the primary complaint, with psychological factors being considered important in the onset, severity, exacerbation, or maintenance of the pain. A somewhat presumptuous assumption underlying these disorders is that the 'appropriate' physical etiology of chronic pain must be identifiable (e.g., free from psychological factor) or patients are suffering from a psychiatric disorder. The assumption ignored the current consensus that pain is not a pure sensory experience but is inherently a biopsychosocial phenomenon resulting in an intense, subjective experience of discomfort. The current version of the DSM (DSM-V; American Psychiatric Association, 2013), therefore, acknowledges that there is a lack of evidence that such distinctions can be made with reliability and validity, and takes the approach that individuals with chronic pain would be appropriately diagnosed as having somatic symptom disorder, with predominant pain. For others, psychological factors affecting other medical conditions or an adjustment disorder would be more appropriate.

Comprehensive Approaches to the Treatment of the Person with Chronic Pain

The recognition of pain, particularly chronic pain, as a major health care issue, has resulted in the development of a growing number of specialists and specialized facilities offering treatments for pain over the last few decades. Comprehensive, multidisciplinary treatments involving teams of professionals (physicians, psychologists, physical therapists, occupational therapists, vocational therapists, and nurses) have been developed for the treatment of patients with the most recalcitrant problems. Flor et al. (1992) summarized the major characteristics of these patients as those with 7 years of pain history and at least one failed surgery. These patients are likely to be unemployed, receiving disability payments, and present an array of psychosocial and motivational issues. In addition to the suffering associated with chronic pain, these patients present socioeconomic difficulties for society due to loss of productivity and costs associated with disability payments and medical care.

In the following, the focus will be on the pain management approaches that have been developed based upon the psychological models described earlier. Although these interventions can be used separately, with few exceptions (e.g., biofeedback for tension-type headache) unimodal approaches are less effective than comprehensive rehabilitation programs (Turk et al., 2011). Typically, the psychological approaches are part of more comprehensive rehabilitation programs in which psychologists work together with therapist from other disciplines (e.g., physicians, physical therapists).

The multidisciplinary rehabilitation programs are philosophically distinct from unimodal medical and surgical treatments. The first and most critical distinction is that the major goals of the multidisciplinary treatment programs extend beyond pain relief. Rather, they focus on physical and psychological functioning as well.

Overall, rehabilitation programs that incorporate a significant psychosocial component are most effective in improving patient functioning, returning patients to work, reducing use of analgesics, reducing health care utilization, and reducing disability costs (Okifuji et al., 1998). 'Cure' of pain (e.g., total relief from pain) is, unfortunately, rarely attained. Reduction of pain is usually only moderate but comparable to that observed with usual medical and surgical approaches that are more invasive and have increased likelihood of iatrogenic consequences (Turk et al., 2011).

Treatment Based on the Operant Model

As noted previously, the operant model of pain focuses upon pain-related behaviors believed to be maintained by social reinforcement. Such behaviors are considered to be maladaptive and thus need to be eliminated. As discussed, reinforcers are the consequences of behaviors that increase the likelihood that behaviors will be repeated. Thus, treatment involves eliminating the rewarding consequences of pain behaviors and increasing the positive responses for activity and other well behaviors. In these programs, moaning, grimacing, and other pain behaviors are ignored. Usual activities of daily living and functional exercises are prescribed and positively reinforced. Medications are made time-contingent, rather than prescribed on an 'as needed' (prn) basis that is believed to reinforce positively the pain behaviors. Simultaneous with the attempts to extinguish pain behaviors, operant treatments are designed to help patients acquire a set of new, more adaptive, behaviors. Quota-based exercise programs with gradually increasing functional activities form the core of operant treatment.

Although operant factors may play an important role in the maintenance of disability, the model has been criticized at two levels. First, it fails to integrate factors other than reinforcement, such as sensory, emotional, and cognitive factors, in the overall pain experience. Second, the assumption that pain behaviors are acquired, maintained, and extinguished solely through environmental reinforcement contingencies is questioned. For example, physical signs, patients' self-efficacy beliefs, and depression are also reportedly related to pain behaviors (Buckelew et al., 1953).

Pain behaviors have been considered to be maladaptive manifestations of pain, guided by an incentive for attention or avoidance of physical activity. However, pain behaviors may be functional if indeed the behaviors protect patients from further injury or exacerbation of pain. Determination of whether an overt pain behavior in a given patient is functional or maladaptive needs to be based upon careful assessment of various factors associated with his or her pain conditions not just the reinforcement contingencies. Despite the criticisms, operant-based treatments have been reported to be successful in reducing overt pain behaviors, increasing well behaviors, and decreasing analgesic medication (e.g., Thieme et al., 2006).

Treatments Based on the Cognitive-Behavioral Model

As noted earlier, the cognitive-behavioral model of pain acknowledges the importance of cognitive variables interacting with sensory, affective, and behavioral factors to establish, maintain, and exacerbate the pain experience. The nature of specific techniques may vary from program to program; however, the primary goals of the cognitive-behavioral approach are relatively uniform (i.e., enhancement of patients' sense of control over their symptoms, increased use of adaptive skills to cope with pain and stress). Cognitive-behavioral therapy is designed to assist patients to identify, evaluate, and correct maladaptive conceptualizations and dysfunctional beliefs about themselves and their predicament. Additionally, patients are taught to recognize the connections linking cognition, affect, and behavior along with their joint consequences. Patients are encouraged to become aware of and to monitor the impact those maladaptive thoughts may have in the maintenance and exacerbation of maladaptive behaviors (see **Table 2** for common maladaptive thoughts). The central emphasis of the cognitive-behavioral approach is on patient self-management.

The cognitive-behavioral approach consists of four interrelated phases. These include (1) reconceptualization, (2) acquisition of coping skills and self-management strategies, (3) skill consolidation, and (4) generalization and maintenance (Flor and Turk, 2011; Turk et al., 1983). The first phase, reconceptualization, uses cognitive restructuring, a method that encourages people to identify and change maladaptive

Table 2 Common examples of cognitive errors

Overgeneralization: Extrapolation from the occurrence of a specific event or situation to a large range of possible situations. 'This coping strategy did not work; nothing will ever work for me'
Catastrophyzing: Focusing exclusively on the worst possibility regardless of its likelihood of occurrence. 'My back pain means my body is degenerating and falling apart'
All-or-none thinking: Considering only the extreme 'best' or 'worst' interpretation of a situation without regard to the full range of alternatives. 'If pain is not completely gone, I cannot do anything right'
Jumping to conclusions: Accepting an arbitrary interpretation without a rational evaluation of its likelihood. 'The doctor did not return my call today. He thinks I am a hopeless case'
Selective attention: Selectively attending to negative aspects of a situation while ignoring any positive factors. 'Physical exercises only make my pain worse'
Negative predictions: Assuming the worst. 'I know I will not get better even with all these therapies and everyone will dislike me for that'
Mind Reading: Make arbitrary assumptions about others without finding out what others are thinking. 'My husband does not talk to me about my pain because he does not care about me'

thoughts and feelings that are associated with the experience of pain. The crucial element in successful treatment is bringing about a shift in the patient's thought processes, away from well-established, habitual, and automatic but maladaptive thoughts toward more hopeful and rational ones. Cognitive restructuring helps foster the reconceptualization by helping patients to become aware of the role thoughts and emotions play in potentiating and maintaining stress and pain. The aim of this phase is to combat the sense of demoralization that many with chronic pain experience. Generally, the process of cognitive restructuring begins with a presentation of a situation or event that provoked a pain-related response. The situation is dissected to identify key thoughts and feelings that precede, accompany, and follow an episode or exacerbation of ongoing pain and pain-related problems. Then, patients are encouraged to challenge the legitimacy of the thoughts – Was it true? Was it reasonable? Was it the only way to respond? What alternatives are available? and so on. Patients are instructed to gather evidence for or against their own maladaptive automatic thoughts. Alternatives are discussed, with the suggestion that different ways of thinking can affect mood, behavior (e.g., reducing physical activity leading to greater disability), and even physiological activity (e.g., increase muscle tension and thereby exacerbate pain).

The second phase is the acquisition of self-management skills. A wide variety of techniques have been shown to be effective for reducing suffering and disability. Some of these strategies are self-regulatory skills (e.g., relaxation, controlled breathing, and attention diversion) that allow pain sufferers to regulate their own physiological responses that may be involved in the maintenance and exacerbation of pain. Other self-management strategies include stress-reduction skills (e.g., problem solving, behavioral rehearsals) that allow people with chronic pain to effectively manage the stress-inducing thoughts, behaviors, and emotions that trigger pain, emotional distress, and other maladaptive responses. Instead of being a passive recipient of a medical intervention (e.g., medication, anesthetic nerve block), patients now learn to use self-management strategies to play an active role in managing the myriad of problems created by the presence of persistent pain. Research has suggested that there is no one specific coping skill that best manages pain and disability (Fernandez and Turk, 1992). It is recommended generally that patients with chronic pain should be taught various types of coping skills to help them acquire a range of options.

Phase 3 is skill consolidation. During the skill-consolidation phase, patients practice and rehearse the skills that they have learned during the acquisition phase and apply them outside the hospital or clinic. Practice may start with mental rehearsal, during which patients imagine using the skills in different situations. Therapists can make use of role-playing, in which patients rehearse learned skills in situations that mirror their home environments. Therapists may start with a relatively easy example and then introduce scenarios that are progressively more realistic. The importance of skill consolidation through home practice cannot be overstated. When patients practice skills at home, it is useful for them to record their experiences including any difficulties that arise. Once problems associated with using the newly acquired skills are identified, these become targets for further discussion.

The final stage is phase 4: the preparation for generalization and maintenance. To maximize the likelihood of maintenance and generalization of treatment gains, therapists focus upon the cognitive activity of patients as they confront problems throughout treatment (e.g., failure to achieve goals, plateaus in progress, recurrent stress). These circumstances are used as opportunities to assist patients to learn how to handle setbacks and lapses because they are probably inevitable parts of life and will occur after the termination of the treatment. In the final phase of treatment, discussion focuses on ways of predicting and dealing with symptoms and related problems following treatment termination. Patients are encouraged to anticipate future problems, stress, and symptom-exacerbating events during treatment and to plan how to respond and cope with these problems.

Since self-initiating pain management is a key factor in pain rehabilitation, some type of cognitive behavioral approach generally is included in a multidisciplinary pain program. Cognitive-behavioral approaches have been demonstrated to be effective with a wide range of debilitating pain syndromes including low back pain (Gatchel and Rollings, 2008), arthritis (Astin et al., 2002), fibromyalgia (Thieme et al., 2006), and temporomandibular disorders (Turner et al., 2006). Moreover, these methods have been shown to be effective with children (Palermo, 2011) and geriatric samples (Lunde et al., 2009).

Summary

This article reviewed the definition of pain and the historical context within which pain has been conceptualized, made a distinction between nociception (a sensory process) and pain (a perception), and described the most common conceptualizations based on the unidimensional and multidimensional models of pain. It was noted that treatment strategies have closely followed the conceptualizations of pain and have included unidimensional modalities (i.e., biomedical, psychogenic) and progressed to multidimensional modalities.

Although pain generally is considered a physical phenomenon, pain involves various cognitive, affective, and behavioral features. These psychological factors are important not only in determining the perception of pain, but also defining disability and patients' general well-being. It should be clear from the review that pain has three main components, physical, psychosocial, and behavioral, that interact to define the unique pain experience. Because the pain experience is subjective and idiosyncratic, it cannot be understood without evaluating how patients perceive and appraise their conditions. A complete clinical picture involves consideration of how patients view their plight. By understanding the phenomenology of chronic pain and disability, effective treatment can be planned to alleviate persistent and debilitating pain, and improve physical and psychological functioning, thereby reducing the disability that accompanies chronic pain.

See also: Chronic Illness, Psychosocial Coping with; Health Research, Qualitative; Injuries and Accidents: Psychosocial Aspects; Mental Health; Pain, Health Psychology of.

Bibliography

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders: DSM-iv. American Psychiatric Association, Washington, DC.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Association, Washington, DC (DSM-5®).

Anon., 1995. Chronic Pain Management Programs: A Market Analysis. Marketdata Enterprises, Tampa, FL.

Astin, J.A., Beckner, W., Soeken, K., Hochberg, M.C., Berman, B., 2002. Psychological interventions for rheumatoid arthritis: a meta-analysis of randomized controlled trials. Arthritis and Rheumatism 47, 291–302.

Bonica, J.J., 1990. The Management of Pain. Lea & Febiger, Philadelphia.

Buckelew, S.P., Murray, S.E., Hewett, J.E., Johnson, J., Huyser, B., 1953. Self-efficacy, pain, and physical activity among fibromyalgia subjects. Arthritis Care Research 8, 43–50.

Cioffi, D., 1991. Beyond attentional strategies: cognitive–perceptual model of somatic interpretation. Psychological Bulletin 109, 25–41.

Deyo, R.A., 1986. Early diagnostic evaluation of low back pain. Journal of General Internal Medicine 1, 328–338.

Fernandez, E., Turk, D.C., 1992. Sensory and affective components of pain: separation and synthesis. Psychological Bulletin 112, 205–217.

Flor, H., Fydrich, T., Turk, D.C., 1992. Efficacy of multidisciplinary pain treatment centers: a meta-analytic review. Pain 49, 221–230.

Flor, H., Turk, D.C., 2011. Chronic pain: An Integrated Biobehavioral Approach. IASP Press, Seattle, WA.

Fordyce, W.E., 1976. Behavioral Methods in Chronic Pain and Illness. Mosby, St. Louis.

Gureje, O., Von Korff, M., Simon, G.E., Gater, R., 1998. Persistent pain and well-being: a World Health Organization study in primary care. The Journal of the American Medical Association 280, 147–151.

Gatchel, R.J., Rollings, K.M., 2008. Evidence-informed management of chronic low back pain with cognitive behavioral therapy. Spine Journal 8, 40–44.

Hart, L.G., Deyo, R.A., Cherkin, D.C., 1995. Physician office visits for low back pain. Frequency, clinical evaluation, and treatment patterns from a U.S. national survey. Spine 20, 11–19.

Jensen, M.C., Brant-Zawadzki, M.N., Obuchowski, N., Modic, M., Malkasian, D., Ross, J., 1994. Magnetic resonance imaging of the lumbar spine in people without back pain. New England Journal of Medicine 331, 69–73.

Lunde, L., Nordhus, I.H., Pallesen, S., 2009. The effectiveness of cognitive and behavioral treatment of chronic pain in the elderly: a quantitative review. Journal of Clinical Psychology in Medical Settings 16, 254–262.

Melzack, R., Wall, P.D., 1965. Pain mechanisms: a new theory. Science 150, 971–979.

Merskey, H., 1986. Classification of chronic pain. Descriptions of chronic pain syndromes and definitions of pain terms. Pain–Supplement 3, S1–S226.

Okifuji, A., Turk, D., Kalauokalani, D., 1998. Clinical outcome and economic evaluation of multidisciplinary pain centers. In: Block, A.R., Kremer, E.F., Fernandez, E. (Eds.), Handbook of Pain Syndromes: Biopsychosocial Perspectives. Erlbaum, Mahwah, NJ, pp. 77–98.

Palermo, T.M., 2011. Cognitive Behavioral Therapy for Chronic Pain in Children and Adolescents. Oxford University Press.

Parker, J.C., Smarr, K.L., Buckelew, S.P., Stucky-Ropp, R., Hewett, J., Johnson, J., Wright, G., Irvin, W., Walker, S., 1995. Effects of stress management on clinical outcomes in rheumatoid arthritis. Arthritis and Rheumatism 38, 1807–1818.

Thieme, K., Flor, H., Turk, D.C., 2006. Psychological treatment in fibromyalgia syndrome: efficacy of operant behavioural and cognitive-behavioural treatments. Arthitis Research and Therapy 8, R121.

Turk, D.C., Meichenbaum, D., Genest, M., 1983. Pain and Behavioral Medicine: A Cognitive–behavioral Perspective. Guilford, New York.

Turk, D.C., Okifuji, A., 2001. Pain terms and taxonomies of pain. In: Loeser, J.D., Butler, S.H., Chapman, C.R., Turk, D.C. (Eds.), Management of Pain, third ed. Lippincott Williams & Wilkins, Philadelphia, pp. 17–25.

Turk, D.C., Okifuji, A., Sinclair, J.D., Starz, T., 1998. Interdisciplinary treatment for fibromyalgia syndrome: clinical and statistical significance. Arthritis Care and Research 11, 186–195.

Turk, D.C., Wilson, H.D., Cahana, A., 2011. Treatment of chronic noncancer pain. Lancet 377, 2226–2235.

Turner, J.A., Mancl, L., Aaron, L.A., 2006. Short- and long-term efficacy of a brief cognitive-behavioral therapy for patients with chronic temporomandibular disorder pain: a randomized, controlled trial. Pain Forum 121, 181–194.

Williams, D.A., Thorn, B.E., 1989. An empirical assessment of pain beliefs. Pain 36, 351–358.

Chronology, Stratigraphy, and Dating Methods in Archaeology

Steven Shackley, University of California, Berkeley, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 1789–1796, © 2001, Elsevier Ltd.

Abstract

Estimating age in the archaeological record is the primary step in understanding the human past. Chronology provides a temporal dimension that distinguishes archaeology, paleontology, and geology from the disciplines of ethnography and ethnology. Increasingly better-defined methods of dating have radically enhanced our ability to address questions of cultural identity and ethnicity, as well as cultural change; an understanding of the primary methods that date the past is still critical to interpreting social process. The facility to correlate an event or episode with an absolute point in time is crucial in solving many archaeological problems. A variety of methods can be used to establish chronology in archaeology. Foremost among these are stratigraphic techniques, dating based on animal and plant remains, chemical accumulation, and temperature effects, as well as radiometric and radiation dating. The various techniques are applicable under certain conditions, and yield variable results based on context and dated material. All these dating techniques are focused on generating an absolute or relative chronology as the first step in understanding the past.

Introduction: Estimating Age in the Archaeological Record

Of the twenty or so dating methods employed in twenty-first-century archaeology, the US Congressional Office of Technology Assessment (OTA) report on *Technologies for Prehistoric and Historic Preservation* (US Congress, OTA, 1986) listed only seven and highlighted two radiocarbon and archaeomagnetic techniques; see Taylor 2000, pp. 75–60. In many regions of the world, radiocarbon, dendrochronology, obsidian hydration, and archaeomagnetism are the most common physical dating methods used, presumably in that order, but often together as a group providing redundant checks on one of the most important goals of archaeology – providing a framework of time, a chronology, upon which a rational reconstruction of the past can be built. Regardless of an archaeologist's theoretical or political perspectives, the first goal is to order regional prehistoric or historic sites, a single site, or a portion of the site into meaningful cultural and chronological segments. Chronology is a primary goal, but should not be confused with the intent of reconstructing the past, regardless of what an archaeologist's definition of the ultimate purpose may be.

During the formative cultural historic perspective in archaeology in the first half of the twentieth century, constructing regional and site-oriented chronologies became the primary goal of archaeology, particularly in North America and the United Kingdom (Trigger, 1989). A dissatisfaction with this rather unilineal explanation of the past stirred archaeologists to look beyond simply constructing chronologies and to begin to understand process and social relationships in the past. Dating methods and the construction of cultural chronologies, however, have remained the basement upon which archaeologists build their understanding of social process through time.

While it would be optimal to discuss all the potential dating methods available to archaeologists, this is not possible in a short synopsis. The most commonly used methods, and those showing promise, will be covered here (see **Table 1** and **Figure 1**). Refer to the publications in the bibliography for more detailed treatments.

Archaeological Stratigraphy

Stratigraphic relations have always been the primary method to infer the relative age of artifacts within a site. Stratigraphy is defined here as the study of the spatial and temporal relationships between the sediments and the soil (Waters, 1992: pp. 60–61). Much of this section's discussion can be more robustly understood by referring to the entry *Geoarchaeology*; an awareness of the geological processes of archaeological site formation is crucial to an understanding of dating archeological sediments. Indeed, the methods used by Quaternary geologists to date sediments include most of the methods used by archaeologists and discussed here (see **Figure 1**). Stratigraphic studies of archaeological sites are designed to define objectively and categorize the sediments and soils, the contact units between them, and the amount of time they represent, as well as their relationship to the surrounding sediment history.

Archaeological stratigraphy is based on the geological concepts of the law of superposition, which states that older sediments are emplaced at a lower level than more recent sediments. Therefore, sites, features, and artifacts residing in lower levels are, by definition, older than those in upper levels. This conceptual framework in archaeology appears to have begun with John Frere in 1797 attempting to understand the relationship between stone axes in a sedimentary sequence in England: "The manner in which ... [the hand axes] lie would lead to the persuasion that it was a place of their manufacture and not of their accidental deposit ... It may be suggested that the different strata were formed by inundations happening at distant periods" (Frere in Rapp and Hill, 1998: p. 5).

This idea that features or artifacts found within a site in the same stratigraphic level are contemporaneous forms the foundation upon which archaeological stratigraphy is based.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.13026-0

Table 1 Datable materials and appropriate dating methods in common use in early twenty-first century

	Magnetism	Hydration	Amino Acid	E.S.R.	Luminescence	Fission Track	Uranium Series	K-Ar; $^{40}Ar/^{39}Ar$	Radiocarbon	Dendrochronology
Obsidian	4	4				4		4		
Other volcanic rocks	4			4	4	4	4	4		
Glass						4				
Unburned sediment	4				4					
Burned sedimentary stone				4	4					
Slag					4	4			4	
Pottery, baked earth	4			4	4				4	
Shell			4	4			4		4	
Tooth enamel			4				4		4	
Bone, antler, ivory, teeth			4				4		4	
Wood/charcoal									4	4
Plant fragments									4	

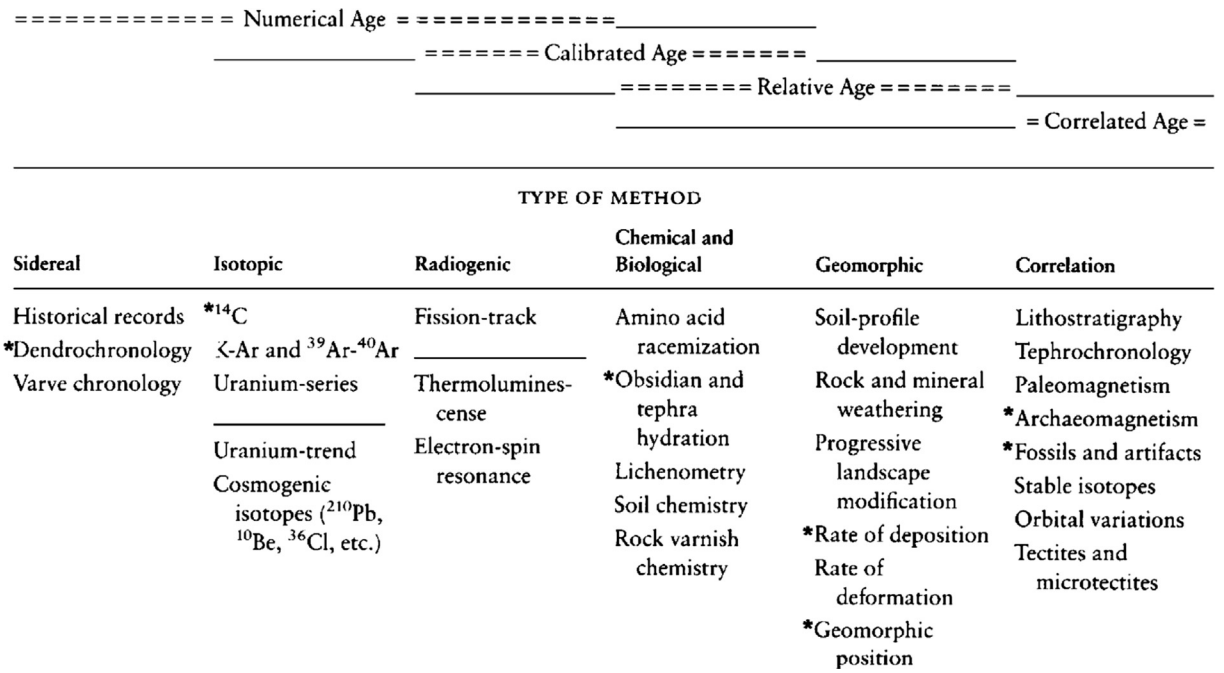

Figure 1 Commonly used dating methods in archaeology. Those marked with (*) are discussed in some detail in this article. Source: Adapted from Waters, M.R., 1992. Principles of Geoarchaeology: A North American Perspective. University of Arizona Press, Tucson, AZ.

In early twenty-first century archaeology, the various relative and absolute dating methods such as radiocarbon or archaeomagnetic dating are used to verify this assumption and produce a site chronology (see **Figure 2**). **Figure 2** shows a simple stratigraphic profile of a single excavation unit with the relative positions of natural and cultural features, and the position of radiocarbon dates recovered from the various levels. Note that the radiocarbon dates indicate that, in general, the stratigraphy is intact in that the oldest dates are at the bottom and the latest dates at the top, except for one date (CAMS 43177) at 3690 ± 50 BP which seems to be out of sequence and suggests 'small-scale' disturbance probably caused by the digging of the pit structure during the later ceramic period (Shackley et al., 2000).

Time-sensitive artifacts, such as previously dated pottery or projectile point types, can be used as *index fossils* within the stratigraphic column to create a relative chronology within the site. The stratigraphic example discussed above also used index fossils, in this case projectile point types, to add another chronological piece of information. In Level 8, where the early radiocarbon date as well as some more recent radiocarbon dates were recovered, Middle Archaic as well as Late Archaic projectile point forms were recovered; both confirming the time range suggested by the radiocarbon date ranges (**Figure 2**, Shackley et al., 2000). In this instance the stratigraphy, radiocarbon dates, and index fossils all contributed to an understanding of the site chronology and the integrity of the archaeological deposit.

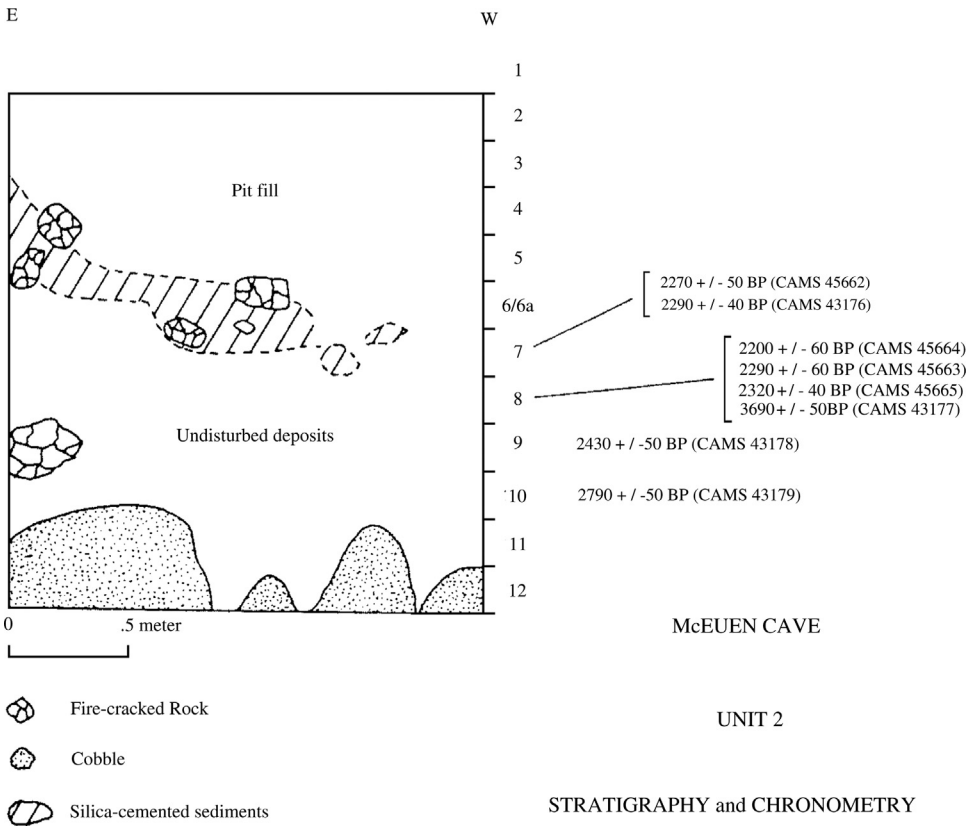

Figure 2 Simple stratigraphic wall profile of a unit at McEuen Cave (AZ W:13:6 ASM) showing position of bedrock, undisturbed strata, pit fill structure, and the position of recovered radiocarbon dates. Source: Shackley, M.S., Huckell, B.B., Huckell, L.W., 2000. Late Preceramic Farmer/Foragers at the Foot of the Mogollon Rim: The McEuen Cave Archaeological Project Testing Report (AZW:13:6 ASM). Report prepared for the Bureau of Land Management, US Department of Interior, Safford Area, AZ.

Constructing Stratigraphic Chronologies: Competing Ideas

It is important to note, however, that there is considerable recent controversy over the proper method to employ in the interpretation of archaeological stratigraphy (Farrand, 1984; Harris et al., 1993; Stein, 1987; Waters, 1992). Archaeological stratigraphic codes similar to geological stratigraphic codes and methodologies have been proposed but not generally adopted, since many geoarchaeologists are convinced that the existing geological code is adequate for archaeological research (see Stein, 1987). More recently, Harris has proposed an 'archaeological tool' to understand archaeological stratigraphy called the 'Harris Matrix,' essentially a simple postexcavation way in which to understand the relationships between stratigraphic units in a single diagram as reflected on paper, or more recently in digital form (see Harris et al., 1993). Farrand (1984) and others, mainly geologists, have questioned Harris's assumption that archaeological stratigraphy is primarily culturally conditioned rather than geologically determined. In a Harris Matrix analysis the relationship between stratigraphic units is based mainly on index fossils, for example the relationship between artifacts and features, rather than geological features or context – essentially emphasizing content rather than structure. Despite criticism from many geoarchaeologists, the Harris Matrix analysis has become increasingly popular with archaeologists working in varieties of geological and cultural contexts worldwide, particularly in Europe. The growing popularity of the Harris Matrix approach will probably continue, particularly in small site settings where a larger geological view is difficult to obtain or not necessarily relevant.

Sidereal Dating Methods

Sidereal dating methods include historical records, glacial varve dating, and dendrochronology. While only dendrochronology will be discussed here, glacial varve dating is quite useful in the upper latitudes, and in some cases may be quite precise. Historical records are a basic dating tool in historic and ethnohistoric archaeology, in addition to the social information they may contain.

Dendrochronology

The science that uses annual tree rings for dating past events and reconstructing past environmental conditions has undergone explosive growth in recent decades (Dean, 1997). While dendrochronology as an archaeological dating tool enjoyed tremendous expansion in the 1990s, particularly in western North America, Europe, Siberia, and the eastern Mediterranean, it is used wherever appropriate trees occur and a sequence has been established. Most importantly,

directly dated tree rings have been instrumental in the calibration of the radiocarbon timescale discussed below. This process indicated that the radiocarbon chronology underestimates the true ages of materials older than 2000 years and that ^{14}C dates must be corrected.

The technique is based on the concept that each year (four seasons) a tree will accumulate one growth ring, and that that ring's attributes reflect the specific climatic regime of that year. Since each year is essentially a unique climatic record, the attributes of the tree ring (thickness of inner and outer bands) are correspondingly unique. In trees, particularly fast-growing conifers, the progression of rings from pith to circumference presents an 'unalterable temporal order, and the production of but one ring per year provides the incremental regularity necessary to establish a fixed [and absolute] time scale' (Dean, 1997: p. 34). More than 180 tree and shrub species worldwide possess the attributes required for successful dendrochronological studies: visible and unambiguous ring definition, production of a set number of rings (generally only one) per year, mainly climate-controlled growth, and the presence of useable morphological features that allow for ring comparison. Cross-dating, matching patterns of ring variation among trees, is the necessary principle of dendrochronology. A sequence within a region is derived from overlap between cut trees as well as archaeological specimens, sometimes with great time depth into the thousands of years. While this process is apparently simple, the application necessarily requires precision. At the turn of the twenty-first century, image analysis technology was being used to develop electronic workstations, including image databases in laptop computers, perform routine comparisons between the database and archaeological wood samples similar to techniques used in obsidian hydration analysis.

Today, in a number of regions of the world (upper altitudes in the North American Southwest, Western Europe, the Eastern Mediterranean) dendrochronology forms the backbone of chronometric dating. When available, dendrochronology provides an absolute chronology that anchors other dating techniques.

Isotopic Dating Methods

Perhaps the one area of dating that has seen the greatest advances recently are techniques based on radioactive decay: ^{14}C, K-Ar, and $^{40}Ar/^{39}Ar$. Advances in accelerator mass spectrometry and laser fusion have propelled these techniques into the forefront of the arsenal of methods used to deal with time in the archaeological record. Importantly, ^{14}C is generally restricted to dating the last 50 000 years and $^{40}Ar/^{39}Ar$, previously restricted to time periods near and over 1 million years, are now nearly overlapping with $^{40}Ar/^{39}Ar$ using the laser fusion method and yielding younger and younger ages. These methods do have limitations, but the organic and mineral restrictions are being transcended almost daily.

Radiocarbon Dating (^{14}c)

Given that the greatest level of human activity occurred after the evolution of modern Homo sapiens, in the last 45 000 years or so, radiocarbon dating has become the most commonly relied upon dating method in archaeology (Taylor in Taylor and Aitken, 1997). Radiocarbon dating, now in its fifth decade of general use, is a primary tool used by archaeologists and Quaternary geologists to date the past.

The Radiocarbon Method

There are three principal isotopes of carbon which occur naturally: ^{12}C, ^{13}C (both stable), and ^{14}C (unstable or radioactive). The radiocarbon method is based on the rate of decay of the radioactive or unstable carbon isotope 14 (^{14}C), which is formed in the upper atmosphere through the effect of cosmic ray neutrons upon nitrogen 14. The reaction is: $^{14}N + n \Rightarrow {}^{14}C + p$ (where n is a neutron and p is a proton).

The ^{14}C formed is rapidly oxidized to $^{14}CO_2$ and enters the Earth's plant and animal lifeways through photosynthesis and the food chain. Plants and animals which utilize carbon in biological food chains take up ^{14}C during their lifetimes. They exist in equilibrium with the ^{14}C concentration of the atmosphere; that is, the number of ^{14}C atoms and nonradioactive carbon atoms stays approximately the same over time. As soon as a plant or animal dies, they cease the metabolic function of carbon uptake; there is no replenishment of radioactive carbon, only decay.

Libby, Anderson, and Arnold (Taylor and Aitken, 1997) first discovered that this decay occurs at a constant rate. They found that after 5568 years half the ^{14}C in the original sample will have decayed, and that after another 5568 years half of that remaining material will have decayed, and so on. The half-life ($t_{1/2}$) is the name given to this value, which Libby measured at 5568 ± 30 years. This became known as the Libby half-life. After 10 half-lives, there is a very small amount of radioactive carbon present in a sample. At about 50 000 to 60 000 years, then, the limit of the technique is reached (beyond this time, other radiometric techniques must be used for dating, such as $^{40}Ar/^{39}Ar$, as discussed above). By measuring the ^{14}C concentration or residual radioactivity of a sample whose age is not known, it is possible to obtain the number of decay events per gram of carbon. By comparing this with modern levels of activity (1890 wood corrected for decay to AD 1950) and using the measured half-life, it becomes possible to calculate a date for the death of the sample.

It follows from this that any material which is composed of carbon may be dated. Herein lies the true advantage of the radiocarbon method – it is able to be uniformly applied throughout the world. A list of the tremendous quantity of organic material that can be dated by radiocarbon is available at the University of Waikato's radiocarbon Web site, a standard for understanding the technique: http://c14.sci.waikato.ac.nz/webinfo/int.html.

Calibration and Radiocarbon Dating

As mentioned above, there are a number of effects that can cause errors in the measurement of radiocarbon dates. For example, shell, in constant contact with more recent atmospheric carbon, will generally yield young dates; conversely a shell artifact deposited in older limestone sediments will obtain a much older date than its actual death. These reservoir effects are in part mitigated by the use of various calibration algorithms, such as the CALIB (2001) and OxCal (2001) programs, both available online.

When a radiocarbon lab returns a date from a sample such as 5568 BP, it does not mean that it dates to 3619 BC, because the

true half-life of radiocarbon is 5730 years, and, more importantly, the proportion of radiocarbon in the atmosphere has varied through time, as discussed above. So the calibration utilities are written to allow for differential in the absorption of ^{14}C by different materials (i.e., marine shell versus wood charcoal), and to allow for different atmospheric effects. Using the CALIB 4.2 calibration, a radiocarbon assay of 5568 BP with a 1 standard deviation of 55 years on wood charcoal yields a date of:

$$68.3\left(1\sum\right)\text{cal } 4454-4416 \text{ BC}$$

$$\text{probability distribution } 0.409$$

$$4408-4354 \text{ BC } 0.591$$

Note that there are two dates with ranges of a number of years. The range includes the one standard deviation, and the two dates are due to multiple intercepts on the calibration curve. We can be 68.3% certain that the dates fall either from 4454 to 4416 BC or from 4408 to 4354 BC. There are multiple possible dates because the radiocarbon date of 5568 BP intercepts the calibration curve at more than one point. Due to the perturbations in the absorption of radiocarbon over the millennia, the variance is sometimes full of 'wiggles' on the curve, so placement can occur at two or more points.

Radiocarbon dates are reported in a standardized method, as shown in Table 2. The convention calls for reporting the provenience of the sample, the laboratory number, the radiocarbon age, and then the calibrated (in this case dendrocalibrated) age at one or two standard deviations (note that most of these dates yielded multiple intercepts). The calibration method used and any further contextual information should also be supplied.

Accelerator Mass Spectrometry (AMS) and Radiocarbon

From the inception of radiocarbon dating, ^{14}C ages of samples were calculated by decay counting in mainly scintillation counters. This requires a relatively large sample, depending on the amount of carbon remaining in that sample. By the late 1970s a number of researchers discovered that when accelerating sample atoms in the form of ions to much higher energies in particle accelerators, a much smaller sample was required to derive confident dates – in most cases only milligrams instead of tens of grams for scintillation counting. Both cyclotron and tandem accelerator mass spectrometers have been used to accomplish this, with tandem accelerators becoming the most popular. One additional advantage of acceleration is that the 'stripping process' disassociates all molecular species with the result that carbon isotopes can be isolated, and contamination minimized. AMS ^{14}C dating theoretically may push the time frame back to 100,000, effectively overlapping $^{40}Ar/^{39}Ar$ laser fusion dating (Taylor and Aitken, 1997).

Radiation Dating Methods

As with most of the other types of dating methods, radiation dating methods have undergone tremendous advances in the last decade, although many of these methods remain somewhat controversial in their applications. Electron spin resonance (ESR) and thermoluminescence (both based on the accumulation of trapped electrons in minerals), and fission track dating (produced when alpha particles are created by spontaneous fission of ^{238}U leaving a damage trail), have all had their detractors, but have recently gained recognition as techniques that are useful in dating materials not readily possible with the more accepted technology, for example glass, tooth enamel, and pigments (see Table 1). Archaeomagnetic dating has gained wide acceptance in many regions of the world and deserves some discussion.

Archaeomagnetic Dating

Archaeomagnetic and paleomagnetic dating both rely on the phenomenon of the frequent and predictable shifts in the Earth's magnetic polarity in space and time. The premise for the method states that the Earth's magnetic poles wander (show

Table 2 Recent pre-3000 BP (RCY) maize and cucurbit AMS ^{14}C dates from the American Southwest showing reporting conventions

Site	*Site type/context*	*Lab number*	*Radiocarbon age BP (1σ)*	*Dendrocalibrated 2σ age ranges BP*[1]
McEuen Cave, AZ	rock shelter	CAMS 4317	3690 ± 50	4151 (4056–3934) 3835
Los Posos, AZ	overbank alluv.	7 CAMS 34923	4050 ± 50	4792 (4517–4442) 4411
Three Fir Shelter, AZ	rock shelter	BETA 26275	3610 ± 170	4417 (3885) 3457
Tornillo Shelter, NM	rock shelter	GX 12720	3225 ±240	4076 (3439–3408) 2784
Cerro Juanaqueña, CHIH[3]	trincheras	INSTARR 3985	3310 ± 60	3685 (3548–3479) 3364
Lukachukai, AZ	pit house	AA 9317	3445 ± 45	3829 (3684–3543) 3508
Lukachukai, AZ	pit house	AA 9321	3050 ± 50	3379 (3315–3212) 3004
Lukachukai, AZ	pit house	AA 9319	3135 ± 45	3443 (3375–3357) 3209
Bat Cave, NM	rock shelter	A 4167	3010 ± 150	3550 (3205–3083) 2778
Bat Cave, NM	rock shelter	A 4189	3060 ± 110	3473 (3317–3214) 2894

[1]The 2σ dendrocalibrated minimums (intercepts) and maximums are based on a bidecadal correction in calendar years BP, rounded to the nearest decade, from Stuiver and Reimer (1993, version 3.0.3c, Method A). The maize dates are not corrected for C4 carbon uptake and can actually be up to 240 years older than reported here assuming –10 0/00. 2 This date and context is somewhat suspect, but was associated with a similar charcoal date at Los Pozos in the Tucson Basin (Freeman A.K.L., 1997. Middle to Late Holocene stream dynamics of the Santa Cruz River, Tucson, Arizona: Implications for human settlement, the transition to agriculture and archaeological site preservation. Ph.D. thesis, University of Arizona, Tucson, AZ, see also Gregory, D.A., Adams, J.L. (Eds.), 1999. Excavations in the Santa Cruz River Floodplain: The Middle Archaic Component at Los Pozos. Center for Desert Archaeology, Tucson, AZ.). 3 This date on a *Cucurbita sp.* is probably from a wild species. However, three specimens of maize were dated to over 2800 RCY BP (Hard, R.J., Roney, J.R., 1998. A massive terraced village complex in Chihuahua, Mexico, 3000 years before present. Science 279, 1661–1664.).
Source: Shackley M.S., Huckell B.B., Huckell, L.W., 2000. Late Preceramic Farmer/Foragers at the Foot of the Mogollon Rim: The McEuen Cave Archaeological Project Testing Report (AZW:13:6 ASM). Report prepared for the Bureau of Land Management, US Department of Interior, Safford Area, AZ.

secular variation) or flip (reverse direction), and that these variations provide a temporal fingerprint that can be detected in rock and sediments. Iron dipoles in minerals in soft sediments align themselves parallel with the Earth's magnetic field at the time of deposition until the sediment solidifies, for example as baked clay in a hearth. This creates detrital remnant magnetism whose direction and intensity precisely reflects the Earth's magnetic field at the time of deposition; and by matching their magnetic orientation to a master record, it is possible to derive a relatively accurate date. This master curve has been derived for the last 10 000 years. Archaeomagnetism is useful in dating material, as appropriate, when other useful dating material, appropriate for other techniques is absent.

Chemical, Temperature, and Water Affected Dating Methods

Some of the more controversial dating methods are those that are based upon changes in chemicals, temperature, and/or water for calibration. Amino acid racemization held great promise for dating organic materials, but is not generally reliable, while obsidian hydration dating, which was also promising, has become equally problematic.

Obsidian Hydration Dating

Obsidian, a quenched rhyolite glass is common along plate boundaries and volcanic arcs where crustal remelting has occurred. A glassy brittle rock with remarkable cutting properties, it has been used throughout human history in the production of stone tools (Shackley, 1998). Obsidian is a noncrystalline glass, and therefore a disordered substance, and moves toward an ordered state by crystallization or, more precisely, perlitization (Friedman et al., 1997; Stevenson et al., 1998). This is accomplished by the absorption of water and a breaking of silicon and aluminum bonds, ultimately devitrifying the glass and forming perlite. As the process proceeds, an absorption front and hydration rim forms that theoretically occurs at a regular, linear rate through time. Measuring this rim with a petrographic microscope in microns theoretically yields an absolute date. On the surface, this would seem to be a remarkable method, resolving a number of dating issues in archaeological contexts in which organic datable material does not occur, but obsidian artifacts are abundant. More recently, the intervening negative effects of variable temperature and humidity through time have been found to influence the rate at which glass will hydrate. Due to this a number of researchers have rejected the method *in toto* (see Morgenstein et al., 1999; Ridding, 1996). Still, a number of researchers have continued to utilize the method, and it has a number of devoted followers, even in the face of discrepancies between radiocarbon dates and obsidian hydration dates in the same stratigraphic position. Stevenson, and others have attempted to derive intrinsic hydration models based on the premise that each single obsidian nodule contains a different proportion of water than any other, and this will produce ages compatible with other dating methods (see Stevenson et al., 1998). Most do agree that obsidian hydration is a useful relative dating method which can be used to determine the extent of mixing in a stratigraphic column; given the model, one expects to see larger rim measurements at the bottom of a site than at the top. If this is not the case, then stratigraphic mixing is indicated. Obsidian hydration holds great promise, but most archaeologists are hesitant to include it in the arsenal of absolute dating methods.

Bibliography

CALIB Radicarbon Calibration Software (2001). http://depts. washington.edu/qil/calib /.

Carver, M.O.H., 1990. Digging for data: Archaeological approaches to data definition, acquisition and analysis. In: Francovich, R., Manacorda, D. (Eds.), Lo scavo archaeologica: della diagnosi all'edizione. Edizioni all'Insegna del Giglio, Firenze, Italy, pp. 45–120.

Colcutt, S.N., 1987. Archaeostratigraphy: A geoarchaeologist's viewpoint. Stratigraphica Archaeologica 2, 11–18.

Dean, J.S., 1997. Dendrochronology. In: Taylor, R.E., Aitken, M.J. (Eds.), Chronometrie Dating in Archaeology. Kluwer/Plenum Publishers, New York, pp. 31–64.

Farrand, W.R., 1984. Stratigraphic classification: Living within the law. Quarterly Review of Archaeology 5, 1–5.

Freeman A.K.L., 1997 Middle to Late Holocene stream dynamics of the Santa Cruz River, Tucson, Arizona: Implications for human settlement, the transition to agriculture and archaeological site preservation. Ph.D. thesis, University of Arizona, Tucson, AZ.

Friedman, I., Trembour, F.W., Hughes, R.E., 1997. Obsidian hydration dating. In: Taylor, R.E., Aitken, M.J. (Eds.), Chronometric Dating in Archaeology. Kluwer/ Plenum Publishers, New York, pp. 297–322.

Göksu, H.Y., Oberhofer, M., Regulla, D., 1991. Scientific Dating Methods. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Gregory, D.A., Adams, J.L. (Eds.), 1999. Excavations in the Santa Cruz River Floodplain: The Middle Archaic Component at Los Pozos. Center for Desert Archaeology, Tucson, AZ.

Hard, R.J., Roney, J.R., 1998. A massive terraced village complex in Chihuahua, Mexico, 3000 years before present. Science 279, 1661–1664.

Harris, E.C., Brown III, M.R., Brown, G.J. (Eds.), 1993. Practices of Archaeological Stratigraphy. Academic Press, London.

Herz, N., Garrison, E.G., 1998. Geological Methods for Archaeology. Oxford University Press, Oxford, UK.

Morgenstein, M.E., Wicket, C.L., Barkatt, A., 1999. Considerations of hydration-rind dating of glass artefacts: alteration morphologies and experimental evidence of hydrogeo-chemical soil-zone pore water content. Journal of Archaeological Science 26, 1193–1210.

OxCal Radiocarbon Calibration Software (2001). http://units.ox.ac.uk / departments / rlaha / orau/06_frm.htm.

Rapp Jr., G., Hill, C.L., 1998. Geoarchaeology: The Earth Science Approach to Archaeological Interpretation. Yale University Press, New Haven, CT.

Ridding, R., 1996. Where in world does obsidian hydration work? American Antiquity 61, 136–148.

Shackley, M.S. (Ed.), 1998. Archaeological Obsidian Studies: Method and Theory. Kluwer Academic/Plenum Publishing, New York.

Shackley, M.S., Huckell, B.B., Huckell, L.W., 2000. Late Preceramic farmer/foragers at the foot of the Mogollon Rim: The McEuen Cave Archaeological Project Testing Report (AZW:13:6 ASM). Report prepared for the Bureau of Land Management, US Department of Interior, Safford Area, AZ.

Stein, J.K., 1987. Deposits for archaeologists. Advances in Archaeological Method and Theory 11, 337–395.

Stevenson, C.M., Mazer, J.J., Scheetz, B.E., 1998. Laboratory obsidian hydration rates: Theory, method, and application. In: Shackley, M.S. (Ed.), Archaeological Obsidian Studies: Method and Theory. Kluwer Academic/Plenum Press, New York, pp. 181–204.

Stuiver, M., Reimer, P.J., 1993. Extended ^{14}C data base and revised CALIB 3.0 ^{14}C age calibration program. Radiocarbon 35, 215–230.

Taylor, R.E., 2000. Science-based dating methods in historic preservation. In: Williamson, R.A., Nickens, P.R. (Eds.), Science and Technology in Historic Preservation. Kluwer Academic/ Plenum Press Publishers, New York, pp. 75–96.

Taylor, R.E., Aitken, M.J. (Eds.), 1997. Chronometric Dating in Archaeology. Kluwer Academic/Plenum Publishers, New York.

Trigger, B.G., 1989. A History of Archaeological Thought. Cambridge University Press, Cambridge, UK.

US Congress, Office of Technology Assessment, 1986. Technologies for Prehistoric and Historic Preservation, OTA-E-319. US Government Printing Office, Washington, DC.

Waters, M.R., 1992. Principles of Geoarchaeology: A North American Perspective. University of Arizona Press, Tucson, AZ.

Chronotype

Ana Adan, Department of Psychiatry and Clinical Psychobiology, School of Psychology, University of Barcelona, Barcelona, Spain; and Institute for Brain, Cognition and Behavior (IR3C), University of Barcelona, Barcelona, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Chronotype (morning-, neither-, evening-type) has been defined in the last decades as an individual difference variable highly relevant for both health and well-being. In the adult population, about 60% of the individuals are neither-type, while the extreme groups have 20% each. Chronotypes differ in their biological and performance circadian rhythmicity, with time differences between morning- and evening-types ranging from 2 to 12 h. Chronotypes also differ in personality traits, health habits, and prevalence of several psychiatric disorders. In general, being evening-type may be a risk factor, especially during adolescence.

Chronotype or circadian typology (i.e., morning-, neither-, or evening-type) has been defined in the last decades as an individual difference variable highly relevant for both health and well-being. The distribution in the adult population is of around 20% for each extreme group and the remaining 60% in the neither-type group. Chronotype groups differ in their circadian functions both in their biological parameters (e.g., sleep–wake, body temperature, melatonin, and cortisol) and in their cognitive performance (e.g., attention, memory, and executive functions), with morning-types showing an advance from 2 to 12 h in the acrophase or peak time when compared to evening-type individuals. The differences among chronotypes can also be seen in the personality traits, health habits, and prevalence of several psychiatric disorders. Neither-type individuals usually are in an intermediate position, although more studies are needed with this larger group to draw stronger conclusions. Being a morning-type may serve as a protective factor for several outcomes, while being evening-type may be a risk factor, especially during adolescence. Thus, chronotype should be taken into account both in health prevention and promotion programs and in therapeutic approaches.

Introduction

Chronotype or circadian typology is an individual difference construct that currently holds the interest of many research groups, since it plausibly provides a more robust explanation of differences in individual functioning than other subjective characteristics, such as extraversion. It has a special relevance in the circadian rhythmicity, which has a 24 h periodicity approximately, synchronized to the light–darkness cycle. The chronotypes (morning-, neither-, or evening-type) show clear phase mismatches in the majority of biological parameters and in cognitive performance.

Morning-type individuals show more synchronicity with the light–darkness cycle, and the peak times (acrophase) of their circadian functions appear earlier – advanced phase – than in evening-types. Thus, morning participants usually go to sleep and wake up earlier and their activity and performance peaks appear in the morning, while evening participants go to sleep and wake up later and their activity and performance peaks happen in the evening. The neither-type are usually in an intermediate position between both extreme groups and are the most prevalent adult chronotype, representing about 60% of the population. Each extreme group (morning- and evening-type) holds 20% of the adult population (Adan et al., 2012). Most of the research done so far has only included participants from the extreme groups, so more studies are needed on the neither-type, not only due to their interest in generalizing the results to the populations, but also to draw the circadian characteristics of the neither-type since their pattern may be mixed and not simply intermediate. That is, sometimes similar characteristics are observed to morning-type and others similar to evening-type.

The implications of chronotype in the academic and work environments, as well as in the clinical area, are very important. Although chronotype is an individual difference variable and belonging to an extreme group should not be considered pathological, currently it is acknowledged that the evening-type shows more difficulty in synchronizing to the social and work time demands usual in our society, as well as more risk of developing nonhealthy habits and suffering diverse psychopathological conditions. In spite of the data collected so far, we have a long road ahead so that health professionals and society in general become sensitive to the need to take into account the meaning of chronotype in both preventive and clinical areas.

Measurement of the Chronotype

Chronotype is usually assessed through self-assessment questionnaires, which take into account factors such as preferred wake-up time, level of activation after getting up, moment of the day when the participant feels best, and time of evening when they feel tired and need to sleep. The 19-item morningness–eveningness questionnaire (MEQ) is the most widely used and validated instrument (Horne and Östberg, 1976). There is also a 5-item short form (reduced morningness–eveningness questionnaire (rMEQ); Adan and Almirall, 1991) with acceptable psychometric properties that is especially useful in the context of large-scale survey administration. The data collection of the rMEQ was carried out in a sample that included a wider age range and a diverse number of occupational groups, and has been adapted in several countries.

Table 1 shows the rMEQ and the correction criterion for each of its items. It allows using an open-question format in the items that ask for specific time (items 1, 3, 4) so that the

Table 1 Reduced morningness–eveningness questionnaire (rMEQ). The correction criterion appears in parentheses

Please read carefully each question before answering and choose only one among the several options. Answer all the questions in order of presentation and with the maximum sincerity. Each question should be answered independently of the rest, without modifying the previous answers.

1. Considering only your own 'feeling beaten' rhythm, at what time would you get up if you were entirely free to plan your day?
   a. 05.00–06.30 h (5)
   b. 06.31–07.45 h (4)
   c. 07.46–09.45 h (3)
   d. 09.46–11.00 h (2)
   e. 11.01–12.00 h (1)
2. During the first half-hour after having woken in the morning, how tired do you feel?
   a. Very tired (1)
   b. Fairly tired (2)
   c. Fairly refreshed (3)
   d. Very refreshed (4)
3. At what time in the evening do you feel tired and, as a result, in need of sleep?
   a. 20.00–21.00 h (5)
   b. 21.01–22.15 h (4)
   c. 22.16–00.45 h (3)
   d. 00.46–02.00 h (2)
   e. 02:01–03.00 h (1)
4. At what time of day do you think that you reach your 'feeling best' peak?
   a. 05.00–08.00 h (5)
   b. 08.01–10.00 h (4)
   c. 10.01–17.00 h (3)
   d. 17.01–22.00 h (2)
   e. 22.01–04.00 h (1)
5. Usually there are two kinds of people, those who are more active during the morning and those more active during the afternoon/evening. Which one of the two adjusts best to your functioning?
   a. Definitely more active in the morning (6)
   b. Quite more active in the morning (4)
   c. Quite more active in the afternoon/evening (2)
   d. Definitely more active in the afternoon/evening (0)

Reproduced from Adan, A., Almirall, H., 1991. Horne and Östberg morningness-eveningness questionnaire: a reduced scale. Personality and Individual Differences 12, 241–253.

participant can reply with the preferred specific hour and minutes. The total score is obtained by adding the score for each item and the three chronotypes are obtained considering the following cut-off scores: evening-type <12, neither-type between 12 and 17, and morning-type >17. In some countries, such as Italy, slight differences have been determined for these cut-off scores. Table 2 summarizes the main rhythmic characteristics of the chronotype groups according to their answers to the rMEQ and other aspects related to their circadian rhythmic expression.

There are other questionnaires to assess chronotype individuals developed subsequent to Horne and Östberg, all with good levels of reliability and adequate validity (see Di Milia et al., 2013 for a review). We can mention the Composite Scale of Morningness (CSM), the Preference Scale (PS), and the Munich Chronotype Questionnaire (MCTQ), among the more established in this field.

Table 2 Main rhythmic characteristics of the three chronotypes. The marked times are derived from the intervals determined by the options in the reduced morningness–eveningness questionnaire (rMEQ)

Morning-type
- wake up early before 7.45 h, feeling very refreshed
- best time of day before 10.00 h
- go to bed by 22.15 h
- define themselves as clearly or more active in the morning
- high regularity of the sleep schedule in working and leisure days
- sleep–wake rhythm very similar to the light–darkness cycle
- more difficulty in adapting to shift work and jet-lag

Neither-type or intermediate-type
- wake up between 7.46–09.30 h, feeling quite refreshed
- best time of day between 10.00–16.00 h
- go to bed between 22.16–00.30 h
- define themselves as more active in the morning
- medium regularity of the sleep schedule in working and leisure days
- sleep–wake rhythm quite similar to the light–darkness cycle
- intermediate capacity for adaptation to shift work and jet-lag

Evening-type
- wake up later than 09.30 h, feeling quite or very tired
- best time of day from 16.00 h on
- go to bed later than 00.30 h
- define themselves as clearly more active in the afternoon/evening
- low regularity of their sleep schedule in working versus leisure days
- sleep–wake rhythm not very similar to the light–darkness cycle
- less difficult adaptation to shift work and jet-lag

Organization of the Circadian System

The chronotype is determined by a different organization of the endogenous biological clock, located in the suprachiasmatic nucleus (SCN) of the hypothalamus. Circadian rhythmic expression is entrained to the environmental light information and, especially, to the cycle of light–darkness, which is the main *zeitgeber* (time giver) of the biological clock. The information from the environmental light is transmitted to the SCN through the retinohypothalamic tract. In addition, a number of other cues have been identified including food, locomotor activity, and social cues. The SCN works closely with the pineal gland, which produces melatonin, the darkness signal in the organism. The endogenous clock is extremely precise in its functioning with the ability to adjust to environmental changes at the fastest possible speed (Rosenwasser, 2009). Figure 1 shows the organization of the circadian system. The differences among chronotypes have been interpreted as coming from the evening-type individuals having a lesser dependency on the environmental synchronizers with respect to the morning-type, and consequently on having a more powerful hypothalamic circadian control.

In the expression of the chronotype there are underlying genetic differences, but also environmental factors (Adan et al., 2012). The genetic basis may account for about 50% of the observed variability, and there have been found associations between the chronotype and polymorphisms of different clock-genes such as the *Period* and the *Clock* genes (Barclay et al., 2010; Ojeda et al., 2013). Among the environmental factors affecting chronotype, we should highlight the perinatal photoperiod to which the individual is exposed during the first months of life and also the light conditions due to the

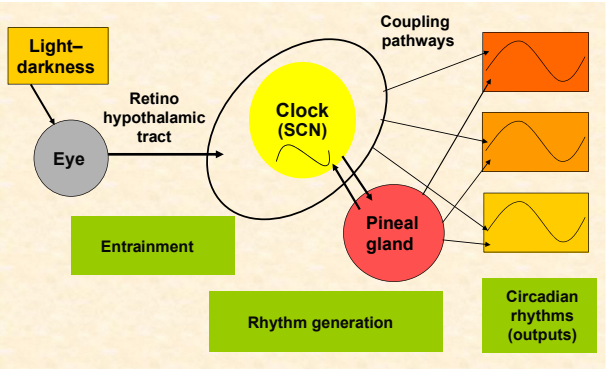

Figure 1 Circadian system organization with the endogenous clock in the suprachiasmatic nucleus (SCN) of the hypothalamus and their coupling with the principal external cues (light–darkness cycle).

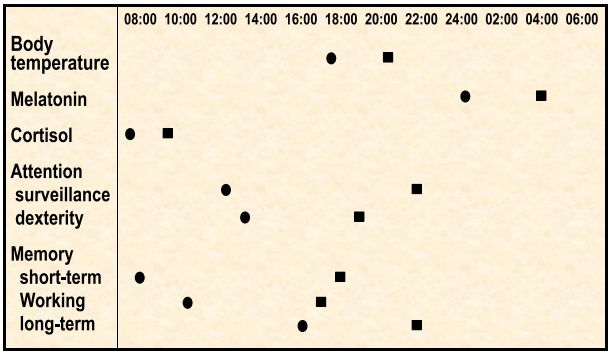

Figure 2 Differences in the peaks or acrophases of the circadian rhythmic functions of biological and cognitive performance markers between extreme chronotype groups (morning-type: ●; evening-type: ■).

longitude (east–west) and latitude (north–south) of the place of residence. The circadian organization of the individuals born within a short photoperiod (fall–winter) tends to be more morning-type, while that of those born within a long photoperiod (spring–summer) tends to be more evening-type. This is more clearly seen in men than in women. Although chronotype is a universal feature, depending on the country of residence, there appear differences in the observed distribution of the three chronotype groups. In a wide sense, we can state that there are more evening-type individuals as we move toward the west and the south.

Two key factors affecting the chronotype are age and sex (Adan et al., 2012; Randler, 2007). The morning chronotype is more frequent until age 10 and from age 50 on. Moreover, most of the studies hold that the tendency toward morningness increases with age. The age-related shift to morningness is observed after controlling for sociodemographic factors. During adolescence (12–17 years) there appears a pronounced tendency to eveningness and its decrease may be considered as a biological marker for the end of adolescence (Roenneberg et al., 2007). The population studies also find a higher evening-type prevalence in boys and men, and a higher morning-type in girls and women. The sex differences disappear following menopause in women, so it is hypothesized that such differences may be explained by the circamensual rhythmicity associated with the menstrual cycle.

Biological Circadian Rhythms

There are several outputs with circadian rhythmicity (biochemical, physiological, behavioral), and the most obvious one is the sleep–wake cycle. Among these, some are considered to be markers of the functionality of the circadian system, thanks to their robustness and implications in both health and well-being, mainly body temperature and the melatonin and cortisol hormones. In all the cases there are differences among chronotypes although depending on the parameter they are of a higher or lesser magnitude (see Figure 2).

Evening-type individuals go to sleep and wake up 2–3 h later than the morning-type, although there are no differences in the total amount of sleep if the individuals are free to choose when to sleep. The sleep–wake rhythm in the morning-type is very similar to the light–darkness cycle, while in the evening-type it shows a delay. Polysomnographic results (sleep quality) show longer sleep latency and delayed appearance of the first REM episode in evening individuals, who also overestimate the latency to fall asleep with subjective assessments. If the evening-types have to wake up early in the morning, they show more difficulties to fall asleep and a shorter sleeping time, together with higher day-somnolence levels and a stronger need to take a nap during the day. Thus, during the working days these individuals accumulate sleep privation and try to compensate for it during nonworking days, usually during the weekend, in which the time devoted to sleep increases and is delayed according to the hourly preference. This phenomenon has been termed 'social jet-lag' (Roenneberg et al., 2007). In contrast, the sleep times of morning-type individuals are more stable or rigid, without significant differences between working and nonworking days. This may be of advantage in environmentally stable schedule conditions but makes it harder to adapt to environmental schedule changes, such as working shifts and jet-lag situations.

Evening-types present a higher level of wrong beliefs and a lesser knowledge and practice of sleep hygiene, whereas morning-types show a more correct level of beliefs and practice of sleep hygiene. Being evening-type is a risk factor to develop a delayed-phase sleep syndrome, especially during adolescence of early adulthood as well as insomnia in adulthood. In contrast, being morning-type is a risk factor to present an advanced-phase sleep syndrome in the last decades of life.

Body temperature has a sinusoidal function whose maximum peak appears between 17 and 20 h and its minimum between 3 and 5 h in the morning. It is considered an estimator of brain activity and its disorganization is well known in several mood disorders, especially in endogenous depression. Morning-type individuals show an earlier circadian temperature phase around 2 h and have a lower amplitude than evening-types.

The melatonin hormone is considered to be the best marker of the endogenous circadian pacemaker. Melatonin presents a secretion with a marked night peak dependent on the duration of the environmental photoperiod, and with minimum values

during the wake period. It is essential to start and keep sleep as well as to produce an acceptable sleep quality (physical and psychological recuperation). Morning-type individuals present an advance in their melatonin peak of approximately 3 h with respect to the evening-type, without differences in amplitude. Cortisol circadian rhythm is characterized by a marked secretion peak near the time of awakening, associated to preparing for the diurnal activity, and a minimum peak during the first half of sleep. The morning-types present an advance of the cortisol peak of about 1 h and higher levels immediately after awakening in comparison to the evening-type.

Cognitive Performance Rhythms

Chronotypes show remarkable differences in the optimum execution moments (acrophase) in cognitive tasks (see Figure 2). Although there are no differences in cognitive performance among chronotypes when evaluated at their optimal times of day, the evening-types are in worse activation conditions during the conventional school or work time periods. We will classify the tasks depending on whether the execution is focused on attention, memory, or executive functions, since these present a diurnal evolution and clearly differentiated acrophases (Schmidt et al., 2007; Valdez et al., 2008).

Attention Tasks

We regard attention tasks as repetitive and monotonous tasks that involve perceptual and motor skills, and whose common characteristic is to prioritize reaction time. Among these we distinguish those based on surveillance and manual dexterity since they present different temporal patterns of execution.

In surveillance tasks, execution speed in the early hours is low and increases along the day, peaking at 19.00–21.00 h, while precision is higher in the early hours (08.00–10.00 h) decreasing along the day. The extreme chronotypes differ in the speed precision of these tasks with phase mismatches of around 7–12 h. The diurnal function for tasks of manual dexterity presents the best executions around midday (12.00–14.00 h), and neither the preceding increase nor the following decrease are so marked as in the surveillance tasks. However, this is the evolution in the morning- and neither-type groups, while for the evening-type the peak of performance happens in the evening (18.00–20.00 h).

Memory and Executive Functions Tasks

Memory tasks involve the presentation of a material that the participant must recall after a certain amount of time, or use it to reply to an exercise that he is presented, or freely narrate its content. These tasks are differentiated according to the capacity and duration of information storage. For the immediate or short-term memory tasks, the interval between presentation–retrieval is short, and the capacity of storage is limited (e.g., retrieving lists of digits or words). The optimal moment of execution is during the first hours of the morning (08.00–11.00 h) for the morning-type, while for the neither-type it happens during midday, and for the evening-type during the afternoon.

Working memory tasks require a longer duration of the information storage and a limited but larger capacity than in the short-term tasks (arithmetic, verbal reasoning, and comprehension). They present a diurnal evolution, in relation to the speed of increase in the morning until they reach their peak at midday (12.00–14.00 h), and then decrease significantly until the last hours of the day. In contrast, precision is superior in the first hours of the morning (09.00–11.00 h). In all the cases, the phase mismatches are observed among the groups, with optimal schedule advances in morning-types with respect to evening-types from 2 to 7 h.

Long-term memory tasks have intervals between presentation and retrieval ranging from several hours to days or even months. The duration of the storage of the information is long and of unlimited capacity, and the material is better remembered if it has been presented toward the end of the afternoon (18.00–20.00 h). This normative pattern is applicable to the neither-type, while it appears during the first hours of the afternoon if they are morning-type and during the first hours of the evening if they are evening-type.

There are fewer works assessing the differences among chronotypes in more complex cognitive tasks, such as executive function tasks. Executive functions include the initiative, inhibition, flexibility, planning, prevision, self-monitoring, verification, and correction components, which may modulate the performance of all tasks requiring decision-making or problem solving. The diurnal functions of executive functions have not been modeled yet, since it is harder to obtain repeated measures with these tasks. However, it can be stated that they meet the synchrony effect. This is, individuals who are more alert in the morning (morning-type) tend to perform better in the morning than in the afternoon/evening and the opposite in the case of individuals who are more alert in the afternoon/evening (evening-type).

The study of the possible relationship between chronotype and the academic outcomes in children, teenagers, and college students is of undoubted interest. In all cases advantage of the morning-type compared to evening-type was observed. Both the attention in class as academic achievement and grading are better in morning-type students (Preckel et al., 2011, 2013; Vollmer et al., 2013).

An interesting approach is to consider the cognitive style, understood as the preferred way of information processing (Adan et al., 2012). It has been observed that morning-type is associated with the left-thinking style and evening-type is associated with the right-thinking style (Fabbri et al., 2007). The left-thinker follows an analytic and sequential mode of reasoning and relies preferably on verbal–abstract representation, while the right-thinker tends to process information in an intuitive, holistic, gestalt-type, synthesized, and visual–motor way. Future research should look deeper into this approach, which will certainly produce interesting theoretical and applied data.

Personality Characteristics

With respect to personality traits, several works have found important associations with the scores in morningness–eveningness or with the chronotypes (Adan et al., 2012;

Tsaousis, 2010). Although none of these relations is in itself a disorder, it is true that all of them together in extreme cases may constitute risk factors for the appearance of several dysfunctional behaviors and mental disorders.

For the dimensions in the model proposed by Eysenck, there is a tendency in morning-type individuals to score toward the introversion pole, while the evening-types score toward the extraversion pole. Using the Big Five Model of personality (extraversion, agreeableness, conscientiousness, neuroticism, and openness), proposed by Costa and McCrae, most studies reported a strong positive association between morningness and conscientiousness and, to a lesser extent, with agreeableness. The observation that morning-types were more conscientious and higher in self-esteem than the evening-types suggest that conscientiousness rather than the extraversion dimension may be the personality factor that best distinguishes chronotypes. The association with the neuroticism constructs or anxiety trait is less obvious, although some works obtain higher scores in the evening-type participants.

More recently, using personality models with a biological basis, such as Cloninger's and Zuckerman's, has provided new and revealing associations for both adolescents and adults. The evening-type participants present higher scores in impulsivity and novelty seeking in comparison with the morning- and neither-types. In contrast, the morning-type showed higher scores in harm avoidance, persistence, self-directedness, cooperation, and activity.

Moreover, more data have been collected using different standardized questionnaires. With the Millon Inventory of Personality Types it has been observed that the morning-types prefer to obtain information from their surroundings based on concrete sources and direct experience, whereas evening-types prefer information based on intangible and abstract data. In addition, evening-types interpret the information using subjective and emotional parameters and tend to elaborate it with knowledge schemata that are newer and more original. The evening-types also express less affinity in their relationships with other people (acquiescence) and less respect for customs and social norms. Recently a more appropriate emotional intelligence pattern (Antúnez et al., 2013) and adaptive emotional profile with higher control, inhibition, and coping (Ottoni et al., 2012) has been observed in the morning-type than in the evening- and neither-types.

Mental Disorders

Mood Disorders

The chronotype is involved in both the prevalence and course of several mood disorders in which the impairment or even loss of the circadian rhythmicity is common. Most of the works in this area have found a higher prevalence of seasonal mood disorder and major depression in evening-type participants, as well as a more severe depressive symptomatology and suicidal ideation (Merikanto et al., 2013). Bipolar disorder is also more prevalent among evening-types, especially in type II. In all cases, remission and asymptomatic periods are associated to a proper reorganization and maintenance of the circadian rhythmicity.

Eating Disorders

When there is no pathology, morning-type participants tend to have their meals in more regular schedules and usually follow healthier diets. The evening-types, in addition to having more variable time patterns, also tend to 'snack' between meals and express a lesser preference to follow a 'Mediterranean' diet and to control the use of normal and spicy seasonings. The available data point at an increased risk for the morning-types to develop eating restrictions and anorexia nervosa, while the evening-types tend to present uncontrolled eating habits, with more perceived hunger and higher body mass index (a common behavior in bulimia). There is a higher proportion of evening-types among the patients diagnosed with bulimia, and remission of the symptoms shows a direct association with decrease of eveningness (Natale et al., 2008).

Drug Use and Dependency

Research by Adan (1994) observed a higher use of legal drugs (alcohol, caffeine, and nicotine) in evening-types when compared to the morning- and neither-types. Since then, several other works have confirmed the same pattern in different countries and with varied samples. Among them we highlight the study by Prat and Adan (2011), which found that the evening-type participants also consume more illegal drugs (e.g., cannabis, ecstasy, and cocaine), have more problems related to alcohol consumption and more severe hangover symptoms (learning problems, tiredness, headaches, sensorial hypersensitive, irritability, and thirst) than the morning-types. **Figure 3** shows the percentage of participants who consume different drugs, classified according to their chronotype.

In drug-dependent participants, the most studied drugs are alcohol and tobacco. These participants present flatter circadian rhythm functions and phase delays, which are related to the

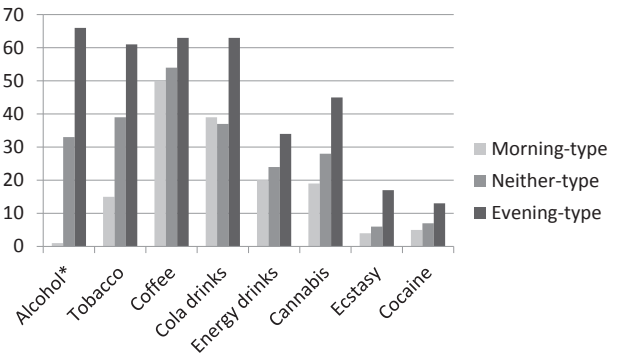

Figure 3 Percentage of drug-consuming participants according to chronotype (morning-, neither-, and evening-type). For the legal drugs excluding alcohol, the habitual consumption has been considered, while for the illegal ones the occasional intake has been taken into account, since it is much less frequent. * Hazardous alcohol use, defined as a pattern of consumption that places at risk for adverse health events (for the sample corresponds to binge drinking pattern). Data revised from the study by Prat, G., Adan, A., 2011. Influence of circadian typology on drug consume, hazardous alcohol use and hangover symptoms. Chronobiology International 28, 248–257.

degree of dependency. In severe cases, the circadian rhythmic expression may even disappear. During the processes of detoxication and dehabituation it is convenient to take into account the rhythmic alterations, since their resolution is associated to the maintenance of abstinence.

Other Disorders

It has been reported that adults suspected of having attention deficit/hyperactivity disorder tend to be in the evening-type (Voinescu et al., 2012), and that this relationship is stronger for the inattention factor than the impulsive hyperactive factor. There is also evidence of the association between the evening-types and higher levels of anxiety symptoms (Hsu et al., 2012), this being more present in the first hours of the day.

Adjusting the Clock: Preventive and Therapeutic Strategies

Preventive and therapeutic approaches should take into account the circadian rhythmic organization of the individuals, with special emphasis on redirecting the time patterns toward a better synchronization with the light–darkness cycle. This is more important in the more vulnerable periods of life, such as adolescence and old age: in adolescence in order to avoid the extreme evening-type pattern, and in old age in order to avoid the extreme morning-type pattern. Chronobiological strategies – regular time patterns, light therapy, and melatonin administration – are effective in several situations and have been proven to be secure even in special populations (Adan, 2013).

In many cases it may suffice to establish regular time patterns of wake–sleep, meals, and daily physical and social activity with a tendency toward a morningness pattern of functioning. Exposition to morning light or light therapy may be considered when scheduled habits alone are not enough to reorganize circadian rhythmic expression. The scientific evidence available on the nonvisual beneficial effects of light is unquestionable. Light therapy stimulates the pineal gland to synthesize an adequate peak of night melatonin for a better sleep–wake cycle tuned to the environmental signals. In areas with an adequate environmental light, it is recommended to get exposure to natural daylight from half an hour to 2 h daily, preferably in the morning. When exposition to natural light is not possible or becomes insufficient, bright light therapy may be used instead. This consists in the exposition to artificial light of total spectrum at an intensity of 10 000 lux, with a standard prescription of application of at least 30 min daily. Finally, the administration of exogenous melatonin at the end of the day is another possibility to consider in order to favor the proper expression and organization of the circadian system.

See also: Addictions: General Considerations; Health Behaviors; Hypothalamus; Mental Disorders, Epidemiology of; Personality Assessment; Personality and Health; Personality, Biological Models of; Sleep Disorders across the Life Span.

Bibliography

Adan, A., 1994. Chronotype and personality factors in the daily consumption of alcohol and psychostimulants. Addiction 89, 455–462.

Adan, A., 2013. A chronobiological approach to addiction. Journal of Substance Use 18, 171–183.

Adan, A., Almirall, H., 1991. Horne and Östberg morningness-eveningness questionnaire: a reduced scale. Personality and Individual Differences 12, 241–253.

Adan, A., Archer, S.N., Hidalgo, M.P., Di Milia, L., Natale, V., Randler, C., 2012. Circadian typology: a comprehensive review. Chronobiology International 29, 1153–1175.

Adan, A., Natale, V., Caci, H., 2008. Cognitive strategies and circadian typology. In: Léglise, A.L. (Ed.), Progress in Circadian Rhythms Research. Nova Biomedical Books. Nova Science Publishers, Inc., New York, pp. 141–161.

Antúnez, J.M., Navarro-Humanes, J.F., Adan, A., 2013. Circadian typology and emotional intelligence in healthy adults. Chronobiology International 30, 981–987.

Barclay, N.L., Eley, T.C., Buysse, D.J., Archer, S.N., Gregory, A.M., 2010. Diurnal preference and sleep quality: same genes? A study of young adult twins. Chronobiology International 27, 278–296.

Di Milia, L., Adan, A., Natale, V., Randler, C., 2013. Reviewing the psychometric properties of contemporary circadian typology measures. Chronobiology International 30, 1261–1271.

Fabbri, M., Antonietti, A., Giorgetti, M., Tonetti, L., Natale, V., 2007. Circadian typology and style of thinking differences. Learning and Individual Differences 17, 175–180.

Horne, J.A., Östberg, O., 1976. A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. International Journal of Chronobiology 4, 97–110.

Hsu, C.Y., Gau, S.S.F., Shang, C.Y., Chiu, Y.N., Lee, M.B., 2012. Associations between chronotypes, psychopathology, and personality among incoming college students. Chronobiology International 29, 491–501.

Merikanto, I., Lahti, T., Kronholm, E., Peltonen, M., Laatikainen, T., Vartiainen, E., Salomaa, V., Partonen, T., 2013. Evening types are prone to depression. Chronobiology International 30, 719–725.

Natale, V., Ballardini, D., Schumann, R., Mencarelli, C., Magelli, V., 2008. Morningness-eveningness preference and eating disorders. Personality and Individual Differences 45, 549–553.

Ojeda, D.A., Perea, C.S., Niño, C.L., Gutiérrez, R.M., López-León, S., Arboleda, H., Camargo, A., Adan, A., Forero, D.A., 2013. A novel association of two non-synonymous polymorphisms in Per2 and Per3 genes with specific diurnal preference subscales. Neuroscience Letters 553, 52–56.

Ottoni, G.L., Antoniolli, E., Lara, D.R., 2012. Circadian preference is associated with emotional and affective temperaments. Chronobiology International 29, 786–793.

Prat, G., Adan, A., 2011. Influence of circadian typology on drug consume, hazardous alcohol use and hangover symptoms. Chronobiology International 28, 248–257.

Preckel, F., Lipnevich, A.A., Schneider, S., Roberts, R.D., 2011. Chronotype, cognitive abilities, and academic achievement: a meta-analytic investigation. Learning and Individual Differences 21, 483–492.

Preckel, F., Lipnevich, A.A., Boehme, K., Brandner, L., Georgi, K., Könen, T., Mursin, K., Roberts, R.D., 2013. Morningness-eveningness and educational outcomes: the lark has an advantage over the owl at high school. British Journal of Educational Psychology 83, 114–134.

Randler, C., 2007. Gender differences in morningness–eveningness assessed by self-report questionnaires: a meta-analysis. Personality and Individual Differences 43, 1667–1675.

Roenneberg, T., Kuehnle, T., Juda, M., Kantermann, T., Allebrandt, K., Gordijn, M., Merrow, M., 2007. Epidemiology of the human circadian clock. Sleep Medicine Reviews 11, 429–438.

Rosenwasser, A.M., 2009. Functional neuroanatomy of sleep and circadian rhythms. Brain Research Reviews 61, 281–306.

Schmidt, C., Collette, F., Cajochen, C., Peigneux, P., 2007. A time to think: circadian rhythms in human cognition. Cognition Neuropsychology 24, 755–789.

Tsaousis, I., 2010. Circadian preferences and personality traits: a meta-analysis. European Journal of Personality 24, 356–373.

Valdez, P., Reilly, T., Waterhouse, J., 2008. Rhythms of mental performance. Mind Brain Education 2, 7–16.

Voinescu, B.I., Szentagotai, A., David, D., 2012. Sleep disturbance, circadian preference and symptoms of adult attention deficit hyperactivity disorder (ADHD). Journal of Neural Transmission 119, 1195–1204.

Vollmer, C., Pötsch, F., Randler, C., 2013. Morningness is associated with better grading and higher attention in class. Learning and Individual Differences 27, 167–173.

Church and State

Gerhard Robbers, Institute for European Constitutional Law, University of Trier, Trier, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.T.S. Madeley, volume 3, pp. 1796–1798, © 2001, Elsevier Ltd.

Abstract

Religion has manifold and differing political functions for different societies. These functions underlie and influence the relationships between religious institutions and the state on the level of the law; their role in history often forms an important part of the collective memory of a people and their identities. Today, it remains a foremost challenge to politics and international law to institutionalize universal foundational claims of freedom of religion or belief and equal treatment in the large variety of differing cultural contexts. While it is possible to identify three basic types of relationships between religions and the state – state religions, separation regimes, and cooperationist systems – these classifications are overlaid by social circumstances which suggest different groupings. Contemporary challenges often relate to migration, state aid to religions, religious autonomy, and pluralism.

Terminology

The term Church and State usually represents the relationships between religious organizations or authorities on the one hand and secular government on the other. Within political science it sometimes also stands, in a somewhat broader approach, for the relationships between religious organizations or authorities with the polity. In this sense, political parties and relevant actors of civil society come into the focus.

In the course of globalization and increasing sensitivity for cultural respect, the term has been questioned as discriminatory against non-Christian beliefs. Church is a genuinely Christian institution not used within the theological context of most other religions, although the Buddhist sangha is sometimes thought to share some features. In the other monotheistic religions, such as Islam and Judaism, and in Hinduism the status of religious authorities and their relation to secular political authority suggest quite different patterns from those typical of areas based in Christian traditions.

Furthermore, it has become quite common in public discourse, not to distinguish between churches and sects or cults, as these latter terms carry for many a somewhat pejorative connotation.

In order to avoid possible sensitivities it has been suggested to use the term Religion and State instead of Church and State. This also takes up the idea that in view of their relations to secular powers, religious institutions and authorities are based in the religious life of individuals. The idea of individual freedom of religion as a concept of human rights forms a basis of this approach; at least, the term Religion and State includes an individual aspect of religion which is less obvious in the more institutional, collective, or corporative approach visible in the term Church and State.

Asking for the relationships between religion and the state (or church and state) implies a general distinction or even a divide between the sacred, for which religion stands, and the secular, identified with the state. This again is a rather modern interpretation based primarily in Christian teaching, referring to the parable as to which one should render unto the emperor the things which are the emperor's and unto God the things that are God's.

Religion–state relationships have come into the focus of political science within recent decades; they have been a subject for legal scholarly work for long. Sociology looks at them primarily from the perspective of religious demography; they have also come to the attention of economists, mostly in regard of financing of religion and in view of religious rules on a proper economy. From the perspective of religions themselves, the relationships between religion and the state is an important matter of theology; their approaches range from promoting intense and direct intervention into the secular sphere to complete abstinence.

Functions of Religions

The relations between religions and states across the world are remarkably diverse. They reflect, among numerous other factors, differences of culture, history, religious demography, constitutional systems, and political circumstances. While it is often suggested that there should be one ideal system of religion–state relationships implemented throughout all states, others insist that the specific patterns of such relationships should be framed according to the diverse needs of the individual states. Differences often even occur between regions within the same state, reflecting developments in history or religious demography. This is the case in France, where in general a far-reaching separation between the state and religious organizations is promoted, while in the eastern parts of the country several religious organizations have a special public law status, receive direct financial support from the state, Christian religious symbols are depicted in public buildings, and the president of the republic appoints the Roman-Catholic bishops of the dioceses of Strasbourg and Metz; in some other parts of France again, different specific rules apply. Likewise, in the United Kingdom, the Anglican Church is the established church in England, but disestablished in Wales, while in Scotland the Calvinist Church of Scotland is the national church, and special regimes apply in several other parts of the

country. In Greece, Islamic law applies for the Muslim minority in Western Thrace. Such circumstances reflect the importance attached to religious needs, history, and cultural differences.

Religion has manifold and differing political functions for different societies. These functions underlie and influence the relationships between religious institutions and the state on the level of the law; their role in history often forms an important part of the collective memory of a people and their identities. Individual freedom of religion is a strong factor of national identity of the US. In China, autonomy of religions is often regarded as a threat to the unity of the country. In Tibet, the Dalai Lama represents for many the idea of a politically independent state. Islam provides a centerpiece of identity for countries of Muslim tradition in defining themselves against what is regarded as 'the West.' Israel defines itself in its Declaration of Independence as a Jewish state. Roman Catholicism is a key factor of national identity in Poland; the Roman Catholic religion and its institutional structures had held together the Polish people in the many partitions of the country between Christian Orthodox Russia and Protestant Prussia; it maintained this identity-providing function also in the times of communist rule and proved to be a decisive factor in regaining independence in the collapse of the Soviet Union to which it contributed. Protestantism emerged as a main factor in forming nation-states in Europe; the Protestant faith remained intrinsically intertwined with the monarchies in Scandinavia, the United Kingdom, the Netherlands, Switzerland, and parts of Germany. Central Europe has experienced 100 years of devastating wars between Protestants and Catholics in the sixteenth and seventeenth centuries. Religious wars culminated in the 30 Years War (1618–48) in Germany, the Huguenots War (1562–98) in France, and to some extent the British civil wars (1642–51). These historic events still form part of the collective memories of the relevant countries and influence the law until today.

Establishment and disestablishment of churches in the United Kingdom mirror self-consciousness and survival of national identities within the country. The synod of the Calvinist Church of Scotland proved to be the center of Scottish political self-identity for lack of a Scottish parliament since the Act of Union in 1707; only in 1999, a Scottish parliament was recreated.

For Ireland, the Roman Catholic denomination was a foremost factor of identity and resistance against the English rule over the country in the nineteenth and early twentieth centuries. Religious affiliations are still relevant within the political antagonisms as well as peace-building attempts in Northern Ireland.

In 1789 in France, the idea of 'la République' together with the principle of 'laïcité' carried a specific identity in promoting a distinct ethos and set of values that is hesitant if not hostile in relation to religion. It emerged in the fight against an overwhelmingly powerful Roman Catholic Church before the 1789 French Revolution and in the course of the nineteenth century.

In Christian Orthodox countries, Orthodox religion often serves as an ideological glue of the society and statehood. The existence of an autocephalous or at least distinct Orthodox Church underlines the independence of the country. It also is of dominant impact in the Russian Orthodox Church filling in the ideological vacuum after the collapse of communist ideology in Russia. In Greece, the Greek Orthodox Church preserved identity of the Greek people against Turkish Ottoman rule over centuries. Somewhat similar, in Serbia, the Orthodox Church as something like a state church in the Muslim Ottoman Empire protected Serbian identity.

The long list of political functions of religion in history and presence underlines that religion is not only a question of individual belief; religion always has a very public, political function for group and national identities and survival. Often times, people turn to the religious sphere and their institutions when they do not find political representation in the secular sphere.

The functions of religion are of ongoing impact. It is relatively easy to separate the state from religious institutions. Moreover, the state can refrain from identifying with a specific set of religious teaching. It is, however, probably impossible for the state to completely separate from religious ideas and traditions: Days of rest such as the Sunday, legal institutions such as matrimony, or basic assumptions such as human dignity are deeply entrenched in religious convictions. Protecting them (and not the holy days, institutions, or fundamental assumptions of other religions) by secular law may well mean some degree of identification; disregarding them is also taking sides and implies negative identification. Democracy entails that the individual can vote on secular issues by reasons of his or her individual religious beliefs; by this way religious convictions will influence the state. While some demand that state decisions should be taken completely independently from religious reasoning and that for every such decision there must be a secular reason, this would amount to an antireligious discrimination, and would probably be impossible, anyway.

Human Rights and Institutional Approaches

A large number of international instruments have taken up the relevance of religion for individuals as well as for the public sphere.

These international instruments mostly promote the idea of human rights. In view of the relationships between state and religious institutions, they, thus, contribute to a development that shifts the basis of these relationships from an institutional approach to an individual human rights approach. Today, a predominant point of reference for the relationships between state and religion is formed by the right to freedom of religion or belief as stated in Art. 18 of the 1948 UN-Universal Declaration of Human Rights: "Everyone has the right to freedom of thought, conscience and religion; this right includes freedom to change his religion or belief, and freedom, either alone or in community with others and in public or private, to manifest his religion or belief in teaching, practice, worship and observance."

Beyond the question of freedom of religion or belief, the issue of equal treatment of religions has become prominent in the field of religion–state relationships. While some claim identical treatment of all religions, others insist on respecting differences based in factors such as tradition, demographic facts, or social and cultural impact of the various religions.

It remains a foremost challenge to politics and international law to institutionalize universal foundational claims of

freedom of religion or belief and equal treatment in the large variety of differing cultural contexts.

Typology of Systems

It is possible to identify three basic types of relationships between religions and the state. The first basic type is characterized by the existence of a state religion or predominant religion. In this system, there are close links between state power and the existence of religious institutions. The systems of England, Denmark, Greece, Malta, and Finland belong to this basic category. On the other hand, there are systems founded on the idea of a strict separation of state and religious organizations, for instance, in the US, in France, with the exception of the three eastern départements and some other regions, and also in the Netherlands. There is to a great extent a legal separation in Ireland also. The third type features the basic separation of state and religions while simultaneously recognizing a multitude of common tasks, in the fulfillment of which state and religious organizations activity are linked: Belgium, Poland, Spain, Italy, Hungary, Austria, the Baltic States, and Portugal belong to this group. While the Australian constitution states – similar to the First Amendment to the United States Constitution – that the Commonwealth shall not make any law for establishing any religion, this has been interpreted as allowing for public funding of religious institutions and for closely cooperating with them. Also, the pattern of 'symphonía,' signifying an intense relationship of loyalty between the state and the church traditionally established in countries of Christian Orthodox tradition, figures mostly in the cooperationist approach.

These classifications according to legal and theoretical considerations are instantly overlaid and rendered questionable by social circumstances which suggest different groupings. The religious influence on the state in mainly Catholic Ireland is probably stronger and more direct than the wording of the constitutional provisions suggests. In the same way, there would be a closer similarity in the social relevance of religion as between Greece, Spain, and Italy than would be revealed in a comparison of Greece with Denmark or the United Kingdom. Likewise, it has been suggested that a 'civic religion' strongly influences US politics.

In some states, agreements between state and religious communities play an important role; in the context of the Roman-Catholic church the most important of those treaties, the concordats, have the status of international law treaties. Several states also conclude treaties with religious organizations as a matter of state law in order to meet the specific individual needs of the religions concerned and to regulate matters of common interest. Therefore, some speak of states with a covenantal system of the relations between state and religious organizations as a special category. However, the impact of such agreements must not be overestimated, important as they are; it seems that they mirror a system of cooperation rather than they would establish it.

A further refined classification of relationships between state and religions that looks from the viewpoint of freedom of religion or belief distinguishes between a large number of typical settings: theocratic states, in which the linkage between state and religion is so close that it is virtually impossible to distinguish state from religious rule; established religions systems in which there is an official state religion; religious status systems where the state recognizes the jurisdiction of a number of religious systems, typically in areas dealing with family law and inheritance; endorsed religions in which the state affirms that a particular religion has a special place in the country's history, culture, and tradition; preferred sets of religions in which multiple religions are favored; cooperationist systems distinguished by the fact that they take a neutral but positive and 'cooperative' stance toward religions in society, typically in fields such as religious education, various humanitarian services, or church finance; accommodationist regimes, accommodationism thought of as cooperation without any direct financial subsidies; separation regimes, covering a wide and diverse range of approaches from protecting religion from the state to protecting the state and society from excessive religious power; the approach of laïcité preferring a rigid separation; secular control regimes which try to exercise strict control over and minimize the influence of religious institutions; and finally abolitionist states in which the secular authorities aim at elimination religion altogether.

The relationships between states and religions are in constant flux. There seems to be a strong tendency to increasingly favor Islam in states with a Muslim tradition by establishing the sharia as the centerpiece of the legal system. In a situation of revolution and civil unrest, religiously inspired groups try to or succeed in taking over state power. In European states, a convergence of systems is predominant leading to a general regime of cooperation between states and religions. In some countries, the earlier antichurch and anticlerical attitudes faded as the centuries passed and their legal consequences are being gradually reduced. Religious communities are allowed greater freedom, and the conditions for meeting religious needs are created by the state. On the other hand, there are clear moves toward the disestablishment of the established churches; thus, Sweden has to a very large extent cut the formerly close ties between the state and the Lutheran Church, and reforming the previous state church systems is a general tendency in the Scandinavian countries.

In the US, emphasis is laid at least recently by the Supreme Court on acknowledging the right of autonomy of religious organizations. Such autonomy is also underlined by the European Court of Human Rights in interpreting the guarantee of freedom of religion or belief as stated in the European Convention of Human Rights, when the court insists that the autonomous existence of religious communities is indispensable for pluralism in a democratic society and is thus an issue at the very heart of the protection which freedom of religion or belief affords.

In many parts of the world, the state supports one or several religious organizations while discriminating against others. A foremost instrument in doing this is a system of registration of religious communities that imposes barriers for specific religions, mostly small or new ones.

Ancient or archaic religiopolitical systems are often understood to combine religious and political functions within a single structure of authority. Divine kingship or theocratic rule can be seen in many cultures, such as in the Roman Empire, when the emperors assumed a godly or godlike status.

Religious functions, however, were from early times usually distinct from functions of secular power.

Religious Approaches

In Christian cultures, the church was from early on conceived as a distinct institution. However, rulers often asserted over the church some sort of sacral or at least political authority, while popes or patriarchs claimed supreme authority also over the secular rulers. Numerous theories provided explanations for manifold approaches such as the medieval theory of the two swords dating back to the Pope Gelasius I, the two kingdoms doctrine based in the Protestant theology of Martin Luther, or the divine-right theory of kingship which assumed sovereignty of the monarch directly held by the grace of god; the Byzantine emperor Justinian I provided a classical description of the concept of symphonia strongly held in Eastern Orthodox Christian theological and political thinking, which claims that church and state are to be complementary and exhibit mutual respect.

Within cultures based in Islam, religious, and secular authority remained even more closely linked. It has been argued that this was because Medina under Muhammad was considered to provide a perfect model for the life of the community of believers. Within this view, the Prophet himself recognized the sovereignty of God alone, and, accordingly, it has not been open to his successors to claim sovereign authority in their own right; rather, it has been the duty of all who exercise authority on behalf of the community faithfully to implement the rules encapsulated in the Qur'an and the traditions of the Prophet as interpreted by legal scholars.

This interpretation sees parallels within Judaism, where the authority of divine law, as interpreted by rabbinical scholars, was held to be supreme.

In a historical perspective, differences between the world religions in their approaches to the relationships between the institutions of secular and sacral powers, while existing, seem to be less prominent than is often assumed. The different approaches to the status of religion vis-à-vis the secular sphere and its authorities seem to surface in any of the world religions from time to time according to the larger political circumstances of the time. Recently, the reassertion of theocratic claims has gained much support in different parts of the world.

Contemporary Challenges

Contemporary challenges to a large extent relate to the developments within states with a Muslim background and culture. Political unrest and revolutionary movements demand and often achieve to implement religious rules in the legal and social systems. Violent confrontation, terrorism, and military conflict are embedded in religious convictions. While this often leads to intense limitations of freedom of religion, the community building and peacemaking power of religion are only gradually perceived.

Migration, mostly also, but in no way exclusively, of Muslim background, challenges the traditional systems of religion–state relationships in many countries, as the newcomers call for equal treatment and accommodation of their religious needs and traditions. This often requires changes in deeply rooted traditions within the receiving society as well as among the newcomers. Points of conflict arise in many social and legal fields such as family and inheritance rules, food requirements, days of rest, burial norms, garment traditions, the construction of religious buildings, or the legitimacy of religious courts competing with state jurisdiction.

In some parts of the world, such as in Europe, former excitement about new and small religions has softened, while in others, such as in Asia, they are still seen as a matter of concern.

Involvement of religious organizations in matters of what is seen as secular politics is often regarded with skepticism or even complete rejection. While in some religious perspectives, such political involvement is seen as a necessary expression of religious activity and freedom of religion or belief, others regard it as an impermissible intrusion of religion into the secular sector.

The education of the young generation is a constant and predominant field of special attention in the relationships between religious organizations and the state. While many states allow for and even support educational institutions maintained by religious organizations, others try to suppress all religious influence on education. The systems of specific religious instruction in state run schools vary considerably. In some states there is complete absence of such instruction, others provide information on various religions, and in many states confessional teaching takes place usually conducted by the individual religions themselves.

Some states, like the US, are very reluctant to afford financial support to religious organizations. In these systems, religious organizations depend on their own assets or private donations. Most other jurisdictions have introduced systems of direct state financial aid to religions. A common feature is providing tax exemptions for religious organizations. In a number of regimes, religious organizations can participate in the state taxation system by having a part of the state income tax assigned to religious or charitable organizations such as Italy, Spain, and Hungary. In some states, religious organizations have the right to levy own taxes on their members such as in Germany, Austria, and Sweden or in a number of Swiss cantons; such church tax is not a state tax that favors certain religions, but the qualified religious organization itself is entitled to impose the tax. Usually, more than one of these approaches can be found in combination with one another. While some regard state financing of religions as a violation of separation between state and religious organizations, others insist that the state renders financial aid to many a group and must not exclude religious organizations, because otherwise the state would discriminate against religion. Critics of mere donation models often claim that this approach allows an undue influence on the religious organization by individual big donors.

A matter of political and legal controversy is often the status of religious and other personnel of religious organizations. Many religions throughout the world operate hospitals and homes, kindergartens, schools, or universities, and they maintain farms and economic enterprises. In doing so, religious organizations are in many cases large and powerful employers. Mostly, religious personnel such as priests, ministers, or members of religious orders who provide services for their

community enjoy a special status as their relation to the religious organization is not primarily regarded as an employment, but as a fundamentally spiritual relationship that requires special rules. Many jurisdictions, such as the US and most European countries, also provide for special regimes in the labor law that applies for religiously oriented institutions and which pays attention to their specific ethos. Usually, employees are expected to respect specific loyalty obligations to their employer which would often not or in a different shape apply in nonreligious institutions.

In such contexts, religious organizations contribute to a fundamental pluralism. They often offer alternatives in a life that is otherwise completely controlled by secular institutions and powers. While increasing antagonisms between competing religious groups in many parts of the world are seen as one of the predominant contemporary challenges, freedom of religion or belief can be endangered. On the other hand, freedom of religion or belief is often seen as one of the foundations of a democratic society. It is regarded as one of the most vital elements that go to make up the identity of believers and their conception of life, but also a precious asset for atheists, agnostics, skeptics, and the unconcerned. The pluralism indissociable from a democratic society, which has been dearly won over the centuries, depends in this view on it. It is mostly assumed that the state has legitimate regulatory power in this sphere and in its relations with the various religions, denominations, and beliefs. International instruments and many national jurisdictions hold that the state has a duty to remain neutral and impartial in exercising this regulatory power. Especially, when different religious groups come into conflict with each other, the preservation of pluralism and the proper functioning of democracy are seen to be at stake, one of the principle characteristics of which is held to be the possibility they offer of resolving a country's problems through dialogue, without recourse to violence, even when they are irksome. Accordingly, the role of the authorities in such circumstances is seen not to remove the cause of tension by eliminating pluralism, but to ensure that the competing groups tolerate each other.

See also: Civil Religion; Globalization and Religion; Personality and Religion; Political Religions: History of the Concept; Political Sociology; Psychiatry and Religion; Reformation and Confessionalization; Religion and Economic Life; Religion and Health: Clinical Considerations and Applications; Religion and New Media; Religion and Politics in the United States; Religion, Anthropology of; Religion, History of; Religion, Sociology of; Religion: Mobilization and Power; Religion: Nationalism and Identity; Religious Fundamentalism: Cultural Concerns; Religious Nationalism and the Secular State; Secular Religions; Secularization; Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns.

Bibliography

Boyle, K., Sheen, J. (Eds.), 1997. Freedom of Religion and Belief. A World Report. Routledge, London.

Durham Jr., W.C., 2011. Patterns of religion state relations. In: Witte, J., Green, M.C. (Eds.), Religion and Human Rights: An Introduction. Oxford University Press, Oxford, UK, p. 360.

Monsma, S.V., Soper, J.C., 1997. The Challenge of Pluralism. Church and State in Five Democracies. Rowman Littlefield, Lanhom, MD.

Robbers, G. (Ed.), 2005. State and Church in the European Union, second ed. Nomos, Baden-Baden, Germany.

Scharffs, B.G., Durham Jr., W.C., 2010. Law and Religion: National, International, and Comparative Perspectives. Aspen Publishers.

Temperman, J., 2010. State–Religion Relationships and Human Rights Law: Towards a Right to Religiously Neutral Governance. Martinus Nijhoff Publishers, Leiden/Boston.

Witte, J., 1999. Church and State: The American Constitutional Experiment. Westview Press, Greenwood, CT.

Cities and the Financial Crisis

Manuel B Aalbers, University of Leuven, Heverlee, BE, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

There is a long history in social science that connects urbanization to capitalism. This article discusses how the work of Manuel Castells and in particular David Harvey informs our understanding of the urbanization of financial crises in the United States and elsewhere. The concepts of capital switching and financialization are central to the argument presented here. The article also includes a discussion of the management of the crisis, focusing on both the failed institutional response to the financial crisis and the global and local grassroots responses to the financial crisis.

Introduction

There is a long history in social science that connects urbanization to capitalism. Perhaps two of the most prominent contributions to this nexus are to be found in the earlier work of sociologist Manuel Castells (1977, 1983) and both the early and contemporary work of human geographer David Harvey (1982, 1985, 2010), but it can also be found in older works such as Friedrich Engels's *The Condition of the Working Class in England* (1844/1999). For these and other authors, urbanization is a necessary condition for capitalist expansion. Another necessary condition of capitalism, according to the same authors, is that capitalism creates its own crises.

To Harvey, one fundamental contradiction in capitalism is that individual capitalists act in a way that, when aggregated, runs counter to their own class interests. This results in overaccumulation and, in time, a variety of crises. The early signs of crisis often include an overproduction of commodities, falling rates of profit, surplus capital (e.g., money capital lacking opportunities for profitable employment), and surplus labor or a rising rate of exploitation of labor. Harvey then distinguishes different kinds of crises: a partial crisis, which affects a particular sector, place, or set of mediating institutions; a switching crisis, which involves a major restructuring of capital flows either between sectors of the economy or between places; and a global crisis, which affects most sectors of the economy and most places within the capitalist system.

The financial crisis that started in 2007–08 and continues to drag on and mutate in various ways and places can be characterized as a global crisis since it affects most sectors of the economy and most places around the world (Aalbers, 2009). The fact that one can mention sectors or places that are not affected much by the crisis should not be seen as proof that this is not a global crisis. It is merely proof that some sectors and places have not yet been fully integrated into the global capitalist system or that they are seen as safe havens to switch investment to. This is 'only' the third global crisis of capitalism, following the one of the 1930s (with the Second World War as its aftermath) and the one that became most visible after 1973 but had been building throughout the 1960s.

The current global crisis had been building throughout the 1990s and early 2000s. Its earlier manifestations could be seen in more localized events such as the first subprime crisis in the late 1990s in the United States, the late 1990s Asian and Russian financial crises, the Argentine crisis at the dawn of the new millennium, and the explosion of the dot-com bubble in the early 2000s. These were all partial crises or switching crises because they hit only certain places or certain sectors of the economy and resulted in money capital being switched to different sectors and different places. Two sectors in particular became the main destinations for money capital to flow out of sectors and places in crisis: real estate and finance itself.

David Harvey: Capital Switching

In Harveyian terms, real estate is seen as an asset in which money can be invested and disinvested by directing capital to the highest and best uses, and by withdrawing and subsequently redirecting capital from low payoffs to potentially higher ones. Capital switching entails the flow of capital from the primary circuit (the production, manufacturing, and industrial sectors) to the secondary circuit, which includes real estate and infrastructure. (Harvey also identifies a tertiary circuit, the circuit of social infrastructure, which is identified as investment in technology, science, conditions of employees, health, and education.) Switching from the primary to the secondary circuit takes place when there is a surplus of capital in the primary circuit and signs of overaccumulation emerge. Capital switching is a strategy to prevent a crisis, but because of the inherent contradictions of capitalism, investment in the secondary circuit will only delay, not take away, the crisis. In empirical terms, the decline in production growth in many Organization for Economic Co-operation and Development (OECD) countries, the overaccumulation in oil-rich countries, the savings glut in countries like China, and the implosion of the dot-com bubble all resulted in capital switching into the real estate sector.

Building on David Harvey's theory, I have argued elsewhere (Aalbers, 2008) that the early 2000s may have seen the rise of capital switching to what I call the quaternary circuit of capital, where the financial sector is not at work to finance production (first circuit), real estate (second circuit), or the building up of welfare states (tertiary circuit), but simply the growth of money capital through finance. Financialization can be characterized as capital switching from the primary, secondary, or tertiary circuit to the quaternary circuit of capital. Financialization implies not only the financialization of existing economies (i.e., the restructuring of existing markets into producer and

consumer markets that are heavily tied to financial markets) but also the rise of financial markets for their own good; that is, the rise of financial markets not for the facilitation of other markets but for trade in money, credit, securities, and so on. Or, in the words of Cox (1992: p. 29), finance is increasingly becoming "decoupled from production to become an independent power, an autocrat over the real economy." The essence of a quaternary circuit is not that power shifts from the nonfinancial to the financial sector, but that financial and nonfinancial firms alike become increasingly involved in financial markets (Aalbers, 2008). And this is not a direct result of a lack of equity, but a direct result of financialization, which rewrites the rules of capital accumulation.

Like capital switching to the secondary circuit, capital switching to the quaternary circuit is liable to lead to overaccumulation and eventually to crisis. But capital switching to the quaternary circuit is in at least one way fundamentally different from capital switching to the secondary circuit: while the secondary circuit generally competes for financing in the capital market, the quaternary circuit represents the capital market as an investment channel in its own right. In that sense, financialization can be characterized as the capitalist economy taken to extremes: it is not a producer or consumer market, but a market designed only to make money. Although it can be argued that there exists a necessary link between production–consumption and financial speculation, this is not necessarily so: many of the recent crises have their roots in unstable income sources and overaccumulation, as both the Enron debacle and the current ongoing financial crisis demonstrate. In a finance-led regime of accumulation, risks that were once limited to a specific actor in the production–consumption chain become risks for all of the actors involved in a specific industry. In that sense, the quaternary circuit, which is in the first place a heuristic tool, is strongly attached to the other circuits of capital. Financialization as capital switching does not deny the necessary link between the different circuits, but, quite the contrary, it stresses the increased interdependence between the circuits of capital. In addition, it highlights how one of these circuits, the quaternary, comes to dominate the other circuits, to the disadvantage of those circuits (Aalbers, 2008).

To see what all of this means for cities, we rely again on the work of Manuel Castells, who identified four basic elements of a capitalist urban structure: production, consumption, exchange, and management (see also Rossi, 2010). Rather than discussing these four elements one by one, we here discuss the impact of the financial crisis on cities, first through a discussion of the link between housing markets and the financial crisis in the United States (including production, consumption, and exchange), second through a discussion of the impact of the crisis more generally speaking, and third through a discussion of the management of the crisis, focusing on both the failed institutional response to the financial crisis and the global and local grassroots responses to the financial crisis.

Subprime Lending in the United States

Real estate is, by definition, local as it is spatially fixed. Mortgage lending, however, has developed from a local to a national market and is increasingly a global market today (Gotham, 2006). An understanding of the financial crisis is ultimately a spatialized understanding of the linkages between local and global. Housing bubbles, faltering economies, and regulation together have shaped the geography of the financial crisis on the state and city levels in the United States. Subprime and predatory lending have affected low-income and minority communities more than others, and we therefore see a concentration of foreclosures not only in certain cities but also in certain neighborhoods (Aalbers, 2009).

The default and foreclosure crisis that was at the origins of the financial crisis has hit American households across the country, but people in some states and cities are more likely to be in foreclosure. The rise in default rates started some years ago in the Rustbelt where housing prices went down and unemployment up. The combination of lack of employment and falling housing prices is perilous as people who lose their job in a high-unemployment area not only have a smaller chance of finding a new job within a few months but also run a bigger chance of not being able to pay off their mortgage loan and might then be faced with negative equity. This implies that lenders will see higher foreclosures not only as a direct result of default but also because homeowners with financial problems in declining housing markets are less likely to sell their house, which would have enabled them to pay off their loan.

The rise of subprime lending started in the early 1990s with refinance loans, often in the poorer parts of Rustbelt cities. Increasing default rates led to a first subprime mortgage crisis in 1997–98, 10 years before the second subprime crisis (Ashton, 2009; Immergluck, 2009). Subprime mortgage lending has been growing fast, from about $35 billion (5% of total mortgage originations) in 1994 to $600 billion (20%) in 2006, 75% of which was securitized. The growth of subprime lending halted for a few years in the late 1990s as a response to the first subprime crisis, but picked up again after 2000 when subprime loans were no longer exclusively targeted at borrowers with low credit scores; 'exotic' mortgages were designed to be sold to middle-class borrowers, in particular in rapidly growing parts of the country. As a result, the fastest increases in defaults and foreclosures since 2007 were not in the Rustbelt but in the Sunbelt, where housing prices had been going up most and exotic mortgages with adjustable interest rates and teaser rates were more common. In some states like Nevada, subprime loans accounted for more than 30% of the loans originated in 2006. Subprime lending is often defined as lending to a borrower with poor credit, but this would be a misrepresentation of the essence of subprime lending, which is lending at higher fees and interest rates whether or not the borrower actually has bad credit (Aalbers, 2012b). Some estimates suggest that more than half of the subprime loans went to prime borrowers.

Up to 2006, the top 10 foreclosure cities almost exclusively consisted of Rustbelt cities. In 2007, when the crisis started, the list was a mix of Rustbelt and Sunbelt cities, but since 2008, the top 10 foreclosure cities are entirely Sunbelt cities, although Rustbelt city Detroit occasionally makes it back to the top 10. There are several such foreclosure lists, and they look slightly different, but generally speaking, cities in California now make up more than half of them, with cities in the South and Florida completing the list. The differences

across the United States are huge: the foreclosure rate in Richmond, Virginia, is almost 100 times as low as in Stockton, California, the foreclosure capital (Aalbers, 2009).

Housing prices can go down because of a structurally faltering economy, like in the Rustbelt, but also because they have been going up extremely fast, like in many cities in the Sunbelt. Housing prices in the Sunbelt were simply more inflated than elsewhere in the United States: the housing bubble was bigger and more likely to bust. In addition, some local and regional economies in the Sunbelt also showed signs of a declining economy, perhaps not structurally, as in the Rustbelt, but conjuncturally. Furthermore, high economic growth also meant a lot of new construction and more homeowners who recently bought a house, thereby increasing the pool of possible victims of falling housing prices. Finally, several states in the United States geared up their own regulations to impede some of the many excesses of subprime lending. As Wyly et al. (2009) have shown, states like North Carolina, New Mexico, Massachusetts, and West Virginia regulate a wide range of practices related to foreclosure rules, loan flipping, prepayment penalties, and other things. New Mexico, for example, introduced the Home Loan Protection Act in 2003.

Most subprime loans are granted to people to refinance their mortgages or as second mortgages (Squires, 2004). This means that most of these loans do not enable homeownership among low-income and minority groups, as is often argued. Subprime, and in particular predatory, loans frequently result in mortgage foreclosures at the individual level and housing abandonment at the neighborhood level. It is not just defaulting borrowers who are hit; in addition, there are severe spillover effects on housing prices, crime, and neighborhood decline. Although some cities in the Sunbelt are now hit harder than those in the Rustbelt, on a neighborhood level the Rustbelt still tops the foreclosure lists. On the list of the most foreclosed zip codes, four are now in Detroit, while Slavic Village in Cleveland has the most foreclosure filings. From the 1950s onward, redlining and suburbanization hit this neighborhood hard. Due to a combined economic and foreclosure crisis, demand for housing has fallen so dramatically that one can now buy many homes in Slavic Village that are priced under $30 000; on eBay, you could even buy one for less than $5000. Slavic Village, now referred to as foreclosure's ground zero, has also seen a rapid increase in crime.

Besides borrowers and neighborhoods, cities are also hit hard because tax income goes down due to foreclosed properties and lower real estate prices, while expenses are increasing as a result of foreclosures and property crime (Dymski, 2010; Immergluck, 2009). Local governments around the country have cut expenses on education, infrastructure, and social services. The public school system in California alone faces a loss of $4 billion in funding. Many cities in these states, but also in countless others, are facing lower incoming taxes (in particular real estate taxes) and cuts in funding of schools, social services, garbage collection, infrastructure, and so on. One complication in the United States, and possibly elsewhere, is that local governments as well as many states are not allowed to run a deficit. While the national government tries to stimulate the economy by spending more, municipalities and many states that are faced with decreasing revenues also have to cut back on expenses. This is by no means a marginal development. State revenues in New York, a state that in no way presents a worst-case scenario, have dropped 36% in 1 year.

Crises around the World

The current financial crisis originates in the housing and mortgage markets in the United States affects financial markets and economic sectors around the world (Aalbers, 2009; French et al., 2009). A few decades ago, most mortgage lenders were local or regional institutions. Today, most mortgage lenders are national lenders who tap into the global credit market. This is not so much the case because lenders are global financial institutions – most lenders are national in scope – but because they compete for the same credit in a global market. The idea was that in the wider credit market, it would be easy for mortgage lenders to get money, as mortgages were considered low risk. Mortgages would be an ideal investment for low-risk investors; cheaper credit, in return, would lower interest rates on mortgage loans. Fannie Mae and Freddie Mac, two government-sponsored enterprises that were meant to spur homeownership rates for low- and middle-income households, already introduced securitization in the 1960s. Securitization enables mortgage lenders to sell their mortgage portfolio on the secondary mortgage market to investors (Aalbers, 2012b).

The credit crisis started in 2007 when foreclosure and default rates went up and housing prices went down. This implied that investing in mortgages was not as low risk as people thought. The value of residential mortgage-backed securities (RMBS) fell even more dramatically. This is the case not only because many people had mortgage loans that were granted without down payments, but also because RMBS were sold with the idea of high returns. These high returns were partly based on high interest rates and not just on the value of the house, and partly on speculation, which increased the value of RMBS beyond what they were actually worth. In sum, not only were risks underestimated, but also returns were overestimated (even if there had not been rising defaults). In addition, even though housing prices on average fell by 20%, the impact on the RMBS market was much bigger. This is a result of not just inflationary prices but also leveraging. Major players in the RMBS market like investment banks basically invested with borrowed money (ratios of 1:20 were not uncommon, with 1:14 being the average) and because of this leveraging both profits and losses would be disproportionally big.

It now becomes easier to understand why the impact of partly local and partly national problems in housing and mortgage markets is global in scope and also affects other credit markets. The crisis hits not just investment banks on Wall Street, European banks, and pension funds that bought RMBS, but also individual investors and cities and towns around the globe. The example of Narvik in the far north of Norway is widely discussed (e.g., Aalbers, 2009). The city council of Narvik (population: 18 000) and three small, nearby municipalities had invested $78 million of the revenues of a nearby hydroelectric plant in RMBS and other products

offered by investment banks – and they lost most of it. The city's investments were meant for the construction of a new school, a nursing home, and a child-care facility. Instead, the city has cut the budget, and as a result several small rural schools will be closed, budgets for elderly care have been cut, the city is behind payments to civil servants, and the fire department will seize their 24/7 service and will switch to daytime service (in a city with mostly wooden houses). The Norwegian state has declared not to help Narvik and other municipalities, as it does not want to set a precedent by which the national state has to bear the losses of local authorities.

Of course, Narvik is not the exception. Cities and towns around the globe have been hit by the financial crisis. We could tell stories of cities in Greece, Spain, Ireland, and Iceland that have all been severely hit, but also of cities in countries that seemed to weather the financial crisis, whether they are located in Germany, Turkey, Brazil, China, or Australia. In some way, the crisis is felt in most economic sectors and in most places around the world. The only partial inclusion of, for instance, China in the global capitalist system has meant that production growth was slowed down but not that the country entered a recession. Yet, the rapid increase in the number of vacant luxury apartments in Chinese cities suggests that Chinese capitalism is also feeding local and national real estate bubbles – bubbles that result from a switching crisis and may one day result in a crisis of real estate and perhaps a larger economic crisis in China and likely elsewhere. It is crucial to link the spatial and temporal dimensions of local and global crises alike.

Cities, pension funds, sovereign wealth funds, and individuals around the world had invested in RMBS or the financial institutions that sank in this crisis such as the American Lehman Brothers or the Icelandic Landsbanki. Narvik illustrates well how connected the world has become in the twenty-first century. It also illustrates well that the world is not flat today: the old geography of local housing markets has not been replaced by a global housing market, but by a chain that starts local (a mortgage loan on a particular property), turns national (through lenders), then turns global (in the RMBS market) and then local again (by the effects in places like Narvik). Nearly everyone in North America, Europe, and Australia, and many in Asia, are in some way involved in this crisis, often as passive investors (e.g., through pension funds or investments by their local governments). We may not all be capitalists now, but most of us are investors, whether we want – or know – it or not.

Crisis Management

The financial crisis and its aftermath are often seen as a crisis of neoliberalism, and although it could be seen this way, what we have witnessed in the last couple of years is not the dismantling of neoliberalism but, by and large, the furthering of neoliberalism (IJURR, 2013). The financial crisis is leading to cutback after cutback. A core element of the state response to the crisis has, in the first place, been the government bailout of private companies, on the one hand, and austerity measures for everybody else, on the other. And although there is now a lot of talk about more regulation, a lot of this will not necessarily be unbeneficial to the neoliberal agenda. Moreover, the end result of this crisis may very well be the further dismantling of the welfare state. There are already many examples that point in this direction, at both the local and the national levels (IJURR, 2010, 2013).

The power of neoliberalism is that it furthers its agenda in economic boom periods as well as in economic bust periods. This is also visible in many European countries with relatively strong welfare states, where the current crisis provides neoliberals with the opportunity to cut money on state spending and to further the commodification of labor power, for example, by adjusting collective labor agreements or by forcing workers to choose between unemployment for some and wage cuts for all. The story is now repeating itself, not just in American cities hit hard by foreclosures such as Stockton, California, and Cleveland, Ohio, but also in many European countries, from hard-hit ones like Greece, Spain, Iceland, and Ireland to less affected ones such as the United Kingdom, France, Germany, and the Netherlands. Both local and national governments around the globe face public cuts. Ironically, neoliberalism is to blame for the problems but also prescribes the 'solution' to the problem: showering giant corporations with public money and cutting all other public spending (IJURR, 2013). As US Secretary of State Timothy Geithner said, "We saved the economy, but we kind of lost the public doing it."

It took surprisingly long before the financial crisis resulted in mass demonstrations. The streets of Dublin remained almost empty after the government had bailed out the banks at the expense of its citizens. Cutbacks in Greece and Spain have resulted in many a demonstration, but it was the Occupy Wall Street (OWS) movement that resulted in something more akin to a global (albeit far from complete) grassroots response to the financial crisis, even though OWS itself was inspired by revolts throughout the Mediterranean region (Aalbers, 2012a). OWS has not physically occupied Wall Street; it has occupied a pocket park that was formerly known as Liberty Plaza, but renamed Zuccotti Park in 2006. At times, Brookfield, the owner of Zuccotti Park, requested New York Mayor Michael Bloomberg and the New York Police Department (NYPD) to evict the demonstrators. The NYPD has not been in favor of eviction as it is easier to contain the occupiers in one easily surveillable place than have the protestors sprawl out around Lower Manhattan. For a few weeks in the late fall of 2011, the strategy of 'Brookberg' or 'Bloomfield' seemed to be to freeze out the protesters, but on the night of 14–15 November 2011, the mayor had Liberty Plaza cleared without prior announcement. That night, about 200 people were arrested. After the plaza was cleaned, the protesters were allowed back in, but they were not allowed to put up tents or sleep on the plaza. At the time of finishing this entry, a small group of protestors continues to occupy the plaza during the day.

The problem that Brookfield has with OWS is that it is hard to discuss with the protestors what should be done for them to leave Zuccotti Park. The reason for this fundamental uncertainty is only in part a result of the heterogeneous nature of the occupiers. All social movements show a degree of heterogeneity, but most still come up with a shared agenda. OWS has

not because it fights against something that is much harder to identify: the contemporary, internationalized, financialized economy. One of the many things that the current global financial and economic crisis demonstrates is that large parts of our economy are really beyond our control and often even beyond our imagination. We used to think of the financial sector as a sector that made other things possible, a sector that enabled people to buy houses, students to go to college, and entrepreneurs to start or expand their businesses. Some might have seen financial institutions as evil, but most would agree that they were necessary evils. What the financial crisis revealed is that financial institutions engage in a great deal of activities that are not useful to the economy at all. A mortgage loan may enable someone to buy a home, but many predatory subprime loans were designed to disrupt rather than to enable homeownership. A credit default swap (CDS) may basically work like insurance, but most of them are simply bets against something. If a CDS were like fire insurance, most of them would not be taken out by homeowners but by others who plan to benefit from someone else's house burning down. The financial sector and media call that innovation, but a lot of financial innovation is simply social destruction (Aalbers, 2012a).

There is a strong political dimension to all this. First, financial institutions heavily lobby states to get what they want, not just in the United States but also elsewhere. They did not simply want less regulation, but sometimes more regulation to enable them to innovate or destruct. This is not only something that happened in the decades prior to the financial crisis, but also something that continues and is undermining reregulatory efforts. Second, states, in particular although not exclusively in the United States, have failed to protect their citizens. Financial institutions have been able to prey not only on homeowners and pensioners but also on businesses and the people who depend on them for their livelihoods. States are complicit in the crisis because they have facilitated financial innovation and destruction. Years of neoliberal restructuring, albeit in different forms and intensities, have not yet been repaired by a little more consumer protection and a little more financial regulation. The fundamental question of what state–market relations should be has hardly been addressed. There is no simple answer, and the answer is of course heavily politicized (as it should be), but the 99% are right that there is a general agreement about the underregulation and underpolicing of financial institutions (Aalbers, 2012a).

Conclusion

While in the past a mortgage bubble or a housing bubble would affect the economy through homeowners, the current bursting of these bubbles affects the economy not just through homeowners but also through financial markets. Because lenders are now national in scope, this no longer affects only some housing markets, but all housing markets throughout a country. Housing markets may still be local or regional; mortgage markets are not. Since primary mortgage markets are national, the bubble in the national mortgage market affects all local and regional housing markets, although it clearly affects housing markets with a greater bubble more than those with a smaller bubble. In addition, secondary mortgage markets are global markets; this means that a crisis of mortgage securitization implies that investors around the globe, and therefore economies around the globe, are affected. The mortgage market crisis affects the US economy on both sides of the mortgage-lending chain – through homeowners and through financial markets – while it affects other economies in the world mostly through financial markets, not just because investors around the globe have invested in RMBS but also because the mortgage market has triggered a whole chain of events that have decreased liquidity, and this affects even agents in financial markets that have never been involved in RMBS.

The financial crisis is redrawing the world in many ways and at many levels. Financial globalization's impacts are, logically speaking, global, but that does not mean the impacts are the same around the globe, although there seem to be few financial institutions that remain unaffected by this crisis. The impact on nonfinancial institutions has been more spread out and has affected cities around the world in different ways and in different degrees. Housing bubbles, faltering economies, and regulation together have shaped the geography of the financial crisis at state and city levels in the United States. Subprime and predatory lending have affected low-income and minority communities more than others, and we therefore see a concentration of foreclosures not only in certain cities but also in certain neighborhoods, often those places inhabited by low-income and minority groups who have been excluded by earlier rounds of exclusion and exploitation. Yet, the meanings of globalization, not unlike the causes and consequences of this crisis, remain geographically uneven. It is important to understand that cities are an essential element in both, and that the fates of places like Stockton and Narvik are related not only to each other but also to those of Wall Street and Raffles Place (Singapore). The space of places is intrinsically linked to the flow of spaces.

See also: Cities: Capital, Global, and World; Economic Geography; Finance, Geography of; Financial Crises, History of; Financial Regulation; Globalization: Geographical Aspects; Housing and Cities; Marxist Geography; Race and Racism, Geography of; Social Movements, Geography of; Urban Geography; Urban Neoliberalism; Urban Policy in North America.

Bibliography

Aalbers, M.B., 2008. The financialization of home and the mortgage market crisis. Competition & Change 12, 148–166.

Aalbers, M.B., 2009. Geographies of the financial crisis. Area 41, 34–42.

Aalbers, M.B., 2012a. Socializing space and politicizing financial innovation/destruction: some observations on Occupy Wall Street. Belgeo 1–2. http://belgeo.revues.org/6155.

Aalbers, M.B. (Ed.), 2012b. Subprime Cities. The Political Economy of Mortgage Markets. Wiley-Blackwell, Oxford.

Ashton, P., 2009. An appetite for yield: the anatomy of the subprime mortgage crisis. Environment & Planning A 41, 1420–1441.

Castells, M., 1977. The Urban Question. A Marxist Approach. Arnold, London.

Castells, M., 1983. The City and the Grassroots. Arnold, London.

Cox, R.W., 1992. Global perestroika. In: Miliband, R., Panitch, L. (Eds.), Socialist Register 1992: New World Order? Merlin Press, London, pp. 26–43.

Dymski, G.A., 2010. Confronting the quadruple crisis. Geoforum 41, 837–840.

Engels, F., 1844/1999. The Condition of the Working Class in England. Oxford University Press, Oxford, UK.

French, S., Leyshon, A., Thrift, N., 2009. A very geographical crisis: the making and breaking of the 2007–2008 financial crisis. Cambridge Journal of Regions, Economy and Society 2, 287–302.

Gotham, K.F., 2006. The secondary circuit of capital reconsidered: globalization and the U.S. real estate sector. American Journal of Sociology 112, 231–275.

Harvey, D., 1982. The Limits to Capital. Blackwell, Oxford, UK.

Harvey, D., 1985. The Urbanization of Capital. Studies in the History and Theory of Capitalist Urbanization. Blackwell, Oxford, UK.

Harvey, D., 2010. The Enigma of Capital and the Crises of Capitalism. Oxford University Press, Oxford, UK.

Immergluck, D., 2009. Foreclosed. High-risk Lending, Deregulation, and the Undermining of America's Mortgage Market. Cornell University Press, Ithaca, NY.

IJURR, 2010. The crisis unfolds. International Journal of Urban and Regional Research 34 (3–4) (special issues).

IJURR, 2013. Neoliberalism in and after the neoliberal crisis. International Journal of Urban and Regional Research 37 (3) (special issue).

Rossi, U., 2010. Capitalist city. In: Hutchinson, R., Aalbers, M.B., Beauregard, R.A., Crang, M. (Eds.), Encyclopedia of Urban Studies. Sage, Thousand Oaks, CA, pp. 109–112.

Squires, G.D. (Ed.), 2004. Why the Poor Pay More: How to Stop Predatory Lending. Praeger, Westport, CT.

Wyly, E.K., Moos, M., Kabahizi, E., Hammel, D., 2009. Cartographies of race and class: mapping the class-monopoly rents of American subprime mortgage capital. International Journal of Urban and Regional Research 33, 332–354.

Cities: Capital, Global, and World

Saskia Sassen, University of Chicago, Chicago, IL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 1808–1816, © 2001, Elsevier Ltd.

Abstract

This is a review of how a new field of inquiry took shape. The article first examines the key features of this new field of inquiry, one that seeks to understand and document the multiple interactions of cities and the global economy in the current period. Next it examines the evolution of the scholarship on the subject, showing how in the early 1980s a large number of scholars began to explore these types of issues independently of each other, typically with very particular questions in mind rather than the more general subject of cities and the global economy. The rest of the article examines the critiques, several of the key hypotheses in this literature, and some of the themes that are emerging as an agenda for research and theory.

The Global City: Introducing a Concept and Its History

Each phase in the long history of the world economy raises specific questions about the particular conditions that make it possible. One of the key properties of the current phase is the ascendance of information technologies and the associated increase in the mobility and liquidity of capital. There have long been cross-border economic processes – flows of capital, labor, goods, raw materials. Over the last century, these took place largely within the interstate system, where the key articulators were national states and colonial empires. The international economic system was ensconced largely in this interstate system. This has changed rather dramatically during the 1990s as a result of privatization, deregulation, the opening up of national economies to foreign firms, and the growing participation of national economic actors in global markets.

It is in this context that we see a rescaling of what are the strategic territories that articulate the new system. With the partial unbundling or at least weakening of the national as a spatial unit due to privatization and deregulation and the associated strengthening of globalization, come conditions for the ascendance of other spatial units or scales. Among these are the subnational, notably cities and regions; cross-border regions encompassing two or more subnational entities; and supranational entities, such as global electronic markets and free-trade blocs. The dynamics and processes that get territorialized at these diverse scales can in principle be regional, national, or global.

We can locate the emergence of global cities in this context and against this range of instantiations of strategic scales and spatial units. In the case of global cities, the dynamics and processes that get territorialized are global. This article examines first some of the key theoretical and empirical elements of the global city model, followed by a brief history of the evolution of the literature on cities in the global economy. Section The Elements of a New Theoretical Framework is a more in-depth discussion of the organizing hypotheses of the global city model.

Sections The Global City Model: Organizing Hypotheses and New Forms of Centrality discuss two specific features: the question of place in a global economy and the question of city-to-city networks in domains other than the economic.

Elements in a New Conceptual Architecture

The globalization of economic activity entails a new type of organizational structure. To capture this theoretically and empirically requires, correspondingly, a new type of conceptual architecture. Constructs such as the global city and the global-city region are, in my reading, important elements in this new conceptual architecture. Arrighi's (1994) analysis is of interest here in that it posits the recurrence of certain organizational patterns in different phases of the capitalist world economy, but at gradually higher orders of complexity and expanded scope, and timed to follow or precede particular configurations of the world economy.

There are today several closely linked terms used to capture a particular intersection between global processes and cities (see Stren, 1996; Savitch, 1996 for overviews). The most common is world cities, a term attributed to Goethe and then relaunched in the work of Peter Hall (1966) and more recently respecified by John Friedmann (Friedmann and Goetz, 1982). Other related terms are 'supervilles' (Braudel, 1984), informational city (Castells, 1989). Choosing how to name a configuration has its own substantive rationality. The decision to formulate the term 'global city' (Sassen, 1984) stemmed out of a recognition of the specificity of the current period. The term world city has precisely the opposite attribute: it refers to a type of city which we have seen over the centuries (Braudel, 1984; Hall, 1966; King, 1990), and most probably also in much earlier periods in Asia than in the West. In this regard it could be said that most of today's major global cities are also world cities, but that there may well be some global cities that are not world cities in the full, rich sense of that term. This is partly an empirical question; further, as the global economy expands and incorporates additional cities into various cross-border networks, it is quite possible that the answer to that particular question will vary over time. Thus the fact that Miami has developed global city functions beginning in the late 1980s

does not make it a world city in that older sense of the term (Nijman, 1996). (See generally Abu-Lughod, 1999; Short and Kim, 1999.)

The Elements of a New Theoretical Framework

By the early 1980s a number of scholars had begun to study cities in the context of globalization (Walton, 1982; Ross and Trachte, 1983; Sassen, 1982; Rodriguez and Feagin, 1986). But it is one article in particular, 'The World City Hypothesis' by Friedmann and Goetz (1982) on which attention centered. This article took a variety of elements that were emerging in the research literature on cities, on the global economy, on immigration, and a number of other subjects, and sought to formalize these into several propositions about the role of cities in the global economy. The key elements in this framework were the emergence of several cities as basing points for global capital, a hierarchy (albeit a shifting one) of such cities, and the social and political consequences for these cities of being such basing points.

With the books by Castells (1989), King (1990), and Sassen (1991/2000), what had been a hypothesis in the early 1980s became a full-fledged theorization and empirical specification. These three books add important and distinct propositions to the general framework: Castells' proposition that globalization as constituted today has engendered a space of flows that reconfigures economic and political power; King's enlargement of the frame of reference to show that the highest levels of internationalization had taken place in the cities of colonial empires rather than in the center of the world economy; Sassen's proposition that it is not simply a matter of global coordination but one of the production of global control capacities and that an examination along these lines allows us to understand the role of global cities as production sites.

It is important to distinguish what is different about this literature from a broader, earlier literature on world cities prominently represented by the work of Peter Hall already in the 1960s, and a new literature on megacities especially focused on Latin America and Asia. These literatures do not have the fact of globalization and the centrality of cross-border networks connecting cities as crucial variables. The earlier literature on world cities is closer to the notion of capitals of empires: one city at the top of the power hierarchy.

In the current literature on global cities the determining factor is a cross-border, global network of cities that function as strategic sites for the management and specialized servicing of global economic operations. There is no such entity as a single global city, as there is with the capital of an empire; by definition, the global city is part of a network of cities. Similarly, an older literature focused on past world cities, as in the work of Braudel (1984), and earlier studies of major centers of world commerce and banking, as well as more recent work on urban hierarchies in the world system (Chase-Dunn, 1984), are to be differentiated from the current literature if we historicize the world economy and specify what is distinct today. Finally, we need to distinguish between a narrowly specified literature on global and world cities today and various literatures that directly or indirectly contribute to our understanding of these cities, notably the research on producer services.

By the mid-1990s the subject had clearly emerged as a rather large field for research among scholars in many different disciplines and countries. We can see this in the variety of authors and themes in several state-of-the-art collections, notably by Fainstein et al. (1993), Knox and Taylor (1995), Noller et al. (1994), Lo and Yeung (1996), and several others that elaborate, critique, expand the empirical base, and generally advance this theoretical and methodological project. We can also see it in several new important books that set the stage for highly focused research on particular variables, notably Meyer (1991), Thrift and Leyshon (1994), Keil (1993), Eade (1996) among others. We also see the creation of several book series by various publishers in different countries: the series on World Cities edited by Knox for Belhaven Press, the series edited by Milton Santos and his colleagues in Sao Paulo for Hucitec, the series edited by Martin Wentz for Campus Verlag, are just some.

It is not only the growth of the research literature but also the growth of a body of critical responses and analyses that signals the strength and vigor of this field of inquiry. There is only space here for the briefest mention, a sort of guide to criticisms: Logan and Swanstrom's (1990) critique of the excessive weight given to global structural processes in comparing internal vs. external factors that shape a city's economic development; Hamnett's (1996) critique of Sassen's proposition that globalization has contributed to social and economic polarization in global cities; Markusen and Gwiasda's (1994) critique of the notion that New York is at the top of the US urban hierarchy and how a comparison with Washington shows that the latter has a higher level of specialization than New York in many advanced specialized services, notably in legal services; critiques of the literature for its neglect of grassroots transnationalism and the new kinds of politics and identity formation that this entails; Beauregard's (1991) critique of the explanatory variables for changes in the built environment and the real estate industry; Simon's (1995) critique of the neglect of the periphery, notably Africa; the debate in *Urban Affairs* (March 1998) on the concept of the global city and a similar one in *Urban Studies* (Summer 2001); the special issue on 'Segregations Urbaines' of *Societes Contemporaines* in 1995; the special issue of *Urban Studies* in 1996, and many others.

There are two types of scholarly literature that intersect with this body of research on cities and the global economy, and indeed often invoke or use it to develop their arguments. They are on the one hand a literature of anthropological and cultural studies on transnationality, globalization, and identity formation (Holston, 1996; Low, 1999). The other is the scholarship by regional economic geographers on the global economy, who have also focused on cities (e.g., Moulaert and Scott, 1997; Gravesteijn et al., 1998). In the last few years there has been a new interest in this subject by geographers (Veltz, 1996; Scott, 2001; Storper, 1997).

In terms of method, a number of strategies have been developed. Even where there are data on intercity flows, it will take a lot of work to constitute the requisite data sets. In this regard, an ambitious initiative by the National Academy of

Sciences of the US examines how we can construct better data sets at the scale of the city (NAS, 2000).

Among the quantitative methodological and data formation strategies are the efforts by Smith and Timberlake (2002) and by Taylor et al. (2002). Smith and Timberlake (2002) conceptualize urban areas as central nodes in multiplex networks of economic, social, demographic and information flows. They use the methodological logic of network analysis, using particularly two measures: one of these is structural or relational equivalence between actors (i.e., cities) in a network; the second measure is centrality. Both of these measures relate to a number of propositions developed in the literature on cities in the global economy.

Taylor et al. (2002) have developed a new and pioneering data set that makes it possible to map the global networks of offices of the leading firms worldwide in several specialized corporate services, such as accounting, law, advertising, and finance. These networks of offices can be used to classify cities in terms of their participation in cross-border networks. The data can be analyzed using a variety of hypotheses and statistical as well as other methods.

There are several other efforts, but space dictates singling out just a few. David Meyer (1991) has developed ways of analyzing international networks through which a variety of exchanges of capital take place. Castells (1989) and Sassen (1991/2000) have developed several techniques of analysis which range from methods to understand the place of cities in global markets to expanding the representation of the global. In *The Informational City* and *The Global City* the authors sought to establish rather broadly what is the array of data sets that can be brought into an analysis of this subject – from international flows of capital and information to very localized social effects. This was an effort to resist the simplification in mainstream accounts which emphasize the global dispersal of activities and telecommunications and exclude most social issues.

Techniques for data analysis traditionally used by economic geographers can also be helpful. For instance Wheeler's (1986) examination of the dispersion of higher-order financial services throughout the US urban hierarchy – which he found had proceeded at a much slower rate than the dispersion of headquarters of other large corporations – can also be used for cross-border hierarchies. Wheeler found that corporations tend to proceed up the urban hierarchy for their advanced service and banking needs.

Elliott (1999) developed a test for the socioeconomic polarization hypothesis in global cities.

The Global City Model: Organizing Hypotheses

There are seven core hypotheses that organize the data and the theorization of the global city model. There follows a brief discussion of each as a way of producing a more precise representation of the model. (See Sassen, 1991/2000.)

First, the geographic dispersal of economic activities that marks globalization, along with the simultaneous integration of such geographically dispersed activities, is a key factor feeding the growth and importance of central corporate functions. The more dispersed a firm's operations across different countries, the more complex and strategic its central functions – that is, the work of managing, coordinating, servicing, financing a firm's network of operations.

Second, these central functions become so complex that increasingly the headquarters of large global firms outsource them: they buy a share of their central functions from highly specialized service firms: accounting, legal, public relations, programming, telecommunications, and other such services. Thus while even 10 years ago the key site for the production of these central headquarter functions was the headquarters of a firm, today there is a second key site: the specialized service firms contracted by headquarters to produce some of these central functions or components of them. This is especially the case with firms involved in global markets and non-routine operations. But increasingly the headquarters of all large firms are buying more of such inputs rather than producing them in-house.

Third, those specialized service firms engaged in the most complex and globalized markets are subject to agglomeration economies. The complexity of the services they need to produce, the uncertainty of the markets they are involved with either directly or through the headquarters for which they are producing the services, and the growing importance of speed in all these transactions, is a mix of conditions that constitutes a new agglomeration dynamic. The mix of firms, talents, expertise from a broad range of specialized fields makes a certain type of urban environment function as an information center. Being in a city becomes synonymous with being in an extremely intense and dense information loop. This is a type of information loop that as of now still cannot be replicated fully in electronic space, and has as one of its value-added features the fact of unforeseen and unplanned mixes of information, expertise, and talent; these can produce a higher order of information. This does not hold for routinized activities which are not as subject to uncertainty and non-standardized forms of complexity. Global cities are, in this regard, production sites for the leading information industries of our time.

A fourth hypothesis, derived from the preceding one, is that the more headquarters outsource their most complex, unstandardized functions particularly those subject to uncertain and changing markets and to speed, the freer they are to opt for any location because the more the work actually done in the headquarters is not subject to agglomeration economies. This further underlines that the key sector specifying the distinctive production advantages of global cities is the highly specialized and networked services sector. In developing this hypothesis I was responding to a very common notion that the number of headquarters is what specifies a global city. Empirically it may still be the case in many countries that the leading business center is also the leading concentration of headquarters, but this may well be because there is an absence of alternative locational options. But in countries with a well developed infrastructure outside the leading business center, there are likely to be multiple locational options for such headquarters.

Fifth, these specialized service firms need to provide a global service which has meant a global network of affiliates or some other form of partnership, and as a result we have seen a strengthening of cross-border city-to-city transactions and networks. At the limit this may well be the beginning of the formation of transnational urban systems. The growth of global

markets for finance and specialized services, the need for transnational servicing networks due to sharp increases in international investment, the reduced role of the government in the regulation of international economic activity, and the corresponding ascendance of other institutional arenas, notably global markets and corporate headquarters – all these point to the existence of a series of transnational networks of cities. One implication of this, and a related hypothesis for research is that the economic fortunes of these cities become increasingly disconnected from their broader hinterlands or even their national economies. We can see here the formation, at least incipient, of transnational urban systems. To a large extent it seems to me that the major business centers in the world today draw their importance from these transnational networks.

A sixth hypothesis, is that the growing numbers of high level professionals and high-profit making specialized service firms have the effect of raising the degree of spatial and socioeconomic inequality evident in these cities. The strategic role of these specialized services as inputs raises the value of top level professionals and their numbers. Further, the fact that talent can matter enormously for the quality of these strategic outputs and, given the importance of speed, proven talent is an added value, the structure of rewards is likely to experience rapid increases. Types of activities and of workers lacking these attributes, notably in manufacturing and industrial services, are likely to get caught in the opposite cycle.

A seventh hypothesis, is that one result of the dynamics described in hypothesis six, is the growing informalization of a range of economic activities which find their effective demand in these cities yet have profit rates that do not allow them to compete for various resources with the high-profit making firms at the top of the system. Informalizing part or all production and distribution activities, including of services, is one way of surviving under these conditions.

The first four hypotheses qualify what has emerged as a dominant discourse on globalization, technology, and cities, which posits the end of cities as important economic units or scales. There is a tendency in that account to take the existence of a global economic system as a given, a function of the power of transnational corporations and global communications. According to the global city model, the capabilities for global operation, coordination, and control contained in the new information technologies and in the power of transnational corporations need to be produced. By focusing on the production of these capabilities we add a neglected dimension to the familiar issue of the power of large corporations and the capacity of the new technologies to neutralize distance and place. A focus on the production of these capabilities shifts the emphasis to the practices that constitute what we call economic globalization and global control.

A focus on practices draws the categories of place and work process into the analysis of economic globalization. These are two categories easily overlooked in accounts centered on the hypermobility of capital and the power of transnationals. Developing categories such as place and work process does not negate the centrality of hypermobility and power. Rather, it brings to the fore the fact that many of the resources necessary for global economic activities are not hypermobile and are, indeed, deeply embedded in place, notably places such as global cities, global-city regions, and export processing zones.

This entails a whole infrastructure of activities, firms, and jobs necessary to run the advanced corporate economy. These industries are typically conceptualized in terms of the hypermobility of their outputs and the high levels of expertise of their professionals rather than in terms of the production or work process involved and the requisite infrastructure of facilities and non-expert jobs that are also part of these industries. This in turn brings with it an emphasis on economic and spatial polarization because of the disproportionate concentration of very high and very low income jobs in the city compared with what would be the case at a larger scale such as the region or the country. A focus on regions, in contrast will lead to an emphasis on broad urbanization patterns, a more encompassing economic base, more middle sectors of both households and firms. Emphasizing place, infrastructure, and non-expert jobs matters precisely because so much of the focus has been on the neutralization of geography and place made possible by the new technologies.

Dealing with place brings with it the problem of boundaries. These are at least of two sorts, the boundary of the territorial scale as such and the boundary of the spread of globalization in the organizational structure of industries, institutional orders, places, and so on. In the case of the global city it is possible to opt for an analytic strategy that emphasizes core dynamics rather than the unit of the city as a container – the latter being one that requires territorial boundary specification. Emphasizing core dynamics and their spatialization (in both actual and digital space) does not completely solve the boundary problem, but it does allow for a fairly clear trade-off between emphasizing the core or center of these dynamics and their spread institutionally and spatially.

Finally, the detailed examination of three particular cities (Sassen, 1991/2000) brought to the fore the extent to which these cities collaborate through their very specific advantages rather than simply competing with each other. In focusing on global finance in the 1980s and 1990s it becomes clear that the growth of the major centers was partly derived from the growing network of financial centers. In looking at the broader network it also becomes clear to what extent it was and remains characterized by a pronounced hierarchy among the growing number of centers that make up the network.

The growth of networked cross-border dynamics among global cities includes a broad range of domains – political, cultural, social, criminal. There are cross-border transactions among immigrant communities and communities of origin and a greater intensity in the use of these networks once they become established, including for economic activities that had been unlikely until now. We also see greater cross-border networks for cultural purposes, as in the growth of international markets for art and a transnational class of art curators; and for non-formal political purposes, as in the growth of transnational networks of activists around environmental causes, human rights, and so on. These are largely city-to-city cross-border networks, or, at least, it appears at this time to be simpler to capture the existence and modalities of these networks at the city level. The same can be said for the new cross-border criminal networks.

In brief, recapturing the geography of places represented by the network of global cities allows us to recapture people, workers, communities, and more specifically, the many different work cultures, besides the corporate culture, involved in the work of globalization. It also brings with it an enormous research agenda, one that goes beyond the by now familiar focus on cross-border flows of goods, capital and information.

Further, by emphasizing the fact that global processes are at least partly embedded in national territories, such a focus introduces new variables in current conceptions about economic globalization and the shrinking regulatory role of the state. That is to say, the space economy for major new transnational economic processes diverges in significant ways from the duality global/national presupposed in much analysis of the global economy. The duality national vs global suggests two mutually exclusive spaces – where one begins the other ends. One of the outcomes of a global city analysis is that it makes evident that the global materializes by necessity in specific places and institutional arrangements a good number of which, if not most, are located in national territories.

The two final sections examine two particular aspects that illustrate some of these issues concerning place in a global economy and city-to-city networks in realms other than the economic.

New Forms of Centrality

Several of the organizing hypotheses in the global city model concern the conditions for the continuity of centrality in advanced economic systems in the face of major new organizational forms and technologies that maximize the possibility for geographic dispersal. Historically, centrality has largely been embedded in the central city. Have the new technologies and organizational forms altered the spatial correlates of centrality?

Today there is no longer a simple straightforward relation between centrality and such geographic entities as the downtown, or the central business district (CBD). In the past, and up to quite recently in fact, the center was synonymous with the downtown or the CBD. The spatial correlate of the center can assume several geographic forms. It can be the CBD, as it still is largely in New York City, or it can extend into a metropolitan area in the form of a grid of nodes of intense business activity, as we see for instance in Frankfurt and Zurich. The center has been profoundly altered by telecommunications and the growth of a global economy, both inextricably linked; they have contributed to a new geography of centrality (and marginality). Simplifying we can identify four forms assumed by centrality today.

First, while centrality can assume multiple spatial correlates, the CBD in major international business centers remains a strategic site for the leading industries. But it is one profoundly reconfigured by technological and economic change (Graham and Marvin, 1996). Further, there are often sharp differences in the patterns assumed by this reconfiguring of the central city in different parts of the world, notably the United States and Western Europe (Veltz, 1996; Kunzmann, 1994).

In the United States, major cities such as New York and Chicago have large centers that have been rebuilt many times, given the brutal neglect suffered by much urban infrastructure and the imposed obsolescence so characteristic of US cities. This neglect and accelerated obsolescence produce vast spaces for rebuilding the center according to the requirements of whatever regime of urban accumulation or pattern of spatial organization of the urban economy prevails at a given time. In Europe, urban centers are far more protected and they rarely contain significant stretches of abandoned space; the expansion of workplaces and the need for intelligent buildings necessarily will have to take place partly outside the old centers. One of the most extreme cases is the complex of La Defense, the massive, state-of-the-art office complex developed right outside Paris to avoid harming the built environment inside the city. This is an explicit instance of government policy and planning aimed at addressing the growing demand for central office space of prime quality. Yet another variant of this expansion of the 'center' onto hitherto peripheral land can be seen in London's Docklands. Similar projects for recentralizing peripheral areas were launched in several major cities in Europe, North America, and Japan during the 1980s. (See Marcuse and van Kempen, 2000.)

Second, the center can extend into a metropolitan area in the form of a grid of nodes of intense business activity. One might ask whether a spatial organization characterized by dense strategic nodes spread over a broader region does in fact constitute a new form of organizing the territory of the 'center', rather than, as in the more conventional view, an instance of suburbanization or geographic dispersal. Insofar as these various nodes are articulated through digital networks, they represent a new geographic correlate of the most advanced type of 'center'. This is a partly deterritorialized space of centrality. Indeed much of the actual geographic territory within which these nodes exist falls outside the new grid of digital networks, and is in that sense partly peripheralized.

This regional grid of nodes represents a reconstitution of the concept of region. Far from neutralizing geography the regional grid is likely to be embedded in conventional forms of communication infrastructure, notably rapid rail and highways connecting to airports. Ironically perhaps, conventional infrastructure is likely to maximize the economic benefits derived from telematics. This is an important issue that has been lost somewhat in discussions about the neutralization of geography through telecommunications.

Third, we are seeing the formation of a transterritorial 'center' constituted, partly in digital space, via intense economic transactions in the network of global cities. These networks of major international business centers constitute new geographies of centrality. The most powerful of these new geographies of centrality at the global level binds the major international financial and business centers: New York, London, Tokyo, Paris, Frankfurt, Zurich, Amsterdam, Los Angeles, Sydney, Hong Kong, among others. But this geography now also includes cities such as Bangkok, Seoul, Taipei, Sao Paulo, Mexico City. The intensity of transactions among these cities, particularly through the financial markets, trade in services, and investment, has increased sharply, and so have the orders of magnitude involved. At the same time, there has been a sharpening inequality in the concentration of strategic

resources and activities between each of these cities and others in the same country, a condition that further underlines the extent to which this is a cross-border space of centrality.

The pronounced orientation to the world markets evident in such cities raises questions about the articulation with their nation-states, their regions, and the larger economic and social structure in such cities. Cities have typically been deeply embedded in the economies of their region, indeed often reflecting the characteristics of the latter; and they still do. But cities that are strategic sites in the global economy tend, in part, to disconnect from their region. This conflicts with a key proposition in traditional scholarship about urban systems, namely, that these systems promote the territorial integration of regional and national economies.

In the case of a complex landscape such as Europe's we see several geographies of centrality, one global, others continental and regional. A central urban hierarchy connects major cities, many of which in turn play central roles in the wider global system of cities: Paris, London, Frankfurt, Amsterdam, Zurich. These cities are also part of a wider network of European financial/cultural/service capitals, some with only one, others with several of these functions, which articulate the European region and are somewhat less oriented to the global economy than Paris, Frankfurt, or London. And then there are several geographies of marginality: the East–West divide and the North–South divide across Europe as well as newer divisions. In Eastern Europe, certain cities and regions, notably Budapest, are rather attractive for purposes of investment, both European and non-European, while others will increasingly fall behind, notably in Rumania, Yugoslavia, and Albania. We see a similar differentiation in the south of Europe: Madrid, Barcelona, and Milan are gaining in the new European hierarchy; Naples, Rome, and Marseilles are not quite.

Fourth, new forms of centrality are being constituted in electronically generated spaces. For instance, strategic components of the financial industry operate in such spaces. The relation between digital and actual space is complex and varies among different types of economic sectors. But it is increasingly becoming evident that the highly complex configurations for economic activity located in digital space contain points of coordination and centralization.

The Global City as a Nexus for New Politico-cultural Alignments

The incorporation of cities into a new cross-border geography of centrality also signals the emergence of a parallel political geography. Major cities have emerged as strategic sites not only for global capital, but also for the transnationalization of labor and the formation of translocal communities and identities (Smith, 1997). In this regard cities are sites for new types of political operations and for a whole range of new 'cultural' and subjective operations (Watson and Bridges, 1999; Allen et al., 1999). The centrality of place in a context of global processes makes possible a transnational economic and political opening for the formation of new claims and hence for the constitution of entitlements, notably rights to place. At the limit, this could be an opening for new forms of 'citizenship' (Isin, 2000; Holston, 1996).

The emphasis on the transnational and hypermobile character of capital has contributed to a sense of powerlessness among local actors, a sense of the futility of resistance. But an analysis that emphasizes place suggests that the new global grid of strategic sites is a terrain for politics and engagement. The loss of power at the national level produces the possibility for new forms of power and politics at the subnational level. Further, insofar as the national as container of social process and power is cracked (Brenner, 1998; Taylor, 1995) it opens up possibilities for a geography of politics that links subnational spaces across borders. Cities are foremost in this new geography. One question this engenders is how and whether we are seeing the formation of a new type of transnational politics that localizes in these cities.

Immigration, for instance, is one major process through which a new transnational political economy and translocal household strategies are being constituted (Portes, 1995; Skeldon, 1997). It is one largely embedded in major cities insofar as most immigrants, certainly in the developed world, whether in the US, Japan, or Western Europe, are concentrated in major cities. It is, in many regards, one of the constitutive processes of globalization today, even though not recognized or represented as such in mainstream accounts of the global economy. This configuration contains unifying capacities across national boundaries and sharpening conflicts within cities. Global capital and the new immigrant workforce are two major instances of transnationalized actors that have unifying properties across borders, and thus internally to each, and find themselves in contestation with each other inside cities. Researching and theorizing these issues will require approaches that diverge from the more traditional studies of political elites, local party politics, neighborhood associations, immigrant communities, and so on through which the political landscape of cities and metropolitan regions has been conceptualized in urban studies.

One way of thinking about the political implications of this strategic transnational space anchored in global cities is in terms of the formation of new claims on that space. The city has indeed emerged as a site for new claims: by global capital which uses the city as an 'organizational commodity', but also by disadvantaged sectors of the urban population, frequently as internationalized a presence in large cities as capital. The 'denationalizing' of urban space and the formation of new claims by transnational actors, raise the question 'Whose city is it?'

This is a space that is both place-centered in that it is embedded in particular and strategic locations; and it is transterritorial because it connects sites that are not geographically proximate yet are intensely connected to each other. If we consider that large cities concentrate both the leading sectors of global capital and a growing share of disadvantaged populations – immigrants, many of the disadvantaged women, people of color generally, and, in the megacities of developing countries, masses of shanty dwellers – then we can see that cities have become a strategic terrain for a whole series of conflicts and contradictions (Allen et al., 1999; Tardanico and Lungo, 1995). We can then think of cities also as one of the sites for the contradictions of the globalization of capital, even though the city cannot be reduced to this dynamic.

Conclusion

An examination of globalization through the concept of the global city introduces a strong emphasis on strategic components of the global economy rather than the broader and more diffuse homogenizing dynamics we associate with the globalization of consumer markets. As a consequence, this also brings an emphasis on questions of power and inequality. And it brings an emphasis on the actual work of managing, servicing, and financing a global economy. Second, a focus on the city in studying globalization will tend to bring to the fore the growing inequalities between highly provisioned and profoundly disadvantaged sectors and spaces of the city, and hence such a focus introduces yet another formulation of questions of power and inequality.

Third, the concept of the global city brings a strong emphasis on the networked economy because of the nature of the industries that tend to be located there: finance and specialized services, the new multimedia sectors, and telecommunications services. These industries are characterized by cross-border networks and specialized divisions of functions among cities rather than international competition per se. In the case of global finance and the leading specialized services catering to global firms and markets – law, accounting, credit rating, telecommunications – it is clear that we are dealing with a cross-border system, one that is embedded in a series of cities, each possibly part of a different country. It is a *de facto* global system.

Fifth, a focus on networked cross-border dynamics among global cities also allows us to capture more readily the growing intensity of such transactions in other domains – political, cultural, social, criminal.

Global cities around the world are the terrain where a multiplicity of globalization processes assume concrete, localized forms. These localized forms are, in good part, what globalization is about. Recovering place means recovering the multiplicity of presences in this landscape. The large city of today has emerged as a strategic site for a whole range of new types of operations – political, economic, 'cultural', subjective. It is one of the nexus where the formation of new claims, by both the powerful and the disadvantaged, materializes and assumes concrete forms.

See also: Culture, Sociology of; Globalization and World Culture; Globalization, Anthropology of; Globalization: Geographical Aspects; Information Society, Geography of; Information Society; International Communication: History.

Bibliography

Abu-Lughod, J.L., 1999. New York, Los Angeles, Chicago: America's Global Cities. University of Minnesota Press, MN.
Allen, J., Massey, D., Pryke, M. (Eds.), 1999. Unsettling Cities. Routledge, London.
Arrighi, G., 1994. The Long Twentieth Century. Money, Power, and the Origins of Our Times. Verso, London.
Beauregard, R., 1991. Capital restructuring and the new built environment of global cities: New York and Los Angeles. International Journal of Urban and Regional Research 15 (1), 90–105.
Braudel, F., 1984. The Perspective of the World, vol. 3. Collins, London.
Brenner, N., 1998. Global cities, global states: global city formation and state territorial restructuring in contemporary Europe. Review of International Political Economy 5 (1), 1–37.
Castells, M., 1989. The Informational City. Blackwell, London.
Chase-Dunn, C., 1984. Urbanization in the world system: new directions for research. In: Smith, M.P. (Ed.), Cities in Transformation. Sage, Beverly Hills, CA.
Eade, J. (Ed.), 1996. Living the Global City: Globalization as a Local Process. Routledge, London.
Elliott, J.R., 1999. Putting 'global cities' in their place: urban hierarchy and low-income employment during the post-war era. Urban Geography 20 (2), 95–115.
Fainstein, S., Gordon, I., Harloe, M., 1993. Divided City: Economic Restructuring and Social Change in London and New York. Blackwell, New York.
Friedmann, J., Goetz, W., 1982. World city formation: an agenda for research and action. International Journal of Urban and Regional Research 6, 309–344.
Graham, S., Marvin, S., 1996. Telecommunications and the City: Electronic Spaces, Urban Places. Routledge, London.
Gravesteijn, S.G.E., van Griensven, S., de Smidt, M.C. (Eds.), 1998. Timing Global Cities. Nederlandse Geografische Studies, vol. 241.
Hall, P., 1966. The World Cities. McGraw-Hill, New York.
Hamnett, C., 1996. Why Sassen is wrong: a response to Burgers. Urban Studies 33 (1), 107–110.
Holston, J. (Ed.), 1996. Cities and Citizenship. Public Culture, vol. 8(2) (Winter).
Isin, E. (Ed.), 2000. Democracy, Citizenship and the Global City. Routledge, London.
Keil, R., 1993. Weltstadt – Stadt der Welt: Internationalisierung und Lokale Politik in Los Angeles. Westfaelisches Dampfboot, Munster, Germany.
King, A.D., 1990. Global Cities: Post-Imperialism and the Internationalization of London. Routledge, London.
Knox, P.L., Taylor, P.J. (Eds.), 1995. World Cities in a World-System. Cambridge University Press, Cambridge, UK.
Kunzmann, K.R., 1994. Berlin im zentrum europaischer Stadtnetze. In: Werner, S. (Ed.), Hauptstadt Berlin. Band 1: Nationale Hauptstadt Europaeische Metropole. Berlin Verlag, Berlin, pp. 233–246.
Lo, F., Yeung, Y. (Eds.), 1996. Emerging World Cities in Pacific Asia. United Nations University, Tokyo.
Logan, J.R., Swanstrom, T. (Eds.), 1990. Beyond the City Limits: Urban Policy and Economic Restructuring in Comparative Perspective. Temple University Press, Philadelphia.
Low, S.M., 1999. Theorizing the city. In: Low, S.M. (Ed.), Theorizing the City. Rutgers University Press, New Brunswick, NJ, pp. 1–33.
Markusen, A., Gwiasda, V., 1994. Multipolarity and the layering of functions in the world cities: New York city's struggle to stay on top. International Journal of Urban and Regional Research 18, 167–193.
Machimura, T., 1998. Symbolic use of globalization in urban politics in Tokyo. International Journal of Urban and Regional Research 22 (2), 183–194.
Marcuse, P., van Kempen, R., 2000. Globalizing Cities. A New Spatial Order. Blackwell, Oxford, UK.
Meyer, D.R., 1991. Change in the world system of metropolises: the role of business intermediaries. Urban Geography 12 (5), 393–416.
Moulaert, F., Scott, A.J., 1997. Cities, Enterprises and Society on the *Eve of the 21st Century*. Pinter, New York.
National Academy of Sciences (NAS), 2000. Panel on Urban Data Sets. Committee on Population, NAS, Washington, DC.
Nijman, J., 1996. Breaking the rules: Miami in the urban hierarchy. Urban Geography 17 (1), 5–22.
Noller, P., Prigge, W., Ronneberger, K. (Eds.), 1994. Stadt-Welt. Campus Verlag, Frankfurt, Germany.
Portes, A. (Ed.), 1995. The Economic Sociology of Immigration. The Russell Sage Foundation, New York.
Rodriguez, N.P., Feagin, J.R., 1986. Urban specialization in the world system. Urban Affairs Quarterly 22 (2), 187–220.
Ross, R., Trachte, K., 1983. Global cities and global classes: the peripheralization of labor in New York city. Review 6 (3), 393–431.
Santos, M., Souze, M.A.A., Silveira, M.L. (Eds.), 1994. Territorio Globalizacao e Fragmentacao. Editorial Hucitec, Sao Paulo, Brazil.
Sassen, S., 1982. Recomposition and peripheralization at the core. Contemporary Marxism 5, 88–100.
Sassen, S., 1984. The new labor demand in global cities. In: Smith, M.P. (Ed.), Cities in Transformation. Sage, CA.
Sassen, S., 1991/2000. The Global City: New York, London, Tokyo. Princeton University Press, NJ.
Savitch, H.V., 1996. Cities in a global era: a new paradigm for the next millenium. In: Cohen, M.A., Blair, A., Ruble, J., Tulchin, S., Garland, A.M. (Eds.), Preparing for the Urban Future. Global Pressures and Local Forces. Woodrow Wilson Center Press, Washington, DC, pp. 39–65.
Scott, A.J., 2001. Global City-Regions. Oxford University Press, Oxford, UK.
Short, J.R., Kim, Y., 1999. Globalization and the City. Longman, Essex.

Simon, D., 1995. The world city hypothesis: reflections from the periphery. In: Knox, P.L., Taylor, P.J. (Eds.), World Cities in a World-System. Cambridge University Press, Cambridge, UK, pp. 132–155.

Skeldon, R., 1997. Hong Kong: colonial city to global city to provincial city? Cities 14 (5), 265–271.

Smith, R.C., 1997. Transnational migration, assimilation and political community. In: Crahan, M., Vourvoulias-Bush, A. (Eds.), The City and the World. Council of Foreign Relations, NY.

Smith, D., Timberlake, M., 2002. Cross-border air traffic patterned networks. In: Sassen, S. (Ed.), Global Networks/Linked Cities. Routledge, London.

Smith, M.P., Feagin, J.R. (Eds.), 1995. The Bubbling Cauldron: Race, Ethnicity, and the Urban Crisis. University of Minnesota Press, MN.

Storper, M., 1997. The Regional World: Territorial Development in a Global Economy. Guilford Press, New York.

Stren, R., 1996. The studies of cities: popular perceptions, academic disciplines, and emerging agendas. In: Cohen, M.A., Blair, A., Ruble, J., Tulchin, S., Garland, A.M. (Eds.), Preparing for the Urban Future. Global Pressures and Local Forces. Woodrow Wilson Center Press, Washington, DC, pp. 392–420.

Tardanico, R., Lungo, M., 1995. Local dimensions of global restructuring in urban Costa Rica. International Journal of Urban and Regional Research 19 (2), 223–249.

Taylor, P.J., 1995. World cities and territorial states: the rise and fall of their mutuality. In: Knox, P.L., Taylor, P.J. (Eds.), World Cities in a World-System. Cambridge University Press, Cambridge, UK, pp. 48–62.

Taylor, P.J., Beaverstock, J.V., Walker, D.R.F., 2002. Introducing GaWC: researching world city network formation. In: Sassen, S. (Ed.), Global Networks/Linked Cities. Routledge, London.

Thrift, N., Leyshon, A., 1994. A phantom state? The de-traditionalization of money, the international financial system and international financial centres. Political Geography 13 (4), 299–327.

Veltz, P., 1996. Mondialisation Villes et Territoires: L'Economie d'Archipel. Presses Universitaires de France, Paris.

Walton, J., 1982. The international economy and peripheral urbanization. In: Norman, I., Fainstein, S. (Eds.), Urban Policy under Capitalism. Sage, CA, pp. 119–135.

Watson, S., Bridges, G. (Eds.), 1999. Spaces of Culture. Sage, London.

Wheeler, J.O., 1986. Corporate spatial links with financial institutions: the role of the metropolitan hierarchy. Annals of the Association of American Geographers 76 (2), 262–274.

Cities: Internal Structure

Elisabeth Lichtenberger, University of Vienna, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article concentrates on big cities in western societies and the changing principles of urban development since the twentieth century. It combines general principles and the dual cycle theory of urban expansion and urban renewal with the impact of political systems and cultural environments.

Three general principles are formative for the internal structure of cities:

1. The accessibility determined by transport technology.
2. The so-called 'distance-decay' from center to periphery, defined by the social gradient and the land values gradient.
3. The hierarchical organization of urban space.

The dual cycle theory of urban development, explains urban decay as a consequence of the time lag between urban expansion and urban renewal. Development differs greatly depending on whether it takes place under private capitalism, in welfare states, or in postsocialist countries.

Three periods of urban development are distinguished. After World War II the system of centered cities dating from the period of Europeanization of the globe, was superseded by city models derived from North America.

During the period AD 1950–2000, a rapidly growing system of suburbia created a sort of extensive – though not ubiquitous – network. The combination of 'urban-like,' suburban structures and new megastructures seems to correspond to postindustrial America's abandonment of the areas of earlier industrial-city development.

In the foreseeable future, however, twenty-first century city development is unlikely to follow North American models much longer. All over the globe – in the geo-strategically and economically newly relevant BRIC (Brazil, Russia, India, and China) countries as well as in all countries of the Islamic world – decision makers have recognized that the inherent functioning of cities (above all of metropolises and capital cities) with the help of state-of-the-art technologies, has gained in importance. Megacities in these regions with populations of 20 million and more surpass all European and North American big cities.

General Principles

Accessibility and Transport Technology

The city is a centered system based on the twofold premise that the city center is the engine of development and the place easiest to access. Accessibility of both the center and the entire city area is determined by its opening up for development and by transport technology (see **Figure 1**).

A city of pedestrians and coaches tended to be circular. When streetcars came into use in Europe, they started mostly at the town gates. This made for star-shaped growth, with the interstitial areas lagging behind in development. Where public transport is predominant, the city center remains the place of highest access to the metro region. The *métropole concentrée* of Paris is an example.

The Regie Autonome de Transports Parisiens transports more than 3 billion passengers (2011) annually, half of them underground (http://wikipedia rg/wiki/RATP; 12 December 2013).

The absolute predominance of private transport reduces access to the center; under a liberal political system and with land abundantly available, a city or metropolitan region designed for car traffic emerges. Los Angeles offers a prototype of this; the City of Los Angeles Transportation Authority (Los Angeles Department of Transportation (LADOT)), for example, moves only about 30 million passengers per year (http://www.discoverlosangeles.com/blog/los-angeles-public-transit; 12 December 2013) although the population of LA is significantly larger than that of Paris.

Distance Decay

There are two theoretical approaches for analyzing gradients from the center to the periphery.

Social Gradients

The central-peripheral organization of urban society may be described by means of centrifugal and centripetal social gradients. These gradients permit the following diagnosis regarding urban growth. Wherever the city center is also the social center and a centripetal social gradient is in evidence, the demand for space of the upper classes extends outward from the center. This causes residential and social upgrading of the adjacent middle class districts. In the middle zone, former lower class quarters are turned into middle class districts. The Founders Period growth of Berlin, Budapest, Copenhagen, Paris, Prague, and Vienna was of this kind. This type of upgrading must not be confused with the recent phenomenon of gentrification in North American cities. Here, middle and upper income urbanites move centripetally into rundown central districts (see **Figure 2**).

Wherever a centrifugal social gradient predominates, as in US cities, a filtering down takes place, with the deterioration of the building fabric in the center; lower class people move outward into adjacent middle and upper class districts. This social downgrading of high-quality living quarters is hard to stop and presented a major problem for British town planners. Having accomplished the redevelopment of rundown early industrial terrace houses next to the old town centers, they were faced with the daunting task of redeveloping devastated mid- and late-nineteenth centuries terrace house districts (the necessary renovation of early twentieth century council housing was partly effected by privatization under Thatcher).

Land Values Gradient

According to the theory of urban land markets (Alonso, 1964), transport costs determine land rents and therefore

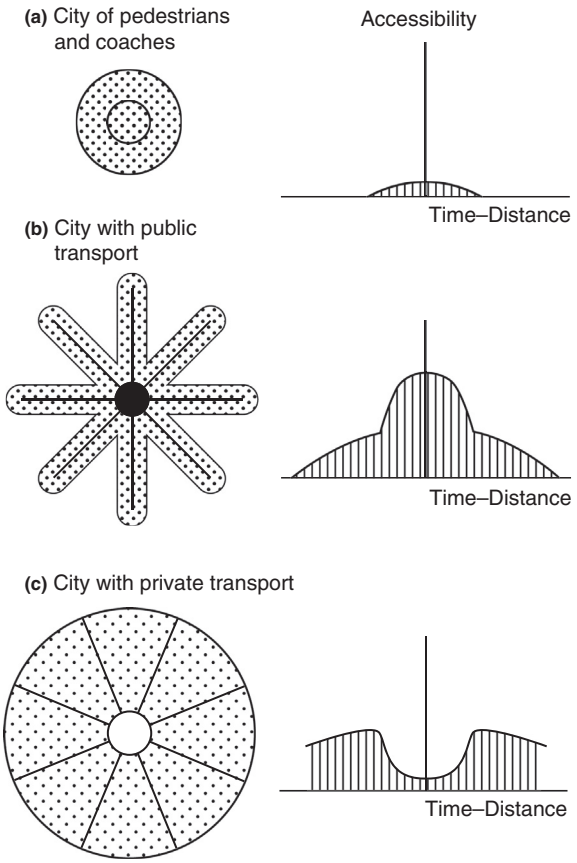

Figure 1 The accessibility of the city center. Lichtenberger, E., 1998. Stadtgeographie (Urban Geography), third ed. Teubner, Stuttgart, Leipzig, Germany. Italian translation: Geografia dello spazio urbano, first ed. Unicopli, Milano, ill. 2.30, p. 103.

urban land uses. Regularly, in socioeconomically intact city centers, there is a center-periphery gradient with several consecutive zones of use, outward from the central business district (CBD). Replacement of historical city-models by the new model of suburbia has made for abandonment of central areas, resulting in 'craters' of land prices and visible decay of inner cities in the United States. As there are no efficient means of public transport, inhabitants of the centers of most metropolitan areas are at a disadvantage. Detroit is the classic example of a dying US-American metropolitan area.

The Alonso model excludes restrictions on planning land use and vertical development. With zoning laws, building categories, and other legal regulations, the gradient of real estate prices is altered. Each category or zone is divided in two, with the inner zone obviously better suited for business purposes and office space than the outer zone that is used for dwelling as it is more profitable (see Figure 3).

Hierarchical Structures

Hierarchical order is one of the basic types of systemic organization. When organizing the physical space of urban areas, a hierarchical order has been attempted wherever planned development was feasible and land in abundant supply. Examples run from seventeenth century Sicily (Granmichele) via Howard's New Town idea (1902) (with city center and garden suburbs in a cluster city), to major planning projects for urban peripheries in Europe at present. The hierarchy is constructed from the bottom upward. The basic units are electoral ward, whose dimensions in many European countries and in North America are derived from the traditional 'pedestrian ideology' of 1 km (i.e., 15 min walking distance) as the determinant of standard size. Comprehensive hierarchical urban structures have been realized only in the former Socialist countries of Eastern Europe, above all the USSR. In the rest of Europe, hierarchical organization of urban space is the exception rather than the rule.

The Impact of Political Systems

Social-Ecological City Models

In their paper, *The Nature of Cities* (1945), Harris and Ullman tried to model internal patterns within cities. The triad of models presented makes use of diverse development phenomena:

1. invasion and succession of social groups in Burgess' zonal model,
2. transport-induced ribbon-developments along traffic lines in Hoyt's sector model, and
3. collective preferences in respect of the allocation of workplaces in the Harris–Ullman multiple-nuclei model.

However, three important underlying assumptions have not yet been made explicit:

1. the political system of liberalism,
2. the historical 'one-dimensionality' of urban development, and
3. the concept of the city center as CBD.

The models mentioned represent ideal types of North American cities of the interwar period (see Figure 4).

Historical-Political Systems and the Concept of the City Center

Changes in the political system alter the conceptions of the city and urban society. The function of the city center is changed. European urban development may be defined as a succession of four types of political systems and corresponding types of cities (Lichtenberger, 1970, 1976).

The concepts of the city center were markedly different in each of the four cases:

1. In the burghers' city of the medieval feudal territorial state, the marketplace was the social center.
2. The center shifted to the ruler's residence in the residence city of the absolutist state. Thus, a social gradient falling from the center toward the periphery is the general rule in preindustrial cities.
3. In the age of liberalism, Great Britain created the prototype of the industrial city. A social gradient rising from the center toward the periphery predominated, a development later paralleled in North American cities.

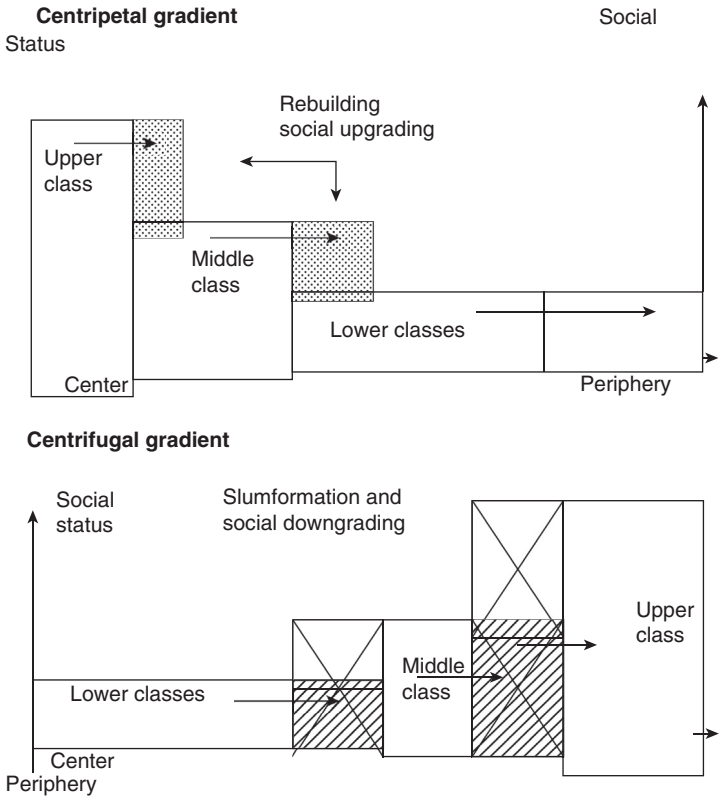

Figure 2 Centrifugal and centripetal social gradients. Lichtenberger, E., 1998. Stadtgeographie (Urban Geography), third ed. Teubner, Stuttgart, Leipzig, Germany. Italian translation: Geografia dello spazio urbano, first ed. Unicopli, Milano, ill. 2.35, p. 110.

4. Again, Great Britain set the rules for the New Town idea; the attempt at structuring the amorphous masses of big industrialized cities on a human scale, with the city divided into parts with different functions. From the outset, spatial segregation of inhabitants was barred from the design of New Towns – a fact that is still influential in urban planning today. European city development is most complex; various superpositions and (to different degrees) the persistence of historical structures, caused the diversification of socioeconomic patterns and the creation of different social gradients.

Vienna in the 1960s offers the model for the traditional continental European city (Lichtenberger, 1993b):

1. Social status is still highest in the core and declines toward the periphery. Depending on site preferences, individual districts deviate from this rule.
2. The CBD has maintained some residential functions and is not surrounded by slums, but by middle and upper class residential districts, encircled by lower class quarters.
3. The wide belt of multistory blocks is followed on its outside by Founders Period industrial quarters, followed in their turn by loosely built-up districts.
4. The fringe zone is not defined by an extensive speculation area of 'vacant land' as in North America, but by sectors of intensive agricultural land use (in keeping with the von Thünen model) such as truck farming and viticulture, as well as allotment gardens and weekend homes (see **Figure 5**).

The Impact of State Socialism

Under state socialism, municipal governments were the local planning authorities, responsible for the mass of multistory housing that was unprofitable because of 'social' low rents. They lost all chances of capital accumulation from real estate property and became dependent on financial endowments from the central government and on central planning. The consequences were:

Big cities' increasing demand for space was met by extensive incorporations.
Public construction (by state, municipal, and other collective institutions) took absolute precedence over private construction.
Architectural design followed totalitarian principles, stressing elements of over-wide avenues and huge squares, both imposing and strategically important.
Massive antisegregation strategies were pursued.
Entire old city centers were listed on historical registers.
City enlargement was effected via New Cities in the shape of hierarchically structured giant multistory blocks.
In compliance with the Charter of Athens, industrial and dwelling zones were strictly separated.
Commuting requirements were taken care of by subway construction.
Forced industrialization created extensive industrial zones with plants located close to railway lines and superhighways.

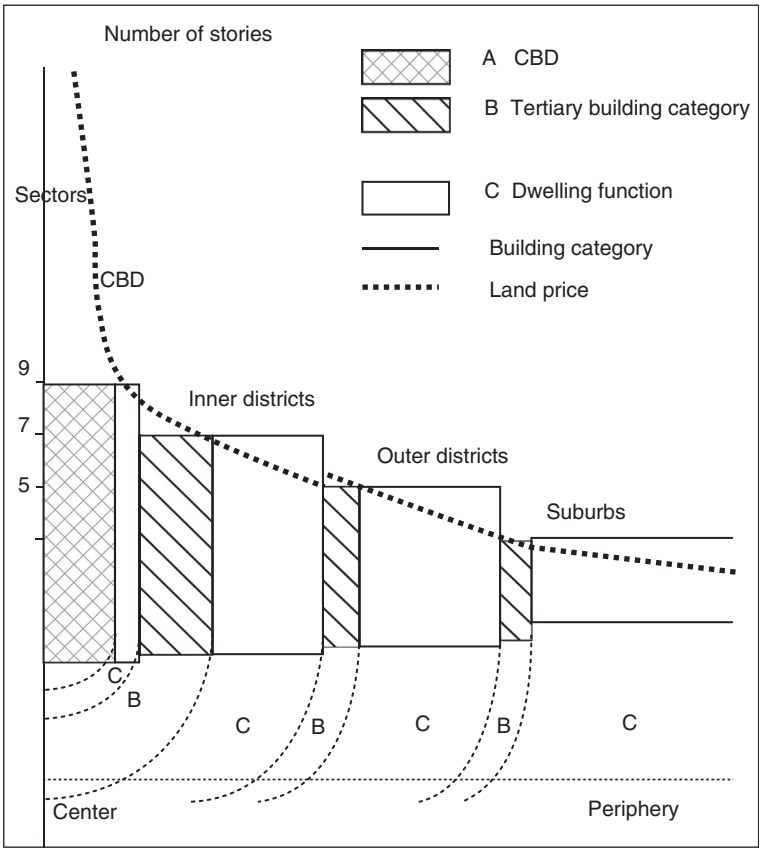

Figure 3 Land values gradient and zoning. Lichtenberger, E., 1998. Stadtgeographie (Urban Geography), third ed. Teubner, Stuttgart, Leipzig, Germany. Italian translation: Geografia dello spazio urbano, first ed. Unicopli, Milano, ill. 2.3, p. 107.

Extensive leisure zones were created on the fringes, with both collective recreation facilities and private second homes (see Figure 6).

City Growth and Political Systems

Cities are growing systems, with increasing populations and/or with increasing demand for space for diverse urban functions (housing, work, education, leisure, traffic, etc.) in a setting of economic growth and technical innovation.

The diverging political-economic effects of central planning and of a free enterprise economy cause significant differences in the physical growth of cities in particular systems. Cities grow in two directions: laterally and vertically (Lichtenberger, 1992).

The Third Dimension

Vertical growth means high-rise construction. Up to the 1970s, when some cities changed their laws, European cities could not expand vertically because of strict zoning laws, but were forced to grow laterally. This necessitated the conversion of dwelling space into office space, e.g., in Paris and Vienna. High-rise construction in European cities started relatively late (see Figure 7). Its location within the city – frequently subject to special permission – is different from that in North America. There, the vertical structure of urban skylines shows that land prices peak in the center while in European cities 'monument protection' bars high-rises from the centers. Thus, the new landmarks of banks, insurance firms, corporation headquarters, and hotels keep a polite distance from the old landmarks of churches, town halls, and palaces.

For the sake of access to the various supply and disposal mains, high-rises are preferably located along urban 'scars': at the interfaces of traditional zones where former boundaries still show in open space or low physical objects. Frequently, new high-rises accent not only the edge of traditional inner cities but also, major points of access to older outer cities and suburbs. High-rises also mark the front of growth of the CBD, busy commuter train stations, as well as 'satellite' districts. They are also instrumental in slum clearing.

Development of Urban Fringes

The development of urban fringes differs materially depending on whether it takes place under private capitalism, in welfare states or, in retrospect, under state capitalism.

1. Under US private capitalism, cities grow as profits from land speculation and rising land prices are invested in land development and in technical improvements. This starts an upward spiral of development and rising prices. North American cities show two wide zones of speculation. The inner one, around the CBD, is marked by decay and slum areas at present. Far more impressive, at least because of its

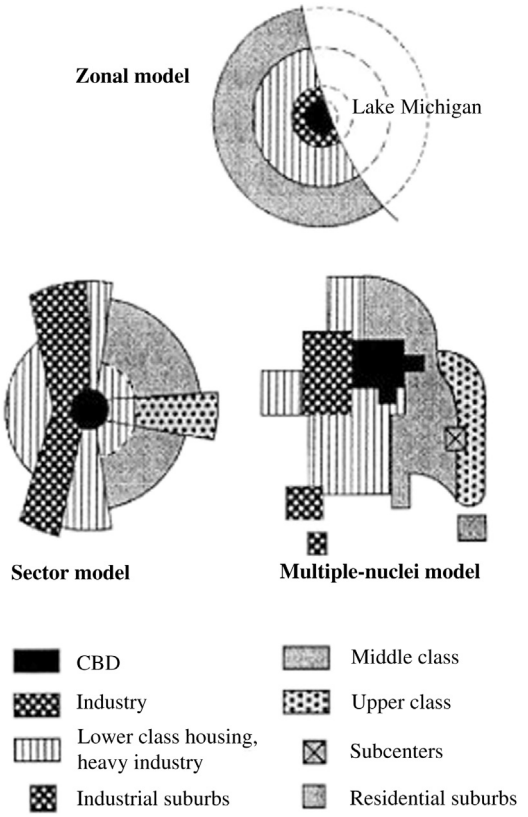

Figure 4 Social ecological models of US cities during the inter-war-period. Lichtenberger, E., 1998. Stadtgeographie (Urban Geography), third ed. Teubner, Stuttgart, Leipzig, Germany. Italian translation: Geografia dello spazio urbano, first ed. Unicopli, Milano, ill. 2.7, p. 58; Hoyt, 1939. The Structure and Growth of Residential Neighborhoods in American Cities. Federal Housing Administration, Washington, DC; and Harris C.D., Ullman, E.L.,1945. The nature of cities. The Annals of American Academy of Political and Social Sciences 242, 7–17.

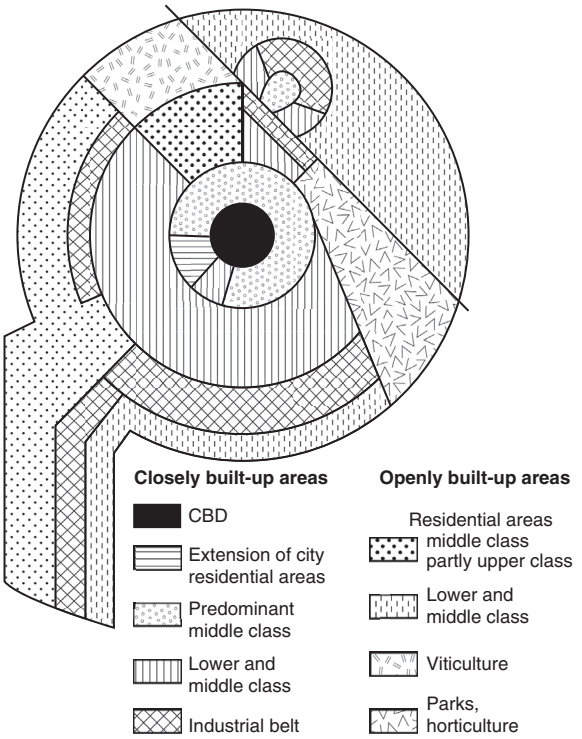

Figure 5 The dual model of Vienna. Lichtenberger, E., 1998. Stadtgeographie (Urban Geography), third ed. Teubner, Stuttgart, Leipzig, Germany. Italian translation: Geografia dello spazio urbano, first ed. Unicopli, Milano, ill. 2.17, p. 78.

extent, is the peripheral zone of vacant land around core cities and suburbs, which, even in the 1950s amounted to 20–60% of core areas (Bartholomew, 1955). Anglo-American urban geography textbooks completely ignore those huge tracts of vacant land – implying that non-utilization of these spaces is taken for granted rather than considered a problem.

2. In continental Europe, too, urban fringes attracted speculation. During World War I, they were occupied by so-called 'emergency gardening plots,' and later turned into allotment gardens. Frequently, these became temporary settlements, partly forerunners of a second-home periphery. All over post-World War I continental Europe, spontaneous settlements typically marked city fringes – a consequence of fundamental political changes. The absence of government checks on land use made for temporary settlements, among them the pavilions of 'chaotic urbanization' in France as well as the often illegal occupation of land around big central European cities (Belgrade, Budapest, Bucharest, Sofia, Vienna, and Warsaw). The succession states were less successful than the Austro–Hungarian monarchy in repelling indigent illegal immigrants. Those postwar squatters were comparable to today's squatter settlements on the fringes of Third World cities beyond the reach of state or municipal authorities.
3. As to planning and regulating the growth of agglomerations, Great Britain set the standard in the early twentieth century, with the two important concepts of the New Town (referred to above) and the 'Green Belt.' The creation of a Green Belt presupposes government control of land use, replacing the market mechanism of the liberal age. Originating in London, Green Belts have also become constituent features of zoning plans in other cities of the former British Empire. The urban development plan of Ottawa shows a Green Belt several miles wide. Persistent urban sprawl caused a characteristic overspill. In the United States too, Green-Belt concepts were introduced into various urban development plans; but in contrast to Europe, the enormous extent of suburbanization could not be contained to fulfill the original purpose of separating urban space from the periphery.
4. In the former Socialist countries, extensive public recreation areas and large private second-home districts exist (neither of these important elements of the periphery of urban regions is to be found in the USA). The public recreation area of Moscow goes back to a Green Belt idea already incorporated in the city development plan of 1935. It is 20–40 km wide and includes major sports and cultural facilities. In most of the big cities of the former USSR, the Green Belt was an integral part of development planning.

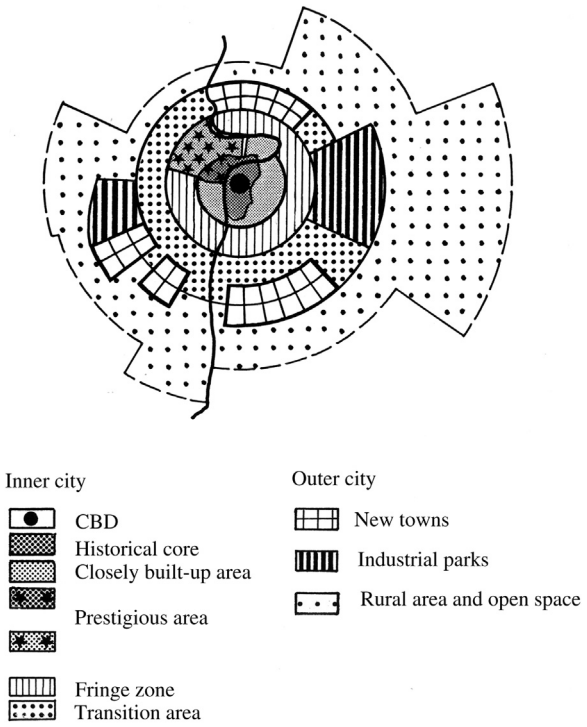

Figure 6 The model of a socialist city: Prague. Lichtenberger, E., 1993b. Wien-Prag: Metropolenforschung (Vienna-Prague: Metropolitan Research). Böhlau, Vienna, Cologne, Weimar, ill. 14, p. 102.

Urban Development in the Twenty-First Century

The New Megacities

The increase in world population from around 2 billion humans in 1925 to 7 billion in 2010, favored the growth of mega-metropolises. New global cities also have prominent quaternary sectors (financial, cultural, and political services).

In the foreseeable future, however, twenty-first century city development is unlikely to follow North American models much longer. All over the globe – in the geo-strategically and economically newly relevant BRIC-countries as well as in all countries of the Islamic world – decision makers have recognized that the inherent function of cities, above all of metropolises and capital cities, is unchanged and, with the help of state-of-the-art technologies, has even gained in importance. In these regions megacities with populations of 20 million and more surpass all European and North American metropolises. Two features are typical of them: monumental structures in their city centers are often also social centers and their adherence to the central places system.

It has never been duly recognized that the Chinese urban system has had the longest continuity by far. Already at the time of the Han dynasty (*c.* 200 BC to AD 200) Chinese regional planning authorities devised hierarchies of cities and these cities were considered 'central places' – thereby anticipating what Christaller would present as his concept of central places in 1930.

At the top of the hierarchy are the megacities, with technological infrastructures that often surpass their western models.

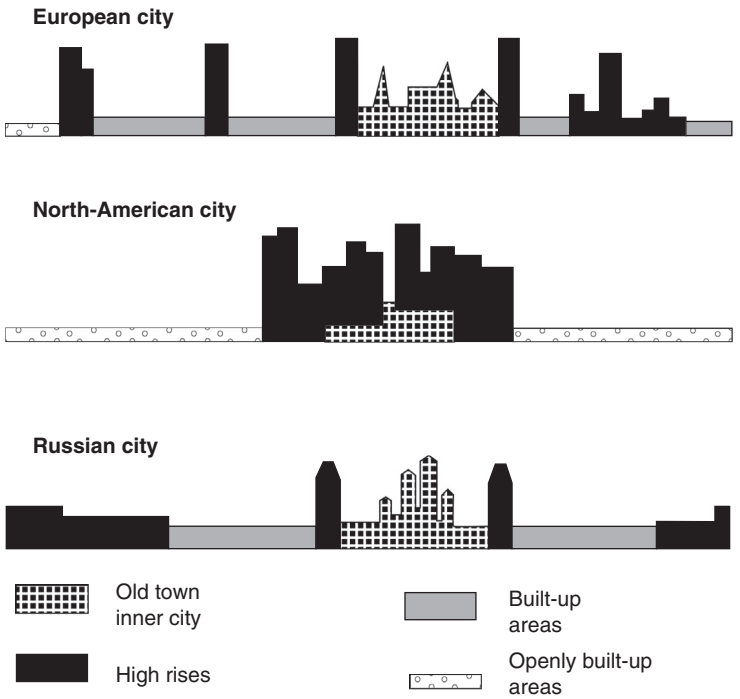

Figure 7 Skylines of European, North-American, and Russian million-cities. Lichtenberger, E., 1998. Stadtgeographie (Urban Geography), third ed. Teubner, Stuttgart, Leipzig, Germany. Italian translation: Geografia dello spazio urbano, first ed. Unicopli, Milano, ill. 3.16, p. 198.

The central places hierarchy can also be found in traditional oriental cities. Here, a spatial hierarchy of family, clan, and local quarters is based on family, religious, and ethnic affiliations. The hierarchical organization of living quarters and subcenters is paralleled in the hierarchy of semi-private, semi-public, and public institutions as well as markets (Wirth, 2000–2001).

Twenty-first century megalopolises exceed the scope and feasibilities of democratic urban planning. For 20 years, Martin Seger and his team studied the case of Istanbul and analyzed the complexity of its spatial structuring (see **Figure 8**). This is affected by a variety of factors, such as: political changes; a ribbon development of more than 80 km along the seaside and simultaneous fragmentation of urban satellites; changes in the use of land, e.g., displacement of industries and removal of workshops; developments along the highway to Ankara; construction of the Bosphorus bridges. A special feature of Istanbul is the chain of CBDs from the traditional center, the one-time core of the Ottoman Empire, to the westernized downtown areas. In addition, Istanbul is a model for other megalopolises in developing countries in respect of the upgrading of dwellings constructed over one night (so-called Gecekondu); these cover wide stretches of the periphery and, under existing law, may not be torn down (Seger, 2012).

New Variants of Urban Development in the Western World

Although the arena of twenty-first century urban development has already been moved away from Europe, from the North American countries settled from Europe, and from Australia, to regions which the twentieth century considered 'underdeveloped,' the Western world, too, will witness new variants of urban development.

Changes in the Life Expectancy of the Physical Structure of Cities in Relation to the Life Expectancy of Their Populations

Without being explicitly recognized by the agencies responsible, the relation of urban development to societal changes has altered. Up to the postwar period it was generally accepted as a fact that the life expectancy of the urban fabric was higher than that of their populations. Research into this relationship, however, has shown that this rule is no longer valid (E. Lichtenberger, 1998. Stadtgeographie (Urban Geography) third ed., p. 198). That discovery also reduces the explanatory value of the 'filtering-down' concept of socio-ecological theory, which presupposes a building fabric lasting long enough for several successive cohorts to use it and eventually run it down.

Since the life expectancy of humans increased by more than 30 years during the twentieth century, the curve of the decreasing life expectancy of the physical fabric of cities intersects the curve of the increasing life expectancy of urban population (see **Figure 9**). The point of intersection must be seen as a turning point in the relation of urban development and population development. The period of (intergenerational) changes in the building fabric of cities is ended; a period of recycling of buildings within the lifetime of a generation begins. When exactly this turning point will be reached depends on various factors, including:

- building regulations and zoning laws,
- quality of buildings: materials used and construction technology, and
- socio-economic succession potential.

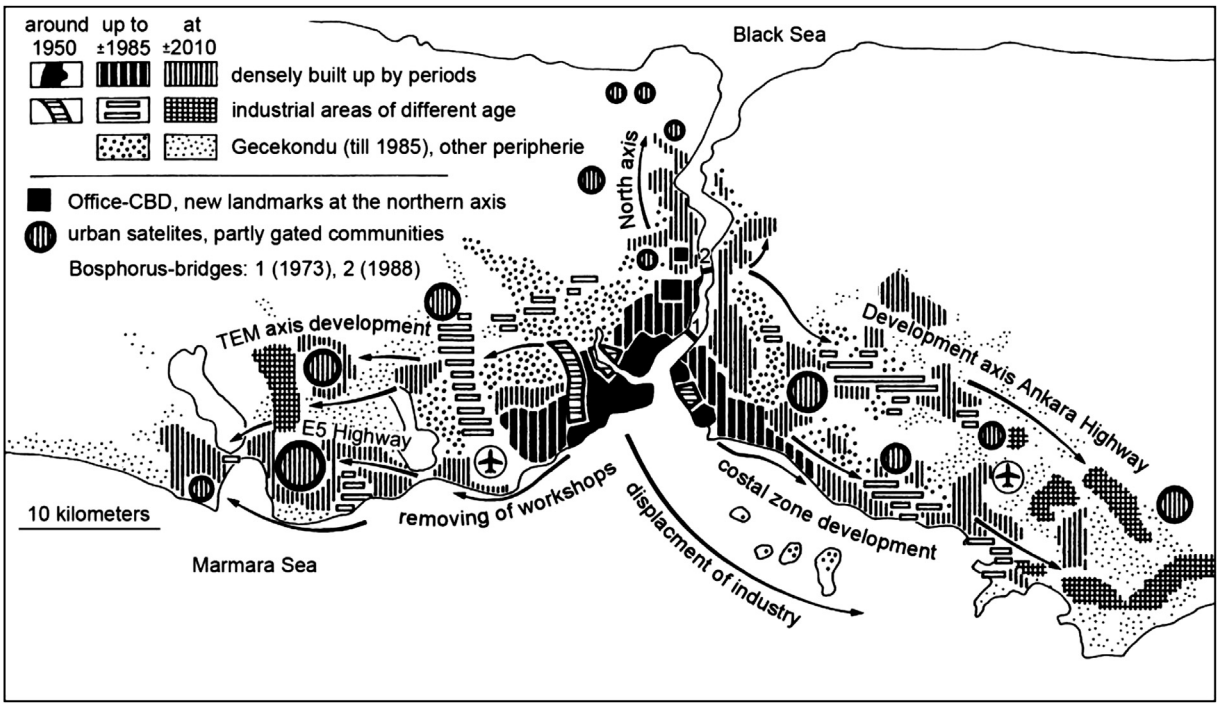

Figure 8 The Megalopolis Istanbul. Seger, M., 2012. Istanbul's backbone – a chain of central business districts (DBDs). In: Polyzos, S. (Ed.), Urban Development. InTech, pp. 201–216.

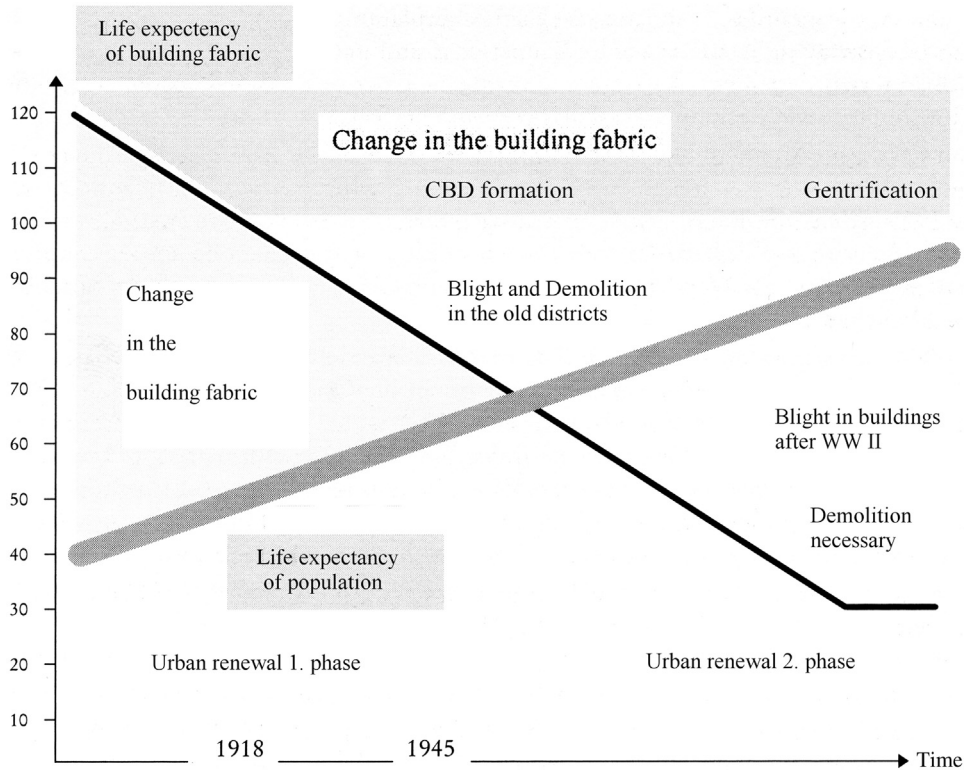

Figure 9 Life expectancy of population versus life expectancy of building fabric. Lichtenberger, E., 1998. Stadtgeographie (Urban Geography), third ed. Teubner, Stuttgart, Leipzig, Germany. Italian translation: Geografia dello spazio urbano, first ed. Unicopli, Milano, ill. 4.1, p. 275.

As a rule, the turning point will be reached at different times in different city quarters, depending on the land price gradient.

The consequences of this change may already be seen in North America, where increasing mobility has led to settlements in increasingly peripheral locations. These locations are abandoned as soon as they become outmoded or outdated.

A Dual Cycle Model of Urban Development

In urban research, the theory of cycles, derived from the theory of evolution and first applied in economics, is applied to the problems of producing urban fabric (Lichtenberger, 1991). The concept of urban development is split in two complementary notions: urban expansion and urban renewal, with urban decay (urban blight) being explained as a consequence of the time lag between the two (see Figure 10). A new urban development cycle is started by a change in the political system or in government or in consequence of a major technological innovation. At the other end, a premature or sudden stop to this process will occur only consequent to changes in the political system – examples are the two World Wars and the turning point of 1989. As a rule, such ruptures leave a considerable part of renewal work undone, with unimproved building fabric to be renewed at best in the next cycle. The fact that buildings from different time periods often exist side by side (a fact that in most cases could not be explained by the local situation alone) can be explained by reference to successive urban development cycles in which urban renewal came to a sudden end before the renewal could be completed. As to their relation in time, urban expansion has always preceded urban renewal.

Urban Decay and Urban Renewal in Different Political Systems

A comparison of political systems shows that urban decay is a feature common to all political systems in the Western world and in threshold countries. On the homepage of the London School of Economics, Director Ricky Burdett informs us: "*Fifty per cent of the world's population* currently live in cities, with 33% of city dwellers currently living in slums. By 2020, 75 per cent will live in cities, with *half the world's population living in slums*."

The following is a comparison of the United States and the transforming countries of the EU.

Urban Decay and Urban Renewal in the USA

Apart from the brief period of colonial dependence, urban development in the United States has taken place in a climate of liberalism, i.e., with a real estate market determined by the forces of supply and demand and governed by land prices, land rent, and speculation. Banks and insurance companies invest only when and where they can expect appropriate profits: rundown sections of core cities, being deserted zones, will be excluded from their investment considerations. Thus, urban renewal means either complete clearance of former residential quarters and their designation as traffic areas or a historically or museum-inspired adaptation of attractive parts as living quarters for an affluent new population segment.

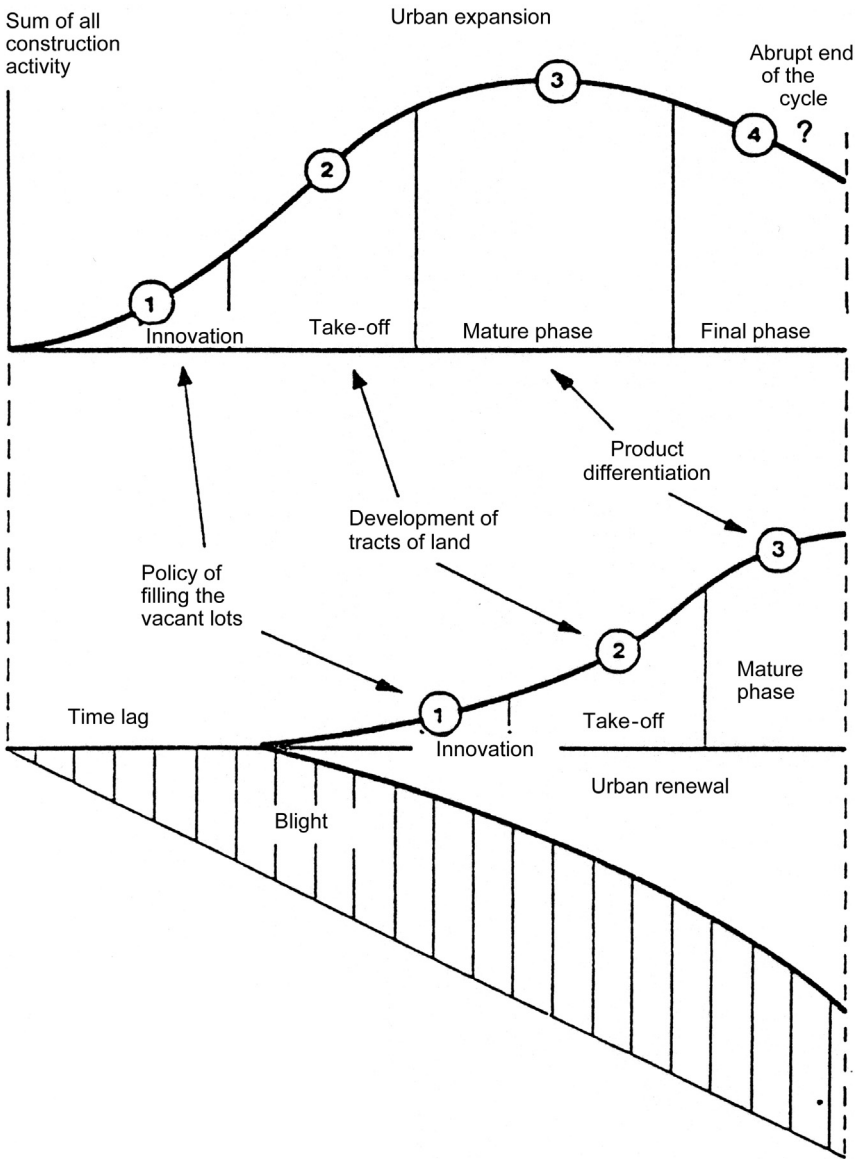

Figure 10 The two-stage model of metropolitan expansion and urban renewal. Lichtenberger, E., 1991. Product cycle theory and city development. In: Pumain, D., Saint-Julien, T., van der Haegen, H. (Eds.), Cities in Movement, Urban Systems on the Eve of the 21st Century. Acta Geographica Lovaniensia 31, pp. 120.

The extraordinary amount of new construction, hand in hand with advances in technology, has reduced the life expectancy of residential buildings to one generation. At present, new buildings invade blighted areas faster than was the case 30 years ago. Deserted buildings 'offend' no one, as there is no sense of collective responsibility for the attractive appearance of cities, apart from one's own neighborhood. Government expenditures in the United States have grown consistently over recent decades, but there is relatively little public interest in spending public money on urban revitalization (see Figure 11).

The tax system is a major factor in urbanization. Real estate taxes are the mainstay of local budgets. There is no way out of the vicious circle of marginal populations, small tax yields for local authorities, poor public services (e.g., schools and hospitals), and the physical decay of the building fabric. The indirect effects of the enormous suburban expansion of metropolitan areas are visible in all sectors of the core cities and have resulted in a staggering amount of commercial blight (documented for the first – and the last – time by Brian Berry for Chicago in 1963), of industrial blight, and residential blight.

In the later 1960s, private business became interested in two problems: the renewal of downtown areas and the provision of residential quarters for a new city population. New commercial superstructures were taken out of the public space at the same time. Thus, private interest took over the requisite areas, which had so far been public space. Eventually, three processes were

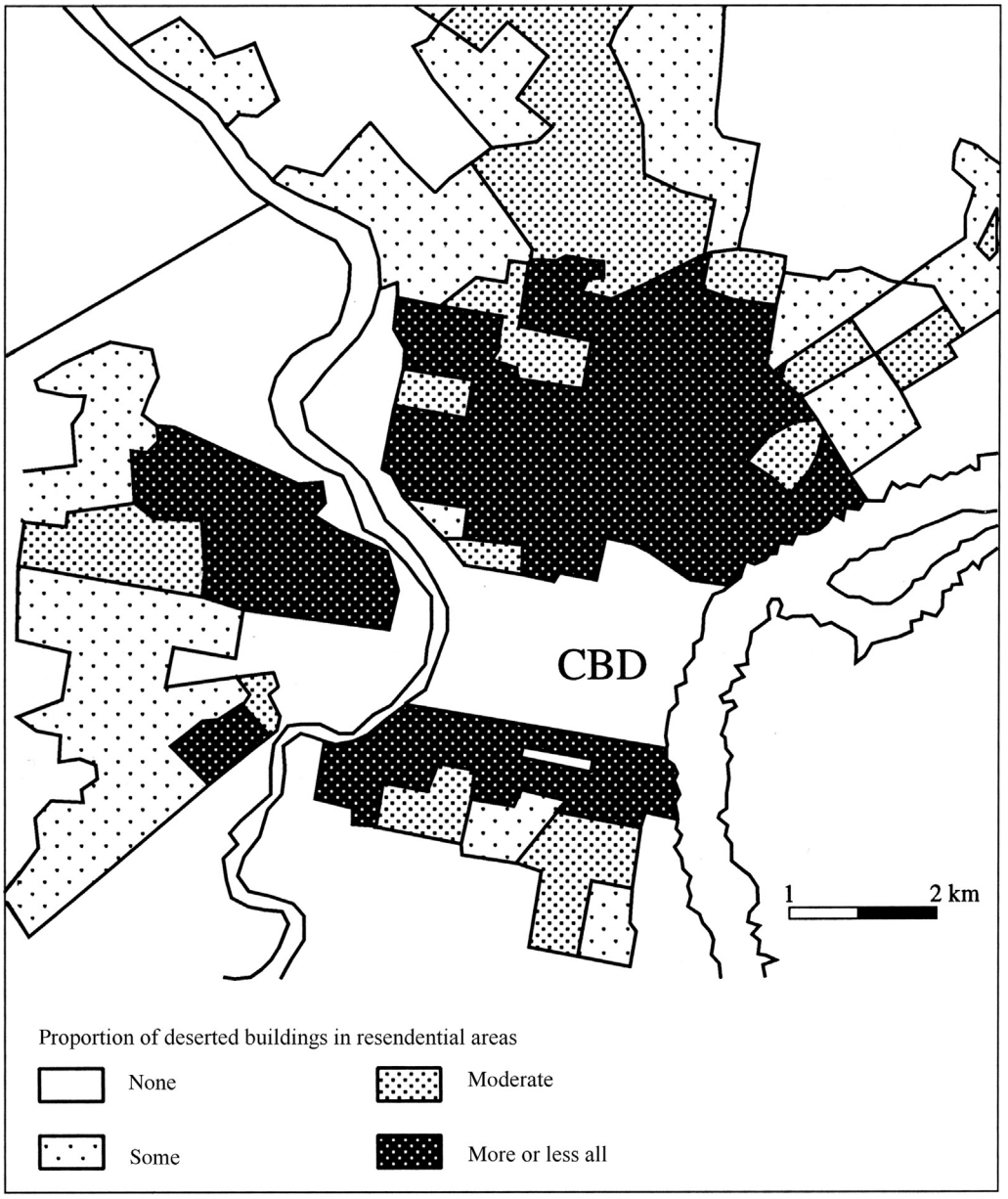

Figure 11 Decaying residential areas in Philadelphia. Bourne, L.S., 1981. The Geography of Housing. Scripta Series in Geography, London, p. 183; and Lichtenberger, E., 1998. Stadtgeographie (Urban Geography), third ed. Teubner, Stuttgart, Leipzig, Germany. Italian translation: Geografia dello spazio urbano, first ed. Unicopli, Milano, ill. 4.4, p. 288.

set in motion, the results of which are now a distinctive feature of many 'modern' cities:

Malling: Shopping malls have become a way of life in America. There are more malls than cities, colleges, or television stations. Shopping center space has increased by a factor of 12 in the last 40 years (Ashish Kumar Sen, 2005). Malls have become social centers, venues for entertainment, and employment hubs in addition to being centers for shopping (http.//www.nielsen com/en/newswire2013/).

Gating: The privatization of public space in various forms from Watched Neighborhoods in the suburbs to Common Interest Developments, to gated communities, and walled cities.

Privatopia, a combination of Gating and Malling: This is the new ghetto of postindustrial American society with its different life styles; this is the ghetto of a group that excludes itself from the 'city' and attempts to create its own rules and administration, centered in 'pseudo local authorities.'

Urban Decay and Urban Renewal in Postsocialist Countries

Postsocialist cities bear a heavy burden from the time of Communism. Core cities show extensive blighted areas: Private homes had been expropriated and administered by public authorities, but government budgets never sufficed to do both erect public housing (and requisite public transport) on a large scale and maintain the building fabric in the centers. Before

1989, the blighted districts adjacent to downtown areas bore comparison with the situation in the United States, apart from the fact that defective building fabric was not met with social marginalization.

After 1989, however, a sort of 'New Founders' Period' began. Privatization of real estate is a major consequence of liberalization; thus the functioning of city centers has become a major objective of urban planning (Lichtenberger et al., 1994).

The Locations of Office Quarters in the USA and in Europe

In North America the office sector is the offspring of business while in Europe it rests on two additional pillars: the demands of government and administrative agencies and of semiofficial institutions. As to the quarters chosen for office buildings, America and Europe differ substantially. North America has been the trendsetter of economic globalization. After World War II skyscrapers became the typical buildings of the metropolitan downtown. But Chicago's Sears Tower, erected in 1974, is nearly synchronous with starting the suburbanization of its office sector in 1975 as the last step in the suburbanization of private homes as well as of industry and business. After industrial parks came office parks and Edge Cities (Garreau, 1991). A case in point is Washington, with 16 major edge centers recently constructed. In consequence of such relocations, more than a quarter of the office spaces available in downtown areas have been deserted (see **Figure 12**).

Developments in Europe have been quite different. There, the construction of metropolitan subcities initiated by public planning (e.g., La Défense in Paris, the City Nord in Hamburg, or the UN City in Vienna) has met only a fraction of the demand. Nor do office buildings erected by big corporations make up more than a fraction of the total supply of office space (**Figures 12** and **13**).

In Europe, the construction of office towers lagged one generation behind North America. They are not located in traditional downtown areas, but rather at the edge of built-up spaces. In all compactly built-up continental cities the conversion of private apartments into offices is still a considerable factor in the creation of office space. The suburbanization of the office

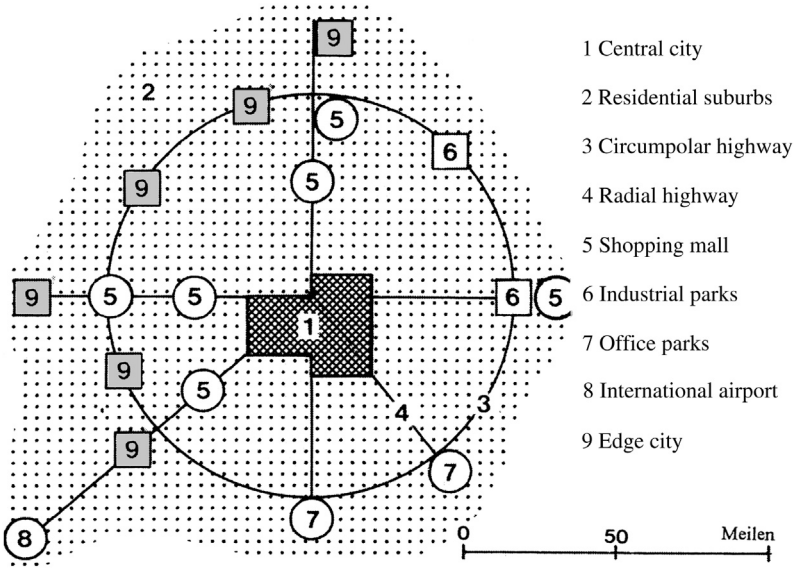

Figure 12 The model of downtown and edge cities in US metropolitan areas. Lichtenberger, E., 1998. Stadtgeographie (Urban Geography), third ed. Teubner, Stuttgart, Leipzig, Germany. Italian translation: Geografia dello spazio urbano, first ed. Unicopli, Milano, ill. 2.5, p. 48.

Figure 13 The changing logo of Syracuse from an industrial to a leisure city. Lichtenberger, E., 1998. Stadtgeographie (Urban Geography), third ed. Teubner, Stuttgart, Leipzig, Germany. Italian translation: Geografia dello spazio urbano, first ed. Unicopli, Milano, ill. 5.8, p. 334.

sector has now advanced as far as the Trade Centers next to international airports. Edge cities are nowhere in sight.

From the City of the Labor Oriented Society to the City of the Leisure Society

Urban planning in Europe is based on the organizational model of a labor oriented society, with the most important functions of the tertiary and quaternary sectors to be discharged in the city center. In recent decades, however, the organizational model of a consumer society has come to the fore, at least in North America, where the consumers are responsible for the takeoff of suburbia.

By now, the city of the consumer society is being rivaled by the city of the leisure society. Wherever it was newly established, like in Florida and California, later in France, in the Alps and around the Mediterranean, this type of city depends on favorable climatic conditions. At present European countries tend to locate their leisure societies in rural areas – with the effects of causing rhythmical over-crowding and desertion of the countryside in the course both of weeks and years. While the labor society depends upon the redistribution mechanisms of the welfare state as to its organization and financing, the commercialization of leisure takes place in a near-perfect capitalist system. By way of direct and indirect government subsidization of the tourism industry, today's leisure society is actually subsidized by the labor society, i.e., by the gainfully employed: e.g., by concessionary rates in public transport or accident insurance services.

Among the most recent developments is "city marketing," a consequence of the discovery that not only the *ville festivale* offers a considerable potential of commercialization in the steadily growing city tourism. The idea is that of marketing the image of a city as an export article. Subsidized by the French government, Paris constructed the event city "Eurodisney" as part of the *agglomération*. In North American metropolises one finds examples of event cities offering all sorts of entertainment, catering and fun. There are also examples of employing a city's "logo" in state-of-the-art advertising (see **Figure 13**). The recreational quality of a city for its own inhabitants is not affected as long as a temporal and spatial separation of residents and foreign leisure population is possible (motto: "When the tourists come, the Romans leave their town").

Acknowledgment

Thanks to Monika Streissler for her assistance with this update.

See also: Access: Geographical; Local Economic Development; Residential Segregation: Geographic Aspects; Urban Geography; Urbanization in China.

Bibliography

Alonso, W., 1964. Location and Land Use. Towards a General Theory of Land Rent. Cambridge, MA.
Ashish Kumar Sen, March/3–6 April 2005. The Malling of America. SPAN.
Bartholomew, H., 1955. Land Uses in American Cities. Cambridge, MA.
Berry, B.J.L., 1963. Commercial Structure and Commercial Bligh. Univ. of Chicago, Dept. of Geography. Research Paper 85.
Bourne, L.S., 1981. The Geography of Housing. Scripta Series in Geography, London, p. 183.
LSE. In: Burdett, Ricky, Sudijc, Deyan (Eds.), Living in the Endless City. Phaidon, London.
Carter, H., 1997. The Study of Urban Geography, fourth ed. Arnold, London.
Garreau, J., 1991. Edge City. Life on the New Frontier. New York.
Harris, C.D., Ullman, E.L., 1945. The nature of cities. The Annals of American Academy of Political and Social Sciences 242, 7–17.
Harris, C.D., 1997. The "nature of cities" and urban geography in the last half century. Urban Geography 18 (1), 15–35.
Heinritz, G., Lichtenberger, L. (Eds.), 1986. The Take-off of Suburbia and the Crisis of the Central City. Erdkundliches Wissen, vol. 76. Steiner, Wiesbaden, Germany.
Herbert, D.T., Thomas, C.J., 1997. Cities in Space: City as Place, third ed. Fulton, London.
Hoyt, H., 1939. The structure and growth of residential neighborhoods in American cities. Federal Housing Administration, Washington, DC.
Knox, P.L., 2009. Urban Social Geography: An Introduction, sixth ed. Longman, Singapore.
Lichtenberger, E., 1976. The changing nature of European urbanization. Urban Affairs Annual Reviews 11, 81–107.
Lichtenberger, E., 1991. Product cycle theory and city development. In: Pumain, D., Saint-Julien, T., van der Haegen, H. (Eds.), Cities in Movement, Urban Systems on the Eve of the 21st Century. Acta Geographica Lovaniensia 31, pp. 88–93.
Lichtenberger, E., 1992. Political systems and city development in western societies. A hermeneutic approach. Colloquium Geographicum 22, 24–40.
Lichtenberger, E., 1993a. Vienna: Bridge between Cultures. Wiley.
Lichtenberger, E., 1993b. Wien-Prag: Metropolenforschung (Vienna-Prague: Metropolitan Research). Böhlau, Vienna, Cologne, Weimar.
Lichtenberger, E., 1994. The future of the European city in the West and the East. European Review 3 (2), 182–193.
Lichtenberger, E., 1998. Stadtgeographie (Urban Geography), third ed. Teubner, Stuttgart, Leipzig, Germany. Italian translation: Geografia dello spazio urbano, first ed. Unicopli, Milano.
Lichtenberger, E., 2011. Die Stadt. Von der Polis zur Metropolis. (The City. From the Greek Polis to the Metropolis), second ed. Wissenschaftliche Buchgemeinschaft, Darmstadt.
Lichtenberger, E., 1970. The nature of european urbanism. Geoforum 4, 45–62.
Lichtenberger, E., Csefalvay, Z., Paal, M., 1994. Stadtverfall und Stadterneuerung in Budapest. (Urban Decay and Urban Redevelopment in Budapest). In: Beiträge zur Stadt- und Regionalforschung, vol. 12. Austrian Academy of Sciences Press, Vienna.
Seger, M., 1978. Teheran. Eine Stadtgeographische Studie. Springer, Vienna, New York.
Seger, M., 1992. Segregation of retail facilities and the bipolar city center of Teheran. In: Adle, Ch., Hourcade, B. (Eds.), Teheran Capitale Bicentenaire. (Paris–Teheran), pp. 281–298.
Seger, M., 2012. Istanbul's backbone – a chain of central business districts (DBDs). In: Polyzos, S. (Ed.), Urban Development. InTech, pp. 201–216.
Seger, M., Palencsar, F., 2006. Istanbul: Metropole zwischen den Kontinenten. In: Urbanisierung der Erde, vol. 10. Bornträger, Berlin-Stuttgart.
Short, J.R., 1996. The Urban Order: An Introduction to Cities, Culture and Power. Blackwell, Oxford, MA.
Wirth, Eugen, 2000–2001. The concept of the Oriental city. Privacy in the Islamic East versus public life in Western culture. Environmental Design: Journal of the Islamic Environmental Design Research Centre, 10–21.

Cities Under Postsocialism

Luděk Sýkora, Charles University in Prague, Prague, Czech Republic

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In postsocialist cities, urban environments formed under socialist regimes are being adapted to new conditions shaped by the transition to capitalism. With the establishment of market rules and pluralistic democracy, social practices of individuals, firms, and governments became imprinted in urban landscapes. Urban change has been especially influenced by internationalization, economic restructuring, social differentiation, postmodern culture, and neoliberal political practices. The main transformations in the spatial organization included the commercialization of city centers, the dynamic and spatially selective revitalization in inner cities, and the radical transformation of suburban hinterland affected by sprawl.

Introduction

This article discusses urban change in postsocialist cities. The term postsocialist refers to societies which had been prior the revolutions of 1989 known as 'socialist' and which have been since 1989 undergoing transformations toward capitalist societies with market-based economy and pluralistic democracy in political decision making. While the former socialist and so called 'second' world (Andrusz, 2001) included countries on most continents, this article mainly builds on the experience and situation in major capital cities of postsocialist Central and Eastern Europe, since these are the cities that have been studied the most. The generalization based on these cities may also be relevant for other postsocialist cities.

The city is understood as complex dynamically evolving sociospatial formation. It is internally differentiated and structured not only in terms of social, economic, or political differences between individuals and social groups, but especially in terms of local sociospatial formations such as residential neighborhood including ethnic enclaves of immigrants, gated communities of better off population, middle-class suburbs, or ghettos of socially excluded, as well as functional zones such as industrial areas, office districts, shopping, and entertainment nodes. The city is also integrated through interaction between people, firms, and organizations, between residents, entrepreneurs, and politicians, and, as these actors are spatially embedded, also between local sociospatial formations integrated into an urban region.

This article is focused on urban change, on processes through which the city and places within urban space are produced, reproduced, and transformed. It emphasizes relations between social processes and spatial structures via social practices. In established societies, the social practices are guided by a set of broadly accepted rules. The unfolding daily life framed within relatively stable institutions results in relatively stable spatial patterns. The correspondence between social institutions, social practices, and structures and spatial arrangements might gradually change in the course of ongoing evolutionary social processes, such as transition from Fordism to post-Fordism, industrial to postindustrial society, and social to neoliberal state or during second demographic transition.

However, the evolutionary relatively gradual processes of urban change may be interrupted by periods of radical urban restructuring influenced by turbulent revolutionary transformations in political and economic principles of societal organization. This is the case of postsocialist cities, where the institutional setup was radically changed in the 1990s. The capitalist political economy leads to distinctively different social practices and urban spatial organizations. Contradictions between the inherited socialist urban patterns and newly established capitalist rules created tensions resulting in the restructuring of existing urban areas and the formation of new postsocialist urban landscapes (Sýkora, 2009). Even though the principles that influence the production of urban environments were changed quickly, the pace of change in the built environment, land-use patterns, and residential differentiation is much slower. The urban transformation of postsocialist cities is an unfolding process that may take decades. Revolutionary transformations in urban spatial organization, such as dynamically developing sprawl and increasing residential segregation, pose significant challenges. They call for anticipation of future developments and critical engagement in debating urban change.

The postsocialist cities offer a chance to trace the impact of transformations in political and socioeconomic forces and changing social practices on the logic of urban space generation under the condition of revolutionary, radical, and rapid transformations framed by the transition from communist totalitarian to democratic political regimes and from command to market economies. We can not only compare how two distinctively different societies of socialism and capitalism produce their own urban spaces, but in an online perspective observe how capitalism appropriates and reconstructs the former patterns to its own needs.

There is a widespread recognition of the diversity of urban transformation in former socialist countries indicating that different postsocialist trajectories may be developing (Tosics, 2005). Differences have emerged between the countries that have managed to apply more comprehensively the Western concept of capitalism, and those whose development is more based on the locally specific recombinations of selected aspects of socialism, capitalism, and unique features that have emerged during postsocialism. However, the rejection of communism and the acceptance of capitalist features have placed the postsocialist societies, countries, and cities on a similar trajectory leading away from communism and toward various forms of

capitalism. In this article, while acknowledging the contextual embeddedness of urban transformations and their consequent local specific features and trajectories, the general and common features developing under the imperative of globally spreading capitalism will be highlighted.

The article starts with the reflection of the trajectories, patterns, and underlying forces of urbanization under socialism. This is followed by a section that overviews the recent transformations toward capitalism, juxtaposing two different political and economic systems with their own logic of urban space generation. Specific attention is devoted to relations between socialist legacies and global capitalism in shaping postsocialist urbanization and to major challenges that postsocialist cities face on their transformation path.

Socialist City

The studies of cities in Central and Eastern Europe suggested that a socialist city existed as a specific entity different from cities in developed capitalist world (French and Hamilton, 1979; Smith, 1996). However, while some scholars emphasized the specific nature of socialist urbanization referring to distinctively different socialist and capitalist political and economic systems (Szelenyi, 1983), others advocated similarities between urbanization in socialist and capitalist countries highlighting the universal influence of industrialization and modernization (Enyedi, 1996). While we can accept that both socialist and capitalist cities were variants of a more general twentieth century modern urban development, important features, such as lower levels of urbanization, less urbanism, and distinct urban spatial structure (Hirt, 2013), distinguished cities in socialist and capitalist countries.

Similarly to capitalist cities, the socialist city was shaped by cultural modernity, industrialization, technological progress, and advanced division of labor. However, the particular character of socialist political and economic organization mediated through the state planned distribution of resources and planned urbanization (Musil, 1980; Smith, 1996) yielded new economic, social, and territorial patterns. They included the lower share of services and higher share of industries in terms of employment and land use, sociospatial disparities specific to socialist production and allocation of housing, and a more compact urban morphology in comparison with capitalist cities. Under the condition of modernity, socialist cities represented the major alternative to capitalist urbanization.

The built environment, land use, and segregation patterns in socialist cities were shaped by political and economic forces that were radically distinct from capitalism. The socialist regime claimed to attain a socially just society with economy serving wider population needs. The essential starting point in forming socialist societies was the elimination of private ownership of means of production and market allocation of resources, which were seen as the key mechanisms being responsible for producing social inequalities. The essential aspects of the socialist system were the common ownership of the means of production and the administrative allocation of resources overseen by the state authorities guided by the power of Communist parties. The flow of resources between governments, firms, and households was regulated by the top-down hierarchically organized system of central planning. The economic growth shall guarantee an ever-increasing quality of life and socially equitable distribution of resources. However, the priority given to production and strategic (mostly heavy) industries, led to the underdevelopment of consumer services. The inefficiencies of central planning, especially its low capability to stimulate innovations, resulted since the 1970s in exhausting inner resources and slowing down economic and social development.

The industrialization of the socialist economy was at the nexus of development priorities. It led to the boom of urban industrial centers that attracted rural migrants pushed away from their villages by the collectivization of agricultural land and the mechanization of agricultural production. As a result, the socialist countries experienced a fast growth in their levels of urbanization. However, the growth of urban jobs was not paralleled by a corresponding supply of new housing. A significant portion of the rural population employed in cities retained their rural residence leading to an increasing share of rural residents commuting to cities, using mass public transit systems. The phenomenon of discrepancy between the urbanization of jobs and housing was described as under-urbanization (Murray and Szelenyi, 1984; Szelenyi, 1996).

Socialist cities and their metropolitan areas were managed through the top-down centralized control of territorial development as a hierarchically organized system. With land development under the control of state authorities, government policies concentrated the spatial allocation of public investments and thus urban development in few target areas. First of them was the expansion of industrial capacity by the development of new and the extension of existing industrial zones stimulated by the preference of production over consumption and political goal of keeping manual workers as the revolutionary force in major cities. Second, the construction of massive housing estates at the urban edges provided standardized supply of new housing. Third, a dense network of public transportation was developed to assure effective commuting between places of residence, work, services, and leisure activities. Fourth, services were provided in hierarchically organized system of neighborhood, district, and city urban centers, in order to guarantee economies of scale as well as even accessibility by population. Finally, city centers were redesigned as monuments of the social and economic prosperity achieved under the leadership of ruling communist parties.

Although central areas concentrated retail and other consumer services, they did not resemble the picture of capitalist urban downtowns dominated by commerce. In societies governed by central planning machinery, there was no room for financial and other business services. Similarly, the state organized supply of basic and standardized good has not required the development of retail and other consumption-oriented services on the scale known from capitalist societies. Consequently, socialist city centers retained substantial proportion of residential function. The political strategy to keep manual workers as the revolutionary force in major cities as well as nonexisting market pressure on a more economically efficient utilization of land purposefully maintained a higher share of industry in terms of both employment and land use within cities. Industrial areas penetrated from old inner city

quarters to newly established outer city zones. Sections of the inner city presocialist neighborhoods decayed due to the disinvestment of the state whose prime concern was to build new socialist areas. Renewal programs applied in last two decades of socialism often led to brutal redesign of presocialist inner-city landscapes. The majority of new construction concentrated at city edges in high-density residential areas of multistory apartment buildings and new industrial zones. In a contrast to sprawling North American and West European metropolitan areas, the socialist cities developed without suburban residential areas and nonresidential ribbon developments typical for capitalist urbanization. The extent of urban industrial land use and the scale of housing estates constructed for several thousands of inhabitants, and the consequently high proportion of industrial labor force and multiple-family housing noticeably distinguished the socialist city from the capitalist city.

Population in urban space was primarily differentiated by its demographic characteristics (Musil, 1987). The urban landscapes build under socialism looked like a mosaic of relatively large housing estates each containing particular age cohort of population with similar positions in family and life cycle. The formation of such pattern was conditioned by the state financed housing construction, which was spatially concentrated in large districts built-at-once, by the administrative allocation of dwellings to young families with children at the start of their family and housing career, and low levels of overall migration after retaining first housing. Particular generations of housing estates corresponded with certain age cohorts of population. Despite the low level of residential socioeconomic differentiation, sociospatial inequalities existed due to the access of people with higher societal and communist party merits to better housing. Large areas, characterized by social mix of households and less segregated and more evenly distributed homes of working class population, were contrasted with socialist elites concentrated in few exclusive neighborhoods. In general, the socioeconomic status declined from city centers toward urban hinterland, with the exception of higher social status of some well-located new housing estates. Most socialist countries were ethnically homogeneous and with the exception of Roma concentrations, ethnic differentiation was not a distinguishing feature of urban social landscapes. However, in the cities within the Soviet Union, the sociospatial pattern was more complex due to the migration of ethnic minorities within USSR (Smith, 1996).

City in Transition: Postsocialist Urban Transformations

Since the collapse of communist regimes in Central and Eastern Europe, cities in the former socialist countries have entered a period of dramatic reconfigurations. In the 1990s, these countries abolished socialism and set on the path toward capitalism with democratic governments and free markets. The installation of market rules and pluralistic democracy began to change social practices of individuals, households, firms, organizations, and governments leading to the development of new social processes reshaping social structures that have been for decades moulded by the socialist regime. The social changes became imprinted in urban landscapes bringing dynamic sociospatial reorganization of former socialist cities. Urban environments formed under the socialist system have been adapting to the new circumstances generated by the political, economic, social, and cultural changes associated with the transition to capitalist market economy and democratic political regimes (Enyedi, 1998; Hamilton et al., 2005; Tsenkova and Nedović-Budić, 2006; Stanilov, 2007).

In the 1990s, former socialist countries abolished socialism and set on the path toward the establishment of capitalism with democratic governments and free markets. The process of societal change is generally referred to as the postsocialist transition. The term transition has been initially associated with the neoliberal agenda of shock therapy, based on the radical replacement of the basic political and economic institutions of socialism with democratic and market arrangements. However, a number of experts have responded to the visions and rhetoric of neoliberal designers of the transition pointing to path dependencies and impacts of socialist legacies (Nielsen et al., 1995; Pickles and Smith, 1998). Using the term transformation they accentuated continuities and highlighted the hybrid nature of postcommunist realities with respect to the recombination of socialist and capitalist elements as constituents of the new postcommunist society (Stark, 1996). This body of work has contributed to the realization that postcommunist transition is not a simple replacement of one social system with another, and that real-life transformations include not only new social forms but also a recombination of new and old elements. The postsocialist transition is now commonly understood as a broad, complex, and a lengthy process of social change (Herrschel, 2007).

The radical change in the political and economic setup has stimulated transformations in virtually all spheres of societal development with its imprint into changing urban structures. The notion of postsocialist city has provided an umbrella concept for insights into urban restructuring stimulated by the government reforms of the 1990s. Two decades since the collapse of former Communist regimes, the relevance of postsocialism began to be put into question. However, while systemic institutional transformation may be over, postcommunist cities are still very much undergoing transformations in their built environment, land-use, and sociospatial patterns (Sýkora and Bouzarovski, 2012). The issue is whether the processes of urban change, such as gentrification or suburban sprawl are new forms of neighborhood change and city development under capitalism or whether they are related to the adjustments of inherited socialist urban legacies to new societal conditions. As the development of these processes is stimulated by structural deficiencies produced under socialism that has sprung up with the establishment of capitalist market economy, they should be interpreted as an integral part of the overall postsocialist transformations.

Transformations take place at many levels and include processes occurring at different speeds (Sýkora and Bouzarovski, 2012). While the systemic institutional transformations have been relatively fast, the pace of changes in social practices has been much slower. The complex environmental systems of cities and urban regions, which embody not only social institutions and practices but also physical artifacts,

have the longest time-span of change and persistence. Distinguishing three levels of postcommunist transition: "*i*) the institutional transformations that created a general societal framework for transition, *ii*) transformations of the social, economic, cultural and political practices exhibited in the everyday life of people, firms and institutions and resulting in social restructuring, and *iii*) the transformation dynamics of urban change"; Sýkora and Bouzarovski (2012) argue that: "The key aspect for understanding post-communist urban change is the distinction between the short-term period when the basic principles of political and economic organization are changed, the medium-term period when peoples' behaviors, habits and cultural norms are adopted to a new environment and transformations in a number of spheres and begin to effect broader societal change, and the long-term period in which more stable patterns of urban morphology, land use and residential segregation are reshaped."

In the first years of the transition, the leading force of transformations was played out by systemic government-controlled reforms that aimed to establish a capitalist system based on a pluralist democracy and a market economy, and to integrate socialist countries into global capitalist economy. The establishment of democratic political system with free elections and the plurality of political forces was followed with economic reforms aimed at the introduction of market system through the privatization of state assets, the liberalization of prices and free foreign trade relations. These reforms were necessary preconditions for other societal transformations to take place through the changing practices of firms and people in turn reflected in changing economic, social, and cultural environments. Urban change has been especially influenced by internationalization and globalization, economic restructuring in terms of deindustrialization and the growth of producer services, increasing social differentiation, new modes of post-modern culture, and neoliberal political practices (Sýkora, 1994; Stanilov, 2007).

International companies and foreign direct investments became leading agents in reshaping local economies and urban geographies. Foreign investment in real estate reflected the demand for space generated by international firms in producer and consumer services as well as in industrial production and started radically reshape urban landscapes and skylines. Foreign managers and high-salary employees as well as low-paid immigrant workers formed new mutually polarizing segments of population on the housing market. Deindustrialization, which presented the downward side of economic restructuring, has been in many placed counterbalanced by the growth of new industries and consumer services. However, both offering mostly low paid-jobs. Only major urban centers, usually capital cities, benefited from the growth in advance producer services. Economic change mediated through labor markets was expressed in increasing wage and income disparities that contributed to social polarization among households, formation of new sociospatial formations, and increasing territorial disparities. Social inequality and conflicts appeared in initially, socially, and ethnically homogenous urban societies.

New conditions allowed for the development of a greater plurality of values, with particularly strong tendency toward individualism and promotion of self-interests. This impacted not only rapid decline of family formation leading to aging urban populations but especially rapid proliferation of consumption oriented life style. National and urban governments often pursued political strategies grounded in a neoliberal discourse, which saw free, unregulated market as the mechanism of allocation of resources that would generate a wealthy, economically efficient, and socially just society. While the state was perceived as the root of principal harms to the economy and society, urban transformations have often been left to the operation of free market partially bound within the framework of traditional rigid physical planning instruments. In such constellation, powerful investors have often steered politicians and planners to facilitate urban development in a manner favorable to capital (Altrock et al., 2006; Horák, 2007)

In postsocialist cities, the urban environments inherited from socialist era, have been adapted to the new conditions created by government reforms and ensuing economic, social, and cultural transformations. With the introduction of capitalist urban economy, the contradictions between capitalist rules of the game and socialist urban environment led to the restructuring of the socialist urban features and structures and the formation of new capitalist urban landscapes. Most urban growth and change concentrated in city centers, selected inner city nodes and areas, and in rapidly expanding suburban zone (**Figure 1**).

First years of transition were characterized by the rapid commercialization and physical regeneration of expanding city centers accompanied by the decline of residential functions, densification of land uses, traffic congestion, and conflicts with historical heritage. In the course of time, urban change spread to inner city areas, which have been affected by both growth and decline. Deindustrialization left behind extensive brownfields, which presents not only the threat of decline but also potential for redevelopment. Urban restructuring was driven by the replacement of inherited land uses with new and economically more efficient activities. In major cities, new commercial growth concentrated in out-of-centre office clusters and retail and entertainment complexes. Besides the formation of new secondary commercial nodes and ribbon development areas along highways, waterfronts received particular attention of development industry.

The shifting housing preferences placed many socialist housing estates and some older presocialist neighborhoods on downwards spiral. On the other hand side, gentrification began to bring some inner city quarter to their former glory and prestige. Together with newly constructed condominium districts and gated residential areas reflected desires of new middle classes to live apart. Considering the formation of immigrant enclaves and ghettoization of Roma and socially vulnerable populations in housing estates in declining old industrial cities, the new residential landscapes began to exhibit bold segmentation in housing supply and evolving segregation processes. With the increase in personal wealth of population, people started to move to urban hinterlands to fulfill their dreams of single family home. Sprawling residential districts have been accompanied with big-box retailing, warehousing, and industrial districts located along highways and on their major intersection. Suburbanization has become the most dynamic process changing the landscapes of postsocialist city regions (Stanilov and Sýkora, 2014).

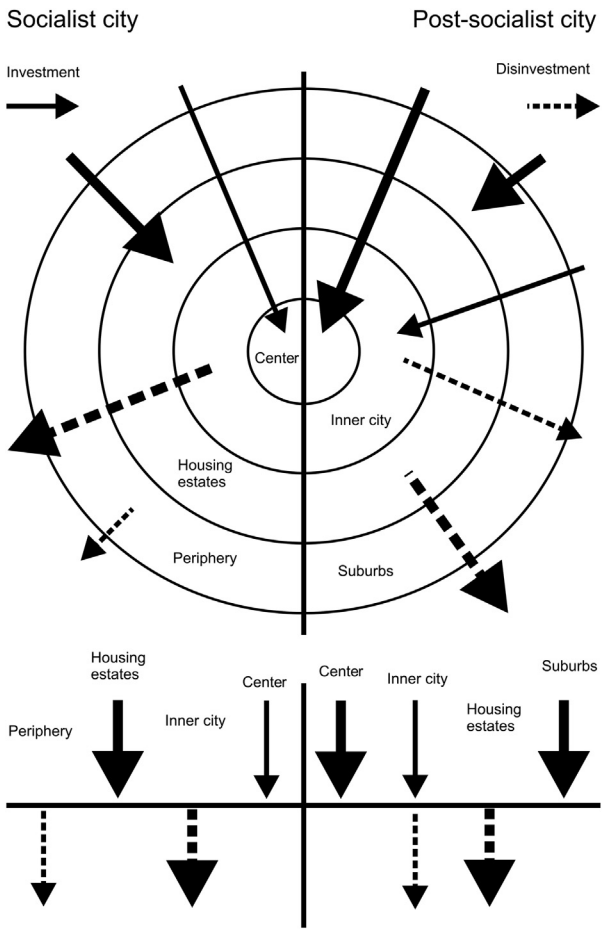

Figure 1 The pattern of investment and disinvestment in socialist and postsocialist city. Source: Sýkora, L., 2009. Postsocialist cities. In: Kitchin, R., Thrift, N. (Eds.), International Encyclopedia of Human Geography. Elsevier, Oxford, pp. 387–395.

Global Capitalism, Socialist Legacies and Challenges of Postsocialist Urbanization

One of the key debates in postsocialist urban studies has been evolving about the issue of interaction between socialist legacies and neoliberal capitalism. Gentile et al. (2012) coin the term heteropolitanization referring to increasing internal urban complexity and heterogeneity of postsocialist city characterized with the simultaneous emergence of new social and spatial patterns in an environment, where powerful legacies preserve the patterns inherited from the previous era" (Gentile et al., 2012: p. 297). Golubchikov et al. (2013) emphasize the "mutual embeddedness of the legacies of socialism and the workings of neoliberal capitalism that jointly produce … the hybrid spatialities of transition" (Golubchikov et al., 2013: p. 618). They go in their argument even further speaking about "the subsumption of the inherited systems into the very logic of capitalist and neoliberal relationships, so that, while capitalist practice feeds on the legacies of state socialism, it also alienates them from their history to make them the infrastructure, and often the agency, for its own expansion" (Golubchikov et al., 2013: 630).

The postsocialist developments bring the re-emergence of some presocialist patterns, transformations in some areas from socialist times and creation of new postsocialist urban landscapes. While the present urban patterns can be interpreted as hybrid (Gentile et al., 2012), in the course of time, new capitalist urban developments will be gaining even more important role in the overall urban organization. Postsocialist city shall be seen as a temporary phenomenon existing in the time-period between the institutional reload through the replacement of the socialist with the capitalist rules of the game and the completion of corresponding transformations in urban built and social environments (Sýkora, 2009). While many features resembling socialism will certainly remain present even in distant future, the essence of daily urban life and city operation will be capitalist.

The dynamic development of suburbanization, a boom of gated communities and gentrification of inner city quarters, the emergence of socially excluded localities and enclaves of immigrants, the establishment of office parks and shopping districts, all are linked with the expansion of capitalism to former second world countries. These similar forms of urban developments, nowadays observable across the globe, are the results of the social practices of firms, households, and governments under the political, economic, and cultural conditions of globally spreading capitalism. Global capitalism has been the most powerful agent in the spread of similar urbanization trends and features across the former socialist countries of Europe, as well as in many other regions around the world. The impact of global capitalist forces on postsocialist urban change is, nevertheless, far from uniform. While the global capitalist system embraces nearly the whole planet, individual countries, regions, cities, and neighborhoods are integrated into it in a highly uneven manner. In postsocialist countries, the legacies of socialism, the character of new capitalist institutional setup, the size and attractiveness of market, etc. produce internally differentiating local outcomes of global capitalism.

The most distinguishing feature of local postsocialist urban restructuring under the imperative of global capitalism is the spatial selectivity of investments and territorially uneven growth. The desire of capital to utilize profit opportunities has not been counterbalanced by public sector striving for more socially and territorially balanced development. The urban development in the postsocialist cities has been managed through a mixture of free market practices of private firms struggling for their share of newly emerging markets and a highly decentralized, locally based, and regionally uncoordinated system of land development in city regions.

Rapidly developing suburban sprawl and residential segregation are two key processes of uneven development in postsocialist cities, symptomatic features of global urbanism spreading across former socialist countries and, at the same time, major challenges for public policies. Residential and commercial activities have decentralized into cities' hinterlands and radically transformed the compact physical morphology of the former socialist city. The dynamics of population and job growth in new suburbs has been remarkable bringing city regions in postsocialist countries among the most sprawling

areas in Europe (EEA, 2006). Suburbanization has increased the range of housing options for the middle and upper income households and brought employment and shopping opportunities closer to residents living in the hinterland of cities. At the same time, sprawl, which has become the typical form of postsocialist metropolitan growth, has undermined the prospect of achieving smart, inclusive, and sustainable development. Suburban growth is rarely recognized as an integral component of the overall process of metropolitan development, leaving it to the decisions of numerous suburban municipalities competing for their stake in the circuits of real estate investments and population flows.

The simultaneous growth of gated communities, gentrified areas, and suburbs of wealthy on the one hand side, and the degradation of some housing estates, formation of socially excluded localities, and spatial concentrations of immigrant workers, reflect tendencies toward segregation and the emergence of sociospatially polarized city (Hirt, 2012). Surprisingly, the level of sociospatial inequality has not increased during the first decades of capitalism (Marcinczak et al., 2013). While inner city gentrification and suburbanization of middle class is lifting up these residential zones from below average social status, socialist housing estates that under socialism concentrated younger and better-educated population are declining from their better-off position toward city average. The simultaneous operation of segregation processes and decline in sociospatial disparities is a temporary paradox of postsocialist transformation. In the course of time, the capitalist urban economy accompanied with neoliberal practices of urban development will strengthen and expand the new localities that concentrate populations from the opposite sides of social spectrum and capitalist segregation will definitely erase socialist legacies.

The emerging social urban problems call for the reconsideration of urban and regional planning, policies, and practices pursued since the early 1990s. The postsocialist governments placed high priority on individual choice, economic freedom, and private property rights, without recognizing that urban development is not just a matter of personal but of societal choice. Fighting suburban sprawl and residential segregation requires the coordination of urban development at the metropolitan level and in interscalar coordination with national and local governments. However, the stronger role of governments in social and territorial development has been resisted.

Conclusions

Postsocialist cities are cities under transformation. Urban landscapes formed under socialism are being adapted and remodeled to new conditions shaped by the political, economic, and cultural transition to capitalism. The development of cities in former socialist countries is now largely governed by market forces and democratically elected governments. However, the reorganization of urban landscapes in postsocialist cities that began with the reforms of the 1990s has not been completed yet. "Their defining feature remains the incessant and relentless process of structural transformation that has started to unfold since the end of communism" (Sýkora and Bouzarovski, 2012: p. 44).

Even though the series of political and economic transformations is already two decades old, the postsocialist transition is a project still not finished. At present, sprawl and segregation dramatically reconfigure urban landscapes. These urban developments can be seen as outcomes of present day choices that will impact the development of society for decades to come. Postsocialist societies are still facing critical junctures. Unfortunately, negligible attention is given to the anticipation of future developments, their critical evaluation, and the discussion of alternative development paths. Postsocialist developments can lead to the formation of new path-dependent lock-ins that these societies will later regret (Sýkora, 2008). They can learn from analogical developments and alternatives known from other countries and pay attention to the problematic developments that may be avoided if alternative development paths are taken sooner rather than later.

See also: Cities: Internal Structure; Gated Communities; Gentrification; Planning During Post-Socialism; Suburbanization and Suburbanism; Urban Neoliberalism; Urban Policy in Europe; Urban Sprawl.

Bibliography

Altrock, U., Günter, S., Huning, S., Peters, D. (Eds.), 2006. Spatial Planning and Urban Development in the New EU Member States: From Adjustment to Reinvention. Ashgate.

Andrusz, G., 2001. Cities: post-socialist. In: International Encyclopedia of the Social & Behavioral Science. Elsevier Science Ltd, pp. 1835–1840.

Enyedi, G., 1996. Urbanization under socialism. In: Andrusz, G., Harloe, M., Szelenyi, I. (Eds.), Cities after Socialism. Urban and Regional Change and Conflict in Post-socialist Societies. Blackwell, Oxford, pp. 100–118.

Enyedi, G. (Ed.), 1998. Social Change and Urban Restructuring in Central Europe. Akadémiai Kiadó, Budapest.

European Environment Agency (EEA), 2006. Urban Sprawl in Europe: The Ignored Challenge. Office for Official Publications of the European Communities, Luxembourg.

French, R.A., Hamilton, F.E.I. (Eds.), 1979. The Socialist City: Spatial Structure and Urban Policy. John Willey & Sons, Chichester.

Gentile, M., Tammaru, T., van Kempen, R., 2012. Heteropolitanization: social and spatial change in Central and East European Cities. Cities 29 (5), 291–299.

Golubchikov, O., Badyina, A., Makhrova, A., 2013. The hybrid spatialities of transition: capitalism, legacy and uneven urban economic restructuring. Urban Studies 51 (4), 617–633.

Hamilton, F.E.I., Dimitrowska-Andrews, K., Pichler-Milanović, N. (Eds.), 2005. Transformation of Cities in Central and Eastern Europe: Towards Globalization. United Nations University Press, Tokyo.

Herrschel, T., 2007. Between difference and adjustment – the re-/presentation and implementation of post-socialist (communist) transformation. Geoforum 38 (3), 439–444.

Hirt, S.A., 2012. Iron Curtains. Gates, Suburbs and Privatization of Space in the Post-socialist City. Wiley-Blackwell, Chichester.

Hirt, S., 2013. Whatever happened to the (post)socialist city? Cities 32 (Suppl. 1), 29–38.

Horák, M., 2007. Governing the Post-communist City: Institutions and Democratic Development in Prague. University of Toronto Press, Toronto.

Marcinczak, S., Gentile, M., Stepniak, M., 2013. Paradoxes of (post)socialist segregation: metropolitan sociospatial divisions under and after socialism in Poland. Urban Geography 34 (3), 327–352.

Murray, P., Szelenyi, I., 1984. The city in the transition to socialism. International Journal of Urban and Regional Research 8 (1), 90–107.

Musil, J., 1980. Urbanization in Socialist Countries. M.E. Sharpe, New York.

Musil, J., 1987. Housing policy and the sociospatial structure of cities in a socialist country: the example of Prague. International Journal of Urban and Regional Research 11 (1), 27–36.

Nielsen, K., Jessop, B., Hausner, J., 1995. Institutional change in post-socialism. In: Hausner, J., Jessop, B., Nielsen, K. (Eds.), Strategic Choice and Path Dependency in Post Socialism: Institutional Dynamics in the Transformation Process. Edward Elgar, Aldershot, pp. 3–44.

Pickles, J., Smith, A. (Eds.), 1998. Theorising Transition: The Political Economy of Post-communist Transformations. Routledge, London.

Smith, D.M., 1996. The socialist city. In: Andrusz, G., Harloe, M., Szelenyi, I. (Eds.), Cities after Socialism. Urban and Regional Change and Conflict in Post-socialist Societies. Blackwell, Oxford, pp. 70–99.

Stanilov, K. (Ed.), 2007. The Post-socialist City: Restructuring of Urban Form and Space in Central and Eastern Europe after Communism. Springer.

Stanilov, K., Sýkora, L. (Eds.), 2014. Confronting Suburbanization: Urban Decentralization in Post-Socialist Central and Eastern Europe. Studies in Urban and Social Change. Wiley-Blackwell, Oxford.

Stark, D., 1996. Recombinant property in East European capitalism. American Journal of Sociology 101 (4), 993–1027.

Sýkora, L., 1994. Local urban restructuring as a mirror of globalisation processes: Prague in the 1990s. Urban Studies 31 (7), 1149–1166.

Sýkora, L., 2008. Revolutionary change, evolutionary adaptation and new path dependencies: socialism, capitalism and transformations in urban spatial organisations. In: Strubelt, W., Gorzelak, G. (Eds.), City and Region: Papers in Honour of Jiří Musil. Budrich UniPress, Leverkusen Opladen, pp. 283–295.

Sýkora, L., 2009. Post-socialist cities. In: Kitchin, R., Thrift, N. (Eds.), International Encyclopedia of Human Geography. Elsevier, Oxford, pp. 387–395.

Sýkora, L., Bouzarovski, S., 2012. Multiple transformations: conceptualising post-communist urban transition. Urban Studies 49 (1), 41–58.

Szelenyi, I., 1983. Urban Inequalities under State Socialism. Oxford University Press, Oxford.

Szelenyi, I., 1996. Cities under socialism and after. In: Andrusz, G., Harloe, M., Szelenyi, I. (Eds.), Cities after Socialism. Urban and Regional Change and Conflict in Post-socialist Societies. Blackwell, Oxford, pp. 287–317.

Tosics, I., 2005. Post-socialist Budapest: the invasion of market forces and the response of public leadership. In: Hamilton, F.E.I., Dimitrowska Andrews, K., Pichler-Milanović, N. (Eds.), Transformation of Cities in Central and Eastern Europe: Towards Globalization. United Nations University Press, Tokyo, pp. 248–280.

Tsenkova, S., Nedović-Budić, Z. (Eds.), 2006. The Urban Mosaic of Post-socialist Europe: Space. Institutions and Policy. Physica-Verlag, Heidelberg.

Citizen Journalism

James F Hamilton, Grady College of Journalism and Mass Communication, The University of Georgia, Athens, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Citizen journalism refers to nonprofessionals taking an increasingly central role in news reporting, writing, editing, publication, and distribution. These roles are exercised both within news organizations and outside them. Studies of citizen journalism are commonly structured around a distinction between citizens and professionals, with the key issue being to what extent citizen journalism fulfills the promises of normative liberal–pluralist press theory.

Citizen journalism refers to nonprofessionals taking an increasingly central role in news reporting, writing, editing, publication, and distribution. These roles are multiple and variable, as Nip (2006) suggests in an influential typology. Although emerging primarily in the West, citizen journalism lays claim to being a global phenomenon; its Western roots have been extended and remade through countless instances in many parts of the world (Allan and Thorsen, 2009). It is part of a global reorganization of media work (Wunsch-Vincent and Vickery, 2007).

Citizen Journalism and the Profession

Citizen journalism draws its significance from a contrast with and resulting criticism of professionalized journalism. General acknowledgment of journalism as a profession aspiring to the same status as practitioners of medicine and law dates in England with the formation of the Chartered Institute of Journalists in 1884 and in the USA from the teaching of courses in journalism in the nineteenth century and by the early-twentieth century the establishment of schools of journalism in universities (Chartered Institute of Journalists, 2012; Schudson, 1978: p. 152).

Recognizing that citizen journalism gains its relevance by being distinguished from professional journalism emphasizes that its core claim is related not to an expansion of participation so much as its challenge to professionalization. The significance of citizen journalism comes from nonprofessional nonemployees taking an increasingly central role in a heretofore professional activity, thus challenging the exclusivity and authority of professionalization itself. Indeed, an institutionalized distinction between a professional class of writers and their readers/consumers can be traced to the beginnings of commercial media production (Hamilton, 2008: pp. 50–53). Forms in which nonprofessionals have routinely been engaged in professional media include as readers, listeners, and viewers; as buyers and subscribers; as contributors of letters, comments, and opinions; and as sources, subjects, and contestants (Griffen-Foley, 2004).

Despite its usefulness, this relatively unproblematic definition is quite complex in practice. One key reason is the multiple positions from which the challenge to professionalization is based. Challenges of citizen journalism emerge from both outside and inside established news organizations (Atton and Hamilton, 2008: pp. 63–76). On one hand, people outside of journalism industries see citizen journalism as a critique of the restrictions and privileges of professionalism and of the insulation more generally of journalism from its readers and society. On the other hand, citizen journalism is also practiced inside as part of management-sanctioned strategies of economic revitalization. For many publishers, managers, and other professionals, citizen journalism can help rescue the professionalized commercial model of news production that has been a dominant form in the West for more than 150 years.

A second key reason for the complexity is that the multiple locations of these challenges are not opposites. Each draws upon the other for its own claims of legitimacy and authority. On one hand, citizen journalism as fostered and practiced outside established news organizations (such as by bloggers and activists) emphasizes its greater professionalism. This is often expressed in claims to greater 'quality' and 'objectivity' compared with underfinanced, deadline-obsessed commercial news organizations that too often simply rewrite public-relations press releases. From this point of view, the problem to which citizen journalism is a response is sloppy unprofessionalism or manipulation by publishers or advertisers.

On the other hand, citizen journalism as fostered by and within established news organizations emphasizes its ostensible democratization of news production. This is often expressed via liberal Anglo-American media traditions as a way to restore the ostensible 'marketplace of ideas' by encouraging the wider diversity of participants and views it requires (Altschull, 1990). From this point of view, the problem to which citizen journalism is a response is insular preoccupation with immediate financial success and over-professionalization (McChesney and Pickard, 2011).

Internal and External Critical Positions

What became citizen journalism emerged through a series of developments and forms spawned by both internal and external critiques. Internal critiques took issue with the abrogation of social responsibility inherent in the disconnection of insulated professionals from the readerships/communities they were to serve.

During the 1960s, significant criticism of the professional convention of journalistic objectivity took place in the form of the so-called 'New Journalism.' This form of reportage

repositioned the reporter, but, unlike citizen journalism, left the reader/viewer's position untouched. Also referred to as 'literary journalism,' such writing placed the reporter as a character, if not the main character, in the story rather than as an outside, neutral observer and recorder (Hollowell, 1977; Weingarten, 2006). 'Participatory journalism' was how the work of author George Plimpton (1966) and documentarian Jon Alpert was described, although not without criticism due to not adhering to the professional form of objectivity (Dickey, 1977; Alter, 1986).

Yet, the new, literary journalism quickly became a global phenomenon (Bak and Reynolds, 2011). By the mid-1990s, additional criticism of professionalized, routinized journalism emerged from within the news industries. In a search for how to make news a more compelling product and perhaps better fulfill its role of empowering citizens to make more-informed choices, efforts emerged to involve readers by making the reporting process more transparent and thus open to public scrutiny. A significant, early effort was the *San Jose Mercury News* experiment with reporter Gary Webb's series 'Dark Alliance' on a CIA-crack cocaine connection (Webb, 1996; the book-length report is Webb, 1998). In the online stories, links allowed readers to view reproductions of a great deal of documentary evidence on which the reporting was based. The hope was that doing so allowed readers to make their own assessment of the veracity and legitimacy of conclusions drawn. The series was subjected to great hostility by the news industry, with the online version coming down and Webb's firing from the *Mercury News* (Webb, 1998: pp. 437–466; Webb, 2002; Paterno, 2005).

By contrast to expanding the writing style of journalism, other industry-centered efforts saw in citizen journalism a viable business strategy for addressing declining readerships. Such efforts built upon pre-Internet projects such as 'public journalism' and 'civic journalism' in the USA (Rosen, 1996; Black, 1997). By 2005, the industry recognized the general trend toward citizen journalism, recognizing it as largely a US and UK phenomenon, although present in other countries (Jesdanun, 2005). Unlike the aspiring national and international reach of citizen-journalism efforts outside of the industry, industry-centered efforts took shape as a hyper-localized form of commercial community journalism (an early US description is Oxfeld, 2005). While this can be seen in other countries such as England in which private ownership and commercial support forms the bedrock of the system, it is most prevalent in its development in the USA (Roberts, 2006). By the mid-2000, the increasing availability and diffusion of consumer-grade digital cameras (and, later, smartphones with video capability) enhanced the attraction of citizen journalism for television-news organizations, which not only used amateur footage more frequently, but also formulated for amateurs online basic training in visual journalism (Romano, 2005).

External critiques took issue with the selectivity and skewed nature of news coverage, which contradicted the professional ethos of impartiality and comprehensiveness. In doing so, they built upon press criticism from its earliest days (McChesney and Scott, 2004; Theobald, 2004; Goldstein, 2007).

The late 1990s brought new capabilities for social movements to address global audiences entirely outside the sphere of corporate news media. Following on the heels of various forms of community and citizen media, ranging from public-access cable channels to a community-run satellite distribution system, Independent Media Centers were established to provide grassroots viewpoints and information during the Seattle meeting of the World Trade Organization in 1999 (Howley, 2005; Halleck, 1993; Hamilton, 2008: pp. 121–159). Its innovation was to create a system of open-publishing, whereby individuals could upload to the website their own news stories and video, for viewing by potentially anyone worldwide (Kidd, 2003).

Developments in the 2000s brought together all these disparate aspects. Additional digital platforms, such as the website Blogger, soon emerged and allowed self-publishing and worldwide distribution in a form far less technical than had previously been the case. New amalgams of professional and nonprofessional news organizations, such as in 2000 the South Korean organization *OhMyNews*, emerged which exemplified a challenge to the mainstream media regime of that country by creating a sustainable hybrid non-professional/professional commercial news organization (Kim and Hamilton, 2006). Proto-national and international systems in addition to Independent Media Centers emerged by late 2004, in such examples as Wikinews (Glasner, 2004). The progressively greater integration of social media such as Facebook and Twitter into not only the organizing of social movements, but also of legitimate reporting, has only furthered these trends (Heflin, 2010).

Postpositivist Studies

Postpositivist studies of citizen journalism search for a means of codifying stable and predictable relationships (Greer, 1969). The typical research design is comparative, in which professional journalism and citizen journalism are compared directly or indirectly in terms of how well each lives up to the normative roles attributed to journalism by Anglo-American liberal theories of the press (Siebert et al., 1956; Altschull, 1990).

One common method of quantitative content analysis seeks to determine whether and to what degree citizen journalism exemplifies the liberal–pluralist normative goal of journalism of providing the broadest range of views and opinions. Examples include Carpenter (2010), which found in a comparison of coded news items that the range of diversity was greater in citizen journalism than professional journalism. Lacy et al. (2011) sought to determine whether and to what degree online citizen journalism was capable of substituting for professional journalism in the economic marketplace. Upon coding the content of online publication websites, they found that "citizen news and blog sites can serve as complements to daily newspapers," thus posing little economic threat to professional news organizations (p. 44).

Research surveys also focus on whether and to what degree citizen journalism fulfills the normative role attributed to the press by liberal–pluralist philosophy. Kaufhold et al. (2010) address the issue of whether online user-generated news informs the public in the same way as professional journalism does. While consumption of user-generated journalism was negatively related to political knowledge but positively related to participation online and off-line, consumption of

professional journalism was positively related to political knowledge and only to off-line participation. Nah and Chung (2009) found in their survey of newspaper editors that they rated the importance of professional journalism more highly than citizen journalism. And Robinson and Deshano (2011a) sought to determine the motivations of people who engaged in citizen journalism. Among other findings, citizen journalists were found to gain a host of positive affective benefits through their participation.

Functionalism

Similarly to postpositivism, a functionalist perspective also concerns itself with a comparison of professional journalism and citizen journalism in the degree to which each fulfills its normative role in a democratic society. While organistic conceptions of society as a set of interrelated systems and functions date from the earliest years of social theory, they were applied to mass communication research initially via the formulation of US political science scholar, Harold Lasswell. Whether single-celled organisms or many membered groups, each similarly seeks to "maintain an internal equilibrium and to respond to changes in the environment in a way that maintains this equilibrium" (Lasswell, 1948: p. 38). Avoiding dysfunction and systemic breakdown requires a consensus in the attention and views of leaders, experts, and various publics (Lasswell, 1948: p. 51). In some studies, the dysfunctions of the current professionalized system such as not enabling the correlation of different social groups (see Yankelovich, 1991) compare negatively to citizen journalism, which in the best cases is seen to create a working if not harmonious fit between agendas of the institution, organization, profession, and publics. Through a functionalist approach, professional journalism and citizen journalism are often compared regarding their ability to bringing about such a consensus.

Functionalist studies of citizen journalism assess it in national or international spheres. To investigate to what degree the introduction of citizen journalism upsets the economic equilibrium of profitability in various local US media markets, Lacy et al. (2011) found that while public-affairs blogs complement radio and TV (and thus pose no economic threat to their viability), they substitute for alternative newsweeklies by causing the readerships and resulting advertising revenue of the latter to drop. Xin (2010) found in China that citizen journalism had been generally incorporated into the workings of professional journalism, and its potential systemic disruption thus contained. However, the study also noted that citizen journalism outside professional journalism also existed, largely to promote various forms of nationalism. In their comparison of eight European countries and the USA, Domingo et al. (2008) found that citizen journalism had also been incorporated by professional journalism with minimal disruption, largely by restricting citizen participation to opinions and feedback, not extending it to news reporting as well.

Technological Determinism and Convergence

The relevance of technological change to citizen journalism is the focus of studies of convergence. Disruptions are attributed largely to the Internet, which provides the potential of open access and worldwide distribution. Studies of citizen journalism in this perspective become part of broader attention to changes throughout the full range of creative industries, thus coinciding with other fields of study such as sociological studies of labor, work, and workplace organization (Hesmondhalgh, 2002; Banks, 2007; Deuze, 2007).

This perspective takes a functionalist interest in systemic analysis, but extends it geographically from discrete localities to a globalized process (Castells, 1996). Indeed, in his discussion of what he calls 'network journalism,' Heinrich (2011) notes as much in his assertion that the emergence of citizen journalism is just one instance of a transformation in media work "driven by two processes: globalization and technological advances in the form of digitization" (p. 2). Studies of convergence pose questions about systemic disruptions or equilibrium, with the evaluative yardstick being that of normative press theory. While scholars such as Jenkins (2006) regard convergence as the steady democratization of cultural production, others such as Cammaerts (2003) argue that user production – which includes citizen journalism – pacifies audiences with a false sense of agency which links participation to consumption instead of to critical reflection.

Whether referring to the consolidation of jobs, roles, content, or companies that earlier had been discrete, convergence as an organizing concept sets aside questions concerning which side – professional journalism or citizen journalism – is carrying the day. Instead it focuses on emerging hybrids of professional–citizen organizations and practices. For example, some studies claim that citizen journalism has been fully absorbed by professional journalism, such as the case of the Community Connection initiative of New Jersey Online studied by Boczkowski (2004). Similarly, Compton and Benedetti (2010) find that professional journalism will maintain its integrity in the face of citizen journalism.

However, conclusions by other scholars emphasize emergent hybridities of production and consumption. Uricchio (2004) refers to the emergence of 'collaborative networks' as the organizational manifestation of convergence driven by technological change, and Bruns (2008) addresses the hybridization of roles through the concept of 'produsage' (an amalgam of 'production' and 'usage'). Deuze et al. (2007) conduct a multinational comparative study and find the emergence of new professional–citizen hybrids. Jouët (2009) finds a similar case of hybrid journalism in France.

Social Constructivism

Social constructivism departs from the prior perspectives in a number of ways. As inferred by the term 'constructivism,' it seeks to account for beliefs, roles, organizations, and systems as ongoing processes rather than established and stable things. Accordingly, studies of citizen journalism with this emphasis seek not to define or measure citizen journalism so much as to account for the processes that bring it into being and that maintain it.

A number of sources inform such work. In their emphasis on building workable public deliberation, studies of civic and public journalism (an earlier cousin of citizen journalism) made their cases via philosophical pragmatism (Coleman, 1997;

Bybee, 1999; Rosen, 1996; Merritt, 1998; Eksterowicz and Roberts, 2000). The relevance of pragmatism for newswork and communication can be traced most overtly in the USA by reference to the work of social philosopher John Dewey (1927; Carey, 1989; Rosenberry and St. John III, 2010; Steiner and Roberts, 2011). However, more recent social-constructivist studies of citizen journalism rely more fully on the constructivism and phenomenology of Berger and Luckmann (1967), which came into relevance in studies of the sociology of newswork by Tuchman (1978) and others (a good overview is Carey, 1989).

Social-constructivist studies of citizen journalism seek to determine whether and to what degree workplace norms and self-understandings of news not only differ between citizen journalism and professional journalism, but also how these understandings either support or dissuade the integration of one into the other (Tilly and Cokley, 2008). Studies rely upon mixes of analysis of news stories (as indications of the norms and beliefs used to report and write them, as well as of the situational resources and restrictions of time, money, access and so on) and analysis of direct interviews, with the presumption that the views expressed indicate relevant norms and practices at work.

One group of studies finds little in common between the two kinds of journalism. Carpenter (2008) finds that citizen journalists and professional journalists use different routines while reporting and writing. While professional journalists relied more on official sources, citizen journalists used opinion and unofficial sources, with the implication that citizen journalists' reports adhere less so to professional norms. In a study based on interviews with citizen journalists and professional journalists about how they went about reporting and writing their stories, Reich (2008) comes to similar conclusions. Citizen journalism differs from professional journalism due to different situations and resources available to each, thus reflecting badly on citizen journalists in relation to professional norms. Other studies investigate more thoroughly professional journalists' perceptions of political–economic pressures that structure their work (Vujnovic et al., 2010).

However, other studies conclude that there exists a good deal of commonality between citizen journalists and professional journalists. Robinson and Deshano (2011a) find that both groups share the liberal–pluralist professional ideology. Williams et al. (2011) find that the BBC accommodated citizen journalism within existing professional routines, thus posing little challenge to professional standards. Lewis et al. (2010) find differing opinions and practices among a sample of community newspaper editors. While some editors oppose citizen journalism in any form other than reader feedback and opinion in order to uphold professionalism (also the conclusion of Singer and Ashman, 2009), other editors approve it due to encouraging closer involvement of public and the newspaper as well as helping provide additional labor at little extra cost. Pantti and Bakker (2009) link the relevance of professional norms to medium. They argue that user pictures are deemed to be more credible and less reliant on skill and judgment than user reporting, and are thus integrated more easily by the profession.

Still others conclude that, far from having little or no impact on professional journalism, citizen journalism is having a dramatic impact. Thurman (2008) concludes that much of the debate among editors and journalists about citizen journalism is spurred by increasing pressures of profitability and their pressures on newsroom routines (see also Hermida and Thurman, 2008). Paulussen and Ugille (2008) find a similarly determinant role of newsroom structures, work routines, and professional beliefs. These, and not intentional resistance to citizen journalism by professional journalists, explain the resulting resistance to its adoption.

Critical Discourse Analysis

Critical discourse analysis introduces a different problem altogether. It posits inequalities as intentional and structurally necessary (not aberrations, distortions, or dysfunctions), which remain opaque to other perspectives (Williams, 1975). By conceiving discourse as a social activity (not as a noun labeling a collection of words or meanings), such studies also share a constructivist predilection with process instead of a positivist preoccupation with attributes and characteristics of fully formed elements. Finally, such studies tend to define citizen journalism and professional journalism relationally instead of as opposites or as mutually exclusive. As such, its assessment of citizen journalism is often posed in terms of struggle and uneven merger instead of the comparative strength of opposing positions or fully realized hybrids.

The extent to which the presumptions of other perspectives are left behind can be seen in a variety of work. Robinson (2009) describes very complex relationships between citizen journalism and professional journalism in a study of news stories written and published after the 2005 Hurricane Katrina disaster in the USA that flooded the city of New Orleans and others on the coast of the Gulf of Mexico. Instead of opposing citizen journalism to professional journalism, the study identifies a complex array of different characters, themes, perspectives, overall messages, and claims of authority by news items written to mark the one-year anniversary of the Katrina disaster. Such findings suggest the inadequacy of claims about which side (professional or citizen) has the upper hand or claims of a synthesis of the two. Contrary to such claims, Robinson suggests the existence of "a patchwork of journalism and citizen work that represents the acting out of the tensions from these role shifts" (p. 810).

Moyo (2011) frames the use of blogging by nonprofessionals within the 2008 election in Zimbabwe within changes not only in technology, but also within changes in the very nature of journalism increasingly characterized by "increased participation, multimodality and multidirectional flows" (p. 756). The emphasis on 'multi' rather than binary (either citizen or professional) broadens consideration of the complexity and interrelatedness of citizen journalism with many other spheres and practices. Citizen journalism is not a clear opposite, nor clearly different from professional journalism, but a provisional activity generated through complex relationships as well as the particular context in which it takes place. While Moyo sees on one hand the emergence of "a hermeneutic, constructivist, and postmodern approach to journalism" very different from the professional mode, this citizen journalism still shares with the professional mode "the idealism of the neoliberal discourse that empowers [both of] them as the watchdog of the state, but this simultaneously

deprives them of the capacity to conceptualize the alternative in truly radical ways" (p. 758).

By contrast to organizations and efforts institutionally outside professional journalism organizations, Kperogi (2011) assesses citizen journalism as enabled by CNN News and its iReport feature. The study traces the "enthusiastic valorization of citizen media" to "the idea that they exist apart from and, in fact, in opposition to, the mainstream [professional] media" (p. 315). By contrast, this study suggests a much more nuanced assessment. On one hand, citizen-submitted items are successfully incorporated not through intentional editorial decision, but discursively, by "construct[ing] the 'normality' of prevailing conventions of news judgment," which people who submit items accept, if only provisionally (p. 320). On the other hand, and despite this largely successful management, other iReport items "openly criticize CNN's coverage of certain news events" and that "give radical perspectives on the US economic recession," suggesting the permeability of means of control (p. 322). Citing theorists such as Gramsci and Zizek, Kperogi (2011) emphasizes the ongoing processual nature of this situation by noting that "there is always contestation and that the volitional consent of the broad segment of subordinate classes has to be perpetually won and re-won" (p. 326).

What thus becomes important is not to assert which side has won, but what the terms of the struggle between the sides is and, furthermore, how this struggle constitutes the very sides in contention.

Historical Studies

Finally, historical studies share many aspects with critical discursive studies. They ask similar questions about the ongoing, processual constitution of citizen journalism (instead of about its imputed and ostensibly fixed and intrinsic nature). However, their distinctive contribution is in expanding the time frame and range of interrelationships that enable this constitution.

A number of historical studies of citizen journalism emphasize its historicity instead of universality by noting its deep relationship with the industrialized West. In doing so, the imputed universal effects of digital technologies is recognized as determined more deeply by social intentions within particular social conditions (Williams, 1975). To underscore this point, Pidduck (2010) argues in favor of "conjunctural studies that trace how social and political agents deploy different media within particular historical, cultural, and political frameworks" (p. 475). It addresses the limits of citizen journalism within the context of the autocratic regime of Burma, "notorious for extensive censorship and for the harassment and imprisonment of journalists" of whatever kind (p. 475). Compared with the view that "that the decentralized, participatory architecture of 'new media' is intrinsically democratic," Pidduck argues that contextual awareness (such as of low internet penetration in Burma) moderates such claims when extended globally.

Hamdy (2009) addresses the effectiveness of citizen journalism as operating outside professional journalism in Egypt prior to the changes wrought by the Arab Spring. The study emphasizes presciently the multitudinous challenges posed by Arab bloggers to the authority of the state, the laws and regulations active at the time to try to limit these challenges, as well as citizen journalists' tactics to avoid such limitations. Bosch (2010) discusses the South African experience, and concludes in that comparatively industrialized country that all can be seen as part of an 'online public sphere,' which is an allusion to Habermas's influential formulation (Habermas, 1989; Calhoun, 1992). Finally, Hamilton (2008) seeks a broader historicization of citizen journalism by locating its dynamic and emergence within the centuries-long formation of capitalist media organization and its steady challenge brought on by accelerating pressures of competition in the contexts of the emergence of modernism and technological change.

Conclusion

Despite the immense variety of perspectives, most studies of citizen journalism are structured by a framework of inquiry defined by normative liberal–pluralist press theory and the assumption of a categorical distinction between professionals and nonprofessionals. That constructivist, critical, and historical perspectives for addressing citizen journalism turn such a framework itself into a problem deserving of theoretical attention continues to move studies of citizen journalism into new and exciting directions.

See also: Journalism and Journalists; Journalism; Mass Media, Representations in; New Media and the Digital Divide; New Media, News Production and Consumption; Printing as a Medium; Television: General.

Bibliography

Allan, S., Thorsen, E., 2009. Citizen Journalism: Global Perspectives. Peter Lang, New York.

Alter, J., March 1986. Taking chances. Newsweek 24, 59.

Altschull, J.H., 1990. From Milton to McLuhan: The Ideas behind American Journalism. Longman, New York.

Atton, C., Hamilton, J.F., 2008. Alternative Journalism. Sage, London.

Bak, J.S., Reynolds, B. (Eds.), 2011. Literary Journalism across the Globe: Journalistic Traditions and Transnational Influences. University of Massachusetts Press, Amherst.

Banks, M., 2007. The Politics of Cultural Work. Palgrave Macmillan, New York.

Berger, P.L., Luckmann, T., 1967. The Social Construction of Reality; a Treatise in the Sociology of Knowledge. Doubleday, New York.

Black, J. (Ed.), 1997. Mixed News: The Public/Civic/Communitarian Journalism Debate. Lawrence Erlbaum, Mahwah.

Boczkowski, P.J., 2004. Digitizing the News: Innovation in Online Newspapers. MIT Press, Cambridge.

Bosch, T., 2010. Digital journalism and online public spheres in South Africa. Communicatio: South African Journal for Communication Theory and Research 36 (2), 265–275.

Bruns, A., 2008. Blogs, Wikipedia, Second Life, and Beyond: From Production to Produsage. Peter Lang, New York.

Bybee, C., 1999. Can democracy survive in the post-factual age? A return to the Lippmann-Dewey debate about the politics of news. Journalism and Communication Monographs 1 (1), 27–66.

Calhoun, C. (Ed.), 1992. Habermas and the Public Sphere. MIT Press, Cambridge.

Cammaerts, B., 2003. Beyond the Digital Divide: Reducing Exclusion, Fostering Inclusion. VUB Brussels University Press, Brussels.

Carey, J.W., 1989. Communication as Culture; Essays on Media and Society. Unwin Hyman, Boston.

Carpenter, S., Autumn 2008. How online citizen journalism publications and online newspapers utilize the objectivity standard and rely on external sources. Journalism & Mass Communication Quarterly 85 (3), 531–548.

Carpenter, S., November 2010. A study of content diversity in online citizen journalism and online newspaper articles. New Media and Society 12 (7), 1064–1084.
Castells, M., 1996. The Rise of the Network Society. Blackwell, Cambridge.
Chartered Institute of Journalists, 2012. Welcome. Online at: http://cioj.co.uk/welcome-to-the-website-of-the-cioj/ (accessed 16.10.12.).
Coleman, R., April 1997. The intellectual antecedents of public journalism. Journal of Communication Inquiry 21 (1), 60–76.
Compton, J.R., Benedetti, P., August 2010. Labour, new media and the institutional restructuring of journalism. Journalism Studies 11 (4), 487–499.
Deuze, M., June 2007. Convergence culture in the creative industries. International Journal of Cultural Studies 10 (2), 243–263.
Deuze, M., Bruns, A., Neuberger, C., September 2007. Preparing for an age of participatory news. Journalism Practice 1 (3), 322–338.
Dewey, J., 1927. The Public and Its Problems. H. Holt and Company, New York.
Dickey, C., 18 November 1977. Cocktails for $2,000. The Washington Post C1.
Domingo, D., Quandt, T., Heinonen, A., Paulussen, S., Singer, J.B., Vujnovic, M., October 2008. Participatory journalism practices in the media and beyond. Journalism Practice 2 (3), 326–342.
Eksterowicz, A.J., Roberts, R.N., 2000. Public Journalism and Political Knowledge. Rowman & Littlefield, Lanham.
Gillmor, D., 2004. We the Media: Grassroots Journalism by the People, for the People. O'Reilly Media, Sebastopol.
Glasner, J., November 2004. Wikipedia creators move into news. Wired 29. Online at: http://www.wired.com/culture/lifestyle/news/2004/11/65819 (accessed 09.10.12.).
Goldstein, T. (Ed.), 2007. Killing the Messenger: 100 Years of Press Criticism. Columbia University Press, New York.
Greer, S., 1969. The Logic of Social Inquiry. Aldine, Chicago.
Griffen-Foley, B., July 2004. From tit-bits to big brother: a century of audience participation in the media. Media, Culture and Society 26 (4), 533–548.
Habermas, J., 1989. The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society (Thomas B. and Frederick L. Trans.). MIT Press, Cambridge.
Halleck, D.D., October 1993. Deep dish TV: community video from geostationary orbit. Leonardo 26 (5), 415–420.
Hamdy, N., June 2009. Arab citizen journalism in action: challenging mainstream media, authorities and media laws. Westminster Papers in Communication & Culture 6 (1), 92–112.
Hamilton, J.F., 2008. Democratic Communications; Formations, Projects, Possibilities. Lexington Books, Lanham.
Heflin, K., 2010 (Unpublished PhD dissertation). In: Twitter and Wikipedia We Trust?: Online Information and the Crisis of Credibility. University of Georgia.
Heinrich, A., 2011. Network Journalism Journalistic Practice in Interactive Spheres. Routledge, New York.
Hermida, A., Thurman, N., October 2008. A clash of cultures: the integration of user-generated content within professional journalistic frameworks at British newspaper websites. Journalism Practice 2 (3), 343–356.
Hesmondhalgh, D., 2002. The Cultural Industries. Sage, London.
Hollowell, J., 1977. Fact and Fiction: The New Journalism and the Nonfiction Novel. University of North Carolina Press, Chapel Hill.
Howley, K., 2005. Community Media; People, Places and Communication Technologies. Cambridge University Press, Cambridge.
Jenkins, H., 2006. Convergence Culture: Where Old and New Media Collide. NYU Press, New York.
Jesdanun, A., October 6, 2005. Traditional Media Experiment with Citizens as News Producers. Associated Press, Lexis-Nexis Academic.
Jouët, J., March 2009. The Internet as a new civic form; the hybridisation of popular and civic web uses in France. Javnost – The Public 16 (1), 59–72.
Kaufhold, K., Valenzuela, S., De Zúñiga, H.G., 2010. Citizen journalism and democracy: how user-generated news use relates to political knowledge and participation. Journalism & Mass Communication Quarterly 87 (3/4), 515–529 (Autumn/Winter).
Kidd, D., 2003. Indymedia.org: a new communication commons. In: McCaughey, M., Ayers, M.D. (Eds.), Cyberactivism: Online Activism in Theory and Practice. Routledge, New York, pp. 47–69.
Kim, E.-G., Hamilton, J.F., July 2006. Capitulation to capital? OhmyNews as alternative media. Media, Culture and Society 28 (4), 541–560.
Kperogi, F.A., March 2011. Cooperation with the corporation? CNN and the hegemonic cooptation of citizen journalism through iReport. New Media and Society 13 (2), 314–329.
Lacy, S., Watson, B., Rife, D. (Fall), 2011. Study examines relationship among mainstream, other media. Newspaper Research Journal 32 (4), 53–67.
Lasswell, H., 1948. The structure and function of communication in society. In: Bryson, L. (Ed.), The Communication of Ideas. Harper and Row, New York, pp. 37–51.
Lewis, S.C., Kaufhold, K., Lasorsa, D.L., April 2010. Thinking about citizen journalism. Journalism Practice 4 (2), 163–179.
McChesney, R.W., Scott, B. (Eds.), 2004. Our Unfree Press: 100 Years of Radical Media Criticism. New Press, New York.
McChesney, R.W., Pickard, V., 2011. Will the Last Reporter Please Turn Out the Lights: The Collapse of Journalism and What Can Be Done to Fix It. New Press, New York.
Merritt, D., 1998. Public Journalism and Public Life: Why Telling the News Is Not Enough. Erlbaum, Hillsdale, NJ.
Moyo, L., August 2011. Blogging down a dictatorship: human rights, citizen journalists and the right to communication in Zimbabwe. Journalism 12 (6), 745–760.
Nah, S., Chung, D., Spring 2009. Rating citizen journalists versus pros: editors' views. Newspaper Research Journal 30 (2), 71–83.
Nip, J.Y.M., April 2006. Exploring the second phase of public journalism. Journalism Studies 7 (2), 212–236.
Oxfeld, J., March 2005. Letting the Blogs Out. Editor & Publisher, 39–42.
Pantti, M., Bakker, P., September 2009. Misfortunes, memories and sunsets: non-professional images in Dutch news media. International Journal of Cultural Studies 12 (5), 471–489.
Paterno, S., June/July 2005. The sad saga of Gary Webb. American Journalism Review. Online at: http://www.ajr.org/article.asp?id=3874 (accessed 09.10.12.).
Paulussen, S., Ugille, P., June 2008. User generated content in the newsroom: professional and organisational constraints on participatory journalism. Westminster Papers in Communication & Culture 5 (2), 24–41.
Pidduck, J., 2010. Citizen journalism in Burma and the legacy of Graham Spry. Canadian Journal of Communication 35 (3), 473–485.
Plimpton, G., 1966. Paper Lion. Harper and Row, New York.
Reich, Z., October 2008. How citizens create news stories. Journalism Studies 9 (5), 739–758.
Roberts, L., September 2006. Newspapers not on the extinction list just yet. Printweek 7, 21.
Robinson, S., August 2009. 'If you had been with us': mainstream press and citizen journalists jockey for authority over the collective memory of Hurricane Katrina. New Media and Society 11 (5), 795–814.
Robinson, S., Deshano, C., October 2011a. Citizen journalists and their third places. Journalism Studies 12 (5), 642–657.
Robinson, S., DeShano, C., November 2011b. 'Anyone can know': citizen journalism and the interpretive community of the mainstream press. Journalism 12 (8), 963–982.
Romano, A., August 2005. Why everyone is a reporter; citizen journalists go mainstream. Broadcasting and Cable 22, 14.
Rosen, J., 1996. Getting the Connections Right: Public Journalism and the Troubles in the Press. Twentieth Century Press, New York.
Rosenberry, J., St John III, B. (Eds.), 2010. Public Journalism 2.0: The Promise and Reality of a Citizen-Engaged Press. Routledge, New York.
Schudson, M., 1978. Discovering the News; a Social History of American Newspapers. Basic Books, New York.
Siebert, F., Peterson, T., Schramm, W., 1956. Four Theories of the Press: The Authoritarian, Libertarian, Social Responsibility, and Soviet Communist Concepts of What the Press Should Be and Do. University of Illinois Press, Urbana.
Singer, J.B., Ashman, I., January 2009. 'Comment is free, but facts are sacred': user-generated content and ethical constructs at the Guardian. Journal of Mass Media Ethics 24 (1), 3–21.
Steiner, L., Roberts, J., 2011. Philosophical linkages between public journalism and civic journalism. In: Papathanassopoulos, S. (Ed.), Media Perspectives for the 21st Century. Routledge, London, pp. 191–211.
Theobald, J., 2004. Radical media critics: the four generations. In: The Media and the Making of History. Ashgate, Aldershot, pp. 19–53.
Thurman, N., February 2008. Forums for citizen journalists? Adoption of user generated content initiatives by online news media. New Media and Society 10 (1), 139–157.
Tilley, E., Cokley, J., April 2008. Deconstructing the discourse of citizen journalism: who says what and why it matters. Pacific Journalism Review 14 (1), 94–114.
Tuchman, G., 1978. Making News. Free Press, New York.
Uricchio, W., March 2004. Beyond the great divide: collaborative networks and the challenge to dominant conceptions of creative industries. International Journal of Cultural Studies 7 (1), 79–90.
Vujnovic, M., Singer, J.B., Paulussen, S., Heinonen, A., Reich, Z., Quandt, T., Hermida, A., Domingo, D., August 2010. Exploring the political-economic factors of participatory journalism. Journalism Practice 4 (3), 285–296.

Wunsch-Vincent, S., Vickery, G., 2007. Participative Web: User-Created Content. Committee for Information, Computer and Communications Policy; Directorate for Science, Technology and Industry; Organization for Economic Co-operation and Development. Report No. DSTI/ICCP/IE(2006)7/FINAL.

Webb, G., 1996. Dark Alliance: The Story Behind the Crack Explosion; a Mercury News Special Report. San Jose Mercury News, 18 August 1996, p. A1; 19 August 1996, p. A1; 20 August 1996, p. 11A; 20 August 1996, p. 1A.

Webb, G., 1998. Dark Alliance; the CIA, the Contras, and the Crack Cocaine Explosion. Seven Stories Press, New York.

Webb, G., 2002. The mighty Wurlitzer plays on. In: Borjesson, K. (Ed.), Into the Buzzsaw: Leading Journalists Expose the Myth of a Free Press. Prometheus Books, Amherst, pp. 141–156.

Weingarten, M., 2006. The Gang That Wouldn't Write Straight: Wolfe, Thompson, Didion, and the New Journalism Revolution. Crown, New York.

Williams, A., Wardle, C., Wahl-Jorgensen, K., February 2011. Have they got news for us? Journalism Practice 5 (1), 85–99.

Williams, R., 1975. Television: Technology and Cultural Form. Schocken, New York.

Xin, X., August 2010. The impact of 'citizen journalism' on Chinese media and society. Journalism Practice 4 (3), 333–344.

Yankelovich, D., 1991. Coming to Public Judgment: Making Democracy Work in a Complex World. Syracuse University Press, New York.

Citizen Mapping

Wen Lin, Newcastle University, Newcastle upon Tyne, Tyne and Wear, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since the mid-2000s, there have been significant shifts in geographic information production and dissemination, marked by the emergence of user-generated geographic information. This form of mapping and spatial data production, facilitated by an array of Web 2.0 technologies and location-aware mobile devices, has allowed ordinary citizens to carry out mapping practices without formal training in Geographic Information Systems or cartography. This article aims to provide an overview of the research on these mapping practices, known as citizen mapping; the focus is on three broad strands of work: spatial data in citizen mapping, citizen mapping as social process, and related methodological challenges and developments.

Introduction

Traditionally, map and spatial data production has been mainly carried out by government agencies and professionals given the technological complexities and associated high costs. Since the mid-2000s, however, the rise of Web 2.0 technologies (such as blogs, wikis, and social networking sites), combined with the increasing availability of mobile devices with global positioning systems (GPSs) and high-resolution remotely sensed imagery, has enabled an unprecedented number of individuals and social groups to generate and share maps and spatial data online (Elwood, 2009). In particular, citizens increasingly play an active role in producing and sharing spatial data at an unprecedented scale (Haklay et al., 2008). Citizen mapping refers to the mapping practices that are carried out by lay persons, enabled by these technological transformations. While forms of such mapping practices are highly diverse, they are characterized by user-generated geographic information production such as contributing to mapping databases through online mapping platforms and generating map mashups that combine multiple sources of spatial and nonspatial data on one shared online interface. A growing range of practices have been undertaken through citizen mapping, such as monitoring environmental change, mapping crises, responding to emergencies, addressing neighborhood change and needs, and mobilizing activist efforts (Elwood, 2009; Miller, 2006; Crutcher and Zook, 2009).

This notion of citizen mapping is highly related to other notions emerged to address this explosive growth of user-generated geographic information such as 'volunteered geographic information' (VGI) (Goodchild, 2007), 'the geoweb' (Scharl and Tochtermann, 2007), 'maps 2.0' (Crampton, 2009), and 'neogeography' (Turner, 2006). All emphasize nonexperts producing and sharing geographic information, facilitated by various online mapping platforms such as OpenStreetMap, Wikimapia, Google Maps, etc. As such, the line between traditional map makers and map readers is increasingly blurred (Budhathoki et al., 2008). This phenomenon can also be seen as part of the broader context of the 'exaflood of big data' (Sui et al., 2012), which has brought profound influences on how geographic information and data are produced and shared (ibid). The following sections will take up major debates around the emergence of citizen mapping and related practices.

Conceptual Debates

A rapidly growing body of work investigates the emergence and impacts of citizen mapping via three major, closely related, themes: (1) the forms and characteristics of the data created through citizen mapping; (2) the social and political implications of these mapping practices; and (3) the methodological challenges of working with and researching these new forms of geographic information and practices.

Spatial Data and Representation in Citizen Mapping

Early discussions focused on the digital data generated through citizen mapping in relation to conventional ways of spatial data production and dissemination, including how data are created and shared and for what purposes as well as data content and characteristics (Elwood, 2009). These discussions have been strongly informed by research in spatial data infrastructure (SDI) and data handling in geographic information science (GIS) and related fields. Research has recognized the significant shifts brought by citizen mapping practices on the scale and scope of geographic information production and the implications for data management, representation, and credibility.

Studies have noted that these technological transformations have led not only to unprecedented volumes of geographic information production and dissemination online by an increasing range of mappers, but also to changes in what is being mapped (e.g., Haklay et al., 2008; Elwood, 2009). These highly open and interactive mapping platforms allow for a much more diverse group of participants, which in turn results in a much wider range of topics in these mapping practices and representations. In particular, citizens can incorporate more localized and everyday information into the mapping platforms such as through mapping mashups. As such, citizen mapping practices generate tremendously vast and heterogeneous data (Elwood, 2009). One of the characteristics in these everyday mapping practices is the tendency to rely more on everyday language, which suggests a more 'platial' context than traditional GIS (Elwood et al., 2012: p. 363). For example, place names and points of interests play a significant role in citizen mapping (ibid). In contrast, traditional GIS tends to focus on planimetric features of mapping and conducting

topological overlay for calculation, visualization, and modeling of spatial data (e.g., Foresman, 1998).

Related to these characteristics of citizen mapping data are questions about how existing SDIs, which usually follow a top-down model, can deal with the distributed and user-generated data derived from citizen mapping (cf Elwood et al., 2012). Researchers have suggested incorporating citizen mapping data as part of a 'patchwork' approach to contribute to existing SDIs (Goodchild, 2007). For example, an initiative of citizen-based measurement of small particles in the air through daily activities in Milan, Italy, shows that the levels of these small particles collected by citizens are 20 times higher than those collected by the monitoring stations (Craglia, 2007). This approach, along with other efforts to develop systems to handle and integrate Web-based geographic information, is informed by research on the long-standing issue of interoperability in GIS (e.g., Craglia, 2007). The idea of building SDIs from the bottom up via grassroots and community organizations has also received attention (e.g., Elwood, 2008). Moreover, issues of the digital divide are important to the development and usage of SDIs and user-generated spatial data, defying a one-size-fits-all approach or a linear conceptualization of data/infrastructure availability and usage (e.g., Sui et al., 2012).

Issues of data quality and credibility have drawn notable attention (Elwood et al., 2012). The quality of citizen mapping practices may vary significantly, while some believe that many of these practices can potentially reach a high level of accuracy ensured by collective intelligence. For example, Haklay (2010) compares citizen mapping data produced through OpenStreetMap with the data produced by Ordnance Survey, the national mapping agency in the United Kingdom. His study shows that the OpenStreetMap data on London and England are quite accurate in this comparison. Yet, despite some successes, producing high quality data through citizen mapping remains a significant challenge (Elwood et al., 2012). At the same time, research has called for approaches that reconceptualize notions of information validity and credibility regarding user-generated geographic information (Flanagin and Metzger, 2008; Poore and Wolf, 2012). For example, Flanagin and Metzger (2008) note that credibility can be viewed as an objective property or a subjective perception, with the former emphasizing accuracy and the latter believability. They suggest that viewing credibility as a perceptual variable helps to address the core issues of publicly VGI that tend to be "more about which information, opinion, or perspective people believe" (p. 141), which is useful for those who use such data for social and political purposes. Meanwhile, credibility as accuracy can be applied to these mapping practices when the results and knowledge produced in such practices can be evaluated through widely shared standards that are less dependent on local contexts. For example, the credibility of locations annotated by WikiMapia users can be assessed using established standards of locational accuracy. Poore and Wolf (2012) use the notion of 'metadata squared' to illustrate the bottom-up approach facilitated by the Web 2.0 technologies, which results in more flexible and accessible metadata generation including a proliferation of ancillary data that describe the map and include discussions and debates around the map.

It is clear that the distributed practices of citizen mapping have called into question "conceptions of authority as centralized, impenetrable, and singularly accurate", which may lead to "a shift from a model of single authority based on information scarcity and hierarchy to a model of multiple distributed authorities based on information abundance and networks of individuals" (Flanagin and Metzger, 2008: p. 143). Simultaneous with efforts to incorporate existing ways of evaluating the credibility of user-generated geographic information is the recognition that data characterized by experiential and quotidian representations gain credibility through practices of transparency and peer-to-peer review such as peer verification and witnessing (Elwood and Leszczynski, 2013). These forms of representing experiential and situated knowledge through citizen mapping have drawn significant attention from scholars, for which these mapping practices are examined as social processes.

Citizen Mapping as Social Processes

A burgeoning body of work investigates citizen mapping as social processes and their social and political implications (cf Elwood et al., 2012). This research strand has been greatly informed by research in critical GIS and critical cartography (Crampton, 2009; Elwood, 2011; Pickles, 2004). Critical GIS and critical cartography investigate the ways in which spatial data and knowledge are generated, represented, and shared by various institutions, organizations, and individuals in different contexts (e.g., Sheppard, 2005). Critical GIS research was stimulated by early critiques of GIS in the early 1990s, in which a number of human geographers called for attention to the social impacts of GIS regarding asymmetries of power and knowledge production, privilege and marginalization, state surveillance, population control, marketing, military operations, and so on (cf Pickles, 1995). Later responses argue that GIS is not inherently positivist (Schuurman, 2002). Subsequent work has broadened access to GIS technologies and data for socially marginalized groups and grassroots community organizations, resulting in rich accounts and discussions of participatory mapping and public participation GIS (PPGIS) (Sieber, 2006).

Together, these studies have provided important insights into understanding how citizen mapping practices may change spatial data production and associated knowledge construction (Elwood et al., 2012; Crampton, 2009). This work has led to discussing the potential of empowerment and facilitating public participation through citizen mapping. Some studies suggest that the greater user friendliness and reduced spatial and temporal constraints of citizen mapping can facilitate greater civic engagement and participation in a number of contexts (e.g., Miller, 2006). Strongly influenced by the PPGIS literature that seeks to tackle issues of uneven GIS technologies usage and spatial knowledge production (cf Sieber, 2006), researchers have examined how these mapping practices might empower citizens and citizen organizations. For example, in the aftermath of Hurricane Katrina in New Orleans, USA, in 2005, citizens used the platform of Google Maps to post information about rescue needs, creating what Miller calls "a map of the disaster" by local people or those who were able to provide local information in this situation

(Miller, 2006: p. 195). Recent years have seen growing efforts to engage with crisis mapping, drawing participants across the world in cases such as responses to the Haiti Earthquake in 2010 where governmental responses and services might be lacking. However, effects of such activism are mixed and varied (Elwood, 2011), with possible unintentional outcomes such as facilitating a form of voyeurism (Parks, 2009).

Continuous efforts have been made to further investigate the sociopolitical implications of these mapping practices and what forms of politics might be emerging or reconfigured in these processes. For example, through investigating five online initiatives that engage with the geoweb for activism and civic engagement, Elwood and Leszczynski (2013) suggest that these new practices entail constructions of visual experiences that are highly individualized, interactive, and exploratory. Through constructing such experiences aiming to affect social and political change, these organizations leverage these mapping platforms to gain legitimacy of knowledge claims, which in turn constitute new forms of knowledge politics.

A related aspect in the area of civic engagement draws upon notions of citizen science and open source software movements. The 'patchwork' approach regarding SDI developments with contributions from citizen mapping noted earlier can be considered in this light. A prime example is the development of OpenStreetMap. This is a free and open source online mapping platform which aims to provide a digital map of the world, a somewhat Wikipedia version of mapping (**Figure 1**; Haklay et al., 2008; Poore and Wolf, 2012). Registered users may upload their own GPS tracks and use out-of-copyright maps and Bing aerial imagery for references to create maps on OpenStreetMap (Haklay, 2010). At the time of writing, there are more than 1 million registered OpenStreetMap users and over 3 billion GPS points uploaded (http://www.openstreetmap.org/stats/data_stats.html). It is expected that these numbers will continue to grow. In general, it is considered that these mapping practices do not require formal training in traditional GIS or cartography. Explanations and discussions of associated steps and procedures are available through respective wiki pages, forums, and email lists. Such discussions help to foster community building to engage with mappers (Poore and Wolf, 2012).

Related to the issue of empowerment and civic engagement are motivations, identities and subjectivities, which shape and are shaped by these mapping practices (Budhathoki et al., 2008; Dodge and Kitchin, 2007; Elwood, 2011). A key issue has been reconceptualizing the notion of 'the user' (e.g., Budhathoki et al., 2008). Attention has also been given to the intersection of subjectivities and highly individualized and interactive electronic communication technologies (e.g., Kingsbury and Jones, 2009). In these discussions, the mapping subject is a hybrid one with constant self-identity construction, in contrast to the traditional view of a rational and coherent mapping subject. These dynamic and fragmented subjectivities, however, can be actively employed to represent alternative and complementary forms of resistance and activism (Cobarrubias and Pickles, 2009).

Meanwhile, researchers have called for more studies investigating how these spatial data and mapping practices may perpetuate existing inequalities, produce new forms of exclusion and surveillance, and raise concerns about privacy encroachment (e.g., Crutcher and Zook, 2009; Elwood and Leszczynski, 2011). Crutcher and Zook (2009), for example, show how racial relations affect the way people use these online mapping services. Elwood and Leszczynski (2011) analyze how the notion of privacy is negotiated and contested in the practices of user-generated geographic information production and dissemination including geotagged images in the age of Web 2.0. Moreover, Leszczynski (2012) examines citizen mapping practices through the lens of political economy. She argues that the significant shift to more user-generated forms of spatial data production indicates new regimes implicated in neoliberal political economy reflected in three dimensions: free labor, private ownership, and unaccountability. In these new regimes, the state is moving toward a role of "one of many producers and facilitator or institutional body of oversight" (Leszczynski, 2012: p. 78). Nonstate agencies, at the same time, are playing a more active role in geographic information provision, dissemination, and utilization.

Citizen mapping can also be situated in the context of pervasive computing (e.g., Dodge and Kitchin, 2007). For example, Dodge and Kitchin (2007) examine the emergence of life-logs as a form of such pervasive computing and their impacts. They note that while these technological transformations have the potential to provide rich autobiographical narratives for individual empowerment and enrichment, they also raise broader social and political questions such as concerns of access to such data by third parties which may result in data abuse and unethical usage.

Methodologies for Researching Citizen Mapping

As discussed earlier, citizen mapping practices have produced tremendous volumes of highly heterogeneous data, which brings new methodological challenges to researchers both in terms of working with these vast and diverse data and investigating these new forms of geographic information. Elwood (2011) provides an insightful overview of methods in research of the geoweb, a term used to describe the vast spatially referenced information produced and shared online, which is highly relevant to citizen mapping practices. Elwood (2011) suggests two important trajectories: one focusing on research *with* geoweb-based information and the other more on research *about* the geoweb. Specifically, regarding research with geoweb-based information, scientific information visualization approaches have been explored to investigate broader patterns in these complex and large datasets (e.g., Jiang, 2012). For example, Jiang (2012) uses geospatial analysis approaches to investigate and visualize mobility patterns detected from user-generated geographic information. In particular, he presents three examples of employing computational geography using OpenStreetMap data, contributed by numerous citizen mappers, for such analyses underpinned by the concepts of topology and scaling. Topology is defined as the study of spatial relationships that remain invariant despite geometric distortions. Scaling refers to the power-law distribution of geographic spaces where there are more small things than large ones, for example, in the case of cities. In one of the examples, the whole Swedish street network database was extracted from

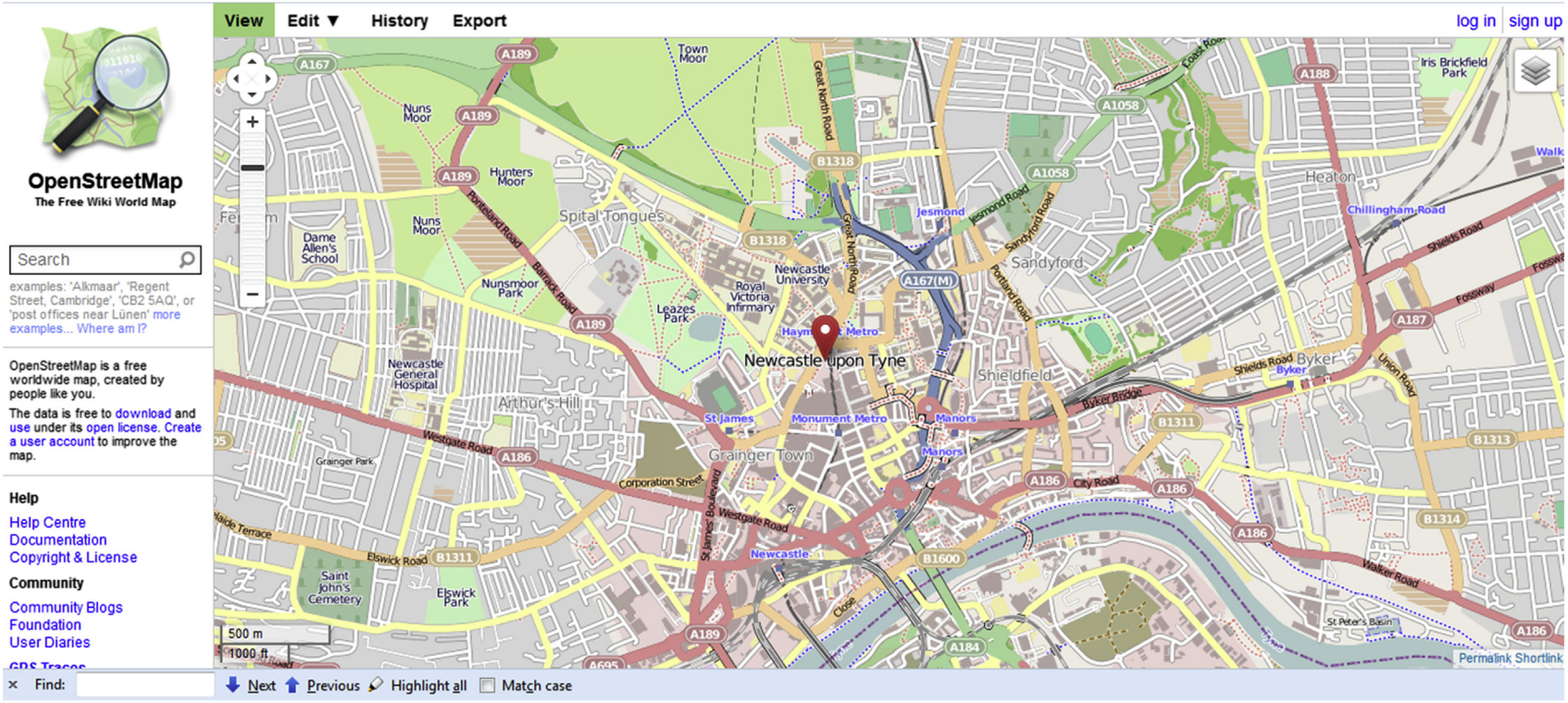

Figure 1 Screenshot of OpenStreetMap. © OpenStreetMap contributors.

OpenStreetMap and used to show hierarchical levels of the street network. Such analyses can have important implications for understanding traffic flow (Jiang, 2012).

Critical visual methods have been employed to examine the everyday mapping practices of the geoweb (e.g., Graham and Zook, 2013). This strand of research emphasizes an epistemological hybridity of visualization (Elwood, 2011). Such an approach may involve visualizing the dynamics of citizen mapping practices as well as understanding these mapping practices as visual practices through engaging with other qualitative methods. For example, Graham and Zook (2013) employ a method of mapping geocoded content on Google Maps to examine the geographies of digital annotations and how these patterns intertwine with constructions of place. Through visualization, they show that the density of digital geocoded content is highly uneven, in which Western Europe and North America have the highest amounts of geographically annotated content per capita, and much of the rest of the world has a small amount of augmented information. While these uneven patterns largely mirror past patterns of the Web in the 1990s, new geographies are also emerging such as the relatively higher densities of some non-Western countries compared to their status in the 1990s. At another level the authors further investigate the spatial footprints of different languages in the geoweb in five particular regions (Canada, Spain, Belgium, France, and Israel and the Palestinian Territories) through visualizing aggregations of certain keywords. The results provide a vivid illustration of the "geolinguistic contours of a place and its representations in the geoweb" (p. 95). Other methods such as discourse analysis, participant observations, interviews, and surveys have also played a role in understanding citizen mapping as visual practices and their associated implications (e.g., Kingsbury and Jones, 2009; Elwood and Leszczynski, 2013).

Summary

In summary, a burgeoning body of work in GIS and geography seeks to grapple with a range of issues and questions derived from the emergence of citizen mapping and other user-generated geographic information production. This body of work has outlined a number of technological, socioeconomic, and subjectivities transformations underpinning the emergence of citizen mapping (e.g., Goodchild, 2007; Sui, 2008; Haklay et al., 2008; Elwood, 2011). In particular, research has sought to analyze the content and characteristics of these data produced in these processes. Studies have also attempted to investigate citizen mapping as social processes and the socio-political implications of such processes. These inquiries are carried out through several broad perspectives including SDI construction, empowerment and disempowerment of different social groups and individuals, and meanings of everyday practices (e.g., Elwood, 2011; Kingsbury and Jones, 2009). These mapping practices and the data they generate present new challenges for researchers with respect to both researching with such data and researching on these data and associated processes. Among many approaches, geovisualizaiton and critical visual approaches have been utilized to understand and interpret these mapping practices. It is important to note that citizen mapping continues to evolve and the volume of data continues to grow on a daily basis. Many key questions remain to be answered including how to better understand ways of incorporating, interpreting, and engaging with these mapping practices in an increasingly shrinking, but also highly divided world in view of rapidly evolving communication and spatial technologies (Sui et al., 2012).

See also: Cartographic Visualization; Cartography; Geographic Information Systems; Global Positioning System; Spatial Analysis in Geography; Spatial Data Infrastructures; Spatial Data; Thematic Maps in Geography.

Bibliography

Budhathoki, N.R., Bruce, B., Nedovic-Budic, Z., 2008. Reconceptualizing the role of the user of spatial data infrastructure. GeoJournal 72, 149–160.

Cobarrubias, S., Pickles, J., 2009. Spacing movements: mapping practice, global justice and social activism. In: Warf, B., Arias, S. (Eds.), The Spatial Turn: Interdisciplinary Perspectives. Routledge, New York, pp. 36–58.

Craglia, M., 2007. Volunteered Geographic Information and Spatial Data Infrastructures: When Do Parallel Lines Converge? http://www.ncgia.ucsb.edu/projects/vgi/participants.html (accessed 24.02.13.).

Crampton, J., 2009. Cartography: maps 2.0. Progress in Human Geography 33, 91–100.

Crutcher, M., Zook, M., 2009. Placemarks and waterlines: racialized cyberscapes in; post-Katrina Google Earth. Geoforum 40, 523–534.

Dodge, M., Kitchin, R., 2007. Outlines of a world coming into existence: pervasive computing and the ethics of forgetting. Environment and Planning B: Planning and Design 34, 431–445.

Elwood, S., 2008. Grassroots groups as stakeholders in spatial data infrastructures: challenges and opportunities for local data development and sharing. International Journal of Geographical Information Science 22 (1), 71–90.

Elwood, S., 2009. Geographic information science: new geovisualization technologies – emerging questions and linkages with GIScience research. Progress in Human Geography 33 (2), 256–263.

Elwood, S., 2011. Geographic information science: visualization, visual methods, and the geoweb. Progress in Human Geography 35 (3), 401–408.

Elwood, S., Goodchild, M., Sui, D., 2012. The prospects VGI research and the emerging fourth paradigm. In: Sui, D., Elwood, S., Goodchild, M. (Eds.), Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice. Springer, London, pp. 361–375.

Elwood, S., Leszczynski, A., 2011. Privacy reconsidered: new representations, data practices, and the geoweb. Geoforum 42 (1), 6–15.

Elwood, S., Leszczynski, A., 2013. New Spatial Media, New Knowledge Politics. Transactions of the Institute of British Geographers 38 (4), 544–559. http://dx.doi.org/10.1111/j.1475-5661.2012.00543.x.

Flanagin, A., Metzger, M., 2008. The credibility of volunteered geographic information. GeoJournal 72 (3), 137–148.

Foresman, T.W. (Ed.), 1998. The History of Geographic Information Systems: Perspectives from the Pioneers. Prentice Hall, Upper Saddle River, NJ.

Goodchild, M., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221.

Graham, M., Zook, M., 2013. Augmented realities and uneven geographies: exploring; the geolinguistic contours of the web. Environment and Planning A 45, 77–99.

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and Planning B 37 (4), 682–703.

Haklay, M., Singleton, A., Parker, C., 2008. Web mapping 2.0: the neogeography of the geoweb. Geography Compass 2, 2011–2039.

Jiang, B., 2012. Volunteered geographic information and computational geography: new perspectives. In: Sui, D., Elwood, S., Goodchild, M. (Eds.), Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice. Springer, London, pp. 125–138.

Kingsbury, P., Jones, J.P., 2009. Walter Benjamin's dionysian adventures on Google Earth. Geoforum 40, 502–513.

Leszczynski, A., 2012. Situating the geoweb in political economy. Progress in Human Geography 36 (1), 72–89.

Miller, C., 2006. A beast in the field: the Google Maps mashup as GIS/2. Cartographica 41, 187–199.

Parks, L., 2009. Digging into Google Earth: an analysis of 'critic in Darfur'. Geoforum 40, 535–545.

Pickles, J. (Ed.), 1995. Ground Truth. Guilford Press, New York.

Pickles, J., 2004. A History of Spaces: Cartographic Reason, Mapping, and the Geo-coded World. Routledge, London.

Poore, B., Wolf, E., 2012. Metadata squared: enhancing its usability for volunteered geographic information and the geoweb. In: Sui, D., Elwood, S., Goodchild, M. (Eds.), Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice. Springer, London, pp. 43–64.

Scharl, A., Tochtermann, K., 2007. The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 Are Shaping the Network Society. Springer, London.

Schuurman, N., 2002. Reconciling social constructivism and realism in GIS. ACME: An International E-Journal for Critical Geographies 1 (1), 75–90.

Sheppard, E., 2005. Knowledge production through critical GIS: genealogy and prospects. Cartographica 40 (4), 5–21.

Sieber, R., 2006. Public participation geographic information systems: a literature review and framework. Annals of the Association of American Geographers 96 (3), 491–507.

Sui, D., 2008. The wikification of GIS and its consequences: or Angelina Jolie's new tattoo and the future of GIS. Computers, Environment and Urban Systems 32, 1–5.

Sui, D., Goodchild, M., Elwood, S., 2012. VGI, the exaflood, and the growing digital divide. In: Sui, D., Elwood, S., Goodchild, M. (Eds.), Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice. Springer, London, pp. 1–12.

Turner, A., 2006. An Introduction to Neogeogrpahy. O'Reilly Media, Sebastapol, CA.

Citizen Participation

Howell S Baum, University of Maryland, College Park, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Participation in making public decisions that affect individual and collective interests is a basic element of democratic citizenship. Citizen participation can make decisions more knowledgeable and legitimate. Besides possible substantive benefits, participants can gain from acquiring skills, knowledge, social capital, and power. Challenges in designing decision processes to provide such benefits include enabling all participants to exercise similar power, including participants representing a wide range of interests and perspectives, accommodating low-income citizens' difficulties participating, and using electronic media to enhance participation and knowledge.

Citizen participation refers to involvement of members of the public in societal governance. In planning, organizing, community development, and related fields, the term includes activities that enable significant numbers of individuals and representatives of groups to influence public (and sometimes private) decisions that affect their interests. Historically, 'citizen participation' emerged from efforts to empower racial minorities and economically disadvantaged groups to take part in government decision making.

Citizen Participation as Citizenship

The growth of industry and the arrival of millions of immigrants to work in manufacturing plants swelled American cities in the late nineteenth century and began to make America an urban nation. Not only the great population growth, but also the physical, social, and cultural changes in cities wrought by new residents from southern and eastern Europe presented challenges regarding how cities would be governed. In response, members of local economic and professional elites spearheaded wide-ranging progressive reforms in cities' physical infrastructure, social welfare institutions, and governance. They created commissions, including planning commissions, and other institutions in an effort to ensure that the wealthy and well educated, rather than the political bosses tied to the immigrant workers, would dominate the local government.

Decades later, when the Johnson Administration launched the 1960s War on Poverty, federal policy makers introduced expectations of 'citizen participation' that would involve the poor and racial minorities in making decisions about urban reforms. An important stimulus for this new orientation was the civil rights movement, which epitomized and pushed for grassroots participation in public institutions and policy making. The designers of the Community Action Program believed that local citizen involvement in the War on Poverty would bring benefits both by introducing new ideas and by bypassing local officials who resisted change. Many local elites resisted new participants, but federal language unambiguously called for expanded public involvement. The Economic Opportunity Act of 1964, the hallmark of the War on Poverty, declared that "The term community action program means a program … which is developed, conducted, and administered with the maximum feasible participation of residents of the areas and members of the groups served." (The Economic Opportunity Act of 1964, P. L. 88-452, August 20, 1964, Sections 202 (a) and (c)) In 1966, the Model Cities Program began with a call for 'widespread citizen participation'. The same year, a US Department of Housing and Urban Development program guide listed 'ethnic or racial groups' among desired participants in local urban renewal citizen advisory committees. The War on Poverty aimed to activate the urban poor, particularly blacks (Marris and Rein, 1982; Moynihan, 1970).

American society had changed conspicuously over the preceding century. The slaves had been freed. The continent was filled out with states and localities. Millions of European immigrants had brought foreign cultures to American cities. The franchise had been extended to all white men, blacks, and women. Government had become a huge, multidepartmental, multilevel bureaucratic welfare state. Two hundred million Americans had interests in governmental actions, and there were no ready channels for them to be heard, much less to have influence. The new citizen participation would be an experiment.

Local government officials, agency professional staffs, and elites typically looked less favorably upon this experiment than did federal policy makers and the intended new participants, and the parties disagreed about the purpose of the experiment. At most, those who had traditionally made decisions commonly believed, 'citizen participation' would mean involvement that did not disturb existing power relations. In contrast, low-income and black residents expected 'citizen participation' to give them power over policies and programs. The War on Poverty coincided with activism to register blacks to vote and gain what they regarded as 'first-class citizenship.' Consistently, racial minorities and the poor understood 'citizen participation' to be central to, and an extension of, citizenship.

People spoke of 'citizen participation' and 'community participation' interchangeably. On the one hand, since individual citizens belong to communities, by implication, communities are involved when citizens participate in decisions. On the other hand, 'community participation' can be more complex and demanding, insofar as it represents an expectation that a community as a whole influence decisions on behalf of collective interests. This broader meaning requires defining communities, identifying their interests, and elaborating ways communities can act in service of their interests. Thus, talk of 'community participation' expanded the meanings and expectations of 'citizen participation.'

A conventional concept of community, compatible with planners' view of the city as a physical place, defines the residents of a geographic territory as a community. In this view, often accompanied by the assumption that community members are homogeneous, participation of citizens is synonymous with community participation. In practice, however, residents of a place are likely to regard themselves as sharing interests only insofar as they have social relations, follow some common cultural norms, and identify with one another. Thus, residents of a place may or may not see themselves as a community and act as one. Moreover, the growth of nonspatial communities based on shared identities or interests has expanded the domain of 'communities' and added complexities to identifying community interests and promoting community action. Not only is citizen participation that involves communities more difficult to carry out than participation of individuals, but community participation also poses hard questions of defining community interests and determining when participants represent them.

Local government officials and their professional staffs have come to regard some kind of citizen participation as a normal part of decision making. They may speak of 'community participation' as well as of 'citizen participation,' although they may not attach a specific meaning to 'community.' Insofar as they distinguish 'community' participation from general participation of a jurisdiction's residents, they are likely to refer to a geographic neighborhood or subdivision. For their part, residents or community members are more likely to think in terms of community identities and interests than those of a whole jurisdiction, and they may organize to direct officials and staffs to look at the specific people or areas that matter to them.

Normative and Practical Arguments for Citizen Participation

Whether government officials and professionals who work with the public support active, not just formal, participation, and whether citizens invest much time and effort in participating depends on what they consider the rationale for and benefits of citizen participation.

The primary normative argument for citizen participation is that, as an exercise in citizenship, it is required by democracy. In this view, people have the right to deliberate on and influence decisions affecting their interests. This right does not depend on whether citizens are experts or especially knowledgeable on issues.

In fact, however, a practical argument for citizen participation emphasizes ways in which a participatory process contributes useful knowledge to decision making. When people feel that current conditions need changing, they may be unsure what exactly the problem is, what values they should use in deciding what to do, what alternatives would benefit them, and which is the best course to take. Experts may have specialized or abstract knowledge on the substance of issues under consideration, but ordinary citizens have knowledge about local conditions. For example, a housing expert may know about markets, policies, and programs in general, but residents may know about the history of housing, the housing market, cultural values related to housing, and influences on homeownership in their particular neighborhood. Both kinds of knowledge together are useful for analyzing housing conditions, framing problems, formulating alternatives, and developing strategies for change. In addition, regardless of who knows what, a process in which public officials, professionals, and ordinary citizens discuss issues together will raise questions, identify problems, and generate new ideas and knowledge that would not have been available in the absence of citizen participation.

A second practical argument for citizen participation is that people who take part in formulating problems, generating ideas, identifying goals, and developing proposals feel invested in the results. They regard plans as legitimate and authorize action to implement them. For example, when residents have an active, influential role in planning to improve public transit, they are more likely to support the final proposals than if they were not consulted or if they only observed deliberations by public officials and agency staff. Experience shows that, even if officials or experts come up with the same plans that residents favor, citizens are less likely to endorse the ideas if they are not involved. Thus, citizen participation serves the practical interests of elected officials and professionals in securing legitimacy for decisions.

A third practical argument for citizen participation is contained within the first two: citizen participation can help resolve conflicts over problems and solutions. A deliberative process including citizens offers an opportunity for people with different positions and interests to reason together. They can compare assumptions and information and can do research to see whether one position is more realistic than another or whether there is another alternative altogether that is preferable to any of the contesting positions. Insofar as participants learn together, they may come to feel that their shared knowledge is authoritative. Moreover, simply the process of discussing ideas and values may enable people who started off in conflict to develop more sympathetic views of one another and form relationships that contribute to agreement. There is no sure method to resolve conflicts, but in these ways well-organized citizen participation contributes to broadly supported, knowledgeable decisions that are likely realistically to benefit many participants' interests (Briggs, 2008; Forester, 1989).

These practical arguments emphasize the tangible benefits of citizen participation: for example, more low-income households can afford decent housing, more inner-city residents can ride public transit to employment, shopping, and social services, and more neighborhoods can have nearby recreational facilities. In addition, the process of participation itself can benefit citizens, communities and organizations to which they belong, and the jurisdictions and larger society of which they are a part.

Individuals benefit from participation in several ways. First, they experience being a citizen, being a significant part of the political system. The activity may bring self-esteem and a sense of potency. In addition, participation enables citizens to acquire knowledge about conditions and policies and to develop skills in collecting information, deliberating with others, resolving conflicts, reaching decisions, and perhaps such specific skills as analyzing or preparing budgets, assessing the historic value of homes, and determining if air or water is safe. Each episode of participation gives citizens problem-solving knowledge and skills that can be used in other cases.

Further, participation enables citizens to develop relations with government officials, agency professionals, community and institutional leaders, and others with whom they can work and from whom they can get benefits in the future. In all these ways, participation enables individual citizens to advance their interests.

The communities and organizations to which citizens belong, which may themselves be represented in a decision-making process, benefit in analogous ways. Individuals bring their experience from participation to their organizations and communities, thus adding to the latter's collective competence. In addition, the relationships individuals have developed become the bases for social capital available to organizations or communities, potentially increasing internal bonds and extending bridges to outside entities with useful resources. As a result, communities and organizations develop increased efficacy and influence over resources that serve collective interests and the interests of members. In turn, more vital and potent communities and organizations are likely to attract members who will work, perhaps contribute resources, and consider the entities legitimate (Baum, 1997; Quick and Feldman, 2011; Mandarano et al., 2011).

Finally, larger jurisdictions and society benefit from the active citizenship of participants in decision making. Individual citizens are more attached to the collectivity, are more likely to contribute their competence and leadership to civic activity, and bring and build relationships that can resolve conflicts and identify common interests. As a result, the society is likely to be more cohesive, and societal institutions are likely to identify and address problems knowledgeably and legitimately in ways that serve many interests.

Challenges in Making Citizen Participation Effective Citizenship

Whether, in fact, citizen participation brings these benefits depends on how it is organized and managed. Designers of participation processes face four challenges: to enable all participants to exercise similar power over decisions, to include participants representing a wide range of interests in conditions affected by decisions, to accommodate low-income citizens' difficulties in participating, and to use electronic media and technologies to enhance participation and knowledge.

Empowering Citizens

By its nature, citizen participation challenges decision-making processes dominated by political and economic elites. The history of citizen participation includes many instances of resistance by elected officials, business executives, and others to including ordinary citizens in deliberations. Arnstein (1969) formulated 'a ladder of citizen participation' to distinguish gradations in the power that citizens exercised in American antipoverty programs:

Nonparticipation

1. Manipulation: placing citizens on powerless advisory committees for the purpose of 'educating' them or gaining their support for others' positions.
2. Therapy: trying to change citizens' beliefs, attitudes, or values as a diversion from analyzing and changing social conditions.

Degrees of tokenism

3. Informing: giving information to citizens but not requesting or listening to their points of view.
4. Consultation: asking citizens' opinions but not making a commitment to act on them.
5. Placation: placing a few citizens on committees or boards that have limited influence over policies and practices.

Degrees of citizen power

6. Partnership: sharing decision-making power with citizens.
7. Delegated power: delegating some decision-making power to citizens.
8. Citizen control: citizens have considerable independent power over decisions.

Such a hierarchy poses the question of what determines how much power a group of citizens have in a particular situation. One factor, implicit in Arnstein's formulation, is whether traditional decision makers are willing to share power and, if so, how much and in what ways they are willing to do so. Here, it is useful to distinguish three levels of power. Manifest, and the usual focus of debate, is influence over specific, overt decisions regarding plans, programs, and policies. Beneath these decisions are others concerning which alternatives get onto the agenda. Thus, citizens may have considerable power in selecting from options on the table without having influence in determining the array of possibilities. Still one level lower are decisions, frequently tacit, about which social conditions are defined as issues or problems and even considered for the agenda. Citizens may have influence in choosing among recognized issues for an agenda but have little power to make certain troubling conditions into recognized problems that might be considered for the agenda. Political and economic elites have been most willing to share power at the top level and progressively less willing at each of the lower two levels.

A second factor affecting citizens' power is their ability to organize to gain influence in relation to formal and traditional decision makers on any of these levels. Citizen groups can use their numbers, knowledge, competence, and connections in forming alliances with some parties or persuading others to go along with their desires in getting problems identified as issues, putting issues on the agenda, or selecting particular solutions for problems.

In recent years, conflicts between organized citizens and formal decision makers, as well as difficulties resolving conflicts, have induced public and nonprofit organizations to employ facilitators with dispute-resolution expertise to design and manage decision-making processes. These processes involve efforts to include representatives of all affected interests, to establish decision rules that offer more or less equal power to all participants, to provide participants with relevant information, and to combine discussion, mediation, negotiation, and other activities that would contribute to collective learning and agreement. In the planning field, such activities may be called 'communicative planning.' Still, while participants' interests in reaching decisions, perhaps fueled by frustration with long impasses, lead them to authorize facilitators to design collaborative processes, more powerful participants

may resist sharing power with ordinary citizens. Hence, a challenge to citizens and those who want to empower them is to organize and negotiate for decision arrangements that involve relatively equal power for all (Forester, 2009; Innes and Booher, 2010).

Nevertheless, citizens who have had little experience making public decisions or negotiating with elected officials, corporate executives, or other powerful actors may be intimidated or may not know how to exercise influence. Under these conditions, a facilitator can help them act more effectively. Another handicap citizens may have is that they have less general knowledge or specific information about issues. This may be particularly so when the issues involve scientific or technical knowledge, such as questions of environmental quality. Industrial leaders may employ or routinely consult with experts, while lay citizens may find issues overwhelming, imagine that they could never understand them, and feel intimidated by others who speak fluently about them. Again, a facilitator can be instrumental in aiding not only laypersons, but everyone, by including experts in deliberations and structuring discussions so that participants can share and analyze information together and come to common understandings about how to assess conditions (Brody et al., 2003; Medoff and Sklar, 1994).

Representing Citizens and Communities

Time and space limitations make it impossible to include everyone affected by a public issue in making decisions. At the same time, not everyone with a stake in an issue wants to participate. Thus, any decision process includes only a sample of interested parties. If everyone has the same interests, then whoever takes part, whether a few or many, can represent these interests. However, unanimity of interests is rare; what makes a condition a public issue is the fact that people see it differently and want different actions to change or preserve it. Hence participation presents a question of representation: Does someone speak effectively for each group that has particular interests in the issue? The answer to this question depends on identifying the different interests and those who hold them. When an issue is relatively simple and the number of affected parties relatively small, it is easier to identify interests and people holding them. Many issues, however, are complex and affect multiple parties, and in these instances, representation poses challenges.

Early public commissions in planning and other fields were elite groups, composed of leaders in business, philanthropy, 'society,' the professions, cultural institutions, politics, religious institutions, and universities. These people claimed to have broad knowledge of local affairs and the ability to identify and speak for various citizens' interests. Advocates for 'citizen participation' in the 1960s challenged the representativeness of such bodies, calling attention to the typical exclusion of the poor and racial minorities. The social movements and poverty programs of that decade contended that the poor and racial minorities were normally the best exponents of their interests and hence should participate directly in making decisions. In the last decades of the twentieth century, government bodies, nonprofit organizations, and community organizations gave greater, more explicit attention to selecting and admitting participants who would represent broad interests.

There is no single correct way to define or measure representation and who or what is to be represented. One simple measure is statistical representativeness: whether participants resemble the overall population demographically, at least by race or ethnicity and by sex, perhaps also by age, income, and/or neighborhood. However, while statistical representativeness is politically prudent, as the sole measure of representation it has limitations. It erroneously assumes that everyone of the same race, sex, age, income, or neighborhood thinks the same and wants the same things. In addition, representation of demographic categories does not guarantee representation of interests. For example, on housing issues homeowners and renters have different interests, as may private developers and neighborhood associations. Similarly, in decisions about community development, merchants, social service providers, religious congregations, employers, schools, and other institutions have various interests. Thus, both types of representations are valuable for increasing the democratic quality of participation, as well as the available knowledge and the likely legitimacy of decisions.

For government agencies and other organizations concerned with 'citizen participation' in a general sense, including people from these categories may be sufficient for representation. However, those who are concerned with 'community participation' must give more specific attention to the identity of participants. To begin with, they must identify which communities have interests and who their members are. Then they must identify the specific groups and individuals within each community that care about or matter with regard to the issues at hand. This task entails identifying leaders (informal as well as formal) and people who are particularly knowledgeable (about issues, the community, and people in the community). With this understanding, those organizing decision making can include specific individuals who are representative of a community not just statistically and in terms of interests, but also in terms of perspectives.

There is no uniquely representative group of participants, but attention to these three types of representation will produce a group of considerable breadth. In this respect, the process of deciding who and what are to be represented and who can represent these people and positions matters. Traditionally, government agencies selected participants. In recent decades, a growing number of public and nonprofit organizations have consulted with community organizations, business associations, social service agencies, and others to consider together whom and what to represent and how. In other words, a representative process of selecting participants contributes to a representative process of deciding on issues.

In practice, it may be unclear at the beginning of a decision process who and what should be represented. Preliminary discussions of an issue may be necessary to bring issues to light and identify relevant parties to learn whom to represent and how. Moreover, representation is fluid. As a process proceeds, participants may identify new issues or interests that call for new representatives, and some issues and interests may recede in importance. As a practical matter, during deliberations of any great length, participants may drop out, and new participants would be necessary to continue representing certain groups, interests, or perspectives. The challenge at such moments, as throughout a decision process, is to be attentive to

who is affected by or interested in issues and to try to include persons who can speak for significant positions. For generating participation that is widely considered legitimate, this spirit matters more than any formula.

Enabling Participation: Predicaments of Low-Income Citizens

Even if a decision process is open to broadly representative participants and even if participants will all have significant and comparable power, citizens will not automatically take part. Participation has requirements and costs, including opportunity costs, and not everyone can afford to get involved. Above all, participation requires time; it depends on having discretionary time to spend in meetings and other activities, plus a willingness to forgo other uses of time that arise from responsibilities or bring pleasure, such as working extra hours, socializing with friends or family, watching television or movies, or helping children with homework. In addition, participation requires certain skills: normally (although not always) fluency in English; an ability to work with diverse other people, endure and work through conflicts, and come to a meeting of minds; an understanding of parliamentary procedure; and a conceptualization of a possibly long-term process of discussion and analysis as a means to solving problems. Further, participants must be at ease speaking in public, comfortable with their appearance and language, and patient with others who disagree and with a process that goes on without immediate solutions for problems.

Well-educated middle-class adults are most likely to have the time, skills, and experience needed for citizen participation. For many professionals, meetings, planning, and problem-solving processes are a normal part of their work, and they have much practice at collective decision making. They may already have considerable information about issues either from their work or from their reading or viewing of the media. If they need child care to attend meetings, they can afford it. Lower income adults, in contrast, are less likely to be able to participate. Economic need may require them to work evening or weekend hours that conflict with meeting times. Single parents do not have a spouse to look after children to permit meeting attendance, and they may be unable to pay for child care, although an extended family member may be available. Lower income adults may be self-conscious about their language or their dress, particularly if other participants are public officials or middle-class professionals. Although they may care greatly about the issues and have considerable information about local conditions, they may feel intimidated by how much more they imagine wealthier or better educated participants will know. They are likely to have limited practice with formal decision making, planning, and conflict resolution. Insofar as their time is scarce and daily needs are pressing, they may find the cost of the time required for participation unacceptably high and may be impatient with talk that does not produce immediate benefits. Whatever the socioeconomic background, people for whom English is not the first language may have difficulty participating in English discussion without interpreters.

Thus, representative participation depends on accommodating low-income and other citizens who face obstacles to participating. People are more likely to turn out when meeting times fit their schedules and locations are at sites where they feel safe and in friendly territory. People are more likely to attend when organizers have contacted and invited them through someone familiar and trusted in their community network. Child care makes it possible for parents to participate, and interpreters aid non-English speakers at meetings conducted in English. Providing clear basic information to all participants contributes to a sense of equality and competence on the part of those who start out with limited information, and bringing in experts helps similarly. Beginning a decision-making process with activities that both introduce participants to one another and describe the anticipated process and ground rules helps the less confident take more active part. Throughout, sensitive and explicit facilitation helps increase people's capacity to participate powerfully and effectively in reaching agreements.

Potentials and Limitations of Virtual Participation

Electronic communication has the potential to overcome some obstacles to face-to-face citizen participation, involve participants in complex activities, make participation more knowledgeable, make it more representative, and reduce power inequalities among participants. Communication technologies make it possible for citizens to interact without being in the same place or even acting at the same time. Most simply, electronic communication enlarges the opportunities for citizens to take part in or follow conventional face-to-face meetings. Individuals whose family responsibilities, work obligations, or health conditions prevent them from attending meetings can listen to, observe, and contribute to discussion electronically. Those who do go can blog reportage or commentary on events to others elsewhere. More complexly, communication technologies, supported by a growing number of online tools, make it possible for large numbers of citizens at face-to-face meetings or linked with Internet access to generate ideas, provide and develop information, express views on or debate ideas and issues, resolve conflicts, draft documents, prioritize options, and commit themselves to action together (Leighninger, 2011).

While electronic communication technologies allow virtual participation in decision making, they do not guarantee it. One constraint is the cost of hardware and some software. Public and nonprofit organizations may not be able to afford necessary equipment and material, and because individuals' computer ownership depends on their income, lower income citizens are least likely to have the consistent access necessary for ongoing participation. Moreover, some of the personal concerns that deter face-to-face participation also discourage virtual participation. Persons who are confident about their knowledge, views, and voice are most likely to take part. In addition, those who are comfortable with computer technologies, particularly those with more education and younger generations, are most likely to venture into virtual participation. Hence, as with conventional deliberation, outreach, especially personal contacts or contacts through community networks, is necessary to encourage and enable people to participate through unconventional means. In general, the complexity of online interaction, especially when participants do not see or know one another, requires careful planning, orientation of participants, and continuing explanation and

feedback to participants. Because virtual participation does not provide the emotional connection of face-to-face deliberation, individuals may feel limited motivation to join in or continue.

Virtual participation experiences have been mixed. Although the amount of communication has increased, often participants are generally the same who attend face-to-face meetings. Online technologies make it possible to obtain and generate considerable knowledge, but participants vary in their openness to new information about issues. Finally, while people may take part in discussions, surveys, games, and even drafting documents, only some are likely to become committed to making or implementing a decision. How virtual participation develops in conjunction with face-to-face participation depends on what technologies are developed, how people learn to use them, and what participation citizens want (Evans-Cowley, 2010).

Conclusion

Citizen participation in making public decisions that affect their individual and collective interests is a basic component of citizenship. However, public organizations may actively or inadvertently discourage citizen involvement, and citizens vary considerably in confidence that they have the political and intellectual authority to take part. As a result, those with higher income and education participate most commonly, and middle-class and white citizens are generally more active than lower income and racial or ethnic minority citizens. Thus, participants often do not represent a wide range of groups, interests, or perspectives, and higher income participants typically exercise more power over decisions than do lower income participants. Hence, deliberate effort is necessary to recruit participants and design processes to create decision making that is broadly representative, in which participants exercise similar power, and where participants act on shared, useful knowledge. Community organizing can draw public officials' and professionals' attention to these challenges and encourage them to address them.

See also: Advocacy and Equity Planning; Governance; Neighborhood; Race: Ethnicity and Urban Planning.

Bibliography

Arnstein, Sherry, 1969. A ladder of citizen participation. Journal of the American Institute of Planners 35, 216–224.

Baum, Howell, 1997. The Organization of Hope: Communities Planning Themselves. SUNY Press, Albany, NY.

Briggs, Xavier de Souza, 2008. Democracy as Problem Solving: Civic Capacity in Communities across the Globe. MIT Press, Cambridge, MA, and London.

Brody, Sam, Godschalk, David, Burby, Ray, 2003. Mandating citizen participation in plan-making: six strategic planning choices. Journal of the American Planning Association 69, 245–264.

Evans-Cowley, Jennifer S, 2010. Planning in the age of Facebook: the role of social networking in planning processes. GeoJournal 75, 407–420.

Forester, John, 2009. Dealing with Differences: Dramas of Mediating Public Disputes. Oxford University Press, New York.

Forester, John, 1989. Planning in the Face of Power. University of California Press, Berkeley, Los Angeles, and London.

Innes, Judith E., Booher, David, 2010. Planning with Complexity: An Introduction to Collaborative Rationality for Public Policy. Routledge, London and New York.

Leighninger, Matt, 2011. Using Online Tools to Engage and Be Engaged by the Public. IBM Center for the Business of Government, Washington.

Mandarano, Lynn, Meenar, Mahbubur, Steins, Christopher, 2011. Building social capital in the digital age of civic engagement. Journal of Planning Literature 25, 123–135.

Marris, Peter, Rein, Martin, 1982. Dilemmas of Social Reform, second ed. University of Chicago Press, Chicago.

Medoff, Peter, Sklar, Holly, 1994. Streets of Hope: The Fall and Rise of an Urban Neighborhood. South End Press, Boston, MA.

Moynihan, Daniel P, 1970. Maximum Feasible Misunderstanding: Community Action in the War on Poverty. Free Press, New York.

Quick, Kathryn S., Feldman, Marsha S, 2011. Distinguishing participation and inclusion. Journal of Planning Education and Research 31, 272–290.

The Economic Opportunity Act of 1964, P. L. 88-452, August 20, 1964, Sections 202 (a) and (c).

Relevant Websites

http://www.cpn.org/ – Civic Practices Network.
https://www.civicus.org/ – Civicus.
http://communitybuildinginstitute.org/ – Community Building Institute.
http://www.community-problem-solving.net/ – Community Problem Solving.
http://www.deliberative-democracy.net/ – Deliberative-Democracy.net.

Citizen Science

Hauke Riesch, Media & Communications, Brunel University London, Uxbridge, Middlesex, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Citizen science (CS) refers loosely to participation in scientific research by members of the public. Diverse programs of CS reflect a host of different aspirations, goals, and ideals. Focusing mainly on the meaning of CS as coined by Bonney, the article will provide a brief account of the different types of projects that have variously been labeled CS, as well as some of the previous classification attempts. The main part will consist of surveying some of the problematic concerns that CS raises, such as the boundaries between lay and scientific expertise, the performance of CS in terms of science communication and other goals, and the ethical issues raised by volunteer research.

Introduction

The concept of citizen science (CS), or public participation in scientific research (PPSR), refers loosely to scientific or environmental monitoring projects involving participation of members of the public. In the mid-1990s, the term was applied by Rick Bonney (Bonney et al., 2009) to refer to a series of hybrid science and science communication projects that were then being pioneered by Bonney and his colleagues at Cornell University's Lab of Ornithology (CLO). These projects were designed to involve lay public participation in the gathering of scientific data, helping scientific research projects in obtaining new data as well as giving interested members of the public an opportunity, not just to learn more about ornithology, but to actively contribute to new and developing science. The initial concept has been widened and elaborated on both the science and the science communication sides, with new projects by the CLO and others attempting to involve members of the public in more activities and other stages of the scientific research process than just the data gathering (Bonney et al., 2009; Shirk et al., 2012), as well as formulation of a wider rationale for the science communication benefits of CS based on ideas that have emerged from science education studies (for example, Abd-El-Khalick et al., 1998, applied to CS at the CLO, see Trumbull et al., 2000; Brossard et al., 2005) that foreground teaching scientific methodology and the 'Nature of Science' as opposed to simply imparting facts. Moreover, learning about the Nature of Science in an experiential and practical manner away from the classroom has also been identified as particularly beneficial (Tuss, 1996).

Although the term CS started with the CLO as an innovative combination of participatory science and science communication, scientific and environmental monitoring projects that involved lay participants had been going on for some considerable time, especially in the field of ornithology (Greenwood, 2007; Toogood, 2011), to some of which the label CS has been applied retrospectively. These have evolved out of enthusiastic amateur natural history societies' activities, such as the annual Audubon Society Christmas Bird Count, which has been going since Christmas 1900 and is now describing itself as "the longest running Citizen Science survey in the world" (Audubon Society, 2013). Individual amateur contributions to science (such as the science of Benjamin Franklin) are often being cited as forerunners of CS (Silvertown, 2009), and although there is almost no relation any longer between the concepts and philosophies of CS in the CLO and eighteenth-century gentlemen scientists, it is probably fair to trace a direct conceptual lineage from interested lone amateur naturalists, to the establishment of amateur natural history societies in the late nineteenth century that draw on and attract a larger membership, and finally to contemporary participatory science research (Miller-Rushing et al., 2012).

CS projects that run along the lines of the CLO's programs have proliferated in recent years (the CLO maintains a non-comprehensive list of other CS projects at Citizen Science Central, 2013), and as the concept has traveled across different disciplines, institutions and countries, the emphases, ideals, and stated goals have often shifted and now produce a large variety of different types of CS projects that are rather loosely connected by an emphasis on combining participatory research with public engagement. The term CS has often been retrospectively applied to some of the predecessor amateur projects discussed above, as well as to other participatory science projects that have been developed independently of the CLO's philosophy, leading to a very diverse family of concepts that have been labeled CS. Added to this, the term was independently coined in the mid-1990s in the UK in an influential book by Alan Irwin (1995), where it refers more to the democratic participation of citizens in science and science policy (as part of the then developing 'upstream engagement' paradigm in public engagement studies). While there are some considerable overlaps between Irwin's and Bonney's concepts of CS (such as emphasis on public participation in science and the importance of dialogue between experts and the public), they are also fundamentally different in that they are intended to address different problems – combining research and science communication in one and producing democratically informed science policy in the other. Consequently, some authors writing within the Irwin tradition have expressed dissatisfaction with the other idea of CS as being more or less irrelevant to the problems they want to address (for example, Stilgoe, 2009).

As a result of the varying meanings attributed to the term, some commentators (including Bonney et al., 2009) now prefer to talk about 'PPSR'; whether for better or worse, CS has become the dominant term in which to describe public

participation in research projects, and this, in turn, has colored perceptions of it. While it appears that the different provenances of the term CS are not very widely known, a situation that can often lead to some confusion when scholars try to provide summaries of the philosophy behind CS, it also has produced some interesting hybrid philosophies of participatory science. These would, for example, emphasize the communal sense of ownership of the local environment that participation in CS could bring about, and which therefore can play a role in enhancing the policy voice of disadvantaged communities, important, for example, in developing countries where the voice of the local community in decision-making on environmental developments has often been marginalized, or which, like University College London's 'Extreme Citizen Science' project, aim to be "a situated, bottom-up practice that takes into account local needs, practices and culture and works with broad networks of people to design and build new devices and knowledge creation processes that can transform the world" (ExCiteS, 2013). Similarly such a hybrid conceptualization can be employed by the statutory sector as a way of simultaneously getting public help in monitoring, for example, the spread of tree diseases as well as providing the public with an avenue to input in any resulting policy (Forestry Commission, 2011). Even as part of the more science communication-oriented CS events, some mostly UK-based writers influenced by the Irwin tradition of public engagement, tend to foreground the value of CS as a democratic rather than scientific exercise (Hemment et al., 2011). While this variety of interpretations of CS is broadly being encouraged by most researchers as showcasing the innovative potential of CS, it has also led to some confusions and uncertainties over what CS should be and can achieve.

It has also led to several attempts to produce a more cohesive and workable classification of the various incarnations of CS. Shirk et al. (2012), developing a scheme they previously published in Bonney et al. (2009), distinguish between the degree of participation ("the extent to which individuals are involved in the process of scientific research") and quality of participation ("the extent to which a project's goals and activities align with, respond to, and are relevant to the needs and interests of public participants"). They thus arrive at five-idealized project models: *Contractual* projects (communities ask experts to conduct specific research, similar to the 'science shop' model, Leydesdorff and Ward, 2005); *Contributory* projects (public contributes data to projects designed by scientists); *Collaborative* projects (designed by scientists, but public also help with project design, data analysis, and dissemination); *Cocreated* projects (designed together by scientists and public); and finally *Collegial* contributions (members of the public conduct research largely independently of institutional science, such as the AIDS activists famously analyzed by Epstein, 1996).

Wiggins and Crowston (2011) distinguish between *Action* projects (grassroots projects "using scientific research as a tool to support civic agendas": p. 5); *Conservation* projects ("engaging citizens as a matter of practicality and outreach" to "support stewardship and natural resource management goals," often combined with "specific educational goals": p. 5); *Investigation* projects (focused on scientific research, with education a valued but not central goal); *Virtual* projects (where the interaction between researchers and public occurs online only); and finally *Education* projects (where the primary goal is science education rather than the science or the monitoring). There is some overlap between these categories, for example, in 'Zooniverse' series of the "Citizen Science Alliance" (Citizen Science Alliance, 2013) projects, which includes *Galaxy Zoo*, where the interaction with the public occurs only online, and which claims academic research as its primary goal, with the public engagement a welcome side effect.

The above typologies center around idealized project types based on the analysis of concrete examples they examined. We can alternatively draw up a map of CS projects that classifies them along several dimensions, such as number of participants, quality of interaction, and project purpose. Here we can observe that usually the quality of interaction decreases as the number of participants increases (because project scientists cannot give large amounts of individual attention to thousands of participants), leading to a trade-off between 'breadth' of and 'depth' of participation (Riesch et al., 2013). Two ideal typical projects lie at opposite ends of a spectrum, where the public contribution either amounts to very small individual contributions or fewer but higher quality and more engaged contributions.

Prainsack (2014) develops a classification of CS based on how the project is structured along six dimensions, or clusters of issues: Who has influence on setting the agenda; who participates and what do they bring to the project; how the project affects the communities and community identities, who decides what counts as success and how it is evaluated, openness and accessibility of the resulting data and publications, and finally project funding. Crucially, she argues that CS projects are not static and that therefore answers to these questions can be different at different stages of a project.

Boundaries and Expertise

The coming together of professional and lay expertise in CS opens up questions surrounding the definition and boundaries of expertise. Ellis and Waterton's (2005) ethnographic analysis of the networks in natural history monitoring projects that operate across the lay/expert boundary emphasizes that this boundary can be very fluid. The projects they analyze are slightly removed from CS projects designed as public engagement exercises that align more with the current CLO tradition, because the amateurs contribute specific lay expertise on a particular species or geographical location; they are involved because they can add unique and specific contributions through their own collection efforts, and therefore the project makes use of amateur experts rather than lay members of the public. In this case the blurring of the boundaries between experts and amateurs may be more obvious because the participating amateurs bring their own expertise to the table without wanting or needing to be educated by the project. In some 'depth' projects, the interaction is prolonged and in-depth so that the participants acquire expertise along the way, rather than necessarily producing a blurring of boundaries. The lay participants in such projects can also develop a discernible lay–expert identity through which they can challenge and engage in the policy process and the professional

expertise as a distinct group (for example, in Cornwell and Campbell's (2012) case study).

Amateur experts, however, can also participate in the 'breadth' educational projects. As scientists working on such projects have experienced, the projects often attract participants such as retired university professors or enthusiastic collectors who can "run rings around many of the scientists when it comes to sort of specifics of what they're interested in" (CS scientist quoted in Riesch et al., 2013). In this sense, the often hoped for aim of crossing boundaries between experts and the public can in some ways be realized, although the boundaries are clearly blurred to begin with.

'Breadth' and 'contributory' style CS projects, especially when intended explicitly as public engagement aimed at people who are not experts, however, might actually feel the need to emphasize the boundaries. In these CS projects, the professional expertise of the participating scientists is often one of the selling points that can entice people to participate. Participating members of the public expect to be guided by people who have expertise in the area and hope to learn science, scientific skills, and techniques. They are explicitly placed in a position of less epistemological and professional authority regarding the data they help produce. On the other side, the participating scientists, who often have years of training and paid positions within the academy or scientific research institutions, are ultimately the decision-makers who decide on which publicly contributed data are worth including and which directions the projects should take. In addition, they may be inclined to resist any blurring of the boundaries, as their professional identity is invested in the expertise they bring to the project. Consequently, greater public participation in the policy process as a sometimes hoped for by-product of educational CS is problematic, as it frames the science as a "ready-made set of issues, bounded and constrained, or pre-framed" (Toogood and Everett, 2013: p. 616), which consequently relegates the role of the public to 'nonnegotiable' and tightly constrained activities. Toogood and Everett thus argue that "the citizenship within such [projects] may actually reduce public awareness of political questions about biodiversity and frame relationships in ways which reinforce the distance between scientists and citizens" (p. 10).

Greater public participation in the actual scientific decision-making is envisaged by the 'cocreated' and in some cases the 'collaborative' CS projects that Shirk et al. (2012) describe and aim to encourage further. Yet these projects are still relatively rare within the universe of CS and the blurred boundaries in decision-making authority between experts and the lay participants may be a contributing factor to their relative scarcity. If intended as public engagement projects, there are inevitable questions to be asked whether cocreated projects can fulfill some or even most of the aims that variously sit behind public engagement: for example, as a way of creating public awareness of the scientific and/or environmental problems, cocreated projects rely on the initiative of people who are already aware of those problems. Similarly, when the aim is about teaching science and scientific practice, these projects also rely on participants with at least a minimal amount of background knowledge already, and although they have potential for educating participants beyond that level, this potentially limits the aim of bringing science to a broader public, regardless of their abilities and background knowledge. The value of cocreated projects thus lies in the science and is more likely to be closer to the CS projects analyzed by Ellis and Waterton (2005), rather than being particularly useful as public engagement tools. Therefore, the way a CS project sets itself up and defines its goals influences how the boundaries between the experts and the participants are defined and how they evolve with the project. While this is in itself not a bad thing, it limits some of the often rather lofty goals that proponents of CS often emphasize; instead of breaking the boundaries between participants and experts, and in the process teaching and/or inspiring the public about (a specific domain of) science, it looks more likely that a project can do either very well, but not really both at the same time.

One of the problems here is that, as described above, CS is a contested concept arising from two very different traditions, with a history that transitioned from enthusiastic amateur participation to public engagement with previously unenthusiastic participants; different goals, aims, norms, and values have been grafted together, and it seems unrealistic that just one particular setup can deliver them all at the same time.

For participating scientists, the data quality issue is seen as crucial, because on this the credibility of the resulting science will ultimately rest. This is not a judgment on the public's ability to gather useful data or the participating scientists' perceptions of that ability, but instead relates to the scientists' need to sell their work to a wider scientific community that may remain very skeptical about the scientific value of publicly gathered data. The peer-review system hinges on the ability of the participating scientists to persuade other scientists who may not share their enthusiasm about publicly gathered data. Consequently, it may also be difficult to report surprising or revolutionary results with CS (Riesch and Potter, 2014), because the default assumption may be that the data are to blame for anomalous results. Attitudes of the wider science community within the relevant field are probably one of the reasons why the scientifically most successful CS projects so far have tended to be in fields that already have an established history of amateurism: Astronomy, natural history (especially the subfield of ornithology), and biomedical sciences.

In order to address data quality, scientists in CS projects can adopt several strategies (see Riesch and Potter, 2014), including validating the data themselves, quizzing the participants on their background knowledge, or, probably most commonly, making the data gathering as easy as possible. For example, asking to count the number of spiders in a patch rather than asking participants to identify spider species, limits the data the project gathers to fairly simple questions, but which nevertheless can contribute to our scientific understanding by allowing a more geographically diverse survey than a small team of scientists could feasibly undertake themselves.

Scientists' and Public Motivations

Despite these issues, it is undoubtedly true that scientists often get excellent data from CS studies, and it is to be hoped that the continued success of these projects will pave the way for making publicly gathered data generally more acceptable within the wider scientific community. This would allow CS

projects to be broadened further to disciplines that do not traditionally partake in such projects, which, in turn, would help also to ameliorate some of the credibility issues.

Participation in public engagement is often by itself a motivator for scientists. Some get genuine enjoyment out of engaging with the public and also often feel a sense of duty (see, for example, Riesch and Potter, 2014). CS affords opportunities for scientists who are not experienced in public engagement to experiment without sacrificing their commitment to research, because when it is going well they can do both at the same time. As such, CS can be advertised as an attractive proposition for scientists, who can get both research and public engagement done at the same time leading to a 'win-win' situation. This situation can enhance personal motivations for scientists to participate in further public engagement projects; something that they are under increasing institutional pressure to do.

Studies that survey scientists' opinions about public engagement and science communication (for example, Burchell et al., 2009) often indicate that while they view such activities as important and often pleasurable duties, they also can cause a lot of problems when treated as an add-on to an already busy workload. Consequently, appropriate institutional reward structures need to be in place to encourage scientists to spend more time in engagement rather than (or as part of) their own research. However, while CS is sometimes seen as a natural solution to this problem (scientists can get both science and public engagement done at the same time), in practice, the right institutional support still needs to be in place to encourage participation in public engagement. Without it, even CS will fall short on the stringent demands for scientists' time, as scientists might feel that they get less research done in a CS project than they would have otherwise. Tasks performed by scientists on CS projects on one half of the equation can end up detracting from work in the other: some interviewees in Riesch et al. (2013), for example, felt that to give the public engagement side of their work the attention it deserves meant that they had to cram the science bit into their free time on weekends, evenings, and holidays.

There are two options on how to split the science and engagement work in a project: Either a scientist works on both sides part-time, or the task is split between dedicated scientists and science communicators within a project. In the former case, as outlined above, scientists often can feel that they are unable to give both sides of CS the attention and time it needs. Without adequate institutional support rewarding public engagement activities, this can have repercussions on their career progressions, because the amount of scientific research papers 'per pound' spent on the project will be less than in pure research projects, which some scientists have worried will look bad on their curriculum vitae when reviewed by faculty promotion or evaluation panels, or when they apply for other positions.

The other way of splitting responsibilities within a CS project is by hiring dedicated science communicators and public engagement professionals to interact with the public. This practice allows the scientists to concentrate on their scientific work, but has the obvious drawback that the public participants no longer get to interact with the leading experts, thus removing one of the primary selling points of CS.

Roy et al. (2012) have investigated the motivations in the statutory sector in promoting CS and similar activities to expand environmental monitoring and, sometimes more implicitly, to increase democratic engagement in science and environmental policy. Summarizing a series of interviews within (mainly UK) government agencies and departments, they list cost-effectiveness of CS environmental monitoring as particularly important, although as with the scientists studied by Riesch et al. (2013), discussed above, a minority of participants were worried about data quality. Overall, however, the utility of CS was seen as very large, with citizen scientists being able to collect more data and in more diverse places than government monitoring could on its own; this was, however, seen to bring risks and ethical issues with it as well (see below for further discussion).

For participating members of the public, motivations are found to be very diverse (for example, Raddick et al., 2010; Rotman et al., 2012). Their motivations depend on the type of CS project in question and can even vary over the course of a single project. They include the attractiveness of participating in real scientific research, the opportunity to learn more about science, contributing to the local community or to environmental conservation, or simply having a great day out. Even the possibly more time consuming and desk-based tasks, such as classifying galaxies for Galaxy Zoo, can be a pleasurable diversion for people who are interested in astronomy. The reasons for why people participate in CS are in many ways as diverse as the field of CS itself – different types of projects appeal to different people. For this reason Roy et al. (2012) argue that "it is not easy for users of data to design and develop a citizen science project that meets the needs of all volunteers," so that "projects must be tailored to match the interests and skill-sets of participants and understanding the motivations and expectations of potential volunteers is crucial to developing successful projects" (p. 5). The more hands-on CS programs that address relatively simple questions (as in the spider example above) are, for example, great for school kids and their teachers, who may be looking for a structured and meaningful activity for their classes. Galaxy Zoo, as an *investigation* project that is more interested in collecting scientific data than in the public engagement aspect (which is more of a side benefit), appeals to people who are already interested in astronomy but have so far not been able to pursue this interest scientifically. Projects aimed at local environmental issues might appeal to communities and community leaders who are interested in creating more ownership of the local environment, thus empowering the local community in their interactions around these issues with local authorities. These varying motivations can, and often do, intersect, leaving any particular CS project with participants who get involved for a variety of reasons.

Ethical Considerations

Although the ethics of CS have not been discussed extensively as a headlining focus of academic literature, practitioners themselves have thought at length about ethical issues, as is evident both in the number of blog posts written by CS scientists, and also by comments made in interviews with them

(see the discussions in Riesch and Potter, 2014). Ethics tend to be discussed in more detail when issues such as commercial exploitation of CS data come into play (see Bonsu and Darmody, 2008; Prainsack, 2014), which is a potential worry in some biomedical CS projects.

Ethical concerns can be raised in relation to both of the two main stakeholders in CS: the public and the professional participants. The former concerns mainly address two issues: ownership of data and remuneration for services given ostensibly for free. Firstly, the public's gathering of scientific data will give them a recognizable claim to ownership as coproducers of those data. Commercial exploitation of the data aside, the concerns relate to wider discussions over scientific transparency with data, but with CS such issues are even more urgent because the direct involvement of citizens as data producers rather than just funders (as taxpayers) of scientific research. These problems may be more easily solved by ensuring that the data and resulting research papers are made publicly accessible, though there may also be practical obstacles such as the cost of publishing open access or potentially sensitive personal information in the biomedical projects. Here discussions within the wider open access movement apply.

Related to this issue though is a second, slightly more intractable, one. Members of the public participate in CS usually without being recompensed, at least monetarily, for their time and effort. For example, criticisms made of *Big Society*, a recent UK policy initiative that aims to increase public volunteering, suggest that it is a matter of outsourcing public sector responsibilities onto unpaid volunteers in the guise of a local empowerment rhetoric (Kisby, 2010).

Instead of paying them in money, CS programs aim for other ways to compensate their participants. In the case of projects aimed at public engagement, such compensation comes in the form of an opportunity to learn, interact with scientists (although see the provisos above), and the enjoyment they get out from participating. Managers of some projects, such as Open Air Laboratories (Davies et al., 2011), are very conscious of the issue, to the extent that they discourage use of the term 'volunteer' for public participants. Although it would be unfair to claim that CS exploits free public labor, the question can still be raised over whether such nonmonetary compensation is enough. The participating experts and scientists are being paid a salary, even if they themselves report that they also getting an enormous amount of enjoyment out of CS projects. Although it can be argued that the participating experts have the experience and education to command such monetary compensation, it opens up the boundary issues discussed above, as CS moves even further away from the ideal that some people see in it of breaking down barriers between professionals and amateurs. Instead of a cooperative collaboration of equals, CS clearly delineates between, on the one hand, the scientists with authority over the direction of a project, who have the qualifications and expertise, and who get paid for their time and get to develop professional careers through scientific publications, and, on the other hand, public participants who are given very little epistemic authority, certified expertise, and monetary compensation.

The reliance on participants' free time, effort, and goodwill also has repercussions on project development and planning itself. Firstly, there may be the effect of public fatigue from being asked to do voluntary work, which can make recruitment of enough participants difficult. Secondly, especially in more small-scale projects, the participating scientists may not be able to rely on volunteers to deliver to targets, or even turn up on time, which can become problematic when strict time frames are involved.

Another form of remuneration that scientists receive for their work on CS is the professional recognition (or symbolic capital) that comes from authoring the resulting scientific papers. Theoretically, the authorship of scientific papers can be extended to all participants who produced data for it, but as a practical matter this is often seen to be impossible, at least for the large-scale 'breadth' projects, where participants number in the thousands. 'Depth' projects, where public participants are fewer and more engaged, offer more obvious instances where coauthorship of the resulting papers may be offered; similarly, such recognition may be possible in those projects that involve extensive public participation, when selected participants have significant input into new discoveries (such as Galaxy Zoo, Clery, 2011).

Finally, CS raises ethical concerns on the science side, both for individual scientists and sometimes for science at large. Asking members of the public to do something for free may end up limiting opportunities for scientists. This may be more of an issue for individual scientists without steady employment, such as postdocs and graduate research students, who might see potential jobs being 'outsourced' to volunteers, an issue that has been at least jokingly mentioned by a postdoc who Riesch and Potter (2014) interviewed. While this may be less of a problem for the 'breadth' projects that rely on large surveys in which many individuals produce small amounts of data, some of the more engaged interactions in the 'depth' projects may intersect with the ethical problems that have been highlighted with scientific research institutions that rely upon unpaid internships (Perlin, 2012). Concerns about such internships were raised that university research performed by unpaid research assistants may crowd out professional scientists – mostly the postdocs who are one of the most financially and professionally insecure groups within the scientific hierarchy. More insidiously, perhaps, the CV-building opportunities afforded by these internships may end up being open only to students who have independent financial means. Parallel to the debates that are going on over unpaid internships within other aspirational professions such as journalism or the banking sector, the concern is raised that poorer people are being denied a route into the profession.

As CS has evolved from the amateur science performed by voluntary membership organizations of the late nineteenth century, which, in turn, derived from the time when science was largely an unpaid hobby of gentlemen with independent means, we may want to look back at the positive aspects of the professionalization of science: By making science a paid profession, the developments from the nineteenth century have opened up science to a wider portion of the population and the vast talent pool of those who cannot pursue science in their spare time. As many commentators have done, we can treat CS as an inheritor of the earlier amateur science tradition, but we can add that it also needs to confront this heritage and to ask itself potentially uncomfortable questions about the actual end effect of what it is trying to do, especially since this effect is

almost the exact opposite of the democratic and progressive intentions of many of its proponents.

See also: Environmental Sciences; Science Education; Science and Technology Studies: Experts and Expertise; Science, Public Engagement with.

Bibliography

Abd-El-Khalick, F., Bell, R.L., Lederman, N.G., 1998. The nature of science and instructional practice: making the unnatural natural. Science Education 82 (4), 417–436.

Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J., Wilderman, C., 2009. Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. Center for Advancement of Informal Science Education (CAISE), Washington, DC.

Bonsu, S.K., Darmody, A., 2008. Co-creating second life: market – consumer cooperation in contemporary economy. Journal of Macromarketing 28 (4), 355–368.

Brossard, D., Lewenstein, B., Bonney, R., 2005. Scientific knowledge and attitude change: the impact of a citizen science project. International Journal of Science Education 27 (9), 1099–1121.

Burchell, K., Franklin, S., Holden, K., 2009. Public Culture as Professional Science: Final Report of the ScoPE Project. London School of Economics, London.

Clery, D., 2011. Galaxy Zoo volunteers share pain and glory of research. Science 333 (6039), 173–175.

Cornwell, M.L., Campbell, L.M., 2012. Co-producing conservation and knowledge: citizen-based sea turtle monitoring in North Carolina, USA. Social Studies of Science 42 (1), 101–120.

Davies, L., Bell, J.N.B., Bone, J., Head, M., Hill, L., Howard, C., Hobbs, S.J., Jones, D.T., Power, S.A., Rose, N., Ryder, C., Seed, L., Stevens, G., Toumi, R., Voulvoulis, N., White, P.C.L., 2011. Open air laboratories (OPAL): a community-driven research programme. Environmental Pollution 159 (8–9), 2203–2210.

Ellis, R., Waterton, C., 2005. Caught between the cartographic and the ethnographic imagination: the whereabouts of amateurs, professionals, and nature in knowing biodiversity. Environment and Planning D 23 (5), 673–693.

Epstein, S., 1996. Impure Science: AIDS, Activism and the Politics of Knowledge. University of California Press, Berkeley.

Forestry Commission, 2011. Action Plan for Tree Health and Plant Biosecurity. http://www.forestry.gov.uk/pdf/Action-Plan-Tree-health-and-plant-biosecurity.pdf/$FILE/Action-Plan-Tree-health-and-plant-biosecurity.pdf (accessed 29.11.13.).

Greenwood, J., 2007. Citizens, science and bird conservations. Journal of Ornithology 148 (Suppl. 1), S77–S124.

Hemment, D., Ellis, R., Wynne, B., 2011. Participatory mass observation and citizen science. Leonardo 44 (1), 62–63.

Irwin, A., 1995. Citizen Science. Routledge, London.

Kisby, B., 2010. The big society: power to the people? The Political Quarterly 81 (4), 484–491.

Leydesdorff, L., Ward, J., 2005. Science shops: a kaleidoscope of science-society collaborations in Europe. Public Understanding of Science 14 (4), 353–372.

Miller-Rushing, A., Primack, R., Bonney, R., 2012. The history of public participation in ecological research. Frontiers in Ecology and the Environment 10 (6), 285–290.

Perlin, R., August 16, 2012. Indentures Are a Sign of Universities' Decline into a Factory System. Times Higher Education. http://www.timeshighereducation.co.uk/420883.article (accesed 29.11.13.).

Prainsack, B., 2014. Understanding participation: the 'citizen science' of genetics. In: Prainsack, B., Wener-Felmayer, G., Schicktanz, G. (Eds.), Genetics as Social Practice. Ashgate, Farnham, UK.

Raddick, M.J., Szalay, A., Vandenberg, J., Bracey, G., Gay, P.L., Lintott, C.J., Murray, P., Schawinski, K., 2010. Galaxy Zoo: exploring the motivations of citizen science volunteers. Astronomy Education Review 9 (1), 010103–1.

Riesch, H., Potter, C., 2014. Citizen science as seen by scientists: methodological, epistemological and ethical dimensions. Public Understanding of Science 23 (1), 107–120.

Riesch, H., Potter, C., Davies, L., 2013. Combining citizen science and public engagement: the open air laboratories programme. Journal of Science Communication 12 (3), A03.

Rotman, D., Preece, J., Hammock, J., Procita, K., Hansen, D., Parr, C., Lewis, D., Jacobs, D., 2012. Dynamic changes in motivation in collaborative citizen-science projects. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work. ACM, pp. 217–226.

Roy, H.E., Pocock, M.J.O., Preston, C.D., Roy, D.B., Savage, J., Tweddle, J.C., Robinson, L.D., 2012. Understanding Citizen Science and Environmental Monitoring: Final Report on Behalf of UK Environmental Observation Framework. NERC Centre for Ecology & Hydrology and Natural History Museum.

Shirk, J.L., Ballard, H.L., Wilderman, C.C., Phillips, T., Wiggins, A., Jordan, R., McCallie, E., Minarchek, M., Lewenstein, B.V., Krasny, M.E., Bonney, R., 2012. Public participation in scientific research: a framework for deliberate design. Ecology and Society 17 (2), 29. http://www.ecologyandsociety.org/vol17/iss2/art29/ (accessed 02.06.14.).

Silvertown, J., 2009. A new dawn for citizen science. Trends in Ecology & Evolution 24 (9), 467–471.

Stilgoe, J., 2009. Citizen Scientists: Reconnecting Science with Civil Society. Demos, London.

Toogood, M., 2011. Modern observations: new ornithology and the science of ourselves, 1920–1940. Journal of Historical Geography 37 (3), 348–357.

Toogood, M., Everett, G., 2013. Engaging publics: biodiversity data collection and the geographies of citizen science. Geography Compass 7 (9), 611–622.

Trumbull, D., Bonney, R., Bascom, D., Cabral, A., 2000. Thinking scientifically during participation in a citizen-science project. Science Education 84 (2), 265–275.

Tuss, P., 1996. From student to scientist: an experiential approach to science education. Science Communication 17 (4), 443–481.

Wiggins, A., Crowston, K., 2011. From conservation to crowdsourcing: a typology of citizen science. In: Proceedings of the 44th Hawaii International Conference on System Sciences. IEEE Computer Science Society, Washington, DC, pp. 2764–2773.

Relevant Websites

http://birds.audubon.org/ – Audubon Society. 2013.
http://www.citizensciencealliance.org/ – Citizen Science Alliance. 2013.
http://www.birds.cornell.edu/citscitoolkit – Citizen Science Central. 2013.
http://www.ucl.ac.uk/excites – ExCiteS. 2013. Extreme Citizen Science: ExCiteS.

Citizenship, Anthropology of

Ayse Caglar, University of Vienna, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article locates anthropology's arrival to citizenship studies and explores the peculiarities of anthropological approaches to citizenship, their contributions to theorization of citizenship as well as their limitations. It argues that the practice-based and processual understandings of citizenship in anthropology put a strong emphasis on the shifting meanings, everyday practices, and the varying and unpredictable sites of citizenship, which in turn makes ethnographic research central for citizenship studies. Finally, it situates anthropology's valorization of 'citizenship from below' within the context of neoliberal governmentality and argues that without 'nonlocal ethnography' anthropology's contribution to citizenship studies remain limited.

Introduction

Citizenship studies always had a solid place in political science and legal studies. However, since the 1980s, there has been a flourishing interest in the concept of citizenship and citizenship arrangements beyond their classical disciplinary bases. Anthropology has been a latecomer to the studies of citizenship. In this article, instead of providing an overview of the vast literature of citizenship studies, I will focus on anthropology's arrival to citizenship studies, on the peculiarities of anthropological approaches to citizenship, and on their contribution to the conceptualization of citizenship as well as on some of their limitations.

The conventional liberal understandings of citizenship are state-centric. They formulate citizenship as a form of (legal) relation between the individual subject and the (nation-) state, as well as a relationship of membership and belonging to a political community, which establishes the ground for the rights and responsibilities. Here, nation is the envisaged political community. The conventional approach to citizenship is a universalized and an abstract one, where legal status and the state are central to its conceptualization. Contrary to sociology, nation-state never loomed large in anthropology, and citizenship was not a central focus of interest for anthropology until the unmooring of citizenship studies from the state-centered frames of liberal and normative approaches. The more scholars documented the disaggregation of citizenship and citizenship rights from the nation-state, the more scholars started to question the conventional conceptions of citizenship. They challenged the hegemonic conceptualizations, which understood citizenship primarily as an issue of formal, legal status. They highlighted those cases which revealed that the nation-state cannot be taken for granted as the source of legitimacy and seedbed of individuals' rights. Several scholars drew attention to the fact that despite the lack of the formal, legal status of citizenship, people within the nation-state enjoyed social, economic, and cultural citizenship rights legitimated by transnational, supranational, and human rights regimes (Soysal, 1994; Bauböck, 1994). Similarly, they noted that citizenship was variegated within the nation-state: not all people who are legal citizens receive the same treatment from the state or are able to claim the same rights.

There are often categories of people who are legal citizens according to the laws of a state yet face various forms of exclusions and the denial of civil rights because they are not considered to be truly part of the nation. These categories include members of lower social classes – both rural and urban – and persons who are racialized or gendered in ways that put them outside the body politic (Marshall, 1964; Yuval-Davis, 1997). These distinctions between legal status and rights illustrated that legal status neither guaranteed the exercise and practice of rights for all those who had this legal status nor that these questions of inclusions and exclusions were static.

In the context of increasing salience of frames and practices of dual, multiple, and transnational citizenship (Goldring, 1998; Glick Schiller, 2005; Bauböck, 1994; Caglar, 2004; Levitt, 2001), the failure of those universalized perspectives of citizenship (which were developed at a time of the hegemony of nation-state in particular parts of the world) to capture the current practices of claim making, participation, and membership became more apparent. It became increasingly more difficult to operate with a concept of citizenship based on a historically constructed isomorphism between nationality and citizenship and which thus disregards the salience of different scales that function as the foundation of the political community upon which the source of rights and membership are grounded (Neveu, 2005; Clarke et al., 2014; Holston, 2008; Sassen, 2001).

Simultaneously, there was an increasing proliferation of forms and sites of citizenship such as 'environmental,' 'sexual,' 'indigenous,' 'post-national,' 'cosmopolitan,' 'multicultural,' 'transnational,' 'intimate,' 'diasporic,' and 'urban.' This pluralization also drew attention to the fact that there were different kinds of communities other than nations upon which claims were based. Consequently, the taken-for-granted connections between rights and states, the static nature of rights, seen as if they were possessions, as well as the universalizing conceptions of citizenship came under scrutiny. Rather than a static view of citizenship, which is inscribed into formal laws, a processual understanding of citizenship with an emphasis on its 'unfinished' nature came to the fore (Balibar, 2003). Anthropological approaches to citizenship followed these perspectives, which underlined the contextual, processual, practice-based, and contested character of citizenship. Ethnographic studies

focused on the citizenship practices and relations in concrete places and times, highlighted the relational, agency-based, and the contentious character of the citizenship within a wider array of mainly state actors. Contrary to the universalizing and normative understandings of citizenship in political sciences, which approached citizenship in a kind of 'empirical void' (Conover et al., 1991) anthropologists' work documented the practices and sites of citizenship and the varying and contested meanings citizenship took in diverse but concrete contexts. Anthropological studies of citizenship, with their emphasis on the diversity of meanings citizenship takes in different contexts contributed to the further "unlocking" of "citizenship from its focus on it as a legal/juridical and political status as corpus of rights and obligations and responsibilities…" (Clarke et al., 2014: 12).

Nation-State Sovereignty, Citizenship, and Migrants

Several scholars noted the disruption of territorial and judicial authority in nation-states in the context of globalization and addressed the contradictions of the nation-state sovereignty-centered citizenship models in our world (This section partly draws on Glick Schiller and Caglar 2008; 332–336). They draw attention to the increasing number of people who live in overlapping or multiple jurisdictions or to people who exercise social, economic, and cultural rights despite the absence of legal citizenship (Soysal, 1994; Sassen, 2001; Benhabib, 2007; Ong, 1999). The meaningfulness of nation-state scale as the suitable frame to capture the current citizenship arrangements is put into question. The weakening of the grip of state-centric citizenship models facilitated the arrival of practice-based processual approaches of anthropology in citizenship studies. Within this context, migrants as groups of people who exemplify the decoupling of rights and legal status as well as those subject to multiple jurisdictions become an important category of people to think with the changing dynamics of state-sovereignty and citizenship.

Consequently, the decoupling of formal rights which people possess as legal citizens and the social and cultural rights people practice without the legal citizenry became an important arena to discuss the problems with the nation-state and legal rights-centric approaches to citizenship (Soysal, 1994; Holston, 2008). Although legal citizens are often denied full civil rights, many states grant a range of rights to those who reside legally in a state but do not have legal citizenship and may even be ineligible to become citizens. As Soysal (1994) has emphasized, although they lacked formal membership in the host nation, migrants and their descendants often have been accorded various social, economic, civil, and cultural rights. The access accorded to these migrants has revealed the increasing disjuncture of formal and substantive citizenship rights and according to her, the sources of these substantive rights are anchored in human rights regimes rather than the nation-state.

Benhabib (2007), on the other hand, traces the source of these discrepancies and the disjuncture between territoriality and jurisdiction to cosmopolitan laws.

Ong (1999) analyzes the varying jurisdictions and the flexibilization of citizenship as part of the requirements of a particular kind of capitalism in the context of global economy. Like others, she also addresses the inroads made into state-sovereignty, but she does not see these heterogeneous political spaces, which are subject to differential regimes of rights and obligations ('graduated sovereignties'), as the indicators of the weakening of state-sovereignty. On the contrary, she approaches them as part of strategies of governmentality through which the state reconfigures its power in the context of global economy. According to her, destabilization and flexibilization of citizenship by deploying different citizenship regimes in the same state territory are part of the changing relations between nation-state and the cultural logics of capitalism. Moreover, this is not a weakness but a strategy for the disciplining of citizen-subjects.

In order to capture the social dynamics of participation and membership of those who do not have legal citizenry, anthropologists of citizenship increasingly have examined the practices and performance of citizenship rights. These discussions have enmeshed with anthropologies of migration, which focused on migrants' practices of social and political participation and membership. They have also explored migrants' vernacular notions of citizenship with a particular emphasis on the local, informal, and the community-based nature of their practices of political participation.

Moreover, scholars have expanded the concept of citizenship by distinguishing between cultural and social citizenships. Rosaldo and Flores (1997: 57) refer to cultural citizenship as "the right to be different with respect to the norms of the dominant national community, without compromising one's right to belong in the sense of participation in the nation-state."

However, cultural citizenship focuses on identities and the diversity of cultural practices, but it does not address the way in which persons who are not citizens participate in the common social, economic, and political life of a specific state and claim rights in these multiple domains. In order to address these practices and claims that facilitate the daily participation in the social life of a locality and generate claims and assertions of belonging that are not confined to the politics of difference and the cosmology of identity politics, scholars have begun to speak of social citizenship.

Social citizenship differs from legal citizenship because of the lack of mesh between formal citizenship and the allocation of rights, benefits, and privileges. When people without official membership make claims to belong to a state or to a city through collectively organizing themselves in order to protect themselves against discrimination, or receive rights and benefits from a state or local authorities, or make contributions to the development of the place of settlement and the life of people within it, they are said to be social citizens. Social citizens assert rights to citizenship substantively through social practice and social presence rather than law (Glick Schiller and Caglar, 2008). This approach to citizenship, which locates citizenship in horizontal social relations is concerned with the performative dimensions of membership beyond the domain of legal rights, and defines the meaning and practices of participation and belonging as it is displayed within the public sphere (Holston and Appadurai, 1999).

Moreover, the emphasis in social citizenship is on the horizontal ties between the people rather than the vertical ties between the individual and the state as formulated in the

formal citizenship (Caglar, 2004). In social citizenship, citizenship practices are located in social relationships. Although social citizenship scholarship operates with an understanding of the social that has been conceptualized almost solely within the conceptual network of social rights that facilitate the full participation of citizens and noncitizens of a polity, some scholars have been drawing attention to the necessity of recasting the meaning of the 'social' in citizenship. Isin (2008) and Brodie (2008) plea for using the broader meaning of the word social that captures its Latin origin *socius*, which refers to fellowship and companionship. In this sense social citizenship refers to "the dense fabric of relations that constitute humans as social beings who coexist with their fellow beings in and through conflict and cooperation that undergird norms, laws, and customs" (Isin, 2008: 282). This is an important point of entry to address sociability and conditions of coevalness and contemporaneity among the members of society beyond legal status.

Many people claim rights in a state and act within its institutional and governmental processes not only without formal citizenship but also while they reside in another state. 'Transnational' or 'transborder' citizenship is used to denote peoples' (mostly migrants') simultaneous participation in citizenship practices within and across borders of nation-states (Aleinikoff and Klusmeyer, 2001; Bauböck, 1994; Faist, 2000; Caglar, 2004; Glick Schiller, 2005; Goldring, 1998). Underlying these transborder practices is some type of claim to membership, rights, and voice in more than one state.

Most of the research and discussion of multiple membership claims and transborder citizenship has focused on migrants' social citizenship practices that connect a new land of settlement and a homeland. In these instances, migrants who become citizens of a new state, continue to influence their homeland, even if they have abandoned their legal citizenship rights. (Caglar, 2004; Levitt, 2001; Glick Schiller, 2005). Migrants who live their lives across borders may claim rights and membership in places where they do not have legal citizenship or in places where they no longer hold full legal citizenship.

These practices that are not necessarily based on formal arrangements of citizenship operate at the disjuncture between legal and substantive citizenship. However, these transnational citizenship practices are shaped in interaction with a number of actors, ranging from state to supranational and global institutions (Goldring, 1998; Caglar, 2004).

Although human rights regimes provide an important basis for social citizenship claims, some others make such claims of social inclusion, participation, and sociability in the name of religion. Religious claims and networks can foster social citizenship. In these instances migrants claim rights to residency, social benefits, respect, and moral and political leadership on the basis of their membership in a global religious project and the social field created by its networks of organizations (Levitt, 2003; Glick Schiller and Caglar, 2008). Scholars who examine migrant (social and political) participation that are anchored at religious organizations, homeland associations, or at seemingly apolitical sites like dance/music associations approach these sites as the new spaces of citizenship from below that function beyond the formal venues of citizenship (Reed-Danahay and Brettell, 2008; Hamidi, 2003).

Citizenship, Alterity and Margins

Citizenship provides the ground to distinguish between citizens and aliens. Scholars of citizenship note that margins and alterity are crucial in the constitutions of citizenship and they argue that the excluded, their practices, and their modes of exclusion are central to the constitution of citizenship. Citizenship creates and is created by its alterity (Isin, 2002). Consequently, within this relational understanding of citizenship and alterity, those who fall into the category of excluded, that is the undocumented, refugees, and the migrants come to the fore as the major constituents of citizenship (Balibar, 2003; Sassen, 1998; Isin, 2002; Holston, 2008; Varsanyi, 2006; Rajaram and Arendas, 2013; Caglar and Mehling, 2013). Building upon this understanding of citizenship, where margins are constitutive of the whole, Clarke et al. (2014: 49) argue for "recentering of citizenship around spaces, sites, and practices that are often described as marginal." Scholars argue that citizenship is defined at its margins because the inclusion of the excluded leads to a rearrangement of the existing structures and to a disruption of the taken-for-granted assumptions upon which the hegemonic understanding of citizenship at that place is based. When the excluded, like the undocumented, refugees, or the poor claim or demand their rights to be included as valid members of the society through informal, unofficial, and/or illegal venues, scholars refer to these practices as 'insurgent citizenship' (Holston, 2008; Varsanyi, 2006). The insurgency of such practices lies in their transformative capacity of social body as well as body of politics, as the granting of such rights to the excluded results in the formation of new categories, laws, and political subjects, in short, in the remaking of citizenship (Clarke et al., 2014: 171).

Isin (2008) refers to such insurgencies as 'Acts of Citizenship.' This is a concept that allows us to capture the moments when subjects constitute themselves as those making claims, asserting and imposing rights and obligations beyond the ones granted to them by law (2008: 5–6). According to him, acts of citizenship 'contest' habitus and citizenship practices by grounding their legitimacy and entitlements not in the existing legal and social frames but in unfamiliar or new grounds, and they introduce rupture to the given order. By instantiating a crack in the institutionalized order of things they become part of contentious politics. Within this perspective it becomes possible to approach sites as being constitutive of the acts of citizenship and explore the impact of the materiality of sites in the formation of such acts (Caglar and Mehling, 2013).

Anthropology and 'Citizenship from Below'

The common ground of anthropological approaches to citizenship is their emphasis on 'politics from below' and on the venues of participation and rights beyond the official forms and state-endorsed rights (Nuijten, 2013: 11). In summary, what differentiates anthropology's take on citizenship from other disciplines is an approach to citizenship as a set of practices mostly located in everyday life; as a process; as belonging and membership anchored at a political community beyond the national scale; as webs of meaning; as horizontal

relations located in social relations with the others and as subjectivity.

All these indicate a broader understanding of citizenship. Rather than seeing it as a relation to a state, where the exercise of political rights derived from the legal status are central, anthropologists view citizenship as a set of practices that enable the participation in social and political life and as a venue for sociability. Here, the emphasis is on the people's everyday experiences of citizenship and on the formation of their subjectivities and their political agency through their claim-making processes within a particular historical and political context (Neveu, 2005; Nuijten, 2013). This contextual and processual understanding of citizenship, which is always in the making, is very different from the static understanding of a liberal conceptualization of citizenship where the rights are seen to be possessions of the individual subjects that are endorsed by the state (Neveu, 2005). However, the most differentiating aspect of anthropological understanding of citizenship lies in the scale of the political community, which establishes the contingency of political membership, belonging, as well as the ground through which the subjectivities are constructed. Rather than equating the political community with the nation, anthropological studies of citizenship draw attention to the different kinds of communities upon which claims are based and membership and belongings are forged. These could be city-based communities, they could be even trade unions as Lazar shows in her study on citizenship and political agency in Buenos Aires, but they could also be border-crossing cosmopolitan or transnational communities (Lazar, 2013; Lazar and Nuijten, 2013). In her study of slum upgrading projects in Brazil, Nuijten (2013) shows how subjectivities, political belonging, and the political community are mutually constituted through the slum dwellers' claim-making processes that are entangled with the creation of a new landscape of empowerment and disempowerment. All these studies underline how citizenship is reworked in daily practices and discourses.

Most importantly, anthropological studies of citizenship put a strong emphasis on the language, vernacular, and the repertoires of citizenship. The language of rights and responsibilities vary. What people understand from rights and responsibilities at a particular place, and the meanings people give to their relationships with the state and with the bureaucracy are central to anthropological studies of citizenship practices (Ouroussoff and Toren, 2005; Stack, 2013). As Ouroussoff and Toren (2005: 209) formulate "[F]for us as anthropologists, the issue is not what model of citizenship we should endorse, but rather what ethnography can do to analyze how these key categories – citizenship and culture – are being constituted anew (and in this process transformed) in the practice of their everyday lives by particular people(s) in particular times and places. The challenge of contemporary ethnography is to render analytical categories people use to talk about themselves, their lives, and their ideas of the world."

The heterogeneity of citizenship's sites, settings, spaces, and scales, which the anthropological studies of citizenship document, shows us that anthropological studies underline the fact that sites and spaces of citizenship and politics could not be known a priori as they vary in space and time. Designating these sites is an empirical question and requires a contextualized analysis to explore in which specific sites, citizenship acts are anchored and in what specific ways. As Clarke et al. (2014: 5) aptly underline "citizenship always takes specific forms that are outcomes of sets of processes and are related to specific political projects, particular social contexts and distinctive cultural configurations." Within this framework, ethnographic research becomes central to the analysis of citizenship practices.

In short, anthropologists approach citizenship as a process from below, thus as a grassroot political process. Given the anthropologists' emphasis on the primacy of everyday interactions for the constitution of individuals' subjectivities, 'everyday citizenship' comes to the forefront as a crucial aspect of anthropological approaches to citizenship (Holston, 2008). Citizenship from below, which characterizes the anthropological approaches to citizenship, then refers to the experienced, lived citizenship that expresses the everyday realities of political belonging (Nuijten, 2013: 11).

Citizenship and Cities

Several scholars who approach citizenship beyond the state-centered liberal model argue that, in the context of globalization, cities emerge as the site and the context in which residents constitute their claims and entitlements and the meaning and practice of citizenship. Cities, rather than nation-states become the axis and the locus for political belonging and membership. Thus they replace states as the primary reference point in citizenship practices. Rather than the legal status, they argue, residence is pivotal for rights and entitlements (Sassen, 2001; Bauböck, 1994; Benhabib, 2007; Varsanyi, 2006). Several ethnographic works on citizenship followed suit. However, no matter how cities become crucial units of governance and strategic sites for generation of wealth and political operations, in the context of neoliberal globalization it is important to approach cities not as entities having a life of their own, but situate the dynamics of urban economies, politics, and culture within global networks of power hierarchies. Especially in the neoliberal context of entrepreneurial cities and governmentality, the valorization of cities as alternative grounds of citizenship practices and entitlement in counterpoint to state-based arrangements might require caution. In reference to the workings of the neoliberal state, Rose and Osborne (2000: 108) draw attention to new forms of governmentality through 'citizenship games.' These games refer to practices, which constitute its players; that is, the citizens with a particular kind of conduct, so that citizens by playing their own parts become part of the games that govern them.

In the context of the neoliberal state where the state shares its powers of control and discipline with other actors and agencies, Rose and Osborne argue that "[C]itizenship is no longer primarily realized in relation with the state," instead through "a set of dispersed and non-totalized practices within which games of citizenship must be played" (2000: 108). Moreover, moving away from an understanding of citizenship as possession to citizenship as capacity is part of the new forms of governance of the neoliberal state (2000: 108). In the context of enabling and constraining the workings of the

neoliberal state, which functions through the activation of powers and agency of citizens, the distinction between state and civil society as well as the clear distinction between 'citizenship from above' and 'citizenship from below' become blurred. Furthermore, communities are the new ethical spaces through which new forms of governmentality operate (Rose and Osborne, 2000). Thus, neoliberalism revives communities. This kind of framing of the workings of the neoliberal state, cities, and governmentality enables us to situate the valorization of 'citizenship from below' and communities in anthropological approaches to citizenship anew.

It is true that anthropologists of citizenship played an important role in moving discussions of citizenship from the normative and the institutional straitjacket of universalized models and put the processual and relational character of citizenship at the center of their studies. However, in the context of neoliberal valorization of imaginaries and practices of community livelihood, anthropologists' emphasis on the local practices of communities as the primary sites that facilitate participation and empowerment, and as sites of citizenship practices and sociabilities becomes problematic. Several works on neoliberal cities have illustrated that the recreation of communitarian obligations and allegiances are crucial for neoliberal governance and especially for that of cities. From this point of view, cherishing communities as sites of mutuality and sociability takes a different meaning once we contextualize such approaches within the novel forms of neoliberal governmentality and governance of cities.

In Conclusion: Everyday Citizenship and Nonlocal Ethnographies

Like all subjectivities, citizenship is also mediated. In order not to be carried away by an over-fluid understanding of citizenship, it is important to refer to the structures and institutions within which subjects are constructed and within which this mediation takes place (Clarke et al., 2014). It is crucial to focus on the meanings of citizenship and the citizenship language, through which demands are articulated and claims are made. However, these meanings are enabled within the constraints of a set of legal, political, and institutional conditions rather than being the outcome of interplay between citizenship and culture. Moreover, the conditions of possibility of meanings and practice cannot be grasped when we remain confined to the descriptions of the everyday practices.

In order to underline the mobilizations and the shifting meanings of citizenship and its plastic capacity of articulation to varying political projects in different places, Clarke et al. (2014: 6) designate citizenship as a keyword in the sense of Williams (1983). By identifying it as a keyword, among other things, they underline the constitutive, productive, and the performative force of citizenship languages and their mobilizing power for social and political action. In fact, citizenship languages, with an imagery and a repertoire, closely tied to human rights discourses very often find their place in the formulation of demands and consequently in the mobilizations construed through these claim-making processes (Nuijten, 2013). However, no matter whether or not their meanings shift in their local use, citizenship languages and the repertoire of human rights are part of global processes of claim making that can neither be confined to that particular locality nor be captured as part of general flows of ideas and discourses. Citizenship languages and repertoires are mediated within a set of institutional and discursive networks and this process cannot be fully grasped through ethnographies of everyday practices.

In order to study the global processes in which relations are mediated between actors that have not been in direct contact with each other, Feldman (2011) argues for a nonlocal ethnography that will go beyond the participant observation bias of ethnography, which is often anchored in the valorization of everyday life. He differentiates connections and relations. While the former involves immediate direct contact, the latter refers to indirect mediated contact. "Nonlocal ethnography" deals with relations and "addresses how people are specifically related when they are not connected in global configurations of power" (Feldman, 2011: 380). This is a similar search for an ethnographic entry to study global dynamics relationally, so that we can reveal, expose, and explain the connections between faraway yet intimately related developments and integrate them into the core of anthropologists' work (Miraftab, 2014). However, in order to see "how seemingly independent processes and locations are intimately connected with each other," we need to situate them into hierarchically structured relationships (Miraftab, 2014). This requires an analysis of the processes and relationship that shape the power structures connecting these places, which urges us to go beyond ethnographies located in everyday life and confined to participant observation. It is true that with its attention to personal narrative and to the contextualized enactments of citizen practices in everyday life, ethnography offers an important entry into the analysis of citizenship practices, their shifting meanings, and the formation of subjectivities beyond the legal grid of citizenship models. It is also true that by exploring the agency-based dynamics of everyday citizenship, anthropological studies on citizenship provide an important intervention to the taken-for-granted scale of liberal understandings of citizenship. However, without a nonlocal ethnography that would allow us to relate the dynamics between everyday citizenship practices and networked institutions of power, such interventions of anthropologies of citizenship fail to grow into their explosive capacities.

See also: Citizenship, Historical Development of; Citizenship, Sociological Aspects of; Immigration, Citizenship, and Integration: Social Work Connections; Immigration, Emigration, and Citizenship; International Migration; Migration History; Migration and Social Capital; Transnational Citizenship.

Bibliography

Aleinikoff, A., Klusmeyer, D.B. (Eds.), 2001. Citizenship Today: Global Perspectives and Practices. Carnegie Endowment for International Peace, Washington.

Balibar, E., 2003. We, the People of Europe? Reflections on Transnational Citizenship. Princeton University Press, Princeton.

Bauböck, R., 1994. Transnational Citizenship: Membership and Rights in International Migration. Edward Elgar Publishing, Aldershot.

Benhabib, S., 2007. Twilight of sovereignty or the emergence of cosmopolitan norms? Rethinking citizenship in volatile times. Citizenship Studies 11 (1), 19–36.

Brodie, Janine, 2008. The social in social citizenship. In: Isin, E. (Ed.), Recasting the Social in Citizenship. Toronto University Press, Toronto, pp. 20–43.

Caglar, A., 2004. "Citizenship light": transnational ties, multiple rules of membership and the "pink card". In: Friedman, J., Randeria, S. (Eds.), Worlds on the Move. I.B. Tauris, London, pp. 273–292.

Caglar, A., Mehling, S., 2013. Sites and the scales of the law: third-country nationals and EU Roma citizens. In: Isin, E., Saward, M. (Eds.), Enacting European Citizenship. Cambridge University Press, New York, pp. 155–177.

Clarke, J., Coll, K., Dagnino, E., Neveu, C., 2014. Conclusion: disputing citizenship. In: Clarke, J., Coll, K., Dagnino, E., Neveu, C. (Eds.), Disputing Citizenship. Policy Press, Bristol, pp. 169–177.

Conover, P.G., Crewe, I.M., Searing, D., 1991. The nature of citizenship in the United States and Great Britain: empirical comments on theoretical themes. The Journal of Politics 53 (3), 800–832.

Faist, T., 2000. Transnationalization in international migration: implications for the study of citizenship and culture. Ethnic and Racial Studies 23 (2), 189–222.

Feldman, G., 2011. If ethnography is more than participant-observation, then relations are more than connections: the case for nonlocal ethnography in a world of apparatuses. Anthropological Theory 11, 375–395.

Glick Schiller, N., 2005. Transnational social fields and imperialism: bring a theory of power to transnational studies. Anthropological Theory 5 (4), 439–461.

Glick Schiller, N., Caglar, A., 2008. And ye shall possess it, and dwell therein: social citizenship, global Christianity, and non-ethnic immigrant incorporation. In: Reed-Danahay, D., Brettell, C.B. (Eds.), Citizenship, Political Engagement, and Belonging: Immigrants in Europe and the United States. Rutgers University Press, New Jersey, pp. 203–225.

Goldring, L., 1998. From Market Membership to Transnational Citizenship? The Changing Politicization of Transnational Social Spaces. Chicano/Latino Research Center, Merrill College, University of California, Santa Cruz.

Hamidi, C., 2003. Voluntary associations of migrants and politics: the case of North African immigrants in France. Immigrants & Minorities: Historical Studies in Ethnicity, Migration and Diaspora 22 (2–3), 317–332.

Holston, J., Appadurai, A., 1999. Introduction: cities and citizenship. In: Holston, J. (Ed.), Cities and Citizenship. Duke University Press, Durham, pp. 1–20.

Holston, J., 2008. Insurgent Citizenship: Disjunctions of Democracy and Modernity in Brazil. Princeton University Press, Princeton.

Isin, E., 2002. Being Political: Genealogies of Citizenship. University of Minnesota Press, Minneapolis.

Isin, E., 2008. Conclusion: the socius of citizenship. In: İsin, E. (Ed.), Recasting the Social in Citizenship. University of Toronto Press, Toronto, pp. 281–286.

Lazar, S., 2013. Citizenship, political agency and technologies of the self in Argentinean trade unions. Critique of Anthropology 33, 110–128.

Lazar, N., Nuijten, M., 2013. Citizenship, the self, and political agency. Critique of Anthropology 33, 3–7.

Levitt, P., 2001. Transnational Villagers. University of California Press, Berkeley.

Levitt, P., 2003. You know, Abraham was really the first immigrant: religion and transnational migration. International Migration Review 37 (143), 847–874.

Marshall, T.H., 1964. Class, Citizenship, and Social Class: Essays by T.H. Marshall. Doubleday and Co., Garden City, NY.

Miraftab, F., 2014. Displacement: framing the global relationally. In: Kahn, H. (Ed.), Framing the Global: Entry Points for the Search. Indiana University Press, Bloomington.

Nuijten, M., 2013. The perversity of the "Citizenship Game": slum-upgrading in the urban periphery of Recife, Brazil. Critique of Anthropology 33, 8–25.

Neveu, C., 2005. Discussion: anthropology and citizenship. Social Anthropology 13, 199–202.

Ong, A., 1999. Flexible Citizenship: The Cultural Logics of Transnationality. Duke University Press, Durham.

Ouroussoff, A., Toren, C., 2005. Discussion: anthropology and citizenship. Social Anthropology 13, 207–209.

Rajaram, P.K., Arendas, Z., 2013. Exceeding categories: law, bureaucracy and acts of citizenship by asylum seekers in Hungary. In: Isin, E., Saward, M. (Eds.), Enacting European Citizenship. Cambridge University Press, New York, pp. 195–219.

Reed-Danahay, D., Brettell, C.B. (Eds.), 2008. Citizenship, Political Engagement, and Belonging: Immigrants in Europe and the United States. Rutgers University Press, New Jersey.

Rosaldo, R., Flores, W., 1997. Identity, conflict, and evolving Latino communities: cultural citizenship in San Jose, California. In: Flores, W., Benmayor, R. (Eds.), Latino Cultural Citizenship: Claiming Identity, Space, and Rights. Beacon Press, Boston, pp. 57–96.

Rose, N., Osborne, T., 2000. Governing cities, governing citizens. In: Isin, E. (Ed.), Democracy, Citizenship and the City: Rights to the Global City. Routledge, London, pp. 95–109.

Sassen, S., 1998. Introduction: whose city is it? Globalization and the formation of new claims. In: Sassen, S. (Ed.), Globalization and Its Discontents. The New Press, New York, pp. xix–xxxvi.

Sassen, S., 2001. Cracked casings: notes towards an analytics for studying transnational process. In: Pries, L. (Ed.), New Transnational Spaces: International Migration and Transnational Companies in the Early Twenty-First Century. Routledge, London, pp. 187–208.

Soysal, Y.N., 1994. Limits of Citizenship: Migrants and Postnational Membership in Europe. University of Chicago Press, Chicago.

Stack, T., 2013. In the eyes of the law, in the eyes of society: a citizenship tradition in west Mexico. Critique of Anthropology 33, 66–90.

Varsanyi, M.W., 2006. Interrogating "urban citizenship" vis-à-vis undocumented migration. Citizenship Studies 10 (2), 229–249.

Williams, R., 1983[1976]. Keywords: A Vocabulary of Culture and Society, revisited ed. Oxford University Press, New York.

Yuval-Davis, N., 1997. Women, citizenship and difference. Feminist Review 57, 4–27.

Citizenship, Historical Development of

Richard Bellamy, London University, London, UK; and European University Institute, Florence, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Historically, the distinctive core of citizenship has been the possession of the formal status of membership of a political and legal entity and having particular sorts of rights and obligations within it. This core understanding of citizenship goes back to classical times and coalesced around two broad understandings of citizenship stemming from ancient Greece and Imperial Rome respectively that later evolved into what came to be termed the 'republican' and 'liberal' accounts of citizenship. This article first examines these two classic views, then looks at how they changed during the Renaissance and Reformation, and finally turns to the ways the two were to some extent brought together following the American and French Revolutions within the liberal democratic nation-state.

Historically, the distinctive core of citizenship – at least within the Western political tradition – has been the possession of the formal status of membership of a political and legal entity and having particular sorts of rights and obligations within it that distinguish one from being either a subject or a casual visitor, on the one hand, or performing some other noncivic social role, such as a friend, a neighbor, or a Good Samaritan, on the other. Different conceptions have offered different views as to what the criteria of membership should be; the nature of the political and legal institutions to which a citizen belongs; the content of their rights and duties; and the character of the norms and attitudes citizens require to exercise and fulfill these civic entitlements and obligations. However, all agree that citizenship is a political and legal artifact that creates a condition of civic equality among those who possess it with regard to the prerogatives and responsibilities it bestows and requires.

As J.G.A. Pocock (1995) has argued, this core understanding of citizenship goes back to classical times and coalesced around two broad understandings of citizenship stemming from ancient Greece and Imperial Rome respectively that later evolved into what came to be termed the 'republican' and 'liberal' accounts of citizenship. In what follows, we shall first examine these two classic views, then look at how they changed during the Renaissance and Reformation, and finally turn to the ways the two were to some extent brought together following the American and French revolutions within the liberal democratic nation-state.

The Two Classic Conceptions

The canonical text of the Greek version of citizenship is Aristotle's *Politics* (335–323 BC) with ancient Athens the model. Aristotle regarded human beings as 'political animals' because it is in our nature to live in political communities – indeed, he contended that only within a *polis* or city–state could human potential be fully realized. However, he believed people played the roles appropriate to their natural station in life, with only some qualifying as *politai* or citizens. Though neither the qualifications Aristotle deemed appropriate for membership of this select group nor the duties he expected of them are regarded as entirely suitable today, they have cast a long shadow over the history of citizenship and their inner rationale still underlies much contemporary thinking.

To be a citizen of Athens it was necessary to be a male aged 20 or over, of known genealogy as being born to an Athenian citizen family, a patriarch of a household, a warrior – possessing the arms and ability to fight, and a master of the labor of others – notably slaves (Finley, 1983). So gender, race, and class defined citizenship and many of the main later debates have turned on how far they continue to do so. As a result, large numbers were excluded: women, though married Athenian women were citizens for genealogical purposes; children; immigrants or 'metics' – including those whose families had been settled in Athens for several generations, although they were legally free, liable to taxation and had military duties; and above all slaves. It is reckoned that the number of citizens in Athens fluctuated between 30 000 and 50 000, while the number of slaves was of the order of 80–100 000. Therefore, citizenship was enjoyed by a minority, though a substantial one. Yet, this was inevitable given the high expectations of citizens. For, their capacity to perform their not inconsiderable citizenly duties rested on their everyday needs being looked after by the majority of the population, particularly women and slaves.

Aristotle described citizens as those who "rule and are ruled by turns" (Aristotle, 1988 [335–323 BC]: 1259b1). Though the duties involved differed between polities and even different categories of citizen within the same polity, at some level citizenship involved "the power to take part in the deliberative or judicial administration" (Aristotle, 1988 [335–323 BC]: 1275b1) In Athens this meant at a minimum participating in the assembly, which met at least 40 times a year and required a quorum of 6000 citizens for plenary sessions, and, for citizens aged over 30, doing jury service – again, a frequent responsibility given that juries required 201 or more members and on some occasions over 501. Though jury service was paid, jurors were chosen by lot from among those who presented themselves to discourage both its becoming a regular income and jury packing. In addition, there were some 140 local territorial units of government, or *demes*, with their own *agorai* or assembly points for public discussion of local affairs and passing local decrees.

Meanwhile, many citizens could not avoid holding public office at some point. Apart from generals, who were elected by the assembly and could serve multiple terms if successful, public officers were chosen by lot, served for 1 or 2 years maximum, with key roles often rotated between office holders. These devices aimed to increase the likelihood that all citizens

had an equal chance of exercising political power, with the short terms of office and the checks operated by the different bodies on each other ensuring this power was severely circumscribed. Yet, though there were no career politicians, citizenship itself, if one adds military service and participation in local affairs, was a fairly full time occupation.

Athens was unusual among Greek city-states in being so democratic. Indeed, Aristotle, who periodically resided in Athens but was not born there and so was not an Athenian citizen, expressed a personal preference for systems that mixed democracy with aristocratic and monarchical elements. However, even in those systems, citizenship remained fairly onerous. Like Plato, Aristotle esteemed the austere citizenship code of Sparta. By contrast to Athens, where the arts, philosophy, and leisure were much admired, Sparta emphasized military service above all else. Male children were separated from their families aged 7, subjected to a rigorous training, and thereafter attached to a 'mess.' Given they still had to attend the Assembly, Spartan citizens became even more permanent public servants than their Athenian counterparts. In fact, it was precisely their limited opportunities to develop private interests that Plato in particular so admired.

Aristotle acknowledged that such forms of citizenship were only possible in small states. That was important not just so everyone could have a turn at ruling and to keep the tasks of government sufficiently simple as to be manageable without a professional bureaucracy or political class, but also because it was only in smaller settings that the requisite civic virtues were likely to be fostered. Although the Athenians probably invented the idea of taking a vote to settle disagreements, unanimity was the ideal and most issues were settled by consensus – if need be following extended debate. Aristotle surmised that such concord or *homonoia* depended on a form of civic friendship among citizens that was only likely in tightly knit communities. Citizens must know each other, share values, and have common interests. Only then will they be able to agree on what qualities are best for given offices and select the right people for them, harmoniously resolve disputed rights, and adopt collective policies unanimously. Even so, agreement rested on citizens possessing a sense of justice, being temperate by exercising self-control and avoiding extremes, having a capacity for prudent judgment, being motivated by patriotism, so they put the public good above private advantage, and being courageous before danger, especially military threats. In sum, a citizen must not belong 'just to himself' but also to 'the *polis*.'

Though the Greek model of citizenship model was the privilege of a minority, it provided a considerable degree of popular control over government. True, the Assembly and Council tended to be dominated by the high born and wealthy, while Aristotle's ideal of concord was often far from the reality, at least in Athens. There were persistent tensions between different classes and factions, with disagreements often bitter and personal, ending with the physical removal of opponents through ostracism and even their execution on trumped up charges of treason. Nonetheless, in a very real sense those people who qualified as citizens did rule, thereby giving us the word democracy from the Greek *demokratia* or people (*demos*) rule (*kratos*). Unsurprisingly, Greek citizenship has appeared to many later thinkers as the epitome of a true condition of political equality, in which citizens have equal political powers and so must treat each other with equal concern and respect. They have viewed the trend toward delegating political tasks to a professional class of politicians and public administrators with foreboding, as presaging a loss of political freedom and equality, and lamented the – in their opinion – short-sighted tendency for ever more citizens to desert public service to pursue personal concerns. By contrast, critics of this model of citizenship argue that it was not so much an ideal as hopelessly idealized. In reality, it was doubly oppressive. On the one hand, it rested on the oppression of slaves, women, and other noncitizens. On the other hand, it was oppressive of citizens in demanding they sacrifice their private interests to service of the state. As we saw, the two forms of oppression were linked: citizens could only dedicate themselves to public life because their private lives were serviced by others.

Later liberal commentators have condemned these last features of Greek republican citizenship as potentially despotic (Constant, 1988 [1819]; Berlin, 1969). They criticize not just the way noncitizens got treated as less than fully human, but also the demand for the total identification of citizens with the state, with all dissent seen as indicative of self-interest rather than an alternative point of view or valid concern. They castigate such regimes as both repressive and corrupt – not least in diverting all talent away from the private sphere of the economy on which the wealth of a society rests. Ironically, making the public sphere the main avenue of personal advancement did not prevent but promoted the abuse of power for private again. They trace these problems to a flawed view of liberty that falsely links freedom with civic participation. Aristotle's defense of this linkage rested on a perfectionist account of human flourishing, with civic involvement a means to human self-realization, whereby individual and collective autonomy can be reconciled by subsuming private interests under the public interest. Many liberals reject such 'positive' conceptions of liberty as suggesting human freedom lies in the pursuit of particular ends. Instead, they advocate a 'negative' conception that consists of being free from interference to pursue one's personal good in one's own way. They claim freedom of this latter sort merely requires a just constitutional regime that limits the power of government to maximizing freedom from mutual interference and has no intrinsic link with democracy.

Imperial Rome offers an important contrast in this respect and one that forms part of the genealogy of the liberal view. Eligibility for Roman citizenship was at first similar to the criteria for Greek citizenship – citizens had to be native free men who were the legitimate sons of other native free men. As Rome expanded – initially within Italy, then over the rest of Europe and finally into Africa and Asia – two important innovations came about. First, the populations of conquered territories were given a version of Roman citizenship while being allowed to retain their own forms of government, including whatever citizenship status they offered. Second, the version of Roman citizenship given was of a legal rather than a political kind – *civitas sine suffragio* or 'citizenship without the vote.' So, the Empire allowed dual citizenship, though it reduced Roman citizenship to a legal status. As a result, the legal and political communities pulled apart. The scope of law went beyond political borders and did not need to be coextensive with a given territorial unit. To cite the famous case of St Paul – on

arrest in Palestine, he proudly declared himself "a Jew of Tarsus, a city in Cilicia, a citizen of no mean city." But not being in Tarsus, it was his additional status as a Roman citizen that allowed him to claim rights against arbitrary punishment, thereby escaping a whipping, and to ask for trial in Rome.

According to the Aristotelian ideal, political citizenship had depended on being freed from the burdens of economic and social life – both in order to participate and to ensure that public rather than private interests were the object of concern. By contrast, legal citizenship has private interests and their protection at its heart. Within Roman law, legal status belonged to the owners of property and, by extension, their possessions. Since these included slaves, a free person was one who owned himself. So conceived, as in many respects it remains to this day, law was about how we could use ourselves and our things and those of others, and the use they may make of us and our things. As the example of St Paul shows, the resulting privileges and immunities, including the right to sue and be sued in given courts, were far from trivial. However, that the rule of law can be detached from the rule of persons, in that those subject to it do not have to be involved in either its making or its administration, creates disadvantages as well as advantages. The advantage is that the legal community can, as we saw, encompass a number of political communities and hold their rulers and officers to account, thereby limiting their discretion to act against the law. Law can be universal in scope and extent, enabling millions of dispersed individuals to pursue their private interests by engaging and exchanging with each other across space and, through such legal acts as bequests, through time, without any direct contact. The disadvantage lies in these same citizens becoming the imperial subjects of the law's empire, who are ruled by it rather than ruling themselves. Yet the rule of law is only ever rule through law by some person or persons. Law can have many sources and enforcers, and different laws and legal systems will apply to different groups of persons and have differing costs and benefits for each of them. If law's empire depends on an emperor, then the danger is that law becomes a means for imperial rule rather than rule of and for the public.

Toward Republicanism and Liberalism

Both these conceptions underwent significant alterations over time in response to changing social and political circumstances and new intellectual preoccupations. As a result, they gradually became reconfigured on rather different ontological and epistemological assumptions. Of particular importance were the struggles between religious and secular political authorities, on the one hand, and, crosscutting this conflict, those between city-states and monarchical rulers, including the imperial pretensions of successive Holy Roman Emperors, on the other. These struggles both shaped and were shaped by political thinking from the Middle Ages to the Reformation. Two related developments emerging from this process were particularly significant: the separation of religion and politics, and the crystallization of notions of political sovereignty – be it of the people or their rulers – in the context of a polity possessing the features of a state: namely, a monopoly of coercive power over those residing within its territorial boundaries. For example, each of them plays a key role in Marsilius of Padua's important tract *Defensor Pacis* (1324). This work draws on both the political and democratic and the legal and imperial conceptions of citizenship explored in the last section, adapting them to the context created by the two sets of struggles mentioned above to ground the legal sovereignty of the Emperor Louis of Bavaria against the claims of Pope John XXII in a doctrine of popular sovereignty.

It is against this background that we need to explore the significant reworkings of the two classic conceptions of citizenship brought about by Machiavelli and Hobbes, respectively. Both offer secular accounts of citizenship that link popular sovereignty to the right to rule of those possessing sovereign political authority. Yet its role within their respective accounts was very different. While Machiavelli appreciated the part laws can play in securing the rights and liberties of citizens under a monarchical regime, he contended liberty was only fully secured within a republic where the people exercised political power in ways that allowed them to mutually check each other, thereby obliging all citizens to act collaboratively and in the public good. By contrast, though Hobbes grounded the right to rule in the mutual consent of the people, he maintained that in the process the ruled surrendered their sovereign power to their rulers. Moreover, he regarded the unlimited and undivided sovereignty of an absolute ruler – preferably, though not necessarily, a monarch – as offering the surest basis for law and the protection of liberty.

Machiavelli's account draws on the Roman republican model of citizenship more than the Greek model associated with Aristotle, explored above. While there are some similarities between the two, there are also striking differences. Though classes existed in Greek society, including among those who qualified as citizens, the ideal of citizenship became classless with the aspiration to 'concord' a product of putting class and other private interests to one side. Instead, the Roman Republic was born of class discord and the struggle of the plebeians to obtain rights against the patricians. For the theorists of the Roman model – Cicero (1991 [44 BC]), the historians of the Roman republic and, drawing on them, Machiavelli (1970 [1531]) – this ongoing class conflict gave politics and citizenship a much more instrumental character than the Greek model theorized by Aristotle. Roman citizens never possessed anything like the political influence of their Athenian counterparts. Despite the creation of Tribunes of the People, elected by a Plebeian Council, true power rested with the Senate. While entry to the Senate ceased to depend on rank around 400 BC, since it was composed instead of the popularly elected magistrates, it was dominated by the patricians – especially among the higher magistracy, particularly the Consuls who formed the executive. The slogan *Senatus Populusque Romanus* ('The Senate and the Roman People,' frequently abbreviated to SPQR) suggested a partnership between the Senate and the people within the popular assemblies. In reality, Senate and people were always in tension, with the influence of the plebeians waxing and waning depending on their importance as support for different factions among the patricians.

Applying these ideas to Renaissance Florence, Machiavelli argued the Roman experience showed how the selfish interests of the aristocracy and the people could only be restrained if each could counter the other. The republic institutionalized such mutual restraint by ensuring no person or institution

could exercise power except in combination with at least one other person or institution, so each could check and balance the other. The need to divide power in this way was elaborated by later republican theorists. It was a key feature of the city-states of Renaissance Italy, especially Florence and Venice, which inspired Machiavelli's writings on the subject, and influenced the political arrangements of the Dutch Republic into the eighteenth century. Republican ideas also informed the constitutional debates of the English Civil War of the seventeenth century, influencing writers such as Milton and Harrington. In the work of Montesquieu and, following him, the American *Federalists*, especially Madison, the check and balance of powers became a central element of the US Constitution (Hamilton et al., 2003 [1787–1788]).

Underlying this account was a distinctively realist view of citizenship, which could be more easily adapted to modern democratic politics than the Greek view. Instead of viewing the private interest and the public interest as diametrically opposed, so that all elements of the first had to be removed from politics, the public interest emerged from the clash and balancing of private interests. Consequently, citizens had self-interested reasons to participate because they could only ensure that their concerns figured in any collective decisions so long as they took part and were counted. Quentin Skinner (1998) and Philip Pettit (1997) have argued that the neo-Roman version of republicanism rejects the 'positive,' Aristotelian view of liberty as self-mastery for a 'negative' account of freedom as the absence of domination or mastery by another. Citizens need not identify their will with that of the polity; merely seek to ensure that government and the laws address the interests of all in an equitable manner through being obliged to 'hear the other side.' Liberty results from a political system where none are the masters of others because all have an equal influence over how public policies are framed and implemented.

Once again, this republican argument can be contrasted with the liberal notion of liberty as freedom from interference. As we noted, this position has its origins in the Imperial view of citizenship centered on providing the legal protection of an individual's civil liberties, especially the right to property. Much as Machiavelli can be regarded as defending the link between political citizenship and liberty on a new basis, Hobbes can be read as criticizing this linkage and putting forward a new defense of the link between legal citizenship and liberty. In framing his argument, Hobbes drew on the contemporary natural law tradition in which individuals are conceived as proprietors of themselves and the world, possessing rights in both for their self-preservation. Although Hobbes regarded humans as capable of perceiving and pursuing laws of nature, he did not believe these precepts allowed them to live peacefully without government. Infamously, he depicted the state of nature as a war of all against all. He ascribed this condition to each person being judge, jury, and executioner in their own case and acting on their own private judgment – itself a product of their exercising their 'right of nature' to do anything which they judged necessary to their preservation. As a result, in the natural state people would live in a permanent condition of insecurity in which neither industry nor any of the activities associated with civilization would be possible. Hobbes contended each person ought to be rational enough to see this state was not conducive to their ability to safely pursue their interests safely and to perform those actions most conducive to peace – a set of practical imperatives he terms the Laws of Nature. The solution was for individuals to lay down a part of their right to all things and to establish an absolute sovereign with sufficient power to hold all in awe. Hobbes contended the passage from the natural to the civil state could be regarded as the product of a mutual covenant or social contract between the members of society, whereby they ceded their right to private judgment concerning those matters most conducive to their preservation to a sovereign political authority. So long as these sovereign authorities could offer effective protection to those subject to their rule, the ruled had an obligation to obey their commands. Nevertheless, the ruled retain a right to self-defense that allows them to resist a sovereign that puts their lives in danger.

Hobbes sketched much of this argument in a book published in Latin in 1642 with the title *De Cive*, or 'On the Citizen,' although it got its most famous statement in the English version of the *Leviathan* published in 1651. From the republican perspective, his argument that citizenship was secured through subjection to an absolute Leviathan seems like a contradiction in terms. However, Hobbes takes issue with this view. Writing in the context of the English Civil War and the religious conflicts of contemporary Europe, Hobbes contended that sovereignty cannot be divided or limited and still offers a reliable source of stability and peace. To divide or seek to limit sovereignty is to risk institutionalizing the disagreements and conflicts that characterize the state of nature and would undermine the effectiveness of the sovereign's authority. Moreover, the citizens of a despotic regime could enjoy as much liberty as, and possibly even more than, those of a self-governing republic, since they too could live under the protection of laws that guaranteed their ability to exercise their rights in pursuit of their private interests.

Both Hobbes' view of the state of nature and his claim that sovereignty must be absolute were disputed by others writing in this tradition. For example, John Locke (1988 [1690]) thought that human nature was more benign than Hobbes and believed he had underestimated the degree to which state power might be an even greater danger to an individual's liberty than other individuals. Nevertheless, the contrast with the republican account of citizenship persists. Two features are especially important. First, rights are commodified as individual possessions, with political society being justified by providing for their preservation. Rights are subjective, pre-social, and prepolitical rather than being grounded in what is objectively or politically determined as right or good for a society, as the Greek and neo-Roman notions contended. Second, political legitimacy rests on a presumed act of consent whereby sovereignty is transferred from the people to the political institutions that govern them. The presumption of consent rests on the supposed rationality and necessity of this transfer once individuals begin to interact with each other on a regular basis. It is then held to persist until such time as the relevant institutions or persons fail to maintain the terms of the original compact, with the continued political participation of citizens an optional extra.

This mode of argument has proved tremendously influential in international law, especially human rights law (Pufendorf, 1991 [1673]; Kant, 1970 [1795]), and fed into

contemporary cosmopolitan conceptions of citizenship. A natural affinity also exists between this account and the liberal defense of the constitutional state. The state and its authorities are deemed to be bound by the act of constitution that justifies and legitimizes their institution, which thereby serves as a higher law to which they can be held to account – ultimately by the people, but in some accounts by their authorized representatives too – be they judges or politicians. This position also fed into views of commerce as resting on a natural system of liberty grounded in the natural rights to ownership of oneself and one's possessions.

Significantly, all three come in for trenchant criticism from J.J. Rousseau (1985 [1755], 1968 [1762]), who subjects the natural law tradition to a radical republican critique and reworking. Rousseau contends that the state of nature only takes on a Hobbesian character once individuals interact and become mutually dependent on each other, yet he agrees with Hobbes that to the extent that states are in this condition in the international sphere they too are in a state of war. However, he regards the Hobbesian social contract as systematically disadvantaging the propertyless and exploited, who only accept it out of fear. His solution was to propose a republican social contract in which the natural freedom each individual enjoyed when independent from others in the state of nature is replaced by a civil freedom that derives from each citizen participating directly in formulating laws that accord with a general will that stems from and applies equally to all. In other words, as with Hobbes sovereignty is single and indivisible, but it remains popular as only then will the laws favor the common interest rather than the partial interests of certain particular individuals over others. The difficulty was that Rousseau doubted that citizens would consistently will what was generally for the good of all – that perhaps such collective goods would not exist – outside of relatively small, undifferentiated political communities of moderate wealth. On his reading of republicanism, therefore, the original dilemma apparently persists whereby a society of equal citizens who rule and are ruled in turn only seems possible in societies that are exclusive in their membership and constrain the private lives of citizens – a vision that seems both anachronistic and coercive. As Adam Smith (1976 [1776]) and Benjamin Constant (1988 [1819]), two of Rousseau's prominent critics, noted, allowing each citizen to pursue their private interests so far as was compatible with a like pursuit by others might result in social and economic inequality but it also fostered both the commerce necessary for the wealth of nations and provided greater opportunities for individuals to exercise their liberty. Yet, both retained republican worries that in a society depleted of civic virtue, the necessary, if minimalist, state regulation risked being exploited by the rich and powerful for their own ends, and Smith especially was concerned for the misery of the poor. Therefore, a central issue was whether a modern commercial republic was possible.

Liberal Democratic Citizenship: Uniting Republican and Legal Citizenship?

The opportunity to create a modern republic confronted the two great revolutions that inaugurated the modern democratic era – the American Revolution of 1776 and the French Revolution of 1789. Both attempted to resolve it by seeing their constitutional settlements as instances of an actual contract between citizens. So, the putative authors of the American Constitution are 'We the People of the United States,' while the French Declaration of the Rights of Man and the Citizen declares "the source of all sovereignty lies essentially in the Nation." However, these formulas preserve a dualism between the 'public' political citizen, who acts as a collective agent – the 'people' or the 'nation' – and the private, 'legal' citizen, who is the subject of the law and the possessor of 'natural' rights to liberty, property, and the pursuit of happiness. Civic virtue gets assigned to a single constitutional moment and enshrined in the institutions that popular act creates, leaving selfish citizens to pursue their personal interests under the law. Meanwhile, a tension between the two models persisted. As Rousseau had noted, it is doubtful that even the most well-designed institutions and laws can economize too much on the virtues of citizens, or that citizens feel they are 'theirs,' if – the founding moment apart – they cannot actively participate in shaping them.

The liberal democratic regimes that emerged during the nineteenth and twentieth centuries struggled with this tension, mixing in their different ways the elements of both the republican and the legal forms of citizenship. Lying midway between a city-state and an empire, the nation-state emerged as their most viable alternative – able to combine certain key advantages while avoiding their disadvantages. If the *polis* was too small to survive the military encroachments of empires, the Empire was too large to allow for meaningful political participation. The nation-state had sufficient size to sustain both a complex economic infrastructure and an army, while being not so large to make a credible – if less participatory – form of democracy impossible. As a result, it became subject to pressures to create a form of citizenship that could successfully integrate popular and legal rule by linking political participation and rights with membership of a national democratic political community.

The sociologists T.H. Marshall (1950) and Stein Rokkan (1974) established what has become the standard narrative of the evolution of modern democratic citizenship. They saw citizenship as the product of the interrelated processes of state building, the emergence of commercial and industrial society, and the construction of a national consciousness, with all three driven forward in various ways by class struggle and war. The net effect of these three processes was to create a 'people,' who were entitled to be treated as equals before the law and possessed equal rights to buy and sell goods, services, and labor; whose interests were overseen by a sovereign political authority; and who shared a national identity that shaped their allegiance to each other and to their state. In a brilliant essay, Marshall argued there had been three periods in the historical evolution of citizenship as a given group fought to attain equal status as a full member of the community. The first period, from the seventeenth to mid-nineteenth centuries, saw the consolidation of the civil rights needed to engage in a range of social and economic activities, from the freedoms to own property and exchange goods, to liberty of thought and conscience. The second period, from the end of the eighteenth century to the start of the twentieth, coincided with the gaining

of political rights to vote and stand for election. The third period, from the end of the nineteenth to the midtwentieth, involved the creation of social rights that gave citizens "the right to share to the full in the social heritage and to live the life of a civilized being according to the standards prevailing in society."

Though modeled on Britain, Marshall's account reflects not just the new liberal and social democratic consensus behind a welfare state fashioned by such British thinkers as T.H. Green, L.T. Hobhouse, and W.H. Beveridge, but also similar intellectual and political movements elsewhere, like the Solidarists in France and progressives in the United States (Bellamy, 1992; Kloppenberg, 1986). Nevertheless, his argument has attracted considerable criticism. He is said to overlook the role external pressures played in promoting rights (Mann, 1987), while the three sets of rights neither arose in quite the order or periods that he mentions, nor proved quite as complementary as he assumed. Thus, social rights emerged in most countries before rather than after political rights – often being offered by the politically dominant class in the hope of damping down demands for political rights. Social and civil rights can also clash, as with the right to property (Bellamy et al., 2004). However, these corrections to the details of his argument are perfectly compatible with its underlying logic, whereby the development of legal rights stems from a subordinate group employing formal and informal political strategies to win concessions from those with power in their fight to be treated with equal concern and respect.

Writing in the 1950s, when the economies of west European countries were in the ascendant and welfare spending expanding, it was natural for Marshall to view social rights as the culmination of the struggle for an ever more inclusive and egalitarian form of citizenship. Needless to say, subsequent events have tended to challenge that optimistic conclusion. For a start, many aspects of the postwar welfare settlements Marshall celebrated got eroded during the economic downturn and restructuring of the 1970s, 1980s, and 1990s. Various New Right politicians and theorists argued for the privatization of numerous public services on the grounds they would be not only cheaper and more efficiently run but also more responsive to consumer pressure via a free market than they had been to the democratic pressures of voters upon politicians and state administrators. They also questioned whether welfare was a right of citizenship (King and Waldron, 1988). Many of the economic and social assumptions on which this settlement rested have also been criticized by those seeking to further expand rather than curtail citizenship. Environmentalists have attacked the emphasis on increasing economic production (Dobson, 2003), feminists its continued overlooking of the subordinate role of women (Lister, 2003), multiculturalists the failure to even mention issues of cultural, religious, or ethnic diversity (Kymlicka, 1995), cosmopolitans its focus on the nation-state (Benhabib, 2004), and so on.

These developments have challenged the view of the sovereign, liberal democratic, nation-state as the context for citizenship. Internally, it has been argued that the people have become too diverse for popular sovereignty not to risk degenerating into the tyranny of the majority unless minority rights have strong legal protection. Externally, state sovereignty has been regarded as both ineffective and unjust. Ineffective, because the state cannot offer citizens economic or physical security in a world dominated by global markets and global threats such as climate change, international terrorism, and nuclear weapons. Unjust, because birth into a rich or a poor, a democratic or a tyrannical state is simply a matter of good or bad luck. State sovereignty merely allows citizens of rich and democratic states to avoid their duties toward the citizens of poor and tyrannous states, often adding to their poverty and tyranny in the process. Yet, if the new forms of multinational and global citizenship attempt to go beyond sovereignty, they have for the most part found inspiration in the presovereign, classical models of ancient Greece and Imperial Rome and remain caught in the dilemma of reconciling the advantages and avoiding the disadvantages of the republican and the liberal, the democratic, and the legal models of citizenship.

See also: American Revolution, The; Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Capitalism; Communism, History of; Dictatorship; Enlightenment; First World War, The; French Revolution, The; Historical Thought and Historiography: Christian Tradition; Historical Thought and Historiography: Classical Period (Especially Greece and Rome); Historical Thought and Historiography: Current Trends; Historical Thought and Historiography: East asia; Historical Thought and Historiography: Indigenous Cultures in the Americas; Historical Thought and Historiography: Islamic Tradition; Historical Thought and Historiography: Jewish Tradition; Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; Historical Thought and Historiography: South Asia; Historical Thought and Historiography: Southeast Asia; Historical Thought and Historiography: Sub-Saharan Africa; Imperialism; Labor, Division of; Marx, Karl (1818–83); National Liberation Movements; Political History: History of Politics; Political Thought, History of; Rousseau, Jean-Jacques (1712–78); Social Inequality in History (Stratification and Classes); Social Movements: A Social Psychological Perspective; Working Classes, History of.

Bibliography

Aristotle, 1988 [335–323 BC]. In: Everson, S. (Ed.), The Politics. Cambridge University Press, Cambridge.

Bellamy, R., 1992. Liberalism and Modern Society. Polity Press, Cambridge.

Bellamy, R., 2008a. Citizenship: A Very Short Introduction. Oxford University Press, Oxford.

Bellamy, R., Castiglione, D., Santoro, E. (Eds.), 2004. Lineages of European Citizenship: Rights, Belonging and Citizenship in Eleven Nation-States. Palgrave.

Benhabib, S., 2004. The Rights of Others: Aliens, Residents and Citizens. Cambridge University Press, Cambridge.

Berlin, I., 1969. Four Essays on Liberty. Oxford University Press, Oxford.

Cicero, 1991 [44 BC]. In: T.Griffin, M., Atkins, E.M. (Eds.), On Duties. Cambridge University Press, Cambridge.

Constant, B., 1988 [1819]. The liberty of the ancients compared with that of the moderns. In: Fontana, B. (Ed.), Political Writings. Cambridge University Press, Cambridge, pp. 308–328.

Dobson, A., 2003. Citizenship and the Environment. Oxford University Press, Oxford.

Finley, M., 1983. Politics in the Ancient World. Cambridge University Press, Cambridge.

Hamilton, A., Madison, J., Jay, J., 2003 [1787–1788]. In: Ball, T. (Ed.), The Federalist. Cambridge University Press, Cambridge.
Hobbes, T., 1998 [1642]. In: Tuck, R. (Ed.), On the Citizen. Cambridge University Press, Cambridge.
Hobbes, T., 1991 [1651]. In: Tuck, R. (Ed.), Leviathan. Cambridge University Press, Cambridge.
Kant, I., 1970 [1795]. Perpetual peace. In: Reiss, H. (Ed.), Political Writings. Cambridge University Press, Cambridge, pp. 93–130.
King, D., Waldron, J., 1988. Citizenship, social citizenship and the defence of welfare provision. British Journal of Political Science 18, 415–443.
Kloppenberg, J.T., 1986. Uncertain Victory: Social Democracy and Progressivism in European and American Thought 1870–1920. Oxford University Press, Oxford.
Kymlicka, W., 1995. Multicultural Citizenship. Clarendon Press, Oxford.
Lister, R., 2003. Citizenship: Feminist Perspectives, second ed. Palgrave, Basingstoke.
Locke, J., 1988 [1690]. In: Laslett, P. (Ed.), Two Treatises of Government. Cambridge University Press, Cambridge.
Machiavelli, N., 1970 [1531]. In: Crick, B. (Ed.), The Discourses. Penguin, Harmondsworth.
Mann, M., 1987. Ruling strategies and citizenship. Sociology 21, 339–354.
Marshall, T.H., 1950. Citizenship and Social Class. Cambridge University Press, Cambridge.
Miller, D., 2000. Citizenship and National Identity. Polity, Cambridge.
Nussbaum, M., 1996. Patriotism and cosmopolitanism. In: Cohen, J. (Ed.), For Love of Country. Beacon Press, Boston, pp. 3–17.
Pettit, P., 1997. Republicanism: A Theory of Freedom and Government. Clarendon Press, Oxford.
Pocock, J.G.A., 1995. The ideal of citizenship since classical times. In: Beiner, R. (Ed.), Theorizing Citizenship. SUNY Press, New York, pp. 29–52.
Pufendorf, S., 1991 [1673]. In: Tully, J. (Ed.), On the Duty of Man and Citizen. Cambridge University Press, Cambridge.
Rokkan, S., 1974. Dimensions of state formation and nation building. In: Tilly, C. (Ed.), The Formation of National States in Western Europe. Princeton University Press, Princeton, NJ, pp. 562–600.
Rousseau, J.J., 1985 [1755]. In: Cranston, M. (Ed.), A Discourse on Inequality. Penguin, Harmondsworth.
Rousseau, J.J., 1968 [1762]. In: Cranston, M. (Ed.), The Social Contract. Penguin, Harmondsworth.
Skinner, Q., 1998. Liberty before Liberalism. Cambridge University Press, Cambridge.
Smith, A., 1976 [1776]. In: Campbell, R.H., Skinner, A.S., Todd, W.B. (Eds.), An Inquiry into the Nature and Causes of the Wealth of Nations. Oxford University Press, Oxford.
Walzer, M., 1989. Citizenship. In: Ball, T., Farr, J., Hanson, R.L. (Eds.), Political Innovation and Conceptual Change. Cambridge University Press, Cambridge, pp. 211–219.

Citizenship: Political

Mark Hoipkemier, University of Notre Dame, Notre Dame, IN, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R.M. Smith, volume 3, pp. 6368–6379, © 2001, Elsevier Ltd.

Abstract

Citizenship is an inherently political status conferring the right to self-governance. Yet recurrent emphasis upon its legal, social, and psychological aspects, often by political actors themselves, creates pressure to depoliticize citizenship. Even when its core of self-rule is obscured, the theory and practice of citizenship follow and respond to political power. Participatory (or republican) and legalistic conceptions of citizenship stand in particular tension but are logically and practically bound together, both in modern national citizenship and in the nascent politics of transnational membership.

It may seem superfluous to subtitle a discussion of citizenship as 'political.' The term is at core ineradicably political: its oldest, most basic, and most prevalent meaning is a certain sort of membership in a political community. There are, nonetheless, good reasons to underline just how deeply political citizenship is. The chief reason is that precisely because citizenship is so profoundly political a term, there are recurrent pressures to depoliticize both its meaning and its accompanying practice. These pressures and the understandings of citizenship they propagate, are best seen as confirmations of the political character of citizenship, not as alternative, apolitical conceptions.

Five Meanings of Citizenship

(1) Perhaps the most familiar meaning of citizenship is, in fact, the seminal one. In both ancient and modern republics and democracies, a citizen has been a person with political rights to participate in processes of popular self-governance. (2) Yet we also commonly speak of citizenship as a more purely legal status. Citizens are people who are recognized legally as members of a particular political community and who, therefore, possess some basic rights to be protected by that community's government, whether or not those rights include rights of political participation. (3) During the last century, many have come to use the term 'citizen' as a way of referring to those who belong to almost any human association. I can be said, metaphorically, to be a citizen of my neighborhood, my fitness club, and my university as well as my broader political community. (4) Two subsidiary uses denote aspects of citizenship that pertain to all of the first three meanings. Citizenship can refer to the collective identity uniting the members of a given association, whether a political community or some other group. The value of such a social identity is a large part of what makes citizenship a central political good. (5) Finally, we often use citizenship to signify, not just membership, but certain standards of proper conduct, implying that only 'good' citizens are truly citizens in the full meaning of the term.

Which of these meanings of citizenship can we properly term 'political'? If we understand politics simply as the distribution and use of power, all three of the initial meanings are political. We commonly speak of 'office politics' just as we speak of the 'politics of labor rights,' or the 'politics of governance' which are politics proper. The three meanings are not political in the same degree, however, and they are listed roughly in descending order. The first meaning makes the citizen both the agent and the beneficiary of public power. Ongoing collective action is at the core of this meaning. The second meaning sees the citizen mainly as the beneficiary of political power. The rights and duties of citizenship here are established by authoritative decision, but the status need not have any further connection to politics. In contrast to the first two, the third meaning involves the citizen in associations that lack final coercive power, but which, nonetheless, may involve structures of authority over goods, rights, and duties. The citizen in this third sense may be agent or beneficiary, depending on the nature of the association; but the power in question is not characteristically political. The same basic ranking holds under a refined conception of a political relationship – either in its ancient sense requiring the exchange of reasons between ruler and ruled or in the classical modern sense of mutual subjection to a common authority. Citizenship as norm and collective identity will concern the use of power to the extent that they correspond to one of the three primary meanings of citizenship.

The word citizen derives from the Latin *civis*, meaning a member of an ancient city-state, preeminently the Roman republic; but *civis* was a Latin rendering of the Greek term *polites*, a member of a Greek *polis*. Innumerable scholars have told how a renowned resident of the Athenian *polis*, Aristotle, defined a *polites* or citizen as someone who rules and is ruled in turn, making citizenship conceptually inseparable from political governance (Aristotle, 1968: p. 1275a23). Aristotle doubtless pleased many Athenians by arguing that such a status and activity, properly performed, represented the highest form of life available to most people. Yet this life was not available to Aristotle himself; he was not an Athenian citizen but a metic, a resident alien. He also suggested that the philosophic life he could and did pursue was, in the end, the highest of all; and the fact that he chose to pursue this life in a city that denied him citizenship may call into question how valuable he really thought citizenship to be.

His fulsome praise of the citizenship his Athenian hosts had created thus suggests how greatly definitions of citizenship have always been shaped by the political structures of power within which they have been offered. It clearly would not have been prudent for Aristotle to denigrate Athenian citizenship. The fact that neither Aristotle nor most of other residents of Athens – including aliens, women, and slaves, – were eligible for citizenship also underlines how citizenship originated; not only as a way of structuring membership but also as a way of distributing power within a particular political regime. That distribution disempowered many more Athenians than it enfranchised. Even so, the ideal of citizenship as self-governance that Athens and Aristotle established, has often served since as an inspiration and instrument for political efforts to achieve greater inclusion and engagement in political life.

As such, this ancient idea of citizenship has often seemed politically threatening to many rulers, who have abolished or redefined the category. It was for this sort of political reason – because the regimes that had created citizenship succumbed to conquest by Alexander the Great's monarchical empire – that ancient Greek citizenship disappeared. And it was for a similar political reason – because the Roman republic gave way to imperial rule generated from within – that Roman citizenship came to have a different meaning than the one Aristotle articulated. In principle, Roman citizenship always carried with it the right to sit in the popular legislative assembly that had been the hallmark of Athenian citizenship. But as participation in that assembly became increasingly meaningless and impractical for most imperial inhabitants, Roman citizenship became essentially a legal status comparable to modern nationality (Pocock, 1995). It provided rights to legal protection by Roman soldiers and judges in return for allegiance to Rome. That status was less 'political' in the sense that it no longer evoked recurrent engagement in practices of self-governance, but it was quite plainly a politically crafted status that represented a new distribution of power of enormous political significance.

Citizenship was then eclipsed in the West by the various feudal and religious statuses of the medieval Christian world, where no community could claim the exclusive allegiance of its members. The claims of what would later develop into sovereign nation-states, were limited from below by the personal authority of local rulers and from above by the universal claims of the Christian church. Yet citizenship did not vanish entirely. 'Burghers,' or the *bourgeoisie*, were citizens of municipalities that often had some special, if restricted, rights of self-governance within feudal hierarchies. Such burghers remained, however, fundamentally subjects of some ruling prince or lord, with their citizenship chiefly providing legal rights of protection in the manner of Roman imperial citizenship. In contrast, during the Renaissance, some Italian cities achieved both independence and a meaningful measure of popular self-governance. They invoked the ancient 'republican' ideals of participatory citizenship to define and defend their regimes. Their experiences, in turn, fed into the antimonarchical revolutions that created the first modern republics, including the short-lived, seventeenth-century English Commonwealth and the late eighteenth-century French Republic, as well as the still-enduring United States (Pocock, 1975).

In complex fashion, those revolutions inaugurated transformations 'from subjectship to citizenship' across much of the globe that are still ongoing. Today most of the world's governments proclaim themselves to be 'republics' of some sort populated by citizens – whether or not their political practices correspond to any historical meanings of these terms. With its political use greatly broadened, the term 'citizen' has been further extended to cover almost every kind of membership in any organization and to equate genuine citizenship with being a good, contributing member of those organizations. These pervasive popularizations of the term reflect, however, political developments that have in some respects diminished the significance of citizenship, even as the term has spread.

The Politics of Apolitical Citizenship

The early modern republics were formed in an international realm that had been organized by the 1648 Treaty of Westphalia into a system of mutual recognition among overwhelmingly monarchical nation-states. In gaining acceptance within that system, the new republics defined their citizens as having the same international status as national monarchical subjects. For international purposes, these citizens, too, were simply persons who owed allegiance to, and could claim protection from particular governments. Thus, Westphalian international law treated modern republican citizenship as akin to the legalistic, protection-oriented version of Roman citizenship.

Americans, especially, forged their republic amid racial and gender hierarchies that few leaders sought to challenge. Hence they felt compelled to argue that though free blacks and women might be citizens, citizenship did not, in fact, inherently entail rights of political participation. It guaranteed, once again, only more limited rights to certain judicial and executive protections (Smith, 1997). For long stretches of time, then, both international and domestic politics worked to strengthen legalistic, as opposed to the more participatory conceptions of citizenship, despite the rise of modern republicanism.

Even though courts made the narrower, protection-centered view of citizenship legally authoritative, the notion that genuine citizenship involved rights of political participation remained a resonant rhetorical tool of legislative and constitutional reformers. Eventually both domestic protest movements and international pressures, including the need for broad support in wartime, converged to work for the expansion of the franchise to all adult citizens in the United States and most of the Western world. In America, blacks won both citizenship and voting rights after the Civil War, even though most came to be effectively disfranchised in the 'Jim Crow' segregation era; and women gained the franchise after World War I. In both cases, arguments emphasizing the public service of the disfranchised, especially in wartime, and the idea that true citizenship must include the franchise, played key roles in their enfranchisement (Foner, 1988; Flexner, 1973). In Britain, and to some degree in other Western European nations that had been politically configured essentially by feudal and industrial class systems, modern citizenship was wrought out via somewhat different struggles. As Marshall famously argued, first middle class and then working class political pressures resulted

in the expansion of the civil rights of property and protection, then in near-universal rights of political participation, and finally and incompletely, in 'social rights' for all national citizens, which included income, housing, medical, and educational guarantees (Marshall, 1950).

But even as the franchise has broadened in the United States, Europe, and elsewhere, the notion of citizenship as active participation in meaningful self-governance has become more remote to many modern citizens. In part, the logistics of large-scale modern societies make effective democratic participation very difficult. In part, the economic and cultural developments that have led to a focus on 'social citizenship,' make political activism seem less important. Engagement in social and economic organizations can appear more pressing and efficacious. Succeeding postwar generations increasingly favor direct action on issues such as global poverty and the environment, bypassing traditional political questions and modes (Dalton, 2009). Perhaps, then, the term citizenship has become common in such contexts because it is there that people find the active social memberships that remain at the heart of the concept of citizenship (Carens, 2013). If so, then the inevitable corollary is that citizenship, understood as political self-governance, has become quite secondary to many modern citizens. Yet this, too, is a signal political development. It is probable that like their predecessors in other regimes, many who wield power in modern republics are content when those they govern think of citizenship chiefly in terms of subnational, often nongovernmental associations; and in terms of the 'good citizen's' civic service rather than vigorous political participation. Certainly few policies within modern republics do much to enhance the feasibility and potency of such participation.

By their very logics, citizenship as self-governance and citizenship as legal status are driven apart and also drawn back together. The history of citizenship has borne out this dual motion. Popular self-governance, famously termed "the liberty of the ancients," does not easily accept limitation by personal legal immunities, "the liberty of the moderns" (Constant, 1988). Neither do these two aspects of citizenship flourish in the same conditions. As citizenship expands to include more members, a genuinely consequential form of self-rule grows less feasible. Each citizen's share of decision diminishes as the citizenry grows. Changing institutional forms, from direct to representative forms of governance, can maintain formal participation but not real influence for citizens of large communities. While citizenship as meaningful self-rule, the 'Greek' form, cannot be indefinitely expanded, legalistic citizenship can extend as far as a polity can project bureaucratic and judicial power. This 'Roman' variant of citizenship not only permits, but even encourages a larger communal setting. As James Madison famously argued, in a large republic, the smaller share of power possessed by discrete members will hamper the use of politics to violate others' rights (Madison, 2001). Participatory and legalistic citizenship are set in tension by the divergent conditions that favor the one or the other.

A different logic explains why the several meanings of citizenship have not broken apart into separate concepts. Legalistic citizenship emphasizes the rights and duties that legal status brings. Yet that status is defined and implemented by political action. Citizenship that does not involve self-rule, i.e., participation in the politics that undergirds the status itself, is precarious. It depends upon the good offices of those exercising political power to maintain its character. There is thus a powerful incentive to insist upon participation as a guarantee of political, social, or civic benefits. This logic underwrote the modern extension of the franchise. Conversely, the practice of political, and especially democratic participation conduces the establishment of legal immunities like those at the heart of modern citizenship (Ober, 2005). The practical connection between the two dominant meanings of citizenship binds them together and helps to explain the continuing appeal of citizenship as self-rule, even in an institutionally inhospitable setting. Although modern citizenship renders participatory self-governance difficult, we should not expect this dimension of citizenship simply to evaporate.

The Prospect of Postnational Citizenships

Though some activists and scholars lament decreasing participation in national politics, others stress that the heightened transnational economic, transportation, and communication systems that we call globalization are in any case making traditional notions of national citizenship obsolete. Regional associations, international legal institutions, and transnational economic, cultural, and political organizations are all said to be more likely to shape humanity's future than existing national regimes. Hence membership in such bodies may represent the most important forms of citizenship in the twenty-first century (Bauböck, 1994). What these changes will entail for a political form of citizenship is as yet unclear, because several avenues are possible.

The postwar rise of human rights ranks among the most important developments for the theory and practice of modern citizenship, with its emphasis on legal status. Many of the benefits once accorded on the basis of shared nationality are increasingly anchored in universal humanity. Such benefits include not only rights against harm, but also positive legal powers and even social entitlements, e.g. to education or health care. As the slate of benefits accorded to all persons expands, the relative worth of national citizenship diminishes. "The logic of personhood supersedes the logic of national citizenship" (Soysal, 1994: p. 164). Formal political rights are not typically granted to noncitizens, but universal rights regimes secure many of the substantive benefits that once made political rights desirable. Global conventions on human rights and the institutions that enforce them, might thus be said to form the basis of a cosmopolitan citizenship. Such a view emphasizes citizenship in its legalistic sense, since the prospect of a global collective identity, much less broad participation in geopolitical institutions, remains remote. Some thinkers, wary of the global scope of this trend, have proposed *ius nexi* (Shachar, 2009), whereby the benefits of citizenship are granted on the basis of social membership in a polity rather than legal status. In both cases, the denationalization of rights already challenges the exclusivity of national citizenship but broader globalizing trends may reconfigure the meanings of citizenship in any of three ways.

First, citizenship may retain the political aspect of self-governance along with legal protections, but in the context of enlarged or redrawn political communities (see Isin and Saward, 2013). This is likely to the extent that "multinational politics" becomes a practical reality for individuals. Increasingly, personal activities and important human goods involve interactions with cross-national groups such as large companies or NGOs, and with international bodies such as the European Union or the United Nations. Citizenship in the sense of self-rule is possible in such groups largely to the extent that individuals participate in, and are governed by them in direct fashion. On the other side, inasmuch as national governmental actors remain major players in transnational organizations, the individuals' relations to such groups constitute a dimension of national citizenship rather than a separate locus of membership. But if power and legitimacy transfer to transnational bodies, rendering them self-governing political communities in a real sense, participatory self-rule can and indeed must concern collective action at this level. Citizenship in transnational communities would likely be legalistic to at least the same degree as modern national citizenship, due to their large scale (although the population of the whole European Union, for instance, is far smaller than that of the most populous states). To the extent that transnational communities become genuinely self-governing, thus encroaching upon the sovereignty of nation-states, citizenship could attach to them and retain its political character.

Second, globalizing trends could also foster a depoliticized 'world citizenship' if claims of membership run ahead of actual contexts of power, legitimacy, and shared meanings. Changing norms of 'engaged citizenship' (Dalton, 2009) point beyond the national political context to issues of global poverty, climate change, and international social justice. Such citizenship is understood as membership in an imagined world community facing these global issues (Cabrera, 2010). Cosmopolitan projects do not belong to political self-rule, however, if they are undertaken in the hope of collective action but lack any structures of authority, reciprocity, or mutual knowledge (see Miller, 2000). Those who merely share a global ethic are not yet fellow citizens. Yet suppose that a durable shift in the norms of good citizenship did take place along these lines, effectively depoliticizing citizenship by putting it out of phase with political reality. One would expect, nonetheless, that a normative shift toward cosmopolitanism would be followed by a corresponding shift in the legal form of citizenship, and ultimately, by some reinvigoration of political participation in the newly imagined and instituted community. The several meanings of citizenship, while distinct, are mutually influential and unlikely to diverge completely. The normative sense of citizenship exerts some pull on its legal and political meaning. On the one hand, then, ideas of global citizenship may be drained of political content if they posit membership in a community that neither exists nor governs as a matter of real human interaction. On the other hand, cosmopolitan norms can themselves serve as reasons to enact the envisioned reality, drawing citizenship and politics together.

Third and finally, globalizing trends that make the nation less salient also open the possibility of local (or subnational) forms of citizenship. If transnational bodies do come to a larger, genuinely political role, they may offer only slight prospects for participation and collective identity. If so, the value of direct participation that historically inhered in the notion of citizenship may favor the reinvigoration of local forms of membership and self-rule (Maas, 2013). The scale and impersonality of national citizenship already encourage private forms of collective action. A global citizenship of human rights would only sharpen this movement. It is possible that organizations formed for localized civic service (more than global activism) might consolidate genuinely political functions and so provide the chance for self-governance that legal citizenship in a far-flung community does not. Local forms of citizenship could offer stronger collective identities by differentiating from other groups (Benhabib, 1996), than a universal citizenship offering no communal distinctness. This possibility envisions currently subpolitical groups taking on genuinely self-governing roles and so becoming political. In this scenario, the participatory and legal aspects of citizenship would diverge in large degree, with rights anchored in large communities and self-rule attached to small ones. This divergence would render the arrangement less stable in the long term, because of the logical and practical forces binding the aspects of citizenship together.

The prospects of postnational membership are various, and it is unclear how far national citizenship will recede in the face of globalization. History suggests that the leaders of political communities rarely give up power willingly, either to encompassing or to constituent groups. It is not surprising that efforts to resist globalizing trends and reinvigorate loyalties to existing nations and regimes are also visible in modern 'citizenship politics,' particularly in regard to immigration policies. None of the avenues for postnational membership, moreover, envision the disappearance of national citizenship – only its dilution by competing communities of identity and self-rule. Citizenship will likely track the foci of political power as they shift or fragment, and so the fate of national citizenship will be bound up with the sovereign state. But whatever forms of citizenship result, they will almost certainly be the products of political contests that result in distributions of powers and memberships to some people and not others, – distributions that will convey to them certain rights and protections, and not to others. Hence, citizenship will remain what it has always been, a fundamentally political status through which human beings partly order both their individual and their collective lives.

See also: Attitudes, Political and Civic Culture; Citizen Participation; Citizenship, Historical Development of; Citizenship, Sociological Aspects of; Politics and Participation.

Bibliography

Aristotle, E., 1968. The Politics of Aristotle. Clarendon Press, Oxford.
Bauböck, R., 1994. Transnational Citizenship. Edward Elgar, Aldershot.
Benhabib, S., 1996. The democratic moment and the problem of difference. In: Benhabib, S. (Ed.), Democracy and Difference. Princeton University Press, Princeton, NJ, pp. 3–18.
Cabrera, L., 2010. The Practice of Global Citizenship. Cambridge University Press, Cambridge, UK.
Carens, J., 2013. The Ethics of Immigration. Oxford University Press, Oxford.

Constant, B., 1988. The liberty of the ancients compared with that of the moderns. In: Fontana, B. (Ed.), Political Writings. Cambridge University Press, Cambridge, UK, pp. 307–328.

Dalton, R., 2009. The Good Citizen: How a Younger Generation Is Reshaping American Politics, second ed. CQ Press, Washington, DC.

Flexner, 1973. Century of Struggle: The Woman's Rights Movement in the United States. Atheneum, New York.

Foner, E., 1988. Reconstruction: America's Unfinished Revolution, 1863–1877. Harper & Row, New York.

Isin, E., Saward, M., 2013. Enacting European Citizenship. Cambridge University Press, Cambridge.

Maas, W., 2013. Varieties of multilevel citizenship. In: Maas, W. (Ed.), Multilevel Citizenship. University of Pennsylvania Press, Philadelphia, pp. 1–21.

Madison, J., 2001. Federalist 10. In: Carey, G., McClellan, J. (Eds.), The Federalist. Liberty Fund, Indianapolis, IN.

Marshall, T.H., 1950. Citizenship and Social Class and Other Essays. Cambridge University Press, Cambridge, UK.

Miller, D., 2000. Citizenship and National Identity. Polity, Cambridge, UK.

Ober, J., 2005. Athenian Legacies: Essays on the Politics of Going on Together. Princeton University Press, Princeton, NJ.

Pocock, J.G.A., 1975. The Machiavellian Moment: Florentine Political Thought and the Atlantic Republican Tradition. Princeton University Press, Princeton, NJ.

Pocock, J.G.A., 1995. The ideal of citizenship since classical times. In: Beiner, R.S. (Ed.), Theorizing Citizenship. State University Press of New York, Albany, NY.

Smith, R.M., 1997. Civic Ideals: Conflicting Visions of Citizenship in U.S. History. Yale University Press, New Haven, CT.

Shachar, A., 2009. The Birthright Lottery: Citizenship and Global Inequality. Harvard University Press, Cambridge, MA.

Soysal, Y.N., 1994. Limits of Citizenship: Migrants and Postnational Membership in Europe. University of Chicago Press, Chicago.

Citizenship, Sociological Aspects of

Thomas Janoski and Sara Compion, University of Kentucky, Lexington, KY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous article by P. Birnbaum, volume 3, pp. 1860–1862, © 2001, Elsevier Ltd.

Abstract

This article defines citizenship and discusses its three basic theoretical traditions. We then delve into seven important areas of citizenship in terms of how: (1) rights and obligations have grown, (2) identities are framed for new citizens; (3) citizenship is universal or particular, or individual or group-based; (4) citizenship is contained within boundaries and externalized to the global; (5) citizenship is structured into various regime types; (6) citizenship has moved from being a status to involving people in a process; and (7) citizenship may develop in the future with new technologies.

In sociological theory, citizenship is defined as the active and passive membership of individuals in a rights-giving and obligations-receiving entity – such as a nation-state, the European Union, and even the world – where specified universalistic rights and obligations approximate a certain level of equality. This definition builds upon the foundation laid by Thomas H. Marshall (1964), who put forward the sociological concept of citizenship as a progression of legal, political, and social rights, framed within a context of the British welfare state of the 1950s. His theoretical framework was largely based on Tocqueville and Durkheim's conceptions of society, and despite his attention to class conflict, many attribute his developmental theory of rights to functionalist influences.

Two decades later, Brian Turner (1986) advanced citizenship theory to explicitly include conflict perspectives that emphasized the role of identity-based social movements aimed at expanding notions of personhood for excluded groups (race, gender, and class) and extending existing or new rights to nascent groups. More recent investigations propose an orientation that seeks to balance the development of rights with obligations at the individual and group levels of society (Janoski, 1998), or seeks to deconstruct citizenship into more particularistic rights for different groups (Lister, 2010). An emerging element in this progression of theory formation is the conceptualization of citizenship in new or developing nation-states.

The sociological discipline offers a broad range of theories that analyze citizenship in diverse relational contexts that are governed by identity and power. Political and liberal theoretical orientations are concerned with individualized rights, property, and voting. They emphasize rights over obligations (Rawls, 1971, 1993). Social democratic or deliberational theories focus on the citizenry, and debate how alternatives will persuade and then benefit the many (Habermas, 1989, 1996). These approaches tend to offer a more balanced view of rights and obligations. Communitarian theories are more concerned with the well-being of communities, and they emphasize obligations more than rights (Etzioni, 1993; Kymlicka, 1995). Postmodern and related theories have argued against essentializing citizenship. They withdraw concepts of universality in favor of emphasizing value change or deconstructing citizenship as a very particular set of rights and obligations for specific groups to pursue (Inglehart, 1997; Bauman, 2007). Some multiculturalists have argued that legal citizenship status doesn't matter since many immigrants and minorities can access all the rights of other citizens; and they suggest that dual nationality seems to be rampant and undermines the construct of citizenship. However, recent events have shown that these rights can be taken away, and naturalization examinations in various European countries ask for some rather specific cultural information (e.g., how to react to having beer spilled on you in a pub in the UK, or having to watch films showing a wide variety of sexual orientations in the Netherlands). So what was once thought to be totally multicultural is looking more like assimilationist socialization. For a review of these theories, see Janoski (1998) and the 'approaches' chapters in Isin and Turner (2002).

Viewed collectively, sociological theories tend to emphasize social and politico-economic aspects of citizenship over and above the legal. In particular, the focus is on: (1) the expansion of rights and obligations; (2) framing the identities of new, existing, or disqualified persons; (3) universal versus particular and individual versus group citizenship; (4) internal versus the external boundaries of citizenship; (5) citizenship variations across regime types; (6) citizenship as a process; and (7) citizenship in the future in the light of emerging technologies. We review these topics in more detail in the rest of this article.

Expanding Rights and Obligations

Numerous books and articles have been written about the development of rights over time across countries. Many of these studies were not necessarily about citizenship, but covered topics about workers forming labor unions, women fighting for freedom of choice, racial groups demonstrating in the civil rights movement, ethnic groups battling for recognition in immigration laws, and so on. In the 1970s, citizenship gained currency and more people working on these different issues espoused a citizenship-focused theoretical framework. This increased interest in citizenship was symbolized by Evelyn Nagano Glenn when she adopted 'citizenship' as the topical theme of her presidential convention for the American Sociological Association (2011). A bit less common are studies that follow the development of the four various kinds

of citizenship rights – legal, political, social, and participation rights. A study that puts the legal and political rights history of the US together over a long period of time is Rogers Smith's *Civic Ideals* (1997). The welfare state literature in political sociology has done extensive work on the development of citizenship rights over time, especially with time-series and pooled analyses of the advanced industrialized countries.

Scholars from western liberal countries, with a few exceptions, tend to ignore or minimize the obligations associated with citizenship, and focus more on rights. Generally, it is the conservative scholars who look at the long-term development of obligations, such as taxing policies, military conscription, jury duty, and so forth. Following the lead of Alexis de Tocqueville, other sociologists study the roots of volunteering, civic engagement, and political participation in relation to the development of citizenship (Musick and Wilson, 2008). Political sociologists also focus on political engagement in the form of campaign contributions, working on campaigns, and running for office as a form of citizenship behavior particular to democratic contexts. While criminology emphasizes the study of obedience to the law, sociology tends toward studying opposition to the law, especially in the form of social movements. In this manner, a sociological perspective is distinct because it underscores the importance of equally understanding rights and obligations in the process of citizenship.

New and/or Disqualified Persons

Personhood – the status of being a citizen – is controversially tied into social and political conceptualizations of citizenships. Acknowledgment of personhood is frequently used in the process whereby an individual from an excluded group gains a status as a citizen. The disqualified are denied personhood because of perceived or real deficiencies in reason, morals, or divergent national or community interests (i.e., allegiance to a foreign land, the Pope, or an Imam). For instance, African-Americans and Asian-immigrants were long denied citizenship on the grounds that either they were irrational, or their loyalties were attached to foreign governments. Most countries have procedures to offer personhood to aliens through naturalization toward citizenship, but fulfilling obligations through military service is often quite effective (Janoski, 2010; Brubaker, 1992). However, many European countries give aliens the right to vote in local elections, despite a seeming denial of personhood.

According to some definitions of citizenship that are based on personhood, 'competence' is a defining feature. Under these criteria, certain disabled or other socially excluded individuals may be denied full citizenship status by law. For example, working class men who were uneducated in seventeenth century England faced a charge of incompetence, and they were also accused of having inadequate interests in protecting property since they had little or none. Similarly, during World War II, Japanese-Americans had their citizenship restricted first on the basis of their questionable competence as reasoned men. And in the two decades after World War II, African-Americans fought the civil rights movement for equal rights after they were nominally granted these rights in the Emancipation Proclamation and the 14th Amendment a century earlier. Increasingly, scholars in the race and ethnicity tradition have been embracing citizenship theory in order to better explain discrimination and prejudice (Gates et al., 2012; Roman, 2010; Somers, 2008). Glenn (2011: 3) recommends that we go beyond personhood as a legally deployed status, and focus more on how others require and construct recognition.

Establishing total or complete competence has been discarded with various physically and mentally challenged groups. While these groups have gone to great efforts to show that they are competent to various degrees, their claim is much more based on a human rights argument that all humans, no matter whether they are 'temporarily abled' or not, are entitled to rights or 'the least restricted environment.' This has also applied extensively to mentally ill persons who make claims to be treated with dignity and compassion in society. Another complication connected to competence, is that many of these challenged or mentally ill groups often require representation by people not in their group (i.e., parents, spouses, or other advocates). Occasionally, exceptional cases emerge such as Temple Grandin who has autism and is a professionally-trained veterinarian. But citizenship for these groups cannot rest on truly exceptional people. It must rely on all disabled citizens and the varieties of experiences that they bring to society through what David Matza calls 'natural deviance' (Carey, 2010).

Individuals can also be denied personhood on the basis of moral grounds. When individuals break the legal, political, social, or religious moral codes in society they are subject to sanctions. These sanctions may deprive them of certain individual citizenship privileges – implying that their behavior is unbecoming of morally civil human beings. For example, prisoners are denied their full entitlements as citizens even after they have served their punishment for violating the rights of others. Once they have paid their debt to society, they should have their rights restored. Similarly, lesbian, gay, bisexual, and transgender (LGBT) persons have been/are denied certain rights as full citizens – the most prominent being the right to marriage – on the basis of moral arguments against their sexual orientation (Fisher, 2008). To gain full personhood recognition, people from these groups have had to argue, demonstrate, or fulfill obligations to make the case for their moral and intellectual worth.

Women too, have long fought for equal rights, and the ways in which women gradually achieved political, civil, and social rights often followed a different pattern from that of men (Lister, 2010). Women have progressively legitimized their personhood through waves of feminist movements, which have served to highlight the gendered nature of state policy and the inadequate theorizing of mainstream citizenship, which ignores a female perspective and discounts the relevance of what happens in the private sphere. Going back to the US and British early nineteenth century, feminist actions brought to light the inequalities facing female citizens when labor policies and development strategies favored men. By the end of that century, feminist campaigns had successfully earned women in most countries the basic legal rights as citizens to vote, hold political office, and take part in the workforce. Moreover, the growing influence of women in public discourse confronted traditionally patriarchal structures of the state, in which

healthy, working citizens are defined in masculine terms. Consequently, as women gain status as a significant interest group, so do the primary issues that concern them, such as health care, maternity rights, birth control and abortion, labor market discrimination, secure housing, infant and old age care, and education (Misra and King, 2005: 529).

Some, including many feminists, see citizenship as being unduly relegated to the public sphere when most of the fight for women's rights concern trying to regulate what goes on in the private sphere (Lister, 2010). Lister suggests that the powerful public–private dichotomy is deeply gendered, and has meant that women's admission to citizenship has been on male terms (Lister, 2003). A feminist perspective regards citizenship as contested; and as a concept in which the legal rights discourse is individualistic and male-inspired. According to Lister (2010) we need a revised discussion of the rights and duties of citizenship in order to acknowledge the performance of men and women's share of work/care in the private realm as a part of a general citizen's obligation.

As has been the case for women, minority racial groups, LGBT persons, and immigrants, the dominant power holders initially discredit all 'new' groups seeking citizenship. There are many discredited groups in society that successfully or unsuccessfully pursue personhood. Pedophiles are one such group generally regarded as being morally reprehensible and devoid of full personhood status – even in prisons where they are singled out for punishment by their fellow prisoners. Sexually abusing children is strongly condemned in most contemporary societies. However, there are variations between countries about the age at which childhood ends, which impact on someone being defined as a child. Beyond those doubts, some could argue that pedophilia is a disease, thus giving pedophiles a right to treatment in their communities and a restoration of personhood and citizenship rights. The medicalization of pedophilia could remove the stigma that is attached to this condition and they could then regain moral standing. In other words, the acceptance or denial of citizenship is heavily dependent on the messy criteria for defining personhood. This case would test the bounds of personhood and how much we might expand the concept of citizen behaviors.

Universal/Particular Rights and Individual/Group Rights

Citizenships rights can be extended to individuals, and also to groups. Following its strong liberal roots, citizenship is most often based on the allocation of individualistic rights that are universally available to all. This approach ensures that conflict between rights and obligations are relatively easy to resolve on a one-to-one or government-to-one basis. Group rights, on the other hand, tend to be framed in connection to honoring diverse human or cultural rights and are allocated to groups of people, not individuals. Theoretically, there are three levels of viewing this latter distinction (Janoski, 1998; Leydet, 2011).

First, many liberal countries have given rights to cultural or linguistic minority populations. For example, in the US, group rights are available to Native American tribes and to Amish dissenting communities who may exist in a separate community – sometimes upholding their own legal system. This protects their more communitarian culture from the individualistic ravages of general liberal society. In Canada and Belgium, similar group rights belong to linguistically defined communities such as the Québécois, Nunavut, Walloon, or the Flemish.

Second, neo-corporatist and social democratic theory gives very strong group rights to labor unions, who gain identity by encircling a large percentage of the population. One example of this is the neo-corporatist structure of bargaining where labor, management, and the state bargain over basic wage structures and the level of work. An example of this is reflected in South Africa's incorporation of the Congress of South African Trade Unions (COSATU) into the nation's parliamentary governance structure. This is particularly effective where 80–90% of the workforce is represented by unions as in Sweden.

Third, communitarianism advocates for multiplex group rights in societies that can be divided up into many subgroups. Communitarianism emphasizes the protection of subcultures, and places importance on obligations over and above individualistic rights. Will Kymlicka (1995) offers a strong example of communitarianism, and is a strong proponent of group rights for immigrant subpopulations in France and Canada. Perhaps, more controversially – such is the argument made by the South African Congress of Traditional Leaders (COTL) – a communitarian vision is one in which the purpose of the nation-state is to accord ethnic groups sovereign autonomy over their realms. In other words, the state functions to "put the dictates of indigeneity before the universal rights of citizens" (Comaroff and Comaroff, 2009: 3). Group rights work best with members living in tightly knit communities where there is little influence from the outside. A cosmopolitan community would increase the conflicts between exclusive rights.

Internal versus External Boundaries

The concept of citizenship has a socially defining characteristic, by ordering people into 'in' and 'out' groups. Rogers Brubaker (1992) addressed the closed nature of citizenship by employing Max Weber's reasoning that one can only have an exclusive status such as citizenship if others do not have it. This resulted in some disputes over citizenship versus human rights. If everyone is a citizen of the world with human rights, how might such a system be enforced? The common answer is that nation-states enforce citizenship rights, and as a result, they do not enforce rights for non-citizens. Hence, citizenship belongs to the nation-state. A large body of literature has developed to cover the processes of naturalization, by which external immigrants become internal citizens (Janoski, 2010; Joppke, 2005).

Four main challenges have emerged to the nation-state being the keeper of internal citizenship status. First, the ability to hold dual citizenships challenged the exclusivity of national membership. As nation-states gradually dropped the insistence that wives take their husband's citizenship, this allowed women to hold two nationalities, and subsequently opened the door for children born from these unions to also lay claim to citizenship in two countries. The Cold War kept a lid on these processes since soldiers were expected to have exclusive

loyalty to one country, but as communism fell, and air travel expanded, states began to liberalize their dual citizenship provisions (Kivisto and Faist, 2008).

The second major challenge to the exclusivity of nation-state citizenship was the formation of the European Union (EU). When the EU developed its unified labor market provisions and recognized citizenship for all members of the EU, it reimagined the purposes of national borders and passports, allowing for a new context of transnational labor, currency creation, economic flows, and citizenship rights. People living within the union could get a job anywhere they wanted in the EU, and their rights could be protected, no matter which EU country they were working in. This fits the concept of a 'transnational state' enforcing citizenship rights (Bauböck, 2003).

Third, the globalization of business and the ability to own property in many countries throughout the world has transformed traditional understandings of national citizenship. Transnationalism has flourished in the global capitalist economy, especially among the Chinese and Indian diaspora, which proliferate throughout Southeast Asia, parts of Africa, and the Americas. In this context, dual citizenship is much sought after. Although holding multiple citizenships are not a requirement for conducting international trade, it certainly eases the burden of travel, passport control, visa applications, and property ownership. Aihwa Ong (2006) contends that transnational citizenship transcends political boundaries for the sake of more effective economic transactions. Ong (2006) connects citizenship to neoliberalism by describing this ideology as an incredibly 'malleable technology of governing,' which can be used in different ways to suit authoritarian, democratic, or communist regimes. As a result, Southeastern and Eastern Asian states are changing their practices to compete more effectively in the global economy.

Furthermore, these perspectives were especially 'interested' in the presentation of self and identities that might develop with these multiple interactions, with formal status for each one. Thus transnational citizenship could include participating in the civic associations and voting in multiple 'interested' countries. Sometimes countries have particular interests in satisfying their diaspora, so that they will continue to send money and remittances back to the home country. Maintaining their original citizenship is one way to do this. But 'sending' countries have also hampered their immigrants in obtaining new citizenship, by causing them to lose property at home if they naturalized to the 'receiving' country. But in the prominent cases of Turkey and Mexico, these restrictions were recently loosened.

The fourth challenge to nation-bound citizenship comes with the emergence of the international aid movement in the development era. In this worldview, every person is a 'global citizen' and is responsible for being civically engaged with matters that concern all of humanity – no matter in which country they take place – such as environmental degradation in the Amazon basin or starving children in Sudan. The concept of cosmopolitanism presents a one-world view, in which every person is a member. This approach has garnered a lot of support from international organizations defending the rights of people in many different countries. International nongovernmental organizations may provide civil surveillance services or act as moral watchdogs (e.g., Amnesty International and Human Rights Watch); or they may provide basic services that protect human rights (e.g., Doctors without Borders and the Red Cross providing health care, food, shelter and other aid to refugees and needy persons throughout the world). The United Nations High Commissioner for Refugees is one instance where the UN takes a strong role in this area.

Status as a global citizen is still highly abstract, and although the League of Nations issued Nansen passports after World War I, the UN has never officially engaged in issuing anything like a world passport. Hence, global or cosmopolitan citizenship usually involves an individual holding multiple-country passports – they are legally citizens of many countries. However, governments disagree about the legitimacy of these individuals, and many are reluctant to issue citizenship to an individual who already has more than one nationality. Despite holding multiple passports, governments can restrict entry to certain people through the issuances of visas. Some countries have downplayed the visa process to the point that visitors – or certain visitors – do not need entry visas. Other countries, like China and Russia have strict visa requirements and are very careful about who comes into their territory and who leaves. Furthermore, China is extremely suspect of cosmopolitan citizenship and does not accept dual nationality, except under very limited conditions. The US does not officially recognize dual citizenship but allows it for children and even recommends informal (but illegal) acceptance i.e., "don't tell us if you did not actually drop your other citizenships" (Kivisto and Faist, 2008; Ong, 1999, 2006; Held, 2010).

Consequently, transnational citizenship is limited to smaller and more ambitious countries in Southeast Asia. Other than the European Union, no supranational entity, or unified set of nations, has bestowed citizenship if an individual nation-state is not willing to do so. Refugees face a precarious situation and the world's nation-states are increasingly resistant to accepting refugees. Jordan, with the world's highest per capita refugee population, is stretched to the limit and there are few countries stepping up to help the Jordanians take care of these individuals.

While not necessarily connected to transnational citizenship, the UN has had a strong impact on nation-states in terms of developing treaties that support various rights of groups, categories of people, and individuals. For example, the UN adopted the International Covenant on Civil and Political Rights, and the International Covenant on Economic, Social, and Cultural Rights in 1966. It has passed other conventions on racial discrimination, women's rights, torture, rights of the child, rights for those with disabilities, and the rights of migrant workers. The Geneva Conventions enacted between 1865 and 1949, and the UN Human Rights Council established in 2006, deal primarily with Human Rights issues, but also with matters of citizenship. While laudable, most of these covenants and conventions need considerable bolstering before they have an enforcement mechanism, and many countries, like the US, don't sign and ratify many of them.

Citizenship Variations in Regime Types

One could hardly expect citizenship rights, obligations, and processes to be the same in all countries across the different continents. Societies vary considerably on the set of citizenship

rights that they offer to their people, and in terms of the institutions that they set up to deal with these agreements. In many ways this often overlaps with various theories of civil society. Studies using an historical time-series analysis, in contrast to a comparative approach, have yielded great insight into the different country conceptions and implementations of citizenship.

Gøsta Esping-Andersen developed a major contribution to citizenship when he came up with the idea of three welfare state regimes under capitalism. He showed quantifiably that welfare states differed consistently between liberal (e.g., US, UK, etc.), traditional (e.g., Germany, France, etc.), and social democratic (e.g., Sweden, the Netherlands, etc.). The average scores for various welfare state characteristics varied considerably, as did the state structure in their level of expenditures, and the way that they structured their institutions for each program. For instance, liberal regimes provide means-tested (i.e., testing for low income) benefits whereas social democratic regimes share benefits with all. Traditional regime pensions rely on insurance programs particular to industrial sectors that are highly regulated by the state, but liberal regime public pensions pay little and workers rely more on private sector pensions that are lightly regulated.

Based on Gøsta Esping-Anderson's (1990) empirically grounded categorization of welfare state, Janoski (1998) constructs a theory of individualized or generalized citizenship regimes. He contends that in 'liberal regimes,' strong legal and moderate political rights protect property and business culture, and uphold a small welfare state. 'Social democratic regimes' tend to support worker organization, the labor force, and very strong welfare protection for all. And 'traditional regimes' have moderate-strong welfare states but they are closed off to protect these gains, and the society has a strong sense of mutual obligation, which lowers or distances interest in transnational human rights. Janoski (2010) followed this up with standardized measurements for these citizenship regime types based on nationality and citizenship laws. These were somewhat different regimes with colonizer (e.g., France and the UK), noncolonizer (e.g., Germany and Italy), settler (e.g., the US, Australia, Canada, etc.), and Scandinavian countries (e.g., Sweden, Norway, etc.).

But there are considerable challenges to measuring citizenship regimes when one moves to countries outside the traditional OECD group of nation-states. In China under Mao, citizenship was a layered construction favoring workers of the privileged state organized enterprises (SOE). These organizations basically operated as the arms of a welfare state, providing the full range of social rights including housing, food, education, and medical care – the 'iron rice bowl.' At the same time, political and legal rights were highly constrained and subject to highly negative sanctions with dissidents sent to jail or killed; and citizenship was segregated between rural and urban environments. In rural villages, citizenship was more akin to subjecthood, with farmers belonging to collectives and families fully supporting themselves without relying on the state at all.

With the opening up of China and its subsequent economic development, this bifurcation remained but in a highly modified form. The cities still have higher citizenship rights; however, the SOEs – except in the areas of energy and utilities – have largely collapsed and their former workers fight for pensions and wages. Nonetheless, wages and benefits have remained higher in the cities and much lower in the villages. The Chinese government has instituted the *Hukou* internal passport system to control the movement of people in the country. When villagers move to the urban areas to work in the factories of the specialized economic zones, their rights do not take effect in the cities. In practice, they become second-class citizens without many pension and housing rights, access only to a secondary form of emergency health care, and they cannot send their children to urban schools. Their situation largely resembles that of illegal aliens in the US, although civic resistance to these norms is growing in China.

In Africa where multiple 'citizenships' coexist, the boundaries separating legitimate members from the rest are murky, and the extension of rights to all individuals and groups may exist in theory but not in practice. This plays out in a scenario where citizenship is officially associated with national identity, but unofficially associated with an ethnic subpopulation. For instance, in many African nations, citizenship rights are extended to all country nationals but also to tribal groups – some of which hold considerable social and political clout. Consequently, two systems of power are institutionalized under a single authority: one urban, based on civil power and rights; and the other rural, where tradition and culture incorporate customary rule. What emerges are two simultaneous systems of direct and indirect rule that correspond to two kinds of citizenship.

This institutionalized bifurcated state, tends to fragment citizenship and civil society along class and ethnic lines, as well as along rural and urban divides (Mamdani, 1996). Rural, tribal dwellers may hold all legal rights of citizenship that correspond to that of their urban counterparts, but in practice, their 'tribal citizenship' may discourage/prevent their full participation in democratic nation-building processes. Similar sets of processes play out for Palestinians who are citizens of the Israeli state but largely prevented from exercising various political rights.

Citizenship as a Process or Activity

There is a moral aspect of citizenship that is somewhat differently conceived from citizenship as a status or an identity. Often, this moral framework refers to cultural notions of 'being a good citizen,' which has two meanings. First, there is an expectation that individuals obey the law and fulfill the required legal obligations that are associated with being a citizen. For instance, a good citizen is not a criminal, a terrorist, or a rebel who intends to bring down the state. The second meaning refers to cultural expectations of morally appropriate actions, such that a good citizen is someone who participates in voluntary associations; someone who volunteers to help his/her community; and someone who generally behaves in ways that better the community. In other words, citizenship is more than a state of being; it is also an action or a process of doing.

Engin Isin (2008, 2012) regards citizenship as a socialization process in the development of citizenship consciousness. Class, race, ethnicity, nationality, gender, family structure, or even place act as determinants of citizenship consciousness by shaping individual learning about voting practice, governance

structures, laws and obligations, and political participation. Isin et al. (2008) contest the conception of citizenship as merely a status; they widen it to include social struggles for recognition, or civic claims by groups and individuals to gain more rights. Citizenship becomes an outcome of interaction processes between those who belong and those struggling to gain access.

In the process of becoming a citizen through birth or naturalization, individuals learn and practice the norms and rituals of being a good citizen. Throughout this process it becomes clear that there rarely is a one-to-one, direct relationship between rights and obligations. For instance, the process of gaining rights may have no direct obligations because they are simply acquired at birth or at the age of majority. Moreover the required obligations are not always clearly distinguishable from rights (e.g., is voting a right or an obligation).

Consequently, citizens make individual choices about how to be citizens. For example, paying taxes is not a simple matter of a citizen fulfilling an obligation by writing out a check to the government. It also involves making many decisions about alternatives to taxation, such as investing in tax free municipal bonds or other tax-deferred accounts, avoiding taxes altogether by hiding money in offshore accounts, protesting about taxes, and so on. In other words, citizenship rights and obligations cannot be reduced to one little pinpoint in a spectrum of behaviors.

When citizenship is framed as a process, behavior and social interaction are moralized. This viewpoint extends the static theories of citizenship in constitutions and laws to include the social- and culture-shaping of citizenship. And as such, citizenship becomes a centrally strategic concept in the analysis of identity formation, civic participation, human rights, and engaged political life (Isin et al., 2008). Hence studying citizenship becomes study of these sorts of processes. Isin and Turner (2002) suggest that one should conceptually frame citizenship around three fundamental axes: extent, content, and depth. The nation-state demarcates the extent of citizenship through rules, laws, and other boundaries. Content looks at how rights and obligations are defined and combined into the actual process of citizenship. And the depth of citizenship refers to how much the practice of citizenship penetrates into the daily lives of individuals.

Citizenship and the Media in the Future

The discipline of sociology pays particular attention to the changing social landscapes of human action. With the rise in popularity of electronic social platforms, and the increase in technology-driven methods of communication, some sociologists began to theorize a new domain of electronic citizenship. According to Gavin Newsome (2013), who has developed the concept of 'citizenville,' individuals already supersede existing avenues of participatory citizenry by using their easily accessible mobile telephones and computers to record/access information and rapidly share it with people in their networks via Twitter and Facebook. He asks the question: "Why is it that people are more engaged than ever with each other – through Twitter, Pinterest, Facebook, text messaging – but less engaged with their government?" During the early Arab Spring movements starting in 2011, civil resistance was spurred on by thousands of individuals tweeting and filming live commentary from the frontlines using their mobile phones. They were engaging with people from all over the world in an online 'citizenville.'

Newsome's argument is that electronic communities approximate real-life citizenship, but in 'citizenville,' individuals have the added benefit of amassing real-life points by doing 'real-life good.' Newsome (2013) highlights five themes that characterize an electronic participation process: government becomes transparent; people use data to create useful apps, devices, and tools; digital aliens and immigrants are made into digital natives; people are allowed to bypass government by taking matters into their own hands; and the process of governing will become much more innovative and entrepreneurial. In the near future, we may more or less become the citizens of Facebook, YouTube, and Twitter.

In this new view of citizenship, there will be some casualties. Luddites and the technologically adverse, or those individuals and organizations whose identities are hacked or breached, will fall victim to the failure of citizenville to guarantee safety. But in actuality, the impact on the first two categories of people will be ignored; while the last one can be countered by having all individuals becoming defended by cloud- or crowd-sourced surveillance. According to this perspective, the political system becomes more voluntary as civil society grows fully interactive and transcends political and economic boundaries. In this decentralized, networked structure, the emergent political order is achieved through local politicians, like mayors, and not through top–down authorities like a president or executive bureaucracies.

However, one might look at this quite differently if we substituted 'computer technology' with 'genetically modified seeds' or 'cloning' for the web. A more serious problem is citizenville's lack of concern for corporations, who after all, tend to erect walls both geographically and economically to parse out how much money goes to owners and how much goes to citizens. Do we believe that these 'Goliaths' of ownership would just let these 'Davids' run the world?

See also: Assimilation of Immigrants; Civil Rights Movements; Communitarianism, Sociology of; Ethnic, Racial, and Nationalist Movements; Identity Movements; Identity Movements; International Migration; Nation, Sociological Aspects of; Nationalism, Sociology of; Political Sociology; State, Sociology of The.

Bibliography

Bauböck, R., 2003. Towards a political theory of migrant transnationalism. International Migration Review 37 (3), 700–723.

Bauman, Z., 2007. Liquid Times: Living in an Age of Uncertainty. Polity Press, Cambridge, UK.

Brubaker, W., 1992. Citizenship and Nationhood in France and Germany. Harvard University Press, Cambridge, MA.

Carey, A., 2010. On the Margins of Citizenship. Temple University Press, Philadelphia, PA.

Comaroff, J.L., Comaroff, J., 2009. Ethnicity, Inc. University of Chicago Press, Chicago.

Esping-Andersen, G., 1990. Three Worlds of Welfare Capitalism. Princeton University Press, Princeton.

Etzioni, Amitai, 1993. Spirit of Community. Crown, NY.

Fisher, J., 2008. We Will Be Citizens. Farland, NY.
Gates, H., Steele, C., Bobo, L., Dawson, M., 2012. The Oxford Handbook of African-american Citizenship. Oxford, NY.
Glenn, E., 2011. Constructing citizenship: exclusion, subordination and resistance. American Sociological Review 76 (12), 1–24.
Habermas, J., 1989. The Structural Transformation of the Public Sphere. MIT Press, Cambridge, MA.
Habermas, J., 1996. Between Facts and Norms. Polity Press, Cambridge, UK.
Held, D., 2010. Cosmopolitanism. Polity Press, Cambridge, UK.
Inglehart, Ronald, 1997. Modernization and Post-modernization. Princeton University Press, Princeton.
Isin, E., 2008. Theorizing acts of citizenship. In: Isin, E., Nielsen, G. (Eds.), Acts of Citizenship. Palgrave-Macmillan, NY, pp. 15–43.
Isin, E., 2012. Citizens without Frontiers. Bloomsbury, London.
Isin, E., Turner, B. (Eds.), 2002. Handbook of Citizenship Studies. Sage, Thousand Oaks.
Isin, E., Nyers, B., Turner, B., 2008. Citizenship between Past and Future. Routledge, London.
Janoski, T., 1998. Citizenship and Civil Society. Cambridge University Press, NY.
Janoski, T., 2010. The Ironies of Citizenship. Cambridge University Press, NY.
Joppke, C., 2005. Selecting by Origin. Oxford University Press, NY.
Kivisto, P., Faist, T., 2008. Dual Citizenship in Global Perspective. Palgrave-Macmillan, Houndsmills, UK.
Kymlicka, W., 1995. Multicultural Citizenship. Clarendon Press, Oxford.
Leydet, D., 2011. Citizenship. fall 2011 ed. In: Zalta, Edward (Ed.), The Stanford Encyclopedia of Philosophy http://plato.stanford.edu/archives/fall2011/entries/citizenship/.
Lister, R., 2003. Citizenship: Feminist Perspectives. Palgrave MacMillan, NY.
Lister, R., 2010. Citizenship, civil society and conflict: a gendered perspective. In: Baert, P., Koniordos, S., Procacci, G., Ruzza, C. (Eds.), Conflict, Citizenship and Civil Society. Routledge, NY, pp. 71–90.
Mamdani, M., 1996. Citizen and Subject: Contemporary Africa and the Legacy of Late Colonialism. Princeton University Press, Princeton.
Marshall, T.H., 1964. Class, Citizenship and Social Development. University of Chicago Press, Chicago.
Misra, J., King, L., 2005. Women, gender and state policies. In: Janoski, T., Alford, R., Hicks, A., Schwartz, M. (Eds.), The Handbook of Political Sociology. Cambridge University Press, USA, pp. 526–545.
Musick, M., Wilson, J., 2008. Volunteers. Indiana University Press, Bloomington.
Newsome, G., 2013. Citizenville. Penguin Press, NY.
Ong, A., 1999. Flexible Citizenship. Duke University Press, Durham.
Ong, A., 2006. Neoliberalism as Exception. Duke University Press, Durham.
Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge, MA.
Rawls, J., 1993. Political Liberalism. Columbia University Press, NY.
Roman, E., 2010. Citizenship and Its Exclusions. NYU Press, NY.
Smith, R., 1997. Civic Ideals. Yale University Press, New Haven.
Somers, M., 2008. Genealogies of Citizenship. Cambridge University Press, NY.
Turner, B., 1986. Citizenship and Capitalism. Allen Unwin, London.

City Marketing

John Rennie Short, University of Maryland Baltimore County, Baltimore, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since the 1970s city marketing has grown in size and importance. A historical contextualization considers the selling of the frontier city, the resort, the suburb, and the industrial city. The article explores the symbolic reconstruction of cities through a discussion of strategies and tactics of city marketing. Marketing strategies include the promotion of the postindustrial city, the global city, the business city, the good city, the green city, the cosmopolitan city, and the city of culture. The tactics of city marketing include namings, slogans, logos, and the construction of iconic buildings. The article discusses what city marketing reveals about social power in cities.

In 2011, officials from the Russian city of Yekaterinburgh submitted a nomination to the Bureau of International Exhibitions to host Expo 2020. It was part of a marketing campaign to rebrand the city that also included a renaming. During the Soviet years the city was called Sverdlovsk; it was a military–industrial city closed to foreigners with entry into and out of the city strictly controlled. It was a closed city not only to visitors but also to discursive representations. It was a supposedly secret city, hidden from international view and global scrutiny. The city's name change and Expo 2020 application embody a more general trend, the marketing of cities to achieve recognition, to change the internal and external image of the city, and to connect to a global flow of positive urban images and upbeat urban imaginaries.

The Historical Context

The marketing of cities has a long history. For just the past 150 years Stephen Ward (1998) identifies several dominant marketing campaigns: selling the frontier town, the resort, the suburb, and the industrial city. We can consider examples from each category.

Marketing campaigns were important for emerging towns in territories opened up for capitalist investment as urban boosters had to convince people that the city, despite its youth, small size, and eccentric location, was on the verge of greatness, increased growth, and vast capital returns. The city needed to project an image to a world that did not know of its existence. In the US West, for example, in the latter half of the nineteenth century, advertisements were placed in newspapers in the populous east of the country and in Europe to attract investors and immigrants. An 1887 promotion map of Wichita, Kansas, for example, uses a subtle spatial distortion to depict the city in the center of the country. We can extend the notion of the frontier town to the contemporary era with the idea of cities emerging from the silences and obscurity created by political and economic exclusion from the wider capitalist world. After the fall of communism in 1989 and the 'opening' up of Eastern Europe and China in the past 20 years, many cities in these countries were suddenly open to foreign investment opportunities and needed to announce their presence and position themselves to attract global flows of investment and to achieve a more positive position in the global flow of urban images. The more the number of cities that marketed themselves the more pressure was placed on the remaining cities in what is described as *place wars*. Shanghai, for example, was marketed by promoting its long association with the wider world, as a cosmopolitan city reconnecting with the global circuits severed by Maoism. Its long cosmopolitan past was promoted. Yekaterinburgh is located more than 1000 miles east of Moscow and only recently emerging from its closed city status. Like a US frontier town in the late nineteenth century, it is seeking to hold an international event. St Louis held the World Fair and the Olympic Games in 1902. Yekaterinburgh applied to host Expo 2020. Hosting international events, constructing new buildings, and initiating urban development plans are important parts of 'frontier' town promotion because they also signal urban sophistication and dispel perceptions of provincialism and unconnectedness. The city authorities of frontier towns such as Denver and Kansas City, desperate to show their sophistication, eagerly adopted the City Beautiful movement of the 1890s that stressed neoclassical architecture, grand vistas, and ambitious landscape designs. Here we can make a more general point. New architectural fashions such as City Beautiful, modernism, or postmodernism are not only shifts in design but also vehicles for city marketing as cities use the new fashions to signal their cutting edge character and the extent to which they are engaged with contemporary fashion and plugged into the most global circuits of innovative architectural design. Constructing a new building in the latest style is an important form of city marketing.

As mass tourism developed in the middle of the nineteenth century, a new category of consumer was created: those searching for play and recreation. Cities responded to the new market by selling themselves as sites of pleasure and entertainment. The city of Blackpool in England advertised itself through a variety of means including building Blackpool Tower to provide an iconic image and elaborate illuminations to extend the season beyond the summer into the longer nights of winter. As the tourist gaze has widened, more cities from around the world are brought into the range of possible destinations. Consider the example of Budapest that emerged from communist domination in 1989. The rebranding of the city from communist city to postcommunist city open for business and tourism was hampered by municipal

fragmentation; however, a Budapest City Identity Office was opened in 2006. Emphasis was placed on attracting tourists, and within that sector particular emphasis was placed on heritage tourism highlighting the historic nature of the city and its architectural legacy (Smith and Puczko, 2010; Szondi, 2011). Niche marketing of tourist cities reflects the variegated market demand. At one extreme are the child-orientated family-friendly destinations such as theme parks. At the other extreme, certain cities sell an image of adult sexual tolerance. Phil Hubbard (2012) documents a number of cases, including San Francisco, Sydney, and Cape Town, of the self-conscious marketing of a gay-friendly city with emphasis on safe gay spaces. For the past 10 years Las Vegas was marketed with the phrase, 'what happens in Las Vegas, stays in Las Vegas,' as an explicit attempt to target an adult market. From the promotion of the genteel spa towns of yesterday to the representation of uninhibited and available sexual pleasure of today's cities, the marketing of the resort city continues. It has widened so that resort cities compete for global markets, both intensifying the competition and homogenizing the set of marketing images for similar destination types. A standard repertoire of images includes beach and sun, play and recreation, escape and discovery, culture and learning, and freedom from the constraints of the ordinary and the banal. The image of the 'exotic' holiday destination also arises as the widening tourist range makes ever more distant cities more available. The global tourist gaze now evaluates urban destinations around the world.

Mass suburbs appear in the mid-nineteenth century. As a new urban form, a market for suburban living had to be created as much as met. Suburbs had to be sold to people long used to living close to the city. In the preauto, walking city, higher income groups lived closest to the center; the poor were banished to the outskirts, forced to travel further and longer. The suburbs represented a new form. While transport improvements shrunk distance, making suburban living more accessible, the newness of the suburbs required image control. The new suburbs of the wealthy needed careful attention. They were sold and marketed as places of exclusivity. To maintain the perception and reality, restrictive covenants were widely advertized and often strictly enforced. In a perceptive study Robert Fogelson (2005) examines the heavy restrictions in US middle and upper income suburbia from 1870 to 1930. He shows that restrictive covenants were a nationwide phenomenon as developers and homeowners sought to minimize any changes that would undermine land values, house prices, and social exclusivity. The restrictive suburbs were envisioned and marketed as a haven of peace and tranquility from all this threatening change.

While suburbanization is now a fact of life, marketing individual suburbs continues. Three themes can be identified that center on the idea of the suburbs as place of retreat. The first is the depiction of the suburb as rural idyll. A massive residential development along Britain's M4 growth corridor was initiated in the 1980s (Short et al., 1986). The development was subdivided into neighborhoods. Part of the marketing involved the self-conscious naming of these neighborhoods. Names such Badger's Walk, Miller's Green, Upper Wood, Meadow Vale, and Swallow's Meadow all sought to evoke the warmth and comfort of the rural preindustrial past. The names evoke a quiet and stable community in a region of noise and rapid change. The second theme, especially important in cities in developing countries, is the evocation of global connections to signal modernity and sophistication. New suburban developments in Shanghai, for example, are called Australia House, Scotland, and Manhattan. These British and United States inspired names represent a marketing of the modern, the international, and the cosmopolitan. Finally, there is a marketing of certain suburbs as gated communities. The gates signal privacy and security. They are a message meant to quiet fears as much as to initiate purchase. The development of gated communities in cities round the world is as much about marketing as actual security. The gates and the guards are meant to reduce urban anxieties of newly emerging middle and upper income groups.

The industrial city emerged in the early nineteenth century. Cities such as Manchester, England, were the crucibles for the creation of an urban industrial economy. Industry was long fixed in place, tied to sources of power, skilled labor, capital, and social networks. In the past 150 years the mobility of industry has increased, freed from the locational constraints of the past. Electricity provides a more dispersed energy, and deskilling and mechanization of the industrial labor process has reduced the pulling power and influence of skilled labor. A globalizing world is embodied in the growing mobility of industry that creates a space of opportunity for newer cities to market themselves as industrial cities. In the Untied States, for example, cities in the Sunbelt long sold themselves as cheaper places to do business. Atlanta has marketed itself to northern and international investors as a place to do business since the late nineteenth century. In a globalized economy the range of industrial cities has widened to include much of the developing world. Since 2003, the city of Ahmedabad, in Gujarat, India, is promoted through a biennial event, the *Vibrant Gujarat: Global Investors Summit*, which targets foreign direct investors, rapidly growing Indian companies, and on-resident Indians with US business interests. The summit aims to strengthen confidence in the state and the city (Nair, 2011).

The promotion of an industrial city revolves around a number of themes: the city as a probusiness environment of low taxes with a skilled, stable, and cheap labor force; good infrastructure; and locational advantages of easy access to regional, national, and global markets. Bailey (1989) suggests a three-stage evolutionary model in the marketing of the industrial city. The first stage, *smokestack chasing*, involves attracting industries on the promise of low operating costs and higher profits. The second, *target marketing*, involves the attraction of industries enjoying profitable growth. Finally, *product development*, folds the first two into a more ambitious strategy to attract the inward investment of cutting edge manufacturing. Marketing shifts through the stages from highlighting lower operating costs in the early stages to greater emphasis on skilled labor, global competitiveness, and quality of life in the later stages.

City marketing is now an important global phenomenon. In the rest of this article I explore the strategies and tactics of contemporary city marketing. A discussion of strategies allows us to answer the question of why cities are being marketed. The

analysis of tactics enables us to ask how cities are being marketed. I will discuss a number of examples in some detail and end with a critical discussion of evaluating effects and noting silences. These latter two points address the issue of what is the result of the marketing and what is left out.

Marketing Strategies

Cities are marketed in order to enhance, improve, and/or change their image, either internally or externally and sometimes both. The term branding is also used, although it is more usefully applied to a specific campaign for an individual city that marks it out as distinctive. There is now substantial literature on city marketing. David Gertner (2011) identifies over 259 articles produced by 280 authors in 43 different journals over the past 20 years.

While each individual city has a unique aim in mind, specific to the city, there are a number of more general strategies that can be identified across a range of cities. They are not mutually exclusive as many cities, especially the very largest cities, project a complex, and sometimes dissident, set of images to the wider world. I discuss seven basic strategies of marketing: promoting the image of the postindustrial city, the global city, the business city, the good city, the green city, the cosmopolitan city, and the city of culture. The divisions between them are fluid and porous. They are best considered as an assemblage of positive urban images that city boosters seek to represent. The most successful and ambitious cities seek to combine most of them, if not all, in one positive urban imaginary. Cities can be ranked by their ability to embody the largest number of elements of this urban imaginary assemblage.

The Postindustrial City

In the wake of deindustrialization of large swathes of urban North America, Australia, and Europe, formerly industrial cities were placed in the unenviable position of losing investment, jobs, and identity. The global shift of manufacturing creates new industrial cities in the developing world but leaves behind industrial cities in the developed world. These industrial cities, or to be more accurate, these newly postindustrial cities, became associated with the old, the polluted, the past, and the failed. A major theme of marketing these postindustrial cities is to distance them from their recent industrial past. Some of the rebranded postindustrial US cities include Pittsburgh and Syracuse. In the case of Syracuse, New York, this involved a new logo for the city that replaced images of smoking factories with postmodern building skyline and a transformation of a local lake from dump site to scene of regeneration and set of recreational opportunities (Short et al., 1993). Formerly industrial cities are rebranded in more attractive packages that emphasize the new rather than the old, the fashionably postmodern rather than the merely modern, the postindustrial rather than the industrial, consumption rather than production, and spectacle and fun rather than pollution and work. Take the case of Wollongong, an industrial city on the New South Wales coast of Australia. A massive steelworks dominated the urban economy. The steelworks shed 15 000 jobs from the early 1980s to the mid-1990s. The rate of job loss was only one major strand in negative imagery associated with the city in the national imagination and among foreign investors. City leaders decided to rebrand the city in the public imagination. An image campaign was built around the idea of 'innovation, creativity, and excellence' (Kerr et al., 2011). An important part of the campaign involved the steelworks planting half a million trees on its site while the council found funds to clean up the beaches and construct cycle ways. The greening of the city is now an integral part of a city's attempt to shed its hard industrial image for a softer postindustrial greener imagery.

The Global City

In a globalizing world the marketing of a city as a global city is now an important strategy. The intense and constant competition to attract advanced producer services is a crucial driver of city marketing campaigns. Because of the growing competition, there is constant need to upgrade and improve a city's image. At the top of the global hierarchy, as measured in terms of the concentration of advanced producer services, sits London and New York. Both cities actively position themselves as global cities. In 1965, New York hosted the World's Fair, and after a period of fiscal crisis and economic decline the city was rebranded in 1977 with the successful *I (heart) New York* campaign. In 2012, London hosted the Summer Olympics Games, and presented itself as a multicultural, cosmopolitan, cool city.

Kim and Short (2008) note that the rhetoric of global city branding is often associated with attempts to justify controversial and costly urban redevelopment projects, to justify tax incentives and business friendly economic policies, and to heighten multiculturalism and cosmopolitanism in the city.

Below the top tier of cities are a number of major cities that compete for dominance in their national or regional space. Darel Paul (2004, 2005) explores the politics of 'global imagineering' in Montreal, Canada, where business interests triumphed, and Minneapolis–St Paul, USA, where similar interests where trumped by populist politics. In St Petersburg, Russia, not only was the old Soviet name of Leningrad dropped but the city was also actively promoted as an international hub of circulatory capital, host to corporate power, and stage for globalist megaprojects. St Petersburg was rebranded as an entrepreneurial, economically competitive, globally connected city (Golubchikov, 2010).

From 1997 to 2007, Honk Hong was specifically marketed as 'Asia's World City' (Shen, 2010). In his critical assessment Stephen Chu (2011) argues that the marketing highlighted the global at the expense of Hong Kong's local creative industries and vernacular hybrid cultures. Emphasizing the global was at the expense of the local.

The Business City

In a capitalist world economy of mobile capital and rapid redirections of investment flows, cities need to promote themselves as probusiness and friendly to capital. Short and Kim (1998) examined the advertisements of US cities in the

business press. Many cities explicitly cited the term 'business' to reinforce the message. The following are some examples:

Dallas	The city of choice for business
Milwaukee	The city that works for your business
New York	The business city that never sleeps
Phoenix	Moving business in the right direction

As capital becomes ever more mobile, more cities enter the competition, so cities need to aggressively promote themselves as good places to do business. More specific themes around this general message include emphasizing a probusiness political climate, a skilled labor force, good infrastructure, and accessible location.

The Good City

It is not enough to market a city's low costs. Cheap labor, for example, can also be associated with poor urban conditions; so cities need to highlight their quality of life. The case study of magazine advertising by Short and Kim (1998) identified a number of specific quality of life themes. These included friendliness, readily available and good quality health services, heritage, cultural amenities, and low cost of living. Investors, tourists, and visitors need to be reassured that the city is a decent place to live, work, and visit.

The case for selling this message to tourists and visitors is self-evident. The needs of business are more complex. While businesses often require low operating costs to maximize profits, they also require stable social conditions, a skilled and happy workforce, and a range of good-quality urban public goods and services such as decent health care. There are also the more pragmatic arguments that business executives may need to follow the inward investment and live in the city. Promoting the city as a good place to live is an important business strategy as well as vital to secure visitors such as conventioneers and tourists.

The Green City

The green city has emerged in the past two decades as an important element in city marketing and urban promotion. To some extent the term green is a loose and wide adjective that indicates a variety of ideas and images including good air quality, low pollution, sustainable urbanism, and accessible and attractive green spaces. The greening of a city can be an important part of promoting the idea of global city since the acceptance in principle, if not in hard practical application, of a green urbanism is now widely shared. It has also become a signifier of development and sophistication. Poor air quality is a sign of low-grade industrialization and unfettered environmental pollution. A green city is now considered a competitive urban advantage. Mexico City, long troubled by poor air quality, now boasts three vertical gardens, giant green walls of 50 000 plants that create shade, absorb pollution, and oxygenate the air. They are the lungs of the city. The living sculptures are the result of a public–private partnership that embodies the new civic consciousness of creating a greener city. In Seoul, another megacity long afflicted with poor air quality, the city government created more green spaces, cheaper public transport, greater use of fuel-efficient cars, and constant monitoring and policing of emissions from factories, cars, and residences. The result is not only a vast improvement in air quality but also a repositioning of Seoul as a global greening city.

The green image takes a number of forms. Seoul's relentless marketing of itself as a global city now incorporates a specifically green theme. It is part of what Blaž Križnik (2011) describes as the *symbolic reconstruction of the city* to stress tourist attractions of global competitiveness. A central part of the strategy was the reconstruction of Cheonggyecheon, an 11-km stream covered up by rapid urban expansion; it was the site of an elevated motor expressway. In 2002, the city government announced plans to demolish the expressway and reconstruct the stream as a green corridor. It was part of a downtown redevelopment plan that also signified a green and attractive Seoul, a global city with the requisite green credentials to go up against the leading cities in the world, part of what Ong (2011) describes as worlding practice.

The Cosmopolitan City

In a globalizing economy it is important to widen the appeal of city and the global competiveness of its goods and services. Cosmopolitan cities appeal to a wider range of investors, migrants, and tourists. The fostering and representation of a cosmopolitan city is now an integral part of the global city imaginary, alongside a probusiness environment and good infrastructure. Cosmopolitanism is now a vital part of the marketing of wannabe global cities. Some cities, such as London and New York, with large immigrant populations, have an inbuilt advantage. Others have to work at it. In Seoul, for example, the reality of a racially homogeneous city and nation meant that a city marketing campaign in 2009 under the slogan of 'Infinitely Yours' had to stress cultural diversity. Tolerance is now seen not only as a civic virtue but also as an important element in the formation and reproduction of the creative city. Marketing a city as cosmopolitan connects with issues of creativity and highlights the ability of the city to generate and attract the leading edge of the cultural creative economies.

Successful cities need to draw upon the global talent pool of labor, especially for the advanced producer services. Selling the city as a tolerant cosmopolitan place makes it easier to attract such labor. Some cities use cosmopolitanism in their marketing. Sheffield in England uses the slogan 'Where everyone matters' (Paganoni, 2012), Amsterdam explicitly used an English rather than a Dutch slogan, 'I am Amsterdam' to signal its linguistic openness (Kavaratzis and Ashworth, 2007), and Trieste promoted multiculturalism during its bid to host the 2008 World Expo (Colombino, 2009).

The Cultural City

An important part of the assemblage of the urban imaginary is the notion of the city as a place of culture, a site of cultural sophistication and a place of cultural practices. Perhaps the most famous case study is the successful rebranding of the city of Bilbao, an old industrial city in the Basque region of Spain,

after the opening of the Frank Gehry-designed Bilbao Guggenheim Museum. The so-called Bilbao effect is the uptick of tourists and investment in the wake of cultural promotion. We also have examples of less successful rebranding such as the case of Rotterdam, which hosted the Cultural Capital of Europe in 2001 yet failed to shift its image as working city (Richards and Wilson, 2004), or the relative failure to promote the cultural image of Milan (De Carlo et al., 2009).

The Tactics of City Marketing

There is now a large literature on the tactics of city marketing, the how of selling a city (Ashworth and Kavaratzis, 2010; Braun, 2008, 2012; Dinnie, 2011; Hanna and Rowley, 2012; Moilanen and Rainisto, 2009). We now have detailed case studies of individual cities ranging from Astana (in Kazakhstan) (Skok Hobbs, 2009), Bogota (Kalandides, 2011), Cleveland (Smith, 2011), New York (Greenberg, 2008), Taipei City (Karvelyte and Chiu, 2011); thematic studies, such as the marketing of smaller cities (Baker, 2012); the role of communication technology and city marketing (Gasco-Hernandez and Torres-Coronas, 2009); and the marketing of the cultural creative city (Harris and Williams, 2011; Haddock, 2010; Kennell, 2010). From this large body of work I will touch on only a few themes.

Marketing can involve namings. In the late eighteenth century the New Town of Edinburgh in Scotland was constructed. In the wake of the Jacobite Revolution against the British Crown, the good burghers of Edinburgh wanted to sell the city as an integral part of the British state. The main thoroughfares were named Queen Street, George Street, Princess Street, and Charlotte Square to forge a direct link the with royal family. In case anyone doubted the city's loyalty, one street was named Hanover Street after the royal family's German lineage.

Today, while naming remains important, the standard tactics of city marketing include the development of a media campaign, and the creation of a slogan or tagline and a logo or set of images. There are now companies that specialize in branding and city branding in particular. Agencies such as Brand Forensics, Corporate Edge, The Hamiltons, and Wolff Olins provide a one stop shopping for cities wanting to brand and rebrand themselves.

Key words are also employed in the slogan adapted by some cities. Here are some from Britain:

Dundee: changing for the future.
A safer, cleaner, ambitious Nottingham: a city we're all proud of.

While the slogan for Dundee is punchy and succinct, the one for Nottingham seems too long and disjointed. Slogans are often used in association with a logo, a visual punch line. The best known logo in the history of city marketing is the 'I (heart) NY' designed by Milton Glaser in 1977. The combination of the emotive heart sign and the geometric letters made for a winning design that was an important part of the successful turnaround in the rebranding of the city. Koller (2008) identifies at least six different types of logos based on different combinations of words and images. The city of Melbourne, for example, employed the design company Landor Associates to come up with a new logo to replace the confusing array of images employed to sell the city; their answer was to use the letter 'M' in variety of graphic designs to embody a singular city with multiple attractions. Short et al. (1993) looked at the rebranding of Syracuse when an image of smoking factories was replaced by an outline of a postindustrial, postmodern city. The slogan and the logo are most successful when they are part of an overall narrative of a marketing campaign. Thomas Sevcik (2011) suggests that city campaigns tend to use the same methods used to sell products, yet cities are more complex and need a more guiding narrative.

The outlets of city marketing include the traditional venues of print and television and increasingly the Internet. Peter Dobers and Anette Hallin (2009) discuss the use of the Web by city managers in Stockholm to promote the image of the city as 'The capital of Scandinavia.' Carl Grodach (2009) highlights the importance of promotional imagery on city homepages.

The marketing of cities can also involve the construction of buildings. Paul Knox (2012) writes about the importance of big-named architects, so-called starchitects, and the centrality of starchitecture to the symbolic capital of cities. To have a new building by a starchitect denotes cultural heft and global intentionality. Individual buildings and assemblages of certain built forms – museums, art galleries, new airports, and the like – built in the latest style by the hottest architects is an important way for a city to position itself. Perhaps the best known examples of branding through building are the Gulf States including Dubai and Abu Dhabi (Bagaeen, 2007; Kanna, 2011).

The hosting of events is now a significant part of marketing a city. So-called spotlight events highlight the city for a wider audience: the larger the event the greater the potential marketing opportunity. I have discussed at length the role of World fairs and Olympic Games in marketing cities (Short, 2012).

The hosting of events – such as film festivals, Expos, and Olympic Games – provides opportunities for a city's elites to stage a widely accessible and globally pervasive marketing campaign. In a detailed case study, Clair Colomb (2011) explores the marketing of Berlin between 1989 and 2011. She argues that the city was marketed through rebuilding and the staging of a new Berlin. In the inner city, sites were turned into tourist attractions. Place marketing and urban reinvention were intertwined in a symbolic politics of representation.

Most successful marketing campaigns involve some form of stakeholder involvement. On the surface this seems self-evident: get as many different interests as possible to buy into the selling of the city. City marketing just to outsiders fails to mobilize local support. However, stakeholder involvement can also generate conflict. Craig Hayden and Efe Sevin (2012), for example, highlight how the marketing of Ankara represents a collective representation that reveals conflicts and fractures well as shared ideas. Ooi and Pedersen (2010) show how the organizers of the Copenhagen Film Festival had different agendas from the brand authorities for the city, and Michele Acuto (2012) highlights the problem in successfully marketing the city of Sydney because of municipal fragmentation and the lack of a metropolitan authority.

The big question is, how affective are city marketing campaigns? There are obvious success stories. One of the most

cited is Barcelona. The success in attracting visitors and creating a positive buzz about the city is in part attributed to good marketing. In the wake of the successful hosting of the 1992 Olympic Games and the promotion of a cultural and especially an architectural legacy, the city is now perceived as cool and accessible, modern with an impressive cultural legacy. However, as Montserrat Pareja-Eastaway (2009) argues, whether this is due to marketing or the embracing and realization of long-term objectives by local actors is a moot point. This raises a more general issue. It has proved difficult to measure the independent role of city marketing in rebranding a city's image and turning around a city's economy. While we have some rough metrics (the Anholt-GfK Roper City Brands Index is based on 20 000 respondents in 20 countries responding to questions about cities), we have few detailed case studies. As one executive is alleged to have said, "I know that half of my advertising budget is a waste of money, I am just not sure which half."

Growing competition ensures that city élites need to be proactive in attracting and generating economic activity. City marketing may or may not work and may be good value or a waste of money, but in an increasingly competitive world it makes little difference. As more cities are marketed, pressure is placed on all cities to be marketed.

There is a growing body of critical work that highlights the silences of city marketing. The emphases of most marketing campaigns are on the city as a place of business, less on social inclusivity, more on the profitable city, and less on the fair and just city. The dominant themes represent the power of the dominant groups. The more marginal are excluded. In an interesting twist on this finding, however, Hernandez and Lopez (2011) suggest how slums can be used to positively brand a city, in this case Bogota. They suggest that highlighting rather than ignoring the city's slums assumes that one can invoke positive responses around ideas of genuine difference, cultural authenticity, and vernacular architecture. The poor and the marginal become less a political threat and more a marketing opportunity. In this case the ethos of marketing is so pervasive that even the silences are uttered and the problems exposed, not as issues to be solved or addressed, but as opportunities to sell and brand. The basic conclusion is that while it is questionable if marketing changes the image of a city, what is not questionable is that marketing represents the success of certain groups, especially political and economic elites, in having 'their' city represented and promoted in a manner friendly to their interests.

Bibliography

Acuto, M., 2012. Ain't about politics? The wicked power-geometry of Sydney's greening governance. International Journal of Urban and Regional Research 36, 381–399.

Ashworth, G., Kavaratzis, M. (Eds.), 2010. Towards Effective Place Brand Management: Branding European Cities and Regions. Edward Elgar, Northampton, MA.

Bagaeen, S., 2007. Brand Dubai: the instant city; or the instantly recognizable city. International Planning Studies 12, 173–197.

Bailey, J.T., 1989. Marketing Cities in the 1980s and beyond. Amercian Economic Development Council, Chicago.

Baker, B., 2012. Destination Branding for Small Cities, second ed. Creative Leap, Tualatin, OR.

Braun, E., 2008. City Marketing: Towards an Integrated Approach. PhD, Thesis. Erasmus School of Economics. Erasmus University Rotterdam.

Braun, F., 2012. Putting city branding into practice. Journal of Brand Management 19, 257–267.

Chu, S.Y., 2011. Brand Hong Kong: 'Asia's world city' as method? Visual Anthropology 24, 46–58.

Colomb, C., 2011. Staging the New Berlin. Routledge, London.

Colombino, A., 2009. Multiculturalism and time in Trieste: place-marketing images and residents' perceptions of a multicultural city. Social and Cultural Geography 10, 279–297.

De Carlo, M., Canali, M.S., Pritchard, A., Morgan, N., 2009. Moving Milan toward Expo 2015: designing culture into a city brand. Journal of Public Management and Development 2, 8–22.

Dinnie, K. (Ed.), 2011. City Branding: Theory and Cases. Palgrave, New York.

Dobers, P., Hallin, A., 2009. The use of internet in building the brand of "Stockhol: The Capital of Scandanavia". In: Gasco-Hernandez, M., Torres-Coronas, T. (Eds.), Information Communication Technologies and City Marketing: Digital Opportunities for Cities Around the World. Information Science Reference, London and Hershey, PA, pp. 265–294.

Fogelson, R., 2005. Bourgeois Nightmares: Suburbia, 1870–1930. Yale University Press, New Haven.

Gasco-Hernandez, M., Torres-Coronas, T. (Eds.), 2009. Information Communication Technologies and City Marketing: Digital Opportunities for Cities Around the World. Information Science Reference, London and Hershey, PA.

Gertner, D., 2011. A (tentative) meta-analysis of the 'place marketing' and 'place branding' literature. Journal of Brand Management 19, 112–131.

Golubchikov, O., 2010. World-city-entrepreneurialism: globalist imaginaries, neoliberal geographies and the production of new St Petersburg. Environment and Planning A 42, 626–643.

Greenberg, M., 2008. Branding New York: How a City in Crisis Was Sold to the World. Routledge, New York.

Grodach, C., 2009. Urban branding: an analysis of city homepage imagery. Journal of Architectural and Planning Research 26, 181–197.

Haddock, S.V. (Ed.), 2010. Brand Building the Creative City: A Critical Look at Current Concepts and Practices. Firenze University Press, Firenze, Italy.

Hanna, S., Rowley, J., 2012. Practitioners views on the essence of place brand management. Place Branding and Public Diplomacy 8, 102–109.

Harris, J., Williams, R. (Eds.), 2011. Regenerating Culture and Society: Architecture, Art and Urban Style within the Global Politics of City Branding. Liverpool University Press, Liverpool.

Hayden, C., Sevin, E., 2012. The politics of meaning and the city brand: the controversy over the branding of Ankara. Place Branding and Public Diplomacy 8, 133–146.

Hernandez, J., Lopez, C., 2011. Is there a role for informal settlements in branding cities? Journal of Place Management and Development 4, 93–109.

Hubbard, P., 2012. Cities and Sexualities. Routledge, New York.

Kalandides, A., 2011. City marketing for Bogota: a case study in integrated place branding. Journal of Place Management and Development 4, 282–291.

Kanna, A., 2011. Dubai: The City as Corporation. University of Minnesota Press, Minneapolis.

Karvelyte, K., Chiu, J.-H., 2011. Planning process of city brands: a case study of Taipei City. Place Branding and Public Diplomacy 7, 257–270.

Kavaratzis, M., Ashworth, G., 2007. Partners in coffeeshops, canals, and commerce: marketing the city of Amsterdam. Cities 24, 16–25.

Kennell, J., 2010. Simulated Cities: Cultural Regeneration, Branding and Representation in Urban Development. Lambert, Saarbrucken, Germany.

Kerr, G., Noble, G., Glynn, J., 2011. The city branding of Wollongong. In: Dinnie, K. (Ed.), City Branding; Theory and Cases. Palgrave Macmillan, New York, pp. 213–220.

Kim, Y.-H., Short, J.R., 2008. Cities and Economies. Routledge, London.

Knox, P., 2012. Starchitects, starchitecture and the symbolic capital of world cities. In: Derudder, B., Hoyler, M., Taylor, P.J., Witlox, F. (Eds.), International Handbook of Globalization and World Cities. Edward Elgar, Northampton, MA, pp. 275–283.

Koller, V., 2008. "The word in one city": semiotic and cognitive aspects of city branding. Journal of Language and Politics 7, 431–450.

Križnik, B., 2011. Selling global Seoul: competitive urban policy and symbolic reconstruction of cities. Revija Za Sociologiju 3, 291–313.

Moilanen, T., Rainisto, S., 2009. How to Brand Nations, Cities and Destinations: A Planning Book for Place Branding. Palgrave, New York.

Nair, S.K., 2011. The city branding of Ahmedabad. In: Dinnie, K. (Ed.), City Branding: Theory and Cases. Palgrave Macmillan, New York, pp. 106–111.

Ong, A., 2011. Worlding cities, or the art of being global. In: Roy, A., Ong, A. (Eds.), Worlding Cities: Asian Experiments and the Art of being Global. Wiley-Blackwell, Chichester, pp. 1–26.
Ooi, C.-S., Pedersen, J.S., 2010. City branding and film festivals: re-evaluating stakeholder's relations. Place Branding and Public Diplomacy 6, 316–332.
Paganoni, M.C., 2012. City branding and social inclusion in the global city. Mobilities 7, 13–31.
Pareja-Eastaway, M., 2009. The Barcelona metropolitan region: from non-existence to fame. Built Environment 35, 212–219.
Paul, D.E., 2004. World cities as hegemonic projects: the politics of global imagineering in Montreal. Political Geography 23, 571–596.
Paul, D.E., 2005. The local politics of 'going global': making and unmaking Minneapolis–St. Paul as a world city. Urban Studies 42, 2103–2122.
Richards, G., Wilson, J., 2004. The impact of cultural events on city image: Rotterdam, Cultural Capital of Europe 2001. Urban Studies 41, 1931–1951.
Sevcik, T., 2011. Strategic urban narratives: beyond conventional city branding. Development 54, 343–344.
Shen, S., 2010. Re-branding without redeveloping: constraints of Hong Kong's 'Asia's World City' brand (1997–2007). The Pacific Review 23, 203–224.
Short, J.R., 2012. Globalization, Modernity and the City. Routledge, London.
Short, J.R., Benton, L.M., Luce, W.B., Walton, J., 1993. Reconstructing the image of an industrial city. Annals of Association of American Geographers 83, 207–224.
Short, J.R., Fleming, S., Witt, S., 1986. Housebuilding, Planning and Community Action: The Production and Negotiation of the Built Environment. Routledge & Kegan Paul, London.
Short, J.R., Kim, Y.-H., 1998. Urban crises/urban representations: selling the city in difficult times. In: Hubbard, P., Hall, T. (Eds.), The Entrepreneurial City. Wiley, Chichester, pp. 55–75.
Skok Hobbs, T., 2009. A World-Class City in the Middle of the Steppe: Place Marketing and the Construction of an Image of Place in Astana, Kazakhstan. PhD, Thesis. Department of Communication. University of Ottowa.
Smith, D.W., 2011. Place Marketing and the Image of Cleveland and Northeast Ohio. MA, Thesis. Department of Geography. Kent State University.
Smith, M., Puczko, L., 2010. Out with the old, in with the new? Twenty years of post-socialist marketing in Budapest. Journal of Town and City Management 1, 288–299.
Szondi, G., 2011. Branding Budapest. In: Dinnie, K. (Ed.), City Branding: Theory and Cases. Palgrave Macmillan, New York, pp. 124–130.
Ward, S.V., 1998. Selling Places: The Marketing and Promotion of Towns and Cities, 1850–2000. Routledge, New York.

Civil Law

Peter Hay, Emory University School of Law, Atlanta, GA, USA; and Technische Universitaet Dresden, Dresden, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The civil law is one of the large legal families in the world. It is widespread in Europe, Central and South America, and in many parts of Asia and Africa. It differs from other legal families not only in its defining history but also in its structure, application, and development. This article provides an introduction to some of its significant aspects and to its continued influence on the development of the law in various parts of the world. It also notes some of the differences in legal education and professional practice.

Structural Aspects

In the comparative study of law, legal systems are often seen as belonging to 'legal families.' These may be religious-based legal systems (for instance, Islamic and Judaic laws), they may be geographically or historically defined (for instance, tribal law on United States Indian reservations and the Roman law origins of Roman–Dutch law in South Africa), or they may be distinct because of their structure and methodology (the civil law and the common law) (see Zweigert and Kötz, 1998, pp. 63–73).

In particular, scholars frequently contrast the families of the common law and the civil law, which differ in terms of structure and methodology. This contrast holds despite the fact that both legal families have their roots in the Corpus Iuris Civilis of Justinian (AD 534, see *infra*). The common law has preserved some of these traditions to this day, while developing its own methodology. At the same time, the civil law has moved closer to the common law in some areas. Hence, the contrast may not be as stark today as it once was.

With respect to both families, it is dangerous to generalize: There can indeed be marked differences between legal systems belonging to the same 'family.' Thus, the use of codes (e.g., the Uniform Commercial Code (UCC)), the existence of a large body of statutory law, and the adoption of the influential Restatements (the summaries of prevailing case law, stylistically in statutory form, prepared by the American Law Institute, a private organization) may make the legal system of the United States appear closer to the civil law than to its English roots (see Farnsworth, 1996, p. 227). Such a conclusion, however, would misconceive the role and function of statutory law and of the methodology in its application in the civil law, on the one hand, and in common law legal systems on the other hand.

Codes and statutes consist of rules in all legal systems. These may be very specific, or they may be stated with some degree of generality. In some areas, common law codes and other legal texts display extraordinary detail, attempting to anticipate every eventuality. The common law judge, as a consequence, will begin with the factual circumstances of the case, and after examining, comparing, and weighing the factual elements of the case, will attempt to find a rule that fits it. This rule may be one of statutory law or derive from prior decisional law, including decisions authoritatively and bindingly interpreting and applying statutory law – the doctrine of *stare decisis* (Hay, 2010, p. 9, also *infra*). The civilian judge, in contrast, sees the case as a problem to be solved within the legal structure of the legal system. First, the problem at hand must be fitted into a legal category. Next, subcategories and sub-subcategories must be identified until 'legal rule' or 'concept' and the problem at hand match. Civilians proceed by deductive reasoning, while the common law approach employs an inductive methodology.

Categorization, as a structural characteristic of the civil law, also results in the drawing of sharp distinctions between different areas of the law. Private law deals with legal problems that arise between natural persons or between natural and legal persons (such as corporations), while public law (for instance, constitutional law, administrative law, and criminal law) addresses the relationship between citizen and state. In some civil law systems, this bifurcation results in the establishment of special courts to deal with these different areas of the law. In Germany, for instance, there are so-called ordinary courts with competence for private law and criminal law, and separate court systems with competence for questions of administrative law, labor law, and social law, with each system having its own supreme court. Since many cases will obviously involve mixed questions (e.g., of labor and private contract or tort law), there may be contradictory rules of law, with no unifying supreme court. In France, the *Conseil d'Etat* stands beside the *Cour de Cassation*, with exclusive competence in certain public fields of law. The exclusivity of the *Conseil's* competence avoids some of the German problems. Germany and Italy, for example, have constitutional courts, while most other civil law countries do not. While the divisions are more pronounced in Germany than elsewhere, categorization of fields of law and the establishment of specialized courts unavoidably leads to a high degree of specialization within the legal profession, among both lawyers and judges. These lines are much more fluid in common law countries.

The central role of the legal rule in civil law legal systems also explains why precedent (i.e., the binding effect of a prior court decision on a different, but similar subsequent case) is quite different from what it is in the common law. The civilian judge applies the law but is not bound (obviously with some exceptions) by earlier decisions of a higher court. In common law countries, in contrast, it is the decision of the highest court that ultimately is the law and therefore binds inferior

courts – the doctrine of *stare decisis*. Subsequent discussion will return to these points.

History

Roman law had no comprehensive codification before Justinian's Corpus Iuris Civilis. Law, such as it was, was divided into *iuscivile* and *iusgentium*. The former applied among Roman citizens; the latter applied to legal relationships among others (Romans and foreigners, foreigners among themselves, and slaves). Judicial functions were exercised by *praetores*. The *praetor peregrinus* administered the *iusgentium* and conceived and developed legal concepts unknown to the strict Roman *iuscivilis*. In a different context, of course, the English chancellor – at the beginning of the equity jurisprudence – performed equally creative functions. Similar ideas – for equity and judicial creativity – underlie the 'general clauses' of the civil law (see *infra*, p. 9).

With population growth and increasing urbanization, a new profession – that of the jurisconsult – arose. They were legal advisors who prepared written opinions for cases. As the number of opinions grew, principles could be derived from them that could be taught to students and serve as a basis for advice to judges. The development of principles fosters more abstract ways of thinking. Categories are conceived, and problems are classified for assignment to various categories. Under Emperor Justinian, opinions, decisions, and other materials were gathered and the Corpus Iuris Civilis was prepared.

The Corpus Iuris stands at the beginning of what Roman law means today. This code evolved in subsequent times as a result of the work of the glossators and commentators. These sources taken together constituted the 'Ius Commune,' the common law of Europe before countries and areas began to grow in different ways and today's 'different families' began to take shape. England developed the common law (although, especially in the early period, with many Roman law elements). The Scandinavian countries, France, and Germany had a leading role in developing the civil law. Once again, 'civil law' is not homogeneous. There are differences among civil law countries, and these differences are also reflected in the legal systems of those countries that modeled their codes or statutes after earlier rules of others. Two codifications in particular replaced the earlier 'Ius Commune' and proved very influential: the French code civil of 1804 and the German *Bürgerliches Gesetzbuch* of 1896 (see also the Austrian *Allgemeines Bürgerliches Gesetzbuch* of 1811). The French code not only influenced the development of legal systems in many European (especially Latin and Eastern) countries but also spread to the Near East, Central and South America, and even to parts of North America (e.g., Louisiana). The German code influenced the law of Eastern and Southern Europe (e.g., Hungary, the Czech Republic, Slovakia, former Yugoslavia, the Baltic States, and Greece) as well as of Japan and China (see Zweigert and Kötz, 1998; pp. 154–156). In 2010, the People's Republic of China promulgated a new law (effective in 2011) on the 'Application of Laws to Foreign-Related Civil Relations,' which mirrors modern European codification on private international law in many respects.

Roman law still exists in Southern Africa as Roman–Dutch law, an admixture of Roman law and old Dutch customary law that interacted with English common law (see Zweigert and Kötz, 1998, p. 232). Codification not only extended, in these and in other countries, to private law and private-law relationships but also encompassed criminal law, commercial law, and both civil (private) and criminal procedures.

Sources of Law

Primary Sources

Civil law systems draw a sharp distinction between primary and secondary sources. Primary sources are enacted law, custom, and 'general principles of law.' Of these, the main source is the enacted (statutory) law; it predominates in civil law systems.

A code in a civil law system consists of general principles, arranged in order of importance. At the beginning, there may be general rules regulating basic problems that need to be addressed before the particular problem can be analyzed. For example, if a plaintiff seeks damages for breach of contract, preliminary analysis must determine whether a contract was validly concluded. Provisions dealing with invalidity and avoidance of contracts usually are found in the general part of a civil code. Such a general part may be followed by particular parts dealing with individual fields of law, such as torts, contracts, property, or the law of succession. The main or basic codes are supplemented in increasing number by special statutes or codes of limited coverage with which the legal system reacts to new societal problems. Examples include consumer protection, telecommunication, and news media.

Custom is also a primary source of law. Customs are unwritten rules, developed and observed over years and now part of social and economic thinking. Customary law tends to be less important in practice because it is often difficult to prove its pervasive observance in society.

'General principles of law' are what the term expresses: Basic principles of the legal system that pervade it and derive from norms of positive law. Civil law judges resort to 'general principles of law' as guidelines in the interpretation of statutory norms both for the purpose of defining their interrelation and for their application. This is of particular importance when dealing with statutory norms that are rather abstract in their formulation. It is tempting to consider this process to be not very different from the case law methodology of the common law. There is an important difference, however. The common lawyer derives the appropriate rule by reliance on precedent and its interpretation and, only lacking precedent, resorts to general principles or to policy considerations. The civilian judge is not so restricted but derives the appropriate rule from the structure of the legal system and from the general principles of law that pervade it; nor will the decision in the present case have a necessary effect on later cases. This is not to say that later cases may not reach the same conclusion: At the point when decisions consistently reach the same result – what French lawyers call *jurisprudence constante* and German lawyers *ständige Rechtsprechung* – such decisional law may itself be regarded as having risen to the position of 'general principles of law.'

Secondary Sources

In the civil law, secondary sources consist of case law (see section Primary Sources) and the legal literature. The legal literature consists of monographs and contributions to the legal periodical literature as well as commentaries. The last are particularly important in civil law countries that follow or are close to the Germanic legal tradition. Commentaries are detailed annotations of each provision of a particular code, consisting of an analysis that brings together all case law dealing with this provision, opinions of others as expressed in the periodical literature, and the commentator's own evaluation and summary.

Case law by itself, as already mentioned in other contexts, does not have the same central importance in a civil law system as it does in a common law system. It is indeed a secondary source, for the judge is bound only by the enacted law, except in the few cases in which – as discussed before – the decisional law has reached the level of *jurisprudence constante*, or in systems in which a 'constitutional court' has power to bind other courts (e.g., Germany). To repeat, however, such cases are rare. Judges will read commentaries and the legal literature in general, just as lawyers do. The process of 'law finding' and its application, however, does not restrict the court to these sources.

Developing the Law

If 'law' is a norm in the form of a statutory codification or a 'general principle' derived from such statutory norms, how can a judge decide a case for which no norm or general principles exist, and, in the absence of legislative action, how can law develop further? (For comparative discussion see Adriaansen, 1998, p. 107; Hazard and Dondi, 2006)

Legislative action is of course the classic instrument for legal change. However, the process, from the drafting of a bill until its ultimate passage and entry into force, can be long. A commission for the revision of the German law of obligations, for instance, worked on this project for more than a decade. The final report of the commission was published in 1992. On January 1, 2002, the so-called Act on the Modernization of the Law of Obligations (*Gesetz zur Modernisierung des Schuldrechts*) came into force. The Act is considered the most important reform of the German Civil Code (*Bürgerliches Gesetzbuch* – BGB) since it came into effect in 1900. Thus, problems may need to be addressed that existing legal norms do not cover. The Swiss Civil Code is unique in its candid grant of discretion to the judge to fill the gap. It provides in Art. 1, §2: "if the Code does not furnish an applicable provision, the judge shall decide in accordance with customary law, and failing that, according to rule which he would establish as legislator." The assumption must be, of course, that even with such a grant of authority, the judge will try to fashion a result that conforms to the general structure and tradition of the legal system. The result will not contravene existing conceptions of what the law should be, but rather will fill the gap, and thereby contribute to the evolution of the law. One way of doing this is to work by analogy, that is, to compare the present problem with other problems that similarly require a weighing of the interests of the parties in the dispute. The analogy then extends the balance struck with respect to other problems of the present case. An example is the treatment of equipment leasing in German law. The German Civil Code contains provisions on leases and for sales (Section 535 et seq. for leases and Section 433 et seq. for sales), but not for leasing. Since leasing is regarded as possessing elements from both fields of law, case law and literature now derive rules from both and apply them to the phenomenon of leasing (even though in most cases the rules for leases are applied primarily, see Graf von Westphalen and Thüsing, 2012, Part 'Klauselwerke,' III, paras. 16–17, 24–26).

The Swiss provision just quoted confers discretion. However, it was suggested that no court will exercise unfettered discretion but will seek to follow the structure and implement the values of the particular legal system. This has also been the experience in other civil law legal systems, for instance, the German. The German Civil Code, as do others, contains provisions of great breadth and, consequently, lack of specificity. These are 'general clauses,' of which Section 242 of the German Civil Code is perhaps the best example. It requires parties to a contract to perform their obligations in 'good faith.' Section 242 has become the source of an extensive body of case law and of legal concepts not addressed specifically in the civil code. When the economic upheavals of the 1930s threatened parties to contracts with economic ruin, the German Supreme Court invoked § 242 to develop a doctrine of 'frustration of contract' (*Wegfall der Geschäftsgrundlage*) that permits courts in appropriate cases to adjust the obligations of the parties. Good faith, so the German court stated, requires the protection of both parties: Cancellation of the contract (by analogizing the situation to one of impossibility) would be unfair to one party; enforcing it in full (against the background of totally different and unexpected circumstances) would be similarly unfair to the other party. The French *Conseil d'Etat* – but not the *Cour de Cassation* – developed a similar remedy, the doctrine of *imprévision* (see Waline, 2010, paras. 452–455). It is interesting that a similar doctrine (and a remedy akin to the German) were never accepted to that extent in the United States despite the greater freedom of common law courts to fashion law through decisions and despite an express provision in the UCC imposing an obligation to deal in good faith (UCC Section 1–203).

The civil law doctrine of *culpa in contrahendo* furnishes another example. The obligations of parties to a contract begin with its conclusion, but expectations will have been created earlier, and their disappointment may cause damage. Such a situation could be dealt with under the law of *Delikt* (tort), but sometimes the preconditions for a remedy in tort will be lacking. Again, the extension of an existing provision may provide relief. Section 276 of the German Civil Code provides for liability for damage caused intentionally and negligently. For liability to arise, there must be some relationship, some duty owed to the other party. Such a relationship, the case law states, again invoking the parties' good faith obligation, is created by the contract negotiations. Parties thus have a precontractual duty of care to each other and may have a remedy for its breach. This is the doctrine of *culpa in contrahendo*. The common law again is less flexible; it generally does not impose contract law-based liability for precontractual misconduct; when it does not, a remedy must be sought in tort for negligent or intentional infliction of damage (Hay and Rasmussen-Bonne, 2011, p. 108).

The civilian judge, despite the different theoretical structure and focus of the legal system, thus can and does display the same creativity as do common law judges, with the principal – perhaps only – distinction that the decision in the individual case does not 'make law.' That new concepts can become part of the law, through repeated practice, is shown by the example just given.

At that point, while sources and methodology may differ, the practical result often no longer does.

The basic conceptual difference between common law and civil law finds reflection in still another area: the role of a judge and the conduct of a case. In common law jurisdictions, particularly the United States, litigation is 'fact driven.' The facts need to be established, and the applicable precedent must then be found. With such an emphasis on the facts of a case, the role of the lawyers is a particularly active one. The judge functions as a neutral arbiter; judges of courts of appeal ultimately decide questions of law (Hay, 2010, p. 90). In a civil law court, the judge starts with the rule of law, as outlined, and searches for the facts that he or she needs for the further categorization of the case. The judge's role thus is much more active. Facts are elicited by the judge; lay judges may participate as members of the court (particularly in commercial matters), but there is never a lay jury in private law matters. The judge also ascertains the applicable law, and, if the applicable rule happens to be one of foreign law because of the international aspect of the case, he or she may even have to ascertain the content of the foreign rule of law *ex officio*. As a result, decisions of civil law appellate courts will read quite differently from those of appellate courts of common law countries. Attention to factual detail will be very slight (this is particularly true of high French courts) and the emphasis will be on legal norms, rather than the other way round.

Legal Education and the Legal Profession

In the United States of the eighteenth and nineteenth centuries, a young person could become a lawyer by 'reading law in the chambers of a lawyer,' in other words by serving as a clerk. In the civil law countries of Continental Europe, legal education has been the monopoly of the universities for centuries. With 'doctrine' – not case law – the heart of the legal system, law faculties served as the only way of entry to the profession, and its faculty members equaled in stature the position of the high judges in common law systems (see Dawson, 1978). In earlier centuries, German appellate and supreme courts routinely sent case files to the prominent law faculties, virtually for final decision on doctrinal grounds. (For the practice of '*Aktenversendung*,' see Rheinstein, 1938, p. 7.) The influence of law professors on the development of legal doctrine thus has a long tradition and continues to this day. The course of study reflected, at least until recent times, both the strict categorization of the legal system into specific fields of law and the high level of abstraction that characterizes the deductive analytic method of the civil law. Practical training of the young lawyer was left to a period of time required to be spent in an official training program (such as the *Referendariat* in Germany) or as a junior lawyer under the direction of a licensed lawyer, combined with additional training (such as in France). Full admission to the practice of law then occurred with the completion of these additional requirements. The mindset of the young lawyer had been formed, of course, by the time the university studies were completed.

In more recent times, law curricula increasingly reflect the offering or requirement of election of interdisciplinary subjects. There is also increasing instruction in the use of new technologies, for instance, how to access and use legal databases. In addition, increased emphasis on international compatibility and equivalence of academic degrees has led to the introduction of baccalaureate programs in law (for instance, in Germany). The resulting bachelor's degree certifies academic study of law but is not the equivalent of the German *Referendar* for purposes of entry into the profession; additional study would be required for the latter. Neither increased 'skills training' nor the new degree programs affect the structure of the legal system.

Outlook

Civil law systems differ among themselves, as noted initially. This also holds true for the course of study and the practical training required to become a lawyer. As a result, a lawyer – as in the case of other professionals – may face impediments in attempting to establish a practice in another country or even to render occasional services there. In the European Union, the 1988 so-called Diploma Directive was one of the first steps to implement the European Community's freedom-of-establishment and free-supply-of-services provisions. The 1998 'Establishment Directive' requires foreign lawyers wanting to work permanently in another European Community country to pass an aptitude test or to have 3 years of practical experience in the state where they want to work. Today, requalification is governed by the 2005 'Recognition of Professional Qualifications Directive,' which replaced the 1988 'Diploma Directive.' In addition, the 2006 'Directive on Services in the Internal Market' also touches on the legal profession. The Council of Bars and Law Societies of the European Community is an organization that includes the national associations of lawyers of the member states of the European Union, the European Economic Area, and Switzerland. It has published the 'Charter of Core Principles of the European Legal Profession and Code of Conduct for European Lawyers.'

In the European Union, civil law meets common law in a shared organizational structure. Administrative regulatory action to facilitate the transborder practice of law has been accompanied by efforts to harmonize areas of private law. Legal scholars are documenting the doctrinal commonalities in various fields of law – under titles such as 'European-wide tort law' (see von Bar, 1996; van Dam, 2006; Koziol and Schulze, 2008), or 'Property law in Europe' (von Bar and Drobning 2004; van Erp et al., 2012) – and there has been discussion of the feasibility of a European Civil Code (see Kötz, 1996).

A number of efforts have borne fruit. In the European Union, the joining of many countries with a civil law tradition and two common law countries (the United Kingdom and Ireland), the 'Brussels I' and 'Brussels II' regulations provide common rules for jurisdiction and judgment recognition in civil and commercial matters and for divorce and child custody,

respectively, while the 'Rome I, II, and III' regulations provide uniform rules of private international law (conflict of laws) for contract, tort, and divorce. The first two of the Rome regulations apply in all EU countries, except Denmark, the third in a more limited number of countries. In 2011, the Commission of the European Union proposed a regulation on a 'Common European Sales Law' as an optional alternative to national law. The European Union thus serves as a laboratory for accommodating the civil law and common law systems. In 2012, effective in 2015, the Brussels Regulation was revised.

While there is, then, some movement toward a new *ius commune* in Europe – which, moreover, may ultimately extend to additional countries as a result of further enlargement of the European Union – many other civil law countries around the world will not participate directly in these developments. The 'new European law' not only seeks to bridge differences among the civil law countries and between them and the common law but may also be expected to effect reforms and innovations within the civil law generally. These may well serve as models for other civil law countries, for instance, for those in South America that grow together through similar processes of economic cooperation and integration.

See also: Common Law; Legal Education; Legal Reasoning and Argumentation; Legal Scholarship; Lex Mercatoria; Supreme Courts.

Bibliography

Adriaansen, R., 1998. Open forum: at the edges of law: civil law v. common law. A response to Professor Richard B Capalli. Temple International and Comparative Law Journal 12, 107.

Brand, P., Getzler, J., 2012. Judges and Judging in the History of Common Law and Civil Law: From Antiquity to Modern Times. Cambridge University Press.

Dawson, J.P., 1978. The Oracles of the Law. University of Michigan Law School, Ann Arbor, MI (Reprinted in 1986 by William S. Hein & Co., Inc. Buffalo, New York).

Farnsworth, E.A., 1996. A common lawyer's view of his civilian colleagues. Louisiana Law Review 57, 227.

Glenn, H., 2007. Legal Traditions of the World: Sustainable Diversity in Law, third ed. Oxford University Press.

Hay, P., 2010. Law of the United States, third ed. Beck Publishing Company, Munich (Paris: Dalloz).

Hay, P., Rasmussen-Bonne, H.-E., 2011. Uniform Interpretation and Application of Brussels I and Lugano, European Legal Forum, 11, 3/4 pp. 105–112.

Hazard, G.C., Dondi, A., 2006. Responsibilities of judges and advocates in civil and common law: some lingering misconceptions concerning civil lawsuits. Cornell International Law Journal 39, 59.

Kötz, H., 1996. Europäisches Vertragsrecht. Mohr, Tübingen, Germany.

Koziol, H., Schulze, R., 2008. Tort Law of the European Community. Springer, Vienna, New York.

Rheinstein, M., 1938. Law faculties and law schools: a comparison of legal education in the United States and Germany. Wisconsin Law Review 5, 42.

van Dam, C., 2006. European Tort Law. Oxford University Press.

van Erp, S., Salomons, A., Akkermans, B., 2012. The Future of European Property Law. Sellier European Law Publishers.

von Bar, C., 1996. Gemeineuropäisches Deliktsrecht. Beck Publishing Company, Munich, Germany.

von Bar, C., Drobning, U., 2004. The Interaction of Contract Law and Tort and Property Law in Europe: A Comparative Study. Sellier European Law Publishers.

Waline, J., 2010. Droitadministrati, twenty third ed. Dalloz, Paris.

Walker, J., Chase, O.G., 2010. Common Law, Civil Law and the Future of Categories. LexisNexis, Markham.

Westphalen, F. Graf von, Thüsing, G., 2012. Vertragsrecht und AGB-Klauselwerke (30th Ergänzungslieferung). Beck Publishing Company, Munich, Germany.

Zweigert, K., Kötz, H., 1998. An Introduction to Comparative Law, third ed. Clarendon Press, Oxford.

Civil Liberties and Human Rights

Donald A Downs, University of Wisconsin, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article addresses the differences between civil liberties, civil rights, and human rights, approaching the subject from a multidisciplinary and global perspective. Among other aspects, the entry focuses on three fundamental tensions in this field: the tension between the philosophical origins of these concepts and their political and historical manifestations and influences; the tension between culture and universalism; and the ironic or ambivalent relationship between rights and state power. A concluding section discusses the special problem of terrorism and rights.

Civil liberties and human rights are the basic freedoms and claims individuals possess as members of political orders or as members of the human race. Civil liberties are essentially 'negative' political rights that stand as shields against state actions and infringements, whereas human rights may include (depending upon the theorist or political posture) these claims as well as broader claims to such things as social and economic rights, cultural rights, and collective rights of peace and environmental health. This article discusses the political and intellectual history of civil liberties and human rights, their contemporary domestic and international political and legal contexts, and major debates surrounding their usage. Emphasis is placed on the aspirations behind these concepts, as well as on the tensions and conflicts that accompany their use.

'Civil liberties' and 'human rights' are closely related terms that embrace the basic freedoms and claims to which individuals are entitled, either as citizens of a particular state or by virtue of being human. The very ideas of civil liberty and human rights presuppose two intertwined convictions: that individuals or groups in civil society have moral status independent of the organized power of society (e.g., the state); and that this power must respect the rights that flow from this status. Though they arose as a response to political and normative claims in the West, at the beginning of the twenty-first century civil liberties and human rights are at least officially endorsed by virtually all countries and the community of international law. The content and scope of these concepts are contested on philosophical and political grounds. Accordingly, the concepts provide a good method of examining the ways in which important philosophical and legal concepts interact with political and historical forces.

Basic Historical Background

The idea of a moral status independent of the state is tied in multifarious ways to notions of higher (or natural) or universal law, democracy, individual conscience, and limited government. These notions have echoes in history as far back as ancient Athens and Rome, and in medieval Christian thought. The modern turn toward individual 'rights' (as opposed to natural 'law') was the product of a complex historical process that included Renaissance and humanist emphases on human achievement and creativity; the Protestant Reformation's stress on individual religious conscience and religious pluralism; the Enlightenment's belief in the power of reason and the individual; the growth of markets; and – most importantly – the rise of democracy (*see* Law and Democracy). The specifically liberal tradition of limited government and natural rights arose in the political and intellectual history of several European countries, notably England, Scotland, France, and the Netherlands, and in the United States, in the seventeenth and eighteenth centuries.

Social contract theory in the seventeenth and eighteenth centuries envisioned 'social contracts' between the government and the citizenry based on the consent of the governed and the protection of natural rights. Building on the theory of Thomas Hobbes, John Locke (1998) maintained in his influential *Second Treatise on Government* (a work defending the Glorious Revolution in England in 1688, and the 1689 English Bill of Rights) that government's primary purpose is to protect the rights of individuals found in the state of nature, specifically rights to life, liberty, and property. The American Declaration of Independence (1776) and the French Declaration of Man and the Rights of Citizens (1789) carried these ideas further, declaring the primacy of civil and political liberties.

Basic Aspects of Civil Liberty and Human Rights

Definitions

Though the distinction between 'positive' and 'negative' rights (first articulated by philosopher Isaiah Berlin, 1969) is often blurred or overstated, civil 'liberties' are often considered negative rights in that they serve as shields that protect the liberty and rights of individuals and members of civil society from state oppression. They represent claims against state action. Classic civil liberties include freedom of speech, the press, assembly, and religion, due process and fairness in legal proceedings (especially criminal process), privacy, and freedom from illegitimate discrimination.

Civil liberties should be distinguished from civil rights. Civil 'rights' are often construed as more 'positive' rights, in that they entail the state bestowing a power to do something affirmative, or taking action to protect fundamental interests or claims against private (nongovernmental) actions. For example, the right to use privately owned public accommodations or

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86008-0

facilities, or the right not to be discriminated against in private employment, can be construed as civil rights. More broadly defined positive rights may include such claims as the right to a job, to obtain adequate housing, and to share in more equal distribution of resources. More aggressive state action is needed to effectuate such rights (*see* Civil Rights).

Civil liberties and rights are generally claims tied to citizenship in particular legal orders. 'Human rights' are more universal in nature; they exist simply because one is a human being. These include the civil liberties discussed previously, as well as freedom from torture, slavery, and degrading treatment; freedom of the family; and basic self-determination. Debate swirls around whether such rights include basic economic, social, and cultural rights and needs, or broader collective goods.

Though human rights claims can be derived from specific domestic or international legal sources, their claims are distinctively moral. As political theorist Jack Donnelley (1989) remarks, "Human rights claims are essentially extralegal; their principal aim is to challenge or change existing institutions, practices or norms, especially legal institutions." Accordingly, Donnelley emphasizes the 'possession paradox': Having a right is most important when "enjoyment of the object of the right is threatened or denied." Thus, human rights claims typically arise when a particular claim is not afforded legal protection by a particular country, such as same-sex sexual expression at one time in some states in the United States, or religious conscience in China today.

The distinct concept of 'human' rights arose in the aftermath of World War II, with the widespread condemnation of the atrocities the Nazis committed against Jews and other minorities. After having declined in the wake of skepticism and new political movements in the nineteenth and twentieth centuries (e.g., utilitarianism, emotivism, nationalism, and Marxism), notions associated with natural rights and natural law enjoyed a revival in the aftermath of the war, as democratic theorists regained respect for more objective moral principles that provide standards by which to evaluate the practices of states. Such theorists as Leo Strauss (1950) and Edward Purcell (1973) have written about a postwar 'crisis' in democratic theory along these lines. The term 'human' rights avoided the intellectual and political baggage associated with 'natural' law and rights, while at the same time pointing to universally held moral principles. In the unprecedented Nuremberg trials held after the war, Allied prosecutors convicted Nazi leaders of crimes against peace and humanity. Also, in the wake of World War II, the new United Nations (UN) made human rights an important part of its agenda, and Japan and West Germany, under the aegis of occupying forces, adopted constitutions that protect basic civil liberties and rights.

Relationship to the State

The concept of individual or natural rights is historically and pragmatically related – both positively and negatively – to the emergence of the modern nation-states from feudalism between the thirteenth and seventeenth centuries. In 1648, the Peace of Westphalia, which ended the murderous Thirty Years' War in Europe between Catholic and Protestant states, constituted the first formal international recognition of the nation-state's autonomy from religious authority. It also established the first official tolerance of religious pluralism, a crucial move in the rise of civil liberty and human rights. Yet the Westphalian model of international law left no room for the international enforcement of individual rights, as its main objective was the recognition of the principle of territorial sovereignty (domestic jurisdiction) of strong states.

Nonetheless, the rise of strong nation-states made individual rights more important than they had been in the past, spawning new theories about the obligations of states to citizens. Indeed, another paradox (which also involves the endemic jurisprudential debate between legal positivism and forms of legal analysis based on natural law) concerns the relationship between rights claims and their enforcement or recognition (*see* Natural Law). Though many theorists persuasively contend that rights claims exist independently of legal protection, rights claims (positive and negative) have to be recognized and enforced by those in power in order to be effective. The Nuremberg trials present a classic example of this fact. (Some have called the verdicts 'victors justice'.) As James Madison wrote, following the logic of Hobbes and Locke, liberal freedom can exist only when the state is strong enough to protect its citizens, but also limited enough so as not to oppress them. Writing in the aftermath of World War II, Hannah Arendt (1951) chillingly portrayed how vulnerable stateless people are to abuse of their humanity. Legal theorist Stephen Holmes (1995) puts the matter succinctly: "Weak-state pluralism is a recipe not for liberalism, but for a proliferation of rival and coercive mafias, clans, gangs, and cults … Liberal government … is meant to solve the problem of anarchy *and* the problem of tyranny within a single and coherent system of rules" (pp. 270–271).

Some Basic Issues

Important questions have been raised about the content and intellectual foundation of human rights and civil liberties. What is the scope of rights? Are such rights derived from political or legal agreement, or are they postulates of theological or philosophical inquiry? Do we grasp them by intuition or reason? Which claims are fundamental, and which less fundamental, and how can we make this determination? Should the list of fundamental claims include only basic political and civil liberties, or should it also include social and economic rights? Are rights claims culturally determined or relative, as the American Anthropological Association officially declared in 1947, or are there general principles that make certain claims universal?

Some thinkers, such as Christian Bay (1982), maintain that human rights stem from human 'needs,' which include shelter, food, and livelihood; others emphasize human beings' distinctive moral nature, stressing human dignity, self-respect, and citizenship. The debate concerns those who define human nature in largely materialistic or naturalistic terms, and those who define human nature in terms of such qualities as rationality, moral capacity, or spirituality.

Writers such as Henry Shue (1980) distinguish basic from less basic rights. A right is basic if its enjoyment is 'essential to the enjoyment of all rights.' These rights include physical security, economic security or subsistence, and liberty to

participate in the economic and political life of the community. Still others, such as Donnelley, argue that this list is insufficient because a fully developed life requires more opportunities and attributes than these minimums. However, history shows that 'negative' civil liberties are necessary to protect us from the state, so these should always be on the short list of basic rights. If we decide to include more rights as basic, we must do so without sacrificing basic civil liberties. (This is the approach of philosopher John Rawls in his classic book *A Theory of Justice* (1971): no scheme of social justice may negate fundamental civil liberties.) And we must understand that the longer the list of basic rights, the greater the potential for conflict among rights and social policies designed to promote them.

Theorists such as French jurist Karel Vasak (1982) posit 'generations' of rights based on historical development. The first generation consists of political and civil liberties, while the second generation embraces egalitarian social and economic rights. The so-called third generation rights involve humanity as a whole, including cultural self-determination, environmental health, solidarity, and peace.

The founding movements and documents in the rise of liberal democracy accentuated civil and political liberties. Yet the rise of socialism, Marxism, and the working class in the nineteenth century spawned the advocacy of social and economic rights in addition to (or instead of) civil and political rights. In the twenty-first century, such rights are found in the constitutions or fundamental laws of communist (and former communist) states and many developing or third world states (*see* Socialist Law; Postcolonial Law). Though developed liberal states are mainly dedicated to political and civil rights, social and economic rights often comprise parts of their social and legislative policy. In recent decades, such internationalist groups as the Lawyers Committee for Human Rights and the Fair Labor Association contend that corporations' use of factories in developing countries has made the protection of economic and social rights in those countries a primary concern.

Human rights covenants in the UN reflect these debates (*see* International Law and Treaties). In 1946, the Economic and Social Council of the UN established the Commission on Human Rights, which led to the Universal Declaration of Human Rights in 1948, a foundational document that has achieved the status of customary international law. In 1966, the International Covenant on Civil and Political Rights (ICCPR) and the International Covenant on Economic, Social, and Cultural Rights (ICESCR) were signed by most states, taking effect in 1976 (*see* Torture: Legal Perspectives). The ICCPR protects such basic civil liberties as freedom from arbitrary punishment, forced servitude, and unfair criminal process; freedom of thought, conscience, and religion; personal liberty and security; freedom of the family; freedom to participate in fair elections; and equal suffrage. An Optional Protocol of the ICCPR commits ratifying states to allow a special committee of experts to examine claims by individuals against them.

The ICESCR covers such rights as the right to work under good conditions; the right to an adequate standard of living; and the right to social security, food, clothing, shelter, and basic health. The ICESCR is less stringent in its wording than the ICCPR. Signatories established no rank ordering of these rights, and enforcement is more a matter of exposure and persuasion than force.

Civil Liberties in Practice

The protection of civil liberties varies (in legal provisions and applications of these provisions) in different countries due to cultural and political factors (*see* Constitutionalism, Comparative). For instance, free speech doctrine and practice in the United States protect the advocacy of illegal action, including racist rhetoric that falls short of directly triggering a disturbance of the peace or inciting violence. In contrast, Canada, Israel, Germany, France, and many other countries prohibit speech that advocates racism (racist rhetoric), regardless of the likelihood of illegal action. In Germany and Austria it is illegal to belong to a Nazi party or to wear a Nazi uniform, while courts in the United States have expressly protected such actions.

Though virtually all countries have basic legal protections for criminal suspects, standards vary widely, especially when we look at practice rather than the letter of the law. In Russia, for example, where legal institutions are poorly developed, preventive detention and criminal procedure rights of criminal suspects are poorly enforced despite formal legal protections; France offers some protections but does not recognize the privilege against self-incrimination. In the area of religion, the United States maintains an exceptionally strict separation of church and state, while such democracies as Ireland, Italy, and Germany allow more accommodation between state and religion. In India, 'personal laws' linked to the major religions (Hindu, Muslim, Christian, and Parsi) are distinguished from normal civil law, thereby accommodating culturally based discrimination against women in such areas as marriage, divorce, and inheritance. Communist China has persecuted such religious groups as Christians and the Falun Gong because their views are allegedly contrary to state ideology. And such Muslim countries as Saudi Arabia, Iran, and Pakistan have severe laws against blasphemy (disparaging religion), a form of speech that the American First Amendment protects without qualification.

Social scientists and legal scholars cite several factors that influence the extent to which countries will support civil liberties and rights on a sustainable basis. Commitment to civil liberty has historically been accompanied by social pluralism, legal institutions based on rule of law, and the differentiation of the state from civil society (*see* Rule of Law (and Rechtsstaat); Law and Development). More specific explanations include such factors as the existence of a bill of rights and judicial independence, judicial leadership and control of case dockets, and a culture of rights consciousness that encourages citizens to think in terms of rights (*see* Judicial Review; Legal Culture and Legal Consciousness; Rights: Legal Aspects).

More recent explanations point to the presence of political and social movements that engender legal change (e.g., the civil rights movement in the United States; the freedom movement of South Africa led by Nelson Mandela). Charles Epp (1998) has pinpointed sustained pressure exerted by a 'support structure for legal mobilization,' which consists of rights-advocacy organizations and lawyers, and sufficient

funding from private and (especially) public sources. Rights revolutions succeeded in the United States and Canada in recent decades because of the presence of these factors, while India's rights movement was thwarted despite a favorable Supreme Court because such factors were absent (*see* Law as an Instrument of Social Change).

Whether the 'Arab Spring' that erupted in Egypt and Tunisia in 2010 and spread to other Middle Eastern countries will result in greater freedoms and civil liberties in that region is an open question as of this writing. It will depend upon the strength of the factors considered herein.

Human Rights in International Practice

Until recently, the international system remained committed to the Westphalian model's strong presumption in favor of the 'domestic jurisdiction' of states. But this situation slowly began to change with the growth of consciousness of human rights, the democratic ethic, and globalization (*see* Globalization: Legal Aspects). The most important events before the end of World War II and the Nuremberg trials include the abolition of slavery in the British empire in the 1830s and 1840s, culminating in the League of Nations' Slavery Convention of 1926; the policy of 'humanitarian invention' by Western states to protect Christian citizens abroad in the nineteenth century; and several conventions and treaties protecting the rights of soldiers in war promulgated in the nineteenth and twentieth centuries.

In the wake of World War II and the humiliating failures of the League of Nations, the world community established the United Nations, which promulgated the UN Charter. Whereas the Westphalian model is premised on the freedom of states over their domestic jurisdictions, the UN Charter, or 'new international law' model, embraces the Kantian model of international relations and law, which emphasizes universal peace and human dignity. In reality, the models coexist in the contemporary world, posing sometimes vexing questions about where to draw the line between state sovereignty and international human rights norms. Such questions are not merely abstract problems, as supranational bodies face them on a regular basis. Yet the conferees who established the charter rejected a proposal to authorize intervention to protect rights, and a clause in the charter expressly prohibits intervention in 'matters which are essentially within the domestic jurisdiction of states.'

In recent decades, many international treaties and forums have been established under the aegis of the UN or other regional and international organizations to promote recognition of human rights (*see* International Law and Treaties). States have signed treaties in conventions against torture, genocide, racial and gender discrimination, and treaties protecting refugees and children. Again, legal, social, and political movements have been indispensable to the promotion of human rights logic and practice, as the noted human rights activist Ayreh Neier chronicled in an authoritative book (2012). In addition to hundreds of regional and intergovernmental organizations (such as the Organization of American States, the Organization of African Unity, etc.), such nongovernmental organizations as Amnesty International, Human Rights Watch, and the International Committee of the Red Cross have played major roles in raising awareness, linking international organizations, and even bringing cases for enforcement in relevant jurisdictions (Keck and Sikkink, 1998).

States have drafted regional agreements to protect rights on all continents except Asia. In Europe, the European Convention of Human Rights and Fundamental Freedoms (based on the ICCPR) formed the European Court of Human Rights, which takes cases after they have been heard by the relevant domestic courts (*see* European Union Law). Member states have agreed to accept all of the court's rulings, leading, for example, to changes in Britain's law of criminal procedure (most notably in the area of pretrial detention) and changes in several states' laws concerning the rights of children born out of wedlock.

Not surprisingly, politics has affected the application of the ICCPR and ICESCR covenants. During the cold war era, Western countries championed the ICCPR, while communist countries supported the ICESCR. Third world countries have advocated rights of self-determination, cultural rights, and collective rights concerning resources (debates over law of the seas, etc.). Cultural relativism remains an issue. In 1993, Asian countries challenged claims about the universality of political and civil rights during the UN World Conference on Human Rights, arguing that such rights can be counterproductive, even dangerous, in the context of economic underdevelopment, fragmented nationalism, and fragile state institutions. Theorists such as Donnelley counter by pointing out that these arguments ignore the actual plights of persecuted minorities in these countries, thereby serving the interests of entrenched elites or tyrants. Similar concerns are raised about the status of women in many states, such as Saudi Arabia and Iran.

Enforcement of human rights at the international level has remained problematic because of the continued normative and prudential reluctance to intervene in the domestic jurisdiction of states. Hobbes and Locke would have predicted such a result in the absence of an international sovereign with sufficient power to enforce protections of rights. As a result, the main support for human rights has been in the form of moral persuasion (far from meaningless, if not always efficacious), exposure of violations through research and publication, and the deployment of such measures as economic sanctions. UN organizations have investigated several countries, including Chile, Rwanda, Somalia, Zaire, several Latin American countries, Iran, Iraq, and South Africa. Such investigations are no substitute for political will, however.

Victories and Defeats: The Continuing Dilemma of Rights and the State

The decline of cold war politics in the UN Security Council enabled the UN to be somewhat more aggressive, sponsoring peacekeeping and actual interventions to protect human rights in Somalia, Iraq, and Bosnia in the early 1990s. Yet the efforts in Somalia and Iraq after 1991 proved short lived, and in 1994 the world stood by while massive genocide took place in Rwanda (Straus, 2006). In 1999, Cuba, China, and Sudan championed the norm of the territorial sovereignty of states in order to shield their abuses of human rights from international intervention, even though these states were at

that time themselves members of the UN Human Rights Commission.

To be sure, Yugoslav president Slobodan Milosevic was defeated in his attempt to take over Kosovo in 1999, yet this victory was won by the military commitment and might of the North Atlantic Treaty Organization under the leadership of the United States and Britain; victory came only after Serbian forces had already carried out massive ethnic cleansing. After Kosovo, Czech president Vaclav Havel wrote in *Kosovo and the End of the Nation-State* that the nation-state would end in the next century, giving away to an international community governed 'by universal or global respect for human rights, by universal civic equality and the rule of law, and by a global civil society.' Writer Leon Wieseltier (1999) replied that no oppressed soul had ever been saved by the forces of 'global civil society.' Kosovo was delivered from Milosevic by the willful acts of allied nation-states. Though the nation-state is a source of evil, 'it is also the nation-state from which we may demand rescue from such evils. The ethical content of a particular sovereignty is what finally matters.'

In 2011, a military intervention sponsored by the UN assisted the revolutionary movement of Libya in its fight against its tyrannical ruler, Muammar Gaddafi. This effort – led by the military power of the West, which regretted its failure to intervene in Rwanda 17 years earlier – succeeded in deposing Gaddafi; but the weakness of the new state in Libya has resulted in a chaotic situation as of this writing. The paradox of state power and rights perseveres.

The Special Problems Posed by International Terrorism

The emergence of violent Islamic extremism, epitomized by the terrorist attacks on the United States on September 11 2001, has created special quandaries for human rights and civil liberties, and for the applicability of governing standards of international law and the laws of war (*see* Law and Terrorism in the US and Beyond; International Law and Treaties) (Cassese, 2001). Terrorist threats before 9/11 were typically addressed through traditional criminal law enforcement, which is based on punishment after the fact and comparatively rigorous protections of rights. Though many states previously provided for greater surveillance and investigatory power in cases involving national security threats before 9/11, the criminal law model served as the predominant model for states in the fight against terrorism before then. But the existential threats posed by twenty-first century terrorism caused nation-states to adopt more aggressive preventative measures that blur the line between war powers and the powers of criminal law enforcement, and that challenge accepted notions of the rules of war because of the contested legal status of terrorists. Such measures have included treating captured terrorists as enemy combatants (rather than as criminal suspects or as normal prisoners of war) who may, therefore, be tried in military tribunals rather than in civil courts; employing more aggressive interrogation techniques in order to obtain intelligence that can prevent a major attack; detaining suspects indefinitely despite the lack of definitive evidence that they are enemy combatants who have violated the rules of war; and expanding the powers of surveillance (Bobbitt, 2008; Hoffman, 2006). Because liberal democracies exist to protect the security and rights of citizens, the battle against contemporary terrorism poses a special problem regarding what scholar Amos Guiora (2005) calls the 'balancing dilemma' between security and rights.

The emergence of twenty-first century terrorism has exacerbated the 'balancing dilemma' because of terrorism's distinctive qualities mentioned previously, and because of the previous growth of international law's commitment to human rights, also discussed earlier in this article. (See, e.g., Geneva Convention Relative to the Treatment of Prisoners of War, arts. 4, 13, August 12, 1949, 6. U.S.T. 3316, 75 U.N.T. S. 135.) Different liberal democracies have responded to the threat of market state terrorism in their own ways, based on such factors as their own understandings of national interest and their own political and legal cultures. Nonetheless, we can discern predominant trends. States typically expanded their powers to promote security, yet modified such powers over time in response to judicial and political pressures, and as more security was attained, however, precariously. In one respect, this trend supports the tenets of 'political realism' in foreign affairs and international relations: states will protect their sovereign interests when they feel they must, including readjusting the balancing dilemma (Posner and Goldsmith, 2005; Morgenthau, 1963).

For example, in the United States, federal courts have upheld detention of terrorists, military tribunals, and new surveillance while also requiring due process protections that the government had not provided (Chesney and Goldsmith, 2008). And in response to public and judicial pressure, the US government abandoned the use of the 'enhanced interrogation techniques' it had originally deployed. Israeli courts affirmed the Israeli government's declaration that it was engaged in armed conflict with Palestinian terrorists in 2000, thereby allowing Israel to engage in proportionate violent attacks against them (*State of Israel v. Marwaan Barghouti*, 2002). The Israeli Supreme Court also upheld the state's right to erect a 385-mile fence along the 'Green Line,' but ordered the state to minimize its effects (*Beit Sourik Village Council v. The Government of Israel & the Commander of the IDF Forces in the West Bank*, HCJ, 2056/04). Meanwhile, India redefined terrorism and expanded its powers of investigation and criminal process after 9/11, but courts constrained these powers in the face of public criticism (Guiora, 2005). As of this article, the balancing dilemma continues to be negotiated.

See also: Civil Rights Movements; Civil Rights; Constitutionalism, Comparative; Discrimination and the Law; European Union Law; Freedom: Political; Globalization: Legal Aspects; Human Rights, Anthropology of; Human Rights, History of; International Law and Treaties; Judicial Review; Law and Democracy; Law and Development; Law as an Instrument of Social Change; Legal Culture and Legal Consciousness; Legal Positivism; Natural Law; Rights: Legal Aspects; Rule of Law (and Rechtsstaat); Socialist Law; Torture: Legal Perspectives.

Bibliography

Arendt, H., 1951. The Origins of Totalitarianism. Harcourt, New York.

Bay, C., 1982. Self respect as a human right: thoughts on the dialectic of wants and needs in the struggle for human community. Human Rights Quarterly 4, 53–75.

Berlin, I., 1969. Two concepts of liberty. In: Berlin, I. (Ed.), Four Essays on Liberty. Oxford University Press, London.

Bobbitt, P., 2008. Terror and Consent: The Wars for the Twenty-First Century. Knopf, New York.

Cassese, A., 1986. International Law in a Divided World. Clarendon Press, Oxford.

Cassese, A., 2001. Terrorism is also disrupting some crucial categories of international law. Electronic Journal of International Law 12 (5), 993–1001.

Chesney, R., Goldsmith, J., 2008. Terrorism and the convergence of criminal law and military detention models. Stanford Law Review 60 (4), 1079–1133.

Conot, R.E., 1983. Justice at Nuremberg. Harper & Row, New York.

Donnelley, J., 1989. Universal Human Rights in Theory and Practice. Cornell University Press, Ithaca, NY.

Epp, C.R., 1998. The Rights Revolution: Lawyers, Activists, and Supreme Courts in Comparative Perspective. University of Chicago Press, Chicago, IL.

Goldsmith, J., 2007. The Terror Presidency: Law and Judgment Inside the Bush Administration. W.W. Norton, New York.

Goldsmith, J., Poser, E., 2005. The Limits of International Law. Oxford University Press, New York.

Guiora, A.N., 2005. Legislative and policy responses to terrorism, a global perspective. San Diego Journal of International Law 7 (1), 125–172.

Henkin, L., 1990. The Age of Rights, second ed. Columbia University Press, New York.

Hoffman, B., 2006. Inside Terrorism. Columbia University Press, New York.

Holmes, S., 1995. Passions and Constraints: On the Theory of Liberal Democracy. University of Chicago Press, Chicago, IL.

Keck, M.E., Sikkink, K., 1998. Activists Beyond Borders: Advocacy Networks in International Politics. Cornell University Press, Ithaca, NY.

Lawson, E. (Ed.), 1991. Encyclopedia of Human Rights. Taylor and Francis, New York.

Locke, J., 1998. In: Laslett, P. (Ed.), Two Treatises of Government. Cambridge University Press, Cambridge, UK.

Morgenthau, H.J., 1963. Politics Among Nations: The Struggle for Power and Peace, third ed. Knopf, New York.

Neier, Aryeh, 2012. The International Human Rights Movement: A History. Princeton University Press, Princeton, NJ.

Purcell Jr, E.A., 1973. The Crisis of Democratic Theory: Scientific Naturalism and the Problem of Value. University Press of Kentucky, Lexington, KY.

Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge.

Straus, S., 2006. The Order of Genocide: Race, Power, and War in Rwanda. Cornell University Press, Ithaca, NY.

Shue, H., 1980. Basic Rights: Subsistence, Affluence, and US Foreign Policy. Princeton University Press, Princeton, NJ.

Steiner, H.J., Alston, P. (Eds.), 1996. International Human Rights in Context: Law, Politics, Morals. Clarendon Press, Oxford, UK.

Strauss, L., 1950. Natural Right and History. University of Chicago Press, Chicago, IL.

Vasak, K. (Ed.), 1982. The International Dimensions of Human Rights. Greenwood Press, Westport, CT.

Wieseltier, L., June 1999. Winning ugly. The New Republic 28, 27–33.

Civil Religion

Todd Nicholas Fuist and Rhys H Williams, Loyola University Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.K. Fenn, volume 3, pp. 1874–1878, © 2001, Elsevier Ltd.

Abstract

The concept of 'civil religion' suggests that a pervasive, nonsectarian understanding of morality and transcendence sacralizes the nation-state, the polity, and the history and destiny of a society. The concept became popular in social science circles following Robert Bellah's (1967) article. Civil religion has generally been credited with functioning as a source of cohesion and coherence in political life. However, recent empirical and theoretical developments have undermined the notion that civil religion is inherently an expression of social solidarity. We first explore definitional distinctions in the literature on civil religion, focusing on the United States, where the concept became important for understanding 'American exceptionalism' with regard to religiosity. We then turn to three developments that complicate the use and application of the civil religion concept as initially formulated: increasing religious pluralism, increasing religio-political polarization, and cross-national comparisons.

The concept of 'civil religion' suggests that a pervasive, nonsectarian understanding of morality and transcendence sacralizes the nation-state, the polity, and the history and destiny of a society. In the American context, where civil religion first gained popularity as a social scientific concept, it was posited to mean the vaguely Judeo-Christian beliefs and practices that provided cohesion and coherence to political life (Bellah, 1967). The question implicit in discussions of civil religion is how the state and religion are connected. In societies where the ruling apparatus is officially 'sacred,' this is fairly straightforward: the rulers are imbued with a divinity or divine sanction that grants them legitimacy. In ostensibly secular or religiously diverse societies, however, the relationship between religion and the state is less clear. Additionally, in societies in which a secular state rules a highly religious populace, the source of institutional legitimacy becomes a salient question, particularly when high levels of religiosity are accompanied by high levels of religious diversity, as in the United States. The concept of civil religion posits an answer to these questions: legitimacy comes through the development of a religious culture that provides solidarity, meaning, and collective identity that are more national than sectarian.

The theoretical discussion that has unfolded around the topic of civil religion for the past 40 years, however, has expanded, challenged, and reworked this assertion in a number of ways. In his seminal essay on the topic, Robert Bellah (1967) suggested that "from the earliest years of the republic [there has been] a collection of beliefs, symbols, and rituals with respect to sacred things … institutionalized in a collectivity," which he termed 'American civil religion.' The phrase 'civil religion' was borrowed from Jean-Jacques Rousseau, the influential eighteenth-century political philosopher, who suggested a top-down, sacred morality imposed by the state on its citizens (Demerath and Williams, 1985; Luchau, 2009). Rousseau was less convinced than either English 'social contract' theorists or his French Enlightenment colleagues that reason and self-interest would be adequate to hold together a society – particularly to bind the masses to the state. He saw the need for some sort of religion, albeit one freed from the patriarchal and monarchal ties that supported the *ancien regime*. Thus, he called for "a purely civil profession of faith, the articles of which it behooves the Sovereign to fix … a body of social *sentiments* without which no man can be either a good citizen or a faithful subject" (quoted in Demerath and Williams, 1985: 155).

Bellah, however, developed his concept owing more to Durkheimian notions about religion as integrative and unifying (Demerath and Williams, 1985; Fenn, 2001; Santiago, 2009). He envisioned American civil religion as something emerging in civil society from the national experience itself, rooted in the nation's religious inheritance but developing separately from it. Nonetheless, despite the differences in Bellah's and Rousseau's understandings of civil religion's sources, they both assumed that culture, in this case religious culture, creates solidarity. Further, they both articulated civil religion as symbols, rituals, and sentiments that evoke loyalty and unity, rather than theology or dogma per se. However, the ongoing debate around the concept of civil religion has undermined the notion that it is inherently an expression of social solidarity. In this article, we will first explore definitional distinctions in the literature on civil religion, focusing on the U.S. context, where the concept became important for understanding American exceptionalism with regard to religiosity. We then turn to three ongoing areas of debate within the field that complicate the use and application of the civil religion concept as initially formulated. We conclude by offering suggestions for the future of work on civil religion that consider these areas of debate.

Definitional Debates

The idea of American civil religion as conceptualized by Bellah quickly proved theoretically useful for understanding faith and politics in the United States, but the concept also generated a great deal of controversy (Wimberley and Swatos, 1998). This is not surprising given the political and cultural changes that the United States was experiencing in the late 1960s and 1970s. The wide-ranging discussions over what, exactly, civil religion is eventually led Bellah himself to lament the concept's muddied analytic usefulness (Bellah, 1980).

Luchau (2009), in fact, has suggested that there are as many definitions of civil religion as there are scholars discussing it. This panoply of definitions renders a clear or essentialist definition of civil religion somewhat challenging, as there is no shortage of concepts to sift through. As early as 1974, Richey and Jones identified five types of civil religion in the literature on the topic: (1) folk religion, (2) religious nationalism, (3) democratic faith, (4) Protestant civic piety, and (5) the transcendent universal religion of the nation (14–18). Of these types, Bellah concentrated on (5), leading much early research that focused on elaborating and demonstrating that particular understanding (see Gehrig, 1981; Hammond, 1980). For example, John A. Coleman (1970), following Bellah, defined civil religion as "the set of beliefs, rites, and symbols which relates a man's role as citizen and his society's place in space, time, and history to the conditions of ultimate existence and meaning" (70). Also in 1974, Martin Marty identified two, broader poles in the understanding of civil religion: 'the nation under God' and 'the nation as self-transcendent.' He identified these dominant understandings of civil religion with the 'priestly' and 'prophetic' strands that Weber (1922/1993) identified in religion's stance toward worldly social arrangements. In the years after Bellah's seminal piece, two things were clear: civil religion had gained traction as a concept but also engendered a multitude of definitions and uses.

Given this definitional variation, perhaps it is easiest to start with what civil religion is *not*. Civil religion is not simply public expressions of religion (Roof, 2009). Merely saying a prayer in public without any sense of connecting the nation to the divine in some way does not qualify as what the current literature understands as civil religion. Additionally, purely sectarian religion doesn't qualify as civil religion (Cristi and Dawson, 2007; Demerath and Williams, 1985). Civil religion is more abstractly defined than most denominations, faiths, or creeds. Rather than suggest 'Methodism,' 'Judaism,' or 'Catholicism,' civil religion is more likely to invoke broadly Judeo-Christian understandings about morality, history, and transcendence. Finally, civil religion is not the 'idolatrous worship of the state' (Bellah, 1992: ix). It represents the state being given legitimacy through the divine rather than turning the state into something divine in and of itself.

Turning to what theoretical understandings of civil religion share, civil religion is generally used to refer to a sense of transcendence and 'overarching religious values' (Porterfield quoted in Hammond et al., 1994: 9) that connect "the society's principles of legitimacy … with divine – or otherwise ultimate – realities" (Moseley quoted in Hammond et al., 1994: 18). While Bellah specifically understood civil religion as involving reference to the nonoffensive presence of a divine being (i.e., God) and as serving an integrative role for society, these specific elements of civil religion have become less important to more recent conceptions (Luchau, 2009). Cristi and Dawson (1996) suggest that understandings of civil religion, at least in the American context, tend to view civil religion as 'existing quasi-independently' of both the state and church, focusing on a sacralized 'set of core values, like liberty, justice, charity and freedom,' being 'carried' by various rhetorics and rituals of the American polity and people.

While there has been more definitional debate around civil religion than attempts at empirical verification of the concept, some empirical work does exist. Ronald Wimberley and colleagues (Christenson and Wimberley, 1978; Wimberley, 1976; Wimberley et al., 1976) provided statistical evidence for the notion that Americans share what could be called civil religious ideas. These studies found agreement across large swaths of the population with statements such as "America is God's chosen nation," "the flag of the United States is sacred," and "holidays like the Fourth of July [are] religious as well as patriotic" (Wimberley, 1976: 343). While later research (Christenson and Wimberley, 1978) complicated the picture by finding that generally religious individuals as well as members of certain religious groups were more likely to express civil religious sentiments than others, the evidence indicated that the widespread agreement with statements such as the above suggested that some kind of civil religious understanding did, in fact, permeate U.S. culture. In sum, civil religion is the generally nonsectarian, semi-independent expression of transcendence that connects the history, values, political institutions, or practices of a nation to some kind of understanding of ultimate reality, meaning, or divinity. It has traditionally been assumed to be unifying, although this has become less common in recent formulations. Additionally, available evidence finds that we can measure civil religious attitudes as correlated with but not dependent on faith tradition and religiosity, suggesting that it holds a position of at least some autonomy from sectarian religious expressions.

Recently, scholars interested in refining Bellah's concept for further use have increasingly asked how applicable any conception of civil religion can be that fails to address or account for three things: (1) the growing diversity of American society, (2) the continued political and religious polarization of the American population, and (3) the utility of the concept outside of the American context that Bellah originally intended it to serve. In this article, we examine foundational and recent work on the concept of civil religion with an eye toward how these three concerns have driven understandings of the concept and should steer future research.

Empirical and Theoretical Developments, Part 1: Diversity and Pluralism

Work on American civil religion has generally revolved around attempts to theoretically situate and refine the concept (Bellah, 1992; Cristi and Dawson, 2007: 273; Porterfield in Hammond et al., 1994), and, contemporarily, much theoretical work rests on what the diversity of American society means for the concept of civil religion. American religion has seen increasing pluralism with regard to ideology, faith tradition, practice, and increasing amounts of secularism and irreligiosity (Putnam and Campbell, 2010; Kniss and Numrich, 2007; Roof, 1999; Wuthnow, 1988). It has become a challenge for scholars to contend that civil religion is a single, unifying idea or set of values in the face of such great diversity of belief and practice.

Religious pluralism, additionally, is not neutral; social groups understand religious pluralism differently and thus put forward competing visions of what contemporary civil religion in the United States looks like. Mark Silk (2007), for example,

presents evidence that approaches to pluralism in the United States are regional, with what it means to be 'pluralistic' differing greatly in both application and appreciation based on social context. While Bellah's original conception of civil religion posited certain understandings of the core elements of American national identity, contemporary work is more likely to point to divergent visions of national identity that are used as cultural resources in conflicts over meaning in a pluralistic society (Cristi and Dawson, 2007; Fenn, 2001; Lichterman, 2008; Long, 1974; Porterfield in Hammond et al., 1994; Williams, 1999). Richard Hecht (2007), in fact, suggests that we have seen a shift from what he terms 'passive pluralism,' the form of pluralism of the mid-twentieth-century United States documented by Will Herberg (1955/1983), to 'active pluralism,' where diverse groups seek full participation while maintaining their unique identities.

Questions about what pluralism means for civil religion have, therefore, become central to the debates about the concept's usefulness. Robert Wuthnow (2004), for example, has suggested that a central question for understanding American civil religion must inherently be how our conceptualization of civil religion will change as increasing numbers of immigrants enter the United States, bringing religious ideas from Buddhist, Hindu, and Muslim contexts with them. It is not, Wuthnow contends, merely that there are more people practicing different faiths in the United States. Rather, it is the 'kind of cultural awareness' that comes with a religiously pluralistic society that drives the changing conception of American civil religion (161). Sarna (in Hammond et al., 1994), likewise, suggests that we continue to see "deep-seated cultural differences that … set Americans at odds with one another" with regard to the expression of religion, complicating how civil religion may be publicly exercised (21).

Bellah understood civil religion as a cultural container that made cohesion possible in the United States despite great diversity. This diversity, however, was largely within the traditions of Judeo-Christianity. The increasing pluralism of American society has posed new questions for contemporary understandings of civil religion. While the directions that American civil religion will take as the divergent conceptions of faith and divinity are incorporated into U.S. culture remain unclear, this pluralism creates the space for critical examinations of what national identity means in a contemporary context. New empirical testing is required to understand if and how expressions of civil religion continue to have unifying logic, and for whom.

Empirical and Theoretical Developments, Part 2: Polarization

The increasing pluralism of the United States anticipates a second concern in parsing out work on civil religion: any understanding of American civil religion must contend with ideological polarization. Recent work has found that Americans have become more ideologically divided during the last several decades, suggesting that the notion of a purely integrative civil religion cannot be squared away with empirical reality, if it ever could (Pew Research Center, 2012). While there has been evidence that civil religious ideas are widespread in the United States (Christenson and Wimberley, 1978; Wimberley, 1976), a belief that polarization may complicate what civil religion means has led researchers to develop a number of different conceptual handles seeking to explain how civil religion may continue to be useful given that its initial formulation suggested consensus. These formulations tend to agree on several key elements: (1) civil religion does continue to be a useful concept; (2) civil religion cannot be thought of as a single, static 'thing,' given the diversity of expressions of it in contemporary society; and (3) the diversity within civil religion is best understood not as a dichotomy but as a continuum that represents different ideological expressions of it.

Widespread agreement that civil religion is a continuum, however, does not lead to widespread agreement about what constitute its poles. Generally speaking, theorists tend to posit poles that largely follow Martin Marty's (1974) discussion that contrasts American priestly civil religion, which suggests that America is a special, sacred nation, with American prophetic civil religion, which suggests that the nation must live up to certain standards or be held accountable. Cristi and Dawson (2007) see enough congruence among different models of civil religion, including Marty's, Wuthnow's (1988), Davis's (1998), and Angrosino's (2002), that they have collected the various models and discuss a 'priestly, orthodox, and conservative account of civil religion' that celebrates nationalism, capitalism, and the United States as a nation of divine providence set against a 'prophetic, progressive, and liberal vision' that demands the United States live up to sacred values lest it face judgment (275).

These are not, however, the only ways that theorists understand the continuum upon which civil religion slides. For example, rather than focus on the 'content' of civil religion, Luchau (2009) focuses on the ontological aspects of civil religion. He draws on Göran Gustafsson (1984) and Demerath and Williams (1985) to suggest that there are two main poles of civil religion: civil religion as phenomenon and civil religion as rhetoric. Civil religion as a phenomenon assumes a values consensus in society and suggests that these values are integrated into the morality of individuals, while civil religion as rhetoric suggests the contested use of symbol systems for public consumption and political debates. Luchau suggests that we have moved away from civil religion as phenomenon to rhetoric, suggesting a move from consensus to competition with regard to values and symbols.

Additionally, we may understand civil religion and polarization with regard to how top-down or bottom-up it is. Williams (1996) suggests that religion in political mobilization can be viewed as 'ideology' (top-down) or 'culture' (bottom-up), and understanding civil religion as one or the other posits a specific understanding of the relationship of citizens to each other and to the state (see also Demerath, 2001; Cristi and Dawson, 2007). Demerath and Williams (1985) additionally point out that the original Rousseauian concept of civil religion was essentially tied to state power, suggesting that the Durkheimian notion of civil religion as an integrative aspect of a unified culture should not be taken as a given. While this understanding is not at the heart of Bellah's (1967) original conception of civil religion, it is not wholly absent. Bellah did understand civil religion as being salient during times of war and crisis (see Swatos, 2006, for a recent example of this

understanding). Marvin and Ingle (1996), in fact, suggest that a civil religion demands the sacrifice of its own through war to maintain its legitimacy. While patriotic rhetoric suggests that war is to destroy an enemy, what actually binds a people is the shared sentiment around the sacrifice of their own. In this way, civil religion is soaked in ritualized bloodletting. In sum, while there is broad agreement that civil religion is expressed in a variety of ways in a polarized nation, and there is implicit agreement that these poles represent wider cultural understandings around political ideologies, the exact meaning of the civil religion continuum continues to be up for debate.

Empirical and Theoretical Developments, Part 3: The Global Context

While Bellah's initial conceptualization specifically discussed 'American civil religion,' researchers have increasingly sought evidence of its existence outside of the United States. Important to the definitional debate is whether civil religion is predicated on the Judeo-Christian cosmos that has been dominant in American culture or whether other ideologies and religious systems, such as Confucianism, operate similarly (Cha, 2000). It may be that officially secular nations develop a sort of civil religion; conversely, having a connection to traditional sectarian religion may be necessary (Lowe, 2001). Researchers have taken up these questions with regard to a number of different social contexts, with varying results.

This research has used a variety of creative approaches. West (2008), for example, studies 'international civil religious pilgrimage' to draw conclusions about how such visitations shape national collective memories. In particular, he examines Australians, predominantly young backpackers, as they visit the Gallipoli battlefields, experiencing both a heightened engagement with a legitimizing myth (Hammond in Hammond et al., 1994) of their nation and a sense of disruption from being presented with a Turkish point of view. In this way, visitation of sacred civil sites can create a complex range of understandings when presented with a variety of national viewpoints. At its core, however, West's work assumes a sense of transcendence associated with the site of pilgrimage, for more than one people.

The notion that civil religion does appear in a variety of national contexts is borne out by additional comparative evidence. Flere and Lavric (2007) find quantitative evidence that civil religion exists in a variety of national contexts by comparing survey results from various Eastern European populations to results from the U.S. state of Alabama. They found that feelings of civil religion are connected to general religiosity and support for traditionalism, but that a sense of civil religion exhibits at least some impact independent of general religiosity and traditionalism across their various samples. This suggests some kind of cross-national understanding of civil religiosity. Lowe (2001) compared expressions of civil religion in the highly religious United States with those in the explicitly atheist Soviet Union and found many similar expressions, including a transcendent collective memory, cults around national heroes, sacred writings, rituals reinforcing the social order, and civil religious virtuosos. Lowe suggests that the U.S. and Soviet cases provide support for a Rousseauian understanding of civil religion in that both nations promoted self-sacrifice and unity, but it also suggests that the Soviet Union was closer to the top-down imposition of state legitimation than the United States.

Similarly, Cristi and Dawson (1996) find evidence for understanding civil religion as coercive in their analysis of Chile under the Pinochet regime. Pinochet, they argue, attempted to impose a civil religious discourse on Chilean society, casting himself in a priestly role with regard to an ethic of war and patriotic sacrifice. Drawing on established religious symbols in Chilean society and given legitimation via the Catholic Church (at least early in Pinochet's regime), Pinochet sacralized the fight against Marxism into a holy war, legitimizing political absolutism. Ultimately, Cristi and Dawson argue, Pinochet's attempts to create this civil religion from the top down were unsuccessful for a variety of reasons, leading them to suggest that analysis of situations in which civil religions succeed, fail, change, or stay the same may generate useful data and theoretical insight. Davie (2001), for example, suggests that increasing religious diversity in Europe is creating more salient national boundaries around lines related to Christendom, even in generally secular settings. This work, taken together, suggests that not only are concepts of civil religion applicable in contexts outside of the United States, but also some of the concerns we see about American civil religion, such as increasing pluralism and concerns over whether civil religion is integrative or coercive, play out in other national settings as well (Hefner, 2001).

The Future of Civil Religion Research

The sheer volume of work on civil religion indicates that researchers continue to find the concept theoretically useful. Few other concepts so evocatively capture the sacred place that national foundational heroes, institutions, values, and practices so often hold with the populace. However, the debates discussed in this article – that the civil religion concept must be theoretically reworked to encompass diversity and polarization as well as be modified for use outside of the United States – have created somewhat muddied waters, with a number of different theoretical assertions and positions vying for dominance. We conclude by offering some suggestions for future ways in which the concept may be refined and tested.

If the work discussed in this article is any indication, there appears to be agreement that civil religion is most useful when understood pluralistically, as something that represents both unity and coercion – an inclusive vision and an exclusive proclamation. Work on civil religion in the American context has tended to examine elite understandings of large historical events. Bellah (1967) himself suggested that civil religion was most apparent during times of trial such as the Revolutionary War, Civil War, and Vietnam War, and looked to how presidents used the divine in their inaugural speeches. If, however, we accept that civil religion may not be exclusively unifying, perhaps looking to elites may not always be the most fruitful way to study the phenomenon or its discourse. Rather, looking at how citizens' groups and social movements draw upon religious rhetoric as they put forward competing visions of society may serve to provide

useful empirical evidence for understanding the competing visions of national identity and its construction (Williams and Alexander, 1994). We might term this 'lived civil religion.' Works by Paul Lichterman (2005, 2008), Mark Warren (2001), and Richard Wood (2002) have explored many ways that religion explicitly enters the activism conducted by faith-based citizens' groups, while Courtney Bender (2003) has examined how a sort of generalized religion often undergirds the work done by volunteers at a nonprofit organization. These intermediate organizations, such as faith-based groups, nonprofit organizations, and social movement groups, are an important way that Americans learn how to understand and talk about civic culture (see Eliasoph, 1999). Additional work could uncover how various visions of national transcendence play into the positions taken by socially active organizations, both those that promote inclusivity or pluralism and those that promote exclusivity or a narrow understanding of who may count as a member of the national community.

Additionally, continued work at the international level could provide necessary scholarship on how the sacred works across cultural borders. Flere and Lavric's (2007) suggestion that there are civil religious understandings across national borders that are connected to traditionalism and general religiosity but not reducible to them is intriguing, and suggests possible avenues for analyzing how different cultures use concepts of the sacred. In particular, if civil religion may be understood as something that can be (1) coercive, (2) top-down, and (3) a failure if met with resistance from the populations it is being imposed on (Demerath and Williams, 1985; Cristi and Dawson, 1996), examining how different groups, both in and out of power, have manipulated sacred symbols and ideas to promote their agendas with varying degrees of success could help us understand how concepts of the sacred are used, disseminated, and modified more broadly. International work could, therefore, help us refine empirically what is at the core of 'the sacred' across a variety of cultural contexts.

See also: Religion: Mobilization and Power; Religion: Nationalism and Identity; Religiosity.

Bibliography

Angrosino, M., 2002. Civil religion redux. Anthropological Quarterly 75, 239–267.

Bellah, R.N., 1967. Civil religion in America. Daedalus, Journal of the American Academy of Arts and Sciences 96, 1000–1021.

Bellah, R.N., 1980. Religion and the legitimation of the American republic. In: Bellah, R.N., Hammond, P.E. (Eds.), Varieties of Civil Religion. Harper and Row, New York.

Bellah, R.N., 1992. The Broken Covenant: American Civil Religion in Time of Trial. University of Chicago Press, Chicago, IL.

Bender, C., 2003. Heaven's Kitchen: Living Religion at God's Love We Deliver. The University of Chicago Press, Chicago, IL.

Cha, S.H., 2000. Korean civil religion and modernity. Social Compass 47, 467–485.

Christenson, J.A., Wimberley, R.C., 1978. Who is civil religious? Sociological Analysis 39, 77–83.

Coleman, J.A., 1970. Civil religion. Sociological Analysis 31, 67–77.

Cristi, M., Dawson, L.L., 1996. Civil religion in comparative perspective: Chile under Pinochet (1973–1989). Social Compass 43, 319–338.

Cristi, M., Dawson, L.L., 2007. Civil religion in America and in global context. In: Beckford, J.A., DemerathIII, N.J. (Eds.), Handbook of Sociology of Religion. Sage, London.

Davie, G., 2001. Global civil religion: a European perspective. Sociology of Religion 62, 455–473.

Davis, D.H., 1998. Civil religion as judicial doctrine. Journal of Church and State 40, 7–24.

Demerath III, N.J., Williams, R.H., 1985. Civil religion in an uncivil society. Annals of the American Academic of Political and Social Science 480, 154–166.

Demerath III, N.J., 2001. Crossing the Gods: World Religions and Worldly Politics. Rutgers University Press, New Brunswick, NJ.

Eliasoph, N., 1999. Avoiding Politics: How Americans Produce Apathy in Everyday Life. Cambridge University Press, New York.

Fenn, R.K., 2001. Civil religion. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Pergamon Press, Oxford, UK.

Flere, S., Lavrič, M., 2007. Operationalizing the civil religion concept at a cross-cultural level. Journal for the Scientific Study of Religion 46, 595–604.

Gehrig, G., 1981. American Civil Religion: An Assessment. Monograph 3. Society for the Scientific Study of Religion, Storrs, CT.

Gustafsson, G., 1984. Civil religion in Sweden: some notes on its contents and local expressions. In: Harmat, B. (Ed.), The Church and Civil Religion in the Nordic Countries of Europe. The Lutheran World Federation, Geneva.

Hammond, P.E., 1980. The rudimentary forms of civil religion. In: Bellah, R.N., Hammond, P.E. (Eds.), Varieties of Civil Religion. Harper & Row, New York.

Hammond, P.E., Porterfield, A., Moseley, J.G., Sarna, J.D., 1994. Forum: American civil religion revisited. Religion and American Culture: A Journal of Interpretation 4, 1–23.

Hecht, R.D., 2007. Active versus passive pluralism: a changing style of civil religion? Annals of the American Academy of Political and Social Science 612, 133–151.

Hefner, R.W., 2001. Public Islam and the problem of democratization. Sociology of Religion 62, 491–514.

Herberg, W., 1955/1983. Catholic, Protestant, Jew. The University of Chicago Press, Chicago, IL.

Kniss, F., Numrich, P.D., 2007. Sacred Assemblies and Civic Engagement: How Religion Matters for America's Newest Immigrants. Rutgers University Press, New Brunswick, NJ.

Lichterman, P., 2005. Elusive Togetherness: Church Groups Trying to Bridge America's Divisions. Princeton University Press, Princeton, NJ.

Lichterman, P., 2008. Religion and the construction of civic identity. American Sociological Review 73, 83–104.

Long, C., 1974. Civil rights – civil religion: visible people and invisible religion. In: Richey, R.E., Jones, D.G. (Eds.), American Civil Religion. Harper and Row, New York.

Lowe, B.M., 2001. Soviet and American civil religion: a comparison. Journal of Interdisciplinary Studies 13, 73–96.

Luchau, P., 2009. Towards a contextualized concept of civil religion. Social Compass 56, 371–386.

Marty, M.E., 1974. Two kinds of civil religion. In: Richey, R.E., Jones, D.G. (Eds.), American Civil Religion. Harper and Row, New York.

Marvin, C., Ingle, D.W., 1996. Blood sacrifice and the nation: revisiting civil religion. Journal of the American Academy of Religion 64 (4), 767–780.

Pew Research Center, 2012. Partisan Polarization Surges in Bush, Obama Years Accessed online on 10.06.12 at: http://www.people-press.org/files/legacy-pdf/06-04-12%20Values%20Release.pdf.

Putnam, R.D., Campbell, D.E., 2010. American Grace: How Religion Divides and Unites Us. Simon and Schuster, New York.

Roof, W.C., 1999. Spiritual Marketplace: Baby Boomers and the Remaking of American Religion. Princeton University Press, Princeton, NJ.

Roof, W.C., 2009. American presidential rhetoric from Ronald Reagan to George W. Bush: another look at civil religion. Social Compass 56, 286–301.

Santiago, J., 2009. From 'civil religion' to nationalism as the religion of modern times: rethinking a complex relationship. Journal for the Scientific Study of Religion 48, 394–401.

Silk, M., 2007. Defining religious pluralism in America: a regional analysis. Annals of the American Academy of Political and Social Science 612, 64–81.

Swatos Jr., W.H., 2006. Implicit religious assumptions within the resurgence of civil religion in the USA since 9/11. Implicit Religion 9, 166–179.

Warren, M.R., 2001. Dry Bones Rattling: Community Building to Revitalize American Democracy. Princeton University Press, Princeton, NJ.

Weber, M., 1922/1993. The Sociology of Religion. Beacon Press, Boston, MA.

West, B., 2008. Enchanting pasts: the role of international civil religious pilgrimage in reimagining national collective memory. Sociological Theory 26, 258–270.

Williams, R.H., 1996. Religion as political resource: culture or ideology? Journal for the Scientific Study of Religion 35, 368–378.

Williams, R.H., 1999. Visions of the good society and the religious roots of American political culture. Sociology of Religion 60, 1–34.

Williams, R.H., Alexander, S.M., 1994. Religious rhetoric in American populism: civil religion as movement ideology. Journal for the Scientific Study of Religion 33, 1–15.

Wimberley, R.C., 1976. Testing the civil religion hypothesis. Sociological Analysis 37, 341–352.

Wimberley, R.C., Clelland, D.A., Hood, T.C., Lipsey, C.M., 1976. The civil religious dimension: is it there? Social Forces 54, 890–900.

Wimberley, R.C., Swatos Jr., W.H., 1998. Civil religion. In: Swatos, W.H. (Ed.), Encyclopedia of Religion and Society. Accessed online on 10.06.12 at: http://Hirr.Hartsem.Edu/Ency/CivilRel.Htm.

Wood, R.L., 2002. Faith in Action: Religion, Race, and Democratic Organizing in America. University of Chicago Press, Chicago, IL.

Wuthnow, R., 1988. The Restructuring of American Religion: Society and Faith since WWII. Princeton University Press, Princeton, NJ.

Wuthnow, R., 2004. Presidential address 2003: the challenges of diversity. Journal for the Scientific Study of Religion 43, 159–170.

Civil Rights

David Kairys, Temple University, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Civil rights are legal protections of individuals or groups from certain forms of oppression that gained widespread acceptance throughout the world in the last half of the twentieth century and continuing into the twenty-first century. The origins of civil rights can be found in practices of governments or powerful individuals or institutions that came to be viewed as oppressive, although these are often obscured in foundational or origin stories that take on an abstract, timeless quality. The most common civil rights are prohibition of discrimination based on race, ethnicity, religion, and gender; the right to personal security, including protections for persons accused or suspected of crimes; the right to vote and to participate in democratic political processes; and freedom of expression, association, and religion. Civil rights lose their significance if they are not available to all people, but the failure to acknowledge conflicting rights and the interconnectedness of rights, interests, privileges, and expectations has made acceptance of civil rights advances more difficult than it has to be, particularly in countries where rights tend to be articulated in absolutist terms. The most troubling questions about civil rights have centered on which rights are protected, their content and formulation, and their enforcement. The model of judicial primacy in the United States produced two periods of historic protection of civil rights available to people of ordinary means, but no single source or enforcement mechanism for civil rights has proved consistently superior to others. Establishment and enforcement seem most effective when accomplished by popular and participatory means, and popular understanding of the importance and history of civil rights is probably the most significant factor in their continued vitality.

Civil rights legally protect individuals or groups from certain forms of oppression. While civil rights are commonly associated with the 1960s movement in the United States to establish equality for people of African descent and – more generally – with the US Bill of Rights, by the end of the twentieth century their reach and recognition was global. In modern political, academic, and public usage, civil rights embody, and provide legal support for, basic concepts of human dignity and respect for individuals and groups in their diverse cultures and ways. In the second decade of the twenty-first century, recognition and enforcement of civil rights, or some assortment of the most fundamental civil rights, are widely understood as a necessary element of freedom, democracy, and equality.

International bodies, treaties, and courts – and many countries – prefer the term 'human rights,' which usually encompasses a broader array of rights and places different priorities on some particular rights than does US civil rights law (*see* Criminal Liability and Criminal Defenses; Human Rights: Legal Perspectives; Civil Liberties and Human Rights). The United Nations Charter, the *Universal Declaration of Human Rights (1948)*, and many other human rights treaties and covenants, including recently the U.N. Declaration on the Rights of Indigenous Peoples (2007), protect social, economic, and cultural rights as well as political and civil rights (*see* Indigenous Rights). These rights often place affirmative obligations on governments to provide all their people with minimal nutrition, healthcare, housing, and education. Eleanor Roosevelt, criticizing the US reluctance to include economic and social rights, put it this way: "You can't talk civil" (see Copelon, 1998).

Civil rights are usually established in a written constitution, statute, or treaty. However, they have been based on unwritten constitutions and have occasionally arisen from the pronouncements of monarchs and conquerors, perhaps most famously Napoleon.

These sources of civil rights protections are not self-executing: some countries have strongly worded civil rights provisions in written constitutions that are not enforced; the meaning and interpretation of civil rights provisions are usually controversial; and despite the prevalence of civil rights aspirations and rhetoric, no country has as yet found a reliable method for systematic, consistent protection of civil rights.

Origins

The origins of some civil rights go back almost a thousand years to the earliest limitations on governments, which at least indirectly protected individuals. Civil rights as such emerged in the seventeenth and eighteenth centuries, as social development and political and philosophical thought emphasized the individual and, later, based national sovereignty and legitimacy on democratic forms of government. However, systematic legal enforcement, and even widespread aspirational recognition, of civil rights did not occur until after World War II. They quickly achieved global acceptance in the last half of the twentieth century; today the laws of almost all countries, and several widely adopted international pacts, at least purport to protect some array of fundamental rights of individuals.

Approached historically and contextually, the various civil rights can be traced to practices by governments or by powerful individuals or institutions that came to be viewed as oppressive. Thus, the Fourth Amendment to the US Constitution limiting police searches was a response to the dreaded house-to-house general searches by British authorities in the colonial period. However, such origins are often obscured, because civil rights – like other basic legal precepts – are usually analyzed and clothed with the rhetoric of foundational principles and tend to take on, in law as well as in politics and fiction, an abstract, timeless quality that transcends history, experience,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86009-2

and human agency. This occurred in the United States with freedom of speech, which has acquired a major place in the national origin story, although its most significant features developed in the mid-twentieth century by common forms of social change (Kairys, 1982; Levy, 1985).

Scope

The most common civil rights are equality based on race, ethnicity, religion, and gender; the right to personal security, including protections for persons accused or suspected of crimes; the right to vote and to participate in democratic political processes; and freedom of expression, association, and religion. Additional civil rights are emerging from political, legislative, and judicial processes in a range of countries and internationally. Privacy, which in the US system encompassing personal security and autonomy as well as control over personal information, is of growing concern throughout much of the world. The rights of indigenous peoples (*U.N. Declaration on the Rights of Indigenous Peoples* (2007)) and sexual preference minorities have recently gained substantial recognition.

Personal security was the first, and remains the primary, civil right. It was a major focus of leading Anglo-American legal authorities, from William Blackstone (1769/1992) and St. George Tucker (1969) to Thomas Cooley (1874). The pathbreaking civil rights text by Thomas Emerson and David Haber (1952) considers the right to personal security first and describes its significance in the introduction: "In a society based upon human dignity and the development of the individual personality, clearly all members are entitled to security of the person – protection from bodily harm, involuntary servitude, and the fear of physical restraint." Although the rights of criminal defendants in the United States are commonly thought to originate in the 1960s Supreme Court rulings, the US Bill of Rights has as its major subject and most numerous protections a series of limits on the government's power to punish individuals criminally.

Although the United States was an early standard bearer and innovator on civil and human rights, the current content of human rights around the world often exceeds or precedes specific US civil rights protections, including prohibition of the execution of children and mentally disturbed adults, protections against domestic abuse, and a range of social, economic and cultural rights. Other nations also place different relative importance on various rights, such as rejecting absolutist free speech rights and emphasizing equality rights (Movement Raelien Suisse vs Switzerland, 2012). While the United States has adopted most of the human rights treaties and covenants (although often with specific reservations), US courts have not consistently enforced them. The United States is also unusual among the nations of the world in its renewed and continuing embrace of capital punishment and in the retrenchment of procedural and fair trial rights to expedite executions.

In most countries, civil rights protect individuals from their governments, but protection for groups and from other individuals have also been recognized (Kichwa Indigenous Community of Sarayaku vs Ecuador, 2012). Prohibitions of slavery around the world limit individuals as well as governments. Equal rights for women have included limits on individually inflicted violence against women and recognition of the rights of women as a group. Group rights have also been recognized in some jurisdictions for indigenous peoples, ethnic minorities, children, and the disabled (*see* Indigenous Rights; Juvenile Justice: International Law and Children's Rights; Law and People with Disabilities).

Universal, Absolute, and Conflicting Rights

Civil rights lose their meaning and importance if they are not extended to all people. However, the universality of civil rights, both within a particular country and across national borders, and the common notion that extension of civil rights to those previously denied them has no effect on others are problematic.

Typically, recognition of civil rights disturbs a status quo in which privileges and hierarchies were entrenched. If previously silenced speakers or voters gain the rights to speak and vote, the effectiveness of the voices and votes of the previously privileged is diminished. Abolishing racial discrimination in employment or college admissions diminishes the prospects of the previously privileged. Abolishing racial discrimination in public accommodations diminishes the freedom of association of those who wish to exclude racially. Recognition of the rights of people who have been oppressed involves a determination that the interests, expectations, tastes, and sometimes civil rights of others constituted an illegitimate or disproportionate privilege which came only at the expense of that oppression, and to which they should not have been, and no longer are, entitled. The failure to acknowledge the interconnectedness of rights, interests, privileges, and expectations has probably made acceptance of civil rights advances more difficult than it has to be. This has been a substantial problem in the United States, where interconnectedness is only acknowledged for affirmative action and where civil rights tend to be articulated in absolutist terms.

Absolutist conceptions of civil rights usually rest on the same thinking, memorably critiqued by Justice Oliver Wendell Holmes' famous example of shouting 'fire' in a crowded theater. Words have effects, otherwise we might not be so fond of them, and in some circumstances those effects outweigh the value of speech. Such occasions are rare, but they do exist – for example, where one person's speech is interfering with or drowning out another's or all others. Nevertheless, the US Supreme Court has protected electoral campaign finances as speech in absolutist terms that make even minimal regulation difficult or impossible (Buckley vs Valeo, 1976; Citizens United vs FEC, 2010). The Court characterized money as speech, and refused to recognize as significant or legitimate countervailing concerns the importance of fair elections to society and other individuals or the speech rights of those whose voices cannot be heard in money-driven elections.

Absolutist conceptions of this sort, though often articulated by staunch advocates of individual rights, are reminiscent of the 'Lochner era' in the United States, in which the civil right to due process was used by the Supreme Court to invalidate a series of economic reforms aimed at protecting working and poor people. The combination of an absolutist formulation and a civil right based on unrestrained free enterprise yielded

one of the most notorious eras in US constitutional law. These rights that people around the world seek and rely on for protection from powerful individuals, groups, and institutions, often at considerable personal sacrifice, can become another instrument for the powerful (Kairys, 1993, 2013).

There is also a fundamental sense in which all civil rights protect individuals by limiting others – either all others collectively, because civil rights prohibit certain actions by majorities and governments, or individually, in the situations in which civil rights protect against private as well as governmental actions. Thus, civil rights both bestow and limit freedom. They limit the ability of even democratically constituted governments, for example, to discriminate against a racial minority or to ban an opposition political party, by placing the individual's or minority's rights above collective power (West Va. Bd. of Educ. vs Barnette, 1943). Civil rights are, in this sense, both a limit on and a necessary feature of democracy.

These tensions permeate heated controversies about the universality of civil rights across national borders. The conventions on the rights of women referred to above have been the subject of an unusual array of national reservations, usually on the ground that they clash with cultures and religions, which also can claim civil rights protection. For example, the UN General Assembly has called on countries to eliminate female genital mutilation because it violates several of the most fundamental and generally accepted civil and human rights, although it is mandated by some cultures and religions.

Formulation and Enforcement: The US Model

The US tradition of extensive judicial power, which includes judicial innovation beyond the specific language of a constitution, statute, or treaty, is gaining some acceptance internationally. This expanded acceptance is in part due to its reputed tendency to protect civil rights, although the role of courts in civil rights matters (and more generally) is controversial in the United States. Many other countries have tended to rely on international human rights treaties or legislation. Courts lack the democratic legitimacy of legislatures, while their very distance from democratic processes could enhance the potential for protection of individual rights against even mobilized majorities. Protection of individuals and groups from majoritarian oppression is the hallmark of meaningful civil rights and provides the best measure of a particular society's civil rights record. A closer look at the history and record of a particular country – the United States, because of its historical role and wide recognition as a leader and model for civil rights – demonstrates both the potential and the difficulty of sustained, systematic protection of civil rights.

The role of courts in the civil rights victories of the 1960s and throughout US history is often exaggerated (Kairys, 1993, 2003, 2006; Cover, 1976). Integration of public schools (Brown vs Bd. of Educ. 1954) was a milestone, but most aspects of equality for African Americans in the twentieth century were established by the US Congress. In the 1960s these included banning racial discrimination in voting, housing, employment, and public accommodations. The major civil rights advances of the nineteenth century were achieved by constitutional amendment and legislation.

The history of US courts reveals them to be more often an obstacle to than a protector of civil rights. For example, the US Supreme Court ruled in 1857 that African Americans have 'no rights which the white man is bound to respect' and are not fully citizens or human beings (Dred Scott vs Sanford), which destroyed the Missouri Compromise on slavery and was a contributing cause of the Civil War; approved segregation even after the Civil War amendments to the Constitution (Plessy vs Ferguson, 1896); negated the 'privileges and immunities' clause of the Fourteenth Amendment (Slaughterhouse Cases, 1873), which applied civil rights protections against the states; refused to protect even the most basic free speech rights until the 1930s (Davis vs Massachusetts, 1897; Kairys, 1982); and approved the imprisonment of all persons of Japanese ancestry on the west coast during World War II without any charges or proof of individual guilt (Korematsu vs US, 1944). Generally, throughout US history, oppressed minorities and individuals facing repression by government or organized majorities – including in modern times the repressive McCarthyism of the 1950s – have not been significantly protected by the courts.

There are only two periods in US history characterized by systematic and sustained judicial protection of such civil rights, from about 1937 to 1944 and from about 1961 to 1973. Both these periods occurred in the context of sustained mass popular support and successful legislative initiatives for civil rights, which in the US system seem necessary (but not sufficient) for systematic protection. During these periods, the Supreme Court established a system of limited government and protection of individual rights available to people of ordinary means probably unrivaled in world history, including strenuous protection of speech, association, and privacy; procedural and substantive rights of persons accused or suspected of crime; the rights of women; voting and participatory rights; and prohibitions of discrimination (Kairys, 1993, 1998, 2006, 2013; W. Va. Bd. of Educ. vs Barnette, 1943).

The main judicial vehicle for protecting civil rights in these two periods was the 'strict scrutiny' standard of judicial review, by which any government action that infringed or restricted these rights was strictly scrutinized and presumptively invalid, surviving only if supported by a 'compelling' government interest that could not be furthered by 'less restrictive' means.

While these decisions seldom articulated absolute rights (although absolutist rhetorical flourishes were common), in practice strict scrutiny meant near certain invalidity.

Starting in the mid-1970s, as the national political mood changed and more conservative justices were appointed to the Court, the Supreme Court repudiated or significantly diluted almost all of the major civil rights advances of the post-World War II period. This retrenchment was led in the political arena by President Ronald Reagan and a conservative movement that emphasized limiting government and espoused the sometimes extreme distrust of government evident in many periods of US history (Wills, 1999).

The thrust was to empower the people by limiting the courts – in the popular phrase, 'judicial restraint' – which, paradoxically, strengthens government and weakens individual rights. Thus, this conservative judicial restraint trend reviled the 1965 case establishing a civil right to privacy (Griswold vs Conn.), a decision that protected the individual and the people generally from

a state law that banned all use of birth control. A judicially based system of protection of civil (or any other) rights requires the courts to intervene when government is infringing on protected rights, and conservative justices have not hesitated to intervene to protect civil rights they value (e.g., Buckley vs Valeo, 1976; Shaw vs Reno, 1993). Historically, the pattern in the United States has been advocacy of judicial restraint by those whose values and approaches are being rejected by the courts. Perhaps the most vehement and successful advocate of judicial restraint was liberal President Franklin Roosevelt, whose reforms were struck down by a Supreme Court intent on imposing unrestricted laissez-faire economics (Kairys, 1998).

In this retrenchment, the rights to exercise nonmainstream religion and against establishment of religion were substantially undercut, including the refusal to protect a Native American religious ceremony that pre-dates Christianity (Empl. Div. vs Smith, 1990). Speech rights available to people of ordinary means were retrenched (for example, the public areas and circumstances available for protest were substantially narrowed), while speech rights available to corporations and wealthy individuals were enhanced, including recognition of corporate free speech rights. The equality rights of minorities were undercut by adoption of virtually insurmountable burdens of proof, resulting in denials of relief for some egregious practices left over from segregation. The equality rights of the white majority were enhanced, resulting in successful challenges to electoral reapportionment and affirmative action, even where they were good faith remedial efforts to end discrimination against African Americans and other minorities. As a result, almost all the winning parties in racial equality cases decided by the Supreme Court in the last few decades of the twentieth century and continuing into the twenty-first were white. The US attempt to end centuries of discrimination that included slavery and forced segregation was short and mostly limited to erasing de jure discrimination; the mass of African Americans live in the same segregated poverty they endured under de jure segregation (Kairys, 1993, 2006, 2013; Memphis vs Greene, 1981; Richmond vs Croson, 1989; Shaw vs Reno, 1993).

This retreat was accomplished by a range of judicial means. In some areas, such as exercise of religion and some aspects of privacy, the Court withdrew strict scrutiny protection. In others, such as equality, establishment of religion, and some aspects of free speech, a range of new rules made it near impossible to prove a violation – such as the requirement for proving that government not only violated a civil right but did so purposely, with animus or the specific motivation to violate the right.

None of these new rules is required by the language of the US Constitution or by rules of legal reasoning or analysis; nor is the strict scrutiny standard that was used to protect civil rights. This is characteristic of the judicial model, which in the United States yielded cyclic protection of civil rights in some (not all) periods in which support for those rights had strong, sustained, and politically mobilized support (see Kairys, 1998, 2003, 2008).

By the end of the twentieth century, the lead in civil rights had passed from the United States to other nations and to international human rights tribunals and agreements. The Constitutional Court of South Africa led a worldwide trend to ban or limit the death penalty (State vs Makwanyane, 1995). The European Court of Human Rights protected the rights of gay men and lesbians (Dudgeon vs United Kingdom, 1981); the Supreme Court of the United States initially rejected such rights (Bowers vs Hardwick, 1986), then accorded them some protection (Lawrence vs Texas, 2003). The rights of racial minorities and women were advanced principally by international human rights agreements. Progress in this direction has come mostly as a result of international human rights agreements and the growing sense they represent that oppression and repression, though they have hardly ceased, are no longer internationally acceptable.

Prospects for Civil Rights

Since civil rights can encompass a variety of possible limits on government to protect individuals and groups (and some limits on other individuals and affirmative obligations of governments), the details of their substance, context, and history are necessary for an understanding of their social meaning and importance. In this sense, it is hard to say that any of us is 'for civil rights,' and perhaps harder to say we are 'against civil rights' (which might mean government without limits), without specifying whose rights, which rights, and the context. For some, government is most oppressive when it limits gun ownership or attempts to overcome the effects of past discrimination; for others, it is most oppressive when it limits exercise of religion, abortion rights, or the rights of historically oppressed minorities. In the post-World War II period, the United States set an international standard for protection of minorities and the enforcement of a range of civil rights that protect the individual from oppression by government and powerful majorities; but US history also demonstrates the fragility of civil rights, and the use of civil rights to insulate the powerful from popular reforms aimed at protecting working and poor people.

Civil rights have played an important role in the advancement of human dignity, respect for the ways of diverse cultures and individuals, and democratic forms of government. In a relatively short time in the last half of the twentieth century, racism moved from acceptable or trivial to wrong (Kairys, 2006), and a deep sense of civil rights and fair play became an international intellectual, moral and political standard. However, this has not stopped – by any reasonable measure – racial, ethnic, religious, and national oppression or violence, or the routine denial of the range of basic civil rights. There is a substantial and troubling divide between our words, aspirations, and moral precepts on the one hand, and our actions on the other. We may be a species whose gift of reason and empathy has outpaced its habits of survival, which still reside – as if helpless to listen to our own pleas – in deep-seated (perhaps significantly genetic) tribalism. If this is so, the struggle for civil rights and human dignity will be long, indeed, but all the more compelling.

No single source or enforcement mechanism for civil rights has proved consistently superior to others. Establishment and enforcement seem most effective when accomplished by the most popular and participatory means available, and popular understanding of the importance and history of civil rights is probably the most significant factor in their continued vitality.

See also: Civil Liberties and Human Rights; Civil Rights Movements, Discrimination and the Law; Discrimination: Racial; Due Process in Law; Employment and Labor, Regulation of; Equality and Inequality: Legal Aspects; Gender and the Law; Human Rights and Social Work; Human Rights, Anthropology of; Human Rights, History of; Human Rights: Legal Aspects; Indigenous Rights; Justice and Law; Juvenile Justice: International Law and Children's Rights; Law and Democracy; Law and People with Disabilities; Race and the Law; Rights: Legal Aspects.

Cases

Bowers vs Hardwick, 478 U.S. 186 (1986).
Brown vs Board of Education, 347 US 483 (1954).
Buckley vs Valeo, 424 US 1 (1976).
Citizens United vs FEC, 558 U.S. 310 (2010).
Davis vs Massachusetts, 167 US 43 (1897).
Dred Scot vs Sanford, 60 US (19 How.) 393 (1857).
Dudgeon vs United Kingdom, 45 European Ct. of Human Rights (ser. A), four E.H.R.R. 149 (1981).
Employment Division vs Smith, 494 US 872 (1990).
Griswold vs Connecticut, 381 US 479 (1965).
Kichwa Indigenous Community of Sarayaku vs Ecuador, Inter-Am. Ct. H.R. (ser. C) No. 245 (27 June 2012).
Korematsu vs United States, 323 US 214 (1944).
Lawrence vs Texas, 539 U.S. 558 (2003).
Lochner vs New York, 198 US 45 (1905).
Miller vs Johnson, 115 S.Ct. 2475 (1995).
Memphis vs Greene, 451 US 100 (1981).
Movement Raelien Suisse vs Switzerland, European Court of Human Rights, No. 16 354/06 (13 July2012).
Plessy vs Ferguson, 163 US 537 (1896).
Richmond vs Croson, 488 US 469 (1989).
Shaw vs Reno, 509 US 630 (1993).
Slaughterhouse Cases, 83 US (16 Wall.) 36 (1873).
State vs Makwanyane, 1995(3) South Africa Law Reports 391 (1995).
West Virginia State Board of Education vs Barnette, 319 U.S. 624 (1943).

Bibliography

Blackstone, W., 1992. Commentaries on the Laws of England. Hein, New York.
Cooley, T., 1874. Constitutional Limitations. Little, Brown, Boston.
Copelon, R., 1998. The indivisible framework of international human rights: bringing it home. Chapter 9. In: Kairys, D. (Ed.), The Politics of Law, third ed. Basic Books, New York, pp. 216–239.
Cover, R., 1976. Justice Accused. Yale University Press, New Haven, CT.
Emerson, T., Haber, D., 1952. Political and Civil Rights in the United States. Dennis, Buffalo, NY.
Kairys, D. (Ed.), 1982. The Politics of Law. Pantheon Books, New York.
Kairys, D., 1993. With Liberty and Justice for Some. New Press, New York.
Kairys, D. (Ed.), 1998. The Politics of Law, third ed. Basic Books, New York.
Kairys, D., 2003. Searching for the rule of law. Suffolk Law Review 36, 307–329.
Kairys, D., 2006. A brief history of race and the supreme court. Temple Law Review 71, 1–12.
Kairys, D., 2008. Philadelphia Freedom, Memoir of a Civil Rights Lawyer. U. Michigan Press, Ann Arbor, MI.
Kairys, D., 2013. The contradictory messages of Rehnquist-Roberts era speech law: liberty and justice for some. Illinois Law Review, 195–220.
Levy, L., 1985. Emergence of a Free Press. Oxford University Press, New York.
Tucker, St. G., 1969. Blackstone's Commentaries, with Notes of Reference to the Constitution and Laws of the Federal Government of the United States and of the Commonwealth of Virginia. Rothman, South Hackensack, NJ.
Universal Declaration of Human Rights. 1948 G.A. Res. 217A, U.N. Doc. A_810
U.N. Declaration of the Rights of Indigenous Peoples 2007
Wills, G., 1999. A Necessary Evil, A History of American Distrust of Government. Simon and Schuster, New York.

Civil Rights Movements

James M Vardaman, School of Culture, Media and Society, Waseda University, Shinjuku-ku, Tokyo, Japan

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A. Morris, volume 3, pp. 1883–1886, © 2001, Elsevier Ltd.

Abstract

This article explores the importance of the US Civil Rights Movement (1954–68), which originated in the American South and used nonviolent direct action to confront Jim Crow laws and regulations in an effort to force the federal government to dismantle the system of racial segregation and inequality.

Black Oppression

As late as the middle of the twentieth century, millions of black citizens in the US were socially oppressed, economically exploited, and politically disenfranchised. These forms of subjugation persisted in a nation viewed as the world's leading democracy. This view of the US was projected nationally and internationally by American leaders. The image of the nation as a democracy contrasted sharply with its treatment of its African-American citizens. Unlike European immigrants, with the exception of forced migrants including the Irish in the seventeenth and eighteenth centuries (Jordan and Walsh, 2007), black Africans were abducted and forcibly transported to America. For a period of two and a half centuries, enslaved blacks supplied a significant portion of the labor that enabled America to become an economic power. Slaves were not merely denied basic rights, they were officially defined as chattel, non-human beings. In apportioning representatives to the individual states, they were counted as three-fifths of a person, but they had absolutely no legal rights themselves. Due to the treatment they suffered, they faced serious health issues and enslaved women were forced to work throughout pregnancy (Fogel, 1994), often endangering their own lives as well as their children's.

The economic and social gap between the states of the North and South reached a breaking point with the election of Abraham Lincoln to the presidency in 1861. Southern states seceded from the Union and war broke out between the two sections. During what came to be called the Civil War (1861–65), President Lincoln issued the Emancipation Proclamation freeing the slaves in the southern states. With the victory of Union forces in 1865, black equality became a real possibility. For a brief period, it appeared that former slaves would be granted rights as American citizens. During the Reconstruction period (1865–77), blacks gained citizenship rights including freedom of movement, restricted male access to the franchise, and access to employment. This was a promising beginning, but it would not continue for long.

In the late 1870s, a system of racial segregation and unequal treatment, known as Jim Crow, emerged with the express purpose of returning blacks to white control. Leon F. Litwack (1979) and others describe in detail how the new form of oppression required segregation of blacks and whites in virtually every public encounter. The races were not allowed to use the same hospitals, schools, hotels, waiting rooms, or barber shops and beauty salons, nor could they sit on the same side of a courtroom, or be sworn in as a witness using the same Bible. One black recalled in an interview (Chafe et al., 2001: p. 5) that as a boy whenever he left home his mother told him to be sure to drink enough water to last him for a sufficient period of time because he would endanger himself if he attempted to drink from a public 'white' water fountain. Perhaps the most contentious locus was public transportation, which first appeared at the end of the nineteenth century and for which there were no customary habits to observe. Moreover, blacks were excluded from the political process, which meant that the law and its enforcers were all white. Toward the close of the nineteenth century, the emerging system called 'Jim Crow' received backing from the highest court of the nation. In 1896 the Supreme Court ruling in *Plessy v. Ferguson* declared that racial segregation did not violate the US Constitution. The court ruled that it was legal to require the races to use separate-but-equal facilities. With this ruling, Jim Crow laws and ordinances became national in scope, although they were more rigorously enforced in the South.

By the middle of the twentieth century, the majority of blacks were disenfranchised. In the South, blacks held no significant political offices and were constant victims of terror and violence, including lynching. In the labor market, blacks were restricted to undesirable low-paying jobs and subject to the will of white employers. Beyond economic subjugation, African Americans experienced daily personal humiliation because racial segregation marginalized them as a people and labeled them as an inferior race. They were stripped of human dignity; simple titles of respect such as 'Mr' or 'Mrs' were withheld, and even white children held symbolic authority over all blacks, however elderly or eminent.

The Movement's Origin

From the earliest days of slavery, African Americans staunchly resisted their harsh treatment by whites. Attempts at coordinated, violent resistance such as the Stono Rebellion (1739), Demark Vesey's Rebellion (1822), and Nat Turner's Rebellion (1831) invariably ended in failure, punishment, and attempts by whites to further restrict black movement and freedom (Berlin, 1998). While these explosive incidents of resistance were collective and public, covert resistance was more common. Slaves disobeyed orders, broke tools, pretended ignorance, slowed work routines, escaped for various lengths

of time, and released their frustrations by singing spirituals and telling stories about the self-importance and superiority felt by white people. Alan Dundee (1990) offers a collection of traditional 'toasts' and examples of 'signifying' that skillfully penetrated white pretentions of superiority and allowed blacks to share their feelings of resentment.

Beginning in the early twentieth century, black leaders like Booker T. Washington, W.E.B. Dubois, and Marcus Garvey attempted in their respective ways to gain rights and respect for black Americans. Perhaps the most significant group formed in this period was the National Association for the Advancement of Colored People (NAACP), founded in 1909 (Sullivan, 2009). The modern Civil Rights Movement drew upon this tradition and embedded itself deeply within the historic struggle for black liberation.

The movement emerged where black oppression was most intense and where racial segregation was most firmly entrenched: in the Southern states. Given the scope of oppression and the power of the white opposition, the birth of a powerful resistance movement was unanticipated, especially by whites who thought that segregation was part of the natural order and that blacks accepted it as such. Several factors contributed to a determination on the part of blacks to change the status quo.

First, after returning from World War II, black soldiers who had risked their lives to free citizens of the Allied nations were determined to work for freedom and equality for themselves in their own country (Wynn, 2010). Second, by the 1950s large numbers of African Americans had migrated to southern cities where they developed tightly knit urban communities and dense, effective communication networks (Dodson and Diouf, 2005: pp. 134–153). These resources made it possible to mobilize people. Third, the NAACP won several legal cases against the Jim Crow system, most significantly the 1954 Supreme Court decision in *Brown v. Board of Education*, which reversed the 1896 *Plessy v. Ferguson*, which had judged segregation to be constitutional. The Brown ruling stated that separate schooling based on race was unconstitutional. The implications of this ruling went far beyond this case for it delegitimized the entire system of racial segregation and encouraged struggles to implement the court orders (Ogletree, 2004).

Additionally, changing international relations were important to the birth of the movement. Former colonies in Africa and Asia were attempting to gain independence and establish new governments. African Americans identified with those struggles, which intensified their own thirst for freedom. Moreover, the context of decolonization and Cold War rivalry rendered the US federal government susceptible to pressure from an oppressed black population because the country sought to persuade these new Third World nations to model themselves after American democracy, not the Soviet alternative. However, racial oppression stood as a barrier to harmonious relations between the US and the emerging Third World nations. The proliferation of television and communication satellites made it possible to show white brutality to blacks to a worldwide audience. America's participation in the two world wars and the Korean War also rendered it vulnerable to black protest, because in those wars the US championed egalitarian values. In this context, black soldiers were radicalized while fighting for democracy on distant shores (Wynn, 2010). Thus, the federal government came under increased and sustained international and domestic pressure to support efforts to overthrow institutionalized racial segregation.

Strategies, Tactics, and Goals

These factors created the fertile soil from which the Civil Rights Movement emerged. By employing the strategy of direct action, this movement mobilized widespread social protest. For such a strategy to succeed, black movements had to disrupt white communities, businesses, and institutions and prevent them from doing business as usual. To this end, civil rights leaders and organizers mobilized thousands of African Americans to confront the Jim Crow institutions through social protest.

The leaders of the Civil Rights Movement succeeded in mobilizing massive nonviolent social protest. Innovative tactics included economic boycotts, beginning with a year-long boycott of a bus company in Montgomery, Alabama, sparked by the arrest of Mrs Rosa Parks, in December 1955, and led by Rev. Martin Luther King Jr. Essential to this early boycott were black churches, which offered the only safe place to hold 'mass meetings' at which the black community could share information that was unavailable through newspapers owned by whites, make plans for action, and encourage one another through sermons, speeches, and group singing of traditional spirituals, hymns, and newly written protest songs (Sullivan, 2009).

The churches cooperated closely with black colleges in the organized sit-in demonstrations that spread quickly when four black college freshmen sat down at a lunch counter in Greensboro, North Carolina in February 1960. When they were refused service, they politely remained seated and refused to move. The next day they sat down at the same counter with the same result. While some have viewed the initial sit-ins as independent, spontaneous acts on the part of impatient college students who were breaking from previous civil rights activities, Aldon Morris (1981) contends that, to the contrary, these sit-ins were carefully considered and resulted from close cooperation between leaders in the colleges and the black churches that allowed the demonstrations to continue and spread rapidly to other store and to other cities. The student sit-in movement of 1960 was a crucial development in that it injected new energy into the Civil Rights Movement, pointing the way to the efficacy of mass participation at the local, everyday level.

The breadth of participation eventually reached down into the lower age groups when dramatic confrontations of schoolchildren with police and firefighters in the streets of Birmingham, Alabama, in 1963, brought serious media attention to the violence white officials were inflicting on peaceful demonstrators (McWhorter, 2001). The largest, most widely reported mass mobilization of supporters of the movement came in August 1963 in the March on Washington, which culminated in Dr King's "I Have a Dream" speech.

The goal of these protests was to overthrow racial segregation and empower African Americans by seizing the franchise. Southern officials utilized their institutional power and the resistance of the larger white community in an intense, and often violent, effort to defeat the movement and to maintain

legally enforced racial segregation. Movement participants – many of them women, children, and college students – were beaten and brutalized by white southern law enforcement officials, and thousands were arrested and jailed for their protest activities. Others paid with their lives, including NAACP Mississippi field secretary Medgar Evers, who was assassinated in 1963 (Evers-Williams and Marable, 2005), and three civil rights workers who were killed near Philadelphia, Mississippi the following year. Nevertheless, the widespread and highly visible confrontations in the streets, which contrasted the brutality and the inhumanity of the white segregationists with the dignity and resolve of black protestors, made the cause of black equality the major issue in America for over a decade. The nation and its leaders were forced to decide publicly whether to grant African Americans their citizenship rights or to side with white supremacists and the undemocratic subjugation of African Americans.

National Achievements

The movement could not be dismissed. Eloquent leaders and their massive followings sustained the pressure on local elites and the federal government. Countless heroic figures inspired and organized a massive following. Among them were Rosa Parks, a dignified older woman and local NAACP activist, who sparked the Montgomery bus boycott when she defied an order to move to the back of the bus to the 'colored seats'; Martin Luther King, Jr., who emerged from the Montgomery bus boycott and the Southern Christian Leadership Conference (SCLC) to assume a position of preeminent moral leadership and national influence; James Forman, executive secretary of the more militant Student Nonviolent Coordinating Committee (SNCC), who was to challenge SCLC's and King's strategy; SNCC's leader, Stokely Carmichael (later Kwame Toure), who introduced the slogan 'Black power'; James Farmer and Floyd McKissick who led the Congress of Racial Inequality (CORE); and Fannie Lou Hamer, a sharecropper who went to a voter registration meeting run by SNCC, was arrested at the courthouse of Indianola, Mississippi, when she tried to register to vote, and then was viciously beaten in prison. Hamer, along with Parks, personified the thousands of women who played important roles in organizing and leading the movement.

These leading figures and thousands of movement participants articulated black suffering and the democratic aspirations of African Americans of every generation and circumstance. Thousands of whites (students, ministers, lawyers, and other civil rights workers) were inspired to participate in the Civil Rights Movement in various ways. Doug McAdam (1986, 1988) describes activities such as signing a petition supporting civil rights as 'low-risk/cost activism.' He examines how motivation for activism and a sense of membership in a group come into play in determining the amount of risk an individual is willing to take on in such a movement. His examination of those who elected to participate in the 1964 Mississippi Freedom Summer Project, a campaign that involved hundreds of volunteers in voter registration drives and the creation of 'freedom schools,' delineates some of the characteristics of participants like Andrew Goodman and Michael Schwerner who were willing to become 'high-risk/cost' activists and eventually lost their lives.

Differences regarding ideology, leadership style, and goals emerged within the movement as the SCLC, NAACP, SNCC, CORE, and other organizations reached different judgments about the effectiveness of nonviolent action, the desirability of racial integration and separatism, the role of whites in the movement, and the influence of black nationalism. Despite these controversies, the moral challenge and the widespread social disruption caused by the economic boycotts, the marches, the sit-ins, and other forms of nonviolent direct action, coupled with the international pressure, created an impasse in the nation that had to be resolved.

As a result, the Civil Rights Movement achieved important legislative victories in Congress. The landmark Civil Rights Act of 1964 outlawed discrimination in public accommodations on the basis of race, color, religion, or national origin and it identified the legal measures to be used to achieve racial integration. Moreover, it barred discrimination in employment practices on grounds of race, color, religion, national origin, or sex.

The passage of the 1965 Voting Rights Act significantly eliminated the use of literacy tests, authorized the attorney general to challenge the constitutionality of poll taxes, and introduced procedures that provided for the appointment of examiners to ensure that all restrictions on black voter registration be ended. In short, the Voting Rights Act enfranchised the southern black population, making it possible for a historic black elected political class to emerge.

International Achievements

The significance of the Civil Rights Movement extends far beyond its historic overthrow of the Jim Crow laws. The movement has affected US politics in fundamental ways. It demonstrated to the oppressed black community how concerted effort could be successful, and it made social protest respectable. The Civil Rights Movement also proved that social protest is capable of generating significant change. Hence, the movement broadened the scope of US politics and inspired diverse movements for citizenship rights and social justice in the US and abroad. Before the movement, many groups in America – including women, Hispanics, Native Americans, farm workers, the physically disabled, gays, and lesbians – were oppressed but unaware of how to resist or galvanize support. The Civil Rights Movement provided a model of successful social protest and produced a host of new tactics and social change organizations. Moreover, the movement influenced freedom struggles around the world. Participants in freedom struggles in Africa, Eastern Europe, the Middle East, Latin America, and China have made it clear that they were deeply influenced by the US Civil Rights Movement.

Continuing Challenges

For all its success and influences, however, the Civil Rights Movement did not solve all of America's racial problems. In the second decade of the twenty-first century, despite a black American being elected to the presidency, many African

Americans and other non-white groups are still at the bottom of the social and economic order. These current conditions are exacerbated by a social climate in which the disadvantaged are often blamed for their own predicament. Thanks to the Civil Rights Movement, a relatively large black middle class has emerged. Yet, over a third of the black community remains trapped in poverty (Wilson, 2009; Klinker and Smith, 1999; Joseph, 2010). Black Americans are disproportionately the victims of police brutality and housed in the rapidly growing prison industry. Thus, some blacks enjoy middle-class comforts while millions of others are victims of lack of opportunity, hopelessness, and high rates of imprisonment. Poverty and inequality are also widespread outside the black community. It may be that organized protest remains the only viable means to achieve greater empowerment. If this is the case, the Civil Rights Movement has left a rich legacy to inspire and inform future struggles. The sparking of a renewed interest in the academic study of resistance and social protest is an unnoticed legacy of the Civil Rights Movement. Through such studies it is possible that new mechanisms of social change will come to light.

See also: Civil Liberties and Human Rights; Civil Rights; Integration, Social; Minorities; Race Identity; Racial Relations; Racism, History of; Racism, Sociology of; Slavery as Social Institution; Slaves and Slavery, History of.

Bibliography

Arsenault, Raymond, 2006. Freedom Riders: 1961 and the Struggle for Racial Justice. Oxford University Press, Oxford.

Berlin, Ira, 1998. Many Thousands Gone: The First Two Centuries of Slavery in North America. Belknap Press of Harvard University Press, Cambridge.

Berlin, Ira, 2003. Generations of Captivity: A History of African-American Slaves. Belknap Press of Harvard University Press, Cambridge.

Berlin, Ira, 2010. The Making of African America: The Four Great Migrations. Viking, New York.

Bordewich, Fergus M., 2005. Bound for Canaan: The Underground Railroad and the War for the Soul of America. Amistad, New York.

Branch, Taylor, 1988. Parting the Waters: America in the King Years 1954–63. Simon and Schuster, New York.

Chafe, William H., et al. (Eds.), 2001. Remembering Jim Crow: African Americans Tell about Life in the Segregated South. The New Press, New York.

Cottrol, Robert J., et al., 2003. Brown v. Board of Education: Caste, Culture and the Constitution. University Press of Kansas, Lawrence.

Daniel, Pete, 2000. Lost Revolutions: The South in the 1950s. University of North Carolina Press, Chapel Hill.

Doyle, William, 2003. An American Resurrection: James Meredith and the Battle of Oxford, Mississippi. Anchor, New York.

Dodson, Howard, Diouf, Sylvaine A., 2005. The Schomburg Center for Research in Black Culture. In: Motion: The African-American Migration Experience. National Geographic, Washington, DC.

Doyle, William, 2003. An American Resurrection: James Meredith and the Battle of Oxford, Mississippi. Anchor, New York.

Dundee, Alan (Ed.), 1990. Mother Wit: From the Laughing Barrel. University of Mississippi Press, Jackson.

Evers-Williams, Myrlie, Marable, Manning (Eds.), 2005. The Autobiography of Medgar Evers. Basic Civitas, New York.

Fogel, Robert W., 1994. Without Consent or Contract: The Rise and Fall of American Slavery. W.W. Norton and Company, New York.

Hendrickson, Paul, 2003. Sons of Mississippi. Alfred A. Knopf, New York.

Joseph, Peniel E., 2010. Dark Days, Bright Nights: From Black Power to Barack Obama. Basic Civitas Books, New York.

Jordan, Don, Walsh, Michael, 2007. White Cargo: The Forgotten History of Britain's White Slaves in America. New York University Press, New York.

Kasher, Steven, 1996. The Civil Rights Movement: A Photographic History: 1954–68. Abbeville Press, New York.

Klinker, Philip A., Smith, Rogers M., 1999. The Unsteady March: The Rise and Decline of Racial Equality in America. University of Chicago Press, Chicago.

Layton, A.S., 2000. International Politics and Civil Rights Policies in the United States 1941–1960. Cambridge University Press, Cambridge, UK.

Litwack, Leon F., 1979. Been in the Storm So Long: The Aftermath of Slavery. Vintage Books, New York.

Litwack, Leon F., 1998. Trouble in Mind: Black Southerners in the Age of Jim Crow. Vintage Books, New York.

Litwack, Leon F., 2009. How Free is Free?: The Long Death of Jim Crow. Harvard University Press, Cambridge.

McAdam, Doug, 1988. Freedom Summer. Oxford University Press, New York.

Marable, Manning, 2011. Malcolm X: A Life of Reinvention. Viking, New York.

McAdam, Doug, July 1986. Recruitment to high-risk activism: the case of Freedom Summer. American Journal of Sociology 92 (1), 64–90.

McWhorter, Diane, 2001. Carry Me Home: Birmingham, Alabama, the Climactic Battle of the Civil Rights Revolution. Touchstone, New York.

Morris, Aldon D., 1981. Black southern student sit-in movement: an analysis of internal organization. American Sociological Review 26 (6), 744–767.

Morris, Aldon D., 1999. A retrospective on the Civil Rights Movement: political and intellectual landmarks. Annual Review of Sociology 25, 517–539.

Ogletree Jr., Charles J., 2004. All Deliberate Speed: Reflections on the First Half-Century of Brown v. Board of Education. W.W. Norton & Company, New York.

Robnett, B., 1997. How Long? How Long? Oxford University Press, New York.

Sullivan, Patricia, 2009. Lift Every Voice: The NAACP and the Making of the Civil Rights Movement. The New Press, New York.

Till-Mobley, Mamie, 2003. Death of Innocence: The Story of the Hate Crime that Changed America. One World, New York.

William, Juan, 1987. Eyes on the Prize: America's Civil Rights Years 1954–1968. Penguin, New York.

Wilson, William J., 2009. More than Just Race: Being Black and Poor in the Inner City. W.W. Norton & Company, New York.

Wynn, Neil A., 2010. The African American Experience during World War II. Rowman and Littlefield Publishers, Lanham, MD.

Civil Service

Itzhak Galnoor, Hebrew University, Jerusalem, Israel; and The Van Leer Institute, Jerusalem, Israel
Jennifer Oser, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The civil service is the generic name given in English to the administrative apparatus of the state. Historically, bureaucratic administrations were developed in Egypt and China to serve the rulers, or the dynasty. These bureaucrats were engaged in activities such as land registration, water allocation, tax collection and above all managing war-related affairs. The emergence of the *modern* civil service is directly connected to the crystallization of the European-style state. A professional, career civil service (based on entrance examinations) was first introduced in Prussia and France and subsequently in Britain and the United States. In democratic states, the professional civil service is assumed to exist as a differentiated and mostly politically neutral institution. By contrast, in nondemocratic regimes the public bureaucracy operates sometimes from the ruler's palace, the military barracks, or the party's headquarters. Civil servants are in charge of many different activities that can be grouped under three headings: shaping and implementing public policy; providing services to individuals, groups, and organizations; and administrating regulatory schemes. These areas of responsibility shift over time. In the 2000s, the modern version of the state is changing and the size, structure, and functions of the civil service are in transition and transformation. Independent public and private entities and agencies are performing what used to be civil service tasks. Other functions have been privatized or 'outsourced' and if a 'skeleton state' will emerge, it may shatter many of the old features of the classical civil service model. Furthermore, new forms of collaborative governance have emerged, which bring together civil service departments and a range of private and voluntary organizations, causing the boundaries between the sectors of government, civil society, and business to become increasingly blurred.

Introduction

The civil service is the generic name given in English to the administrative apparatus of the state. The term was first introduced in the British administration in India and then in the UK (1854), and has become almost universally synonymous with civilian (i.e., nonmilitary and nonjudicial) administrators employed by central governments. The civil service is also referred to by different terms (e.g., 'public bureaucracy'), and it differs from other bureaucracies by virtue of its public missions, however defined. It has a collective justification, namely, doing something otherwise unattainable for a certain public end, and therefore, civil servants have been entrusted with the task of being the guardians, or at least the interpreters, of the common good. Another term loosely used is 'the public sector' as distinguished from the private (business) sector. It may include all the not-for-profit organizations, governmental and nongovernmental, as well as local and regional authorities. The modern civil service is above all a 'state institution' and this article will focus on the changing relationship between the state and its specialized bureaucratic institution – the civil service.

History

There were bureaucratic administrators long before they were known as 'civil servants.' Historical antecedents can be found mainly in centralized states and empires where public bureaucracies were developed to serve the rulers or the dynasty. Despite their subservience to the emperors and kings, however, these bureaucracies became 'professional,' that is, they manifested some conception of themselves as servants of the 'state,' the polity, or even the community (Eisenstadt, 1965). Such embryonic civil service institutions existed in different forms first in the Egypt and China ancient empires, and later, for instance, in the Byzantine, Sassanid, Abbasid, and Ottoman empires. Sculptures of the ancient Egyptian scribes date back to the Old Kingdom (*c*.2600 BC). They belonged to a body of officials who served the Kingdom, as did their Assyrian and Babylonian counterparts. Even more professional were the Chinese functionaries employed in the imperial court. These officials – better known as 'mandarins' – were selected by rigorous examinations, paraded through the villages with servants carrying their desks and proudly displaying their rank in the shape and color of their hat tassels. The term 'mandarin' was later used pejoratively to indicate the aloofness of civil servants, the official jargon, and the sheltered life of their long careers.

Broadly speaking, the history of public administration testifies that there has always been a need to link state (or ruler) goals, with the official expertise of a bureaucracy that tends to be centralized. These bureaucratic entities were engaged in three main types of activities: technical services such as land registration or water allocation; social and political regulation such as administrating justice, collecting taxes, or supervising weights and measures; and above all – managing war-related affairs (Finer, 1997: vol. 1, pp. 59–72).

The European patterns of the civil service have emerged gradually from the Greek and especially the Roman law and administrative traditions of various public functions such as public construction, food supply regulation, and population census. They were later influenced by the medieval feudal system of financial and judicial administration. Yet, the most strenuous collective effort has always been related to mobilizing money, soldiers, and equipment for the military. Accordingly, throughout the history of most countries, public

and private organizational practices have been influenced directly by the most readily available example – the military model. This model was further developed to encompass the colonial expansion of some European countries, which required new organizational expertise, both in content and in scope. The influence of the military is most readily seen through such administrative concepts as 'chain of command,' or 'line and staff,' so much so that attempts to 'debureaucratize' the civil service (e.g., creating less hierarchical departments) are essentially an attempt to break away from the dominance of the military bureaucratic model.

The Modern Civil Service

The emergence of the modern civil service is connected to the crystallization of the European-style state, first in France and Prussia and later in all the 200 or so states that populated the globe in the early twenty-first century. As an institution, the civil service dates back to the middle seventeenth century with the inauguration of competence entrance examinations in Prussia under Friedrich Wilhelm I, analogous to the old Chinese practice. Equally important, the German Cameralists were the first modern scholars–practitioners to study public administration and to offer a set of principles for the 'management of the state' (Gross, 1964: vol. I, p. 108). The next stages in the development of the civil service were diffused, but a few milestones can be presented, all of them related to the attempt to create an institutional instrument with acquired experience in running state affairs. The dissimilar features should also be noted as each civil service is also a product of its particular environment, and the human composition of each of them reflects recurring attempts of social groups to enter the ruling elites of their respective states.

The Napoleonic Reforms in the early nineteenth century converted the previous French royal service into a State public service. France, and before that Prussia, created the continental model of a central, professional, tenured, and in-house trained civil servant. More than in other countries, continental civil servants embody the state, and the continuity of state affairs, and operate within a specific legal framework of administrative law (Crozier, 1964; Suleiman, 1974).

The Northcote–Trevelyan report (1854) in Britain introduced competitive examinations into the civil service and uniform methods of recruitment and promotion across departments. It eventually succeeded in eliminating patronage. As the senior administrative class became a professional life career it attracted educated professionals who were given the task of shaping public policy and advising politically elected ministers. Below senior civil servants were the executive and clerical classes, entrusted respectively with managerial and routine responsibilities. The UK was the first to establish a civil service commission to oversee the entire operations of the civil service. Politically neutral and very secretive, the British senior home civil service acquired a reputation of integrity and impartiality, and of being rather conservative guardians of crown affairs, as distinct from the governments of the day.

The Pendelton Act (1883) in the United States marked the beginning of federal civil service reform, aimed at abolishing the previous 'spoils system' in the government. Patronage in the federal government was closely related to electoral turnovers in the Presidency and Congress. Unlike the UK, the American system has never achieved the same degree of uniformity, political neutrality, or continuity. However, the American civil service has been much more open, with recruitment to the senior positions less dependent on social class and on elite education systems. Consequently, the U.S. civil service has never had the aura of 'officialism' as in Europe, and has not acquired the prestige and the consequent status of being a permanent agent of the federal state.

Nondemocratic Regimes

In nondemocratic regimes, the position of the civil service institution is related directly to the scope of the state's monopoly of the different spheres of life. The stronger the state, the more dominant is the public bureaucracy – operating sometimes from the ruler's palace, the military barracks, or the party's headquarters. In such regimes, the civil service hardly exists as a differentiated professional institution. In the former Soviet Union, for example, the lines between state and party bureaucracies, and indeed between political and administrative decisions, were (and are still today in China) practically nonexistent. In this respect, the most troubling phenomenon appeared with the total submission of the highly developed German civil service (including the military and the judicial) to Nazi leaders. This raises the question: Can there be safeguards against turning professional competence – and loyalty – into a sharp tool in the hands of evil political masters?

New States

In many new states, the civil service is a mixture of colonial heritage and unique local features. In India, and in some other former British colonies, the civil service is a strong institution, and it helps to keep the country going, despite recurring crises and accusations that it impedes development. In other new countries, particularly where the military took over, the civil service hardly functions. It cannot maintain law and order, let alone carry the task of providing social services. The weaker the state, the more helpless is the civil service institution; the weaker the civil service, the lesser its ability to maintain stability and introduce positive changes. However, nondemocratic states may have very strong military or even civilian bureaucracies. When state-building efforts are proposed, the importance of the civil service institution is duly stressed (Fukuyama, 2004). Accordingly, the popular revolutions in the Arab world that started in 2011 used the slogan of 'democratizing' the state civil service.

The Rational Model

Max Weber (1864–1920) delineated three 'ideal-types' of legitimate authority – the legal-rational, the traditional, and the charismatic, and suggested that the first one requires the existence of a 'rational' administrative staff. Strongly influenced by the Prussian example, Weber drew a list of the characteristics of a modern bureaucracy and the conditions that contributed to its emergence and growth, particularly in capitalist market

economies (Weber, 1947). His ideas reflected, and to some extent strengthened, the heavy reliance – both in continental Europe and in the United States – of civil service institutions upon administrative law: legal codes, rules, regulations, and precedents. The notable exception was the UK that (unlike Canada, Australia, and its other former dominions) did not have a civil service act of Parliament until the Constitutional Reform and Governance Act of 2010.

The rational ('Weberian') model of the civil service as a permanent institution – influenced by the military model – also includes the following common features, which are assumed to exist in the public bureaucracies of the developed states: the centralization of authority; formal modes of operation; a central commission or agency in charge of civil service affairs; established procedures for recruitment, tenure, promotion, compensation, job evaluation, discipline, and the right to associate and strike; and arrangements for securing political neutrality. In addition, civil servants must adhere to a special service code of conduct which specifies their obligations, and constitutes the ethos (or perhaps the myth) of impersonality, neutrality, and guardianship of state secrets.

The important point about the rational model of the civil service is that there is a connection between the development of the modern state and this new form of bureaucratic organization. Indeed, in the twentieth century the civil service and the entire public sector expanded rapidly in response to growing roles of the state, particularly with the growth of the welfare state. Governments and their civil service apparatus have been given the task of providing more services and of finding answers to new problems such as environmental degradation. However, at the same time, the 'administrative' or 'bureaucratic' state has also come under attack, in terms of the size, structure, and functions of the civil service. New post-Weberian questions have been floated: to whom should the civil service be responsive and accountable: to the state, the law, the government of the day, the multicultured conglomeration of sectors, groups, and citizens, or perhaps to the individual customers of its services? These questions paved the way for new democratic demands: citizen involvement and participation in policy making, in addition to the indirect supervision of the civil service through its political masters.

What Do Civil Servants Do?

Civil servants are in charge of a rich menu of government activities ranging from artificially 'seeding clouds' in the sky to induce rainfall, to operating schools, or administrating programs for the increasing number of aged people. In between, civil servants, as in ancient times, continue to collect taxes and perform state functions that cannot be trusted to or executed by social (voluntary) or private market institutions. These various activities can be grouped under the following headings: shaping and implementing public policy decisions; providing services to individuals, groups, and organizations; and administrating regulatory functions in areas such as aviation, drugs, and election financing.

The main structural entities for carrying out these activities could be divided roughly into: (1) the regular civil service departments and agencies – responsible for the generic governmental functions; (2) the special statutory agencies, authorities and commissions (e.g., the U.S. Securities and Exchange Commission) responsible for specific tasks removed from the regular civil service, increasingly including regulatory tasks; and (3) government corporations entrusted with running mainly utilities and other commercial enterprises regarded as natural monopolies (e.g., electricity companies; regional development projects), or in some other way related to merit public goods and national interests (e.g., national forests, postal services, public broadcasting). The combinations of the various types of activities and the different structures reflect the myriad and ever-changing scope of civil service responsibilities, including 'joined-up government' or 'collaborative governance' (Bogdanor, 2005; Donahue and Zeckhauser, 2011) – aimed at bringing together different government agencies and private and voluntary bodies, thus blurring the previous boundaries of the civil service.

It is impossible to prescribe the right number of civil servants per capita, given the differences in their functions, and in the number of intrastate governmental levels in different states. In economically developed states, the tendency is to cut down the scope of civil service activities (and consequently its size as well) through privatization, outsourcing, and delegation of responsibilities to nonprofit civil society organizations. In new states, the pressure still exists on the civil service to do much more for their hard-pressed societies.

Studying the Civil Service

Woodrow Wilson (1887) was among the pioneers who attempted to create 'a practical science of administration,' aimed at improving not only the personnel, but also the organization and methods of government offices. He saw the field of administration as a field of business that should be outside the sphere of politics. Hence his famous dictum: separate policy making from administrative execution. However, there were other scholarly currents as well. From the 'scientific management' movement in the United States (*c.*1900s) came the notion of 'efficiency,' aimed at increasing workers' productivity through detailed time and motion studies, standardization of tools, and careful attention to training. From the 'human relations' school (*c.*1930s) came the behavioral message that in all organizations, individual and group motivations are critical factors. A little later, and more directly applied to public administration, attempts were made to develop prescriptive principles. For instance, establish one top executive; fit people to organizational structures; ensure unity of command; use specialized staff; maintain homogeneity in organizational subdivisions; delegate authority; match responsibility with authority; and limit span of control (Gulick and Urwick, 1937).

Theoretically, these 'principles' were not only culturally restricted, but inherently anecdotal and Herbert Simon (1946) regarded them as no more than 'proverbs of administration.' Simon's concept of 'bounded rationality' that governs the life of all organizations, was a step away from prescriptions toward observations, and from practical advice on how to run organizations to laying the foundation of public administration theory and of a comparative perspective. First came the formal studies of the civil service as an institution within the executive

branch, or in the context of administrative law. Later, mainly because of the impact of studies in political science, the context as well as the horizon were expanded, away from the previous formalism. For example, power, group theory, and communication were introduced into the vocabulary of public administration research.

The related area of public policy has recently emerged with new emphasis on economic models and game theory, as well as on the redefinition of values, public goods, and the role of politics in decision-making (Majone, 1989). New foci were introduced such as implementation, social regulation, and comparative policy analysis in areas such as welfare, health, and environment (Goodin and Klingemann, 1996). More recently, the nature of policy making has been studied empirically with emphasis on policy processes and transformations (Jones and Thomas, 2013).

The most challenging and inadequately researched new area has remained the role of the civil service institution in democracies. This is a practical issue as well: how to reform the civil service within a new democratic framework in the era of state weakening (Silberman, 1993).

A New Version of the Civil Service?

In the 1980s, a number of Organization for Economic Cooperation and Development countries adapted a 'New Public Management' (NPM) doctrine that was founded on principles of organizational best practices and public accountability (Hood et al., 1999). In NPM reforms were introduced, a new post-public bureaucracy model was assumed to emerge with entirely new roles for the civil service (Peters and Wright, 1996). A key assumption was that the NPM reforms would lead to a diminished public bureaucracy, and perhaps even a nonexistent civil service (Olsen, 2008).

The institute of the state continues to change rapidly in the twenty-first century, and if there is going to be a 'skeleton state,' there will be a skeleton civil service as well. This process has been gradually occurring both because of internal changes such as the weakening of public trust in political institutions, and under the impact of globalization. In recent years there are states still struggling to build a British-type civil service (which does not exist in the UK anymore); other states in which few changes have taken place in their public bureaucracies; and economically rich states in which the traditional role of the civil service is visibly contracting. There are also new suprastate bureaucracies, not only in international organizations such as the United Nations or the European Union, but also in interstate agencies and in joint public management of regional projects. These developments have shattered many of the old features of the classical Weberian-type civil service.

Examples:

1. Civil service monopoly on official information is no longer feasible and government secrecy has yielded to freedom of information acts enacted in almost all democratic states (Galnoor, 1989). State secrets too are no longer immune, as exemplified by the disclosure of numerous highly classified U.S. government documents by E. Snowden in 2013 (Harding, 2014).
2. With modern information technology and social networks, civil servants' discretion in shaping public policy (and perhaps anonymity too) has been gradually disappearing.
3. The previous civil service – interest groups network, has been replaced by a much more intricate and uncontrollable system of organization in the mushrooming 'third sector' (nongovernmental, nonprofit) of public groups, NGOs, and philanthropic funds.
4. The civil service is under great pressure to be more 'representative' in its membership, or at least to assist in achieving equal opportunity through 'affirmative action' and other 'open government' measures.
5. Direct accountability of the civil service to the public is demolishing the last walls of insulation. One example is the spread of the Swedish Ombudsman institution. Another is the authority given to special investigation commissions to probe the decision-making discretion of public officials.

These developments were followed by changes in the traditional features of the civil service. For example, in many states the civil service is no longer a life-long career and new opportunities for entrance to (and exit from) senior positions have been introduced. The emphasis on management skills has downgraded the policy-advice roles of civil servants. The results-oriented business approach has diverted attention to the quality of service aspect, including the idea that civil servants should be rewarded according to their standing in customers' satisfaction surveys.

The increased prominence of a business approach in the civil service has influenced the relationship between the civil service and the public to whom it is accountable (Vigoda, 2002). Even as the civil service increasingly relates to citizens as clients, recent channels have developed for including citizen input in policy making and decision making, whether through networks, nongovernmental organizations, or through direct democracy initiatives (Fung, 2006). A related development is the slow emergence of 'e-Government' as a tool for public management and innovation. The original use of Information and Computing Technologies (ICT) was for increased efficiency of internal civil service functioning, but they became used more frequently to improve the quality of service provision. The most promising ICT use is to support citizen involvement and participation in democratic policy making. Indeed, there is evidence that these technologies facilitate innovation in the public sector (Tolbert et al., 2008). Yet, it remains to be seen whether these tools will meaningfully change the previous one-sided relationship between civil servants and the public (Hood, 2006).

The Impact of Privatization on the Civil Service

As part of the process of 'reshaping the state,' the size, structure, and functions of the civil service are in transition, with independent public and private entities and agencies performing what used to be civil service tasks. Other functions have been 'privatized,' 'marketed,' or 'outsourced': running prisons, computer services, or even recruitment to the civil service itself. Privatization in its narrow, technical sense is the transfer of assets, services, and goods from the ownership and/or

management and/or financing by state and substate governmental organizations to market organizations, or to organizations of the nonprofit sector. From a broader perspective, privatization is a new ideological definition of state responsibility aimed at shifting the boundaries between public and private (or market) allocation. The premise behind this approach is that goods and services are produced more efficiently and with less opportunity for corruption by private organizations. Based on these assumptions, a wide range of public services has been privatized since the 1980s in many countries, thus reducing the role and responsibility of the state, particularly the intervention of the executive branch, in economic and social affairs (Galnoor, 2011: pp. 150–158).

Privatization and outsourcing have led to increasing ambiguity about the boundaries of responsibility between the private and public sectors, as well as to some attempts at cooperation between the two. Another outcome has been the growth of a powerful third sector that has appropriated many areas traditionally considered the responsibility of the state, either willingly or out of necessity. The results of these processes in many countries have been that government ministries and local authorities virtually ceased functioning in certain areas, while private and voluntary organizations operate in their stead. The broader context is society's transitions from one based on a civil orientation that places heavy obligations on the state to one based on a consumer orientation that perceives its citizens as clients. This has resulted in increased regulation, but in many cases, privatization did not produce real competition or better, cheaper services. The ability of the state to regulate lagged behind the increased power and expertise of the bodies in charge of privatized and contracted-out goods and services. In addition, the assumption that privatization will reduce corruption has not been at all substantiated. Areas of privatization can be categorized as follows:

1. The Privatization of Public Goods: the government's exit from activities such as road building, mining, and communications. The assumption is that privatization will lead to increased competition and thus greater efficiency and improved services. Examples are the sale of government-owned corporations, such as oil refineries, ports, roads, and railways. Privatizations such as these evoke questions about the transparency of the process, the appropriate compensation for public assets transferred, the danger of creating private monopolies accountable to no one, the balance between beneficiaries and losers, the violation of labor rights and agreements, and the effectiveness of postprivatization regulation.
2. The Privatization of Public Services: usually contracting out health services, education and welfare previously delivered by the state or local authorities. This type of privatization comes about because of a gradual reduction in services, cutbacks in funding, creating gaps often filled by private providers, for example, cutbacks in law enforcement have led well-to-do homeowners to install security cameras and employ private security firms.
3. The Privatization of Operations: appears under different names such as outsourcing, partnerships, franchises, subcontracting, and voucher distribution. This type of privatization may be full or partial, involving either the acquisition of private services by the government or the transfer of operations to the private or voluntary sector. The responsibility remains in the hands of the state, sometimes effectively regulating the delivery of the product or the services, and sometimes helplessly watching its deterioration. These arrangements, prevalent in the business sector, are often motivated by a desire to scale down the public sector or increase so-called 'flexibility' in management and employment. Privatization of operations can lead, however, to the violation of workers' rights, as may occur when work is outsourced to private service contractors. If outsourcing involves the transfer of policy making out of the civil service, one must consider the potential injury to the public and particularly to helpless people.

Thus, privatization has been a profound reform in the structure and functions of the civil service, raising many questions about the role of the state and its regulation performance.

The Civil Service and Regulation

Regulation refers to laws and rules that are binding on the conduct of organizations and individuals, and the regulator is responsible for formulating and enforcing these rules. This definition refers primarily to the oversight of business organizations by the state. A broader definition emphasizes the social–political role: focused and continuing supervision by a public body over socially significant activities (Selznick, 1985). In this definition, regulation is more than law enforcement and its aim is to attain social goals, including the proper functioning of the economy.

The civil service deals with regulation through legislation, rules, and other regulatory mechanisms. Regulation laws are concerned with licensing professionals in sensitive occupations (such as accounting), safeguarding the individual's property rights (patents), ensuring the functioning of the financial market (securities), preventing epidemics (food), and supervising the use of a collective good such as space (satellites). Other regulations seek to cope with 'market failures,' such as the absence of competition, the marketing of dangerous products, the lack of transparency to consumers, and environmental pollution. The list could also include common social motivations such as regulating the public labor force by the use of affirmative action, supervising the use of drugs, and control over the use of public resources such as the coastline.

Opponents of regulation associate it with bureaucratic red tape, and point out the inflexibility, convoluted procedures, excessive control and demand for permits, and its high cost. Proponents of regulation point out that we, as citizens, want the state to oversee greater regulation in many key areas, such as protection from building violations, prevention of safety and traffic hazards, environmental preservation, and so forth. Others believe that a 'regulatory state' is about to replace the 'dirigiste state' that preceded it and regulation will thus supersede state ownership of organizations; rather than government intervening in planning and implementation, it will focus on oversight and control (Majone, 1994; Levi-Faur, 2011).

Blurred Boundaries of the Civil Service and Continued Relevance

As the enthusiasm for New Public Management reforms faded in the twenty-first century, the three main tasks of public servants mentioned above (i.e., shaping and implementing policy decisions; providing services; and regulating public functions) are still within the responsibility of the public bureaucracy. Even as these tasks are increasingly contracted out, the expertise of the civil servants who contract for and regulate these tasks is still required. The most important change that has occurred, therefore, is the blurring of the old boundaries between the public and private sectors, and between politics and administration.

Merriam (1944) pointed out the blurred boundary long ago: the civil service has constituencies, and in many countries, corporations exert a great deal of influence on public policy. However, with increased regulation, the private sector became even more involved, and hence the phenomenon of excessive clientelism (Nadel and Rourke, 1975). Increased globalization and the diffusion of public management practices between countries have also blurred the autonomous domain of the civil service (Matthiasan, 2005).

As for the politics–administration dichotomy, for many years it disappeared from public administration textbooks. Everybody knew that civil servants exercise state power both formally (law application) and informally (policy shaping). However, starting in the 1980s the distinction returned as part of NPM. The textbooks dealing with the civil service are called 'introduction to public sector management,' and the civil service is conceived to be market-oriented, customer-driven, deregulated and decentralized. Thus, the old Wilsonian dichotomy has returned through the back door: an institutional separation within the civil service between the 'departments' in charge of policy and the 'performance-based organizations' or 'executive agencies' responsible for delivering services to customers.

What is therefore the outlook for the institution of the civil service? New democracies have learned the hard way that state-building means institution-building and that a key institution is an old-fashioned autonomous civil service. In advanced countries, crippling the civil service resulted in diminished state capacity to deal with natural disasters, or with financial corruption in the private sector. Moreover, it seems to curtail the collective ability to invest in innovation and there are calls for reinstating the 'entrepreneurial state' (Mazzucato, 2013). If this line of thinking will prevail, the continued relevance of the civil service will be assured.

See also: Agency Theory; Bounded Rationality; Bureaucracy, Sociology of; Bureaucratization and Bureaucracy, History of; Executive Government; New Managerialism and New Public Sector Management; Public Administration: Organizational Aspects; Weber, Max (1864–1920); Weberian Social Thought, History of.

Bibliography

Aberbach, J.D., Putnam, R.D., Rockman, B.A., 1981. Bureaucrats and Politicians in Western Democracies. Harvard University Press, Cambridge, MA.
Barnard, C., 1938. The Functions of the Executive. Harvard University Press, Cambridge, MA.
Bogdanor, V. (Ed.), 2005. Joined-Up Government. Oxford University Press, Oxford.
Crozier, M., 1964. The Bureaucratic Phenomenon. University of Chicago Press, Chicago.
Donahue, J.D., Zeckhauser, R.J., 2011. Collaborative Governance: Private Roles for Public Goals in Turbulent Times. Princeton University Press, Princeton.
Dror, Y., 1971. Design for Policy Sciences. Elsevier, New York.
Eisenstadt, S.N., 1965. Essays in Comparative Institution. Wiley, New York.
Ferlie, E., Lynn Jr., L., Pollitt, C. (Eds.), 2005. The Oxford Handbook of Public Management. Oxford University Press, Oxford.
Finer, S.E., 1997. The History of Government from the Earliest Times. Oxford University Press, Oxford.
Fukuyama, F., 2004. State-Building: Governance and World Order in the Twenty First Century. Cornell University Press, Ithaca.
Fung, A., 2006. Varieties of participation in complex governance. Public Administration Review 66, 66–75.
Galnoor, I., 1989. Government secrecy. In: International Encyclopedia of Communications. Oxford University Press, Oxford, pp. 34–37.
Galnoor, I., 2011. Public Management in Israel. Routledge, London.
Goodin, R.E., Klingemann, H.-D. (Eds.), 1996. Public policy and administration. A New Handbook of Political Science. Oxford University Press, Oxford, pp. 551–641.
Gross, B.M., 1964. The Managing of Organizations. The Free Press, New York.
Gulick, L., Urwick, L. (Eds.), 1937. Papers on the Science of Administration. Institute of Public Administration, New York.
Harding, L., 2014. The Snowden Files: The Inside Story of the World's Most Wanted Man. Vintage Books, London.
Hood, C., 2006. The tools of government in the information age. In: Moran, M., Rein, M., Goodin, R.E. (Eds.), The Oxford Handbook of Public Policy. Oxford University Press, Oxford, pp. 469–481.
Hood, C., et al., 1999. Regulation Inside Government. Oxford University Press, Oxford.
Jones, B.D., Thomas, H.F., 2013. Bounded rationality and public policy decision-making. In: Howlett, M., Ramesh, M., Wu, X. (Eds.), Routledge Handbook of Public Policy. Routledge, London, pp. 273–286.
Levi-Faur, D. (Ed.), 2011. Handbook on the Politics of Regulation. Edward Elgar, Cheltenham.
Majone, G., 1989. Evidence, Argument and Persuasion in the Policy Process. Yale University Press, New Haven, CT.
Majone, G., 1994. The rise of the regulatory state in Europe. West European Politics 17 (3), 77–101.
Mazzucato, M., 2013. The Entrepreneurial State: Debunking Public vs. Private Sector Myths. Anthem Press, London.
Matthiasan, D., 2005. International public management. In: Ferlie, E., Lynn Jr., L., Pollitt, C. (Eds.), The Oxford Handbook of Public Management. Oxford University Press, Oxford, pp. 644–670.
Merriam, C.E., 1944. Public and Private Government. Yale University Press, New Haven, CT.
Nadel, M.V., Rourke, F.E., 1975. Bureaucracies. In: Handbook of Political Science, vol. 5, pp. 373–440.
Olsen, J.P., 2008. The ups and downs of bureaucratic organization. Annual Review of Political Science 11 (1), 13–37.
Peters, B.G., Wright, V., 1996. Public policy and administration: old and new. In: Goodin, R.E., Klingemann, H.-D. (Eds.), A New Handbook of Political Science. Oxford University Press, Oxford, pp. 628–641.
Pressman, J.L., Wildavsky, A., 1984. Implementation, third ed. University of California Press, Berkeley, CA.
Selznick, P., 1985. Focusing organizational research on regulation. In: Noll, R.G. (Ed.), Regulatory Policy and the Social Sciences. University of California Press, Berkeley, pp. 363–367.
Silberman, B., 1993. Cages of Reason: The Rise of the Rational State in France, Japan, the United States and Great Britain. University of Chicago Press, Chicago.
Simon, H.A., 1946. The proverbs of administration. Public Administration Review 6, 53–67.
Simon, H.A., 1957. Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations, second ed. Macmillan, New York.
Suleiman, E.N., 1974. Politics, Power, and Bureaucracy in France: The Administrative Elite. Princeton University Press, Princeton, NJ.
Tolbert, C.J., Mossberger, K., McNeal, R., 2008. Institutions, policy innovation, and e-government in the American states. Public Administration Review 68 (3), 549–563.
Vigoda, E., 2002. From responsiveness to collaboration: governance, citizens, and the next generation of public administration. Public Administration Review 62 (5), 527–540.
Weber, M., 1947. The Theory of Social and Economic Organization. Oxford University Press, New York.
Wildavsky, A., 1987. Speaking Truth to Power. Transaction, New Brunswick, NJ.
Wilson, J.O., 1989. Bureaucracy. Basic Books, New York.
Wilson, W., 1887. The study of administration. Political Science Quarterly 2, 197–222.

Civil Society and the Public Sphere: History of the Concept

Craig Calhoun, New York University, NY, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 1897–1903, © 2001, Elsevier Ltd.

Abstract

The concept of the public sphere refers to the capacity of the members of civil society to coordinate their common affairs through collective discourse which transcends the private interests of each. The concept is associated especially with the theories of Hannah Arendt and Jurgen Habermas. Current work focuses especially on the relationship between rational-critical communication and other dimensions of culture-formation and expressive communication; on issues of diversity in participation, topics, and communicative styles; and on the relationship of the public sphere to state-centered politics and/or the transcendence of the state in an international public sphere.

The closely related concepts of civil society and public sphere developed in the early modern era to refer to capacities for social self-organization and influence over the state. Civil society usually refers to the institutions and relationships that organize social life at a level between the state and the family. Public sphere is one of several linked terms (including 'public space,' simply 'public,' and the German *Öffentlichkeit*, or publicness) that denote an institutional setting distinguished by openness of communication and a focus on the public good rather than simply compromises among private goods. Located in civil society, communication in the public sphere may address the state or may seek to influence civil society and even private life directly. Key questions concern the extent to which it will be guided by critical reason, and how boundaries between public and private are mediated.

Civil Society and Self-Organization

The distinction of 'civil society' from the state took its modern form in the seventeenth and eighteenth centuries. Prior to this separation, political and social realms were seldom clearly distinguished. When they were, the social was exemplified by the family and often subordinated as the realm of necessity or mere reproduction to the broader public character and possibilities for active creation that lay in the state. The Greek conception of the *polis*, for example, usually referred to both, but when a distinction was made, it clearly favored the state.

Roman law contributed the idea of *civitas* and a stronger sense of relations among persons that were neither narrowly familial nor specifically about constituting the political society through the state. Medieval political and legal theory developed this theme, especially in relation to the freedoms claimed by medieval cities but also in relation to the Church. Some strands juxtaposed the notion of legitimacy ascending from 'the people' to the eventually dominant idea of divine right of kings, with its notion of legitimacy descending from God. Also influential was the distinction of civil from criminal law (in which the former governs relations formed voluntarily among individuals and the latter the claims of the whole society against malefactors). Nonetheless, it was only in the course of early modern reflection on the sources of social order that civil society came to be seen as a distinct sphere.

A crucial step in this process was the 'affirmation of ordinary life' (Taylor, 1989). Whereas the Greek philosophers had treated the private realm – including economic activity – as clearly inferior to the public realm associated with affairs of state, many moderns placed a new positive value on family and economic pursuits. They argued that both privacy and civil society needed to be defended against encroachments by the state. In this context, it was also possible to conceive of a public sphere that was not coterminous with the state but rather located in civil society and based on its voluntary relations. In this communicative space citizens could address each other openly, and in ways that both established common notions of the public good and influenced the state.

Social contract and natural law theories – especially as joined in the work of John Locke – contributed to this shift by suggesting ways in which the creation of society conceptually preceded the creation of government. From this it was only a short step to say that the legitimacy of government depended on its serving the needs of civil society (or of 'the people'). Thomas Paine and other advocates of freedom from unjust rule advanced an image of the freedoms of Englishmen which was influential not only in England and America, but in France, notably in Montesqueiu's account of the 'spirit' of laws which combined an appreciation of English division of powers with an older tradition of republican (and aristocratic) virtue. From Rousseau through Tocqueville, Comte, and Durkheim, this French tradition developed an ever-stronger account of the autonomy of the social (resisting not only the claims of the state but the Cartesian postulate of the primacy of the individual subject).

A crucial innovation was to understand society as at least potentially self-organizing rather than organized only by rulers. If there was a single pivotal intellectual source for this, it lay with the Scottish moralists. In Adam Smith's (1776) notion of the invisible hand, the market exemplified this self-organizing capacity but did not exhaust it. In his *Essay on the History of Civil Society* (1767), Adam Ferguson presented human history as a series of social transformations leading to modern society.

This prompted Hegel (1821) to treat civil society as a field in which the universal and particular contended; their reconciliation depended on the state. The idea of civil society also shaped classical political economy and ideas of social evolution, and informed Marx's account of the stages of historical development as combinations of productive capacity and (conflict-ridden) social relations. Marx also challenged the notion that markets were neutrally self-organizing, emphasizing the role of historical accumulations of power.

Though the actual analyses differed, what had been established was the notion of society as a distinct object of analysis, not reducible to either state or individual. People formed society impersonally as actors in markets, more personally as parties to contracts. The idea of civil society hearkened back to the sort of social life that emerged among the free citizens of medieval cities because this was largely self-regulated – as distinct from direct rule by ecclesiastical or military authorities. It also suggested 'civility' in interpersonal relations. This meant not just good manners, but a normative order facilitating amicable or at least reliable and nonthreatening relationships among strangers and in general all those who were not bound together by deep private relations like kinship. Equally important, the idea of civil society included – in some versions – the notion that communication among members might be the basis for self-conscious decisions about how to pursue the common good. This notion is basic to the modern idea of public sphere.

The Idea of a Public Sphere

Rousseau (1762) famously sought to understand how social unity could result from free will rather than external constraint. This depended, he argued, on transcending the particular wills of many people with a general will that was universal. Kant admired Rousseau's pursuit of unity in freedom as distinct from mere social instinct (as in Aristotle's notion of a political animal) or imposition of divine authority. He relied implicitly on the idea of a collective conversation through which individual citizens reach common understandings. Likewise the development of representative institutions in eighteenth century England informed and anchored a public discourse directed at bringing the will and wisdom of citizens to bear on affairs of state. Finally, the idea of the people as acting subject came to the fore in the American and French revolutions. The idea of a public sphere anchored democratic and republican thought in the capacity of citizens in civil society to achieve unity and freedom through their discourse with each other.

Kant, like many eighteenth-century philosophers, lacked a strong notion of the social. This Hegel (1821) supplied, rejecting social contract theory because even in Rousseau's notion of a general will it suggested that the union achieved in the state depended not on its own absolute universality but on a development out of individual wills. Nationalism also shaped ideas of society and political community in holistic ways well matched to unitary states (Calhoun, 1999). Marx's (1843/1975, 1927/1975) critique of politics based on bourgeois individual rights further challenged the adequacy of civil society as a realm of freedom and unity. Where Hegel thought that the state in itself might overcome the tension between necessity and freedom and the clash of particular wills, Marx held that only a transformation of material conditions including the abolition of private property could make this possible. As a result, theories stressing stronger ideas of the social were apt to offer weaker notions of public life. The Marxist tradition denigrated 'mere democracy' as an inadequate means of achieving either freedom or unity.

The ideas of public sphere and civil society developed primarily in liberal theory. These were not always seen in the manner of Hegel as merely 'educative' on the way to a more perfect latter unity. Nor was political unity necessarily left to the workings of an invisible hand or other unchosen system, but freedom was treated commonly as a matter of individual rather than collective action. This accompanied the rise of relatively asocial understandings of the market (Polanyi, 1944). In addition, the emerging notion of the public sphere was not clearly distinct from other usages of 'public.' State activity, for example, was sometimes described as public without regard to its relationship to democracy or its openness to the gaze or participation of citizens. This usage survives in reference to state-owned firms as 'public' regardless of the kind of state or the specifics of their operation.

More important was the overlapping concept of 'public opinion' (*see* Public Opinion: Social Attitudes). The dominant eighteenth-century usage emphasized open expression and debate, contrasting free public opinion to absolutist repression. At the same time, it generally treated public opinion as a consensus formed on the basis of reasoned judgment. 'Opinion' was something less than knowledge, but especially where it had been tested in public discourse, it was not simply sentiment and it gained truth-value from reflexive examination. Various euphemisms like 'informed opinion' and 'responsible opinion,' however, reflected both a bias in favor of the opinions of elites and an anxiety about the possibly disruptive opinions of the masses. During the nineteenth century, this anxiety came increasingly to the fore. Tocqueville (1840/1844/1961) and Mill (1859), thus, both contrasted public opinion to reasoned knowledge; Mill especially worried about 'collective mediocrity' in which the opinion of debased masses would triumph over scientific reason. While advocates of the public sphere saw rational-critical discourse producing unity, critics saw mass opinion reflecting psychosocial pressures for conformity. Implicitly, they associated reason with individuals rather than any collective process. The distinction between 'public' and 'crowd' or 'mass' was lost in such views (Splichal, 2000). Early positivist research into public opinion approached it as explicable on the basis of social psychology rather than as a species of reasoned argument. Toennies (1922/2000) sought a way to discern when each approach ought to apply.

Conversely, in the late nineteenth and twentieth centuries, a new field of public opinion research developed that approached public opinion as an aggregation of individual opinions. The shift was based largely on the development of empirical polling methods. It brought a renewal of attention to differences within public opinion, and thus to the distinction between public and crowd (Blumer, 1948; Key, 1961). It also focused attention on patterns of communication among members of the public rather than the more generalized notions of imitation or emotional contagion. New media – first

newspapers, and then broadcast – figured prominently in efforts to understand public communication. While Lippman (1960) and a variety of social psychologists worried that the new media would produce the descent to a lowest common denominator of public opinion that liberals had long feared, Dewey (1927) and other pragmatists defended the capacity for reason in large-scale communication. In this, they hearkened back to the eighteenth-century hopes of Kant and Rousseau.

Even before the apotheosis of the opinion poll, Cooley (1909) had argued emphatically that public opinion ought to be conceived as 'no mere aggregate of individual opinions, but a genuine social product, a result of communication and reciprocal influence.' A key question was whether this communication and reciprocal influence amounted to the exercise of reason. Peirce (1878/1992) had argued that among scientists the formation of consensus on the basis of openness and debate was the best guarantee of truth. Could this view be extended into less specialized domains of public discourse? This has been an enduring focus for Jurgen Habermas, the most influential theorist of the public sphere.

Habermas

In the context of some cynicism about democratic institutions, Habermas (1962/1991) set out to show the unrealized potential of the public sphere as a category of bourgeois society. He challenged most directly the tendencies in Marxism and critical theory to belittle democratic institutions – and also the collapsing of public into state characteristic not only of Hegel but of actually existing socialism. Habermas celebrated the emancipatory potential of a collective discourse about the nature of the public good and the directions of state action. This could be free insofar as it was rational – based on the success of argument and critique rather than the force of either status or coercion – and could achieve unity by disregarding particular interests – like particular statuses – in favor of the general good. The best version of the public sphere was based on 'a kind of social intercourse that, far from presupposing the equality of status, disregarded status altogether.' It worked by a 'mutual willingness to accept the given roles and simultaneously to suspend their reality' (Habermas, 1962/1991, p. 131).

The basic question guiding Habermas' exploration of the public sphere was: to what extent can the wills or opinions guiding political action be formed on the basis of rational-critical discourse? This is a salient issue primarily where economic and other differences give actors discordant identities and conflicting interests. For the most part, Habermas took it as given that the crucial differences among actors were those of class and largely political-economic status; in any case, he treated them as rooted in private life and brought from there to the public. He focused on how the nature, organization, and opportunities for discourse on politically significant topics might be structured so that class and status inequalities were not an insuperable barrier to political participation. The first issue, of course, was access to the discourse. This was not so simple as the mere willingness to listen to another's speech, but also involved matters like the distribution of the sorts of education that empowered speakers to present recognizably 'good' arguments. Beyond this, there was the importance of an ideological commitment to setting aside status differences in the temporary egalitarianism of an intellectual argument.

The public sphere joined civil society to the state by focusing on a notion of public good as distinct from private interest. It was however clearly rooted in civil society and indeed in the distinctive kind of privacy it allowed and valued.

The bourgeois public sphere may be conceived above all as the sphere of private people coming together as a public; they soon claimed the public sphere regulated from above against the public authorities themselves, to engage them in a debate over the general rules governing relations in the basically privatized but publicly relevant sphere of commodity exchange and social labor. The medium of this political confrontation was peculiar and without historical precedent: people's public use of their reason (Habermas, 1962/1991, p. 27).

This public use of reason depended on civil society. Businesses from newspapers to coffee shops, for example, provided settings for public debate. Social institutions (like private property) empowered individuals to participate independently in the public sphere; forms of private life (notably that of the family) prepared individuals to act as autonomous, rational-critical subjects in the public sphere. But the eighteenth-century public sphere was also distinguished by its normative emphases on openness and rational political discourse. Habermas' concern focused on the way later social change brought these two dimensions into conflict with each other.

The idea of publicness as openness underwrote a progressive expansion of access to the public sphere. Property and other qualifications were eliminated and more and more people participated. The result was a decline in the quality of rational-critical discourse. As Habermas later summed up:

> Kant still counted on the transparency of a surveyable public sphere shaped by literary means and open to arguments and which is sustained by a public composed of a relatively small stratum of educated citizens. He could not foresee the structural transformation of this bourgeois public sphere into a semantically degenerated public sphere dominated by the electronic mass media and pervaded by images and virtual realities.
>
> (Habermas, 1998, p. 176)

While Habermas' account of the continuing value of the category of public sphere evoked by the eighteenth-century ideal set him apart from Horkheimer and Adorno (1944/1972) and their pessimistic turn in critical theory, he largely incorporated their critique of 'mass society' as 'administered society' into his survey of twentieth-century developments and with it many of the fears of nineteenth-century liberals. He held that the public sphere was transformed not only by simple increase of numbers but by the success of various new powers at re-establishing in new form the power to 'manage' public opinion or steer it from above. Public relations agents and public opinion polls replaced rational-critical debate; electronic media allowed openness but not the give and take conversation of the eighteenth-century coffee houses. At the same time, rising corporate power and state penetration of civil society undermined the distinction of public and private, producing a 'refeudalization' of society.

Arendt

Hannah Arendt also focused on the problem of collapsing distinctions between public and private. Arendt emphasized the capacity of action in public to create the world that citizens share in common. The term 'public,' she wrote, 'signifies two closely interrelated but not altogether identical phenomena: It means, first, that everything that appears in public can be seen and heard by everybody and has the widest possible publicity. … Second, the term "public" signifies the world itself, in so far as it is common to all of us and distinguished from our privately owned place in it' (Arendt, 1958, pp. 50, 52). Public action, moreover, is the crucial terrain of the humanly created as distinct from the natural world, of appearance and memory, and of talk and recognition. Such action both requires and helps to constitute public spaces – spaces held in common among people within which they may present themselves in speech and recognize others. Public action is thus a realm of freedom from the necessity – notably of material reproduction – that dominates private life.

Arendt's usual term, 'public space,' leaves the 'shape' of public life more open than the phrase public sphere. Public action can create institutions, as in the founding of the American Republic, but as action it is unpredictable. Its publicness comes from its performance in a space between people, a space of appearances, but it is in the nature of public action to be always forming and reforming that space and arguably the people themselves. This conceptualization offers clear advantages for thinking about the place of plurality in the public sphere. As Arendt wrote of America, 'since the country is too big for all of us to come together and determine our fate, we need a number of public spaces within it' (1972, p. 232).

Arendt saw this plurality threatened not just by mass conformity but by the reduction of public concerns to material matters. A focus on sex as much as on the economy threatens the public–private distinction. It not only intrudes on intimacy and private life but impoverishes public discourse. Arendt (1951) saw this problem as basic to totalitarianism, which could allow citizens neither privacy nor free public discourse. Totalitarianism is distinguished from mere tyranny by the fact that it works directly on private life as well as limiting public life. This is not just a matter of contrasting intentions, but of distinctively modern capacity. Modern sociological conditions offer rulers the possibility to reach deeply into the family in particular and personal life in general, to engineer human life in ways never before imagined.

This potential for collapsing the public and private realms is linked to Arendt's unusually negative view of civil society. 'Society,' she writes, is 'that curious and somewhat hybrid realm which the modern age interjected between the older and more genuine realms of the public or political on one side and the private on the other' (1990, p. 122). Civil society is first and foremost a realm of freedom *from* politics. But public freedom is freedom *in* politics. This calls for action that creates new forms of life, rather than merely attempting to advance interests or accommodate to existing conditions. This distinguishes Arendt's view, and republicanism generally, from much liberal thought: 'Thus it has become almost axiomatic even in political theory to understand by political freedom not a political phenomenon, but on the contrary, the more or less free range of nonpolitical activities which a given body politic will permit and guarantee to those who constitute it' (1990, p. 30).

The founding of the United States was a favorite example of such action for Arendt. The American Founders imagined and created a new kind of society, a new set of institutions. This relied on citizens' public commitments to each other rather than assumptions about human nature or mere external application of law. The Founders 'knew that whatever men might be in their singularity, they could bind themselves into a community which, even though it was composed of "sinners," need not necessarily reflect this "sinful" side of human nature' (1990, p. 174). Arendt's vision of public life as central to a moral community shares much with a republican tradition that deplores the modern decline of the public sphere – generally associated with the rise of particular interests at the expense of concern for the general good, the deterioration of rational public discourse about public affairs, or outright disengagement of citizens from politics (*see* Public Sphere: Nineteenth- and Twentieth-Century History). Republican accounts of the public sphere place a strong emphasis on the moral obligations of the good citizen; recent scholarship has often questioned whether citizens lived up to significantly higher standards in earlier eras (Schudson, 1998).

Differentiation in the Public Sphere and Civil Society

Habermas' account of the public sphere has been enduringly influential (see Calhoun, 1992). Its delayed translation into English in 1989 ironically contributed to an invigorating new reading shaped by both the fall of communism and widespread projects of privatization in the West. Critics within communist societies had revived the notion of civil society (as distinct from simply 'society') in order to speak of the realm outside state control and its relative absence in communist societies. Likewise, transitions away from right-wing dictatorships were often treated in terms of a 'return of civil society' (Perez-Diaz, 1993). In the US, the idea of civil society was linked not only to democracy but to reliance on voluntary organizations and philanthropy (Powell and Clemens, 1998; Putnam, 2000).

What civil society signifies in contemporary political analysis is the organization of social life on the basis of interpersonal relationships, group formation, and systems of exchange linking people beyond the range of intimate family relations and without reliance on direction by the government. As a number of scholars of Africa have noted, it incorporates an unfortunate understanding of family privacy that underestimates the positive and supraprivate social roles that African kin organizations can play (see essays in Harbeson et al., 1994). Even more basically, references to civil society often fail to distinguish adequately between systemic capitalist economic organization and much more voluntary creation of social organization through the formation of civic associations, interest groups, and the like – a distinction Habermas has sought to stress. This has sometimes been a source of confusion in use of the public sphere concept to analyze distinctive institutional developments in diverse political and cultural settings (Calhoun, 1993).

Civil society has been important to defenders of free market economics because it suggests the virtues of an economy in

which participants' choices are regulated by their interests rather than their official statuses. In principle, such an economy is able to effectively produce and circulate goods on the basis of prices rather than government direction. Civil society has been equally important to advocates of democracy because it signifies the capacity of citizens to create amongst themselves the associations necessary to bring new issues to the public agenda, to defend both civil and human rights, and to provide for an effective collective voice in the political process. This involves both a free press and political mobilization on the basis of parties and interest groups (see Cohen and Arato, 1992 for the most detailed review; also Chandhoke, 1995; Seligman, 1992; Alexander, 1998; Keane, 1999). Habermas (1992, p. 367) summarizes the recent usage: 'civil society is composed of those more or less spontaneously emergent associations, organizations, and movements that, attuned to how societal problems resonate in the private life spheres, distill and transmit such reactions in amplified form to the public sphere. The core of civil society comprises a network of associations that institutionalizes problem-solving discourses on questions of general interest inside the framework of organized public spheres.' Habermas' work more generally, however, reveals this to be a minimally theorized as well as optimistic usage. It highlights one aspect of civil society but does not make clear the most basic issue.

While part of the heritage of the idea of civil society has been the effort to organize society through public discourse, an equally influential part has been the claim to privacy, the right to be left alone, the opportunity to enter into social relations free from governance by the state or even the public. The idea of business corporations as autonomous creatures of private contract and private property thus reflects the heritage of civil society arguments as much as the idea of a public sphere in which citizens joined in rational-critical argument to determine the nature of their lives together. Civil society refers to the domains in which social life is self-organizing, that is, in which it is not subject to direction by the state. But this self-organization can be a matter of system function or of conscious collective choice through the public sphere (Calhoun, 2001).

Habermas' account of the public sphere drew a variety of important critical responses. One of the first focused on the extent to which he focused on the bourgeois public sphere and correspondingly neglected nonbourgeois public life and failed to clarify some of the conditions built into the bourgeois ideal. Negt and Kluge (1972/1993) responded with an account of the proletarian public sphere. Clearly, workers have at many points built their own institutions, media, and networks of communication, and entered into contention with bourgeois elites and other groups over the collective good. But if this is a discursive competition – that is, if workers and bourgeois argue over what constitutes the collective good rather than only fighting about it – then this implies an encompassing public sphere, albeit an internally differentiated one.

Nancy Fraser (1992) has influentially emphasized the importance of 'subaltern counterpublics' such as those framed by race, class, or gender. Some publics – even very partial ones – may claim to represent the whole; others oppose dominant discursive patterns and still others are neutral. Not all publics that are distinguished from the putative whole are subaltern. As Michael Warner (2001) has suggested, the deployment of claims on an unmarked public as *the* public sphere is also a strategy, generally a strategy of the powerful. Yet, it is important to keep in mind both that the existence of counterpublics as such presupposes a mutual engagement in some larger public sphere and that the segmentation of a distinct public from the unmarked larger public may be a result of exclusion, not choice. Feminist scholars especially have drawn attention to both the gender biases within family life that disempower women and the historically strong gender division between public and private realms on which male political freedom has generally rested (Elshtain, 1993; Young, 2000).

Conclusion

Theories of civil society focus on the capacity for self-organization of social relations, outside the control of the state and usually beyond the realm of family. The basic question posed by theories of the public sphere is to what extent collective discourse can determine the conditions of this social life. Contemporary research on civil society and the public sphere turns on the breadth of political participation, the extent to which capitalist markets limit other dimensions of self-organization in civil society, the existence of multiple or overlapping public spheres, the impact of new communications media, and the quality of rational-critical discourse and its relationship to culture-forming activities. These issues also inform discussions about international civil society and its public sphere.

The concepts of civil society and public sphere took on their primary modern dimensions in the late eighteenth and early nineteenth centuries in Western Europe and to a lesser extent the United States. They have become important in a variety of other settings, including in conceptualizing social autonomy in relationship to communist and authoritarian states. They inform democratic projects as well as academic research in a variety of settings and are in turn themselves informed by cultural creativity and social action.

See also: Civil Society, Concept and History of; Individuals in Society: History of the Concept; Public Sphere: Nineteenth and Twentieth-Century History; State, History of.

Bibliography

Alexander, J.C., 1998. Real Civil Societies: Dilemmas of Institutionalization. Sage, Thousand Oaks, CA.
Arendt, H., 1951. The Origins of Totalitarianism. Harcourt Brace, New York.
Arendt, H., 1958. The Human Condition. University of Chicago Press, Chicago.
Arendt, H., 1972. Crises of the Republic. Harcourt Brace Jovanovich, New York.
Arendt, H., 1990. On Revolution. Penguin, New York.
Blumer, H., 1948. Public opinion and public opinion polling. American Sociological Review 13, 542–554.
Calhoun, C. (Ed.), 1992. Habermas and the Public Sphere. MIT Press, Cambridge, MA.
Calhoun, C., 1993. Civil society and public sphere. Public Culture 5, 267–280.
Calhoun, C., 1999. Nationalism, political community, and the representation of society: Or, why feeling at home is not a substitute for public space. European Journal of Social Theory 2 (2), 217–231.

Calhoun, C., 2001. Constitutional patriotism and the public sphere: Interests, identity, and solidarity in the integration of Europe. In: De Greiff, P., Cronin, P. (Eds.), Transnational Politics. MIT Press, Cambridge, MA.
Chandhoke, N., 1995. State and Civil Society: Explorations in Political Theory. Sage, New Delhi.
Cohen, J., Arato, A., 1992. The Political Theory of Civil Society. MIT Press, Cambridge, MA.
Cooley, C.H., 1909. Social Organization: A Study of the Larger Mind. Scribner, New York.
Dewey, J., 1927. The Public and its Problems. Ohio State University Press, Columbus, OH.
Elshtain, J.B., 1993. Private Man, Public Woman. Princeton University Press, Princeton, NJ.
Ferguson, A., 1767. Essay on the History of Civil Society. Transaction Publishers, New Brunswick, NJ.
Fraser, N., 1992. Rethinking the public sphere: a contribution to the critique of actually existing democracy. In: Calhoun, C. (Ed.), Habermas and the Public Sphere. MIT Press, Cambridge, MA, pp. 109–142.
Habermas, J., 1962/1991. The Structural Transformation of the Bourgeois Public Sphere: An Inquiry into a Category of Bourgeois Society [trans. Burger T]. MIT Press, Cambridge, MA.
Habermas, J., 1992. Between Facts and Norms. MIT Press, Cambridge, MA.
Habermas, J., 1998. In: Cronin, C., De Grieff, P. (Eds.), The Inclusion of the Other. MIT Press, Cambridge, MA.
Harbeson, J.W., Rothchild, D., Chazan, N. (Eds.), 1994. Civil Society and the State. Lynne Rienner, Boulder, CO.
Hegel, G.W.F., 1821. The Philosophy of Right [trans. Knox T M].
Horkheimer, M., Adorno, T.W., 1944/1972. Dialectic of Enlightenment. Herder and Herder, New York.
Keane, J., 1999. Civil Society. Stanford University Press, Stanford, CA.
Key, V.O., 1961. Public Opinion and American Democracy. Knopf, New York.
Lippman, W., 1960. Public Opinion. Macmillan, New York.
Marx, K., 1843/1975. On the Jewish question. In: Marx–Engels Collected Works, Vol. 3. Lawrence and Wishart, London, pp. 146–174.
Marx, K., 1927/1975. Critique of Hegel's philosophy of law. In: Marx–Engels Collected Work, Vol. 3. Lawrence and Wishart, London, pp. 1–129.
Mill, J.S., 1859. On Liberty. Penguin, London.
Negt, O., Kluge, A., 1972/1993. The Public Sphere and Experience. University of Minnesota Press, Minneapolis, MN.
Peirce, C.S., 1878/1992. The Essential Peirce: Selected Philosophical Writings, 1867–1893. Indiana University Press, Bloomington, IN.
Polanyi, K., 1944. The Great Transformation: the Political and Economics Origins of Our Time. Beacon, Boston.
Powell, W., Clemens, E. (Eds.), 1998. Private Action and the Public Good. Yale University Press, New Haven, CT.
Prez-Daz, V.M., 1993. The Return of Civil Society: The Emergence of Democratic Spain. Harvard University Press, Cambridge, MA.
Putnam, R., 2000. Bowling Alone. Simon and Schuster, New York.
Rasmussen, T., 2000. Social Theory and Communication Technology. Ashgate, London.
Rousseau, J.-J., 1762. The Social Contract. Everyman Paperback Classics, London.
Schudson, M., 1998. The Good Citizen: A History of American Civic Life. Free Press, New York.
Seligman, A., 1992. Civil Society. Free Press, New York.
Sennet, R., 1977. The Fall of Public Man. Knopf, New York.
Smith, A., 1776. On the Wealth of Nations. Penguin, Harmondsworth, UK.
Splichal, S., 2000. Defining public opinion in history. In: Hardt, H., Splichal, S. (Eds.), Ferdinand Toennies on Public Opinion. Rowman and Littlefield, London, pp. 11–48.
Taylor, C., 1989. Sources of the Self. Harvard University Press, Cambridge, MA.
Tocqueville, A., 1840/1844/1961. Democracy in America, Vol. 2. Schocken, New York.
Toennies, F., 1922/2000. Kritik der ffentlichen Meinung; selections translated. In: Hardt, H., Splichal, S. (Eds.), Ferdinand Toennies on Public Opinion. Rowman and Littlefield, London, pp. 117–210.
Warner, M., 2001. Public and Counterpublics. Zone Books, Cambridge, MA.
Young, I.M., 2000. Inclusion and Democracy. Oxford University Press, Oxford, UK.

Civil Society, Concept and History of

H Islamoglu, Bogazici University, Istanbul, Turkey

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 1891–1897, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

Historically, the idea of civil society has taken two very different forms. In the first, civil society is 'political community' (*societas civilis* or *koinonia politike*) encompassing a state undifferentiated from society. Here, civil society is coterminous with the state – that is, power relations ordered through law and institutions with the objective of ensuring social harmony. In the second, civil society is a self-regulating, self-governing body outside and often in opposition to the state, represented both as the nexus of societal associations expected to generate civility, social cohesion, and morality, and as the site of reciprocal economic relations among individuals engaged in market exchange activity. The article, emphasizing social, economic, and state transformations between the sixteenth and the eighteenth centuries, focuses on the differentiation in social and political theory, between state and civil society. Since the nineteenth century, this differentiation has been a leitmotiv of Western political discourse notwithstanding the blurring of the lines between state and society amid social and political struggles out of which emerged the social states (Bismarckian welfare state), corporatist states, between the two world wars, and the Keynesian welfare states.

Historically, the idea of civil society takes two very different forms. In the first, civil society is 'political community' (*societas civilis* or *koinonia politike*) encompassing a state undifferentiated from society (Ellis, 2000). Here, civil society is coterminous with the state, that is, power relations ordered through law and institutions with the objective of ensuring social harmony. In the second, civil society is a self-regulating, self-governing body outside and often in opposition to the state, represented both as the nexus of societal associations expected to generate civility, social cohesion and morality, and as the site of reciprocal economic relations among individuals engaged in market exchange activity.

Koinonia Politike and Its Historical Modulations

In Greek and Roman political thought, the notion of civil society as political community was not limited to the legal category of citizenship. Central to Aristotle's (384–22 BC) conception of political community was the recognition that people lived in different social spheres and their status varied in terms of property, skills, and abilities. The art of politics was the use of laws and institutions to organize activities within these different spheres with the objective of attaining a harmonious or 'just' social environment. Aristotle regarded the citizen as one who shared in the administration of justice and held office to this end (Aristotle, 1965). In this sense, citizenship in the Athenian city-state was as much a moral as a legal category (Ehrenberg, 1999). The household was not excluded from Aristotle's moral scheme. Household subsistence production was 'natural,' while production for commercial exchange and profit was 'unnatural' and subversive of the moral order. The exercise of justice presupposed constraints on commercial activity.

For Cicero (106–43 BC) and the Roman lawyers, civil society (*societas civilis*) was the equivalent of *res publica* (commonwealth), or "an assemblage of people in large numbers associated in an agreement with respect to justice and a partnership for the common good" (Cicero, 1988). Cicero viewed justice as rooted in man's natural 'social spirit,' which was informed by reason and induced individuals to forego a measure of self-interest in the interest of common good (Ehrenberg, 1999). Cicero was writing when Rome had ceased to be a harmonious 'commonwealth' run by its citizens and their subsistence-producing households, but had become an imperial city in which landowners, financiers, and merchants coexisted uneasily with a vast and hungry population of peasants and slaves. Cicero saw the realization of justice as the "even balancing of rights, duties and functions" within the state (Cicero, 1988). *Societas civilis* thus represented groups and individuals united by laws and institutions, which organized their activities and sought to achieve a flexible equilibrium among them.

Until the end of the eighteenth century, the various societies of Western Eurasia generally conformed to the Aristotelian model of a politically constituted moral community. In Western Europe, however, the idea of *koinonia politike* virtually disappeared during the Middle Ages when extreme political fragmentation led to the Church assuming a central role in government and social life.

On the other hand, in the Byzantine Empire, the idea of a politically constituted community persisted and the Church was subordinated to the political and moral authority of the emperor. The Ottoman Empire succeeded to this mode of political discourse, but added ideas of bureaucratic and hierarchical organization derived from the political traditions of the agrarian empires farther east, notably Persia. This synthesis, which predated the Ottomans, was formalized in the ninth and tenth centuries by Islamic philosophers well versed in the works of Plato and Aristotle, as well as by the bureaucratic elites of the Islamic states who were educated in the Persian tradition. It was resisted by some sections of the Islamic establishment on the grounds of the equality of all before God. The ensuing power struggles did not, however, result in an autonomous

'Islamic Church,' but rather the religious establishment was subsumed within the political hierarchy.

In the Ottoman Empire, *koinonia politike* was epitomized by the figure of the just ruler, whose ability to establish 'good order,' or *shariah* in the Islamic sense, predicated upon an absence of social strife, constituted justice. Exercise of justice involved laws and institutions accommodating different interest groups in order to sustain the conditions for household subsistence production and to achieve equitable distribution of resources among power holders. Thus, laws and institutions represented negotiated settlements between the ruler and different interest groups.

In Western Europe, the idea of community created by God and comprising all mankind (i.e., all Christians) was a challenge to the notion of differentiated community constituted through politics to achieve social harmony and justice. St Augustine's *City of God* (*De Civitate Dei*, 413–26), where men were equal before God and united in the universal church, stood in contrast to the Greek and Roman *polis* of differentiated spheres of life. The Church and the temporal rulers responsible for its government derived their right to rule from God and stood outside the community.

However, struggles between the Church and temporal rulers, coinciding with the rise of monarchies, resulted in the division of social activities into spiritual and temporal spheres governed, respectively, by the Church and the kings. The concept of *koinonia politike* found a new foothold in the institution of monarchy. Between the fourteenth and seventeenth centuries, European monarchs, endowed with divine right and with absolute powers, nevertheless engaged in constant negotiations with different interest groups. Political society rested on the exchange of entitlements to different groups and individuals in return for their obedience and/or services they rendered.

Koinonia Politike Transformed

In the sixteenth and the seventeenth centuries, following a period of religiously driven civil wars, the idea of a Christian *koinonia politike* was replaced by the concept of the overriding sovereignty of the secular ruler who subordinated religious claims and restored peace. This was the Hobbesian moment in which the modern state rose to define, limit, and enforce moral alternatives to God's order (Maier, 1987). Competition among European monarchies, warfare, civil war, and peasant uprisings all contributed to centralization of political power. By the end of the seventeenth century, this was reflected in the institution of the sovereign monarch as a site of bureaucratic administration and regulation. Commercial expansion in the seventeenth and eighteenth centuries meant that economic activity became a primary target of state regulation. The administrative, regulatory state was represented in discourses on economic administration that combined cameralism, mercantilism, political arithmetic, and bullionism.

Thus, political or civil society was no longer the politically constituted community characterized by a diffusion of political power. Instead, political society referred to the sphere of absolute sovereignty. Social harmony, which remained the legitimizing objective of rule, was the product of bureaucratic regulations and no longer the outcome of reciprocal exchanges or negotiated settlements between the ruler and the different sectors of society.

Attempts to formalize the sovereign state also had the effect of pointing to the existence of a sphere outside the political domain. In the latter part of the eighteenth century, a new conception of civil society as the site of self-regulation developed, referring to voluntary associations freed from the corporate grip of the Church and urban institutions and to the sphere of economic activity. This concept of a self-regulatory, self-governing society was often in opposition to the regulatory, political domain of the state.

The dichotomous conception of state and civil society, was, however, preceded by the distinction in Natural Law theory between *status civilis* and *status naturalis* (Tribe, 1988). The latter was the sphere of discord that the English philosopher Thomas Hobbes (1588–1674) described as *bellum omnium contra omnes* – barter society in which individuals contracted with and against one another (Hobbes, 1949), and to which his countryman John Locke (1632–1704) assigned the contentious process of the formation of private property. For Hobbes, the state of nature ended when it submitted to the 'civil union' 'called state or civil society'. For the Prussian Cameralists of the same period, organization of economic activity in political society, or its regulation, ensured the security and prosperity of the commonwealth, including providing the population with subsistence and productive employment. For the English mercantilist, James Steuart (1712–90), self-interest had to be restrained and directed by the 'reason of state' if it were not to damage public good, which he defined in terms of England's increasing commercial success.

Concentration of political power in administrative monarchies had resulted in a decline in the importance of the corporate bodies and deliberative institutions that had earlier served to mediate power. Groups that were excluded from the political process, as well as new bureaucratic elites and commercial classes, sought inclusion through societal associations and political institutions that could provide them with a voice against absolutism. In France, Britain, and Germany, respectively, Montesquieu (1689–1755), Ferguson (1723–1816), and Kant (1724–1804) proposed conceptions of civil society focused on delineation of associational spaces, of environments for societal negotiations, of civility, and of publicity (Jacob, 1991).

Responding to the despotism of the eighteenth-century French ancient regime, Montesquieu saw in civil society (*l'état civil*) a context for the societal negotiation of the absolute power of the monarch that was not a domain separate from the monarchy. His concept of division of powers addressed a situation in which the monarchy had severed itself from the network of power relations within the political society and had appropriated for itself a separate sphere of power (*l'état politique*). It sought to reintroduce the monarch into the *l'état civil* by means of institutions that would check the absolute authority of the ruler and balance it against the authority of the landed aristocracy, their advocates in the judiciary, and commercial interests (Richter, 1998).

Kant first employed the notion of civil society, or *bürgerliche Gesellschaft*, in the sense of political society inseparable from the Prussian absolutism of Frederick the Great (1712–86),

which he considered indispensable for social stability. Second, *bürgerliche Gesellschaft* referred to the public sphere or a domain of literate citizenry that was separate from the arena of political power and action. Kant regarded the political arena as the reserve of the state, or the ruler. *Bürgerliche Gesellschaft* referred to a sphere 'beyond the political order,' 'beyond the particularistic concerns of political action' where practical issues of governance could be debated on the basis of universal principles of reason. For Kant, the critical practice of exposing actual state policies to the light of universal reason could act to restrain the absolute power of the ruler as well as legitimizing his power (Ellis, 2000). Kant's *bürgerliche Gesellschaft* was composed of individuals from the Prussian bureaucratic and bourgeois elites, educated and trained in state schools and administrative offices, as well as members of social clubs and associations, who could by dint of reason rise beyond the trappings of class or official status.

The thinkers of the Scottish Enlightenment who addressed the issue of civil society did not limit themselves to the issue of restraints to centralized state authority. Adam Ferguson pointed to the corrosion of civic spirit in political society, where the successful commercial classes became servile to the administrative state, which provided them with a 'rule of law' but deprived them of their traditional rights (Keane, 1988; Ferguson, 1995). He conceived of civil society as networks of self-governing and self-regulating voluntary associations, such as self-help groups and 'friendly' or charitable societies, which had expanded rapidly in the eighteenth century and, in Britain, played an important part in poverty relief efforts. Ferguson pointed to the potential of voluntary associations for engendering civility beyond the special interests of state administration and the commercial classes. The central question facing Ferguson and other eighteenth-century European thinkers was how society increasingly differentiated administratively and economically, could remain integrated and harmonious. For Ferguson, as for his compatriot, David Hume (1711–76), civility was the basis for social cohesion, and he saw it as rooted in sociability or moral and emotional communication among persons "that fostered social bonds and friendships and cultivated manners and moral tastes" (Trentmann, 2000). Ferguson did not place civil society in opposition to the state; rather, civil society constituted a protective shield from the uncertainties of social and political life.

Civil Society as a Separate Domain

For Adam Smith (1723–90), also a thinker of the Scottish Enlightenment, civilized society consisted in self-regulating and interdependent networks of economic relations among individuals and groups, originating in the decisions of individuals competing in markets for goods, labor, and capital. Economic activities of self-interested individuals were guided by the universal, natural laws of supply and demand. For Smith, the civilizing of economic society and the harmonizing of individual interests presupposed a mode of sociability and reciprocal sympathies among individuals (Trentmann, 2000).

Smith proposed a separation between the civilized society of economic activity and the political sphere of the state and insisted on the liberation of labor, capital (including land), and goods from the network of relations of political society. The doctrine of *laissez-faire, laissez passez*, first introduced by the eighteenth-century French Physiocrats and taken up by Smith, was a plea for the removal of regulations that privileged the landed classes through restrictions on the grain trade, merchants through grants of monopolies, and the 'poor' through grain subsidies provided by the Poor Laws. Smith was an advocate of the new industrial order and opposed to regulations that represented interventions by the state in the economic process, including the free exchange of labor, capital (including land), and goods.

Smith argued that if individuals, unimpeded by privilege and state regulation, could act in accordance with their self-interest, markets could be trusted to allocate resources equitably in the form of wages, rents, and profits. This, for Smith, was the key to progress, economic growth, and national prosperity (Smith, 1974). On the other hand, he was not entirely indifferent to the dislocations generated by markets and he assigned the civil state the duties of administering justice and providing security, including the protection of private property and correction of severe inequalities created by market forces.

Smith's civilized society, like Ferguson's civil society, Kant's *bürgerliche Gesellschaft*, and Montesquieu's *l'état civil*, represented positive images of civil society, which, in the course of encounters with hitherto unknown regions of Asia and the Americas during the seventeenth and eighteenth centuries, became part of Europe's definition of itself as the domain of the 'civilized' (Kocka, 2000). The discourse of civilized Europe created its opposite in images of an uncivilized non-Europe, notably the East, which was perceived as the domain of despotism *pace* Montesquieu, of chafing regulations over economic activity *pace* Smith, and of the absence of private property *pace* Marx. Following the last decade of the eighteenth century, with the onset of revolutionary upheavals against absolutist regimes, the terms *bürgerliche Gesellschaft* and civil society were increasingly equated with bourgeois society and acquired a negative connotation as 'the reign of dissoluteness, misery, and physical and ethical corruption' (Hegel, 1967; Kocka, 2000). The European idea of itself as 'civilized' was effectively distanced from understandings of civil society as bourgeois society. No matter how 'barbaric' Europe became in Europe, it would remain 'civilized' vis-à-vis non-Europe, its optimistic image of itself having been fixed in a vocabulary of domination.

Hegel (1770–1831) responded to this situation with the idea of the universal state that was to end the separation between the political and the economic. He subscribed to the political economists' notion of civil society as a nexus of economic interests and relationships that continuously reproduce themselves. He did not, however, share Smith's utopian faith in the mechanistic operations of the market. Revolutionary upheavals had revealed a discrepancy between models of an orderly economy subject to universal laws of nature and the chaotic reality of contemporary society. Conceptualizations of civil society as an autonomous domain capable of generating order and progress through its own dynamics now appeared dubious.

Hegel's civil society was not the civilized society of Smith, but bourgeois society that manifested extremes of wealth and poverty that threatened to destroy the productivity of

individuals pursuing their self-interest (Reidel, 1984). The destructive potential of civil society, he felt, could be restrained through redirecting individual self-interest by administrative means, including justice, police, and moral measures.

Marx (1818–83) subscribed to Hegel's critique of political economy and his conception of civil society as an arena of self-interest and divisiveness with a potential for self-destruction (Bottomore, 1983). But unlike Hegel, Marx was not concerned with the containment of bourgeois society's destructive potential. Instead, he focused on the revolutionary transformation of economic relations in civil society, which he believed was possible through the mobilization of conflicts between different interest groups. For Marx, the state, distinct from civil society or *bürgerliche Gesellschaft*, is largely limited to formalistic and negative activities since in his view, civil society both preceded and determined the state. Civil society required no regulation but was ruled through the contingencies of class struggle.

Civil Society and the Liberal State

Political and academic debates of the nineteenth century addressed the question of how to achieve social stability in market societies where the logic of economic activity overrode moral and political concerns regarding equity. Civil, or liberal, society was perceived as a domain shaped by or reformed through the practices of a central state representing the public interest. This constitutes a major departure from eighteenth century thinking. On the one hand, civil society, its actors, its activities, and the economic relations that characterized it, were viewed as inseparable from their legal and administrative formulations. This resulted in the creation of legal entities including trade unions, corporations, and family, and of voluntary and charitable associations, which, while autonomous from the state, remained within the bounds of its administrative–legal vision (Neocleous, 1996). On the other hand, the state, 'objectified' in its administrative and legal practices, was understood to stand apart from civil society and to mediate divergent interests and needs.

From the perspective of the English utilitarians, most notably Jeremy Bentham (1748–1832), and of the German social economist Lorenz von Stein (1815–90), civil society, or the market economy of the political economists, needed to be actively constituted. It was not enough simply to remove the obstacles of obsolete privilege and restrictive policies of mercantilism of the *ancien régime*. For Bentham, the new economic order required positive state intervention, and government was inseparable from an "art of directing the national industry to purposes to which it may be directed with greatest advantage." Von Stein, agreeing with Hegel's perception of civil society as a site of conflict and oppression, identified the 'social problem' as the main obstacle to economic progress. Progress required a market economy, or civil society; however, the injustices and inequalities it generated must be ameliorated through the state's administrative activity. The creation of social citizenship and the active participation of the citizen in the state's decision making were central to this process (Pasquino, 1981). Thus, civil society represented an outcome of collective struggles and clashes among divergent interests that were mediated through the state's administrative practices, so that for von Stein, as for Bentham, the state was not the *Rechtsstaat* ('rule of law' state) that stood outside civil society, but the *Sozialstaat* (social state) representing a process whereby society was continuously formed and reformed.

Not all observers of nineteenth-century Europe saw the centralized administrative state as the liberator of civil society. Conservative Romantics rejected the notion of the state or politics shaping society and assigned self-policing power to society and church. Alexis de Tocqueville (1805–59) in his *De la democratie en Amerique* (1835–40) saw the real danger to modern society in the new despotism of the all-pervasive state administrations rather than in class conflicts. Anticipating his fellow Frenchman, Michel Foucault (1926–84), de Tocqueville pointed to the administrative suffocation of civil society as evidenced in the state's monopoly of public education, health care, and social services to the poor and unemployed that subjected all aspects of citizens' lives to state scrutiny (Keane, 1988). De Tocqueville stressed the importance of voluntary associations in placing checks on administrative despotism by providing the services that people expected from government, thus preempting government intervention in civil society. Recalling the ideas of the Scottish Enlightenment, de Tocqueville saw voluntary associations as generating and expressing community civic values.

Civil society all but disappeared from political and academic debates in the late nineteenth century and the period following World War I, a time of slackening economic growth, rising working-class activism, imperial rivalries, and wars, which witnessed the displacement of political power from state administrations to major organized groups in society. The European welfare states, heirs to the liberal states, increasingly became sites for centralized and bureaucratic bargaining among political parties, labor unions, and business cartels vying for economic and political power and blurring the distinction between state and civil society, public and private (Maier, 1987).

Antonio Gramsci (1891–1937), leader of the Italian Communist Party, represented a notable exception to the lack of interest in the notion of civil society in the post-World War I era. He reformulated the Marxist–Hegelian understanding of civil society in terms of the corporatist Zeitgeist of Italy in the 1920s. For Gramsci, civil society was not merely the sphere of individual needs but also of organizations where the hegemony of the ruling class and consent to that rule was negotiated. In this sense, civil society comprised not only all material (economic) but also political and cultural relations. While Marx insisted on separation between state and society, Gramsci held the two were interrelated. Hegemony, which was basic to Gramsci's notion of civil society, presupposed 'interpretations' of the economic structure, that is, the political and cultural mediation of different interests (Bobbio, 1988). In that sense, Gramsci had a kindred spirit in Hegel rather than Marx.

Late-Twentieth Century Understandings of Civil Society

After World War II, the European political and economic order focused politics on the state with welfare states

subsuming civil society to state regulation. As in the late nineteenth century and the post-World War I era, economic growth, distribution, and security represented the core themes of politics. However, now these themes were generalized to regions outside Western Europe, including Eastern Europe and postcolonial societies in the so-called Third World. The different regions varied greatly in terms of how growth was attained and sustained, and the efficiency of realizing their economic and social goals.

They also varied in terms of their political institutions and modes of resolution of social and political conflict. From the late 1940s to the mid-1970s sustained economic growth in the capitalist economies of Western Europe and North America made challenges to economic redistribution less urgent, and the internationalization of the security problem (e.g., the formation of North Atlantic Treaty Organization) led to a consensus regarding security. Capital transfers from a booming United States economy, including military and economic aid, helped to sustain the developing states of the Third World, while the statist socialist regimes relied on the ideological and political disciplining of their populations to generate surpluses centrally distributed by the state.

During the 1950s and 1960s, some social scientists in Western Europe and the United States predicted that under conditions of long-term growth, political problems could be largely transformed into noncontroversial administrative routines resolvable by experts. Socialist regimes perfected this functionalist perception of politics where government excluded all competing forms of political organization, including political parties, trade unions, and collective bargaining. However, this perspective precluded discussion of civil society. Socialist ideologues identified civil society with bourgeois society and held it was superseded by the advent of socialism.

In the mid-1970s, slowing economic growth in the advanced capitalist societies resulted in constraints on the ability of the state to tax business and distribute assistance to the working class in the form of social welfare benefits. The fiscal crisis was further aggravated by rising costs of welfare provisions, and called into question the validity of ideas that subsumed society to state regulation. Since the 1970s, social and political debates have largely focused on the critique of social statism and the differentiation of civil society from the state.

From a neoconservative stance, overpoliticization of the state and the fusion of political and nonpolitical spheres of social life has eroded political authority, weakened governability, and destroyed the autonomy and authority of nonpolitical spheres, including family, religion, and the market. Neoconservatives proposed the boundaries between state and civil society be redrawn. Civil society, in opposition to the state, represented the depoliticized sphere of market activity and all else, including religion and family. It should be reintegrated through a cultural model that cohered the exchange activities of self-interested individuals. This is a model of the self-regulating civil society of nonprofit voluntary associations reminiscent of the eighteenth-century Scottish Enlightenment. During the 1980s and 1990s, social science paradigms based on individual self-interest and rational expectations presented models for social programming that would render a politicized civil society unnecessary. In the neoconservative view, the state is a lean structure characterized by effective authoritarian forms of action, approximated by its Thatcherite interpretation in 1980s Britain.

For proponents of the voluntary social movement paradigm, statism had disastrous environmental, political, and social consequences. They proposed an understanding of civil society as an autonomous politicized sphere independent of regulation and constraint by bureaucratic political institutions. Social movements such as the environmentalist, antinuclear, and women's and gay movements of the 1970s, were politically motivated by concerns about the environment and quality of life, and issues of equity, authenticity, and participation. Their agendas required 'spontaneous' organization and not 'officially' sanctioned institutions of political parties and trade unions. Here, civil society referred to an intermediate institutional space between private (personal) and public (the object of official political institutions and actors).

Jurgen Habermas' (1929) idea of public sphere parallels this understanding of civil society in the context of ideas and communications; it includes the organization of civic or bourgeois opinion as represented by associations and media. Public sphere is a space where communication about collective values takes place. In advanced capitalist societies where political and economic domains merge, the state became a major actor in the market economy and thus no longer able to advance the common good. This, for Habermas, signals the legitimation crisis of advanced capitalist societies and raises the issue of the formulation of a public discourse outside of the spheres of market economy and the welfare state and in doing so (in the spirit of Kant), subjects both spheres to the critical scrutiny of communicative rationality.

In the mid-1970s, the crises in socialist states, initially Poland, provoked Central and Eastern European intellectuals to address the issue of the politicization of civil society. After failed revolutions in Hungary (1956) and Czechoslovakia (1968), the new evolutionism of Adam Michnik and Jacek Kuron proposed the bottom-up construction of a civil society to repulse state intrusions into social life. For other dissident socialist intellectuals in the region, most notably Vaclav Havel, civil society represented the domain of antipolitics; it was a vision of society not simply independent of the state but opposed to it.

During the 1980s and 1990s in many parts of the world, including the former socialist states of Central and Eastern Europe, the expansion of the market system, including privatization programs, led to two divergent perceptions of civil society. The first represented civil society in opposition to global capitalism and to the state as the implementor of market reforms. The second emphasizes the centrality in civil society of religion, family, and voluntary associations for generating moral, economic, and cognitive norms.

See also: Attitudes, Political and Civic Culture; Civil Society and the Public Sphere: History of the Concept; Democracy: Normative Theory; Democratic Theory; Hegel, Georg Wilhelm Friedrich (1770–1831); Kant, Immanuel (1724–1804); Liberalism: Historical Aspects; Politics and Participation; Smith, Adam (1723–90).

Bibliography

Aristotle, E., 1965. In: Barker, N. (Ed.), The Politics. Oxford University Press, New York.

Bobbio, N., 1988. Gramsci and the concept of civil society. In: Keane, J. (Ed.), Civil Society and the State: New European Perspectives. Verso, London, pp. 73–99.

Bottomore, T. (Ed.), 1983. Civil Society. A Dictionary of Marxist Thought. Blackwell, Oxford, UK.

Calhoun, C. (Ed.), 1992. Habermas and the Public Sphere. MIT Press, Cambridge, MA.

Chambers, S., Kymlicka, W., 2002. Alternative Conceptions of Civil Society. Princeton University Press, Princeton, NJ.

Cicero, 1988. The Republic. Cambridge University Press, Cambridge, UK.

Cohen, J.L., Arato, A., 1992. Civil Society and Political Theory. MIT Press, Cambridge, MA.

Deng, Z., 2010. State and Civil Society: The Chinese Perspective. World Scientific, Singapore.

Donzelot, J., 1984. L'Invention du social. Fayard, Paris.

Dörner, W., List, R., 2012. Civil Society, Conflict and Violence: Insights from the CIVICUS Civil Society Index Project. Bloomsbury Academic, London.

Edwards, M., 2004. Civil Society. Polity Press, Malden, MA.

Ehrenberg, John, 1999. Civil Society: The Critical History of an Idea. New York University Press, New York.

Ellis, E., 2000. Immanuel Kant's two theories of civil society. In: Trentmann, F. (Ed.), Paradoxes of Civil Society: New Perspectives on Modern German and British History. Bergham, New York, pp. 105–131.

Ferguson, A., 1995. An Essay on the History of Civil Society. Transaction, New Brunswick, NJ.

Gierke, O., 1958. Political Theories of the Middle Age. Beacon, Boston.

Havel, V., 1988. Anti-political politics. In: Keane, J. (Ed.), Civil Society and the State: New European Perspectives. Verso, London, pp. 381–398.

Hegel, G.W., 1967. The Philosophy of Right London (T.M. Knox, English Trans.). Oxford University Press, London.

Hobbes, T., 1949. De Cive or the Citizen (S.P. Lamprecht, English Trans.). Greenwood, New York.

Hoffmann, S.-L., 2006. Civil Society, 1750–1914. Palgrave Macmillan, Basingstoke, England.

Howell, J., Lind, J., 2010. Civil Society under Strain: Counter-terrorism Policy, Civil Society, and Aid Post-9/11. Kumarian Press, Sterling, VA.

Jacob, C.M., 1991. The enlightenment redefined: the formation of modern civil society. Social Research 58 (2), 475–495.

Juergensmeyer, M., 2005. Religion in Global Civil Society. Oxford University Press, Oxford.

Kant, I., 1963. What is enlightenment? In: Beck, L.W. (Ed.), On History. The Bobbs-Merrill Company, Indianapolis, IN.

Keane, J., 2003. Global Civil Society? Cambridge University Press, Cambridge.

Keane, J. (Ed.), 1988. Civil Society and the State: New European Perspectives. Verso, London.

Keane, J.H., 2006. Civil Society: Berlin Perspectives. Berghahn Books, New York.

Kocka, J., 2000. Zivilgesellschaft als historisches problem und Versprechen. In: Hildermeier, M., Kocka, J., Conrad, C. (Eds.), Europaische Zivilgesellschaft in Ost und West: Begriff, Geschichte, Chancen. Campus Verlag, Frankfurt/Main, Germany, pp. 13–39.

Maier, C. (Ed.), 1987. Changing Boundaries of the Political: Essays on the Evolving Balance between State and Society, Public and Private in Europe. Cambridge University Press, Cambridge, UK.

Maley, W., Sampford, C.J.G., Thakur, R.C., 2003. From Civil Strife to Civil Society: Civil and Military Responsibilities in Disrupted States. United Nations University Press, Tokyo.

Neocleous, M., 1996. Administering Civil Society: Towards a Theory of State Power. MacMillan, London.

Offe, C., 1987. Challenging the boundaries of institutional politics: social movements since the 1960s. In: Maier, C. (Ed.), Changing Boundaries of the Political: Essays on the Evolving Balance between State and Society, Public and Private in Europe. Cambridge University Press, Cambridge, UK, pp. 63–105.

Pasquino, P., 1981. Introduction to Lorenz von Stein. Economy and Society 10 (1), 1–6.

Reidel, M., 1984. Between Tradition and Evolution: The Hegelian Transformation of Political Philosophy. Cambridge University Press, Cambridge, UK.

Richter, M., 1998. Montesquieu and the concept of civil society. The European Legacy 3 (66), 33–41.

Rosanvallon, P., 1988. The decline of social visibility. In: Keane, J. (Ed.), Civil Society and the State: New European Perspectives. Verso, London, pp. 199–220.

Rosenblum, N.L., Post, R., 2002. Civil Society and Government. Princeton University Press, Princeton, NJ.

Schuler, D., Day, P., DIAC (Conference), 2004. Shaping the Network Society: The New Role of Civil Society in Cyberspace. MIT Press, Cambridge, Mass.

Smith, A., 1974. The Wealth of Nations. Penguin, Harmondsworth, UK.

Trentmann, F. (Ed.), 2000. Paradoxes of Civil Society: New Perspectives on Modern German and British History. Bergham, New York.

Tribe, K., 1988. Governing Economy: The Reformation of German Economic Discourse, 1750–1840. Cambridge University Press, Cambridge, UK.

Civil Wars

Bill Kissane, London School of Economics, Houghton St., London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article summarizes the social and especially political science literature as it pertains to the wave of civil or internal wars which have emerged since 1945. It does so by; (1) asking how well the concept of civil war travels, (2) locating this literature with the general treatment of war and revolution historically, and (3) discussing the nature of causality and the recent reliance on purely structuralist accounts of civil war.

Only recently has the study of civil war become a specific subfield within social science, a development reflecting the fact that as war between states became rarer after 1945, civil wars have proliferated. The dominant historical attitude was disapprobation. Civil wars threatened both the collapse of states and those social relations that made states possible. Between the French Revolution and the end of the Cold War, civil wars were also treated as appendices to other forms of war, such as interstate war. Only the end of the Cold War opened up a vista for civil war studies in their own right. Scholarship became less overtly partisan, the connection with colonial legacies intriguing, and a strong emphasis placed on the consequences in terms of economic underdevelopment, state collapse, population displacement, and human rights violations. A huge literature has also emerged on establishing peace.

The Concept

In an influential definition Small and Singer (1981, p. 201) considered civil wars as armed conflicts which involved military action internal to the metropole, with government participation but effective resistance by both sides. Such wars were different from wars that involved external intervention or resistance to colonial rule by virtue of their internality: the domestic government was involved and fighting was on its sovereign territory. By adding the criterion of a threshold of deaths, (to exceed more than 1000), one could distinguish civil from other forms of internal war.

Such a definition defines civil war in terms of its necessary elements and allows for comparison between a large number of cases. Yet it considers as one conflicts leading to total polarization within a state and insurgencies which do not. Conflicts which occur within an integrated political community, and those where no such community exists at the outset of conflict, are also conflated. The problem is one of 'lumping' types of conflict together, rather than 'splitting' them into different subtypes in order to reveal distinct characteristics.

The issue is how well civil, an adjective given elevated status only in the West's political thinking, actually travels. In Europe, the essence of civil war was presumed to lie in this civil character. The German concept of *Burgerkrieg* reflects the idea of civil war taking place within a society of fixed social orders (Brunner et al., 1972). It suggests that civil refers to the participants. This continued the Roman way of categorizing wars in terms of the enemy being fought: to them we owe the term civil war (*bellum civile*) (Armitage, 2009). Civil has as antonyms military and ecclesiastical, and can pertain to the community, ordinary life, or individual citizens. In law it covers private relations among citizens. Within this sphere civil life can flourish, but conflict between citizens also produces the worst polarization. The Roman conception is still influential. The Chambers English Dictionary defines civil war was simply as a war between citizens of the same state.

Kalyvas (2006, 18) sees two further ways of thinking about the civil in civil war which have wide application. The first, rooted in the Greek historian Thucydides, focused on the split nature of the polity, and reflects different groups controlling different territories, or rival claims to sovereign authority. The second, which emerged in larger entities such as Empires, uses concepts of sedition and rebellion to characterize civil war. Hence the focus on legality. Thus in England a distinction emerged between rebellion, which was an action against an established authority without legal justification, and revolution which meant the restoration or renovation of an earlier lawful authority. Cromwell's actions were first seen as the Great Rebellion, while the events of 1688 were termed the Glorious Revolution. These ideas are not obsolete. In Costa Rica, the idea of civil war as a means of restoring a constitutional *status quo ante* had a powerful influence on the outcome of the 1948 civil war. In Rwanda however, the split nature of the populace is far more important than any constitutional framework. Yet the American civil war emerged in a split polity, under a constitutional framework invoked by both sides as a source of legitimacy. It was a civil war in both senses.

In ancient Greece and Rome, civil conflicts threatened afixed social orders. The French Revolution changed this, and by mobilizing the nation in arms, made the people as a whole the collective agent. Civil war ceased to be defined as a war among citizens of a city, country, or state, but became a war of a part of a nation against another, or a war among hostile parties 'within the same state.' Hence the link with fratricide. What first set it apart from other conflicts was the concept of citizenship (promoted by the French Revolution, but reviving an old Roman idea). Yet modernization soon broke up those fixed orders and established state and civil society as the two

poles around which contemporary politics were organized (Brueilly, 1996, pp. 164–169). Civil war would happen between those poles. For Carl Schmidt (1996), the state's primary role was to make sure that the total enmity of 'friend enemy' distinctions in interstate wars would not be reproduced within the state. His greatest fear was thus of internal war.

Internal war became a concept used by political scientists in the 1960s (Eckstein, 1965). The neutrality of the term suggests that civil war was an internalized form of war in general. The Cold War was the crucial backdrop, since after 1945 all wars were transformed into civil wars (Koselleck, 2004, p. 56). The Stanford Encyclopedia of Philosophy defined war as "an actual, intentional, and widespread armed conflict between political communities": civil war, between rival groups or communities 'within a state,' is its internal variant (Orend, 2008 Eckstein used 'internal war' as a generic term, with civil wars, coups, riots, rebellions, massacres, and revolutions, species of it. This concept could include conflicts not considered 'classic' civil wars, and avoided the problem of legitimizing all internal wars as 'civil wars' (Payne, 2011, p. 5).) The internal war conception appealed in a world where territorially concentrated insurgencies against the state became its typical form. There was a 'spontaneous' type, where violence occurred without much leadership and planning, and provided opportunities for the seizure of power. These occurred mainly in societies with no tradition of stable institutions within which political changes could take place. The more common 'planned' type – where channels for the settling of grievances are blocked – involves an attempt at seizing power violently, involving a guerrilla group and an organized campaign. This assumes some centralized power as a starting point (International Encyclopaedia of the Social Sciences, 1968, p. 499).

Eckstein wanted the internal war conception to facilitate comparative analysis of conflict transformation (how terrorism became insurgency for example). This could not be done by categorizing conflicts into distinct types in advance. Yet what makes them internal in the first place? They were conflicts 'within a society' that usually resulted from an attempt to seize power and symbols of legitimacy by extralegal means. For Renzato (1994v) 'internal war' takes place within a state or a city, among citizens of the same state or city, as opposed to external war that is fought between sovereign states. Brown (1996) sees them as conflicts whose origins are *primarily* due to domestic rather than international factors, and in which violence takes place *primarily* within the boundaries of a single state. Others stress that the government must be a protagonist, with control of the state the key objective for both parties. Hence, the struggle for legitimacy is crucial, unlike in conventional interstate wars, where the status of the opposed states is usually uncontested. The stress on legitimacy is consistent with Clausewitz's conception of war as politics by other means.

The internal war approach was appropriate to a world where decolonization had created more than 120 new states, and when the nuclear standoff during the Cold War had created an international environment in which more internal war took place. Yet these were often proxy wars, where the boundaries between internal and external politics were actually porous. Going back as far as Thucydides, the worst consequences of internal wars often result from the involvement of external powers. In the 1990s, doubts then emerged about the robustness of both state and nation in an era of globalization. The 'New Wars' literature refers to a type of war which developed in Africa and Eastern Europe in the 1980s and 1990s. Their newness consists in blurring the boundaries between war, organized crime, and large-scale human rights violations (Kaldor, 1999, p. 2). Such conflicts are not necessarily fought for political ends (to acquire state power). Related concepts: 'wars of the third kind,' 'complex emergencies,' 'postmodern wars,' and 'never-ending wars' suggest rather a blurring of categories, between combatants and noncombatants, between war and humanitarian crises, and between periods of war and peace.

The approach is especially relevant to interethnic wars which are fought between citizens of the same state, but often take the forms of interstate war; with regular armies, battles, and soldiers in uniforms. The substitution of 'new' for civil reflects 'the erosion of the state's monopoly of violence' (ibid, 40), from above by the growth of international military forces, and from below by transnational crime networks, diasporas, and paramilitary actors. Indeed the paramilitary has become the embodiment of such new wars. Bosnia Herzegovina is the paradigm case. Such wars "are part of a process which is more or less a reversal of the processes through which the modern state evolved" (ibid, 5). Historically, as the modern state developed, it grew through war and had to mobilize domestic resources to finance this effort. As war became the province of the state, and competitors to the state were eliminated, its growing destructiveness was accompanied by growing security at home: "hence the way in which the term 'civil' came to mean internal" (ibid). This distinction between internal and external politics had allowed for a sphere within which civil wars could occur. As it collapsed civil war has ceased to be a major category of warfare.

The new wars approach reflects the tendency for conceptions of civil war to reflect the regime trends of their era. It is appropriate to a world where many states were failing, just as the ancient Greek concept of 'stasis' (faction or the consequences of faction) reflected an acute sensitivity to the morphology of states (de Romilly, 1991). The newness of such wars is debatable, but if neither state nor nation is ontologically robust, the question it raises is whether concepts of conflict assume an integrated political community for them to be categorized as civil or internal. Kalyvas (2006, 17) wants civil war to remain the master concept, which encompass peasant rebellion, ethnic insurgency, revolution, and anticolonial wars. 'Civil' refers primarily to a divided polity, but the people must also be members of a common community at the outset. If so, the relationship between state and nation usually defines what is civil new or internal about domestic conflicts.

The Literature

Civil wars are ubiquitous in human history, but little has been written systematically about them. What exists reflect different intellectual traditions. On the English civil wars of the seventeenth century, historians stress constitutional factors as the prism through which the conflicts should be viewed. In contrast, Thucydides (1910) focused on the psychological

consequences of internal strife, believing constitutional ideals were pretexts for other (venal) ambitions. In his study of the Peloponnesian Wars he sees stasis as resulting in a deterioration of manners and morals, as a war between two city-states became the vehicle for personal ambition and an occasion for private vengeance.

Civil war has always suffered from comparisons: nowhere in nineteenth-century Europe was the almost universal approbation of war applied to civil war. With the emergence of the modern conception of revolution, civil war was also left bereft of a positive connotation. It invoked the religious wars of the sixteenth and seventeenth centuries, with their hatred and blind fanaticism. The idea that revolutionary violence could legitimately be motivated by understandable political grievances, also contrasted with the fanaticisms of such wars. Thus while revolution promised change, after 1789 civil war acquired rather the meaning "of a senseless circling upon itself, with respect to which Revolution sought to open up a new vista" (Koselleck, 2004, p. 49).

The negative judgment also reflects the simple conviction that war within is worse than conflict with an external enemy. Aggression that comes from close by is more harmful and more likely to transgress norms. Yet the contrast also reflects a persistent tendency to romanticize wars fought by states. Such wars, being fought by regular armies, subject to greater centralized control and doctrines of war, are somehow less barbarous. The fact that such contrasts are always made (without testing them) shows the extent to which our understanding of what is political has been shaped by the modern state. Only for internal wars do we ask *Why Men Rebel?* (Gurr, 1970). We never do so for interstate wars, whose political rationality is conferred upon by the participation of the state.

The treatment of civil war in the history of Western political thought has been sporadic. We have no treatise to compare with Burke's reflections on the revolution or Kant's on peace. In the Greek city-states, when systematic thought on the nature of power flourished, the concept of stasis suggested that fragmentation was the worst consequences of civil war. The Romans also tended to see the succession of civil wars which led to the establishment of the Empire in the first century B.C. in terms of a deterioration in public life. Both show that the contribution of civil war to the history of political thought has largely been to make people reveal their vision of the good society when faced with the challenge of reconstruction (MacIntyre, 1971, 132). One example is Hobbes vision of a Leviathan based on obedience and protection. What is ours?

During the Cold War most universities taught courses on revolutions. Special attention was placed on big cases, such as the French Revolution, linked to ideas of social transformation, or those like the Iranian Revolution of 1979, which left their mark on the international relations of the era. Little systematic attention was paid to civil war, despite it often being the 'hard nucleus' of most revolutions in reality. One school of thought attempted 'a natural history of revolutions' which they thought went through regular patterns and cycles, This was very similar to current scholarship on processes of polarization, radical or deradicalization, and mobilization. Later scholars used comparative historical analysis of a small number of cases to identify preconditions for revolution, and this work on crises and state breakdown has been extended to more recent civil wars (Goldstone, 2009).

There was also an urban bias. Everyone has heard of Che Guevara, Malcolm X, and the Baader Meinhof. The lack of attention to rural insurgency was rectified by the 'internal' wars approach which continues in the form of large N research today. Yet the fundamental catalyst for the growth of this field was the end of the Cold War. The collapse of Yugoslavia refocused the attention of security studies onto domestic sources of conflict. Moreover, as many Soviet experts found themselves unemployed, the end of the Cold War revealed that the number of domestic conflicts had actually been steadily accumulating since the 1940s. What was striking about many proxy wars, such as in Southern Africa, was that they lasted for decades. As a consequence, there was a revival of interest factors that made the new states vulnerable to internal wars: the connection between ethnic diversity and violence, the concept of state weakness, and the 'conflict trap' spoken about by development economists were explored.

The literature on revolution focused on big cases. Even those which emerged on the world's periphery (China, Mexico, Turkey, and Russia) set in train processes which have taken more than half a century to work themselves out. In civil war studies attention focused poor and formerly colonized territories in Asia and Africa where civil war was seen as a problem for political development. Whereas the standard curriculum 30 years ago included the Spanish civil war, or Fascism, the imagination of students of conflict is now captured by parts of the world very remote from their experience. This has meant that any serious research requires fieldwork, language acquisition, and that a blurring of boundaries between conflict and development studies has taken place. Many influential studies have been done by economists. Some still argue for a longer-term perspective on legacies of imperialism and slavery (Richards, 2008), but the general interest has been in structural factors that make these new states vulnerable to conflict after independence.

Thus since 1989 the field of civil war studies has acquired a North–South perspective, as that of revolution did during decolonization. The globalization of news media and the number of international organizations involved in crisis management have led to a whole subdiscipline in 'conflict and peace' studies. The terminology of 'complex emergencies,' 'crisis states,' and 'humanitarian disaster' reflects this. A specific literature on peace agreements, post-civil war democracy, transitional justice, and reconstruction has also flourished. Here, the gap between the academic and the policy worlds is becoming narrow. Subdisciplines like that on transitional justice, first concerned with the legal aspects of war to peace transitions, now take nearly every legacy of conflict within their remit. Normative perspectives are strongly in debt to Western liberalism and raise questions uneven power relationships in the academic world. Scholars from poorer countries are reduced to the role of mere suppliers of data in an international division of scholarly labor, which privileges theorists from Western countries whose *recent* historical experience has not been one of internal war.

Methodologically, case studies remain the norm, but the more prestigious political science journals focus almost exclusively on cross-national statistical studies. These trace patterns

of conflict globally, but few good theories on the regional or international sources of civil war exist. There is almost no importance attached to ideology. The general assumption is that the sources of civil war are primarily internal. Yet some landmark books focus not on causal analysis at all, but on the changing nature of warfare, the variation in violence within conflicts, and whether conflict retards development or not (Cramer, 2006; Holsti, 1996; Kalyvas, 2006). Are the actual causes of civil war may too numerous too complex for theoretical treatment in comparative studies?

The Causes

Mainstream causal theories usually try to go beyond the focus on bad leaders where the blame game usually begins. This was true of the American civil war, after which explanations, conspiratorial in character, soon flourished. The collapse of Yugoslavia was also blamed on the behavior of politicians like Slobodan Milosevic, although subsequent studies have tried to specify the conditions under which such demagogues hold sway. An emergent field explores the connection between regime transitions in general, elections, and civil war (Hegre et al., 2001).

Manifestations of conflict can be explained either by the motives of which the quarreling parties are conscious, or by objective situations and processes – in the latter tradition: "we try to explore the unconscious complexes to which these motives relate like mere symptoms of an illness to the organic disturbances themselves" (Buber: 2002, p. 20). The current literature even uses medical language, for example, on the "onset, duration, and termination" of civil war (Hegre, 2004, pp. 243–252). Theories from international relations have been deployed to explain the Yugoslav conflicts as a security dilemma; state weakness has been found to be more important than ethnic differences; and micro-economic explanations have stressed the costs and benefits of participation in civil war. Together these constitute a 'structural turn' in theorizing (Kissane and Sitter, 2012).

The security dilemma provides an explanation of how a relatively stable society, like Yugoslavia, rapidly deteriorated into civil war without assuming that this process reflected prewar divisions (Posen, 1993, pp. 27–47). Anarchy is central. In the absence of central authority, the security requirements of one group are seen as a threat by another group. If the authority of the state collapses, two or more groups may come into conflict, not because they actively pursue this course of action, but because the measures they take to protect themselves constitute a threat to the other groups. Hence the suggestion that mutual fear caused the Bosnian Civil War (Woodward, 1999).

Most of the post-1945 civil wars have been produced in poor countries, by the increasing contradiction between new political and economic aspirations, often fed by the example of other countries, and the relatively oppressive governments with which people are confronted (Luard, 1972, p. 14). However Fearon and Laitin (2003) argue that this contradiction results in civil war mainly when the state is weak. Their statistical results find that low GDP makes civil war more likely, and they explain this in terms of the low capacity of 'financially and bureaucratically weak states.' Mountainous terrain and large populations make civil war more likely because they make state control more difficult. 'Grievance' factors such as ethnic diversity and the low level of democracy, do not affect civil war: they are seen as the product of civil wars, not their cause.

For Fearon and Laitin civil wars in failed states reflect *permanent* "conditions that favour insurgency – in particular state weakness marked by poverty, a large population and instability" (Fearon, 2004). Collier and Hoeffler add a stress on the costs and benefits to the actors in insurgencies. Greed is key, and is linked to the 'resource curse': rebels' quest for economic gain, notably in the shape of 'predation of the rents from primary commodity exports,' is the most important factor. 'Grievance,' defined as dissatisfaction caused by ethnic or religious differences, political repression or unequal economic distribution, is less important. Like Fearon and Laitin they find a strong statistical correlation between GDP and civil war, but their explanation lies in the actors' quest for private gains and the opportunities that civil war provides for this. As their subsequent study put it: 'where rebellion is feasible it will occur: motivation is indeterminate, being supplied by whatever agenda happens to be adopted by the fist social entrepreneur to occupy a viable niche' (Collier et al., 2009).

Instead of focusing on the 'triggering events' which lead to civil war, the latter two studies try to relate the variation in civil war to countries' longer term social, economic, and institutional features. Rather than going into detailed analysis of the various micro processes which cause civil war, such studies hang their analyses on the two tail ends of the causal chain; initial structural conditions on the one hand, and outcomes (violence or nonviolence) on the other. What connects these structural attributes to the onset of civil wars in these theories are (rationalist) stipulations (greed not grievance) regarding human motivations which are employed to substitute for the complex causal processes behind civil wars which this literature is unable or unwilling to specify (Ringmar, 1996, p. 36). In the words of one critic "homo economicus goes to war" (Cramer, 2006, pp. 124–139).

These studies combine a root cause approach to conflict, once prominent in sociology, with statistical methods. Yet root cause explanations require a clear hierarchy between variables, which is hard to establish. In Yugoslavia had the state remained intact, the spiral of violence would have been avoided. Yet the political salience of ethnic identity in the Balkans, was also a root cause. Causal hierarchy could be established through process-tracing and establishing which came first in the sequence of events. The problem is that these two factors were mutually constituted. Table 1 represents this mutual constitution.

What about the grievances the actors themselves articulate as the reason for conflict? Collier, Hoeffler, and Rohner (2009) develop a radical 'feasibility hypothesis,' which holds that where rebellion, driven by greed, is financially and militarily viable, it will occur (as policy entrepreneurs sooner or later step forward and formulate a suitable grievance). Others argue that there can be no clear rational course of action in the absence of interpretative filters. Work on Indonesia, Northern Ireland, and the Philippines suggests the prior existence of myths and memories was a necessary condition for the eruption of conflict over material resources and state power (Aspinall, 2007; Kaufmann, 2006). There is also philosophical difference. The realist premise of the first approach is that 'reality' – in the form of ethnic groups, material resources,

terrain – is something 'out there.' However, actors' perceptions are social facts too. We live in a material world which we interpret, and it is on the basis of these interpretations that we present various options to ourselves (Ringmar, 1996, p. 37). The option to engage in civil war is one such decision.

Ultimately, structural approaches can only investigate underlying factors which make societies prone to violence, but are weak when it comes to identifying catalytic factors (Brown, 1996). They should distinguish between the factors that bring a conflict into existence and those that make it intractable, since those factors identified by large N studies are usually of the latter kind. The age-old question of sociology – do structures really determine human action? – returns. Two countries with similar social structure may have very different experiences of conflict and a given society may experience phases of peace and of violent conflict despite the underlying fault lines remaining constant. These issues are acknowledged in Collier and Sambanis (2005a,b) series of cases studies, including the difficulty of identifying causal mechanisms with general models, and the question of why some political conflicts escalate into civil war while others do not.

The alternative (constructivist) approach focuses on relationships between actors. Interests and identities are not fixed by the material world, but emerge in an unpredictable way through processes of conflict and interaction. Moreover, perceptions can change during conflict, causing a shift in the underlying conditions of conflict that has significant consequences for its duration. For this reason process-based explanations are increasingly popular. Tilly (1973, p. 439) argued that theories that linked conflict to modernization for example, "fall wide of the mark," because they miss "the extent to which violence is a contingent outcome of interaction among contenders and governments." Tarrow's (2007, pp. 587–600) review of four studies identified a "(re)turn to case-study methods," and also concluded: 'it is not quantities but *interactions* that matter.' Increasing emphasis has been placed the way identities are formed during conflict.

Root cause approaches begin with a checklist of potentially relevant causal factors and then reorder them hierarchically for each case. Yet their importance varies over time. Process-based theories capture conflicts dynamics better. The Chechen struggle was essentially one of secession, but at different stages of the conflict it was viewed through different ideological frames. Hughes (2007) study shows that how an issue is reframed – from being about the desire for a territorial state to religious jihad – altered states' views about whether the Chechan issue was negotiable (ibid, 203. Thus a temporal perspective was required in order to know when a conflict was ripe for resolution. Moreover, identifying the causes of civil war is not the same as identifying the causes of its violence. Kalyvas' (2006) study of the Greek Civil War of the 1940s compares two villages in the plain of Argos, Manesi, and Gerbesi (now Midea), which were the same size, of a similar social outlook, and also shared a common descent and social structure. Egregious episodes of violence during the civil war (including of children and the elderly) took place in only one, in Gerbesi in August 1944. In the other, they were repelled by the villagers. The differences cannot be explained in terms of ideology, the claims of groups, or different economic structures.

Conclusion

If process-based approaches capture the dynamic nature of conflicts and their sensitivity to initial conditions, what remains of cause and effect? Processes are not necessarily causal: to say a situation produced a civil war because it polarized to that extent is tautological. Structures in contrast are 'emergent properties' which are the product of past actions giving them causal status in any society. The causal theories summarized here do not tell us enough about where these structures came from, who built and captured them, and for what purpose. Structures (including states) do not determine action, but set the parameters for action and carry meanings which inform the terms of any conflict about them. The relationship between state weakness and civil war must thus be affected by the meanings of structures for different groups. If their construction had meaning during decolonization; so too do their subsequent failures.

The civil war literature is marked by porous scholarly boundaries. The concept itself lacks conceptual autonomy and can legitimately be the province of international relations scholars. Many can only define civil war by contrasting it with other types of war. On the one hand, only science has the need to eliminate rival theories as a criterion of scholarly progress. Yet employing a single causal model for all the various types of internal war may actually obscure rather than illuminate the distinct causal factors at work in these types of conflict (Wood, 2003, p. 256). While the literature on revolution struggled to accommodate ideological factors (often specific to cases) in comparative analysis, that on civil war now tends to reduce identities and ideas to material motivations. Yet the state is a conceptual as well as material variable, the importance of which varies from society to society, so much that a loss of state capacity in one country will not have the same consequences as in another. A recent study of four internal wars in Latin America found no relationship between violence and attitudes to either state or nation (Centeno et al., 2013).

The field of civil war studies is new and interdisciplinary. Works of theoretical synthesis are badly needed. Ultimately, civil war should be regarded as a distinct form of internal war. Such conflicts can be internal in two ways; they take place within a political unit recognized as such by the international community, or in a society united by shared norms and values. Since levels of social integration, and of political development generally, are not constant, civil war studies have to say more about the political communities within which they take place. This is as important an ontological question as that of the contrast with other types of war. Throughout history such conflicts have polarized precisely because they put into doubt the basis of political communities. Whether this basis is institutional, economic, or cultural, we should accept that the foundations of social existence in our region will not be those of the rest of the world.

See also: Conflict and War: Anthropological Aspects; Insurgency; Military Sociology; Military, War, and Politics; Nation-State and War; Revolution; War and Nationalism; Warfare in History; Wars among Nation-States: Patterns and Causes.

Bibliography

Armitage, D., 2009. Ideas of civil war in seventeenth Century England. Annals of the Japanese Association for the Study of Puritanism 4, 4–18.

Aspinall, E., 2007. The construction of grievance: natural resources and identity in a separatist conflict. Journal of Conflict Resolution 51, 950–971.

Buber, M., 2002. The Way of Man. Routledge, London, New York.

Breuilly, J., 1996. Approaches to nationalism. In: Balakrishan, G., Anderson, B. (Eds.), Mapping the Nation. Verso, London, pp. 146–175.

Brown, M.B., 1996. The International Dimensions of Internal Conflict. Centre for Science and International Affairs. MIT Press, Cambridge, MA.

Brunner, O., Conze, W., Koselleck, R., 1972. Geschichtliche Grundbegriffe: Historiches Lexikon zur politisch-sozialer Sprache in Deutschland. Stuttgart.

Centeno, M.A., Cruz, J.M., Flores, R., Gustavo Silva Cano., 2013. Internal wars and Latin American Nationalism, In: Hall, J.A., Malešević, S. (Eds.), Nationalism and War. Cambridge, Cambridge University Press, pp. 279–306.

Cramer, C., 2006. Civil War Is Not a Stupid Thing. Hurst & Co., London.

Collier, P., Hoeffer, A., 2000. Greed and Grievance in Civil War. Policy Research Paper 2355. The World Bank, Washington, DC.

Collier, P., Hoeffler, A., Rohner, D., 2009. Beyond greed and grievance: feasibility and civil war. Oxford Economic Papers 61, 1–27.

Collier, P, Sambanis, N. (Eds.), 2005a. Understanding Civil Wars: Evidence and Analysis. Africa, vol 1. World Bank. Washington, DC.

Collier, P., Sambanis, N. (Eds.), 2005b. Understanding Civil Wars: Evidence and Analysis. Europe, Central Asia and Other Regions, Vol. 2. World Bank. Washington, DC.

De Romilly, J., 1991. The Rise and Fall of States According to Greek Authors. University of Michigan Press, Ann Arbour.

Eckstein, H., 1965. On the etiology of internal war. History and Theory 4 (2), 133–163.

Fearon, J.D., Laitin, D.D., 2003. Ethnicity, insurgency and civil war. American Political Science Review 9 (1), 75–90.

Fearon, J.D., 2004. Why do some civil wars last so much longer than others? Journal of Peace Research 41 (3), 275–301.

Goldstone, J.A., 2009. Pathways to state failure. In: Starr, H. (Ed.), Dealing with Failed States: Crossing Analytic Boundaries. Routledge, London, pp. 5–16.

Gurr, T., 1970. Why Men Rebel? Princeton University Press, Princeton.

Hegre, H., Ellengsen, T., Gates, S., Gladitsch, N.P., 2001. Toward a democratic civil peace? Democracy, political change, and civil war, 1816–1992. American Political Science Review 95 (1), 16–33.

Hegre, H., 2004. The duration and termination of civil war. Journal of Peace Research 41 (3), 243–252.

Holsti, K.J., 1996. The State, War, and the State of War. Cambridge University Press, Cambridge.

Hughes, J., 2007. Chechnya: From Nationalism to Jihad. University of Pennsylvania Press, Philadelphia.

Kaldor, M., 1999. New and Old Wars: Organised Violence in a Globalised Era. Cambridge University Press, Cambridge.

Kaufman, S.J., 2006. Symbolic politics or rational choice: testing theories of ethnic violence. International Security 30 (4), 45–86.

Kalyvas, S., 2006. The Logic of Violence in Civil War. Cambridge University Press, Cambridge.

Kissane, B., Sitter, N., 2012. Ideas in Conflict: Nationalism Literature and The Comparative Study of Civil War. Nationalism and Ethnic Politics 19 (1), pp. 38–53.

Koselleck, R., 2004. Historical criteria of the modern concept of revolution. In: Koselleck, R. (Ed.), Futures Past: On the Semantics of Historical Time. Columbia University Press, New York, pp. 43–58.

Luard, E., 1972. Civil conflicts in modern international relations. In: Luard, E. (Ed.), The International Regulation of Civil Wars. Thames and Hudson, London, pp. 7–25.

MacIntyre, A., 1967. A Short History of Ethics: A History of Moral Philosophy from the Homeric Age to the Twentieth Century. Routledge and Kegan Paul, London.

MacIntyre, A., 1971. A Short History of Ethics. Routledge, Kegan Paul, London.

Orend, B., Fall 2008. War, the Stanford Encyclopedia of Philosophy. In: Zalta, E.N. (Ed.). URL: http://plato.stanford.edu/archives/fall2008/entries/war/ (accessed 31.01.13.).

Payne, S., 2011. Civil War in Europe, 1905–1949. Cambridge University Press, Cambridge.

Posen, B.R., 1993. The security dilemma and ethnic conflict. Survival 35 (1), 27–47.

Ranzato, G., Corbain, A., 1994. Guerre fratricide: le guerre civili in eta contemporanea. Bollati Boringhieri, Torino.

Ringmar, E., 1996. Identity, Interest and Action: A Cultural Explanation of Sweden's Intervention in the Thirty Years War. Cambridge University Press, Cambridge.

Richards, P., Chauvea, J.P., 2008. West African insurgencies in Agrarian perspective: Côte d'Ivoire and Sierra Leone compared. Journal of Agrarian Change 8 (4), 515–552.

Sambanis, N., 2005. Conclusion: using case studies to refine and expand the theory of civil war. In: Collier, P., Sambanis, N. (Eds.), Understanding Civil War: Evidence and Analysis, Two Volumes. The World Bank, Washington, DC, pp. 303–335.

Schmidt, C., 1996. The Concept of the Political. Chicago University Press, Chicago.

Sills, D., 1968. International Encyclopaedia of the Social Sciences. MacMillan, New York.

Small, M., Singer, J.D., 1981. Resort to Arms: International and Civil Wars, 1816–1980. Sage, Beverly Hills, CA.

Tarrow, S., 2007. Inside insurgencies: politics and violence in an age of civil war. Perspectives on Politics 5 (3), 587–600.

Tilly, C., 1973. Does modernization breed revolution? Comparative Politics 5 (3), 425–447.

Thucydides, 1910. History of the Peloponnesian Wars (R. Crawley, Trans.) J.M. Dean. London.

Woodward, S.L., 1999. Bosnia and Herzegovina: how not to end civil war. In: Walter, B., Snyder, J. (Eds.), Civil Wars, Insecurity and Intervention. Columbia University Press, New York, pp. 73–116.

Wood, E.J., 2003. Civil wars: what we don't know. Global Governance 9, 247–260.

Civilization, Concept and History of

A Al-Azmeh, Yale University, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 1903–1909, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

The concept of civilization as it is recognizable today, emerged with the rise of historical consciousness in Europe in the eighteenth and nineteenth centuries, and achieved global spread in the twentieth century. Civilization came to constitute a primary unit of historical discourse, in association with cognate terms such as culture, quite apart from its indication of certain morphological features of human society, such as urbanism. Broadly conceived, the notion of civilization has served two schemas of world history, the one universalist and evolutionist and the other particularistic and vitalist. Both notions have ideological implications, and were often deployed in conflicts between universalist-progressivist and conservative-nationalist political creeds, the former laying emphasis on the normative and morphological continuities in human societies, while the latter stressing openness to historical becoming as well as societal and historical transformism. Quite apart from the normative implications of both notions, one valorizing abiding resources of particularist national and civilizational character and the other speaking for an open notion of progress, recent historical research has rendered possible the concrete and properly historical consideration of the notion of historical continuity beyond the boundaries of the ideological commitment of the two notions of civilization that have profoundly marked the categorization of historical material and historical periodization in general.

The concept of civilization is inextricably connected with the conditions of its emergence, most notably with the rise of historical consciousness in Europe in the eighteenth and nineteenth centuries, as well as the globalization of this form of historical understanding and correlative forms of intellectual practice. The concept is complex and imprecise in its definition, but ubiquitous in its uses, and inextricably imbricated with other categories by which historical materials are organized, such as culture, nation, and race. Apart from designating certain morphological features of human society, particularly with reference to urbanism and urbanity, civilization has been a schema for historical categorization and for the organization of historical materials. Here it has generally taken two forms, the universalist evolutionist, and the romantic particularist. The latter was tending to regain, in the ascendant context of identity politics, a certain hegemonic primacy worldwide at the close of the twentieth century. In all, the concept of civilization forms a crucial chapter in the conceptual, social, and political history of history; it, or its equivalents are presupposed, implicitly or explicitly, in the construal and writing of almost all histories.

Prehistory

The mental and social conditions for speaking about civilization in a manner recognizable in the year 2000 were not available before the middle of the eighteenth century. Hitherto, in Europe as elsewhere, large-scale and long-term historical phenomena, which later came to be designated as civilizations, had been categorized in a static manner that precluded the consciousness of directional or vectorial historicity as distinct from the mere register of vicarious change.

Hitherto, the succession of large-scale historical phenomena, such as Romanity or Islam, had been regarded (1) typologically, most specifically in the salvation-historical perspective of monotheistic religious discourse, in which successive events are taken for prefigurations and accomplishments of each other; (2) in terms of the regnal succession of world empires; and (3) in the genre of regnal succession, which started with the Babylonian king lists and the earliest stages of Chinese historical writing, and culminated in medieval Arabic historical writing. Not even the schema of state cycles evolved by the celebrated Ibn Khaldun (d. 1406), where civilization (*'umrân*) was quasisociologically identified with various organizational forms of human habitation and sociality, could meaningfully escape from this finite repertoire of possible historical conceptions.

In the perspective of typology, the continuity of historical phenomena was expressed in the repetition of prophecies successively reaffirming divine intent and inaugurating a final form of order whose *telos* would be the end of time. Thus the Jewish prophets repeat each other and are all figures for Abraham; Jesus is at once the repetition and termination of this unique cycle of terrestrial time and is prefigured in Jewish prophecies. Muhammad is the final accomplishment and the consummation of earlier prophetic revelations, prefigured in Jewish and Christian scriptures; his era inaugurates the consummation of time with the Apocalypse. The structure of time in the Talmud, in the Christian writings of Eusebius (d. 339), of Augustine's Spanish pupil Orosius (fl. 418), of Bernard de Clairvaux (1153), as in Muslim writings such as Muhammad's biography by Ibn Ishâq (d. after 761), or the universal histories of Tabarî (d. 923) and Ibn Kathîr (d. 1373) is homologous.

All but Jewish typology is independent of ethnic origin or geographical location, and construes historically significant units as religious communities of ecumenical description. This yields the second mode of organizing historical phenomena, the regnal. Thus, the regnal categorization of long-term historical phenomena of broad extent was expressed in terms of the succession of four ecumenical world empires, succeeding one another as the central actors in world history: the

Assyrio-Babylonian, the Median-Persian, the Alexandrian-Macedonian, and the Roman. This perspective was shared by the *Book of Daniel* and by ecclesiastical works, especially Syrian and Byzantine apocalypses influenced by it, albeit with minor variations, as in Orosius' substitution of the Carthaginian world empire for the Median-Persian. Muslim caliphs considered their own ecumenical world empire to be the fifth and final order of world history, a conception shared by Muslim apocalypticism and, with many complications and nuances, by universal histories written in Arabic, all of which regarded dynastic succession as both prophetic inheritance and as the renewal of ecumenical imperial ambition, transferred from one line to another.

Analogously, medieval Christian polities, Byzantine as well as Frankish, subscribed to the same theory of *translatio imperii* by regarding themselves as being in a direct line of typological continuity with Rome, variously through Byzantium, the 'New Rome,' or the Holy Roman Empire. In both, Romanity was the worldly cement of Christianity. This was a conception developed by Eusebius for his contemporary overlord Constantine, and was to remain effective until the dawn of modern times. In all cases, the past was understood to have been completed at its inception, with subsequent polities reenacting the foundational event.

Finally, mention must be made of the disjunction between these metahistorical and transcendental realms of typological continuity, and the all-too-human chaos of particular histories. No movement or qualitative change is discernible in the context of these, only the predictable succession of wars, rapine, pestilence, and occasionally of praiseworthy acts, without connection with an order of reality that might transcend the events themselves and render sense unto them.

Two roughly contemporary events heralded a new conception of history that made the modern notion of civilization conceivable. Both were specifically European, but their conceptual consequences were globalized in the course of the nineteenth and twentieth centuries. The first was Humanism, particularly in Italy, which for the first time broke the spell of Roman continuity by construing the immediate past as an age of darkness, and by confining Roman grandeur to the republican and early imperial ages. Thus Petrarch's (d. 1374) project of classicism counterposed to living tradition; Flavio Biondo's (d. 1463) anticipation of division of history into the classical, the medieval, and the modern; and Lorenzo Valla's (d. 1457) refutation of medieval documentary forgeries such as the *Donation Constantini*, based on an argument from anachronism: together, these laid the ground for a view of history as the domain of change rather than of repetition.

Needless to say, the notion of *translatio imperii* was no longer tenable in this context. It is at this point that Humanism converged with the other event foundational of the modern historical consciousness, namely, the Reformation. Criticism of the Church by Wycliff (d. 1384) and Luther, and the historiographic expression of this antitraditionalist fundamentalism in John Foxe's *Acts and Monuments* (1563), were in crucial ways congruent in their conception of the past with Humanism. The identification of the Pope with the Antichrist, the designation of the greater part of the history of Christianity as a history of falsehood, and the devalorization of the immediate past as abiding tradition and its construal as degeneration, also led to the rejection of the notion of *translatio imperii*, and the substitution of notions of Reformation and renovation to that of transfer in continuity.

Thus the ground was prepared for notions of rise, decline, and fall – most notably the decline and fall of Rome and the reclamation of Roman republicanist models in a spirit of revivalism – that were finally to mature in the eighteenth century, with Gibbon and Montesquieu among others, spurred along with the development and ultimately, in the late eighteenth and during the nineteenth centuries, the institutional transformation of philology and antiquarianism into history as a topic of research detached from rhetoric. This was far beyond the late flowering of medievalizing typology with Bossuet (d. 1704), and it made the past tangible in its having been (*Vergangenheit*), most graphically represented in the establishment of museums during the eighteenth century in many European capital cities.

With Voltaire and other eighteenth-century figures like Volney and Chardin, another notion crucial for speaking of civilization was developed. The notion of qualitative societal and cultural difference (*les moeurs*) – quite apart from the dynastic and the religious – was now available in the eighteenth century, as Europe was accounting for her differences from the Ottomans, the Persians, the Chinese, and tribal peoples in the Americas.

Whereas previously the notion of historical senescence may have been used in a tragical and rhetorical sense, the new notion of decadence required the correlative notion of progress and amelioration. These notions are the very conditions of possibility for conceiving of civilization as the accomplishment of a continuous line of historical development in which origins and beginnings are transcended rather than repeated.

Word and Concept

The terms civilization and culture are intimately related in their reference, and in many instances are used almost interchangeably, according to national and linguistic conventions. Both are terms of ancient vintage, which underwent a gradual lexical expansion until, in the eighteenth century, they came to designate meanings that are recognizable in 2001.

In the course of the seventeenth and eighteenth centuries, the meaning of the term culture expanded, in English, French, and German (as *Kultur*) from a medieval sense indicating the cultivation of land and of the religious cult, figuratively to denote the maintenance and cultivation of arts and letters. This figurative sense was further extended in the course of the eighteenth century to encompass the nonmaterial life of human societies in a very broad sense encompassing the refinement of manners no less than intellectual and artistic accomplishments associated with the Enlightenment – the former sense still persists in terms such as *haute couture* and *Kulturbeutel* (vanity case).

The term was used both locally and in contrast to societies adjudged still living in a state of nature, although in German the accent had been on an esthetic of the lofty and the sublime as distinct from the crassly material, in association with a correlative emphasis on cultivation, *Bildung*, both individual and collective. In this way, the term was opened up to

impregnation by the emergent notion of progress, the progress of individual societies as of humanity in general regarded both as a process natural to human society and as a principle of normative ranking among societies.

Not dissimilarly, and in imitation of 'culture,' in the eighteenth century the term civilization underwent – especially in France and somewhat later in England – a figurative expansion in its lexical reference from the Latin *civilis*, life under reputable forms of government, to the broader designation of order, civil, and governmental. This order befitted developed societies that might regard themselves as civilized in contradistinction to other, barbarian or savage, yet-to-be-civilized societies. The contrastive connotations of 'civilization' were far more accentuated than those of 'culture,' which was more commonly used in Germany, a land then with little or no experience of the world outside Europe. 'Culture' also came to be used in France and more saliently in England, as in Germany, decidedly to signal social distance and social distinctions within particular countries.

In all cases, these two terms increasingly came to be associated with a developmental perspective on history: not only the linear and cumulative course traversed by historical phenomena in time but also of languages, geological layers, plant and animal species, and human societies generally (and later, races), toward greater differentiation, complexity, and accomplishment. Correlatively, the meanings conveyed by these terms were implied by other terms or by none at all, as, for example, with Rousseau and Voltaire.

The nineteenth century witnessed a complicated relationship between 'culture' and 'civilization' whose fields of connotation and denotation shade into each other in a manner that has not helped the distinctive clarity and definition of either. The crucial player here was Germany, where *Kultur* took a decidedly romantic-nationalist turn in the early nineteenth century, dwelling on national uniqueness and individuality, and buttressed by the emergence of *Kulturwissenschaft* as a discipline and by studies of folklore.

The further developments of this politicocultural impulse toward the end of the nineteenth century led not only to the profuse discourses on decadence, *Entartung*, ('disnaturation') but also to the extension to France of this particularitic understanding of culture – and of civilization – under the influence of de Joseph de Maistre and the Catholic Counter-Enlightenment. As a result of Franco-German conflicts and of the severe stresses within France, a battle was waged between the advocates of 'culture,' upholders of national particularism, and of 'civilization,' champions of Enlightenment universalism accused by their detractors of crass materialism, a battle that reached its apogee during the First World War in the polemics between Romain Rolland and Thomas Mann.

Yet 'civilization' itself had been increasingly more receptive to particularism and nationalism, most specifically in historical writing. Books on the civilization of France, Germany, Europe, Italy, and England, emerged from the second quarter of the nineteenth century, and in certain ways, the very currency of the term made it open to divergent uses. Toward the end of that century, the term came to be used under the influence of the German notion of *Kultur*, a term often rendered as 'civilization' in French translations of German works. One additional but decisive factor was English anthropology, with the appearance in 1871 of Edward Tylor's *Primitive Culture* (1958), and the subsequent predominance of the term 'culture' to designate a condition, on a ladder leading from savagery to barbarism and finally to civilization, in Anglo-Saxon anthropology. With the discovery and mystique of classical Greece toward the end of the eighteenth century, of the unity of the 'West' from Greece, through Rome, on to the Romano-Germanic peoples in the Middle Ages, culminating in modern European civilization, was to become the *locus classicus* of this notion of civilization.

In the later part of the nineteenth century and throughout the twentieth century, both the terms and conceptions outlined were subsequently taken over and became crucial instruments of historical categorization in India, China, and the Arab countries and elsewhere. In Arabic, *thaqafa*, an equivalent to the German *Bildung*, came to stand for culture, normally taken to designate intellectual and artistic life in a manner strongly elitist in character, and *hadara* was used for civilization, taken as a more general concept indicating the entire life of society including material life.

Continuities: Relativism

Linear developmentalism had, in general, underlain the entire body of diverse discourses on civilization and of its sister concept of culture. This developmentalism bifurcated along lines that may be characterized broadly as German and French in original inspiration.

Of the two, the former had been more intimately associated with romantic national and, the latter, with civilizational particularism, producing a natural history of human groups regarded in analogy to organic species. These human groups were thus conceived of as self-subsistent, continuous over time, largely impermeable, essentially intransitive, and according to many representatives of this view, almost congenitally given to conflict and war.

Originating in conservative reactions to the Enlightenment, with a strong anti-Gallic political impulse, this theory of historical and social development was associated with figures such as Johann Gottfried Herder in Germany and Edmund Burke in England, although it did have a strong representation in France among royalist, Catholic, and other antirevolutionary (1789, 1848, 1871) currents represented by figures such as de Maistre, Gobineau, and Gustave Le Bon. The principal conceptual feature of this antimechanistic concept of history was insistence on individuality in the histories of different nations, races, and civilizations – terms often conflated in various combinations, in analogy with biological organisms, and perhaps best captured in the capacious semantic field of the German word *Volk*. In this sense, one may speak of continuity with pre-Enlightenment concepts of a social organism modeled upon the integrated somatic unity of the human body, as had been previously thought in medieval Arabic historicopolitical writings and in medieval European conceptions from the time of John of Salisbury (d. 1180). The fear of decline and decadence, conceived as a breakdown of a natural order, was the specific point at which continuity with medieval organismic concepts of the historicopolitical order made itself evident and conceptually formative. This was at a time when the idea of progress – and in contrast to it – had become a genuinely historical category, involving a consequential

conception of change, of evolutionism, of the temporality specific to events (*Verzeitlichung*), and of a distanciation correct between human and natural histories.

The course of a particular history was seen to reside in a number of essential features, which Herder termed *Kräfte*, resulting in a history which, increasingly elevated and evolutive in the course of time as it may be, was still governed by principles that were, in essence, changeless, principles that imparted individuality upon these intransitive histories.

Whereas the Enlightenment provided, and the nineteenth century elaborated, a notion of genetic development along an axis of cumulative time to which civilizations and other historical masses are subject, the organismic, particularist notion conceived of a civilization as bound to a self-enclosure inherent in its origins. Although this did not necessarily lead to a historical cyclism, it constituted its conceptual condition of possibility, and facilitated culturalist notions of nationalism that spoke in terms of 'revival.'

It was a cyclical notion of the history of cultural and civilizational circles, *Kulturkreise*, according to Ernst Troeltsch (1920) that was made conceivable by the manifold discourses on decadence, malfunction, and historical pathology, notions that implicitly involve a measure against a more consummate state of organic health and well-being implicit in the foundations of each. It was the systematic elaboration of the decadence/normalcy structure that led to great schemas of world history, divided into intransitive civilizations, of Oswald Spengler's *Untergang des Abendlandes* (Spengler, 1922) and of the Spenglerian heresy represented by Arnold Toynbee's *A Study of History* (Toynbee, 1934–1959).

Described by Claude Chaunu (1984 as 'a samsara of historical forms,' these vitalist theories of the rise and inevitable decline of civilizations constitute a naturalistic morphology of historical becoming. From this perspective, civilizations were seen as historical phenomena that are perpetually in conflict with one another, each endowed with a particular ethos or animating principle. Such were Spengler's (1922) Magian culture (the Perso-Islamic) and Promethean culture (the European). Such were also Toynbee's (1934–1959) Syriac and other civilizations, although with this latter author the inner definition of civilizations was less clearly predetermined, and founded on a firmer and far more scrupulous empirical foundation than with Spengler. Nevertheless, Toynbee does characterize civilizations in terms of particularistic impulses, such as the estheticism of Greek civilization, the religious spirit of the Indian, and the mechanistic ethos of the West. Each of these is an integrated pattern of daily life, on attitude toward the holy, a style of jurisprudence, a manner of government, an artistic style, and much more.

With both authors, the historical phenomena, respectively, designated as 'cultures' and 'societies,' obey an iron law of rise and decline, of glory and senescent atrophy, the terminal phases of which Spengler, in keeping with German usage of the day, derisively termed 'civilization.' It is noteworthy that this morphology of historical masses, be they called civilizations, cultures, or societies, that this monocausal description in terms of basic traits such as the Promethean or the esthetic, was congruent with certain developments in anthropology, particularly in American anthropology in the first half of the twentieth century (Franz Boas, Alfred Kroeber, Ruth Benedict, and Edward Sapir), which identified separate societies according to self-consistent and intransitive personality profiles that they ostensibly gave rise to. This had a decisive influence on the introduction of organismic thinking into the human sciences in general. After a long period of disrepute, the 'culturalized' notions of human collectivities, of self-enclosure, and of continuity have come back to center stage at the close of the twentieth century, correlatively with the politics of identity worldwide.

The organismic and vitalist notion of civilization was, and still is, extremely effective in the writing of history and in the late twentieth century has had a certain political salience in terms of Samuel Huntington's 'war of civilizations' and the mirror-image riposte in terms of the 'dialogue of civilizations.' It denies the possibility of a general human history, which it regards, in the words of Ernst Troeltsch (1920), as being 'violently monistic,' and construes the task of a history of civilizations as separate histories of Europe, China, India, Islam, Byzantium, Russia, Latin, and Protestant Europe, and others, in various possible permutations and successions, as separate cultural and moral spheres that are merely contiguous in space. Correlatively, this constituted the conceptual armature of certain forms of violently particularistic history that was mirrored or endogenously paralleled, among others in India, for instance (Savarkar, 1969), or in the writings of radical Muslim ideologues.

Novelties: Universalism

The deficit in historicity evident in romantic and vitalist historism, with its emphasis on organic continuity, was to a great extent made up in the more consistently evolutionist accounts of historicism. It was this conception of history, at once evaluative, evolutive, and vectorial, which bore the burden of the universalist notion of civilization. Whereas in historism the bearers of civilization – or rather of 'culture' – were particular peoples or individual historical itineraries, such as the West or Islam, the whole of humanity partook of the development of civilization, which, in the historicist perspective, was universal and continuous.

Several versions of this were in evidence, of which two are of particular note due to their wide conceptual incidence and the social and political influence they exercised through worldwide social movements, both revolutionary and gradualist, inspired by the Enlightenment. According to this conception of a universal civilization, human societies pass through a uniform series of progressive developments, which result in intellectual and moral elevation. They also result in superior social, economic, and political levels of development, marked by a higher and wider order of rationality, social differentiation, and control over nature whose instrument is science.

The more consistently universalist of these versions – exemplified, among others, by Jean Marquis de Condorcet, Auguste Comte, and Herbert Spencer, and in the evolutionist anthropology of Tylor (1958) – saw the whole of humanity as being predisposed to this upward movement. Nevertheless, according to this view some peoples may still subsist in a condition that others (Europe) had already surpassed, being still captive to superstition, to a weak organization of society and the economy, and to undeveloped political institutes

(despotism, as in the case of Orientals, or otherwise various forms of acephalic organization, as is the case with primitive societies). The only caveat here is that many theorists of the evolution of human societies did not see human improvement as uniform in developed societies themselves, but that developed societies were not only internally differentiated in the levels of accomplishment attained by different social groups, but also liable to fall far below the moral ideals that development makes possible. Nevertheless, much writing on the history of universal civilization, especially in countries influenced by Marxism, such as the Soviet Union or the People's Republic of China, saw the historical itineraries they led as crucial and exemplary steps in the development of humanity at large. In this way, the histories of Russia and China come to recuperate and tap world history at large by acting as the vanguard and exemplar of its future consummation.

Correlatively, the other conception of universalism, and this is one of profound conceptual and political importance, recognized the past contributions of various civilizations – the Mesopotamian, Egyptian, and Islamic, and in some instances the Indian and the Chinese as well – to the course of human civilization, which eventually lodged itself in Europe, defined as the abode of Romano-Germanic history, or even in small parts of Europe, such as France or Prussia. This writing of universal history, much in evidence in the plethoric literature of universal histories both popular and learned, was expressed in what is perhaps its most accomplished general formulation by Georg Hegel (1956) in the nineteenth century, and by Karl Jaspers (1949) in the twentieth – the latter is little read today, but he is the one who, nevertheless, captures this notion with special clarity.

Not infrequently, this second type of universalist history is allied to an important element derived from the romantic theory of the history of civilizations treated in Section Novelties: Universalism, namely, the presumption of very long-term individual continuity. Thus the point is habitually made, with varying shades of emphasis and nuance, that universal civilization made Europe its eventual home because of some abiding characteristics possessed *ab initio* by 'the West,' such as rationality, the spirit of freedom, vigor, and dynamism. This changeless West is counterposed to an eternal East, such as Mesopotamia, Egypt, and Islam, which, their relative erstwhile merits apart, constitute, in a categorical degradation, a mere prehistory to fully developed civilization. East and West – not to speak of Islam, the name of a religion transmuted into an atopic location – are metageographical notions that do not allow, except with anachronistic violence, for projection into the antique and late antique worlds. However, this does not disturb the ideological coherence of this notion of civilization.

Beyond Totality

Writing about civilization according to the manners outlined contained both large-scale abstraction and a great deal of precise empirical historiography. Under the influence of Marxism, acknowledged as such as well as implicit, historical scholarship, and most specifically social and economic history in the traditions of Max Weber and of the Annales school, produced specifications regarding material and other aspects of civilization that allayed, to a considerable degree, the rhetorical force of thinking civilizations in terms of purely moral and ideological continuities, betokening exclusive sociohistorical groups. The notion of a Judeo-Christian civilization, as distinct from textual typology within the Bible, is an excellent illustration of this, having been born and expanded in specific circumstances following the Second World War.

In this way, civilizations for contemporary historical writing have come to comprise the total historical conditions that exist in a specific place and time: functional and organizational forms of the state; the *longue durée* of demographic, agricultural, economic, social, urban, ecological, climatic, and other forces underwritten by geographical structures and relations; and nonmaterial culture, such as arts, letters, cognitive structures, and religions. The possibilities of a total history of a civilization is, it must be stressed, an expansion of a previous and more limited form of the cultural history of a particular epoch, exemplified by Jakob Burckhardt's studies of Constantine and Renaissance Italy (Burckhardt, 1955: Vols. 1 and 3).

As a result, since about 1950 it has become possible historically to specify the material elements that constitute the history of a civilization; however, its temporal and geographical boundaries may be defined. Correlatively, it has become possible to conceive of the specific differences between histories – China and Europe for instance, as in the work of Jacques Gernet (1988) – beyond a discourse on immobility and other immanent characteristics ascribed to this history or that, and to think of specificity in proper historical terms, such as the relative weight of various elements of the rural economy, the relation between state and the economy, the impact of metallurgy, and much more.

In this context, civilizational continuities came to be reconsidered in terms of historically determinant factors of a predominantly geographical nature, not so much in the spirit of geographical determinism as described by the German school of Friedrich Ratzel (1895), but with a greater degree of temporal specification, mediated by Lucien Febvre's (1925) consideration of historical geography and culminating in Fernand Braudel's study of the Mediterranean (1972–73). By the same token, it has become possible squarely to face the nominalist caution required in thinking about civilizations, and to think of their constitution, specification, and collapse in terms of the concrete historical investigation of demography, economy, and society without recourse to the metaphysical rhetoric of decline (Tainter, 1988).

These specifications apart, it remains true that the construal of civilizational intransitivity, with Braudel as with others, still needs to resort to a new redaction of the rhetoric of permanence, most particularly with regard to nonmaterial culture, now underscored and almost overdetermined by considerations of relief, soil, water supply, and means of transport – all of which are undeniable factors, albeit ones that modern technology and economy, most poignantly the postmodern economy, have rendered questionable.

Nevertheless, recent historical research has made it concretely possible to tap the genial formulations made by Marcel Mauss in 1930 (Febvre et al., 1930) concerning the categorization of historical masses: of civilizations as a 'hypersocial systems of social systems,' as transsocietal and extranational units of historical perception and categorization. These

are conceived of in opposition to specific social phenomena, and this conception of civilization valorizes the distinction between civilization, society, and culture, freeing the first of the deterministic and totalizing rhetorical glosses of metahistory, and making possible a veritable history of civilizations. Civilization may thus be considered as at once a particular instance of historical becoming, and a specific ideological redaction of the past whose relation to historical reality can be questioned and rendered historical. In this context, the historian may also be able to valorize the extremely expansive *longue durée* implied by such theories as Dumézil's (1958) Indo-European tri-functionalism without recourse to organismic and totalizing figures of particularity and continuity (Le Goff, 1965). A historian may be similarly able to valorize recent studies that stress the occurrence and communicability of recurrent phenomena of the imaginary order across vast spaces, cultures, histories, and times, as instanced by Carlo Ginzburg's study of the European witch hunts (Ginzburg, 1990). Finally, given the accent on complexity, one might be able to take the precise and nuanced study of levels and modes of socio-economic, political, institutional, ideational, and other instances of complexity – rather than criteria of simple continuity – as crucial to the delimitation of historical phenomena that one designates as 'civilizations.'

See also: Civilizations; Enlightenment; Global History: Universal and World; History: Overview; Modernization and Modernity in History; Societies, Types of; Society and People: History of the Concept; State and Society: History of the Concept; State, History of; Time, Chronology, and Periodization in History.

Bibliography

Allawi, A.A., 2009. The Crisis of Islamic Civilization. Yale University Press, New Haven, CT.
Auerbach, E., 1984. Figura. Auerbach E Scenes from the Drama of European Literature. University of Minnesota Press, MN.
Bénéton, P., 1975. Histoire des Mots: Culture et Civilisation. Fondation Nationale des Sciences Politiques, Paris.
Braudel, F., 1972–73 (S. Reynolds, Trans.). The Mediterranean and the Mediterranean World in the Age of Philip II, 2 vols. Collins, London.
Braudel, F., 1993. Grammaire des Civilisations. Flammarion, Paris.
Burckhardt, J., 1955. Renaissance Italy. Gesemmete Werke, vols. 1 and 3. Schwab, Basle, Switzerland.
Chaunu, P., 1981. Histoire et Décadence. Librairie academique Perrin, Paris.
Collingwood, R.G., 1946. The Idea of History. Clarendon Press, Oxford, UK.
Dumézil, G., 1958. L'idéologie Tripartie des Indo-Européens. Collection Latours, Bruxelles, Belgium.
Enciclopedia Einaudi, 1977–84. G. Einaudi, Torino (s.v. 'Cultura materiale', 'Civilit').
Encyclopædia Universalis, 1968. Encyclopædia Universalis France, Paris (s.v. 'Civilization' 'Culture et civilization').
Encyclopedia of Religion and Ethics, 1980. T and T Clark, Edinburgh, UK (s.v. 'Civilisation').
Febvre, L., 1925. A Geographical Introduction to History. Kegan Paul, Trench, Trubner & Co., London.
Febvre, L., Mauss, M., Tonnelat, E., Nicophoro, A., 1930. Civilisation. Le Mot et l'Idée. Renaissance du Livre(Premire Semaine internationale de Synthèse, Fascicule 2), Paris.
Fehl, E. (Ed.), 1971. Chinese and World History. The Chinese University of Hong Kong, Hong Kong.
Ferguson, N., 2011. Civilization: The West and the Rest. Penguin Press, New York.
Gat, A., 2006. War in Human Civilization. Oxford University Press, Oxford. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=186606.
Gernet, J., 1988. A History of Chinese Civilization (J.R. Forster, Trans.). Cambridge University Press, Cambridge, UK.
Ginzburg, C., 1990. Ecstasies: Deciphering the Witches' Sabbath. Hutchinson, London.
Hegel, G.W.F., 1956. Philosophy of History (G. Sibree, Trans.). Dover, New York.
Herder, J.G.von, 1968. Reflections on the Philosophy of History of Mankind (Abridged by F. Manuel). University of Chicago Press, Chicago and London.
Herder, J.G.von, 1969. J.G. Herder on Social and Political Culture (F.M. Barnard, Ed. and Trans.). Cambridge University Press, Cambridge, UK.
Jaspers, K., 1949. Vom Ursprung und Ziel der Geschichte. Artemis Verlag, Zrich, Switzerland.
Kemp, A., 1991. The Estrangement of the Past. A Study in the Origin of Modern Historical Consciousness. Oxford University Press, Oxford and New York.
Khaldun, Ibn, 1958. The Muqaddimah: An Introduction to History (F. Rosenthal, Trans.). Pantheon Books, New York.
Kosellek, R., Widmer, P. (Eds.), 1980. Niedergang. Studien zu einem geschichtlichen Thema. Klett-Cotta, Stuttgart, Germany.
Kroeber, A.L., 1963. An Anthropologist Looks at History. University of California Press, Berkeley and Los Angeles.
Kroeber, A.L., Kluckhohn, C., 1952. Culture: A Critical Review of Concepts and Definitions. The Peabody Museum, Cambridge, MA.
Le Goff, J., 1965. La Civilisation au l'occident médieval. Arthaud, Paris.
Lewis, M.W., Wigen, K.E., 1997. The Myth of Continents. A Critique of Metageography. University of California Press, Berkeley and Los Angeles.
Marrou, H.I., December 1938. Culture, civilization, décadence. Revue de Synthèse 133.
Nasr, S.H., 2003. Islam: Religion, History, and Civilization. HarperSanFrancisco, San Francisco.
Ratzel, F., 1895. Anthropogeographische Beitrage. Duncker und Humblot, Leipzig, Germany.
Rsen, J., Gottlob, M., Mittag, A. (Eds.), 1998. Die Vielfalt der Kulturen. Erinnerung, Geschichte, Identität, 4. Suhrkamp, Frankfurt am Main, Germany.
Savarkar, V.D., 1969. Hindutva: Who Is a Hindu?, fifth ed. Veer Savarkar Parakashan, Bombay, India.
Schlanger, J.E., 1971. Les Métaphores de l'Organisme. Vrin, Paris.
Searle, J.R., 2010. Making the Social World: The Structure of Human Civilization. Oxford University Press, Oxford.
Solomon, S., 2010. Water: The Epic Struggle for Wealth, Power, and Civilization. Harper, New York.
Spengler, O., 1922. Der Untergang des Abendlandes, 2 vols. W. Braumller, Vienna and Leipzig.
Tainter, J.A. (Ed.), 1988. The Collapse of Complex Societies. Cambridge University Press, Cambridge, UK.
Toynbee, A., 1934–1959. A Study of History, 12 vols. Oxford University Press, London.
Troeltsch, E., 1920. Der Aufbau der europäischen Kulturgeschichte. Schmollers Jahrbuch für Gesetzgebung, Verwaltung, und Volkswirtschaft im Deutschen Reiche 44 (3), 1–48.
Tylor, E.B., 1958. Primitive Culture, 2 vols. Harper, New York.
Wells, S., 2010. Pandora's Seed: The Unforseen Cost of Civilization. Random House, New York.
Zureiq, K., 1969. Nahnu Wa't-tarikh. Dar al- Ilm lil-Malayin, Beirut, Lebanon.

Civilizations

Samuel N Eisenstadt, Van Leer Jerusalem Institute, Jerusalem, Israel

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 1915–1921, © 2001, Elsevier Ltd.

Abstract

The central analytical core of the term civilization as employed here is the combination of ontological or cosmological visions, of visions of transmundane and mundane reality, with the definition, construction, and regulation of the major arenas of social life and interaction. The full development of the distinct ideological and institutional dimensions, and of some awareness of their distinctiveness, has emerged in some very specific historical settings – namely, the so-called Axial Civilizations. The crystallization of these civilizations constitutes a series of some of the greatest revolutionary breakthroughs in human history, which have shaped contours of human history in the last two to three millennia. The central aspect of these breakthroughs was the emergence and institutionalization of new ontological metaphysical conceptions of a chasm between the transcendental and mundane orders.

Introduction

The term 'civilization' has been used in modern social science and historical literature in several different ways. One such way, developed above all in Germany from about the end of the nineteenth century through the period up to World War II and perhaps best represented by scholars like Alfred Weber (and taken over to a certain extent in the English-speaking world by R. M. Mclver), designated 'civilization' – as distinct from 'society' and above all from 'culture' – as encompassing above all the technological, material factors and to some extent organizational aspects of social life as against the deeper, more 'spiritual' cultural and aesthetic ones.

Another designation of the term, made famous by Norbert Elias in his *Über den Prozess der Zivilisation* (1939) focused on the 'socializing' process through which the image of the civilized person, as constructed in the courtly and also early bourgeois society in Europe, was promulgated and institutionalized. This designation of civilization was related to an earlier one rooted in the French Enlightenment, in which civilization was seen as the opposite of barbarism. However, in later works by Elias's followers, for instance Goudsblom, this view of civilization was extended to cover many other societies and historical periods, going back even to the impact of the presumably first domestication of fire.

The third and most extensive designation of civilization was promulgated by scholars such as Max Weber, Emile Durkheim, Oswald Spengler, Pitirim Sorokin, Arnold Toynbee, A.L. Kroeber, Carroll Quigley, Cristopher Dawson, Fernand Braudel, William H. McNeill, Adda Bozeman, or Immanuel Wallerstein, and lately very forcefully by Samuel Huntington. However great the differences in perspective, methodology, focus, and concepts that pervade the works of these scholars, they share the use of the term civilizations as distinct societal-cultural units which share some very important, above all cultural, characteristics. Here we shall use the term civilization in a way very close to, but also distinct from, such a designation.

> Civilization as combination of ontological or cosmological visions, of conceptions of trans-mundane and mundane reality, with the definition, construction, and regulation of the major arenas of social life and interaction

The central analytical core of the term civilization as employed here – as distinct from such social formations as political regimes, different forms of political economy or collectivities like 'tribes,' ethnic groups or nations, and from religion or cultural traditions – is the combination of ontological or cosmological visions, of visions of trans-mundane and mundane reality, with the definition, construction, and regulation of the major arenas of social life and interaction.

The central core of civilizations is the symbolic and institutional inter-relation between the formulation, promulgation, articulation, and continuous reinter-pretation of the basic ontological visions prevalent in a society, its basic ideological premises and core symbols on the one hand, and the definition and regulation of major arenas of institutional life on the other. Such definitions and regulations construct the broad contours, boundaries, and meanings of the major institutional formations and their legitimization, and greatly influence their organization and dynamics.

The impact of such ontological visions and premises on institutional formation is effected through the processes of interaction and control that develop in a society. Such processes of control – and the opposition to them – are not limited to the exercise of power in the 'narrow' political sense. Rather, they are activated by major elites in a society. The most important such elite groups are the political, the cultural, and the economic ones and those which construct the solidarity and collective images of the major groups, all of which have different cultural visions and represent different interests.

The structure of such elite groups is closely related, on the one hand, to the basic cultural orientations prevalent in a society; that is, different types of elite groups bear different orientations or visions. On the other hand, and in connection with the types of cultural orientations and their respective transformation into basic premises of the social order, these elite groups tend to exercise different modes of control over the allocation of basic resources.

Such combination of ontological visions and of structuration of institutional formations and collective identities constitutes an inherent component of the formation of any society, and is always closely interwoven with the more

organizational aspect of any institutional formation – political, economic, or family and kinship.

The very implementation or institutionalization of such premises and the concomitant formation of institutional patterns through processes of control, symbolic and organizational alike, also generate tendencies to conflict and change. The crystallization of these potentialities of change usually takes place through the activities of secondary elite groups who attempt to mobilize various groups and resources to change aspects of the social order.

The full development of the distinct ideological and institutional dimensions, and of some awareness of their distinctiveness, has emerged in some very specific historical settings – namely, the so-called Axial Civilizations – even if some very important kernels thereof can be identified in some archaic civilizations such as those of ancient Egypt, Assyria, or Mesoamerica.

Axial Age Civilizations: the Reconstruction of the World and the Crystallization of Distinct Civilizational Complexes

By Axial Age civilizations (to use Karl Jaspers' nomenclature), we mean those civilizations that crystallized during the thousand years from 500 BC to the first century of the Christian era, within which new types of ontological visions, of conceptions of a basic tension between the transcendental and mundane orders, emerged and were institutionalized in many parts of the world. Examples of this include ancient Israel; later in Second-Commonwealth Judaism and Christianity; Ancient Greece; possibly Zoroastrianism in Iran; early imperial China; Hinduism and Buddhism; and, beyond the Axial Age proper, Islam.

The crystallization of these civilizations constitutes a series of some of the greatest revolutionary breakthroughs in human history, which have shaped contours of human history in the last two to three millennia. The central aspect of these breakthroughs was the emergence and institutionalization of new ontological metaphysical conceptions of a chasm between the transcendental and mundane orders.

The development and institutionalization of these ontological conceptions entailed the perception of the given mundane order as incomplete, inferior – oftentimes as evil and polluted. It gave rise in all these civilizations to attempts to reconstruct the mundane world, from the human personality to the sociopolitical and economic order, according to the appropriate 'higher' transcendental vision.

The revolutionary conceptions, which first developed among small groups of autonomous, relatively unattached 'intellectuals' (a new social element at the time), were ultimately transformed into the basic 'hegemonic' premises of their respective civilizations, and were subsequently institutionalized. That is, they became the predominant orientations of both the ruling elites and of many secondary elites, fully embodied in the centers or subcenters of their respective societies.

One of the most important manifestations of such attempts in all these civilizations was the strong tendency to construct societal centers to serve as the major autonomous and symbolically distinct embodiments of respective ontological visions, as the major loci of the charismatic dimension of human existence. But at the same time the 'givenness' of the centers could not necessarily be taken for granted. The construction and characteristics of the center tended to become central issues under the gaze of the increasing reflexivity which focused above all on the relations between the transcendental and mundane orders. The political dimension of such reflexivity was rooted in the transformed conceptions of the political arena and of the accountability of rulers. The political order as one of the central loci of the mundane order had to be restructured according to the precepts of the transcendental visions. The rulers were usually held responsible for organizing the political order according to such precepts.

At the same time the nature of rulers became greatly transformed. The king-god, embodiment of the cosmic and earthly order alike, disappeared, and a secular even if often semisacral ruler appeared. Thus there emerged the conception of the accountability of rulers and community to a higher authority, God, Divine Law, or a metaphysical vision. Accordingly, the possibility of calling a ruler to judgement appeared. One such dramatic appearance of this conception occurred in ancient Israel, in the priestly and prophetic pronunciations. 'Secular' conceptions of such accountability to the community and its laws appeared in both the northern shores of the eastern Mediterranean, in ancient Greece, as well as in the Chinese conception of the Mandate of Heaven.

Concomitantly with the emergence of these conceptions of accountability of rulers, autonomous spheres of law began to develop as somewhat distinct from purely customary law. Such developments could also entail some beginnings of a conception of rights even if the scope of these spheres of law and rights varied greatly. Of special importance from the point of view of our analysis is the fact that one of the most important manifestations of the attempts to reconstruct the social order was the strong tendency to define certain collectivities and institutional arenas as most appropriate for the implementation of transcendental visions.

Autonomous Elites as Bearers of Civilizational Visions: Change, Protest, and Heterodoxies

The development of new ontological metaphysical conceptions in the Axial civilization was closely connected with the emergence of a new type of elite, carriers of models of cultural and social order. These were often autonomous intellectuals, such as the ancient Israelite prophets and priests, and later on the Jewish sages, the Greek philosophers and sophists, the Chinese literati, the Hindu Brahmins, the Buddhist Sangha, the Islamic Ulema. Initial small nuclei of such groups of cultural elites developed the new ontologies, the new transcendental visions and conceptions, and were of crucial importance in the construction of the new 'civilizational' collectivities.

The new type of elites differed greatly from the ritual, magical, and sacral specialists in the pre-Axial Age civilizations. They were recruited and legitimized according to autonomous criteria, and were organized in settings distinct from those of the basic ascriptive political units of the society. The acquired a potentially countrywide and also trans-country status of their own. They also tended to become potentially independent of other categories of elites, social groups, and sectors.

At the same time a far-reaching transformation of other elites, such as political elites, took place. All these elites saw themselves not only as performing specific technical activities – be they those of scribes, ritual specialists, and the like – but also as potentially autonomous carriers of a distinct order related to the prevalent transcendental vision. They saw themselves as the autonomous articulators of the new order, and rival elites as both accountable to them and as essentially inferior. Moreover, each of these groups of elites was not homogeneous, and within each of them a multiplicity of secondary influentials developed.

These new groups became transformed into relatively autonomous partners in the major ruling coalitions. They also constituted the most active elements in the movements of protest and processes of change that developed in these societies and which evinced some very distinct characteristics at both symbolic and organizational levels.

First, there was a growing symbolic articulation and ideologization of the perennial themes of protest found in any human society, such as rebellion against the constraints of division of labor, authority, and hierarchy, and of the structuring of time dimension, the quest for solidarity and equality, and for overcoming human mortality.

Second, utopian orientations were incorporated into the rituals of rebellion and the double image of society. It was this incorporation that generated alternative conceptions of social order and new ways of bridging the distance between the existing and the 'true' resolution of the transcendental tension.

Third, new types of protest movements appeared. The most important were intellectual heterodoxies, sects, or movements which upheld the different conceptions of the resolution of the tension between the transcendental and the mundane order, and of the proper way to institutionalize such concepts. Since then, continuous confrontation between orthodoxy on the one hand, and schism and heterodoxy on the other, has been a crucial component in the history of mankind.

Fourth, and closely related to the former, was the possibility of the development of autonomous political movements and ideologies usually oriented against an existing political but possibly also religious center.

All these developments ushered into the arena of human history the possibility of the conscious ordering of society, and also the continuous tension that this possibility generated. The new dynamics of civilization transformed group conflicts into potential class and ideological conflicts, cult conflicts into struggles between the orthodox and the heterodox. Conflicts between tribes and societies could become missionary crusades. The zeal for reorganization informed by each civilization's transcendental vision made the entire world at least potentially subject to cultural- political reconstruction.

The Expansion of Axial Civilizations

Concomitantly with the institutionalization of Axial civilizations, a new type of intersocietal and intercivilizational world history emerged. To be sure, political and economic interconnection have existed between societies throughout human history. Some conceptions of a universal kingdom emerged in many post-Axial civilizations, like that of Genghis Khan, and many cultural interconnections developed between them, but only with the institutionalization of Axial civilizations did a more distinctive ideological and reflexive mode of expansion develop, with potentially strong semimissionary orientations.

It was indeed in close connection with the Axial civilizations' tendency to expansion that there developed new 'civilizational' collectivities, distinct from political and from 'primordial' ones, yet impinging on them, continuously challenging them, and provoking continual reconstruction of their respective collective identities. Such processes were effected by the interaction between the new autonomous cultural elites and the various carriers of solidarity and political elites of the different continually reconstructed 'local' and political communities.

In the continuous encounter of Axial civilizations with non-Axial or pre-Axial civilizations it was usually the Axial that came out victorious, without however necessarily obliterating many of the symbolic and institutional features of the non-Axial. The most important case of an encounter of non-Axial with Axial civilization in which the former absorbed the latter has been Japan.

The Multiplicity of Axial Civilizations and World Histories

The general tendency to reconstruct the world and to expand was common to all the post-Axial age civilizations. But the concrete implementation varied greatly. There emerged a multiplicity of different, divergent, yet mutually impinging world civilizations, each attempting to reconstruct the world in its own mode, and either to absorb the others or to segregate itself from them.

Two sets of conditions were of special importance in shaping these different modes of institutional creativity and expansion. One such set are variations in the basic cultural orientations. The other is the concrete structure of the social arenas in which these institutional tendencies can be played out.

Among the different cultural orientations the most important have been differences in the very definition of the tension between the transcendental and mundane orders and the modes of resolving this tension. There is the distinction between the definition of this tension in relatively secular terms (as in Confucianism and classical Chinese belief systems and, in a somewhat different way, in the Greek and Roman worlds) and those cases in which the tension was conceived in terms of a religious hiatus (as in the great monotheistic religions and Hinduism and Buddhism).

A second distinction is that between the monotheistic religions in which there was a concept of God standing outside the Universe and potentially guiding it, and those systems, like Hinduism and Buddhism, in which the transcendental, cosmic system was conceived in impersonal, almost metaphysical terms, and in a state of continuous existential tension with the mundane system. The 'secular' conception of this tension was connected, as in China and to some degree in the ancient world, with an almost wholly this-worldly conception of salvation.

A third major distinction refers to the focus of the resolution of the transcendental tensions. Here the contrast is between purely this-worldly, purely other-worldly, and mixed this-and other-worldly conceptions of salvation. The metaphysical nondeistic conception of this tension, as in Hinduism and

Buddhism, tends towards an other-worldly conception of salvation, while the great monotheistic religions emphasize different combinations of this-and otherworldly conceptions of the transcendental vision.

Another set of cultural orientations which influenced the expansion of the various Axial civilizations was the extent to which the access to their centers and major attributes of the sacred within them was open to all members of the community or was mediated by specific institutions.

In addition, there are differences in the way in which relations between the attributes of cosmic and social order of civilizational collectivities and those of the major primordial ascriptive collectivities are conceived – the extent to which there is a disjunction between the two, to which these respective attributes are mutually relevant, each serving as a referent of the other.

But the concrete working out of all such tendencies depends on the second set of conditions – namely the arenas for their concretization. These conditions included, first, the respective concrete economic political-ecological settings, whether they were small or great societies, whether they were societies with continuous compact boundaries, or with cross-cutting and flexible ones. Second was the specific historical experience of these civilizations – especially in terms of mutual penetration, conquest, or colonization.

Internal Transformation of the Axial Civilization: Secondary Breakthroughs and the Crystallization of Modern Civilization

One of the most important aspects of the dynamics of Axial civilizations was the possibility of development within them of internal transformation, of what has been designated as secondary breakthroughs, the most important illustrations of which have been Second Temple Judaism and Christianity; later Islam, Buddhism, and to a lesser extent Neo-Confucianism, all of which developed out of heterodox potentialities inherent in the respective 'original' Axial civilizations.

But the most dramatic transformation from within one of the Axial civilizations has probably been that of modernity as it first emerged in Western Europe and as it expanded – encompassing most parts of the world, giving rise to development of multiple, continually changing modernities.

The cultural and political program of modernity constituted in many ways a sectarian heterodox breakthrough in the West and Central European Christian Axial civilization. Such transformation took place through the Reformation and in the Great Revolutions, in which there developed very strong emphasis on the bringing together of the City of God and the City of Man. It was in these revolutions that sectarian activities were taken out from marginal or segregated sectors of society and became interwoven not only with rebellions, popular uprisings, movements of protest but also with the political struggle at the center, and were transposed into the general political movements and the centers thereof.

It was above all in the French Revolution that the fully secular transformation of the sectarian anti-nomian orientation with strong Gnostic components took place. It was epitomized in the Jacobin orientations which became a central component of the modern political program – to reappear yet again forcefully, as Alain Besancon has shown, in the Russian Revolution, and later in the Chinese and Vietnamese revolutions.

The strong sectarian roots of modernity and of the tensions between totalistic Jacobin and pluralistic orientations which developed in Europe find very strong resonance in the utopian sectarian traditions of the Axial civilizations. It is also the religious roots of the modern political program that explain the specific modern characteristics of what may be seen as the most antimodern contemporary movements – namely the various fundamentalist movements which, contrary to the view which defines them as traditional, constitute a new type of Jacobin movement constructing tradition as a totalistic ideology.

The Cultural and Political Program of Modernity: Premises and Antinomies

The cultural and political program of modernity, as it crystallized first in Western Europe from around the seventeenth century, was rooted in the premises of the European civilization and European historical experience and bore these imprints – but at the same time it was presented and perceived as being of universal validity and bearing.

The radical innovation of this cultural program as it developed in Europe lay first in the 'naturalization' of man, society, and nature; second in the promulgation of the autonomy and potential supremacy of reason in the exploration and even shaping of the world; and third the emphasis on the autonomy of man, of his reason and/or will.

In connection with these orientations there took place far-reaching transformations of the symbolism and structure of modern political centers, as compared with their predecessors in Europe or with the centers of other civilizations. The crux of this transformation was first the charismatization of the political centers as the bearers of the transcendental vision of the cultural program of modernity; second the development of continual tendencies to permeation of the peripheries by the centers and of the impingement of the peripheries on the centers, of the concomitant blurring of the distinctions between center and periphery; and third was the incorporation of themes of protest as basic, legitimate components of the premises of these centers. These themes became central components of the project of emancipation – a project which sought to combine equality and freedom, justice and autonomy, solidarity and identity of modern political discourse and practice. The program also entailed a distinctive mode of the construction of the boundaries of collective identities. Such identities were not taken as preordained by some transcendental authority, but continually constructed and continually problematized, becoming also foci of political struggle by national and ethnic movements.

The civilization of modernity as it developed in the West was from its very beginning beset by internal contradictions, giving rise to critical discourse which focused on the tensions and contradictions between its premises, and between these premises and institutional development. The most important such tensions in this program were first that between totalizing and more pluralistic conceptions of its major components – of the very conception of reason and its place in human life and society, and of the construction of nature, of human society and

its history; second, between reflexivity and active construction of nature and society; third, those between different evaluations of major dimensions of human experience; and fourth between control and autonomy.

These basic tensions, contradictions, and antinomies inherent in the cultural program of modernity were continually played and worked out in major institutional arenas.

Continually Changing Multiple Modernities

It was out of the conjunction of these cultural orientations with the development of market, commercial, and industrial economies; with the crystallization of a new political and state order; and with military and imperialist expansion, that the civilization of modernity emerged. Its crystallization and expansion were not unlike those of the expansion of all historical civilizations. What was new was first that the great technological advances and the dynamics of modern economic and political forces made this expansion, the changes and developments attendant on them and their impact on the societies to which it expanded much more intensive. The expansion, through the use of military, political, and economic forces, of modern civilization which took place first in Europe and then beyond it continually combined economic, political, and ideological aspects and forces, and its impact on the societies to which it expanded was much more intense than in most historical cases. It spawned a tendency – rather new and practically unique in the history of mankind – to the development of universal, worldwide institutional, cultural, and ideological frameworks and systems. Yet all of these frameworks were multicentered and heterogeneous, each generating its own dynamics.

Of special importance in this context was the relative place of the non-Western societies in the various – economic, political, ideological – international systems which differed greatly from that of the Western ones. It was not only that it was Western societies which were the 'originators' of this new civilization. Beyond this and above all was the fact that the expansion of these systems, especially insofar as it took place through colonialization and imperialist expansion, gave to the Western institutions the hegemonic place in these systems. Yet it was in the nature of these international systems that they generated a dynamics which gave rise both to political and ideological challenges to existing hegemonies, as well as to continual shifts in the loci of hegemony within Europe, from Europe to the United States, then also to Japan and East Asia.

But it was not only the economic, military-political, and ideological expansion of modernity from the West throughout the world that was important in this process. Of no lesser significance was the fact that this expansion has given rise to continual confrontation between the cultural and institutional premises of Western modernity and those of other civilizations. Thus, while the spread or expansion of modernity has indeed taken place throughout most of the world, it did not give rise to just one civilization, one pattern of ideological and institutional response, but to at least several basic variants – and to continual refracting thereof.

Consequently, multiple modernities have emerged. These civilizations, which share many common components and which continually constitute mutual reference points, have been continually unfolding, giving rise to new problematiques and reinterpretations of the basic premises of modernity. All these attest to the growing diversification of the visions and understanding of modernity, of the basic cultural agendas of different sectors of modern societies – far beyond the hegemonic vision of modernity that was prevalent before. The fundamentalist – and the new communal-national – movements constitute one of such new developments, in the unfolding of the potentialities and antinomies of modernity.

Such developments may indeed also give rise to highly confrontational stances – especially to the West – but these stances are promulgated in changing modern idioms, and they may entail a continual transformation of the cultural programs of modernity. At the same time the new diversity was closely connected – perhaps paradoxically – with the development of new multiple common reference points, and with a globalization of cultural networks and channels of communication far beyond what existed before.

See also: Civilization, Concept and History of; Cultural History; Cultural Psychology; Elias, Norbert (1897–1990); Elites: Sociological Aspects; Hegemony and Cultural Resistance; National Character; Societies, Types of.

Bibliography

Breuer, B., 1994. Kulturen der Achsenzeit. Leistung und Grenzen eines geschichtsphilosophischen Konzepts. Saeculu 45, 1–33.
Durkheim, E., Mauss, M., 1971. Note on the notion of civilization. Social Research 38, 808–813.
Elias, E., 1939. Über den Prozess der Zivilisation. Haus zum Falken, Basel.
Eisenstadt, S.N., 1973. Tradition, Change and Modernity. John Wiley & Sons, New York.
Eisenstadt, S.N., 1982. The axial age: the emergence of transcendental visions and rise of clerics. European Sociology 23 (2), 294–314.
Eisenstadt, S.N. (Ed.), 1986. The Origins and Diversity of Axial Age Civilizations. State University of New York Press, Albany, NY.
Eisenstadt, S.N., 1982. Heterodoxies and dynamics of civilizations. Diogene 120, 3–25.
Eisenstadt, S.N., Achlama, R., 1992. Kulturen der Achsenzeit II, ihre institutionelle und kulturelle Dynamik. 3 vols. Suhrkamp, Frankfurt am Main.
Eisenstadt, S.N., 1996. Japanese Civilization: A Comparative View. University of Chicago Press, Chicago.
Eisenstadt, S.N., 1999. Fundamentalism, Sectarianism and Revolution. The Jacobin Dimension of Modernity. Cambridge University Press, Cambridge, UK.
Huntington, S.P., 1996. The Clash of Civilizations andthe Remaking of World Order. Simon and Schuster, New York.
Kroeber, A.L., Kluckhohn, C., 1952. Culture: A Critical Review of Concepts and Definitions. The Museum, Cambridge, MA.
MacIver, R.M., 1931. Society: Its Structure and Changes. R. Long & R. R. Smith, Inc, New York.
Melko, M., 1969. The Nature of Civilizations. Porter Sargent, Boston.
Nelson, B., 1981. On the Roads to Modernity. Rowman and Littlefield, Totowa, NY.
Ogburn, W.F., 1922. Social Change: With Respect to Culture and Original Nature. Huebsch, New York.
Schluchter, W., 1979. Die Entwicklung des okzidentalen Rationalis- mus—Eine Analyse von Max Webers Gesellschaftsgeschichte. Mohr, Tübingen.
Schluchter, W., 1981 (1985). The Rise of Western Rationalism, Max Weber's Developmental History. University of California Press, Berkeley.
Schluchter, W., 1989. Rationalism, Religion and Domination. A Weberian Perspective. University of California Press, Berkeley.
Weber, A., 1921. Prinzipielles zur Kultursoziologie. Archiv für Sozialwissenschaft und Sozialpoliti XL (VII), 1–49. J-C-B. Mohr, Tübingen.
Weber, A., 1931. Kultursoziologie. In: Enke, F. (Ed.), Handwörterbuch der Soziologie. Alfred Vierkandt, Stuttgart, pp. 284–294.
Weber, M., 1922–3. Gesammelte Aufsätze zur Religionssoziologie. Mohr, Tübingen.

Clan

Philip Burnham, University College of London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The definition of a clan as a group of persons who believe themselves to be related by unilineal descent but who are unable to trace genealogical connections linking all members of the group emerged out of a complex intellectual history. Nineteenth-century social evolutionists disagreed whether early stages of society were patrilineal or matrilineal but generally accepted that clans were a typical form of primitive social structure. Clanship was also seen as linked with totemism, as an early form of religion, although Lévi-Strauss later rejected any necessary association between clanship, religious belief in animal or plant ancestors, and group exogamy. Radcliffe-Brown, Fortes, and Evans-Pritchard analyzed clanship as one level within a system of segmentary lineages, but there is analytical value in maintaining a distinction between lineages, with their comprehensive genealogical structures, and clans. The diverse forms of clanship that occur in many parts of the world can be understood in relation to several key variables. Of prime importance is the extent of their capacity for organized collective action, as well as their social networking function for a dispersed clan's members. Many cultures' understandings of clanship, which are often rooted in myth, posit a primordial clan identity that is seen as immutable, much like ethnicity.

The modern definition of the clan, now widely accepted, as a group of persons who believe themselves to be related by unilineal descent but who are unable to trace genealogical connections linking all members of the group, has emerged out of a complex intellectual history stretching back to the Enlightenment. Many social theorists of the eighteenth and nineteenth centuries who were concerned to understand the origins of the state and democracy favored evolutionist explanations Kuper (1988). Some viewed the family as the original form of society which, over the course of human history, had aggregated into progressively larger and more complex kinship-based groupings of clans and tribes, which eventually were transformed into the territorially based political formation of the state. Drawing on the work of Barthold Niebuhr (1828), George Grote (1851), and other historians of Greece and Rome, social theorists such as Sir Henry Maine (1861), Numa Fustel de Coulanges (1876), and Lewis Henry Morgan (1877) used the ancient Greek and Roman kinship groups, the *genos* and the *gens*, as models on which they based their understanding of early stages of social evolution.

Maine's (1861) analysis of the evolution of legal systems, using the example of the Roman *gens* and the concept of agnation, that is, kinship traced through exclusively male, patrilineal descent links, posited a primitive stage of social development in which membership in patriarchal kinship groups defined a person's social status. Maine emphasized that the *gens* was a corporate group which had a legal personality that endured beyond the lives of its individual members, an insight which has been very influential in subsequent work on clans and other forms of descent grouping.

In contrast to Maine's theory of primitive patriarchy, John Ferguson McLennan (1865) and Lewis Henry Morgan (1877) argued that the early stage of social evolution was characterized by group marriage, which meant that paternity was uncertain and kinship was traced in the maternal line. For some authors favoring the theory of primitive matriliny, the term 'clan' was applied exclusively to matrilineal descent groups while *gens* referred to patrilineal groups, a terminological distinction that has fallen out of use in the modern literature. Morgan, whose own research among the Iroquois in New York State had been influenced by Grote's analysis of the ancient Greek *genos*, saw a descent group organization, which he termed the 'gentile' system, as characteristic of much of early human social history. McLennan, basing his reasoning on contemporary reports of female infanticide in India, claimed that primitive societies were obliged, in their struggles for survival, to engage in this population limiting practice, which in turn led kin groups to procure wives by capture from outside their own group. McLennan termed this marriage outside the group 'exogamy,' which came to be seen, by many nineteenth and early twentieth century social theorists, as another characteristic attribute of clanship.

McLennan was also influential in developing the theory of totemism, which built on E.B. Tylor's notion that primitive religion was based on animism, the worship of inanimate objects or fetishes which were believed to be the abodes of spirits. Along with his follower W. Robertson Smith and Smith's student Sir James Fraser, McLennan defined totemism as a system which originally combined animistic religious beliefs with matrilineal descent and exogamy. According to this view, through their ignorance of paternity caused by group marriage and their belief in animism, primitive societies posited links to original animal or plant ancestors. These natural species, which were treated as sacred emblems, or 'totems,' of different clans, were periodically worshipped by clan members in totemic rituals. Emile Durkheim (1912), in his work on the social origins of religion which drew extensively on early ethnographic reports of aboriginal Australian societies, reproduced many of these classic arguments concerning totemism, emphasizing particularly the concept of a sacred social solidarity based on a belief in a shared substance between clan members and their ancestral totemic species. Claude Lévi-Strauss's (1962) subsequent general critique of theories of totemism, which emphasized the lack of a necessary coincidence between totemic species as emblems, exogamous clans, and religious sacrifice, served to unhitch the clanship concept from the pseudohistorical

hypotheses of nineteenth century evolutionism while, at the same time, helping to clarify the classificatory logic operative in cultural categories such as the clan.

The work of the British structural-functionalist A.R. Radcliffe-Brown (1950) and his followers E.E. Evans-Pritchard, Meyer Fortes (Fortes and Evans-Pritchard, 1940), and others between the two world wars, did much to put the study of unilineal descent systems on a firmer ethnographic footing, while refining the conceptual framework for their study. The model of the segmentary lineage, consisting of a nested set of increasingly inclusive corporate unilineal descent groups linked by a comprehensive genealogical structure, is the most well-known outgrowth of their work, with the term 'clan' often being used by these authors to label a particular segmentary level within a lineage system. Although Radcliffe-Brown's (1950: pp. 39–40) definition of the clan, given in the first sentence of this article, has come to be widely used, this application of the term to characterize one of the levels in a segmentary lineage system is not always appropriate and reflects a continuing influence of earlier understandings of totemism, in which descent group exogamy and totemic taboos were taken to be defining attributes of clanship.

Although a definitional contrast between clan and lineage is thus still not always clearly made in the more modern literature, there is analytical value in maintaining a clear distinction since the organizational implications of the two forms of descent grouping are different. Members of a lineage know or claim to know the genealogical connections interlinking all members of the group, and these links, viewed in terms of generation and relative birth order, often provide a basis for calculating relations of seniority or degrees of relatedness between individuals and segments within a lineage. Lineage genealogy can also be the basis for internal segmentation of groups, according to the segmentary lineage model. In contrast, lacking such a comprehensive internal genealogical armature, relations of clanship are categorical in character and typically nonhierarchical within the clan. Persons are members of a clan because of being the offspring either of their fathers or their mothers, with the terms 'patriclan' and 'matriclan' often being used to indicate a clan's mode of recruitment of its members. Although ascending descent links beyond the grandparental or great-grandparental generation are normally of little or no organizational significance, clan members in many societies do nonetheless recognize a founding clan ancestor, who is often of mythical or nonhuman status. Thus, when clans do segment according to putative genealogical links, this is typically 'from the top' with reference to the founding ancestor and his or her children, producing a set of subclan categories that also lack comprehensive internal genealogical structures.

Terminological confusion is also possible in the case of the descent group known as the conical clan, or ranked lineage, which combines characteristic features of both lineages and clans. Such descent groups are internally differentiated into a high ranked, lineage-like, chiefly or noble descent line, and a lower ranked and internally undifferentiated clan-like category of commoners (Kirchhoff, 1959; Friedman, 1979). Chiefly rank in a conical clan is typically based on relative birth order in present and ancestral generations, so that senior sons or daughters of senior ranking ancestors keep careful track of their pedigrees to validate their noble status. Junior offspring of junior ancestors, on the other hand, have little motivation to remember their genealogies, and their affiliation to the group is more categorical, or clan-like, in character. Numerous ethnographers have drawn attention to the structural potential of conical clanship, particularly as a transitional formation between uncentralized, kin-based political systems and centralized chiefdoms or states. Conical clan structures are relatively common in Polynesian societies, such as the Maori, and in southeast and central Asian societies, such as the Mongols. The clans of the Scottish Highlands also displayed such internal ranking (Dodgshon, 1998) during their hey-day from the sixteenth to eighteenth centuries. The Scottish clan chief and his close relatives enjoyed high status due to their close patrilineal connection to the founding clan ancestor. The commoner members of the clan, on the other hand, did not necessarily have blood ties to their chiefs but were bound to their clan by diverse social ties including feudal land tenancy or relations of marriage, as well as real or fictive kinship expressed by a common clan surname.

The term 'clan' is derived from the Gaelic word *clann* meaning 'children of' and came into medieval English usage to describe the kin-based character of early Irish and Scottish societies. But as explained above, members of Scottish clans were not necessarily all blood kin, despite a strong ideology of shared kinship. Recent research by Basu (2005) and others on modern Scottish heritage tourism and the numerous Scottish clan societies in present-day North America has emphasized the inventedness of much of the modern understanding of the Scottish clan, which imagines a clan's members as sharing blood ties, an ancestral territory, a tartan, and other symbols of clan unity.

Systems of clanship occur in many parts of the world and their diversity of form can be understood in relation to several key variables. Of prime importance is their often weak capacity for organized collective action as corporate groups, which is frequently limited or absent due to the spatial dispersal of clan memberships. Particularly in territorially extensive societies with large populations, a clan's members usually reside in many different locations, and there are limited possibilities for the full membership of a clan to meet as a total group. In such circumstances, clans may act collectively on only a few occasions, such as during annual rituals like the aboriginal Australian *corroboree* ceremonies, which clans celebrate to insure the continued fertility of their totemic species. In other cases of spatially dispersed clan membership, a clan may never meet as a totality but simply act as a named category from which localized sets of clan members may be mobilized to pursue collective action as necessary or advantageous.

In addition to their variable capacities for organizing group action, numerous authors have also drawn attention to the potential of clanship systems to serve a social networking function. For example, in West Africa among speakers of the Mande family of languages (Jackson, 1974), a limited set of clan names is found throughout this large zone, and a clan member making a long journey can feel confident of finding members of his or her clan who will provide assistance at the distant destination. Even in cases where the set of clan names changes from one society to an adjacent one, conventional equivalences are often established between clan names to

facilitate the extension of such networks of mutual aid. Similar clan networks have also been noted among North American Amerindian and aboriginal Australian societies.

Although analysts of clanship have frequently conflated it with lineage structure, the logic of clanship is often more a matter of 'sub-ethnicity' rather than 'super-lineage.' Many cultures' understandings of clanship, which are often rooted in myth, posit a primordial clan identity which is seen as immutable, much like many cultures' conceptualizations of ethnicity. It is this theme that Lévi-Strauss (1962) explored in his work on totemism, in which the classificatory logic of clans as social species is seen as sharing the same pattern as cultural classifications of animal and plant species. In such cases, the total number of clans in a society may be quite small and be viewed as unchanging, with the various clanship categories standing in long-term interrelationships of marriage alliance, ritual cooperation, or other modes of mutual exchange or solidarity. In societies where the total number of clan categories is only two, these groupings are conventionally referred to as 'moieties.'

In contrast, some societies have systems of proliferating clans in which it is evident to the analyst, from the presence of large numbers of clans with relatively small memberships in the context of overall population growth, that clans are undergoing segmentation, although this is often not acknowledged by clan members themselves. Such systems of proliferating clanship tend to be found in politically uncentralized societies with high rates of spatial mobility, where local groups frequently split as a mode of dispute settlement. In such cases, it may be posited that, following an episode of intra-clan dispute which has been resolved by group fission and spatial displacement, the members of the resultant new segments have an active interest in denying knowledge of former ancestral interlinkage. Such systems of proliferating clanship, characterized by many small clans that are widely dispersed spatially, are not conducive to the maintenance of the patterns of long-term inter-clan alliance mentioned above.

In some cases, the term 'clan' has also been applied to territorial groups which are recruited both on the basis of unilineal descent and long-term coresidence. Such a usage is common in New Guinea. Here, in-migrant strangers hailing from other clans may initially be welcomed in order to strengthen a localized clan grouping. Over time, cultural theories of personal identity may posit, for example, that through continued consumption of foodstuffs cultivated on the clan's land, the physical nature of such incomers becomes transformed into that of true clansmen, thus converting coresidence in a single territory into shared descent. In effect, as argued by the American kinship specialist George P. Murdock (1949), such a 'compromise kin group' reflects the difficulty of constituting a large scale, residentially unified, and collectively functional kinship grouping on the basis of unilineal descent alone.

As explained above, the concept of clanship is based on a categorical understanding of social grouping which classifies the members of a society into mutually exclusive units, or 'types,' of people. Recent studies such as Rupp's (2011) of the Bangando of the Cameroonian rainforest in Central Africa show how this classificatory logic may be employed by members of a society to conceptualize their relations with the wider world. Thus, while the Bangando themselves are divided into diverse clans (People of the Elephant, People of the Monitor Lizard, People of the Monkey, etc.), their clanship terminology is extended today to encompass new categories of individuals the Bangando now have contact with, such as wildlife conservationists (People of the World Wide Fund for Nature (WWF)), loggers (People of the Timber Camp), or soccer teams (People of the Ball).

Finally, it should be recognized that, in common parlance today, the term 'clan' is often used in a metaphorical way to refer to any group of persons who act toward each other in a particularly close and mutually supportive way. Thus, criminal organizations such as the Mafia may be referred to as 'clans,' in recognition of the ideology of kin-like solidarity that binds their members together, while in post-socialist Eastern European countries, the term is often applied to networks of politically and economically influential persons.

Bibliography

Basu, P., 2005. Macpherson country: genealogical identities, spatial histories and the Scottish diasporic clanscape. Cultural Geographies 12, 123–150.

Dodgshon, R., 1998. From Chiefs to Landlords. Edinburgh University Press, Edinburgh.

Durkheim, E., 1912. Les formes élémentaires de la vie religieuse. Alcan, Paris. (1915. The elementary forms of the religious life. Allen & Unwin, London.)

Fortes, M., Evans-Pritchard, E.E. (Eds.), 1940. African Political Systems. Oxford University Press for the International African Institute, London.

Friedman, J., 1979. System, Structure and Contradiction in the Evolution of 'Asiatic' Social Formations. National Museum of Denmark, Copenhagen.

Fustel de Coulanges, N., 1876. La Cité Antique. Hachette, Paris. (1980. The ancient city. Johns Hopkins University Press, Baltimore.)

Grote, G., 1851. A History of Greece, third ed. John Murray, London.

Jackson, M., 1974. The structure and significance of Kuranko clanship. Africa 44 (4), 397–415.

Kirchhoff, P., 1959. The principles of clanship in human society. In: Fried, M. (Ed.), Readings in Anthropology, vol. II. Crowell, New York.

Kuper, A., 1988. The Invention of Primitive Society. Routledge, London.

Lévi-Strauss, C., 1962. Le totemisme aujourd'hui. Presses Universitaires de France, Paris. (1969. Totemism. Penguin, Harmondsworth, UK.)

Maine, H., 1861. Ancient Law. John Murray, London.

McLennan, J.F., 1865. Primitive Marriage. Adam & Charles Black, Edinburgh.

Morgan, L.H., 1877. Ancient Society. Henry Holt, New York.

Murdock, G.P., 1949. Social Structure. Macmillan, New York.

Niebuhr, B., 1828. The History of Rome. Taylor, Cambridge.

Radcliffe-Brown, A.R., 1950. Introduction. In: Radcliffe-Brown, A.R., Forde, D. (Eds.), African Systems of Kinship and Marriage. Oxford University Press for the International African Institute, London.

Rupp, S., 2011. Forests of Belonging: Identities, Ethnicities and Stereotypes in the Congo River Basin. University of Washington Press, Seattle.

Class Actions, Legal

Deborah R Hensler, Stanford Law School, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G.C. Hazard, Jr., volume 3, pp. 1924–1927, © 2001, Elsevier Ltd.

Abstract

At one time legal scholars and practitioners viewed the class action as a quintessentially American civil procedure. Now class actions have spread around the world, to at least two dozen jurisdictions on every continent but Antarctica. All class action procedures share some common features, most importantly that they allow one or more people or entities to represent a much larger group of similarly situated claimants who are not before the court. Specific provisions vary, however. The goals of class actions include efficient case management, access to justice and regulatory enforcement, but priorities assigned to these goals vary across jurisdictions, and are reflected in the design of procedures. Class actions have been used in the United States and elsewhere to seek a wide variety of monetary and nonmonetary remedies for claims ranging from human rights abuse, to environmental disasters to financial wrongdoing. The extensive literature on class actions is largely theoretical but the field has recently attracted the attention of empirical scholars.

A class action is a legal procedure that enables one or a few people or entities to pursue a claim in court on behalf of themselves and other similarly situated claimants who are not themselves before the court. What distinguishes class actions from other forms of collective litigation is its representative character. Because the procedure may extinguish as well as secure the rights of absent parties, its adoption and implementation typically evoke controversy. Although not always articulated, some of the opposition to class actions stems from the understanding that permitting collective litigation has the potential to increase the number of certain types of lawsuits that are too expensive for plaintiffs to litigate on an individual basis. Most of the scholarly literature on class actions addresses normative or technical legal issues but there is a growing empirical literature on economic and socio-legal aspects of class action litigation.

History

The notion that a group of people united by family ties, geography, socio-economic status or other social identity characteristics might have a collective interest in pursuing a claim against a public authority or private person appears to have deep historical roots. Traditional societies' 'blood feuds' are a violent manifestation of a social norm legitimating collective redress for harms inflicted by one member of a clan on another clan's member (Otterbein and Otterbein, 1965). Yeazell (1987) reports multiple instances of group litigation against feudal institutions over property rights in medieval England. Although group litigation gradually died out in England, the tradition lived on in the American colonies.

Early American courts permitted representative group litigation when numerous parties shared a common legal interest. When the U.S. federal court rules were substantially revised in 1938 the revised rules included an updated procedure for what were by then called 'class actions.' But the 1938 class action rule was technically demanding, leading to post-World War II calls to revise it, particularly to facilitate lawsuits seeking the end of racial discrimination. The current form of federal Rule 23 was adopted in 1966 (Hensler et al., 2000). Since that time there have been frequent calls for amendment to address one or more perceived problems with class action practice. Amendments in 2003 left the structure of rule intact, but clarified the timing of class certification and the judge's responsibilities regarding appointment of class counsel, settlement approval and attorney fee awards. All but one U.S. state have class action rules modeled on federal Rule 23.

Until the early 1990s, there were no other countries that permitted collective litigation along the lines of the American class action. Over the next 20 years the global landscape of class actions shifted dramatically, and class actions spread first to Australia and Canada, then to Brazil, and subsequently to Northern and Western Europe, the Middle East and Asia. Today more than two dozen countries have adopted some sort of collective litigation procedure; similar procedures are under debate in other countries (Hensler et al., 2009).

Goals

Efficient case management is the first goal of most class action laws. When large numbers of claims arise out of the same factual and legal circumstances, the result often is to flood one or a few courts with caseloads far beyond their capacity to resolve expeditiously. By allowing a court to decide common issues in a single representative action, the class action eliminates the need to decide essentially the same issue over and over again in individual litigation, saving expensive judicial time. For the same reason, class actions can also save money for a defendant who is the target of hundreds or more almost identical legal claims. Collecting claims in a single action also facilitates settlement, a goal of most courts and many parties. Class settlements are attractive to defendants who prefer capping their liability exposure and avoiding protracted individualized litigation.

A second goal of some class action laws is to increase access to justice for individuals who cannot easily afford legal representation unless they come together as a group to pursue their legal rights. Since 1966, the class action has been the primary

vehicle for bringing claims against the United States federal and state governments charging violations of constitutional and statutory rights (Chayes, 1976; Fiss, 2009). Canadians have also used the class action procedure to pursue rights claims (Kodar, 2012). Whether social impact litigation via class actions and other procedures is effective in achieving social policy goals is sharply contested (Bell, 2004: Llewellyn, 2002; Lobel, 2003: Rosenberg, 2008).

A different sort of access to justice argument is made in favor of individuals and small businesses who have lost small amounts of money as a result of alleged legal violations by defendants who stood to make large sums of money as a result. Proponents of class actions argue that in these circumstances collective proceedings are the only practical way for injured parties to obtain compensation for their losses. This goal of class actions has been dubbed 'collective redress' by European policy-makers and policy advocates (European Parliament, 2012).

The wisdom of litigating over small monetary claims strikes some policy-makers and citizens as questionable, however. Critics of 'small value' class actions find it particularly troubling that the attorneys who represent class members in such suits earn far more in fees than any individual class member receives in compensation for his or her loss. They label these suits 'negative value class actions.' The fact that in some of these cases the class action attorney wins fees that are equal to or are larger than the *total* sum paid to class members makes this use of class actions seem particularly egregious to class action critics. Some class action proponents argue to the contrary, emphasizing that the small value class action is an important tool for regulatory enforcement. According to this 'private attorney general' theory, class actions make an important contribution to deterrence by forcing defendants to pay substantial penalties for illegal behavior (Coffee, 1986; Rubenstein, 2004; Lehav, 2011). From the perspective of law and economics, the division of the defendant's payment between class counsel and class members is immaterial: the total payment constitutes the sanction. However, this argument is often lost on politicians and the general public, and private regulatory enforcement is the most controversial goal of class actions in every jurisdiction that has debated whether to adopt such a procedure.

Description of Class Action Procedures

The common features of contemporary class action procedures are that (1) they allow one or more people or entities to bring a lawsuit on behalf of a class of people or entities that share legal and factual claims arising out of the same event or practice, and (2) the outcomes of the lawsuit will bind the members of the class. The key distinction between class actions and other forms of collective action is that, save for the class representative(s), class members are not before the court. Other forms of collective litigation generally involve aggregating similar claims, each of which has been filed in court.

Generally, jurisdictions only permit representative class actions when the class members are so numerous that it is impractical for all to appear in court ('numerosity') and when the claims share common legal and factual issues ('commonality'). Jurisdictions define numerosity and commonality differently. Some jurisdictions, including the United States, require an onerous certification process to assure that only those civil actions that satisfy the relevant criteria are allowed to proceed in class form; in other jurisdictions, such as Australia, this process is less demanding (Grave et al., 2012). Other key features vary from country to country. The most important of these features are:

- *Scope* (For what sorts of substantive claims may class actions be used?)
- *Standing* (Who is permitted to come forward on behalf of the class?)
- *Remedies* (What will class members obtain if their suit is successful?) and
- *Whether* eligible class members need to 'opt in' proactively if they want to join the class, or need only 'opt out' proactively if they wish to exclude themselves.

U.S. procedural rules are *trans*-substantive, meaning class actions may be filed for a wide variety of constitutional, statutory and common law claims. Most other jurisdictions limit the use of class actions to one or a few domains, such as securities, consumer or anti-trust violations. U.S. class action rules permit any person or organization (e.g., a small business) to represent a class of similarly situated claimants, subject only to a judge's finding that the proposed representative's interests are 'typical' of other class members and that the individual or organization will 'adequately represent' the class members. In many other jurisdictions, only preexisting organizations such as consumer or investor associations or specially authorized nongovernmental organizations or public officials may sue on behalf of a class. In the United States, class members may obtain declaratory or injunctive relief (i.e., a judicial declaration as to the constitutionality or legality of a policy or practice or an order to end or adopt a practice) or monetary compensation. Outside the United States, class members generally are limited to nonmonetary remedies in all or most circumstances. In the United States, in class actions for money damages all people or entities that meet the class definition (e.g., all investors who purchased a certain stock during a specified time period) are deemed to be class members unless they proactively exclude themselves – i.e., 'opt out.' Most other jurisdictions have adopted 'opt in' rules instead (Hensler et al., 2009).

The consequences of these variations in class action features are profound. By permitting any typical class member with adequate resources to represent the class, allowing class members to seek monetary as well as nonmonetary remedies for a wide variety of alleged harms, and not requiring eligible class members to pro-actively declare their interest in joining the litigation, U.S. class action rules facilitate broad use of class action procedures for diverse purposes. By limiting scope, standing and remedies and requiring that class members 'opt in,' other jurisdictions limit the use of class actions (Hensler, forthcoming).

Class action procedural rules operate within distinctive legal regimes, with different substantive laws. Some statutes specify that individuals may use a class action to claim remedies for violation of rights granted by the statute; other statutes specifically *prohibit* use of a class action procedure to claim remedies for statutory violations, forcing those who wish to

litigate to file individual claims. Judicial interpretations of the relationship between substantive laws and class action procedural rules also affect whether and how frequently class actions are filed. For example, if each investor who claims to have lost money in trading on the stock market as a result of a corporation's failure to properly disclose relevant information has to prove individually that she relied on whatever information was disclosed, then investors are unlikely to be able to proceed in class form. Securities class actions in the United States became more viable in 1988 when the U.S. Supreme Court held in *Basic v. Levinson* that instead of each class member needing to prove individual reliance, the class as a whole needed only to show that the corporation perpetrated a 'fraud on the market' that affected stock prices. But U.S. consumer class actions became less viable in 2011 when the Supreme Court held in *AT&T Mobility v. Concepcion* that under the Federal Arbitration Act of 1925, an arbitration provision in a consumer contract could preclude consumers from proceeding in class form both in court and in private arbitration.

The substantive legal rules that most affect the use of class action procedures, however, are those that pertain to litigation financing. Every jurisdiction has rules about how to apportion the cost of litigation that were fashioned for individual litigation. In the United States, each party bears its own costs, win or lose, except in certain circumstances specified by statute. In most other countries, the losing party is required to pay the prevailing party's expenses. To assure that the defendant's costs will be covered if it prevails, the plaintiff may be required to post a bond at the time she files her lawsuit. Usually the actual amount paid by the loser is determined by a judge or other court official, sometimes according to a published cost schedule. In the United States, a lawyer is allowed to represent a party on a contingent or speculative basis, meaning that the lawyer is only paid if her client prevails. Typically, in individual litigation, the plaintiff lawyer enters into a contingent fee contract with her client, specifying that her client will pay her a share of the damages (e.g., one-third) if the client prevails and nothing if the client does not prevail (Kritzer, 2004). Defense lawyers are most often paid on an hours-and-expense basis, although sometimes defendants enter into contingent contracts. In most other jurisdictions lawyers are not allowed to enter into contingent fee contracts in which their pay will depend on the amount of money the plaintiff wins, although some jurisdictions do allow attorneys to represent plaintiffs on a conditional ('no win, no pay') basis (Emons and Garoupa, 2006), and some parties pay 'success fees' to their lawyers.

In the United States, special rules on selecting and paying lawyers have evolved for class actions. These rules incentivize lawyers to bring class actions and ensure that all class members contribute to the cost of successful class litigation. Except in securities class actions that are governed by a special statute, the judge appoints the class counsel and if the class prevails at trial or negotiates a settlement with the defendant, the judge decides how much the lawyer will be paid. (U.S. rules for securities class actions are governed by the *Private Securities Litigation Reform Act of 1995*. The legislation directs the judge to appoint as class representative the investor with the greatest stake in the litigation; that investor is then free to select the lawyer for the class.) Often, the judge specifies an amount that is a fraction of the total amount the class obtained (Eisenberg and Miller, 2010). The result is that all of the class members contribute to paying the attorneys' fees.

In most other jurisdictions, the same legal financing rules that apply to individual litigation apply to class actions. These rules make it much less attractive, compared to the United States, for individuals to come forward to represent a class and for lawyers to agree to serve as class counsel. The class representative contracts with the lawyer to represent the class, is subject to adverse costs if the class does not prevail, and may have to post bond when the class action is filed. None of the other class members share these costs or risks. If conditional fees are permitted, the lawyer will fund the litigation while it is ongoing; otherwise, the class representative will have to pay legal fees and expenses until the case is resolved. The successful class counsel who accepted the case on a conditional basis may or may not receive any risk premium for doing so.

To address financial constraints on class actions (and other complex commercial litigation), an increasing number of jurisdictions outside the United States have endorsed the introduction of third-party litigation financing into class action practice (Kalajdzic et al., 2013). Litigation financers view litigation as an investment opportunity. They contract with class members to take on the risk of adverse costs and pay the class counsel to maintain the litigation; in return, if the class prevails, the financers – typically publicly traded companies, hedge funds or other private investment companies – take a share of each class member's compensation. The introduction of third-party litigation financing facilitates class actions with good potential for large monetary damages but does nothing to encourage social impact litigation seeking declaratory or injunctive relief.

Uses

Most of what we know about how class actions are used in different jurisdictions derives from reported judicial decisions and commercial media. No jurisdiction maintains a comprehensive database of class action filings (as distinct from nonclass civil filings) and no jurisdiction tabulates the number of class actions filed by substantive case type. Not all judicial decisions relating to class actions are published. Mass media focus on cases arising out of vivid circumstances, such as environmental disasters and financial scandals. As a result we know about examples of the use of the class action procedure, but not how representative these examples are.

Social Impact Litigation against Governments

In the United States class actions have been used to shape public policies governing schools (e.g., racial segregation, school financing), prisons (e.g., race, ethnic and religious discrimination, health services, crowding), social safety net programs (e.g., aid to women and children), taxation and a host of other issues (Hensler et al., 2000; Burke, 2003; Farhang, 2011). In Canada class actions have been used to shape policies regarding the treatment of indigenous people (Kodar, 2012) and same sex couples. However, outside the U.S.

and Canada, this use of class actions appears to be rare. In Israel the class action statute limits class actions against the state (Magen and Segal, 2009) and in China social activists who have attempted to bring class actions that implicate government policies have been jailed.

Social Policy Litigation against Industry

American lawyers have long used class actions against individual businesses to shape public policies governing the workplace (e.g., racial, ethnic, gender, age and disability discrimination, sexual harassment). More recently, social activists in many jurisdictions ave teamed with trial lawyers and government officials to bring lawsuits against the tobacco, gun, alcohol and food industries seeking changes in production and marketing practices as well as monetary compensation. These suits combine 'cause lawyering' strategies (Scheingold, 2004; Sarat and Scheingold, 2006) with entrepreneurial litigation tactics. Activists argue that legislative efforts to regulate behaviors that are hazardous to public health and safety have been stymied by corporate lobbyists. They see litigation as a potentially effective tool of overcoming rent-seeking behavior. The resulting litigation has taken diverse forms, including (in the United States) coordinated lawsuits brought by state attorneys general on behalf of their citizens (sometimes termed *parens patriae* suits), single lawsuits brought by local government officials, individual tort suits seeking compensation for wrongful death and personal injury and private class actions. The multi-billion settlement with the tobacco industry in the United States was the result of coordinated state attorneys general actions, not class actions (Mollenkamp et al., 1998; Zegart, 2000). However, in Brazil, Bulgaria, Canada, Israel and Spain class actions have been filed against the tobacco industry. U.S. litigation against the gun industry has been a mix of public and private individual and class action lawsuits (Lytton, 2009). A large fraction of these social policy lawsuits have failed. However, by calling attention to industry practices that activists allege are harmful to the public, they may have long-term effects on market behavior and public policy.

Mass Accidents and Environmental Disasters

In the United States, the judiciary has long thought that the large numbers of personal injury and property damage claims that often result from mass catastrophes are not suitable for class treatment because of the likelihood of significant differences among class members (Hensler et al., 2000). Notwithstanding this general policy, U.S. federal judges have certified money damage class actions in some high profile environmental catastrophes, such as the *Exxon Valdez* tanker oil spill in Alaska and the British Petroleum *Deepwater Horizon* explosion and subsequent oil spill in the Gulf of Mexico. In some other jurisdictions, class action procedures have been adopted specifically to address environmental contamination or are provided for in environmental protection statutes. For example, Indonesia adopted a class action procedure in response to an environmental nongovernmental organization's legal activism (Santosa, 2009). A massive collective litigation on behalf of Ecuadorian indigenous peoples seeking compensation for environmental contamination from Chevron proceeded in Ecuadorian courts under an environmental protection statute (Gomez, 2013).

Product Safety and Marketing

Although mass product defect claims have been a feature of United States civil litigation for several decades, consistent with the general policy disfavoring class certification of personal injury and property damage claims, few of these mass tort suits have been certified as class actions. An attempt to certify a massive asbestos injury class action was overturned by the U.S. Supreme Court in 1997 (Hensler, 2005–06) and more recent attempts to obtain class certification of pharmaceutical product liability claims have also failed (Silver and Miller, 2010). The Netherlands collective settlement program, however, was adopted specifically to enable a mass settlement of personal injury claims associated with the pharmaceutical product DES (Tzankova, 2012). Chinese activists have attempted to use collective litigation to obtain compensation for injuries associated with contaminated milk and other products but were prevented by the national government from doing so (Liu, 2009).

Economic Transactions

Worldwide the most common use of class action procedures appears to be with regard to market activities. Hensler et al. (2000) estimate that securities and consumer class actions accounted for more than half of all U.S. class actions in the mid-1990s, the most recent period for which these data are known. About half of all Australian federal class actions filed since 2004 pertained to securities, investment advice, and consumer transactions (Morabito, 2009). Although a mass product defect litigation was the inspiration for the adoption of the Netherland's collective settlement act, the five class actions that have followed since all pertained to securities or investment products (Hensler, 2011).

Reparations

Human rights activists in several jurisdictions have attempted to use the class action procedure to obtain compensation from governments and private actors. Although all of these suits seek money damages, those who bring or promote them often are seeking public admission and recognition of harms they or their ancestors suffered as a result of the defendants' denial of their fundamental rights. African-American activists have sought, so far unsuccessfully, to obtain reparations for the enslavement of their ancestors by private parties acting under public law (Fortson, 2004). African-American farmers and Native Americans have been more successful in obtaining compensation from the U.S. federal government for discrimination in payment of agricultural benefits and management and distribution of trust funds (Fortson, 2004). Holocaust survivors and in some instances their descendents have obtained compensation for bank policies that denied children and other relatives access to accounts that had been set up by victims of Nazi atrocities and for industrial slave labor (Authers and Wolffe, 2002: Bazyler, 2003; Neuborne, 2002: Schapiro, 2003). Canadians have received compensation for abusive

practices in residential schools to which the government consigned indigenous children. In class and non-class litigation, victims of sexual abuse by Catholic priests have obtained compensation from Catholic churches in the United States and Canada, in amounts that in some instances resulted in diocesan bankruptcies. Some class action suits brought in U.S. courts against multi-national corporations alleging they were complicit in killing, torture and other abuse of workers in third-world countries have been successful, although others were not. In all these instances, although the direct objective of the class litigation was monetary damages, an equally important – if not more important – goal was bringing culpable parties to account. Because of the media attention they receive, class actions of this sort may achieve some measure of success in meeting their social justice goals even when they fail to obtain monetary compensation for victims (Lobel, 2003).

Scholarship

Perhaps not surprising given that the class action is a creature of statutory law and legal procedural rules, doctrinal analyses dominate the scholarly literature. In addition, there is a vast theoretical literature on class actions. Early work was authored almost exclusively by Americans but today the subject is attracting scholars from other parts of the world (Klemont and Procccia, 2013). Law and economics scholars, who dominate the theoretical literature, focus on the agency problems posed by representative class litigation, particularly in small value class actions where the class counsel's interests outweigh the interests of any individual class member (Coffee, 1983, 1985; Hay, 1997; Macey and Miller, 1991). Legal scholars writing from a legal ethics perspective have also focused on these issues, in some instances drawing attention to what the authors view as particularly egregious examples of a mismatch between the class counsel's and class members' interests (Koniak, 1995; Koniak and Cohen, 1996; Wasserman, 2000). With a few exceptions (Silver and Miller, 2010; Erichson, 2000, 2003) these scholars have ignored similar agency problems posed by nonclass mass litigation in which large numbers of individual claims are aggregated for pretrial development and resolution.

Empirical scholarship on class actions has emerged more recently. The U.S. class action rule requires the presiding judge to approve any class settlement negotiated by the class counsel and defendant and also to award fees and expenses to the prevailing class counsel. Federal judicial decisions including these amounts are published and similar state judicial decisions are increasingly available. As a result there is more information available on class action settlements than for the settlements of ordinary civil litigation, which are regarded as private in the United States as elsewhere. Law and economics scholars have assembled datasets containing this outcome information and analyzed the relationship between settlement amounts and attorney fee awards, by type of case (Eisenberg and Miller, 2010; Fitzpatrick, 2010). The consistent finding for these U.S. cases is that the percentage of settlements awarded to attorneys as fees is negatively correlated with the amount of the settlement. Most settlements offer cash payments to class members, sometimes in addition to other remedies. Mean and median settlement values vary substantially across case types, with commercial matters such as securities and anti-trust class actions yielding the highest average values and labor and employment and civil rights cases yielding the lowest average values (Fitzpatrick, 2010). (In the United States, unlike many other jurisdictions, federal and state courts of first instance have jurisdiction over labor and employment suits.)

Although quantitative analyses of published judicial opinions approving class action settlements shed empirical light on one of the key class action controversies, the relationship between settlement amounts and attorney fees awards, they leave many other important questions unanswered. For example, information on disposition patterns that might indicate the proportion of cases deemed nonmeritorious by judges is scanty. Information on the relationship between negotiated settlement amounts and aggregate losses class members claim is virtually nonexist. Nor is there much information on what proportion of negotiated and judicially approved settlement funds is actually paid by defendants to class members. Because of the lack of comprehensive reliable data on dispositions and pay-outs, Hensler (forthcoming) concludes that it is impossible currently to assess whether class actions generally play a beneficial regulatory enforcement role.

A second line of empirical scholarship on class actions is more qualitative, comprising case studies of small and nonrandom samples of class action lawsuits (Hensler et al., 2000). Because of the strong doctrinal orientation of legal scholarship in most non-US jurisdictions, scholars outside the U.S. are just beginning to conduct these sorts of socio-legal studies of collective litigation (Hensler et al., forthcoming; Liu, 2009; Llewellyn, 2002; Voet, 2013). As a result, the extant literature largely focuses on U.S. litigation.

Qualitative case studies provide thick descriptions of class action litigation processes, highlighting the tensions within a legal regime that incentivizes private lawyers to bring litigation in pursuit of public goals such as structural reform and regulatory enforcement. Scholars as well as journalists have written book length case studies of selected high-profile class action lawsuits that provide vividly detailed stories of how such litigation enfolds. These tend to focus more on the heroic dimension of cause lawyering but also provide insights into the principal-agent problems inherent in representative litigation that have been the focus of more theoretically oriented law and economics scholars (Roberts and White, 1999; Bingham and Gansler, 2003; Schuck, 1986; Zegart, 2000; Schapiro, 2003; Kluger, 2011). To date, neither the quantitative nor qualitative scholarship has been sufficient to test fully the theoretical propositions that have preoccupied class action legal scholars since the advent of the modern U.S. class action rule.

See also: Civil Law; Courts and Adjudication; Disputes, Social Construction and Transformation of; Justice, Access to; Litigation.

Bibliography

AT&T Mobility v. Concepcion, 131 S. Ct. 1740 (2011).

Authers, John, Wolffe, Richard, 2002. The Victims's Fortune; Inside the Epic Battle over the Debts of the Holocaust.

Bingham, Clara, Gansler, Laura, 2003. Class Action: the Landmark Case that Changed Sexual Harassment Law.

Basic v. Levinson, 485 U.S. 224 (1988).

Bazyler, Michael, 2003. Holocaust Justice: The Battle for Restitution in America's Courts.

Bell, Derrick, 2004. Silent Covenants: Brown V. Board of Education and the Unfilled Hopes for Racial Reform.

Burke, Thomas, 2003. Lawyers, Lawsuits and Legal Rights: The Battle over Litigation in American Society.

Chayes, Abram, 1976. The role of the judge in public law litigation. Harvard Law Review 89, 1281–1316.

Coffee, John, 1986. Understanding the plaintiff attorney: the implications of economic theory for private enforcement of law through class and derivative actions. Columbia Law Review 86, 669–727.

Coffee, John, 1985. The unfaithful champion: the plaintiff as monitor in shareholder litigation. Law & Contemporary Problems 48, 5–82.

Coffee, John, 1983. Rescuing the private attorney general: why the model of the lawyer as Bounty Hunter is not working. Maryland Law Review 82, 215–288.

Emons, Winand, Garoupa, Nuno, 2006. US-style contingent fees and UK-style conditional fees: agency problems and the supply of legal services. Managerial and Decision Economics 27, 379–386.

Eisenberg, Theodore, Miller, Geoffrey, 2010. Attorneys' fees and expenses in class action settlements: 1993–2008. Journal of Empirical Legal Studies 7, 248–281.

Erichson, Howard, 2000. Informal aggregation: procedural and ethical implications of coordination among counsel in related suits. Duke Law Journal 50, 381–471.

Erichson, Howard, 2003. Beyond the class action: lawyer loyalty and client autonomy in non-class collective representation. University of Chicago Legal Forum 2003, 519–575.

European Parliament, February 2, 2012. Toward a Coherent European Approach to Collective Redress, 2011/2089 INI.

Farhang, Sean, 2011. The Litigation State: Public Regulation and Private Lawsuits in the United States.

Fiss, Owen, 2009. The history of an idea. Fordham Law Review 78, 101–108.

Fitzpatrick, Brian, 2010. An empirical study of class action settlements and their fee awards. Journal of Empirical Legal Studies 7, 811–846.

Fortson, Ryan, 2004. Correcting the harms of slavery: collective liability, the limited prospects of success for a class action suit for slavery reparations and the reconceptualization of white racial identify. Berkeley Journal of African-American Law and Policy 6, 71–127.

Gomez, Manuel, 2013. The global chase: seeking the enforcement of the Lago Agrio judgment outside ecuador. Stanford Journal of Complex Litigation 1 (Spring), 429–466.

Grave, Damian, Adams, Ken, Betts, Jason, 2012. Class Actions in Australia, second ed.

Hay, Bruce, 1997. Asymmetric rewards: why class actions (May) settle for too little. Hastings Law Journal 48, 479–508.

Hensler, Deborah, Pace, Nicholas, Dombey-Moore, Bonita, Giddens, Beth, Gross, Jennifer, Moller, Erik, 2000. Class Action Dilemmas: Pursuing Public Goals for Private Gain.

Hensler, Deborah, 2005–06. Asbestos litigation in the United States: triumph and failure of the civil justice system. Connecticut Insurance Law Journal 12, 255–279.

Hensler, Deborah, 2011. The future of mass litigation: global class actions and third-party financing. George Washington Law Review 79, 306–323.

Hensler, Deborah. Can private class actions enforce market regulations? Do they? Should they. In: Francesca, Bigsami, David, Waring. Handbook of Comparative Law and Regulation. Elgar, forthcoming.

Hensler, Deborah, Hodges, Christopher, Tulibacka, Magdalena, March 2009. The globalization of class actions. Annals of the American Academy of Political and Social Science 622.

Hensler, Deborah, Christopher, Hodges, Ianika, Tzankova. Class Actions In Context: How Culture, Economics and Politics Shape Collective Litigation, forthcoming.

Kalajdzic, Jasminka, Cashman, Peter, Longmoore, Alana, 2013. Justice for profit: a comparative analysis of Australian, Canadian and U.S. third party litigation funding. American Journal of Comparative Law 61, 93–147.

Klement, Alon, Procccia, Yuval, December 2013. An Economic Analysis of Reliance in Market Fraud and Negligent Misrepresentation. Available at http://ssrn.com/abstract=2372922.

Kluger, Richard, 2011. Simple Justice: The History of Brown V. Board of Education and Black America's Struggle for Equality, revised ed.

Kodar, Freya, 2012. The Woodland school litigation and settlement: further complications for historical abuse claims. University of British Columbia Law Review 45, 393–405.

Koniak, Susan, 1995. Feasting while the widow weeps: Georgene v. Amchem products. Cornell Law Review 80, 1045–1158.

Koniak, Susan, Cohen, George, 1996. Under cloak of settlement. Virginia Law Review 82, 1051–1280.

Kritzer, Herbert, 2004. Risks, Reputations and Rewards: Contingency Fee Legal Practice in the United States.

Lehav, Alexandra, 2011. Two view of the class action. Fordham Law Review 79, 1939–1963.

Liu, Chenlin, 2009. Profits above the law: China's melamine tainted milk incident. Mississippi Law Journal 79, 371–417.

Llewellyn, Jennifer, 2002. Dealing with the Legacy of Native Residential School Abuse in Canada: Litigation, ADR and Restorative Justice, vol. 52. University of Toronto Law Review, 253–300.

Lobel, Jules, 2003. Success without Victory: Lost Legal Battles and the Long Road to Justice in America.

Lytton, Timothy, 2009. Suing the Gun Industry: A Battle at the Crossroads of Gun Control and Mass Torts.

Macey, Jonathan, Miller, Geoffrey, 1991. The plaintiffs' attorney role in class action and derivative litigation: economic analysis and recommendations for reform. University of Chicago Law Review 58, 1–118.

Magen, Amichai, Segal, Peretz, March 2009. Israel. In: Hensler, Deborah, Hodges, Christopher, Tulibacka, Magdalena (Eds.), The Globalization of Class Actions, Annals of the American Academy of Political and Social Science, vol. 622, pp. 244–253.

Mollenkamp, Carrick, Levy, Adam, Menn, Joseph, Rothfelder, Jeffrey, 1998. The People VS. Big Tobacco: How the States Took on the Cigarette Industry.

Morabito, Vince, 2009. An Empirical Study of Australia's Class Action Regimes: First Report: Class Action Facts and Figures.

Neuborne, Burt, 2002. Preliminary reflections on aspects of the Holocaust-era litigation in American courts. Washington University Law Review 80, 795–834.

Otterbein, Keith, Otterbein, Charlotte, December 1965. American Anthropologist 67, 1480–1482.

Roberts, Bari-Ellen, White, Jack, 1999. Roberts VS. Texaco: A True Story of Race and Corporate America.

Rosenberg, Gerald, 2008. The Hollow Hope: Can Courts Bring about Social Change?, second ed.

Rubenstein, William, 2004. On what a 'private attorney general' is – and why it matters. Vanderbilt Law Review 57, 2129–2173.

Santosa, Mas Achmad, March 2009. Indonesia. In: Hensler, Deborah, Hodges, Christopher, Tulibacka, Magdalena (Eds.), The Globalization of Class Actions, Annals of the American Academy of Political and Social Science, vol. 622, pp. 310–317.

Schapiro, Jane, 2003. Inside a Class Action: The Holocaust and the Swiss Banks.

Scheingold, Stuart, 2004. The Politics of Rights: Lawyers, Public Policy and Political Change.

Schuck, Peter, 1986. Agent Orange on Trial: Mass Toxic Disasters in the Courts.

Silver, Charles, Miller, Geoffrey, 2010. The quasi-class action approach of managing multi-district litigation: problems and a proposal. Vanderbilt Law Review 63, 107–177.

Sarat, Austin, Scheingold, Stuart, 2006. Cause Lawyers and Social Movements.

Tzankova, Ianika, 2012. Funding of mass disputes: lessons from the Netherlands. Journal of Law, Economics and Policy 8, 549–591.

Voet, Stefaan, 2013. Cultural dimensions of group litigation: the Belgian case. Georgia Journal of International and Comparative Law 41, 433–479.

Wasserman, Rhonda, 2000. Dueling class actions. Boston University Law Review 80, 461–542.

Yeazell, Stephen, 1987. From Medieval Group Litigation to the Modern Class Action.

Zegart, Dan, 2000. Civil Warriers: The Legal Siege on the Tobacco Industry.

Class Consciousness

Jesus de Felipe-Redondo, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by P. Ansart, volume 3, pp. 1931–1933, © 2001, Elsevier Ltd.

Abstract

'Class consciousness' has been employed in two different but related ways: (1) as the awareness that the members of a class (generally the working class) have of their 'objective' class interests and aspirations, and (2) as a sense of class belonging that adopts different shapes in specific times and places. Both concepts, especially the first one, were fundamental categories of analysis and debate in Marxist and academic studies on class. The evolution of the different explanations they generated mirrors the theoretical changes in sociological and historical theory since the end of the nineteenth century.

Class Consciousness

The term 'class consciousness' represents two different, but related, concepts. The first one refers to the awareness that the members of a social class, usually the working class, have of their common class interests (what is called the 'principle of class identity'). These shared interests are considered to be opposed to those of other classes, especially the bourgeois class (the 'principle of class opposition'). Awareness of this antagonism makes workers fight to end class conflict by transforming social organization. This final goal leads to identification of class consciousness with socialist (particularly Marxist) aspirations that provide a general vision of a classless society (the 'principle of totality').

This first concept has a theoretical character. It plays a crucial role in socialist, especially Marxist, explanations of human action and historical change. Indeed, Marxist thinkers coined and developed the term at the end of the nineteenth century. Conversely, the second meaning is eminently descriptive. It refers to a set of political attitudes, cultural traditions, and moral values shared by the members of a class. Here 'consciousness' does not necessarily imply the awareness of underlying, 'objective' class interests, but rather the specific and contingent ways in which individuals perceive their class position. For this reason, this second concept is employed as synonym of 'class identity,' 'class subjectivity,' or 'class mentality' and, unlike the first concept, it does not articulate a general theory about human action – although it constitutes an object of sociological and historical analysis.

Despite their differences, the two senses are related. The first concept has guided socialist debates since the nineteenth century and has had a significant impact on academic research, especially from the 1940s to the 1980s. It became inseparable from the question of class, and thus it was at the center of the debates on class belonging and human action in sociology and history. By the same token, its gradual disappearance since the 1980s was a consequence of the crisis of class theories and objectivist epistemology. The second meaning spread in part as a response to this crisis, though its current limited use is also a result of the ongoing effects of that crisis (for a different analysis of these two meanings see Wright, 1985).

Marxist Origins

The first concept of class consciousness emerged within the materialist epistemological framework in which most socialist theories were rooted in the second half of the nineteenth century and has been crucial to the development of Marxist philosophy and political theory since then. However, interest in the workers' class conscience was not exclusively Marxist. Many non-Marxist socialists focused on the process through which workers 'became aware' of their supposed class interests. Pierre–Joseph Proudhon, for instance, analyzed "how the working class had distinguished itself from the bourgeois class since 1789 and how it has became conscious of it" (Proudhon, 1865). Karl Marx, on the other hand, did not use the term 'class consciousness,' although his reflection on the relationship between class position and consciousness set the foundation for the concept's elaboration at the turn of the century.

The concept of class consciousness was a consequence of the way in which Marxist thinkers faced a crucial problem: the lack of correspondence between their predictions about workers' conduct and workers' real actions, that is to say, between Marxist theory and empirical observation. Marxist predictions were based on a specific conception of social organization and its effects on individuals' conduct. According to Marx's classic formulation, the implementation of capitalist social relations divided individuals into two classes: the exploitative and parasitical class, represented by the bourgeoisie, and the exploited and dispossessed class, identified with the working class. Since the very existence of the former was based on the systematic exploitation of the later, Marx deduced that capitalist social organization created classes with irreconcilable social interests that spurred workers to fight against the bourgeoisie. Thus, class 'objective' interests rooted in relations of production shaped individuals' consciousness and, therefore, their actions. As Marx (1904) put it in 1859:

> In the social production of their existence, men inevitably enter into definite relations, which are independent of their will, namely relations of production […]. The totality of these relations of production constitutes the economic structure of society, the real foundation on which arises a legal and political superstructure and to which correspond definite forms of social consciousness. The mode of production of material life conditions the general process of social,

> political and intellectual life. It is not the consciousness of men that determines their existence, but, on the contrary, their *social existence determines their consciousness*.

This is the classical materialist-objectivist assumption that led Marxists to explain workers' consciousness and conduct as being determined by workers' position in productive relations. However, there was a critical flaw in this explanation: the majority of workers did not participate in labor struggles, and those who did were more interested in the improvement of particular labor conditions than in changing social organization. Marx was aware of this problem, but this did not lead him to reconsider the aforementioned materialist-objectivist assumption. Rather, this assumption provided him with an explanation of the workers' conduct. He pointed out that workers' perceptions of themselves and their interests were not necessarily a faithful representation of their 'real being' or 'objective class interests.' Marx (Marx and Engels, 1956) wrote in 1844 that:

> The question is not what this or that proletarian, or even the whole of the proletariat at the moment considers its aim. The question is what the proletariat is, and what, consequent on that being, it will be compelled to do. Its aim and historical action is irrevocably and obviously demonstrated in its own life situation as well as in the whole organization of bourgeois society today.

From this perspective, class position and class consciousness were two different, although related, things. This led Marx to differentiate between 'class against capital' and 'class for itself,' a distinction later interpreted by Marxist intellectuals as the dichotomy between 'class in itself' – the class that emerged in the economic structure – and 'class for itself' – the workers' awareness of their class position and interests (Andrew, 1983). At the turn of the century, class for itself was usually referred to as 'class consciousness.'

Marx argued that class position eventually produced class consciousness – in fact, he considered this process as historically inevitable. However, such an emergence was neither automatic nor immediate, but the product of a gradual process through which workers gradually became aware of their alleged class interests. Marx and later Marxists defined two different sets of factors that hindered this process. First, the slow implementation of capitalism and class relations that delayed the emergence of class in the economic structure in certain countries, for class consciousness could not exist without class in itself. Second, capitalism produced its own 'ideological' justification that misled workers regarding their class interests. Marx mentioned organized religion and nationalism as examples of these ideologies, but for him the most prominent was 'commodity fetishism.' Commodity fetishism masked exploitation by making workers perceive productive relations as an exchange of commodities. Marx used the term 'alienation' to express the consequences of this phenomenon: commodity fetishism 'alienated' workers from the act of production and product of their work, from their own nature as human beings and from other workers, preventing them from acquiring class consciousness. A decade after Marx's death, Friedrich Engels (1893) coined the concept 'false consciousness' to refer to the distorting effect of capitalist ideologies on workers' consciousness – a concept that Georg Lukács later explored and developed.

Thus, concepts such as 'ideology,' 'alienation,' and 'false consciousness' accompanied the formulation of class consciousness as category and object of analysis. Indeed, all of them were inextricably linked because the study of class consciousness was the analysis of its absence. Marxist thinkers assumed that class consciousness did not exist in the present, but would appear in the near future. Meanwhile, workers had an alienated, false consciousness. Therefore, the notion of class consciousness foreshadowed the concepts that referred to its very absence.

The explanation of this absence spurred the development of Marxist analysis and debates. The longer that absence endured, the more sophisticated the thesis of false consciousness. In 1844, Marx was confident about the spontaneous character of the emergence of class for itself. He argued that capitalist development and class struggle were already making workers progressively aware of their class interests: "a large part of the English and French proletariat is already conscious of its historic task and is constantly working to develop that consciousness into complete clarity" (Marx and Engels, 1956). At the end of the century, however, Marxist thinkers were more skeptical. In Karl Kautsky's words, the 'error' of Marx and Engels was that they had 'expected revolution too soon' (Kautsky, 1909). Kautsky, Georgi Plekhanov, and other Marxists observed that, despite capitalist development, the growth of labor conflicts, and intensive socialist propaganda, most unionized workers did not follow their supposed class interests. Moreover, many of those who participated in the growing socialist unions and parties disagreed with Marxist tenets, and even many of those supposedly 'conscious workers' who controlled socialist organizations and proclaimed revolutionary objectives acted in reformist ways.

It was then that class consciousness became a pivotal object of study for Marxists, who went deeper than Marx in exploring the factors that prevented workers from becoming class conscious. Marxist intellectuals identified capitalist ideologies as the main factor that hindered the emergence of class consciousness. As Vladimir I. Lenin wrote in 1902, "The working class spontaneously gravitates towards socialism, but the more widespread (and continuously revived in the most diverse forms) bourgeois ideology spontaneously imposes itself upon the working class still more" (Lenin, 1987). Marxist thinkers began to examine and elaborate Marx' and especially Engels' previous ideas about ideological distortion of workers' consciousness. Lenin's development of Engels' thesis of 'labor aristocracy' and 'embourgeoisement' was one of the results of the new reflections. Marxist intellectuals also focused on questions such as the class consciousness of peasants and why socialism – considered the highest expression of class consciousness – had not developed or had done it weakly and differently in each national case. This was the origin of the question of historical exceptionalism that later spread across academic labor studies (the paradigmatic case is Sombart, 1976).

This process of theoretical development brought new and contradictory analytical perspectives on the very nature of class

consciousness that implied antagonistic political strategies. Eduard Bernstein argued that since the majority of workers from the most developed capitalist countries supported social reforms, working-class consciousness was not revolutionary, but reformist. This particular conception of class consciousness became the basis for Marxist 'revisionism' and social democracy during the twentieth century (Crompton, 1993). In rejecting Bernstein, Rosa Luxemburg stated that political and labor struggles would lead workers to spontaneously overcome reformism and develop a revolutionary consciousness. Lenin, in turn, argued that there were different 'levels of class consciousness' and that organized workers could only spontaneously develop one of them, the limited and reformist 'trade-union consciousness.' He maintained that there was a superior level of consciousness, namely 'socialist' or 'revolutionary' consciousness, that workers could only acquire through the guidance of Marxist intellectuals (the socialist vanguard or intelligentsia). This conception of class consciousness became the theoretical pillar for radical Marxist left and communism in the twentieth century (Mayer, 1997).

In this debate, the concept of class consciousness underwent two significant transformations. First, its meaning changed according to the new approaches that divided Marxism. From Marx's perspective, there was only one kind of consciousness that reflected workers' 'real' interests. At the turn of the century, Marxist intellectuals talked of several kinds of class consciousness (reformist, revolutionary, socialist, communist, etc.). This implied that workers could develop different forms of representation of their class experiences and aspirations and that it was necessary to study how these forms of consciousness changed (or could be changed) into others. In a certain way, this idea was one of the roots of the second, descriptive meaning of class consciousness since it entailed the idea that the same class position could be represented in dissimilar ways (through different kinds of consciousness) that produced specific types of conduct.

Second, the Leninist perspective brought the paradoxical idea of 'party vanguard,' that is, a group mainly composed of middle-class individuals that became class conscious before the mass of workers (Bottomore, 1971). This idea led some Marxist intellectuals to argue that 'revolutionary'/'socialist' class consciousness was not a mere product (and could appear independently) of the grade of development of class in itself due to the influence of certain cultural, political, and ideological factors. From this perspective, the relationship between class belonging and class consciousness was more complex than previously thought. Consciousness itself was an autonomous factor that interacted with class position in a dialectical way.

The Russian Revolution in 1917 had a decisive impact on this change of perspective. The triumph of a socialist revolution in a country with a weak industrial working class led Marxist intellectuals such as Lukács, Karl Korsch, and Antonio Gramsci to revise crucial aspects of previous theories of class consciousness in the 1920 and 1930s. These authors gave greater importance to the ideological, political, and cultural factors – including the Marxist theoretical tradition itself – that shaped class consciousness. Furthermore, they described the relationship between consciousness and class position in dialectical terms, thus nuancing explanations based on economic determinism (Goldman, 1971). Lukács (1971), the most important analyst of the question of class consciousness, used the concept of 'reification' to explain the strong influence of bourgeois ideologies on workers' minds and distinguished between 'psychological' and 'ascribed'/'imputed' consciousness. Drawing from a Leninist perspective, he analyzed how the socialist intelligentsia played a crucial role in transforming workers' psychological (false) consciousness into class (ascribed) consciousness. Lukács thought that subjective factors contributed to this process, including Marxism itself. He pointed out that while workers experienced capitalist relationships in their workplace, it was the mediation of Marxist theory that provided them with an explanation of these experiences of exploitation and a solution to them. Thus, Marxism was an active, mediating force that contributed to creating, rather than to merely representing, class consciousness – a thesis ultimately rejected by Soviet Marxists (Eyerman, 1981; Mészáros, 1971).

Lukács, like other Marxists, emphasized the historical and necessary transition from false consciousness to class consciousness (Wright, 1985). Gramsci's view represented a crucial difference on this point. From Gramsci's perspective, historical necessity and class economic determinism had less importance, while cultural factors were crucial for the formation of class consciousness. For Gramsci, class consciousness was not a mere subjective representation of the workers' class position, but a cultural product of workers' reflections on their labor situations (Eyerman, 1981). In other words, if workers became agents of historical change it was not because class position determined them to be so, but because of the way they made sense of that position. Workers were agents of change not because they were proletarian, but because they thought of themselves as proletarian:

> One has only to think of the prominent use that Socialists make of the terms 'consciousness', 'class consciousness', and 'socialist and proletarian consciousness'; implicit in this language is the philosophical conception that one 'is' when one 'knows', or 'has consciousness of', his own being. A worker 'is' proletarian when he 'knows' his truth and works and thinks according to this 'knowing'. (Gramsci, 1984)

Gramsci thought of class consciousness as largely a cultural phenomenon derived both from the workers' experiences, and common practices and reflections about their class position through the cultural traditions and concepts that constituted their perspectives of the world. Workers' consciousness was composed of several cultural and ideological elements, some of which were shared by the dominant classes. Gramsci argued that some ideological trends aspired to organize these reflections and perceptions, creating a prevailing 'common sense' that determined workers' interpretations and actions. The crucial goal for Marxist organizations was precisely to build a 'cultural hegemony' through these elements, fighting against the hegemonic aspirations of other trends and institutions – such as Catholicism and the Church in the Italian case (Eyerman, 1981).

These dialectical perspectives on class consciousness had a limited influence in the Marxist debate. Mainstream explanations assumed the Leninist thesis that (socialist/revolutionary)

class consciousness mirrored class position, and that when this did not occur it was because of capitalist ideological distortion. However, Lukács' and Gramsci's contributions shared an important commonality with the Leninist approach: all of them introduced or further elaborated concepts that referred to the absence of class consciousness (labor aristocracy, levels of class consciousness, reification, false consciousness, hegemony, etc.). Thus, in the 1930s studies on class consciousness had developed a complex and refined set of analytical tools and explanatory theses in order to explain the nonexistence of such consciousness and its possible modes of development.

Academic Debates

Marxism's growing influence in the universities from the 1930s onward contributed to the introduction of the concept of class consciousness into historical and sociological academic studies. Class consciousness became one of the central categories in the analysis of the relationship between social structure and human action. From the 1940s through the 1980s, it was particularly used in studies on social stratification and the history of the labor movement (Crompton, 1993). These analyses also introduced the problem of the absence of class consciousness and spurred debates on questions such as development of class consciousness among white collar workers, peasants, and the bourgeoisie. The problem of the lack of class consciousness had been particularly evident since the 1950s, when several sociological studies showed that most workers in several western countries pragmatically accepted the capitalist order. These studies pointed out that although many workers presented themselves as members of the working class in social or cultural terms, class consciousness was quite limited. Moreover, any radical ideas held were closer to left-populist ideologies than to socialism. Some authors even stated that working-class culture and workers' enthusiasm for collective action were fading away due to the impact of mass society and commercial capitalism (Savage, 2000).

However, since Marxist researchers assumed that there was a direct relationship between class position and consciousness, they kept looking for factors that promoted or hindered development of class consciousness. As sociologist David Lockwood (1958) argued in his study of the blackcoated workers, "To study the class consciousness of the clerk is to study the factors affecting his sense of identification with, or alienation from, the working class." In dealing with this question, researchers incorporated new perspectives and concepts such as the Weberian notion of 'status.' Thus, in trying to explain the British workers' gradual withdrawal of support to the Labor Party in the 1960s, Lockwood, John H. Goldthorpe, and other British Marxist sociologists argued that different perceptions of labor status divided the English working class and thus prevented the rise of class consciousness (Goldthorpe et al., 1968).

The political and social context of the 1960 and 1970s, with the crisis of the labor movement and the rise of the new social movements, made researchers reexamine the relationship between class position and consciousness and consider the impact of cultural, political, and subjective factors on workers' class consciousness. Some authors distinguished between class consciousness and other types of workers' consciousness and studied the mutual influences between them. A famous 1966 article by Lockwood initiated a debate on the distinction between class consciousness and other forms of 'class imagery' (Bulmer, 1975). Alain Touraine (1968) differentiated between 'workers' consciousness' and 'class consciousness' and argued that the latter was not a consequence of class position but the result of a specific and rarely found combination of the principles of class identity, opposition, and totality. Michael Mann (1973) added a fourth principle, the 'vision of an alternative society without classes.' He stated that each of the four principles could exist without any of the others and that their multiple combinations explained the diversity of workers' kinds of consciousness. Anthony Giddens (1973) separated 'class awareness' from 'class consciousness' and 'conflict consciousness' from 'class consciousness,' and focused on the factors that determined the transformation of one form of consciousness into another.

The new perspectives assumed that certain ideological forms previously considered 'false' were active forces in molding workers' subjectivities. Eventually, this idea led some researchers to distance themselves from the old dichotomy of class/false consciousness. According to Tom Bottomore (1971), social researchers had "passed out of the era in which the real consciousness of social groups, expressed in their beliefs and actions, could be discussed as mere 'psychological', 'false' consciousness, and be contrasted with the 'rational consciousness' enshrined in the ideology of a Communist party." In the 1960s and 1970s, many authors criticized economic determinism and leaned toward the study of the dialectic influences between class position and consciousness. In so doing, they explored the arguments developed by Lukács and Gramsci – as well as Marx' recently discovered Grundrisse. Class consciousness began to be perceived as a complex form of subjectivity that emerged in relation to several cultural and political factors and had a dynamic and relational character derived from its relationship to labor experiences and conflicts. Some authors began to argue that class was not a 'thing' (a static object of analysis), but a relational, historical process of creation of a collective political subject, which was mirrored in the ongoing formation of class consciousness. These authors felt increasingly uncomfortable with conventional quantitative methodology and defended qualitative approaches (Savage, 2000, 2005; Fantasia, 1995).

A new conception of class consciousness that entailed a greater attention to the hybridization between class position and cultural and political factors progressively arose from these changes. The most influential author in this 'cultural turn' was the historian Edward P. Thompson (1968). In 1963, Thompson argued that economic conditions did not have a direct impact on workers' minds. Rather, they were 'experienced' in specific ways by workers, who interpreted these experiences through the available cultural and political traditions. Class consciousness was the outcome of this interpretation.

> Class consciousness is the way in which these experiences are handled in cultural terms: embodied in traditions, value systems, ideas, and institutional forms. If the experience appears as

> determined, class consciousness does not. We can see a logic in the responses of similar occupational groups undergoing similar experiences, but we cannot predicate any law. Consciousness of class arises in the same way in different times and places, but never in just the same way. (Thompson, 1968)

From this viewpoint, class consciousness was a historical and relational process in which cultural and economic factors intertwined in complex ways. Class, as subject of historical change, appeared only when workers developed class consciousness, and thus the distinction between class in itself/for itself was called into question. Thompson stated that political languages and cultural traditions molded class consciousness into specific cultural shapes. This meant that it was not possible to distinguish between 'pure' and 'spurious,' 'truer' and 'falser' forms of class consciousness. Consciousness was neither false nor true, but simply what it was (Thompson, 1977). There was not a single model of class consciousness, but different variations depending on cultural and historical circumstances. This entailed the transformation of the problem of exceptionalism in the history of labor movement into the analysis of regional 'variations' of class consciousness and formation (Katznelson and Zolberg, 1986).

The abandonment of the class/false consciousness dichotomy meant a break with previous perspectives. Thompson's work inspired a new generation of historians and sociologists in the 1970s and 1980s (Savage, 2000). These researchers defined and incorporated into their analysis a growing set of cultural factors and sources of identity, including race and gender (Fantasia, 1995). However, this theoretical change eventually put the concept of class consciousness into question. In the 1980s and 1990s, new studies showed the greater importance of non-class identities in explaining workers' actions. In what is considered a general crisis of the class and objectivist paradigm, these studies not only called the category of class consciousness into question, but also criticized the validity of the objectivist epistemic framework in which that category was inscribed. Thus, many authors stated that class consciousness guided analysis toward economic determinism and class teleology and essentialism, including the case of 'culturalist' perspectives based on the Thompsonian notion of experience (Crompton, 1993; Joyce, 1994). The concept of class consciousness was considered to be a key element in what Margaret Somers has called 'epistemology of (class) absence' (Somers, 1997), which prevented researchers from developing deeper analyses of workers' actual identities.

As a result of this criticism, many researchers stopped using the category of class consciousness in the 1980s and 1990s. Class consciousness was replaced by other concepts that mirrored the emergence of new objects of analysis and perspectives on the study of workers' subjectivity with less obvious objectivist connotations, such as 'class identity,' 'class subjectivity,' 'class mentality,' 'class culture,' or 'class dispositions.' This change 'from consciousness to identity' reflected the new orientation of culturalist perspectives that distanced themselves from traditional class issues – although class essentialism and teleology kept guiding the use of these new concepts in many studies (Fantasia, 1995; Savage, 2000, 2005; Bottero, 2004; Somers, 1997).

This does not mean that the term 'class consciousness' has completely disappeared from the academic literature. It is occasionally used in some studies, although it has lost its previous theoretical connotations and assumed a more descriptive sense, referring to the different cultural shapes adopted by the 'sense of belonging to a group.' This is the second concept of class consciousness mentioned in the introduction. Although this second sense had been frequently mixed with the first concept, its relevance grew in the 1990s accompanying the crisis of the latter. Moreover, in this process the term 'consciousness' was gradually identified with that of 'identity,' that is, a set of cultural, moral, and political principles of a particular group of workers that is related to a more general vision of the world (Wright, 1985). This explains the current use of the concept in recent works. According to Mike Savage, "Studying class consciousness (today) involves going beyond studying how someone identifies as a member of a class, and examines the extent to which a person's view about class forms part of a coherent social outlook which can be said to be consistent and organized in terms of class" (Savage, 2000: p. 37). However, since the term 'class consciousness' is still strongly related to the objectivist perspective, some researchers prefer to add adjectives that emphasize the culturally constructed character of such consciousness, creating new terms such as 'discursive class consciousness' (Sayer, 2005).

See also: Action, Theories of Social; Alienation, Sociology of; Class: History of the Concept; Class: Social; Communism, History of; Ideology, Sociology of; Industrialization; Labor History; Marxism in Contemporary Sociology; Marxism–Leninism: The Ideology of Twentieth-Century Communism; Objectivity: Philosophical Aspects; Reification: History of the Concept; Social Identity in Sociology; Social Inequality in History (Stratification and Classes); Social Movements, History of: General; Social Movements: A Social Psychological Perspective; Social Stratification; Socialism: Historical Aspects; Socialism; Working Classes, History of.

Bibliography

Andrew, E., 1983. Class itself and class against capital: Karl Marx and his classifiers. Canadian Journal of Political Sciences 16 (3), 577–584.

Bottero, W., 2004. Class identities and the identity of class. Sociology 38 (5), 985–1003.

Bottomore, T., 1971. Class structure and social consciousness. In: Mézáros, I. (Ed.), Aspects of History and Class Consciousness. Routledge & Kegan Paul, London, pp. 49–76.

Bulmer, M. (Ed.), 1975. Working-Class Images of Society. Routledge & Kegan Paul, London.

Crompton, R., 1993. Class and Stratification: An Introduction to Current Debates. Polity Press, Cambridge.

Engels, F., 1893. Letter to Franz Mehring in Berlin (London, 14 July). In: Marx, K., Engels, F. (Eds.), 1942. Selected Correspondence (D. Torr, Trans.). International Publishers, New York, p. 511.

Eyerman, R., 1981. False Consciousness and Ideology in Marxist Theory. Almqvist & Wiksell International, Stockholm.

Fantasia, R., 1995. From class consciousness to culture, action, and social organization. Annual Review of Sociology 21, 269–287.

Giddens, A., 1973. The Class Structure of the Advanced Societies. Hutchinson, London.

Goldman, L., 1971. Reflections on history and class consciousness. In: Mézáros, I. (Ed.), Aspects of History and Class Consciousness. Routledge & Kegan Paul, London, pp. 65–84.

Goldthorpe, J.H., Lockwood, D., Bechhoofer, F., Platt, J., 1968. The Affluent Worker: Political Attitudes and Behaviour. Cambridge University Press, Cambridge.

Gramsci, A., 1984. Misteri della cultura de della poesia. October 19, 1918. Quoted in: Germino, D., 1990. Antonio Gramsci: Architect of a New Politics. Louisiana State University Press, Baton Rouge, p. 80.

Joyce, P., 1994. Introduction. In: Democratic Subjects. The Self and the Social in Nineteenth-Century England. Cambridge University Press, Cambridge.

Katznelson, I., Zolberg, A., 1986. Working Class Formation: Nineteenth-Century Patterns in Western Europe and the United States. Princeton University Press, Princeton.

Kautsky, K., 1909. The Road to Power. Samuel A. Bloch, Chicago.

Lenin, V.I., 1987. What is to be done? In: Lenin, V.I., Christman, H.M. (Eds.), Essential Works of Lenin. Dover, New York.

Lockwood, D., 1958. The Blackcoated Worker: A Study in Class Consciousness. George Allen & Unwin Ldt., London.

Lukács, G., 1971. History and Class Consciousness: Studies in Marxist Dialectics. Merlin Press, Cambridge, MA.

Mann, M., 1973. Consciousness and Action Among the Western Working Class. MacMillan, London.

Marx, K., 1904. A Contribution to the Critique of Political Economy. Trans. from the second German edition. International Library Publishing Co., Chicago.

Marx, K., Engels, F., 1956. The Holy Family or Critique of Critical Critique. Foreign Languages Publishing House, Moscow.

Mayer, R., 1997. Plekhanov, Lenin and working-class consciousness. Studies in East European Thought 49, 159–187.

Mészáros, I., 1971. Contingent and necessary class consciousness. In: Mézáros, I. (Ed.), Aspects of History and Class Consciousness. Routledge & Kegan Paul, London, pp. 85–127.

Proudhon, P.J., 1865. De la capacité politique des classes ouvrières. E. Dentu Libraire-Éditeur, Paris.

Savage, M., 2000. The limits to class consciousness. In: Class Analysis and Social Transformation. Open University Press, Buckingham, pp. 23–42.

Savage, M., 2005. Working-class identities in the 1960s: revisiting the affluent worker study. Sociology 39 (5), 929–946.

Sayer, A., 2005. The Moral Significance of Class. Cambridge University Press, Cambridge.

Sombart, W., 1976. Why Is There No Socialism in the United States? Shape, New York.

Somers, M., 1997. Deconstructing and reconstruction class formation theory: narrativity, relational analysis, and social theory. In: Hall, J.R. (Ed.), Reworking Class. Cornell University Press, Ithaca, pp. 72–105.

Thompson, E.P., 1968. The Making of the English Working Class. Penguin, Aylesbury.

Thompson, E.P., 1977. Alcune osservazioni su classe e 'falsa coscienza'. Quaderni Storici 36, 900–908.

Touraine, A., 1968. La Conscience Ouvrière. Seuil, Paris.

Wright, E.O., 1985. Classes. Verso, London.

Class: History of the Concept

Jan Pakulski, University of Tasmania, Hobart, TAS, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The history of the concept of class – the evolution of its meaning and deployment – is divided into five periods: (1) the 'prehistory of class'; (2) the formative period of the early nineteenth century, when the concept enters the stage in its Marxist formulations; (3) the period of spread and diversification at the beginning of the twentieth century; (4) the period of ascendancy and proliferation in the second half of the twentieth century; and, finally, (5) the period of decomposition and eclipse starting during the last decades of the twentieth century and lasting till now.

Although 'class' is a very old concept, with its roots in ancient Rome, its contemporary meanings were shaped and reshaped in the context of modern debates triggered by the Industrial Revolution in Europe and its worldwide consequences. In this context, and in the context of Marxist theory and ideology, class started to denote large economically based and polarized (opposed, antagonistic) social groupings, marking the main lines of structured social inequality, conflict, and change in industrializing societies. The concept subsequently evolved together with the changing patterns of social inequality and contestation, as well as the theoretical debates that pitched class against its main concepts-competitors: estate, rank, race, nation(ality), occupational group, stratum, elite-mass, etc. Its popularity had initially increased, and then gradually faded, especially following the declining popularity of Marxism – the theoretical cradle in which the concept of class was embedded – and the collapse of communism at the end of the twentieth century. It is therefore convenient to see the concept of class as (1) theoretically embedded; (2) diverse in its meaning; (3) ideologically implicated, and therefore always politically contested; and (4) historical, that is, evolving together with major social upheavals and their theoretical accounts. It is convenient to divide the history of the concept – its changing meaning and deployment – into four periods:

1. the 'prehistory' of class – the early period when the concept acquired its hierarchical-classificatory meaning;
2. the formative period of the early nineteenth century, when the concept entered the stage in its original (but diverse) Marxist formulations;
3. the period of reshaping at the turn of the nineteenth and twentieth centuries, when the concept was reworked within and outside Marxism;
4. the period of ascendancy and proliferation, when 'class' entered popular discourses and vocabularies – but in an increasingly vague and diverse form; and, finally,
5. the period of further diversification, decomposition, and eclipse that started during the last decades of the twentieth century and continues at present.

Prehistory of Class

The earliest references to 'class' – in its hierarchical-classificatory meaning – are found in ancient Rome, where the summons to military service, called *classis*, ranked citizens into hierarchical divisions (subsequently called *classes*) according to the value of property and expected military equipment and strength. The landless *proletarii* were excluded from the service – they formed the equivalent of ancient 'underclass.' The propertied (land-owning) citizens – those with the right to carry arms and vote – formed seven classes, ranging from armed infantry warrior to those who could afford entire cavalry detachments (Cowell, 1962). This early usage shaped the meaning of the term that became a synonym of a social rank, a position in social hierarchy, a location in the stratification system typically based on wealth and income. While in the medieval and early modern period the term 'class' was seldom used – overshadowed by more traditional notions, such as 'estates' and 'orders' – it was, nevertheless, mentioned as a generic classificatory concept, an equivalent of 'social rank' or 'stratum.' The modern usage, pioneered by Karl Marx, oscillates between this generic meaning of the main 'polar' social division, and a more specific meaning, as a type of social division emerging together with modern capitalism. The latter – the modern class based on capital ownership and employment status – was contrasted with the old social divisions ('estate' or 'order') based on the ownership of land and the accompanied status (honor). While the old 'estates' were closed and ascriptive, the new class divisions were seen as open and achievement based. It was the 'first estate' – the despised, privileged, and exclusive aristocracy – that provoked the ire of French revolutionaries.

Class in Marxism

The contemporary meaning of 'class' arose in the context of the Marxist theory of social conflict and change formulated in response to the Industrial Revolution in Britain and Western Europe or, strictly speaking, to the disruptive social consequences of the 'Great Transformation' (Polanyi, 1944). Modern sociologists, political scientists, social historians, and anthropologists made 'class' a central focus of their analysis, mainly under the influence of rising and evolving Marxism. Class was the central concept in Marxist historical materialism, referring to three theoretically interrelated analytic entities, usually organized into 'polar' (opposed) pairs: (1) the major structural socioeconomic categories differing in their relation to the key 'means of production' or the key productive resources (capital

vs labor) – sometimes referred to by Marx as *Klasse an sich*; (2) the large social groupings, the social 'carriers' of the employment relations, forming on the basis of the main ownership/employment relations (bourgeoisie vs proletariat); and finally (3) the principal collective 'social actors': class-conscious, solidarity, and organized 'social forces,' formed in the process of class struggles (ruling class vs challenging class) – seen by Marx as 'mature' *Klasse für sich*.

This multiplicity of meanings reflected the Marxist theoretical aspiration to explain the social–historical evolution of societies (the 'rules of motion of capitalist society'), and bridged the gap between structural accounts of social hierarchy on the one hand, and the 'actor-centered' accounts of social change on the other. Class became a pivotal 'bridging concept' in this attempt. Social change always started, according to Marx, within the socioeconomic 'base' of society (technology and know-how). It then articulated itself in conflicts between the main antagonistic classes. These principal class conflicts – in the case of capitalism between the bourgeoisie (capital owners–employers) and the proletariat (workers–employees) – were gradually intensifying leading to revolutionary changes.

In addition to major 'polar' or 'antagonistic' classes, Marx also distinguished some 'intermediate' and 'transient' classes bound to gradually disappear in the processes of class polarization and struggle (Table 1). The class structure of modern capitalism was thus seen not only as complex, but also as gradually simplifying and polarizing. The most developed capitalist societies were to be sharply divided into the mature ruling class (bourgeoisie) that typically controlled and managed the state, and the expanding, impoverished, and dominated – and also increasingly conscious, antagonistic, and organized – working class. This class polarization and contestation was supposed to reach its peak when the dominated class – now class-conscious, politically organized, and backed by bourgeois defectors – took over power in a revolutionary coup and completed a revolutionary reconstruction of society. Note that in this evolutionary–revolutionary cycle major classes would change from nominal categories (*Klasse an sich*) to real grouping and organized collective actors (*Klasse für sich*), while minor classes would wither away. Note also that the concept of class was ideologically 'loaded' because Marxism claimed to be not just an explanatory theory but also – and principally – a program for change, a radical recipe for social reconstruction.

Table 1 The Marxist class scheme

Classes	*Trends*
Bourgeoisie	Growing in power and organization but declining in number
Old dominant classes (aristocracies & nobilities)	Declining in number and power
New middle classes (officialdom)	Growing in number but joining the bourgeoisie
Old middle classes (petit bourgeois, craftsmen, peasant-farmers)	Declining in number and power
Working class	Growing in number, consciousness, and organization

Based on Pakulski, J., 2004. Globalising Inequalities. Allen & Unwin, Sydney.

At the end of the nineteenth century, the process of conceptual refashioning of class started in the most industrialized European societies, with the realization that class polarization was arrested, class radicalism gave way to reformism, and the large 'middle classes' are expanding, rather than shrinking. Moreover, the 'old' social divisions ('estates') had not disappeared, and the new social divisions along ethnonational (and racial) lines started to dominate popular identifications.

Thus, the concept of class started to evolve in competition with four main conceptual rivals. The first one was 'estate' or 'order' that described hierarchical 'social ranks': the nobility, the clergy, and the residual 'third estate' of peasants, traders, and burghers, as well as a new status group of 'the intelligentsia,' emerging from the *déclassé* nobility. Class also competed with the increasingly popular concepts of 'people' (nation), referring to groups sharing a similar culture and tradition, as well as 'races,' referring to social categories with physical similarities. Another conceptual competitor was an equally polar pair of concepts, 'elites vs the masses,' popularized by mainly Italian critics of Marxism (Vilfredo Pareto, Gaetano Mosca, and Roberto Michels). This competition was particularly intense in Europe. In the United States, where ethnoracial divisions were more salient than socioeconomic ones, class was overshadowed by race.

Class in the Twentieth Century

Class emerged as the dominant concept–descriptor of social inequality and conflict at the beginning of the twentieth century. There were three reasons for this triumphant ascendancy: class was embedded in an increasingly popular (mainly in Europe) Marxist theory of social conflict and change; it became part of an attractive ideology promising a happy resolution of the main 'social question'; and class mobilization became a goal of both the radical social movements (socialist, revolutionary), as well as the moderate social-reformist push (social democrats). Moreover, the Bolshevik Revolution in Russia, an event that demonstrated the ideological mobilizing potential of the class concept, enhanced the popularity of 'class,' especially among socialists.

This ascendancy, however, was accompanied by theoretical and semantic diversification. As pointed out by Ossowski (1963/1958), the meaning of 'class' diverged within three theoretical–political contexts.

Polar Classes

Most Marxists, as well as radical social democrats, advocated 'polar' class schemes that stressed class exploitation and conflict: exploiters vs exploited, rich vs poor. Thus the Bolshevik leader, Vladimir Lenin, famously defined classes as "large groups of people differing from each other by the place they occupy in a historically determined system of social production, by their relation (in most cases fixed and formulated in law) to the means of production, by their role in the social organization of labor, and, consequently, by the dimensions of the share of social wealth of which they dispose and the mode of acquiring it" (Lenin, 1919/1928: 408).

Occupational Classes

The functional theorists, by contrast, identified classes with occupational groupings. Thus defined, classes were seen as elements of social differentiation, rather than social stratification. They complemented each other, even if competing for material rewards, privileges ('rents'), and social recognition. Many contemporary scholars, such as Grusky and Sorensen (1998), follow this analytic tradition and see social classes as 'hierarchically organized occupational microcommunities.' Contemporary analysts of such occupational classes tend to stress competition between occupational classes, rather than class conflict. Studies of labor market differentiation and segmentation, as well as analyses of professionalization, often refer to such concept of class (e.g., Grusky, 2005).

Class Stratification

Finally – and perhaps most significantly – class was redefined by Max Weber and his followers as a 'market-based social stratification concept.' Weberian classes, forming modern 'social ladders,' consisted of large social groupings that shared similar types and levels of 'market endowments,' that is, property, skills, and labor power. They differed in degree of 'social formation,' that is, the degree to which class members shared mobility chances, lifestyles, and identities. The typical class structure of industrialized societies, according to Weber, consisted of four large and hierarchical 'social classes': the owner and manager class, the *petit bourgeoisie* or small owner class (mainly shopkeepers and family businessmen), the white collar (mainly clerical) class, and the manual working class. Social classes seldom gained a high level of class consciousness, identity, and organization – a view that contrasted with the Marxist belief in the inevitable evolution of the major classes into organized *Klasse für sich*. Thus, the Weberian concept(s) of class became the major element of the burgeoning field of stratification and mobility studies.

Initially, the Weberian reformulations were overshadowed by the more radical Marxist vocabularies of class. In fact, the early twentieth century marked the most dramatic period in the history of class concept. World War I, the Bolshevik Revolution, and the Great Depression radicalized national populations, thus forming receptive audiences for the Marxist concept of class in its polar and radical versions. In the highly industrialized western Europe, antagonistic class divisions, class movements, and elaborate class ideologies formed a powerful stream of turbulent 'class politics' competing with equally violent racist–nationalist politics. This resulted not only in mass radicalization, favoring the polar class schemes, but also in 'reversals of democratization,' emergence of radical dictatorships ('revolutionary dynamo states'), brutal wars, and genocides: the Soviet–Stalinist 'classicides' and the Nazi 'ethnoracial genocides.'

Proliferation, Decomposition, and Decline

The class concept reached the peak of its popularity in the 1950s–70s as the flagship concept in the social sciences and in popular discourses on social stratification and mobility. It was seen as the main element of modern social structure, and as the key element of occupational class stratification. This widening popularity accompanied the ascendancy of 'class analysis' – a type of social analysis focusing on class inequality and antagonism. In the United States, class entered popular vocabularies mainly via 'community studies,' where 'class' meant a social rank in a stratified community – a group of families sharing a similar social status or 'standing' (e.g., Warner, 1960).

The 'Golden Age' of class was in the 1960s–80s when comparative national studies of class inequality, class stratification, and class mobility reached their peak. In such studies, social inequalities were portrayed as pyramids of hierarchical occupational classes. Class mobility was measured as circulation between those class locations, while mobility barriers defined 'class closure.' Class identifications were treated as universal 'master identities'; politics in industrialized 'advanced societies' were analyzed as 'democratic class struggles'; and cultural divisions were construed in terms of 'class cultures' (e.g., Giddens, 1973; Parkin, 1979; Goldthorpe, 1980; Wright, 1978; Lipset, 1960/1981). Perhaps the best known and most frequently used was a collapsible class-stratification scheme popularized worldwide by John Goldthorpe in international comparative studies of class structure and mobility (Erikson and Goldthorpe, 1992) (**Table 2**).

The concept of 'class' was also applied – although seldom officially – to analyses of inequalities and conflicts in the Soviet-type societies. Thus Wesolowski (1978) charted class and 'strata' inequalities in Poland, while dissident scholars, such as Djilas (1957) in Yugoslavia and Konrad and Szelenyi (1979) in Hungary singled out *the intelligentsia* and the party-state *nomenklatura* (patronage list of top party-state officials)

Table 2 Goldthorpe's class scheme

Full 11-class version	*Seven-class scheme*	*Five-class scheme*	*Three-class scheme*
1. Higher Professional			
2. Lower Professional	Service class (1 + 2)	White collar (1–3)	
3a. Higher administrator			
3b. Sales & service	Sales & service (3)		Nonmanual (1–4a)
4a. Small owner/ employer			
4b. Small owner	Petty bourgeois (4a, 4b)	Petty bourgeois (4a, 4b)	
4c. Farmer	Farmer (4c)	Farm worker (4c)	Farm worker (4c)
5. Lower Technician			
6. Skilled Manual	Skilled worker (5 + 6)	Skilled worker (5 + 6)	Manual (5–7)
7a. Semi- and unskilled	Unskilled (7a)	Unskilled worker (7)	
7b. Agricultural worker	Agricultural labor (7b)		

Based on Erikson, R., Goldthorpe, J., 1992. Constant Flux. Oxford University Press, Oxford.

as new and exploitative state–socialist 'ruling classes' pitched against the 'working classes.'

Amid this proliferating class analyses, however, there appeared some critical voices pointing to major transformations in class structure and the declining relevance of class. The advocates of '*embourgeoisement*' suggested that spreading affluence, welfare reforms, and intense mobility lead to the formation of 'middle class' societies. The early advocates of 'industrial society,' like Ralf Dahrendorf (1959), pointed to a progressive 'decomposition' of major social classes that resulted in proliferation of intraclass divisions, expansion of the new and amorphous 'middle classes,' intense mobility, declining class identity, consciousness and solidarity, and waning industrial class conflict. The advocates of 'post-industrial' developments, like Daniel Bell (1973), in turn, stressed the declining silence of class divisions and the growing relevance of knowledge/education and 'sector location' ('social situs'). At the same time, the neo-Weberian scholars, like Giddens (1973), pointed to the variability and flexibility of 'class structuration,' and started to supplement class schemes with rival concepts, such as 'elites,' 'strata,' and ethnoracial 'underclass.' In the meantime, students of occupational attainment and class mobility (e.g., Blau and Duncan, 1967; Goldthorpe, 1980; Heath, 1981) stressed the porous nature of class boundaries and high intensity of interclass movements. Even the neo-Marxist analysts, like E.O. Wright (1978), who initially defended the Marxist class scheme, started to multiply classes (from 4 to 12), stress the multidimensionality of class structure, and admit the blurring of class divisions – changes that marked a turning point in the popularity of class and the growing convergence between 'class' and its conceptual rivals (Wright, 1996, 1997; **Table 3**).

This convergence, however, was accompanied by confusion. Thus, E.P. Thompson (1964: 9) defined class as a rather nebulous "historical phenomenon, unifying a number of disparate and seemingly unconnected events." "I do not see class as a 'structure' nor even as a 'category'" – Thompson declared – "but as something which in fact happens … when some men, as a result of common experiences (inherited or shared) feel and articulate the identity of their interests as between themselves, and against other men whose interests are different from (and usually opposed to) theirs." Class concept, in other words, started to lose consistency and clarity, and with it also its analytic utility.

The advocates of class did not give ground without a fight. The prominent class analysts were busy adapting the concept to proliferating social and cultural studies. Prominent sociologists tried to explain the fuzziness and fluidity of 'new class divisions' by pointing to changing work conditions and intense mobility (Goldthorpe), diverse patterns of 'class structuration' (Giddens), variation in class closure and class formation (Parkin), and by the emergence of altogether new classes (Poulantzas, Wright).

A vigorous defense of class as a relevant and useful concept was also undertaken by the neo-Marxist 'world system' or 'dependency' theorists, such as Immanuel Wallerstein (1974), and by the 'culturalists' like Pierre Bourdieu (1984/1979). Wallerstein and his followers (including Christopher Chase-Dunn and Daniel Chirot) argued that class divisions and conflicts should be reconceptualized within a global 'world-systemic' theoretical framework as structural world-market divisions between families of nation-states. They distinguished between internal classes (within nation-states) and global classes: capitalist 'core' (consisting of the most developed nation-states) and the underdeveloped 'periphery' (consisting of developing regions). Between them there were 'semiperipheral' societies, such as Australia and Canada, as well as noncapitalist societies of the Soviet type. Popularity of these global (world-market) class schemes faded with the collapse of Soviet-type societies and the rapid development of the industrializing nations – both contradicting the predictions of dependency theorists. Even when S.M. Lipset, the doyen of class analysis in America, abandoned his original scheme of 'class politics' and 'democratic class struggle' (Clark and Lipset, 1991), the defenders of class and class analysis were confirming its new relevance (e.g., Hout et al., 1995; Wright, 1996).

While losing its popularity among mainstream analysts, class continued to be deployed by 'global class' theorists who suggested the 'denationalization' of classes, such as the 'transnational bourgeoisie' (Robinson, 2004: 49–57). Sassen (2007: 167–170) has distinguished three such 'denationalized' classes: transnational corporate professionals and executives, top state executives, and 'an emergent class of disadvantaged or resource-poor workers and activists.'

Bourdieu's (1984, 2001)'culturalist' reconceptualization of class has opened the debate about the importance of 'cultural capital' (closely identified with education, taste, and lifestyle) and cultural domination in generating new class divisions. According to Bourdieu, diverse lifestyles reflect distinctive *habituses*, which, in turn, are organized in the process of social–cultural contest into hierarchical (in relation to 'legitimate culture') and antagonistic classes. The subordinate/working classes contest social and cultural domination of dominant/

Table 3 Wright's class scheme

		Relations to the means of production				
		Owner		*Employees*		
Number of Employees	Many	Capitalist	Expert manager	Skilled manager	Nonskilled manager	**Relation to authority**
	Few	Small employer	Expert supervisor	Skilled supervisor	Nonskilled supervisor	
	None	Petty bourgeois	Expert worker	Skilled worker	Nonskilled worker	
			Expert	*Skilled*	*Nonskilled*	
			Relations to scarce skills			

Based on Wright, E.O., 1997. Class Counts: Comparative Studies in Class Analysis. Cambridge University Press, Cambridge, UK.

ruling classes. But classes are many, and they are diverse, fuzzy, and fickle. Their composition and boundaries constantly change. Moreover, as stressed by Weininger (2005), class divisions and conflicts analyzed by Bourdieu always fuse with much older and more entrenched racial and gender divisions and conflicts. The latter cannot be separated from contemporary class antagonisms, thus generating a complex web of crosscutting hierarchical divisions.

Bourdieu's approach has not only popularized – but also further diluted – the concept of class. It now applies to any socially recognized division: market-occupational, cultural-lifestyle, status-consumption, etc. – and it acquired gender and racial 'dimensions' that further complicate class analysis. Alas, the initial enthusiasm greeting Bourdieu's analyses started to fade when it became apparent that his 'flexible and inclusive' class scheme cannot be operationalized in a clear way.

The Proliferation of Classes or the 'Death of Class'?

As far as the concept of class is concerned, the current (2014) state of affairs is variously characterized as either an 'adaptive proliferation' or as the 'death of class,' a progressive waning of the class concept and its disappearance from academic and popular vocabularies.

The prominent neo-Marxist scholar E.O. Wright (Wisconsin–Madison), together with and Richard Breen (Oxford), neo-Ricardian analysts, such as Aage Sorensen (Harvard), and the neo-Durkheimian David Grusky (Stanford), have recently suggested that the concept of class and the accompanying class analysis do not decline in relevance, but simply diversify and adapt to new global and postindustrial conditions (Wright, 2005). Consequently, what the term 'class' means and denotes depends on the social-historical and theoretical contexts within which the concept is embedded. And these contexts continue to multiply. For the British neo-Weberian, Anthony Giddens (2006: 300), class is "a large scale grouping of people who share common economic resources, which strongly influence the type of life they are able to lead. Ownership of wealth, together with occupation, are the chief bases of class differences." For an American ('analytic') neo-Marxist, Wright (2005: 16–21), class refers to structural category of people located in similar (unequal and exploitative) ownership/work relations. For a follower of neoclassical economics, Sorensen (2005: 119–152), class denotes a category of people who extract similar economic 'rents,' that is, similar privileges and disadvantages. For a neo-Durkheimian, David Grusky (2005: 51–81), class is identical with a cohesive occupational group community and it forms the main element of contemporary social stratification. Followers of Bourdieu, by contrast, see classes as blurred but antagonistic social-cultural groupings, and they draw complex 'class maps' combining economic and cultural capitals. Students of globalization add to the widening gallery of competing class concepts 'transnational classes.' Structural and hierarchical delineations of 'classes' compete with those that denote a capacity for solidary collective action. You can pick your class concept and theoretical–analytic framework – whichever you find relevant and attractive.

Other scholars, the 'class skeptics,' claim that class is 'dying' and class analysis has been losing its relevance and utility (e.g., Beck, 1992; Clark et al., 1993; Pakulski and Waters, 1996; Clark and Lipset, 1991, 2001). This does not mark the end of social hierarchy, but a change in the forms of this hierarchy, division, and conflict. The old divisions – for which the concept of class was originally forged – have been decomposing, the accompanying class identities decline, and the lifestyles and politics associated with those divisions have withered away. Class inequalities, in other words, are gradually replaced by fluid and diverse – although often steep – 'classless inequalities.' Class is no longer a useful analytic concept-tool, especially in studies of the contemporary stratification resembling a 'status bazaar.' Contemporary cultures diversify into fluid subcultures and they spawn cultural 'omnivores' whose tastes and lifestyles defy Bourdiesque *habituses*. Politics also lose their class 'moorings.' Class–voter–party dealignment has resulted in the rapid decline of 'class identification' and 'class voting.' This is accompanied by a gradual withering away of 'class issues,' 'class parties,' and 'class politics.' They are replaced by 'celebrity' issues, risk concerns, cartel-type parties, and populist politics. Patterns of political association and action detach themselves from the old class-related ideological packages. So do the issue–agendas of social movements and mass parties. Both abandon class issues, vocabularies, and ideological programs (e.g., Beck, 1992; Giddens, 2000; Pakulski, 2005). Consequently, the major developments in the last three decades – the crisis and collapse of the Soviet Union, the unification of Europe, the ascendancy of market liberalism, the mobilization of powerful 'new' movements, (left-libertarian and right populist), the eruptions of 'terror-containment wars,' and the trans-Atlantic Great Recession – all defy analysis and explanation in class terms.

The major public and academic sources today define 'class' in most general terms, often detached from theoretical contexts. Thus the *Princeton Web Encyclopedia – Glossary* (2013) defines class as "people having the same social, economic, or educational status," while *Wikipedia* (2013) – an increasingly popular source – sees "**Social class** (or simply 'class'), as in a **class society**, [as] a set of concepts in the social sciences and political theory centered on models of social stratification in which people are grouped into a set of hierarchical social categories, the most common being the upper, middle, and lower classes." Moreover, the popularity of the concept continues to decline. While class dominated as a topic and referent in sociological journals in the 1970s and 1980s, it was mentioned in only 15% of British Sociological Association (BSA) articles at the turn of the century, and in only 8% of articles in *Sociology* in 2010 (Payne, 2011).

See also: Capitalism: The History of the Concept; Class Consciousness; Class: Social; Income Inequality; Inequality, Social; Inequality: Comparative Aspects; Marxism in Contemporary Sociology; Marxism–Leninism: The Ideology of Twentieth-Century Communism; Revolution, Theories of; Revolutions, History of; Revolutions, Sociology of; Russian Revolution, The; Social Inequality in History (Stratification And classes).

Bibliography

Beck, U., 1992. Risk Society. Sage, London.
Bell, D., 1973. The Coming of Post-Industrial Society. Basic Books, New York.
Blau, P., Duncan, O.D., 1967. The American Occupational Structure. Wiley, New York.
Bourdieu, P., 2001. Masculine Domination. Stanford University Press, Stanford.
Bourdieu, P., 1984. Distinction. Harvard University Press, Cambridge, MA.
Breen, R., 2005. Foundation of a neo-Weberian class analysis. In: Wright, E.O. (Ed.), Approaches to Class Analysis. Cambridge University Press, Cambridge, UK, pp. 31–50.
Chase-Dunn, C., 1998. Global Formations, updated ed. Rowman & Littlefield, Lanham, MD.
Chirot, D., 1977. Social Change in the Twentieth Century. Harcourt Brace Jovanovich, New York.
Clark, T.N., Lipset, S.M., 1991. Are social classes dying? International Sociology 6 (4), 397–410.
Clark, T.N., Lipset, S.M., 2001. The Breakdown of Class Politics. The John Hopkins University Press, Baltimore.
Clark, T.N., Lipset, S.M., Rempel, M., 1993. The declining political significance of social class. International Sociology 8 (3), 279–293.
Konrad, G., Szelenyi, I., 1979. The Intellectuals on the Road to Class Power. Harcourt Brace Jovanovic, New York.
Cowell, F.R., 1962. The Revolutions of Ancient Rome. Thames & Hudson, London.
Dahrendorf, R., 1959. Class and Class Conflict in Industrial Society. Routledge, London.
Djilas, M., 1957. The New Class. Wiley, New York.
Erikson, R., Goldthorpe, J., 1992. Constant Flux. Oxford University Press, Oxford.
Franklin, M.N., 1985. The Decline of Class Voting in Britain. Clarendon Press, Oxford.
Franklin, M.N., Macke, T., Velen, H., 1992. Electoral Change: Responses to Evolving Social and Attitudinal Structures in Western Countries. Cambridge University Press, Cambridge.
Giddens, A., 1973. The Class Structure of Advanced Societies. Hutchinson, London.
Giddens, A., 1996. Beyond Left and Right. Polity, Cambridge.
Giddens, A., 2000. The Third Way and Its Critics. Polity, Cambridge, UK.
Giddens, A., 2006. Sociology, fifth ed. Polity, Cambridge, UK.
Goldthorpe, J., 1980. Social Mobility and Class Structure in Modern Britain. Clarendon Press, Oxford.
Goldthorpe, J., Golden, M., Pontusson, J. (Eds.), 1992. Bargaining for Change: Union Politics in North America and Europe. Cornell University Press, Ithaca, NY.
Grant, J.A., 2001. Class, definition of. In: Jones, R., Barry, R.J. (Eds.), Routledge Encyclopedia of International Political Economy: Entries A-F. Taylor & Francis, p. 161.
Grusky, D., Galescu, G., 2005. Foundations of a neo-Durkheimian class analysis. In: Wright, E.O. (Ed.), Approaches to Class Analysis. Cambridge University Press, Cambridge, UK, pp. 52–81.
Grusky, D.B., Sorensen, J.B., 1998. Can class analysis be salvaged? American Journal of Sociology 103 (5), 187–234.
Heath, A., 1981. Social Mobility. Fontana, London.
Hout, M., Brooks, C., Manza, J., 1995. The democratic class struggle in the United States, 1948–1992. American Sociological Review 60, 805–828.
Lenin, V.I., 1919/1928. A Great Beginning. In: Collected Works, vol. 29. Progress Publishers, Moscow.
Lipset, S.M., 1960/1981. Political Man, second ed. John Hopkins University Press, Baltimore, MD.
Lipset, S.M., 1991. No third way: a comparative perspective on the left. In: Chirot, D. (Ed.), The Crisis of Leninism and the Decline of the Left. University of Washington Press, London, pp. 183–232.
Ossowski, S., 1963/1958. Class Structure in the Social Consciousness. Routledge, London.
Pakulski, J., 1993. The dying of class or of Marxist class theory. International Sociology 8 (3), 279–292.
Pakulski, J., 2004. Globalising Inequalities. Allen & Unwin, Sydney.
Pakulski, J., 2005. Foundations of a post-class analysis. In: Wright, E.O. (Ed.), Approaches to Class Analysis. Cambridge University Press, Cambridge, pp. 152–179.
Pakulski, J., Waters, M., 1996. The Death of Class. Sage, London.
Parkin, F., 1979. Marxism and Class Theory. Tavistock, London.
Payne, G., 2011. Sociological Class Analysis. http://www.sagebsa.co.uk/code/resources/pdf/Payne_Class%20Analysis.pdf (accessed 10.02.13.).
Polanyi, K., 1944. The Great Transformation. Basic Books, New York.
Robinson, W., 2004. A Theory of Global Capitalism. John Hopkins University Press, Baltimore, MD.
Sassen, S., 2007. A Sociology of Globalization. W.W.Norton, New York & London.
Sorensen, A.B., 2005. Foundations of a rent-based class analysis. In: Wright, E.O. (Ed.), Approaches to Class Analysis. Cambridge University Press, Cambridge, pp. 119–151.
Thompson, E.P., 1964. The Making of the English Working Class. Victor Gollancz, London.
Wallerstein, I., 1974. The Modern World – system. Academic Press, New York.
Warner, W.L., 1960. Social Class in America. Harper and Row, New York.
Weininger, E.B., 2005. Foundations of Pierre Bourdieu's class analysis. In: Wright, E.O. (Ed.), Approaches to Class Analysis. Cambridge University Press, Cambridge, pp. 82–118.
Wesolowski, W., 1978. Classes, Strata and Power. Routledge, London.
Wright, E.O., 1978. Class, Crisis and the State. New Left Books, London.
Wright, E.O., 1996. The continuing relevance of class analysis. Theory and Society 25 (5), 693–716.
Wright, E.O., 1997. Class Counts: Comparative Studiers in Class Analysis. Cambridge University Press, Cambridge, UK.
Wright, E.O. (Ed.), 2005. Approaches to Class Analysis. Cambridge University Press, Cambridge, UK.

Relevant Websites

Princeton University. Social Class. WordNet Search 3.1. http://wordnetweb.princeton.edu/perl/webwn?s=social+class&sub (accessed 12.02.13.).
Business Dictionary. Social class. http://www.businessdictionary.com/definition/social-class.html (accessed 10.02.13.).

Class Size

Diane W Schanzenbach, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article summarizes the literature on the impact of class size on student achievement. Following a brief discussion of the theory linking class size to student outcomes, I review the experimental and quasi-experimental evidence on the impact of class size. Research using a variety of approaches across a variety of settings indicates that class size is an important factor in student outcomes.

Introduction

Average class size in primary grades varies widely across countries, from over 30 in Chile and Korea to under 17 in Luxembourg, Greece, and the Russian Federation (OECD, 2010). Simple correlations between class size and student outcomes without accounting for the selection process of students to class size may be extremely misleading, as they may provide a biased estimate of the true impact of class size on student outcomes. This article begins with an overview of the theory linking class size to student outcomes, followed by a discussion of the empirical strategies used in research to identify the causal impact of class size on student achievement. It concludes with a review the empirical literature on class-size reduction focusing on studies using experimental and quasi-experimental techniques because these rely on the most credible strategies for identifying the cause-and-effect relationship between class size and student achievement.

Class Size as an Educational Input

Theories linking smaller class sizes to better student achievement generally posit that the relationship is mediated through increased time on task, the use of more effective instructional techniques such as individualized instruction, or both. None of the well-identified research studies described below are able to trace out the mechanisms through which smaller class sizes produce higher student achievement, and there is substantial uncertainty about the mechanism at work (Blatchford et al., 1998, 2002). Nonetheless, there is some research evidence suggesting that teachers spend less time on classroom management and discipline and more time on instruction activities in smaller classes (Bourke, 1986; Pate-Bain et al., 1992; Blatchford and Mortimore, 1994; Zahorik et al., 2003; Blatchford and Martin, 1998). In addition, the use and size of classroom groupings appears to be an important mechanism through which class size affects learning (Blatchford et al., 2001).

Lazear (2001) formalizes a useful theory of class size in which any given student is behaving in class at a given moment with probability p, and misbehaving with probability $(1\text{-}p)$. With n children in the classroom, p^n is the probability that the entire class is behaving at a given time (assuming that p is independent across children). Assuming a constant disruption rate, fewer students in class – holding total time in class constant – implies more time on task. As a result, the impact of a reduction in class size is a function of both the size of the class and also the behavior level of students. The theory therefore predicts that class-size effects will be larger for classes with more poorly behaved students.

Studying the Impact of Class Size

A naïve approach to modeling the relationship between student achievement and class size would be to estimate a regression analysis where a student's outcomes are a function of school-level inputs (such as class size and teacher quality), family inputs (such as income and parental education) plus an unexplained error term. Using such an approach, a negative estimated coefficient on class size indicates that student achievement tends to decline as class size increases.

The drawback of estimating such a model is that the estimated coefficient on class size will not reflect the causal impact of smaller class size if a student's class size is systematically related to other factors that determine achievement. If these other factors are not controlled, then the equation will suffer from omitted variables bias and the estimated coefficient may be severely misstated. For example, if weaker students are assigned to small classes in order to remediate their achievement, but if the researcher cannot adequately account for this information, then the estimated impact of school resources will be biased (Boozer and Rouse, 2001). In general, studies that control for observable factors but do not have a credible explanation for why students experience different class sizes are unlikely to yield reliable estimates of the impacts of class size.

The importance of using a research design that can estimate the causal impact of class size has long been acknowledged in the literature. In their influential meta-analysis of the research on class size and achievement, Glass and Smith (1979) conclude that there is a clear, strong relationship between class size and achievement, and that the relationship is "strongest in well-controlled studies in which pupils were randomly assigned to classes of different sizes." Subsequent literature reviews by Hanushek (1986, 1997) came to the opposite conclusion that the literature did not find a 'strong or consistent relationship' between student performance and class size. The Hanushek reviews have been sharply criticized both for the quality of studies included and for the manner in which they were aggregated (Krueger, 2003; Hedges et al., 1994). In

particular, many of the underlying studies included in the surveys employed research designs that would not allow researchers to isolate the causal effects of class size, but were nonetheless given substantial weight. Below, I limit the article to work that uses experimental or quasi-experimental research designs such as policy changes to identify the impact of class size.

Experimental Research

The most influential studies of class-size reduction are based on data from Project STAR (Student-Teacher Achievement Ratio), a large-scale randomized experiment in the U.S. state of Tennessee conducted from 1985 to 1989. In it, students were randomly assigned to attend one of three treatment types: a small-size class (target of 13–17 students), a regular-size class (target of 22–25 students), or a regular-size class with a full-time teacher's aide. Importantly, teachers were also randomly assigned to class types. The experiment took place in 79 public schools across a variety of geographic locations (inner-city, suburban, and rural; predominantly low-income and middle-class) for a single cohort of students in kindergarten through third grade. More than 11 500 students and 1300 teachers took part in the experiment. Students were randomly assigned within schools (or, more precisely, school interacted with the first year a student entered the treated cohort which was for most students either kindergarten or first grade). By design, students were to remain in the same randomly assigned class-type from kindergarten through the end of third grade. In practice, there were several major sources of deviation from this model such as nonrandom switches across class types, but the findings of the experiment have been shown to be robust to careful controls for these deviations (see Krueger, 1999).

Many researchers have shown that students randomly assigned to a small class perform 0.15–0.20 standard deviations higher on math and reading achievement tests compared with students in their school and grade who were randomly assigned to a regular-size class (Word et al. 1990; Krueger, 1999; Krueger and Whitmore, 2001; Schanzenbach, 2007). When the results are disaggregated by student characteristics, it appears that African-American students and low-income students benefit more from small classes than the overall population.

In fourth grade, the class-size reduction experiment concluded and all students were returned to regular-sized classes. Since participants have been tracked through approximately age 30, a good deal is known about long-term impacts of class-size reduction in the early grades. Overall, there is a persistent positive impact of small-class assignment that is, statistically significant (or borderline significant) through eighth grade, but the magnitude of the gain is one-third to one-half the size that was observed while the students were enrolled in the experimental classes (see Krueger and Whitmore, 2001). There is also some evidence of improvements in nonacademic outcomes during the teenage years, such as reduced criminal behavior and teen pregnancy (Krueger and Whitmore, 2002). Finn et al. (2005) find that small-class students more likely to graduate from high school. Students who had been randomly assigned to a small class were more likely to take a college-entrance exam in high school (Krueger and Whitmore, 2001).

Using information on college enrollment, new studies have shown that small-class attendance increases the rate of postsecondary attendance by between 1.8 (Chetty et al., 2011) and 2.7 percentage points (Dynarski et al., 2013), and increases the probability of earning a college degree by 1.6 percentage points. Small-class students have better outcomes across an index measured in early adulthood and including measures such as whether they are married, homeowners, have retirement savings, earnings, and neighborhood quality. Impacts on earnings from age 25–27 are not statistically significantly different from zero, but have a wide confidence interval. Across all dimensions, the long-term impacts have generally been found to be the strongest among disadvantaged children.

An important concern for any experimental results is whether the results may be generalized to other settings. Along many measures, Tennessee in the mid-1980s looks similar to other places that might be interested in implementing a class-size reduction policy, so it would be reasonable to expect similar effects as those in the experiment. However, compared with the United States overall, Tennessee has lower levels of educational inputs, such as education spending and teachers' educational attainment. If adding resources has a greater impact when baseline levels are already low, it might suggest that schools with higher levels of spending could experience a smaller impact of class-size reduction (Schanzenbach, 2007).

Quasi-Experimental Research

Because true randomized experiments are rare, researchers must also look for quasi-experimental approaches that allow isolation of the causal impact of class-size reduction. The most famous quasi-experimental approach to studying class-size reduction comes from Angrist and Lavy's (1999) use of a strict maximum class size rule in Israel and a regression discontinuity (RD) approach. In Israel, Maimonides' Rule dictates that no more than 40 students shall be in one class. As a result, the maximum class size faced by a student drop changes dramatically when enrollment is close to multiples of 40. For example, if a grade has 80 students, then a school could offer only 2 classrooms, with 40 students in each. If a grade has 81 students, however, in order to abide by the maximum class-size rule the school must offer (at least) 3 classrooms and maximum average class size falls to 27 students. In practice, some schools add an additional classroom prior to hitting the 40-student cap. Nonetheless the maximum class-size rule is a good predictor of actual class sizes, and can be used as an instrumental variable to isolate the causal impact of class size on student achievement. Using local variation around the enrollment sizes that are multiples of 40 students, Angrist and Lavy find strong improvements overall in both math and reading scores, of a magnitude that is consistent with Project STAR's experimental results. Consistent with the experiment results, they also find larger improvements among disadvantaged students.

There have been several subsequent papers that have identified the impact of smaller class sizes using maximum class size rules in other settings. Most recently, Fredriksson et al. (2013) evaluate the long-term impact of class size using data from Sweden and measuring class sizes among students between ages 10 and 13 who were facing a maximum class size rule of 30 students. At age 13, students who were in smaller classes had higher cognitive and noncognitive skills. In adulthood (measured between ages 27 and 42), students in smaller classes had higher levels of completed education, wages, and earnings. In addition, Urquiola (2006) uses a similar regression discontinuity approach in Bolivia and finds that a one standard-deviation reduction in class size (about 8 students in his data) improves test score performance by 0.2–0.3 standard deviations. Browning and Heinesen (2007) find similar results on data from Denmark, even though average class size is much smaller in their study (20 pupils per classroom, compared to 31 students in Angrist and Lavy's Israeli data). Urquiola and Verhoogen (2009) provide a cautionary tale about possible endogenous responses of schools to class-size caps, and show that in Chilean data that these endogenous responses of schools lead to violations of the assumptions of the RD design.

Another quasi-experimental approach was used to evaluate Wisconsin's targeted class-size reduction program called the Student Achievement Guarantee in Education (SAGE) program (Molnar et al., 1999). In SAGE, high-poverty school districts could apply to implement a pupil–teacher ratio of 15 to 1 in grades K-3. While most participating schools reduced class sizes, some schools chose to attain the target pupil–teacher ratio using two-teacher teams in classes of 30 students. Test scores of first graders in SAGE schools were higher in math, reading, and language arts compared with scores of first graders in selected comparison schools in the same districts, which had average pupil–teacher ratios of 22.4 to 24.5.

Another quasi-experimental approach is to use variation in enrollment that is driven by random fluctuations in cohort sizes across different years. Hoxby (2000) takes this approach using data from the U.S. state of Connecticut. After removing any preexisting trends in enrollment that might signify that a school district is waning or booming, the effect of class size is identified by the variation in cohort size that reflects a temporary random shock in population size that may have been caused by an unusually small (or large) birth rate in a given year. Using this approach, there is no statistically significant positive effect of smaller class size, and the estimates have the statistical precision to rule out an effect as large as about one-fifth the size found in Project STAR. The discrepancy between these results and those of other well-identified experimental and quasi-experimental studies remains a puzzle.

Relatively few well-identified studies have estimated the impact of class-size reductions in later grades. A notable exception is Dee and West (2011), which uses a fixed-effect approach to estimate within-student and within-teacher class-size effects. The study finds that small classes improve several measures of students' engagement in school. Furthermore, the fixed effects estimates are substantially larger than the ordinary least squares regression results, which is consistent with weaker students being assigned to smaller classes.

Statewide Policy-Induced Variation

Based in part on the research evidence on the impact of class-size reduction, several U.S. states including California, Texas, and Florida have implemented class-size caps. The most widely studied of the state policies was California's 1996 adoption of a law that gave strong monetary incentives to schools to reduce class size in grades K-3 to 20 or fewer students. The take-up of the policy was nearly universal within a short period of time, so there was very little scope to evaluate its impact. In addition, test scores are only available starting in grade 4, so any evaluation of the policy was forced to use subsequent test scores instead of scores during the year that the reduced class size was experienced. Evaluations suggest positive impacts on achievement that were probably somewhat offset because many inexperienced teachers had to be hired to staff the new classrooms (Jepsen and Rivkin, 2009; Unlu, 2005).

Discussion

The bulk of the research using credible identification of the impacts of class-size reduction suggests that reducing class size in the early grades improves both short- and long-run student outcomes. In addition, benefits appear to be larger for disadvantaged groups such as African-American students and children from low socioeconomic status families. There are many important research questions that have not been adequately addressed in the literature. For example, while class-size effects across a variety of studies appear to be stable and roughly linear in the range of 12–40 students, little is known about the potential effects of class size outside of this range. Similarly, little is known about the impact of class size on students older than elementary school.

An important question that policy makers must ask prior to embarking on class-size reduction is whether the projected benefits outweigh the costs. A cost–benefit study of Project STAR found that the overall benefits outweighed the costs (Krueger and Whitmore, 2001). The answer in other cases may depend on context, such as what is the level and distribution of current resources, can additional teachers of comparable quality be hired, and what is the best alternate use of available resources.

See also: Cognitive Development: Child Education; Cognitive Development: Mathematics Learning and Instruction; Education, Economics of; Educational Productivity; Educational Systems: North America; Socioeconomic Status and Academic Achievement; Vygotsky's Theory of Human Development and New Approaches to Education.

Bibliography

Angrist, Joshua D., Lavy, Victor, 1999. Using maimonides' rule to estimate the effect of class size on scholastic achievement. Quarterly Journal of Economics 114 (2), 533–575.

Blatchford, Peter, Mortimore, Peter, 1994. The issue of class size for young children in schools: what can we learn from research? Oxford Review of Education 20 (4).

Blatchford, Peter, Goldstein, Harvey, Mortimore, Peter, 1998. Research on class size effects: a critique of methods and a way forward. International Journal of Education Research 29, 691–710.

Blatchford, Peter, Martin, Clare, 1998. Research review: the effects of class size on classroom processes: it's a bit like a treadmill – working hard and getting nowhere fast! British Journal of Education Studies 46 (2), 118–137.

Blatchford, Peter, Baines, Ed, Kutnick, Peter, Martin, Clare, 2001. Classroom contexts: connections between class size and within class grouping. British Journal of Education Psychology 71, 283–302.

Boozer, Michael, Rouse, Cecilia, 2001. Intraschool variation in class size: patterns and implications. Journal of Urban Economics 50 (1), 163–189.

Bourke, Sid, 1986. How smaller is better: some relationships between class size, teaching practices, and student achievement. American Educational Research Journal 23, 558–571.

Browning, Martin, Heinesen, Eskil, 2007. Class size, teacher hours and educational attainment. Scandinavian Journal of Economics 109 (2), 415–438.

Chetty, Raj, Friedman, John N., Hilger, Nathaniel, Saez, Emmanuel, Schanzenbach, Diane Whitemore, Yagan, Danny, 2011. How does your kindergarten classroom affect your earnings? Evidence from project star. Quarterly Journal of Economics 126 (4), 1593–1660.

Dee, Thomas S., West, Martin, 2011. The non-cognitive returns to class size. Educational Evaluation and Policy Analysis 33 (1), 23–46.

Dynarski, Susan, Hyman, Joshua M., Schanzenbach, Diane Whitmore, 2013. Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion. Journal of Policy Analysis and Management 34 (2), 692–717.

Finn, Jeremy D., Gerber, Susan B., Boyd-Zaharias, Jayne, 2005. Small classes in the early grades, academic achievement, and graduating from high school. Journal of Educational Psychology 97 (2), 214.

Fredriksson, Peter, Öckert, Björn, Oosterbeek, Hessel, 2013. Long-term effects of class size. Quarterly Journal of Economics 128 (1), 249–285.

Glass, Gene V., Smith, Mary Lee, 1979. Meta-analysis of research on class size and achievement. Educational Evaluation and Policy Analysis 1 (1), 2–16.

Hanushek, Eric A., 1986. The economics of schooling: production and efficiency in public schools. Journal of Economic Literature 24, 1141–1177.

Hanushek, Eric A., 1997. Assessing the effects of school resources on student performance: an update. Educational Evaluation and Policy Analysis 19 (2), 141–164.

Hedges, Larry V., Laine, Richard D., Greenwald, Rob, 1994. Does money matter? A meta-analysis of studies of the effects of different school inputs on student outcomes. Educational Researcher 23, 5.

Hoxby, Caroline M., 2000. The effects of class size on student achievement: new evidence from population variation. Quarterly Journal of Economics 115 (4), 1239–1285.

Jepsen, Christopher, Rivkin, Steven, 2009. Class size reduction and student achievement the potential tradeoff between teacher quality and class size. Journal of Human Resources 44 (1), 223–250.

Krueger, Alan B., 1999. Experimental estimates of education production functions. Quarterly Journal of Economics 114 (2), 497–532.

Krueger, Alan B., 2003. Economic considerations and class size. Economic Journal 113 (485), F34–F63.

Krueger, Alan B., Whitmore, Diane M., 2001. The effect of attending a small class in the early grades on college-test taking and middle school test results: evidence from project STAR. Economic Journal 111, 1–28.

Krueger, Alan B., Whitmore, Diane, 2002. Would smaller classes help close the black-white achievement gap? In: Chubb, John E., Loveless, Tom (Eds.), Bridging the Achievement Gap. Brookings Institution Press, Washington.

Lazear, Edward P., 2001. Educational production. Quarterly Journal of Economics 116 (3), 777–803.

Molnar, Alex, Smith, Philip, Zahorik, John, Palmer, Amanda, Halbach, Anke, Ehrle, Karen, 1999. Evaluating the AGE program: a pilot program in targeted pupil-teacher reducation in wisconsin. Educational Evaluation and Policy Analysis 21 (2), 165–177.

Mosteller, Frederick, 1995. The tennessee study of class size in the early school grades. The Future of Children 5 (2), 113–127.

OECD, 2010. Education at a Glance 2010: OECD Indicators. OECD Publishing.

Pate-Bain, Helen, Achilles, Charles M., Boyd-Zaharias, Jayne, McKenna, Barbara, 1992. Class size does make a difference. Phi Delta Kappan 74 (3), 253–256.

Schanzenbach, Diane Whitmore, 2007. What have researchers learned from project STAR? Brookings Papers on Education Policy.

Unlu, Fatih, 2005. California Class Size Reduction Reform: New Findings from the NAEP. White paper, Princeton University.

Urquiola, Miguel, 2006. Identifying class size effects in developing countries: evidence from rural Bolivia. Review of Economics and Statistics 88 (1), 171–177.

Urquiola, Miguel, Verhoogen, Eric, 2009. Class-size caps, sorting, and the regression-discontinuity design. American Economic Review 99 (1), 179–215.

Word, Elizabeth R., Johnston, John, Bain, Helen P., Fulton, Dewayne B., Achilles, Charles M., Lintz, Martha N., Folger, John, Breda, Carolyn, 1990. The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Technical Report 1985-1990. Tennessee State Department of Education.

Zahorik, John, Halbach, Anke, Ehrle, Karen, Molnar, Alex, 2003. Teaching practices for smaller classes. Educational Leadership 61 (1), 75–77.

Class: Social

David B Grusky, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P. Saunders, volume 3, pp. 1933–1938, © 2001, Elsevier Ltd.

Abstract

The labor market of contemporary societies is rife with various types of 'classes' that impede the free flow of labor by restricting entry to those who have the requisite degrees, certificates, memberships, or capital. These classes take the form, for example, of occupations (e.g., economist, carpenter), aggregates of occupations (e.g., manager, farmer), or groups that represent competing factors of production (e.g., workers, capitalists). The purpose of this article is to review the conceptual foundation of class models and to introduce some of the research based on a class analysis of inequality.

Given the complexity of contemporary inequality, one might expect most scholars to adopt a multidimensional approach to characterizing it, an approach that might begin by specifying the multivariate 'inequality space' defined by the distribution of income, power, honor, and other valued resources. This instinct, however reasonable it might seem, has not in fact been the dominant one in the field. It instead remains conventional, especially among sociologists, to describe stratification systems in terms of discrete classes or strata whose members are presumed to have similar levels or types of assets. In the most extreme versions of this approach, the resulting classes are assumed to be real entities that preexist the distribution of assets, and many scholars therefore refer to the 'effects' of class location on the assets that their incumbents control.

The claim that inequality takes on a 'class form' is one of the distinctively sociological contributions to inequality measurement and stands as the main alternative to approaches that either focus exclusively on income inequality or analyze the many dimensions of inequality independently and separately. The main advantage of class-based measurement, as argued by some sociologists, is that conventional class categories (e.g., professional, manager, clerk, craft worker, laborer, farmer) are institutionalized within the labor market and are accordingly more than purely nominal or statistical constructions. The labor market, far from being a seamless and continuous distribution of incomes, is instead understood as a deeply lumpy entity, with such lumpiness mainly taking the form of institutionalized groups (i.e., 'classes') that constitute prepackaged combinations of such valued goods as income, power, or honor.

Within sociology, the implicit critique, then, of income-based approaches rests not so much on the argument that the income distribution is just one of many distributions of interest (i.e., multidimensionalism), but rather on the argument that measurement strategies based on the income distribution alone impose an excessively abstract, analytic, and statistical lens on a social world that has much institutionalized structure to it. This structure takes the tripartite form of a set of social classes that are privileged under capitalist labor markets (e.g., capitalists, professionals, managers), a set of social classes that are less privileged under advanced capitalism (e.g., routine nonmanuals, craft workers, operatives), and an 'underclass' that stands largely outside the labor market and is accordingly deeply disadvantaged in market systems. The rise of class models should therefore be understood as a distinctively sociological reaction to the individualism of the income paradigm and other unidimensional approaches to measuring inequality.

The foregoing account, which is a largely consensual rendition of the rationale for social class measurement, nonetheless conceals much internal debate within the field on how best to identify and characterize the boundaries dividing the population into classes. It is useful to review the three main phases in which these debates have played out.

Structuralist Phase (*c.*1945–85)

The class models of the postwar period rested implicitly or explicitly on the assumption that classes are bundles of endowments (e.g., education levels), working conditions (e.g., amount of autonomy), and reward packages (e.g., income) that tend to cohere together (Giddens, 1973; Parkin, 1979). The middle class of 'craft workers,' for example, comprises individuals with moderate educational investments (i.e., secondary school credentials), considerable occupation-specific investments in human capital (i.e., on-the-job training), average income coupled with substantial job security (at least until deindustrialization), middling social honor and prestige, quite limited authority and autonomy on the job, and comparatively good health outcomes (by virtue of union-sponsored health benefits and regulation of working conditions). By contrast, the underclass may be understood as comprising a rather different package of conditions, a package that combines minimal educational investments (i.e., secondary school dropouts), limited opportunities for on-the-job training, intermittent labor force participation and low income, virtually no opportunities for authority or autonomy on the job (during those brief bouts of employment), relatively poor health (by virtue of lifestyle choices and inadequate health care), and much social denigration and exclusion. The other classes appearing in conventional class schemes (e.g., professional, managerial, routine nonmanual) may likewise be understood as particular combinations of 'scores' on the most important assets or resources.

For the purposes of illustration, consider a simplified case in which the multidimensional 'inequality space' comprises only three individual-level variables (e.g., education, autonomy, income), thus allowing the class hypothesis to be readily

graphed. Additionally, assume that the class structure can be represented by six classes (e.g., professional, managerial, sales and clerical, craft, laborer, farm), signified in Figure 1 by six different symbols (dark squares, light squares, dark circles, etc.). As shown in this figure, the two main claims underlying the class hypothesis are that (1) the structural conditions of interest tend to cluster together into characteristic packages, and (2) these packages of conditions correspond to occupational or employment groupings. For structuralists, the inequality space is presumed to have a relatively low dimensionality, indeed a dimensionality neither more nor less than the number of postulated classes. The individuals falling within the classes constituting this scheme will accordingly have endowments, working conditions, and reward packages that are close to the averages prevailing for their classes. Moreover, even when individual scores deviate from class averages, the conventional class-analytic assumption (albeit largely untested) is that the contextual effect of the class is dominant and overcomes any such individual-level deviations. For example, a full professor who lacks a PhD is presumably just as marketable as a fully credentialed (but otherwise comparable) full professor, precisely because membership in the professorial class is a master status that tends to dominate all others.

The postwar period also was notable for a flourishing of gradational measurement approaches that again treated occupations as the fundamental units of analysis but assumed that such occupations may be ordered into a unidimensional socioeconomic or prestige scale (e.g., Treiman, 1976; Goldthorpe and Hope, 1972). In Figure 1, it is assumed that the class structure cannot be understood in simple gradational terms, meaning that at least some classes were formed by combining high values on one dimension with low values on another. It is possible, however, that the structural conditions of interest tend to covary linearly, thus generating a class structure of the very simple type represented in Figure 2. In a regime of this sort, inequality becomes rather stark (i.e., crystallized), as privilege on one dimension (e.g., autonomy) implies very

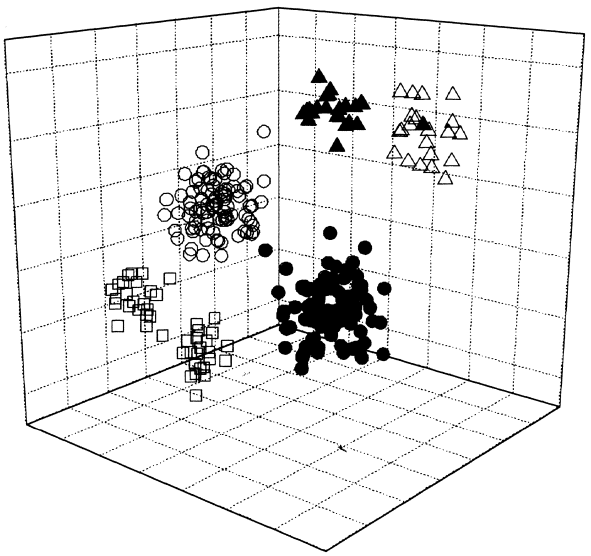

Figure 1 Class regime.

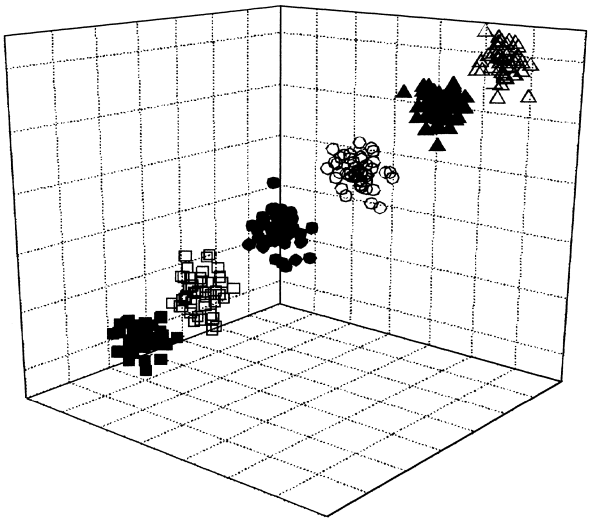

Figure 2 Gradational regime.

reliably privilege on another (e.g., income). There should accordingly be much interest in determining whether inequality indeed takes this form. Unfortunately, inequality scholars of the postwar period did not typically test the crystallization assumption, but rather simply assumed that it held and proceeded to develop socioeconomic scales that treated education and income as the main dimensions of interest (and ranked occupations by averaging scores on these two dimensions).

It may be noted that many neo-Marxian scholars during this period also deviated from a strict multidimensional stance by nominating particular dimensions as being theoretically crucial and hence the appropriate basis upon which social classes might be defined. There are nearly as many claims of this sort as there are dimensions of inequality. To be sure, Marx is most commonly criticized for placing "almost exclusive emphasis on economic factors as determinants of social class" (Lipset, 1968), but in fact much of what passed for stratification theorizing during this period amounted to reductionism of some kind, albeit often an expanded version of reductionism in which two or three dimensions were nominated as especially crucial. When a reductionist position is adopted, the rationale for a class model is not typically that classes are coherent packages of conditions (as represented in Figure 1), but rather that the nominated dimension or dimensions are crucial in defining interests and will accordingly come to be the main sources of social action. The classic Marxian model, for example, has workers ultimately appreciating that their status as workers (i.e., nonowners) defines their interests.

Culturalist Phase (*c.*1985–95)

In the mid-1980s, Bourdieu (1984) and other sociologists (esp. Wilson, 1978) sought to develop a culturalist rationale for class models, a rationale that rested on the claim that classes are not merely constellations of structural conditions (e.g., working conditions, rewards) but are also socially closed groupings in

which distinctive cultures emerge and come to influence attitudes, behaviors, or even preferences of class members. Throughout this period, many sociologists continued to work with more narrowly structuralist definitions of class (e.g., Wright, 1984; Erikson and Goldthorpe, 1992), but Bourdieu (1984) and Wilson (1978) were instrumental in pressing the claim that class-specific cultures are a defining feature of inequality systems (also, Lareau, 2003).

The two main forms of closure that serve to generate class-specific cultures are residential segregation (e.g., urban ghettos) and workplace segregation (e.g., occupational associations). As Wilson notes, members of the underclass live in urban ghettos that are spatially isolated from mainstream culture, thus allowing a distinctively oppositional culture to emerge and reproduce itself. The effects of residential segregation operate, by contrast, in more attenuated form for other social classes; after all, residential communities map only imperfectly onto class categories (i.e., the demise of the 'company town'), and social interaction within contemporary residential communities is in any event quite superficial and cannot be counted upon to generate much in the way of meaningful culture. If distinctive cultures emerge outside the underclass, they do so principally through the tendency for members of the same occupation to interact disproportionately with one another in the workplace and in leisure activities. In accounting, for example, for the humanist, antimaterialist, and otherwise left-leaning culture and lifestyle of sociologists, class analysts would stress the forces of social closure within the workplace, especially the liberalizing effects of (1) lengthy professional training and socialization into the 'sociological worldview' and (2) subsequent interaction in the workplace with predominantly liberal colleagues.

When occupations or classes are allowed to have cultures in this fashion, one naturally wishes to better understand the content of those cultures and, in particular, the relationship between such content and the structural conditions (i.e., endowments, outcomes, institutional setting) that a class or occupational situation implies. At one extreme, class cultures may be understood as nothing more than 'rules of thumb' that encode optimizing behavioral responses to prevailing institutional conditions, rules that allow class members to forgo optimizing calculations themselves, and rely instead on cultural prescriptions that provide reliable and economical shortcuts to the right decision. For example, Breen and Goldthorpe (1997) argue that working-class culture is disparaging of educational investments not because of some maladaptive oppositional culture, but because such investments expose the working class (more so than other classes) to a real risk of downward mobility. In most cases, working-class children lack insurance in the form of substantial family income or wealth, meaning that they cannot easily recover from an educational investment gone awry (i.e., dropping out); and those who nonetheless undertake such an investment therefore face the real possibility of substantial downward mobility. The emergence, then, of a working-class culture that regards educational investments as frivolous may be understood as encoding that conclusion and thus allowing working-class children to undertake optimizing behaviors without explicitly engaging in decision tree calculations. The behaviors that a 'rule of thumb' culture encourages are, then, deeply adaptive because they take into account the endowments and institutional realities that class situations encompass (see Morgan, 2008).

The foregoing example may be understood as one in which a class-specific culture instructs recipients about appropriate (i.e., optimizing) means for achieving ends that are widely pursued by *all* classes. Indeed, the prior 'rule of thumb' account assumes that members of the working class share the conventional interest in maximizing labor market outcomes, with their class-specific culture merely instructing them about the approach that is best pursued in achieving that conventional objective. At the other extreme, one finds class-analytic formulations that represent class cultures as more overarching worldviews, ones that instruct not merely about the proper means to achieve ends but additionally about the proper valuation of the ends themselves. For example, some class cultures (e.g., aristocratic ones) place an especially high valuation on leisure, with market work disparaged as 'common' or 'polluting' (see Veblen, 2008). This orientation presumably translates into a high reservation wage within the aristocratic class. Similarly, 'oppositional cultures' within the underclass may be understood as worldviews that place an especially high valuation on preserving respect and dignity for class members, with of course the further prescription that these ends are best achieved by (1) withdrawing from and opposing conventional mainstream pursuits, (2) representing conventional mobility mechanisms (e.g., higher education) as tailor-made for the middle class and, by contrast, unworkable for the underclass, and (3) pursuing dignity and respect through other means, most notably total withdrawal from and disparagement of mainstream pursuits. This is a culture that advocates that respect and dignity deserve an especially prominent place in the utility function and that further specifies how those ends might be achieved.

It should by now be clear that sociologists operating within the class-analytic tradition have adopted very strong assumptions about how inequality and poverty are structured. As noted, intrinsic to the class concept are such claims as (1) the space of outcomes and capabilities has a (low) dimensionality equaling the number of social classes, (2) the class locations of individuals become master statuses that dominate (or at least supplement) the effects of individual-level endowments, and (3) such class locations are socially closed and come to be associated with adaptive or maladaptive cultures. These claims, although frequently made, have not typically been explicity tested.

Postmodernist Phase (*c.*1995–Present Day)

The third phase of conceptual work within sociology has been marked, however, by an increased willingness to challenge the assumptions underlying the class-analytic status quo. Over the last two decades, such criticisms of the class-analytic enterprise have escalated, with many postmodernist scholars now feeling sufficiently emboldened to argue that the concept of class should be abandoned altogether. Although the postmodern literature is notoriously fragmented, the variant of postmodernism that is most relevant here proceeds from the assumption that the labor movement is rooted in the old and

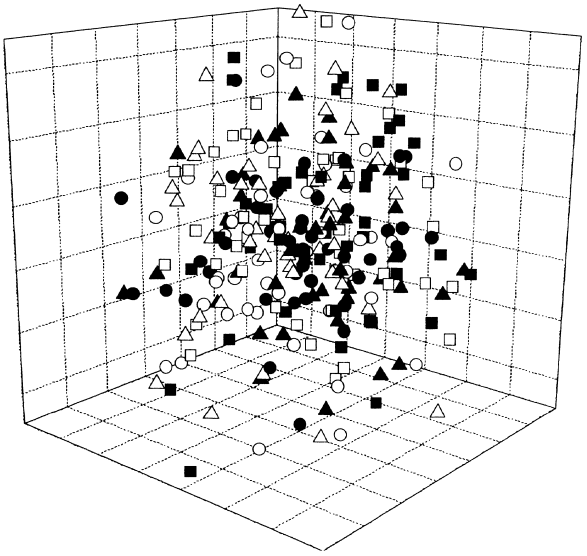

Figure 3 Disorganized inequality.

increasingly irrelevant conflicts of industrial capitalism, that political parties have abandoned class-based platforms in favor of those oriented toward values and lifestyles, and that class-based identities accordingly become ever weaker and more attenuated (see Hout and Laurison, 2014). The resulting 'individualization of inequality' (e.g., see Pakulski and Waters, 2008) implies that lifestyles and consumption practices are becoming decoupled from work identities as well as other status group memberships. The stratification system may be regarded, then, as a 'status bizarre' (Pakulski and Waters, 2008) in which identities are actively constructed as individuals select and are shaped by their multiple statuses.

This hypothesis, which is represented in extreme form by **Figure 3**, may well prove to be premature (but see Jonsson et al., 2011; Weeden and Grusky, 2012). Moreover, even if lifestyles and life chances are truly decoupling from economic class, this ought not be misunderstood as a more general decline in inequality per se. The brute facts of inequality will still be with us even if social classes of the conventional form are weakening. Most obviously, income inequality is clearly on the rise (e.g., Atkinson et al., 2011), while other forms of inequality show no signs of withering away. The postmodernist hypothesis speaks, then, to the way in which inequality is organized as opposed to the overall amount of such inequality.

The 'Class Effect' Rationale

We have to this point represented the class principle as a hypothesis about the clustering of observations in the inequality space. As an alternative motivation for the class hypothesis, it is sometimes claimed that classes are social contexts that affect attitudes, behavior, and individual action of many kinds. When this motivation is adopted, classes are not typically construed as information-rich social containers that capture many life conditions of interest, but rather as analytic categories that single out a particular social context that is presumed to be very consequential in defining interests. Under such a formulation, a class analyst will therefore typically nominate a single variable (e.g., authority, ownership) as especially useful in understanding the sources of social behavior, with the class categories then defined so as to capture differences across workers on that underlying variable of presumed consequence. The Marxian model, for example, famously embodies the claim that classes are best defined in terms of employment status alone (i.e., employed, self-employed), with the rationale for this definition being that employment status putatively defines interests and hence attitudes and behavior (e.g., Marx, [1894]1972). In contemporary labor markets, the class of employed workers is of course very heterogeneous, thus motivating class analysts to introduce further distinctions within that class that are presumed to be consequential in defining interests and action. There is a profusion of such contemporary class models on offer (e.g., Wright, 2005).

When a class model is motivated in terms of presumed class effects, it is important to establish that such effects are indeed truly causal. If, for example, one finds that seeming differences in the politics of professionals, managers, craft workers, and other social classes disappear when income is controlled, then presumably one can only refer to an income effect on politics, not a true class effect. Why might net effects of class be detected even with rigorous controls? In addressing this question, what must first be stressed is that, even when classes are defined in terms of a single analytic variable (e.g., form of regulation of employment), they are nonetheless often organic packages of conditions; and the constituents of these packages may combine and interact in ways that lead to an emergent logic of the situation (often expressed via a 'class culture'). The underclass, for instance, may be understood as a combination of negative conditions (e.g., intermittent labor force participation, limited education, low income) that, taken together, engender a sense of futility, despondency, or learned helplessness that is more profound than what would be expected from a model that simply allows for independent effects of each constituent class condition. To be sure, a committed reductionist might counter that one merely needs to include the appropriate set of interactions between the constituent variables, but insofar as classes define the relevant packages of interacting conditions this caveat just becomes an unduly complicated way of sidestepping the reality of classes.

Whatever the mechanism that underlies class effects, the common assumption is that classes are meaningful social contexts, just as neighborhoods are likewise understood within the 'neighborhood effects' literature as meaningful social contexts. These contexts are expected in both cases to have causal effects that are not reducible to mere selective processes. Again, we have to stress that such a 'class effects' rationale for class models is best treated as a hypothesis, as there is little in the way of substantiating evidence at this point (cf Weeden and Grusky, 2005).

Conclusions

Over the long history of class analysis, the class concept has been motivated in two main ways, either by (1) claiming that

the inequality space has a (low) dimensionality equaling the number of social classes or (2) claiming that the class locations of individuals have a true causal effect on behaviors or attitudes of interest. The foregoing claims have been unstated articles of faith among class analysts in particular and sociologists more generally. In this sense, class analysts have behaved rather like stereotypical economists, the latter frequently being criticized (and parodied) for their willingness to assume most anything provided that it leads to an elegant model.

See also: Income Inequality; Inequality Analysis: Overview; Inequality, Social; Inequality: Comparative Aspects; Poverty in History; Poverty, Geography of; Poverty, Sociology of; Religion and Stratification; Social Inequality and Schooling; Social Inequality in Cultural Consumption Patterns; Social Inequality in History (Stratification and Classes); Social Stratification; Urban Poverty in Neighborhoods.

Bibliography

Atkinson, A.B., Piketty, T., Saez, E., 2011. Top incomes in the long run of history. Journal of Economic Literature 49 (1), 3–71.

Bourdieu, P., 1984. Distinction: A Social Critique of the Judgment of Taste. Harvard University Press, Cambridge.

Breen, R., Goldthorpe, J.H., 1997. Explaining educational differentials: towards a formal rational action theory. Rationality and Society 9 (3), 275–305.

Erikson, R., Goldthorpe, J.H., 1992. The Constant Flux: A Study of Class Mobility in Industrial Societies. Clarendon Press, Oxford.

Giddens, A., 1973. The Class Structure of the Advanced Societies. HarperCollins, New York.

Goldthorpe, J.H., Hope, K., 1972. Occupational grading and occupational prestige. In: Hope, K. (Ed.), The Analysis of Social Mobility: Methods and Approaches. Oxford University Press, Oxford.

Hout, M., Laurison, D., 2014. The realignment of U.S. presidential voting. In: Grusky, D.B. (Ed.), Social Stratification: Class, Race, and Gender in Sociological Perspective, fourth ed. Westview Press, Boulder, pp. 1037–1045.

Jonsson, J.O., Grusky, D.B., Di Carlo, M., Pollak, R., 2011. It's a decent bet that our children will be professors too. In: Grusky, D.B. (Ed.), The Inequality Reader: Contemporary and Foundational Readings in Race, Class, and Gender, second ed. Westview Press, Boulder, pp. 499–516.

Lareau, A., 2003. Unequal Childhoods: Class, Race, and Family Life. University of California Press, Oakland.

Lipset, S.M., 1968. Social class. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences. Macmillan, New York, pp. 296–316.

Marx, K., [1894]1972. Capital. Lawrence and Wishart, London.

Morgan, S.L., 2008. A new social psychological model of educational attainment. In: Grusky, David B. (Ed.), Social Stratification, Class, Race, and Gender in Sociological Perspective, third ed. Westview Press, Boulder, pp. 542–549.

Pakulski, J., Waters, M., 2008. The death of class. In: Grusky, David B. (Ed.), Social Stratification, Class, Race, and Gender in Sociological Perspective, third ed. Westview Press, Boulder, pp. 1022–1029.

Parkin, F., 1979. Marxism and Class Theory: A Bourgeois Critique. Tavistock, London.

Treiman, D.J., 1976. A standard occupational prestige scale for use with historical data. The Journal for Interdisciplinary History VII, 285–290.

Veblen, T., 2008. The theory of the leisure class. In: Grusky, David B. (Ed.), Social Stratification, Class, Race, and Gender in Sociological Perspective, third ed. Westview Press, Boulder, pp. 862–869.

Weeden, K.A., Grusky, D.B., 2005. The case for a new class map. American Journal of Sociology 111 (1), 141–212.

Weeden, K., Grusky, D.B., 2012. The three worlds of inequality. American Journal of Sociology 117 (6), 1723–1785.

Wilson, W.J., 1978. The Declining Significance of Race: Blacks and Changing American Institutions. University of Chicago Press, Chicago.

Wright, E.O., 2005. Approaches to Class Analysis. Cambridge University Press, Cambridge.

Wright, E.O., 1984. A general framework for the analysis of class structure. Politics and Society 13 (4), 383–423.

Classical Archaeology

Alain Schnapp, Institute of Art and Archeaology, Paris, France

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 1938–1942, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The notion of 'classical archaeology' is of relatively recent origins in the vocabulary of the humanities. Its common usage dates to the second half of the nineteenth century, when archaeology, a newly formed university discipline, sought its place within the broader intellectual framework of the sciences of antiquity. Seen as the foundation of occidental history, classical archaeology has its roots in the passion of Renaissance 'antiquarian' collectors of Greco-Roman works of art. With the Enlightenment came a remarkable appreciation of the history of plastic forms in antiquity (Winckelmann), as well as an increased awareness of the technical achievements of the past (Caylus). By the nineteenth century, classical archaeology had begun to distance itself from collections and from the social milieu of collecting, advocating instead the necessity of observing finds in the field, and of recognizing them as coherent wholes open to rigorous analysis. It was thus a development of archaeology writ large, a universal discipline bearing on all epochs and regions of the world, which cast light on the specificity of classical archaeology. Out of this confrontation classical archaeology has emerged today as a fully-fledged branch of archaeology, based equally on written sources and on the methods of orientalist and prehistoric research.

The notion of 'classical archaeology' is of relatively recent origins in the vocabulary of the humanities. Its common usage dates to the second half of the nineteenth century, when archaeology, a newly formed university discipline, sought its place within the broader intellectual framework of the sciences of antiquity. Seen as the foundation of occidental history, classical archaeology has its roots in the passion of Renaissance 'antiquarian' collectors of Greco-Roman works of art. With the Enlightenment came a remarkable appreciation of the history of plastic forms in antiquity (Winckelmann), as well as an increased awareness of the technical achievements of the past (Caylus). By the nineteenth century, classical archaeology had begun to distance itself from collections and from the social milieu of collecting, advocating instead the necessity of observing finds in the field, and of recognizing them as coherent wholes open to rigorous analysis. It was thus a development of archaeology writ large, a universal discipline bearing on all epochs and regions of the world, which cast light on the specificity of classical archaeology. Out of this confrontation classical archaeology has emerged today as a fully-fledged branch of archaeology, based equally on written sources and on the methods of orientalist and prehistoric research.

The notion of 'classical archaeology' is of relatively recent origins in the vocabulary of the humanities. Its common usage dates to the second half of the nineteenth century, when archaeology, a newly formed university discipline, sought its place within the broader intellectual framework of the sciences of antiquity, the *Altertumswissenschaft* of German scholarship. For knowledge of antiquity to be scientific, it had to encompass all the creations of human genius. Thus, just as philology had emancipated itself from theology, so was archaeology to free itself from philology in order to constitute, together with the latter and with ancient history, the third pillar of *Altertumswissenschaft*. This disciplinary triangle effectively constitutes the foundations of the modern conception of the classical past—that is, the past of the Greco-Roman world.

The Definition of Classical Archaeology

Such a definition of classicism is quite evidently arbitrary. It rests on a very specific and thoroughly European experience of antiquity, in which Greco-Roman literature and its concepts are overwhelmingly privileged. In specifying the chronological and geographical boundaries of the classical world, nineteenth century historians and archaeologists have confirmed a spatio–temporal distinction already initiated during antiquity, and notably in the Alexandrian golden age of the third and second centuries BC. Following this definition, the classical world effectively begins with the dispersal of the Greeks in the Mediterranean during the eighth century BC, and ends with the fall of the occidental Roman empire in 476 AD. In terms of its geographical extension, the classical world reaches as far as the farthest expansion of the Roman empire, at the end of the first century AD.

As can be seen, this mode of historical representation clearly focuses on the Mediterranean as the zone of contact between Europe, Africa and Asia. But where does this Mediterranean space actually ends? Is it somewhere in the vast European plains? On the shores of the Black Sea? At the foot of the mountain ranges bordering Anatolia or Africa? According to their readings of both the classical sources and the surrounding landscapes, experts have drawn different boundaries to the classical world. Given this diversity of opinions, German archaeologists have proposed the notion of *Randkulturen* to designate all those civilizations known to have been in contact with the classical world without fully merging with it: Scythians, Parthians, Germans, Phoenicians, Egyptians, and many others.

Chronology is not much better established. When the Archaic epoch was first identified, archaeologists knew nothing of the Minoan world and the palace civilizations of Greek antiquity, civilizations which later on would end being called preclassical. At the other end of the scale, the date commonly chosen to signal the conclusion of Greco-Roman history—the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.13015-6

fall of the occidental Roman empire in 476 AD—proves to be equally problematic: it takes no account of the persistence of the oriental Roman empire until the fall of Constantinopolis in 1453.

The History of Classical Archaeology

With its uncertain chronological and geographical boundaries (a problem admittedly shared with protohistory), classical archaeology stands somewhat apart from other, more global, strands of archaeology. It is both the oldest among these archaeologies, and the one most influenced by tradition: the ancient Greek word *archaiologia* retains within it all its meaning as a discourse on antiquity. It was, after all, in fifth century Greece that the historical genre first appeared, the reasoned discourse on the past to which the Greeks gave the name of *historia* (enquiry). And since Varron in the first century BC, the title of antiquary has been given to the man who seeks to interpret and classify ancient objects or monuments. In Greece as well as Rome, these antiquaries contributed to elucidate the past; they assembled series of objects and monuments, collected inscriptions, and then put them in order and sought to understand them.

In this respect, the collapse of the Occidental Roman Empire was not without consequence, for with it a whole social class of knowledge- producers and users came to disappear. Secular men of letters were gradually replaced by clerics, whose function was to educate the masses and the elite in the Christian faith. To be sure, there were among these clerics some antiquarians who undertook to collect and investigate the monuments of antiquity. But one has to wait until the end of the Middle Ages to find an antiquarian movement comparable to that of Greco-Roman times.

It was in Renaissance Italy that a new passion for antiquity emanated which was also a means for reinventing the contemporary world. Collections of Greek and Roman manuscripts arose passions, ancient coins were coveted, monuments begun to be excavated for the sculptures and architectural remains they may have contained. The past that was sought after and revived was that of Greece and Rome, if only because a reference to this world was seen as a means to do away with a present seen as barbarous, illiterate and indeed 'gothic.'

For Renaissance scholars, antiquity was quintessentially Greco-Roman: the literary works, sculptures and architecture of these periods were seen as unsurpassable productions, to be studied and emulated by the best contemporary artists. This classical focus dominated scholarly and artistic production until the end of the eighteenth century, and it explains that 'barbarian' antiquities received only marginal interest. The Renaissance has thus contributed to conflating the knowledge of the past with the investigation and understanding of the Greco-Roman heritage—the part of universal culture believed to be 'classic' by virtue of being the veritable pedestal of humanist scholarship. There were some scholars who manifested considerable interest for the oriental world, Egypt or even Mesopotamia, and others, particularly in Scandinavia, Britain and Germany, who paid attention to local antiquities, but the model for the organization of knowledge derived directly from the prestigious Greco-Roman tradition.

Renaissance scholars and their Enlightenment successors assembled together coins, inscriptions, sculptures, and sometimes even ceramics or other objects of daily life. These collections were displayed in 'cabinets of curiosity' and eventually came to adorn the first public museums, such as the Venetian collections and the Ashmolean Museum in Oxford (from 1683). However, since no careful observations were made regarding the place of discovery and the mode of manufacture of these collected works and monuments, it was only with the greatest difficulty that scholars were able to classify them in terms of their dates and provenance. In fact, it was only the antiquarians of the eighteenth century who were first able to lay the grounds of a critical analysis of archaeological monuments.

These advances occurred at different paces, and they were due both to discoveries in the field and to new attitudes towards finds and their study. In this respect, the fortuitous discovery of the buried cities of Pompeii and Herculaneum at the beginning of the eighteenth century did much to awaken the interest of the elite in the Greco-Roman world. What was being unearthed there by the excavating teams of the King of Naples were no longer usual ruins, but entire cities, covered by volcanic eruptions and thus miraculously preserved from the ravages of destruction. Closely guarded by the Neapolitan government, the exploration of Pompeii gave rise to an unprecedented craze for the painting, architecture and implements of daily life of the time. Given the context of their discoveries, these items could also for the first time be attributed a secure chronological position. Naples thus played the role of an antiquarian capital, and gradually came to compete with Rome in the esteem of travelers and connoisseurs.

Without interpretation, however, discovery is nothing. In the second part of the eighteenth century, two men revolutionized the understanding of antiquity. The first of these was Johannes Winckelmann (1717–68), a German scholar from a small Prussian town who achieved tremendous influence with his work. His *History of Art in Antiquity*, published in Leipzig in 1764, offered to amateurs of Greco-Roman art the first ever systematic and chronological treatise on the subject. Aided by a thorough familiarity with textual sources, Winckelmann depicted and commented on antique objects scattered throughout European collections. Drawing on his personal study of the most important Roman collections, he could propose to his readers a remarkable history of plastic forms in antiquity. The style of this work, the quality of its descriptions, and the philosophical spirit which animated it, all explain its unprecedented success: with it, the occident could at last discover the sources of Renaissance art in the classical world. Winckelmann, otherwise an enemy of the aristocracy and determined critic of the ancient regime, received his highest acclaim from the courts of kings and princes, among those who sought antiquity for the aesthetic pleasure it procured them.

One of the most notable aristocrats in the court of France, the Comte de Caylus (1692–1765), was instrumental in drawing the attention of his contemporaries to another aspect of the study of the past. For Caylus, what justified the quest for antiquity is not the aesthetic, but the technical achievements of the past. He therefore developed an approach to the study of past techniques in which can be recognized our current concerns with artifact morphology and archaeometry. Each in

his way, Winckelmann for the appreciation of art and Caylus for the understanding of techniques, laid down the bases for a history of the *oeuvres* of antiquity, a research programme developed and followed throughout the nineteenth century.

The Rules of Classical Archaeology

The exploration of antiquity thus gradually transformed itself into a new discipline of archaeology. Under this framework, nineteenth century scientists elaborated and imposed new rules for the extraction, analysis and publication of evidence from the past. Archaeology distanced itself from collections and from the social milieu of collecting, and advocated instead the necessity of observing finds in the field, and of recognizing them as coherent wholes open to rigorous analysis. While the antiquarian channeled his passion for the past to the collection of objects, the archaeologist aimed to valorise these objects and monuments in the context of their discovery. This approach found its most systematic expression in Germany with the research programme of E. Gerhard (1795–1867), who effectively fought for an autonomous archaeology, free from collectors, philologists and artists. Archaeologists must take the place of connoisseurs with their various intermediaries, and themselves seek for the evidence on the ground.

They must also distance themselves from philologists, and challenge the primacy of written over nonwritten sources. Finally, they must take leave of aesthetic considerations, and study ancient productions in all their material and technical dimensions. Building on advances in historical studies, this positivist program also drew on developments in geology and natural history to forge for itself its specific scientific instruments.

Three complementary pillars came then to form then backbone of archaeology. One, *stratigraphy*, consists of observing the conditions of deposition of objects and monuments in the ground. The second, *typology*, seeks to identity in the objects themselves some evidence regarding the time and place of their manufacture. The third, *technology*, deals with the modes of production, the raw materials, the procedures of fabrication. It was in fact only by the end of the nineteenth century, when these three strands were mastered and marshaled together, that archaeology became a fully-fledged discipline able to undertake the reliable and methodical investigation of the past. The body of doctrines and institutions which accompanied this development contributed to giving archaeology, and specifically to classical archaeology, its distinctive identity. In this respect E. Gerhard created a novel kind of institution which prefigures in many ways those of today: the *Instituto di corrispondenza archeologica*, established in 1829 in Rome through the activities of aristocrats and scholars from across Europe, took it as its aim to collect, excavate and publish as thoroughly as possible all the antiquities of the Mediterranean world.

By drawing archaeologists to the field, by elaborating a modern strategy of publication based on detailed sections and descriptions, the *Instituto* virtually launched the new discipline that was embodied in this dedicated research institution located in Rome, one of the capitals of the classical world.

Until the middle of the nineteenth century, the exploration of the Greco-Roman world had been a matter of individual enterprise with occasional royal support. From then on, archaeology would become what the discoverer of Greek Asia Minor T. Wiegand (1864–1936) has called a *Grosswissenschaft*, a state-sponsored science. The leading European powers established then various institutes or schools in Rome and Athens, and these encouraged archaeological expeditions throughout the main urban sites of the Mediterranean. The enormous success encountered by the private initiatives of the German businessman Heinrich Schliemann (1822–90) at Mycenae and Troy fueled a veritable scramble for new excavating territories among the countries of Europe and even America. This competition contributed notably to the growth of collections in the main European and American museums. The rise of such museums as the British Museum, the Louvre, the Berlin museums and even the Metropolitan Museum in New York, all catering for an ever-increasing public, corresponded with the ascendancy of a classical archaeology closely entwined with the economic developments of capitalism and colonialism. The conquest of the past was effectively an instrument of foreign policy in the hands of all main powers, beginning with the core area of the Greco-Roman world and reaching towards the Orient, Africa, and the globe as whole.

Archaeological schools, museums and, of course, universities were all part of this expansion, from which they have benefited. Whereas in the mid nineteenth century only Germany was endowed with a network of university Chairs in classical archaeology, by 1914 all European nations as well as the United States had established university curricula in classical archaeology. This three-pronged movement, involving the development of museum collections, the systematic excavation of sites, and the setting of an academic theoretical and educational framework, effectively gave to classical archaeology its modern appearance.

The Paths of Classical Archaeology

Set out during the second half of the nineteenth century, this disciplinary model witnessed during the first half of the twentieth century a considerable expansion and consolidation in both qualitative and quantitative terms. Alongside a marked increase in the number of excavations carried out, the main archaeological institutions followed the German model by systematically publishing catalogues of finds and collections according to specific descriptive protocols. This increased and refined considerably the existing knowledge on artistic and artisanal productions: sculptures, architecture, ceramics were all researched, and so were numismatics and glyptics. Thus classical archaeology forged itself a vast corpus of systematic reference, bearing equally on chronology, typology and technology.

Some criticisms and contradictions emerged, however, during the second half of the twentieth century. The implicit supremacy of Greco-Roman culture came under challenge; the comparative study of cultures raised questions over the Hellenocentrism characteristic of most Classical archaeologists, while advances in the study of mentalities promoted a critique of Eurocentrism.

The Greek and Roman worlds are now understood as melting-pots rather than *hochkulturen*. Under the influence of economic history, scholarly interest has shifted from a history of art in the narrow sense towards a wider history of productions. The better chronological grasp allowed by typological studies has served to advance studies on colonization, on long distance contacts and trade. Studies of the relations between centre and periphery, between Greeks, Roman and 'barbarians' have benefited from developments in the field of protohistory. It may indeed be said that the research program of classical archaeology has undergone a wide-reaching redefinition during the 1960s.

No longer considered as the elder of the sciences of antiquity, it is now one discipline among others, sharing resources and addressing issues similar to other strands of archaeology. In this context, such topics as the history of the landscape, or that of the movements of populations and the exchanges of goods and ideas, all open up new perspectives which contribute to extricate classical archaeology from its somewhat marginal position.

At the same time, advances in the iconography and the sociology of representations make it possible to cast a new light and reach a more satisfying understanding of otherwise well-known bodies of evidence. Thus, building on the strength of its centuries-long traditions as a science dedicated to Ancient art, Classical archaeology is developing now into to a human science in the full sense of the term.

Editor's Note

Archaeological work on Classical Period Europe continues apace. Indeed, every time a new construction is planned in Athens or Rome, classical archaeologists have a new excavation site to examine. Some recent works in the Classical world include: Alcock and Osborne (2012), Graves-Brown et al. (2013), Lulof (2014), Becker (2012).

See also: Christianity, Origins of; Chronology, Stratigraphy, and Dating Methods in Archaeology; Nationalism, Historical Aspects of: East Asia.

Bibliography

Alcock, Susan E., Osborne, Robin (Eds.), 2012. Classical Archaeology, second edition. Blackwell, Oxford.

Becker, Jeffrey A., 2012. Roman Republican Villas: Architecture, Context, and Ideology. University of Michigan Press, Ann Arbor.

Bérard, C., Vernant, J.-P., 1989. A City of Images. Iconography and Society in Ancient Greece. Princeton University Press, Princeton, NJ.

Bordein, A., Hölscher, T., Zanker, P. (Eds.), 2000. Klassiche Archaeologie, eine Einführung. Dieter Reimer Verlag, Berlin, Germany.

Elsner, J.A., 1955. Art and the Roman Viewer, the Transformation of Art from the Pagan World to Christianity. Cambridge University Press, Cambridge, UK.

Goldhill, S., Osborne, R. (Eds.), 1994. Art and Text in Ancient Greek Culture. Cambridge University Press, Cambridge, UK.

Graves-Brown, Paul, Jones, Sian, Gamble, Clive (Eds.), 2013. Cultural Identity and Archaeology: The Construction of European Communities, second edition. Routledge, London.

Lulof, Patricia S., 2014. Terracotta Architectural Sculpture in Classical Archaeology. In: Smith, Claire (Ed.), Encyclopedia of Global Archaeology. Springer Verlag, pp. 7273–7278.

Marchand, S.L., 1996. Down from Olympus. Archaeology and Philhellenism in Germany. Princeton University Press, Princeton, NJ.

Morris, I. (Ed.), 1994. Classical Greece, Ancient Histories and Modern Archaeologies. Cambridge University Press, Cambridge, UK.

Parslow, C.C., 1998. Rediscovering Antiquity, Karl Weber and the Excavations of Herculaneum, Pompeii and Stabiae. Cambridge University Press, Cambridge, UK.

Schnapp, A., 1997. The Discovery of the Past. Harry Abrams, New York.

Thomson de Grummond, N. (Ed.), 1996. An Encyclopedia of the History of Classical Archaeology. Greenwood Press, Westport, CT.

Classical Conditioning Methods in Psychotherapy

William C Follette and Georgia Dalto, University of Nevada Reno, Reno, NV, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Classical conditioning describes associative learning where stimuli are sometimes paired to produce clinical problems including most anxiety disorders. Extinguishing problematic responses that arise through classical conditioning is the focus of many psychotherapy procedures. This article describes the basic principles of classical conditioning, how understanding these principles helped clarify our understanding of the etiology of clinical problems, and how exposure-based treatments were developed to reduce or eliminate these problems. Recent cognitive explanations regarding the change process in humans are described. Finally, some attempts to add pharmacological adjuncts to improve the efficacy of exposure are reviewed.

One of the most common reasons people seek psychotherapy is because of anxiety or fear about some situation or object. There are many accounts for how these emotional states develop, but when people present for therapy it is often because their avoidance of these fearful situations interferes with some aspect of their desired role functioning. One of the most commonly used technical procedures in psychotherapy is exposure, and it is particularly effective in the treatment of these types of clinical presentations. The effectiveness of exposure hinges primarily on an understanding of classical conditioning, also called Pavlovian or respondent conditioning.

Classical conditioning is commonly defined as a process whereby a previously neutral stimulus comes to exert control over a response through pairing of the neutral stimulus with a stimulus that naturally (i.e., with no prior training) elicits the response. Since its accidental discovery by the Russian physiologist Ivan Pavlov in the late-nineteenth century, classical conditioning has been applied to many normal as well as clinically significant behaviors such as those characterizing anxiety disorders and addiction. Using principles derived from classical conditioning theory and contemporary extensions of the theory, behavior therapies have been developed to reduce the fear and anxiety responses that interfere with normal functioning.

The Procedures and Processes of Classical Conditioning

Classical conditioning is the most straightforward example of associative learning where stimulus–stimulus associations develop. According to the above description, the stimulus that naturally elicits the response is known as the unconditional stimulus (US), the naturally occurring response is known as the unconditional response (UR), the previously neutral stimulus is known (following conditioning) as the conditional stimulus (CS), and the response that comes under the control of the CS is known as the conditional response (CR). The process was first outlined by Pavlov during his study of the salivary response in dogs. Food in the mouth naturally produces a salivary response to aid digestion, but Pavlov noticed that his subjects began to salivate to the sound of a bell that accompanied the opening of the laboratory door during feeding time. In a series of now classic experiments, Pavlov demonstrated that the bell became a CS for the food (US) and that the CR of salivation was very similar to the natural salivary response (UR) elicited by the food.

For the better part of a century, contiguity between CS and US was commonly cited as an essential determinant of the successful conditioning of a response. Optimal conditions were said to be those in which the CS simply occurred immediately before onset of the US. However, more recently, the focus has shifted from contiguity between the CS and US to the amount of information supplied by the CS about the occurrence of the US (Rescorla and Wagner, 1972). For example, if a US occurs as frequently (or more frequently) in the absence of the CS as in its presence, very little conditioning will occur. The CS in this case is simply a poor predictor of the onset of the US. To exemplify the basic procedure by which a conditional response to a previously neutral stimulus is established, consider the following example. A small child with no fear of dogs is playing near a dog's food bowl. As the dog approaches, the unsuspecting child reaches out playfully to pet it. The dog (CS), presumably guarding its food, nips the child (US), causing minor tissue damage and producing a fear response in the child (UR; e.g., crying, retreating, etc.). If the child emits a fear response (CR) when next confronted with the sight of a dog (CS) and in the absence of another nip, conditioning is shown to have occurred. Earlier descriptions of classical conditioning also stated that the CR was generally topographically similar to the UR. That is, the CR looked the same. It is now recognized that the CR may look different from the UR depending on the species of the organism. In humans, the nature of the CR may vary substantially from the UR.

Generalization and Discrimination

Two other processes are relevant to a discussion of classical conditioning in clinical psychology: generalization and discrimination. In stimulus generalization, the CR occurs to stimuli that are similar in some way to the CS (but to which the response has never been conditioned). If the child was bitten by a Golden Retriever, but then comes to emit the fear response in the presence of German Shepherds and Chihuahuas, stimulus generalization is said to have occurred. In nonverbal organisms, generalization occurs on the dimension of some shared property of the stimulus. For example, if a dog were

 http://dx.doi.org/10.1016/B978-0-08-097086-8.21052-0

classically conditioned to a tone of a certain frequency, then similar sounding tones could also elicit a CR. In humans, language allows other generalization to occur along dimensions that may share no formal properties with the original CS but have acquired a relationship with the CS that can produce a CR. For example, only humans can show conditioning to the sight of a dog (CS) and also to the word 'dog' or the name 'Rover' or a leash, without prior conditioning once the animal has become a CS. In stimulus discrimination, the CR does not occur to stimuli that differ sufficiently from the original CS. In this case, the child would only fear Golden Retrievers, and would not emit that response to other breeds.

Exposure to the US or the CS: Habituation and Extinction

The process by which a fear response can be learned was described above. This same response can also be inhibited. One or both of the following processes can be used clinically to reduce fear responding. The first process is called habituation and it occurs through the repeated presentation of the US without the CS. The CS comes to have no predictive value for the occurrence of the US such that later presentations of the CS do not elicit a CR. The other process that can lead to inhibition of the CR is called extinction. Extinction is when the CS is presented without an accompanying US. A tone may be presented to a dog but food no longer follows. Eventually, the predictive value of the CS is diminished such that the CS no longer produces a CR.

Classical Conditioning and Avoidance in the Development of Psychopathology

The example of the child and the dog demonstrates the potential for classical conditioning to play a part in the etiology of a severe and persisting fear of dogs, or simple phobia, and it identifies a pathway through which many other fear or anxiety related disorders are thought to develop. Watson and Rayner (1920), in their classic Little Albert experiment, effectively conditioned fear in an infant. After allowing the child to play with a white laboratory rat (to which he showed no previous fear), the researchers followed presentation of the rat (CS) with a loud bang on an iron bar behind Little Albert's head. The loud noise (US) elicited a startle response from the child (UR), accompanied by crying. After a number of trials with this pairing, Little Albert began to emit a fear response (CR) to presentation of the rat alone, a response that also generalized to other white, furry objects. The clinical significance of this response (i.e., its persistence and the degree of functional impairment it produced) is not known, as Little Albert did not participate in follow-up studies and is thought to have died at age six (Beck et al., 2009). However, one of Watson's later graduate students, Mary Cover Jones, did demonstrate that phobic responses could be reduced by gradually exposing a child to the feared object (Jones, 1924).

People can also learn fear responses without directly experiencing the CS–US pairing, but can instead learn through observation. When someone observes a dog bite a person resulting in injury and distress, the observer may acquire a conditioned response. This is called vicarious learning. One can acquire even more subtle associations such as when a child is holding a parent's hand and notices a hand squeeze and protective movement when the parent sees a strange, large dog. Though many adults with specific phobias can recall a CS–US pairing, a large number of such people have no such recall. While it is possible that some of these reports are simply failures of memory, this also suggests that vicarious learning can account for some proportion of episodes of conditioning.

Not all stimulus–stimulus relationships are equally easily established (Garcia and Koelling, 1966). Certain modalities of associations are learned more readily than others. For example, pain is more readily associated with visual or auditory stimuli than gustatory stimuli (Rachman, 1991). Conversely, internal stimuli are more readily conditioned with gustatory stimuli. Thus, it is common for taste aversions to be acquired after an episode of nausea or vomiting (Bernstein, 1999).

Individual Differences

Not everyone who experiences a potentially traumatic event will go on to demonstrate the pathological avoidance or exaggerated startle responses characteristic of post-traumatic stress disorder (PTSD), nor indeed does everyone exposed to a fear-inducing US develop conditioned fear or a phobic response to the accompanying CS. Again, classical conditioning principles may help shed some light on this individual difference. The concept of 'conditioned inhibition' suggests that when the CS is a compound stimulus (e.g., all the myriad stimuli that make up the context in which an event occurs), the stimuli that signal the absence of the US will inhibit conditioned responding. Consider the example of a child who is with her mother when a bomb goes off on a bus. Because the presence of the child's mother has, through previous conditioning, come to signal safety (or the absence of danger), the child may not develop a conditioned fear response to other stimuli present at the time of the explosion (e.g., loud noises, buses, etc.). In this case, the child's mother has served as a conditioned inhibitor and effectively protected the child from developing a pathological fear response to the neutral stimuli.

The principles of stimulus generalization and discrimination can also help explain individual differences in psychopathology. When fears become pervasive (such as when a rape victim comes to fear all men), it is likely due to extensive stimulus generalization. Many things have come to participate in a stimulus class with the original CS because they are similar on some relevant dimension (e.g., they are all male). Alternatively, very circumscribed fears are the result of appropriate stimulus discrimination (as when an individual avoids the intersection at which he had a serious automobile accident but is still able to drive through all other parts of the city).

Individual differences in reactivity to potentially fear-inducing events also depend on the unique learning histories of the individual. One way histories differ is in terms of prior experience with stimuli. The Kamin 'blocking' effect states that a stimulus that is part of a compound stimulus will fail to produce conditioning if other stimuli in the compound have already been conditioned (Kamin, 1969). A boy who has come to emit a fear reaction at the sight of soldiers in uniform because they were present when he arrived home to find that his mother had been killed in an airstrike, may not come to fear news reporters who are present (in addition to soldiers) when

the rest of his family is killed in a similar attack 2 weeks later. The blocking effect is thought to occur because the second stimulus does not provide any new information about the US beyond that supplied by the original CS.

Avoidance

If classically conditioned fear responses reduce on their own when either the CS is presented without the US, or the US is presented without the CS, then why don't fear responses naturally ameliorate? In fact, some do (Staley and O'Donnell, 1984). Epidemiologic studies do show that children endorse a higher number of feared objects than do adults, though some fears develop later (Merckelbach et al., 1996). What accounts for the diminution of the feared objects as people get older? The answer is likely natural exposure that occurs during the course of childhood. If a child is showing a fear of a dog known to be friendly, parents or the dog owner may well slowly present the dog for petting by the child or may model petting after which the child does the same. This kind of natural exposure produces extinction and is likely why the prevalence of specific phobias decreases over time.

What accounts for when fear responses persist? A version of the answer was offered by O.H. Mowrer (1956) called two-factor learning theory. In its simplest form, the theory argues that fear responses are classically conditioned in roughly the way described above. It is the second factor that interferes with the natural reduction in the fear response: avoidance. Avoidance occurs when someone begins to encounter either the feared stimulus, the US, or stimuli that have become conditioned to elicit the feared response. As the fear rises, the person avoids contact with either the CS or US. This avoidance behavior is negatively reinforced, i.e., made more likely to occur (see operant learning) by the reduction or removal of fear or anxiety. The clinical implication of this avoidance is that the person does not experience either habituation or extinction, because the aversive stimulus in never repeatedly contacted so that habituation or extinction would occur. Going back to the simple example of the person fearful of dogs, when such a person sees the dog from some distance away, a mild aversive response may be noted. As the dog gets closer, the anxiety rises. If the person turns around and exits the situation, the anxiety or fear quickly reduces. This reduction in the aversive emotion negatively reinforces avoidance. In the short run, the aversive state is removed thereby strengthening the avoidance. Had the person continued to approach to dog in spite of the growing anxiety, the person would have come right up to the feared object, the dog, and not have experienced a bite. If the approach behavior continued in other episodes, the fearful response to dogs would have habituated. Though Mowrer's original formulation has been extended to include cognitive components to address how widespread avoidance of various stimuli can become, its treatment implications are still generally relevant, as will be discussed below.

Exemplars

People with PTSD show exaggerated startle responses and report intense fear reactions to many stimuli that were present at the time of the original traumatic incident (e.g., the sound of a helicopter for a combat veteran, the presence of crowds for a survivor of a mass shooting, darkness or the smell of alcohol for someone who was raped in an alleyway outside a bar late at night). In all of these examples, previously neutral stimuli have come to elicit a fear response very similar to that elicited at the time of the trauma. The stimuli that elicit the fearful response are external to the person.

Panic disorder (sometimes described as fear of fear) is characterized by intense fear responses to interoceptive (as well as external) cues such as a pounding heart, dizziness, shortness of breath, sweaty palms, or symptoms that may have preceded a full-blown panic attack in the past. In this case, the initial (and extremely unpleasant) panic attack can be thought of as the US, fear of dying may be the UR, the early symptoms or environmental conditions as the CS, and the fear response as the CR. Since panic attacks may be perceived as occurring randomly, any aspect of the environment may become a CS as might any unpleasant physical sensation.

In some cases, panic attacks may lead to the development of agoraphobia, literally a fear of public (open) spaces, but clinically a fear of being in a public place where one feels trapped or faces embarrassment should a panic attack ensue. The person tries to identify situational cues that will allow him or her to avoid a panic attack. Since the cues are difficult to detect, the person generalizes those situations broadly and avoids more and more activities that take place away from the relative safety of one's home.

Anxiety disorders in general rely on treatment components that derive from classical conditioning models. An individual with social phobia may learn to avoid all manner of social interactions or performance situations, because characteristics of these contexts have been associated with humiliating experiences in the past.

Certainly, addictive behaviors have many classically conditioned elements to them. The sight of needles, drug paraphernalia, the olfactory or visual cues, or the appearance of another drug user may elicit cravings in alcohol, opioid, or stimulant abusers.

Using Classical Conditioning in Psychotherapy

In addition to helping explain the etiology of many forms of psychopathology, an understanding of classical conditioning principles has given rise to several forms of behavioral psychotherapy. Beginning with Mary Cover Jones, 'counter-conditioning' involved the pairing of a feared stimulus with an appetitive (desirable) stimulus such that the response to the former was replaced with the response to the latter.

Sensitization

One line of therapy based on classical conditioning was developed to reduce behaviors desired by patients but that were recognized to be unhealthy or otherwise undesirable. One of the more common examples can be found in the history of the treatment of alcohol dependence where conditioned taste aversions are used to treat this difficult problem. As with many clinical problems, treatment can be understood from either a classical or operant conditioning perspective (or both). One

treatment strategy has been to classically condition the taste or smell of alcohol with nausea by using conditioned taste aversion procedures (Revusky, 2009; Lemere, 1987; Reily and Schachtman, 2009). As mentioned earlier, gustatory–nausea relations are more easily learned than some other sensory modalities. Two procedures are most common. In one, after the risk of alcohol withdrawal has been managed, the patient is given a taste of a preferred alcohol followed by an emetic drug (Lemere, 1987). Emetics induce nausea or vomiting. The notion is that alcohol when paired with the emetic will produce a response of nausea to the alcohol in the absence of the emetic. A variation on this approach is the use of disulfiram (Antabuse™) in alcohol treatment programs (cf Krampe et al., 2011; Jørgensen et al., 2011; Alharbi and el-Guebaly, 2013). Disulfiram is another commonly used drug that provides a similar learning history. Disulfiram interferes with the normal metabolism of alcohol resulting in a buildup of acetaldehyde. This can lead to nausea, vomiting, dizziness, and other unpleasant side effects when one drinks alcohol, thus creating a conditioned taste aversion.

Alcohol treatment programs have also tried to use aversive electrical stimulation paired with alcohol taste to create a conditioned aversion to alcohol as well. There is less evidence that the shock–alcohol pairing is clinically as effective when compared to the emetic–alcohol pairing, though the limited evidence is not consistent (Smith et al., 1997; Cannon et al., 1981; Lamon et al., 1977). There have been ethical objections to conditioned aversive therapies (Nathan, 1985; Wilson, 1987), but some patients find them a better alternative than other treatment programs or compared to continued addiction.

Conditioned aversion therapies have been applied to a variety of behaviors, the consequences of which can have deleterious effects for those who exhibit them, or for society. Aversion therapy has been adapted to smoking, cocaine use, sexual deviations including child molestation, and paraphilias, to name some examples (Rachman and Teasdale, 1969).

Covert Sensitization

Though not purely involving classical conditioning procedures, covert sensitization, also called covert conditioning, has been used to treat maladaptive approach behaviors including but not limited to problematic compulsions, sexual behaviors, alcoholism, obesity, nail biting, and smoking (Cautela and Kearney, 1986; Cautela, 1971). This procedure was developed as an alternative to the application of aversive stimuli such as an electric shock or an emetic (cf Krampe et al., 2011; Jørgensen et al., 2011; Alharbi and el-Guebaly, 2013). Covert sensitization makes use of aversive imagery that may be easier to produce outside of the laboratory or treatment environment. In brief, the subject is taught to imagine himself as about to engage in some desired but problematic behavior such as taking a drink. Vivid cues are provided to enhance the image. As the person is about to take the drink, he is instructed to imagine an uncomfortable feeling taking over that results in a further image of feeling profound nausea followed by vomiting. The vomiting is imagined to result in vomiting on oneself, the table, and perhaps others. The person is instructed to imagine odors, being embarrassed, and perhaps other aversive consequences. This process is repeated. In some versions, the person may be taught to imagine being about to take a drink, feeling the beginning of the nausea, and stopping, which is followed by a feeling of relief and pride that one resisted the drink.

It is not easy to measure how well the covert stimuli are produced or to assess how well or how willing one is able to produce the aversive covert stimuli. Nevertheless, this procedure can be useful when one behaves under the control of a reinforcer that is inappropriate or leads to undesirable consequences.

Systematic Desensitization

Joseph Wolpe was among the first to use the term 'systematic desensitization' for his approach to reducing fear responses to anxiety-producing stimuli (Wolpe, 1961). In this treatment, a relaxation response is trained in advance of exposure to the feared stimulus. When the feared stimulus is introduced, the client is instructed to engage in the relaxation response, which is believed to be physiologically incompatible with the fear response (Wolpe originally use the term 'reciprocal inhibition'; see Wolpe, 1958 for an early explanation of the intervention). Typically, there are three steps to this treatment. One is to identify a hierarchy of situations that are increasingly fear-provoking for the patient. In the case of acrophobia (a fear of heights), the patient and therapist list a series of such scenes from looking at a short step ladder to standing in front of this ladder to stepping on the first step. Additional scenes are constructed culminating with the most challenging scene that might be standing on the ledge of a tall building and looking down at the street below. In a commonly practiced version of this treatment, the scenes are sorted from the lowest arousing scene to the highest. Some number of sessions are used to teach the subject relaxation skills. Once those skills are learned, the therapist has the person imagine approaching the first element of the hierarchy until they notice some uneasiness, at which point they are told to use their relaxation skills until they become comfortable. This is repeated until that element of the hierarchy no longer produces anxiety or fear, and then the next scene is presented. This process is repeated until the client completes the hierarchy. Some have proposed that the process involves extinction, while others suggest habituation takes place (Watts, 1979). In either event, the previously avoided stimuli are contacted and the anxiety response is sufficiently reduced to allow normal functioning.

When treatment is designed as described above, it is often experienced as more palatable for both the client and therapist. As research has shown, it is not actually necessary for the hierarchy of scenes to be presented in a particular order; nor is it essential that the client have mastered a relaxation response; and some data indicate that *in vivo* exposure to elements in the hierarchy are perhaps more effective than imaginal techniques (see Marks, 1978 for a review). Thomas Stampfl introduced the technique of 'flooding,' in which the client is exposed to large doses of the feared stimulus and prevented from escaping until the fear response subsides (Stampfl and Levis, 1967). Contemporary treatments such as prolonged exposure for trauma and exposure and response prevention for obsessive-compulsive disorder (OCD) were built on this tradition of harnessing the power of classical conditioning to replace maladaptive responses with more adaptive ones. What all these

techniques have in common is that they involve exposing the client to the feared stimulus instead of allowing him or her to continue to avoid it.

Over the last 25 years, several difficult-to-treat problems have been successfully addressed by creative exposure procedures. Panic disorders have been treated by making use of interoceptive exposure where some of the symptoms of panic are produced but without the panic attack itself (Barlow et al., 1989; Barlow and Craske, 1989). Many who experience panic attacks become hypersensitive to normal physiological responses such that when they occur, fear of a panic attack ensues. In interoceptive exposure procedures, a variety of exercises are used to bring about some of those internal sensations so that the fear of a panic attack does not occur when some cues occur. For example, patients are taught to hyperventilate to experience some lightheadedness. Similarly, a patient may sit on a chair that spins around sufficiently to induce mild dizziness. A variety of exercises are utilized to expose patients to bodily cues that do not become panic attacks.

OCD is another clinical problem that has been usefully treated by exposure and response prevention (Franklin and Foa, 2011). In OCD, patients are exposed to that about which they obsess and are prevented from exhibiting the compulsive behavior that they use to reduce the obsessions. For instance, someone who obsessed about germs might be exposed to a dirty article of clothing for long periods of time and not allowed to wash his hands.

Debate exists with regard to the mechanisms by which exposure reduces fear and anxiety (McSweeney and Swindell, 2002). Traditionally, this process has been described as extinction, whereby the CR fails to occur following repeated presentation of the CS without the US. The CR is said to extinguish as a result of this procedure. The idea is that the link between CS and US is severed, so that the CS no longer predicts the US. According to Rescorla–Wagner model, this procedure would diminish the information about the US provided by the CS. However, McSweeney and Swindell (2002) examined the substantial body of literature available at the time and concluded that there is considerable evidence to suggest that the process known as extinction actually relies on the even more basic principle of habituation. Habituation is defined as "a decrease in responsiveness to a stimulus when that stimulus is presented repeatedly or for a prolonged time" (p. 364–365). When applied to the senses, habituation is known as 'sensory adaptation' and it is a process so pervasive that we often take it for granted (consider the experience of walking into a room with a strong offensive odor and suddenly realizing half an hour later that you can no longer smell it). The case presented by McSweeney and Swindell suggests that repeated or prolonged exposure to the CS will cause a decrease in the probability or likelihood of the CR through this process of habituation. It remains an empirical question whether habituation occurs to the CS through repeated or prolonged exposure even when the CS continues to be followed by the US.

Treatments based on classical conditioning principles also consider such effects as stimulus generalization and discrimination, blocking, and conditioned inhibition. The tendency of fears to generalize to additional stimuli can make treatment challenging, as many more stimuli than those involved in the original fear-inducing event may need to be targeted in treatment. On the other hand, one expects that treatment produces new learning (i.e., to not fear the stimuli used in exposure) that will generalize to additional feared stimuli.

However, the situation is complicated, as habituation and extinction appear to generalize less readily than the original conditioning (McSweeney and Swindell, 2002). Stimulus discrimination can be encouraged by training the individual to distinguish between the original feared stimulus and similar (but different) stimuli. This process may help prevent the generalization of the fear response following the original conditioning event. Blocking could potentially prevent new learning from occurring such that it may be more effective to conduct exposures to one feared stimulus at a time; pairing a reconditioned stimulus with one that is still feared may render the still feared stimulus redundant (i.e., it does not provide any new information about the situation).

The impact of conditioned inhibition on psychotherapeutic techniques informed by classical conditioning may be more complicated. While the presence of a stimulus that has come to signal safety may protect the individual from fear conditioning to begin with, such 'safety signals' may also impede the process of extinction/habituation during exposure. It is thought that safety signals (such as empty medication bottles or being accompanied by a significant other during exposure exercises) prevent the individual from fully contacting the feared stimulus so that when exposed to the stimulus in the absence of the safety signals, any apparent positive effects of the exposure disappear.

Cognitive Considerations

Certainly, exposure-based therapies are among the most robust intervention for the treatment of anxiety disorders. Exposure relies on some aspects of classical conditioning principles augmented by an operant understanding of the role of avoidance in the maintenance of anxious or fearful response. Two observations have stimulated theorizing about a cognitive explanation for how exposure works. First, not everyone who undergoes an exposure-based treatment shows improvement (Choy et al., 2007). This implies that our understanding of the dimensions of the CS or CR is not fully adequate. Second, the treatment of two difficult clinical problems, PTSD and OCD, has engendered theorizing about what happens in humans who can verbalize their experiences (Foa and Kozak, 1986). In PTSD, a wide variety of stimuli, not easily categorized, can produce fearful responding. In OCD, obsessions are cognitions and patients provide elaborate explanations for the source of the obsessions and the functions of the compulsions.

One of the more elaborated information processing/cognitive models was called emotional processing theory (EPT) (Foa and Kozak, 1991; Foa et al., 2006). EPT suggests that effective treatments for fear and anxiety, including exposure, produce their effects by introducing accurate information to alter existing, or create new, fear structures. A fear structure contains memories about sensory information about the feared situation, the avoidance behaviors and one's reactivity to the situations, and information about the interpretations relating the stimuli and responses. The basis of the prolonged exposure, an effective but not the only treatment for PTSD, is that fear

structures associated with the traumatic memory entail incorrect associations between stimuli and responses during the trauma and its subsequent meaning. The set of stimuli recognized and responded to as dangerous is too broad, meaning that fear is experienced to a wider set of stimulus conditions. The person also understands their responses as incompetent in indicating they cannot cope. This results in an inability to take in new experiences that would alter their emotional processing. Prolonged exposure putatively works by promoting alterations of the fear structures through systematic confrontation of trauma related stimuli (using imaginal exposure) with discussion of the experience in order to help disconfirm the problematic beliefs. The result is the modification of the problematic fear structure or the creation of another more adaptive one.

Not everyone agrees with the need for or the accuracy of this conceptualization and some suggest genetic and individual difference variables warrant attention in addition to remaining aware of advances in associative learning (e.g., Mineka and Thomas, 1999; Mineka and Oehlberg, 2008). In either case, as the verbal constructions of the fear-evoking stimuli get more elaborated, so too does theorizing. This perhaps follows from the natural language patients use to discuss fearful situations and responses to them, as well as the Rescorla and Wagner use of the concept of information contained in the relationship between the CS and US, again a use of language that promotes a more mentalistic interpretation of emotional responding and its treatment.

Future Directions

New technologies such as virtual reality are being applied to simulate *in vivo* exposure to otherwise difficult-to-create stimulus conditions (see Virtual Reality in Psychotherapy). For example, virtual reality goggles and sounds have been used to present combat situations for those who show clinically significant fear and anxiety responses following wartime experiences. This same technology has been used to present cues for high-risk situations such as gambling opportunities and smoking situations where the cues themselves may be conditional stimuli for these inappropriate behaviors or discriminative stimuli from an operant learning perspective. Virtual reality scenarios can be programmed for the treatment of generalized anxiety disorder where many different kinds of stimuli can elicit anxiety. Increasingly, technology is expanding into the therapy environment.

It is clear from the principles of change purported to underlie many of the empirically supported treatments for anxiety that classical conditioning principles have contributed significantly to the arsenal of interventions currently available to psychotherapists. However, the field would likely benefit from the further explication of the precise mechanisms that account for the improvements seen with these therapeutic techniques. Among these are the issue of whether extinction is a distinct process from habituation, how the unique learning history of the client can assist or impede treatment, and the extent to which classical and operant conditioning interact both in the etiology and the resolution of psychopathology.

Interesting pharmacologic interventions that possibly affect mechanisms of memory (re)consolidation are currently being researched. There is some evidence that propranolol, a β-adrenergic blocker, may reduce PTSD symptoms, but the ethics of such interventions have been debated (Donovan, 2010). There is evidence that the extinction of fear is mediated in the basolateral amygdala by the activation of N-methyl-D-aspartate (NMDA) receptors. The drug D-cycloserine can indirectly enhance that activity and has been shown to enhance fear extinction in exposure procedures in both animals and humans, though that effect may only be present if fear is lowered during a session (Smits et al., 2013). Given the ubiquity of anxiety disorders and the public health costs in terms of treatment, resource utilization, and lost productivity, it is likely that procedures to make extinctions procedures more available and cost-effective will continue to evolve.

See also: Behavior Therapy: Background, Basic Principles, and Early History; Exposure Therapies and Stress Inoculation: A Brief Overview; Obsessive-Compulsive Disorder across the Life Span; Panic Disorder and Agoraphobia Across the Lifespan; Phobias Across the Lifespan; Reinforcement, Principle of; Social Phobia across the Lifespan; Virtual Reality in Psychotherapy.

Bibliography

Alharbi, F.F., el-Guebaly, N., 2013. The relative safety of disulfiram. Addictive Disorders & Their Treatment 12, 140–147.

Barlow, D.H., Craske, M.G., 1989. Mastery of Your Anxiety and Panic. Graywind Publishing Company, New York.

Barlow, D.H., Craske, M.G., Cerny, J.A., Klosko, J.S., 1989. Behavioral treatment of panic disorder. Behavior Therapy 20, 261–282.

Beck, H.P., Levinson, S., Irons, G., 2009. Finding little Albert: a journey to John B. Watson's infant laboratory. American Psychologist 64, 605–614.

Bernstein, I.L., 1999. Taste aversion learning: a contemporary perspective. Nutrition 15, 229–234.

Cannon, D.S., Baker, T.B., Wehl, C.K., 1981. Emetic and electric shock alcohol aversion therapy: six- and twelve-month follow-up. Journal of Consulting and Clinical Psychology 49, 360–368.

Cautela, J.R., 1971. Covert conditioning. In: The Psychology of Private Events: Perspectives on Covert Response Systems. Academic Press, New York, pp. 109–130.

Cautela, J.R., Kearney, A.J., 1986. The Covert Conditioning Handbook. Springer, New York.

Choy, Y., Fyer, A.J., Lipsitz, J.D., 2007. Treatment of specific phobia in adults. Clinical Psychology Review 27, 266–286.

Donovan, E., 2010. Propranolol use in the prevention and treatment of posttraumatic stress disorder in military veterans: forgetting therapy revisited. Perspectives in Biology and Medicine 53, 61–74.

Foa, E.B., Huppert, J.D., Cahill, S.P., 2006. Emotional processing theory: an update. In: Rothbaum, B.O. (Ed.), Pathological Anxiety: Emotional Processing in Etiology and Treatment. Guilford Press, New York.

Foa, E.B., Kozak, M.J., 1986. Emotional processing of fear: exposure to corrective information. Psychological Bulletin 99, 20–35.

Foa, E.B., Kozak, M.J., 1991. Emotional processing: theory, research, and clinical implications for anxiety disorders. In: Safran, J.D., Greenberg, L.S. (Eds.), Emotion, Psychotherapy, and Change. Guilford Press, New York.

Franklin, M.E., Foa, E.B., 2011. Treatment of obsessive compulsive disorder. Annual Review of Clinical Psychology 7, 229–243.

Garcia, J., Koelling, R.A., 1966. Relation of cue to consequence in avoidance learning. Psychonomic Science 4, 123–124.

Jones, M.C., 1924. The elimination of children's fears. Journal of Experimental Psychology 7, 383–390.

Jørgensen, C.H., Pedersen, B., Tønnesen, H., 2011. The efficacy of disulfiram for the treatment of alcohol use disorder. Alcoholism: Clinical and Experimental Research 35, 1749–1758.

Kamin, L.J., 1969. Predictability, surprise, attention and conditioning. In: Campbell, B.A., Church, R.M. (Eds.), Punishment and Aversive Behavior. Appleton-Century-Crofts, New York.

Krampe, H., Spies, C.D., Ehrenreich, H., 2011. Supervised disulfiram in the treatment of alcohol use disorder: a commentary. Alcoholism: Clinical and Experimental Research 35, 1732–1736.

Lamon, S., Wilson, G.T., Leaf, R.C., 1977. Human classical aversion conditioning: nausea versus electric shock in the reduction of target beverage consumption. Behaviour Research and Therapy 15, 313–320.

Lemere, F., 1987. Aversion treatment of alcoholism: some reminiscences. British Journal of Addiction 82, 257–258.

Marks, I., 1978. Behavioral psychotherapy of adult neurosis. In: Garfield, S.L., Bergin, A.E. (Eds.), Handbook of Psychotherapy and Behavior Change: An Empirical Analysis, second ed. Wiley, New York.

McSweeney, F.K., Swindell, S., 2002. Common processes may contribute to extinction and habituation. Journal of General Psychology 129, 364.

Merckelbach, H., de Jong, P.J., Muris, P., van den Hout, M.A., 1996. The etiology of specific phobias: a review. Clinical Psychology Review 16, 337–361.

Mineka, S., Oehlberg, K., 2008. The relevance of recent developments in classical conditioning to understanding the etiology and maintenance of anxiety disorders. Acta Psychologica 127, 567–580.

Mineka, S., Thomas, C., 1999. Mechanisms of change in exposure therapy for anxiety disorders. In: Dalgleish, T., Power, M.J. (Eds.), Handbook of Cognition and Emotion. John Wiley & Sons Ltd, New York.

Mowrer, O.H., 1956. Two-factor learning theory reconsidered, with special reference to secondary reinforcement and the concept of habit. Psychological Review 63, 114–128.

Nathan, P.E., 1985. Aversion therapy in the treatment of alcoholism: Success and failure. Annals of the New York Academy of Sciences 443, 357–364.

Rachman, S., 1991. Neo-conditioning and the classical theory of fear acquisition. Clinical Psychology Review 11, 155–173.

Rachman, S., Teasdale, J., 1969. Aversion Therapy and Behaviour Disorders: An Analysis.

Reily, S., Schachtman, T.R. (Eds.), 2009. Conditioned Taste Aversion: Behavioral and Neural Processes. Oxford Press, New York.

Rescorla, R.A., Wagner, A.R., 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A.H., Prokasy, W.F. (Eds.), Classical Conditioning II. Appleton-Century-Crofts, New York.

Revusky, S., 2009. Chemical aversion treatment of alcoholism. In: Schachtman, T.R. (Ed.), Conditioned Taste Aversion: Behavioral and Neural Processes. Oxford University Press, New York.

Smith, J.W., Frawley, P.J., Polissar, N.L., 1997. Six- and twelve-month abstinence rates in inpatient alcoholics treated with either faradic aversion or chemical aversion compared with matched inpatients from a treatment registry. Journal of Addictive Diseases 16, 5–24.

Smits, J.A.J., Rosenfield, D., Otto, M.W., Marques, L., Davis, M.L., Meuret, A.E., Simon, N.M., Pollack, M.H., Hofmann, S.G., 2013. D-cycloserine enhancement of exposure therapy for social anxiety disorder depends on the success of exposure sessions. Journal of Psychiatric Research 47, 1455–1461.

Staley, A.A., O'Donnell, J.P., 1984. A developmental analysis of mothers' reports of normal children's fears. Journal of Genetic Psychology 144, 165–178.

Stampfl, T.G., Levis, D.J., 1967. Essentials of implosive therapy. A learning-theory based psychodynamic behavioral therapy. Journal of Abnormal Psychology 72, 496–503.

Watson, J.B., Rayner, R., 1920. Conditioned emotional reactions. Journal of Experimental Psychology 3, 1–14.

Watts, F.N., 1979. Habituation model of systematic desensitization. Psychological Bulletin 86, 627–637.

Wilson, G.T., 1987. Chemical aversion conditioning as a treatment for alcoholism: a re-analysis. Behaviour Research and Therapy 25, 503–516.

Wolpe, J., 1958. Psychotherapy by Reciprocal Inhibition. Standford University, Stanford, CA.

Wolpe, J., 1961. The systematic desensitization treatment of neuroses. The Journal of Nervous and Mental Disease 132, 189–203.

Classical Legacies

Norman Vance, University of Sussex, Falmer, Brighton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Latin and Greek social and political thoughts, particularly as preserved and discussed in the international language of Latin, influenced ideas of natural law, liberty, public virtue, and republic and empire as models of government, particularly in medieval and early modern Europe, colonial America, and imperial Britain. Plato was invoked for and against twentieth-century political absolutisms and the emerging sciences of anthropology and psychoanalysis drew on classical sources for ancient legend and evidences of primitive religion and ritual as source material and conceptual models. Plato, Aristotle, and, earlier, pre-Socratic, Greek philosophy have remained reference points for phenomenologists and other modern thinkers.

The philosophers, orators, and historians of ancient Greece and Rome, particularly Plato, Aristotle, the Stoics, Cicero, Livy and Tacitus, together with Virgil as the poet of empire and Homer as witness to the dawn of European civilization, are the main sources of the ideas and attitudes of classical antiquity which had a significant afterlife in the social and behavioral sciences.

Ancient Rome, the first and greatest beneficiary of Greek culture and thought, bequeathed its laws, its language and much of its ordered civilization to the west. The Greek city-states, particularly Athens and Sparta, the Roman Republic and the empires of Alexander the Great and of the Caesars, with all their strengths and weaknesses, provided later states and statesmen as well as political thinkers with contestable models of political organization. Particularly in the era of the French Revolution, invocations and discussions of political freedoms lost and won in Greek and Roman history reflected and expressed modern fears and preoccupations. The very vocabulary of political discourse continued to be strongly classical: 'politics', 'democracy', 'economics', and their equivalents in most European languages come from the Greek; 'republic', 'empire', and 'senate' are from Latin, and Latin continued to be the international language of intellectual exchange, including social and political theory, down to the end of the eighteenth century.

The classical architecture of town halls and other civic or national buildings across Europe and North America indicates continuing respect for classical traditions of public life. The more or less exemplary lives of classical heroes, and occasionally heroines, many of them recorded by Livy or collected by Plutarch, were frequently retold and revisited in the schoolroom. Parallels with celebrated classical achievements in word and deed were invoked in city squares and other public places by bronze or sculpted statesmen in togas, or soldiers on horseback on the model of Roman equestrian statues. The massive nineteenth-century Statue of Liberty in New York Harbor represents the classical goddess Libertas whose temple had stood on the Aventine Hill in Rome. Triumphal arches such as the Brandenburg Gate in Berlin (1788–91), the Arc de Triomphe in Paris (1806–36), and the Marble Arch in London (1825), all based on surviving classical models in Athens and Rome, celebrated modern victories and victors. The *fasces*, or bundles of rods bound with red thongs carried in procession to symbolize the power of ancient Roman magistrates, lent their symbolism and their name to Italian Fascism. The myths and rituals actually or arguably represented or implied in classical sources, particularly in the poets and in Greek tragedy, stimulated and helped to formulate new kinds of understanding in the course of the nineteenth century, particularly in the emerging sciences of anthropology and psychoanalysis.

Plato's influence as a moral and political philosopher, particularly in the *Republic*, was widely acknowledged in antiquity and supplied a basis for subsequent discussions such as Aristotle's *Nicomachean Ethics* and *Politics*. Plato and Aristotle between them provided later ages with nuanced conceptual vocabulary, models of dialectical, skeptical and logical inquiry, analytical tools and specific arguments. But the influence was sometimes indirect. After the fall of Rome in the fifth century, there was comparatively little direct access to Greek sources in Western Europe until the Renaissance. Aristotle, read in Latin versions, suffered the further misfortune of being too closely associated with medieval prescientific approaches to the natural world, in time vigorously repudiated by Francis Bacon (1561–1626) and others. But Plato and Aristotle lived on in the teachings and speculations of the Latin Church fathers such as Augustine. Works of Aristotle and other Greek writers which might otherwise have been lost were preserved in Arabic versions, eventually translated into Latin. The Greek-educated Cicero, reflective statesman, orator, lawyer, and exemplary Latin stylist, continued to be widely read and respected, and this helped to preserve awareness of Greek thought. His works on ethics (particularly *De officiis*), politics (*De republica*), and law (*De legibus*) preserved arguments, examples, and ideas which drew extensively on the Greek tradition.

Cicero and Law

Plato had developed a discussion of justice in the second book of the *Republic*, and Aristotle had considered it in the fifth book of the *Nicomachean Ethics* as an aspect of moral virtue. But Cicero was probably the most important direct source of ideas and aphorisms about justice and law invoked by later social thinkers, a major influence on the development of natural law theory in Thomas Aquinas's *Summa Theologiae* (1274) and the Dutch jurist Hugo Grotius' *De Iure Belli et Pacis* (1625). Cicero features much more prominently than Aristotle in the Swiss

philosopher and diplomat Emer de Vattel's *Droit de gens* (1758). Cicero's contention that *salus populi suprema lex*, the safety of the people should be the supreme law (*De legibus* 3.3.8) was invoked in John Locke's *Second Treatise on Government* (1690) as a 'just and fundamental' principle (13.158) and quoted as axiomatic in the radical republican Algernon Sidney's *Discourses concerning Government* (1698) (2.7) and elsewhere. Particularly admired in Britain and America, the phrase was adopted as the motto of the London Borough of Lewisham and of the State of Missouri. Drawing on the Stoic theory of a world soul or God as the ultimate sanction for law and morality, Cicero posited a natural law, grounded in the reason which was shared by God and man (*De legibus* 1.6,7), unwritten but innate in humanity (*non scripta, sed nata lex*), to which men might appeal, given the imperfections of formal law (*Pro Milone* 4). His principle that law, derived from the will of the gods, commanded what is honorable and forbade the contrary (*Philippics* 11.12.28), was quoted (in simplified form) by the thirteenth-century English jurist Henry de Bracton in his *De Legibus et Consuetudinibus Angliae*. It was repeated by Algernon Sidney (*Discourses* 3.10, 3.21) and invoked in William Blackstone's discussion 'Of the Absolute Rights of Individuals' at the beginning of his magisterial *Commentaries on the Laws of England* (1765–69) studied in rights-conscious America as well as in Britain (Sellers, 1998). It was a kind of backhanded compliment to Cicero's enduring reputation and influence as a theorist of political freedom under the law and opponent of arbitrary government that he was unfairly attacked by Thomas Hobbes (*De Cive*, 1642, 12.3) as one of the patrons of Greek and Roman 'anarchy', together with Aristotle and others.

Cicero, with Aristotle, contributed significantly to the developing political thought of colonial America (Botein, 1978). The patriotic Patrick Henry's famous rallying-cry 'give me liberty or give me death', spoken at the second Virginia Convention in 1775, was modeled on Cicero's alternatives of *libertas* and *mors* in *Philippics* 8.10.29. Thomas Jefferson specifically mentioned Cicero, along with Aristotle, as contributing to a tradition of 'public right' informing his draft of the American Declaration of Independence, and in his *Defense of the Constitutions of Government of the United States of America* (1787–88) John Adams invoked the authority of Cicero as a statesman as well as a philosopher who understood that the state needed to set bounds to ordinary factional behavior and find a balance of interests.

Livy, Tacitus, and Republican Virtue

In colonial America, as in Florence and Geneva in earlier days, political theorizing often fed on more or less stylized accounts of political life and public virtue in the ancient world. Livy's *History of Rome* provided Dante with examples of political virtue in *Monarchia* (*c.*1318), his Latin treatise on political philosophy. The Florentine Niccolo Machiavelli's *Discorsi* (1517), based on the first 10 books of Livy, which deal with the Roman republic, is in effect a treatise on the problems of establishing a durable republican polity in Florence. The radical Genevan Pastor Antoine-Jacques Roustan, who admired the work of Rousseau, resurrected the praise of classical republicanism as Livy described it in his *Dialogues entre Annibal et Tite-Live* (Geneva, 1789). Lucius Junius Brutus, prototype of republican virtue in that he delivered Rome from the monarchical tyranny of the Tarquins and helped to inaugurate the Roman Republic, perhaps the supreme Roman exemplar of uncompromising public duty (Livy 2.5) was celebrated in Voltaire's tragedy *Brutus* (1730) and became something of a cult figure in the early days of the French Revolution.

Tacitus began his *Annals* by saluting this Brutus as the father of Roman liberty. Laconic, epigrammatic, shrewdly pessimistic about Roman political culture as he had experienced it, Tacitus was much admired both for style and substance by the essayists Bacon and Montaigne in the sixteenth century. Tacitus lived under the Roman Empire and accepted that it was a permanent institution, but he was savagely critical of the more vicious emperors and claimed at the beginning of his *History* that Empire brought about the vanishing of great minds. His combination of moral indignation, political pragmatism, and rhetorical skill in conveying different points of view allowed him to be used on different sides in political argument. Leonardo Bruni, in his *Panegyric to the City of Florence* (1404), found ample support in Tacitus for his commitment to the republican form of government. He objected to Dante's admiration for Julius Caesar as founder of the Roman Empire and his consequent assigning of Caesar's assassins (Marcus) Brutus and Cassius to the deepest part of Hell in the *Inferno*. Though the foundation of Florence was usually associated with Caesar, Bruni went out of his way to attribute it to the republican general Sulla, at a time when the Roman people still flourished in an allegedly free and unconquered state. Other scholars, notably the Dutch humanist Justus Lipsius, read Tacitus rather differently. In his *Politicorum sive civilis doctrinae libri sex* (1589), much reprinted in Latin and translated into English in 1595, he insisted that Tacitus showed that the interests of peace were best preserved if all power was concentrated in one person: in an imperfect world, government by a single ruler was the best solution for a country in discord.

The success of the revolt of the Dutch United Provinces against imperial Spain helped to stimulate interest in Tacitus and his respect for Germanic peoples and heroes such as the German Arminius and the British Caractacus who had tried to resist the might of imperial Rome. The Dutch scholar Isaac Dorislaus was invited to be the first Professor of Ancient History at Cambridge in 1627, and gave two lectures on Tacitus. He drew alarming parallels with contemporary politics, suggesting that even in the Roman Empire sovereignty rested ultimately with the people and the absolutist claims of tyrannical emperors were illegitimate. The England of Charles I was not ready for such teaching: he was not allowed to finish the course of lectures. The full extent of his Tacitean radicalism became apparent in 1649 when he took part in the prosecution of Charles I which led to his execution. He was murdered by British royalists in the Hague on his return to his own country (Mellor, 2004).

Radical or 'red' Tacitism may subsequently have had a rather underground existence in England. But in his *Discourses concerning Government* (3.9) Algernon Sidney, champion of English republicanism, invoked Tacitus and his condemnation of the bad Roman emperors such as Nero 'and the rabble of succeeding monsters' as part of his argument against Robert Filmer's extreme monarchism in *Patriarcha*.

A generation later, Thomas Gordon, author with John Trenchard of the strenuously Whiggish *Cato's Letters or Essays on Liberty* (1720–23), drew on Sidney and on Tacitus, Livy, and Cicero in defense of freedoms of speech and warnings against the natural encroachments of power. Gordon devoted many years to a translation of *The Works of Tacitus* (1728–32), dedicated to Sir Robert Walpole, with various prefatory discourses critical of arbitrary power and empires ancient and modern, Roman, Turkish, and Spanish.

Classical Empire

Despite republican and Tacitean resistance to autocratic imperialism, the grandeur and longevity of the Roman Empire and the poetic anticipation and celebration of it in Virgil's *Aeneid* were influential in succeeding ages of colonial expansion and European empire-building. By the middle of the sixteenth century, the Portuguese had somehow contrived to acquire a fragile overseas empire stretching across the world from Brazil to Japan. The poet Luis Vaz de Camoens celebrated this achievement in his Virgilian epic *The Lusiads* (1572) which begins by inviting comparisons with classical adventurers overseas such as Virgil's Aeneas and the famous victories of the emperors Alexander and Trajan before settling down to epic celebration not of a single hero but of a whole race of them. Spanish imperial adventures and the Spanish Empire, at its height in the reign of Philip II (1556–98), had been sharply criticized by the Tacitean Thomas Gordon, but attracted Virgilian epic celebration in Alonso de Ercilla's *Araucana* (1569–89) which engages with the suppression of a revolt by Araucano Indians and the Spanish capture of a small valley in Chile (Quint, 1993: pp. 155–157). The successful Dutch revolt against Spanish rule facilitated seventeenth-century mercantile triumphs. This era stimulated writers such as the dramatist and national poet Joost van den Vondel (1587–1679) who hailed it in Virgilian terms (*Eclogue* 4) as a return to the Golden Age of classical mythology. But in 'Aen de Beurs van Amsterdam', his poem addressed to the immensely successful Amsterdam Stock Exchange, his allusion to Virgil's Dido and ruined Carthage, another great trading nation, hinted that such good fortune might not be permanent (Schenkeveld, 1991). Carthage, ultimately defeated by the Romans, was also associated with Britain, a mere nation of shopkeepers to the French who tended to see themselves in more heroic terms as modern Romans. Virgil was a useful starting for dreams of nation-building against the odds. As far away as the Ukraine Ivan Kotliarevsky burlesqued the *Aeneid* in his *Eneïda* (1798–1820), which links Aeneas and his defeated Trojan followers, ancestors of the imperial Romans, with the Zaporozhian Cossacks forcibly disbanded by Catherine the Great in 1775.

Victorian and Edwardian Britain, conscious of imperial hopes, glories and responsibilities, was also disposed to consider possible Roman parallels, though Edward Gibbon's endlessly reprinted *Decline and Fall of the Roman Empire* (1776–88) was a powerful reminder that even the most celebrated of empires eventually disintegrated. The Earl of Cromer, retired from colonial administration in Egypt, addressed the Classical Association on the subject of *Ancient and Modern Imperialism* (published in book form in 1910) and Sir Charles Lucas wrote thoughtfully about *Greater Rome and Greater Britain* (1912). But the parallels were not exact and uncomfortable aspects of the Roman experience such as overextension, corrupting luxury, and slavery, and unavoidable reliance on military force, were as much a warning as an inspiration.

The legacy of classical empire had always been tricky and ambiguous. The early church developed and spread within the Roman Empire and benefited from its communication systems. But it also suffered intermittent persecution, encouraging a certain caution. Early church fathers such as Tertullian, Irenaeus, and Origen prudently praised the stability, unity, and tolerance of the Roman empire (apart from a few persecutions under Nero and Domitian) and the fabled 'peace of Augustus', the *pax Romana* particularly associated with the emperor Augustus, seeing them as enabling conditions for the development of Christianity; Christianity in its turn, it was argued, had been a blessing to the empire, despite occasional calumnies against it. Augustine was much more critical in *The City of God*, written between 413 and 426 as the empire was crumbling, and so was the Spanish historian Orosius, whose *History against the Pagans* (417), written at Augustine's request, was the standard history of the ancient world available to the Middle Ages. But, even though Dante defended the Roman Empire as a model of government in his *Monarchia*, he was still content to follow Augustine and Orosius up to a point, endorsing their moral critique of the violent and tyrannical exercise of power. This is a theme in the *Inferno*, which places Alexander the Great among the Violent in the Seventh Circle of Hell, but Dante uncouples the violence of rulers from the imperial context which often sanctioned it. This partly ignores the informal moral and religious tradition stemming from Tacitus and Augustine which condemned military despotism and territorial aggrandizement and provided ammunition for later political commentators critical of aggressive imperialism in theory or in practice (Vance, 2011). There were formidable challenges to be negotiated by rhetorical champions of modern empire.

Benjamin Disraeli, novelist, fantasist, and spell-binding political orator, British Prime Minister in 1868 and again from 1874 to 1880, was just such a champion. In the 1870s he made political capital out of explicitly identifying Britain as a great imperial country and annexing the classical slogan *Imperium et Libertas* which he rather vaguely attributed to a great but unspecified Roman. The phrase was enthusiastically adopted as a motto by the Conservative Primrose League, founded in 1883, numbering over a million members at its peak and committed to maintaining the 'imperial ascendancy of the British Empire'. It can be traced back to Tacitus (*Agricola* 3) or possibly Cicero (*Philippics* 3.37), but it had long been misquoted and taken out of context. Though it was often translated as 'Empire and Liberty' Tacitus and Cicero had not intended to praise Empire as such but to draw attention to a seldom-resolved tension at the very heart of power politics: the need to uphold the rule of law without endangering civil freedoms (Hammond, 1963). Liberals such as W.E. Gladstone, Disraeli's great parliamentary opponent, four times Prime Minister, were not at all sure about empire, particularly empire on the grand scale as Disraeli imagined it. Gladstone was a more accomplished classicist than Disraeli and had read St Augustine as well as Tacitus and Cicero. He knew enough about Roman imperial history to be deeply worried about unchecked territorial acquisition. In speech after speech

in the 1870s he insisted that revival of the Roman imperial analogy, applied to contemporary adventures in Africa or Afghanistan, was morally and practically indefensible. No people, whatever their state of civilization, deserved to be subjected to military aggression to extend an overseas empire.

City and Satire

Theorizing and praising republican constitutions and city states was all very well, but actually living in the overcrowded cities of the ancient world such as Rome or Alexandria was often unpleasant. Greek and Latin pastoral poetry, much imitated in the vernacular literature of early modern Europe, had idealized the countryside and the lives of shepherds and country people to point a contrast with city life. Verse satire, harsher and darker, a Roman rather than a Greek literary form, took the critique of urban life and of public and private morals to much greater extremes. 'Imitations' or adaptations of Roman satirists such as Juvenal subjected aspects of the life of Paris or London to classical scorn. Nicholas Boileau's updated imitations of classical satire (1660–1708) enjoyed a considerable vogue in France. Samuel Johnson's poems 'London' (1738) and 'The Vanity of Human Wishes' (1749) are self-conscious imitations of Juvenal. Eighteenth-century Russia, influenced by the language and culture of France and encouraged by Peter the Great to be more receptive to western influences, turned to classical satire as mediated by the French as a new literary form. Prince Antiokh Kantemir (1709–44), pioneer Russian satirist and philosopher, Russian ambassador in London and then in Paris, wrote nine satires modeled on Horace and Boileau, first published in French translation and published in Russian only in 1762. He also translated Horace into Russian verse. He was one of the first Russian thinkers to invoke the metaphor of 'Liberty' as a heavenly goddess, as enshrined in the temple of Libertas on the Aventine Hill in Rome.

There was often a strong satirical component in classically influenced Utopian literature lampooning not just the politics but the intellectual and social fashions of the day. Fantastic imaginary voyages to the moon or the Isles of the Blest, as described by Lucian in the ancient world mingled with parodies of contemporary travels and voyage narratives to transport the reader to the surreal, black-comic invented worlds of François Rabelais' *Pantagruel* (1532) and *Gargantua*, (1534–35), Jonathan Swift's *Gulliver's Travels* (1726), and Voltaire's *Candide* (1759). In eighteenth-century Russia several utopian novelists followed the example of Lucian in colonizing the moon for satirical purposes. Mikhail Kheraskov's novels such as *Numa, or Flourishing Rome* (1768) described a utopian land curiously resembling his own country as a way of praising the Russia of Catherine the Great. On the other hand Prince M.M. Shcherbatov's *Journey to the Land of Ophir* (1783–84) uses his imaginary city of Peregab (Petersburg) to denounce Peter the Great's Westernization of Russia.

Plato to Popper

Plato was read in nineteenth-century Russia as well as everywhere else. His *Republic* included extended discussion of the education of the true philosophers who would alone be fit to be the Guardians or magistrates of the ideal state. With substantial supplementation from Cicero, particularly the *De officiis*, this contributed to recurring Renaissance considerations of the education of the prince or public official, such as Erasmus's *Education of a Christian Prince* (1516), Castiglione's much-translated *Book of the Courtier* (1528), and Thomas Elyot's *Boke named the Governour* (1531). Juan Luis Vives' *Declamationes Sullanae* (1520) approached the question of political leadership from another direction, presenting a fictional and rhetorical expansion of an actual historical situation by writing speeches for the Roman dictator Sulla which addressed questions of political principle and alleged political necessity. Plato's *Republic* also contributed to the construction of Thomas More's imaginary state in *Utopia* (1516) and, usually more indirectly, to the making of the countless utopian fictions that followed it.

Platonism and the quasi–religious, even mystical, Neoplatonism associated with the *Enneads* of Plotinus (*c*.205–270) had considerable influence on the development of religious thought in the early church and among the Cambridge Platonists in seventeenth-century England. It also influenced poetry and painting, particularly in renaissance Italy and in France. But it was not really until the nineteenth century that serious interest in Plato himself significantly revived among social and political thinkers. The legal reformer Jeremy Bentham had called him 'the master manufacturer of nonsense' in his posthumously published *Deontology* (1834), but (Demetriou, 2011) he was carefully studied by the Benthamite banker, politician, and historian George Grote, author of *Plato and the other Companions of Sokrates* (1865). John Stuart Mill, who was taught Greek at the age of three by his father in order to enable him to read Plato, admired his approach to truth through critical discussion rather than dogmatic assertion and reread the dialogues in order to review Grote's work.

Friedrich Hegel, convinced that reality itself was essentially dialectical, had responded positively to Plato's dialectical method, discussed also by the later Hegelian Herbert Marcuse in his *Reason and Revolution* (1941). The growing popularity of Hegelian idealism, particularly in nineteenth-century Oxford, encouraged renewed attention to Plato, translated by Benjamin Jowett of Balliol College who described Plato as the father of idealism. Plato was studied and cited by John Ruskin and others in support of late-Victorian moves to rethink the nature of the British state and state responsibility for its citizens.

There was something of a reaction against Plato in the twentieth century, though that could be seen more as reaction against totalitarian appropriations of Plato. In *Plato Today* (1937) the young Oxford tutor Richard Crossman, subsequently a senior figure in the British Labor Party and a Cabinet Minister, represented Plato as sharply critical of the political uses to which he had been put by Nazi and communist ideologues. After the Second World War, Karl Popper devoted the first volume of his *Open Society and Its Enemies* (1945) to attacking a repressively authoritarian version of Plato, though the main enemy was really Hegel rather than Plato (Bamborough, 1962).

Stoicism and Epicureanism

Christianity, influenced by Platonism, was also quite kind to the pre-Christian Stoic tradition stemming from the teaching of

Zeno of Citium (*c*.333–262 BC), particularly as it was represented by Seneca and, to a lesser extent, Epictetus. The rather grim determinism of Ecclesiastes in the Old Testament had a certain affinity with Stoic teaching though direct influence is unlikely. Seneca's brother is mentioned in the New Testament and Seneca and St Paul were contemporaries. The alleged correspondence between Seneca and St Paul, apparently accepted as genuine by Augustine and by Jerome, who thought it indicated that Seneca was a Christian, turned out to be a forgery, but there were some broad similarities in their beliefs. While Stoicism was never fully endorsed by Christianity, church fathers such as Tertullian and Augustine were attracted to aspects of it. The medieval philosopher and theologian Peter Abelard (1079–1143) was interested in Stoic ethics, particularly in his *Dialogus inter Philosophum, Judaeum et Christianum*. In the fourteenth century the Italian humanist Petrarch drew mainly on Seneca and Cicero for his *De Remediis Utrisque Fortunae*, a collection of philosophical dialogues arguing for Stoic calm. The English translation of 1579 was entitled *Phisicke against Fortune, as well prosperous as adverse*. In the sixteenth century John Calvin was a close student of Seneca and in his *Institutes* (3.8.9) tried to make a perhaps oversubtle distinction between Stoic *apatheia*, impassivity or willed indifference to suffering, and the Christian virtue of heroic endurance of God-sent suffering. The Dutch scholar Justus Lipsius' Stoic dialogues *De Constantia* (1584), the more popular French account of Stoic moral philosophy from the same period by Guillaume Du Vair, Bishop of Lisieux and Francisco de Quevedo's translation of Epictetus into Spanish (1635) as well as his own *Doctrina Estoica* (1635) all testify to enthusiastic neo-Stoicism in the early modern period (Morford, 1991). But this produced its own reaction. The materialism and extreme determinism associated with Stoicism alienated some Christian thinkers, all the more as the ever-controversial Spinoza had drawn on the Stoics for his view that true freedom could be found in reconciling oneself to the necessities of nature. In the *Encyclopédie*, Diderot prised Stoicism away from Christianity altogether by deeming it to be materialist, fatalist, and atheist. Seneca's reputation as a moral teacher suffered increasingly from his association with his former pupil Nero, but Seneca still commended himself to Ralph Waldo Emerson and American Transcendentalism in nineteenth-century America and the uncompromising moral teaching of Stoicism appealed to Victorian religious doubters such as Matthew Arnold anxious to cling on to morality without the metaphysical support of theology.

Epicureanism, the ancient rival concept of Stoicism, had a less distinctive afterlife as it was often misperceived as mere pleasure-seeking and did not commend itself to austere Christian moralists. Epicurus (341–279 BC) had been an influential teacher in Athens but very little has survived of his writings. Epicurean teaching and the atomic theory which Epicurus had adapted from Democritus have come down to us largely through Lucretius' didactic poem *De Rerum Natura*, known in antiquity but largely forgotten in the Middle Ages, rediscovered in the fifteenth century. Lucretius, imaginatively as well as intellectually committed to the beauties, energies, and intricate workings of the material universe, was variously respected as a scientific theorist and a nature poet, particularly in the increasingly scientific eighteenth century. He had assigned a purely nominal role to the gods, argued for human free will and attacked theological views of the world and the needless misery caused by religious superstition, which endeared him to freethinkers from Giordano Bruno to Percy Bysshe Shelley even if it angered theologians such as John Calvin. He featured prominently in the doctoral thesis of Karl Marx (1841) on the difference between the Democritean and Epicurean philosophy of nature, attributing to Epicurus an assertion of absolute freedom of spirit. Lucretius (5.113-25) had criticized the appalling punishment the gods gave to rebellious Titans such as Prometheus, comparing them with ground-breaking thinkers more generally, and this encouraged Marx's enduring admiration for Prometheus as the archetypal rebel-hero (Megill, 2002: pp. 75–76).

Anthropology and Psychoanalysis

Classical studies could provide sources and examples for new sciences. Classical authors and classical archaeology provided evidence, of disputed reliability and significance, about long-forgotten ancient cultures and cultic practices. The Greek historian Herodotus, often acclaimed as the 'father of history', acquired a new reputation in the nineteenth century as the 'father of anthropology'. His traveler's observant curiosity about local customs and traditions in Egypt and elsewhere anticipated that of Tacitus in Germany or the Greek geographer Pausanias in different parts of Greece many centuries later (Elsner, 2010). Pioneer scholars of primitive customs and cultures such as E.B. Tylor (1832–1917), first Professor of Anthropology at Oxford, needed to range widely across time and space to collect and compare their evidence, and Herodotus and Tacitus were useful early observers. Myth and legend preserved in very ancient sources such as Homer's epic narratives or Hesiod's systematic accounts of the gods of Greece in his *Theogony* could also be pressed into service and accommodated to larger theoretical structures. The fourteenth-century French *Ovide moralisé* was just one of many attempts, some going back to the early church fathers, to read classical mythology as coded religious and moral discourse, perhaps as versions or variants of biblical story, so that the mighty Hercules was claimed as a version of the Old Testament Samson and Deucalion's flood in Ovid's *Metamorphoses* was interpreted as a muddled recollection of Noah's flood in Genesis. Eighteenth-century freethinkers and Deists wondered whether all ancient myths reflected some underlying system of religious belief (Kuhn, 1956). Classical and biblical story both fed into Jacob Bryant's eccentrically syncretist *New System, or, an Analysis of Ancient Mythology* (1774–76), but (Blok, 1994) speculative work of this kind gradually yielded to attempts at a more rigorously scientific approach, notably by the German scholar Karl Ötfried Müller (1797–1840). Yet speculation was not dead: Tylor in *Primitive Culture* (1871) considered that the Greek stories of Odysseus or Orpheus going down into the underworld could be interpreted as solar myth, a departure at the end of day into a dark unknown, and noted there were comparable stories in other cultures, Polynesian, Scandinavian, and native American.

As Tylor acknowledged, alleged solar myths in classical and other sources, mocked and parodied even in their Victorian heyday, were particularly associated with the work of the

great Sanskrit scholar Friedrich Max Müller (1823–1900), a comparative philologist particularly interested in Indo-European languages and belief-systems. His philological approach was challenged by his arch-rival, the classical scholar, translator, poet, and polymath Andrew Lang (1844–1912), but Lang's much-reprinted *Custom and Myth* (1884), dedicated to Tylor, and other works such as *Myth, Ritual and Religion* (1887), again make use of classical materials, particularly Homer, Hesiod and Herodotus. Classical sources are also important in J.G. Frazer's massively ambitious comparative anthropological study of magic and religion, *The Golden Bough*. Between the first edition (1890) and the greatly expanded third edition (1911–15) he published a substantial anthropologically enriched edition of Pausanias in 1898. Frazer provided a starting-point for the so-called Cambridge Ritualists such as the classical scholars Jane Harrison (1850–1928) and Gilbert Murray (1866–1957) who tried to use anthropology and the study of ancient fertility rituals to elucidate ancient Greek culture and the origins of tragedy, a topic already approached in more speculative form in Friedrich Nietzsche's *The Birth of Tragedy* (1872). Frazer, like Karl Marx with Prometheus, encouraged symbolic readings of classical myth and legend. Sigmund Freud's 'Oedipus complex', first proposed in 1909, used the story of Oedipus murdering his father and marrying his mother in Sophoclean tragedy to universalize his famous theory of male psychosexual development (Leonard, 2013). Carl Jung, aware that in tragedies by Sophocles and Euripides Electra had killed her mother Clytemnestra, suggested a corresponding 'Electra complex' as an aspect of female psychosexual development in 1913, but Freud was unimpressed (Dervin, 1998). In a late essay, posthumously published in 1963, the psychiatrist Melanie Klein, influenced by Gilbert Murray's translation of Aeschylus' *Oresteia*, used the ancient legend of Agamemnon and his family, the doomed house of Atreus, to illustrate her views on fear, guilt, aggression, and hatred in the family (Alford, 1990).

The Classics and the Moderns

Classical literature continues to be a familiar reference-point in modern thought. In the second part of *Dialectic of the Enlightenment* (1947), Max Horkheimer and Theodor Adorno claimed the adventures of Homer's Odysseus by land and sea enacted the struggle for self-preservation and autonomy and the development of human subjectivity through repression and sacrifice and resistance to the primal forces of nature. Derrida's 'Plato's Pharmacy' (1968) and Ronald Dworkin (Dworkin, 2011) show new arguments about meanings and morals can still involve rereadings or invocations of Plato and Aristotle. The Aristotelian distinction between *techne* ('craft' or 'art') and *praxis* ('acting') (*Nicomachean Ethics* 1.1) is harnessed to critical social theory in Jurgen Habermas's *Theorie und Praxis* (*Theory and Practice*) (1973). Martin Heidegger's radical interrogation of the nature of Being in *Being and Time* (1927) begins by invoking Plato's *Sophist* and *Parmenides* and Aristotle's *Metaphysics* on the subject. But in attempting to probe deeper he pushes his enquiries further back in time, like Husserl before him, to the enigmatic fragments of relatively neglected pre-Socratic Greek philosophers, particularly Parmenides and Heraclitus, on whom he lectured in the 1940s and 1950s. He makes use of Parmenides' innovative attempt to deduce the character of what is from the requirements of thought and language and Heraclitus' celebration of constant flux and disturbance as the due and necessary order of things. His concerns with Phenomenology and with Truth leaned hard on the etymologies of the underlying Greek terms, associating 'phenomenon' with light (*phos*) to bring out the meaning of that which lies in the light of day or can be brought to light. The Greek term *aletheia*, usually translated 'truth', is the formal opposite of *lethe*, 'forgetfulness' or 'oblivion', what escapes notice, which enables an understanding of truth as something like 'unconcealedness'. Greek terminology and this etymological method have proved useful to others. Giorgio Agamben, who had attended some of Heidegger's seminars, discussed the Greek terms *poiesis* and *praxis*, 'making' and 'acting', in *The Man without Content* (1999), expanding on the distinction Aristotle makes between them (*Nicomachean Ethics* 6.4) by distinguishing between artistic or creative production on the one hand and action or immediate practical expression of the will on the other. It seems likely that contemporary thinkers will continue to exploit their classical legacy.

See also: Freedom: Political; Imperialism; Justice: Philosophical Aspects; Natural Law; Phenomenology: Philosophical Aspects; Plato (428 BCE–347 BCE); Political Thought, History of; Popper, Karl Raimund (1902–94); Republicanism: Philosophical Aspects.

Bibliography

Alford, C. Fred, 1990. Melanie klein and the "Oresteia complex": love, hate and the tragic worldview. Cultural Critique 15, 167–189.
Bamborough, R., 1962. Plato's modern friends and enemies. Philosophy 37, 97–113.
Blok, Josine H., 1994. Quest for a scientific mythology: F. Creuzer and K.Ö. Müller on history and myth. History and Theory 33, 26–52.
Botein, Stephen, 1978. Cicero as role model for early american lawyers. Classical Journal 73, 313–321.
Demetriou, Kyriakos, 2011. Studies on the Reception of Plato and Greek Political Thought in Victorian Britain. Farnham, Ashgate.
Dervin, Daniel, 1998. The Electra complex: a history of Misrepresentations. Gender & Psychoanalysis 3, 451–470.
Dworkin, Ronald, 10 February 2011. What Is a Good Life. New York Review of Books.
Elsner, Jaś, 2010. Reception of pausanias: from Winckelmann to frazer. Special Issue Classical Receptions Journal 2.2.
Hammond, Mason, 1963. "Res olim dissociabiles: principatus ac libertas": liberty under the Roman Empire. Harvard Studies in Classical Philology 67, 93–113.
Kuhn, Albert J., 1956. English Deism and the development of Romantic Mythological Syncretism. Publications of the Modern Language Association of America 71, 1094–1116.
Leonard, Miriam, 2013. Freud and tragedy: oedipus and the gender of the universal. Classical Receptions Journal 5, 63–83.
Megill, Allan, 2002. Karl Marx: The Burden of Reason. Rowan & Littlefield, Lanham, MD.
Mellor, Ronald, 2004. Tacitus, academic politics and regicide in the reign of charles I: the tragedy of Dr Isaac Dorislaus. International Journal of the Classical Tradition 11, 153–193.
Morford, Mark, 1991. Stoics and Neostoics: Rubens and the Circle of Lipsius. Princeton University Press, Princeton.
Quint, David, 1993. Epic and Empire: Politics and Generic Form from Virgil to Milton. Princeton University Press, Princeton.
Schenkeveld, Maria A., 1991. Dutch Literature in the Age of Rembrandt. John Benjamins, Amsterdam.

Sellers, M.N.S., 1998. The Sacred Fire of Liberty: Republicanism, Liberalism and the Law. Macmillan, Basingstoke.
Vance, Norman, 2011. Anxieties of Empire and the moral tradition: Rome and Britain. International Journal of the Classical Tradition 18, 246–261.

Relevant Websites

http://classics.mit.edu/Aristotle/nicomachaen.html – Aristotle, Nicomachean Ethics.
http://catdir.loc.gov/catdir/samples/cam032/98049660.pdf – Cicero, *De Republica, De Legibus*.
http://ancienthistory.about.com/library/bl/bl_text_livy_1.htm – Livy.
http://classics.mit.edu/Plato/republic.html – Plato, Republic.
http://classics.mit.edu/Tacitus/annals.html – Tacitus, Annals.This page intentionally left blank

Classical Mechanics and Motor Control

Pietro G Morasso, Istituto Italiano di Tecnologia, Genoa, Italy

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S. Mussa-Ivaldi, volume 3, pp. 1951–1955, © 2001, Elsevier Ltd.

Abstract

The article analyzes the application of basic principles of classical mechanics to the human body, namely biomechanics. The multireferential nature of the human movements is emphasized, considering the multiplicity of 'motor spaces' that describe human motion: tool space, end-effector space, joint space, and muscle space. In this framework, coordinate transformations and kinematics are analyzed, taking into account the causality constraint in the development of descriptive models, namely well-posed transformations from one kinematic descriptor to another in spite of the global redundancy of the motor system. Using the Lagrangian formalism, we derive the dynamic equation of a kinematic chain, which may model a human upper or lower limb. The complexity of the dynamics is considered, with emphasis on nonlinearity and anisotropy of the different dynamic components: inertial forces, Coriolis forces, gravity forces, external force fields, and muscle forces. The question is formulated about the necessity, for the brain, to be a 'conscious' expert of such mechanical complexity. Tentative answers are suggested while outlining a range of control schemes that are known in the repertoire of the motor control system: feed-forward, model-based control; stiffness control; and feedback control. As regards the last class of control mechanisms, the critical influence of delayed sensory afferences is taken into account, showing how intermittent, event-driven control in many situations may be more robust and more biologically plausible than classical, continuous-time feedback control.

As a physical system made of flesh and bones, the moving human body can be described in terms of classical mechanics. This concept is captured by the term 'biomechanics,' which deals with the relative and global motions of different body parts, the forces and moments transmitted by the muscles, as well as the associated flows of energy. The first scientist to address the study of the moving human body in this framework was Giovanni Alfonso Borrelli (De Motu Animalium, 1680) who applied to human physiology the mechanical concepts and formalisms developed by Galileo Galilei a few decades before. However, this is only part of the story, because goal-oriented coordination of the different body parts requires the brain, which is interfaced with the body in a bidirectional manner through muscle actuators and multimodal sensory channels (see Motor Control Models: Learning and Performance). The question then arises to which extent the brain must be and/or is fully aware of the complexity of biomechanical dynamics. Biology indeed provides a large number of examples, such as insects, where neural control is extremely simplified, in face of a very complex dynamics of the insect's body.

In fact, there is a range of possible answers, supported by a variety of experimental and theoretical evidence. On one extreme it is suggested that the brain develops internal representations of the mechanics of the body coupled with its environment (internal models Wolpert and Kawato, 1998; Kawato, 1999) through a process of motor learning, thus becoming implicitly an 'expert' in classical mechanics. On the opposite extreme there are examples of how purely reactive natural or artificial agents–agents that have no internal states but only direct connections between sensors and motors–can exhibit complex, adaptive behaviors. Consider, for example, the *Aplysia depilans* (Kandel and Tauc, 1965) or the vehicles of Valentino Braitenberg (1986). In the realm of artificial intelligence, Rodney Brooks (1986) introduced a computational framework (subsumption architecture) based on decentralized control and generalized the concept by proposing the idea of "intelligence without representation" (Brooks, 1991).

Moreover, even if one accepts the hypothesis that the brain is an expert in the mechanics of its own body, it still remains unclear where is the focus of the supposed internal models used by the motor controller. In much of the literature, it is taken for granted that 'internal models' means 'internal models of limb dynamics,' with the underlying assumption that dynamics is the most fundamental building block in the neural control of movements. However, this hypothesis is challenged by recent evidence in motor imagery made available by advanced techniques of brain imaging. It appears indeed that the brain activity recorded during imagined movements ('motionless' motor actions) is quite similar to the activity of actual movements, except for the specific muscle activation patterns conveyed in the final common pathway of the motor controller (Jeannerod, 2001; Kranczioch et al., 2009; Munzert et al., 2009). This suggests that although dynamics is likely to play an important role in motor control, on top of it there is a computational layer that deals with 'synergy formation' independently of forces and masses that are experienced only during real, overt movements.

Synergy Formations: Motor Redundancy, Coordinate Systems, and Kinematics

A significant computational challenge for the brain comes from the need to perform changes of representation – or more technically, coordinate transformations – between the description of a task and the specification of the body motions. Tasks, such as hitting a ball with a racket, are described most efficiently and parsimoniously with respect to some fixed reference points in the environment (frequently called' endpoint'). In the example, the racket is the site at which one interacts with the environment. The position of the racket

is fully determined by six coordinates. These coordinates may be measured with respect to three orthogonal axes originating, for example, from the shoulder. Then, a position in endpoint coordinates may be specified as a point $\vec{p} = (x, y, z, \theta_x, \theta_y, \theta_z)$ in six dimensions. The subset of coordinates (x, y, z) determines a translation with respect to the orthogonal axes; the other subset of angular coordinates $(\theta_x, \theta_y, \theta_z)$ determines an orientation with respect to the same axes. Consistent with this notation, a force in endpoint coordinates, for example originating from the interaction of the racket with a ball, is a vector with three linear and three angular components, $\vec{F} = (F_x, F_y, F_z, \tau_x, \tau_y, \tau_z)$.

A different way of describing the position of an arm is to provide the set of joint angles that define the orientation of each skeletal segment either with respect to fixed axes in space or with respect to each other. Joint angles are a particular instance of generalized coordinates. According to the standard definitions of analytical mechanics, generalized coordinates or degrees of freedom (DoF) are independent variables, which are suitable for describing the kinematic and dynamics of a system (see, for example, Goldstein, 1980). Once we have defined a set of generalized coordinates, we may also define a set of corresponding generalized forces. For example, if we use joint angles as generalized coordinates $\vec{q} = (q_1, q_2, \ldots, q_n)$, the corresponding generalized forces are the torques generated at each joint $\vec{\tau} = (\tau_1, \tau_2, \ldots, \tau_n)$; n is the number of DoFs.

Movements are executed by the central nervous system activating a multitude of muscles. The motor system is indeed highly redundant at least at two levels: the number of DoF is much greater than the number of variables that describe a task, and the number of muscles is much greater than the number of DoF. Moreover, muscles are composed of a large number of motor units (collections of motor fibers innervated by the same motor neuron) and their activation patterns have some degree of independence, although the overall recruitment process is characterized by Hennemann size principle (Henneman et al., 1965). This means that each task-oriented action can be conceived, imagined, coordinated, and activated in an infinite number of ways. In spite of this, and on top of an intrinsic variability that characterizes even the most practiced skills, the observed motor patterns, at both kinematic and muscular levels can be described by a small number of components, somehow correlated with the relevant task variables.

In general, we may say that coordination of movements or synergy formation is a multireferential problem that involves multiple 'spaces' at the same time: task space, end-effector space, joint space, and muscle space. The dimensionality of the spaces grows from the task space to the joint and muscle spaces, and such dimensionality imbalance is the source of what is known as motor redundancy. In each space we can identify motion variables and effort variables. Causality, in the sense of well-posed nonlinear mappings, runs in opposite directions as regards movement and effort:

- Joint space → End-effector space → Task space (motion variables)
- Task space → End-effector space → Joint space (effort variables)

This means, for example, that an elementary joint rotation pattern can be uniquely mapped into an elementary motion of the end-effector and then into a motion in the task space (for example, the motion of a handheld tool). In contrast, an effort applied by means of a tool in task space can be uniquely mapped into an effort of the end-effector, and then an effort of the joints. The latter causal chain can be considered an inverse model because it makes it possible to compute the effort that is needed for obtaining a desired effect at the end of the kinematic chain; the former chain is a direct model, in the sense that it makes it possible to predict the task-relevant consequence of delivering a coordinated motor pattern. A pair of direct/inverse models is considered 'inseparable twins' for skilled motor control (Kawato and Wolpert, 1998). More technically, the transformations quoted above can be expressed by means of the Jacobian matrices of the corresponding nonlinear mappings:

$$\begin{cases} dq_j \rightarrow dp_e = J_j\, dq_j \rightarrow dp_t = J_t\, dp_e \\ df_t \rightarrow df_e = J_t^T\, df_t \rightarrow d\tau_j = J_j^T df_e \end{cases} \quad [1]$$

where dq_j is an elementary motion in the joint space, dp_e is the corresponding motion in the endpoint space, and dp_t is the motion in the tool space. J_j, J_t are the two Jacobian matrices that describe the mappings between the joint/endpoint space and the endpoint/tool space, respectively. τ_j, f_e, f_t represent the corresponding efforts. The 'T' symbol implies that the corresponding matrix is 'transposed' (the rows are exchanged with the columns), and this is required by the duality between effort and motion and the fact that work is an invariant when passing from one space to the other.

Some experimental studies (see Morasso, 1981; Flash and Hogan, 1985) have suggested that actions are planned by the brain in terms of endpoint coordinates. For example, one can mentally formulate (and execute) commands such as "move the hand 10 cm to the right" without being concerned with the set of muscle commands that are involved with this action. However, once one has decided a plan of action, one must somehow choose how to coordinate the joint rotations, which muscles to activate, and in what temporal order. The latter part of the plan (generation of muscle synergies) is only operant in overt movements, whereas the former part (generation of joint synergies) must be operant both in covert and overt actions.

Dynamics

The well-known Newton's equation ($\vec{F} = m\,\vec{a}$) tells us that a point mass (i.e., a point of negligible size and with a mass m) will accelerate when a force vector is applied to it, with an acceleration magnitude that is directly proportional to the force and inversely proportional to the mass. Such a statement expresses the direct dynamic problem: from the cause (the applied force) to the effect (the acceleration and, ultimately, the actual motion by integrating the differential equation). Moreover, the same equation can be read in the opposite direction, expressing the inverse dynamic problem: from the desired motion (the time course of the acceleration vector) to the required effort (the time course of the applied force).

If the object's size is not negligible, force vectors can be applied to many points of the object (for example, many

muscles can apply forces to different parts of the same bone); in this case, Newton's equation is split into two components, one related to the motion of the center of mass (CoM) of the object and the other to the rotation around an axis through the CoM:

$$\begin{cases} \vec{F}_{\mathrm{R}} = m\dfrac{\mathrm{d}\vec{v}}{\mathrm{d}t} \\ \vec{M}_{\mathrm{R}} = I\dfrac{\mathrm{d}\vec{\omega}}{\mathrm{d}t} \end{cases} \qquad [2]$$

Here $\vec{F}_{\mathrm{R}}$ is the resultant vector of all the forces applied to the object; $\vec{M}_{\mathrm{R}}$ is the resultant moment; $\vec{v}$ is the velocity vector of the CoM; and $\vec{\omega}$ is the angular velocity of the object. The human body is usually modeled as a set of linked rigid bodies, one for each body part. For example, a limb like the arm can be modeled as a kinematic chain with three links, namely the humerus, the forearm, and the hand, and several DoF. The dynamic equations of such a system can be derived by generalizing the cardinal equations above, that is, by writing the pair of equations for each body segment and taking into account the constraints introduced by the joints. This is the so-called Newton–Euler recursive algorithm, which makes it possible to derive the distribution of forces in a kinematic chain given precise measurements of the kinematics (joint angles, angular velocities, and accelerations) and the endpoint forces. It is an efficient algorithm of dynamic computation, typically used in gait analysis, but is somehow difficult to interpret.

A more compact form is provided by the Lagrangian formalism, which focuses on energy rather than force. For any complex dynamic system, with scleronomous constraints (constraints in which time is not an explicit variable), the following Lagrangian function can be defined: $L(q,\dot{q}) = K(q,\dot{q}) - P(q)$, where K is the kinetic energy and P the potential energy possessed by the system at any given instant of time. Both elements of the Lagrangian function have an additive property, that is, they are computed by adding the kinetic and potential energies of each link, respectively. The dynamics of the kinematic chain can then be represented in a compact way by the following equation:

$$\tau_m = I(q)\ddot{q} + C(q,\dot{q})\dot{q} + G(q) + J_{\mathrm{e}}^{T}F_{\mathrm{e}} \qquad [3]$$

The n-dimensional vector on the left-hand side of the equation is the set of torques (or generalized forces) that must be applied to the joints in order to realize a given law of motion, that is, a given time course of the joint rotations $q = q(t)$. The joint torque vector is determined by the activation of the actuators that control the kinematic chain, namely the muscles in the biological case. The equation says that such actuator forces must take into account four contributions: (1) inertial forces, proportional to acceleration ($I(q)$ is the $n \times n$ inertial matrix of the kinematic chain); (2) Coriolis forces, proportional to angular velocity; (3) gravity forces, which only depend on position; and (4) external forces (if the kinematic chain interacts with the environment through the end-effector, otherwise this contribution is null). In discrete, point-to-point movements, where speed is generally bell-shaped (Morasso, 1981), inertial forces are predominant in the initiation and termination phases, whereas Coriolis forces are predominant in midflight. For very slow movements (or in the case of imagined movements that do not involve any kinematic energy), inertial and Coriolis forces can be neglected and thus the brain can only focus on quasistatic elements: gravity and external forces.

The inertial matrix generally is not proportional to the identity matrix and this implies that the inertial properties are not isotropic but change with movement direction. The anisotropy of the matrix can be well illustrated in geometric terms if we consider the ellipsoid that relates a torque vector to the corresponding acceleration vector. This ellipsoid has a specific orientation in space and its major axes coincide with the eigenvectors of the matrix. In the case of the arm, it can be shown that the ellipsoid is oriented with its major axis approximately aligned with the forearm (Mussa-Ivaldi et al., 1985). This means that the 'apparent inertia' perceived by an external agent interacting with the subject's arm is not constant but varies with the direction of interaction and has a maximum value if the interaction direction is aligned with the subject's forearm. Another consequence of such inertia anisotropy is that in general if a force is applied to the end-effector, the resulting acceleration is not collinear but there is a deviation between force and acceleration, in contrast with the dynamics of the point mass.

Feed-Forward, Stiffness, and Feedback Control

The brain can affect the time course of vector τ_m in eqn [3] by selecting appropriate muscle activation patterns, according to a control scheme. In particular, we will briefly touch upon three main classes: feed-forward, feedback, and stiffness control schemes.

We should take into account that muscles are not pure force or torque actuators but have a complex internal dynamics that links chemical, electrical, and mechanical variables. Nonetheless, the three-element muscle model proposed by Hill (1938) is generally considered an acceptable approximation: it includes a nonlinear contractile element, a series, and a parallel elastic element. The essence is that muscles can generate force but with an intrinsic mechanical impedance. In particular, the elastic component of impedance, that is, stiffness, is a function of both muscle length and neural drive level. Moreover, we must take into account that muscles are arranged in antagonistic sets around the joints of the body: the forces of antagonist muscles are subtracted from each other, whereas the stiffness values are added. Therefore, the brain can choose reciprocal activation patterns if it is concerned with the production of torque around a joint, or coactivation patterns if it needs to affect the joint impedance. By properly tuning the two activation paradigms, the brain can control the torque τ_o delivered at the joints, the corresponding joint stiffness K_{j}, and the reference or equilibrium trajectory $q_o(t)$:

$$\tau_m = \tau_o + K_{\mathrm{j}}(q - q_o) \qquad [4]$$

τ_o, q_o, and K_{j} are functions of the neural drive and of the mechanical state of the body $(q, \dot{q})$. Therefore, the overall dynamics described by eqn [3] implies a bidirectional interaction between the neuromuscular system (left part of the equation) and the body–environment system (right part of the equation).

Suppose now that a movement must be produced with a desired law of motion $q = q_d(t)$. In principle, the brain can achieve the plan by means of three main control mechanisms:

1. *Feed-forward control*, based on an accurate internal model of the limb dynamics and/or external disturbances:

$$\begin{cases} q_o = q_d \\ \tau_o = I(q_d)\ddot{q}_d + C(q_d, \dot{q}_d)\dot{q}_d + G(q_d) + J_e^T F_e(q_d, \dot{q}_d) \\ \quad = \text{Model}(q_d, \dot{q}_d) \end{cases} \quad [5]$$

   In the ideal case, where the internal model is flawless and there is no noise, then applying feed-forward control will result in the actual movement $q(t)$ being identical to the planned one and the contribution to the overall dynamics of stiffness will be null. According to such a scheme, the brain generates neural commands in anticipation of the actual movement, on the basis of interacting neural networks that replicate the external dynamics of the body–environment system. Of course, it is not conceivable that such internal model is innate because the dynamics of the body–environment system is not constant but varies with age, loading conditions, and task-dependent environmental conditions. Therefore, internal models must be the result of a learning process. This has been and still is the subject of many studies, beginning with the seminal work of Kawato and Gomi (1992). Nonetheless, the acquired internal model will be only an approximation and there will be noise, thus causing a discrepancy between the planned and the real trajectory: $\Delta q(t) = q_d(t) - q(t)$. This is where joint stiffness comes into play, moderating the effect of noise and modeling error: $K_j\,\Delta q(t) = \Delta\text{Model}(q, \dot{q}) + \text{noise}$.

2. *Stiffness control*, based on a high level of coactivation. This control strategy does not require any internal model of the body–environment dynamics and regards such dynamics as a kind of disturbance or noise to be 'resisted' by muscle stiffness. This strategy is related to the equilibrium point hypothesis (Asatryan and Feldman, 1965; Feldman, 1966; Bizzi et al., 1976, 1991, 1992; Feldman and Levin, 1995), which assumes that in the control of movement the brain is only concerned with the equilibrium between internal and external force fields and compensates the disturbances due to dynamics by means of muscle stiffness. In order to be effective, the control scheme requires sufficiently high values of stiffness, and the question has been raised to which extent and/or in which conditions this can occur. Also in the case of the stiffness strategy, a learning process is required, because the brain must learn the stiffness values that are required in given conditions. However, this is much simpler than learning a full-blown internal model of body–environment dynamics. Also the run-time control is much more complex in feed-forward than in stiffness control: in the former case the brain must produce a very detailed profile of control signals, whereas in the latter it must simply 'stiffen up' to the required level before the movement start and then generate the equilibrium patterns. On the other hand, a stiffness strategy is energetically more expensive because it requires high coactivation levels.

3. *Feedback control*. Both feed-forward and stiffness control are open-loop strategies, which do not exploit the sensory feedback coming from proprioception or other types of receptors of body movement. In feedback control paradigms, the motor control commands originate from a comparison between ideal or reference values of body motion parameters and the actual values transmitted by the sensory channels, namely the explicit computation of a motor error signal. There are at least three different ways in which such error signals can be used by the neural controller: (1) reflex control, which reacts 'one-shot' to an unpredicted disturbance; (2) sustained, closed loop, continuous-time control, similar to engineering servomechanisms; and (3) intermittent, closed-loop control, such as the balancing actions in the manual stabilization of an inverted pendulum.

It is also possible to conceive mixed strategies. For example, feed-forward model-based control can compensate for only part of the dynamics, such as the static component due to gravity, which does not depend on the speed of the movement. In this case, stiffness control must compensate for only part of the unaccounted dynamics and a lower level of cocontraction is necessary.

Stiffness control is a kind of feedback control because the driving force is proportional to the mismatch between the planned and actual position of the different joints. However, in stiffness control the mismatch is not computed explicitly on the basis of feedback sensory signals, but implicitly by the viscoelastic properties of muscles. Therefore, it is an instantaneous feedback. Explicit feedback control, particularly in the case of continuous-time feedback, is affected by a major problem: feedback sensory signals are far from instantaneous. Summing up the propagation delay and the contraction delay in the segmental reflex loop, the functional delay is over 100 ms. In control engineering, it is well known that delays of this order of magnitude severely limit the frequency bandwidth and the feedback gain that can be used for compensating internal/external disturbances without triggering instability. Therefore, explicit feedback control tends to be a mechanism of last resort for unpredictable disturbances or for dynamic tasks that a subject did not have enough time to practice.

Evidence of Feed-Forward, Model-Based Control

Strong experimental evidence for the biological and behavioral relevance of internal models has been offered by experiments that involved the adaptation of arm movements to a perturbing force field generated by an instrumented manipulandum (Sabes et al., 1998; Shadmehr and Mussa-Ivaldi, 1994). This corresponds to the external force of eqn [3]. In order to investigate the possible emergence of feed-forward control, the external force must be deterministic and predictable but characterized by a dynamics which was completely unknown to the subjects. A typical example is a curl viscous field, which disturbs point-to-point reaching movements with a force which is proportional to the instantaneous speed and is directed perpendicularly to the direction of the speed vector. People are quite unlike to have experienced such kind of dynamics, which tends to

deviate sideways the movement to the target. Thus, in the initial exposure to the field only muscle stiffness operates in such a way to limit the entity of the motor error. However, with practice people can recover the original path and kinematics and consolidate such knowledge (Brashers-Krug et al., 1996). Since the movements are rather fast, at least in the first part there is no time for feedback to influence the mechanism that compensates the disturbing field. Therefore two possible explanations must be considered: stiffness control and feed-forward control. The former explanations can be ruled out for two reasons: (1) no evidence has been found of the cocontraction that would be required by a substantial increase of stiffness; (2) if the force field is unexpectedly turned off (catch trials) after the subject learned to compensate the disturbance, the trajectory is distorted in a mirror way with respect to the field. One would not expect such counterdeformation in the absence of disturbance if the subject was exploiting stiffness; in contrast, the recorded behavior is consistent with an anticipatory modification of the term $\tau_o(t)$ in eqns [3] and [4], incorporating the acquired knowledge of the disturbing field.

Evidence of Stiffness Control

Suppose that an instrumented manipulandum, similar to the one mentioned in the previous section, is programmed in such a way to generate a divergent force (DF) field during point-to-point reaching movements. If a subject moves along the ideal straight path with a bell-shaped speed profile in accordance with spontaneous human behavior, then the field is null. If the hand deviates laterally, the field pushes the hand sideways proportionally to the lateral error. In other words, the simulated dynamic environment is unstable.

The critical point is the relation between endpoint stiffness and the rate of growth of the DF, which is a kind of negative stiffness. Endpoint stiffness can be represented geometrically by means of an ellipsoid (or ellipse in the case of planar movements) in a similar way to the endpoint inertia of the arm. The difference is that the former ellipse can be modulated by neural activation patterns whereas the latter ellipse cannot be, for obvious physical reasons. The stiffness ellipse that can be recorded with spontaneous tonic activity of arm muscles is rather small and is oriented approximately with a polar pattern (i.e., is aligned with the line that joins the hand to the shoulder: Mussa-Ivaldi et al., 1985). If all the arm muscles increase proportionally their activity level with respect to the basis level, the ellipse size will increase without affecting its orientation. Rotation of the ellipse requires a more sophisticated rearrangement of the relative activation patterns of the muscles.

Coming back to the DF experiment, if the lateral rate of growth of the field is smaller than the lateral component of the stiffness ellipse in the spontaneous activity level of the muscles, then the destabilizing effect of the field will be compensated by the natural muscle stiffness, allowing the subject to reach the target, although with some error. However, if the DF is stronger than such critical value, the subject will 'fall' on one side or the other and will be unable to reach the target. Indeed this was found by Burdet et al. (2001), but they also found that subjects can learn to compensate for the instability after a suitable number of trials. In these experiments, no miscompensation was observed during catch trials, thus ruling out the possibility of a feed-forward compensation strategy. In contrast, it was found that in the course of learning the subjects managed to tune coactivation patterns of the muscles in such a way to obtain a stiffness ellipse of sufficient size and oriented optimally, namely aligned with the direction of instability. In other words, a stiffness strategy can be used by the brain in suitable experimental conditions and can be learned by human subjects.

Evidence of Intermittent Feedback Control

In the DF experiment described in the previous section, human subjects demonstrate that a stiffness strategy can be learned and applied for carrying out an unstable task. However, this is not a *universal* solution to tasks of this kind. There are unstable tasks where a stiffness strategy is impossible due to the nature of the task or the properties of human tissues. For example, when balancing an inverted pendulum placed on top of the hand's palm, subjects are unable to modify the stiffness of the pendulum virtual joint, which remains close to nil throughout experiments. Human upright standing can also be likened to an inverted pendulum. In this case, the ankle stiffness is not null and in principle can be modulated by coactivating the ankle muscles. But this is not what people do during quiet standing. The reason is that the Achilles tendon is too compliant (see Loram et al., 2005a for direct measurements of the tendon stiffness) in relation to both the toppling gravity torque and the range of stiffness values of the ankle plantar flexors. This is a quite peculiar situation, because in most sites of the human musculoskeletal system tendons, connecting muscle groups to bones, are much stiffer than the physiological range of muscle stiffness values. Considering that tendons and muscles are connected in a serial fashion, this makes it possible to modulate joint stiffness by coactivating antagonistic muscle groups. But this is not functional in the ankle case because the low stiffness of the tendons radically reduces the range of ankle stiffness values achievable by means of cocontraction.

As a consequence, in both situations sketched above, the stiffness strategy successfully adopted by human subjects in the DF experiments is ineffective and thus is ruled out by the subjects. What is observed, instead, is the presence of intermittent control patterns (Cabrera and Milton, 2002 in the human balancing task; Bottaro et al., 2005; Loram et al., 2005b in the upright standing synergy). Intermittent, closed-loop control should not be confused with sampled-data control, which is an engineering technique dictated by the use of digital computers in the control of dynamic processes. Rather, as concisely expressed by Gawthrop et al. (2011), the control mechanism is 'continuous observation, intermittent action.' In other terms, in intermittent control paradigms, a sensory observer is supposed to continuously analyze the flow of multimodal sensory signals in order to spot critical events and then to 'fire' at the right moment control bursts that temporarily stabilize the system. The crucial point is that even in unstable tasks it is possible to

single out in the state space stable and unstable manifolds that have different degrees of instability. Therefore, an event can be a situation in which the system state comes too close to an unstable manifold, requiring an urgent stabilizing action, or close enough to a stable manifold, suggesting that control signals turn off because intrinsic dynamics is sufficiently 'well behaved.' This kind of stabilization strategy is effective, even with delayed feedback, and indeed it has been demonstrated for upright standing that it is much more robust that a corresponding continuous-time control strategy, with the same value of delay in the feedback loop, in the sense that it can guarantee stability for a much larger range of drift/uncertainty of the control parameters (Asai et al., 2009; Suzuki et al., 2012).

Generally speaking, since the time of Woodworth (1899), it has been well known from the observation of rapid limb and eye movements that the human repertoire includes control mechanisms with a discrete, preprogrammed, ballistic nature that also includes some kind of psychological refractory period. Later on, the analysis of human operators in a loop that implies sustained, continuous feedback from some external process has prompted the emergence of explanations based on various forms of intermittent, predictive control (Craik, 1947; Bekey, 1962; Navas and Stark, 1968; Miall et al., 1993; Hanneton et al., 1997), challenging traditional engineering explanations based on continuous-time control. Apart from empirical aspects, such as the degree of data fitting, the rationale of favoring intermittent predictive over continuous predictive control in the biological domain is that the former control paradigm provides a simpler algorithm than the latter due to the very simple form of the predictor. However, as well explained by Gawthrop et al. (2011), intermittent control has a 'masquerading property,' which makes it indistinguishable from continuous control under some experimental conditions.

Concluding Remarks

Classical mechanics has certainly an important role to play in human motor control, but this role is mediated by the cybernetics of purposive actions that adapts control mechanisms to the nature of the task and the constraints imposed by the body and the environment. This implies a kind of hierarchy between geometric, kinematic, and dynamic aspects in the generation of purposive actions, in order to maintain a degree of uniformity between covert and overt actions. When it comes to carrying out overt actions, the main problem to face is the delay in the acquisition of sensory information about the ongoing action. A variety of control mechanisms have been developed for this purpose. At the lowest level, in the case of unpredicted disturbances, the intrinsic stiffness of muscles provides some zero-delay feedback, which may be insufficient but is supported, after some delay, by segmental reflexes. Thus muscle stiffness and segmental reflexes are synergistic in many cases.

If the source of internal/external disturbance is predictable, training can make it possible to develop internal models of the disturbances that can drive a feed-forward control mechanism. Again, error models and/or additional sources of noise can be partially compensated for by muscle stiffness and reflexes.

In some case, as in the DF experiments, the detailed time course of the disturbance cannot be the predicted and/or learned but some global parameters can be acquired during training, such as the range and directionality of the disturbance. This is enough for the brain to program an optimal muscle coactivation pattern that implements predictive stiffness control. This mechanism is a combination of feed-forward and feedback controls: feed-forward because muscle coactivation is chosen and becomes effective before the initiation of the movement; feedback because the actual compensatory forces are determined by the instantaneous response to positional errors.

In similar situations, where the detailed time course of the disturbance cannot be predicted, a form of feedback control more sophisticated than segmental reflexes is in the repertoire of human control: event-driven intermittent control, based on a short-term prediction of relevant behavioral features such as the stability/instability of the ongoing action.

See also: Motor Control Models: Learning and Performance.

Bibliography

Asai, Y., Tasaka, Y., Nomura, K., Nomura, T., Casadio, M., Morasso, P., 2009. A model of postural control in quiet standing: robust compensation of delay-induced instability using intermittent activation of feedback control. PLoS One 4 (7), e6169.

Asatryan, D.G., Feldman, A.G., 1965. Functional tuning of the nervous system with control of movements or maintenance of a steady posture. Biophysics 10, 925–935.

Bekey, G.A., 1962. The human operator as a sampled-data system. IRE Transactions on Human Factors in Electronics HFE-3 (2), 43–51.

Bizzi, E., Polit, A., Morasso, P., 1976. Mechanisms underlying recovery of final head position. Journal of Neurophysiology 39, 435–444.

Bizzi, E., Mussa-Ivaldi, F.A., Giszter, S., 1991. Computations underlying the execution of movement: a biological perspective. Science 253, 287–291.

Bizzi, E., Hogan, N., Mussa Ivaldi, F.A., Giszter, S.F., 1992. Does the nervous system use equilibrium-point control to guide single and multiple movements? Behavioral and Brain Sciences 15, 603–613.

Borrelli, G.A., 1680. De Motu Animalium. Typographia A. Bernabò, Rome, Italy.

Bottaro, A., Casadio, M., Morasso, P., Sanguineti, V., 2005. Body sway in quiet standing: is it the residual chattering of an intermittent stabilization process? Human Movement Science 24, 588–615.

Braitenberg, V., 1986. Vehicles – Experiments in Synthetic Psychology. MIT Press.

Brashers-Krug, T., Shadmehr, R., Bizzi, E., 1996. Consolidation in human motor memory. Nature 382, 252–255.

Brooks, R.A., 1986. A robust layered control system for a mobile robot. IEEE Journal of Robotics and Automation 2, 14–23.

Brooks, R.A., 1991. Intelligence without representation. Artificial Intelligence 47, 139–159.

Burdet, E., Osu, R., Franklin, D.W., Milner, T.E., Kawato, M., 2001. The central nervous system stabilizes unstable dynamics by learning optimal impedance. Nature 414, 446–449.

Cabrera, J.L., Milton, J.G., 2002. On-off intermittency in a human balancing task. Physical Review Letters 89, 158702.

Craik, K.J.W., 1947. Theory of the human operator in control systems. I. The operator as an engineering system. British Journal Psychology, General Section 38 (Pt 2), 56–61.

Feldman, A.G., 1966. Functional tuning of the nervous system with control of movement or maintenance of a steady posture, II: controllable parameters of the muscles. Biophysics 11, 565–578.

Feldman, A.G., Levin, M.F., 1995. The origin and use of positional frames of references in motor control. Behavioral and Brain Sciences 18, 723–745.
Flash, T., Hogan, N., 1985. The coordination of arm movements: an experimentally confirmed mathematical model. Journal of Neuroscience 7, 1688–1703.
Gawthrop, P., Loram, I., Lakie, M., Gollee, H., 2011. Intermittent control: a computational theory of human control. Biological Cybernetics 104, 31–51.
Goldstein, H., 1980. Classical Mechanics. Addison–Wesley, Reading, MA.
Hanneton, S., Berthoz, A., Droulez, J., Slotine, J.J., 1997. Does the brain use sliding variables for the control of movements? Biological Cybernetics 77, 381–393.
Henneman, E., Somjen, G., Carpenter, D.O., 1965. Functional significance of cell size in spinal motoneurons. Journal of Neurophysiology 28, 560–580.
Hill, A.V., 1938. The heat of shortening and dynamics constants of muscles. Proceedings of the Royal Society of London B 126 (843), 136–195.
Jeannerod, M., 2001. Neural simulation of action: a unifying mechanism for motor cognition. Neuroimage 14, 103–109.
Kandel, E.R., Tauc, L., 1965. Heterosynaptic facilitation in neurones of the abdominal ganglion of Aplysia depilans. Journal of Physiology 181, 1–27.
Kawato, M., 1999. Internal models for motor control and trajectory planning. Current Opinion in Neurobiology 9, 718–727.
Kawato, M., Gomi, H., 1992. A computational model of four regions of the cerebellum based on feedback-error learning. Biological Cybernetics 68, 95–103.
Kawato, M., Wolpert, D., 1998. Internal models for motor control. In: Bock, G.R., Goode, J.A., (Eds.), Sensory Guidance of Movement, Novartis Foundation Symposium, vol. 218. J. Wiley, New York, USA, pp. 291–307.
Kranczioch, C., Mathews, S., Dean, J.A., Sterr, A., 2009. On the equivalence of executed and imagined movements. Human Brain Mapping 30, 3275–3286.
Loram, I.D., Maganaris, C.N., Lakie, M., 2005a. Active, non-spring-like muscle movements in human postural sway: how might paradoxical changes in muscle length be produced. Journal of Physiology 564 (Pt 1), 281–293.
Loram, I.D., Maganaris, C.N., Lakie, M., 2005b. Human postural sway results from frequent, ballistic bias impulses by soleus and gastrocnemius. Journal of Physiology 564 (Pt 1), 295–311.
Miall, R.C., Weir, D.J., Stein, J.F., 1993. Intermittency in human manual tracking tasks. Journal of Motor Behavior 25, 53–63.
Morasso, P., 1981. Spatial control of arm movements. Experimental Brain Research 42, 223–227.
Munzert, J., Lorey, B., Zentgraf, K., 2009. Cognitive motor processes: the role of motor imagery in the study of motor representations. Brain Research Reviews 60, 306–326.
Mussa-Ivaldi, F.A., Hogan, N., Bizzi, E., 1985. Neural, mechanical, and geometric factors subserving arm posture in humans. Journal of Neuroscience 10, 2732–2743.
Navas, F., Stark, L., 1968. Sampling or intermittency in hand control system dynamics. Biophysical Journal 8, 252–302.
Sabes, P.N., Jordan, M.I., Wolpert, D.M., 1998. The role of inertial sensitivity in motor planning. Journal of Neuroscience 18, 5948–5957.
Shadmehr, R., Mussa-Ivaldi, F.A., 1994. Adaptive representation of dynamics during learning of a motor task. Journal of Neuroscience 14, 3208–3224.
Suzuki, Y., Nomura, T., Casadio, M., Morasso, P., 2012. Intermittent control with ankle, hip, and mixed strategies during quiet standing: a theoretical proposal based on a double inverted pendulum model. Journal of Theoretical Biology 301, 55–79.
Wolpert, D.M., Kawato, M., 1998. Internal models of the cerebellum. Trends in Cognitive Science 2, 338–347.
Woodworth, R., 1899. The accuracy of voluntary movement. Psychological Review 3, 1–114.

Classical (Psychometric) Test Theory

Rolf Steyer, Friedrich-Schiller-Universität, Jena, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Classical test theory (CTT) has been developed to quantify measurement error and to solve related problems such as correcting observed dependencies between variables (e.g., correlations) for attenuation due to measurement errors. Basic concepts of CTT are true score and measurement error variables. These concepts are defined as specific conditional expectations and their residuals, respectively. The definitions of these concepts already imply a number of properties that were considered axioms in early presentations of CTT. Models of CTT consist of assumptions about the true score and error variables allowing to identify the theoretical parameters (such as true score variance and error variance) from the variances and covariances of the observable measurements (test score variables). A number of implications of the assumptions defining models of CTT may be tested empirically via structural equation modeling. Hinting at more recent theories and their goals such as item response theory, generalizability theory, and latent state-trait theory concludes this article.

Introduction

One of the most striking and challenging phenomena in the social sciences is the unreliability of its measurements: measuring the same attribute twice often yields two different results. If the same measurement instrument is applied twice such a difference may sometimes be due to a change in the measured attribute itself. Sometimes these changes in the measured attribute are due to the mere fact of measuring. People learn when solving tasks, and they change their attitude when they reflect on statements in an attitude questionnaire, for instance. In other cases, the change of the measured attribute is due to developmental phenomena or to learning between occasions of measurement. However, if change of the attribute can be excluded, two different results in measuring the same attribute can be explained only by measurement error.

Classical (psychometric) test theory (CTT) aims at studying the reliability of a (real-valued) test score variable (measurement, test) that maps a crucial aspect of qualitative or quantitative observations into the set of real numbers. Aside from determining the reliability of a test score variable itself CTT allows answering questions such as

1. How do two random variables correlate once the measurement error is filtered out (correction for attenuation)?
2. How dependable is a measurement in characterizing an attribute of an individual unit, i.e., which is the confidence interval for the true score of that individual with respect to the measurement considered?
3. How reliable is an aggregated measurement consisting of the average (or sum) of several measurements of the same unit or object (Spearman–Brown formula for test length)?
4. How reliable is a difference, e.g., between a pretest and posttest?

Basic Concepts of Classical Test Theory

Primitives

In the framework of CTT, each measurement (test score) is considered being a value of a random variable Y consisting of two components: a true score and an error score. Two levels, or more precisely, two random experiments may be distinguished: (1) sampling an observational unit (e.g., a person) and (2) sampling a score within a given unit. Within a given unit, the true score is a parameter, i.e., a given but unknown number characterizing the attribute of the unit, whereas the error is a random variable with an unknown distribution. The true score of the unit is defined to be the expectation of this intra-individual distribution.

Taking the across units, perspective, i.e., joining the two random experiments, the true score is considered itself being a value of a random variable (the true score variable). The error variable is again a random variable, the distribution of which is a mixture of the individual units, error distributions. Most theorems of CTT (see, e.g., Lord and Novick, 1968) are formulated from this across units' perspective allowing to talk about the correlation of true scores with other variables, for instance.

More formally, CTT refers to a (joint) random experiment of (1) sampling an observational unit u (such as a person) from a set Ω_U of units (called the population) and (2) registering one or more observations out of a set Ω_O of possible observations. The set of possible outcomes of the random experiment is the set product: $\Omega = \Omega_U \times \Omega_O$. The elements of Ω_O, the observations, might be qualitative (such as "answering in category a of item 1 and in category b of item 2"), quantitative (such as reaction time and alcohol concentration in the blood), or consisting of both qualitative and quantitative components. In psychology, the measurements are often defined by test scoring rules prescribing how the observations are transformed into test scores. (Hence, these measurements are also often called tests or test score variables.) These scoring rules may just consist of summing initial scores of items (defining a psychological scale) or might be more sophisticated representations of observable responses of the units. CTT does not prescribe the definition of the test score variables. It just additively decomposes them into true score variables and error variables. Substantive theory and empirical validation studies are necessary in order to decide whether or not a given test score variable is meaningful. CTT only helps disentangling the variances of its true score and error components.

Table 1 Basic concepts of classical test theory

Primitives	
The set of possible outcomes of the random experiment	$\Omega = \Omega_U \times \Omega_O$
Test score variables	Y_i: $\Omega \rightarrow \mathrm{IR}$
Projection	U: $\Omega \rightarrow \Omega_U$
Definition of the theoretical variables	
True score variable	$\tau_i = E(Y_i \mid U)$
Measurement error variable	$\varepsilon_i = Y_i - \tau_i$

Referring to the joint random experiment described above the mapping U: $\Omega \rightarrow \Omega_U$, $U(\omega) = u$ (the unit or person projection) may be considered a qualitative random variable having a joint distribution with the test score variables Y_i. Most theorems of CTT deal with two or more test score variables (tests) Y_i and the relationship between their true score and error components. (The index i refers to one of several tests considered.)

The Core Concepts: True Score and Error Variables

Using the primitives introduced above, the true score variable $\tau_i = E(Y_i \mid U)$ is defined by the conditional expectation of the test Y_i given the variable U. The values of the true score variable τ_i are the conditional expected values $E(Y_i \mid U = u)$ of Y_i given the unit u. They are also called the true scores of the unit u with respect to Y_i. Hence, these true scores are the expected values of the intraindividual distributions of the Y_i. The measurement error variables ε_i are simply defined by the difference $\varepsilon_i = Y_i - \tau_i$. Table 1 summarizes the primitives and definitions of the basic concepts of CTT.

Properties of True Score and Error Variables

Once the true score variables and error variables are defined a number of properties (Table 2) can be derived, some of which are known as the axioms of CTT. At least since Novick's (1966) and Zimmerman's (1975, 1976) papers it is well known, however, that all these properties already follow from the definition of true score and error variables. They are not new and independent assumptions as has been originally proposed (see, e.g., Gulliksen, 1950). All equations in Table 2 are no assumptions. They are inherent properties of true scores and errors. Hence, trying to test or falsify these properties empirically would be meaningless just in the same way as it is meaningless to test whether or not a bachelor is really unmarried. The property of being unmarried is an inherent part or logical consequence of the concept of a bachelor.

Only one of the 'axioms of CTT' does not follow from the definition of true score and error variables: uncorrelatedness of error variables among each other. Hence, uncorrelatedness of errors has another epistemological status as the properties displayed in Table 2 (the other 'axioms'). Uncorrelatedness of errors is certainly a desirable and useful property; but it might be wrong in specific empirical applications (see, e.g., Zimmerman and Williams, 1977). It in fact is an assumption and it plays a crucial rule in defining models of CTT.

Equation [1] of Table 2 is a simple rearrangement of the definition of the error variable. Equation [2] shows that the variance of a test score variable, too, has two additive components: the variance of the true score variable and the variance of the error variable. This second property follows from Eqn [3] according to which a true score variable is uncorrelated with a measurement error variable, even if they pertain to different test score variables Y_i and Y_j. Equation [4] states that the expected value of an error variable is zero, whereas Eqn [5] implies that the expected value of an error variable is zero within each individual observational unit u. Finally, according to Eqn [6] the conditional expectation of an error variable is also zero for each mapping of U. This basically means that the expected value of an error variable is zero in each subpopulation of observational units.

Additional Concepts: Reliability, Unconditional, and Conditional Error Variances

Although the true score and error variables defined above are the core concepts of CTT, it is impossible, in general, to learn something about the true score and error scores themselves in empirical applications. What is possible, however, is to estimate the variances of the true score and error variables in a random sample (consisting of repeating many times the random experiment described earlier). The variance $\mathrm{Var}(\varepsilon_i)$ of the measurement error may be considered a gross parameter representing the degree of unreliability. A normed parameter of unreliability is $\mathrm{Var}(\varepsilon_i)/\mathrm{Var}(Y_i)$, the proportion of the variance of Y_i due to measurement error. Its counterpart is $1 - \mathrm{Var}(\varepsilon_i)/\mathrm{Var}(Y_i)$, i.e.,

$$\mathrm{Rel}(Y_i) = \mathrm{Var}(\tau_i)/\mathrm{Var}(Y_i) \qquad [7]$$

Table 2 Properties of true score and error variables implied by their definition

Decomposition of the variables	$Y_i = \tau_i + \varepsilon_i$	[1]
Decomposition of the variances	$\mathrm{Var}(Y_i) = \mathrm{Var}(\tau_i) + \mathrm{Var}(\varepsilon_i)$	[2]
Other properties of true score and error variables implied by their definition	$\mathrm{Cov}(\tau_i, \varepsilon_j) = 0$	[3]
	$E(\varepsilon_i) = 0$	[4]
	$E(\varepsilon_i \mid U) = 0$	[5]
for each (measurable) mapping of U:	$E[\varepsilon_i \mid f(U)] = 0$	[6]

the reliability of Y_i. This coefficient varies between zero and one. In fact, most theorems and most empirical research deal with this coefficient of reliability. The reliability coefficient conveniently informs about the dependability of the measurement in one single number.

In early papers on CTT, reliability of a test has been defined by its correlation with itself (see, e.g., Thurstone, 1931). However, this definition is only metaphoric, because a variable always correlates perfectly with itself. What is meant is to define reliability by the correlation of 'parallel tests' (see below). The assumptions defining parallel tests in fact imply that the correlation between two test score variables is the reliability. Note that the definition of reliability via Eqn [7] does not rest on any assumption other than $0 < \mathrm{Var}(Y_i) < \infty$.

Reliability is useful in comparing different instruments with each other if they are applied to the same population. Used in this way, reliability in fact helps in evaluating the quality of measurement instruments. It may not be useful, however, under all circumstances, to infer the dependability of measures of an individual unit. For the latter purpose, one might rather look at the conditional error variance $\mathrm{Var}(\varepsilon_i \mid U = u)$ given a specific observational unit u or at the conditional error variances $\mathrm{Var}(\varepsilon_i \mid \tau_i = t)$ given the subpopulation with true score $\tau_i = t$.

Models of Classical Test Theory

The definitions of true score and error variables have to be supplemented by assumptions defining a model if the theoretical parameters such as the reliability are to be computed by empirically estimable parameters such as the means, variances, covariances, or correlations of the test score variables. Table 3 displays the most important of these assumptions and the most important models defined by combining some of these assumptions.

The assumptions (a_1)–(a_3) specify in different ways the assumption that two tests Y_i and Y_j measure the same attribute. Such an assumption is crucial for inferring the degree of reliability from the discrepancy between two measurements of the same attribute of the same person. Perfect identity or 'τ-equivalence' of the two true score variables is assumed with (a_1). With (a_2) this assumption is relaxed: the two true score variables may differ by an additive constant. Two balances, for instance, will follow this assumption if at least one of them yields a weight that is always one pound larger than the weight indicated by the other balance, irrespective of the object to be weighed. According to assumption (a_3), the two tests measure the same attribute in the sense that their true score variables are linear functions of each other.

The other two assumptions deal with properties of the measurement errors. With (b) one assumes measurement errors pertaining to different test score variables to be uncorrelated. In (c) equal error variances are assumed, i.e., these tests are assumed to measure equally well.

Parallel Tests

Definition: The most simple and convenient set of assumptions is the model of parallel tests. Two tests Y_i and Y_j are defined to be parallel if they are τ-equivalent, their error variables are uncorrelated, and they have identical error variances. Note that assumption (a_1) implies that there is a uniquely defined latent variable being identical to each of the true score variables. Hence, one may drop the index i and denote this latent variable by η. The assumption of τ-equivalence may equivalently be written $Y_i = \eta + \varepsilon_i$, where $\varepsilon_i = Y_i - E(Y_i \mid U)$.

Identification: For parallel tests the theoretical parameters may be computed from the parameters characterizing the distribution of at least two test score variables, i.e., the theoretical parameters are identified in this model if $m \geq 2$. According to Table 4 the expected value of η is equal to the expected value of each of the tests, whereas the variance of η can be computed from the covariance of two different tests. The variance $\mathrm{Var}(\varepsilon_i)$ of the measurement error variables may be computed by the difference $\mathrm{Var}(Y_i) - \mathrm{Cov}(Y_i, Y_j)$, $i \neq j$. Finally, the reliability $\mathrm{Rel}(Y_i)$ is equal to the correlation $\mathrm{Corr}(Y_i, Y_j)$ of two different test score variables.

Testability: The model of parallel tests implies several consequences that may be tested empirically. First, all parallel tests Y_i have equal expectations $E(Y_i)$, equal variances $\mathrm{Var}(Y_i)$, and equal covariances $\mathrm{Cov}(Y_i, Y_j)$ in the total population. Second, parallel tests also have equal expectations in each subpopulation (Table 4).

Note that these hypotheses may be tested separately and/or simultaneously as a single multidimensional hypothesis in the framework of simultaneous equation models via AMOS (Arbuckle, 2006), EQS (Bentler, 2004), LISREL (Mels, 2006),

Table 3 Assumptions and some models of CTT

Assumptions used to define some models of CTT	
(a_1) τ-Equivalence	$\tau_i = \tau_j$
(a_2) Essential τ-equivalence	$\tau_i = \tau_j + \lambda_{ij}$, $\lambda_{ij} \in \mathrm{IR}$
(a_3) τ-Congenerity	$\tau_i = \lambda_{ij0} + \lambda_{ij1}\,\tau_j$, $\lambda_{ij0}, \lambda_{ij1} \in \mathrm{IR}$, $\lambda_{ij1} > 0$
(b) Uncorrelated errors	$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$
(c) Equal error variances	$\mathrm{Var}(\varepsilon_i) = \mathrm{Var}(\varepsilon_j)$
Models defined by combining these assumptions	
Parallel tests are defined by assumptions (a_1), (b), and (c).	
Essentially τ-equivalent tests are defined by assumptions (a_2) and (b).	
Congeneric tests are defined by assumptions (a_3) and (b).	

The equations refer to each pair of tests Y_i and Y_j of a set of tests $Y_1, \ldots, Y_m$, their true score variables, and their error variables, respectively.

Table 4 The model of parallel tests

Definition	*Assumptions (a_1), (b), and (c) of Table 3*
Identification	$E(\eta) = E(Y_i)$
	$\mathrm{Var}(\eta) = \mathrm{Cov}(Y_i, Y_j)$, $i \neq j$
	$\mathrm{Var}(\varepsilon_i) = \mathrm{Var}(Y_i) - \mathrm{Cov}(Y_i, Y_j)$, $i \neq j$
	$\mathrm{Rel}(Y_i) = \mathrm{Corr}(Y_i, Y_j)$, $i \neq j$
Testability	
in the total population	$E(Y_i) = \mu$
	$\mathrm{Var}(Y_i) = \sigma^2_Y$
	$\mathrm{Cov}(Y_i, Y_j) = \sigma^2_\eta$
within each subpopulation s	$E^{(s)}(Y_i) = \mu_s$

The indices i and j refer to tests and the superscript s to a subpopulation. The equations are true for each test Y_i of a set of parallel tests $Y_1, \ldots, Y_m$ or for each pair of two such tests, their true score variables, and their error variables, respectively.

Mplus (Muthén and Muthén, 2007), and others. Such a simultaneous test may even include the hypotheses about the parameters in several subpopulations (Table 4). What is not implied by the assumptions of parallel tests is the equality of the variances and the covariances of the test score variables between different subpopulations.

For parallel tests $Y_1 + \ldots + Y_m$ as defined in Table 4, the reliability of the sum score $S = Y_1 + \ldots + Y_m$ may be computed by the Spearman–Brown formula for lengthened tests:

$$\mathrm{Rel}(S) = \mathrm{Rel}(S/m) = \frac{m \cdot \mathrm{Rel}(Y_1)}{1 + (m-1) \cdot \mathrm{Rel}(Y_1)}$$

Using this formula the reliability of an aggregated measurement consisting of the sum (or average) of m parallel measurements of the same unit can be computed. For $m = 2$ and $\mathrm{Rel}(Y) = 0.80$, for instance, the Spearman–Brown formula yields

$$\mathrm{Rel}(S) = 2 \cdot 0.80 / 1 + (2-1) \cdot 0.80 \approx 0.89$$

The Spearman–Brown formula may also be used to answer the opposite question. Suppose there is a test being the sum of m parallel tests and this test has reliability $\mathrm{Rel}(S)$. What would be the reliability $\mathrm{Rel}(Y_1)$ of the m parallel tests? For example, if $m = 2$, what would be the reliability of a test half if $\mathrm{Rel}(S) = 0.90$?

Essentially τ-Equivalent Tests

Definition: The model of essentially τ-equivalent tests is less restrictive than the model of parallel tests. Two tests Y_i and Y_j are defined to be essentially τ-equivalent if their true score variables differ only by an additive constant (assumption (a_2) in Table 3) and if their error variables are uncorrelated (assumption (b) in Table 3). Assumption (a_2) implies that there is a latent variable η that is a translation of each of the true score variables, i.e., $\eta = \tau_i + \lambda_i$, $\lambda_i \in \mathrm{IR}$, such that $Y_i = \eta + \lambda_i + \varepsilon_i$, where $\varepsilon_i = Y_i - E(Y_i \mid U)$ and $\lambda_i \in \mathrm{IR}$.

Also note that the latent variable η is uniquely defined up to a translation. Hence, it is necessary to fix the scale of the latent variable η. This can be done by fixing one of the coefficients λ_i (e.g., $\lambda_1 = 0$) or by fixing the expected value of η (e.g., $E(\eta) = 0$).

Table 5 summarizes the assumptions defining the model and the consequences for identification and testability. In this model, the reliability cannot be identified any more by the correlation between two tests. Instead the reliability is identified by $\mathrm{Rel}(Y_i) = \mathrm{Cov}(Y_i, Y_j)/\mathrm{Var}(Y_i)$, $i \neq j$. Furthermore, the expected values of different tests are not identical any more within each subpopulation. Instead, the differences between the expected values $E^{(s)}(Y_i) - E^{(s)}(Y_j)$ of two essentially τ-equivalent tests Y_i and Y_j are the same in each and every subpopulation. All other properties are the same as in the model of parallel tests. Again, all these hypotheses may be tested via structural equation modeling.

For essentially τ-equivalent tests $Y_1, \ldots, Y_m$, the reliability of the sum score $S = Y_1 + \ldots + Y_m$ may be computed by Cronbach's coefficient α:

$$\alpha = \frac{m}{m-1}\left(1 - \frac{\sum_{i=1}^{m} \mathrm{Var}(Y_i)}{\mathrm{Var}(S)}\right)$$

This coefficient is a lower bound for the reliability of S if only uncorrelated errors are assumed.

Table 5 The model of essentially τ-equivalent tests

Definition	*Assumptions (a_2) and (b) of Table 3*
Fixing the scale of η	$E(\eta) = 0$
Identification	$\mathrm{Var}(\eta) = \mathrm{Cov}(Y_i, Y_j)$, $i \neq j$
	$\mathrm{Var}(\varepsilon_i) = \mathrm{Var}(Y_i) - \mathrm{Cov}(Y_i, Y_j)$, $i \neq j$
	$\mathrm{Rel}(Y_i) = \mathrm{Cov}(Y_i, Y_j)/\mathrm{Var}(Y_i)$, $i \neq j$
Testability	
in the total population	$\mathrm{Cov}(Y_i, Y_j) = \sigma_\eta^2$
in each subpopulation s	$E^{(s)}(Y_i) - E^{(s)}(Y_j) = E^{(s)}(Y_i - Y_j) = \lambda_{ij}$

The indices i and j refer to tests and the superscript s to a subpopulation. The equations are true for each test Y_i of a set of tests $Y_1, \ldots, Y_m$ or for each pair of two such tests, their true score variables, and their error variables, respectively.

τ-Congeneric Tests

Definition: The model of τ-congeneric tests is defined by the assumptions (a_3) and (b) in Table 3. Hence, two tests Y_i and Y_j are called τ-congeneric if their true score variables are positive linear functions of each other and if their error variables are uncorrelated. Assumption (a_3) implies that there is a latent variable η such that each true score variable is a positive linear function of it, i.e., $\tau_i = \lambda_{i0} + \lambda_{i1}\eta$, λ_{i0}, $\lambda_{i1} \in \mathrm{IR}$, $\lambda_{i1} > 0$, or equivalently: $Y_i = \lambda_{i0} + \lambda_{i1}\eta + \varepsilon_i$, where $\varepsilon_i = Y_i - E(Y_i \mid U)$.

The latent variable η is uniquely defined up to positive linear functions. Hence, in this model, too, it is necessary to fix the scale of η. This can be done, e.g., by fixing a pair of the coefficients (e.g., $\lambda_{i0} = 0$ and $\lambda_{i1} = 1$) or by fixing the expected value and the variance of η (e.g., $E(\eta) = 0$ and $\mathrm{Var}(\eta) = 1$).

Table 6 summarizes the assumptions defining the model and the consequences for identification and testability assuming $E(\eta) = 0$ and $\mathrm{Var}(\eta) = 1$. Other ways of fixing the scale of η would imply different formula. As can be seen from the formulas in Table 6, the model of τ-congeneric variables and all its parameters are identified if there are at least three different tests for which assumptions (a_3) and (b) hold. The covariance structure in the total population implied by the model may be tested empirically if there are at least four test score variables. Only in this case the model has fewer theoretical parameters determining the covariance matrix of the test score variables than there are elements in this covariance matrix. The implications for the mean structure are testable already for three test score variables provided the means of the test score variables are available in at least four subpopulations.

Other Models of CTT

The models treated above are not the only ones that can be used to determine the theoretical parameters of CTT such as reliability, true score variance, and error variance. In fact, the models dealt with are limited to unidimensional models. True score variables, however, may also be decomposed into several latent variables. Confirmatory factor analysis provides a powerful methodology to construct, estimate, and test models with multidimensional decompositions of true score variables. Note, however, that not each factor model is based on CTT. For instance, there are one-factor models that are not models of τ-congeneric variables in terms of CTT. A model with

Table 6 The model of τ-congeneric tests

Definition	*Assumptions (a_3) and (b) of Table 3*
Fixing the scale of η	$E(\eta) = 0$ and $\mathrm{Var}(\eta) = 1$
Identification	$\lambda_{i1} = \sqrt{\dfrac{\mathrm{Cov}(Y_i, Y_j)\ \ \mathrm{Cov}(Y_i, Y_k)}{\mathrm{Cov}(Y_j, Y_k)}}, \quad i \neq j, \quad i \neq k, \quad j \neq k$
	$\mathrm{Var}(\varepsilon_i) = \mathrm{Var}(Y_i) - \lambda_{i1}^2$
	$\mathrm{Rel}(Y_i) = \lambda_{i1}^2 / \mathrm{Var}(Y_i)$
Testability in the total population	$\dfrac{\mathrm{Cov}(Y_i, Y_k)}{\mathrm{Cov}(Y_j, Y_k)} = \dfrac{\mathrm{Cov}(Y_i, Y_l)}{\mathrm{Cov}(Y_j, Y_l)}, \quad i \neq k, i \neq l, j \neq k, j \neq l$
between subpopulations	$\dfrac{E^{(1)}(Y_i) - E^{(2)}(Y_i)}{E^{(1)}(Y_j) - E^{(2)}(Y_j)} = \dfrac{E^{(3)}(Y_i) - E^{(4)}(Y_i)}{E^{(3)}(Y_j) - E^{(4)}(Y_j)}$

The indices i and j refer to tests and the superscripts to one of four subpopulations.

one common factor and several specific but uncorrelated factors is a counter example. The common factor is not necessarily a linear function of the true score variables, and the specific factors are not necessarily the measurement error variables as defined in CTT.

Some Practical Issues

Once a measurement or test score has been obtained for a specific individual, one might want to know how dependable that individual measurement is. If the reliability of the measurement is known and if one assumes a normal distribution of the measurement errors, which is homogeneous for all individuals, the 95% confidence interval for the true score of that individual with respect to the measurement Y_i can be computed by

$$Y_i \pm 1.96 \cdot \sqrt{\mathrm{Var}(Y_i) \cdot (1 - \mathrm{Rel}(Y_i))}$$

Another result deals with the correlation between two true score variables. If the reliabilities for two test score variables Y_1 and Y_2 are known, and assuming uncorrelated measurement errors, one may compute

$$\mathrm{Corr}(\tau_1, \tau_2) = \frac{\mathrm{Corr}(Y_1, Y_2)}{\sqrt{\mathrm{Rel}(Y_1)} \cdot \sqrt{\mathrm{Rel}(Y_2)}}$$

This equation is known as the correction for attenuation.

Another important issue deals with the reliability of a difference variable, for instance, a difference between a pretest Y_1 and a posttest Y_2. Assuming equal true score and error variances between pre- and posttest implies identical reliabilities, i.e., $\mathrm{Rel}(Y_1) = \mathrm{Rel}(Y_2) = \mathrm{Rel}(Y)$. If, in addition, uncorrelated measurement errors are assumed, the reliability $\mathrm{Rel}(Y_1 - Y_2) = \mathrm{Var}(E(Y_1 - Y_2) \mid U)/\mathrm{Var}(Y_1 - Y_2)$ of the difference $Y_1 - Y_2$ may be computed by

$$\mathrm{Rel}(Y_1 - Y_2) = \frac{\mathrm{Rel}(Y) - \mathrm{Corr}(Y_1, Y_2)}{1 - \mathrm{Corr}(Y_1, Y_2)}$$

According to this formula, the reliability of a difference between pre- and posttest is always smaller than the reliability of the pre- and posttest, provided the assumptions mentioned above hold. In the extreme case in which there is no differential change, i.e., $\tau_2 = \tau_1 + \text{constant}$, the reliability coefficient $\mathrm{Rel}(Y_1 - Y_2)$ will be zero. Of course, this does not mean that the change is not dependable. It only means that there is no variance in the change, since each individual changes in the same amount. This phenomenon has lead to much confusion about the use of measuring change (see, e.g., Cronbach and Furby, 1970; Harris, 1963; Rogosa, 1995). Most of these problems are now solved by structural equation modeling, allowing the inclusion of latent change variables such as in growth curve models (see, e.g., McArdle & Epstein, 1987; Willet & Sayer, 1996) or, more directly, in true change models (Steyer, Eid, & Schwenkmezger, 1997; Steyer, Partchev, & Shanahan, 2000). Models of this kind are not hampered any more by reliability problems and allow one to explain inter differences in intra change.

Discussion

It should be noted that CTT refers to the population level, i.e., to the random experiment of sampling a single observational unit and assessing some of its behavior (Table 1). CTT does not refer to sampling models that consist of repeating this random experiment many times. Hence, no questions of statistical estimation and hypothesis testing are dealt with. Of course, the population models of CTT have to be supplemented by sampling models when it comes to applying statistical analyses, e.g., via structural equation modeling.

Aside from this more technical aspect, what are the limitations of CTT? First, CTT and its models are not really adequate for modeling answers to individual items in a questionnaire. This purpose is more adequately met by models of item response theory (IRT), which specify how the probability of answering in a specific category of an item depends on the attribute to be measured, i.e., on the value of a latent variable.

A second limitation of CTT is the exclusive focus on measurement errors. Inspired by generalizability theory (Cronbach et al., 1972; Shavelson and Webb, 1991), Tack (1980) and Steyer et al. (1989) presented a generalization of CTT, called latent state-trait theory, which explicitly takes into account person-specific situation effects, introduces formal

definitions of states and traits, and presents models that allow one to disentangle person effects, situation and/or interaction effects and measurement error. More recent presentations are Steyer et al. (1992), Steyer et al. (1999), as well as Steyer et al. (2012). Eid (1995, 1996) extended this approach to the normal ogive model for analyses on the item level.

The parameters of CTT are often said to be 'population dependent', i.e., meaningful only with respect to a given population. This is true for the variance of the true score variable and the reliability coefficient. The reliability (coefficient) of an intelligence test is different in the population of students than in the general population. This is a simple consequence of the restriction of the (true score) variance of intelligence in the population of students. However, such a restriction exists neither for the true score estimates of individual persons nor for the item parameters λ_i of the model of essentially τ-equivalent tests, for instance. The population dependence critique has often been forwarded by proponents of IRT models. They contrast it with the 'population independence' of the person and the item parameters of IRT models. However, 'population independence' also holds for the person and the item parameters of the model of essentially τ-equivalent tests, for instance.

In applications of CTT it is often assumed that the error variances are the same for each individual, irrespective of the true score of that individual. This assumption may be wrong, indeed, in many applications. In IRT models no such assumption is made. However, it is possible to assume different error variances for different (categories of) persons in CTT models as well. In this case, the unconditional error variance and the reliability coefficient are not the best available information for inferring the dependability of individual true score estimates. In this case one should seek to obtain estimates of conditional measurement error variances for specific classes of persons. It is to be expected that persons with high true scores have a higher measurement error variance than those with medium true scores and that those with low true scores have a higher error variance again. (This would be due to floor and ceiling effects.) Other patterns of the error variance depending on the size of the true score may occur as well.

Such phenomena do not mean that true scores and error scores would be correlated; only the error variances would depend on the true scores. None of the properties listed in Table 2 would be violated. As mentioned before, the properties listed in Table 2 cannot be wrong in empirical applications. What could be wrong, however, is the true score interpretation of the latent variable in a concrete structural equation model. Misinterpretations of this sort can be prevented most effectively by empirical tests of the hypotheses listed in the testability sections of Tables 4–6.

The most challenging critique of many applications of CTT is that they are based on rather arbitrarily defined test score variables. If these test score variables are not well chosen, any model based on them is not well founded, too. Are there really good reasons to base models on sum scores across items in questionnaires? Why take the sum of the items as test score variables Y_i and not another way of aggregation such as a weighted sum, a product, or a sum of logarithms? Why aggregate and not look at the items themselves?

There is no doubt that IRT models are more informative than CTT models if samples are big enough to allow their application, if the items obey the laws defining the models, and if detailed information about the items (and even about the categories of polytomous items, such as in ratings scales) is sought. In most applications, the decision how to define the test score variables Y_i on which models of CTT are built is arbitrary, to some degree. It should be noted, however, that arbitrariness in the choice of the test score variables cannot be avoided all together. Even if models are based on the item level, such as in IRT models, one may ask, "Why these items and not other ones"? Whether or not a good choice has been made will be proved only in model tests and in validation studies. This is true for models of CTT as well as for models of alternative theories of psychometric tests.

See also: Generalizability Theory; Reliability: Measurement; Structural Equation Modeling.

Bibliography

Arbuckle, J.L., 2006. Amos 7.0 User's Guide. SPSS, Chicago, IL.

Bentler, P.M., 2004. EQS 6. Structural Equations Program Manual. Multivariate Software, Encino.

Cronbach, L.J., Furby, L., 1970. How should we measure "change" – or should we? Psychological Bulletin 74, 68–80.

Cronbach, L.J., Gleser, G.C., Nanda, H., Rajaratnam, N., 1972. The Dependability of Behavioral Measurements: Theory of Generalizability of Scores and Profiles. Wiley, New York.

Eid, M., 1995. Modelle der Messung von Personen in Situationen [Models of measuring persons in situations]. Psychologie Verlags Union, Weinheim.

Eid, M., 1996. Longitudinal confirmatory factor analysis for polytomous item responses: model definition and model selection on the basis of stochastic measurement theory. Methods of Psychological Research 1, 65–85. http://www.dgps.de/fachgruppen/methoden/mpr-online/.

Gulliksen, H., 1950. Theory of Mental Tests. Wiley, New York.

Harris, C.W. (Ed.), 1963. Problems in Measuring Change. University of Wisconsin Press, Madison, WI.

Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison Wesley, Reading, MA.

McArdle, J.J., Epstein, D., 1987. Latent growth curves within developmental structural equation models. Child Development 58, 110–133.

Mels, G., 2006. LISREL for Windows. Getting Started Guide. Scientific Software, Chicago, IL.

Muthén, L., Muthén, B., 2007. Mplus User's Guide. fifth ed. Muthén & Muthén, Los Angeles, CA.

Novick, M.R., 1966. The axioms and principal results of classical test theory. Journal of Mathematical Psychology 3, 1–18.

Rogosa, D., 1995. Myths and methods: "Myths about longitudinal research" plus supplemental questions. In: Gottman, J.M. (Ed.), The Analysis of Change. Lawrence Erlbaum, Mahwah, NJ.

Shavelson, R.J., Webb, N.M., 1991. Generalizability Theory. A Primer. Sage, Newbury Park.

Steyer, R., Eid, M., Schwenkmezger, P., 1997. Modeling true intraindividual change: true change as a latent variable. Methods of Psychological Research-Online 2, 21–33.

Steyer, R., Majcen, A.-M., Schwenkmezger, P., Buchner, A., 1989. A latent state-trait anxiety model and its application to determine consistency and specificity coefficients. Anxiety Research 1, 281–299.

Steyer, R., Ferring, D., Schmitt, M.J., 1992. States and traits in psychological assessment. European Journal of Psychological Assessment 8, 79–98.

Steyer, R., Partchev, I., Shanahan, M., 2000. Modeling true intra-individual change in structural equation models: the case of poverty and children's psychosocial adjustment. In: Little, T.D., Schnabel, K.U., Baumert, J. (Eds.), Modeling Longitudinal and Multiple-Group Data: Practical Issues, Applied Approaches, and Specific Examples. Erlbaum, Hillsdale, NJ.

Steyer, R., Schmitt, M., Eid, M., 1999. Latent state-trait theory and research in personality and individual differences. European Journal of Personality 13, 361–387.
Steyer, R., Geiser, C., Fiege, C., 2012. Latent state-trait models. In: Cooper, H. (Ed.), Handbook of Research Methods in Psychology. APA, Washington, DC.
Tack, W.H., 1980. Zur Theorie psychometrischer Verfahren: Formalisierung der Erfassung von Situationsabhängigkeit und Veränderung [On the theory of psychometric procedures: formalizing the assessment of situational dependency and change]. Zeitschrift für Differentielle und Diagnostische Psychologie 1, 87–106.
Thurstone, L.L., 1931. The Reliability and Validity of Tests. Edwards Brothers, Ann Arbor, p. 3.
Willet, J.B., Sayer, A.G., 1996. Cross-domain analyses of change over time: combining growth modeling and covariance structure analysis. In: Marcoulides, G.A., Schumacker, R.E. (Eds.), Advanced Structural Equation Modeling: Issues and Techniques. Erlbaum, Mahwah, NJ.
Zimmerman, D.W., 1975. Probability spaces, Hilbert spaces, and the axioms of test theory. Psychometrika 40, 395–412.
Zimmerman, D.W., 1976. Test theory with minimal assumptions. Educational and Psychological Measurement 36, 85–96.
Zimmerman, D.W., Williams, R.H., 1977. The theory of test validity and correlated errors of measurement. Journal of Mathematical Psychology 16, 135–152.

Classical Tests for Speech and Language Disorders

Yves Turgeon, Clinique de mémoire, Centre de santé et de services sociaux du Rocher-Percé, QC, Canada
Joël Macoir, Université Laval, Québec, QC, Canada
Anne Lafay, Centre de recherche de l'Institut universitaire en santé mentale de Québec, Québec, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the most commonly used tests for the assessment of speech and language processes in adults and children. We present reference models for speech and language evaluations as well as widely used procedures for the assessment of aphasia. The assessment of functional communication is also briefly presented. Finally, issues pertaining to assessing speech and language in special populations are discussed.

Introduction

The evaluation of speech and language is one of the most important tasks of speech–language pathologists and professionals from a variety of disciplines and backgrounds (neuropsychologists, physicians, nurses, etc.). The assessment session is often the first contact with clients and also constitutes the starting point of all clinical interventions. Because of the absence of biological markers or simple assessment methods, the early detection or diagnosis of speech and language problems remains dependent on various indirect assessments (i.e., speech or language functioning must be inferred from the client performance in various tasks devised to explore the different areas of this functioning) performed to identify specific impairments and eliminate other possible causes.

There are various purposes to conduct speech–language assessments. The main goal of screening is to determine whether a client has a problem or not. The output of this type of assessment is a 'pass' or 'fail' result, based on an established criterion that could lead to a more extensive or a follow-up assessment. Diagnosis and differential diagnosis assessments are usually performed to label the communication problem and/or to differentiate it from other disorders in which similar characteristics are usually reported.

Another important purpose to evaluation is to provide clinicians with a detailed description of the client's baseline level of functioning in all areas of communication in order to identify affected and preserved components, to plan for treatment, to establish treatment effectiveness, or to track progress over time through periodic reevaluations. These types of assessment require the clinician to consider all aspects of communication, including the different areas of speech (e.g., articulation, voice, resonance) and language (e.g., lexical access, comprehension, written spelling), but also important related abilities and components such as pragmatics, cognitive functions (e.g., attention, memory, visual perception), emotions, awareness of deficits, etc. The selection of evaluation tools is also conditioned by the specific objectives of assessments. Screening for a speech or language disorder is usually performed with standardized screening measures, whereas standardized norm-referenced tests are used for diagnosis and differential diagnosis assessments as well as for clinical treatment purposes (baseline, effectiveness, progress).

Reference Models for the Assessment of Speech and Language Impairment

The choice of a particular method of assessment, the selection of evaluation tools as well as the interpretation of results, is highly dependent not only on the clinician's own conception of speech and language functioning but also on the reference to a clinicopathological or cognitive model of assessment.

In the clinicopathological model, speech and language problems are considered as essential characteristics of clinical syndromes. These clinical syndromes are organized and classified according to neurological–neuropathological characteristics (e.g., deterioration of cortical tissue in a specific brain area) and according to semiology (e.g., sensitive and motor deficits, visuospatial deficits, language deficits, etc.). For the purpose of assessment, the emphasis is put on the precise identification of the diagnostic label that best corresponds to the observed deficits as well as to the identification of the possible etiology. For example, within this model, the general assessment process of an aphasic person essentially consists of (1) gathering case history data (e.g., cerebrovascular accident in the left frontal area), (2) administrating a specific test battery (e.g., the *Boston Diagnostic Aphasia Examination* (BDAE); Goodglass et al., 2000), (3) confronting the results and description of behavior (e.g., impaired fluency, impaired articulatory agility, relatively good auditory comprehension, agrammatism) with the classification of neurogenic acquired deficits of language, and (4) specifying the precise aphasic label (Broca's aphasia) that best fits these characteristics. If screening or labeling is the main goal of the assessment, the clinicopathological model is probably the best option. It is, however, certainly not so if the purpose of the evaluation is to localize the functional origin of deficits or to guide clinical practice. Knowing that a person presents with a Broca's aphasia may not be much help in identifying the specific components of language that are totally or partially affected or preserved. It also does not tell the clinician what intervention goals are appropriate, what treatment approaches will succeed best. Instead of resorting to a medical assessment model, clinicians may also use cognitive neuropsychological models, directly derived from information-processing theories, to evaluate language. In these models, cognitive functions, including language, are sustained by

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54048-3

specialized interconnected processing components, represented in functional architecture models. For example, as shown in Figure 1, the ability to orally produce a word in picture naming is conceived as a staged process in which the activation flow is initiated in a conceptual-semantic component and ends with the execution of articulation mechanisms.

An assessment process based on cognitive neuropsychological models consists in the localization of the impaired and preserved processing components for each language modality. This localization is performed through the administration of specific tasks or test batteries (e.g., *Psycholinguistic Assessments of Language Processing in Aphasia* (PALPA); Kay et al., 1992) aiming at the evaluation of each component and route of the model. For example, the evaluation of naming abilities in an aphasic person could be performed by the administration of tasks exploring the conceptual-semantic (e.g., semantic questionnaire), phonological output lexicon (e.g., picture-naming task controlled for frequency, familiarity, etc), and phonological output buffer (e.g., repetition of words and nonwords controlled for length) components. Important information regarding the level of impairments also arises from error analysis. With the same example, an anomic behavior could arise from distinct underlying deficits (e.g., in the activation of conceptual-semantic representations or in retrieving phonological forms of words in the output lexicon), leading to distinct types of errors (e.g., semantic substitutions, phonemic errors). The complete cognitive assessment process should allow the clinician to understand the client's deficits (i.e., surface manifestations, underlying origins, affected components) as well as to identify the strengths and weaknesses in his communication abilities. When recommended, the treatment may focus on the impaired levels of processing (i.e., function restoration) or on alternative processing routes (i.e., function reorganization) that will allow the client to communicate successfully.

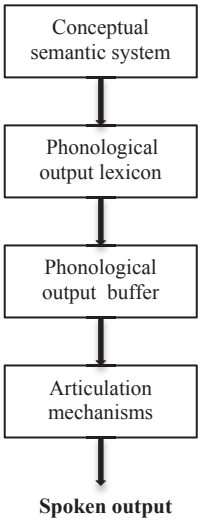

Figure 1 Schematic depiction of the cognitive neuropsychological model of spoken picture naming.

Classical Tests for the Assessment of Aphasia

Aphasia is the most common disorder of communication resulting from brain damage (i.e., stroke, brain tumor, head trauma, infections). This affection mainly involves language problems of production and comprehension as well as disturbances in reading and spelling.

Bedside and Screening Tests

The patient's symptoms change rapidly during the first days and weeks following the brain damage. Moreover, patients are often too ill to complete an exhaustive aphasia examination and bedside or screening instruments may be useful to advise relatives and health care professionals about the global communication profile and the best means to communicate in functional situations. These instruments are also useful to help clinicians to determine the necessity of performing a more thorough and extensive assessment of language or to establish the priority of patients on a waiting list. In addition to actual screening tests (e.g., Aphasia Screening Test; Reitan, 1991; for an extensive list see Murray and Chapey, 2001; Spreen and Risser, 2003), clinicians also may administer shortened versions of comprehensive tests of aphasia (e.g., short form of the Token Test; Spellacy and Spreen, 1969; for an extensive list see Murray and Chapey, 2001). As pointed out by Spreen and Risser (2003), although bedside and screening tests may be used to identify language impairments in moderate and severe aphasics (language is obviously affected, even in simple and natural communication situations), they are inappropriate or of little use to distinguish the responses of individuals with mild deficits from those with normal language skills.

Comprehensive Examinations and Aphasia Batteries

As compared to bedside and screening tests, the main purpose of comprehensive examinations of aphasia is to provide an extensive description of language skills through the administration of tests designed to explore the different areas of language (i.e., spontaneous speech, naming, oral expression, auditory and written comprehension, repetition, reading, and writing). According to the reference model of assessment, the output of a comprehensive examination may consist in the identification of a particular diagnostic of aphasia with the description of severity of deficits in each language area (clinicopathological approach), or in the localization of specific impairments affecting functional processing components of language skills (cognitive neuropsychological approach).

There are several classical comprehensive examinations and aphasia batteries. The most widely used in clinical and research settings in English are BDAE (Goodglass et al., 2000), the *Western Aphasia Battery* (Kertesz, 1982), and the *Aphasia Diagnostic Profiles* (Helm-Estabrooks, 1992). All these standardized test batteries comprise different subtests (e.g., BDAE has 27 subtests) that assess all the dimensions of language in order to diagnose and classify aphasic syndromes according to clinical localization-based classifications (i.e., Broca's, Wernicke's aphasia, etc.). For a complete description and a critical review

of these instruments, and others not reported here, see Spreen and Strauss (1998), Murray and Chapey (2001), and Spreen and Risser (2003). PALPA (Kay et al., 1992) is a comprehensive test battery directly derived from the cognitive neuropsychology approach of assessment. This aphasia battery, commonly used in the United Kingdom, consists in a set of resource materials comprising 60 rigorously controlled tests that enable the user to select tasks "that can be tailored to the investigation of an individual patient's impaired and intact abilities." The scoring and analysis of errors give the clinician a detailed profile of language abilities, including reading and written spelling, which can be interpreted within current cognitive models of language. As compared to classical batteries of aphasia, the versatile and flexible nature of PALPA is, however, lessened by the lack of standardization and validity/reliability measures.

Tests for the Assessment of Specific Aspects of Language

Specific aspects of language behavior can also be assessed through the administration of several tests. These are often used to complete aphasia batteries but some of them also are used as screening tests. Clinicians may select these tests according to the different aspects of language they want to explore in depth, but also according to the underlying theoretical model of assessment. For example, comprehension may be tested through the administration of specific tests aiming at the discrimination of phonemic sounds (Phoneme Discrimination Test; Benton et al., 1994), semantics (*Pyramids and Palm Trees Test*; Howard and Patterson, 1992), sentence length and syntactic complexity (Auditory Comprehension Test for Sentences; Shewan, 1979), commands (Token Test; De Renzi and Vignolo, 1962), or narrative discourse (*Discourse Comprehension Test*; Brookshire and Nichols, 1993). Other tests are available for measuring verbal expression, spoken and written naming, verbal fluency, reading, writing, gestural abilities, etc. An extensive list of specific language function tests can be found in Spreen and Strauss (1998), Murray and Chapey (2001), and Spreen and Risser (2003).

The Assessment of Functional Communication

Traditional tests provide useful information on linguistic abilities and language impairments in aphasia. However, performance on these tests does not necessarily predict how a person will communicate in more naturalistic settings and everyday life. Instead of focusing on the importance and the nature of deficits, the functional communication approach of assessment aims at the impact of these deficits on the person's activities and participation in society. Functional communication skills may be assessed with specific structured tests or by rating scales and inventories of communication profiles. Structured tests such as *Communication Activities of Daily Living, Second Edition* (Holland et al., 1999) and the Amsterdam–Nijmegen Everyday Language Test (Blomert et al., 1994) have been devised to explore functional communication skills using role-play in daily life activities (shopping, dealing with a receptionist, etc.) and have shown themselves to be useful to track progress over time. However, while they are certainly more ecological than comprehensive examinations and tests for specific aspects of language, structured tests of functional communication do not necessarily give reliable views of the communication skills of a person in real-life situations. In this respect, rating scales and inventories of communication profiles are closer to functional situations. For example, the *Functional Assessment of Communication Skills for Adults* (Frattali et al., 1995) is a rating protocol of 43 items, on a seven-point scale, based on the observations made by the speech–language pathologist or other significant person in the following four domains: social communication (e.g., 'refers to familiar people by name'); communication of basic needs (e.g., 'makes needs to eat'); reading, writing, and number concepts (e.g., 'writes messages'); and daily planning (e.g., 'tells time'). For a more extensive description of these functional communication tools, and others not described here, see Murray and Chapey (2001) and Spreen and Risser (2003).

Classical Tests for the Assessment of Speech, Language, Reading, and Writing Impairment in Children

The assessment of language and communication in children can take place from infancy through adolescence, when cognitive abilities are developing. Therefore, the language assessment process must not only inform on current specific abilities, but has also to capture changes over time in the level, sequence, and rate of acquisition. The interrelationship between language and other cognitive and social skills is also of primary importance. As part of a larger process, usually performed by different professionals, the evaluation of language in children should be completed by an assessment of nonverbal communication, play and social skills, perception, attention and memory, behavior, etc. Moreover, because of the major influence it has on child development, the evaluation also has to consider the familial and social environment, especially with respect to adult–child interaction. The different components (e.g., sensitivity, promptness) as well as the context (e.g., physical settings, types of play, activities) in which this interaction takes place should be analyzed through specific assessment tools or through direct observation.

The assessment of preschool children (children aged 2–5 years) and school-age children (5–10 or 12 years) is usually based upon a combination of parent interviews, standardized tests, criterion-referenced instruments, developmental scales, and observations. All these tools and methods aim to explore both receptive and expressive language abilities in semantics, morphology, syntax, phonology, and pragmatics. Collecting a communication sample is also a frequently used method to analyze communication in terms of sentence length, intelligibility of speech, vocabulary, and conversational strengths and weaknesses. Similarly to tests for aphasia, preschool and school-age tests can be divided into two major categories: screening and diagnostic tests. The purpose of screening tests is to determine if the child's communication should be explored more extensively for the presence of a possible impairment. On the other hand, the main purposes of diagnostic tests

are to establish the presence or absence of a deficit in one or more areas of language, to identify a possible difference in language development, to determine the child's eligibility for clinical services, and to identify the targets for intervention. These instruments are devised to assess language development by reference to the parameters of the normal range.

Screening Tests

Screening tests are usually inexpensive and require minimal time for administration and interpretation of results. Many norm-referenced standardized instruments may be used to establish the child's general level of expressive and receptive language functioning as well as other areas of functioning. The following are a few examples of such screening tests. The *Denver Developmental Screening Test II* (Frankenburg et al., 1990), a standardized screening battery for children from birth to age 6, is designed to test the child's abilities in the following four domains: personal-social, fine motor, gross motor, and language (including expressive–receptive vocabulary). The *Transdisciplinary Play-Based Assessment, Second Edition* (Linder, 2008) is an informal assessment for children from birth to age 6. Through observation of the child's play with family members, peers, and professionals, children can be assessed in four critical developmental areas, namely sensorimotor, emotional and social, communication, and cognition. The *Bayley Scales of Infant and Toddler Development, Third Edition* (Bayley, 2005) is a screening test designed to assess children from 1 to 42 months in many aspects of their development, including language. The *Bracken Basic Concept Scale, Third Edition* (Bracken, 2006) is an assessment tool for language skills, cognitive development, and school readiness for children from 3 to 7 years comprising a receptive and expressive evaluation. Screening tests may also consist in large batteries exploring language and cognitive functions through tasks of general verbal and nonverbal intellectual abilities. For example, the *Wechsler Intelligence Scale for Children, Fourth Edition* (Wechsler, 2004) is the most widely used measure of verbal and nonverbal intelligence in individuals from age 6 years 0 months to 16 years 11 months. As a screening tool, this battery consists in 16 subtests of verbal comprehension, perceptual reasoning, working memory, and processing speed skills.

For school-age children, some large screening batteries specifically concern academic achievement. That is, for example, the case with the *Peabody Individual Achievement Test – Revised* (PIAT-R; Markwardt, 1998), which provides a screening measure of achievement in the areas of mathematics, reading recognition and comprehension, spelling, and general information. However, most of the tasks of these large screening batteries are multifactorial and are therefore not appropriate to assess specific language or cognitive processes. For this purpose, clinicians may select among various specific screening tests for preschool and school-aged children that focus only on language. Most of these instruments are designed to explore the different language components. That is the case, for example, with the *Fluharty Preschool Speech and Language Screening Test II* (Fluharty, 2001), which explores articulation, expressive and receptive vocabulary, and composite language in children from 2 to 6 years old. An exhaustive list of norm-referenced standardized screening tests of language can be found in Paul (2001).

Comprehensive Examinations and Batteries

As for screening, some diagnostic tools are designed to explore language skills as well as other aspects of development. That is the case, for example, with the *Communication and Symbolic Behavior Scales Developmental Profile* (Wetherby and Prizant, 1998), which includes tasks exploring expressive and receptive language, symbolic play, and nonverbal communication in children from 6 to 24 months old. Another example is the *Rossetti Infant–Toddler Language Scale* (Rossetti, 1990), which is used to assess attachment, play, gestures, and pragmatics, as well as language comprehension and expression in children from birth to 3 years old. There are also several standardized comprehensive batteries of language processing that comprise tests exploring exclusively some or all of the language areas. That is the case, for example, with the *Preschool Language Scale, Fourth Edition* (Zimmerman et al., 2002), which is used to identify specific strengths and weaknesses in receptive and expressive language skills in children from birth to 6 years 11 months. The *Test of Early Language Development, Third Edition* (Hresko et al., 1999) is a standardized tool used to assess receptive and expressive language development in children aged 2 through 8. The *Test of Language Development – Fourth Edition – Primary and Intermediate* (Hammill and Newcomer, 2008) is also a classic test for spoken language used to identify oral language proficiency in children aged 4 years 0 months to 17 years 11 months. The test comprises nine subtests (e.g., syntactic comprehension, word discrimination), which measure various aspects of oral language: vocabulary, semantics, syntax, morphology, phonology, and articulation. The *Clinical Evaluation of Language Fundamentals, Fourth Edition* (Semel et al., 2003) is a multidimensional battery that can be used in individuals between the ages of 5 and 21 years to explore semantics, expressive and receptive language, and syntax, as well as working memory. The *Comprehensive Assessment of Spoken Language* (Carrow-Woolfolk, 1999b), designed for children from age 3 to 21, is another comprehensive battery of language skills, comprising 15 tests that provide an assessment of expressive and receptive skills in four language categories: lexical/semantic, syntactic, supralinguistic, and pragmatics. An extensive list of available comprehensive examinations and batteries of language for children can be found in McCauley (2001), Paul (2001), Mattis and Luck (2002), and Haynes and Pindzola (2003).

Tests for Specific Aspects of Language

Different components of language can be affected with more or less intensity in children according to the origin of developmental disorders. Therefore, the in-depth assessment of language and communication disorders in children is a critical component in the clinical process. Core tests can be used to evaluate each of the language areas in order to identify specific impairments, establish baselines, and identify precise therapeutic and intervention goals. For example, there are several core tests and instruments for the evaluation of word retrieval (e.g., naming and verbal fluency tests), phonology (e.g., word and nonword repetition tests), receptive and expressive vocabulary (e.g., word definition tests), receptive and expressive

syntax and morphology (e.g., sentence-to-picture matching tests), and pragmatic skills (e.g., narrative production, story comprehension tests).

A combination of different tests, each focusing on specific language components, may also be used to establish such a language and communication profile. To assess articulation, clinicians may select the following standardized norm-referenced specific tests: the *Goldman–Fristoe Test of Articulation 2* (Goldman and Fristoe, 2000) for children from 2 years 0 months through 21 years 11 months, the *Structured Photographic Articulation Test II* (Dawson and Tattersall, 2001) for children from 3 to 9 years old, or the *Verbal Motor Production Assessment for Children* (Hayden, and Square, 1999), which assess the integrity of the motor speech system for children 3 through 12 years. There exist various tests to assess phonology in children. One can mention, for example, the *Phonological Awareness Test 2* (Robertson and Salter, 2007) for children from 5 to 9 years old, which assesses all the pre-reading skills, like explicit phonological knowledge, that are early indicators of reading success; the *Test of Auditory Processing Skills, Third Edition* (Martin and Brownell, 2005) for children from 4 to 18 years old; or the *Hodson Assessment of Phonological Patterns, Third Edition* (Hodson, 2004) for children 2 through 8 years old. With respect to vocabulary, clinicians may select the following standardized norm-referenced tests: the *Expressive One-Word Picture Vocabulary Test Revised* (Gardner, 2000), designed to explore expressive vocabulary in children from 2 to 18 years 11 months; the *Peabody Picture Vocabulary Test* (Dunn and Dunn, 1997), an assessment tool to explore receptive vocabulary in individuals from age 2 years 6 months to adulthood; the *Montgomery Assessment of Vocabulary Acquisition* (Montgomery, 2008), a test for expressive and receptive vocabulary in children from 3 to 12 years 11 months; the *Expressive Vocabulary Test, Second Edition* (Williams, 2007), a measure of expressive vocabulary and word retrieval for children from 2.5 years old to adulthood. Syntactic abilities can be assessed with the *Structured Photographic Expressive Language Test 3* (Dawson et al., 2003), which examines expressive use of morphology and syntax in children from 4 through 9 years 11 months, or with the *Test for Auditory Comprehension of Language, Third Edition* (Carrow-Woolfolk, 1999a), which is an individually administered measure of receptive spoken grammar and syntax for children from 3 to 9 years 11 months. Finally, the *Test of Narrative Language* (Gillam and Pearson, 2004) and the *Edmonton Narrative Norms Instrument* (Schneider et al., 2005) can be used to assess narrative discourse abilities in children from 5 to 11 years 11 months and children from 4 to 9 years old, respectively. A more complete description of available diagnostic tests adapted to preschool and school-age children can be found in McCauley (2001), Paul (2001), Mattis and Luck (2002), and Haynes and Pindzola (2003).

The Assessment of Reading and Writing

The relationship between language acquisition and academic achievement is well established. Developmental disorders of language in preschool children are frequently associated with later difficulties in learning to read and write. The most common referral for a speech–language pathology assessment concerns school-age children who encounter problems in progressing beyond the developing language phase and present with difficulties in learning and acquiring communicative and academic skills. As for other populations, but especially at this stage of development, a significant difficulty in assessing school-age children arises because of important comorbidity between language and learning disorders and other cognitive and clinical pathological profiles, such as attention deficit/hyperactivity or executive function disorders. Therefore, the assessment process should include specific tests of language and communication but also instruments designed for exploring other cognitive functions, such as attention, working memory, and executive functions. In addition to formal tests, another important source of information also comes from structured interviews of the child himself, his parents, and his teacher. With respect to language, phonological processing deficits are considered as an underlying cause of dyslexia and also play a role in developmental disorders of spelling. For example, dyslexic children often show problems with word and nonword repetition tasks, phonological awareness tasks (e.g., word and nonword segmentation tasks, phoneme manipulation, etc.), and working memory tasks for verbal material (i.e., word or digit span tasks). The semantic processing is another cognitive area highly related to reading and writing. School-age children usually learn new words through reading and writing. Those who encounter problems in reading and writing often present with poor vocabulary as well as with difficulty in word association and comprehension. Therefore, the assessment procedure for written language problems should be part of a more exhaustive evaluation of language and cognition. It should also include a close control of psycholinguistic parameters (e.g., orthographic regularity, lexical frequency) that are known to play an important role in written and spoken word recognition, reading comprehension, phoneme–grapheme conversion, decoding, etc. In addition to experimental tasks, one can resort to standardized achievement or specific tests of reading and writing skills. Among the most used of these achievement tests are *PIAT-R* (Markwardt, 1998), which comprises subtests of reading comprehension, reading recognition, and spelling, and the *Wide Range Achievement Test* (Wilkinson, 1993), a brief test measuring reading recognition, spelling, and arithmetic computation. Finally, to specifically assess reading and/or writing in children, clinicians may select the *Test of Early Written Language, Third Edition* (Hresko et al., 2012) for children 4 to 10 years 11 months, or the *Oral and Written Language Scales, Second Edition* (Carrow-Woolfolk, 2011) for children from 5 to 21 years 11 months, which assess reading comprehension and written expression. For a description of specific clinical tests of reading and writing, see Spreen and Strauss (1998) and Bailet (2001).

Classical Tests for the Assessment of Speech and Language Impairment in Special Populations

Referral for speech–language assessment not only concerns aphasia and developmental deficits of language but also involves individuals of different age groups presenting with various language and communication problems. In children

and adolescents, these references include language deficits in pervasive developmental disorders (e.g., autism, Asperger's disorder), mental retardation, attention deficit/hyperactivity disorder, specific language impairment, sensory deficits (hearing loss, blindness), acquired disorders (e.g., traumatic brain injury), stuttering, etc. In adults, referral for a speed–language evaluation may be required for language and communication deficits following right hemisphere damage, traumatic brain injury, Alzheimer's disease and other forms of dementia (e.g., primary progressive aphasia, semantic dementia), stuttering and other fluency problems, etc.

In children, adolescents, and adults, clinical assessments may also concern such speech problems as dysarthria, following a stroke and neurodegenerative illnesses or accompanying cerebral palsy, acquired or developmental apraxia of speech, etc. In addition to the conventional evaluation of basic language (or speech) skills, the assessment procedure in all these special populations involves specific aspects and particularities of speech and language. For example, because of the absence of biological markers or simple diagnosis methods, the early detection of dementia often depends on various assessment tools, including language and cognitive tests performed to exclude other possible disease processes or identify specific forms of a given disease. In that particular domain, the assessment of speech and language usually includes tests allowing for differential diagnosis. For example, tests that specifically tap either semantic processing or written spelling can contribute to differentiating common disease processes in the elderly population. Semantic deficits are prominent characteristics of individuals diagnosed with Alzheimer's disease and these individuals usually differ from patients diagnosed with vascular dementia or with the semantic variant of primary progressive aphasia because of the presence of surface dysgraphia, a specific spelling disorder.

It is obviously not possible to exhaustively describe here the various tests adapted to special populations. The reader will find a more complete description of such tests in McCauley (2001), Paul (2001), Haynes and Pindzola (2003), and Spreen and Risser (2003).

Conclusion

Language production and comprehension are complex cognitive skills that should not be considered in isolation in assessment procedures. The interrelation between language and other cognitive functions has to be captured, particularly with respect to the possible influence of attention, working memory, and executive functions on linguistic abilities. If possible, clinicians should always select valid and reliable norm-referenced tests to assess language and communication. Resorting to theoretical models of language functioning also appears of primary importance and may sometimes condition the utilization of experimental, well-controlled, assessment tasks. A comprehensive assessment of language and communication is more than just an evaluation of specific skills in terms of preservation or impairment of processing components and surface structures. The scope of assessment should be widened in order to provide information about physical, social, and emotional contexts of communication, cultural differences, and economic factors. The combination of these data, obtained through assessment tools and direct observations, should then allow the clinician to establish a complete portrait of functional communication abilities.

See also: Chinese Literacy Acquisition: A Multidimensional Puzzle; Intellectual Functioning, Assessment of; Neural Basis of Reading; Neuropsychological Testing; Writing Systems, Psychology of.

Bibliography

Bailet, L.L., 2001. Written language test reviews. In: Bain, A.M., Bailet, L.L., Moats, L.C. (Eds.), Written Language Disorders. Pro-Ed, Austin, pp. 221–248.
Bayley, N., 2005. Bayley Scales of Infant and Toddler Development, third ed. (Bayley-III). Pearson, San Antonio. TX.
Benton, A.L., Hamsher, K., Varney, N.R., Spreen, O., 1994. Contributions to Neuropsychological Assessment, second ed. Oxford University Press, New York.
Blomert, L., Kean, M.L., Koster, C., Schokker, J., 1994. Amsterdam–Nijmegen everyday language test: construction, reliability, and validity. Aphasiology 8, 381–407.
Bracken, B., 2006. Bracken Basic Concept Scale, third ed. Pearson, San Antonio, TX.
Brookshire, R., Nichols, L.E., 1993. The Discourse Comprehension Test. BRK, Minneapolis.
Carrow-Woolfolk, E., 1999a. Test of Auditory Comprehension of Language, third ed. (TACL-3). Pearson Assessments, Minneapolis, MN.
Carrow-Woolfolk, E., 1999b. Comprehensive Assessment of Spoken Language. American Guidance Service, Circle Pines, MN.
Carrow-Woolfolk, E., 2011. Oral and Written Language Scales, second ed. Pearson, San Antonio, TX.
Dawson, J.I., Tattersall, P.J., 2001. Structured Photographic Articulation Test II. Janelle, DeKalb, IL.
Dawson, J., Stout, C., Eyer, J., 2003. Structured Photographic Expressive Language Test-3. Janelle Publications, DeKaub, IL.
De Renzi, E., Vignolo, L., 1962. The token test: a sensitive test to detect receptive disturbances in aphasics. Brain 85, 665–678.
Dunn, L.M., Dunn, L.M., 1997. Peabody Picture Vocabulary Test. American Guidance Service, Circle Pines, MN.
Fluharty, N.B., 2001. Fluharty Preschool Speech and Language Screening Test II. Pro-Ed, Austin.
Frankenburg, W., Dodds, J., Archer, P., Bresnick, B., Maschka, P., Edelman, N., Shapiro, H., 1990. Denver II: Screening Manual. Denver Developmental Materials, Denver.
Frattali, C.M., Thompson, C.K., Holland, A.L., Wohl, C.B., Ferketic, M.M., 1995. Functional Assessment of Communication Skills for Adults. American Speech-Language-Hearing Association, Rockville, MD.
Gardner, M.F., 2000. Expressive One-Word Picture Vocabulary Test – Revised. Academic Therapy, Novato, CA.
Gillam, R.A., Pearson, N.A., 2004. Test of Narrative Language. Pro-Ed, Austin, TX.
Goldman, R., Fristoe, M., 2000. Goldman-Fristoe Test of Articulation 2. Pearson, San Antonio, TX.
Goodglass, H., Kaplan, E., Barresi, B., 2000. Boston Diagnostic Aphasia Examination, third ed. Lippincott Williams & Wilkins, Philadelphia.
Hammill, D.D., Newcomer, P.L., 2008. Test of Language Development – Primary and Intermediate, fourth ed. Pearson, San Antonio, TX.
Haynes, W.O., Pindzola, H.R., 2003. Diagnosis and Evaluation in Speech Pathology, sixth ed. Allyn & Bacon, Boston.
Hayden, D.A., Square, P.A., 1999. Verbal Motor Production Assessment for Children. Psychological Corporation, San Antonio, Texas.
Helm-Estabrooks, N., 1992. Aphasia Diagnostic Profiles. Riverside Publishing, Chicago.
Hodson, B., 2004. Hodson Assessment of Phonological Patterns, third ed. Pro-Ed, Austin, TX.
Holland, A.L., Frattali, C.M., Fromm, D., 1999. Communication Activities of Daily Living, second ed. Pro-Ed, Austin.
Howard, D., Patterson, K.E., 1992. The Pyramids and Palm Trees Test. Harcourt Assessment, Oxford.

Hresko, W.P., Reid, D.K., Hammill, D.D., 1999. Examiner's Manual: Test of Early Language Development, third ed. Pro-ed, Austin, TX.
Hresko, W., Herron, S., Peak, P., Hicks, D.L., 2012. Test of Early Written Language, third ed. Pro·Ed, Austin, TX.
Kay, J., Lesser, R., Coltheart, M., 1992. Psycholinguistic Assessments of Language Processing in Aphasia (PALPA). Lawrence Erlbaum Associates, Hove, England.
Kertesz, A., 1982. Western Aphasia Battery. Grune & Stratton, New York.
Linder, T.W., 2008. Trans-disciplinary Play-Based Assessment, second ed. Paul H. Brookes Publishing Co., Inc, Baltimore.
Markwardt, F.C., 1998. Peabody Individual Achievement Test Revised. American Guidance Service, Circle Pines, MN.
Martin, N.A., Brownell, R., 2005. Test of Auditory Processing Skills, third ed. (TAPS-3). Academic Therapy Publications, Novato, CA.
Mattis, S., Luck, D.Z., 2002. Neuropsychological assessment of school-aged children. In: Segalowitz, S.J., Rapin, I. (Eds.), Handbook of Neuropsychology, Child Neuropsychology, vol. 1. Elsevier Science, New York.
McCauley, R.J., 2001. Assessment of Language Disorders in Children. Lawrence Erlbaum Associates, Mahwah, NJ.
Montgomery, K., 2008. MAVA-Montgomery Assessment of Vocabulary Acquisition. Super Duper Publications, Inc, Greenville, South Carolina.
Murray, L.L., Chapey, R., 2001. Assessment of language disorders in adults. In: Chapey, R. (Ed.), Language Intervention Strategies in Aphasia and Related Neurogenic Communication Disorders, fourth ed. Lippincott Williams & Wilkins, Philadelphia, pp. 55–126.
Paul, R., 2001. Language Disorders from Infancy through Adolescence: Assessment and Intervention, second ed. Mosby, Toronto.
Reitan, R.M., 1991. Aphasia Screening Test. Reitan Neuropsychology Laboratory, Tucson.
Robertson, C., Salter, W., 2007. Phonological Awareness Test 2. LinguiSystems, East Moline, IL.
Rossetti, L., 1990. Rossetti Infant–Toddler Language Scale. LinguiSystems, East Moline, IL.
Schneider, P., Dubé, R.V., Hayward, D., 2005. Edmonton Narrative Norms Instrument. Retrieved from University of Alberta Faculty of Rehabilitation Medicine website: http://www.rehabresearch.ualberta.ca/enni.
Semel, E., Wiig, E.H., Secord, W.A., 2003. Clinical Evaluation of Language Fundamentals, fourth ed. Psychological Corporation, Toronto.
Shewan, C.M., 1979. Auditory Comprehension Test for Sentences (ACTS). Biolinguistics Clinical Institutes, Chicago.
Spellacy, F., Spreen, O., 1969. A short form of the token test. Cortex 5, 390–397.
Spreen, O., Risser, A.H., 2003. Assessment of Aphasia. Oxford University Press, New York.
Spreen, O., Strauss, E., 1998. A Compendium of Neuropsychological Tests, second ed. Oxford University Press, New York.
Wechsler, D., 2004. The Wechsler Intelligence Scale for Children, fourth ed. integrated. Harcourt Assessment, San Antonio, TX.
Wetherby, A.M., Prizant, B.M., 1998. Communication and Symbolic Behavior Scales-Developmental Profile. Applied Symbolix, Chicago.
Wilkinson, G.S., 1993. Wide Range Achievement Test 3. Harcourt Assessment, San Antonio, TX.
Williams, K.T., 2007. Expressive Vocabulary Test, second ed. Pearson, San Antonio, TX.
Zimmerman, I.L., Steiner, V.G., Pond, R.E., 2002. Preschool Language Scale, fourth ed. Harcourt Assessment, San Antonio, TX.

Classifications in Popular Music

Alex van Venrooij, University of Amsterdam, Amsterdam, The Netherlands
Vaughn Schmutz, UNC Charlotte, Charlotte, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The categorical system of popular music, such as genre categories, is a highly differentiated and dynamic classification system. In this article we present work that studies different aspects of these categorical systems in popular music. Following the work of Paul DiMaggio, we focus on four questions: How does the position of popular music vary in terms of its place in the cultural hierarchy and the manner in which it is legitimated? How do social and symbolic boundaries mutually shape each other? What are the consequences of boundary crossing in popular music, and do these effects vary across different subfields of popular music? How do classification systems in popular music change over time?

Introduction

From a sociological perspective, musical categories are based on cultural conceptions about types of music that are similar to or dissimilar from one another. Rather than being based on formal musical properties or sonic qualities alone, such conceptions of similarity or dissimilarity are influenced by a range of factors, including social relations, artistic conventions, characteristics of producers and audiences, broad systems of classification, and even production technologies. By describing musical categories as cultural, we highlight the shared meanings that enable a common system for classifying and evaluating music. Although they can become so taken for granted that they appear natural and inevitable, musical categories are constructed and upheld by social practices and cultural assumptions (Roy and Dowd, 2010). Yet while the boundaries between musical categories are symbolic in character, they create consequential distinctions between genres as well as between legitimate and illegitimate musical forms and musicians.

Perhaps the most familiar type of musical category is a genre, which Paul DiMaggio (1987: p. 441) defines as a set of 'artworks classified together on the basis of perceived similarities.' Relative to other cultural fields, popular music is characterized by a large and growing number of genres that vary considerably in terms of how much popularity they achieve and how long they endure (Lena, 2012). Indeed, some features of classification systems change and others persist as genres emerge, transform, and disappear. Despite the seemingly idiosyncratic characteristics of this process, genres often follow similar trajectories (Lena, 2012) and their boundaries both reflect and reinforce broader patterns of social relations (Roy, 2004; Roy and Dowd, 2010).

The symbolic boundaries that differentiate music genres are not the only categorical distinctions of interest to sociologists. Symbolic divisions also separate music that is legitimate from music that is illegitimate, music that is authentic from music that is inauthentic, music that is newsworthy from music that is unworthy of media attention, music that wins prestigious awards from music that is undeserving of such esteem, and so on. Such divisions may transcend generic boundaries, but reinforce boundaries based on artistic conventions or associated with social characteristics like race or gender. Musical classification systems vary not only in terms of how many genres they include, but also in the degree to which their categories are strongly bounded or hierarchically ranked (DiMaggio, 1987; van Venrooij, 2009). Indeed, the cultural standing of popular music itself and of particular popular genres can vary greatly across time and space.

Classification Systems

The work of DiMaggio (1987) has provided the theoretical foundation for much research on classification systems. His concept of artistic classification systems offers a number of insights. First, the study of classification systems needs to take into account both the consumption and production of art. According to DiMaggio (1987: p. 441), artistic classification systems consist of "the way that the work of artists is divided up both in the heads and habits of consumers *and* by the institutions that bound the production and distribution of separate genres" [italics in original]. On the one hand, classification systems arise out of processes of social distinction, whereby consumers use cultural objects to mark social boundaries. These 'ritual classifications' can thus be influenced by social structural factors at the societal level – such as stratification systems, elite cohesion, social and geographic mobility, etc. – that generate demand for cultural boundaries. On the other hand, classification systems are also influenced and mediated by 'classification processes' at the production side. DiMaggio (1987) identifies *commercial* classifications (the classifications used by commercial producers to market their products), *administrative* classifications (created and enforced by the state), and *professional* classifications (classifications driven by the incentives of artists to differentiate and mark boundaries). The study of classification systems thus needs to be attentive to the extent to which the categorical demands of consumers and the categories used by producers overlap, diverge, or mutually reinforce each other.

Second, DiMaggio also argues that classification systems have both a cognitive/symbolic as well as organizational/material dimension. This analytical distinction allows for analyzing processes of institutionalization whereby symbolic

categories can become 'objectified' in organizational practices, network structures of artists, or unequal resource distributions. If the symbolic and material boundaries of categories strongly overlap, chances are these categories become seen as taken for granted and 'natural.'

Third, the concept of classification *systems* implies the study of collections of categories, not of isolated genres or categories. It thus focuses on the relationship between categories – since categories are always defined vis-a-vis other categories. A single category is a logical impossibility, similar to clapping with one hand (see Mohr and White, 2008).

Fourth, artistic classification systems therefore are also studied in their formal characteristics. DiMaggio distinguishes four such formal characteristics. For one, artistic classification systems can vary in their *differentiation*, or the degree to which genres are institutionally bounded. Fields that are highly segmented with many identifiable genres are highly differentiated. Second, artistic classification systems can vary in the degree to which genres are ranked by prestige, which is an indicator of *hierarchy*. In more hierarchical systems, genres diverge widely in prestige and command unequal resources. A third dimension of variation is *universality*, or the degree to which there is agreement among members of a society in the ways they recognize and classify genres. Finally, artistic classification systems can vary in their *boundary strength*, or the degree to which genre boundaries are highly ritualized and difficult to transgress.

Although DiMaggio has mostly developed this theory for the analysis of the construction of the 'high culture' category in the United States, it has also proven relevant for understanding the categorical position and categorical systems of other fields (e.g., Baumann, 2007 on film; Verboord and van Rees, 2008 on literature; Zhao, 2005, 2008 on wine), including popular music. Below we will first discuss our research that builds on DiMaggio's work in comparing the legitimation of popular music in cross-national comparative perspective (cf Janssen et al., 2011). His work not only suggests reasons why the hierarchical classification system has been challenged in recent decades, but also why this process might vary between nations, and how this affects the trajectory of popular music. Second, we focus on research that investigates the homology between aesthetic and social boundaries, which develops out of the confluence of ritual and other types of classifications. Third, we investigate the dimension of boundary strength and whether there are penalties associated with crossing genre boundaries in popular music. Finally, we discuss research that looks at how classification systems in popular music develop over time, and how this work can be extended to analyze how new categories within popular music develop.

The Legitimacy of Popular Music in Comparative Perspective

Paul DiMaggio's (1987) influential theory of artistic classification systems has proven useful in analyzing cross-national comparative differences in the rise of popular music in the cultural hierarchy. His theory proposes a number of social structural factors and institutional mechanisms that might influence the differentiation, hierarchy, universality, and boundary strength of ritual classification systems. He has also described how the 'high culture system' in the United States gave way to a more differentiated, less hierarchical, less universal, and more weakly bounded classification system. His theory, however, also suggests that this process of 'declassification' might vary across nations, and thus for the position of popular music within the classification system.

Comparative research by Susanne Janssen et al. (2011) shows, indeed, that the legitimacy of popular music as a cultural category varied greatly over time and across countries. Such research supports the idea that social structural characteristics are associated with variation in classification systems and, therefore, with the relative position of popular music. Our own related research shows that these cross-national differences in classification systems also influence *how* a previously suspect cultural form, like popular music, is increasingly legitimated. Building on the work of Motti Regev (1994) and Shyon Baumann (2007) who focus on the legitimating ideology that promoted the aesthetic mobility of rock music and film, we investigated the evaluative criteria used to assess popular music releases in three countries; however, we criticized the idea that the rise of popular music in the cultural hierarchy could only be legitimated by assessing the worth of this cultural form as a 'serious art form' or, in other words, by the use of autonomous criteria of art (van Venrooij and Schmutz, 2010). This idea of artistic legitimation, we argued, takes for granted that the process of declassification in itself challenges the dominance and acceptance of 'art' as the single legitimate category. The differentiation of the music field and the corresponding flattening of its hierarchy could also result in the 'autonomization' of popular music and its aesthetic, thus making it less likely that popular music needs to be legitimated as art, but could also be legitimated by a 'popular aesthetic.'

In our comparison of popular music reviews in the United States, Germany, and the Netherlands, we find that differences in the boundary strength, hierarchy, and universality of their respective classification systems – based on differences in the educational and stratification systems in each country as well as their historic national repertoires – indeed lead to variation in ways of legitimating popular music. In Germany, which has a more hierarchical, universal, and ritually bounded classification system, music reviewers draw more exclusively on high art criteria to legitimate popular music albums. By contrast, the United States is characterized by a less hierarchical, universal, and ritually bounded classification system; as expected, American critics are much more likely to mix high art and popular aesthetic criteria in their reviews. Finally, Dutch reviewers are more likely to employ exclusively popular aesthetic criteria in their discourse about popular music albums. So although popular music gained much in attention from established critical institutions in all three countries, the differences in how they evaluated popular music were striking. In general, comparative research of this kind thus clearly shows that popular music itself is a cultural category and that its meanings vary over time and from one place to another.

Social and Symbolic Boundaries

The symbolic boundaries that distinguish one musical category from another are shaped by the social boundaries that divide

groups of people from one another. However, as argued by DiMaggio, the homology between aesthetic and social boundaries can be mediated and influenced by the commercial classifications used by for-profit firms. Genre classifications in the United States, for example, have historically varied in how tightly coupled they are with social boundaries based on race, class, or gender. In the early American recording industry, the dominant categorical system was largely focused on a class-based distinction between classical and popular music, yet Bill Roy (2004) demonstrates how musical categories also became overtly racialized and thereby more closely linked to social categories. 'Race records' and 'hillbilly music' were constructed musical categories by which the major recording firms explicitly linked musical styles to social categories based on race and class. The separation of the two categories had little to do with musical differences – rather, it was a business decision introduced by the Okeh Recording Company in its 1921 catalog – but other recording firms quickly adopted it and the categories soon became institutionalized. Thus, symbolic boundaries (i.e., music genres) and social boundaries (i.e., race) became tightly coupled in the field of American popular music with long-term consequences.

Although musical categories have become less overtly racialized over time, the *de facto* segregation of the musicians' union and of radio airplay continued to hinder the success of African American performers in the mainstream US music market (Dowd and Blyler, 2002). Furthermore, Keith Negus (1999) argues that this homology between racial and musical categories has implications for the cultural and organizational divisions of contemporary recording firms. In addition to being isolated from the resource-rich and decision-making centers of the major labels, genres like rap or salsa are also constrained by the cultural assumptions about their performers and their audiences. With regards to rap, major labels produce a 'discourse of the street' that constrains the range of musicians they embrace through a narrow cultural conception of authentic rap production, distribution, and performance. Similarly, the genre culture of salsa remains distant from the dominant interests of the major labels and is contained in distinct Latin music divisions that correspond to broader social divisions based on commodified Latin identities (Negus, 1999). In short, cultural assumptions about racial and ethnic groups are linked with musical categories in the contemporary recording industry.

Likewise, symbolic boundaries linked to gender have shaped the opportunities available to men and women in the popular music industry. Typically, the connection between gender and genre has constrained the prospects for women in popular music (Roy and Dowd, 2010). For instance, scholars across disciplines have called attention to the ways in which the masculinization of many popular music genres has limited the inclusion of women in jazz ensembles, rock and heavy metal bands, rap performance, and so on. While some opportunities have opened for women in popular music, cultural assumptions about gender frequently shape the roles and styles that are available to female musicians. For example, alternative rock bands in the 1990s were more open to women than their rock music predecessors, evidenced by a notable increase in female bass players during this period. Yet despite this slight decoupling between masculinity and rock music, Mary Ann Clawson (1999) found that the feminization of bass playing limited women to a more peripheral role as men gravitated toward higher status positions like lead singer or guitarist. Beyond alternative rock, Patti Donze (2011) has demonstrated that women represent a narrower range of artistic types relative to men in the mainstream music field. She uses latent class analysis of the labels attributed to male and female musicians to show how gendered and sexualized schemas relate to categorical distinctions in popular music.

The interplay between gender and symbolic boundaries can also be seen in research on cultural legitimacy and consecration in popular music. First, consider newspaper coverage of popular musicians in the United States, France, Germany, and the Netherlands from 1955 to 2005 (Schmutz, 2009). As genres gained greater legitimacy in the popular music hierarchy, measured by the amount of attention received in elite newspapers, their media coverage became increasingly male dominated. In other words, the more legitimate a genre became, the more its media coverage focused on male musicians and this pattern was consistent regardless of country or the genre that was gaining greater attention. Beyond the sheer quantity of media attention, cultural frameworks about art and gender influence the ways that music critics write about male and female musicians (Schmutz and Faupel, 2010). Not only are female musicians less likely to have their albums consecrated among the 'greatest of all time,' but also the few who do achieve such status are legitimated in subtly different ways. Whereas men are more likely to be described as autonomous, historically important, and with 'high art' terms, women are more likely to be legitimated through their emotional authenticity and their ties to others (Schmutz and Faupel, 2010).

The Consequences of Categories

The sociology of art considers the creation of artistic works as a collective and social enterprise (Becker, 1982). To account for the collective nature of the production, distribution, and reception of art, sociologists often point to the taken for granted and shared understandings that constrain and enable this collective production of art (cf Roy and Dowd, 2010). Especially relevant for the collective production of art are the shared understandings of categorical systems such as genres. Genres can be seen as consisting of rules and conventions shared by participants in particular genre worlds about ways of playing an instrument (the 'swing' of jazz drummers), the use of the voice (the 'rawness' of rock singers), the inclusion of instruments (the 'banjo' in bluegrass), or forms of presentation (the 'long hair' of heavy metal bands) or who can be legitimately part of the genre world (whether or not 'white' rappers are 'real,' etc. – see below). Genre categories can function as shorthand for a collection of conventions and enable the interaction between musicians – as when asked to play the rhythm 'more funky,' or put on a 'rock voice.' Genre categories are therefore frequently invoked in decision-making processes in the production and distribution of art (Negus, 1999), the evaluation and assessment of worth by critics (Schmutz, 2009; van Venrooij, 2009) as well as by consumers to organize their taste patterns (Bourdieu, 1984).

Confusion about the characteristics of a genre can impede its development. For instance, Christophe Boone et al. (2012) show how attempts to institutionalize modernistic music in Brussels was hampered by the fragmentation of the cultural elites and their ideas about which example should be followed: Francophone, Germanic, or hybrid form of modernistic music. The importance of consensus building on the key characteristics could also play a role in the mobilization and legitimation of popular musical forms. Genres that are 'hybrids' might be less easily institutionalized than genres that only marginally differ from already established genres. Such mutual understandings and expectations also tend to restrain the opportunities of producers to deviate from 'categorical imperatives' without sanction. Although musicians often do not want to be confined by genre categories, the collaborative nature of music production implies that musicians are bounded by the expectations others have of their genre identity. Musicians who, for example, depart from genre expectations can expect negative reactions from other participants in the art world, such as audiences, critics, or industry personnel. The history of popular music provides many anecdotal examples of this mechanism. Well-known cases include the criticism Bob Dylan received for deviating from the folk genre by introducing electronic instruments at the Newport Folk Festival or Neil Young being sued by his record label Geffen in 1983 for making genre-deviating albums that they considered musically uncharacteristic of Young's previous recordings.

A burgeoning literature on 'multiple category membership' indeed shows the negative consequences of transgressing categorical boundaries (for a review, see Hannan, 2010). Products and producers who do not have a 'crisp' identity but combine or cross categorical boundaries are found to be penalized in terms of several indicators of success, such as market appeal and critical ratings (Hsu, 2006). The main argument is that audiences are confused by 'fuzzy' identities, and react negatively or ignore products or producers they cannot make sense of with the schematic understandings of their existing classification systems.

These studies, however, tend to assume and study the effects of fuzziness in fields and markets that have established, stable, well-institutionalized classification systems. When classification systems are emergent, weak or in flux, and have not yet crystallized in well-developed categorical schemas, fuzziness might not lead to the same negative penalties as in mature, institutionalized classification systems. When classification systems are weakly bounded, the penalties for products that do not have a clear categorical fit tend to be lower. In other words, the work of DiMaggio (1987) provides reasons to assume that the penalties for 'boundary crossing' might be contingent on the boundary strength of the classification systems.

To investigate this possibility, we compared the constraining effects of genres within two subfields of popular music that are likely to have varying boundary strength of their classification systems (van Venrooij and Schmutz, 2012). We argue that classification systems in fields of large-scale production – the mainstream market – are more strongly institutionalized and therefore fuzziness tends to detract from being successful in that subfield. In fields of restricted production, the 'independent market,' however, the classification system – we argue – is weaker and more ambiguous, and thus fuzziness tends to not be penalized, and even valorized within this subfield.

We give several reasons why this difference between subfields might occur. The *first* reason concerns the entry barriers and competitive pressures in both subfields. DiMaggio (1987: p. 451) suggests that classification systems will be strongly differentiated, but weakly institutionalized in fields with lower entry barriers. If more participants enter the field than the field can absorb, the incentive among artists increases to 'make one's name' by attaching themselves to a new genre and thereby create more finely grained classification systems. In low entry fields, the classification system can therefore become highly differentiated but poorly institutionalized and only weakly bounded. Even though these differentiated categories might make sense to insiders and the ones who proposed them, the more specific the categories become, the less broadly accepted are the boundaries between them. As one of DiMaggio's (1987: p. 448) propositions reads, 'the more differentiated the system of classification, the less universal.' In the 'independent' subfield, the entry barriers are arguably lower than in the 'major' subfield. The 'indie labels' are, for example, more closely connected to the different music scenes and generally provide more opportunities for new artists than the major labels who invest more in their 'stars.'

The *second* reason follows from differences in the structure of organizations that produce and distribute the cultural products of artists in the restricted and large-scale fields. The 'autonomous' subfield consists of small entrepreneurial organizations that follow a 'professional logic' and delegate more discretion to artists or brokers, grant them more autonomy, and assign lesser influence to managers. This balance between artists and managers is important since they often have conflicting interests. Artists have a professional incentive to innovate, challenge existing tastes, and generate new cultural forms. Managers, however, are mostly concerned with routinely sorting the unfamiliar into the familiar to assure survival of the product within the 'decision chain' of the organization or tailor the sound of new bands in the mold of accepted genres in order to secure acceptance of the music 'downstream' by radio and retailers. The 'professional logic' of small entrepreneurial organizations places more weight on the importance of the former, and the greater autonomy of the artists can therefore strengthen the possibility that artists create music that does not fit with established categories or assign their own classifications onto their work, and thereby push toward greater differentiation of the genre classification system. The 'commercial logic' of larger bureaucratic culture-producing organizations grants less artistic autonomy to artists, gives more power to managers and marketers who are therefore more likely to uphold established genre categories. Moreover, the practice of 'portfolio management' at major labels assigns genres to different divisions within the company, which are monitored and controlled as separate units – a practice which also works against 'genre crossing' (Negus, 1999).

The *third* reason lies with the intended audiences of both subfields. The heteronomous field orients itself mainly toward success among mainstream audiences and the 'modal' consumer, whereas the autonomous field focuses on symbolic success among other producers, 'insider audiences,' aficionados, and critics. The insider audiences of the autonomous

field are equipped with the necessary 'code to decode' to understand and value the cultural products from the viewpoint of the field, and are therefore less likely to be confused by aesthetic innovation and experimentation than mainstream audiences. Rather than penalizing artists for crossing categorical boundaries, insider audiences are therefore more likely to value their quest for distinction and innovativeness.

Our study found that combining or hybridizing genre categories indeed leads to more severe penalization for albums in the mainstream subfield than in the independent subfield. Negative reactions by audiences and critics occur especially for musicians produced by major record labels and whose music is targeted at mainstream audiences. However, our study also shows that genre spanning has no negative, and even positive consequences for musicians produced by 'independent' record labels and who aim at specialized audiences. This finding thus corroborates DiMaggio's theory that art worlds might vary in terms of how strongly they reinforce genre categories.

The Dynamics of Categories

How do classification systems in popular music develop over time? Using DiMaggio's framework, the classification system of popular music could be described as showing trends in the direction of becoming more differentiated, less hierarchical, less universal, and more weakly bounded. The genre diversity, for example, seems to have increased considerably in recent decades and the field has become associated with a plethora of genres – especially those fields such as electronic dance music (which according to some counts consists of more than 300 subcategories) that have absorbed a large number of newcomers. This is on the one hand simply the result of the cumulative development of new genres, but also signals changes in the institutional logic of the recording and radio industry that have focused on more and more specialized genre niches. The way hierarchical boundaries are drawn within the field also seems to have changed. Until the mid-1980s, hierarchical distinctions between 'serious' and 'authentic' rock music and 'frivolous' commercial pop were more firmly based on the application of the principles of distinction that mirrored the boundary between 'art' and 'commerce.' In contrast, the contemporary field of popular music apparently is characterized by an 'intermediate aesthetic,' which combines commercial and artistic values. Genre boundaries have apparently also weakened. The rise of hybrid genres have been the outcome of the search for artistic innovation, which especially in the more autonomous sectors of the field, have been valorized rather than penalized, as we described in the previous section. Finally, the universality or consensus could also be said to have weakened, as a result of the increased differentiation in the field. As mentioned, one of DiMaggio's (1987: p. 448) propositions suggests that differentiation leads to loss of universality. People will, in other words, find it easier to agree on broad categories such as 'rock,' 'country,' 'rap,' rather than on very specific categories such as 'neoelectro' or 'futuristic breakbeat.'

To study whether these dynamic changes occurred in the *cognitive* classifications systems of critics, van Venrooij (2009) compared the meanings that critics associated with different genre terms between 1985–86 and 2004–05. By coding and identifying the various 'aesthetic codes' used by critics to describe and evaluate musical characteristics, this research was able to model the position of genres in an 'aesthetic discourse space.' By analyzing relations of similarity and difference between genres in this discourse space, the cognitive classification system of reviewers could be visualized.

The analysis of the relational structure of the classification system in 1985–86 and 2004–05 reveals that – in the way critics interpret and associate aesthetic elements with genres – the boundaries between genres have weakened. In general the overlap between genre categories increased, and more specifically the genre categories of 'pop' and 'rock' became more similar and synonymous. These two genres were also joined by 'R&B' and 'hip-hop' in the core of the classification system, suggesting a blurring of boundaries between these often socially differentiated categories, and signaling the movement of hip-hop into the center of the popular music field. This analysis also allowed the comparison of genres in their 'fuzziness' over time. For instance, jazz as a category seemed to hold relatively specific meanings for critics in 1985 but became a much more general and broadly defined category in 2005. Funk, on the other hand, went from a fuzzy category in 1985 to having more restricted meaning in 2005. Hence, this type of relational analysis of critical discourse can help assess changes over time in the meaning of categories.

To model the 'discourse space,' van Venrooij (2009) made use of the ecological notion of a multidimensional space – or Blau space. Although this research used ecological ideas to *describe* and model cultural classification systems, the use of ecological theory also holds the promise of possibly *explaining* changes in classification systems. Ecological theory – as applied to organizational forms – assumes that 'forms' occupy niches and compete with each other in a bounded resource space. Regions within this resource space have limited 'carrying capacities' and can only sustain a certain number of forms until they are saturated. New forms therefore develop within unsaturated niches that already contain similar forms to increase their 'legitimacy' but this 'emulation' turns to 'differentiation' when the 'carrying capacity' for certain niches is reached.

Drawing on this ecological approach to model how cultural forms, such as genres, are positioned in a bounded 'aesthetic space' may enable researchers to determine whether the dynamics of genres also follows ecological mechanisms. For example, do new genres emerge 'nearby' existing genres – where 'nearby' indicates similarity in terms of aesthetic conventions, or do existing genres change over time in their characteristics by adopting element from 'unsaturated' areas in the aesthetic space? This would thus allow the analysis of *endogenous* mechanisms of cultural change, which might occur through processes of mutual orientation and differentiation among actors within the field, mediated by the information provided by critics, magazines, and journalists.

Conclusion

The work of Paul DiMaggio has stimulated much interest in the study of classification systems. Although his theory was

predominantly developed to account for the rise and fall of the high culture system in the United States, here we have presented recent work that shows the usefulness of his thinking to the study of classification systems in popular music. This overview has, of course, not been exhaustive. Many more examples of research on classifications in popular music could be given. Here we have focused on our own contributions to this literature concerning the cross-national variations in the position of popular music in the cultural hierarchy, the construction of homologies between social and symbolic boundaries, and the consequences of boundary crossing and changes in the meaning of genre categories.

See also: Cultural Mediators and Gatekeepers; Culture, Cognition and Embodiment; Culture, Production of: Prospects for the Twenty-First Century; Symbolic Boundaries.

Bibliography

Baumann, S., 2007. Hollywood Highbrow. Princeton University Press, Princeton, NJ.
Becker, H.S., 1982. Art Worlds. University of California Press, Berkeley, CA.
Boone, C., Declerck, C.H., Rao, H., van den Buys, K., 2012. Out of tune: the rise and fall of modernistic music in Brussels, 1919–1939. Poetics 40, 44–66.
Bourdieu, P., 1984. Distinction: A Social Critique of the Judgment of Taste. Routledge, London.
Clawson, M.A., 1999. When women play the bass. Gender and Society 13, 193–210.
DiMaggio, P., 1987. Classification in art. American Sociological Review 52, 440–455.
Donze, P.L., 2011. Popular music, identity, and sexualization: a latent class analysis of artist types. Poetics 39, 44–63.
Dowd, T.J., Blyler, M., 2002. Charting race: the success of black performers in the mainstream recording market, 1940 to 1990. Poetics 30, 87–110.
Hannan, M.T., 2010. Partiality of memberships in categories and audiences. Annual Review of Sociology 36, 159–181.
Hsu, G., 2006. Jacks of all trades and masters of none: audiences' reactions to spanning genres in feature film production. Administrative Science Quarterly 51, 420–450.
Janssen, S., Kuipers, G., Verboord, M., 2011. Comparing cultural classification: high and popular arts in European and US elite newspapers, 1955–2005. Kölner Zeitschrift Für Soziologie Und Sozialpsychologie 63 (51), 139–169.
Lena, J.C., 2012. Banding Together: How Communities Create Genres in Popular Music. Princeton University Press, Princeton, NJ.
Mohr, J.W., White, H.C., 2008. How to model an institution. Theory and Society 37, 485–512.
Negus, K., 1999. Music Genres and Corporate Cultures. Routledge, London.
Regev, M., 1994. Producing artistic value: the case of rock music. Sociological Quarterly 35, 85–102.
Roy, W.G., 2004. 'Race records' and 'hillbilly music': institutional origins of racial categories in the American commercial recording industry. Poetics 32, 265–279.
Roy, W.G., Dowd, T.J., 2010. What is sociological about music. Annual Review of Sociology 36, 183–203.
Schmutz, V., 2009. Social and symbolic boundaries in newspaper coverage of music, 1955–2005: gender and genre in the US, France, Germany, and the Netherlands. Poetics 37 (4), 298–314.
Schmutz, V., Faupel, A., 2010. Gender and cultural consecration in popular music. Social Forces 89, 685–707.
van Venrooij, A., 2009. The aesthetic discourse space of popular music: 1985–86 and 2004–05. Poetics 37 (4), 315–332.
van Venrooij, A., Schmutz, V., 2010. The evaluation of popular music in the United States, Germany and the Netherlands: a comparison of the use of high art and popular aesthetic criteria. Cultural Sociology 4 (3), 395–421.
van Venrooij, A., Schmutz, V., 2012. Logics and Consequences of Boundary Crossing in the Field of Popular Music. Paper presented at the Annual Meeting of the American Sociological Association. Denver, United States.
Verboord, M., van Rees, K., 2008. Cultural classifications in literary education: trends in Dutch literary textbooks, 1968–2000. Cultural Sociology 2 (3), 321–343.
Zhao, W., 2005. Understanding classifications: empirical evidence from the American and French wine industries. Poetics 33, 179–200.
Zhao, W., 2008. Social categories, classification systems, and determinants of wine price in the California and French wine industries. Sociological Perspectives 51, 163–199.

Classification, Understandings of in the Social Sciences

Hubert Feger, Free University of Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 1966–1973, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article covers the use of classification in the social sciences. Although in the past this tool was considered more suitable for the natural sciences, it has been applied to social phenomena more recently. The fundamental goal of classification is to find structures common to a group of objects, using properties to classify the objects into subgroups based on the similarity of their properties. This article details each step in the classification process and concludes with a discussion of how to evaluate a classification.

Classification is the assignment of objects to classes. For example, an educational researcher might want to establish a taxonomy of teaching styles that covers all possible approaches to teaching. A psychologist studying personality might be interested in whether children can be grouped into categories according to their patterns, or profiles, of personality traits. A sociologist might be interested in whether certain combinations of characteristics of urban areas (such as average socioeconomic status, crime rate, and building types) occur much more often than other combinations. A biologist might want to study whether animals showing a particular phenotype have specific combinations, or patterns, of genetic codes. In all these cases, objects (such as teachers, children, urban areas, and animals) are classified based on their patterns of some observable characteristics (such as teaching behaviors, personality traits, city characteristics, and genes).

The task of classifying objects poses several problems. The objects to be classified, the properties based on which they are classified, and the way of assessing similarities among objects have to be specified. The aim is to identify individual classes, to decide how many classes are warranted, and to establish procedures for identifying which class each object should be assigned to. Eventually, the success of the classification system must be evaluated. There are problems associated with each of these steps, and different solutions have been suggested for each problem.

In a sense, this article re-creates the process of establishing and testing a classificatory system. First, we define the topic more precisely, clarify the terminology, and provide some examples. Then, we present an outline of the first steps in a classification process and discuss one of the most debated issues: What is the concept of a class? After this, we deal briefly with procedures for assigning objects to classes and with the – frequently neglected – question of how to evaluate the resulting classification system.

Definition of the Topic

Terminology

Depending on the research tradition, the objects to be classified into a system are called elements, cases, units, exemplars, specimens, or items. They are the sources or 'carriers' of properties, characteristics, or variables. These properties may be dichotomous or polytomous, qualitative or quantitative. A property can only be useful in a classification, if it varies within the set of objects, that is, if at least two different values (e.g., categories, states, and labels) on the respective property occur in the sample. When more than one property is used to characterize an object, the object can be described as a vector of values, a profile, a set of symptoms, or a pattern of features.

Sometimes, the data do not consist of objects and their properties, but of measures of relations between objects, such as their similarity, likeness, or belonging together. For example, a researcher might ask participants to rate the similarity between different politicians. The similarity ratings can then be used to classify politicians into groups. Based on this classification, the researcher can then study on which features people's perceptions of similarities between politicians are based.

The crucial assumption underlying classification is that objects are elements of a class, of a set, of a partition, or – in biology – of a taxon. In other terminologies, the terms 'category' or 'cluster' are also used. Classification is the process of finding classes and of assigning entities to these classes. The end product of this order-creating process, however, is often also referred to as 'classification.' To stress this distinction, the term 'classification system' can be used for the end product, although in clinical psychology and biology the word 'taxonomy' is more common. Identification is the assignment of a specific case or object to (usually only) one of the classes.

Limits of Discussion

In studying classification in the behavioral and social sciences, we have to distinguish between two questions. (1) What theory and empirical evidence are available about how people classify objects? (2) What theories and methods are used to create classification systems? The first question will not be dealt with here (*see* Concept Learning and Representation: Models). The second question refers to the formal and empirical procedures used for defining classes and the rules that have evolved for assigning cases to classes. This second question is what concerns us here; it will be discussed from a conceptual rather than a statistical point of view. For more statistically-oriented discussions of classification methods, *see* Statistical Clustering; Mixture Models in Statistics.

One commercially 'booming' application of classification that this article does not refer to is in biometric authentification. The aim of these methods is to use combinations of characteristics of individuals to uniquely identify each individual. Thus, conceptually, the task is to assign one of very many cases into a class, when there might be as many cases as classes. In this process, an observed pattern of features, or even an observed set of such patterns, is matched against a stored list of patterns. For example, DNA is used in this way in forensic criminology, as are records of the past *modus operandi* of individual criminals in detective work. This is, however a fairly unique example of classification, because the number of classes is intended to be equal to the number of cases, which is rarely the case in the social and behavioral sciences.

Purposes of Classification

The fundamental purpose of classification is to find structure. Typically, a large number of objects is reduced to a much smaller set of classes without too much loss of information about the objects. The data thus summarized allow objects to be identified, at least in part, through the class to which they belong. Specifying the boundaries describing a class has several advantages. One is that limits to generalization can be established, and another is that it becomes possible to generate predictions about how different classes are composed and how class membership relates to other variables.

Some Examples of Classifications

The most well-known examples of classifications are from the natural sciences, rather than the social sciences. A well-known, still used, and expanding classification is Mendelejew's *Table of Elements*. It can be viewed as a prototype of all taxonomies in that it satisfies the following evaluative criteria: (1) Theoretical foundation: A theory determines the classes and their order. (2) Objectivity: The elements can be observed and classified by anybody familiar with the table of elements. (3) Completeness: All elements find a unique place in the system, and the system implies a list of all possible elements. (4) Simplicity: Only a small amount of information is used to establish the system and identify an object. (5) Predictions: The values of variables not used for classification can be predicted (number of electrons and atomic weight), as well as the existence of relations and of objects hitherto unobserved. Thus, the validity of the classification system itself becomes testable.

Another successful classification system is biological taxonomy. Indeed, most attempts to formalize classification have some intellectual roots in this tradition (Sokal and Sneath, 1963). The result of such classification is frequently depicted as a 'phylogenetic tree,' today often the result of comparative genomics. In biological taxonomy, however, theory is not so strong as to warrant completeness, as in the *Table of Elements* (e.g., how should one deal with archaebacteria?). Moreover, the identification of a specimen requires information from morphology and sometimes from behavioral observation. In addition, the system abounds with nested criteria. Further, compared with physics, predictions of future developments, or of 'missing links' in biological taxonomy are vague. However, the classes of the phylogenetic system are still useful because, at the very least, they indicate boundaries to generalization.

In the behavioral and social sciences, hundreds of classifications are published every year. Noteworthy examples are Bloom's taxonomy of educational objectives (Krathwohl et al., 1964), as well as the *DSM* (*Diagnostic and Statistical Manual of Mental Disorders*) and *ICD* (*International Classification of Diseases*) classification systems used in psychology and psychiatry. None of these systems have been formally derived, however. Instead, they were generated based on 'experience.' The resulting classes are so heterogeneous that they acknowledge many exceptions. Also, a phenomenon called 'comorbidity' shows that these classification systems are not optimal yet. It refers to the simultaneous existence of two or more disturbances in the same patient. If comorbidity is the rule rather than the exception, then the classification system loses plausibility and practicability.

Preparing the Basis of a Classification

Selecting the Cases

In the beginning of the process of developing a classification, two main questions arise: (1) Which elements are to be differentiated in a classification? One searches for a list (complete, if possible) of cases to be classified. This list is called the 'extension' of a classification. (2) Which properties characterize the cases? The list of these properties is called the 'intension' of the classification. The answers to these questions already determine, in part, the results of the classification. To quote Hartigan (1982: p. 2): "Clearly, the selection of variables to measure will determine the final classification. Some informal classification is necessary before data collection: deciding what to call an object, deciding how to classify measurements from different objects as being of the same variable, deciding that a variable on different objects has the same value." Sometimes, no well-defined population of objects is available from which to sample, and a preliminary selection has to be made intuitively. In such cases, future applications of the classification system may result in more or different classes from those originally obtained.

Specifying the Properties

The question of feature selection (selection of the properties on which the classification will be based) arises at two points in the process of classifying. First, as mentioned above, it arises at the very beginning of the process. The second opportunity to select properties comes when an established procedure is tested for identification. This problem is very similar to one in regression analysis: Which variables should be retained because they discriminate best, between the classes (see Pankhurst, 1991). Even if computational problems do not play a role, use of too many properties can still be problematic, particularly if measuring these variables is expensive or dangerous. In both instances of selecting properties, reliability is a very important issue. With decreasing reliability of the measurement of the properties, the identification of classes becomes more difficult.

Another important question is whether the values of the properties should be transformed before searching for classes.

The results of most classification procedures will be influenced by transformations. If differences in variability between the variables are of substantive importance, no transformations that equate variability across variables should be used. The use of transformations is also called 'a priori weighting.' '*A posteriori* weighting,' on the other hand, refers to cases in which different variables are given different emphasis in the identification process.

Especially in routine applications, a good strategy for selecting properties to be retained in the final classification might be to find a minimal set of variables sufficient to discriminate between all the classes. Relative to the set of all variables, the minimal set may not be unique. If this is the case, one will often prefer a set with few practical problems, and replace properties and take other aspects, such as minimization of costs, into account (see Pankhurst, 1991 for an algorithm).

Determining the Similarity

After the objects to be classified and their relevant properties have been selected, the similarity between objects is determined. Similarity is a key concept in classification. As was mentioned earlier, there are two basic ways to obtain similarity measures: The researcher can either collect similarity judgments from participants, or derive similarity measures from the empirical co-occurrence of properties. Methods for obtaining similarity judgments in the context of the first approach, 'similarity in the eyes of the participants,' are discussed in the context of data theory (Coombs, 1964). 'Proximity' measures can also be derived from confusion or generalization data, association probabilities, substitutability ratings, sorting procedures, and so on. The second case, 'similarity in the mind of researchers,' amounts to comparing the feature patterns of objects and describing the similarity between objects using *similarity coefficients*. A large, and still growing, number of these coefficients exist, and monographs on classification devote considerable space to them. Therefore, the choice of one particular coefficient should be explicitly justified. An analysis of the metric properties of coefficients is given by Gower and Legendre (1986). To choose a coefficient, one may refer to their axiomatic foundation (Baulieau, 1999).

Another important distinction in the selection of similarity coefficients refers to 'negative matching' i.e., deciding whether to include observations stating that two objects agree that a property is absent rather than present; that is, whether similarity between two objects in the absence, rather than the presence, of a property should be included in the similarity measure. Jaccard's (1908) coefficient excludes negative matchings.

Although most classification methods make use of similarity information, clustering models exist that do not refer to similarity. Another aspect that might be taken into account is the concept of similarity used: Why only use pairwise co-occurrences, and not higher-order contingencies (Daws, 1996)?

Establishing a Classification

After a measure of similarity has been selected, the next step is the actual classification of the objects based on the similarities between them. Formally (e.g., Biggs, 1999), classes can be thought as equivalent to (1) partitioning a set into subsets, (2) classifying a set of objects, and (3) distributing a set of objects into a set of 'boxes.' These various perspectives differ markedly in their implications for classification. For example, in most mathematical conceptualizations, an element is classified into exactly one class. Some clustering procedures, however, allow for residual elements, which are not considered clusterable.

Depending on the approach to classification that a researcher has chosen, certain considerations and precautions are necessary. For example, similarity judgment data may not fulfill some necessary assumptions; generally, related objects are to be located in the same class. The relationship xRx when 'x is related to x' has the properties of reflexivity: xRx, symmetry: $xRy \rightarrow\!\!\!\rightarrow yRx$, and transitivity: xRy and $yRz \rightarrow\!\!\!\rightarrow xRz$. These are the properties of an equivalence relation. If the empirical data are similarity judgments, they do not necessarily fulfill this relation. Some properties of this relation can be tested statistically. Another important consideration applies if the objects are classified by using rules referring to features. In such cases, these rules need to be free of contradictions (for a test, see Feger, 1994).

Only a few substantive theories in the behavioral and social sciences allow one to deduce the number and kind of classes needed to describe a given range of phenomena. Therefore, in many cases inductive procedures have to be used to generate classes. For this, one needs a concept of what constitutes a class. Many researchers apply inductive classification methods without ever considering explicitly the class concept that their method implies. The following part of the article gives a brief discussion of the class concepts implied in frequently used methods for finding classes. The list is not complete, and the 'cluster analysis proper' dominates all other approaches, because of the frequency of its use.

Before discussing class concepts in detail, one more general distinction needs to be made. If classes are defined by properties of objects, two levels of definition can be distinguished. A general definition specifies the relationship between the properties and the classes. Specific definitions provide detailed translations of the general definition into formal operations for assigning the objects to the classes. Obviously, there can be many different specific definitions. General definitions can be ordered by the kind and amount of variability they allow among objects within the class. There are two general positions with respect to within-class variability. The 'monothetic' position (Sutcliffe, 1993) assumes that a class is defined by one or a few necessary properties. The 'polythetic' counter-position (Gyllenberg and Koski, 1996) assumes that some properties of a specified total set, not necessarily the same for every object, are sufficient. According to this position, a property is shared by most, but not necessarily all objects of a given class. Proponents of the monothetic camp tend to stress that some properties are more important than others, and that these properties should be used to establish the classification. The opposite position assumes equal importance of all properties. As a third type of general definition, one may add definitions referring to a 'prototype,' that is, the most typical example of a class or a hypothetical mean object. In this last case, 'closeness' or similarity decides about class membership, and the prototype may be defined with or without allowing for variation in its properties.

Given properties as the base for a classification, the actual observations often are represented as a data matrix containing, for example, the objects as the columns and their properties as the rows. The cells of the matrix contain either the values '0' or '1' to indicate the absence or presence of properties, or they contain frequencies, durations, intensities, or symbols (in the case of qualitative polytomous items) that indicate the type or degree of the respective property in the respective object. As this enumeration shows, the procedure can accommodate data of all scale types. The goal now is to find a 'feature by classes' matrix, called – corresponding to its purpose – the reference or identification matrix, or simply 'a classification.'

Concepts of Classes

Cluster analysis proper. When authors (e.g., Everitt, 1993) illustrate the concept of a cluster, they often use two-dimensional graphs to show the clusters as clouds of points (representing the objects). The clouds can have various forms; generally there are 'gaps' between the clusters that contain no data points, so that the clusters are isolated from one another. While such explanations of the cluster concept seem intriguing as they invoke classical 'gestalt' concepts, it is important to remember that the properties of (good) figures are defined by several 'laws,' not just one or two axioms or rules, as in cluster concepts.

Helpful as visualizations are, the more general definition of a cluster does not refer to any particular conception of space, be it dimensional, metric, or Euclidean. Set theory defines a cluster as the maximal subset of elements for which proximities within this subset are larger than between any elements of the subset and elements not contained within it. As was discussed above, proximities are information about the extent to which objects 'belong together,' and could be expressed in many different ways, for example, as similarities, distances, ranks, or binary information about set membership. More than one subset may exist; subsets may be disjointed or overlapping; and they may or may not be hierarchically ordered. Given this very broad conceptualization, social scientists have access to a large number of clustering procedures. The large number of options reveals that no 'one and only' definition of a cluster can be found. Presumably, the availability of so many approaches is one reason for the paucity of comparative studies on methods of clustering.

Clustering procedures can be classified as 'leading to a structure that is either hierarchical or non-hierarchical.' The most frequently applied classification procedures are hierarchical, disjointed, and provide exactly one class for each object. The best known hierarchical procedures are agglomerative; that is, in a series of partitions they successively and with increasing dissimilarity, fuse the objects into classes. Each step provides a set of classes, from which the researcher has to make his choice. Once a fusion is made, it is irrevocable, so the early fusions should be very reliable. *Additive clustering* (Shepard and Arabie, 1979) is a hierarchical partitioning allowing membership of objects in any number of classes. Here, the classes might be interpreted as properties (Lee, 1999).

The cluster concept treated thus far is based on similarity as formally represented either in a space or by set theory. A close relative is prototype theory, popular in cognitive research. A prototype can be defined as a vector of values of selected properties; usually a list of cases as exemplars of this prototype is also available. One fundamental assumption of the prototype-oriented approach can be formulated as follows: If there is high similarity among a set of patterns, these patterns are also similar to an observed or inferred prototypical pattern. An inferred pattern could, for example, be the vector of mean values. This pattern has high or maximal similarity to every other pattern. The idea of inferring the prototypical pattern from the data forms a bridge to the similarity-based conception. But the researcher has to be more active in abstracting and defining a specific instance as the prototype.

Contingencies of higher order than similarities between the properties are exploited in some other generalizations of the concept of similarity-based clustering, such as Configural Frequency Analysis (Krauth and Lienert, 1973) and Pattern-Analytic Clustering (McQuitty, 1987). For example, Configural Frequency Analysis identifies combinations of properties that occur more often than expected from some specified base model.

A recent trend, increasing in strength, is to use mixture models for clustering. The original purpose of these methods was to base classification on a model that allows for inference-statistical treatment. But they have since found wider purposes. The basic idea of mixture models can be illustrated using the following example: Assume that a sample of measurements of body height is drawn from a human population. While it is known that all the cases are male or female, gender is not recorded for individual respondents. It is, however, possible, based on the distribution of heights in the total sample, to estimate the coefficients of the separate height distributions for men and women. This is done by interpreting each measurement as a sum of weighted height measurements for women and for men. These weights are the probabilities for each measurement to be from a man and from a woman. "Thus the density function of height has been expressed as a superposition of two conditional density functions; it is known as a finite mixture density." (Everitt, 1993: p. 110).

Mixture models are based on a 'space' concept rather than a 'similarity' concept; clusters are regions of relative point densities in this space. The assumptions for mixture models are comparable with those of the general linear model: cardinal scale level and multivariate normal (or similar) distributions of the data. A comparatively common mixture model for categorical data is Latent Class Analysis (De Soete, 1993).

To conclude this classification of class concepts, one further concept needs to be mentioned. This conception, models for block structure, is close to the raw data matrix and the Aristotelian tradition. A block is a maximal rectangular submatrix combining some objects and some properties with the same (or similar) values in the cells of the data matrix. The scale level of the values is not fixed, and the similarity concept is not invoked in the analytical procedure. In a block, the set of partially similar objects corresponds to the extension of a concept or class. The set of partially similar properties corresponds to the intension. The symmetry in the definitions of intension and extension is fully exploited and preserved (see Feger and De Boeck, 1993).

Evaluation of a Clustering Result

Although model evaluation is only a part of the overall evaluation of a classification (see Section Evaluating a Classification), it is an important one. As Dunn and Everitt (1982: p. 94) state: "Since clustering techniques will generate a set of clusters even when applied to random, unclustered data, the question of validating and evaluating becomes of great importance." Jain and Dubes (1988) classify the criteria of validation as follows:

> External criteria measure performance by matching a clustering structure to a priori information …. Internal criteria assess the fit between the structure and the data, using only the data themselves … Relative criteria decide which of two structures is better in some sense, such as being more stable or appropriate for the data.

Considerable progress has been made in internal statistical cluster evaluation. Statistical procedures exist for testing the existence of 'natural' clusters, for testing the adequacy of computed classifications, and for the determination of a suitable number of clusters (see, e.g., Bock, 1996). A very plausible way to evaluate any solution, independent of the clustering approach used, is to reproduce, or 'derive,' from the solution, all information that the solution gives about raw data that would fit with the solution, and then to compare this information with the actual raw data.

Procedures to Assign Cases to Classes

Procedures to assign single cases to classes are needed for two purposes. One purpose is to assign newly observed cases to the classes of an already existing classification. The other purpose is to evaluate a classification by taking 'old' cases from the original sample on which the classification was based, and checking which class they would be assigned to. In both cases, the question is: Into which class should the case be placed? In practice, experts (e.g., physicians) are often consulted for the answer of this question. In other cases, numerical procedures ('automatic classification') are used. Here, the properties may be used, either sequentially, as in a diagnostic key, or simultaneously by some type of matching between the case and the existing classes (see Dunn and Everitt, 1982; especially on diagnostic keys). Quite often, as identification with certainty is impossible "either because too many characters are variable within taxa or because all assessments of character states are subject to error, probabilistic identification methods are often used" (Dunn and Everitt, 1982: p. 112). Of the probabilistic procedures, the Bayes approach (*see* Bayesian Statistics) and discriminatory analysis (*see* Multivariate Analysis: Classification and Discrimination) are especially well known.

Other placement rules can be used if they are transparent, unambiguous, and do not lead to contradictions. For example, the principle of 'nearest neighbor' computes the distances of a new pattern to all existing classes. It assigns the new case to the class to which the distance is shortest (for details and other rules, see Looney, 1997). Rules may also include options such as rejecting a case as 'not classifiable' or postponing a decision until more information is available. Most rules currently applied are compensatory, but rules could also be disjunctive or conjunctive, requiring at least one value to reach a high amount, or all values to surpass a given minimum (Coombs, 1964).

Different rules lead to different results, especially if the classes vary in their a priori probability, if the distributions and covariances of the variables are very different, and if the number of observations is small. The single most important criterion for evaluating an assignment procedure is the number of correct classifications. But this 'apparent error rate' is optimistically biased, because it does not take into account the probability of correct assignments by chance. If base rates of class membership are known, the predictions have to perform better than the base rate (Pires and Branco, 1997).

Evaluating a Classification

While every step of the classification process can be evaluated (Milligan, 1996), two stages have received special attention: the evaluation of class definitions and of the identification procedure. Both have been mentioned previously. There also exist procedures to evaluate the overall performance of a classification system. The main method is 'cross validation,' using a new sample of data comparable to the old one, or splitting the original sample randomly into two halves and using one half to evaluate the classification obtained in the other half.

Another way of evaluating the results of a classification process is by comparing the results of using different classification procedures. Usually the researcher has several choices in the classification process, and is not forced by theory to select just one option. Examples include multiple options about the selection of cases and variables, of similarity coefficients, of clustering models, and of identification rules. With computers, it is easy to try several combinations of such choices. Confidence that the classification captures substantial information in the data grows with the amount of agreement in the results from different combinations of choice options (for the evaluation and comparison of solutions, see Everitt, 1993). To aid the interpretation of the resulting class structure, Milligan (1996: p. 346) suggests deliberately adding 'ideal types,' that is, characteristic patterns constructed by the researcher, to the data, and to assess what clusters these patterns are assigned to.

Can a classification be wrong? In most cases, a classification is just a systematic description, and as such, may, or may not, be useful. New observations may require changes in the classification. But if there exists a theory about the definitions of the classes, and the theory is strong enough to allow for specific predictions, then these predictions can be falsified and/or lead to revisions of the classification system.

Conclusions

More in the past than in the present, opinions fundamentally critical of the possibility of classification in the social sciences have been expressed. For example, Galt and Smith (1976: p. 58) stated: "Because they usually lack measurable dimensions, social entities are difficult to classify, and any given system of classification will inevitably be arbitrary." Indeed, numerical classification definitely requires measurement, or more generally, the interpretation of observations as variables.

Indeed, every variable is itself a classification, defined as a set of disjointed, exclusive and together sufficient classes, and the 'categories' of variables are referred to as their values. Without variables in this formal sense, one might consider a heuristic equivalent in abstraction from, and ordering of, observations: the 'ideal type,' as introduced by Max Weber.

Variables used to establish a classification may be discrete or continuous. For a classification to be justified, the frequency or density distributions of the properties should not be equal distributions, but show one or several peaks. Then, when the joint distributions of more than one variable are considered, some combinations of values may be more frequent than other combinations, perhaps even more frequent than would be expected based on the marginal distributions. This is one of the fundamental phenomena enabling the formal definition of classes.

While classification is a 'process,' the temporary result is a 'structure.' Dynamic aspects, such as the development of a property and other trends and changes, might be included in the definitions of variables. This does not, however, make a formal classification a process model. In this sense, classification is static. It temporarily fixes the – sometimes turbulent – streams of information. Changing a classification means the (re-)interpretation of some substantive area.

See also: Measurement Theory: Conjoint; Mixture Models in Statistics; Multivariate Analysis: Classification and Discrimination; Statistical Clustering.

Bibliography

Arabie, P., Hubert, L.-J., De Soete, G. (Eds.), 1996. Clustering and Classification. World Scientific, Singapore.

Baulieau, F.B., 1999. Two variant axiom systems for presence/absence based dissimilarity coefficients. Journal of Classification 14, 159–170.

Biggs, N.L., 1999. Discrete Mathematics, rev. ed. Clarendon Press, Oxford, UK.

Bock, H.-H., 1996. Probability models and hypothesis testing in partitioning cluster analysis. In: Arabie, P., Hubert, L.J., De Soete, G. (Eds.), Clustering and Classification. World Scientific, Singapore, pp. 377–453.

Coombs, C.H., 1964. A Theory of Data. Wiley, New York.

Daws, J.T., 1996. The analysis of free-sorting data: beyond pairwise cooccurrences. Journal of Classification 13, 57–80.

De Soete, G., 1993. Using latent class analysis in categorization research. In: van Mechelen, I., Hampton, J., Michalski, R.S., Theuns, P. (Eds.), Categories and Concepts. Academic Press, London, pp. 309–330.

Dunn, G., Everitt, B.S., 1982. An Introduction to Mathematical Taxonomy. Cambridge University Press, Cambridge, UK.

Everitt, B.S., 1993. Cluster Analysis, third ed. Edward Arnold, London.

Feger, H., 1994. Structure Analysis of Co-occurrence Data. Shaker, Aachen, Germany.

Feger, H., De Boeck, P., 1993. Categories and concepts: introduction to data analysis. In: van Mechelen, I., Hampton, J., Michalski, R.S., Theuns, P. (Eds.), Categories and Concepts. Academic Press, London, pp. 203–223.

Galt, A.H., Smith, L.J., 1976. Models and the Study of Social Change. Wiley, New York.

Gower, J.C., Legendre, P., 1986. Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification 3, 5–48.

Gyllenberg, M., Koski, T., 1996. Numerical taxonomy and the principle of maximum entropy. Journal of Classification 13, 213–229.

Hartigan, J.A., 1982. Classification. In: Kotz, S., Johnson, N.L., Read, C.B. (Eds.), In Encyclopedia of Statistical Sciences, vol. 2. Wiley, New York, pp. 1–10.

Jaccard, P., 1908. Nouvelles recherches sur la distribution florale. Bulletin de la Socit Vaudoise de Science Naturelle 44, 223–270.

Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, NJ.

Krathwohl, D.R., Bloom, B.S., Masia, B.B., 1964. Taxonomy of Educational Objectives. Longman, London.

Krauth, J., Lienert, G.A., 1973. KFA—Die Konfigurationsfrequenz-analyse. Alber, Freiburg, Germany.

Lee, M.D., 1999. An extraction and regularization approach to additive clustering. Journal of Classification 16, 255–281.

Looney, C.G., 1997. Pattern Recognition Using Neural Networks. Oxford University Press, New York.

McQuitty, L.L., 1987. Pattern-Analytic Clustering: Theory, Method, Research and Configural Findings. University Press of America, New York.

Milligan, G.W., 1996. Clustering validation: results and implication for applied analyses. In: Arabie, P., Hubert, L.J., De Soete, G. (Eds.), Clustering and Classification. World Scientific, Singapore, pp. 341–375.

Pankhurst, R.J., 1991. Practical Taxonomic Computing. Cambridge University Press, Cambridge, UK.

Pires, A.M., Branco, J.A., 1997. Comparison of multinomial classification rules. Journal of Classification 14, 137–145.

Shepard, R.N., Arabie, P., 1979. Additive clustering representations of similarities as combinations of discrete overlapping properties. Psychological Review 86, 87–123.

Sokal, R.R., Sneath, P.H.A., 1963. Principles of Numerical Taxonomy. Freeman, San Francisco.

Sutcliffe, J.P., 1993. Concept, class, and category in the tradition of Aristotle. In: van Mechelen, I., Hampton, J., Michalski, R.S., Theuns, P. (Eds.), Categories and Concepts. Academic Press, London, pp. 35–65.

Classifiers, Linguistics of

Colette Grinevald, University of Lyon, Lyon, France

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C. Grinevald, volume 3, pp. 1973–1978, © 2001, Elsevier Ltd.

Abstract

The study of classifiers was concerned at first by classification systems in the nominal domain (with systems labelled numeral, noun, genitival and verbal classifiers according to their morphosyntacticloci), and only later with classifiers in the verbal domain (with verb classifiers). Beyond the spoken modality, classifiers exist in sign languages and Egyptian hieroglyphs. To account for the great variety of individual systems, a multidimensional approach is adopted. This approach includes an inventory of forms and a semantic profile of the system, the classificatory scope of individual classifiers, and the system's origin and stage of evolution. Pending are issues of lumping versus nonlumping of classifier and other types of systems, and a terminology to handle the great variety of systems.

In the literature concerned with linguistic categorization systems, the term 'classifiers' is actually used in two ways, either for a great variety of systems or for a much narrower range of such systems. The wide-scope approach, called here the lumping position, finds its justification in the multiple ways in which many linguistic categorization systems share some similarities of either forms or semantics. The narrower approach taken here, or nonlumping position, means to isolate a particular set of those systems and to only use the term 'classifiers' for them. This set is characterized by its nonconcordial morphosyntactic nature, its not being involved in lexicogenesis (or creation of new words in the lexicon), its being of clear lexical origin and persistent semantic motivation, and its being subject to specific discourse-pragmatic conditions of use. In this narrower sense of the term 'classifiers,' there are no classifier systems in Indo-European languages and the most widely known exemplars of such systems are most likely the numeral classifier systems (see Gil, 2013), found in Asian languages like Chinese or Japanese (pp. 279–303), but also in America (see early study by Berlin, 1965). They are illustrated in **Table 1**.

History of Classifier Studies: From the Classics on …

Classifiers became of interest to general linguists in the 1970s, following proposals to capture their universal semantic properties. Three articles stand out as classics of the field. Adams and Conklin (1973) were the first to wade through much comparative data and to claim the existence of some universal semantic properties of systems of classification. They used primarily data from Asian numeral classifier systems and established the primacy of three basic shapes, semantic combinations of one of the major dimensional outlines of objects (1D, 2D, 3D) with a secondary characteristic of consistency and/or size, the primary shapes being 1D rigid, 2D flexible, and 3D full. They showed how these combinations of features were directly inherited from the most common lexical sources for the most basic set of classifiers, designating the primary elements of the physical world being handled for the survival of human communities, as shown in **Table 2**.

Denny (1976) was the work of a psychologist whose main contribution was the appealing proposal that what classifiers are 'good for' (the title of his paper) is to signal how humans interact with the world, as indicated by the organization of the semantic traits of classifiers into three kinds, those of 'social, physical, and functional interaction.' The 'social interaction' domain is that of animate entities of our world, principally fellow human beings, classified by sex, social rank, or other categorization schemata, as well as other entities such as divinities and other powers specific to a culture. The 'physical interaction' domain is the one of manipulatable and manipulated objects of the world classified along certain parameters linked to their nature, principally of shape. Finally, the 'functional interaction' domain is that of entities of the world being classified by the use to which they are put, such as food, clothing, or transportation, for instance.

Allan (1977) was the first typological study based on a broad database of 50 classifier languages covering different types of nominal classification systems. Although the reliability of the data was variable, there was still a remarkable overlap between the seven 'categories of classification' (material, shape, consistency, size, location, arrangement, and quanta) he proposed and Denny's (1976) three functional domains. Furthermore, two of Allan's original statements were of particular interest

Table 1 Examples of numeral classifier systems

	Japanese		*Tzotzil*
a.	enpitsu ni-hon pencil 2-CL(1D) 'Two pencils'	a.	j-ch'ix kantela 1-CL(1D) candle 'One candle'
b.	ringo ni-ko pomme 2-CL(3D) 'Two apples'	b.	j-p'ej alaxa 1-CL(3D) orange 'One orange'

Table 2 Lexical sources of basic classifiers

Lexical origin	*Classifiers*
Tree/trunk	1D: long-rigid
Leaf	2D: flat-flexible
Fruit	3D: round

for later discussions on the nature and purpose of such categorization systems. They were the note of a total absence of color classifiers, and of a constraint that the characteristics picked up in the classification system be perceivable by more than one of the senses alone, such as sight and touch, where sight means primarily perception of shape.

Several collective publications later expanded the field of nominal classification studies in general from various disciplines. Seiler and Lehmann (1982), Seiler and Stachowiack (1982) and Seiler (1986) report on the major typological project known as UNITYP, which studied the various techniques that languages have to capture the function of apprehending objects, including numeral classifiers and noun classes. The collection in Craig (1986a) covers various approaches from different fields, and includes articles on the acquisition, historical development, discourse function, semantics, and cognitive value of these systems. The collection in Senft (2000, actually based on a 1993 working conference) is a further elaboration of issues of typology, grammaticalization, and function of nominal classification systems.

Typology of Classifier Systems

While the first studies of so-called classifiers had a general tendency to overlook the existence of multiple types of classification systems, at some point linguists' attention turned to the typological characterization of classifier systems per se. The specificity of these particular systems was then considered at two levels: at one level the issue is to distinguish them from other types of classification systems, and at the other it is to recognize several subtypes among them.

Classifiers as One Type of Nominal Classification System

As argued early in Craig (1987) and worked out in more detail in Grinevald (2000), classifier systems can be identified as specific types of linguistic systems both on lexical and morphosyntactic grounds, although it is true that in some languages the different systems commonly share lexical sources and seem to be intertwined, as will be considered below.

In the lexical domain, classifiers may be distinguished from *measure terms and class terms* with which they are often either confused or consciously lumped.

All languages have lexical sets of measure terms, the expression 'measure terms' being used here as a cover term for what are strictly speaking measures and types of arrangements. Examples of English measure terms include actual measure terms such as a glass of water, a pound of sugar, a slice of bread, a sheet of paper, and arrangements such as a pile of books, a group of children, a line of cars. Class terms are sets of lexical items used in lexicogenesis; they participate in compounding processes of word formation that are functionally equivalent to derivational processes. English has class terms like '-berry' (as in strawberry, blueberry, boysenberry, gooseberry, loganberry), '-man' (as in mailman, policeman, garbageman), or '-tree' (as in apple tree, banana tree, cherry tree), for which the functional equivalent in French can be derivational suffixes, such as '-ier' (as in pommier 'apple tree,' bananier 'banana tree,' cerisier 'cherry tree').

Table 3 Different systems of overt linguistic categorization

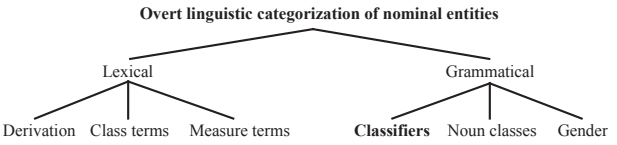

In the morphosyntactic domain, classifiers can be distinguished from concordial systems such as *gender* best known from Indo-European languages and *noun class systems* of Bantu languages. Classifiers therefore are one among several systems of overt linguistic categorization of nominals as shown in Table 3.

Various Subtypes of Classifier Systems

Beyond distinguishing classifier systems from other systems of nominal classification, it is further necessary to acknowledge the existence of several subsystems of classifiers.

Craig (1987) and Grinevald (2000) proposed a typology based on the morphosyntactic locus of the classifier, on grounds of better reliability of this criterion over one of semantics. It was clearly conceived as a guide to fieldwork on yet undocumented systems of oral tradition languages. The proposal presented here stands therefore in contrast to studies where all systems are mixed in a lumping position, as was the case in Allan (1977) but also still in Croft (1994), for instance, in which discussions on universal semantics miss the specificities of the distinct systems in their more prototypical forms.

In the study of classifiers, which was first limited to the domain of nominal classification, i.e., the categorization of nouns, the numeral classifiers (numeral + CL) used in quantifying expressions were the best known and those considered as the prototype of classifiers (and are still considered that way by many). But three other subtypes of classifiers were later identified. From languages of Oceania linguists learned of genitival or possessive classifiers (poss + CL) as constituents of possessive constructions, and from languages of Mesoamerica and Australia came the identification of noun classifiers (CL noun), so-called for appearing with a bare noun and not linked to the expression of quantification or possession, although sometimes labeled generic classifiers because of their semantics. Finally, North American languages exhibited yet another type of classifier, labeled verbal classifiers (verb-CL) for being morphemes referring to nominal arguments inside the verb form, often linked to linguistic processes of incorporation.

These four major subtypes of nominal classification systems identified in the languages of the world are schematized in Table 4.

Table 5 gives sentence examples to illustrate these four major subtypes of classifier systems.

To the major noun/numeral/genitival/verbal classifiers shown in this diagram, one should add two much rarer secondary types that have been identified more recently, mostly in Amazonian languages, and have been found on demonstratives and locative adpositions (Aikhenvald, 2000).

Table 4 Loci of major subsystems of nominal classifiers

$_{NP}[$POSS+**CL**$_{genitival}$	Numeral+**CL**$_{numeral}$	**CL**$_{noun}$+Noun$]_{NP}$	//	Verb-**CL**$_{verbal}$

Table 5 Examples of major subtypes of nominal classifier systems

1.	**Noun classifiers**; Jakaltek-Popti' (Craig, 1986, p. 264)	a.	xil **naj** xuwan **no7** lab'a saw CL John CL snake '(man)John saw the (animal)snake'
2.	**Numeral classifiers**; Ponapean (Rehg, 1981, p. 130)	a.	pwihk riemen pig 2+CL: animate 'two(-animate) pigs'
		b.	tuhke rioapwoat tree 2+CL: long 'two(-long) trees'
3.	**Genitive classifiers**; Ponapean (Rehg, 1981, p. 184)	a.	kene-i mwenge CL-GEN.1 food 'my(-edible) food'
		b.	were-i pwoht CL-GEN.1 boat 'my(-transport) boat'
4.	**Verbal classifiers**; Cayuga (Mithun, 1986, pp. 386–388)	a.	ohon'atatke: ak-hon'at-a:k it-potato-rotten past.I-CL-eat 'I (potato-)ate a rotten potato'
		b.	skitu ake'-treh-tae' skidoo I-CL: vehicle-have 'I (vehicle-)have a car'

Arguments for Subtypes

The existence of different subtypes of classifiers was argued on several grounds (as in Grinevald, 2000). First is the characteristic of different loci for the classifier, as mentioned above in Table 4, with a specific inventory for each locus. Second is the fact that several independent systems can co-occur in the same language with different inventories and different semantics in different morphosyntactic loci. Such is the case of the coexistence of numeral and genitival (or possessive) classifier systems of Micronesian languages like Ponapean, as exemplified above in Table 5. The third argument is semantic, and relies on a strong correlation between the major morphosyntactic subtypes of classifiers presented and their semantic profiles, i.e., the semantic basis of their categorizing principle. It is not a question of all or nothing, since, as discussed below, the dynamics of such systems allow for much variation and overlap, but of strong tendencies. Concentrating on the categorization of nonhuman entities (and borrowing from Denny (1976) the notions of physical and functional properties) the following pattern can be seen: physical properties such as shape seem to be the dominant semantic parameter of numeral classifier systems, functional properties such as food and transportation the major one of genitival classifier systems, while material and essence of objects constitute the major concern of noun classifier systems, as schematized in Table 6.

Two notes of clarification about these correlations between semantic profiles and subtypes of classifiers are needed. One is that the numeral classifier systems have two kinds of elements within the same inventory and in the same slot, known as mensural (two-truck loads of/lines of trees) and sortal (two-long trees) elements, and that only the sortal ones are considered here. The other is that no particular correlation seems to hold between verbal classifier subtype and semantic profile (as illustrated in the Cayuga examples of Table 5).

While the first studies of the phenomenon of classifiers were restricted to the domain of nominal classification, other systems were identified that classified verbs through verb semantics, first on the basis of languages of Australia (Schultze-Berndt, 2000; McGregor, 2002), but also in parts of

Table 6 Different semantic profiles of subtypes of nominal classifiers

Subtype	*Semantic profile*	*Examples*
Numeral classifiers	Physical categories	One-LONG RIGID canoe; Two-ROUND oranges; Three-FLAT FLEXIBLE blankets
Genitive classifiers	Functional categories	My-VEHICLE canoe; Your-EDIBLE fish; His-DRINKABLE potion
Noun classifiers	Material/essence categories	(a/the) WOOD canoe; ANIMAL (woollen) blanket; PLANT coffee (drink)

Table 7 Verbal versus verb classifiers

Verbal CL (classifier of patient object)	*Verb CL (classifier of verb semantics)*	
a. I [LONG RIGID]-put a knife on the table	c. You shout-[SAY] to me	(Intransitive)
b. I [VEHICLE]-have a canoe	d. Put-[DO] a knife on the table!	(Transitive)

Latin America (Dickinson, 2000). The contrast between verbal (locus) and verb (domain and locus) classifiers is outlined in Table 7 where in (a) the classifier attached to the verb refers to the characteristic shape of the knife, while in (b) the classifier categorizes the semantics of the verb itself.

Schultze-Berndt labels these verb classifiers generic verbs. They constitute a closed class that take verbal inflections and are accompanied by so-called 'coverbs' corresponding to the verb class of European languages but including also adverbs. Table 8 shows three such verb classifiers.

The inventory of these verb classifiers (as in Schultze-Berndt, 2000, pp. 402–404) contains 26 elements, constituting the categorization of events given in Table 9.

Table 9 Inventory of classifier VERBS in Jamingjung

	Semantics of categorization	*CL of intransitive verbs*	*CL of transitive verbs*
a.	Location, existence, possession, and change of locative relation	BE, FALL	HAVE, PUT
b.	Translational motion	GO, COME	TAKE, BRING, LEAVE, APPROACH, FOLLOW
c.	Contact/force	–	HANDLE, HIT, CHOP, KICK/STEP, POKE, BITE, THROW
d.	Burning, cooking	BURN	COOK
e.	Polyfunctional	–	SAY/DO
f.	Caused change of possession	–	GIVE, TAKE AWAY
g.	Other major verbs	–	SEE, EAT, MAKE

Schultze-Berndt, E., 2000. Simple and complex verbs in Jamingjung: a study in event categorization in an Australian language, Ph.D. thesis, University of Nijmegen, pp. 402–404.

Classifiers in Other Modalities

Furthermore, the discussion of classifiers has been extended to languages of other modalities, such as visual modalities, dynamic for sign languages, and static for hieroglyphs, but both interesting for their iconic nature. Emmorey (2002) is a collection of articles on the phenomenon of classifiers found across the sign languages of the world, shows that their dominant type is of the form of verbal classifiers, discussed in sign language literature under the label of classifier constructions. It has also been argued that some sign languages, such as Swedish sign language, exhibit not only verbal but also noun classifiers (Berman and Wallin, 2003). The latest addition to the discussion of classifier studies comes from the writing system of some ancient languages, such as Ancient Egyptian and Sumerian. Goldwasser (2006) argues that the elements of the Egyptian hieroglyphic writing system known in the literature as 'determinatives' (Champollion, 1836) can be reanalyzed as classifiers. Goldwasser and Grinevald (2012) further specify that, on closer inspection, this particular system of classifiers can be shown to resemble known classifier systems of oral languages although it presents the unique characteristic, not described for other languages as yet, of combining within one system the classification of both nouns and verbs, in their respective morphosyntactic loci with their expected semantic profiles (categorization of humans, animals, and inanimates of the Egyptian culture vs categorization of types of events). Table 10 offers a sample of such coexisting noun and verb classifiers.

This system sometimes also applies to other word categories, such as adverbs and adpositions more rarely attested in classifier systems of the world (but found in Australian Jamingjung verb classifier system too). It would seem that what appears as a mixed noun and verb system is most likely a natural phenomenon of classification of a root system. Roots are a central concept of this language in which it may prevail over that of word categories (noun and verbs) of prime importance in Indo-European languages. The analysis of determinatives as classifiers is not yet accepted by all Egyptologists who do not consider it as part of the language, on the ground that the system is restricted to the hieroglyphic writing system and appears not to have been part of the spoken language. This is evidenced by the fact that when hieroglyphs were transcribed into Coptic writing these classifiers were ignored. But the system is so strikingly similar in its categorizing schemata to classifier systems of spoken languages that drawing the parallel argues interestingly for the cognitive power of such systems, with their mix of universal and language specific categorization processes, independent of the modality of expression.

Therefore, in the end, the variety of subsystems of classifiers is much greater than thought before, the classifier systems of

Table 8 Classification of verbs in Jamingjung (Schultze-Bernd, 2000, p. 404)

a. yeah,	**dalb**	guyug	yirr-**arra**-m=ngarndi	
yes	**light.fire**	fire	1pl.excl:3sg-**PUT**-PRS=Sfoc2	
'yes, we set fire to the firewood \'				
b. burra-**ngayi**-rna=yirrag	wirib-di	**jarl,**	malajagu	
3pl:3sg-**SEE**-IMPF=1pl.excl.OBL	dog-ERG	**track**	goanna	
'the dogs used to track them for us, the goannas'				
c. ga-**jga**-ny=ni	wagurra-bina	**burduj**		
3sg-**GO**-PST=SFOC1	rock-ALL	**go.up**		
'he went up on a rock'				

Table 10 Classifiers of Egyptian hieroglyphs

	Classifier	*Hieroglyph name & code*	*Examples*	
		of NOUNS		
1.	'HUMAN FEMALE' (CL:HF)	seated woman B1	**a.** *ḥwnwt*-CL:HF 'young girl'	**b.** *mnʿt*-CL:HF 'wet-nurse'
2.	'ANIMAL' (CL:ANIMAL)	hide and tail[1] F27	**a.** *miw*-CL:ANIMAL 'cat'	**b.** *pnw*-CL:ANIMAL 'mouse'
3.	'WOOD' (CL:WOOD)	branch M3	**a.** *ꜣṯwt*-CL:WOOD 'bed'	**b.** *hbny*-CL:WOOD 'ebony'
		of VERBS		
4.	'INGESTION/COGNITION/ PERCEPTION' (CL:ICP)	man with hand to mouth A2	**a.** *swr*-CL:ICP 'to drink'	**b.** *mrí*- CL:ICP 'to love'
5.	'POWER' (CL:PWR)	man striking with stick A24	**a.** *ḫti*-CL:PWR 'to carve/ to sculpture'	**b.** *sbꜣ*-CL:PWR 'to teach'
6.	'MOVEMENT' (CL:MVMT)	legs walking D54	*nʿi*-CL:MVMT 'to travel'	
7.	'MOVEMENT BY BOAT' (CL:MVMT/BOAT)	boat on water P1	*nʿi*-CL:MVMT/BOAT 'to travel (by boat)'	

After Goldwasser, O., 2002. Lovers, Prophets and Giraffes: Wor[l]d Classification in Ancient Egypt. Göttinger Orient Forschungen, Otto Harrassowitz, Wiesbaden, pp. 57–89.

the major lexical categories of N and V attested in the languages of the world today being as presented in Table 11.

A Multidimensional Approach to the Study of Linguistic Classification Systems

One of the major challenges of classifier studies is not only that there is a great variety of classification systems, but that there is also great variability within the different types of systems and subsystems. These varieties are taken to reflect the intermediate nature of these systems as secondary linguistic systems of clear lexical origins, some borrowed, some native, and all at different degrees of grammaticalization. The multiple dynamics at work in those systems make the descriptive task onerous, but at the same time very productive when proceeding to comparative and typological work. It is of prime importance, therefore, to consider a number of dynamic variables, from synchronic and diachronic points of view, for any description of a specific system along the line of topics listed below.

The Inventory of Forms and Its Semantic Profile

Systems of classifiers vary greatly as to the number and the specificity of the classes that compose a system. The inventory of classificatory items can be small or large, from a handful to hundreds, with some systems open. The inventory generally has a dominant semantic profile, corresponding to one of the three functional domains mentioned (social, physical, functional).

The Classificatory Scope of Individual Items

The classes defined can vary from very simple homogeneous with transparent semantic motivation to very complex, heterogeneous ones. The latter are organized around a core of prototype elements to which others have been added through various means of extension, such as with the famous class of 'women, fire, and dangerous things' of Dyirbal (interpretation from Lakoff, 1986, 1987; or original analysis from Dixon, 1982). Another well-known example is the case of the Japanese numeral classifier [hon], used prototypically for long, thin objects but in modern times classifying telephone calls and baseball hits as well (Matsumoto, 1993).

The items of an inventory are therefore of very different classificatory scope. The most common classifiers categorize at some generic level (e.g., animal/plant or 1D–2D–3D or food/transport), while others, fewer, at a more specific one (e.g., a corn and corn product class next to a generic plant class), with the specificity going to the extreme situation of so-called 'unique' classifiers, meaning classifiers heading a class of just one item (e.g., 'elephant' or 'tiger' or 'crocodile' in a class of its own, next to an animal classe). These specific and unique

Table 11 Diversity of subtypes of classifiers (and sample languages)

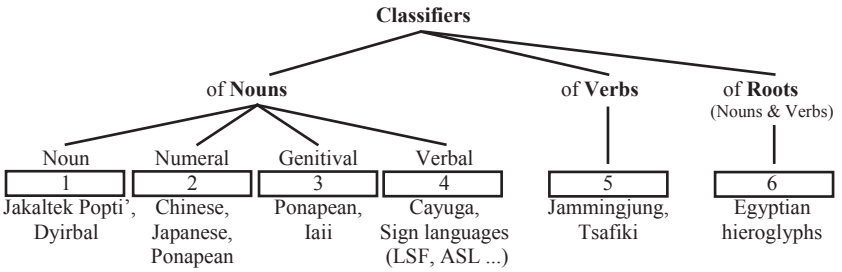

classificatory items are particularly interesting for their cultural significance (as demonstrated for Jakaltek Popti' in Craig, 1986b). At the other extreme, the inventory of some systems also includes a totally desemanticized classifier, or default classifier, for any number of items with no semantic links (meaning something like 'a thing'). This default item is often traceable through time to some specific noun turned classifier (as e.g., bamboo in Chinese and papyrus scroll in Egyptian hieroglyphs).

Origins: Internal Development or Borrowing of an Idea and Areal Spread

Classification systems follow various paths of evolution. Some emerge language internally, from already existing lexical classification systems like class terms, and recycled lexical items of the language. Any classification system needs to be assessed further in the context of a common phenomenon of areal spread. The spread can operate either through the actual borrowing of a system, morphology included, or through the borrowing of the idea and motivation for the development of such systems with native lexical items. A well-known case of areal spread is that of the expansion of the original Chinese numeral classifier system into its surrounding regions (Bisang, 1999); more recent literature treats the spread of noun class and classifiers in the Amazon region (Seifart and Payne, 2007; Aikhenvald, 2012).

Age, Life Cycles, and Degree of Grammaticalization

Systems studied today vary greatly in age. Some are very old, such as the Chinese numeral classifier system or the Egyptian hieroglyphic system, while others can be argued to be only several centuries old, like the Q'anjob'alan-Mayan noun classifiers, or even much less for some of the recent sign language verbal systems.

An early proposal (Dixon, 1982; repeated in Dixon, 1986; echoed in Grinevald, 2000) had suggested a correlation between type of classification system in general and degree of grammaticalization based on comparisons of the better known systems of the time, i.e., Bantu noun class systems and Asian numeral classifier systems. However, new data on other noun class systems and classifier systems per se have led to the proposition of disconnecting the question of the degree of grammaticalization from the type of system. Guarding against some Eurocentric posture and the problem of considering as prototypical the first systems described, Grinevald (2002) argues against this correlation, on the basis of a sample of classifier systems of different subtypes. They include incipient to well-established systems of noun classifiers from the Australian (Sands, 1995) and American continents, and more or less grammaticalized numeral classifiers of Asian languages as well as very grammaticalized ones of Chibchan languages of Central America. In a similar vein, Grinevald and Seifart (2004) contrast incipient noun class systems of Amazonia to more grammaticalized Bantu noun class systems (which in fact exhibit more variability than acknowledged in the literature and can be very old, sometimes even decaying).

The parameters considered to establish the degree of grammaticalization of a system are as listed in Table 12.

The Productivity of the System

Independently of its origin, its type or its age, any system can be shown to be more or less productive, adapting or not to linguistic and cultural development by incorporating new items in its classification schemata. For instance, the Thai

Table 12 On the grammaticalization of classification systems

	Less grammaticalized systems	*More grammaticalized systems*
System properties	Do not classify all nouns (nonexhaustive)	Classify ALL nouns (exhaustive)
	Few classes if emergent, to very LARGE NUMBER of classes if fully developed	SMALLER number of classes
	OPEN and productive	CLOSED
Use	Possibility of assigning N to DIFFERENT CLASSES	N assigned to a UNIQUE CLASS
	Determined by speaker's pragmatic CHOICE	NO speaker VARIATION
	Varies in formal/informal contexts	No variation according to register
Morphosyntax	Not affixed to the noun	Possibility of marking on the noun itself
	INDEPENDENT constituent without fusion	FUSED with other grammatical categories (def, nb, case)
	No or limited agreement schema: marked for pragmatic purpose	Part of rigid AGREEMENT schemata (beware of effect of standardization; freer in speech!)

numeral classifier system, which is very old, is also very productive: it is open and adapting to the language of modern life, in contrast to the noun classifier system of Jakaltek-Popti', which is not very old, but seemed frozen and unable to cope with the classification of modern imports and products when studied in the 1970s (Craig, 1986b).

All these parameters of variation make it clear that no two classification systems, even of the same type or subtype, will resemble each other closely, although they remain comparable when caution is taken to factor in those parameters of variation. What is striking in fact, in the midst of the surface variation of all these classification systems in terms of origins and paths of evolution, is precisely their shared basic principles of semantic categorization of the world humans live in. In their classificatory schemata, made at once of cognitive universal patterns and cultural specificities, they are of great interest to various disciplines beyond linguistics itself.

Discussion

The situation in classifier studies today is that of a pervasive confusion about what is being discussed, due to unsettled issues of terminology (that persist even at the time of this writing).

Classifiers from a Lumping versus Nonlumping Position

As set out from the start, a primary issue to clarify would be the question of lumping all classification systems together (Aikhenvald, 2000, 2012; Croft, 1994; Seifart and D. Payne, 2007) or not lumping them and identifying a subset of systems called classifiers (the line mainly argued in Grinevald, 2000 inter alia). Much depends on an a priori positioning, clearly influenced by the nature of the data that individual linguists handle. The fact is that there can be overlap among systems, of lexical sources, of categorization schemata, when systems sometimes emerge one from another, while discussions of dynamics and levels of categorization can apply equally to both noun class and classifier systems. However, from a strictly linguistic point of view, a nonlumping morphosyntactic approach facilitates a concentration on the specificities of 'classifier systems' per se as one-place grammatical classification systems, and allows for better attention to their own varied subtypes, as shown above.

It is worth noting, in addition, that if all systems of linguistic categorization were to be called 'classifiers,' in a lumping approach, then another term would need to be invented for the particular set of classification systems labeled 'classifiers' in a nonlumping approach.

Further Terminological Issues

Independent of the lumping issue, some researchers question altogether the appropriateness of the term 'classifier' or the pertinence of talking of classification in general, since the term classifier can be seen as a misnomer indeed, if taken literally as necessarily implying classifying. The problem often raised is that of the phenomena of 'repeater' and 'unique' classifiers, where 'repeater' refers to a classifier that is a copy of a noun and 'unique' to one that heads a class of just one item, neither appearing to classify sensu stricto. But classification is far from the only or even, at times, the main function of classifiers, which lay important and diversified roles in discourse (as discussed in the collections of Craig, 1986a; Senft, 2000, for instance, and further explored in Contini-Morava and Kilarski, 2013).

The term classifier cannot be taken either as applying to a specific linguistic category, since there is so much variation along the parameters outlined above, so that no two systems are indeed the same. The appropriate way to use the term is therefore as a cover term, maybe best written in caps as some type of grammatical gloss, as CLASSIFIER, to signify its status of abstract comparative concept to permit a typological approach through the great language specific diversity of such phenomena locally described in family specific terms.

At a more local level, the terminological problem extends to the naming of the subtypes of classifier at two levels, the lexical domain of classification (of nouns or verbs, that could also be labeled nominal vs verbal classification) versus the morphosyntactic locus of the classifying element (again potentially either set of terms). But the present situation remains one of nonalignment of terms between these two levels which comes from the evolution of the discussions over time, from language particular descriptions to a more global view of all systems for typological purposes. The choice was made not to correct this problem in this article since one of its major goals was to report on the existing literature, but the task remains, particularly now that an even wider range of systems has been described, inviting another round of typological studies.

Another unsettled question of terminology is that of the labeling of the different levels of categorization identified, with alternative naming existing at each level identified. For instance, the most desemanticized classifier is called either 'default' or 'general,' while the most common one is called either 'general' or 'generic,' or even 'specific,' although some reserve the term specific for a lower level of categorization. Only the level of unique classifier is uniformly recognized and labeled. The most important point has been made, however, that classifiers define more or less heterogeneous classes, through processes of extension and progressive desemanticization of existing classifiers, with a very language specific selection of specific items of culture for special treatment.

Additionally, there is sometimes a confusion between two types of classifiers, where 'unique' refers to a classificatory function (of class contains of only one item), and repeater to the form of the classifier (whether full or truncated form of its lexical source, with no particular link to the process of classification). The nature of unique and repeater classifiers was illustrated in Table 13.

Table 13 Repeater versus unique in Jakaltek Popti'

CL		*Repeater*	*Unique*	*CL*		*N*
atz'am	'Salt'	+	+	*atz'am*	*atz'am*	'(The) salt'
metx'	'Dog'	−	+	*metx'*	*tx'i'*	'The/a dog'
ha	'Water'	+	−	*ha*	*ha'*	'The water'
				ha	*nhab'*	'The rain'
				ha	*pam*	'The/a lake'
				etc.		

Conclusion

The domain of linguistic classification systems is very rich and varied, marked by a fluidity across a number of possible systems which share their being secondary systems of lexical origin forming more or less grammaticalized systems. This article has concentrated on the systems known as classifiers in a nonlumping position and proposed a morphosyntactic approach to the typological study of classifiers, with the view of encouraging the linguistic description of as many new systems as possible. The lumping position has been taken more recently by linguists dealing generally with the intricate interweaving and overlapping of classificatory systems typical of many Amazonian systems, more focused on semantic similarities than on the structural specificities of the systems. Such a lumping position is also the approach of a general study of the functions of the different nominal classification systems: semantic, such as expanding the referential power of the lexicon; and discourse/pragmatic, such as establishing and manipulating the status of discourse referents (Contini-Morava and Kilarski, 2013). The challenge of accounting for the nature and specificities of linguistic classification systems persists, but interest in them remains high as they provide a unique window on issues of cognitive and cultural categorization, and, as emphasized here, on the dynamics of grammar-making. Language systems in other modalities, be they contemporary sign languages or ancient hieroglyphic writings have confirmed in recent times how essential linguistic categorization is for any type of human language.

See also: Australian Aboriginal Society and Culture: An Overview; Classification, Understandings of in the Social Sciences; Historical Thought and Historiography: Southeast Asia; Sign Language: Psychological and Neural Aspects.

Bibliography

Adams, K.L., Conklin, N.F., 1973. Towards a theory of natural classification. Papers from the Regional Meeting of the Chicago Linguistic Society 9, 1–10.

Aikhenvald, A., 2000. Classifiers: A Typology of Noun Categorization Devices. Oxford University Press, Oxford.

Aikhenvald, A., 2012. The Languages of the Amazon. Oxford University Press, Oxford.

Allan, K., 1977. Classifiers. Language 53, 285–310.

Berlin, B., 1965. Tzeltal Numeral Classifiers. Mouton, The Hague.

Bergman, B., Wallin, L., 2003. Noun and verbal classifiers in Swedish sign languages. In: Emmorey, K. (Ed.), Perspectives on Classifier Constructions in Sign Languages. L Erlbaum Associates, Mahwah, NJ, pp. 32–52.

Bisang, W., 1999. Classifiers in East and Southeast Asian languages: counting and beyond. In: Gvozdanovic, J. (Ed.), Numeral Types and Changes Worldwide. Mouton de Gruyter, Berlin, pp. 113–185.

Champollion, J.-F., 1836. Principes généraux de l'écriture sacrée égyptienne. Institut d'Orient/Michel Sidhom, Paris.

Contini-Morava, E., Kilarski, M., 2013. Functions of nominal classification. Language Sciences 40, 263–299.

Craig, C.G. (Ed.), 1986a. Noun Classes and Categorization. John Benjamins, Amsterdam.

Craig, C., 1986b. Jacaltec noun classifiers: a study in language and culture. In: Craig, C.G. (Ed.), Noun Classes and Categorization. John Benjamins, Amsterdam, pp. 263–295.

Craig, C.G., 1987. Jacaltec Noun Classifiers: A Study in Grammaticalization, Lingua 70, 241–284.

Croft, W., 1994. Semantics universals in classifier systems. Word 45, 145–171.

Denny, J.P., 1976. What are noun classifiers good for? Papers from the Regional Meeting of the Chicago Linguistic Society 12, 122–132.

Dickinson, C., 2000. Complexpredicates in Tsafiki. Proceedings of the Annual Meetings of the Berkeley Linguistic Society 26 (2), 27–37.

Dixon, R.M.W., 1982. Noun classifiers and noun classes. In: Dixon, R.M.W. (Ed.), Where Have All the Adjectives Gone? and Other Essays on Semantics and Syntax. Mouton, The Hague, pp. 211–233.

Dixon, R.M.W., 1986. Noun classes and noun classification in typological perspective. In: Craig, C.G. (Ed.), Noun Classification and Categorization. John Benjamins, Amsterdam, pp. 105–112.

Emmorey, K., 2002. Perspectives on Classifier Constructions in Sign Languages. L Erlbaum Associates, Mahwah, NJ.

Gil, D., 2013. Numeral classifiers. In: Dryer, M.S., Haspelmath, M. (Eds.), The World Atlas of Language Structures Online. Max Planck Institute for Evolutionary Anthropology, Leipzig. Available online at: http://wals.info/chapter/55 (accessed on 05.08.14.).

Goldwasser, O., 2006. A comparison between classifier language and classifier script: the case of ancient egyptian. In: Goldenberg (dir), G. (Ed.), A Festschrift for Hans Jakob Polotsky. Magnes Press, Jerusalem, pp. 16–39.

Goldwasser, O., Grinevald, C., 2012. What are determinatives good for?. In: Grossman, E., Polis, S., Winand, J. (Eds.), Lexical Semantics in Ancient Egyptian, vol. 9. Widmaier. Lingua Aegyptia Studia Monographica, Hamburg, pp. 17–53.

Grinevald, C., 2000. A morphosyntactic typology of classifiers. In: Senft, J. (Ed.), Nominal Classification. Cambridge University Press, Cambridge, UK, pp. 50–92.

Grinevald, C., 2002. Making sense of nominal classification systems: noun classifiers and the grammaticalization variable. In: Wischer, I., Diewald, G. (Eds.), New Reflections on Grammaticalization. John Benjamins, Amsterdam, pp. 259–275. Typological Studies in Language 49.

Grinevald, C., 2004. Classifiers. In: Lehmann, C., Mugdan, J., Skopeteas, S. (Eds.), Morphology: An International Handbook on Inflection and Word-Formation, vol. 2. Walter de Gruyter, Berlin/New York, pp. 1016–1031.

Grinevald, C.G., Seifart, F., 2004. Noun classes in African and Amazonian languages: Towards a comparison. Linguistic Typology 8, 243–285.

Lakoff, G., 1986. Classifiers as a reflection of mind. In: Craig, C. (Ed.), Noun Classes and Categorization. John Benjamins, Amsterdam, pp. 13–51. Typological Studies in Language 7.

Lakoff, G., 1987. Women, Fire and Dangerous Things: What Categories Reveal about the Mind. The University of Chicago Press, Chicago.

Matsumoto, Y., 1993. Japanese numeral classifiers: a study of semantic categories and lexical organization. Linguistics 31, 667–713.

McGregor, W., 2002. Verb Classification in Australian Languages. Mouton de Gruyter, Berlin.

Mithun, M., 1986. In: Craig, C.G. (Ed.), The convergence of noun classification systems, pp. 379–397.

Rehg, K., 1981. Ponopean Reference Grammar. Pali LanguageTexts: Micronesia. UniversityPress of Hawaii, Honolulu.

Seifart, F., Payne, D., 2007. Nominal classification in the North West Amazon: issues in areal diffusion and typological characterization. International Journal of American Linguistics 73 (4), 381–387.

Sands, K., 1995. Nominal classification in Australia. Anthropological Linguistics 37, 247–346.

Seiler, H., 1986. Language, Object, and Order: The Universal Dimension of Apprehension. Gunter Narr, Tübingen.

Seiler, H., Lehmann, C. (Eds.), 1982. Apprehension. Das Sprachliche Erfassen von Gegenstäden. Bereich und Ordnung der Phänomene, vol. 1. Narr, Tübingen.

Seiler, H., Stachowiack, F.J. (Eds.), 1982. Apprehension. Das Sprachliche Erfassen von Gegenstäden. Die Techniken und ihr Zusammenhang in Einzelsprachen, vol. 2. Narr, Tübingen.

Senft, G., 2000. Systems of Nominal Classification. Cambridge University Press, Cambridge.

Schultze-Berndt, E., 2000. Simple and Complex Verbs in Jamingjung: A Study in Event Categorization in an Australian Language. Ph.D. thesis. University of Nijmegen.

Classroom Assessment

James H McMillan, Virginia Commonwealth University, Richmond, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P.W. Airasian, volume 3, pp. 1978–1983, © 2001, Elsevier Ltd.

Abstract

This article defines and discusses the essential characteristics of classroom assessment. Classroom assessment consists of procedures teachers use to gather information about student performance. This comprises both informal and formal techniques, aligned to learning objectives and used for documenting student learning (summative assessment) and improving student learning (formative assessment), as well as influencing student motivation. Characteristics of high-quality classroom assessment are presented, including validity, reliability, and fairness. Grading is also discussed.

Classroom assessment is a set of strategies, techniques, and procedures that teachers and students engage in to collect, evaluate, and report student achievement. It is a tool teachers use to gather data and information about how students are progressing in what they know and in the development of their skills. It is generally focused on K-12 education, although many principles are also applicable to higher education. Classroom assessment is used for a variety of purposes – to document what students know and can do, diagnose strengths, weaknesses, and misunderstandings, improve and enhance student learning, motivate students, assign grades, and provide feedback to parents. It is much more than 'testing,' although tests have been used extensively as one type of classroom assessment.

Classroom assessment is different from other kinds of assessments that are used in education. They are primarily controlled and used by classroom teachers, and are often unique to each class. Teachers use different types of classroom assessments on a daily basis, as well as over weeks or even months. A different type of assessment could be classified as *standardized*. Standardized tests have specific and set procedures, formats, and scoring levels, and are used for many students across different classes, schools, and states. They are often called *large-scale* tests. While there are many types of standardized tests that have been used for decades, including standardized achievement and aptitude tests, the most common ones now used in schools are externally prepared and scored accountability tests. Such tests are mandated by states and used to document student achievement at the end of the year. Psychological tests, such as personality tests, are also used with students, as are a multitude of other types of assessments that are used for research purposes.

Components of Classroom Assessment

There are five components of classroom assessment: purpose, learning objectives or targets, measurement, interpretation, and use. Purpose is the reason for the assessment, why it is given. This includes such uses as obtaining diagnostic information, providing feedback to students, documenting levels of achievement, motivating students, grading, reporting to parents, instruction, and, recently, preparation for taking accountability tests.

Learning Objectives or Targets

Learning objectives or targets are statements that define what students are expected to learn. Since the early 1990s the term *standards* has been used to designate what students should learn at different grade levels in each subject. It is important for these statements to be clear and reflect the range of content and skills to be emphasized, as well as the level of cognition required to demonstrate what has been learned. States have identified standards that are the basis for accountability tests. They tend to be broad, covering an entire year of learning. Objectives and learning targets are more specific and refer to what is to be achieved over short units of instruction.

The level of cognition or thinking needed for meeting learning objectives is a critical part of classroom assessment. Cognition can be very simple and consist of recalling or remembering something, or more complex in requiring knowledge utilization with skills such as application, analysis, reasoning, critical thinking, or evaluation of something (these are often called *higher-order* skills) (Marzano and Kendall, 2007). Educators often use Bloom's taxonomies to define levels of skills (Anderson and Krathwohl, 2001; Bloom, 1956), although there are others that reflect more recent research (e.g., Marzano and Kendall, 2007). As long as the objective is clear it will provide a sound basis for the assessments and how results will be scored.

Measurement

Measurement is the process of quantification of behavior or performance, the assigning of numbers to suggest a level of achievement. Measurement is used to differentiate levels or degrees of knowledge, understanding, and skills. There are many types of assessment that are used to enable the measurement process, including *selected-response* tests, such as multiple-choice, true–false, and matching, where students select an answer from choices that are provided, and *constructed-response* assessments, such as essay, papers, and projects, in which students must provide their own work without being given choices. The result of measurement is typically a score, reported as either a number, or a grade, or a level of performance.

Interpretation

Once a result of student performance is obtained it needs to be interpreted or evaluated to determine what it means. This involves the use of a variety of benchmarks and other information that assists teachers in making professional judgments about the meaning of the results. For example, what does achieving 70% correct on a test mean? Is this average, less than average, or even failing? Does it suggest an incomplete understanding of what was learned? Teachers' professional judgments are informed from their experience and guidelines, as well as from comparing students, and are needed to make correct decisions. 'Objective' assessments such as multiple-choice tests are objective only in the manner in which they are scored, not interpreted. There is still a need for interpretation.

Use

Use of the results should be consistent with purpose. Once results are interpreted teachers use them in many different ways, including grading, reporting to parents, planning subsequent instruction, and feedback to students.

Summative and Formative Classroom Assessment

Perhaps the most significant development in classroom assessment over the past two decades is the development of what is called assessment *for* learning, or *formative assessment* (Andrade and Cizek, 2010; Black et al., 2003; Stiggins, 2008). It is contrasted with *summative* assessment, which is used at the end of a period of student learning to document student performance, to show student achievement. Formative assessment is typically ongoing, with the primary purpose of enhancing student learning. While summative assessments tend to be formal, objective, and conducted only at the end of a unit of instruction, assessment *for* learning tends to be informal, occurring throughout and embedded in learning activities. Evidence of student learning is used by teachers to give students feedback and to make adjustments in their instruction. Learning is actively monitored as it occurs to help provide hypotheses about progress toward learning goals and objectives and instructional effectiveness. As such, formative assessment is integrated with instruction, rather than being something that occurs after instruction has concluded to provide a summative and static measure of achievement.

The goal of formative assessment is to improve student learning and motivation by enacting five essential steps:

1. Establishing criteria for evaluating student learning and understanding;
2. Collecting evidence of student learning and understanding;
3. Providing feedback to students on their progress toward reaching learning targets;
4. Involving students in self and peer assessments; and
5. Using instructional adjustments to help students learn.

Criteria are the elements of student performance that are important determinants of success. They are contained in student responses to evaluate progress toward meeting the learning targets; the dimensions or traits in products or performance that are used to determine level of understanding. For example, criteria to evaluate a persuasive essay could include the following criteria:

- Appropriateness of language for the audience
- Plausibility and relevance of supporting arguments
- Level of detail presented
- Evidence of innovative thinking
- Clarity of expression
- Organization of ideas

Criteria for an evaluation of graphic display of data could involve:

- Title
- Labels
- Accuracy
- Neatness

Once criteria are established and communicated to students, evidence of student understanding is gathered (Heritage, 2013). For some situations this could be structured as a pretest or pre-assessment, given to students before instruction is started. Pretests can be effective, as long as students are motivated to do their best. Other approaches to pre-assessment, such as giving students a short assignment or task and observing their performance, can be useful. During instruction, pre-assessment is often accomplished with the effective use of questioning or assigning short tasks. Verbal responses are evaluated, as are nonverbal indications such as facial expressions and body posture. Here the teacher needs to ask the right kind of questions, ones that engage and challenge students and elicit their thinking. These typically are not questions simply requiring right and wrong answers (e.g., 'yes' or 'no'), and teachers need to be patient to allow for student thinking before answering. Often, it is effective to have students share potential answers with each other. Short quizzes or other tasks, as well as interviewing with students, are also useful in obtaining evidence of student progress.

Once the evidence is gathered, feedback is given to students about their performance. There is much research on the nature of feedback that is effective in helping students learn (Brookhart, 2008). The timing and nature of the feedback needs to be appropriate, coming as soon as feasible after the performance and being specific and individualized. General evaluative statements such as 'good job' or 'correct' or 'excellent work' are not considered specific and individualized, and may in fact demotivate students. Rather, the information provided to the students must be directly tied to the criteria and their specific responses, and should be targeted to each student or small groups of students. The idea is to help the student see how their performance is making progress toward the learning target, in specific ways, and to help students understand that their lack of achievement is primarily due to lack of effort. Effort attributions are very important for the development of perseverance by helping students understand that success is more determined by what they do in the future than what has been done.

In general, effective feedback from information obtained from assessment has the following characteristics (McMillan, 2014):

- Focused on connecting evidence of learning to progress toward successful outcomes

- Positive without sugarcoating or effusive praise
- Honest and accurate
- Specific, individualized, and differentiated to students as appropriate
- Timely, typically immediately
- Descriptive of work accomplished, not about the student more generally
- Actionable, so students can take next steps
- Engaging so that students generate feedback
- Differentiated to produce the best motivation for each student
- Highlight student errors or omissions
- Focused to push students to next levels of learning
- Encourages a mastery rather than performance orientation toward learning

The final step in effective assessment *for* learning is to provide instructional adjustments or correctives. In this step the teacher helps students understand what else can be done to learn the material or deepen understanding. This typically means more learning activities that are different from what was done previously. Often students can come up with their own good ideas about what they can do to improve their understanding, which is excellent in promoting self-monitoring and self-determination.

The idea of assessment *for* learning is that it makes explicit the way assessments are designed and used to have a positive effect on student achievement and motivation. It is a philosophy of how curriculum, instruction, and assessment can be integrated to impact what students learn and their engagement in learning. It is especially helpful for the development of student dispositions needed for success in their careers and lives because it focuses on mastery of knowledge and skills, not simply performance for a good test score, and because it enhances students' abilities to evaluate their understanding of something, identify where the gap is between what they know and the target, and then design what further learning activities are needed to reach the achievement goals. It is also a frame of mind for teachers – how committed they are to improving student learning and motivation.

Characteristics of High-Quality Classroom Assessments

For classroom assessment to be successful in meeting its goals there are several indicators of quality that translate into effective practice. Classroom assessments that have high quality result in accurate, trustworthy verification of learning, appropriate motivation of students for learning, and helpful information to improve instruction. One characteristic already indicated is the determination of clear objectives and learning targets. It is best if these are understood by students and parents as well as teachers.

Alignment

Alignment is the degree to which different components of assessment that are essential to produce intended outcomes are in agreement, supportive, or consistent. There are several important ways that alignment needs to be demonstrated, beginning with the learning objectives or targets.

Assessment is most effective when it is aligned with a teacher's philosophy of education and style of instruction. Some teachers are formal in their approach to instruction, while others take a more facilitative, personal mode of interaction with students. It is important for teaching styles to be aligned with student learning styles, and to be aligned with learning targets and assessments. Teachers with a formal approach may tend to lecture and give traditional tests. Teachers with a more collaborative style include more discussion, project-based activities, cooperative learning, and assessments that are performance-based, authentic, and require students to apply their knowledge.

Classroom assessments need to be structured, administered, and scored in alignment with contemporary theories of human learning and motivation. The research in these areas has emphasized the critical role that students' existing knowledge plays in making connections between what is new information and what is currently understood. Learning is most effective when it is meaningful; motivation is most effective when it is mastery oriented and develops self-monitoring and self-efficacy in students (students' beliefs that they are capable of learning). Implications of these theories for assessment are clear for both instructions and assessment, calling for more authentic, performance-based, and project-based assessments rather than objective tests, more variety in assessment methods, more formative assessment, and more individualized feedback. Mastery goal setting is promoted by the use of clear targets and criteria for scoring assessments prior to testing, the use of exemplars, and student self-assessment. The goal is to maximize student engagement and meaningfulness in learning to establish effort attributions for both success and failure.

Another important type of alignment is between the nature of learning targets and type of assessment. Assessments can be categorized broadly into four major types: selected response, constructed response, teacher observation, and student self-assessment (McMillan, 2014). (Peer assessment could be listed as a fifth category, although it is not widely used.) Selected-response tests consist of multiple-choice, matching, true–false, short answers, and completion sections. Constructed-response assessments include performance, essay, and oral questioning. *Performance assessments* require students to construct an extensive, complex answer or response, often in the form of a presentation or something that is constructed or accomplished. The assessment is a product such as a speech, athletic demonstration, musical recital, or project. These types of assessments require the development of a meaningful task to complete and a scoring rubric that specifies different levels of performance with criteria. Teacher observation can be formal but is more often informal. *Student self-assessment*, or self-reflection, refers to students rating their own performance.

Different types of assessment are best at measuring certain kinds of learning objectives or targets. Well-constructed selected-response items are best at measuring subject matter and procedural knowledge and comprehension, especially when students need to recognize or remember facts, dates, people, historical events, spellings, and concepts. Deep understanding and reasoning is best assessed with essays and performance assessments. Products and performances are only well assessed with performance assessments. Teacher observation is effective in measuring student dispositions, such as attitudes,

cooperativeness, effort, trust, and responsibility. Student self-assessments are effective in gathering data on student dispositions and students' abilities to understand learning criteria, what progress has been made, and what else needs to be learned. This self-regulation process contributes to self-determination and a sense of self-control, which provides intrinsic motivation and stronger positive expectations for future learning.

Finally, classroom assessments need to be well aligned with state standards. This requires a thorough understanding of the levels of cognition required in state standards, the types of items in tests used to assess state standards, and how classroom assessment formats and procedures match these dimensions. A rough match can be made simply on the basis of content; a more sophisticated, refined match is achieved by understanding the depth, range, scope, and balance of knowledge and skills needed.

Validity

Validity is the extent to which the use of student assessments is appropriate. It is a measure of the reasonableness of the interpretations and inferences that are made from the results, as well as the consequences that follow. For example, if a conclusion is reached that a student's performance indicates that there is poor achievement, is that true? Or, if an assessment indicates that a student is proficient, what is the chance that is wrong? If a consequence of an assessment is that students learn only to study to pass the test and not to master the content, is that what is desired?

While for standardized tests technical approaches to establishing validity are well defined, validity for classroom assessments relies more on common sense, comparisons, logic, and careful examination of the nature of the assessments. This evidence is gathered by constructing a *test blueprint* or *table of specifications*. The blueprint is a matrix that shows the type of learning that is measured by different areas of the test. The number and percent of items covering different cognitive tasks in each area provides a picture to be sure that the content and levels of cognition required covers what is meant to be covered. This could be called *content-related evidence* for validity, and it is the most common type of evidence for classroom assessments. A related kind of evidence focuses on *instructional sensitivity* by comparing what is on the test with what was taught. Often both content-related evidence and instructional sensitivity are established by having others review the assessments in the context of what was taught during the instructional segment.

Another type of evidence for validity comes from several measures of the same thing when the results point toward the same conclusion. This occurs when the quiz, homework, test, and project grades on the same thing are similar, or when teachers' observations of understanding from questioning match how students perform on a unit test. It is best to use multiple measures of the objective or learning target if possible, ideally with different methods of measurement.

Reliability

All classroom assessments contain some degree of error or imprecision, some amount of 'noise' that interferes with an absolutely correct result. In measurement, the concept of *reliability* is used to define and estimate the amount of error that is contained in a single assessment. Reliability, then, is an indication of stability, consistency, or dependability of the scores that are obtained from an assessment. The less the error in the measure, the greater the reliability. The amount or degree of error in a classroom assessment, like validity, is usually a matter of professional judgment. Error can come from the nature of the assessment and scoring procedures, such as from ambiguous directions and poorly worded items, to biased or tired scorers. Error also comes from the student, with such factors as being overly tired or sick on the testing day, not being motivated to do the best work, anxiety, or how lucky the student was in answering questions. Usually teachers are not aware of how these sources of error affect results, so a general understanding of conditions and students is needed. A more reliable score is obtained when several questions are asked about the same thing. This occurs when a math test has several two-digit multiplication items to measure this skill. If all the items are answered correctly by one student who clearly knows how and incorrectly by a student who clearly does not know how, there is evidence that at least the test does not contain much error. However that does not address other sources of error, so even with the best of tests final scores should be interpreted with a margin of error, much like what is reported in political polling. That means that a score of 85% is more accurately something like 82–87% or 83–86%. The students' actual or 'true' knowledge or capability is something within the range. Assessment practices that can improve reliability include providing clear guidelines for answering and scoring, providing several items or tasks to measure the same thing, using independent raters or scorers if possible, keeping interruptions when students take a test to a minimum, using exemplar answers and papers for scoring, using specific scoring rubrics, and using shorter assessments more frequently rather than longer assessments less often.

Fairness

A *fair* assessment is one that provides all students, regardless of their characteristics and circumstances, a reasonable opportunity to demonstrate achievement. In this sense, fair assessments are unbiased and nondiscriminatory (Tierney, 2013). A *biased* assessment is influenced by factors that are not associated with the knowledge or skill to be demonstrated. This can result from biased assessment content that may unfairly penalize a student or be offensive due to culture, income, ethnicity, or gender (Popham, 2008). Unfair penalization occurs when some students, because of their status, experience, or characteristics, are at a disadvantage in taking the test. Conversely, some students may have an advantage. Suppose, for example, that a math test item uses a sailing example that makes it easier for those students who have sailed and more difficult for those who have never sailed. This would be a biased item. Cultural differences reflected in vocabulary, values, and conventions can unfairly influence the ability of some students to show what they know and can do. Even students' learning style preferences can influence the performance of some students on some types of assessments. Students can also be offended by culturally based, demeaning content in assessments, such as portraying a given race as having only low-status jobs. Finally, fairness is reflected in the extent to which appropriate

accommodations are included for students with special needs and English Language Learners.

Practicality and Efficiency

A continuing challenge to good teaching is to balance the potential complexity and extensiveness of many assessments with the need to be both practical and efficient. Teachers have limited time for assessment. The ideal is to use just enough assessment to reach valid conclusions and promote positive consequences for students (e.g., enhancing their motivation, self-reflection, and overall achievement). It is best to consider the time needed to construct, administer, and score assessments. Some constructed-response assessments may be great for promoting thinking skills, but there needs to be consideration of what will be required to do the scoring and give feedback. Objective tests are easily scored, although much time is needed to produce good multiple-choice questions, ones where one or two alternatives are not obviously incorrect. Other things being equal, it is best to use the shortest possible assessments, as long as they provide valid, reliable, and fair results.

Grading and Reporting

Grading (marking) and reporting are long-standing practices that are closely aligned to classroom assessment. The results of classroom assessments are used, sometimes exclusively, to determine grades. Grades are an indication of a student's achievement or performance (Brookhart, 2009). Marks or symbols are used to convey a message that has a predetermined meaning. Letter marks, from A to F, are used extensively in grading. It is also common to use percentage correct to correspond to each grade (e.g., 90–100% – A, 80–89% – B, and 70–79% – C), and often pluses and minuses are added to the letters when the scores are at the ends of each range of percentages. The grade is then labeled with a descriptor such as 'excellent,' 'good,' 'fair,' 'poor,' or 'failing.' Traditionally, letter grades are not distributed evenly to a group of students. That is, fewer students typically receive As, Ds, and Fs, and most students receive Bs or Cs, although the distribution can vary significantly from one class to another. This is often left to the option of the teacher. While teachers try to be 'objective' in their grading and use measures of academic achievement, some level of subjectivity and professional judgment is needed. Naturally, this can vary from one teacher to another, even for the same course or same group of students. Some teachers make hard tests, others easy tests. Many teachers use some form of extra credit, and improvement and effort is often used as part of what determines the grade. Grading essays, papers, and projects is essentially a matter of professional judgment. In the end, each teacher uses his or her experience and knowledge in constructing and grading classroom assessments. Grades in higher-level classes, such as honors and advanced placement classes, are often given higher points than are given in lower-level classes.

A more recent trend in grading is to focus on absolute levels of performance. This is called *standards-based* grading. Standards-based grading is very similar to what was called criterion-referenced grading in the sense that grades are based on preestablished levels of performance that are thoroughly and clearly described and understood. The standards are sometimes state standards, although typically teachers use descriptions of different levels of performance in classroom assessments as well. The levels are differentiated by using a rubric that includes descriptors to indicate increasingly more sophisticated understanding or skills. For example, the scale could begin with 'needs improvement,' progress to 'approaching,' then 'proficient,' and finally 'advanced.' The descriptors in the rubric vary considerably, but each has a definition to indicate in general what the student knows, understands, and is able to do.

With standards-based grading the basis of comparison is to examine student work and match it with the appropriate levels of performance. This is in contrast to what is called *norm-referenced* grading, in which the basis of comparison is how other students perform. In norm-referenced grading only the best achieving students, as compared to others, get the highest grades. This is a long-standing approach to grading and is still important for ranking students for such purposes as achieving honors, admittance to college, and employer evaluations.

A final and less frequent type of grading is *self-referenced*. The essence of self-referenced grading is to compare students' final performance after instruction to their performance in the past, or, less frequently, to their aptitude. The logic here is that what is most important is growth in performance, not simply the level of performance. Thus, students may initially have a very weak understanding and after instruction advance to an adequate or proficient level. This could indicate a significant and important amount of growth, and be graded highly. A student who begins at the proficient level but does not progress much would receive a lower grade. While measuring student progress is now critical to evaluating teachers and schools, it is not without disadvantages. Measuring growth with classroom assessments, or any kind of assessment for that matter, suggests a probable doubling of testing time, relies on having reliable measures, assumes appropriate student motivation when taking 'pretests,' and somehow getting 'aptitude' right if that factor is used as a premeasure.

Final semester or course grades are determined from classroom assessments as a summary indicator of student performance. Web-based and other software grading programs are used to combine grades, points, and percentages of all the assessments to result in a final score, which is then usually converted to a grade. Teachers combine the individual assessments, deciding how much weight each one should be given and whether scores are averaged or added (total points method). Exams are typically given more weight than are quizzes and homework. Both students and parents are usually able to see grades and final calculations online.

Along with reporting grades, teachers indicate to students and parents strengths, weaknesses, and areas that need further improvement. This is accomplished through parent–teacher and student–teacher conferences, and/or in writing (Stiggins, 2008). In some instances, the student takes leadership for the conference, taking parents through a detailed review of their work for the grading period. Often this includes a portfolio of student products. The portfolio, which can be in a paper or electronic file, will contain periodic examples of student performance, such as tests, papers, and quizzes, along with student reflections of the meaning of the results. This type of student self-assessment is one of the most recent developments in classroom assessment, along with peer assessment. Both of these approaches are well

aligned with contemporary theories of motivation and learning, and will probably be commonplace in the future.

Summary

Classroom assessment is something all teachers do as part of their responsibility for enhancing student learning and motivation, documenting student performance, and reporting results for accountability purposes. There are many different types of classroom assessments that can be used, each with advantages and disadvantages. To the extent that teachers base their assessments on known indicators of high quality, such as reliability and validity, align their assessments with clear and appropriate learning objectives and targets, use formative assessment with meaningful feedback, and grade student work to provide meaningful summary indications of performance, student learning and motivation will be enhanced, as well as documented.

See also: Academic Achievement Motivation, Development of; Attitude Measurement; Cognitive Development, Educational Theories of; Educational Learning Theory; Equivalence and Transfer Problems in Cross-Cultural Research; Intelligence: Assessments of; Intrinsic Motivation: A Cultural Perspective; Motivation, Learning, and Instruction; School Achievement as Moderated by Sociodemographic Status: The US Case; School Achievement: Motivational Determinants and Processes; Self-Determination Theory; Self-Efficacy: Education Aspects; Self-Efficacy; Self-Regulated Learning: Theories, Measures, and Outcomes; Self-Regulated Learning; Self-Regulation During Early Childhood Across Cultures, Development of; Test Anxiety and Academic Achievement; Vygotsky's Theory of Cognitive Development.

Bibliography

Anderson, L.W., Krathwohl, D.R., 2001. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Longman, New York.

Andrade, H.L., Cizek, G.J. (Eds.), 2010. Handbook of Formative Assessment. Routledge, Oxford.

Black, P., Harrison, C., Lee, C., Marshall, B., Wiliam, D., 2003. Assessment for Learning-Putting It into Practice. Open University Press, Buckingham, UK.

Bloom, B.S. (Ed.), 1956. Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1. Cognitive Domain. David McKay, New York.

Brookhart, S.M., 2008. How to Give Effective Feedback to Your Students. Association for Supervision and Curriculum Development, Alexandria, VA.

Brookhart, S.M., 2009. Grading, second ed. Merrill, New York.

Heritage, M., 2013. Gathering evidence of student understanding. In: McMillan, J.H. (Ed.), Sage Handbook of Research on Classroom Assessment. Sage Publishing, Thousand Oaks, CA, pp. 1179–1196.

Marzano, R.J., Kendall, J.S., 2007. The New Taxonomy of Educational Objectives, second ed. Corwin Press, Thousands Oaks, CA.

McMillan, J.H., 2014. Classroom Assessment: Principles and Practice for Effective Standards-Based Instruction, sixth ed. Pearson, Boston.

Popham, W.J., 2008. Transformative Assessment. Association for Supervision and Curriculum Development, Alexandria, VA.

Stiggins, R.J., 2008. Student-Involved Assessment for Learning. Pearson Merrill Prentice Hall, Upper Saddle River, NJ.

Tierney, R.D., 2013. Fairness in classroom assessment. In: McMillan, J.H. (Ed.), Sage Handbook of Research on Classroom Assessment. Sage Publishing, Thousand Oaks, CA, pp. 125–144.

Classroom Climate

Barry J Fraser, Science and Mathematics Education Centre, Curtin University, Perth, WA, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews four decades of research on classroom climate, with a major thrust on the development, validity, and use of several frequently used and extensively validated questionnaires. Another purpose is to review several lines of past classroom climate research, including evaluation of educational innovations, associations between student outcomes and classroom climate, and teachers' action research attempts to improve their classroom climates.

Over 40 years ago, Herbert Walberg and Rudolf Moos began seminal independent programs of research which form the starting points for the work reviewed in this article. Walberg developed the Learning Environment Inventory (LEI) as part of the research and evaluation activities of Harvard Project Physics (Walberg and Anderson, 1968), whereas Moos developed numerous social climate scales, including those for use in psychiatric hospitals and correctional institutions, which ultimately led to the development of the Classroom Environment Scale (CES; Trickett and Moos, 1973). Walberg's and Moos' work built upon Lewin's (1936) seminal field theory, which recognized that both the environment and its interaction with personal characteristics of the individual are potent determinants of human behavior, and Murray's (1938) needs-press model, which allows the analogous representation of person and environment in common terms. The way in which this important pioneering work on perceptions of classroom climate spawned other research is reflected in books (Fisher and Khine, 2006; Fraser, 1986; Fraser and Walberg, 1991), literature reviews (Fraser, 2007), the American Educational Research Association's Special Interest Group on Learning Environments which began in the mid-1980s, the initiation in 1998 of Kluwer/Springer's *Learning Environments Research: An International Journal*, and the birth in 2008 of Sense Publishers' book series *Advances in Learning Environments Research* (Aldridge and Fraser, 2008).

Following the pioneering research of Walberg and Moos in the USA, two further programs of learning environments research emerged. In the Netherlands, Theo Wubbels and his colleagues began ambitious programmatic research focusing specifically on the interactions between teachers and students in the classroom and often involving use of the Questionnaire on Teacher Interaction (QTI). This research program is described in more detail below. In Australia, Barry Fraser and his colleagues began programmatic research, which first focused on student-centered classrooms and involved use of the Individualized Classroom Environment Questionnaire (ICEQ; Fraser, 1990). The ICEQ differs from the LEI and CES, which focus on teacher-centered classrooms, in that it assesses those dimensions that are salient in open or individualized classroom settings. Subsequently, this led to other specific-purpose classroom climate instruments in Australia and cross-validating and applying them for a variety of research purposes around the world, as discussed in detail later in this article.

Instruments for Assessing Classroom Climate

Historically Significant Questionnaires: LEI, CES, ICEQ, and MCI

As noted above, the LEI, CES, and ICEQ were important in the evolution of the field of classroom climate even though they are seldom used in contemporary educational research. On the other hand, the My Class Inventory (MCI) – a simplified version of the LEI for younger students – is still used to some extent today because of its particularly low reading level.

Fraser et al.'s (1982) version of the MCI was developed for use among children aged 8–12 years, but subsequently Fisher and Fraser (1981) simplified the original version of the MCI and later Fraser and O'Brien (1985) evolved and used a short 25-item version. Although the MCI was developed originally for use at the primary school level, it has also been found to be very useful with students in the junior high school, especially with those who might experience reading difficulties with other instruments. The MCI contains only five of the LEI's original 15 scales, item wording has been simplified to enhance readability, the LEI's four-point response format has been reduced to a two-point (Yes–No) response format, and students answer on the questionnaire itself instead of on a separate response sheet to avoid errors in transferring responses from one place to another.

Goh and Fraser (1998) modified the MCI to involve a three-point frequency response format (Seldom, Sometimes, and Most of the Time) and a Task Orientation scale, and they used it in research in Singapore among primary mathematics students. In Brunei Darussalam, Majeed et al. (2002) used an English-language version of the MCI among 1565 lower-secondary mathematics students in 81 classes in 15 government schools and reported sex differences in classroom climate perceptions and associations between students' satisfaction and the classroom environment. In a small-scale evaluation of a K-5 program that integrates mathematics and children's literature, Mink and Fraser (2005) used the MCI in Florida and found congruence between students' actual and preferred classroom climate. In Texas, Scott Houston et al. (2008) used the MCI in an evaluation of science kits among a sample of 588 grade 3–5 students. As well as attesting to the validity of the MCI, using science kits was associated with a more positive classroom climate in terms of student satisfaction and cohesiveness. Sink and Spencer (2005) advocate the use of the MCI as an accountability tool for elementary-school counsellors. Using a large sample of 2835 grade 4–6 students in an

urban school district in Washington State, these researchers found that an 18-item revision of the MCI was psychometrically sound. Implications for elementary-school counseling programs and practices and their evaluation were provided by the authors.

Questionnaire on Teacher Interaction

As noted above, pioneering and programmatic research which originated in The Netherlands focuses on the nature and quality of interpersonal relationships between teachers and students (Wubbels and Brekelmans, 2012). Drawing upon a theoretical model of proximity (cooperation–opposition) and influence (dominance–submission), the QTI was developed to assess student perceptions of eight behavior aspects. Each item has a five-point response scale ranging from Never to Always.

Although research with the QTI began at the senior high school level in The Netherlands, cross-validation and comparative work has been completed at various grade levels in the USA (Wubbels and Levy, 1993) and Australia (Fisher et al., 1995) and a more economical 48-item version has been developed and validated in Singapore (Goh and Fraser, 1996). In Brunei Darussalam, Scott and Fisher (2004) validated a version of the QTI in Standard Malay with 3104 students in 136 elementary-school classrooms and showed that achievement was related positively to cooperative behaviors and negatively to submissive behaviors. In Singapore, Quek et al. (2005) validated an English version of the QTI with 497 gifted and non-gifted secondary school chemistry students and reported some stream (i.e., gifted and non-gifted) and sex differences in QTI scores. In Korea, a translated version of the QTI was validated and used by Lee et al. (2003) among 439 science students and by Kim et al. (2000) among 543 students. In Indonesia, a translated version of the QTI was validated with a sample of 422 university students by Fraser et al. (2010b). Brekelmans et al. (2005) traced changes in teacher–student relationships across the teaching career for samples from various countries.

Constructivist Learning Environment Survey

Taylor et al. (1997) developed the Constructivist Learning Environment Survey (CLES) to assist researchers and teachers to assess the degree to which a particular classroom's climate is consistent with a constructivist epistemology, and to assist teachers to reflect on their epistemological assumptions and reshape their teaching practices. Taylor and his colleagues reported sound factorial validity and internal consistency reliability for the CLES for several diverse samples.

Working with 1079 students in 59 science classes in North Texas, Nix et al. (2005) reported strong support for the validity of the CLES. An evaluation of an innovative science teacher professional development program revealed that the students of these teachers perceived their classrooms more favorably than did the students of other teachers. In a follow-up study in Texas, Nix and Fraser (2010) used a shorter version of the CLES in an evaluation of the implementation of the same professional development model over three semesters involving 17 teachers and 845 students. Changing teachers' classroom climate at the university level was found to foster similar changes in their students' middle-school classroom climates.

In a cross-national study of junior high school science classroom climates, the English version of the CLES was administered to 1081 students in 50 classes in Australia while a Mandarin translation was administered to 1879 students in 50 classes in Taiwan. Aldridge et al. (2000) reported sound validity for both English and Mandarin versions, and also that Australian classes were perceived as being more constructivist than Taiwanese classes (especially in terms of Critical Voice and Student Negotiation).

Peiro and Fraser (2009) modified the CLES, translated it into Spanish, and administered the English and Spanish versions to 739 grade K-3 science students in Miami, USA. Analyses supported the validity of the modified English and Spanish versions when used with these young children. Strong and positive associations were found between students' attitudes and the nature of the classroom climate, and a 3-month classroom intervention led to large and educationally important changes in classroom environment.

In South Africa, Aldridge et al. (2004b) administered the English version of the CLES to 1864 grade 4–6 mathematics learners in 43 classes. This led to the cross-validation of this version of the CLES for this population in terms of factorial validity, internal consistency, reliability, and ability to differentiate between classrooms. The primary focus of this study was to assist South African teachers to become more reflective practitioners in their daily classroom teaching. Through the use of the CLES in teacher action research, some improvements in the constructivist orientation of classrooms were achieved during a 12-week intervention.

When Kim et al. (1999) translated the CLES into the Korean language and cross-validated it with a sample of 1083 students in 24 grade 10 science students, results supported the factor structure and reliability of the Korean version, revealed statistically significant relationships between classroom climate and students' attitudes to science, and confirmed that students exposed to a new curriculum perceived a more constructivist learning environment than did students who had not been exposed to this curriculum.

Science Laboratory Environment Inventory

Because of the critical importance and uniqueness of laboratory settings in science education, an instrument specifically suited to assessing the climate of science laboratory classes at the senior high school or higher education levels was developed (Fraser et al., 1995; Fraser and McRobbie, 1995; Fraser et al., 1993). The Science Laboratory Environment Inventory (SLEI) has five scales (each with seven items) and the five frequency response alternatives are Almost Never, Seldom, Sometimes, Often, and Very Often. The SLEI was field tested and originally validated simultaneously with a sample of over 5447 students in 269 classes in six different countries (USA, Canada, England, Israel, Australia, and Nigeria). Subsequently, it was cross-validated in Australia with 1594 students in 92 classes by Fraser and McRobbie (1995) and 489 senior high school biology students by Fisher et al. (1997).

Fraser and Lee (2009) translated the SLEI into the Korean language in a study of differences between the classroom

climates of three streams (science-independent, science-oriented, and humanities) among a sample of 439 high school students. The Korean version of the SLEI exhibited sound factorial validity and internal consistency reliability, and was able to differentiate between the perceptions of students in different classes. Generally, students in the science-independent stream perceived their laboratory classroom climates more favorably than did students in either of the other two streams. Working with a sample of 761 high school biology students in 25 classes in southeastern USA, Lightburn and Fraser (2007) used the SLEI in an evaluation of the effectiveness of using anthropometry activities. Data analyses supported not only the SLEI's validity (in terms of factor structure, internal consistency reliability, and ability to differentiate between classrooms), but also suggested that there was a positive influence of using anthropometric activities in terms of both classroom climate and student attitudes.

WIHIC Questionnaire

According to Dorman (2008, p. 181), "the WIHIC has achieved almost bandwagon status in the assessment of classroom environments." The What Is Happening In This Class? (WIHIC) combines modified versions of the most salient scales from a wide range of existing questionnaires with additional scales that accommodate contemporary educational concerns (e.g., equity and constructivism). Also, the WIHIC has a separate Class form (which assesses a student's perceptions of the class as a whole) and Personal form (which assesses a student's personal perceptions of his or her role in a classroom).

Developed by Fraser et al. (1996), the original 90-item nine-scale version was refined by both statistical analysis of data from 355 junior high school science students, and extensive interviewing of students about their views of their classroom environments in general, the wording and salience of individual items, and their questionnaire responses. The field testing of a second version of the WIHIC involved junior high school science classes in Australia and Taiwan. While the Australian sample of 1081 students in 50 classes responded to the original English version, a Taiwanese sample of 1879 students in 50 classes responded to a Chinese version that had undergone careful procedures of translation and back translation (Aldridge et al., 1999). For both the Australian and Taiwanese samples, strong factorial validity and internal consistency reliability were reported.

A comprehensive validation of the WIHIC was conducted by Dorman (2003) using a cross-national sample of 3980 high school students from Australia, the UK, and Canada. Confirmatory factor analysis supported the seven-scale a priori structure, with fit statistics indicating a good fit of the model to the data. The use of multisample analyses within structural equation modeling substantiated invariant factor structures for the three grouping variables of country, grade level, and student sex. In a second study, Dorman (2008) used both the actual and preferred forms of the WIHIC with a sample of 978 secondary school students from Australia. Separate confirmatory factor analyses for the actual and preferred forms supported the seven-scale a priori structure, with fit statistics again indicating a good fit of the models to the data. The use of multitrait–multimethod modeling with the seven scales as traits and the two forms of the instrument as methods supported the WIHIC's construct validity.

The WIHIC has been used in various countries and in various languages. Cross-national research was conducted in Australia, the UK, and Canada in English by Dorman (2003), in Australia and Indonesia in two languages by Fraser et al. (2010a), and in Australia and Canada by Zandvliet and Fraser (2005). The WIHIC was used in English in Singapore by Chionh and Fraser (2009) and Khoo and Fraser (2008), in India by Koul and Fisher (2005), in Australia by Dorman (2008), and in South Africa by Aldridge et al. (2009). Other researchers used the WIHIC, respectively, in the Korean language in Korea (Kim et al., 2000) and in the Indonesian language in Indonesia (Wahyudi and Treagust, 2004). Other studies involved the use of an Arabic translation of the WIHIC in the UAE (Afari et al., 2013; MacLeod and Fraser, 2010).

Research that involved the use of the WIHIC in the USA includes three studies in California by den Brok et al. (2006b), Martin-Dunlop and Fraser (2008), and Ogbuehi and Fraser (2007); a study in New York by Wolf and Fraser (2008); and studies in Florida by Pickett and Fraser (2009), Allen and Fraser (2007), Robinson and Fraser (2013), and Helding and Fraser (2013). Although all the four studies in Florida involved the use of an English-language version of the WIHIC, it is noteworthy that three of them offered students the option of responding to a version of the WIHIC in either Spanish or in English.

Instruments Based on WIHIC: OBLEQ, TROFLEI, and COLES

Numerous researchers have incorporated WIHIC scales into specific-purpose questionnaires tailored to particular contexts and purposes. For example, working with a sample of 2638 grade 8 science students from 50 classes in 50 schools in the Limpopo Province of South Africa, Aldridge et al. (2006) developed and validated the Outcomes-Based Learning Environment Questionnaire (OBLEQ) in the Sepedi language for monitoring the implementation of outcomes-based classroom climates. As well as validating a widely applicable questionnaire suited for outcomes-based education, the researchers used case studies to support and check the accuracy of profiles of OBLEQ scores for specific classes.

Outcomes-focused education is an approach to school reform in which planning, delivery, and assessment all focus on the student's outcomes/results from teaching rather than on a syllabus or curriculum. Aldridge and Fraser (2008) conducted a study of an innovative new postsecondary school, whose emphases included an outcomes focus and the use of ICT in program delivery, during its first year of operation. As part of the formative and summative evaluations of this new school, they designed and used the Technology-Rich Outcomes-Focused Learning Environment Inventory (TROFLEI), which not only incorporates all of the WIHIC's seven scales, but also includes three other important scales that were salient in the context of this new school: Differentiation, Computer Usage, and Young Adult Ethos. The TROFLEI has eight items in each of the eight scales that are responded to using a five-point frequency scale (Almost Never, Seldom, Sometimes, Often, and Almost Always). An innovative aspect of the TROFLEI is that it employs a side-by-side response

format which enables students to provide their separate perceptions of actual and preferred classroom climate in an economical way.

Using a large sample of 2317 students from 166 grade 11 and 12 classes from Western Australia and Tasmania, Aldridge and Fraser (2008) reported strong factorial validity and internal consistency reliability for both the actual and preferred forms of the TROFLEI. Aldridge et al. (2004a) used multitrait–multimethod modeling with a subsample of 1249 students, of whom 772 were from Western Australia and 477 were from Tasmania (compared with 2317 students in our entire sample). When the 10 TROFLEI scales were used as traits and the actual and preferred forms of the instrument as methods, the results supported the TROFLEI's construct validity and sound psychometric properties, as well as indicating that the actual and preferred forms share a common structure.

When the TROFLEI was used in monitoring and evaluating the success of the new school in promoting outcomes-focused education, changes in students' perceptions of their classroom climates over four years supported the efficacy of the school's educational programs (Aldridge and Fraser, 2008). Using structural equation modeling with a sample of 4146 grade 8–13 students, Dorman and Fraser (2009) used the TROFLEI to establish associations between students' affective outcomes and their classroom climate perceptions. In an investigation of some determinants of classroom climate involving the use of TROFLEI with 2317 students, Aldridge and Fraser (2008) reported interesting differences in classroom climate perceptions between male and female students and between students enrolled in university-entrance examinations and in wholly school-assessed subjects. With the sample of 4146 students, Dorman et al. (2006) used cluster analysis with TROFLEI responses to identify five relatively homogeneous groups of classroom environments.

The Constructivist-Orientated Learning Environment Survey (COLES) incorporates numerous scales from the WIHIC into an instrument that is designed to provide feedback as a basis for reflection in teacher action research. In constructing the COLES, Aldridge et al. (2012) were especially conscious of the omission in all existing classroom environment questionnaires of important aspects related to the assessment of student learning. The COLES incorporates six of the WIHIC's seven scales (namely Student Cohesiveness, Teacher Support, Involvement, Task Orientation, Cooperation, and Equity). Like the TROFLEI, the COLES also includes the scales of Differentiation and Young Adult Ethos. In addition, the COLES includes the Personal Relevance scale from the CLES (the extent to which learning activities are relevant to the student's everyday out-of-school experiences). The two new COLES scales related to assessment are Formative Assessment (the extent to which students feel that the assessment tasks given to them make a positive contribution to their learning) and Assessment Criteria (the extent to which assessment criteria are explicit so that the basis for judgments is clear and public).

For a sample of 2043 grade 11 and 12 students from 147 classes in 9 schools in Western Australia, data analyses supported the sound factorial validity and internal consistency reliability of both actual and preferred versions of the COLES. A noteworthy methodological feature of this study was that the Rasch model was used to convert data collected using a frequency response scale into interval data suitable for parametric analyses. Interestingly, when analyses were performed separately for raw scores and Rasch scores, Aldridge et al. (2012) found that the differences between the validity results (e.g., reliability, discriminant validity, and ability to differentiate between classrooms) were negligible. During action research with teachers, feedback based on students' responses to both the actual and preferred versions of the COLES was used, in conjunction with reflective journals, written feedback, discussion at a forum, and teacher interviews. Aldridge et al. (2012) reported the experiences of these teachers concerning the viability of using feedback from the COLES as part of their action research aimed at improving their classroom climates.

Research Involving Educational Environment Instruments

The main types of past research considered in this section are (1) use of climate dimensions as criterion variables in the evaluation of educational innovations, (2) associations between student outcomes and classroom climate, (3) teachers' practical attempts to improve their classroom and school environments, and (4) other types of research with classroom climate instruments.

Evaluation of Educational Innovations

Classroom climate instruments can be useful as a source of process criteria in the evaluation of educational innovations. For example, by incorporating a classroom climate instrument within an evaluation of the use of a computerized database, Maor and Fraser (1996) found that students perceived that their classes became more inquiry oriented during the use of the innovation. In two studies in Singapore, classroom climate measures were used as dependent variables in evaluations of computer-assisted learning by Teh and Fraser (1994) and computer application courses for adults by Khoo and Fraser (2008).

Nix et al. (2005) used the CLES in their evaluation of an innovative science teacher development program. Programs were evaluated in terms of the types of school classroom climates created by these teachers as perceived by their 445 students in 25 classes. Students provided their perceptions of THIS classroom (the students' current class with the teacher who had experienced the professional development) and OTHER classroom (other classes at the same school taught by different teachers). Students of teachers who had experienced the professional development perceived their classrooms as having appreciably higher levels of the scales of Personal Relevance and Uncertainty relative to the comparison classes.

Martin-Dunlop and Fraser (2008) evaluated an innovative science course for prospective elementary teachers in a large urban university in California. When classroom climate scales selected from the WIHIC and SLEI were administered to 525 females in 27 classes, very large differences were found on all scales (over 1.5 standard deviations) between students' perceptions of the innovative course and that of their previous courses.

In a study of 761 high school biology students in southeastern USA, Lightburn and Fraser (2007) used the SLEI in an

evaluation of the effectiveness of using anthropometric activities. Relative to a comparison group, the anthropometry group had significantly higher scores on some SLEI and attitude scales.

Pickett and Fraser (2009) argued that the litmus test of the success of any teacher professional development program is the extent of changes in teaching behaviors and ultimately student outcomes in the participating teachers' school classrooms. Consequently, their evaluation of a two-year mentoring program in science for beginning elementary-school teachers drew on the field of classroom climate in gauging this program's success in terms of participants' classroom teaching behavior as assessed by their school students' perceptions of their classroom climates. The efficacy of the mentoring program was supported in terms of some improvements over time in the classroom climate.

In New York, Wolf and Fraser evaluated the effectiveness of using inquiry-based laboratory activities in terms of classroom climate, attitudes, and achievement. Administration of the WIHIC to 1434 middle-school science students in 71 classes supported the validity of the WIHIC and analyses for a subsample of students revealed that inquiry instruction promoted more Student Cohesiveness than did non-inquiry instruction (effect size of one-third of a standard deviation). As well inquiry-based instruction was differentially effective for male and female students.

Associations between Student Outcomes and Climate

Numerous research programs have shown that student perceptions account for appreciable amounts of variance in learning outcomes, often beyond that attributable to background student characteristics. For example, using the SLEI, associations with students' cognitive and affective outcomes have been established for a sample of approximately 80 senior high school chemistry classes in Australia (Fraser and McRobbie, 1995; McRobbie and Fraser, 1993), 489 senior high school biology students in Australia (Fisher et al., 1997), and 1592 grade 10 chemistry students in Singapore (Wong and Fraser, 1996). Using the QTI, associations between student outcomes and perceived patterns of teacher–student interaction were reported for samples of 489 senior high school biology students in Australia (Fisher et al., 1995), 1512 primary school mathematics students in Singapore (Goh et al., 1995), and 7484 grade 9 to 11 students from 278 classes in 55 public schools in 13 major Turkish cities (Telli et al., 2010).

While many past learning environment studies have employed techniques such as multiple regression analysis, few have used the multilevel analysis in order to take cognizance of the hierarchical nature of classroom settings (den Brok et al., 2006a). Two studies of outcome–environment associations compared the results obtained from multiple regression analysis with those obtained from an analysis involving the hierarchical linear model. In Wong et al.'s (1997) study involving 1592 grade 10 students in 56 chemistry classes in Singapore, associations were investigated between three student attitude measures and a modified version of the SLEI. In Goh et al.'s (1995) study with 1512 grade 5 mathematics students in 39 classes in Singapore, scores on a modified version of the MCI were related to student achievement and attitude. Most of the significant results from the multiple regression analyses were replicated in the HLM analyses, as well as being consistent in direction.

Dorman and Fraser (2009) investigated classroom climate, antecedent variables (gender, grade level, and home computer and Internet access), and student affective outcomes using the TROFLEI among 4146 high school students from Western Australia and Tasmania. The student outcome measures were attitude to the subject, attitude to computer use, and academic efficacy. Confirmatory factor analysis using LISREL supported the 10-scale a priori structure of the TROFLEI. When structural equation modeling using LISREL was used to test a postulated model involving antecedent variables, classroom environment and outcomes: improving classroom climate had the potential to improve student outcomes; antecedents did not have any significant direct effect on outcomes; and academic efficacy mediated the effect of several classroom climate dimensions on attitude to subject and attitude to computer use.

Teachers' Attempts to Improve Classroom Climate

Although much research has been conducted on educational climate, less has been done to help teachers to improve the climates of their own classrooms. Fraser and Fisher (1986) proposed five steps (assessment, feedback, reflection and discussion, intervention, and reassessment) for change attempts.

Yarrow et al. (1997) introduced 117 pre-service education teachers to the field of classroom climate through being involved in action research aimed at improving their university teacher education classes and their 117 primary school classes during teaching practice. Improvements in classroom climate were observed, and the pre-service teachers generally valued both the inclusion of the topic of classroom climate in their pre-service programs and the opportunity to be involved in action research aimed at improving classroom climates.

Two studies in South Africa employed this approach with teachers who used action research in an attempt to improve their classroom climates. In Aldridge et al.'s (2009) study, 31 in-service teachers undertaking a distance-education program administered a primary school version of the WIHIC in the Isi-Zulu language to 1077 grade 4–7 learners in the KwaZulu-Natal province. Different teachers were able to use feedback from the WIHIC with varying degrees of success in their attempts to improve their classroom climates. In Aldridge et al.'s (2004b) study, a group of 29 mathematics teachers administered the English version of the CLES to 1864 grade 4–9 learners in 43 classes. During an intervention phase in this study, some teachers were able to increase the constructivist orientation of their classrooms, thus supporting the efficacy of using the CLES to provide feedback to guide change.

Using the COLES, Aldridge et al. (2012) explored the viability of teachers using feedback based on their students' actual and preferred classroom climate perceptions for reflection in action research aimed at improving their classrooms. Reflective journals, written feedback, forum discussions, and teacher interviews also were used to provide feedback. Both actual and preferred forms of the COLES were administered on two occasions – first as a pretest prior to commencing the action research and as a posttest six weeks later after the implementation of classroom strategies aimed at reducing

actual–preferred discrepancies on selected COLES scales. Overall, teachers involved found that feedback information based on students' responses to the COLES prompted valuable reflection that led to implementing classroom changes that resulted in improvements in their classroom climate. An interesting feature of Aldridge et al.'s (2012) study was the use of the circular profiles as a means of providing each teacher with a comparison of mean actual and preferred responses to the COLES for his/her class.

Other Applications

Differences between students and teachers in their perceptions of the same actual classroom climate and of differences between the actual climate and that preferred by students or teachers were investigated by Fisher and Fraser (1983). Students preferred a more positive classroom climate than was actually present for all five ICEQ dimensions and teachers perceived a more positive classroom climate than did their students in the same classrooms on four of the ICEQ's dimensions.

Using both actual and preferred forms of educational climate instruments permits exploration of whether students achieve better when there is a higher similarity between the actual classroom climate and that preferred by students. Using a person–environment interaction framework, Fraser and Fisher (1983) reported that actual–preferred congruence (or person–environment fit) could be as important as the classroom climate per se in predicting student achievement of important affective and cognitive aims. The practical implication of these findings is that class achievement of certain outcomes might be enhanced by attempting to change the actual classroom climate in ways which make it more congruent with that preferred by the class.

Although most studies of educational climate have tended to focus on a single environment, there is potential in simultaneously considering the links between, and the joint influence of, two or more environments. In South Africa, Aldridge et al. (2011) used a school climate instrument with 50 secondary school science teachers from 50 different schools, together with a classroom climate questionnaire based on the WIHIC with the 2638 grade 8 students in the 50 classes of these 50 teachers. Although there emerged a small number of interesting specific relationships, overall, the school climate did not have a strong influence on what happens in classrooms. When Dorman et al. (1997) administered a classroom climate instrument to 2211 students in 104 classes and a school climate instrument to the 208 teachers of these classes, only weak associations between classroom climate and school climate were found. Although school rhetoric often would suggest that the school ethos would be transmitted to the classroom level, it appears that classrooms are somewhat insulated from the school as a whole.

Using secondary analysis of a large database from a Statewide Systemic Initiative in the USA, Fraser and Kahle (2007) examined the effects of several types of environments on student outcomes. Over three years, nearly 7000 students in 392 middle-school science and mathematics classes in 200 different schools responded to a questionnaire that assesses class, home, and peer environments as well as student attitudes. Students also completed an achievement measure. Rasch analyses allowed comparison across student cohorts and across schools. Findings confirmed the importance of extending research on classroom climates to include the learning climates of the home and the peer group. Although all three environments accounted for statistically significant amounts of unique variance in student attitudes, only the class climate (defined in terms of the frequency of use of standards-based teaching practices) accounted for statistically significant amounts of unique variance in student achievement scores.

The creation and empirical investigation of typologies of classroom climates has been pursued in only a handful of past studies. Working in Turkey with a Turkish translation of the WIHIC den Brok et al. (2010) created classroom climate profiles for a sample of 1474 high school biology students in 52 classes. The six distinct classroom profiles that emerged were self-directed learning; task-orientated cooperative learning; mainstream; task-orientated individualized; low-effective learning; and high-effective learning. The most common profile was the mainstream classroom for which all WIHIC scales had medium–high scores. Based on a sample of 4146 Australian students from 286 grade 8–13 classes, Dorman et al. (2006) used the 10-scale TROFLEI to develop a classroom typology. The five relatively homogeneous groups of classes that emerged were exemplary; safe and conservative; non-technological teacher centered; contested technological; and contested non-technological.

See also: Cross-Cultural Study of Education; Educational Evaluation: Overview; Environments for Education; Mathematics Education; Science Education.

Bibliography

Afari, E., Aldridge, J.M., Fraser, B.J., Khine, M.S., 2013. Students' perceptions of the learning environment and attitudes in game-based mathematics classrooms. Learning Environments Research 16, 131–150.

Aldridge, J.M., Dorman, J.P., Fraser, B.J., 2004a. Use of multitrait-multimethod modelling to validate actual and preferred forms of the Technology-Rich Outcomes-Focused Learning Inventory (TROFLEI). Australian Journal of Educational and Developmental Psychology 4, 110–125.

Aldridge, J.M., Fraser, B.J., 2008. Outcomes-Focused Learning Environments: Determinants and Effects (Advances in Learning Environments Research Series). Sense Publishers, Rotterdam, The Netherlands.

Aldridge, J.M., Fraser, B.J., Bell, L., Dorman, J.P., 2012. Using a new learning environment questionnaire for reflection in teacher action research. Journal of Science Teacher Education 23, 259–290.

Aldridge, J.M., Fraser, B.J., Huang, I.T.-C., 1999. Investigating classroom environments in Taiwan and Australia with multiple research methods. Journal of Educational Research 93, 48–62.

Aldridge, J.M., Fraser, B.J., Laugksch, R.C., 2011. Relationship between the school-level and classroom-level environment in secondary schools in South Africa. South African Journal of Education 31, 127–144.

Aldridge, J.M., Fraser, B.J., Ntuli, S., 2009. Utilising learning environment assessments to improve teaching practices among in-service teachers undertaking a distance education programme. South African Journal of Education 29, 147–170.

Aldridge, J.M., Fraser, B.J., Sebela, M.P., 2004b. Using teacher action research to promote constructivist learning environments in South Africa. South African Journal of Education 24, 245–253.

Aldridge, J.M., Fraser, B.J., Taylor, P.C., Chen, C.-C., 2000. Constructivist learning environments in a cross-national study in Taiwan and Australia. International Journal of Science Education 22, 37–55.

Aldridge, J.M., Laugksch, R.C., Seopa, M.A., Fraser, B.J., 2006. Development and validation of an instrument to monitor the implementation of outcomes-based learning environments in science classrooms in South Africa. International Journal of Science Education 28, 45–70.

Allen, D., Fraser, B.J., 2007. Parent and student perceptions of classroom learning environment and its association with student outcomes. Learning Environments Research 10, 67–82.
Brekelmans, M., Wubbels, Th, van Tartwijk, J., 2005. Teacher–student relationships across the teaching career. International Journal of Educational Research 43, 55–71.
den Brok, P., Brekelmans, M., Wubbels, Th, 2006a. Multilevel issues in research using students' perceptions of learning environments. The case of the questionnaire on teacher interaction. Learning Environments Research 9, 199–213.
den Brok, P., Fisher, D., Rickards, T., Bull, E., 2006b. Californian science students' perceptions of their classroom learning environments. Educational Research and Evaluation 12, 3–25.
den Brok, P., Telli, S., Cakiroglu, J., Taconis, R., Tekkaya, C., 2010. Learning environment profiles of Turkish secondary biology students. Learning Environments Research 13, 187–204.
Chionh, Y.H., Fraser, B.J., 2009. Classroom environment, achievement, attitudes and self-esteem in geography and mathematics in Singapore. International Research in Geographical and Environmental Education 18, 29–44.
Dorman, J.P., 2003. Cross-national validation of the What Is Happening In This Class? (WIHIC) questionnaire using confirmatory factor analysis. Learning Environments Research 6, 231–245.
Dorman, J.P., 2008. Use of multitrait-multimethod modelling to validate actual and preferred forms of the What Is Happening In This Class? (WIHIC) questionnaire. Learning Environments Research 11, 179–197.
Dorman, J.P., Aldridge, J.M., Fraser, B.J., 2006. Using students' assessment of classroom environment to develop a typology of secondary school classrooms. International Education Journal 7, 909–915.
Dorman, J.P., Fraser, B.J., 2009. Psychosocial environment and affective outcomes in technology-rich classrooms: testing a causal model. Social Psychology of Education 12, 77–99.
Dorman, J.P., Fraser, B.J., McRobbie, C.J., 1997. Relationship between school-level and classroom-level environments in secondary schools. Journal of Educational Administration 35, 74–91.
Fisher, D.L., Fraser, B.J., 1981. Validity and use of my class inventory. Science Education 65, 145–156.
Fisher, D.L., Fraser, B.J., 1983. A comparison of actual and preferred classroom environment as perceived by science teachers and students. Journal of Research in Science Teaching 20, 55–61.
Fisher, D.L., Henderson, D., Fraser, B.J., 1995. Interpersonal behaviour in senior high school biology classes. Research in Science Education 25, 125–133.
Fisher, D., Henderson, D., Fraser, B., 1997. Laboratory environments & student outcomes in senior high school biology. American Biology Teacher 59, 214–219.
Fisher, D.L., Khine, M.S. (Eds.), 2006. Contemporary Approaches to Research on Learning Environments: Worldviews. World Scientific, Singapore.
Fraser, B.J., 1986. Classroom Environment. Croom Helm, London.
Fraser, B.J., 1990. Individualised Classroom Environment Questionnaire. Australian Council for Educational Research, Melbourne, Australia.
Fraser, B.J., 2007. Classroom learning environments. In: Abell, S.K., Lederman, N.G. (Eds.), Handbook of Research on Science Education. Lawrence Erlbaum, Mahwah, NJ, pp. 103–124.
Fraser, B.J., Aldridge, J.M., Adolphe, F.S.G., 2010a. A cross-national study of secondary science classroom environments in Australia and Indonesia. Research in Science Education 40, 551–571.
Fraser, B.J., Aldridge, J.M., Soerjaningsih, W., 2010b. Instructor-student interpersonal interaction and student outcomes at the university level in Indonesia. The Open Education Journal 3, 32–44.
Fraser, B.J., Anderson, G.J., Walberg, H.J., 1982. Assessment of Learning Environments: Manual for Learning Environment Inventory (LEI) and My Class Inventory (MCI) (Third version). Western Australian Institute of Technology, Perth, Australia.
Fraser, B.J., Fisher, D.L., 1983. Use of actual and preferred classroom environment scales in person-environment fit research. Journal of Educational Psychology 75, 303–313.
Fraser, B.J., Fisher, D.L., 1986. Using short forms of classroom climate instruments to assess and improve classroom psychosocial environment. Journal of Research in Science Teaching 5, 387–413.
Fraser, B.J., Fisher, D.L., McRobbie, C.J., April 1996. Development, validation, and use of personal and class forms of a new classroom environment instrument. In: Paper Presented at the Annual Meeting of the American Educational Research Association, New York.
Fraser, B.J., Giddings, G.J., McRobbie, C.J., 1995. Evolution and validation of a personal form of an instrument for assessing science laboratory classroom environments. Journal of Research in Science Teaching 32, 399–422.
Fraser, B.J., Kahle, J.B., 2007. Classroom, home and peer environment influences on student outcomes in science and mathematics: an analysis of systemic reform data. International Journal of Science Education 29, 1891–1909.
Fraser, B.J., Lee, S.S.U., 2009. Science laboratory classroom environments in Korean high schools. Learning Environments Research 12, 67–84.
Fraser, B.J., McRobbie, C.J., 1995. Science laboratory classroom environments at schools and universities: a cross-national study. Educational Research and Evaluation 1, 289–317.
Fraser, B.J., McRobbie, C.J., Giddings, G.J., 1993. Development and cross-national validation of a laboratory classroom environment instrument for senior high school science. Science Education 77, 1–24.
Fraser, B.J., O'Brien, P., 1985. Student and teacher perceptions of the environment of elementary-school classrooms. Elementary School Journal 85, 567–580.
Fraser, B.J., Walberg, H.J. (Eds.), 1991. Educational Environments: Evaluation, Antecedents and Consequences. Pergamon, London.
Goh, S.C., Fraser, B.J., 1996. Validation of an elementary school version of the questionnaire on teacher interaction. Psychological Reports 79, 512–522.
Goh, S.C., Fraser, B.J., 1998. Teacher interpersonal behaviour, classroom environment and student outcomes in primary mathematics in Singapore. Learning Environments Research 1, 199–229.
Goh, S.C., Young, D.J., Fraser, B.J., 1995. Psychosocial climate and student outcomes in elementary mathematics classrooms: a multilevel analysis. Journal of Experimental Education 64, 29–40.
Helding, K.A., Fraser, B.J., 2013. Effectiveness of NBC (National Board Certified) teachers in terms of learning environment, attitudes and achievement among secondary school students. Learning Environments Research 16, 1–21.
Khoo, H.S., Fraser, B.J., 2008. Using classroom psychosocial environment in the evaluation of adult computer application courses in Singapore. Technology, Pedagogy and Education 17, 67–81.
Kim, H.B., Fisher, D.L., Fraser, B.J., 1999. Assessment and investigation of constructivist science learning environments in Korea. Research in Science and Technological Education 17, 239–249.
Kim, H.B., Fisher, D.L., Fraser, B.J., 2000. Classroom environment and teacher interpersonal behaviour in secondary science classes in Korea. Evaluation and Research in Education 14, 3–22.
Koul, R.B., Fisher, D.L., 2005. Cultural background and students' perceptions of science classroom learning environment and teacher interpersonal behaviour in Jammu, India. Learning Environments Research 8, 195–211.
Lee, S.S.U., Fraser, B.J., Fisher, D.L., 2003. Teacher-student interactions in Korean high school science classrooms. International Journal of Science and Mathematics Education 1, 67–85.
Lewin, K., 1936. Principles of Topological Psychology. McGraw, New York.
Lightburn, M.E., Fraser, B.J., 2007. Classroom environment and student outcomes among students using anthropometry activities in high school science. Research in Science and Technological Education 25, 153–166.
MacLeod, C., Fraser, B.J., 2010. Development, validation and application of a modified Arabic translation of the What Is Happening In This Class? (WIHIC) questionnaire. Learning Environments Research 13, 105–125.
Majeed, A., Fraser, B.J., Aldridge, J.M., 2002. Learning environment and its association with student satisfaction among mathematics students in Brunei Darussalam. Learning Environments Research 5, 203–226.
Maor, D., Fraser, B.J., 1996. Use of classroom environment perceptions in evaluating inquiry-based computer assisted learning. International Journal of Science Education 18, 401–421.
Martin-Dunlop, C., Fraser, B.J., 2008. Learning environment and attitudes associated with an innovative course designed for prospective elementary teachers. International Journal of Science and Mathematics Education 6, 163–190.
McRobbie, C.J., Fraser, B.J., 1993. Associations between student outcomes and psychosocial science environment. Journal of Educational Research 87, 78–85.
Mink, D.V., Fraser, B.J., 2005. Evaluation of a K–5 mathematics program which integrates children's literature: classroom environment and attitudes. International Journal of Science and Mathematics Education 3, 59–85.
Murray, H.A., 1938. Explorations in Personality. Oxford University Press, New York.
Nix, R.K., Fraser, B.J., 2010. Using computer-assisted teaching to promote constructivist practices in teacher education. In: Morris, B.A., Ferguson, G.M. (Eds.), Computer-Assisted Teaching: New Developments. Nova Science Publisher, New York, pp. 93–115.
Nix, R.K., Fraser, B.J., Ledbetter, C.E., 2005. Evaluating an integrated science learning environment using the constructivist learning environment survey. Learning Environments Research 8, 109–133.
Ogbuehi, P.I., Fraser, B.J., 2007. Learning environment, attitudes and conceptual development associated with innovative strategies in middle-school mathematics. Learning Environments Research 10, 101–114.

Peiro, M.M., Fraser, B.J., 2009. Assessment and investigation of science learning environments in the early childhood grades. In: Ortiz, M., Rubio, C. (Eds.), Educational Evaluation: 21st Century Issues and Challenges. Nova Science Publishers, New York, pp. 349–365.

Pickett, L.H., Fraser, B.J., 2009. Evaluation of a mentoring program for beginning teachers in terms of the learning environment and student outcomes in participants' school classrooms. In: Selkirk, A., Tichenor, M. (Eds.), Teacher Education: Policy, Practice and Research. Nova Science Publishers, New York, pp. 1–15.

Quek, C.L., Wong, A.F.L., Fraser, B.J., 2005. Student perceptions of chemistry laboratory learning environments, student-teacher interactions and attitudes in secondary school gifted education classes in Singapore. Research in Science Education 35, 299–321.

Robinson, E., Fraser, B.J., 2013. Kindergarten students' and parents' perceptions of science classroom environments: achievement and attitudes. Learning Environments Research 16, 151–167.

Scott, R.H., Fisher, D.L., 2004. Development, validation and application of a Malay translation of an elementary version of the Questionnaire on Teacher Interaction (QTI). Research in Science Education 34, 173–194.

Scott Houston, L., Fraser, B.J., Ledbetter, C.E., 2008. An evaluation of elementary school science kits in terms of classroom environment and student attitudes. Journal of Elementary Science Education 20, 29–47.

Sink, C.A., Spencer, L.R., 2005. My class inventory – short form as an accountability tool for elementary school counsellors to measure classroom climate. Professional School Counseling 9, 37–48.

Taylor, P.C., Fraser, B.J., Fisher, D.L., 1997. Monitoring constructivist classroom learning environments. International Journal of Educational Research 27, 293–302.

Teh, G., Fraser, B.J., 1994. An evaluation of computer-assisted learning in terms of achievement, attitudes and classroom environment. Evaluation and Research in Education 8, 147–161.

Telli, S., den Brok, P., Cakiroglu, J., 2010. The importance of teacher-student interpersonal relationships for Turkish students' attitudes towards science. Research in Science and Technological Education 28, 261–276.

Trickett, E.J., Moos, R.H., 1973. Social environment of junior high and high school classrooms. Journal of Educational Psychology 65, 93–102.

Wahyudi, Treagust, D.F., 2004. The status of science classroom learning environments in Indonesian lower secondary schools. Learning Environments Research 7, 43–63.

Walberg, H.J., Anderson, G.J., 1968. Classroom climate and individual learning. Journal of Educational Psychology 59, 414–419.

Wolf, S.J., Fraser, B.J., 2008. Learning environment, attitudes and achievement among middle-school science students using inquiry-based laboratory activities. Research in Science Education 38, 321–341.

Wong, A.L.F., Fraser, B.J., 1996. Environment-attitude associations in the chemistry laboratory classroom. Research in Science and Technological Education 14, 91–102.

Wong, A.F.L., Young, D.J., Fraser, B.J., 1997. A multilevel analysis of learning environments and student attitudes. Educational Psychology 17, 449–468.

Wubbels, Th, Brekelmans, M., 2012. Teacher–students relationships in the classroom. In: Fraser, B.J., Tobin, K.G., McRobbie, C.J. (Eds.), Second International Handbook of Science Education. Springer, New York, pp. 1241–1255.

Wubbels, Th, Levy, J. (Eds.), 1993. Do You Know What You Look Like: Interpersonal Relationships in Education. Falmer Press, London.

Yarrow, A., Millwater, J., Fraser, B., 1997. Improving university and primary school classroom environments through preservice teachers' action research. International Journal of Practical Experiences in Professional Education 1 (1), 68–93.

Zandvliet, D.B., Fraser, B.J., 2005. Physical and psychosocial environments associated with networked classrooms. Learning Environments Research 8, 1–17.

Classroom Level in Educational Effectiveness Research

Christopher A Hafen and Erik A Ruzek, University of Virginia, Charlottesville, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Valid measurement of effective classrooms requires a developmental approach to conceptualizing classroom processes. This article presents several theoretical models (e.g., teaching through interactions framework; connective instruction), which emphasize the importance of teacher–student interactions as a central driver for student learning and socioemotional development. Included in this discussion is research detailing how effective classrooms have teachers that engage students in developmentally appropriate ways, and what practices increase student engagement. The article ends with some discussion of professional development practices that might increase teacher effectiveness.

Understanding and describing an effective classroom across prekindergarten (pre-K) to grade 12 environments is challenging. How does one equate a high-quality classroom in kindergarten with a high-quality classroom in 10th grade? The answer, in part, comes from designating the classroom as a context, and placing the term 'effectiveness' in a broader developmental arena. In other words, what makes teachers effective, regardless of the grade level they teach, is the extent to which they offer opportunities for developmentally appropriate interactions for their students. This approach is strongly supported by decades of research on the power and importance of daily interactions as a driving force of child and adolescent growth and development (Bronfenbrenner and Morris, 1998). This point is oftentimes lost in the current educational setting, which rewards or punishes teachers and schools on the basis of student test scores, without considering the broader impact of a truly effective teacher. This article will discuss recent research on educational effectiveness with a particular focus on how developmentally appropriate interactions are pivotal for optimal student development.

Understanding the Goal of an Effective Classroom

While there is much debate on the degree to which standardized test scores and achievement data should be used for tracking student learning, there is little debate that the purpose of school is to aid in the social and intellectual learning of developing children and adolescents. That suggests that at any model for understanding an effective classroom should have as its foundation student engagement (Ames, 1992; Shernoff et al., 2003). Student engagement is the vehicle through which rich content is learned, positive developmental lessons take root, and general learning occurs. In other words, all roads toward an effective classroom travel through student engagement.

Focusing on whether or not a classroom is engaging to students should make understanding an effective classroom pretty straightforward. That is, if it were straightforward to measure and make sense of student engagement. However, there is much debate among researchers on how best to capture student engagement, as it is a multifaceted construct with behavioral, emotional, and persistence oriented markers (see Appleton et al., 2008 for review). Ultimately, what matters most for social and intellectual learning is the behavior a student endorses by either reporting that they are involved in classwork and engaged in a class or by actually physically demonstrating engagement. The specific interactions in a classroom that lead individual students to maintain engagement may differ (see Cooper, 2014 for illustrative examples), but in an effective classroom there needs to be clear and consistent evidence of engagement across the majority of students.

Assuming that student engagement is the key to an effective classroom leads to an obvious question, how does a teacher foster student engagement in the classroom? The answer may come in the form of the interactions a classroom provides. At the start of a school year, there are a wide range of student perceptions, abilities, and past experiences presenting a challenge to any teacher. Due to this variability, no single classroom approach will engage every student for every minute throughout the year. However, there are strong theoretical arguments that the most successful approaches focus efforts on connecting students emotionally and instructionally to the teacher, the content, and each other. This process has recently been coined connective instruction (Martin and Dowson, 2009).

One approach to studying connective instruction is provided by the Teaching through Interactions framework, which defines an effective classroom through the quality of three broad domains of interactions that students experience in the classroom (see **Table 1**). These domains include emotional, organizational, and instructional interactions. This structure is powerful as it allows stakeholders (researchers, educators, and parents) to use the same language to understand and discuss classroom quality, whether the classroom contains elementary-aged children or older students in middle or high school classrooms. For example, the Emotional Support domain of the framework includes the dimensions of Positive Climate, Teacher Sensitivity, and Regard for Student Perspectives. In a kindergarten classroom having high-quality Regard for Student Perspectives might involve creating more opportunities for students to move freely around the classroom, while having high quality Regard for Student Perspectives in a 10th grade classroom might entail creating opportunities for autonomy and leadership. Importantly, research on the

Table 1 Description of teaching through interactions framework

Domain	*Dimension*	*Description*
Emotional support	Positive climate	Reflects the emotional connection and relationships among teachers and students, and the warmth, respect, and enjoyment in verbal and nonverbal interactions.
	Negative climate	Reflects the level and intensity of expressed negativity among teachers and students in the classroom.
	Teacher sensitivity	Encompasses the teacher's awareness and responsiveness to the academic, social-emotional, and developmental needs of individual students and the entire class.
	Regard for student perspectives	Focuses on the extent to which the teacher is able to meet and capitalize on the social and developmental needs of adolescents by providing opportunities for student autonomy and leadership. Also considered are the extent to which student ideas and opinions are valued and content is made useful and relevant to adolescents.
Classroom organization	Behavior management	Encompasses the teacher's use of clear behavioral expectations and effective methods to prevent and redirect misbehavior.
	Productivity	Considers how well the teacher manages time and routines so that instructional time is maximized.
	Instructional learning formats	Focuses on the ways in which the teacher maximizes students' interest and engagement in learning. This includes teachers' use of interesting and engaging lessons and materials, active facilitation, and clarity of learning objectives.
Instructional support	Content understanding	Refers to both depth of the lesson content and the approaches used to help students comprehend the framework, key ideas, and procedures in an academic discipline. At a high level, this refers to interactions among the teacher and students that lead to an integrated understanding of facts, skills, concepts, and principles.
	Analysis and inquiry	Assesses the degree to which the teacher facilitates students' use of higher-level thinking skills, such as analysis, problem solving, reasoning, and creation through the application of knowledge and skills. Opportunities for demonstrating metacognition, i.e., thinking about thinking, are also included.
	Quality of feedback	Assesses the degree to which feedback expands and extends learning and understanding and encourages student participation. (At the secondary level, significant feedback may be provided by peers.)
	Instructional dialogue	Captures the purposeful use of dialogue-structured, cumulative questioning and discussion which guide and prompt students to facilitate students' understanding of content and language development. The extent to which these dialogues are distributed across all students in the class and across the class period is important to this rating.

Teaching through Interactions framework has yielded evidence for the consistency and utility of the framework across levels of the education system (Hamre et al., 2013; Hafen et al., in press).

This still leaves the question of what do effective teachers do that allows them to engage students? In the emotional domain, effective teachers demonstrate trust, respect, and warmth toward their students, along with a promotion of a sense of belonging by showing a genuine interest in students' lives. This creates an atmosphere in the classroom such that students see their teacher as willing to help, fair, knowledgeable, and easy to approach when issues arise socially or emotionally. In the organizational domain, effective teachers manage behavior proactively and maximize learning time by utilizing clear routines and quick transitions. This creates an atmosphere in the classroom where students take ownership and share control of the classroom. In the instructional domain, effective teachers provide clear learning targets, incorporate variety and novelty into their instruction, provide relevant and varied examples, encourage metacognitive processes, and provide timely feedback to scaffold student learning. This creates an atmosphere in the classroom in which students engage in higher-order thinking using novel concepts, yet feel comfortable asking questions and voicing their thought processes when learning new content. These processes can be studied at multiple levels (teacher, observer, student), each with its own insight toward understanding classroom effectiveness.

Understanding the Classroom from a Teacher Perspective

The simplest way to understand the environment of the classroom is to ask the teacher. They are able to report on what they do, the interactions they create, and the extent to which their students are engaged in the classroom. The key is how accurate are they in reporting on these elements. There is ample evidence to suggest that teachers are much more accurate in describing quantity information (how often they use a strategy) than quality information (how well its strategies are implemented). This is particularly true when teachers are asked to report on more specific information like their interactions with students, and their students' perceptions of engagement. Typically, correlations between teacher and student reports of engagement are very low. However, there is evidence to suggest that aspects of teacher report, particularly those related to connection with students, have an influence over student reports of engagement and enjoyment in the classroom (see Skinner and Belmont, 1993).

There is also the question of how does the impact of capturing the teacher's perspective change depending upon the age of the students. In surveying the participants in a K-3 classroom, the indicators of an effective classroom may only be reliably captured by asking the teacher. Students at that age are likely to only be able to accurately report on 'do you like your class/teacher' rather than different features or aspects of the classroom environment. Thus, capturing the dynamics of an effective classroom in pre-K or early elementary grades should involve a combination of teacher report and as much observation as time/cost allows, while the utility of teacher reports may wane as students get older.

Understanding the Classroom from an Observer Perspective

A focus on classroom interactions is the centerpiece of observational measures that are strongly linked to the positive development of students. Though observations of classrooms are more costly than teacher or student report, they offer an independent view into what makes a classroom environment effective. With the rise in available technology over the last decade, video-recorded observations can be used to visually show teachers the behaviors and interactions that are most strongly linked to increases in student engagement and learning. Two systems that have recently seen vast implementation in pre-K-12 settings are the Classroom Assessment Scoring System (CLASS; Pianta et al., 2010) and the Framework for Teaching (FFT; Danielson, 2007).

The CLASS is a tool for observing and rating the interaction quality among teachers and children in classrooms. Following the Teaching through Interactions framework, it measures teachers' emotional, organizational, and instructional supports that research indicates contribute to children's social development and academic achievement. The tool is used to assess interactions between teachers and children for a variety of purposes, including teacher professional development, monitoring and evaluation, and research. The CLASS focuses on interactional processes, and captures *how* materials and curriculum are implemented. The CLASS is based on both developmental and educational theory. It evolved from instruments developed and used in the large-scale National Institute of Child Health and Human Development Study of Early Child Care, including the Observational Record of the Caregiving Environment and the Classroom Observation System. The CLASS has since been validated by over 10 years of research in educational settings, directly connecting teacher–child interactions with growth in children's academic and social skills. An early version of the CLASS was used in the National Center for Early Development and Learning Multi-State Pre-K Study and Study of Statewide Education Programs. These two studies evaluated over 700 state-funded pre-K classrooms in 11 states, and included classroom observations; child assessments; and surveys of teachers, parents, and administrators (e.g., Pianta et al., 2005). The extensive use of the CLASS in these studies and others has led to further refinements in the tool's development and construction.

While the CLASS was initially developed for use in pre-K classrooms, it has since been expanded for use in classrooms for both older- and younger-aged children. The majority of studies on the reliability and validity of CLASS come from its roots in pre-K settings. Widely employed in research and evaluation studies, the CLASS has been affirmed as a reliable and valid measure through use in over 3000 classrooms in Head Start, child care, and pre-K settings (both center- and school-based). However, recent studies including the large Measures of Effective Teaching (MET; Kane et al., 2012) study funded by the Bill and Melinda Gates Foundation have tested the reliability and validity of CLASS in K-12 classrooms. There is now considerable evidence that the dimensions of the CLASS, particularly those tapping emotional and instructional interactions, are consistently linked to the social, emotional, and academic development of students (see Hamre and Pianta, 2010 for review).

The FFT was also included in the MET study and has been implemented as both a research and an evaluative tool (see Danielson, 2011 for review). The FFT is a research-based set of components of instruction, grounded in a constructivist view of learning and teaching. It is organized as a holistic approach to the school setting, where the complex activity of teaching is divided into 22 components (and 76 smaller elements) clustered into four domains of teaching responsibility: (1) planning and preparation; (2) classroom environment; (3) instruction; and (4) professional responsibilities. The structure of the classroom environment and instruction domains is very similar to the CLASS in that it focuses on behaviors that promote a positive and respectful climate, effectively managing student behavior, and consistently using questioning and discussion to promote engagement in the classroom.

Understanding the Classroom from a Student Perspective

One of the most prevalent sources of information about classrooms in educational research comes from student reports. Indeed, one might argue that the best source of information about the effectiveness of a classroom environment comes from the individual students in that classroom (Weinstein, 1983; Ames, 1992). However, caution should always be used when analyzing student reports as they are prone to false interpretation. Particularly since as children get older, their reports of the classroom (at least those taken early in the school year) may be more strongly influenced by previous experiences. For example, one study found that average classroom engagement levels in the beginning of the school year (taken from student reports of engagement) were negatively correlated with observations of classroom engagement during the same period ($r = -0.19$), whereas those same correlations measured at the end of the school year indicated much greater agreement ($r = 0.39$; Hafen et al., 2012). The bulk of the fluctuation in these correlations is due to variation in how students within the same classroom view the effectiveness of classroom interactions. Despite this, student reports of the classroom have indicated several markers through which healthy development occurs.

Once students reach the upper elementary grades, they are more capable of reliably reporting on their experiences in the classroom. In other words, older students are more likely to

agree on the relative degree to which they experience certain teacher practices (although there is rarely perfect agreement; see Miller and Murdock, 2007 for a discussion of the reliability of student reports). This provides an excellent opportunity to understand an effective classroom from the student perspective. Approaching classroom quality from the premise that an effective classroom will meet the developmental goals of its students, self-determination theory provides a useful developmental framework to conceptualize a student's self-motivation and well-being (Ryan and Deci, 2000). Self-determination theory includes three basic needs: competence, relatedness/connection, and autonomy (Reis et al., 2000). The need for competence is satisfied by the feeling that one can successfully produce desired effects and outcomes. The need for relatedness is derived from the experience of feeling close and connected to others one considers significant. The need for autonomy involves the perception that one's activities and environment are matched and endorsed with the self. Each of these three needs outlined in self-determination theory have been linked to various indicators of achievement and adjustment inside and outside the classroom (Deci et al., 2001; Denissen et al., 2007; Roeser et al., 1998).

The conceptualization of student experiences in the three core areas of self-determination theory is similar to the Teaching through Interactions framework in that the actions of the teacher could reasonably translate into perceptions by the student: higher levels of emotional support from a teacher would likely result in higher student reports of relatedness and connection; higher levels of organizational support would presumably lead to higher student reports of autonomy; and higher levels of instructional support should lead to students reporting more competence over time. This process of connective instruction should be the core of any measure of classroom quality, and the most effective strategy for assessing the effectiveness of the classroom environment is to look for overlap between multiple perspectives: teacher, student, and observer.

Understanding the Classroom from Multiple Perspectives

As detailed above, studies of the effective components of the classroom have yielded support for frameworks like Teaching through Interactions and self-determination theory because they highlight the importance of eliciting a student's emotional connection to the *teacher*, the *content*, and the *instruction*. The process of eliciting student connections, also called connective instruction, emphasizes the importance of these three facets in understanding and predicting student engagement (Martin and Dowson, 2009). Forming an emotional connection to the teacher occurs through interactions in which the teacher affirms and displays a genuine interest in the student's perspective and gets to know each student as an individual. Forming an emotional connection to the content involves opportunities to work with material that is relevant to a student's life and that has meaningful purpose. Forming an emotional connection to instruction is characterized by opportunities to learn from mistakes and develop a sense of competence. This framework aligns with the Teaching through Interactions framework, as they both posit that an effective classroom environment is created by the interactions teachers have with students along emotional, organizational, and instructional domains.

A recent study extends this theoretical argument by positing that identity development sits at the center of the connective instruction model of effective classrooms (Cooper, 2014). In this study, a mixed-methods approach was employed collecting data from 581 different classrooms totaling 1132 of the school's 1420 students through a self-report survey, and then following up with case studies of five classes that were intended to be representative and "provide insight into how the various types of teaching practices related to engagement" (p. 12). Results suggested quite strongly that connective instruction practices emphasizing the uniqueness of students by connecting understanding, affirmation, and humor with expressions of relevance and self-expression were most strongly associated with student engagement. This study highlights the important point that emotional engagement achieved through positive affect and emotional connections with teachers are only predictive of greater learning when paired with high-quality instruction that pushes higher-order thinking (Marzano and Pickering, 2011).

Impacting the Effectiveness of the Classroom with Professional Development

In order to enhance students' experiences in school, it is necessary to not just measure but also to improve interactions between teachers and students. However, improvement is often difficult, and independent reviews of professional development often find little to no evidence of validated and consistent approaches to improving teacher–student interactions (see http://ies.ed.gov/ncee/wwc/PDF/WhtPapers/wwcreviewprocess.pdf). Increasing evidence from the research literature suggests three features are vital to improving teacher–student interactions: (1) all individuals in a school must have a common vision and shared goals about how to create an effective classroom; (2) there needs to be a reliable and valid mechanism in place to measure classrooms that aligns with the shared goals within a school; and (3) teachers and other school personnel need access to professional development experiences that are effective in making progress toward these goals. A well-defined classroom observation system can assist in addressing all of these factors.

To the first feature, that of a shared vision and goals for creating effective classrooms, we know that in a typical school there are probably as many definitions of effective practice as there are professionals in the building. Without a shared vision of effective practice, communication between teachers and administrators and among teachers is likely to be unproductive. A reliable and validated observational protocol provides school personnel with a shared definition of effective practice and a common method for looking for these practices in the classroom. Importantly, by using a reliable and validated tool (i.e., one that has shown capacity for consistent use across observers and across time, as well as a relation to desired student outcomes), teachers and administrators can have confidence that their shared definition of quality practice is

focusing them on the types of classroom interactions that really make a difference for students.

With regard to the second feature, which is the need to have an appropriate measure, the explicit purpose of a classroom observation system is to assess the standing of classrooms in relation to the types of classroom interactions deemed relevant for student success. Using reliable and valid tools to observe and rate classrooms provides a research-based mechanism that can facilitate systematic increases in educational quality. Direct assessments of teaching practices based on reliable and valid metrics combined with commonly utilized assessments of effective teaching (such as student standardized test scores, or teacher value-added metrics) can create a robust basis for accountability efforts. Taken together, observational and value-added metrics can provide data that directly inform individualized and targeted programs of professional development, discussed below.

The third feature that facilitates improvement in teaching practice is access to professional development that offers ongoing training, support, and feedback. Identifying areas of strengths and challenges based on classroom observation scores can provide information about school-level instructional weaknesses that can be targeted for intensive in-service trainings. Furthermore, profiles of the strengths and areas of challenge for individual teachers provide a useful starting point for individualized professional development plans that target improvements in specific areas of teaching practice through the provision of support and constructive challenge.

The frameworks presented (Teaching through Interactions, self-determination theory, connective instruction) in this article indicate that an effective model of professional development for teachers must include a focus on interactions. One such model that has received empirical support for improving the classroom environment and ultimately student achievement is MyTeachingPartner (MTP). MTP targets the quality of teacher–student interactions. The program is designed to guide teachers in creating emotionally positive, motivating, and cognitively challenging classrooms characterized by sensitivity to individual students' socioemotional and academic needs (see Pianta et al., 2008). The additional value of this model is that it has been successfully implemented with teachers from pre-K to 12th grade. A key focus of MTP is that it offers teachers *ongoing, personalized* coaching and feedback that is grounded in a common instructional framework. MTP was originally designed to accompany the CLASS framework, though recent iterations have utilized frameworks that are idiosyncratic and district-specific. The model uses video from the classroom, and thus the focus is always on specific interactions that impact students in the classroom.

Research evidence suggests that MTP is successful at producing desired outcomes across grade levels. Early work demonstrated effectiveness in preschool programs (Mashburn et al., 2008; Pianta et al., 2008). More recent work has demonstrated efficacy in middle and high school, and the ability to deliver MTP consistently at scale (LoCasale-Crouch et al., 2011). In a randomized control trial of MTP conducted with 78 middle and high school teachers and over 1400 of their students, teachers who participated in the MTP condition had significantly better interactions as measured by the CLASS, controlling for where they were at the beginning of the year (Allen et al., 2011). Critically, end-of-course, standardized state exam scores for students whose teachers were in the MTP condition were higher than those for students in control classrooms 1 year after the intervention. This significant difference equated to an average increase in student achievement from the 50th to the 59th percentile for a student moved from the control condition to the MTP condition. MTP participation was also associated with increases in observed student engagement (Gregory et al., 2014) and positive peer interactions (Mikami et al., 2011). This evidence provides promise in that teachers are able to improve classroom effectiveness when provided specific, ongoing, and relevant feedback using a consistent framework.

Conclusions

The most effective classrooms have teachers that recognize the developmental needs of their students and structure learning opportunities accordingly. This is not typically taken into account in the current educational system, as teachers are typically pressured to improve their instruction around content for standardized tests rather than being given opportunities to learn how to deliver content in ways that will engage and motivate their students. A renewed focus on developmentally appropriate interactions (e.g., increased autonomy for adolescent students) has the power to yield increases not only on student learning and achievement, but also on students' emotional and social development. Through this focus on classroom interactions, it is clear that effective classrooms are those that help students develop into effective people.

See also: Classroom Climate; Education and Learning: Lifespan Perspectives; Learning Theories and Educational Paradigms; Metacognitive Development: Educational Implications; Piaget's Theory of Human Development and Education; School Effectiveness Research.

Bibliography

Allen, J., Pianta, R.C., Gregory, A., Mikami, A.Y., Lun, J., 2011. An interaction-based approach to enhancing secondary school instruction and student achievement. Science 333, 1034–1037.

Ames, C., 1992. Classrooms: goals, structures, and student motivation. Journal of Educational Psychology 84, 261–271.

Appleton, J.J., Christenson, S.L., Furlong, M.J., 2008. Student engagement with school: critical conceptual and methodological issues of the construct. Psychology in the Schools 45, 369–386.

Bronfenbrenner, U., Morris, P.A., 1998. The ecology of developmental processes. In: Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology, Theoretical Models of Human Development, fifth ed., vol. 1. John Wiley & Sons, New York, pp. 993–1029.

Cooper, K.S., 2014. Eliciting engagement in the high school classroom: a mixed-methods examination of teaching practices. American Educational Research Journal 51, 363–402.

Danielson, C., 2007. Enhancing Professional Practice. Association for Supervision and Curriculum Development, Alexandria, VA.

Danielson, C., 2011. Evaluations that help teachers learn. Educational Leadership 4, 35–40.

Deci, E.L., Ryan, R.M., Gagné, M., Leone, D.R., Usunov, J., Kornazheva, B.P., 2001. Need satisfaction, motivation, and well-being in the work organizations of a former eastern bloc country: a cross-cultural study of self-determination. Personality and Social Psychology Bulletin 27, 930–942.

Denissen, J.J.A., Zarrett, N.R., Eccles, J.S., 2007. I like to do it, I'm able, and I know I am: longitudinal couplings between domain-specific achievement, self-concept, and interest. Child Development 78, 430–447.

Gregory, A., Allen, J.P., Mikami, A.Y., Hafen, C.A., Pianta, R.C., 2014. Effects of a professional development program on behavioral engagement of students in middle and high school. Psychology in the Schools 51, 143–163.

Hafen, C.A., Hamre, B.K., Allen, J.P., Bell, C.A., Gitomer, D.H., Pianta, R.C. Teaching through interactions in secondary school classrooms: revisiting the factor structure and practical application of the classroom assessment scoring system-secondary. Journal of Early Adolescence. http://dx.doi.org/10.1177/0272431614537117, in press.

Hafen, C.A., Allen, J.P., Mikami, A.Y., Gregory, A., Hamre, B., Pianta, R.C., 2012. The pivotal role of adolescent autonomy in secondary school classrooms. Journal of Youth and Adolescence 41, 245–255.

Hamre, B.K., Pianta, R.C., 2010. Classroom environments and developmental processes: conceptualization and measurement. In: Meece, J., Eccles, J. (Eds.), Handbook of Research on Schools, Schooling, and Human Development. Routledge, New York, pp. 25–41.

Hamre, B.K., Pianta, R.C., Downer, J.T., DeCoster, J., Mashburn, A.J., Hamagami, A., 2013. Teaching through interactions: testing a developmental framework of teacher effectiveness in over 4,000 classrooms. The Elementary School Journal 113, 461–487.

Kane, T., Staiger, D., McCaffrey, D., Cantrell, S., Archer, J., Buhayar, S., Parker, D., 2012. Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains (Technical Report). Bill & Melinda Gates Foundation, Seattle, WA. Measures of Effective Teaching Project.

LoCasale-Crouch, J., Kraft-Sayre, M., Pianta, R.C., Hamre, B.K., Downer, J.T., Leach, A., Scott-Little, C., 2011. Implementing an early childhood professional development course across 10 sites and 15 sections: lessons learned. NHSA Dialog: Research-to-Practice Journal for the Early Childhood Field 14, 275–292.

Martin, A.J., Dowson, M., 2009. Interpersonal relationships, motivation, engagement, and achievement: yields for theory, current issues, and educational practice. Review of Educational Research 79, 327–365.

Marzano, R., Pickering, D., 2011. The Highly Engaged Classroom. Marzano Research Laboratory, Bloomington, IN.

Mashburn, A.J., Pianta, R.C., Hamre, B.K., Downer, J.T., Barbarin, O., Bryant, D., et al., 2008. Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. Child Development 79 (3), 732–749.

Mikami, A.Y., Gregory, A., Allen, J.P., Pianta, R.C., Lun, J., 2011. Effects of a teacher professional development intervention on peer relationships in secondary classrooms. School Psychology Review 40, 367–385.

Miller, A.D., Murdock, T.B., 2007. Modeling latent true scores to determine the utility of aggregate student perceptions as classroom indicators in HLM: the case of classroom goal structures. Contemporary Educational Psychology 32, 83–104.

Pianta, R.C., Hamre, B., Mintz, S., 2010. Classroom Assessment Scoring System – Secondary (CLASS-S). University of Virginia.

Pianta, R.C., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, C., et al., 2005. Features of pre-kindergarten programs, classrooms, and teachers: do they predict observed classroom quality and child-teacher interactions? Applied Developmental Science 9, 144–159.

Pianta, R.C., Mashburn, A.J., Downer, J.T., Hamre, B.K., Justice, L., 2008. Effects of web-mediated professional development resources on teacher-child interactions in pre-kindergarten classrooms. Early Childhood Research Quarterly 23, 431–451.

Reis, H.T., Sheldon, K.M., Gable, S.L., Roscoe, J., Ryan, R.M., 2000. Daily well-being: the role of autonomy, competence, and relatedness. Personality and Social Psychology Bulletin 26, 419–435.

Roeser, R.W., Eccles, J.S., Sameroff, A.J., 1998. Academic and emotional functioning in early adolescence: longitudinal relations, patterns, and prediction by experience in middle school. Development and Psychopathology 10, 321–352.

Ryan, R.M., Deci, E.L., 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist 55, 68–78.

Shernoff, D.J., Csikszentmihalyi, M., Schneider, B., Shernoff, E.S., 2003. Student engagement in high school classrooms from the perspective of flow theory. School Psychology Quarterly 18, 158–176.

Skinner, E.A., Belmont, M.J., 1993. Motivation in the classroom: reciprocal effects of teacher behavior and student engagement across the school year. Journal of Educational Psychology 85, 571–581.

Weinstein, R.S., 1983. Student perceptions of schooling. The Elementary School Journal 83, 287–312.

What Works Clearinghouse, 2008. WWC Review Process. National Center for Education Statistics, Institute for Education Sciences, Washington, DC. Retrieved from: http://ies.ed.gov/ncee/wwc/PDF/WhtPapers/wwcreviewprocess.pdf (14.06.11.).

Climate Change

Ian Bailey, School of Geography, Earth and Environmental Sciences, Plymouth University, Plymouth, UK
Piers Revell, School of Government, Plymouth University, Plymouth, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Climate change has become an important social and policy concern in recent years, chiefly as a result of scientific warnings about the potential effects of anthropogenic climate forcing on ecological systems, human health, water supplies, agriculture and land-use, economic performance, and security. Three main themes in social and behavioral scientific research on climate change are surveyed: the challenges of communicating the complexities and uncertainties of climate science to political and public audiences; developments in the governance of climate change at the international, national, and transnational levels; and individual perceptions of, and responses to, climate change.

Introduction

Climate change is recognized as a major environmental problem by a majority of the international scientific community. According to the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (IPCC, 2007: 2): "global atmospheric concentrations of carbon dioxide, methane and nitrous oxide have increased markedly as a result of human activities since 1750 and now far exceed pre-industrial values." The report further suggests that most of the observed increase in global average temperatures since the mid-twentieth century is very likely due to rising anthropogenic greenhouse gas concentrations and that, if not controlled, climate effects such as rising sea levels, disruption to weather patterns, and ocean acidification pose serious to human health, water supplies, agricultural systems, economic performance, and global security. Such projections have triggered calls for prompt and coordinated action to reduce greenhouse gas emissions and adapt to changes in the climate system (Stern, 2007).

Research on climate change in the social and environmental sciences in many ways mirrors these concerns. Three main themes emerge within this body of research and form the mainstay of this article. The first concerns the communication of climate science to political, business, and public audiences. Although human-induced climate change is almost universally recognized by scientists, it is not a discrete or easily observable entity; rather, it is a social construct used to represent an array of phenomena, cause-and-effect relationships, and projections about the future. A key focus for social scientists has thus been how the technical and uncertain science of climate change is being translated and conveyed to nonspecialist audiences and how different messages and communication methods influence how climate change is comprehended and viewed.

A second theme concerns the governance of climate change. The long-term and potentially far-reaching effects of climate change suggest the need for a global-level response but, so far, a strong international deal to reduce greenhouse gas emissions has proven elusive. The reasons for this have been an obvious preoccupation, as has analysis of climate governance at the national level. Since the late 1990s, however, state-focused analyses have been complemented by studies examining the involvement of subnational authorities, private corporations, and civil society groups in climate governance, drawing from other work on the governance turn in the social sciences. Third, disciplines such as behavioral psychology have explored how individuals comprehend and respond to climate change. Building on general models of environmental behavior developed from the 1970s onward, this work has also investigated opportunities and barriers to integrating 'climate-friendly' behaviors, such as reducing energy consumption and low-carbon transportation, into everyday life practices.

Communicating Climate Change

Scientific Reporting

One aspect of climate-change communication that has received considerable research interest focuses on the role of scientific authority and consensus in supporting the construction of an effective global response to climate change. The theory of epistemic communities upon which much of this research is based rests on the observation that environmental problems, like many other global problems confronting contemporary leaderships, are highly complex and their effects extend beyond the issue area to which they ostensibly relate. In this particular case, climate change is not just an environmental policy problem; it has far-reaching implications for policies affecting energy production, transport, manufacturing, water supplies, and land-use and agriculture. This complexity breeds uncertainty in the mind of policy-makers on questions of risk and how to define national interests in the formulation of foreign and domestic policy. Under such conditions, politicians and their bureaucracies have strong incentives to seek advice from recognized bodies of expertise held in common by *epistemic communities*.

An epistemic community is defined as a transnationally organized but nationally embedded network of professionals that possess an authoritative claim to policy-relevant knowledge within their domain of expertise. In order for reporting to governments and policy elites to be effective in building successful environmental regimes, epistemic communities need to have four main qualities (Haas, 1992):

1. a shared set of normative and principled beliefs;
2. shared causal beliefs derived from an analysis of practices that have led to an environmental problem or set of problems;

3. shared notions of validity based upon internally defined criteria for weighing and validating authoritative knowledge; and
4. a common policy enterprise prompted by conclusions arising from the above.

These qualities have been observed in numerous studies of relatively successful attempts at environmental governance, such as the Mediterranean Action Plan and the Ozone Layer Protection regime. Quite naturally, it was hoped that such examples might be reproducible in other areas of global environmental governance.

Haas' model can help explain why most of the world's governments took scientific reporting of climate change seriously and signed the United Nations Framework Convention on Climate Change (UNFCCC), developed as part of the Rio Earth Summit held in 1992. The discipline of meteorology established the preconditions for a climate epistemic community in that it developed norms and institutions for international scientific cooperation well before questions of anthropogenic influence upon the climate system were discussed. The common interest among European maritime states in understanding climate patterns further enabled meteorology to develop norms and institutions for sharing, validating, and weighing knowledge. Arrhenius theorized the possibility of human activity as a cause of climate change as early as 1896, but it was the epistemic community that had developed norms for collecting, sharing, and evaluating meteorological knowledge that was responsible for verifying climate data, recognizing climate change as a problem, and bringing it to the attention of the international community at the World Meteorological Conference in Toronto in 1988 and subsequent international meetings.

The epistemic community model became less compelling once formal international climate negotiations started under the UNFCCC (Paterson, 1996). In some cases an epistemic consensus may diagnose an environmental problem that can be solved through technocratic means. In other cases environmental problems appear to cross a threshold of complexity where the epistemic community founders upon normative and political issues. One way to illustrate this with regard to climate change is to examine the division of labors between the three main working groups of the IPCC, the international scientific body appointed by governments to investigate and synthesize scientific research on climate change.

With painstaking work, and despite well-financed opposition from some sections of the fossil-fuel lobby, Working Group I has delivered a series of scientific consensus reports on the scientific basis of climate change and the need to mitigate its anthropogenic causes in 1991, 1997, 2001, 2007, and 2013. These reports have been an important factor in persuading national politicians and bureaucracies in signing up to a 'motherhood' convention for stabilizing the growth in concentration of greenhouse gases, the UNFCCC. It has proven more difficult, however, for working groups to secure consensus at high levels of scientific certainty with regard to forecasting the differential impacts of climate change upon human societies living across different regions of the world. This is a key issue for international negotiations because the lack of certainty creates high levels of difficulty for securing a political agreement on issues such as whether developing countries that bear the least responsibility for historical emissions of greenhouse gases should attempt to curb their emissions or whether developed countries should take a policy lead or compensate developing countries for the effects of climate change.

It has proven even harder for Working Group III to draft a convincing report that conveys 'a common policy enterprise' on appropriate responses to mitigate and adapt to climate change. The main reason for this is that developing policy recommendations requires the mediation of many ontological positions (e.g., by what measures can the effectiveness of a policy be measured?), ideological tensions (e.g., strong government intervention or more market-based approaches), and political conflicts (e.g., when policy responses impose costs of citizens and industry) that are highly impermeable to resolution using positivist methods. Particularly but not exclusively in relation to offering policy solutions, there are limits beyond which a climate epistemic consensus can fragment into a plurality of *communities*. Working Group III has often attracted criticism for the timidity of its proposed policy options, and for creating a tacit boundary line that excludes some more radical proposals for reforming institutions and the structure of international political and economic governance from consideration by governments (O'Riordan and Jordan, 1997).

Nongovernment Organization Campaigns

A second area of communicating climate change focuses on nongovernment organization's (NGO) involvement in communicating climate change to public audiences as part of efforts to build social movements for change. Some research on this topic comes with an explicit or implicit critique of epistemic or 'policy network' approaches alleging that they overestimate the role of experts and understate that of environmental movements (Toke, 1999). According to this line of reasoning, scientists and policy experts can produce reports for governments but real progress in mitigating and adapting to climate change requires political leaders to be compelled to override countervailing interests by sustained pressure from below.

Diani and Donati (1999) map out some ways NGOs might mobilize civil society in the pursuit of environmental objectives and the quandaries these tactics create. For instance, groups seeking to campaign for action commonly face a dilemma between seeking cash membership for funding a professional lobby and seeking mass membership to help enable participatory action. Likewise, activists may face a choice between advocating conventional protest and pressure, and an alternative repertoire that promotes civil disobedience or disruptive protest. Each choice has advantages and disadvantages, so activists have tended to carve out niches within the NGO ecosystem. First, there is the public-interest lobby or think-tank model with professional staff and expertise but low member participation. Second, there is the participatory-pressure model, which seeks to involve rank-and-file membership but uses conventional pressure tactics. Third, there is the professional protest organization model that cultivates mainly passive membership to finance professional activism, using confrontational and disruptive tactics as well as conventional ones. Finally, there is the participatory-protest organization approach, which uses a mass membership,

low-hierarchy model and utilizes grassroots protest, civil disobedience, and disruptive tactics to incur lower costs.

Wapner (1995) further examines the ways in which a large number of NGOs have pursued opportunities to expand their role through the emergence of global civil society. This approach envisages that NGOs may seek both to pressurize national governments on climate change and to become international actors in their own right. Some environmental NGOs command larger annual budgets than those of intergovernmental organizations (IGOs), such as the United Nations Environmental Programme. They can also mobilize a level of cross-national support that many states and IGOs would struggle to match. Maurice Strong, General Secretary of the United Nations Conference on the Human Environment (held in Stockholm in 1972) and the Rio Earth Summit of 1992, consciously pushed for greater involvement by NGOs in these summits to pressurize state delegations and raise the profile of NGOs in the global media. This initiated a tradition that has continued in climate diplomacy, with engagement by NGOs beyond the protests and fringe debates that have become regular features of international climate summits. NGOs now regularly attend UNFCCC Conference of the Parties meetings as observers, advisors to delegations, critical commentators, and as conveners of side events. Many environmental NGOs also receive funding from IGOs for delivering or assisting sustainable development projects linked to climate mitigation and adaptation. One concern in the literature relates to the trade-offs between independence and influence that accompany the institutionalization of NGOs (Chatterjee and Finger, 1994). Other debates arise about differences in objectives and tactics between climate-related environmental movements based in the global North and South. Much North–South NGO networking has had very positive results, however, such as the discovery that the reporting of indigenous knowledge of climate change can be extremely powerful. The intimacy of many indigenous communities with nature and their dependence upon it lends their observations a human dimension that reaches audiences in ways that scientific reports never could.

Media Reporting of Climate Change

A third arena of climate-change communication relates to mass media reporting of climate issues. Public audiences obtain most of their information about scientific and political issues from newspapers, television, or the Internet. As such, how different mass media report climate change has strong potential to influence not just public attitudes toward the existence and seriousness of human-induced climate change, but also toward government performance in managing the issue as political life is increasingly conducted in the media spotlight (Gavin, 2009). The complexities of the relationship between science, public opinions, and politics have led to keen interest in how the mass media reports climate change and how this reporting affects social and political debate. Boykoff and Boykoff (2007: 1190) argue that: "mass-media coverage of climate change is not simply a random amalgam of newspaper articles and television segments … it is a social relationship between scientists, policy actors and the public" that is mediated through a series of journalistic norms. These include the *personalization* and *dramatization* of stories to increase their appeal, and the notion of *journalistic balance*, whereby the media seek to provide 'fair' and 'objective' coverage on issues where there is significant uncertainty or dispute by giving both sides of the argument roughly equal attention (Boykoff, 2011).

The scientific uncertainties encircling climate science potentially reduce its newsworthiness; however, personalizing and dramatizing climate change by focusing on geo-hydrological hazards, such as flooding or 'freak' weather events, and human interest stories during such events, provide powerful devices for gaining public attention and boosting sales. Dramatization has also become prominent in media reporting of the future effects of climate change. Doulton and Brown (2009) analyzed UK 'broadsheet' newspaper reports on climate change between 1997 and 2007 and report that discourses of *optimism* (climate change may bring benefits to parts of the world) and *rationalism* (other development priorities should be addressed first) have steadily been displaced by discourses emphasizing *climate change as a potential catastrophe* for developing countries to press for greater action by wealthier nations to reduce emissions and assist poorer countries in adapting to future climate changes.

Balanced reporting is a further device which can be used to add drama and conflict to stories. However, it can also create a false sense of controversy about climate change by placing the minority of scientists who are skeptical about climate change or human influence on the climate on a roughly equal footing with the majority of scientists who accept the major tenets of anthropogenic climate change. Under such circumstances, balanced reporting may produce biases in public opinion and political debate, particularly where media outlets have political motivations and use it to criticize governments. Geographical variations in balanced reporting have formed a further focus for scholars interested in the influence of the mass media on national climate policy. Boykoff (2011) notes a greater prevalence of balanced reporting on climate change in sections of the United States newspaper and television media compared with the United Kingdom that may help to explain the difficulties in creating domestic climate legislation in the United States. Dunlap and McCright (2011) similarly note the growing influence of the blogosphere as an unregulated arena for 'balanced' reporting even where major outlets follow the scientific consensus on climate change.

Just how far mass media influences public concerns about climate change, though, remains an open question. Boykoff (2011: 71) argues that: "the media don't tell people what to think, but they tell them what to think *about*." Thus, media reporting may shape public opinion on climate change less directly than is sometimes presumed but may affect it in subtler ways, via the level of attention given to the topic and the specific issues citizens are encouraged to consider, e.g., scientific uncertainty or the case for political or social action. The situation is complicated further by the so-called issue–attention cycle popularized by Anthony Downs (1972). Downs argued that environmental issues may fade from public and political attention before they are resolved (**Figure 1**), but suggested that environmental concerns may be eclipsed at a slower rate than most issues because of their visibility and widespread effects. The long-term nature of climate change may therefore give rise to periods of more intensive media coverage followed by

Figure 1 The issue–attention cycle.

interludes of relative inattention, with each cycle producing different effects on public and political opinion.

Governing Climate Change

International Climate Negotiations

In a comparative analysis of global governance, political theorists from the liberal institutionalist tradition have observed that successful international regimes tend to evolve through three main stages (O'Neill, 2009). Initially, there is an agenda-setting phase where common problems are identified and reasons for international action are developed. In the second phase, states meet to agree the principles, norms, rules, and decision-making procedures for a legal treaty or framework convention. Third, states move into a phase of regime strengthening, whereby negotiations work toward supplementing the framework convention incrementally through further agreements.

The history of climate diplomacy appears to fit the first two stages of this model. A dialogue between scientists and policymakers initiated climate negotiations during the 1980s and informed the drafting of the UNFCCC signed at the Rio Earth Summit in 1992. The climate regime then sought to embark on stage three in 1997 with the creation of the Kyoto Protocol. Thereafter, the story darkens, arguably reaching a nadir at the failed climate negotiations in Copenhagen in 2009 from which ongoing climate negotiations under the UNFCCC have struggled to recover (Falkner et al., 2010).

Political theorists from the neorealist tradition tend to direct their analysis toward less successful regimes. Neorealists acknowledge the existence of international cooperation but stress the systemic obstacles and structural barriers to its long-term success and longevity (Grieco, 1988). A key barrier is the anarchical structure of world politics, whereby states are compelled to compete with each other for wealth, power, and status in the pursuit of state security. This results in the construction of national interests that make it difficult to sustain genuine international cooperation. Such analysis could arguably be applied to the progress of climate diplomacy since 1997, where reasons of national interest are identified as contributing to durable fault-lines in the politics behind the negotiations.

First, conflicts arise out of the structure of energy interdependence in a hydrocarbon age (Paterson, 1996). States dependent on fossil fuels for export income such as the Oil Producing and Exporting Countries (OPEC), for example, are often found amongst the laggards in climate negotiations. Similarly, political leaders in of states that possess hydrocarbon-intensive economies based on ample domestic supplies, such as the United States, Canada, and Australia, have frequently cited their national interests in their resistance to binding targets on carbon emissions. Conversely, states with economies that are highly dependent upon hydrocarbon imports (notably some member states of the European Union) tend to be more inclined to support initiatives to mitigate climate change through multilateral agreements to reduce dependency on fossil fuels.

Second, there are tensions arising from inequalities in levels of vulnerability or adaptive capacity to climate change. Small island states, states with large low lying coastal areas (e.g., Bangladesh), states facing water-scarcity problems, or states which are predicted to be most vulnerable and least able to respond to climate change, such as many African countries, have advocated more rapid progress in strengthening the climate regime after the Kyoto Protocol. States that perceive themselves to be less threatened by the consequences of climate change, either through geographical accident or because of their wealth and adaptive capacity, arguably pursue negotiations with less urgency (Leiserowitz, 2007).

Third, North–South questions of climate justice have also complicated progress in the UNFCCC negotiations. This was already acknowledged in the Framework Convention principle of 'common but differentiated responsibilities.' Key articles in the Framework Convention also assimilate the norm of sustainable development and stress the need for poorer countries to pursue development while the world's richer nations lead efforts to decarbonize their economies. Attempts to mediate this fault-line have only been marginally successful, however, as differences perennially reemerge whenever attempts are made to translate these principles into specific carbon-emission commitments. The United States, for example, has long maintained that when it comes to interpreting the idea of common but differentiated responsibilities, its own historic, present, and future carbon emissions should be placed into perspective against the burgeoning emissions of newly industrializing countries, particularly China and India. Conversely, these countries and members of the G77 (the negotiating collective for developing nations in the UNFCCC system) have argued that common but differentiated responsibilities should take into account the question of historical contributions and national per capita emissions (Hurrell and Sengupta, 2012). North–South tensions have also emerged over aid for capacity building, technology transfer, and adaptation. Some southern states have remonstrated that their capacity to act is being eroded by unequal political and economic relations with the North, and are reluctant to pay excessive premiums for mitigation technologies developed in the North. They have also argued that agreements should recognize the need for adaptation finance as reparations for climate change already are in the pipeline, and that aid should also be provided where southern states are asked to forgo development to preserve natural carbon sinks such as the Amazon rain forest. Conversely, some northern states have

been reluctant to concede to climate change related aid demands, partly on the suspicion that they attempt to revive proposals made by developing countries for a New International Economic Order of reformed trade relations and increased development assistance that were roundly rejected by developed countries in the 1970s, and partly on the grounds that the emergence of countries like China, India, and Brazil as economic powerhouses make these calls redundant.

While neorealist explanations for the slow progress of international climate diplomacy are well recognized, some utilize constructivist theory to argue that the anarchic nature of the international system may drive states to pursue their national interest while pointing out that national interests are contingent rather than fixed. States are capable of learning, and of redefining their interests, for example, by placing greater emphasis on security risks presented by climate change, as they receive and interpret new information.

Governing Climate Change at the National Level

The UN climate negotiations establish the basic structure of commitments and mechanisms for international cooperation on climate change. However, substantive policies to reduce greenhouse gas emissions are primarily enacted at the national level. Several strands of enquiry have been pursued to understand the processes through which states seek to govern climate change. First, studies in the comparative politics tradition attempt to identify how general national characteristics, such as per capita wealth, education, dependency of fossil fuels for energy production, press freedom, and domestic environmental groups explain the strength or weakness of national climate policies. For instance, Harrison and Sundstrom (2007) examine the effects of electoral incentives, compliance costs, pressure from domestic interest groups, the concentration of governing authority within the country, and international influences (e.g., from UN negotiations or multinational corporations). The conclusions from such research have been somewhat inconclusive. Compliance cost emerges as a significant determinant of whether countries introduce domestic climate policy but is not decisive in all cases. Similarly, public opinion appears to have been important in persuading politicians to act on climate change in the EU, Japan, and Australia but has been less influential in the US, Canada, and Russia.

A second strand of investigation has examined the learning and diffusion processes involved in developing new ways of governing climate change. Hall (1993) distinguishes three main types of policy learning contributing to policy change: *technical learning* (learning within fixed policy objectives to improve policy effectiveness, cost-efficiency, and/or equity); *conceptual learning* (where decision-makers engage in deeper reflection about policy problems, policy goals, and the strategies used to pursue them); and *social learning* (where social debates on policy problems lead to learning by policy-makers about society's preferences and their adoption in policy). Lenschow (2005) identifies that such processes may be informed by some combination of past experiences (*lesson drawing*); experiences of others (*diffusion and emulation*); new information; or technological developments affecting the policy options available to decision-makers. Utilizing such concepts, Voβ (2007) examines how the framing of carbon trading as an effective and cost-efficient policy instrument for reducing greenhouse gas emissions contributed to the European Union's decision to adopt an EU-wide carbon-trading scheme in 2005. Related analysis has sought to explain the conditions causing the uneven uptake of new climate policies in different countries. Among the main factors examined are: attributes of the policy (including its complexity, trialability, and relative advantage over previously used instruments); and whether new types of climate policy 'fit' or clash with existing national institutional procedures and preferences (Makse and Volden, 2009).

Despite the insights provided by these approaches, both could be criticized for taking a rather apolitical view of climate policy when, in reality, climate change has become a politically charged issue in countries like the United States and Australia, where (generally) right-wing political parties and energy-intensive industries have enjoyed considerable success in blocking new climate legislation. Similarly, internal opposition to measures that threaten the development ambitions of industrializing countries (e.g., China and India) has been noted by Stern (2007). Most of the literature that has engaged more explicitly with the political dimensions of climate policy has adopted a descriptive or critical approach, charting current trends or censuring governments for prioritizing economic performance over climate protection while examining climate policy through the lens of existing political and economic mind-sets (Harris, 2009). Compston (2009) and Bailey and Compston (2012) provide more direct analysis of political barriers facing national governments on climate policy. These include: fear of electoral retribution, loss of competitiveness if other countries do not introduce comparable measures, business threats to withdraw investment, and the influence of climate 'sceptics' on political and public debates on climate change. Political strategies to counter these risks include: (1) ensuring new climate policies are designed to be as effective and cost-efficient as possible in reducing emissions; (2) emphasizing the cobenefits of climate policy (e.g., competitive advantage in low-carbon technologies and the health benefits of reducing emissions) rather than climate change itself to create a broader support base for policies; (3) developing 'package deals,' where politicians appease opponents by offering concessions in other policy areas; and (4) increasing the political clout of climate change relative to other interests by amalgamating climate and energy ministries and appointing high-profile scientists to high-level committees to encourage balanced debate between science and business interests on climate change.

Governing Climate Change Transnationally

A final sphere of climate governance to come under scrutiny is so-called *transnational* climate governance, a term coined to describe various types of public–private partnerships and networks involved in governing aspects of climate change that operate outside the control of national governments. Interest in transnational governance has stemmed principally from the recognition that globalization and the introduction of neoliberal reforms across numerous sectors and jurisdictions has resulted in a significant dispersal of governance activities (and

authority) away from nation-states to subnational authorities and nonstate actors, including corporations and NGOs.

Many lines of enquiry have again been pursued, but much interest has centered on: (1) the types of actors involved in these 'new' governing arrangements; (2) the ways in which new institutions, forms of cooperation, and political spaces are emerging in response to the limitations of state-centered climate governance; and (3) the governance functions performed by these networks. Andonova et al. (2009) provide a typology of the main actors and governance functions involved in transnational climate governance, where they stress the multiple forms of public- and private-sector authority and the use of multiple governing mechanisms, ranging from informal cooperation (e.g., information sharing) to capacity building and implementation (e.g., providing finance, expertise, labor, technology, or monitoring) to more formal rule-setting.

A first pivot of investigation has been the rapid growth since the mid-1990s in *subnational governments* and *local communities* working to place climate change on local agendas, and the consolidation of individual efforts by initiatives like Cities for Climate Protection that promote international information and resource pooling, and the creation of political identities that can enable 'local' actors to press for stronger national action on climate change (Betsill and Bulkeley, 2007). North American subnational administrations have been particularly active in developing local and regional initiatives such as the Californian emissions trading scheme and the Western Climate Initiative, a regional scheme covering several US states and Canadian provinces aimed at developing market-based solutions to climate change and laying the ground for a North American carbon market. Significant attention has also been directed at private-sector initiatives. These range from schemes encouraging individuals to pay to offset the carbon emissions from activities such as flying and the Carbon Disclosure Project (CDP), a not-for-profit organization that encourages companies and cities to disclose their emissions performance and share information about their climate-related activities (Newell and Paterson, 2010). The philosophy behind the CDP is that revealing such information helps companies to manage climate risks strategically while providing decision-makers with data and insights to develop policies that meet business and environmental needs.

Alongside chronicling the rise of climate governance networks, more critical questions have been posed about their effectiveness and implications for the state as a 'traditional' source of governing authority. Market-based approaches like the Clean Development Mechanism and voluntary carbon offsets that operate by financing emissions-reduction projects in developing countries have been criticized for methodological and governance deficits. For instance, determining whether projects such as afforestation produce additional emissions reductions depends on counterfactual arguments about what would have happened if the project had not gone ahead. A further problem is that markets tend to draw investment toward the most lucrative emissions-reduction options (the 'low-hanging fruit') and arguably reduces the pressure on developed countries to reduce their own emissions (Newell and Paterson, 2010). The broader concern is that the interests of smaller developing countries are poorly served compared with those of more powerful countries, major corporations, and multilateral institutions. The credibility of market-based approaches thus rests on effective governance in relation to target setting, accounting methodologies, lines of accountability, and sanctions against those that infringe the rules. The value of city and regional initiatives, meanwhile, rests on their ability to expand sufficiently that they make a worthwhile difference to global emissions trends or provoke stronger national or international action.

Most commentators agree that transnational governance is creating more interconnected and in some ways less top-down relationships between state and nonstate actors but has not diminished the state as a critical actor in climate governance (Bulkeley and Jordan, 2012). State involvement is even intensifying in some areas as governments and international bodies develop new legal and monitoring regimes to ensure carbon markets achieve climate objectives. Equally, states possess formidable informational and implementation capacities that provide essential legal support, credibility, and financial and other resources to transnational partnerships. As such, the growth of private and hybrid systems of climate governance may be transforming statehood, but does not appear to be undermining it.

Climate Change and Personal Behavior

Much work on the communication of climate change and climate governance rests upon certain assumptions about the mainsprings of human behavior. One common assumption is that the critical challenge in engineering more 'climate-friendly' forms of human behavior is to address current information deficits about the climate issue. According to this line of reasoning, once people have accurate information, attitudes will change and more appropriate patterns of behavior will emerge (**Figure 2**). Another common assumption is the idea of the 'rational actor model,' which assumes that people, states, and businesses are rational beings that calculate their self-interest based on all the information available to them. Sociobehavioral research in disciplines such as behavioral psychology supplements and in some respects challenges these assumptions.

A core strand of this research focuses on the so-called 'value-action gap.' This term refers to cases where individuals are well

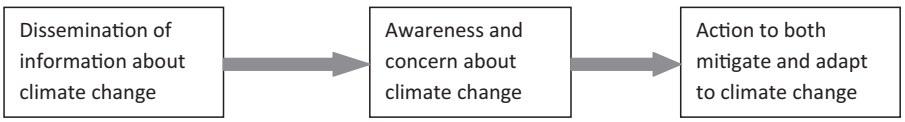

Figure 2 The information deficit model.

informed about climate science and share common values about what needs to be done about climate change, but this fails to have any impact on personal behaviors (Kollmuss and Agyeman, 2002). The question therefore arises as to how to explain situations where perceived deficits in information about climate change have been filled but where expected transformations in behavior, based on enlightened self-interest, have not materialized. Answers to this tend to fall into the realm of external and internal factors (**Figure 3**).

One external factor is where there are institutional barriers to forming a change in human behavior. Appropriate changes in behavior are difficult if enabling institutional or physical infrastructure is not in place on a macroscale (e.g., legal and binding international agreements, national policies on energy efficiency, recycling and public transport services, etc.). In other cases, infrastructure may be in place on a macrolevel but have not percolated down into the institutions with which humans interact on a routine basis (e.g., businesses, schools, universities, armed forces, and hospitals). Another factor can be lack of peer support. Changing consumption habits, for example, can be difficult or intimidating if it involves breaking inherited cultural norms and expectations. Another external factor is cost. An individual might wish to choose carbon-efficient options but feels that the cost is prohibitive.

External barriers can also have a negative impact upon individual perceptions about being the locus of control (i.e., do individuals have the power to do anything about a global problem like climate change anyway?). However, there is ambiguity over the interpretation of expressions of powerlessness, as this emotion can also have internal drivers. For example, people who have a strong predisposition to feel that they are agents of change are less likely to accept external barriers passively than those who are more resigned to accepting external constraints (Kollmuss and Agyeman, 2002). Perceptions of powerlessness can also be used as a strategy to deny personal responsibility (Barr et al., 2011a).

Other internal reasons explaining the value-action gap focus upon human cognition. People absorb new information through preexisting cognitive structures. Therefore, human recognition of climate change can sometimes be culturally overdetermined (Moser, 2007). Many people naturally associate climate with weather and are accustomed to seeing the weather as an uncontrollable force of nature. Therefore, there is a cultural propensity for cognitive spillover in the sense that climate change is interpreted as being 'like the weather' (something which individuals cannot possibly control). Evolution has also left Homo sapiens with all-too-human cognitive disabilities when making rational choices about the risks of climate change. Human judgment is colored by emotions that, from an evolutionary perspective, have become hardwired. For good prehistorical reasons, human emotions overdetermine the cognition of risk by prioritizing immediate personal danger. Climate change is a long-term process with almost unfathomable lags between cause and effect. It is incremental and intangible. It also lacks immediacy because its impacts are 'global' and can appear unthreateningly remote. Therefore, climate change may be placed in the 'less-urgent threat' category until the deadline for acting has passed.

Emotions can also contribute to the value-action gap as defense mechanisms. In some cases, individuals may recognize accounts of climate change but filter them out of consciousness to reduce cognitive dissonance (Kollmuss and Agyeman, 2002). According to the theory of cognitive dissonance, humans seek consistency in their beliefs and selectively commit information to memory based on whether it complies with or challenges existing mental frameworks. They also display archetypal fear reactions, namely: fight, flight, or freeze (Moser, 2007). If climate change information stubbornly demands

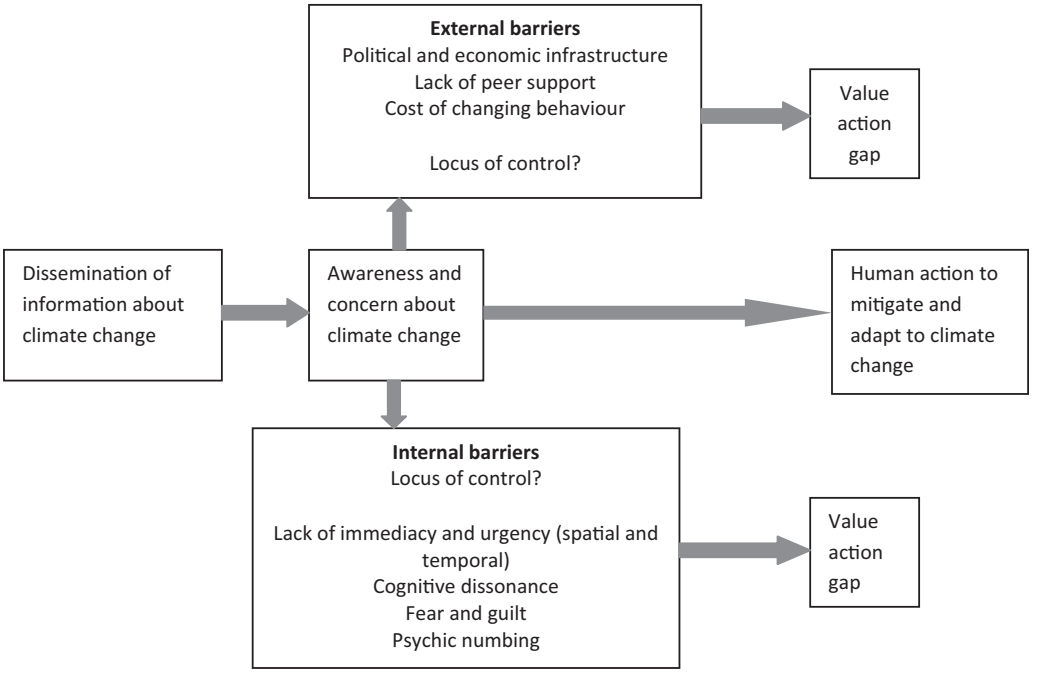

Figure 3 Information deficit model and possible reasons for 'value-action gap.'

attention (in effect another confrontation), one instinctive reaction is to contemplate putting up a fight. Flight from climate change information can come in many forms that are not always logically consistent: it does not exist; it may exist but we do not know enough yet to act; it does exist but is not a threat; it is a threat but it will not happen to us; anyway it is someone else's responsibility; technological solutions are just around the corner. A freeze response to knowledge of climate change might express itself in the form of emotive fatalism or, as Moser (2007) puts it: 'the capitulatory imagination.'

For these reasons, Moser and others caution that it can be counterproductive to manipulate emotions like fear of climate change. Attempts to evoke fear through media campaigns, for instance, may result in 'maladaptive reactions' unless target audiences feel they have a clear mitigation or adaptation plan and a sense of retaining control. Similar arguments are invoked to claim that trading on guilt can also be misguided. There is a temptation to apply shame or responsibility because guilt has some potential for changing behavior. On the other hand, humans also have a capacity for building immunity to guilt and may experience compassion fatigue, especially if they suspect it is being used manipulatively. There also comes a point where guilt can be a barrier to action if it erodes individuals' sense of identity as morally upright beings.

Another key lesson from sociobehavioral research is that effective communication of climate change requires some knowledge about the characteristics of intended target audience. To this end, social marketing techniques have been identified as a potentially fruitful way of understanding barriers to individual responses to climate change (Barr et al., 2011b). Surveys and focus groups can be used to segment public opinion according to whether individuals are committed or casual environmentalists, partly or completely disengaged, or according to their level of belief that technology will provide solutions to climate change. Thus, rather than treating audiences as an undifferentiated mass, those seeking to stimulate changes in attitudes and behavior can strategically hone climate messages to target specific audiences and deconstruct their particular barriers to behavior change. Public opinion can also arguably be segmented not just in terms of general worldviews and opinions on climate change, but also in terms of sites and practices of consumption. One study by Barr et al. (2011b) revealed that one lifestyle group that professed careful attention to 'climate-friendly' behavior in and around the home found it challenging to carry over similar behaviors when it came to tourism and air travel.

Conclusion

Although ostensibly a natural scientific issue, the social and economic effects of climate change have made it an obvious and important subject of social and behavioral enquiry. Its prominence has persisted despite ebbs and flows in political (and to some extent public) attention to climate change as a result of events like the failure of the 2009 Copenhagen Climate Conference and the external pressures and agendas created by the 2008–09 economic crisis. The multiple intertwinements that exist between the physical processes and effects associated with climate change and key aspects of human society nevertheless assure that climate change will retain a prominent, and quite conceivably preeminent, place in social-scientific and behavioral research in the coming decades.

See also: Anthropocene Epoch; Climate Change and Global Poverty; Climate Change and the Media; Environment and Development; Limits to Growth.

Bibliography

Andonova, L., Betsill, M., Bulkeley, H., 2009. Transnational climate governance. Global Environmental Politics 9, 52–73.
Barr, S., Gilg, A., Shaw, G., 2011a. Helping people make better choices: exploring the behaviour change agenda for environmental sustainability. Applied Geography 31, 712–720.
Barr, S., Shaw, G., Coles, T., 2011b. Times for (un)sustainability? Challenges and opportunities for developing behaviour change policy. A case-study of consumers at home and away. Global Environmental Change 21, 1234–1244.
Bailey, I., Compston, H., 2012. Feeling the Heat: The Politics of Climate Policy in Rapidly Industrializing Countries. Palgrave Macmillan, Basingstoke.
Betsill, M., Bulkeley, H., 2007. Looking back and thinking ahead: a decade of cities and climate change research. Local Environment 12, 447–456.
Boykoff, M., 2011. Who Speaks for the Climate: Making Sense of Media Reporting of Climate Change. Cambridge University Press, Cambridge.
Boykoff, M., Boykoff, J., 2007. Climate change and journalistic norms: a case-study of US mass-media coverage. Geoforum 38, 1190–1204.
Bulkeley, H., Jordan, A., 2012. Transnational environmental governance: new findings and emerging research agendas. Environment and Planning C 30, 556–570.
Chatterjee, P., Finger, M., 1994. The Earth Brokers: Power, Politics and World Development. Routledge, London.
Compston, H., 2009. Networks, resources, political strategy and climate policy. Environmental Politics 18, 727–746.
Diani, M., Donati, P., 1999. Organisational change in Western European environmental groups: a framework for analysis. Environmental Politics 8, 13–34.
Doulton, H., Brown, K., 2009. Ten years to prevent catastrophe? Discourses of climate change and international development in the UK press. Global Environmental Change 19, 191–202.
Downs, A., 1972. Up and down with ecology: the 'issue attention cycle'. Public Interest 28, 28–50.
Dunlap, R., McCright, A., 2011. Organised climate change denial. In: Dryzek, J., Norgaard, R., Schlosberg, D. (Eds.), The Oxford Handbook of Climate Change and Society. Oxford University Press, Oxford, pp. 144–160.
Falkner, R., Hannes, S., Vogler, J., 2010. International climate policy after Copenhagen: towards a 'building blocks' approach. Global Policy Journal 1, 252–262.
Gavin, N., 2009. Addressing climate change: a media perspective. Environmental Politics 18, 765–780.
Grieco, J., 1988. Anarchy and the limits of cooperation: a realist critique of the newest liberal institutionalism. International Organization 42, 485–507.
Haas, P., 1992. Epistemic communities and international policy coordination. International Organization 46, 1–35.
Hall, P., 1993. Policy paradigms, social learning, and the state: the case of economic policy making in Britain. Comparative Politics 25, 275–296.
Harris, P., 2009. Beyond bush: environmental politics and prospects for US climate policy. Energy Policy 37, 966–971.
Harrison, K., Sundstrom, L.M., 2007. The comparative politics of climate change. Global Environmental Politics 7, 1–18.
Hurrell, A., Sengupta, S., 2012. Emerging powers, north-south relations and global climate politics. International Affairs 88, 463–484.
IPCC, 2007. Summary for Policymakers. Cambridge University Press, Cambridge.
Kollmuss, A., Agyeman, J., 2002. 'Mind the gap': why do people act environmentally and what are the barriers to pro-environmental behaviour? Environmental Education Research 8, 239–260.
Leiserowitz, A., 2007. Communicating the risks of global warming: American risk perceptions, affective images, and interpretive communities. In: Moser, S., Dilling, L. (Eds.), Creating a Climate for Change. Cambridge University Press, Cambridge, pp. 44–63.

Lenschow, A., 2005. New regulatory approaches in 'greening' EU policies. In: Jordan, A. (Ed.), Environmental Policy in the European Union. Earthscan, London, pp. 295–316.
Makse, T., Volden, C., 2009. The role of policy attributes in policy diffusion. Journal of Politics 73, 108–124.
Moser, S., 2007. More bad news: the risk of neglecting emotional responses to climate change information. In: Moser, S., Dilling, L. (Eds.), Creating a Climate for Change. Cambridge University Press, Cambridge, pp. 64–80.
Newell, P., Paterson, M., 2010. Climate Capitalism: Global Warming and the Transformation of the Global Economy. Cambridge University Press, Cambridge.
O'Neill, K., 2009. The Environment and International Relations. Cambridge University Press, Cambridge.
O'Riordan, T., Jordan, A., 1997. The intergovernmental panel on climate change: consensual knowledge and global politics. Global Environmental Change 7, 77–79.
Paterson, M., 1996. Global Warming and Global Politics. Routledge, London.
Stern, N., 2007. The Economics of Climate Change: The Stern Review. Cambridge University Press, Cambridge.
Toke, D., 1999. Epistemic communities and environmental groups. Politics 19, 97–102.
Voβ, J.P., 2007. Innovation processes in governance: the development of 'emissions trading' as a new policy instrument. Science and Public Policy 34, 329–343.
Wapner, 1995. Politics beyond the state: environmental activism and world civic politics. World Politics 47, 311–340.

Climate Change and Global Poverty

Abigail Jones and Joshua Hermias, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although the effects of climate change will be felt world over, studies broadly agree that the distribution of impacts will be skewed against the world's poorest countries which have limited capacity to adapt. This differential impact is both a function of geography and of a confluence of economic, social, and political factors. This article highlights projected effects of climate change across a number of key sectors in low-income countries as a forecast across the literature. We conclude with a discussion of key research challenges for the decade of the 2020s.

Climate Change and Global Poverty

Although the effects of climate change will be felt across the entire planet, studies broadly agree that the distribution of negative impacts are skewed toward the world's poorest countries. Climate-related changes to the environment are especially disproportionate near the equatorial belt, an area that mostly comprises low-income countries. Compounding this heterogeneity of environmental effects is the low-income countries' acute vulnerability to climate change. As these countries continue to address long-standing economic, social, and political development needs, their capacity to recover from and adapt to climate change impacts is limited.

Both an individual's and a country's ability to cope with the effects of climate change are fundamentally connected to economic and human development. When faced with rising food or fuel prices, for example, wealthy individuals can cope by dipping into savings; however, for the poorest families, who spend 45–80% of their income on food, rising prices force stark tradeoffs (Banerjee and Duflo, 2011). Similarly, low-income countries often lack sufficient resources and technical capacity to adequately invest in protective infrastructure, social safety nets, and recovery operations, especially when compared to the capacities of high-income countries.

Adverse climate impacts are also expected to threaten positive trends in growth and poverty reduction over the medium and long term. In poorer countries, a 1 °C rise in regional temperature in a given year dampens growth in that year by about 1.3 percentage points, meaning that the adverse effects of climate change will be felt throughout the economies of poor countries for years to come (Dell et al., 2012). This article highlights projected effects across a number of key sectors in the developing world as forecast across the literature, and concludes with a discussion of key research challenges for the coming decade. The impacts discussed come from a diverse body of scholarship that is based on differing assumptions and underlying models. However, before turning to sector-specific impacts, it is important to situate this challenge within the broader frame.

Sustained growth, climate change, and related impacts comprise a complex system that remains a topic of intense scrutiny. A key ingredient to economic growth to date has been carbon-intensive energy. In turn, carbon-intensive activities exacerbate climate change; the impacts of these changes eventually feedback into growth potential. Thus, many observers point to an 'energy-growth puzzle'. Although many pollutants seem to follow the so-called environmental Kuznets curve, rising with income until a certain threshold is met and then declining thereafter, greenhouse gas emissions do not follow this pattern; though emission intensity tends to diminish at high-income levels, emissions rise monotonically with income at every level (Jones et al., 2009).

Increasing access to energy remains an important development imperative to many countries' growth strategy. Today, about 1.3 billion people lack access to electricity (one in five people), while unreliable electricity networks serve another 1 billion people. More than 95% of the energy poor are either in sub-Saharan Africa or developing Asia, while 84% are in rural areas – the same regions that are the most vulnerable to the adverse effects of climate change (International Energy Agency, 2011). At the same time, successfully confronting climate change – the impacts of which are already being felt disproportionately by the poor – will require a global shift toward sustainable energy. Half the world's greenhouse gas emissions come from developing countries, however in 2030, carbon dioxide emissions from non-OECD countries are projected to exceed those from OECD countries by 72%. Because unmanageable climate change risks threaten to reverse development gains, efforts to alleviate energy poverty are sometimes at odds with mitigating climate impacts; sustainable and noncarbon-intensive energy sources are an often-cited alternative for solving portions of the energy-growth puzzle.

Water

Climate change is likely to alter precipitation patterns (Meehl et al., 2007). Already, increased precipitation has been observed in the eastern parts of North America and South America, Northern Europe, and Northern and Central Asia, while drying has been observed in the Sahel, Southern Africa, the Mediterranean, and parts of Southern Asia. Models predict that global warming will affect water availability and demand across the developing world with some regions experiencing increased drought and others experiencing severe flooding as a result. These stresses will present challenges for human health, agricultural productivity, and local infrastructure, with attendant consequences for poverty alleviation and economic growth.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64042-4

Increased water stress is expected across many parts of the developing world with more precipitation in high latitudes and less precipitation in subtropical regions. Throughout Africa, warming is likely to be larger than the global annual mean, which in part will contribute to the complete disappearance of the ice caps of Mount Kilimanjaro by 2020 (Thompson et al., 2002). Between 75 and 250 million people, or 6–20% of the projected population, will be at risk by 2020; another 275–350 million people are expected to join their ranks by 2050 leaving anywhere between 16 and 30% of the continent's population facing water stress from climate change alone (Arnell, 2004). Three degrees of warming may push many African nations beyond the limits of their economically usable, land-based water resources within the next 15 years (Ashton, 2002). In Central and South America, annual mean warming is likely to be above the global mean, leaving between 12 and 81 million people at risk of increased water stress by 2020 and 79–178 million people at risk by 2050 – up to 25% of Latin Americans (Magrin et al., 2007). In Asia, the melting of glaciers due to warming temperatures could seriously affect half a billion people in the Himalaya-Hindu-Kush region and a quarter of a billion people in China who depend on glacial melt for their water supplies. While river runoff will initially increase with glacial melt, over time water supplies will fall as a result of limited ice resources with detrimental effects for downstream agriculture in most countries of South Asia (Cruz et al., 2007).

Climate change will also affect water quality through both direct and indirect effects (Nyong, 2009). The global average sea level rose at an average rate of 1.8 mm a year, over the period 1961 to 2003, however the rate of sea-level rise between 1993 and 2003 was considerably faster, at about 3.1 mm a year (Solomon et al., 2007). As sea levels rise due primarily to ice sheet melt, the coastal regions of Africa, Bangladesh, China, India, and the small island states will see increased salinity in their ground and surface water. Furthermore, years of over-exploiting groundwater will leave it vulnerable to saltwater intrusion in many coastal areas (UNEP, 2007).

Water sustains the livelihoods of society and makes productive economic activity possible. With the hydrological changes anticipated in a warmer world, future water scarcity may put at risk many societies' capacity to feed their growing populations, power industrial development, for example, for thermal power plants, as well as the functioning of natural ecosystems absent innovation and technological change (World Bank, 2012).

Agriculture and Food Security

The impact of climate change on global agricultural activity will have acute affects in the developing world. Seventy-five percent of the world's poor live in rural areas, with 86% of rural people engaged in agriculture-related activities. Agricultural activities are especially important to development because work occurs both in the formal labor market, where individuals are employed as workers, and as a form of self-employment or subsistence activity (World Bank, 2007).

From a macroeconomic perspective, developing countries remain dependent on the farming sector as an engine for growth and development. Agriculture accounts for about 30% of gross domestic product (GDP) in developing economies. The World Bank (2007) estimates, for example, that GDP growth generated in agriculture, on average, is four times more effective in benefiting the poorest half of the world's population than growth generated outside agriculture. However, the performance of the agricultural sector is sensitive to climate change impacts. This sensitivity is especially acute for developing countries, which have less fiscal and technical capacity to adapt to climate impacts yet face some of the more severe projections in terms of agricultural productivity.

Climate change impacts on temperature and rainfall affect a variety of factors important to agricultural environments and ecosystems: the location and area of land suitable for farming, the selection of crop varieties that are viable, and the length and productivity of the growing season. These several factors result in a wide variety of predicted outcomes. Global warming is predicted to favorably impact agricultural environments in current temperate and even arctic areas of the Northern Hemisphere; rising temperatures and better rainfall patterns will lengthen growing seasons in some areas and open new lands to farming in other areas. China, for example, may experience gains in cereal outputs of more than 20%. Climate change models also suggest, however, negative impacts on agricultural activity, especially for many tropical and subtropical developing countries. In these areas, the area of land suitable for farming decreases amid rising temperatures and increased water scarcity. Under these conditions, it is thought that wheat crops, for example, will completely collapse in sub-Saharan Africa. At the global level, the net impact of climate change on agricultural GDP from projection baselines are small, ranging between −1.5 and 2.6%; however, for most developing countries, the impact is certainly negative, with Africa seeing the largest losses, ranging between −2 and −9% of GDP (Fischer et al., 2002).

Agricultural yields may also be impacted by climate change through secondary affects. Warmer environments are often more hospitable to crop-destroying pests and plant disease. Erratic and volatile weather patterns make it difficult for farmers to plan – and at the most extreme, severe storms, floods, or droughts can directly damage crops.

Agricultural activities also include aquaculture and livestock. Like plant crops, these activities are also sensitive to environmental conditions and impacts will be heterogeneous. Temperature increases of 1.5 °C, for example, can cause fish stocks to shrink or move to different locations like plant crops – some areas will gain and some areas will lose (McLean et al., 2001).

From a human development perspective, agriculture also remains vitally important to the poor, who both depend on agricultural employment for wages as well as agricultural outputs for food security. While agricultural work can be part of a robust livelihood, there is significant overlap between those who are poor and those who are engaged in agriculture. In Banerjee and Duflo's (2011) 18-country data set, the percentage of rural households living in extreme poverty with a household member working in agriculture ranged from 0.6% in Ghana to 93% in Cote d'Ivoire. Households also reported a wide range of self-employment in the agricultural sector: 32% of rural households living in extreme poverty in Mexico report

that at least one household member is self-employed in agriculture, while this figure in Bangladesh was roughly 92%.

Health

Climate change will pose additional challenges for human health and delivery systems across the developing world with adverse consequences for development and economic growth, as healthy populations are more productive and resilient in the face of adverse shocks. It will do so through four primary channels including: direct weather events, for instance, heat waves and severe storms, changed microclimates that alter the range of diseases, altered precipitation and drought that vary food production, and increased air pollutants and airborne allergens. Already, 9 million children under 5 die each year across the developing world from preventable diseases such as diarrhea and malaria (Banerjee and Duflo, 2011). While mortality attributable to climate change accounted for only 0.3% of global deaths and 0.4% of disability-adjusted life years lost (a measure of the total health burden) in 2000, further warming is projected to aggravate these burdens, complicating household decision-making about everything from education to investment and adding pressures to nations' health infrastructure and budgets (Ebi, 2009).

Direct weather events will pose new challenges for health systems across the developing world. Tropical cyclones, storm surges, heat stress, flooding, and drought will directly cause injuries and mortalities, as well as displace affected populations. Migration resulting from climatic threats provides an avenue for the spread of diseases, while natural disasters are expected to take a heavy toll on the mental health of the people involved (Hajat et al., 2003).

Climate change is projected to alter the range of vector (tick, rodent) and waterborne diseases as precipitation patterns, temperatures, and humidity change, including malaria and dengue fever, increasing the absolute number of people at risk. Estimates suggest by 2050, over 200 million people will be at risk of contracting malaria due to climate change pushing the total figure to roughly 5.2 billion. In some areas in South America, sub-Saharan Africa, and Asia, the malaria transmission rate is projected to be 50% higher (Béguin et al., 2011).

Increasing temperatures and decreasing water availability due to climate change may increase burdens of diarrheal and other infectious diseases in many parts of the developing world. This is especially true in some small island developing states (Singh et al., 2001), regions of sub-Saharan Africa, and in Southeast Asia (McMichael et al., 2004). Widespread contraction of diarrhea could exacerbate malnutrition, compromising peoples' ability to work and attend school.

New air pollutants and airborne allergens will increase across regions experiencing more frequent forest fires, including some countries in Latin America. Forest fires have been associated with an increased risk of respiratory diseases and breathing problems, which could adversely affect the labor force in effected countries (WHO, 2000).

The extent of these impacts will depend not only on the magnitude and rate of climate change, but also on factors including wealth, education, the status of each nation's public health infrastructure, the provision and scope of medical care, access to adequate nutrition, safe water and sanitation, and the timeliness and effectiveness of interventions to address the risks of climate change (Ebi, 2009). It will also depend on whether current barriers to the adoption of cheap, effective, and readily available treatments (for instance, bed nets to curb malaria transmission) can be overcome in the near term (Banerjee and Duflo, 2011).

Coastal Areas and Settlements

There has been sustained population growth in coastal areas over the past two decades. Currently, 40% of the world's population lives 100 km from the sea; 73% of this population live in developing countries.

Coastal populations have long been centers of poverty. Campbell et al. (2006) posit that the poor are attracted to the coast because of access to common-pool resources like fish and because of a diversity of livelihoods that are often eschewed by more privileged groups. Of course, these livelihoods tend to be more fragile, depending on hard labor and risk taking, as well as exogenous factors commonly associated with natural resources. Even apart from climate change impacts, for example, fishing livelihoods are increasingly vulnerable to shocks from pollution, competition from global markets, and overfishing.

The most profound impact of climate change on coastal areas is sea-level rise. Global warming leads to higher sea level from melting ice caps and glaciers, as well as through the thermal expansion of water as it becomes warmer. Projecting future sea-level rise remains a considerable challenge for the scientific community, with various methodologies producing a large range of possibilities: at a 4 °C warming by 2012, the mean estimate ranges between 47 and 96 cm of sea-level rise, depending on methodology (World Bank, 2012). Absent massive infrastructure investments like seawalls, as sea levels rise, low-lying coastal areas will become inundated with water leading to increased flooding, erosion, and, in the most extreme cases, submersion of coastlands and mangrove forests. These outcomes can lead to population displacement, loss of life and property, contamination of groundwater, and destruction of infrastructure and natural features – all of which impact the poor both directly and through general disturbances to economic activities they may depend on.

Sea-level changes, like many climate impacts, are not evenly distributed. Numerous factors including sea floor morphology, Earth's rotation and gravitational forces, and air and water currents affect the distribution of ocean water. Low-lying tropical areas are more vulnerable, for example, as sea-level rise may be 20% higher in this region than near the poles (World Bank, 2012). Similarly, the geography of small islands leave them acutely exposed to extreme flooding, salinization of drinking water, and changes in surrounding ecosystems.

The increased levels of atmospheric carbon dioxide associated with global warming projections also lead to projections of ocean acidification, as CO_2 dissolves into water. Ocean acidification can exacerbate coastal flooding because it can lead to the destruction of protective barrier reefs. The loss of reefs and changing habitat environments may also devastate fish

populations and tourism, and thus threaten coastal livelihoods (World Bank, 2012).

Natural Ecosystems and Biodiversity

Almost one-third of the wealth of low-income nations is in their 'natural capital' – their rivers, forests, grasslands, and oceans (World Bank, 2006). Yet the value of this capital per person is lower in developing nations than in high-income countries, which are far less forested in absolute terms (World Bank, 2006). Natural ecosystems are equally important for human development, as forest resources directly contribute to the livelihoods of almost 90% of the extreme poor in developing nations (World Bank, 2003). However, recent assessments have identified significant effects that climate change will have on terrestrial ecosystems.

Terrestrial systems will be affected to varying degrees across and within regions of the developing world. In Latin America, numerous tree species will likely be lost. By mid-century, 24% of 138 tree species in the savannas of central Brazil will be threatened with extinction, while by the century's end, 34% of 69 tree plant species could become extinct in Amazonia. Up to 40% of the Amazonian forests could react drastically to even a slight reduction in precipitation with attendant consequences for tropical vegetation, hydrology, and the climate system across South America (Magrin et al., 2007). In Africa, forest productivity will likely decline (UNEP, 2007) and between 25 and 42% of plant species could lose all suitable range by 2085 (World Bank, 2012). In Asia, up to 50% of the region's total biodiversity could be at risk, with projections showing that 105 to 1522 plant species in China alone could become extinct by the end of the century (Cruz et al., 2007).

These degraded and lost sources of natural capital will place the poor at further risk, as forests provide a natural defense against flooding and drought, and can improve water quality by filtering out harmful pollutants and pathogens. Forests also provide other less tangible benefits to poor communities. For example, in the Amazon River Basin indigenous peoples consider forests to be essential to their cultures, traditional livelihoods, and religious practices (Jones et al., 2009).

Disasters

A warmer world is projected to be a more violent one, with more frequent and intense extreme weather events potentially setting back economic and human development gains. Between 1990 and 1999, an estimated 188 million people a year were affected by extreme weather, droughts, floods, and other such disasters, six times more than the 31 million annually affected by armed conflict. Immense storms, like the massive cyclone that destroyed much of coastal Burma in May 2008, uproot millions of people from their homes. Some estimates suggest that the number of environmental migrants and refugees could grow by tens of millions in the coming decades. In addition to direct displacement, disaster victims often end up in makeshift camps or dysfunctional communities plagued by crime and violence (Jones et al., 2009).

Substantial scientific evidence has begun to link recent, extreme weather events to global warming with medium to high levels of confidence. In 2011, record breaking heat and drought in the United States induced damages estimated between US$6 and US$8 billion and proved that developed countries are also vulnerable to extreme weather. The European heat wave of 2003 claimed more than 70 000 lives (World Bank, 2012). Yet in the future, poor countries and communities, particularly in Africa and Asia, will face the gravest risks. Their vulnerability is both a function of their geography, as places near the equator and the poles and along seacoasts are likely to experience the most pronounced effects of climate changes, and of weak institutions. Developed economies have greater capacity to adapt and absorb displaced populations (CAN Corporation, 2007), while states with weak institutions stand a higher risk of being destabilized by extreme weather events.

Conclusion

Climate change and development are fundamentally linked. How nations grow affects the climate, yet whether nations grow determines how economies and citizens are equipped to cope with climate impacts. If countries develop along traditional, carbon-intensive energy pathways, climate impacts will mount, all else equal. However, the ability to develop robust economies, political systems, and societies is also critical in determining local, regional, and global responses to climate change: adverse climate impacts will be magnified if developing countries are too poor to invest in adaptive measures and infrastructure or weather, the global economic consequences of mitigating climate change.

In turn, global warming has enormous implications for the lives of the millions living on less than a dollar a day across the developing world. What is in store for the world's poorest people as global mean temperature rises depend in part how much the world is able to abate greenhouse gas emissions. Yet many known adverse effects, including sea-level rise, heat waves, and rainfall variation are already underway. Although quantifying the full scale of consequences on livelihoods is not yet possible, it is clear that climate change has the potential to seriously undermine poverty alleviation in many regions. Both past trends and future projections show that adverse climate impacts have a negative effect on economic growth in developing countries (World Bank, 2012).

Future researchers will need to grapple with the many feedback loops and interactive effects between the sectors as impacts mount, greenhouse gas emissions continue to accumulate, and populations grow across the developing world. For instance, greater water scarcity in regions experiencing decreased annual precipitation will have significant effects on national economies through a number of channels including health and agriculture. However, greater water insecurity could also compromise national energy systems, as they require water to power generation and transmission. The dynamic effects between sectors have not yet been fully modeled, and such work would help provide a clearer understanding of how both low-income and high-income economies will need to adapt to flourish in a warmer world.

See also: Climate Change: Economics; Climate Change; Development Theory in Geography; Eco-cities and Sustainable Urbanism; Energy, Anthropology of; Geographies of Food: Politics, Place, Power Reprised; Health in Low-Income Countries; Poverty, Geography of; Poverty: Measurement and Analysis; Resilient City; Resource Geography; Sustainable Lifestyles; Sustainable Practices; Water Resources and Sustainable Water Management; Water, Health and Social Inequality.

Bibliography

Adger, W.N., Adhikary, S., Ando, M., Anokhin, Y., Cruz, R.V., Ilyas, M., Kopaliani, Z., Lansigan, F., Li, C., Patwardhan, A., Safriel, U., Suharyono, H., Zhang, X., 2001. Asia. In: McCarthy, J.J., Canziani, O.F., Leary, N.A., Dokken, D.J., White, K.S. (Eds.), Climate Change 2001: Impacts, Adaptation, and Vulnerability—Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge UP, Cambridge, pp. 533–590.

Arnell, N.W., 2004. Climate change and global water resources: SRES emissions and socio-economic scenarios. Global Environmental Change 14, 13–52.

Ashton, P.J., 2002. Avoiding conflicts over Africa's water resources. Ambio 31, 236–242.

Banerjee, A., Duflo, E., 2011. Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty. PublicAffairs, New York.

Béguin, A., Hales, S., Rocklöv, J., Åström, C., Louis, V.R., Sauerborn, R., 2011. The opposing effects of climate change and socio-economic development on the global distribution of malaria. Global Environmental Change 21, 1209–1214.

Campbell, J., Whittingham, E., Townsley, P., 2006. Responding to coastal poverty. In: Hoanh, C.T., Tuong, T.P., Gowing, J.W., Hardy, B. (Eds.), Environment and Livelihoods in Tropical Coastal Zones. CAB International, Wallingford, Oxon.

CAN Corporation, 2007. National Security Threat of Climate Change. CAN Corporation, Washington.

Cruz, R.V., Harasawa, H., Lal, M., Wu, S., Anokhin, Y., Punsalmaa, B., Honda, Y., Jafari, M., Li, C., Huu Ninh, N., 2007. Asia. In: Parry, M.L., Canziani, O.F., Palutikof, J.P., van der Linden, P.J., Hanson, C.E. (Eds.), Climate Change 2007: Impacts, Adaptation and Vulnerability—Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge UP, Cambridge, pp. 469–506.

Dell, M., Jones, B.F., Olken, B.A., 2012. Temperature shocks and economic growth: evidence from the last half century. American Economic Journal: Macroeconomics 4, 66–95.

Ebi, K.L., 2009. Public health adaptation to climate change in low-income countries. In: Brainard, L., Jones, A., Purvis, N. (Eds.), Climate Change and Global Poverty: A Billion Lives in the Balance? Brookings Institution Press, Washington, pp. 130–141.

Fischer, G., Shah, M., Tubiello, F.N., van Velhuizen, H., 2002. Climate Change and Agricultural Vulnerability. International Institute for Applied Systems Analysis, Special Report.

Fischer, G., Shah, M., Tubiello, F.N., van Velhuizen, H., 2005. Socio-economic and climate change impacts on agriculture: an integrated assessment 1990–2080. Philosophical Transactions of the Royal Society 360, 2067–2083.

Food and Agriculture Organization, 2009. Executive Summary. Report from Expert Meeting on "How to Feed the World in 2050". FAO, Rome.

Hajat, S., Ebi, K.L., Edwards, S., Haines, A., Kovats, S., Menne, B., 2003. Review of the human health consequences of flooding in Europe and other industrialized civilizations. Applied Environmental Science and Public Health 1, 13–21.

International Energy Agency, 2011. Energy for All: Financing Access for the Poor, Special Early Excerpt. World Energy Outlook 2011. OECD/IEA, Paris.

Jones, A., LaFleur, V., Purvis, N., 2009. Double jeopardy: what the climate crisis means for the poor. In: Brainard, L., Jones, A., Purvis, N. (Eds.), Climate Change and Global Poverty: A Billion Lives in the Balance? Brookings Institution Press, Washington, DC, pp. 10–42.

Magrin, G., Gay García, C., Cruz Choque, D., Giménez, J.C., Moreno, A.R., Nagy, G.J., Nobre, C., Villamizar, A., 2007. Latin America. In: Parry, M.L., Canziani, O.F., Palutikof, J.P., van der Linden, P.J., Hanson, C.E. (Eds.), Climate Change 2007: Impacts, Adaptation and Vulnerability—Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge UP, Cambridge, pp. 581–615.

McLean, R.F., Tsyban, A., Burkett, V., Codignotto, J.O., Forbes, D.L., Mimura, N., Beamish, R.J., Ittekkot, V., 2001. Coastal zones and marine ecosystems. In: McCarthy, J.J., Canziani, O.F., Leary, N.A., Dokken, D.J., White, K.S. (Eds.), Climate Change 2001: Impacts, Adaptation and Vulnerability—Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge UP, Cambridge, pp. 343–379.

McMichael, A.J., Campbell-Lendrum, D., Kovats, S., Edwards, S., Wilkinson, P., Wilson, T., Nicholls, R., Hales, S., Tanser, F., le Sueur, D., Schlesinger, M., Andronova, N., 2004. Global climate change. In: Ezzati, M., Lopez, A., Rodgers, A., Murray, C. (Eds.), Comparative Quantification of Health Risks: Global and Regional Burden of Disease Due to Selected Major Risk Factors. World Health Organization, Geneva, pp. 1543–1649.

Meehl, G.A., Stocker, T.F., Collins, W.D., Friedlingstein, P., Gaye, A.T., Gregory, J.M., Kitoh, A., Knutti, R., Murphy, J.M., Noda, A., Raper, S.C.B., Watterson, I.G., Weaver, A.J., Zhao, Z.C., 2007. Global climate projections. In: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L. (Eds.), Climate Change 2007: The Physical Science Basis—Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge UP, Cambridge, pp. 747–846.

Nyong, A., 2009. Climate change impacts in the developing world: implications for sustainable development. In: Brainard, L., Jones, A., Purvis, N. (Eds.), Climate Change and Global Poverty: A Billion Lives in the Balance? Brookings Institution Press, Washington, pp. 43–64.

Parry, M.L., Canziani, O.F., Palutikof, J.P., van der Linden, P.J., Hanson, C.E. (Eds.), 2007. Climate Change 2007: Impacts, Adaptation and Vulnerability—Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge UP, Cambridge.

Singh, R.B.K., Hales, S., deWet, N., Raj, R., Hearnden, M., Weinstein, P., 2001. The Influence of climate variation and change on diarrheal disease in the Pacific Islands. Environmental Health Perspectives 109, 155–159.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L. (Eds.), 2007. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge UP, Cambridge.

Thompson, L.G., Mosley-Thompson, E., Davis, M.E., Henderson, K.A., Brecher, H.H., Zagorodnov, V.S., Mashiotta, T.A., Lin, P.N., Mikhalenko, V.N., Hardy, D.R., Beer, J., 2002. Kilimanjaro ice core records: evidence of Holocene change in tropical Africa. Science 298, 589–593.

United Nations Environment Program, 2007. Global Environmental Outlook—GEO4: Environment for Development. UNEP, New York.

World Bank, 2003. Sustaining Forests: A World Bank Strategy. World Bank, Washington.

World Bank, 2006. Where Is the Wealth of Nations? Measuring Capital for the 21st Century. World Bank, Washington.

World Bank, 2007. World Development Report 2008: Agriculture for Development. World Bank, Washington.

World Bank, 2012. Turn Down the Heat: Why a 4 Degree World Must Be Avoided. World Bank, Washington.

World Health Organization, 2000. Vegetation Fires, Fact Sheet 254. WHO, Geneva.

Climate Change and the Media

Mike S Schäfer, Universität Zürich, Institut für Publizistikwissenschaft und Medienforschung, Zürich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Media communication – and the coverage of mass media such as TV, radio, newspapers, or the Internet in particular – is an important source for people's awareness of, and knowledge about, anthropogenic climate change. Accordingly, many scholars have analyzed the emergence, characteristics, uses, and effects of mediated communication about climate change in recent years. This article reviews the respective field, presents its major findings, and outlines future directions for research.

The Relevance of the Media in Climate Change Communication

Climate change, compared to many other issues, is an 'unobtrusive' topic. This is because the climate and its past, current, or future developments are not easily observable first hand. The term 'climate' refers to statistical averages of weather indicators, such as temperatures, precipitation, wind, and so on. To speak of 'climate,' these indicators have to be monitored and described on large scales both in terms of space and time. The World Meteorological Organization, for example, proposes to speak of 'climate' when referring to average weather over at least 30 years. Spatially, 'climates' are usually described for entire countries, continents, hemispheres, or the whole world. Such dimensions lie far beyond most people's life-worlds and biographical horizons. In addition, the main anthropogenic causes of climate change – greenhouse gas emissions – remain invisible.

A second reason for the issue's unobtrusiveness is its complexity. The climate and its changes are observed and (re)constructed primarily by scientists, whose results are rather complex and difficult to understand for many people. This is due, among other factors, to the growing number of disciplines participating in climate science. Each comes with its own kinds of data, measures, models, and heuristics, which results in climate models including more and more variables and interrelations between these. Moreover, the findings of climate science, as those of any other science, are not unanimously agreed upon. Although a consensus seems to exist within the respective scientific community about the basic features and causes of anthropogenic climate change, dissent, and uncertainty can be found in many details.

Thirdly, climate change is also an unobtrusive issue because its social effects and the measures to mitigate them are debated, complex, and, at times, difficult to understand: A lot of the consequences of climate change, and most likely the more severe ones, are not occurring here and now. They will appear in the future and affect countries in the 'Global South' more than those in North America or Western Europe. In addition, stakeholders advocate different actions that are often based on elaborate, difficult to understand rationales and justifications. The most important strand of current climate politics involves international efforts at 'mitigation,' that is, a reduction of greenhouse gas emissions in developed countries, in particular, constituting a political process that is far removed from people's life-worlds.

As a result, the causes and consequences of climate change and its implications are not directly and easily perceivable, and what most people know about them stems from media communication. The mass media – means of communication that distribute content such as text, pictures, and sound to an anonymous and spatially diverse public via technical means, such as newspapers, TV, or radio – are particularly relevant in this respect. Surveys in the US, the UK, Australia, and Germany show that mass media – particularly television, followed by newspapers and, increasingly, the Internet – are the most important sources from which people draw information about climate change. These studies also show mass media to be more important than people's interpersonal communication with family members and friends, as well as workshops and school classes. Accordingly, they are "important arenas and important agents in the production, reproduction, and transformation of the meaning" of climate change (Carvalho, 2010: 172).

Overview of the Research Field

The outlined importance of mass media communication on climate change has long been acknowledged by the scientific community. A number of introductory texts into the field are available, such as Moser's (2010) historical overview of climate change communication that spans mass media as well as other kinds of communication. Carvalho's (2010) description of the politics of climate change communication, Schäfer's (2012) overview of online communication about the issue O'Neill et al. 's (2013) review of visual communication about climate change, and Anderson's (2009) or Olausson and Berglez's (2014) programmatic formulation of a future research agenda are further examples.

Taken together as a whole, the research field reveals a number of specifics. A recent *meta*-analysis of some 130 relevant publications shows the growth of the field (Schäfer and Schlichting, 2014): After only a few analyses were published in the 1980s and 1990s, research activity increased and remained on a moderate level until the mid-2000s, when annual publication numbers rose considerably. Today, analyses of media representations of climate change constitute a lively and growing field.

This growth is accompanied by a diversification of research approaches and objects. A relatively large number of earlier single-country case studies (e.g., Boykoff, 2008; Gavin et al., 2011) have increasingly been supplemented by comparative

and longitudinal research. The theoretical foundations of the respective studies range from politically oriented agenda-setting approaches to critical discourse analysis and adopt quantitative, qualitative, and mixed method research designs.

In spite of this diversification, however, the research field is still strongly focused on a small number of countries and media: European and North American countries have received by far the largest share of research attention and continuously do so. Asian, African, and Latin American countries, many of whom are strongly affected by the outcomes of climate change, have not often been analyzed (Schäfer and Schlichting, 2014). In addition, most studies focus on print media, particularly broadsheet newspapers. Even though the number of analyses taking up online and social media is increasing, other media are largely neglected. The importance of television as a source of information about climate change, in particular, has not led to an equally strong amount of scholarly attention for climate change portrayals on TV.

Furthermore, not all steps of the communication process are taken up equally. Most studies focus on the media presentation of climate change, i.e., on how much attention media devote to the issue, or how they evaluate and frame it. In contrast, analyses of production processes, i.e., how different stakeholders try to position themselves in media coverage and on media consumption and its effects are less common.

The following article will focus on these subdimensions. It will organize the literature into three distinct sections: agenda-building efforts related to climate change, the role of journalists and the characteristics of the respective media representations, and their uses and effects.

Stakeholder Communication about Climate Change

The existence of climate change itself, its extent and urgency, as well as the appropriate responses and measures toward it are "deeply contested[, with] considerable competition among (and between) scientists, industry, policymakers and non-governmental organizations (NGOs), each of whom is likely to be actively seeking to establish their particular perspectives on the issues" (Anderson, 2009: 166). Using different kinds of strategic communication, these stakeholders often aim for media representation because media are the main forum for public debates in modern societies and important sources for information about climate change for many people. Therefore, stakeholders try to position themselves in the media in order to be seen as relevant and viable actors in the field, and to inject their positions and views into coverage in order to influence the societal perceptions of climate change and the subsequent decisions about paths of action.

A considerable number of studies from political science, sociology, communications, and public relations have analyzed these efforts using structuration theory, agenda-building theory, theories of the public sphere, or social movement theories (for overviews see Hall and Taplin, 2007; McCright and Dunlap, 2003). They have shown that many concerned parties exist in the case of climate change, which devote large amounts of monetary, personal, cultural, and/or symbolic resources to the issue. The strategic communication of these stakeholders can be organized in different 'frames,' i.e., ways of interpreting climate change from a certain perspective that correspond with preferences for specific solutions (Table 1).

Anthropogenic Climate Change as a Global Problem

The first frame, which already emerged in the early 1990s and has been powerful ever since, emphasizes the existence of anthropogenic climate change, and advocates various measures to fight it. Its proponents consist of renowned individual scientists such as James Hansen (US), Nicholas Stern (UK), and Hans Joachim Schellnhuber (GER), prominent activists like Al Gore, hybrid scientific/political organizations such as the United Nations' Framework Convention for Climate Change's (UNFCCC) Intergovernmental Panel on Climate Change (IPCC), large international environmental NGOs such as Greenpeace or the World Wildlife Fund, and, in many countries, green parties. These stakeholders use their considerable resources to argue that the "warming of the climate system is unequivocal, as is now evident," and "very likely due to the observed increase in anthropogenic [greenhouse gases]" (IPCC). It is further emphasized that these developments lead to more extreme weather, intensify resource shortages, and are "already harming people and ecosystems" (Greenpeace). Therefore, the need for action to counteract climate change is emphasized, such as the '2-degree goal' to limit global warming at 2 °C compared to preindustrial levels.

Scientific Uncertainty

In the mid-1990s, and particularly among representatives of the US and Australian oil and automobile industries, a counterframe to the abovementioned perspective emerged that has also been used in other science-related public debates, for example, around the effects of smoking and secondhand smoking or genetic engineering: the 'scientific uncertainty' frame. It was sponsored by international corporations like Shell or Elf as well as their representatives such as the 'Global Climate Coalition,' conservative think tanks, and various administrations and political parties (especially in the US and in Australia). The frame was less prominent in Europe and other countries and has lost importance since the 2000s. The frame argues that while global warming may exist, "Scientists do not agree on man's effect on the climate" (Citizens for a Sound Economy Foundation). Those scientists who argue otherwise are portrayed as 'alarmists,' and their work is discredited as 'junk' or 'pseudo' science that has a 'political purpose' (Global Climate Coalition). If anything, a proper response to the alleged state of affairs would be more research, but any political regulation is rejected as being premature.

Economic Development

A third frame emerged in the wake of negotiations leading up to the Kyoto Protocol. It accepts global warming as existent and human activity as its major cause, but stresses the high economic costs of action. "We are now trying to be more positive with the science, while still pointing to the high cost of precipitate action" (Ford), which might "cost American families billions of dollars and millions of jobs" (Global Climate Coalition) and "damage competitiveness" (Australian Industry

Table 1 Frames of climate change and their sponsors

Frame	*Anthropogenic climate change as a global problem*	*Scientific uncertainty*	*Economic development*	*Ecological modernization*
Central organizing idea	Scientific evidence shows climate change exists and is man-made; action is needed	Scientific evidence about climate change is inconclusive; therefore, action would be hasty	Climate change does exist, but measures will seriously hamper (domestic) economic development	Technological development is key to fighting climate change, industry, and developed countries as pioneers
Framing devices	Scientific evidence points to 'climate catastrophe'; atmosphere as global 'common good'; historical responsibility of developed countries; global 'climate justice'	Climate science as 'alarmist' and 'junk science'; its background as a political complot; climate change is 'a theory, not a fact'	Climate change regulation leads to 'crippled economy' and 'restricted lifestyle' as well as a 'transfer of wealth'	Corporations acknowledge 'responsibility,' reframe them as 'business opportunities' for 'bridging technologies'; 'Ecoimagination'
Main frame sponsors	Prominent individual scientists and activists, scientific institutions and organizations, IPCC, environmental NGOs, green parties	Fossil fuel, coal, automotive, and electric utilities industry, their associations; think tanks, conservative politicians, especially in the US	US fossil fuel, coal and automotive industry, Australian coal industry, conservative politicians	European multinational corporations, later adopted by many industry sectors in many world regions
Occurrence	Emerged in 1990s and is still widely used	Early to mid-1990s, relatively widespread geographically, most strongly present in Anglo-Saxon world	From Kyoto negotiations in 1997 to early 2000s	Started around Kyoto negotiations, first in Europe; strong in many countries since mid-2000s

NGO, non-governmental organizations; IPCC, intergovernmental panel on climate change.
Sources: Schlichting, I., 2013. Strategic framing of climate change by industry actors: a meta-analysis. Environmental Communication: A Journal of Nature and Culture 7 (4), 493–511; Schmidt, A., Schlichting, I., in press. In: Hemmer, I., Müller, M. (Eds.), Sustainability and Climate Change: Interpretations and Claims by Societal Actors from Germany, India and the United States. Oekom, Rio+20, Munich with additions from McCright, A.M., Dunlap, R.E., 2003. Defeating kyoto. The conservative movement's impact on US climate change policy. Social Problems 50 (3), 348–373; Nisbet, M.C., 2009. Communicating climate change: why frames matter for public engagement. Environment 51 (2), 514–518; Weingart, P., Engels, A., Pansegrau, P., 2000. Risks of communication. Discourses on climate change in science, politics, and the mass media. Public Understanding of Science 9 (3), 261–283; and from own research.

Greenhouse Network). Accordingly, it is argued that political measures have to take this into account, and that the responsibility for tackling climate change cannot solely lie on industrialized, developed countries. To be as effective as possible, it calls for a global effort that included countries like China, India, Brazil, and others. Accordingly, international agreements are rejected as long as they do not include all countries. This frame is especially prominent among political and economic factors from developed countries, especially from countries with a strongly fossil-dependent economy like Australia or Norway, but it also includes the US.

Ecological Modernization

The fourth frame acknowledges climate change as "the single biggest environmental challenge facing the planet" (Hong Kong Shanghai Bank of China), but focuses on technological development as the key to fighting climate change and describes industry actors and developed countries as pioneering this development. It highlights the potential of an ecological modernization to create jobs and national welfare. Accordingly, the frame resonated not only among industry representatives from BP, Shell, other companies, and conservative politicians, but also among trade unions and NGOs, in civil society and among liberal-to-left politicians like the Blair administration in the UK.

In some variation, these frames are prevalent among many stakeholders and in many societies. In some countries, particularly Anglo-Saxon ones, stakeholders fall into two fiercely competing camps of climate change: proponents and skeptics. Whereas in countries like Germany, the Netherlands or India, the different frames and their sponsors are found in a less conflictive situation.

'Climate Journalists' and Their Role

Agenda-building efforts directed at the media are not always, and not equally, successful. Publishing houses and journalists play important roles as an intermediaries in this respect. Being seen by journalism theory as 'gatekeepers' of the news flow, they ultimately decide which topics, stakeholders, and perspectives are represented in the media. Accordingly, the role of journalists in climate change communication has received some scholarly interest (for overviews see Bodker and Neverla, 2013; Brüggemann and Engesser, 2014). Most of these studies are content analyses that measure journalistic output and extrapolate information about the journalists from there, while surveys among journalists and in-depth ethnographic studies in newsrooms are largely missing.

Who 'Climate Journalists' Are

Even though little is known about the sociodemographic characteristics and attitudes of 'climate journalists,' studies have shown that they are a diverse group of professionals and

that their positions within the media differ strongly. As climate change is a cross-sectional topic that includes scientific as well as political, economic, cultural, and other facets, the issue is taken up by numerous media desks and their journalists. As a result, most journalists writing about climate change are not monothematic experts on this issue. Still, 'climate journalists' seem to have become relatively knowledgeable about the issue over time. While an early US journalist survey showed considerable gaps in their knowledge about climate change, later studies came to different results and documented that journalists' (average) knowledge about the issue is better than that of decision-makers, NGO representatives and others. Moreover, recent findings indicate that most 'climate journalists' are in line with the IPCC view on climate change.

How 'Climate Journalists' Work

Some aspects of climate journalists' work differ strongly from country to country. The sources they tend to use, for example, have been shown to differ considerably according to varying national cultures of journalism – in countries such as Germany and India, scientists are the major sources of climate change coverage, whereas in developing countries media and news agencies from other parts of the world and domestic political elites seem to be more important. Other aspects of journalists' work seem to be rather stable across countries: journalists' working routines consist, among other things, of finding newsworthy topics and highlighting their newsworthiness in their coverage, while adhering to general journalistic norms such as thoroughly researching information and providing a balanced account of an issue to the readers. It has been shown that these general characteristics also apply when journalists cover climate change, but that they sometimes lead to problematic results. On the one side, because the unobtrusive and complex issue of climate change does not cater well to news values, journalists have to find and emphasize its newsworthy aspects in order to cover it. Therefore, they occasionally tend to play down or omit uncertainties connected to scientific information to comply with the news value of 'unambiguity'; or to exaggerate the expected negative outcomes of climate change to comply with the 'negativity' news value. Accordingly, newspaper headlines like "Our Earth is Dying" (German tabloid 'Bild,' 3 February 2007) or, in turn, "Climate Hysteria is a Scam" (Australia's Narrogin Observer, 30 June 2010) have been criticized as being sensationalist and biased. On the other side, established journalistic norms have been shown to be counterproductive at times when covering climate change. The prime example is Boykoff's 'balance as bias' thesis, which argues that the norm of balanced reporting led to a coverage in US media in the 1990s where a minority of climate skeptics was portrayed as being equally important as the (large majority of) proponents of anthropogenic climate change (cf. Boykoff, 2011: esp. 99ff.).

Media Representations of Climate Change

Analyses of media representations of climate change are most prominent in the respective literature, mostly coming from communication and political sciences as well as sociology. These studies cover a large number of different countries – even though still primarily from the Western world – different media outlets, many years, and several facets of climate change. Their main findings are discussed in the following sections.

Climate Change Is a Relevant Media Issue Worldwide

Many studies have focused on the amount of attention media devote to climate change. They show that the issue is a relevant topic in many countries on all continents, outnumbering other hotly debated science issues like stem cell research or genome sequencing. They also demonstrate that media attention for climate change has risen considerably in many countries since the mid-2000s (Schmidt et al., 2014; see **Figure 1**).

Media Coverage Is Event-Driven

In addition, studies have shown that media attention for climate change, as for many other issues as well, strongly fluctuates over time and that it peaks around specific events (Schmidt et al., 2014). Some of these are weather events, particularly extreme weather; coverage about climate change peaked around Hurricane Katrina, the 'flood of the century' in Germany and the heat waves in Europe. Journalists, looking for newsworthy angles of climate change and visible manifestations of the invisible phenomenon, occasionally interpret such events as symptoms for climate change, even though these claims often lack a scientific basis. Other events triggering climate change coverage are sociopolitical ones. Particularly influential are the proceedings of international climate politics, namely the Conferences of the Parties to the UNFCCC process, where international climate politics is debated by high-ranking politicians, and where numerous large NGOs provide additional interpretations and photo ops for journalists. These high-profile political events occur annually, can be prepared for, and systematically focus large amounts of media coverage on climate change and climate politics. But media attention for climate change has also been triggered by the recurring releases of the IPCC Assessment Reports or the Stern Review on the economic impacts of climate change, as well as by cultural events such as the 'Live Earth' concerts and movies like 'An Inconvenient Truth' or 'The Day after Tomorrow' (Schäfer et al., 2014).

Coverage Has Moved from the Science of Climate Change to Its Societal Impacts

Longitudinal content analyses in various countries demonstrate that the placement of climate change in media coverage has changed over time. While it was mostly dealt with on the science pages (or segments) in the early 1990s, the issue moved from being a primarily scientific issue into other sections of the media, most notably to politics and the economy. This is a symptom of a changing focus – a 'societal turn' – in climate change coverage. While the scientific basis of climate change is still being reported upon at times (and, especially in Anglo-Saxon countries, sometimes debated; see Painter and Ashe, 2012), its relative importance has decreased, while climate change has increasingly become a political, economic, and social issue. Aspects like the political efforts toward a global

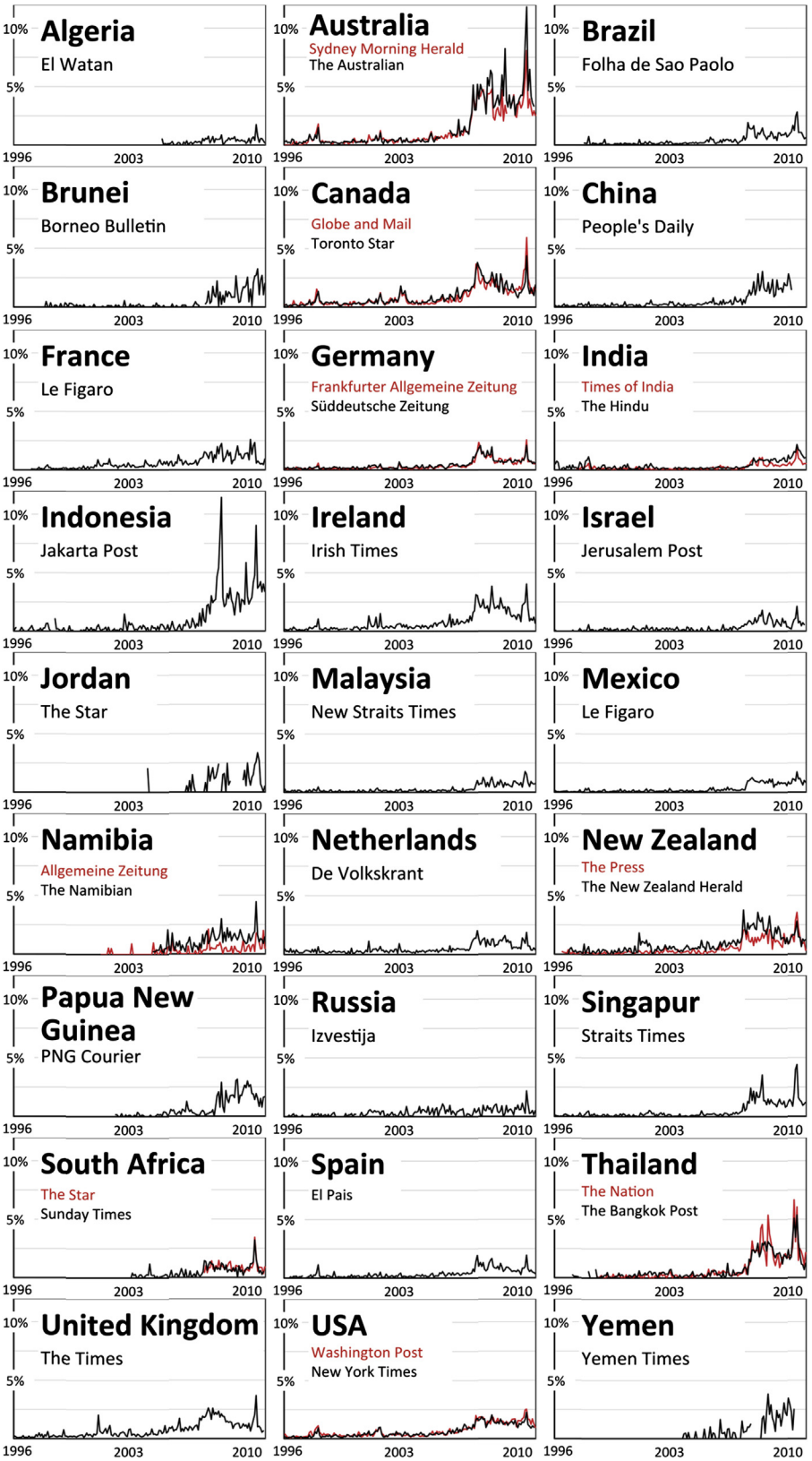

Figure 1 Media attention for climate change in 27 countries worldwide. For each country, one or two leading print media representing different political positions were chosen for analysis. Leading print media were defined as national media with a high circulation, being published daily and widely regarded as high-quality papers. The final sample of the analysis contained 153 261 articles. Percentages indicate the amount of articles dealing with climate change in relation to all articles published in the analyzed newspapers per month; 1996–2010; gaps in graphs are due to missing data. Source: Schmidt et al., 2013.

treaty on emissions reductions, the costs of moving toward low-carbon societies for national economies, or the humanitarian and security implications connected to potentially intensifying resource conflicts and climate-induced migration, are increasingly taking center stage when media cover climate change.

The Content of Climate Change Coverage Differs Considerably from Country to Country

The increased focus of climate change coverage on societal impacts also seems to come with an increasing 'domestication' of the issue. While the timing and, to some extent, the occasions of coverage are still quite similar internationally – particularly among Western European and North American countries – the content of coverage often differs considerably. While coverage in the Netherlands, for example, focuses more on technical solutions in line with the described 'ecological modernization' frame, Swedish media emphasize moral and ecological aspects to a larger extent. German media display rather 'warmist' standpoints similar to the 'anthropogenic climate change as a global roblem' frame, whereas media coverage in the US and Australia is stronger polarized between skeptics and warmists who devote more attention to the 'scientific uncertainty' frame. In countries beyond the Western world, such as India or Bangladesh, postcolonial arguments can be found, which interpret climate change as mainly of Western concern (e.g., Billett, 2010).

Uses and Effects of Media Representations of Climate Change

Compared to the analyses of media representations of climate change themselves – which are often justified with their alleged effects on people's issue awareness, knowledge, or behavior – their actual uses and effects have received less scholarly attention (cf Taddicken, 2013). Furthermore, scholarship on these issues has many gaps: Studies tend to focus on cognitive effects, but not on affective and other effects; on information-oriented formats and not on entertaining or fictional ones; on print media and not on TV (which has been shown to have the strongest effects in general and to be the main source of people's information about climate change); on individual level effects and not on meso (institutional) and macro (societal) effects; and on the general public and not on special audiences such as scientists, politicians, or journalists. In addition, the results of the few existing studies paint an unclear, incoherent, and sometimes contradictory picture. Nevertheless, a number of recurring results can be extracted.

Mass Media Are Important and Trusted Sources for Climate Change Information

Regarding the uses of media representations of climate change, it has been shown repeatedly that mass media – particularly television and newspapers, but increasingly the Internet as well – are the most important sources for people's information about climate change. Furthermore, mass media also seem to be credible sources, whom people trust more on the issue than, for example, family and friends. Beyond these rather basic facts, however, little is known about uses and reception processes. The social contexts in which users engage with climate-related media content, their motivations, and gratifications as well as their interpretations of the content have not yet been taken up properly. Motivational theories like uses-and-gratifications and theories of interpersonal communication and information diffusion could be useful to shed light on these facets.

Agenda-Setting Effects on the Audience's Issue Awareness Exist

Studies on the effects of media representations of climate change have mainly focused on individual level effects on people's problem awareness, their level of information, and their willingness to act, i.e., on potential first-level agenda-setting effects and on persuasion. They have demonstrated that the amount of climate change representations that people encounter in mass media is linked to their awareness of the issue, pointing toward an agenda-setting effect. This result has mainly been established through surveys of people's self-assessed media use, which are rarely supplemented with analyses of actual media content or use. Future studies should take these factors into account, and also aim to distinguish the specific contribution of different media with regards to agenda setting.

Mass Media Use Increases Knowledge about Climate Change

Cognitive effects of media representations of climate change have also been repeatedly shown. People acquire information about the issue from the media and learn something about it. This includes factual information about the phenomenon itself – the extent and causes of climate change – as well as potential solutions and options for (individual) action (Stamm et al., 2000; Taddicken, 2013). These processes of information acquisition have been well established for cases in which people turn to media specifically for information about climate change. Their existence and extent is less certain for general, unspecific media use, i.e., in cases where people use media habitually and come across climate change information involuntarily.

Behavioral Effects Are Less Clear

The existence of media effects on climate change-related behavior has not yet been established. Studies have found effects of media use on people's information-seeking; those using media and learning from them are more likely to search for more information about climate change in the future (Zhao, 2009). Regarding climate-related behavior and action, however, only some weak media effects have been found, and in practically all cases refer to behavioral intentions rather than actual behavior. Even in these cases, the evidence is mixed: whereas effects could be shown for the intention to engage in 'societally relevant activities,' no media influence could be shown for self-reported investment decisions or lifestyle changes (Arlt et al., 2011). Therefore, more robust studies in longitudinal behavioral effects are necessary.

Outlook

While media portrayals of climate change have been taken up in many studies, scholarship still has many gaps. Analyses of the production, presentation, and consumption of these media portrayals need to be balanced better in the future, with a particular emphasis on the uses and effects of media portrayals of climate change. More studies on media uses and effects could also help assess, and maybe answer, normative calls for social sciences to play a role in the transition toward low-carbon societies. In order to do so, the objects of analysis, namely the countries and media, which are analyzed should be more diverse, with more studies taking up non-Western countries and audiovisual and online media. As climate change is a transnational phenomenon affecting countries around the world, it would also be worthwhile to engage in more comparative research in order to better understand the (potentially) varying perspectives on the issue, and to look for a potential transnationalization or globalization of the respective debates.

See also: Agenda Setting, Media Effects on; Audiences, Media; Environmental Movements; Environmental Sciences; Mass Media, Representations in; Media Effects; Media and Social Movements; News: General; Public Sphere and the Media; Science Communication; Science and Media.

Bibliography

Anderson, A., 2009. Media, politics and climate change. towards a new research agenda. Sociology Compass 3 (2), 166–182.

Arlt, D., Hoppe, I., Wolling, J., 2011. Climate change and media usage. Effects on problem awareness and behavioural intentions. International Communication Gazette 73 (1–2), 45–63.

Billett, S., 2010. Dividing climate change. global warming in the Indian mass media. Climatic Change 99 (1), 1–16.

Bodker, H., Neverla, I., 2013. Environmental Journalism. Routledge, London.

Boykoff, Maxwell T., 2008. Lost in translation? United States television news coverage of anthropogenic climate change, 1995–2004. Climatic Change 86, 1–11.

Boykoff, M.T., 2011. Who Speaks for the Climate? Making Sense of Media Reporting on Climate Change. Cambridge University Press, Cambridge.

Brüggemann, M., Engesser, S., 2014. Between consensus and Denial: climate journalists as interpretive community. Science Communication. Online first under. http://scx.sagepub.com/content/early/2014/05/20/1075547014533662.abstract.

Carvalho, A., 2010. Media(ted)discourses and climate change. A focus on political subjectivity and (dis)engagement. WIREs Climate Change 1 (2), 172–179.

Gavin, Neil T., Leonard-Milsom, Liam, Montgomery, Jessica, 2011. Climate change, flooding and the media in Britain. Public Understanding of Science 20 (3), 422–438.

Hall, N.L., Taplin, R., 2007. Revolution or inch-by-inch? Campaign approaches on climate change by environmental groups. The Environmentalist 27 (1), 95–107.

McCright, A.M., Dunlap, R.E., 2000. Challenging global warming as a social problem. An analysis of the conservative movement counter-claims. Social Problems 47 (4), 499–522.

McCright, A.M., Dunlap, R.E., 2003. Defeating kyoto. The conservative movement's impact on US climate change policy. Social Problems 50 (3), 348–373.

Moser, S.C., 2010. Communicating climate change: history, challenges, process and future directions. WIREs Climate Change 1 (1), 31–53.

Nisbet, M.C., 2009. Communicating climate change: why frames matter for public engagement. Environment 51 (2), 514–518.

Olausson, U., Berglez, P., 2014. Media and climate change: four long-standing research challenges revisited. environmental communication. A Journal of Nature and Culture 8 (2), 249–265.

O'Neill, S.J., Boykoff, M.T., Niemeyer, S., Day, S.A., 2013. On the use of imagery for climate change engagement. Global Environmental Change 23 (2), 413–421.

Painter, J., Ashe, T., 2012. Cross-national comparison of the presence of climate scepticism in the print media in six countries, 2007–10. Environmental Research Letters 7 (4), 1–8.

Schäfer, M.S., 2012. Online communication about climate change and climate politics. A literature review. WIREs Climate Change 3 (6), 527–543.

Schäfer, M.S., Ivanova, A., Schmidt, A., 2014. What Drives media attention for climate change? Explaining issue attention in australian, german and Indian print Media from 1996 to 2010. International Communication Gazette 76 (2), 152–176.

Schäfer, M.S., Schlichting, I., 2014. Media representations of climate change. A meta-analysis of the research field. Environmental Communication: A Journal of Nature and Culture 8 (2), 142–160.

Schlichting, I., 2013. Strategic framing of climate change by industry actors: a meta-analysis. Environmental Communication: A Journal of Nature and Culture 7 (4), 493–511.

Schmidt, A., Ivanova, A., Schäfer, M.S., 2014. Media attention for climate change around the world. A comparative analysis of newspaper coverage in 27 countries. Global Environmental Change 23 (5), 1233–1248.

Schmidt, A., Schlichting, I., in press. In: Hemmer, I., Müller, M. (Eds.), Sustainability and Climate Change: Interpretations and Claims by Societal Actors from Germany, India and the United States. Oekom, Rio+20, Munich.

Stamm, K.R., Clark, F., Reynolds Eblacas, P., 2000. Mass communication and public understanding of environmental problems: the case of global warming. Public Understanding of Science 9, 219–237.

Taddicken, M., 2013. Climate change from the user's perspective: the impact of mass media and internet use and individual and moderating variables on knowledge and attitudes. Journal of Media Psychology 25 (1), 39–52.

Weingart, P., Engels, A., Pansegrau, P., 2000. Risks of communication. Discourses on climate change in science, politics, and the mass media. Public Understanding of Science 9 (3), 261–283.

Zhao, X., 2009. Media use and global warming perceptions – a snapshot of the reinforcing spirals. Communication Research 36, 698–723.

Climate Change: Economics

Gary W Yohe, Wesleyan University, Middletown, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

There are two sides to the economics of climate change – the economics of the economic damages, risks, and opportunities associated with the impacts of climate and the economics of the benefits and costs of mitigation. The benefits of mitigation are generally estimated as net damages that can be avoided or delayed by actions designed to reduce emissions of greenhouse gases of all kinds. All of these calculations are uncertain given our inability to understand fully the science of, to comprehend fully the links between global change and local impacts, to translate rigorously impacts into economic consequences that account adequately for adaptation, and the necessity to look many decades into the future. This article reports briefly in successive sections on the economic costs associated with climate change and the economic costs of mitigation, and it concludes with some coverage of caveats and open research questions as of the fall of 2013.

There are, fundamentally, two sides to the economics of climate change – the economics of the economic damages of the impacts of climate change (most appropriately calculated in terms of residual damages after adaptation, but including the cost of adaptation) and the economics of the benefits and costs of mitigation (most appropriately calibrated in terms of the net benefits from reduced damages and the cost of mitigation polices taking account of the design and implementation of those policies); that is, the benefits of mitigation are generally estimated as net damages that can be avoided or delayed by actions designed to reduce emissions of greenhouse gases of all kinds (carbon dioxide, methane, fluorocarbons, black carbon, etc.).

Despite years of concentrated work around the world, all of these calculations are highly uncertain given our inability to understand fully the science of change (the uncertain link between cumulative emissions and changes in temperature), to comprehend fully the links between global change and local impacts, to translate impacts into economic consequences that account adequately for adaptation, and the necessity to look many decades into the future. Perhaps most importantly, these calculations are complicated by the imperfections of economic and social institutions that make assumptions that support estimates in a first best world tenuous at best – except that first-best estimates indicate what could be accomplished if everything worked perfectly. First-best analyses therefore play the same role in the climate change arena as much as standard economic analyses of this or that conducted under the assumption that all markets are perfectly competitive and all negative (or positive) externalities are internalized (see Ebi and Yohe, 2013)

The Economic Cost of Climate Change Impacts

Estimates of the economic damages from climate change have been reported in three different ways. In the first, damages were computed as a percent of gross domestic product (GDP) for a specified rise in global mean temperature. The second tracked damages along specific scenarios that define future socioeconomic and climatic conditions and reports their present value. The third built on a marginality concept called the social cost of carbon (SCC) – estimates of the discounted damages estimated to be attributable to an extra ton of carbon emissions. Researchers calculate SCC by summing the discounted value of extra damage caused each year for as long as it remains in the atmosphere – a process, which requires a model of atmospheric residence time in addition to a sophisticated economic model. The results of the second two approaches are highly dependent not only on scientific parameters like climate sensitivity (the warming associated with a doubling of atmospheric concentrations from preindustrial levels) and atmospheric residence time, but also on assumptions about how to discount the future, how to reflect attitudes toward risk, and how to incorporate attitudes toward income inequality across the globe.

The history of damage estimates begins with the first of these approaches. In 1996, the Intergovernmental Panel on Climate Change (IPCC) reported preliminary estimates of the annual economic impact of a doubling of concentrations of greenhouse gases (~550 ppmv) and an associated 2.5 °C increase in global mean temperature (IPCC, 1996). The estimates reported for the United States, for example, ranged from 1 to 2.5% of anticipated GDP in 2065, respectively (see Nordhaus, 1991 and Titus (1992), for example). For developed countries more generally, damage estimates for the same time period on the order of 1% of GDP were also reported, but larger percentages were projected for developing countries. As a result, global mean damages were thought to lie in the range of 1.5–3.5% of global GDP. The year 2065 was chosen as a benchmark because that was when the specified doubling of concentrations was anticipated to occur. It is interesting to note that IPCC (2001) reported almost exactly the same results.

Stern (2007) was a notable contribution to the literature spawned by the second approach because of its comprehensive coverage of available impact studies, its treatment of uncertainty and risk, its reliance on utility-based damage estimates, its expanded time horizon (the models were run through 2200), and its assumed very low discount rate. Four reported estimates of the economic effects of unmitigated climate change were expressed in terms of equivalence to a persistent average reduction in global per capita consumption; they ranged from 5% to 20% from the present into the indefinite future.

Early calculations based on the third approach – the SCC – were reported in IPCC (1996) wherein estimates ranged from $5 to $125 per ton of carbon in 1990 dollars. As reported in IPCC (2007a), Tol (2005) gathered over 100 estimates of the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.91078-X

SCC from 28 published studies. His survey showed a median estimate of $14 per ton of carbon, a mean of $93 per ton, and a 95th percentile estimate equal to $350 per ton. His survey also showed that 10% of the estimates were negative (for low climate sensitivity and underlying estimates of benefits to modest warming in some sectors). Perhaps more importantly, the 90th percentile SCC estimates were $60 per ton for studies that assumed a 3% pure rate of time preference (ppt defined as the minimum rate at which present consumption would be traded for consumption in 1 year's time) and more than $160 per ton from those who assumed a 1% ppt. The equivalent calculation for the low Stern (2007) runs produced an estimate of more than $1600 per ton.

Clearly, SCC estimates are sensitive to the pure rate of time preference; but they are also highly sensitive to other equity weighting of regional damages. Watkiss (2011) indicated, on the basis of runs of Tol's own integrated assessment model, that adding utility-based weighting would essentially double the median estimate of SCC for any ppt between 0% and 3%.

Turning now to other greenhouse gases for which there are fewer estimates, **Table 1** records mean and 5th–95th percentile ranges from Hope (2006b). It should also be noted (see Watkiss et al., 2005) that the social cost of all greenhouse gases are expected to rise over time for unmitigated futures – for example, SCC by 2–4% per year and methane by 4–8% per year (because of its shorter atmospheric lifetime).

The Economic Cost of Mitigating Climate Change

The economic cost of mitigation depends critically upon the pace and flexibility of underlying economic activity, its geographic distribution, the rate at which the future is discounted and its intertemporal distribution. Mitigation costs also depend on the design of the policy or regulatory interventions themselves; so they depend critically on the political economy of international negotiation and internal national decision-making procedures. Dynamic and spatial efficiency can reduce the cost of limiting atmospheric concentrations of greenhouse gases by 50–80%, but excessive near-term limitations of emissions can exacerbate long-term costs by comparable proportions.

Many researchers have investigated the cost of climate change mitigation. They have, in particular, focused their attention on energy consumption and the resulting emission of carbon dioxide and other greenhouse gases. Carbon dioxide, for example, is a product of burning fossil fuel, and its emission varies from fuel to fuel. Burning coal emits 25% more carbon per unit energy than burning oil, and burning oil emits 43% more carbon than natural gas. Burning hydrogen emits no carbon. Hydroelectric, wind, solar, and nuclear power are similarly carbon-free sources of energy. Mitigation therefore involves reducing energy consumption where appropriate *and* substituting carbon-free sources of energy for carbon-based fuels *and* low-carbon fossil fuels such as natural gas for high-carbon fuels such as coal. Our ability to effect and to sustain this sort of conservation and/or substitution over the very long run depends upon the availability of new technology, the supply of low-carbon and carbon-free sources of energy, as well as the political will to provide incentives, when necessary, to promote both.

Table 1 Estimates of the social cost of three greenhouse gases (in $ per ton) from a single model (PAGE 2002)

Gas	*Mean*	*5th Percentile*	*95th Percentile*
Methane	$105	$25	$250
SF6	$200 000	$45 000	$450 000
Carbon dioxide	$14	$4	$50

Source: Hope, C., 2006. The marginal impacts of CO2, CH4 and SF6 emissions. Climate Policy 6, 537–544.

The most effective means of conveying the cost of mitigation working through either route is to track estimates of the marginal cost of emission reductions. These estimates are highly model specific and they depend on the mix of energy sources assumed or determined endogenously. They all show that marginal costs increase at an increasing rate, as driven by conservation and fuel switching. **Figure 1**, included below in Section How can Mitigation Targets be Framed with such Limit Information about Damages? to illustrate a different point, clearly displays this observation for a specific example.

Article 2 of the United Nations Framework Convention on Climate Change (UNFCCC) committed the globe in 1992 to holding concentrations of greenhouse gases below levels that would imply 'dangerous anthropogenic interference with the climate system'. The precise concentration target that corresponds with this imperative has not yet been identified to everyone's satisfaction, even though two degrees has been articulated as an appropriate target (the association with a concentration target is uncertain – see NRC (2010a)). As a result, many researchers have investigated the economic cost of reducing emissions over the next 100 years so that concentrations do not exceed a range of thresholds. Manne and Richels (1997) estimated that the cost of achieving the most popular threshold among academic contributors, 550 ppmv (since lower targets now seem to be almost impossible to achieve), could be as high as $3.5 trillion (1990 dollar) or as low as $650 billion (1990 dollar). Estimates from other researchers were comparable then, and still are; but they differed from one another in specific detail for the reasons as mentioned above; again, see **Figure 1**.

Cost estimates for meeting concentration thresholds also depend critically on the location and timing of each unit of emission reduction. Wigley et al. (1996) have called this the 'where' and 'when' flexibility components of cost. Their 'WRE' results emphasize that costs would be minimized if each ton of emission reduction were taken from the least costly source regardless of where it is located; this implies that the marginal cost of the last ton of greenhouse gas emission removed from the effluent of every source should be equal. The Kyoto Protocol from the third Conference of the Parties of the UNFCCC had this goal in mind with its emission reduction targets for each country, but the inability of international negotiations to implement those plans has lead to a collection of independent bilateral agreements and regulations that do not guarantee 'where' efficiency.

The WRE results also demonstrated that emissions should be reduced at any point in time only if the discounted value of the associated cost were in line with all the discounted costs of all other reductions at all other times. These intertemporal

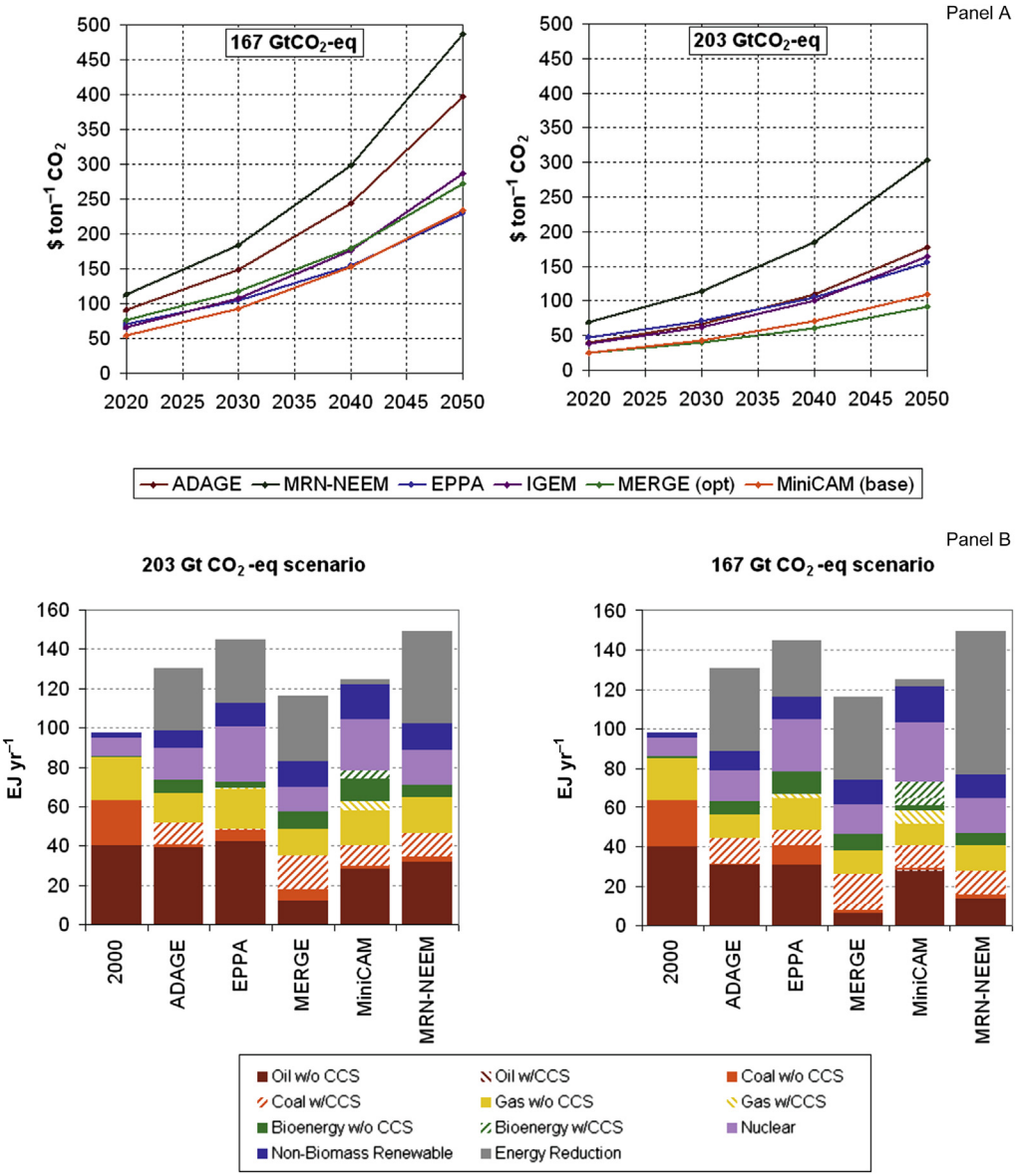

Figure 1 Panel A (top panel) shows trajectories for the prices of carbon (the marginal cost of emission reduction) that are consistent with achieving cumulative emission goals (increasing reductions over time against a baseline of no intervention) for six different integrated assessment models; each trajectory supports the model specific projected energy source mixes displayed in Panel B (bottom panel). Source: Figures 2.12 and 3.3 in NRC, 2010b. Limiting the Magnitude of Climate Change. National Academies Press, Washington, DC. http://www.nap.edu/catalog.php?record_id=12785, derived from Fawcett, A.A., Calvin, K.V., De La Chesnaye, F.C., Reilly, J.M., Weyant, J.P., 2009. Overview of EMF-22 U.S. transition scenarios. Energy Economics 31 (Suppl. 2), S198–S211; see www.emf.stanford.edu.

'WRE' results were controversial when they were published because 'when' flexibility implies that early reductions in emissions should be small and grow over time. The intuition behind this result is a product of the basic science of climate change. Increases in global mean temperature depend entirely on the concentrations of greenhouse gases and concentrations depend entirely on *cumulative* emissions (NRC, 2010a). As a result, determining how to achieve any concentration target at least cost is equivalent to determining how to allocate the consumption of a scarce resource (total carbon emissions over a specific or indefinite time period) optimally over time. This is the Hotelling (1931) problem for which the solution is widely known – calculate a scarcity rent and allow it to rise over time at the rate of interest so that consumption gradually falls (i.e., in this application, emission reductions climb).

Some Caveats and Open Questions

Because the analysis of climate change damages and opportunities continue to evolve and will do so for the foreseeable future, this article should be read as a brief snapshot of the state

of play as of the summer of 2013. It is fitting, therefore, to end with some coverage of issues that are on the forefront of current work.

What Is Missing from the Damage Estimates?

As noted above, current estimates of the SCC vary widely in response to the parameterization of the climate and economic systems as well as parameters that reflect normative judgments on time preference, risk aversion, and inequality aversion. This is problematic, but reasonably well understood. It must be emphasized, however, that current estimates are plagued by incomplete coverage of all of the possible effects of climate change. Figure 2 shows that many estimates of market damages for gradual climate change as well as some estimates for bounded risk and system surprises exist, and so these inform SCC calculations. A smaller number of estimates for nonmarket damages for the first two sources of risk also exist, but there is almost nothing in the literature on system surprises for this category of impact. Calculations of the SCC cannot, however, include any rigorous estimates of socially contingent damages because there are none. It follows that all SCC calculations systematically underestimate the true marginal cost of carbon, and so it also follows that the value of information derived from future research along the right-hand column and the bottom row of Figure 2 would be very high.

How Can Mitigation Targets Be Framed with such Limit Information about Damages?

Nordhaus (1991) was the first researcher to weigh the present value of the benefits of mitigation policy against the present value of their costs to compute an economically optimal policy trajectory over the long term. His results were based on the anticipation that impacts would be smooth and amount to roughly 1% of gross world product if the global mean temperature rose by 2.5 °C. Corroborated by his own subsequent work and parallel work by others, optimality calculations in the first-best deterministic analyses support modest early intervention followed by smooth and gradual tightening of emission restrictions. Indeed, most optimality exercises propose pricing carbon at between $10 and $40 per ton in 2013. All report that carbon prices should increase over time at about 3% per year (a reflection of the Hotelling result mentioned above), but none stabilize concentrations any time within the next 150 years.

As noted above, first-best calculations are suggestive, but they miss uncertainty and they miss everything that was identified in Figure 2 as an omission in the knowledge base. IPCC (2007b) includes the following language in response to this unsettling observation: "Responding to climate change involves an iterative risk management process that includes both adaptation and mitigation and takes into account climate change damages, co-benefits, sustainability, equity, and attitudes toward risk" (pg. 22). These words were unanimously agreed upon word for word by all of the countries that participate in IPCC (essentially all of the countries who have signed onto the UNFCCC).

NRC (2010b) shows how this language can be put into practice. Authors there assumed that 550 ppm was an appropriate long-term target given current information about associated economic and social risk, and they contemplated how the United States would participate over an initial 40-year time period from 2010 until 2050. The idea is that an adjustment in

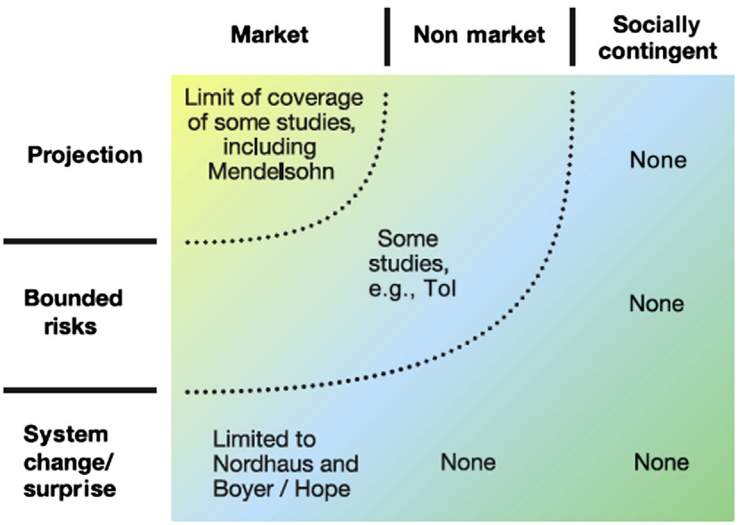

Figure 2 A schematic of coverage of studies that support calculations of the social cost of carbon against source of uncertainty and socio-economic context. There are many studies in the upper left, but essentially none for combinations that would capture catastrophes (low probability – high consequence possibilities) and other extreme events (that are already being observed and, to some degree, attributed). Source: Figure 20.4 in IPCC, Yohe, G.W., Lasco, R.D., Ahmad, Q.K., Arnell, N., Cohen, S.J., Hope, C., Janetos, A.C., Perez, R.T., 2007a. Perspectives on climate change and sustainability. In: Parry, M.L., Canziani, O.F., Palutikof, J.P., Hanson, C.E., van der Linden, P.J. (Eds.), Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, from Watkiss, P., Anthoff, D., Downing, T., Hepburn, C., Hope, C., Hunt, A., Tol, R., 2005. The Social Costs of Carbon (SCC) Review: Methodological Approaches for Using SCC Estimates in Policy Assessment. Final Report, Defra, p. 124.

the long-run concentration target would occur in 2050 on the basis of new science, evidence of intervening warming, evidence of economic damages distributed around the world, and evidence about the success of the global community to achieve its collective goal. In practical terms, it meant that the globe would have a cumulative emissions budget through 2050, and so the response problem in every country was to minimize the cost of holding cumulative emissions below their share of this total budget.

This allocation exercise led to the expectation that U.S. emissions should sum to something between 167 and 203 gigatons of carbon between now and 2050. Panel A of Figure 1 shows carbon-pricing trajectories that could achieve these limiting national budgets that were derived from six different integrated assessment models that included detail representations of the economics of the energy sector. Panel B shows the underlying energy mixes that each model projected over time. Diversity in what to expect in terms of energy development is clear in Panel B, but the reassuring conclusion was that any of them would be economically feasible if carbon were priced as shown in Panel A.

How or Why Should Carbon Be Priced

Pricing carbon is required not only to characterize mitigation policy, but also to internalize the value of carbon emission reductions in the evaluation of other regulations like CAFÉ standards and programs designed to promote energy efficiency – or other actions that might increase carbon emissions. Specific SCC values chosen from across the range of estimates have been used for this purpose, but they are always a bit suspect because the range is so large and they are based on incomplete knowledge.

Two other approaches have been offered. The first is supported by evidence of the type displayed in Panel A of Figure 1 – use the shadow price of carbon that is consistent with achieving mitigation targets that are derived from acceptable levels of risk. There is still no single estimate, but at least the range is smaller. The second looks at the specific technologies required to achieve those targets (from evidence of the type displayed in Panel B) and asks economists to calculate the price of carbon that would be required to make them economically feasible on a large scale. Hotelling again comes into play, because this approach suggests simply calculating a current price of carbon that would, if it were to climb at the rate of interest, reach perhaps the highest of these feasibility threshold prices at the requisite time in the future.

See also: Agricultural Sciences and Technology; Agriculture, Economics of; Ecological Economics; Economic Geography; Environmental Planning; Globalization: Geographical Aspects.

Bibliography

Ebi, K., Yohe, G., 2013. Adaptation in first- and second-best worlds. Current Opinion on Environmental Sustainability. http://dx.doi.org/10.1016/j.cosust.2013.06.004.

Fawcett, A.A., Calvin, K.V., De La Chesnaye, F.C., Reilly, J.M., Weyant, J.P., 2009. Overview of EMF-22 U.S. transition scenarios. Energy Economics 31 (Suppl. 2), S198–S211.

Hope, C., 2006. The marginal impacts of CO2, CH4 and SF6 emissions. Climate Policy 6, 537–544.

Hotelling, H.H., 1931. The economics of exhaustible resources. Journal of Political Economy 39, 137–175.

IPCC, 1996. Climate Change 1995: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Second Assessment Report of the Intergovernmental Panel on Climate Change. In: Watson, R.T., Zinyowera, M.C., Moss, R.H. (Eds.). Cambridge University Press, Cambridge, p. 880.

IPCC, 2001. Climate Change 2001: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change. In: McCarthy, J.J., Canziani, O.F., Leary, N.A., Dokken, D.J., White, K.S. (Eds.). Cambridge University Press, Cambridge, p. 1032.

IPCC, 2007a. Yohe, G.W., Lasco, R.D., Ahmad, Q.K., Arnell, N., Cohen, S.J., Hope, C., Janetos, A.C., Perez, R.T. Perspectives on climate change and sustainability. In: Parry, M.L., Canziani, O.F., Palutikof, J.P., Hanson, C.E., van der Linden, P.J. (Eds.), Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge.

IPCC, 2007b. In: Bernstein, L., Bosch, P., Canziani, O., Chen, Z., Christ, R., Davidson, O., Hare, W., Huq, S., Karoly, D., Kattsov, V., Kundzewicz, Z., Liu, J., Lohmann, U., Manning, M., Matsuno, T., Menne, B., Metz, B., Mirza, M., Nicholls, N., Nurse, L., Pachauri, R., Palutikof, J., Parry, M., Qin, D., Ravindranath, N., Reisinger, A., Ren, J., Riahi, K., Rosenzweig, C., Rusticucci, M., Schneider, S., Sokona, Y., Solomon, S., Stott, P., Stouffer, R., Sugiyama, T., Swart, R., Tirpak, D., Vogel, C., Yohe, G. (Eds.), Climate Change 2007: Synthesis Report (For the Fourth Assessment Report of the Intergovernmental Panel on Climate Change). Cambridge University Press, Cambridge.

Manne, A., Richels, R., 1997. On stabilizing CO2 concentrations—cost-effective emission reduction strategies. In: Cameron, O.K., Fukuwator, K., Morita, T. (Eds.), Proceedings of the IPCC Asia–Pacific Workshop on Integrated Assessment Models. Center for Global Environmental Research, Ibaraki, Japan, pp. 439–459.

NRC, 2010a. Climate Stabilization Targets: Emissions, Concentrations, and Impacts of Decades to Millennia. National Academies Press, Washington, DC. http://www.nap.edu/catalog.php?record_id=12877.

NRC, 2010b. Limiting the Magnitude of Climate Change. National Academies Press, Washington, DC. http://www.nap.edu/catalog.php?record_id=12785.

Nordhaus, W.D., 1991. To slow or not to slow: the economics of the greenhouse effect. Economic Journal 101, 920–937.

Stern, N., 2007. The Economics of Climate Change: The Stern Review. Cambridge University Press, Cambridge, p. 692.

Titus, J.G., 1992. The cost of climate change to the United States. In: Majumdar, S.K., Kalkstein, L.S., Yarnal, B., Miller, E.W., Rosenfeld, L.M. (Eds.), Global Climate Change: Implications, Challenges and Mitigation Measures. Pennsylvania Academy of Science, Easton, PA, pp. 217–235.

Tol, R.S.J., 2005. The marginal damage costs of carbon dioxide emissions: an assessment of the uncertainties. Energy Policy 33, 2064–2074.

Watkiss, P., Anthoff, D., Downing, T., Hepburn, C., Hope, C., Hunt, A., Tol, R., 2005. The Social Costs of Carbon (SCC) Review: Methodological Approaches for Using SCC Estimates in Policy Assessment. Final Report, Defra, UK, p. 124.

Watkiss, P., 2011. Aggregate economic measures of climate change damages: explaining the differences and implications. Interdisciplinary Review of Climate Change 2, 356–372.

Wigley, T., Richels, R., Edmonds, J.A., 1996. Economic and environmental choices in the stabilization of atmospheric CO2 concentrations. Nature 379, 240–243.

Climate, History of

Christian Pfister, University of Bern, Bern, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A survey on climate change over the last 10 000 years is provided in this article. It highlights natural (orbital elements, solar irradiance, volcanic explosions) and anthropogenic (greenhouse) forcing. The article introduces the reader to Historical Climatology, which reconstructs weather and climate from documentary evidence on daily to seasonal resolutions. Climate change over the last millennium in China, Europe, and the Mediterranean, South America, Japan and Africa is summarized. Dealing with the societal impacts of climate requires the consideration of social and cultural factors, which underlie people's ongoing vulnerability. This article emphasizes the need for case studies combining quantitative and narrative approaches rather than demonstrating the deterministic synchrony of climate and historical events.

Introduction

Past climate has become an important issue in the climatic change debate. Though hundreds of scholars have assembled robust evidence that the world's climate has significantly changed in the past due to the anthropogenic emission of greenhouse gases, their findings are still put in doubt by a handful of scientists from outside the field of climatology (Oreskes and Conway, 2010). The first section of this article presents the main types of evidence used to reconstruct past climate and summarizes scientifically accepted results of global climate change since the end of the last Ice Age. The second section deals with the approach of Historical Climatology. The third section reviews the main trends of regional climate variability over the past millennium. Links between past climatic impacts, human perception, and social action are addressed in the last section.

Global Climatic Changes and Forcing Factors

This paragraph mainly draws on the review written by Wanner et al. (2008). For the past one million years the Earth's climate has been characterized by an alternation of glacial and interglacial episodes, marked by the waxing and waning of continental ice sheets. Over the last 9000 years the distribution of total solar irradiance was progressively reduced in the Northern Hemisphere and enhanced in the Southern Hemisphere, due to changes in the orbital parameters (**Figure 1**). After the Climatic Optimum 6000 years before the present, these tendencies led to a southward shift of the climatic zones resulting (among other effects) in the desertification of the once green Sahara.

Globally unevenly distributed, decadal to century long, warm and cold fluctuations caused by natural forcing (solar irradiance activity and tropical volcanic eruptions) and variability due to feedbacks within the climate system, were superimposed on this long-term north hemispheric cooling trend. During the Roman Period (250 BC to AD 400) temperatures in Europe were somewhat higher than at present due to orbital and solar forcing, which is an important, but often overlooked argument in the climatic change debate. The subsequent Migration Period (AD 400 to 800) was cooler. A sharp cold relapse in the AD 530s due to volcanic forcing, led to several 'years without a summer' in Europe and to drought in other parts of the world (Gunn, 2000). Climate in Europe was mild during the early millennium due to solar and orbital forcing, albeit not so warm as during the late twentieth century. Much of humanity, particularly in the Americas, Northern China, as well as in the Saharan Sahel, the Nile Valley, and Eastern Africa suffered from long periods of severe aridity (Fagan, 2008). A subsequent period of cooling, the Little Ice Age (about AD 1300 to 1900) connected to orbital forcing, reduced solar irradiance and frequent tropical volcanic eruptions and led to repeated far-reaching glacier advances around the globe. The Little Ice Age is composed of manifold monthly and seasonal temperature and precipitation patterns, also including warm phases and droughts. The global temperature rise of the last 150 years is attributed to the anthropogenic greenhouse effect, which accelerated since the 1950s in step with the increased use of fossil fuels (Pfister, 2010). "Humans are now a global force that rivals the geophysical forces of nature in many aspects" (Costanza et al., 2005: 12). Since 1990, global mean temperatures have very likely been higher than at any previous time during the last 1000 years.

The Approach of Historical Climatology

Three kinds of evidence are used for reconstructing past climate. First, proxy evidence stored in archives of nature (e.g., tree-rings and ice-cores) and second, documentary proxy evidence preserved in the archives of societies (e.g., chronicles, diaries, account books). Third, instrumental measurements of climatic elements carried out by individuals from the late seventeenth century. Such observations were institutionalized within national and global networks from the nineteenth century (Brázdil et al., 2010). Climate reconstructions in the instrumental period are reviewed in reports of the Intergovernmental Panel on Climate Change (IPCC) (Trenberth and Jones, 2007).

Climatic change is a statistical artifact that cannot be readily related to human perception and societal action. Climate sciences are directed at an improved understanding of the climate system. Starting from models of the physical world they

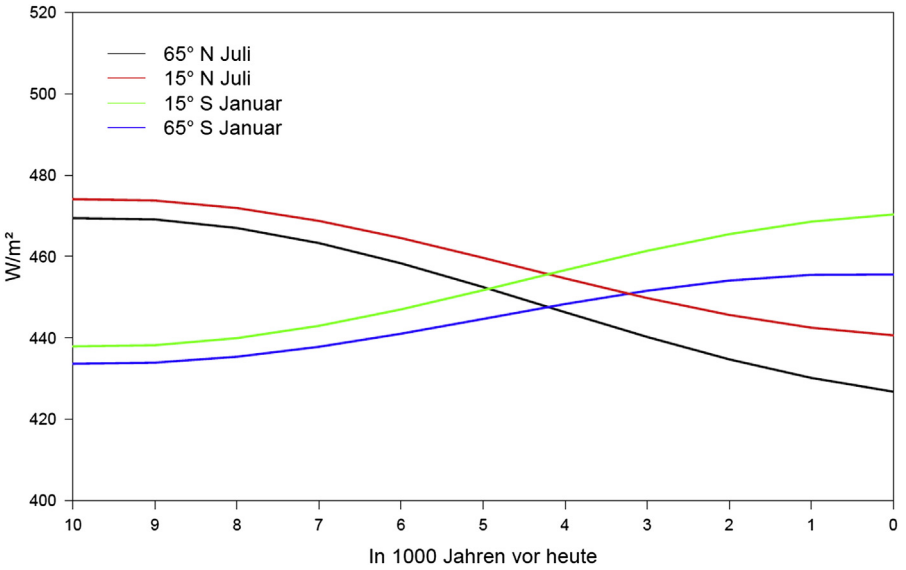

Figure 1 Insolation changes based on the orbital forcing during the last 10 000 years at 65° and 15° North and South during the corresponding summer season. Wanner, Heinz, Neukom, Raphael, 2010. Changes of the climate system in Europe. Nova Acta Leopoldina 112, 45–58.

extend the climate record as far back in time as possible on the basis of statistical procedures. Such macro-approaches are not suited for demonstrating people–climate interactions. Societies are sensitive and responsive to (extreme) weather anomalies taking place within daily to multi-seasonal time windows, which calls for micro-approaches. This awareness gained ground in the 1990s because an increase in weather extremes and disasters led to an additional focus on climate variability and extremes. This trend promoted a rise in historical disaster studies (Schenk and Engels, 2007) and an increasing consideration of short-term events within Historical Climatology.

The field of Historical Climatology is situated at the interface of climatology and (environmental) history, using methods and skills from both disciplines. Its objectives are threefold: First, past weather and climate of the period prior to the creation of national meteorological services are reconstructed. Second, the vulnerability of past societies to climate variations and extremes is investigated. Thirdly, debates on, and social representations, of weather and climate are explored (Pfister, 2007). Over the last 40 years Historical Climatologists (most of them being geographers, historians, and physicists) directed most of their time and effort to the collection of documentary data and the reconstruction of past climatic conditions. Given the short documentary record available to them, cooperation with archaeologists is very important for Climate Historians in North America (Carey, 2012). Until very recently climatic change was rarely addressed in syntheses of (European) environmental history, early exceptions being John F. Richards (2001) and John McNeill (2003).

The discussion on the social significance of climatic variations was long discredited by environmental determinism carried to an extreme by the geographer Ellsworth Huntington in the early twentieth century and by his followers to the present day (Fleming, 1998). The debate was relaunched from the late 1950s by the historian Emmanuel Le Roy Ladurie, a leading scholar of the French Annales School. His pioneering work was widely read and received, particularly in the English translation (Le Roy Ladurie, 1972). The English meteorologist Hubert Lamb, who took an active interest in history, became Le Roy Ladurie's most prominent opponent. He was convinced that weather and climate had affected human affairs in the past and that humankind would do well to examine some of the lessons provided by nature (Lamb, 1995).

Documentary sources contain two different kinds of evidence, namely: (1) narrative accounts on (unusual) weather spells and weather-induced disasters; and (2) biophysical evidence such as advances or delays of vegetation in the boreal summer half year and the presence or absence of frost, ice, and snow in the boreal winter half year. Another important distinction is related to the agents who kept the records, distinguishing between individuals and institutions. Individuals usually put an emphasis on describing extreme events such as floods, droughts, frosts, and hailstorms, their socio-economic impact, and their perception. They referred to biophysical evidence as a means to provide evidence that could be compared over time. Sources created by individuals are singular and relatively short, ending, at the latest, with the death of the observer. Institutions are bodies maintaining administrative functions within existing territorial structures. Their officeholders kept year by year records on weather dependent resources and activities (Pfister et al., 2009), such as the time of grape and grain harvests or the opening or closing of harbors in the Baltic (e.g., Leijonhufvud et al., 2010). Institutional functions involved more or less standardized procedures and these often worked in the same way for centuries, often generating multi-secular, long, continuous, and quasi-homogeneous series of proxy data. Ship's logbooks covering large parts of the world's oceans, including the high latitudes, are probably the most important institutional source. All officers on board of a (war) ship had to keep a logbook, which involved laying down regular observations on wind direction, wind force, weather, and – eventually – sea and ice conditions.

Logbooks usually adopt a standardized vocabulary and system of recording. Hundreds of thousands of logbooks are contained in the archives of naval powers, of which just a sample has so far been investigated in the framework of the EU Project CLIWOC (Wheeler et al., 2006).

Historical Climatologists developed a specific approach of climate reconstruction tailored to investigations on climate impacts as well as extreme value and disaster analysis. It allows integrating data on daily, monthly, and seasonal time scales into systematic monthly and seasonal reconstruction of temperatures and precipitation. Data fields containing direct narrative and calibrated indirect data are assigned synthetic ordinal (ranked) variables called temperature and precipitation indices taking a finite number of 3 to 7 values. This approach, elaborated by the Swiss historian and geographer Christian Pfister (1984) and based on preliminary work by Lamb (1977), became standard in Historical Climatology. Dobrovolný et al. (2010) used the palaeoclimatological calibration and verification approach to assess monthly and seasonal temperatures for Central Europe (Germany, Czechia, and Switzerland) from the so called Pfister Indices for the period from AD 1500 to present. The Middle Ages are included in millennium records of temperature variations for the Netherlands (Shabalova and van Engelen, 2003) as well as for Germany and Central Europe (Glaser and Riemann, 2009).

Main Trends of Climatic Variability in the Last Millennium

Today, there are six major foci of Historical-Climatologic research, namely China, Europe and the Mediterranean, South America, Japan, Africa, and the World's Oceans. For North-America, no review is available yet. In the Arab world an annalistic tradition is known to exist, but research on past weather and climate has hardly begun (Weintritt, 2009). In India, South-East Asia, and Oceania the potential for Historical Climatology still needs to be explored.

China

Ge et al. (2008) have composed a review article on climate history of China introducing the climatic information recorded in Chinese historical documents in easily understandable form and including a glossary. Systematic documentary records were kept beginning in the Qin dynasty (221 BC). The emperors strived to secure reports on current weather events and disasters from the provinces, with a view to taking timely measures to forestall the outbreak of famine-driven riots. Records mainly cover the eastern part of the country. Under the Ming Dynasty (AD 1368 to 1644) the central government called local scholars to compile the body of knowledge on past and current geography and to report on natural disasters, population, and agricultural resources. Under the Qing Dynasty (AD 1644 to 1911) administrative centers had to submit memos on their daily administrative activities to the Emperor. From AD 1736, officials were required to compile data on the duration of rainfall, including its implications for crop growth. The reports had to be prepared by the responsible officials in person and their reliability was checked. Local chroniclers recorded extraordinary phenomena in the natural world, assuming that disasters reflected on people's behavior and the government's performance. A discourse describing China as a country especially prone to natural disaster became dominant during the second half of the nineteenth century (Janku, 2009).

In South China the thirteenth century was the warmest of the last millennium. Three cold periods – AD 1470 to 1520; AD 1620 to 1740; and AD 1840 to 1890 – have been identified in the Little Ice Age. In North China two cold periods – AD 1500 to 1690 and AD 1800 to 1860 – stand out over the last six centuries. Climate variability increased markedly throughout the nineteenth century to a maximum in the early twentieth century (Wang et al., 1991).

Europe and the Mediterranean

Research in Europe was reviewed by Brázdil et al. (2010). Though some documentary evidence exists in most European countries, only a few of them have been engaged in systematic, historical-climatologic investigations: these being the Czech Republic, Estonia, Germany, Hungary, Italy, the Netherlands, Norway, Poland, Portugal, Spain, Switzerland, and the UK. Progress in climate reconstruction in these countries was spurred through EU projects. Syntheses in the form of monthly and seasonal charts, spatial charts of seasonal sea-level pressure, temperature, and precipitation dating back to AD 1500, were produced by the team led by Jürg Luterbacher (Luterbacher et al., 2002, 2004; Pauling et al., 2006).

In a diachronic perspective, the known evidence is summarized according to four cumulative time frames:

1. AD 1000 to 1200: individual reports of socio-economically significant anomalies and (weather induced) disasters.
2. AD 1200 to 1500: more or less continuous clusters of individual reports on summer and winter, partly on spring, including references to 'average' conditions. A wealth of administrative/institutional records exists in England for the period *c.* AD 1270 to 1400, which contains both direct weather references as well as institutional proxy information (Pribyl et al., 2012).
3. AD 1500 to 1800: almost complete monthly weather descriptions exist for Switzerland (Pfister, 1999), Germany (Glaser, 2008), and Czechia (Brázdil et al., 2010). The keeping of weather diaries was promoted by the rise of planetary astronomy and its application in astro-meteorology. Thirty-two sixteenth century weather diaries are so far known for Central Europe. The data obtained from these sources provided the backbone for setting up monthly precipitation statistics (Pfister et al., 1999). Institutional sources giving long series of dates of grape harvests (Daux et al., 2011) and grain harvests (Mozny et al., 2010; Wetter and Pfister, 2011) are abundant. Records of rogations (i.e., standardized religious ceremonies to put an end to meteorological stress situations) were widely used for climate reconstructions in Spain (Barriendos, 2009).
4. AD 1650 to 1860: the establishment of short-lived networks of early instrumental observations, the first being 'Rete Medicea' (AD 1653 to 1677) based in Florence and comprising 10 stations (Camuffo, 2002). A later network, the 'Societas Meteorologica Palatina,' founded 1780 by Karl

Theodor, Prince Elector of the Palatinate, included 39 individual scientists resident across Europe, in Siberia, Greenland, and North America. The Second Coalition War put an end to this organization in 1795 (Lingelbach, 1980). Regional instrumental networks were set up in the early nineteenth century, which were integrated into national and global networks after 1850.

Subsequent climatic trends are summarized: after a cold phase in the twelfth century, winters were prevailingly warm from AD 1180 to 1300. From AD 1300 to 1900, the winter half-year was colder than today, due to more frequent advection of cold, dry air-masses from the Northeast. Springs were extremely cold in the 1690s and the 1740s. Summers in the thirteenth century were prevailingly warm and dry. In the fourteenth century, clusters of cold and wet summers occurred repeatedly (e.g., in the 1310s and 1340s). April to July temperatures in England dropped by 0.6 °C from AD 1256 to 1431, affecting the economy by a higher frequency of extremely cold 'growing seasons' (Pribyl et al., 2012). From AD 1380 to 1430 and again from AD 1530 to 1565 the summer half-year was as warm as it is today.

Cold spells and long rains in midsummer became more frequent in the late sixteenth century (Pfister, 2007). The period AD 1640 to 1650 witnessed harsh weather and severe famines in many countries, also outside Europe (Parker, 2008). Warm decades stand out in the 1720s, the 1730s, and in the 1780s in Western and Central Europe, whereas the first half of the nineteenth century, particularly the 1810s, were markedly cooler (Bradley and Jones, 1996; Pfister, 1999; Glaser, 2008).

Winters in Russia became more severe from the end of the sixteenth century, in particular from AD 1620 to 1680 and during the first half of the nineteenth century. The first half of the sixteenth century was warm in all seasons. Subsequently, cold spells occurred more often from AD 1590 to 1620 and from AD 1690 to 1740, with a peak in the 1730s. Droughts were frequent from AD 1201 to 1230, AD 1351 to 1380, and AD 1411 to 1440; and again from AD 1640 to 1659, from AD 1680 to 1699, and from AD 1801 to 1860. Summers from AD 1890 to 1920 were the coldest in the last 500 years and extremely dry and wet seasons were frequent (Bradley and Jones, 1996).

Climatic trends in the Mediterranean since AD 1500 were reviewed by Luterbacher et al. (2006). Winter temperatures were, with respect to AD 1961 to 1990 averages, almost consistently colder until the 1950s. In Catalonia (Northeastern Spain) dry spells in the winter half-year were frequent in the mid sixteenth century, but almost absent from AD 1580 to 1620 (Barriendos, 2009). The Mediterranean witnessed increased precipitation in the form of torrential rain and, at times, increased snowfall at higher altitudes (Tabak, 2008).

South America

The Argentinean historian and geographer Maria Rosario Prieto reviewed the studies in this continent up to AD 1810. Most work was done in Ecuador, Bolivia, Peru, Chile, and Argentina, although evidence also exists in other countries. The Spanish colonization of America from the sixteenth century relied on a rampant bureaucracy. The resulting documentation includes weekly reports on notable incidents including weather, harvest expenditures, and food supplies. Sources from priests, military officers, or civilians, such as letter collections, travel journals, annals, or diaries provide more continuous information. From AD 1784 the Spanish Crown asked each municipality to submit a half-yearly report about weather (impacts), agriculture, and trade. Logbooks of ships sailing the surrounding oceans provided daily information on winds and weather conditions. The transition to the Republican period after AD 1810 led to gaps in the record. Newspapers became a useful source of information in the postcolonial period. The first meteorological stations were established in the late nineteenth century.

In most regions, cold conditions prevailed from the late sixteenth to the early seventeenth century (e.g., in Central Europe). The Mendoza river streamflow in Argentina was smaller from AD 1600 to 1670, due to lower summer temperatures and reduced melting of snow in the Andes. A long, cold interval from AD 1520 to 1660 stands out in the Northern Patagonian Andes (Rosario Prieto and García-Herrera, 2008).

Japan

Historical Climatology in Japan was reviewed by the geographer Takehiko Mikami. Climatic trends in the 'pre-instrumental' period are based upon three kinds of proxy-data. (1) The annual dates when Lake Suwa in central Japan froze have been recorded since the fifteenth century, and these dates are highly correlated to December and January temperatures. When Lake Suwa freezes, the shrinkage and expansion of the ice due to diurnal temperature variations produces a specific feature (Omiwatari) resembling a bridge over the lake. The annual appearance of Omiwatari gave rise to a celebration the date of which was recorded and retained. (2) Cherry blossom is known in Japanese culture as a symbol for the ephemeral nature of life. Picnicking under a blossoming cherry tree (hanami) became first a custom of the Imperial elite in Kyoto and spread then gradually to other strata of society (http://en.wikipedia.org/wiki/Cherry_blossom). Year-to-year variations in full flowering dates of cherry trees in Kyoto were recorded since AD 1001 and were used as a proxy for estimating March temperatures. (3) Families in many parts of the country kept weather diaries from the early eighteenth century. Based on this evidence, July temperatures in Tokyo from 1721 to the present, have been estimated.

Reconstructed mean December–January temperatures were 1 to 1.5 °C lower in the early 1600s than in the years AD 1961 to 1990. Winters then warmed up between AD 1750 and 1850. March temperatures were higher during the eleventh to the thirteenth centuries and relatively lower during the Little Ice Age. Estimated July temperatures from AD 1721 to 1790 were 1 to 1.5 °C lower than in the years AD 1961 to 1990. Midsummers were rather warm in the nineteenth century. A cold relapse down to the low level of AD 1740 occurred around 1900 (Mikami, 2008).

Africa

The evidence for the last two centuries was reviewed by the climatologist Sharon E. Nicholson. Few instrumental records

exist prior to the twentieth century, and most of these dealt with rainfall. However, documentary proxy information permits a quasi-continuous record on precipitation patterns since the early eighteenth century. Nineteenth century rainfall variability in Lesotho (Southern Africa) was investigated by Nash and Grab (2010).

During the Medieval Warm Period and again from the sixteenth through the eighteenth centuries, Northern and Southern Africa benefited from rather humid conditions. From the late eighteenth century arid conditions prevailed again, particularly during the 1820s and 1830s. Since then, several periods of abundant rainfall (AD 1860 to 1894; in the 1920s and 1930s; in the 1950s) alternated with periods of severe drought (in the 1910s and the 1960s) to the present. Nicholson argued that the recent droughts in Africa are more driven by changing rainfall than by land-use change and desertification (Nicholson, 2001).

Weather, Climate and History

Human history is usually cast in terms of the rise and fall of great civilizations, wars, and specific human achievements, leaving out the ecological and climate context, which shaped and mediated these events (Costanza et al., 2005). However, "the inclusion of the climate factor in the study of history should not be regarded as a search for an alternative, and deterministic, explanation of the past, but as an expansion of the context in which the workings of past societies are to be understood." (De Vries, 1985: 274). Indeed, the effects of climatic fluctuations or 'climate' on the course of 'history' are difficult to demonstrate, because both climate and history are blanket terms, situated on such a high level of abstraction that relationships between them cannot be investigated in a scientifically meaningful way. On a very general level, it could be said that beneficial climatic effects tend to enlarge the scope of human action, whereas climatic shocks tend to restrict it. The sequences of climatic situations that mattered, depended upon the vulnerability of the impacted groups and societies to climatic hazards (Pfister, 2007). Social (or internal) vulnerability encompasses the entirety of structural factors that determine the outcome of an event of a given nature, as well as the severity of factors such as poverty, dependence on outside support, and the ability to cope with specific hazards (Füssel, 2007). Striking differences of vulnerability may arise within very small areas, as the example of Switzerland in the European subsistence crisis of 1816/17 demonstrates. Whereas the easternmost proto-industrialized parts of the country, cut off by embargoes from their main breadbasket in neighboring Germany, suffered from famine, the western region being almost self-sufficient, only underwent a mild form of food deprivation (Krämer, 2012).

In many cases, however, investigations of climate and history limit themselves to hinting at the synchronicity of climatic and economic events. A recent example is a tree ring-based study linking the temperature depression of the early seventeenth century in Europe with the Thirty Years War (Büntgen et al., 2011). With regard to China, Zhang et al. (2007) correlated analyses of natural proxy data from national archives with hypothesized data on harvests and demography, to demonstrate interrelations between climatic change, the frequency of conflicts, and dynastic rise and fall. A study on the relationship between frequency of wars and climate in Europe was attempted by Tol and Wagner (2010). Historian Ka-wai Fan (2010) argues that scientists too often fail to acknowledge the complexity of history and the overwhelming role played by human beings. Indeed, the authors of all three above mentioned papers overlooked the truism recalled by Martin Parry (1981) that "an investigation of the synchronicity of climatic and economic events infers a space–time coincidence rather than causality." (p. 321). Nico Stehr and Hans von Storch (2000) note in this context that "a large proportion of today's climate impact research is genuine climate determinism" (p. 187).

A human history of climate should highlight those extreme climate anomalies that are known to have affected everyday life and disrupted daily routines. In particular this relates to those conditions on which humans depend for their basic needs and well-being, i.e., the availability of water, food, and energy. Every vulnerability analysis needs an appropriate focus, considering of course, the interrelationships of the many dimensions of vulnerability (see **Figure 2**).

Food availability is a major focus of (historical) climate impact studies. A history linking climate and society should draw on extreme events and highlight changes in the frequency of weather spells that were recognized by contemporaries to be responsible for widespread weather-related crop failures. In any case, the contexts of historical events need to be highlighted to provide meaningful explanations.

There is agreement that societies do not just react to climatic stimuli. Brian Fagan (2008) was amazed "how remarkably flexible human societies were in much of the world one thousand years ago" (p. 236). Climatic impacts become critical, "when a society reached a critical mass, a density of urban population and nonfarmers that was unsustainable in drought cycles or an agrarian economy that had exhausted the land and all opportunities for diversification" (p. 237). As Jared Diamond (2005) put it: "One of the main lessons to be learned from the collapse of past societies […] is that a steep decline may begin only a decade or two after a society reached its peak numbers, wealth and power" (p. 509). The reaction of the ruling elites to climate and environmental change may be decisive for the survival of states and empires in such situations. Besides the well-known paradigm of Easter Island (e.g., Diamond, 2005), the example of the Western Roman Empire, (a large, multicultural state entity of the size of the European Union) is worth recalling in this context: The cooling trend from the late second century AD affected the size of harvests and thus the revenues of the Empire. The imperial bureaucracies and the army were so much expanded, that the taxable peasants became unable to bear the tax burden in the context of falling harvests; and they fled the land leaving large arable tracts uncultivated. In the end, the Emperor could not pay his armies any more, whereupon these seized power and deposed him (Tainter and Crumley, 2005).

In a seminal article Robert W. Kates (1985) distinguished several levels of climatic impacts (see **Figure 3**). Their sequence and the size of the intermediate arrows indicate, how closely these effects are related to climatic impact.

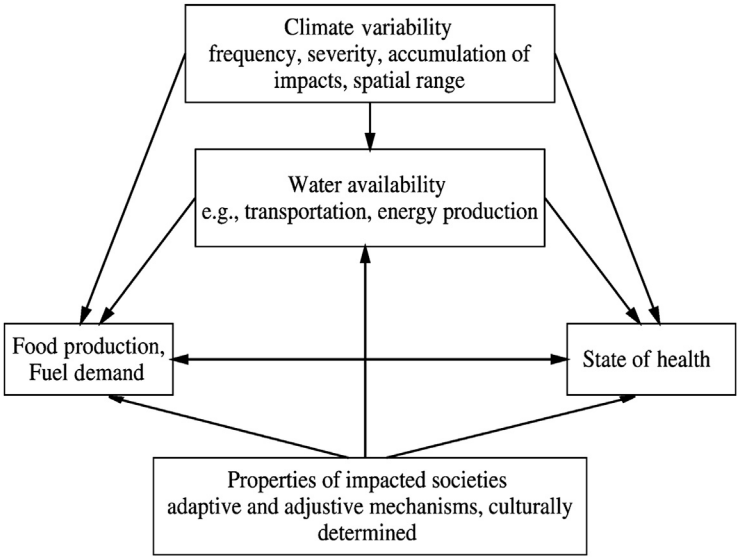

Figure 2 Impacts of extreme climate events on societies. Pfister, Christian, 2007. Climatic extremes, recurrent crises and witch hunts: strategies of European societies in coping with exogenous shocks in the late sixteenth and early seventeenth centuries. The Medieval History Journal 10, 1–41.

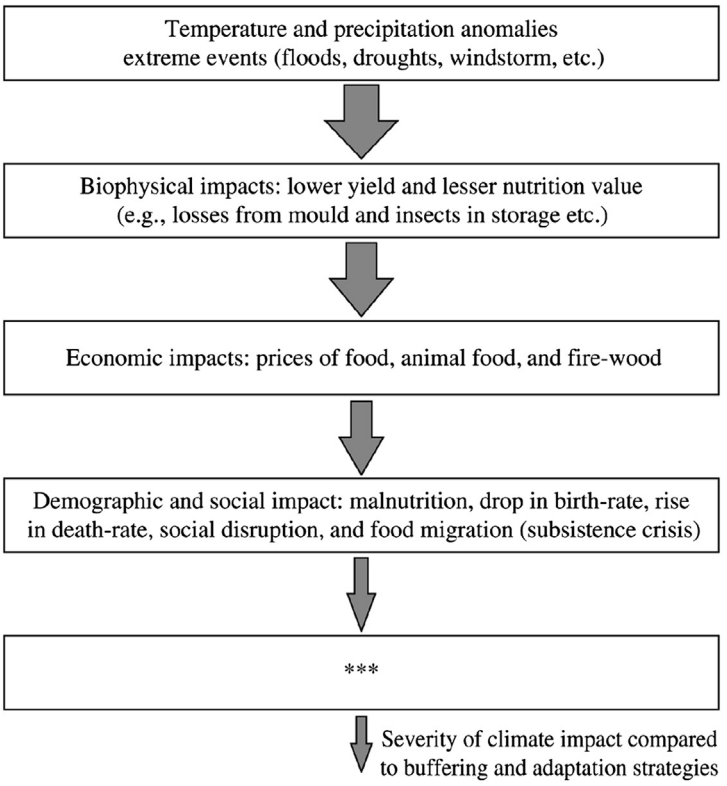

Figure 3 A basic model of climatic impacts on society. Kates, Robert W., 1985. The interaction of climate and society. In: Kates, Robert W., Ausubel, Jesse H., Berberian, Mimi (Eds.), Climate Impact Assessment: Studies of the Interaction of Climate and Society, vol. 27. John Wiley & Sons, Chichester, 3–36; Pfister, Christian, 2007. Climatic extremes, recurrent crises and witch hunts: strategies of European societies in coping with exogenous shocks in the late sixteenth and early seventeenth centuries. The Medieval History Journal 10, 1–41.

The quantity and storability of crops situated on the first level are most immediately related to the weather. However, most members of premodern societies had developed sophisticated strategies (e.g., horizontal and vertical crop diversification) to cope with the risk of harvest failure. Fluctuations in food prices are among the most obvious second-order consequences. Climatic effects on this level are, however, masked by intervening factors, such as trade, the stocking of

public grain, hoarding, and speculation. Third-order effects, such as demographic effects, are intertwined with epidemiological and public health contexts in ways that are highly controversial. Malnutrition is known to be contingent on distributional systems, in turn dependent on power relations, factors of class and gender, and the supportive capacity of social networks. Whether famines resulted in social upheaval largely depended on the quality of governance. The farther we move away from first-order effects, the broader are the options open to individual or collective actors. Likewise, the web of factors masking the climatic effect gets to be more complex. Of course, the one-directional, top-down arrays are heuristically motivated oversimplifications. As is well known, the role of feedback is crucial in complex socio-ecological systems, and a it is major reason why simple cause and effect paradigms often have little explanatory power (Costanza et al., 2005: 14).

A major difference in dealing with climate impact studies relates to the fact that natural sciences mainly depend on quantitative or quantified data, while human sciences mainly draw on narrative or qualitative evidence. Case studies in terms of absorbing narratives are closest to people's perceptions and experiences, but they are hardly suited for generalizations. In some cases it may be possible to model relationships between weather, harvests, and prices (e.g., Pfister and Brázdil, 2006). In general, cold periods in March and April and rainy midsummers affected the quantity and quality of all sources of agricultural produce in temperate Europe (Pfister, 2007). Cultural analyses of climate change illuminate the response of 'real people' to climate change. At present, it is scientifically accepted "that climate is both the physical characteristics of the climate system and a cultural construction that emerges from perceptions, meanings, spirituality, discourse, and distinct knowledge bases that vary in time and space" (Carey, 2012; Daniels and Endfield, 2009).

See also: Climate Change and the Media; Climate Change: Economics; Consumption, History of; Environmental Archaeology; Environmental Conservation and Development: Critical Perspectives; Environmental Sociology; Hazards and Disaster Research in Contemporary Anthropology; Limits to Growth; Societal Impact Assessment; Sustainable Development: An Economic Perspective.

Bibliography

Barriendos, Mariano, 2009. Historical climatology in Spain. In: Kirchhofer, André, Krämer, Daniel, Maria, Merki Christoph, et al. (Eds.), Nachhaltige Geschichte. Festschrift für Christian Pfister. Chronos, Zürich, pp. 49–64.

Bradley, Raymond S., Jones, Philip D., 1996. Climate Since A.D. 1500. Routledge, London.

Brázdil, Rudolf, Dobrovolný, Petr, Luterbacher, Jürg, Moberg, Anders, et al., 2010. European climate of the past 500 years: new challenges for historical climatology. Climatic Change 101, 7–40.

Büntgen, Ulf, Tegel, Willy, McCormick, Michael, et al., 2011. 2500 Years of European climate variability and human susceptibility. Science Express 331, 1–4.

Camuffo, Dario, 2002. History of the long series of daily air temperature in Padova (1725–1998). Climatic Change 53, 7–75.

Carey, Mark, 2012. Climate and history: a critical review of historical climatology and climate change historiography. WIREs Climate Change 3, 233–249.

Clark, Peter (Ed.), 1985. The European Crisis of the 1590's. Essays in Comparative History. Allen & Unwin, London.

Costanza, Robert, Graumlich, Lisa J., Steffen, Will, 2005. Sustainability or collapse? Lessons from integrating the history of humans and the Rest of nature. In: Tainter, Joseph A., Crumley, Carole I. (Eds.), Climate, Complexity and Problem Solving in the Roman Empire. In: Costanza, Robert, Graumlich, Lisa J., Steffen, Will (Eds.), Sustainability or Collapse? An Integrated History and Future of People on Earth. MIT Press, Cambridge, pp. 5–18.

Daniels, Stephen, Endfield, Georgina H., 2009. Narratives of climate change: introduction. Journal of Historical Geography 35, 215–222.

Daux, Valerie, García, Inaki, Yiou, Pascal, et al., 2011. An open-database of Grape Harvest Dates for climate research: data description and quality assessment. Climate of the Past Discussions 7, 3823–3858.

De Vries, Jan, 1985. Analysis of historical climate-society interaction. In: Kates, R.W., Ausubel, J., Berberian, M. (Eds.), Climate Impact Assessment, SCOPE 27, Studies of the Interaction of Climate and Society. John Wiley & Sons, Chichester, pp. 273–292.

Diamond, Jared M., 2005. Collapse. How Societies Choose to Fail or Survive. Penguin Books, London.

Dobrovolný, Petr, Moberg, Anders, Brázdil, Rudolf, et al., 2010. Monthly, seasonal and annual temperature reconstructions for Central Europe derived from documentary evidence and instrumental records since AD 1500. Climatic Change 101, 69–107.

Fagan, Brian, 2008. The Great Warming. Climate Change and the Rise and Fall of Civilizations. Bloomsbury Press, New York.

Fan, Kai-wai, 2010. Climatic change and dynastic cycles in Chinese history: a review essay. Climatic Change 101, 565–573.

Fleming, James Rodger, 1998. Historical Perspectives on Climate Change. Oxford University Press, Oxford.

Füssel, Hans-Martin, 2007. Vulnerability: a generally applicable conceptual framework for climate change research. Global Environmental Change 17, 155–167.

Ge, Quansheng, Zheng, Jingyun, Tian, Yanyu, et al., 2008. Coherence of climatic reconstruction from historical documents in China by different studies. International Journal of Climatology 28, 1007–1024.

Glaser, Rüdiger, 2008. Klimageschichte Mitteleuropas. 1200 Jahre Wetter, Klima, Katastrophen. Wissenschaftliche Buchgesellschaft (WBG); Primus Verlag, Darmstadt.

Glaser, Rüdiger, Riemann, Dirk, 2009. A thousand-year record of temperature variations for Germany and Central Europe based on documentary data. Journal of Quaternary Science 24, 437–449.

Gunn, Joel, 2000. The Years without Summer: Tracing A.D. 536 and Its Aftermath. Archaeopress, Oxford.

Janku, Andrea, 2009. What Chinese biographies of moral exemplars tell us about disaster experiences (1600–1900). In: Kirchhofer, André, Krämer, Daniel, Maria, Merki Christoph, et al. (Eds.), Nachhaltige Geschichte. Festschrift für Christian Pfister. Chronos, Zürich, pp. 129–148.

Kates, Robert W., 1985. The interaction of climate and society. In: Kates, Robert W., Ausubel, Jesse H., Berberian, Mimi (Eds.), Climate Impact Assessment: Studies of the Interaction of Climate and Society, vol. 27. John Wiley & Sons, Chichester, pp. 3–36.

Krämer, Daniel, 2012. Der kartierte Hunger. Räumliche Kontraste der Verletzlichkeit in der Schweiz während der Hungerkrise 1816/17. Schweizerisches Jahrbuch für Wirtschafts- und Sozialgeschichte 27, 113–123.

Lamb, Hubert Horace, 1977. Climate History and the Future. Methuen, London.

Lamb, Hubert Horace, 1995. Climate, History and the Modern World. Methuen, London. Routledge, New York.

Le Roy Ladurie, Emmanuel, 1972. Times of Feast, Times of Famine: A History of Climate Since the Year 1000. Allen & Unwin, London.

Leijonhufvud, Lotta, Wilson, Rob, Moberg, Anders, et al., 2010. Five centuries of Stockholm winter/spring temperatures reconstructed from documentary evidence and instrumental observations. Climatic Change 101, 109–141.

Lingelbach, Ernst, 1980. Vom Messnetz der Societas Meteorologica Palatina zu den weltweiten Messnetzen heute. Annalen der Meteorologie, Neue Folge 16, 1–27.

Luterbacher, Jürg, Xoplaki, Elena, Dietrich, Daniel, et al., 2002. Reconstruction of sea level pressure fields over the Eastern North Atlantic and Europe back to 1500. Climate Dynamics 18, 545–561.

Luterbacher, Jürg, Dietrich, Daniel, Xoplaki, Elena, et al., 2004. European seasonal and annual temperature variability, trends, and extremes since 1500. Science 303, 1499–1503.

Luterbacher, Jürg, Xoplaki, Elena, Casty, Carlo, et al., 2006. Mediterranean climate variability over the last centuries: a review. In: Lionello, Piero, Malanotte-Rizzoli, Paola, Boscolo, Roberta (Eds.), The Mediterranean Climate: An Overview of the Main Characteristics and Issues. Elsevier, Amsterdam, pp. 27–148.

Mauelshagen, Franz, 2010. Klimageschichte der Neuzeit 1500–1900. Wissenschaftliche Buchgesellschaft, Darmstadt.

McNeill, John R., 2003. The nature of environmental history. History and Theory. Studies in the Philosophy of History 42, 5–43.

Mikami, Takehiko, 2008. Climatic variations in Japan reconstructed from historical documents. Weather 63, 190–193.

Mozny, Martin, Brázdil, Rudolf, Dobrovolný, Petr, et al., 2010. March-June temperature reconstruction in the Czech Republic based on cereal harvest dates in the 1501–2008 period. Climatic Change 110 (3-4), 801–821.

Nash, David J., Grab, Stefan W., 2010. 'A sky of brass and burning winds': documentary evidence of rainfall variability in the Kingdom of Lesotho, Southern Africa, 1824–1900. Climatic Change 101, 617–653.

Nicholson, Sharon E., 2001. Climatic and environmental change in Africa during the last two centuries. Climate Research 17, 123–144.

Oreskes, Naomi, Conway, Erik M., 2010. Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming. Bloomsbury Press, New York.

Parker, Geoffrey, 2008. Crisis and catastrophe: the global crisis of the seventeenth century reconsidered. American Historical Review October, 1053–1079.

Parry, Martin L., 1981. Climatic change and the agricultural frontier. A research strategy. In: Wigley, Tom M.L., Ingram, Martin J., Farmer, Greg (Eds.), Climate and History. Studies in Past Climates and Their Impact on Man. Cambridge University Press, Cambridge, pp. 319–336.

Pauling, Andreas, Luterbacher, Jürg, Casty, Carlo, et al., 2006. Five hundred years of gridded high-resolution precipitation reconstructions over Europe and the connection to large scale circulation. Climate Dynamics 26, 387–405.

Pfister, Christian, 1984. Das Klima der Schweiz von 1525 bis 1860 und seine Bedeutung in der Geschichte von Bevölkerung und Landwirtschaft. Haupt, Bern.

Pfister, Christian, 1999. Wetternachhersage. 500 Jahre Klimavariationen und Naturkatastrophen 1496–1995. Haupt, Bern.

Pfister, Christian, 2007. Climatic extremes, recurrent crises and witch hunts: strategies of European societies in coping with exogenous shocks in the late sixteenth and early seventeenth centuries. The Medieval History Journal 10, 1–41.

Pfister, Christian, 2010. The "1950s Syndrome" and the Transition from a Slow-going to a Rapid Loss of Global Sustainability. In: Uekötter, Frank (Ed.), Turning Points in Environmental History. University of Pittsburgh Press, pp. 90–117.

Pfister, Christian, Brázdil, Rudolf, 2006. Social vulnerability to climate in the "Little Ice Age": an example from Central Europe in the early 1770s. Climate of the Past 2, 115–129.

Pfister, Christian, Bokwa, Anita, Brázdil, Rudolf, et al., 1999. Daily weather observations in sixteenth-century Europe. Climatic Change 43, 111–150.

Pfister, Christian, Luterbacher, Jürg, Wanner, Heinz, Wheeler, Dennis, et al., 2009. Documentary evidence as climate Proxies. Proxy-specific white paper produced from the PAGES/CLIVAR Workshop, Trieste, 2008. In: PAGES (Past Global Changes). http://www.pages.unibe.ch/products/latest.

Pribyl, Kathleen, Cornes, Richard, Pfister, Christian, 2012. Reconstructing medieval April–July mean temperatures in east Anglia, 1256–1431. Climatic Change 113, 393–412.

Prieto, Maria Rosario, García-Herrera, Ricardo, 2008. Documentary sources from South America: potential for climate reconstruction. Palaeography, Palaeoclimatology, Palaeoecology 281, 196–209.

Schenk, Gerrit J., Engels, Jens Ivo (Eds.), 2007. Historical disaster research. Concepts, methods and case-studies. Historical Social Research 32, Special Issue.

Shabalova, Marina V., van Engelen, Aryan F.V., 2003. Evaluation of reconstruction of winter and summer temperatures in the Low Countries, AD 764–1998. Climatic Change 58, 219–242.

Stehr, Nico, von Storch, Hans, 2000. Von der Macht des Klimas. Ist der Klimadeterminismus nur noch Ideengeschichte oder relevanter Faktor gegenwärtiger Klimapolitik. Gaia 9, 187–195.

Tabak, Faruk, 2008. The Waning of the Mediterranean, 1550–1870. Johns Hopkins University Press, Baltimore.

Tainter, Joseph A., Crumley, Carole I., 2005. Climate, complexity and Problem Solving in the Roman Empire. In: Costanza, Robert, Graumlich, Lisa J., Steffen, Will (Eds.), Sustainability or Collapse? an Integrated History and Future of People on Earth. MIT Press, Cambridge, pp. 61–76.

Tol, Richard S.J., Wagner, Sebastian, 2010. Climate change and violent conflict in Europe over the last millennium. Climatic Change 99, 65–79.

Trenberth, Kevin E., Jones, Philip D., 2007. Observations: surface and atmospheric climate change. In: Climate Change 2007: Working Group I: The Physical Science Basis. Cambridge University Press, Cambridge, New York.

Wang, Risheng, Wang, Shaowu, Fraedrich, Klaus, 1991. An approach to reconstruction of temperature on a seasonal basis using historical documents from China. International Journal of Climatology 11, 381–392.

Wanner, Heinz, Neukom, Raphael, 2010. Changes of the climate system in Europe. Nova Acta Leopoldina 112, 45–58.

Wanner, Heinz, Beer, Jürg, Bütikofer, Jonathan, et al., 2008. Mid- to late Holocene climate change: an overview. Quaternary Science Reviews 27, 1791–1828.

Weintritt, Ottfried, 2009. The floods of Baghdad: cultural and technological responses. In: Mauch, C., Pfister, C. (Eds.), Natural Disasters, Cultural Responses. Case Studies toward a Global Environmental History. Lexington Books, Plymouth, pp. 165–182.

Wetter, Oliver, Pfister, Christian, 2011. Spring-summer temperatures reconstructed for northern Switzerland and south-western Germany from winter rye harvest dates, 1454–1970. Climate of the Past 7, 1307–1326.

Wheeler, Dennis, García-Herrera, Ricardo, Koek, Fritts, et al., 2006. CLIWOC. Climatological Database for the World's Oceans 1750 to 1850. Results of a research project. European Commission, Luxemburg.

Zhang, David D., Brecke, Peter, Lee, Harry F., et al., 2007. Climate change and war frequency in eastern China over the last millennium. Human Ecology 35, 403–414.

Clinical Psychology in Europe, History of

I Lunt, University of London, London, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 2025–2029, © 2001, Elsevier Ltd.

Abstract

Although psychology as a discipline is held to have originated in Europe with Wundt's laboratory in Leipzig in 1879, clinical psychology emerged in Europe substantially after the World War II in the 1940 and 1950s when Europe was rebuilding itself and its health and welfare systems. Europe is characterized by the diversity of its 48 or so countries which have different histories and social, intellectual, and cultural traditions which have affected the development of the profession and practice of clinical psychology. Nevertheless, since the 1950s with the increasing professionalization of clinical psychology, leading to increased length of education and training, the development of ethics codes, growing regulation, and legislation in relation to the profession, and in 1981 the formation of the federation EFPPA (European Federation of Professional Psychologists Associations), there has been a gradual move toward more common standards and an attempt to develop a European framework for different areas of training and practice. This trend is likely to develop well into the mid-twenty-first century with a greater unification of Europe.

Introduction

Europe is characterized by its social, linguistic, and cultural diversity (e.g., Drenth et al., 1990; Poortinga, 1996), which is rooted in very different histories and traditions of the 48 or so countries. Even the geographical or political extent of Europe is a matter for some debate, though it is assumed that the growth of the European Union (currently 15 states, all of them in western Europe) will clarify this in future years. Although psychology as a discipline originated with Wundt's laboratory in Leipzig, Germany, in 1879, the profession of clinical psychology was considerably slower to develop in Europe than in the USA (Sexton and Misiak, 1976; Eysenck, 1990; Sexton and Hogan, 1992).

The major influence for the development of clinical psychology in many western European countries was the rebuilding of Europe after World War II in the 1940s and 1950s and the development of mental health facilities by European states, as the challenge of mental problems and mental health presented itself to health authorities (Lunt, 2000). In eastern European countries clinical psychology developed much later as these countries emerged in the 1980s from political regimes which were critical of the discipline (e.g., Pawlik, 1996). It is only relatively recently that these countries have begun to develop a profession of clinical psychology.

The Field of Clinical Psychology

Although there are differences between countries, it is possible to provide a general definition of the field which would be agreeable to all European countries; clinical psychologists may be defined as psychologists who

> apply psychology in a clinical context, usually a hospital, medical or community setting, with people (patients or staff) who consider themselves to be in need of a psychological perspective on their lives. In practice, the majority of clinical psychologists contribute to the assessment and treatment of people who see themselves as having psychological problems, such as those with mental health difficulties, but they also work with the handicapped, families, those with learning difficulties, and, more widely, with staff and organisations.
>
> Llewelyn, 1994

There are some differences in the practice of clinical psychology, from countries where the majority are employed by the state such as the United Kingdom where most clinical psychologists work for the National Health Service (NHS) and there is relatively little private practice, to countries such as Switzerland and Germany where many clinical psychologists work in private practice, with charges reimbursed by medical insurance companies. Furthermore, there are differences in the dominant activity of clinical psychologists: in many European countries the main activity is psychotherapy, whereas in others there is a broader role which includes assessment and other forms of intervention, and also a role in training and consulting with other staff. While the dominant paradigm informing clinical psychology in the United Kingdom is cognitive behavior therapy, other European countries such as France and Italy historically have been more influenced by psychoanalytic ideas. Although all European countries have been influenced by the USA, this influence has been perhaps most striking in the Scandinavian countries which embraced the 'scientist practitioner' approach to practice from an early stage.

Europe

As mentioned, the political face of Europe is developing and changing rapidly (see Lunt, 1998). Many of the central and eastern European countries have applied for membership of the European Union, and are working toward criteria which will enable them to join. This will mean an expanded and changed Europe in the not too distant future, bringing together the early members of the European Union and nonmember states; many of the latter in Eastern Europe and formerly belonging to the Soviet Union. The countries of Europe

constitute a very diverse group that may nevertheless be grouped into broad regions with some commonality in their organization and practice of clinical psychology. For these purposes, at a very broad level it is possible to identify a Nordic group of countries with considerable commonalities in their practice (the Scandinavian countries), Great Britain which has much in common with education and training in the USA and other parts of the English-speaking world, the German speaking countries, a southern European group, and an eastern European group. Of course, within these regions, the countries are characterized by their individuality and their diversity, and the different histories that have had a profound influence on the development of clinical psychology.

The Early Years of Clinical Psychology

There are a number of models in the field of clinical psychology, a field which emerged in European countries substantially after World War II (see above). It emerged as a recognizable profession at different times in different countries. In the United Kingdom the two World Wars provided a significant impetus to the emergence of this profession, initially through the need to develop psychological tests to recruit suitable personnel, and later in 1948 with the formation of the NHS that provided a considerable impetus to the development of this new profession. In these early days, the role of the clinical psychologist was largely one of laboratory technicians administering psychometric and other tests, usually for medical practitioners (Eysenck, 1990). However, in the United Kingdom, the development of behavior therapy, in particular under the influence of Hans Eysenck who established clinical psychology as a profession in England in 1949, led to clinical psychologists developing a therapeutic role, and by the 1960s they had become clinical practitioners in their own right.

In the Nordic countries, also, clinical psychology emerged substantially after World War II, in the 1940s and early 1950s; the Norwegian Psychological Association was founded in 1934, the Danish Psychological Association in 1947, the Swedish Psychological Association in 1955, and the Finnish Union of Psychologists in 1957. The formation of these Psychological Associations reflected the emergence in these countries of a profession of psychology, mainly clinical psychology, which progressed through similar broad major phases as the UK; that is, a phase focusing on diagnostic examination and testing of individual patients, a phase focusing on therapeutic work involving mainly psychotherapy, and a phase focusing on indirect work with other professional groups through techniques such as consultation and training.

In Spain, on the other hand, clinical psychology has had a shorter history (Belloch and Olabarria, 1994) and is said to have emerged in the 1970s along with fundamental changes in Spanish society. Similarly in Italy, clinical psychology as a professional application emerged substantially in the 1960s and there remain some tensions over the autonomy and role of clinical psychologists and their relationship with medical practitioners, especially psychiatrists. The Association of Greek Psychologists, founded in 1963, reflects the existence of professional psychology in Greece, although at that time and until very recently, all clinical psychologists received their training overseas, and the profession was very much dominated by the medical profession.

As mentioned above, clinical psychology emerged much later in Eastern European countries, and it was not until the 1980s that psychotherapy was fully recognized as a profession (Pawlik, 1996). Until recently, nearly all psychologists were employed in state institutions, though in recent years there has been a growth in demand which has resulted in a growth in private practice.

In all European countries, clinical psychologists represent the largest group of psychologists, and experienced a rapid growth in their number between the 1960s and the 1990s with the expansion of mental health provision and the growing awareness of the contribution of clinical psychology to a wide and diverse range of areas of work (EFPPA, 1997). Indeed, the 31 Member Associations of EFPPA (see below) represent 150 000 psychologists, the majority of whom are clinical psychologists.

Since the 1990s, in many European countries the emergence of the specialty of health psychology, with a focus on prevention rather than treatment, and the promotion of health rather than a more therapeutic or curative function, has led to attempts to define and differentiate a new field of psychological activity within the mental health field. This has been supported by the WHO commitment to Health for All by the Year 2000; EFPPA has had three Task Forces, focusing respectively on clinical psychology, health psychology, and psychotherapy, which have drawn up a model defining overlap and separate areas of activity within the health field, with corresponding commonalities and differences in education and training (EFPPA, 1997).

Training of Clinical Psychologists

Education and training of clinical psychologists varies in European countries though, again, there are major regional groupings. Across Europe, as in other regions of the world, the education and training period has increased, with moves in some European countries for doctorate training for clinical psychologists, and more demanding requirements in all countries for the internship period. In the United Kingdom, there is a strong commitment to a scientist-practitioner model: 'the clinical psychologist is first and foremost an 'applied scientist' or 'scientist-practitioner' who seeks to use scientific knowledge to a beneficial end' (Marzillier and Hall, 1990). This commitment also characterizes the Nordic countries which have been influenced substantially by the USA, and also by the UK. These countries would sign up to the definition that:

> clinical psychologists share several common attributes. They are psychologists because they have been trained to use the guidelines of knowledge of psychology in their professional work. They are clinicians because they attempt to understand people in their natural complexity and in their continuous adaptive transformations … they are scientists because they utilize the scientific methods to achieve objectivity and precision in their professional work. Finally, they are professionals because they render important human services by helping individuals, social groups, and communities to solve psychological problems and improve the quality of life.
>
> Kendall and Norton-Ford, 1982: p. 4

This model is also espoused by Spanish clinical psychology where Belloch and Olabarria (1994) state that clinical psychology training is 'very similar to that proposed in the 1940s, in the famous Boulder Conference, organised by the APA.' In France, where there is a strong clinical psychoanalytic tradition, there is less of a commitment to the 'Boulder' model, and more of a philosophical or hermeneutic tradition in relation to education and practice.

However, since 1957 when the Treaty of Rome provided the foundation for the European Community (later Union), there have been requirements on individual countries ('states') to provide procedures for the mutual recognition of psychologists' qualifications across national boundaries (see McPherson, 1988). Wider moves within the European Union such as the Bologna Agreement, which was signed by 29 Ministers in 1999 and commits them to greater convergence in terms of university degree structures, mean that even within Europe there is likely to be greater similarity in terms of structures of education and training for clinical psychologists.

Professionalization of Clinical Psychology

The period since the 1950s has seen a greater professionalization of clinical psychology in all European countries, with the development of codes of ethics (Lindsay, 1996) and increased regulation and laws for clinical psychologists (Lunt, 2000) across European countries. These political and professional developments have been supported by EFPPA which has brought together clinical psychologists from all over Europe to work on the professional aspects of practice at a European level and to support individual European countries seeking to develop their ethical codes, laws protecting the title of psychologist, and education and training in clinical psychology.

Organization of Clinical Psychology

It is also possible to trace the history of clinical psychology through its organization in Europe (Gilgen and Gilgen, 1987; Lunt, 1998). Many European countries founded scientific societies for psychologists in the early twentieth century whose purpose was to foster research and psychological science. At the time of World War II, separate professional associations to meet the needs of professional psychologists, often mainly clinical psychologists, were founded in a number of European countries. These associations had, as a focus, issues concerning professional practice and emerged in some countries as Trade Unions, negotiating salaries and terms and conditions of work for clinical psychologists as well as wider professional issues such as regulation, legislation, and ethical codes. In 1981, at a time when the provisions of the European Community demanded that European member states encourage mobility of professionals across Europe, the European Federation of Professional Psychologists Associations was formed to bring together professional associations in Europe and to collaborate on matters of common professional concern.

EFPPA currently has 31 member associations representing all the countries of the EU, all other countries in western Europe, and a growing number of member associations from central and eastern Europe. The federation provides a unique opportunity for comparison between the practices of different European countries and a forum for discussion and debate of important issues. The formation of EFPPA in 1981, when matters of mobility and mutual recognition were becoming more pressing, was due to a realization by psychologists and psychology associations of member states that a federation would provide a professionally and politically useful way to move forward and to begin to develop common policies in this area. Clinical psychologists, in particular, were faced with the growing prospects of mobility between countries in the European Union, and the implications of the Treaty of Rome that provided the foundation for the European Community in 1957.

As a federation of Professional Psychology Associations representing around 150 000 professional psychologists in Europe, EFPPA spends most of its efforts on clinical psychology and clinical psychologists, also the largest group of psychologists within Europe (as in the rest of the world). In many countries psychology has become one of, if not the, most popular subjects to study at university. The majority of students studying psychology aspire to become clinical psychologists, and for this reason, in many European countries, there is an oversupply of qualified practitioners. Many countries now operate a so-called *numerus clausus*, either at the start of the psychology study, or during the study. This controls the numbers in training. In the United Kingdom, where specialist training in clinical psychology is funded by the NHS – where the vast majority of clinical psychologists work – the number of 'trainee' posts is strictly limited and is planned according to staffing needs in the different regions of the country. In other countries, where there is a tradition of predominantly private practice in clinical psychology, there are large numbers of qualified psychologists unable to find work. In all European countries, the ratio of female to male students is between 6:1 and 3:1, leaving the profession in danger of becoming an almost feminized profession in the future (Schorr and Saari, 1995). There has also been some difficulty in many countries in recruiting students from the range of ethnic groups represented in Europe's increasingly multiethnic population. This clearly has implications for the clinical treatment of different client groups.

The Future

As new fields of practice in the health field evolve, in particular health psychology, there are pressures on clinical psychology. For example, in some countries it has been said that there may no longer be a field of clinical psychology, since there are strong moves toward a broader field of health psychology and a greater focus on preventive work. These newer areas, such as health psychology, neuropsychology, and forensic psychology, are leading to greater specialization within and outside clinical psychology. Although European countries differ in the extent and nature of their specialisms within the health field, there is an increasing trend for specialization and demands for higher qualifications. In one respect, psychologists working within the health system could be said to be becoming more generic,

while on the other hand there are increasing specializations within this field of work.

Summary

The brief 50-year history of clinical psychology in Europe has seen an enormous increase in numbers both of students and of practitioners, such that the majority of psychologists are now clinical psychologists. This rapid professionalization has been accompanied by higher qualifications, greater regulation, the development of ethical codes, and all the characteristics of traditional professions. However, as the number of clinical psychologists increases, and the question of mobility across Europe becomes more pressing, there will be increasing attempts to develop more common frameworks and standards for education and practice; the challenge will be to achieve a balance between allowing individual countries their own autonomy which reflects their differing history and culture ('subsidiarity' as it is called), and developing more common agreed frameworks of practice which reflect a possible future 'federalization' of Europe.

See also: Clinical Psychology in North America, History of; Freud, Sigmund (1856–1939); Psychoanalysis, History of; Psychotherapy, History of: Psychiatric Aspects; Training in Clinical Psychology in the United States: Practitioner Model; Training in Clinical Psychology in the United States: Scientist–Practitioner Model.

Bibliography

Belloch, A., Olabarria, B., 1994. Clinical psychology: current status and future prospects. Applied Psychology: An International Review 43 (2), 193–211.

Drenth, P.J.D., Sergeant, J.A., Takens, R.J., 1990. European Perspectives in Psychology, vol. 2. Wiley, Chichester, UK.

EFPPA, 1997. Report to the General Assembly of EFPPA of Clinical Psychology Task Force. Available from EFPPA secretariat.

Eysenck, H., 1990. Clinical psychology in Europe and in the US: development and future. In: Drenth, P.J.D., Sergeant, J.A., Takens, R.J. (Eds.), European Perspectives in Psychology, vol. 2. Wiley, Chichester, UK.

Gilgen, A.R., Gilgen, C.K. (Eds.), 1987. International Handbook of Psychology. Greenwood Press, New York.

Kendall, P.C., Norton-Ford, J.D., 1982. Clinical Psychology. Scientific and Professional Dimensions. Wiley, New York.

Llewelyn, S.P., 1994. Assessment and therapy in clinical psychology. In: Spurgeon, P., Davies, R., Chapman, A. (Eds.), Elements of Applied Psychology. Harwood Academic Publishers, Chur, Switzerland.

Lindsay, G., 1996. Developing an ethical psychological practice. In: Georgas, J., Manthouli, M., Besevegis, E., Kokkevi, A. (Eds.), Contemporary Psychology in Europe. Hogrefe & Huber, Göttingen.

Lunt, I., 1998. Psychology in Europe: challenges and opportunities. The European Psychologist 3 (2), 93–101.

Lunt, I., 2000. Psychology as a profession. In: Pawlik, K., Rosenzweig, M. (Eds.), The International Handbook of Psychology. Sage, London.

Marzillier, J., Hall, J., 1990. What Is Clinical Psychology?, second ed. Oxford University Press, Oxford, UK.

McPherson, F., 1988. Psychologists and the EEC. The Psychologist 9, 353–355.

Pawlik, J., 1996. The situation of the psychologists in Eastern European countries today. In: Georgas, J., Manthouli, M., Besevegis, E., Kokkevi, A. (Eds.), Contemporary Psychology in Europe. Hogrefe & Huber, Göttingen.

Poortinga, Y., 1996. Cultural diversity in Europe: extrapolations from cross-cultural research for professional psychology. In: Georgas, J., Manthouli, M., Besevegis, E., Kokkevi, A. (Eds.), Contemporary Psychology in Europe. Hogrefe & Huber, Göttingen.

Pilgrim, D., Treacher, A., 1992. Clinical Psychology Observed. Routledge, London.

Schorr, A., Saari, S. (Eds.), 1995. Psychology in Europe. Hogrefe, Göttingen.

Sexton, V.S., Hogan, J.D. (Eds.), 1992. International Psychology. Views from Around the World. University of Nebraska Press, Lincoln & London.

Sexton, V.S., Misiak, H. (Eds.), 1976. Psychology Around the World. Brooks/Cole, Monterey, CA.

Clinical Psychology in North America, History of

Donald K Routh, University of Miami, Coral Gables, FL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 2029–2036, © 2001, Elsevier Ltd.

Abstract

The first psychology clinic in the world was founded by Lightner Witmer at the University of Pennsylvania in 1896. Clinical psychologists in the United States first organized themselves in 1917. This field was greatly expanded after World War II with funds from the US government for training. In Canada and Mexico, clinical psychology developed mostly after World War II. Thus, a 1965 conference in Canada endorsed the doctoral training model previously developed at a 1949 conference in the US. In French-speaking Canada, masters level training was more typical, whereas in Mexico a 6-year 'licentiate' degree prevailed. Today, in North America as elsewhere in the world, clinical psychology is concerned with the study of psychopathology and with its diagnosis and treatment.

Clinical psychology is concerned primarily with the study of psychopathology and with its diagnosis and treatment. It shares this domain with several other mental health disciplines, including psychiatry, social work, nursing, and various types of counseling. Compared to these other disciplines, clinical psychology is distinctive for its training in research and for its expertise in psychometrics and the behavior therapies. North America played an important role in the emergence of clinical psychology. The field usually dates its origin from the founding of the first psychology clinic in 1896 by Lightner Witmer (1867–1956) at the University of Pennsylvania (Routh, 1996). Clinical psychology in the English-speaking parts of Canada developed in a pattern similar to that seen in the US (with doctoral training required for independent practice) but somewhat later in time. The field developed in French-speaking Canada and in Mexico in a way more resembling that of European countries, with master's or licenciate-level training required for independent practice. The North American Free Trade Agreement now exerts pressure on all three countries and their states and provinces to coordinate these differences to a greater degree, to permit more freedom of movement to qualified clinical psychologists.

The Prehistory of the Mental Health Field

The need for humans to deal with the problems now called mental illness did not emerge suddenly a century ago. It seems reasonable to assume that such problems have existed in some form in every society through all the millenia of human experience. The ancient literature of India, Egypt, China, Greece, and Rome contain descriptions of disturbed behavior, often interpreted in religious terms as some type of retribution by magical or divine forces. Legal systems as they developed in all of these civilizations necessarily included provisions for seeing to the management of the affairs and property of persons who were temporarily or permanently unable to manage for themselves (Routh, 1998).

Greek Ideas Concerning Psychopathology

Western concepts of psychopathology have their roots in those of the ancient Greeks, including the writings attributed to the physician Hippocrates (460–377 BC). These Hippocratic writings include terms such as melancholia, mania, paranoia, and dementia, with meanings not all that different from their present ones, albeit with different explanations. For example, in Greek, the word 'melancholia' simply means 'black bile.' In the Hippocratic theory of the humors, a person suffering from severe mental depression had an overabundance of black bile, a substance thought to be produced by the spleen. Within this system, one aspect of treatment quite reasonably aimed to reduce the amount of black bile by administering a purgative such as hellebore. Another disorder was that of 'phrenitis,' literally meaning an inflammation of the mind. This referred to mental disturbance accompanied by fever, and the approach taken was simply to wait for the fever to abate. Though some of these Hippocratic ideas may seem strange to us now, it is worth reflecting why some of them lasted well into the eighteenth century and beyond.

Emergence of Psychiatry

Although according to Herodotus (*c*.484–425 BC), medical specialties existed even in ancient Greece, the one we now call psychiatry did not emerge until the late eighteenth century in Europe. The ancients did not conceptualize mental disorders as a separate category but regarded them as being illnesses like any other. When psychiatry did emerge, with the work of such pioneers as Philippe Pinel (1745–1826), Benjamin Rush (1745–1823), and Vincenzo Chiarugi (1759–1820), it was associated with the development of mental asylums or hospitals as a separate locations for the care of those with mental derangements. The theory of 'moral treatment' that was typical of that time tried to minimize the use of coercive methods such as chaining patients to restrain them and instead insisted that they be treated with kindness and courtesy. It was often found that even some very disturbed patients responded positively to such a regimen.

Modern Psychology and the Study of Psychopathology

Long before a formal discipline of psychology existed, people in every society still no doubt reflected upon human experience

and behavior. As was the case of Hippocrates in relation to medicine and psychiatry, the influence of ancient Greek philosophers such as Plato (427–347 BC) and Aristotle (384–322 BC) upon our present psychological concepts was pervasive. There is conventional agreement that psychology emerged as a formal academic discipline only in the mid-nineteenth century in Europe. Wilhelm Wundt (1832–1920) of the University of Leipzig is usually named as the founder of the field, and 1879, the year in which he set up his psychology laboratory there, is celebrated as the key event in its origin. Wundt's work and those of the other early psychologists often focused on sensory processes, reaction time, and memory. It is also noteworthy that the study of psychopathology was a possible topic of psychological study even in those days. The eminent psychiatrist Emil Kraepelin (1856–1926) was influenced by Wundt's writings and was interested in psychology. His main motive for becoming a psychiatrist was that this was the only way he could see to make a living while doing psychological research. He later actually studied under Wundt, who encouraged him to keep on with his work combining psychology and psychiatry. Kraepelin set up psychology laboratories at his psychiatric clinics in both Heidelberg and Munich.

Many of the pioneers in psychology in both Europe and the US were trained as medical doctors. In France these included Theodule Ribot (1839–1916), who wrote about diseases of memory and about personality disorders. An important French colleague was Pierre Janet (1859–1947), who studied anxiety, hysteria, and obsessions and developed concepts of dissociation that continue to be influential today. In the US the leading pioneer in psychology, William James (1842–1910) was originally trained in medicine but wrote a psychology textbook that proved to be the most influential of all. In 1896, James gave his Lowell lectures on exceptional mental states, much influenced by the work of Janet. Boston neurologist Morton Prince also became interested in Janet's writings and published a description of a woman with multiple personalities who had been his patient. Prince established the *Journal of Abnormal Psychology* in 1906 and later gave it to the American Psychological Association. Subsequently, in 1926, he established the Harvard Psychological Clinic, which was a research facility rather than one delivering mental health services. The most influential medically trained student of psychology of this time was no doubt Sigmund Freud (1856–1939). Breuer and Freud's book *Studies in Hysteria*, was published in 1895 and Freud's book on the interpretation of dreams in 1900. The first international psychoanalytic meeting was held in Salzburg in 1908. In 1909, Freud came to the US for the first and only time.

Lightner Witmer and Clinical Psychology

As the above paragraphs make clear, Witmer was hardly the first to suggest that psychologists study psychopathology. Instead, his main contribution was to go beyond that to advocate that psychologists try to help people as well as study them. Witmer had been an undergraduate at the University of Pennsylvania and then for a time, before going to Leipzig to obtain his PhD under Wundt, served as a school teacher. He had as a student a young man with marked difficulty in reading and was able to help the youngster succeed in school and go on to attend college. This turned out to be a formative experience for Witmer. After Witmer had obtained his PhD and returned to his *alma mater* as a psychology professor, a school teacher named Margaret Maguire asked his advice about one of her pupils with a spelling problem. Witmer reasoned that if psychology was of any practical use, it should be able to be of help in a case of this kind. Thus was the psychological clinic and the field of clinical psychology launched.

Witmer's clinic worked more with children than with adults and tended to concentrate on academic difficulties such as reading, spelling, or general backwardness in school as opposed to emotional or behavioral problems. His historic forebears are thus not Hippocrates and Pinel but rather eighteenth and nineteenth century French physicians and special educators such as Jacob Pereire (1715–80) (who taught deaf-mutes to speak), J.M.G. Itard (1775–1838) (who worked with the 'wild boy' of Aveyron), and Edouard Seguin (1812–80) (a physician who devised a 'physiological method' of sensory and motor training in an attempt to remediate mental retardation). Witmer used existing laboratory procedures including the Seguin formboard and sensory/motor procedures adapted from Wundt to evaluate the children referred to him and often tried to teach them simple tasks as a part of his diagnostic efforts. In his treatment activities he often collaborated with school teachers, as well as with physicians, thus serving as more of a consultant than doing anything resembling present-day psychotherapy. As a matter of fact, he was little influenced by the activities of the Boston School of psychotherapy that was contemporary with his work, nor later by Freud and his psychoanalytic movement.

Witmer is not remembered for any noteworthy scientific discoveries but rather for his persistence in enacting this new role of the clinical psychologist. At the University of Pennsylvania's PhD program, he essentially trained most of the first generation of clinical psychologists. He maintained his clinic as a service and training facility and in 1907 began a journal, the *Psychological Clinic*, to publicize these activities (Witmer, 1907).

The Binet Test

There is still no consensus among psychologists as to precisely how to interpret its findings. Still, Alfred Binet's 'metric scale' of intelligence (Binet and Simon, 1905) may be the most noteworthy piece of technology developed by psychology in its first century. Certainly it had a major impact on the new field of clinical psychology. In fact, before World War II, probably the most characteristic activity of the typical clinical psychologist was the administration of the Binet test and other similar measures (Routh, 1994). This was true despite the fact that Lightner Witmer, the founder of the field, was quite critical of the Binet test and used it only as one part of his extensive battery of laboratory procedures.

In the light of how influential his test was, it is interesting to note that Alfred Binet himself was not particularly identified with the field of clinical psychology. Originally trained as a lawyer, Binet became part of the circle around the influential neurologist Jean Charcot at the Salpetriere in Paris. Much of his psychology was self-taught, through extensive reading at the

Bibliothéque Nationale. In France, Binet became known as one of the founders of the entire field of psychology (often characterized worldwide as 'experimental psychology') and edited the influential journal, *Annee psychologique*. As is well-known, Binet's successful attempts to devise an intelligence test departed from the conventional approach of using relatively 'pure' sensory and motor tasks to use complex worksamples of the kinds of things school children might be expected to know and do.

For some reason, Binet's new test did not create as much of a stir in his homeland as it did in the US. Psychologist Henry Goddard, who directed the psychology laboratory at the Vineland Training School, in New Jersey, had the Binet test translated and soon confirmed its impressive validity in identifying persons with mental retardation. The use of the Binet spread like wildfire among the early clinical psychologists in the US, beginning with those employed in the field of mental retardation. Goddard founded the first psychology internship in 1908 at Vineland, NJ. Goddard went on to become the first professor of clinical psychology at Ohio State University, like the University of Pennsylvania an important early training center in the field (Routh, 1994).

Lewis M. Terman (1877–1956) at Stanford University developed a standardized version of Binet's test, collected normative data for it, and introduced certain refinements such as the ratio IQ score (originally suggested by Wilhelm Stern of Hamburg). The 1916 Stanford-Binet, as it was called, dominated this field for many years. At about the same time, 1915, Robert Yerkes pointed out the unsuitability of the concept of mental age and of this testing format for use with adults and introduced his own 'point scale' as a substitute for it. Yerkes and his colleagues were also responsible for the development of group intelligence tests, the Army Alpha (for those who could read) and Army Beta (for the illiterate), used for mass testing of military recruits during World War I. Another wartime development was Robert S. Woodworth's Personal Data Sheet published in 1917. This was one of the first rationally developed self-report questionnaires intended to detect neurotic tendencies (Routh, 1994).

The Child Guidance Center Movement

The development of child guidance centers was another factor that influenced early clinical psychologists in the direction of working with children more than with adults. The first child guidance clinic was the Institute of Juvenile Research, established in 1909 by physician William Healy in conjunction with the juvenile court of Chicago. The idea behind such facilities was that careful clinical study of children engaging in antisocial activities could assist in guiding them away from crime. Healy was joined at first by clinical psychologist Grace Fernald and subsequently by her replacement, Augusta Bronner (Healy and Bronner, 1926). The child guidance clinic was the origin of the 'clinical team' of psychiatrist, psychologist, and social worker that later spread to other settings. The typical pattern was that the psychiatrist saw the child, the social worker saw the family, and the psychologist did the testing. The child guidance movement was supported by the Harkness family's philanthropy in the form of the Commonwealth Fund and replicated in many US cities and abroad.

The First Clinical Psychology Organization

In 1917, in Pittsburgh, a group of eight clinical psychologists organized themselves into what they called the American Association of Clinical Psychologists and invited 48 colleagues to join them (Routh, 1994). They were led by J.E.W. Wallin (1876–1969) and Leta Hollingworth (1886–1939). Incidentally, Hollingworth was the first to suggest in 1918 that a person trained in clinical psychology receive a distinctive type of degree, the doctor of psychology. The new organization was viewed by many as divisive and was soon incorporated into the American Psychological Association as its Clinical section. An effort by the same group to introduce procedures for certifying qualified clinical psychologists failed, however.

Psychometric Developments

The interwar years were a fertile time for the emergence of various new psychometric procedures, many of which continue to be in use today. For example, in 1921, the Swiss psychiatrist Herman Rorschach (1884–1922) published his well-known inkblot test. It was brought to the US by a child psychiatrist who taught it to a clinical psychology graduate student at Columbia named Samuel Beck (1896–1976). Beck then proceeded to do his dissertation on this new test and eventually to develop his own system for administering and scoring it. Psychologist Bruno Klopfer (1900–71), a disciple of Carl Jung (1875–1961), also introduced the Rorschach to the US and developed a separate system for administering and scoring it (*see* Projective Methods in Psychology). In 1936 the Thematic Apperception Test was introduced by Henry A. Murray (1893–1988) of the Harvard Psychological Clinic, and a colleague. Also in 1935 Edgar A. Doll (1889–1969), introduced the Vineland Social Maturity Scale, an interview-based method involving informants familiar with the person, for assessing the social competence of individuals suspected of mental retardation. David Wechsler (1896–1981) published the original version of his Wechsler-Bellevue intelligence test for adults. This was but the first of many Wechsler tests of intelligence and memory. It introduced the use of the deviation IQ, a standard score comparing the individual to age-matched normative subjects. In 1943, psychologist Starke R. Hathaway and psychiatrist J.C. McKinley introduced the first edition of the Minnesota Multiphasic Personality Inventory (MMPI). The MMPI had novel 'validity' indicators, and its measures of psychopathology were empirically keyed to psychiatrically defined groups (*see* Minnesota Multiphasic Personality Inventory).

Organizational Activities

In 1937 a new organization known as the American Association of Applied Psychology (AAAP) split off from the American Psychological Association (APA) to provide a home for various

professionally oriented groups including the clinical psychologists, who at this same time dissolved the Clinical Section of the APA. The AAAP began to publish the *Journal of Consulting Psychology*, which subsequently developed into a high-prestige clinical psychology journal (Routh, 1994).

It was also at this time that some preliminary developments in psychology began in other parts of North America. In 1937, for example, the first psychology curriculum was devised at UNAM, the National Autonomous University of Mexico, in Mexico City. In 1939, the Canadian Psychological Association was founded. It had 38 members to begin with, and it has been estimated that there were only 53 psychologists in all of Canada at the time. Needless to say, there were few prewar developments in Canada or Mexico specifically relating to clinical psychology.

The Post-World War II Boom in US Clinical Psychology

Clinical psychology expanded so greatly in the US after World War II that many brief historical accounts of the field even consider its development to have begun at that time. The war effort tended to draw everyone into it, either on the battlefield or on the home front. Many psychologists whose interests prior to the war had been strictly in research and in the academic side of the field found themselves assigned to carry out psychological testing or to help medical staff in treating psychiatric casualties. After the war, it was clear that the Veterans Administration (VA) would have to be vastly expanded to deal with the need for residential care, psychotherapy, or at least vocational counseling of some of those returning from military service.

In 1945, the VA and the newly established National Institute of Mental Health in the US came to the APA to ask it to establish a system for accrediting training programs in clinical psychology. The government intended to pour millions of dollars into training such individuals and needed to know which programs were competent to carry this out. In response, APA created a system of accreditation, and for the first time, it began to be possible to say who was a well trained clinical psychologist and who was not. David Shakow (1901–81) was the architect of the 1949 conference held in Boulder, Colorado, which ratified what has come to be called the 'scientist–practitioner' model of training clinical psychologists (Raimy, 1950) (*see* Training in Clinical Psychology in the United States: Scientist–Practitioner Model).

Postwar clinical psychologists continued in their role as mental testers, but gave more emphasis to the assessment of personality and psychopathology, not just cognitive status. This was the heyday of projective tests, and at least for a time the Rorschach inkblot was an appropriate symbol for the clinical practitioner of psychology. A well-known exemplar of the clinical psychologist as projective tester in this era was David.

Rapaport (1911–60), chief psychologist at the Menninger Clinic in Kansas. A two volume set of books published at this time by Rapaport and coworkers (Rapaport et al., 1945, 1946) established the Rorschach, the TAT, and the Wechsler test to be a 'full test battery' for almost a half century to come.

The clinical psychologists of this era were also eager to become full-fledged psychotherapists as well as mental testers. Their route to therapeutic training was blocked to some extent by the American Psychoanalytic Association's 1938 policy decision (contrary to Freud's own views) that only psychiatrists were to be trained in psychoanalysis. Clinical psychologists thus became very ingenious in devising ways of becoming therapists. The best known of them was perhaps Carl Rogers (1902–87), whose original therapy supervisor was the social worker Jessie Taft. Taft, in turn, had received her training from Otto Rank, a psychologist from Vienna who had received orthodox psychoanalytic training there and had been a close colleague of Freud's. Rogers was successfully assertive in other ways. At one point in his career he was director of a child guidance clinic when such administrative positions were supposed to be held only by physicians. Rogers also was determined to combine his psychology training with his role as a therapist. He was among the first to produce recordings of actual psychotherapy sessions, and was a pioneer in doing controlled research on the outcome of psychotherapy. Rogerian therapy ('client centered' or 'person centered', therapy, as it was later called) is still practiced and studied both in North America and elsewhere (Routh, 1994, 1998).

In 1945, the first state law certifying psychologists for independent practice was passed by Connecticut. By 1977, all states in the US had passed such certification or licensing laws regulating the use of the title, 'psychologist' or the practice of psychology (Routh, 1994).

The Behavior Therapy Movement

Some psychologists were of the opinion that clinical psychologists in their professional activities should not simply try to duplicate the activities of psychiatrists. In 1913, John Watson had boldly proclaimed a behavioristic approach to psychology, which was widely influential at least in academic psychology in the US. In a famous paper reporting research carried out under Watson's supervision, Jones (1924) described the case of 'Peter', whose fear of rabbits she desensitized. Not even Jones herself realized the wider implications of this study at the time, but in the light of events many years later some considered her to have been 'the mother of behavior therapy' (*see* Behavior Therapy: Psychiatric Perspectives).

The behavior therapy movement progressed not only by following its own agenda but by attacking its opponents. Psychologist Eysenck (1952) in England thus skeptically reviewed the evidence for the effectiveness of psychotherapy. It was not enough, he noted, to show simply that psychotherapy patients improved. One also needed to consider the rate of spontaneous improvement of patients who did not receive psychotherapy.

In 1962 in Charlottesville, Virginia, a behavior therapy conference was held, sponsored by psychiatrist Joseph Wolpe and psychologists Andrew Salter and Leo Reyna. At the time Wolpe had just published a book on his success in treating patients with phobias using the behavioral method of systematic desensitization. Soon afterward, Lang and Lazovik (1963) published the first controlled study of desensitization, in treating snake phobia. Soon the behavior therapy movement was in full swing, with its own organizations, journals, and many

adherents. Sidney Bijou and his colleagues established behavioral treatments based on the research of B.F. Skinner. These came to be known as applied behavior analysis and were especially influential in work with the behavior disorders of children and of those with mental retardation (*see* Behavior Analysis: Applied).

Canadian Clinical Psychology

It was in the 1960s that clinical psychology finally came of age in Canada. It was and is the largest applied specialty in psychology numerically in that country, as it is in the rest of the world. In 1965, the Couchiching Conference basically endorsed the 'Boulder model' of scientist–practitioner training for clinical psychology. Some Canadian doctoral programs such as the one at McGill even sought accreditation by the American Psychological Association. There was even something of a boom north of the border. It is said that by 1966, more than half the doctoral psychologists in Canada were either American born or trained in the US. In 1983, the Canadian Psychological Association established its own program of accreditation. The first CPA-accredited doctoral programs in clinical psychology were those at McGill, Concordia, and Simon Fraser Universities. The success of doctoral training in clinical psychology tells only part of the story there, however. In Quebec and some other eastern provinces, the master's degree was accepted as the entry level of training for independent clinical practice. By 1996, Canada had 88 graduate training programs in professional psychology (including clinical): 57 doctoral and 31 terminal masters.

Canadian clinical psychology is noted for particular strength in the area of neuropsychology, which built on Canadian strengths in the neurosciences, including the work of neurosurgeon Wilder Penfield at the Montreal Neurological Institute. In academic psychology, the research and writings of Donald Hebb concerning the CNS (conceptual nervous system) were influential. It was at McGill University that Olds and Milner published their famous 1954 paper on the reinforcement of an animal's behavior by electrical stimulation of its brain. On the clinical side, Ronald Melzack elaborated his theory of gating mechanisms influencing the experience of pain. Brenda Milner explored the role of the hippocampus in semantic memory, including work with her famous patient, 'H.M.' This man developed permanent memory deficits after surgery inadvertently destroyed his hippocampus bilaterally. Doreen Kimura documented the left cerebral hemisphere advantage in dichotic listening.

Clinical Psychology in Mexico

In 1997, the Division of Clinical Psychology of the APA held its midwinter board meeting in Mexico City, hosted by Juan Jose Sanchez-Sosa, the director of the school of psychology at UNAM, the National Autonomous University of Mexico. Sanchez-Sosa provided his US colleagues with a tour of this school, itself as large as many an entire college campus in the US. The school offers both Master's and PhD degrees in psychology, but as in much of Europe, these degrees are intended for those headed for academic and research careers. Those who intend to practice psychology, including clinical psychology, in Mexico need only a 'licentiate' or diploma to do so, which is awarded after 6 years of what to persons trained in the US seems to be undergraduate training. But students in such a program spend essentially full time on psychology, without the need for a broad liberal arts distribution of courses. This includes a significant amount of practicum experience. The psychology clinic, one of the practicum facilities used at UNAM, features a variety of clinical activities, including psychological testing, psychodynamic therapy, behavior therapy, group therapy, and even biofeedback. Since there is no certification or licensing system beyond the licentiate degree itself, it is difficult to be sure how many of these graduates are practicing clinical psychology in a way parallel to what would be seen in the US or Canada.

Independent Practice of Clinical Psychology

Although Lightner Witmer in his clinical work often collaborated with school teachers, physicians, or others, psychologists working in his clinic were never supervised by members of any other profession. This tradition of independent work has continued within the field of clinical psychology, somewhat in contrast with social work and nursing. The post-World War II expansion of the field in the US was primarily in the public sector, typically VA hospitals, but also child guidance centers and eventually community mental health centers. The large government training grants supporting clinical psychology programs at the time presupposed that the graduates would go to such public sector jobs or teach in colleges and universities. In the 1980s, the Reagan administration made most such training grants a thing of the past (Routh, 1994).

David Mitchell, a PhD student of Lightner Witmer, was one of the first individuals to make his living primarily in the private practice of psychology. Eventually, he was joined by many others. After all, the state and provincial laws that developed to regulate psychology after 1945 specified what qualifications were necessary to offer one's services to the public as a psychologist. Many psychologists trained in Boulder-model PhD programs did no research after graduation, and eventually the idea of training psychologists as practitioners rather than scientist–practitioners emerged. Beginning with the University of Illinois in 1966, a number of programs began to offer the doctor of psychology (Psy.D.) degree rather than the PhD The conference in Vail, Colorado in 1973 officially legitimized such practitioner training for the first time. In fact, a number of nonuniversity-affiliated schools of professional psychology sprang up beginning in the 1970s, many of them offering Psy.D. degrees. Often these programs were supported only by student tuition, and many students assumed substantial loans to finance their education (Routh, 1994). No such private school of professional psychology has emerged in Canada (nor in Mexico) (*see* Training in Clinical Psychology in the United States: Practitioner Model).

As psychologists in private practice emerged in larger numbers, they also became more active politically. The first practitioner became president of the APA in 1977, and within 20 years the first Psy.D. was elected to this position.

The practitioners began to dominate both the APA and the Canadian Psychological Association to a greater and greater extent. In response, many academic psychologists retreated to form national organizations of their own that were more research-oriented. Thus, the American Psychological Society was founded in 1988. Similarly, in 1989, academic and research psychologists in Canada founded the Canadian Society for Brain, Behavior, and Cognitive Science.

Practicing psychologists both in the US and Canada battled psychiatrists for their share of the mental health 'market.' Thus, they fought to obtain hospital privileges. They supported 'freedom of choice' legislation to become eligible as health providers reimbursable by health insurance companies and Health Maintenance Organizations. In 1988, a law suit by clinical psychologist Bryant Welch and others forced the American Psychoanalytic Association to begin to admit psychologists for training at its local institutes (Routh, 1994). Most recently, a number of practicing psychologists in the US, led by Patrick DeLeon, have been trying to obtain the right to prescribe medications for mental health conditions, so far without much success.

The Continued Commitment of Clinical Psychology to Research

Well before the founding of Witmer's clinic, there was a strong interest on the part of psychology in research on psychopathology, including its diagnosis and treatment. This continues to be the case. In fact, psychologists are far more likely than psychiatrists or those in other fields to be principal investigators on research grants from the National Institute of Mental Health. Such research is international in scope and is a collaborative interdisciplinary enterprise. The turf battles that characterize the marketplace of practice are much less typical in the research arena, where clinical psychologists often cooperate smoothly with experimental psychologists and statisticians as well as with medical colleagues. In research, it is neither possible nor necessary to draw any bright line to show the boundaries between these fields.

The diversity of such research is so great that it would be impossible to cover it in an article as brief as this one. Instead, a few examples must suffice. In 1954, psychologist Evelyn Hooker received her first NIMH grant to study homosexuality. Her work, much of it using the types of projective testing that were typical of clinical psychological work at the time, suggested that homosexuals might be essentially normal psychologically. This research was related to the later decisions to delete homosexuality as a pathological category from the Diagnostic and Statistical Manual of the APA.

In the 1950s, psychologist Leonard Eron began collecting peer-rating data on aggressive behavior in 8-year-old children. The subsequent longitudinal research he and his colleagues did helped establish the fact that aggressive behavior is highly stable well into adulthood. Psychologist Gerald R. Patterson studied aggressive behavior in children using direct behavioral observations, relating it to coercive processes in the parent–child dyad and doing controlled intervention studies showing how it could be reduced.

In the late 1950s, psychologist C. Keith Conners devised a simple teacher rating scale for the assessment of varied types of disordered behavior in school children. Together with child psychiatrist Leon Eisenberg he helped carry out the first controlled studies of the effects of stimulant medications on children's disruptive behavior (Conners and Eisenberg, 1963). Such research formed an important basis of present concepts of Attention Deficit Hyperactivity Disorder, the most commonly diagnosed type of child psychopathology and one still typically treated with stimulant medications. By the 1970s, Thomas Achenbach and his colleagues had begun to develop the use of parent, teacher, and self-ratings for child behavior into the most widely used forms of assessment by mental health workers.

In 1962, Meehl published a classic paper on 'schizo-taxia, schizotypy, and schizophrenia,' elaborating his concepts as to how genetic factors might be involved in the development of this disorder (Meehl, 1962). In 1973, psychologist Holzman and coworkers announced their discovery of smooth pursuit eye-movement difficulties in patients with schizophrenia (Holzman et al., 1973). This neurological symptom proved to be an important 'trait' marker in the first degree relatives of schizophrenics as well, whether or not they manifested any overt psychopathology. In Denmark, psychologist Sarnoff A. Mednick carried out a series of studies using the excellent public registers that characterize that country to do longitudinal, epidemiological studies of schizophrenia, implementing the type of research that Meehl had only been able to imagine.

Problems of the dysregulation of affect and emotion are the most ancient in the field of psychopathology, having been with us since Hippocrates. Beginning in the 1960s, psychologist Richard Lazarus elaborated his concepts of stress, appraisal, and coping and demonstrated experimentally how his subjects' use of different coping strategies could dampen or heighten their physiological stress (Lazarus, 1966). Beginning in the 1970s, Charles Spielberger and his colleagues began the development and validation of measures of state as well as trait anxiety and later of state and trait anger as well. In the 1970s, Martin E.P. Seligman and his colleagues showed how his studies of learned helplessness in dogs could be used to reconceptualize human depression.

In conclusion, clinical psychology has both international and important North American roots. In its first century, it has developed into a viable science and profession, and there is good reason to suppose that its trajectory will continue in the twenty-first century.

See also: Clinical Psychology in Europe, History of; Psychoanalysis, History of; Psychotherapy, History of: Psychiatric Aspects; Training in Clinical Psychology in the United States: Practitioner Model; Training in Clinical Psychology in the United States: Scientist–Practitioner Model.

Bibliography

Binet, A., Simon, T., 1905. A new method for the diagnosis of intellectual level of abnormal persons. Annee Psychologique 11, 191–244.

Conners, C.K., Eisenberg, L., 1963. The effects of methylphenidate on symptomatology and learning in disturbed children. American Journal of Psychiatry 120, 458–464.
Eysenck, H.J., 1952. The effects of psychotherapy: an evaluation. Journal of Consulting Psychology 16, 319–324.
Healy, W., Bronner, A.F., 1926. Delinquents and Criminals: Their Making and Unmaking. Macmillan, New York.
Holzman, P.S., Proctor, L.R., Hughes, D.W., 1973. Eye tracking patterns in schizophrenia. Science 181, 179–181.
Jones, M.C., 1924. The elimination of children's fears. Journal of Experimental Psychology 7, 382–390.
Kraepelin, E., 1987. Memoirs. Springer, Berlin.
Lang, P.J., Lazovik, A.D., 1963. Experimental desensitization of a phobia. Journal of Abnormal and Social Psychology 66, 519–525.
Lazarus, R.S., 1966. Psychological Stress and the Coping Process. McGraw-Hill, New York.
Meehl, P.E., 1962. Schizotaxia, schizotypy, and schizophrenia. American Psychologist 17, 827–838.
Raimy, V.C. (Ed.), 1950. Training in Clinical Psychology. Prentice-Hall, New York.
Rapaport, D., Gill, M.M., Shafer, R., 1945. Diagnostic Psychological Testing, vol. 1. Yearbook, Chicago.
Rapaport, D., Gill, M.M., Shafer, R., 1946. Diagnostic Psychological Testing, vol. 2. Yearbook, Chicago.
Routh, D.K., 1994. Clinical Psychology Since 1917: Science Practice and Organization. Plenum, New York.
Routh, D.K., 1996. Lightner Witmer and their first 100 years of clinical psychology. American Psychologist 51, 244–247.
Routh, D.K., 1998. Hippocrates meets Democritus: a history of psychiatry and clinical psychology. In: Bellack, A.S., Hersen, M. (Eds.), Comprehensive Clinical Psychology, vol. 1. Pergamon, New York, pp. 1–48.
Witmer, L., 1907. Clinical psychology. Psychological Clinic 1, 1–9.

Clinical Psychology, Psychiatry and Homosexuality

Frank Muscarella, Barry University, Miami Shores, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The modern concept of homosexuality as a sexual orientation emerged with the field of descriptive psychiatry in Europe in the late 1800s. It was formally classified as a mental disorder in 1886 with its inclusion in *Psychopathia Sexualis*. Pressure from new research on homosexuality, public advocacy, and criticism of the earlier versions of the psychiatric diagnostic system resulted in its declassification in 1973. Major areas of current research in homosexuality include sexual orientation change efforts, the prevalence of depression and anxiety in homosexual men and women secondary to social stigma and prejudice, and theories on the etiology of homosexuality.

Introduction

The term homosexuality refers to a specific sexual orientation and is the integrated sexual, emotional, and psychological attraction to individuals of the same sex (Ellis and Mitchell, 2000). Today, homosexuality per se is not categorized as a mental disorder in the major psychiatric diagnostic classification systems that are used for the majority of the world's population: the American Psychiatric Association's (APA) *Diagnostic and Statistical Manual* (5th ed.) (*DSM-5*) (APA, 2013); the 2010 version of the World Health Organization's (WHO) *International Classification of Diseases* (ICD-10) (WHO, 2010); and the Chinese Classification of Mental Disorders (3rd ed.) (CCMD-3) (Lee, 2001; Wu, 2003). Rather, in Western culture, homosexuality is considered a normal variation in human sexuality (Drescher, 2010; Lee, 2001). However, homosexuality has been categorized as a mental disorder in the recent past, and presently across the world it is viewed with varying levels of intolerance that too often leads to harassment and persecution (Hadler, 2012; Reading and Rubin, 2011). Despite the fact that homosexuality has been declassified as a mental disorder, current psychotherapeutic approaches designed to change homosexual orientation to heterosexual orientation are based on the assumption that homosexual orientation is pathological (Drescher, 1998; Flentje et al., 2013; Gonsiorek, 2004; Grace, 2008).

Disagreement about the conceptualization of homosexuality as inherently pathological has existed since homosexuality first came under the scrutiny of the emerging field of descriptive psychiatry in Europe in the mid-1800s (De Block and Adriaens, 2013). Critics of the psychiatric diagnostic system argue that it too often serves as a culturally approved method to pathologize, control, and sometimes punish (cf Silverstein, 2008) socially unacceptable behavior that is more accurately described as nonconforming rather than pathological (Conrad and Angell, 2004). As such, the determination of pathology has been interpreted as partly a political process which, throughout the history of the development of the European and American psychiatric diagnostic frameworks, has resulted in socially unacceptable sexual behavior being consistently labeled as pathological and targeted for treatment and cure (De Block and Adriaens, 2013; Silverstein, 2008). There is no better example of this pathologizing and political process than the history of homosexuality in European and American psychiatry, the outcomes of which were ultimately exported to the world through readily adopted diagnostic classification systems (e.g., Lee, 2001; Mendelson, 2003).

Early History

Homosexual expression has been documented in cultures across the world and across recorded history; and cultural beliefs about it and responses to it have ranged dramatically from negative to positive (Crompton, 2003; Greenberg, 1988). Specifically, the conceptualization of homosexuality in Western culture has changed radically across historical periods. Ancient Greek and Roman cultures are considered the foundation of Western civilization. Homosexual behavior was common among and appreciated by the Ancient Greeks and Romans, and even their Gods engaged in homosexual behavior (Boswell, 1980, 1994; Cantarella, 1992; Dover, 1978). The consensus among historians is that the Ancient Greeks and Romans did not define people in terms of their sexual orientation, rather they described people in terms of the roles that they played in sexual interactions and the preferences they had for sexual partners. Social disapproval resulted not from the sex of the partner but from violations in sexual role (Cantarella, 1992; Dover, 1978). The other ancient European lands were made up of numerous cultures and there is very little documented evidence of their attitudes about homosexual expression. It appears from the few, mainly ancient Roman reports (e.g., Boswell, 1994; Tacitus, 2009), that attitudes in those cultures ranged from negative to positive, depending on the culture and the circumstances.

According to Boswell (1980, 1994), the advent of Christianity in Europe initiated a homogenization of societal attitudes, and consequently rules, toward sex. According to natural law, the only sex that was acceptable was between a husband and wife for procreation. Nonprocreative sexual acts were considered particularly sinful, but the Church in practice was relatively tolerant of unacceptable or 'deviant' sexual acts until around the AD 1200s. At that time, due to social change and perceived threats, the Church embraced orthodoxy with a concomitant intolerance and discouragement of nonprocreative sexual behavior, and in particular, homosexual behavior.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.21026-X

Crompton (2003) disagrees with Boswell and argues that from its inception, Christianity, for reasons that remain unclear, was extremely hostile toward homosexuality. The hostility worsened in terms of condemnation, persecution, and punishment with Christianity's development and spread across Europe and the world. Crompton states that Christianity emphasized and magnified the association of the biblical story of the destruction of Sodom with homosexuality. According to Crompton, Christianity's strongly negative conceptualization of homosexuality and the persecution that followed this, arose from the close association of homosexuality with the belief in God's anger against it and resulting disasters. Finally, he argues that this intense hostility manifested itself both in ecclesiastical laws against homosexual conduct and cruel punishments for it. As the ancient European lands developed into the modern European countries, ecclesiastical law and its penalties for homosexual acts were incorporated into civil law with the same strength of negative attitudes and responses and ultimately exported to the United States and across much of the world (e.g., Fradella, 2002).

The Mid-1800s

The definition of sexual deviance, including homosexual conduct, remained in the realm of moral, legal, and theological considerations until the mid-1800s with the advent of descriptive psychiatry (De Block and Adriaens, 2013). In the mid-1800s, across much of Europe, there was very severe criminalization of homosexual conduct, and some psychiatrists and early sexologists saw this as unjust and attempted social action (Conrad and Angell, 2004; De Block and Adriaens, 2013; Drescher, 2010). Critics from both fields found common ground and argued that people who engaged in homosexual conduct did so due to an innate, congenital condition and therefore punishment was not the appropriate social response.

The innate characteristic argument set the stage for the development of the concept of homosexuality as a condition that people could be identified as having. However, there were two different lines of reasoning about the meaning of the innate quality of homosexuality. Sexologists, including some physicians and psychiatrists, argued that not only was homosexuality a congenital condition but as such, it was part of the natural variation in sexuality and not inherently pathological or a social problem. Therefore, they argued that it should be decriminalized (Conrad and Angell, 2004; De Block and Adriaens, 2013; Drescher, 2010; Scasta, 1998). In fact, the modern term 'homosexuality' was coined in 1869 by a Hungarian journalist and social activist, K.M. Kertbeny (born Benkert) as part of his treatise against Paragraph 143, a Prussian law criminalizing homosexual conduct (Conrad and Angell, 2004; Drescher, 2010). Psychiatry, on the other hand, argued that homosexuality was congenital but pathological. Thus, homosexuality became a topic of interest in psychiatry, with the goal of bringing it into the medical realm for the 'humane' purpose of treatment and cure (Conrad and Angell, 2004; De Block and Adriaens, 2013; Drescher, 2010; Scasta, 1998). The psychiatric conceptualization soon dominated, introduced homosexuality into Western psychiatric nosology, and maintained its dominance through most of the twentieth century (Mendelson, 2003).

The Late 1800s

De Block and Adriaens (2013) report that in the mid-1800s both the general public and civil authority quickly turned to psychiatry for guidance in the determination of normal and deviant sexuality. Psychiatry's influence increased significantly by the late 1800s. One of the most influential books in psychiatry related to sexual deviance was first published in 1886 by the forensic psychiatrist Richard Von Krafft-Ebing: *Psychopathia Sexualis*. It contained three major groupings of sexual perversions, one of which included homosexuality. Although Krafft-Ebing used the terminology coined by Kertbeny, he did not adopt his conceptualization of it as normal variation. In fact, homosexuality was described as a degenerative disorder (Drescher, 2010). This initiated the formal psychiatric conceptualization and categorization of homosexuality as a mental disorder, and it remained as such until 1973 when the American Psychiatric Association declassified homosexuality as a mental disorder.

The development of the term 'homosexuality' and psychiatry's interest in it set the stage for a major cultural shift in focus from an act that was considered criminal to a sexual orientation and state of mind considered to be pathological. Boswell states that most historians agree that "…no Western legal or moral tradition – civil or ecclesiastical, European, English, or Anglo-American – has ever attempted to penalize or stigmatize a 'homosexual person' apart from the commission of external acts" (Boswell, 1993: p. 40 cited in Fradella, 2002: p. 280). However, the conceptualization of homosexuality as a mental disorder did contribute to the stigmatization of homosexual people (Silverstein, 2008). The understanding of and response to homosexuality moved across history from the theological to the legal to the psychiatric, but the strongly negative cultural attitudes about homosexuality remained relatively unchanged. This is best illustrated by the fact that although there was a change in the conceptualization of homosexual acts, and ultimately homosexuality, from sin to crime to mental illness, society's responses to homosexual behavior and homosexuality remained punitive and socially and personally destructive for those categorized as homosexual (Berube, 1990; Conrad and Angell, 2004; Crompton, 2003; Scasta, 1998; Silverstein, 2008).

The Early 1900s

Although *Psychopathia Sexualis* and the dominance of psychiatry formalized homosexuality as a mental disorder in the late 1800s, sexologists and social activists continued to press their point that homosexuality was not inherently pathological or a social problem, to smaller audiences. This disagreement about its conceptualization is best illustrated by the opinion of Sigmund Freud himself whose thinking was influenced by the work of sexologists (De Block and Adriaens, 2013). In 1903 the newspaper *Die Zeit* published an interview with Freud in which he stated "I am…of the firm conviction that homosexuals must not be treated as sick people…Homosexual persons are not sick. They also do not belong in a court of law!" (cited in Isay, 1989: p. 3).

De Block and Adriaens (2013) indicate that in the twentieth century in Europe and the United States two voices continued to

speak about homosexuality: psychiatry and sexology. Psychiatry maintained the pathology model. Sexologists, who included anthropologists, biologists, historians, sociologists, and physicians, became stronger and more well-defined in their approach. Sexologists brought to bear cross-species, cross-cultural, and historical research, as well as observations of essentially normal populations. The sexologists argued more forcefully, with new evidence in hand, that deviations from the sexual norm, such as homosexuality, were neither pathological nor dangerous to society, and they promoted social change in attitudes toward homosexuality. Again, their influence is expressed by Freud. In the famous letter to the mother of a homosexual man written in 1935, Freud stated: "Homosexuality is assuredly no advantage, but it is nothing to be ashamed of, no vice, no degradation; it cannot be classified as an illness; we consider it to be a variation of the sexual function, produced by a certain arrest of sexual development" (Freud, 1935/1960a: p. 423, cited in De Block and Adriaens, 2013: p. 283).

The Mid-1900s

American Psychoanalytic Influence

The prevailing voice regarding homosexuality remained that of mainstream psychiatry that held to the pathology perspective. This group was largely psychoanalytic and even though Freud himself did not see homosexuality as inherently pathological, after Freud's death in 1938, his followers became increasingly strident in their opinion that it was (Drescher, 1998). They based their study of homosexuality on the patients they saw in their private and institutional clinical practices (De Block and Adriaens, 2013; Silverstein, 2008). By contrast, one of the reasons that dissenting professionals disagreed with the classification of homosexuality as a mental disorder was that they saw people who were homosexual who lived essentially normal, contented lives. They recognized that although there could be psychopathology in homosexuals, a state of homosexuality was not inherently pathological itself. Freud himself had urged mental health professionals to take a wide perspective on homosexuals, specifically those who did not seek treatment: "If we disregard the patients we come across in our medical practice, and cast our eyes round a wider horizon, we shall come in two directions upon facts which make it impossible to regard inversion as a degeneracy:…" (Freud, 1962: p. 5).

However, other social forces appear to have contributed to some psychoanalysts' commitment to the pathological view of homosexuality. Isay (1989) reports that from its inception as a radical movement, psychoanalysis tended to be conservative, but on the topic of homosexuality it was the most conservative of all the mental-health professions. Isay states that although this is not a uniquely American phenomenon, this attitude is the most fixed in the American psychoanalytic tradition for two reasons.

The first reason is that during World War II (WW II), many analysts held important positions in the US military (cf Berube, 1990). Upon their return after the war they assumed positions of leadership in departments of psychiatry. They were very motivated to link psychoanalysis to medicine and to the disease model, to increase the prestige and influence of psychoanalysis and economic benefit to its members. The second reason is that shortly before the outbreak of WW II many European analysts fled to the United States. They were anxious about their professional and economic situation and quickly gravitated toward the movement to link psychoanalysis with medicine and the disease model to give stability and prestige to the profession. The McCarthy era followed WW II in the early 1950s in the United States. The goal of McCarthy and his adherents was to identify and remove from public life and influence communists, homosexuals, and other nonconforming 'enemies' of the United States. Many psychoanalysts, themselves recent arrivals to the United States, felt vulnerable and fearful and again gravitated to established and conforming professional organizations and doctrine (Isay, 1989). Some critics have also suggested that since psychoanalysts argued that the only effective cure for homosexuality was long-term psychoanalysis, the issue of secondary gain in their adherence to the disease model of homosexuality is open to question (e.g., Silverstein, 2008).

The DSM

The American Psychiatric Association, dominated by psychoanalytic psychiatrists, standardized the psychiatric classification system begun by the Army in 1945 (Berube, 1990), with the creation and publication of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*). The predecessor of the first edition of the *DSM*, the *Statistical Manual*, and the first edition of the *DSM* (1952) were relatively ambiguous about classifying sexual deviations as mental disorders (De Block and Adriaens, 2013). Although the first edition in 1952 said relatively little about homosexuality, it was classified as a 'sociopathic personality disturbance' that was part of a general category of personality disorders (De Block and Adriaens, 2013; Drescher, 2010). Reportedly, it was possible to diagnose even a contented homosexual as a sociopath because sociopathy was interpreted as a type of personality disorder characterized by a lack of distress or anxiety associated with the observed pathological condition (O'Donohue and Casselles, 2005 cited in Grace, 2008).

The DSM-II

De Block and Adriaens (2013) state that the *DSM-II* was an adapted version of the mental health disorder section of the World Health Organization's eighth *International Classification of Diseases* (*ICD-8*) which itself was an attempt to create a common ground for psychiatrists of different nationalities, institutional backgrounds, and theoretical orientations. In the first six printings (May 1968–October 1973) of the *DSM-II*, homosexuality was categorized unambiguously as a mental disorder (De Block and Adriaens, 2013); specifically, it was reclassified as a sexual deviation (Drescher, 2010). De Block and Adriaens (2013) report that this was the logical conclusion of a predominantly psychoanalytic body creating a classification system for mental disorders based on theory and etiology. That is, since the psychoanalytic theory held by this body stipulated that heterosexuality was the norm and that developmental trauma caused deviation from the norm, homosexuality could only be pathological (in contrast to Freud's thinking about the issue). However, it is argued here that the historically long negative religious and social

attitudes about homosexuality (Crompton, 2003) and the historically recent conditions in American psychoanalysis (Isay, 1989; Silverstein, 2008) must have had a combined effect that contributed to the resistance to a paradigm shift in the psychoanalytic profession. In fact, Drescher (1998) reports that post-Freud, psychoanalytic theories couched moral condemnation of homosexuality in scientific and pseudoscientific metaphors.

The Late 1900s

In the early 1970s, in the United States, there were a number of social forces that were developing, converging, and challenging the conceptualization of homosexuality as pathological. The voices of the sexologists and social activists, present since the beginning in the long discussion of homosexuality, became louder and stronger and appear to have made significant inroads into psychiatry. According to De Block and Adriaens (2013), shortly after the publication of the *DSM-II* in 1968, disagreement about the pathological conceptualization of homosexuality as well as other conditions, became more compelling in psychiatry. One factor influencing this disagreement was a new generation of scientifically trained psychiatrists who, seeing the benefits of psychotropic medications, began to question traditional psychoanalytic conceptualizations and wanted diagnostic categories that were data driven. The first two editions of the *DSM* had not included an explicit definition of mental disorder but, reportedly, relied on an implicit definition that included two criteria: distress and disability or functional impairment. However, most homosexuals exhibited neither, and if they did, an increasing number of mental health professionals understood that their problems resulted not from the state of being homosexual but rather from society's response to it. This realization logically led to the need to clarify the difference between socially unacceptable or deviant behavior and truly abnormal or pathological behavior (De Block and Adriaens, 2013).

Results from new research began to challenge the conceptualization of homosexuality as pathological. Kinsey's research published in 1948 and 1953 demonstrated that homosexual behavior was not the same as homosexuality since a large percentage of the population engaged in the behavior, and a larger percentage than believed had predominantly homosexual orientations. Chiang (2008) reports that Kinsey's work had a significant influence on psychiatrists' attitudes about homosexuality. For example, despite the antihomosexual fervor of the 1950s McCarthy era, the Group for the Advancement of Psychiatry, working with governmental agencies on a report about homosexuality, was relatively supportive of homosexuality. According to Chiang (2008), the report began by citing Kinsey's statistics of homosexual behavior in America and concluded that homosexuals functioned very well in civilian life and government settings, without problems; and suggested that investigations should be carried out on a case by case basis taking into account a variety of circumstances. Evelyn Hooker's (1957) ground-breaking study comparing the assessment results of heterosexual and homosexual men initiated an ongoing body of literature showing that there is no difference in psychological adjustment and functioning based on sexual orientation itself (Morin and Rothblum, 1991).

Another factor that contributed to a challenge of the status quo was a direct outcome of WW II (Berube, 1990). In the United States, many homosexual men and women served in the Armed Forces in WW II, and they served valiantly and effectively, often receiving military honors. They were pathologized by the mainstream of the military psychiatrists (mostly psychoanalytic), and some were persecuted and denied veterans' benefits after discharge. Individuals began to fight this discrimination themselves and with advocates in the veterans hospitals, military, and among local political representatives. After WW II there was an increase in the number of clinical psychologists due to the need to treat veterans. Chiang (2008) reports that clinical psychologists were less accepting than psychiatrists of the pathological view of homosexuality. After the war, many homosexual veterans, men and women, also chose to relocate to large urban areas to live their lives more freely rather than to return to small towns and rural areas and conform to local norms. All of this contributed to a change in their images of themselves from being pathological and victims, to being different but normal and political. This change set the stage for the postwar gay rights movement in the United States, which became a visible political force with the Stonewall riots in 1969. The movement was fueled by a younger generation, a cultural zeitgeist of liberation and antiauthority attitudes, and new research on homosexuality. This social movement confronted the establishment in psychiatry and psychology directly and forcefully, with new research and the devastating social and personal impact of the disease model (De Block and Adriaens, 2013; Drescher, 2010; Silverstein, 2008). The movement was particularly hostile and confrontational toward the psychoanalytic establishment because of its continued promulgation of the disease model and because of lingering resentment of its treatment of homosexual men and women in the Armed Services in WW II (Berube, 1990; Silverstein, 2008).

Declassification

In December 1973, after a long, raucous, and contentious battle with gay rights activists and their allies outside and within the discipline, the Board of Trustees of the APA unanimously accepted a proposal to remove homosexuality from the *DSM-II* (De Block and Adriaens, 2013; Drescher, 2010; Silverstein, 2008). Later, in response to protests from a number of leading psychoanalysts, a referendum was held with 58% of the membership of the APA accepting the proposal and 37% voting against (Silverstein, 2008). Homosexuality was eliminated as a mental disorder in the seventh printing (July 1974) of the *DSM-II* (De Block and Adriaens, 2013). The American Psychological Association and the National Association of Social Workers publicly supported the resolution of the American Psychiatric Association shortly after it was passed (Silverstein, 2008).

Critics of the declassification of homosexuality as a mental disorder argued that the decision was largely a political one (De Block and Adriaens, 2013). It is undeniable that cultural, social, and political forces converged and contributed to the process to depathologize homosexuality, just as they did in the process to pathologize it. Within the socio-political context of Europe in 1886, homosexuality was essentially proclaimed a mental disorder by Krafft-Ebing in *Psychopathia Sexualis*. Homosexuality

was declassified as a mental disorder by popular vote within the socio-political context inhabited by the American Psychiatric Association in 1973. Neither process employed rigorous empirically based standards, and both processes were arguably more socio-political than scientific (cf Grace, 2008). The entire pathologizing–depathologizing issue is consistent with Bayer's (1981) interpretation that "The status of homosexuality is a political question, representing a historically rooted, socially determined choice regarding the ends of human sexuality" (cited in Drescher, 1998: p. 5).

Both processes supported the contention of critics that the psychiatric diagnostic system too often served as a vehicle to pathologize socially unacceptable behavior that was arguably not pathological (Conrad and Angell, 2004; Scasta, 1998; Silverstein, 2008). However, this scientifically questionable pathologizing-depathologizing of homosexuality led to three positive outcomes all present in the subsequent version of the *DSM*, the *DSM-III*: the development of a more data-driven diagnostic classification system; a clear distinction between a mental disorder and socially deviant behavior; and, most importantly, an explicit operational definition of the term 'mental disorder' was derived, which noted that subjective distress and functional impairment were essential features of this condition (De Block and Adriaens, 2013). Further, De Block and Adriaens argue that the reconceptualization of homosexuality was an essential component for the development of the definition of mental disorder, and this has reverberated across proposed diagnostic categories through all subsequent versions of the *DSM*.

Although homosexuality per se was deleted from the *DSM-II*, another, related disorder was added under the diagnostic category 'Sexual Deviations'. A reading of the eleventh printing (January 1978) of the *DSM-II* (APA, 1968) describes "Sexual orientation disturbance [Homosexuality]" as follows: "This is for individuals whose sexual interests are directed primarily toward people of the same sex and who are either disturbed by, in conflict with, or wish to change their sexual orientation. This diagnostic category is distinguished from homosexuality, which by itself does not constitute a psychiatric disorder" (p. 44).

Subsequent Editions of the DSM

The *DSM-III* (APA, 1980) included an explicit conceptualization of a mental disorder. Parenthetically, the relationship between a mental disorder and socially deviant behavior was clarified: ("When the disturbance is *limited* to a conflict between an individual and society, this may represent social deviance, which may or may not be commendable, but is not by itself a mental disorder") (p. 6). This clarification has run through the later editions of the *DSM*. The DSM-5 (APA, 2013) is more explicit: "Socially deviant behavior (e.g., political, religious, or sexual) and conflicts that are primarily between the individual and society are not mental disorders unless the deviance or conflict results from a dysfunction in the individual, as described above" (p. 20).

The *DSM-III* did not include homosexuality as a mental disorder, but it did include a reformatting of the *DSM-II* sexual orientation disturbance with the new diagnostic category of 'Ego-dystonic Homosexuality' listed in a section titled 'Other Psychosexual Disorders'. This comprised people who had little to no heterosexual interest and experienced homosexual arousal that was unwanted and distressing. This diagnostic category appeared to ignore the fact that distress about one's homosexual orientation was the result of social stigmatization, and this caused distress, not homosexuality itself. This diagnostic category was greeted with great disapproval, rarely used, and eliminated in the revision, *DSM-III-R* (APA, 1987) (De Block and Adriaens, 2013; Drescher, 2010; Silverstein, 2008). However, the association between homosexuality and a mental disorder continued to exist, albeit in a very subtle way. The *DSM-III-R* included the diagnostic category 'Sexual Disorder Not Otherwise Specified' listed in a section titled 'Other Sexual Disorders'. This could cover a number of conditions but included the explicit example of persistent distress over sexual orientation. The number of heterosexuals who complained of this appears to have been zero. This category remained in the *DSM-IV* (APA, 1994) and in the *DSM-IV-TM* (APA, 2000) in a section titled 'Sexual and Gender Identity Disorders'. The *DSM-5* (APA, 2013) does not contain the category and distress about sexual orientation is not described. The slow exit of the association of homosexuality with mental disorder appears to have been due to the lingering effects of psychoanalytic interpretations (Conrad and Angell, 2004; De Block and Adriaens, 2013; Drescher, 2010; Mendelson, 2003; Silverstein, 2008).

Under the *DSM-5*, should debilitating distress about sexual orientation occur in an individual, it could be listed under the diagnostic category of 'Unspecified Mental Disorder' which reads as follows: "This category applies to presentations in which symptoms characteristic of a mental disorder that cause clinically significant distress or impairment in social, occupational, or other important areas of functioning predominate but do not meet the full criteria for any mental disorder" (p. 708). In theory, this diagnostic category would allow a clinician to acknowledge a clinically relevant degree of distress and impairment due to one's sexual orientation without any implication that it is the orientation itself that is responsible for the distress. Thus, it appears that the *DSM-5* has completely severed the association between homosexuality and mental disorder, 127 years after its first appearance in *Psychopathia Sexualis*.

Other Psychiatric Classification Systems

The World Health Organization's International Classification of Diseases *ICD-10*, published in 1992, removed the diagnostic category 'homosexuality' (Mendelson, 2003), but added some other categories. The 2010 version of the *ICD-10* (WHO, 2010) has a general category (F66) titled "Psychological and Behavioural Disorders Associataed with Sexual Development and Orientation". It is specified that "Sexual orientation by itself is not to be regarded as a disorder". It then lists the following diagnoses: sexual maturation disorder, egodystonic sexual orientation, sexual relationship disorder, other psychosexual development disorders and psychosexual development disorder, unspecified. 'Sexual maturation disorder' is a term that can be applied to those who experience anxiety or depression because of their gender identity or sexual orientation, and is most applicable to adolescents. Egodystonic sexual orientation is the term for those who are distressed by their gender identity or sexual orientation and want to change it. Sexual relationship disorder is the term for those whose gender identity or sexual

orientation is responsible for difficulties in developing or maintaining a relationship with a sexual partner.

It appears that classification in the *DSM-5* is more evolved than that in the *ICD-10* in several ways: the *DSM* separates sexual orientation from gender identity, and it has done away with explicit diagnostic categorization of distress related to sexual orientation. As has been reviewed, essentially distress occurs only in cases of a homosexual orientation, and it is an accepted fact that this problem results from social stigma, not from the orientation itself. It has been argued that the presentation of sexual orientation in the *ICD-10* should be revised (Mendelson, 2003).

The Chinese Classification of Mental Disorders (3rd ed.) (*CCMD-3*) is used in China, which contains 20% of the world's population. The *CCMD-3*, published in 2001, has been heavily influenced by the *DSM* and the *ICD*, but incorporates elements unique to Chinese culture (Lee, 2001). The third edition dropped the diagnostic category of 'homosexuality' but added 'ego-dystonic homosexuality' (Wu, 2003). More specifically, Lee (2001) reports that despite an understanding that homosexuality is considered a normal variant in the Western world, exclusive homosexuality is considered a mental disorder, but only if it causes distress and a desire to change it. Lee states that it is not clear why heterosexuals who are distressed with their sexual orientation are not considered mentally ill. His interpretation is that the Chinese psychiatric establishment is going through a step-by-step paradigm shift similar to that experienced by the American psychiatric establishment.

Contemporary Issues in Research

The subject of homosexuality generates a considerable amount of research from a variety of perspectives. Three major areas are considered here. These are the outcomes of sexual orientation change efforts, the prevalence and causes of mental illness in homosexual men and women, and theories on the etiology of homosexuality.

Sexual Orientation Change Efforts

Therapies known as conversion therapy, reparative therapy, and sexual reorientation are collectively referred to as sexual orientation change efforts (SOCE). According to Drescher (1998), despite the declassification of homosexuality by the APA in 1973, a small group of conservative psychoanalysts maintained that homosexuality was indeed a pathological condition and a homosexual orientation could be changed to heterosexual orientation. SOCE were linked primarily to psychodynamic and behavioral theories in the 1960s and 1970s (Flentje et al., 2013). SOCE today do not appear to have close association with any psychological theory. However, they do appear to be the direct descendants of earlier psychoanalytic change efforts that have been adopted and modified by some fundamentalist Christian groups (Drescher, 1998). Currently, the change therapies use a group of verbal interventions that essentially discourage homosexuality and encourage heterosexuality (Flentje et al., 2013). Most of the individuals who seek out such treatment are those who are having problems reconciling their fundamentalist Christian faith with their homosexual orientation (Drescher, 1998; Flentje et al., 2013; Gonsiorek, 2004; Grace, 2008). Some of these individuals are minors who have been forced into therapy by their parents.

SOCE have come under scrutiny and criticism. It has been argued that behavioral science and the psychological techniques derived from it are not good to help people reconcile issues of religious doctrine, because of the inherent conflict between faith and science regarding the source of truth (Gonsiorek, 2004; Grace, 2008). Major mental health organizations in the United States including the American Psychiatric Association, the American Psychological Association, and the National Association of Social Workers have all made policy statements rejecting SOCE (Flentje et al., 2013; Grace, 2008). Half a century of research on SOCE has not shown that they work. The American Psychological Association has stated that there is no demonstrated evidence that SOCE are effective and there is some evidence that they may cause harm (Hancock et al., 2012). Relying partly on this argument, both the states of California and New Jersey passed laws in 2013 banning the use of SOCE for individuals under the age of 18.

Homosexuality and Mental Illness

Although homosexuality is not a mental disorder, the evidence suggests that homosexual adolescents (Mustanski and Liu, 2013) and men and women (Cochran et al., 2003) may be more vulnerable than their heterosexual counterparts to some forms of mental illness such as anxiety and depression. An enormous body of psychological research, beginning with Evelyn Hooker (1957) and up to the present, has consistently shown that there are no psychological differences between people based on sexual orientation alone; and when differences do occur they are the result of society's response to homosexuality, not of homosexuality itself. Research indicates that differences in psychological functioning that do exist between homosexuals and heterosexuals result from the stress associated with the societal pressures on homosexuals as a marginalized group (Grace, 2008; Meyer, 2013).

Homosexual men and women across the world live under varying levels of social marginalization and intolerance that lead to varying levels of aggression against them (Hadler, 2012; Reading and Rubin, 2011; Meyer, 2013; Sue, 2010). Western countries have the lowest levels of gross aggression toward homosexual men and women and greater tolerance (Hadler, 2012; Reading and Rubin, 2011). However, even in Western countries, homosexual men and women, like individuals of many socially marginalized groups, are more likely to suffer from an accumulation of microaggressions in daily life (Sue, 2010). Microaggressions are verbal, behavioral, and environmental insults that communicate hostility, dislike, and disdain and are detrimental to psychological and physical well-being. Meyer (2013) conducted meta-analyses and found that homosexual men and women have a higher prevalence of mental disorders than do heterosexuals. He attributes this to what is termed 'minority stress' or the stress of living as a minority group suffering from stigma and social prejudice. Stigma and prejudice generate micoaggressions (Sue, 2010). Meyer recommends that interventions at both the personal level (e.g., improving individual coping methods) and the structural level (e.g., changing oppressive social and legal

policies) may be useful in alleviating minority stress. Consistent with Meyer's observation, research has shown that socializing agents such as laws affect the individual's attitudes toward homosexuality with more legal protection of homosexuality contributing to greater acceptance (van den Akker et al., 2013).

Theories on the Etiology of Homosexuality

At this time, it is not scientifically known how a particular sexual orientation develops nor can deviation from heterosexual orientation development be reliably predicted. The consensus among most of the scientific community is that sexual orientation develops in an individual as the result of an interaction between genetic, biological, developmental, psychological, cultural, and unique experiential factors. However, there are a number of theories on the development of homosexual orientation. These share the implicit assumption that sexual orientation development would be expected to be heterosexual or fall on a continuum from heterosexual to bisexual because of the direct relationship of these orientations to reproduction. Because exclusive homosexuality is not directly associated with reproduction, its prevalence has generated interest in its possible origins. Comprehensive reviews and critiques of the theories on the etiology of homosexuality are presented by Rosario and Schrimshaw (2014) and LeVay (2011). Muscarella (2006) presents a review of evolutionary theories. A summary of all of these theories follows.

The theories on the etiology of homosexuality have emerged historically with the predominant scientific paradigms that developed across time and were applied to human sexuality. The major contemporary theories of the etiology of homosexuality begin with Freud in his psychoanalytic theory. He theorized that homosexuality in males was one possible product of an unresolved oedipal complex, in which a boy failed to identify with his father. In the little he wrote about female homosexuality, its origins appear to be due, not to an unresolved Electra complex, but rather to a later turning away from the father because of anger and disappointment. The behavioral theories attribute the development of homosexual orientation to learning and the result of classical and operant conditioning. The advent of research on prenatal hormonal influences on the development of brain organization and sexual behavior in nonhuman animals led to a number of neurohormonal theories. These hold that various factors may disrupt the normal level of prenatal hormones during gestation, resulting in the organization of the brain for homosexual orientation. The "exotic becomes erotic" model by Bem includes both neurohormonal and social factors. Bem theorizes that males and females exhibit sex-typed behavior that makes them different from each other, or exotic; and across development, the exotic becomes sexually attractive or erotic. Prenatal hormonal variation causes some children to develop some gender-nonconforming behavior. Thus, the behavior of their own gender is exotic, and ultimately, they become attracted to members of their own gender.

The identification of genetically linked characteristics generated theories on the possibility of an inherited gene, or set of genes, that resulted in a homosexual orientation.

Since, in theory, these genes could not be passed on directly, other theories were developed to explain a possible process. These theories can be categorized as evolutionary theories because they try to explain how genes associated with a homosexual orientation would have been indirectly reproductively beneficial and thus selected for during human evolution (Muscarella, 2006). The first evolutionary theory was presented by E.O. Wilson who stated that among human ancestors, homosexual men and women may have held particularly high social positions such as the group shaman. This would have raised the status of their families who may have then reproduced more successfully as a result.

Edward Miller theorized that homosexuality is a polygenic trait that results from selection for a number of characteristics that each contribute to fitness. These include tenderness, empathy, and kindness which result from a more feminized brain, which would have made ancestral males more attractive to females as mates and better fathers. In some cases the brain organization is extremely feminized giving rise to a homosexual orientation. Miller uses the maternal immune hypothesis to explain the purported fraternal birth order effect, which holds that the probability of a male developing a homosexual orientation increases with the number of older brothers. The maternal immune hypothesis holds that mothers carry a 'biological memory' (in the form of an H–Y antigen) of the number of sons they have gestated that leads to changes in the intrauterine environment – and results in the feminization of later born males. According to Miller, younger sons, being more sensitive and flexible, would compete less with older brothers for resources. Homosexuality in males would be the rare by-product of genetic variability for the capacity to flexibly exploit various social niches, outweighing the cost of a few individuals who do not reproduce. Another line of research holds that maternally inherited factors are associated with both male homosexuality and increased fecundity in female relatives.

The alliance theory holds that during the course of human evolution, homosexual behavior may have reinforced alliances between same-sex conspecifics, which increased their chances of survival and reproduction (Muscarella, 2006). Thus, there was selection for homosexual behavior in all humans. Genetic variability underlying this behavior may result in individuals with a preference for exclusive homosexual behavior and ultimately the development of a homosexual orientation.

Simon LeVay (2011) looks at biological and psychological data and suggests a compelling neurohormonal theory for the development of homosexual orientation. He argues that homosexuality is a 'package' of mental traits, many of which are gender-atypical, while heterosexuality is a package of mental traits that are gender-typical. He states the association of gendered traits with sexual orientation arises from sexual differentiation of the brain under the influence of sex hormones. LeVay speculates that a number of social and biological factors across human history may have influenced prenatal hormonal conditions and may have influenced the prevalence of homosexuality over time. His theory complements evolutionary theories and could be interpreted as a proximate mechanism for variation in sexual orientation.

Summary and Conclusions

Homosexuality appears to have existed cross-culturally over recorded history. Societies' reactions to homosexuality have

ranged from negative to positive as functions of the historical and cultural traditions of each society. Christianity introduced a very negative view of homosexuality that remained relatively constant across history. With cultural evolution, the conceptualization of homosexuality changed from sin to crime to mental illness. The inclusion of homosexuality in *Psychopathia Sexualis* formally introduced homosexuality to the Western world as a mental disorder, and this conceptualization was then exported across the world and across generations in psychiatric diagnostic classification systems. Disagreement about the inherent pathology and social danger of homosexuality existed from the beginning of the conceptualization of homosexuality. However, it was not until the late twentieth century that research supporting this view was abundant and that Western society was culturally prepared to accept it and act upon it. Although the American Psychiatric Association declassified homosexuality as a mental disorder in the DSM-II in 1973, it is argued here, that it was not until 2013 with the publication of the DSM-5 that the final vestiges of the link between homosexuality and psychopathology finally vanished.

Despite the declassification of homosexuality as a mental disorder in the major psychiatric diagnostic classification systems of the world, some societies and subcultures continue to view it as a mental disorder and support 'treatments' to change it. This is clearly outside of mainstream psychiatry and psychology and more harmful than helpful. Research has also demonstrated that homosexual adolescents, men, and women tend to exhibit more mental disorders such as depression and anxiety as a result of the major and minor psychological stresses associated with stigma, prejudice, harassment, and in some cases persecution. Research shows that psychological interventions at the personal level can help individuals to cope with this minority stress and, consequently, to function more effectively. However, research also shows that interventions at the societal level aimed at changing socializing agents such as law and government policy can lead to more tolerant societies that generate less stress for homosexual men and women. Thus, for the purpose of attaining this goal, efforts to change the socializing agents in societies and subcultures oppressive to homosexual men and women are to be encouraged and supported. Finally, there are numerous theories on the etiology of homosexual orientation. They reflect the predominant paradigms of the times in which they developed. Currently, the neurohormonal and evolutionary theories provide a compelling rationale.

See also: Gender Differences in Personality and Social Behavior; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Mental Illness, Etiology of.

Bibliography

American Psychiatric Association, 1968. Diagnostic and Statistical Manual of Mental Disorders (second ed., 11th printing, January 1978). American Psychiatric Association, Washington, DC.

American Psychiatric Association, 1980. Diagnostic and Statistical Manual of Mental Disorders, third ed. American Psychiatric Association, Washington, DC.

American Psychiatric Association, 1987. Diagnostic and Statistical Manual of Mental Disorders, third ed., rev. American Psychiatric Association, Washington, DC.

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders, fourth ed. American Psychiatric Association, Washington, DC.

American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders, fourth ed., text rev. American Psychiatric Association, Washington, DC.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Association, Arlington, VA.

van den Akker, H., van der Ploeg, R., Scheepers, P., 2013. Disapproval of homosexuality: comparative research on individual and national determinants of disapproval of homosexuality in 20 European countries. International Journal of Public Opinion Research 25 (1), 64–86.

Berube, A., 1990. Coming Out under Fire: The History of Gay Men and Women in World War Two. Free Press, New York.

Boswell, J., 1980. Christianity, Social Tolerance, and Homosexuality. University of Chicago Press, Chicago.

Boswell, J., 1994. Same-Sex Unions in Premodern Europe. Villard Books, New York.

Cantarella, E., 1992. Bisexuality in the Ancient World. Yale University Press, New Haven.

Chiang, H.H.-H., 2008. Effecting science, affecting medicine: homosexuality, the Kinsey reports, and the contested boundaries of psychopathology in the United States, 1948–1965. Journal of the History of the Behavioral Sciences 44 (4), 300–318.

Cochran, S.D., Sullivan, J.G., Mays, V.M., 2003. Prevalence of mental disorders, psychological distress, and mental health services use among lesbian, gay and bisexual adults in the United States. Journal of Consulting and Clinical Psychology 71 (1), 53–61.

Conrad, P., Angell, A., 2004. Homosexuality and remedicalization. Society 41 (5), 32–39.

Crompton, L., 2003. Homosexuality & Civilization. Belknap Press, Cambridge, MA.

De Block, A., Adriaens, P.R., 2013. Pathologizing sexual deviance: a history. Journal of Sex Research 50 (3–4), 276–298.

Dover, K.J., 1978. Greek Homosexuality. Harvard University Press, Cambridge, MA.

Drescher, J., 1998. I am your handyman: a history of reparative therapies. Journal of Homosexuality 36 (1), 5–24.

Drescher, J., 2010. Queer diagnoses: parallels and contrasts in the history of homosexuality, gender variance, and the diagnostic and statistical manual. Archives of Sexual Behavior 39, 427–460.

Ellis, A.L., Mitchell, R.W., 2000. Sexual orientation. In: Szuchman, L.T., Muscarella, F. (Eds.), Psychological Perspectives on Human Sexuality. John Wiley & Sons, New York, pp. 196–231.

Flentje, A., Heck, N.C., Cochran, B.N., 2013. Sexual reorientation therapy interventions: perspectives of ex-ex-gay individuals. Journal of Gay & Lesbian Mental Health 17, 256–277.

Fradella, H.F., 2002. Legal, moral, and social reasons for decriminalizing sodomy. Journal of Contemporary Criminal Justice 18 (3), 279–301.

Freud, S., 1962. Three Essays on the Theory of Sexuality (S. James, Trans.). Basic Books, New York.

Gonsiorek, J.C., 2004. Reflections from the conversion therapy battlefield. The Counseling Psychologist 32 (5), 750–759.

Grace, A.P., 2008. The charisma and deception of reparative therapies: when medical science beds religion. Journal of Homosexuality 55 (4), 545–580.

Greenberg, D.F., 1988. The Construction of Homosexuality. University of Chicago Press, Chicago.

Hadler, M., 2012. The influence of world societal forces on social tolerance. A time comparative study of prejudices in 32 countries. The Sociological Quarterly 53, 211–237.

Hancock, K.A., Gock, T.S., Haldeman, D.C., 2012. Science meets practice in determining effectiveness of sexual orientation change efforts. American Psychologist 67 (6), 499–500.

Hooker, E., 1957. Male homosexuality in the Rorschach. Journal of Projective Techniques 22, 957–963.

Isay, R.A., 1989. Being Homosexual: Gay Men and Their Development. Farrar Straus Giroux, New York.

Lee, S., 2001. From diversity to unity: the classification of mental disorders in 21st-century China. Cultural Psychiatry: International Perspectives 24 (3), 421–431.

LeVay, S., 2011. Gay, Straight, and the Reason Why. University Press, Oxford.

Mendelson, G., 2003. Homosexuality and psychiatric nosology. Australian and New Zealand Journal of Psychiatry 37, 678–683.

Meyer, I., 2013. Predjudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. Psychology of Sexual Orientation and Gender Diversity 1 (S), 3–26. Reprinted from: Psychological Bulletin 129(5), pp. 674–697 (2003).

Morin, S.F., Rothblum, E.D., 1991. Removing the stigma: fifteen years of progress. American Psychologist 46 (9), 947–949.

Muscarella, F., 2006. The evolution of male-male sexual behavior in humans: the alliance theory. Journal of Psychology and Human Sexuality 18 (2/3,4), 275–311.

Mustanski, B., Liu, R.T., 2013. A longitudinal study of predictors of suicide attempts among lesbian, gay, bisexual, and transgender youth. Archives of Sexual Behavior 42, 437–448.

Reading, R., Rubin, L.R., 2011. Advocacy and empowerment: group therapy for LGBT asylum seekers. Traumatology 17 (2), 86–98.

Rosario, M., Schrimshaw, E.W., 2014. Theories and etiologies of sexual orientation. In: Tolman, D.L., Diamond, L.M. (Eds.), APA Handbook of Sexuality and Psychology, vol. 1. American Psychological Association, Washington, DC, pp. 555–596.

Scasta, D., 1998. Historical perspectives on homosexuality. Journal of Gay & Lesbian Psychotherapy 2 (4), 3–17.

Silverstein, C., 2008. Are you saying homosexuality is normal? Journal of Gay & Lesbian Mental Health 12 (3), 277–287.

Sue, D.W., 2010. Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation. Wiley, Hoboken.

Tacitus, 2009. Agricola and Germany. A new translation by A.R. Birley. Oxford University Press, Oxford.

World Health Organization, 2010. http://apps.who.int/classifications/icd10/browse/2010/en.

Wu, J., 2003. From "*Long Yang*" and "*Dui Shi*" to tongzhi: homosexuality in China. Journal of Gay & Lesbian Psychotherapy 7 (1/2), 117–143.

Closed and Open Systems: Organizational

Thomas G Cummings, University of Southern California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7237–7241, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editors.

Abstract

Closed and open systems refer to whether organizational entities, such groups and organizations, are viewed as relatively closed or open to their environment. Such a perspective has a profound influence on how organizations are described and studied. When treated as closed systems, organizations are unaffected by their environment, and attention is directed inward to internal structures and behaviors. As open systems, organizations are interdependent with their environment, and focus is outward to how such interaction is managed. Knowledge of closed and open systems derives from the broad framework of general systems theory which seeks to explain the structure and behavior of complex wholes called systems. This article shows how general systems theory is applied to organizations. It describes the systemic properties of organizations as closed and open systems, and draws implications for how they are structured and managed.

Closed and open systems refer to whether organizational entities, such as groups and organizations, are viewed as relatively closed or open to their environment. Such a perspective has a profound influence on how organizations are described and studied. When treated as closed systems, organizations are unaffected by their environment, and attention is directed inward to internal structures and behaviors. As open systems, organizations are interdependent with their environment, and focus is outward to how such interaction is managed.

Knowledge of closed and open systems derives from the broad framework of general systems theory (GST) which seeks to explain the structure and behavior of complex wholes called systems (e.g., von Bertalanffy, 1956; Miller, 1978). This broad metatheory is based on related research from the physical, biological, and social sciences, and seeks to discover general laws which apply to all levels of systems from single cells to societies.

This article shows how GST applies to organizations. It describes the systemic properties of organizations as closed and open systems, and explains how they are structured and managed.

Organizations as Systems

Organizational scholars use GST to describe the general properties of organizational systems. This includes defining their constituent parts and how they are interrelated to form a system. It also involves identifying different levels of organizational systems and explaining how they interact with each other.

Organizational Parts, Relationships, and Wholes

A key premise of GST has to do with the definition of a system and how it forms an organized whole. A system is composed of parts and relationships among them. The system provides the framework or the organizing principle for structuring the parts and relationships into an organized whole. This makes systems capable of behaving in ways that are greater than merely the sum of the behaviors of their parts, thus leading to the common adage: "the whole is greater than the sum of its parts." Organizational scholars have expended considerable effort in identifying the constituent members or subunits of organizational systems and examining relationships among them. They have sought to discover the organizing principles through which the parts are arranged into a coherent whole. For example, job design researchers have identified different elements of jobs; they have shown how they can be combined to affect employee motivation. Group dynamics scholars have spent considerable time addressing issues of group membership and member interaction. They have discovered different ways of structuring groups to perform tasks that members could not do working alone. They have shown how, under certain conditions, groups can outperform individuals, thus leading to various high-performing group designs, such as self-managing teams, quality circles, and cross-functional teams. Similarly, organization theorists have identified the different components of organizations and examined relations among them. They have found different ways to organize the components and relationships for competitive advantage (Cummings and Worley, 2015).

Levels of Organizational Systems

A second principle of GST has to do with the multilevel nature of systems. Systems exist at different levels; the levels exhibit a hierarchical ordering, with a higher level of system being composed of systems at lower levels. For example, societies are composed of organizations; organizations are composed of groups; groups are composed of individuals; and so on. Higher-level systems provide constraints and opportunities for how a system organizes its parts, and the nature of those parts affects the system's organizing possibilities. Thus, to describe a system at a particular level and explain its behavior, it is necessary to look both upward to the higher level system

within which it is embedded and downward to its constituent parts.

This multilevel perspective has led scholars to identify different levels of organizational systems, and to focus on understanding them and how they interact with each other. Considerable attention has been directed at specifying appropriate levels of analysis, both for conceptualizing about organizational systems and for aggregating and disaggregating data that apply to different levels. For example, scholars have developed analytical guides for aggregating data from individuals to devise measures of group and organization functioning. As researchers have developed more comprehensive theories and more powerful analytical methods, they have made finer distinctions among levels of organizational systems, particularly above the organization level. This has led to at least five levels of organizational systems:

1. individual member, role, or job;
2. group;
3. organization;
4. population of organizations and/or alliance among organizations;
5. community of populations and/or community of alliances.

Organizations as Closed Systems

Closed systems do not interact with the environment, and consequently their behavior depends largely on the internal dynamics of their parts. They seek to maintain a steady state or equilibrium among their parts while performing goal-directed behaviors. Because the environment is inconsequential for goal achievement, system behavior is highly specified and maximally controlled within the system.

Until the late 1960s, organizational scholars tended to employ a closed-system perspective for studying organizations. Organizational environments were seen as relatively simple and predictable, and, thus, were not problematic or significant for how organizations behaved. Attention was directed at the internal dynamics of groups and organizations, particularly at how member behavior was controlled to achieve specific objectives and goals. This led to extensive knowledge of organizational control mechanisms and search for the best way to structure organizational systems. For example, at the individual level, job design researchers discovered how to analyze and maximally specify the most efficient way to perform various tasks. At the group level, researchers showed how group structures and processes could contribute to member conformity to group norms. At the organization level, scholars studied a variety of devices for controlling member behavior, such as managerial hierarchy, rules/procedures, and functional design.

Organizations as Open Systems

In the late 1960s, organizational scholars began to broaden their focus to external forces affecting organizational systems. This open systems view was fueled by growing applications of GST to the social sciences and by realization that the behavior of organizational systems could not be adequately explained without examining environmental relationships and their effects on the system (e.g., Aldrich, 1979; Buckley, 1967; Katz and Kahn, 1966; Lawrence and Lorsch, 1967, Scott and Davis, 2007; Thompson, 1967). It has led to considerable research about organizational environments, their dynamics and effects, and how organizational systems interact with them. An open systems perspective draws attention to how organizations exchange information and resources with their environment, and how the two mutually influence each other. Moreover, it provides a number of useful concepts for understanding how organizations maintain functional autonomy while influencing and adapting to external forces.

Critical Functions

To survive and prosper, open systems need to perform at least four critical functions:

1. transformation of inputs of energy and information to produce useful outputs;
2. transaction with the environment to obtain needed inputs and to dispose of outputs;
3. regulation of system behavior to achieve stable performance;
4. adaptation to changing conditions.

Because these different functions often place conflicting demands and tension on open systems, system viability depends on maintaining a dynamic balance among them. Organizational researchers have devoted considerable time identifying and explaining how these four functions operate and contribute to organizational survival and effectiveness. This has led to knowledge about how organizations and groups produce products and services through operating and developing different technologies; how they protect their technologies from external disruptions while acquiring raw materials and marketing finished products; how they regulate themselves for stable performances while initiating and implementing innovation and change. This research defines a key role of management in organizational systems as sustaining a dynamic balance among these four functions; one that allows the organization sufficient stability to operate rationally yet requisite flexibility to adapt to changing conditions.

Information and Resource Flows

As open systems, organizations seek to sustain a cycle of activities aimed at taking inputs of information and resources from the environment, transforming them into outputs of goods and services, and exporting them back to the environment. This cycle enables organizations to replenish themselves continually, so long as the environment provides sufficient inputs and the organization delivers valued outputs. Considerable research has gone into understanding how organizations manage these information and resource flows. One perspective focuses on how organizations process information to learn how to improve themselves and to relate to their environments. Organizations must be capable of learning from their experiences and of disseminating such knowledge widely if

they are to change themselves to respond to emerging conditions. Consequently, research has been directed at organizational learning and knowledge management, particularly at discovering mechanisms to enhance learning capability, such as shared databases, groupware, and decision-support systems. Another view concentrates on how organizations compete for resources through managing key resource dependencies. Attention has been directed at how organizations gain access to resources without becoming overly dependent on those who supply them. Still another perspective focuses on how organizations gain legitimacy from environmental institutions, so they can continue to function with external support. Researchers have sought to identify the institutional demands of powerful resource providers, and how organizations respond to them.

Boundaries

In managing information and resource flows, organizations, like all open systems, seek to establish boundaries around their activities. These organizational boundaries must be sufficiently permeable to permit necessary environmental exchange, yet afford the organization adequate protection from external demands to allow for rational operation (Santos and Eisenhardt, 2005). Organizational scholars have devoted considerable attention to understanding the dual nature of organizational boundaries. They have studied various boundary-spanning roles that relate the organization to its environment, such as sales, public relations, and purchasing. They have examined how organizational members perceive and make sense out of environmental input, and how organizational boundaries vary in sensitivity to external influences. Researchers have also identified different strategies for protecting transformation processes from external disruptions while remaining responsive to suppliers and customers.

Self-Regulation

Viewed as open systems, organizations use information about how they are performing to modify future behaviors. Referred to as cybernetics, this information feedback enables organizations to be self-regulating (e.g., Ashby, 1956). They can adjust their behavior to respond to deviations in expected performance. To be effective, however, organizations must have sufficient diversity of responses to match the variety of disturbances encountered. Thus, as technologies and environments become more complex and uncertain, organizations seek to become more flexible and nimble. They can enhance their adaptive capabilities through such innovations as downsizing, employee empowerment, and reengineering.

Extensive research has been devoted to understanding how organizations regulate themselves. Using modern information technology, organizations have developed a variety of methods for setting goals, obtaining information on goal achievement, and making necessary changes. For example, database technologies enable organizations to store large batches of information and to logically connect events, actions, and outcomes. Employees can access and revise this information through online, common databases.

Equifinality

As open systems, organizations display the property of equifinality. They can achieve objectives with varying inputs and in different ways. Consequently, there is no one best way to design and manage organizations, but there are a variety of ways to achieve effective performance. This contrasts sharply with a closed system perspective which seeks the one best way to structure and control organizations.

Organizational researchers have devoted considerable attention to identifying different choices for designing and managing organizations. This has led to a virtual revolution in new organizational designs aimed primarily at making organizations leaner, more flexible, and more responsive to human resources. For example, traditional bureaucratic structures that emphasize efficiency and control have increasingly been supplanted with designs that emphasize flexibility and innovation, such as matrix organizations, horizontal organizations, network organizations, and virtual organizations. Moreover, researchers have developed contingency theories which specify under what technological, environmental, and human conditions different organizational designs are likely to be most successful. For example, these newer, more innovative designs are best suited to situations where technologies are complex, environments are unpredictable, and people have high growth needs.

Promising Trends

Organizational researchers have recently applied systems concepts to understand how organizations can adapt to rapidly changing, unpredictable environments. They have borrowed heavily from complexity theory which seeks to explain the behaviors and changes that can occur when the parts of complex systems interact (e.g., Holland, 1995; Brown and Eisenhardt, 1998; Eisenhardt and Piezunka, 2011) in nonlinear and dynamic ways and can invent entirely new responses to external forces. These concepts provide a dynamic, change-oriented perspective on organizations. They help to explain how organizations can restructure themselves continually to keep pace with fast-changing environments.

See also: Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Leadership in Organizations, Sociology of; Organization: Overview; Organizational Climate; Organizational Decision Making; Organizations and the Law; Organizations, Sociology of.

Bibliography

Aldrich, H.E., 1979. Organizations and Environments. Prentice-Hall, Englewood Cliffs, NJ.
Ashby, W.R., 1956. An Introduction to Cybernetics. Wiley, New York.
von Bertalanffy, L., 1956. General systems theory. General Systems 1, 1–10.
Brown, S.L., Eisenhardt, K.M., 1998. Competing on the Edge. Harvard Business School Press, Boston.
Buckley, W., 1967. Sociology and Modern Systems Theory. Prentice-Hall, Englewood Cliffs, NJ.

Cummings, T., Worley, C. (Eds.), 2015. Organization Development and Change, tenth ed. Cengage Learning, Stamford.

Eisenhardt, K.M., Piezunka, H., 2011. Complexity theory and corporate strategy. In: Allen, P., Maguire, S., McKelvey, B. (Eds.), The Sage Handbook of Complexity and Management. Sage Publications, London, pp. 506–523.

Holland, J.H., 1995. Hidden Order: How Adaptation Builds Complexity. Addison-Wesley, Reading, MA.

Katz, D., Kahn, R.L., 1966. The Social Psychology of Organizations. Wiley, New York.

Lawrence, P.R., Lorsch, J.W., 1967. Organization and Environment. Harvard University Press, Boston.

Miller, J.G., 1978. Living Systems. McGraw-Hill, New York.

Santos, F.M., Eisenhardt, K.M., 2005. Organizational boundaries and theories of organization. Organization Science 16 (5), 491–508.

Scott, W.R., Davis, G.F., 2007. Organizations and Organizing: Rational, Natural and Open System Perspectives. Pearson Education, Upper Saddle River.

Thompson, J.O., 1967. Organizations in Action. McGraw-Hill, New York.

Cluster Randomized Trials, Ethical Implications of

Monica Taljaard, Ottawa Hospital Research Institute, Ottawa Hospital, Ottawa, ON, Canada; and University of Ottawa, Ottawa, ON, Canada
Charles Weijer, Rotman Institute of Philosophy, Western University, London, ON, Canada
Jeremy M Grimshaw, Ottawa Hospital Research Institute, Centre for Practice-Changing Research (CPCR), The Ottawa Hospital, Ottawa, ON, Canada; and University of Ottawa, Ottawa, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

While cluster randomized trials (CRTs) have unique features that differentiate them from conventional (individually) randomized trials, standard research ethics guidelines by and large assume individual randomization. The Ottawa Statement on the Ethical Design and Conduct of Cluster Randomized Trials provides recommendations specific to CRTs to assist researchers and research ethics committees in the ethical design, conduct, and review of CRTs. In this article we describe the unique characteristics of CRTs, identify the main ethical issues raised, and discuss the application of the Ottawa Statement to CRTs in health, social, and behavioral research using a series of examples.

Introduction

Cluster randomized trials (CRTs), also known as place or group randomized trials, are an increasingly common research design in the areas of health services, public health, education, welfare, criminal justice, and international health and development research (Boruch et al., 2004; Donner and Klar, 2000; Murray, 1998). In CRTs, the individuals whose outcomes are of interest in the trial are randomized – not in the conventional way as independent individuals – but as naturally occurring groups or sites such as communities, housing developments, schools, or worksites. All individuals in a particular group (i.e., 'cluster') are allocated to the same study arm. It has become widely recognized that CRTs have unique methodological requirements because observations from multiple individuals in the same cluster are usually positively correlated; less well recognized is that CRTs also have unique implications for research ethics (Edwards et al., 1999; Hutton, 2001; Klar and Donner, 2007). Because these implications have not been explicitly addressed in major research ethics codes such as the Declaration of Helsinki (World Medical Association, 2008) and the International Ethical Guidelines for Biomedical Research Involving Human Subjects (Council of International Organizations of Medical Science (CIOMS), 2002), research ethics committees (RECs) and researchers have been left uncertain as to the best ethical practices in CRTs. The Ottawa Statement on the Ethical Design and Conduct of Cluster Randomized Trials (Weijer et al., 2012) was developed to address this gap. It was the end-product of a 5-year international collaboration (Taljaard et al., 2009), which included a series of empirical studies documenting ethical issues in CRTs (Chaudhry et al., 2013; McRae et al., 2013a,b; Taljaard et al., 2011, 2014a,b), a detailed ethical analysis rooted in ethical principles and international regulations (McRae et al., 2011a,b; Binik et al., 2011; Gallo et al., 2012), and a consensus process that included an international interdisciplinary expert panel to write ethics guidelines for CRTs (Weijer et al., 2012; Taljaard et al., 2013). In this article we describe the unique characteristics of CRTs, identify the main ethical issues raised, and discuss key considerations in the application of the Ottawa Statement for CRTs in health, social, and behavioral research using a series of examples.

General Ethical Principles

Four fundamental ethical principles govern all research involving human participants: respect for persons, beneficence, justice, and respect for communities (World Medical Association, 2008; CIOMS, 2002). These principles are grounded in moral theories, the research ethics literature, and convention. Respect for persons implies the need to take seriously the autonomous choices of prospective research participants and to protect those who are incapable of making autonomous choices. The principle of beneficence obliges researchers not to inflict unnecessary harm and, where possible, to promote the good for research participants. The principle of justice may be defined as the ethical obligation to distribute the benefits and burdens of research fairly. Respect for communities implies that researchers have an obligation to respect communal values, protect and empower communities, and, where applicable, abide by the decisions of legitimate communal authorities. There is considerable debate about the applicability of biomedical ethical principles to social and behavioral research (Hoeyer et al., 2005). In the context of international research, in particular, a variety of alternative principles have been proposed that include reinterpretation of the principle of justice and novel notions such as international solidarity (Benatar, 2004).

Characteristics of CRTs

A key characteristic of CRTs is that they have a hierarchical structure, which allows the units of randomization, intervention, and observation to differ within a single study. Conventional trials in psychology and medicine have the individual as the unit of randomization, intervention, and observation.

Because CRTs allocate intact clusters with multiple observations from each cluster, they often randomize a relatively small number of independent units, but the number of observations per cluster may be relatively large. Study interventions are often complex, and may include interventions targeted at one or more levels: the cluster organization, cluster-level participants (e.g., health professionals, employers, and teachers), and individual-level participants (e.g., patients, employees, or students). Data collection procedures may similarly be at the individual level and/or the cluster level, but some CRTs use routinely collected data for outcome assessment without interaction with individuals. In such studies, outcomes may be transmitted to researchers anonymously, in the form of rates or cluster-level summaries. Analyses in CRTs may be conducted at either the individual or cluster level.

Table 1 describes characteristics of a random sample of 300 CRTs in health research published between 2000 and 2008, identified using a validated electronic search strategy implemented in Medline. Further details of the search strategy and sample are presented elsewhere (Ivers et al., 2011). The units of randomization were diverse, including primary care practices, hospitals, nursing homes, residential areas, schools, classrooms, worksites, and groups such as sports teams or social clubs. Approximately one-third of trials had interventions targeted solely at the cluster level, one-third had interventions targeted solely at the individual level, and one-third had complex interventions targeted at multiple levels. The majority of trials (72%) included data collection via questionnaires administered to individual-level participants, but one-fifth included questionnaires administered to cluster-level participants. The use of routinely collected data sources for outcome assessment was common. In just over one-fifth of studies there was no recruitment of individuals to the trial.

Reporting of ethical issues in the sample was poor (Taljaard et al., 2011): approximately one-quarter of trial reports did not include a statement about research ethics review, and one-third did not report on informed consent. In a survey of the corresponding authors of these trials (Chaudhry et al., 2013; Taljaard et al., 2014a), 70% of respondents indicated that the trial included some level of informed consent for study interventions, and 88% that it included some level of consent for data collection. Just over half of respondents indicated that consent procedures applied to participants at multiple levels. The use of 'gatekeepers' was common with 83% of respondents indicating the presence of a cluster gatekeeper in some role. Gatekeepers, also called 'guardians' or 'cluster representation mechanisms' are unique to CRTs, and are individuals or bodies that are seen to represent the interests of cluster members, clusters, or institutions.

Table 1 Characteristics of a random sample of 300 published CRTs between 2000 and 2008

Country of study recruitment	
Canada or USA	130 (43.3%)
UK or Ireland	50 (16.7%)
Other developed country	75 (25.0%)
Developing country	45 (15.0%)
Unit of randomization	
Primary care practices/professionals	121 (40.3%)
Hospitals/wards	25 (8.3%)
Nursing homes/wards	16 (5.3%)
Residential areas	47 (15.7%)
Schools/classrooms	66 (22.0%)
Worksites	16 (5.3%)
Sports teams/social groups	9 (3%)
Number of clusters randomized	
Median (IQR)	21.0 (12–52)
Range	2–605
Average cluster size	
Median (IQR)	33.9 (12.5–88.5)
Range	1.7–122855
Study interventions[a]	
Aimed at participants at cluster level	143 (47.7%)
Aimed at cluster organization	106 (35.3%)
Educational intervention aimed at individual-level participants	165 (55.0%)
Therapeutic intervention aimed at individual-level participants	51 (17.0%)
Target of study interventions	
Cluster level only	99 (33.0%)
Individual level only	105 (35%)
Both cluster and individual levels	96 (32%)
Types of control interventions	
Not reported	42 (14.0%)
No active intervention	156 (52.0%)
Minimal version of active intervention	52 (17.3%)
Placebo or sham intervention	25 (8.3%)
Other active intervention	12 (4.0%)
Wait-list or delayed intervention	13 (4.3%)
Data collection procedures[a]	
Routinely collected data	140 (46.7%)
Physical examination or other intrusive method of data collection	63 (21.0%)
Questionnaires administered to individual-level participants	215 (71.7%)
Questionnaires administered to cluster-level participants	63 (21.0%)
Levels of data collection	
Individual-level only	192 (64%)
Cluster-level only	23 (8%)
Both individual and cluster level	40 (13%)
Routine data only	45 (15%)
Any recruitment of individuals into the trial?	
Yes	233 (77.7%)
No	67 (22.3%)

[a]A trial can be classified into more than one category.

Methodological Implications of CRTs

CRTs are more complex to design, analyze, interpret, and report than individually randomized trials (Campbell et al., 2012). In a CRT, observations from the same cluster are usually positively correlated; as a result, CRTs require a larger total number of individuals to yield the same power as in an individually randomized trial (Donner and Klar, 2000). They also require special methods of analysis as standard methods assume statistical independence. Results from CRTs are more difficult to interpret because they are particularly susceptible to various sources of bias. For example, CRTs are at increased risk of baseline imbalances because the number of randomization units is often limited. For that reason, it is important to use stratification, matching, or some other method of promoting

balance in the design of the trial. CRTs are more susceptible to selection bias as individual participants may need to be identified or enrolled after clusters have already been randomized (Eldridge et al., 2009; Hahn et al., 2005). This is an important concern in CRTs evaluating behavioral interventions, as it is more difficult to blind participants to the study interventions.

A review of the methodological and reporting quality of the sample of 300 published CRTs showed that methods used in CRTs are often suboptimal (Ivers et al., 2011). Because methodologically flawed studies expose individuals to experimental interventions without the potential for scientific gain, failure to adhere to these basic requirements for CRTs may lead to studies that are inherently unethical. Therefore, it is important that investigators adopt statistical methods appropriate for this design. Multiple resources are now available to help investigators plan, conduct, analyze, and report CRTs (Campbell and Walters, 2014; Donner and Klar, 2000; Murray, 1998).

Reasons for Adopting Cluster Randomization

Because of their inherent methodological challenges, the decision to employ the CRT design requires careful justification. There are usually cogent reasons to adopt a CRT: in the case of cluster-level or professional-level interventions, the reasons are obvious; that is, the trial simply cannot be done in any other way. But in the case of individual-level interventions, the trial could – in principle – have used individual randomization. In such trials, cluster randomization may nevertheless be selected for scientific reasons (e.g., to avoid contamination and to estimate population-level effects of an intervention), or for other logistical, practical, or political reasons.

The need to avoid contamination is a common concern in CRTs. Contamination occurs when individuals in the control arm are partially exposed to the intervention through interaction with individuals receiving the study intervention, thus biasing the results toward smaller effect sizes. The risk of contamination is particularly great in the case of unblinded or behavioral interventions. For example, in a school-based CRT of an intervention to promote healthy behaviors, entire schools may be randomized to minimize the likelihood of children in different intervention groups sharing information about the trial.

Even in the case of individual-level therapeutic interventions with blinding, there may be good reasons to adopt the cluster randomized design. For example, the ObaapaVitA trial (Kirkwood et al., 2010) was a double-blind, placebo-controlled trial to evaluate the effect of weekly, low-dose vitamin A supplementation on pregnancy-related and all-cause female mortality in Ghana. A total of 1086 small clusters of compounds were randomized to either vitamin A or placebo capsules. Fieldworkers visited compounds to recruit women (approximately 100 000 per arm), distribute capsules, and gather data on pregnancies, births, and deaths. Although individual randomization could in principle have been used, the use of cluster randomization considerably simplified the trial organization and fieldwork. Randomization of clusters also minimized the possibility of women receiving the wrong capsules as fieldworkers only had one type of capsule in their possession during any week. It may also have helped to prevent the risk of sharing of capsules among women (in the hope of getting some benefit should they be randomized to placebo).

Other common reasons to adopt the CRT design include the need to secure cooperation from cluster decision makers and facilitate recruitment of participants. In Progresa, a poverty alleviation program in Mexico (Parker and Teruel, 2005), a sample of 500 rural villages were randomly assigned to intervention or control arms. In the intervention arm, poor families were offered cash transfers in exchange for sending their children to school and for regular attendance at health clinics. Officials decided against randomization at the household level as they were concerned that perceived favoritism in the selection process would have created discontent and resentment between intervention and control households within the same communities.

Ethical Implications of CRTs

While the ethics of individually randomized trials is well understood, CRTs differ from individually randomized trials in important ways; this complicates the application of established research ethics guidelines. For example, in an individually randomized trial, the units of randomization, intervention, and outcome assessment are the same (e.g., the citizen, student, and patient) – in a CRT, the units of allocation, intervention, and outcome measurement may differ within any given trial. This raises questions as to who ought to be considered research participants and whose informed consent is required. In an individually randomized trial, participants are identified and approached for consent prior to study randomization – in a CRT, clusters may be randomized before individual cluster members can be identified and approached for informed consent. This raises the question whether and from whom consent for randomization is required. In some CRTs, gatekeepers or decision makers may exist within each cluster, but it is unclear whether gatekeepers may provide consent to randomization. In individually randomized trials, the study intervention is directed at the individual, but in a CRT interventions may be directed at the level of the individual, the cluster, or both. Cluster-level interventions (e.g., public educational messages) may be difficult for individuals to avoid; in such cases, meaningful refusal of study participation is impossible. While conventional randomized trials involve individuals, CRTs involve social groups as well as individuals. However, our current understanding of the moral status of social groups, and of the gatekeepers who speak on their behalf, is incomplete. Group interests may conflict with individual interests and this complicates benefit–harm assessments. Finally, some CRTs involve very large clusters, and requiring informed consent may make the study infeasible. Further in-depth discussion of the ethical issues raised by CRTs is presented elsewhere (Weijer et al., 2011).

Examples

It is convenient to distinguish among three kinds of CRTs, depending on the level of the intervention: cluster–cluster trials (interventions delivered at cluster level), professional–cluster

trials (interventions delivered to professionals in each cluster), and individual–cluster trials (interventions delivered directly to individual cluster members) (Eldridge et al., 2005). Some CRTs may involve more than one kind of intervention.

Example 1: Cluster–Cluster Trial

The COMMIT study (Community Intervention Trial for Smoking Cessation, 1995a,b) was designed to test the effectiveness of a comprehensive, community-oriented approach to influence citizens' smoking behaviors. Twenty-two communities with populations between 50 000 and 250 000 in the USA and Canada were randomly assigned to intervention or control arms. The researchers interacted with the communities via a community advisory board, consisting of key community representatives. The intervention included activities focused on public education using mass media and organized community events, promotion of policies to restrict the sale of tobacco to youth, and development of smoking cessation resources and activities. Population-based surveys, using random digit telephone dialing, were used to measure outcomes. Before randomization and at the end of the study, cross-sectional samples of approximately 2500 households per community were surveyed about their smoking behaviors. In addition, cohorts of approximately 550 heavy and 550 light-to-moderate smokers were identified in each community and followed to assess quit rates. Although the intervention significantly improved quit rates among light-to-moderate smokers, there was no significant effect on quit rates among heavy smokers or on the community prevalence of smoking.

Example 2: Professional–Cluster Trial

The Guidelines Trial (Althabe et al., 2008) was a CRT to evaluate a multifaceted behavioral intervention to improve the use of episiotomy and management of the third stage of labor, targeted at birth attendants in Latin America. Nineteen hospitals (approximately 500 birth attendants) in Argentina and Uruguay were randomized to either intervention or control arms. In the intervention arm, teams of birth attendants (physicians, residents, and midwives) were identified as opinion leaders and participated in a workshop to develop clinical guidelines. Teams returned to their respective hospitals, disseminated guidelines, trained other birth attendants, and implemented a system of reminders and feedback. Birth attendants in the control arm received no intervention. Rates of episiotomy and prophylactic use of oxytocin in the third stage of labor were collected from approximately 5500 births using a standard clinical record form. No information identifying individual women was transmitted from the hospitals. Birth attendants completed self-administered questionnaires measuring their readiness to change their behaviors. Responsible authorities from all the hospitals provided written permission, and birth attendants provided written informed consent for study participation. Rates of prophylactic oxytocin use increased significantly; rates of episiotomy use decreased significantly in intervention versus control hospitals. Approval for the study was obtained from eight institutional review boards.

Example 3: Individual–Cluster Trial

Robroek et al. (2012) evaluated a CRT to evaluate the effectiveness and cost-effectiveness of an individually tailored worksite health promotion program designed to improve employees' physical activity and nutrition. The intervention was delivered at the individual level, but worksites were chosen as the units of randomization to enhance compliance and promote adherence using the worksite as a natural social network. Worksites randomized to the intervention arm received computer-tailored physical activity and nutrition advice, a monitor function that allowed individuals to chart their progress, a personal coach who gave continuous feedback and support through monthly e-mails, and the opportunity to seek advice from a range of professionals. The primary outcomes were compliance with national recommendations on physical activity and fruit and vegetable intake, assessed by means of questionnaires and physical examination. After obtaining agreement from the employer, 74 eligible worksites were randomized; a total of 924 employees who provided written informed consent were enrolled in the study. No statistically significant effects due to the intervention were observed. The local REC approved the study.

The Ottawa Statement

The Ottawa statement sets out 15 recommendations for the ethical design and conduct of CRTs, covering the justification of cluster randomization, the need for REC review, the identification of research participants, obtaining informed consent, the role of gatekeepers in protecting group interests, the assessment of benefits and harms, and the protection of vulnerable participants. The specific recommendations are summarized in **Table 2**. Here we interpret recommendations in the Ottawa Statement with respect to CRTs in health, social, and behavioral research using three examples.

Justification for the CRT Design

Recommendation 1 of the Ottawa Statement requires that the choice of cluster randomization be carefully justified. The design should not be adopted in the mistaken belief that the need for informed consent can be avoided. In example 1 (COMMIT) the choice was obvious as it involved a community-wide smoking cessation campaign; likewise in example 2, it would have been impossible for birth attendants to avoid letting the training influence treatment of all women under their care. In example 3, the intervention was an individually tailored health promotion intervention, and employees could, in principle, have been randomized individually. In this trial, entire worksites were randomized to avoid contamination, as well as to enhance compliance with the intervention.

Need for Research Ethics Review

National and international research ethics guidelines stipulate that all human research must be submitted to and

Table 2 Ottawa statement: summary of recommendations

	Ethical issue	*Recommendation*
1	Justifying the cluster randomized design	Researchers should provide a clear rationale for the use of the cluster randomized design and adopt statistical methods appropriate for this design
2	Research ethics committee review	Researchers must submit a CRT involving human research participants for approval by a research ethics committee before commencing
3	Identifying research participants	Researchers should clearly identify the research participants in CRTs. A research participant can be identified as an individual whose interests may be affected as a result of study interventions or data collection procedures – that is, an individual who is the intended recipient of an experimental (or control) intervention; or who is the direct target of an experimental (or control) manipulation of his or her environment; or with whom an investigator interacts for the purpose of collecting data about that individual; or about whom an investigator obtains identifiable private information for the purpose of collecting data about that individual. Unless one or more of these criteria is met, an individual is not a research participant
4	Obtaining informed consent	Researchers must obtain informed consent from human research participants in a CRT, unless a waiver of consent is granted by a REC under specific circumstances
5		When participants' informed consent is required, but recruitment of participants is not possible before randomization of clusters, researchers must seek participants' consent for trial enrollment as soon as possible after cluster randomization – that is, as soon as the potential participant has been identified, but before the participant has undergone any study interventions or data collection procedures
6		A research ethics committee may approve a waiver or alteration of consent requirements when the research is not feasible without a waiver or alteration of consent, and the study interventions and data collection procedures pose no more than minimal risk
7		Researchers must obtain informed consent from professionals or other service providers who are research participants unless conditions for a waiver or alteration of consent are met
8	Gatekeepers	Gatekeepers should not provide proxy consent on behalf of individuals in their cluster
9		When a CRT may substantially affect cluster or organizational interests, and a gatekeeper possesses the legitimate authority to make decisions on the cluster or organization's behalf, the researcher should obtain the gatekeeper's permission to enroll the cluster or organization in the trial. Such permission does not replace the need for the informed consent of research participants
10		When cluster randomized trial interventions may substantially affect cluster interests, researchers should seek to protect cluster interests through cluster consultation to inform study design, conduct, and reporting. Where relevant, gatekeepers can often facilitate such a consultation
11	Assessing benefits and harms	The researcher must ensure that the study intervention is adequately justified. The benefits and harms of the study intervention must be consistent with competent practice in the field of study relevant to the cluster randomized trial
12		Researchers must adequately justify the choice of the control condition. When the control arm is usual practice or no treatment, individuals in the control arm must not be deprived of effective care or programs to which they would have access, were there no trial
13		Researchers must ensure that data collection procedures are adequately justified. The risks of data collection procedures must be minimized consistent with sound design and stand in reasonable relation to the knowledge to be gained
14	Protecting vulnerable participants	Clusters may contain vulnerable participants. In these circumstances, researchers and RECs must consider whether additional protections are needed
15		When individual informed consent is required, and there are individuals who may be less able to choose participation freely because of their position in a cluster or organizational hierarchy, RECs should pay special attention to recruitment, privacy, and consent procedures for those participants

Source: Weijer, C., Grimshaw, J.M., Eccles, M.P., McRae, A.D., White, A., Brehaut, J.C., et al., 2012. The Ottawa statement on the ethical design and conduct of cluster randomized trials. PloS Medicine 9 (11), e1001346. http://dx.doi.org/10.1371/journal.pmed.1001346.

approved by a REC (World Medical Association, 2008; CIOMS, 2002). Research may be defined as a systematic investigation designed to produce generalizable knowledge. CRTs generally meet the definition of research and must be reviewed by an REC (recommendation 2). All of the three examples meet the definition of research and hence should have been reviewed by RECs. RECs ought to take a proportional approach to the review of study protocols; thus, protocols not involving vulnerable participants and posing only low risk to research participants may be eligible to undergo expedited or delegated review. While there is a common perception that socio-behavioral research poses only low risk to research participants, examples of unintended harm to participants as a result of seemingly innocuous research exist (Phillips, 1994).

Identification of Research Participants

The identification of human research participants is an ethical requirement before consent procedures and harm–benefit issues can even be considered. Failing to accurately identify research participants in CRTs may have differing

consequences: on the one hand, being overly inclusive may lead to unnecessary regulatory burdens that can hamper important research; on the other hand, failing to identify someone as a research participant may lead to inadequate protections of their interests. A defining feature of a 'research participant' is someone whose interests are affected by study interventions or data collection procedures (McRae et al., 2011a). Based on this definition, the Ottawa Statement provides four criteria for identifying research participants in a CRT (recommendation 3). Individuals classified as research participants are entitled to regulatory protections, including assessment by an REC of study benefits and harms, informed consent provisions, and subject selection procedures.

In example 1 (COMMIT), the entire community was being intervened upon through the cluster-level intervention, and thus all residents were research participants. Although residents in control communities were not targeted by any study interventions, their interests may be affected by being denied access to a potentially beneficial intervention, and thus, they were also research participants. Community residents participating in the surveys were research participants as they interacted with investigators for the purpose of data collection. Birth attendants in example 2 were intervened upon through training sessions and reminders and completed questionnaires; thus, birth attendants were research participants. However, women giving birth were not research participants: although the women may have been indirectly affected by the interventions directed at the birth attendants, the birth attendants continued to have an obligation to act in their patients' best interests and the fidelity of the health provider–patient relationship was preserved. Since women were not directly intervened upon by researchers and did not provide researchers with identifiable private information for research purposes, the women were not research participants (McRae et al., 2011a). Employees in example 3 were research participants as they were directly intervened upon and interacted with researchers for the purpose of data collection.

Informed Consent Requirements

The ethical principle of respect for persons generally requires that researchers seek the informed consent of prospective research participants. The consent process allows potential participants to adopt the ends of the study as their own instead of being treated merely as a means to an end. In CRTs, consent is required only from those cluster members who are research participants (recommendation 4). When cluster members are not research participants, their informed consent for research is not required.

National and international research ethics guidelines provide detailed disclosure requirements for consent processes, including information about the purpose of the study, study interventions and data collection procedures, the potential benefits and risks of study participation, and alternatives to participation. In individually randomized trials, participants usually consent to randomization, intervention, and data collection prior to study enrollment. Because the units of allocation, intervention, and data collection may be different in CRTs, consent for these three aspects of the trial may need to be sought separately. For example, professionals may need to provide consent for study interventions, while individuals may need to consent to data collection.

In example 2, birth attendants were research participants and their informed consent was required for both study interventions and completion of questionnaires; however, the informed consent of patients was not required as they were not research participants. In example 3, employees were research participants, and required to consent to the study interventions as well as data collection procedures.

Timing of Consent

Recommendation 5 of the Ottawa Statement requires researchers to identify participants and seek their consent before cluster allocation. However, in many CRTs, it is not possible to approach or even identify eligible participants before allocation of clusters. In such cases, researchers should seek potential participants' consent as soon as possible after the participant has been identified, before administering any study interventions or data collection procedures. Seeking informed consent as soon as possible and before any study interventions or data collection procedures satisfies the moral purpose of informed consent (McRae et al., 2011b). In these circumstances, consent to randomization is not required. However, researchers should be aware that post-randomization consent increases the risks of selection bias (Eldridge et al., 2009; Hahn et al., 2005). In example 3, it would have been both ethically and methodologically advantageous to identify and obtain employees' informed consent prior to randomization, if this were feasible. When investigators believe that practical or logistical issues mean that obtaining such consent is infeasible, they must justify this to the satisfaction of the responsible REC and obtain individual consent as soon as feasible and prior to the study interventions or data collection procedures.

Feasibility of Consent

In CRTs with cluster-level interventions (e.g., COMMIT), individual informed consent for study intervention is usually not feasible because there is no meaningful way for an individual to opt out or avoid the intervention. In such cases, researchers may apply for a waiver of consent for the study intervention, provided it presents no more than minimal risk to research participants (recommendation 6). Waivers of consent apply in some jurisdictions, and imply that the REC removes the requirement to obtain informed consent. 'Minimal risk' refers to the risks of daily life of the study population. Various study procedures are commonly regarded as presenting only minimal risk, including questionnaires and interviews covering nonsensitive topics (e.g., topics other than mental health, sexually transmissible diseases, or violent or criminal behavior), physical examination, noninvasive medical monitoring, and review of medical records (assuming adequate protections of confidentiality are in place). The researcher is responsible for adequately justifying to the REC that obtaining informed consent is infeasible and that study participation poses only minimal risk. However, a waiver of consent may not extend to all study procedures. For example, even if a waiver of consent is granted for study intervention,

researchers may nevertheless be required to seek participants' consent for data collection.

In some CRTs, seeking individual consent for data collection may be logistically impossible (e.g., due to extremely large cluster sizes), but data may nevertheless be available from secondary data sources. If informed consent to the use of individuals' identifiable information is not feasible, the researchers may apply for a waiver of individual consent on the grounds that the research could not practicably be carried out otherwise.

In some CRTs, researchers may be concerned that information disclosed during the informed consent process may lead to either selection or response bias. In such cases, they may appeal to an REC for a waiver or *alteration* of consent, which means that the REC permits changes to or removal of some of the standard elements of disclosure in the informed consent, for example, by allowing incomplete disclosure of the nature of a behavioral intervention. Alterations of consent procedures are subject to the same restrictions as waivers; that is, the risk of bias must make the study infeasible and study interventions must pose no more than minimal risk. In example 1 (COMMIT), cohort members were not explicitly notified of their status as cohort members, but were informed that they would be contacted annually. In example 3, employees were blinded as to the exact nature of the intervention in order to minimize the risks of participation or response bias. As far as possible, researchers should avoid having differential consent procedures in the intervention and control arms of the study to minimize the risks of bias (Eldridge et al., 2009; Hahn et al., 2005).

Informed Consent in Organizations

CRTs taking place in organizations or worksites may contain vulnerable cluster members whose presence is not obvious. For example, individuals in subordinate positions may be less able to freely choose whether they wish to participate in the trial. Researchers should take special care to identify and protect vulnerable participants in CRTs (recommendation 14) and pay special attention to their recruitment, privacy, and consent procedures (recommendation 15). When professionals (such as physicians, nurses, teachers, and managers) are research participants in CRTs, their informed consent may be required unless conditions for a waiver of consent obtain (recommendation 7). Recruitment of professionals in CRTs should be free of coercive influence from supervisors, administrators, or organization leaders. In such cases, consent procedures should include discussions about career-related risks (e.g., due to detection of possible negligence or incompetence). Data about professional performance should be kept confidential within the research team, except in cases of gross negligence or incompetence where statute or professional obligation mandates disclosure to a professional certifying or licensing body. This possibility should be disclosed to potential research participants in the informed consent process.

The Role of Cluster Gatekeepers

Gatekeepers are individuals or bodies who may be called upon to represent the interests of cluster members, clusters, or institutions. Examples of gatekeepers in CRTs include village elders, mayors, school principals, employers, or hospital administrators. Although gatekeepers have traditionally played an important role in CRTs due to challenges in seeking individual informed consent, such concerns may be mitigated by careful identification of research participants and application of waivers of consent (Gallo et al., 2012). Gatekeepers do not have the authority to provide proxy consent on behalf of cluster members, and CRTs should not proceed on the basis of such 'consent' (recommendation 8). However, gatekeepers do have an important role with respect to the protection of group or organizational interests affected by a CRT. As long as a gatekeeper's role within the cluster or organization endows them with the authority to make decisions on behalf of the cluster, and as long as cluster members recognize this authority, researchers may approach a gatekeeper for permission to enroll a cluster in a trial. A school principal, for example, may provide permission to researchers to access the school after considering availability of staff, resource implications of participation, and the likelihood that teachers or students would be willing to participate.

Gatekeepers may also facilitate consultation between researchers and cluster members about the goals, design, and implementation of the study, as well as dissemination of research findings. Consultation with cluster members may protect group interests by subjecting the study to examination and discussion by those whose interests may be affected. Gatekeepers may have the authority to withdraw a cluster from an ongoing CRT. As cluster withdrawals can have serious consequences for participants as well as the scientific validity of a CRT, researchers should make every reasonable effort to retain all enrolled clusters in the trial. In some trials, the intervention may be offered to clusters allocated to the control arm as a way to help secure their cooperation and prevent attrition.

In example 1 (COMMIT), the researchers recognized the need for community permission and consultation: the communities (via the community advisory boards) were closely involved in the planning and implementation of the trial. In example 2, hospital authorities legitimately provided permission to conduct the study in their facility and provided access to the anonymized data from individual births. They may also have facilitated consultation with birth attendants, doctors, and nurses about specific aspects of the trial, but this was not a substitute for the informed consent of birth attendants. In example 3, employers legitimately had a role to protect the organizational interests, and provided permission to include sites in the trial (but not to enroll individual employees in the trial).

Conclusion

CRTs are clearly distinguishable from individually randomized trials; they have unique methodological and ethical requirements. Once the decision to use cluster randomization has been adequately justified, investigators must adopt statistical methods appropriate for the CRT design. CRTs may involve multiple kinds of participants, not all of whom will meet the criteria for research participants: informed consent is not required from those who are not research participants, while waivers of consent may obtain in many other cases. Gatekeepers cannot legitimately provide consent for randomization or study procedures on behalf of cluster members and have

a limited role in protecting individual interests. They do in some cases have a legitimate role in protecting group or institutional interests.

Prior to publication of the Ottawa Statement, there were no comprehensive ethics guidelines specific to CRTs; as a result, researchers and RECs had to rely on ambiguous interpretation of standard research ethics guidelines designed with individually randomized trials in mind. This contributed to considerable uncertainty and variability in the conduct and ethics review of CRTs within and among different jurisdictions. Here we interpreted key recommendations in the Ottawa Statement with respect to CRTs in health, social, and behavioral research. There is increasing interest in the use of CRTs in education, welfare, crime prevention, and economic development (Boruch, 2005). Future research may focus on the application of the Ottawa Statement to CRTs in these areas. Finally, given the rapidly expanding nature of the field, the authors of the Ottawa Statement anticipate the need for significant revisions and additions to the guidelines over the next 5 years.

See also: Common Rule: US Regulatory Protections for Human Research Participants; Ethics for Biomedical Research Involving Humans: International Guidelines; Informed Consent of Research Subjects; Privacy of Individuals in Social Research: Confidentiality; Psychiatry: Informed Consent and Care; Research Ethics, Cross-Cultural Dimensions of; Research Ethics: Research; Research Publication: Ethical Aspects.

Bibliography

Althabe, F., Buekens, P., Bergel, E., Belizán, J.M., Campbell, M.K., Moss, N., Hartwell, T., Wright, L.L., Guidelines Trial Group, 2008. A behavioral intervention to improve obstetrical care. New England Journal of Medicine 358 (18), 1929–1940.

Benatar, S., 2004. Towards progress in resolving dilemmas in international research ethics. The Journal of Law, Medicine & Ethics 32 (4), 574–582.

Binik, A., Weijer, C., McRae, A., Grimshaw, J.M., Boruch, R., Brehaut, J.C., Donner, A., Eccles, M.P., Saginur, R., Taljaard, M., Zwarenstein, M., 2011. Does clinical equipoise apply to cluster randomized trials? Trials 12, 118.

Boruch, R., 2005. Preface: better evaluation for evidence-based policy: place randomized trials in education, criminology, welfare, and health. In: Annals of the American Academy of Political and Social Science. Place Randomized Trials: Experimental Tests of Public Policy, vol. 599. Sage Publications, Inc., pp. 6–18.

Boruch, R., May, H., Turner, H., Lavenberg, J., Petrosino, A., de Moya, D., Grimshaw, J., Foley, E., 2004. Estimating the effects of interventions that are deployed in many places: place randomized trials. American Behavioral Scientist 47 (5), 608–633.

Campbell, M.J., Walters, S.J., 2014. How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research. Wiley, Chichester.

Campbell, M.K., Piaggio, G., Elbourne, D.R., Altman, D.G., 4 September 2012. CONSORT 2010 statement: extension to cluster randomised trials for the CONSORT Group. British Medical Journal 345, e5661.

Chaudhry, S.H., Brehaut, J.C., Grimshaw, J.M., Weijer, C., Boruch, R., Donner, A., Eccles, M.P., McRae, A.D., Saginur, R., Skea, Z.A., Zwarenstein, M., Taljaard, M., 2013. Challenges in the research ethics review of cluster randomized trials: international survey of investigators. Clinical Trials 10 (2), 257–268.

Council of International Organizations of Medical Science (CIOMS), 2002. International Ethical Guidelines for Biomedical Research Involving Human Subjects. Geneva, Switzerland. http://www.cioms.ch/publications/layout_guide2002.pdf (accessed 14.07.14.).

Donner, A., Klar, N., 2000. Design and Analysis of Cluster Randomization Trials in Health Research. Arnold, London.

Edwards, S.J.L., Braunholtz, D.A., Lilford, R.J., Stevens, A.J., 1999. Ethical issues in the design and conduct of cluster randomised controlled trials. British Medical Journal 318 (7195), 1407–1409.

Eldridge, S., Kerry, S., Torgerson, D.J., 2009. Bias in identifying and recruiting participants in cluster randomised trials: what can be done? British Medical Journal 339.

Eldridge, S.M., Ashby, D., Feder, G.S., 2005. Informed patient consent to participation in cluster randomized trials: an empirical exploration of trials in primary care. Clinical Trials 2, 91–98.

Gallo, A., Weijer, C., White, A., Grimshaw, J.M., Boruch, R., Brehaut, J.C., Donner, A., Eccles, M.P., McRae, A.D., Saginur, R., Zwarenstein, M., Taljaard, M., 2012. What is the role and authority of gatekeepers in cluster randomized trials in health research? Trials 13, 116.

Hahn, S., Puffer, S., Torgerson, D.J., Watson, J., 2005. Methodological bias in cluster randomised trials. BMC Medical Research Methodology 5, 10.

Hoeyer, K., Dahlager, L., Lynöe, N., 2005. Conflicting notions of research ethics: the mutually challenging traditions of social scientists and medical researchers. Social Science & Medicine 61, 1741–1749.

Hutton, J.L., 2001. Are distinctive ethical principles required for cluster randomized controlled trials? Statistics in Medicine 20 (3), 473–488.

Ivers, N., Taljaard, M., Dixon, S., Bennett, C., McRae, A., Taleban, J., Skea, Z., Brehaut, J.C., Boruch, R.F., Eccles, M.P., Grimshaw, J.M., Weijer, C., Zwarenstein, M., Donner, A., 2011. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000–08. British Medical Journal 343, d5886.

Kirkwood, B.R., Hurt, L., Amenga-Etego, S., Tawiah, C., Zandoh, C., Danso, S., Hurt, C., Edmond, K., Hill, Z., Ten Asbroek, G., Fenty, J., Owusu-Agyei, S., Campbell, O., Arthur, P., ObaapaVitA Trial Team, 8 May 2010. Effect of vitamin A supplementation in women of reproductive age on maternal survival in Ghana (ObaapaVitA): a cluster-randomised, placebo-controlled trial. Lancet 375 (9726), 1640–1649.

Klar, N., Donner, A., 2007. Ethical challenges posed by cluster randomization. In: Wiley Encyclopedia of Clinical Trials, pp. 1–5.

McRae, A., Bennett, C., Weijer, C., Brehaut, J., Belle Brown, J., Donner, A., Eccles, M.P., Grimshaw, J.M., Zwarenstein, M., Chaudhry, S., Taljaard, M., 2013a. Researchers' perceptions of ethical challenges in cluster randomized trials: a qualitative analysis. Trials 14, 1.

McRae, A.D., Bennett, C., Skea, Z., Boruch, R., Brehaut, J.C., Donner, A., Eccles, M.P., Grimshaw, J.M., Weijer, C., Taljaard, M., 2013b. Reporting of patient consent in healthcare cluster randomized trials is associated with the type of study interventions and publication characteristics. Journal of Medical Ethics 39, 119–124.

McRae, A.D., Weijer, C., Binik, A., White, A., Grimshaw, J.M., Boruch, R., Brehaut, J.C., Donner, A., Eccles, M.P., Saginur, R., Zwarenstein, M., Taljaard, M., 2011a. Who is the research subject in cluster randomized trials in health research? Trials 12, 183.

McRae, D., Weijer, C., Binik, A., Grimshaw, J.M., Boruch, R., Brehaut, J.C., Donner, A., Eccles, M.P., Saginur, R., White, A., Taljaard, M., 2011b. When is informed consent required in cluster randomized trials in health research? Trials 12, 202.

Murray, D.M., 1998. Design and Analysis of Group-Randomized Trials. Oxford University Press, New York.

Parker, S.W., Teruel, G.M., 2005. Randomization and social program evaluation: the case of progresa. Annals of the American Academy of Political and Social Science 599, 199–219.

Phillips, S.R., 1994. Asking the sensitive question: the ethics of survey research and teen sex. IRB: Ethics and Human Research 16 (6), 1–7.

Robroek, S.J., Polinder, S., Bredt, F.J., Burdorf, A., 2012. Cost-effectiveness of a long-term Internet-delivered worksite health promotion programme on physical activity and nutrition: a cluster randomized controlled trial. Health Education Research 27 (3), 399–410. http://dx.doi.org/10.1093/her/cys015.

Taljaard, M., Brehaut, J.C., Weijer, C., Boruch, R., Donner, A., Eccles, M.P., McRae, A.D., Saginur, R., Zwarenstein, M., Grimshaw, J.M., 2014a. Variability in research ethics review of cluster randomized trials: scenario-based survey in three countries. Trials 15, 48. http://dx.doi.org/10.1186/1745-6215-15-48.

Taljaard, M., Chaudhry, S.H., Brehaut, J.C., Weijer, C., Boruch, R., Donner, A., Eccles, M.P., McRae, A.D., Saginur, R., Zwarenstein, M., Grimshaw, J.M., 2014b. Survey of consent practices in cluster randomized trials: improvements are needed in ethical conduct and reporting. Clinical Trials 11 (1), 60–69.

Taljaard, M., McRae, A., Weijer, C., Bennett, C., Dixon, S., Taleban, J., Skea, Z., Brehaut, J., Eccles, M.P., Donner, A., Saginur, R., Boruch, R.F., Grimshaw, J.M., 2011. Inadequate reporting of research ethics review and informed consent in cluster randomized trials: review of a representative sample of published trials. British Medical Journal 342, d2496.

Taljaard, M., Weijer, C., Grimshaw, J.M., Belle Brown, J., Binik, A., Boruch, R., Brehaut, J.C., Chaudhry, S.H., Eccles, M.P., McRae, A., Saginur, R., Zwarenstein, M., Donner, A., 2009. Ethical and policy issues in cluster randomized trials: rationale and design of a mixed methods research study. Trials 10, 61.

Taljaard, M., Weijer, C., Grimshaw, J.M., Eccles, M.P., The Ottawa Ethics of Cluster Randomised Trials Consensus Group, 2013. The Ottawa statement on the ethical design and conduct of cluster randomised trials: a précis for researchers and research ethics committees. British Medical Journal 346, f2838. http://dx.doi.org/10.1136/bmj.f2838.

The COMMIT Group, 1995a. Community intervention trial for smoking cessation (COMMIT): II. Changes in adult cigarette smoking prevalence. American Journal of Public Health 85 (2), 193–200.

The COMMIT Group, 1995b. Community intervention trial for smoking cessation (COMMIT): I. Cohort results from a four-year community intervention. American Journal of Public Health 85 (2), 183–192.

Weijer, C., Grimshaw, J.M., Eccles, M.P., McRae, A.D., White, A., Brehaut, J.C., Taljaard, M., The Ottawa Ethics of Cluster Trials Consensus Group, 2012. The Ottawa statement on the ethical design and conduct of cluster randomized trials. PloS Medicine 9 (11), e1001346. http://dx.doi.org/10.1371/journal.pmed.1001346.

Weijer, C., Grimshaw, J.M., Taljaard, M., Binik, A., Boruch, R., Brehaut, J.C., Donner, A., Eccles, M.P., Gallo, A., McRae, A.D., Saginur, R., Zwarenstein, M., 2011. Ethical issues posed by cluster randomized trials in health research. Trials 12, 100.

World Medical Association, 2008. Declaration of Helsinki. http://www.wma.net/en/30publications/10policies/b3/index.html (accessed 14.07.14.).

Clustering and Cohesion in Networks: Concepts and Measures

James Moody and Jonathan Coleman, Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social networks are not homogeneous but typically grouped into subsets of strongly reconnected groups. Here we review the literature on structural cohesion and clustering in networks. We divide our review into sections based on overall measures of cohesion and approaches to finding subgroups in larger networks.

Introduction

Network cohesion and clustering are important for understanding how social networks shape communities, facilitate norm maintenance, or form the basis of categorical group identity, among many other things (Freeman, 1992; Pescosolido and Rubin, 2000; Martin, 2009). Substantively, we expect that norms and diffusion circulate more readily within cohesive networks, and that clusters of people connected to each other – peer groups or crowds – are likely to be similar. If we seek to measure cohesion and identify clusters in networks, then we must answer two related methodological questions. First, how best to operationalize cohesion on a network and second, how to identify the naturally occurring 'cohesive subgroups' that emerge in most social settings. This article reviews current work on the related problems of structural cohesion and clustering in social networks.

We understand social networks as graphs of social relations, G(V,E), where the nodes (V) of the graph are typically people and the relations (E) are some positive social connection, such as friendship, love, or communication, linking pairs of nodes (for the substantive problem of cohesion, relations should be thought of as carrying positive social meaning rather than mere contact or negative relations. See Friedkin (2004) for further reflections on the relations between individual and group cohesion). Relations can be directed and valued, although most of the discussion below will focus on binary undirected networks for simplicity. A path in the network is a (possibly directed) sequence of nodes and edges starting and ending with a node but never repeating. Two nodes are said to be reachable if there is a path from one node to the other, and the graph is connected if all pairs are reachable. Two paths are node (edge) independent if they share the same start and end nodes but overlap on no other nodes (edges). The geodesic distance is the shortest path connecting two nodes. If there is no path connecting a pair, we define the distance as infinite. A cycle is a path that starts and ends on the same node. The density of a graph is the average value of edges calculated over all pairs: $(\sum_{ij} A_{ij})/[N(N-1)]$.

Structural Cohesion

Cohesive networks should be difficult to separate. For cohesive networks, the relations should bind the collective together. In such settings, we expect that people have many ties with others and that the ties are widely distributed (rather than routing through one node). In cohesive networks, people are generally close to each other (Moody and White, 2003; Friedkin, 2004).

The archetypical structural pattern capturing these intuitive cohesion ideas is the clique (Luce and Perry, 1949; Alba, 1973): a collection of nodes with edges connecting every pair in the network. Every node in a clique of size n is connected to all $n-1$ other nodes. The complete nature of a clique leads to maximal values on other features, which have typically been used as indicators of cohesion or clustering. For example, the density of a binary graph is 1, the distance between all nodes is 1, and every pair has $n-2$ contacts in common, implying that all triads within the graph are complete. Disconnecting a clique requires deleting at least $n-1$ edges or nodes.

The complete adjacency requirement of cliques means they are substantively difficult to use in real-world data settings where some element of randomness is usually present. First, known substantive groups are rarely complete cliques, as any missing edge undermines the complete connectivity requirement. (Moreover, if a survey design for collecting network data has limited the respondent to nominating k alters, then the maximal size of clique in the network must be k.) Second, since cliques can overlap (by as many as $k-1$ nodes), attempts to use cliques as markers for cohesion frequently must aggregate overlapping clique memberships (Palla et al., 2005; Evert and Borgatti, 1998). This leads researchers to search for groups that are 'clique like,' but not complete. Understanding exactly what counts as 'clique like' has been the heart of the methodological challenge in this area, but most have focused on relaxing one of the dimensions that defines a clique, effectively capturing how 'close' to a pure clique a given graph might be.

The simplest (because it involves only summing edges) such generalization has been to use the graph's density. While intuitive (cohesive groups should be dense), density is an insufficient measure of cohesion. Consider Figure 1, which shows three networks with the same density, but differing path structures, and thus substantively different cohesion (for how density nonetheless constrains structure, see Faust, 2006) This general sort of problem, that networks have the same score on a given clique-maximal characteristic nonetheless have different structures that clearly imply varying cohesion, is at the heart of the difficulty in generalizing clique features to measure cohesion. This is true for many attempts at measuring cohesion, such as the distance between nodes in a group (N-Clique, Bron and Kerbosh, 1973), the average number of nodes each node is connected to (the k-plex or k-core, Seidman and Foster,

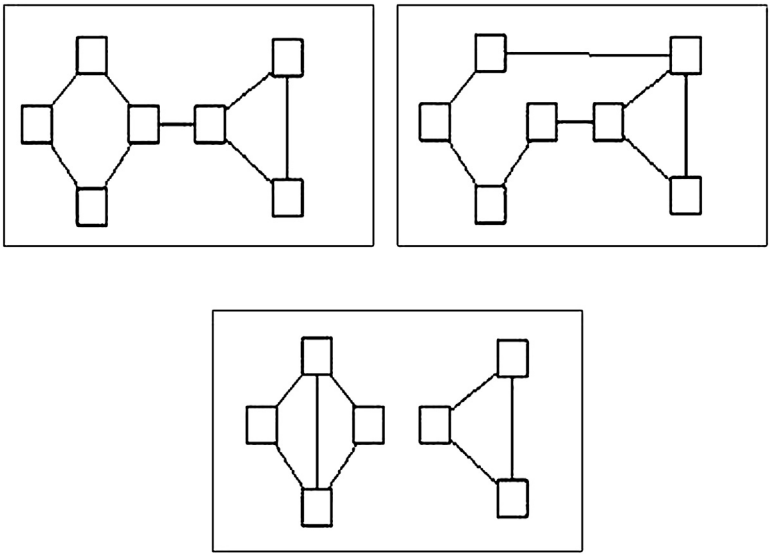

Figure 1 Three graphs with same density but different cohesion.

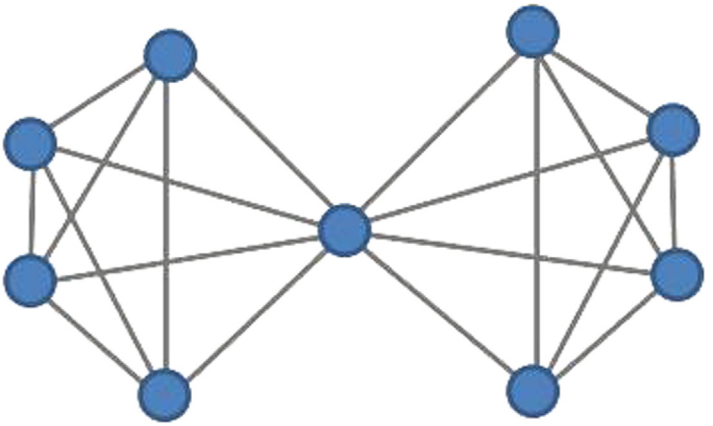

Figure 2 Graph with high edge connectivity but low node connectivity.

1978), or the proportion of transitive triples (such as the clustering coefficient (Watts and Strogatz, 1998). At the heart of the problem is the understanding that cohesion rests on levels of connectivity – on the pattern of paths as much as the volume of ties.

In response to the ambiguity of these many measures, Moody and White (2003) introduced node connectivity as a necessary measure of structural cohesion. Starting with the simple idea that cohesion must at a minimum imply connectivity, Moody and White argued for generalizing connectivity based on the difficulty of disconnecting a network. A graph can be disconnected by removing either nodes or edges. Edge connectivity is the minimum number of edges that one has to remove to disconnect the graph, while node connectivity is the minimum number of vertices one has to remove to disconnect the graph. While these might seem similar, it turns out that one can easily have high edge connectivity even if node connectivity is low, while the opposite is not true: high node connectivity implies high edge connectivity. As an illustration, consider Figure 2. Here the graph is four-edge connected but one-node connected, meaning that graph resilience is dependent on the center star-like cut-node. Since social cohesion turns on the supraindividual character of the setting (Simmel, 1950), node connectivity better characterizes the intuitive notion that a cohesive group is held together but not dominated by a single node, and thus Moody and White take biconnectivity (two-node-separable graphs) as the fundamental starting point for structural cohesion.

Node connectivity implies other features that similarly map well to our understanding of social cohesion. First, a k-connected component implies that every pair of nodes is connected by at least k-node-independent paths and implies $(k-1)$ node-independent cycles among all pairs in the graph. This means that k-components (for $k > 1$) contain multiple paths between pairs that never cross the same set of nodes. For social networks, the existence of multiple paths implies many alternative routes through the network that could generate high levels of information diffusion (see Centola and Macy, 2007

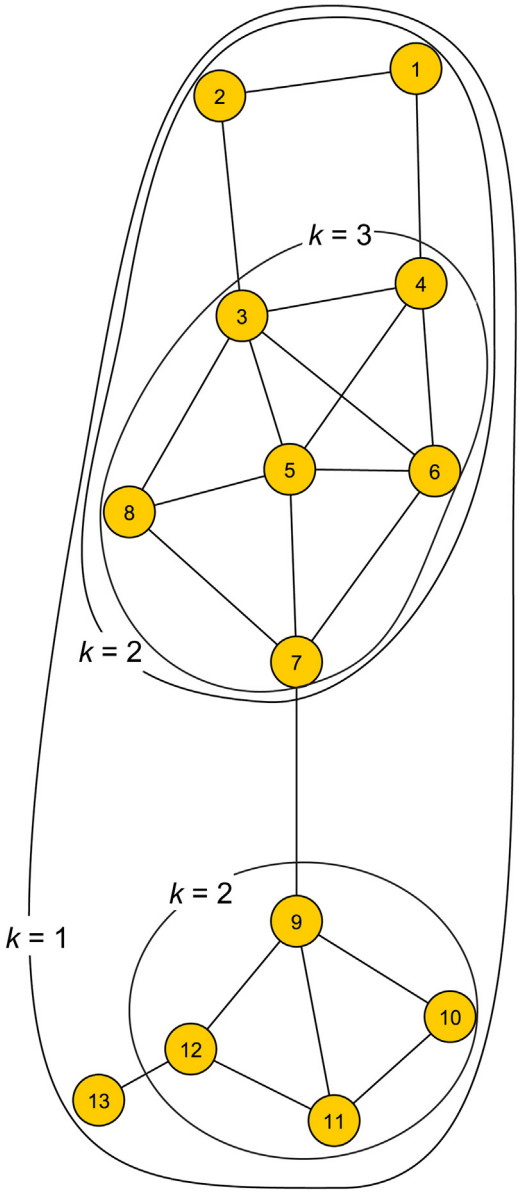

Figure 3 Nested *k*-components of a network.

for related ideas with respect to diffusion) and promote the conditions necessary for enforceable trust, a key element for norm formation (Mark, 1998).

Node connectivity implies a hierarchical ordering of the groups in the network, since higher valued *k*-components are nested within lower valued *k*-components. For example, all nodes in Figure 3 are members of a (1-)component. Nested within this are two bicomponents (sets {1–8}, {9–12}) and nested within the largest bicomponent is a three-component ({3–8}). The full enumeration of the nested node components provides a cohesive blocking of the network and is a complete summary of the structural cohesion of the graph. The depth of involvement within this hierarchical structure is a natural operationalization of network embeddedness (Grannovetter, 1985; Uzzi, 1999), and thus the highest value *k*-component within which any given node is nested is recommended. Moments, such as the mean, of the distribution of pairwise joint maximal cohesion provide simple global summaries, although the minimum is the most theoretically consistent measure of overall network cohesion (the network is as strong as the weakest link).

Relative Density Measures

Moving from the network as a whole to subgroups within a network, many recognize that ties within a group typically come at the expense of ties between groups (Freeman, 1992). Thus, we expect that group (we follow the sociological convention and refer to these collections of nodes as 'groups,' the other common terms used in statistical physics or biology include 'communities' or 'modules', respectively) members interact with each other more often than with others, and measures based on this balance between intra- and intergroup ties are common. Theoretically, we should distinguish the relations within the group (what is best thought of as cohesion per se) from the structural distinctiveness of the group – the level of boundary crossing represented by between-group ties. (To give an intuitive sense of the difference, imagine two systems with multiple clusters but no ties between the clusters. In that case the relative density measures would be the same for any internal cluster density above zero.) Node cohesion implies a particular path structure within groups (*k*-node-independent paths between all pairs) and maximal contact between groups (cannot be more than $k-1$). Building on this basic balance of within versus between-group ties, a natural measure for how grouped a network is would be a relative score for the ties within compared to the ties between. One can easily summarize the group-level tie counts with a mixing matrix, where each row and column represents a group, and cell values are the count of relations within/between groups. Relative density measures are some function of the values on the diagonal compared to the values off the diagonal. An example is given in Figure 4, with the groups indicated by the shaded ellipses under the network. The mixing matrix cells provide tie counts (top number in each cell) and expected values under independence (bottom).

There had been some debate over the optimal function to weight within and between-group ties. Freeman (1978) introduced the segregation index, as $(E-O)/E$, where E is the number of expected cross-group ties and O is the number of observed cross-group ties. The logic of this index is that segregation implies difference from random mixing: If ties were distributed randomly with respect to the group labels, we would find as many observed as expected, and the score would be zero (completely nonsegregated). If, on the other hand, groups were strongly bounded we would observe very few cross-group ties, and the score would approach one. Thus the measure interpolates between negative values (more cross-group ties than within-group ties), random mixing (0), and strongly bounded groups (1.0). Similar scores simply take the ratio of diagonal cells to off-diagonal or the odds ratio of a within-group tie (overall or group specific, see Frank, 1995). Each of these scores can be calculated for the network as a whole or for specific groups.

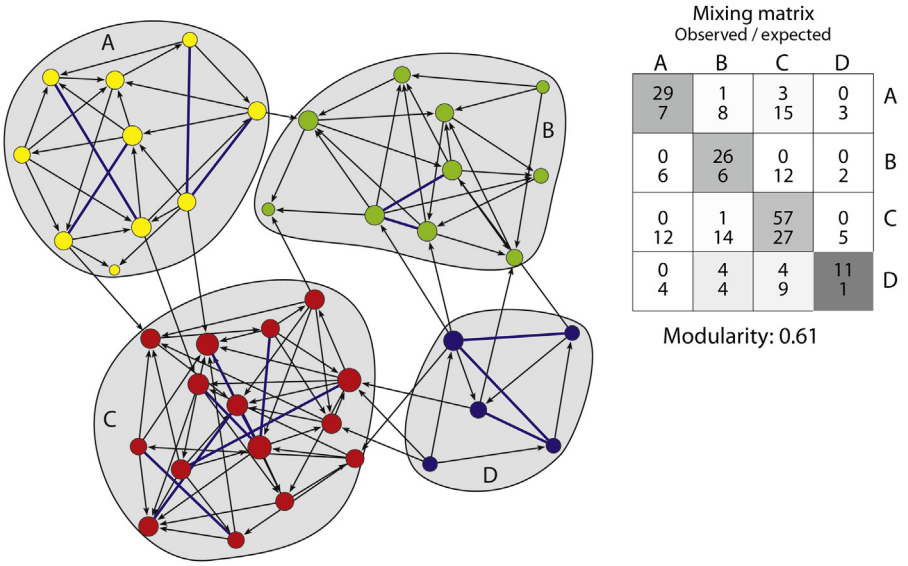

Figure 4 Example of subgroups in a network and mixing table used for relative density measures.

Of late, the most common relative density measure is the modularity index (Newman and Girvan, 2004). The modularity index is calculated as

$$Q = \sum_{ij} \frac{1}{2m}\left(A_{ij} - \gamma\frac{k_i k_j}{2m}\right)\delta(C_i C_j)$$

where m is the number of edges, k is the degree, A is the adjacency matrix, δ is an indicator for whether node i and j are in the same subgroup or not, and γ is a 'resolution parameter' that identifies the scale at which clustering is observed (Fortunato and Barthélemy, 2007; Reichardt and Bornholdt, 2006). Substantively, $[k_i k_j/2m]$ represents the null model – the expected likelihood of contact between two nodes – here simply dependent on the degrees of the pair, so $\left(\mathbf{A}_{ij} - \left[\gamma\frac{k_i k_j}{2m}\right]\right)$ is the connectivity above random expectation, normalized by the total volume of ties in the network. This normalization is key, as it ensures that the index takes a maximum value if all ties fall within separate groups. This is the same logic as Freeman's segregation index, but because Q takes a value of 0 if there is only one group, the function must have a maximum value for multiple group partitions, which makes the quantity a very useful guide for identifying unknown communities in networks (see Porter et al., 2009 for review).

Three features of the modularity score are important to note. First, since the function can be searched for a maximum value, it had been seen as a solution to the problem of identifying the number of cohesive groups in a network. Identifying the number of naturally occurring clusters in a dataset is a long-standing problem, and for networks, most prior work required the user to specify the number of groups, or relied on principled but known-to-be arbitrary stopping criteria (Moody, 2001; Frank, 1995). Since modularity captures how grouped a network is and has a clear maximum, people had assumed that maximizing modularity automatically revealed the number of groups in the data. Unfortunately, the discovery of the resolution parameter issue (Fortunato and Barthélemy, 2007) has undone this sense of natural maximization, since changing the resolution parameter will change the number of groups that maximize Q for any given network, reintroducing an arbitrary parameter that controls the number of clusters in the group. (This is a common problem, not unique to finding groups in networks. Any stopping/selection criteria for creating distinctions in a continuum would be similar. For example, the notion that principle components extracted from a factor analysis are retained when the eigenvalue is greater than 1.0 is similarly arbitrary (Kim, 1978). The advantage of principled-but-arbitrary selection rules is that they provide a basis for consistency across studies, rather than leaving it entirely at the discretion of investigators.) Since a resolution parameter of one (i.e., ignoring the parameter) is arbitrary, the notion that the score determines the 'natural' number of found groups is not likely true.

Second, having a high modularly score only ensures that the groups as observed are distinct, not that they are internally cohesive. The modularity score will always increase if disconnected groups are assigned to distinct communities, since the score rests on the number of intragroup ties. This means that any assignment to groups that truly maximizes the modularity (see point three below) should identify internally connected groups. Beyond this simple connectivity requirement, however, the score is based merely on volume and does not ensure any formal limit on the group's internal structure.

Third, while the modularity score can be maximized, it is not necessary that the maximum modularity value correspond to a unique assignment of nodes to groups. That is, there may be many partitions on a network that all maximize modularity, but imply different substantive assignments (Bagrow and Bollt, 2005; Clauset, 2005; Good, et al., 2010). While such assignments may overlap significantly, lack of a one-to-one correspondence between maximum modularity

and group assignment has further lessened the extent to which one can depend on modularity in defining cohesive groups.

Cohesive Group Detection

While modularity is a general index for the relative volume of ties within and between specified groups, an obvious related question turns on how one identifies cohesive subgroups in a network. The field of approaches for identifying cohesive groups – called 'community detection' within the computer science and statistical physics literature – is growing rapidly. For a full review see Porter et al. (2009) or Schaeffer (2007). Here, we review the history of subgroup search procedures and outline a few of the more popular approaches. We divide the field by a basic distinction between deterministic, often graph-theoretic based, approaches and heuristic approaches that often include some level of randomization.

Deterministic Group Finding Approaches

The simplest approaches are the direct extensions of clique models discussed above. A common approach is to identify k-cores – maximal subgraphs where all nodes are connected to at least k other nodes in the set (Siedman, 1983). While every k-node-connected group is a k-core, not all k-cores are k-node connected (Moody and White, 2003; White and Harary, 2001) as such, k-cores are computationally efficient, but ambiguous. In practice, this efficiency can be useful when trying to find k-connected groups in large networks (Powell et al., 2005).

An obvious deterministic group-finding approach is to start with cliques, but real-world networks tend to have many small overlapping cliques that do not obviously represent groups. One approach is to count the number of shared cliques between nodes and treat this as a new higher order valued matrix, which can then be clustered using standard cluster or factor analysis tools (Everett and Borgatti, 1998). A nice generalization of this approach can be found with the clique percolation method (CPM) (Palla et al., 2005). CPM is a deterministic group-finding algorithm, which allows groups to overlap in a very limited way. The CPM starts with cliques of size k (k-cliques) and defines two k-cliques as adjacent if they overlap by $k-1$ nodes. A community is then the maximal union of k-cliques that can be reached through adjacent k-cliques. The communities so identified can overlap, but by no more than $k-2$ nodes (Derenyi et al., 2005), and becomes difficult to use in large networks with high values of k.

A second deterministic approach focuses on cutting the graph in careful ways. Early work used the min-cut, max-flow features to break the network into edge-connected sets (see, for example, Borgatti et al., 1990). The most popular recent work in this area divides the network removing key edges until the graph is no longer connected (Girvan and Newman, 2002). The trick to this method is to calculate 'edge-betweenness' as the number of times the shortest path between any pair of nodes passes through a given edge. One then deletes the edge with the highest score, recomputes the edge-betweenness score for the remaining edges, and repeats until the graph is disconnected. This is the first cut of the graph. One then repeats the method on the resulting components. The algorithm is computationally intensive, but seems to perform quite well (note the procedure is theoretically deterministic, but ties in the edge-betweenness level will require a tie-breaking rule, which is often random).

A third sort of deterministic clustering rests on functions of the underlying eigenstructure of the network. This has emerged repeatedly, variously seen as applications of principal component analysis (Cairns and Cairns, 1994), peer influence models (Moody, 2001), or spectral clustering (Newman, 2006). Early social network researchers used principal components or factor analysis to identify groups from interaction (e.g., Wright and Evitts, 1961) or nomination matrices (e.g., Bock and Husain, 1952; MacRae, 1960; Bagwell et al., 2000). The approaches share the premise that groups can be conceptualized as individuals whose friendships are similar (i.e., correlated). A common practice is to run factor analysis tools on the observed network, then retain all factors with eigenvalues greater than 1.0 and factor loadings above a predetermined threshold (see Gest et al., 2007 for comparison).

This approach has clear links to the block-modeling traditions rooted in CONCOR (The CONvergence of iterated CORrelations, White et al., 1976), where actors are classified as similar if they have similar nomination patterns to/from others in the network. At their core, these models share a use of eigenvalues (or close approximations of eigenvalues) as the key information source to identify the partition. Such models work because the eigenstructure captures an underlying random-walk process that maps well onto diffusion ideas, which are captured by cycles and redundancy in the network. As such, these models are likely well suited to link groups to behavior, as the underlying model is concordant between the two features of groups.

Search/Optimization Approaches

Another general approach to finding groups in networks involves searching over possible partitions to optimize some function that identifies the best assignment of nodes to cohesive groups. In theory, one could imagine assigning nodes to groups exhaustively: trying all possible assignments until the best possible solution is found. Since such exhaustive search is computationally prohibitive for any real-world sized networks, we instead use heuristic approaches that aim to efficiently approximate an ideal assignment of nodes to groups. In general, these approaches must balance between computation costs, accuracy, and arbitrariness. For example, a simple way to reduce computation cost is to have users specify the number of groups, but that is often an arbitrary choice. One can also reduce computation by using graph approximations, which sacrifices accuracy. Since most of the metrics used to optimize heuristics are based on some version of a relative density score (modularity, Freeman's segregation index, odds ratio), all the problems associated with those scores carry over to heuristics designed to optimize them.

Direct optimization routines seek to sort nodes into cohesive groups in a manner that maximizes some clustering objective function. The Newman (2004) greedy method is one such agglomerative technique that consecutively combines nodes to form large groups. Starting with each node in its own group, the routine joins groups together in pairs, choosing the

pairing that maximizes the increase in modularity. The Louvain method, implemented by Blondel et al. (2008), is an extension of this idea, locally optimizing modularity then replacing communities with 'supernodes' and links between communities as weighted ties, yielding a smaller network. This iterates until modularity is maximized, leading to a hierarchical nesting of nodes.

In a similar vein on a different metric, Ken Frank's Klique-Finder (Frank, 1995) optimizes the odds of an in-group tie directly. The objective function can then be represented as a logistic regression model of the form

$$\log\left(\frac{p(A_{ij} = 1)}{p(A_{ij} = 0)}\right) = \theta_0 + \theta_1 \text{ samegroup}_{ij}$$

So the task is to identify the group assignment labels in a way that maximizes the mixing parameter θ_1. The procedure for making this assignment involves identifying a seed triad, adding nodes to the group until no improvement is made, then starting with a new seed, and repeating this assignment process until all nodes are in initial groups. A similar approach is used with latent space group models (Krivitsky et al., 2009).

In many of these approaches, initial decisions and ties in clustering values can create assignments that are locally optimal but not globally optimal. A higher order pass through the data can then be used to move members between identified groups, merge groups, or split groups. A formal model for this sort of higher order mixing was proposed by Kernighan and Lin (1970) (see also Newman, 2006), but similar ideas are implemented in Frank's Kliquefinder and Moody's (2001) algorithm, among others.

A new development in group-finding tools extends the model to dynamic networks (Mucha et al., 2010). Here the task is to identify clusters based on the evolution of the network. The key insight builds on the random walk/diffusion notions embedded in eigenmodels discussed above, but applies the process to a graph containing two sorts of links: those connecting nodes to each other within time slices and those connected nodes to themselves across time slices. The result is a 'multislice' network, and modularity maximization tools can then be run over the entire compiled network.

Summary and Conclusion

Cohesion is a seeming simple idea that becomes more complicated once we attempt to operationalize it in real data. The problem is fundamental to the phenomena: Real-world data are ambiguous and messy, but analysts prefer simple (exhaustive and mutually exclusive) groups with strict graph-theoretic definitions or searches without arbitrary parameter choices. While we have made progress in operationalizing structural cohesion overall (Moody and White, 2003), the algorithm is complex (i.e., slow in large networks), and the resulting cohesive blocking does not lend itself to a simple scalar representation. Work on subgroup detection methods is developing rapidly, so it is too soon to conclude that line of reasoning. However, the fundamental challenge here is that we appear unlikely to identify a single detection method that has no arbitrary parameters and can identify real underlying groups with certainty. Instead, there is always the risk of multiple equivalent partitions and only locally optimal searches, with even these distinctions dependent to some degree on user-specified parameters. This implies that any search for groups in networks needs to be empirically informed by theory and knowledge of the setting, to help provide external validation for claimed group assignments. This is, then, a rich and open area for future research.

See also: Core–Periphery Patterns in Networks; Social Network Algorithms and Software; Social Networks, Algebra of; Structural Equivalence: Meaning and Measures.

Bibliography

Alba, R.D., 1973. A graph-theoretic definition of a sociometric clique. The Journal of Mathematical Sociology 3 (1), 113–126.

Bagrow, J.P., Bollt, E.M., 2005. A local method for detecting communities. Physical Review E 72 (4), 046108.

Bagwell, C.L., Coie, J.D., Terry, R.A., Lochman, J.E., 2000. Peer clique participation and social status in preadolescence. Merrill-Palmer Quarterly: Journal of Developmental Psychology 46 (2), 280–305.

Blondel, V.D., Jean-Loup, G., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, Volume 10, page 10008.

Bock, R.D., Husain, S.Z., 1952. Factors of the tele: a preliminary report. Sociometry 15 (3/4), 206–219.

Borgatti, S.P., Everett, M.G., Shirey, P.R., 1990. LS sets, lambda sets and other cohesive subsets. Social Networks 12 (4), 337–357.

Bron, C., Kerbosch, J., 1973. Algorithm 457: finding all cliques of an undirected graph. Communications of the ACM 16 (9), 575–577.

Cairns, R.B., Cairns, B.D., 1994. Lifelines and Risks: Pathways of Youth in Our Time. Cambridge University Press.

Centola, D.J., Macy, M., 2007. Complex contagions and the weakness of long ties. American Journal of Sociology.

Clauset, A., 2005. Finding local community structure in networks. Physical Review E 72 (2), 026132.

Derenyi, I., Palla, G., Vicsek, T., 2005. Clique percolation in random networks. Physical Review Letters 94 (16), 160202.

Everett, M.G., Borgatti, S.P., 1998. Analyzing clique overlap. Connections 21, 49–61.

Faust, K., 2006. Comparing social networks: size, density, and local structure. Metodološki Zvezki 3 (2), 185–216.

Fortunato, S., Barthélemy, M., 2007. Resolution limit in community detection. Proceedings of the National Academy of Sciences 104 (1), 36–41.

Frank, K.A., 1995. Identifying cohesive subgroups. Social Networks 17 (1), 27–56.

Freeman, L.C., 1978. Segregation in social networks. Sociological Methods and Research 6, 411–429.

Freeman, L.C., 1992. The sociological concept of 'group': an empirical test of two models. American Journal of Sociology 98 (1), 152–166.

Friedkin, N.E., 2004. Social cohesion. Annual Review of Sociology 30, 409–425.

Gest, S.D., Moody, J., Rulison, K.L., 2007. Density or distinction? The roles of data structure and group detection methods in describing adolescent peer groups. Journal of Social Structure 8 (1).

Girvan, M., Newman, M.E.J., 2002. Community structure in networks. Proceedings of the National Academy of Science 99, 7821–7826.

Good, B.H., de Montjoye, Y.-A., Clauset, A., 2010. The performance of modularity maximization in practical contexts. Physical Review E 81, 046106.

Granovetter, M., 1985. Economic action and social structure: the problem of embeddedness. American Journal of Sociology 91, 481–510.

Kernighan, B.W., Lin, S., 1970. An efficient heuristic procedure for partitioning graphs. Bell Systems Technical Journal 49, 291–307.

Kim, J.-O., Mueller, C.W., 1978. Introduction to factor analysis. SAGE.

Krivitsky, P., Handcock, M., Raftery, A., Hoff, P., 2009. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. Social Networks 31, 204–213.

Luce, R., Perry, A., 1949. A method of matrix analysis of group structure. Psychometrika 14 (2), 95–116.
MacRae, D., 1960. Direct factor analysis of sociometric data. Sociometry 23 (4), 360–371.
Mark, N., 1998. Beyond individual differences: social differentiation from first principles. American Sociological Review 63 (3), 309–330.
Martin, J.L., 2009. Social Structures. Princeton University Press.
Moody, J., White, D.R., 2003. Structural cohesion and embeddedness: a hierarchical concept of social groups. American Sociological Review 68 (1), 103–127.
Moody, J., 2001. Peer influence groups: identifying dense clusters in large networks. Social Networks 23 (4), 261–283.
Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., Onnela, J.-P., 2010. Community structure in time-dependent, multiscale and multiplex networks. Science 328, 876–878.
Newman, M.E.J., 2004. Detecting community structure in networks. The European Physical Journal B 38, 321–330.
Newman, M.E.J., 2006. Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America 103 (23), 8577–8582.
Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Physical Review E 69 (2), 026113.
Palla, G., Derenyi, I., Farkas, I., Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435 (7043), 814–818.
Pescosolido, B., Rubin, B.A., 2000. The web of group affiliations revisited. American Sociological Review 65, 52–76.
Porter, M.A., Onnela, J.-P., Mucha, P.J., 2009. Communities in networks. Notices of the American Mathematical Society 56, 1082–1166.
Powell, W.W., White, D.R., Koput, K.W., Owen-Smith, J., 2005. Network dynamics and field evolution: the growth of interorganizational collaboration in the life sciences. American Journal of Sociology 110 (4), 1132–1205.
Reichardt, J, Bornholdt, S., 2006a. Statistical mechanics of community detection. Physical Review E 74 (1), 016110.
Reichardt, J, Bornholdt, S., 2006b. When are networks truly modular? Physica D: Nonlinear Phenomena 224 (1–2), 20–26.
Schaeffer, S.E., 2007. Graph clustering. Computer Science Review 1, 27–64.
Seidman, S.B., 1983. Network structure and minimum degree. Social Networks 5 (3), 269–287.
Seidman, S.B., Foster, B.L., 1978. A graph-theoretic generalization of the clique concept. The Journal of Mathematical Sociology 6 (1), 139–154.
Simmel, G., 1950. In: Wolff, Kurt H. (Ed.), The Sociology of Georg Simmel. Free Press, New York.
Uzzi, B., 1999. Embeddedness in the making of financial capital: how social relations and networks benefit firms seeking financing. American Sociological Review 64 (4), 481–505.
Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. Nature 393 (6684), 440–442.
White, D.R., Harary, F., 2001. The cohesiveness of blocks in social networks: node connectivity and conditional density. Sociological Methodology 31, 305–359.
White, H.C., Boorman, S.A., Breiger, R.L., 1976. Social structure from multiple networks. I. Blockmodels of roles and positions. American Journal of Sociology, 730–780.
Wright, B., Evitts, M.S., 1961. Direct factor analysis in sociometry. Sociometry 24 (1), 82–98.

Co-constructivism in Educational Theory and Practice

Kurt Reusser, Institute of Education, University of Zurich, Zurich, Switzerland
Christine Pauli, Department of Education, University of Fribourg, Fribourg, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Though it is widely accepted that learning is a constructive process, it is also a social activity. People should not be considered as solo learners, but as interacting with each other and as being shaped through collaboration and discourse. The basis of personal development, therefore, is not isolated construction of knowledge, but rather its coconstruction in a socially shared, cultural space. While classrooms, consequently, should develop from aggregations of solo learners to communities engaged in dialogues and collaborative learning, the teachers' role extends from direct instruction to adaptive guidance and scaffolding of students.

Ever since Piaget's dynamically Kantian epistemology (Piaget, 1925, 1975), it has been widely accepted as a pervasive assumption that learning is a constructive process. In contrast to the epistemological assumption of empiricism that what we know is a direct reflection of ontological reality, learning is considered as an active construction of knowledge. Learners, as they strive to make sense of their world, do not passively receive stimulus information matching independent physical structures, but genuinely interpret their experience by (re)organizing their mental structures in successive cycles of assimilation and accommodation, while *interacting* with the physical and symbolic environment. According to Piaget and most of his successors in cognitive, developmental, and educational psychology, this process of adaptive and viable reality construction is enabled and constrained both by biologically grounded structures (the strength and scope of which, however, are not yet well known) and by the already existing prior knowledge (concepts, operative schemas, and structures) of the individual.

Even though the constructivist assumption makes some traditional problems in both psychology and education easier to solve, it also raises some new ones. An important problem is how we can think of achieving intersubjectivity while individually constructing meaning and knowledge. How can individuals who personally build their knowledge independently of each other come to the same or similar cognitive structures and common world views? How can we *share the* knowledge of our culture if people are conceived of as being solo learners, and socially isolated Robinson Crusoe figures?

The answer to this seminal question, challenging traditional (Western) epistemological constructivism which is foremost based on the concept of individual learning and learners, stems from symbolic interactionism (Mead, 1934) and from the explanatory framework of sociocultural theory (Vygotsky, 1962). It claims that the minds of human beings are fundamentally social products and that learning is a *social activity*. Learning and enculturation are not bounded by the individual brain or mind but are intrinsically social endeavors, embedded in a society and reflecting its knowledge, perspectives, and beliefs. People construct their knowledge, not only from direct interaction with the physical world and personal experience, but also from interacting with each other, including being told by others and by being shaped through social experience, collaboration, and discourse. The basis of personal development and enculturation, thus, is not the socially isolated construction of knowledge, but its coconstruction in a social and cultural space. Or, as Bruner puts it: "Most learning in most settings is a communal activity, a sharing of the culture. It is not just that the child must make his knowledge his own, but that he must make it his own in a community of those who share his sense of belonging to a culture" (Bruner, 1986: p. 86). Knowledge, from this perspective, is no longer seen as solely residing as a kind of possession in the head of each individual, but as being created and shared by communities, and distributed across individuals whose coordinated joint commitment to shared goals, mutual interactions, and continual (re)negotiation of meaning determine decisions, the formation of concepts, and the solution of problems.

Concept and Process of Coconstruction

No precise and widely accepted definition of the concept and process of coconstruction can be found in psychological or educational literature. What has been provided is very diverse and depends on the theoretical context in which it is embedded. Differences can be found with regard to at least three aspects:

1. the social type of discourse eligible to be called coconstructive: mother–child dialogue, peer interaction, teacher–student interaction, problem solving in pairs, learning in teams or learning communities, computer-supported collaborative work;
2. the psychopedagogical processes involved in productive coconstructive activity: productive dialogue such as exploratory talk and collective argumentation, collaborative negotiation after sociocognitive conflict or as a process of reciprocal sense-making, joint construction of a shared understanding, elaboration on mutual knowledge and ideas, giving and receiving help, instructional dialogue with a teacher or expert, tutoring and scaffolding;
3. the expected outcomes of collaboration: taken-as-shared individual vs. socially shared cognitions; convergence and intersubjectivity; academic task fulfillment, student motivation, and conceptual development; effects on skills in listening, discussion, disputation, and argumentation.

Common to most theoretical contexts of coconstructivism is the implication of some kind of social interaction, collaborative activity and, through joint patterns of awareness, of seeking

some sort of convergence, synthesis, intersubjectivity, or shared understanding, with language as the central tool and mediator for the negotiation of meaning. Theorists, moreover, largely converge in the adopted methodology of microgenetic analysis that has been used to examine the inherently fragile processes of coconstruction.

Piagetian Perspectives: The Role of Sociocognitive Conflict

In a (Neo-)Piagetian framework, genuine dialogue becomes possible and facilitates the individual cognitive construction of operational structures when children are able to take other persons' points of view into consideration (decentration) and when they are able to resolve socio-cognitive conflict. The early Piaget (2000) in particular regarded social interaction, – specifically, peer interaction – as a significant factor in individual cognitive development: "Criticism is born of discussion and discussion is only possible amongst equals" (2000: p. 409). According to later experimental studies carried out by coworkers of Piaget (e.g., Doise and Mugny, 1984; Perret-Clermont et al., 1991), social factors, such as the need to deal with conflicting perspectives in the context of pair or small group problem solving, can have a productive impact on cognitive behavior. In a Piagetian conservation task, for example, pupils more easily progressed to a subsequent level of development after having been confronted with contradictory judgments put forth by an adult or another child.

Vygotskian Perspectives of Coconstruction

In Vygotsky's cultural–historical and sociocultural view of the development of the 'higher mental functions' – viz., thinking, reasoning, and understanding – as a process of meaningful appropriation of culture, the interactive foundation of the cognitive is at the core of the developmental process. In contrast to Piaget's view, however, "the constructivist principle of the higher mental functions lies outside the individual – in psychological tools (such as 'language') and interpersonal relations" (Kozulin, 1998: p. 15). According to Vygotsky's claim that interpersonal interactions on a social plane serve as prototypes for intrapersonal processes, i.e., for functions to be internalized, coconstruction can be seen as (asymmetrical) adult–child interaction, or interaction between a child and a more advanced and capable peer or member of the society, in the 'zone of proximal development.' "What a child can do today in cooperation, tomorrow he will be able to do on his own" (Vygotsky, 1962: p. 87). The quality and development of higher order thinking is prepared by the coconstructive patterns and distinctive properties of social interaction encompassing guidance and contributions from experts or peers. Meaningful new learning emerges by embedding mental functions (like logical argumentation, proof, reflection, or problem solving) into specific forms of goal-directed interaction and dialogue, where more knowledgeable individuals adaptively tailor a task and provide cultural tools and resources in such a way that a child can successfully coperform it. The acquisition of a new concept or mental function becomes progressively more skillful as the child – in a process of apprenticeship and appropriation – learns to respond in gradually more sophisticated and personally more meaningful ways to the coconstructive, sense-mediating context of adult regulations, and eventually takes over responsibility for his or her own learning.

Perspective of Situated and Socially Shared Cognition

Situated learning theory views human cognition as being embedded in and inseparable from specific sociocultural contexts and its resources for sense-making. The goal of learning is to enter a community of practice and its culture, i.e., to learn, like an apprentice, to use cultural tools, artefacts, technologies, and rituals as practitioners and skilled members of society use them (Brown et al., 1989; Greeno et al., 1998; Rogoff, 1990; Lave and Wenger, 1991). As a process, learning takes place through the interaction and transaction between people and their environments. Coconstruction, from a situated cognition perspective, can be seen as involving two or more individuals who collaboratively construct a shared understanding, or a solution to a problem, which neither partner entirely and necessarily possesses beforehand (Chi, 1996). In a widely quoted definition proposed by Roschelle and Teasley (1995), "[c]ollaboration is a coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem" (p. 70).

At the heart of this concept of coconstruction lie two coexisting activities: solving the problem collaboratively, and constructing and maintaining a *joint problem space*. Both activities require constant negotiations and recreations of meaning, i.e., trying to find out what can reasonably be said about the task in hand, and occur in structured forms of conversation and discourse, utilizing language and physical actions as their most important mediators and resources. With the use of symbolic tools, it becomes possible for the conversants to express and objectify meanings, to compare and change them deliberately, to exchange and renegotiate them with others, to build on, defend, or challenge their partner's contributions, and to reflect on the organization of judgments and arguments (see van Oers, 1996; Chi, 2009). However, as observational studies show, coconstructive learning and dialogue are hardly homogeneous but inherently fragile processes in the service of convergence and mutual intelligibility. The achievement of a shared conceptual structure cannot be reliably predicted, nor does the iterative construction of a joint problem space through cycles of displaying, confirming and repairing occur by simply putting two students together. As Roschelle and Teasley (1995: p. 94) remark:

> [s]tudents' engagement with the activity sometimes diverged and later converged. Shared understanding was sometimes unproblematic and but oftentimes troublesome. The introduction of successful ideas was sometimes asymmetric, although it succeeded only through coordinated action. These results point to the conclusion that collaboration does not just happen because individuals are co-present: individuals must make a conscious, continued effort to coordinate their language and activity with respect to shared knowledge.

Context of Linguistic Discourse Analysis

From the perspective of communication and conversation analysis, coconstructive learning requires individuals to

collaboratively establish, exhibit, maintain, and update some degree of mutual understanding. Its accountability is granted "through a plurality of displays, claiming and demonstrating understanding thanks to the mobilization of linguistic and embodied resources at specific sequential positions" (Mondada, 2011: p. 550). One of the processes by which this is accomplished between individuals is called *grounding* (Clark and Brennan, 1991). Grounding as a basic form of collaboration means the moment-by-moment coordination and synchronization of the content-specific as well as the procedural aspects and steps of coconstructive activity. There is no need, however, to fully ground every aspect of an utterance in communication. Clark and Brennan (1991: p. 148) frame a pragmatic criterion for establishing common ground: The conversants "mutually believe that they have understood what [they] meant well enough for current purposes." Thus, the techniques that are used for grounding are shaped by the goal and the medium of communication. That is, the criterion of grounding and the techniques exploited for its maintenance dramatically change according to the purpose of communication (e.g., planning a party, swapping gossip, or gaining deep understanding) and the constraints of its medium (copresence, audibility, simultaneity, and visibility in face-to-face communication; sequentiality and reviewability in written communication, e-mail, or computer-supported collaborative work).

Pedagogical Concepts of Coconstruction and Its Facilitation

An important aspect of concern for the social nature of learning – and for a pivotal way in which it is supported by culture – is instructional dialogue or conversation. This term refers to a discursive activity in classrooms that permits the coconstruction of meaning between teachers and students, tutors and tutees, the more and the less experienced, and between students. Probably, the most important feature of a culture of coconstruction in the classroom is dialogue as opposed to, e.g., solo learning and teacher monologues. Emphasis on joint learning and instructional conversation among peers, and between teachers and students, is associated with the internal mediating processes that are essential for an understanding of how coconstruction through discourse operates and influences outcomes. The pedagogical cultivation of processes such as negotiation of meaning, reciprocal sense-making, revising one's cognitions in situations of sociocognitive conflict, precise verbalization of reasoning and knowledge, listening to others' lines of argumentation, tuning one's own information to that of a partner, lending and receiving help, or modeling cognitive and metacognitive activities to be internalized by the participating individuals should, thus, be placed at the core of instructional design.

Teacher–Student Talk: Tutoring and Scaffolding

Tutoring is commonly defined as a situation in which a more knowledgeable and skilled person (in the classroom typically the teacher) provides one-to-one support to a student. Consistent with Vygotsky's theory of the constructive role played by adults in children's knowledge acquisition, the teacher's goal of assistance can be seen as trying to get the students to share his or her understanding and knowledge. Because of the asymmetrical distribution of knowledge between teachers and students, understanding might be expected to be less jointly constructed in tutorial dialogues than it is observed to be in peer dialogues. Actions that tutors or teachers can take in order to elicit responses, including some coconstructive behavior from a tutee, are described in literature on tutoring (e.g., Chi et al., 2001; Graesser et al., 2012) and cognitive apprenticeship (Collins et al., 1989). They can be subsumed under two broad categories: (1) modeling, scaffolding, and fading as content-specific ways of providing hints, strategies, and situational forms of coaching and guidance that are tailored to the needs of individual students; and (2) prompting as a more content-neutral invitation by the tutor to elicit elaborations, reflections, and self-explanations from students (Chi et al., 2001). The metaphor of *scaffolding* in particular serves as a description of a productive way of supporting student thinking and problem solving. Specifically put, the teacher helps the students make sense of a difficult task in a highly responsive way that allows as much active student participation as possible. Moreover, as Kleine Staarman and Mercer (2010) notice, "the notion of scaffolding is not merely providing a learner with help to accomplish the task, but actively and temporarily providing a learner with just the right amount of support that is geared towards bringing him or her closer to a state of individual competency" (Kleine Staarman and Mercer, 2010: p. 79).

Teacher–Student Talk: Instructional Conversations and Dialogic Teaching

A rich body of research conducted in connection with pedagogical concepts of coconstruction and its facilitation studied instructional conversation during whole-class teaching. Numerous analyses of classroom talk have shown that in most cases it is characterized by a highly asymmetric distribution of utterances in favor of the teacher, which leaves students with only few opportunities to play an actively participating and responsible role in the process of knowledge construction. Classroom talk often follows the typical pattern of IRE or IRF sequences (Mehan, 1979; Sinclair and Coulthard, 1975). Such an IRE sequence starts with a teacher question (Initiation), followed by the response of the student and the evaluation or feedback of the teacher. Usually, the majority of teachers' questions requires only brief, 'correct' answers and do not invite students to think for themselves, to generate and externalize their own ideas about the object under consideration, or to take up and examine contributions of other students. It is for this reason that there seems to be only little space left for coconstructive activities such as negotiation of meaning and reciprocal sense-making. Referring to Bakhtin's (1981) distinction between monologic and dialogic discourse, some authors deemed this structure of classroom interaction to be rather monologic than dialogic (e.g., Ford and Forman, 2006). From a sociocultural perspective on teaching and learning, however, teaching approaches which are of a more dialogic nature are required in order to foster students' deep understanding of concepts and the appropriation of cognitive and communicative strategies of the discipline being taught.

Implementing a more dialogically geared teaching approach does not mean that the teacher completely abandons his or her leading role in the discourse. Depending on the part assigned to the students and the quality of the teacher's question and feedbacks, dialogic structures are possible even within a teacher-directed setting. So, by making use of interactional moves like 'revoicing' instead of evaluating a student utterance, the teacher can position a student as "a thinker or theorizer, the holder of a noteworthy idea, or explanation" (O'Connor and Michaels, 2007: p. 281), and as an authoritative and accountable member of a knowledge building community. The distinction between monologic and dialogic discourse might therefore be too simplified, as O'Connor and Michaels notice. They distinguish two notions of dialogic discourse, namely dialogic discourse as an ideological stance and dialogic discourse considered as a linguistic discourse structure. Thus, a discourse sequence may appear rather monologic with respect to its linguistic form, but at the same time it invites students to listen to each other, to generate arguments and to negotiate a shared understanding. Mortimer and Scott (2003) pursued a similar line of argument and proposed a two-dimensional model drawing a distinction between an authoritative versus dialogic dimension on the one hand and an interactive versus noninteractive dimension on the other hand.

As research on teacher–student talk indicates, it is challenging for teachers to develop an approach to teaching which is more dialogue-oriented. Yet there are theoretically substantiated concepts and related tools available to teachers that can support such a development, for instance, 'Dialogic Teaching' (Alexander, 2008), or 'Accountable Talk' (Michaels et al., 2008). Both concepts provide teachers with a collection of concrete recommendations and strategies or tools that can support the implementation of a more dialogic culture of interaction in the classroom.

Student–Student Talk: Collaborative Learning and Patterns of Dialogues

Student–student talk during a collaborative learning activity is perhaps the most obvious example of knowledge coconstruction in the classroom. In this context, collaborative learning is defined as an activity in the course of which "participants are engaged in a coordinated, continuing attempt to solve a problem or in some other way construct common knowledge" (Mercer and Littleton, 2007; see also Section Perspective of Situated and Socially Shared Cognition). In contrast to teacher–student talk, student collaboration is usually characterized by a (more or less) symmetrical relationship between the participants as regards hierarchy and knowledge level. In pursuing different theoretical perspectives on coconstructivism (see Sections Piagetian Perspectives: The Role of Sociocognitive Conflict; Vygotskian Perspectives of Coconstruction; Perspective of Situated and Socially Shared Cognition; and Context of Linguistic Discourse Analysis), research on the processes of collaborative learning has put forward several descriptions of patterns which guide productive student–student talk in the course of which both partners substantially contribute to a topic in a relatively equal and balanced way. One of the most prominent concepts in this regard is exploratory talk (as opposed to disputational and cumulative talk; Mercer and Littleton, 2007), which is characterized as "constructive and critical negotiation of views, offering constructive criticism and building on challenges and counter-challenges" (Vass and Littleton, 2010: p. 113). Exploratory talk is likely to incorporate other characteristics reported in the literature, such as regulation of sociocognitive conflicts, argumentation, or giving and receiving elaborated explanations (see also Webb, 2009). Another concept of productive student–student talk refers to patterns of joint dialogues (Chi, 2009) "which occur when both peers make substantive contributions to the topic or concept under discussion, such as by building on each other's contribution, defending and arguing a position, challenging and criticizing each other on the same concept or point, asking and answering each other's questions" (p. 82f).

Quite a lot of both theoretical and empirical work has addressed the processes that underlie dialogues and the conditions under which collaborative interaction is most likely to occur, and a whole range of possible ways to enhance its quality has been provided. Among the input characteristics that exert a complex influence upon the quality of interaction are: the preparation of the students for collaborative learning (including training for cooperation and discourse prior to the collaborative learning event), the establishment of a culture of dialogue and of problem-based learning, group characteristics (composition, size, ability, and sex), the goal and incentive structure of the task, and the structuring of group interaction (see, for reviews Chi, 2009; Vass and Littleton, 2010; Webb, 2009).

The question as to how dialogues can be elicited is quite important and needs to be considered in this connection. Beyond simply letting students work together in pairs – which is a means to be located on the surface level of instruction – an effective way of improving the quality of collaborative thinking is the explicit structuring of the interaction. Examples are scripted collaboration, 'Reciprocal Teaching' (Palincsar and Brown, 1984) or 'Guided Reciprocal Peer Questioning' (King, 2008). A commonality all these techniques share is the fact that a set of cognitive and metacognitive strategies (e.g., sentence openers, posing questions with generic prompts, instructing students how to proceed in a dialogue; see Chi, 2009), which have to be applied in a more or less prescribed way, is provided. Another possibility for making learning environments more supportive for collaborative activities is to enrich them with technology. Well-designed *computer-based cognitive tools* provide users with both process-related and task-related means for thinking and communication, irrespective of whether they are to be applied in face-to-face or computer-mediated interaction. Functioning as mediating resources and cognitive tools for the representation, negotiation, and modeling of concepts and activities, educational software has the potential – by making conceptual structures and processes visible, accessible, and manipulable on a computer screen – to facilitate processes of sharing understanding, of achieving convergence and intersubjectivity and of argumentation (e.g., Noroozi et al., 2012; Reusser, 1993).

Although collaborative learning emphasizes the joint construction of knowledge among students, the teacher has still an important role to play. In relation to the coconstructive activities of learners this role can be described within the

didactic framework of 'cognitive apprenticeship' (Collins et al., 1989). According to the ethnographic model in which practices and principles of traditional craftsmanship are applied to cognitive learning activities, teachers, experts, or more capable peers provide guidance and support to learners as they participate as apprentices in authentic and task-related, structured social interactions. As opposed to a transmissionist view of instruction, teachers should provide aid in the intellectual development of students in ways that leave room for negotiation and joint expansion of meaning: (1) as scaffolds and role models for the behavior that students are expected to engage in; (2) as active participants in learning groups aiming at shaping the group's dialogue; (3) as monitors of coconstructive norms in social interactions in which negotiation of taken-as-shared meaning is essential (Webb, 2009); (4) as advocates of content-specific standards and of the achievement of convergence and intersubjectivity in understanding and problem solving.

Associated with this shift in the pedagogical orientation of teachers is a shift in the role of learners and the organization of classrooms. In the wake of a view that sees learning essentially as sociocultural interaction, classrooms should develop from aggregations of solo learners to communities engaged in coconstructive learning. That is, individuals should become acculturated members of a culture and community through collaboration and negotiation. Or, as Bruner (1986: p. 123) has put it: culture as "a *forum* for negotiating and renegotiating meaning and for explicating action … is constantly in process of being recreated as it is interpreted and renegotiated by its members."

See also: Apprenticeship and School Learning: Lessons from Germany; Communities of Practice; Cooperative Learning in Schools; Instructional Design; Learning and Instruction: Social-Cognitive Perspectives; Piaget's Theory of Human Development and Education; Situated Learning (Learning In Situ); Social Constructivism; Vygotsky's Theory of Human Development and New Approaches to Education.

Bibliography

Alexander, R., 2008. Towards Dialogic Teaching: Rethinking Classroom Talk, fourth ed. Dialogos, York.

Bakhtin, M., 1981. The Dialogic Imagination: Four Essays. University of Texas Press, Austin.

Brown, J.S., Collins, A., Duguid, P., 1989. Situated cognition and the culture of learning. Educational Researcher 18, 32–42.

Bruner, J.S., 1986. Actual Minds, Possible Worlds. Harvard University Press, Cambridge, MA; London.

Chi, M.T.H., 1996. Constructing self-explanations and scaffolded explanations in tutoring. Applied Cognitive Psychology 10 (7), 33–49.

Chi, M.T.H., 2009. Active-constructive-interactive: a conceptual framework for differentiating learning activities. Topics in Cognitive Science 1, 73–105.

Chi, M.T.H., Siler, S.A., Jeong, H., Yamauchi, T., Hausmann, R.G., 2001. Learning from human tutoring. Cognitive Science 25, 471–533.

Clark, H.H., Brennan, S.E., 1991. Grounding in communication. In: Resnick, L.B., Levine, J.M., Teasley, S.D. (Eds.), Perspectives on Socially Shared Cognition. American Psychological Association, Washington, DC, pp. 127–149.

Collins, A., Brown, J.S., Newman, S.E., 1989. Cognitive apprenticeship: teaching the crafts of reading, writing, and mathematics. In: Resnick, L.B. (Ed.), Knowing, Learning, and Instruction. Erlbaum, Hillsdale, NJ, pp. 453–494.

Doise, W., Mugny, G., 1984. The Social Development of the Intellect. Pergamon Press, Oxford.

Ford, M.J., Forman, E., 2006. Redefining disciplinary learning in classroom contexts. Review of Research in Education 30, 1–32.

Graesser, A.C., Conley, M.W., Olney, A., 2012. Intelligent tutoring systems. In: Harris, K.R., Graham, S., Urdan, T. (Eds.), APA Educational Psychology Handbook, vol. 3. American Psychological Association, Washington, DC, pp. 451–474.

Greeno, J.G., 1998. The situativity of knowing, learning, and research. American Psychologist 53, 5–26.

Greeno, J.G., Benke, G., Engle, R.A., Lachapelle, C., Wiebe, M., 1998. Considering conceptual growth as change in discourse practices. In: Gernsbacher, M.A., Derry, S. (Eds.), Proceedings of the Twentieth Annual Conference of the Cognitive Science Society. Erlbaum, Mahwah, pp. 442–447.

King, A., 2008. Structuring peer interaction to promote higher-order thinking and complex learning in cooperating groups. In: Gillies, R.M., Ashman, A.F., Terwel, J. (Eds.), The Teacher's Role in Implementing Cooperative Learning in the Classroom. Springer, New York, pp. 74–91.

Kleine Staarman, J., Mercer, N., 2010. The guided construction of knowledge: talk between teachers and students. In: Littleton, K., Wood, C., Kleine Staarman, J. (Eds.), International Handbook of Psychology in Education. Emerald, Bingley, pp. 75–104.

Kozulin, A., 1998. Psychological Tools. A Sociocultural Approach to Education. Harvard University Press, Cambridge, MA.

Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, Cambridge, MA.

Mead, G.H., 1934. Mind, Self, and Society. University of Chicago Press, Chicago, IL.

Mehan, H., 1979. Learning Lessons: Social Organization in the Classroom. Harvard University Press, Cambridge, MA.

Mercer, N., Littleton, K., 2007. Dialogue and the Development of Children's Thinking. A Sociocultural Approach. Routledge, London.

Michaels, S., O'Connor, C., Resnick, L., 2008. Deliberate discourse idealized and realized: accountable talk in the classroom and civic life. Studies in Philosophy and Education 27, 283–297.

Mondada, L., 2011. The embodied and negotiated production of assessments in instructed actions. Research on Language and Social Interaction 42, 329–361.

Mortimer, E., Scott, P., 2003. Meaning Making in Secondary Science Classrooms. Open University Press, Maidenhead.

Noroozi, O., Weinberger, A., Biemans, H.J.A., Mulder, M., Chizari, M., 2012. Argumentation-based computer supported collaborative learning (ABCSCL): a synthesis of 15 years of research. Educational Research Review 7, 79–106.

O'Connor, C., Michaels, S., 2007. When is dialogue 'dialogic'? Human Development 50, 275–285.

van Oers, B., 1996. Learning mathematics as meaningful activity. In: Steffe, L.P., Nesher, P., Cobb, P., Goldin, G.A., Greer, B. (Eds.), Theories of Mathematical Meaning. Erlbaum, Mahwah, NJ, pp. 91–113.

Palincsar, A., Brown, A.L., 1984. Reciprocal teaching and comprehension-fostering and comprehension-monitoring activities. Cognition and Instruction 1, 117–175.

Perret-Clermont, A.N., Perret, J.-F., Bell, N., 1991. The social construction of meaning and cognitive activity in elementary school children. In: Resnick, L.B., Levine, J.M., Teasley, S.D. (Eds.), Perspectives on Socially Shared Cognition. American Psychological Association, Washington, DC, pp. 41–62.

Piaget, J., 1925. Psychologie et critique de la connaissance. Archives de Psychologie 19, 193–210.

Piaget, J., 2000. The Moral Judgement of the Child. Routledge, London.

Piaget, J., 1975. L'Equilibration des structures cognitives. Presses Universitaires de France, Paris (Engl., 1985. The Equilibration of Cognitive Structures: The Central Problem of Intellectual Development. University of Chicago Press, Chicago, IL).

Reusser, K., 1993. Tutoring systems and pedagogical theory: representational tools for understanding, planning, and reflection in problem-solving. In: Lajoie, S.P., Derry, S. (Eds.), Computers as Cognitive Tools. Erlbaum, Hillsdale, NJ, pp. 143–177.

Rogoff, B., 1990. Apprenticeship in Thinking. Cognitive Development in Social Context. Oxford University Press, Oxford.

Roschelle, J., Teasley, S.D., 1995. The construction of shared knowledge in collaborative problem solving. In: O'Malley, C. (Ed.), Computer Supported Collaborative Learning. Springer, Berlin, pp. 69–97.

Sinclair, J., Coulthard, R.M., 1975. Towards an Analysis of Discourse: The English Used by Teachers and Pupils. Oxford, London.

Vass, E., Littleton, K., 2010. Peer collaboration and learning in the classroom. In: Littleton, K., Wood, C., Kleine Staarman, J. (Eds.), International Handbook of Psychology in Education. Emerald, Bingley, pp. 105–135.

Vygotsky, L.S., 1962. Thought and Language. Harvard University Press, Cambridge, MA.

Webb, N.M., 2009. The teacher's role in promoting collaborative dialogue in the classroom. British Journal of Educational Psychology 79, 1–28.

Code Switching: Linguistic

Shana Poplack, Linguistics Department, University of Ottawa, Ottawa, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by S. Poplack, volume 3, pp. 2062–2065, © 2001, Elsevier Ltd.

Abstract

Though code-switching (CS) has been exceptionally well researched, controversy continues to reign over its identity, structure, and the rules governing its use. This article relates the state of the field to epistemological differences among researchers with respect to data, methods, the nature of evidence, and principles of scientific proof. Progress toward developing accountable theories of CS will be achieved once greater efforts are made to distinguish (1) apparently homologous but fundamentally distinct manifestations of language contact (currently subsumed under the label of CS) and (2) robust community-based usage patterns from isolated counterexamples.

Introduction

No outcome of language contact has garnered more scholarly attention than code-switching (CS). In its broadest sense, CS refers to the mixing of two or more languages in discourse. What prompts bilingual speakers to engage in such mixing? A variety of different motivations have been identified (e.g., Auer, 1995; Bhatt and Bolonyai, 2011; Gumperz, 1976/1982), often involving the bilingual abilities of the interlocutors (speak Croatian to grandma because she doesn't speak English), the context (call it *pizza* or *latte*, because there's no good English word for it), or the establishment of identity (as in emblematic CS), among many other sociocultural determinants. These contrast with what some have called 'true' or 'internally generated' CS (Gullberg et al., 2009), which has the particularity that two (or more) languages are spontaneously brought to bear in a single utterance with no change in interlocutor, situation, or even topic, i.e., with no external trigger. In some tightly knit bilingual communities, particularly Spanish–English communities in the United States, CS actually functions as a discourse *mode* (Pfaff, 1979; Poplack, 1980; Torres Cacoullos and Travis, 2015). This article surveys the linguistic treatment of such *intrasituational* switching.

Languages may, in principle, be combined at any level of linguistic structure, but community studies reveal that three are favored. These display varying degrees of syntactic involvement of the languages involved, and therefore require differing degrees of bilingual proficiency to achieve a 'felicitous' outcome (Poplack, 1980). The other-language element may be a *tag*, as in the Finnish–English example in (1), which can typically be freely inserted anywhere in the sentence with few if any syntactic repercussions; it may consist of a full *clause* or major sentence constituent, as in the French–English example in (2), where the syntactic requirements of either language are not a consideration, or it may appear *intrasententially* within the confines of a single sentence or constituent, as in the Spanish–English example in (3). It is this last type of CS that has attracted the most linguistic attention.

1. Mutta en mä viittinyt, *no way!* (9b.134)
   But not-1p. I bothered
   "But I'm not bothered, no way!" (Poplack et al., 1987, p. 394)
2. Je vas voir la fille, je dis, "*You got two tickets for Ottawa*?" (OH.105.3294)
   "I go see the girl, I say, 'you got two tickets for Ottawa?'" (Poplack and Dion, 2012, p. 300)
3. Pero los *grades are so bad* tambien, oiga. (088.10)
   "But the grades are so bad too, you know?" (Torres Cacoullos and Aaron, 2003, p. 291)

In combining languages intrasententially, various problems of incompatibility may arise. The most obvious ones derive from word order differences: under what conditions, if any, can the boundary between constituents ordered differently in two languages host a switch between them? Other potential combinatorial difficulties involve cross-language mismatches in grammatical categories, subcategorization patterns, and idiomatic expressions. Systematic examination of the spontaneous speech of bilinguals resident in a wide range of communities, however, suggests that the vast majority of sentences containing intrasentential CS remain 'grammatical,' in the sense that the appearance in them of two distinct languages does not contravene the syntax of either. This observation eventually led to the general consensus that these combinations are not effected randomly, as per earlier convictions (e.g., Weinreich, 1953/1968), but are rule governed. Discovery of the nature and locus of these rules, and the mechanisms by which they operate to produce well-formed, albeit 'mixed' utterances, has been the major goal of linguistic research on CS. Central questions include locating permissible switch sites, ascertaining the nature (hierarchical or linear, variable or categorical) of the constraints on switching, and determining when other-language items of one language (La) come to assume the structure of another (Lb) into which they are incorporated, or alternatively, retain their original La identity.

Close inspection of intrasentential mixing actually produced by bilingual speakers shows that the vast majority of other-language material appears in two major forms and results from two distinct processes. It may consist of a *lone* content word, such as those italicized in the Turkish–English example in (4) or be part of a *multiword fragment* of La in otherwise Lb discourse, as are the French and Wolof portions of (5). (All examples are reproduced verbatim from actual speaker utterances.)

4. … you go to visit your family … and visit your *nene* ('grandmother') and your *amcas* ('uncle-PL') and your *teyzes* ('aunt-PL') and that's about it really (EliF.153.5B) (Adalar and Tagliamonte, 1998, p. 154)
5. des fois da nga y xool un film avec des sous-titres en français (1.319)
   sometimes AUX you ASP watch a film with IND subtitles in French
   "Sometimes you watch a film with subtitles in French" (Poplack and Meechan, 1995, p. 213)

Lone grammatical elements (e.g., affixes, determiners, prepositions) are not generally involved. The La content words may have been *actively* drawn from La by the bilingual in the course of speaking Lb, or simply retrieved from an Lb lexical stock containing elements drawn from La at some point in the past, but now fully incorporated into the native Lb lexicon. (These are in fact equally accessible to monolingual speakers of Lb.) This is illustrated in (6), in which every 'English' noun is etymologically French. Such items, known as *loanwords*, are exceedingly common in languages of the world (e.g., Haspelmath and Tadmor, 2009).

6. The *judge* called on the *jury*, *lawyers*, and other *members* of the *court* to stand.

Loanwords are particularly instructive to students of language contact because of the evidence they provide on the synchronic result of the diachronic process of *borrowing*, which turns out to be the major strategy for language mixing. That evidence converges overwhelmingly on the fact that established loanwords come to assume the (morphosyntactic) properties of the language into which they were borrowed (the *recipient* language), following its morphological requirements and adopting its word order, while retaining none of their donor-language identity other than their etymology, and optionally, some elements of the phonology. Moreover, the available evidence suggests that this outcome is universal: identity with recipient-language morphosyntax is virtually always achieved, regardless of the typological features of the languages involved. CS, in contrast, retains the distinctive grammatical properties of its language of origin. As such, loanwords offer a crucial structural *benchmark* against which to identify mixed elements whose status is not so clear (see Section Evaluating CS Theories). (Because phonological integration is gradient, in both long-attested (e.g., *bar* [bɑɹ]) and more recent borrowings (Poplack, 2012; Poplack and Dion, 2012; Poplack et al., 1988), as well as in CS and unmixed speech more generally, on its own it is a poor predictor of language status.)

A 'Grammar' of CS

Bilinguals tend to switch intrasententially at certain (morpho) syntactic boundaries and not at others. Early efforts to explain these preferences proceeded by proscribing certain switch sites, e.g., between pronominal subjects and verbs or between conjunctions and their conjuncts (Gumperz, 1976/1982; Timm, 1975). However, these particular sites were soon reported to figure among the regular CS patterns of some bilingual communities (e.g., Pfaff, 1979; Poplack, 1981).

The first more general account of the distribution of CS stemmed from the observation that CS is favored at the kinds of syntactic boundaries which occur in both languages. The *Equivalence Constraint* (Poplack, 1980, 1981) states that switched sentences are made up of concatenated fragments of alternating languages, each of which is grammatical in its respective language of origin. The boundary between adjacent fragments occurs between two elements that are ordered in the same way in both languages, ensuring the linear coherence of sentence structure without omitting or duplicating lexical content (see also Lipski, 1977; Muysken, 2000; Pfaff, 1979).

That general principles, rather than atomistic constraints, govern CS is now widely accepted, though there is little consensus as to what they are or how they should be represented. Some research assumes that the mechanisms for language switching follow directly from general principles of (monolingual) grammar. Theories based on this assumption tend to appeal to abstract grammatical properties; since Klavans' (1985) proposal that CS was constrained by structural relations, the formal linguistic theories successively in vogue in the monolingual arena have each been extended to encompass CS. di Sciullo et al. (1986), for example, identified the relevant relations as C-command and government: CS cannot occur where a government relation holds, e.g., between verb and object, or preposition and NP complement. Later, the notion of feature agreement led to a parallel focus on feature matching in CS studies, as in the Functional Head Constraint (Belazi et al., 1994, 1995), which adds language choice to the features instantiated in functional and lexical categories, prohibiting CS where a mismatch occurs.

The distinction between lexical and functional categories is a hallmark of theories invoking the complement structure of individual lexical items to characterize permissible CS sites (e.g., Bentahila and Davies, 1983; Joshi, 1985; Mahootian, 1993). Perhaps the most detailed such model is the *Matrix Language Frame* model (e.g., Azuma, 1993; Myers-Scotton, 1993, 2002). Here, structural constraints on CS result from a complex interaction between a dominant 'matrix' language, which determines word order and supplies 'system' morphemes, and an 'embedded' language, which may only contribute content elements. The many exceptions to this generalization, notably all those consisting of 'embedded language islands' (i.e., the *multiword* fragments of La), are treated by a large number of subsidiary principles.

A more recent proposal in the minimalist framework dismisses the foregoing theories, not only as 'descriptively invalid,' but also as theoretically undesirable, because they posit CS-*specific* mechanisms. The 'Null Theory' (Mahootian, 1993; Mahootian and Santorini, 1996), like the minimalist approach (Chan, 2003, 2009; MacSwan, 1999, 2009), holds that there are no constraints or principles specific to CS in the language faculty of bilinguals. Following the dictate that only the 'minimal theoretical assumptions may be made to account for linguistic data,' and 'privileging more simplistic and elegant accounts over complex and cumbersome ones,' MacSwan (2009, p. 325) proposed the *Phonetic Form Interface Condition* (PFIC) to account for a purported 'ban' on CS in word-internal and head-movement contexts.

The PFIC was rapidly deemed descriptively invalid in its own right, on the basis of counter evidence brought by

Schindler et al. (2008) and Legendre and Schindler (2010), who in turn adopt the Optimality-Theoretic (OT) framework advocated by Bhatt (1997) for bilingual syntax to propose that "CS emerges from the resolution of conflicting constraints," and that different patterns of CS result from alternative rankings of a small set of universal but violable constraints. According to its proponents, Optimality Theory would explain why certain language pairs seem to violate constraints that account for CS in other language pairs. Though the methodology and specific assumptions differ, the Optimality-Theoretic approach is closest in spirit to the variationist perspective (e.g., Poplack, 1993), according to which a prediction like that made by the Equivalence Constraint is seen as universally applicable, but differentially instantiated according to the specific properties of the language pair – allowing CS between DET and noun in French–English (both of which feature DET + N order), for example, but prohibiting it between Swedish (N + DET) and English – and *variably* implemented.

As of this writing, however, not only is there no general consensus on what linguistic conditions constrain CS, but some scholars insist that the very search for them is futile (e.g., Bokamba, 1989; Gardner-Chloros and Edwards, 2004); others maintain that there are none (Chan, 2009; MacSwan, 1999; Mahootian, 1993), and still others, that constraints on CS may vary from one community to another (Bhatt, 1997; Muysken, 2000; Schindler et al., 2008).

There has been remarkably little cross-fertilization among these theories, and few researchers have accepted, let alone built upon, existing findings. Instead, scholarship in this field has consisted largely of successive dismissals of previous proposals, often on the basis of counter-evidence of uncertain provenance. Should the field proceed along the path of the last several decades, new models for CS will continue to be proposed and discarded, in tandem with new developments in (monolingual) syntactic theory. Whether any of these will come any closer than their predecessors to accounting for the data of actual CS productions will depend on the extent to which they succeed in addressing and resolving the following core issues, as well as the considerations outlined in Section Why Consensus Remains Elusive.

- Are all types of language mixture instantiations of the same process or do they result from different processes with distinct outcomes?
- Do lone other-language elements and multiword fragments behave in the same way and can they all be accommodated under a single theory?
- Is the process giving rise to multiword CS best characterized as alternation, insertion, or something else?
- Under what conditions are other-language elements integrated into a recipient language, and why do some of them resist integration?
- Are CS and borrowing related in the sense that the former gives rise to the latter?

Why Consensus Remains Elusive

Testing the fit of competing models against the data of CS should be a straightforward matter, especially since they often make competing predictions. But their disparate assumptions, goals, and domains of application have hindered such efforts. In what follows we attempt to pinpoint some of the reasons why consensus remains so elusive, despite decades of intense scholarly attention.

The Quest for an 'Optimal' Theory of CS

Lack of a widely accepted evaluation metric for CS theories, coupled with the range of their predictions, their volatility, and pervasive doubt over the extent to which they account for the data (and if so, which portions thereof), have all conspired in the general failure of any one to gain widespread acceptance. Section A 'Grammar' of CS detailed the tradition of appealing to the syntactic theories currently in vogue as models for CS behavior. But the assumption that bilingual syntax can be explained by general principles inferred from the study of monolingual grammar has not yet been substantiated. While formal theories of grammar may account well for monolingual language structure, including that of the monolingual fragments in CS discourse, there is no evidence to suggest that the act of *combining* two languages can be explained in the same way.

Theory-internal constraints play a more insidious role. For example, blind adherence to Occam's razor (e.g., Chan, 2009; MacSwan, 1999) and the quest for a 'unified account' (Gardner-Chloros, 2009; Myers Scotton, 2006; Winford, 2009) enjoin us to ignore fundamental differences among different manifestations of language contact, and to treat them all – at our peril! – as one and the same. The arbitrary requirement that an optimal theory must necessarily eschew CS-specific mechanisms diverts researchers from recognizing and attending to community trends. Different patterns of adapting monolingual resources in code-mixing strategies have been reported in different bilingual communities, even when the same language pair is involved; these are not predictable through purely linguistic considerations (Poplack, 1987; Muysken, 2000; Torres Cacoullos and Travis, 2015).

Variability, Probability, and Quantitative Reasoning

The field of contact linguistics has not fully acknowledged that the property of *inherent variability* (Labov, 1969), now generally accepted to characterize spoken language, extends as well to the bilingual context. The existence of variability renders the quest for 'categorical' constraints (obeyed 100% of the time) on CS futile. Instead, the behavior of bilingual speakers will be characterized by *patterns*, some very pervasive, others far more tenuous, as well as any number of outlier tokens, slips of the tongue, misspeakings, and even the odd speech error, *just as is the speech of monolinguals*. Distinguishing robust community patterns from the idiosyncratic utterance of any provenance requires in the first instance identifying a stable bilingual community where CS occurs regularly, and second, conducting accountable quantitative analyses of sufficiently large corpora of spontaneous CS production.

The Data

The data on which theories of CS have been built derive from a wide variety of (often incommensurable or even unidentified) sources. They may be intuited, invented, overheard, remembered, elicited in a laboratory setting, or observed during regular interactions among bilinguals in a bilingual speech community. The provenance of the many purportedly 'prohibited' CS which figure so prominently in CS theory building and demolition is likewise often unstated. Clearly, all kinds of data (with the exception of outright fabrications) have their place in CS research (e.g., Gullberg et al., 2009), but crucially, they are not interchangeable. The gold standard remains the (standard sociolinguistic-style) bilingual *corpus*: a principled assemblage of actual spontaneous productions of systematically sampled bilingual speakers resident in a well-defined bilingual speech community. Where these are stratified, randomized, and subjected to other basic requirements of social science research they yield invaluable information not only about the types and distribution of bilingual mixing strategies, but also about the associations they entertain with different sectors of the society and types of interaction.

A further consideration is that, from an analytical perspective, corpus data will almost always contain a variety of different types of language mixture. As detailed in Section Evaluating CS Theories, these do not necessarily behave in the same way. We cannot evaluate a theory of CS if we cannot assess its fit to a well-circumscribed body of unambiguous CS data.

Methods: Cherry-Picking versus Accountability

To be sure, many scholars of language contact have elicited bilingual data of varying degrees of spontaneity, and some have even constructed sizable corpora. But exploitation of this material is too often restricted to a few selected examples, or to declarations in lieu of demonstrations that all the data satisfy some constraint.

The real value of a corpus can be realized only through the *principle of accountability* (Labov, 1972), where every token of all constructions carrying out a certain role is counted and placed in statistical context. Only then are patterns discernable and general trends emergent from the complex production of spontaneous discourse. With few exceptions, however, linguistic analyses of CS have not been characterized by quantitative reasoning (in marked contrast to many psycholinguistic and neurolinguistic studies of the same topic). Remarkably, some linguistic researchers actually advocate *against* the use of quantitative methodology in the study of CS (Gardner Chloros, 2009). This is unfortunate, because where accountable analyses are *replicated*, striking commonalities have emerged (cf the papers in Poplack and Meechan, 1998a; Torres Cacoullos and Travis, 2015). These regularities are obscured by the emphasis accorded exceptions by the anecdotal reports that pervade the field.

The Pre-Eminence of the Counterexample

Indeed, the literature on CS is largely characterized by the 'rule-and-exception' paradigm. Each successive theory has been greeted with an onslaught of counterexamples, but very few have been systematically tested against principled collections of data of any sort. Even more rarely have the counterexamples been shown to represent quantitatively meaningful *patterns* of language use in a well-defined bilingual speech community. Instead, not only the theories but also tests of their applicability tend to be based on isolated examples, drawn from judgments, informant elicitation, and linguist introspection. Such data do not permit us to distinguish between the recurrent and systematic patterns of everyday interaction and utterances which may have been judged 'acceptable' in some sense, but which rarely or never occur. Even where theories are enthusiastically received, they seldom inspire the type of quantitative empirical work required to test their predictions on novel data sets. This lacuna, coupled with the tolerance for isolated counterexamples as a surrogate for proof, and the relative paucity of systematic quantitative analyses of large data sets and associated measures of significance and goodness of fit, is another key contributor to the lack of consensus. Failure to get on the bandwagon of replicable quantitative studies of spontaneous CS dooms the field to an endless round of rule-and-exception, minimizes our chances of arriving at cumulative findings, and hinders advancement of our understanding of CS.

Evaluating CS Theories

Assessment of the descriptive adequacy of a theory of CS requires that at least two methodological issues be resolved. One involves the proper classification of other-language phenomena; the other, systematic quantitative confrontation of the predictions of the theory with the data of actual bilingual behavior.

CS versus Borrowing

CS differs from the other major manifestation of language contact: *lexical borrowing*. Despite etymological identity with the donor language, established loanwords assume the morphological, syntactic, and often, phonological identity of the recipient language. They tend to be recurrent in the speech of the individual and widespread across the community. The stock of established loanwords is available to monolingual speakers of the recipient language, who access them normally along with the remainder of the recipient-language lexicon. Loanwords further differ from CS in that there is no involvement of the morphology, syntax, or phonology of the lexifier language.

Nonce Borrowing

It has now been established that borrowing is actually much more productive than implied above (e.g., Poplack and Dion, 2012; the papers in Poplack and Meechan, 1998a; Poplack et al., 1988). In particular, the social characteristics of recurrence and diffusion are not always satisfied. This results in what has been called, after Weinreich (1953/1968), *nonce borrowing* (Sankoff et al. 1990). Like its established counterpart, the nonce borrowing tends to involve lone lexical items,

generally major class content words, from La (acting here as donor language), and to assume the morphological, syntactic, and optionally, phonological identity of Lb, the recipient. Like CS, on the other hand, nonce borrowing is neither recurrent nor widespread, and involves *actively* drawing from La, necessarily requiring a certain level of bilingual competence. The Nonce Borrowing Hypothesis (Poplack, 2012; Sankoff et al. 1990) captures the empirical observation that speakers not only CS spontaneously, but may also *borrow* spontaneously, and these spontaneous borrowings assume the morphological and syntactic, and, optionally, phonological, identity of the recipient language *prior to* and *independently of* achieving the social characteristics of established loanwords (recurrence in the speech of the individual and diffusion across the community). The example in (7) illustrates how the nonce borrowing *coper* (<Eng. *cope*) is fully adapted morphologically and syntactically into French, despite having been unattested at the time of utterance, and produced only once by a single speaker.

7. Je serais pas capable de *coper* avec. (OH.037.1086)
   "I wouldn't be able to cope with it" (Poplack et al., 1988, p. 52)

In this, nonce borrowings mirror the linguistic behavior not only of established loanwords (e.g., *upgrader* 'to upgrade'), but of native counterparts (e.g., *échouer* 'to fail'). As noted in the Introduction, phonological integration, by virtue of its variability, is not diagnostic here.

Distinguishing Language Contact Phenomena

The classification of lone items is at the heart of a fundamental disagreement among CS researchers over (1) whether there is a distinction between CS and borrowing, (2) whether this should be formally recognized in a theory of CS, (3) whether these and other manifestations of language contact can even be identified in bilingual discourse, and (4) operational criteria for determining whether a given item was switched or borrowed. Researchers who consider lone other-language items to be CS tend to posit – for all CS – an *asymmetrical* relationship, in which one language dominates and other-language items are *inserted* (e.g., Joshi, 1985; Myers-Scotton, 1993). Where the class of CS is (in the first instance) limited to unambiguous *multiword* fragments, the languages are seen to alternate (Belazi et al., 1994; Sankoff, 1998a,b; Sankoff and Poplack, 1981; Woolford, 1983). Muysken (2000) admits both insertion and alternation as strategies.

Quantitative analysis of language mixing phenomena in both typologically similar and distinct language pairs shows that lone other-language items, especially major-class content words, are typically the most important component of mixed discourse by far (e.g., Berk-Seligson, 1986; Nortier, 1989; Treffers-Daller, 1994; Shin, 2002; among many others). The severe quantitative disproportions among single-word and multiword other-language items, coupled with the recent discovery (Poplack and Dion, 2012) that the vast majority of the former are integrated *abruptly* into recipient-language grammar, renders any analysis treating them as the same in effect a study of borrowing. They also explain the persistence of (descriptively inadequate) insertion analyses as blanket models for both CS and (nonce) borrowing. They do account for the bulk of the language mixing data, but not for the data of the less common multiword CS.

Contextualizing Language Contact Phenomena with Respect to Relevant Benchmarks

Distinguishing nonce borrowings from single-word CS is conceptually easy but methodologically difficult, especially when they surface bare or uninflected, giving no apparent indication of language membership. Insofar as CS and borrowing are based on some principled combination of elements of the monolingual (i.e., *unmixed)* vernaculars of the bilingual community, it is crucial to have as explicit an idea as possible of the structure of these vernaculars before concluding that a code-mixed element is behaving like one or the other. Such contextualization is usually omitted from CS research. The relevant monolingual benchmarks are not prescribed standards but the (unmixed) contact languages *as they are actually spoken*, and as noted above, spoken language is characterized by structural variability. In dealing with, rather than evading this variability, Sankoff et al. (1990) and Poplack and Meechan (1998b) developed a method to compare bilingual structures with the unmixed source languages of the same speakers. Making use of the framework of linguistic variation theory (Labov, 1969), the patterning of variant choice is used to determine their status. If the rate and distribution of, for example, case marking of the contentious lone other-language items show quantitative parallels to those of their counterparts in the (unmixed) recipient language, while at the same time differing from relevant patterns in the donor language, the lone other-language items are inferred to be largely *borrowed*, since only the grammar of the recipient language is operative. If they pattern with their counterparts in the (unmixed) donor language, while at the same time differing from the patterning in the unmixed recipient language, the lone other-language items must result from CS.

The last several decades have witnessed the accumulation of a coherent body of variationist studies on a large number of language pairs, involving large corpora of spontaneous bilingual speech, and a rich variety of morphosyntactic diagnostics such as determiner expression, adjective placement, case-marking, word order, and verb formation, among many others. These have demonstrated that the vast majority of lone items show the same fine details of quantitative conditioning of morphological and syntactic variability as dictionary-attested loanwords, both of which in turn parallel their unmixed counterparts in the recipient language (e.g., Aaron, 2014; Adalar and Tagliamonte, 1998; Blas Arroyo and Tricker, 2000; Ghafar Samar and Meechan, 1998; Poplack and Dion, 2012; Shin, 2002; Torres Cacoullos and Aaron, 2003; Torres Cacoullos and Travis, to appear). This tendency is apparent, regardless of the typological properties of the languages involved.

These quantitative studies have confirmed beyond any doubt that CS and borrowing are indeed two distinct processes, which are governed by different rules, and which – using the appropriate methodology – may be operationally distinguished as such (except in inherently language-neutral constructions). They have also converged on two

groundbreaking discoveries, correcting widely espoused assumptions in the field:

- Morphosyntactic integration of other-language items is independent of phonological integration. (Phonology, of both CS and borrowing, is variable, and thus cannot reliably be used to distinguish between them.)
- Morphosyntactic integration is not contingent on frequency and diffusion (i.e., achievement of loanword status), but rather occurs abruptly at the nonce borrowing phase.

Other recurrent findings to emerge from the systematic quantitative analysis of tens of thousands of tokens of spontaneous bilingual production data are

- the vast majority of all instances of language mixing are *single* other-language items and
- nearly all of them are immediately integrated morphologically and syntactically into the recipient language.

This means that a disproportionate number of lone other-language items have the linguistic features of established loanwords, independent of their social characteristics of frequency in the individual or recurrence across the community. Single-word CS, on the other hand, is exceedingly rare. (This finding does not license the a priori identification of *all* lone items as borrowings, however; the variationist literature offers clear discovery procedures to determine their status). Most important, the linguistic properties of CS – of whatever length – have been consistently shown in quantitative studies to contrast diametrically with those of borrowings: both the internal constituency and positioning in the clause of *borrowed* words come from the recipient language, while the internal constituency of CS is that of the language of origin, but its placement in the sentence tends to respect the word order requirements of *both* languages in the mix. This is the basic insight behind the Equivalence Constraint.

The Equivalence Constraint has been quantitatively verified as a major tendency in a number of language pairs, both typologically similar and different (e.g., Spanish–English (Poplack, 1980), Finnish–English (Poplack et al., 1987), Arabic–French (Naït M'Barek and Sankoff, 1988), Tamil–English (Sankoff, Poplack and Vanniarajan 1990), Fongbe–French, Wolof–French (Meechan and Poplack, 1995; Poplack and Meechan, 1995), Igbo–English (Eze, 1998), French–English (Turpin, 1998), and Ukrainian–English (Budzhak-Jones, 1998)). But most of the voluminous literature on CS, especially of the 'insertional' type, is based on data which represents, properly speaking, lexical borrowing. Thus, virtually all the published 'counterexamples' to the Equivalence Constraint involve single content words, which of course do not display the homologous word order that characterizes most true CS, but instead pattern syntactically and morphologically with the recipient language, just as predicted by the Nonce Borrowing Hypothesis (Poplack, 2012; Sankoff et al. 1990). Once the data of language mixing have been accurately circumscribed, and accountable analyses of large bodies of bilingual production data undertaken, nonce borrowing and CS under equivalence emerge as the quantitatively preponderant active mixing strategies employed in bilingual communities.

Future Directions

Over the past half-century, interest in CS has not only not abated, but to judge by the now vast amount of literature on the topic, it seems in fact to be growing. In welcome and instructive developments, approaches to its study have spread across linguistic subdisciplines from sociolinguistics to syntax to psycholinguistics, neurolinguistics, and beyond. In this context it is most disconcerting to the discipline that answers to the core *linguistic* questions continue to elude us.

This article has related the lack of consensus characterizing the linguistic study of CS to a number of epistemological problems. Foremost among them is failure to distinguish multiword CS from other types of language mixture, which, despite some similarities in surface manifestation, are fundamentally different mechanisms for combining languages. The current state of knowledge suggests that borrowing, nonce or established, is the major manifestation of language contact in most bilingual communities. Intrasentential CS involving multiword fragments of two or more languages is also attested in some communities. Achievement of consensus regarding an empirically verifiable characterization of the rules for *combining* such fragments within the sentence remains an important goal for CS research. To achieve it, we must improve the fit between theories and data. This would be facilitated by a broader empirical base.

In the first instance, existing theoretical, experimental, and anecdotal accounts of CS should be supplemented with many more large-scale empirical studies of actual bilingual behavior in a much wider variety of well-defined bilingual communities. As has been amply documented in the monolingual case, bilingual communities may exhibit widely different patterns of adapting monolingual resources in their code-mixing strategies, and these are not predictable through purely linguistic considerations. Thus, language pairs with the same typological characteristics do not necessarily display the same CS patterns in different communities (Poplack, 1987), the same language may behave differently when combined with different languages (Meechan and Poplack, 1995; Naït M'Barek and Sankoff, 1988; Nortier, 1989), and even the same language pair may display different patterning from one community to the next (Pfaff, 1979; Poplack, 1980; Torres Cacoullos and Travis, 2015). To account for these unruly facts, much more speaker-, context-, and community-based information must be incorporated into our analyses than has heretofore been the case. Consideration of the data of actual bilingual interactions in the context of the speech community in which they were produced would permit researchers to situate bilingual behavior with respect to the monolingual vernaculars implicated in language mixing, account for the disparate CS strategies that have evolved in different bilingual communities, and distinguish among incommensurable manifestations of bilingual language contact. Many more such studies are accumulating, but they are still in the minority.

To develop a theory of CS that is truly accountable to the facts, greater efforts must be made to tease apart the idiosyncratic and/or exceptional uses of individuals from community strategies. This can only be achieved via more accountable analyses of principled collections of naturally occurring bilingual data,

using accepted social-science standards of proof, including data-driven reports of rates of occurrence, conditioning of variant choice, and measures of statistical significance. This will necessarily involve distinguishing among the different manifestations of language contact and contextualizing the different mixing strategies with respect not only to each other, but crucially, to the monolingual benchmark varieties actually involved in the process (as opposed to some standard varieties thereof) as well.

We would then be in a position to carry out *more linguistic analyses* of the type of CS that all analysts can agree to constitute CS (i.e., multiword fragments of one language alternating with multiword fragments of another), free of the distorting effects of other manifestations of language contact. As detailed above, the latter, in particular lexical borrowing, have been demonstrated to far exceed unambiguous CS, and more importantly, display different linguistic behaviors from them, with the predictable result that including them in the pool of CS can only blur their structure. Recall that the major goal of linguistic analyses is to uncover, not the *constitution* of CS, which is already well understood (it corresponds exactly to that of its language of origin), but rather the *positioning* of CS in the sentence. Only by paying close attention to the directions outlined above, in conjunction with more experimental studies on multiword fragments, can we hope to achieve a true understanding of this endlessly fascinating recourse uniquely available to bilinguals.

See also: Diglossia; Language Contact; Language and Society; Multilingualism; Quantitative Linguistics; Sociolinguistics.

Bibliography

Aaron, J.E., 2014. Lone English-origin nouns in Spanish: the precedence of community norms. International Journal of Bilingualism. http://dx.doi.org/10.1177/1367006913509900.

Adalar, N., Tagliamonte, S., 1998. Borrowed nouns; bilingual people: the case of the "Londrali" in Northern Cyprus. International Journal of Bilingualism 2 (2), 139–160.

Auer, P., 1995. The pragmatics of code-switching: a sequential approach. In: Milroy, L., Muysken, P. (Eds.), One Speaker, Two Languages: Cross-Disciplinary Perspectives on Code-Switching. Cambridge University Press, Cambridge/New York, pp. 115–135.

Azuma, S., 1993. The frame-content hypothesis in speech production: evidence from intrasentential code switching. Linguistics 31, 1071–1093.

Belazi, H., Rubin, E., Toribio, A.J., 1994. Code switching and X-bar theory: the functional head constraint. Linguistic Inquiry 25 (2), 221–237.

Bentahila, A., Davies, E.E., 1983. The syntax of Arabic-French code-switching. Lingua 59 (4), 301–330.

Berk-Seligson, S., 1986. Linguistic constraints on intrasentential code-switching: a study of Spanish/Hebrew bilingualism. Language in Society 15, 313–348.

Bhatt, R., 1997. Code-switching, constraints, and optimal grammars. Lingua 102, 223–251.

Bhatt, R.M., Bolonyai, A., 2011. Code-switching and the optimal grammar of bilingual language use. Bilingualism: Language and Cognition 14 (4), 522–546.

Blas Arroyo, J.L., Tricker, D., 2000. Principles of variationism for disambiguating language contact phenomena: the case of lone Spanish nouns in Catalan discourse. Language Variation and Change 12 (2), 103–140.

Bokamba, E., 1989. Are there syntactic constraints on code-mixing? World Englishes 8 (3), 277–292.

Budzhak-Jones, S., 1998. Against word-internal code-switching: evidence from Ukrainian-English bilingualism. International Journal of Bilingualism 2 (2), 161–182.

Chan, B.H.-S., 2003. Aspects of the Syntax, the Pragmatics, and the Production of Code-Switching. Peter Lang, New York.

Chan, B.H.-S., 2009. Code-switching between typologically distinct languages. In: Bullock, B.E., Toribio, A.J. (Eds.), The Cambridge Handbook of Linguistic Code-Switching. Cambridge University Press, Cambridge, pp. 182–198.

Eze, E., 1998. Lending credence to a borrowing analysis: lone English-origin incorporations in Igbo discourse. International Journal of Bilingualism 2 (2), 183–201.

Gardner-Chloros, P., 2009. Code-Switching. University Press, Cambridge.

Gardner-Chloros, P., Edwards, M., 2004. Assumptions behind grammatical approaches to code-switching: when the blueprint is a red herring. Transactions of the Philological Society 102 (1), 103–129.

Ghafar-Samar, R., Meechan, M., 1998. The null theory of codeswitching versus the nonce borrowing hypothesis: testing the fit in Persian-English bilingual discourse. International Journal of Bilingualism 2 (2), 203–220.

Gumperz, J.J., 1976/1982. Conversational code-switching. In: Gumperz, J.J. (Ed.), Discourse Strategies. Cambridge University Press, Cambridge, pp. 59–99.

Gullberg, M., Indefrey, P., Muysken, P., 2009. Research techniques for the study of code-switching. In: Bullock, B.E., Toribio, A.J. (Eds.), The Cambridge Handbook of Linguistic Code-Switching. Cambridge University Press, Cambridge, pp. 21–39.

Haspelmath, M., Tadmor, U. (Eds.), 2009. Loanwords in the World's Languages: A Comparative Handbook. Walter de Gruyter, Berlin/New York.

Joshi, A.K., 1985. Processing of sentences with intrasentential code-switching. In: Dowty, D.R., Karttunen, L., Zwicky, A.M. (Eds.), Natural Language Parsing: Psychological, Computational and Theoretical Perspectives. Cambridge University Press, Cambridge, pp. 190–204.

Klavans, J.L., 1985. The syntax of code-switching: Spanish and English. In: King, L.D., Maley, C.A. (Eds.), Selected Papers from the XIIIth Linguistic Symposium on Romance Languages. Benjamins, Amsterdam, pp. 213–221.

Labov, W., 1969. Contraction, deletion, and inherent variability of the English copula. Language 45 (4), 715–762.

Labov, W., 1972. Sociolinguistic Patterns. University of Pennsylvania Press, Philadephia.

Legendre, G., Schindler, M., 2010. Code switching in Urban Wolof: a case for violable constraints in syntax. Revista virtual de estudos em linguistica (ReVEL) 8 (Special Issue number 4), 47–75 [www.revel.inf.br/eng].

Lipski, J., 1977. Code-switching and the problem of bilingual competence. In: Paradis, M. (Ed.), Aspects of Bilingualism. Hornbeam Press, Columbia, SC, pp. 250–264.

Mahootian, S., 1993. A Null Theory of Codeswitching (Doctoral dissertation). Northwestern University.

Mahootian, S., Santorini, B., 1996. Code switching and the complement/adjunct distinction. Linguistic Inquiry 27 (3), 464–479.

MacSwan, J., 1999. A Minimalist Approach to Intrasentential Code Switching. Garland Press, New York.

MacSwan, J., 2009. Generative approaches to code-switching. In: Bullock, B.E., Toribio, A.J. (Eds.), The Cambridge Handbook of Linguistic Code-Switching. Cambridge University Press, Cambridge, pp. 309–335.

Meechan, M., Poplack, S., 1995. Orphan categories in bilingual discourse: adjectivization strategies in Wolof-French and Fongbe-French bilingual discourse. Language Variation and Change 7 (2), 169–194.

Muysken, P.C., 2000. Bilingual Speech: A Typology of Code-Mixing. Cambridge University Press, Cambridge, MA.

Myers-Scotton, C., 1993. Duelling Languages. Clarendon Press, Oxford.

Myers-Scotton, C., 2002. Contact Linguistics: Bilingual Encounters and Grammatical Outcomes. University Press, Oxford.

Myers-Scotton, C., 2006. Multiple Voices: An Intro to Bilingualism. Blackwell Publishers, Malden, MA. Lexical borrowing, pp. 208–232 (Chapter 8).

Naït M'Barek, M., Sankoff, D., 1988. Le discours mixte Arabe/Français: des emprunts ou des alternances de langue? Revue Canadienne de Linguistique 33 (2), 143–154.

Nortier, J., 1989. Dutch and Moroccan Arabic in Contact: Code-Switching Among Moroccans in the Netherlands (Doctoral dissertation). University of Amsterdam.

Pfaff, C.W., 1979. Constraints on language mixing. Language 55 (2), 291–318.

Poplack, S., 1980. Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching. Linguistics 18 (7/8), 581–618.

Poplack, S., 1981. Syntactic structure and social function of code-switching. In: Duran, R.P. (Ed.), Latino Discourse and Communicative Behaviour. Ablex Publishing Corporation, New Jersey, pp. 169–184.

Poplack, S., 1987. Contrasting patterns of code-switching in two communities. In: Wande, E., Anward, J., Nordberg, B., Steensland, L., Thelander, M. (Eds.), Aspects of Bilingualism: Proceedings from the Fourth Nordic Symposium on Bilingualism, 1984. Borgström, Motala, Uppsala, pp. 51–76.

Poplack, S., 1993. Variation theory and language contact. In: Preston, D. (Ed.), American Dialect Research: An Anthology Celebrating the 100th Anniversary of the American Dialect Society. Benjamin, Amsterdam, pp. 251–286.

Poplack, S., 2012. What does the Nonce Borrowing Hypothesis hypothesize? Bilingualism: Language and Cognition 15 (3), 644–648.
Poplack, S., Dion, N., 2012. Myths and facts about loanword development. Language Variation and Change 24 (3), 279–315.
Poplack, S., Meechan, M., 1995. Patterns of language mixture: nominal structure in Wolof-French and Fongbe-French bilingual discourse. In: Milroy, L., Muysken, P. (Eds.), One Speaker, Two Languages. Cambridge University Press, Cambridge, pp. 199–232.
Poplack, S., Meechan, M. (Eds.), 1998a. Instant Loans, Easy Conditions: The Productivity of Bilingual Borrowing. International Journal of Bilingualism 2 (2) (Special issue).
Poplack, S., Meechan, M., 1998b. How languages fit together in code-mixing. International Journal of Bilingualism 2 (2), 127–138.
Poplack, S., Sankoff, D., Miller, C., 1988. The social correlates and linguistic processes of lexical borrowing and assimilation. Linguistics 26 (1), 47–104.
Poplack, S., Wheeler, S., Westwood, A., 1987. Distinguishing language contact phenomena: evidence from Finnish-English bilingualism. In: Lilius, P., Saari, M. (Eds.), The Nordic Languages and Modern Linguistics. University of Helsinki Press, Helsinki, pp. 33–56.
Rubin, E.J., Toribio, A.J., 1995. Feature-checking and the syntax of language contact. In: Amastae, J., Goodall, G., Montalbetti, M., Phinney, M. (Eds.), Contemporary Research in Romance Linguistics. John Benjamins, Amsterdam and Philadelphia, pp. 177–185.
Sankoff, D., 1998a. A formal production-based explanation of the facts of code-switching. Bilingualism: Language and Cognition 1 (1), 39–50.
Sankoff, D., 1998b. The production model of code-mixed discourse. In: COLING-ACL '98: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics. University of Montreal Press, Montreal, pp. 8–21.
Sankoff, D., Poplack, S., 1981. A formal grammar for code-switching. Papers in Linguistics: International Journal of Human Communication 14 (1), 3–45.
Sankoff, D., Poplack, S., Vanniarajan, S., 1990. The case of the nonce loan in Tamil. Language Variation and Change 2 (1), 71–101.
Schindler, M., Legendre, G., Mbaye, 2008. Violations of the PF interface condition in Urban Wolof. Proceedings from the Annual Meeting of the Chicago Linguistic Society 44 (2), 169–184.
Shin, S., 2002. Differentiating language contact phenomena: evidence from Korean-English bilingual children. Applied Psycholinguistics 23, 337–360.
di Sciullo, A.-M., Muysken, P., Singh, R., 1986. Government and code-mixing. Journal of Linguistics 22 (1), 1–24.
Timm, L.A., 1975. Spanish-English code-switching: el porque y how-not-to. Romance Philology 28 (4), 473–482.
Torres Cacoullos, R., Aaron, J.E., 2003. Bare English-origin nouns in Spanish: rates, constraints, and discourse functions. Language Variation and Change 15 (3), 287–326.
Torres Cacoullos, R., Travis, C.E., 2015. Gauging convergence on the ground: code-switching in the community. In: Travis, C.E., Torres Cacoullos, R. (Eds.), International Journal of Bilingualism, Special issue.
Torres Cacoullos, R., Travis, C.E. Two languages, one effect: structural priming in code-switching. In: Deuchar, M. (Ed.), Bilingualism: Language and Cognition (Special issue), to appear.
Treffers-Daller, J., 1994. Mixing Two Languages: French–Dutch Contact in a Comparative Perspective. Mouton de Gruyter, Berlin/New York.
Turpin, D., 1998. "Le français c'est le last frontier": the status of English-origin nouns in Acadian French. International Journal of Bilingualism 2 (2), 221–233.
Weinreich, U., 1953/1968. Languages in Contact. Mouton, The Hague.
Winford, D., 2009. On the unity of contact phenomena and their underlying mechanisms: the case of borrowing. In: Isurin, L., Winford, D., de Bot, K. (Eds.), Multidisciplinary Approaches to Code-Switching. John Benjamins, Amsterdam/Philadelphia, pp. 279–306.
Woolford, E., 1983. Bilingual code-switching and syntactic theory. Linguistic Inquiry 14 (3), 519–536.

Coeducation and Single-Sex Schooling

Sum Kwing Cheung, Early Childhood and Elementary Education Division, School of Continuing Education, Hong Kong Baptist University, HKSAR, China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In many countries, coeducation schooling is predominant today, though single-sex schooling was once the norm in history. From the 1960s to 2010s, dozens of studies have been conducted to compare the relative effects of single-sex to coeducation schooling on students' academic- and career-related outcomes, as well as affective and social development. At its root, the controversy centers on the issue of gender equity in education. Despite its abundant literature, inconsistent findings were observed, and some findings were confounded by selection and school quality effects. Future studies can therefore investigate more carefully in which context single-sex schooling can be effective.

Coeducation and Single-Sex Schooling

Coeducation schooling, also called mixed-sex schooling, is the practice of educating males and females together. Single-sex schooling, on the other hand, refers to a form of education in which the two sexes are separated when teaching is conducted. It can be implemented at the school level (i.e., all-girls and all-boys schools) or classroom level only (i.e., single-sex classes for some or all subjects within coeducational schools) (Wiseman, 2008).

Among past literature, the debate on coeducation vs single-sex schooling often centers on the issue of how gender equity can be achieved in education. In the following, I would first describe the prevalence of coeducation and single-sex schooling around the world, followed by an outline of research trends in the literature on coeducation and single-sex schooling. Then, I would review empirical studies on the relative effects of single-sex to coeducation schooling on students' academic- and career-related outcomes, as well as affective and social development. Finally, I would conclude the article with a discussion on the limitations of existing studies and directions for future research in the field.

Prevalence of Coeducation and Single-Sex Schooling

Over the history of many countries, single-sex schooling had once been the norm. Nonetheless, its dominant position was later replaced by coeducation schooling. For example, in the United States, formal schooling was offered exclusively to boys at the beginning (Bracey, 2006). Later, during the post-Revolutionary War years, in response to the call for more systematic schooling for females by the public, girls started to have the chance to attend elementary schools in a sex-segregated manner (Salomone, 2003). In the mid-nineteenth century, common schooling for the masses began to take place; and coeducation schooling gradually became widespread, because it was often more economic to run schools for boys and girls to attend together than for a particular sex to attend (Anfara and Mertens, 2008; Shmurak, 1998). In 1972, with the passage of Title IX legislation, the setup of new single-sex public schools and classes was restricted (Bracey, 2006). Since then, single-sex schooling retained in the private and denominational sector only (Anfara and Mertens, 2008). Similarly, in England, receiving education at school was the privilege of boys of the wealthy in the early nineteenth century; and more schools for girls were established starting from the late nineteenth century (Leonard, 2006). However, unlike the case of the United States, the introduction of mass education in the nineteenth century led to the establishment of many single-sex schools (Leonard, 2006). It was only until the mid-twentieth century, when many selective secondary schools were replaced by large comprehensive schools, coeducation schooling began to become pervasive in England (Leonard, 2006).

Nowadays, coeducation schooling remains predominant in many parts of the world. Single-sex schooling occurs, but its prevalence varies considerably across countries. For example, in a study on the prevalence of sex-segregated eighth grade mathematics classrooms across 46 nations, Wiseman (2008) found that the level of sex segregation was nil in 12 nations (e.g., Sweden, Norway, Romania), very low (about 1–10% of schools) in 17 nations (e.g., the United States, South Africa, Japan, the Netherlands), moderate (about 15–20% of schools) in seven nations (e.g., Malaysia, Australia, Singapore, England), and high (about 25–100% of schools) in 10 nations (e.g., Israel, New Zealand). In the high sex-segregated group, eight nations (e.g., Egypt, Jordan, Iran, Saudi Arabia) even had sex segregation in more than half of the schools (Wiseman, 2008). In general, these variations in the rates of single-sex schooling arose from a wide array of sociocultural, economic, and political factors (Wiseman, 2008). For example, in countries like Australia, Singapore, and New Zealand, the practice of single-sex schooling was the legacy of colonialism and extensive borrowing of educational policies when implementing mass education (Wiseman, 2008). Nevertheless, it should be noted that the figures provided by Wiseman (2008) were only rough estimates of the prevalence of single-sex schooling in the aforementioned nations, because the percentage of single-sex schooling might vary by the level of education.

In recent years, single-sex schooling begins to return back to the educational landscape in the United States; and there have been a growing number of single-sex classes and schools (Hayes et al., 2011). The major reason is that the Title IX regulations had been amended in 2006, and single-sex education opportunities are now permitted in public schools

(Hayes et al., 2011). Therefore, it is generally expected that single-sex schooling would continue to blossom in the United States in the next decade (Hayes et al., 2011).

Trends in Research on the Effects of Coeducation and Single-Sex Schooling

As proposed by the National Association for Single Sex Public Education (NASSPE) in 2013, research studies that compared the effects of coeducation with single-sex schooling can broadly be classified into three types. The first type is nationwide studies, which involved samples of thousands of students or above (NASSPE, 2013). One example is the study on 2954 high school students' academic attainments in England conducted by Spielhofer, O'Donnell, Benton, Schagen, and Schagen from the National Foundation for Educational Research team in 2002. To address the problem that students are not randomly assigned to attend coeducation and single-sex schools, some researchers would develop samples for their research studies by randomly selecting schools or students from these large databases (Salomone, 2003). The second type is 'before and after' studies, which compared students' performance before and after the introduction of single-sex classes in particular schools (NASSPE, 2013). Since the same schools are involved when making comparisons, the effects of extraneous variables related to school characteristics are supposed to be eliminated in this type of studies (NASSPE, 2013). Streitmatter (1999, as cited in Mael et al., 2005), for example, had published a study of this type to investigate female students' attitudinal changes after they had participated in a single-sex class in coeducational public schools. The third type of studies is smaller-scale scientific studies that are commonly found in academic journals (NASSPE, 2013). Though the pool of participants involved in these studies was relatively smaller as compared to those involved in nationwide studies, researchers of these studies would adopt a number of statistical techniques to control the effects of extraneous variables (e.g., student–teacher ratios, school climate) when conducting data analyses (NASSPE, 2013; Salomone, 2003).

In the mid-nineteenth century, when coeducation schooling began to spread throughout the United States, there was an emerging controversy over the advantages and disadvantages of coeducation schooling (Shmurak, 1998). Feminists at that time generally supported the implementation of coeducation schooling because they feared that gender equality could not be achieved if separate schools were run for boys and girls (Shmurak, 1998). Other proponents of coeducation schooling also suggested that with the presence of the members of the other sex, boys showed better behaviors and girls became more independent (Shmurak, 1998). Middle- and upper-class parents, however, were reluctant to send their daughters to coeducation schools, because they did not want their daughters to be in touch with sons of the lower classes (Shmurak, 1998). Despite their concerns, the majority was inclined to coeducation schooling (Shmurak, 1998).

In the late 1960s and early 1970s, the feminist movement sparked off a debate on the advantages and disadvantages of coeducation schooling, which sustained till late 1980s and early 1990s (Shmurak, 1998). On the one hand, proponents of coeducation schooling advocated that coeducation schooling provided a happier learning atmosphere and promoted optimal adjustment to life (Gill, 2004), as students from coeducation schools were found to have more pleasurable social contacts and better interactions with members of the other sex (Gill, 2004; Leonard, 2006; Mael et al., 2005). On the other hand, feminists criticized that girls were often disadvantaged in coeducation schooling environments (Gill, 2004; Shmurak, 1998). As support for their claim, they demonstrated that in mixed-sex classrooms, girls received less attention from teachers, acted as school captains less often, and involved less in team sports (Gill, 2004). Girls were also found to receive sex-stereotyped advice from counselors, which made them less likely to enroll in mathematics and science courses (Gill, 2004; Shmurak, 1998).

In the meantime, in view of the situation that gender bias prevailed in coeducational classrooms, feminists began to regard single-sex schooling as a solution of achieving gender equity in education (Shmurak, 1998). From 1980s to 1990s, there was an increasing amount of research evidence demonstrating that single-sex schooling had some positive impacts on girls. For example, Lee and Bryk (1986, 1989, as cited in Shmurak, 1998) analyzed data from a longitudinal study conducted in the United States and found that girls from all-girls schools showed higher level of interest in mathematics and educational aspirations than their peers from coeducation schools. Riordan (1990, as cited in Shmurak, 1998) also found that girls from all-girls school performed better than in four curriculum-specific tests. However, there were some critics that Lee and Bryk did not control adequately the effects of background variables during their data analyses (Shmurak, 1998). They argued that after taking consideration of the preexisting differences among students, the achievement attainment gaps observed between girls from coeducation and single-sex schools would diminish (Salomone, 2003).

Since the late 1990s, there was an upsurge of concerns in countries like the United States, England, and Australia that boys were somehow being marginalized in the current schooling practices (Gill, 2004). Despite that some boys excelled well in coeducation schools, the dropout rate was higher among boys than girls; and teachers often complained about the disruptive behaviors of boys (Gill, 2004). Single-sex schooling is therefore regarded by some researchers (e.g., Gurian et al., 2009), a means to save boys from their crisis.

Effects of Coeducation and Single-Sex Schooling on Academic- and Career-Related Outcomes

The arguments put forward to support single-sex schooling primarily rest on two rationales (Hayes et al., 2011). The first rationale is the belief that there are some substantial differences in the biological predispositions between the two sexes; therefore, single-sex schooling environment can allow teachers to adopt more effective instructional strategies that best fit the learning abilities and styles of males and females, respectively (Hayes et al., 2011). Gurian et al. (2009), for example, suggested that teachers should provide more movement-based learning opportunities when educating boys because boys tend to have higher energy levels. On the other hand, given that

it is often very important for girls to feel belonged, teachers can create more collaborative and cooperative learning opportunities when educating girls (Gurian et al., 2009). Nevertheless, some researchers argued that innate differences between the two sexes in learning abilities and styles were negligible (Hayes et al., 2011). The second rationale is the belief that compared to coeducation schooling, single-sex schooling can provide an environment that is more conducive to the learning for the two sexes (Hayes et al., 2011). For example, there was some empirical evidence that in coeducation classrooms, boys tended to receive most of the teachers' attention whereas girls were often viewed and treated in a sexist manner by their teachers and peers (Hayes et al., 2011). All these factors, in turn, could have detrimental effects on girls' achievement attainments. Moreover, with the absence of members of the opposite sex, there were neither sexual tensions nor attractions in the classroom; and students could thus concentrate more on their learning (Bracey, 2006).

Over the last few decades, a number of research studies had been conducted to compare the effects of coeducation and single-sex schooling on students' academic- and career-related outcomes. Mixed results, however, were observed. Whereas some studies demonstrated that students from single-sex schools had better academic- and career-related outcomes than their peers from coeducation schools, some studies reported findings in the exactly opposite direction. Furthermore, there were some studies showing that the sex composition of the school did not affect students' academic- and career-related outcomes. In the remaining of the section, I would discuss the relations of coeducation and single-sex schooling to different outcomes of students in the academic and career aspects by making reference to the results of previous empirical studies.

The first area that often receives researchers' attention is students' attitudes toward different academic subjects and their subject choices (Gill, 2004). Advocates of single-sex schooling criticized that in coeducation schools, subjects were often polarized into the domains of males and females (e.g., mathematics, science, and technology as male subjects; language and history as female subjects) (Gill, 2004). This might, in turn, lower students' interest in subjects that belong to the domain of their own sex and their likelihood in choosing those subjects (Gill, 2004). Basically, a number of empirical studies provided support for this claim by demonstrating the effects of single-sex schooling on countering students' sex-stereotyped beliefs toward academic subjects. For example, there was evidence that girls who had attended single-sex science classes for a year showed more positive attitudes to science than girls who had only attended coeducation science classes (Gill, 2004). Some research findings also suggested that compared to girls from coeducation schools, girls from single-sex schools showed greater interest in mathematics (Shmurak, 1998). Furthermore, they were less likely to regard English as female subject and science as male subject, and were more likely to enroll in physics, chemistry, or biology (Shmurak, 1998; Spielhofer et al., 2002). Insignificant results, however, were sometimes reported. One example was the large-scale Australian study conducted by Ainley and Daly (2002, as cited in Gill, 2004), which did not find any significant differences between students from single-sex and coeducation schools in enrolling in physics courses. Herein, it should be noted that the presence or absence of differences in students' subject choice might sometimes be confounded by the subject-taking patterns offered by the school (Gill, 2004).

Another area that is often addressed when comparing the effects of single-sex and coeducation schooling is students' academic performance, especially girls' performance in subjects that are traditionally regarded as masculine (e.g., mathematics, science). Single-sex schooling, according to its advocates, can help improve students' educational outcomes, because pedagogical methods that are favored by a particular sex group can easily be implemented when the classroom is sex-segregated (Bracey, 2006). Nonetheless, empirical findings in this area were rather inconsistent. On the one hand, some classic studies conducted in the United States showed that compared to girls from coeducation Catholic schools, girls from single-sex Catholic school showed better academic achievements (Shmurak, 1998). On the other hand, Daly (1996, as cited in Shmurak, 1998) found that in Northern Ireland, girls from single-sex and coeducation schools did not differ in their scores on state examination in English and mathematics. In the United States, Shmurak (1998) also could not find any significant differences in SAT (a standardized test that is commonly used for admission to American colleges) between girls from single-sex and coeducation schools. Similarly, in New Zealand, Harker (2000, as cited in Mael et al., 2005) found that high school students from single-sex and coeducation schools did not differ in their scores on mathematics and science achievement tests. The findings of Spielhofer et al. (2002) in England revealed more complicated relations between the type of schooling and students' academic achievement. Girls in single-sex comprehensive schools have better academic achievements than those from coeducation comprehensive schools (Spielhofer et al., 2002). In contrast, boys from single-sex and mixed-sex comprehensive schools did not differ in their academic performance, though boys from single-sex grammar schools had significantly better academic performance than boys from coeducation grammar schools (Spielhofer et al., 2002).

The third area that researchers in the field focus on is students' classroom behaviors and general attitudes to school. Proponents of single-sex schooling often criticize that in coeducation classrooms, sexism is common during teacher–student as well as peer interactions (Bracey, 2006). In consequence, girls may suffer from a poor learning environment (Bracey, 2006). Such claims, however, are partially supported as well as rejected by past research findings. Apparently, girls could behave in a less sex-stereotypical manner in single-sex classrooms, but this did not necessarily lead to more favorable attitudes toward school. For example, though there was some research evidence that girls in single-sex classrooms displayed more risk-taking behaviors and asked more questions than girls in coeducation classrooms (Anfara and Mertens, 2008), Gill (1996, as cited in Shmurak, 1998) could not find any significant differences between girls from single-sex and coeducation schools in their attitudes toward school. Some research studies even found that students from coeducation schools held more positive attitudes to school than their peers from single-sex schools (Gill, 2004; Shmurak, 1998). One reason given by students from coeducation school was that they found interactions with opposite-sex members make school experiences more fun (Shmurak, 1998).

The last area that is of the interest of researchers in the field is students' educational and career aspirations. Supporters of single-sex schooling propose that in all-girls school, girls can get rid of sex stereotypes; and they are allowed to develop more ambitions in their educational and career goals (Bracey, 2006). Among past research findings, there are again no clear patterns. Though girls from single-sex schools were found to set higher educational goals to attain than girls from coeducation schools in some studies, such differences were not observed in some other studies (Mael et al., 2005; Shmurak, 1998). Shmurak (1998) even found in her longitudinal study that compared to girls who studied in single-sex schools before, girls who studied in coeducational schools before got into colleges at higher rank. In regard to career aspirations, some research studies showed that compared to girls from coeducation, girls from single-sex schools perceived the roles of women in the workplace in a less stereotyped manner and set higher yet more realistic career goals (Mael et al., 2005; Shmurak, 1998). Shmurak (1998), however, found that there were no significant differences between girls from single-sex and coeducation schools in their aspiration to careers that were traditionally male-dominated. Hoffnung (2011, as cited in Bigler and Signorella, 2011) also showed that graduates of all-girls and coeducation colleges did not differ in their attitudes toward women, career aspirations, marital timing as well as family planning.

In sum, though there are some studies documenting the relative merits of single-sex over coeducation schooling on students' academic- and career-related outcomes, there seems to be no clear evidence that single-sex schooling is more superior than coeducation schooling all the times in every single aspect. The positive effects of single-sex schooling on boys are also rarely stated.

Effects of Coeducation and Single-Sex Schooling on Affective and Social Development

Compared to the vast number of research studies that addressed the effects of coeducation and single-sex schooling on academic- and career-related outcomes, the literature on the effects of coeducation and single-sex schooling on affective and social development is rather sparse (Bigler and Signorella, 2011). Though theoretical arguments on the advantages of each type of schooling on students' affective and social development have sometimes been put forth, not all of the arguments have been validated by empirical studies (Bigler and Signorella, 2011). Among the arguments on the effects of coeducation and single-sex schooling on students' affective and social development that had been tested, just like the case of academic- and career-related outcomes, inconsistent findings were yielded. Below, I would discuss the arguments and empirical findings that are in support for each type of schooling in accordance with the area of concern they addressed.

The first issue that draws the attention of many researchers in the field is students' self-esteem. Claiming that sexism is prevalent in coeducation classrooms, proponents of single-sex schooling argued that students, in particular girls, can have higher self-esteem in sex-segregated classrooms (Gurian et al., 2009). Results of past research findings, however, are rather mixed. Some studies showed that students from single-sex schools reported higher levels of self-esteem than their peers from coeducation schools; whereas some other studies, including both cross-sectional and longitudinal studies, could not find any significant differences in the self-esteem level of girls from single-sex and coeducation schools (Anfara and Mertens, 2008; Shmurak, 1998). Some researchers thus proposed that significant relations between the type of schooling and students' self-esteem, if there was any, were more likely to be attributed to factors like students' academic abilities and socioeconomic status (Shmurak, 1998).

The second area that arouses researchers' interest is students' gender identity. Advocates of single-sex schooling suggested that girls can have more leadership opportunities in all-girls school, which in turn might help girls develop assertiveness (Campbell and Sanders, 2002; Shmurak, 1998). Moreover, some girls found that they could be themselves by the absence of boys in the classrooms (Shmurak, 1998). The study conducted by Sadker and Sadker (1994, as cited in Hayes et al., 2011), and Shmurak (1998) provided some empirical support for the claims. However, in the study of 6427 third-year secondary school students in Belgium by Brutsaert (1999), compared to girls from single-sex schools, girls from coeducation schools were found to be more likely to describe themselves as possessing more feminine as well as masculine traits.

The third issue that researchers center on is students' body image. Though advocates of single-sex schooling proposed that sexist incidents were less likely to occur in sex-segregated classrooms (Bracey, 2006), single-sex environments could promote more rigid definitions of 'masculinity' and 'femininity' in some aspects (Bigler and Signorella, 2011). One example of these aspects was body image. Some past research findings suggested that compared to girls from coeducation schools, girls from single-sex high schools were more dissatisfied with their body type, displayed more eating disorder behaviors, and showed a stronger desire for thinness (Mael et al., 2005; Shmurak, 1998). Davey, Jones, and Harris (2011, as cited in Bigler and Signorella, 2011), however, found that girls from single-sex and coeducation schools did not report any significant differences in their indices of body dissatisfaction and eating disorders, though the ideal body shape endorsed by girls from single-sex schools was thinner than that endorsed by girls from coeducation schools.

The last area of concern of researchers in the field is students' interpersonal relations. Proponents of coeducation schooling proposed that mixed-sex learning environments can promote 'optimal adjustment to life' because students would have experiences in interacting with members of the opposite sex (Gill, 2004). Several studies provided empirical support for this argument. For example, Harris (1986, as cited in Gill, 2004) showed that compared to graduates of single-sex schools, coeducation school graduates found interaction with opposite-sex members in the university easier. Foon (1988, as cited in Cheung and McBride-Chang, 2007) also found that coeducation school students were more affiliated with their peers than single-sex school students. Furthermore, in a study on Hong Kong 10th grade students' cross-sex friendship patterns by Cheung and McBride-Chang (2007), coeducation school students were found to have higher levels of friendship satisfaction toward their best cross-sex friends than single-sex school

students did. On the other hand, Brutsaert (2006) found that compared to feminine boys in coeducation schools, feminine boys in single-sex schools were less accepted by their peers.

In short, the literature on the effects of single-sex and coeducation schooling on students' affective and social development was not so abundant. Despite this, the empirical findings were not conclusive. There is some evidence of the merits of coeducation as well as single-sex schooling. **Table 1** summarized the relative merits of coeducation and single-sex schooling mentioned in the literature reviewed above.

Rethinking the Research Findings of the Effects of Coeducation and Single-Sex Schooling

Hayes et al. (2011) remarked that empirical studies that compared the effects of coeducation and single-sex schooling rarely employed random assignment of participants into the two types of school programs. In other words, many studies in the literature could not provide direct tests of the causal relations of the two types of school programs to the developmental outcomes of their students (Hayes et al., 2011). Their results should therefore be interpreted with caution, because they might be confounded by the following two effects.

The first one is the selection effects (Hayes et al., 2011). On the one hand, students who choose to study in single-sex schools may be more achievement-oriented than those who opt to study in coeducations schools (Hayes et al., 2011). On the other hand, the coeducation and single-sex schools under investigation may set different admission criteria (Hayes et al., 2011). As a result, the characteristics of students admitted to single-sex schools may be qualitatively different from those admitted to coeducation schools. For example, in England and Australia, single-sex schools were usually private schools with higher academic standards (Anfara and Mertens, 2008). They might therefore attract brighter students from more privileged backgrounds.

Table 1 Summary of the relative merits of coeducation and single-sex schooling mentioned in past literature

Type of schooling	*Relative merits*
Coeducation schooling	● Students held more positive attitudes toward school (Gill, 2004; Shmurak, 1998), showed more affiliation with their peers (Foon, 1988, as cited in Cheung and McBride-Chang, 2007), and found their best cross-sex friendship more satisfying (Cheung and McBride-Chang, 2007). Graduates found interaction with opposite-sex members easier (Harris, 1986, as cited in Gill, 2004). ● In particular, girls possessed more feminine and masculine traits (Brutsaert, 1999), felt more satisfied with their body type, displayed fewer eating disorder behaviors, and showed less desire for thinness (Mael et al., 2005; Shmurak, 1998).
Single-sex schooling	● Students held fewer sex-stereotyped beliefs as well as attitudes toward academic subjects (Gill, 2004; Shmurak, 1998), showed better academic performance (Shmurak, 1998; Spielhofer et al., 2002), and showed higher self-esteem (Anfara and Mertens, 2008; Shmurak, 1998). ● In particular, girls behaved in a less sex-stereotypical manner in classroom (Anfara and Mertens, 2008), set higher educational and career goals (Mael et al., 2005; Shmurak, 1998), and could be themselves more self-esteemed. (Sadker and Sadker, 1994, as cited in Hayes et al., 2011; Shmurak, 1998).

Note: Given that mixed results were yielded among past studies examining the effects of coeducation and single-sex schooling, the above merits were inconclusive and should be interpreted with great caution (e.g., by taking account of the context in which the studies were conducted).

The second one is the school quality effects (Hayes et al., 2011). Apart from the sex composition of students, the coeducation and single-sex schools under investigation in a study may vary greatly in other aspects of the school (e.g., academic standard, class size) (Bracey, 2006; Hayes et al., 2011). For example, Hayes et al. (2011) commented that many studies conducted in the United States involved comparison between students enrolled in private single-sex schools and public coeducational schools. However, private and public schools in the United States differed greatly in terms of their financial resources and other school characteristics (e.g., academic standards, class size) (Hayes et al., 2011). Thus, any observed differences in student outcomes can be caused by factors other than the sex composition of the school. For example, there were some speculations that the success of the California's experiment with single-sex schooling should, in fact, be attributed to the positive changes made in curriculum and instruction rather than sex composition of the school only (Anfara and Mertens, 2008).

Moreover, when interpreting research findings on the effects on single-sex and coeducation schooling, one should not generalize any observed effects too much. Instead, one should always pay attention to the context (including both the cultural and chronological contexts) in which the studies were conducted, as well as the educational level of the participants involved (i.e., preschool, elementary, secondary, or university). Given that the schooling system varies across countries and different countries have different cultural values, effects found in single-sex schools in the United States, for example, may not be necessarily existed in single-sex schools in other countries. Following the same reasoning, since some societal values have been evolved over the past few decades, it is possible that significant effects of single-sex schooling observed in 1980s, for instance, may magnify or diminish now. Similarly, one cannot assume that students at different ages have unique developmental characteristics, one cannot assume that the effectiveness of single-sex schooling on secondary education, for example, also applies in other levels of education.

To conclude, among past empirical findings, there is no clear evidence whether coeducation or single-sex schooling is a dominantly better option than the other. This perhaps suggests that researchers should stop asking the simple question, whether coeducation or single-sex schooling should be adopted. Instead, researchers should move a step further by investigating the interaction of the effects of coeducation and single-sex schooling with various contextual variables (Shmurak, 1998). In view of the above point, three possible directions for future studies are listed below.

First, in light of the mixed findings in the literature, researchers should admit that there are both advantages and

disadvantages for each type of schooling. To move on, researchers can investigate the relative benefits to the costs for each type of schooling in both the short and long run (Shmurak, 1998). Second, researchers can investigate which types of students can be benefited more from each type of schooling (Shmurak, 1998). Examples of student characteristics that may be relevant include students' academic abilities, socioeconomic status, and possessions of gender-related personality traits. Last, if single-sex schooling is thought to have some potential benefits of students, researchers can investigate how single-sex schooling can be implemented so as to maximize their benefits (Shmurak, 1998). For example, researchers can take into the consideration of the school or class size that is favorable for implementing single-sex schooling. The quality of teachers (including their beliefs, attitudes, and abilities to use various types of instructional strategies) may also be crucial factors for the success of single-sex schooling.

See also: Academic Self-Concept and Achievement; Education (Primary and Secondary Schools) and Gender; Educational Institutions, History of; Educational Systems: North America; Equity and Education; Gender and Education; Sex Role Development and Education; Teacher Behaviours and Student Outcomes.

Bibliography

Anfara, V.A., Mertens, S.B., 2008. Do single-sex classes and schools make a difference? Middle School Journal 40, 52–59.

Bigler, R.S., Signorella, M.L., 2011. Single-sex education: new perspectives and evidence on a continuing controversy. Sex Roles 65, 659–669.

Bracey, G.W., 2006. Separate but Superior? A Review of Issues and Data Bearing on Single-Sex Education. Educational Policy Research Unit (EPRU), Tempe.

Brutsaert, H., 1999. Coeducation and gender identity formation: a comparative analysis of secondary schools in Belgium. British Journal of Sociology of Education 20, 343–353.

Brutsaert, H., 2006. Gender-role identity and perceived peer group acceptance among early adolescents in Belgian mixed and single-sex schools. Gender and Education 18, 635–649.

Campbell, P.B., Sanders, J., 2002. Challenging the system: assumptions and data behind the push for single-sex schooling. In: Datnow, A., Hubbard, L. (Eds.), Gender in Policy and Practice: Perspectives on Single-Sex and Coeducational Schooling. RoutledgeFalmer, New York, pp. 31–46.

Cheung, S.K., McBride-Chang, C., 2007. Correlates of cross-sex friendship satisfaction in Hong Kong adolescents. International Journal of Behavioral Development 31, 19–27.

Gill, J., 2004. Beyond the Great Divide: Single Sex or Coeducation? UNSW Press, Sydney, NSW.

Gurian, M., Stevens, K., Daniels, P., 2009. Successful Single-Sex Classrooms: A Practical Guide to Teaching Boys and Girls Separately. Jossey-Bass, San Francisco, CA.

Hayes, A.R., Pahlke, E.E., Bigler, R.S., 2011. The efficacy of single-sex education: testing for selection and peer quality effects. Sex Roles 65, 693–703.

Leonard, D., 2006. Single-sex schooling. In: Skelton, C., Francis, B., Smulyan, L. (Eds.), The Sage Handbook of Gender and Education. Sage Publications, Thousand Oaks, CA, pp. 190–204.

Mael, F., Alonso, A., Gibson, D., Rogers, K., Smith, M., 2005. Single-Sex versus Coeducational Schooling: A Systematic Review (Report No. 2005-01). U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, Washington, DC. Retrieved from: http://www2.ed.gov/rschstat/eval/other/single-sex/single-sex.pdf.

National Association for Single Sex Public Education, 2013. Single-Sex vs. Coed: The Evidence. http://www.singlesexschools.org/research-brain.htm/.

Salomone, R.C., 2003. Same, Different, Equal: Rethinking Single-Sex Schooling. Yale University Press, New Haven.

Shmurak, C.B., 1998. Voices of Hope: Adolescent Girls at Single Sex and Coeducation Schools. P. Lang, New York.

Spielhofer, T., O'Donnell, L., Benton, T., Schagen, S., Schagen, I., 2002. The Impact of School Size and Single-Sex Education on Performance. The National Foundation for Educational Research, Slough.

Wiseman, A.W., 2008. A culture of (in)equality? a cross-national study of gender parity and gender segregation in national school systems. Research in Comparative and International Education 3, 179–201.

Relevant Websites

http://www.easse.org/eng/index.htm – European Association for Single-Sex Education.

http://www.singlesexschools.org/home-introduction.htm – National Association for Single Sex Public Education.

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 4

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser–Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 4

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 4 *xlv*

VOLUME 4

Cognition, Evolution of 1
Christopher D Peterson and Harold C Barrett

Cognition: Knowledge Spaces 6
Jean-Paul Doignon

Cognitive Aging 10
K Warner Schaie

Cognitive Anthropology 16
Jürg Wassmann and Andrea Bender

Cognitive Archaeology 23
Marc A Abramiuk

Cognitive Behavioral Therapy 30
David Kingdon and Helen Mander

Cognitive Control (Executive Function): Role of Prefrontal Cortex 33
Jonathan D Cohen

Cognitive Development: Child Education 38
Zhenlin Wang and Lamei Wang

Cognitive Development During Infancy and Early Childhood across Cultures 43
Olivier Houdé

Cognitive Development, Educational Theories of 51
Hafiz M Iqbal

Cognitive Development in Infancy, Neural Mechanisms of 58
Laura A Edwards

Cognitive Development: Mathematics Learning and Instruction 66
Clarissa A Thompson and Julie L Booth

Cognitive Dissonance 76
Joel Cooper and Kevin M Carlsmith

Cognitive Maps 79
Rob Kitchin

Cognitive Modeling: Connectionist Approaches 84
Bruce MacLennan

Cognitive Neuropsychology, Methods of 90
Saskia Kohnen and Lyndsey Nickels

Cognitive Neuroscience 95
James L McClelland and Matthew AL Ralph

Cognitive Psychology: History 103
Edward E Smith

Cognitive Styles and Learning Styles 110
Stephen G Rayner

Cohabitation: United States 118
Pamela J Smock and Rhiannon Kroeger

Cold War: Impact on Social and Behavioral Sciences 122
Hamilton Cravens

Cold War, The 128
Gustav Schmidt

Coleman, James Samuel (1926–95) 135
Aage B Sorensen

Collaborative Consumption 139
Rebeca Perren and Liz Grauerholz

Collective Action 145
Bert Klandermans

Collective Behavior, Social Psychology of 151
Stephen Reicher and John Drury

Collective Behavior, Sociology of 157
Neil J Smelser

Collective Beliefs, Sociological Explanation of 163
Michel Dubois

Collective Efficacy and Crime 169
John R Hipp and James C Wo

Collective Identity 174
David A Snow and Catherine Corrigall-Brown

Collective Memory, Anthropology of 181
Uli Linke

Collective Memory, Psychology of 188
Alin Coman

Collective Memory, Social Psychology of 194
Laurent Licata and Aurélie Mercy

Collective Memory, Sociology of 200
Masahiro Ogino

Collectivism and Individualism: Cultural and Psychological Concerns 206
Harry C Triandis

Collins, Randall (1941–) 211
Juan M Iranzo

Colonialism, Anthropology of 218
David Scott

Colonization and Colonialism, History of 223
Wolfgang Reinhard

Color Classification and Symbolism 228
Anita Jacobson-Widding

Color Terms, Linguistics of 231
Paul Kay

Commemoration 235
Barry Schwartz

Commodity Chains 243
Elaine R Hartwick

Common Law 247
Simon Roberts

Common Rule: US Regulatory Protections for Human Research Participants 252
Carl H Coleman

Common Sense, Anthropology of 258
Michael Herzfeld

Communalism 263
Peter van der Veer

Communes, Anthropology of 266
Terence MS Evens

Communication and Transportation, History of 273
Franz Baltzarek

Communication Research and Media Studies, History of 278
Michael Meyen

Communication Skills in Social Work 284
Barbara Staniforth

Communication, Twostep Flow of 291
Gabriel Weimann

Communicative Competence: Linguistic Aspects 294
Constanze Vorwerg

Communism, History of 302
Jerzy Holzer

Communitarianism 308
Walter Reese-Schäfer

Communitarianism, Sociology of 311
Henry Tam

Communitarians 317
Matteo Bortolini

Communities and Crime 322
Lesley W Reid and Christopher Pell

Communities of Practice 328
Jacquelin McDonald

Community and Media 332
Gustavo S Mesch

Community and Society: History of the Concepts 337
Bo Stråth

Community Care in England 344
Caroline Glendinning

Community Genetics 350
Leo Pieter ten Kate

Community Resilience to Disasters 354
Ben Wisner and Ilan Kelman

Community Sociology 361
Brian Jacobs

Community Studies: Anthropological 369
Dan Rabinowitz

Community Supervision and Diversion in the United States 372
Lincoln B Sloas, Jennifer Lerch, and Faye S Taxman

Community Work 378
Kieron Hatton

Community, Social Context of: The US Case 385
Rachel Garshick Kleit

Co-morbidity of Mental and Physical Conditions 390
Kate M Scott

Comparative Evolutionary Psychology: Current Status and a Proposal for a More Integrated Future 394
Jennifer Vonk and Todd K Shackelford

Comparative Health Systems 400
Linda M Whiteford and Cecilia Vindrola-Padros

Comparative History 405
Heinz-Gerhard Haupt

Comparative Method in Anthropology 411
Andre Gingrich

Comparative Method in Evolutionary Studies 415
Chris Venditti and Mark Pagel

Comparative Method in Linguistics 423
Lyle Campbell

Comparative Research in Education: IEA Studies 427
John P Keeves and Petra Lietz

Comparative Studies: Method and Design 439
Dirk Berg-Schlosser

Compliance 445
Jing Cheng and Dylan S Small

Compliance and Legal Authority 456
Jonathan Jackson, Tom R Tyler, Mike Hough, Ben Bradford, and Avital Mentovich

Compression of Morbidity 463
Siegfried Geyer

Computation: Expectation-Maximization Algorithm 469
Geoffrey J McLachlan

Computational Approaches to Model Evaluation 475
Daniel R Cavagnaro

Computational Neuroscience 480
Terrence J Sejnowski

Computational Organizational Theory 485
Geoffrey P Morgan and Kathleen M Carley

Computer-Assisted Instruction 493
Joke Voogt and Petra Fisser

Computer Games and Academic Achievement 498
Christopher J Ferguson

Computer Mediated Communication 504
Yuhua (Jake) Liang and Joseph B Walther

Computerized Test Construction 510
Bernard P Veldkamp

Computers: Impact on the Social Sciences 515
Gerd Gigerenzer

Comte, Auguste (1798–1857) 519
Mary Pickering

Concentrated Disadvantage 525
Paul A Jargowsky and Natasha O Tursi

Concept Formation in Childhood: Cross-Cultural Perspectives 531
Dušan Krnel and Darija Skubic

Concept Learning and Representation: Models 538
David R Shanks

Concepts and Methods of Cognitive Therapies 542
Keith S Dobson and Liza Mastikhina

Conceptual History: *Begriffsgeschichte* 548
Mario Wimmer

Conditional Cash Transfers: Influence on Marriage and Fertility 555
Sarah Baird and Berk Özler

Confabulation 560
Ana Bajo, Aikaterini (Katerina) Fotopoulou, Simon Fleminger, and Michael D Kopelman

Confederations, History of 570
Lakatos Artur

Confidentiality and Statistical Disclosure Limitations 577
George T Duncan

Conflict and Conflict Resolution, Social Psychology of 582
Ronald J Fisher

Conflict and Consensus 589
Katy Hayward

Conflict and Socioemotional Development 594
Brett Laursen and Amy C Hartl

Conflict and War, Archaeology of: Behavior and Social Organization 600
David H Dye

Conflict and War, Archaeology of: Weapons and Artifacts 607
Helle Vandkilde

Conflict and War: Anthropological Aspects 614
Douglas P Fry

Conflict: Anthropological Aspects 620
Georg Elwert and Günther Schlee

Conflict Mediation 627
Isak Svensson

Conflict: Education and Youth 632
Sarah Dryden-Peterson, Michelle J Bellino, and Vidur Chopra

Conflict: Organizational 639
Stewart Clegg, Elisabeth Naima Mikkelsen, and Graham Sewell

Conformity: Sociological Aspects 644
Pierre Demeulenaere

Congress: United States 649
John B Gilmour

Connectionist Models of Concept Learning 654
Koen Lamberts

Conscientiousness 658
Anissa Mike, Kelci Harris, Brent W Roberts, and Joshua J Jackson

Conscious Control during Childhood, Development of 666
Vladimir A Fomichov

Consciousness, Neural Basis of 673
Tamar Weber and Hakwan Lau

Consequentialism (Including Utilitarianism) 679
Philip Pettit

Conservatism: Historical Aspects 684
JZ Muller

Conservatism: Theory and Contemporary Political Ideology 689
Paul E Sigmund

Consociationalism 692
Rudy B Andeweg

Conspicuous consumption and distinction, history of 695
Connor J Fitzmaurice

Constitutional Courts 700
Brun-Otto Bryde

Constitutionalism, Comparative 704
Heinz Klug

Constitutions 710
Kevin L Cope and Mila Versteeg

Constraint-Satisfaction Models 716
László Egri and Thomas R Shultz

Constructivism and Constructionism: Methodology 724
Robert A Neimeyer and Carlos Torres

Constructivist Theory and Psychotherapy 729
Robert A Neimeyer and David A Winter

Consumer Policy 733
Christoph Strünck

Consumer Psychology 738
Jacob Jacoby and Maureen Morrin

Consumption, Geography of 744
Charles Mather

Consumption, History of 751
Emanuela Scarpellini

Consumption, Sociology of 757
Miki Nakai

Contemporary History 762
Pat Thane

Content Analysis 769
Carl W Roberts

Content Validity 774
Linda Crocker

Context Dependent Memory 778
William E Hockley and Tyler D Bancroft

Contextual Studies: Methodology 786
Laurence Steinberg

Continental Philosophy of Language 790
Cristina Lafont

Continuing Professional Development 796
Helen S Timperley

Contractarianism 803
Alan Hamlin

Contracts: Legal Perspectives 809
Leon E Trakman

Control Behavior: Psychological Perspectives 813
Jutta Heckhausen

Control Beliefs: Health Perspectives 819
Kenneth A Wallston

Control, Social 822
Maurice Cusson

Control, Syntactic 828
Klaus-Uwe Panther

Control Theories and Crime 834
Chester L Britt

Control Variables in Research 840
Paras D Mehta

Conventions and Norms: Philosophical Aspects 844
Richard H McAdams

Conversation Analysis: Sociological 850
Jack Sidnell

Cooperation and Competition 856
David W Johnson and Roger T Johnson

Cooperation: Sociological Aspects 862
Andreas Diekmann and Siegwart Lindenberg

Cooperative Game Theory 867
Toru Hokari and William Thomson

Cooperative Learning in Schools 881
Robert E Slavin

Coping across the Lifespan 887
Ellen A Skinner and Melanie Zimmer-Gembeck

Coping, Personality and 895
Gerald Matthews, Moshe Zeidner, and Richard D Roberts

Coping with Job Stress 902
Wilmar B Schaufeli

Core Self-Evaluations 905
Russell E Johnson and Chu-Hsiang (Daisy) Chang

Core–Periphery Patterns in Networks 911
Martin G Everett

Coronary Heart Disease: Psychosocial Aspects 917
Andrew Steptoe

Corporate Culture 923
Edgar H Schein

Corporate Finance: Financial Control 927
Marta Renzetti

Corporate Governance 932
Brett H McDonnell

Corporate Law: Publicly Traded Corporations 939
Arthur R Pinto

Corporate Taxation 946
James R Hines

Corpus Linguistics 949
Chu-Ren Huang and Yao Yao

This page intentionally left blank

CONTRIBUTORS TO VOLUME 4

Marc A. Abramiuk
USA

Rudy B. Andeweg
Leiden University, Leiden, The Netherlands

Lakatos Artur
George Baritiu Institute for History, Romanian Academy of Sciences, Cluj-Napoca, Romania

Sarah Baird
Milken Institute School of Public Health, George Washington University, Washington, DC, USA

Ana Bajo
Institute of Psychiatry, London, UK

Franz Baltzarek
University of Vienna, Vienna, Austria

Tyler D. Bancroft
Wilfrid Laurier University, Waterloo, ON, Canada

Harold C. Barrett
University of California, Los Angeles, CA, USA

Michelle J. Bellino
University of Michigan, Ann Arbor, MI, USA

Andrea Bender
University of Freiburg, Freiburg, Germany

Dirk Berg-Schlosser
Philipps University, Institut fur Politikwissenschaft Universität Marburg, Marburg, Hessen, Germany

Julie L. Booth
Temple University, Philadelphia, PA, USA

Matteo Bortolini
Università di Padova, Padova, Italy

Ben Bradford
University of Oxford, Oxford, UK

Chester L. Britt
School of Criminology and Criminal Justice, Northeastern University, Boston, MA, USA

Brun-Otto Bryde
Justus Liebig University Giessen, Giessen, Germany

Lyle Campbell
University of Canterbury, Christchurch, New Zealand

Kathleen M. Carley
Carnegie Mellon University, Pittsburgh, PA, USA

Kevin M. Carlsmith[†]
Princeton University, Princeton, NJ, USA

Daniel R. Cavagnaro
California State University Fullerton, Fullerton, CA, USA

Chu-Hsiang (Daisy) Chang
Michigan State University, East Lansing, MI, USA

Jing Cheng
University of California San Francisco, San Francisco, CA, USA

Vidur Chopra
Harvard Graduate School of Education, Cambridge, MA, USA

Stewart Clegg
University of Technology Business School, Sydney, NSW, Australia

Jonathan D. Cohen
Princeton University, Princeton, NJ, USA

Carl H. Coleman
Seton Hall University School of Law, Newark, NJ, USA

Alin Coman
Princeton University, Princeton, NJ, USA

Joel Cooper
Princeton University, Princeton, NJ, USA

Kevin L. Cope
Department of Political Science, University of Michigan, Ann Arbor, MI, USA

[†] Deceased.

Catherine Corrigall-Brown
University of Western Ontario, London, ON, Canada

Hamilton Cravens
University of Minnesota, Minnesota, MN, USA

Linda Crocker
University of Florida, Gainesville, FL, USA

Maurice Cusson
Université de Montréal, Montréal, QC, Canada

Pierre Demeulenaere
Université Paris-Sorbonne, Paris, France

Andreas Diekmann
Swiss Federal Institute of Technology, ETH-Zurich, Chair of Sociology, Zurich, Switzerland

Keith S. Dobson
University of Calgary, Calgary, AB, Canada

Jean-Paul Doignon
Université Libre de Bruxelles, Bruxelles, Belgium

John Drury
School of Psychology, University of Sussex, Brighton, UK

Sarah Dryden-Peterson
Harvard Graduate School of Education, Cambridge, MA, USA

Michel Dubois
French National Research Center, CNRS, Paris, France

George T. Duncan
Carnegie Mellon University, Pittsburgh, PA, USA

David H. Dye
University of Memphis, Memphis, TN, USA

Laura A. Edwards
Harvard Graduate School of Education, Boston Children's Hospital, Boston, MA, USA

László Egri
McGill University, Montreal, QC, Canada

Georg Elwert[†]
Free University of Berlin, Berlin, Germany

Terence M.S. Evens
University of North Carolina, Chapel Hill, NC, USA

Martin G. Everett
University of Manchester, Manchester, UK

Christopher J. Ferguson
Stetson University, DeLand, FL, USA

Ronald J. Fisher
School of International Service, American University, Washington, DC, USA

Petra Fisser
Netherlands Institute for Curriculum Development, Enschede, The Netherlands

Connor J. Fitzmaurice
Boston University, Boston, MA, USA

Simon Fleminger
Glenside Hospital, Salisbury, UK

Vladimir A. Fomichov
National Research University Higher School of Economics, Moscow, Russia

Aikaterini (Katerina) Fotopoulou
University College London, London, UK

Douglas P. Fry
University of Alabama at Birmingham, Birmingham, AL, USA

Siegfried Geyer
Hannover Medical School, Hannover, Germany

Gerd Gigerenzer
Max Planck Institute for Human Development, Berlin, Germany

John B. Gilmour
College of William and Mary, Williamsburg, VA, USA

Andre Gingrich
Austrian Academy of Sciences, Vienna, Austria

Caroline Glendinning
University of York, Heslington, York, UK

Liz Grauerholz
University of Central Florida, Orlando, FL, USA

Alan Hamlin
University of Manchester, Manchester, UK

Kelci Harris
Washington University in St. Louis, St. Louis, MO, USA

Amy C. Hartl
Florida Atlantic University, Boca Raton, FL, USA

Elaine R. Hartwick
Framingham State University, Framingham, MA, USA

Kieron Hatton
School of Health Sciences and Social Work, University of Portsmouth, Portsmouth, UK

[†] Deceased.

Heinz-Gerhard Haupt
European University Institute, Florence, Italy

Katy Hayward
School of Sociology, Social Policy and Social Work, Queen's University Belfast, Belfast, UK

Jutta Heckhausen
University of California, Irvine, CA, USA

Michael Herzfeld
Harvard University, Cambridge, MA, USA

James R. Hines
University of Michigan, Ann Arbor, MI, USA; and NBER, New York, NY, USA

John R. Hipp
University of California, Irvine, CA, USA

William E. Hockley
Wilfrid Laurier University, Waterloo, ON, Canada

Toru Hokari
Department of Economics, Keio University, Keio, Japan

Jerzy Holzer
Polish Academy of Sciences, Polna, Warszawa, Poland

Olivier Houdé
Paris-Descartes University, Sorbonne-Paris-City Alliance for Higher Education and Research, CNRS, Paris, France

Mike Hough
University of London, London, UK

Chu-Ren Huang
Hong Kong Polytechnic University, Hong Kong

Hafiz M. Iqbal
University of the Punjab (formerly), Lahore, Pakistan

Juan M. Iranzo
Public University of Navarre, Pamplona, Spain

Jonathan Jackson
London School of Economics and Political Science, London, UK

Joshua J. Jackson
Washington University in St. Louis, St. Louis, MO, USA

Brian Jacobs
Staffordshire University, Stoke-on-Trent, UK

Anita Jacobson-Widding
University of Uppsala, Uppsala, Sweden

Jacob Jacoby
New York University, New York, NY, USA

Paul A. Jargowsky
Center for Urban Research and Education, Rutgers University — Camden, Camden, NJ, USA

David W. Johnson
University of Minnesota, MN, USA

Roger T. Johnson
University of Minnesota, MN, USA

Russell E. Johnson
Michigan State University, East Lansing, MI, USA

Paul Kay
University of California, Berkeley, CA, USA

John P. Keeves
The School of Education, University of Adelaide, Adelaide, SA, Australia

Ilan Kelman
Institute for Risk and Disaster Reduction (IRDR) and Institute for Global Health (IGH), University College London (UCL), London, UK

David Kingdon
College Keep, University of Southampton, Southampton, UK

Rob Kitchin
National University of Ireland Maynooth, Maynooth, Ireland

Bert Klandermans
VU University, Amsterdam, The Netherlands

Rachel Garshick Kleit
The Ohio State University, Columbus, OH, USA

Heinz Klug
University of Wisconsin Law School, Madison, WI, USA; and University of Witwatersrand, Johannesburg, South Africa

Saskia Kohnen
Macquarie University, Sydney, NSW, Australia

Michael D. Kopelman
King's College London, London, UK; and St Thomas' Hospital, London, UK

Dušan Krnel
University in Ljubljana, Ljubljana, Slovenia

Rhiannon Kroeger
Population Research Center, University of Texas at Austin, Austin, TX, USA

Cristina Lafont
Northwestern University, Evanston, IL, USA

Koen Lamberts
University of Warwick, Coventry, UK

Hakwan Lau
Columbia University, Los Angeles, CA, USA

Brett Laursen
Florida Atlantic University, Boca Raton, FL, USA

Jennifer Lerch
Center for Advancing Correctional Excellence (ACE!), George Mason University, Fairfax, VA, USA

Yuhua (Jake) Liang
Chapman University, Orange, CA, USA

Laurent Licata
Research Center for Social and Cultural Psychology, Unité de Psychologie Sociale, Université Libre de Bruxelles, Bruxelles, Belgium

Petra Lietz
The Australian Council for Educational Research (ACER), Adelaide, SA, Australia

Siegwart Lindenberg
University of Groningen, Groningen, The Netherlands; and Tilburg University, Tilburg, The Netherlands

Uli Linke
Rochester Institute of Technology, Rochester, NY, USA

Bruce MacLennan
University of Tennessee, Knoxville, TN, USA

Helen Mander
Southern Health NHS Trust, Southampton, UK

Liza Mastikhina
University of Calgary, Calgary, AB, Canada

Charles Mather
Memorial University, St John's, NL, Canada

Gerald Matthews
Institute for Simulation & Training, University of Central Florida, Orlando, FL, USA

Richard H. McAdams
University of Chicago Law School, Chicago, IL, USA

James L. McClelland
Center for Mind, Brain, and Computation, Stanford University, Stanford, CA, USA

Jacquelin McDonald
Learning and Teaching Services (LTS), The University of Southern Queensland, Toowoomba, QLD, Australia

Brett H. McDonnell
University of Minnesota Law School, Minneapolis, MN, USA

Geoffrey J. McLachlan
University of Queensland, Brisbane, QLD, Australia

Paras D. Mehta
University of Illinois, Chicago, IL, USA

Avital Mentovich
University of California Los Angeles, Los Angeles, CA, USA

Aurélie Mercy
Research Center for Social and Cultural Psychology, Unité de Psychologie Sociale, Université Libre de Bruxelles, Bruxelles, Belgium

Gustavo S. Mesch
University of Haifa, Haifa, Israel

Michael Meyen
Institute for Communication and Media Research, University of Munich, München, Germany

Anissa Mike
Washington University in St. Louis, St. Louis, MO, USA

Elisabeth Naima Mikkelsen
Copenhagen Business School, Frederiksberg, Denmark

Geoffrey P. Morgan
Carnegie Mellon University, Pittsburgh, PA, USA

Maureen Morrin
Temple University, Philadelphia, PA, USA

J.Z. Muller
Catholic University of America, Washington, DC, USA

Miki Nakai
Ritsumeikan University, Kyoto, Japan

Robert A. Neimeyer
University of Memphis, Memphis, TN, USA

Lyndsey Nickels
Macquarie University, Sydney, NSW, Australia

Masahiro Ogino
Kwansei Gakuin University, Nishinomiya, Japan

Berk Özler
Development Research Group, The World Bank, Washington, DC, USA

Mark Pagel
University of Reading, Reading, UK

Klaus-Uwe Panther
Nanjing Normal University, Nanjing, China; and University of Hamburg, Hamburg, Germany

Christopher Pell
University of Tennessee at Chattanooga, Chattanooga, TN, USA

Rebeca Perren
College of Business Administration, University of Central Florida, Orlando, FL, USA

Christopher D. Peterson
University of California, Los Angeles, CA, USA

Philip Pettit
Princeton University, Princeton, USA; and Australian National University, Canberra, ACT, Australia

Mary Pickering
San Jose State University, San Jose, CA, USA

Arthur R. Pinto
Brooklyn Law School, New York, NY, USA

Dan Rabinowitz
Tel Aviv University, Tel Aviv, Israel; and Central European University, Budapest, Hungary

Matthew A.L. Ralph
School of Psychological Sciences, The University of Manchester, Manchester, UK

Stephen G. Rayner
Newman University, Birmingham, UK

Walter Reese-Schäfer
University of Goettingen, Goettingen, Germany

Stephen Reicher
School of Psychology and Neuroscience, University of St. Andrews, St. Andrews, UK

Lesley W. Reid
University of Alabama, Tuscaloosa, AL, USA

Wolfgang Reinhard
University of Freiburg im Breisgau, Freiburg im Breisgau, Germany

Marta Renzetti
SDA Bocconi-Corporate Finance, Milan, Italy

Brent W. Roberts
University of Illinois at Urbana-Champaign, Champaign, IL, USA

Carl W. Roberts
Iowa State University, Ames, IA, USA

Richard D. Roberts
Center for Innovative Assessments, Professional Examination Service, New York, NY, USA

Simon Roberts
London School of Economics and Political Science, London, UK

Emanuela Scarpellini
Università degli Studi di Milano (University of Milan), Milano, Italy

K. Warner Schaie
University of Washington, Seattle, WA, USA

Wilmar B. Schaufeli
Utrecht University, Utrecht, The Netherlands

Edgar H. Schein
MIT Sloan School of Management, Cambridge, MA, USA

Günther Schlee
Free University of Berlin, Berlin, Germany; and Max Planck Institute for Social Anthropology, Halle (Saale), Germany

Gustav Schmidt
Ruhr University Bochum, Herdecke, Germany

Barry Schwartz
University of Georgia, Athens, GA, USA

David Scott
Columbia University, New York, NY, USA

Kate M. Scott
University of Otago, Dunedin, New Zealand

Terrence J. Sejnowski
Howard Hughes Medical Institute, Chevy Chase, MD, USA; Salk Institute for Biological Studies, La Jolla, CA, USA; and University of California, San Diego, La Jolla, CA, USA

Graham Sewell
University of Melbourne, Melbourne, VIC, Australia

Todd K. Shackelford
Oakland University, Rochester, MI, USA

David R. Shanks
University College London, London, UK

Thomas R. Shultz
McGill University, Montreal, QC, Canada

Jack Sidnell
University of Toronto, Toronto, ON, Canada

Paul E. Sigmund
Princeton University, Princeton, NJ, USA

Ellen A. Skinner
Portland State University, Portland, OR, USA

Darija Skubic
University in Ljubljana, Ljubljana, Slovenia

Robert E. Slavin
Johns Hopkins University, Baltimore, MD, USA; and University of York, York, England

Lincoln B. Sloas
Center for Advancing Correctional Excellence (ACE!), George Mason University, Fairfax, VA, USA

Dylan S. Small
University of Pennsylvania, Philadelphia, PA, USA

Neil J. Smelser
University of California, Berkeley, CA, USA

Edward E. Smith
University of Michigan, Ann Arbor, MI, USA

Pamela J. Smock
Population Studies Center, Institute for Social Research, Ann Arbor, MI, USA

David A. Snow
University of California, Irvine, CA, USA

Aage B. Sorensen
Harvard University, Cambridge, MA, USA

Barbara Staniforth
School of Counselling, Human Services and Social Work, Faculty of Education, The University of Auckland, Auckland, New Zealand

Laurence Steinberg
Temple University, Philadelphia, PA, USA

Andrew Steptoe
University College London, London, UK

Bo Stråth
University of Helsinki, Helsinki, Finland

Christoph Strünck
School of Arts and Humanities, University of Siegen, Siegen, Germany

Isak Svensson
Uppsala University, Uppsala, Sweden

Henry Tam
University of Cambridge, Cambridge, UK

Faye S. Taxman
Center for Advancing Correctional Excellence (ACE!), George Mason University, Fairfax, VA, USA

Leo Pieter ten Kate
VU University Medical Center, Amsterdam, The Netherlands

Pat Thane
King's College London, London, UK

Clarissa A. Thompson
University of Oklahoma, Norman, OK, USA

William Thomson
University of Rochester, Rochester, NY, USA

Helen S. Timperley
Faculty of Education, School of Learning, Development and Professional Practice, The University of Auckland, Epsom, Auckland, New Zealand

Carlos Torres
University of Memphis, Memphis, TN, USA

Leon E. Trakman
University of New South Wales, Sydney, NSW, Australia

Harry C. Triandis
University of Illinois, Champaign, IL, USA

Natasha O. Tursi
Center for Urban Research and Education, Rutgers University — Camden, Camden, NJ, USA

Tom R. Tyler
Yale University, New Haven, CT, USA

Peter van der Veer
Max Planck Institute for the Study of Religious and Ethnic Diversity, Göttingen, Germany

Helle Vandkilde
Institute of Culture and Society, Aarhus University, Hoejbjerg, Denmark

Bernard P. Veldkamp
University of Twente, Twente, The Netherlands

Chris Venditti
University of Reading, Reading, UK

Mila Versteeg
School of Law, University of Virginia, Charlottesville, VA, USA

Cecilia Vindrola-Padros
London South Bank University, London, UK

Jennifer Vonk
Oakland University, Rochester, MI, USA

Joke Voogt
University of Amsterdam/Windesheim University, Amsterdam, The Netherlands

Constanze Vorwerg
Center for the Study of Language and Society, University of Bern, Bern, Switzerland; and Institut für Sprachwissenschaft, University of Bern, Bern, Switzerland

Kenneth A. Wallston
Vanderbilt University, Nashville, TN, USA

Joseph B. Walther
Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore

Lamei Wang
The Hong Kong Institute of Education, New Territories, Hong Kong

Zhenlin Wang
The Hong Kong Institute of Education, New Territories, Hong Kong

Jürg Wassmann
Institute of Anthropology, University of Heidelberg, Heidelberg, Germany

Tamar Weber
University of California Los Angeles, Los Angeles, CA, USA

Gabriel Weimann
University of Haifa, Haifa, Israel

Linda M. Whiteford
University of South Florida, Tampa, FL, USA

Mario Wimmer
University of California Berkeley, Berkeley, CA, USA

David A. Winter
University of Hertfordshire, Hatfield, UK

Ben Wisner
Aon-Benfield UCL Hazard Research Centre, University College London, London, UK

James C. Wo
University of California, Irvine, CA, USA

Yao Yao
Hong Kong Polytechnic University, Hong Kong

Moshe Zeidner
Laboratory for Research in Personality, Emotions, and Individual Differences, University of Haifa, Haifa, Israel

Melanie Zimmer-Gembeck
School of Applied Psychology and Griffith Health Institute, Behavioural Basis of Health, Griffith University, Gold Coast, QLD, Australia

This page intentionally left blank

Cognition, Evolution of

Christopher D Peterson and Harold C Barrett, University of California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by H.C. Barrett, volume 7, pp. 4996–5000, © 2001, Elsevier Ltd.

Abstract

Cognition is the product of brain mechanisms that have been shaped by the evolutionary process. Understanding the form and function these mechanisms take, how they develop, and how they are distributed across species requires situating these mechanisms in an evolutionary context, and studying them using the diversity of methods available to the evolutionary sciences. Here we review how the mechanisms underlying cognition can be studied from an evolutionary point of view.

Evolutionary theory is widely agreed to be the best explanation for biological complexity and organization yet conceived, but the power of this theoretical approach has only in the last few decades been fully realized in explanations of human and nonhuman behavior (*see* Sociobiology: Overview; Sociobiology: History of). The behavior of humans and other animals is generated by functionally specialized nervous systems that develop and operate within the specific ecology of the organism (Tooby and Cosmides, 1992). Since the work of Niko Tinbergen, biologists have recognized that in order to fully understand an organism's behavior, explanations at several complementary (i.e., nonmutually exclusive) causal levels are required (Tinbergen, 1958).

On the proximate level, behaviors can be explained mechanistically as the product of neural activity, which itself responds to both external states, such as temperature, and internal ones, such as goals, emotion states, and physiological conditions. Neurobiology, cellular biology, and related disciplines have produced explanations for behaviors at this direct, causal level that are critical to understanding how behavior is proximally caused. Also at the proximate level are developmental explanations. An organism's psychological and physical traits are the result of developmental processes, initiated at conception, that shape the growing organism as a function of its interaction with the environment. Developmental biology and psychology have produced explanations of behavior that consider the timing of the appearance of adaptations over a single lifetime and the roles of learning and plasticity in shaping them (*see* Evolutionary Developmental Psychology).

Evolutionary or ultimate explanations are of two kinds. Phylogenetic explanations take into account the evolutionary history of descent with modification that has led a given organism to possess some traits and not others. This history can shape and constrain adaptations underlying behavior in powerful ways. Paleontologists, comparative anatomists, and geneticists have reconstructed the lineages of many species, and these phylogenies allow us to understand the evolution of brains and behavior in comparative and historical perspective (*see* Comparative Evolutionary Psychology: Current Status and a Proposal for a More Integrated Future; Comparative Method in Evolutionary Studies; Brain, Evolution of; Body, Evolution of). Another kind of ultimate explanation is functional or adaptive explanation, in which the traits of organisms are analyzed in terms of the roles they play in promoting survival and reproduction (fitness).

Functionalist or adaptationist approaches to cognition have made significant contributions to our understanding of how particular forms of behaviors and neural mechanisms reflect their adaptive functions. Until recently, researchers studying behavior have tended to focus exclusively on one level of analysis. More recent approaches to the evolution of cognition, such as evolutionary psychology, have sought to incorporate multiple levels of causation by explaining behavior not merely in adaptive terms (i.e., in terms of their immediate effects on fitness in current environments), but as the product of underlying proximate mechanisms, shaped by evolution, that might or might not produce adaptive behavior depending on the conditions under which they develop and operate.

Psychological Adaptations

An adaptationist approach to the mind begins with a simple premise: just like lungs, livers and flagella, brains are evolved organs with functional parts. Each of these parts, which are sometimes called psychological adaptations or modules, has evolved for something. They exist because they have been selected for, meaning they had positive effects on survival and reproduction over evolutionary time. They are 'for' something, in that they have consistently solved some adaptive problem, whether it was to discern light from darkness, to discriminate loud noises from soft ones, or to identify an object as food or as inedible. Any process in the mind that has evolved through natural selection to carry out an information-processing function can be regarded as a psychological adaptation (Barrett and Kurzban, 2006).

Formal descriptions of psychological adaptations include descriptions of their inputs, information-processing procedures, and outputs. The set of inputs that a psychological adaptation processes is called its domain, and adaptations vary in the kinds of information they have evolved to use. Similarly, the information-processing features of mechanisms can vary substantially, from the time series analyses performed by classical conditioning mechanisms, to Bayesian inference performed by systems underlying word learning and categorization, to if-then rules characterizing reasoning

systems. Finally, the outputs of psychological adaptations can exhibit special design; for example, different emotion systems cause a variety of diverse behavioral and physiological responses in organisms, from sexual attraction to nausea (*see* Emotion, Evolution of).

At the outset of the study of psychological adaptations, many researchers held that they should possess a common set of features, such as narrow information domains, encapsulation (inaccessibility to conscious processes), lack of environmental influences on development, and automaticity (Fodor, 1983). While some researchers continue to view psychological adaptations as strictly 'modular' in this sense, others argue that the hallmark of adaptation is likely to be functional diversity rather than functional uniformity (Barrett and Kurzban, 2006). This 'functional diversity' view of psychological adaptations is able to accommodate many types of mechanisms under the rubric of specialized adaptation, including mechanisms of learning and cultural transmission.

Domains

All the problems mentioned above are solved via the processing of information, but not all psychological processes use the same information. The information that a psychological mechanism uses is called its domain. For example, the visual system processes information about light energy transduced by photoreceptors in the retina, whereas the auditory system processes information about sound waves transduced by the cochlea. Within vision, some mechanisms respond to color, whereas others respond to shape, and even more domain-specific mechanisms can exist, such as mechanisms specialized to process the configuration of facial features in humans. In each case, the mechanism was shaped by the evolutionary process because of the effects it had on the organism's survival and reproduction. Visual mechanisms, for example, evolved because of their benefits for detecting predators and mates, for spatial navigation, and so on.

The behavior of complex organisms such as humans is not likely to be carried out by a single, general-purpose mechanism. In brains, information processing is made more efficient by cognitive division of labor: delegating tasks among smaller subsystems, whose operations can be coordinated through hierarchical organization and feedback loops. The nervous system is therefore a mosaic of many evolved mechanisms, often with different evolutionary histories. However, the fact that brains are complex networks of neural tissue means that investigators often cannot distinguish one subsystem from the next using anatomy alone.

Methods for Investigating Psychological Adaptations

Laboratory and Naturalistic Experiments

The signature of natural selection is functional design, characterized by a fit between form and function. In the case of cognition, this means that the brain mechanisms that perform specific information-processing functions will exhibit processing features designed to carry out those functions. To investigate adaptive design, rigorous empirical methods from a variety of fields including cognitive psychology, developmental psychology, anthropology, and neuroscience are in order, because these methods allow some measure of control over the multitude of variables interacting within a thinking mind. Laboratory experiments are tightly constrained and often artificial, which has advantages and disadvantages. It is informative to complement laboratory studies of cognition with studies in the field, and across diverse settings. An advantage of field studies is increased ecological validity, i.e., resemblance of the task to the natural setting in which it evolved.

As an example, standard experimental methods from cognitive psychology have been used in ecologically valid contexts in work investigating sex differences in human spatial cognition. Silverman and Eals (1992) proposed that sex differences in human spatial cognition may reflect an ancestral division of labor between hunting and foraging. Prior work has shown that women are better than men at certain abstract spatial tasks involving memory for object location, but since most of the data were gleaned from laboratory experiments, ecological validity was lacking. New et al. (2007) studied spatial cognition in an outdoor setting in a farmers market in Southern California, and showed that women are more accurate than men when pointing in the direction of newly learned locations for food items, but that both men and women are equally good in locating calorically rich food items in the same market. This finding is consistent with the existence of two distinct adaptive mechanisms: one to rank the value of food resources and one to remember the locations of those resources. By approximating foraging in a simple, natural design with modest controls, New and colleagues were able to offer a stronger argument for functional design. While sex differences may exist in spatial foraging ability, they are attenuated when their caloric value is higher, consistent with a mechanism that prioritizes the location of high-value food resources above low-value ones. This example shows how the use of multiple methods, within and outside the lab, can be used to provide converging evidence for or against evolutionary hypotheses.

Neuropsychological Dissociations

Just as multiple functionally specific adaptations can independently solve problems that confront the organism, so too can they independently malfunction, to varying degrees. Much of our knowledge of neurology and neuroanatomy has come from studies that examine impairments associated with localized brain tissue damage. The fact that small amounts of damage can result in specific deficits or dissociations in cognitive ability is powerful evidence for specialized design in the brain. Prosopagnosia, for instance, is a specific dissociation that impairs the facial recognition system, rendering patients unable to recognize people they know from facial features alone, while leaving relatively intact social knowledge of those individuals and the ability to identify them through other cues such as voice, hair, and gait. However, there is debate regarding the exact nature of the mechanisms impaired in prosopagnosia, with some arguing that they are evolutionarily specialized for face recognition, and others arguing that they are specialized for developing expertise in object recognition, with faces as a special developmental case (Gauthier and Nelson, 2001;

Kanwisher, 2010). This debate illustrates an important feature of domain specificity. Mechanisms are not either 'domain specific' or 'domain general,' but rather, can be described in terms of the nature of their domain: faces and objects are both domains, albeit different ones.

Developmental Studies

One mistaken, though widely held, belief about psychological adaptations is that they cannot be shaped by developmental input: that they must be present at birth and develop identically across individuals (i.e., 'innate' in a narrow sense). Instead, virtually all adaptations exhibit some degree of plasticity, though they vary in the nature and degree of that plasticity (*see* Evolutionary Developmental Psychology). Thus, hypotheses about psychological adaptations should include descriptions of how the mechanism in question is proposed to develop. This would include age trajectories of development, as well as hypotheses about how the mechanism in question might vary across developmental environments. For example, experiments with infants suggest that some mechanisms, such as mechanisms underlying the perception of objects, are thought to develop early in childhood, and similarly across children (Spelke, 1990). Other mechanisms, such as mechanisms shaping notions of fairness and moral judgment, appear to develop later, and can vary significantly depending on the culture in which the child is raised (House et al., 2013). This suggests that controlled studies across ages and populations are necessary to test hypotheses about the adaptive nature of cognition.

Cross-Cultural Studies

Cultural variation in cognition does not necessarily imply that adaptations are not at work. Even physiological and morphological traits that have been clearly shaped by natural selection, such as the immune system and the heart, vary across individuals. In the case of cognitive mechanisms, learning and other forms of external input are likely to play a role, in addition to – especially in the case of humans – culture (*see* Cultural Evolution: Overview). Thus, the existence or lack of variation does not by itself rule in or rule out the possibility that an aspect of cognition has been shaped by natural selection. Instead, theories of cognitive adaptations should include hypotheses about how and whether variation in the trait is expected to be manifest.

For example, there seems little doubt that spatial cognition in humans and other animals has been shaped by natural selection. In some cases, highly specialized spatial navigation systems are known to exist, such as dead reckoning in ants (Gallistel, 1990) and celestial navigation in indigo buntings (Emlen, 1975). Even in these cases, however, environmental input and learning play a key role: for example, the bunting celestial navigation system is calibrated through repeated exposure to the night sky during the course of development. In the case of humans, cross-cultural studies have shown that language is likely to play an important role in shaping some aspects of spatial cognition. For example, speakers of languages that customarily use absolute instead of body-centered frames of reference (equivalent to using, e.g., 'north' versus 'south' instead of 'left' and 'right') show greater awareness of their orientation relative to environmental landmarks and frames of reference (Levinson, 2003). This cross-cultural variation does not suggest that human spatial cognition is not the product of adaptations; it does, however, suggest that these adaptations are shaped by linguistic and/or other experiential input. More broadly, dimensions of cross-cultural variation, or lack thereof, are important sources of evidence for testing specific hypotheses about the features of psychological adaptations.

Cross-Species Comparisons

As Darwin recognized, adaptations evolve through descent with modification, and can therefore exhibit homologies, or similarities across species as a result of descent from common ancestors. On the other hand, specific taxa can have derived traits or traits that are unique to the taxon (and therefore do not have homologs in other taxa). And finally, distantly related taxa may share analogous versions of traits – traits that are similar – not due to common ancestry, but due to convergent evolution to solve similar adaptive problems. Because of this, the comparative method – comparing traits across species – is an important source of evidence for testing hypotheses about psychological adaptations, including hypotheses about their functions and evolutionary origins (*see* Comparative Method in Evolutionary Studies). Among the earliest cross-species comparisons were Darwin's analyses of similarities and differences in emotional expression across species.

An example of recent comparative work on psychological mechanisms is research examining the nature of prosociality, or cooperation, in humans and other great apes (*see* Human Cooperation, Evolution of). Key to this work are methods that can be compared across taxa, such as simple choices of how to allocate resources between oneself and a social partner (Silk et al., 2005). This work provides evidence for both similarities and differences in prosocial decision-making between humans and our closest relatives, suggesting some homologies and other derived features in humans. For example, humans and other primates respond prosocially to some signs of need, but only humans perform altruistic, prosocial acts in the absence of explicit signals of need from another party, suggesting a difference in the mechanisms underlying cooperation (Jaeggi et al., 2010).

Genes, Fossils, and Other Sources of Evidence

Finally, new technologies are beginning to provide data about cognitive evolution from novel sources. Genetic studies, for example, are beginning to provide evidence about genes that have diverged in humans and chimpanzees, and genes that have been under recent selection in our lineage. These include genes potentially related to language, motor control, brain connectivity, and other aspects of brain development. Studies of gene expression in humans and other animals also potentially provide a window into the nature and development of cognitive mechanisms, with a variety of studies showing substantial divergence in how genes are expressed in the brains of humans and other apes (Enard et al., 2002). Finally, paleoanthropological and paleoarchaeological evidence potentially bear on hypotheses about human cognition. For example,

anatomical evidence has been used to examine hypotheses about the origins of language in humans, and data from stone tools have been used to evaluate hypotheses about the cognitive mechanisms of early toolmakers, such as their working memory and spatial cognition (Wynn and Coolidge, 2004). Given the speed of technological change, it is likely that even more previously unknown sources of evidence will be brought to bear on our understanding of cognitive evolution.

A Case Study: Theory of Mind

Few cognitive adaptations have been investigated using all or most of the potential sources of evidence described above. For example, there are many claims about human universals that are supported by evidence from only a small and perhaps nonrepresentative sample of human societies. Ideally, multiple and converging sources of evidence are necessary in order to be confident about the nature of the evolved mechanisms underlying any particular aspect of cognition.

A useful case study for how diverse sources of evidence can be brought to bear in understanding the evolution of a cognitive ability is the case of 'Theory of Mind,' also variously known as mentalizing, mind reading, folk psychology, and psychological reasoning. Broadly, these terms refer to the ability to represent and make inferences about, and decisions based on, the mental states of others, such as their goals, desires, beliefs, intentions, attitudes, and motivations. While some theorists regard Theory of Mind as an 'all or none' capacity – you either have it or you don't – many researchers regard Theory of Mind as a continuum of abilities, distributed in different forms and combinations across species. Many agree that humans have particularly well-developed and elaborate Theory of Mind abilities, which develop early in childhood and are deployed across a range of contexts, from cooperation, to communication, to social learning.

A variety of sources of evidence have been used to examine how Theory of Mind is distributed across species, how it develops, how it operates on the neural level, and what happens when it is impaired. In humans, hundreds of studies have been conducted using the false belief task, a task designed to measure the ability to represent another's belief that is different from one's own (i.e., a false belief). The majority of studies have used so-called elicited false belief tasks in which children are asked to report about another's beliefs or behavior, and these generally show an age of development around four years, with substantial cross-cultural variation (Wellman et al., 2001). More recently, 'spontaneous' response tasks have shown that tracking of others' beliefs develops much earlier, and shows a similar developmental trajectory in early childhood across cultures (Barrett et al., 2013). Across species, no other species has been shown to track false beliefs, but our closest relatives, chimpanzees, represent the knowledge or ignorance of others and use it to predict their behavior (Call and Tomasello, 2008). Moreover, other mind-reading skills, such as the ability to track gaze, to understand when one has been seen, and to understand communicative intent, have been found in species ranging from scrub jays to domesticated dogs (Dally et al., 2006; Hare et al., 2002). Importantly, this work shows that the mind-reading skills of a species tend to be correlated with the adaptive problems facing the species, such as avoiding food theft (Dally et al., 2006). In humans, brain mapping studies have begun to reveal the neural substrates underlying Theory of Mind (Saxe and Powell, 2006). Impairment of mind-reading abilities, for example in autism, have been used to test hypotheses about the specificity of the underlying mechanisms (Baron-Cohen, 1995). In all, these studies converge on the conclusion that Theory of Mind is a specialized ability, produced by psychological adaptations that exhibit characteristic developmental timing and that are distributed across species in ways that make sense, given the social ecology of the species. While Theory of Mind is perhaps atypical in the diversity and breadth of techniques that have been used to study it, it provides a good example of how converging work across many fields can be used to understand the evolution of cognitive mechanisms.

Conclusion

There can be little doubt that the cognitive abilities of all species are products of the evolutionary process. They are the products of neural systems that have evolved through descent with modification, shaped by the success or failure of their information-processing properties over evolutionary time. The question, then, is not whether cognition is the product of evolution. Instead, we should ask how the mechanisms giving rise to a species' cognitive abilities have been shaped by the evolutionary process, how those mechanisms develop, and how they interact with the species' environment to produce the behaviors we observe.

In many ways, the study of cognitive evolution is still in its infancy. Contemporary psychology is only a little more than a century old, and evolutionary theory is only slightly older than that. The blending of the two, although it began in the work of Darwin, has only recently begun to take shape as a mainstream form of scientific inquiry. As a result, there is as yet no standard toolkit of methods for studying the evolution of cognitive mechanisms, but rather a diverse assortment of approaches drawn from the many branches of the biological and behavioral sciences. However, with the rise in popularity of evolutionary psychology and related fields, the possibility of a mature science of cognitive evolution is now in sight. Our map of the brain's evolved mechanisms still contains many blank spots, but the development of new and innovative ways of testing hypotheses about psychological adaptations suggests that much more will be known about the evolutionary basis of human cognition and that of other species, in the coming decades.

See also: Body, Evolution of; Brain, Evolution of; Comparative Evolutionary Psychology: Current Status and a Proposal for a More Integrated Future; Comparative Method in Evolutionary Studies; Cultural Evolution: Overview; Emotion, Evolution of; Evolutionary Developmental Psychology; Human Cooperation, Evolution of; Sociobiology, History of; Sociobiology: Overview.

Bibliography

Baron-Cohen, S., 1995. Mindblindness: An Essay on Autism and Theory of Mind. MIT Press, Cambridge.

Barrett, H.C., Kurzban, R., 2006. Modularity in cognition: framing the debate. Psychological Review 113, 628–647.

Barrett, H.C., Broesch, T., Scott, R.M., He, Z., Baillargeon, R., Wu, D., Bolz, M., Henrich, J., Setoh, P., Wang, J., Laurence, S., 2013. Early false-belief understanding in traditional non-western societies. Proceedings of the Royal Society of London, Series B 280 (1755), 20122654.

Call, J., Tomasello, M., 2008. Does the chimpanzee have a theory of mind? 30 years later. Trends in Cognitive Science 12, 187–192.

Dally, J.M., Emery, N.J., Clayton, N.S., 2006. Food-caching western scrub-jays keep track of who was watching when. Science 312, 1662–1665.

Emlen, S.T., 1975. The stellar-orientation system of a migratory bird. Scientific American 233, 102–111.

Enard, W., Khaitovich, P., Klose, J., Zöllner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R., Doxiadis, G.M., Bontrop, R.E., Pääbo, S., 2002. Intra-and interspecific variation in primate gene expression patterns. Science 296, 340–343.

Fodor, J., 1983. The Modularity of Mind. MIT Press, Cambridge.

Gallistel, C.R., 1990. The Organization of Learning. Bradford Books/MIT Press, Cambridge.

Gauthier, I., Nelson, C.A., 2001. The development of face expertise. Current Opinion in Neurobiology 11, 219–224.

Hare, B., Brown, M., Williamson, C., Tomasello, M., 2002. The domestication of social cognition in dogs. Science 298, 1634–1636.

House, B.R., Silk, J.B., Henrich, J., Barrett, H.C., Scelza, B., Boyette, A., Hewlett, B., McElreath, R., Laurence, S., 2013. The ontogeny of prosocial behavior across diverse societies. Proceedings of the National Academy of Sciences 110 (36), 14586–14591.

Jaeggi, A.V., Burkart, J.M., Van Schaik, C.P., 2010. On the psychology of cooperation in humans and other primates: combining the natural history and experimental evidence of prosociality. Philosophical Transactions of the Royal Society B: Biological Sciences 365 (1553), 2723–2735.

Kanwisher, N., 2010. Functional specificity in the human brain: a window into the functional architecture of the mind. Proceedings of the National Academy of Sciences 107, 11163–11170.

Levinson, S.C., 2003. Space in Language and Cognition. Cambridge University Press, Cambridge.

New, J., Krasnow, M.M., Truxaw, D., Gaulin, S.J., 2007. Spatial adaptations for plant foraging: women excel and calories count. Proceedings of the Royal Society B: Biological Sciences 274 (1626), 2679–2684.

Saxe, R., Powell, L., 2006. It's the thought that counts: specific brain regions for one component of theory of mind. Psychological Science 17 (8), 692–699.

Silk, J.B., Brosnan, S.F., Vonk, J., Henrich, J., Povinelli, D.J., Richardson, A.S., Lambeth, S., Mascaro, J., Schapiro, S.J., 2005. Chimpanzees are indifferent to the welfare of unrelated group members. Nature 437 (7063), 1357–1359.

Silverman, I., Eals, M., 1992. Sex differences in spatial abilities: evolutionary theory and data. In: Barkow, J.H., Cosmides, L., Tooby, J. (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford Press, New York, pp. 531–549.

Spelke, E., 1990. Principles of object perception. Cognitive Science 14, 29–56.

Tinbergen, N., 1958. Curious Naturalists. Basic Books, New York.

Tooby, J., Cosmides, L., 1992. The psychological foundations of culture. In: Barkow, J., Cosmides, L., Tooby, J. (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, New York.

Wellman, H.M., Cross, D., Watson, J., 2001. A meta-analysis of theory of mind development: the truth about false belief. Child Development 72, 655–684.

Wynn, T., Coolidge, F.L., 2004. The expert Neandertal mind. Journal of Human Evolution 46 (4), 467–487.

Cognition: Knowledge Spaces

Jean-Paul Doignon, Université Libre de Bruxelles, Bruxelles, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Nomenclature

- $\mathcal{P}$ A relation on the ground set Q.
- $\varnothing$ The empty set.
- $\boldsymbol{\sigma}$ A mapping that associates to any item q the collection of clauses for q, that is, the various possible ways to acquire the mastery of q.
- $\boldsymbol{d}$ The symmetric-difference distance. For two sets K and L, the expression $d(K,L)$ gives the number of elements in which K and L differ.
- $K^{\mathcal{I}}$ The inner fringe of the knowledge state K, consisting of all items q in K such that $K\backslash\{q\}$ is again a knowledge state.
- $K^{\mathcal{O}}$ The outer fringe of the knowledge state K, consisting of all items q outside K such that $K\cup\{q\}$ is again a knowledge state.

Abstract

Knowledge assessment should be a central component in any computerized course. Procedures for knowledge assessment have been designed in the theoretical setting of knowledge spaces and learning spaces. They allow for a better outcome than just a crude number, giving deeper information on what is mastered together with advices for further study. Knowledge space theory thus offers a multidimensional extension of classical test theory. The present article introduces knowledge spaces under their combinatorial aspects, while the article Markov Processes in Learning Spaces describes the probabilistic aspects and the related assessment procedures.

This article introduces the combinatorial aspects of knowledge spaces and their special cases, learning spaces, by considering toy examples and stating precise definitions. It provides the required background for another article, *Markov Processes in Knowledge Spaces*, which introduces probabilistic aspects of the theory as well as knowledge assessment procedures.

Knowledge spaces are combinatorial models for encoding the conceptual organization of a given body of knowledge. Their theory leads to the design of efficient algorithms for assessing the concepts acquired by an individual in the given domain. Computer software implementing such algorithms are now in use in primary and secondary schools (e.g., Aleks). Their output is by far richer than just a crude number (e.g., 8 of 10) or mark (e.g., B). Indeed, it describes the present state of knowledge of the individual together with advice for further progress. Thus, by its combinatorial emphasis, knowledge space theory appears as a multidimensional generalization of classical test theory. It can also be applied to other settings (for instance, system failure analysis). Probabilistic extensions of the theory are covered in the article Markov Processes in Learning Spaces, while here the basic, deterministic concepts are introduced and some applications are sketched.

Organizing a Body of Knowledge

To start with an example, suppose that a teacher wants to assess the acquired capabilities of the pupils in a classroom as regards arithmetical multiplication. A list of items will be set up, starting from a simple question such as

$$\text{Item } a \quad 9 \times 8 = ?$$

and ending with a more complicate problem such as

$$\text{Item } b \quad \begin{array}{rr} & 278.9 \\ \times & 3.98 \\ \hline = & ? \end{array}$$

In classical practice, all items will be proposed to all pupils. The outcome of such a test boils down to a mark assigned to each pupil, which provides no other information than the number of correct answers. The performance of such a test is very poor. First, the pupils have to work on all problems (be these problems fairly easy or on the contrary out of reach for their actual competencies). Second, the marks convey a rather inaccurate description of the concepts and techniques mastered by each individual pupil.

These drawbacks are circumvented in oral testing. Here, the teacher can improve the examination procedure by taking into account answers previously collected in order to carefully select the next question to be asked. Moreover, this flexibility helps in building a more accurate description of the pupils' strengths and weaknesses. To illustrate these two points with our example, a teacher who checks that a pupil masters Item b will probably avoid asking Item a. On the other hand, if the pupil does not master Item b, other items will be proposed in order to diagnose which mistakes or misconceptions hamper the pupils' acquisitions. By carefully selecting each time the new item, the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.43049-7

teacher will gradually refine a description of the concepts and techniques mastered by the pupil. Such a description can be complemented with advice for remediation and a plan for further study.

Which are the factors that potentially make oral examination superior to written testing? Notice first that our analysis assumes a decomposition of the body of knowledge under testing into elementary pieces. To make things simple, we may suppose that each such piece corresponds to an item. Then, we assume some structure on the set of all these items. Back to our example, we declare Item a to be a prerequisite for Item b. As a consequence, we take for granted the following assertion:

> From a correct answer to Item b, we may surmise a correct answer to Item a (if the latter is asked).

In short, mastery of Item a is surmised from mastery of Item b. By contraposition, this rule is reformulated as

> a false answer to Item a implies a false answer to Item b.

Mathematically speaking, we have a (binary) relation $\mathcal{P}$ on the set Q of all items. The above assertions can now be written as $a\mathcal{P}b$. We call $\mathcal{P}$ the surmise relation on Q and assume that $\mathcal{P}$ is a quasi order, that is, $\mathcal{P}$ is reflexive ($q\mathcal{P}q$ holds for all q in Q) and transitive ($p\mathcal{P}q$ and $q\mathcal{P}r$ imply $p\mathcal{P}r$, for all $p, q, r \in Q$). Such a relation puts severe restrictions on the possible states of knowledge.

Knowledge States

Given the set Q of items, we identify the state of knowledge of an individual with the subset K of items that this person masters, thus $K \subseteq Q$. However, in view of the surmise relation $\mathcal{P}$, not every subset of Q can be a state of knowledge. Indeed, if a knowledge state contains q and $p\mathcal{P}q$ holds, then K must also contain p. We illustrate this with an example where $Q = \{a, b, c, d\}$ and $\mathcal{P}$ is the relation depicted in Figure 1. Subset $\{a, c\}$ is not a knowledge state since $b\mathcal{P}c$ holds, while $\{a, b, c\}$ is a knowledge state as can be checked. The full family $\mathcal{K}$ of knowledge states in this example is as follows:

$$\mathcal{K} = \{\varnothing, \{a\}, \{b\}, \{a, b\}, \{b, d\}, \{a, b, c\}, \{a, b, d\}, Q\}$$

where $\varnothing$ denotes the empty set. Notice $\{a, b\} \in \mathcal{K}$ and $\{b, d\} \in \mathcal{K}$, and also $\{a, b\} \cap \{b, d\} = \{b\} \in \mathcal{K}$.

This example can be used to illustrate some general properties of the family $\mathcal{K}$ of knowledge states of any surmise relation. The empty set $\varnothing$ and the whole set Q belong to $\mathcal{K}$; moreover, if $K, L \in \mathcal{K}$, then $K \cap L \in \mathcal{K}$ and $K \cup L \in \mathcal{K}$, that is, $\mathcal{K}$ is closed under intersection and under union. It turns out that these properties characterize families of subsets deduced from surmise relations. Indeed, this is a reformulation of Birkhoff's theorem (for infinite ground sets Q, closure under general intersection and union, not restricted to two subsets, need to be considered). We now argue that a more general setting is required.

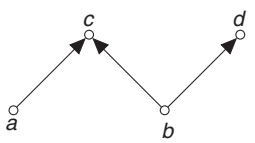

Figure 1 Diagram of a surmise relation.

Knowledge Spaces

Suppose that a student produces correct solutions to systems of linear equations in several unknowns. The examiner cannot infer from this information whether the student has mastered computation of determinants, Gauss elimination procedure, the substitution scheme, or still another method of resolution. At least one of these methods is mastered, but which one is not clear from the student's answers. A generalization of the concept of a surmise relation is thus needed. Again, let Q be a set of items. To each item q, several subsets $C_1, C_2, \ldots, C_k$ of Q will be attached. These clauses for q represent all possible ways of acquiring the mastery of q; their number k may of course vary with q. We denote by σ the mapping associating to q the collection $\{C_1, C_2, \ldots, C_k\}$. For example, assume $Q = \{a, b, c, d\}$ and

$$\sigma(a) = \{\{a, b, c\}, \{a, d\}\} \quad [1]$$

$$\sigma(b) = \{\{b, c\}\} \quad [2]$$

$$\sigma(c) = \{\{c\}\} \quad [3]$$

$$\sigma(d) = \{\{d\}\} \quad [4]$$

The expression for $\sigma(a)$ thus encodes the following fact: mastery of item a implies mastery of a itself plus (b and c) or d.

In fact, a mapping σ is another formalization of an AND/OR graph from artificial intelligence. We will now make the concept more precise by taking into consideration four axioms (satisfied by the last example).

1. For each item q, there is at least one clause for q;
2. For each item q and each clause C for q, $q \in C$;
3. For each pair of items q, r and each clause C for q containing r, there is a clause D for r such that $D \subseteq C$;
4. Any two clauses C, C' for the same item are incomparable, i.e., neither $C \subset C'$ nor $C' \subset C$ hold.

Axiom (2) clearly generalizes the reflexivity axiom of a surmise relation, while axiom (3) extends transitivity. Axiom (4) translates our assumption that a clause for item q encodes a minimal way of acquiring the mastery of q.

Given a surmise system σ on Q, we define a knowledge state as any subset K of Q satisfying the following:

> For each item q in K, there is a clause C for q that is contained in K.

In the four-item example above, the collection $\mathcal{K}$ of all knowledge states is as follows:

$$\mathcal{K} = \{\varnothing, \{c\}, \{d\}, \{a, d\}, \{b, c\}, \{c, d\}, \{a, b, c\}, \{a, c, d\}, \{b, c, d\}, Q\}$$

Any surmise relation $\mathcal{P}$ on Q becomes a surmise system if we attach to any item q the single clause $C = \{r \in Q : r\mathcal{P}q\}$. Notice that the knowledge states defined from $\mathcal{P}$ coincide with those defined from the surmise system. We have moreover a neat variant of Birkhoff's theorem. Indeed, a family $\mathcal{K}$ of subsets of a finite set Q constitutes the collection of all knowledge states of a surmise system iff $\mathcal{K}$ is closed under union, and contains $\varnothing$ and Q (see Doignon and Falmagne, 1999, which presents also an infinite version of the result). A structure $(Q,\mathcal{K})$ satisfying these two conditions is called a knowledge space.

Closure under union is not required in most practical applications. A knowledge structure $(Q,\mathcal{K})$ is a finite set Q of items together with a family $\mathcal{K}$ of subsets of Q, with $\varnothing$ and Q belonging to $\mathcal{K}$. The elements of $\mathcal{K}$ are the knowledge states.

Wellgradedness

Knowledge assessment in this context assumes that a knowledge structure $(Q,\mathcal{K})$ has been built; the goal then is to design an algorithm that determines which state in $\mathcal{K}$ represents the knowledge of an actual student. This algorithm will efficiently select questions to be asked, thus reducing at each step the remaining possible states. Clearly, if all answers collected truly reflect the student's knowledge, the process will end after a few steps with a precise element of $\mathcal{K}$.

In real life, things are not so simple. Answers may be biased by lucky guesses (even if the items are not of the multiple-choice type this can happen) or by careless errors (a slight flaw in computation may alter the correctness). To take these undeterministic aspects into account, a probabilistic extension of the theory was conceived (see Markov Processes in Learning Spaces). It relies on a combinatorial notion that we now sketch.

After some time, an assessment procedure will have proposed one state K from $\mathcal{K}$ as the accurate representation in the structure $(Q,\mathcal{K})$ of the student's knowledge. A further answer of the student, say to item q, may call for a modification of K. If $q \notin K$ and q is correctly answered, then $K \cup \{q\}$ should become the state selected in $\mathcal{K}$ – provided $K \cup \{q\}$ belongs to $\mathcal{K}$. Similarly, if $q \in K$ and q is incorrectly answered, then $K \setminus \{q\}$ should be selected – provided it belongs to $\mathcal{K}$. We may not assume that always $K \cup \{q\}$ and $K \setminus \{q\}$ belong to $\mathcal{K}$, because otherwise $\mathcal{K}$ would be the collection of all subsets of Q (which is a much too large collection for our purposes). On the other hand, the assessment procedure should be able to explore the whole collection $\mathcal{K}$ by elementary steps consisting in the addition or deletion of a single item. To state a precise definition, we use for any two subsets K, L of Q their (symmetric-difference) distance

$$d(K,L) = \left|\{q \in Q : q \in K \setminus L \text{ or } q \in L \setminus K\}\right|$$

which counts the number of items q in which K and L differ. By definition, a knowledge structure $(Q,\mathcal{K})$ is well graded when for any two knowledge states K, L in $\mathcal{K}$ with $d(K,L) = h$, there exists a sequence $K = K_0, K_1, \ldots, K_h = L$ of states in $\mathcal{K}$ such that $d(K_j, L_{j+1}) = 1$ for $j = 0, 1, \ldots, h-1$.

For K, a state in a knowledge structure $(Q,\mathcal{K})$, we define the inner fringe and outer fringe of K as

$$K^{\mathcal{I}} = \{q \in K : K \setminus \{q\} \in \mathcal{K}\} \quad [5]$$

$$K^{\mathcal{O}} = \{q \in Q \setminus K : K \cup \{q\} \in \mathcal{K}\} \quad [6]$$

For example, if

$$\mathcal{K} = \{\varnothing, \{c\}, \{d\}, \{a,d\}, \{b,c\}, \{c,d\}, \{a,b,c\}, \{a,c,d\}, \{b,c,d\}, Q\}$$

then

$$\{a,d\}^{\mathcal{I}} = \{a\} \quad [7]$$

$$\{a,d\}^{\mathcal{O}} = \{c\} \quad [8]$$

In real-life applications, the outer fringe $K^{\mathcal{O}}$ collects those items that a student in knowledge state K could directly learn without any further prerequisite. This notion plays a crucial role in the design of the Aleks software.

It can be proved that a knowledge structure $(Q,\mathcal{K})$ is well graded iff any state in $\mathcal{K}$ is characterized by its inner and outer fringes. The latter condition can be stated as follows: if $K, L \in \mathcal{K}$, and $K^{\mathcal{I}} \subset L$, $K^{\mathcal{O}} \subset Q \setminus L$, then $K = L$.

Learning Spaces

Well-graded knowledge spaces are called learning spaces (cf. Falmagne and Doignon, 2011). They are also characterized by conditions of an intuitive nature. Let $\mathcal{K}$ be a collection of subsets, called states, of a finite set Q that contains $\varnothing$ and Q – that is, $(Q,\mathcal{K})$ is a knowledge structure. The pair $(Q,\mathcal{K})$ is a learning space if it moreover satisfies.

[L1] Learning smoothness. For any two states K, L such that $K \subset L$, there exists a finite chain of states

$$K = K_0 \subset K_1 \subset \cdots \subset K_p = L \quad [9]$$

such that $|K_i \setminus K_{i-1}| = 1$ for $1 \le i \le p$ (and so $|L \setminus K| = p$).

[L2] Learning consistency. If K, L are two states satisfying $K \subset L$ and q is an item such that $K \cup \{q\} \in \mathcal{K}$, then $L \cup \{q\} \in \mathcal{K}$.

The two conditions have simple interpretations in pedagogical language: [L1] if the actual state K of a learner is included in some other potential state L, then the learner can reach state L by mastering the missing items one at a time; [L2] knowing more does not prevent learning a new item.

Aleks

Aleks is a software initially designed at the University of California, Irvine, by a team headed by Jean-Claude Falmagne (other projects with similar goals were also conceived; see below the pointer to the bibliography). Now maintained and developed by a company with the same name, Aleks routinely assesses online the mathematical knowledge of pupils and students from kindergarten to bachelor programs. Presently used in schools, colleges, and universities, it covers arithmetic, algebra, precalculus, etc., and also, outside mathematics, statistics, business, etc. For a user, a session typically presents about 10 items, carefully selected on the basis of previously

acquired information on that particular user, and also on collected answers. At the end of the session, students receive information on their progress and advice for faster study (they can even access online explanation on relevant topics). Of course, the teacher is also informed of the monitoring process. Aleks is more than an assessment system, since it provides on demand the user with explanation and access to a dictionary. From experiments, it appears that students using the Aleks system progress much further in the material than students in a conventional class.

Bibliographic Orientation

The first publication on knowledge spaces was a 1985 paper by Doignon and Falmagne. The same authors published a survey book in 1999, and a revised, enlarged edition in 2011 under the title *Learning Spaces*. Various applications of the theory are covered in two books edited by Albert (1994) and by Albert and Lukas (1998). An evolutive and searchable bibliography is maintained by Cord Hockemeyer at the following Web address:

http://wundt.kfunigraz.ac.at/hockemeyer/bibliography.html

while the Aleks Website is at

http://www.aleks.com

See also: Markov Processes in Learning Spaces.

Bibliography

Albert, D., (Ed.), 1994. Knowledge Structures. Springer–Verlag, Berlin–Heidelberg.

Albert, D., Lukas, J. (Eds.), 1998. Knowledge Spaces: Theories, Empirical Research, Applications. Lawrence Erlbaum Association, Hillsdale, NJ.

Doignon, J.-P., Falmagne, J.-Cl., 1999. Knowledge Spaces. Springer–Verlag, Berlin–Heidelberg.

Falmagne, J.-Cl., Doignon, J.-P., 2011. Learning Spaces. Springer–Verlag, Berlin–Heidelberg.

Relevant Websites

http://www.aleks.com/
http://wundt.kfunigraz.ac.at/hockemeyer/bibliography.html

Cognitive Aging

K Warner Schaie, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by K.W. Schaie, volume 3, pp. 2072–2075, © 2001, Elsevier Ltd.

Abstract

Cognitive aging is concerned with the experimental study of memory functions and language and the descriptive study of adult intellectual development. Methodological issues discussed include the issues of age-comparative versus age change designs and the role of response speed. Basic findings are presented on age differences in memory and language, age changes in intellectual competence, and on the question whether cognitive aging can be slowed or reversed by means of cognitive training.

Cognitive aging is concerned with age-related changes in adulthood in the basic processes of learning and memory, as well as the complex higher-order processes of language and intellectual competence or executive functioning. Although most of the literature has been concerned with explaining the mechanism of cognitive decline, there is also a substantial interest in issues such as compensation and the role of external support, including collaborative problem solving.

Definition of Cognitive Aging

There have been two distinct traditions in the study of cognitive aging. The first grew out of experimental child psychology while the second derives from psychometric roots.

Experimental Study of Memory Functions and Language

The concern in this literature is to explicate possible causal variables that would explain why many adults suffer memory loss and decline in the complex manipulation of language variables such as text processing. The typical approach here is to design experiments testing the effects of single variables in carefully controlled laboratory settings requiring only limited numbers of subjects. Because there is often little interest in individual differences, or population parameters, study participants.

Descriptive Study of Adult Intellectual Development

Descriptive studies of adult intellectual development often stem from the longitudinal follow-up of samples first assessed in childhood or adolescence (Schaie and Hofer, 2001). Other such studies may represent carefully stratified samples from defined populations, first assessed at a particular life stage, whether in early adulthood or in early old age. Although descriptive studies often begin as cross-sectional inquiries, they are most frequently conducted as longitudinal analyses since the primary interest of such studies is often in individual differences in intraindividual change or in the elucidation of typologies of individuals who follow different growth trajectories (*see* Adulthood: Dependency and Autonomy). These are frequently large sample studies, and the use of correlational or quasi-experimental approaches is typical (Baltes and Mayer, 1999; Schaie, 2013).

Methodological Issues

Age-Comparative versus Age Change Designs

Much of the experimental cognitive aging literature is based on age-comparative (or cross-sectional studies, which typically contrast a group of young adults, typically college students) with convenience samples of community-dwelling older adults in their 60s and 70s. It should be recognized that such comparisons are fraught by the problem that it is often unreasonable to assume that the two age groups can be adequately matched for other status variables that might provide a rival explanation for any observed age difference on the dependent variable. This creates particular problems for identifying the mechanisms that may be implicated in age-related decline from young adulthood into old age. Age-comparative designs are also inadequate in explaining individual differences in age changes (*see* Age Stratification). The latter can only be investigated by means of longitudinal paradigms (Ferrer and Ghisletta, 2011; Hofer and Sliwinski, 2006; Schaie, 1965, 1977, 2008, 2013; Sliwinski et al., 2010).

The Age-Cohort-Period Model

Early students of normal and pathological aging thought that the comparison of groups of individuals at different ages (cross-sectional data) could be used to predict and understand age changes within the same individuals (longitudinal data). A paradigm shift occurred when it was shown that such inference was not possible except under very unusual circumstances (Schaie, 1965, 1977, 2013).

The Model

The age-cohort-period model specifies that any age-related or time-dependent behavior can be assigned three temporal characteristics, such that

$$b\,(\text{Behavior}) \;=\; f\,(A + C + P)$$

where the behavior b is observed at the chronological age A, for individuals over the calendar period P who have entered the environment as members of cohort C. Just as is true for the relation among the physical variables of volume, pressure, and temperature, here also the third component can always be stated as a function of the other two components. Thus, $A = C + P$, $C = A + P$, and $P = A + C$. However, each of the three components may be of primary interest for some scientific questions in the developmental sciences, and one may therefore want to be able to estimate the specific contribution attributable to each component. In the behavioral sciences, in

particular, we typically want to differentiate effects that change across age (intraindividual change) from those effects that differ across cohorts or generations (interindividual differences).

Data Collection Strategies

Empirical studies in the developmental sciences involve age and/or cohort comparisons either at one point in time or at successive time intervals. Traditional strategies used for this purpose are represented by cross-sectional, longitudinal, and time-lag designs.

The cross-sectional strategy investigates the hypothesis that there are differences in one or more characteristics for samples drawn from different cohorts but measured at the same point in time. This strategy is most appropriate for the study of interindividual differences. Age differences in behavior at a particular point in historical time may be relevant for policy decisions that lead to differential societal responses regardless of the antecedent conditions responsible for the age differences. Age differences detected in a cross-sectional data set, however, are inextricably confounded with cohort differences. Since cross-sectional subsamples are measured only once, no information is available on intraindividual change. Unless there is independent evidence to suggest that older cohorts performed at the same level as younger cohorts at equivalent ages, it would be most parsimonious to assume, at least in comparisons of adult samples, that cross-sectional age differences represent estimates of cohort differences that may be either inflated or reduced by maturational changes occurring over a specified age range.

The longitudinal strategy investigates whether age-related changes have occurred within the same population cohort measured on two or more occasions. This strategy is appropriate when the investigator wishes to predict age differentiation in behavior that occurs over time. But longitudinal data do not always provide unambiguous estimates of intraindividual change. A single-cohort longitudinal study confounds age-related (maturational) change with period effects that are specific to the particular historical period over which the behavior is monitored. The period effects could either mask or grossly inflate estimates of maturational changes.

The time-lag strategy compares two or more samples of individuals drawn from successive cohorts at successive points in time at the same chronological age. The hypothesis tested is whether there are differences in a given behavior for samples of equal age but drawn at different points in time. This strategy is of particular interest to social and educational psychologists. It is particularly appropriate when one wishes to study performance of individuals of similar age in successive cohorts (e.g., comparing baby boomers with the preceding generation). The simple time-lag design confounds cohort effect with period effects and may provide inflated or reduced cohort estimates depending on whether the temporal interval between the cohorts represents a period of favorable or adverse environmental influences.

Sequential Strategies

Several alternative sequential strategies are available that might differentiate effects of maturational characteristic for a particular developmental period from the attainment of different levels of functioning attributable to differences in socialization and/or other life experiences characteristic for successive generation by assessing the behavior of more than one cohort over a given age range (Schaie, 1977; Schaie and Willis, 1999: Chapter 5; specific implications for the study of emotion are discussed in Schaie, 2001).

The term 'sequential' implies that the sampling strategy used to study generational differences must include the acquisition of a sequence of samples taken across several measurement occasions. Perhaps the most widely used sequential strategy is the 'cross-sequential' design, in which two or more cohorts are followed by an identical time period. This approach permits the direct comparison of longitudinal and cross-sectional data (provided that the calendar time ranges are similar for age and cohort). The advantage of this approach is that only two points in time are needed; hence the early appearance in the literature of studies using this design. For purposes of studying generational differences, however, this approach represents a 'model misspecification' because it does not allow comparing each cohort over the same age range.

Geropsychologists and other developmental scientists often find the cohort-sequential design of greatest interest because it explicitly differentiates intraindividual age changes that occur within a generation from interindividual differences between generations. This design also permits a check of the consistency of age functions over successive generations, thereby offering greater external validity than would be provided by a single-cohort longitudinal design. A cohort-sequential study consists of two or more generations (however defined) being followed over two or more similar age levels. The minimum design for such a study involves three measurement points, allowing each of two cohorts to be followed over the same age range (also see Sliwinski et al., 2010).

The Role of Response Speed

A number of theorists have argued that changes in the central nervous system are the primary common cause for the observed age-related declines in cognitive performance. In fact, there have been many published analyses that show a substantial reduction in age differences, if some measure or measures of reaction time or perceptual speed is partialled out of the relation between a given cognitive process and chronological age (Hartley, 2006; Madden, 2001; Salthouse, 1999). This issue is of particular concern because it is not clear whether the observed average increase in reaction time (generally assumed to be of the magnitude of approximately 1.6 from the early 20s to the late 60s), while reliably demonstrable in the laboratory, is of significance in many or most tasks of daily living.

Basic Findings from the Experimental Literature on Cognitive Aging

Most of this literature is cross-sectional in nature and usually consists of a comparison of convenience samples of young adults (often sophomore psychology students) and of community-dwelling older adults (often

participants in adult education programs). The major findings regarding age differences in cognitive performance include the following.

Memory

It is currently thought that older persons are at a disadvantage in retrieving information from memory when the information to be retrieved is complex and when there are few cues or other environmental support (*see* Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences). Hence, age differences are far greater in recall than in recognition of information. It is also thought that the magnitude of age difference in memory is far greater when a task involves effortful processing than when automatic processing is involved. Hence, greater age differences have been found for explicit than implicit memory. Older persons are also thought to have greater difficulty in integrating the context of information they are trying to remember. It is also thought that working memory capacity (i.e., the information kept in immediately accessible store) becomes reduced with increasing age. On the other hand, there is little evidence for age differences in long-term storage. Memory deficits occurring with age include nonverbal tasks such as memory for spatial location, faces, and actions and activities. Studies of prospective memory (i.e., remembering something to be done in the future) suggest that older people do well in remembering simple and event-based tasks, but are at a disadvantage when tasks become complex or are time-based. In sum, it appears that age differences are known to increase in magnitude as a function of the processing requirement of a given task (Hoyer and Verhaeghen, 2006; Salthouse, 1999; Smith, 1996). More recently, reductions in cortical volume particularly in the hypothalamus and reduction in functional brain activity have also been implicated in the reduction of memory function in the elderly (cf Nyberg and Bäckmann, 2011).

Language

Age-related differences in language behavior are closely related to the processes of encoding and retrieving verbal materials discussed above. But in addition there appear to be greater age differences in textual tasks that involve recent connections than in those that involve recollection of older connections.

Language production also seems to be adversely affected in older persons under intense time pressure. The interesting tip-of-the-tongue phenomenon involving word-finding difficulty, however, seems to be more likely with infrequently used words. Significant age differences have also been found in language planning, that is, in planning what one intends to say and how to say it during language production. Hence, older persons are more likely to engage in hesitations, false starts, and repetitions. Age-linked deficits in story recall are thought to be more of a general deficit in connection formation than in specific communication ability. Older persons tend to benefit from textual material that provides priming of association because it contains learned semantically linked information (McKay and Abrams, 1996; Thornton and Light, 2006).

Basic Findings from the Descriptive Literature on Age Changes in Intellectual Competence

Changes in intellectual competence over the adult life span have been studied primarily with either the Wechsler Intelligence Scale or with ability batteries derived from the Thurstonian Primary Mental Ability framework.

Distinctions are made between fluid abilities thought to be innate and crystallized abilities, which involve the utilization of culturally acquired knowledge (Cattell, 1963). More recently further distinctions have been introduced between the mechanics (or basic processes) of intellectual competence and the pragmatics that involve cultural mediation (Baltes et al., 1984).

What has been found in most longitudinal studies is that the adult life course of mental abilities is not uniform. The so-called fluid abilities (sometimes defined as cognitive mechanics or primitives) peak in early midlife and begin to decline in the early 60s. By contrast, crystallized abilities, or the pragmatics of intellectual competence that represent abilities acquired in a given cultural context (particularly verbal abilities), do not usually peak until the 50s are reached and begin to show significant decline only in the 70s and often show only minimal decline even in the 80s (Schaie, 2013). See **Figure 1**.

At any particular time, a cross-sectional snapshot may yield very different ability profiles because of the fact that subsequent population cohorts reach different asymptotes in midlife. For example, there has been a positive linear cohort trend in the twentieth century for inductive reasoning, the basic component of most problem-solving tasks, while there has been a negative trend in numeric skills. The magnitude of cohort differences in abilities since the 1950s has been comparable to the average age changes observed from young adulthood into the 70s. Hence, many older persons may appear to have declined markedly in comparison to young peers, even though the age difference may be primarily due to what might be called the obsolescence of earlier cohorts. See **Figure 2** for an example from the Seattle Longitudinal Study (Schaie, 1996, 2013).

Studies of individual differences suggest that while most persons have declined on some aspect of intellectual functioning from their own peak as the 60s are reached, that specific pattern of decline may well depend on complex patterns of individual life experience. Most healthy community-dwelling persons are able to maintain a high level of function until advanced old age (but see Baltes and Mayer, 1999 for the consequences of sensory dysfunctions). Because most tasks of daily living represent complex combinations of basic cognitive processes, many individuals can maintain their abilities above the minimally necessary threshold level for independent functioning by often rather complex compensatory processes (cf Baltes et al., 1984, 1999).

Can Cognitive Aging be Slowed or Reversed?

Once the course of adult intellectual development had been described and a number of antecedents of individual differences had been identified, it then became useful for researchers to think about ways in which normal intellectual

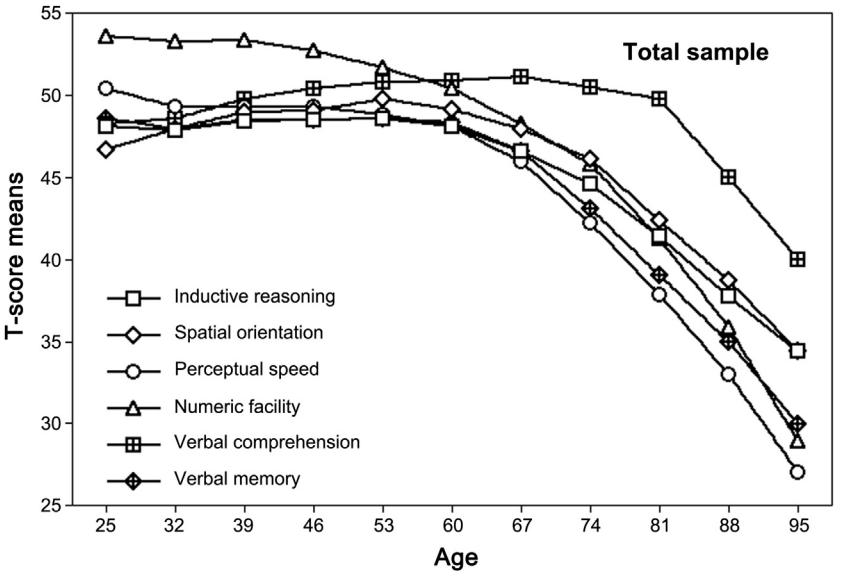

Figure 1 Longitudinal age changes for six latent mental abilities. From: Schaie, K.W., 2013. Developmental Influences on Adult Intelligence: The Seattle Longitudinal Study. Oxford University Press, New York, Figure 5.7, p. 162. Reproduced by permission of the publisher.

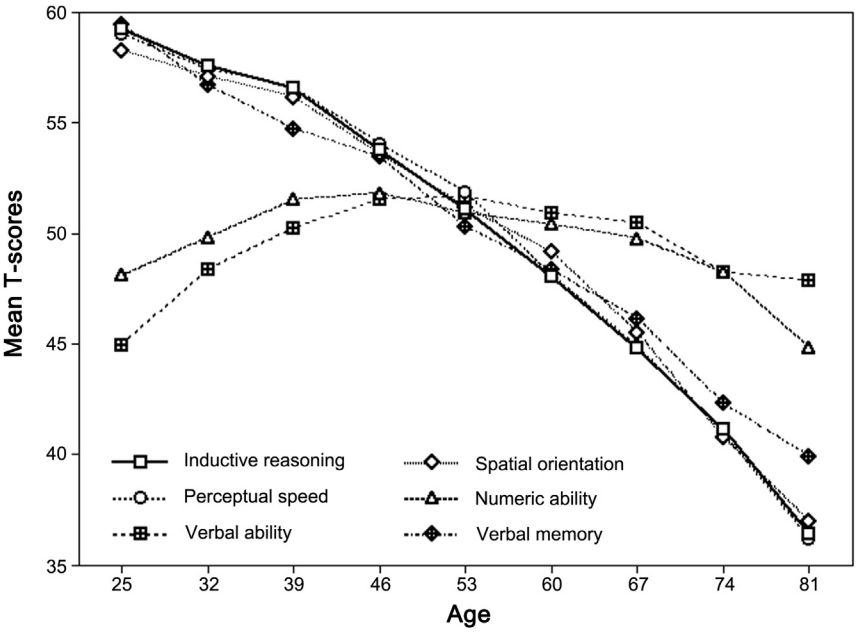

Figure 2 Cross-sectional age difference patterns for six latent mental abilities. From: Schaie, K.W., 2013. Developmental Influences on Adult Intelligence: The Seattle Longitudinal Study. Oxford University Press, New York, Figure 4.7, p. 120. Reproduced by permission of the publisher.

aging might be slowed or reversed (Stine-Morrow and Basak, 2011; Willis, 2001).

In a number of laboratories (primarily in the United States and in Germany), cognitive training programs have been developed that have been applied in the laboratory and more recently in cooperative multisite intervention trials. In contrast to training young children, where it can be assumed that new skills are conveyed, older adults are likely to have access to the skills being trained, but through disuse have lost their proficiency. Information from longitudinal studies is therefore particularly useful in distinguishing individuals who have declined from those who have remained stable. In the former, training is directed toward remediation of loss, while in the latter the enhancement of previous levels of functioning are sought with the intention of compensating for possibly cohort-based disadvantage of older persons.

Results from such cognitive interventions allow the conclusion that cognitive decline in old age, for many older persons, is likely to be a function of disuse rather than of the deterioration of the physiological or neural substrates of

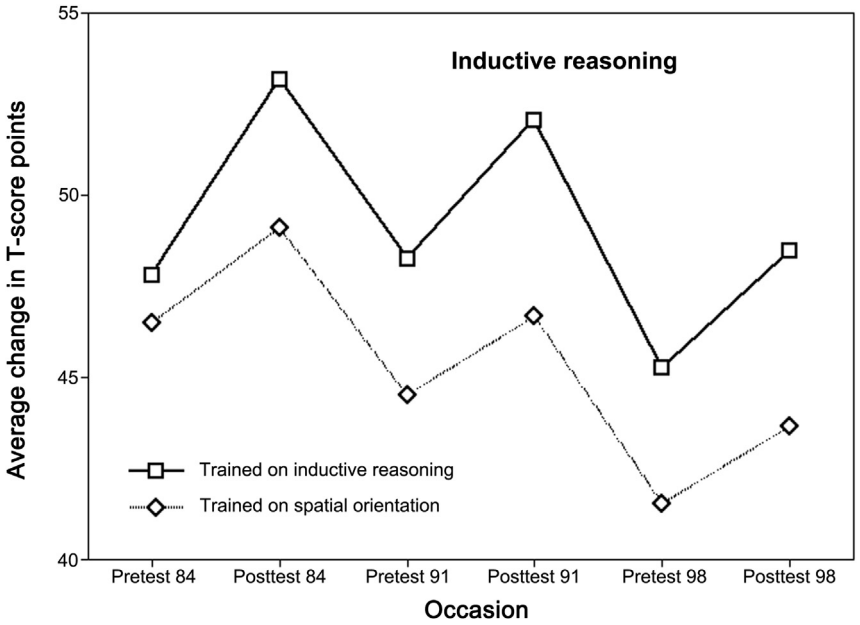

Figure 3 Pre- and post-test findings at initial training and for the 7- and 14-year follow-up for the mental ability of inductive reasoning. From: Schaie, K.W., 2013. Developmental Influences on Adult Intelligence: The Seattle Longitudinal Study. Oxford University Press, New York, Figure 7.12, p. 221. Reproduced by permission of the publisher.

cognitive behavior. For example, a brief 5-hr training program for persons over 65 years of age resulted in average training gains of about one-half SD on the abilities of spatial orientation and inductive reasoning. Of those for whom significant decrement could be documented over a 14-year period, roughly 40% were returned to the level at which they had functioned when first studied. **Figure 3** shows pre- and post-test performance levels at initial training and at 7- and 14-year follow-up for the primary ability of inductive reasoning.

The analyses of structural relationships among the ability measures prior to and after training further allow the conclusion that training does not result in qualitative changes in ability structures and is thus highly specific to the targeted abilities (Schaie et al., 1987). A 7-year follow-up further demonstrated that those subjects who showed significant decline at initial training do remain at substantial advantage over untrained comparison groups (Schaie, 2013; Willis and Schaie, 1994).

Multiple site randomized trials of cognitive interventions have also been proved to be successful for programs seeking to enhance elders' performance on working memory, inductive reasoning, and speed of processing (e.g., Willis et al., 2006). It should be noted, moreover, that while cognitive training may improve performance in the elderly and may function to reduce effects of age decrement, such training will also be effective in enhancing the performance of young adults such that age differences tend to remain robust (cf Baltes and Kliegl, 1992). More recently, there has also been considerable interest on the effectiveness on collaborative learning in increasing cognitive performance and the question of extending cognitive training directly to activities of daily living.

Other Related Topics in Cognitive Aging

Much of the work on cognitive aging in the past has been concerned with age-related development in the mechanics and basic processes of cognition. It should be recognized that current attention in the study of cognitive aging is turning to the discovery of how these basic processes operate within more complex domains. Of particular interest here are the study of wisdom (e.g., Ardelt, 2011; Baltes and Staudinger, 1993; Sternberg, 1990), the application of the basic processes to social cognition (e.g., Staudinger, 1999), the development of expert systems (e.g., Charness and Bosman, 1990), and in everyday problem solving (Willis, 1996). The extensive literature on these topics is beyond the scope of this article.

See also: Aging Mind: Facets and Levels of Analysis; Aging and Health in Old Age; Aging and Memory; Aging, Theories of; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Differential Aging; Ecology of Aging; Old Age and Centenarians: The Human 'Warranty Period'; Social Cognition.

Bibliography

Ardelt, M., 2011. Wisdom, age and well-being. In: Schaie, K.W., Willis, S.L. (Eds.), Handbook of the Psychology of Aging, seventh ed. Academic Press, San Diego, CA, pp. 279–291.

Baltes, P.B., Dittmann-Kohli, F., Dixon, R.A., 1984. New perspectives on the development of intelligence in adulthood: toward a dual process conception and a model of selective optimization with compensation. In: Baltes, P.B., Brim Jr., O.G. (Eds.), Lifespan Development and Behavior, vol. 6. Academic Press, New York, pp. 33–76.

Baltes, P.B., Kliegl, R., 1992. Further testing of limits of cognitive plasticity: negative age differences in a mnemonic skill are robust. Developmental Psychology 28, 121–125.
Baltes, P.B., Lindenberger, U., 1997. Emergence of a powerful connection between sensory and cognitive functions across the adult life span: a new window at the study of cognitive aging. Psychology and Aging 12, 12–21.
Baltes, P.B., Mayer, K.U. (Eds.), 1999. The Berlin Aging Study: Aging from 70 to 100. Cambridge University Press, Cambridge, UK.
Baltes, P.B., Staudinger, U.M., 1993. The search for a psychology of wisdom. Current Directions in Psychological Science 2, 75–80.
Baltes, P.B., Staudinger, U.M., Lindenberger, U., 1999. Lifespan psychology: theory and application to intellectual functioning. Annual Review of Psychology 50, 471–507.
Cattell, R.B., 1963. Theory of fluid and crystallized intelligence: a critical experiment. Journal of Educational Psychology 54, 1–22.
Charness, N., Bosman, E.A., 1990. Expertise and aging: life in the lab. In: Hess, T.H. (Ed.), Aging and Cognition: Knowledge Organization and Utilization. Elsevier, Amsterdam, pp. 343–385.
Ferrer, E., Ghisletta, P., 2011. Methodological and analytical issues in the psychology of aging. In: Schaie, K.W., Willis, S.L. (Eds.), Handbook of the Psychology of Aging, seventh ed. Academic Press, San Diego, CA, pp. 25–40.
Hartley, A., 2006. Changing role of the speed of processing construct in the cognitive psychology of human aging. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, sixth ed. Academic Press, San Diego, CA, pp. 183–208.
Hofer, S.M., Sliwinski, M.J., 2006. Design and analysis of longitudinal studies on aging. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, sixth ed. Academic Press, San Diego, CA, pp. 17–40.
Hoyer, W.J., Verhaeghen, P., 2006. Memory aging. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, sixth ed. Academic Press, San Diego, CA, pp. 209–232.
Kramer, A.F., Fabiani, M., Colcombe, S.J., 2006. Contributions of cognitive neuroscience to the understanding of behavior and aging. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, sixth ed. Academic Press, San Diego, CA, pp. 57–84.
Madden, D.J., 2001. Speed and timing in behavioral processes. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, fifth ed. Academic Press, San Diego, CA, pp. 288–312.
Margrett, J.A., Willis, S.L., 2006. In-home cognitive training with older married couples: individual versus collaborative learning. Aging, Neuropsychology and Cognition 13, 173–195.
McKay, D.G., Abrams, L., 1996. Language, memory, and aging: distributed deficits and the structure of new-versus-old connections. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, fourth ed. Academic Press, San Diego, CA, pp. 251–265.
Nyberg, L., Bäckmann, L., 2011. Memory changes and the aging brain: a multitmodal imaging approach. In: Schaie, K.W., Willis, S.L. (Eds.), Handbook of the Psychology of Aging, seventh ed. Academic Press, San Diego, CA, pp. 121–132.
Salthouse, T., 1999. Theories of cognition. In: Bengtson, V.E., Schaie, K.W. (Eds.), Handbook of Theories of Aging. Springer Publishing Co., New York, pp. 196–208.
Schaie, K.W., 1965. A general model for the study of developmental problems. Psychological Bulletin 64, 92–107.
Schaie, K.W., 1977. Quasi-experimental designs in the psychology of aging. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging. Van Nostrand Reinhold, New York, pp. 39–58.
Schaie, K.W., 2001. Emotional intelligence: psychometric and developmental characteristics: Comments on Roberts, Zeidner and Matthews. Emotion 1, 243–246.
Schaie, K.W., 2006. The concept of event time in the study of aging. In: Baars, J., Visser, H. (Eds.), Concepts of Time and Aging. Baywood Publishing Co., Amityville, NY, pp. 121–135.
Schaie, K.W., 2008. A lifespan developmental perspective of psychological aging. In: Laidlaw, K., Knight, B.G. (Eds.), The Handbook of Emotional Disorders in Late Life. Oxford University Press, Oxford, UK, pp. 3–32.
Schaie, K.W., 2013. Developmental Influences on Adult Intellectual Development: The Seattle Longitudinal Study, second rev. ed. Oxford University Press, New York.
Schaie, K.W., Hofer, S.M., 2001. Longitudinal studies in research on aging. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, 5th ed. Academic Press, San Diego, CA, pp. 53–77.
Schaie, K.W., Willis, S.L., 1999. Theories of everyday competence. In: Bengtson, Schaie, K.W. (Eds.), Handbook of Theories of Aging. Springer, New York, pp. 174–195.
Schaie, K.W., Willis, S.L., Hertzog, C., Schulenberg, J.E., 1987. Effects of cognitive training upon primary mental ability structure. Psychology and Aging 2, 233–242.
Sliwinski, M., Hoffman, L., Hofer, S.M., 2010. Evaluating convergence of within-person change and between-person age differences in age-heterogeneous longitudinal studies. Research in Human Development 7, 45–60.
Smith, A.D., 1996. Memory. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, fourth ed. Academic Press, San Diego, CA, pp. 236–250.
Staudinger, U.M., 1999. Social cognition and psychological approach to an art of life. In: Blanchard-Fields, F., Hess, T.B. (Eds.), Social Cognition, Adult Development and Aging. Academic Press, San Diego, CA, pp. 343–375.
Stine-Morrow, E.A.L., Basak, C., 2011. Cognitive interventions. In: Schaie, K.W., Willis, S.L. (Eds.), Handbook of the Psychology of Aging, seventh ed. Academic Press, San Diego, CA, pp. 153–171.
Sternberg, R.J., 1990. Wisdom and its relation to intelligence and creativity. In: Sternberg, R.J. (Ed.), Wisdom: Its Nature, Origins, and Development. Academic Press, New York, pp. 142–149.
Thornton, R., Light, L.L., 2006. Language comprehension and production in normal aging. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, sixth ed. Academic Press, San Diego, CA, pp. 262–288.
Willis, S.L., 1996. Everyday problem solving. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, fourth ed. Academic Press, San Diego, CA, pp. 287–307.
Willis, S.L., 2001. Methodological issues in behavioral intervention research with the elderly. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, 5th ed. Academic Press, San Diego, CA, pp. 78–108.
Willis, S.L., Schaie, K.W., 1994. Cognitive training in the normal elderly. In: Forette, F., Christen, Y., Boller, F. (Eds.), Plasticité cerebrale et stimulation cognitive. Fondation Nationale de Gerontologie, Paris, pp. 91–113.
Willis, S.L., Tennstedt, S.L., Marsiske, M., Ball, K., Elias, J., Koepke, K.M., Morris, 2006. Long terms effects of cognitive training on everyday functional outcomes in older adults. In: Rebok, G., Unverzagt, F.W., Stoddard, A.M., Wright, E. (Eds.), Journal of the American Medical Association 296, 2805–2814.

Cognitive Anthropology

Jürg Wassmann, Institute of Anthropology, University of Heidelberg, Heidelberg, Germany
Andrea Bender, University of Freiburg, Freiburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Situated at the intersection of cultural anthropology and cognitive science, the main concern of cognitive anthropology is to explain the mental representations and processes that make up cultural content. To this end, cognitive anthropologists combine approaches and methods from cognitive science with traditional ethnography and with techniques specifically developed for this very purpose. In this article, we first present a brief history of cognitive anthropology and of its core concepts and main methods. This is then rounded off with a perspective on the contributions that cognitive anthropology has to offer to cognitive science, including, among others, a concern not only with knowledge, but also with knowing, insights into the relation of language and thought, and a corrective for the (often untested) assumption of universality in cognitive processes.

One of the central objects of research in cultural anthropology is culture, in at least two respects: as cultural phenomena in a material sense, and as the cultural content based on mental representations. Although most cultural phenomena are public and thus accessible to ethnographic documentation, this is not true for cultural content that, instead, is largely situated inside people's heads. Explaining these mental representations and processes is the main concern of *cognitive anthropology*, and it is also where cultural anthropology intersects with other cognitive sciences. Together with cognitive psychology, philosophy, cognitive neuroscience, artificial intelligence, and linguistics, cognitive anthropology aims at understanding how people perceive, learn, process, store, and use knowledge of the world or, more generally, how the information received through the senses is interacting with thought processes (Gardner, 1985; Sternberg, 1999; Boden, 2006). In this alliance, however, anthropology has become a rather shadowy partner, which is unfortunate given that some of the theories emerging in cognitive science are also central to the concerns of anthropology.

The prevailing approach in the cognitive sciences is to analyze the human experience of the world and human thinking as information processing, including perception, motivation, emotion, problem solving, and memory. The objective is to make transparent, among other things, the mental faculties that enable all humans to internalize the social and cultural characteristics of the society into which they are born. And this interest overlaps substantially with central topics of anthropological research:

> …anthropologists' concerns place them right in the middle of the cognitive sciences, whether they like it or not, since it is cognitive scientists who have something to say about learning, memory and retrieval (Bloch, 1991, p. 184).

Cognitive anthropology seeks to contribute to these issues by combining cognitive science with traditional ethnography (Keesing, 1987; Tyler, 1969; Wassmann, 1994; Wassmann et al., 2011; Wassmann, n.d.; Bender and Beller, 2013). This bridge building is particularly valuable, given that the definitions of cognition (and many related concepts) used by mainstream anthropologists are often too sociological to be compatible with the approaches used by cognitive scientists. Lave (1988), for instance, sees cognition as originating in social practices of individuals, and Geertz (1973, p. 45) proposes that "thought is, in a great part anyway, a public activity." It can therefore not be surprising that anthropologists' theories of thinking, meaning, memory, and retrieval do not converge with those of cognitive scientists.

The limitations of the purely cultural focus was realized, for instance, by (psychologist) Rogoff and (anthropologist) Lave (1984), who advocated a shift away from Durkheim's cultural *représentation collective* toward a more central role for the individual. However, they also pointed out that relying on single individuals in special social roles (e.g., village principals) as 'omniscient informants' is not sufficient, but needs to be complemented by including individuals in nonexposed roles, the so-called plain folks. Even in common worlds, single lives differ (Keck, 1998). In fact, the tremendous variability of individual biographies renders the idea of one homogenous culture (at any scale, applied to an individual or the society) nonsensical.

Of course, the plain-folk individual in isolation, stripped of any cultural ties, is a limited model system too. And yet, it is this view that has been – and to some extent still is – characteristic of most cognitive science research. Apart from some socio-psychological strands, as in cross-cultural psychology (Berry et al., 2002; Mishra and Dasen, 2007), experimental work is often carried out on individuals in isolation and in highly nonnatural (but carefully controlled) environments. But this is only one of several characteristics that anthropologists consider critical. Another point for critique regards the computational model of the mind that has led scholars to neglect the interaction of content and process and the relevance of the social and material context of cognitive activities (Bender and Beller, 2011b, p. 1). Meanwhile, scholars have come to understand that cognition cannot be fully understood or explained without reference to a specific body (embodiment) and to a specific cultural environment (situatedness).

One of the implications of the computational model was the assumption that cognitive processes are universal – an

assumption compatible with the conception of 'psychic unity of humankind' proposed by anthropological founding fathers such as Franz Boas. Along the same lines, Cole and Scribner tried to identify basic capacities that seemed to be invariant across different cultural context. Although we may indeed assume that most cognitive and affective competencies do exist in all humans, the way in which they manifest as performance and emerge as 'cognitive styles' (Dasen and Mishra, 2010) depends on the cultural and ecological environment. This implies that the cognitive processes themselves may be subject to cultural influences (Atran and Medin, 2008; Bender and Beller, 2011b).

As a consequence, precise knowledge of and about such general cognitive competencies and processes is indispensable for any scientific field concerned with human biology and psychology or human nature and culture. It could also provide guidelines for finding the right balance between an individual and a cultural perspective on cognition and culture which, in turn, would be of the utmost relevance for both the field-working anthropologist and the brain-scanning neurobiologist. In order to understand better what cognitive anthropology has to offer in this debate, however, we first present a brief history of the discipline and an overview of some of its core concepts and methods.

From Ethnoscience to the Cognitive Sciences

The year of the 'cognitive revolution' to which, apart from cultural anthropology, various human sciences contributed, is considered to be 1956 (Gardner, 1985). In that year, a conference concerning information theory took place at the Massachusetts Institute of Technology, where Newell and Simon delivered a paper on computer programs. Miller presented his famous treatise, 'The Magical Number Seven,' and the 28-year-old Chomsky read excerpts from his thesis, 'Three Models of Language.' In the same year, Bruner's book *A Study of Thinking* came out, and two anthropologists, Goodenough and Lounsbury, published the first programmatical articles on what then came to be known as cognitive anthropology: 'Componential Analysis and the Study of Meaning' and 'A Semantic Analysis of the Pawnee Kinship Usage.'

With cognitive anthropology – or, more precisely, with the ethnoscientific phase of cognitive anthropology – a new field of research came to the fore. Its goal was to describe other cultures in their own conceptualization, that is, in *emic* terms, or 'frominside.' Because various cultures categorize the world differently and appear to apply different types of logic in dealing with it, these differences need to be recorded and analyzed to explore the differing cognitive worlds. The underlying question was an old one, posed by another founding father of anthropology, Malinowski: What does 'order out of chaos' look like? More precisely, scholars were interested in how (other) cultures label the things in their environments and how those labels are related to each other. The underlying assumption was that, with the help of cultural phenomena, the mental representations that make up the cultural content could also be understood.

Three Premises

Three premises formed the background of this ambitious program (see Dougherty, 1985):

Premise 1: Culture is Common (Shared) Knowledge

The first and most influential definition of culture as knowledge – and thus as, in principle, a *mental* phenomenon – was given by Goodenough:

> A society's culture consists of whatever it is one has to know or believe in order to operate in a manner acceptable to its members, and to do so in any role that they accept for any one of themselves.… It is the forms of things that people have in mind, their models for perceiving, relating, and otherwise interpreting them.… Culture does not exist of things, people, behaviour, or emotions, but in the forms or organizations of the things in the minds of people (1957, pp. 167–168).

Premise 2: Knowledge Systems Provide a Cultural Grammar

On the basis of the statements made by their informants, anthropologists inductively discover this abstract and shared knowledge as a systematic mental representation. In principle, this could be done with only one informant, just as learning a foreign language requires only one native speaker as one's teacher. The knowledge system of a culture is understood to be a conceptual model that embraces the organizational principles of the culture and the behavior of its members. In this sense, the model is a 'cultural grammar.'

Premise 3: Language is the Best Means of Access to Mental Phenomena

With the 'reduction of chaos,' certain phenomena and characteristics are selected from the environment as being significant are named and given a classificatory meaning. The main (but not the only) proof of a category's existence is its label, and this label conversely provides access to the category.

The equation of culture and knowledge initially proved to be very fruitful, and in the 1960s, ethnoscience experienced rapid success. Innumerable studies were published about terminologically densely structured individual fields, such as kinship, colors, ethnozoology, ethnobotany, or illness (for an overview, see D'Andrade, 1995). One succumbed, as it were, to the great theoretical temptation to reduce complex and ostensibly heterogeneous domains to a few rules (inclusion, exclusion, and intersection) and to present them as elegant models (taxonomies or paradigms; see below) that were looked upon as mental representations of individuals.

At the beginning of the 1970s, however, only a few studies of this descent were still published, and in 1972, Keesing felt justified to start a paper with the sentence: "Whatever happened to ethnoscience?" (1972, p. 299). Ironically, when, at the end of the 1950s, ethnoscience had taken over the phonological model from structural linguistics to become the focus in the 1960s, this same model had already given way in linguistics itself to Chomsky's new generative linguistics.

As a consequence, anthropology had to open up and turn toward modern trends in neighboring disciplines. Computer science exerted a particularly strong influence. When the first computer programs started to 'play' chess, cognitive

anthropologists adopted an information-processing approach and attempted to transfer the concept of programs to describe and explain the way in which culture is both structured and structuring human thinking and behavior. The underlying assumption was that cognitive processes could be modeled in the same way, regardless of whether they take place in humans, animals, or machines. In other words, the software was assumed to be the same everywhere, irrespective of the hardware in which it was processed.

Revised Premises

In light of this theoretical and conceptual reorientation, the three premises of ethnoscience had to be reconsidered and extended: While the basic question remained the same, namely, how knowledge is structured (to create 'order out of chaos'), the novel twist to it was where knowledge is sought and how it is represented.

Revised Premise 1: Turning Toward the Individual

With the insight that individual cognitive processes or representations cannot be directly inferred from cultural phenomena came a shift in focus away from the collective knowledge system as a whole (to be recorded as the ideal type) toward the scattered, variable knowledge acquired, memorized, and applied by individuals in their everyday lives.

Revised Premise 2: Turning Toward Operationalization Instead of Categorization

As soon as knowledge is no longer conceived of as an isolated, static system (that is, as 'grammar') but as revealing itself (verbally or nonverbally) in everyday use by individuals, many categories and semantic fields were regarded as fuzzy sets that lacked precise or fixed boundaries. These categories can be organized by best exemplars, or prototypes, or they can be organized according to what a person typically does in daily life: 'taskonomy,' instead of 'taxonomy.' *Taskonomies* refer to everyday cognition that a housewife needs when shopping, a milkman when distributing dairy products to his customers according to a certain pattern, a Yakan in the Philippines when entering a house correctly, and it is this everyday cognition that anthropologists then became interested in.

Revised Premise 3: Turning away from Language as the Only Instrument to Code Knowledge

Knowledge is expressed not only through language but also by means of actions and emotions. Habitual actions, in particular, can be very eloquent in revealing tacit knowledge. To describe stereotyped sequences of actions in certain situations, the computer scientist Schank and the social psychologist Abelson, in 1977, introduced the notion of *scripts*. Language remained of interest but was treated differently, no longer as a lexicon (or grammar) but as discourse from which inferences had to be drawn about the intended message. No longer was it assumed by default that the structure of knowledge (as represented in the mind) has to be language-like:

> Knowledge organized for efficiency in day-to-day practice is not only non-linguistic, but also not language-like in that it does not take a sentential logical form (Bloch, 1991, p. 189–190).

Some pushed this critique even further by calling for ignoring language more frequently because certain kinds of knowledge cannot be expressed linguistically, or only with great difficulty.

The fact that individuals and how they act in their daily lives are now in the focus of interest is a consequence of a paradigmatic shift. Cognition is no longer regarded as an expression of culture as a whole, abstracted from linguistic material, but as the mental activity of individuals who make flexible use of their knowledge in different contexts, in that they think, generalize, draw inferences, perceive, recognize, and categorize; analyze, combine, assess possibilities, solve problems, and make decisions; classify, differentiate, and choose; remember and master new situations. Although these activities are performed individually, or between individuals (Hutchins, 1995), they always and everywhere take place within culture.

The Representation of Knowledge

The most important models of representing knowledge in cognitive anthropology are briefly described below.

Taxonomy

The inner order of a semantic field (such as furniture or colors) depends on a small number of ordering principles that structure the lexical components (lexemes) of the field (e.g., 'table,' 'red'). The ordering principles are inclusion, exclusion (contrast), and intersection.

A taxonomy provides a description of such a semantic field; it lists the categories (lexemes) and shows how they are connected with each other: that is, according to the principles of exclusion and inclusion in hierarchic order. Categories at the same level *exclude* each other, whereas categories at the lower level are *included* in the categories at the higher level (Example: Table 1).

The grouping shows that dining tables are different from smokers' tables but that both are kinds of tables. Chairs, on the other hand, are distinguished from tables. The grouping does not say which are the distinguishing characteristics.

Paradigm

If a semantic field is structured according to the principle of intersection, we have a paradigm. Hierarchy, inclusion, and exclusion are missing; instead, the *distinctive features* (or criterial attributes) distinguish the categories. Paradigms are constructed by a componential analysis. A complete lexicon of a semantic

Table 1 Taxonomy

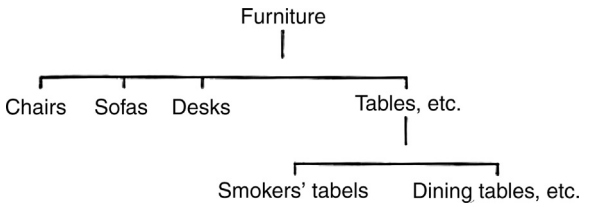

Table 2 Paradigm

		Sex	
		a1	*a2*
Generation	b1	Grandfather	Grandmother
	b2	Father	Mother
	b3	Ego	
	b4	Son	Daughter
	b5	Grandson	Granddaughter

field (e.g., blood relationship) is established and then those characteristics (components) are looked for, according to which the lexemes differ, for example, male and female in the dimension sex, or the distance from Ego in the dimension generation. Every single lexeme is now defined by a bundle of components. In a graphic representation, these components intersect at the defined lexeme (Example: Table 2).

Taxonomy and paradigm were consistently emic and were the classic models of ethnoscience. They were strongly idealized and understood as appropriate tools to characterize a culture. As we know today, however, they are only to a small degree cognitive and not really instruments of thinking.

Prototype

A prototype is the object within a category that is thought to be its best example. In the category furniture, for instance, chair is a better example than, say, radio. Of course we could define *furniture* as a category of objects having certain semantic characteristics or attributes in common (and whatever does not have these characteristics does not belong). However, this checklist definition does not allow any grading, such as a chair being a better example, more of a prototype of furniture, than radio (although both belong to the category of furniture). It may happen that there are no criterial attributes common to all parts (i.e., no semantic field, no class, according to ethnoscience) but only a large number of characteristics that may match some but by no means all category members. As a consequence, they only have a 'family resemblance' structure – a term Rosch and Mervis adopted from Wittgenstein. Following Lévi-Strauss, prototypes are particularly 'good to think' (Example: Table 3).

Script, Schema, and Cultural Model

The information theory approach forces the anthropologist to be explicit about what normally remains implicit, and some of this is captured by scripts. Scripts are used to represent and/or describe procedural knowledge, including the structure of the activities or events they represent. Scripts describe the tacit knowledge that enables us also to understand incomplete descriptions and suggestions; we automatically add what is missing by an inference process. Every situation requires specific knowledge and, accordingly, there are scripts for eating in a restaurant, playing football, and attending a birthday party. But not only our actions are based on scripts, our language is as well – a story told with all the details would be tedious.

Table 3 Prototypes

Prototype	*Category*
Chair	Furniture
Car	Vehicle
Orange	Fruit
Gun, etc.	Weapon, etc.

Example:
I am in New York and somebody asks me the way to Coney Island; I tell him to take the N-train to the terminus.
This instruction makes sense only "if this improperly specified algorithm can be filled out with a great deal of knowledge about how to walk, pay for the subway, get in the train and so on" (Schank and Abelson, 1977, p. 20).

If the typical characteristics of a situation are grasped – hence the stereotypical, the standard-like is stressed, but raised to a higher level of abstraction – we can talk about schemata and cultural models (which partially replace the older term, *folk models*). All the knowledge we acquire, remember, and communicate about this world is not a simple reflection of this world, nor does it consist of a series of categories (as ethnoscience assumed), but it is organized into different situation-relevant, prototypical, simplified sequences of events (Quinn and Holland, 1987, p. 32). What is more, these models are probabilistic and partial; they are actual frames we can also use to react to new situations. They are world-proposing "yet cannot be directly observed since they are not presented, but merely represented by the behavior of the people. They are models *of* the mind and in the mind" (D'Andrade, 1989, p. 824). The organizing principle behind these models is metonymy: a conspicuous part represents the whole (cf frame theory by Minsky, story grammar by Rumelhart).

Mental Images

Knowledge may also be represented by inner images. The image schema consists of schematized, simplified images. These images make comprehensible and literally imaginable physical objects or logical relationships that are otherwise difficult to conceptualize. The organizing principle behind them is the metaphor: through analogy, information from the physical world is introduced into the non-physical world.

Examples:
Rage can be imagined as a hot liquid in a container, evaporation as rising molecules springing out of the water like popcorn, and electricity as a crowd of people (in front of the gate of a stadium at a sports event) or as a flood forcing itself through some obstacle.

Trajectories for Future Research

A broad overview of current topics and trends in the field is given in *Companion to Cognitive Anthropology* (Kronenfeld et al.,

2011). Some of these trajectories, however, deserve explicit mentioning because they are closely related to the theoretical notions discussed in this article.

The theoretical developments of the structure of knowledge representations, as summarized in sections Prototype; Script, Schema, and Cultural Model; and Mental Images, allow us to understand better the great deal of order that appears to exist in the natural world we experience – much of which is there only because we put it there. They allow us to conceptualize cultural content as a number of general-purpose models, which are instantiated in line with the special purposes in daily life. They explain how the enormous amount of knowledge can be mastered, namely by storing only its prototypical parts, which are then actualized on demand and enriched with details. And they allow us to consider how people interpret new experiences because these models not only represent knowledge but also afford inferences to new situations.

Besides these achievements, three fields of research remain productive for theorizing in anthropology and for exchange across disciplinary boundaries.

Knowledge and Knowing

Most of what people know is not directly accessible because it is rather practical in character, incorporated in encounters, and encoded in the routines of social life (Giddens, 1984). When asked to describe how to ride a bicycle, one realizes that the traditional conception of *knowledge* – that which is known, as an abstract pool of information, declarative and verbalized – needs to be modified. *Knowing* how to do something in practice – being implicit and hidden, accessed primarily though performance – is at least as interesting for anthropological inquiry. For this reason, Borofsky (1994) suggested that we reorient our analysis and reverse the process, starting with knowing and then exploring how knowledge is constituted from it.

These insights have led to new trends, in both anthropology and the cognitive sciences more broadly. One has given rise to the parallel distributed processing models, also called connectionistic models or neuronal networks, developed by Rumelhart and McClelland (1986) and adopted for anthropological theorizing by Strauss and Quinn (1997). What intrigues anthropologists about this approach is that it allows modeling the functioning of patterns not "as loading in a set of instructions, but as gradually building up associative links among repeated or salient aspects of experience" (Strauss and Quinn, 1997, p. 50) or, in other words, as being dependent on our everyday lives, including the cultural phenomena.

A related trend in this vein is an increasing interest by anthropologists in how ways of knowing are embodied (Ingold, 2011), and this converges with the increasing interest by cognitive scientists in how cognitive concepts and processes may also be embodied (Barsalou, 2008). The underlying assumption is that information is memorized and represented in the same way in which it was originally encoded by the sensomotoric system, which provides a direct link to the environment (Gibbs, 2005). This assumption, backed up by neuroscientific findings, goes beyond previous ideas, in that cognition is no longer regarded as a collection of abstract representations stored solely in the brain, but as grounded in perception and behavior, and thus eventually embodied (De Vega et al., 2008; Fuchs, 2008; Klatzky et al., 2008).

Language and Cognition

The question of whether language shapes thought was dismissed in the wake of the cognitive revolution, but re-emerged some 20 years ago (Gumperz and Levinson, 1996) as one of the most promising fields for exchange between anthropology, linguistics, and other cognitive sciences. A prime example in this field is the research on spatial orientation and communication. While certainly belonging to the basic cognitive and linguistic equipment of all people around the world, systems of orientation differ across languages and cultures, and even apparently basic spatial categories such as 'left' and 'right' cannot be taken for granted (Levinson, 2003). Several languages that lack these terms use a geocentric system instead – based, for example, on the cardinal points – and not only in navigation, but also in everyday life (the glass is then not placed 'to the left of' the plate but, say, 'to the east'). These linguistic and cultural differences influence the perception of spatial relationships as well as their storage in memory (Wassmann and Dasen, 1998; Majid et al., 2004). Here, cognitive anthropology contradicts a school of thought for which the worldwide diversity of languages is only a cultural phenomenon and hence one of surface, and this gives rise to an ongoing dispute (Li and Gleitman, 2002 vs Levinson et al., 2002; and see Haun et al., 2011; Li et al., 2011). A medium position in this dispute is that although language may be taken as an important part of cultural transmission, it does not determine cognition (in the sense of a strong linguistic relativism) but is only one aspect of the overall ecocultural framework (Dasen and Mishra, 2010, 2013).

Universal or Cultural Variable?

Cognitive processes, such as perception, categorization, and memorization, are the foundations of knowledge acquisition and of the structure of its content. In principle, they are thought to be universal, albeit applied to particular contexts in different ways. Segall et al. (1990, p. 184), for instance, report cross-cultural differences in habitual strategies for classifying and problem solving, in cognitive styles, and in progression rates during development; but they consider them to be differences in performance rather than in competence.

Exploring under which circumstances and to what extent these differences occur remains one of the core tasks for the future (Henrich et al., 2010a,b) and one of those fields in cognitive science to which anthropology can – and should – contribute with its unique expertise (Bender and Beller, 2011b; Bender et al., 2012). Domains for which this is already under way include space and time (Bender et al., 2010; Haun et al., 2011; Núñez et al., 2012a); causality (Bender and Beller, 2011a); numerical cognition (Beller and Bender, 2008; Bender and Beller, 2012; Núñez et al., 2012b; Pica et al., 2004); biological and ecological reasoning (Atran and Medin, 2008; Bang et al., 2007; Medin and Atran, 1999; Medin et al., 2006); and religious concepts such as the notion

of an afterlife (Astuti et al., 2004; Whitehouse and McCauley, 2005; Astuti and Harris, 2008). This also includes an upsurge of interest in more complex topics that had been excluded from the cognitive sciences in their early phase (Gardner, 1985), and some of these issues are brought to the fore by anthropologists: the relation between emotion and cognition (Röttger-Rössler, 2004); the development of moral reasoning or theory of mind in specific cultural settings (Wassmann et al., 2013); and the phylogenetic evolution and ontogenetic development of complex phenomena such as religion and social cooperation (Henrich et al., 2010a,b).

More and more cognitive scientists, meanwhile, acknowledge the need to scrutinize universality rather than simply to postulate it. To further sensitize cognitive scientists to this endeavor and to make progress in this direction is a task for which no discipline is better suited than anthropology.

Bibliography

Atran, S., Medin, D., 2008. The Native Mind and the Cultural Construction of Nature. MIT Press, Cambridge, MA.

Astuti, R., Harris, P.L., 2008. Understanding mortality and the life of the ancestors in rural Madagascar. Cognitive Science 32, 713–740.

Astuti, R., Solomon, G.E.A., Carey, S., 2004. Constraints on Conceptual Development: A Case Study of the Acquisition of Folk Biological and Folk Sociological Knowledge in Madagascar. Monographs of the Society for Research in Child Development 69 (3), 1–135. Blackwell, Boston.

Bang, M., Medin, D.L., Atran, S., 2007. Cultural mosaics and mental models of nature. Proceedings of the National Academy of Sciences 104, 13868–13874.

Barsalou, L.W., 2008. Grounded cognition. Annual Review of Psychology 59, 617–645.

Beller, S., Bender, A., 2008. The limits of counting: numerical cognition between evolution and culture. Science 319, 213–215.

Bender, A., Beller, S., 2011a. Causal asymmetry across cultures: assigning causal roles in symmetric physical settings. Frontiers in Psychology 2 (231), 1–10.

Bender, A., Beller, S., 2011b. The cultural constitution of cognition: taking the anthropological perspective. Frontiers in Psychology 2 (67), 1–6.

Bender, A., Beller, S., 2012. Nature and culture of finger counting: diversity and representational effects of an embodied cognitive tool. Cognition 124, 156–182.

Bender, A., Beller, S., Bennardo, G., 2010. Temporal frames of reference: conceptual analysis and empirical evidence from German, English, Mandarin Chinese, and Tongan. Journal of Cognition and Culture 10, 283–307.

Bender, A., Beller, S., Medin, D.L. (Eds.), 2012. Does cognitive science need anthropology? Topics in Cognitive Science 4 (3) (special issue).

Bender, A., Beller, S., 2013. Die Welt des Denkens. Kognitive Einheit, kulturelle Vielfalt, Bern, Huber.

Berry, J., Poortinga, Y., Segall, M., Dasen, P., 2002. Cross-cultural Psychology. Cambridge University Press, Cambridge, UK.

Bloch, M., 1991. Language, anthropology and cognitive science. Man 26, 183–198.

Boden, M., 2006. Mind as Machine. A History of Cognitive Science. Oxford University Press, Oxford.

Borofsky, R. (Ed.), 1994. On the knowledge and knowing of cultural activities. Assessing Cultural Anthropology. McGraw-Hill, New York, pp. 331–346.

D'Andrade, R.G., 1989. Cultural cognition. In: Posner, M.I. (Ed.), Foundations of Cognitive Science. MIT Press, Cambridge, MA, pp. 795–830.

D'Andrade, R.G., 1995. The Development of Cognitive Anthropology. Cambridge University Press, Cambridge.

Dasen, P.R., Mishra, R., 2013. Cultural differences in cognitive styles. In: Kar, B.R. (Ed.), Cognition and Brain Development. APA, Washington, DC, pp. 231–249.

Dasen, P.R., Mishra, R., 2010. Development of Geocentric Spatial Language and Cognition. Cambridge University Press, Cambridge.

De Vega, M., Glenberg, A., Graesser, A.C., 2008. Symbols and Embodiment. Oxford University Press, Oxford.

Dougherty, J. (Ed.), 1985. Directions in Cognitive Anthropology. University of Illinois Press, Urbana, IL.

Fuchs, T., 2008. Das Gehirn – ein Beziehungsorgan. Eine phänomenologisch-ökologische Konzeption. Kohlhammer, Stuttgart.

Gardner, H., 1985. The Mind's Mew Science. A History of the Cognitive Revolution. Basic Books, New York.

Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.

Gibbs, R., 2005. Embodiment and Cognitive Sciences. Cambridge University Press, Cambridge.

Giddens, A., 1984. The Constitution of Society. University of California Press, Berkeley and Los Angeles.

Goodenough, W., 1957. Cultural anthropology and linguistics. In: Garvin, P.L. (Ed.), Reports of the Seventh Annual Round Table Meeting on Linguistics and Language Studies, Monograph Series on Language and Linguistics No. 9. Georgetown University, Washington, DC, pp. 167–173.

Gumperz, J., Levinson, S. (Eds.), 1996. Rethinking Linguistic Relativity. Cambridge University Press, Cambridge.

Haun, D.B.M., Rapold, C.J., Janzen, G., et al., 2011. Plasticity of human spatial cognition: spatial language and cognition covary across cultures. Cognition 119, 70–80.

Henrich, J., et al., 2010a. Markets, religion, community size, and the evolution of fairness and punishment. Science 327, 1480–1484.

Henrich, J., Heine, St. J., Norenzayan, A., 2010b. The weirdest people in the world? Behavioral and Brain Sciences 33, 61–135.

Hutchins, E., 1995. Cognition in the Wild. MIT Press, Cambridge, MA.

Ingold, T., 2011. Being Alive: Essays on Movement, Knowledge and Description. Routledge, London.

Keck, V. (Ed.), 1998. Common Worlds and Single Lives. Berg, Oxford.

Keesing, R., 1987. Models, 'folk' and 'cultural' paradigms regained? In: Holland, D., Quinn, N. (Eds.), Cultural Models in Language and Thought. Cambridge University Press, Cambridge, pp. 368–393.

Keesing, R., 1972. Paradigm lost. The new ethnography and the new linguistics. Southwestern Journal of Anthropology 28 (4), 299–332.

Klatzky, R.L., MacWhinney, B., Behrman, M. (Eds.), 2008. Embodiment, Ego-Space and Action. Psychology Press, Hove.

Kronenfeld, D., et al. (Eds.), 2011. A Companion to Cognitive Anthropology. Wiley-Blackwell, Chichester.

Lave, J, 1988. Cognition in Practice Mind, Mathematics and Culture in Everyday Life. Cambridge University Press, Cambridge.

Levinson, S.C., 2003. Space in Language and Cognition. Explorations in Cognitive Diversity. Cambridge University Press, Cambridge.

Levinson, S.C., Kita, S., Haun, D.B.M., et al., 2002. Returning the tables: language affects spatial reasoning. Cognition 84, 155–188.

Li, P., Abarbanell, L., Papafragou, A., et al., 2011. Spatial reasoning in Tenejapan Mayans. Cognition 120, 33–53.

Li, P., Gleitman, L., 2002. Turning the tables: language and spatial reasoning. Cognition 83, 265–294.

Majid, A., Bowerman, M., Kita, S., et al., 2004. Can language restructure cognition? the case for space. Trends in Cognitive Sciences 8 (3), 108–111.

Medin, D.L., Atran, S. (Eds.), 1999. Folkbiology. MIT Press, Cambridge, MA.

Medin, D.L., Ross, N.O., Cox, D., 2006. Culture and Resource Conflict: Why Meanings Matter. Russell Sage Foundation, New York.

Mishra, R.C., Dasen, P.R., 2007. The methodological interface of psychology and anthropology. In: Wassmann, J., Stockhaus, K. (Eds.), Experiencing New Worlds. Berghahn, Oxford, pp. 21–35.

Núñez, R., Cooperrider, K., Doan, D., Wassmann, J., 2012a. Contours of time.Topographic construals of past, present, and future in the Yupno Valley of Papua New Guinea. Cognition 124, 25–35.

Núñez, R., Cooperrider, K., Wassmann, J., 2012b. Number concepts without number lines in an indigenous group of Papua New Guinea. PLoS ONE.

Pica, P., Lemer, C., Izard, V., Dehaene, S., 2004. Exact and approximate arithmetic in an Amazonian indigene group. Science 306, 499–503.

Quinn, N., Holland, D., 1987. Culture and cognition. In: Holland, D., Quinn, N. (Eds.), Cultural Models in Language and Thought. Cambridge University Press, Cambridge, pp. 3–40.

Rogoff, B., Lave, J. (Eds.), 1984. Everyday Cognition: Its Development in Social Context. Harvard University Press, Cambridge.

Röttger-Rössler, B., 2004. Die kulturelle Modellierung des Gefühls. LIT, Münster.

Rumelhart, D.E., McClelland, J.L. (Eds.), 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge, MA.

Schank, E., Abelson, R., 1977. Scripts, Plans, Goals and Understanding. Lawrence Erlbaum, Hillsdale, NJ.

Segall, M., Dasen, P., Berry, J., Poortinga, Y., 1990. Human Behavior in Global Perspective. Pergamon Press, New York.

Sternberg, E.J. (Ed.), 1999. The Nature of Cognition. MIT Press, Cambridge, MA.

Strauss, C., Quinn, N., 1997. A Cognitive Theory of Cultural Meaning. Cambridge University Press, Cambridge.

Tyler, S. (Ed.), 1969. Cognitive Anthropology. Holt, Rinehart, and Winston, New York.
Wassmann, J., 1994. Worlds in mind: the experience of an outside world in a community of the Finisterre Range of Papua New Guinea. Oceania 64 (2), 117–145.
Wassmann, J., Dasen, P.R., 1998. Balinese spatial orientation: some empirical evidence of moderate linguistic relativity. Journal of the Royal Anthropological Institute 4 (4), 689–711.
Wassmann (n.d), J. The ideal of the slightly bent person: an interdisciplinary approach in collaboration with musicologist R. Ammann, cross-cultural psychologist P.R. Dasen, anthropologist V. Keck, cognitive scientist R. Núñez, and linguist J. Slotta.
Wassmann, J., Kluge, C., Albrecht, D., 2011. The cognitive context of cognitive anthropology. In: Kronenfeld, D., et al. (Eds.), A Companion to Cognitive Anthropology. Wiley-Blackwell, Oxford, pp. 47–60.
Wassmann, J., Träuble, B., Funke, J. (Eds.), 2013. Theory of Mind: Reasoning across Cultures. Berghahn, New York and Oxford.
Whitehouse, H., McCauley, R.N. (Eds.), 2005. Mind and Religion. Altamira, Walnut Creek, CA.

Cognitive Archaeology

Marc A Abramiuk, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D.S. Whitley, volume 3, pp. 2085–2089, © 2001, Elsevier Ltd.

Abstract

Cognitive archaeology, also called the archaeology of mind, is a burgeoning branch of archaeology that has as its focus of investigation the mind in the past. Cognitive archaeology makes use of several forms of evidence to infer past peoples' cognitive capabilities and mind frames (e.g., concepts and percepts), the most obvious being the material remains constituting the archaeological record. From these remains and their distributions, cognitive archaeologists attempt to make informed claims for understanding how people in the past thought and what they thought about. Cognitive archaeological inquiry is cross-disciplinary, such that in addition to relying on inferences made from archaeological data, cognitive archaeologists also depend on evidence that other disciplines, such as physical anthropology, cognitive anthropology, cognitive science, neuropsychology, linguistics, and philosophy, can bring to bear on the question of mind frames and mind functioning in the past. Despite the cross-disciplinary nature of cognitive archaeology, cognitive archaeologists are first and foremost archaeologists and, as such, maintain a predominant concern with what can be learned from material culture. For this reason, cognitive archaeologists have become mutually invested in the study of how humans interact with material culture, and the different ways the mind might be constituted through this engagement. Over the past decade, cognitive archaeologists have begun to form their own distinctive views regarding how the mind operates in relation to the material world, and this has contributed significantly to our general knowledge of the nature of the mind.

Definitions

Although the term 'cognitive archaeology' was used to refer to mind-related archaeological research as early as the 1970s (Flannery and Marcus, 1976: p. 374), it was not until the early 1980s that an agenda was articulated consisting of several questions with which cognitive archaeology was meant to concern itself (Renfrew, 1982: p. 2). One of these questions pertained to how archaeologists infer 'intelligent behavior' from material remains (Renfrew, 1982: p. 2), a concern that would formally lead to defining cognitive archaeology as "the study of past ways of thought as inferred from material remains" (Renfrew, 1994: p. 3). This definition was spawned to account for archaeologists working in the area of cognitive evolution (e.g., Wynn, 1979) as well as archaeologists who recognized the need to account for culturally specific belief systems when attempting to explain the archaeological record (e.g., Flannery and Marcus, 1976). The purpose of cognitive archaeology as envisioned by these archaeologists, it should be noted, was distinctive, albeit related to one of the main objectives that postprocessual archaeologists were pursuing contemporaneously, namely to uncover the 'meanings' or mind frames that were theorized to constitute the archaeological record (Hodder, 1982a,b). Where the approaches differed was on an epistemological level and, in particular, on the kinds of questions concerning the mind that were proposed to be illuminated through archaeological research. To highlight the difference, Colin Renfrew, for example, used the term cognitive-processual archaeology often in place of the term cognitive archaeology to delineate how he envisioned conducting mind-related archaeological research (Renfrew, 1994). In emphasizing the processual component, Renfrew was stating that cognitive archaeology was a scientific endeavor rather than a series of contextual and empathetic efforts, like those being explored by postprocessualists.

Today, cognitive archaeology may mean slightly different things to a cognitive archaeologist depending on his or her specialty. For example, it may mean the study of cognitive evolution, ancient belief systems, or experience and perception in the past. A recently published definition of cognitive archaeology that attempts to include these various foci, sees cognitive archaeology as "a subdiscipline of archaeology comprised of several practical approaches to studying the mind in the past" (Abramiuk, 2012: p. 15). This definition, in its generality, acknowledges the variety of mind frames and cognitive capabilities archaeologists examine and, consequently, the need for a number of investigative approaches; it also recognizes the need to consider other forms of inferential evidence (e.g., genetic and neuroscientific) to enter into the inquiry process in addition to archaeological data.

Philosophical Commitments

In the history of archaeology, inquisitiveness about past people's thoughts goes back to the beginnings of antiquarianism, but legitimate methods for pursuing this interest do not become apparent until the mid-twentieth century with the first speculative but grounded forays into understanding Upper Palaeolithic cave art (e.g., Breuil, 1952). With the exception of a few scholars who contended – or implied – that past human mind frames could be uncovered in the features and artifacts they studied (e.g., Leroi-Gourhan, 1964; Spaulding, 1960: p. 76), the 1960s and 1970s were dominated by the 'New Archeology' movement. The 'New Archeology,' more widely known later as functional-processual archaeology (processual archaeology), focused on *explaining* the archaeological record and utilized a logical empiricist methodology adopted from Carl Hempel (1962) to accomplish this objective. Processualism's commitment

to a hard-lined empiricist – if not positivistic – stance, however, would disallow inquiries about the mind in the past and the mind's role in archaeological explanation, cynically referring to such attempts as 'paleopsychology' (Binford, 1965: p. 204). As a result, it would not be until the mid-1970s and 1980s that mind-related studies in archaeology – beginning to be referred to as cognitive archaeology at this time – would begin in earnest.

The emergence of cognitive archaeology and its associated philosophical commitments can be fundamentally retraced to the same epistemological challenge processual archaeology faced early in its development, as it sought not merely to describe culture change evincible in the archaeological record, but to explain it. This challenge consequently was imparted to processual archaeology through its adoption of the covering law model found in logical empiricism. In processual archaeology, the covering law model (also known as the deductive-nomological model) was seen as a way of explaining patterns in the archaeological record deductively – that is, by virtue of expecting such patterns in light of general laws governing human behavior and certain observable initial conditions. The challenge in using the covering law model, however, lies in the fact that the model does not require that the observable initial conditions be connected in a *causally relevant* or substantive manner to the patterns in the archaeological record, and as such does not preclude alternative explanations for the archaeological patterns (Salmon, 2001: p. 235); this, of course, makes processual archaeological explanations difficult to logically falsify. Indeed, it implies that conditions that are unobservable or simply not considered by the processual archaeologist (e.g., the beliefs people had in the past) may also explain the archaeological patterns in question and, in fact, may explain the patterns more adequately. The realization that the explanatory model used in processual archaeology lacked sufficient scope effectively opened the door to mind-related archaeological research and, in turn, cognitive archaeology.

The rise of cognitive archaeology, however, would not have been secured were it not also for a divergence from the semantic position taken by processual archaeologists as it pertained to unobservable conditions. Processual archaeologists, in following the logical empiricists, treated unobservable conditions (e.g., cognitive states and beliefs) either in a nonliteral way (Binford, 1973: p. 227) or as meaningless because archaeologists were believed to be ill-equipped to delve into such conditions (Binford, 1965: pp. 204–205). As a result, processual archaeologists were inclined to focus on behavior, explaining cultural continuities and changes in the past as adaptations to external stimuli often taken to be environmental in origin.

In diverging from processual archaeologists, cognitive archaeologists take a semantic stance in which the value of considering unobservable phenomena is not denied. In cognitive archaeology, unobservable phenomena, such as cognitive capabilities and cultural beliefs, are treated explicitly as subjects of study or as factors in explaining observable patterns in the archaeological record. In other words, cognitive archaeologists assert that the mind is something to be taken literally and that it furthermore can be reconstructed or understood through material remains.

Where cognitive archaeologists may semantically differ is in how they interpret the relationship between mind and material, which stems from different ontological commitments. Notable ontologies that have gained favor in cognitive archaeology are those that merge mind and material; and these have, in turn, contributed to the varying degrees in which cognitive archaeologists literally construe the mind as contiguous with the brain, body, and material world. Cognitive archaeologists, therefore, may take slightly different semantic lines when speaking of the mind in its relation to brain, body, and material world that can range from realist to phenomenological (*see* Archaeology, Theory in).

Changes in the Focus of Mind-Related Studies in Archaeology Over Time

Beginning in the 1970s and continuing through the 1980s, as mind-related studies in archaeology were seeping into archaeological research, two trends can be seen to have emerged. The first trend included archaeologists who retained a processually derived systemic view of ancient societies, and who considered the relevance of the mind, but mainly as a factor to help explain culture change evidenced in the archaeological record. The second trend included archaeologists who were significantly more radical. These archaeologists adopted linguistic and psychological models of the mind, which encouraged treating the mind as the subject of study in archaeological investigation.

The first trend in mind-related archaeological research remained committed to the goal of providing explanations for past phenomena, just as it always had been in processual archaeology. The difference between the two approaches, however, lay in the nature of the *explanans* – a term derived from Hempel and Oppenheim (1948) and adopted by some archaeologists to describe the factors used in explaining a particular archaeological phenomenon. In mind-related archaeological studies, an ideational-themed *explanans* was simply substituted for an environmental- or technological-themed *explanans*, such as those popularly employed in processual archaeology. The research objective continued to emphasize explaining cultural continuity and change in the archaeological record; what changed were the factors used to do the explaining. Instead of resorting to material factors, archaeologists began to consider ideational alternatives such as past beliefs. These ideational factors were typically revealed through a process of ethnographic comparison and then assessed against the archaeological evidence. Flannery and Marcus (1976) exemplify this trend in mind-related archaeological research by explaining Zapotec agricultural practices, village life, and burial patterns in Formative Oaxaca by factoring in proposed religious beliefs recorded in ethnohistoric sources. In another example of this trend, Hall (1976) explains how the belief in the ability to control for the movement of spirits – a belief that is widely described in ethnographic accounts of certain contemporary Native American peoples – might have been the impetus for the construction of ditches and earthworks during the Eastern Woodland period.

The second trend in mind-related archaeological studies made extensive use of linguistic and psychological theories to understand what and how people thought in the past. It involved reconstructing mind frames and thought processes evidenced in the archaeological remains; however, rather than solely being directed by ethnographic comparison, it was guided by theories concerning the cross-cultural nature and development of the human mind. For this reason, the structuralist notion of a universal, underlying set of operations for relating concepts in the mind became a common theoretical platform for archaeologists seeking to piece together thoughts from the past. McGhee (1977), for example, examined the way in which the Thule of Arctic Canada selected tool-making raw materials by proposing that their selections were motivated by binary oppositional relations, which he proceeded to reconstruct. In this second research trend, inferences regarding what and how people were thinking often became the ends of research, and the archaeological remains, the means for inferring the mind frames and thought processes. Some of the earliest research regarding reconstructing thought processes was conducted by Thomas Wynn. Inspired by Piagetian psychological theory, Wynn (1979) was able to gain valuable insight into the cognitive capabilities of Acheulean toolmakers through an extensive analysis of Acheulean tools.

Both trends in mind-related archaeological research would persist into the next decade and beyond; however, they would largely be caught up in a contemporaneous movement pervading archaeological theory – the objective of which was to uncover social life in the past. Political power, social relations and divisions, among other social phenomena, became aspects of social life that archaeologists attempted to uncover from the archaeological record, and this facilitated mind-related archaeological studies on several levels. For instance, the related notions of ideology and consciousness – notions critical for understanding power and social divisions – naturally implied a need to understand the mind frames of the people being studied (Hodder, 1982a), as well as those of the archaeologists (see Leone, 1982: p. 751). Delving into the social aspects of ancient life, specifically through a Marxist lens, was nothing new in archaeology as it had been proponed by Gordon Childe and others decades before; however, social theory's marriage to structuralism to uncover the social mind through artifact assemblages was distinctive. Hodder (1982b), for instance, showed through successive ceramic sequences that aspects of a Dutch Neolithic ideology, serving to legitimize social relations and contradictions, could be identified through a structural analysis of pot decorations. Such studies that incorporated social theory as well as explored mind frames soon came under the umbrella of postprocessual archaeology.

Similarly motivated but alternative attempts at gleaning insight into the social circumstances of people in the past were proposed by Colin Renfrew. In contrast to advocating theory-driven understandings of social life, emphasis was placed on what empirically driven inferences – largely through spatial patterning – could succinctly say about sociality in ancient times. For Renfrew (1982), the acknowledgment that social life in the past could be teased from the archaeological record was encouraging and helped pave the way for accepting that cognitive aspects of people in the past also could be investigated. For the first time an explicit agenda for cognitive archaeology was put in place in which systems of thought (e.g., measurement, planning, and religion) could and would be examined only insofar as these systems were relevant for the functioning of ancient society and clearly expressed in material remains (Renfrew, 1982: p. 11). By the end of the 1980s, it was evident that postprocessualists were concerned with instances of thought and processualists were concerned with systems of thought. In the United Kingdom, these different ways of looking at cognition would become gradually more divergent in the following decade. In other parts of the world, such as South Africa, where significant research on cognitive evolution and the origins of art was being conducted (e.g., Lewis-Williams and Dowson, 1988), cognitive archaeologists were less affected by the polarization.

In the 1990s, a statement was formally issued that would distinguish two camps with distinct aims and ways for studying the mind in the past (Renfrew, 1994). The postprocessual camp would have as one of its purposes the uncovering of meanings, namely *what* people thought in the past. Its plan for elucidating these meanings or mind frames would rely on structural (Hodder, 1982a,b), historical (Hodder, 1987), phenomenological (Thomas, 1996; Tilley, 1994), and other interpretive approaches. The cognitive-processual camp, on the other hand, would strive to understand *how* people thought in the past inclusive of their cognitive capabilities and symbolic systems (Renfrew, 1994). One area of research that lay squarely within the mission statement of cognitive-processualists was cognitive evolution, a topic that saw a surge of interest and gained serious attention from specialists in the cognitive sciences (e.g., Donald, 1991). Cognitive-processual archaeology would be conducted from a scientific realist position, and rely on a method of logical inference by which systems of thought would be guardedly arrived at through supporting evidence derived from cultural remains. Eventually the term 'cognitive archaeology' would come to be associated with the cognitive-processual camp's aim and approach, while the study of the mind in the postprocessual school would increasingly come under the sway of archaeological phenomenology (Gosden, 1994; Tilley, 1994; Thomas, 1996) and material culture theory (Miller, 1987; Tilley, 1999). Although, the cognitive-processual and postprocessual schools of thought continued along separate paths through the dawn of the new millennium, certain convergent viewpoints became noticeable. The similarity of viewpoints can largely be attributed to a major undercurrent in archaeology that stressed materiality.

Just as the sociality movement helped give impetus to studying the mind in archaeology in the 1980s, so too did the materiality movement help propel cognitive archaeology in the 2000s. The kernels of the materiality movement could be traced back to Hodder's (1982a,b) work on the active role of material culture in society; however, the role of material culture in social interactions and its relation to the mind was fostered extensively in the late 1980s and 1990s in the postprocessual work of Daniel Miller, Christopher Gosden, and Christopher Tilley. By the early 2000s, the notion of materiality was ripe for study by cognitive archaeologists as well as by the

postprocessual archaeologists who helped bring materiality to the forefront of archaeological investigation. Whereas postprocessual archaeologists tended to focus on the meaning of material culture in and for a society, cognitive archaeologists preferred to address the broader, generalizable ramifications of material culture's role in the functioning of the mind.

In cognitive archaeology, one of the fruits of labor that resulted from exploring materiality was the formulation of material engagement theory (MET). First propounded by Colin Renfrew (2001, 2007) and elaborated upon by Lambros Malafouris (2004, 2013), MET goes beyond the statement that advocates of materiality make regarding the primacy (or at least equivalency) of material culture in its relationship with the mind. MET makes the claim that the mind and the material that comprise us, our surroundings, and our actions within these surroundings are ontologically indivisible; as such, it views the mind as embodied, extended, and enacted. MET does not make it a priority to struggle to uncover the meanings or instances of thought *behind* material culture, particularly, since thought is seen as inseparable from material culture. Rather, the aim of MET is to understand in a holistic manner how the human mind works through material things and action and, in so doing, aids in its development.

Current Thinking about the Field

Historically, archaeologists have tended to gravitate toward one approach or another for understanding the mind in the past – often out of theoretical inclination and without due attention to the approach's substantiation. Recent research, however, has placed six such approaches on more solid ground by detailing their empirical and logical foundations (Abramiuk, 2012). This research finds that these approaches can be utilized productively depending on the precise context and objective of the cognitive archaeological inquiry. These six approaches, which can be used alone or in certain cases jointly, are briefly described below.

The General Comparative Approach

Early attempts at utilizing what could be regarded as a general comparative approach in mind-related archaeological studies date to the dawn of the twentieth century (e.g., Reinach, 1903). However, it has only been in the past several decades, particularly through the work of anthropologists and archaeologists at the University of Witwatersrand, that the general comparative approach has been developed for the purposes of cognitive archaeological research (e.g., Lewis-Williams and Dowson, 1988). Grounded in cognitive scientific research and applied strategically to illuminating specific mind frames, the general comparative approach has proven to be constructive for understanding the motivations for and significance of early parietal art in several parts of the world, from the Upper Palaeolithic cave art of the Franco-Cantabrian region (Lewis-Williams, 2002) to the late prehistoric rock art in far western North America (Whitley, 1992).

The objective of the general comparative approach, as it is used in cognitive archaeology, is to ascertain the mind frames of people in the past by the way of ethnographic comparison. This is accomplished by guardedly projecting beliefs and perceptions of a well-documented group of people onto a target group – the people in the past whose material remains are the subject of inquiry by the archaeologist. The procedure first entails identifying a relevant societal analog to the early society being investigated by the archaeologist, and then drawing inferences from this analog. General comparative inference is supported by cognitive scientific evidence for the universality and, hence, translatability across cultures of certain properties of the human mind in similarly situated social contexts. For this approach to work, however, it is of course paramount to demonstrate that the social environments of the two peoples being compared share conditions that have understood cognitive implications that can aid in the reconstruction of mind frames.

The Direct Historical Approach

The direct historical approach has been used as a way of reconstructing ancient practices from modern ones since at least the early twentieth century. However, its formal recognition as a means of inferring ancient mind frames can be attributed to work of Marcus and Flannery (1994), who utilize the approach for the sole purpose of gleaning insight into ancient Zapotec religious beliefs during the Formative period.

The direct historical approach is related to the general comparative approach in that it seeks to reconstruct mind frames through ethnographic comparison with a well-documented societal analog. However, instead of selecting the analog to be an arbitrary group that shares social organizational attributes to the group in the past being investigated, the direct historical approach bases its inferences on an analog that is culturally – if not, genealogically – related to the target group. The direct historical approach gains insight into the concepts (e.g., religious beliefs) of past peoples then by retracing the concepts that would have been passed down directly – generationally – to the cultural analog. Cognitive anthropological research confirms the saliency of concepts reproduced intraculturally, and therefore supports the long-term continuity or preservation of culturally bounded concepts. Accordingly, culturally bounded concepts can be envisaged as notions from the past that have been passed down through time while resisting significant alterations to their original form.

The Associative Approach

Although, the associative approach has been used in archaeology since the early nineteenth century to relatively date cultural remains and elucidate ancient practices, the implication that it could be utilized to infer mind frames from archaeological assemblages is more recent (e.g., Spaulding, 1960: p. 76). The associative approach is extremely useful in cases where all that is known about the people in the past being studied is that which can be gleaned from the archaeological data. In this way, the associative approach differs substantially from the ethnographic comparative approaches discussed above.

The associative approach takes a drastically different track than the general comparative approach in particular. Rather

than imparting mind frames – often in the form of intact belief systems – from one society onto another, attempts are made to reconstruct concepts through an understanding of how concepts universally relate to one another in the mind.

The associative approach commences by observing and analyzing the cultural remains left behind. Through the contextual analysis of the remains, the associative approach proceeds to identify clearly evincible concepts associated with the remains and the categorical relations between them so that larger semantic conceptual constructs known as categories can be reconstructed from the past. Whereas semantic concepts, such as categories and the concepts, that make them up are often culturally specific, the fundamental manner in which categories form is cross-cultural, and it is this cognitive scientifically substantiated insight that enables the archaeologist to reconstruct them directly from the archaeological record. Kim (1994) implicitly exemplifies this procedure as he principally utilizes an associative approach to argue that pigs served as prestige items in the minds of certain Neolithic peoples in China.

The Structural Archaeological Approach

Advocated as early as the mid-twentieth century by Leroi-Gourhan (1964), the structural archaeological approach, like the general comparative approach, was adopted to uncover the meaning behind Upper Palaeolithic cave art. It maintains many of the same features as the associative approach, particularly the capacity to infer mind frames directly from the archaeological record, and it remains a useful approach where oppositionally related concepts can be detected.

The structural archaeological approach is similar to the associative approach in that it seeks to reconstruct concepts from archaeological data; however, the structural archaeological approach's distinct emphasis is on reconstructing *composites* of concepts known as opposite analogies. The procedure follows a similar sequence of steps as the associative approach, beginning with the observation of cultural remains and a subsequent contextual analysis for the concepts associated with the remains and their relations. What is different in the structural archaeological approach, however, is that oppositionally related concepts – rather than categorically related concepts – are sought among archaeological data for analytical reconstruction. An example of this can be seen in McGhee's (1977) study of the Thule in which he identifies and associates concepts attached to male and female cultural remains. He then uses the cross-culturally recognized oppositional concepts 'male' and 'female' to form the backbone for piecing together all other concepts associated with the cultural remains. Whereas the meaning of these latter concepts may be culturally specific and therefore tentatively unknown to the archaeologist, how these concepts relate has a known cross-cultural aspect since they are connected through the male–female oppositional relation. Cognitive scientific research supports the ability to analytically reassemble oppositionally related concepts (i.e., opposite analogies), since core stages in the analogical reasoning process responsible for the formation of opposite analogies – and the steps with which cognitive archaeologists are concerned as they pertain to reconstructing opposite analogies – are cross-cultural.

The Conditional Approach

Although the term 'conditional approach' was never actually used in mind-related archaeological research, it was Colin Renfrew (1982, 1994) who first laid out an agenda advocating the use of deductive reasoning to infer cognitive capabilities as logical consequences of the presence or absence of certain cultural remains. The term 'conditional approach' is used to refer to this reasoning strategy and, in particular, how arguments are often expressed using this approach – namely in the form of conditional statements (Abramiuk, 2012: pp. 141–152). It continues to be one of the most intuitive as well as adaptable approaches available to the cognitive archaeologist. The conditional approach is mainly concerned with ascertaining *how* people thought in the past, rather than what they thought about. As such, it stands in stark contrast to the other approaches discussed.

Rather than being used to reconstruct mind frames, the conditional approach is commonly utilized to postulate the existence of certain cognitive capabilities, whether they are general memory systems or more particular systems of thought such as measurement and planning. The conditional approach undertakes this endeavor by inferring the cognitive capabilities that would have been needed to produce the behavior responsible for the cultural products left behind. More specifically, the conditional approach sets up logical arguments in a conditional manner, such that some behavior or remain is *contingent* on a certain cognitive capability. Establishing an argument this way, allows the archaeologist to present conclusions about mental functioning in the past in a manner that can be analyzed objectively and systematically verified or rebutted. Because the conditional approach is logic-based, rather than empiric-based, and because it is concerned with thought processes rather than instances of thought, it constitutes a substrate-independent approach. As such, the conditional approach can be used to infer cognitive capabilities of virtually any species, and this makes it particularly valuable in cognitive evolutionary studies (e.g., Wynn, 1979; Mithen, 1996).

The Materiality Approach

Despite having earlier roots (e.g., Barrett, 1994; Miller, 1987; Tilley, 1994, 1999), the materiality approach formally entered cognitive archaeology through the work presented in an edited volume entitled *Rethinking Materiality: The Engagement of the Mind with the Material World* (DeMarrais et al., 2004). The term, 'materiality approach' refers to a spectrum of cognitive archaeological inquiries, each comprised of nuanced differences, but all emphasizing the material world and its primacy over or equivalency with the mind. The materiality approach constitutes one of the more recent analytical attempts at understanding how the human mind functioned in the past (as well as functions in the present). It achieves this understanding by concerning itself with how humans in their natural settings engaged with material things and how these interactions shaped their minds.

Although the materiality approach, like the conditional approach, is concerned with how people thought in the past, it achieves this understanding through the dynamics of perception–action which can, in turn, facilitate the recovery of certain percepts (Abramiuk, 2012: pp. 105–112). More specifically, the materiality approach analytically simulates the process by which people in the past 'thought' as they interacted with their surroundings by focusing on the ways in which people universally 'perceive' things through engagement. Provided the surroundings in question have not changed significantly from ancient times, some of people's accompanying percepts of material forms in the past can become relatable to an archaeologist in the present. A similar line of reasoning can be used to derive meanings that would have been afforded by these ancient human–object interactions (e.g., Knappett, 2005: pp. 133–166). Much of the extradisciplinary support for the materiality approach and how it can facilitate the recovery of certain percepts and meanings can be found in the cognitive anthropological and ecological psychological research on direct perception (e.g., Warren, 1984).

A variant of the materiality approach that has come to the forefront of cognitive archaeological research in recent years is MET. MET's calling card is its stance on the unity of mind and material, and it concerns itself less with attempting to elucidate mind frames and more with tracing thought 'through' the object and/or action with that object. More generally, MET is concerned with understanding how the mind works as a totality, inclusive of the body and the material world. Because of MET's advocacy of an embodied, extended view of the mind, MET finds much support in the recent literature on the philosophy of mind (e.g., Clark, 1997) and is gaining some empirical backing in neuroscientific studies (see below).

Emerging Directions in Theory, Practice, and Research

Cognitive archaeology is a rapidly expanding field with many advancements being made in both theory and practice. At present, few cognitive archaeologists would disagree with the statement that despite advancements in the field, the mind in the past – let alone, the mind in the present – remains a complex and incompletely understood phenomenon. Where specialists involved in the development of cognitive archaeological theory and practice differ, is with regard to the lines of inquiry they pursue in contributing to the field. For some cognitive archaeologists, the starting point for such an inquiry is a metaphysical position regarding the nature of the mind, and for others it is an epistemological concern for the utility of a particular cognitive archaeological approach. The archaeologist who begins with a metaphysical stance concentrates on developing an approach for studying the mind in the past based on his/her understanding of the nature of mind. The epistemologically concerned cognitive archaeologist works in reverse by initially focusing on the approach utilized, and then analyzes the validity and scope of the approach to determine its basis. Both ways of contributing to cognitive archaeology are apparent in the literature to date, and both ways add to our overall knowledge of the subject, which consequently helps to advance research. Among the research areas emerging as the result of taking the former (metaphysical) line of inquiry and latter (epistemological) line of inquiry are neuroarchaeology and foundations, respectively.

The development of neuroarchaeology in recent years can partly be seen as a consequence of a recently introduced perspective regarding the nature of the mind, which sees the mind as embodied and extended. What this view has accomplished for cognitive archaeology, and which is articulated in MET, is to link archaeology with neuroscience, thereby bridging the gap between culture and the brain. This marriage between archaeology and neuroscience, referred to as neuroarchaeology, shows much promise in advancing cognitive studies since both archaeology and neuroscience have the ability to examine the relationship between brain and culture from distinct vantage points. The field of archaeology studies human lives over significant time frames through material remains. This makes archaeology a field particularly suited to studying a mind that is inseparable from (material) culture and engaged in specific activities or practices over the long term. Neuroscience, on the other hand, is equipped to study the immediate and short-term effects of engagement with culture on the brain through neuroimaging and other monitoring techniques (Stout and Chaminade, 2007). Provided researchers avoid embracing a crude neurocentric position, the subfield of cognitive archaeology known as neuroarchaeology has the potential to provide keen insight into the study of cognitive evolution and, in particular, the notion of brain plasticity as it relates to culture and the material world (Malafouris, 2013: pp. 168–169, 229–330).

A different emerging research trend is foundational studies in cognitive archaeology. Like many disciplines, cognitive archaeology has undergone an epistemological dilemma that is natural for a field that has been fueled mainly by the need for a variety of approaches to study the mind, often at the expense of establishing solid bases for their utilization. Approaches to studying the mind in the past, since the emergence of mind-related studies in archaeology, often involved assumptions regarding how the mind works without substantial cross-checking of empirical research in related disciplines. Foundational research in cognitive archaeology consequently focuses on the epistemological issues that have emerged in cognitive archaeology. Instead of devising new approaches, foundational research synthesizes and analyzes the approaches that have been utilized throughout the last several decades, determines their limitations, and tries to situate them more firmly on logical and empirical grounds. Abramiuk (2012) has taken one of the first steps toward such a foundational inquiry, and others (e.g., Garofoli and Haidle, 2013) similarly have been quick to realize that epistemological issues need to be considered before any knowledge claims are to be made in cognitive archaeology.

See also: Archaeology, Theory In; Archaeology: Philosophy and Science; Body, Evolution of; Brain, Evolution of; Cognitive Anthropology; Cognitive Neuroscience; Comparative Method in Anthropology; Consciousness, Neural Basis of; Culture,

Cognition and Embodiment; Embodiment Theory; Ethnohistory; Human Cognition, Evolution of; Ideological Constraint: History and Current Status of the Concept; Intelligence, Genetics of: Cognitive Abilities; Learning and Memory, Neural Basis of; Logical Positivism and Logical Empiricism; Materiality and Culture; Mental Imagery: Visual Cognition; Phenomenology: Philosophical Aspects; Prehistoric Art; Relativism, Cognitive; Ritual and Symbolism, Archaeology of; Scientific Explanation; Structuralism.

Bibliography

Abramiuk, M.A., 2012. The Foundations of Cognitive Archaeology. MIT Press, Cambridge.

Barrett, J.C., 1994. Fragments from Antiquity: An Archaeology of Social Life in Britain, 2900–1200 BC. Blackwell, Oxford.

Binford, L., 1965. Archaeological systematics and the study of culture process. American Antiquity 31, 203–210.

Binford, L., 1973. Interassemblage variability – the Mousterian and the "functional" argument. In: Renfrew, A.C. (Ed.), The Explanation of Culture Change: Models in Prehistory. Duckworth, London, pp. 227–254.

Breuil, H., 1952. Four Hundred Centuries of Cave Art. Centre d'Etudes et de Documentation Prehistoriques. Montignac.

Clark, A., 1997. Being There: Putting Brain, Body, and World Together Again. MIT Press, Cambridge.

DeMarrais, E., Gosden, C., Renfrew, A.C. (Eds.), 2004. Rethinking Materiality: The Engagement of Mind with the Material World. McDonald Institute Monographs, Cambridge.

Donald, M., 1991. Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition. Harvard University Press, Cambridge.

Flannery, K., Marcus, J., 1976. Formative Oaxaca and the Zapotec cosmos: the interactions of ritual and human ecology. American Scientist 64 (4), 374–383.

Garofoli, D., Haidle, M.N., 2013. Epistemological problems in cognitive archaeology: an anti-relativistic proposal towards methodological uniformity. Journal of Anthropological Sciences 91, 1–35.

Gosden, C., 1994. Social Being and Time. Blackwell, Oxford.

Hall, R.L., 1976. Ghosts, water barriers, corn, and sacred enclosures in the eastern woodlands. American Antiquity 41, 360–364.

Hempel, C., 1962. Explanation in science and in history. In: Colodny, R.C. (Ed.), Frontiers of Science and Philosophy. University of Pittsburgh Press, Pittsburgh, pp. 7–33.

Hempel, C.G., Oppenheim, P., 1948. Studies in the Logic of Explanation. Philosophy of Science, vol. XV, pp. 135–175.

Hodder, I.R., 1982a. Theoretical archaeology: a reactionary view. In: Hodder, I.R. (Ed.), Symbolic and Structural Archaeology. Cambridge University Press, Cambridge, pp. 1–16.

Hodder, I.R., 1982b. Sequences of structural change in the Dutch Neolithic. In: Hodder, I.R. (Ed.), Symbolic and Structural Archaeology. Cambridge University Press, Cambridge, pp. 162–177.

Hodder, I.R., 1987. The contribution of the long term. In: Hodder, I.R. (Ed.), Archaeology as Long-Term History. Cambridge University Press, Cambridge, pp. 1–8.

Kim, S.-O., 1994. Burials, pigs, and political prestige in Neolithic China. Current Anthropology 35 (2), 119–133.

Knappett, C., 2005. Thinking through Material Culture: An Interdisciplinary Perspective. University of Pennsylvania Press, Philadelphia, PA.

Leone, M., 1982. Some opinions about recovering mind. American Antiquity 47, 742–760.

Leroi-Gourhan, A., 1964. Les Religions de la Prehistoire. Presses Universtaires de France, Paris.

Lewis-Williams, D.J., 2002. The Mind in the Cave. Thames and Hudson, London.

Lewis-Williams, D.J., Dowson, T.A., 1988. The signs of all times: entoptic phenomena in upper paleolithic art. Current Anthropology 29, 201–245.

Malafouris, L., 2004. The cognitive basis of material engagement: where brain, body, and culture conflate. In: DeMarrais, E., Gosden, C., Renfrew, A.C. (Eds.), Rethinking Materiality: The Engagement of Mind with the Material World. McDonald Institute Monographs, Cambridge, pp. 53–62.

Malafouris, L., 2013. How Things Shape the Mind: A Theory of Material Engagement. MIT Press, Cambridge.

Marcus, J., Flannery, K., 1994. Ancient Zapotec ritual and religion: an application of the direct historical approach. In: Renfrew, A.C., Zubrow, E.B.W. (Eds.), The Ancient Mind: Elements of Cognitive Archaeology. Cambridge University Press, Cambridge, pp. 55–74.

McGhee, R., 1977. Ivory for the sea woman: the symbolic attributes of a prehistoric technology. Canadian Journal of Archaeology 1, 141–148.

Miller, D., 1987. Material Culture and Mass Consumption. Blackwell, Oxford.

Mithen, S., 1996. The Prehistory of the Mind: The Cognitive Origins of Art and Science. Thames and Hudson, London.

Reinach, S., 1903. L'art et la magie: à propos des peintures et des gravures de l'âge du renne. L'Anthropologie 14, 257–266.

Renfrew, A.C., 1982. Towards and Archaeology of the Mind: An Inaugural Lecture Delivered before the University of Cambridge. Cambridge University Press, Cambridge.

Renfrew, A.C., 1994. Towards a cognitive archaeology. In: Renfrew, A.C., Zubrow, E.B.W. (Eds.), The Ancient Mind: Elements of Cognitive Archaeology. Cambridge University Press, Cambridge, pp. 3–12.

Renfrew, A.C., 2001. Symbol before concept: material engagement and the early development of society. In: Hodder, I.R. (Ed.), Archaeological Theory Today. Polity Press, London, pp. 122–140.

Renfrew, A.C., 2007. Prehistory: The Making of the Human Mind. Weidenfeld and Nicolson, London.

Salmon, M.H., 2001. Explanation in archaeology. In: Hon, G., Rakover, S.S. (Eds.), Explanation: Theoretical Approaches and Applications. Kluwer Academic Publishers, Dordrecht, pp. 231–248.

Spaulding, A.C., 1960. Statistical description and comparison of artifact assemblages. In: Heizer, R.F., Cook, S.F. (Eds.), The Application of Quantitative Methods in Archaeology. Viking Fund Publications in Anthropology, No. 28. Wenner Gren Foundation for Anthropological Research, New York, pp. 60–83.

Stout, D., Chaminade, T., 2007. The evolutionary neuroscience of tool making. Neuropsychologia 45, 1091–1100.

Thomas, J., 1996. Time, Culture and Identity. Routledge, London.

Tilley, C., 1994. A Phenomenology of Landscape. Berg, Providence.

Tilley, C., 1999. Metaphor and Material Culture. Blackwell, Oxford.

Warren, W.H., 1984. Perceiving affordances: visual guidance of stair climbing. Journal of Experimental Psychology: Human Perception and Performance 10, 683–703.

Whitley, D.S., 1992. Shamanism and rock art in far western North America. Cambridge Archaeological Journal 2 (1), 89–113.

Wynn, T., 1979. The intelligence of later Acheulean hominids. Man 14, 371–391.

Cognitive Behavioral Therapy

David Kingdon, College Keep, University of Southampton, Southampton, UK
Helen Mander, Southern Health NHS Trust, Southampton, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C.J. Mace, volume 3, pp. 2075–2080, © 2001, Elsevier Ltd.

Abstract

Cognitive behavioral therapy (CBT) is a form of psychotherapy focused on the 'here and now,' which is based on the assumed interaction between our emotions, cognitions (mental events including thoughts, memories, images, etc.), and our behaviors. CBT aims to promote change by helping an individual to understand and then alter the way they think, and/or the things they do, leading to indirect changes in the way they feel.

History and Development

In the early twentieth century, attempts to understand the human mind were based on introspective studies of consciousness. Behaviorism evolved with the ambition of creating an objective science with which to study observable behavior (i.e., Watson, 1913; Skinner, 1974). Based on the principles of learning theory, behaviorist explanations of human behavior asserted that behavior is more or less likely in a particular context, due to prior learning (or conditioning) about the consequences that follow that behavior.

In the 1970s – a time of computer modeling and study of cognition – behaviorism, with its absence of an account of conscious mental processes, was felt by some to be limited, overly descriptive, and deterministic. Beck (1976), originally a psychoanalyst, built on the behavioral assumption that an individual's history and experiences will guide their behavior by suggesting a means of how this influence may be transferred. Beck's cognitive theory of emotional disorders (Beck et al., 1979) describes how learned principles are stored in cognitive structures or 'schemas.' Emotional distress is, therefore, caused not directly by environmental contingencies rather by experiences mediated through distorted or unhelpful cognitions.

In line with this theory, Beck developed cognitive therapy, based on the premise that, given the guidance to do so, an individual could be able to identify within consciousness their maladaptive reactions, and use logical problem-solving skills to correct these themselves. Beck's cognitive theory gained prominence through experimental validation of its principles, and through positive evaluation of the resulting therapy (e.g., DeRubeis et al., 2005).

Over the last 20 years, a range of third-wave cognitive and behavioral therapies have developed in a continuing effort to further hone the cognitive theory by identifying the active components and mechanisms and to improve treatment for emotional difficulties. For example, mindfulness-based cognitive therapy (MBCT, Teasdale et al., 2000) was developed in an attempt to combine the short-term pragmatic and financial benefits of pharmacotherapy for treatment of depression with the relapse-protective benefits of cognitive therapy. MBCT assumes that cognitive therapy achieves its effects through helping an individual to alter their extreme, absolutist, dichotomous thinking style, thus changing an individual's relationship with their thoughts and modifying the reciprocal relationship between negative thoughts on low mood. Third-wave theories may vary in philosophical and pragmatic positions; however, they generally focus on the function, rather than the form of cognitions in understanding their role in distress. For example, dialectical behavior therapy (Linehan, 1993) combines CBT with skills training groups and team consultations and has been developed for the treatment of borderline personality disorder. Acceptance and commitment therapy (Hayes and Smith, 2005) uses acceptance and mindfulness strategies to encourage an individual to commit to engagement in valued activities.

Cognitive Behavioral Theory

Beck's cognitive theory of depression (Beck et al., 1979) describes how, through early experiences, people develop beliefs about themselves, the world, and the future. This information is stored as stable, global, core beliefs, not immediately accessible to conscious reflection. Core beliefs may be activated by environmental triggers, such as a situation that is relevant to, or consistent with its content. This leads to activation of conditional rules or 'dysfunctional assumptions' – strongly held personal rules, containing information about how the individual or environment should be in order to satisfy the core belief. Phrased as absolute demands, dysfunctional assumptions are often preverbal, excessive, and unreasonable. Active dysfunctional assumptions then interact with the environment to generate situation-specific statements known as negative automatic thoughts, which are accessible to awareness and lead to low mood and to the symptoms of depression including affective, cognitive, behavioral, and motivational responses.

Although early experiences are considered particularly important in the formation of core beliefs, through our lives we constantly interact with our environments, thus generating new information to update and correct them. However, certain cognitive processes may mean that unhelpful information persists despite new contradictory information. For example, individuals with depression show an increase in cognitive processes, such as mood-congruent thoughts or selective

attention, which manipulate interpretations of events, making it more likely that new information will be consistent with existing schemas.

Since the introduction of Beck's theory, specific cognitive accounts or models have been developed to explain emotional difficulties in other areas. These models are based on clinical observations, formed into hypotheses and tested through experimental methods. The overriding commonality between models is the assumption that distress is the result of an unhelpful interpretation of an environmental or internal trigger. The nature of this interpretation will determine the emotional response. Those characterized by overestimation of threat are likely to lead to anxious responses, interpretations of loss lead to a sad or depressed emotional response, whereas interpretations connected to injustice may lead to angry responses. These interpretations are then unwittingly and unfortunately maintained through a combination of cognitive, behavioral, affective, and interpersonal responses, the relative contribution of which are differentially emphasized across various different psychological problems. For example, an individual presenting with a phobia of dogs may experience their distress as a result of their interpretation that "all dogs are dangerous and will bite me," which may be triggered every time that a dog is present in their immediate environment. This interpretation of threat may then be maintained through an avoidance of all dogs (thus precluding any opportunity for disconfirmation of this anxious prediction), and hypervigilance and cognitive biases around dog-threat related stimuli (therefore making it more likely that any barking in the area will be noticed and interpreted as a sign that a dog is out of control).

Cognitive Behavioral Therapy

Cognitive behavioral therapy (CBT) is delivered through a collaborative, empathic, nonjudgmental therapeutic relationship between therapist and client. CBT varies in duration and delivery, according to the target difficulty and therapeutic setting, however generally involves between 5 and 20 weekly sessions, lasting between 30 and 60 min. CBT is a focused psychotherapy – directed toward explicit meaningful targets identified by the client. CBT sessions themselves are also focused, making use of an 'agenda' to agree on shared goals for a session and ensure that adequate attention is allowed for these. Outcome-driven CBT involves regular evaluation toward its therapeutic targets, which may include the use of standardized or idiosyncratic measures.

CBT generally begins with an individualized assessment of current and historical factors, which may be contributing to the development and maintenance of the identified difficulty. From this assessment, a formulation is developed and shared. The aim of this formulation is to understand and explain the factors that may be influential in the focal difficulty, and based on this, to provide a plan for intervention. This formulation may then be updated and developed throughout the therapeutic intervention, as additional information is gathered.

Based on the formulation, intervention strategies will be used to support the individual to make changes. As such, a formulation, which indicates that an unhelpful interpretation is influential in leading to distress, may lead to strategies designed to update this interpretation with a more functional alternative. A formulation indicating that continued avoidance of a feared situation may be involved in maintaining a fear through lack of opportunity to disconfirm this appraisal would be likely to lead to an intervention that includes gradually and collaboratively eliminating the avoidance behavior to allow the client to become accustomed to the feared stimulus.

Strategies used within CBT are intended to facilitate the client to arrive at more helpful and functionally adaptive conclusions. Rather than reaching this change through teaching or offering insight, the therapist uses method such as 'guided discovery' and 'Socratic questioning' to draw information from the client themselves. Regular feedback is used to increase the shared understanding, promote insight, and focus on therapeutic issues. Within CBT, clients are encouraged to use discussion and education to seek further information with which to update their unhelpful beliefs. An intervention is also likely to include encouraging the client to behave in new ways or to conduct 'behavioral experiments,' in order to explore the consequences of alternative actions and make sense of these within their emerging framework. Other CBT techniques may include the use of role plays and diaries. Central to the aforementioned aim of supporting the client to identify and correct their own unhelpful reactions is the use of 'homework.' Clients are encouraged to develop and carry out between-session tasks to continue and expand their therapeutic change and increase the generalization of any within-session changes.

Applications

The majority of evidence around the effectiveness of CBT has been gathered through trials of treatment efficacy in different therapeutic areas. There is a strong evidence base for the use of CBT in the treatment of depressive disorders; CBT is equally effective to antidepressant medication, but with more enduring effects (e.g., Paykel et al., 1999). CBT is also the psychological treatment of choice for anxiety disorders (Butler et al., 2006) including social anxiety, panic disorder, obsessive-compulsive disorder, post-traumatic stress disorder, and generalized anxiety disorder. CBT is recommended in the treatment of eating disorders such as bulimia nervosa (Agras et al., 2000) and in the treatment of chronic fatigue (Malouff et al., 2008).

More recently, the evidence base has grown for the effectiveness of CBT in more severe and enduring mental health difficulties including treatment of schizophrenia (Tarrier and Wykes, 2004) with some benefit observed using CBT in the treatment of bipolar disorder (Scott et al., 2006) and in the treatment of personality disorders (e.g., Davidson et al., 2006).

As a time-limited, evidence-based, and focused therapy, CBT has been recommended extensively for use within the UK National Health Service. Recent initiatives designed to increase dissemination of CBT have led to the implementation of 'Improving Access to Psychological Therapies' services, designed to provide short-term outcome-driven psychological therapy to large numbers of clients within a primary care setting (Clark, 2011). This initiative has since been expanded in order to increase provision of these therapies for other groups including children and young people, and those with severe and enduring mental health difficulties. As such, emphasis has

been placed on development of more brief and focused interventions.

CBT is traditionally carried out by mental health professionals (e.g., nurses, psychologists, psychiatrists, etc.) with specific training in providing CBT. More specific packages and individual CBT skills can be used by practitioners without specialist training. The majority of outcome research into CBT has used individual therapy; however, CBT has also been applied in group formats (Tucker and Oei, 2007). It has been shown to lead to effective reduction of symptoms in a computer-delivered format (e.g., Hayward et al., 2007) and delivered through the Internet (Kiropoulos et al., 2007).

Criticisms of Cognitive Behavioral Therapy

In common with many psychotherapies and treatments, CBT is not suitable for everyone. The structured, present-focused stance of CBT may be uncomfortable for individuals who are unable to identify a focal difficulty, or who are interested mainly in gaining greater insight into their experiences. Furthermore, the collaborative attitude means that clients are required to undertake much of the responsibility and effort involved in making changes. There have also been questions about its cultural applicability, although, with adaptation, it can be successful (Rathod et al., 2013). This change necessarily involves moving outside of one's comfort zone, and as such is likely to be difficult for those with difficulties around low motivation or high avoidance.

See also: Anxiety and Anxiety Disorders; Behavior Therapy: Background, Basic Principles, and Early History; Bipolar (Manic-Depressive) Disorder; Borderline (Patient) Personality; Depression; Mania; Obsessive-Compulsive Disorder; Panic Disorder; Post-Traumatic Stress Disorder.

Bibliography

Agras, W.S., Walsh, B.T., Fairburn, C.G., Wilson, G.T., Kraemer, H.C., 2000. A multicenter comparison of cognitive-behavioural therapy and interpersonal psychotherapy for bulimia nervosa. Archives of General Psychiatry 57, 459–466.

Beck, A.T., 1976. Cognitive Therapy and the Emotional Disorders. International Universities Press, Michigan.

Beck, A.T., Rush, A.J., Shaw, B.F., Emery, G., 1979. Cognitive Therapy of Depression. Guildford Press, New York.

Butler, A.C., Chapman, J.E., Forman, E.M., Beck, A.T., 2006. The empirical status of cognitive-behavioral therapy: a review of meta-analyses. Clinical Psychology Review 26, 17–31.

Clark, D.M., 2011. Implementing the NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. International Review of Psychiatry 23, 375–384.

Davidson, K., Norrie, J., Tyrer, P., Gumley, A., Tata, P., Murray, H., et al., 2006. The effectiveness of cognitive behavior therapy for borderline personality disorder: the results from the borderline personality disorder study of cognitive therapy (BOSCOT) trial. Journal of Personality Disorders 20, 450–465.

DeRubeis, R.J., Hollon, S.D., Amsterdam, J.D., Shelton, R.C., Young, P.R., Salomon, R.M., et al., 2005. Cognitive therapy vs medications in the treatment of moderate to severe depression. Archives of General Psychiatry 62, 409–416.

Hayes, S.C., Smith, S., 2005. Get Out of Your Mind and into Your Life: The New Acceptance and Commitment Therapy. New Harbinger Publications.

Hayward, L., MacGregor, A., Peck, D.F., Wilkes, P., 2007. The feasibility and effectiveness of computer-guided CBT (FearFighter) in a rural area. Behavioural and Cognitive Psychotherapy 35, 409–419.

Kiropoulos, L.A., Klein, B., Austic, D.W., Gilson, K., Pier, C., Mitchell, J., et al., 2007. Is internet-based CBT for panic disorder and agoraphobia as effective as face-to-face CBT? Journal of Anxiety Disorders 22, 1273–1284.

Linehan, M.M., 1993. Cognitive Behavioral Treatment of Borderline Personality Disorder. Guilford Press, New York.

Malouff, J., Thorsteinsson, E., Rooke, S., Bhullar, N., Schutte, N., 2008. Efficacy of cognitive behavioural therapy for chronic fatigue syndrome: a meta-analysis. Clinical Psychology Review 28, 736–745.

Paykel, E.S., Scott, J., Teasdale, J., et al., 1999. Prevention of relapse in residual depression by cognitive therapy: a controlled trial. Archives of General Psychiatry 56, 829–835.

Rathod, S., Phiri, P., Harris, S., Underwood, C., Thagadur, M., Padmanabi, U., et al., 2013. Cognitive behaviour therapy for psychosis can be adapted for minority ethnic groups: a randomised controlled trial. Schizophrenia Research 143 (2–3), 319–326.

Scott, J., Paykel, E., Morriss, R., Bentall, R., Kinderman, P., Johnson, T., et al., 2006. Cognitive–behavioural therapy for severe and recurrent bipolar disorders. Randomised controlled trial. British Journal of Psychiatry 188, 313–320.

Skinner, B.F., 1974. About Behaviorism. Vintage, New York.

Tarrier, N., Wykes, T., 2004. Is there evidence that cognitive behaviour therapy is an effective treatment for schizophrenia? A cautious or cautionary tale? Behaviour Research and Therapy 42, 1377–1401.

Teasdale, J.D., Segal, Z.V., Williams, J.M.G., Ridgeway, V.A., Soulsby, J.M., Lau, M.A., 2000. Prevention of relapse/recurrence in major depression by mindfulness-based cognitive therapy. Journal of Consulting and Clinical Psychology 68, 615–623.

Tucker, M., Oei, T.P.S., 2007. Is group more cost effective than individual cognitive behaviour therapy? the evidence is not solid yet. Behavioural and Cognitive Psychotherapy 35, 77–91.

Watson, J., 1913. Psychology as the behaviourist views it. Psychological Review 20, 158–177.

Cognitive Control (Executive Function): Role of Prefrontal Cortex

Jonathan D Cohen, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp. 2089–2094, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Cognitive control is the ability to direct mental function and behavior in accord with internally represented intentions or goals. This ability is fundamental to all higher cognitive faculties, such as language, reasoning, problem solving, and social behavior. While this ability no doubt relies on interactions between many areas of the brain, neuropsychological, neurophysiological, neuroimaging, and neural network modeling studies all indicate strongly that the prefrontal cortex plays a critical role in cognitive control. This article reviews these findings briefly, and suggests that the prefrontal cortex executes control by guiding the flow of neural activity along pathways in other parts of the brain responsible for performing an intended task.

Control refers to the ability to direct mental function and behavior in accord with an internally represented set of intentions. This is manifest in higher cognitive function in many forms: the ability to direct attention to a specific stimulus within a large array of other competing, and perhaps more salient stimuli (e.g., finding the face you are looking for within a crowd), maintaining a new and important piece of information in mind against distraction (e.g., remembering a telephone number you just got from the directory assistance until you dial the same, while in a phone booth at a noisy airport terminal), overcoming a compelling but undesirable behavior (e.g., suppressing the urge to scratch a mosquito bite) or pursuing a complex but unfamiliar behavior (e.g., playing a new piano piece) and, perhaps most importantly, the ability to respond flexibly and productively in novel circumstances (e.g., mentally explore the consequences of a complex sequence of moves in a game of chess).

The distinction between controlled and automatic processing is one of the central concepts within modern cognitive psychology. Controlled processing is considered to be effortful, and to rely on a limited capacity system, while automatic processing is assumed to occur independently of this system (Posner and Snyder, 1975; Shiffrin and Schneider, 1977). This concurs with common experiences, such as the ability to carry on a conversation while driving a car (an automatic process) but not while conducting multidigit arithmetic in one's head (a process that relies on control). Many contemporary theories posit that there is actually a continuum between controlled and automatic processing (for a review, see Cohen et al., 1990). Nevertheless, virtually all theorists acknowledge the need for some mechanism, or set of mechanisms, responsible for the coordination of processing in a flexible fashion – particularly in novel or demanding tasks. This idea figures centrally in Baddeley's classic theory of working memory (Baddeley, 1986), which postulates two critical components: a storage component responsible for the active maintenance of information in a short-term store, and an executive control component responsible for the manipulation and coordinated use of this information. For example, in a multidigit multiplication problem, the storage component maintains the intermediate products while the executive carries out the arithmetic operations.

The postulation of a central executive closely paralleled theorizing regarding the nature of frontal lobe function (for a review, see Shallice, 1982), based on the clinical observation that patients with frontal lesions exhibit a 'dysexecutive syndrome.' The frontal lobes are the areas of the brain most highly expanded in humans relative to other species (see Figure 1). Damage to this area is associated with impairments in characteristically human cognitive functions that are directly

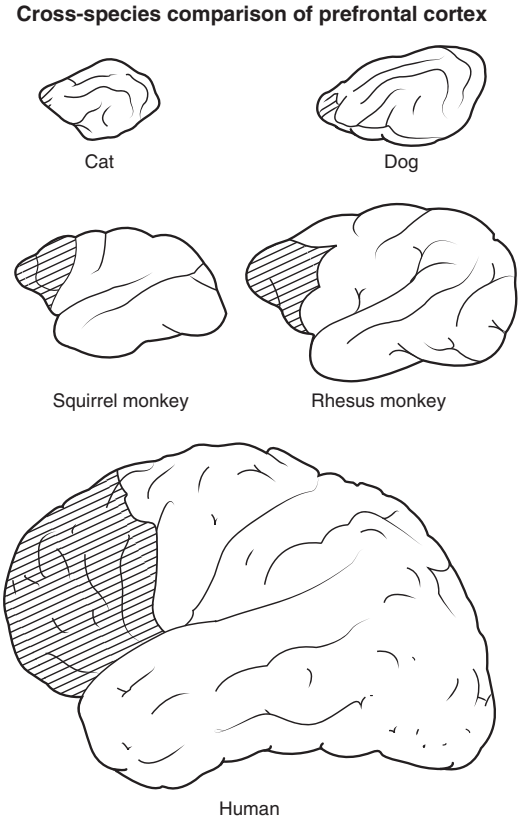

Figure 1 Relative size of prefrontal cortex (gray region) in five species. http://www.psychol.ucl.ac.uk/kate.jeffery/c567/lecture13_frontal/overheads/img003.jpg.

dependent upon control (such as planning or the ability to respond adaptively in novel circumstances), and disturbances of this structure have been implicated consistently in neuropsychiatric diseases such as schizophrenia that also appear to be uniquely human. Indeed, the earliest neurologists, neuroscientists, and neuropsychologists recognized the importance of the frontal lobes in the control of behavior. Perhaps this observation was made most dramatically in the classic case of Phineas Gage, the unfortunate railroad foreman who suffered damage to his prefrontal cortex (PFC) when a 1½″-diameter rod penetrated his skull in a construction accident. Originally someone who was considered to be thoughtful, responsible, and of sound judgment, following the accident was described as "capricious…and unable to settle on any of the plans he devised for future action" (Harlow, 1848). Such changes have been observed repeatedly in patients with damage to the frontal cortex (for a review, see Stuss and Benson, 1986). Furthermore, patients with frontal damage perform poorly in tasks that require even the simplest forms of cognitive control, such as the Wisconsin Card Sorting Task (WCST) (Milner, 1964), and, most recently, brain imaging studies consistently have revealed increased activity of the PFC while subjects are performing tasks that demand cognitive control (for a review, see Miller and Cohen, 2001). Although this accumulation of findings has led most investigators to assume that the PFC plays a critical role in cognitive control, they have provided little insight into the specific contribution that it makes, or the mechanisms by which it operates. However, more detailed neurobiological studies have begun to shed light on this question.

Using single unit recording techniques, Fuster and Alexander (1971) and Niki (1974) reported the remarkable finding that some neurons in the PFC continue to fire during the delay between a stimulus and a contingent response, even after the stimulus has disappeared. Goldman-Rakic and her colleagues followed up on this finding and, in a series of elegant studies, demonstrated that the firing of such neurons was stimulus selective and was critical for performance in delayed-response tasks (Goldman-Rakic, 1990).

These findings strongly implicate neuronal activity in PFC as the site of temporary storage of information in working memory. Recent neuroimaging studies of humans have provided strong convergent support for this idea, demonstrating sustained activity in PFC during the performance of simple working memory tasks (for a review, see Smith and Jonides, 1999). However, this discovery seemed to pose an interesting puzzle. Most of the neuropsychological data suggested that the PFC housed the central executive which, at least within Baddeley's influential theory of working memory, was clearly distinguished from the storage component thought by most to be housed primarily in more posterior structures. Computational models of cognitive control have offered a different perspective on this problem that may reconcile the role of the PFC in control and storage.

Computational Models

Shallice (1982) presented the first account of the role of PFC in cognitive control within an explicitly computational framework. He proposed that the PFC housed a supervisory attentional system (SAS) – a mechanism by which PFC coordinates complex cognitive processes. Shallice's theory was described in terms of a production system architecture. This has appeal, as it relates the executive functions of the frontal cortex to the well-characterized mechanisms of other production system theories that include the active representation of goal states to coordinate the sequences of production firings involved in complex behaviors (e.g., Anderson, 1983). One feature of goal representations is that they must be maintained actively throughout the course of a sequence of actions to direct behavior effectively. This coincides with the observation that PFC appears to be specialized for the active maintenance of task-relevant information. Thus, it is possible that PFC is specialized for the maintenance of a *particular type* of information, as a means of executing control over performance.

While Shallice's theory was not implemented originally as a functioning model, Kimberg and Farah (1993) proposed a model of frontal function – also using a production system architecture – that simulated performance in a variety of tasks considered to rely on frontal lobe function (including the WCST), and illustrated that damage to this component of the model produced impairments similar to those observed in patients with frontal damage. However, while such models have produced insights into the functional role of PFC in cognitive control, their components do not have a transparent mapping on to specific neurobiological mechanisms. Neurobiological plausibility is not a requirement, per se, of a theory that seeks to explain the cognitive functions of PFC. Nevertheless, the question of how goal-directed behavior arises from the firing of millions of neurons presents a mystery in its own right. Furthermore, understanding how different forms of cognitive impairment arise from different forms of damage to PFC (e.g., the effects of stroke or injury vs. schizophrenia) provides an important motivation for understanding how cognitive control may arise from the specific neurobiological mechanisms housed within PFC.

Recently, investigators have begun to use neural network models to better understand how PFC may carry out its functions. Such models (also known as connectionist or parallel-distributing processing models) simulate the behavioral performance of human subjects (or animals) in cognitive tasks using neurally plausible processing mechanisms (e.g., the spread of activity among simple processing units along weighted connections; see Rumelhart and McClelland, 1986). The goal of this effort is to identify principles of neural function that are most relevant to behavior. Using this approach, Dehaene and Changeux (1989), Levine and Prueitt (1989), and Cohen and Servan-Schreiber (1992) have all described models of prefrontal function, and used these to simulate the performance of normal and frontally damaged patients in tasks that are sensitive to PFC damage, such as the WCST and others. All of these models simulate PFC function as the activation of a set of units that represent the 'rules' of the task; that is, units whose activation leads to a task-relevant response, even when this may not be the one most strongly associated with the stimulus. However, in most models, the PFC units themselves are not responsible for generating the response directly. Rather, they influence the activity of other units whose responsibility this is.

This is clearly illustrated by a model of the Stroop task developed by Cohen et al. (1990).

In the Stroop task (Stroop, 1935), subjects respond to words either by reading them, or by naming the color in which they are displayed. In the critical condition, the word itself conflicts with the color in which it is displayed (e.g., the word 'GREEN' displayed in red). Subjects have no trouble reading such words (e.g., saying 'green'). However, when they are asked to name the color (e.g., say 'red'), they are significantly slower, and sometimes even make errors. This reflects the highly practiced, and therefore prepotent tendency to read written words, which interferes with the ability to name the color of a Stroop conflict stimulus. The ability to name the color in the face of such interference (i.e., to produce the weaker, but task-relevant response) is a simple but clear example of cognitive control.

Cohen et al. (1990) constructed a model of this task, as shown in **Figure 2**. The connections among the units in this model defined two processing pathways, one for word reading and another for color naming. The connections in the word-reading pathway were stronger, capturing the assumption that this was the more practiced task. Because of these stronger connections, information flowing along the word pathway interfered with color naming, simulating the interference effects observed when human subjects perform this task. Indeed, the model's ability to produce a response to the color in the face of such interference required the addition of a set of units (labeled 'Dimension' in **Figure 2**), which provided additional activation of the units in the color-naming pathway. This additional activation biased processing in favor of this pathway, allowing it to compete more effectively with, and prevail over, activity flowing along the stronger word pathway. This biasing effect corresponds precisely to the role of top-down control in neurophysiological models of attention (such as the Biased Competition Model of Desimone and Duncan, 1995), and has been proposed as a mechanism by which PFC exerts control over processing (Cohen and Servan-Schreiber, 1992; Miller and Cohen, 2001). Models such as this have been used to simulate the performance of both normal subjects and patients with frontal damage in a wide range of tasks that tap cognitive functions commonly associated with PFC function, such as working memory, attention, behavioral inhibition, planning, and problem solving.

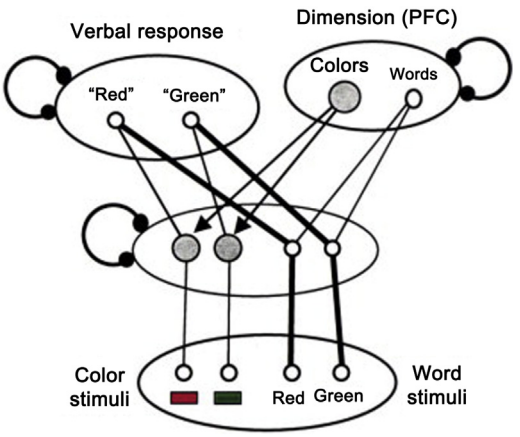

Figure 2 Model of the Stroop task, showing the pathways for word reading, color naming, and top-down control. Connections between units in different layers are excitatory and within layers are inhibitory (closed circles). Shown with color dimension unit active, biasing color-naming pathway. Adapted from Cohen, J.D., Dunbar, K., McClelland, J.L., 1990. On the control of automatic processes: a parallel distributed processing account of the Stroop effect. Psychology Review 97, 332–361.

Guided Activation as a Method of Control

The Stroop model brings several features of PFC function into focus. First, it emphasizes the view that the role of the PFC in control is modulatory, guiding the flow of activity along pathways in other parts of the brain that are responsible for task performance. For example, activating the color unit does not in itself transmit information about a particular response (red or green). Rather, it simply insures that activity flowing along the color-naming pathway will have a greater influence over the response than activity flowing along the word pathway. In this way, representations within the PFC can function as intentions, rules, or goals (comparable to those in a production system architecture), by setting up the appropriate relationship between a stimulus (or category of stimuli) and an associated response (or set of responses) through the proper guidance of activity along pathways in other parts of the brain. Recent neurophysiological findings regarding the firing properties of neurons in PFC have provided strong support for this view (for a review, see Miller and Cohen, 2001). Recent neuroimaging studies have also provided convergent support, indicating that PFC activity occurs when behavior relies upon explicit knowledge about rules or arbitrarily determined conjunctions of stimulus features (for a review, see Miller and Cohen, 2001). Note that this function is not necessarily restricted to mappings from stimuli to responses, but applies equally well to mappings involving internal states (e.g., thoughts, memories, and emotions). Note also that this function is not necessarily unique to PFC. There may also be more local forms of control, responsive to regionally specific needs for the biasing of processing. However, the wide range of anatomic connections that the PFC shares with virtually all associative areas of the brain places it in a strategic position to guide the flow of activity between and among these other regions (for a review, see Miller and Cohen 2001) – a position well suited to its presumed role in the control of higher cognitive processes.

The emphasis of this guided activation model of PFC on its modulation of other brain areas actually responsible for task execution is consistent with the classic pattern of neuropsychological deficits associated with frontal lobe damage: the individual elements of a complex behavior are usually left intact, but the subject is not able to coordinate them in a task-appropriate way (e.g., a patient, who when preparing coffee, first stirred and then added cream; Shallice, 1982). The guided activation model also captures another critical feature of theories about the role of PFC in executive function: the importance of sustained activity as a critical component of control. For a representation to have a biasing influence, it must be activated. For this influence to endure (e.g., over the course of performing a task), its activity must be sustained. For example, to continue color naming, the activity of the color

unit must be maintained, lest the word begins to dominate processing. Similarly, an increase in the demand for control requires greater or more enduring activity of the corresponding units in PFC. This concurs with accumulating evidence from the neuroimaging literature, that tasks thought to rely on controlled processing consistently engage the PFC (for a review, see Smith and Jonides, 1999). It is also consistent with both behavioral and neuroimaging findings regarding the effects of practice on automaticity and the involvement of PFC. Increased practice on a task should strengthen its underlying pathway, reducing its reliance on control. Indeed, consistent practice on a task reduces the amount of PFC activity observed, and PFC damage can impair new learning but spare performance on well-practiced tasks (for a review, see Miller and Cohen, 2001).

Finally, this approach helps to unify the role that PFC plays in the variety of cognitive functions with which it has been associated, such as selective attention, behavioral inhibition, and executive function in working memory. These can all be seen as varying reflections, in behavior, of the operation of a single underlying mechanism of cognitive control: the biasing effects of representations in PFC on processing in pathways responsible for task performance. For example, selective attention and behavioral inhibition may be viewed as two sides of the same coin: attention is the effect of biasing competition in favor of task-relevant information, and inhibition is the consequence that this has for the irrelevant information (cf. the Biased Competition Model of Desimone and Duncan, 1995). This assumes that inhibition occurs due to local competition between conflicting representations (e.g., between the two responses in the Stroop model), rather than centrally by the PFC. The 'binding' function of selective attention can also be explained by such a mechanism, by assuming that PFC representations can select the desired combination of stimulus features (over other competing combinations) to be mapped on to a response.

Outstanding Questions

The guided activation model provides an integrative and mechanistically explicit framework for considering the role of PFC in cognitive control. At the same time, it brings into focus a number of important and unanswered questions. For example, despite the longstanding observation of delay-period activity for PFC neurons, and the centrality of this property in neural network models of PFC, little is known about the actual mechanisms by which neuronal activity is sustained. This could reflect a cellular property of PFC neurons (e.g., bistability), or a circuit-level phenomenon (e.g., recirculation of activity within the PFC, or between PFC and other structures). Several models have proposed the latter, assuming that representations are maintained in PFC as attractor states (for a review, see O'Reilly et al., 1999). However, neurobiological studies are needed to confirm this hypothesis. A closely related question concerns the mechanisms by which patterns of activity are updated in PFC. These must be able to satisfy two conflicting demands: on the one hand, they must be responsive to relevant changes in the environment; and on the other, they must be resistant to updating by irrelevant changes. Neurophysiological studies suggest that PFC representations are selectively responsive to task-relevant stimuli and robust to interference from distractors (for a review, see Miller and Cohen, 2001). Not surprisingly, two hallmarks of damage to PFC are perseveration (failure to update) and distractibility (inappropriate updating). These observations suggest the operation of mechanisms that insure the appropriate updating of PFC activity in response to behavioral demands.

Recent modeling work has suggested that brainstem dopaminergic systems and the basal ganglia may play an important role both in updating representations in PFC and learning how and when to do so (for a review, see Miller and Cohen, 2001). Other work has suggested that the anterior cingulate cortex – a midline structure within the frontal lobes – may play an important role in monitoring task performance, and identifying the need to allocate control (Botvinick et al., 2001). Such theories make important predictions about the neural mechanisms underlying cognitive control that serve as valuable challenges to future neurobiological research in this area.

An equally fundamental question concerns the nature of representations in PFC and how they arise. For example, the Stroop model assumes there are units that represent each of the two dimensions of the stimulus (color and word), with the appropriate connections to the corresponding pathways. However, are there such units in the PFC for every possible combination of stimulus and response of which tasks may be composed? This seems unlikely. Yet, there must be sufficient representational richness to support the flexibility in behavior that the PFC seems to afford. What principles define this set of representations, how they are organized, and how they are learned are important questions at both the computational and neurobiological levels.

A better understanding of the mechanisms underlying active maintenance may also provide insight into one of the most striking and perplexing properties of cognitive control: its severely limited capacity. This has long been recognized in cognitive psychology (Posner and Snyder, 1975; Shiffrin and Schneider, 1977), and is painfully apparent to anyone who has tried to talk on the phone and read an email at the same time. The resource limitation of cognitive control has played an explanatory role in many important models of human cognition. However, to date, no theory has provided an explanation of the limitation *itself*. This is a *sine qua non* of cognitive control, and, therefore, provides an important benchmark for any theory that seeks to explain its underlying mechanisms.

Finally, it is important to recognize that the PFC is certainly not the only brain structure involved in cognitive control. As noted above, mechanisms similar to those within PFC may operate locally in other parts of the brain. Furthermore, there are certainly other types of mechanisms critical to cognitive control. For example, the mechanisms responsible for keeping an immediate goal in mind (e.g., working on a book chapter) are not likely to be the same ones responsible for realizing long-term goals (e.g., getting the book published). While the former may be guided by representations actively maintained in PFC, the latter almost certainly engages mechanisms of long-term storage. Some suggestions have been made about how the PFC may interact with the hippocampus to orchestrate the storage and retrieval of goal representations at appropriate times (O'Reilly et al., 1999); however, this remains another area in need of further research.

Conclusion

One of the great mysteries of the brain is how purposeful goal-directed behavior emerges from the millions of relatively simple processing units that are its basic computational elements. Behavioral, neuropsychological, and neurobiological data converge on the idea that the frontal lobes play a critical role in cognitive control. Neural network models have begun to suggest how this may be carried out, as sustained patterns of activity within PFC modulate, or 'guide' the flow of activity along pathways in other parts of the brain responsible for performing a task. However, many fundamental questions remain to be addressed. The human brain is arguably the most complex device in the known universe, and its capacity for higher cognitive function continues to be one of its deepest mysteries. Unraveling this mystery stands as one of the most exciting challenges in science, and the rapid development of sophisticated new empirical methods (such as functional brain imaging) and theoretical tools (such as neural network modeling) offer hope that this challenge can be met.

See also: Dysexecutive Syndromes; Prefrontal Cortex Development and Development of Cognitive Function; Prefrontal Cortex.

Bibliography

Anderson, J.R., 1983. The Architecture of Cognition. Harvard University Press, Cambridge, MA.

Baddeley, A., 1986. Working Memory. Clarendon Press, Oxford, UK.

Botvinick, M.M., Braver, T.S., Carter, C.S., Barch, D.M., Cohen, J.D., 2001. Conflict monitoring and cognitive control. Psychological Review.

Cohen, J.D., Dunbar, K., McClelland, J.L., 1990. On the control of automatic processes: a parallel distributed processing account of the Stroop effect. Psychology Review 97, 332–361.

Cohen, J.D., Servan-Schreiber, D., 1992. Context, cortex and dopamine: a connectionist approach to behavior and biology in schizophrenia. Psychology Review 99, 45–77.

Dehaene, S., Changeux, J.P., 1989. A simple model of prefrontal cortex function in delayed-response tasks. Journal of Cognitive Neuroscience 1, 244–261.

Desimone, R., Duncan, J., 1995. Neural mechanisms of selective visual attention. Annual Review of Neuroscience 18, 193–222.

Fuster, J.M., Alexander, G.E., 1971. Neuron activity related to short-term memory. Science 173, 652–654.

Goldman-Rakic, P.S., 1990. Cellular and circuit basis of working memory in prefrontal cortex of nonhuman primates. Progress in Brain Research 85, 325–335.

Harlow, J.M., 1848. Passage of an iron rod through the head. Boston Medical and Surgical Journal 39, 389–393.

Kimberg, D.Y., Farah, M.J., 1993. A unified account of cognitive impairments following frontal lobe damage: the role of working memory in complex organized behavior. Journal of Experimental Psychology 122, 411–428.

Levine, D.S., Prueitt, P.S., 1989. Modeling some effects of frontal lobe damage – novelty and perseveration. Neural Networks 2, 103–116.

Miller, E.K., Cohen, J.D., 2001. An integrative theory of prefrontal cortex function. Annual Review of Neuroscience 24, 167–202.

Milner, B., 1964. Some effects of frontal lobectomy in man. In: Warren, J.M., Akert, K. (Eds.), The Frontal Granual Cortex and Behavior. McGraw-Hill, New York, pp. 313–331.

Niki, H., 1974. Prefrontal unit activity during delayed alternation in the monkey. 1. Relation to direction of response. Brain Research 68, 185–196.

O'Reilly, R.C., Braver, T.S., Cohen, J.D., 1999. A biologically-based computational model of working memory. In: Miyake, A., Shah, P. (Eds.), Models of Working Memory: Mechanisms of Active Maintenance and Executive Control. Cambridge University Press, New York.

Posner, M.I., Snyder, C.R.R., 1975. Attention and cognitive control. In: Solso, R.L. (Ed.), Information Processing and Cognition. Erlbaum, Hillsdale, NJ.

Rumelhart, D.E., McClelland, J.L., 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge, MA.

Shallice, T., 1982. Specific impairments of planning. Philosophical Transactions of the Royal Society of London Series B 298, 199–209.

Shiffrin, R.M., Schneider, W., 1977. Controlled and automatic human information processing: II. Perceptual learning automaticity, attending and a general theory. Psychology Review 84, 127–190.

Smith, E.E., Jonides, J., 1999. Storage and executive processes in the frontal lobes. Science 283, 1657–1661.

Stroop, J.R., 1935. Studies of interference in serial verbal reactions. Journal of Experimental Psychology 18, 643–662.

Stuss, D.T., Benson, D.F., 1986. The Frontal Lobes. Raven, New York.

Cognitive Development: Child Education

Zhenlin Wang and Lamei Wang, The Hong Kong Institute of Education, New Territories, Hong Kong

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 2094–2098, © 2001, Elsevier Ltd.

Abstract

Learning is a process of hypothesis testing based on probabilistic models. Children learn through play, observation, and through explicit teaching. The domain-general cognitive abilities of self-regulation and executive function help children succeed in school by setting learning goals, focusing their attention on the learning tasks at hand, resisting temptations, and regulating their emotions. Children's learning strategies show great variability, both within and across individuals. Education should provide children with opportunities of supported exploration and experimentation, intentional demonstration and instruction, as well as guided discovery.

Understanding children's cognitive development lays the foundation for child education. On the one hand, the recent advancement in cognitive development challenges some of the conventional views of how children learn and how they succeed in school; on the other hand, it also provides theoretical underpinning and empirical support for some of the well-recognized ideologies and practices in early education. This article highlights the contributions of cognitive development research to child education in three areas. The first area focuses on domain-specific developmental mechanisms for learning including theory theory and probabilistic learning model, theory of mind, testimony, and conceptual change. The second area is the domain-general developmental mechanisms focusing on self-regulation and executive function, as well as their relation to school readiness and success. Although pretense is arguably associated with both domain-general and domain-specific development, its relation with self-regulation is highlighted. The last area discusses children's learning process itself and the elaborate relation between learning and development.

Domain-Specific Developmental Mechanisms for Learning

Theory Theory and Probabilistic Learning

Children's learning mechanism is a topic deeply embedded in the debate of nature vs nurture. Nativists view learning as unfolding of innate, modulelike structures when triggered by environmental stimuli, while empiricists deny the very existence of abstract, coherent structures and see learning as building associations among information that is context dependent. Rooted in Piaget's constructivism, developmental 'theory theorists' have suggested that the way children learn resembles the process of scientific theory formation (e.g., Gopnik and Meltzoff, 1997). The recent development of the theory theory (Gopnik, 2012) attempts to consolidate the nativists' and the empiricists' views of the mechanism of children's learning by applying probabilistic learning model based on Bayesian statistics. Bayesian statistics conceive that the truth or falsity of propositions is uncertain and the evidence of the true state of the world is expressed as degrees of belief. Statements concerning the causal structures are considered as hypotheses. Prior probability of a hypothesis is specified, and then constantly updated in the light of new data. The probabilistic learning account proposes that young children, even infants, can infer causal structure from statistical information. Structures of the real world can be represented mathematically by generative models that allow the learner to compute the patterns of evidence generated by the structure and make inferences. The learner can also invert that process and learn about the structure of the real world from evidence. What is unique about the probability model is that it helps the learner to pick out the most likely hypothesis among all the hypotheses that are compatible with the pattern of evidence based on probability. Consequently, children are not constrained by the innately determined representations, nor do they start from a blank slate; instead, they are constantly assessing their data and testing hypotheses as scientists do. Learning becomes a process of integrating prior knowledge and new evidence to revise the representations of the world based on experience. In a recent review, Gopnik and Wellman (2012) provided empirical evidence that children from 16 months to 4 years of age adopt the probabilistic learning mechanisms such as informal experimentation through play, imitation, and informal pedagogy in their physical and psychological causal inferences.

One of the most appealing implications of the probabilistic learning account for educators is that it advocates explorative play as informal experimentation to generate causally relevant and informative evidence for children to make causal inferences. Gopnik and Wellman (2012: p. 1095) call this type of learning as 'learning from interventions.' They argued that by intervening first hand, one could rapidly narrow down the number of possible hypotheses in order to search more efficiently. These interventions need not to be systematically controlled experiments. Play, as informal experimentation, is sufficiently systematic to help children actively seek out evidence and discover causal structure. In the education sector, play has long been held as the driving force for learning during early childhood. The probabilistic learning account provides the theoretical underpinning from cognitive development perspective for the play-based learning viewpoint for the first time. Furthermore, this account also backs up the inquiry-based scientific education in general.

Theory of Mind

Children not only learn from firsthand exploration, they also learn new causal relations by observing others' interventions and through imitation. Gopnik and Wellman (2012) argued that the theory formation mechanism of children's learning not only applies to the causal structure of the world, but also the mind of the person teaching them. Indeed, teaching and learning are enterprises involving a theory of mind, the intuitive understanding of one's own and other people's mental states and activities, such as thoughts, beliefs, perceptions, knowledge, intentions, desires, and emotions, and of behavioral consequences of such mental states. A significant developmental milestone during early years of life, the awareness of people's mental life, and its relationship with behavioral outcome, is being proposed to be critical for learning and being taught more generally (Wellman and Lagattuta, 2004).

Recognizing existing differences and changes in knowledge is a theory of mind reasoning skill that is essential in teaching and learning. Young children will not yet be able to assume an intentional stance in learning until they see the possibility of acquiring knowledge. Research on the relationship between theory of mind and teaching and learning (e.g., Frye and Ziv, 2005; Wang, 2010) had shown that young children were sensitive to both the epistemic and motivational mental states in teaching and learning, such as the differences and changes in knowledge state, beliefs about knowledge state, and teaching and learning intentions. Furthermore, children's understanding of teaching and learning developed with their evolving theory of mind. Those who acquired false belief understanding also had a grasp of more sophisticated teaching and learning concepts such as imitation, embedded teaching, and discovery learning. In contrast, younger children who were yet to understand the independent nature of reality and mental representation failed to recognize that the teaching intention embedded in play, and treat pure discovery through coincidence as intentional learning. It seems that young children have mistaken 'learning by doing' as 'learning is doing.'

This line of research has several implications for early education. Firstly, it brought the precursors of metacognition during early childhood to light. Theory of mind enables children to think reflectively on their learning. Along with the developing executive functioning and self-regulation around the same time in life, which is discussed later in this article, these developmental achievements prepare children to become conscious learners with the capacity to actively monitor and regulate their own learning. Secondly, this line of research echoes Vygotsky's idea of social learning. Children learn in a network of knowledge exchange. In addition to firsthand intervention, social interactions between teachers and learners and among learners provide the much needed intentional teaching and learning such as demonstration, instruction, and guided discovery. Last but not least, young children's confusion between doing and learning suggests that effective teaching and learning have to be effortful, with sustained attention control, and working memory resources. Children can be busy all the time with hands-on activities and experiments in classroom. But only when children are mentally engaged in the tasks will they actually start to learn. In order to mentally engage children, teachers need to make the teaching intention and the knowledge state change explicit for them to take full advantage of the teaching event (Frye and Wang, 2008).

Testimony

Sobel (2010) argued that probabilistic learning account is an integration of both bottom-up processes, such as extracting knowledge from observed data, and top-down processes, such as direct instruction. Harris (2002) calls the top-down learning processes as learning from 'testimony.' Children rely extensively on adults' testimony to learn, especially about entities that are not readily observable, such as metaphysical entities like God and tooth fairy, as well as scientific entities like germs and the shape of the Earth. The degree to which children commit to these entities varies according to the exposure they receive about them. More generally, Sobel suggested that children learn causal structures simply by being told. Nevertheless, children do not readily trust everything they are told; instead, they selectively trust particular informants' testimony from an early age. By observing the informants at work, children constantly access the informants' motive, reliability, competence, and knowledgeability when learning; they also compare testimony with their own observation and make sensible choices (e.g., Clément, 2010; Liu et al., 2013).

Testimony research highlights the role of direct instruction, which is compatible with the probabilistic learning account. Gopnik (2012) cautioned the use of direct instruction in teaching by emphasizing a balanced view of pedagogy. On the one hand, children know when they are being taught, which enables them to learn faster and more efficiently. On the other hand, explicit teaching limits the range of possible hypotheses for children to consider. A balanced view should consider direct instruction as one of the tools for teaching, yet at the same time provide examples, anomalies, encourage exploration and experimentation, and ask for explanations. Exploration and experimentation should be supported with appropriate scaffolding.

Conceptual Change

Another contribution of theory theory to child education is conceptual change (Vosniadou et al., 2008). Scientific theories develop in the context of paradigms (Kuhn, 1962), which refer to webs of shared concepts, beliefs, and practices, in the science community. The scientific theory change is not cumulative but revolutionary: old paradigms are rejected and replaced by new ones. The radical changes between theories make it difficult to establish rationally that one is better than another. Theory theory compares the process of children's cognitive development to that of theory development in the science community. Knowledge is acquired in domain-specific, theorylike structures. Conceptual change involves significant changes in concepts and theoretical frameworks. In the process of conceptual change, prior knowledge or misconceptions become stepping-stones toward more accurate knowledge and scientific explanations, although they may coexist with accurate ideas in the explanation and prediction. An individual's understanding can vary across contexts and domains, and there is not necessarily always linear improvement across age.

Conceptual change posits a relatively coherent body of domain-specific knowledge characterized by a distinct ontology and a causality that can give rise to explanation and prediction. Children have a rich set of ideas, conceptual frameworks, and reasoning skills. Organized into different domains, their naive theories explain and predict how things ought to behave, and solve problems. Conceptual change can be either bottom-up or top-down. Some misconceptions will self-correct without instruction as children gain more experience, such as one will not grow into a giant if eat more. These types of conceptual change are conservative and additive, largely unconscious. Yet some aspects of modern scientific knowledge are so counterintuitive that children are unlikely to achieve understanding without instruction, such as germs, particles, and the shape of the Earth. The instruction-induced conceptual changes are usually radical and deliberate. The process involves hypothesis testing, analogy, and metarepresentation, sometime thought experiments.

Rather than simple accumulation of information, proficiency in science is theory building based on prior knowledge and misconceptions. Factual knowledge must be placed in a conceptual framework to be well understood. Children need to learn how concepts are related to each other and their implications and applications. For spontaneous conceptual change, teachers need to provide an enriched environment for children's explorations and scaffold scientific thinking by positing task-appropriate questions at the right moment. For instruction-induced conceptual change, it is important to confront children with conflicts in their framework. Teachers need to focus on a few key concepts and core ideas and explain causality.

Domain-General Developmental Mechanisms for Learning

Self-Regulation and Executive Function

If theory theory is a domain-specific learning mechanism, self-regulation is rather a broad construct that lays the foundation for development across all domains. Self-regulation is an integrated, multidimensional construct that incorporates a broad range of psychological processes including biological processes, attention, emotion, behavior, and cognition. As a goal-directed behavior over time and in varying contexts, self-regulation generally entails three aspects: representing goals, motivation to achieve goals, and capacity to achieve goals despite of obstacles. Executive function is a construct subserving self-regulation that is discussed mainly by cognitive psychologists. It encompasses several distinctive cognitive constructs, including working memory, mental set shifting, and inhibitory control (Hofmann et al., 2012). While the measures for self-regulation can be both teacher or parent reports or behavior tasks such as delay of gratification, executive function measures are mainly behavioral tasks for children, including working memory tasks, stroop-type tasks for measuring inhibitory control, and set shifting tasks such as the Dimensional Change Card Sort (DCCS) task (Zelazo et al., 2003).

School Readiness and Success

While the focus of school readiness has traditionally been on academic preparedness, more recently it has been recognized as a multidimensional construct that includes cognitive and social emotional competences such as self-regulation (Blair, 2002) and theory of mind (Astington and Pelletier, 2005). The very components of self-regulation such as attention, inhibitory control, planning, abiding by rules, and working memory can explain why children's ability to self-regulate is strongly related to their performance in school. Virtually any effortful academic task requires attention control ability including focused attention, and suppression of interference or conflict from other, irrelevant information, which improves significantly during the early school years. Blair (2002) proposed that inefficient emotional regulation inhibits the use of higher order cognitive regulation, therefore affects academic achievement. Behavioral regulation supports the development of internally driven conduct norms like sitting quietly in the classroom, which children need in order to function independently in the school. Cognitive regulation measured by executive function tasks not only enables children to hold multiple stimuli and/or perspectives in mind and switch between two incompatible classification rules flexibly, it also moderately correlates and predicts later theory of mind development (Devine and Hughes, in press) that in turn affects children's learning.

Both correlational and longitudinal data have demonstrated the connection between self-regulation and academic success. A recent meta-analysis (Allan et al., 2013) found a modest overall relation between self-regulation and the academic skills of preschool and kindergarten children. Consistent with the distinction between the cognitive-laden cool self-regulation and the affect-laden hot self-regulation, academic skills were found to be more associated with measures of cool self-regulation than with those of hot self-regulation. Both behavior tasks and teacher reports were related to academic skills. Furthermore, self-regulation in general was more strongly associated with early math skills than with early literacy skills. Longitudinal data showed that self-regulation abilities strongly predicted later success in school and life. Self-regulation measured in preschool predicted children's academic success in the early primary years, even when variations in their intelligence or family backgrounds were controlled (e.g., Blair and Razza, 2007).

Class activities that demand higher levels of executive control might mask young children's mastery of the academic skills that the activities aim to assess. Evidence from adult literature suggests that 'depleted' self-control subsequently lowered one's cognitive function (Baumeister et al., 2007). Knowing this challenge children face in learning, teachers ought to consider the general cognitive demands of a lesson in addition to its content-specific requirements in lesson planning. There is also evidence that attention control skill is trainable for children as young as 4–6 years of age using computer games (Rueda et al., 2005) that might suggest an alternative pathway for academic enhancement of general executive function training.

Pretense

The cognitive flexibility required for self-regulation and executive function is readily available in children's pretend play. Assuming a pretend character in play helps children to take on another perspective. Even more so, when the child steps out

from the play frame to plan, coordinate, and negotiate roles and play scenarios, she attends to other players' perspective while assuming her own identity. In the process of back and forth shifting of perspectives, children practice executive control skills related to inhibiting the reality, holding multiple perspectives at the same time, and switching from one perspective to another flexibly. Carlson and associates (cf Carlson and White, 2013) found children's pretend play was indeed related to executive function measured by DCCS, controlling for working memory and verbal ability.

Carlson and White (2013) brought attention to the psychological distancing effect of symbolic thought. They argued that substituting direct experiences with symbols distances one from the sought after 'hot' reward in the delay of gratification tasks, which allows behavior to be controlled in light of the symbol rather than the stimulus itself. In the Less is More task, children need to indicate that they want the tray with less treats to actually get the tray with more treats. Young children found it rather difficult to resist the temptation of wanting the tray with more treats. When symbols were used to represent the real treats, however, young children's performance significantly improved. Attention control strategies such as the generalization from the symbolic representation to the real stimuli help children to manage the delay more efficiently.

Pretend play and symbolic representation and reasoning carry a lot of weight in early childhood education. The idea that these intuitive teaching practices actually support higher order executive function is reassuring and encouraging. A study by Bodrova and Leong (2007) had found that teacher-supported socio-dramatic play was an effective strategy for developing preschool children's self-regulation. More research is needed to fully understand the link between good teaching practices and children's cognitive development.

Pretense is believed to facilitate theory of mind development due to its metarepresentational nature (Leslie, 1987). Rich in counterfactual reasoning, it is also essential to probabilistic model learning (Walker and Gopnik, 2013). However, a recent comprehensive review (Lillard et al., 2013) found that the empirical evidence on the impact of pretend play on children's cognitive and social development was inconclusive. More rigorous research is needed to fully understand how pretend play facilitates learning. Meanwhile, play-based learning, according to Lillard et al., is still the most positive means yet known for child development.

How Do Children Learn?

So far the article has been focusing on the developmental mechanisms for learning. How about the learning process itself? How do children really learn? According to Siegler (2000), research on children's learning has reemerged after the decline of associationism focusing on learning of meaningful concepts and skills that are important in children's lives. One of the leading theories on children's learning is the overlapping waves theory (Siegler, 1996, 2000, 2005) that centers on the variety of strategies and ways of thinking in learning and problem solving. According to this theory, multiple approaches coexist simultaneously, with new strategies being added and old strategies cease to function sometimes, and certain single strategy could exist for a prolonged period of time, even after more advanced strategies emerge. It emphasizes cognitive diversity in learning, both with individuals and across them, even within an individual solving the same problem in different trials or within the same trial. A schematic depiction of the overlapping waves model (Figure 1) outlines the diversity of learning strategies in terms of the levels of sophistication, the timing, and the complex relationships among the strategies.

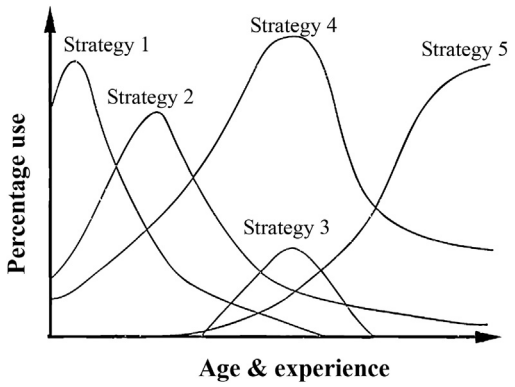

Figure 1 Schematic depiction of the overlapping waves model. Adapted from Emerging Minds: The process of change in children's thinking (Figure 4.4: p. 89), by Siegler, R.S., 1996. Emerging Minds: The Process of Change in Children's Thinking. Oxford University press, New York, Oxford. With permission from Oxford University Press, USA.

To study the detailed learning process with ecological validly, microgenetic methods of learning are adopted in the overlapping waves theory. These methods make dense observations in relation to the rate of change over a period of rapidly changing competence and intense analysis of these observations. The richness of the microgenetic data allows the overlapping waves theory to analyze learning along five dimensions, the path, rate, breadth, source, and variability of learning.

One of the major contributions of the overlapping waves theory is that it marries learning and development. Siegler argues that the learning and development processes have a lot in common in terms of variability, choice, and change, therefore cannot be distinguished from one and another. Some of the learning analyses echo the age-related changes described by developmentalists like Piaget. Siegler calls his theory as a developmental approach to learning that integrates both qualitative and quantitative changes. The educational implication of microgenetic analysis of learning is that it provides evidence for how teaching instruction unfolds children's learning therefore is helpful for designing more effective future instructions.

Conclusions

The learning process is inseparable from child development. Learning is a process of hypothesis testing based on probabilistic models. Children learn through play, observation, and through explicit teaching, during which children are constantly

testing and adjusting their hypotheses based on data collected from spontaneous experimentation, observation of other people's intervention, as well as knowledge and intentional attribution in teaching. The domain-general cognitive abilities of self-regulation and executive function help children succeed in school by setting learning goals, focusing their attention on the learning tasks at hand, resisting temptations, and regulating their emotions. Children's learning strategies show great variability, both within and across individuals. Teaching and learning are goal-directed, mindful, and effortful enterprises. Education should provide children with opportunities of supported exploration and experimentation, intentional demonstration and instruction, as well as guided discovery.

See also: Bayesian Statistics; Cognitive Development, Educational Theories of; Cognitive Development: Mathematics Learning and Instruction; Executive Functions During Childhood, Development of; Neo-Piagetian Theories of Cognitive Development; Piaget's Theory of Cognitive Development; Pretend Play and Cognitive Development; Self-Regulation During Early Childhood Across Cultures, Development of; Theory of Mind During Infancy and Early Childhood Across Cultures, Development of; Vygotsky's Theory of Human Development and New Approaches to Education.

Bibliography

Allan, N.P., Hume, L.E., Sims, D.M., Farrington, A.L., Lonigan, C.J., 2013 April. Relations between Self-regulation and the Development of Academic Skills in Preschool and Kindergarten: A Meta-analysis. Paper Presented at the Biennial Meeting of the Society for Research in Child Development, Seattle, WA.

Astington, J.W., Pelletier, J., 2005. Theory of mind, language, and learning in the early years: developmental originals of school readiness. In: Homer, B.D., Tamis-LeMonda, C. (Eds.), The Development of Social Cognition and Communication. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 205–230.

Baumeister, R.F., Schmeichel, B.J., DeWall, C.N., Vohs, K.D., 2007. Is the conscious self a help, a hindrance, or an irrelevance to the creative process. In: Columbus, A.M. (Ed.), Advances in Psychological Research, vol. 53. Nova Science Publishers, New York, pp. 137–152.

Blair, C., 2002. School readiness. American Psychologist 57 (2), 111–127.

Blair, C., Razza, R.P., 2007. Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. Child Development 78 (2), 647–663.

Bodrova, E., Leong, D.J., 2007. Tools of the Mind: The Vygotskian Approach to Early Childhood Education, second ed. Prentice-Hall, Upper Saddle River, NJ.

Carlson, S.M., White, R.E., 2013. Executive function, pretend play, and imagination. In: Taylor, M. (Ed.), The Oxford Handbook of the Development of Imagination. Oxford University Press, New York, pp. 161–174.

Clément, F., 2010. To trust or not to trust? Children's social epistemology. Review of Philosophy and Psychology 1 (4), 531–549.

Devine, R., Hughes, C., in press. Theory of Mind and Executive Function: Meta-analytic and Longitudinal Findings. Paper Presented at the Biennial Meeting of the Society for Research in Child Development.

Frye, D., Wang, Z., 2008. Theory of mind, understanding teaching, and early childhood education. In: Thurman, S.K., Fiorello, C.A. (Eds.), Applied Cognitive Research in K-3 Classrooms. Routledge, New York, pp. 85–108.

Frye, D., Ziv, M., 2005. Teaching and learning as intentional activities. In: Homer, B.D., Tamis-LeMonda, C. (Eds.), The Development of Social Cognition and Communication. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 205–230.

Gopnik, A., 2012. Scientific thinking in young children: theoretical advances, empirical research, and policy implications. Science 337 (6102), 1623–1627.

Gopnik, A., Wellman, H.M., 2012. Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. Psychological Bulletin 138 (6), 1085–1108.

Gopnik, A., Meltzoff, A.N., 1997. Words, Thoughts, and Theories. MIT Press, Cambridge.

Harris, P.L., 2002. What do children learn from testimony? In: Carruthers, P., Siegal, M., Stich, S. (Eds.), The Cognitive Basis of Science. Cambridge University Press, New York, pp. 316–334.

Hofmann, W., Schmeichel, B.J., Baddeley, A.D., 2012. Executive functions and self-regulation. Trends in Cognitive Sciences 16 (3), 174–180.

Koenig, M.A., Clément, F., Harris, P.L., 2004. Trust in testimony: children's use of true and false statements. Psychological Science 15 (10), 694–698.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.

Leslie, A.M., 1987. Pretense and representation: the origins of "theory of mind". Psychological Review 94 (4), 412–426.

Lillard, A.S., Lerner, M.D., Hopkins, E.J., Dore, R.A., Smith, E.D., Palmquist, C.M., 2013. The impact of pretend play on children's development: a review of the evidence. Psychological Bulletin 139, 1–34. http://dx.doi.org/10.1037/a0029321.

Liu, D., Vanderbilt, K.E., Heyman, G.D., 2013. Selective trust: children's use of intention and outcome of past testimony. Developmental Psychology 49 (3), 439–445.

Rueda, M.R., Rothbart, M.K., McCandliss, B.D., Saccamanno, L., Posner, M.I., 2005. Training, maturation and genetic influences on the development of executive attention. Proceedings of the National Academy of Sciences of the USA 102, 14931–14936.

Siegler, R.S., 1996. Emerging Minds: The Process of Change in Children's Thinking. Oxford University press, New York.

Siegler, R.S., 2000. The rebirth of children's thinking. Child Development 71 (1), 26–35.

Siegler, R.S., 2005. Children's learning. American Psychologist 60 (8), 769–778.

Sobel, D.M., 2010. Integrating top-down and bottom-up approaches to children's causal inference. In: Johnson, S.P. (Ed.), Neoconstructivism: The New Sciences of Cognitive Development. Oxford University Press, Oxford, pp. 159–179.

Vosniadou, S., Vamvakoussi, X., Skopeliti, I., 2008. The framework theory approach to the problem of conceptual change. In: Vosniadou, S. (Ed.), International Handbook of Research on Conceptual Change. Routledge, New York, pp. 3–34.

Walker, C.M., Gopnik, A., 2013. Pretense and possibility – a theoretical proposal about the effects of pretend play on development: comment on Lillard et al. (2013). Psychological Bulletin 139 (1), 40–44.

Wang, Z., 2010. Mindful Learning: Children's Developing Theory of Mind and their Understanding of the Concept of Learning (Doctoral dissertation). University of Pennsylvania. http://repository.upenn.edu/edissertations/157.

Wellman, H.M., Lagattuta, K.H., 2004. Theory of mind for learning and teaching: the nature and role of explanation. Cognitive Development 19, 479–497.

Zelazo, P.D., Müller, U., Frye, D., Marcovitch, S., 2003. The development of executive function in early childhood. Monographs of the Society for Research in Child Development 68 (3) i-151.

Cognitive Development During Infancy and Early Childhood across Cultures

Olivier Houdé, Paris-Descartes University, Sorbonne-Paris-City Alliance for Higher Education and Research, CNRS, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Jean Piaget, the great pioneer of this field in the twentieth century, described children as active learners who, through interactions with their environments, construct a complex understanding of the world, from the sensorimotor stage in infancy to the logical and formal operations stage in adolescence. However, Piaget underestimated the cognitive capabilities of infants, preschoolers, and elementary schoolchildren and overestimated the capabilities of adolescents and adults, which are often biased by illogical intuitions and overlearned strategies. We highlight several discoveries in the field from Piaget to the new theoretical ideas and empirical research reported over the last three decades. The crucial question is to understand why, despite rich precocious knowledge about physical and mathematical principles observed in infants and young children (from object principle to pure reasoning on abstract principles), older children, adolescents, and adults often have poor reasoning ability. We state that inhibition of less sophisticated solutions (or heuristics) by the prefrontal cortex is a critical component of children's conceptual insights associated with more advanced Piagetian stages. We conclude the article with the issue of cultural variations and suggestions for educational policy.

The scientific study of cognitive development in young children traces its roots back to Jean Piaget, a pioneer of this field in the twentieth century (Piaget, 1954, 1983). Piaget described children as active learners who, through numerous interactions with their environments, construct a complex understanding of the physical world around them. From infancy to adolescence, children progress through four psychological stages: (1) the sensorimotor stage from birth to 2 years (when cognitive functioning is based primarily on biological reactions, motor skills, and perceptions); (2) the preoperational stage from 2 to 7 years (when symbolic thought and language become prevalent, but reasoning is illogical by adult standards); (3) the concrete operations stage from 7 to 12 years (when logical reasoning abilities emerge but are limited to concrete objects and events); and (4) the formal operations stage at approximately 12 years (when thinking about abstract, hypothetical, and contrary-to-fact ideas becomes possible). Like the logician or mathematician, the child 'models' objects, their properties, and their relations through a succession of cognitive frameworks, from primary biological reactions and motor skills to high-order formal thinking. According to Piaget, after the age of 12, children model a formal hypothetico-deductive logic that resembles the rational logic of scientists and mathematicians. Piaget was the first psychologist to take children's thinking seriously. His genius was based on the idea of building his child development theory on triple roots in epistemological, biological, and logicomathematical foundations. Consequently, Piaget is now recognized as one of the precursors of cognitive science during the last century (Fischer and Kaplan, 2003).

Piaget underestimated the cognitive capabilities of infants, preschoolers, and elementary schoolchildren, and he overestimated the capabilities of adolescents and adults, which are often biased by illogical intuitions and overlearned strategies (or heuristics) they fail to inhibit (Houdé, 2000; Kahneman, 2011). During the last three decades, detailed behavioral studies of children's problem solving led to a reconceptualization of cognitive development, from discrete Piagetian stages to one that is analogous to overlapping waves (Siegler, 1996, 1999). The latter is consistent with a neo-Piagetian approach of cognitive development, in which more and less sophisticated solutions compete for expression in the human brain. In this approach, inhibition of less sophisticated solutions by the prefrontal cortex is a critical component of children's conceptual insights associated with more advanced Piagetian stages (Houdé et al., 2000, 2011; Poirel et al., 2012). At any point in time, children and adults potentially have available to them heuristics (i.e., intuitions) and logicomathematical algorithms, or as Kahneman described, multiple levels of 'thinking fast and slow' (Kahneman, 2011). Heuristics are rapid, often global or holistic, useful strategies in many situations, *but sometimes they are misleading*, whereas algorithms are slow, demanding, and analytical strategies that *necessarily lead to a correct (i.e., logical) solution in every situation*. In general, children and adults prefer using fast heuristics spontaneously, but that choice does not indicate that they are illogical per se (Houdé, 2000) or that they are 'happy fools' (De Neys et al., 2013). Psychologists had to be careful to avoid false negatives (Gelman, 1997), which is a strong tendency to say that those children or adults who fail a task are incompetent in the target domain of knowledge. A 'presumption of rationality' is sometimes the best assessment.

Contrary to Piaget's theory, infants learn more about the outside world through information that is captured by their perceptual systems than through motor skills development (Baillargeon, 1995; Mandler, 1988; Spelke, 2000). Infant cognition studies evaluated both the capacity to interpret sensory data and the faculty for understanding and reasoning about complex events. In the last 10 years, theoretical ideas and empirical research in the field have demonstrated that very young children's learning and thinking mechanisms do remarkably resemble the basic inductive processes of science, i.e., probabilistic models and Bayesian learning methods (Gopnik, 2012). Infants can implicitly reason statistically (Téglás et al., 2011). From this point of view, the very young child is already seen as a 'scientist in the crib' (Gopnik et al., 1999).

We highlight several discoveries in the field, from Piaget to the new theoretical ideas and empirical research. We will conclude the article with the issue of cultural variations and suggestions for educational policy.

Beyond Piaget's Description of Infant Cognition

The most important advances were associated with a change in methodology, made possible by the use of videotapes and computers. Instead of analyzing infant action skills (or 'schemes'), as advocated by Piaget, the infant's gaze behavior was examined. Experiments using the visual habituation technique and recordings of visual fixation time were conducted to study infant reactions to novelty and the detection of unexpected (or 'impossible') events. These experiments led to the discovery of early abilities unknown to Piaget, such as object unity and object permanence by the age of 4 or 5 months (along with other physical principles or core knowledge: cohesion, contact, continuity, containment, and support; Baillargeon, 1995; Spelke, 2000; Spelke et al., 1995), the existence of numerical abilities by that same age (Wynn, 1992, 1998), and 'pure reasoning' through statistical expectations by the age of 12 months (Téglás et al., 2011). Other studies have described neonatal imitation of social partners during the first few days of life and cross-modal perception in the early months (Meltzoff and Moore, 1977; Meltzoff and Prinz, 2002). Moreover, new discoveries about early social cognition, which were underestimated by Piaget, showed that very young children can perform moral evaluation of others' actions: 6- and 10-month-old infants prefer an individual who helps others to one who hinders others, prefer a helping individual to a neutral individual, and prefer a neutral individual to a hindering individual (Hamlin et al., 2007). These findings constitute evidence that preverbal infants have the capability to assess individuals on the basis of their behavior toward others. This capability may serve as the foundation for moral thought and action. By the age of 12 months, infants know that only intentional agents can create physical order between objects (Newman et al., 2010). Finally, in the same vein, new discoveries demonstrated that preverbal 15-month-old infants can understand false beliefs in theory-of-mind tasks, which were adapted for the visual detection of unexpected psychological events (Onishi and Baillargeon, 2005). In the next section, we will focus on strictly cognitive development, from early object principle to pure reasoning on abstract principles.

From Early Object Principle to Pure Reasoning on Abstract Principles

Children as Physicists: Early Object Permanence

Piaget described children as active learners who, through numerous interactions with their environments, construct an increasingly complex understanding of the physical world around them. In infancy, the first step of this construction is object permanence (Piaget, 1954). We are indebted to Piaget for his clear statement of the problem of how objects act as the basic unit for constructing reality: Does an object continue to exist for infants once it is out of sight? Does it have permanence outside of the immediate experience? Six substages (within Stage 1: Sensorimotor) stand out from Piaget's observations; these substages delineate the laborious path that infants take toward object permanence, essentially during the first year of life.

According to Piaget, the systematic search for a disappeared object does not begin until Substage IV of sensorimotor development, at approximately 8 or 9 months. This substage still suffers from the location error, which can be demonstrated using the following device: the experimenter places the infant in front of two screens (A and B) that are equally easy to reach, and he hides an object under screen A. The 8- to 9-month-old has no trouble finding the object. After a few repetitions of this situation, the object is conspicuously transferred from under screen A to under screen B. If there is a delay before the infant can respond, he continues to look for the object under screen A: this is the so-called A-not-B error. The Piagetian interpretation is that infants have not yet mastered the true object permanence principle. It is not until the end of the first year that this principle is acquired, and hence the disappearance of the A-not-B error at about 12 months (with a few remaining refinements during the second year). This interpretation is contested today in light of new data on the existence of object permanence by the early age of 4 or 5 months. The new data came with a change of methodology from the Piagetian study of the infant's action skills to the post-Piagetian study of the infant's looking behavior, i.e., analysis of perceptual activity measured by visual fixation time.

With this new approach, Baillargeon and colleagues (Baillargeon, 1987, 1995; Baillargeon et al., 1985) provided evidence of the early existence of object permanence using the so-called impossible-event paradigm (or violation-of-expectation paradigm). This famous experimental procedure runs in two phases: (1) The infant is facing a presentation device with a screen fastened to its base by a hinge in such a way that the screen can be rotated 180° along the horizontal axis. The infant is habituated to the alternating backward–forward rotation. (2) A box is placed behind the screen's rotation axis in such a way that the screen can now only be rotated 112°, at which point it completely hides the box (the disappeared object). The infant is then shown a possible event, a 112° rotation of the screen, and an impossible event, a 180° rotation. In Phase 2, 4- and 5-month-old infants look longer at the impossible event than at the possible one, although the impossible event is identical to the habituation situation (180° rotation) and the possible event is new (identical results have been found in infants at three and a half months of age, but not in all infants). Baillargeon concluded that 4- and 5-month-old infants realize that the event is impossible, which suggests that they can conceive of the violated property, namely, the permanence of the box hidden by the screen. By this young age, the concept of object permanence may already be in place, along with the cohesion principle (experiments on this problem were conducted by Baillargeon and other investigators: Baillargeon, 1995; Baillargeon and DeVos, 1991). Moreover, Wynn (1992, 1998) showed that in protonumerical activities, a permanent object (one that is completely out of sight) can be integrated by 4- or 5-month-old infants as a physical entity that is connected to other discrete and quantifiable entities.

The divergence between the experimental data just presented and Piaget's theory, closely tied to the paradigms used, is problematic for infant psychologists. If by 4 or 5 months of age, there is the concept of object permanence in infants, how can one explain the A-not-B error classically found in 8- to 12-month-old children? For some authors, this question is not a critical one, whereas for others, it is one of the fascinating puzzles of the psychology of cognitive development (Houdé, 2000). A range of interpretations have been proposed, along with ingenious experimental situations, e.g., the role of memory capacity, spatial organization, and motor programming constraints (motor perseveration and inhibition). One of the most impressive interpretations was proposed by Diamond (1991, 1998); Diamond and Goldman-Rakic (1989) in their conclusion of a comparative study in neurocognitive animal psychology (conducted on monkeys after prefrontal cortex ablation). Diamond and Goldman-Rakic suggested that the location error made by 8- to 12-month-old infants results from ineffective inhibition of a prepotent gesture (moving toward A) brought about by the insufficient development of the prefrontal cortex. Single-cell recordings in monkeys (Watanabe, 1996) and the neural models devised by Dehaene and Changeux (1989) corroborated the hypothesis by Diamond and Goldman-Rakic that the frontal system plays a decisive role. Complementary studies by Bell and Fox (1992) established a relationship between the electroencephalographic recordings of the frontal system and the cognitive development of infants, in particular, their performance in the A-B situation. The recordings confirmed the existence, in infants, of the link between the A-not-B error and the frontal system established by Diamond and Goldman-Rakic for nonhuman primates. The authors concluded that success in the A-B situation is achieved through inhibition of the initial programmed gesture, in conjunction with a sufficient memory capacity and the ability to resist distraction during the delay (several experimental findings, however, argued against the hypothesized role of frontal inhibition, and thus of the perseveration of a to-be-inhibited gesture, as the sole explanatory factor; for meta-analyses, see Marcovitch and Zelazo, 1999; Wellman et al., 1987).

Children as Mathematicians: Early Numerical Abilities

A heated debate topic in psychology is how children come to understand numbers? Piaget's answer was that number is constructed in children through the logicomathematical synthesis of classification and seriation operations. Number borrows its inclusion structure from classes (1 is included in 2, 2 in 3, etc.); because it disregards qualities by transforming objects into units, it brings a serial order into play, the sole means of distinguishing one unit from the next: 1 then 1, then 1, etc. The serial ordering of units is combined with the inclusion of the sets that result from their union (1 is included in 1 + 1, 1 + 1 is included in 1 + 1 + 1, etc.) to constitute number. The task Piaget used was the conservation of number. When children are shown two rows of objects that contain an equal number of objects but that differ in length (because the objects in one of the rows have been spread apart), young children think the longer row has more objects. Piaget's interpretation was that preschool children are still fundamentally intuitive (their reasoning is illogical by adult standards), or as he called them, 'preoperational' (Stage 2), and hence limited to a perceptual way of processing information (based on length or, in certain cases, on density). At the age of 6 or 7 years, children understand the equivalency of quantities, regardless of apparent transformations. At this point, they are called 'operational' or 'conserving' (Stage 3), the criterion for mastery of number. Piaget also worked on determining whether the conservation of number develops simultaneously with inclusion (classification) and order relations (seriation).

After this founding work on the genesis of number, research in this domain proliferated, and criticisms of Piaget's theory were far from scarce. First, the synchronous development of classification, seriation, and conservation was not validated in experimental verifications. Second, it became clear that Piaget's view of the logicostructural aspect of number is overly polarized and overshadows the more functional aspects of numerical development, such as counting.

A radical change in perspective began with Gelman (Gelman and Meck, 1983; Gelman et al., 1986), who not only turned the attention toward counting but also postulated the early existence of five fundamental principles of counting: stable order (order of the number words), strict one-to-one correspondence (between the number words and the items counted), cardinality (the number word corresponding to the last item counted is equal to the total number of items), abstraction (any kind of item can be counted), and order irrelevance (items can be counted in any order). Gelman demonstrated the presence of these principles in young children by having them say whether they thought a doll was counting correctly or incorrectly. Knowledge or lack of knowledge of a given principle was deduced from whether the child detected the corresponding type of counting error (unstable order, violation of the one-to-one correspondence, cardinal number referred to by an ordinal word number, etc.). The results indicated that 3-year-old children have already acquired the basic principles of counting. This led Gelman to distinguish three components in the ability to count: a conceptual component ('knowing why' or understanding the five principles), a procedural component ('knowing how' or understanding the structure and order of counting), and a utilization component ('knowing when' or understanding the relevance of using the first two components in a given context). Defending the principles-before-skills hypothesis, Gelman suggested that the numerical difficulties of preschool children lie essentially in the procedural and utilization components. Another of Gelman's original contributions was her use of the so-called magic task to demonstrate that 3- to 4-year-old children are surprised by transformations that affect the cardinal number of a set (adding and subtracting items) but not by transformations that do not (spreading and grouping) (Gelman, 1972). She concluded that despite their failure in Piaget's conservation of number task, the children at this age are already capable of seeing through irrelevant transformations and treating the number of items as invariable (for a seminal study on this point, see Mehler and Bever, 1967). This new conclusion was corroborated by the discovery of the perception of numerical invariance in neonates (Antell and Keating, 1983) and in 5- and 8-month-old infants (Loosbroek and Smitsman, 1990; Lipton and Spelke, 2003).

The most striking example of infants as 'mathematicians' is found in the famous work by Wynn (1992). Wynn recorded the looking time of 4- and 5-month-old infants in the 'impossible-event' procedure (or violation-of-expectation procedure) and demonstrated that infants were surprised by (looked longer at) impossible numerical events (e.g., $1 + 1 = 1$ and $1 + 1 = 3$, or $2 - 1 = 2$) but were not surprised at the corresponding possible events ($1 + 1 = 2$ and $2 - 1 = 1$) (the events were staged with Mickey Mouse figures). She concluded that infants are endowed with a mechanism that calculates the exact outcome of simple arithmetic operations. She claimed that infants at this age are already able to encode ordinal information and possess genuine numerical concepts that cannot be reduced to holistic percepts derived from a pattern recognition process (for a brain imaging confirmation of Wynn's results, see Berger et al., 2006). Like Gelman's stance, Wynn's position is strong and seems to run counter to what we know about the numerical difficulties of preschool children. Wynn's empirical results are robust and consistent (Wynn, 2000), but they have sparked theoretical debates (Simon, 1997, 1998). The task of the following research was to devise a developmental model of logicomathematical operations (conservation, counting, and elementary arithmetic) that accounts for both early abilities (Gelman, Wynn, etc.) and late inabilities (Piaget), without denying the reality of the former but raising the question of the factors that explain the latter (see Section What Develops? Heuristics, Algorithms, Cognitive Control, and Inhibition of Misleading Strategies).

Children as Bayesian Statisticians: Early Reasoning Abilities

All of the discoveries summarized before and for most of them achieved during the 1980s and 1990s are key examples that illustrate the post-Piagetian revolution in our understanding of early cognitive development. A new trend in the field is to see children as Bayesian statisticians, i.e., thinkers who already infer abstract hierarchical structure from data. According to Gopnik (2012), if children learn in this way, they could drastically revise their initial representations of the world on the basis of their experience, as scientists do. In this framework, researchers have recently shown that infants and young children not only detect statistical patterns but also use those patterns to test causal hypotheses about objects and people. For example, using a 'looking-time' technique, Xu and Garcia (2008) showed that 8-month-old infants were sensitive to statistical patterns in a ping-pong ball experiment. In this study, the experimenter showed the infants a box full of white and red balls. She closed her eyes and randomly took several balls from the box and put them in a small bin. If the sample was random, then the distribution of balls in the bin should match the distribution of the balls in the box. The infants saw a sample that either matched (i.e., expected event) or did not match (unexpected event) the probabilistic distribution, and they looked longer at the nonmatching sample. In a control condition, the infants saw the same sequence of events, but the experimenter took the balls out of her pocket rather than from the box, and the looking-time difference disappeared.

In the same vein, a subsequent study by Xu and colleagues has demonstrated that children under 2 years of age could infer an underlying mental state, i.e., a 'preference' from a statistical pattern (Kushnir et al., 2010). Another group of researchers illustrated a type of 'pure reasoning' in 12-month-old infants (Téglás et al., 2011). Using complex displays of multiple moving objects, this elegant study showed that infants' probabilistic reasoning is surprisingly rich, powerful, and coherent: infants form time-varying expectations about future events that are a systematic and rational function of several stimulus variables, such as the number of objects, their physical arrangement, and the duration of their occlusion. Infants' looking times are consistent with a Bayesian ideal observer embodying abstract principles of object motion. The power and originality of Bayesian inference are to consider both new evidence and the prior probability of hypotheses through a generative mathematical model, such as a map or tree structure or a graphical network. This gives Bayesian reasoning a combination of stability and flexibility for learning several patterns of evidence rather than others, such as a particular set of visual images, or spoken sentences, or statistical contingencies between events.

Pure and abstract reasoning starts earlier than Piaget had predicted in his stages theory (Stage 4 for formal reasoning operations in adolescents). Hence, all of these new theoretical ideas and empirical research on infancy and early childhood refine one old and simple question (see Karmiloff-Smith, 1992, for a discussion of Fodor's nativism opposed to Piaget's constructivism): Finally, *what develops* after and beyond this rich first set of cognitive abilities?

What Develops? Heuristics, Algorithms, Cognitive Control, and Inhibition of Misleading Strategies

The crucial question now is to understand why, despite these precocious abilities about physical and mathematical principles, older children, adolescents, and adults so often have poor reasoning ability. How might we explain the systematic errors observed in humans at all stages of their cognitive development by Piaget (1954, 1983) in children and by post-Piagetian experimental psychologists of reasoning in adults (Evans, 1989, 2003; Kahneman, 2011; Kahneman et al., 1982; Kahneman and Tversky, 2000)?

How might we explain the famous number-conservation error observed in children until the age of 7 by Piaget and, after him, by all developmental psychologists around the world? It is an intriguing question because we know today that very young children are already capable of treating the number of items as invariable through irrelevant transformations (Antell and Keating, 1983; Gelman, 1972; Loosbroek and Smitsman, 1990; Mehler and Bever, 1967; Lipton and Spelke, 2003) and that they possess other protonumerical skills (Gelman and Meck, 1983; Gelman et al., 1986; Wynn, 1992, 1998). One of the main current explanations is that children learn heuristics, which are often useful in a large set of situations, but fail to inhibit them when, contrary to general practice, they are misleading (Houdé, 2000). In the case of Piaget's number-conservation algorithm, the overlearned competing heuristic is 'length-equals-number' (Houdé and Guichart, 2001). This new theoretical approach is in line with Diamond's explanation of the A-not-B error in infants (see Section Children as Physicists: Early Object Permanence) and assumes that cognitive development relies not only on the

acquisition of knowledge of incremental complexity (Piaget, 1983) but also on the ability to inhibit previously acquired knowledge (Bjorklund and Harnishfeger, 1990; Dempster and Brainerd, 1995; Diamond, 1991, 1998; Harnishfeger, 1995; Houdé, 2000). Increasing evidence shows that the ability to inhibit previous knowledge is critical for developmental milestones, such as those defined by Piaget's theory (Houdé, 2000). Inhibitory control of misleading strategies, an executive function performed by the prefrontal cortex, has been claimed necessary for acquisition and use of motor or cognitive algorithms in the fields of object permanence in infants (Bell and Fox, 1992; Diamond, 1991, 1998; Diamond and Goldman-Rakic, 1989), number conservation and class inclusion in preschool and schoolchildren (Borst et al., 2012, 2013; Houdé and Guichart, 2001; Perret et al., 2003), and logical reasoning in adolescents and adults (Houdé, 2007; Houdé and Tzourio-Mazoyer, 2003; Houdé et al., 2000).

One of the challenges of today's developmental research, in all domains of cognition ranging from motor programming to high-order logical reasoning, is to account not only for a general and incremental process of coordination-activation capacities of structural units, schemes, or skills through ages and stages (Piaget, 1983, and all the 1980s neo-Piagetians: see the review book by Demetriou, 1988) but also for a general process of selection-inhibition of competing strategies, i.e., heuristics (or intuitions) and logicomathematical algorithms, occurring with different weights at any point in time, depending on the context, in a nonlinear dynamical system of growth (Houdé, 2000; Siegler, 1996, 1999). Such cognitive model introduces less regular developmental curves containing perturbations, bursts, and collapses. O'Reilly (1998) described six principles for biologically based computational models of cognition, one of which is inhibitory competition (see also Johnson, 2010). Resolving this 'inhibition issue' is an important task for both developmental psychology and cognitive neuroscience. The most compelling magnetic resonance imaging (MRI) reports of structural changes with brain development during childhood and adolescence showed a sequence in which the higher-order association area, such as the prefrontal cortex sustaining inhibitory control, matures last (Casey et al., 2005). The sequence in which the cortex matures parallels the cognitive milestones in human development. First, the regions subserving primary functions, such as motor and sensory systems, mature the earliest; the temporal and parietal association cortices associated with basic language skills and spatial attention mature next; and the last to mature are the prefrontal cortex and its inhibitory-control ability.

Using fMRI (functional MRI), from this theoretical perspective, we reexamined what occurs in the developing brain when schoolchildren are tested for their performances in Piaget's number-conservation task. Remember that when children are shown two rows of objects that contain an equal number of objects but that differ in length (because the objects in one of the rows had been spread apart), young children think that the longer one has more objects. Piaget's interpretation was that preschool children are still fundamentally intuitive (their reasoning being illogical by adult standards), or as he called them, 'preoperational' (Stage 2), and hence limited to a perceptual way of processing information (here, based on length or, in certain cases, on density). When they are approximately 6 or 7 years old, children understand the equivalency of quantities, regardless of apparent transformations. At this point, they are called 'operational' or 'conserving' (Stage 3), the criterion for logicomathematical mastery of number. Our new hypothesis was that their main cognitive difficulty (beyond logicomathematical cognition per se) was to efficiently inhibit through their prefrontal cortex the overlearned 'length-equals-number' strategy, a heuristic that is often used both by children and adults in many school and everyday situations.

In a first fMRI study, we found that the cognitive change allowing children to access conservation (i.e., the shift from Stage 2 to Stage 3 in Piaget's theory) was related to the neural contribution of a bilateral parietofrontal network involved in numerical and executive functions (Houdé et al., 2011). These imaging results highlighted how the behavioral and cognitive stages that Piaget formulated during the twentieth century manifest in the brain with age. In a second fMRI study (Poirel et al., 2012), we demonstrated that the prefrontal activation (i.e., the blood-oxygen-level-dependent signal) observed when schoolchildren succeeded at the Piaget's number-conservation task was correlated to their behavioral performance on a Stroop-like measure of inhibitory function development (Wright et al., 2003). These new results in schoolchildren fit well with previous brain imaging data from our laboratory showing a key role of prefrontal inhibitory-control training when adolescents or adults (belonging to Stage 4 in Piaget's theory) spontaneously fail to block their perceptual intuitions (or bias, heuristics) to activate logicomathematical algorithms (i.e., deductive rules) in reasoning tasks (Houdé, 2007; Houdé et al., 2000).

If we have 'two minds in one brain' as stated by Evans (2003) or, in other words, two ways of thinking and reasoning, i.e., 'fast and slow' (Kahneman, 2011), currently called 'System 1' (intuitive system) and 'System 2' (analytic system), then the crucial challenge is to learn to inhibit the misleading heuristics from System 1 when the more analytic and effortful System 2 (logicomathematical algorithms) is the way to solve the problem (Borst et al., 2013; Houdé, 2000). Within this post-Piagetian theoretical approach, we can now understand why, despite rich precocious knowledge about physical and mathematical principles observed in infants and young children (from object principle to pure reasoning on abstract principles), older children, adolescents, and adults so often have poor reasoning. The cost of blocking our intuitions is high and depends on the late maturation of the prefrontal cortex. Moreover, this executive ability remains delicate throughout our lifetime, and adults may sometimes need 'prefrontal pedagogy' to learn inhibiting intuitive heuristics (or biases) in reasoning tasks (Houdé, 2007).

An innovative research question now is to better understand the cognitive roots of such powerful heuristics (intuitions and bias from System 1) that children and adults have so much difficulty inhibiting in some cases. New heuristics may appear and be overlearned at any time in the course of development (Houdé, 2000) because our brain is an irrepressible detector of regularities from its perceptual and cultural environment. For example, preschool children (more than infants) are often exposed, in 'math books' in the classroom or in everyday scenes, to patterns of objects in which number and length covary (e.g., the 1–10 Arabic numbering series is frequently

illustrated by increasing lines of drawn animals or fruits: one giraffe, two hippopotamus, three crocodiles, and so on), hence the overlearned and misleading 'length-equals-number' heuristic, which is overactivated in Piaget's conservation of number task. A new avenue of research would be to assess the role of early sensitivity to statistical patterns (i.e., probability of hypotheses) and Bayesian inference (Gopnik, 2012) in the psychological construction of perceptual, motor, and cognitive heuristics. Moreover, the power of Bayesian learning (see the previous section on this topic) might require, in some conflict situations, a strong antagonist process of inhibition for blocking heuristics when they are misleading.

Cultural Variations

In the past, there has been research investigating the universality of Piaget's stages, from infancy (Stage 1: the sensorimotor stage) to adolescence (Stage 4: the formal operations stage). According to Piaget, cultural factors should only affect the age at which the stages (which are qualitatively different from each other) are attained, but the sequence should be universal. Generally, there is stronger support for the universality of the less advanced stage (sensorimotor) than the more advanced stage (formal operations) (Dasen, 1977). This linear Piagetian approach to cultural variations precludes dynamic and nonlinear view of cognitive development.

Today, the question is reframed within a different and exciting research program on geometry and number (in line with the general 'Kantian research program' in the *Cognitive Sciences*: Dehaene and Brannon, 2010; Izard et al., 2011). In a famous paper, Dehaene and colleagues explored whether geometry constitutes a core set of intuitions present in all humans, regardless of their language or schooling (Dehaene et al., 2006). They used two nonverbal tasks to probe the conceptual underpinnings of geometry in the Mundurukú, an isolated Amazonian indigenous group (already studied by Pica et al., 2004; for arithmetical abilities). Mundurukú children and adults spontaneously used basic geometric concepts such as points, lines, parallelism, or right angles to detect intruders in simple pictures, and they used distance, angle, and sense relationships in geometrical maps to locate hidden objects. Dehaene's results provided evidence for geometrical intuitions in the absence of schooling, experience with graphic symbols or maps, or a rich language of geometrical terms.

In a subsequent study, the same authors showed distinct intuitions of the number scale in Western and Amazonian indigenous cultures (Dehaene et al., 2008). This study was rooted in previous developmental experiments that showed a remarkable shift in the child's conception of how numbers map onto space, from logarithmic to linear mappings, between the first and fourth grades, depending on experience and the range of numbers tested (Booth and Siegler, 2006; Siegler and Booth, 2004; Siegler and Opfer, 2003). However, a logarithmic strategy may remain dormant in all of us for very large numbers or whenever we approximate numbers (Banks and Coleman, 1981), including prices (Dehaene and Marques, 2002). Consequently, the log and linear scales may be deeply embedded in all of our mental activities. Moreover, the mapping of numbers onto space through the logarithmic strategy plays an essential role in mathematical sciences. In their cross-cultural study, Dehaene et al. (2008) tested whether this mapping is a cultural invention or a universal intuition shared by all humans, regardless of culture and education. They probed number-space mappings in the Mundurukú, the same Amazonian indigenous culture as in the previous study (a group with a reduced numerical lexicon and little or no formal education). At all ages, the Mundurukú mapped symbolic and nonsymbolic numbers onto a logarithmic scale, whereas the Western adults used linear mapping with small or symbolic numbers and logarithmic mapping when the numbers were presented nonsymbolically under conditions that discouraged counting. According to Dehaene, this demonstration indicated that the mapping of numbers onto space is a universal intuition and that this initial intuition of number is logarithmic. The concept of a linear number line appeared to be a cultural invention that failed to develop in the absence of formal education.

These new data on cultural variations illustrate, through a dynamical system of cognitive growth, how cultural experiences may alter the weight of strategies used by children or adults (see Geary, 1996; Lubin et al., 2006; Saxe, 1999; for previous cross-cultural or cross-linguistic studies on mathematical abilities; see also Dasen and Mishra, 2013; for cross-cultural studies on the development of geocentric spatial language). Future studies can explore systematically how cultural variations may modulate the competition between cognitive strategies in the brain, with the idea that heuristics (sometimes, algorithms) are culture-dependent strategies. From this point of view, educational policy should focus on interventions that can aid executive function, particularly inhibitory control, in children (Diamond et al., 2007; Diamond and Lee, 2011; Houdé, 2007). Prefrontal executive control is probably a general culture-independent potential shared by all humans in their genetic heritage (Thompson et al., 2001), even if its ontogenetic maturation is late (Casey et al., 2005), although the respective weights of cognitive strategies to inhibit or activate, case by case, might depend on context and culture.

See also: Breast-feeding: Effects on Cognitive and Neural Development; Bruner's Theory of Cognitive Development; Cognitive Development, Educational Theories of; Decision and Choice: Heuristics; Heuristics in Social Cognition; Heuristics, History of; Neo-Vygotskian Developmental Theory; Piaget's Theory of Cognitive Development; Pretend Play and Cognitive Development; Vygotsky's Theory of Cognitive Development.

Bibliography

Antell, S., Keating, D., 1983. Perception of numerical invariance in neonates. Child Development 54, 695–701.

Baillargeon, R., 1987. Object permanence in 3.5 and 4.5-month-old infants. Developmental Psychology 23, 655–664.

Baillargeon, R., 1995. Physical reasoning in infancy. In: Gazzaniga, M.S. (Ed.), The Cognitive Neurosciences. MIT Press, Cambridge, MA, pp. 181–204.

Baillargeon, R., DeVos, J., 1991. Object permanence in young infants. Child Development 62, 1227–1246.

Baillargeon, R., Wang, S., 2002. Event categorization in infancy. Trends in Cognitive Sciences 6, 85–92.

Baillargeon, R., Spelke, E., Wasserman, S., 1985. Object permanence in five-month-old infants. Cognition 20, 91–208.
Banks, W., Coleman, M., 1981. Two subjective scales of number. Perception and Psychophysics 29, 95–105.
Bell, M., Fox, N., 1992. The relations between frontal brain electrical activity and cognitive development during infancy. Child Development 63, 1142–1163.
Berger, A., Tzur, G., Posner, M., 2006. Infant brains detect arithmetic errors. Proceedings of the National Academy of Sciences 103, 12649–12653.
Bjorklund, D., Harnishfeger, K., 1990. The resources construct in cognitive development: diverse sources of evidence and a theory of inefficient inhibition. Developmental Review 10, 48–71.
Booth, J., Siegler, R., 2006. Developmental and individual differences in numerical estimation. Developmental Psychology 41, 189–201.
Borst, G., Poirel, N., Pineau, A., Cassotti, M., Houdé, O., 2012. Inhibitory control in number-conservation and class-inclusion tasks: a neo-Piagetian inter-tasks priming study. Cognitive Development 27, 283–298.
Borst, G., Pineau, A., Poirel, N., Cassotti, M., Houdé, O., 2013a. Inhibitory control efficiency in a Piaget-like class-inclusion task in school-age children and adults: a developmental negative priming study. Developmental Psychology 49, 1366–1374.
Borst, G., Moutier, S., Houdé, O., 2013b. Negative priming in logicomathematical reasoning: the cost of blocking your intuition. In: De Neys, W., Osman, M. (Eds.), New Approaches in Reasoning Research – Current Issues in Thinking & Reasoning. Psychology Press, New York, pp. 34–50.
Casey, B., Tottenham, N., Liston, C., Durston, S., 2005. Imaging the developing brain: what have we learned about cognitive development? Trends in Cognitive Sciences 9, 104–110.
Dasen, P., 1977. Piagetian Psychology: Cross-Cultural Contributions. Gardner, New York.
Dasen, P., Mishra, R., 2013. Development of Geocentric Spatial Language and Cognition: An Eco-cultural Perspective. Cambridge University Press, Cambridge.
De Neys, W., Rossi, S., Houdé, O., 2013. Bats, balls, and substitution sensitivity: cognitive misers are no happy fools. Psychonomic Bulletin & Review 20, 269–273.
Dehaene, S., Brannon, E., 2010. Space, time, and number: a Kantian research program. Trends in Cognitive Sciences 14, 517–519.
Dehaene, S., Changeux, J.-P., 1989. A simple model of prefrontal cortex function in delayed-response tasks. Journal of Cognitive Neuroscience 1, 244–261.
Dehaene, S., Marques, J.-F., 2002. Cognitive euroscience: scalar variability in price estimation and the cognitive consequences of switching to the euro. Quarterly Journal of Experimental Psychology 55, 705–731.
Dehaene, S., Izard, V., Pica, P., Spelke, E., 2006. Core knowledge of geometry in an Amazonian indigene group. Science 311, 381–384.
Dehaene, S., Izard, V., Spelke, E., Pica, P., 2008. Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. Science 320, 1217–1220.
Demetriou, A. (Ed.), 1988. The Neo-Piagetian Theories of Cognitive Development. North-Holland, Amsterdam.
Dempster, F., Brainerd, C. (Eds.), 1995. Interference and Inhibition in Cognition. Academic Press, New York.
Diamond, A., 1991. Neuropsychological insights into the meaning of object concept development. In: Carey, S., Gelman, R. (Eds.), The Epigenesis of Mind: Essays on Biology and Cognition. Lawrence Erlbaum, Hillsdale, NJ, pp. 67–110.
Diamond, A., 1998. Understanding the A-not-B error: working memory vs. reinforced response, or active trace vs. latent trace. Developmental Science 1, 185–189.
Diamond, A., Goldman-Rakic, P., 1989. Comparison of human infants and rhesus monkeys on Piaget's AB task. Experimental Brain Research 74, 24–40.
Diamond, A., Lee, K., 2011. Interventions shown to aid executive function development in children 4 to 12 years old. Science 333, 959–964.
Diamond, A., Barnett, W., Thomas, J., Munro, S., 2007. Preschool program improves cognitive control. Science 318, 1387–1388.
Evans, J., 1989. Bias in Human Reasoning. Erlbaum, London.
Evans, J., 2003. In two minds: dual-process accounts of reasoning. Trends in Cognitive Sciences 7, 454–459.
Fischer, K., Kaplan, U., 2003. Jean Piaget. In: Nadel, L. (Ed.), The Encyclopedia of Cognitive Science, vol. 1. Nature Publishing Group, Macmillan, London, pp. 679–682.
Geary, D., 1996. International differences in mathematical achievement: their nature, causes, and consequences. Current Directions in Psychological Sciences 5, 133–137.
Gelman, R., 1972. Logical capacity of very young children. Child Development 43, 75–90.
Gelman, R., 1997. Constructing and using conceptual competence. Cognitive Development 12, 305–313.
Gelman, R., Meck, E., 1983. Preschooler's counting. Cognition 13, 343–359.
Gelman, R., Meck, E., Merkin, S., 1986. Young children's numerical competence. Cognitive Development 1, 1–29.
Gopnik, A., 2012. Scientific thinking in young children: theoretical advances, empirical research, and policy education. Science 337, 1623–1627.
Gopnik, A., Meltzoff, A., Kuhl, P., 1999. The Scientist in the Crib. William Morrow and Cie, New York.
Hamlin, J., Wynn, K., Bloom, P., 2007. Social evaluation by preverbal infants. Nature 450, 557–560.
Harnishfeger, K., 1995. The development of cognitive inhibition: theories, definition, and research evidence. In: Dempster, F., Brainerd, C. (Eds.), Interference and Inhibition in Cognition. Academic Press, New York, pp. 176–204.
Houdé, O., 2000. Inhibition and cognitive development: object, number, categorization, and reasoning. Cognitive Development 15, 63–73.
Houdé, O., 2007. First insights on neuropedagogy of reasoning. Thinking & Reasoning 13, 81–89.
Houdé, O., Guichart, E., 2001. Negative priming effect after inhibition of number/length interference in a Piaget-like task. Developmental Science 4, 71–74.
Houdé, O., Tzourio-Mazoyer, N., 2003. Neural foundations of logical and mathematical cognition. Nature Reviews Neuroscience 4, 507–514.
Houdé, O., Zago, L., Mellet, E., Moutier, S., Pineau, A., Mazoyer, B., Tzourio-Mazoyer, N., 2000. Shifting from the perceptual brain to the logical brain: the neural impact of cognitive inhibition training. Journal of Cognitive Neuroscience 12, 721–728.
Houdé, O., Pineau, A., Leroux, G., Poirel, N., Perchey, G., Lanoë, C., Lubin, A., Turbelin, M.-R., Rossi, S., Simon, G., Delcroix, N., Lamberton, F., Vigneau, M., Wisniewski, G., Vicet, J.-R., Mazoyer, B., 2011. Functional MRI study of Piaget's conservation-of-number task in preschool and school-age children: a neo-Piagetian approach. Journal of Experimental Child Psychology 110, 332–346.
Izard, V., Pica, P., Spelke, E., Dehaene, S., 2011. Flexible intuitions of Euclidean geometry in an Amazonian indigene group. Proceedings of the National Academy of Sciences 108, 9782–9787.
Johnson, M., 2010. Interactive specialization: a domain-general framework for human functional brain development? Developmental Cognitive Neuroscience 1, 7–21.
Kahneman, D., 2011. Thinking Fast and Slow. Allen Lane, London.
Kahneman, D., Tversky, A., 2000. Choices, Values, and Frames. Cambridge University Press, New York.
Kahneman, D., Slovic, P., Tversky, A., 1982. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, New York.
Karmiloff-Smith, A., 1992. Beyond Modularity: A Developmental Perspective on Cognitive Science. MIT Press, Cambridge.
Kushnir, T., Xu, F., Wellman, H., 2010. Young children use statistical sampling to infer the preferences of others. Psychological Science 21, 1134–1140.
Lipton, J., Spelke, E., 2003. Origins of number sense: large number discrimination in human infants. Psychological Science 14, 396–401.
Loosbroek, E., Smitsman, A., 1990. Visual perception of numerosity in infancy. Developmental Psychology 26, 916–922.
Lubin, A., Pineau, A., Hodent, C., Houdé, O., 2006. Language-specific effects on number computation in toddlers: a European cross-linguistic cartography. Cognitive Development 21, 11–16.
Mandler, J., 1988. How to build a baby: on the development of an accessible representational system. Cognitive Development 3, 113–136.
Marcovitch, S., Zelazo, P., 1999. The A-not-B error: results from a logistic meta-analysis. Child Development 70, 1297–1313.
Mehler, J., Bever, T., 1967. Cognitive capacity of very young children. Science 158, 141–142.
Meltzoff, A., Moore, M., 1977. Imitation of facial and manual gestures by human neonates. Science 198, 75–78.
Meltzoff, A., Prinz, W., 2002. The Imitative Mind: Development, Evolution, and Brain Bases. Cambridge University Press, Cambridge.
Newman, G., Keil, F., Kuhlmeier, V., Wynn, K., 2010. Preverbal infants appreciate that only agents can create order. Proceedings of the National Academy of Sciences 107, 17140–17145.
Onishi, K., Baillargeon, R., 2005. Do 15-month-old infants understand false beliefs? Science 308, 255–258.
O'Reilly, R., 1998. Six principles for biologically based computational models of cortical cognition. Trends in Cognitive Sciences 11, 455–462.
Perret, P., Blaye, A., Paour, J.-L., 2003. Respective contributions of inhibition and knowledge levels in class inclusion development: a negative priming study. Developmental Science 6, 283–286.
Piaget, J., 1952. The Child's Conception of Number. Humanities Press, New York.
Piaget, J., 1954. The Construction of Reality in the Child. Basic Books, New York.
Piaget, J., 1983. Piaget's theory. In: Mussen, P.H. (Ed.), Handbook of Child Psychology, vol. 1. Wiley, New York, pp. 103–128.
Pica, P., Lemer, C., Izard, V., Dehaene, S., 2004. Exact and approximate arithmetic in an Amazonian indigene group. Science 306, 499–503.

Poirel, N., Borst, G., Simon, G., Rossi, S., Cassotti, M., Pineau, A., Houdé, O., 2012. Number conservation is related to children's prefrontal inhibitory control: an fMRI study of a Piagetian task. PLoS One 7 (7), e40802. http://dx.doi.org/10.1371/journal.pone.0040802.

Saxe, G., 1999. Cognition, development, and cultural practices. In: Turiel, E. (Ed.), Development and Cultural Change. Jossey-Bass, San Francisco, pp. 19–35.

Siegler, R., 1996. Emerging Minds: The Process of Change in Children's Thinking. Oxford University Press, New York.

Siegler, R., 1999. Strategic development. Trends in Cognitive Sciences 3, 430–435.

Siegler, R., Booth, J., 2004. The development of numerical estimation skills in young children. Child Development 75, 428–444.

Siegler, R., Opfer, J., 2003. The development of numerical estimation evidence for multiple representations of numerical quantity. Psychological Science 14, 237–243.

Simon, T., 1997. Reconceptualizing the origins of number knowledge: a non-numerical account. Cognitive Development 12, 349–372.

Simon, T., 1998. The foundations of numerical thinking in a brain without numbers. Trends in Cognitive Sciences 3, 363–365.

Spelke, E., 2000. Core knowledge. American Psychologist 55, 1233–1243.

Spelke, E., Vishton, P., von Hofsten, C., 1995. Object perception, object-directed action, and physical knowledge in infancy. In: Gazzaniga, M. (Ed.), The Cognitive Neurosciences. MIT Press, Cambridge, MA, pp. 165–179.

Téglás, E., Vul, E., Girotto, V., Gonzales, M., Tenenbaum, J.B., Bonatti, L., 2011. Pure reasoning in 12-month-old infants as probabilistic inference. Science 332, 1054–1059.

Thompson, P., Cannon, T., Narr, K., van Erp, T., Poutanen, V., Huttunen, M., Lönnqvist, J., Standertskjöld-Nordenstam, C., Kaprio, J., Khaledy, M., Dail, R., Zoumalan, C., Toga, A., 2001. Genetic influences on brain structure. Nature Neuroscience 4, 1253–1258.

Watanabe, M., 1996. Reward expectancy in primate prefrontal neurons. Nature 382, 629–632.

Wellman, H., Cross, D., Bartsch, K., 1987. Infant search and object permanence: a meta-analysis of the A-not-B error. Monographs of the Society for Research in Child Development 51, 1–51 (Serial 214).

Wright, I., Waterman, M., Prescott, H., Murdoch-Eaton, D., 2003. A new Stroop-like measure of inhibitory function development: typical developmental trends. Journal of Child Psychology and Psychiatry, and Allied Disciplines 44, 561–575.

Wynn, K., 1992. Addition and subtraction by human infants. Nature 358, 749–750.

Wynn, K., 1998. Psychological foundations of number: numerical competence in human infants. Trends in Cognitive Sciences 2, 296–303.

Wynn, K., 2000. Findings of addition and subtraction in infants are robust and consistent. Child Development 71, 1535–1536.

Xu, F., Garcia, V., 2008. Intuitive statistics by 8-month-old infants. Proceedings of the National Academy of Sciences 105, 5012–5015.

Cognitive Development, Educational Theories of

Hafiz M Iqbal, University of the Punjab (formerly), Lahore, Pakistan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since very early days, the phenomenon of learning by individuals has fascinated both psychologists and philosophers alike. The complexity of the human learning phenomenon gave rise to different theories or schools of thoughts. One of the most influential schools of thought in educational psychology is cognitivism.

Cognitive theories explain how behavior changes as a result of people's perception, transformation, processing, storage, and retrieval of information in contrast to behaviorist theories that put emphasis only on observable behavior of individuals. Although the origin of cognitive development theories can be traced back to the early twentieth century, it became known to psychologists and academicians through the work of the Swiss psychologist Jean Piaget and the Russian psychologist Lev Vygotsky. Cognitive development theories provide a rich and descriptive data to enhance learning among pupils and to promote their healthy cognitive and emotional developmental needs.

Introduction

The phenomenon of human learning has fascinated humans of all ages. Psychologists have tried to understand and explain how human beings acquire and accumulate information and how their thinking ability develops over a period of time. Psychology as a discipline of human inquiry was mainly concerned with the study of functioning of the mind, including learning. Thus, many psychologists, particularly during the nineteenth and twentieth centuries tried to explain the process of human learning and to discover the laws governing this important human phenomenon. It was believed that proper understanding of the phenomenon will lead to guiding teachers to implement effective teaching and thus improving the teaching–learning process. Many psychologists contributed toward development of the discipline and proposed different theories. For the sake of clear understanding, different authors have grouped these theories into three categories or schools of thoughts. One important school of thought and the most dominant in recent years is cognitivism, the school of thought that tries to explain the origin of intellect in human beings (Piaget, 1952) and how their ability to learn and acquire new information evolves from birth to maturity.

Emergence of Cognitivism

Cognitivism as a school of thought emerged in response to behaviorism in mid-1970s (Bell-Gredler, 1986). Two factors contributed toward the development of cognitive psychology. First, the behaviorist theories could not offer plausible explanation of higher order mental processes like thinking and processing of information. Thus psychologists became frustrated with the limitation of behaviorist psychology to explain internal mental processes of perception, insight, and meaning (Iqbal, 1997). The second factor was the availability of computers and research in areas like linguistics (Baars, 1986; Weisberg and Reeves, 2013) that provided the model for explaining human information processing. In cognitive psychology, the emphasis is on the observable behavior of the individuals, but it is equally important to consider whatever goes inside the mind of that individual (Flavell et al., 1993). However, there are different views to explain that mechanism. Cognitive psychologists believed that human learning is a complex phenomenon and hence cannot be understood by breaking it down into simpler parts.

Early Development

The founder of this school of thought is said to be a German psychologist Max Wertheimer (1880–1943). This school of thought is known as classical Gestalt psychology. Gestalt means configuration. According to this view, learning is the change in figure–ground relationship or the development of a better Gestalt. In Gestalt psychology, the learner sees a total picture with all its parts interrelated (Reilly and Lewis, 1991). The way in which a learner sees any part of the picture depends on its relationship with other parts. Seeing only part/s of a picture will not be meaningful (Pillsbury, 1933). According to this view, learning cannot be reduced to simple observable behavior only.

The second psychologist who contributed toward the development of Gestalt psychology was Wolfgang Kohler (1887–1967). He studied problem solving in apes by putting them in a cage with a bunch of bananas hanged in the cage from the ceiling. The apes could not reach the bananas. There were two boxes in the cage. The apes stacked these boxes together, climbed on top of the boxes, and reached the bananas. Kohler did not see any evidence of trial and error or gradual conditioning process or any stimulus-response (S-R) association as envisioned by behaviorists. On the other hand, the ape perceived a problem situation and tried to solve it. This finding of solution resembled a burst of insight or understanding, following a moment of little pondering, in turn followed by a sudden correct behavior. In the beginning, the ape could not understand the relationship between ceiling, boxes, and bananas. But, when it was put in a problem situation, it had a new perception of all these parts and restructured the situation in the mind and arrived at a solution. It is this new perception, 'this better gestalt,' that allowed the apes to stack the boxes and reach the bananas. The important point of this perception is how different elements are seen relative to the total picture.

The third contributor to the development of cognitive model of learning was Kurt Lewin (1890–1947). Lewin (1936)

theorized that behavior is the result of the situation as perceived by individuals, as individuals relate it to their own needs and purpose. According to Lewin, individuals' perception of a situation is called their 'cognitive field' and the world as they sees or experiences it, including goals as well as barriers to them, is called their 'life space.' Lewin elaborates that an individual faces conflicts in reaching his or her goals. For example, an individual seeks to reach a particular goal that he or she sees in his or her life space. But, the individual finds that there are certain barriers in the achievement of that desired goal. These are the forces that push the individual in an opposite direction. If positive valences are strong enough, the individual will overcome the barrier and reach the goal. However, if the negative valances are strong enough, the individual will find some reason to abandon the task. According to the field theory, the important element is the situation as perceived by the individual. This is the key element in changing behavior. An individual's behavior can be understood and predicted if how the individual perceives the situation and what he or she expects out of it can be found. Similarly, if an individual is able to change his or her perception, he or she will be able to change his or her behavior as well.

The *Gestalt psychology* could not get due attention of educated circles. Instead, behaviorism became a dominant tradition, particularly in American institutions. However, toward the second half of the twentieth century, the school of thought alternative to behaviorism reemerged as cognitivism because of the work of Jean Piaget, Lev Vygotsky, and later of Jerome Bruner. This article briefly describes cognitive theory of Jean Piaget and Lev Vygotsky and its implication for education and teaching. For detailed description of these theories and that of Jerome Burner, the readers are referred to other articles, particularly Piaget's Theory of Cognitive Development (*see* Piaget's Theory of Cognitive Development), Vygotsky's Theory of Cognitive Development (*see* Vygotsky's Theory of Cognitive Development), and Jerome Bruner's Theory of Cognitive Development (*see* Bruner's Theory of Cognitive Development).

Jean Piaget's Theory of Cognitive Development

The intellectual development theory that has had a great bearing on the practice of education during the second half of twentieth century is that of the Swiss psychologist Jean Piaget. His contribution to the development of cognitive tradition is unparalleled. His propositions have generated a lot of interest and research in the field. His theory of cognitive development is most encompassing and comprehensive than any of his predecessors. Though he was researching and writing on this issue of human importance since the beginning of the twentieth century, his findings got recognition nearly 50 years later. Much of the current understanding of children's thinking comes from Piaget's notions.

The basic premise of Piaget's theory is that the complex form of human behavior cannot be reduced to a simple form. He was of the view that complex problem-solving behavior cannot be reduced to a set of conditioned operant, as Skinner would have us believe. Neither can such complex behaviors be reduced to classically conditioned responses, as Pavlov envisioned. Complex phenomena have to be studied in their entirety and with all of their complexity, and there are laws of complex forms of behavior as there are laws of simple forms of behavior, but these are different laws (Iqbal, 1997). Piaget does not deny the existence of either operant or classical conditioning as phenomena, but he views them as referring to trivial aspects of human behavior. Thus, he understands that behavior is influenced by what operant psychologists call reinforcements and others call incentives, but he does not see these as having any relevance to human intellectual performance.

According to Piaget, human beings are not passive recipients of stimuli from the outside world; rather they have to play an active role in their learning. In order to learn, the learner needs to process the information that he receives through senses in the form of stimuli. Active interaction of the child with his or her environment is the salient feature of Piaget's theory. This interaction plays an important role in determining the child's progress from one stage to the other. He did not conceptualize behavior and reinforcement as behaviorists do (Wadsworth, 1989). Being a developmental psychologist, he was more interested in the internal processes of brain and in studying the changes in the cognitive functioning of individuals with the passage of time (Piaget, 1970). He was of the view that thought processes in children are quite different from those in adults as they are not 'little adults.' They have their own unique way of thinking and interpreting the world, which may be different from adults' view but is true with respect to the age group they belong. The term cognition means the product as well as the process of knowing. From birth to maturation, an individual changes the perception of world around him or her. With the passage of time the individual also changes the strategies of knowing the world. This process is called cognitive development (Inhelder and Piaget, 1958).

Piaget contends that the cognitive development of an individual is the outcome an interaction between three factors: heredity, maturation, and the environment in which the child operates. Heredity is the genetic potential of an individual that is received from parents. Maturation is the unfolding of that genetic potential with the passage of time. The environment, both physical and intellectual, provides the necessary stimuli, inputs, and experiences for cognitive stimulation through which the individual constructs his or her knowledge (see **Figure 1**).

According to Piaget cognitive development has two important aspects: the process or mechanism of cognitive development and the outcome of that developmental process (Weisberg and Reeves, 2013).

The Mechanism of Cognitive Development

Before discussing Piaget's stages of development, it is very important to take a note of his main concepts. Piaget used the constructs of schemata, assimilation, accommodation, and equilibration to describe the process of cognitive development and consequently learning (Wadsworth, 1989).

'Schemes' are patterns of behavior that children show in dealing with objects in space. In case of children, schemes can be simple like the pattern of behavior shown by a child in grasping an object within his or her reach. It is complex in case of adults. They show it when they try to solve a complex mathematical problem.

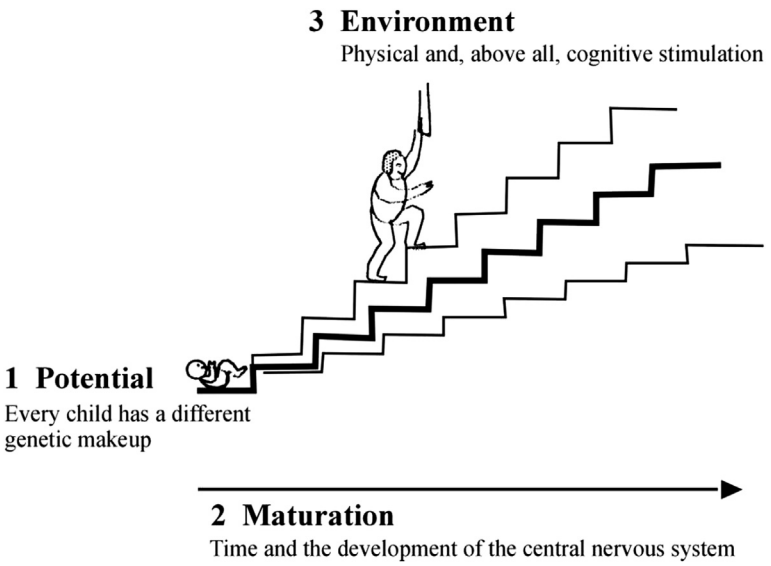

Figure 1 Factors affecting cognitive development.

Assimilation is an internal organizational process used by individuals to make sense of the world around them. Through the process of assimilation, new incoming information becomes part of the old one. However, it is very important for the new knowledge to have best similarity with the old or existing structure/scheme (Gallagher and Reid, 2002; Fischer and Hencke, 1996). In contrast, 'accommodation' occurs when one tries to modify the old schemas to fit the new information in his or her cognitive structures. Piaget says that when the organism fails to handle the new situation with the help of the previously learned pattern of behavior, some sort of inequilibrium or imbalance is created. However, the individual tries to reduce such imbalance. The individual does so by focusing attention on the stimuli that has caused the imbalance. The individual develops new schemes or adopts old ones until the equilibrium is restored. This process of restoring balance is called equilibration (Gallagher and Reid, 2002). Thus the individual uses either assimilation or accommodation to make new information part of his or her schemes and hence learning occurs (Fischer and Hencke, 1996).

Piaget argues that these processes are invariant and continue throughout the life span of an individual. However, as a result of these processes of assimilation and accommodation, the cognitive system itself changes and cognitive development occurs (Piaget, 1972). The mechanism of cognitive development is illustrated in **Figure 2**.

Stages of Cognitive Development

According to Piaget, a child is in a constant state of cognitive change throughout her entire span of life, particularly from her date of birth to the period of adolescence. During this period, an individual passes through different and clearly noticeable stages of cognition. Piaget divides this developmental period into four stages. Each stage is reached or attained by an individual at a particular age. In other words, cognitive development of every individual is age and stage bound. All children invariably go through these stages of cognition in a fixed order. Piaget called these four stages as sensory motor, preoperational, concrete operational, and formal operational. Development of a new stage incorporates the former stage. These stages and the ages at which the individual attains these are given in **Table 1**.

These periods of attainment of different stages are just suggestive. Piaget (1972) himself was of the view that the age at which an individual child attains a particular stage is not fixed.

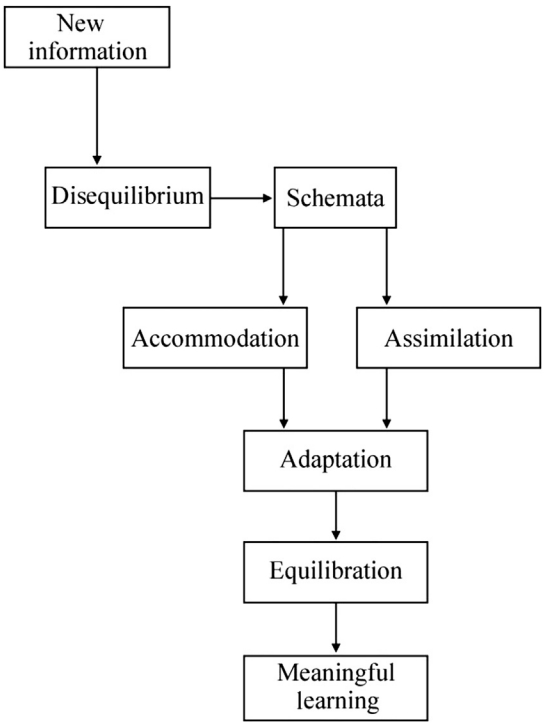

Figure 2 Mechanism of cognitive development.

Table 1 Stages of cognitive and their characteristics

Stage/age	*Characteristics*
1. Sensorimotor (0–2 years)	1. Begins to make use of imitation, memory, and thought 2. Begins to recognize that objects do not seem to exist when they are hidden 3. Moves from reflex action to goal-directed activity
2. Preoperational (2–7 years)	1. Gradually develops use of language and ability to think in symbolic form 2. Able to think logically in one direction 3. Has difficulties seeing another person's point of view
3. Concrete operational (7–11 years)	1. Able to solve concrete (hands-on) problems in a logical fashion 2. Understand conservation and is able to classify and seriate 3. Understand reversibility
4. Formal operational (11–adult)	1. Able to solve abstract problems in a logical fashion 2. Becomes more scientific in thinking 3. Develops concerns about social issues and identity

He seems to accept the idea that the age at which the stages occur can vary from individual to individual depending on the nature of both an individual's experiences and his or her hereditary potential. What is more important is the idea that children can demonstrate specific thinking abilities at the specific age. They can perform a set of mental functions at one stage and not at the other. A higher stage of mental function is built upon and incorporates the lower stage or stages and not the other way round. Piaget calls these mental functions as operations. Thus, a child at the preoperational stage is able to demonstrate operations specific to that age and is not able to perform operations of concrete operational stage.

Vygotsky's Theory of Cognitive Development

Piaget's is not the only theory of children's cognitive development. Another theory that has received increased attention in recent years was proposed by Lev Vygotsky. Like Piaget, the Russian psychologist Lev Vygotsky (1896–1934) also believed that children actively construct their knowledge, but he put more emphasis on social interaction. Both Piaget's and Vygotsky's ideas went unnoticed by the academic community for many years. These were introduced to American audiences through English translations in the 1960s. Since then the psychologists and educators, not only in America but also in other parts of the world, have shown an increased interest in Vygotsky's views and its implications for education (Newman and Holzman, 1993).

Vygotsky's views about cognitive development can be summarized in three points (Daniels, 2008).

1. The child's cognitive skills can be understood only when they are developmentally analyzed and interpreted. For him, it is very important to measure how children form concepts and transform them from one stage to another, and such processes must be observed throughout their development gradually and subsequently (Daniels, 2008).
2. Cognitive skills are mediated through psychological tools or mediators that facilitate in transforming and assessing mental processes and functions such as language, words, counting systems, mnemonic techniques, algebraic symbols, artwork, writing patterns, maps, diagrams, drawings, and other sorts of signs and discourse. Such tools nature is social and artificial rather than organic. He proposed that language is the most important tool in this regard and for children it serves as a tool to plan activities and solve problems in daily life (Daniels, 2008).
3. Vygotsky claimed that cognitive skills have their origins in social relations and are embedded in a sociocultural backdrop (Daniels, 2005). He emphasized that a child's development cannot be studied in isolation from social and cultural activities. Language, for example, is a very important social tool that plays a very important role in development. Similarly, various cultures use different techniques as memory strategies and developing mathematical concepts (Reid et al., 1982).

Vygotsky coined different terminologies to propose the concepts that explain the relationship between society, learning, and development of an individual, which reflect his conceptualization about social origin of cognitive functioning. Three key concepts capture the essence of what Vygotsky means by social origin of cognitive functioning. These are (1) zone of proximal development, (2) scaffolding, and (3) the role of language as a tool for promoting cognitive development (Vygotsky, 1978).

Zone of Proximal Development

- Zone of proximal development, now called the zone of potential development (ZPD) is a term that Vygotsky used to explain the range of development. He is of the view that every individual has two levels of development: a lower level that each is able to attain under normal circumstances and an upper level that each can attain under the influence of an adult or capable peer. The difference between these two levels is the ZPD (Daniels, 2005). Vygotsky's articulation of the ZPD underscores his belief in the importance of social influences, especially instruction, on children's cognitive development.

Scaffolding

- Scaffolding is a concept closely linked with the zone of proximal development. It explains the nature and kind of help and guidance afforded to the child by an adult in solving a relatively difficult task. Vygotsky emphasizes that instead of providing a direct solution to the problem, an adult or more knowledgeable person should adjust the level of guidance to help the learner to arrive at the solution. Direct instruction may be provided when the task is new or more difficult.

  Dialog, discussion, questions, prompts, and clues are different tools that are used during scaffolding. By using

these techniques, the teacher or peer provides logical direction to the thought processes of the more novice learner.

- Language and thought

  Vygotsky (1978) puts an added emphasis on the use of language by children from an early age. He believes that children's use of language is intimately linked with the development of their thought. He believed that young children use language not only to communicate with adults in their surroundings or care givers, but regulate their thinking and behavior as well. Thus, for Vygotsky, children's inner or private speech is an important tool for development of thoughts.

  Vygotsky believed that children's use of language has two dimensions, external and internal. External dimension refers to their use of language to communicate with others and internal dimension refers to their inward focus on their own thoughts.

Vygotsky's views about language are in direct contrast with those of Piaget. Piaget views children's language as nonsocial or egocentric speech and their self-talk being immature thought. In contrast, Vygotsky believed that the language is socially based and governs and guides the children from very early stage (Scholnick et al., 1999). It is equally evident that both Vygotsky and Piaget have given due attention to social relationships and cultural aspects in the process of learning and knowledge enhancement (Smith et al., 1997).

Role of Culture in Cognitive Development

Cognitive development theorists, particularly Vygotsky, have put an added emphasis on the role that a particular culture plays in the cognitive development of individuals. Although Piaget recognizes the role of interaction of children with the environment in unfolding their innate abilities, he puts an upper limit on such development, which is determined by the phenomenon called maturation. Piaget believes the young child to be interacting with the environment on its own and developing the concepts of conservation, seriation, and object permanence without the help of others solely on the basis of his or her observation. This age-bound developmental nature of cognition necessitates that a teacher should teach an age-appropriate and developmentally appropriate content and should adopt instructional strategies accordingly. Vygotsky, on the other hand, believes that culture in the form of social interaction plays a major role in the cognitive development of individuals from very early age. He views cognitive development as a social and cultural process. As a matter of fact, it is this idea that distinguishes Vygotsky from Piaget. Vygotsky's quoted statement reads:

> … every function in the child's … development appears twice: first, on the social level, and later, on the individual level; first between people (interpsychological), and then inside the child (intrapsychological) …
>
> Vygotsky, 1978: p. 57

It was this cultural aspect of human's cognitive development that inspired Vygotsky to propose the concept of zone of proximal (potential) development. Vygotsky defines the ZPD in the following words.

> … the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers.
>
> Vygotsky, 1978: p. 86

Contrary to Piaget, Vygotsky is of the view that learning leads cognitive development, and does not follow it. Through his sociocultural theory and particularly the concept of ZPD of Vygotsky, he in fact emphasized the role that the society can play in the process of cognitive development of an individual. Later research in this field provided further evidence regarding the role of culture in cognitive development (Arnett and Maynard, 2013; Rogoff, 2003; Sternberg, 2004).

Analyzing her research spread over a period of two to three decades, Rogoff (2003) concludes that cultural processes are of crucial importance in human development. She argues that cognitive development not only involves skills and knowledge at an individual level but also is a collective and collaborative endeavor, involving other people around an individual who are part of his or her immediate environment. She describes learning as a process of 'guided participation' shared by the child and other adults in the community. Similarly, Arnett and Maynard (2013) are of the view that human beings are biologically and culturally destined to language and other cultural tools and to learn from each other. They argue that the Western notion of cognitive development cannot be overgeneralized and applied to other cultures because each culture has different bearings on the cognitive developmental process depending upon its values, customs and traditions, and practices. Likewise, Sternberg (2004) emphasizes studying the cognitive development process in cultural context. He argues that if one tries to study intelligence acontextually, one will not be able to understand the phenomenon but will impose one's own view on the rest of the world.

It is basically the recognition of role of culture and adults in the intellectual development, particularly through mediation that provided the perspective and paved the way to explore the possibility of accelerating cognitive development of children and adolescents by way of human intervention (Adey and Shayer, 2014; Iqbal and Shayer, 2000; Shayer, 1999).

Educational Implications of Cognitive Development Theories

Although Piaget was interested in describing how the ability to think and making sense of the world develops in children, he did not talk explicitly about what schools should do in making the learning process more effective. He, however, has described some general principles of learning and that how the environment should be shaped to facilitate cognitive development of children. For example, Piaget believes that learning is an internal process of construction. He argues that learning cannot be deduced from observation or experience alone but is subordinated to developmental needs (Collins, 2002).

Therefore, for children to learn, they must be able to assimilate in order to respond to new concepts or experiences. Piaget thinks that children learn not only by observation but also through a process of self-regulation – a process called equilibration. In this process, they not only record what is going on but also reorganize and coordinate their own activities to form rules and principles.

Vygotsky, on the other hand, argues that learning is a collaborative process where meanings are constructed through the process of social interaction, questioning, discussion, and dialog. Subsequent research in these two domains leads to the conclusion that these two theories are in fact compatible and complement each other. Further work by von Glasersfeld (1990, 1995), Gergen (1985), Salomon and Perkins (1998), and Richardson (1997) in this perspective gave rise to a new school of thought called constructivism. Piaget's and Vygotsky's theories (constructivism) have implications for education, including teaching, learning, curriculum, and assessment. Some of these are as follows:

- Promote inquiry learning
  Piaget emphasized that children learn best when they are active and seek solutions to their questions for themselves. Piaget is opposed to the teaching methods that treat children as passive recipient of information. Hence teachers should promote inquiry learning by providing students with opportunities of exploring the material, setting up experiments, making observations, and drawing conclusions. Teachers should not only respect students' ideas and accept their autonomy, but also encourage them to take initiatives.
- Facilitate learning, do not direct teaching
  Instead of direct teaching and passing information to students in transmission mode, teachers should act as facilitators, helping students to construct their own meanings. Teachers should ask open-ended questions and generate discussion in the class.
- Promote students' questioning and cooperative learning
  In addition to questioning and discussion initiated by the teacher, students should also be encouraged to put questions to the teacher and to one another, seek clarification, and offer explanations. Vygotsky believed that it is not just adults who are important in helping students learn important skills. Students also benefit from the support and guidance offered by more capable peers. Group work, tasks, and assignments should be given to afford students opportunities to work cooperatively.
- Consider child's prior knowledge and level of thinking
  Students come to class with many ideas about the physical and natural world. They have concepts of space, time, quantity, and causality. Although these ideas differ from the ideas of an adult, they are very important from a child's point of view. Teachers need to ascertain children's prior ideas, interpret them, and address those ideas in a manner that is within the cognitive reach of the children.
- Use scaffolding
  Teachers should look for opportunities to use scaffolding when students need help for completing tasks and assignment. They should refrain from providing them correct answers or solutions. They should provide just enough help to the students to arrive at conclusion by themselves.
- Teach in the zone of proximal development
  Vygotsky's concept of zone of proximal development has a good practical value. He believes that social interaction in the class takes the students to the next higher mental level. Teachers should not be able to determine students' current level of development but arrange activities that lead toward further development.
- Aim at students' intellectual growth
  For Piaget, children's learning should occur naturally. He did not endorse the idea of pushing children to achieve too much too early in their development, before they are mature enough to process the given information. When Piaget came to lecture in the United States, he was asked, "What can I do to get my child to a higher cognitive stage sooner?" He was asked this question so often that he called it the 'typical American question.' But as mentioned earlier, subsequent work involving Vygotsky's concept of social construction and mediation made it possible to achieve such objectives. Teachers need to provide a stimulating environment by setting challenging tasks, discrepant events, cognitive dissonance, discussion, and dialog that provoke thinking in students and hence promote intellectual growth.
- Use continuous assessment and assessment for learning
  Constructivists believe that students construct their own meanings to make sense of the world around them. One standardized method will not fit for all. Teachers need to use a variety of techniques, both oral and written, to ascertain students' learning and feed that information for improving learning.
- Assess the ZPD, not the intelligence quotient (IQ)
  Teachers should learn the methods, procedures, and competence to measure children's ZPD to be able to adopt appropriate pedagogical skills and to take them to the next higher level of thinking. Vygotsky argues that formal, standardized assessment tests or IQ tests cannot provide measures to ascertain ZPD.

Acknowledgment

The contributor of this article is thankful to his PhD student Mr Muhammad Anwar for helping to develop the outline and to Ms Mona Khalid and Syeda Farhat for identifying and locating some useful material and references and help prepare the initial draft.

See also: Bruner's Theory of Cognitive Development; Neo-Piagetian Theories of Cognitive Development; Vygotsky's Theory of Cognitive Development.

Bibliography

Adey, P., Shayer, M., 2014. The Effects of Cognitive Acceleration and Speculation about Causes of These Effects. http://modeling.asu.edu/modeling/AdeyPCognitiveAccel2011.pdf (retrieved on 02.04.14.).

Arnett, J., Maynard, A., 2013. Child Development: A Cultural Approach. Pearson, New York.

Baars, B.J., 1986. The Cognitive Revolution in Psychology. Guilford Press, New York.
Bell-Gredler, M.E., 1986. Learning and Instruction: Theory into Practice. Macmillan, New York.
Collins, A., 2002. How students learn and how teachers teach. In: Bybee, R.W. (Ed.), Learning Science and the Science of Learning. NSTA Press, Arlington, VA, pp. 3–11.
Daniels, H., 2005. An Introduction to Vygotsky. Routledge, New York.
Daniels, H., 2008. Vygotsky and Research. Routledge, New York.
Fischer, K.W., Hencke, R.W., 1996. Infants construction of action in context: Piaget's contribution to research on early development. Psychological Science 7 (4), 204–210.
Flavell, J.H., Miller, P., Miller, S., 1993. Cognitive Development, third ed. Prentice Hall, New Jersey.
Gallagher, J.M., Reid, D.K., 2002. The Learning Theory of Piaget and Inhelder. Authors Choice Press, New York.
Gergen, K.J., 1985. The social constructionist movement in modern psychology. American Psychologist 40, 266–275 (cited in Burr, V. An Introduction to Social Constructionism. Routledge, London).
von Glasersfeld, E., 1990. An exposition of constructivism: why some like it radical. In: Davis, R.E., Maher, C.A., Noddings, N. (Eds.), Constructivist Views on the Teaching and Learning of Mathematics. National Council of Teachers of Mathematics, Reston, VA, pp. 19–29 (JRME, Monograph # 4).
von Glasersfeld, E., 1995. Radical Constructivism: A Way of Knowing and Learning. Falmer Press, London.
Inhelder, B., Piaget, J., 1958. The Growth of Logical Thinking: From Childhood to Adolescence. Basic Books, New York.
Iqbal, H.M., 1997. Behaviourism and cognitivism: their relevance and implications for teachers. Bulletin of Education and Research XIX (1), 68–91.
Iqbal, H.M., Shayer, M., 2000. Accelerating the development of formal thinking in Pakistan secondary school students: achievement effects and professional development issues. Journal of Research in Science Teaching 37, 259–274. http://dx.doi.org/10.1002/(SICI)1098–2736(200003)37:3<259::AID-TEA3>3.0.CO;2-W.
Lewin, K.A., 1936. Principles of Topological Psychology. McGraw-Hill, New York.
Newman, F., Holzman, L., 1993. Lev Vygotsky. Routledge, London and New York.
Piaget, J., 1970. Genetic Epistemology. Columbia University Press, New York.
Piaget, J., 1952. The Origin of Intelligence in Children. International University Press, New York.
Piaget, J., 1972. Intellectual evolution from adolescence to adulthood. Human Development 15, 1–12.
Pillsbury, W.B., 1933. The units of experience—meaning or Gestalt. The Psychological Review 40 (6), 481–497.
Reid, T., Forrestal, P., Cook, J., 1982. Small Group Work in the Classroom: Language and Learning Project. Education Department, Western Australia.
Reilly, R.R., Lewis, E.L., 1991. Educational psychology: application for classroom learning and instruction. Macmillan, New York.
Richardson, V., 1997. Constructivist teaching and teacher education: theory and practice. In: Richardson, V. (Ed.), Constructivist Teacher Education: Building a World of New Understandings. Falmer Press, London.
Rogoff, B., 2003. The Cultural Nature of Human Development. Oxford University Press, New York.
Scholnick, E.K., Nelson, K., Gelman, S.A., Miller, P.H. (Eds.), 1999. Conceptual Development: Piaget's Legacy. Lawrence Erlbaum Associates, Inc., New Jersey.
Salomon, G., Perkins, D.N., 1998. Individual and social aspects of learning. Review of Research in Education 23, 1–24.
Shayer, M., 1999. Cognitive acceleration through science education II: its effects and scope. International of Science Education 21 (8), 883–902. http://www.letsthink.org.uk/documents/ca_approach/CASE_II_Effects_and_scope.pdf (retrieved on 28.03.14.).
Sinclair, H., Kamii, C., 1970. Some implications of Piaget's theory for teaching young children. The School Review 78 (2), 169–183.
Smith, L., Dockrell, J., Tomlinson, P. (Eds.), 1997. Piaget, Vygotsky and beyond: Future Issues for Developmental Psychology and Education. Routledge, New York.
Sternberg, R.J., 2004. Culture and intelligence. American Psychologist 59 (5), 325–338.
Vygotsky, L.S., 1978. Mind in Society. The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA.
Wadsworth, B.J., 1989. Piaget's of Theory of Cognitive and Affective Development, fourth ed. Longman, New York.
Weisberg, R.W., Reeves, L.M., 2013. Cognition: From Memory to Creativity. John Wiley & Sons Inc, Hoboken, NJ.

Relevant Websites

http://www.edudemic.com/2012/12/a-simple-guide-to-4-complex-learning-theories/.
http://www.ibe.unesco.org/publications/innodata/inno02.pdf.
http://www.letsthinkinenglish.org/wp-content/uploads/2012/06/LTIS_efficacy.pdf.
http://peoplelearn.homestead.com/beduc/chapter_5.pdf.
http://www.princeton.edu/~achaney/tmve/wiki100k/docs/Learning_theory_(education).html.

Cognitive Development in Infancy, Neural Mechanisms of

Laura A Edwards, Harvard Graduate School of Education, Boston Children's Hospital, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.H. Johnson, volume 3, pp. 2103–2108, © 2001, Elsevier Ltd.

Abstract

The first few years of life are a period of rapid brain and cognitive development. This article outlines the general structural and functional changes in the brain that underlie cognitive development during infancy, and uses examples from the domains of face processing, language development, memory, and numerical cognition to illustrate these developmental processes. The article also briefly introduces the imaging technologies and research methods that are currently used to study brain and behavioral development in infancy.

Introduction

From the first moments of life, infants are developing the neural and behavioral capacities that will form the foundation upon which their future development rests. Newborns only a few minutes to a few hours old, for example, preferentially attend to human speech sounds over nonspeech sounds (Ecklund-Flores and Turkewitz, 1996), and to faces or facelike stimuli over other visual aspects of their surroundings (Johnson et al., 1991). The capacity for rapid auditory processing develops early in infancy, and is believed to be an important precursor of speech perception and processing and later language development (Sheridan et al., 2010). Infancy is, thus, a critical period of rapid cognitive development, which is both produced by and capable of influencing the maturation of brain structures and functional neural circuits. The past decade has seen a vast expansion of our knowledge on the neural mechanisms of infant cognitive development, the major principles of which are discussed in detail in the following sections.

Structural Brain Development in Infancy

A discussion of the neural basis of cognitive development necessarily begins with development of the brain structures and overall architecture that are the substrates of cognition and behavior. Structural brain development begins shortly after conception, and – depending on how it is defined – continues late into adolescence or adult life. Our most detailed understandings of how structural brain development contributes to cognition come from studies using animal models, postmortem brain tissue analyses, and, more rarely, those cases in which fetal or infant brain development is disrupted, leading to atypical cognitive outcomes. As such, it is important to note that our knowledge in this area may not reflect the most normative human developmental trajectories. However, newer neuroimaging technologies – some of which are introduced in the following sections – are already being used to study the links between brain and cognitive development more directly in awake and active infants and young children, and thus to deepen our understanding of the structural substrates of cognition in typical human development.

The precursor of the entire central and autonomic nervous system is the neural plate, which arises from the outermost layer of embryonic tissue, formed from multiplication of the single-celled zygote. Around the 4th week of gestation, this neural plate folds inward and forms the neural tube, which then differentiates into spinal cord and brain tissue, including the structures that will become the developed forebrain, midbrain, and hindbrain. Differentiation of the neural tube into the distinct structures of the nervous system is called *neurulation*, and is induced by chemical growth factors. Evidence of the importance of neurulation for future brain and cognitive development comes from studies of disorders of this process such as holoprosencephaly, in which the forebrain fails to develop cleanly into right and left hemispheres. In rare cases, in which children born with holoprosencephaly survive beyond infancy, these children show profound cognitive (and physical) delays starting in infancy; they usually develop little to no expressive language or higher level cognitive functions, although they have been observed to develop some receptive language at a delayed pace (Barr and Cohen, 1999).

Beginning around the 5th gestational week, neural progenitor cells proliferate rapidly, then develop axons and dendrites, a process known as neurogenesis. The proliferation of progenitor cells peaks by the 3rd or 4th month of gestation, and the development of axons and dendrites continues well into the 2nd year of postnatal life. New neurons formed prenatally migrate from the inside out, eventually giving rise to the six-layered cortex – the seat of cognition (Nelson et al., 2006). In humans, prenatal neurogenesis and neuronal migration are responsible for the majority of neurons comprising the brain throughout the lifetime, although neural progenitor cells present in the subventricular and hippocampal subgranular zones contribute to neurogenesis in the olfactory bulb and hippocampus into adulthood (Bonfanti and Peretto, 2011), and the prefrontal cortex up to 18 months of age (Sanai et al., 2011).

Disorders of neurogenesis in humans tend to result either in macroencephaly (large brain) or microencephaly (small brain; Nelson et al., 2006). Both these disorders and those implicated as disorders of abnormal dendritic growth (such as Rett syndrome and tuberous sclerosis) are frequently associated with mental retardation, although they usually cooccur with significant modifications in neuronal differentiation and

connectivity, making it difficult to tease apart the distinct effects of cortical neurogenesis on cognitive development. Our knowledge of the distinct effects of neurogenesis on cognition instead comes from animal studies such as those in which the antimitotic agent methylazoxymethanol (MAM) is delivered to rats during discrete periods of gestation. The offspring of rats treated with MAM at the period of gestation during which neurogenesis is at its peak experience a 50% reduction in cortical thickness, but exhibit surprisingly moderate alterations in cognitive abilities (Berger-Sweeney and Hohmann, 1997). In fact, the most consistently observed abnormality in these rats is hyperactivity, which may account for their altered performances in other cognitive tasks, such as deficits in spatial recognition, memory, and social interaction (Hazane et al., 2009). Thus the cerebral cortex might be sufficiently plastic to compensate for the loss of early neuron populations later in development.

Synaptogenesis – which typically begins in the 23rd week of human gestation – refers to the formation of points of contact between neurons. This process continues throughout postnatal life, resulting in an overproduction of synapses during or near the first year of life in most regions of the brain. Synaptic overproduction is then followed by a protracted period of synaptic pruning, which is highly experience-dependent; synapses are relatively unstable when they are first formed, and they are either reinforced and maintained by continued activation, or pruned away until adult numbers and densities of synapses are reached. Different brain structures follow different developmental courses in terms of synaptic overproduction and pruning, however. Synaptic densities in primary sensory areas such as the auditory cortex of the temporal lobe, for example, typically reach their peak at about 3 months of age, whereas the density of synapses in the middle frontal gyrus of the prefrontal cortex peaks after 15 months (Huttenlocher and Dabhoklar, 1997). Synapses in the visual cortex decrease to about half their maximum level when children are between 1 and 5 years of age, remaining stable after age 5, while synaptic density in the prefrontal cortex typically only begins to decrease around age 7 (Huttenlocher, 1999) and continues well into the third decade (Petanjek et al., 2011).

During and proceeding from synaptogenesis, spontaneous neuronal firing – independent of external stimuli and thus likely under genetic control – occurs throughout the brain. Additionally, sensory activity, chemical and cellular factors, and environmental stimuli may produce impulse transmission between neurons. Such neuronal firing, if temporally and/or spatially synchronized, forms neural circuits that become integrated and reinforced with continued activation. Brain function arises from the ongoing modulation, enhancement, synchronization, and integration of these neural circuits. Moreover, the plasticity that synaptic overproduction affords and the experience-dependent nature of synaptic pruning likely underlie the developmental course of individual cognitive abilities over the life span. In the domain of language acquisition, for example, as the auditory cortex matures over the first 3 months of life, infants are increasingly able to discriminate the sounds of most languages in the world, regardless of their own language environments: in the first hours and days of life, they show preferences for their native languages (Moon et al., 1993) and they can discriminate their native languages from those of relatively dissimilar language families (Mehler et al., 1988; Mehler and Christophe, 1994; Nazzi et al., 1998); by 4 or 5 months, when synaptic production in the auditory cortex has typically peaked, infants are able to discern their native languages from very close neighbors (Nazzi et al., 1998). In their hallmark studies, Werker and Tees (1983) demonstrated that infants then develop from being able to discriminate both native and nonnative phonemic contrasts to showing an adult-like inability to discriminate nonnative contrasts around 10–12 months of age – coinciding with the onset of synaptic pruning in sensory brain regions. A similar developmental course of cognitive maturation followed by 'perceptual narrowing' has also been observed in the domain of face processing (Nelson, 2001). The 'face-inversion effect' – in which inverted faces are harder to recognize than upright faces – is often used as a marker of expertise in face processing, as it indicates that upright faces belong to a special class of visual stimuli that are rapidly and automatically processed, although they are identical in perceptual properties to inverted faces. Infants develop the face-inversion effect around 4 months of age (Fagan, 1972) as synaptic connections in their visual system are growing and being reinforced, and the ability of infants to discriminate between different faces of nonhuman species declines between 6 and 9 months of age (Pascalis et al., 2002), when synaptic pruning occurs in the visual cortex.

Alongside synaptogenesis and pruning, myelination – the wrapping of an insulating fatty myelin sheath around axons – begins during the 5th month of fetal development. Myelination follows an organized and predictable temporal sequence, beginning at the brainstem and progressing from posterior to anterior regions, then from deep structures to superficial layers of the brain (Ballesteros et al., 1993). The myelination process continues through to the second decade of life, although its most dramatic changes occur during the first year of postnatal development: the base of the brain is myelinated around birth, and deep structures throughout the brain become myelinated by 6 months of age. The cerebral cortex – the last region to myelinate – begins to develop myelin first in the occipital lobe, and then in the frontal, parietal, and temporal lobes, respectively, between 8 and 12 months of age (Paus et al., 2001).

Myelination facilitates more efficient transmission of neural signals and thus establishes rapid communication pathways throughout the brain. It also allows for more synchronized activity within and between neural networks (Deoni et al., 2012). Increases in myelin are associated with domain-general cognitive developments such as the increased information encoding capabilities and processing speeds observed in infants over the first year of life (Richmond and Nelson, 2007; Webb et al., 2005). Additional evidence for the role of myelination in cognition comes from studies of developmental and psychiatric disorders such as autism, attention deficit hyperactivity disorder (ADHD), schizophrenia, and bipolar disorder, in which aberrant myelin development has been detected, and appears to contribute to the underconnectivity or desynchronized brain activity that may underlie cognitive dysfunction (Just et al., 2012; Meda et al., 2012; Mills et al., 2012).

Technologies such as structural magnetic resonance imaging (MRI) – which produce macrostructural images of the brain and can be used to measure white matter density and

cortical thickness associated with myelination – and diffusion tensor MRI – which is sensitive to the architecture of tracts and fibers throughout the brain – have made it possible to trace the development of myelin throughout even the deepest structures in the brains of infants and young children. Our current understanding of myelin progression in the typically developing infant brain is thus becoming more and more extensive; research in the near future may even be able to answer questions of individual differences in myelination, in order to provide a more precise and extensive picture of the role of myelin in cognitive development. For example, increasing sophistication in imaging resolution and analysis techniques may enable the quantification of variations in myelination between individuals, and investigations of the relationships between these differences and observable cognitive differences between infants and young children.

Functional Bases of Cognitive Development in Infancy

Studies such as those explored above, in which changes in brain morphology are linked to cognitive changes in humans, tend to be correlational in nature. In order to move beyond the level of association, it is necessary to observe the brain directly during specific cognitive tasks. The development of noninvasive imaging technologies – such as electroencephalography (EEG) and derived event-related potentials (ERPs), functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), functional near-infrared spectroscopy (fNIRS), or multimodal applications that may include combinations of the above – has enabled researchers to do just this.

EEG and ERPs use electrodes placed on the scalp to detect the electrical activity produced by neuronal firing within the brain. MEG provides a similar measure of brain activity, by recording the magnetic fields associated with these electrical currents. As such, MEG, EEG, and ERPs allow researchers to detect correlates of brain function that occur on the timescale of synaptic transmission, and studies utilizing these techniques thus have very high temporal resolution. fMRI, on the other hand, measures metabolic changes, specifically the blood-oxygen-level-dependent (BOLD) change, associated with increased neural circuit activation. The BOLD response occurs on a protracted timescale compared to neuronal activity, and while fMRI lacks the temporal resolution of EEG/ERP, it is useful for providing measures of brain activation with high spatial resolution. fMRI also has the advantage of providing detailed images of the brain without exposure to ionizing radiation, such as that used in positron emission tomography (PET). fNIRS is a newer imaging technology that, similar to fMRI, measures metabolic changes associated with brain activity in the cerebral cortex. While fNIRS lacks the spatial resolution for imaging deep structures that fMRI affords, it produces individual measures of oxyhemoglobin, deoxyhemoglobin, and total hemoglobin flow to the cortex, and due to its minimalistic hardware, noiseless data acquisition properties, and relative tolerance of participant movement, has distinct advantages for the functional neuroimaging of waking infants in active cognitive tasks. Although most of the research using these imaging technologies has so far been carried out on adult populations, these tools are increasingly being used to investigate the neural mechanisms underlying cognition in infants and children, and thus provide a developmental perspective on brain function over time (see **Figure 1**).

Developmental neuroimaging studies have elucidated several general principles of functional neural development, which underlie the cognitive development that is observable over the first few years of life. At a fundamental level, functional activity associated with cognition tends to progress from subcortical to cortical localization over infancy. Face processing, for example, is likely subserved by subcortical regions in newborns, who have been shown to preferentially attend to faces, facelike stimuli and direct eye contact from as early as the first hours of postnatal life (Farroni et al., 2002; Johnson et al., 1991); basic visual cortical function, however, is not thought to

Figure 1 Advances in noninvasive neuroimaging technologies are enabling researchers to elucidate the neural substrates of cognitive development in typically developing infants and young children. On the left, a 7-month-old is fitted with a high density electrode cap for the acquisition of EEG and ERP data. At right, minimalistic fNIRS hardware fits into a cap and enables the measurement of oxyhemoglobin and deoxyhemoglobin concentration changes in the cortex, while infants and young children (pictured here, a 2-year-old boy) carry out active tasks.

develop until the 3rd to 4th months of life (Braddick et al., 1980). By 2 months of age, infants in PET studies showed specific responses to face stimuli within the occipital cortex (Tzourio-Mazoyer et al., 2002). Although some functional activation thus occurs in these cortical regions before they are fully anatomically and metabolically developed, the progression of functional correlates of cognitive activity from subcortical to cortical regions is likely related to the fact that subcortical regions develop structurally before cortical regions; cognitive functions thus develop as their underlying cortical substrates mature and become functionally active.

As different regions of the cortex mature at different rates, the development of specific cognitive abilities is often associated with changing regions of functional activation over time. Changes in memory – arguably the basis of all learning and cognitive development – are good illustrations of this principle. In adults, declarative memory, or memory for facts or events, depends on a distributed network of medial temporal and cortical structures, and their intervening connections. Sensory inputs from a variety of origins converge on the hippocampus (a structure that lies deep in the medial temporal lobe) where they are bound together to create a memory trace; various cortical structures are implicated in the long-term storage of memories, and the prefrontal cortex is thought to be critical to memory retrieval (Bauer, 2013). PET studies of infants indicate that areas of the temporal lobe show peak metabolic activity by 3 months of age, then slowly decline to adult levels; in the medial temporal regions of the hippocampus, adult numbers, and densities of synapses are reached by approximately 6 months (Bauer, 2009, 2013). Accordingly, declarative memory capacities are low in very young infants, and ERP studies on 4-month-olds produce no evidence that these children differentiate between familiar and novel stimuli (Nelson and Collins, 1991). By about 6 months of age, however, many infants can remember an average of one action of a 3- or 5-step action sequence for a period of 24 h (Barr et al., 1996; Collie and Hayne, 1999), and ERP studies indicate that 8-month-olds show distinct responses to familiar vs. unfamiliar stimuli, even after infrequent exposure to the 'familiar' (Nelson and Collins, 1992). As the dentate gyrus of the hippocampus matures (showing peak synaptic density around 24 months and pruning to adult levels by 8 years of age; Seress and Ábrahám, 2008) and major frontal lobe development occurs (blood flow to this region peaks between 8 and 14 months of age (Chugani, 1994) and synapses in this region develop adult morphology around 24 months (Huttenlocher, 1979)) over a longer period, they are accompanied by changes in various aspects of declarative memory, such as encoding and retention periods. For example, 16-month-olds exhibit less immediate recall than 20-month-olds, and 12-month-olds require more learning trials to achieve the same level of recall as 18-month-olds (Bauer and Dow, 1994; Howe and Courage, 1997). In terms of long-term retrieval, by 9 months of age infants can remember single actions for up to 5 weeks and by 13–14 months of age, they may remember over delays of up to 4–6 months (Carver and Bauer, 2001). In the latter half of the second year of life, infants may remember entire 4-step action sequences for up to 12 months (Bauer et al., 2000). Improvements in encoding and lengths of recall periods then improve, albeit less dramatically, until adolescence, when the medial temporal and cortical structures and their intervening connections have reached near adult structural and functional maturity.

Over the course of infant cognitive development, brain activity also changes from being diffuse (recruiting larger or more areas) to becoming more focal (activating smaller regions with increased magnitudes; Durston and Casey, 2006). In the domain of face processing, for example, 2-month-old infants who were shown static pictures of unknown women's faces and imaged using PET, showed neural activation over a distributed network of cortical areas; these areas included the inferior temporal gyrus and the occipitotemporal sulcus, which are likely homologues of the fusiform face area in adults, as well as the left superior temporal and inferior frontal gyri, which are not recruited for face processing, but are associated with language processing in adults (Tzourio-Mazoyer et al., 2002). ERP studies of face processing also reveal that there are two distinct components of the infant ERP waveform that show sensitivity to faces (the N290 and the P400). These components are more medially distributed and smaller in amplitude than the adult face response (the N170; de Haan et al., 2002). While the infant N290 is larger to human than monkey faces by 3 months of age, this component does not differentiate between upright and inverted faces until 12 months, when its amplitude is higher in response to upright relative to inverted faces. By about 4 years of age, as a result of cortical specialization, a single adult-like N170 is observed during face processing, and the amplitude of this component continues to increase, reaching adult levels in adolescence (Nelson et al., 2006).

The change from more diffuse to more focal brain activation may reflect the functional consequences of synaptic pruning and other regressive brain maturation processes, in concert with the strengthening and fine tuning of relevant neural connections. Often, as brain activation becomes more focal over development, activation patterns may also change from being bilaterally distributed to becoming increasingly lateralized within one hemisphere of the brain. Recent advances in speech perception research provide a prime example of this phenomenon. Two- to five-day-old neonates who were played forward and backward recordings of infant-directed speech (Peña et al., 2003) or speech and music (Kotilahti et al., 2010) and imaged using fNIRS showed significant increases in total hemoglobin in the channels overlying the left and right temporal regions during forward speech in comparison to silence. The total hemoglobin increase was also significantly greater in the left compared to the right hemisphere when neonates heard forward speech as compared to reversed speech (Peña et al., 2003). In an fMRI study of 3-month-olds, similar lateralization was observed to forward vs. reversed speech in the left superior temporal and angular gyri, and significant activity was observed in the right dorsolateral prefrontal cortex in response to both types of speech stimuli in waking but not sleeping infants (Dehaene-Lambertz et al., 2002). Thus, although infants show hemispheric lateralization (a bias toward left hemisphere dominance) in speech processing from birth, this left hemisphere activity can be nonspecific, and is accompanied by significant functional activation in the right hemisphere. By about 6 months of age, however, infants imaged using fNIRS showed significant oxyhemoglobin increases in the left hemisphere, but not in

the right hemisphere, to speech in the presence of visual animations as compared to only visual stimuli (Bortfeld et al., 2009), and significant oxyhemoglobin changes in the right hemisphere to stimuli that differed in pitch (Homae et al., 2012). This developmental trajectory – with speech processing becoming more localized to regions of the left hemisphere, and pitch and prosody processing increasingly associated with right hemispheric activation as cortical specialization increases – continues through infancy and childhood into adolescence. In an fMRI study of children aged 7–18 years, activation to a verb generation task occurred in Broca's area, Wernicke's area, the cingulate gyrus, and the dorsolateral prefrontal cortex, with significant increases in lateralization and lateralized signal strength in Broca's area associated with increasing age across the participants (Holland et al., 2001).

The timing of neural responses that subserve cognitive functions also changes over development. In the aforementioned face-processing studies, in addition to ERP amplitude changes over development, researchers observed changes in the latency of ERP responses, with the general tendency for responses to occur more quickly, or to decrease in latency, as infants develop. Both the N290 and the P400 ERP deflections are so named because they occur at 290 and 400 ms, respectively, after a face is seen. These two components are the putative precursors of the adult ERP response to faces, the N170, which is observed 170 ms after a face stimulus is presented. Even before the more adult-like N170 response is observed in children, the latency of the P400 also changes over development. This component does not discriminate between upright and inverted human faces at 6 months of age, but by 12 months of age it responds to upright human faces with shorter latency than that observed in response to inverted faces (Halit et al., 2003).

Decreases in latency of neural signals might result from increased myelination throughout regions of the developing brain, as myelin increases the efficiency with which signals are transmitted along synapses, particularly over long distances, or between spatially distinct neural circuits. Increasing myelination along with the reinforcement and strengthening of neural circuits relevant to specific cognitive capacities results in enhanced structural connectivity over development. This structural connectivity increase is mirrored by an increase in functional connectivity, observed through various means, including the increasing prevalence and magnitude of EEG and resting fMRI coherence (synchronous oscillations measured over spatially distinct regions of the brain). In a study of EEG coherence changes in theta band frequencies over development, coherence increases were observed over neighboring regions (separated by 6 cm interelectrode distances) over the parietal and central- to lateral-temporal regions around 1.5 years of age, and then along the rostral caudal dimension (over 12 cm interelectrode distances) at 2.5 years of age (Thatcher, 1992). Functional connectivity continued to increase in sudden spurts, over longer ranges, and over both ipsilateral and bilateral distances as participants increased in age to 7 years. Using fMRI, Gao et al. (2009) more recently found little connectivity in 2-week-old infants, a marked increase in the percentage of connectivity and the number of regions exhibition connectivity in the brains of 1-year-olds, and a steady increase from then until 2 years of age, at which point the default network had attained adult-like connectivity. Such increases in long-range functional connectivity occur along with the aforementioned focalization of neural activity over development, so that smaller regions of increasingly specialized brain activation may be integrated. The cognitive consequences of increased functional connectivity are thus likely observed in the development of higher order cognitive processes such as visual working memory, which has shown evidence of increasing rapidly in change-preference tasks (in which infants with the working memory capacity to detect changes in visual stimuli are predicted to preferentially look at one or more of a set of displays that has changed) between 6 and 10 months of age (Perone et al., 2011).

Several other patterns of change in brain activation are observed during infancy. As cognitive abilities develop, for example, functional activation increases in the areas of the brain known to be critical for these competencies in adulthood. Bell and Fox (1997) found individual differences in EEG power in a group of 8-month-olds, such that infants who were successful at object permanence tasks showed increased EEG power over frontal and occipital regions as compared to those who were unsuccessful at the tasks. In a longitudinal study using fNIRS, increases in oxyhemoglobin were observed over the frontal cortex between the period of 5 and 12 months of age, as object permanence developed. Significant frontal activation was only observed in infants once they had attained object permanence, as compared to the previous time point in which they had not yet shown behavioral evidence of object permanence. Infants in this participant pool developed object permanence and the associated change in functional brain activation around 7 months of age (Baird et al., 2002).

In addition to changes in neural activation and cognition that occur as a result of different brain regions coming 'online' at varying rates, top-down shifts in cognitive strategy may alter the neural substrates of cognition over development. In the domain of mathematical cognition, for example, infants are able to discriminate quantities with increasing proficiency over development starting long before they begin to learn language and develop a speech-based numerical symbol system (Feigenson et al., 2004). Once language abilities start to develop and numbers are taught through a verbal symbol system, however, the substrates of numerical cognition expand to include new regions such as the left angular gyrus and perisylvian areas, which are activated during both phoneme detection and calculation tasks (Dehaene et al., 2003). Although longitudinal research tracking the neural correlates of mathematical skills over development remains to be done, recent studies using fNIRS technology have begun to chart the precursors of mathematical cognition over development. Hyde et al. (2010), for example, have demonstrated that the right parietal cortex is sensitive to discrimination of quantities in a 1:2 ratio, in infants as young as 6-months-old.

Summary and Considerations for Future Research

The first few years of life are a period of rapid cognitive development. Within 2 years of being born, children typically

develop the capacity to communicate with their parents and peers through language, to rapidly recognize and distinguish between the faces of other humans, and to remember events up to a year after they have occurred. Both structural and functional changes in the brain underlie and give rise to the cognitive developments, as well as the changes in behavior that we observe over the course of a person's lifetime; additionally and in turn, changes in behavior also bring about changes in the brain. For example, as the cortex becomes increasingly active and specialized, infants develop the ability to hear and then produce an array of distinct sounds; as children increasingly produce and listen to some speech sounds and not others though, those neural circuits corresponding to the sounds that they practice become stronger, both functionally (increasingly efficient neural signal transmission) and anatomically (increasing synaptic density, changes in the numbers of dendrites), and those circuits associated with sounds that they do not practice weaken (as neuronal signaling slows at these synapses) and disappear (as neurons die naturally by apoptosis). In understanding the ontogeny of human cognition then, it is necessary to take a dynamic, developmental approach that focuses on the trajectories that brain activity and behavior take over time, rather than to assume that the adult brain is a template that guides brain and cognitive development over the life span. Such a developmental focus has important implications for the design and interpretation of future research into human cognition, as well as profound practical implications for the ways in which educational and therapeutic programs are developed for individuals at different stages of life.

Increasingly sophisticated and diverse imaging technologies are now being used to elucidate the substrates of cognition and the changes therein in infants and young children. Such technologies are particularly useful for understanding development during infancy, when the precursors of various cognitive capacities are being formed, but cannot be directly observed in behavior. Neuroimaging has been used to show, for example, that typically developing infants can discriminate even closely related languages from their native languages by 4 or 5 months of age, long before they utter their first words. These methods may thus enable researchers and clinicians to monitor the development of typical abilities in infants and to identify those at risk for difficulties or impairments long before they might have been able to do so behaviorally. As research into the cognitive development of infants and young children moves forward, neuroimaging may become an important tool for facilitating early diagnosis and enabling early intervention for those children at risk of developmental delays or deficits.

Relatedly, neuroimaging may enable researchers to illuminate the extent to which certain cognitive capacities are experience-dependent and perhaps more malleable, or experience-independent, and thus unlikely to be affected by interventions. For example, although behavioral data indicate that humans are born with a proclivity for human faces, ERP studies have shown that the neural bases of these responses to faces are distinct from those used by adults, and may develop for the faces of other species to which infants are given increased exposure (Nelson et al., 2006). In the case of a child who is at increased risk of exhibiting face processing deficits or who does not show typical patterns of orienting to faces then, these studies indicate a potential role for human intervention – perhaps by increased exposure to relevant faces and/or the use of specific ERP components to monitor this ability over development – in improving the child's developmental outcomes. Understanding those areas of brain and cognitive development that appear to be experience-independent and thus less changeable by human intervention is just as important, in preventing the expenditure of time and money on educational or therapeutic programs that may ultimately prove fruitless.

In addition to experience-dependence and -independence, caregivers and scholars of human cognitive development are also invested in determining the relative role of genetic and environmental influences on the emergence of cognitive abilities. Such distinctions are notoriously difficult to tease apart, because of the reciprocal relationship that exists between the brain and behavior. Although we often talk about the neural substrates underlying behavior, it is important to keep in mind that behavior – or one's environment – in turn, may influence aspects of brain structure and function, even at the level of the genes that guide these processes. While studies of this nature are currently conducted using cross-rearing designs in animal models, experimental designs involving identical twins, or rarer foster care and adoption studies in humans, the application of infant neuroimaging techniques to the field of 'transcultural neuroimaging' may provide an additional means of investigating the relative roles of the brain and the environment in human cognitive development. In the emerging field of transcultural neuroimaging, studies are designed to directly compare the neural correlates of specific behaviors and functions in two or more participant groups from separate cultures. Here, culture may be defined in a number of ways depending on the aims of the study – from as broadly as members of Western and non-Western cultures, to as narrowly as individuals who share identical geographies and ethnicities, but are members of different religious groups (Han and Northoff, 2008). In an interesting study using this approach, Gutchess et al. (2006), found that the neural substrates of object processing differed between Chinese and American participants in ways that paralleled the semantics of their native languages. If extended downward to infant populations, it may be possible to distinguish whether the early neural correlates of these differences existed from birth, or developed from common neural precursors, diverging only later, upon exposure to culturally specific environmental inputs.

Although the last decade of research has produced impressive advances in our understanding of the neural mechanisms of typical human cognitive development from infancy, most of these studies analyze data at the group level, and report measures of brain and cognitive development on average in the population, rather than on a more individualized basis. This trend is in part a limitation inherent to many of the imaging modalities and analytic methods currently available for investigating infants, which derive the power necessary to make statistically significant claims from large sample sizes. Average population measurements, however, run the risk of masking tremendous variation in typically developing populations, and particularly characteristic of infant populations. As such, future

research into the neural mechanisms of cognitive development in infants should employ designs that enable findings to be interpreted on both a group and individual level, and an emphasis should be placed on the use and development of imaging technologies that enable visualization and analysis of individual-level data. In addition to more accurately representing the range of neural and behavioral development inherent in typical populations, studies that relate quantitative measures of structural or functional brain activity to individual differences in cognitive performance may also provide stronger evidence than is possible through group-level correlational studies, for the role of particular neural mechanisms in the development of behavior.

See also: Brain Development: Ontogenetic Neurobiology of; Cognitive Development During Infancy and Early Childhood across Cultures; Cognitive Neuropsychology, Methods of; Neural Basis of Language Development; Neural Correlates of Social Development; Stress: Effects on Neonatal and Child Development.

Bibliography

Baird, A.A., Kagan, J., Gaudette, T., Walz, K., Hershlag, N., Boas, D.A., 2002. Frontal lobe activation during object permanence: data from near-infrared spectroscopy. NeuroImage 16 (4), 1120–1126. http://dx.doi.org/10.1006/nimg.2002.1170.

Ballesteros, M.C., Hansen, P.E., Soila, K., 1993. MR imaging of the developing human brain. Part 2. Postnatal development. Radiographics: A Review Publication of the Radiological Society of North America, Inc. 13 (3), 611–622.

Barr, M., Cohen, M.M., 1999. Holoprosencephaly survival and performance. American Journal of Medical Genetics 89 (2), 116–120. http://dx.doi.org/10.1002/(SICI)1096-8628(19990625)89:2<116::AID-AJMG10>3.0.CO;2–4.

Barr, R., Dowden, A., Hayne, H., 1996. Developmental changes in deferred imitation by 6- to 24-month-old infants. Infant Behavior & Development 19 (2), 159–170. http://dx.doi.org/10.1016/S0163-6383(96)90015-6.

Bauer, P.J., 2009. The cognitive neuroscience of the development of memory. In: Courage, M.L., Cowan, N. (Eds.), The Development of Memory in Infancy and Childhood, second ed. Psychology Press, New York, NY, USA, pp. 115–144.

Bauer, P.J., 2013. Event memory: neural, cognitive, and social influences on early development. In: Legerstee, M., Haley, D.W., Bornstein, M.H. (Eds.), The Infant Mind: Origins of the Social Brain. Guilford Press, New York, NY, USA, pp. 146–166.

Bauer, P.J., Dow, G.A., 1994. Episodic memory in 16- and 20-month-old children: specifics are generalized but not forgotten. Developmental Psychology 30 (3), 403–417.

Bauer, P.J., Wenner, J.A., Dropik, P.L., Wewerka, S.S., 2000. Parameters of remembering and forgetting in the transition from infancy to early childhood. Monographs of the Society for Research in Child Development 65 (4), v–204.

Bell, M.A., Fox, N.A., 1997. Individual differences in object permanence performance at 8 months: locomotor experience and brain electrical activity. Developmental Psychobiology 31 (4), 287–297.

Berger-Sweeney, J., Hohmann, C.F., 1997. Behavioral consequences of abnormal cortical development: insights into developmental disabilities. Behavioural Brain Research 86 (2), 121–142. http://dx.doi.org/10.1016/S0166-4328(96)02251-6.

Bonfanti, L., Peretto, P., 2011. Adult neurogenesis in mammals–a theme with many variations. The European Journal of Neuroscience 34 (6), 930–950. http://dx.doi.org/10.1111/j.1460-9568.2011.07832.x.

Bortfeld, H., Fava, E., Boas, D.A., 2009. Identifying cortical lateralization of speech processing in infants using near-infrared spectroscopy. Developmental Neuropsychology 34 (1), 52–65. http://dx.doi.org/10.1080/87565640802564481.

Braddick, O., Atkinson, J., Julesz, B., Kropfl, W., Bodis-Wollner, I., Raab, E., 1980. Cortical binocularity in infants. Nature 288, 363–365. http://www.nature.com/nature/journal/v288/n5789/abs/288363a0.html.

Carver, L.J., Bauer, P.J., 2001. The dawning of a past: the emergence of long-term explicit memory in infancy. Journal of Experimental Psychology. General 130 (4), 726.

Chugani, H.T., 1994. Development of regional brain glucose metabolism in relation to behavior and plasticity. In: Dawson, G., Fischer, K.W. (Eds.), Human Behavior and the Developing Brain. Guilford Press, New York, NY, USA, pp. 153–175.

Collie, R., Hayne, H., 1999. Deferred imitation by 6- and 9-month-old infants: more evidence for declarative memory. Developmental Psychobiology 35 (2), 83–90. http://dx.doi.org/10.1002/(SICI)1098-2302(199909)35:2<83::AID-DEV1>3.0.CO;2-S.

Dehaene, S., Piazza, M., Pinel, P., Cohen, L., 2003. Three parietal circuits for number processing. Cognitive Neuropsychology 20 (3–6), 487.

Dehaene-Lambertz, G., Dehaene, S., Hertz-Pannier, L., 2002. Functional neuroimaging of speech perception in infants. Science (New York, NY) 298 (5600), 2013–2015. http://dx.doi.org/10.1126/science.1077066.

Deoni, S.C.L., Dean, D.C., O'Muircheartaigh, J., Dirks, H., Jerskey, B.A., 2012. Investigating white matter development in infancy and early childhood using myelin water faction and relaxation time mapping. NeuroImage 63 (3), 1038–1053. http://dx.doi.org/10.1016/j.neuroimage.2012.07.037.

Durston, S., Casey, B.J., 2006. What have we learned about cognitive development from neuroimaging? Neuropsychologia 44 (11), 2149–2157. http://dx.doi.org/10.1016/j.neuropsychologia.2005.10.010.

Ecklund-Flores, L., Turkewitz, G., 1996. Asymmetric headturning to speech and nonspeech in human newborns. Developmental Psychobiology 29 (3), 205–217. http://dx.doi.org/10.1002/(SICI)1098-2302(199604)29:3<205::AID-DEV2>3.0.CO;2-V.

Fagan 3rd, J.F., 1972. Infants' recognition memory for faces. Journal of Experimental Child Psychology 14 (3), 453–476.

Farroni, T., Csibra, G., Simion, F., Johnson, M.H., 2002. Eye contact detection in humans from birth. Proceedings of the National Academy of Sciences of the United States of America 99 (14), 9602–9605. http://dx.doi.org/10.1073/pnas.152159999.

Feigenson, L., Dehaene, S., Spelke, E., 2004. Core systems of number. Trends in Cognitive Sciences 8 (7), 307–314. http://dx.doi.org/10.1016/j.tics.2004.05.002.

Gao, W., Zhu, H., Giovanello, K.S., Smith, J.K., Shen, D., Gilmore, J.H., Lin, W., 2009. Evidence on the emergence of the brain's default network from 2-week-old to 2-year-old healthy pediatric subjects. Proceedings of the National Academy of Sciences of the United States of America 106 (16), 6790–6795. http://dx.doi.org/10.1073/pnas.0811221106.

Gutchess, A.H., Welsh, R.C., Boduroglu, A., Park, D.C., 2006. Cultural differences in neural function associated with object processing. Cognitive, Affective & Behavioral Neuroscience 6 (2), 102–109.

de Haan, M., Pascalis, O., Johnson, M.H., 2002. Specialization of neural mechanisms underlying face recognition in human infants. Journal of Cognitive Neuroscience 14 (2), 199–209. http://dx.doi.org/10.1162/089892902317236849.

Halit, H., de Haan, M., Johnson, M.H., 2003. Cortical specialisation for face processing: face-sensitive event-related potential components in 3- and 12-month-old infants. NeuroImage 19 (3), 1180–1193.

Han, S., Northoff, G., 2008. Culture-sensitive neural substrates of human cognition: a transcultural neuroimaging approach. Nature Reviews. Neuroscience 9 (8), 646–654. http://dx.doi.org/10.1038/nrn2456.

Hazane, F., Krebs, M.-O., Jay, T.M., Le Pen, G., 2009. Behavioral perturbations after prenatal neurogenesis disturbance in female rat. Neurotoxicity Research 15 (4), 311–320. http://dx.doi.org/10.1007/s12640-009-9035-z.

Holland, S.K., Plante, E., Weber Byars, A., Strawsburg, R.H., Schmithorst, V.J., Ball Jr., W.S., 2001. Normal fMRI brain activation patterns in children performing a verb generation task. NeuroImage 14 (4), 837–843. http://dx.doi.org/10.1006/nimg.2001.0875.

Homae, F., Watanabe, H., Nakano, T., Taga, G., 2012. Functional development in the infant brain for auditory pitch processing. Human Brain Mapping 33 (3), 596–608. http://dx.doi.org/10.1002/hbm.21236.

Homae, F., Watanabe, H., Otobe, T., Nakano, T., Go, T., Konishi, Y., Taga, G., 2010. Development of global cortical networks in early infancy. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 30 (14), 4877–4882. http://dx.doi.org/10.1523/JNEUROSCI.5618-09.2010.

Howe, M.L., Courage, M.L., 1997. Independent paths in the development of infant learning and forgetting. Journal of Experimental Child Psychology 67 (2), 131–163.

Huttenlocher, P.R., 1979. Synaptic density in human frontal cortex – developmental changes and effects of aging. Brain Research 163 (2), 195–205.

Huttenlocher, P.R., 1999. Dendritic synaptic development in human cerebral cortex: time course and critical periods. Developmental Neuropsychology 16 (3), 347–349. http://dx.doi.org/10.1207/S15326942DN1603_12.

Huttenlocher, P.R., Dabhoklar, A.S., 1997. Regional differences in synaptogenesis in human cerebral cortex. The Journal of Comparative Neurology 387, 167–178.

Hyde, D.C., Boas, D.A., Blair, C., Carey, S., 2010. Near-infrared spectroscopy shows right parietal specialization for number in pre-verbal infants. NeuroImage 53 (2), 647–652. http://dx.doi.org/10.1016/j.neuroimage.2010.06.030.

Johnson, M.H., Dziurawiec, S., Ellis, H., Morton, J., 1991. Newborns' preferential tracking of face-like stimuli and its subsequent decline. Cognition 40 (1–2), 1–19. http://dx.doi.org/10.1016/0010-0277(91)90045-6.

Just, M.A., Keller, T.A., Malave, V.L., Kana, R.K., Varma, S., 2012. Autism as a neural systems disorder: a theory of frontal-posterior underconnectivity. Neuroscience and Biobehavioral Reviews 36 (4), 1292–1313. http://dx.doi.org/10.1016/j.neubiorev.2012.02.007.

Kotilahti, K., Nissilä, I., Näsi, T., Lipiäinen, L., Noponen, T., Meriläinen, P., Fellman, V., 2010. Hemodynamic responses to speech and music in newborn infants. Human Brain Mapping 31 (4), 595–603. http://dx.doi.org/10.1002/hbm.20890.

Meda, S.A., Gill, A., Stevens, M.C., Lorenzoni, R.P., Glahn, D.C., Calhoun, V.D., Pearlson, G.D., 2012. Differences in resting-state functional magnetic resonance imaging functional network connectivity between schizophrenia and psychotic bipolar probands and their unaffected first-degree relatives. Biological Psychiatry 71 (10), 881–889. http://dx.doi.org/10.1016/j.biopsych.2012.01.025.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., Amiel-Tison, C., 1988. A precursor of language acquisition in young infants. Cognition 29 (2), 143–178.

Mehler, Jacques, Christophe, A., 1994. Language in the infant's mind. Philosophical Transactions: Biological Sciences, 13–20.

Mills, K.L., Bathula, D., Dias, T.G.C., Iyer, S.P., Fenesy, M.C., Musser, E.D., Fair, D.A., 2012. Altered cortico-striatal-thalamic connectivity in relation to spatial working memory capacity in children with ADHD. Frontiers in Psychiatry/Frontiers Research Foundation 3, 2. http://dx.doi.org/10.3389/fpsyt.2012.00002.

Moon, C., Cooper, R.P., Fifer, W.P., 1993. Two-day-olds prefer their native language. Infant Behavior and Development 16 (4), 495–500.

Nazzi, T., Bertoncini, J., Mehler, J., 1998. Language discrimination by newborns: toward an understanding of the role of rhythm. Journal of Experimental Psychology. Human Perception and Performance 24 (3), 756–766.

Nelson, C.A., 2001. The development and neural bases of face recognition. Infant and Child Development 10, 3–18.

Nelson, C.A., Collins, P.F., 1991. Event-related potential and looking-time analysis of infants' responses to familiar and novel events: implications for visual recognition memory. Developmental Psychology 27 (1), 50–58.

Nelson, C.A., Collins, P.F., 1992. Neural and behavioral correlates of visual recognition memory in 4- and 8-month-old infants. Brain and Cognition 19 (1), 105–121. http://dx.doi.org/10.1016/0278-2626(92)90039-0.

Nelson, C.A.I., Thomas, K.M., de Haan, M., 2006. Neural bases of cognitive development. In: Kuhn, D., Siegler, R.S., Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology, Cognition, Perception, and Language, sixth ed., vol. 2. John Wiley & Sons Inc., Hoboken, NJ, USA, pp. 3–57.

Pascalis, O., de Haan, M., Nelson, C.A., 2002. Is face processing species-specific during the first year of life? Science (New York, NY) 296 (5571), 1321–1323. http://dx.doi.org/10.1126/science.1070223.

Passarotti, A.M., Paul, B.M., Bussiere, J.R., Buxton, R.B., Wong, E.C., Stiles, J., 2003. The development of face and location processing: an fMRI study. Developmental Science 6 (1), 100–117.

Paus, T., Collins, D., Evans, A., Leonard, G., Pike, B., Zijdenbos, A., 2001. Maturation of white matter in the human brain: a review of magnetic resonance studies. Brain Research Bulletin 54 (3), 255–266. http://dx.doi.org/10.1016/S0361-9230(00)00434-2.

Peña, M., Maki, A., Kovačić, D., Dehaene-Lambertz, G., Koizumi, H., Bouquet, F., Mehler, J., 2003. Sounds and silence: an optical topography study of language recognition at birth. Proceedings of the National Academy of Sciences of the United States of America 100 (20), 11702–11705. http://dx.doi.org/10.1073/pnas.1934290100.

Perone, S., Simmering, V.R., Spencer, J.P., 2011. Stronger neural dynamics capture changes in infants' visual working memory capacity over development. Developmental Science 14 (6), 1379–1392. http://dx.doi.org/10.1111/j.1467-7687.2011.01083.x.

Petanjek, Z., Judas, M., Simic, G., Rasin, M.R., Uylings, H.B.M., Rakic, P., Kostovic, I., 2011. Extraordinary neoteny of synaptic spines in the human prefrontal cortex. Proceedings of the National Academy of Sciences of the United States of America 108, 13281–13286.

Richmond, J., Nelson, C.A., 2007. Accounting for change in declarative memory: a cognitive neuroscience perspective. Developmental Review 27 (3), 349–373. http://dx.doi.org/10.1016/j.dr.2007.04.002.

Sanai, N., Nguyen, T., Ihrie, R.A., Mirzadeh, Z., Tsai, H.-H., Wong, M., Gupta, N., Berger, M.S., Huang, E., Garcia-Verdugo, J.M., Rowitch, D.H., Alvarez-Buylla, A., 2011. Corridors of migrating neurons in the human brain and their decline during infancy. Nature 478 (7369), 382–386. http://dx.doi.org/10.1038/nature10487.

Seress, L., Ábrahám, H., 2008. Pre- and postnatal morphological development of the human hippocampal formation. In: Nelson, C.A., Luciana, M. (Eds.), Handbook of Developmental Cognitive Neuroscience, second ed. MIT Press, Cambridge, MA, USA, pp. 187–211.

Sheridan, C., Draganova, R., Ware, M., Murphy, P., Govindan, R., Siegel, E.R., Eswaran, H., Preissl, H., 2010. Early development of brain responses to rapidly presented auditory stimulation: a magnetoencephalographic study. Brain and Development 32 (8), 642–657. http://dx.doi.org/10.1016/j.braindev.2009.10.002.

Thatcher, R.W., 1992. Cyclic cortical reorganization during early childhood. Brain and Cognition 20 (1), 24–50.

Tzourio-Mazoyer, N., De Schonen, S., Crivello, F., Reutter, B., Aujard, Y., Mazoyer, B., 2002. Neural correlates of woman face processing by 2-month-old infants. NeuroImage 15 (2), 454–461. http://dx.doi.org/10.1006/nimg.2001.0979.

Webb, S.J., Long, J.D., Nelson, C.A., 2005. A longitudinal investigation of visual event-related potentials in the first year of life. Developmental Science 8 (6), 605–616. http://dx.doi.org/10.1111/j.1467-7687.2005.00452.x.

Werker, J.F., Tees, R.C., 1983. Developmental changes across childhood in the perception of non-native speech sounds. Canadian journal of Psychology 37 (2), 278–286.

Cognitive Development: Mathematics Learning and Instruction

Clarissa A Thompson, University of Oklahoma, Norman, OK, USA
Julie L Booth, Temple University, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cognitive development has yielded many findings with implications for improving classroom instruction and student learning. Improving learning in mathematics is a priority for the United States. This article describes stumbling blocks that preclude children from showing optimal learning and generalization of mathematics skills including misconceptions about fractions and features of equations, like the equals sign. After identifying these learning challenges, fruitful interventions, such as 'just-in-time' feedback, that have improved learning outcomes are described. The article concludes with information about individual differences in stumbling blocks and the effectiveness of interventions.

The field of cognitive development has yielded many findings with implications for improving classroom instruction and student learning. Improving learning in mathematics is a priority for the United States: "International and domestic comparisons show that American students have not been succeeding in the mathematical part of their education at anything like a level expected of an international leader. Particularly disturbing is the consistency of findings that American students achieve in mathematics at a mediocre level by comparison to peers worldwide." (NMAP, 2008: p. xii).

Results from a recent study that linked findings from the 'Nation's Report Card,' the National Assessment of Educational Progress, and the Trends in International Mathematics and Science Study showed that six US states performed below the international average in mathematics (National Center for Education Statistics, n.d.). The National Mathematics Advisory Panel (NMAP, 2008: p. xii) further points out that, "A strong grounding in high school mathematics through Algebra II or higher correlates powerfully with access to college, graduation from college, and earning in the top quartile of income from employment." In the following sections, several barriers that impede optimal student learning in mathematics are identified and successful interventions are documented.

First stumbling blocks that preclude children from showing optimal learning and generalization of mathematics skills are described. Research on children's misconceptions about fractions and features of equations, like negative signs, is summarized. Then, the article focuses on instances where 'practice makes imperfect' because a faulty mental set gets entrenched. Finally, the authors detail how a common teaching practice – providing students with concrete instead of abstract examples – can introduce a barrier to generalization.

After these learning challenges are identified, fruitful interventions that have helped students overcome adversities are described. For example, the authors detail the ways targeted explicit and implicit feedback has improved children's estimation abilities and how students' algebraic word problem-solving abilities were improved when they viewed worked examples interleaved with problems they needed to solve. Additionally, promising findings from fun, gamelike interventions that have brought low-income students' numerical understanding up to par with their middle-income peers, are discussed. Finally, how stereotype threat has been successfully combatted when students receive lessons reminding them that intelligence is malleable is described.

The last section concludes with information about individual differences in stumbling blocks and the effectiveness of interventions. The authors focus on individual differences in use of addition strategies and cross-cultural differences in mathematics learning, as well as factors that influence the effectiveness of a well-established intervention – worked examples.

Stumbling Blocks that Preclude Optimal Learning in Mathematics

Misconceptions in Learning

The National Mathematics Advisory Panel (2008: p. 18) noted, "The most important foundational skill not presently developed appears to be proficiency with fractions (including decimals, percent, and negative fractions). The teaching of fractions must be acknowledged as critically important and improved before an increase in student achievement in Algebra can be expected." Subsequent research has confirmed this assertion, demonstrating that fraction knowledge is predictive of Algebra readiness, performance, and learning (Booth and Newton, 2012; Booth et al., 2014). In this section, common misconceptions that impede children's fractions understanding (for a recent review, see Siegler et al., 2013) and algebraic problem-solving abilities are discussed.

Both children and adults struggle to accurately represent the value of fractions (Opfer and DeVries, 2008; Thompson and Opfer, 2008). Siegler et al. (2011) hypothesized that faulty fractions understanding stems from overextending automatic processing of whole numbers to all other numbers, including fractions. However, characteristic features (Keil and Batterman, 1984) of whole numbers (i.e., that whole numbers have unique successors, include a finite number of entities within any given interval, can be expressed as a single symbol, etc.) are not defining features of all numbers. For example, there are an infinite number of fractions between the interval 0 and 1. However, the defining feature of all numbers is their possession of magnitudes that can be represented on a number line. Siegler

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92146-9

et al. (2011) found that sixth and eighth graders who transformed fractions into easier-to-handle formats (e.g., 1/3 is a small number because it is less than half) and subjectively segmented the number line (e.g., 2.5 is the halfway point on a 0–5 number line) made more accurate fraction estimates than children who did not use such effective strategies. As with whole numbers, tasks tapping fractions magnitude understanding correlated strongly with standardized achievement test and fractions arithmetic performance.

Siegler and Thompson (2014) found that experimenter-imposed landmarks can help, hinder, or fail to impact fractions estimates. Some objective landmarks (e.g., 0, 1, 2, 3, 4, and 5 in the 0–5 range) promote encoding of structurally important features of numbers and use of relevant strategies. For example, use of the mixed number strategy, translating 18/7 into 2 and 4/7, and placing this mixed number between 2 and 3 on a 0–5 number line, resulted in accurate number line estimates. However, other objective landmarks hindered students' estimation accuracy by promoting encodings and strategies that were not structurally important. For instance, children treated numerators and denominators as independent whole numbers when they estimated the location of fractions on a number line with quintile landmarks (0, 1/5, 2/5, 3/5, 4/5, and 1). A child using the independent components strategy might incorrectly match the numerator of the to-be-estimated fraction, 3/8, to the landmark with the matching numerator, 3/5. Other landmarks, like 1/2 in the 0–1 range, were redundant with children's processing, and these objective landmarks did not prove more or less advantageous than a number line without objective landmarks. For example, children often subjectively segmented 0–1 lines without objective landmarks at the halfway point and considered whether the to-be-estimated fraction was greater than or less than 1/2.

Middle school students also hold persistent misconceptions about critical features of algebraic equations (Booth and Koedinger, 2008). For instance, students lack conceptual understanding of the equals sign and believe that it means that an operation should be performed, or an answer should be found, rather than exhibit an understanding that one side of the equation is equivalent to the other side. Booth and Koedinger (2008) found that students often perform an operation on one side of the equation, but fail to carry out the operation on the other side. Another common misconception that impacts algebraic problem solving is the meaning of negative signs. This misconception was best illustrated in an in-depth case study where one student claimed that $-4x + 3$, $4x + 3$, and $4x - 3$ were equivalent expressions. This student failed to realize that the negative sign modifies the term to which it is tied and moving the negative sign changes the meaning of the equation. A more recent study revealed that students with these types of misconceptions not only have difficulty solving equations but also fail to encode the equations correctly (Booth and Davenport, 2013).

In summary, children maintain widespread misconceptions about mathematics. For instance, children overextend what they know about whole numbers and mistakenly assume these characteristic features are defining features of all numbers, including fractions. Also, children misunderstand the meaning of the equals and negative sign; these misconceptions are often detrimental to children's Algebra performance (**Table 1**).

Table 1 Summary of misconceptions in learning

Finding	*Example*
Children overextend what they know about whole numbers to fractions.	When asked to estimate 3/19 on a 0–1 number line, a child may place this fraction closer to 1 than to 0 and provide the explanation that 19 is a big number.
Children misunderstand the meaning of the equals sign and negative sign.	Children often believe that the equals sign means to carry out an operation, and do not realize that it means that one side of the equation should be equivalent to the other. Further, children do not realize that the negative sign is tied to an algebraic term, and moving the negative sign modifies the entire meaning of the equation.

Mindlessness and Automaticity

The old adage, 'practice makes perfect,' speaks to the beneficial impact of automaticity. However, research reviewed in this section suggests that automaticity can be detrimental in some learning contexts, like mathematical equivalence problems, but children's gesture–speech mismatches can identify those children on the brink of a breakthrough in understanding.

Children are drilled on simple arithmetic problems of the form: $2 + 3 + 4 =$ ___. This extensive practice may lead children to misinterpret the meaning of the equals sign (e.g., carry out an operation). Later, when children are asked to find the answer to a problem like $2 + 3 + 4 =$ ___ $+ 4$, they may mistakenly answer 9, because they added all numbers on the left side of the equation, or 13, because they added all numbers on both sides of the equation. Interestingly, U-shaped development is evident when 7-year-olds outperform 9-year-olds on equivalence problems (McNeil, 2007). Older children do worse on these equivalence problems because of their resistance to change; they have practiced the simple equations to mastery and this extensive practice makes it very difficult to inhibit adding everything to the left of the equals sign and placing that number in the blank on the right side of the equation. Subsequent research (McNeil, 2008) found that children who practiced typical problems ($15 + 13 = 28$) produced fewer correct responses on a follow-up equation-solving task ($5 + 4 + 7 =$ ___ $+ 3$) than did those children who completed training on atypical problems ($28 = 28$) that focused on the conceptual meaning of the equals sign (e.g., the left and right sides of the equation must be equivalent). McNeil suggested the results were a product of 'mindlessness' whereby students relied on strategies that had worked for them in the past but failed to recognize that these strategies were not appropriate for the task at hand.

Children on the cusp of learning often show variability in their performance (Siegler, 1994). This variability is evidenced in gesture–speech mismatches on equivalence problems (Perry et al., 1988). For instance, even though a child may provide the incorrect answer of 9 on the problem $2 + 3 + 4 =$ ___ $+ 4$, a gesture–speech mismatch may occur when the same child simultaneously points to the 2 and 3 on the left of the equation and the blank on the right of the equation indicating she has an implicit understanding that 2

and 3 can be added together and correctly inserted in the blank on the right (Goldin-Meadow, 2001).

Similarly, participants may rely on inappropriate mental sets (Luchins and Luchins, 1950), or *Einstellungsmuster*, when they attempt to transfer what they have learned about numbers from one context to the next. Opfer and Thompson (2008) tested whether uncorrected practice on a pretest might result in entrenched, inappropriate numerical representations. One group of participants completed a number categorization pretest in which children were asked to identify which category – really small, small, medium, big, or really big – best described particular numbers in the 0–1000 range, then received training on the correct locations of numbers on number lines, and finally completed an identical number categorization posttest. Other children received training on a number line estimation task and then completed a number categorization posttest. The group that completed a pretest did not receive any feedback on the correctness of their category judgments, so they did not know that they were practicing an inappropriate representation for the task. Even when they subsequently received feedback on a related estimation task during training, participants failed to transfer the more appropriate representation to the categorization posttest. That is, when these pretested participants completed a posttest identical to their uncorrected pretest, their initial, albeit inaccurate, representation was activated and led to diminished transfer of learning.

In summary, at times 'practice makes imperfect' when children's numerical representations become entrenched with extensive practice and children mistakenly generalize an overpracticed representation to a context in which it is not relevant (**Table 2**).

Table 2 Summary of mindlessness and automaticity

Finding	*Example*
Children make mistakes on mathematics equivalence problems because they cannot inhibit their overly practiced responses and mindlessly rely on strategies that are not appropriate for the task at hand.	Children mistakenly fill in the blank with 9 or 13 in the following problem: $2 + 3 + 4 = ___ + 4$.
Children show gesture–speech mismatches in mathematics equivalence problems indicating that they have some implicit understanding of how to solve the problem even if they cannot explicitly answer it.	Children verbally provide the wrong answer in the above equivalence problem, but they may simultaneously point to the 2 and 3 on the left side of the equation and the blank on the right side of the equation.
Inappropriate representations can become entrenched with practice.	Those children who practiced an inappropriate representation while completing an unsupervised numeric category judgment pretest failed to transfer a newly learned representation to an equivalent numeric categorization posttest.

Concrete versus Abstract Instantiations

Instructors frequently provide students with concrete examples to make abstract concepts more accessible and meaningful. For instance, the Institute of Education Sciences (IES) Fractions Practice Guide (Siegler et al., 2010) suggests that teachers should build on their students' naive understanding of sharing and proportionality to promote early fractions understanding. However, could there be instances in which concrete examples hinder learning and transfer to new situations? In this section, the authors summarize views in favor of the use of concrete examples, such as Piaget's constructivist view of knowledge acquisition, and the recommendation by the National Council of Teachers of Mathematics to incorporate concrete examples into the curriculum. Then, the authors summarize research that suggests concrete instantiations may inhibit optimal transfer, and usefulness of concrete examples might depend on the skill level of learners.

Piaget's (1960) theory of cognitive development maintains that active manipulation of physical objects is important for children's learning in the concrete operational stage of development; from these experiences, children are thought to be able to garner more abstract concepts (Kamii, 1974). Consistent with this, manipulatives, or concrete objects designed to help children learn mathematics (Uttal et al., 1997), are used widely in elementary school classrooms. However, research on the effectiveness of using manipulatives (e.g., counters for simple arithmetic, Dienes blocks for place value) or other concrete instantiations (hundreds squares for arithmetic with larger numbers, a pizza cut into multiple equivalent pieces to represent fractions) in the classroom has had mixed results, with some findings asserting they are beneficial (e.g., Sowell, 1989) and others suggesting no value added (e.g., Friedman, 1978).

The idea of bridging instruction from concrete to abstract instantiations does not end in elementary school, and the technique is supported by a number of empirical studies (e.g., Moreno and Mayer, 1999; Goldstone and Son, 2005). In fact, the NCTM (2000) recommends that teachers include multiple forms of representations – including concrete examples – when teaching mathematical concepts.

However, recent research calls into question whether concrete instantiations best support student learning. Undergraduate students, who were given either concrete or symbolic (abstract) representations, learned basic properties of number equally, but those studying abstract representations displayed improved transfer to novel problem-solving contexts (Kaminski et al., 2008). This study suggests that concrete instantiations may constrain learners' understanding of a concept to the presented context, thus hindering transfer.

Similarly, diagrams are often successfully used to facilitate understanding of text (Mayer, 2005), and combining them with algebraic word problems supports problem-solving performance for older and higher ability middle school students (Booth and Koedinger, 2012). However, the same representations may actually hinder younger, lower ability students, who show greater problem-solving success with textual problems alone (Booth and Koedinger, 2012). Younger and lower ability students are often unable to correctly map

diagram components to the relevant pieces of text and become increasingly confused by the problems.

In summary, not every student will universally reap the benefits of concrete examples. These and other studies caution that the relative benefit of concrete and abstract representations may vary with characteristics of the user, the representation, or the goals of the task (Ainsworth, 2006; Larkin and Simon, 1987) (**Table 3**).

Table 3 Summary of concrete versus abstract instantiations

Finding	*Example*
Adults may learn better from abstract instantiations than concrete ones.	Undergraduate students learning about number properties from abstract instantiations were better able to translate their knowledge to solve real-world problems.
Although much emphasis is placed on using concrete instantiations to help children learn, children do not always learn better from concrete instantiations.	When solving algebraic word problems, younger middle school students and those with less prior knowledge succeed more when a corresponding diagram is not provided.

Interventions that Improve Mathematics Learning Outcomes

Explicit and Implicit Feedback

Is it possible for children to learn as much from direct instruction as from discovery learning? In this section, the relative benefits of explicit and implicit feedback to improve children's estimation skills are investigated.

According to Piaget's (1970: p. 715) constructivist approach, "each time one prematurely teaches a child something he could have discovered for himself, that child is kept from inventing it and consequently from understanding it completely." Contrary to Piaget's view, Klahr and Nigam (2004) claim, 'what is learned is more important than how it was taught.' For instance, students who received direct instruction and answered questions that probed their knowledge were more successful at learning to create unconfounded experiments by using the Control of Variables Strategy (CVS, Chen and Klahr, 1999) than peers who attempted to discover this principle on their own (p. 662). Most importantly, direct instruction allowed children to generalize the CVS strategy to analogous contexts.

The question remains, however, are explicit and implicit direct instruction equally effective at bringing about conceptual change and facilitating problem-solving skill? Research investigating children's and adults' analogical problem solving suggests that providing explicit hints or reminding participants to draw parallels between key components of isomorphic problems leads to improved accuracy (Gick and Holyoak, 1980; Brown et al., 1986). NMAP (2008) recommended that it is particularly important for low-achieving students and children diagnosed with mathematics disabilities to receive frequent direct instruction. This direct instruction might take the form of explicit feedback about the accuracy of students' performance, discussion of the appropriate strategies for solving problems, and encouragement to ask and answer questions and talk aloud as students attempt to solve mathematics problems. Participants can also benefit from implicit instruction when various example problems are aligned to facilitate surface-level comparisons (Kotovsky and Gentner, 1996). This comparison process allows participants to move beyond perceptual similarities and recognize commonalities at a more abstract, conceptual level (Gentner et al., 2003).

In the domain of numerical cognition, children also benefit from explicit and implicit instruction aimed at improving their estimation abilities. Children and adults know and use multiple representations (e.g., linear and logarithmic) depending on the numerical context at hand (Siegler and Opfer, 2003; Siegler et al., 2009; see Barth and Paladino, 2011 for an opposing view and Opfer et al., 2011 for a meta-analysis). That is, although second graders accurately estimated the location of 75 on a 0–100 number line, these same children tended to wildly overestimate the location of 75 on a 0–1000 number line. This pattern of overestimation of relatively small numbers and failure to differentiate the locations of relatively large numbers characterizes the logarithmic representation. With increasing age, children exhibit a logarithmic-to-linear shift as they begin producing more accurate, adultlike, estimates that are equally spaced across the number line. This conceptual change across development unfolds in a piecemeal manner; with age, children gradually scale up their linear representation from smaller to larger numerical contexts (Siegler and Opfer, 2003; Siegler and Booth, 2004; Siegler et al., 2009, 2010).

Children who produce more linear estimates show better performance on standardized achievement tests (Booth and Siegler, 2006); they also display greater accuracy in recalling numerical information presented in meaningful vignettes (Thompson and Siegler, 2010; Thompson and Opfer, 2011) and answers to novel arithmetic problems (Booth and Siegler, 2008; Siegler and Ramani, 2009). These associations speak to the importance of developing a linear representation of numerical magnitudes, and have prompted investigation of the usefulness of targeted interventions to improve children's estimation abilities. To this end, Opfer and Siegler (2007) provided children with feedback about their estimates in the 0–1000 range. In this intervention, children learned that they had greatly overestimated the location of 150, the maximally discrepant point between a logarithmic and linear function forced to pass through 0 and 1000. On receiving one trial of corrective feedback, children abruptly adopted a linear representation as they estimated the location of subsequent numbers in the 0–1000 range.

However, explicit feedback does not always lead to positive learning outcomes. Children who received explicit feedback on a whole number estimation task inappropriately transferred their newly learned linear representation to an estimation task with unit fractions (e.g., those with a numerator of 1: Opfer and DeVries, 2008; Thompson and Opfer, 2008). Both children and adults failed to represent the fractional magnitudes (e.g., 1/60 is a very small number) and instead only represented the denominators as independent whole numbers (e.g., 60). While use of a linear representation optimizes accuracy when estimating whole numbers (e.g., 60 is located closer to 1 than to 1440), the linear representation does not optimize accuracy when used to estimate the location of unit fractions (e.g., 1/60 placed nearer to 1/1 than to 1/1440). This cognitive illusion

likely stems from children's and adults' familiarity and extensive practice with whole numbers.

Implicit feedback, like explicit feedback, can cause representational change. Thompson and Opfer (2010) used progressive alignment methodology (Kotovsky and Gentner, 1996) to highlight the underlying structural similarity of magnitudes within the numerical decimal system. A small, familiar range (0–100) was aligned with large, unfamiliar ranges (0–1000, 0–10 000, and 0–100 000) prompting second graders to draw comparisons across surface-level features (e.g., number of zeros in the 0–100 and 0–1000 ranges). These comparisons highlighted an analogy (e.g., 15:100::150:1000) that could be drawn between small and large ranges and was not as evident to children who did not observe the alignment of small and large ranges. The progressive alignment group produced more accurate estimates in large numerical ranges (0–10 000 and 0–100 000); this representational change persisted on a posttest without alignment.

In summary, numerical cognition research suggests that children's estimation performance can be improved with explicit as well as implicit feedback, and this targeted experience prompts children to exhibit adultlike estimation abilities. However, under some circumstances children may inappropriately overgeneralize corrective feedback (**Table 4**).

Table 4 Summary of explicit and implicit feedback

Finding	*Example*
Children's numerical representations have been improved via explicit feedback.	Children who produced an inaccurate, logarithmic series of estimates abruptly switched to producing an accurate, linear series of estimates after receiving corrective feedback on the location of 150 in the 0–1000 numeric range.
Children's numerical representations have been improved via implicit feedback.	Children who saw 0–100 number lines aligned with 0–1000 number lines learned to produce a more accurate series of estimates by drawing a relevant analogy (e.g., 15:100::150:1000).

Learning from Worked Examples

Attempting to solve long strings of problems is not the optimal method for having students hone their problem-solving skills. While practicing, students should also study examples that illustrate how to solve problems (Sweller and Cooper, 1985; Sweller, 1999). In this section, how examples have been used to target critical skills, such as procedural fluency, conceptual understanding, and problem-solving flexibility, is summarized.

Using examples in the traditionally suggested way, by replacing approximately half of practice problems in a set with fully worked-out examples for students to study (Sweller, 1999), improves students' procedural knowledge, largely because it reduces students' cognitive load and allows them to focus on learning the steps, and underlying meaning of those steps, necessary to understand a problem (Zhu and Simon, 1987; Sweller, 1999). Scaffolding has also been used to improve students' procedural fluency by having students study complete examples and then gradually fading the completeness of the examples throughout the practice session (Renkl et al., 2000).

Another line of research, focused more on conceptual understanding, suggests that students should be asked to *explain* examples, not just study them. Such self-explanation is thought to improve learning by forcing students to make their newly acquired concepts explicit (Chi, 2000; Roy and Chi, 2005). While most worked-example research focuses on the use of correct examples, recent work suggests that a combination of correct and incorrect examples (with prompts for self-explanation) can be even more effective (Siegler, 2002; Rittle-Johnson, 2006; Siegler and Chen, 2008). Explaining errors can be beneficial because it helps struggling students accept that they have chosen incorrect procedures and realize why those strategies are inappropriate. For instance, consider a student attempting to solve $4x - 3 = 6$ by subtracting 3 from both sides, and ending up with $4x = 3$. When the student explains how the procedure used led to the incorrect answer, it forces the student to notice that the negative sign before the 3 in $4x - 3 = 6$ is the feature that makes the procedure inappropriate. Consistent with this assertion, explaining incorrect examples has been shown to lead to improvements in conceptual understanding in Algebra (Booth et al., 2013).

Examples can also be used to improve students' flexibility in problem solving, or their ability to recognize more than one way to solve a problem and choose the most effective one (Star and Seifert, 2006; Rittle-Johnson and Star, 2007). For this purpose, directly comparing and contrasting multiple examples of how to correctly solve a particular problem can be more effective than studying such examples sequentially (Rittle-Johnson and Star, 2007). While introduction of multiple solution methods has been shown to benefit even struggling students (Newton et al., 2010), the contrasting cases method may be most useful for students who already have achieved an adequate level of prior knowledge in the domain (Rittle-Johnson et al., 2009).

In summary, interleaving worked examples with practice problems improves children's procedural understanding of the steps necessary to solve the problems, explaining correct and incorrect answers in worked examples improves children's underlying conceptual knowledge, and comparing and contrasting simultaneously presented worked examples can show students that multiple strategies can be used to solve given problems (**Table 5**).

Table 5 Summary of learning from worked examples

Finding	*Example*
Studying worked examples promotes students' procedural learning.	Students who study worked examples during their practice sessions learn more of the content or learn the same amount faster.
Explaining worked examples, especially incorrect ones, promotes students' conceptual understanding.	Students who explained common errors for solving algebraic equations showed greater gain in conceptual knowledge of the features in the problems.
Comparing and contrasting worked examples promotes flexibility in problem solving.	Studying contrasting cases of two ways to solve the same problem leads students to recognize multiple solution paths and adopt more efficient solution methods.

Learning via Games and Cognitive Tutors

Academic engagement is the degree to which students are attentive, interested, and invested in learning (Marks, 2000). Not surprisingly, student engagement is positively related to achievement; a lack of engagement is typically detrimental to learning (Finn, 1989). Much research has been focused on finding ways to make instruction more engaging to improve learning in a variety of domains. In this section, learning gains resulting from children engaging with numerical board games and Cognitive Tutors are reviewed.

One of the ways researchers have tried to increase engagement in mathematics is to use games to introduce and strengthen core math skills. For example, the Number Worlds' curriculum uses interactive games to strengthen preschoolers' number sense by providing experience with different facets of number (e.g., quantities, counting, formal notation) and opportunities to understand the relations between those facets (Griffin, 2004; Griffin et al., 1994). Use of this curriculum has improved learning and achievement for at-risk children; its use in the RightStart program led to gains that lasted a year (Griffin et al., 1994). More recently, research by Siegler and colleagues (Siegler and Ramani, 2008, 2009; Ramani and Siegler, 2008, 2011) has shown that children from low-income homes can catch up with middle-income age mates after playing a linearly spaced numeric board game, akin to Chutes and Ladders, for about an hour. This experience is thought to reinforce numerical magnitude understanding across multiple modalities (visual, auditory, tactile, etc.), and the resulting improvements in numerical magnitudes' understanding remain evident more than 2 months later. Given the mounting evidence for the importance of magnitude knowledge for learning in math (e.g., Booth and Siegler, 2008), the use of games as instructional tools has the potential to help young children gain the foundations necessary for later success in mathematics.

Another instructional method that has been used to increase student engagement and learning is the use of intelligent tutors to provide individualized instruction for students (Anderson et al., 1985). One such tool, the Cognitive Tutor, was originally developed to improve student learning in Algebra (Koedinger et al., 1997), but has since been developed for other courses, including Pre-Algebra, Algebra II, and Geometry (see www.carnegielearning.com). The Cognitive Tutor allows students to work through Algebra I content at their own pace while providing just-in-time feedback and hints as needed during practice sessions. For example, if a student chooses to begin solving the problem $3x - 5 = 1$ by subtracting 5 from both sides, the step will turn red and a pop-up message may appear, pointing out that the 5 is already negative, so it should not be subtracted from both sides. The Tutor also tracks student learning across problems, monitoring the student's strengths and weaknesses and providing appropriate types of problems until the student achieves mastery on necessary skills (Koedinger et al., 1997). Evidence suggests this approach is superior to traditional methods of mathematics instruction and practice (Ritter et al., 2007), and has been identified by the What Works Clearinghouse (2009) as a potentially effective method for middle school math instruction.

In summary, researchers have created fun ways for children to engage in mathematics curricula via games and Cognitive Tutors. Numerical board game interventions bring low-socioeconomic status (SES) children's numerical understanding up to par with that of their middle-income age mates. Further, Cognitive Tutor's 'just-in-time' feedback has been proved to be more effective than traditional classroom lessons where individualized and targeted feedback is not always present (**Table 6**).

Table 6 Summary of learning via games and cognitive tutors

Finding	*Example*
Numerical board games enhance children's numerical understanding.	Playing a numerical board game for just 1 h closed the gap in numerical knowledge between low-SES and middle-SES children.
Individualized instruction from intelligent tutors promotes learning.	Algebra students enrolled in a Cognitive Tutor Algebra 1 course learn more over the course of the year than students in traditional classrooms.

Interventions that Combat Stereotype Threat

Most of the factors impacting learning discussed here are cognitive aspects of the learner and interventions aimed at facilitating student comprehension of content material. However, psychosocial factors are also important in predicting whether an individual will learn and succeed. In this section, we detail ways stereotype threat may substantially impact student performance, especially in mathematics and other Science, Technology, Engineering, and Mathematics (STEM) domains (NMAP, 2008).

Research suggests that stereotype threat, or the increased anxiety and decreased performance by members of groups, which are stereotypically considered to perform poorly in a given domain, is induced when the individual's race (or gender, SES, etc.) is made salient just prior to assessment (Steele and Aronson, 1995; Croizet and Claire, 1998). However, more recent research suggests that stereotype threat hinders learning, even though the characteristic is not made particularly salient, and may be the default mode for stereotyped individuals when approaching mathematics tasks (Good et al., 2008).

Thus, reduction of stereotype threat may not be as easy as ensuring that students are not reminded of stereotypes. However, it can be combatted through measures to change the way students think about themselves and their potential to succeed. For example, one type of intervention focuses on helping students understand that intelligence is malleable – that individuals do not have a fixed amount of intelligence. One reason that stereotype threat may be detrimental to students is that it makes them feel that their innate abilities are being questioned. Students who believe that these innate abilities are limited and fixed are often anxious about being assessed on those abilities, and may become less motivated to succeed compared with students who believe intelligence

can be improved (Blackwell et al., 2007). Explicitly encouraging students to believe that intelligence is malleable can lead to improved standardized test scores (Good et al., 2003) and ease the transition to middle school (Blackwell et al., 2007).

Other interventions capitalize on the idea that a sense of belonging is necessary for success in a domain (Steele, 1997), and attempt to counteract the very message conveyed in stereotypes – that members of the stereotyped group simply do not belong in that domain (Steele, 1997). Convincing students that anyone can belong in mathematics classes, not just those who are typically expected to be successful at math, can have positive effects on students' motivation, engagement, and achievement (Good et al., 2012).

In summary, some initial interventions aimed at eradicating the effects of stereotype threat have been successful. These interventions aimed to show students that intelligence is malleable and to convince them that all students belong in and can succeed in mathematics classes (**Table 7**).

Table 7 Summary of interventions that combat stereotype threat

Finding	*Example*
For students in underrepresented minority groups, belief in the malleability of intelligence is important for achievement.	Teaching students that working hard can increase one's intelligence in math led to improvements in math achievement, especially in middle school.
For students in underrepresented minority groups, a feeling of belonging in a class or domain is important for achievement.	Convincing students that anyone can belong in math class led to improvements in math achievement, as well as in students' engagement and motivation to succeed.

Mathematical Stumbling Blocks and Effective Interventions May Differ from Student to Student

Individual Differences in Learning Mathematics

Not every child solves every mathematics problem in the same way on every trial in every context. In this section, individual differences in learning mathematics are discussed including cross-cultural differences in estimation, memory, and understanding of fractions, and strategy use in basic arithmetic.

Children in the United States lag behind their East Asian peers in many mathematical tasks. Research investigating what underlies this learning gap has shown that even before Chinese children receive formal schooling, they already outperform their US peers on counting and arithmetic tasks (Geary et al., 1993; Geary et al., 1996), likely because Chinese parents provide their preschoolers with extensive practice on these skills. Interestingly, Chinese kindergartners not only excelled at a practiced task, arithmetic of single-digit numbers, but also at an unpracticed task, estimating the location of numbers in the 0–100 range (Siegler and Mu, 2008). For both US and Chinese children, performance on the arithmetic and estimation tasks was strongly correlated. Siegler and Mu (2008) hypothesized that Chinese children received greater practice with counting fingers and sets of objects and this led to the accumulation of kinesthetic, visual, verbal, and temporal cues that mutually reinforced the magnitudes of numbers. Further, Chinese children tested early in the academic year were better able to identify Arabic numerals and recall and recognize numbers than their US peers (Thompson et al., unpublished data). US children tested at the end of their kindergarten year performed equivalently to the early year Chinese children on estimation, numeral identification, recall, and recognition. Basic memory abilities cannot explain the early year differences, as Chinese and US children recalled a list of words equally well.

Can cross-cultural differences in fractions understanding be explained by the hypothesis of linguistic determinism? Miura et al. (1999) claimed that the transparency of part–whole relations in Korean fractions names (e.g., of four parts, one) accounted for Korean children's optimal performance on a fractions identification task. However, Paik and Mix (2003) showed that neither Korean nor US children performed at above chance levels when particular foils were included. That is, if children were asked to choose a visual representation of 2/3, both groups were likely to choose incorrectly on the basis of the independent components of the fraction (2 parts shaded dark and 3 parts shaded light instead of 2 parts shaded dark and one part shaded light). However, when these independent component foils were not included, US children who received English language translations of the part–whole relations actually outperformed Korean children on the fractions identification task, likely because the Korean children were not as familiar with the formal term (jo-gak) for parts. Subsequent work by Mix and Paik (2008) showed that when a more familiar word for parts (boon) was used for the Korean translation, Korean children made accurate fractions identifications.

Individual differences are present in arithmetic performance, too. Children use multiple strategies to solve simple addition problems (Siegler and Shrager, 1984), choosing among the strategies in adaptive ways (Siegler, 1996). For example, while retrieval of simple addition facts often leads to the most accurate and least time-consuming responses, sometimes children use less adaptive problem-solving strategies, such as counting on their fingers, counting from the largest addend, or decomposing the problems into more manageable forms.

Siegler (1988) highlighted individual differences in first graders' arithmetic performance noting that students could be characterized as 'good,' 'not so good,' and 'perfectionists.' Good students made quicker and more accurate responses, performed better on standardized achievement tests, and showed more mastery of retrieval and other backup addition strategies than did not-so-good students. Interestingly, perfectionists were very accurate and did well on standardized achievement tests; however, they only used retrieval when they were very sure of their answers and often reverted to less efficient backup strategies to answer problems when they were less confident. Perfectionists effectively retrieved answers, although they used the strategy less often than did not-so-good students. More recent research has shown that children diagnosed with mathematical disabilities

knew fewer addition facts and executed addition strategies less effectively (e.g., committed more retrieval errors) than did typically achieving children (Geary et al., 2007).

In summary, significant cross-cultural differences exist in mathematics, at least in part because of varied amounts of practice with counting and arithmetic. Even individual children in the United States know and use many different strategies for solving simple arithmetic problems. Students do not always choose the most optimal strategy to solve a problem; this is especially true for children diagnosed with a mathematical disability (Table 8).

Table 8 Summary of individual differences in learning mathematics

Finding	*Example*
Cross-cultural differences in whole numbers understanding exist.	Chinese children produce more accurate estimates in the 0–100 range and remember numbers within this range more accurately, too.
Cross-cultural differences in fractions understanding exist.	The Korean language may make fractions more transparent (e.g., of four parts, one).
Children know and use various strategies to answer simple addition problems.	Children may answer the problem 3 + 4 by employing various strategies such as retrieval, counting on from the smallest or largest addend, counting their raised fingers, or decomposing the problem into a more manageable form. Children who have mathematical disabilities often make more mistakes than typically developing children when executing addition strategies.

Individual Differences in Intervention Effectiveness

In light of individual differences in students' prior knowledge and abilities, it is not surprising that the benefit of instructional interventions also varies for students with different backgrounds. In this section, three instances are described where the effectiveness of worked examples depends on characteristics of the student experiencing the intervention.

Perhaps the most intuitive of these findings is a case in which the effectiveness of the presented worked examples depended on the ability level of the learner. Grosse and Renkl (2007) presented college students, who possessed varying levels of prior knowledge about probability, with correct and incorrect worked examples. For correct examples, students explained what procedure was used and why it was correct; in incorrect examples, students located the error and fixed it. This combination led to improved learning for students with sufficient prior knowledge. However, students with low prior knowledge did not benefit from inclusion of incorrect examples. Further results from this study suggested that highlighting the portion of the problem containing the error increased the benefit for low-ability learners.

Another instance in which students' prior knowledge affects the degree to which they benefit from instruction shows the opposite effect. Kalyuga et al. (2001) demonstrated that inexperienced students benefited more from studying worked examples than did higher ability students. This type of finding is referred to as an 'expertise reversal effect,' in which interventions that provide instructional guidance are effective for low-knowledge learners who may not have internal schemas activated at the time, but are less effective or even harmful for more experienced learners who become overloaded by simultaneous activation of internal and external schemas (Kalyuga et al., 2003).

Finally, Booth et al. (under review) demonstrate that the benefit of worked examples may vary for students with different ethnic backgrounds, regardless of prior ability level. In their study, minority (Black/Latino) students benefited more than nonminority (White/Asian) students from explaining correct and incorrect examples while practicing solving equations. An achievement gap between minority and nonminority students existed at pretest, and was closed at posttest. However, both high- and low-ability minority students benefited more from the inclusion of worked examples than from problem-solving practice alone, suggesting that the differential benefit was not due to cognitive differences between the groups. In this particular case, incorrect examples were designed to facilitate correct encoding of problem features and to illustrate that errors are an important part of the learning process, perhaps decreasing stereotype threat and increasing minority students' self-efficacy for the task.

In summary, even for interventions that improve learning, individual differences in student characteristics impact their effectiveness. It is clearly important for researchers and practitioners to familiarize themselves with typical stumbling blocks inherent in mathematics learning and interventions that successfully overcome them. However, it is also crucial to be aware that individual differences in cognitive and motivational factors (e.g., self-efficacy) may influence the degree to which students will be able to benefit from the intervention and overcome their stumbling blocks (Table 9).

Table 9 Summary of individual differences in intervention effectiveness

Finding	*Example*
The effectiveness of instruction varies based on students' prior knowledge.	High-ability undergraduates learned more from finding and fixing algebraic mistakes; however, lower ability undergraduates learned more from studying worked examples.
The effectiveness of instruction varies based on students' motivation for learning.	Explaining correct and incorrect examples was more beneficial for students from underrepresented minority groups who held lower beliefs about their own self-efficacy.

See also: Computer-Assisted Instruction; Cross-Cultural Study of Education; Gender and Education; Intelligence, Prior Knowledge, and Learning; Mathematics Education; School Learning for Transfer; Socioeconomic Status and Academic Achievement.

Bibliography

Ainsworth, S., 2006. DeFT: a conceptual framework for considering learning with multiple representations. Learning and Instruction 16 (3), 183–198.

Anderson, J.R., Boyle, C.F., Reiser, B.J., 1985. Intelligent tutoring systems. Science 228, 456–462.

Barth, H., Paladino, A.M., 2011. The development of numerical estimation: evidence against a representational shift. Developmental Science 14, 125–135.

Blackwell, L.S., Trzesniewski, K.H., Dweck, C.S., 2007. Implicit theories of intelligence predict achievement across an adolescent transition: a longitudinal study and an intervention. Child Development 78, 246–261.

Booth, J.L., Davenport, J.L., 2013. The role of conceptual encoding and feature knowledge in algebraic equation-solving. Journal of Mathematical Behavior 32, 415–423.

Booth, J.L., Koedinger, K.R., 2008. Key misconceptions in algebraic problem solving. In: Love, B.C., McRae, K., Sloutsky, V.M. (Eds.), Proceedings of the 30th Annual Cognitive Science Society. Cognitive Science Society, Austin, TX, pp. 571–576.

Booth, J.L., Koedinger, K.R., 2012. Are diagrams always helpful tools? Developmental and individual differences in the effect of presentation format on student problem solving. British Journal of Educational Psychology 82, 492–511.

Booth, J.L., Lange, K.E., Koedinger, K.R., Newton, K.J., 2013. Using example problems to improve student learning in algebra: differentiating between correct and incorrect examples. Learning and Instruction 25, 24–34.

Booth, J.L., Newton, K.J., 2012. Fractions: could they really be the gatekeeper's doorman? Contemporary Educational Psychology 37, 247–253.

Booth, J.L., Newton, K.J., Twiss-Garrity, L., 2014. The impact of fraction magnitude knowledge on algebra performance and learning. Journal of Experimental Child Psychology 118, 110–118.

Booth, J.L., Oyer, M.H., Paré-Blagoev, J.E., Barbieri, C., Elliot, A., Augustine, A.A., Koedinger, K.R., (under review). Learning algebra by example in real-world classrooms. Journal for Research on Educational Effectiveness.

Booth, J.L., Siegler, R.S., 2006. Developmental and individual differences in pure numerical estimation. Developmental Psychology 41, 189–201.

Booth, J.L., Siegler, R.S., 2008. Numerical magnitude representations influence arithmetic learning. Child Development 79, 1016–1031.

Brown, A.L., Kane, M.J., Echols, K., 1986. Young children's mental models determine analogical transfer across problems with a common goal structure. Cognitive Development 1, 103–122.

Chen, Z., Klahr, D., 1999. All other things being equal: acquisition and transfer of the control of variables strategy. Child Development 70, 1098–1120.

Chi, M.T.H., 2000. Self-explaining expository texts: the dual processes of generating inferences and repairing mental models. In: Glaser, R. (Ed.), Advances in Instructional Psychology. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 161–238.

Croizet, J.C., Claire, T., 1998. Extending the concept of stereotype threat to social class: the intellectual underperformance of students from low socioeconomic backgrounds. Personality and Social Psychology Bulletin 24, 588–594.

Finn, J.D., 1989. Withdrawing from school. Review of Educational Research 59, 117–142.

Friedman, M., 1978. The manipulative materials strategy: the latest pied piper? Journal for Research in Mathematics Education 9, 78–80.

Geary, D.C., Bow-Thomas, C.C., Fan, L., Siegler, R.S., 1993. Even before formal instruction, Chinese children outperform American children in mental addition. Cognitive Development 8, 517–529.

Geary, D.C., Bow-Thomas, C.C., Liu, F., Siegler, R.S., 1996. Development of arithmetical competencies in Chinese and American children: influence of age, language, and schooling. Child Development 67, 2022–2044.

Geary, D.C., Hoard, M.K., Byrd-Craven, J., Nugent, L., Numtee, C., 2007. Cognitive mechanisms underlying achievement deficits in children with mathematical learning disability. Child Development 78, 1343–1359.

Gentner, D., Loewenstein, J., Thompson, L., 2003. Learning and transfer: a general role for analogical encoding. Journal of Educational Psychology 95, 393–408.

Gick, M.L., Holyoak, K.J., 1980. Analogical problem solving. Cognitive Psychology 12, 306–355.

Goldin-Meadow, S., 2001. Giving the mind a hand: the role of gesture in cognitive change. In: McClelland, J., Siegler, R.S. (Eds.), Mechanisms of Cognitive Development: Behavioral and Neural Perspectives. Earlbaum Associates, Mahwah, NJ, pp. 5–31.

Goldstone, R.L., Son, J.Y., 2005. The transfer of scientific principles using concrete and idealized simulations. The Journal of the Learning Sciences 14, 69–110.

Good, C., Aronson, J., Harder, J.A., 2008. Problems in the pipeline: stereotype threat and women's achievement in high-level math courses. Journal of Applied Developmental Psychology 29, 17–28.

Good, C., Aronson, J., Inzlicht, M., 2003. Improving adolescents' standardized test performance: an intervention to reduce the effects of stereotype threat. Journal of Applied Developmental Psychology 24, 645–662.

Good, C.D., Rattan, A., Dweck, C.S., 2012. Why do women opt out? Sense of belonging and women's representation in mathematics. Journal of Personality and Social Psychology 102, 700–717.

Griffin, S., 2004. Building number sense with number worlds. Early Childhood Research Quarterly. 19, 173–180.

Griffin, S., Case, R., Siegler, R., 1994. Rightstart: providing the central conceptual prerequisites for first formal learning of arithmetic to students at-risk for school failure. In: McGilly, K. (Ed.), Classroom Lessons: Integrating Cognitive Theory and Classroom Practice. Bradford Books MIT Press, Cambridge, MA, pp. 24–49.

Grosse, C.S., Renkl, A., 2007. Finding and fixing errors in worked examples: can this foster learning outcomes? Learning & Instruction 17, 617–634.

Kalyuga, S., Ayres, P., Chandler, P., Sweller, J., 2003. The expertise reversal effect. Educational Psychologist 38, 23–31.

Kalyuga, S., Chandler, P., Tuovinen, J., Sweller, J., 2001. When problem solving is superior to studying worked examples. Journal of Educational Psychology 93, 579–588.

Kamii, C., 1974. Pedagogical principles derived from Piaget's theory: relevance for educational practice. In: Schwebel, M., Raph, J. (Eds.), Piaget in the Classroom. Basic Books, New York, NY, pp. 199–215.

Kaminski, J.A., Sloutsky, V.M., Heckler, A.F., 2008. The advantage of abstract examples in learning math. Science 320, 454–455.

Keil, F.C., Batterman, N., 1984. A characteristic-to-defining shift in the development of word meaning. Journal of Verbal Learning and Verbal Behavior 23, 221–236.

Klahr, D., Nigam, M., 2004. The equivalence of learning paths in early science instruction. Effects of direct instruction and discovery learning. Psychological Science 10, 661–667.

Kotovsky, L., Gentner, D., 1996. Comparison and categorization in the development of relational similarity. Child Development 67, 2797–2822.

Koedinger, K.R., Anderson, J.R., Hadley, W.H., Mark, M.A., 1997. Intelligent tutoring goes to school in the big city. International Journal of Artificial Intelligence in Education 8, 30–43.

Larkin, J., Simon, H., 1987. Why a diagram is (sometimes) worth ten thousand words. Cognitive Science 11 (1), 65–100.

Luchins, A.S., Luchins, E., 1950. New experimental attempts at preventing mechanization in problem solving. Journal of General Psychology 42, 279–297.

Marks, H.M., 2000. Student engagement in instructional activity: patterns in the elementary, middle, and high school years. American Educational Research Journal 37, 153–184.

Mayer, R., 2005. Cognitive theory of multimedia learning. In: Mayer, R. (Ed.), The Cambridge Handbook of Multimedia Learning. Cambridge University Press, Cambridge, pp. 31–48.

McNeil, N.M., 2007. U-shaped development in math: 7-year-olds outperform 9-year-olds on equivalence problems. Developmental Psychology 43, 687–695.

McNeil, N.M., 2008. Limitations to teaching children $2 + 2 = 4$: typical arithmetic problems can hinder learning of mathematical equivalence. Child Development 79, 1524–1537.

Miura, I.T., Okamoto, Y., Vlahovic-Stetic, V., Kim, C.C., Han, J.H., 1999. Language supports for children's understanding of numerical fractions: cross-national comparisons. Journal of Experimental Child Psychology 74, 356–365.

Mix, K.S., Paik, J.H., 2008. Do Korean fraction names promote part-whole reasoning? Journal of Cognition and Development 9, 145–170.

Moreno, R., Mayer, R.C., 1999. Multimedia-supported metaphors for meaning making in mathematics. Cognition and Instruction 17, 215–248.

National Center for Education Statistics, n.d. Eighth graders in most U.S. states perform above international average on mathematics and science assessment, press release. Retrieved from: http://nces.ed.gov/nationsreportcard/subject/studies/naep_timss/pdf/naep_timss_news_release.pdf

National Council of Teachers of Mathematics (NCTM), 2000. Principles and Standards for School Mathematics. Author, Reston, VA.

Newton, K.J., Star, J.R., Lynch, K., 2010. Understanding the development of flexibility in struggling algebra students. Mathematical Thinking and Learning 12, 282–305.

NMAP, 2008. Foundations for Success: The Final Report of the National Mathematics Advisory Panel. U.S. Department of Education, Washington, DC.

Opfer, J.E., DeVries, J., 2008. Representational change and magnitude estimation: why young children can make more accurate salary comparisons than adults. Cognition 108, 843–849.

Opfer, J.E., Siegler, R.S., 2007. Representational change and children's numerical estimation. Cognitive Psychology 55, 169–195.

Opfer, J.E., Siegler, R.S., Young, C.J., 2011. The powers of noise-fitting: reply to Barth and Paladino. Developmental Science 14, 1194–1204.

Opfer, J.E., Thompson, C.A., 2008. The trouble with transfer: insights from microgenetic changes in the representation of numerical magnitude. Child Development 79, 790–806.

Paik, J.H., Mix, K.S., 2003. U.S. and Korean children's comprehension of fraction names: a reexamination of cross-national differences. Child Development 74, 114–154.

Perry, M., Church, R.B., Goldin-Meadow, S., 1988. Transitional knowledge in the acquisition of concepts. Cognitive Development 3, 359–400.

Piaget, J., 1960. The Psychology of Intelligence. Harcourt, Brace, New York.

Piaget, J., 1970. Piaget's theory. In: Mussen, P. (Ed.), Carmichael's Manual of Child Psychology, vol. 1. John Wiley & Sons, New York, pp. 703–772.

Ramani, G.B., Siegler, R.S., 2008. Promoting broad and stable improvements in low-income children's numerical knowledge through playing number board games. Child Development 79 (2), 375–394.

Ramani, G.B., Siegler, R.S., 2011. Reducing the gap in numerical knowledge between low- and middle-income preschoolers. Journal of Applied Developmental Psychology 32, 146–159.

Renkl, A., Atkinson, R.K., Maier, U.H., 2000. From studying examples to solving problems: fading worked-out solution steps helps learning. In: Gleitman, L., Joshi, A.K. (Eds.), Proceeding of the 22nd Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates, Inc, Mahwah, NJ, pp. 393–398.

Ritter, S., Anderson, J.R., Koedinger, K.R., Corbett, A., 2007. Cognitive Tutor: applied research in mathematics education. Psychonomic Bulletin & Review 14, 249–255.

Rittle-Johnson, B., 2006. Promoting transfer: effects of self-explanation and direct instruction. Child Development 77, 1–29.

Rittle-Johnson, B., Star, J.R., 2007. Does comparing solution methods facilitate conceptual and procedural knowledge? an experimental study on learning to solve equations. Journal of Educational Psychology 99 (3), 561–574.

Rittle-Johnson, B., Star, J., Durkin, K., 2009. The importance of prior knowledge when comparing examples: influences on conceptual and procedural knowledge of equation solving. Journal of Educational Psychology 101 (4), 836–852.

Roy, M., Chi, M.T.H., 2005. Self-explanation in a multi-media context. In: Mayer, R. (Ed.), Cambridge Handbook of Multimedia Learning. Cambridge University Press, New York, pp. 271–286.

Siegler, R.S., 1988. Individual differences in strategy choices: good students, not-so-good students, and perfectionists. Child Development 59, 833–851.

Siegler, R.S., 1994. Cognitive variability: a key to understanding cognitive development. Current Directions in Psychological Sciences 3, 1–5.

Siegler, R.S., 1996. Emerging Minds: The Process of Change in Children's Thinking. Oxford University Press, New York.

Siegler, R.S., 2002. Microgenetic studies of self-explanations. In: Granott, N., Parziale, J. (Eds.), Microdevelopment: Transition Processes in Development and Learning. Cambridge University Press, New York, pp. 31–58.

Siegler, R.S., Booth, J.L., 2004. Development of numerical estimation in young children. Child Development 75, 428–444.

Siegler, R., Carpenter, T., Fennell, F., Geary, D., Lewis, J., Okamoto, Y., Thompson, L., Wray, J., 2010. Developing Effective Fractions Instruction: A Practice Guide. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Siegler, R.S., Chen, Z., 2008. Differentiation and integration: guiding principles for analyzing cognitive change. Developmental Science 11, 433–448.

Siegler, R.S., Fazio, L.K., Bailey, D.H., Zhou, X., 2013. Fractions: the new frontier for theories of numerical development. Trends in Cognitive Science 17, 13–19.

Siegler, R.S., Mu, Y., 2008. Chinese children excel on novel mathematics problems even before elementary school. Psychological Science 19, 759–763.

Siegler, R.S., Opfer, J.E., 2003. The development of numerical estimation: evidence for multiple representations of numerical quantity. Psychological Science 14, 237–243.

Siegler, R.S., Ramani, G.B., 2008. Playing board games promotes low-income children's numerical development. Developmental Science 11 (5), 655–661.

Siegler, R.S., Ramani, G.B., 2009. Playing linear number board games – but not circular ones – improves low-income preschoolers' numerical understanding. Journal of Educational Psychology 101, 545–560.

Siegler, R.S., Shrager, J., 1984. Strategy choices in addition and subtraction: how do children know what to do? In: Sophian, C. (Ed.), The Origins of Cognitive Skills. Erlbaum, Hillsdale, NJ, pp. 229–293.

Siegler, R.S., Thompson, C.A., 2014. Numerical landmarks are useful—except when they're not. Journal of Experimental Child Psychology 102, 39–58.

Siegler, R.S., Thompson, C.A., Opfer, J.E., 2009. The logarithmic-to-linear shift: one learning sequence, many tasks, many time scales. Mind, Brain, & Education 3, 143–150.

Siegler, R.S., Thompson, C.A., Schneider, M., 2011. An integrated theory of whole number and fractions development. Cognitive Psychology 62, 273–296.

Sowell, E.J., 1989. Effect of manipulative materials in mathematics instruction. Journal for Research in Mathematics Education 20, 498–505.

Star, J.R., Seifert, C., 2006. The development of flexibility in equation solving. Contemporary Educational Psychology 31, 280–300.

Steele, C.M., 1997. A threat in the air: how stereotypes shape intellectual identity and performance. American Psychologist 52, 613–629.

Steele, C.M., Aronson, J., 1995. Stereotype threat and the intellectual test performance of African Americans. Journal of Personality and Social Psychology 69, 797–811.

Sweller, J., 1999. Instructional Design in Technical Areas. ACER Press, Camberwell, Australia.

Sweller, J., Cooper, G.A., 1985. The use of worked examples as a substitute for problem solving in learning algebra. Cognition and Instruction 2, 59–89.

Thompson, C.A., Opfer, J.E., 2010. How 15 hundred is like 15 cherries: effect of progressive alignment on representational changes in numerical cognition. Child Development 81, 1768–1786.

Thompson, C.A., Opfer, J.E., 2008. Costs and benefits of representational change: effects of context on age and sex differences in magnitude estimation. Journal of Experimental Child Psychology 101, 20–51.

Thompson, C.A., Opfer, J.E., 2011. Learning linear spatial-numeric associations improves memory for numbers. In: Carlson, L., Hölscher, C., Shipley, T. (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society. Cognitive Science Society, Austin, TX, pp. 1503–1508.

Thompson, C.A., Siegler, R.S., 2010. Linear numerical magnitude representations aid children's memory for numbers. Psychological Science 21, 1274–1281.

Thompson, C.A., Xin, Z., Siegler, R.S. Cross-national Differences in Numerical Recall and Recognition, unpublished data.

Uttal, D.H., Scudder, K.V., DeLoache, J.S., 1997. Manipulatives as symbols: a new perspective on the use of concrete objects to teach mathematics. Journal of Applied Developmental Psychology 18, 37–54.

What Works Clearinghouse, 2009. Middle School Math: Cognitive Tutor Algebra 1.

Zhu, X., Simon, H.A., 1987. Learning mathematics from examples and by doing. Cognition & Instruction 4, 137–166.

Cognitive Dissonance

Joel Cooper and Kevin M Carlsmith[†], Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp. 2112–2114, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The theory of cognitive dissonance states that inconsistency between beliefs or behaviors creates an aversive motivational state akin to hunger or thirst. This tension is typically reduced by changing one of the dissonant elements, or adding new ones, until mental consonance is achieved. There have been three major revisions relevant to modern interpretation of dissonance phenomena. Aronson's self-consistency model (1968) proposed that dissonance resulted from behaviors that were discrepant with one's conception of oneself as a decent and sensible person. Steele's self-affirmation theory (1988) proposed that dissonance emerged from threats to the overarching self-system, and that dissonance reduction relied on reestablishing the integrity of the global self-concept. Cooper and Fazio's 'New Look' model (1984) proposed that dissonance resulted from creating unwanted aversive consequences and did not require cognitive inconsistency. A recent synthesis discussed by Cooper (1999) and Stone (1999) suggests that dissonance is caused by a discrepancy between the outcome of a behavioral act and the standard to which it is compared. According to this self-standards model, contextual variables determine the comparison standard, and it is this standard that determines which dissonance process is most likely to be operative.

Foundations of Dissonance Theory

The theory of cognitive dissonance is elegantly simple: it states that inconsistency between two cognitions creates an aversive state akin to hunger or thirst that gives rise to a motivation to reduce the inconsistency. According to Leon Festinger (1957), cognitions are elements of knowledge that people have about their behavior, their attitudes, and their environment. As such, a set of cognitions can be unrelated, consonant, or dissonant with each other. Two cognitions are said to be dissonant when one follows from the obverse of the other. The resultant motivation to reduce dissonance is directly proportional to the magnitude and importance of the discrepant cognitions, and inversely proportional to the magnitude and importance of the consistent cognitions. This tension is typically reduced by changing one of the cognitions, or adding new cognitions until mental 'consonance' is achieved. Festinger's original formulation proved to be one of the most robust, influential, and controversial theories in the history of social psychology. Although a number of challenges and revisions have been suggested, the basic behavioral observation remains uncontested and continues to stimulate fresh research.

Application of this theory has yielded many surprising and nonintuitive predictions. For example, conventional wisdom suggests that behavior follows from attitudes; dissonance theory, however, identifies conditions under which just the opposite occurs. An early and often replicated experiment illustrates the power and counterintuitiveness of the theory. In what is now known as the induced compliance effect, Festinger and Carlsmith (1959) asked individuals to perform 30 min of a mind-numbingly tedious activity, and then to persuade a waiting participant that the activity was in fact quite interesting. This situation created cognitive dissonance in most individuals – they believed that the task was boring, yet inexplicably found themselves arguing quite the opposite. Half of the participants were given a ready excuse for telling this lie – they were paid $20 to do so – while the other half, paid only $1, had no such excuse. Those with a clear justification for their odd behavior experienced no dissonance and, as one would expect, later reported that the task was rather boring. The other half, however, given insufficient justification for their behavior, experienced dissonance between the knowledge that the task was boring and the reality that they were misleading a fellow participant into believing the opposite. Rather than endure the aversive experience of believing one thing but saying another, these individuals changed their opinion and convinced themselves that the task was actually interesting. In other words, their attitude was shaped by their behavior.

Subsequent studies have confirmed the basic theory of cognitive dissonance and demonstrated its far-reaching impact. For example, cognitive dissonance explains the increased commitment so frequently observed following a severe initiation into a group. The theory also explains why, when faced with a choice among several desirable options, we observe the tendency to highlight positive aspects of the chosen option and negative aspects of the rejected alternatives after (and only after) the choice has been made. In the course of such studies, we have learned much about the boundary conditions associated with the theory and have identified anomalies not easily explained by the original theory. Since the 1960s, a number of theoretical revisions have sought to subsume these limitations under a unifying theory. This article summarizes briefly the leading reformulations of dissonance theory and speculates on future directions.

An Early Theoretical Challenge: Self-Perception Theory

Dissonance did not come quietly into psychology, nor did controversy begin over the finer points of the underlying process. Rather, the theory challenged the reigning theoretical paradigm of behaviorism of the 1950s by questioning the sovereign utility of basic learning theory. Rather than increased rewards leading to more positive attitudes, Festinger and Carlsmith (1959) had shown quite the opposite: participants who received a smaller reward for counterattitudinal advocacy

[†] Deceased

developed more positive attitudinal responses. The majority of early criticisms, therefore, focused not on the details of the dissonance theory but rather on its fundamental legitimacy. During parts of the next two decades, one of the enduring intellectual feuds in social psychology debated whether dissonance phenomena were the result of complex cognitive processes in the mind of the participant, or whether they were merely the result of complex cognitive processes in the mind of the experimenter. Daryl Bem (1972) took the position that one could derive similar predictions through far more parsimonious behavioral processes. He argued that participants in dissonance experiments were not experiencing negative psychological tension due to inconsistency, but rather were simply inferring their attitudes from their behavior and the situation in which it occurred. In essence, he suggested that people view their own behavior as though they were outside observers, and infer their underlying attitude from an analysis of their behavior. In support of this position, Bem replicated Festinger and Carlsmith (1959) and demonstrated that independent observers, aware of the monetary inducements, attributed attitudes to the participants that were nearly identical to the participants' actual attitudes. The critical test between these theories revolved around the search for physiological arousal. Festinger was quite clear that inconsistent cognitions created an aversive motivational state that could presumably be measured. Bem's self-perception theory, by contrast, predicted neither psychological nor physiological tension. However, Zanna and Cooper (1974) found indirect evidence for physiological arousal by showing that if arousal were misattributed to an irrelevant source, the effects of dissonance disappeared. The debate over dissonance versus self-perception was finally laid to rest by a series of experiments that identified the precise conditions under which each was operative. Fazio et al. (1977) found that small discrepancies between attitude and behavior (defined as those within a person's latitude of acceptance) tended to elicit self-perception processes, but larger discrepancies (those that fall in the person's latitude of rejection) were more likely to generate dissonance processes. Croyle and Cooper (1983) added direct support for Festinger's original position by showing that engaging in counter-attitudinal advocacy, at least outside people's latitude of acceptance, is marked by measurable increases in people's skin conductance responses. Ultimately, the importance of self-perception theory lay in its contributions to identifying boundary conditions of dissonance processes and in provoking research that established physiological arousal as one of the hallmarks of dissonance.

Introduction of the Self

It may be that all cognitive inconsistencies are not psychologically equivalent. According to some theorists, inconsistencies that implicate aspects of the self-maintain a privileged position. This suspicion led Elliot Aronson (1968) and Claude Steele (1988) to consider the role of self-concept in the dissonance process. Aronson concurred with Festinger's view that dissonance was caused by inconsistencies, but argued that it was a particular inconsistency that mattered most in arousing dissonance, i.e., the discrepancy between a person's general expectations for the self and his or her actual behavior. In other words, the arousal due to dissonance came about when a person's belief that he or she was a good and rational individual was called into question by behavior that was neither good nor rational. Aronson predicted that dissonance arousal would be more frequent and more powerful among those with high self-esteem – that is, among those whose past history had led them to believe that their high internal standards of behavior were likely to be achieved. By contrast, Aronson predicted that those with low self-esteem, who were accustomed to behaving less competently, would not be surprised or discomfited to find themselves once again behaving in an incompetent manner.

To illustrate his point, Aronson (1968) argued that the dissonance aroused in Festinger and Carlsmith's (1959) experiment was not due to inconsistency between the thoughts "I believe the task was dull" and "I told someone the task was interesting." Instead, Aronson proposed that dissonance was aroused by inconsistency between cognitions about the self (e.g., "I am a decent and truthful human being") and cognitions about the behavior (e.g., "I have misled a person … (and) conned him into believing something that just isn't true," p. 24). Aronson concluded: "at the very heart of dissonance theory, where it makes its clearest and neatest prediction, we are not dealing with just any two cognitions; rather we are usually dealing with the self-concept and cognitions about some behavior. If dissonance exists it is because the individual's behavior is inconsistent with his self-concept" (1968, p. 23).

Claude Steele (1988) suggested a different interpretation of the role of the self in creating dissonance. Like Aronson, Steele viewed inconsistent cognitions as a threat to the self; but unlike Aronson's self-consistency model, he suggested that the primary function of dissonance reduction was not to rescue the specific self-cognitions threatened by a behavioral outcome, but instead to restore the completeness of the overarching self-system. The difference, then, was not about the origin of dissonance arousal, but rather about the purpose and mechanism underlying dissonance reduction. Whereas Aronson focused on individual susceptibility to dissonance arousal, Steele's self-affirmation theory focused on individual resilience to dissonance. Interestingly, from Aronson's point of view, it is those individuals who have a positive self-regard that are most likely to experience dissonance arousal. Steele, by contrast, asserted that it is those same individuals – wrapped in their armor of self-resources – who feel immune from the need to reduce dissonance. Based on the data provided by Aronson and Steele, it is probably fair to say that those with high self-esteem are both more susceptible to dissonance arousal and more resistant to its effects because they can focus on their many other strengths. Holding all else constant, high self-esteem serves both as a catalyst that generates dissonance and as a buffer that mitigates the need to reduce dissonance.

A New Look at Dissonance

In 1984, Cooper and Fazio provided a comprehensive review of the dissonance literature and challenged the dominant assumption that dissonance was driven by a need for psychological consistency. According to their 'New Look' model, dissonance is aroused when people perceive that their behavior has been responsible for bringing about consequences that are

unwanted or aversive. If there are no such consequences, then inconsistent behavior will not produce the state of dissonance. For example, Cooper and Worchel (1970) replicated Festinger and Carlsmith's (1959) study with a condition in which the waiting participant was not convinced by the subject's lie. In this condition, the aversive consequence of misleading a fellow participant was removed along with all evidence of the dissonance process. Cooper and Fazio concluded from this and many other studies that responsibility for an aversive event rather than cognitive inconsistency plays the vital role in producing cognitive dissonance.

Future Directions – Self-Standards Model of Dissonance

Each of these reformulations of dissonance theory differs with respect to one major issue: what is the role of self-concept in dissonance processes? Is it a problem, a benefit, or completely irrelevant to the arousal and reduction of cognitive dissonance? There are data in favor of each position – data that are not easily reconcilable. However, a recent synthesis discussed by Cooper (1999) and Stone (1999) suggests that dissonance is caused by a discrepancy between the outcome of a behavioral act and the standard to which it is compared. According to the self-standards model, sometimes the standard that people use to measure their behavioral outcomes are personal and idiosyncratic. In such cases, people's views of themselves will play a crucial role. At other times, the assessment of an act is based on broad, normative standards that are shared in the culture. At these times, the self will not play a role in the dissonance process.

In summary, it is still useful to think of dissonance as involving inconsistency among cognitive elements and to conclude that inconsistency produces motivation for change. It is also fair to conclude that the future of dissonance theory will include a role for behavioral consequences, an assessment of the self, and an analysis of the contextual variables that make different standards the basis of judgment for behavioral outcomes. The evolution of cognitive dissonance calls for an integration that will likely include insights from the currently dominant perspectives.

Editor's Note

This article appeared in the first edition of the Elsevier *Encyclopedia of Social and Behavioural Sciences* (2001). Since then cognitive dissonance continues to interest social psychologists and to feed research on different fields such as health attitudes (smoking, eating habits), political attitudes (voting behavior, opinion polls, and election campaigns), and marketing. Dissonance theory also fed theories about system justification (see also Kay et al., 2002; Jost et al., 2003). Moreover, developments on attitudes theory showed that explicit but not implicit attitudes are affected by dissonance (Gawronski and Strack, 2004). There are still debates around the theory that can be traced in Brehm's (2007) review and a more recent Kenworthy et al. (2011) trans-paradigm theoretical synthesis of cognitive dissonance theory. In this article the authors using modern statistical techniques aim to test five research paradigms. The recent interest in and the development of neuroscience techniques showed brain activity at the moment of the decision phase in a decision-based cognitive dissonance paradigm (Jarcho et al., 2011) and suggest that the mere act of making a choice can change self-report preference and its neural representation (Izuma et al., 2010). It seems that cognitive dissonance is a theoretical field that will continue to trigger basic and applied research.

See also: Attitudes and Behavior; Motivation and Actions, Psychology of; Motivation: History of the Concept; Social Comparison, Psychology of.

Bibliography

Aronson, E., 1968. Dissonance theory: progress and problems. In: Ableson, R.P., Aronson, E., McGuire, W.J., Newcomb, T.M., Rosenberg, M.J., Tannenbaum, P.H. (Eds.), Theories of Cognitive Consistency: A Sourcebook. Rand McNally, Chicago.

Bem, D.J., 1972. Self-perception theory. Advances in Experimental Psychology 6, 1–62.

Brehm, J.W., 2007. A brief history of cognitive dissonance theory. Social and Personality Psychology Compass 1 (1), 381–391.

Cooper, J., 1999. Unwanted consequences and the self: in search of the motivation for dissonance reduction. In: Harmon-Jones, E., Mills, J. (Eds.), Cognitive Dissonance: Progress on a Pivotal Theory in Social Psychology. American Psychological Association, Washington, DC, pp. 149–173.

Cooper, J., Fazio, R.H., 1984. A new look at dissonance theory. Advances in Experimental Psychology 17, 229–262.

Cooper, J., Worchel, S., 1970. Role of undesired consequences in arousing dissonance. Journal of Personality and Social Psychology 16, 199–206.

Croyle, R., Cooper, J., 1983. Dissonance arousal: physiological evidence. Journal of Personality and Social Psychology 45, 782–791.

Fazio, R.H., Zanna, M.P., Cooper, J., 1977. Dissonance and self-perception: an integrative view of each theory's proper domain of application. Journal of Experimental Psychology 13, 464–479.

Festinger, L., 1957. A Theory of Cognitive Dissonance. Row, Peterson, Evanston, IL.

Festinger, L., Carlsmith, J.M., 1959. Cognitive consequences of forced compliance. Journal of Abnormal Social Psychology 58, 203–210.

Gawronski, B., Strack, F., 2004. On the propositional nature of cognitive consistency: dissonance changes explicit but not implicit attitudes. Journal of Experimental Social Psychology 40 (4), 535–542.

Izuma, k., Matsumuto, M., Muramaya, K., Samejima, K., Sadato, N., Matsumoto, K., 2010. Neural correlates of cognitive dissonance and choice-induced preference change. PNAS. Proceedings of the National Academy of Sciences of the United States of America 107 (51), 22014–22019.

Jarcho, J.M., Berkman, E.T., Lieberman, M.D., 2011. The neural basis of rationalization. Cognitive dissonance reduction during decision-making. Social Cognitive and Affective Neuroscience 6 (4), 460–467.

Jost, J.T., Pelham, B.W., Sheldon, O., Sullivan, B.N., 2003. Social inequality and the reduction of ideological dissonanceon behalf of the system: evidence of enhanced system jutification among the disadvantaged. European Journal of Social psychology 33 (1), 13–36.

Kay, A.C., Jimenez, M.C., Jost, J.T., 2002. Sour grapes, sweet lemons and the anticipatory justification of the status quo. Personality and Social Psychology Bulletin 28 (9), 1300–1312.

Kenworthy, J.B., Miller, N., Collins, B.E., Read, S.J., Earleywine, M., 2011. A trans-paradigm tehoretical synthesis of cognitive dissonance theory: illuminating the nature of discomfort. In: Stroebe, W., Hewstone, M. (Eds.), European Review of Social Psychology, vol. 22, pp. 36–113.

Steele, C.M., 1988. The psychology of self-affirmation: sustaining the integrity of the self. Advances in Experimental Psychology 12, 261–302.

Stone, J., 1999. What exactly have I done? The role of self-attribute accessibility in dissonance. In: Harmon-Jones, E., Mills, J. (Eds.), Cognitive Dissonance: Progress on a Pivotal Theory in Social Psychology. American Psychological Association, Washington, DC, pp. 175–200.

Zanna, M.P., Cooper, J., 1974. Dissonance and the pill: an attribution approach to studying the arousal properties of dissonance. Journal of Personality and Social Psychology 29, 703–709.

Cognitive Maps

Rob Kitchin, National University of Ireland Maynooth, Maynooth, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A cognitive map is a representative expression of an individual's cognitive map knowledge, where cognitive map knowledge is an individual's knowledge about the spatial and environmental relations of geographic space. For example, a sketch map drawn to show the route between two locations is a cognitive map – a representative expression of the drawer's knowledge of the route between the two locations. This map can be analyzed using classification schemes or quantitatively using spatial statistics. The drawing of a cognitive map is supplemented by a range of other techniques designed to measure various aspects of cognitive map knowledge, including distance and direction tasks and more recently brain scans. These measures are important because they reveal fundamental aspects of cognition and the cognitive processes that underlie spatial decision and choice making.

Introduction

A cognitive map is a representative expression of an individual's cognitive map knowledge, where cognitive map knowledge is an individual's knowledge about the spatial and environmental relations of geographic space. For example, a sketch map drawn to show the route between two locations is a representative expression of the drawer's cognitive map knowledge. While the term cognitive map suggests a representation that corresponds cartographically to how geographic knowledge is held within the mind, the term is used more metaphorically to capture the notion that people and animals possess such knowledge and use it to navigate and make spatial decisions and choices that affect spatial behavior – why we choose certain routes, places to visit, locations to live, and so on. Indeed, there is much debate as to the extent to which cognitive map knowledge consists of imagery and representations or other forms of knowledge, encodes spatial relations (such as distance, scale and direction), links spatial and nonspatial information, how it is mentally organized and stored, and is shaped by cognitive, neurological, and external factors (see Kitchin and Blades, 2002). Because cognitive map knowledge represents and demonstrates an individual's geographic knowledge, geographers, psychologists, and others seek to capture such information through various tests in order to assess how we learn, process, and store geographic information gained from primary (e.g., walking through an area) and secondary (e.g., reading a map) sources.

Cognitive Mapping

The combined processes by which we learn, store, and use information relating to the geographic world are known as cognitive mapping. Cognitive mapping as a field of enquiry is relatively young. While a handful of studies predate 1960, the majority of research has occurred since the publication of Kevin Lynch's seminal work *The Image of the City* (Lynch, 1960). This book examined how people understood and remembered the geography of US cities. The cognitive mapping research that followed has been a multidisciplinary endeavor, undertaken by geographers, psychologists, anthropologists, computer and information scientists, and neurologists. By examining how different populations (e.g., varying ages of children, people without sight, or the two sexes) learn and employ geographic knowledge, research has sought to understand how people and animals acquire cognitive mapping abilities across the lifespan and to document and understand fundamental processes of spatial and environmental cognition. Despite the vast amount of research that has been conducted, as yet there is only limited consensus as to the nature of cognitive mapping. As such, there are a number of theories that seek to explain how we construct, process, and store cognitive map knowledge and how this knowledge is used to make spatial decisions and choices.

Until recently, most of these theories have been grounded in cognitive and environmental psychology and concern fundamental processes of spatial cognition and how people and animals come to spatially understand the world (Golledge and Stimson, 1997; Kitchin and Blades, 2002). In such cases, an interactionalist perspective has largely been adopted wherein there is a Cartesian separation of mind from the external world, of individual from environment, with the elements of the latter understood as inputs and the source of causal effects on spatial knowledge and cognition. More recently, such theories have been challenged from two predominant sources. One set of studies have focused not on psychological processes but on neurological processing and how the brain functions to enact spatial cognition and behavior (Byrne et al., 2007). The second set of studies are transactional in nature, in that they do not separate individuals from environment, but rather see sociocultural context as constitutive of spatial cognition; spatial choices and decisions are not simply the result of disembodied spatial reasoning but are grounded in the unfolding transaction between individual and the world (Heft, 2013). All of these theories are underpinned by studies that seek to capture cognitive map knowledge (see Downs and Stea, 1977; Gärling and Golledge, 1993; Portugali, 1996; Golledge and Stimson, 1997; Kitchin and Blades, 2002; Montello, 2009).

Cognitive Maps

Traditionally, the bases and processes of human cognitive mapping was exclusively measured and assessed through

sketch maps and other external forms of knowledge representation (e.g., estimating distances). More recently, these methods of data generation have been supplemented by qualitative and neurological approaches. Moreover, there has been a transference of experimental setting from the laboratory to natural and virtual environments.

Sketch Maps

Sketch maps have long been used as a method to capture cognitive map knowledge, consisting of a drawing of how an individual understands the spatial relations of an area. In experimental conditions, subjects are given a sheet of paper and are asked to draw a map of a certain location, area, or route between locations (alternatively, they might be asked to make a map using some building blocks). The scale of the geographic realm to be drawn can vary substantially from the global (e.g., draw a map of the world) to the local (e.g., draw a map of your street or neighborhood) (Appleyard, 1970). Variants on this simple sketch mapping exercise include providing respondents with a small portion of the map to provide a scale and reference (e.g., Pearce, 1981) and teaching subjects a sketch map language where specific symbols are used to denote particular features (Beck and Wood, 1976). By aggregating together the cognitive maps of several individuals, it is possible to determine their shared level of knowledge and which elements of an environment are most salient. This is the technique pioneered by Lynch (1960). He analyzed individuals' maps by classifying their elements into five different classes (see **Table 1**) which he then used to produce a composite map where the symbol size/shading density is proportional to the number of times an element appeared on the individual maps. Using this technique, he aggregated together the sketch maps of residents in Boston, Jersey City, and Los Angeles to create composite cognitive maps of these cities (see **Figure 1**).

The analysis used by Lynch is a content classification. A number of other classification schemes have been used to analyze sketch maps. For example, there have been classifications that assess map style, structure, and accuracy. In these cases, the focus moves beyond what elements an individual draws to assessing the relationship between the elements and their relativity to the real world. In addition, the accuracy of the spatial relations portrayed can be analyzed statistically using spatial statistics. For example, in many studies bidimensional regression has been used to compare the geometry of the cognitive map to a cartographic map (Golledge and Spector, 1978; Waterman and Gordon, 1984). Bidimensional regression is a two-dimensional equivalent of linear regression that quantifiably assesses scale, rotation, and translation differences between an actual and an estimated pattern of responses.

Using the technique of sketch mapping to generate data about a person's cognitive map knowledge and their ability to use this knowledge is not without criticism. For example, a number of researchers have argued that sketch maps have a number of qualities that make them unreliable and inaccurate measures of spatial knowledge (note not geographic knowledge): they are dependent upon drawing ability and familiarity with cartographic conventions; they suffer from associational dependence where later additions to the sketch will be influenced by the first few elements that are drawn; their content and style are influenced by the size of paper used for sketching; they are difficult to subjectively score and code; and they often show less information than the respondent knows (Beck and Wood, 1976; Blades, 1990; Hart, 1981). As a consequence of these criticisms, it is becoming less common for researchers to use sketch mapping as an analytical tool to assess individual and collective cognitive mapping. Instead, researchers are turning to a range of other techniques.

Table 1 Lynch's classification

Category	*Description*
Paths	Paths are the channels along which an individual moves. They may include streets, walkways, and railways
Edges	Edges are the linear elements not considered as paths. They are the boundaries between two phases, linear breaks in continuity such as shores or walls
Districts	Districts are the medium-to-large scale sections to the city, conceived as having a two-dimensional extent, which the observer mentally enters, and which have some common identifiable character
Nodes	Nodes are points, the strategic spots in the city into which an observer can enter, and which are the intensive foci to and from which (s)he is traveling. They may be primarily junctions, transportation changeovers, a crossing, or convergence of paths
Landmarks	Landmarks are another type of point reference. They are usually a physical object such as a building, sign, store, or mountain

Other Action Measures

Other forms of action (e.g., drawing) measures can be divided into those that, like sketch mapping, are two-dimensional in nature and those that are unidimensional. Other two-dimensional measures include completion tasks and recognition tasks. Completion tasks require an individual to complete a task that has already been started for them. For example, spatial cued response tests require subjects to place locations in relation to locations that are preplaced (Thorndyke and Hayes-Roth, 1982). A highly cued version of this are cloze procedure tests that require a subject to 'fill in' a missing space (an aspatial example of which would be, 'a dog barks but a cat ------?') (Robinson, 1974). Recognition tasks measure subjects' success at identifying spatial relationships. Iconic tests require the respondent to identify correctly features on a map or aerial photograph (Matthews, 1984). Configuration tests require a subject to identify correctly which configuration, out of several, displays the correct spatial relations (Evans et al., 1984). Verifiable statement tests require subjects to identify whether a textual description of a spatial relationship is true or false (Ferguson and Hegarty, 1994). Spatial cued response data are often analyzed like sketch maps using bidimensional regression. Cloze procedure and recognition tests are analyzed by constructing an accuracy score that reveals as a percentage the number of correct placements or recognitions (Kitchin and Blades, 2002).

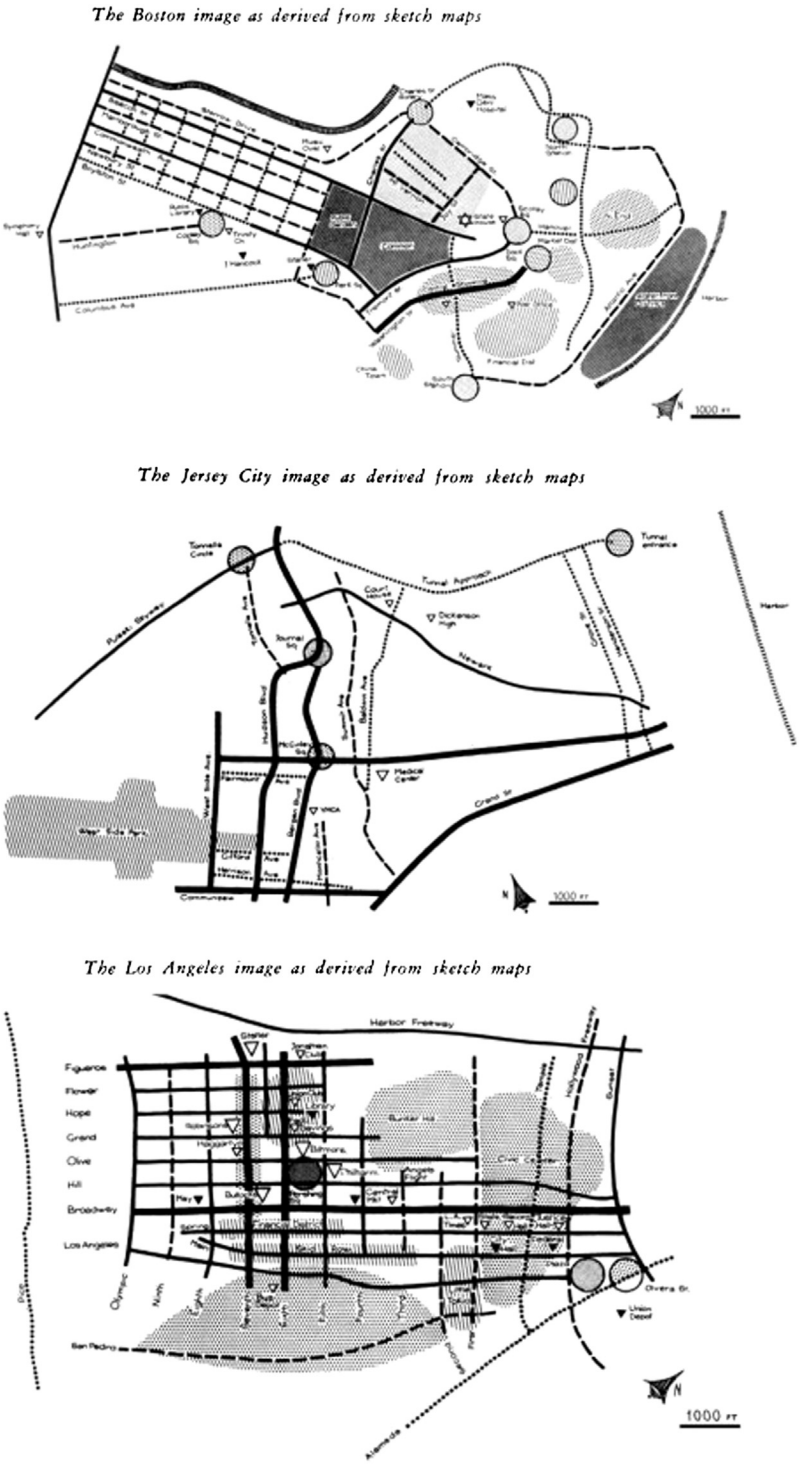

Figure 1 Cognitive maps of Boston, Jersey City, and Los Angeles.

Unidimensional tests seek to uncover one-dimensional aspects of cognitive map knowledge such as distance and direction. These dimensions are thought to be representative of spatial knowledge in general but are particularly useful for measuring levels of route (procedural) knowledge. Distance tasks are used to assess an individual's knowledge of the distance between locations. In his review, Montello (1991) identifies five groups of tests designed to measure cognitive distance estimates: ratio scaling, interval and ordinal scaling, mapping, reproduction, and route choice. Ratio scaling adapts traditional psychophysical scaling techniques to a distance context, with subjects estimating the distance to

a location as a ratio of some other known distance, such as an arbitrary scale or the length of a ruler. Interval and ordinal scaling are similar to ratio scaling but differ in their level of measurement: paired comparison requires a respondent to decide which one of a pair of distances is longer; ranking requires a respondent to rank various distances in order along the dimension of length; rating requires a respondent to assign the distance between places to a set of predetermined classes that represent relative length; partition scales require a respondent to assign distances to classes of equal-appearing intervals of length. Mapping is the measurement of distances from a sketch map for comparison with the actual distances (Montello, 1991). Reproduction requires a respondent to provide distance estimates at the scale of the estimated distance (Herman and Klein, 1985). Route choice consists of inferring judgments of cognitive distance from the choice of route an individual makes when asked to take the shortest route between two locations.

Like sketch map data, distance data can be analyzed individually or aggregated to provide data about a group. A common method to analyze ratio scaling, mapping, and reproduction data is to regress the cognitive distance estimates onto the objective distance values, observing the relationship between the two (Cadwallader, 1973). Another common strategy, and one also used with interval or ordinal distance data, is to analyze the data using multidimensional scaling techniques (MDS) (Harman and Betak, 1976). MDS techniques explore the latent structure of a set of distance estimates by assessing the dimensionality of the data. They do this by constructing a two-dimensional space from one-dimensional data using a series of algorithms. In essence, they construct a 'map' showing the relationships among a number of objects. This 'map' can then be compared to an actual map using techniques such as bidimensional regression (Kitchin, 1996).

Direction tasks assess an individual's knowledge of the direction between two locations. The most common direction task is pointing. Pointing involves standing at, or imagining being at, a location and pointing to another location (Kirasic et al., 1984). An alternative technique involves drawing on a compass the direction to a location from another (Tversky, 1981). Direction estimates have been analyzed by comparing the estimates to the actual directions, often through a simple subtraction process. In other cases, a technique of projective convergence has been used to construct a 'map' from estimates by calculating where estimates to the same location but from different sites intersect (Kitchin, 1996).

Qualitative Approaches

In recent years, there has been an increase in the use of qualitative methodologies to investigate cognitive map knowledge. In some cases, this has involved a scientific approach and in others an interpretative approach. The scientific approach continues the tradition described earlier, but rather than externally representing their knowledge through actions (e.g., drawing, pointing) individuals are required to describe verbally routes or layouts in experimental conditions (Taylor and Tversky, 1992). Like sketch maps these data can be analyzed for content, style, structure, and accuracy. The interpretative approach, however, is less structured in terms of data collection. It posits that talking to and observing individuals as they interact with an environment reveals information concerning spatial behavior. Such an approach might seek to gain a spatial understanding of an area by adopting a strategy of in-depth interviews, discussing the reasoning behind spatial decision making (Kitchin, 1997).

Neurological Approaches

In contrast to the earlier approaches, which seek to understand the process of cognitive mapping by examining external measures (action or verbal), neurological approaches measure neural activity within the brain (Dudchenko, 2010). A set of brain-scan techniques exist, differentiated by their temporal and spatial resolution. Magnetic electroencephalography has a high temporal resolution (0–300 ms) but low spatial resolution (providing only a general indication of neural activity). Positron emission tomography is the converse, with functional magnetic resonance imagery (fMRI) providing a middle ground for both parameters. To reveal information about spatial cognition, scans are taken as the individual undertakes a series of spatial, problem-solving tasks. As such, neurological approaches seek to explain spatial thought and spatial behavior by identifying its neural, physiological bases rather than its psychological basis. To date few studies have sought to marry the findings of neurological and psychological research to provide a comprehensive, physiological, and psychological model of cognitive mapping.

Naturalistic Settings

Many studies, particularly from psychology, which seek to understand cognitive map knowledge are conducted in controlled laboratory settings. For example, respondents may be required to learn the layout of objects in an experimental laboratory and then to map the location of objects or estimate distance and directions between objects. The laboratory is attractive to researchers because they can control/monitor all the variables that may affect spatial learning and processing and thus isolate those that are particularly salient. Some researchers, however, question the ecological validity of this approach given that cognitive map knowledge concerns the geographic environment and it is this environment in which spatial behavior occurs (Golledge and Stimson, 1997; Kitchin and Blades, 2002). The real world is inherently messy and full of factors that cannot easily be controlled and that work in concert with each other. Data collected in a laboratory then do not mirror real life, and salient variables do not work in isolation. For these reasons, they suggest that testing should occur in the natural environment, and increasingly this is becoming more common (Gale et al., 1990; Jacobson et al., 1998).

Virtual Environments

The development of sophisticated virtual, immersive 'game spaces' has provided a new means to assess cognitive mapping. Such virtual environment are attractive to researchers because they can appear highly naturalistic through the use of photo-realistic graphics but are nevertheless highly controlled

environments; the researcher can limit and isolate inputs and contextual factors, and virtual environments are entirely measurable in precise ways with regards to location and individual reactions (Peruch et al., 2000). Such systems can also measure physiological factors such as eye and body movement. For these reasons, studies using virtual environments have become quite common in the past decade (e.g., Mou et al., 2013). It is likely that these studies will be supplemented by those utilizing the recent development of augmented reality – virtual displays overlaid on natural environments – as a measurement technique, an approach that allows the blending of naturalistic and virtual terrains.

Summary

Cognitive map knowledge concerns how we learn, store, and process knowledge of the geographic environment. A variety of theories designed to explain such cognitive mapping have been tested using a variety of measures, applied across different groups and settings. The outputs of these measures are, in essence, cognitive maps – externalized representations of cognitive map knowledge. Such representations are important because they reveal fundamental aspects of cognition and the cognitive processes that underlie spatial decision and choice making.

See also: Hippocampus and Related Structures; Knowledge (Explicit, Implicit and Tacit): Philosophical Aspects; Spatial Cognition.

Bibliography

Appleyard, D., 1970. Styles and methods of structuring a city. Environment and Behavior 2, 100–117.

Beck, R.J., Wood, D., 1976. Comparative developmental analysis of individual and aggregated cognitive maps of London. In: Moore, G.T., Golledge, R.G. (Eds.), Environmental Knowing. Dowden, Hutchinson and Ross, Stroudsberg, PA, pp. 173–184.

Blades, M., 1990. The reliability of data collected from sketch maps. Journal of Environmental Psychology 10, 327–339.

Byrne, P., Becker, S., Burgess, N., 2007. Remembering the past and imagining the future: a neural model of spatial memory and imagery. Psychological Review 114 (2), 340–375.

Cadwallader, M.T., 1973. A methodological examination of cognitive distance. In: Preisner, W.F.E. (Ed.), Environmental Design Research, vol. 2. Dowden, Hutchinson and Ross, Stroudsberg, PA.

Downs, R.M., Stea, D., 1977. Maps in Minds: Reflections on Cognitive Mapping. Harper & Row, New York.

Dudchenko, P.A., 2010. Why People Get Lost: The Psychology and Neuroscience of Spatial Cognition. Oxford University Press, Oxford.

Evans, G.W., Skorpanich, M.A., Gärling, T., Bryant, K., Bresolin, B., 1984. The effects of pathway configuration, landmarks, and stress on environmental cognition. Journal of Environmental Psychology 4, 323–335.

Ferguson, E.L., Hegarty, M., 1994. Properties of cognitive maps constructed from texts. Memory and Cognition 22, 455–473.

Gale, N., Golledge, R.G., Pellegrino, J.W., Doherty, S., 1990. The acquisition and integration of neighborhood route knowledge. Journal of Environmental Psychology 10, 3–25.

Gärling, T., Golledge, R.G. (Eds.), 1993. Behavior and Environment: Psychological and Geographical Approaches. North Holland, Amsterdam.

Golledge, R.G., Spector, A.N., 1978. Comprehending the urban environment: theory and practice. Geographical Analysis 9, 403–426.

Golledge, R.G., Stimson, R.J., 1997. Spatial Behavior: A Geographic Perspective. Guildford Press, New York.

Harman, E.J., Betak, J.F., 1976. Behavioral geography, multidimensional scaling and the mind. In: Golledge, R.G., Rushton, G. (Eds.), Spatial Choice and Spatial Behavior. Ohio State University Press, Columbus, OH, pp. 3–22.

Hart, R., 1981. Children's representation of the landscape: lessons and questions from a field study. In: Liben, L., Patterson, A.H., Newcombe, N. (Eds.), Spatial Representation across the Life Span. Academic Press, New York, pp. 195–223.

Heft, H., 2013. Environment, cognition, and culture: reconsidering the cognitive map. Journal of Environmental Psychology 33, 14–25.

Herman, J.F., Klein, R.V., 1985. The effect of travel on children's distance estimation. British Journal of Developmental Psychology 3, 353–361.

Jacobson, R.D., Kitchin, R.M., Garling, T., Golledge, R., Blades, M., 1998. Learning a complex urban route without sight: comparing naturalistic versus laboratory measures. In: Proceedings of Mind III: Spatial Cognition Conference. Dublin City University, Dublin, August 18.

Kirasic, K.C., Allen, G., Siegel, A.W., 1984. Expression of configurational knowledge of large-scale environments: student's performance of cognitive tasks. Environment and Behavior 16, 687–712.

Kitchin, R.M., 1996. Methodological convergence in cognitive mapping research: investigating configurational knowledge. Journal of Environmental Psychology 16, 163–185.

Kitchin, R.M., 1997. Exploring spatial thought. Environment and Behavior 29, 123–156.

Kitchin, R., Blades, M., 2002. The Cognition of Geographic Space. IB Tauris, London.

Lynch, K., 1960. The Image of the City. Technology Press, Cambridge, MA.

Matthews, M.H., 1984. Cognitive maps: a comparison of graphic and iconic techniques. Area 16, 33–40.

Montello, D.R., 1991. The measurement of cognitive distance: methods and construct validity. Journal of Environmental Psychology 11, 101–122.

Montello, D.R., 2009. Cognitive geography. In: Kitchin, R., Thrift, N. (Eds.), The International Encyclopedia of Human Geography, vol. 2. Elsevier, Oxford, pp. 160–166.

Mou, W., McNamara, T.P., Zhang, L., 2013. Global frames of reference organize configural knowledge of paths. Cognition 129, 180–193.

Pearce, P.L., 1981. A study of travellers' perceptions of a section of countryside. Journal of Environmental Psychology 1, 141–155.

Peruch, P., Gaunet, F., Thinus-Blanc, C., Loomis, J., 2000. Using virtual technology to investigate mental spatial representations. In: Kitchin, R., Freundschuh, S. (Eds.), Cognitive Mapping. Past, Present, and Future. Routledge, London, pp. 108–124.

Portugali, J. (Ed.), 1996. The Construction of Cognitive Maps. Kluwer, Dordrecht.

Robinson, M.E., 1974. Cloze procedure and spatial comprehension test. Area 9, 137–142.

Taylor, H.A., Tversky, B., 1992. Description and depictions of environments. Memory and Cognition 20, 483–496.

Thorndyke, P.W., Hayes-Roth, B., 1982. Differences in spatial knowledge acquired from maps and navigation. Cognitive Psychology 14, 560–589.

Tversky, B., 1981. Distortions in memory for maps. Cognitive Psychology 13, 407–433.

Waterman, S., Gordon, D., 1984. A quantitative-comparative approach to analysis of distortion in mental maps. Professional Geographer 36, 326–337.

Cognitive Modeling: Connectionist Approaches

Bruce MacLennan, University of Tennessee, Knoxville, TN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Connectionist approaches to cognitive modeling make use of large networks of simple computational units, which communicate by means of simple quantitative signals. Higher-level information processing emerges from the massively parallel interaction of these units by means of their connections, and a network may adapt its behavior by means of local changes in the strength of the connections. Connectionist approaches are related to neural networks and provide a distinct alternative to cognitive models inspired by the digital computer.

Definitions

To facilitate the following discussion, it will be helpful to define some terms. A typical connectionist network comprises a (potentially large) number of simple processing 'units.' The units are often called (artificial) neurons, but that terminology begs the question of their relation to biological neurons, so it will be avoided here. In the most common case, the units form a weighted sum of their (quantitative) inputs and pass the result through a simple, nonlinear 'activation function,' which limits the range of possible outputs. The resulting value is considered the 'activity' of the unit, which may be transmitted to other units (through outgoing connections). In some cases the activity of a unit is a combination of its inputs and previous activity, which provides a kind of short-term memory residing in the collective activities of the units.

The weighted sum results from the fact that each connection in the network has an associated 'weight' (analogous to synaptic efficacy in biological neural networks), which multiplies the quantity transmitted by that connection. Positive weights correspond to excitatory connections and negative weights to inhibitory; zero-valued weights correspond to the absence of a connection. Mathematically, connection weights are often treated as a 'weight matrix' $\mathbf{W}$, with element W_{ij} being the weight of the connection to unit i from unit j. Learning and adaptation take place by modification of the weights according to some 'learning algorithm' (described later); thus the connections constitute the network's long-term memory. 'Connectionism' derives its name from the fact that knowledge resides in the patterns and weights of the connections.

Many connectionist networks are organized into 'layers,' analogous to functional areas in the brain; information usually moves in lockstep from layer to layer. Although many networks are 'feed-forward,' that is, the information moves through successive layers from input to output, other networks are 'recurrent,' which means that there may be feedback connections from a layer to itself or to earlier layers. Recurrent networks are able to recognize and process 'temporally extended patterns,' that is, sequences of related inputs.

It must be stressed that there are exceptions to all of the preceding general statements about connectionist networks, and connectionist approaches are best viewed as forming a Wittgensteinian family resemblance.

History

A short history of connectionist approaches will be presented, first in the narrower context of cognitive science and artificial intelligence (AI), and then in the broader context of epistemology, linguistics, and the philosophy of mind. Although the size of this article does not permit detailed citations of the literature, many of the seminal articles are collected in Anderson and Rosenfeld (1988) and Anderson et al. (1990). *See* Cognitive Psychology: History.

Narrower History

According to the second edition of the *Oxford English Dictionary*, the term 'connectionism' was first used by E. L. Thorndike in his *Fundamentals of Learning* (1932) to refer to the reduction of mental processes to the connections between stimuli and responses, that is, to a form of associationism, and so connectionist theories have been set in opposition to cognitive theories. The term is used somewhat differently now (so that 'connectionist cognitive science' is not an oxymoron), but retains some similarities to associationism. However, to understand the relation it is better to look at connectionism from the perspective of neural network models of cognition.

In the early 1940s, W. S. McCulloch and W. Pitts investigated the computation of logical functions by simple neuron-like elements; in effect they showed that these elements could compute logical 'and,' 'or,' 'not,' and so forth. In his *Organization of Behavior*, D. O. Hebb (1949) suggested that learning takes place by the formation of 'cell assemblies,' and that this occurs through the strengthening of connections between simultaneously active neurons in neural networks, which are initially randomly connected. His description of this process inspired one of the simplest connectionist learning rules (described later).

In the late 1950s, F. Rosenblatt began to investigate the application of simple neuron models called 'perceptrons' to perceptual problems such as classifying printed letters. He developed a learning algorithm for simple (single-layer) perceptron networks, which iteratively adjusted the connection weights (strengths) whenever the network made a mistake. He proved that if the network were capable of solving the problem at all, then the algorithm would eventually find the connection weights to solve it. However, there are many problems that

 http://dx.doi.org/10.1016/B978-0-08-097086-8.43021-7

a single-layer network cannot solve, and Rosenblatt never succeeded in finding a multilayer learning algorithm.

A key event in the history of connectionism was the publication of M. Minsky and S. Papert's *Perceptrons* (1969), which demonstrated limitations of simple perceptron networks. Specifically, they proved that single-layer perceptron nets could discriminate only those categories that are 'linearly separable'. Their proof did not apply to multilayer nets (for which there was still no learning algorithm), but they suggested that similar limitations would be found for these too. Nevertheless, their book was widely interpreted as showing the impotence of neural networks in general, and is commonly blamed for discouraging research in the field for a decade. (The extent to which it actually did so is perhaps a topic for historians of science.)

It will be worthwhile to comment on the role of holography in the development of connectionist approaches. As early as 1929, K. S. Lashley had conducted experiments suggesting that individual memory traces were not localized in any one place, and that degradation of memory was proportional to the amount of cortical mass destroyed, thus implying that individual traces were distributed over large areas of cortex. In a well-known 1950 paper, he despaired of ever understanding how such a nonlocal memory could operate (Lashley, 1950). The principles of holography had been described by D. Gabor in 1949, but it was not until the advent of optical holography in the early 1960s that it began to be seen as a solution to Lashley's dilemma. Although an analogy between holography and memory had been suggested as early as 1963 by P. J. van Heerden, the holographic hypothesis has been developed most extensively by K. H. Pribram and his colleagues since 1966 (see, e.g., Anderson et al., 1990, Chapter 7). In the late 1960s and early 1970s, holographic and holography-inspired models of associative memory were also investigated by H. C. Longuet-Higgins, D. J. Willshaw, and others (Hinton and Anderson, 1989). Some connectionists were influenced by the critiques of traditional, rule-based AI by the phenomenologist philosopher H. L. Dreyfus (*What Computers Can't Do*, 1972). Although he stressed the limitations of rule-based systems, he also explained that some of these limitations would not apply to analog systems operating on principles similar to holography. *See* Distributed Cognition; Memory Trace, Nature of; Models of Neural Basis of Learning and Memory.

Although a number of investigators (including J. A. Anderson, S. Grossberg, and T. Kohonen) continued connectionist research through the 1970s, the field was quiet until rejuvenated by the work of D. E. Rumelhart, J. L. McClelland, and other members of the PDP (Parallel Distributed Processing) Working Group. Many of their publications were collected in a widely read two-volume set (Rumelhart et al., 1986). The credibility of connectionist approaches was also enhanced by J. Hopfield's publication in 1982 of a simple recurrent net capable of associative memory and pattern completion.

Broader History

Although connectionism can be viewed as an approach to knowledge representation and inference of relevance only to cognitive science, in fact it has much broader implications, for it challenges assumptions about knowledge that have been largely unquestioned since ancient Greek philosophy. Already in the philosophies of Socrates, Plato, and Aristotle, there is a preference for knowledge expressed as logical relations among discrete, language-like structures, and for a view of cognition as mechanized deduction. These ideas influenced many later philosophers, including Hobbes (who equated thinking with computation), Leibniz (who experimented with formalized systems of knowledge representation and mechanical deduction), and Boole (who invented mathematical logic). These ideas were also influential in the development of logical positivism, which dominated the philosophy of science in the first half of the twentieth century. The idea persisted in the assumption that there must be a 'language of thought,' because it is 'the only game in town' (i.e., no one could think of an alternative). Similarly, most research in AI took for granted that intelligence resides in the structures of a 'knowledge representation language' and in deduction-like formal rules for their manipulation. Throughout the 1970s, AI researchers concentrated their attention on expert systems, which depended on expertise represented symbolically. Disappointment with the performance of these systems was one of motivations for the connectionist renaissance.

In summary, the Western tradition (with some exceptions) has displayed a kind of linguistic chauvinism, which presumes that all knowledge and cognition can be expressed in language-like structures. Knowledge is expressed at a 'symbolic' level, that is, in terms of atomic (indivisible), word-level categories related by sentence-like logical structures. On the other hand, most connectionist approaches represent knowledge at a 'subsymbolic' level, that is, in terms of minute, quantitative features related by low-level, often statistical, connections. In other words, knowledge is more akin to an image than to a sentence (*see* Mental Imagery: Visual Cognition). Therefore, some of connectionism's advocates see it as a fundamentally new view of knowledge and cognition, which is leading to a paradigm shift in cognitive science and philosophy and is engendering a new AI. Specific innovations of the connectionist approach are discussed below. *See* Artificial Intelligence: Connectionist and Symbolic Approaches; Knowledge Representations, Theory of.

Mechanisms of Adaptation and Learning

Virtually all connectionist approaches incorporate adaptive mechanisms or 'learning algorithms,' which allow the network to improve its performance; here a few will be discussed briefly.

Correlational Learning

'Correlational learning' ('Hebb's rule'), the simplest connectionist learning algorithm, takes its inspiration from Hebb's hypothesis that the simultaneous activity of two neurons strengthens the connection between them. It makes a change in connection weight proportional to the product of activities of the units it connects. Thus, the change in the weight W_{ij} of the connection to unit i from unit j is proportional to $\gamma_i x_j$, where γ_i

is the activity of unit i and x_j is the activity of j. This learning rule can be viewed as a highly simplified model of long-term potentiation. *See* Long-term Potentiation and Depression (Cortex); Models of Neural Basis of Learning and Memory; Regulation of Synaptic Efficacy.

The effect of this rule is that the weight becomes a correlation coefficient between the activities of the units it connects. That is, the connection will become stronger (more positive) to the extent that the units are simultaneously positive or simultaneously negative. The connection will become more inhibitory (more negative) to the extent that one unit is positive while the other is negative. If there is, on average, no systematic relation between the activities of the two units, then the weight will tend toward zero, effectively disconnecting the units.

The correlational learning rule is the basis of a simple associative memory known as a 'linear associator.' In this connectionist network, there are two layers of linear units, an input layer and an output layer; each input unit is connected to every output unit, so that the output is a linear function of the input, $\mathbf{y} = \mathbf{M}\mathbf{x}$. A series of pairs of pattern vectors $(\mathbf{y}_1, \mathbf{x}_1),\ldots,(\mathbf{y}_p, \mathbf{x}_p)$ may be presented to the input and output layers of such a network and the weights adjusted according to the learning rule. The goal of the linear associator is that the network associate each $\mathbf{x}_k$ with the corresponding $\mathbf{y}_k$. In fact, it can be shown that if the set of input patterns is orthogonal, then the output $\mathbf{W}\mathbf{x}_k$ will be proportional to the desired $\mathbf{y}_k$.

Delta Rule

An improvement called the 'delta rule' can be made in the linear associator; it illustrates a fundamental approach to connectionist learning. The idea is that we define the error, as a function of the weight matrix, as the sum of the differences between the desired and actual outputs of the network, $E(\mathbf{W}) = \sum_{k=1}^{p} D(\mathbf{y}_k, \mathbf{W}\mathbf{x}_k)$. The weight matrix is then changed by 'gradient descent,' which means that the elements of $\mathbf{W}$ are changed in the relative proportion that causes a maximal incremental decrease of $E(\mathbf{W})$.

If the difference between patterns is measured by Euclidean distance squared, $D(\mathbf{y}, \mathbf{y}') = \|\mathbf{y} - \mathbf{y}'\|^2$ (i.e., the sum-of-squares error), then the delta rule is essentially equivalent to linear regression. If the input patterns are linearly independent, then the delta rule will converge to a weight matrix that associates perfectly, $\mathbf{y}_k = \mathbf{W}\mathbf{x}_k$. If they are not, then the weight matrix will be that which minimizes the total error; that is, it will be the best linear prediction of the output patterns from the input patterns.

The delta rule also illustrates an important characteristic of most connectionist networks: their ability to generalize to inputs other than those upon which they have been trained. The delta rule provides only linear generalization, but other algorithms, such as backpropagation (see the next section), can make nonlinear generalizations. Typically, connectionist categories are represented by concrete prototypes rather than by definitions in terms of necessary and sufficient conditions or other abstract symbolic structures. Network behavior then depends on similarity to the prototypes rather than on formal manipulation of symbolic structures.

Backpropagation

Gradient descent may be applied also to multilayer networks of nonlinear units, so long as the activation function is differentiable. The 'backpropagation algorithm' (also called the 'generalized delta rule') efficiently computes the weight changes by starting with the last layer and working backward layer by layer. It has been rediscovered a number of times, perhaps first by P. Werbos in 1974, but its importance in connectionism began with its rediscovery in the early 1980s. There are also special adaptations of backpropagation for recurrent networks. Backpropagation has a number of limitations, including: (1) it may be quite slow, (2) it does not necessarily take the shortest path to an error minimum, and (3) it may get trapped in local minima. Nevertheless, it remains a fundamental learning algorithm and has been subject to many practical improvements. *See* Backpropagation.

Other Learning Algorithms

The preceding are examples of 'supervised learning' procedures, which means that the correct answer $\mathbf{y}_k$ is available for each training input $\mathbf{x}_k$. Although this is appropriate for modeling some cognitive processes and for many practical problems, for others 'unsupervised learning' is preferable. In these procedures, the network is not trained to produce any specific outputs, but it allowed to group or categorize inputs according to standards inherent in its design. Thus unsupervised learning is often equivalent to some kind of statistical clustering. Between these two extremes lies 'reinforcement learning,' in which the algorithm is told whether or not the output is correct, but not what the correct output is. In unsupervised and reinforcement learning, as in supervised learning, the network is normally expected to generalize to novel inputs reasonably. There are now hundreds of connectionist learning algorithms, of greater and lesser relevance to cognitive science and neuroscience, but this must suffice for an introduction. *See* Artificial Neural Networks: Associative and Self-organizing.

Example: Learning Past Tenses

Since the early 1980s, connectionist networks have been used for an enormous number of practical applications and for modeling many aspects of cognition. One notable example must suffice here; see Rumelhart et al. (1986) for additional examples.

Rumelhart et al. (1986, vol. 2, Chapter 18) trained a connectionist network to produce the past tenses of English verbs. The inputs were vectors representing phonological features of the present tenses and the outputs were vectors representing the phonological features of the corresponding past tenses. Learning was observed to pass through three stages. In the first stage, the most common verbs were learned, essentially by rote as individual special cases. In the second stage, the network learned how to form regular past tenses (for it was able to generalize to novel regular verbs), but overgeneralized by treating the (previously correctly processed) irregular verbs as though they were regular. In the

third stage, the network (re) learned the correct formation of the irregular past tenses without losing its ability to form regular past tenses. It is interesting and highly suggestive that children pass through these same three stages; the model also exhibited errors of the same kinds made by children.

Although this experiment can be (and has been) criticized on a number of grounds as a model of language learning, it is perhaps more valuable as a demonstration of how a connectionist network can exhibit apparently rule-like behavior, but not be following any explicit rules. In particular, exceptions to the rules are handled automatically without explicit accommodation. Thus it is paradigmatic of connectionist cognitive models.

Issues

Connectionist Representations

Connectionist approaches represent and process information in a way that is fundamentally different from symbolic approaches, in which knowledge is represented in discrete structures relating atomic lexical-level 'features' (i.e., categories of the sort for which natural languages have words). In symbolic approaches, information is processed by formal logic-like rules, which rearrange these atomic units of meaning. In connectionist approaches, on the other hand, information is represented by patterns of activity distributed over large numbers of units, which individually have no lexical-level meaning. The latter are often termed 'microfeatures,' but they are fundamentally different from features, which are supposed to be complete, context-free units of meaning. Microfeatures are just components of distributed representations and are usually individually uninterpretable.

Instead of rules, connectionist information processing is defined by quantitative connections between microfeatures and takes place at a subsymbolic level (Smolensky, 1988). Cognition then is an emergent effect of large numbers of these interactions (which therefore constitute the 'microstructure' of cognition). Indeed, according to this account, apparent symbolic processing is just such an emergent effect of these subsymbolic interactions.

Traditional (symbolic) models of knowledge have been criticized for their 'brittleness.' That is, when a set of rules is formulated in an attempt to model the behavior of some expert, it is generally found that the rules do worse than the expert, since the expert applies rules flexibly and makes ad hoc exceptions to them as required by the particulars of the situation. Of course, additional rules can be formulated to cover the exceptions, but then these are likewise found to have exceptions, and so forth. The attempt to reduce flexible behavior to inflexible rules leads to a combinatorial explosion of possibilities, which exceeds the capacities of brains and computers.

Connectionist approaches seem better able to account for the flexibility and context-sensitivity of natural intelligence. This is because the connection weights function as a large number of 'soft constraints,' none of which are individually necessary or sufficient to produce a result. Therefore, connectionist networks can accommodate inputs that are exceptional in various ways, either by ignoring the anomalous aspects or by corresponding adjustment of their output. As a consequence, their performance is also robust in the face of noise in the input or damage to the network. Further, to the extent that microfeatures of the environment are represented in the input, the network can process information in a way that is sensitive to the context. *See* Artificial Intelligence: Connectionist and Symbolic Approaches; Concept Learning and Representation: Models; Connectionist Models of Concept Learning; Knowledge Representation.

Rules and Constituent Structure

Connectionist networks have been criticized for their 'opacity' or 'uninterpretability.' That is, when a connectionist network has been trained to perform some task, it is difficult to extract from the network human-interpretable rules; although the network performs correctly, we cannot see the rules it is apparently following. The reason, of course, is that it is not following rules, and the individual units and weights represent microfeatures and constraints that are lexically meaningless (i.e., have no lexical-level meaning). There are mathematical procedures for extracting rule-like information from networks, but they give only approximations to the network's behavior. This is analogous to what has been observed during knowledge acquisition for expert systems: after the fact, human experts can account in terms of rules for their decision-making, but the rules do not adequately predict the expert's future decisions.

Quite naturally, some of the severest criticisms of connectionism have come from linguists and other cognitive scientists committed to a 'language of thought,' that is, to the hypothesis that cognition must be understood in terms of the manipulation of propositional or sentential symbolic structures. These criticisms have focused on the alleged inability of flat connectionist representations to capture the rich hierarchical symbolic structure of human language and propositional attitudes, and the related sensitivity of cognition to their constituent structure. However, experiments in connectionist symbol processing have shown that connectionist networks can be sensitive to the constituent structure of representations without explicit representation of that structure and without the use of explicit symbolic rules (as in the learning of past tenses). There is much more to the issue than this, however, and the early collection by Pinker and Mehler (1988) is still a good introduction to the anti-connectionist position. *See* Artificial Intelligence: Connectionist and Symbolic Approaches.

Computability

If connectionism is viewed as a fundamentally new approach to information representation and processing, then the question arises of its power relative to conventional digital computation, as modeled by the Turing machine. At one level, this question is easy to answer. On one hand, since connectionist networks are routinely simulated on digital computers, it is apparent that they have no greater power than a Turing machine. On the other hand, researchers have shown that various sorts of connectionist networks can simulate Turing machines, which therefore have no greater power than

the networks. The conclusion would seem to be that connectionist networks are equivalent to Turing machines in computing power.

However, at a different level, the question is problematic, for the Turing machine model is based on certain idealizing assumptions about what is significant and insignificant in models of computation, assumptions that are questionable when applied to connectionist models. In particular, the Turing machine model is based on the assumption that computation proceeds by the recognition of atomic tokens of definite type according to the discrete application of finite, definite rules; these processes are assumed to operate with complete reliability. These assumptions are a poor match to connectionist approaches, in which information is represented in distributed patterns of continuous activity, and in which recognition is a matter of degree. Nevertheless, questions of computability are also important in connectionism, but relevant answers may require the development of a new theory of computation that makes idealizing assumptions more relevant to connectionist approaches.

Relation to Biological Neural Networks

Connectionist networks are often called 'neural networks' and described in terms of (artificial) neurons connected by (artificial) synapses, but is this more than a metaphor? Generally, connectionist models have reflected the contemporary understanding of neurons. For example, McCulloch and Pitts focused on the all-or-nothing character of neuron firing, and modeled neurons as digital logic gates. Newer connectionist models have had a more analog focus, and so the activity level of a unit is often identified with the instantaneous firing rate of a neuron. However, these models still ignore many important properties of real neurons, which may be relevant to neural information processing (Rumelhart et al., 1986, vol. 2, Chapter 20). As a consequence neuroscientists have stressed the differences between biological neurons and the simple units in connectionist networks; the relation between the two remains an open problem. Nevertheless, it is much easier to envision neural implementations of connectionist networks than of symbol-processing architectures.

Another divergence between biological neural nets and many connectionist models arises from the topology of the layers. Although the neurons in a layer may be numbered 1,2,…,N, these labels are arbitrary, and there is no implication that neurons 1 and 2 are any closer together in space than neurons 1 and 3. In the brain, in contrast, many neural populations are organized into 'topographic' or 'computational maps' in which the physical positions of neurons are systematically correlated with properties of some sensory or motor domain (e.g., position on the skin, location is the visual field, edge orientation, sound frequency, reach direction). Neural representations often take the form of spatially distributed fields of activity across these maps, and neural information processing in them depends crucially on spatial relations through such mechanisms as local excitation and lateral inhibition. Some connectionist models incorporate spatial relations, but many do not, and so they might not be adequate models of natural cognitive representations and processes.

Situated Embodiment

Connectionist approaches share a weakness with symbolic computational approaches in that they often treat cognitive processes as computations whose aim is to produce an appropriate output given some input; cognition is treated as an abstract process. However, natural cognition is both situated and embodied.

'Situated' refers to the fact that natural agents are embedded and engaged purposefully in a structured physical environment (often including other agents). The situated approach focuses on the effect that the structure of the environment and the agent's ongoing interaction with it has on the emergent cognitive structures that coordinate that interaction. Therefore, unsituated models, such as many connectionist models, are fundamentally incomplete and unlikely to be able to account for natural cognition. These issues can be addressed by embedding connectionist models into situated robots or sufficiently realistic simulated environments.

Cognition is 'embodied' in the sense, first, that it is implemented by neurophysiological processes, and so connectionist models should take account of that fact (see above). More importantly, however, cognition is embodied in the sense that its principal function is the control of a physical body in its environment. On the one hand, this means that biological neural networks must be capable of controlling in real time the dynamical complexities of living bodies interacting with their physical environments. On the other, it means that a biological control system can exploit the dynamical properties of the body, thereby decreasing the computational load on its nervous system. Both of these are important constraints on the architecture of biological neural networks and on the structure of emergent neural representations – indeed, on the structure of knowledge. Due to its neural orientation and respect for performance issues, connectionism is generally compatible with an embodied approach, but most connectionist models do not address embodiment.

The benefits and limitations of embodiment are investigated in such disciplines as embodied cognitive science and embodied AI; see Pfeifer and Bongard (2007) for a useful review.

Dynamic Systems Approach

The requirements of situated embodiment imply that an adequate theory of cognition must address the real-time control of the coupled dynamics of the body and its environment. However, typical connectionist models incorporate a notion of sequence, but not of time. That is, the network is understood as a functional composition in which each layer of units takes as input, the output of the preceding layer: $y = f_n(f_{n-1}(\cdots f_2(f_1(x))\cdots))$. The network defines the sequence of computations, but not their durations. Some recurrent connectionist networks produce sequences of outputs $y_1, y_2, \ldots$ from corresponding sequences of inputs $x_1, x_2, \ldots$ by feedback of an internal state variable r: $y_i = f(x_i, r_{i-1})$, $r_i = g(x_i, r_{i-1})$. However, such models do not address the essential dynamical, real-time coupling of an agent's nervous system, body, and

environment. To address these issues, the 'dynamic systems approach' views neural networks as dynamical systems and investigates issues such as stability and attractor dynamics that lead to the emergence of representations (e.g., Thelen and Smith, 1994). In particular, dynamic field theory incorporates dynamics and computational maps into connectionism. *See* Classical Mechanics and Motor Control; Motor Control Models: Learning and Performance; Self-Organizing Dynamical Systems.

See Churchland (1986) and Quinlan (1991) for an introduction to connectionist approaches. *See* Connectionist Models of Concept Learning.

See also: Artificial and Natural Computation; Connectionist Models of Concept Learning; Learning and Memory, Neural Basis of; Learning: Mathematical Learning Theory, History of; Long-Term Potentiation and Long-Term Depression; Mathematical Psychology: History; Mathematical Psychology.

Bibliography

Anderson, James A., Pellionisz, Andras, Rosenfeld, Edward (Eds.), 1990. Neurocomputing 2: Directions for Research. MIT Press, Cambridge, MA.

Anderson, James A., Rosenfeld, Edward (Eds.), 1988. Neurocomputing: Foundations of Research. MIT Press, Cambridge, MA.

Churchland, Patricia S., 1986. Neurophilosophy: Toward a Unified Science of the Mind-Brain. MIT Press, Cambridge, MA.

Dreyfus, Hubert L., 1972. What Computers Can't Do. Harper & Row, New York, NY.

Hebb, Donald O., 1949. The Organization of Behavior. Wiley, New York, NY.

Hinton, Geoffrey E., Anderson, James A. (Eds.), 1989. Parallel Models of Associative Memory, revised ed. Lawrence Erlbaum Associates, Hillsdale, NJ.

Hopfield, John J., 1982. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences of the United States of America 79, 2554–2558.

Lashley, Karl S., 1950. In search of the engram. Symposia of the Society for Experimental Biology 4, 454–482.

Minsky, Marvin, Papert, Seymour, 1969. Perceptrons. MIT Press, Cambridge, MA.

Pfeifer, Rolf, Bongard, Josh C., 2007. How the Body Shapes the Way We Think: A New View of Intelligence. MIT Press, Cambridge, MA.

Pinker, Steven, Mehler, Jacques (Eds.), 1988. Connections and Symbols. MIT Press, Cambridge, MA.

Quinlan, Philip T., 1991. Connectionism and Psychology: A Psychological Perspective on New Connectionist Research. University of Chicago Press, Chicago, IL.

Rumelhart, David E., McClelland, James L., and the PDP Research Group, 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 2. MIT Press, Cambridge, MA.

Smolensky, Paul, 1988. On the proper treatment of connectionism. Behavioral and Brain Sciences 11, 1–74.

Thelen, Esther, Smith, Linda B., 1994. A Dynamic Systems Approach to the Development of Cognition and Action. MIT Press, Cambridge, MA.

Relevant Websites

Anon, (Ed.). Embodiment and situated cognition. PhilPapers: Online Research in Philosophy – http://philpapers.org/browse/embodiment-and-situated-cognition.

Anon, (Ed.). Philosophy of connectionism. PhilPapers: Online Research in Philosophy – http://philpapers.org/browse/philosophy-of-connectionism.

Garson, James, 2010. Connectionism. Stanford Encyclopedia of Philosophy – http://plato.stanford.edu/entries/connectionism/.

Rescoria, Michael (Ed.). Computation and representation. PhilPapers: Online Research in Philosophy – http://philpapers.org/browse/computation-and-representation.

Cognitive Neuropsychology, Methods of

Saskia Kohnen and Lyndsey Nickels, Macquarie University, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by T. Shallice, volume 3, pp. 2128–2133, © 2001, Elsevier Ltd.

Abstract

This article provides an overview of some of the most commonly used methodologies in the discipline of cognitive neuropsychology: dissociations, associations, single case studies, case series, intervention research, computational modeling and the cognitive neuropsychological approach to studying developmental disorders of cognition.

Cognitive neuropsychologists study impaired cognition to inform theories of normal cognition. They do so by observing cognitive behavior in people with impaired cognition, for example, people who have acquired a brain injury. Cognitive neuropsychologists are interested in what data from these individuals reveal about the processes and representations of cognitive systems (McCloskey and Caramazza, 1988).

Dissociations and Associations

Cognitive neuropsychological studies have made major contributions to our understanding of the functional architecture of cognition (e.g., reading: Coltheart et al., 2001; Marshall and Newcombe, 1973; spoken language production: Badecker et al., 1995; Dell et al., 1997; spelling: Miceli and Capasso, 2006; Tainturier and Rapp, 2001). Under assumptions of modularity (Fodor, 1983; see Caramazza, 1992; Coltheart, 1999 for different versions of this assumption), the architecture of cognitive systems is composed of information-processing components that each fulfill a different purpose. These relatively autonomous components can also be independently damaged, for example, as the result of a brain injury. When such a functionally independent component is selectively impaired this will result in a 'dissociation.' Shallice (1988) distinguishes between classical, strong, and trend dissociations. In the most powerful dissociation, the classical dissociation, performance on one task is in the normal range, while performance on another task is very impaired. Strong and trend dissociations consist of less marked differences in task performance. These three dissociations rely on the use of control subjects (who are unimpaired in performing the tasks in question) to determine what score should be considered normal or impaired. John Crawford and his colleagues have provided the statistical means to allow detection of a deficit in a single patient compared to a (relatively small) control group, deliver a point estimate of the abnormality of the patient's score and determine the difference between scores on two tasks (Crawford and Garthwaite, 2006; Crawford and Howell, 1998; Crawford et al., 2009).

An example of an interesting, theoretically important dissociation is that of JOH, who showed an intriguing pattern of performance (Costello and Warrington, 1987). JOH presented with a reading disorder – left neglect dyslexia – that led to reading errors at the beginning of words (e.g., reading *make* as 'cake' or *least* as 'beast'). However, this was not due to a general neglect of the left hand side of space. In fact, he showed evidence of a right-sided visuospatial neglect. That is, when processing (e.g., copying) pictures or figures, he omitted information on the right hand side. For example, when asked to copy a row of three shapes he would omit the last (rightmost) shape. This dissociation shows that there is not a single mechanism underlying neglect in both visuospatial tasks and reading, but rather there are two separate mechanisms, which can be separately impaired.

Note that this dissociation is particularly clear as the argument is not in terms of the extent of the impairment in one task or another (both tasks were impaired) but the type of impairment (reading impaired on the left and intact on the right, visuospatial tasks impaired on the right and intact on the left). In cases where the dissociation is in terms of one task being impaired and the other intact, it could be that this is due to a difference in the difficulty of the two tasks rather than independent mechanisms. For example, interpretation can be open to the potential argument that one task is simply harder than the other (requires more processing resources). Imagine a dissociation between recognition of faces and recognition of printed words. Could it be that this dissociation in fact reflects damage to a single system that impairs face recognition without impairing visual word recognition, because faces are visually more complex, and so stress this system more. However, if there is a case where printed word recognition is impaired and face processing is intact, this would refute the difficulty argument. This pair of opposite dissociations makes up a 'double dissociation,' which has long held to be the strongest evidence for separate processing components.

While cognitive neuropsychology has been accused of being overly interested in dissociations (Laine and Martin, 2012), the study of human cognition often requires testing for associations. Hence, many cognitive neuropsychological studies set out with the explicit aim to discover associations. One example of looking for associations can be found in a multiple single case study on letter position dyslexia (Friedmann and Rahamim, 2007). The main symptom of letter position dyslexia is a disproportionately high number of anagram reading errors (e.g., *slime* read as 'smile'). Friedmann and Rahamim proposed that this error type is caused by an impairment early on in the reading process (the

orthographic visual analyzer). Hence, they predicted that anagram errors should occur in all reading tasks (whether reading of nonwords or words, whether they included reading aloud or not). This association was indeed what they found (also see Section Computational Cognitive Neuropsychology for more examples of studying systematic associations).

Single Case Studies

The investigation of a single case is considered an important (but not the only) methodology in cognitive neuropsychology. Single cases may present with a behavior that cannot be explained by current theories and hence may be used to test, extend, and/or change current understanding in an area of cognition. One of best known single cases to inform cognitive theory is probably that of HM (Scoville and Milner, 1957). The first report of HM has not only attracted over 4000 citations (by 2013) but also has been argued to represent the beginning of the modern study of memory (Squire, 2009). HM was an epilepsy patient who after brain surgery had lost his capacity to remember daily events. At the same time, his intellectual and perceptual functioning remained intact. Before Brenda Milner and her colleagues began to study HM, memory function was thought to be integrated with intellectual and perceptual processes. HM's case helped establish the theory that memory is independent from these other processes.

The main reason why single case studies are considered both powerful and valid rests on the finding that individuals with cognitive disorders are by their nature variable (Caramazza and Coltheart, 2006). Due to the high number of possible functional lesions, there is only a theoretical chance that two patients will ever display the exact same impairment (Caramazza, 1986; Caramazza and Coltheart, 2006; Coltheart, 1984). In commenting on the huge variability in the actual presentation of hemispatial neglect in patients, Buxbaum (2006) evokes the analogy of snowflakes, each of which, of course, is unique. The snowflake analogy is probably valid for most areas of cognition. For example, Schmalzl and colleagues (Schmalzl et al., 2008) studied seven members of a family who had face-processing difficulties (prosopagnosia). These individuals all had difficulties recognizing the faces of family members, four of them could not even identify their own face reliably in a photograph. However, a more detailed investigation showed different underlying impairments for these family members' difficulties in recognizing face. In fact, there were no two family members that showed the exact same pattern of difficulties. For example, some but not all of the family members had difficulties detecting facial expressions. Similarly, some but not all family members had problems detecting when spacing between features of the faces (e.g., nose, mouth) was changed. This is an example of a study that shows how face processing difficulties can be caused by different underlying processes, even within the same family.

Due to this variability, cognitive neuropsychologists tend not to study syndromes. Labels such as prosopagnosia, surface dyslexia, or even Broca's aphasia may be used but they are mere 'short hand' terms for a gross level of symptomatology (i.e., problems recognizing faces; problems reading irregular words; limited syntactic structure in speech production) that hide the complexity of each patient's presentation and the high variability between patients. Studying the average performance of groups of patients that belong to the same syndrome complex could mask important differences between individuals as well as produce averages that do not describe a single individual in this group. However, single case studies have been criticized for their lack of generalizability, lack of control, and poor statistical methods (but see e.g., Robey et al., 1999). Lack of control and statistical problems are easily taken care of by using appropriate designs including converging evidence from replications, multiple experiments of the function in question, and employing some of the myriad of appropriate statistics (e.g., McNemar's, Fisher's exact, Wilcoxon). Generalizability can be improved by studying symptoms (rather than syndromes) in multiple single cases (Friedmann and Rahamim, 2007; Warrington, 1975). Indeed, many reports of new dissociations within a syndrome describe multiple cases, which are examined for their differences and similarities in performance; both in reference to previous cases and within the sample reported in the new paper.

If patients are so highly variable in terms of their functional deficits, how is it ever possible to generalize beyond the level of the individual? This relies on two assumptions. First, the assumption of 'universality' poses that neurologically intact people do not differ from each other in regard to their cognitive architectures (Caramazza, 1986). It is worthwhile to note that this assumption underlies all studies of cognitive psychology, otherwise inferences about cognition could not be made from small study samples drawn from the general population. Moreover, some degree of individual difference is not incompatible with the universality assumption (Rapp, 2011). Second, if inferences can be drawn from neurologically impaired patients to normal cognition, it needs to be the case that the injury simply subtracted one or more cognitive functions without also altering or adding new functions (Saffran, 1982). Therefore, studying an individual with a selective impairment provides a unique experimental technique for studying human cognition, in particular, a system that misses exactly one function. However, single case studies are only one of the methodologies available to the cognitive neuropsychologist.

Case Series

Another form of study with multiple participants that is popular in cognitive neuropsychology is the case series (Schwartz and Dell, 2010). McCloskey and Caramazza (1988) distinguish between two ways of studying groups of people: *Group studies*, which report average performance of the group in a way that individual results are invisible vs *case series*, in which individual data for each participant are reported. Averaging is only appropriate when variability in the data is assumed to be due to noise (McCloskey and Caramazza, 1988). As noted above, one of the problems in reporting agglomerate data in classical group studies is that what describes the whole group may not describe any single individual. Case series do not suffer from this problem. Case

series designs have the potential to allow the identification of central tendencies in the data, which reflect the actual performance of several individuals, while still including variability typical of larger populations (Shallice and Buiatti, 2011).

In patient data, heterogeneity may be informative in that it may reveal systematic relationships among different variables (Schwartz and Dell, 2010). While heterogeneity in the dependent variable is required to carry out such investigations, it is important that the source of the variability is relevant to the question in hand (Nickels et al., 2011). For example, when investigating word retrieval in a group of aphasic speakers, the individuals should vary in terms of their naming accuracy. At the same time, factors that could potentially impact on performance on the dependent variable (e.g., picture naming), but are not subject of the particular investigation, such as attention or object recognition need to be kept similar between the individuals. Which exact variables need to be kept similar and which should vary depends on what theory a study aims to investigate.

There are certain theoretical questions that require a case series methodology. Rapp (2011: p. 437) summarizes them as "As X increases (or decreases), Y should increase (or decrease)." Hence, the study would select patients with varying degree of deficits in skill X and test their ability to do Y. Rapp's examples of X and Y include the severity of a semantic impairment and accuracy for exception word reading (Woollams et al., 2007); or severity of sublexical damage and perseveration errors in spelling (Fischer-Baum and Rapp, 2012). Note that all of these studies are interested in a theoretically relevant association between two functions.

Ideally, case series should include a combination of standard background assessment and experimental tests. These should be performed on all participants in a case series and also take into account items and tasks that have been used in previous studies (Nickels et al., 2011). Using similar tasks and items not only makes it easier to compare cases across studies but also helps in testing and designing computational models of cognition.

Computational Cognitive Neuropsychology

The architecture of cognitive processes is often described in verbal or diagrammatic models. These consist of verbal descriptions or box and arrow type diagrams. Seidenberg and McClelland (1989) were the first to design a computational model of reading aloud. Since then, more cognitive theories have been implemented computationally, including one of the most well-known reading models, the Dual Route model (e.g., implemented as the Dual Route Cascaded model: Coltheart et al., 2001), but also models of spelling (e.g., Houghton and Zorzi, 2003) and spoken language (Dell et al., 1997; Levelt et al., 1999; Roelofs, 1997). A good model, be it verbal, diagrammatic, or computational, should be able to account for the full body of evidence pertaining to both normal and impaired cognition.

Cognitive neuropsychological principles can also be applied to these computational models. They can be 'lesioned' in order to display symptoms described in patients (Coltheart, 2006). Lesioning a model can be achieved in different ways, depending on the model. For example, the strengths of connections can be adjusted or noise may be introduced. Whatever change to the 'normal' computational model best fits the patient data, can be regarded as an account of the type of impairment to the cognitive system. Damaging a computational model in ways that will match the data of patients can provide converging evidence for a certain theory, and it can also allow a specification in how exactly cognitive processes are impaired when a certain pattern of impairment is shown.

Case series (see Section Case Series) can play a great role in testing these models. Nickels (1995) examined the errors produced in picture naming by a case series of people with aphasia. Specifically, she examined the relationship between production of semantic (doctor named as 'nurse') and phonological (doctor named as 'doptor') errors and the psycholinguistic variables that affected these errors. Imageability was found to affect the production of semantic errors (but not subjective frequency, or length). Whereas for phonological errors subjective frequency and number of phonemes showed significant effects (but not imageability). Nickels used these patterns to adjudicate between the predictions of highly interactive and strictly feedforward models, arguing that the data were inconsistent with the interactive models of speech production.

Theoretical Impact of Treatment Studies

In its aim to understand human cognition, cognitive neuropsychological studies can assess patients' performance on a multitude of tasks pertaining to the area of cognition under investigation. One way to adjudicate between different theories of cognition is to design an experiment in which different theories predict different outcomes. The experimental design can be an intervention. Hence, another powerful tool to test theories is the use of treatment (Nickels et al., 2010). For example, Biedermann and Nickels (2008a,b) investigated the lexical representation of homophones using treatment (see Nickels et al., 2010 for other examples of studies using intervention to investigate cognitive theories). Homophones are words that are pronounced in the same way but have different meanings (e.g., tank: fish-tank, military tank; flour/flower). One theory proposes that each version of a homophone has its own lexical representation (Caramazza, 1997), while an opposing theory posits homophones share a representation (Levelt et al., 1999). Biedermann and Nickels designed an intervention study to provide new insights into this debate. Intervention can help people with word finding difficulties to become better at retrieving the words practiced during intervention. However, usually, effects are restricted to practiced words and will not benefit words that were not explicitly practiced during the intervention. Following this logic, Biedermann and Nickels made the following differential predictions: if homophones share representations, treating one partner of a homophone (e.g., flower; (fish) tank) should lead to improvements in retrieving the untrained partner of the homophone (i.e., flour; (military) tank). In contrast, if

each homophone partner has its own lexical representation, only the trained words (e.g., flower; (fish) tank) should improve. The results favored the theory that homophones share lexical representations since not only the trained homophones but also their untrained partners improved after training. Note that a set of control words that sounded similar to the treated words (but were not homophone partners) or were similar in meaning did not improve.

Other than providing insights into theoretical issues cognitive neuropsychological approaches to treatment can be used to assess the efficacy of treatment protocols and procedures. Moreover, there are many examples of treatment studies where the aim was to explain how a particular treatment works within a particular cognitive architecture. Often, these explanations in effect provide converging evidence for one particular theory. An example of this is a treatment study conducted with a person who had both deep dyslexia and anomia (Nickels, 1992; see Nickels et al., 2010 for further examples). Deep dyslexia refers to impaired reading that is marked by the presence of semantic reading errors (e.g., reading *wine* as 'beer'), visual errors (e.g., *late* as 'lake'), morphological errors (e.g., *swims* as 'swimming'), a word class effect (i.e., function words worse than content words), and an inability to read nonwords. Nickels' participant TC also made many semantic errors when asked to name pictures. The intervention consisted of reteaching letter–sound correspondences. While this did not lead to improved nonword reading (because the participant remained unable to blend sounds together to form a word), the frequency of semantic errors decreased after intervention. These results were later used as supporting evidence for the 'summation theory' according to which phonological information feeds back into the phonological lexicon (Hillis and Caramazza, 1995). After the intervention, TC started to 'self-cue' by using his knowledge of the initial letter of a word he was supposed to say and generating the sound (Nickels, 1992). For example, if he was asked to name the picture of a *tiger*, TC would activate semantically related words in addition to the correct word (e.g., leopard, panther, tiger, lion). Being able to retrieve the written word form, and sounding out the initial letter T, provided additional activation to the correct response allowing selection, and reduction of semantic errors.

The cognitive neuropsychological approach to intervention also advocates that, with use of a strong design and appropriate statistics, single case intervention studies can provide a high level of evidence for the efficacy of treatment. Many cognitive neuropsychologists would argue, once again, that a series of single case studies can provide stronger and more valid evidence than evaluating average effects over a group study (such as advocated in randomized controlled trials) (also see Wright, 2012 for a discussion of this issue).

Acquired and Developmental Cognitive Neuropsychology

Traditionally, the cognitive neuropsychological approach has been applied to studies with individuals with acquired disorders, that is, people who are now impaired in a skill that they had fully acquired prior to their brain injury. However, in the 1980s, the methodologies of cognitive neuropsychology were extended to the developmental domain, where a cognitive system is still in the process of being acquired. This work was pioneered by John Marshall, Jane Holmes, and Christine Temple. Developmental cognitive neuropsychology investigates developmental disorders using models of the cognitive architecture of the adult system. The strong claim underlying this approach is that in developmental disorders, highly selective disorders should be observable (Castles et al., 2006; Marshall, 1984; Temple, 1997). This is in contrast to a view according to which a disorder early in development distorts the system in more than one way since development is an interactive process (see Bishop, 1997). For example, if a child has difficulties learning to say words, they may also have difficulties to learn the sounds of a language. The degree to which development of different cognitive functions happens independently is an ongoing topic of research. However, there is now a large literature detailing the vast differences in the presentation of children with developmental dyslexia (see Castles et al., 2006; Marshall, 1984): for almost every processing component in the reading architecture (in this case dual route models of reading), there are published case reports of developmental cases who have a specific difficulty acquiring the skill that is governed by this processing component (for overviews of these disorders see Coltheart and Kohnen, 2012; Jones et al., 2011). These selective difficulties can be taken as evidence that development of cognitive skills can, to some degree, be modular.

In the case of developmental dyslexia, cognitive neuropsychologists refer to dual route architecture. Evidence for this architecture is based on data from adult readers. Hence, one could call dual route models 'adult models' of reading. What is somewhat controversial is whether developmental disorders should be studied within the framework of an adult system; a cognitive system that is presumably static and not learning and developing. Bishop (1997) argues that the course of development is typically highly interactive, with skills and competencies influencing each other. Therefore, children usually present with complex patterns of associated impairments, rather than highly selective deficits. While pure cases of selective difficulties do indeed tend to be rare (Castles et al., 2006), they nevertheless make contributions to our understanding and advance developmental theories of cognition. For example, one of the most prominent current theories of how children acquire a written vocabulary posits that lexical acquisition depends largely on nonlexical skills (Share, 1999). After one or more exposures where a novel word is sounded out using nonlexical skills, a representation of the word will be developed in the orthographic lexicon. However, some children acquire a reading vocabulary that is at a normal level while their nonlexical skills are severely impaired. This is called developmental phonological dyslexia. While Share's theory is plausible, it is cases of phonological dyslexia that clearly show that this theory alone cannot explain how children acquire a reading vocabulary. This is just one example of how the study of selective deficits in developmental disorders is important in furthering our understanding of the acquisition of cognitive functions.

Conclusion

Cognitive neuropsychology aims to understand and uncover the cognitive architecture and processing that underlies both normal and impaired cognition. It does so by studying children and adults with cognitive disorders, or lesioning computational models of cognition. Note that many of the methodologies described above are not only used by cognitive neuropsychologists. For example, single case studies form an important methodology in medical science. Similarly, the use of treatment studies to inform theory is also used in many other disciplines, including, for example, in developmental psychology. Embracing multiple methodologies is one way to further understanding of human cognition and ultimately the interaction of brain and mind.

See also: Case Study: Methods and Analysis; Cognitive Neuroscience.

Bibliography

Badecker, W., Miozzo, M., Zanuttini, R., 1995. The two stage model of lexical retrieval: evidence from a case of anomia with selective preservation of grammatical gender. Cognition 5, 193–216.

Biedermann, B., Nickels, L., 2008a. The representation of homophones: more evidence from remediation of anomia. Cortex 44, 276–293.

Biedermann, B., Nickels, L., 2008b. Homographic and heterographic homophones in speech production: does orthography matter? Cortex 44, 683–697.

Bishop, D.V.M., 1997. Cognitive neuropsychology and developmental disorders: uncomfortable bedfellows. The Quarterly Journal of Experimental Psychology: Section A 50, 899–923.

Buxbaum, L.J., 2006. On the right (and left) track: twenty years of progress in studying hemispatial neglect. Cognitive Neuropsychology 2, 156–173.

Caramazza, A., 1986. On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: the case for single-patient studies. Brain and Cognition 5, 41–66.

Caramazza, A., 1992. Is cognitive neuropsychology possible? Journal of Cognitive Neuroscience 4, 80–95.

Caramazza, A., 1997. How many levels of processing are there in lexical access? Cognitive Neuropsychology 14, 177–208.

Caramazza, A., Coltheart, M., 2006. Cognitive Neuropsychology twenty years on. Cognitive Neuropsychology 23, 3–12.

Castles, A., Bates, T.A., Coltheart, M., 2006. John Marshall and the developmental dyslexias. Aphasiology 20, 871–892.

Coltheart, M., 1984. Acquired dyslexias and normal reading. In: Malatesha, R.N., Whitaker, H.A. (Eds.), Dyslexia: A Global Issue. Martinus Nijhoff, The Hague, The Netherlands.

Coltheart, M., 1999. Modularity and cognition. Trends in Cognitive Sciences 3, 115–120.

Coltheart, M., 2006. Acquired dyslexias and the computational modelling of reading. Cognitive Neuropsychology 23, 96–109.

Coltheart, M., Kohnen, S., 2012. Acquired and developmental disorders of reading and spelling. In: Faust, M. (Ed.), The Handbook of the Neuropsychology of Language. Wiley-Blackwell Publishers, Chichester, pp. 892–920.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., Ziegler, J., 2001. DRC: a dual route cascaded model, of visual word recognition and reading aloud. Psychological Review 108, 204–256.

Costello, A.D., Warrington, E.K., 1987. The dissociation of visuospatial neglect and neglect dyslexia. Journal of Neurology, Neurosurgery & Psychiatry 50, 1110–1116.

Crawford, J.R., Garthwaite, P.H., 2006. Methods of testing for a deficit in single case studies: evaluation of statistical power by Monte Carlo simulation. Cognitive Neuropsychology 23, 877–904.

Crawford, J.R., Garthwaite, P.H., Howell, D.C., 2009. On comparing a single case with a control sample: an alternative perspective. Neuropsychologia 47, 2690–2695.

Crawford, J.R., Howell, D.C., 1998. Comparing an individual's test score against norms derived from small samples. The Clinical Neuropsychologist 12, 482–486.

Dell, G.S., Schwartz, M.F., Martin, N., Saffran, E.M., Gagnon, D.A., 1997. Lexical access in normal and aphasic speech. Psychological Review 104, 801–838.

Fodor, J.A., 1983. The Modularity of Mind. MIT Press, Cambridge.

Fischer-Baum, S., Rapp, B., 2012. Underlying causes of letter perseveration errors in dysgraphia. Neuropsychologia 50, 305–318.

Friedmann, N., Rahamim, E., 2007. Developmental letter position dyslexia. Journal of Neuropsychology 1, 201–236.

Hillis, A.E., Caramazza, A., 1995. Converging evidence for the interaction of semantic and sublexical phonological information in accessing lexical representations for spoken output. Cognitive Neuropsychology 12, 187–227.

Houghton, G., Zorzi, M., 2003. Normal and impaired spelling in a connectionist dual-route architecture. Cognitive Neuropsychology 20, 115–162.

Jones, K., Castles, A., Kohnen, S., 2011. Subtypes of developmental dyslexia: recent developments and directions for treatment. Acquiring Knowledge in Speech, Language and Hearing 13, 79–83.

Laine, M., Martin, N., 2012. Cognitive neuropsychology has been, is, and will be significant to aphasiology. Aphasiology 26, 1362–1376.

Levelt, W.J.M., Roelofs, A., Meyer, A.S., 1999. A theory of lexical access in speech production. Behavioral & Brain Sciences 22, 1–75.

Marshall, J.C., 1984. Toward a rational taxonomy of the developmental dyslexias. In: Malatesha, R.N., Whitaker, H.A. (Eds.), Dyslexia: A Global Issue. Martinus Nijhoff, The Hague.

Marshall, J.C., Newcombe, F., 1973. Patterns of paralexia: a psycholinguistic approach. Journal of Psycholinguistic Research 2, 175–199.

McCloskey, M., Caramazza, A., 1988. Theory and methodology in cognitive neuropsychology: a response to our critics. Cognitive Neuropsychology 5, 583–623.

Miceli, G., Capasso, R., 2006. Spelling and dysgraphia. Cognitive Neuropsychology 23, 110–134.

Nickels, L., 1992. The autocue? Self–generated phonemic cues in the treatment of a disorder of reading and naming. Cognitive Neuropsychology 9, 155–182.

Nickels, L.A., 1995. Getting it right? Using aphasic naming errors to evaluate theoretical models of spoken word production. Language and Cognitive Processes 10, 13–45.

Nickels, L., Kohnen, S., Biedermann, B., 2010. An untapped resource, treatment as a tool for revealing the nature of cognitive processes. Cognitive Neuropsychology 27, 539–562.

Nickels, L., Howard, D., Best, W., 2011. On the use of different methodologies in cognitive neuropsychology: drink deep and from several sources. Cognitive Neuropsychology 28, 475–485.

Rapp, B., 2011. Case series in cognitive neuropsychology: promise, perils, and proper perspective. Cognitive Neuropsychology 28, 435–444.

Robey, R.R., Schultz, M.C., Crawford, A.B., Sinner, C.A., 1999. Single-subject clinical-outcome research: designs, data, effect sizes, and analyses. Aphasiology 13, 445–473.

Roelofs, A., 1997. The WEAVER model of word-form encoding in speech production. Cognition 64, 249–284.

Saffran, E.M., 1982. Neuropsychological approaches to the study of language. British Journal of Psychology 73, 317–337.

Scoville, W., Milner, B., 1957. Loss of recent memory after bilateral hippocampal lesions. Journal of Neurology, Neurosurgery, and Psychiatry 20, 11.

Schmalzl, L., Palermo, R., Coltheart, M., 2008. Cognitive heterogeneity in genetically based prosopagnosia: a family study. Journal of Neuropsychology 2, 99–117.

Schwartz, M., Dell, G., 2010. Case series investigations in cognitive neuropsychology. Cognitive Neuropsychology 27, 477–494.

Seidenberg, M.S., McClelland, J.L., 1989. A distributed, developmental model of word recognition and naming. Psychological Review 96, 523.

Shallice, T., 1988. From Neuropsychology to Mental Structure. Cambridge University Press, Cambridge, UK.

Shallice, T., Buiatti, R., 2011. Types of case series: the anatomically-based approach. Cognitive Neuropsychology 28, 500–514.

Share, D.L., 1999. Phonological recoding and orthographic learning: a direct test of the self-teaching hypothesis. Journal of Experimental Child Psychology 72, 95–129.

Squire, L.R., 2009. The legacy of patient HM for neuroscience. Neuron 61, 6–9.

Tainturier, M.-J., Rapp, B., 2001. The spelling process. In: Rapp, B. (Ed.), The Handbook of Cognitive Neuropsychology: What Deficits Reveal about the Human Mind. Psychology Press, Philadelphia.

Temple, C., 1997. Developmental Cognitive Neuropsychology. Psychology Press, Hove, UK.

Warrington, E.K., 1975. The selective impairment of semantic memory. The Quarterly Journal of Experimental Psychology 27, 635–657.

Woollams, A.M., Lambon Ralph, M.A., Plaut, D., Patterson, K., 2007. SD-squared: on the, association between semantic dementia and surface dyslexia. Psychological Review 114, 316–339.

Wright, M., 2012. Cognitive neuropsychological approach to aphasia treatment: implications for evidence-based clinical practice. Aphasiology 26, 1391–1396.

Cognitive Neuroscience

James L McClelland, Center for Mind, Brain, and Computation, Stanford University, Stanford, CA, USA
Matthew AL Ralph, School of Psychological Sciences, The University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.L. McClelland, volume 3, pp. 2133–2140, © 2001, Elsevier Ltd.

Abstract

Cognitive neuroscience explores the neural basis of cognition, including perception, attention, language understanding, memory, problem solving, and decision-making. The field draws on findings on how neurons process and represent information, and on ideas about how learning may occur through the modification of properties of neurons and their connections. It is clear that there is specialization of function in the brain, yet brain areas appear to work together, interactively, to support emergent cognitive functions. The article discusses these points, reviews research methods used in this field, and touches on open questions, as well as the future of cognitive neuroscience.

What is the neural basis of cognition? How do our thoughts, perceptions, beliefs, and intentions arise from the activity of the vast numbers of neurons in the brain? The discipline of cognitive neuroscience emerged in the 1990s at the interface between neurobiological, cognitive, and computational sciences to answer these questions. On one side, the discipline grows out of the traditions of both cognitive psychology and neuropsychology, which use behavioral experiments to uncover the processes and mechanisms underlying normal and impaired human cognitive processes, linking work of this kind to computational modeling approaches to develop explicit mechanistic accounts of these functions and dysfunctions. On the other side, it emerges from the traditions of neuroscience, which use neurophysiological and neuroanatomical methods to explore the mechanisms underlying complex functions and draws on findings and principles of cellular and molecular neuroscience. It joins these approaches with the use of both structural and functional brain imaging methods, and the application of formal mathematical models (computational neuroscience) to explore how neural processes give rise to cognitive outcomes.

The Microstructure of Cognition

Patterns of Activity Arising in Ensembles of Simple Elements

A starting point for cognitive neuroscience is the idea that a cognitive or mental state consists of a pattern of activity distributed over many neurons. For example, the experience of holding, sniffing, and viewing a rose is reflected in a complex pattern of neural activity, distributed over many brain regions, including the participation of neurons in somatosensory, olfactory, and visual regions, and possibly extending to language areas (that represent the words we use and hear in relation to this object) and other neural regions that code information associated with the experience.

These patterns of activation arise from excitatory and inhibitory interactions among the participating neurons, mediated by connections called synapses. The inputs that neurons receive cause them to 'fire' or emit electrical pulses, called spikes or action potentials, which travel down their axons to synaptic terminals where they cause the release of chemicals (neurotransmitters) that then have excitatory or inhibitory influences on the neurons on the other side of the synapse. The combined effect of the incoming signals to each neuron, together with its recent history, determines whether it will fire at a particular moment. Figure 1 indicates something

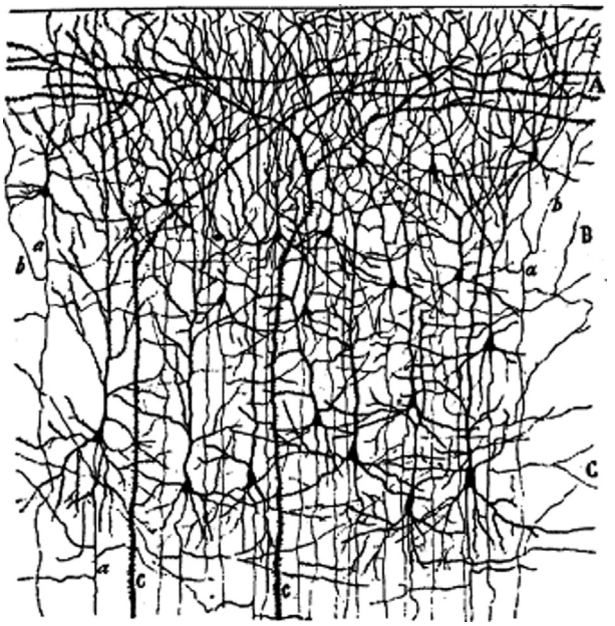

Figure 1 An early camera lucida drawing of the circuitry of the neocortex, based on the Golgi stain method, which impregnates just one out of every 100 cortical neurons. The diagram depicts the rich dendritic branching structure of the individual neurons present, whose cell bodies appear as small, pyramid-shaped blobs. The dendrites (and the little spines visible on the surfaces of some of the larger dendrites) are the structures on which the neurons receive most of their inputs from other neurons. This figure comes from a book by the Spanish neuroscientist Santiago Ramon y Cajal. I recently retrieved it from Wikimedia Commons on 2009-11-20. It was originally published in Santiago Ramon y Cajal 'Comparative study of the sensory areas of the human cortex' published 1899, ISBN 9781458821898. I believe it is in the public domain.

of the fundamental circuitry involved, though it should be noted that only one out of 100 of the neurons in the tiny region shown (about 1×1 mm) are indicated, and the synaptic connections onto the neurons illustrated are not visible. While the computations performed by individual neurons should not be underestimated (*see* Neurons and Dendrites, Integration of Information in), it seems likely that what gives the system its power and complexity is the number of neurons involved (most estimates place the number in the human brain between 10 and 100 billion) plus the density and pattern of connections among them (typical cortical neurons receive between 10 000 and 100 000 individual synapses from other neurons).

Distributed Representations

A great deal of research has concerned the nature of the active representations that the brain uses (cf rose example above). There is now a great deal of support for the view that the brain's representations typically consist of activity patterns involving many thousands or millions of neurons. Individual neurons are often described as 'detectors' for particular stimulus or situational features or conjunctions of features (e.g., the 'edge detectors' introduced by Hubel and Weisel (1962) in their seminal studies in visual cortex), but most such neurons are fairly broadly tuned, so that they will also be partially activated by a wide range of stimuli overlapping in one way or another with the optimal stimulus, and thus will participate at least partially in the representation of many different inputs.

A prime example of distributed representation is that of the direction of arm movements in the motor cortex. It appears that the representation of a particular direction of reaching reflects a pattern of distributed activity over a large population of neurons, each of which responds maximally to a particular preferred direction, but responds to a lesser degree to neighboring directions, and thus participates partially in the representation of many different directions of reaching (Georgopoulos et al., 1986). There are other types of distributed representations used in the brain, in which a neuron can participate in two different representations, without there being a clear shared feature or other similarity between the situations that cause the neuron to fire. For example, in the hippocampus, individual neurons participate in distributed representations of the animals' location in external space and other aspects of the current behavioral situation. Interestingly, the same neurons may participate in different ways in the representation of different environments, or even of two distinct representations of the same environment when the animal performs different tasks (Markus et al., 1995). Some studies suggest that there may be neurons specialized for items or persons who are highly familiar (e.g., a current actress frequently described in the popular media, Quiroga et al., 2005).

Knowledge and Learning in the Strengths of Connections

The particular pattern of activation that arises in experiencing an input (or in reconstructing a memory or formulating an imagined experience) is determined by the connections among the neurons. A key issue, then, is to understand the processes that lead to the formation of the specific excitatory and inhibitory connections that shape the processes of perception, cognition, and action. Generally, it is thought that largely activity-independent processes establish an initial skeleton framework of connectivity early in development, for example, causing connections to form between neurons in the retina of the eye and other neurons in the lateral geniculate nucleus, a way station for visual information on the way to the cortex. Then, activity-dependent processes selectively refine and stabilize some of the connections, and perhaps cause new ones to form, while other connections are pruned away.

Activity-dependent processes continue throughout life, at least in many parts of the brain, and appear to provide the basis of both explicit and implicit learning. They have been the subject of intense scrutiny in neuroscience. Donald Hebb, the mid-twentieth-century neuropsychologist, proposed that if one neuron participates in firing another, the connection from the first to the second will be strengthened (Hebb, 1949). Hebb's idea has been encapsulated in the phrase 'cells that fire together wire together.' While there is no direct proof that this is a principle of learning in the brain, the idea has received a great deal of experimental support in experiments that have been carried out in slices of brain tissue (*see* Neural Plasticity). It should be understood that there may also be plasticity at the level of the whole neuron (in some specialized brain areas, neurons are continually created and incorporated into circuits while others are continually being lost). There is also likely to be some plasticity at the level of the branches of axons and/or dendrites, which provide the scaffolding underlying the formation and loss of synaptic connections.

System-Level Organization: The Macrostructure of Cognition in the Brain

Specialization of Brain Regions

A central and important fact about the organization of cognition in the brain is that individual brain regions are specialized. The cerebral cortex can be partitioned conceptually into primary, secondary, and tertiary cortical zones (Luria, 1966). According to this conception, the primary areas contain neurons whose responses can be largely characterized as reflecting relatively simple, local properties of inputs or outputs within a given modality, such as the presence of an oriented line segment at a particular position on the retina of the eye, the presence of acoustic energy in a particular frequency band, or the presence of a tactile stimulus at a particular point on the skin surface. Corresponding motor areas contain neurons whose responses may correspond to the activation of specific muscles or elementary movement elements. Secondary areas contain neurons whose responses represent higher order stimulus attributes within a given modality, such as conjunctions of features, and the representations in these areas may be relatively invariant over some lower level properties, such as position of the stimulus containing the feature on the sensory surface (Tanaka, 1996). Tertiary areas are responsible for representations that transcend individual modalities, such as representations of the current task context, representation of one's location in extrapersonal space, or representation of semantic content. It should be noted that this picture is only a very crude approximation, and many so-called primary areas appear to participate in the representation of the global

structure of a stimulus or response situation, and many areas that are treated as modality-specific can be modulated by influences from other modalities (see below). It should also be noted that structures outside the neocortex also play very important roles in cognitive functions. Among these are the diffuse neuromodulatory systems that regulate behavioral/cognitive states such as alertness, wakefulness, and mood; and subcortical systems in the thalamus, basal ganglia, and cerebellum that play important roles in sensory motor processes, skills and habit formation, and motor learning. The amygdala plays a central role in emotion and its influence on cognition, and the hippocampus plays a crucial role in the formation of new memories.

Modular versus Interactive Approaches to the Organization of Function

The mechanisms and processes noted above provide only the starting place for the formulation of an understanding of how cognitive processes arise from neural activity. There are two contrasting views: (1) The modular approach, championed by David Marr for vision and Noam Chomsky for language, and systematized as a general approach to understanding brain organization by Fodor (1983), holds that the brain consists of many separate modules that are informationally encapsulated in that their operation is informed only by a very limited range of constraining sources of information. The modular view also holds that the principles of function are specific to each domain, and that distinct and individualized mechanisms are used to subserve each distinct function. For example, the initial assignment of the basic grammatical structure to a sentence is thought to be based only on the syntactic classification of words and their order, and to be governed by the operation of a system of structure-sensitive rules. The module that carries out this assignment is considered to be structured specifically so that it will acquire and implement structure-sensitive rules, and to contrast in the principles that it employs internally with other modules that carry out other tasks, including other aspects of language processing, such as the assignment of meanings to the words in a sentence. In Fodor's view, there are many specialized modules (corresponding approximately to primary and secondary cortical areas and their subcortical inputs and outputs). These are complemented by a general-purpose cognitive system that is completely open-ended in the computations that it can undertake and in the range of informational sources that it can take into consideration.

The alternative, interactive approach, has its seeds in the ideas of Luria (1966), and has been championed by Mesulam (2000) and Rumelhart et al. (1986), and overlaps with the ideas of Damasio (1989). On this view, cognitive outcomes, such as the assignment of an interpretation to a sentence, arise from mutual, bidirectional interactions among neurons in populations representing different types of information. An example of a system addressing the representations and interactions involved in reading individual words aloud is shown in Figure 2. As the Figure suggests, the formation of the pronunciation from a visual orthographic input arises from an interactive process involving orthographic (i.e., letter identity), semantic, phonological, and contextual information. Both the modular and the interactive view are consistent with the idea that neurons in the brain are organized into populations

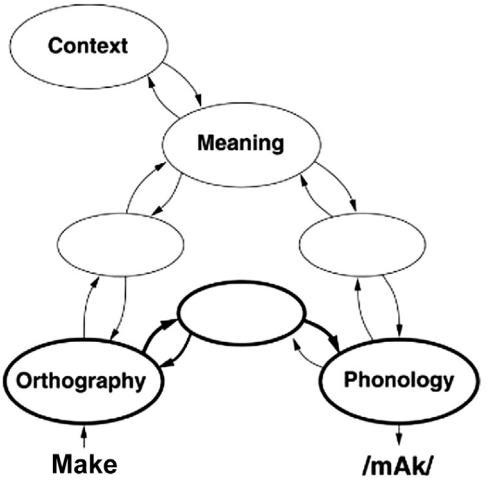

Figure 2 The interactive, distributed framework for modeling individual word reading of Seidenberg and McClelland, 1989. All of the relevant processing pathways are assumed to be bidirectional. Figure 2 comes from Figure 1, p. 526, of Seidenberg, M.S., McClelland, J.L. 1989. A Distributed, Developmental Model of Word Recognition and Naming. Psychological Review 96, 523–568. Copyright 1989 American Psychological Association. Reprinted with permission (automatically granted by APA for a single figure).

specialized for representing different types of information (though the degree of specialization remains hotly debated; see Kanwisher, 2010). Where they differ is in the extent and the role of bidirectional interactions among participating brain areas and in the extent to which the principles governing computation in each area are domain-general or domain-specific.

Evidence of Interactive Processes in the Brain

The debate between modular and interactive approaches is a long-standing one, and can be seen as the modern legacy of a history of diverse views on the localization of functions within the brain (Luria, 1966). While the debate is likely to continue to evolve with additional empirical evidence, it may be worth considering a few elements of evidence that support the idea that processing may be interactive. One relevant anatomical point is the fact that connectivity within and between brain areas is generally reciprocal: when there are connections from region A to region B, there are nearly always return connections (Maunsell and van Essen, 1983).

While there is no consensus on the function of reciprocal connections, there is some evidence that they subserve distributed, interactive computations, at least in particular cases. For example, there is evidence that interactive processes influence the activation of individual neurons in primary visual cortex (Area V1). Traditionally, individual neurons in this area have been seen as encoding the presence of segments of oriented edges at particular positions in a visual display. Recent evidence suggests, however, that primary visual cortex participates in a distributed and interactive process that contributes to the representation of global stimulus properties such as figure-ground organization. The firing of neurons in primary visual

cortex is strongly affected by temporary inactivation of corresponding portions of secondary visual cortex, suggesting that reciprocal interactions between these areas shape neuronal responses in V1 (Hupe et al., 2001). Although the initial response of neurons in V1 is determined by appropriate oriented line segments at a specific location, by about 80 ms their firing is heavily dependent on the global structure of the display (Lee et al., 1998). Furthermore, neurons in V1 respond to illusory contours (contours that are not present in the image but are implied by other contours and perceived by the viewer) that fall in their receptive field. The response occurs at a lag of about 80 ms, suggesting an indirect source, perhaps arising from feedback from higher cortical areas (Lee and Nguyen, 2001). There is also considerable evidence of between-modality interactions. For example, activity in auditory processing areas associated with speech perception is enhanced by visible speech (Callan et al., 2001). There are many corresponding examples of cross-modal influences in single neuron recording studies in animals.

Methods and Approaches in Cognitive Neuroscience

Cognitive neuroscience is a highly interdisciplinary endeavor and draws on a wide range of research methods and approaches, each with its own history and underlying theoretical frame of reference. One important challenge for the field is to find ways of integrating the insights gained from the different methods to allow the field as a whole to converge on a common theoretical framework. Here the predominant research approaches are briefly described, and some of the prospects for integration are considered.

Lesion and Behavior Approaches (Cognitive Neuropsychology, Behavioral Neuroscience, and Transcranial Magnetic Stimulation)

These research approaches have the oldest historical roots, based as they often are in the assessment of the effects of naturally occurring brain damage on cognitive function. A seminal case study was the report by Broca (1861) of a man with a severe disturbance of language arising from a large brain lesion in the posterior portion of the left frontal lobe. Since Broca's day, neurologists and neuropsychologists have investigated the effects of accidental or therapeutic brain lesions in humans, and many key insights have arisen from these studies (*see* Agnosia; Amnesia: General; Aphasia; Dyslexias, Acquired and Agraphia). The subdiscipline of cognitive neuropsychology arose specifically around the study of the effects of brain lesions (*see* Cognitive Neuropsychology, Methods of). A complementary endeavor has grown up around the use of brain lesions in animals carried out with specific experimental intent. This tradition is relevant to human cognitive neuroscience in view of the very close homology between many structures in the human brain and corresponding structures in the primate and rodent brains. This work has obvious advantages in that lesions can be carefully targeted to particular brain areas to test specific hypotheses. Many key insights have emerged from this work, including the discovery of complementary processing streams in the visual system (Ungerleider and Mishkin, 1982; *see* Neural Plasticity). A third technique, which shares many of the same features, is transcranial magnetic stimulation (TMS). TMS utilizes a high magnetic field to generate stimulation in the cortex immediately below the coil. When combined with careful behavioral experiments, TMS can be used to test the role of targeted (human) brain regions. Unlike human neuropsychology, of course, the transiently stimulated brain regions are under experimental control not only in terms of location but also the timing of the interference. Consequently, both the function and the chronometry of brain regions can be probed (Walsh and Cowey, 2000). While all three techniques have provided important and fundamental insights about brain function, the approaches are not without their pitfalls, since lesions or localized stimulation may have unintended and unobserved effects in other brain regions; the refinement and extension of experimental lesion techniques is ongoing (including the combination with functional neuroimaging).

Neuronal Recording Studies

Studies relying on microelectrodes to record from neurons in the brains of behaving animals can allow researchers to study the representations that the brain uses to encode information and the evolution of these representations over time. Several fundamental observations have been made using this technique. As discussed above, these studies indicate, among other things, that the brain relies on distributed representations and that neurons participate dynamically and interactively in the construction of representations of external inputs. As one example, neuronal recording studies have had a profound impact on our understanding of the nature of representations of extrapersonal space. There are neurons in the brain that encode the location of objects in extrapersonal space simultaneously in relation to many different parts of the body, including the limbs and head (Duhamel et al., 1991; Graziano and Gross, 1993) and other neurons that encode the locations of objects in relation to other objects (Olson and Gettner, 1995). Furthermore, recordings from neurons in parietal cortex suggest that when we move our eyes from one location to another, we update our internal representations of the locations of important objects in space, based on where we anticipate that they will be after the upcoming eye movement (Duhamel et al., 1992).

An important recent development is the ability to record from hundreds or thousands of individual neurons at a time. A key finding that has come out of this work is the confirmation in studies with rodents that the simultaneous and successive patterns of activity acquired during behavior may be reactivated in the brain during subsequent sleep (Wilson and McNaughton, 1994). The potential of such methods to shed light on the moment-by-moment relations between activations of different neurons and between distributed brain representations and specific inputs and outputs makes them essential to the future of cognitive neuroscience.

Although primarily utilized in nonhuman situations, some important findings have arisen from the rare opportunity to measure directly from the human cortex. As a part of preoperative neurosurgical investigations, some patients have subdural grid or depth electrodes inserted. In addition to providing key clinical information (e.g., localizing important language and motor functions, or epileptic foci), these studies have generated direct insights about the location of human cognitive functions (including language), and about the timing of the underlying neural activity and interactivity between regions. For example, although often considered to be a high-order component of the visual 'ventral stream,' these methods have shown that the anterior fusiform/inferior temporal gyrus region is important in human language function (the 'basal temporal language area': Lüders et al., 1991) and can probe the presence and timing of cortico-cortical connectivity (Matsumoto et al., 2004).

Functional and Structural Brain Imaging

Cognitive neuroscience arose as a distinct discipline at the same time as the emergence of functional brain imaging methods (PET, fMRI, EEG, and MEG) as major tools for the analysis of human cognition. Indeed, the prospect of visualizing human cognitive activity has been a major catalyst for the field. First used to analyze cognitive functions by the St Louis group (Petersen et al., 1988; *see* Functional Brain Imaging), these methods have come into widespread use. While these methods have low temporal and spatial resolution compared to neuronal recording studies, they still provide our best opportunity to explore the neural mechanisms underlying distinctly human cognitive functions.

The broad adoption of functional neuroimaging techniques since 1988 means that a vast array of methods and findings has emerged from this part of cognitive neuroscience (Bandettini, 2012). The full range of cognitive functions spanning from perception to higher cognitive functions including language, social interaction, and consciousness have been explored using these methods. The first and still very common use of functional neuroimaging is to localize function often with an aim to corroborate findings from other methods, and/or to explore commonalities and differences in human and animal brain organization. As an obvious extension to classical neuropsychology, such approaches have the advantages that, depending on the imaging technique, function can be probed across the entire brain simultaneously and for 'normal' function. Consequently, studies are not confined either to the limited number of recording locations offered with single cell recording or to the neural locations implicated by neurological disease or TMS. In addition to this, brain imaging has begun to reveal a great deal about the plasticity of the brain, since patterns of brain activation can change dramatically with practice (Karni et al., 1998) or after brain damage (Saur et al., 2006). The counterpoint limitation of functional neuroimaging, however, is the method does not identify which region(s) are necessary for that function – something that can only be derived from neuropsychology and TMS (Price and Friston, 2002).

As noted above, at the heart of cognitive neuroscience has been a debate between modular and distributed approaches to understanding human cognition. These same themes are apparent within functional neuroimaging. Thus, the early and still common use of the 'subtraction method' in neuroimaging (where brain activity for a specific cognitive component is identified by contrasting target and control activities that utilize the same assumed cognitive elements minus the key cognitive component) is based, at least implicitly, on a version of a modular view of brain function, which may be incorrect for a variety of reasons (Price and Friston, 2002). Functional neuroimaging, however, has tended to support a more distributed approach given that studies of cognitive function almost never elicit a single region but rather a broadly distributed 'network' of cortical and subcortical areas. Thus, more recently, alternative acquisition and analysis methods have been adopted, which consider cognitive function to be inherently distributed in nature either at whole brain or local levels. At the whole-brain systems level, many studies explore or extract the pattern of correlated activity across a set of regions from either spontaneous ('resting state') or task-based brain activity (referred to as 'functional connectivity') or investigate how the relationship between brain regions changes with respect to task- or stimulus-related factors ('effective connectivity'). At the local level, other approaches have adopted the principle that representations are likely to be distributed and may generate only small variations in local activity (i.e., within $\sim$1 mm^2 *voxels*) in keeping with multielectrode neurophysiological studies (see above). Accordingly, multivoxel pattern analyses investigate the pattern of activity observed across a patch of cortex in order to reveal how different types and levels of information for a set of stimuli are neurally coded (Kriegeskorte, 2011).

In addition to the investigations of neural function, neuroimaging techniques also provide important information about the anatomy of the brain. Indeed, before the advent of neuroimaging, attempts to relate patients' impairments to the pattern of underlying damage (neuropsychology) or investigations of the white-matter connectivity of the human brain were limited to postmortem studies. Modern structural neuroimaging has provided a method for investigating neuroanatomy in ever-increasing detail such that with high-field scanning and analysis methods, it is now becoming possible, for example, to image the layered structure of cortex in the living human brain (*in vivo*). This technological shift provides multiple novel opportunities for cognitive neuroscience including: (1) the ability to relate function and dysfunction (after brain damage or in neurodevelopmental disorders) to brain structure not only *in vivo* for each individual but across a much larger sample of individuals than is practically possible for postmortem studies, licensing correlational as well as intergroup comparisons (e.g., voxel-based lesion-symptom mapping: Bates et al., 2003); (2) the opportunity to map the white-matter connectivity between neural regions including the major long-distance neural pathways or *fasciculi*, and relate these to function and dysfunction; and (3) emerging possibilities to map subtle anatomical features of different brain areas *in vivo*.

Computational and Mathematical Modeling Approaches

While investigations relying on lesion and behavior approaches, neuronal recording studies, and functional brain

imaging have provided and will continue to provide the empirical evidence on which to build our understanding of the basis of cognitive functions in the brain, these approaches, even when used in a convergent way, may still fail to provide a complete understanding of how cognitive functions emerge from underlying neural activity. This may require the use of additional tools provided by mathematical modeling and computer simulation. These approaches allow researchers to formulate possible accounts of specific processes in the form of explicit models that can be analyzed mathematically or simulated using computers to determine whether they can account for all of the relevant neural and behavioral evidence. In short – like other sciences – purely analytical techniques are insufficient to build a true understanding, instead it is important to test and refine theories through synthesis.

Three examples of cases in which computational models have already led to new thinking will be briefly considered. First, a number of computational modeling studies have shown that many aspects of the receptive field properties of neurons and their spatial organization in the brain can arise through the operation of very simple activity-dependent processes shaped by experience and a few rather simple additional constraints (Linsker, 1986; Miller et al., 1989; *see* Neural Development: Mechanisms and Models). Second, models may aid in the understanding of the pattern of deficits seen in patients with brain lesions. Certain patients with an acquired dyslexic syndrome known as deep dyslexia make a striking type of error known as semantic errors; for example, the patient may misread *apricot* as 'peach.' In addition, all such patients also make visual errors, for example, reading *sympathy* and 'symphony.' Early, noncomputational accounts postulated that there must be multiple lesions underlying this disorder. However, computational models of the reading process (Hinton and Shallice 1991; *see* Cognitive Functions (Normal) and Neuropsychological Deficits, Models of) have shown that a single lesion affecting either the visual or the semantic part of an interactive neural network will lead to errors of both types. Thus, the coexistence of these errors may be an intrinsic property of the underlying processing architecture rather than a reflection of multiple distinct lesions. A third area where computational models have shed considerable light is in the interpretation of the receptive field properties of individual neurons (Zipser and Andersen, 1988). While initial interpretations were based on verbally describable features such as oriented bars or edges, such properties are not always apparent, and even when they are, a more detailed characterization may be possible in computational terms (Pouget et al., 1999). A further area of fertile research is in the use of computational models to explain and catalog the ways in which neuronal activation changes dynamically in the course of task performance (Moody et al., 1998).

Like all single methodologies in cognitive neuroscience, an important future step will be the fusion of computational and mathematical models with complementary neuroscience methods. Computational models have the capacity to include and test the implications of new findings at different levels of analysis, ranging from cellular/molecular influences on neuronal assemblies through to the question of how sophisticated high-order human cognition arises from interactions between widely distributed, interconnected cortical and subcortical regions. Perhaps even more excitingly, computational models may offer a formal mechanism to bridge between these different levels of analytical neuroscience (Ueno et al., 2011), ultimately offering a tangible link between brain and mind, which is the central, essential goal of cognitive neuroscience.

Open Issues in Cognitive Neuroscience

Cognitive neuroscience is young, and there is a great deal of work to be done. No aspect of cognition is fully understood, and in general, the more abstract or advanced the cognitive function, the less is known about its neural basis. A few of the most important and interesting issues that remain to be addressed are considered briefly here.

How Does the Brain Learn?

There is a great deal known about the basic mechanisms of synaptic plasticity, but typically these are studied in highly reduced preparations, such as brain slices. The basic processes that are studied in slices surely play a role in the shaping of neural connections in the whole, living brain, but they are also undoubtedly modulated by processes that are usually eliminated in slices. We know that attention and engagement in processing is essential for learning, and there is good reason to believe that learning is gated by various neuromodulatory mechanisms in the brain, but the details of the modulation and gating processes are only beginning to be explored.

What Makes an Experience Conscious?

Although some considerable progress has been made in characterizing the concomitants of consciousness (*see* Consciousness, Neural Basis of), there is no overall understanding of exactly what it is about the activity of the brain that gives it the attribute of consciousness. It appears likely that consciousness will not be localizable; although it may be highly dependent on specific brain structures (e.g., those that regulate sleep vs wakefulness, etc.), it may well depend on the intact functioning of many interacting parts of the brain. Exactly why or how consciousness arises from these interactions is not at all understood.

What Is the Basis for the Unique Cognitive Capacities of the Human Brain, Relative to that of Other, Simpler Organisms?

The issue of what sets humans apart from other organisms remains one of the central unresolved questions. The similarity of the human genome to that of closely related species can be taken in different ways. It can suggest to some that a very small number of specific faculties have been added that differentiate the human from, say, the chimpanzee; or it could suggest that rather than new faculties, the human brain really differs only in the expansion and extension of structures already present to a degree in other organisms. The idea that the highest cognitive functions are emergent functions rather than localizable or

locally encoded in genes remains an attractive, though elusive possibility.

The Future of Cognitive Neuroscience

Nobel laureate Eric Kandel has suggested that cognitive neuroscience will increasingly assume center stage in the neurosciences in the twenty-first century, and it has begun to make dramatic inroads into the field of cognitive psychology, where many leading investigators have redirected their research to exploit ideas and methods from neuroscience. Future research in cognitive neuroscience will address the general issues raised above as well as many other topics. What makes the future of the field so exciting is the prospect of further development of a number of important contributing methodologies. Breakthroughs in functional brain imaging and other related methods are likely to provide far greater spatial and temporal resolution of brain activity. Another, very important area of methodological advance is the ability to create genetically altered brains especially in small mammals and invertebrates, and thereby to explore the consequences of these alterations for function (*see* Genetic Approaches to Memory). These methods have already reached the point where it is possible to allow an organism to develop normally, and then induce a region-specific gene knockout, thereby providing the opportunity to investigate, for example, the effect of the alternation of synaptic plasticity in a specific part of the brain. Breakthroughs should be expected in many other areas of cognitive neuroscience as well, including neuronal recording, functional imaging, and computational modeling approaches. Together, these methods will lead to a deeper understanding of how the highest capabilities of the human mind arise from the underlying physical and chemical processes in the brain.

See also: Brain, Evolution of; Cerebral Cortex; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Cognitive Neuropsychology, Methods of; Computational Neuroscience; Evolutionary Social Psychology; Human Cognition, Evolution of; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures.

Bibliography

Bandettini, P.A., 2012. Twenty years of functional MRI: the science and the stories. NeuroImage 62, 575–588.

Bates, E., Wilson, S.M., Saygin, A.P., Dick, F., Sereno, M.I., Knight, R.T., Dronkers, N.F., 2003. Voxel-based lesion-symptom mapping. Nature Neuroscience 6, 448–450.

Broca, P., 1861. Remarques sur le siege de la faculte de la parole articulee, suives d'une observation d'aphemie (perte de parole). Bulletin de la Societe d'Anatomie 36, 330–357.

Callan, D.E., Callan, A.M., Kroos, C., Vatikiotis-Bateson, E., 2001. Multimodal contribution to speech perception revealed by independent component analysis: a single-sweep EEG case study. Cognitive Brain Research 10 (3), 349–353.

Damasio, A.R., 1989. Time-locked multiregional retroactivation: a system-level proposal for the neural substrates of recall and recognition. Cognition 33, 25–62.

Duhamel, J.-R., Colby, C.L., Goldberg, M.E., 1991. Congruent representations of visual and somatosensory space in single neurons of monkey ventral intraparietal cortex (area VIP). In: Paillard, J. (Ed.), Brain and Space. Oxford University Press, Oxford, UK, pp. 223–226.

Duhamel, J.-R., Colby, C.L., Goldberg, M.E., 1992. The updating of the representation of visual space in parietal cortex by intended eye movements. Science 255, 90–92.

Fodor, J.A., 1983. Modularity of Mind: An Essay on Faculty Psychology. MIT Press, Cambridge, MA.

Georgopoulos, A.P., Schwartz, A.B., Kettner, R.E., 1986. Neuronal population encoding of movement direction. Science 233, 1416–1419.

Graziano, M.S.A., Gross, C.G., 1993. A bimodal map of space: somatosensory receptive fields in the macaque putamen with corresponding visual receptive fields. Experimental Brain Research 97, 96–109.

Hebb, D.O., 1949. The Organization of Behavior. Wiley, New York.

Hinton, G.E., Shallice, T., 1991. Lesioning an attractor network: Investigations of acquired dyslexia. Psychological Review 98 (1), 74–95.

Hubel, D.H., Weisel, T., 1962. Receptive fields, binocular orientation and functional architecture in the cat's visual cortex. Journal of Physiology 166, 106–154.

Hupe, J.M., James, A.C., Girard, P., Lomber, S.G., Payne, B.R., Bullier, J., 2001. Feedback connections act on the early part of the responses in monkey visual cortex. Journal of Neurophysiology 85, 134–145.

Kanwisher, N., 2010. Functional specificity in the human brain: a window into the functional architecture of the mind. Proceedings of the National Academy of Sciences of the United States of America 107, 11163–11170.

Karni, A., Meyer, G., Rey-Hipolito, C., Jezzard, P., Adams, M.M., Turner, R., Ungerleider, L.G., 1998. The acquisition of skilled motor performance: fast and slow experience-driven changes in primary motor cortex. Proceedings of the National Academy of Science of the United States of America 95, 861–868.

Kriegeskorte, N., 2011. Pattern-information analysis: from stimulus decoding to computational-model testing. NeuroImage 56, 411–421.

Lee, T.S., Mumford, D., Romero, R., Lamme, V.A.F., 1998. The role of primary visual cortex in higher level vision. Vision Research 38, 2429–2454.

Lee, T.S., Nguyen, M., 2001. Dynamics of subjective contour formation in early visual cortex. Proceedings of the National Academy of Science of the United States of America 98 (4), 1907–1911.

Linsker, R., 1986. From basic network principles to neural architecture, I: emergence of orientation columns. Proceedings of the National Academy of Sciences of the United States of America 83, 7508–7512.

Lüders, H., Lesser, R.P., Hahn, J., Dinner, D.S., Morris, H.H., Wyllie, E., Godoy, J., 1991. Basal temporal language area. Brain 114, 743–754.

Luria, A.R., 1966. Higher Cortical Functions in Man. Basic Books, New York.

Markus, E.J., Qin, Y., Leonard, B., Skaggs, W.E., McNaughton, B.L., Barnes, C.A., 1995. Interactions between location and task affect the spatial and directional firing of hippocampal neurons. Journal of Neuroscience 15, 7079–7094.

Maunsell, J.H., van Essen, D.C., 1983. The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. The Journal of Neuroscience 3, 2563–2586.

Matsumoto, R., Nair, D.R., LaPresto, E., Najm, I., Bingaman, W., Shibasaki, H., Lüders, H., 2004. Functional connectivity in the human language system: a cortico-cortical evoked potential study. Brain 127, 2316–2330.

Mesulam, M.M., 2000. Principles of Behavioral and Cognitive Neurology. Oxford University Press, New York.

Miller, K.D., Keller, J.B., Stryker, M.P., 1989. Ocular dominance column development: Analysis and simulation. Science 245, 605–615.

Moody, S.L., Wise, S.P., di Pellegrino, G., Zipser, D., 1998. A model that accounts for activity in primate frontal cortex during a delayed matching-to-sample task. Journal of Neuroscience 18 (1), 399–410.

Olson, C.R., Gettner, N., 1995. Object-centered direction selectivity in the supplementary eye field of the macaque monkey. Science 269, 985–988.

Petersen, S.E., Fox, P.T., Posner, M.I., Mintun, M., Raichle, M.E., 1988. Positron emission tomographic studies of the cortical anatomy of single-word processing. Nature 331, 585–589.

Pouget, A., Deneve, S., Sejnowski, T.J., 1999. Frames of reference in hemineglect: a computational approach. Progress in Brain Research 121, 81–97.

Price, C.J., Friston, K.J., 2002. Degeneracy and cognitive anatomy. Trends in Cognitive Sciences 6, 416–421.

Quiroga, R.Q., Reddy, L., Kreiman, G., Koch, C., Fried, I., 2005. Invariant visual representation by single neurons in the human brain. Nature 435, 1102–1107.

Rumelhart, D.E., McClelland, J.L., the PDP Research Group, 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. In: Foundations, vol. 1. MIT Press, Cambridge, MA.

Saur, D., Lange, R., Baumgaertner, A., Schraknepper, V., Willmes, K., Rijntjes, M., Weiller, C., 2006. Dynamics of language reorganization after stroke. Brain 129, 1371–1384.

Seidenberg, M.S., McClelland, J.L., 1989. A distributed, developmental model of word recognition and naming. Psychological Review 96, 523–568.

Tanaka, K., 1996. Inferotemporal cortex and object vision. Annual Review Neuroscience 19, 109–139.

Ueno, T., Saito, S., Rogers, S.S., Lambon Ralph, M.A., 2011. Lichtheim 2: synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. Neuron 72, 385–396.

Ungerleider, L.G., Mishkin, M., 1982. Two cortical visual systems. In: Ingle, D.J., Goodale, M.A., Mansfield, R.J.W. (Eds.), Analysis of Visual Behavior. MIT Press, Cambridge, MA.

Walsh, V., Cowey, A., 2000. Transcranial magnetic stimulation and cognitive neuroscience. Nature Reviews Neuroscience 1, 73–80.

Wilson, M.A., McNaughton, B.L., 1994. Reactivation of hippocampal ensemble memories during sleep. Science 265, 676–679.

Zipser, D., Andersen, R.A., 1988. A back propagation programmed network that simulates response properties of a subset of posterior parietal neurons. Nature 331, 679–684.

Cognitive Psychology: History

Edward E Smith, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 2140–2147, © 2001, Elsevier Ltd.

Abstract

Modern cognitive psychology began in the late 1950s, when American and British psychologists rejected behaviorism and adopted a model of the mind based on the computer. The overthrow of behaviorism came not so much from ideas within psychology as from three research approaches external to the field, including: (1) communications research including information theory; (2) the computer modeling approach, particularly the work of Newell and Simon; and (3) the generative linguistics approach, particularly the work of Chomsky. The merging of these approaches with substantive psychology led to the information-processing approach to cognition, which emphasized the representations and processes needed to give rise to mental activities ranging from pattern recognition, attention, categorization, memory, reasoning, decision making, problem solving, and language.

Since the beginning of experimental psychology in the nineteenth century, there had been interest in the study of higher mental processes. But something discontinuous happened in the late 1950s, something so dramatic that it is now referred to as the 'cognitive revolution,' and the view of mental processes that it spawned is called 'cognitive psychology.' What happened was that American psychologists rejected behaviorism and adopted a model of mind based on the computer. The brief history that follows (adapted in part from Hilgard (1987) and Kessel and Bevan (1985)) chronicles mainstream cognitive psychology from the onset of the cognitive revolution to the beginning of the twenty-first century.

Beginnings

From roughly the 1920s through the 1950s, American psychology was dominated by behaviorism. Behaviorism was concerned primarily with the learning of associations, particularly in nonhuman species, and it constrained theorizing to stimulus-response notions. The overthrow of behaviorism came not so much from ideas within psychology as from three research approaches external to the field.

Communications Research and the Information-Processing Approach

During World War II, new concepts and theories were developed about signal processing and communication, and these ideas had a profound impact on psychologists active during the war years. One important work was Shannon's (1948) paper about *Information Theory*. It proposed that information was communicated by sending a signal through a sequence of stages or transformations. This suggested that human perception and memory might be conceptualized in a similar way: sensory information enters the receptors, then is fed into perceptual analyzers, whose outputs in turn are input to memory systems. This was the start of the 'information-processing' approach – the idea that cognition could be understood as a flow of information within the organism, an idea that continues to dominate cognitive psychology.

Perhaps the first major theoretical effort in information-processing psychology was Donald Broad-bent's *Perception and Communication* (Broadbent, 1958). According to Broadbent's model, information output from the perceptual system encountered a filter, which passed only information to which people were attending. Although this notion of an all-or-none filter would prove too strong (Treisman, 1960), it offered a mechanistic account of selective attention, a concept that had been banished during behaviorism. Information that passed Broadbent's filter then moved on to a 'limited capacity decision channel,' a system that has some of the properties of short-term memory, and from there on to long-term memory. This last part of Broadbent's model – the transfer of information from short- to long-term memory – became the salient point of the dual-memory models developed in the 1970s.

Another aspect of Information theory that attracted psychologist's interest was a quantitative measure of information in terms of 'bits' (roughly, the logarithm to the base two of the number of possible alternatives). In a still widely cited paper, George Miller (1956) showed that the limits of short-term memory had little to do with bits. But along the way, Miller's and others' interest in the technical aspects of information theory and related work had fostered mathematical psychology, a subfield that was being fueled by other sources as well (e.g., Estes and Burke, 1953; Luce, 1959; Garner, 1962). Over the years, mathematical psychology has frequently joined forces with the information-processing approach to provide precise claims about memory, attention, and related processes.

The Computer Modeling Approach

Technical developments during World War II also led to the development of digital computers. Questions soon arose about the comparability of computer and human intelligence (Turing, 1950). By 1957, Alan Newell, J.C. Shaw, and Herb Simon had designed a computer program that could solve difficult logic problems, a domain previously thought to be the unique province of humans. Newell and Simon soon followed with programs that displayed general problem-solving skills

much like those of humans, and argued that these programs offered detailed models of human problem solving (a classic summary is contained in Newell and Simon (1972)). This work would also help establish the field of artificial intelligence (AI).

Early on, cross-talk developed between the computer modeling and information-processing approaches, which crystallized in the 1960 book *Plans and the Structure of Behavior* (Miller et al., 1960). The book showed that information-processing psychology could use the theoretical language of computer modeling even if it did not actually lead to computer programs. With the 'bit' having failed as a psychological unit, information processing badly needed a rigorous but rich means to represent psychological information (without such representations, what exactly was being processed in the information-processing approach?). Computer modeling supplied powerful ideas about representations (as data structures), as well as about processes that operate on these structures. The resultant idea of human information processing as sequences of computational processes operating on mental representations remains the cornerstone of modern cognitive psychology (see, e.g., Fodor, 1975).

The Generative Linguistics Approach

A third external influence that lead to the rise of modern cognitive psychology was the development of generative grammar in linguistics by Noam Chomsky. Two of Chomsky's publications in the late 1950s had a profound effect on the nascent cognitive psychology. The first was his 1957 book *Syntactic Structures* (Chomsky, 1957). It focused on the mental structures needed to represent the kind of linguistic knowledge that any competent speaker of a language must have. Chomsky argued that associations per se, and even phrase structure grammars, could not fully represent our knowledge of syntax (how words are organized into phrases and sentences). What had to be added was a component capable of transforming one syntactic structure into another. These proposals about transformational grammar would change the intellectual landscape of linguistics, and usher in a new psycholinguistics.

Chomsky's second publication (1959) was a review of *Verbal Behavior*, a book about language learning by the then most respected behaviorist alive, B.F. Skinner (Skinner, 1957). Chomsky's review is arguably one of the most significant documents in the history of cognitive psychology. It aimed not merely to devastate Skinner's proposals about language, but to undermine behaviorism as a serious scientific approach to psychology. To some extent, it succeeded on both counts.

An Approach Intrinsic to Psychology

At least one source of modern cognitive psychology came from within the field. This approach had its roots in Gestalt psychology, and maintained its focus on the higher mental processes. A signal event in this tradition was the 1956 book *A Study of Thinking*, by Bruner, Goodnow, and Austin (Bruner et al., 1956). The work investigated how people learn new concepts and categories, and it emphasized strategies of learning rather than just associative relations. The proposals fit perfectly with the information-processing approach – indeed, they were information-processing proposals – and offered still another reason to break from behaviorism.

By the early 1960s all was in place. Behaviorism was on the wane in academic departments all over America (it had never really taken strong root in Europe). Psychologists interested in the information-processing approach were moving into academia, and Harvard University went so far as to establish a Center for Cognitive Studies directed by Jerome Bruner and George Miller. The new view in psychology was information processing. It likened mind to a computer, and emphasized the representations and processes needed to give rise to activities ranging from pattern recognition, attention, categorization, memory, reasoning, decision making, problem solving, and language.

The Growth of Cognitive Psychology

The 1960s brought progress in many of the above-mentioned topic areas, some of which are highlighted below.

Pattern Recognition

One of the first areas to benefit from the cognitive revolution was pattern recognition, the study of how people perceive and recognize objects. The cognitive approach provided a general two-stage view of object recognition: (1) describing the input object in terms of relatively primitive features (e.g., 'it has two diagonal lines and one horizontal line connecting them'); and (2) matching this object description to stored object descriptions in visual memory, and selecting the best match as the identity of the input object ('this description best matches the letter A'). This two-stage view was not entirely new to psychology, but expressing it in information-processing terms allowed one to connect empirical studies of object perception to computer models of the process. The psychologist Ulrich Neisser (1964) used a computer model of pattern recognition (Selfridge, 1959) to direct his empirical studies and provided dramatic evidence that an object could be matched to multiple visual memories in parallel.

Other research indicated that the processing underlying object perception could persist after the stimulus was removed. For this to happen, there had to be a visual memory of the stimulus. Evidence for such an 'iconic' memory was supplied by Sperling in classic experiments in 1960 (Sperling, 1960). Evidence for a comparable brief auditory memory was soon provided as well (e.g., Crowder and Morton, 1969). Much of the work on object recognition and sensory memories was integrated in Neisser's influential 1967 book *Cognitive Psychology* (Neisser, 1967). The book served as the first comprehensive statement of existing research in cognitive psychology, and it gave the new field its name.

Memory Models and Findings

Broadbent's model of attention and memory stimulated the formulation of rival models in the 1960s. These models assumed that short-term memory (STM) and long-term

memory (LTM) were qualitatively different structures, with information first entering STM and then being transferred to LTM (e.g., Waugh and Norman, 1965). The Atkinson and Shiffrin (1968) model proved particularly influential. With its emphases on information flowing between memory stores, control processes regulating that flow, and mathematical descriptions of these processes, the model was a quintessential example of the information-processing approach. The model was related to various findings about memory. For example, when people have to recall a long list of words they do best on the first words presented, a 'primacy' effect, and on the last few words presented, a 'recency' effect. Various experiments indicated that the recency effect reflected retrieval from STM, whereas the primacy effect reflected enhanced retrieval from LTM due to greater rehearsal for the first items presented (e.g., Murdock, 1962; Glanzer and Cunitz, 1966). At the time these results were seen as very supportive of dual-memory models (although alternative interpretations would soon be proposed – particularly by Craik and Lockhart, 1972).

Progress during this period also involved empirically determining the characteristics of encoding, storage, and retrieval processes in STM and LTM. The results indicated that verbal material was encoded and stored in a phonologic code for STM, but a more meaning-based code for LTM (Conrad, 1964; Kintsch and Buschke, 1969). Other classic studies demonstrated that forgetting in STM reflected a loss of information from storage due to either decay or interference (e.g., Wickelgren, 1965), whereas some apparent losses of information in LTM often reflected a temporary failure in retrieval (Tulving and Pearlstone, 1966). To a large extent, these findings have held up during over 30 years of research, although many of the findings would now be seen as more limited in scope (e.g., the findings about STM are now seen as reflecting only one component of working memory, e.g., Baddeley (1986), and the findings about LTM are seen as characterizing only one of several LTM systems, e.g., Schacter (1987)).

One of the most important innovations of 1960s research was the emphasis on reaction time as a dependent measure. Because the focus was on the flow of information, it made sense to characterize various processes by their temporal extent. In a seminal paper in 1966, Saul Sternberg reported (Sternberg, 1966) that the time to retrieve an item from STM increased linearly with the number of items in store, suggesting that retrieval was based on a rapid scan of STM. Sternberg (1969) gave latency measures another boost when he developed the 'additive factors' method, which, given assumptions about serial processing, allowed one to attribute changes in reaction times to specific processing stages involved in the task (e.g., a decrease in the perceptibility of information affected the encoding of information into STM but not its storage and retrieval). These advances in 'mental chronometry' quickly spread to areas other than memory (e.g., Fitts and Posner, 1967; see also Schneider and Shiffrin, 1977).

The New Psycholinguistics

Beginning in the early 1960s there was great interest in determining the psychological reality of Chomsky's theories of language (these theories had been formulated with ideal listeners and speakers in mind). Some of these linguistically inspired experiments presented sentences in perception and memory paradigms, and showed that sentences deemed more syntactically complex by transformational grammar were harder to perceive or store (Miller, 1962). Subtler experiments tried to show that syntactic units, like phrases, functioned as units in perception, STM, and LTM (Fodor et al. (1974) is the classic review). While many of these results are no longer seen as critical, this research effort created a new subfield of cognitive psychology, a psycholinguistics that demanded sophistication in modern linguistic theory.

Not all psycholinguistic studies focused on syntax. Some dealt with semantics, particularly the representation of the meanings of words, and a few of these studies made use of the newly developed mental chronometry. One experiment that proved seminal was reported by Collins and Quillian (1969). Articipants were asked simple questions about the meaning of a word, such as 'Is a robin a bird,' and 'Is a robin an animal?'; the greater the categorical difference between the two terms in a question, the longer it took to answer. These results were taken to support a model of semantic knowledge in which meanings were organized in a hierarchical network, e.g., the concept 'robin' is directly connected to the concept 'bird,' which in turn is directly connected to the concept 'animal,' and information can flow from 'robin' to 'animal' only by going through 'bird' (see the top of **Figure 1**). Models like this were to proliferate in the next stage of cognitive psychology.

The Rise of Cognitive Science

Memory and Language

Early in the 1970s the fields of memory and language began to intersect. In 1973 John Anderson and Gordon Bower published *Human Associative Memory* (Anderson and Bower, 1973), which presented a model of memory for linguistic materials. The model combined information processing with recent developments in linguistics and AI, thereby linking the three major research directions that led to the cognitive revolution. The model used networks similar to that considered above to represent semantic knowledge, and used memory-search processes to interrogate these networks (see the middle of **Figure 1**). The Anderson and Bower book was quickly followed by other large-scale theoretical efforts that combined information processing, modern linguistics, and computer models. These efforts included Kintsch (1974), which focused on memory for paragraphs rather than sentences, and Norman et al. (1975), Anderson (1976), and Schank and Abelson (1977), which took a more computer-science perspective and focused on stories and other large linguistic units.

As psychologists became aware of related developments in linguistics and AI, so researchers in the latter disciplines become aware of pertinent work in psychology. Thus evolved the interdisciplinary movement called 'cognitive science.' In addition to psychology, AI, and linguistics, the fields of cultural anthropology and philosophy of mind also became involved. The movement eventuated in numerous interdisciplinary

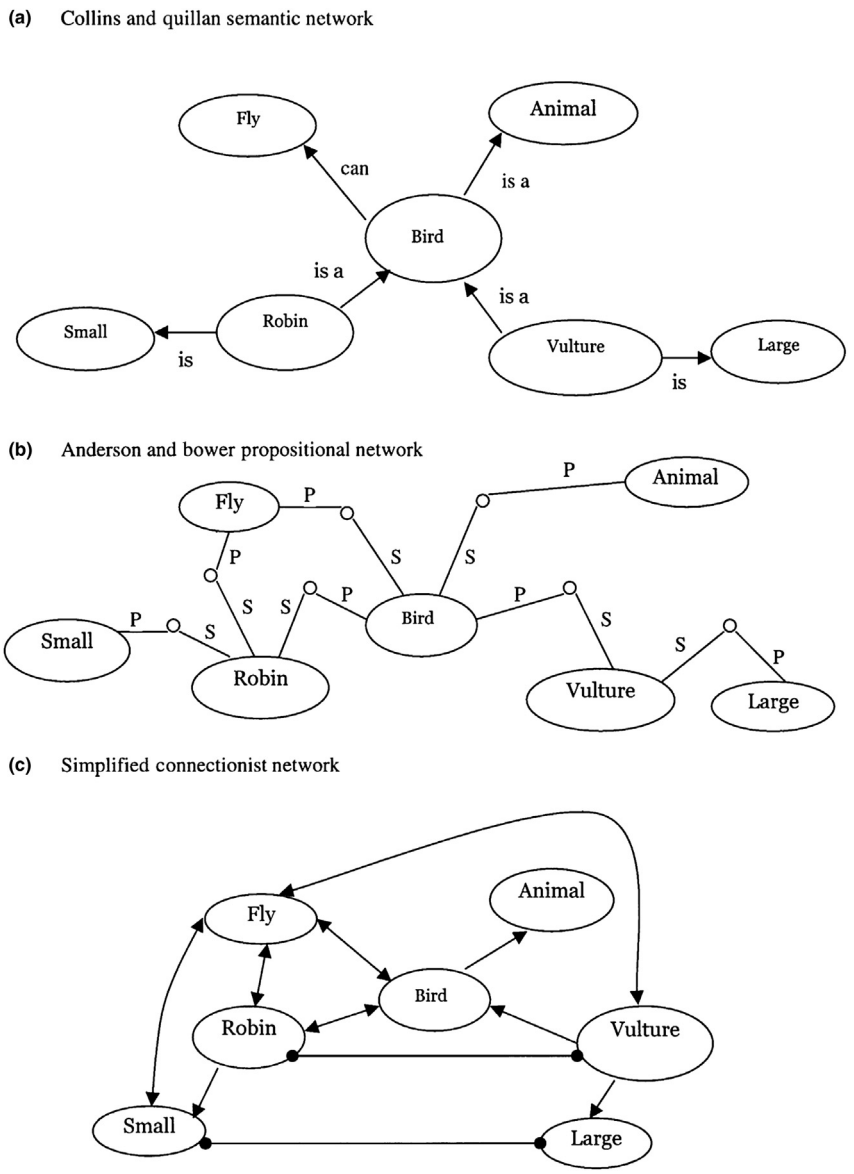

Figure 1 (a) Part of a Collins, A.M., Quillian, R.M., 1969. Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior 8, 240–247 semantic network. Circles designate concepts and lines (arrows) between circles designate relations between concepts. There are two kinds of relations: subset-superset ('Robin is a bird') and property (e.g., 'Robins can fly'). The network is strictly hierarchical, as properties are stored only at the highest level at which they apply. (b) Part of an Anderson, J.R., Bower, G.H., 1973. Human Associative Memory. Winston, Washington, DC propositional network. Circles represent concepts and lines between them labeled relations. All propositions have a subject-predicate structure, and the network is not strictly hierarchical. (c) Part of a simplified connectionist network. Circles represent concepts, or parts of concepts, lines with arrowheads depict excitatory connections, and lines with filled circles designate inhibitory connections; typically numbers are put on the lines indicate the strength of the connections. The network is not strictly hierarchical, and is more interconnected than the preceding networks.

collaborations (e.g., Rumelhart et al., 1986), as well as in individual psychologists becoming more interdisciplinary.

Representational Issues

In the 1970s and early 1980s, cognitive science was much concerned with issues about mental representations. Whereas the memory-for-language models described earlier had assumed representations that were language-like, or propositional, other researchers argued that representations could also be imaginal, like a visual image. Shepard and Cooper (1972) provided evidence that people could mentally rotate their representations of objects, and Kosslyn (1980) surveyed numerous phenomena that further implicated visual imagery. In keeping with the interdisciplinariness of cognitive science, AI researchers and philosophers entered the debate about propositional versus imaginal representations (e.g., Block, 1981; Pylyshyn, 1981). In addition to questions about the modality of representations, there were concerns about the structure of representations. While it had long been assumed that

propositional representations of objects were like definitions, researchers now proposed the representations were *prototypes* of the objects, fitting some examples better than others (Tversky, 1977; Mervis and Rosch, 1981; Smith and Medin, 1981). Again the issues sparked interest in disciplines other than psychology (e.g., Lakoff, 1987).

The cognitive science movement affected most areas of cognitive psychology, ranging from object recognition (Marr, 1982) to reasoning (e.g., Johnson-Laird, 1983) to expertise in problem solving (e.g., Chase and Simon, 1973). The movement continues to be influential and increasingly focuses on computational models of cognition. What has changed since its inception in the 1970s is the kind of computational model in favor.

Newer Directions: Connectionism and Cognitive Neuroscience

Connectionist Modeling

The computer models that dominated cognitive psychology from its inception used complex symbols as representations, and processed these representations in a rule-based fashion (e.g., in a model of object recognition, the representation for a frog might consist of a conjunction of complex properties, and the rule for recognition might look something like 'If it's green, small, and croaks, it's a frog'). Starting in the early 1980s, an alternative kind of cognitive model started to attract interest, namely 'connectionist' (or 'parallel distributed processing') models. These proposals have the form of neural networks, consisting of nodes (representations) that are densely interconnected, with the connections varying in strength (see the bottom of **Figure 1**).

In 1981 Hinton and Anderson published a book surveying then existent connectionist models (Hinton and Anderson, 1981), and in the same year McClelland and Rumelhart (1981) presented a connectionist model of word recognition that explained a wide variety of experimental results. The floodgates had been opened, and connectionist models of perception, memory, and language proliferated, to the point where they now dominate computational approaches to cognition. Why the great appeal? One frequently cited reason is *neurological plausibility*: the models are clearly closer to brain function than are traditional rule-based models. A second reason is that connectionist models permit *parallel constraint satisfaction*: different sources of activation can converge simultaneously on the same representations or response. A third reason is that connectionist models manifest *graceful degradation*: when the model is damaged, performance degrades slowly, much as is found in human neurological disorders (Rumelhart et al., 1986).

From a historical viewpoint, there is an ironic aspect about the ascendancy of connectionist models. Such models return to the pure associationism that characterized behaviorism. While connectionist models hardly fit with all behaviorist dictums – their representations are not restricted to stimuli and responses, and they routinely assume massively parallel processing – still their reliance on associations runs counter to Chomsky's arguments that sheer associationism cannot explain language (Chomsky, 1957, 1959). This issue has formed part of the basis for critiques of connectionist models (e.g., Fodor and Pylyshyn, 1988). Newell (1990), one of the founders of traditional computational models, suggested a plausible resolution: lower-level cognitive processes like object recognition may be well modeled by connectionist models, but higher-level cognitive processes like reasoning and language may require traditional symbolic modeling.

Cognitive Neuroscience

The other major new direction in cognitive psychology is the growing interest in the neural bases of cognition, a movement referred to as 'cognitive neuroscience.' There had been little interest in biological work in the research that brought about the cognitive revolution. That early work was as much concerned with fighting behaviorism as it was with advancing cognitive psychology, and consequently much of the research focused on higher-level processes and was completely removed from anything going on in the neurobiology of its day. Subsequent generations of cognitive psychologists solidified their commitments to a purely cognitive level of analyses, by arguing that the distinction between cognitive and neural levels of analyses was analogous to that between computer software and hardware, and that cognitive psychology (and cognitive science) was concerned primarily with the software. Since the early 1990s, views about the importance of neural analyses have changed dramatically. There is a growing consensus that the standard information-processing analyses of cognition can be substantially enlightened by knowing how cognition is implemented in the brain.

While many factors may have been responsible for the change in view, three will be mentioned here. First, the rise of connectionist models that were loosely inspired by brain function had the side effect of increasing interest in what was known in detail about brain function. Indeed, these two recent movements have become increasingly intertwined as connectionist models have increasingly incorporated findings from cognitive neuroscience. Second, in the 1970s and 1980s there were breakthroughs in systems-level neuro-science that had implications for mainstream cognitive psychology. As one example, research with neurological patients as well as with nonhuman species established that structures in the medial temporal lobe were essential for the formation of memories that could serve as the basis for recall ('explicit memory'), but not for other kinds of 'implicit memories' (e.g., Schacter, 1987; Squire, 1992). As another example, research with nonhuman primates showed that two distinct systems were involved in the early stages of visual perception: a 'where' system that is responsible for spatial localization, and a 'what' system that is responsible for recognition of the object (e.g., Ungerleider and Mishkin, 1982). Both of these discoveries had major implications for the cognitive level of analyses, e.g., computational models of object recognition or memory would do well to divide the labor into separate subsystems.

The third factor responsible for the rise of cognitive neuroscience is methodological: the development of neuroimaging techniques that produce maps of neural activity while the brain is performing some cognitive task. One major technique is positron emission tomography (PET). In 1988,

Michael Posner, Marcus Raichle, and colleagues used PET in a groundbreaking experiment to localize neurally different subprocesses of reading – PET images obtained while participants looked at words showed activated regions in the posterior temporal cortex; images obtained while participants read the words revealed additional regions activated in the motor cortex; and images obtained while participants generated constrained associations to the words revealed still additional regions in frontal areas (Posner et al., 1988). PET analyses of object recognition, memory, and language soon followed. A more recent technique is functional magnetic resonance imaging. It is now being used to study virtually every domain of human cognition.

Conclusion

This article has given short shrift to important contributions that tend to fall off the mainstream of cognitive psychology. One such case is the work done by Dan Kahneman and Amos Tversky (e.g., Kahneman and Tversky, 1973; Tversky and Kahneman, 1983) on the use of heuristics in decision making, which can result in deviations from rational behavior. Another example is the cognitively inspired study of memory and language deficits in neurological patients (Shallice (1988) provides a review). There are other cases like these which deserve a prominent place in a fuller history of cognitive psychology.

Bibliography

Anderson, J.R., Bower, G.H., 1973. Human Associative Memory. Winston, Washington, DC.

Anderson, J.R., 1976. Language, Memory, and Thought. Erlbaum, Hillsdale, NJ.

Atkinson, R.C., Shiffrin, R.M., 1968. Human memory: a proposed system and its control processes. In: Spence, K., Spence, J. (Eds.), The Psychology of Learning and Motivation, vol. 2. Academic Press, San Diego, CA.

Baddeley, A.D., 1986. Working Memory. Oxford University Press, Oxford, UK.

Block, N.J., 1981. Imagery. MIT Press, Cambridge, MA.

Broadbent, D.E., 1958. Perception and Communication. Pergamon, New York.

Bruner, J.S., Goodnow, J.J., Austin, G.A., 1956. A Study of Thinking. Wiley, New York.

Chase, W.G., Simon, H.A., 1973. Perception in chess. Cognitive Psychology 4, 55–81.

Chomsky, N., 1957. Syntactic Structures. Mouton, The Hague, The Netherlands.

Chomsky, N., 1959. Review of B F Skinner's verbal behavior. Language 35, 26–58.

Collins, A.M., Quillian, R.M., 1969. Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior 8, 240–247.

Conrad, R., 1964. Acoustic confusions in immediate memory. British Journal of Psychology 55, 75–84.

Craik, F.I., Lockhart, R.S., 1972. Levels of processing: a framework for memory research. Journal of Verbal Learning and Verbal Behavior 11, 671–684.

Crowder, R.G., Morton, J., 1969. Precategorical acoustic storage (PAS). Perception and Psychophysics. 5, 365–373.

Estes, W.K., Burke, C.J., 1953. A theory of stimulus variability in learning. Psychological Review 60, 276–286.

Fitts, P.M., Posner, M.I., 1967. Human Performance. Brooks/Cole, Belmont, CA.

Fodor, J.A., Pylyshyn, Z.W., 1988. Connectionism and cognitive architecture: a critical analysis. Cognition 28, 3–71.

Fodor, J.A., Bever, T.G., Garrett, M.F., 1974. The Psychology of Language: An Introduction to Psycholinguistics and Generative Grammar. McGraw-Hill, New York.

Fodor, J.A., 1975. The Language of Thought. Crowell, New York.

Garner, W.R., 1962. Uncertainty and Structure as Psychological Concepts. Wiley, New York.

Glanzer, M., Cunitz, A.R., 1966. Two storage mechanisms in free recall. Journal of Verbal Learning and Verbal Behavior 5, 351–360.

Hilgard, E.R., 1987. Psychology in America: A Historical Survey. Harcourt, Brace Jovanovich, San Diego, CA.

Hinton, G.E., Anderson, J.A., 1981. Parallel Models of Associative Memory. Erlbaum, Hillsdale, NJ.

Johnson-Laird, P.N., 1983. Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness. Cambridge University Press, Cambridge, MA.

Kahneman, D., Tversky, A., 1973. Psychology of prediction. Psychological Review 80, 237–251.

Kessel, F.S., Bevan, W., 1985. Notes toward a history of cognitive psychology. In: Buxton, C.E. (Ed.), Points of View in the Modern History of Psychology. Academic Press, New York.

Kintsch, W., Buschke, H., 1969. Homophones and synonyms in short-term memory. Journal of Experimental Psychology 80, 403–407.

Kintsch, W., 1974. The Representation of Meaning in Memory. Erlbaum, Hillsdale, NJ.

Kosslyn, S.M., 1980. Image and Mind. Harvard University Press, Cambridge, MA.

Lakoff, G., 1987. Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. University of Chicago Press, Chicago.

Luce, R.D., 1959. Individual Choice Behavior: A Theoretical Analysis. Wiley, New York.

Marr, D., 1982. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Freeman, San Francisco.

McClelland, J.L., Rumelhart, D.E., 1981. An interactive activation model of context effects in letter perception: I. An account of basic findings. Psychological Review 88, 375–407.

Miller, G.A., et al., 1960. Plans and the Structure of Behavior. Holt, New York.

Miller, G.A., 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review 63, 81–97.

Miller, G.A., 1962. Some psychological studies of grammar. American Psychologist 17, 748–762.

Murdock Jr., B.B., 1962. The serial position effect in the recall. Journal of Experimental Psychology 64, 482–488.

Neisser, U., 1964. Visual search. Scientific American 210, 94–102.

Neisser, U., 1967. Cognitive Psychology. Appleton-Century-Crofts, New York.

Neisser, U., 1968. The processes of vision. Scientific American 219, 204–214.

Newell, A., Simon, H.A., 1972. Human Problem Solving. Prentice-Hall, Englewood Cliffs, NJ.

Newell, A., Shaw, J.C., Simon, H.A., 1958. Elements of a theory of human problem solving. Psychological Review 65, 151–166.

Newell, A., 1990. Unified Theories of Cognition. Harvard University Press, Cambridge, MA.

Norman, D.A., Rumelhart, D.E., LNR Research Group, 1975. Explorations in Cognition. Freeman, San Francisco.

Posner, M.I., Petersen, S.E., Fox, P.T., Raichle, M.E., 1988. Localization of cognitive operations in the human brain. Science 240, 1627–1631.

Pylyshyn, Z.W., 1981. The imagery debate: analogue media versus tacit knowledge. Psychological Review 88, 16–45.

Mervis, C.B., Rosch, E., 1981. Categorization of natural objects. Annual Review of Psychology 32, 89–115.

Rumelhart, D.E., McClelland, J.L., PDP Research Group, 1986. Parallel Distributed Processing: Explorations in the Micro-structure of Cognition. MIT Press, Cambridge, MA.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536.

Schacter, D.L., 1987. Implicit memory: history and current status. Journal of Experimental Psychology: Learning, Memory, and Cognition 13, 501–518.

Schank, R.C., Abelson, R.P., 1977. Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures. Erlbaum, Hillsdale, NJ.

Schneider, W., Shiffrin, R.M., 1977. Controlled and automatic human information processing: I. Detection, search, and attention. Psychological Review 84, 1–66.

Selfridge, O.G., Neisser, U., 1960. Pattern recognition by machine. Scientific American 203, 60–68.

Selfridge, O.G., 1959. Pandemonium: A Paradigm for Learning. The Mechanization of Thought Processes. Her Majesty's Stationery Office, London.

Shallice, T., 1988. From Neuropsychology to Mental Structure. Cambridge University Press, Cambridge, UK.

Shannon, C.E., Weaver, W., 1949. The Mathematical Theory of Communication. University of Illinois Press, Urbana, IL.

Shannon, C.E., 1948. A mathematical theory of communication. Bells Systems Technical Journal 27, 379–423, 623-656.

Shepard, R.N., Cooper, L.A., 1972. Mental Images and Their Transformations. MIT Press, Cambridge, MA.

Skinner, B.F., 1957. Verbal Behavior. Appleton-Century-Crofts, New York.

Smith, E.E., Medin, D.L., 1981. Categories and Concepts. Harvard University Press, Cambridge, MA.

Sperling, G., 1960. The information available in brief visual presentations. Psychological Monographs 77 (3, no. 478).
Squire, L.R., 1992. Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. Psychological Review 99 (2), 195–231.
Sternberg, S., 1966. High-speed scanning in human memory. Science 153, 652–654.
Sternberg, S., 1969. The discovery of processing stages: extensions of Donders' method. Acta Psychologica 30, 276–315.
Treisman, A.M., 1960. Contextual cues in selective listening. Quarterly Journal of Experimental Psychology 12, 242–248.
Tulving, E., Pearlstone, Z., 1966. Availability versus accessibility of information in memory for words. Journal of Verbal Learning and Verbal Behavior 5, 381–391.
Turing, A.M., 1950. Computing machinery and intelligence. Mind 59, 433–460.
Tversky, A., Kahneman, D., 1983. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. Psychological Review 90, 293–315.
Tversky, A., 1977. Features of similarity. Psychological Review 84, 327–352.
Ungerleider, L.G., Mishkin, M., 1982. Two cortical visual systems. In: Engle, D.J., Goodale, M.A., Mansfield, R.J. (Eds.), Analysis of Visual Behavior. MIT Press, Cambridge, MA, pp. 549–586.
Waugh, N.C., Norman, D.A., 1965. Primary memory. Psychological Review 72, 89–104.
Wickelgren, W.A., 1965. Acoustic similarity and intrusion errors in short-term memory. Journal of Experimental Psychology 70, 102–108.

Cognitive Styles and Learning Styles

Stephen G Rayner, Newman University, Birmingham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The psychological construct of style in personality, cognition, and learning is explained in this article. The development of a styles theory is interpreted as the evolution of a generic concept of individuality and its status as an individual difference in cognition and learning. The emergence of popular applications of learning styles as well as a wave of critical revisionism and new directions in researching style differences is related to this development. An application of style differences for lifelong learning in both education and workplace are then considered, including implications for further study of differential psychology, pedagogy, training, and the nature of an individual's personal approach to learning (style).

Style Differences and Individuality: Cognition, Behavior, and Affect

The commonsense idea of style is one of the distinctive and personal signatures or hallmarks woven into an individual's performance (Rayner, 2000, 2001). Less clear, however, is use of the term 'style' when applied in personal or differential psychology, involving aspects of affect, behavior, cognition, intelligence, personality, self-perception, and social identity (Rayner and Peterson, 2009; Peterson et al., 2009). The implications of style differences for understanding how affect, behavior, and cognition in the person combine to form individuality is reflected in a traditional contribution to psychometric assessment and factor analytic research across three key areas of human performance: first, profiling individual differences in workplace and educational settings (Messick, 1976; Jonassen and Grabowski, 1993); second, exploiting the presence and effect of style features as variables in learning and teaching (Kolb, 1976; Entwistle, 1981; Schmeck, 1988); and third, reasserting the conceptual basis of the style construct in personal and social behavior (Witkin et al., 1962; Sternberg, 1996; Rayner and Riding, 1997; Peterson et al., 2009).

An aim in the research over time has arguably been to describe a holistic but valid and reliably consistent theory of style, distinct from and separate to other individual differences (constructs) such as personality, intelligence, identity, and gender (Miller, 1987; Curry, 1987; Rayner and Riding, 1997). In more recent work (Rayner and Cools, 2011; Zhang et al., 2012), an account is made reexamining the basis of styles as traits, states, identities, and intellectual processes. Renewed interest in style research has also coincided with calls for developing alternative ways of researching style combining new forms of research methodology and epistemology (Rayner and Peterson, 2009; Cools and Rayner, 2011; Rayner, 2011).

The Style Construct in the Psychology of Education

The exact structure and nature of the style construct remains contested (Messick, 1976, 1994, 1996; Ritter, 2007). Similarly, the effect size associated with various measurements of a style factor in educational research (variously described as cognitive, thinking, learning, or intellectual styles) remains disappointingly small when applied to contexts of learning and work-related behavior (Kavale and Forness, 1987; Coffield et al., 2004; Pashler et al., 2008; Peterson et al., 2009). This is an important issue to which we will again return later in this discussion. The utility and integrity of style as a construct, however, is clearly reliant upon a well-established theory, which is both coherent and consistent in terms of conceptual rigor, distinctive validity, independence as a construct, and epistemic robustness.

The following recategorization of cognitive styles begins by distinguishing research in the mid part of the twentieth century, as the production of a styles construct associated with: (1) personality-based styles reflecting type and trait as key indicators of an individual's innate behavior across different contexts of performance; (2) more stable cognition-centered structures and/or dimensions of a person's cognitive style associated with traitlike differences in an individual's tendency to think, feel, or act; (3) less stable cognitive processes linked to the ways in which cognitive controls, strategies, and processes combine in an individual's approach to demand or task in a range of contexts; and (4) preferences for learning and thinking affected by modality, cognitive processes, and learning experiences, while engaged in completing an assigned activity. The first is related to constructs associated with 'personality-based styles,' the second to 'cognition-centered styles,' the third learning strategies, learning patterns, and an orientation to learning, and the fourth, generic 'learning-based styles.'

The fourth part of this style recategorization which involves the identification and development of 'learning styles' reflects a refocused interest upon the question of 'person-environment fit,' and so is associated with a widening range of executive functions in the psychology of the individual when engaged with a task-related performance. This includes intellectual styles related to the following: multiple intelligences or abilities, problem solving, creativity, intuition, decision-making, and a wider range of style-based learning behaviors integrating learning styles with learning strategies as part of an effort to create a framework for applying 'holistic' style profiles (for example, combining personal learning style, learning preferences, experiential learning, learning motivation, and forms of accelerated, extended, or enhanced learning).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92008-7

Personality Styles – Types of Individuality

An early influence in the development of style as type may be linked to the work of Jung's typology of personality constructs and 'psychoanalytic ego psychology' (Jung, 1923). The most influential style model emerging from this paradigm was the Myers–Briggs Type Indicator (Myers, 1978; Myers-Briggs, 1980). This test is generally recognized as a measure of personal style but conversely is described as a personality test (Isaksen et al., 2003; Hough and Ogilvie, 2005). The Myers–Briggs Type Indicator is purported to have become one of the most widely used measures of psychological types in business and public sector organizations. Few studies, however, have been published examining its efficacy (Kelley, 2005).

A separate development to both the theorizing and measuring aspects of personality as individual *preferences* was constructed by Holland (Holland, 1973, 1985; Gottfredson and Holland, 1996), focusing upon individual career choices or planning. Holland's theory is based upon the idea of defining personality types in terms of occupational preferences: these include Realistic, Investigative, Artistic, Social, Enterprising, and Conventional (Holland, 1973). Much of this 'type' theory reflects a similar set of ideas about preferences and trait underpinning the application of a range of style theories including the fundamental idea of the 'matching hypothesis,' which supposes that a deliberate style-led matching of different personality types with task, mode of work, grouping, and context can enhance performance.

According to Sternberg (1997), intellectual styles function at the interface between ability and personality. The psychological structure of style posited in this way describes a relatively stable but complex interactive system that embodies psychological controls, processes, and states but also reflects an executive function regulating personal, physiological, and sociological aspects of the individual's mind at work. For some psychologists, it is self-evident that style and personality represent very similar explanations of individuality and simply are manifestations of one and the same construct (see Messick, 1994; Chamorro-Premuzic and Furnham, 2006; Furnham et al., 2008). It is argued here, however, that the structure and function of the style construct is closely implicated with rather than an aspect of personality. It contributes to a holistic structure (or model) shaping as well as sustaining and reflecting an individual's psychology, including structural forms of their personality, cognition, and behavior (Riding and Rayner, 1998).

Cognitive Styles – Types of Process in Human Cognition

The following section provides a summary of the work related to styles defined as internal control processes and strategies in human cognition. A useful account of the historical development of this work can be found in Kozhevnikov (2007). This is further complemented by analytic reviews offered by Grigerenko and Sternberg (1995), Rayner and Riding (1997), and Zhang and Sternberg (2005). The best introduction to the original styles experiments and source for the construct can be found in Messick (1976).

Perceptual Differentiation

Witkin and coworkers produced seminal work over a period of more than 50 years – originating in an interest with the 'regularities' of information processing, derived from the German gestalt school of perceptual psychology, and which led to an early development of the 'style construct' of 'Field Dependence-Independence' (FDI). The practical issues related to flying aircraft through dense cloud formation were the original spur to investigating what and how individual intradifferences were found to influence crucial variations in pilots maintaining course and control in their aircraft. Individuals, when reorienting an object relative to the vertical, were found to rely upon the surrounding 'field' or 'context' to a greater or lesser extent. This was subsequently found to correlate with competence in disembedding shapes from a perceptual field, and experimental participants were found to either rely heavily on the field for orientation or shape discrimination (field-dependent) or rely little or not at all (field-independent).

The assessment tools developed as part of this research included an original rod and frame tilting chair experiment and then the 'pencil and paper' embedded figures test. While a great deal of critique was subsequently leveled at the FDI theory for confounding abilities and style, Witkin and Ash (1948) extrapolated a comprehensive theory of psychological differentiation to explain how it affected a wide range of human performance across a breadth of social and personal contexts.

Cognitive Controls and Cognitive Processes

Psychologists interested in cognitive control processes and structure produced a second alternative approach to theorizing intradifferences in personality and style. Early experimental work at the Messinger Clinic in the USA developed notions of ego adaptation to the environment in the individual (Gardner, 1953; Gardner et al., 1959). This led to further experimental research hypothesizing style as a key factor in facilitating stable responses (styles) in the individual to the situation, producing patterns of response and function in thinking and behavior (Messick, 1976).

The program completed by these coworkers over more than 20 years produced a series of 'styles' (to function rather like a set of cogs and balances in the cognitive machinery of the mind), usually comprising a bipolar structure and implicated in a wide range of adaptive responses to differing demands in the experimental context. These included, for example, stimuli differentiation ('levelers' and 'sharpeners') identified by Klein (1958), which subsequently developed to include differences in other forms of cognitive scanning, categorization, conceptual articulation, and degree of cognitive control (Gardner et al., 1959; Messick and Fritzky, 1963). A third focus provided by Kagan and coworkers investigating types and levels of response to cognitive uncertainty typically recorded during problem-solving task completion. This latter work identified two types of typical response along a continuum of conceptual tempo, from the impulsive to reflective (Kagan, 1966). It led in turn to the claim for a unique style dimension identified using the Matching Familiar Figures Test (Kagan, 1965).

Mental Imagery, Memory, and Thinking

Similar research around the same time involved exploring individual differences in mental representation and visual information processing: this included looking at visual and/or verbal ways of representing information in thought, as well as visual or imaginal processes used in dealing with new information. Research led by Pask (1972, 1976) on style and information processing introduced the idea of serial or holistic processing. This research identified how the individual would split and separate (serialist type) or group and synthesize (holist type) new information when dealing with a problem-solving task (Pask and Scott, 1972).

At the same time, other work investigating mental imagery revealed how some people have a predominantly verbal way of representing information in thought, while others are more 'visual' or 'imaginal.' Paivio (1971) further developed this theory with a dual coding measurement of mental imagery. Further research completed by Richardson (1977) using the Verbalizer–Visualizer Questionnaire revealed habitual methods of visual thinking. Most recent work investigating the way in which these style dimensions may function has been completed by Kozhevnikov et al. (2005), and secondly, Miller et al. (2012), providing new neurological evidence of style-related mental activity, elicited from the use of functional magnetic resonance imaging (fMRI) and data analysis employing *t*-statistic mapping of brain function.

Riding and Taylor (1976) initially identified, as fundamental to the style construct, a verbal-imagery dimension of cognitive style. This led to the construction of the 'cognitive styles assessment' (CSA), used to support further research in developing an integrated theory of cognitive style (Riding, 1991). The model of style presented subsequently by Riding and Rayner (1998) described an orthogonal arrangement of two fundamental but independent dimensions of cognitive style: the verbal-visualizer style continuum and the wholist–analytic style continuum. These dimensions are assessed by the CSA, a computerized test locating a respondent on each of the wholist–analytic and verbal-imagery dimensions of a cognitive style (Riding, 1991).

Research into Executive Functioning and Differences in Thinking

A widening field of research that was interested in how executive functioning and individual differences in thinking and action influence key aspects of human performance emerged in the latter part of the twentieth century. The work involved many researchers independently constructing new models and measures, thereby adding to a broad and disparate theoretical base, and fueling in turn an increasing number of loosely linked conceptions of style in thinking, learning, and performance.

These various style labels reflecting aspects of cognitive functioning included a model of phenomenological style (Gregorc, 1982); the Adaptor–Innovator cognitive style of decision-making (Kirton, 1994); an intuitive–analytic cognitive style of problem solving (Hayes and Allinson, 1998); cognitive styles analysis focusing upon information processing/mental representation (see Riding, 1991; Riding and Rayner, 1998); and a cognitive style of creativity (Kaufmann, 1989). Another more elaborate model of mental self-government presented by Sternberg (1996, 1997), represented a theory of style derived from notions of a holistic regulation in the mind, integrating constructs of intelligence, abilities, and self-government. According to this theory, individual differences in styles or mental processes can be understood in terms of function, form, level, scope, and leaning, which impact upon thinking and learning.

The final word on the importance of this phase of research development belongs to Kozhevnikov (2007). She explains that its significance principally lies with an attempt to both expand and clarify the mechanisms of cognitive styles in the context of an information processing approach. She argues that there is sufficient evidence to suggest a theoretical basis for the hierarchical classification of cognitive styles according to the level of information processing in which the individual is engaged (from simple perceptual decisions to complex problem solving).

Learning Styles: Personal Differences in Approach and Performance

The development of learning styles research was preceded by work on the assessment of the learner, learning, and teaching. This work, epitomized by the activity of the American National Association of Secondary School Principals (NASSP) task force looking at learning styles (Keefe, 1985), triggered a wave of new models of style, with a great many populist versions, and more recently accessible from the Internet. The result was a proliferation in the range of user-sourced tools of assessment, hybrid forms of style theory, and many commercial packages advocating all in one approaches to personalizing educational intervention. This development was further reinforced by popular interpretations of advances in neuroscience, producing different versions of learning styles applied to teaching and learning in the classroom (Geake, 2008).

In spite of this popularizing trend, several developing theories explaining approaches to individual differences in learning reflected cautious and well-established applied research. This section includes a summary of some key theories of learning styles organized into the following four groups: models focusing on experiential learning; models based on orientation to study; models based upon instructional preferences; and models based upon learning skills associated with neuroscience. The latter research in particular reflected attempts to explore respectively and in combination interaction between inter- and intra-differences in learning and the learner when engaged in learning.

Models Focusing on Experiential Learning

Models derived from a theory of experiential learning, of which most influential is the work of Kolb (1976), described learning styles as the individual's preferred method for assimilating information, in an active learning cycle. Kolb constructed a two-dimensional model comprising perception (concrete/abstract thinking) and processing (active/reflective information processing), treating these as fundamental aspects of an experiential learning cycle. The assessment tool used for this learning styles model is the Learning Styles Inventory (LSI)

(Kolb, 1999). Recent research further developing this model in studies of undergraduate MBA courses in the USA is reported by Sharma and Kolb (2011), a new form of related assessment reported by Boyatzis and Mainemelis (2011) and related pedagogic development (Portugal and the UK) reported by Almeida et al. (2011).

Models Based on Orientation to Study

These derived from earlier theories of information processing and learning processes; the most influential has been that of Entwistle (1981), in developing the early work of Pask (1972) to produce a measure called the Approaches to Study Inventory. This paved the way for a good deal of subsequent research exploring how learners construct an approach to learning and how this interacts with motivation (Marton and Saljo, 1976; Vermunt, 1996; Entwistle and Peterson, 2004). The 'approaches to learning' model is widely referred to in University education (Biggs, 2001; Vermunt, 2011), and typically describes ways in which students engage with deep or surface learning, thus reflecting the adoption of one or more orientations to study (meaning, reproducing, achieving, and holistic learning). These choices of approach reflect a series of orientations that are thought to be relatively stable differences in study behavior but remain susceptible to change as patterns of learning.

Models Focusing upon Instructional Preference

These set out to present a comprehensive profile of learning differences in the typical response of an individual to instructional contexts. Learning style (Keefe, 1989) was typically perceived as a multimodal construct and understood to describe a range of emotional and intellectual functioning relating to the learning activity. Examples of this approach included the 'Cognitive Profile,' (Letteri, 1980) and the Learning Styles Profile (Keefe, 1985, 1989). It is, however, more widely associated with the program of teacher education promoted by Rita Dunn – the LSI – developed by Dunn et al. (1989).

The learning style elements identified when using the Dunn and Dunn LSI are self-reported responses eliciting information for modal and learning preferences located in the individual student. These are: *environmental stimulus* (light, temperature); *emotional stimulus* (persistence, motivation); *sociological stimulus* (peers, adults); *physical stimulus* (perceptual strengths, time of day – morning vs. after-noon); and *psychological stimulus* (global-analytic, or impulsive reflective). A comprehensive range of learning and teaching materials has been developed by Dunn over a 30-year period to support a learning styles-based differentiation of curricula and pedagogy across the full range of educational phases. It is summarized in Dunn and Griggs (2003), and further theorized in Dunn and Honigsfeld (2011).

Models Based upon Learning Skills and Neuroscience

A fourth group of learning style labels is focused upon an individual's developing cognitive ability and repertoire of cognitive skills or capacity to learn, together with related behavioral characteristics. These were understood to comprise an individual's learning profile. More recent versions in what has amounted to a populist approach to learning styles include the application of a widely established 'visual, auditory, and kinesthetic (VAK)' sensory-based approach to learning and teaching, combining multiple intelligences theory (Gardner, 1983), and neuro-linguistic theory (O'Connor and McDermott, 1996). Other similar 'practical applications' of neuroscientific research and teacher development also popular during the last decade include 'accelerated learning' (Rose and Nicholl, 1997), 'cognitive enhancement' (Adey et al., 1999), and 'personalized learning' utilizing 'learning power' (Ritchie and Deakin Crick, 2005).

Revisionist Theories and Problematizing the Style Construct

An original interest in researching styles from a practitioner's perspective reflects a concern for enhancing learning and teaching that in essence seeks to help 'teach the hard to reach and reach the hard to teach.' For this to happen, styles theory must offer and secure clearly stated relevance, utility, and worthwhile effects for any intervention in the workplace or classroom. The nature of such take up, however, may arguably need to be much more than only the 'Aptitude-by-Treatment Interactions' (ATI) based experimental research advocated by Pashler et al. (2008), and identified as the evidentiary gold standard required to confirm the status of style research.

Rayner (2001), moreover, has argued that the main purpose and value for importing style theory into the design of pedagogy and curricula is the proper use of profiling and assessment as a tool for formative learning. Rosenfeld and Rosenfeld (2011), reflecting a similar 'practitioner' perspective, also argue a need for paradigm shift from second to third person in research design, and call for research directly involving students as individuals working with their own learning styles. They very deliberately challenge the 'ATI template' and traditional experimental design. Such an approach reflects a concern for more applied, humanistic understanding of style differences, as well as providing evidence of impact on 'value for instruction' and 'pedagogic development.'

There are a few but prominent and far reaching critiques of style published in the past 50 years reflecting a more general dissatisfaction with what is described as an unexpectedly gargantuan, complex, and disparate field of research. These criticisms include, for example, the following: no consensual or coherent field theory and consistent psychometric failings across the field in respect to measures of validity, reliability, and effect sizes produced in empirical research (Coffield et al., 2004; Pashler et al., 2008); an incestuous research tradition practiced in a field almost entirely engaged with self-affirming replication (Coffield et al., 2004); a flawed theory of the psychology of learning (Nixon et al., 2007; Geake, 2008); misplaced approaches to pedagogy, curriculum, and scholarship (Sharp et al., 2008); a shortfall in clearly stated or reliable and coherent accounts of value, relevance, and utility of learning styles in the applied context (Reynolds, 1997; Coffield et al., 2004; Peterson et al., 2009); resultant misplaced and restrictive typing, labeling, and harmful individualized practices when applying styles theory to education, instruction, pedagogy, or the curriculum (Coffield et al., 2004; Ritter, 2007; Sharp et al., 2008); reference to alternative more compelling theories and

explanations of the learning process in social psychology and sociology, which in itself is claimed to offer a better prospect of return for resource in the educational setting (Coffield et al., 2004; Nixon et al., 2007; Ritter, 2007).

The issues of these critiques of style research persist to date and invalidate by association a good deal of rigorous and robust work. Nonetheless, continuing criticism of styles research reflecting a mix of ideological conflicts and perspectival differences has led at times to a rejection of any need for styles research. These disputes, however, are often reflected in writers from opposing research epistemologies articulating differing conceptualizations of efficacy and equity for the educational setting. It is salutary and interesting to note Messick's comment that such intensity of conflict was indeed part of the academic debate over 50 years ago. He commented that critique was often "*… excessively polarized in either supporting or undercutting styles as meaningful constructs. This polarization seems to reflect different stances not just with reference to scientific evidence but also with respect to ideology*" (Messick, 1994: 121). There has not been much change over time with the substance of style critique.

Implications of Styles for Lifelong Learning

The extent to which an awareness of cognitive style or learning styles and the self as a learner is currently considered and managed within the learning context raises key questions for the design of training, instruction, and pedagogy. These suggest in turn a need to consider the following key aspects of education construed as contributing to the idea of lifelong learning.

- *Intellectual styles and metacognition* – involving research looking at how intellectual styles (cognition and learning) interact with an awareness of self as a learner and thinker. An obvious implication of this approach to valuing style as a concept lies with its effect upon influencing motivation, behavior, and attribution governing responses to personal achievement. The recent work of Sadler-Smith (2009, 2012) provides an interesting start to a conceptualization linking intuition styles, motivation, and experience with duplex theory and intellectual styles with metacognition.
- *Style and self-regulation in learning* – involving work linking theories of self-regulation with awareness and perceptions of self as a learner. A great deal of integrating work completed by researchers in the area of self-attribution and/or regulation (Zimmerman, 2008) offer a useful model as well as related work for further developing pedagogic approaches to working with student motivation and approaches to learning (see Rayner, 2011; Vermunt, 2011).
- *Assessment-led learning and curriculum differentiation* – meeting the need to further demonstrate relevance and positive effects of style awareness for instructional design in work around the original interest in the *matching hypothesis* (aligning or contrasting style with task or social context to enhance flexibility in the learner and learning outcome). Indeed many writers seem to regard this aspect of 'style' as the only serious implication of style in education (for example, Pashler et al., 2008, devote most of their criticism for the *meshing hypothesis*). Recent work in several areas of style research has reflected confirmation of flexible approaches to learning styles in application (Dunn and Honigsfeld, 2011; Sharma and Kolb, 2011). Similarly, work looking at curriculum differentiation has produced new integrative approaches drawing upon both a cognition and learning centered theory of styles (Rayner, 2000; Evans and Waring, 2011; Zhang et al., 2012).
- *Personalizing education and learning technologies* – a great deal of research is concerned with developing an understanding and thereby facilitating the design of personalized learning. Mourlas et al. (2009) provide an overview of this area exploring 'pedagogic' aspects of the interaction between the individual and web-based or digital infrastructure and/or architecture, with a view to enhancing ways in which an alignment between cognitive style and demands may enhance learning. This not only embraces remote or mobile-digital learning, but learning and teaching applications of social media, intelligent design of technologies, user interface with digital technology and more widely personalizing computer-mediated learning and teaching.
- *Organizational learning, training & business management* – the intuitive appeal of style differences in both cognition and learning continues to generate research and development in the workplace. The key themes emerging in this area are a consideration of the role of style differences in organizational behavior and human resource development including: selection and person-environment fit, vocational choice and career success; diversity, group processes, and conflict management; intuition and emotion in the workplace; training and development; personal styles' profiling; and career management in global organizations (Armstrong and Cools, 2009; Armstrong et al., 2012; Cools, 2012).
- *Pedagogic differences and professional development* – personal and social differences identified in style research impact directly upon the training and the development of professional expertise in professional practitioners (e.g., teachers, trainers, and educators). Much of the research into this area reflects adherence to one or more models of learning styles (for example, Dunn and Griggs, 2003; Sharma and Kolb, 2011). An interesting development in further developing a holistic theory of style, however, is reflected in the earlier work of Sadler-Smith (1996), Peterson et al. (2010) as well as Evans and Waring (2008, 2011), focusing upon teacher training and the effects of raising self-awareness of cognitive style in the teacher's developing pedagogic practice.

The significance of 'style' as reflected in this range of research for the life cycle is clearly about much more than 'style' simply used as a term that has linguistic or conceptual versatility in describing personal performance. While this common usage of the term clearly denotes a popular as well as intuitive appeal and so clearly reflects a universal value, the import of this discussion is that style as an individual difference marks something much more specific in a range of learning associated with the idea of the person engaged in lifelong learning.

Conclusion: A Matter of Style

Why do style differences matter? Knowing more about how *another* or *I* as an individual(s) tend to think, feel and act in

a learning context provides rich possibilities for new ways of setting up, facilitating and leading to any learning activity. Profiling differences and potentials in human performance affords greater awareness for self-directed learning and participation or efficacy in a project group. It is equally relevant to informing differentiated design in programs of study, pedagogic practices, and formative curricula. A style-based contribution to education is also one way in which consideration may be given to further customizing digital technology in better serving personal and social learning across contexts and community.

These implications for further study in differential or personal psychology (individual differences), pedagogy, training and the better management of an individual's personal approach to learning (style) should finally require a closer focus upon types of epistemology and methodology in research design. Rayner (2011) and Rayner et al. (2012), for instance, have argued the need to deliberately attempt paradigm shift to supplement the traditional styles research agenda, which has been largely located in a positivistic paradigm, and more usually employed as an experimental or survey based design associated with the science of cognitive psychology. Whilst this experimental work remains central to style research (Kozhevnikov, 2007), there is a conjoint need for more pluralistic approaches in the design and use of psychological assessment or theorizing of style in cognition, learning, and management.

Sternberg et al. (2012: 416) on the other hand, in a retrospective comment on the recent past in style research makes an argument for much greater emphasis in renewing a factor analytic mode in style research. He acknowledges an urgent need for 'sorting out the theories of style,' and developing 'better forms of style assessment,' but finally stresses the need to reinforce the psychological status of style research utilizing 'random-assignment studies with control groups.' He comments that *The research that has been conducted has been of variable quality and styles would get more research if they were more rigorously researched*. What is very apparent in this view is a need, as suggested earlier, for new ways of constructing the methodology of assessment or testing to better treat complex models of style differences, as have, for example, more recently been occurring in the field of personality (see Hopwood and Donnellan, 2010).

Finally, and as described by Rayner (2013), concern for validity, efficacy, and utility as key markers in a theoretical revisionism within future research and development for the field of 'style differences' rests upon attempting a paradigmatic step change and establishing new ways of modeling and measuring the style construct. Such an approach requires a reexamining of epistemic issues underpinning both the style construct and, as crucially, careful thought about research methodology and project design. It also more fundamentally means responding with greater clarity and certainty to a question about the epistemological status of a 'style' construct asked by Schmidt (2012).

Schmidt points out that the style construct to date is on the one hand conceptualized as a typology in the sense of Hempel's 'aspects of scientific explanation' and more recently tested by procedures such as Latent Class or Latent Profile Analysis (Hempel, 1962, 1965), and, on the other hand, the more traditional approach to style is one in which a classical psychometric method, using exploratory or confirmatory factor analysis for testing reliability and validity, is applied and then classical multivariate models used to test the specific contribution of style. Schmidt questions, therefore, if style research might usefully benefit from adopting an approach combining both perspectives and thereby following a three-step methodology in mixture modeling (see, Jung and Wickrama, 2007; Vermunt, 2010), and which is implemented in programs like Latent Gold or MPLUS (Asparouhov and Bengt, 2013).

This pertinent suggestion offered by Schmidt merits serious consideration as part of what Rayner (2013) argues is a need for problematizing the style construct as a basis for moving forward in the field, thereby leaving a fourth period of research and development (see, Kozhevnikov, 2007), to enter a fifth wave of style research (Rayner, 2013). In many ways this arguably reflects a need to move 'back to the future' and robustly reconsider the question of construct validity at the heart of Messick's work on psychological testing (Schiffman and Messick, 1963; Messick, 1988), and more widely any factorial interaction between personality, style, and abilities (Messick, 1994, 1996).

In conclusion, using style as a psychological construct can aid organizing and developing teams of practitioners in a wide range of organizational activity. But we do need to know much more and as such clearly reestablish or confirm the validity of key style differences, as well as related models or measures, and further affirm a 'grand theory' (general understanding) of the meaning or implications of style. The key theme for utilizing the style construct is captured perfectly in the popular song, *It is not what you do, it is the way that you do it*. The point is not that content is unimportant, but that being aware and knowing more about 'how I tend to do it' can immeasurably help in 'my doing it better.' A greater understanding and improvement of human performance, as psychological process, social behavior, and educational achievement, is the touchstone for future work in applied research aimed at developing style theory, research, and practice.

See also: Academic Self-Concept and Achievement; Computer-Assisted Instruction; Education for the Gifted and Talented; Instructional Design; Instructional Psychology; Learning Theories and Educational Paradigms; Learning To learn; Personality Development and Temperament; Self-Concepts: Educational Aspects; Self-Efficacy: Education Aspects; Self-Regulated Learning: Theories, Measures, and Outcomes; Tangible User Interfaces in Learning and Education.

Bibliography

Adey, P., Fairbrother, R., Wiliam, D., 1999. A Review of Research on Learning Strategies and Learning Styles. King's College, London.

Almeida, P., Pedrosa de Jesus, H., Watts, C., 2011. Kolb's learning styles and approaches to learning through the use of students' critical questions. In: Rayner, S., Cools, E. (Eds.), Style Differences in Cognition, Learning and Management. Routledge, New York, pp. 115–128.

Armstrong, S.J., Cools, E., 2009. Cognitive styles and their relevance for business and management: a review of development over the past two decades. In:

Zhang, L.F., Sternberg, R.J. (Eds.), Perspectives on the Nature of Intellectual Styles. Springer, Heidelberg, pp. 253–290.
Armstrong, S.J., van der Heijden, B.I.J.M., Sadler-Smith, E., 2012. Intellectual styles, management of careers, and improved work performance. In: Zhang, L.F., Sternberg, R.J., Rayner, S.G. (Eds.), Handbook of Intellectual Styles. Springer Pubs, NY, pp. 273–294.
Asparouhov, T., Bengt, M., 2013. Auxiliary variables in mixture modeling: 3-step approaches. Using Mplus. Mplus Web Notes 15 (7), 1–48.
Biggs, J., 2001. Enhancing learning: a matter of style or approach? In: Sternberg, R.J., Zhang, L.F. (Eds.), Perspectives on Thinking, Learning and Cognitive Styles. Springer Pubs, NY, pp. 73–102.
Boyatzis, R.E., Mainemelis, C., 2011. Assessing the distribution of learning and adaptive styles in an MBA program. In: Rayner, S., Cools, E. (Eds.), Style Differences in Cognition, Learning and Management. Routledge, New York, pp. 99–114.
Chamorro-Premuzic, T., Furnham, A., 2006. Intellectual competence and the intelligent personality: a third way in differential psychology. Review of General Psychology 10 (3), 251–267.
Coffield, F.C., Moseley, D.V.M., Hall, E., Ecclestone, K., 2004. Learning Styles and Pedagogy in Post-16 Learning: Findings of a Systematic and Critical Review of Learning Styles Models. Learning and Skills Research Centre, London.
Cools, E., Rayner, S., 2011. Researching style: more of the same or moving forward. In: Rayner, S., Cools, E. (Eds.), Style Differences in Cognition, Learning and Management. Routledge, New York, pp. 295–306.
Cools, E., 2012. Understanding styles in organizational behaviors: a summary of insights and implications. In: Zhang, L.F., Sternberg, R.J., Rayner, S.G. (Eds.), Handbook of Intellectual Styles. Springer Pubs, NY, pp. 329–352.
Curry, L., 1987. Integrating Concepts of Cognitive or Learning Style: A Review with Attention to Psychometric Standards. Canadian College of Health Service Executives, Ottawa, ON, Canada.
Dunn, R., Griggs, S., 2003. Synthesis of the Dunn and Dunn Learning Styles Model research: Who, What, When, Where and So What – The Dunn and Dunn Learning Styles Model and Its Theoretical Cornerstone. St John's University, New York.
Dunn, R., Honigsfeld, A., 2011. Learning-styles-based differentiating instruction for at-risk students: rationale, research documentation, assessment tools, and alternative approaches. In: Rayner, S., Cools, E. (Eds.), Style Differences in Cognition, Learning and Management. Routledge, New York, pp. 263–277.
Dunn, R., Dunn, K., Price, G.E., 1989. Learning Styles Inventory. Price Systems, Lawrence, KS, USA.
Entwistle, N., Peterson, E.R., 2004. Learning styles and approaches to studying. In: Spielberger, C. (Ed.), Encyclopedia of Applied Psychology. Academic Press, New York, pp. 537–542.
Entwistle, N.J., 1981. Styles of Teaching and Learning: An Integrated Outline of Educational Psychology for Students, Teachers and Lecturers. Wiley, Chichester, NY.
Evans, C., Waring, M., 2008. Trainee teachers' cognitive styles and notions of differentiation. Education and Training 50, 140–154.
Evans, C., Waring, M., 2011. Enhancing feedback practice: a personal learning styles pedagogy approach. In: Rayner & E Cools, S. (Ed.), Style Differences in Cognition, Learning, and Management: Theory, Research and Practice. Routledge, New York, pp. 188–203.
Furnham, A., Christopher, A., Garwood, J., Martin, N.G., 2008. Ability, demography, learning style, and personality trait correlates of student preference for assessment method. Educational Psychology 28 (1), 15–27.
Gardner, R.W., Holzman, P.S., Klein, G.S., Linton, H.B., Spence, D.P., 1959. Cognitive control: a study of individual consistencies in cognitive behaviour. Psychological Issues 1 (4).
Gardner, R.W., 1953. Cognitive styles in categorizing behaviour. Journal of Personality 22, 214–333.
Gardner, H., 1983. Frames of Mind: The Theory of Multiple Intelligences. Basic Books, New York.
Geake, J., 2008. Neuromythologies in education. Educational Research 50 (2), 123–133.
Gottfredson, G.D., Holland, J.L., 1996. Dictionary of Holland Occupational Codes, third ed. Psychological Assessment Resources, Odessa, FL.
Gregorc, A.R., 1982. Style Delineator. Gabriel Systems, Maynard, MA.
Grigerenko, E.L., Sternberg, R.J., 1995. Thinking styles. In: Saklofske, D.H., Zeidner, M. (Eds.), International Handbook of Personality and Intelligence. Plenum Press, NY, USA, pp. 205–229.
Hayes, J., Allinson, C.W., 1998. Cognitive style and the theory and practice of individual and collective learning in organizations. Human Relations 51, 847–871.
Hempel, C., 1962. Explanation in science and history. In: Colodny, R.C. (Ed.), Frontiers of Science and Philosophy. The University of Pittsburgh Press, Pittsburgh, pp. 9–19.
Hempel, C., 1965. Aspects of Scientific Explanation. The Free Press, New York.
Holland, J.L., 1973. Making Vocational Choices: A Theory of Careers. Prentice-Hall, Englewood Cliffs, NJ.
Holland, J.L., 1985. Manual for the Vocational Preference Inventory. Psychological Assessment Resources, Inc, Odessa FL.
Hopwood, C.J., Donnellan, M.B., 2010. How should the internal structure of personality inventories be evaluated? Personality and Social Psychology Review 14 (3), 332–346.
Hough, J.R., Ogilvie, D.T., 2005. An empirical test of cognitive style and strategic decision outcomes. Journal of Management Studies 42 (2), 417–448.
Isaksen, S.G., Lauer, K.J., Wilson, G.V., 2003. An examination of the relationship between personality type and cognitive style. Creativity Research Journal 15 (4), 343–354.
Jonassen, D.H., Grabowski, B.L., 1993. Handbook of Individual Difference, Learning and Instruction. Lawrence Erlbaum Associates, Hillsdale, NJ.
Jung, T., Wickrama, K.A.S., 2007. An introduction to latent class growth analysis and growth mixture modeling. Social and Personality Psychology Compass 2 (1), 302–317.
Jung, C., 1923. Psychological Types. Harcourt Brace, NY.
Kagan, J., 1965. Individual difference in the resolution of response uncertainty. Journal of Personality and Social Psychology 2, 154–160.
Kagan, J., 1966. Reflection-impulsivity: the generality and dynamics of conceptual tempo. Journal of Abnormal Psychology 71, 17–24.
Kaufmann, G., 1989. The Assimilator-explorer Inventory. University of Bergen, Bergen, Norway.
Kavale, K.A., Forness, S.R., 1987. Substance over style: assessing the efficacy of modality testing and teaching. Exceptional Children 54 (3), 228–239.
Keefe, J.W., 1985. Assessment of learning style variables: the NASSP task force model. Theory into Practice 24, 138–144.
Keefe, J.W., 1989. Profiling and Utilising Learning Style. National Association of Secondary School Principals, Reston, VA.
Kelley, K.N., 2005. Using the Myers-Briggs type indicator. PsycCRITIQUES 50 (33). http://dx.doi.org/10.1037/05159912.
Kirton, J.W. (Ed.), 1994. Adaptors and Innovators, second ed. Routledge, London.
Klein, G.S., 1958. Cognitive control and motivation. In: Lindzey, G. (Ed.), Assessment of Human Motives. Holt, Rinehart & Winston, New York, pp. 87–118.
Kolb, D.A., 1976. Learning Style Inventory: Technical Manual. Prentice-Hall, Englewood Cliffs, USA.
Kolb, D.A., 1999. The Kolb Learning Style Inventory Version 3. Hay Group, Boston, MA.
Kozhevnikov, M., Kosslyn, S.M., Shephard, J., 2005. Spatial versus object visualizers: a new characterization of visual cognitive style. Memory & Cognition 33, 710–726.
Kozhevnikov, M., 2007. Cognitive styles in the context of modern psychology: toward an integrated framework of cognitive style. Psychological Bulletin 133, 464–481.
Letteri, C.A., 1980. Cognitive profile – basic determinants of academic achievement. Journal of Educational Research 73, 195–199.
Marton, F., Saljo, R., 1976. On qualitative differences in learning outcome and process. British Journal of Educational Psychology 46, 4–11.
Messick, S., Fritzky, F.J., 1963. Dimension of analytic attitude in cognition and personality. Journal of Personality 31, 346–370.
Messick, S., 1976. Individuality in Learning: Implications of Cognitive Styles and Creativity for Human Development. Jossey Bass, San Francisco, CA.
Messick, S., 1988. The once and future issues of validity: assessing the meaning and consequences of measurement. In: Wainer, H., Braun, I.H. (Eds.), Test Validity. Lawrence Erlbaum, Hillsdale, NJ, pp. 33–45.
Messick, S., 1994. The matter of style: manifestations of personality in cognition, learning and teaching. Educational Psychologist 29, 121–136.
Messick, S., 1996. Bridging cognition and personality in education: the role of style in performance and development. European Journal of Personality 10, 353–376.
Miller, M.B., Donovan, C.L., Bennett, C.M., Aminoff, E.M., Mayer, R.A., 2012. Individual differences in cognitive style and strategy predict similarities in the patterns of brain activity between individuals. NeuroImage 59, 83–93.
Miller, A., 1987. Cognitive styles: an integrated model. Educational Psychology 7 (4), 251–268.
Mourlas, C., Tsianos, N., Germanakos, P. (Eds.), 2009. Cognitive and Emotional Processes in Web-Based Education: Integrating Human Factors and Personalization. IGI Global Pub, Hershey, PA.
Myers, I., 1978. Myers-Briggs Type Indicator. Consulting Psychologists Press, Palo Alto, CA.

Myers-Briggs, I., 1980. Gifts Differing: Understanding Personality Type. Davies-Black Publishing, NY.
Nixon, L., Gregson, M., Spedding, T., 2007. Pedagogy and the intuitive appeal of learning styles in post compulsory education in England. Journal of Vocational Education and Training 59 (1), 39–50.
O'Connor, J., McDermott, I., 1996. Principles of NLP. Thorsons, London.
Paivio, A., 1971. Styles and strategies of learning. British Journal of Educational Psychology 46, 128–148.
Pashler, H., McDaniel, M., Rowher, D., Bjork, R., 2008. Learning styles: concepts and evidence. Psychological Science in the Public Interest 9, 105–119.
Pask, G., Scott, B.C.E., 1972. Learning strategies and individual competence. International Journal of Man-machine Studies 4, 217–253.
Pask, G., 1972. A fresh look at cognition and the individual. International Journal of Man-machine Studies 4, 211–216.
Pask, G., 1976. Styles and strategies of learning. British Journal of Educational Psychology 46, 128–148.
Peterson, E.R., Rayner, S.G., Armstrong, S.J., 2009. Researching the psychology of cognitive style and learning style: is there really a future? Learning and Individual Differences 19, 518–523.
Peterson, E.R., Brown, G.T.L., Irving, S.E., 2010. Secondary school students conceptions of learning and their relationship to achievement. Learning and Individual Differences 20 (3), 167–176.
Rayner, S., Cools, E. (Eds.), 2011. Style Differences in Cognition, Learning and Management. Routledge, New York.
Rayner, S., Peterson, E., 2009. Re-affirming style as an individual difference – global paradigm or knowledge diaspora? In: Zhang, L.F., Sternberg, R.J. (Eds.), Perspectives on the Nature of Intellectual Styles. Springer Pubs, New York, pp. 107–134.
Rayner, S., Riding, R., 1997. Towards a categorisation of cognitive styles and learning styles. Educational Psychology 17, 5–28.
Rayner, S., Roodenburg, J., Roodenburg, E., 2012. Understanding an integrated theory of intellectual styles: models, measures, and meaning. In: Zhang, L.F., Sternberg, R.J., Rayner, S.G. (Eds.), Handbook of Intellectual Styles. Springer Pubs, NY, pp. 47–66.
Rayner, S., 2000. Reconstructing style differences in thinking and learning: profiling learning performance. In: Riding, R., Rayner, S. (Eds.), International Perspectives on Individual Differences, Cognitive Styles, vol. 1. Ablex, Stanford, CT, pp. 115–177.
Rayner, S.G., 2001. Cognitive styles and learning styles. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopaedia of the Social and Behavioural Sciences. Elsevier Press, Oxford, pp. 2171–2175.
Rayner, S., 2011. Researching style: epistemology, paradigm shifts and research interest groups. Learning and Individual Differences 21 (2), 255–262.
Rayner, S.G., 2013. Problematizing style differences theory and professional learning in educational psychology. The Australian Educational and Developmental Psychologist 30 (1), 13–35.
Reynolds, M., 1997. Learning styles: a critique. Management Learning 28, 115–133.
Richardson, A., 1977. Verbalizer–visualizer: a cognitive style dimension. Journal of Mental Imagery 1, 109–126.
Riding, R.J., Rayner, S.G., 1998. Cognitive Styles and Learning Strategies. David Fulton Pub, London.
Riding, R.J., Taylor, E.M., 1976. Imagery performance and prose comprehension in 7-year-old children. Educational Studies 2, 21–27.
Riding, R.J., 1991. Cognitive Styles Analysis. Learning and Training Technology, Birmingham.
Ritchie, R., Deakin Crick, R., 2005. Distributing Leadership for Personalizing Learning. Network Educational Press, Stafford.
Ritter, L., 2007. Unfulfilled promises: how inventories, instruments and institutions subvert discourses of diversity and promote commonality. Teaching in Higher Education 12 (5&6), 569–579.
Rose, C., Nicholl, M.J., 1997. Accelerated Learning for the 21st Century. Dell Publishing, NY.
Rosenfeld, M., Rosenfeld, S., 2011. Illustrating a complementary paradigm for styles research: from a third-person to a second-person perspective. In: Rayner, S., Cools, E. (Eds.), Style Differences in Cognition, Learning and Management. Routledge, New York, pp. 143–159.
Sadler-Smith, E., 1996. Learning styles: a holistic approach. Journal of European Industrial Training 20 (7), 29–36.
Sadler-Smith, E., 2009. A duplex model of cognitive style. In: Zhang, L.F., Sternberg, R.J. (Eds.), Perspectives on the Nature of Intellectual Styles. Springer, Heidelberg, pp. 3–28.
Sadler-Smith, E., 2012. Metacognition and styles. In: Zhang, L.F., Sternberg, R.J., Rayner, S.G. (Eds.), Handbook of Intellectual Styles. Springer Pubs, NY, pp. 153–172.
Schiffman, H., Messick, S., 1963. Scaling and measurement theory. Review of Educational Research 33 (5), 533–542.
Schmeck, R.J., 1988. Learning Strategies and Learning Styles (Perspectives on Individual Differences). Springer, New York.
Schmidt, P., 2012. The epistemological status of the style construct? Private Correspondence Received: 12.12.2013.
Sharma, G., Kolb, D., 2011. The learning flexibility index: assessing contextual flexibility in learning style. In: Rayner, S., Cools, E. (Eds.), Style Differences in Cognition, Learning and Management. Routledge, New York, pp. 60–77.
Sharp, J.G., Bowker, R., Byrne, J., 2008. VAK or VAK-uous? towards the trivialisation of learning and the death of scholarship. Research Papers in Education 23 (3), 293–314.
Sternberg, R.J., Zhang, L.F., Rayner, S.G., 2012. Epilogue. In: Zhang, L.F., Sternberg, R.J., Rayner, S.G. (Eds.), Handbook of Intellectual Styles. Springer Pubs, NY, pp. 415–418.
Sternberg, R.J., 1996. Styles of thinking. In: Baltes, P.B., Staudinger, U.M. (Eds.), Interactive Minds. Cambridge University Press, Cambridge, pp. 347–365.
Sternberg, R.J., 1997. Thinking Styles. Cambridge University Press, Cambridge.
Tiedemann, J., 1989. Measures of cognitive styles: a critical review. Educational Psychology 24, 261–275.
Vermunt, J.D., 1996. Metacognitive, cognitive and affective aspects of learning styles and strategies: a phenomenographic analysis. Higher Education 31, 25–50.
Vermunt, J.K., 2010. Latent class modeling with covariates: two improved three-step approaches. Political Analysis 18, 450–469.
Vermunt, J.D., 2011. Patterns in student learning and teacher learning: similarities and differences. In: Rayner, S., Cools, E. (Eds.), Style Differences in Cognition, Learning and Management. Routledge, New York, pp. 173–187.
Witkin, H.A., Ash, S.E., 1948. Studies in space orientation: IV further experiments on perception of the upright with displaced visual field. Journal of Experimental Psychology 43, 58–67.
Witkin, H.A., Goodenough, D.R., 1981. Cognitive Style: Essence and Origins. International Universities Press, New York.
Witkin, H.A., Dyk, R.B., Faterson, H.F., Goodenough, D.R., Karp, S.A., 1962. Psychological Differentiation. John Wiley & Sons, New York.
Zhang, L.F., Sternberg, R.J., 2005. A threefold model of intellectual styles. Educational Psychology Review 17, 1–53.
Zhang, L.F., Sternberg, R.J., Rayner, S.G. (Eds.), 2012. Handbook of Intellectual Styles. Springer Pubs, NY.
Zhang, L.F., Sternberg, R.J., Rayner, S.G., 2012. Intellectual styles: challenges, milestones, and agenda. In: Zhang, L.F., Sternberg, R.J., Rayner, S.G. (Eds.), Handbook of Intellectual Styles. Springer Pubs, NY, pp. 1–20.
Zimmerman, B.J., 2008. Investigating self-regulation and motivation: historical background, methodological developments, and future prospects. American Educational Research Journal 45 (1), 166–183.

Cohabitation: United States

Pamela J Smock, Population Studies Center, Institute for Social Research, Ann Arbor, MI, USA
Rhiannon Kroeger, Population Research Center, University of Texas at Austin, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Over the past several decades family patterns have changed markedly in the United States and numerous other countries. Prominent increases in nonmarital cohabitation have been central to this change. It is now the norm for individuals to cohabit, whether or not the cohabitation leads to marriage, and the vast majority of marriages now begin as cohabiting relationships. While cohabitation is now common across most population subgroups, there remain economic and racial ethnic differences in the extent to which cohabitation is tied to marriage. Those with higher incomes and education are more likely to transform their cohabiting relationships into marital ones. Major research issues about cohabitation include understanding the striking increase in its popularity, measurement, the implications of cohabitation for children, and the relationship between premarital cohabitation and marital stability.

Cohabitation: An Overview

Only a few decades ago, marriage was the main form of intimate coresidential partnership for heterosexual couples. Beginning in the 1970s, in many Western countries including the United States, couples began living together outside of marriage. Since that time, family demographers and other scholars have been tracking the continued rise in cohabitation and attempting to assess its consequences (Smock, 2000). While this article focuses on the United States, many of the trends and patterns discussed here pertain to other Western contexts.

While marriage remains popular and the vast majority of people express the desire to marry someday, it is no longer the case that marriage is the dominant family form. One indication of this is that cohabitation is not only a step toward marriage. While many cohabiting couples do go on to marry, most others break up within a few years, while a select few remain cohabiting for a significant number of years. Moreover, many cohabiting families include children, either by virtue of a child being born to a cohabiting couple or via a type of 'stepfamily' arrangement in which one parent, usually the mother, begins a cohabiting relationship.

Trends

What was once uncommon has now become the experience of the majority. Currently, there are over 7.8 million heterosexual unmarried couples living together unmarried in the United States, more than twice the number in 2000 (U.S. Bureau of the Census, 2012). One way to get a sense of the trend is to examine the percentages of marriages that begin as cohabiting unions. Four decades ago (1965–74) only 10% of marriages were preceded by cohabitation. This has increased to about two-thirds for couples marrying recently. Marrying without living together is thus now a much less common route to marriage, characterizing only one-third of marriages. In other words, while cohabitation has become increasingly separate from marriage in recent decades, the route to marriage increasingly includes cohabitation. An alternative way to understand the trend is the proportion of first coresidential unions that are cohabitations rather than marriages. About one-half of women ages 15–44 experienced cohabitation as their first union in 2006–10; this is substantially higher than in 1995 when that percentage was 34% (Copen et al., 2013).

Family scholars also use measures of whether an individual has ever cohabited to understand trends in cohabitation. For example, recent surveys indicate that among women ages 30–34, 70% report having ever cohabited compared to just 20% in 1987 (Manning, 2010). In addition, while the likelihood of remarriage after divorce has declined in recent decades, cohabitation appears to be replacing remarriage to some extent. In the 1970s, 24% of women cohabited after divorce. This fraction rose to nearly 50% by the early 2000s (Zeng et al., 2012). It is also important to note that cohabitation is quite common among very young adults, with about half having lived in a cohabiting relationship by the time they reach age 25 (Payne, 2011).

Differentials

When scholars first began studying cohabitation, researchers were interested in the characteristics of people who cohabit versus those who do not. Emphasis was placed on attitudinal factors (e.g., liberal vs conservative gender ideologies) and socioeconomic differentials (Smock, 2000). Today, because most people cohabit, questions about differentials tend to revolve around the issue of who does not cohabit and the different role that cohabitation plays in family formation by social class and for different racial and ethnic groups. Regarding the former, couples marrying without living together first tend do so because of religious beliefs, including individuals living in geographic areas with high levels of religious homogamy.

Regarding the latter issue, there is little difference in the likelihood of cohabiting along racial and ethnic lines, with two exceptions. One is that Asians are less likely to live together outside of marriage. A second is that more U.S.-born Hispanics report having ever cohabited (56%) than foreign-born Hispanics (49%) (Kennedy and Bumpass, 2008). What does differ in terms of race and ethnicity is whether a cohabiting

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31081-9

union evolves into a marriage. Specifically, fewer cohabiting non-Hispanic blacks or Hispanics marry their partners than whites (Zeng et al., 2012). Among Hispanics, this pattern is largely due to more couples remaining in long-term cohabitations than transitioning to legal marriage. Among non-Hispanic blacks, this pattern is explained by more couples dissolving their cohabiting unions than remaining together.

One important issue interconnected with race and ethnicity is economic circumstances. In the United States, on average, whites are economically better off than blacks or Hispanics. Researchers use various terms when studying the connections between family patterns and economic circumstances (e.g., social class, socioeconomic status) and measure it in various ways (e.g., educational attainment, income).

The bottom line is that people with more money and education are more likely to transform their cohabiting unions into marital ones. This finding is consistent with a large body of research showing that getting married, and staying married, is positively related to economic well-being. Consider the differential in the percentage of 22- to 44-year-old women whose first union was cohabitation. While cohabitation increased for all educational groups, there is a striking educational gradient; recent data indicate that 48% of women with college degrees experienced cohabitation as their first union compared to 70% for women who did not complete high school. People with college degrees are also more likely to use cohabitation as a step toward marriage; they have the highest chances of marrying their cohabiting partners compared to their less educated counterparts (Copen et al., 2013).

Major Research Issues about Cohabitation

The Rise in Cohabitation

The steep rise in cohabitation has several sources. Scholars view cohabitation as part of a constellation of family change that has swept the US and other Western countries over the past several decades. These changes include rising levels of divorce, increases in nonmarital childbearing, and delays in marriage. Scholars consider both cultural and economic factors to be important in understanding the increase in cohabitation (Smock, 2000). The former includes factors such as rising individualism and secularism, with a growing emphasis on the importance of achieving individualistic goals (Lesthaeghe, 1983). Economic factors include macro-level changes such as industrialization and women's changing role in the labor market. Also important has been the sexual revolution. Once premarital sex became acceptable, cohabitation became destigmatized (Bumpass, 1990).

An additional factor is feedback loops. What this means is that the various trends are reinforcing. High levels of marital disruption demonstrate that marriage is not necessarily permanent, resulting in more people opting for cohabitation. If childbearing outside of marriage is common, it then becomes clear that marriage is not the only possible or appropriate context for childbearing.

At a more micro-level, couples appear to choose cohabitation for numerous reasons. Some are practical matters such as convenience (e.g., the couple is already spending significant time together), the ability to live more cheaply, or as a test to evaluate whether the relationship is marriage material (Sassler, 2004; Huang et al., 2011). Importantly, it offers many of the benefits of marriage such as income-pooling, companionship, sexual access, childbearing, and childrearing. Studies also indicate that many young adults may want to marry but choose cohabitation or remain cohabiting until they feel ready. In particular, young adults wish to be in a financially stable situation before considering marriage. The perceived requirements include having enough money to afford a house and mortgage, having a stable job, saving sufficient money for a wedding and reception, and not living paycheck to paycheck (Gibson-Davis et al., 2005; Smock et al., 2005).

Measurement

To study a phenomenon, it has to be measured. National surveys began to gather detailed information about the cohabitation experiences of individuals in the 1980s. Arguably, the survey that spawned an enormous amount of knowledge about cohabitation was the 1987–88 National Survey of Families and Households. It obtained full cohabitation histories from a nationally representative sample of adult women and men ages 19 and up, setting the standard for future surveys. Still, until the 1990s, many national data sources, such as the U.S. census, did not directly measure cohabitation. Instead, scholars inferred cohabitation from other types of information about household members and the composition of the household. For example, if there was a two-person household comprised of a male and a female and whose ages were fairly close, researchers would infer cohabitation.

Though numerous national data sources now directly measure cohabitation, many challenges to accurate measurement remain. Marriage is relatively easy to measure. People may have a wedding or go to the courthouse, and there is a legal document with a specific date of marriage. This is not the case for cohabitation. Family scholars have thus studied how cohabitation is measured in surveys and in other data sources, and have played a role in changing how it is measured. Qualitative work has been extremely useful in this endeavor. By asking open-ended questions, qualitative work can delve more deeply into people's cohabitation experiences than close-ended survey questions.

Several conclusions have emerged from this work. First, cohabitation is often a gradual slide than a living arrangement associated with a particular start date. It may begin by spending weekends together and slowly move toward living together full time. Thus, people do not necessarily know exactly when they started cohabiting. Second, cohabiting couples are not accurately measured in surveys and census data that define relationships through ties between each household member and a specific household 'head' or reference person. That is, the convention was that every person living in the household was identified in terms of his or her relationship to the household head or reference person. If a cohabiting couple existed in the household but the reference person was not a member of the cohabiting couple, the existence of that couple was missed. This includes situations in which the couple resides with roommates, parents, or other relatives. Research showed that these missed couples were more likely to be economically disadvantaged and more likely to have children, groups of great

interest to scholars interested in families and their well-being. Third, major data sources such as the Census use the term 'unmarried partner'; qualitative work suggested that this was not a term people understood, let alone one that cohabitors would use to describe their relationships (Manning and Smock, 2005).

At least in part as a result of this work, the Current Population Survey (CPS), an ongoing survey that is part of the U.S. Bureau of the Census, made some major alterations regarding cohabitation measurement in 2007. It added a direct question about cohabitation, asking if the individual was currently living with a 'boyfriend/girlfriend/partner,' allowing researchers to capture cohabitations in which the reference person was not part of the couple. There were also improvements that led to being able to identify the relationship of all children in the household to their parent(s). Such changes have been significant. In 2008, 5% of all cohabiting unions in the CPS were between two household members that did not include the reference person. Moreover, the addition of the 'boyfriend/girlfriend/partner' question captured 20% more cohabitations than the 'unmarried partner' question (Kennedy and Fitch, 2012).

Continuing challenges include the lack of a clear start date, and, not surprisingly, inconsistency in reports of whether a couple is actually cohabiting depending on which member of the couple is asked (Waller and McLanahan, 2005). Another issue is the timing of data collection. The accuracy of cohabitation histories decreases as the span of time between the cohabitation and the date of interview increases (Hayford and Morgan, 2008).

Children and Cohabitation

Cohabitation commonly involves children. About 40% of cohabiting-couple households include children; this is about the same as the percentage of children in married-couple households. Up to one-half of children born in the early 1990s will live in a cohabiting-parent household at some point in their childhood (Bumpass and Lu, 2000; Kennedy and Bumpass, 2008). One way for children to experience this family form is by being born to a cohabiting couple. This trend is on the upswing. The percentage of all U.S. births that were to cohabiting parents increased from 6% in the 1980s compared to about 20% today (Bumpass and Lu, 2000; Lichter, 2012). Overall, the percentage of births taking place outside of marriage is 40%. Now the majority of these births are to cohabiting couples rather than to single women.

A central focus of research on cohabitation and children is whether cohabitation is harmful to children. In general, comparisons of children growing up in cohabiting versus married households suggest that cohabitation has some negative effects on child well-being. While many children fare well in cohabiting households, parental cohabitation has been linked to a range of negative outcomes such as increased odds of teen pregnancy for girls, poorer educational outcomes, and more behavioral problems.

At the same time, it is not clear that parental cohabitation itself causes worse outcomes for children. In other words, children experiencing parental cohabitation have some disadvantages to begin with compared to children born to and raised by married couples. One difference is economic; cohabiting families have lower incomes than married-couple families. The connection between parental cohabitation and child well-being is also partly explained by family instability, that is, by the parents dissolving their relationship. It is well documented that family instability tends to have negative consequences for children, with children typically doing best in stable family structures. Children born to cohabiting parents are five times more likely to see their parents breakup within 3 years of their birth than children born to married parents (Osborne et al., 2007). It is important, however, to be cautious about the generalization that cohabitation may not be in children's best interest. The quality of the cohabiting parents' relationship has been found to be very important and positively linked to parental engagement with the child (Carlson et al., 2011).

Cohabitation and Marital Stability

For many years, studies showed that premarital cohabitation raised the likelihood of divorce: the chances of divorce were higher for people who had previously cohabited either with their future spouse or someone else. This was demonstrated in many studies and most of those studies took into account other factors that might also raise the risk of divorce (Smock, 2000). Two explanations were posed to account for this pattern. The first was that cohabitation affected people in ways that made them more likely to end a marriage. The idea was that living in a nonmarital family form could change people's attitudes about the centrality and permanence of marriage. The other explanation was that premarital cohabitation drew people with values, attitudes, or other qualities that potentially made their marriages more fragile. For example, cohabitors tended to be less religious with less conventional family values than people who did not cohabit.

A shift in what appeared to be a solid finding has occurred as cohabitation has become the modal experience. Indeed, studies using recent data are overturning conclusions about premarital cohabitation and increased marital instability. One study reports that the positive link between premarital cohabitation and marital instability has decreased for more recent birth and marriage cohorts and that cohabitation does not appear selective of individuals with higher chances of divorce. Another finds that for marriages begun in the late 1990s and later, premarital cohabitation is not associated with divorce for either women or men (Manning and Cohen, 2012; Copen et al., 2012). It has also recently been shown that premarital cohabitation may even make remarriages more stable and less likely to dissolve (Reinhold, 2010).

Conclusion

Recent decades have seen sweeping changes in the family, perhaps none more striking than those related to nonmarital cohabitation. In addition to tremendous increases in the prevalence of cohabitation, it is also increasingly a context for childbearing, suggesting erosion in the dominance of marriage. Indeed, it has been argued that marriage is becoming 'deinstitutionalized' (Cherlin, 2004). However, another way to understand family trends is that neither marriage nor

cohabitation is institutionalized in the way that marriage once was. That is, while there still may be differences between cohabitation and marriage, as demonstrated by most people valuing marriage, hoping to marry, and eventually doing so, people are increasingly writing their own scripts about family formation and the meaning of families.

Family patterns are constantly evolving. As scholars have noted, studying families is akin to studying a continually evolving phenomenon (Seltzer, 2004; Smock, 2004; Smock and Greenland, 2010). Thus, in the years to come our understanding of cohabitation and its implications for individuals and for the broader family landscape will shift as well. All that seems certain is that cohabitation is not going away.

See also: Families and Households, Formal Demography of; Family Theory: Competing Perspectives in Social Demography; Family as Institution; Fertility of Single and Cohabiting Women; Partnership Formation and Dissolution in Western Societies; Repartnering and Stepchildren.

Bibliography

Bumpass, L.L., 1990. What's happening to the family? Interactions between demographic and institutional change. Demography 27, 483–498.

Bumpass, L.L., Lu, H.H., 2000. Trends in cohabitation and implications for children's family contexts in the United States. Population Studies: A Journal of Demography 54, 29–41.

Carlson, M.J., Pilkauskas, N., McLanahan, S.S., Brooks-Gunn, J., 2011. Couples as partners and parents over children's early years. Journal of Marriage and Family 73, 317–334.

Cherlin, A.J., 2004. The deinstitutionalization of American marriage. Journal of Marriage and Family 66, 848–861.

Copen, C.E., Daniels, K., Vespa, J., Mosher, W.D., 2012. First marriages in the United States: data from the 2006–2010 national survey of family growth. U.S. Department of Health and Human Services, National Health Statistics Reports 49.

Copen, C.E., Daniels, K., Vespa, J., Mosher, W.D., 2013. First premarital cohabitation in the United States: 2006–2010 national survey of family growth. U.S. Department of Health and Human Services, National Health Statistics Reports 64.

Gibson-Davis, C.M., Edin, K., McLanahan, S., 2005. High hopes but even higher expectations: the retreat from marriage among low-income couples. Journal of Marriage and Family 67, 1301–1312.

Hayford, S.R., Morgan, S.P., 2008. The quality of retrospective data on cohabitation. Demography 45, 129–141.

Huang, P.M., Smock, P.J., Manning, W.D., Bergstrom-Lynch, C.A., 2011. He says, she says: gender and cohabitation. Journal of Family Issues 32, 876–905.

Kennedy, S., Bumpass, L.L., 2008. Cohabitation and children's living arrangements: new estimates from the United States. Demographic Research 19, 1663–1692.

Kennedy, S., Fitch, C.A., 2012. Measuring cohabitation and family structure in the United States: assessing the impact of new data from the current population survey. Demography 49, 1479–1498.

Lesthaeghe, R., 1983. A century of demographic and cultural change in Western Europe: an exploration of underlying dimensions. Population and Development Review 9, 411–435.

Lichter, D.T., 2012. Childbearing among cohabiting women: race, pregnancy, and union transitions. In: Booth, A., Brown, S.L., Landale, N.S., Manning, W.D., McHale, S.M. (Eds.), Early Adulthood in a Family Context. Springer, NY, pp. 209–219.

Manning, W.D., 2010. Trends in Cohabitation: Twenty Years of Change, 1987–2008. http://ncfmr.bgsu.edu/pdf/family_profiles/file87411.pdf (accessed 28.04.13.).

Manning, W.D., Cohen, J.A., 2012. Premarital cohabitation and marital dissolution: an examination of recent marriages. Journal of Marriage and Family 74, 377–387.

Manning, W.D., Smock, P.J., 2005. Measuring and modeling cohabitation: new perspectives from qualitative data. Journal of Marriage and the Family 67, 989–1002.

Osborne, C., Manning, W.D., Smock, P.J., 2007. Married and cohabiting parents' relationship stability: a focus on race and ethnicity. Journal of Marriage and Family 69, 1345–1366.

Payne, K.K., 2011. On the Road to Adulthood: Forming Families. http://ncfmr.bgsu.edu/pdf/family_profiles/file101794.pdf (accessed 28.04.13.).

Reinhold, S., 2010. Reassessing the link between premarital cohabitation and marital instability. Demography 47, 719–733.

Sassler, S., 2004. The process of entering into cohabiting unions. Journal of Marriage and the Family 66, 491–505.

Seltzer, J.A., 2004. Cohabitation in the United States and Britain: demography, kinship, and the future. Journal of Marriage and Family 66, 921–928.

Smock, P.J., 2000. Cohabitation in the United States: an appraisal of research themes, findings, and implications. Annual Review of Sociology 26, 1–20.

Smock, P.J., 2004. The wax and wane of marriage: prospects for marriage in the 21st century. Journal of Marriage and Family 66, 966–973.

Smock, P.J., Manning, W.D., Porter, M., 2005. "Everything's there except money": how money shapes decisions to marry among cohabitors. Journal of Marriage and Family 67, 680–696.

Smock, P.J., Greenland, F.R., 2010. Diversity in pathways to parenthood: patterns, implications, and emerging research directions. Journal of Marriage and Family 72, 576–593.

U.S. Census Bureau, 2012. Table UC3. Opposite Sex Unmarried Couples by Presence of Biological Children under 18, and Age, Earnings, Education, and Race and Hispanic Origin of Both Partners. http://www.census.gov/hhes/families/data/cps2012.html (accesssed 23.03.13.).

Waller, M.R., McLanahan, S.S., 2005. "His" and "her" marriage expectations: determinants and consequences. Journal of Marriage and Family 67, 53–67.

Zeng, Y.S., Morgan, S.P., Wang, Z., Gu, D., Yang, C., 2012. A multistate life table analysis of union regimes in the United States: trends and racial differentials, 1970–2002. Population Research and Policy Review 31, 207–234.

Cold War: Impact on Social and Behavioral Sciences

Hamilton Cravens, University of Minnesota, Minnesota, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Cold War encouraged the development of engineering and behaviorally controlling assumptions, concepts, and methods in the social sciences, especially in the United States and the nations in its orbit. Globalization and the multiple perspectives it revealed ultimately overrode Cold War notions and nostrums.

Cold War's Impact

What was the Cold War's impact on the social and behavioral sciences? Many historians argue that the Cold War began in 1947 and ended in 1991. It was the bipolar competition between the United States and the Soviet Union, together with each nation's allies, between each alliance, and for the allegiances of the remaining 'neutral' nations. (Gaddis (1972), especially pp. 283–362, the standard conservative view; for provocative and contrary views of the Cold War's duration and characteristics, see several of the essays in Joel Issac and Duncan Bell (2012), especially pp. 19–50, 91–141, et passim. It could be argued that the Cold War's impact on the social sciences lasted much longer than the Cold War itself.) In the late 1940s and early 1950s the United States cemented Western Europe in its orbit, together with several Asian states, including Japan, South Korea, and South Viet Nam. The Soviet government kept almost all of the Red Army's wartime gains in Eastern Europe, although in time it permitted West Berlin to remain free and Austria to be neutralized. Each superpower attempted to impose or insinuate its preferred social, political, and economic systems on each of its subordinate or allied states. Those associated with the United States formed democratic, market capitalist societies, with social welfare systems, while in the Russian bloc, Soviet forms of government, enterprise, science, education, culture, and social welfare were imposed from Moscow.

In assessing the impact of the Cold War on the social and behavioral sciences, we must realize that there were important differences between the impact of the Cold War in the Western world, and in the Communist nations, chiefly the Soviet Union, its Eastern European satraps, and the Peoples' Republic of China. National intellectual and political differences made for different results. And there were differences in the 'underdeveloped world,' as in Latin America, and in the areas emancipated by post-World War II decolonization, as in parts of Asia, notably India, and the Middle East. The picture is very complex. Not all nations experienced the development of the social sciences at the same time in history. Most social sciences existed before the Cold War began. The Cold War had an enormous influence in reshaping and expanding the social sciences' character: their approaches, research agendas, funding possibilities, and all manner of other uses for public and private enterprises. It also sparked critical, even opposite, reactions, such as democratic Marxist movements in the West and its allies, and dissent in the Communist world.

Impact of the Cold War in the Americas

In the United States, the federal government underwrote much behavioral science research in economics, political science, area studies, psychology and sociology and related endeavors through various agencies and departments of defense, intelligence, science, and medicine. The number of dollars available for social and behavioral science research and practice from private philanthropy was perhaps one-tenth of that available from federal sources in a typical year during the Cold War.

Within the U.S. social sciences, there was the emergence of the behavioral sciences, clearly a Cold War product, from its Second World War origins as a kind of engineering social science to devise methods and approaches to win the global conflict. The behavioral scientist, as distinct from the social scientist, sought to create rigorous methods, often informed by high mathematics, for understanding and predicting the behavior of individuals, of groups, of institutions, and of entire nations; the social scientist, traditionally, sought to gather information and to explain what happened or what would happen. Social scientists had used descriptive statistics since the mid-nineteenth century. Behavioral scientists went further and deployed sophisticated probability techniques and applied methods from inferential statistics. In the 1950s and 1960s, to take one example, political scientists experienced a 'behavioral revolution.' They imitated natural scientists in the sense that they employed formal classifications, correlations, appeals to synchronic systems and structures, as well as the formal location and purpose of units within larger political systems. They moved away from prior developmental historicism and toward a neopositivist empiricism. They often used one or another modernization scheme when talking about comparisons between the developed and the underdeveloped worlds – between political cultures in the industrialized West and the preindustrial 'third world.' Implicit was often the shopworn assumption of the general superiority of the West over the preindustrial regions. Thus were absolutist patriotism and scientific positivism parallel underlying assumptions of American social and behavioral science. The domestic and international tensions of the Cold War simply magnified these attitudes. In the 1950s, American fears of potential Soviet nuclear parity – later proven to be false – drove Americans, whether social scientists or not, to intensify their opposition to Soviet aggression on all fronts, real and perceived.

Although the Cold War introduced new issues and inspired new techniques, such as futures studies or game theoretical

nuclear war strategies, basic political attitudes and cleavages were not so much changed as intensified, especially a heightened patriotism and a worshipful faith in scientific and engineering rationality – scientific positivism, that is. Thus American sociology was divided into various camps. The structural–functional school of Harvard's Department of Social Relations, led by sociologist Talcott Parsons, emphasized the adjustment of the individual to the larger social system, and Robert K. Merton, of Columbia, promoted a similar but more modest version adjustment according to theories of middle range application. These might well characterize the posture of the postwar liberal in American politics. More radical were sociologists who followed the social conflict ideas of Lewis Coser, who insisted that conflict among differing social interests, or groups, constituted the warp and woof of society. Even more radical were C. Wright Mills, of Columbia, and his hardy followers, who declared that the social system had been taken over by powerful oligarchies and elites and that the average 'mass man' had little agency in society. Alvin Gouldner insisted that mainstream sociology simply upheld the status quo. All were fierce critics of Cold War conservatism, as well as what they believed was social science that buttressed Cold War conservatism.

Economists were similarly split into several camps. The macroeconomists who followed John Maynard Keynes had the experience of the 1940s to the 1970s to show that their aggregate demand economics did lift employment and produce prosperity; World War II and the Cold War sparked such great government spending which, together with certain tools, such as national income accounting and specific mathematical theorems, spelled unparalleled economic growth for the United States. Yet there was at least one other kind of economics, neoclassical microeconomics, or the study of the individual agent or firm in a market, which feted a philosophy of individualism, not the aggregate, whose proponents pushed it into the terrain of abstract mathematics, and rigorous axiomatic logic, thus laying claim to a higher level of scientific, positivistic economics. With the problems of 'stagflation' of the 1970s in the economy, the neoclassicists bested the Keynesians in their discipline and in the public arena. Hence partisans of conservative neoclassical economics won out in the 1980s and early 1990s, as the Reagan Administration scuttled détente with the Soviets and the Cold War heated up, as it had not since the later 1960s. There were also 'heterodox economists': they included libertarians, who wanted to wish government away; Marxist economists, who pined for social democracy; and 'post-Keynesian' economists, who worked to convince others that Keynes's doctrines had a rigorous mathematical, as well as a policy basis, and better for all than the dog-eat-dog ideology they accused the neoclassical economists of peddling. But this strange assortment of 'heterodox economists' could only agree among themselves that the neoclassical conservatives were wrong; by the 1990s they were split into warring groups themselves among post-Keynesians, libertarians, feminists, and Marxists – among others.

Development of the neutral or 'underdeveloped' world was also a part of hemispheric social and behavioral science. The trailblazers were likely the anthropologists, who during the 1940s and 1950s, turned from the study of man's origins with preliterate tribes to the study of contemporary peoples in exotic areas of strategic importance to the United States. The Cold War made such work even more crucial to American policymakers, who believed the nation was engaged in a twilight struggle, as President Kennedy put it, for the 'underdeveloped' world's allegiance against the Communist menace. According to Walt W. Rostow, sometime economic historian and presidential adviser, all nations followed certain stages of economic growth to urbanization and industrialization; this became high policy with which to thwart Communism and to promote democratic capitalism around the globe. During the Cold War, the United States' military, intelligence, and international development agencies funded numerous social science investigations in the 'underdeveloped nations' seeking, in the case of South Viet Nam, to support the Diem dictatorship, or in the case of Chile, to carry out through 'Project Camelot,' a study of how to design a general systems model to predict and control internal strife and promote counterinsurgency in peripheral countries. In any given time there were perhaps several dozen such covert social science operations functioning with governmental support.

Some social science expertise was indigenous in Latin America. The nineteenth century independence movements fostered champions of liberal statism, Comtean positivism, and German historical or institutional economics. Between the Second World War and the 1970s, as Latin America entered the world economy, its economists, whether proficient academically or self-trained, wrote much about the economics of underdeveloped countries and regions – that is, their region. For example, the left-liberal Argentine economist Raúl Prebisch developed from the 1940s on a center–periphery theory of economic underdevelopment that was both theoretical and historical, that social scientists and governmental *apparatchiks* throughout Latin America took up. He also created a network of like-minded economists through the Economic Commission for Latin America (ECLA), located in Santiago, Chile, and which nurtured a Latin American economics profession anxious to promote its region. Certain groups of United States economists, as the neoclassical economists at the University of Chicago, established rival groups of conservative economists, as at the Catholic University of Chile, who went on to advise the region's rightist regimes, such as Pinochet's in Chile. Here the regional stimulus was Fidel Castro's Cuba, against which both the United States and the rightist regimes reacted, plotted, and schemed. The Cold War definitely spilled over into the Western Hemisphere. The social scientists, as retainers of various governments, smartly followed behind their patrons.

The Impact of the Cold War in Western Europe

Before the Second World War, social science's intellectual center was Europe. On the European continent, dictatorships arose in the interwar and war years from Iberia to the Urals, save for England, France, and a few neutrals, such as Switzerland and Sweden. Whether these dictatorships were Fascist, National Socialist, or Marxist-Leninist, their leaders destroyed or drove into exile most of their social sciences – and practitioners. Only those social sciences that pleased the dictators, such as eugenics in Nazi Germany, or eugenics and psychology in the Soviet Union, survived the war years. Thereafter consequently the United States became a world power in the social

sciences, thanks in addition to its expansion and mobilization of social scientists. Europe's distress was America's advantage.

In postwar Western Europe – the areas allied with the United States and liberated from Nazi or Fascist control – there was a rebuilding of society and economy, thanks in no small measure to the Marshall Plan. Everywhere the Americans there behaved much as they did at home, so that national autonomy and democracy emerged in Western Europe. There the masses demanded, and got, sometimes with the threat of Communist influence, a new deal of social welfare legislation which in turn necessitated the training and employment of social scientists to administer the new states' social welfare, economies, political democracies, and defense forces, in the form of the *North Atlantic Treaty Organization* (NATO), formed as a *cordon sanitaire* against the engorged postwar Soviet empire to the east. The Western European countries demanded NATO from the U.S. as the price of the founding of the West German Republic in 1949 and its rearmament in 1955. Thus an American empire was created there, but by mutual consent and invitation, not terror, coercion, or imperial rule from Washington, as was the case in Soviet-controlled Eastern Europe.

Hence the social sciences in Western Europe followed both national traditions and American influence, with the latter usually secondary. Thanks to the American influence, or lack thereof in other cases, the Cold War stimulated an intellectual renaissance. In Great Britain, for example, sociology came into its own, even appearing in Oxford and Cambridge by the 1960s; initially British sociology followed American patterns, but after 1970 other, more indigenous traditions, often ruggedly empiricist, reasserted themselves. The American influence could be felt mainly in institutional forms, as in disciplinary departments and formal associations. On the Continent, American and European traditions clashed and melded. Thanks to American influence in the United Nations Economic, Social, and Cultural Organization (UNESCO), American forms of social science organization, as in disciplinary departments and associations transferred to the Continent, but theory was a different matter. In France, the Marxist influence on sociology was evident, as were the ideas of Michel Foucault and others. In West Germany, American intellectual traditions vied with indigenous ones, such as the returned social science exiles, such as the Frankfurt School, but also conflict theorists such as Ralf Dahrendorf. And there were democratic Marxist social scientists as well, thanks in some measure to the Social Democratic Party. Anthropology in America became focused on contemporary cultures, whereas anthropology in Europe tended to focus on the effects of colonization. American anthropology became useful to the nation's cold war social sciences, whereas European anthropology was largely oriented away from the Cold War's direct influences.

Two disciplines in particular, psychology and economics, betrayed Cold War influences but also became international technical behavioral sciences. Psychology became increasingly Americanized. In West Germany, professors from the Nazi era retired only in the 1960s, when a younger generation emerged, insisting on American-style psychology, meaning data-driven research and statistical treatment of results. In Britain, psychology closely paralleled the American version. And by the 1970s, everywhere else in Western Europe, the Anglo-American synthesis of cognitive behaviorism dominated. Economics also mutated into an increasingly technical behavioral science, which split into conflicting macroeconomic and microeconomic branches, often with the former pushing a Keynesian or post-Keynesian agenda, and the latter an updated version of Austrian individualistic economics. The technical methods psychologists and economists used made their sciences international, but the methods themselves were largely American and Cold War-inspired in origin. If the technical innovations were largely a Cold War legacy, their cosmopolitanism betrayed both American and European elements. Thus the Europeans retained a sense of society as a collectivity, whereas the Americans' underlying attitudes always retained some agency for individuals – and individualism.

The Impact of the Cold War in the Soviet World

Josef Stalin suppressed most of the social sciences in the Soviet Union well before the Cold War. He would not tolerate sociology, law, political science, or economics, unless derived from his notions of Party orthodoxy. The construction of a Soviet centralized system of science permitted only one social science – psychology – some latitude, largely because Ivan Pavlov's work had been reconciled with dialectical materialism, the Party's official philosophy. Dialectical materialists sought to explain the world by insisting that all that exists is real; the real world exists as matter–energy; and this matter–energy develops in accordance with universal regularities or laws. They thus refuted idealism, especially the Western versions. Yet ironically dialectical materialism permitted Soviet scientists to engage in philosophical jargon – 'newspeak' – that in turn permitted considerable flexibility, as well as camouflage for scientific discourse safe from Party censors' onslaughts. Thus Soviet psychologists rehabilitated consciousness and other mental constructs justifying attempts to create a new 'Soviet personality,' a Party priority, and Soviet cybernetics, again a patriotic necessity, developed from 'newspeak' discussions among those to compete with America in computers (Slava Gerovitch, 2002: Chapters 1–6). Cybernetics was indeed a Cold War interdisciplinary behavioral science.

The Soviet imprint was manifest on the social sciences in its satellites in the Cold War era. The social sciences received less funding and more intellectual control than other sciences, thanks to the monopoly of the Communist Party in each country. The Russian occupiers made sure their hierarchical and centralized system of government institutes of science was created in each satellite. The Party governments in each country used the social sciences to enforce Party hegemony. Results differed from country to country. In East Germany, for example, until the 1960s, Party officials permitted Nazi-era psychologists to retain their chairs on the Leninist ground that bourgeois scientists could work for the Party; as this generation retired, indigenous devotees of Soviet psychology replaced them. Yet even these professors mixed Soviet-style cognitive research with American information-processing approaches as a new 'Marxist' approach congruent with the indispensable current 'scientific–technical revolution.' Poland's

national consciousness was the strongest in Soviet Europe. Catholicism and repressed Polish national pride kept Polish intellectuals free in their private thoughts, at least; and whatever had the imprint of the Russian bear was distasteful to most Poles. Polish social science during the whole period was more up to date with Western social science than social science in other Soviet satellites. Psychology flourished in Poland as the study of national character, seen as a collective entity. Even before the Soviet collapse, Poland was an unreliable ally, as the emergence of the labor union Solidarity in the 1980s suggested. Often the social sciences were weak in some satellites, such as Bulgaria or Yugoslavia, and the imposition of Soviet orthodoxy came easily. Yet in all the satellite countries, the social scientists turned inward, toward studies of their own national character; commonly they paid mere lip service to Moscow. In this way they gained some distance, some breathing room, from Marxist-Leninist orthodoxy. Hence the Cold War had an enormous impact, institutionally and, to a lesser extent, intellectually, upon Central and Eastern Europe, especially with the establishment of Communist governance, something that the Soviets had known since Stalin consolidated his rule in the late 1920s in Russia.

The Impact of the Cold War in Asia

Among Asiatic Communist states – chiefly the People's Republic of China and the Democratic Republic of North Korea – the social sciences were completely under the thumb of their dictators and Party officials. Neither Mao Zedong nor Kim Il-sung would permit competition with their social doctrines. In the 1930s, however, Chinese social scientists had imported Western social science to facilitate the country's modernization, especially in sociology and economics – precisely the fields that Mao would commandeer and submit to his ideas. Hence in the Cold War era, non-Communist social scientists retreated to Taiwan and Hong Kong. In Taiwan, those who retained an ideological affinity with the mainland were suppressed. In Hong Kong, thanks to British colonial traditions of political stability and academic freedom, Chinese social scientists developed their teaching and research often along the lines of American and British social sciences, where many had received their postgraduate training. In South Korea, at first the social sciences were a pale imitation of the American social sciences, especially sociological structural–functionalism and conservative microeconomics. Eventually, as South Korea's political culture became increasingly democratic, so its social scientists began to study problems facing their nation, such as national division, the status of a postcolonial society in the world economy, and related issues.

Two major Asian states, Japan and India, demonstrated much about the Cold War's impact beyond the Americas and Europe. In the early postwar years Japan's American occupiers democratized society, politics, and the economy, although Japanese cultural traditions of collectivity, obedience, and central authority remained below the surface. Social scientists then wished to eliminate Japan's negative history and reputation, and adopted many American social science ideas and methods. Yet in the 1970s and beyond, Japanese Marxism, a product of the 1920s, was revived as a way to reassert Japan's intellectual identity apart from Western influences. Japan was doubtless an American ally in world politics, as were the nations of Western Europe, which tilted the Cold War toward the United States. Yet there was a reassertion of Japanese cultural and intellectual identity and independence, married with American-style disciplinary formation and institutionalization.

With British rule in the eighteenth century came a rich and complex tradition of social research – for administrative and imperial purposes, to be sure – in India. After independence in 1947, Indian sociologists and anthropologists at first gravitated to American-style structural–functionalist analysis of modernization. They focused on rural communities undergoing social change. The village they assumed was a functional unit with different caste groups, factionalism, class antagonisms, patron–client relations, and the relation of the village to the outside world. The University of Chicago's sociology programs heavily influenced these scholars. One could also see the influence of structuralism in studies of caste, kinship structure, ritual, and religious practices and beliefs. Yet there was also a growing interest in the use of Marxian methods of social and economic analysis. In economics, independence brought forth the problem of national economic development, and in the early postwar years there was much interest in government planning of the economy. Indian social scientists sought to conform to international standards of excellence, which meant they looked toward the United States, Western Europe, and Japan, not the Communist world, for their ideas. In part due to the British influence, Indian economists used statistical investigations of Indian economic life to help central planning. Economists such as Amartya Sen and others asked basic questions about equity and justice in Indian modernization. From the 1970s on, government planning fell into disfavor and economists began asking more specific questions about, for example, whether agriculture promoted or retarded development, about whether government promotion of industrialization was beneficial or harmful, the role of technology, and the impact of the world economy and of foreign investment on Indian economic growth. Indian political thinking was predominantly liberal, and this was reflected in its governmental structures and among its political scientists. Indian political scientist Rajni Kothari, for example, insisted in 1970 there was a dynamic core in the Congress Party's dominance of national democratic politics. Using a structural–functional analysis, Kothari argued that the "Congress system" connected government and party at different levels, from the nation's capital to the cities and the rural hinterlands. This accommodated dissent and helped stimulate needed correctives. Yet, perhaps inevitably, there was criticism from an indigenous Marxist social democratic left, not unlike developments in Western Europe and Japan, and left-liberal Marxian analysis informed dissent among Indian social scientists and politicians.

Conclusion

What, then, was the impact of the Cold War on the social and behavioral sciences? It encouraged, from developments in the Second World War, a positivistic, manipulative, and engineering approach to human behavior in all its dimensions. While the Cold War was in its early, intense phases – until the

resolution of the Berlin and Cuban missile crises of the early 1960s – the social and behavioral sciences closely mirrored the values and remedies of their sponsoring central governments. These political and intellectual straitjackets began to fray and dissolve outside the Communist world by the 1970s, thanks to liberal and radical dissent in the United States, in Western Europe, in Japan, and even in India, and if the Cold War idea of using social science for social engineering maintained its vitality, the Cold Warrior attitudes diminished considerably, and national differences and preferences asserted themselves. In the Communist world, it was only when the Soviet Union collapsed, and China came under Mao's successors, Marxist-Leninist rhetorically but nationalistic and capitalistic in fact, that matters changed. In Central and Eastern Europe, each former satellite retained, to some extent, Russian-style centralization of cultural life, but the expansion or reinvigoration of universities promoted Western-style standards of social science teaching and research, again, with each nation's elite asserting its distinctive national cultural identity. In the People's Republic, social science that supported the nation's development, for example, sociology and demography emerged from the Party's stranglehold.

An indirect influence of the Cold War, and not just in the United States, was the rise of ideologically driven think tanks, such as the right-wing Heritage Foundation in the United States, which employed social scientists as spokespersons for their ideas to influence public opinion and governmental policy. As such the think tanks came to challenge the intellectual authority of the social and behavioral sciences in academic and even government academy circles. Here was Cold War social engineering, often financed by private interests, creating potentially powerful Non-Governmental Organizations (NGOs) that could direct or redirect public policy along risky, even dangerous pathways, with little, if any, influence from ordinary citizens. The United States likely had the most such think tanks. There was no small irony in this, for throughout the nineteenth century, and into the twentieth, American social science was often a folk movement, democratic to the core.

In conclusion, the Cold War dominated certain aspects of the social and behavioral sciences around the globe for some years, but clearly after 1970 or so, and especially with the passing of Mao in China and much later, the collapse of the USSR, the Cold War was in the main a memory, and not always a pleasant one to those involved. Yet in the long perspective, the world's peoples had become increasingly aware that they were living in a global civilization. The Cold War obscured and retarded that perception for several decades. But the fact of globalization trumped Cold War perspectives. Now the world appeared highly diverse, constituted of *n*-dimensions, proportions, and images.

See also: Coser, Lewis A. (1913–2003); Dahrendorf, Ralf (1929–2009); Keynes, John Maynard (1883–1946); Marxist Approaches to Social Work; Mills, Charles Wright (1916–62).

Bibliography

Backhouse, Roger E., Fontaine, Phillippe (Eds.), 2010. The History of the Social Sciences since 1945. Cambridge University Press, New York.

Berghahn, Volker R., 2009. America and the Intellectual Wars in Cold War Europe. Princeton University Press, Princeton, NJ.

Collander, David C., Lendreth, Harvey (Eds.), 1991. The Coming of Keynesianism to the United States. Conversations with the Founders of Keynesian Economics. Edward Elger, London.

Coser, Lewis, 1956. The Functions of Social Conflict. The Free Press, New York.

Davidson, Paul, 1991. Controversies in Post Keynesian Economics. Edward Elger, London.

Dahrendorf, Ralf, 1969. Society and Democracy in Germany. Anchor Books, Doubleday and Company, New York.

Debreu, Gerard, 1959. Theory of Value. Yale University Press, New Haven.

Doyle, Michael W., Ikenberry (Eds.), 1997. New Thinking in International Relations Theory. Westview Press, Boulder, CO.

Duxbury, Neil, 1995. Patterns of American Jurisprudence. Cambridge University Press, Cambridge, UK.

Engerman, David C., 2004. Know Your Enemy. The Rise and Fall of America's Soviet Experts. Oxford University Press, New York.

European Science Foundation. Status and Developments of Social Science Research in Central and Eastern Europe. www.esf.org/file/admin/PublicDocuments/Publications/MOCEE-Oli.pdf (accessed 15.04.14).

Erickson, Paul, Klein, Judy L., Daston, Lorraine, Lemov, Rebecca, Sturm, Thomas, Gordon, Michael D., 2013. How Reason Almost Lost Its Mind. The Strange Career of Cold War Rationalism. University of Chicago Press, Chicago and London.

Ferguson, Yale H., Mansbach, Richard W., 2002. The Elusive Quest Continues. Theory and Global Politics. Prentice Hall, Saddle River, NJ.

Fleck, Christian, 2013. Lewis A. Coser – a stranger within more than one gate. Czech Sociological Review 49 (6). Prague.

Gaddis, John Lewis, 1971. The United States and the Origins of the Cold War 1941–1947. Columbia University Press, New York.

Gaddis, John Lewis, 1972. The United States and the Origins of the Cold War 1941–1947. Columbia University Press, New York.

Gaddis, John Lewis, 2002. Now We Know. Rethinking Cold War History. Oxford University Press, New York.

Gendzier, Irene L., 1985. Managing Political Change. Social Scientists in the Third World. Westview Press, Boulder and London.

Gilman, Nils, 2003. Mandarins of the Future. Modernization Theory in Cold War America. The Johns Hopkins University Press, Baltimore.

Gerovitch, Slava, 2002. From Newspeak to Cyberspeak. A History of Soviet Cybernetics. The Massachusetts Institute of Technology Press, Cambridge.

Graham, Loren R., 1967. The Soviet Academy of Sciences and the Communist Party, 1927–1932. Princeton University Press, Princeton, NJ.

Graham, Loren R., 1972. Science and Philosophy in the Soviet Union. Alfred A. Knopf, New York.

Graham, Loren R., 1987. Science, Philosophy, and Human Behavior in the Soviet Union. Columbia University Press, New York.

Graham, Loren R., 1993. Science in Russia and the Soviet Union. A Short History. Cambridge University Press, New York.

Habermas, Juergen, 1971[1968]. Knowledge and Human Interests. Beacon Press, Boston.

Haney, Paul David, 2008. The Americanization of Social Science. Intellectuals and Public Responsibility in the Postwar United States. Temple University Press, Philadelphia.

Issac, Joel, Bell, Duncan (Eds.), 2012. Uncertain Empire. American History and the Idea of the Cold War. Oxford University Press, New York.

Joravsky, David, 1961. Soviet Marxism and Natural Science 1917–1932. Routledge and Kegan Paul, London.

Joravsky, David, 1989. Russia Psychology. A Critical History. Basil Blackwell, Oxford, UK.

Kahn, Herman, 1960. On Thermonuclear War. Princeton University Press, Princeton, NJ.

Kent, Sherman, 1949. Strategic Intelligence for American World Policy. Princeton University Press, Princeton, NJ.

Kissinger, Henry A., 1957. Nuclear Weapons and Foreign Policy. Published for the Council on Foreign Relations by Harper and Brothers, New York.

Knutsen, Torbjorn L., 1997[1992]. A History of International Relations Theory, second ed. Manchester University Press, Manchester and New York.

Lee, Frederick, 2009. A History of Heterodox Economics. Challenging the Mainstream in the Twentieth Century. Routledge, New York.

Lubrano, Linda L., Solomon, Susan Gross (Eds.), 1980. The Social Context of Soviet Science. Westview Press, Boulder, CO.

Mazower, Mark, 2000[1998]. Dark Continent. Europe's Twentieth Century. Alfred A. Knopf, New York.

Mazower, Mark, 2012. Governing the World. The History of an Idea, 1815 to the Present. The Penguin Group, New York.

Medvitz, Thomas, 2012. Think Tanks in America. University of Chicago Press, Chicago.
Milne, David, 2008. America's Rasputin. Walt Rostow and the Vietnam War. Hill and Wang, New York.
Mirowski, Philip, 2002. Machine Dreams. Economics Becomes a Cyborg Science. Cambridge University Press, New York.
Morgan, Mary S., 2012. The World in the Model. How Economists Work and Think. Cambridge University Press, New York.
Ninkovitch, Frank, 2009. Global Dawn. The Cultural Foundations of American Internationalism 1865–1890. Harvard University Press, Cambridge.
Ninkovitch, Frank, 1994. Modernity and Power. A History of the Domino Theory in the Twentieth Century. University of Chicago Press, Chicago.
Ninkovitch, Frank, 1981. The Diplomacy of Ideas. U.S. Foreign Policy and Cultural Relations, 1938–1950. Cambridge University Press, Cambridge, UK.
Navarro, Juan Jose, 2011. Cold war in Latin America: the Camelot project 1963–1965 and the political and academic reactions of the Chilean left. Comparative Sociology 10, 807–825. Leiden.
O'Neill, William, 2005[1971]. Coming Apart. An Informal History of America in the 1960s. Ivan Dee, Chicago.
Porter, Theodore M., Ross, Dorothy (Eds.), 2003. The modern social sciences. The Cambridge History of Science, vol. 7. Cambridge University Press, New York.
Price, David H., 2008. Anthropological Intelligence. The Deployment and Neglect of American Anthropology in the Second World War. Duke University Press, Durham, NC.
Price, David H., 2011. Weaponizing Anthropology. Social Science in the Service of the Militarized State. Counterpunch, Petrolia, CA.
Rogers, Daniel T., 2012. Age of Fracture. Harvard University Press, Cambridge.
Rosen, George, 1985. Western Economists and Eastern Societies. Agents of Change in South Asia, 1950–1970. Oxford University Press, Dehli.
Rostow, Walt W., 1960. The Stages of Economic Growth. A Non-communist Manifesto. Cambridge University Press, Cambridge, UK.
Schriffin, Andre (Ed.), 1987. The Cold War and the University. Toward an Intellectual History of the Postwar Years. The New Press, New York.
Schulzinger, Robert D., 1984. Wise Men of Foreign Affairs: The History of the Council on Foreign Relations. Columbia University Press, New York.
Smith, Roger, 1997. The Human Sciences. W.W. Norton and Company, New York.
Solovey, Mark, 2013. Shaky Foundations. The Politics-Patronage-Social Science Nexus in Cold War America. Rutgers University Press, New Brunswick, NJ.
Solovey, Mark, Cravens, Hamilton (Eds.), 2012. Cold War Social Science. Knowledge Production, Liberal Democracy, and Human Nature. Palgrave Macmillan, New York.
Turner, Stephen Park, Turner, Jonathan H., 1990. The Impossible Science. An Institutional Analysis of American Sociology. Sage Library of Social Research 181. Sage Publications, Newbury Park; London; and New Delhi.
Vucinich, Alexander, 1984. Empire of Knowledge. The Academy of Sciences of the USSR 1917–1970. University of California Press, Berkeley and Los Angeles.
Weintraub, E. Roy, 2002. How Economics Became a Mathematical Science. Duke University Press, Durham, NC.
Weintraub, E. Roy, 1979. Microfoundations. The Compatibility of Microeconomics and Macroeconomics. Cambridge University Press, Cambridge, UK.

Cold War, The

Gustav Schmidt, Ruhr University Bochum, Herdecke, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term refers to the epoch between 1947–48 and 1989–90. The persistent central feature of the Cold War is the global contest between the United States (US) and the Soviet Union (USSR). The major changes in the structure of the contest are the transformation of the confrontation in Europe into stability, cold peace on the basis of the status quo since the mid-1950s, and the reversal in China's posture from principal ally of the USSR into expedient cooperation with the US. This article explains the longevity of the Cold War concept, the different perspectives on how the major powers ended and reignited the tensions, and what threat perceptions and resulting actions produced the structures of the East–West conflict.

The term is widely accepted in historical writings. It refers to the epoch between 1947–48 and 1989–90. The nature of the Cold War, its causes and effects, and the reasons for its long duration are, however, controversial. Its meaning would exclude Asia because of the outbreak of wars there; the wars in Korea, Vietnam, and Afghanistan made a difference to how the United States (US) and the Soviet Union (USSR), the two superpowers, waged their global contest in Europe. The situation in Europe resembled neither peace nor war; *East–West Conflict: Confrontation and Détente* expresses the ambiguity, but this label is less eye-catching.

The Concept and Its Salience

The persistent central features of the *Cold War* are the global contest between the US and the USSR, the dependence of the allies on the security guarantee of their respective superpower, and bipolarity. The latter was reinforced in the core area, the arms race, due to the widening gap between the extensive and highly diversified weapon systems of the superpowers and the arsenals of Britain, France, and the Peoples' Republic of China (PRCh), the other three established nuclear powers. This explains that American and Russian writings hold to *Cold War* as a concept fitting the entire epoch. It allows for distinguishing between moments of imminent war – Cuba, October 1962, and Yom Kippur War, October 1973 – times of high tensions and war outside Europe – Korea, 1950–53; Vietnam, 1964–75; Afghanistan, 1979–88 – long intervals of *détente* (Relaxation of tensions. Détente policy aims at making agreements that are trustworthy because both sides have incentives to keep them. The assumption is that 'security' could only be found through political means, and on a mutual basis) – highlighted by summit meetings and major arms control agreements – and even moments of concerted crisis management (June 1967 Near East war) or collaboration (during the Laos crisis and Vietnam War in the 1960s and ending wars in Africa in the late 1980s).

This version pays less attention to the fact that throughout the four decades not all parties to the Cold War were involved on the same issue-at-stake and at the same time. France joined Britain (UK) and the US in 1948 in founding the West German state as the Western allies' response to Stalin's anchoring of Poland, Bulgaria, Romania, Hungary, and Czechoslovakia firmly into the Soviet security-zone, and maintained staunch opposition to proposals for a neutralized, but nationally rearmed united Germany (Hitchcock, 1998). Since the mid-1960s, France became the spokesman for a Europe less dependent on the US and discussing terms of settlement with 'Russia.' The Federal Republic of Germany (FRG) converted *Cold War* into *cold peace* with the ratification of *Ost-und Deutschlandverträge* (1970–73). Both governments cooperated in defending *détente* in and for Europe; by implication, this aspiration for a European peace zone rejected the American concept of the indivisibility of the global contest between East and West. The PRCh, the only principal ally the USSR ever had (1950–58/59), made its peace with the US in 1972 on Pejing's term of 'One China.' While the PRCh had provoked the US during the 1950s to continuous nuclear saber-rattling, it turned its antihegemonic posture throughout the 1970s and 1980s against the USSR.

The Components of the Term

The definition refers to four aspects of the conflict: (1) the ideological antagonism between the Western concept of freedom of choice for the people of the domestic regime and external alignments of their state and the Soviet-imposed monopoly of the 'workers' and land labor party and subjugation to the community of socialist states; (2) the geostrategic struggle for bases of power-projection; (3) the domestic political contest, but almost exclusively within Western countries, about commitments to the military alliance or equidistance to both of the superpowers; (4) the dispute about where to draw the line between permissible and noncompatible elements in the policy mix of market-oriented and government-controlled economies. Of these four components, the last became least important for refueling the confrontation; the convergence of the mixed economies rather promoted the idea that *Osthandel*, credit facilities, and technology transfer might enhance not only 'liberalization' in the East European economies, but also generate transformation of the political system.

The Ideological Antagonism

The ideological antagonism was the first impulse for *confrontation*, but it was also subject to changing threat

 http://dx.doi.org/10.1016/B978-0-08-097086-8.96005-7

perceptions and shifts of emphasis in the balance between *confrontation* and *détente* in the overall relationship. The wider notion of *East–West Conflict* posits the Cold War as a distinctive period into the ideological struggle, originated in 1917–18, between the *Wilsonian Impulse* and Lenin's urge for peoples' democracy as the basis for securing peace (Link, 1980). From this perspective, Stalin and Truman likewise took the lead in splitting the world along 'alternative ways of life': Western definitions of *democracy* could not be reconciled with Soviet definitions of security (Lundestad); the latter implied that "whoever occupies a territory also imposes his own social system" (M. Djilas). Pursuing different policies, 'East' as well as 'West' nourished suspicions of the opposite's intention to extend its influence into the other's emerging bloc, the Soviets through encouraging Communist parties and the Americans via 'dollar imperialism.' Western leaders argued that the West had to learn its lessons from the vain hope to appease an expansionist regime. 'No more Munichs' informed the confrontation stance of *containment* as the West's answer to the external threat. (The term refers to the Munich Conference on 29–30 September 1938, where N. Chamberlain and E. Daladier conceded A. Hitler to proceed with the annexation of the Sudetenland and informed the Czechoslovakian government that it must either fight alone or abide by the Agreement. The notion 'Munich' conveys the message that dictatorships are unsatiable, that aggression feeds on weakness, that appeasement merely postpones the day of reckoning, etc.) The concomitant stance at the home front is based on the thesis that authoritarian regimes abuse state-power against large parts of their own population; therefore Communists must not be given a chance to occupy crucial positions of government such as Defense, Interior, Public Transport, or Justice.

Conversely, Communist ideology maintained that *capitalism* on the one hand inspires the state to prosecute the working-class and thus breeds *fascism*, and on the other hand is inherently expansionist (imperialist) and therefore knows no boundary to its domination. The belief, however, that capitalist nations are bound to fight for defeating rivals, made Stalin expect that Britain would resist the US' aspiration to become heir to the British Empire. The *new feature* of the post-World War II international system is that the US and the UK were competitive partners in founding the International Organizations destined to develop and monitor rules of conduct for international trade, currency exchange, and development aid (Gardner, 1969). The UK and France, but also Italy, expected the up-to-then reluctant US government to come to the rescue and infuse capital into the European economies as the only strategy which could render support to the parties willing to exclude Communists from governments (Lundestad, 1986). The formation of Western-oriented governing coalitions ensured that the European allies, as well as Japan, cooperated with the US in the evolution of the International Economy, to which the USSR and the PRCh were not negotiating parties after the outbreak of the Cold War. Most allies did not comply, however, with the American demand to restrict 'trading with the Communist enemy,' except in conspicuous moments of Soviet war activities.

The ideological component became a wasted asset after its overexploitation during the first phase (1947–53). Khrushchev's condemnation of Stalin's death toll in February 1956 and the NATO allies' perception that the Kremlin was unlikely to launch a general war on European territory (1956/57) put a break on the momentum of ideology as a driving force in the East–West conflict. But it was on the Kremlin to end the system conflict. In 1987, Gorbachev, determined to promote domestic reform and improve economic performance, rescinded the Breshnev doctrine and made it known that hard-liners in Eastern European governments should not reckon with Moscow calling on the Red Army to back up unpopular Communist regimes (Adomeit, 1998).

The Long Duration of the Cold War

For explaining the longevity of the Cold War, the account therefore has to focus on the other two factors: the structures which emerged with the consolidation and reaffirmation of the divisions of Europe and Germany, and the peculiarities of the military–geostrategic balance between the two sides.

The Divisions of Europe and Germany

The key feature is the hinge between the two *divisions*. The stark contrast between Stalin's refusal to consider concessions on Poland, Bulgaria, and Romania in exchange for American economic assistance and his demand to have a say on questions concerning primarily intra-West relations, e.g., the status of the Ruhr area or of Norway, provoked the US and the UK to tie their zones firmly to Western institutions (Deighton, 1990). The implementation of this basic strategy was, however, complicated by clashes of interest between the US, France, and Britain. These tensions prevented them from advancing as far as the USSR had with incorporating its German state, the German Democratic Republic (GDR), into the Soviet Empire. Hence they could not, as wanted, negotiate from a position of strength, but expected to be asked to give away on what 'the West' did not yet have, e.g., the FRG's defense contribution, whereas the USSR would persist in its refusal to put its reign over Eastern Europe on the agenda. The stalemate was compounded, when the reimposition of the Ulbricht regime in 1953 revealed that the Kremlin considered the GDR as its west-side lever for control over Poland and Czechoslovakia.

The structural impediment to a negotiated settlement disappeared when Moscow tacitly bowed to Germany's entry into NATO (1955) and when the 'West' acknowledged the fact that the USSR was strong enough to prevail in Eastern Europe (1956). German *Ostpolitik* put the final stamp to the 'normativity of the facts.' (Initiated by the Grand Coalition government under Kiesinger and Brandt and actively pursued by the SPD–FDP government under Brandt and Scheel, the FRG made an effort to end the special West German conflict with the USSR and the GDR within the overall East–West conflict. *Ostpolitik* was based on the assumption that the 'giant' to the East was an everlasting world power; besides, as the winner of the 'great patriotic war,' the USSR had to be accommodated. Bonn was prepared to acknowledge Moscow's quest for security 'from' Germany and to admit to the invulnerability of the post-World War II frontiers.) On this platform, the so-called Helsinki process, the Conference on

Security and Cooperation in Europe, lingered on throughout the 1973–89 period. Neither side was ready to proceed toward peacemaking.

The Cold War might have ended earlier if Gorbachev's predecessors had responded to the US view that Germany's double bind into a US-centered NATO and European integrated structures would prevent Germany from positing a threat to the security of the USSR (Schmidt, 1993–95). Instead, Moscow's attempts to push the divisiveness of nuclear diplomacy within NATO rather than wait for the outcome of such conflicts provoked the 'Atlanticists' to close ranks and reiterate the standard thesis that the West could prevent the USSR from winning the Cold War if its members resisted the temptation to court Moscow and make separate deals with the USSR. The Soviet leadership's 'stupidity' is said to have saved the US or NATO to get rearmament projects – such as NSC-68 or the 1979 dual-track decision – through.

The Peculiarities of the Military–Geostrategic Balance

The key feature is the fundamental asymmetry between the nature of America's and Russia's predominance in their respective sphere. This asymmetry is reinforced by the imbalance of military forces between NATO-Europe and the USSR's forward deployed forces (Kugler, 1993).

The advantage of the US in its global contest with the USSR was that all other principal powers, including Germany and Japan, were allied to the US. This helped in the buildup of the international economy, but not necessarily in defense. Consequently, the US diplomacy was absorbed as much in intra-West crisis management as in the context of East–West relations. The allies wanted to be assured that there were no long-term security risks involved for them in America's option for sponsoring the resurgence of the enemies and occupiers of World War II as strongholds of the West. Having imposed limits on Germany's and Japan's military status, the US could not reckon with a defense contribution for some years to come; both were prohibited to engage in 'out-of-area' defense activities (Schmidt and Doran, 1996).

By contrast, the USSR gained a formidable ally in the PRCh, who was keen to test the credibility of the American commitment to South Korea and Taiwan. Although the USSR copied the West's community building by setting up COMECON (1949) and the Warsaw Treaty Organization (1955), the Kremlin relied for exerting influence *de facto* on the penetration of the bloc partners' central agencies: the State Party, Secret Service, and top military echelon. The USSR also made sure through logistic measures that the Soviet military could operate from its allies' territory independently, and if need be without the consent of the incumbent governments (Wolfe, 1970).

The military imbalance between NATO-Europe and the USSR existed throughout 1947 until the mid-1980s. The US found no takers – except the FRG from 1961 onward – for its concept of balanced collective forces in the sense that the US were the sole provider of strategic forces, that is, weaponry projected to hitting targets located in the USSR, whereas the allies would recruit the ground forces and tactical air forces required for the defense of the region against infiltration or surprise attacks. Britain and France preferred to duplicate the US strategic role, albeit on a much minor scale, and took the relaxation of tensions in Europe and their entanglement in the legacies of their imperial rule as the rationale for reducing their force assignments to NATO.

The Intersection of Economic and Strategic Decisions: The European Structure

The failure of the 'Big Four' to cooperate on a German Peace Treaty is identified as a crucial turning point toward the Cold War (Leffler, 1992). Stalin did not want to give the impression that the US, thanks to the atomic bomb and its economic wealth, could impose its will on the USSR; he therefore pressed his claims on Iran, Turkey, during the Peace Conferences with Germany's wartime allies and in the Allied Control Council. This in turn was taken in Washington as a signal that hope of Russian cooperation must be abandoned. Truman and Byrnes had been reluctant to confront the USSR when the Cold War started over the fate of Eastern European nations (1944–47). Now, in view of the havoc, the 1946–47 crisis wreaked on the economies of Britain and France, the US government conceived that Western Europe could not save itself (Hogan, 1992).

From there on, the 'West' displayed a dynamic of its own. The first act was that the economic recovery of Europe was said to require the inclusion of (West) Germany; the price for getting France to change its German policy was the assurance that the US and the UK would back up the French quest for security against a resurgent Germany. The second act followed: Britain and France pleaded with Washington that the Marshall Plan initiative would not suffice to attain stability without an American security guarantee. Against the background of the Berlin blockade, Truman in October 1948 ordered that the US provide the major counterbalance to the 'ever-present threat of the Soviet military power.' In September 1948, the Brussels Treaty Organization (BTO) had already resolved that the defense of Western Europe should be as far to the east (of West Germany) as possible. Because this was beyond the means available to the BTO and not yet compatible with the military strategy of American and British defense planning, it was only a question of developments in the Cold War that the third issue be placed on the agenda: German rearmament. The logic behind the buildup of conventional forces sounded compelling: why do the old allies expect the US to provide ground and air forces and the Supreme Allied Commander to NATO's integrated forces with a view to realize the concept of forward defense when they deny the indispensibility of a sensible German defense contribution, on whose territory the French, in particular, wanted to stop the enemy's offensive? At that time, however, all the US military could offer was an air offensive to deter the USSR; they did not yet want to rely on 'the bomb' (Ross), and the USSR had acquired by then (Fall, 1949) an atomic capability. On 22 September 1949, the US-UK-Canadian governments jointly announced that an atomic explosion has taken place in the USSR.

The outbreak of the Korean War generated the fear of a similar war-by-proxy in divided Germany. The US did not only reverse its stance on what countries were of absolute importance to US security by now declaring Korea the test case

of what became the domino syndrome, that no US ally should fall prey to a Communist invader, but also designated the National Security Paper NSC-68 as the platform for a massive conventional rearmament of the US, adequate force deployment abroad, and military aid to upgrade the defense capabilities of its European and Asian allies (May, 1993).

The issue of German rearmament was divisive, but the US compromised with France on the understanding that German forces be integrated into a European Defense Community (EDC), whereas the EDC would delegate strategic planning and fixing force requirements to NATO. The project became the victim of French domestic politics; the US and the UK had to comply with the French request that no German soldier be officially recruited until France had ratified the EDC and German Contractual Agreements. The demise of the EDC meant that Washington had to forego the hope that US force deployment would be a temporary stopgap until the European allies, and especially the FRG, were able to do more toward meeting NATO's minimum force requirements.

The parallel staged fourth act (1952–54) produced the *definitive structure of the Cold War in Europe*: the NATO, to which the FRG was finally admitted, had made itself dependent on the credibility of the US (and the UK) strategic deterrent. This provoked the issue how the USSR would react to the logic that NATO's new member would have to be equipped with the same weaponry as the allied forces into which the German divisions were to be incorporated. Chancellor Adenauer made it clear that the FRG did not want to provoke Moscow unnecessarily by dislocating medium-range ballistic missiles on German soil, whose later replacements might be able to hit targets in Russia, but he and his successors insisted on having at least a crucial German vote in all decisions on the use of nonstrategic weapons in NATO's custody in case deterrence to all types of war failed (Heuser, 1998; Trachtenberg, 1999).

This final act of incorporating the two German states into alliances dominated by the nuclear-strategic superpowers had a far-reaching effect: plastering the 'frontline' states with short-range atomic weapons was the best insurance premium that the Germans would not launch war from their soil (Ullman, 1991). As long as the US and the USSR stationed troops there and resolved to maintain control over the warheads, the danger of accidental war could be excluded. The parallel to the military factor written into the strategic landscape is the interest of the Western nuclear powers in confirming the status quo and discuss on that territorial basis the questions of putting ceilings on rearmament. Separately, the two German states relaunched their rivalry and demanded (1955–73) allies and Third World countries alike to subscribe to the *Alleinvertretungsanspruch* of the Western democratic (FRG) or peoples' GDR. (Due to the resumption of diplomatic relations with the USSR (in 1955) as the last of the four powers responsible for Germany and Berlin as a whole, the (West) German Foreign Office (W. Grewe, W. Hallstein) considered it necessary to stipulate that the recognition of the GDR as an independent state (rather than the Soviet occupied zone) by states other than the USSR and the members of the Warsaw Pact must be taken by the FRG as an unfriendly act; undermining West Germany's pursuit of reunification. Subsequently the FRG would be forced to interrupt diplomatic relations and suspend foreign economic assistance. Conversely, the GDR urged members of the Warsaw Pact to support its claim for being the second German state. With a few, but crucial exceptions (Yugoslavia), the FRG succeeded in stalling the formal recognition of the GDR until the conclusion of the '*Deutschlandverträge*' and the admission of the two Germanys into the United Nations in 1973.)

The History of the Concept, Major Developments, and Empirical Results

The invention of the term is accredited to Walter Lippmann who took issue with G.F. Kennan's 'X'-article 'Sources of Soviet Conduct.' Kennan's long telegram of 22 February 1946 molded the agenda of America's containment policy. His arguments about Stalin's hostility of the West and unapproachability to conventional diplomacy were selectively used by top officials in Washington to assert that the 'Soviets' will develop all means and methods ranging from threat of military aggression via propaganda warfare to clandestine activities to a degree without precedent in history. Therefore, the US had to strengthen its executive branch, introduce a national security policy, establish a professional intelligence agency, and develop an unassailable military–industrial base (Leffler, 1992; Paterson, 1979).

The opposite Communist view emphasizes three aspects: the economic aggressiveness of the long-standing hegemonic US project to attain a 'one-world market' and abuse of the open-door doctrine for penetrating other nations' economy; the manipulation of anticommunism as a club to suppress claims for social justice and equality of rights within the American and other Western capitalist systems; and America's acquisition of bases bordering on the USSR for purposes of encirclement and monitoring inside the USSR. *In geostrategic terms the US became a direct neighbor of the USSR*, whereas the USSR required long-distance bombers (it introduced such jets by 1955) and intercontinental ballistic missiles to show the US its vulnerability. The strong engagement of the US in the post-1990 contests about Caucasian and Middle East oil and gas concessions and pipeline routes indicates that the US–USSR rivalry did not end with the Cold War.

The 'Revisionists' (LaFeber, 1991; Paterson, 1979) disclaim that Stalin pursued an offensive strategy aiming at extending Soviet rule; they discern rather the US' responsibility for causing the Cold War. Truman and his advisers (Acheson, Harriman, Lovett, McCloy, Nitze), intending to contain the self-opinionated coalition of leftists and isolationists at home and to exploit the economic preponderance abroad, is attributed to launch a military–industrial complex and a national security state in order to perpetuate American hegemony; this view is said to lead the US to march toward world power. The postrevisionist approach accounts for the Soviet aspect of the fundamental change in the Cold War since about 1954–55, and explains the US interest in *détente*, i.e., the coinciding interest in *Long Peace* (Gaddis, 1987). Moscow settled the Austrian question, tolerated the FRG's entry into NATO, and opted for reconciliation with Tito, the only East European leader who had got away with breaking ranks with Stalin. The cumulation point was the declaration of the doctrine of *peaceful coexistence*. The Western allies' match was tacit satisfaction with consolidating the status quo after they had absorbed the FRG.

Stability in Europe turned the UK's, US's, and France's attention to the Third World. In reaction to Communist China's revolutionary foreign policy, the Soviet leadership in 1960 declared wars of liberation from colonial rule legitimate and thus distinguished peaceful coexistence in the developed world from just wars in the 'southern' part of the globe. In this sense, the East–West conflict was exported to the Third World.

The US resolved on noninterference in the internal affairs of the Soviet bloc and relaxation of tensions. The US was somewhat dependent on Russia's resistance to become the air and nuclear strategic arm of Pejing's violent anti-American activities in the Far East. The emerging picture is that the US, in parallel with the Berlin and Cuba crises (1958–62), wanted to concert the superpowers' activities in China's *'hinterland'* (former 'Indochina') in the sense that both exert pressure on the parties to a conflict amenable to their respective influence and through such agreement attain the neutralization of the conflict area, including the installment of all-party or power-sharing coalition governments, however unstable (Nelson, 1995).

European Perspectives on the Cold War

A different strand in the history of the concept is the perception of the impact of the other Western powers, especially Britain and France, on how the struggle was waged (Greenwood, 2000; Bozo, 1991). The contribution is twofold: (1) the *Cold War* is viewed as a new stage in great power rivalry. These studies stress the Europeans' interest in devoting resources to restoring or preserving their assets and commitments overseas (Kent, 1993; Bossuat, 1992). (2) The second contribution takes a different direction: it challenges the assumption that Western Europe by 1947–48 was in such a critical state that the leading nations had to 'invite' the US to reconstruct and protect Western Europe. Examining the economic potential of the European nations and how their governments used the financial means provided under the Marshall Plan, Milward (1984) presents the thesis of the rescue of the nation-state. His complex argument also draws on the observation that at that time the West did not reckon with an immediate Soviet threat. The implication is that an unwarranted haste was imposed on a process which in any case depended on what the European countries did to and for another, formally, US assistance was tied to the Uniting of (Western) Europe and to parallel bilateral agreements between the US and the recipient country.

Competitive Cooperation between the Superpowers

By the end of the 1960s, researchers discerned the convergence of five developments that demonstrated the continuing relevance of the 'competitive cooperation' between the superpowers for the changing structures of the *Cold War*:

1. Soviet policy shifted toward accepting the US's and Canada's presence in Germany
2. German *Ostpolitik* presumed the engagement of the US in NATO
3. de Gaulle realized in the context of the Prague, 1968 crisis – 'Prague' refers to the invasion of Warsaw Pact troops into Czechoslovakia on 20–21 August 1968 and the arrest of party leader Dubcek, who initiated domestic reforms ("Prague spring" – "Prager Frühling") – that Breshnev's Russia was no party to his vision that dissolving Cold War structures in the West, i.e., France's disengagement from NATO's military organization, might induce 'Russia' to allow for more evolution of national Communist and then independent states in Eastern Europe
4. The US embroilment in the Vietnam War caused a change in the intellectual climate; the abuse of the USSR being the cause of every evil became obsolete; instead the US became the villain in the piece
5. New developments in arms technology – e.g., high-precision weaponry as a substitute for atomic weapons, MIRV-technique (multiple independently targetable reentry vehicles), antiballistic missiles – were beyond Britain's and France's capability to follow on; this reinforced NATO Europe's self-elected dependency on the US with respect to security and defense (Hanrieder, 1989).

Renewed US pressures on its allies to extend their conventional military capabilities and share the burdens for the 'Defense of the West' more evenly, the reescalation of the strategic arms race between the superpowers under the aegis of limitation treaties, and the evidence presented by German *Ostpolitik* that negotiations with Russia generate tolerable results, combined to raise the basic questions: What are the costs of the ongoing *Cold War*? On what terms could the conflict be ended and converted into politicoeconomic competition? (Garthoff, 1985).

From the Second Cold War to the End of the Cold War

The USSR, after getting Germany's pledge to observe the invulnerability of the territorial status quo in Europe, revived the global contest with the US, which was immersed in domestic turbulences *and* disputes with all its allies about oil and the dollar. The Kremlin demonstrated its self-confidence by expanding Soviet naval forces and establishing bases, however short-lived, stretching from Vietnam via Mozambique, Somalia, and Angola to Central America. In contrast to the 1950–53 period, the US catchword 'arch of crises' did not resonate well with its allies. The latter wanted to develop *détente*, notwithstanding the fear, expressed by German Chancellor Helmut Schmidt in 1977, that the Soviets' new equipment (Backfire bombers; SS-20) made NATO-Europe hostage to the Kremlin's whip of the will. This raised the question whether the USSR was about to win the Cold War or whether the Kremlin was overstretching the country's resources and would thus expose the USSR sooner rather than later to the need for radical changes in her system. The wishful thinking that the USSR would fall victim to its inability of continuous adaptation worked out after three decades of aspiring military strategic and political parity with the US. Some authors argue that Reagan's defense buildup deliberately forced the Soviet leadership to acknowledge the failure of its domestic system and hence the USSR's inability to persist in the contest and sustain bipolarity (Adomeit, 1998; Gaddis, 1987; Wells' article in Hogan, 1992). The way the global contest ended invited Americans to believe that they had won the Cold War. By contrast, Garthoff and Lebov/Stein argue that only a Soviet leader could have ended the Cold War, "and Gorbachev set out

deliberately to do so." His initiatives in dismantling the Soviet command (military–industrial) economy and emphasizing 'common security' jettisoned the USSR's Cold War apparatus.

Methodological Problems

The immersion of the concept *Cold War* in the perpetual clash of interests between East and West and within each 'bloc' subjects the researcher on the one hand to the political climate of his own times, and may thus reload or de-emphasize the contentions of the past; on the other hand, access to newly available records reveals new insights into previous phases of the Cold War and thus demands rethinking the past but in a different way. Getting the balance right between these two operations is difficult when knowledge about one party to the conflict is the result of generations of research, whereas the more recent presentations of findings from Soviet, GDR, Czech, or Polish sources are both less systematic or structured and more exposed to be taken instantly as evidence for one interpretation, e.g., of Stalin's or Krushchev's reasonableness, or another, appraising both their internal and external actions and rhetoric. The task to study the records comprehensively, but also be prepared to modify one's assessments in response to newly available empirical evidence is prone to collide with the other obligation, namely to explain what were the basic causes of a conflict and which causes – meaning in politics: sins of commission and omission – are accountable for what developments.

A second main problem is the exact phase-by-phase and overall intersection between the interpretation of the collaboration leading to the evolution of 'Western' structures in the international economy and the assessment of the concomitant events in the politicostrategic (military) sphere, where the USSR appeared to be the winner, but then ended the Cold War on Western terms.

Future Direction of Research

'East' and 'West' waged the struggle by all means and methods except 'hot' war in the area stretching from Vancouver to Wladivostok. Hence, research must attend to different subject areas and to the expertise developed in the many scholarly communities. Some areas of research, e.g., diplomatic or intellectual history and biography, are more established than others. The impact of *intelligence* on policymakers is a relatively new area of systematic research. International cooperation projects have done much to promote *nuclear history* and *strategic studies*, but the history of the military alliances and of national defense organizations and policies depend still on the authorities' grant of access as well as clearance and permission for publishing inspected material. The studied ambiguity of the political leaderships and the top military echelon about the worth of strategic nuclear weapons in case deterrence should fail and about the use of short-range atomic weapons deserves further study in order to know the implications of nuclear weaponry on the conduct of Cold War diplomacy and governmental guidance to their military.

Future research should be more systematic in the sense of *extending* the conceptualization beyond the national and bilateral focus to the regional context and *intensifying* the approach by addressing the fundamental questions about the changing nature of the struggle, its costs, the persistence or recreation of patterns of conflict, and above all the question whether the Cold War structure affected all other bilateral or intraregional conflicts or whether reemerging older conflicts pervaded the East–West conflict, so that the parties to such conflicts used the *Cold War* for the purpose of engaging wealthy allies on their side.

See also: Arms Control; Communism, History of; Contemporary History; Eastern European Studies: History; Military, War, and Politics; National Security Studies and War Potential of Nations; Revolutions of 1989–90 in Eastern Central Europe; Russian and Post-Soviet Studies: Politics; Second World War, The; War, Sociology of; Warfare in History.

Bibliography

Adomeit, H., 1998. Imperial Overstretch: Germany in Soviet Policy from Stalin to Gorbachev. Nomos Verlagsgesellschaft, Baden-Baden, Germany.
Bossuat, G., 1992. La France, l'aide américaine et la construction européenne, 1944–1954. Impr. nationale, Paris.
Bozo, F., 1991. La France and l'OTAN. De la guerre froide au nouvel ordre européen. Masson, Paris.
Brands, H.W., 1993. The Devil We Knew. Americans and the Cold War. Oxford University Press, New York.
Cold War History, 1999. Routledge, London.
Deighton, A., 1990. The Impossible Peace: Britain, the Division of Germany and the Origins of the Cold War. Clarendon Press, Oxford, UK.
Fleron, F., Hoffmann, E.P., Laird, R.F. (Eds.), 1991. Soviet Foreign Policy: Classic and Contemporary Issues. Transition Publishers, New York.
Gaddis, J.L., 1987. The Long Peace: Inquiries into the History of the Cold War. Oxford University Press, New York.
Gaddis, J.L., 1997. We Know Now: Rethinking Cold War History. Oxford University Press, New York.
Gardner, R.N., 1969. Sterling–Dollar Diplomacy. The Origins and the Prospects of Our International Economic Order (New expanded ed.). McGraw-Hill, New York.
Garthoff, R.L., 1985. Détente and Confrontation: American–soviet Relations from Nixon to Reagan. Brookings Institution, Washington, DC.
Gori, F., Pons, S. (Eds.), 1993. The Soviet Union and Europe in the Cold War, 1945–1953. St. Martin's Press, London.
Greenwood, S., 2000. Britain and the Cold War 1945–91. Macmillan, Houndsmill, UK.
Hanrieder, W.F., 1989. Germany, America, Europe: Forty Years of German Foreign Policy. Yale University Press, New Haven, CT.
Heuser, B., 1998. NATO, Britain, France and the FRG. Nuclear Strategies and Forces in Europe, 1949–2000. Macmillan, Basingstoke, UK.
Hitchcock, W.I., 1998. France Restored. Cold War Diplomacy and the Quest for Leadership in Europe, 1944–1954. University of North Carolina Press, Chapel Hill, NC.
Hogan, M. (Ed.), 1992. The End of the Cold War. Cambridge University Press, Cambridge, UK.
Journal of Cold War Studies, 1999. MIT Press.
Kent, J., 1993. British Imperial Strategy and the Origins of the Cold War, 1944–1949. Leicester University Press, Leicester, UK.
Kugler, R.L., 1993. Commitment to Purpose. How Alliance Partnership Won the Cold War. RAND, Santa Monica, CA.
LaFeber, W., 1991. America, Russia, and the Cold War, 1945–1990, sixth ed. McGraw-Hill, New York.
Lebov, R.N., Stein, J.G., 1994. We All Lost the Cold War. Princeton University Press, Princeton, NJ.
Leffler, M.P., 1992. A Preponderance of Power: National Security, the Truman Administration, and the Cold War. Stanford University Press, Stanford, CA.
Link, W., 1980. Der Ost-West-Konflikt. Die Organisation der internationalen Beziehungen im 20. Jahrhundert. Kohlhammer, Stuttgart, Germany.

Lundestad, G., 1986. East, West, North, South: Major Developments in International Politics, 1945–1986. Oxford University Press, New York.
May, E.R. (Ed.), 1993. American Cold War Strategy: Interpreting NSC 68. Bedford Books, Boston.
Milward, A.S., 1984. The Reconstruction of Western Europe, 1945–1951. Methuen, London.
Nelson, K.L., 1995. The Making of Détente: Soviet–American Relations in the Shadow of Vietnam. Johns Hopkins University Press, Baltimore, MD.
Paterson, T.G., 1979. On Every Front: The Making of the Cold War. Norton, New York.
Schmidt, G. (Ed.), 1993–1995. Ost-West-Beziehungen: Konfrontation und Détente 1945–1989, 3 vols. Brockmeyer, Bochum, Germany.
Schmidt, G., Doran, C.F. (Eds.), 1996. Amerikas Option für Deutschland und Japan. Die Position und Rolle Deutschlands und Japans in regionalen und internationalen Strukturen. Die 1950er und die 1990er Jahre im Vergleich. Brockmeyer, Bochum, Germany.
Trachtenberg, M., 1999. A Constructed Peace. The Making of the European Settlement 1945–1963. Princeton University Press, Princeton, NJ.
Ullman, R.H., 1991. Securing Europe. Adamantine Press, Twickenham, UK.
Wolfe, T.W., 1970. Soviet Power and Europe, 1945–1970. Johns Hopkins University Press, Baltimore, MD.
Zubok, V., Plechanov, C., 1996. Inside the Kremlin's Cold War: From Stalin to Krushchev. Harvard University Press, Cambridge, Mass.

Coleman, James Samuel (1926–95)

Aage B Sorensen, Harvard University, Cambridge, MA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 4, pp. 2200–2204, © 2001, Elsevier Ltd.

Abstract

James S. Coleman, born on 12 May 1926 in Bedford, Indiana, was among the most important American sociologists of his generation. Coleman was unique among sociologists by importantly influencing both intellectual and policy debate, especially educational policy. He made lasting contributions to several areas of scholarly concern, including our understanding of school effects on learning, the development of rational choice theory in sociology, and the use of mathematics as a tool for the development of theory in sociology. James S. Coleman died on 25 March 1995 in Chicago, Illinois.

James S. Coleman was born on 12 May 1926 in Bedford, Indiana. He died on 25 March 1995 in Chicago, Illinois. Coleman was among the most important American sociologists of his generation. By importantly influencing both intellectual and policy debate, Coleman was unique among sociologists. By making important contributions to a large range of areas of scholarly concern, including mathematical sociology, sociology of education, and social theory, Coleman was unique among social scientists.

Life

James Samuel Coleman was born into a family of teachers, with roots in the landed gentry of the South on his father's side, and in various places in the MidWest on his mother's side. His grandfather, Samuel, was a minister. The family moved frequently in Coleman's youth between places in Ohio, Arkansas, and Kentucky to settle, finally, in Louisville, Kentucky, where Coleman graduated from Manual High School in 1941. The origin is important for his basic sociological interests and positions. The marginality from moving around between southern and northern American cultures and the diversity of his origins instilled a strong curiosity about social relations. The teacher occupations of both parents created a strong interest in schools and education. Samuel, the minister grandfather, and the mother were important for Coleman's preoccupation with moral issues that profoundly influenced his work.

Coleman's choice of sociology as a vocation came quite late. He graduated from Purdue University in 1949, with a degree in Chemical Engineering, and his first job was as a chemist with Eastman Kodak. He had almost no undergraduate education in any social science. Nonetheless, in 1951 he began graduate study in sociology at Columbia University. Coleman's dual attraction to science and moral engagement made sociology an impeccable choice, or so it would seem in 1951. He had found industry frustrating and a likely eventual career in management unappealing. He wanted to devote his life to discovery and concluded it could only be about people, their relationships, and their social organization.

Columbia's sociology department gave Coleman four intense years and three important teachers: Paul Lazarsfeld, Robert Merton, and Seymour Martin Lipset. Coleman is usually regarded as Paul Lazarsfeld's student. This is not quite correct. Lazarsfeld was not his dissertation advisor; it was Lipset. Lazarsfeld was not the teacher who had the most profound intellectual influence on Coleman; it was Merton. Lazarsfeld did involve and use Coleman for the development of mathematical and statistical tools for social analysis, and these activities created the point of departure for some of Coleman's most important later work. However, there is an important difference between Coleman's *Introduction to Mathematical Sociology* (1964) and Lazarsfeld's branch of mathematical thinking in the social sciences. Coleman's main objective with the use of mathematics was the development of theoretical insights and conceptual development. Lazarsfeld's major contributions were to the codification of research procedures, that is, methodology. Coleman has made important contributions to methods, but his most remarkable quality as a sociologist was his ability to develop sociological ideas and sustain them with empirical evidence. This is much closer to Merton's style of theorizing about empirical matters, though Merton often relied on evidence produced by others.

Coleman obtained his first faculty position in 1956 as an Assistant Professor at the University of Chicago. In 1959 he moved to Johns Hopkins University to create his own sociology department. He developed a small organization with an intellectual intensity and excitement that was truly remarkable. It was perhaps unsustainable and, for a variety of reasons, Coleman went back to the University of Chicago in 1973, now as a University Professor, and stayed there until his death in 1995.

Scholarship

The body of scholarly work is large. It includes 30 books and monographs and more than 300 articles and chapters in books. These contributions have profoundly influenced and, in several cases, defined the agenda for important subfields of sociology: sociological theory, sociology of education, sociology of the family, mass-communication, social stratification, political sociology, mathematical sociology, and sociological methods. The contributions include three major pieces of policy-oriented research on schools and educational opportunities that are the major examples of 'sociology making a difference' in the 1960s, 1970s, and 1980s. The scholarly work covers a phenomenal range of approaches to social research. There is work about social systems (Coleman, 1961) and about individual behavior (Coleman, 1990b). There is basic research as well as applied. There is quantitative (Coleman et al., 1966;

Coleman, 1987) as well as qualitative analysis (Coleman, 1961). There are contributions to economics (Coleman, 1966), political theory (Coleman, 1982), moral philosophy (Coleman, 1974), statistics and probability theory (Coleman, 1981), and education (Coleman, 1990a).

It is difficult to characterize James Coleman's contributions to the discipline because there are so many components. It includes two major and two very different paradigms of what sociology is about. One is a Durkheimian paradigm: it sees the task as studying how social structure creates and constrains individual action and causal social processes. The other paradigm takes departure in individual purposeful action for developing properties of social systems and structures. The former project governs most of Coleman's empirical work on educational processes and social processes in educational institutions. The latter project, the development of a sociological, rational action theory, moves, so to speak, in the opposite direction. It is theory aimed at understanding social systems themselves, their development, and properties, beginning with a theory of individual action. It occupied most of Coleman's attention in the last years of his life.

Almost all of Coleman's empirical research was about growing up, about schools, and learning and educational opportunity. His first study of schools was a study of adolescent subcultures based on surveys of students in ten high schools in northern Illinois. The report on this research was published in 1961 as *The Adolescent Society* (1961). In the preface, Coleman notes two reasons for undertaking this study. One was his wish to study how schools might be more effective; he states that the study had its origin in 'a deep concern I have had, since my own high school days, with high schools and with... ways to make an adolescent experience with learning more profitable...' (1961, p. viii). The second concern was an interest in different types of social systems and the value systems they reflect. These two interests, in creating effective schools and other social organizations and in the values and cultures of social systems, would guide Coleman's research throughout his life.

The wish to study the effects of high schools on learning did not receive much attention in *The Adolescent Society*, possibly because the empirical evidence Coleman obtained about these effects, and their relationship to status systems, is quite difficult to interpret. *The Adolescent Society* is primarily a study of adolescent subcultures. The policy prescriptions Coleman drew from his analysis were about how to change these subcultures to encourage learning and academic achievement. This would be done by changing the structure of competition so that students would see academic achievement as a collective good, like sports.

The interest in the effects of schools and school social organizations on learning became the main preoccupation of Coleman's educational research for almost four decades. The link between the two concerns was always clear, and the interest in the social systems created in schools re-emerges as the dominant concern again in his latest research. All of his educational research sees the effectiveness of schools related to the social systems created in schools and in the family and community that form the context for the school social systems. Coleman was very much a sociologist of schools.

Coleman performed three major studies of schools and educational processes after *The Adolescent Society*. The first so-called 'Coleman Report' (for a while referred to by many as the Coleman – Campbell report) was a report on the massive *Equality of Educational Opportunity* survey (Coleman et al., 1966). It is probably the largest social science survey research ever undertaken with some 600 000 students surveyed. The Report's main finding was that the instructional facilities and resources, including per pupil expenditures, have small and comparatively insignificant effects on academic achievement relative to the influence of family background. This result created considerable controversy, a controversy that continues today. A very large number of studies have focused on the issue since then. The best of this research has replicated the main conclusion: there is very little relationship between the amount of expenditure on schools and achievement outcomes, net of the family backgrounds of students. Next to family background, the family background of a student's peers, as measured for example by the racial composition of schools, seems most relevant, especially for minorities. This conclusion had an important consequence for American schools as it was used to justify racial desegregation in public schools. The second Coleman Report focused on this effect on policy: the use of busing to integrate schools and reap some of the benefits of desegregation on the academic achievement of minorities. The Report concluded that the policy had been ineffective. This conclusion was widely interpreted as denying the benefits of desegregation documented in the First Report and was strongly attacked by many, including many of Coleman's fellow sociologists. The interpretation was not correct. Coleman did not doubt the relevance of social systems created by student body composition for schools' educational climates. He doubted the long-term benefits of busing as a remedy for segregation because it encouraged white flight from desegregated public schools.

The allegation that Coleman had reversed his position would be repeated with the publications from the third major piece of educational research published in the early 1980s. In a series of publications, for example *Public and Private High Schools* (Coleman and Hoffer, 1987), Coleman and co-authors concluded that Roman Catholic parochial schools produced more learning and superior educational outcomes than public schools because of superior social resources, or social capital created by the involvement of parents with schools and their personnel. These conclusions resulted in another set of published controversies, now about the existence of school effects – the very effects that were said to have been denied by Coleman in the first Report. Now, the tables had been reversed. The critiques now denied the existence of school effects, in particular the apparent superiority of Catholic schools.

Coleman strongly believed in the use of mathematics as a language for the development of social theory. His *Introduction to Mathematical Sociology* (1964) remains unsurpassed as a text on how to use mathematics, in particular stochastic process models, to mirror social processes of diffusion and influence in groups. It is an enormously rich collection of tools for the analysis of social processes. Coleman remained committed to the use of mathematics for the development of theory. Over time, he changed the type of theory focused upon, from modeling the mechanism of social processes in his early work, to modeling the outcomes of purposive action in his later work. Throughout, Coleman's contributions provide perhaps the most important demonstration of the power of mathematics as a tool for social theory in contemporary sociology.

Already in the 1960s, Coleman published several articles on a model for collective decisions where actors trade control over events and resources to maximize their interests. This work came about as a result of Coleman's interest in developing games for learning and for social theory, where players engage in transactions designed to mirror important learning tasks or to simulate basic rules of social systems. One ofthese games was called the legislative game, and was designed to simulate the transactions going on in a legislative body with vote trading. Coleman formalized this game in a simple model that provided equilibrium outcomes providing measures of the power of actors and the value of events. These ideas became the basis for what became his major preoccupation in the 1970s and 1980s: creating a rational action basis for social theory where the emergence of macro social structures are clearly linked to micro foundations.

The outstanding result of many years of effort became *Foundations of Social Theory* (1990). It is a major book in ambition, achievement, and size. It provides theory and theoretical tools for the analysis of existing society and the creation of better societies. The book moves from the analysis of elementary action and social relations of trust and authority to the development of norms and more complex authority systems or corporate actors, such as large corporations, and then on to the analysis of the major institutions of modern society. A major concern for Coleman was the distribution of power in society and the problem of holding corporate actors responsible for their actions. *Foundations of Social Theory* is an unfamiliar contribution to a discipline where theory often seems to be an exegeses about the theory created by those who are safely dead, or abstract conceptual development with little if any empirical reference. This major book aims to shape the discipline by providing a theory and a mathematical structure that may have extraordinary potential for research. The realization of this potential depends not only on the quality of the ideas, but also on the discipline's ability to retool. Some concepts from the book have become prominent in sociology, for example Coleman's ideas about the micro – macro link and about social capital. However, this is only a part of this enormously rich contribution. Coleman knew that an extraordinary educational effort was needed, and he devoted much of his attention to this effort in his last years, for example, by founding and editing the journal *Rationality and Society*.

Coleman's theoretical work created a landmark in social theory. His concern for developing a mathematical language for the analysis of empirical processes created a rich set of tools that eventually may change the way sociologists work. His empirical work often includes imaginative interpretation and brilliant insights. Though his emphasis increasingly was on the project that resulted in *Foundations of Social Theory*, Coleman repeatedly returned to empirical research on causal processes with individual actions and lives as the outcome. The synthesis was accomplished in some of Coleman's latest empirical work, on schools, family, and community.

Impact

There was no lack of recognition of Coleman's scholarly contributions by the bodies that confer the highest prestige upon scientists. Coleman was elected to the American Academy of Arts and Sciences in 1966, the National Academy of Education in 1966, the American Philosophical Society in 1970, the National Academy of Sciences in 1972, and the Royal Swedish Academy of Sciences in 1984. He was a Fellow at the Center for Advanced Study in the Behavioral Sciences, a Guggenheim Fellow, and Fellow at the Wissenschaftskolleg zu Berlin. He received numerous honorary degrees from universities in the USA and Europe.

Despite all the recognition, American sociology had, and still has, considerable ambivalence toward Coleman. His international reputation, particularly in Germany, the Netherlands, and Scandinavia, and his reputation outside of sociology seemed often more solid than his reputation in American sociology. He became President of the American Sociological Association (ASA) from 1991 to 1992, but he never held another elected office in the ASA. His election to the Presidency in 1990 was the result of a write-in campaign and not of a nomination by the Association. Indeed, the leadership of the ASA, including the ASA President, tried to censor him in the mid-1970s for producing subversive sociology, because his work on the impact of busing on white flight threatened traditional liberal beliefs. It was an ignominious act that almost succeeded.

The ambivalence of the profession toward Coleman had two main sources. One was Coleman's unwillingness to specialize in one of the usual three main roles for sociologists: as theorist, methodologist, or researcher. This contradicts the implicit theory most of us have that one cannot be outstanding in all three roles. Coleman was outstanding in all three, but specialists in theory, methods, or research had difficulties recognizing his achievements. The second, and probably major, source of ambivalence was Coleman's use of research to draw policy inferences. Coleman stated what his research meant for policy and was not hesitant to do so when it contradicted conventional wisdom. He loved controversy. Each of the three Coleman Reports stated a conclusion that infuriated many: that school resources have little impact on academic achievement compared to the family resources of a child; that busing to achieve racial integration speeds up the process of white flight from our central cities; that schools organized as many private Roman Catholic schools produce more learning and less inequality in learning than schools organized like the typical public school. In every instance, an army of researchers tried to find fault with the evidence for these conclusions and largely failed. They concentrated on specific statistical issues. This was a mistake when confronting Jim Coleman. He anticipated criticisms by demonstrating the main finding in several ways. Moreover, his powerful intuitions about the mechanisms that create observed outcomes produced coherent arguments that Coleman effectively integrated with empirical analysis, while the opposition failed to formulate an alternative argument that could be sustained with evidence.

Sociology is profoundly influenced by James Coleman. So are many individual lives and careers. None of his students and followers moved with him through all the projects, but remain in one or the other of his many areas of interests. Those who had the privilege to know him were always profoundly touched by his excitement of translating into empirical analysis theoretical ideas about how social structure affects individuals. Coleman's enormous mental (and physical) energy never ceased to amaze. This energy and creativity were sustained by

his certitude about the importance of the project of making sociology a better tool for a better society.

Appraisals of Coleman's scholarly and policy contributions may be found in Sørensen and Spilerman (1993) and Clark (1996). The latter volume, published after Coleman's death, includes a complete bibliography and several autobiographical fragments.

See also: Adolescence, Sociology of; Affirmative Action: Empirical Work on Its Effectiveness; Educational Sociology; Family and Schooling; Family as Institution; Public Policy Schools; Racial Relations; Rational Choice Theory in Sociology; Sociological Theory; Sociology, History of.

Bibliography

Clark, J. (Ed.), 1996. James S. Coleman. Falmer Press, London.
Coleman, J.S., 1961. The Adolescent Society. Free Press, Glencoe, IL.
Coleman, J.S., 1964. Introduction to Mathematical Sociology. Free Press, Glencoe, IL.
Coleman, J.S., 1966. The possibility of social welfare function. American Economic Review 56 (5), 1105–1112.
Coleman, J.S., 1974. Inequality, sociology, and moral philosophy. American Journal of Sociology 80, 739–764.
Coleman, J.S., 1981. Longitudinal Data Analysis. Basic Books, New York.
Coleman, J.S., 1982. Re-contracting, trustworthiness, and the stability of vote exchanges. Public Choice 40, 89–94.
Coleman, J.S., 1990a. Equality and Achievement in Education. Westview Press, Boulder, CO.
Coleman, J.S., 1990b. Foundations of Social Theory. Belknap Press, Cambridge, MA.
Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A., Weinfeld, F.D., York, R.L., 1966. Equality of Educational Opportunity. Government Printing Office, Washington, DC.
Coleman, J.S., Hoffer, T., 1987. Public and Private High Schools: The Impact on Communities. Basic Books, New York.
Sørensen, A.B., Spilerman, S. (Eds.), 1993. Social Theory and Social Policy: Essays in Honor of James S. Coleman. Praeger, Westport, CT.

Collaborative Consumption

Rebeca Perren, College of Business Administration, University of Central Florida, Orlando, FL, USA
Liz Grauerholz, University of Central Florida, Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

We review a marketing reality borne of the digital era, collaborative consumption, in which individuals actively engage in the production of service offerings for the benefit of others. This phenomenon is rapidly gaining momentum with the advent of new technology and firms that seek to develop profitable business models by leveraging their Web platforms to engender trust among strangers and facilitate transactions among consumers. As a result, collaborative consumption has helped push traditional consumption communities from localized marketplaces with limited economic activity to collaborative global communities with important economic, environmental, and social effects.

Since the widespread adoption of Web 2.0 (2000s) there has been an enormous shift in the ability of consumers to provide services and coproduce consumption experiences for themselves and others. The proliferation of networked technologies has transformed exchanges among individuals, as online and mobile platforms are deployed to equip ordinary people with the ability to monetize their resources and skills. Reduced economic costs, time, and effort required for consumer participation in the production of market offerings has rendered exchange among individuals convenient, easy, and as readily available as Internet access. This phenomenon, which Botsman and Rogers (2010) call collaborative consumption, is rapidly gaining momentum. Collaborative consumption has been featured as one of 10 ideas that will change the world in the coming years (Walsh, 2011).

Collaborative consumption, which was estimated at more than $3.5 billion in 2013 (Geron, 2013), allows revenues to flow directly into individuals' pockets. For example, Konrad Marshall, a journalist, spent a week immersed in collaborative consumption, earning $335 for about 20 h of work that included hauling bricks for a suburban resident building a backyard chicken coop ($50), writing a bio for a band ($80), unpacking boxes and organizing their contents in a new home ($60), and a few other jobs (Marshall, 2014). Similarly, Larson Frederic generated about $3000 income a month renting his home and transforming his Prius into a *de facto* taxi (Geron, 2013). As Konrad and Larson illustrate, collaborative consumption allows individuals to leverage online and mobile platforms to monetize their skills and their idle assets. Such exchanges also operate largely within an informal economy as profits often go unreported and local taxes or fees uncollected.

Background on Collaborative Consumption

Collaborative consumption is also referred to as the 'sharing economy' because individuals are sharing access to resources (for a fee or other compensation), or 'peer-to-peer' exchange because both the service provider and recipient are individuals rather than businesses. Although exchange among individuals has taken place as long as people have been trading, bartering, and swapping, these traditional face-to-face peer exchanges had limited appeal and were restricted by geographic bounds. Moreover, exchange models such as garage sales or swap meets are temporary in nature and people who participate in these exchanges tend to oppose traditional marketplaces (Belk et al., 1988). As such, their attractiveness is limited to a niche market.

With the advent of new technology, however, these traditional consumption communities have evolved from localized marketplaces with limited economic activity to collaborative global communities with significant economic, environmental, and social consequences. Chalmers et al. (2013) analyzed nearly 100 articles published in marketing and consumer research and identified several dimensions on which consumption communities vary. Based on this structural classification, collaborative consumption communities are scalable, nongeographically bound, and enduring platforms that operate synergistically with traditional marketplaces. These new platforms of peer exchange are attractive to a broad market because they bring the convenience and access associated with traditional business-to-consumer practices to consumer-to-consumer exchange.

Beyond technology, what has propelled collaborative consumption practices is the involvement of firms that facilitate exchanges among peers. Older, established firms, such as *Craigslist* and *eBay*, have given way to a wide variety of firms that serve to connect or facilitate peer exchanges. These firms are growing tremendously in number, size, and profits. For example, *Lyft* provides on-demand peer-driven rides in cities across the United States and has recorded more than one million rides since it was founded in 2012 (http://blog.lyft.com); *TaskRabbit*, which outsources household errands and skilled tasks in the United States and the United Kingdom, boasted 1.25 million new users in 2013 alone (http://blog.taskrabbit.com); *thredUP* is a clothing resale platform earning sellers $3.2 million and saving shoppers $21 million in 2013 (http://www.thredup.com/resale); the 1 million members of *Quirky*, a community of inventors who collaborate in developing unique products, have developed over 350 products (https://www.quirky.com/about); and *LendingClub*, a peer-to-peer lending platform, has provided more the $5 billion in total loans, earning investors nearly $500 million in interest as of September 2014 (https://www.lendingclub.com). Indeed, firms and marketers have good reasons to be interested; a 2012 consumer panel survey found that generation X consumers with household incomes exceeding $75 thousand per year find

collaborative consumption most appealing, indicating a healthy market potential (Franz, 2012).

Collaborative consumption markets are structurally different from the traditional buyer-seller dyad; they are triadic, rather than dyadic (Perren et al., 2014). Firms in this market serve as intermediaries between an individual providing a service and the person benefiting from the service. Each actor in this triadic exchange is interdependent and actively involved in the coproduction of a unique consumption experience. Importantly, the roles of the firms, sellers, and consumers can differ across types of markets (e.g., some require buyers to contact sellers directly, in others buyers and sellers have no contact; some firms assume responsibility if a product or service fails, others do not). As such, collaborative consumption markets are more dynamic, flexible, and less institutionalized than traditional markets.

In addition, collaborative consumption markets are challenging existing business models and current regulatory environments. Collaborative consumption has emerged as a viable alternative to traditional business in industries such as transportation, hospitality, retail, and banking. As the popularity of these practices has grown, so has its economic, environmental, and societal impact. Recently, discussion of collaborative consumption has risen to the level of heated debate, along with increased attention from legal and regulatory bodies. Some collaborative practices have been deemed illegal (e.g., short-term rentals are prohibited in cities like New York and San Francisco) and lack of regulation and oversight of these informal transactions in areas traditionally regulated such as food preparation, banking, or transportation could expose collaborative consumers to risks.

Although collaborative consumption has gained the most popularity in wealthy countries like the United States, the United Kingdom, and Australia, developing nations are also embracing these new business models. Moreover, developing economies have a great potential to benefit from collaborative practices to alleviate lack of access to resources. Firms like *Airbnb* and *Uber* have already entered markets in Latin America, Asia, the Middle East, and Africa, while local entrepreneurs have sprung-up businesses to tackle problems such as access to education and financing.

Collaborative Consumption and the Evolution of Consumer Roles

Marketing research has documented the evolution of consumer roles in market exchanges. Service marketing scholars recognized the participatory role of the customer in value creation by highlighting the importance of 'partial employees' in the conceptualization of service quality (Kelley et al., 1990), as well as the value of 'commercial friendships' in interpersonal exchange between customers and service providers (Price and Arnould, 1999). Consumers' helpful, discretionary behaviors were conceptualized as 'customer voluntary performance' (Bettencourt, 1997), and later as 'customer citizen behaviors' (Rosenbaum and Massiah, 2007). A second rich body of literature in consumer research centered on the concept of 'brand community' (Muniz and O'Guinn, 2001) recognized that individuals can be an important information source for other individuals, providing social benefits and representing a form of consumer agency (Schau et al., 2009). Finally, scholars have recently attempted to subsume many of the behavioral manifestations toward a brand or firm under the concept of 'customer engagement behavior' (VanDoorn et al., 2010).

None of these conceptualizations captures the full extent of activities involving individuals in collaborative consumption, however. For example, partial employee, customer voluntary performance, and customer citizen behaviors are conceptualized in the context of consumer receiving a service from a firm, not being the service provider themselves. Similarly, customer engagement behaviors are theorized as customers' behavioral manifestation toward a brand or firm *beyond purchase* but don't necessarily address the service provider role that customers adopt in collaborative consumption. Collaborative consumption challenges all these traditional views of consumers (Vargo and Lusch, 2004) and point to a new stage in the evolution of the consumer role.

Likewise, this evolution of consumers from passive recipients of goods and services produced by businesses to active participants working collaboratively with enterprise in the marketplace is underscored by the proliferation of literature in sociology and consumer culture literature advocating its significance as well as criticizing its merits. Consumer culture theorists in particular have critically examined the marketing discourse and worklike activities carried out by consumers (Ritzer and Jurgenson, 2010; Dujarier, forthcoming; Cova and Cova, 2012; Zwick et al., 2008). Nonetheless, the concepts of 'prosumer' and 'prosumption' have become central to understanding the growing role of collaborative practices in shaping economic relations in contemporary capitalist society (Comor, 2010; Denegri-Knott and Zwick, 2011; Ritzer et al., 2012).

Differentiating Collaborative Consumption from Other Online Collectives

Collaborative consumption marketplaces where people coordinate the acquisition and distribution of a resource in exchange for a fee or other compensation (Belk, 2014) can be differentiated from other consumption communities studied in marketing and consumer behavior literature. **Table 1** provides a summary of the similarities and differences among key online collectives. First, brand communities such as online groups for Star Trek fans (Kozinets, 2001), are virtual meeting spaces where individuals can connect with each other to share experiences and advice; these groups focus on a central brand or product and foster a sense of collective belonging (Muniz and O'Guinn, 2001). Second, collective innovation communities refer to collaborative platforms where collective participatory actions of online consumers give rise to creativity and innovation (Kozinets et al., 2008). Similarly, virtual peer-to-peer problem solving communities focus on collaborative problem-solving activities related to consumption experiences of any type (Mathwick et al., 2008). The latter two online collectives provide a space for consumers to collaborate, innovate, and problem solve. Third, digital content sharing communities focus on sharing of digital 'e-goods' distributed for free among peers (Plouffe, 2008). Web sites such as Kazaa and Napster provide a meeting space to share digital content.

Table 1 Differences and similarities among online collectives

	Collaborative consumption	*Brand communities*	*Collective innovation communities*	*Digital content sharing communities*
Definition	Market model that enables individuals to coordinate the acquisition and distribution of a resource for a fee or other compensation (Belk, 2014), where the interaction is at least partially supported or mediated by technology	Specialized, nongeographically bound community, based on a structured set of social relationships among admirers of a brand. At its center is a branded good or service and it is marked by a shared consciousness, rituals and traditions, and a sense of moral responsibility (Muniz and O'Guinn, 2001)	Online creative consumer communities are collaborative platforms where collective, participatory actions of online consumers give rise to innovation (Kozinets et al., 2008). Peer-to-peer problem solving communities are networks of practice where consumers are 'working' to solve problems related to their shared consumption experiences (Mathwick et al., 2008)	Sometimes called P2P systems or P2P networks. Described as a form of exchange between individual consumers who operate in conjunction with one another, all outside the realm of traditional channels and the value chain, broadly construed (Plouffe, 2008)
Prototypes	Craiglist, ebay, Airbnb, Lending Club, Lyft, Uber	Ford Bronco, Macintosh and Saab online communities Star Trek fan communities	Flickr, Wikipedia, Crash the Super Bowl Lonely Planet's Thorn Tree Forum and VirtualTourist.com	Digital downloading Web sites such as Kazaa, Napster, Morpheus, and AudioGalaxy
Community purpose	Alternative marketplace to exchange of goods/services	Virtual meeting space to connect with others, share experiences, and advice	Virtual space to collaborate, innovate, or receive consumption problem solving advice	Virtual meeting space to share digital content
Type of activity	Commodity exchange	Sharing	Sharing	Sharing
Economic activity	Underground economy (often taxes are not paid, income goes unreported)	N/A – no economic activity takes place	N/A – no economic activity takes place	Illegal economic activities (often users do not have rights to distribute downloaded music)
Characteristics (*based on* Chalmers et al. (2013) *structural classification of consumption communities*)	● Can be online or both online and face-to-face ● Appeal can vary from limited to broad ● Focused on consumption activity ● Easy access ● Heterogeneous participants	● Online ● Appeal can vary from limited to broad ● Focused on a brand ● Access can vary from easy to limited ● Homogenous participants	● Online ● Appeal can vary from limited to broad ● Can be focused on brand or consumption activity ● Access can vary from easy to limited ● Homogenous participants	● Online ● Broad appeal ● Focused on consumption activity ● Easy access ● Heterogeneous participants
Service provider role	Varying roles depending on the extent to which it facilitates the exchanges	N/A – not involved	N/A – not involved	Provides platform for exchange
Key differences from collaborative consumption		● Sense of collective belonging ● Built around brand loyalty ● Not an exchange/marketplace	● Sense of collective belonging ● Built around shared desire to solve consumption-related problems ● Not an exchange/marketplace	● Not a marketplace ● Participants do not derive income from activities

Collaborative consumption communities differ from brand communities, collective innovation communities, and digital content sharing communities in three key aspects: community purpose, type of consumption activity, and type of economic activity. Collaborative consumption communities are formed with the explicit purpose of facilitating economic exchange of goods and services among individuals and rarely have a brand-specific focus. In contrast, the aforementioned communities have more prosocial purposes and their members engage in noneconomic activities such as sharing experiences, advice, or digital content.

The Effects of Collaborative Consumption

The effects of collaborative consumption can be felt at multiple levels, including the environment, the economy, individual consumer, and society at large. Consider, for example, potential effect on the environment. Marketing scholars argue that consumption turns into problematic behaviors – overconsumption – when the level of consumption becomes unacceptable due to environmental consequences, unaffordable due to economic consequences, or when it negatively affects personal and collective well-being (Sheth et al., 2010). Participation in collaborative consumption mitigates overconsumption by altering the consumption cycle and allowing individuals to acquire, use, and dispose of their assets in a way that positively influences the three pillars of sustainability: environmental, economic, and social (Huang and Rust, 2011). Environmental benefits are realized by extracting more use from assets that would otherwise go unused. For example, most power tools are used less than 13 min in their lifetime (Botsman and Rogers, 2010); when consumers are enabled to rent out or sell a tool to other consumers, the environmental impact can be minimized while maximizing its utility during the product's lifespan. In short, collaborative consumption potentially results in more sustainable consumption behaviors that benefit individual consumers, businesses, and society (Botsman and Rogers, 2010).

Collaborative consumption has also had a major economic effect on traditional markets, disrupting existing institutions in key industries like transportation, hospitality, and banking. For example, hotels lost revenue from the 9 million guests who used *Airbnb*, banks failed to collect interest from the $5 billion loans funded through *Lending Club*, and taxi and car rental companies lost millions of customers thanks to *Uber*, *Lyft*, and other ride-sharing businesses. Zervas et al. (2014) did find that *Airbnb* penetration into the state of Texas was negatively correlated with hotel revenues, with lower-end hotels and those with little conference space incurring the most financial cost. Their model projects that in the state of Texas, a 1% increase in the market size of *Airbnb* will result in 0.05% decrease in total hotel revenues. In short, collaborative consumption represents an important societal shift by altering the traditional exchange between consumers and firms, requiring even well-established markets to adapt their existing practices or marketing strategies.

Significant effects can also be seen at the individual level. Collaborative consumption may serve as a platform upon which individuals can enact and reinforce their political and personal ideologies. For example, proponents of collaborative consumption view it as a prosocial consumerist movement whereby individuals engage in political consumerism, using conscious consumption as a political tool (Parsons, 2014). Since these exchanges are less costly than traditional marketplaces, individual consumers benefit economically by fulfilling consumption needs at lower costs. The consumer providing the product or service also benefits from receiving supplementary income (as does the business facilitating the exchange). These marketplaces may also enhance social and individual well-being by strengthening a sense of community as consumers collaborate with each other. Yet, recent findings suggest that this may not always be the case. In a case study of car sharing conducted in Boston, Massachusetts, Bardhi and Eckhardt (2012) found that *Zipcar* users do not have an interest in meeting and socializing with other member despite the company's efforts to foster a sense of community. Others have noted that relationships with neighbors may be strained as individuals rent out their homes to strangers who may not ascribe to community norms of noise or cleanliness, disrupting the process of community building and sense of safety among local residents (Zervas et al., 2014).

There is little doubt that collaborative consumption will continue to have an important effect on society and that firms participating in these exchanges will continue to grow. Start-ups such as *Airbnb* and *Lyft* have raised $130 and $83 million respectively, in venture capital funding (Owyang, 2013). Similarly, *Lending Club* received $57 million from investors in 2013, placing the company's valuation at $2.3 billion and creating anticipation about going public in 2014 (Calvey, 2013). Most recently, *Uber* raised $1.2 billion in a round of funding with a valuation of over $18 billion, making the 5-year-old San Francisco start-up more valuable than companies such as Whole Foods and United Airlines (Knowledge@Wharton, 2014).

Future Research Directions

Despite its importance, scholarly research investigating this consumption phenomenon remains scant (Prothero et al., 2011), and the implications for firms, individuals, and societies are not well understood. Research is needed to understand the structure of collaborative consumption markets. What role do consumers and firms play within this market? What 'logics' dominate? How are various collaborative consumption markets similar or different from one another? Answers to such questions would provide practical guidelines for service marketers and help develop theory to advance theoretical understanding of the collaborative consumption phenomenon. Collaborative consumption communities also are fertile grounds for academic research to examine the evolving role of consumers as coproducers of their consumption practices and the strategies organizations deploy to harness its power. In terms of sustainable consumption, Prothero et al. (2011) argue that the full consumption cycle needs to be further investigated, as individuals' postpurchase behaviors (e.g., product usage, product life extension, and disposal) all have significant effects on the sustainability of consumption. Thus, how does the increased adoption of collaborative consumption markets affect sustainability? At the same time, alternative outlets that

allow consumers to alter how they use and dispose of their belongings could have important consequences for nonprofit organizations that rely on donations.

Participation in these marketplaces could enable individuals to pursue independent business opportunities instead of traditional employment, creating a new wave of micro-entrepreneurs. Research is needed to investigate whether collaborative consumption gives rise to new forms of consumer agency and how that shapes the labor force. Moreover, the regulatory environment remains in flux, and both firms and lawmakers could benefit from quality research that sheds light on the role of laws and regulation in assisting or inhibiting the growth of these business practice. For example, how can research better assess the effect of the largely informal economy, including the income that probably goes unreported? What are the consequences for local governments as sales and service taxation goes uncollected? Some evidence has surfaced to suggest that the total impact of collaborative consumption activities could be more positive than expected. A report commissioned by *Airbnb* claims that the company contributed an estimated $56 million to San Francisco's economic activity in the period from June 2011 to May 2012, providing locals with substantial income and having a positive effect on local neighborhoods that are 'off the beaten path' (Triple Pundit, 2013). However, more extensive independent studies are needed in order to measure and assess the real effects of this collaborative economy.

Organizational, legal, and community-level questions and concerns also arise. What effect will collaborative consumption have on organizations? For the collaborative consumption firms themselves, there are unique challenges such as how to contend with the complexities of serving two separate parties in the exchange and manage the increasingly volatile regulatory environment. More generally, will certain businesses be better able than others to integrate collaborative consumption models into their existing platforms and thereby tap into these lucrative markets? If so, will collaborative consumption be co-opted by more powerful markets (e.g., *Avis* acquired *Zipcar* in 2013)? The power struggle between emerging collaborative consumption markets and more established, traditional ones, is already being played out in a number of communities. For example, in 2012 the California Public Utilities Commission issued fines against car-sharing firms (*Lyft*, *SideCar*, *Uber*), yet American Express card members can now use their reward points for rides through *Uber*. Even organizations unrelated to the collaborative consumption movement may find themselves struggling with the realities of networked peer platforms. The University of California (UC) system found itself in the middle of a media storm when its director of travel services sent a message to faculty members warning them not to use the new peer-to-peer or sharing businesses such as *Uber*, *Lyft*, and *Airbnb* when traveling on UC business because "these services are not fully regulated and do not protect users to the same extent as commercially regulated business" (N.B., 2014). After the California lieutenant governor sent a letter urging university officials to reconsider such policy, UC's office of travel services issued a retracting statement.

For individual consumers, we might ask how and when does participation in collaborative consumption shape or change identities? There is an inherent incompatibility between prosocial values and the self-serving benefits linked to collaborative consumption. Does increased participation diminish the importance of moral identity, and thus weaken associated values and motives? Amid the aftermath of the global financial crisis during the 2000s, a heated debate has been brewing in the popular press about the nature of collaborative consumption practices, questioning whether the so-called 'sharing economy' is a manifestation of the empowered and entrepreneurial consumer or just the latest form of contemporary capitalist exploitation. The dualism between individual voluntaristic action and constrained deterministic behavior echoed in this contemporary discourse has long been at the center of scholarly debate regarding the primacy of structure or agency in shaping human behavior. Yet we know little about how peers experience these coproduction practices or how they perhaps construct entrepreneurial identities as a way of negotiating key tensions in this structure–agency dichotomy. Moreover, prior research investigating interactions among consumers surrounding marketer generated word-of-mouth campaigns suggest that the "intrusion of commercial 'hype' presents a type of moral hazard when it contains the inappropriate and unsanctioned mixing of dominant norms, such as sharing, caring or market exchange" (Kozinets et al., 2010: p. 85). Thus, consumer research is needed to determine if and how tensions arise due to both the social and commercial nature of exchange between peers. For example, how does monetary versus nonmonetary compensation affect exchange behaviors in collaborative consumption markets? What are the antecedents necessary to develop trust in other peers? And what effect does collaborative consumption have on interpersonal relationships, including neighbors?

What is needed is a broad research approach that includes different levels of analyses (structural, organizational, and individual) to understand how social structures, market institutions, and consumer practices interact to generate, maintain, and enforce a social order that combines the social domain of peers with the economic domain of market exchange. Such an approach could advance theoretical understanding of the collaborative consumption phenomenon and provide practical implications for managerial practice, public policy, and communities.

See also: Boundaries and New Organization Forms; Business Models; Consumption, History of; Entrepreneurship and Organizations; Ethical Consumers; Marketing Strategies; Networks and Cultural Consumption; Sustainable Practices; Sustainable Production and Consumption; Technology and Organization; Time Use and Resource Consumption.

Bibliography

Bardhi, F., Eckhardt, G.M., December 2012. Access based consumption: the case of car sharing. Journal of Consumer Research 39, 881–898.

Belk, R.W., 2014. You are what you can access: sharing and collaborative consumption online. Journal of Business Research 67 (8), 1595–1600.

Belk, R.W., Sherry, J.F., Wallendorf, M., 1988. A Naturalistic Inquiry into buyer and seller behavior at a swap meet. Journal of Consumer Research 14 (4), 449–470.

Bettencourt, L., 1997. Customer voluntary performance: customers as partners in service delivery. Journal of Retailing 73 (3), 383.

Botsman, R., Rogers, R., 2010. What's Mine Is Yours: The Rise of Collaborative Consumption. HarperCollins, New York.
Calvey, M., 2013. Lending club hits $3 billion milestone. San Francisco Business Times. Available at: http://www.bizjournals.com/sanfrancisco/blog/2013/12/lending-club-prosper-daric-san-francisco.html?page=all.
Chalmers, T., Price, L.L., Schau, H.J., 2013. When differences unite: resource dependence in heterogeneous consumption communities. Journal of Consumer Research 39 (5), 1010–1033.
Comor, E., 2010. Contextualizing and critiquing the fantastic prosumer: power, alienation and hegemony. Critical Sociology 37 (3), 309–327.
Cova, B., Cova, V., 2012. On the road to prosumption: marketing discourse and the development of consumer competencies. Consumption Markets & Culture 15 (2), 149–168.
Denegri-Knott, J., Zwick, D., 2011. Tracking prosumption work on eBay: reproduction of desire and the challenge of slow re-McDonaldization. American Behavioral Scientist 56 (4), 439–458.
Dujarier, M.-A. The three sociological types of consumer work. Journal of Consumer Culture. http://dx.doi.org/10.1177/1469540514528198.
Franz, L., 2012. Collaborative consumption: should marketers be afraid? Forbes. Available at: http://www.forbes.com/sites/onmarketing/2012/03/12/collaborative-consumption-should-marketers-be-afraid/; http://onforb.es/xWnTk5.
Geron, T., 2013. Airbnb and the unstoppable rise of the share economy. Forbes. Available at: http://www.forbes.com/sites/tomiogeron/2013/01/23/airbnb-and-the-unstoppable-rise-of-the-share-economy/; http://onforb.es/XxuTa1.
Huang, M.-H., Rust, R.T., 2011. Sustainability and consumption. Journal of the Academy of Marketing Science 39 (1), 40–54.
Kelley, S.W., Donnelly, J.H.J., Skinner, S.J., 1990. Customer participation in service production and delivery. Journal of Retailing 66 (3), 315.
Knowledge@Wharton, 2014. Kicking the Tires on Uber's $17 Billion Valuation: Is It Worth That Much?. Available at: http://knowledge.wharton.upenn.edu/article/kicking-tires-ubers-17-billion-valuation-worth-much/.
Kozinets, R.V., 2001. Utopian enterprise: articulating the meanings of Star Trek's culture of consumption. Journal of Consumer Research 28 (1), 67–88.
Kozinets, R.V., Hemetsberger, A., Schau, H.J., 2008. The wisdom of consumer crowds: collective innovation in the age of networked marketing. Journal of Macromarketing 28 (4), 339–354.
Kozinets, R.V., 2010. Networked Narratives: Understanding Word-of-Mouth. Journal of Marketing 74 (2), 71–89.
Marshall, K., et al., 2014. A week immersed in the "sharing economy." The Sidney Morning Herald. Available at: http://www.smh.com.au/digital-life/digital-life-news/a-week-immersed-in-the-sharing-economy-20140709-zt0a1.html.
Mathwick, C., Wiertz, C., De Ruyter, K., 2008. Social capital production in a virtual P3 community. Journal of Consumer Research 34 (6), 832–849.
Muniz, A.M.J., O'Guinn, T., 2001. Brand community. Journal of Consumer Research 27 (4), 412–432.
N.B., 2014. Steer clear of peer-to-peer. The Economist. Available at: http://www.economist.com/node/21606137.
Owyang, J., 2013. The Collaborative Economy. The Altimeter Group.
Parsons, A., 2014. The sharing economy: a short introduction to its political evolution. openDemocracy, pp.1–8. Available at: http://www.opendemocracy.net/transformation/adam-parsons/sharing-economy-short-introduction-to-its-political-evolution.
Perren, R., Massiah, C., He, X., 2014. The next generation of peer-to-peer (P2P) marketplaces: conceptualizing the dynamic interaction between firms and consumers in facilitating exchanges. In: Paper Presented to the American Marketing Association Winter Educator's Conference, Orlando FL, 21 February 2014.
Plouffe, C.R., 2008. Examining "peer-to-peer" (P2P) systems as consumer-to-consumer (C2C) exchange. European Journal of Marketing 42 (11/12), 1179–1202.
Price, L.L., Arnould, E.J., 1999. Commercial friendships: service provider-client relationships in context. Journal of Marketing 63 (4), 38–56.
Prothero, A., et al., 2011. Sustainable consumption: opportunities for consumer research and public policy. Journal of Public Policy 30 (1), 31–38.
Ritzer, G., Dean, P., Jurgenson, N., 2012. The coming of age of the prosumer. American Behavioral Scientist 56 (4), 379–398.
Ritzer, G., Jurgenson, N., 2010. Production, consumption, prosumption: the nature of capitalism in the age of the digital "prosumer." Journal of Consumer Culture 10 (1), 13–36.
Rosenbaum, M.S., Massiah, C.A., 2007. When customers receive support from other customers: exploring the influence of intercustomer social support on customer voluntary performance. Journal of Service Research 9 (3), 257–270.
Schau, H., Muniz, A.M.J., Arnould, E.J., September 2009. How brand community practices create value. Journal of Marketing 73, 30–51.
Sheth, J.N., Sethia, N.K., Srinivas, S., 2010. Mindful consumption: a customer-centric approach to sustainability. Journal of the Academy of Marketing Science 39 (1), 21–39.
Triple Pundit, 2013. The Rise of the Sharing Economy.
VanDoorn, J., et al., 2010. Customer engagement behavior: theoretical foundations and research directions. Journal of Service Research 13 (3), 253–266.
Vargo, S.L., Lusch, R.F., 2004. Evolving to a new dominant logic for marketing. Journal of Marketing 68 (1), 1–17.
Walsh, B., 2011. 10 ideas that will change the world. Today's smart choice: don't own. Share. Times Magazine 1–4. Available at: http://content.time.com/time/specials/packages/article/0, 28804, 2059521_2059717_2059710,00.html.
Zervas, G., Proserpio, D., Byers, J.W., 2014. The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry. Boston University School of Management Research Paper Series, No. 2013–1.
Zwick, D., Bonsu, S.K., Darmody, A., 2008. Putting consumers to work: "co-creation" and new marketing govern-mentality. Journal of Consumer Culture 8 (2), 163–196.

Relevant Websites

http://www.collaborativeconsumption.com.
http://meshing.it.
http://ouishare.net/en.
http://www.peers.org.
http://p2pfoundation.net.
http://www.thepeoplewhoshare.com.
http://www.shareable.net.
http://www.triplepundit.com/topic/rise-of-the-sharing-economy/.

Collective Action

Bert Klandermans, VU University, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A.R. Oberschall, volume 1, pp. 49–54, © 2001, Elsevier Ltd.

The article answers the questions of what is collective action, why do people participate in collective action, and how is preparedness to participate turned into action? As most collective action concerns acts of political protest, I concentrate on collective political action. Three motives are distinguished: instrumentality referring to participation as an attempt to influence one's situation; identity referring to participation as an expression of group identification; and ideology referring to participation as an expression of one's views. Anger strengthens motives to participate and increase the likelihood that motivation turns into action. Social embeddedness is crucial in the transformation of readiness into action.

Protests in the 'new' democracies in Central Europe about 'stolen elections', street demonstrations in the 'old' democracies against austerity measures, ongoing protests in the Arab world for more democracy, and occupied city squares throughout the world to rally against inequality and to claim better governance. Men and women of all ages express their grievances and aversion. They are mobilized through the media, through organizations they are members of, or by people they know. They come alone, or with family or friends. This article is about collective political action, that is to say, collective action aiming to influence politics. Almost daily our news media report on how people try to influence politics by collective action. This is not to say that political protest is something people regularly do. In fact, participants in political protest are most of the time a minority. Even mass mobilization rarely encompasses more than a few percent points of the population. This raises a question that has always occupied students of collective political action, especially social and political psychologists: *Why do some individuals participate in collective action while others do not?*

What Is Collective Political Action?

Collective political actions are challenges by people with common purposes and solidarity in interaction with elites and authorities (Tarrow, 1998: p. 4).

This definition includes two key elements that deserve some elaboration. First, collective political actions are *challenges*. They concern disruptive direct action against elites, authorities, other groups, or cultural codes. There is an obvious reason why this is the case. Collective political action is typically staged by people who lack access to politics. Had they had access there would have been no need for the action. Disruption forces authorities to pay attention to the claims brought forward. Second, it concerns people with a common purpose and solidarity. Participants rally behind common claims; they want authorities to do something, to change a state of affair or to undo changes. Such common claims are rooted in feelings of collective identity and solidarity. If these challenges gain sustainability, they constitute a social movement.

At the same time, a decline of participation in party politics is reported (Dalton and Wattenberg, 2000). Some labeled this trend 'movimentization of politics' (Neidhardt and Rucht, 1993), while others coined the term 'movement society' (Meyer and Tarrow, 1998; Johnston, 2011). Some facts: for example, in 1975, 22% of the British people signed some petition. In 1990, the figure was 75%. In 1974, 9% of the British participated in a demonstration against 25% in 1990. Between 1979 and 1993, the number of protest events in the French city of Marseille more than doubled from 183 to 395 events per annum. In that same period, considerable increases in protest events in Germany were reported, a result that is confirmed by Kriesi, Koopmans, Duyvendak, and Giugni's now classical study on new social movements in Europe (1995). According to the European Values Surveys, between 1981 and 2008, the proportion of the French population declaring they had at some time or another actually signed a petition, has climbed from 43 to 64%, as has participation in a lawful demonstration (25–42%), or a boycott (10–16%). Finally and somehow against the observed trend, the five rounds of the European Social Survey between 2002 and 2010 reveal a fairly stable 6–7% of the population of 16 European countries reporting participation in lawful demonstration in the past 12 months; 25–26% signed a petition, while 17–18% boycotted some products.

Why Do People Participate in Collective Political Action?

The social psychology of protest suggests that three fundamental motives plus anger account for participation in collective political action (van Stekelenburg and Klandermans, 2007): people may want to change their circumstances, they may want to act as members of their group, they may want to express their views, or they may want to vent their anger. I suggest that together these four factors account for most of the participation in collective political action. I refer to the three motives as: instrumentality, identity, and ideology. *Instrumentality* refers to participation as an attempt to influence one's social and political situation; *identity* refers to participation as an expression of identification with a group; and *ideology* refers to participation as an expression of one's views. *Anger* amplifies and accelerates; it strengthens motives to participate and makes it more likely that motivation turns into action. I hold that approaches that neglect any of those factors are fundamentally flawed. To be sure, individual participants may participate because of a single motive, but all three plus anger are needed

to understand why people take part in collective political action.

The Motivation to Participate in Collective Political Action

Instrumentality

Instrumentality begins with grievances (Klandermans, 1997), be it the experience of illegitimate inequality, feelings of relative deprivation, feelings of injustice, moral indignation about some state of affairs, or a so-called suddenly imposed grievance (an event imposed on people, for example, austerity measures taken by government). Social psychological grievance theories such as relative deprivation theory, or social justice theory have tried to specify how and why grievances develop. Resource mobilization theory and political process theory, the two approaches that dominate the field since those days, have always taken as their point of departure that grievances are ubiquitous and that the key question in action participation research is not so much whether people who participate are aggrieved, but whether aggrieved people participate. The instrumentality paradigm holds that participants are people who believe that they can change their social and political situation at affordable costs.

The perceived costs and benefits of collective action can be distinguished in *selective* and *collective* incentives (Klandermans, 1997). Selective incentives relate to costs and benefits that differentiate between participants and nonparticipants. You are only spending time or money if you participate; you only run the risk of being beaten up by the police if you participate; your friends will only blame you for not participating if you stay home, and so on. Sometimes movement organizations try to make participation more attractive by providing selective benefits: a popular music group, a train ticket to the city where the demonstration is held, a T-shirt, etc. Authorities or opponents on their part can try to make participation less attractive by imposing costs upon participants. Collective incentives are related to achievement of the movement's goals and the extent to which participation in a specific activity contributes to goal achievement. Obviously, it is not enough for a goal to be important to a person, some likelihood of success is also needed (Corcoran et al., 2011). The problem with collective action is that it is difficult to know to what extent an activity will have any influence on authorities. In any event, chances are low that an activity will have any impact if only a few people participate. Therefore, the likelihood of success is influenced by the expected behavior of others (Klandermans, 1984). If too few people participate it is unlikely that the activity will make any difference. As a consequence, expectations about the behavior of others play an important role in the decision to participate. If someone expects that only a few people participate his or her motivation to participate will be low. In a way, the expectation about other people's behavior functions as a self-fulfilling prophecy: if people believe that only few will participate, they will not be motivated to participate and thus make their own expectation true.

Instrumentality presupposes an effective movement that is able to enforce some wanted changes or at least to mobilize substantial support. Making an objective assessment of a movement's impact is not an easy task, but movement organizations will try to convey the impression of an effective political force. They can do so by pointing to the impact they have had in the past, or to the powerful allies they have. Of course, they may lack all this, but then, they might be able to show other signs of movement strength. A movement may command a large constituency as witnessed by turnout on demonstrations, or by membership figures, or large donations. It may comprise strong organizations with charismatic leaders who have gained respect, and so on. Instrumentality also implies the provision of selective incentives. Movements may vary considerably in the selective incentives for participation they provide. This is, obviously, also a matter of the resources a movement commands. Surprisingly little comparative information is available on the resources movements have at their disposal. In a similar vein, systematic documentation is lacking on the way in which the larger political system and the alliances and opponents of movement organizations influence movement participation. It was Tilly, who in 1978 coined the now classic terms 'repression' and 'facilitation' to distinguish between political systems that increase or decrease the costs of participation. Indeed, repressive political environments may increase the costs of participation considerably: people may lose friends, they may risk their jobs, or otherwise jeopardize their sources of income, they may be jailed, and they may even lose their lives.

From an instrumental perspective, a solution must be found to the dilemma of collective action. In 1965, Mancur Olson published his *The Logic of Collective Action*. The core of the book is the argument that rational actors will *not* contribute to the production of a collective good unless selective incentives persuade them to do so. Collective goods are characterized by 'jointness of supply'. That is to say, if they are made available to one person, they become available to everybody irrespective of whether people have contributed to the production of the collective good (for example, a law against discrimination, or measures against pollution). Therefore, according to Olson, a rational actor will choose to take a free ride, unless selective incentives prevent him from doing so. Olson's argument was soon applied to social movement participation. It helped to explain why so often people do not participate in social movements, despite the interest they have in the achievement of the movement's goals. Movement scholars argued that movement goals typically are collective goods. If the goal is achieved people will enjoy the benefits irrespective of whether they have participated in the effort. In view of a goal for which achievement is uncertain, but for which benefits – if materialized – can be reaped anyway, rational actors will take a free ride.

However, social movement scholars quickly discovered that reality was more complex than Olson's simple model suggested. The problem with Olson's logic of collective action is that it provides an explanation for why people do not participate, but fares poorly in explaining why people *do* participate. Moreover, Olson's solution that people participate for selective incentives is fundamentally flawed, as it does not give a satisfactory answer to the question of where the resources needed to provide selective incentives come from (Oliver, 1980). If these must be collected from individual citizens the same collective action dilemma arises again. This is not to say that selective incentives are irrelevant, but that in the final instance they cannot solve the collective action dilemma. In other words, if

collective and selective incentives do not provide a sufficient explanation of movement participation, what else might make the difference? A recurring criticism was that Olson's model assumes that individuals make their decisions in isolation, as if there are no other people with whom they consult, with whom they feel solidarity, and by whom they are kept to their promises. This pointed to the significance of collective identity as a factor in movement participation.

Identity

It soon became clear that instrumentality was not the only reason to participate. After all, many goals are only reached in the long run if at all. Similarly, when it comes to material benefits, costs frequently outweigh benefits. Apparently, there is more in being a movement participant than perceived costs and benefits. Indeed, one of those motives relates to belonging to a valued group (Simon et al., 1998).

Identity is described as a place in society. People occupy many different places. They are student, unemployed, housewife, soccer player, politician, farmer, and so on. The different roles and positions a person occupies form his or her personal identity. At the same time, every place a person occupies is shared with other people. I am not the only professor of social psychology, nor the only Dutch, or the only European. I share these identities with other people – a fact that turns them into collective identities at the same time.

Most of the time collective identities remain latent. Self-categorization theory hypothesizes that an individual may act as a unique person, that is, display his personal identity, *or* as a member of a specific group, that is, display one of the many collective identities he has (Turner, 1999). Obviously, this is often not a matter of free choice. Circumstances may force a collective identity into awareness whether people like it or not, as the Yugoslavian and South African histories have illustrated dramatically and there are other equally or even more dramatic examples throughout human history. But also in less extreme circumstances, collective identities can become significant. Take, for example, the possible effect of an announcement that a waste incinerator is planned next to a neighborhood. Chances are that in no time the collective identity of the people living in that neighborhood will become salient.

Self-categorization theory proposes that people are more prepared to employ a social category in their self-definition, the more they identify with that category. Identification with a group makes people more prepared to act as a member of that group. This assertion refers, of course, to identity strength. Social identity literature tends to neglect that real-world identities vary in strength. But, the strength of group identification may make a real difference especially in political contexts. We may expect that strong identities make it more likely that people act on behalf of their group. The basic hypothesis regarding collective identity and movement participation is fairly straightforward: a strong identification with a group makes participation in collective political action on behalf of that group more likely. The available empirical evidence overwhelmingly supports this assumption.

Movements offer the opportunity to act on behalf of one's group. This is most attractive if people identify strongly with their group. The more farmers identify with other farmers the more appealing it is to take part in farmers' protest. The more women identify with other women the more attractive it is to participate in the women's movement. In addition to the opportunity to act on behalf of the group, collective political action participation offers the opportunity to identify with the movement's cause, the people in the movement, the movement organization, the group one is participating in, or the leader of the movement. Not all these sources of identification are always equally appealing. Movement leaders can be more or less charismatic, or the people in the movement or in someone's group can be more or less attractive. Moreover, movements and movement organizations may be, and in fact often are, controversial. Hence, becoming a participant in a movement organization does not mean taking a respected identity upon oneself, on the contrary. Within the movement's framework, this is, of course, completely different. There the activist does have the status society is denying him. And, of course, for the activist in-group–out-group dynamics may turn the movement organization or group into a far more attractive group than any other group 'out there' that is opposing the movement. Indeed, it is not uncommon for militants to refer to the movement organization as a second family, a substitute for the social life society is no longer offering them. Movement organizations not only supply sources of identification, but they also offer all kinds of opportunities to enjoy and celebrate the collective identity: marches, rituals, songs, meetings, signs, symbols, and common codes.

A complicating matter is the fact that people simultaneously hold multiple identities while movements tend to emphasize a single identity and refer to a single place in society. As a consequence, people may experience that conflicting identities steer behavior in different directions. Individuals might find themselves under cross pressure when two groups they identify with are on opposite sides of a controversy. Indeed, movement activists who challenge their government are often accused of being disloyal to the country. Gonzalez and Brown (2003) coined the term 'dual identity' to point to the concurrent workings of supra- and subordinated identities. They argue that identification with a subordinate entity (e.g., ethnic identity) does not necessarily exclude identification with a supra-ordinate entity (e.g., national identity). In fact, they claim that dual identity is a healthy configuration, as it implies sufficient identification with one's subgroup to experience basic security *and* sufficient identification with the overarching group to preclude divisiveness. There is evidence that indeed people who hold a dual identity are more satisfied with their situation than people who do not (Gonzalez and Brown, 2003). However, if they are dissatisfied, individuals who hold a dual identity were more likely to participate in collective action (Klandermans et al., 2008).

In 2001, Simon and Klandermans published an influential paper on the politicization of collective identity (PCI). In order to become the vehicle of collective action collective identity must politicize, they argued. Shared grievances, common enemies, and a search for third party support are the building stones of PCI, the authors maintain. Some sense of identification with the superordinate political entity seems to be a basic requirement of social and political mobilization, in that it ensures that this entity is acknowledged as one's own social or political habitat. Therefore, politicized identity is by

definition dual identity. More specifically, to the extent that one identifies with the superordinate entity one should feel entitled to make political claims. Similarly, one should feel motivated to get actively involved in the political game, because it becomes one's own game, and one should feel encouraged to approach third parties as potential allies, because they can be viewed as in-group members at the superordinate level, so Simon and Ruhs (2008). Politicization divides people's social environment into allies and opponents and results in polarization. Polarization concerns the process of distancing of the opposing camps. Eventually, this may result in radicalization. Simon (2011) suggests that in a polarized situation, to the extent that PCI is a dual identity including identification with the superordinate polity, PCI has a pacifying effect on politicization and associated collective action in that it prioritizes claims and actions that stay within the limits of normative acceptance in the larger polity. In contrast, collective identities lacking this pacifying effect, such as separatist identities based on more exclusive cultural, ethnic, or religious allegiances, should be more prone to escalation and radicalization. Collective identities must politicize to become the engine of collective action. Typically, politicization of identities begins with the awareness of shared grievances. Next, an external enemy is blamed for the group's predicament, and claims for compensation are leveled against this enemy. Unless appropriate compensation is granted, the power struggle continues. Politicization of identities and the underlying power struggle unfold as a sequence of politicizing events that gradually transform the group's relationship to its social environment, whereby the tactical choices are again shaped by identity. Hence, workers strike and anarchists fight the police.

Ideology

The third motive – wanting to express one's views – refers at the same time to a longstanding theme in the social movement literature, and to a recent development. In classic studies of social movements, the distinction was made between instrumental and expressive movements, or protest. In those days, instrumental movements were seen as movements that aimed at some external goal, for example, the implementation of citizenship rights. While expressive movements had no external goals. Participation was a goal in itself, for example, the expression of anger in response to experienced injustice. Movement scholars felt increasingly uncomfortable with the distinction, as it was thought that most movements had both instrumental and expressive aspects and that the emphasis on the two could change over time. Therefore, the distinction lost its use. Recently, however, the idea that people might participate in movements to express their views has received renewed attention. Attention arising this time from movement scholars, who were unhappy with the overly structural approach of resource mobilization and political process theory. These scholars began to put an emphasis on the creative, cultural, and emotional aspects of social movements, such as music, symbols, rituals, narratives, and moral indignation (Goodwin et al., 2000). People who are angry develop feelings of moral indignation about some state of affairs or some government decision, and wish to make that known. They participate in a social movement not only to enforce political change, but also to gain dignity in their lives through struggle and moral expression.

Social movements play a significant role in the diffusion of ideas and values. Rochon (1998) makes the important distinction between 'critical communities', where new ideas and values are developed, and 'social movements' that are interested in winning social and political acceptance for those ideas and values. "In the hands of movement leaders, the ideas of critical communities become ideological frames" (p. 31), states Rochon, who continues to argue that social movements are not simply extensions of critical communities. After all, not all ideas developed in critical communities are equally suited to motivate collective action. Social movement organizations, then, are carriers of meaning. Through processes such as consensus mobilization or framing, they seek to propagate their definition of the situation to the public at large. A study of flyers produced by the various groups and organizations involved in the protests against the International Monetary Fund and the World Bank in Berlin is an excellent example in this respect (Gerhards and Rucht, 1992). The study shows how links are constructed between the ideological frame of the organizers of the demonstration and those of the participating organizations in order to create a shared definition of the situation. In the social movement literature, such definitions of the situation have been labeled 'collective action frames'.

Social movements do not invent ideas from scratch; they borrow from the history of ideas. They build on an ideological heritage as they relate their claims to broader themes and values in society. In so doing they relate to societal debates that have a history of their own and that history is usually much longer than that of the movement itself. Gamson (1992), for example, refers to the 'themes' and 'counterthemes' that, in his view, exist in every society. One such pair of a theme and countertheme he mentions is 'self-reliance' vs 'mutuality', that is the belief that individuals should take care of themselves vs the belief that society is responsible for its less fortunate members. In my own work, I have demonstrated how in the Netherlands these two beliefs became the icons that galvanized debate and spurred protest over disability payments. While, 'self-reliance' became the theme of those favoring restrictions in disability payment, 'mutuality' was the theme of those who defended the existing system (Klandermans and Goslinga, 1996).

Emotions

Ideology has a significant affective component. Acting on one's ideology is deemed to be one of the fundamental motives of action participation and necessarily charged with emotion. Appraisal and action are socially constructed, that is to say, are formed in interpersonal interaction, especially in the case of politically relevant emotions. Cultural and historical factors play an important role in the interpretation of the state of affairs by which politically relevant emotions are generated. Obviously, appraisal can be manipulated. Activists work hard to create moral outrage and anger, and to provide a target against which these can be vented. They must weave together a moral, cognitive, and emotional package of attitudes. But also, in the ongoing activities of the movements, emotions play an important role. In the literature, two kinds of collective emotions are distinguished – reciprocal emotions and shared emotions – that reinforce each other. Each measure of shared

outrage against an injustice reinforces the reciprocal emotion of fondness for others precisely because they feel the same way. Conversely, mutual affection is a context in which new shared emotions are easily created. Anger and indignation are emotions that are related to a specific appraisal of the situation.

Recent work in sociology and social and political psychology has brought emotions to the study of social movements (Goodwin et al., 2000; van Zomeren et al., 2004). For those of us who have been part of protest events or watched reports on protest events in the news media, this is hardly surprising. Indeed, it is hard to conceive of protest detached from emotions. Emotions can be avoidance or approach oriented. Fear, which makes people refrain from taking action, is an example of an avoidance-oriented emotion. Anger is an approach-oriented emotion and is known to be an antecedent of protest participation (van Zomeren et al., 2004). There appears to be a relation between emotions and efficacy. When people do not feel efficacious, they are more likely to experience fear; feeling efficacious, on the other hand, is associated with experiencing anger. Findings from a study among migrants we conducted confirm this, feelings of efficacy reinforced anger and reduced fear, while in their turn, anger fostered collective action participation while fear undermined it (Klandermans et al., 2008). van Zomeren et al. (2004) show that anger is an important motivator of protest participation of disadvantaged groups. Anger and fear are not the only emotions relevant in the context of movement participation; indeed, other emotions such as hope and despair are proposed as well (Gould, 2009; Stürmer and Simon, 2009). Anger moves people to adopt a more challenging relationship with authorities than subordinate emotions such as shame and despair or fear.

The Transformation of Potential into Action

People who are prepared to take part in collective action do not automatically participate. Potential must be transformed into action (Klandermans and Oegema, 1987). This is what mobilization campaigns are about. Targeting potential participants and turning them into actual participants. Obviously, a campaign will never be 100% successful in that regard. In the course of a campaign, people may loose their sympathy for the movement, or the movement may fail to target or to motivate its sympathizers. As the previous section concentrated on motivating, I will focus now on targeting.

Targeting sympathizers implies answering two strategic questions: Who are the sympathizers? And, How can they be reached? These are two questions every attempt to mobilize must find an answer to. Social networks are of crucial importance in this regard. People are embedded in the social fabric of society (Klandermans et al., 2008). Targeting sympathizers implies knowing where to find them and commanding communication channels that reach them. Movement organizations have two options in that respect: they can try to co-opt existing networks or they can build new networks. Both strategies are mobilization efforts in itself (Boekkooi, 2012). Co-optation is the easier strategy of the two, because it builds on existing commitments to organizations and networks that are the movement organization's allies. There are risks, however. The co-opted organization may use the campaign for its own ends or negotiations with the leadership fail so that it decides to not collaborate. The latter makes it more difficult for the rank and file to cooperate. Yet, co-opting existing networks such as churches, unions, political parties, youth organizations, and the like is frequently applied, if only because it implies an answer to both strategic questions at the same time. On the one hand, it works from the assumption that most members of the allies sympathize with the movement, and on the other hand, it is assumed that these sympathizers can be reached through the allies' communication channels.

Building new networks implies the recruitment of people, who are willing to spend sometimes considerable amounts of time for a prolonged period as a movement activist. It will, therefore, require more effort on the part of the organizer than co-optation of existing networks, but once established the networks are more reliable. The recruitment of such movement activists is a process, which is determined, on the one hand, by factors that influence who is being asked, and on the other hand, by factors that influence who agrees to serve as an activist when asked. As for the first type of factors, a crucial determinant is someone's embeddedness in networks linked to the movement organization, or more specific to the movement organizer, who is undertaking the recruitment effort. Movement organizers tend to recruit first among the people they know and often that suffices. You need activists to maintain the network, and once you have those the return of having additional activist is rapidly diminishing. In fact, long-term activism is one of the forms of activism that must cope with free rider behavior. The people who are asked to serve as an activist understand perfectly well that they are giving most of the sympathizers to the movement a free ride, but that they are prepared to do so because they care. Only people who really care a lot are prepared to sacrifice for the others. They make the effort because they feel that "If [they] do not do it, nobody else will" (Oliver, 1984).

Once the mobilizing structure is in place, the actual mobilization for collective action can proceed (Boekkooi, 2012). Again embeddedness in the social fabric of society plays a crucial role. People may be directly or indirectly connected to the networks of a movement's mobilization potential or not connected at all. Such embeddedness or unembeddedness is of crucial influence on the likelihood of being targeted and if targeted on the likelihood of participation.

See also: Collective Behavior, Social Psychology of; Intergroup Relations; Political Psychology; Social Identity in Social Psychology; Social Movements: A Social Psychological Perspective; Social Psychology.

Bibliography

Boekkooi, Marije, 2012. Mobilizing Protest. The Influence of Organizers on Who Participates and Why. VU-University, Amsterdam.

Corcoran, K.E., Pettinicchio, D., Young, J.T.N., 2011. The context of control: a cross-national investigation of the link between political institutions, efficacy, and collective action. British Journal of Social Psychology 50, 575–605.

Dalton, R.J., Wattenberg, M.P., 2000. Parties Without Partisans. Oxford University Press, Oxford.

Dodson, Kyle, 2011. The movement society in comparative perspective. Mobilization 16, 475–494.

Gamson, W.A., 1992. Talking Politics. University of Cambridge Press, Cambridge.

Gerhards, Jürgen, Rucht, Dieter, 1992. Mesomobilization: organizing and framing in two protest campaigns in West Germany. American Journal of Sociology 98, 555–596.

Goldstone, J.A., 2004. More social movements or fewer? Beyond political opportunity structures to relational fields. Theory and Society 33, 333–365.

Gonzalez, Roberto, Brown, Rupert, 2003. Generalization of positive attitudes as a function of subgroup and superordinate group identifications in intergroup conflict. European Journal of Social Psychology 33, 195–214.

Goodwin, Jeff, Jasper, J.M., Polletta, Francesca, 2000. Return of the repressed: the fall and rise of emotions in social movement theory. Mobilization 5, 65–82.

Gould, Deborah, 2009. Moving Politics: Emotion and ACT UP's Fight against AIDS. University of Chicago Press, Chicago.

Johnston, Hank, 2011. States and Social Movements. Polity Press, Cambridge.

Klandermans, Bert, 1984. Mobilization and participation: social psychological expansions of resource mobilization theory. American Sociological Review 49, 583–600.

Klandermans, Bert, 1997. The Social Psychology of Protest. Blackwell, Oxford.

Klandermans, Bert, Goslinga, Sjoerd, 1996. Media discourse, movement publicity and the generation of collective action frames: theoretical and empirical exercises in meaning construction. In: McAdam, Doug, McCarthy, John, Zald, Mayer (Eds.), Opportunities, Mobilizing Structures, and Frames: Comparative Applications of Contemporary Movement Theory. Cambridge University Press, Cambridge, pp. 312–337.

Klandermans, Bert, Oegema, D., 1987. Potentials, networks, motivations and barriers: steps toward participation in social move. American Sociological Review 5, 519–531.

Klandermans, Bert, Toorn, Jojanneke van der, Stekelenburg, Jacquelien van, 2008. Embeddedness and identity how immigrants turn grievances into action. American Sociological Review 73, 992–1012.

Kriesi, Hanspeter, Koopmans, Ruud, Duyvendak, J.W., Marco, Giugni, 1995. New Social Movements in Western Europe: A Comparative Analysis. University of Minnesota Press, Minneapolis.

Leach, C.W., Iyer, A., Pedersen, A., 2006. Anger and guilt about in-group advantage explain the willingness for political action. Personality and Social Psychology Bulletin 32, 1232–1245.

McCarthy, J.D., Rafail, Patrick, Gromis, Ashley, 2013. Recent trends in public protest in the U.S.A.: the social movement society thesis revisited. In: Stekelenburg, Jacquelien van, Roggeband, Conny, Klandermans, Bert (Eds.), The Future of Social Movement Research: On Dynamics, Mechanisms, and Processes. University of Minnesota Press, Minneapolis.

Meyer, Davld, Tarrow, Sidney, 1998. The Social Movement Society: Contentious Politics for a New Century. Rowman & Littlefield, Boulder, CO.

Neidhardt, Friedhelm, Rucht, Dieter, 1993. Auf dem Weg in die Bewegungsgesellschaft? Uber die Stabilisierbarkeit sozia Bewegungen. Sozialer Welt 44, 305–326.

Oliver, P.E., 1980. Rewards and punishments as selective incentives for collective action: theoretical investigations. American Journal of Sociology 85, 1356–1375.

Oliver, P.E., 1984. If you don't do it, nobody else will: active and token contributors to local collective acti. American Sociolo Review 49, 601–610.

Olson, Mancur, 1965. The Logic of Collective Action. Public Goods and the Theory of Groups. Harvard University Press, Cambridge, MA.

Rochon, T.R., 1998. Culture Moves: Ideas, Activism, and Changing Values. Princeton University Press, Princeton, NJ.

Simon, Bernd, 2011. Collective identity and political engagement. In: Azzi, Assad E., Chryssochoou, Xenia, Klandermans, Bert, Simon, Bernd (Eds.), Identity and Participation in Culturally Diverse Societies. Wiley-Blackwell, Malden, MA, pp. 137–157.

Simon, Bernd, Klandermans, Bert, 2001. Towards a social psychological analysis of politicized collective identity: conceptualization, antecedents, and consequences. American Psychologist 56, 319–331.

Simon, Bernd, Loewy, Michael, Stürmer, Stefan, Weber, Ulrike, Kampmeier, Claudia, Freytag, Peter, Habig, Corinna, Spahlinger, Peter, 1998. Collective identity and social movement participation. Journal of Personality and Social Psychology 74, 646–658.

Simon, B., Ruhs, D., 2008. Identity and politicization among Turkish migrants in Germany: the role of dual identification. Journal of Personality and Social Psychology 95 (6), 1354–1366.

Stürmer, Stephan, Bernd, Simon, 2009. Pathways to collective protest: calculation, identification, or emotion? A critical analysis of the role of group-based anger in social movement participation. Journal of Social Issues 65, 681–705.

Tarrow, Sidney, 1998. Power in Movement. Social Movements, Collective Action and Mass Politics in the Modern State, second ed. Cambridge University Press, Cambridge.

Tilly, Charles, 1978. From Mobilization to Revolution. Addison-Wesley, Reading, MA.

Turner, J.C., 1999. Some current issues in research on social identity and self-categorization theories. In: Ellemers, Naomi, Spears, Russell, Doosje, Bertjan (Eds.), Social Identity. Blackwell, Oxford, pp. 6–34.

van Stekelenburg, J., Klandermans, Bert, 2007. Individuals in movements: a social psychology of contention. In: Klandermans, B., Roggeband, C. (Eds.), The Handbook Social Movements across Disciplines. Springer, New York, pp. 157–204.

van Zomeren, M., Spears, R., Fischer, A.H., Leach, C.W., 2004. Put your money where your mouth is! Explaining collective action tendencies through group-based anger and group efficacy. Journal of Personality and Social Psychology 87 (5), 649–664.

Collective Behavior, Social Psychology of

Stephen Reicher, School of Psychology and Neuroscience, University of St. Andrews, St. Andrews, UK
John Drury, School of Psychology, University of Sussex, Brighton, UK

Published by Elsevier Ltd.

Abstract

In this article, we first review classic models of crowd and collective psychology, second outline a contemporary approach to crowd phenomena based on social identity theorizing, and third apply this approach to understand the dynamics of social conflict and social change. We finish by arguing that crowd and collective psychology needs to be moved from the periphery to the center of social psychology and the social sciences and that its continuing exclusion reflects a tendency to under-represent and undervalue resistance to and change of the existing social order.

Introduction

The term 'collective behavior' has many different meanings in different disciplines. It has variously been used to refer to the behavior of the masses in society, to 'publics' or groups organized around a particular issue, to social movements which are organized to effect change around an issue, and to crowds which are physically copresent groups characteristically (but not necessarily) organized around some event (such as a sporting crowd) or some shared concern (such as a protest crowd).

Psychologists, however, have generally concentrated on the last of these meanings. Collective behavior and crowds become largely synonymous. This is the position we shall follow here. Additionally we shall use the term 'collective action' to refer to instances where crowd members act together in a coordinated way.

The psychological study of crowds and collective action attests to a profound conservatism at the core of the discipline. Crowds have tended to be feared and reviled as a threat to the status quo in society. They are an eruption of the primitive, the passionate, and the pathological which has the potential to shatter the order and reason of everyday life.

This leads to two problems: one of commission and one of omission. On the one hand, where crowds and collective action are studied, they are viewed as something exotic and exceptional which reveals the breakdown of ordinary psychological processes and hence has little to tell us about these processes. On the other hand, when it does come to the bread and butter issues that dominate the discipline, crowds and collective action are all but ignored. In psychology we are predominantly concerned with what individuals see, think, and feel and our interventions are designed to change perceptions, cognitions, and emotions. Collective action, it appears, has no place in this at all.

Here, we will examine the traditional psychology of crowds and collective action, how it came into being, and what it has to say about these phenomena. We will then outline a body of recent work which challenges the pathologizing perspective of this traditional work and which shows collective action to be meaningful, purposeful, and both reflecting and also changing the perspectives of participants. Moreover, we will point to the ways in which collective action plays a pivotal role in shaping the perspectives which regulate mundane social life. We will conclude by examining how, despite this, collective action is excluded from mainstream social psychological analysis partly through the way in which change and collective resistance is ignored in the discipline and partly through the way that, even when change is addressed, it is treated more in terms of individual changes among the privileged than in terms of the collective action of the oppressed.

Overall, we shall argue that any social psychology which recognizes the unequal and contested nature of social reality, and which is open to the possibility of social change, must move the study of crowds and collective action from the periphery to the center of the discipline.

Studying Crowds and Collective Action

The Origins of Collective Psychology

There are three layers of influence which shaped the emergence of a systematic collective psychology, predominantly in France, in the late nineteenth century. First, through the centuries, there has been a constant flow of elite invective directed at the crowd. Emblematic is a comment by the Danish philosopher, Soren Kierkegaard: "you must be aware of the opinion of the masses in order to know what to do: the opposite" (cited in Reiwald, 1949: p. 325. Translation by the authors). Second, the industrialization of western Europe through the nineteenth century led to a rapid growth of the masses, concentrated in the cities. This posed an urgent question of social order: would the masses accept their place or would they rise up and overthrow the existing order (Giner, 1976)? Moreover, if the masses in general were a cause of concern, the crowd, as the mass in action, was a particularly rich repository of elite fears (Barrows, 1981). Third, some countries were less stable than others and more vulnerable to mass agitation. Notably, the elites of the French Third Republic were permanently on guard against a wave of oppositional mobilizations: clerical, populist, and, above all, politically radical (Nye, 1975).

Of all the early crowd psychologists, one name stands out, that of Gustave Le Bon. This is less because of his theoretical originality (Le Bon was an expert synthesizer of others' ideas) and more because of his practical and activist bent. Whereas other authors simply noted and bemoaned what they saw to be

the dangerous and destructive tendencies of the crowd, Le Bon's project was to use his understanding in order to harness the power of the crowd for the elite and against the radicals. His book *The Crowd* (1895/1947) reads like a primer on how to manipulate crowds. It was enormously successful, influencing the like of Mussolini, Goebbels, and Hitler, and has been described as the most influential psychology text of all time (Moscovici, 1980).

Classic Crowd Psychology: Debating the Site of Pathology

For Le Bon, crowd psychology is a litany of loss. When people become submerged in a crowd, they lose their individual identity, the source of their standards, and values. Hence they lose their ability to judge and reason. As a consequence, crowds are marked by contagion whereby people simply go along with any passing idea or (because people become incapable of dealing with complex concepts) any passing emotion. These impulses largely derive from an underlying racial unconsciousness which remains once our conscious personalities cease to function. Because this unconsciousness is atavistic and primitive, the behavior of crowds is primitive, sometimes heroic, but above all destructive. Crowd members, having lost their selves and their minds, also lose their morality. People 'descend several rungs on the ladder of civilization' upon entering the crowd, avers Le Bon. Ordinarily reasonable people become senseless beasts. These ideas have filtered into, and continue to influence modern social psychology through the notion of 'deindividuation' (for a critique see Reicher et al., 1995).

However influential Le Bon's ideas may have been outside psychology, in many ways his greatest influence inside the discipline came through those who reacted against them. Notably, Floyd Allport used a large part of his foundational text 'Social Psychology' (first published in 1924) to refute Le Bon's ideas. He denounced the idea of a 'racial unconscious' or indeed any notion of mind separate from the individual as a nonsense. He argued that "there is no psychology of groups which is not essentially and entirely a psychology of individuals" (1924: p. 4). And, most famously, he insisted that "the individual in the crowd behaves just as he would alone only more so" (p. 295).

While the details of Allport's crowd psychology have been largely forgotten, the individualism that he championed has, over time, become orthodoxy throughout social psychology. In crowd research it takes the form of 'convergence' models (cf Turner and Killian, 1972), which propose that the character of crowd action derives from personal characteristics of crowd members. Specifically, violent and antisocial crowds reflect the convergence of violent antisocial or 'riot prone' individuals, the so-called 'riffraff' of society (e.g., US Riot Commission, 1968).

If Le Bon's position can be characterized in terms of the crowd rendering people 'mad,' so Allport's position can be characterized in terms of the crowd as an assembly of the 'bad.' And despite the apparent opposition between the two perspectives, they share a number of fundamental assumptions in common. Both view individual identity or personality as the sole basis for reasoned action. Both see the operation of a reasoning self as lacking in the crowd. Where they differ is whether this lack derives from the crowd process itself (Le Bon) or the nature of the individuals who are attracted to the crowd (Allport). Traditional crowd psychology of whichever variant sees crowds as pathological, the difference lies in whether the site of this pathology lies at the collective or the individual level.

Perhaps the most fundamental criticism of these ideas is that they cannot explain the empirical characteristics of collective action. While both Le Bon and Allport suggest that crowds are characteristically and indiscriminately violent, the historical record tells a very different story. Not only do very few crowd events result in violence but even when they do, behavior is highly patterned. What is more, these patterns are highly meaningful in terms of social belief systems. A case in point is Davis's analysis of the St Bartholomew's Day massacre of 1572 (Davis, 1978). This event, the culmination of French religious wars between Catholics and Protestants, was extremely brutal. Members of each side killed members of the other. Catholics then concentrated on desecrating Protestant bodies, while Protestants concentrated more on desecrating the objects of the holy mass. This, Davis argues reflects the fact that, in the struggle over imposing the 'true' doctrine, Catholics saw heresy as residing in the individual Protestant, whereas Protestants saw heresy as residing more in the rites of Catholicism – the 'demonic mass.' Albeit in the most brutal of manners, each side literally inscribed its beliefs upon the opponent.

The question, then – a question which classic crowd psychology cannot answer – is how can crowd members come together, often spontaneously and in unstructured gatherings, yet still act in coherent and meaningful ways? The obvious answer is that they are governed by social norms, and that is indeed the approach of emergent norm theory (Turner and Killian, 1972). According to ENT, crowd events begin with a period of milling during which prominent individuals (keynoters) determine norms of behavior. But the theory does not explain what does and does not become normative (beyond rooting it in the personal charisma of keynoters and their interpersonal interactions with crowd members). Hence, while ENT can explain the restrained and patterned nature of crowd action, it provides no basis for explaining how the local norms developed in the event connect to larger belief systems and hence it cannot account for the socially meaningful nature of these patterns. It is precisely this issue which led to the development of social identity models which have come to dominate the study of crowds and collective action in recent decades.

Social Identity and Collective Action

The starting point for social identity researchers is a reformulation of the concept of self. Rather than being a singular entity, selfhood is a system that operates at different levels of abstraction. At different times and places, different elements of the system become salient and guide what we think and do (Turner et al., 1987). Specifically, we can either think of ourselves in terms of what makes us, as individuals, distinctive from other individuals ('I' vs 'you' – personal identity) or in terms of what makes the various groups that we belong to different from other groups ('we' vs 'they' – social identity). Moreover, when a given social identity (associated with a specific group membership, such as 'Catholic') is salient, we behave in terms of the beliefs and values that characterize that group.

The concept of social identity has been proposed as an antecedent of participation in collective action (thus we are more likely to participate when we identify with a social group, believe that we have to struggle for the rights of the group; feel angry because we consider that the group suffers from unfair social disadvantage; and also believe that, as a group, we have the power to rectify that disadvantage – see van Zomeren et al., 2012). It has also been studied as a consequence of collective participation (both in terms of the content of identity and the strength of attachment to the group – see Drury and Reicher, 2009). Most of all, however, it has been used to account for the nature of collective action itself, and to contest the pathological perspective of classic crowd psychology.

The starting point for social identity accounts lies in a shift (rather than a loss) of identity when people become crowd members. They begin to perceive themselves in terms of social rather than personal identity. This is at the root of three psychological transformations. First, there is a *cognitive transformation*. Thus, people begin to evaluate their interests and their priorities in relation to group-level beliefs and values (Reicher, 2001). Given the ambiguous and novel circumstances that attend many crowd events, however, this is not a matter of mechanically applying predetermined rules of action. Instead, it is a matter of developing situational norms that translate superordinate categorical beliefs into contextually appropriate behavior. So, there needs to be a process of norm construction, as suggested in Turner and Killian's emergent norm theory. However, critically, only those proposals which are seen as consonant with identity beliefs are liable to become normative and have influence at the collective level. This explains how the patterns of crowd behavior, albeit based on locally elaborated norms, reflect broader belief systems (Reicher, 1984).

Second, there is a *relational transformation*. This derives less from the fact that people start to see themselves as group members and more that they start to see *each other* as fellow group members. They cease to regard those around them as 'other' and more as 'one of us.' This increases the closeness between crowd members: shared identity in the crowd leads to greater trust, respect, cooperation, and mutual support – for a review, see Reicher and Haslam, 2009. Such closeness, combined with a common perspective on the world, allows crowd members to align their efforts, to become more organizationally effective and hence more powerful (Reicher, 2011).

The combination of these two transformations is at the root of a third *emotional transformation*. Many authors, including Le Bon, have focused on the passionate nature of crowds. Durkheim coined the term 'effervescence' to denote the fizzing bubbling excitement of such events. In contrast to Le Bon, though, who considered that heightened emotionality flows from the loss of reason, Durkheim (1912/1995) saw effervescence as flowing from the ability of people to express their ideas in the crowd. This latter approach is very much in-line with social identity accounts. Thus both the development of a common perspective and the development of intimate and empowering social relations between crowd members are in themselves a source of positive emotions. However, in combination, these two factors enable people to enact their collective worldview, to translate beliefs and priorities into lived social practice – even against the opposition of others. It is this, which has been termed 'collective self-objectification' (Drury and Reicher, 2009), which is a crucial source of positive affect (Reicher, 2011).

Far from being mindless and meaningless, crowds are one of the few contexts in which people have both the inclination and the ability to express their social identities in full (Reicher, 2001). Far from being peripheral to everyday social process, crowds provide a privileged site in which to explore the identities that underpin our everyday lives. However, to leave the analysis at that is to miss three crucial things about crowds. First, thus far we have discussed crowds and collective action in terms of a set of people who share a common identity. However, many collective gatherings may involve people who lack shared identity (say shoppers in a high street) or else a number of different groups sharing different identities (say fans of different clubs at a football match). Second, we have examined how collective action is based on the nature of the relevant social identities. However crowds do not just reflect existing social identities and social realities, but they are also a source of change both at individual and social levels (Dixon et al., 2012). Third, our discussion has focused on intragroup processes – the transformations which happen *within* crowds. Yet, as the discussion of collective self-objectification has intimated, crowd events are characteristically intergroup affairs.

As we shall show next, the elaborated social identity model of crowds (ESIM; Drury and Reicher, 2009) seeks to connect these three elements by analyzing how categories are formed and how identities change through the intergroup dynamics of crowd events.

Intergroup Dynamics and Social Change: An ESIM of Crowd Action

As we have seen, different approaches to crowds turn on different conceptions of the nature and dynamics of selfhood – is self-lost in the crowd (Le Bon), accentuated in the crowd (Allport), or else shifted from the personal to the social level? Likewise, the ESIM begins by elaborating the very notion of social identity that underpins social identity analyses of collective action. That is, the construct is defined as a model of our position in a set of social relations along with the actions which are proper and possible in that position. Social identity, then, is a rich representation of social reality. To define oneself as black, for instance, is to see the world in terms of racialized categories, to be subordinated and oppressed, and to be constrained in one's ability to flourish by the acts of powerful white individuals and institutions.

This definition makes the definition of identity into an interactive accomplishment. Our social identity depends not only upon the way we see ourselves but also how we are positioned by others. Most of the time, the way we see ourselves and the way we are seen by others coincide. Hence our interactions consolidate our social position and our sense of self. However, there are times in some crowd contexts where there is an asymmetry between auto- and hetero-perceptions. When that happens we can see dramatic changes in the social identities of crowd members (Drury and Reicher, 2009).

One example of such change occurs in heterogeneous crowds which consist of a plurality of social groups some of whom see themselves as positively aligned with the authorities

in general and the police in particular, others who see themselves as in an antagonistic relationship to the police and authorities. To the extent that the police themselves see the entire crowd as antagonistic and to the extent that they use their power to *act* on this perception (for instance, using coercive tactics to contain or disperse the crowd as a whole), then a series of consequences are set in motion.

First, those in the crowd who previously saw themselves as aligned with the authorities are repositioned as antagonistic and hence begin to redefine themselves, their relationship with the authorities and with fellow crowd members. They begin to see the police as an oppressive out-group acting in illegitimate and oppressive ways. Next, whereas the crowd as a whole was previously segmented into separate groups, their common positioning by the police leads to the emergence of a sense of shared identity and hence the sense of solidarity and empowerment described above. This sense of empowerment enables crowd members to challenge what they see as oppressive crowd action. Finally, the actions of the crowd confirm and consolidate the original police perceptions of the crowd as a homogeneous danger. In this way, a spiral of escalating radicalization in the crowd and of escalating conflict between crowd members and the police is set in motion (for analyses of such dynamics, see Stott and Drury, 2000; Stott et al., 2001).

As can be seen, this dynamic sets in train a whole series of changes. Most obviously, it changes participants' sense of their social position and hence their social identity. As a consequence, it changes the way in which people see their relationship with others – not just the police and authorities. Most immediately, members of the crowd who might previously have been seen as 'other' come to be seen as fellow crowd members. Thus Drury charts how an antiroad building campaign was initially split between locals and seasoned environmentalist campaigners. Locals viewed the environmentalists with suspicion, as unrespectable and as outsiders. However common experiences of harsh policing led to the formation of a single category and to intimate in-group relations between erstwhile opponents (Drury and Reicher, 2009). Such recategorization was not limited to the actual participants in the 'local' campaign. More generally, locals came to see others involved in antiroad campaigns around the country and in conflict with the authorities as in a similar social position to themselves and hence as part of a broad 'oppositional' category. As a result, support and solidarity for their struggles developed. The emergence of more global categories – both within the specific campaign and more generally – went hand in hand with a greater sense of empowerment and hence a greater confidence in taking action (Drury and Reicher, 2009).

Willingness to take action was not just a matter of empowerment. It is also derived from a reconceptualization of what the campaign was all about and hence a change in understanding as to what constituted 'success.' Whereas, to begin with, the locals saw the campaign as a matter of convincing reasonable authorities to protect their community, as they began to see the authorities as unreasonable and antagonistic, so they saw the mere act of resistance and exposing the nature of the State as a crucial achievement in the longer-term campaign. Hence, even if the road was built through the community, the campaign could still be seen as part of a longer-term success in terms of changing government roads policy.

In detailing these various dimensions of change, ESIM does not suggest that crowds always – or even often – result in change. To the contrary, the model specifies stringent conditions which must be satisfied for change to occur: (1) there must be an asymmetry between the self-perceptions of crowd members and the hetero-perceptions of observers; (2) the observer group must have and use their power to impose their perspective on crowd members and hence reposition them in practice. What is more, for the change process to become a self-reinforcing cycle, (3) as a result of observer action and recategorization within the crowd, members must in turn feel empowered to impose their new perspective upon the observer group. ESIM thereby helps to explain both when change does and does not occur. The model addresses the potential for change in and through crowds – a crucial requirement for any adequate psychology of collective action.

Equally, ESIM does not suggest that change, when it does occur, always goes in the same direction and leads to increased antagonism or radicalization. It follows from the model that, where crowd members see police and authorities as an antagonistic out-group but are seen and treated by the police as being respectable and reasonable members of the general community, so there may be shift in identity and action toward inclusion and harmony (Stott et al., 2008).

Finally, ESIM, although specifically developed through the study of crowds, also addresses a whole series of issues which are central to psychology in general: the nature of identity, our perceptions and evaluations of in-groups and others, and our sense of efficacy and its relationship to action. Irrespective of the particular merits or demerits of the model itself it recognizes the crowd as a particularly valuable site in which to study the interactive processes which underpin social understanding and social action.

The Continuing Exclusion of Collective Action

Conformity Bias and the Neglect of Collective Resistance

In recent years, there have been some signs of a growing interest in collective action in social psychology (e.g., van Zomeren and Klandermans, 2011). However, first, the majority of the work concerns the factors that lead people to participate in various forms of protest action and not in the nature or dynamics of what happens when people do actually gather together. Second, for all its growth, collective action remains a niche topic, rarely if ever invoked when more central issues in the discipline (such as intergroup relations, prejudice, and stereotyping) are addressed.

Part of the reason for this can be put down to a strong 'conformity bias' which has marked and continues to mark social psychology (see Moscovici, 1976). This suggests that people generally follow those with status and power. It implies that social understandings are consensual rather than contested. It rules out a study of how people challenge and resist dominant points of view. As we shall see, it is through the neglect of resistance that the importance of collective action is likewise obscured.

Perhaps the most powerful examples of conformity bias can be found in two of the best known studies in all of psychology, Milgram's 'obedience' studies (Milgram, 1974) and Zimbardo's Stanford Prison experiment (Haney et al., 1973). The former is

generally represented as showing that people will obey the orders of an authority figure, no matter how extreme. The Zimbardo experiment is popularly represented as showing that, even without overt orders, people 'naturally' slip into their assigned roles: those assigned to the position of a guard in a simulated prison setting rapidly became brutal. Those assigned to the role of a prisoner became passive and increasingly disturbed.

Yet, in contrast to the received accounts, closer inspection tells a very different picture. Both studies are replete with examples of resistance and of people rejecting orders and roles (see Smith and Haslam, 2012 for detailed critiques of both studies). Obedience varies from 0% to almost 100% in different versions of Milgram's studies, so it is very far from automatic. Equally, in Zimbardo's study, many guards resisted their role, refusing to be brutal and even siding with the prisoners. The prisoners also resisted their allotted position and challenged the guards throughout the study.

What is more, in both studies, the resistance was most likely and most effective when people had collective support. In Milgram's studies, only 10% of participants went up to 450v when they were in the company of two others who dissented. In Zimbardo's study, the dissent was greatest at the start because the prisoners acted in a coordinated manner. This analysis is supported by the results of the BBC Prison Study which demonstrated that the extent to which participants dissented from and were willing to challenge the system was a function of their degree of shared identification with and expected support from fellow group members (Reicher and Haslam, 2006).

Of course, it is arguable that the degree of resistance that was found reflects the artificial nature of such studies and that, in the real world, resistance is rarely found in prisons and other situations of extreme inequality. But in fact dissent and even prisoner control is easily found in even the most extreme carceral (see Haslam and Reicher, 2012 for examples). In the real world, as in psychological studies, such resistance is dependent upon the development of shared identity, mutual support, and effective coordination among the subordinated and oppressed. All in all, the possibility of resistance is always present, however repressive the defense of the status quo. And the emergence of resistance is characteristically bound up with the emergence of collective action. To privilege conformity and stability in everyday life over resistance and change is therefore to obscure the incidence and the significance of collective action in the formation and reformation of groups and intergroup relations.

The Decollectivization of Social Change

It would be wrong, however, to suggest that mainstream social psychology ignores social change. In the period since World War II, the question of prejudice and how to reduce it has been at the very core of social psychology. The dominant approach to the issue of prejudice reduction has been work on the effects of contact between groups. Drawing upon Gordon Allport's foundational text (Allport, 1954), the question has been how and under what conditions contact (notably between black and white people) leads to more positive attitudes. By now there is an overwhelming amount of evidence that contact will generally lead to greater liking between dominant and subordinate group members (Pettigrew and Tropp, 2006). Such findings have led researchers to conclude that contact is the royal road to change our unequal society. But there is a problem with this work. It does not lie so much in the well-documented link between contact and liking, but in the assumption of a link between liking and social change. Indeed there are a variety of reasons for thinking that liking may not improve relations of social inequality (Dixon et al., 2012).

There is ample evidence that dominant group members can like their subordinates without believing that they should be equals or indeed supporting moves toward equality (Dixon et al., 2012). This is most obvious in the case of gender relations: men can like women while treating them as inferiors – something that is encapsulated in the notion of 'benevolent sexism' (Glick and Fiske, 2001). However, even in more extreme cases, like that of slavery, there is evidence that many owners looked kindly on their slaves and felt that ownership protected inferior creatures from the depredations of the free-market. Thus, they could depict slavery as "a blessing to both master and slave" (Fox-Genovese and Genovese, 2005: p. 515). However such fondness lasted only as long as slaves accepted their place. Should they challenge it, benevolence would rapidly change to anger and brutality. But despite that, there were multiple challenges and indeed slave revolts were critical to the eventual abolition of slavery itself.

The same is true for most entrenched forms of inequality. The end of apartheid, the introduction of civil rights to African-Americans, and the granting of votes to women, all were a result of collective action. To cite Barack Obama's speech in Selma on 4 March 2007 "I'm here because someone marched." It is hard to think of circumstances where equality has been freely granted through the individual kindness of the dominant group members. On the whole it has been claimed through the collective mobilization of subordinate group members. That is why it is particularly worrying that contact may have a demobilizing effect on them. Thus there is growing evidence that increasing contact and increasing the extent to which subordinate group members like dominant group members may actually decrease the extent to which the former will demand and agitate for change in part at least because it reduces their perception of racial discrimination and increases their perceptions of social mobility (Wright and Lubensky, 2009).

So, even where social psychology does address issues of social change, collective action tends to be neglected. This is not just a conceptual issue. The focus on psychological rather than sociostructural outcomes leads to the employment of strategies that demobilize people and therefore leave structural inequalities in society intact.

Conclusion

In this article we have considered both how crowds and collective action have been studied in social psychology and also how they have been neglected. We have encountered evidence to show both that the study of collective action allows us to understand the nature of social action and social change, and also that collective action is critical to achieving social change. Our conclusion is that a social psychology which

marginalizes collective action will be both intellectually deficient and socially conservative. For both reasons, we reiterate that the study of collective action needs to be moved from the periphery to the core of the discipline.

See also: Collective Action; Deindividuation, Psychology of; Intergroup Relations; Political Psychology; Self-Categorization Theory; Social Identity in Social Psychology; Social Movements: A Social Psychological Perspective; Social Psychology.

Bibliography

Allport, F., 1924. Social Psychology. Houghton Mifflin, Boston.

Allport, G.W., 1954. The Nature of Prejudice. Doubleday, Garden City, NY.

Barrows, S., 1981. Distorting Mirrors. Yale University Press, New Haven.

Davis, N.Z., 1978. The rites of violence: religious riot in sixteenth century France. Past and Present 59, 51–91.

Dixon, J., Levine, M., Reicher, S.D., Durrheim, K., 2012. Beyond prejudice: are negative evaluations the problem and is getting us to like each other more the solution? Behavioural and Brain Sciences 35, 411–425.

Drury, J., Reicher, S.D., 2009. Collective psychological empowerment as a model of social change: researching crowds and power. Journal of Social Issues 65, 707–725.

Durkheim, E., 1912/1995. The Elementary Forms of Religious Life (K.E. Fields, Trans.). Free Press, New York.

Fox-Genovese, E., Genovese, E.D., 2005. The Mind of the Master Class. Cambridge University Press, New York.

Giner, S., 1976. Mass Society. Martin Robertson, London.

Glick, P., Fiske, S.T., 2001. An ambivalent alliance: hostile and benevolent sexism as complementary justifications for gender inequality. American Psychologist 56, 109–118.

Haney, C., Banks, C., Zimbardo, P., September 1973. A study of prisoners and guards in a simulated prison. Naval Research Review 1–17. Office of Naval Research, Washington, DC.

Haslam, S.A., Reicher, S.D., 2012. When prisoners take over the prison: a social psychology of resistance. Personality and Social Psychology Review 16, 154–179.

Le Bon, G., 1895/1947. The Crowd: A Study of the Popular Mind. Ernest Benn, London.

Milgram, S., 1974. Obedience to Authority: An Experimental View. Harper & Row, New York.

Moscovici, S., 1976. Social Influence and Social Change. Academic Press, London.

Moscovici, S., 1980. L'Age des Foules. Fayard, Paris.

Nye, R.A., 1975. The Origins of Crowd Psychology. Sage, London.

Pettigrew, T.F., Tropp, L.R., 2006. A meta-analytic test of intergroup contact theory. Journal of Personality and Social Psychology 90, 751–783.

Reicher, S.D., 1984. The St. Pauls' Riot: an explanation of the limits of crowd action in terms of a social identity model. European Journal of Social Psychology 14, 1–22.

Reicher, S.D., 2001. The psychology of crowd dynamics. In: Hogg, M., Tindale, S. (Eds.), Blackwell Handbook of Social Psychology: Group Processes. Blackwell, Oxford, pp. 182–208.

Reicher, S.D., 2011. Mass action and mundane reality: an argument for putting crowd analysis at the centre of the social sciences. Contemporary Social Science 6, 433–450.

Reicher, S.D., Haslam, S.A., 2006. Rethinking the psychology of tyranny: the BBC Prison Study. British Journal of Social Psychology 45, 1–40.

Reicher, S.D., Haslam, S.A., 2009. Beyond help: a social psychology of social solidarity and social cohesion. In: Snyder, M., Sturmer, S. (Eds.), The Psychology of Prosocial Behaviour. Wiley-Blackwell, Oxford.

Reicher, S.D., Spears, R., Postmes, T., 1995. A social identity model of deindividuation phenomena. European Review of Social Psychology 6, 161–198.

Reicher, S.D., Stott, C., 2011. Mad Mobs and Englishmen? Myths and Realities of the 2011 Riots. Constable & Robinson, London.

Reiwald, P., 1949. De L'Esprit des Masses. Delachaux & Niestle, Neuchatel.

Smith, J., Haslam, S.A., 2012. Social Psychology: Revisiting the Classic Studies. Sage, London.

Stott, C., Adang, O.M., Livingstone, A., Schreiber, M., 2008. Tackling football hooliganism: a quantitative study of public order policing and crowd psychology. Psychology, Public Policy and Law 52, 111–138.

Stott, C., Drury, J., 2000. Crowds, context and identity: dynamic categorization processes in the 'poll tax riot'. Human Relations 53, 247–273.

Stott, C., Hutchison, P., Drury, J., 2001. 'Hooligans' abroad? Inter-group dynamics, social identity and participation in collective 'disorder' at the 1998 World Cup Finals. British Journal of Social Psychology 40, 359–384.

Turner, J.C., Hogg, M.A., Oakes, P.J., Reicher, S.D., Wetherell, M., 1987. Rediscovering the Social Group: A Self-categorization Theory. Basil Blackwell, Oxford.

Turner, R.H., Killian, L.M., 1972. Collective Behaviour, second ed. Prentice-Hall, Englewood Cliffs, NJ.

US Riot Commission, 1968. Report of the National Advisory Commission on Civil Disorders. Bantam, New York.

Wright, S.C., Lubensky, M., 2009. The struggle for social equality: collective action vs. prejudice reduction. In: Demoulin, S., Leyens, J.P., Dovidio, J.F. (Eds.), Intergroup Misunderstandings: Impact of Divergent Social Realities. Psychology Press, New York, pp. 291–310.

van Zomeren, M., Klandermans, B., 2011. Towards innovation in theory and research on collective action and social change. British Journal of Social Psychology. Special Issue: Innovation in Theory and Research on Collective Action and Social Change 50 (4), 573–574.

van Zomeren, M., Leach, C.W., Spears, R., 2012. Protestors as 'passionate economists': a dynamic dual pathway model of approach coping with collective disadvantage. Personality and Social Psychology Review 16, 180–199.

Collective Behavior, Sociology of

Neil J Smelser, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.L. Wood, volume 4, pp. 2204–2208, © 2001, Elsevier Ltd.

Abstract

The article begins with a selective summary of contributions to the study of collective behavior by four historical figures: Charles Mackay, Gustave Le Bon, Alexis de Tocqueville, and Karl Marx. They are considered both as figures on their own and, taken together, as expressing dilemmas that have characterized the field throughout its history. These dilemmas are scope of the subject matter of collective behavior, psychological versus social explanations, the role of 'irrational' forces, and analytic versus evaluative assessments of collective behavior. The remainder of the article reviews the past half-century, both on its own and as expressing the ongoing dilemmas. The 'approaches' summarized are 'collective dynamics,' emergent norms, Smelser's general theory, resource mobilization theory, contentious politics, opportunity structure, framing and frame alignment, and the role of emotions in explanation. The article closes with a few summary observations on diverse tendencies that have characterized the development of the sociology of collective behavior.

Four Classical Approaches

It is instructive to consider a number of historical figures who made the field of collective behavior a preoccupation. These are Charles Mackay, Gustav Le Bon, Alexis de Tocqueville, and Karl Marx. To my knowledge, these figures have not been grouped together before. I do so because of their importance for the genesis of the field and because, together, their works lay down surviving emphases and generate several dilemmas in scholars' descriptions and explanations of collective behavior.

Charles Mackay (1812–89)

One of the most dramatic if not sensational treatments of collective phenomena is found in a book by this nineteenth-century Scottish poet, songwriter, author, and anthologist. He called it *Extraordinary Popular Delusions and the Madness of Crowds* (2012 [1841]), a title that reveals much about the book's tone. He identified his focus as nations' and individuals' "whims," "peculiarities," and "excitement and recklessness" (2012 [1841], Preface). The book consists of colorful historical descriptions of an incredible array of phenomena including financial crazes, alchemy, prophecies, the Crusades, witchcraft, dread of haunted houses, speech fads, popular adoration of thieves, dueling, and ordeals. Reading the book 170 years later, one remains glued to its stories, its wit, and its ironical observations.

We should not hold Mackay to standards of scientific explanation and evidence that were not available to him. His time was mainly prescientific, except for nineteenth-century economics and psychology. Yet if we examine his accounts from the standpoint of scientific desiderata, they are wanting in several respects, which will now be described.

Delineation of Subject Matter

Mackay appeared to include anything he found in the historical record that kindled his interest, with the result that his methodology is that of a historical vacuum sweeper. The boundaries and subdivisions of his subject matter – called for in scientific explanations – are absent, except for their apparently common characteristics of folly and madness.

Interest in Variation

Because Mackay regarded the tendency to madness to be basic to human nature, the fundamental reason for their occurrence of follies appears to be that they express this. This yields an unsatisfactory scientific handle, because all science rests on the explanation of variation in phenomena. To treat them as recurrent expressions of a universal human propensity cannot discriminate between their occurrence and nonoccurrence or other variations.

Although this diagnosis is generally correct, a few observations in the volume suggest not only variations but also crude explanatory mechanisms. Mackay observed, for example, that "Men … go mad in herds, while they only recover their senses slowly, and one by one" (Preface), presaging Le Bon's later emphasis on the independent power of the group. Another instance: "Credulity is always greatest in times of calamity" (p. 144). This is a version of deprivation theory (cryptic, to be certain), at least noticing the problem of why collective delusions cluster more in some circumstances than others. As a general rule however, Mackay's descriptions are devoid of what we would now call causal forces or mechanisms.

Scientific Neutrality

Mackay's work reveals no discernible line between his descriptions of his chosen phenomena and his moral stance toward them. As such they are simultaneously both objective and silly, delusional, misguided, and often dangerous. In the generations since Mackay we have come to objectify such phenomena and insulate them from praise or blame, but as will become evident later, the field of collective behavior has had an impossible time of escaping moral and political preferences as well as conflicts about its subject matter, particularly its apparent irrationality and emotionality.

Gustave Le Bon (1841–1931)

In 1895 a small book, called *The Crowd* (2001 [1895]), was written by this French social psychologist, sociologist, and amateur physicist. It is probably the most influential work in the history of the study of collective behavior. Part of its impact lay in how he framed his study; not only did he advertise his work on crowds as established psychological knowledge, but he also used the most extreme psychological language to describe them. Moreover, he related his work to the sweep of history, proclaiming contemporary times to be an "era of crowds" (p. x). He attributed this era to the rise of democracy and the demise of stabilizing religious and political traditions (dramatized by the French Revolution). This alone assured that the book would be the source of continuing controversy.

Le Bon dwelt on the characteristics of the crowd mind: mentally inferior, incapable of reason, impulsive, fickle, suggestible, credulous, incapable of compromise, subject to illusions, intolerant, dictatorial, and amoral (but capable of both moral degradation and lofty morality). Above all this mentality was unconscious, reflecting in Le Bon's work the late nineteenth- and early twentieth-century preoccupation with unconscious forces. All these characteristics are not to be found in individuals; they are a product of the crowd as such (Le Bon's version of the 'group mind'), and the immediate causes are found in the mechanisms of a sense of omnipotence, contagion, and suggestion. Reflecting a bias of Victorian times, Le Bon singled out primitives, Latin races, women, and children as especially susceptible to crowd-like thinking.

Among the 'remote' (i.e., historical) factors that gave rise to the era of crowds were a people's racial characteristics, the destruction of traditions, and the decay of regulating political and social institutions, all of which check the formation and activity of crowds. Consistent with this diagnosis, he singled out juries, election campaigns, and parliamentary assemblies – all associated with democracies – as outright 'crowds,' demonstrating the same excesses.

For half a century *The Crowd* and its adaptations dominated psychoanalytic group psychology (Freud, 1955 [1921]), social psychology (e.g., Trotter, 1916), and sociology (echoing into Park and Burgess, 1924; Blumer, 1951). Since World War II its influence has continued in sociology, though mainly in a negative sense, in that developments external and internal to the field shifted its emphasis in more rational directions, with Le Bon as its principal casualty.

From a methodological point of view, Le Bon's treatment of crowds also suffers from indistinct and sprawling conceptual boundaries and a looseness of causal explanations, though he did organize multiple causes into remote and immediate, and social and psychological. Unlike Mackay, moreover, he proclaimed his method to be scientific, that is, not influenced by "opinions, theories and doctrine," even though his topics generated 'impassioned controversy.' Though he betrayed this outlook by flooding his work with gross moral and political judgments, it is notable that he also appealed to the logic of science and dispassion as legitimizing criteria.

Alexis de Tocqueville (1805–59)

More than 20 years after his immortal *Democracy in America*, Tocqueville developed an account of the French Revolution, under the title, *The Old Regime and the French Revolution* (1945 [1856]). In his analysis he also distinguished between remote and more recent causes. Among the former he specified (1) the tendency for governments to centralize; which led to (2) the partial deterioration of feudal institutions – retention of privileges but loss of authority; (3) at the same time, certain classes, notably the bourgeoisie and the peasants, had gained in wealth, but were excluded from other advances within the social order; this produced widespread feelings of relative deprivation in those classes; this produced (4) an isolation and conflict among groups; which was aggravated by (5) a general tendency of group members to imitate one another; out of these forces grew (6) a general ideology calling for wholesale social change, based on equality and simplicity; this kind of thinking was also an expression (7) of a 'French temperament' that favored total diagnoses of society and comprehensive programs of change. As more proximate causes, Tocqueville mentioned repressive activities and hasty, ill-conceived reforms by the government that affected timing and direction in the actual revolutionary overthrow.

Though ad hoc in appeal to causes, Tocqueville's study reveals several special features. He relied on an abiding distinction between longer term predisposing causes and immediate precipitants. He developed a multicausal analysis that included social–structural (e.g., government, relations among classes), psychological (e.g., relative deprivation), even political (state aggrandizement) and cultural ('French temperament') forces. Finally, though Tocqueville is generally heralded as a 'liberal' in political preference, his analysis of the French revolution is remarkably dispassionate, free of intemperate value judgments and emotional reactions. His summary of the accumulation of causes was that the revolution was "a foregone conclusion."

Karl Marx (1818–83)

As with Tocqueville, we note a specific social movement central to Marx's economic and social theory – the rise of a worker revolution that will destroy capitalism as a system and usher in a propertyless and stateless society, the communist system.

Marx's ultimate explanation of the proletarian revolution traces to fundamental economic principles, specifically the labor theory of value and its derivatives. The highly elaborated theory's essence is that the amount of labor imparted to a commodity determines its value. In the organization of capitalist production, it becomes possible, even necessary for capitalists to enter this process and expropriate a portion of this value, in the form of what Marx calls "surplus value" by using strategies such as lengthening the working day and using female and child labor, but above all by gaining technological advantage in the competitive capitalist system. This extraction is translated directly into profits, and more importantly, the extraction of surplus value is the principal instrument and determinant of labor exploitation. This ever-increasing exploitation is the engine for worker reaction and ultimately for their destruction of the capitalist system through violent revolution.

In one sense Marx's theory of revolution is mechanical, reflecting the immutable forces of capitalist production. However, he developed a kind of 'learning theory' whereby workers progress from a condition of false consciousness of

their class position to a truly revolutionary consciousness. In the early stages, workers are isolated, in competition with one another, and disorganized; they are capable of only isolated forays against the means of production (for example, machine breaking) and misguided attacks against the remnants of the feudal ruling classes. As workers gather into greater proximity with one another, and as the advance of industry obliterates their differences, they become more mobilizable into collective action groups such as trade unions, and are able to compel their recognition and attain legislative victories such as limitations on hours. Finally, as the contradictions of the capitalist system worsen, a revolutionary party is formed, consisting of the proletariat and that part of the bourgeoisie that has been driven into the ranks of the workers. That party ultimately overthrows the capitalist system.

I include Marx – seldom considered a scholar of collective behavior – among the illustrative historical scholars for two reasons. First, his explanation of revolution is deterministic in the extreme, relying on the unfolding of the economic laws of capitalist society. Human agency plays a role in mobilizing for and executing the revolution, but that agency is more profoundly determined by structural forces. Second, the emphasis is overwhelmingly macroscopic, ignoring any special 'mentality' of participants in the worker movement beyond its progress toward increasing class consciousness.

Enduring Dilemmas Posed by These Historical Accounts

Assuming it is a legitimate exercise to compare and contrast these historical figures, it is also appropriate to analyze several dilemmas that their work – considered together – poses for the study of collective behavior. These dilemmas, moreover, extend to the present and impart understanding to the directions, strengths, limitations, and ambiguities of this subfield of sociology:

Dilemma 1

What subject matter is covered by the term, 'collective behavior'? Specifically, does it refer to episodic collective outbursts or does it include social movements? Most treatments have avoided this classificatory dilemma by creating open-ended categorical lists including panic, craze, stampede, crowd ecstasy, mass hysteria, mass delusion, riot, lynching, fad, fashion, spread of collective mood, social protest, reform movement, revolutionary movement, religious cult, and religious revival. Some have argued that there is a continuity among collective behavior and social movements and have grouped them together; others find that the two classes of social behavior are separate beasts, to be explained separately. The general confusion over proper terrain has also produced diverse generic terms – mass phenomena, mass movements, mass behavior, collective action, and collective dynamics, as well as lesser candidates such as collective moods and opinion drifts – none of which have 'stuck.'

There is no clear resolution of this dilemma, because inclusion and exclusion vary according to the defining characteristics assigned to the diverse phenomena. Empirical reality alone does not dictate a solution; the mode of conceptualization also plays a role. If spatial contiguity and face-to-face interaction are evoked as defining characteristics, for example, generic differences between episodic outbursts and social movements are evident; if common emotions or belief systems are the defining characteristics, then more similarities may be found. In this article I include reference to social movements as well as collective outbursts for two reasons: first, in the history of investigations they have sometimes been lumped together; and second, despite the development of separate strands of research in recent decades, the two classes of phenomena raise common issues of internal dynamics, explanation, and consequences.

Dilemma 2

The two most common explanatory frameworks for collective-behavior phenomena are the psychological and the social (although cultural factors have appeared and political explanations have surged in recent decades). In principle the distinction between the psychological and the social is straightforward: psychological refers to internal orientations of actors, perhaps intensified in groups through processes of social interaction but still psychological in essence; sociological explanations rely on social–structural dimensions, including cleavages, dislocations, and malintegrations in the larger society. In practice, however, these two types of explanation have been so fused in emphasis that is impossible to separate them neatly and without distortion.

In this article I bow to the realities of history, which has refused to honor any neat or consistent separation of psychological and sociological approaches to collective behavior, however defined and explained. Different research traditions stress one aspect or the other, of course, and these stresses have different implications, as we will demonstrate. But to honor the complexity of this field, it is obligatory to include both levels of classification and explanation, as well as others.

Dilemma 3

Insofar as causes are psychological, are the driving forces conceived as irrational (real, based on distortion, blind, misdirected) or rational (reasonable, calculated, deliberative, intelligible, or right because they reflect the laws of history)?

Dilemma 4

The phenomena identified as included in collective behavior themselves generate – perhaps universally – strong emotional and evaluative reactions among those who participate in such behavior, those who witness it, and those who study it. As a result, social observers, including social scientists, forever allow evaluative reactions – open or subtle – to become a part of their assessments. These reactions are emotional, moral, and political, or some mix of these. The history of the field, in consequence, has been one of bias imparting, bias exposing, bias denying, denying of bias denials, and rebiasing, despite contravening scientific dictates to be objective and dispassionate. Other sociological specializations – family, class, religion – reveal the same tendencies, but collective

behavior seems a strong competitor for first place in this questionable race.

Historical Threads during the Past Half-Century

For reasons not evident, the years surrounding 1960 produced several general, ambitious treatments of collective behavior. (At the beginning of that decade, Lang and Lang (1961, p. vi) declared that up to that time the field of collective behavior had "somewhat fallen into disrepute" and "had not kept pace with developments in other fields.") Taken together, these works simultaneously constitute shifts in emphasis, innovation, and consolidation. The first was a general work on social processes, by Turner and Killian (1957). Accepting the premise of the Chicago school (especially Blumer, 1951) that collective behavior arises from a breakdown of social control and integration, and itself is an unstructured, noninstitutionalized expression, they developed a more positive theme whose influence has persisted: collective behavior is a *process* in that in its course of development it develops emergent and increasingly explicit norms that shape the goals, strategies, tactics, and envisioned outcomes of any given episode or movement. In this way Turner and Killian, while retaining residues of the 'irrationalist' elements that survived in the Chicago school, added a more purposive, positive, and creative element. A second major book by Lang and Lang (1961), a synthetic work, emanated from the Chicago school as well. It was mainly social–psychological in emphasis, though acknowledging social–structural background influences; it also was influenced by the mass-society and mass-behavior perspectives, salient at the time. Their book was widely recognized as a valuable comprehensive statement, but was not received as innovative, and has not persisted in influence. However, it was sufficiently visible at the time that it earned them the authorship of the entry on collective behavior in the *International Encyclopedia of the Social Sciences* later in the decade (Lang and Lang, 1968).

The third work, by Smelser (1962), involved a self-conscious shift from the social–psychological tradition of collective behavior to a more explicit social–structural or sociological direction. He defined collective behavior as "mobilization on the basis of a belief that redefines social action," and made an effort to integrate collective outbursts such as panics, crazes, and hostile outbursts as well as reform and revolutionary movements under this definition. He developed a comprehensive "value-added" explanatory framework that conceived episodes of collective behavior as multiply determined by a narrowing series of conditions (causes): structural conduciveness, or the possibility (opportunity) for such episodes; structural strain (which encompassed a range of deprivations, malintegrations, and ongoing social cleavages); growth and spread of a generalized belief; precipitating factors; mobilization of participants for action; and responses to collective actions by agencies of social control. His emphases on conduciveness, mobilization, and agencies of social control survived into subsequent, more politically oriented approaches. Smelser himself eschewed psychological definitions of collective behavior, and in particular, implications that "persons involved … are irrational, that they lose their critical faculties, that they experience psychological regression, that they revert to some animal state" (Smelser, 1962, p. 11). However, in his exposition of generalized beliefs, he identified a tendency for these to "short-circuit" from generalized hopes and aspirations into specific promises and solutions (suggesting, perhaps, a magical element), and, used the colorful but unhappy phrase "action of the impatient" to describe episodes of collective behavior. Many critical reactions to his work, developed in the years after the book's appearance, seized on these descriptions and consigned Smelser's work to the irrationalist (and implicitly politically conservative) tradition.

In the late 1960s the study of collective behavior took a decisive turn. The origins of this emanated from the larger society, as movement after movement cascaded through the nation – civil liberties, violent urban protests, student activism, antiwar agitation, affirmative action, new activity in one minority group after another, women's liberation, and gay rights. It was a decade-and-a-half of ferment sandwiched between the quiescent Eisenhower years and the conservative Reagan years. Most of the movements came from leftish territory and spoke on behalf of the disprivileged. Accordingly, they excited the sympathy and support of many social scientists, already leftish in their political leanings. Sociology, long attuned to social problems and social inequalities, was perhaps most affected.

In the study of collective episodes and social movements, the 'turn' was a complex psychological and political one. In general attention turned *away* from the psychological characteristics of participants, but in particular it evolved toward a psychology that involved purposive and reasonable if not rational pursuit of legitimate goals. This augmented representation of rationality informed the turn of the 1970s, and in a few cases (Olson, 1965; Oberschall, 1973), stronger assumptions deriving from models of economic rationality were adopted. The turn was political in four senses:

- Political variables began to take over (the brief definitional flurry over whether the Watts disturbances were 'riots' or 'protests,' and the continuing ambivalence toward the term 'riots' remains). (1) Movements were treated as purposive efforts to mobilize resources such as financial support, member commitment, the media, personal networks, and supportive groups (McCarthy and Zald, 1977). One consequence was a shift away from a focus on causes of social movements toward a focus on reasons for their success or failure. (2) The 'opportunity structure' for movements – meaning mainly political opportunities and constraints (e.g., Tarrow, 1983) – gained salience. (3) The political emphasis reached a kind of apex with the treatment of collective phenomena as 'contentious politics' (McAdam et al. (2001). This last effort appeared, in effect, to treat much of what had been considered collective behavior as part of the political process. In these reorientations the emphases on participant motivation, social psychology, ideology, culture, and affective dimensions as characteristics and/or determinants gravitated toward the invisible (Chazel, 2001).
- Scholars and teachers in the field became politically friendlier toward uninstitutionalized protest, even though

this shift in affect was often muted or disguised in the face of continuing force of the importance of scientific dispassion in social science inquiry.

- In the profession and in university and college curricula the study of collective behavior and social movements assumed greater salience, manifested in particular by the formation of a separate section on Collective Behavior and Social Movements in the American Sociological Association.
- The 1960s–1970s involved a wholesale assault on almost all past traditions of collective behavior, which were regarded as 'irrationalist' psychologically, misguided conceptually, and implying a sympathy or identification with the ruling classes, the status quo, the 'establishment,' and authorities. Some of this might be written off as the general tendency of social science innovators to declare the past mistaken and irrelevant, but in this case a major paradigmatic reorientation was accomplished. These multisided political emphases produced the greatest surge of literature in the history of the field – a flood of conceptualization, theory, case studies, comparative research, readers, handbooks, and texts on social movements in the last several decades of the twentieth century. The interpretation of large-scale historical revolutions, highlighted by the work of Skopcol and the continuing influence of Tilly, emerged into a subfield of its own. The study of short-term episodes such as panic, scapegoating, and crazes revealed little innovation, but the occurrence of and preoccupation with disasters (such as 9/11 and Hurricane Katrina, produced bursts of empirical studies and commentaries (Dynes and Tierney, 1994; Brunsma et al., 2010).

As the dominant political themes of the 1970s either became established research traditions or subsided in the face of criticisms, new emphases appeared. Many of these changes, it can be argued, constituted a 'cultural turn,' reflecting wider trends toward multiculturalism, diversity, and more vaguely, postmodernism in the larger society and in many humanities and social science fields. In sociology in particular, this impulse found expression in a heightened preoccupation with 'identity movements,' in which personal cultural self-consciousness and personal dignity are conspicuous (see Reger et al., 2008). These movements mainly include those based on the assertion of racial, ethnic, gender, and sexual preference identities. Emphasis on them paralleled a surge of interest in 'identity politics' in political science and the concept of 'self-identity' in social psychology (see Tajfel, 1982).

In Western Europe a related cultural thread emerged. It went under the name of 'new social movements' (Melucci, 1980; Touraine, 1981). Of course, the adoption of the term 'new' implied an 'old.' In this case the old were the labor and socialist movements, bred by and ranged against capitalist exploitation; these were assumed to have lost their energy and resonance as societies evolved into postindustrial forms. The new were seen as reflecting continuing and new contradictions of postindustrialism, including inequality, domination, and repression. However, they did not stress income and economic security, but rather 'life-world' concerns of personal identity, personal life, neighborhood, environment, gender, and peace, and antiwar themes (Kriesi et al., 1995). Scholars' enthusiasm for these two cultural approaches appears to have subsided from a torrent to a more routine preoccupation.

A third development in late twentieth century, also in many ways 'cultural,' goes under the name of 'framing.' Invented decades ago, and tracing ultimately, perhaps to W.I. Thomas' definition of the situation, the concept was vitalized in the 1980s in relation to social movements (Snow and Benford, 1988). Basically, the concept refers to ideas that diagnose issues, envision solutions, motivate participants, provide identity, define agency, generate strategies, and specify grievances for social movements. The concept has become highly elaborated into a subfield and widely applied, and criteria for effective and ineffective framing have evolved. Proponents of the perspective acknowledge its kinship to 'ideology' (Snow and Benford, 1996) and its link to 'culture' (Zald, 1996), but also insist on its uniqueness and advantage. The concept may also be regarded as a return of themes thought to be neglected in the politically oriented approaches of the 1970s.

Though not expressive of a cultural turn, a new, or rather revived interest in the role of emotions in social movements has also appeared, in part critical of the perceived over rationality of recently dominant approaches, in part reflective of the general revitalization of affective dimensions in psychology in recent decades, and in part a reflection of a more general trend reflecting, for example, 'emotions in economics' and 'emotions in organizations.' While remaining generally inchoate, the 'emotions' perspective has revitalized reflexive emotions, affective bonds, moods, and moral reactions as relevant dimensions (Goodwin et al., 2004).

Finally, notice might be taken of a recent instance of the general tendency of the field of collective behavior to take an interest in and give emphasis to particular novelties as they appear in the real world. I refer in this case to the emergence of both interest and enthusiasm about the Internet as an instrument of collective mobilization – especially under repressive regimes that have difficulty in controlling it. The burst of interest is especially evident in the Arab Spring and other Middle-Eastern uprisings. More generally, electronic communication may well be a major redefining force for social movements and politics in general.

Concluding Remarks

The sociology of collective behavior, over its history, has revealed a number of diverse and not altogether consistent tendencies: (1) a stubborn persistence, and in some periods, impressive growth and accumulation of knowledge; (2) a certain dialectic, with competing perspectives, explanations, and theories waxing and waning but none disappearing altogether; (3) unusually high levels of paradigmatic conflict; (4) a struggle between scientific and evaluative impulses; (5) a tendency to 'chase' emerging phenomena in the larger society; and (6) an assured future, given the centrality of its subject matter in the continuing processes of democratization and globalization, both of which nurture uninstitutionalized political protest in the political process.

See also: Civil Rights Movements; Collective Behavior, Social Psychology of; Collective Beliefs, Sociological Explanation of; Collective Memory, Sociology of; Crowds in History; Movements, Social; Social Movements: A Social Psychological Perspective; Tocqueville, Alexis de (1805–59).

Bibliography

Blumer, Herbert, 1951. The Field of Collective Behavior. In: Lee, A.M. (Ed.), Principles of Sociology. Barnes and Noble, New York, pp. 167–222.

Brunsma, David L., Overfel, David, Picou, Steven (Eds.), 2010. The Sociology of Katrina: Perspectives on a Modern Catastrophe, second ed. Rowman & Littlefield., Lanham, MD.

Chazel, F., 2001. Social movements, sociology. In: Smelser, Neil J., Baltes, Paul B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences, vol. 21. Elsevier, Oxford, pp. 14371–14375.

Dynes, Russell R., Tierney, Kathleen J. (Eds.), 1994. Disaster, Collective Behavior, and Social Organization. University of Delaware Press., Cranbury, NJ.

Freud, Sigmund, 1955 [1921]. Group Psychology and the Analysis of the Ego. In: Strachey, James (Ed.), The Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. XVIII. The Hogarth Press, London, pp. 69–143.

Goodwin, Jeff, Jasper, James M., Polletta, Francesa, 2004. Emotional dimensions of social movements. In: Snow, David A., Soule, Sarah S., Kriesi, Hanspeter (Eds.), The Blackwell Companion to Social Movements. Blackwell, Malden, MA, pp. 413–432.

Kriesi, Hanspeter, Koopmans, Rudd, Duyvendak, Jan Willem, Giugni, Marco, 1995. New Social Movements in Western Europe: A Comparative Analysis. University of Minnesota Press, Minneapolis, MN.

Lang, Kurt, Lang, Gladys E., 1961. Collective Dynamics. Crowell., New York.

Lang, Kurt, Lang, Gladys E., 1968. Collective behavior. In: Sills, David L. (Ed.), International Encyclopedia of the Social Sciences, vol. II. The Macmillan Company and the Free Press, New York, pp. 556–564.

Le Bon, Gustave, 2001 [1895]. The Crowd: A Study of the Popular Mind. Dover Publications, Minneola, NY.

Mackay, Charles, 2012 [1841]. Extraordinary Popular Delusions and the Madness of Crowds. Maestro Reprints, San Bernardino, CA.

McAdam, Doug, Tarrow, Sidney, Tilly, Charles, 2001. Dynamics of Contention. Cambridge University Press, Cambridge.

McCarthy, John D., Zald, Mayer N., 1977. The Trend of Social Movements in America; Professionalization and Resource Mobilization. General Learning Press, Morristown, NJ.

Melucci, Alberto, 1980. The new social movements: a theoretical approach. Social Science Information 19, 199–226.

Oberschall, Anthony, 1973. Social Conflicts and Social Movements. Prentice-Hall, Englewood Cliffs, NJ.

Olson, Mancur, 1965. The Logic of Collective Action: Public Goods and the Theory of Groups. Harvard University Press, Cambridge, MA.

Park, Robert E., Burgess, Ernest W., 1924. Introduction to the Science of Sociology. University of Chicago Press, Chicago.

Reger, Jo, Myers, Daniel J., Einwohner, Rachel L. (Eds.), 2008. Identity Work in Social Movements. University of Minnesota Press, Minneapolis, MN.

Smelser, Neil J., 1962. Theory of Collective Behavior. The Free Press of Glencoe, New York.

Snow, David A., Benford, Robert D., 1988. Framing processes, ideology, and discursive fields. International Social Movement Research 1, 197–218.

Snow, David A., Benford, Robert D., 1996. Clarifying the relationship between framing and ideology. In: Johnston, Hank, Noakes, John A. (Eds.), Frames of Protest: Social Movements and the Framing Perspective. Rowman & Littlefield, Lanham, MD, pp. 205–212.

Tajfel, Henri (Ed.), 1982. Social Identity and Intergroup Relations. Cambridge University Press, New York.

Tarrow, Sidney, 1983. Struggling for Reform: Social Movements and Political Change during Cycles of Protest. New York Center for International Studies, Ithaca, NY.

Tocqueville, Alexis de, 1945 [1856]. The Old Regime and the French Revolution. Doubleday Anchor, Garden City, NY.

Touraine, Alain, 1981. The Voice and the Eye: An Analysis of Social Movements. Cambridge University Press, Cambridge.

Trotter, W., 1916. Instincts of the Herd in Peace and War. T. F. Unwin, London.

Turner, Ralph H., Killian, Lewis M., 1957. Collective Behavior. Prentice-Hall., Englewood Cliffs, NJ.

Zald, Meyer N., 1996. Culture, ideology, and strategic framing. In: McAdam, Doug, McCarthy, John D., Zald, Mayer (Eds.), Comparative Perspectives on Social Movements: Political Opportunities, Mobilizing Structures, and Cultural Framing. Cambridge University Press, Cambridge, pp. 275–290.

Collective Beliefs, Sociological Explanation of

Michel Dubois, French National Research Center, CNRS, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The explanation of collective beliefs is a central question for sociologists and generally for all social scientists. If the 'market' of ideas and beliefs is frequently described in terms of individualization, this should not be overinterpreted by asserting that the realm of collective beliefs is over. The dissemination of individual beliefs frequently converges in collective trends that can be sociologically identified and analyzed. We propose to distinguish five main modes of sociological explanation of collective beliefs – positional, interactional, emotional, cultural, and rational – and analyze their conditions of scientific robustness.

The explanation of collective beliefs is a central question for sociologists and generally for all social scientists. Alexis de Tocqueville, Karl Marx, Max Weber, Gabriel Tarde, Emile Durkheim, Vilfredo Pareto, Max Scheler, and Karl Mannheim, to name a few founding fathers of the sociological tradition, all have devoted a lot of time and energy to this question. Beyond their differences of methods and concepts, they share the general idea formulated by Durkheim in his *Elementary forms of religious life* (1912/1995): the explanation of collective beliefs constitutes the center of gravity of the overall sociological discipline. Although of decisive importance for past and contemporary sociologists, this project remains confronted by certain difficulties, some more challenging than others. At first we will review three of the most commonly discussed difficulties. Next, a detour to classical and contemporary sociological studies will demonstrate the diversity of the sociological modes of explanation of collective beliefs. We will focus on five main factors or social mechanisms. Finally we will identify some of the general criteria used by sociologists to evaluate the respective merits of these explanations. We will be interested in the reasons why many contemporary sociologists consider the generalization of the principle of 'symmetry' as representing significant progress in the sociological explanation of collective beliefs.

Why Is It Difficult to Explain Collective Beliefs?

A first difficulty related to the sociological explanation of collective belief is the existence of collective belief itself. At the beginning of the twenty-first century, are there any collective beliefs left to explain by sociologists? It is a well-known fact that the landscape of collective beliefs has dramatically changed since the period of the founding fathers of sociology. The market of ideas and beliefs is nowadays described in terms of balkanization and individualization (Boudon, 1999; Hervieu-Léger, 2010; Bronner, 2011). Whereas past collective beliefs, religious or political, were massive in terms of scale, cohesion, and diffusion, contemporary collective beliefs are subject to a double process of deinstitutionalization and subjectivization. The modern propensity to 'cognitive tinkering' regularly described by sociologists weakens the long-established relation between believing and belonging. Individuals tend more and more to 'believe without belonging' (Davie, 1994), hence a greater variety of social beliefs. This of course should not be overinterpreted by asserting that the realm of collective beliefs is over. The dissemination of individual beliefs frequently converges in collective trends that can be sociologically identified and analyzed. Collective beliefs still exist, but their form and capacity to circulate have obviously changed. New information technologies, such as the Internet, modify the way individuals produce and diffuse their ideas.

A second difficulty regularly associated with sociological explanation of collective beliefs is their extreme diversity. What is the common thread in beliefs related to the competence differences between male and female (Ridgeway and Correll, 2004; Rashotte and Webster, 2005), God's influence in everyday life (Schieman, 2010), origins of poverty (Lee et al., 1990; Hunt, 1996) and social or racial inequality (Kluegel and Smith, 1981; Kluegel, 1990; Krahn and Harrison, 1992; Evans, 1997; Schuman and Krysan, 1999; Bailey, 2002; Galland et al., 2013), life after death (Greeley and Hout, 1999), the existence of Santa Claus (Bronner, 2004) and paranormal phenomena (Boy and Michelat, 1986; Markovsky and Thye, 2001; Renard, 2010), the origins and prevalence of AIDS (Natras, 2013), and personal sensitivity to medications (Horne et al., 2004)? Sociologists have different methods to characterize the state of minds and representations about all kinds of topics, but in what sense can we consider that they all are studying various aspects of the same general object – collective beliefs?

One way to answer this question is to go beyond appearances and consider what is common to all of these peculiar objects. Whatever the topic, the term 'belief' implies the act of believing a proposition x as true and thus the conscious feeling of confidence about the reality of x. Believing constitutes a type of action and a form of intentionality. But are all beliefs really equal in terms of action and intentionality? Do we believe in the same way in the competence differences between male and female as we do about the existence or non-existence of life after death? Sociologists generally distinguish beliefs that are consciously endorsed with a strong level of individual or collective reflexivity from others that are more or less taken for granted with a low level of individual or collective reflexivity. Hence the tendency to disassemble the general category of 'collective belief' into multiple subcategories to account for the diversity of the modes of production and acceptance of opinions and representations (Clément, 2010).

This leads us to a third difficulty: confronted by the internal complexity of the category of collective belief, sociologists have sometimes multiplied the explanatory categories, and in so doing have forgotten that the vocation of any explanatory theory is to establish the simplicity behind the apparent complexity. In a well-known article intended to provide a general theoretical framework for the sociology of knowledge, Merton (1945) summarizes the nature of the sociological explanation in the following way: to establish the correlations between collective beliefs, conceived as dependent variables, and "the other existential factors of society and culture" conceived as independent variables. If such a formula makes it possible to unify many sociological studies devoted to the explanation of collective beliefs, this unity remains, however, purely 'formal.' In practice, an overview of the sociological literature shows that there is no real consensus among sociologists on the nature of the social mechanisms at work in the production and dissemination of beliefs. Collective beliefs are shared representations by individuals or groups. But what do individuals or groups have in common that generates the collective nature of their beliefs? What are the 'existential factors,' to use Mertonian terminology, that explain at least partly the ability of certain beliefs to be produced, diffused, and shared by multiple individuals? Without trying to be exhaustive, we may analytically distinguish between five main factors or social mechanisms, albeit slightly overlapping, generally related to the production and diffusion of collective beliefs.

Collective Beliefs and Social Mechanisms

Historically, one of the first factors discussed by sociologists is social position or social status. Why do individuals share the same beliefs? Because they share the same social position or the same socioeconomic status. Marx's main contribution to the sociology of ideology is the general thesis that individuals having a similar position in the social order (dominant or dominated) also have similar social interests and by extension a similar set of impressions, illusions, ways of thinking, and some philosophical conceptions. There is no need to return to the traditional theory of ideology; contemporary research has repeatedly documented a consistent relationship between social position and representations about poverty or socioeconomic inequality (Nilson, 1981; Kluegel and Smith, 1986; Lee et al., 1990; Hadler, 2005). For example, two types of representation on the origins of poverty are generally opposed: individualist accounts and structuralist accounts. The former posits that social inequality is due to disavantadged individuals themselves and is associated with inaction regarding that inequality. The structuralist category holds that there are external factors systematically disfavoring the disadvantaged individual. According to numerous studies, notably in the United States, being white, older, male, and of higher income increases the propensity to favor individualistic over structural beliefs. Inversely, black Americans, who often occupy lower social positions, are more critical about the social order and tend to favor structural beliefs over individualistic beliefs. This relationship between belief and social position is of course not limited to socioeconomic representations. In their study on the diffusion of parascientific beliefs, Boy and Michelat (1986) distinguish two sets of beliefs: astrology and paranormal phenomena. These two sets are not equally distributed among the various socioeconomic status groups. High socioeconomic groups tend to favor paranormal beliefs whereas middle or low socioeconomic groups are more inclined toward astrology. This inequal socioeconomic distribution is also observed by Schieman (2010) in his study on beliefs about God's influence in everyday life: "Compared with high socioeconomic status (SES) groups, low SES individuals tend to profess higher levels of divine involvement and control even when both groups share similarly low levels of religious involvement. These patterns concur with the claim that low SES groups tend to hold more orthodox beliefs than high SES individuals-regardless of their actual participation in religious activities" (p. 44). All of the collective representations – socioeconomic beliefs, parascientific beliefs, religious beliefs, etc. – are discussed in terms of positional effect. However, if a simple psychology of interest gives real support to the positional explanation of socioeconomic beliefs (each position is supposed to determine a sense of self-interest for each individual), the relation between social position and parascientific or religious beliefs remains more enigmatic. Sociologists generally try to overcome this enigma by combining positional factors with other types of factors or mechanisms.

The second 'existential factor' generally related to collective belief is social interaction. Why do individuals share the same beliefs? Because, living generally not as hermits, they share the same ability to interact with others, and through these interactions, they have the same capacity to be influenced by others in the nature of their beliefs. There are of course numerous ways to conceptualize this social influence. In his classical *Laws of Imitation* (1890), G. Tarde uses the notion of imitation to describe one of its modalities. Defined as an 'interpersonal suggestion,' imitation acts as a social mechanism by which the individual's actions and beliefs are able to converge collectively in the same direction. Without necessarily endorsing the epidemiological approach of ideas developed by Sperber (1996) in the continuity of Tarde, sociologists frequently emphasize the importance of day-to-day interactions on the formation and diffusion of beliefs. It is well established that parents largely contribute, by explicit teaching and by example, to shaping their children's beliefs, moral values, and political preferences, at least until a certain age. Devout parents contribute to the reproduction of religious beliefs through their children (Myers, 1996; Kelley and De Graaf, 1997). But parents are only a small part of the interactional environment of children. Studies on primary and secondary forms of socialization describe the variety of interactions through which collective beliefs are not only transmitted but also produced and transformed. Kuhn (2012 [1962]) and Zuckerman (1977), in the tradition of the sociology of science, observe that the daily interaction between the masters and apprentices in science is essential for the production and transmission of normative and aesthetic beliefs. More generally, the importance of interaction has been analyzed for 'status beliefs,' that is, collective representations about inequality between people having

different kinds of occupation, education, age, race/ethnicity, or gender (Ridgeway et al., 1998). Interaction among individuals has a creative dimension: "People carry [their] status beliefs to subsequent encounters with individuals from the other category and, by acting on the beliefs, induce some of those others to take on the status beliefs as well. This creates a diffusion process that, under some structural conditions, will create roughly consensual (i.e., widely shared) status beliefs" (Ridgeway and Erickson, 2000: p. 581). If social influence is somehow inevitable, certain situations seem more likely to make people particularly receptive to it. Rydgren (2011) argues that social influence is at its strongest when two elements combine in a single situation: (1) a real uncertainty about the content of the belief and (2) a feeling of deviation from the belief held by most others in the immediate social surrounding. In this situation, the diffusion of beliefs depends heavily on the use of epistemic authorities and the tendency of individuals to harmonize their beliefs with those of significant others in order to reduce their own feeling of deviation.

This reference to feelings and emotions leads us to a third mode of explanations of collective beliefs, frequently combined with the two previous ones. If individuals or groups share the same beliefs it is because they are moved by the same collective feelings or the same emotional currents. In the famous *Traité de sociologie générale*, Pareto (1916) analyzes the social diffusion of the religions as the consequence of social 'feelings.' He affirms the importance of clearly separating the respective influence of 'derivations' – ideas, theories, theologies – and 'residues' – deep-rooted feelings. "The social value of the religions," Pareto writes, "depends very little on derivations, enormously on the residues. In several religions, there is a significant group of residues (…) which correspond to feelings of discipline, submission, hierarchy" (section 1854). For Pareto, the social actor wants to be convinced of the value of his or her passion, and produces derivations for this purpose. Religions or ideologies are just epiphenomenal expressions of dominant passions. If the Paretian model of collective beliefs remains generally ignored by contemporary sociologists, some of them do analyze the emergence and success of certain beliefs as the consequence of certain feelings, in particular relative deprivation feelings. As recalled by Merton (1964), the concept of relative deprivation comes from the study of the 'sentiments and attitudes' of American soldiers (Stouffer et al., 1949). The central premise of relative deprivation theory (Runciman, 1966) is that individuals or groups will feel deprived when their current circumstances are negatively compared to (1) one's own previous or anticipated future situation (self-referenced relative deprivation), (2) the (real or imagined) situation of members of one's in-group (e.g., friends and relatives), or (3) the (real or imagined) situation of members of external reference groups. Relative deprivation theory can be used to explain different kind of beliefs. How do personal experiences of economic recession affect the distribution of economic beliefs (Krahn and Harrison, 1992)? What accounts for the apparently 'universal' fact that women are more religious than men (Miller and Stark, 2002)? How do the new radical right-wing parties in Europe try to channel the feelings of anxiety and resentment of the social 'losers' of post industrialization processes (Rydgren, 2007)? Why are low socioeconomic status individuals more likely to seek God's will through prayer (Schieman, 2010)? To what extent are the representations of social inequality in the French population produced by comparison mechanisms (Galland et al., 2013)? In all of these studies, sociologists try to measure the social influence of feelings resulting from interpersonal comparison on the acceptance of religious, economic, or ideological beliefs.

A fourth sociological approach of collective beliefs can be briefly identified as culturalist. It proposes that belief is a piece of a wider sociocultural system that imposes itself on the social actors. Why do individuals share the same beliefs? Because living in the same collectivity, they are raised within the same dominant culture and thus tend to select their beliefs according to their degree of elective affinity with it. Moreover, beliefs are not independent from one another, but interconnected in a dominant cultural system. This culturalist approach is explicit in Sorokin's pioneer studies on *Social and Cultural Dynamics* (1937). Unlike others sociologists trying to explain belief by various states of the social structure – positional or interactional factors – Sorokin proposes to derive every aspect of shared knowledge from underlying 'culture mentalities.' Two 'pure types' of culture mentalities are identified: the *ideational* and the *sensate*. The former conceives of reality as 'nonmaterial' and defines human needs as primarily spiritual. The latter conceives of reality as only what can be perceived through the senses and defines human needs as primarily physical. Sorokin argues that what a given society regards as true or false, scientific or unscientific, is fundamentally conditioned by the long-term historical fluctuation of influence of these culture mentalities (Barber and Merton, 1963). If contemporary sociologists generally ignore the macro-sociohistorical perspective formulated by Sorokin, they still show strong interest in the relationship between culture and beliefs. In his study of public opinion on racial inequality, Bailey (2002) opposes Brazil to the United States. Why is there a Brazilian consensus – the majority of the population favors a structuralist representation – on the origins of racial inequality? According to Bayley, part of the answer comes from a specificity of the Brazilian culture. One important element that may separate Brazil from the United States is the lower salience of color as a marker of social belonging: "Brazil's color awareness responds to 'marca' or appearance, as opposed to "origem" or ancestry, the explicitly racial U.S. version. The difference between the two may be subtle, but not unimportant. (…) distinctions based on discrete racial identities may create caste-like or essentialist distinctions (…) But when socially constructed boundaries are porous or ambiguous, their significance may be reduced as markers of discrete, self-identifying social segments with diverging worldviews" (p. 429). Another recent example is sociological study of beliefs about illness and medicines. People differ not just in their beliefs about the nature and uses of medicines, but also in the degree to which they perceive themselves to be sensitive to side effects. In their study over a population of university students, Horne et al. (2004) show that "students who identified themselves as Asian in origin were more likely than students with a European origin to perceive medicines as intrinsically harmful. These findings support previous research, which

found that students who identified themselves as Asian were generally less satisfied with "Western" medical care compared with "Western" students" (p. 1312).

Finally, a fifth type of sociological explanation of collective beliefs focuses on the reasons or systems of reasons that social actors have to endorse their own beliefs. This approach to collective belief is generally combined with positional and interactional factors. It is generally described in terms of 'cognitive rationality' or 'ordinary rationality' (Boudon, 1994, 2010). Why do individuals share the same beliefs? Because they have the same capacity to develop a form of ordinary rationality and to adopt, in a deliberative manner or by simple aggregation, a common system of reasons. This cognitive explanation of beliefs is generally associated with Durkheim and Weber's classical writings on magic beliefs. "The acts motivated by magic, Weber writes (1922/1979), are acts at least relatively rational (…): they follow the rules of the experience even if they are not necessarily acts according to the means and ends." The same point is emphasized by Durkheim (1912/1995), in particular when he compares the rationality of traditional 'rites' and the rationality of modern 'techniques': "the rites which [the primitive] use to ensure the fertility of the ground (…) are not, to him, more irrational than are, to us, the technical processes used by our agronomists (…). The powers linked to these rites do not appear particularly mysterious. For those who believe in them, those forces are not more unintelligible than are gravity or electricity to a contemporary physicist." The explanatory strategy proposed by Weber and Durkheim invites the sociologist to reconstruct the set of reasons that the actor gives to himself to act. This cognitive approach is opposed to those involving the illusory nature of reasons or to those assuming that the behavior of the social actor is primarily emotional or mechanically forged by the cultural environment. Applied to various beliefs – the effectiveness of public policy on drug addiction (Bergeron, 1999), the existence of Santa Claus, the social resistance to Darwinism (Bronner, 2004, 2007), or paranormal beliefs that allege the existence of extraordinary beings and extraordinary phenomena (Renard, 2010) – the rationalist approach to collective beliefs follows four general principles summed up by Raymond Boudon (2011) in one of his last public conferences. These principles are that (1) ideas, and not interests, are the essential component of social action; (2) the causes of beliefs are consciously or meta-consciously present in the mind of the believer; (3) rationalization of ideas plays a central role in modern societies life; and (4) normative beliefs, just like descriptive beliefs, have their origin in ordinary rationality and are subject to a process of rationalization.

Robustness and Symmetry of Sociological Explanations

As shown by the examples discussed in the previous section, sociologists refer in the course of their studies to multiple factors or social mechanisms. However, which general conditions do sociologists need to respect in order to produce a convincing or robust explanation of collective beliefs? To conclude our discussion, four elementary conditions deserve to be briefly discussed.

First condition: *The sociological explanation of collective beliefs must be more than the translation of the phenomenon to explain a new idiom*. The sociologists of beliefs generally start from the identification of an enigmatic phenomenon. Why did the industrial bourgeoisie of the nineteenth century believe in the virtues of parliamentary monarchy (Marx)? Why do Zuni Indians adopt a representation of the world in seven categories (Durkheim)? Why were corpuscular and ondulatory theories of light widely accepted during the nineteenth century (Sorokin)? But it sometimes happens that their answers are only apparent. The methodological analysis of the theory of 'primitive mentality' thus shows that if the theory of Lévy-Bruhl (1922/1960) is not convincing anymore, this is essentially because it does not explain anything. The key concepts of this theory, Boudon emphasizes (1994), "only paraphrase the beliefs. They explain the confusion of the primitives between verbal associations and causal relations by the primitive's tendency to make this confusion, tendency whose reality is guaranteed by the fact that they do make it indeed." A similar point is made by Merton (1945) concerning Sorokin's theory. "It seems tautological, Merton writes, to say, as Sorokin does, that in a sensualist society and culture, the sensualist system of the truth based on the testimony of the sensory organs must prevail. Because he had already defined the sensualist mentality as one which conceives reality as being only what is perceived with the sensory organs." More recently, Cole (1996) pointed out that the analysis of the production of scientific beliefs by certain constructivist sociologists of sciences is primarily of a circular nature. These sociologists generally claim "that it is impossible to separate the technical from the social; that all science is inherently social. This turns their entire argument into a tautology. If science is inherently social, this means that the technical aspects of scientific discoveries by necessity must be determined socially. This is indeed the question we are examining."

Second condition: *The sociological explanation of collective beliefs must avoid ambiguous and/or occult concepts*. Sociologists use a great number of conceptual categories. Many proposals produced on the basis of the combination of these categories make clear and univocal sense; some, however, remain more ambiguous. For example, certain sociological explanations rest on an ambiguous concept of 'social,' or more precisely a confusion between the social and historical dimensions of a collective belief. To show, however, that a belief appears in a particular historical context is not the same as to demonstrate its social character. Laudan (1977) is right to observe that "the fact that certain assertions are contextual, that one adhered to it only to a certain time and in a certain place establishes neither that they are necessarily of social origin, nor even that they are opened to the sociological analysis." The sociological use of the concept of 'interest' constitutes another well-known example of ambiguity. This concept is sometimes used by sociologists of science in order to describe the origin of the prestige or credibility of scientific beliefs. In an under-determined situation, contemporary sociologists of science say, scientists choose between two concurrent theories the one that appears to conform to their best interests. However, how should we

conceive the exact nature of these interests? Are they material, symbolic, technical, or cognitive interests? As long as the concept of interest remains ambiguous, the sociologist cannot legitimately claim to explain the observed beliefs. In addition, following a Weberian principle, many sociologists closely associate the solidity of an explanation with the manner in which these explanations avoid using occult categories, that is, invisible forces – such as the Paretian concept of 'residue' or the 'primitive mentality' of Lévy-Bruhl – that determine the behaviors of the social actors. The methodological principle is simple and easily acceptable: the sociologist seeking to account for the origin of a collective belief may refer to the existence of an occult cause only insofar as she cannot explain this belief more easily, in particular by identifying her meaning. The obvious difference between the explanations of magic beliefs proposed on the one hand by Lévy-Bruhl and on the other by Durkheim and Weber seems to be directly linked to the application of this general principle by the latter.

Third condition: *The sociological explanation of collective belief must be compatible with observable facts*. This third condition can be interpreted in various ways. The first and simplest one consists of assuming that explanatory theories of scientific beliefs are not different from other scientific theories. They must not only dispose of various internal qualities (absence of circularity, precision, and so on) but also be able to be confronted with precise facts, and further still to resist this empirical test. These facts might be elementary observations. The observation, for example, of the persistence of the magic beliefs in Western societies seems hardly compatible with the assumption of a causal link between the magic beliefs and a 'primitive mentality' (Lévy-Bruhl). In the same way, the observation of the development of new religious movements in large cities openly contradicts the hypothesis that the contemporary process of urbanization is essentially a vector of disintegration of collective belief (Borhek and Curtis, 1975). These facts can also be more complex products of the interpretative reconstruction work done by the sociologist or anthropologist. When Horton (1993) argues, for example, against the tenants of symbolist theory of magic, he points out that this theory, for which magic beliefs are mainly the expression of a symbolic 'desire,' enters into flagrant contradiction with what the members of the traditional societies say about their own rituals: "(…) in denying the paramount importance of explanation, prediction and control as guiding aims of traditional African religious life, "symbolists" (…) are committing the cardinal interpretative sin of flouting the actor's point of view." Another way to interpret this third condition consists of associating the solidity of an explanatory theory with its capacity to give an account of other beliefs than the one for which it was initially elaborated. For example, is it possible to apply the explanatory scheme imagined by Durkheim and Weber for magic beliefs to other types of beliefs, such as ideological beliefs? The fact that this transposition is not only possible but brings an objective gain of intelligibility to the comprehension of ideological phenomenon (Boudon, 1986, 1989) demonstrates the real value of their general conception of the sociological explanation.

Sociologists frequently add to these three general conditions a fourth and final one: a condition known as the 'principle of symmetry.' This principle implies that the explanation of collective beliefs should be equally applied to phenomena of various natures. This last condition is not reducible to the conditions previously mentioned (the third one in particular) insofar as, to identify the nature of this plurality, it introduces a criterion of truth and falseness, rationality and irrationality. A sociological theory of collective beliefs is said to be symmetrical if it explains in the same terms true beliefs and false beliefs, rational beliefs and irrational beliefs. This principle, implicitly used by Sorokin (1937) or more recently explicitely by Boudon (2010), is directly opposed to the "a-rationality principle" according to which the sociological explanation of collective beliefs cannot be indifferent to the content of these beliefs. The interest of the principle of symmetry is twofold. From a methodological point of view, it requires the suspension of all evaluation related to the potential validity or invalidity of the analyzed collective belief. It thus represents a guarantee of neutrality. From the point of view of empirical research, it makes it possible to follow in unified and thus simplified terms the manner in which collective beliefs obtain social recognition, or conversely an absence of social recognition. The sociological use of this principle follows, however, two paths of unequal importance. The first, chosen by some sociologists of science, exists mainly to explain the scientific beliefs in terms of explanatory categories previously devoted to false beliefs. Linked to an exclusively causal approach of scientific belief production, this first use of the principle of symmetry has greatly contributed to the diffusion of a relativistic representation of science (Dubois, 2001). There is, however, another way to apply this principle. It consists of following the major intuition of the theory proposed by Weber and Durkheim: to seek behind all collective beliefs, true or false, socially validated or socially marginal, the influence of 'reasons,' objective and or subjective. Thusl, as Horton (1993) is reconstructing the limited rationality of certain African rituals, he defines the major nature of its explanatory project as elaborating "an explanatory framework that would deal 'symmetrically' with the traditional and the modern, the prescientific and the scientific."

See also: Action, Theories of Social; Belief, Anthropology of; Cognitive Dissonance; Collective Memory, Anthropology of; Culture, Sociology of; Durkheim, Emile (1858–1917); Ideology, Sociology of; Knowledge, Sociology of; Mannheim, Karl (1893–1947); Rational Choice Theory in Sociology; Social Movements: A Social Psychological Perspective; Sociological Theory; Weber, Max (1864–1920).

Bibliography

Bailey, S., 2002. The race construct and public opinion: understanding Brazilian beliefs about racial inequality and their determinants. American Journal of Sociology 108 (2), 406–439.

Barber, B., Merton, R., 1963. Sorokin's formulation in the sociology of science. In: Allen, J. (Ed.), Pitirim A. Sorokin in Review. Duke University Press, pp. 332–368.

Bergeron, H., 1999. Politiques publiques et croyances collectives. Analyse socio-historique de la politique française de soins aux toxicomanes de 1970 à 1995. Déviance et Société 23 (2), 131–147.

Boudon, R., 1994. The Art of Self-Persuasion: The Social Explanation of False Beliefs. Polity Press, London.
Boudon, R., 1986. L'Idéologie ou l'origine des idées reçues. Fayard, Paris; English translated.
Boudon, R., 1989. The Analysis of Ideology. University of Chicago Press, Chicago.
Boudon, R., 1999. Local vs general ideologies: a normal ingredient of modern political life. Journal of Political Ideologies 4 (2), 141–161.
Boudon, R., 2010. La rationalité ordinaire: colonne vertébrale des sciences sociales. L'Année Sociologique 60 (1), 19–40.
Boudon, R., October 2011. Choisit-on ses Croyances? Conference University Paris Sorbonne. Text available: http://www.gemass.fr/spip.php?article3099.
Borhek, J.T., Curtis, R.F., 1975. A Sociology of Belief. Wiley, New York.
Boy, D., Michelat, G., 1986. Croyances aux parasciences: dimensions sociales et culturelles. Revue Française de Sociologie 27 (2), 175–204.
Bronner, G., 2004. Contribution à une théorie de l'abandon des croyances :la fin du père noël. Cahiers Internationaux de Sociologie 116 (1), 117–140.
Bronner, G., 2007. La résistance au darwinisme: croyances et raisonnements. Revue Française de Sociologie 48 (3), 587–607.
Bronner, G., 2011. Ce qu'Internet fait à la diffusion des croyances. Revue Européenne de Sciences Sociales 49 (1), 35–60.
Clément, F., 2010. De la nature des croyances collectives. L'Année Sociologique 60 (1), 63–91.
Cole, S., 1996. Voodoo sociology: recent developments in the sociology of science. Annals of the New York Academy of Sciences 775, 274–287.
Davie, G., 1994. Religion in Britain since 1945. Oxford, Blackwell.
Dubois, M., 2001. La Nouvelle Sociologie des Sciences. Presses Universitaires de France, Paris.
Durkheim, E., 1912/1995. The Elementary Forms of Religious Life. The Free Press, New York.
Evans, G., 1997. Political ideology and popular beliefs about class and opportunity: evidence from a survey experiment. The British Journal of Sociology 48 (3), 450–470.
Galland, O., Lemel, Y., Frénod, A., 2013. La perception des inégalités en France. Essai d'explication. Revue Européenne des Sciences Sociales 51 (1), 179–211.
Greeley, A., Hout, M., 1999. American's increasing belief in life after death: religious competition and acculturation. American Sociological Review 64 (6), 813–835.
Hadler, M., 2005. Why do people accept different income ratios? A multi-level comparison of thirty countries. Acta Sociologica 48 (2), 131–154.
Hervieu-Léger, D., 2010. Le partage du croire religieux dans des sociétés d'individus. L'Année Sociologique 60, 41–62.
Horne, R., Graupner, L., Frost, S., Weinman, J., Wright, S., Hankins, M., 2004. Medicine in a multi-cultural society: the effect of culture background on beliefs about medications. Social Science & Medicine 59, 1307–1313.
Horton, R., 1993. Patterns of Thought in African and the West. Cambridge University Press, Cambridge, UK.
Hunt, M., 1996. The individual, society or both? A comparison of black, Latino and White Beliefs about the causes of poverty. Social Forces 75 (1), 293–322.
Kelley, J., De Graaf, N., 1997. National context, parental socialization, and religious belief: results from 15 nations. American Sociological Review 62 (4), 639–659.
Kluegel, J., 1990. Trends in whites' explanations of the black-white gap in socioeconomic status, 1977–1989. American Sociological Review 55 (4), 512–525.
Kluegel, J., Smith, E., 1981. Beliefs about stratification. Annual Review of Sociology 7, 29–56.
Kluegel, J., Smith, E., 1986. Beliefs about Inequality. Americans' Views of What Is and What Ought to Be. Aldine de Gruyter, New York.
Krahn, H., Harrison, T., 1992. "Self-referenced" relative deprivation and economic beliefs: the effects of the recession in Albert. Canadian Review of Sociology & Anthropology 29 (2), 191–209.
Kuhn, T., 2012 [1962]. The Structure of Scientific Revolutions. 50th Anniversary. Ian Hacking (Intro.). University of Chicago Press.
Laudan, L., 1977. Progress and Its Problems. University of California Press, Berkeley, CA.
Lee, B., Jones, S., Lewis, D., 1990. Public beliefs about the causes of Homelessness. Social Forces 69 (1), 253–265.
Lévy-Bruhl, L., 1922/1960. La Mentalité Primitive. Presses Universitaires de France, Paris.
Markovsky, B., Thye, S., 2001. Social influence on paranormal beliefs. Sociological Perspectives 44 (1), 21–44.
Merton, R.K., 1945. Sociology of knowledge. In: Gurvitch, G., Moore, W.E. (Eds.), Twentieth Century Sociology. Freeport, New York.
Merton, R., 1964. Social Theory and Social Structure. The Free Press of Glencoe.
Miller, A., Stark, R., 2002. Gender and religiousness: can socialization explanations be saved? American Journal of Sociology 107 (6), 1399–1423.
Myers, S., 1996. An interactive model of religiosity inheritance: the importance of family context. American Sociological Review 61, 858–866.
Natras, N., 2013. Understanding the origins and prevalence of AIDS conspiracy beliefs in the United States and South Africa. Sociology of Health & Illness 35 (1), 113–129.
Nilson, L., 1981. Reconsidering ideological lines: beliefs about poverty in America. Sociological Quarterly 22, 53148.
Pareto, V., 1916. Trattato di Sociologia Generale. Firenze, G. Barbéra.
Rashotte, L., Webster, M., 2005. Gender status beliefs. Social Science Research 34, 618–633.
Renard, J.-B., 2010. Croyances fantastiques et rationalité. L'Année Sociologique 60 (1), 115–135.
Ridgeway, C., Boyle, E., Kuipers, K., Robison, D., 1998. How do status beliefs develop? The role of resources and interactions experience. American Sociological Review 63 (3), 331–350.
Ridgeway, C., Correll, S., 2004. Unpacking the gender system: a theoretical perspective on gender beliefs and social relations. Gender and Society 18 (4), 510–531.
Ridgeway, C., Erickson, K., 2000. Creating and spreading status beliefs. American Journal of Sociology 106 (3), 579–615.
Runciman, W., 1966. Relative Deprivation and Social Justice. University of California Press.
Rydgren, J., 2007. The sociology of the radical right. Annual Review of Sociology 3, 241–262.
Rydgren, J., 2011. Beliefs. In: Hedström, P., Bearman, P. (Eds.), The Oxford Handbook of Analytical Sociology. Oxford University Press, pp. 72–93.
Schieman, S., 2010. Socioeconomic status and beliefs about god's influence in everyday life. Sociology of Religion 71 (1), 25–51.
Schuman, H., Krysan, M., 1999. A historical Note on White's Beliefs about racial inequality. American Sociological Review 64 (6), 847–855.
Sorokin, P., 1937. Social and Cultural Dynamics, Fluctuation of Systems of Truth, Ethics, and Law, vol. 2. American Book Company, New York.
Stouffer, S., Suchman, E., DeVinney, L., Star, S., Williams, R., 1949. The American Soldier: Adjustment during Army Life. Princeton University Press.
Sperber, D., 1996. La Contagion des Idées. Théorie Naturaliste de la Culture. Odile Jacob.
Weber, M., 1922/1979. Economy and Society. University of California Press.
Zuckerman, H., 1977. Scientific Elite. Nobel Laureates in the United States. The Free Press.

Collective Efficacy and Crime

John R Hipp and James C Wo, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, we explore the literature on collective efficacy and crime, but pay attention to three major challenges that confront this literature, and researchers moving forward with the concept in studies of neighborhoods. These challenges are (1) precisely defining and measuring the concept of collective efficacy; (2) determining whether the notions of general cohesion and trust in neighborhoods are really components of collective efficacy, or determinants of it; and (3) whether neighborhood crime can have feedback effects on the level of collective efficacy.

Collective efficacy is a construct that originally comes out of the psychology literature and the writings of Bandura (1982, 1986). Although the psychology literature has explored the empirical effects of collective efficacy for various relatively small groups, arguably its largest impact on social science literature comes from the introduction of the concept to the neighborhoods and communities literature by Sampson et al. (1997). A voluminous literature that is largely sociological and criminological has explored the effect of collective efficacy on a number of community processes, including neighborhood crime rates.

In this article, we explore the literature on collective efficacy and crime, but pay attention to three major challenges that confront this literature, and researchers moving forward with the concept in studies of neighborhoods. These challenges are (1) precisely defining and measuring the concept of collective efficacy; (2) determining whether the notions of general cohesion and trust in neighborhoods are really components of collective efficacy, or determinants of it; and (3) whether neighborhood crime can have feedback effects on the level of collective efficacy.

A Brief History of the Concept of Collective Efficacy

Collective efficacy was initially introduced as a concept by Bandura (1982, 1986). Building on his notion of self-efficacy, Bandura noted that "Perceived collective efficacy will influence what people choose to do as a group, how much effort they put into it, and their staying power when group efforts fail to produce results" (Bandura, 1982: p. 143). His early consideration of the idea mostly conceptualized collective efficacy as a property of very large groups, such as nations and social movements (Bandura, 1986: pp. 449–453). This spawned a number of studies in the psychology literature in the 1990s and into the 2000s focusing on collective efficacy as a property of small groups, such as sports teams and work organizations (Zaccaro et al., 1995). Collective efficacy is conceptualized in this literature as a fundamentally group phenomenon that "… fosters groups' motivational commitment to their missions, resilience to adversity, and performance accomplishments …," and an emergent phenomenon in which "the locus of perceived collective efficacy resides in the minds of group members" (Bandura, 2000: p. 76)

The key twist in the trajectory of collective efficacy as a concept in the social sciences arguably came when Sampson and colleagues introduced the concept to the neighborhoods and communities literature. In what has since become a citation classic, their study in *Science* introduced the concept as a property of neighborhoods that might be able to reduce the level of violence (Sampson et al., 1997). This seminal study focused on neighborhoods in the city of Chicago, and proposed defining collective efficacy as the combination of a general sense of residents' willingness to provide informal social control along with high levels of trust and solidarity among residents. They argued for the plausibility of this broader conceptualization of collective efficacy compared to that developed in the psychology literature because "the willingness of local residents to intervene for the common good depends in large part on conditions of mutual trust and solidarity among neighbors" (Sampson et al., 1997: p. 919). The cross-sectional study indeed found that neighborhoods with higher levels of collective efficacy exhibited lower levels of violence (whether measured as perceived violence, violent victimization, or homicide events). This single study has fostered an army of subsequent research.

Sampson himself has further refined the concept of collective efficacy in subsequent work. In particular, he has argued that it is a concept that helps focus on social mechanisms, and moves away from a risk-factor approach (Sampson, 2006a,b). In this sense, he considers it a mediating variable between neighborhood structural characteristics and crime. He emphasizes that this concept is fundamentally about repeated interactions, which then impact expectations about the future (residents' beliefs about the actions of their fellow residents) (Sampson, 2006a,b). Sampson's work on collective efficacy and neighborhood crime can be seen in his recent book that chronicles the Project on Human Development in Chicago Neighborhoods (PHDCN) as a large study of neighborhoods in Chicago (Sampson, 2012).

Collective Efficacy: Challenges of Measurement

Measuring informal social control in neighborhoods is actually quite a challenging problem. The challenge is that informal social control activity is only on display in moments in which there are challenges to the social order. This potential response on the part of residents is typically hypothetical, or *latent*.

Thus, attempting to measure the behavior of informal social control is difficult because of its rarity. This is both a methodological challenge (for measurement) as well as a theoretical challenge. More commonly, studies have relied on questions that are hypothetical about how the respondent would behave in a particular instance (Warner, 2007). In the case of studies measuring collective efficacy, the hypothetical questions ask respondents how they believe *their neighbors* would behave in various situations (Sampson et al., 1997). Answering such questions will sometimes be difficult if residents have never seen their neighbors be confronted by situations requiring informal social control behavior and require them to speculate on the answer. This speculation is particularly likely in neighborhoods with low levels of crime or disorder (St. Jean, 2007: p. 211). Besides these issues around uncertainty is the challenge of measuring a group concept by surveying individuals. This issue remains unresolved in the literature, as Zaccaro et al. (1995) note, "… a more useful approach to defining collective efficacy is to consider both judgments of members' abilities and perceptions of how well group members work together in achieving collective outcomes" (p. 309).

Other approaches have instead attempted to measure collective efficacy using behavioral measures. For example, one study suggested that parental monitoring may represent collective efficacy in action (Rankin and Quane, 2002). Another study used the presence of registered voters as a proxy for the collective efficacy in a streetblock (Weisburd et al., 2012). One unique and clever approach to measuring collective efficacy in neighborhoods constructed a website that allowed residents of neighborhoods to start webpages, and then assessed the usage of the pages as one measure of collective efficacy (Hampton, 2010). An interesting finding was that there were two extremes of neighborhoods that generated such webpages – new middle-class suburban neighborhoods and truly disadvantaged neighborhoods – suggesting that such a measure may be tapping some different processes than the more conventional measures of collective efficacy.

Distinguishing between Cohesion and Informal Social Control: One Construct, or Causally Related?

Although Sampson and colleagues defined collective efficacy in neighborhoods as being a combination of both cohesion/trust and expectations of informal social control, there is debate about this strategy. It is worth noting that the psychology literature has typically conceived of cohesion and collective efficacy as two separate constructs, which may separately impact one another in a causal fashion (Zaccaro et al., 1995). Empirical evidence in the communities and crime literature suggests that these may indeed be distinct constructs in many locations. Horne (2004) notes that the two measures are not always highly correlated and that high levels of cohesion do not necessarily increase norm enforcement. A study in Chicago suggested that informal social control attitudes mediated the relationship between neighborhood cohesion and crime (Rhineberger-Dunn and Carlson, 2011). A study of Tianjin, China found a correlation of just 0.43 between neighborhood cohesion and informal control, suggesting that they are in fact distinct constructs (Zhang et al., 2002). A study that split cohesion and expectations of informal social control found that while cohesion had a negative effect on perceived incivilities in the neighborhood, expectations of informal social control had no impact (Reisig and Cancino, 2004). A study of Mesa, AZ found minimal factor analysis evidence for combining these two constructs, and also found that only cohesion had a negative effect on violent crime when both were included in the model simultaneously (Armstrong et al., 2010). A study of Brisbane neighborhoods using a confirmatory factor analysis strategy also found no evidence for combining these two constructs; furthermore, this study used measures capturing collective efficacy for three distinct tasks, and all of them were distinct from a measure of cohesion (Wickes et al., 2013). Another study measured neighborhood attachment rather than cohesion, and found that higher levels of attachment led to higher levels of expectations of informal social control (Burchfield, 2009; Silver and Miller, 2004).

A key assumption of the neighborhood collective efficacy literature is that these norms and attitudes within a neighborhood that lead residents to perceive higher levels of collective efficacy then translate into actual higher levels of informal social control behavior. Thus, "Social movements and collective efficacy theory share a common orienting framework – a focus on the mobilization of action for an intended purpose" (Sampson et al., 2006: p. 4). As Steenbeek and Hipp (2011: p. 839) noted: "expectations that others will intervene (potential social control) need not necessarily result in people actually intervening more (actual social control behavior), even though this is implicitly assumed by social disorganization theory." However, only a few studies have addressed this question empirically, and the evidence so far appears somewhat weak. For example, one study in the Netherlands did not have a robust measure of collective efficacy, but did use longitudinal data to test and find that the potential for social control (measured by a single question asking about residents' willingness to improve the neighborhood) in fact did *not* increase social control behavior two years later (Steenbeek and Hipp, 2011). A study of neighborhoods in Brisbane focused on residents who reported a high level of disorder in the neighborhood, and found that higher levels of reported collective efficacy did not in fact lead to more actual informal social control behavior (Wickes et al., 2012). Instead, this study found that higher levels of neighboring increased the likelihood of engaging in informal social control behavior. As another example, a study using an experimental design concluded that norms are more likely to be enforced when there are benefits that increase social interaction, which suggests that cohesion might be more important for fostering action than residents' perceptions of informal social control potential (Horne, 2004).

Studies Assessing the Consequences of Collective Efficacy in Neighborhoods

A number of studies have assessed the cross-sectional relationship between neighborhood collective efficacy and crime, with studies almost always finding a negative relationship. The seminal study of Chicago found such an effect (Sampson et al., 1997), and subsequent studies in Chicago have also detected this negative relationship (Sampson and Raudenbush, 1999; Morenoff et al., 2001;

Schreck et al., 2009). This relationship has also been detected in other locations, including Stockholm (Sampson and Wikstrom, 2008), Los Angeles (Burchfield and Silver, 2013), Tianjin, China (Zhang et al., 2007), and Brisbane (Mazerolle et al., 2010). A challenge with some cross-sectional studies is the use of victimization reports as the outcome measure; in such studies the independent variable of collective efficacy is measured at one point in time but the dependent variable of victimization is assessed based on incidents that occurred during the previous year, introducing a temporal problem.

Studies have also assessed the relationship between neighborhood collective efficacy and various other outcomes, typically in cross-sectional designs. For example, studies in Chicago have found a negative relationship between collective efficacy and partner violence (Browning, 2002; Wright and Benson, 2011), risky sexual behavior by adolescents (Browning et al., 2008), sexual initiation by adolescents (Browning et al., 2005), self-control by children (Gibson et al., 2010), self-rated physical health (Browning and Cagney, 2002), and mortality during a heat wave (Browning et al., 2006). However, one study in Chicago found countervailing evidence, as impulsivity had an even stronger effect on offending in neighborhoods with high levels of collective efficacy (Zimmerman, 2010). This negative relationship between collective efficacy and various outcome measures has also been detected in other cities, including such outcomes as substance use by adolescents (Erickson et al., 2012) and bullying in school (Williams and Guerra, 2011). A longitudinal study found that collective efficacy increased authoritative parenting, which reduced delinquency among adolescents; it also found that authoritative parenting had the strongest effect in high collective efficacy neighborhoods (Simons et al., 2005). One study explored how school social controls were weakened in low collective efficacy neighborhoods when measuring the outcome of suspensions and arrest (Kirk, 2009). However, a study in England only found a negative relationship between collective efficacy and adolescent delinquency in economically deprived neighborhoods, but not in more economically advantaged neighborhoods (Odgers et al., 2009).

Sampson (2006a,b) points out that a distinction can be made between a theory of neighborhood crime rates and one of individual development, which is a distinction between enduring effects and situational effects. In this framework, collective efficacy may be more appropriately considered a situational theory. Indeed, a study in Chicago found no relationship between collective efficacy and violence committed by specific racial group members (Sampson et al., 2005), suggesting that collective efficacy may operate as a situational factor impacting crime events, and not as a factor creating offenders. Other research has found that robbers are less likely to offend in a high collective efficacy neighborhood, again emphasizing the role of collective efficacy in situational theories (Bernasco and Block, 2009).

The Possible Feedback Effect from Neighborhood Crime to Collective Efficacy

Relatively understudied is the possibility that crime may have a feedback effect on neighborhood-level collective efficacy. This feedback possibility has been discussed by Sampson (2006a,b) and Sampson and Raudenbush (1999). Indeed, St. Jean (2007: p. 211) argued that research needs "to pay close attention to the reciprocal relationship between high collective efficacy and low crime rates" because "… under conditions of low crime, neighbors are quite likely to report high estimates of trust and solidarity among themselves, and of the willingness of others to intervene" even though we do not know if they will actually intervene.

Nonetheless, studies have rarely accounted for this likely endogenous relationship between collective efficacy and crime, and instead often simply assume a one-directional causal relationship from collective efficacy to crime. One study tested and found such a negative feedback effect from crime to collective efficacy in a cross-sectional analysis utilizing instrumental variables (Sampson and Raudenbush, 1999). A cross-sectional study of neighborhoods in Mesa, AZ also used instrumental variables and found a negative feedback effect from crime to collective efficacy (Armstrong et al., 2010), and a cross-sectional study of 100 urban neighborhoods found a similar feedback effect on residents' expectations of informal social control with the use of instrumental variables (Bellair, 2000). Occasional research has ignored the possible effect of collective efficacy on crime and simply tested the reverse posited causal path; for example, one study found that violent crime had a negative relationship with neighborhood collective efficacy (Duncan et al., 2003), whereas another study found that violent crime was associated with lower levels of trust in a neighborhood (Garcia et al., 2007).

Social Networks and Collective Efficacy

Another thorny theoretical and methodological issue concerns disentangling the effects of the social network of residents in a neighborhood, the level of cohesion, and the level of collective efficacy. Besides the considerable conceptual overlap in these concepts, there is also a considerable overlap in how they operate in neighborhoods. An additional challenge is that actually measuring the social network of ties among residents in a neighborhood is extremely difficult, and studies are typically constrained to proxying neighborhood ties by asking a sample of residents to report how many neighbors they know. Such measures are simply estimates of mean degree in a neighborhood and do not capture any richer information about the full network (Bursik, 1999). Therefore, one recent study adopted the novel approach of *simulating* the networks existing in neighborhoods based on certain well-understood properties of network tie formation, and detected strong negative relationships between various structural network measures and crime in micro-units (Hipp et al., 2013).

Beyond these measurement challenges, there are at least three perspectives on how the social network of a neighborhood and the level of collective efficacy might be related: (1) a dense network of ties in a neighborhood leads to more collective efficacy; (2) a dense network of ties is necessary for, but not sufficient for, neighborhood collective efficacy; (3) there is no consistent relationship between the network of ties and collective efficacy. The first perspective builds on the insights of systemic theory and the importance of the network of ties among residents for fostering behavior (Kasarda and Janowitz,

1974). It is difficult to conceive of a neighborhood with high levels of cohesion but lacking dense social networks. Indeed, a study focusing on the actual provision of informal social control behavior in response to observed problems among residents of Brisbane neighborhoods found that the most important predictor of such behavior was a higher density of neighborhood ties, and not assessments of collective efficacy (Wickes et al., 2012).

In the second perspective, social networks are a necessary, but not sufficient, characteristic for a neighborhood to develop collective efficacy. Sampson (2006a,b) argues that the relationship between dense networks and informal social control capability in neighborhoods is more tenuous, and that in some neighborhoods dense ties can actually inhibit social control behavior. Empirically, Sampson (2006a,b) showed that indicators of friendship ties loaded on a different factor than the collective efficacy indicators. In short, networks may provide social capital for offenders, as well as others in the neighborhood (Browning et al., 2004). For example, Browning (2009) suggested that dense networks can bring together offenders and conventional residents, which then weakens the effect of collective efficacy in this 'negotiated coexistence' model.

In the third perspective, there is enough inconsistency in how neighborhood networks and collective efficacy relate that no pattern can be distinguished. For example, a study using in-depth interviews in two high collective efficacy neighborhoods found that strong social ties were not necessary: instead symbols of community and collective representations generated higher levels of collective efficacy (Wickes, 2010). Notably, this same study noted that key neighborhood institutions and organizations were crucial for responding to problems, suggesting that formal organizations may be understudied in this literature.

Theoretical Extensions of the Collective Efficacy Construct

There have been various theoretical extensions to the collective efficacy model in recent years. One important direction is recent research that has emphasized the importance of the *task-specific* nature of collective efficacy. Just as self-efficacy is in respect to a specific task, likewise collective efficacy is fundamentally in respect to a specific task. Despite the importance of the task-specific nature of the collective efficacy construct, nearly all studies in the communities and crime literature have tested it as a unitary phenomenon. Thus, researchers often refer to the 'collective efficacy' of a neighborhood as if it is a unitary phenomenon, rather than the 'collective efficacy regarding X specific task.' Not only does this lose the conceptual richness of the construct, but it also increases the risk of allowing the concept of collective efficacy to slide into a measure of 'everything good' about a neighborhood, a charge sometimes leveled against the concept of social capital (Portes, 1998). In response to this lacuna, a study of Brisbane neighborhoods measured collective efficacy for three distinct tasks and not only found important differences in the determinants of each of these, but also found that these tasks were distinct from the more global construct of cohesion and trust (Wickes et al., 2013). Thus, making a distinction between residents' collective efficacy regarding addressing violence in the neighborhood, vs. addressing confronting misbehaving children, and vs. addressing political/civic matters demonstrated that neighborhoods and individuals differed regarding these particular tasks.

There are also other developments. In an example of theoretical integration, some research has attempted to integrate the legal cynicism literature with collective efficacy. For example, one study found that collective efficacy mediated the relationship between highly cynical neighborhoods and the probability of arrest (Kirk and Matsuda, 2011). Another strand of research has developed the notion of 'street efficacy' for residents. In this view, adolescents with high street efficacy choose the streets that they frequent in order to avoid violence, which has consequences for the environments in which they find themselves (Sharkey, 2006). Another research strand focuses on how perceptions of collective efficacy might impact the perceptions of other neighborhood characteristics – even the neighborhood's boundaries. For example, a study suggested that residents in neighborhoods with higher levels of collective efficacy perceive their neighborhoods to be larger in size (Coulton et al., 2013).

Conclusion

Collective efficacy is clearly a hugely important concept in the communities and crime literature, which has spawned considerable theoretical and empirical excitement. There is a large body of literature that has detected a negative relationship between the level of collective efficacy in a neighborhood and the crime rate. However, important directions need to be explored. One important direction is distinguishing whether collective efficacy operates differently in high versus low disorder neighborhoods. St. Jean argued that collective action is only necessary if the block is attractive to offenders: if the block is not desirable to offenders, there will not be crime regardless of what residents do (St. Jean, 2007: p. 164). A second important direction is longitudinal research of this relationship. Given the very plausible expectation of a feedback effect from neighborhood crime to perceptions of collective efficacy, as well as the suggestive evidence from cross-sectional studies, there is clearly a strong need for longitudinal research to determine whether the collective efficacy and crime relationship is robust.

Bibliography

Armstrong, Todd A., Katz, Charles M., Schnebly, Stephen M., 2010. The relationship between citizen perceptions of collective efficacy and neighborhood violent crime. Crime & Delinquency.

Bandura, Albert, 1982. Self-efficacy mechanism in human agency. American Psychologist 37 (2), 122–147.

Bandura, Albert, 1986. Social foundations of thought and action: a social cognitive theory. In: Bandura, A. (Ed.), Series in Social Learning Theory. Prentice-Hall, Englewood Cliffs, NJ.

Bandura, Albert, 2000. Exercise of human agency through collective efficacy. Current Directions in Psychological Science 9 (3), 75–78.

Bellair, Paul E., 2000. Informal surveillance and street crime: a complex relationship. Criminology 38 (1), 137–167.

Bernasco, Wim, Block, Richard L., 2009. Where offenders choose to attack: a discrete choice model of robberies in Chicago. Criminology 47 (1), 93–130.

Browning, C.R., Leventhal, T., Brooks-Gunn, J., 2005. Sexual initiation in early adolescence: the nexus of parental and community control. American Sociological Review 70 (5), 758–778.

Browning, Christopher R., 2002. The span of collective efficacy: extending social disorganization theory to partner violence. Journal of Marriage and Family 64 (4), 833–850.

Browning, Christopher R., 2009. Illuminating the downside of social capital. American Behavioral Scientist 52 (11), 1556–1578.

Browning, Christopher R., Burrington, Lori A., Leventhal, Tama, Brooks-Gunn, Jeanne, 2008. Neighborhood structural inequality, collective efficacy, and sexual risk behavior among urban youth. Journal of Health and Social Behavior 49 (3), 269–285.

Browning, Christopher R., Cagney, Kathleen A., 2002. Neighborhood structural disadvantage, collective efficacy, and self-rated physical health in an urban setting. Journal of Health and Social Behavior 43 (4), 383–399.

Browning, Christopher R., Feinberg, Seth L., Dietz, Robert D., 2004. The paradox of social organization: networks, collective efficacy, and violent crime in urban neighborhoods. Social Forces 83 (2), 503–534.

Browning, Christopher R., Wallace, Danielle, Feinberg, Seth L., Cagney, Kathleen A., 2006. Neighborhood social processes, physical conditions, and disaster-related mortality: the case of the 1995 Chicago heat wave. American Sociological Review 71 (4), 661–678.

Burchfield, Keri B., 2009. Attachment as a source of informal social control in urban neighborhoods. Journal of Criminal Justice 37 (1), 45–54.

Burchfield, Keri B., Silver, Eric, 2013. Collective efficacy and crime in Los Angeles neighborhoods: implications for the Latino paradox. Sociological Inquiry 83 (1), 154–176.

Bursik, Robert J., 1999. The informal control of crime through neighborhood networks. Sociological Focus 32 (1), 85–97.

Coulton, C.J., Jennings, M.Z., Chan, T., 2013. How big is my neighborhood? individual and contextual effects on perceptions of neighborhood scale. American Journal of Community Psychology 51 (1–2), 140–150.

Duncan, Terry E., Duncan, Susan C., Okut, Hayrettin, Strycker, Lisa A., Hix-Small, Hollie, 2003. A multilevel contextual model of neighborhood collective efficacy. American Journal of Community Psychology 32 (3–4), 245–252.

Erickson, Patricia G., Harrison, Lana, Cook, Steven, Cousineau, M-Marthe, Adlaf, Edward M., 2012. A comparative study of the influence of collective efficacy on substance use among adolescent students in Philadelphia, Toronto, and Montreal. Addiction Research & Theory 20 (1), 11–20.

Garcia, R. Marie, Taylor, Ralph B., Lawton, Brian A., 2007. Impacts of violent crime and neighborhood structure on trusting your neighbors. Justice Quarterly 24 (4), 679–704.

Gibson, Chris L., Sullivan, Christopher J., Jones, Shayne, Piquero, Alex R., 2010. 'Does it take a village?' Assessing neighborhood influences on children's self-control. Journal of Research in Crime and Delinquency 47 (1), 31–62.

Hampton, Keith. 2010. Glocalization: internet use and collective efficacy. American Behavioral Scientist 53, 1111–1132.

Hipp, John R., Butts, Carter T., Acton, Ryan M., Nagle, Nicholas N., Boessen, Adam, 2013. Extrapolative simulation of neighborhood networks based on population spatial distribution: do they predict crime? Social Networks 35 (4), 614–625.

Horne, Christine, 2004. Collective benefits, exchange interests, and norm enforcement. Social Forces 82 (3), 1037–1062.

Kasarda, John D., Janowitz, Morris, 1974. Community attachment in mass society. American Sociological Review 39 (3), 328–339.

Kirk, David S., 2009. Unraveling the contextual effects on student suspension and juvenile arrest: the independent and interdependent influences of school, neighborhood, and family social controls. Criminology 47 (2), 479–520.

Kirk, David S., Matsuda, Mauri, 2011. Legal cynicism, collective efficacy, and the ecology of arrest. Criminology 49 (2), 443–472.

Mazerolle, Lorraine, Wickes, Rebecca, McBroom, James, 2010. Community variations in violence: the role of social ties and collective efficacy in comparative context. Journal of Research in Crime and Delinquency 47 (1), 3–30.

Morenoff, Jeffrey D., Sampson, Robert J., Raudenbush, Stephen W., 2001. Neighborhood inequality, collective efficacy, and the spatial dynamics of urban violence. Criminology 39 (3), 517–559.

Odgers, Candice L., Moffitt, Terrie E., Tach, Laura M., Sampson, Robert J., Taylor, Alan, Matthews, Charlotte L., Caspi, Avshalom, 2009. The protective effects of neighborhood collective efficacy on British children growing up in deprivation: a developmental analysis. Developmental Psychology 45 (4), 942–957.

Portes, Alejandro, 1998. Social capital: its origins and applications in modern sociology. Annual Review of Sociology 22 (1), 1–24.

Rankin, Bruce H., Quane, James M., 2002. Social Contexts and urban adolescent outcomes: the interrelated effects of neighborhoods, families, and peers on African-American youth. Social Problems 49 (1), 79–100.

Reisig, Michael D., Cancino, Jeffrey Michael, 2004. Incivilities in nonmetropolitan communities: the effects of structural constraints, social conditions, and crime. Journal of Criminal Justice 32 (1), 15–29.

Rhineberger-Dunn, Gayle M., Carlson, Susan M., 2011. An analysis of the mediating effects of social Relations and controls on neighborhood crime victimization. Western Criminology Review 12 (1), 15–34.

Sampson, Robert J., 2006a. Collective efficacy theory: lessons learned and directions for future inquiry. In: Cullen, F.T., Wright, J.P., Blevins, K.R. (Eds.), Taking Stock: The Status of Criminological Theory. Transaction, New Brunswick.

Sampson, Robert J., 2006b. How does community context matter? Social mechanisms, and the explanation of crime rates. In: Wikstrom, P.-O.H., Sampson, R.J. (Eds.), The Explanation of Crime: Context, Mechanisms, and Development. Cambridge University Press, New York.

Sampson, Robert J., 2012. Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago, Chicago.

Sampson, Robert J., McAdam, Doug, MacIndoe, Heather, Weffer-Elizondo, Simon, 2006. Civil society reconsidered: the durable nature and community organizational context of collective civic action. American Journal of Sociology 111 (3), 673–714.

Sampson, Robert J., Morenoff, Jeffrey D., Raudenbush, Stephen W., 2005. Social anatomy of racial and ethnic disparities in violence. American Journal of Public Health 95 (2), 224–232.

Sampson, Robert J., Raudenbush, Stephen W., 1999. Systematic social observation of public spaces: a new look at disorder in urban neighborhoods. American Journal of Sociology 105 (3), 603–651.

Sampson, Robert J., Raudenbush, Stephen W., Earls, Felton, 15 August 1997. Neighborhoods and violent crime: a multilevel study of collective efficacy. Science 277, 918–924.

Sampson, Robert J., Wikstrom, Per-Olof H., 2008. The social order of violence in Chicago and Stockholm neighborhoods: a comparative inquiry. In: Kalyvas, S.N., Shapiro, I., Masoud, T. (Eds.), Order, Conflict, and Violence. Cambridge University Press, Cambridge, MA.

Schreck, Christopher J., McGloin, Jean Marie, Kirk, David S., 2009. On the origins of the violent neighborhood: a study of the nature and predictors of crime-type differentiation across Chicago neighborhoods. Justice Quarterly 26 (4), 7710794.

Sharkey, Patrick T., 2006. Navigating dangerous streets: the sources and consequences of street efficacy. American Sociological Review 71 (5), 826–846.

Silver, Eric, Miller, Lisa L., 2004. Sources of informal social control in Chicago neighborhoods. Criminology 42 (3), 551–584.

Simons, Ronald L., Simons, Leslie Gordon, Burt, Callie Harbin, Brody, Gene H., Cutrona, Carolyn, 2005. Collective efficacy, authoritative parenting and delinquency: a longitudinal test of a model integrating community- and family-level processes. Criminology 43 (4), 989–1029.

St. Jean, Peter K.B., 2007. Pockets of Crime: Broken Windows, Collective Efficacy, and the Criminal Point of View. University of Chicago, Chicago.

Steenbeek, Wouter, Hipp, John R., 2011. A longitudinal test of social disorganization theory: feedback effects among cohesion, social control and disorder. Criminology 49 (3), 833–871.

Warner, Barbara D., 2007. Directly intervene or call the authorities? a study of forms of neighborhood social control within a social disorganization framework. Criminology 45 (1), 99.

Weisburd, David, Groff, Elizabeth, Yang, Sue-Ming, 2012. The Criminology of Place. Oxford, New York.

Wickes, Rebecca, Hipp, John R., Sargeant, Elise, Homel, Ross, 2013. Collective efficacy as a task specific process: examining the relationship between social ties, neighborhood cohesion and the capacity to respond to violence, delinquency and civic problems. American Journal of Community Psychology 52 (1–2), 115–127.

Wickes, Rebecca, Hipp, John R., Sargeant, Elise, Mazerolle, Lorraine, 2012. Social Ties, Collective Efficacy and Actions: What Influences Informal Social Control Behaviors? Under review.

Wickes, Rebecca L., 2010. Generating action and responding to local issues collective efficacy in context. Australian and New Zealand Journal of Criminology 43 (3), 423–443.

Williams, Kirk R., Guerra, Nancy G., 2011. Perceptions of collective efficacy and bullying perpetration in schools. Social Problems 58 (1), 126–143.

Wright, Emily M., Benson, Michael L., 2011. Clarifying the effects of neighborhood context on violence 'Behind Closed Doors'. JQ: Justice Quarterly 28 (5), 775–798.

Zaccaro, Stephen J., Blair, Virginia, Peterson, Christopher, Zazanis, M., 1995. Collective efficacy. In: Maddux, J.E. (Ed.), Self-efficacy, Adaptation, and Adjustment: Theory, Research, and Application. Plenum, New York.

Zhang, Lening, Messner, Steven F., Liu, Jianhong, 2007. A multilevel analysis of the risk of Household Burglary in the city of Tianjin, China. British Journal of Criminology 47 (6), 918–937.

Zhang, Lening, Messner, Steven F., Liu, Jianhong, Zhuo, Yue Angela, 2002. Guanxi and fear of crime in contemporary urban China. British Journal of Criminology 49 (4), 472–490.

Zimmerman, Gregory, 2010. Impulsivity, offending, and the neighborhood: investigating the Person–context nexus. Journal of Quantitative Criminology 26 (3), 301–332.

Collective Identity

David A Snow, University of California, Irvine, CA, USA
Catherine Corrigall-Brown, University of Western Ontario, London, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D.A. Snow, volume 11, pp. 2212–2219, © 2001, Elsevier Ltd.

Abstract

This article provides an analytic overview of scholarly work on the concept of collective identity by considering its conceptualization and various empirical manifestations, the analytic approaches informing its discussion and analysis, and a number of theoretical and empirical issues, including a synopsis of the symbolic means through which collective identity is expressed and asserted. Although the scholarly roots of the concept can be traced to classical sociologists such as Marx and Durkheim, and more recently to the midcentury work of scholars such as *Erik Erickson* and *Erving Goffman*, it was not until the latter quarter of the past century that the concept generated an outpouring of work invoking the concept directly or referring to it indirectly through the linkage of various collectivities and their identity interests via such concepts as *identity politics*, *contested identities*, *insurgent identities*, *nationalism*, *imagined communities*, *identity movements*, and even *social movements* more generally. Conceptually, the essence of collective identity resides in a shared and interactive sense of 'we-ness' and 'collective agency.' Although the concept is distinguished analytically from both *personal identity* and *social identity*, the three types of identity clearly overlap and interact. Empirically, collective identity can surface in a variety of contexts, although the preponderance of research has focused on its connection to *gender*, *ethnicity*, *religion*, *nationalism*, and particularly *social movements*. Analytically, collective identity has generally been discussed from a primordial, structural, and/or constructionist standpoint. Primordial and structural approaches are discussed as variants of *essentialism*, which is contrasted to *constructionism*. Among other things, constructionism focuses attention on the *symbolic expression* and maintenance of collective identities.

If key concepts or expressions can be identified that function to capture the animating spirit of different epochs, then certainly one candidate concept for the last 40–50 years is the concept of collective identity. Indeed, it is a concept that came of age in the latter part of the past century, as reflected in the outpouring of scholarly work invoking the concept directly or referring to it indirectly through the linkage of various collectivities and their identity interests via such concepts as identity politics, contested identities, insurgent identities, and identity movements. This article provides an analytic overview of scholarly work on the concept by considering, in order, its conceptualization, its various empirical manifestations, the analytic approaches informing its discussion and analysis, and several unresolved theoretical and empirical issues.

Conceptualization

The concept of collective identity, just as the base concept of identity, is rooted in the observation that interaction between two or more sets of actors minimally requires that they be situated or placed as social objects. To do so is to announce or impute identities. Hence, interaction among individuals and groups, as social objects, is contingent on the reciprocal attribution and avowal of identities. This process holds for both individuals and collectivities, and it probably has always been a characteristic feature of human interaction. To note this is not to ignore the sociological truism that the issue of identity becomes more problematic and unsettled as societies become more structurally differentiated, fragmented, and culturally pluralistic (Castells, 1997; Giddens, 1991). But historical variation in the extent to which matters of identity are problematic does not undermine the double-edged observation that the reciprocal imputation and avowal of identities is a necessary condition for social interaction and that identities are thus rooted in the requisite conditions for social interaction.

Delineating the interactional roots of identities does not explain what is distinctive about collective identity, as there are at least three conceptually distinct types of identity: personal, social, and collective. Although they often overlap, one cannot be inferred from the other. Hence, the necessity of distinguishing among them.

Social identities are the identities attributed or imputed to others in an attempt to situate them in social space. They are grounded typically in established social roles, such as 'teacher' and 'mother,' or in broader and more inclusive social categories, such as gender categories or ethnic and national categories, and thus are often referred to as 'role identities' (Stryker, 1980) and 'categorical identities' (Calhoun, 1997). Whatever their specific sociocultural base, social identities are fundamental to social interaction in that they provide points of orientation to 'alter' or categorize 'other' as a social object.

Personal identities are the attributes and meanings attributed to oneself by the actor; they are self-designations and self-attributions regarded as personally distinctive. They are especially likely to be asserted during the course of interaction when other imputed social identities are regarded as contradictory, as when individuals are cast into social roles or categories that are insulting and demeaning (Snow and Anderson, 1987). Thus, personal identities may derive from role incumbency or category-based memberships, but they are not necessarily comparable since the relative salience of social roles or category membership with respect to personal identity can be quite variable.

Just as social and personal identities are different yet typically overlapping and interacting constructs, such is the relationship between collective and social and personal identities. Although there is no consensual definition of collective identity, discussions of the concept invariably suggest that its essence resides in a shared sense of 'one-ness' or 'we-ness' anchored in real or imagined shared attributes and experiences among those who comprise the collectivity and in relation or contrast to one or more actual or imagined sets of 'others.' Embedded within the shared sense of 'we' is a corresponding sense of 'collective agency.' This latter sense, which is the action component of collective identity, not only suggests the possibility of collective action in pursuit of common interests, but even invites such action. Thus, it can be argued that collective identity is constituted by a shared and interactive sense of 'we-ness' and 'collective agency.' This double-edged sense can be culled from classic sociological constructs such as Durkheim's 'collective conscience' and Marx's 'class consciousness,' but is reflected even more clearly in most conceptual discussions of collective identity, although the agentic dimension is sometimes implied rather than directly articulated (e.g., Castells, 1997; Cerulo, 1997; Jasper and Polletta, 2001; Jensen, 1995; Melucci, 1989, 1995).

A common theme running throughout a segment of the literature on collective identity is the insistence that collective identity is, at its core, a process rather than a property of social actors. Such work acknowledges that collective identity is 'an interactive and shared definition' that is evocative of "a sense of 'we,'" but then highlights the process through which social actors recognize themselves as a collectivity, contending that this process is more vital to conceptualizing collective identity than any resultant product or property (e.g., Melucci, 1989: pp. 34, 218, passim). Few scholars would take exception with the importance of the process through which collective identities develop, but it is both questionable and unnecessary to contend that the process is more fundamental than the product to understanding the character and functionality of collective identity. Not only is the product or 'shared we' generative of a sense of agency that can be a powerful impetus to collective action, but it functions, as well, as the orientational identity for other actors within the field of action. More concretely, it is the constructed social object to which the movement's protagonists, adversaries, and audience(s) respond (Hunt et al., 1994), and which, in turn, may have implications for the operation of its organizational carrier, affecting the availability and character of allies, resources, and even tactical possibilities (Jensen, 1995).

The initial projected collective identity may be short lived and transient, subject to modification and even transformation during the course of ongoing collective (inter) action. Whatever the character of the collective identity, the underlying collective interaction often occurs through participation in social groups. For example, Jones' (2011) work on collective identity building in a mixed-race student organization highlights how identity formation can be a group process involving strategies to achieve a sense of belonging and cohesion through the micropractices that are essential to group formation. And, these emerging collective identities can, reflexively, shape the personal and social identities of group members.

Collective identities are distinguished from social and personal identities in several ways. First, collective identities may or may not be embedded in existing social identities, since they are often emergent and evolving rather than firmly rooted in prior social categories. This is often the case with the collective identities that emerge in the course of dynamic social protest events (for illuminating examples, see Calhoun's and Zhao's accounts of the Chinese student movement of 1989) or over time in relation to larger historical changes (e.g., Roth's work on the emergence of collective identity in the US women's movement 2008). Second, the collective, shared 'sense of we' is animating and mobilizing cognitively, emotionally, and sometimes even morally. The shared perceptions and feelings of a common cause, threat, or fate that constitute the shared 'sense of we' motivate people to act together in the name of, or for the sake of, the interests of the collectivity, thus generating the previously mentioned sense of collective agency. That potential inheres within social identities, but they typically function more like orientational markers as the routines of everyday life are negotiated. When they are activated or infused affectively and morally, it is arguable that they have been transformed into collective identities. Third, the emergence and operation of collective identities mean that other social identities have subsided in relevance and salience for the time being. In other words, collective identities, when they are operative, generally have claims over – not so much normatively as cognitively and emotionally – other identities in terms of the object of orientation and character of corresponding action. Examples abound, as observed frequently in the case of many protest gatherings, gripping fads, joyous and celebratory sports crowds, and the concerted campaigns and actions associated with social movement activism. Fourth, while collective identities and personal identities are obviously different, they are still very much interconnected in the sense that collective identities are predicated, in part, on constituents' embracement of the relevant collective identity as a highly salient part of their personal identity and sense of self (Gamson, 1991). And finally, while the attribution or avowal of all identities is interactionally contingent, collective identities tend to be more fluid, tentative, and transient than either categorically based social identities or even personal identities.

Empirical Manifestations

Empirical manifestations of collective identity can vary in a number of significant ways. One important axis of variation is the size of the collectivity, and the corresponding scope of its claims. If the essence of collective identity resides in a sense of 'we-ness' associated with real or imagined attributes in contrast to some set of others, then it follows that collective identities can surface among almost any grouping or aggregation in a variety of contexts. These could range from relatively small cliques and gangs, to sports fans and celebrity devotees, to laborers and occupational groupings, to neighborhoods and communities, and even to broader categories such as sexual and gender categories, religions, ethnic groups, and nations. The preponderance of empirical research on collective identity has focused on the last, more inclusive set of categories – sexuality and gender, religion, ethnicity, and nationality.

Illustrative is Ward's (2008) research on gay and lesbian identity within organizations; Wilkes' (2006) analysis of the importance of collective identity for mobilization among Indigenous peoples; Cornell and Hartmann's (1998) overview of the construction of ethnic and racial identities in the modern world; and Anderson's (2006) and Calhoun's work on nationalism, which the latter defines, in part, as one "way of constructing collective identities" (1997: p. 29).

Social movement mobilizations also tend to be both generative of and dependent on collective identities (e.g., Bernstein, 2005; Gamson, 1991; Hunt et al., 1994; Jasper and Polletta, 2001; Melucci, 1989; Snow and McAdam, 2000; Van Stekelenberg, 2013). In fact, one of the most consistent findings in this area of research is that identifying with a social movement cause or collectivity enhances the probability of participation for individuals across a variety of movements (Corrigall-Brown et al., 2009; Dixon and Roscigno, 2003; Liss et al., 2004).

Although collective identities can congeal in various aggregations and contexts, they appear not to do so on a continuous basis historically. Instead, their emergence and vitality appear to be associated with conditions of sociocultural change or challenge, socioeconomic and political exclusion, and political breakdown and renewal, thus suggesting that they cluster historically in social space. The past 50 years has generally been regarded as one such period of collective identity effervescence and clustering, with some scholars characterizing this period in terms of identity crises and collective searches for identity (e.g., Castells, 1997; Gergen, 1991; Giddens, 1991; Klapp, 1969). In *The Power of Identity*, Castells captures both this characterization and the kinds of conditions thought to be associated with the various manifestations of collective identity during this period:

> Along with the technological revolution, the transformation of capitalism, and the demise of statism, we have experienced, in the last quarter of the century, the widespread surge of powerful expressions of collective identity that challenge globalization and cosmopolitanism on behalf of cultural singularity and people's control over their lives and environment. (Castells, 1997: p. 2)

Analytic Approaches

To note that expressions of collective identities cluster historically according to the conjunction of various social conditions does not specify the character or content of the emergent collective identities. This issue has been addressed from the vantage point of three contrasting perspectives: primordialism, social structuralism, and social constructionism.

Both the primordialist and structuralist views can be construed as variants of an overarching essentialist perspective which posits that a collectivity's identity basically flows naturally from some underlying set of characteristics, often reduced to a single determinative attribute regarded as the collectivity's 'defining essence.' From the primordialist point of view, the defining characteristic is typically an ascriptive attribute, such as race, gender, or sexual orientation, or sometimes a deep, underlying psychological or personality disposition.

From a structuralist perspective, the critical characteristic is typically a kind of master social category implying structural commonality, such as social class, ethnicity, or nationality; a set of relational ties or networks suggesting structural connectedness; or a mixture of both. Individuals who are similarly situated structurally, such that they are incumbents of similar roles, work in similar enterprises, are linked to the same social networks, or are members of the same social class, religion, or ethnic group, are presumed to have a shared collective identity or at least be candidates for such.

The constructionist perspective, in general, rejects both the primodialist and structuralist variants of the essentialist argument, seeing the presumed link between identities and their ascriptive or structural moorings as being more indeterminate than postulated. Instead, attention is shifted to the construction and maintenance of collective identities. Collective identities are seen as invented, created, reconstituted, or cobbled together rather than being biologically preordained or structurally or culturally determined.

Assessment of the three perspectives in terms of their relative analytic utility for explaining the character and content of collective identities reveals considerable support for the constructionist thesis. This may be due in part to the currents of fashion, influenced by the winds of multiculturalism, postmodernism, and identity politics, but it is also due, more importantly, to other factors. One is that the hypothesized link between identities and the primordial attributes or structural categories in which they are presumably anchored is too mechanistic. In its hard version, it is contradicted by the sociological observation that people are often members of the same categories or groups in different ways and with varying degrees of commitment and identification, thus suggesting that inferring correspondence between personal, social, and collective identities solely on the basis of primordial or structural categories is empirically suspect. Additionally, the claims of primordialist and structuralist arguments do not fare well when confronted with the observation that people generally have multiple identities (e.g., family, work, leisure, gender, ethnic, religious, political, and national) that are differentially invoked or avowed depending on their relative salience and their situational pervasiveness. Salience refers to the relative importance of an identity in relation to other identities (Stryker, 1980); pervasiveness or comprehensiveness refers to the situational relevance or reach of any particular identity and the corresponding degree to which it organizes social life, including collective action (Cornell and Hartmann, 1998; Snow and McAdam, 2000). Given the fact that increasing numbers of individuals live in a world in which they are the carriers of multiple and often conflicting identities, what determines any particular identity's relative salience and pervasiveness, and thus the influence of its claims, vis-à-vis others? Clearly such matters are not determined solely by an identity's primordial roots or structural footing. And finally, much of the empirical evidence is consistent with the constructionist argument. Two highly evocative examples include Trevor-Roper's (1983) account of the retrospective invention of the distinctive Highland culture and tradition so redolently associated with all of Scotland, and James' parallel conclusion, based on extensive archeological and archival research, regarding the origins of the modern Celts:

> …the idea of a race, nation or ethnic group called Celts in Ancient Britain and Ireland is indeed a modern invention. It is an eighteenth- and nineteenth-century reification of a people that never existed, a factoid…assembled from fragments of evidence drawn from a wide range of societies across space and time. This reification served the interests of a range of cultural expectations, aspirations, and political agendas – and still does. (James, 1999: p. 136)

Such conclusions should not be read as unequivocal refutations of the primordialist and structuralist arguments, since constructed identities are not fabricated whole cloth but typically knit together by drawing on threads of past and current cultural materials and traditions, structural arrangements, and even primordial attributes. These materials and attributes constitute the kinds of things from which collective identities – particularly ethnic, religious, and national ones – are fashioned, and thereby function, in varying degrees, to constrain the construction process. With this in mind, it is advisable to consider these theories as not competing but, instead, as potentially complementary. This complementarity is illustrated in Eriksen's work on the emergence of nationalism (2004) which shows that both kinship (ethnicity) and place (territory) – basic elements of the more traditional primordial and structuralist perspectives – are important elements in the construction of a national identity.

Interpretative constraint also may be exercised by the institutional contexts and relations of power in which contestants are embedded (Castells, 1997; Jensen, 1995). Additionally, analyses of the relationship between collective identity and participation in social movements repeatedly point to the experience of collective action itself as a fertile seedbed for the generation of collective identities (Calhoun, 1991; Fantasia, 1988; Melucci, 1989) providing what Dufour and Giraud (2007) describe as "collective identity moments." Thus, while collective identities are undeniably constructed, they rarely are constructed *carte blanche*; rather, they typically are forged not only with the materials suggested by the primordialist and structuralist perspectives, but with and through the experience of collective action itself.

Theoretical and Empirical Issues

Relevant to a thoroughgoing understanding of collective identity, whatever its empirical locus, are several theoretical and empirical issues that require more careful consideration than often accorded.

Identity Work (the Expression of Collective Identities)

Fundamental to understanding collective identity, particularly from a constructionist standpoint, are the processes through which it is created, expressed, sustained, and modified. These processes have been conceptualized as variants of 'identity work,' which encompasses the range of activities people engage in, both individually and collectively, to signify and express who they are and what they stand for in relation or contrast to some set of others (Einwohner, 2013; Schwalbe and Mason-Schrock, 1996; Snow and Anderson, 1987; Snow and McAdam, 2000). In some times and contexts, identity work can be more difficult than at other times and in other contexts. In essence, the 'identity workload' can be heavier (Reger et al., 2008). At its core, identity work is the generation, invocation, and maintenance of symbolic resources used to bound and distinguish the collectivity both internally and externally by accenting commonalities and differences (Schwalbe and Mason-Schrock, 1996; Taylor and Whittier, 1992).

While these collective movement identities are often oppositional, recent work cautions that the line between 'us' and 'them' is not always clear and impermeable (Reger et al., 2008). Ghaziani's (2011) study of gay identity over time illustrates the dynamic relationship between movement collective identities and 'others,' arguing that we live in a 'post-Gay' society in which the gay and lesbian community has moved from an oppositional (us versus them) identity to an inclusive (us and them) one. This work highlights how collective identities are constructed and reconstructed over time.

Symbolic resources include the interpretive frameworks (or frames), avowed and imputed names, and dramaturgical codes of expression and demeanor (e.g., particularistic styles of storytelling, dress, adornment, and music) that are generated and employed during the course of a collectivity's efforts to distinguish itself from one or more other collectivities. Hunt and Benford (2004) argue that "fundamentally, collective identities are *talked* into existence … personal and collective identities shape and are shaped by collective action and the subsequent identity talk" (p. 445). In addition to identity talk, groups use particular songs and styles of music that invite participation and that are politically and emotionally evocative, such as 'We Shall Overcome' (Eyerman and Jamison, 1998; Roy, 2010); key words and slogans that function in a similar fashion, such as '*Liberte*, *Fraternite*, and *Egalite*' and 'Workers of the World Unite'; and systems of gestures and signs, such as the raised clinched fist and the peace sign, that function similarly to the tradition of heraldry. These and other symbolic resources function as boundary markers of collective differentiation, distinguishing insiders from outsiders, or protagonists from antagonists, in a fashion that heightens awareness of in-group commonalities and connections and out-group differences. Together they congeal into a kind of 'semiotic bricolage' (Schwalbe and Mason-Schrock, 1996) that gives symbolic substance to the claimed distinctive 'we,' and it is largely through this bricolage that collective identity is expressed and known publicly. While the boundary-making and -maintenance functions of these symbolic resources, or bricolage, are widely acknowledged, what accounts for the differential resonance or carrying power of different symbolic markers is less well understood.

The Problem of Identity Correspondence

A not uncommon problem with analyses of collective identity is the tendency to reify the collective identity, and thus take-for-granted the link between the individuals that make-up the collectivity and the shared, overarching identity. This gloss is particularly troublesome in light of the observation that people typically have multiple identities that vary in salience and pervasiveness. Thus, how is any particular collective identity reconciled with other identities adherents possess? Such

questions allude to what has been referred to as the problem of 'identity correspondence' – the alignment or linkage of personal and collective identities (Snow and McAdam, 2000).

One general answer is that this alignment "is accomplished by enlarging the personal identities of a constituency to include the relevant collective identities as part of their definition of self" (Gamson, 1991: p. 41). But what are the processes through which prospective adherents come to embrace the relevant collective identity, such that personal and collective identities are correspondent or congruous? Two broad processes have been suggested: identity convergence and identity construction.

Identity convergence refers to the union of personal and collective identities when both are congruent, such that an extant collectivity provides a venue for an individual to act in accordance with his or her personal identity. The analytic problem is not one of identity construction or transformation, but one of linkage or bridging and the identification of the mechanisms that facilitate the convergence. A number of such mechanisms have been identified. One operates at the organizational level, entailing the occasional appropriation of existing solidary networks by movement organizations (Snow and McAdam, 2000); the other mechanisms are variants of rational choice processes. One is based on a 'tipping' or 'threshold' model, which posits that collective identities are assumed by individuals when the perceived actions of others reach a point that suggests that the payoffs for adopting, or at least acting in accordance with, the collective identity outweigh doing otherwise. Illustrative is the contention that such tipping points played a critical role in explaining language and identity shifts among Russian-speaking immigrants 'beached' in four of the republics of the former Soviet Union (Laitin, 1998). A related mechanism is the existence of intergenerational investments in personal identities that may have implications for the embracement of future collective identities (Laitin, 1998). A third rational choice explanation holds that the collective identities associated with social movements – and, by implication, with other collectivities as well – can be regarded as 'selective incentives' for those who seek to express and affirm their personal identities (Friedman and McAdam, 1992). The explanatory utility of these arguments, as well as the network appropriation thesis, is contingent on two underlying assumptions: that there is preestablished congruence among some number of personal identities and the proffered or available collective identities; and that collective identities are constituted by the aggregation or convergence of parallel personal identities. While the first assumption is empirically tenable, the second is questionable from the vantage point of many scholars of collective identity (e.g., Jasper and Polletta, 2001; Melucci, 1989, 1995).

In the absence of correspondence between personal identities and collective identities, some variety of identity work is necessary in order to facilitate their alignment. This alignment can vary significantly, ranging from the elevation of the salience of a particular identity to a fairly dramatic change in one's sense of self. Four identity construction processes have been identified that capture this variation: identity amplification, identity consolidation, identity extension, and identity transformation (Snow and McAdam, 2000; see also Robnett, 2005). Identity amplification affects a change in an individual's identity salience hierarchy, such that an existing but lower-order identity becomes sufficiently salient to ensure engagement in collective action, as in the many cases in which in the identity of woman was elevated and expanded in conjunction with the Women's Movement; identity consolidation involves the adoption of an identity that is a blend of two prior but seemingly incompatible identities, as in the case of the union of environmentalists and labor activists and 'Jews for Jesus'; identity extension entails the expansion of the situational pervasiveness of an individual's personal identity so that its reach is congruent with the collectivity's, as when individuals come to see themselves as representatives for a specific cause that transcends other role obligations and identities; and identity transformation involves a dramatic change in identity, such that individuals now see themselves as remarkably different than before, as often occurs in the case of conversion to a new group or movement.

The mechanisms or processes underlying these various forms of identity construction include framing processes in which identities are announced or renounced, embraced or rejected, and modified or reframed in the course of various interactions with adherents, antagonists, and bystander audiences (Benford and Snow, 2000; Hunt et al., 1994); engagement in collective action, as when direct observation or experience functions as a demonstration event that gives rise to a situationally specific collective identity or affirms collective claims and thus helps to render salient, and perhaps pervasive, what was previously a secondary or marginal personal or social identity (Calhoun, 1991; Melucci, 1989); or some combination of both framing and actual engagement.

Given the variety of ways in which identity correspondence can be affected, the question arises as to whether the relevance of the convergence and construction processes varies by type of collectivity. In the case of social movements, for example, it has been hypothesized that movements that are culturally different, greedy in terms of the cognitive and behavioral demands, and/or politically radical are likely to rely more on identity construction than convergence processes (Snow and McAdam, 2000). Whether this is the case is an empirical question, but it does caution against presuming that what accounts for identity correspondence and shifts in one context necessarily holds for another. That context is clearly significant is highlighted in Reger's (2002) work analyzing different chapters of the National Organization for Women (NOW). She shows that the distinct cultural and geographical contexts of the local chapters affected the feminist identity constructed by women at each chapter. The relationship of context and identity construction is a fruitful avenue for future research.

Just as there may be variation between types of collectivities and the processes of identity convergence and identity construction, so it is reasonable to ask whether these processes might vary in importance at different points in the life course of a social movement or ethnic or nationalist mobilization? Rather than assuming that a particular process, such as the tipping process or identity amplification, operates routinely with respect to the emergence of a collective identity, might not these processes be more relevant at particular junctures in the career of a movement's collective identity? Preliminary consideration of such questions suggests that network appropriation, rational choice, and constructionist explanations, rather than being mutually exclusive and antithetical, may

interact and combine in interesting ways in explaining the emergence, institutionalization, and diffusion of collective identities across different contexts (Snow and McAdam, 2000).

Dimensions, Layers, and Types of Collective Identity

Although there is an extensive literature on collective identity, with considerable discussion regarding its conceptualization and sources, this literature has been relatively mute regarding variation in its form. The concept most often is invoked as if it is an invariant, uniform collective phenomenon. This is not the case, however, as collective identities can be multidimensional and multilayered within a specific locus, and they may also vary by type. The multidimensionality of collective identity is indicated by reference to its cognitive, emotional, and moral dimensions (Jasper and Polletta, 2001; Melucci, 1989). The relative importance of each of these dimensions to the vitality and motivational force of a collective identity has not been elaborated, however. Presumably, the presence of each of these dimensions yields a more robust and vital collective identity. Clearly a collective identity in which the boundaries between 'us' and 'them' are unambiguously drawn, in which there is strong feeling about those differences, and in which there is a sense of moral virtue associated with both the perceptions and feelings, should be a more potent collective identity than one in which either the emotional or moral dimensions are weakly developed.

Similarly, several analyses have noted that collective identity can be multilayered, such that there can be variation in its locus. Three such layers have been noted with respect to social movements (Gamson, 1991; Stoeker, 1995). They include, beginning with the broadest and potentially most inclusive layer: the social movement community or solidary group, which can be thought of as the constituent layer, as in the case of black Americans in relation to the civil rights movement; the social movement layer as in the case of the civil rights movement; and the organizational layer, as represented by the Southern Christian Leadership Conference (SCLC), the Congress for Racial Equality (CORE), and the Student Nonviolent Coordinating Committee (SNCC) in the context of the civil rights movement (alternatively divided as the 'activist,' 'organizational' and 'tactical' identity by Jasper (1997). In principle, each successive layer may be embedded in the larger more inclusive layer, giving rise to a generalized, cohesive collective identity at the community or national level. But clearly the existence of a collective identity at one level does not automatically generalize to or incorporate another level and each level of identity may exist alongside both movement identities and preexisting collective identities, interweaving with them in complex ways. Thus, collective identities can be built around the organizational carriers of a movement, as in the case of SCLC and SNCC, without necessarily representing the broader movement, which indicates the potential for identity conflicts at the collective level and the potential for schism and factionalization. Such observations suggest the need for more careful consideration of the often multilayered character of collective identities and of greater specification of the ways they can interact and combine, and with what consequences. As well, these observations call for caution in generalizing about the scope and functioning of collective identities, particularly with respect to broader social categories, such as ethnicities and nationalities.

Finally, it is reasonable to wonder if collective identities vary by type. At the most general level, Hunt et al. (1994) distinguish among protagonist, antagonist, and audience or bystander identities, arguing that even though protagonist or oppositional identities have received most of the scholarly attention, each type or field of identity is fundamental to understanding the interactive dynamics underlying the emergence, character, and functioning of a collective identities. Noting as well that collective identities arise and operate within an interactive context 'marked by power relationships,' Castells distinguishes among legitimizing, resistance, and project collective identities (1997: pp. 7–10). Legitimizing identities are associated with dominant institutions or the state, whereas both resistance and project identities represent two basic forms of the antagonist identity, the former generated by devalued and stigmatized collectivities, and thus constituting the seedbed for identity movements and politics, and the latter associated with movement beyond resistance to the construction not only of alternative identities but also a new system that valorizes rather than defiles the new identity. The important issue is not whether such typologies of collective identities are exhaustive, but the emphasis on their contextually embedded and interactional character and their different consequences.

See also: Collective Behavior, Social Psychology of; Collective Beliefs, Sociological Explanation of; Collective Memory, Social Psychology of; Essentialism; Gay and Lesbian Movements; Globalization and World Culture; Identity Movements, Identity Movements; Nation-State as Symbolic Construct; Nationalism: General; Personal Identity: Philosophical Aspects; Social Identity in Sociology; Social Movements: A Social Psychological Perspective; Symbolic Boundaries.

Bibliography

Anderson, B., 2006. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London, UK.

Benford, R.D., Snow, D.A., 2000. Framing processes and social movements: an overview and assessment. Annual Review of Sociology 26, 611–639.

Bernstein, M., 2005. Identity politics. Annual Review of Sociology 31, 47–74.

Calhoun, C., 1991. The problem of identity in collective action. In: Huber, J. (Ed.), Macro-micro Linkages in Sociology. Sage Publications, Newbury Park, CA.

Calhoun, C., 1997. Nationalism. University of Minnesota Press, Minneapolis, MN.

Castells, M., 1997. The Power of Identity. Blackwell Publishers, Oxford, England.

Cerulo, K.A., 1997. Identity construction: new issues, new directions. Annual Review of Sociology 23, 385–409.

Cornell, S., Hartmann, D., 1998. Ethnicity and Race: Making Identities in a Changing World. Pine Forge Press, Thousand Oaks, CA.

Corrigall-Brown, C., Snow, D.A., Quist, T., Smith, K., 2009. Explaining the puzzle of homeless mobilization: an examination of differential participation. Sociological Perspectives 52, 309–335.

Dixon, M., Roscigno, V., 2003. Status, networks, and social movement participation: the case of striking workers. American Journal of Sociology 108, 1292–1327.

Dufour, P., Giraud, I., 2007. The continuity of transnational solidarities in the World March of Women, 2000 and 2005: a collective identity-building approach. Mobilization: An International Journal 12 (3), 307–322.

Eriksen, T.H., 2004. Place, kinship and the case for non-ethnic nations. Nations and Nationalism 10, 49–62.
Einwohner, R., 2013. Identity work processes. In: Snow, D.A., della Porta, D., Klandermans, B., McAdam, D. (Eds.), Wiley-Blackwell Encyclopedia of Social and Political Movements. Wiley-Blackwell, Oxford, UK.
Eyerman, R., Jamison, A., 1998. Music and Social Movements: Mobilizing Traditions in the Twentieth Century. Cambridge University Press, Cambridge, UK.
Fantasia, R., 1988. Cultures of Solidarity: Consciousness, Action, and Contemporary American Workers. University of California Press, Berkeley, CA.
Friedman, D., McAdam, D., 1992. Collective identity and activism: networks, choices, and the life of a social movement. In: Morris, A.D., Mueller, C.M. (Eds.), Frontiers in Social Movement Theory. Yale University Press, New Haven, CT.
Gamson, W.A., 1991. Commitment and agency in social movements. Sociological Forum 6, 27–50.
Gergen, K., 1991. The Saturated Self. Dilemmas of Identity in Contemporary Life. Basic Books, New York.
Ghaziani, A., 2011. Post-Gay collective identity construction. Social Problems 58, 99–125.
Giddens, A., 1991. Modernity and Self-identity: Self and Society in the Late Modern Age. Stanford University Press, Stanford, CA.
Hunt, S.A., Benford, R.D., 2004. Collective identity, solidarity, and commitment. In: Snow, D.A., Soule, S., Kriesi, H. (Eds.), The Blackwell Companion to Social Movements. Blackwell Publishing, Malden, MA.
Hunt, S.A., Benford, R.D., Snow, D.A., 1994. Identity fields: framing processes and the social construction of movement identities. In: Larana, E., Johnston, H., Gusfield, J.R. (Eds.), New Social Movements: From Ideology to Identity. Temple University Press, Philadelphia, PA.
Jasper, J.M., 1997. The Art of Moral Protest. University of Chicago Press, Chicago, IL.
Jasper, J.M., Polletta, F., 2001. Collective identity and social movements. Annual Review of Sociology 27, 283–305.
James, S., 1999. The Atlantic Celts: Ancient People or Modern Invention? British Museum Press, London.
Jensen, J., 1995. What's in a name? Nationalist movements and public discourse. In: Johnston, H., Klandermans, B. (Eds.), Social Movements and Culture. University of Minnesota Press, Minneapolis, MN.
Jones, J., 2011. Who are we? Producing group identity through everyday practices of conflict and discourse. Sociological Perspectives 54, 139–161.
Klapp, O.E., 1969. Collective Search for Identity. Holt, Rinehart and Winston, New York.
Laitin, D.D., 1998. Identity in Formation: The Russian-speaking Populations in the Near Abroad. Cornell University Press, Ithaca, NY.
Liss, M., Mary, C., Danielle, P., 2004. Predictors and correlates of collective action. Sex Roles 50, 771–779.
Melucci, A., 1989. Nomads of the Present: Social Movement and Identity Needs in Contemporary Society. Temple University Press, Philadelphia, PA.
Melucci, A., 1995. The process of collective identity. In: Johnston, H., Klandermans, B. (Eds.), Social Movements and Culture. University of Minnesota Press, Minneapolis, MN.
Reger, J., 2002. Organizational dynamics and the construction of multiple feminist identities in the national organization for women. Gender & Society 16, 710–727.
Reger, J., Meyers, D.J., Einwohner, R.L. (Eds.), 2008. Identity Work in Social Movements. University of Minnesota Press, Minneapolis, MN.
Robnett, B., 2005. We don't agree: collective identity justification work in social movement organizations. Research in Social Movements, Conflict and Change 26, 201–237.
Roth, B., 2008. The reconstruction of collective identity in the emergence of the U.S. white women's liberation. In: Reger, J., Myers, D.J., Einwohner, R.L. (Eds.), Identity Work in Social Movements. University of Minnesota Press, Minnesota, MN.
Roy, W.G., 2010. Reds, Whites, and Blues: Social Movements, Folk Music, and Race in America. Princeton University Press, Princeton, NJ.
Schwalbe, M.L., Mason-Schrock, D., 1996. Identity work as group process. Advances in Group Processes 13, 113–147.
Snow, D.A., Anderson, L., 1987. Identity work among the homeless: the verbal construction and avowal of personal identities. American Journal of Sociology 92, 1336–1371.
Snow, D.A., McAdam, D., 2000. Identity work processes in the context of social movements: clarifying the identity/movement nexus. In: Stryker, S., Owens, T., White, R.W. (Eds.), Self, Identity and Social Movements. University of Minneapolis Press, Minneapolis, MN.
Stoeker, R., 1995. Community, movement, organization: the problem of identity convergence in collective action. Sociological Quarterly 36, 111–130.
Stryker, S., 1980. Symbolic Interactionism: A Structural Version. Benjamin/Cummings, Menlo Park, CA.
Taylor, V., Whittier, N.E., 1992. Collective identity in social movement communities: lesbian feminist mobilization. In: Morris, A.D., Mueller, C.M. (Eds.), Frontiers in Social Movement Theory. Yale University Press, New Haven, CT.
Trevor-Roper, H., 1983. The invention of tradition: the Highland tradition of Scotland. In: Hobsbawm, E., Ranger, T. (Eds.), The Invention of Tradition. Cambridge University Press, Cambridge, London.
Van Stekelenberg, J., 2013. Collective identity. In: Snow, D., della Porta, D., Klandermans, B., McAdam, D. (Eds.), Wiley-Blackwell Encyclopedia of Social and Political Movements. Wiley-Blackwell, Oxford, UK.
Wilkes, R., 2006. The protest actions of indigenous peoples: a Canada-US comparison of social movement emergence. American Behavioral Scientist 50, 510–525.
Ward, 2008. Diversity discourse and multi-identity work in lesbian and gay organizations. In: Reger, J., Myers, D.J., Einwohner, R.L. (Eds.), Identity Work in Social Movements. University of Minnesota Press, Minneapolis, MN.

Collective Memory, Anthropology of

Uli Linke, Rochester Institute of Technology, Rochester, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The rapid emergence of global memory archives presents a novel challenge for anthropological research. Media industries, commodity capitalism, and digital technologies have altered the practices and possibilities of collective remembering. Competing representations of the past are forged, deposited, and exchanged in virtual space. How do anthropologists investigate these modalities of social remembering and forgetting when the machinations of past experience and the formations of historical consciousness are increasingly uncoupled from local memory communities? Through a critical engagement with anthropological concepts, research agendas, and debates about the multiple linkages between remembrance, past experience, and history, this article attempts to uncover the impact of globalization on collective memory practices.

Focused on socially shared, collective memory practices, this article examines the linkages between memory and history with attention to the fields of power in which the struggle for domination over remembrance, retrievable historical consciousness, and collective forgetting takes place. Through a critical engagement with anthropological concepts, research agendas, and debates, the article uncovers the impact of globalization on memory practices. Media industries, commodity capitalism, and digital technologies have altered the possibilities of collective remembering. Uncoupled from memory communities, competing representations of the past are deposited and exchanged in virtual space. These fluid, expansive global memory archives pose a new challenge for anthropological research.

Definition of Concepts: Memory and History

The term *history* is often used to describe representations of the past that appear in written or narrative form, a primary medium through which states, elites, or dominant descent groups confiscate linear time and proclaim official chronologies as master narratives. The writing of history is a 'colonization' of time 'by the discourse of power' (de Certeau, 1988). The term *memory*, by contrast, is conventionally applied to those oral, visual, ritual, and bodily practices through which a community's collective remembrance of the past is produced or sustained (Connerton, 1989). It includes the vast complex of unofficial, noninstitutionalized knowledge not yet sedimented into formal traditions and which represent the 'collective consciousness' of whole groups (Halbwachs, 1980), forming a counterweight to the knowledge that is privatized and monopolized by certain elites for the defense of established interests. While memory may be a moving reservoir of history, furnishing the 'raw material' for representing the past, it is not the same as history (Le Goff, 1992). Individual remembrance, collective memory, and narrative history interact in highly complicated ways, shaping each other as different versions of the past are constructed and reconstructed, modified, and invented.

The Politics of Time: History and Anthropology

The theories, models, and methods developed by anthropologists to study non-European worlds tended toward a 'rejection of historical research of any kind' (Evans-Pritchard, 1962), at least within the limits of the discipline. Having as its central theme 'the unlettered and forgotten peoples,' the anthropological endeavor was inclined to disregard the existence of indigenous histories, which also included, in particular, the trauma of the colonial encounter. But if *making history* is a social practice that produces peoples, bodies, and places, as de Certeau (1988) suggests, then the very erasure of such a universe of time and being must be understood as a figuration of power. The effacement of historicity is a political operation.

Nineteenth-century anthropology was deeply implicated in the projects of Western modernity: the construction of nation-states and the creation of colonial systems with which Europeans came to dominate other worlds. During this process of global expansion and the establishment of modern hegemonic polities, the past was increasingly mobilized as a symbolic discourse 'in the definition of, and the marking of, the boundaries of states and nations' (Cohn, 1981, p. 228). Anthropologists of this period contributed to the colonial undertaking by a reliance on analytic tools that used Western notions of time as an index of unequal evolutionary development. The prerogative to know and control the past, by defining it as 'history' (an order of chronological or linear time), was an entitlement of civilization. The social worlds outside of Europe came to be marked as timeless: a closed symbolic universe inhabited by 'the people without history' (Wolf, 1982). Displaced into the sphere of alterity (Fabian, 1983), colonial subjects were incarcerated by the temporal logic of an expansive European world system.

The past as history became a way to construct schemes of classification that differentiated Europeans from others. Anthropologists began to delineate stages of human development via the comparative method: by identifying phenomena of a general type (taboo, totemism, polygamy), their goal was to map out a global history of social institutions. In this endeavor, however, analytic units were 'taken out of the flux of time' (Evans-Pritchard, 1962, p. 175): social forms, stripped of their unique features and context-bound meanings, escaped temporality, appearing static and unchanging. The comparative method, in its attempt to achieve conceptual stability as a sociological proposition, was explicitly antihistorical in

emphasis and thereby effectively obscured the effects of Western contact and domination.

By the early twentieth century, anthropologists had decisively abandoned the comparative paradigm. But in spite of the shift to new methods of analysis, the discipline's aversion to historical research intensified. Scholars in Britain, Germany, and the United States asserted that the customs of native peoples were to be found not in colonial archives, missionary reports, or dubious travel accounts but *in the field*: reliable data about non-Western societies demanded first-hand observation. This was accomplished at first through brief expeditions and later by long-term residence of the anthropologist in native communities. British scholars, in their interpretation of the ethnographic venture, reasserted the enlightenment program: anthropology was to become "a natural science of society, based on a descriptive analysis of the structure and functioning of social forms" (Cohn, 1981, p. 232). Later theorists deemed such a description of societal surfaces insufficient for a postcolonial anthropology of temporality. French critics like Claude Lévi-Strauss advocated for a humanist-philosophical approach that examined the deeper systemic framework whereby a people lived, signified, and imagined their own historical situatedness in cultural terms (Harkin, 2009). American scholars, by contrast, adopted a romantic idealist stance: in North America, anthropology emerged as a hermeneutic science, focused on the discursive reconstruction and symbolic capture of entire cultural systems. Although the nation-bound projects of anthropology began to diverge, the ethnographic field method was adopted as a common research standard, thereby imparting and codifying the now conventional disciplinary gaze: the observers' eyes were trained to *see* society only at a particular point in time. Ethnography, in its obsession with reifying the authenticity of native peoples, fabricated a closed temporal frame: a *present present* or a *past present*, as Evans-Pritchard (1962) termed it in a critical commentary on the discipline's 'breach with history', illuminating the distortions which flow from the imposition of a priori temporal limits upon indigenous societies.

But it was precisely this aversion to historical research that led anthropologists to discover 'memory' as a valuable ethnographic tool. Evans-Pritchard (1962), for instance, suggested that the past was always "encapsulated in a context of present thought": the past, as *memory*, was embedded in material, ritual, and narrative practices and therefore "part of the social life which the anthropologist can directly observe" (pp. 177–178). The pursuit of collective memory was to constitute a major change in ethnographic practice: the 'life story' of local worlds – of bodies, magic, and markets – could be rendered intelligible through the oral archive. But in the colonizing imagination of Western anthropologists, the work of memory was consigned to the uneventful register of *structural*, *mythical*, and *sacred* time: an order of signs without history. Peoples' means of remembering traumatic events (famine, migration, conquest) "were always seen to reinforce the system in place, never to transform it" (Comaroff and Comaroff, 1992, p. 21). In a shift from historical past to ethnographic present, anthropologists erected 'counterfeit signposts' of a primitive world of enduring traditions – an analytic fiction.

The anthropological fascination with 'unchanging' societies persisted. Signs of cultural duration, repetition, and reversibility were interpreted as evidence of timeless worlds that existed as antithetical counters to the West. Such assertions, following Mbembe (2001), postulated the possibility of temporal realities without historical agency: social life purportedly proceeded 'by the blind force of custom,' moved by 'charms, spells, and prodigies,' and "time [was] supposedly stationary," governed by "the importance of repetition and cycles, and the alleged central place of witchcraft and divination procedures," thereby defining a social landscape inhabited by people who were "caught in a relation of pure immediacy to the world and to themselves," living life as 'libidinal machines' (pp. 4, 15). Such notions endorsed the authority of European world-domination by claiming empirical objectivity. In turn, the colonial gaze attached sociotemporal otherness to a visual frame, 'an aesthetic of respect for the status quo' (Fanon, 1994, p. 3). Anthropological subjects were thus apprehended in terms of a 'complex of visuality,' as Mirzoeff (2011) elaborates, which incarcerated temporal difference in an 'aesthetics of the proper,' perceived as 'pleasing, ultimately even beautiful' (pp. 3–4). While the abrogation of indigenous histories and temporal experiences was transformed into an aestheticized artifact by colonial intervention, creating the figure of the 'ritualistic, tradition-bound native,' critical anthropological inquiry into postcolonial pasts was censured by an affective charge: 'hostility,' 'revulsion,' and, at best, 'inattention'(Evans-Pritchard, 1962, pp. 172–191). The anthropology of collective memory, in consequence, began as a kind of antihistory.

The Shape of Memory: Modes of Historical Consciousness

In their ethnographic explorations, anthropologists encountered indigenous memory forms that did not seem to fit the European idea of historical time: an objective, unmediated, linear account of past events. Divergent temporalities were interpreted as premodern forms of consciousness, modeled on the periodicity of nature: seasonal changes, plant fertilities, female rhythms, and lunar cycles. These assertions have since been contested (Munn, 1992). A reevaluation of the ethnographic evidence suggests that indigenous memory practices do not negate time by simulating nature. Rather, historical experiences are often spatially encoded: memory archives are attached to landscapes to recall place-specific events or the movement of people, bodies, and objects across a topographic terrain (Greenhouse, 1996, pp. 39–48). This mode of time keeping is not nature based. The morphologies of space are meaningfully mapped to record political, economic, or social events. Oral histories and collective memories can thus be archived and retrieved by reference to spatial markers. Ironically, by naturalizing non-European constructions of historical time, anthropologists began to scrutinize the symbolic fabric of society for signs and indexes of such mimetic practices. These investigations led to unforeseen discoveries.

By a focus on myths, ritual, oratory, and body practices, anthropologists uncovered the narrative and performative possibilities of *a history of the present*: the social sense of time whereby people attempted "to grasp the world as both a synchronic and a diachronic totality" (Lévi-Strauss, 1966, p. 263). In the realms of memory, past and present – although

theoretically distinct – were conjoined to link, from generation to generation, the living with the dead. Among Australian aborigines, for instance, *historical* or *commemorative rites* "recreate the sacred and beneficial atmosphere of mythical times—the "dream age"—mirroring its protagonists and their great deeds," and transport "the past into the present"; *rites of mourning* 'correspond to an inverse procedure: instead of charging living men with the personification of remote ancestors, these rites assure the conversion of men who are no longer living into ancestors,' thereby bringing "the present into the past" (Lévi-Strauss, 1966, pp. 236–237).

These images of time are different, perhaps, but not 'other.' In the rites performed, past and present are propelled into a social space – the body. The act of remembering takes corporeal form: memory is embodied (Connerton, 1989). When inhabited or possessed by the past, performative bodies articulate "the means by which a narrated ancestral event is given 'presence'" (Munn, 1992, p. 113). But this transposition of temporal worlds, which connects the living and dead to the ancestors via the mimetic power of embodiment, is not an abrogation of time. It is a generative reconstitution of society: a practice of empowerment, a political engagement with the matrix of life (Stoller, 1995). This attentiveness to indigenous body-time paradigms galvanized memory research in Europe, thereby closing the postulated rift between 'Western' and 'non-Western' approaches to temporal meaning (see Badone, 1989; Herzfeld, 1991; Kertzer, 1989; Kligman, 1990; Verdery, 1999). In all societies, representations of the past are forged by the synergies of symbolic processes and the contingencies of life.

Anthropological studies further suggest that collective memory may be systematized on the basis of *relative temporalities*, that is, a flexible system of polymorphous reference points. Among certain peoples of the Ivory Coast, like the Guéré, "the consciousness of a historical past has developed alongside a multiplicity of other times" (Le Goff, 1992, pp. 8–9): *mythical*, *genealogical*, *historical*, *lived*, and *projected time*. Such a register of memory forms belongs to the aftermath of the colonial project. This observation is congruent with the historical reverberations of struggle 'in the postcolony,' as Mbembe (2001) asserts, where the 'languages of life' always consist of 'a combination of several temporalities' – brought into being by *memories of living* in the concrete world and the entanglement of *existence and experience* in times of 'instability and crisis' (pp. 15–17). Under the impact of prolonged social trauma, the disparate logics of time and memory chart the precarious matrices of life: a synchrony of struggle, violence, and perseverance. In the 'afterlife of collectively experienced catastrophes,' these knowledge of the past can be deployed as a 'political language' to affect future change (Radstone and Schwarz, 2010, p. 3). By the twenty-first century, anthropologists shifted their investigative focus, turning attention from ethnographic inventories of polymorphous memory archives to the politicization of historical consciousness.

Memory sites: The Localization of Remembrance

The anthropological analysis of collective memory moves toward a renunciation of linear temporality in favor of multiple kinds of time as experienced at the levels where the individual takes root in the social: language, demography, kinship, economics, and politics. Memory production is in this sense not dependent on a single individual's experience of the past. A remembrance, a social understanding of events that is represented as memory, can be constructed by sharing with others "sets of images that have been passed down to them through the *media of memory*—through paintings, architecture, monuments, ritual, storytelling, poetry, music, photos, and film" (Watson, 1994, p. 8). In every society, we can identify an array of memory sites or places of commemorative record and practice where remembrance anchors the past: *topographical* places (archives, libraries, museums); *monumental* places (cemeteries, architectural edifices); *symbolic* places (commemorative rites, pilgrimages, emblems); *functional* places (manuals, autobiographies, associations); and places of *power* (states, elites, milieux) which "constitute their historical archives in relation to the different uses they make of memory" (Le Goff, 1992, pp. 95–96, Nora, 1986). These memory sites furnish a series of locations where knowledge of the past are conveyed and sustained by a circulation of signs that focuses meaning on pertinent ideas, values, and feelings.

In this era of globality, the relation of collective memory to locality and place has been reframed. Conventional sites of remembrance are replaced by virtual images. Digitalized knowledge of the past travel across national borders and communal solidarities. New information technologies provide historical imaginaries with alternate vectors of space–time orientation. In her work on the contested state in Cyprus, Navaro-Yashin (2002) tracks the political implications of such altered memory capacities. Her research shows not only how local histories are complicated by military violence, law enforcement practices, and international legal discourse, but also the limitations of subjecthood imposed by regimes of 'documented lives.' State-issued documents such as passports, birth certificates, visas, identity cards, and work permits are sites of political memory and indexes of state power, which capture and reveal life histories. While these portable docu-memories of the state may be authenticated in police archives or judicial record rooms, enhanced identity profiles are stored in electronic space – together with biometric scans, fingerprints, and criminal histories. Digital technologies can transmit a state's 'memory of a people' across the globe, permitting instant and total recall of archived data – from citizenship portfolios to security profiles (Matsuda, 1996, p. 121). Located in the virtual realms of electronic space, the political memory banks of belonging and exclusion have become transterritorial and transglobal.

In a globalizing world, the anthropology of collective memory acquires a new vantage point. Entrenched inequalities are evident across continents. Local experiences are identifiable as worldwide patterns. Postcolonial and posttotalitarian strands of historical consciousness can be globally mapped, permitting an analysis of cognate memory traces from different existential life-worlds of the *Global South*: an imagined territory and a relational space that has been configured "by its structural and tropic situation in the history of the ongoing global present" (Comaroff and Comaroff, 2012). From this perspective, collective memory practices must be understood

as situated political formations. Across the spaces of the global south, postindependence representations of colonial pasts may become oppositional, liberatory, or revolutionary. In each case, the multiple registers of temporality and historical experience serve as a weapon of choice against the mythologies of premodernity (*Orient, Third World, Primitive Other*) that were so violently propagated by the empires of the world's north (Comaroff and Comaroff, 2011). A global vantage point also sheds new light on the creation of 'affective communities' across political spaces, as Casey (2009) shows in her analysis of citizenship, Hausa Muslims, and violence in northern Nigeria: media flows shape translocal solidarities by infusing alternate registers of history and truth with composite memories of identity, affect, and belonging.

Unsanctioned Archives: Memory and Countermemory

The anthropology of collective remembrance is not only concerned with affirmation and heritage, but with processes of contestation, the aftermath of violence, and the emergence of countermemories (Stoller, 1995). *A social history of remembering* must take account of 'the mechanisms that make unsanctioned remembrance possible' under repressive political regimes: in asking how people remember what is meant to be forgotten, 'matters of transmission' and the procedures for creating shared memories and hidden histories take center stage (Watson, 1994). Here, interpersonal memories, commemorations, theater, drama, folklore, secret, and oppositional histories are 'the venues within which alternative remembrances' – unauthorized and unapproved memories of the past – can be located and analyzed (Watson, 1994, p. 2). Collective memory practices are linked not only to sites of domination that attempt to legitimate a given social order, but also to unsanctioned sites of struggle, opposition, and transformation.

In attempts to decipher the underground historical consciousness of ordinary people, anthropologists began to investigate the unarticulated truths or 'hidden transcripts' (Scott, 1990) that emerged communities stripped of agency and power by a centralized state apparatus. Initial research on these hidden collective memory forms or 'public secrets' (Taussig, 1999) in Latin America, the Soviet Union, and Eastern Europe triggered further interest in 'perilous' or dangerous memories that had been forged in contexts of war – the silenced remembrances of suffering that posed a challenge to official histories: examples include local memories staged against the allied liberation narratives propagated in the Asia-Pacific Wars (Fujitani et al., 2001) or memories of 'evasive spaces' in the domestic interior of Dutch colonialism in Southeast Asia (Stoller and Strassler, 2000). With attention to the possibilities of 'effacing' dominant screen memories (Taussig, 1999), anthropologists began to examine the stratigraphies of remembered truths in the aftermath of genocide and militarism (Feldman, 1997; Hinton, 2005; Linke, 1999; Robben, 2005). Such studies point to the complexities of societal recovery after the trauma of mass violence or colonial terror. "At issue is the political life of imperial debris, the longevity of structures of dominance, and the uneven pace with which people can extricate themselves from the colonial [or political] order of things" (Stoler, 2008, p. 193). Countermemory studies analyze the impact of unsanctioned archives on the emergent sense of transnational justice in posttrauma contexts.

The Traffic in Memory: Global Hegemonies and Local Struggle

In all societies, there exist distinct moments when alternative representations of the past are forged, contested, and used for new cultural or ideological projects. Battles waged over ethnic origin or national identity may be linked to the constitution of radically new kinds of memory archives (Hobsbawm and Ranger, 1983). The creation of nations, ethnicities, and subjects inevitably alters the shape of the past. "Memory" is therefore "not a generic term of analysis, but itself an object appropriated, transformed, and politicized" (Matsuda, 1996, p. 6). Or, put differently, memory can be nationalized, aestheticized, and commodified, offered up for purchase or branded for consumer attention.

Historical consciousness shaped by nation-building, scientific discourse, and global capitalism plays a significant role in the formation of social identities. Detailing the politics of remembrance in postsocialist China, Schein (2000) shows how state-orchestrated reforms, staged against the chaos of the Cultural Revolution and implemented during a period of increasing exchange and fascination with the west, produced a popular appetite for exotic others: commodified images of minority peoples, like the Miao, were fashioned by an acute nostalgia for 'ancient' traditions and those feminized cultural essences on which to recraft a distinctive Chinese nationness. The result was a boom in memory recuperation among China's urbanites and Miao elites, who simultaneously played to a Western valorization and high-spending tourist consumption of dwindling pasts. But such a newly forged cultural attachment to memory artifacts, the 'fossil bed' phantasms of folk heritage, also entails transformation. In other words, the work of symbolic restoration 'may conceal very new projects' (Le Goff, 1992, p. 9). The recruitment of the past for revolutionary or political ends, as in the case of China, is entangled in a twofold venture: the forging of a distinctive ethnonational imaginary and the renegotiation or recalibration of *place* in the global order.

The pursuit, rescue, and celebration of collective memory is always socially motivated and has, thus, to be understood in positional terms. The global context – the theater of transnational consumer capitalism – is, according to Friedman (1994), particularly relevant in understanding how local memory forms are affected by the formative impact of world systemic processes. In the case of modern Greece, as Friedman shows, national selfhood was crafted by internalizing the ways in which European elites, in constructing their own 'civilized' origins, mythologized classical antiquity as the cradle of Western civilization. Greek nationalists found their past in the institutional memory of the West, thereby accelerating incorporation into an expanding European world system. Such acts of cosmological repositioning are symptomatic of a transglobal system that fuses socioeconomic transformation and collective

memory to "the reconfiguration of the map of the world's peoples" (Friedman, 1994, p. 123).

Yet even in the global arena, there exist other possibilities of memory production. In the Hawaiian case, Friedman's instructive contrast, indigenous communities attempted to extricate themselves from Western dominance by projecting a value system, produced in the modern context, onto an aboriginal past. The result was a body of memories composed of mythic history, a work of folklore and folklorization 'rife with simulacra of tropical fantasies.' Social and economic marginalization, in the aftermath of World War II, gave rise to an intense period of Hawaiian cultural revival. Recuperation efforts, centered on the reconstitution of ethnic history by retrieving precolonial social forms, brought into play a *museology of memory* that was staged in opposition to the West: a precontact Hawaii stocked with all the items to be found in Western historical archives – libraries, museums, and missionary records. The Hawaiian confiscation of these colonial memory stores proceeded, as Friedman documents, by forging an indigenous past uncoupled from the larger world, a Western-imposed modernity, which had obliterated a population and absorbed its social history into the projects of global economic systems.

Anthropologists have firmly established the need to attend to these paradoxes of mobile cultural and symbolic forms – this *traffic in memory* – during periods of transnational crisis and restructuring (Appadurai, 1996). In the context of global systems, where the practice of identity is brought on stage, the logic of time and the meaning of memory have been radically transposed.

Global Memory Markets: Commodifying History, Commercializing Violence

In the global arena, the production of historical consciousness becomes increasingly entangled with the commodity form: encoded by temporal longings and mythic value, it furnishes the common stock for a *memory market* saturated by cargo-type products of surplus and lack, bearing the deep imprint of the globally hegemonic mode in which people's ethnonational remembrances are manufactured, exchanged, and consumed (Taussig, 1980). Inserted into the media circuits of commodity capitalism, historical consciousness is rendered portable. "The global reach of Hollywood allows certain images of the past to be exported around the world" (Morris-Suzuki, 2005, p. 243). But such memory forms are not transported innocently. Representations of the past have become commodity objects, made for profit. In the contemporary spaces of cultural production, photography, comic books, performative theater, the museum, entertainment industries, and cyberspace serve as conduits for nourishing the consumptive appetites of global publics. Select trauma histories circulate as universal tropes for other human atrocities. The Holocaust is an exemplary case. By exposing the linkages between modernity, racialization, and violence, as Hyussen (2003) argues, Europe's genocide has acquired a totalizing dimension that is continuously appropriated and meaningfully attached to other "local situations that are historically distant and politically distinct from the original event" (pp. 13–14). Such emergent memory markets do not necessarily inhibit critical inquiry but may pose "fundamental questions about human rights violations, justice, and collective responsibility" (p. 15). Nevertheless, the mass-mediated traffic of *present pasts* encourages a habitus of selective remembering, rendering some aspects of traumatic events hypervisible while muting the recall of others.

In a consumption-driven capitalist world, memory industries obliterate temporal boundaries and destabilize the precarious relation between remembering and forgetting. Mass-mediated productions of historical events flourish on narrative plots scripted by corporate interest. Representations of the past are fabricated, staged for popular entertainment. Such a collage of fiction and memory, unlike remembrances produced by collective experience or lived time, is not 'embodied in the social' and therefore more easily forgotten: commodified memory "is sucked into the timeless present of the all-pervasive virtual space of consumer culture" (Hyussen, 2003, pp. 10, 28). These transient, fleeting 'memories of the modern' (Matsuda, 1996) tend to destabilize a secure sense of the past. Guided by market principles, mediated historical consciousness is a product of 'prosthesis' memories (Nora, 1989, p. 14): a repertoire of simulated temporalities, composed of recycled image fragments and fictional texts that conjure an affective engagement with the past via shock-effect or mass-produced sentiments of nostalgia and commemoration. As 'prosthetic' implants, such consumable memory artifacts have the "capacity to create shared social frameworks for people who inhabit different social spaces, practices, and beliefs" (Landsberg, 2004, p. 8). In a global media market, translocal identities and transnational solidarities are forged by consumer-directed memory productions that mimic a sensuous engagement with imagined historical realities (Neiger et al., 2011). Stocked with prosthetic memories and consumable pasts, global commodity culture is increasingly marked by fears of forgetting the phantasmatic and surreal.

What sorts of memory markets have emerged to counter these popular anxieties about the slippage of temporal experience? Through the interventions of commercial media, trauma history has come to the attention of a bourgeoning tourist industry. Commercial sightseeing tours and venture capitalism flourish in areas of former violence, suffering, and mass destruction: examples include the battlefields of Europe, sites of genocide at Auschwitz-Birkenau, the Killing Fields of Cambodia, and the ruins of New York's Ground Zero. These are the negative spaces of history: the nonvirtual, archaeological sites of a bygone world or past event defined as inhumane, catastrophic. This memory genre belongs to a broader pattern of commodity consumption variously described as 'dark tourism,' 'disaster tourism,' and 'trauma tourism' (Lennon and Foley, 2000; Stone, 2006; Stone and Sharpley, 2008; Gotham, 2007). Such practices are focused on sites of human tragedy that can be branded, showcased, and meaningfully reimagined by consumer travelers. Organized venture tours not only promise to deliver memorable experiences: by touring sites of past violence, memory is socially constructed, from 'lived' life (however episodic), which is authenticated by the event's sightseeing experience shared by tourist participants. Through the ready-made vignettes of space and time, as Sturken (2007) has argued, trauma tourism markets

'comfort culture': it propagates fear to sell safety; it offers prepackaged emotions to replace traumatic memory formations with kitsch sentiments for purchase. Collective memory practices are thereby arrested in time, permitting a constant replay or reenactment of history with the promise of return to a secure present that is "somehow distanced and unimplicated in the global strife of the world" (Sturken, 2007, pp. 3–4). Violence remembered remains in the past, as an artifact of a distant time, a different temporality. Trauma tourists participate in a shared experience, consuming prosthetic memories as ready-made versions of history. The social construction of collective memory is thereby simulated by time-travel into the contact zones of the past: touristic encounters with history are mediated by the hegemonic representations of commodity realism.

In the global marketplace, memory consumption can also be a subordinate production, one that creatively and tactically resists the homogenizing tendency of centralized regimes. With a focus on the political economy of memory culture in Latin America, Bilbija and Payne (2011) scrutinize the posttrauma labor of remembrance in Argentina, Brazil, Chile, Mexico, Peru, and Uruguay, in countries where traces of repressive political regimes continue to reverberate in everyday life. The commercial use of memory artifacts (e.g., testimonials, places of terror, prisons, paraphernalia) transports accounts of past atrocities and state terror into the public domain. Such publicity can inspire alternative engagements with existing memory stores or turn attention to human rights agendas. But not all commercial memory projects advocate for justice. The commercial recall of past violence promotes consumer culture. This is suggested by the conversion of a former prison complex into a shopping center, the staging of state terror as a backdrop for the love-triangle plots in Brazilian (Globo) television novellas, or the use of atrocity images in corporate ads for jeans in Argentina or Chile that 'render torture sexy and trendy' (Bilbija and Payne, 2011, p. 24). Although past violence is commodified, capitalist memory markets acquire political momentum when cross-coded with transglobal ideas of justice or human right agendas captured from beyond the borders of the nation-state. Inflected with political meaning, and reinterpreted as an ontological discourse of visibility, the presence of historically charged trauma images in the public sphere 'defaces' official screen memories, thereby lifting the veil of secrecy and silence placed upon an unutterable, muted past (Taussig, 1999). By contesting the state's authority of censorship, collective memory sharing in the marketplace is thereby equated with 'the right to look' (Mirzoeff, 2011), even though this demand for freedom and truth is mediated by the logic of commercial culture. Commodity practices not only render muted histories perceptible. As past violence is made consumable, it becomes memorable. What is collectively remembered about past terror may be dictated by capitalist intervention. Nevertheless, the 'politics of interpretive license' (Stoler and Strassler, 2000, p. 5) introduce unconventional tactics of meaning making and signification, and as such remembering. This suggests, as Schein (2000) observed, an alternate sense for collective memory production: "the myriad ways in which people make use of received cultural products presents a measure of possibility, not only for autonomy, but also for subversion of the dominant order" (p. 16). Put differently, the modalities of collective memory practice attest to the possibilities of *symbolic creativity* – even in the province of globalizing modernities or consumer capitalism.

Memory of the World: Globalizing Collective Remembrance

How are collective memories constituted in a global world? Under globalization, the 'spaces of memory and the composition of memory communities' have been reconfigured (Assmann and Conrad, 2010, p. 1): not only are memories themselves rendered portable but events may be simultaneously witnessed in real time by differently constituted publics. While representations of the past continue to be framed by national agendas, transglobal memory archives are stocked with data snippets pooled from socially unrelated subjects worldwide. Memory is no longer a stable repository of the residual traces of shared life experience. The planetary proliferation of electronic media has altered the possibilities of memory recall and input across the globe: moments in time can be recorded, stored, then 'brought back to life, restaged, or replayed' (Winter, 2012, p. 4). Individualized access to the Internet has changed practices of remembering. Digital forms of communication (email, Facebook, Twitter, YouTube, Skype, blogging) offer new venues for memory sharing beyond conventional boundaries of nation, culture, and community. Transported through cyberspace, bits of 'digital network memory' (Hoskins, 2009) reside in virtual timescapes, where temporal records can be speeded up or brought to a standstill, creating parallel universes of memory, time, and space.

In the "wake of the information explosion and the marketing of memory" (Hyussen, 2003, p. 18), the societal dialogue with lived historical experience has been disrupted. Global memory stores are expansive, gaining momentum through new forms of representation, memorializing, knowledge recycling, instant recall and replay of temporal experiences, data simulation, copying, and replication. Although "memory has become a cultural obsession of monumental proportions across the globe" (Hyussen, 2003, p. 16), it is driven by consumer culture and corporate desire rather than the existential machinations of ethnosocial communities. Global media industries favor 'random access memories' and episodic attention to news worthy events (Morris-Suzuki, 2005). Competing versions of past experience rise to the surface of data-saturated awareness by the media hype of spectacle, advertisement, and entertainment. The formation of global memory markets productively engages postmodern instabilities: a world of shifting imaginaries, inhabited by flows, fragmentation, and forgetting. The eventual shrinkage of collective memory in the global theater poses a challenge for conventional anthropological research. What happens to the discipline's humanistic endeavor in light of this emergent 'future history of the world' (Lévi-Strauss, 1952, p. 49)? Twenty-first century anthropology needs to grapple with the implications of fluid temporalities and shifting memory imprints in a virtual world of global culture politics.

See also: Anthropology and History; Belief, Anthropology of; Colonialism, Anthropology of; Cultural Heritage; Cultural Relativism, Anthropology of; Culture: Contemporary Views; Ethnography; Ethnohistory; Fieldwork in Social and Cultural Anthropology; Globalization, Anthropology of; Human Rights, Anthropology of; Identity in Anthropology; Internet and Social Media: Anthropological Aspects; Knowledge, Anthropology of; Postcoloniality; Ritual; Time, Anthropology of.

Bibliography

Appadurai, A., 1996. Modernity at Large. University of Minnesota Press, Minneapolis.
Assmann, A., Conrad, S. (Eds.), 2010. Memory in a Global Age. Palgrave Macmillan, New York.
Badone, E., 1989. The Appointed Hour. University of California Press, Berkeley.
Bilbija, K., Payne, L.A. (Eds.), 2011. Accounting for Violence. Duke University Press, Durham.
Casey, C., 2009. Mediated hostility. In: Hinton, A.L., O'Neill, K.L. (Eds.), Genocide. Duke University Press, Durham, pp. 247–278.
Cohn, B.S., 1981. Toward a rapprochement. Journal of Interdisciplinary History 12 (2), 227–252.
Comaroff, J., Comaroff, J.L., 2011. Theory from the South. Paradigm Publishers, Boulder.
Comaroff, J., Comaroff, J.L., 24 February 2012. Theorizing the contemporary—theory from the South. Cultural Anthropology. Forum. http://www.culanth.org/?q=node/500.
Comaroff, J.L., Comaroff, J., 1992. Ethnography and the Historical Imagination. Westview Press, Boulder.
Connerton, P., 1989. How Societies Remember. Cambridge University Press, Cambridge.
de Certeau, M., 1988. The Writing of History. Columbia University Press, New York.
Evans-Pritchard, E.E., 1962. Social Anthropology and Other Essays. The Free Press of Glencoe, New York.
Fabian, J., 1983. Time and the Other. Columbia University Press, New York.
Fanon, F., 1994. The Wretched of the Earth (R. Philox, Trans.). Grove, New York.
Feldman, A., 1997. Formations of Violence, Second ed. University of Chicago Press, Chicago.
Friedman, J., 1994. Cultural Identity and Global Process. Sage, London.
Fujitani, T.T., White, G.M., Yoneyama, L., 2001. Perilous Memories. Duke University Press, Durham.
Gotham, K.F., 2007. (Re)branding the big easy. Urban Affairs Review 42 (6), 823–850.
Greenhouse, C., 1996. A Moment's Notice. Cornell University Press, Ithaca.
Halbwachs, M., 1980. The Collective Memory. Harper & Row, New York.
Harkin, M.E., 2009. Lévi-Strauss and history. In: Wiseman, B. (Ed.), The Cambridge Companion to Lévi-Strauss. Cambridge University Press, Cambridge, pp. 39–58.
Herzfeld, M., 1991. A Place in History. Princeton University Press, Princeton.
Hinton, A.L., 2005. Why Did They Kill? University of California Press, Berkeley.
Hobsbawm, E., Ranger, T. (Eds.), 1983. The Invention of Tradition. Cambridge University Press, Cambridge.
Hoskins, A., 2009. Digital network memory. In: Erli, A., Rigney, A. (Eds.), Mediation, Re-mediation, and the Dynamics of Cultural Memory. Walter de Gryuter, Berlin, pp. 91–106.
Hyussen, A., 2003. Present Pasts. Stanford University Press, Stanford.
Kertzer, D.I., 1989. Ritual, Politics, and Power. Yale University Press, New Haven.
Kligman, G., 1990. The Wedding of the Dead. University of California Press, Berkeley.
Landsberg, A., 2004. Prosthetic Memory. Columbia University Press, New York.
Le Goff, J., 1992. History and Memory. Columbia University Press, New York.
Lévi-Strauss, C., 1952. Race and History. UNESCO, Paris.
Lévi-Strauss, C., 1966. The Savage Mind. University of Chicago Press, Chicago.
Lennon, J., Foley, M., 2000. Dark Tourism. Continuum, London.
Linke, U., 1999. German Bodies. Routledge, New York.
Matsuda, M.K., 1996. The Memory of the Modern. Oxford University Press, New York.
Mbembe, A., 2001. On the Post-colony. University of California Press, Berkeley.
Mirzoeff, N., 2011. The Right to Look. Duke University Press, Durham.
Morris-Suzuki, T., 2005. The Past Within Us. Verso, London.
Munn, N., 1992. The cultural anthropology of time. Annual Review of Anthropology 21, 93–123.
Navaro-Yashin, Y., 2002. Faces of the State. Princeton University Press, Princeton.
Neiger, M., Meyers, O., Sandberg, E. (Eds.), 2011. On Media Memory. Palgrave Macmillan, London.
Nora, P., 1986. Les lieux de mémoire, vols. 3. Gallimard, Paris.
Nora, P., 1989. Between memory and history. Representations 26, 7–25.
Radstone, S., Schwarz, B. (Eds.), 2010. Memory. Fordham University Press, New York.
Robben, A., 2005. Political Violence and Trauma in Argentina. University of Pennsylvania Press, Philadelphia.
Schein, L., 2000. Minority Rules. Duke University Press, Durham.
Scott, J.C., 1990. Domination and the Arts of Resistance. Yale University Press, New Haven.
Stoler, A.L., Strassler, K., 2000. Castings for the colonial. Comparative Studies in Society and History 42 (1), 4–48.
Stoler, A.L., 2008. Imperial debris. Cultural Anthropology 23 (2), 191–219.
Stoller, P., 1995. Embodying Colonial Memories. Routledge, New York.
Stone, P.R., 2006. A dark tourism spectrum. Tourism 54 (2), 145–160.
Stone, P.R., Sharpley, R., 2008. Consuming dark tourism. Annals of Tourism Research 35 (2), 574–595.
Sturken, M., 2007. Tourists of History. Duke University Press, Durham.
Taussig, M.T., 1980. The Devil and Commodity Fetishism in South America. University of North Carolina Press, Chapel Hill.
Taussig, M.T., 1999. Defacement. Stanford University Press, Stanford.
Verdery, K., 1999. The Political Lives of Dead Bodies. Columbia University Press, New York.
Watson, R.S. (Ed.), 1994. Memory, History, and Opposition under State Socialism. School of American Research Press, Santa Fe.
Winter, A., 2012. Memory. University of Chicago Press, Chicago.
Wolf, E.R., 1982. Europe and the People Without History. University of California Press, Berkeley.

Collective Memory, Psychology of

Alin Coman, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Prager, volume 4, pp. 2223–2227, © 2001, Elsevier Ltd.

Abstract

Collective memory has been the subject of productive interactions within the social sciences ever since its initial conceptualization. Employing a social interactionist framework in which large-scale phenomena are empirically investigated as interactions between cognition and social dynamics, the article presents psychology's contribution to understanding the formation and the dynamics of collective memories.

Collective memory has had a long history within the social sciences. It served, in the period since Halbwachs's landmark conceptualization (1980/1925), as a meeting place for interdisciplinary conversations involving sociology, history, anthropology, and psychology (Hirst and Manier, 2008). As any constructive dialog, conceptualizing collective memory has been ridden with theoretical disputes ever since the emergence of the concept. Methodological individualism, an approach aimed at deterministically deriving collective phenomena from the psychological make up of individuals and their interactions, constituted the main issue of contention (Udéhn, 2002). Although, it is beyond the scope of this article to provide a review of methodological individualism and its nuances, the burgeoning literature on the psychology of collective memory provides substantial evidence that bridging between micro and macroprocesses is a worthy endeavor (DiMaggio, 1997). In what follows, I will briefly present approaches within the social sciences that allow for a psychological exploration of collective memory, I will then articulate the psychological assumptions of such an endeavor, and finally, I will review the burgeoning psychology literature aimed at bridging between micro and macroprocesses.

Collective Memory: A Case for Its Psychological Foundations

It is widely accepted that just as individuals have memories of their own personal pasts, so do collectivities have memories of the events that are central to their identity. What is contentious, though, is whether to understand these collective memories, one needs to take into account how individuals remember, or simply restrict oneself to the study of the social processes involved in their formation. Widely credited with the articulation of the concept, Maurice Halbwachs was mindful of the fact that individual memory is critical to understanding collective memory. "It is individuals as group members who remember," he claims (Halbwachs, 1980: 46). The critical notion that individuals remember as group members stands at the crux of his conceptualization of memory as inherently social: from the very basic usage of a particular group's language to reminisce to the social dimension that accompanies the act of communicating a memory. Given the belief that it is impossible to separate what is individual and what is social about memory, he cautions against looking into the brain or in some 'nook of (the) mind' to understand memory, for memories "are recalled by me externally, and the groups of which I am part at any given time give me the means to reconstruct them" (Halbwachs, 1980: 46). It is in light of this unbreakable link between the individual and its social embeddedness that Halbwachs proposes the study of the social frameworks in which memory is manifested.

This ambiguity of locating collective memory in either individual minds or in the social processes that govern their creation has received extensive attention among social scientists. In contrast with Halbwachs's acceptance of the ambiguity, most current approaches seek to resolve it by clearly differentiating between endeavors that commit to one location or the other. Young (1993), for instance, differentiates between 'collected memories,' defined as the aggregation of individually produced social memories, and 'collective memory,' seen as collectivistic cultural activities or productions (see Olick, 1999; for a similar distinction). In final analysis, he argues, "societies cannot remember in any other way than through their constituents' memories" (1993: xi). In a similar attempt to provide clarity to the concept of collective memory, Jan Assmann distinguishes between communicative and cultural memories. The former are conceptualized as conversationally constructed memories that have limited time horizons, the latter are seen as long lasting, publicly available cultural products (e.g., monuments, texts, images, rituals) designed to emphasize meaningful events in the history of the group. It is only in those instances in which communicative memories get transformed into 'objectivized culture' that they become cultural memories (Assmann, 1995: 130). Individuals and their interactions are, according to Assmann, critical only when considering transient communicative memories. The focus of collective memory scholars should be, he claims, the stable cultural memories and the struggles behind their enactment.

What emerges in the social sciences is, then, a dichotomy as to whether understanding collective memories requires the integration of memory processes that are specific to individual memory (Olick, 1999). On the one hand, there are efforts to carve out the 'spaces' where individuals are critical units when discussing collective memory, in concepts such as collected memories and communicative memories. On the other hand, approaches that follow an antiindividualistic

stance (Assman, 1995; Connerton, 1989) advocate that the study of collective memory be completely divorced from its anchoring within individual minds. To resolve this antagonistic dichotomy, Olick suggests a framework, under the heading of 'social memory studies,' that would allow disciplines to carry out their independent research programs in a way that respects their theoretical and methodological differences. Collected memory scholars would study the aggregation of memories, whereas collective memory scholars would carry out investigations of institutionalized objects.

A social interactionist approach stands in contrast to these dichotomous perspectives. It accepts that one cannot easily separate, as Halbwachs contended, the individual from the social. Social processes are constituted by individuals' interactions, and so they are dependent on individual characteristics, whereas individual behavior is dramatically affected by the social contexts in which it is manifested. Integrating this interdependence, a social interactionist position insists that one cannot understand how social artifacts and practices affect memory without considering individual cognition (Hirst and Echterhoff, 2012). This approach builds on recent philosophical developments in the theory of mind, specifically, work on extended mind and distributed cognition (Clark and Chalmers, 1998; Michaelian and Sutton, 2013). According to this view, the mind is part of a larger system that incorporates the environment, including the social contexts in which individuals function. To understand the navigation of a ship, for instance, one has to understand the complex system that produces the act of navigation. A system that includes the interdependence of individuals who are responsible for the navigation, the communication system that coordinates the interdependent individuals, and the equipment that allows for the integration and calibration of the different components (Hutchins, 1995).

In this context, psychology has set out to investigate how a perspective that takes into account both individual cognition and social dynamics could contribute to understanding the formation and the dynamics of collective memories. In what follows, I will trace the history of this social interactionist approach, explore how it leads to an empirically based investigation of collective memory, and finally present the recent empirical advances in the field.

Social Dimensions of Memory

Ever since psychology embraced the experimental method to study the mind, the individual has been, for the most part, the exclusive focus of its investigations. Given this focus, a topic such as the formation and dynamics of collective memories seems to be orthogonal to both theoretical and methodological advances within psychology, particularly with recent developments in neuroscience (Mather et al., 2013). Despite the prevailing individual reductionism, the social nature of memory has been a topic of rigorous psychological research from the early days of experimental psychology. In an extreme form of individual reductionism, one of the first experimental psychologists to explore memory, Herman Ebbinghaus (1913), carefully designed his stimulus material to strip away its contextual and social dimensions. He used nonsense syllables and took precautions to make these syllables as distinct from natural language as possible. The reasoning was that any associations that these syllables might trigger would interfere with one's ability to understand how memory works.

Stripping these dimensions away, Sir Frederic Barlett (1932) claimed, left nothing to be studied. Taking a perspective consistent with Halbwachs's notion that memory is inherently social, Bartlett embarked on programmatically studying the social dimensions of memory. For Bartlett, memory was akin to a tennis stroke. Just as when serving a ball, different contextual factors, such as wind speed, position of the sun, and the location of the opponent, will affect one's serve, so is remembering influenced by the context in which it is manifested. No two strokes, even by the same individual, will be exactly the same because of the variations in the different contextual factors. It follows that no two acts of remembering, even of the same memory by the same individual, will be exactly the same. In this conception, memory is less of a reproductive replica of the past, and more of a reconstructive process affected by a host of contextual factors. Consistent with this realization, Bartlett introduces the concept of *schemata*, which he defines as organized structures of knowledge that act as scaffolds for remembering. It is through these schemata, he argues, that the profoundly social nature of memory becomes manifest. They are acquired and fine-tuned within the groups that one belongs to, and they constitute the generative engines of remembering (Bartlett, 1932). The cultural schemata we develop during socialization processes, Bartlett argues, not only guide what we pay attention to, but also circumscribe how and what we remember (Wertsch, 2002).

Collective Memory from a Psychological Perspective

Building on the idea that memories are socially constructed (Bartlett, 1932), and emboldened by a wider acceptance of an extended mind philosophical tradition (Clark and Chalmers, 1998), psychologists have recently made collective memory a matter of not just sociological, but psychological concern (Hirst and Echterhoff, 2012). The framework that allows for these advances is one in which collective memories are defined as shared individual memories that bear on people's identities (Coman et al., 2009; Hirst and Echterhoff, 2012). By conceptualizing collective memories as shared individual memories, psychologists have committed themselves to answering several questions. First, how do memories become shared, and what are the psychological mechanisms that account for their sharedness? This set of questions falls under investigations of the process of the formation of collective memories. A second set of questions has to do with the form that these shared representations take. The focus here is not on the processes that underlie remembering, but on the outcome of these processes at a collective level. Psychology has made significant progress in answering these questions, but before we delve into it, I will briefly articulate the psychological assumptions that allow for a programmatic investigation of the psychology of collective memory.

Psychological Assumptions

I have thus far argued that individual cognition and social dynamics interactively shape collective memories. In this section, I will discuss three claims that have received extensive support within psychology and are fundamental to understanding the emergence of shared representations of the past:

1. Memory is reconstructive. As evidenced in the tennis strike metaphor (Bartlett, 1932), remembering is seen as a flexible reconstruction of the past, in which the audience, future goals, and the context in which it happens, are all critical factors that determine what form memory would have (Schacter, 1999).
2. Individuals conform. When placed in a social context, individual behavior is drastically affected by the constraints of such a context. Individuals' memories are, for example, affected by suggestions or recalls offered by others in a social group (Hirst and Echterhoff, 2012). And the reconstructive nature of memory is what allows for these suggestions to become completely internalized and remembered as valid representations of the past, even though individuals never experienced them (Edelson et al., 2011).
3. Individuals identify with groups. Extensive research within social psychology has explored the processes involved in group identification, but also the implications that such identification entails. Favoring the in-group and derogating the out-group, preferential attention to group-relevant information, and the motivation to construct a common ground with fellow group members represent biases that favor the emergence of collective memories (Ellemers et al., 2002).

It is the combination among memory's flexibility, and the individuals' tendencies to conform to and identify with groups that give force to an approach aimed at understanding the formation of collective memories. What results from the integration of these psychologically grounded claims is a social interactionist approach that takes into account both the individual cognitive mechanisms, such as memory's flexibility, but also the social dynamics surrounding people's interactions, such as the specific ways in which they influence one other. In focusing on memories growing out of human interactions, the social interactionist approach sees remembering as embedded in social activities. Conversations, as such central social activities, provide a fertile ground for psychological investigations of collective memories (Hirst and Echterhoff, 2012).

Collective Memory as a Process: A Social Interactionist Approach

Shared representations are created through complex processes. While recognizing the intricate linkage between the cognitive and social dimensions governing their formation, researchers have recently taken on the challenge of studying these processes empirically. In this section, I will first describe work on shared attention that allows for the emergence of shared representations in the absence of communication, I will then turn to the larger literature on the impact of conversations on memory in dyads and small groups, and I will end the section with recent work aimed at understanding how shared representations emerge in social networks.

Noncommunicative Processes

Most of the process accounts of collective memory emphasize the conversational nature of their emergence. Bartlett's notion of schema is, however, consistent with the idea that shared representations of a particular event could emerge even in the absence of social interactions. By using similar schemas to filter incoming information, individuals can form similar representations of the past. These schemata are, of course, acquired by way of interactions among individuals, but once they are part of a person's repertoire, they will automatically guide the encoding and retrieval of information. Group membership and situation typicality increase the chances that independent individuals use similar schemas to make sense of an event (Mandler, 1984).

Shared Attention Mechanisms. Recent investigations show that shared representations of the past could emerge in the absence of communication. Individuals, it was found, pay preferential attention to, and better remember, information that they believe similar others have access to at the same time (Shteynberg, 2010). In this particular instance, this shared attention mechanism makes people focus on the presented information unselectively, just by virtue of it being attended by similar others. This increase in shared attention could also guide participants as to what specifically to attend to. The resultant selective attention could create similar representations of an event, as when one is exposed to media coverage of causes of group-relevant conflict, for instance, simply because group members use their shared heuristics to decide what is salient for one's own group (Hastorf and Cantrill, 1954; Vallone et al., 1985).

Communicative Processes

Most typically, however, researchers have focused on the way people influence one another in a more direct fashion. An extant literature has focused, for instance, on how communicating about past events influences people's memories and renders them more similar.

Audience Tuning/Shared Attention. When we communicate to an audience, be it another person or a large group, we often have expectations as to what the audience wants to hear. These expectations affect not only what one communicates to others, but also the communicator's subsequent memories. Speakers, for example, often recount stories of the past in a biased manner to fit the audience's expectations. Interestingly, though, this effort to align one's message to the audience's expectations only results in altered memories consistent with the communicated message in situations when the speaker is motivated to create a connection with the audience. That is, only when speakers are motivated to share their beliefs, values, and attitudes with their fellow interactants does audience tuning result in consistent memory changes (Echterhoff et al., 2008). These motivations to create a shared reality can facilitate, thus, convergent memories between speakers and listeners.

Conversational/Dyadic Interactions. Most of our interactions are, however, in the form of extensive conversations in which content is exchanged bidirectionally. There is extensive

empirical research on how these conversations affect our memories. In these conversational paradigms, participants are instructed to study material such as newspaper article, a video, or a story and then are asked to recall the story first individually, then jointly as a pair, and finally, again, individually. Analysis of the content of the final individual recall reveals how the processes that unfold during joint remembering affect participants' individual memories.

Employing this experimental paradigm to study how individuals influence each other's memories in a conversational setting resulted in the discovery of three important psychological processes. First, there is a significant literature on the social contagion of memory. According to this research, information that only one participant knew emerges in conversations and is taken over by the other conversational participants (Roediger et al., 2001). Critically, the contagious information is often misleading, that is, an inaccurate recollection of the past. Second, information mentioned in the conversation is reinforced through practice effects (Karpicke and Roediger, 2007). That is, repeatedly listening to certain pieces of information will make this information more memorable, and more likely to emerge in future recalls for both speakers and listeners. And third, information that is not mentioned in the conversation is more likely to be forgotten by the interactants (Cuc et al., 2007). Critically, information that is related to what is discussed in the conversation is more likely to be forgotten than information that is unrelated to what is discussed. This phenomenon is known as retrieval induced forgetting. Information that the speaker recounts in the conversation results in induced forgetting in his/her related memories (*within-individual retrieval-induced forgetting*, or WIRIF), and also in the listeners' related memories (*socially shared retrieval-induced forgetting*, or SSRIF). Given that in most situations WIRIF and SSRIF occur simultaneously, they lead to forgetting of the same material between speakers and listeners, and thus increase in the overlap among conversational participants' memories is reached. By and large, all three discussed effects have cumulative effects insofar as mnemonic convergence is concerned.

Group Processes

Instead of focusing on the influence one person has on the memory of another, some researchers have examined how group processes affect what people remember as the group attempts to jointly recount the past. Several findings emerge. First, individuals distribute the task of remembering, a process called transactive memory (Wegner, 1987). Couples, for instance, tacitly coordinate during the memory task, such that each member of the pair focuses attention on a subset of the material to be remembered. They use their knowledge of the other interactant to infer what information they should pay attention to and later remember. It is this knowledge, an essential aspect of a transactive memory system, which allows for increased group recall performance.

Second, by analyzing a group's memory performance, researchers found that a group remembers more than any individual member on their own (collaborative facilitation), but less than the sum of the nonoverlapping individual memories remembered in isolation by group members (collaborative inhibition). Several explanations have been proposed for these group phenomena (for a review, see Hirst and Echterhoff, 2012). Suffice it to say that the surfacing of memories in joint remembering was found to impact group members' memories making them more similar to one another's (Barnier et al., 2008; Rajaram and Pereira-Pasarin, 2010).

Information Propagation and Extension to Social Networks

Most of the research reviewed under this section has focused on dyadic, or small group interactions. We know, however, that mnemonic effects go beyond these small group interactions. For instance, in an experiment designed to understand information propagation in social networks Harber and Cohen (2005) studied how a morgue visit by 33 students propagated in conversations in their social networks and reached nearly 900 people within 10 days. The finding that the more emotionally arousing one's experience the deeper into one's social network the information spreads, showcases the importance of individual processes in as much as propagation is concerned. This emotional contagion effect is due to the fact that the individual is more likely to talk about emotional material in vivid terms, which might make the information more likely to propagate (Berger and Milkman, 2012; Sperber and Hirschfeld, 2004).

Following up on the idea that the networks people are embedded in are critically important when it comes to understanding collective phenomena, recent investigations have started to reveal how cognition interacts with structural features of the network to impact mnemonic convergence. Coman et al. (2012), for instance, implemented the results of their empirical studies on conversational influences on memory into computer simulations. They used agent-based computer simulation, a method that allows for the creation of agent societies, akin to human societies. By manipulating the characteristics of agents (e.g., memory, identity) and their interactions (e.g., the degree of mnemonic influence), Coman et al. set out to study the parameters that affect mnemonic convergence. They showed that the degree of dyadic influence, network structure, network size, and number of conversations are all critical factors when it comes to creating collective memories. This work advocates for an approach that seeks to establish general principles by operationalizing collective memory and examining the contribution of cognitive mechanisms as well as the social dynamics to the emerging collective representations.

While acknowledging that studying large groups of interacting individuals is challenging from both practical and methodological perspectives, researchers interested in collective dynamics have made meaningful progress in understanding the spread, stability, and the convergence of memories within both small and large communities.

Collective Memory as a Product: The Role of Identity

Even though the relation between collective memories and identity is widely theorized within the social sciences, it has not yet been explored in much detail in the psychological literature. This is surprising, inasmuch as identity, and particularly collective identity has been the topic of empirical investigations in social psychology for much of its history (Ellemers et al., 2002). The exploration of the relation between collective memory and collective identity has been successfully undertaken, however,

by a small group of psychologically minded sociologists. They employ an aggregation method to study collective memory. For instance, Schuman and his collaborators asked participants to name important public events that happened during a period of 50 years. By computing the proportion of the sample that remembered public events within this time span, they conclude that different cohorts have different collective memories. Events that happened when participants were in their adolescence and young adulthood were much more likely to be mentioned as important when compared to events outside of this critical period. The study of individual autobiographical memories reveals the same pattern, referred to as the *reminiscence bump* (Berntsen and Rubin, 2002). Critically, the explanation that applies to both the individual and public reminiscence bumps is that identities become crystalized during this transitional period in one's life, which facilitates the integration of these events into our identities (Brown et al., 2009).

Linking process and product, and building on Bartlett's schemata theory, Wertsch (2002) proposed that culturally dependent schemas lead to the emergence of different collective memories for different cultures. He proposed that individuals employ schematic narrative structures to guide remembering of public events. For instance, the narrative structures that Russians employ when remembering the past start with a context in which Russia is unrightfully attacked by outsiders, then, overwhelmed by them, and finally, through courage and solidarity, the Russians triumph over their enemies. This schematic narrative structure is then, employed by individuals to make sense of and remember past and present events. This approach is one that proposes an explanatory process to the observed collective memories.

In order to explore the process–product link, scholars have started to take seriously the claims that collective representations are distributed in the population and that they change over time (Michelian and Sutton, 2013). Attempts to operationalize collective memory in a positivistic manner would not only provide a framework in which these claims could be rigorously investigated, but would also advance our current understanding of the factors that drive the emergence of collective memory.

Conclusion

This article presented the historical background as well as recent empirical advances in the psychology of collective memory. After tracing the roots of the psychological investigations of collective memory to Bartlett's visionary approach to memory, I have succinctly presented the current empirical work on the psychology of collective memory.

The advances in this burgeoning field show that integrating psychological insights into approaches aimed at understanding large-scale dynamics is a promising avenue for research. The scope of this investigation is larger than simply understanding the formation of collective memories, since these theoretical and methodological advances could be used to construct a scaffold for understanding how communities mobilize their identities, how they make decisions and solve problems collectively. Scholars have already convincingly made the case for an approach aimed at bridging between micro and macroprocesses when it comes to understanding culture (Cole, 1996; DiMaggio, 1997). Even though approaches that psychologized culture to an extreme, such as Parson's attempt to study it as a collection of beliefs, values, and norms, are cautioned against as unrealistic (DiMaggio, 1997), an effort to identify potential bridges between micro and macroprocesses is certainly within the reach of social sciences today. A social interactionist approach in which both individual cognition and social dynamics are interactive units of a system has produced, as shown in this contribution, meaningful advances within the social sciences.

See also: Collective Behavior, Social Psychology of; Collective Behavior, Sociology of; Collective Beliefs, Sociological Explanation of; Collective Memory, Anthropology of; Culture and Networks; Culture, Cognition and Embodiment; Experimentation in Psychology, History of; Methodological Individualism: Philosophical Aspects; Nation-State as Symbolic Construct; Network Analysis; Reconstructive Memory, Psychology of.

Bibliography

Assmann, J., 1995. Collective memory and cultural identity. New German Critique 65, 125–133.
Barnier, A.J., Sutton, J., Harris, C.B., Wilson, R.A., 2008. A conceptual and empirical framework for the social distribution of cognition: the case of memory. Cognitive Systems Research 9 (1), 33–51.
Bartlett, F., 1932. Remembering: A Study in Experimental and Social Psychology. Cambridge University Press, New York.
Berger, J., Milkman, K., 2012. What makes online content viral? Journal of Marketing Research 49 (2), 192–205.
Bernsten, D., Rubin, D.C., 2002. Emotionally charged autobiographical memories across the life span: the recall of happy, sad, traumatic, and involuntary memories. Psychology & Aging 17, 636–652.
Brown, N.R., Lee, P.J., Krsiak, M., Conrad, F.G., Havelka, J., Reddon, J.R., 2009. Living in history: how war, terrorism, and natural disaster affect the organization of autobiographical memory. Psychological Science 20, 399–405.
Clark, A., Chalmers, D., 1998. The extended mind. Analysis 58, 10–23.
Cole, M., 1996. Cultural Psychology. A Once and Future Discipline. Harvard University Press, Cambridge, MA.
Coman, A., Brown, A.D., Koppel, J., Hirst, W., 2009. Collective memory from a psychological perspective. International Journal of Politics, Culture, and Society 22 (2).
Coman, A., Kolling, A., Lewis, M., Hirst, W., 2012. Mnemonic convergence: from empirical data to large-scale dynamics. Lectures in Computer Science 7227, 256–265.
Connerton, P., 1989. How Societies Remember. Cambridge University Press, New York.
Cuc, A., Koppel, J., Hirst, W., 2007. Silence is not golden: a case for socially-shared retrieval induced forgetting. Psychological Science 18, 727–733.
DiMaggio, P., 1997. Culture and cognition. Annual Review of Sociology 23, 263–287.
Ebbinghaus, H., 1913. Memory: A Contribution to Experimental Psychology. Teachers College, Columbia University, New York original work published in 1885.
Echterhoff, G., Higgins, E.T., Kopietz, R., Groll, S., 2008. How communication goals determine when audience tuning biases memory. Journal of Experimental Psychology: General 137, 3–21.
Edelson, M., Sharot, T., Dolan, R.J., Dudai, Y., 2011. Following the crowd: brain substrates of long-term memory conformity. Science 333 (6038), 108–111.
Ellemers, N., Spears, R., Doosje, B., 2002. Self and social identity. Annual Review of Psychology 53, 161–186.
Halbwachs, M., 1980. The Collective Memory (F.J. Ditter & V. Y. Ditter, Trans). Harper and Row, New York (Original work published 1950).
Harber, K.D., Cohen, D.J., 2005. The emotional broadcaster theory of social sharing. Journal of Language and Social Psychology 24, 382–400.

Hastorf, A.H., Cantril, H., 1954. They saw a game: a case study. Journal of Abnormal and Social Psychology 49 (1), 129–134.
Hirst, W., Echterhoff, G., 2012. Remembering in conversations: the social sharing and reshaping of memory. Annual Review of Psychology 63, 55–79.
Hirst, W., Manier, D., 2008. Towards a psychology of collective memory. Memory 16, 183–200.
Hutchins, E., 1995. Cognition in the Wild. MIT Press, Cambridge, MA.
Karpicke, J.D., Roediger, H.L., 2007. Repeated retrieval during learning is the key to long-term retention. Journal of Memory and Language 57, 151–162.
Mather, M., Cacioppo, J.T., Kanwisher, N., 2013. Introduction to the special section: 20 years of fMRI – what has it done for understanding cognition? Perspectives on Psychological Science 8, 41–43.
Mandler, J.M., 1984. Stories, Scripts, and Scenes: Aspects of Schema Theory. Lawrence Erlbaum Associates, Hillsdale, NJ.
Michaelian, K., Sutton, J., 2013. Distributed cognition and memory research: history and current directions. Review of Philosophy and Psychology 4 (1), 1–24.
Olick, J.K., 1999. Collective memory: the two cultures. Sociological Theory 7, 333–348.
Rajaram, S., Pereira-Pasarin, L., 2010. Collaborative memory: cognitive research and theory. Perspectives on Psychological Science 5, 649–663.
Roediger, H.L., Meade, M.L., Bergman, E., 2001. Social contagion of memory. Psychonomic Bulletin Review 8, 365–371.
Schacter, D.L., 1999. The seven sins of memory. American Psychologist 54, 182–203.
Shteynberg, G., 2010. A silent emergence of culture: the social tuning effect. Journal of Personality and Social Psychology 99, 683–689.
Sperber, D., Hirschfeld, L., 2004. The cognitive foundations of cultural stability and diversity. Trends in Cognitive Sciences 8, 40–46.
Udéhn, L., 2002. The changing face of methodological individualism. Annual Review of Sociology 28, 479–507.
Vallone, R.P., Ross, L., Lepper, M.R., 1985. The hostile media phenomenon: biased perception and perceptions of media bias in coverage of the Beirut Massacre. Journal of Personality and Social Psychology 49, 577–585.
Wegner, D.M., 1987. Transactive memory: a contemporary analysis of group mind. In: Mullen, B., Goethals, G.R. (Eds.), Theories of Group Behavior. Springer-Verlag, New York, pp. 185–208.
Wertsch, J.V., 2002. Voices of Collective Remembering. Cambridge University Press, New York.
Young, J.E., 1993. The Texture of Memory: Holocaust Memorials and Meaning. Yale University Press, New Haven and London.

Relevant Websites

http://www.johnsutton.net.
http://www.psychology.sunysb.edu/srajaram-/index.html.
http://artsci.wustl.edu/∼pboyer/LuceWebSite/IndexGrey2.html.
http://www.culturomics.org.

Collective Memory, Social Psychology of

Laurent Licata and Aurélie Mercy, Research Center for Social and Cultural Psychology, Unité de Psychologie Sociale, Université Libre de Bruxelles, Bruxelles, Belgium

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Prager, volume 4, pp. 2223–2227, © 2001, Elsevier Ltd.

Abstract

Collective memories are shared representations of a group's past based on a common identity. Social psychological approaches of collective memory have addressed their antecedents and functions. Their formation is affected by cognitive and emotional factors, but it takes place in the context of human interactions with other humans or with cultural artifacts. They are shaped by, and transmitted through, narratives. Because they intervene in the definition, maintenance, and mobilization of social identities, they have a strong impact on intergroup relations. Collective memories influence the present, but they are also influenced by present psychological states and needs.

Introduction

Through the expression 'collective memory,' a psychological faculty – memory – is attributed to a social group, whether small (a family, a group of friends) or large (a whole nation). Collective memories are instrumental in defining, maintaining, and mobilizing social identities. As a consequence, they are involved in intergroup relations, particularly in intergroup conflict and reconciliation processes. For example, Barack Obama began his second inaugural address (21 January 2013) by reminding the principles of equality and freedom on which the USA were founded. Then he emphasized the continuity of their history: "Today we continue a never-ending journey, to bridge the meaning of those words with the realities of our time" (White House, 2013). And he listed some of the lessons learned from this history. The first one was about racial equality: "Through blood drawn by lash and blood drawn by sword, we learned that no union founded on the principles of liberty and equality could survive half-slave and half-free." This was an allusion to Abraham Lincoln's second inauguration address: "until every drop of blood drawn with the lash shall be paid by another drawn with the sword." Through this speech, Obama depicted the USA as a united people that, beyond social and racial divisions, share the same identity and strive to achieve the same goals. He did it by drawing on the collective memory shared by Americans and by stressing the continuity of the nation, as well as by casting himself as the continuator of the founding fathers of the nation. Thereby, he asserted his legitimacy as the President of the USA and tried to mobilize Americans toward the realization of his political agenda.

Most scholars refer to the sociologist Maurice Halbwachs (1980, 1992) as the founding father of collective memory studies, although he did not coin the term (Olick et al., 2011). Halbwachs was influenced both by Emile Durkheim's sociology of collective representations and by Henri Bergson's philosophy of subjectivity. Accordingly, Halbwachs tried to conciliate an individual-based and a collective-based accounts of collective memory. But Halbwachs defended different, apparently incompatible, theses on the nature of collective memories. In *The collective memory*, he wrote "it is individuals as group members who remember" (Halbwachs, 1980: p. 48) whereas, in *The social frameworks of memory*, he stated "it is only natural that we consider the group in itself as having the capacity to remember" (Halbwachs, 1992: p. 54). Thus, Wertsch (2002) distinguishes two interpretations of collective memory: the strong and the distributed versions. According to the strong version, groups have memories above and beyond the minds of group members. Wertsch views this version as a metaphor that can be useful in some circumstances – i.e., when dealing with very cohesive social groups – but should not be taken literally. In contrast, the distributed version assumes that group members share some representations of the past, without positing that the group itself has a memory. Among distributed accounts of collective memory, one can further distinguish homogeneous and complementary variants. The homogeneous variant assumes that all group members share the same representation of the past, whereas the complementary variant posits that group members hold different representations, but that these are coordinated to form a system made of complementary pieces.

Frederic Bartlett (1932) is generally credited for having introduced collective memory in the field of psychology. In Remembering: a study in experimental and social psychology, he defended a distributed account of collective memory, and criticized Halbwachs for having reified the concept. However, according to Bartlett, remembering is social in nature. Far from being enclosed in the individual's mind, it is a skilled activity that takes place in the framework of human interactions (Sutton, 2008). This activity does not amount to retrieve information that was previously stocked in the individual memory system, but rather implies an active reconstruction of memory traces as a function of the present situation. More precisely, it is the interaction between schemata inherited from our cultural background or from our personal experiences and the situational context that determines this reconstruction. This reconstruction is therefore influenced by the social context in which remembering takes place, which implies that representations of the past can be distorted, both at the individual and at the collective levels. Indeed, social groups such as nations 'flatter and deceive themselves' (Baumeister and Hastings, 1997).

As an object of scientific inquiry, collective memory lies at the crossroads between behavioral and social sciences. Historians, sociologists, political scientists, philosophers, cultural scientists, anthropologists, and philologists, as well as psychologists have contributed to develop what has been referred to as 'memory studies' (Olick et al., 2011). Somehow paradoxically, the contribution of psychologists, and particularly of social psychologists, appeared relatively late in the history of this movement, and it still represents only a small portion of the bulk of studies and theoretical developments on collective memory. In psychology, interest in collective memory and social representations of history sprang in the 1930s with the work of Bartlett (1932) then waned only to regain some visibility from the 1990s on. However, as we shall see, this contribution has been continuously increasing ever since and, as we shall argue, this contribution is significant.

Psychologists have addressed different categories of questions regarding collective memory: how do collective memories form? What are their social psychological functions? And what is their influence on intergroup relations?

The Psychological Antecedents of Collective Memory

Cognition

Coman et al. (2009) anchor their approach of collective memory in cognitive psychology, in line with Bartlett's interactionist and constructivist stance. They reckon that collective memory should not be conflated with individual psychology. However, they emphasize that collective memories should not only be studied at the collective level – publicly available symbols maintained by society – because it is the interplay between these public symbols and individual psychology that makes these memories collective. This interplay is necessary because of the very nature of human memory. Coman et al. advocate a social interactional approach to memory according to which internal and external factors should be considered: "When it comes to the study of memory, the claim would be that, in order to understand how a person remembers what he ate last night, a researcher needs to consider not only the neural mechanisms underlying memory, but also, for instance, the visual cues in the environment, such as the pile of dishes in the sink" (2009: p. 127). Human beings reconstruct their memories by using cues available in their environment. For example, bartenders use the shape of glasses as cues to remember complex orders. They place the appropriate glass on the bar when they receive the order; the presence of different glass shapes then guides their memory. The use of social artifacts is seen as a response to the limitation of human mnemonic capacities. As Sutton (2008: p. 220) pointed out, "it's just because isolated items aren't stored atomically in the brain that our relatively vulnerable biological memories are supplemented by more stable external scaffolding. Brains like ours need media, objects, and other people to function fully as minds." Likewise, society restructures the world so that the group better remembers. For example, the Washington Lincoln memorial was built in order to guide people's remembering of this president as a godlike figure. Thus, the memories of people who visited the Lincoln memorial are influenced in a similar way, which gradually changes their collective memory. Thus, the interactions that group members have with social artifacts shape their memories (Coman et al., 2009).

With his 'lay historian' metaphor, Klein (2013) also advocates for the taking into account of the role of cognitive processes in the formation of collective memory. Lay people are regularly facing the necessity of explaining the past, either for themselves or for others. Klein parallels the construction of lay knowledge about the past with the three steps involved in historical research: archival (access to knowledge), explanatory (elaboration of causal reasoning about past events), and representational (elaboration and communication of a representation adapted to an audience). Cognitive or sociocognitive processes intervene at each of these steps when lay historians form their historical knowledge. These processes may lead to biases, thus hampering the epistemic function of collective memory. For example, people's evaluation of the truthfulness of testimonies depends on their evaluation of the reliability of the source: they tend to use ready-made heuristics when the source is deemed trustful, or adopt a critical posture when it is not, thus influencing the archival step. At the explanatory step, other cognitive phenomena may then interfere. For example, people generally try to explain unexpected or unpleasant events, thus neglecting events that appear to follow the 'normal course of history' or that are pleasant. They also underestimate chance and overestimate causal explanations. In addition, they often fall prey to the retrospective illusion or hindsight bias: when the outcome of a sequence of events is known, people tend to overestimate the probability of that particular outcome over all other possible ones. This might lead to reconstructing the sequence of events in order to explain this outcome through a coherent narrative that logically leads to it, thus obliterating elements of the past that do not contribute or contradict this narrative. Finally, social cognitive biases can also interfere at the representational step, when a representation of the past is communicated to others through narration. Explaining is always explaining something to someone. Therefore, the representation that one communicates may be influenced by the characteristics of the audience one is addressing. A study showed that participants provided different explanations of a mass murder as a function of whether they addressed a personality psychologist or a nonspecialist: their explanations were more dispositional in the former than in the latter case. Thus, people adapt their message as a function of the way they perceive their audience. In brief, various cognitive and social cognitive processes known by psychologists may shape laypeople's understanding of the past and, indirectly, contribute to the formation of collective memories.

Collective memory approaches anchored in cognitive psychology or social cognition already pay due attention to interactional factors. Some approaches have gone a step further in that direction by stressing the importance of conversations in the formation and transmission of collective memories.

Conversation

According to Middleton and Edwards' (1990) discourse-analytical approach, in the course of conversations, people

coconstruct their memories or contest other people's accounts of an event. Versions of events are seen as pragmatic speech acts deemed to achieve rhetorical functions, so that a version cannot be taken as a person's real memory. Collective remembering is then studied as a joint activity rather than as the product of this activity: memory is the act of remembering and forgetting together.

Coman et al. (2009) also contend that conversations play a pivotal role in the formation of collective memories. People often talk about past events that had important consequences for their social group. These conversations can thus be instrumental in the spreading of memories across the group. One speaker, especially if she is seen as an expert, may influence a listener's memory that could, in turn, influence another listener, and so forth. For example, individual memories of family members were collected before and after a conversation about a story they had just read. Individual recollections were much more similar after than before the collective remembering, and story details that were shared during the conversation were disproportionately included in individual accounts of the story. Conversely, collective remembering can also lead to the forgetting of some elements of a narrative: listening to someone who communicates some items of a narrative, but fails to communicate other related items, facilitates recall of the communicated items while inhibiting the recall of unmentioned ones (Stone et al., 2010).

Narratives

Bartlett (1932) studied the formation of collective memories by asking participants to recount a story they had just read to others. According to Jerome Bruner (1990), human thought is organized by storytelling, but this cognitive mechanism is validated and maintained socially. As accounts of events involving some temporal or causal coherence (Laszlo, 2013), narratives have been viewed as means *par excellence* for forming and transmitting collective memories. In cultures with literacy, these historic narratives are mainly conveyed through texts. Written accounts of the distant or close past of nations are very influential in shaping and maintaining a sense of national identity. These narratives are not univocal: they are debated and contested in the public sphere, where several versions of the same events often compete (Jovchelovitch, 2012; Reicher and Hopkins, 2001). But this very process of contestation ensures that the memory of the event is kept alive. Laszlo (2013) has studied the historical trajectories of nations as remembered by laypeople. For example, these studies showed that the history of Hungary is remembered as a story of grandeur (anti-Ottoman, then anti-Austrian fights) followed by decline (Trianon Treaty, wars), then by a slow but incomplete redress (end of Soviet rule). Hungarian collective memory is familiar with the 'First victory, then failure' pattern, but not with the 'Failure, then victory' one, which prevents Hungarians from making sense of the fall of the Berlin wall as it rather belongs to the latter.

Emotions

When an event affects the lives of large numbers of individuals, it triggers emotions. As these emotions are shared among group members, they may facilitate the formation of collective memories about that event. Hence, according to Pennebaker and Banasik (1997), collective events are more likely to lead to collective memories when they bring social change; are emotionally loaded; elicit social sharing; are socially rehearsed by the mass media; and are associated with collective behaviors such as rituals, demonstrations, or commemorations.

Flashbulb memories are vivid recollections of the personal circumstances in which one receives the news about a shocking event. For example, many people remember how was the weather, where they were, or what tune was playing on the radio when they heard about the 9/11 attacks. Research on flashbulb memories concentrated quasi exclusively on negative – often violent – collective events, such as the assassination of political leaders, terrorist attacks, disasters, and catastrophes (Paez et al., 2009). These events prompt emotions in most, if not all, members of a population. Berntsen (2009) has shown that social identity relevance, rather than personal relevance, was central for the formation and maintenance of flashbulb memories. Flashbulb memories and collective memories share many characteristics and are probably two ways of studying the same phenomenon (Paez et al., 2009).

Emotions are therefore crucial determinants of the formation of collective memories, but their role extends far beyond this formation phase. Collective rituals are particularly effective in maintaining collective memories. To be sure, commemorative ceremonies are designed for this very function. Those rituals, through their form and content, are effective in reminding the event through the repetitive reactivation of the cognitive representation. However, rites are more than reminders, they are reenactments of the event (Connerton, 1989) and, as such, they also engage bodies. In collective emotional gatherings, people perform the same gestures, adopt the same postures, sing the same songs at the unison. This synchrony leads to an important social sharing of emotions, to the point of eliciting feelings of 'identity fusion.' This experience then revives the common identity and the beliefs associated with it (Paez and Rimé, 2014).

The Social Psychological Functions of Collective Memory

Cognitive, social interactional, and emotional factors thus play a role in the formation and maintenance of collective memories. Beyond the question of their formation, social psychologists have devoted much attention to their functions. Rather than providing an accurate account of the past, their main function is to provide a *usable past* (Wertsch, 2002); a representation of the past that can fulfill some role in the present. Collective memory "retains from the past only what still lives or is capable of living in the consciousness of the groups keeping the memory alive" (Halbwachs, 1980: p. 80). According to psychologists, collective memories are kept alive if they meet the needs and motivations of group members. These needs have been described as epistemic – providing a clear representation of the group's past – and/or identitary – defining the group's identity. Social identity (Tajfel and Turner, 1986) and social representations theories (Moscovici,

2008), either independently or in combination, have often been mobilized in this framework.

Social Identity Approaches to Collective Memory

Tajfel and Turner (1986: p. 16) defined social identity as "those aspects of an individual's self-image that derive from the social categories to which he perceives himself as belonging." Collective memory fulfills various functions with regard to social identity (Klein et al., 2011). First, collective memories contribute to the group's definition. According to Liu and Hilton (2005: p. 537), "History provides us with narratives that tell us who we are, where we came from and where we should be going. It defines a trajectory which helps construct the essence of a group's identity, how it relates to other groups, and ascertains what its options are for facing present challenges." Representations of the group's past, which are constantly renegotiated as a function of the present situation, thus provide content to the social identity. In addition, these representations have a normative function as they define how that group should behave. Paralleling Malinowski's views on myths, Liu and Hilton contend that social representations of history serve as 'charters' that define how the group should react to a present situation, because group members draw a 'moral' from their interpretation of a past event. Beyond these cognitive and normative aspects, the simple fact that those representations are shared brings a sense of cohesion among group members. Thus, this imagined past simultaneously creates a 'horizontal' solidarity between contemporary group members and a 'vertical' solidarity with past and future members of the group. In brief, sharing representations of the group's past confers psychological reality to that group.

In particular, representing the group as continuous through time facilitates identification with it (Sani, 2008). Conversely, presenting the group's history as discontinuous may elicit collective angst regarding the group's future, which can lead to group protective attitudes, such as opposing immigration (Jetten and Wohl, 2012; Smeekes and Verkuyten, 2013).

Second, collective memory contributes to set the group's value. According to Social identity theory (Tajfel and Turner, 1986), people are motivated to hold a positive view of their group because their social identity is an intimate part of their self-concept. The group's value is assessed through intergroup comparisons. As a consequence, one needs to belong to a group that compares positively with relevant out-groups in order to derive a positive collective self-esteem. A group's past – its successes and failures; its moral or immoral actions – hence contribute to defining its relative value. Social groups thus compare their respective pasts, and generally strive to derive a positive sense of their identity through this comparison. This helps explain why collective memories are often biased toward positive accounts of the in-group's past and/or negative accounts of out-groups' actions in history (Baumeister and Hastings, 1997). In addition, it accounts for the fact that some representations of the group's past – such as reminders of past misdeeds – elicit strong defensive reactions (Branscombe and Doosje, 2004).

Third, collective memories can be used for legitimizing past, present, or planned actions of the group. For example, Serb nationalists referred to memories of the Kosovo battle in 1389, after which the Serbs had lost control over the Kosovar territory, in order to legitimize the deportation of Muslim Kosovars in the 1990s.

Fourth, collective memories can be used in order to mobilize members of a social group to accomplish a particular collective – often political – project in the name of their shared identity (Reicher and Hopkins, 2001). For example, the Congolese leader Patrice Lumumba adapted his discourse as a function of the audience he was addressing: in front of a Congolese audience, he would represent the history of colonization as undermined by the Belgians' cupidity and contempt whereas, in front of a Belgian audience, he would emphasize the friendship that the colonial experience had allowed to build between the two nations. Far from being a simple example of opportunism, this apparent lack of coherence actually reflected a stable political project, which demanded different actions from different actors. Hence, he had to mobilize the Congolese against the paternalistic colonial ideology, as well as make sure that the Belgians would collaborate with the new independent State (Klein and Licata, 2003).

Collective Memory and Reconciliation Processes

The management of collective memories of a past conflict is crucial to the quality of the present and future intergroup relations. Indeed, narratives shared inside the implicated groups often diverge. Given that social groups are motivated to perceive themselves in a positive way, these narratives are fundamentally selective, including and excluding historical events in order to characterize one's group in a positive way, at the expense of the former enemy's image (Bar-Tal, 2003).

In a postconflict context, the status attributed to a group during the conflict is determinant for the contents of the collective memory. These contents will be very different whether the group is casted as the victim or as the perpetrator in the historical narrative. Thus, if the positive actions of a group allow its members to value, and assert their moral character, its negative actions pose a significant threat to collective self-esteem, especially among highly identified group members (Branscombe and Doosje, 2004).

Memories of past misdeeds committed by the group are therefore difficult to integrate into the group's history. Being casted as a perpetrator group has consequences both within the social group and between social groups. Individuals seek to differentiate themselves from members who have committed negative actions. Therefore, the perpetrator status is associated with important tension, undermining social cohesion within the group. However, this status may also have a positive impact at the intergroup level. Indeed, group members may experience negative group-based moral emotions, such as collective guilt, shame, or regret (Giner-Sorolla, 2012). For example, memories of colonialism have been shown to trigger collective guilt or shame among young generations in formerly colonizing countries (Volpato and Licata, 2010). These aversive emotions motivate individuals to implement acts of reparation toward out-group members (Branscombe and Doosje, 2004).

The victim status also affects the social dynamics within and between groups. This position can be traumatic for the group, and hence also threaten its cohesion. However, if the traumatic event is commemorated – which highlights the common destiny of group members and accentuates distinction from other groups – the integration of past in-group sufferings in collective memory may have the opposite effect and increase social cohesion (Bar-Tal et al., 2009).

Victimhood also involves a competitive component: people draw comparisons between the sufferings of the groups involved in the conflict (Noor et al., 2012). This competitive victimhood implies a vision of the suffering of one's group as being more important and less legitimate compared to that of the out-group, and represents a strong obstacle to intergroup reconciliation.

However, perpetrator and victim statuses are not always incompatible. Indeed, in intergroup conflicts such as wars, and especially civil wars, group members are often involved in violent actions against the out-group, while themselves undergoing violence. In that case, the recognition of one's group's victim status might prevent group members from implementing reconciliation mechanisms. Indeed, perceived collective victimhood tends to inhibit empathy toward the out-group and provides an opportunity to discard any responsibility in the conflict (Bar-Tal et al., 2009). It then tends to reduce propensity to experience collective guilt. Yet, victimhood can be reframed: group members can include their former enemies into a common group, thus sharing a common identity of individuals having suffered from the conflict (Vollhardt, 2012). This inclusive view of victimhood has the opposite effects on intergroup relations: it leads to more empathy and increases willingness to forgive the out-group.

Statuses of victim and perpetrator are associated with a gain or loss of power in intergroup dynamics. In the aftermath of the conflict, the victimized group may appeal to the international community to recognize its status, or to obtain compensation from its former opponent, which can increase its relative power. The relationship between collective memory and power is thus mutual: the dominant group often has control over the official memory, particularly through public commemorations. In contrast, low-status groups strive for recognition of their status in the official narrative. Therefore, discourses about memory should be interpreted as a function of the group's current position and of its objectives (Reicher and Hopkins, 2001).

Collective Memories as Social Representations

The way a social group relates to its past can be envisioned in two different manners: collective memories are either depicted as weighing on the group's present, or as being influenced by the group's present (Prager, 1998). The past can be undergone or it can be chosen. For example, one can view memories of a collective trauma as bearing on the present, independently of people's will. Survivors of genocides and their descendants have to live with these sore memories, as they are included in their social identity. However, collective memories can also be tailored to fulfill current identity needs, or to legitimize behaviors. This rejoins Bartlett's contention that "with the individual as with the group, the past is continually re-made, reconstructed as a function of present interests" (1932: p. 309). For example, after the 9/11 attacks, American participants prompted to cite significant events of the last 70 years mainly listed events involving war or an attack against an American institution (Schuman and Rodgers, 2004). This suggests that people remember events that are in line with their present mental state: present to past. However, representations of the past can also influence the present. Indeed, drawing analogies between the 9/11 attacks and the Munich agreement of 1938 or to Pearl Harbor yielded an understanding of these attacks through the prism of a war metaphor. In turn, this metaphor helped legitimizing attacks of Afghanistan, then Iraq (Kruglanski et al., 2007). Thus, representations of history influence the interpretation of current situations: past to present. This example suggests that both directions of influence are part of the same global dynamic: present situations influence the choice and interpretation of particular memories, which in turn are used to make sense of present situations and to tailor people's attitudes and behaviors.

Moscovici's social representations theory (2008), provides a theoretical framework that accounts for both directions of influence. It posits that social representations form through two processes: objectification and anchoring. Objectification is the process through which complex, often abstract, concepts are gradually transformed, through social exchange, into concrete, visual images. Anchoring relates to the integration of a new knowledge into a group's preexisting knowledge system. The latter process is useful for understanding how the past weighs on the present (Liu and Hilton, 2005) as – already shared – collective memories of past events are used for making sense of current events. The objectification process is useful to account for the fact that the past is continually reconstructed as a function of present stakes (Clémence, 2002). Hence, contemporary attitudes, norms, and values are projected onto past events and historical characters in a process of objectification. This creates the illusion that current states of mind or ideological choices are based on objective historical facts, which endows them with legitimizing power (Klein et al., 2011). Finally, these past events or figures, which actually objectify present concerns, can then serve to anchor new knowledge. As Jan Assmann (1997: p. 9) pointed out "The past is not simply 'received' by the present. The present is 'haunted' by the past and the past is modeled, invented, reinvented, and reconstructed by the present."

See also: Collective Memory, Anthropology of; Collective Memory, Psychology of; Conflict and Conflict Resolution, Social Psychology of; Cultural History; Emotions and Intergroup Relations; Ethnic Identity, Psychology of; Intergroup Relations; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Narrative, Sociology of; Representations, Social Psychology of; Social Identity in Social Psychology; Social Psychology.

Bibliography

Assmann, J., 1997. Moses the Egyptian: The Memory of Egypt in Modern Monotheism. Harvard University Press, Cambridge, MA.

Bar-Tal, D., 2003. Collective memory of physical violence: its contribution to the culture of violence. In: Cairns, E., Roe, M.D. (Eds.), The Role of Memory in Ethnic Conflict. Palgrave Macmillan, Houndmills, pp. 77–93.

Bar-Tal, D., Chernyak-Hai, L., Schori, N., et al., 2009. A sense of self-perceived collective victimhood in intractable conflicts. International Review of the Red Cross 91, 229–258.
Bartlett, F.C., 1932. Remembering: A Study in Experimental and Social Psychology. Cambridge University Press, Cambridge.
Baumeister, R.F., Hastings, S., 1997. Distortions of collective memory: how groups flatter and deceive themselves. In: Pennebaker, J.W., Paez, D., Rimé, B. (Eds.), Collective Memory of Political Events: Social Psychological Perspectives. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 277–293.
Berntsen, D., 2009. Flashbulb memories and social identity. In: Luminet, O., Curci, A. (Eds.), Flashbulb Memories: New Issues and New Perspectives. Psychology Press, New York, pp. 187–205.
Branscombe, N.R., Doosje, B., 2004. Collective Guilt: International Perspectives. Cambridge University Press, Cambridge.
Bruner, J.S., 1990. Acts of Meaning. Harvard University Press, Cambridge, MA.
Clémence, A., 2002. Prises de position et dynamique de la pensée représentative: les apports de la mémoire collective. In: Laurens, S., Roussiau, N. (Eds.), La mémoire sociale: identités et représentations sociales. Presses Universitaires de Rennes, Rennes, pp. 51–61.
Coman, A., Brown, A.D., Koppel, J., et al., 2009. Collective memory from a psychological perspective. International Journal of Politics, Culture, and Society 22, 125–141.
Connerton, P., 1989. How Societies Remember. Cambridge University Press, Cambridge, NY.
Giner-Sorolla, R., 2012. Judging Passions: Moral Emotions in Persons and Groups. Psychology Press, London, UK.
Halbwachs, M., 1980. The Collective Memory. Harper & Row, New York.
Halbwachs, M., 1992. On Collective Memory. University of Chicago Press, Chicago.
Jetten, J., Wohl, M.J.A., 2012. The past as a determinant of the present: historical continuity, collective angst, and opposition to immigration. European Journal of Social Psychology 42, 442–450.
Jovchelovitch, S., 2012. Narrative, memory and social representations: a conversation between history and social psychology. Integrative Psychological and Behavioral Science 46, 1–17.
Klein, O., 2013. The lay historian: how ordinary people think about history. In: Cabecinhas, R., Abadia, L. (Eds.), Narratives and Social Memory: Theoretical and Methodological Approaches. Communication and Society Research Centre, Braga, pp. 25–45.
Klein, O., Licata, L., 2003. When group representations serve social change: the speeches of Patrice Lumumba during the Congolese decolonization. British Journal of Social Psychology 42, 571–593.
Klein, O., Licata, L., Van der Linden, N., et al., 2011. A waffle-shaped model for how realistic dimensions of the Belgian conflict structure collective memories and stereotypes. Memory Studies 5, 16–31.
Kruglanski, A.W., Crenshaw, M., Post, J.M., et al., 2007. What should this fight be called? Metaphors of counterterrorism and their implications. Psychological Science in the Public Interest 8, 97–133.
Laszlo, J., 2013. Historical Tales and National Identity: An Introduction to Narrative Social Psychology. Routledge, New York.
Liu, J.H., Hilton, D.J., 2005. How the past weighs on the present: social representations of history and their role in identity politics. British Journal of Social Psychology 44, 537–556.
Middleton, D., Edwards, D., 1990. Collective Remembering, Inquiries in Social Construction. Sage Publications, Thousand Oaks, CA.
Moscovici, S., 2008. Psychoanalysis: Its Image and Its Public. Polity Press, Cambridge.
Noor, M., Shnabel, N., Halabi, S., et al., 2012. When suffering begets suffering: the psychology of competitive victimhood between adversarial groups in violent conflicts. Personality and Social Psychology Review 16, 351–374.
Olick, J.K., Vinitzky-Seroussi, V., Levy, D., 2011. The Collective Memory Reader. Oxford University Press, New York.
Paez, D., Bellelli, G., Rimé, B., 2009. Flashbulb memories, culture, and collective memories. In: Luminet, O., Curci, A. (Eds.), Flashbulb Memories: New Issues and New Perspectives. Psychology Press, New York, pp. 227–245.
Paez, D., Rimé, B., 2014. Collective emotional gatherings. Their impact upon identity fusion, shared beliefs and social integration. In: von Scheve, C., Salmela, M. (Eds.), Collective Emotions. Oxford University Press, Oxford, UK, pp. 204–216.
Pennebaker, J.W., Banasik, B.L., 1997. On the creation and maintenance of collective memories: history as social psychology. In: Pennebaker, J.W., Paez, D., Rimé, B. (Eds.), Collective Memory of Political Events: Social Psychological Perspectives. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 3–20.
Prager, J., 1998. Presenting the Past, Psychoanalysis and the Sociology of Misremembering. Harvard University Press, Cambridge, MA.
Reicher, S., Hopkins, N., 2001. Self and Nation. Sage, London.
Sani, F., 2008. Self Continuity: Individual and Collective Perspectives. Psychology Press, New York.
Schuman, H., Rodgers, W.L., 2004. Cohorts, chronology, and collective memories. Public Opinion Quarterly 68, 217–254.
Smeekes, A., Verkuyten, M., 2013. Collective self-continuity, group identification and in-group defense. Journal of Experimental Social Psychology 49, 984–994.
Stone, C.B., Barnier, A.J., Sutton, J., et al., 2010. Building consensus about the past: schema consistency and convergence in socially shared retrieval-induced forgetting. Memory 18, 170–184.
Sutton, J., 2008. Remembering. In: Robbins, P., Murat, A. (Eds.), The Cambridge Handbook of Situated Cognition. Cambridge University Press, New York, pp. 217–235.
The White House, 2013. Inaugural Address by President Barack Obama. http://www.whitehouse.gov/the-press-office/2013/01/21/inaugural-address-president-barack-obama (accessed 31.01.13.).
Tajfel, H., Turner, J.C., 1986. The social identity theory of intergroup behavior. In: Worchel, S., Austin, W.G. (Eds.), The Psychology of Intergroup Relations. Nelson-Hall, Chicago, pp. 7–24.
Volpato, C., Licata, L., 2010. Collective memories of colonial violence: introduction to the focus section. International Journal of Conflict and Violence 4, 4–10.
Vollhardt, J., 2012. Collective Victimization. Oxford Handbook of Intergroup Conflict. Oxford University Press, New York, 136–157.
Wertsch, J.V., 2002. Voices of Collective Remembering. Cambridge University Press, Cambridge, UK; New York.

Collective Memory, Sociology of

Masahiro Ogino, Kwansei Gakuin University, Nishinomiya, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of collective memory introduced by Maurice Halbwachs is one of the key concepts to understand how a society maintains its order and how social change occurs. In this sense, the sociology of collective memory is not a simple branch of sociology and can contribute to an elaboration of general social theory all the more so as late-modern society is becoming the age of preservation when even negative memory like war experiences becomes an object of preservation and that of globalization politics of memory occupy an important or even primordial place.

How can social order be maintained? This is indisputably one of the most crucial questions in sociology. There are two approaches to answer this question. One is to focus on the durable form of society by resorting to terms like social structure or social system to designate the continuation of social order. Another approach is to try to understand by which means a society can last. Collective representation introduced by Émile Durkheim is one of the founding concepts of this approach (Durkheim, 1937). According to Durkheim, collective representation refers to a symbol having common-shared meaning to members of a social group or society. Sociology of collective memory introduced by Maurice Halbwachs, one of the disciples of Durkheim, is on the line of Durkheimian sociological theory.

Collective Memory in Halbwachs

Memory is a complex totality of knowledge and image and, generally, is considered as belonging to the individual. For Halbwachs, however, individual memory is not completely closed insofar as it is constructed through interactions with others. Individual memory is, then, influenced by the collective representation of the past. In this sense, there is no individual memory as such. Halbwachs designates this collective representation of the past as collective memory.

Jeffey K. Olick remarked quite rightly that collective memory in Halbwachs indicates two district sorts of phenomena: socially framed individual memories, and collective commemorative representations and mnemonic traces (Olick, 2007: 20). In the first part of his *The Social Framework of Memory* (Halbwachs, 1925), Halbwachs tends to discuss on the collective memory from the point of view of an individual. In particular when he treats familial memory, his argument focuses essentially on the way in which an individual constitutes his/her memory in relation with others, and through what he calls *framework*. Framework gives opportunities to remember the past related to one's individual life. Some members of the group and/or some important events constitute a framework which offers a reference point for remembering. Collective memory defined as such is not well demarcated. It is something uncertain and difficult to comprehend. Furthermore, collective memory changes with time, because those who compose the group, change. In other words, it is closely linked to the living past of those who compose the group. If the past loses this vivacity for the members of the group, it will be outside of the collective memory.

However, when Halbwachs analyzes religious collective memory, he treats this as something quite distinct, insofar as religious memory belongs to a group and is considered as fixed once and for all by the group. For example, the Church has its own memory which is religious doctrine. It seems to the members of the Church that the doctrine is unchanging. Christians believe that the memory of Christians faithfully keeps the life of Christ. In fact what they believe is not unchanging, because collective memory is essentially a reconstruction of the past. Halbwachs (1971) goes so far as to say, in particular, in his study on topography and religious memory, that knowledge of what was original is secondary, or even unimportant. Collective memory considered in this way comes close to collective representation, or even becomes synonymous with collective conscience. Religious memory considered in this way is more than a simple aggregation of individual memories.

Past in the Present or Past Constructed by the Present

The English-speaking and German-speaking tradition of social theory uses the term, *knowledge* rather than the concept of memory in the sense that it, in particular, takes an interest in the process of cognition and comprehension. George Herbert Mead pointed out that the past is in the present (Mead, 1932). He affirmed in a similar way as Halbwachs that the past is reconstructed by the present point of view. An individual thinks of the past always only at present and the knowledge about the past determined by the present may change (Adam, 1995: 81). Alfred Schutz inspired by the phenomenology attempts to show how an individual forms his/her stock of knowledge. He uses the term, 'retention' to designate the process by or in which an individual constitutes his/her stock of knowledge through his/her experiences. He also mentions the world of 'predecessors' who play a role in the formation of the knowledge on the 'social world' (Schutz and Luckmann, 1973).

Mead and Schutz as well as Halbwachs emphasize the constitutive aspects of the past while Michael Polanyi treats of unconscious acquirement of knowledge by resorting to the concept of 'tacit knowledge.' The essence of tacit knowledge

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32191-2

lies in Polanyi's statement, "We can know more than we can tell" (Polanyi, 1966: 4). That is, knowledge is not limited to that which we can consciously analyze using language; knowledge perceived physically precedes it, and Polanyi considered that this knowledge is more fundamental. For instance, students preparing their sociology graduation theses hear other students' interim reports about their theses and feel that they are interesting, but they cannot explain why the reports are interesting. This is because the students are not focusing on and analyzing the values of the specific themes of the theses and the problems they propose, but are simply making intuitive judgments. To use a term coined by Polanyi, the students have already 'internalized' their sociological knowledge. To wit, the knowledge as yet without form is indwelled, but the students are not aware of it, and therefore the students can only judge whether sociological problems have been constructed or not intuitively. Polanyi's theory illustrates that the indefinite portion of knowledge that exists prior to being given form by language (including things such as intuition, gut feeling, and senses) holds considerable significance for everything from our daily activities through to scientific discovery. Part of human intellectual work is indefinite and cannot be given form, and that the acquisition and creation of knowledge is impossible without this tacit part.

Nonetheless, Polanyi placed too much emphasis on the point that knowledge is created by the individual and ignored the relationship between the individual person and the external environment. Individuals act within a predetermined environment but are not necessarily aware of the relationship between that environment and their actions, and because people do not always act based on the most rational judgments, it is not given that they can clearly express the reality in which they exist using their own expressions.

There are two different ways of thinking about the past. One emphasizes the constitutive aspects of the past. The other insists on the unaware aspects of the transmission of the past. If there exists a tacit part of knowledge, this part has to be socially constituted and accumulated. The tacit knowledge thus presuppose some accumulative dimension. Tacit knowledge understood in this way comes close to the concept of 'habitus' in the social theory of Pierre Bourdieu. Habitus conceived by Bourdieu allows a given situation to be appropriate without any reflection, since habitus embedded in our body makes it possible to sustain the social structure (Bourdieu, 1980). In other words, by means of habitus, repetitive and unconscious practice in daily life assure the 'reproduction.' This question of cultural practice or quite simply of tradition in the sense given by Edward Shils (1981) is distinguished from that of collective memory insofar as tacit knowledge and habitus constitute an unconscious and corporal way of cultural transmission. Nevertheless, corporal transmission of culture analyzed by Bourdieu may supplement the study in sociology of collective memory in the strict sense of the word.

History and Collective Memory

The constructive nature of collective memory pointed out by Halbwachs gives rise to numerous empirical studies in sociology and history. However, according to Halbwachs, collective memory is different from history even if history influences the formation of collective memory, because collective memory holds only a part of the past that is still alive in the conscience of the group. On the contrary, history chooses important events and tries to give these events an order to simplify them for the didactic purpose. Traditional conception of history defined as such by Halbwachs is revised nowadays insofar as Annales School inspired by the Durkheimian school brought a great change to the conception of history. The title of its journal, *Annales d'histoire économique et sociale*, insists on the importance of economic and social history and shows well the nature of the Annales School. Lucian Febvre and Marc Bloch, two founders of the Annales School, declared that historians should find problems from the interests of the present to write a living past. This indicates they shared the same problematic as that of their intimate colleague at Strasbourg University, Halbwachs. In this sense, history comes up to or surfaces in the collective memory. The difference lies in the fact that collective memory is more likely to fade out with the disappearance of those directly concerned.

Structuralism also contributed to the change of perspective by moving away from the modern, straight-line view of history and questioning the origins of phenomena that can be explained by the existence of certain structures. While, by definition, the structures behind the phenomena may lack any historical factor, the fact is that every phenomenon has an origin. Lévi-Strauss (1962) asserted that rituals of commemoration were held to reproduce the past in the present to resurrect mythological origins in the present.

The idea of collective memory introduced by Halbwachs and the Annales School is picked up again from the 1970s. The historian Pierre Nora notes that modern society is losing the sense of memory and needs to actively set up institutions that will allow it (society) to keep its memories. He designates these institutions as *lieux de mémoire* (site of memory) in French, an expression difficult to translate into other languages. Some institutions developed in modern society play an important role in the reconstruction of collective memory. Public institutions like archives, libraries, and museums are thus established since the nineteenth century. Important events are commemorated, anniversaries celebrated, and monuments are constructed to remember these events. Autobiographies also offer *lieux de mémoire* that historians can explore.

The same idea can be found in its embryonic stages in Halbwachs; One of the frameworks that orients that the constitution of memory is spatial (Halbwachs, 1925: 97). Memory is closely related to a particular space, for example, a room, a garden, a corner of a street etc., all of which are familiar and gain an affective value and allow us to remember the past. However, the attempt of the historian has to be distinguished from that of the sociologist; the historian makes historiographies on institutions for remembering; the sociologist, on the contrary, insists on the constructive aspect of the memory. The historian tends to accept historical data as such, while the sociologist deliberately brings a critical point of view, given that what the latter is interested in is not to find new historical facts, but to know how an individual or a group

remembers the past. Each group has its traditions and retains its memories on events and persons that made a strong influence on the group. Without these memories shared by the members of the group, the group could not survive. And if some members of the group are influenced by anyone or anything outside of the group, the group reacts in various ways. The group can accept new elements based on memories of the outside openly or in taking the former elements into account. The group can also reject completely new elements to defend its traditions. Conflicts, thus, can occur concerning memory and the reaction of the group can give rise to a change. The sociology of collective memory is, therefore, an attempt to understand how a group maintains its order and how social change occurs. It is not a simple branch of sociology. It seeks for making a sociological theory in which memory plays an essential role.

Social Order and Social Memory

Halbwachs noticed that the memory of ancestors is alive by means of worship. The worship of ancestors is institutionalized, because the living tries to hold the memory of the dead. The ancestor, indeed, plays an important role for the maintenance of the social order. Contemporary of Halbwachs and Durkeimian like him, Marcel Mauss was very sensitive to this issue. He pointed out that "among the first groups of beings with whom men must have made contacts were the sprits of the dead and the gods." (Mauss, 1969: 13) The primordial exchange is an exchange between the living, the gods, and the spirits of the dead. Before Mauss, Friedrich Nietzche made a similar observation in *Genealogy of Morality*:

> …the living generation always acknowledged a legal obligation towards the earlier generation, and in particular towards the earliest, which founded the tribe (abbrev) …There is a prevailing conviction that the tribe *exists* only because of the sacrifices and deeds of the forefathers, – and that these have to be *paid back* with sacrifices and deeds.
>
> Nietzche, 1994: 65

There is an imbalance between a sense of 'obligation' toward our ancestors, and 'us,' and as a result this imbalance becomes a source of control; by demonstrating that one possesses special powers that enables communication with the gods and the spirits of the dead. The order based on the recollection of the dead is therefore established through a never-ending exchange between the living and the spirits of their ancestors.

This mechanism of maintaining order can be described as "the order of recollection" (Ogino, 2006: 59–61).

The order of recollection is the result of man's greatest concern – how to deal with the death of a loved one. In the order of recollection, there is no clean break between family members who have died and those who are still living. Order is sustained by their memories of the dead (their ancestors) who dwell in a sacred place. The dead are seen as protecting the living, and ancestors can at times symbolically communicate with the living. For example, in Japanese Buddhist tradition, the spirits return to *shigan* (the human world) during the *obon* season and therefore the living carry out rituals to receive their ancestors, such as making daily offerings to the *butsudan* (Buddhist altar).

In the order of recollection there is an attempt to restore order and overcome a crisis, by arriving at an understanding of the cause of the crisis by relating it to gods and spirits. A crisis could be a natural disaster such as a flood, or a deterioration of social order triggered by a crime such as theft. *The Legends of Tono* (2008) and its supplement (1989), gathered by ethnologist Kunio Yanagita, shows memories about the area of Tono and its habitants. Whether they are true or not, they show in detail what the habitants remember, i.e., their collective memory. It contains a typical anecdote of theft that demonstrates how the recollection of order functions.

> In the event of an adversity such as theft, the villagers call on the powers of a god named *mitsuminesama* in order to identify the perpetrator. A man called Yoshitaro Sasaki called on the powers of *mitsuminesama* when *watagase* [ceremonial dress made of cotton] was stolen from his house. All the lights in the house were turned off and the *goshintai* [body of god] placed in the *okuzashiki* [innermost room of the house]. Each of the villagers who had gathered in the house then entered the room to pray to *mitsuminesama*. One woman, however, was terrified and refused to enter the room. When the villagers tried to force her into the room, the woman vomited blood and lost consciousness. The villagers therefore concluded that she was the thief. The woman handed over the stolen goods to the villagers that night.
>
> Yanagita, 1989: 101–102

As is clear from this anecdote, the villagers believed that it was *mitsuminesama* and not they themselves who have the ability to solve crimes of theft. Therefore, a series of rituals were carried out in preparation for this, commencing with the welcoming of the *goshintai* by the villagers. This ritual is clearly only carried out "when there are already several suspects" (Yanagita, 1989: 101) and the villagers already have some idea of who is guilty. It should actually be possible to identify the perpetrator without having to carry out such rituals. However, the act of reconfirming the function of the order of recollection through the ritual of identifying the perpetrator is of more significance to the villagers than simply identifying the perpetrator for the Sasaki family and returning the *watagase* to them. Thus, in the event of an incident, the villagers must ask the spirits of the dead for assistance.

Violence was indirectly exerted in this example of theft by forcing the female suspect to vomit blood through the act of ritual. The community exerts violence directly or indirectly against those who it believes have caused a deterioration of order. In each case, the violence is understood to be promoted by a god such as *mitsuminesama* or a priestess who has the ability to communicate with the gods. The exertion of violence is therefore validated because of a god or deity who inhabits a space found beyond the boundaries of the village, or a priestess who possesses the powers to communicate with the gods. The villagers believe the priestess' oracle to be true because she is a mediator who is able to communicate with the gods and the spirits of the dead. This signifies that the villagers are hierarchizing the world that they live in, including places accessible to any person through to those

only accessible to exceptional people or beings. The ability to demonstrate one's power to communicate with gods becomes a source of control. What we can see from the above is that violence is not the trigger for the establishment of order, but that it is the recollection of order that actuates the device that, in turn, enables the exertion of violence in response to a crisis.

Modern Society and Memory

The order of recollection should be inherent in all community. But the more the society is complex, the more the order of recollection is imperiled since the complex society provides more opportunities to come across unknown people who do not share the same memory or memories. Families and communities tend to lose autonomy and submit to the order of the State, designed to be the only peacemaker of different groups facing conflict situations concerning memory. To play this role, the nation-state has tried to construct a new framework of collective memory which Nora called *lieux de mémoire*.

The perception of collective memory and of its constructed aspect by Nora is not a pure coincidence. In fact Henri-Pierre Jeudy pointed out that from 1980s, industrial societies began to constitute their regional and urban ethnology (Jeudy, 1986: 17). Similarly, in sociology a new method based upon life history or life story was highlighted. This phenomenon common to industrial societies was, in part, due to the accelerated innovation in industry, which inevitably caused the social transformation and the loss of 'old' industry and its value; old industrial sites were recognized as cultural heritage to fill the gap. Therefore, we have entered the age in which we are actively, and as exhaustively as possible, trying to preserve as shared heritage the various experiences of people and the culture they have created. This is signaling the arrival of the "age of preservation."

In this age, what is saved is more than just cultural heritage that is recognized as having aesthetic value, or assets from the distant past such as ancient Egyptian mummies. Things that are preserved include events that invoke abhorrent memories that may still be vivid, and that some people may want to quickly expunge, as well as items connected to these. The A-Bomb Dome is registered on UNESCO's World Heritage List, and can be considered one of these. In addition to monuments or memorials of war, other legacies with negative aspects, including abandoned mines that have resulted in environmental pollution and industrial accidents, and buildings from colonial times can be called 'negative heritage.' These differ from the safeguarding of cultural heritage from ancient periods, for which obtaining political agreement is easy, and can cause controversy and political conflict.

If so, this is related to the ambivalent nature of 'negative memory.' On the one hand, people naturally are subject to the urge to forget unpleasant and unfortunate experiences as soon as possible. Even if people are unable to fully forget these matters, they want to remove themselves from them as much as possible. In the event that they are forced to confront sad memories, they try to keep them bottled up deep inside themselves. Feelings such as these mean that people remain tight-lipped about negative memories, and this is tied to the impulse to expunge all items that could invoke them. The former headquarters of the Governor General of Korea in Seoul was demolished because it evoked memories of the former colonial period. When East and West Germany were unified, the first thing torn down was the Berlin wall although a part of it was preserved.

However, even if this negative memory is of an abhorrent past, this also invokes the desire to record that past, and to express it in different ways. At times, people may feel the desire to talk about the negative experiences that were once so unpleasant to put into words. If this is so, it is because at the core of negative experiences are death and memories of the departed. For instance, the experiences described by survivors of the Hiroshima atomic bomb depict a scene of mass death and suffering as told by those who happened to survive by pure chance. The same goes for those describing the massive destruction of the Great Hanshin-Awaji Earthquake in 1995; houses facing each other across a street suffered differing degrees of damage. This kind of experience serves to make survivors aware that a single step in the wrong direction could have ended in their death rather than survival, and the remembrance of those who died shared by the survivors inspires a will to talk about those experiences – even though they are negative experiences – to pass them on to subsequent generations. In other cases, heritage that no longer exist may be reconstructed, such as in the case of the Urakami Cathedral, a part of the Nagasaki Atomic Bomb Museum. In Korea, the former Governor General's headquarters was demolished, but the Independence Memorial Hall of Korea in Cheonan was constructed, showing dioramas that depict torture carried out by Japanese soldiers.

The question of negative heritage by excellence shows how the late-modern society inclines toward the conservation of the recent past including a negative one that people do not necessarily like to remember. We are forced to keep memory even about events we hope to forget. From the point of view of the sociology of collective memory, this is due to the absence of the order of recollection, which allows the living to mourn for the dead. Through the order of recollection there is an attempt to restore order and overcome crisis. This is achieved by arriving at an understanding of the cause of the crisis through attributing it to gods and spirits. Modern society, in contrast, has lost the smooth functioning of the order of recollection. Faced with a secure future transformed suddenly into uncertainty, no solution is available. In this situation, disaster victims tend to regard their experience as 'absurd.' The loss of the loved ones cannot be accepted. It arouses worry when there are not adequate measures against disasters. The late-modern society seeks to find a pertinent system for mourning by means of commemorations, monuments, museums, and the collection of various stories in vain. The impossibility of mourning would be one of the currents issues in sociology of collective memory.

Postmodern Time and Memory

Philosopher Jean-François Lyotard, who proposed reflections on modernity, threw light on the rise of interests about the

collective memory. According to Lyotard (1988), modernity defines the series of events grouped in the name of the universal history of mankind. He holds that modernity must be formed under the philosophy of liberation, i.e., the perception that mankind is being liberated. Of course this philosophy of liberation – "the Grand Narrative" – can be framed in many ways; liberation from exploitation and alienation under Marxism, liberation from poverty by technology and industry under capitalism, and redemption through the love of Christ, are all modern reiterations of the philosophy of liberation.

However, according to Lyotard, the predominance of western European thinking in modern philosophy, which has been dominant for two centuries, is beginning to crumble. This was triggered by Auschwitz. If history was heading toward the liberation of mankind then modern philosophy has no answer for the question of how it was that so many Jews were allowed to be massacred. Thus, while the perception that modern philosophy is powerless becomes more deeply rooted, the hidden underbelly of modern philosophy is being gradually exposed. This two-faced nature of modernity is one of the defining traits of postmodernism. As a result, postmodernism (as has been pointed out many times) does not refer to a new zeitgeist that follows on from modernism; indeed, the very idea is that a new zeitgeist must follow on from modernism is a modernist view of history. Seeing postmodernism as simply a zeitgeist is to deprive the concept of its defining characteristic.

How, then, should we regard postmodernism? Using the example of a modernist painting, Lyotard (1988) said that a painting can only be modern by first being postmodern. For instance, Cezanne sought to create a new art movement by rejecting impressionism. That could be said to be a postmodern move because it came after impressionism, a modern movement. However, for artists like Braque and Picasso, who came after Cezanne, Cezanne was already modern. So, whether it be Cezanne or Picasso or even Duchamp, who came later still, they first had to arrive on the scene as postmodern artists and remain faithful to the modernist philosophy of "questioning everything that came before" in order to be perceived as modern artists. If this is so, then their elevation as modernist masters was consequential rather than intentional, given that, according to Lyotard (1988: 28) "postmodernism is not the end of modernism but the process by which modernism is born."

The interesting thing about Lyotard's postmodernism theory is that whereas modernism gives meaning to the future by defining history as the process of human liberation, postmodernity has no such preassigned standards and is itself the process of creating new standards. Postmodern works therefore do not seek to be evaluated in accordance with preprepared standards; they are a quest for new standards of value. In that sense, the very act of unveiling a new piece of work takes on the appearance of an event in itself; postmodernism is the tendency toward the chaotic and lies beyond the bounds of established standards.

Seeking new standards does not mean making light of the modernism that gave rise to the established standards. Rather, it is the process of questioning the nature of modernity. It could also be said that the very stance of questioning the nature of modernity is postmodern. Based on a positive image of the future in which it is assumed that people will eventually be liberated, postmodernism abandons the sequencing of discrete events and elements under a straight-line historical perspective. Modern awareness positions the past as a 'passé' age that has been surpassed by the present. In other words, whereas, in the modernist view, the past is defined by the present and the present is defined by the future, postmodernism dismantles the historical view and seeks meaning or to express the very process by which meaning is generated. According to Lyotard, it is similar to psychoanalysis. In this treatment, the analyst seeks to understand the sequence of memories that come flooding back without conscious manipulation, as a means of identifying the root cause of the disorder, the memories that the patient has tried so hard to repress. Postmodernism could be said to have a certain historical awareness and, just like the way images from the past come flooding back in psychoanalysis, postmodernism is the fixation on the memories revived when modernist historical awareness is dismantled.

In modernism, there are two perceptions of time. One is that of modern philosophy, in which history is seen as the process of the liberation of mankind. The second is the postmodern view, which moves away from that kind of past-present-future, straight-line perception of time. In the postmodern view, the past exists only as memories. The important thing here is that for modernity to exist, it requires not just the modernist historical view, which draws the schematics of human liberation, but also the postmodern perspective, which questions those plans. The postmodern perception of time does not, however, oppose the modern view of history. Modernity, after all, has both modern and postmodern aspects and would not exist without both of them (Ogino, 2013).

Halbwachs launched the sociology of collective memory as a critical approach to modern society and the overall view of history on which modernity is founded at the age where the modern notion of history was dominated. This challenge of Halbwachs still remains current at the age of preservation where even negative memory becomes an object of preservation and war heritage is considered as 'cultural' heritage and at the age of globalization politics of memory occupy an important or even primordial place.

See also: Collective Behavior, Sociology of; Collective Beliefs, Sociological Explanation of; Collective Memory, Psychology of; Culture, Sociology of; Death and Dying, Sociology of; Identity Movements; Postmodernism in Sociology; Race Identity; Sociological Traditions; Tradition: Social.

Bibliography

Adam, B., 1995. Timewatch. Polity Press, Cambridge.
Bourdieu, P., 1980. Le Sens pratique. Les Éditions de Minuit, Paris.
Durkheim, E., 1937. Les Règles de la méthode sociologique. P.U.F., Paris.
Halbwachs, M., 1925. Les Cadres sociaux de la mémoire. Alcan, Paris.
Halbwachs, M., 1971. La Topographie légendaire des évangiles en terre sainte, second ed. P.U.F., Paris.
Jeudy, H.-P., 1986. Mémoires du social. P.U.F., Paris.
Lévi-Strauss, C., 1962. La Pensée sauvage. Plon, Paris.
Lyotard, J.-F., 1988. Le Postmoderne expliqué aux enfants. Galilée, Paris.
Mauss, M., 1969. The Gift: Forms and Functions of Exchange in Archaic Societies. Cohen & West, London.

Mead, G.-H., 1932. The Philosophy of the Present. The University of Chicago Press, Chicago.
Nietzche, F., 1994. On the Geneaology of Morality. Cambridge University Press, Cambridge.
Nora, P. (Ed.), 1984. Les Lieux de mémoire I La République. Gallimard, Paris.
Ogino, M., 2006. Scams and Sweeteners: Sociology of Fraud. Trans Pacific Press, Melboune.
Ogino, M., 2013. Structuralism, post-structuralism and thereafter. In: Elliott, A., et al. (Eds.), Japanese Social Theory: From Individualization to Globalization in Japan Today. Routledge, New York.
Olick, J.-K., 2007. The Politics of Regret. Routledge, New York.
Polanyi, M., 1966. The Tacit Dimension. Routeledge & Kegan Paul Ltd., London.
Shils, E., 1981. Tradition. The University of Chicago Press, Chicago.
Schutz, A., Luckmann, T., 1973. The Structure of the Social World. Northwestern University Press, Evanston.
Yanagita, K., 2008. The Legends of Tono. Lexington Books, Lexington.
Yanagita, K., 1989. Tono monogatari shûi, Tono monogatari. Kadokawa, Tokyo.

Collectivism and Individualism: Cultural and Psychological Concerns

Harry C Triandis, University of Illinois, Champaign, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Collectivism is a cultural pattern found in most traditional societies, especially in Asia, Latin America, and Africa. It contrasts with individualism, which is a cultural pattern found mostly in Western Europe, North America, Australia, and New Zealand. The dimension of collectivism–individualism is also a psychological one that has been examined as a personal and social value. This article describes some of the characteristics of this cultural pattern, its geographic and demographic distribution, and its antecedents and consequences. It also describes collectivism as a psychological phenomenon. It ends with an examination of the implications of globalization and economic development for changes in this cultural pattern, which may involve a weakening of some of its aspects. Especially important are the implications of these changes for well-being and social control. An evaluation of this cultural pattern depends on the criteria that one wishes to maximize.

Introduction

Collectivism is a cultural pattern found especially in East Asia, Latin America, and Africa. It is usually contrasted with individualism found in the West, e.g., in Western and Northern Europe, the USA, Canada, Australia, and New Zealand. Any typology is an oversimplification. Societies are not purely individualist or collectivist, but some mixture of the two. The construct is widely used by those interested in studying cultural differences. It is also used to describe the values of individuals in these differing societies.

Among the most important characteristics of this pattern are the following individual orientations:

1. Individuals define themselves as aspects of a collective, interdependent with some ingroup, such as the family, tribe, coworkers, nation, or religious, political, ideological, economic, or aesthetic group.
2. They give priority to the goals of that collective rather than to their personal goals.
3. Their behavior is determined more often by the norms, roles, and goals of their collective than by their personal attitudes, perceived rights, or likes and dislikes.
4. They stay in their groups and prefer these relationships even when the costs of staying in these relationships exceed the advantages of remaining.

These are some of the defining attributes of collectivism. There are as many kinds of collectivism as there are collectivist cultures. To distinguish among collectivist cultures, additional attributes are necessary (Triandis, 1995).

History

The term was first used in political philosophy in the eighteenth century in contrast to the individualism advocated by John Locke and Adam Smith. An early exponent was Jean-Jacques Rousseau, who in his *Social Contract* (1762) advocated submission to the 'general will,' which was conceived as the common core of societal opinion. Georg Fredrick Hegel and Karl Marx also proposed collectivist philosophic systems.

The term was introduced into psychology by Hofstede (1980) in a study of the values of IBM employees in some 80 countries. He found that by summing the responses to value items of individuals from each country, and using the country as the unit of analysis, it was possible to identify a cluster of responses that reflected collectivism and individualism. Triandis (1995) summarized the literature, which used that construct in social psychology, and Kagitcibasi (1997) provided a critical evaluation of the construct. Oyserman et al. (2002) also expressed some concerns about the degree to which this construct has been used to characterize whole societies, without taking variation within societies into account. Triandis and Gelfand (2012) focused on the theory of collectivism.

Historical analysis suggested that collectivism was the typical cultural pattern in many traditional societies (mainly agricultural and pastoral ones, but less so in hunting and gathering societies). Individualism began emerging in England as early as in the thirteenth century. Traces of individualism have also been found in ancient Greece.

Distribution

Within any society there are individuals who behave like persons in collectivist cultures. They are called *allocentrics*, and contrast with those who behave like persons in individualist cultures who are called *idiocentrics*. It is assumed that there are both allocentrics and idiocentrics in every society, but their distributions are different, with more allocentrics found in collectivist cultures. The distinction between collectivism and allocentrism is important because in some cases a phenomenon observed at the cultural level is quite different from the phenomenon at the individual level of analysis.

Collectivism is maximal in relatively homogeneous societies, such as theocracies and monasteries, while individualism is maximal in heterogeneous societies that are very affluent.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.24008-7

Thus, there will be few idiocentrics in monasteries and few allocentrics among Hollywood stars.

All individuals have access to cognitive systems that include both allocentric and idiocentric cognitions but they sample them with different probabilities, depending on the situation. For example, if the ingroup is under attack most individuals become allocentric. In the company of other allocentrics the norms for allocentric behavior become salient and individuals are more likely to sample allocentric cognitions. Some situations provide very clear norms about appropriate behavior (e.g., in a house of worship), while other situations do not (e.g., at a party). Individuals will be more allocentric in the former than in the latter situations. When the ingroup can supervise an individual's behavior, norms are more likely to be observed, and the individual will be more allocentric. When individuals are asked to think for 2 min about what they have in common with their family and friends they are likely to sample more allocentric cognitions. Situations are more important than the personality of individuals as determinants of allocentric behaviors.

Collectivism can appear in all or none of the domains of social life. For example, it can be found in politics, religion, aesthetics, social life, economics, or philosophy.

Following is an outline of findings in the empirical literature that describes the antecedents and consequents of collectivism, as characterized by Triandis and Gelfand (2012).

Antecedents

Homogeneity of the Collective and the Culture

One antecedent of collectivism is the homogeneity of the collective. The more homogeneity the more collectivism. By contrast when members of the collective differ on their attitudes and values or do not agree concerning what are the group's norms of behavior, it is left up to the individual to decide which normative system to follow, and thus there is individualism.

Interdependence Imposed by the Situation

Collectivism is higher when members of the collective are interdependent, for example when they engage in activities that must be coordinated and when what one does affects the success of others, or where cooperation is essential for success. When the goal is to execute large projects such as irrigation canals, or defensive walls, collectivism is more likely. In societies where people are financially interdependent, collectivism is high.

Population Density

In dense social environments many rules develop that are designed to reduce conflict and ensure the smooth functioning of the group. People make sure that the rules are observed. When group norms are more important than attitudes as determinants of social behavior, collectivism is high. Urban environments often include members of different groups, and individuals can decide for themselves whether to observe the norms of one or another group, thus they become idiocentric. On the other hand, an urban setting may have neighborhoods of culturally homogeneous individuals, which increases the collectivism of that group.

Social Stability

When the collective is stable, it is more likely to develop agreements about norms, and to make sure that the norms are observed. By contrast, individuals who have migrated from their societies or have experienced social mobility are more likely to be idiocentric. When social change is slow, the norms can be observed more reliably than when social change is rapid. There is evidence that the older members of all societies are more allocentric than the younger members.

Low Economic Status

Since the poor have fewer choices in life, and tend to do what norms specify, they tend to be more conforming to social norms than members of the upper classes. When resources are limited, one often depends on group members for assistance, especially in emergencies. These factors increase collectivism. On the other hand, those in positions of leadership tend to be idiocentric.

One Large Ingroup Rather than Many Ingroups

Those who only have one ingroup can channel all their energy into that group. Also, they cannot afford to develop poor relationships with members of that group, so they are more likely to observe its norms. Those who have many ingroups have superficial relationships with most of them, and if they do not have good relationships with one of these ingroups they drop it and spend more of their energy with some other ingroup. Intense involvement in relatively large ingroups (extended family, religious, or national ingroup) is more common among collectivists than individualists. Intensive involvement in individualist cultures is usually limited to first-degree family members.

Low Exposure to Other Cultures

People who only know one culture are unlikely to have more than one set of norms for social behavior. They tend to be maximally ethnocentric, authoritarian, and submissive to ingroup authorities. Those who are more educated, traveled, and have lived with more than one cultural group develop idiocentric tendencies.

Salient Common Fate

Common fate with members of the ingroup (e.g., the ingroup is under attack) increases collectivism. Time pressure for decisions has similar consequences. When there is time pressure an ingroup authority must make quick decisions that do not allow time for discussion and that places most members of the ingroup in a condition of common fate, as well as requiring conformity to these decisions.

Collective Mobility Is Advantageous

If individual upward social mobility is not possible, then collective mobility may be used. Thus, individuals invest their

energy in promoting the status of their ingroup. For example, people who do not get promoted on their own might join a union and invest their energy in promoting the union.

Large Families

In large families it is not practical to allow each child to follow idiosyncratic schedules, or have much privacy. Many rules are enforced, and that creates collectivism.

Strict Religious Upbringing

Most religions use a large number of norms that are reliably observed. That increases collectivism.

Little Exposure to Modern Mass Media

Analyses of the content of television programs made in the USA show that the themes used are often highly individualistic (emphasis on pleasure, doing what the individual wishes to do even if that is inconsistent with the wishes of groups or authorities). Countries where people have little exposure to such mass media tend to be more collectivist.

Consequences

Socialization

In collectivist societies, children are usually socialized to be obedient, reliable, to follow traditions, do their duties, sacrifice for the ingroup, and be conforming. In contrast, self-reliance, self-control, and independence are usually emphasized in individualist societies (Berry, 1979; Bond and Smith, 1996),

Low Well-Being

While members of all cultures are satisfied with their life (Tov and Diener, 2007), members of collectivist cultures are less satisfied than members of individualist cultures. Well-being tends to be correlated with individualism (Diener et al., 1995). Even when controlling statistically for income, this relationship persists.

Consequences for Cognition, Emotion, and Motivation

Markus and Kitayama (1991) have reviewed an extensive literature that shows that members of collectivist cultures differ from members of individualist cultures in their cognition, emotion, and motivation.

Social Behavior

Collectivists are extremely concerned with saving their face, and even attempt to save the face of the persons with whom they are interacting. Preserving harmony within the ingroup and being pleasant in relationships with ingroup members are often important values. On the other hand, collectivists do not care much about their relationships with outgroups. They usually behave quite differently toward ingroup and outgroup members, showing sacrifice and extreme cooperation with ingroup members, and suspicion and hostility toward outgroup members.

Receive Much Social Support if They Conform

Allocentrics report that they receive more social support and a better quality of social support than idiocentrics. However, those who do not conform to ingroup norms are rejected, and may even be killed (Triandis, 1995).

Have Poor Skills in Entering New Groups

Members of collectivist cultures do not have good skills for entering new groups because they are not socialized to do this. They are comfortable as members of their ingroups, thus they do not need to enter other groups. On the other hand they have more intimate relationships with those whom they interact with than idiocentrics who tend to have superficial relationships (e.g., cocktail parties).

Are Easy to Mobilize for Confrontations with Outgroups

When ingroups are in conflict it is easier to arouse members of collectivist cultures to fight for the ingroup than it is to do so with members of individualist cultures.

Describe the Self by Using Roles, Groups

When asked to complete 20 statements that begin with 'I am …,' members of collectivist cultures provide sentence completions that have more social content (I am a brother, I am a member of the communist party) than members of individualist cultures.

Attributions

Allocentrics make external attributions in explaining events (e.g., norms, group pressure), while idiocentrics show the opposite pattern (e.g., personality, attitudes). Thus, idiocentrics make the fundamental attribution error (seeing internal more often than external causes for the behavior of other people, while seeing external causes more often than internal causes for themselves) more frequently than allocentrics. Also, allocentrics attribute their successes to the help they have received from others and their failures to personal shortcomings (e.g., I did not try hard enough).

Have the Ideal of Fitting-In Rather than Shining

Allocentrics prefer to fit in and be like most others in their ingroup rather than shine and sticking out. Idiocentrics tend to be high in self-enhancement (report that they are better than most other people on most desirable attributes). Allocentrics tend to be modest when they compare themselves to others.

Show Little Social Loafing

Allocentrics show little evidence of social loafing (the tendency to produce less when they work in a group than when working alone, if their output is not identifiable). They

usually produce as much as possible when they work with ingroup members, but they show social loafing if their coworkers are outgroup members.

Within the Ingroup Distribute Resources Equally or According to Need

When interacting with ingroup members, members of collectivist cultures tend to distribute resources equally or according to need. When interacting with outgroup members, however, they use the equity principle (to each according to contribution).

Recreation Occurs in Groups

Members of collectivist cultures seek recreation in groups rather than alone to a greater extent than members of individualist cultures (e.g., bowling rather than skiing).

Communication

During interpersonal communication, members of collectivist cultures rely on the *context* (such as distance between and position of bodies, eye contact, gestures, tone of voice) more than on the *content* of communications. For example, rather than saying 'no' they may serve incongruous foods, to put across their negative message without causing the other person to lose face. When they describe a person they are likely to use context, e.g., 'he is intelligent in the market place,' rather than decontextualized statements such as 'he is intelligent.' Silence is used more than in individualist cultures. Compliments are not used as frequently as in individualist cultures. The languages of collectivist cultures do not require the use of 'I.' Collectivists use more action verbs (e.g., he asked for help) rather than state verbs (e.g., he is helpful), which suggests that they use communications that make greater use of context, are more concrete, and place less emphasis on internal attributes of the person.

Conflict Resolution

Members of collectivist cultures avoid conflict with ingroup members, and in the case of disagreement prefer silence to argument. They prefer mediation to confrontation.

Conformity

Members of collectivist cultures show more conformity in experiments that measure conformity.

Emotions

Members of collectivist cultures get angry when something unpleasant happens to ingroup members more often than when something unpleasant happens to themselves. They do not express negative emotions as often as members of individualist cultures. Collectivists often have emotions that reflect relationships (e.g., sympathy), while individualists often express emotions that reflect independence (e.g., pride).

Apologize More for Transgressions of the Ingroup

Members of collectivist cultures apologize more when an ingroup member commits a crime than is common in individualist cultures.

Globalization and Collectivism

In the process of globalization, is collectivism to be avoided and individualism to be promoted? This issue is especially relevant as societies become more affluent, because they tend to acquire some of the attributes of the West, including individualism. The group is rejected, and the individual becomes the center of social change. People in cities, in small apartments that do not have room for extended families, thus drift toward individualism.

Is 'economic development' necessarily associated with more individualism and less collectivism? My view is that the problem is much more complex than just stated. First, one can be both modern (use computers and fax machines) and collectivist. Second, societies are complex, i.e., they have many aspects. Some aspects might change and become more Western, while most aspects may stay the way they were. In general superficial attributes, dress, public behavior, and economic life might change while fundamental values may not.

In any case, for collectivist societies, one might ask: is it desirable to 'promote' individualism, or better to stay with collectivism? As is the case with most complex issues, the answer is 'it depends.' A major problem can be the extremity of collectivism and individualism. Extreme collectivism is stifling. One becomes a tool manipulated by others, unable to reach personal goals. Extreme individualism is isolating, and leads to ego-centered narcissism. Furthermore, extremely individualistic societies are often extremely competitive, and that results in remarkably high levels of stress. Escapes from stress may include drug abuse.

It is possible to list both positive and negative factors associated with each cultural pattern. In the case of individualism we find affluence, self-determination, and self-actualization. These factors are associated with feeling good about oneself, and being satisfied with life and being happy. Also, in individualist cultures there is an emphasis on individual rights, and people are expected to shape the direction of societal change to fit their personal needs.

On the other hand, Naroll (1983) has assembled much empirical information suggesting that the weak interpersonal bonds that are associated with individualism result in higher rates of civil disturbance, riots, murder, delinquency, theft and fraud, drug and alcohol abuse, suicide, brawling, wife beating, child battering, and mental illness. The competition found in individualist cultures often results in some segments of the society remaining economically behind and becoming especially susceptible to social pathologies.

By contrast, collectivism is associated with social controls, including extreme punishment for crimes, prompt punishment, and active gossip. Social control is also achieved through emotional closeness, social support, appeals to memorable myths, and a plausible ideology that supports the ingroup's

norms. People feel embedded in the social system, so they are less likely to try to change it, to commit crimes or suicide, to steal from the ingroup, or use drugs that isolate them from it. One of the best predictors of drug abuse is the combination of high pressure for achievement with low social support, a combination often found in individualist cultures.

On the negative side, collectivism is often found in societies where groups assume that the individual is relatively unimportant and must do what is best for the group. This is especially significant when the group is the state. People are expected to be 'patriotic,' and sacrifice themselves for the good of the state. However, very often, what is good for the state is in fact only good for those in power, who frequently take advantage of the masses to further their personal goals. As we have seen, leaders in all cultures are idiocentric, and the history of the world offers too many examples where they order their followers to do what is good for the leader rather than what is good for the followers. Also, collectivism makes it easier to convince the ingroup to fight outgroups, and wars are more frequent between collectivist countries than between individualist cultures. In a world with nuclear weapons, there may be much to be said in favor of reduced collectivism.

Note also that while the social pathology associated with individualism is formidable, it does not affect most individuals in a society. Pathology by definition is found among a small minority in any society. Thus, the majority can enjoy the benefits of affluence without the negative consequences of individualism.

In short, an evaluation of the desirability of collectivism or individualism depends on what criteria we wish to maximize. If we wish to emphasize the creativity that individualism allows – which often results in more affluence, the freedom and self-actualization of individuals, democratic processes, and the independence of individuals from the social pressures of groups – we may wish to see globalization leading to more individualism. But if we wish to emphasize family values and the values of the ingroup, we may resist the weakening of collectivism associated with affluence. These values include religious beliefs, the social support that one can get from the ingroup in moments of crisis, and the security, both material and emotional, that the ingroup can provide.

In addition to the above considerations, there are issues of individual differences. For those who are intelligent, educated, easily changed, and adaptable, an individualistic society can offer a great deal. Creativity requires freedom to pursue different courses of action and in a society where one is supposed to consult others, and act according to well-established customs, creativity is not likely to flourish. On the other hand, for those who value especially family life, harmony, security, and social stability, a collectivist society may be optimal.

See also: Acculturation; Collective Action; Collective Behavior, Social Psychology of; Collective Memory, Social Psychology of; Cross-Cultural Psychology; Cultural Influences on Interpersonal Relationships; Cultural Psychology; Culture and Emotion; Culture and the Self: Implications for Psychological Theory; Indigenous Psychology; Individualism versus Collectivism: Philosophical Aspects; Levels of Analysis in Social Psychology; Values, Psychology of.

Bibliography

Berry, J.W., 1979. A cultural ecology of social behaviour. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. XII. Academic Press, New York, pp. 177–206.

Bond, R., Smith, P.B., 1996. Culture and conformity: a meta-analysis of studies using Asch's line judgment task. Psychological Bulletin 119, 111–137.

Diener, E., Diener, M., Diener, C., 1995. Factors predicting the subjective well-being of nations. Journal of Personality and Social Psychology 69, 851–864.

Hofstede, G., 1980. Culture's Consequences. Sage, Beverly Hills, CA.

Kagitcibasi, C., 1997. Individualism and collectivism. In: Berry, J.W., Segall, M.H., Kagitcibasi, C. (Eds.), Handbook of Cross-cultural Psychology, second ed. Allyn & Bacon, Boston, pp. 1–50.

Markus, H., Kitayama, S., 1991. Culture and self: implications for cognition, emotion, and motivation. Psychological Review 98, 224–253.

Naroll, R., 1983. The Moral Order. Sage, Beverly Hills, CA.

Oyserman, D., Coon, H.M., Kemmelmeier, M., 2002. Rethinking individualism and collectivism: evaluation of theoretical assumptions and meta-analyses. Psychological Bulletin 128, 3–72.

Tov, W., Diener, E., 2007. Culture and subjective well-being. In: Kitayama, S., Cohen, D. (Eds.), Handbook of Cultural Psychology. The Guilford Press, New York, pp. 691–713.

Triandis, H.C., 1995. Individualism and Collectivism. Westview Press, Boulder, CO.

Triandis, H.C., Gelfand, M., 2012. A theory of individualism and collectivism. In: Van Lange, P.A.M., Kruglanski, A.W., Higgins, E.T. (Eds.), Handbook of Social Psychological Theories, vol. 2. CA Sage, Los Angeles, pp. 498–520.

Collins, Randall (1941–)

Juan M Iranzo, Public University of Navarre, Pamplona, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Randall Collins is a leading social theorist in the late-twentieth/early twentieth century. Born into a diplomatic family, he soon gained an equanimous view of, and an interest in both global conflict and the dynamics of social appearances. After studying with Parsons, Blumer, and Goffman, he made a double breakthrough simultaneously reshaping conflict sociology, and detecting credentialism. He has developed and integrated a microsociology of interaction ritual chains, a macrohistorical sociology focused on market and geopolitics effects on state dynamics, and a mesosociology of networks – illustrated by his overarching historical-sociology study of philosophers' networks. Presently he investigates violence at all three levels.

Introduction

Randall Collins (Knoxville, Tennessee, 1941) is one of the most insightful, innovative, and prolific social theorists in sociology at the beginning of the twenty-first century. Together with more than 300 papers delivered in Universities, academic conferences, and professional meetings all over the world, and almost 200 articles and book chapters, he is author of 13 academic books (plus a novel), and co-author, editor, or co-editor of another 5, with 14 of them translated to different European and Asian languages. He has contributed fundamentally to several research fields, particularly conflict theory, the sociology of education, the sociology of the intellectuals, microsociology, the macrohistorical sociology of political, economic, and cultural change, and the sociology of violence. His constructivist view of social theory as a logically coherent body of ideas, which explains the variations of social phenomena in terms of general principles inferred from comparisons among rigorously obtained empirical data: his idea of the epistemic integration of the micro- and macrosocial realms through the mesolevel of networks; his theory of interaction ritual chains; his geostrategic view of world history; his notion of democracy (defined as collegial government plus the extent of franchise); and concepts as credential inflation, rapid-discovery science, emotional energy, stratified attention space, confrontational tension/fear or forward panic have significantly enriched contemporary social theory.

Early Years

Collins' family background, early life experiences, and university years deeply influenced his subsequent intellectual and research career. His maternal grandfather, a German immigrant in the 1890s, was a noncommissioned officer in the US army for 30 years. His father, originally a professor of German before joining the US Army intelligence, and then the State Department, was trained in Russian during World War II and sent to Berlin in 1946, and later, to Moscow (1949–51). As a diplomat there, he was in the political section of the embassy – his mother performing heavy duties as a 'diplomatic hostess.' That is why some of Collins' first nonfamily memories are about a bombed-out city, unexploded shells, unburied dead soldiers' bodies, and Russian children throwing stones against their family dacha after the outbreak of the Korean War. As a child, he reckoned those as normal facts of his environment, not as fearful or cruel ones. That gave him soon a clear sense of violence, or at least conflict as a common feature of life.

Similarly, witnessing the military often scarce interest in fighting – their main interest is maintaining the routine–, his mother's recurrent need to retire temporally to a hotel to rest and read novels (because of what he would later call 'interaction fatigue'), and his father 'decoding' activities (he became what was then called a kremlinologist), and by his sheer acceptance of all of it as familiar facts, and not necessarily morally labeling conflict as disgrace or strategic behavior as hypocrisy, prepared him for the functionalist view of conflict as a sometimes change-triggering but also sometimes homeostatic event, and for the Goffmanian idea of a front stage/backstage partition of social life. Although his father rarely spoke out his own personal opinions, his performance in his official position during a cold war marked by the arms race, and the possibility of a nuclear war gave Collins a strong consciousness of international and military actions, and a firm affinity with the peace movement, that would lay the foundations of his life-long interest in, and his geostrategic understanding of macrosociology.

A University Student in the Sixties: Harvard, Stanford, Berkeley

After living his earlier years mostly overseas, his family sent him to a New England preparatory school where he was trained to attend Harvard College – and his contact with upper-class students there made him for the first time acutely aware of class differences. As an undergraduate there Collins briefly tried studies in engineering, mathematics, and literature, before opting for social relations – a blend of sociology, social psychology, and anthropology. He soon felt attracted by research, and he started investigating on Piaget's ideas looking for a way to synthesize them with Freudian theories – then strongly connected with the fashionable philosophy of existentialism through Sartre's attempts to build an existential psychology by integrating and expanding out some of Freud (although Camus' *The Rebel* was personally a more inspiring reading). At the same time he shared the recent and growing

interest in the culturally and historically oriented work of Luria, Vygostki, and their circle – which he felt had strong parallels with the sociological view on language and thinking – as introduced to Harvard students by cognitive, language, and educational psychologist Jerome Bruner.

Collins also attended Talcott Parsons classes at a time when he was brilliantly lecturing about Durkheim and Weber's theories of religion, and also trying to integrate Freudianism within his scheme of the social system as the set of mechanisms of the personality system that would induce the individual to interiorize the cultural system, thus providing the basis for the structuring of society – an early attempt at what Collins would later call 'micro-macro integration.' Another Harvard teacher was then George Homans, who since the 1950s was stating, against Parsons, that society is an epiphenomenon of individual actions (at their best chosen by rational calculus) and that sociological explanation can be reduced to behaviorist psychology – ideas that would form the theoretical nucleus of Exchange Theory and then Rational Choice Theory. Later in his career Collins would question the empirical base and the explanatory scope of these paradigms but then the relevance for him of this coexistence in the same university department of such diametrically opposed views was that it gave him a taste of how science flourish in – actually, thanks to – a milieu of theoretical contest and intellectual rivalry, a key concept in his theory of intellectual networks.

Having obtained a BA in psychology in 1963 Collins moved to Stanford University, where he took an MA in psychology that he expected would correct what he felt as Harvard's somewhat narrow theoretical emphasis through Stanford's insistence on laboratory and experimental type studies – as, among others, Leon Festinger, who had recently formulated his theories on social comparison and cognitive dissonance, was promoting there. Nonetheless, Collins had already decided to shift to sociology. Thus, he moved to Berkeley where he would be awarded an MA (1965) and a PhD (1969) in sociology. Berkeley Sociology Department was then intensely attractive because it was a crossroads of paradigms and a hub of professors of fascinating scientific height and engaging personal enthusiasm – what Collins would later call 'energy stars': Joseph Ben-David, Reinhard Bendix, Herbert Blumer, Erving Goffman, Seymour Martin Lipset, or Neil Smelser, among others.

Collins had co-authored just one psychology article (published in *Nature*, though) before becoming Ben-David's research assistant. Usually identified as a member of Robert K. Merton functionalist school, this sociologist of science and expert on the history of university systems and their transformations was at that time visiting Professor in Berkeley, and he was doing research for a book on the world history of the role of scientist. With him Collins did a network analysis of the origins of experimental psychology (mainly in nineteenth-century Germany), and derivatively grasped that the philosophers who led the Idealist movement (Kant, Fichte, Hegel, Schelling…) were connected in a network, and were involved in the struggle to reform their received university system into a system based on the research university. This investigation, published in two articles co-authored with Ben-David, showed him that both the networks of vertical relationships from masters to pupils, and horizontal among these, in intellectual fields, and the material-organizational base of intellectual work strongly affect intellectual production. This was the first seed of what would eventually become *The Sociology of Philosophies*, a book on which Collins would be working the following three decades. His association with Ben-David also reinforced his interest (already nurtured by Parsons) in wide historical comparisons as a means to uncover macropatterns – something decisive for his following endeavor, in collaboration with Reinhard Bendix.

Bendix had exiled in his early twenties from Nazi Germany and attended the University of Chicago, where he took a BA and a PhD in sociology. He first got acknowledged for his work on social mobility with ex-Marxist Seymour Lipset, and shortly after he set a strong Weberian position with his book *Max Weber: An Intellectual Portrait* (1960). He enrolled Collins into a research project that led to the publication of a co-edited reader: *State and Society* (1968). Collins' contribution to this issue, "A Comparative Approach to Political Sociology," where he explicitly adopted an anti-Parsonian stance (then a brave move), became his first influential text. Both Parsons and Bendix followed Weber, the former emphasizing his cultural side, the latter his political side; and perhaps it was the accident of having studied with both of them that enabled Collins to devise a more multidimensional view, one that took into account both stratification and conflict as crucial social phenomena.

In Collins' view, Weber had focused as no other sociologist on the state, and the collective struggles to monopolize the organization of force and the means of generating legitimacy, as key social phenomena. In addition, he had asserted an enduring tenet of sociology: that the three basic dimensions of human social structures are those concerned with material/economic things (class), communities of cultural membership (status), and organizational social control (power: including both politics and violence). Defying both functionalist and Marxian approaches to political history, Collins multidimensional perspective offered a more balanced analysis of those dimensions, and their mutual interactions. His view was later shared and developed by those scholars that created the military-fiscal theory of the upcoming and expansion of the modern state, its most important revolutions and other geopolitical dynamics of historical change, and also the field of social movements studies.

At the other wing of the antifunctionalist front were those professors whose common denominator was their interest in the close analysis of everyday life. In Berkeley, their most respected and authoritative figures was Herbert Blumer, then finishing his classic book *Symbolic Interactionism: Perspective and Method* (1969). Apparently a strong individualist (already famous for his attacks on statistics and 'the variable' – because variables actually do nothing – and for asking his students: "Where is social class? Where do you see it?"), and an old-fashioned gentleman who neither took part nor encouraged student mobilizations and demonstrations, he was nonetheless one of the few prominent members of the university staff who supported their movement and even gave them a theoretical rationale for it. An inspiring teacher, he was widely admired because in the theory course required for all graduate students that he taught he set up a theoretical link with campus-based civil rights, antiwar, and prouniversity reform campaigns,

giving them intellectual tools to understand what was happening: he stressed that the social order is always the result of all the groups of people who mobilize themselves to define what social institutions are; that an organization exists only to the extent that people act it out – when they stop interpreting it as existing it stops existing; or that power is a social construct and it can change very rapidly if a collectively sustainable reinterpretation occurs.

Blumer had been G.H. Mead's teaching assistant and took over his class when Mead died in 1933. He deeply identified himself with Mead and he often lectured as if he was just reporting Mead's ideas, but he actually created much of symbolic interactionism: he thought of its name, designed empirical research strategies for it, and refined its concepts, substantially emphasizing actors' definition of their situation, and strongly stressing the possibility of reinterpreting the situation, thus allowing for the relative volatility to social life. These are the roots of Collins' radical microsociology, and of his ambition to find macrosocial dynamics in microinteractional situations; to show how, through organizational and/or communicational networks, both poles of social dynamics influence each other, and their middle ground.

Nevertheless, the star of interaction studies was Erving Goffman. Twenty years younger than Blumer and then at the height of its theoretical creativity, Goffman's lectures were always crowded, not because he was a charismatic personality but because of the quiet but extreme intensity with which he expounded his empirical observations and his concepts: front stage and backstage, face work, total institutions, interaction rituals, frames, etc. They were an eye-opener experience: almost everything people were doing was socially constructed; it was all social rituals, and they all operated (more or less unconsciously) to enact a certain kind of social order – which, most students thought (Collins among them), just the same could also be torn down. Personally, though, Goffman was much more difficult to get along with. He did not allow many people to be personally close to him, and he often played 'one-upmanship' on other people – trying to embarrass them while maintaining one's own demeanor. Collins thus decided to keep politely distant from him and confine their connections to intellectual matters, where they shared more affinities.

Goffman had also started as a Freudian, but he gathered that conscious and unconscious processes could be empirically better investigated by looking at them anywhere in the everyday life of children and adults – if the researcher had enough mental discipline and observational insight to grasp the details of people's behavior, and sufficient acumen to interpret them theoretically. Feeling himself apt for the task, Collins changed allegiances from Freudianism to micro-interaction analysis. Also, Goffman disliked to be tagged a symbolic interactionist as, having studied in Chicago, he often was; but, because of his knowledge of Parsons, Collins was able to recognize his Durkheimian side, thus interpreting his theory in the way he preferred – although he said Collins was turning it into Marxism because he focused on how social rituals change historically, and how they apply to conflict and violence, among other themes alien to Goffman interests.

Eventually, their relationship became strained. Goffman upheld that society exists and is primary and, citing Durkheim in his support, also that society constructs the individual self and makes it a sacred object – who returns the favor, in its own benefit, by itself frequently giving ritual respect to other selves, which reciprocally uphold the former one. Collins was deeply skeptic of this explanatory undermechanism, and when, partly to test the scope of this assumption, he once asked Goffman what he thought would be a microsociology of the intellectual world, Goffman brushed the question aside with a characteristically sarcastic remark. Finally, Collins was perhaps the first theorist to attach to the micro-macro dichotomy a new sociological sense in his article "The micro-foundations of macro-sociology" (1981a), in which, from a nonreductionist stance, he asserted the theoretical primacy of microsituations. Possibly in response to it, Goffman (1983: 7) stated in his Presidential Address to the *American Association of Sociology* that "(some) argue reductively that all macro-sociological features of society, along with society itself, are an intermittently existing composite of what can be traced back to the reality of encounters, a question of aggregating and extrapolating interactional effects. (…) I find these claims uncongenial," which may be read as an utter rejection of Collins' claim.

Political differences added to their increasing lack of mutual rapport. Absorbed in his research, Goffman stayed out of politics – not like young Collins. Soon after getting to Berkeley, Collins had joined a fairly idealistic civil rights group and participated in nonviolent sit-in demonstrations for the Free Speech Movement, and against racial discrimination and the Vietnam War. Some of the group members were Marxists: Stalinists, Maoists, and Trotskyites – and Collins associated with the latter. Nonetheless, as the antiwar movement actions became soon more violent many people left what was actually a political movement of intellectuals (most members were graduate students) and drifted toward its spreading into popular culture through popular musicians and the drug culture: they became more interested in the hippie counter-culture revolution, the quasi-religious psychedelic drugs movement, the encounter group psychotherapy movement, Eastern mystical religious movements, transcendental meditation, etc., all of it dressed up with rock music, and an erotic/sexual revolution. Collins became briefly a hippy, went to group therapy, and learned quite a lot about Buddhism studying with Tibetan and Japanese monks – which gave him a basis for his later sociological research on Eastern philosophers for *The Sociology of Philosophies*.

An Academic Career: Early Breakthroughs and Collins' Discontent

After reading his PhD dissertation in 1969 Collins got appointed as an assistant professor, then associate professor by the University of California–San Diego, where he stayed till 1977. There he met and became friend of Visiting Professor Alvin W. Gouldner. After being simultaneously invited by an American and a Dutch publisher respectively to start a theory-oriented journal, both agreed to combine the two enterprises, asked Pierre Bourdieu to join them, and founded *Theory and Society*. In 1979 he would be also founding editor of American Sociological Association journal *Sociological Theory*. It was also an extremely fruitful writing time: the three books Collins

published while being in San Diego gained him an enduring reputation as an insightful and path-breaking scholar.

Co-authored with Michael Makowsky, *The Discovery of Society* (1972, eighth edition 2009), now a classic introductory text to the history of sociology and to sociological theory, appeared first. Written as series of sociobiographically well contextualized intellectual and conceptual portraits of those nineteenth- and twentieth-century thinkers most significant for social thought – from first and second generation founding fathers (Comte, Marx, Tocqueville, Nietzsche, Durkheim, Weber, Pareto, Freud, Mead), through the age of sociology academic institutionalization (the Chicago School, Parsons, Goffman) to a selection of present theorists mainly focused on macro- and micro-investigations about conflict (Bourdieu, Skocpol, Wallerstein, Elijah Anderson, Donald Black), it included two chapters on evolutionism and racism, and on the emergence of African-American sociology – matters rarely treated by textbooks.

Then came *Conflict Sociology* (1975; updated edition, with Stephen K. Sanderson, 2010), that gestalt-shifted the perception of numerous authors and studies, formerly considered to be heterogeneous, into a new understanding that they constituted a coherent research field – which became named after this title. Writing from a coherence theory of truth, opposed to positivist empiricism, to dispel the prejudice that sociology was not a cumulative science, Collins arrayed it in terms of basic propositions, their logical relationships, and the phenomena – empirically documented through rigorous historical, ethnographic, and observational evidence – which they explained. It covered several areas: state development and dynamics, markets evolution, stratification and social mobility, class struggles, and deference and demeanor, among others. He wanted to show that an integrated conflict theory that synthesized the best insights of the Marxian, Weberian, and (through Goffman) Durkheimian traditions could explain a wider range of sociological issues than other views. This approach was later vindicated, since the fields of social movements (Tilly, McAdam, Tarrow, etc.) and historical sociology of the state (Skocpol, Tilly, Goldstone, Mann, etc.) set out and developed successfully precisely from a similar Weberian perspective. It is also congruent with a world-system view. Nonetheless at the time the book's viewpoint failed in a disciplinary sense because in the highly conflicting early 1970s sociologists were not for synergies and synthesis, but polarizing into several varieties of factionalized Marxism, hard-line well-funded positivism armed with more advanced statistics and bigger data sets, radical social constructivism as ethnomethodology, and post-modernism. After this setback Collins would never try again such a direct strategic syncretistic move – in the future he would rather work from both macro- and micropoles toward each other simultaneously.

Lastly, in 1979a, came out *The Credential Society: A Historical Sociology of Education and Stratification*, which basically held that just expanding contemporary educational system could never produce social equality. Instead of that, its inflationary dynamic would only cause longer pursuits of more and higher degrees in the often misguided hope that they would help to achieve better jobs and higher class positions, and that this demand would fuel public and private offers – only checked eventually by its increasing costs, and decreasing returns. Facts have essentially borne out Collins' thesis globally for more than 30 years. Furthermore, Collins' convictions guided his personal life: he had then a revolutionary orientation and believed the bases of stratification should be somehow overthrown, and so while writing *The Credential Society* he came to feel that it would be hypocritical to keep working in the credential-creating academic system. Consequently, he quit his academic post in 1977 and worked in commercial writing for a year before accepting a professorship in the University of Virginia, which he held up to 1982. Then he quit again and worked as a professional writer for another 3 years.

(Incidentally, Collins also published in 1979b his novel *The Case of the Philosophers' Ring*, written in 3 weeks after he bet against a friend that he could write a Sherlock Holmes story in that time using a well-tried formula: 'Holmes meets…' – here famous 1900s cultural figures: Wittgenstein, Russell, Keynes, Ramanujan, Virginia Woolf, and many others, as he took advantage of his work on intellectual networks.)

During those days of independent scholarship alternated with resident and visiting professorships, Collins made his living mainly by writing textbooks – which admittedly makes profit of the academic system too; as a moral compensation, he tried to raise his works at this generalist-synthesizer's trade above the common level. And he worked hard: 5 books in 5 years. The most theoretically noteworthy of them is *Sociological Insight* (1982, second edition 1992), where he head-on criticized rational action theory. Collins stated that in any socially meaningful context the key mechanisms of human choice rather more frequently than the cognitive mechanisms emphasized by the utilitarian tradition involve decision habits and practices developed and learned in interaction rituals involving emotions, a microinteractional focus and mutual entrainment. To show the explanatory fruitfulness of his ritual–interactional approach Collins applied it there to several macroscale phenomena: religion, power, crime, sex/love, and even artificial intelligence. Another major scholarly successful book was *Sociology of Marriage and Family: Gender, Love and Property* (1985b, fifth edition 2000; from the third on co-authored with Scott Coltrane), an extensive, historically informed, and conflict-sensitive handbook which ranges from the geopolitical and economic conditionings of kinship-family structures through love and marriage markets (influenced also by social class, racial, ethnic, and other social and cultural variables – such as the rise of feminism) to intimate issues such as erotic ties, contraception, childbearing, marital cohabitation, violence, etc.

The other three books dwelled on the history of sociological theory. In 1981b he published *Sociology since Mid-Century*, a collection of papers and reviews on some of his constant themes (the empirical support for conflict theory; geopolitics, social change, state power, and revolution; symbolic and physical violence; schooling in the USA) together with sharp analysis and balanced criticism of several contemporary social thinkers (Blau, Goffman, Gouldner, Levi-Strauss, Merton, Touraine, Wallerstein, etc.) In 1985a *Three Sociological Traditions* (and its accompanying edited reader) followed, a bold comparative assessment of Marx–Weber–Coser conflict-oriented viewpoint, the ritual solidarity/functionalist paradigm of Durkheim–Parsons–Goffman, and the interactionist view of Mead–Blumer–Garfinkel – expanded later with rational choice

theory as *Four Sociological Traditions* in 1994. Finally his brief *Max Weber: A Skeleton Key* (1985c) appeared in a collection of short biographies of classical social theorists.

An Academic Career: An Integrative and Innovative Sociological Theory

In 1985 Collins returned to the University as Professor of Sociology in California–Riverside, where he spent the next 12 years – but he did not work on credentialism again for decades. There he began to fully develop his overarching view of Sociology. If *Sociological Insight* had stated his sophisticated Durkheimian position regarding microsociology, *Weberian Sociological Theory* (1986) did the same for macro- and meso-sociology from a Weberian perspective. Collins showed that Weber's late work, based on massive historical comparisons, offered theories of the origin of capitalism (in rationalizing practices of medieval Christian and Buddhist monasteries), and of the development of the state (as a rationalizing by-product of a world-system model of geopolitics) that had hitherto been utterly neglected. More importantly, Collins showed how to apply the more fruitful lines of Weber's approach – state/market macrolevel dynamics – to the analysis of cultural issues such as dissidence (whether religious heresy or intellectual alienation), or the evolution of family structures and women's social status. This book also included a chapter devoted to the prognosis of the future decline and eventual demise of the URSS – originally formulated in a 1978 article, and that most people found conspicuously preposterous then but which would soon enhance Collins academic fame perhaps more than any other of his former contributions.

Theoretical Sociology (1988), perhaps Collins' most ambitious book to that date, was simultaneously a textbook on sociological theory and an attempt to assess what valid theoretical, and most importantly empirical knowledge there was in several theoretical schools: at the macrolevel evolutionism, functionalism/system theories, Marxism and its kin, and conflict theory; Mead–Blumer's symbolic interactionism, Homans-Blau's exchange theory, and Goffman's interaction rituals at the microlevel; and what was the main innovation of the book, a section on mesolevel theories, including chapters on the micro–macro link, network theories, and organization theories – conceiving organizations as network-structures operating at that level, whose mechanisms shape the conditioning circumstances of their composing microinteractions with different kinds of task and communication problems. On the whole, the book was not very successful, mainly because of its criticism of fashionable theories such as postmodernism.

An Academic Career: An Overarching View and Violence Studies

In 1997 Collins became and has been since the Dorothy Swaine Thomas Professor in Sociology at the University of Pennsylvania, and it has been there where his *oeuvre*, i.e., his attempt to build a comprehensive sociological theory has flourished and finally earned him world recognition. In 1998, culminating more than 30 years of research, he published *The Sociology of Philosophies*, a full analysis of more than two millennia of Eurasian philosophical networks. This massive historical comparative exercise in mesosociology yielded impressive discoveries such as the 'law' of the small number of philosophical positions (often six) viable simultaneously in an intellectual attention space; the basis of 'rapid-discovery science' in genealogies of laboratory equipment connected by networks of persons who modify and recombine them into equipment that produces a stream of new phenomena to study; the concept of ideas as social coalitions of thoughts guided by their prospective relevance in valued intellectual networks; the pragmatic view of intellectual greatness as sociocognitive usefulness for next generations of thinkers; or the dependence of creativity on a thinker's network position, which determines the cooperative-competitive rituals and consequently the emotionally charged ideas she/he has access to, and the emotional energy she/he gets there to combine them in new potentially successful ways, etc.

In 1999 it was the time of *Macro-History*, a collection of essays aligned with the most successful modern historical sociology, where Collins presents a theory of market dynamics across history including early kinship structures based on marriage markets, ancient military states based on conquest of slaves, religious corporate capitalism in medieval Christianity and Buddhism, and modern capitalism with its expansion and crises. Collins also shows how wide historical comparisons can lead to solid generalizations on, for instance, the variations of ethnic dominance within states, the extent and health of democracy, or the chances of success of social revolutions. Finally, after co-editing with Mauro Guillén and others *The New Economic Sociology* (2002), and as if rounding off a trilogy devoted to the three levels of social dynamics, Collins published *Interaction Ritual Chains* (2004), a work which presented a general theory of social rituals as a general purpose tool to understand microsituations, based on a full analysis of the functioning of ritual interaction situations as mechanism for building collective identity and solidarity.

Developing and combining insights from both the tradition of minute conversational analysis (Sacks, Schegloff, Sudnow) and the sociology of emotions (Jack Barbalet, Arlie Hochschild, Jonathan Turner, Thomas Scheff), Collins explained interaction rituals' intellectual, moral, and behavioral efficacy by their build up or draining down of a measurable and quantifiable individual variable he calls emotional energy, a distinctive social emotion subjectively felt as strength, self-confidence, and thrust to take the initiative and participate in social action. Emotional energy is, together with emotionally charged symbols of one's group, a feeling of collective belonging and a moral stance on behalf of that group, the ordinary fruit of successful rituals. A market-like dynamics where agents compete in order to participate in the best rituals their material/personal/cultural capital gives them access to, and to foster those they find most rewarding, sustains the network of ritual occasions that make up the interactional structure of a culture. Collins applies this theory to the evolution of sex/love and to tobacco rituals, to historical variations in modern Western elites rituals of power/deference, to economic/social/status-fame stratification at large, and to the social construction of the innermost parts of the individual self: the very process of thinking of an individual and her/his personality type – even

the seemly less social one, the introvert, as shown in his historical diversity.

Recently Collins has embarked in what will be another series, now on social violence. Going beyond the analytical approach of conflict theory, he studies the actual dynamics of fighting at microlevel – where shared focus of attention and a common (negative) emotion generates tension/fear instead of rapport – and its effects at every other social scale. Its first installment was *Violence: A Micro-Sociological Theory* (2008), an investigation on actual moments of fighting and almost-fighting based on first-hand empirical data (eye-witness written testimony, real-life photos, and videos), which shows that humans tend to be averse to face-to-face violence and very incompetent when they resort to it. This would be a consequence of an evolution-developed innate and socialization-enhanced human propensity to collectively entrain in cooperative ritual behavior (conflict being a result of variously motivated failures of this dynamic). So, in potentially violent situations humans feel inhibiting tension-fear, and violence only proceeds when they find a way to overcome it – sometimes in severely destructive ways, as in forward-panics, when a group or an individual which have been under prolonged and serious threat suddenly get emotional dominance over its rival and charges viciously against her/him. History of violence is the history of techniques to train humans to be effective fighters. All this shows the falsity of current popular views of violence based in images and accounts edited by news media and stylized and glamorized by the entertainment industry.

The second book of this series deals with the time-dynamics of conflicts and its relation with collective dynamics of emotional motivation and domination. Collins (2012) Presidential Address to the American Sociological Association (he served as ASA president in 2011–12) may be considered as the book's forerunner article. The third book is expected to be devoted to recently developed forms of high-tech military violence, and the differences it purports for the emotional dynamics of war at all human levels.

Before that, though, he has co-authored with other four leading scholars from the field of macrohistory *Does Capitalism Has a Future?* (2013). Acknowledging that society at macroscale is too complex to allow for detailed predictions, they also believe that macrostructural conditions may be forecasted. Collins contribution offers a definite prognosis of dire straits for capitalism. Basically, he expects an underconsumption crisis as a consequence of a massive destruction of middle-class employment in the services sector due to the advance and world deployment of information technologies – which technology, spreading of markets, financing sector, government investment, or educational expansion will not be able to counterbalance. In order not to collapse, the world-system would evolve, through more or less peaceful or violent transitions, into a variety of central command economies whose governments, whether democratic, authoritarian, or something in between, would likely manage difficult redistributions of wealth by regulating or bypassing markets, endlessly balancing subsidiary capitalist and core central command – perhaps often socialist – features.

A closing remark: since 2010 Collins writes a blog (http://sociological-eye.blogspot.com.es/) where to date (September 2014) he has published 20 article-long essays on current social issues from the vantage point of microsociology (interaction rituals and the new electronic media, Mona Lisa's smile, USA immigration crisis), sociology of intellectual life (emotional energy and belief in free-will, Jesus' charisma, changing love plots y XXth narrative or film and the sexual revolution), political dynamics (recent tipping point revolutions, the hero of Tiananmen), sociological theory (sociologically significant quantities, ambiguity of material interests, network morphology dynamics, the intellectual descent of Goffman and Garfinkel), and sociology of violence (bullying, mass rampage killers, animal violence, organized crime, Alexander's and Napoleon's carriers). These texts illustrate once more the broad applicability and heuristic fruitfulness of Collins' three-tiered approach (Interaction-Ritual Theory, Network Theory, Macro-History Geopolitical (state-and-market) Theory) to comprehensively understand the courses and recourses of social life: how under macro-world-market/geopolitical and mesoorganizational constraints micro-meso-and-macro data get processed in microsituations process resulting in patterned recurrent individual and group behaviors that renew or change the micro-meso-and-macro structures and dynamics of cultures and societies. (This short biography is based in Maclean and Yocom (2000), Santoro (2012), and Waldstein (2007), in Randall Collins' books, articles, and blog posts, and in personal communication with him.)

See also: Balance of Power, History of; Bendix, Reinhard (1916–91); Bourdieu, Pierre (1930–2002); Comparative History; Coser, Lewis A. (1913–2003); Democracy, History of; Durkheim, Emile (1858–1917); Gender History; Goffman, Erving (1922–1982); History and the Social Sciences; Intellectual History; Mann, Michael (1942–); Marx, Karl (1818–83); Parsons, Talcott (1902–79); Skocpol, Theda (1947–); Social Science, The Idea of; State, History of; Wallerstein, Immanuel (1930–); Weber, Max (1864–1920).

Bibliography

Bendix, Reinhard, Collins, Randall (Eds.), 1968. State and Society. Little, Brown, Boston. Re-issued, University of California Press, 1974.

Collins, Randall, Makowsky, Michael, 1972. The Discovery of Society, eighth ed. 2009. Random House, New York.

Collins, Randall, 1975. Conflict Sociology: Toward an Explanatory Science. Academic Press, New York. Collins, Randall, 2010. Conflict Sociology: A Sociological Classic Updated. (Abridged and updated by Stephen K. Sanderson). Paradigm Publishers, Boulder, CO.

Collins, Randall, 1979a. The Credential Society: An Historical Sociology of Education and Stratification. Academic Press, New York.

Collins, Randall, 1979b. The Case of the Philosophers Ring (Fiction), British ed. Crown Publishers, New York. Harvester Press, London, 1980; reissued, Ostara Publications, 2009.

Collins, Randall, 1981a. The micro-foundations of macro-sociology. American Journal of Sociology 86, 984–1014.

Collins, Randall, 1981b. Sociology since Mid-century: Essays in Theory Cumulation. Academic Press, New York.

Collins, Randall, 1982. Sociological Insight: An Introduction to Non-obvious Sociology. Oxford University Press, New York.

Collins, Randall, 1985a. Three Sociological Traditions & Three Sociological Traditions: Selected Readings, ed. volume. Oxford University Press, New York. Collins, Randall, 1994. Four Sociological Traditions & Four Sociological Traditions: Selected Readings (ed. volume); second ed., Oxford University Press, New York.

Collins, Randall, 1985b. Sociology of Marriage and Family: Gender, Love and Property. Fifth ed. 2000 (with Scott Coltrane). Nelson-Hall, Chicago.
Collins, Randall, 1985c. Max Weber: A Skeleton Key. Sage, Beverly Hills.
Collins, Randall, 1986. Weberian Sociological Theory. Cambridge University Press, Cambridge and New York.
Collins, Randall, 1988. Theoretical Sociology. Harcourt, Brace, Jovanovich, San Diego.
Collins, Randall, 1998. The Sociology of Philosophies: A Global Theory of Intellectual Change. Harvard University Press, Cambridge.
Collins, Randall, 1999. Macro-history: Essays in Sociology of the Long Run. Stanford University Press, Stanford.
Collins, Randall, 2002. The new economic sociology. In: Guillén, Mauro, England, Paula, Meyer, Marshall (Eds.), Developments in an Emerging Field. Russell Sage Foundation, New York.
Collins, Randall, 2004. Interaction Ritual Chains. Princeton University Press, Princeton, NJ.
Collins, Randall, 2008. Violence: A Micro-sociological Theory. Princeton University Press, Princeton, NJ.
Collins, Randall, 2012. C-Escalation and D-Escalation: a theory of the time-dynamics of conflict. American Sociological Review 77, 1–20.
Collins, Randall, 2013. Does Capitalism Have a Future? (Co-authors: Immanuel Wallerstein, Michael Mann, Georgi Derluguian, Craig Calhoun) Oxford University Press, New York.
Goffman, Erving, 1983. The interaction order. American Sociological Review 48, 1–17.
Maclean, Alair, Yocom, James, September 2000. Interview with Randall Collins. http://www.ssc.wisc.edu/theoryatmadison/papers/ivwCollins.pdf (accessed 25.11.13.).
Santoro, Marco, 2012. La sociologia in (inter)azione. Una conversazione con Randall Collins. Rassegna Italiana Di Sociologia 2012–2014, 719–750.
Waldstein, Maxim, March 2007. Interview with Professor Randall Collins. http://www.academia.edu/905324/Interview_with_Professor_Randall_Collins_March_2007_ (accessed 24.11.13.).

Colonialism, Anthropology of

David Scott, Columbia University, New York, NY, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 4, pp. 2232–2237, © 2001, Elsevier Ltd.

Abstract

The article surveys aspects of the history of the anthropology of colonialism conceived as a story of successive contexts of questions in which the problem about colonialism for anthropology has been shaped and reshaped. It begins in the 1970s with the concern with the role of the discipline in the colonial/imperial project, and with the ethics/politics of its practitioners. These concerns with the effects of world politics are displaced in the 1980s by questions that turn on the problem of colonialist representation, the imbrication of power and discourse in the making of anthropological knowledge. In the context of the realignment of disciplines in the 1980s, there is also a reformulation of relation between anthropology and history. In the 1990s there is another shift, one away from an interest in textuality towards a concern with the relationship between colonialism and modernity. Here issues such as nationalism and the colonial state are reconceived. Finally, it is suggested that because the anthropology of colonialism has been oriented towards understanding the present as much as the past it has been importantly a *critical* enterprise. The article urges that to retain this critical purchase it has to more adequately specify the domain of contemporary *postcolonial* questions that theorizing colonialism is meant to illuminate.

The story of the relationship between colonialism and anthropology (or more specifically, of the anthropology of colonialism) is not the story of a neatly unfolding self-consciousness. It is, rather, the story of successive contexts of questions that have shaped the intellectual space in which the problem of colonialism for anthropology has been formulated and engaged.

Anthropological self-consciousness of colonialism, that is, of colonialism *as a distinctive problem for anthropological knowledge and anthropological practice*, begins roughly in the 1970s. In the early years of that decade two edited volumes appeared dealing with anthropology's relation to the social and political world of which it was a part. One came out of the tradition of US cultural anthropology, *Reinventing Anthropology* edited by Hymes (1972); and the other out of the tradition of British social anthropology, *Anthropology and the Colonial Encounter* edited by Asad (1973). Both collections perceived the discipline to be in serious crisis; and both were radical in the sense that they were explicitly concerned to subject that crisis to a fundamental self-scrutiny.

Hymes's volume urged that the social and political conditions in which anthropology had emerged as a professional academic discipline in the US at the turn of the twentieth century had, by the late 1960s, significantly altered. First, there was the overwhelming fact that with political decolonization there was a fundamental change in the conditions in which field- work could be conceived and undertaken. At the same time, the relation between the US government and the Third World (as the colonized or recently decolonized part of humanity came to be called from the 1950s) had given rise to troubling moral-political questions about the state of the anthropological enterprise. In particular, US involvement in attempted counter- insurgency in Latin America, and the role of anthropologists as cultural experts in the war in Indochina, served to pose the question of the relationship between 'objectivity' of ethnographic data and political interest in a very acute way (see Stocking, 1991). What place would anthropology have in raising fundamental questions about injustice and inequality in the world it inhabited, and what was its role in trying to transform that world? Was anthropology to be employed as an instrument of Western domination or as a means of liberation? Moreover, the publication of Malinowski's field diaries in 1967 and the revelation of his barely veiled racism further focussed attention on the ethics of the relation between ethnographer and informant.

In this context, the challenge *Reinventing Anthropology* sought to pose was whether the discipline would be able to transform – to *reinvent* – itself in a way that reflected the demands for political accountability, ethical concern, and critical social commitment on the part of anthropologists. Two strains in US anthropology inspired the volume's contributions: one – reflected especially in the articles by Hymes and Stanley Diamond – was the radical Boasian tradition in the work of people like Edward Sapir, Ruth Benedict, and Paul Radin, who thought of anthropology as a reflexive critique of civilization. The other strain – reflected in Eric Wolf's contribution – was a post World War II Marxism (much influenced by the cultural-evolutionary materialism of Julian Steward) which was concerned to highlight the plight of peasant societies caught in the juggernaut of capitalist imperialism.

The intellectual tradition of British social anthropology has of course a very different historical context than US cultural anthropology, one far more intimately interconnected with the subordination, management and rule of colonial subjects. Not surprisingly, then, the concerns of Asad's *Anthropology and the Colonial Encounter* were somewhat differently inflected than those of Hymes's volume. To be sure, they shared a concern that since World War II fundamental changes had occurred in the world inhabited by social anthropology. The emergence of new nations, especially in Africa where much British social anthropological fieldwork was carried out – Sudan (1956), Ghana (1957), Nigeria (1960) – made it increasingly difficult to avoid the pertinence of a more historical approach that brought the larger political and economic system of colonialism into view. And in consequence it was increasingly difficult not to be critical (as indeed Edmund Leach was in

a limited way) of the ahistorical functionalism and empiricism that characterized British anthropological theorizing.

In the middle 1960s, New Left critics had in fact launched an attack on British social anthropology. And part of the project of Asad's volume was to respond to these and other radical claims, not simply to affirm or deny them, but to more adequately specify and formulate them. There was a sense that perhaps, while undeniable, the relation between colonialism and social anthropology was more complicated than was allowed by the polemical tone of accusation and disavowal. In her contribution, for example, Wendy James argued that the very existence of social anthropology in the colonial period constituted a source of *potential* radical criticism of the colonial order, and she sought to provide a more nuanced account of Malinowski's own active social reformism during the interwar years. Stephan Feuchtwang, in his contribution, discussed the relation between the social and institutional conditions of British anthropology and the knowledges it produced. In an allied way, Asad examined 'objectifications' of colonial power produced in two authoritative domains of systematic knowledge – the functionalist anthropology of Africa and Islamic orientalism. Asad was concerned to understand the connections between these different images of rule (one largely of 'consent,' the other largely of 'repression'), their ideological roots, and conceptual consequences.

In subsequent decades, the anthropological preoccupation with colonialism has tended to be largely (if not solely) a phenomenon of the US academy. At any rate, in the 1980s there was a gradual shift away from the concern with the *role* of the discipline in the colonial/imperial enterprise and the politics of its practitioners, toward a concern with colonial forms of knowledge about non-European worlds. The register in which the problem of colonialism for anthropology was conceived was now less a political than an epistemological one. Or rather, linking power and knowledge, colonialism was now conceived as a problem of (in the language of the day) the politics of representation. This shift has many interconnected sources. I shall only be able to touch on a few of the more salient ones.

At one level, of course, the shift is connected to the decline of anti-imperialism and anticolonialism as going *politico-ideological* concerns. On the one hand, beginning with the end of the Vietnam War and accelerating through the years of Ronald Reagan's US Presidency there is a marked decline in the First World intelligentsia's interest in revolutionary change in the Third World. On the other hand, in the Third World itself there is a decline (indeed a rapid unraveling) of the experiments with radical social transformation in many of the independent countries of South Asia, Africa, the Caribbean, and Latin America. At another level this shift was connected to a number of *intellectual* transformations in the understanding and practice ofthe human sciences. Very broadly speaking, these may be characterized in terms of a general drift away from the determinism (Leslie White, Julian Steward) and functionalism (Radcliffe-Brown) of postwar anthropology and toward a social/cultural constructionism. Principal among these transformations was the so-called 'turn to meaning' in social and cultural analysis. In many ways, of course, the work of Geertz (1973) embodied – indeed, helped to constitute – this constructionist moment. In his work, culture came to be understood as a web of semiotic processes or relations, and cultural analysis as a practice less of uncovering determinate causes, or of explaining functions, as of deciphering meanings. Two features of this anthropology are of special importance here: one is that its criticism of objectivism brought about a renewed sensitivity to 'the native's point of view' and 'local knowledges'; and the other is that the critique of positivist social science led to a greater openness to interdisciplinarity, to literature and history in particular (Geertz's famous 'blurring of genres').

This work, in turn, set the stage for the emergence in the 1980s of an interest in the *textuality* of ethnographic description and analysis, in the senses in which these are artifacts of language – hence, in their way, *literary* – and therefore open to an investigation of their rhetorical and poetic structures (Clifford and Marcus 1986). How is ethnographic authority constructed? What are the linguistic devices that enable anthropological representations to produce their distinctive 'realistic' effects? Some of this work sought explicitly to reconceive the task of a critical anthropology (Marcus and Fischer, 1986).

This preoccupation with modes of representation in social and cultural analysis was, in part at least, the context for the rise of what would come to be called colonial studies. An interdisciplinary field (drawing largely on literary critics, anthropologists, and historians), colonial studies would help to institutionalize the renewed scholarly focus on colonialism. The central inspiring text here was Said's *Orientalism* (1978). Borrowing in some measure from the middle work of Foucault (1979), this book thematized the idea of colonialism as a *discursive* formation. Said set out to demonstrate the ways in which, in the (especially French and British) colonial discourse of orientalism, the Orient was invented. The Orient, therefore, was less a geo-political location than a figure in the cultural-political imagination of the West. *Orientalism* opened up a rich vein ofresearch and a critical demand to interrogate the West's authoritative discourses on the non-West – anthropology chief among them – and to unmask the ways in which they produced and reproduced their hegemonic knowledges. Interestingly Said himself (1988) would later express doubts about the value and future ofa discipline so deeply implicated in the making of such hegemonic knowledge.

One anthropological work that drew together some of these themes – though in a way less pessimistic about the possibility of a revised anthropological enterprise – was Fabian's *Time and the Other* (1983). The subtitle of the work was significant: *How Anthropology Makes its Object*. It focussed attention precisely on the 'constructedness' of anthropological discourse, specifically on the uses of Time in anthropology's constitution of its object (the savage, the primitive, the native, the Other). Fabian sought to show how the existential copresence of the Other in the dialogical interaction of ethnographic fieldwork gets transformed into a temporal distancing – what he called a denial of coevalness or cotemporality – in the theoretical construction of authoritative anthropological knowledge. This 'conjuring trick' served perpetually to consign the native to a temporal zone outside of the Time of anthropology, which is to say the privileged Time of the West. It is easy to see that the lessons of Fabian's book are wider than the issue of temporality alone. What it demonstrated was that unless anthropology subjects its own discourse to critical scrutiny, unless it pays attention to the conceptual assumptions that go into the *making* of its objects, it is liable to reproduce colonial

knowledge. Other work would later seek to extend and develop this inquiry into the formation of anthropological objects (see Scott, 1994).

Another set of anthropological concerns with colonialism in the 1980s was more closely connected to the antinomies of Marxism. The defeat of the communist project in Western Europe after 1920 had given rise to a strain of Marxism (soon to be called Western Marxism) focussed on what had hitherto been dismissed as superstructural issues. Marginalized until the rise of the New Left in the 1960s, this strain became more prominent in the 1970s, especially in the North Atlantic academy. Antonio Gramsci's theorization of hegemony, Louis Althusser's reconceptualization of ideology as representation, and Raymond Williams's displacement of the base/superstructure model for the analysis of literature (and cultural production more generally) offered less reductionistic critical languages in which to situate social and cultural life in relation to capitalism. An interest in rethinking the problem of the articulation of modes of production, and of precapitalist (and precolonial) formations was one direction of Marxist work that had actually begun earlier. But in another direction, folded into the anthropological concern with local knowledges and finite histories, this cultural Marxism offered the prospect of an antiuniversalistic account of hitherto invisible idioms of resistance to dominant power (including, of course, globalizing capital), and of hitherto unacknowledged forms of (often informal) agency in the quotidian making of histories. In their very titles, James Scott's *Weapons of the Weak* (1985), and Jean Comaroffs *Body of Power, Spirit of Resistance* (1985), offer exemplification of this interest in the anthropology of anticolonial resistance.

Throughout the 1980s history was becoming a site of profound scholarly preoccupation. Indeed within the discipline of history itself there was already a growing interest in the intersection of modes of cultural and historical representation. Work in feminist history, in particular, helped to reformulate relations between power, difference, and representation. But perhaps the most significant contribution to the argument for more than a nodding relationship between anthropology and history – and this too on the terrain of colonialism – has been the pioneering work of Cohn (1987). Cohn's essays were only brought together in 1987, under the title, *An Anthropologist among the Historians and Other Essays*, but from the early 1960s he had been urging the theorization of the intersection of history and anthropology. In a number of wide-ranging essays on British India – from concerns with the formation of historical, legal, and statistical modes of colonial knowledge, to an interest in ritualized representations of political authority – Cohn demonstrated that historical analysis of colonial rule was indispensable for anthropological understanding. This work inspired a number of innovative studies in the historical anthropology of colonial India (among them Dirks 1987). Significantly too, India was the conceptual and intellectual-political site of another important contribution to the debate about history for an anthropology of colonialism – the Subaltern Studies Collective's re-interrogation of liberal and nationalist accounts of colonial rule in India and the anticolonial struggle against it (Guha, 1984). Trans-disciplinary in perspective, the work of the Collective resonated with the anthropological discontent with universalism and Eurocentrism and its interest in the agency of non-dominant idioms, knowledges, and practices.

In the 1990s one can discern another shift in the questions driving the anthropology of colonialism. These new questions turn less on the textuality of colonial representation or the agency of resistance than on the problematization of the relationship between colonialism and modernity, and the implications of this relationship for understanding the postcolonial present (Asad, 1991, 1992; Scott, 1997). The profoundly epochal sense attending the *fin de siecle*, the sense of a crisis of modernity (and for some, an emerging *post*modernity), has made for a growing preoccupation with a re-examination of the formation of the modern world and its moral, epistemic and political modes of legitimization. Much of this re-historicizing work has sought to criticize the progressivist story of modernization.

There are several dimensions to this anthropological preoccupation with colonialism and modernity. One of the earliest has been the concern with the problem of nation and nationalism. There had of course been an older multidisciplinary interest in nations and nationalisms, namely, the 'new nations' projects of the early 1960s. These studies were primarily concerned with inquiring into the modernizing and democratizing challenges facing the newly independent nation-states. However, two books more than any others reopened and significantly reshaped the debate about the nation: Gellner's *Nations and Nationalism* (1983), and Anderson's *Imagined Communities* (1983). Each was focussed differently, but they were both inspired by the problem of the relation between nation/nationalism and *modernity*. If the 'new nations' work tended teleologically to take nationhood as a normal developmental threshold, what Gellner and Anderson foregrounded was the idea that nations were not natural entities nor nationalisms primordial sentiments or ideologies. Both, they argued, were profoundly historical, and that historical character, moreover, owed most to the making of the *modern* world.

Gellner and Anderson would both be criticized for their emphasis on the modernity of nations and nationalisms, but it was precisely this connection that those interested in colonialism were concerned to explore. Anticolonial nationalisms and the making of postcolonial sovereignties clearly had a very complex – and as yet, perhaps, unresolved – relation to the modernizing project of late colonialism. In this regard the decisive work was Chatterjee's seminal book, *Nationalism and the Colonial World*, originally published in 1986 but becoming influential in debates from the early 1990s. A noticeable feature of this work was its interdisciplinarity, and especially the place of anthropology (as the authoritative discourse on Europe's Others) within it. While indebted to Said's *Orientalism* for its opening deconstructive move, Chatterjee undertook a sustained examination of anthropological assumptions (e.g., about rationality) undergirding the distinction between Western and non-Western nationalisms.

Connected to this focus on rethinking the relations between modernity and nation, has been the renewed attention given to the problem of modern colonial power generally, and the colonial state in particular. Especially important here has been the later work of Foucault on governmentality (1991) that raised the question of the place of the state in a larger and much

transformed understanding of power. Questioning a deeply influential tradition of social and political analysis owing much to both Marx and Weber, Foucault's concept of governmentality sought to think (conceptually) about a field of power in which the state was not the central analytical domain; and (historically) about a concrete moment in European history when political power becomes more dispersed, less dependent upon the institutions of coercion. This concept opened up the possibility of reconceiving the questions through which the problem of colonial power was formulated. Where modernization narratives (liberal or Marxist) had plotted a story of the progressive rationalization of the state (the more or less of power; the counter-positioning of power and freedom), what now emerged was the possibility of a story of colonial power told in terms of the political rationalities by which it was organized. The point here was to urge an historicization in which the emergence of a colonial rationality of government (i.e., a form of colonial power directed at transforming conditions so as to produce effects on colonial conduct) could come more sharply into view. And the significance of this is that this rationality is central to the making of our postcolonial worlds (see Scott, 1999).

There have been critics, of course, who are skeptical about this thematization of governmentality in the anthropology of colonialism. The more sympathetic of them argue that neither the old master-narrative of progress and modernization nor the new Foucauldian revisionism are adequate because neither appreciates the inherent *doubleness* of colonial rule. The essential paradox of the colonial state, it is argued, is that it was always caught between *two* principal imperatives – ruling subjects and making citizens – and, therefore, always moved simultaneously in two directions: coercion and civilization. The crucial hinge on which this doubleness hung was the colonial language of law and legality (Comaroff, 1998). In a related direction some critics have been concerned that the whole relation between metropole and colony has been ill-conceived (Cooper and Stoler, 1997). The worry here is the assumption among many students of empire of a Manichaeism that sees the problem of colonialism as a binary – colonizer/colonized – one. These critics argue that, to the contrary, the conception of an encounter between an empty colony simply remade by Europe and a Europe that remains ever unmarked by its empire is untenable. They urge a view that places metropole and colony in a single (if internally differentiated) analytical field, and focuses precisely on the multiple tensions – inclusions/exclusions, hierarchies, reversals, etc. – through which Europe and its empire were (unevenly and asymmetrically) co-constructed.

Since its emergence as a distinctive theme in anthropology in the 1970s, the problematization of colonialism has been thought of as part of an orientation towards the present at least as much as, if not more than, the past. In this sense there has always been about it an animating ethos of criticism – whether criticism of our worlds or criticism of our disciplines. The anthropological problem about the colonial past, in other words, has been connected to raising critical questions about the postcolonial present: it is part of the making of a distinctively *postcolonial* anthropology. However this connection and its implications have been more often assumed than systematically thought out. Consequently, anthropology has been less concerned than it might, not merely about the ways in which understandings of the colonial past are conditioned by contemporary 'scenes of inquiry,' but more importantly about the ways in which a self- consciousness about alterations in the latter ought *strategically* to alter the kinds of questions we pursue about the nature of the former. The point, then, is that for anthropological criticism to have purchase on the present it is not enough to insist that the purview of colonial studies be enlarged (to encompass Europe, for instance, or gender, or the state), it is necessary to specify the domain of contemporary questions that are being addressed in so doing. This is the only way to gauge the extent to which our critical exercises are meeting a contemporary demand as opposed to simply expanding an already established cognitive field.

This is the challenge. It is an important challenge because the crisis upon which the formerly colonized world entered in the 1980s and 1990s (unlike those of the 1960s and 1970s) is a crisis not merely of policy orientations or development options or even ideological alternatives. It is a crisis of the legitimacy of the modern/modernizing assumptions that have undergirded the very making of these colonial and postcolonial societies and polities (assumptions, for example, about the relation between culture and politics; between nation and sovereignty; between individual rights and historical traditions, and so on). Our *conjuncture,* in short, is shifting. And therefore, if it is to more adequately theorize the relation between the problem-space in which it produces knowledge and the questions through which it does so, the anthropology of colonialism has to attend more systematically to the changing contours and character of this postcolonial present and its crisis.

See also: Colonization and Colonialism, History of; Nationalism: General; Racism, History of.

Bibliography

Anderson, B., 1983. Imagined Communities. Verso, New York.

Asad, T. (Ed.), 1973. Anthropology and the Colonial Encounter. Humanities Press, New York.

Asad, T., 1991. From the history of colonial anthropology to the anthropology of western hegemony. In: Stocking Jr., G.W. (Ed.), Essays in the Contextualization of Ethnographic Knowledge. University of Wisconsin Press, Madison, WI Colonial Situations.

Asad, T., 1992. Conscripts of western civilization. In: Gailey, C. (Ed.), Dialectical Anthropology: Essays in Honor of Stanley Diamond. University Press of Florida, Gainesville, FL.

Chatterjee, P., 1986. Nationalism and the Colonial World: A Derivative Discourse? Zed, London.

Clifford, J., Marcus, G.E. (Eds.), 1986. Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley, CA.

Cohn, B.S., 1987. An Anthropologist among the Historians and Other Essays. Oxford University Press, Delhi.

Comaroff, J., 1985. Body of Power, Spirit of Resistance. University of Chicago Press, Chicago.

Comaroff, J., 1998. Reflections on the colonial state, in South Africa and elsewhere: Factions, fragments, facts and fictions. Social Identities 4 (3), 321–361.

Cooper, F., Stoler, A.L. (Eds.), 1997. Tensions of Empire: Colonial Cultures in a Bourgeois World. University of California Press, Berkeley, CA.

Dirks, N.B., 1987. The Hollow Crown: Ethnohistory of an Indian Kingdom. Cambridge University Press, Cambridge, UK.

Fabian, J., 1983. Time and the Other: How Anthropology Makes its Objects. Columbia University Press, New York.

Foucault, M., 1979. Discipline and Punish (Sheridan, A., Trans.). Vintage, New York.
Foucault, M., 1991. Governmentality. In: Burchell, G., Gordon, C., Miller, P. (Eds.), The Foucault Effect: Studies in Governmentality. University of Chicago Press, Chicago.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Gellner, E., 1983. Nations and Nationalism. Blackwell, Oxford, UK.
Guha, R. (Ed.), 1984. Subaltern Studies. vol. 1. Oxford University Press, Delhi.
Hymes, D. (Ed.), 1972. Reinventing Anthropology, First ed. Pantheon, New York.
Marcus, G.E., Fischer, M.J., 1986. Anthropology as Cultural Critique. University of Chicago Press, Chicago.
Said, E.W., 1978. Orientalism, 1st ed. Pantheon, New York.
Said, E.W., 1988. Representing the colonized. Critical Inquiry 15, 205–225.
Scott, D., 1994. Formations of Ritual: Colonial and Anthropological Discourses on the Sinhala Yaktovil. University of Minnesota Press, Minneapolis, MI.
Scott, D., 1997. Colonialism. International Social Science Journal 154 (December): 517–526.
Scott, D., 1999. Refashioning Futures: Criticism after Post-coloniality. Princeton University Press, Princeton, NJ.
Scott, J.C., 1985. Weapons of the Weak: Everyday Forms of Peasant Resistance. Yale University Press, New Haven, CT.
Stocking Jr., G.W., 1991. Colonial situations. In: Stocking Jr., G.W. (Ed.), Colonial Situations: Essays in the Contextualization of Ethnographic Knowledge. University of Wisconsin Press, Madison, WI, pp. 314–324.

Colonization and Colonialism, History of

Wolfgang Reinhard, University of Freiburg im Breisgau, Freiburg im Breisgau, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article begins with an exploration of the semantic field of 'colony' and its theoretical implications. On this basis, an outline of the history of colonization and colonialism hints first at analogous cases in antiquity and elsewhere in world history, then describes European expansion since the early Middle Ages down to twentieth-century imperialism. Next, a typology of expansion processes, of individual motives of participants and of colonies is presented. Finally, after the history of decolonization, the consequences of colonialism in the postcolonial world are discussed, postcolonialist theory included.

Introduction: Problems of Definition and the Semantic Field

"Colonization is […] a phenomenon of colossal vagueness" (Osterhammel, 1995: p. 4), because it covers large and rather different parts of the world and its history. Therefore, a convincing theory of colonization and colonialism in general or of European colonialism in particular does not exist and will probably never exist, because it is not feasible. But scholars have still to define the concepts they use, because an agreement upon their meaning makes communication possible. Colonization has to do with migration, because it describes the movement of people from one part of the world to another to establish a settlement, quite often an agrarian one. In this sense, the term has a neutral connotation. In contrast, colonialism has become a general invective against Western policy, especially since the Bandung conference of recently decolonized Asian countries in 1955. In the nineteenth century, however, it was used more or less neutrally to characterize the condition of colonies and the (speech) habits of colonials (Fieldhouse, 1981: p. 6). One has no choice but to accept the change of meaning that colonialism has undergone, although one can try to neutralize political emotions. In this sense, colonialism can be defined as the control of one people by another culturally different one, an unequal relationship that exploits differences of economic, political, and ideological development between the two (Reinhard, 2008: p. 1). Even in European countries, unbalanced internal group relations as in the case of Britain's 'Celtic fringe' were qualified as internal colonialism (Hind, 1984). People instead of nation or state is used because basically no sophisticated political organization is necessary on either side. And the terms difference and development are used in a strictly descriptive sense and without any value judgment. They do not suggest that it is more desirable to have nuclear weapons instead of bow and arrow or that there exists a general and 'normal' path of human development with the West at the end. But differences of development are essential to distinguish colonial rule from empire in general. Roman rule over ancient Greece and Russian control of East Germany were imperial, not colonial.

Nevertheless, colonial empires are still a variety of empire in general. One uses empire worldwide to designate any large political unit before the modern state reached its maturity in Europe and started to expand from there. In contrast to the state's homogenous structure and consequential outstanding concentration of power, empires remained loosely structured and more often than not organized in concentric circles of decreasing control by the center. The Ottoman Empire, China, Siam, Madagascar, and initially also Japan remained empires in that sense, whereas during the eighteenth and nineteenth centuries their European counterparts became powerful modern states. Their notorious superiority enabled them to build up additional colonial empires around the hard core of European national power states. Only this exceptional unbalanced constellation made ephemeral European world hegemony possible. In addition, Western colonies in the Americas, in South Africa, and in Australia established a secondary colonialism in their parts of the world.

Colonialism, as an unequal relationship between human groups, very often was the outcome of imperialism, defined as a political activity with the intention to establish colonialism, in other words, a colonial empire. But if the meaning of imperialism is limited to expansive policy of the nineteenth and twentieth centuries, colonial expansion has to be used to designate earlier policies. Only the similarly unbalanced power relations between Amerindian stone age empires and Iberian monarchies allowed the formation of colonial empires in the New World as early as the sixteenth century.

During nineteenth-century European world hegemony colonialism changed, under certain conditions, into the varieties of informal control or even of informal empire. Semi-colonies such as China or the Ottoman Empire about 1900 were formally independent but in reality under the economic and therefore also political control of the West. In contrast, in the nineteenth century, West Africa under British paramountcy was a case of informal empire. After decolonization, colonialism as a form of political dependency has become a mere phenomenon of history. But economic and cultural domination by former colonial powers such as Britain and France on the one hand, and by the United States and the US-controlled world economy on the other, still exists. It has been labeled neocolonialism. In the 1960s and 1970s, theories of structural dependency explained economic underdevelopment as a consequence of Western economic domination and as self-reproducing without chance of escape because of Western control of the world economy. Several success stories of former

colonies such as Brazil or Taiwan have falsified these theories, but their lesson on informal control as an element of colonialism remains.

The History and Geography of Colonization and Colonialism

The expansion of Europe between the fifteenth and twentieth centuries was one of the most momentous processes of history. World history in the sense of global history became possible only through that process, because no part of the world escaped the direct or indirect impact of Europe. In 1914, half of the land surface of the earth and one-third of its population were still under direct colonial rule. But this was only the last chapter of a long story, because colonization and colonialism are essential components of history in general, especially, if one does not omit continental expansion in Russia, the Americas, and elsewhere in favor of overseas colonialism. Colonization in the sense of expansion of settlement and agrarian land use could even be considered the quintessence of human history before industrialization. Chinese colonialism is perhaps the most remarkable case because of the relative continuity of Han-Chinese expansion starting from the lower Huang basin thousands of years ago and leading to the present penetration of Tibet, Mongolia, and Xinjiang. The ancient world of the Mediterranean was full of Phoenician, Greek, and Roman settlement colonies.

Europe was expansive right from the outset; it was a creation of its own expansion. To a large extent, European history consists of successive waves of expansion. The Vikings were followed by colonization in Eastern Europe. People from central and Western Europe settled in the East, a process to be continued by Russia in Siberia until the Pacific Ocean was reached. The French who were excluded from the colonization of the East became protagonists of the crusades instead. The first European colonies established by the crusaders in today's Middle East were a failure. But in the same context, Italians developed a system of maritime trade and factories combined with the colonial production of high-value commodities such as sugar. For some time, they participated in a world trading system extending from Greenland to China. On the other hand, Spain and Portugal colonized the reconquered South of the Iberian Peninsula and the Canaries, whereas England did the same in Wales and Ireland. Such experiences of the Middle Ages became predecessors and models of the later expansion of Europe. Not by chance America was 'discovered' by a Genoese and 'baptized' after a Florentine in the service of Castile.

Beginning in 1415, the Portuguese followed the coast of Africa until they found their way to India in 1498. To control the spice trade of the Indian Ocean, they established a trade empire based on fortified ports, factories, and colonial rule over spice producers such as Ceylon and the Moluccas, extending from East Africa to China and Japan with Goa in the center. During the European conflicts of the early seventeenth century, the Dutch East India Company, in contrast to Portuguese crown capitalism a privileged private corporation of shareholders, conquered most of the Portuguese system, but established a new capital in Batavia on Java. Arriving at the same time, the French and English East India Companies created their bases on the coast of India, a comparative advantage when European demand changed in the seventeenth century and Indian textiles became the leading commodity. In the eighteenth century, they were replaced by tea and coffee with the China trade as core business. But trade remained under strict control of the Chinese and Japanese empires and 'Western barbarians' were only tolerated until the middle of the nineteenth century when China and Japan were 'opened' for free trade by Western military aggression. On Java and in India, however, the Mataram and Mughal empires began to disintegrate in the eighteenth century. This led to the control of Java by the Dutch and the conquest of all India by the British. Already in the later eighteenth century "the parallels with today's corporate leviathans soon became overpowering, with the Company outstripping Wal-Mart in terms of market power, Enron for corruption and Union Carbide for human devastation" (Robins, 2006: p. XI). Most of it was the result of contingency, but besides the interest in profits the ambition of local European military, the 'men on the spot,' was a constant factor. For some time, they could also rely on the superiority of soldiers drilled in the European way.

Discovery and colonialism went hand in hand down to the twentieth century. Since 1492, Castilians competing with the Portuguese discovered and conquered their India in the West. Spaniards emigrated to live in the cities of a New World. But despite a tremendous loss of native population, mostly through infectious diseases imported by the conquerors, the labor force consisted mostly of Amerindians, especially when, after the discovery of silver mines in northern Mexico and highland Peru, America became the treasure house of the Spanish monarchy. A protobureaucratic system of colonial government was established instead of traditional feudal lordship desired by the conquerors. In theory, the Spanish empire consisted of territories with equal legal status. The term colony was not used for overseas possessions before the eighteenth century. But Iberian America was to remain a racially stratified society. Metropolitan Spain did not profit much from American silver, which, because of differential inflation rates and overambitious policy, flowed out of the country to feed the Asian trade of the Dutch and the English. Via Europe and directly via the Philippines, much of the American silver ended up in India and China and favored economic and political growth there. A kind of world trade system originated, but probably with rather marginal economic impact on Europe and on Asia. Brazil became a Portuguese colony almost by chance, economically important only in the later sixteenth century when sugar cane was grown on its abundant fertile soils. Indian labor was scarce here but the Portuguese controlled the opposite coast and therefore were able to provide the sugar industry with African slaves. When the sugar cycle reached its end, gold and diamonds created a new boom in Brazil during the eighteenth century, but with similar negative consequences for Portugal as for Spain 200 years earlier.

In the seventeenth century, the Dutch transferred the Brazilian plantation economy to the Caribbean where the English, French, and Dutch had conquered several islands from Spain. No Amerindian labor force was left there, but the Dutch had occupied some of the Portuguese strongholds on the West African coast with the British and the French to follow. Thus, the West Indies could be provided with African slaves. The

English were the leading slave traders of the eighteenth century when the sugar business reached its culmination point. Large parts of the Caribbean were converted into an agroindustrial complex with a completely artificial society of black slaves, white lords, and an increasing number of mulattos.

In the meantime, another type of artificial new world had originated in North America, which in the long run became a white man's country, the first and most successful of several 'new Europes' created by about 60 million Europeans who emigrated between the sixteenth and twentieth centuries. North American Indians were few in number and not able to defend the land white immigrants desired. French Canada with its small settler population could at least attempt their integration, whereas British North America had no room for them. They were removed, marginalized, or wiped out. Around 1780, British North America had more than 2.6 million inhabitants, compared with about 8 million inhabitants of England and Wales.

In 1775–1823, Europe lost most of her American colonies. But Britain, which dominated the colonial scene for most of the nineteenth century, compensated for the loss with the conquest of India and with several new colonies, some of them acquired during the wars against France, like Canada and South Africa, others new foundations in Australia and New Zealand, which had been explored during the late eighteenth century. But they were accepted rather reluctantly, because in an age of British-dominated free trade colonial rule was considered unnecessary and too expensive. Therefore, self-financed self-government of white settler-colonies was now welcome. Between 1840 and 1931, these British 'dominions' became quasi-independent. This was the age of the so-called free trade imperialism when political pressure and military violence were used only exceptionally where free trade relations had to be enforced as in the cases of China and Japan (Gallagher and Robinson, 1953).

This relaxed attitude was to change in the age of imperialism, when steamships, deep sea cables, telegraph, and finally wireless transmission improved worldwide communication and Western superiority increased tremendously with military technology such as the machine gun (Maxim gun). Now, nervous competition of old and new powers under the ideological impact of nationalistic social Darwinism led to the division of Africa within a few decades based upon the legal fiction of *terra nullius* (land belonging to no one) and the right to take possession of such supposedly unoccupied territory. Most of what was left of Asia was also occupied or at least controlled economically. Governments were now much more engaged in colonial activities than ever before. Besides Britain's notorious rival France and the old colonial powers Portugal and Spain, new competitors entered the stage: Russia and the King of Belgium, Germany and Italy, the United States and Japan, the first non-Western imperialist power, which a few decades earlier had been close to becoming a semicolony. After World War I, the German colonial empire and the Ottoman Empire were divided among the victors. The formal qualification as mandates of the League of Nations did not make much difference. Neither was the League of Nation able to prevent the Japanese occupation of Manchuria in 1931–32 and the Italian conquest of Ethiopia in 1935. Ethiopia had defended herself successfully against Italian aggression in 1896, but as the last African country had to yield to colonialism – if only for a few years. Times were beginning to change.

Less spectacular than expansion overseas, and therefore often ignored by historians of colonialism, was the 'quiet' continental expansion of Russia and the United States; of Canada, Australia, and South Africa; and of Brazil, Argentina, and Chile. This expansion was possible only at the expense of indigenous populations and through the exploitation of differences of development. This turned out a very effective variety of colonialism, because the result was most often a white man's country, another 'new Europe,' where the former population was marginalized or wiped out. Only in few cases such as South Africa, Russian Central Asia, and Caucasia, non-Western groups were numerically and culturally strong enough for later successful decolonization.

Typology of Colonization and Colonialism

There was no master plan for colonial expansion. Quite often, it happened as unintentional consequence of some other action. Nevertheless, typical sequences of actions occurred again and again, for instance, preventive occupation to keep out possible competitors. The Portuguese used this in sixteenth-century Brazil, as did most European powers in nineteenth-century Africa. Military intervention in Africa was sometimes not intended as permanent occupation, because it was much too expensive. But when African resistance made a retreat impossible without loss of national prestige, the temporary occupation became a permanent colony. If not a matter of prestige, colonial policy was always based on a kind of cost-benefit analysis. Colonies were expected to be profitable. At least they had to finance their own government. The famous British model of indirect rule did originate much less from political wisdom or respect for other civilizations than from the necessity to keep administration costs down. In reality, national budgets only profited from colonial empires in exceptional cases, because infrastructural costs were considered the responsibility of the government, whereas colonial profits remained private.

The most common individual motives for participating in colonial activities were indeed the desire for extra profits and the improvement of social status. These were certainly not the only motives, but they were almost never absent. Of course, their character changed over time. The conquerors of British India were capital accumulating profit seekers, but like those of Spanish America still with the intention to invest in social status and not in capitalist enterprise in the way of the nineteenth and twentieth centuries. The second most common motive, missionary zeal, also changed over time. Instead of spreading the gospel and European civilization, modern colonizers followed the urge of a civilizing mission to educate barbarians. They fought slavery and tried to transform 'lazy natives' into industrious workers. In the age of imperialism, however, it became common conviction that inferior races could not be 'improved,' but were doomed to serve or to die. Thus, colonial adventurers could feel entitled to pose without restraint as the master race.

From the fifteenth to the twentieth century, colonies usually represented one of three basic types, sometimes as a mixture.

1. Trade and/or military bases such as the stronghold system of the Portuguese trade empire or Britain's naval bases in the nineteenth century or the frontier forts of the United States. Spanish ecclesiastical missions such as the Jesuit reductions of Paraguay sometimes served a similar purpose. Of course, systems of bases were also an essential part of informal empires.
2. Colonies of settlement, probably the primitive type of colony, especially when combined with a kind of stronghold as in the case of the Roman 'Colonia.' Larger settlement colonies could only be founded at the expense of indigenous people who were either removed (New England) or transformed into a dependent labor force for an 'improved' economy (Spanish America and South Africa) or replaced by workers of foreign origin (Africans in the Caribbean and East Indians in Guyana).
3. Colonies of exploitation with only a small number of members of the ruling people present most of them only temporarily, to run business, administration, and defense. Parts of colonial Latin America resembled that type, but it became the dominating type during nineteenth- and twentieth-century colonialism with British India as the prototype and much of Africa to follow. The administrative penetration of this type of colony tended to be rather limited.

In consequence, colonies of exploitation could not be ruled without massive collaboration from the indigenous people (Robinson, 1974). But collaboration had already been a necessary precondition for the Spanish conquest. In colonial wars, natives quite often fought on both sides. To consider such collaborators as traitors, however, is an anachronistic value judgment of modern nationalist historiography. People like the Tlazcaltecans quite rationally served their own interest regardless of a Mexican nation that did not yet exist. Recent research on the colonial situation from the perspective of people under colonial rule has corrected such simplifications. Non-European history no longer looks just like an inversion of the eurocentric pattern when the colonialist image of benevolent Western heroes ruling inferior races for their own benefit is simply inverted in the anticolonial story of Western rascals systematically abusing helpless non-Western victims. To some extent, the image of the helpless native itself is a creation of latent racism. In reality, a broad range of options to cope quite successfully with colonial domination was available to indigenous people, to some extent even to African slaves. The traditional contrast between colonial lords as active initiators of colonial processes and colonial subjects who had no choice but to suffer passively or at best to react against the actions of their betters has been questioned. But to believe after decolonization that colonial rule was only a short episode without much impact on the colonies seems to be an anticolonialist oversimplification, even in Africa.

Decolonization

The term decolonization was created in 1932 (Albertini, 1966: p. 28) but the process is much older. A first wave of decolonization, the independence movements of most American settler colonies and of the Afro-Americans of Haiti 1775–1823, reduced European colonial empires drastically. But Britain at least knew how to compensate for these losses. On the other hand, some lessons of the first wave were learned and made the second wave, the gradual increase of self-government in Britain's white dominions 1840–1931 much more easy going. Between the World Wars, anticolonial movements emerged in several exploitation colonies, with British India in the lead. Japanese occupation of Asian colonies during World War II fostered independence movements that were successful after the war in a third wave of decolonization in the 1940s and 1950s. At that time, Africa was apparently not yet ready for decolonization. Colonial powers once again tried instead to compensate for their losses in Asia by economic development of their possessions in Africa. However, under growing political pressure, some African countries became independent in the 1950s, and most in what was the fourth wave of decolonization during the 1960s. A fifth wave began in the Portuguese colonies in 1974–75 and accelerated political change in South Africa. But the sixth wave of decolonization in the 1990s, this time internal, included not only the end of South Africa's white minority regime, but also quite unexpectedly the breakdown of Soviet rule in Central Asia and Caucasia and Israel's first attempted arrangements with her Arabs.

Decolonization was a truly dialectical process, because colonial rule produced its own contradiction, which is politically frustrated new indigenous elites. Everywhere, Westernized groups and not the traditional authorities were in the lead of independence movements. But an independence movement or even an independence war was not enough. In most cases, a triangular constellation of political factors was the essential precondition; besides an internal independence movement, favorable international conditions such as allies, anticolonial policies of the world powers, and the United Nations, and the willingness of the colonial power to accept the unavoidable were also necessary.

The Consequences of Colonialism and the Postcolonial World

It is simply not true that colonial processes left no impact on former colonies. The postcolonial world is not the same as before, even in Europe. Several American plants, the potato in particular, became essential for Europe's population and economic growth. Colonial bullion and colonial profits contributed to the rise of Europe, if rather marginally and indirectly. They were certainly not the sufficient precondition of industrialization. Colonial problems influenced European power relations, but rarely dominated them. Europe assimilated new cultural elements from all over the world, but without changing her character. In the colonies, however, ecology, economy, politics, society, and culture often changed more thoroughly, sometimes even fundamentally. However, this does not imply a one-sided modernization directed by Europe but rather processes of interaction.

Intentionally or unintentionally, colonial rule has transformed ecological systems. After new plant foods had been introduced, American savannahs changed into corn fields suffering from soil erosion and new weeds. Domesticated animals formerly unknown in the Americas, Australia, and

New Zealand revolutionized working and eating habits, transportation systems, and even the whole way of life of such groups as the Indians of the North American plains. Unknown infectious diseases killed millions who were replaced by an artificial population of immigrants and imported slaves.

The deliberate diffusion of European and especially American cultivated plants such as maize and manioc over Asia and Africa created a new agriculture and allowed an expansion of food production, a necessary precondition of modern population growth in the colonies. But colonial economies were first of all based upon the systematic exploitation of mineral resources such as American silver, African diamonds, gold and other ores, and more recently of oil in the Middle East and other parts of the world. In addition, they specialized in the production of high-value agrarian products such as sugar, coffee, and tea on large plantations. Colonial railways were a harbor-centered instrument of this export economy. Mines and plantations more often than not ruthlessly exploited a colonial labor force of slaves and contract and forced laborers. Because, as a rule, colonies could not industrialize, their independent successor states remained dependent on the world market price of their raw materials, which only in the case of oil took a turn to their advantage with lasting consequences.

The essential political heritage of colonialism is the modern state with its legislative, administrative, judicial, military, and educational apparatuses. Every former colony has changed into a modern nation-state, but strictly within the arbitrary and sometimes absurd colonial border lines. Therefore, some are extremely poor and very few are powerful enough for an independent foreign policy. But in contrast to former periods of history, weak states today have their existence guaranteed by the international community (Jackson, 1990). Internally, only few of them can build upon the tradition of a historical community. And instead of a modern civil society, their colonial heritage is ethnic fragmentation, to some extent artificially created by colonialism. Quite often, they had no alternative to the language of the former colonial power as national language. Under these circumstances, the modern state in general and Western democracy in particular could not become a success and were replaced by the parastateliness of self-organizing groups based on remarkable combinations of traditional and modern elements. Without generally accepted legitimacy and reliable civil servants, the political leaders had little other choice but to rely upon a neo-patrimonial system of clients and upon the military.

Everywhere, colonialism has created new social groups living in Western-style urban environments: workers and employers, school teachers and professionals, and civil servants and professional soldiers. Women have found new roles and quite often emancipated themselves from traditional restrictions. All this is based on Western science and technology, on Western ideas and ideologies like rationalism and individualism, the rule of law and human rights, and Christianity and socialism. From this state of things, so-called postcolonialists have drawn the conclusion that political decolonization may be more or less finished and economic decolonization well under way, whereas mental or cultural decolonization has not even begun because nowadays nobody outside Europe is able to think in a non-European way. Even attempts to deconstruct this state of things have no choice but to employ intellectual instruments of European origin. However, paradoxically their very assessment of the interaction of colonizers and colonized in terms of hybridity, of third space, of in-between demonstrates that it is completely inadequate to assume a lingering continuity of mental colonialism. Because this state of things no longer implies cultural dependency, but rather the opposite: non-Western people in the meantime have appropriated a cultural heritage of Western origin. For example, the role of English as common language of the world stems from historic British and American dominance. Increasingly, however, the acceptance of 'new Englishes' all over the world indicates that the Anglo-Saxons have been expropriated of their language.

See also: Civil Society, Concept and History of; Civilization, Concept and History of; Communication and Transportation, History of; Frontiers in History; Global History: Universal and World; Meiji Revolution, The; Racism, History of; Slaves and Slavery, History of; Warfare in History.

Bibliography

Albertini, R.v., 1966. Dekolonisation. Die Diskussion über die Verwaltung und Zukunft der Kolonien 1919–1960. Westdeutscher Verlag, Köln-Opladen (1971. Decolonization: The Administration and Future of the Colonies, 1919–1960. Doubleday, Garden City, NY).

Albertini, R.v., 1987. Europäische Kolonialherrschaft 1880–1940, third ed. Steiner, Stuttgart. (1982. European Colonial Rule, 1880–1940: The Impact of the West on India, Southeast Asia, and Africa. Greenwood Press, Westport).

Burbank, J., Cooper, F., 2010. Empires in World History. Power and the Politics of Difference. Princeton UP, Princeton and Oxford.

Childs, P., Williams, R.J.P., 1997. An Introduction to Post-colonial Theory. Prentice Hall, London/New York.

Fieldhouse, D.K., 1981. Colonialism 1870–1945. An Introduction. Weidenfeld and Nicholson, London.

Forsyth, J., 1992. A History of the Peoples of Siberia, Russia's North Asian Colony, 1581–1990. University Press, Cambridge.

Gallagher, J., Robinson, R., 1953. The imperialism of free trade. The Economic History Review VI (1), 1–15.

Hind, R.J., 1984. The internal colonial concept. Comparative Studies in Society and History 26, 543–568.

Histoire de la Colonisation Française, vol. 2, 1991. Fayard, Paris.

Histoire de la France Coloniale, vol. 2, 1990–1991. Colin, Paris.

Holland, R.F., 1985. European Decolonization, 1918–1981. Macmillan, Basingstoke.

Jackson, R.H., 1990. Quasi-states. Sovereignty, International Relations, and the Third World. University Press, Cambridge.

Lockhart, S., Schwartz, S.B., 1983. Early Latin America. A History of Colonial Spanish America and Brazil. University Press, Cambridge.

Milner, C.A., O'Connor, C.A., Sandweiss, M.A. (Eds.), 1994. The Oxford History of the American West. Oxford University Press, New York.

Olson, J.S. (Ed.), 1991. Historical Dictionary of European Imperialism. Greenwood Press, Westport and New York.

Osterhammel, J., 1995. Kolonialismus. Geschichte - Formen - Folgen. C. H. Beck, München (1997. Colonialism. A Theoretical Overview. M. Wiener, Princeton).

Reinhard, W., 1983–1990. Geschichte der Europäischen Expansion, vol. 4. Kohlhammer, Stuttgart.

Reinhard, W. (Ed.), 1999. Verstaatlichung der Welt? Europäische Staatsmodelle und Außereuropäische Machtprozesse. Oldenbourg, München.

Reinhard, W., 2008. Kleine Geschichte des Kolonialismus, second ed. Kröner, Stuttgart. (1996. first ed.; 1997. Petite Histoire du Colonialisme. Belin SUP, Paris; 2002. Storia del Colonialismo. Einaudi, Torino; 2011. A Short History of Colonialism. Manchester UP, Manchester).

Robins, N., 2006. The Corporation that Changed the World: How the East India Company Shaped the Modern Multinational. Pluto Press, London.

Robinson, R.E., 1974. Non-European foundation of European imperialism. In: Owen, R., Sutcliffe, B. (Eds.), Studies in the Theory of Imperialism. Longman, London.

The Oxford History of the British Empire, vol. 5, 1998–1999. University Press, Oxford.

Color Classification and Symbolism

Anita Jacobson-Widding, University of Uppsala, Uppsala, Sweden

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 4, pp. 2245–2248, © 2001, Elsevier Ltd.

Abstract

The history of research into color classification is traced, looking at both the variation of linguistic categorization *of* colors in different cultures and at the work of anthropologists concerned with classification *by* colors and its symbolic meaning. The triadic classification by colors (black, white, and red), which is especially prevalent in sub-Saharan Africa, is discussed. The color triad is interpreted as a device for philosophical categorizations and the 'magic' or 'sacred' power of red is noted in other contexts – such as within the Christian church. However, any combination of colors from the triad may have different concrete references in different situations in both Europe and Africa.

'Color classification' is an ambiguous concept. It may refer to 'classification *of* colors,' or 'classification *by* colors'. Most research on color classification has opted for the first approach, which is concerned with the variation of linguistic categorization of colors in different cultures. This kind of research has been undertaken by linguists, psychologists, and cognitive anthropologists. By contrast, the few anthropologists who have opted for the second approach have been concerned rather with symbolic anthropology, or have worked in a structuralist tradition of sociocultural anthropology. Research into color classification has a long history, dating back to the 1870s (Allen, 1879; Gladstone, 1877; Magnus, 1877). However, it was not until Berlin and Kay published their seminal book on basic color terms (1969) that research in this field virtually exploded, especially within cognitive and linguistic anthropology. Over the last 25 years, a wealth of data on color categories has been meticulously collected in different parts of the world, mainly with the help of the 'Munsell color chips.' As a general rule, randomly chosen informants have been asked to define the colors on these chips, and to 'map' them. The findings have then been tested as 'hard data,' often using quantitative methods (e.g. MacLaury, 1986; Kay et al., 1991). This 'hard-wire' approach to color classification research has been further strengthened by the general assumption that color perception is a matter of neurobiology, universality, and evolution.

Although the evolutionary paradigm had featured in research on color classification for almost 100 years, Berlin and Kay were the first scholars to present substantial evidence for a theory concerning the universality and evolution of basic color terms. In the original formulation of their theory, they defined a particular sequence of developmental 'stages' of basic color terms, and suggested that these stages would correspond to the stages of technological evolution. The few cultures in which only two basic color terms were used would represent Stage I on the evolutionary ladder, while the prevalence of three color terms would place a particular culture one step higher up (Stage II). Four basic color terms would imply Stage III, five terms Stage IV, and so on. Societies with the maximum number of color terms (11) would represent the highest level of technological evolution.

Although Berlin and Kay have modified the original formulation of their theory (Kay et al., 1991), the main features of the paradigm have been generally accepted by scholars in their own and related fields. Part of the reason for this is that Berlin and Kay backed up their theory with a wide range of cross-cultural linguistic data, which showed a remarkable regularity with respect to the sequence of basic color terms: if only two terms were found in a particular culture, they were invariably white and black. If three terms were recognized, these were always white, black, and red. The fourth basic color term to be introduced was either yellow or green, the fifth either green or yellow, the sixth blue, and so on. In spite of some revisions and later developments of the theory (see Hardin and Maffi, 1997; MacLaury, 1999), there is still a general consensus among anthropologists about the sequence of basic color terms – even though the hypothesis about a corresponding technological evolution has yet to be proved. It would, for example, be difficult to maintain the theory about technological evolution, given that many hunting/gathering societies recognize more basic color terms than agriculturalists. In Africa, even the most sophisticated Bantu-speaking agriculturalist and agropastoral societies have only three basic color terms, whereas their hunting and gathering neighbors may have five or more.

However, strictly cognitive and linguistic research on color classification has met with criticism on several other accounts. The main criticism has been leveled at the 'hard-wire' methods employed in this kind of work, and at the lack of 'emic' approaches to the issue of color classification. It has been argued that the 'meanings' of color terms cannot be evaluated adequately when the data have been collected under conditions reminiscent of a laboratory situation (Saunders, 1999; cf Tornay, 1978), and that 'affective features of meaning' may contribute to Berlin and Kay's evolutionary sequence of color terms (Osgood et al., 1975; Taft, 1997). The problem of 'meaning' is perhaps the most controversial issue associated with the concept of color classification. Though a few representatives of the 'hard-wire' approach have attempted to account for the 'meanings' of individual color terms (Kay and McDaniel, 1978; MacLaury, 1992), the focus has still been on the implications of the informants' task of 'mapping' color chips (Lucy, 1997: p. 327). By way of contrast, Lucy directs attention to Conklin's pioneering work on color categories in the 1950s (Conklin, 1955). Applying a structural approach, Conklin discovered that the four-color categories recognized by the Hanunóo corresponded to an underlying four-way classification of other visual stimuli as well.

Conklin's findings correspond to the ideas proposed by the Swiss-French school of semiology, as represented by Saussure and Derrida. According to them, the meanings of signs depend on their positions in a larger system of signs (Layton, 1997). Meanings are thus anchored in the syntagmatic relations among signs, but if we also take into consideration the meanings which Barthes (1967) has referred to by the concept of 'the symbolic relation between the sign and the signified,' it may be possible to identify congruities of symbolic meanings when 'parallel' syntagmatic sequences of signs are compared. Such a 'parallel' structure of signifying, syntagmatic sequences of signs may be referred to as a paradigmatic structure of meaningful categories.

To identify such paradigmatic structures in cultural symbolism is the core of anthropological research on symbolic classification. This approach has characterized symbolic anthropology since the late 1960s. When colors have been included in this kind of symbolic analysis, the research has thus been concerned with classification *by* colors, rather than *of* colors. However, at the same time as this structurally oriented research developed, there was also a growing awareness among some anthropologists of the importance of contextualization in the study of symbolic meanings.

The pioneering research in the domain of symbolic classification by colors was carried out by Victor Turner, who in 1966 wrote an article on the symbolic meanings of white, black, and red among the matrilineal Ndembu of Zambia (Turner, 1966). Not only did Turner attempt to trace the basic symbolic meanings of the three colors, he also analyzed their variable references in different ritual and social contexts, and thus became one of the first anthropologists to put contextual analysis of cultural symbols on the agenda.

Turner's article stirred up an enthusiastic interest in color classification, especially among anthropologists working in the structuralist tradition. No matter whether the anthropologists found two, three, or four colors to be key symbols in a particular culture, they usually attempted to identify a dualistic system of symbolic classification, of which the colors formed a part. Thus, for instance, most of the contributions to Needham's book on dual classification included references to colors as 'parallel' devices of classification, alongside the right/left division that was the main topic of the book. While the concept of 'paradigmatic structures of signs' was not mentioned, there was an insistence on the congruity between different levels of symbolic classification. In many articles referring to different areas of the world, colors were regarded as part of this dual classification.

Needham's book was cast in the kind of structuralist tradition that marked the 1970s, and which was inspired by Levi-Strauss (1963). The basic assumption was that human beings everywhere tend to classify their world into binary oppositions. However, even if a dualistic approach could account for the choice of two opposing categories of classification in certain specific situations in any society, it cannot solve the enigma of triadic classification, which is typical of the ritual use of colors in many parts of the world.

In most African cultures south of the Sahara, triadic classification by colors appears to permeate the ritual practice of healers and diviners (Baumann, 1935; Straube, 1960; White, 1961; Ngubane, 1977; Jacobson-Widding, 1979; Kirby, 1999). This applies particularly to the Bantu-speaking societies, where only three basic color terms are recognized, namely white, black, and red. In a systematic investigation of the symbolic references of the three colors in the matrilineal belt of Central Africa, Jacobson-Widding (1979) found that the only situations in which all three colors were visually represented were in connection with certain rituals, when the ritual specialist manipulated the colors as symbols with particular, contextually dependent references. In profane contexts in which the official ideology of social and moral order was referred to as a matter of principle, on the other hand, white and black were the only colors demonstrated (cf Jacobson-Widding, 1991). In Central Africa, there is even a formal prohibition against showing the color red in such contexts, for instance in connection with lawsuits. This implies that white and black are contrasted as binary oppositions only in certain contexts, like those in which the dominant structure of order is supposed to be manifested, and in which the rationality of a logic based on two opposite values is to be applied. In contrast, red is represented in situations where indeterminate values are dealt with, and where a simplistic two-value logic would fail to account for the truth.

The concrete references of the color red may vary according to the context. However, whatever tangible phenomenon it may refer to in any particular situation, most ritual experts in central and southern parts of Africa will employ this color when they wish to represent ambiguity. In particular, red is used to represent phenomena that seem to defy classification into any of the 'officially' recognized categories of determinate values that are represented by white and black. Such phenomena, or states of affairs, may consist of bewildering emotions, anomalies, or mysterious spirits. Or they may be 'twilight' situations like dusk and dawn, birth and death, or any state of 'betwixt and between' – such as the state of initiands in a bush camp, or the state of adolescent girls, whose bodies are painted red during a period of seclusion. In essence, the color red can be viewed as a symbol of the kind of ambiguity that may result from conceptual boundary crossings. That is, when a particular phenomenon, or state of affairs, is considered to blur the boundaries of the concepts on which a 'rational' order is based, it is depicted as 'red.' This implies that the color red is associated with irrational feelings, and with the unknown.

Considering the ways in which ritual experts in Africa tend to interpret and manipulate colors, we may regard the color triad as a device for philosophical categorizations. At the most general level of meanings, the color triad appears to classify different kinds of phenomena according to three basic values of truth and morality. Whereas the combination of white and black is supposed to represent clear-cut values as logical opposites (i.e. 'good' vs 'bad,' 'true' vs 'false,' 'us' vs 'them,' 'sky' vs 'earth,' and so on), red will be introduced when the logical boundary distinguishing A from B is impossible to define. The color red will thus represent a third, indeterminate value, which can be classed neither as A, nor as B (nor as non-A), but rather as C (cf Leach, 1975).

This implies that the two-value logic of the Aristotelian philosophy of knowledge is supplemented with a third term, which in the case of African ritual practice is regarded as an *a priori* neutral value, in relation to the determinate values of truth and morality. However, this 'three-value logic' is assumed to be valid only in ritual thought and praxis, and not in

contexts where profane, 'rational' reasoning is supposed to be applied.

Since red is used to symbolize an unknown factor, which by definition is initially value-neutral, this color will be associated with unknown potentialities and acquire a dynamic quality. Accordingly, it is sometimes associated with powerful emotions, sometimes with 'magic' power – a magic power that many people may regard as dangerous, polluting, or taboo, whereas others may associate it with miracles and miraculous healing. It is this 'magic' power that is exploited by the ritual expert in Africa, whenever he or she manipulates the color red in any of the many possible combinations or oppositions with white or black. As a general rule, the ritual expert will begin any 'sacred,' ritual seance with a visual demonstration of all three colors, as if to show the cards he or she intends to play with.

If we understand the concept of 'magic power' in a broad sense, some of the African associations of the color red may also be discerned in sacred rituals outside Africa, even within the Christian church in Europe. When red is combined with white, it usually points to connotations of the 'sacred,' or the 'holy.' Correspondingly, we often find that the combination of white and black is resorted to in situations where the order of society, or its moral dogma, is supposed to be demonstrated – not only in Africa, but in Europe and other parts of the world as well. However, whether used to represent 'sacred' values in the church or 'profane' values in the courtroom, any combination of colors from the triadic set of symbols constituted by white, black, and red may have different concrete references in different situations, in both Europe and Africa. The 'meanings' of these colors as culturally instituted symbols are thus to be found on different levels of interpretation. Colors may be used as concrete signifiers with context-dependent meanings and, at the same time, as representations of more general values of 'truth' and 'power.'

See also: Color Terms, Linguistics of; Visual Memory, Psychology of.

Bibliography

Allen, G., 1879. The Colour Sense: Its Origin and Development. Trubner, London.
Barthes, R., 1967. Elements of Semiology. Cape, London.
Baumann, H., 1935. Die Lunda. Bei Bauern und Jägern in Inner-Angola. Reimer, Berlin.
Berlin, B., Kay, P., 1969. Basic Color Terms. Their Universality and Evolution. University of California Press, Berkeley.
Conklin, H., 1955. Hanunóo color categories. Southwestern Journal of Anthropology 11, 339–344.
Gladstone, W.E., 1877. The color sense. Nineteenth Century 2, 366–388.
Hardin, C.L., Maffi, L. (Eds.), 1997. Color Categories in Thought and Language. Cambridge University Press, New York.
Jacobson-Widding, A., 1979. Red-White-Black as a Mode of Thought. A Study of Triadic Classification by Colours in the Ritual Symbolism and Cognitive Thought of the Peoples of the Lower Congo. Uppsala Studies in Cultural Anthropology 1. Almqvist & Wiksell International, Uppsala, Sweden.
Jacobson-Widding, A., 1991. The encounter in the water mirror. In: Jacobson-Widding, A. (Ed.), Body and Space. Symbolic Models of Unity and Division in African Cosmology and Experience. Uppsala Studies in Cultural Anthropology 16. Almqvist & Wiksell International, Uppsala, Sweden.
Kay, P., Berlin, B., Merrifield, W.R., 1991. Biocultural implications of systems of color naming. Journal of Linguistic Anthropology 1, 12–25.
Kay, P., McDaniel, C., 1978. The linguistic significance of the meanings of basic color terms. Language 54, 610–646.
Kirby, J.P., 1997. White, red and black: Color classification and illness management in Northern Ghana. Social and Medicine Science 44 (2), 215–230.
Layton, R., 1997. An Introduction to Theory in Anthropology. Cambridge University Press, London.
Leach, E., 1975. Culture and Communication. The Logic by which Symbols are Connected. Cambridge University Press, Cambridge.
Lévi-Strauss, C., 1963. Structural Anthropology. Basic Books, New York.
Lucy, J., 1997. The linguistics of 'color.' In: Hardin, C.L., Maffi, L. (Eds.), Color Categories in Thought and Language. Cambridge University Press, New York.
MacLaury, R.E., 1986. Color in Meso-America. A theory of composite categorization. Doctoral Dissertation. University of California, Berkeley.
MacLaury, R.E., 1992. From brightness to hue: An explanatory model of color-category evolution. Current Anthropology 33, 137–186.
MacLaury, R.E., 1999. Basic color terms: Twenty-five years after. In: Borg, A. (Ed.), The Language of Color in the Mediterranean. Almqvist & Wiksell International, Uppsala, Stockholm, pp. 1–37.
Magnus, H., 1877. Die geschichtliche Entwicklung des Farbensinnes. Veit, Leipzig.
Needham, R. (Ed.), 1973. Right & Left. Dual Symbolic Classification. The University of Chicago Press, Chicago.
Ngubane, H., 1977. Body and Mind in Zulu Medicine. Academic Press, London.
Osgood, C.E., May, W.H., Miron, M.S., 1975. Cross-cultural Universal of Affective Meaning. University of Illinois Press, Urbana, IL.
Saunders, B., 2000. Revisiting 'Basic Color Terms.' Journal of the Royal Anthropological Society 6, 81–99.
Straube, H., 1960. Gedanken zur Farbensymbolik in afrikanischen Eingeborenenkulturen. Studium Generale 1, Göttingen.
Taft, C., 1997. Generality Aspects of Color Naming and Color Meaning. Department of Psychology, Göteborg University.
Tornay, S. (Ed.), 1978. Voir et nommer les couleurs. Laboratoire d'Ethnologie et de Sociologie Comparative, Nanterre.
Turner, V., 1966. Colour classification in Ndembu ritual. In: Banton, M. (Ed.), Anthropological Approaches to the Study of Religion. ASA Monographs, vol. 3. Tavistock, London, Humanities Press, New York, pp. 47–84.
White, C.M.N., 1961. Elements in Luvale Beliefs and Rituals Rhodes-Livingstone Papers 32.

Color Terms, Linguistics of

Paul Kay, University of California, Berkeley, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 4, pp. 2248–2252, © 2001, Elsevier Ltd.

Abstract

The study of the linguistics of color terms has mostly been concerned with isolating the 'basic' color terms of a language: the smallest set of linguistically simplex expressions of the language with which a speaker can name any color.

The words that different languages use to name colors have furnished a prime focus for research on the doctrine of *linguistic relativity*. This doctrine holds that (1) different languages categorize the world very differently and (2) the nonlinguistic thought and behavior of different peoples are deeply influenced by these differences in linguistic categorization. Words for colors have been considered an ideal area of lexicon for testing this doctrine because there exist scientific color order systems, which provide a precise metalanguage for describing color sensations independent of the words of any natural language. Such a metalanguage furnishes the means to characterize precisely the sameness or difference in meaning of color words in different languages.

Restricting the Domain

It is clear, however, that every language has immense – probably infinite – resources for denoting color sensations. First, every language so far studied can productively create complex expressions, which describe the color of an object by indicating similarity to another kind of object that is known to saliently display the color. An example from English is 'like the color of a male mallard duck's neck feathers.' (Although most languages do not have a word that straightforwardly translates the English word 'color,' the speaker of such a language who wishes to characterize, say, a kind of fish as mallard green can usually say that, for example, the fish 'looks like' the neck feathers of a male mallard, relying on context and background knowledge to provide the interpreter with color as the relevant dimension of visual similarity. Although relativist critiques of color naming research have maintained that absence of a word meaning 'color' in a language indicates that its speakers see and/or conceive of colors very differently from, say, speakers of European languages, the experiences of workers who have actually studied color vocabularies in the field do not provide significant support for this claim See, e.g., Conklin, 1955: p. 340; Turton, 1980: p. 322). Second, many language have words like 'blond' or 'bay,' which denote a color only as restricted to a class of objects (hair and horses, respectively). Third, many languages have morphological and syntactic processes which create complex color terms out of simple color terms. English examples include 'blue–green,' 'yellowish,' 'pale reddish purple,' and so on. Finally, some languages, English included, have many linguistically simple terms which denote subtypes or 'shades' of colors denoted by other simple terms. For example, 'scarlet,' 'crimson,' 'vermillion,' 'puce,' 'magenta,' 'burgundy,' and 'maroon' are among the more commonly named shades of 'red.' The linguistics of color terms has become a focus of research and a source of controversy in part because of the practical need to arrive at a manageable set of 'color terms' in each language to enable comparative study.

Color Terms and Linguistic Relativity

Before taking up in greater detail the ways researchers have attempted to delimit the linguistic domain of color, it is useful to consider briefly the role color naming has played in the long-running debate over linguistic relativity versus semantic universals (q.v.). In the second half of the nineteenth century it had become common knowledge among scholars that the major color words of different languages were not always perfectly translatable into one another. To some, this represented prima facie evidence of biologically based, perceptual differences among different groups of human beings, in fact the last traces of Darwinian evolution observable among humans. For example, William Gladstone, Homeric scholar and Prime Minister of Britain, observing differences between the color words of classical Greek and the modern European languages, wrote,"… the organ of color and its impressions were but partially developed among the Greeks of the heroic age" (1858: pp. 457–499).

Although this view was widespread in the late nineteenth century, it was not unanimous. For example, in order to study the discrimination and naming of colors independently, the German ophthalmologist Hugo Magnus circulated a questionnaire to missionaries and traders with 10 standardized color samples and instructions for using them in the 1870s. These instructions contained an explicit warning that failure of a language to distinguish lexically between two colors did not necessarily imply that speakers of that language did not distinguish the two colors perceptually. Magnus received completed questionnaires on 25 African, 15 Asian, 3 Australian, and 2 European languages. He concluded in part, "As regards the range of the color sense of the primitive peoples tested with our questionnaire, it appears in general to remain within the same bounds as the color sense of the civilized nations. At least, we could not establish a complete lack of the perception of the so-called main colors as a special racial characteristic of any one of the tribes investigated for us. We consider red, yellow, green, and blue as the main

representatives of the colors of long and short wavelength; among the tribes we tested not a one lacks the knowledge of any of these four colors" (Magnus, 1880: p. 6, as trans. in Berlin and Kay, 1969: p. 141). Magnus did find widespread *lexical* neutralization of green and blue, that is, a single word covering both these colors, as have all subsequent comparative studies of color lexicons.

While nineteenth century and early twentieth century researchers saw differences among vocabularies in an evolutionary perspective (biological and/or cultural), in the second quarter of the twentieth century the tide began to turn against evolutionary explanations of human variation and toward cultural and linguistic relativity. This was especially true in America, in the tradition of the anthropologists and linguists F. Boas, E. Sapir, and B.L. Whorf, although there were precursors in the views of the German romantics, who had tended to associate the *geist* of a people strongly with its language. According to Whorf, "The categories and types that we isolate from the world of phenomena we do not find there because they stare every observer in the face. On the contrary the world is presented in a kaleidoscopic flux of impressions which have to be organized in our minds. This means, largely, by the linguistic system in our minds" (1956: p. 212). Throughout the 1930s and 1940s this view gained increasing prominence and in the 1950s and 1960s achieved virtual hegemony in American anthropology and linguistics, with considerable influence in Britain and elsewhere outside of the US. (Chomskian universalism and nativism were to become influential only at the end of this period.)

The parade example chosen by the mid-century American relativists to make the twin points of semantic arbitrariness and linguistic determinism was the lexicon of color. The most influential introductory linguistics textbook of the period, for example, stated: "There is a continuous gradation of color from one end of the spectrum to the other. Yet an American describing it will list the hues as red, orange, yellow, green, blue, purple, or something of the kind. There is nothing inherent either in the spectrum or the human perception of it which would compel its division in this way" (Gleason, 1961: p. 4). This sweeping conclusion appears to have been based on two small-scale empirical studies of color terminology, whose results have subsequently been interpreted differently, one on the Hanunoo language (Philippines) (Conklin, 1955) and the other a comparison of the color terminologies of languages (Ray, 1953). Despite the paucity of evidence, the idea that different languages divide up the perceptual color space in mutually incomparable ways was widespread at the time, doubtless because it fits so well with the prevailing doctrine of radical cultural and linguistic relativity.

Color Terms as Linguistic Objects

Against this background a study appeared in 1969, which reached exactly contrary conclusions. After examining the color term systems of 98 languages, Berlin and Kay (1969) (B&K) proposed that (1) the differences in color lexicons of the languages of the world could be explained as different combinations of a few landmark colors (which were probably human perceptual universals) and (2) the differences in color lexicons of the world's languages could be arranged in an evolutionary progression. In order to carry out this study, B&K found it necessary to limit the domain of inquiry to the 'basic color terms' (bcts) of the languages being compared. (They examined 20 languages experimentally, using standard color stimuli, and interpreted another 78 reports of color terminologies from the literature based on their formulation of the results of these 20.) As a preliminary to this research, B&K offered what they called a definition of 'basic color term' and much of the subsequent research on the linguistics of color terms has had to do with attacking, defending, revising and evaluating this definition.

The B&K study was not without direct precursors. Lenneberg and Roberts (1956) (L&R) compared the color terms of Zuni monolinguals with those of Zuni–English bilinguals and English monolinguals. After isolating the words whose denotation was to be studied, L&R had each informant indicate on a representative array of 230 color samples the denotation of each color word as well as the color sample which best exemplified the reference of that word. B&K used essentially the same method and the same stimulus array as L&R (except for adding black, white, and shades of gray) with the exception that the color words whose denotation was to be studied were selected differently by L&R and by B&K. L&R proceeded by getting speakers to list as many color words as they could think of without color stimuli present, which produced a list of 52 terms. They also elicited lists of color words by asking for names for 24 selected color stimuli, producing a list of 105 terms. The color terms for which L&R show mappings are those most frequently elicited by the second method. L&R show mappings for seven color terms for Zuni monolinguals, corresponding closely to English pink, red, brown, yellow–orange, green, blue, and purple (recall they had no white, black, or gray stimuli). They show mappings for eight color terms for Zuni bilinguals, adding a term for orange, 'olechin-anne' ('orangelike' < English 'orange' + 'anne' 'like') and retracting the area of yellow accordingly. They show 10 mappings for English color terms (pink, red, orange, brown, yellow, green, light green, blue, light blue, purple). No terms like greenish, blue–green, or crimson were mapped, but light green and light blue were apparently ruled in by naming frequency considerations. (L&R's text is not explicit on this point.) Clearly in L&R, as in Conklin's, Ray's, and other early investigations of color word systems some tacit, if hazy, notion of major color term or principal color term is at work. It was this tacit notion that B&K tried to operationalize in their 'definition' of bct.

B&K's Definition of bct

B&K offered what they unfortunately called a definition of bct. Their intent was not to postulate and define a new theoretical entity but to operationalize, insofar as possible, a tacit concept that was already in general use. The guiding intuition of the B&K 'definition' of bct was that each language has a small set of simplex lexemes (or word *senses*, as in the case of English 'green (color)' and 'orange (color)') of pure color meaning (i.e., unlike words such as 'blond'), which collectively name all the color sensations. The bcts are the members of that set. However, B&K stated the definition in terms of eight criteria,

four primary and four subsidiary, the latter to be used in cases where application of the former is not sufficient to decide the bcts of a language. The four primary criteria for a bct were that it (1) be monolexemic (its meaning not predictable from the meanings of its parts, eliminating 'greenish,' etc.), (2) not be a hyponym (eliminating, e.g., 'crimson'), (3) be predictable of any kind of object (eliminating, e.g., 'blond'), and (4) be 'psychologically salient' (eliminating e.g., 'puce' and 'like the neck feathers of a male mallard duck'). Two telling criticisms that have been leveled against the fourth criterion are (1) that it is not a linguistic criterion, while the other three are and (2) that 'psychological salience' is neither a conceptually limpid nor a readily measurable notion. The four subsidiary criteria, designed for use in *ceteris paribus* situations, were that a questionable term (1) have a morphological distribution similar to that of the established bcts, (2) not be the name of a kind of object (which would eliminate English 'orange' if subsidiary criteria were allowed to override primary criteria), (3) not be a recent borrowing from another language, and (4) not be morphologically complex. The B&K definition of bct has led both to practical problems in determining the bcts of a language and to conceptual and theoretical problems regarding the semantics of color in natural languages.

Practical Problems with the B&K Definition of bct

The main practical problem in applying the B&K criteria to given languages is that since language change is always gradual there are many transitional cases. For example, there are many languages with bcts for black, white, red, yellow, and 'grue,' i.e., a single word covering green and blue. There are also many languages with bcts for black, white, red, yellow, green, and blue. When a language changes from the first type to the second type, it is not the case that everyone goes to bed Sunday night speaking the 'grue' language and gets up Monday morning speaking the 'green' versus 'blue' language. Many kinds of in-between situations arise that make it difficult to say which words of the language naming part or the entire green/blue region of color space are bcts. Moreover, there is virtually always widespread interspeaker variation in such cases, complicating further the task of deciding the bcts of 'the language' (*see* Sociolinguistics). Second, the four primary criteria of B&K do not always correlate and the definition gives no clue as to how the different criteria should be weighted in such cases. Also, as mentioned, cultural salience is not a straightforwardly observable characteristic. Workers who have taken the B&K 'definition' as a recipe rather than a heuristic have complained of problems in application in particular cases. Usually, however, these problem cases have involved a single term or a single area of the color space, indicating that the area is one undergoing change.

A third practical issue has arisen in connection with the first B&K criterion of nonpredictability of meaning of a bct from the meaning of its parts. It has been noted that in several Oceanic and Australian languages many of the presumably bcts are in fact reduplicated forms of roots that denote objects notoriously exhibiting the corresponding color. For example, in Yélîdnye (Rossel Island, Melanesia (Non-Austronesian)), the word for white *kpaaîkpaapî* represents reduplication of the nominal root *kpaapî* denoting a pure white cockatoo species. A similar situation holds for the black and red words, the only other plausible candidates for bcts in this language (Levinson, 2000). The question arises whether *Yélîdnye*, and similar languages, can be said to have any bcts at all. The answer is yes. The criterion of nonpredictability of meaning requires not only that from knowing the rest of the *Yélîdnye* language – including the word *kpaapî* and the pattern of reduplication which yields a salient attribute of the denotatum of a noun from reduplication of the noun stem – one could as a *listener* guess the meaning of *kpaapîkpaaî*, but also that as a *speaker* one would know an advance that *kpaapîkpaaî* was the way to express the idea of white. The latter of course would not be possible without memorizing this particular fact, because not all attributive words are formed by reduplication and even if they were the meaning 'white' might be based on some other object or substance.

Theoretical Issues Concerning the B&K Definition of bct

It has been argued, principally by John Lucy (e.g., Lucy, 1997), that a set of bcts isolated by B&K or associated researchers does not correspond to any linguistically legitimate system because there is no insistence in the B&K tradition of research that the set of bcts for a given language be definable by strictly formal, distributional criteria. This critique of the bct concept and practice is based on the questionable assumption that all morphosyntactic categories correspond to semantic categories and vice versa (see Kay, 1999). Moreover, careful empirical studies have shown that application either of the explicit B&K criteria or of the underlying intuition of a small set of linguistically simplex lexemes jointly exhaustive of the color domain can isolate sets of bcts which obviously form a coherent semantic system and whose members belong to several different form classes (see, for example, Maffi, 1990 on Somali). It has also frequently been found that morphological category, including native versus borrowed origin, can correlate with the evolutionary sequence posited by B&K ({black, white} > {red} > {yellow, green} > {blue} > {other}). The Somali case is again illustrative: the words for black, white, and red are intransitive verbs, the words for yellow and green are denominal adjectives and the word for blue (*buluug*), also a bct, is a noun. To say that something is blue, one has to say that it 'has' blue (Maffi, 1990). Such neat correlations between morphology and evolutionary status are, however, sporadic.

Recently, the question has arisen whether all languages do in fact have a small set of lexemes whose denotations are restricted to color sensations and which jointly name all the color sensations. That is, does every language have a full set of bcts? This idea has been dubbed the Emergence Hypothesis. (The question was first raised in unpublished work by Luisa Maffi. See also Lyons, 1995, 1999; Kay, 1999; Levinson, 2000; Kay and Maffi, 2000.) Research on this hypothesis has just begun but it appears that a small fraction of the world's languages may lack complete lexical partitions of the color domain (Levinson, 2000; Kay and Maffi, 2000). Preceding, and in some cases overlapping temporally, with the evolution of bct systems there may exist an evolution toward bct systems.

Nonbasic Color Terms and Noncolor Meanings of Color Terms

The vast majority of studies of color terms, especially since the early 1970s, have restricted themselves to the basic or major color terms, considering nonmajor terms only with a view to excluding them from further consideration. The L&R study described above typifies this practice. There have, however, been studies covering all the lexicalized color expressions of a language, a classic example being Andrés (1949) study of Latin color terms and a more recent example being Bricker's (1999) study of productive and semiproductive morphological process in the formation of secondary color words in two Mayan languages. Bricker also provides a thorough analysis of the noncolor aspects of meaning of Mayan nonbasic color terms, often involving texture and other aspects of surface appearance. Color words frequently also display metaphorical associations to moral, emotional, supernatural, and other kinds of nonphysical entities and beliefs. Such associations have been widely reported in the ethnographic literature.

See also: Color Classification and Symbolism; Comparative Method in Linguistics.

Bibliography

André, J., 1949. Étude sur les termes de couleur dans la langue latine. G. Klincksieck, Paris.

Berlin, B., Kay, P., 1969. Basic Color Terms: Their Universality and Evolution. University of California Press, Berkeley.

Bricker, V.R., 1999. The morphology of color and texture words in the Maya language of Yucatan. Anthropological Linguistics 41 (3), 283–307.

Conklin, H.C., 1955. Hanunóo color categories. Southwestern Journal of Anthropology 11, 339–344.

Gladstone, W.E., 1858. Studies on Homer and the Homeric Age. Tubner, London.

Gleason, H.A., 1961. An Introduction to Descriptive Linguistics. Holt, Rinehart and Winston, New York.

Kay, P., 1999. The emergence of basic color lexicons hypothesis. In: Borg, A. (Ed.), The Language of Colour in the Mediterranean. Almqvist and Wiksell International, Stockholm.

Kay, P., Maffi, L., 2000. Color appearance and the emergence and evolution of basic color lexicons. American Anthropologist 101 (4), 743–760.

Lenneberg, E.H., Roberts, J.M., 1956. The language of experience: a study in methodology. International Journal of American Linguistics Memoir 13.

Levinson, S.C., 2000. Yélîdnye and the theory of basic color terms. Journal of Linguistic Anthropology 10 (1), 1–53.

Lucy, J.A., 1997. The linguistics of 'color.'. In: Hardin, C.L., Maffi, L. (Eds.), Color Categories in Thought and Language. Cambridge University Press, Cambridge, UK, pp. 320–346.

Lyons, J.A., 1995. Colour in language. In: Lamb, T., Bourriau, J. (Eds.), Colour: Art and Science. Cambridge University Press, Cambridge, UK, pp. 194–224.

Lyons, J.A., 1999. The vocabulary of colour with particular reference to ancient Greek and classical Latin. In: Borg, A. (Ed.), The Language of Colour in the Mediterranean. Almqvist and Wiksell International, Stockholm.

Maffi, L., 1990. Somali color term evolution: grammatical and semantic evidence. Anthropological Linguistics 32, 316–334.

Magnus, H., 1880. Die geschichtliche Entwicklung des Farben-*sinnes*. Leipzig, Viet.

Ray, V., 1953. Human color perception and behavioral response. Transactions of the New York Academy of Sciences 16 (2), 98–104.

Turton, D., 1980. There's no such beast: cattle and colour naming among the Mursi. Man 15, 320–338.

Whorf, B.L., 1956. Science and linguistics. In: Carroll, J.B. (Ed.), Language, Thought and Reality. MIT Press, Cambridge, MA.

Commemoration

Barry Schwartz, University of Georgia, Athens, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Commemoration employs narratives, icons, music, monuments, shrines, naming conventions, and ritual observances to awaken ideas and feelings about those parts of the past that embody current ideals. Two categories of commemoration are distinguishable. Positive commemoration, the traditional mode of marking past events and persons, transmits images that affirm society's ideals and grand achievements. Negative commemoration, the marking of events weighing on the present as sources of shame and dishonor, became conventional during the late twentieth-century human rights movement. The early twenty-first century is thus distinctive in its commemoration of guilt and victimhood. To explain these developments, the theories of the 'politics of memory' and 'memory as a cultural system' are invoked.

Commemoration refers to the mobilization of symbols to awaken and preserve beliefs and feelings about the past. The word commemoration derives from Latin (com – together + memorare – to remember: to remember together). Remembering together occurs in every sphere of life – family, association, occupation, community, and society – but its dynamics are most visible in the most inclusive, societal, sphere. Commemoration is society's moral memory, distinguishing events its members believe to be deserving of remembrance from those deserving of being merely recorded. Commemoration promotes society's sense of itself as it affirms its members' mutual affinity and common identification with its past.

Commemorative Symbolism

The primary vehicles of commemoration include written and oral hagiography, icons, music, monuments, shrines, relics, naming conventions, and ritual observances. These 'cultural objects' (Griswold, 1987) lift from the historical record those events that represent a given collectivity's conception of its highest ideals and lowest misdeeds.

Commemorative writing includes biographies, eulogies, poems, plays, and commentaries motivated by admiration or enmity. Emphasizing the moral rather than factual aspects of the past, commemorative texts tend toward hagiography or condemnation and are thus distinguishable from analytic history and chronicling. The line between commemoration and analytic history blurs, however, when the latter is written and published with a view to marking the anniversary of the event or person, which is its subject. First-century Christian lectors read to illiterate Christian communities written accounts of Jesus' life during their Sabbath observances. Newspapers played the major role in the anniversary of Hurricane Katrina (Robinson, 2009).

Music, consisting of anthems, hymns, and inspirational songs, aligns melodic and harmonic structures and lyrics to historical events in such a way as to dramatize their significance. National anthems, school, and military songs are cases in point.

Icons, signs resembling what they represent, commemorate the past by bringing it to visual presence through paintings, statues, prints, photographs, motion picture film, television, and online video.

Monuments, including obelisks, antique temples, and other memorial structures, are designed to elevate the public imagination by dramatic reference to both grand events and atrocities.

Shrines, such as birthplaces, residences, state buildings, military headquarters, battlefields, cemeteries, and places of surrender, execution, and extermination – all play a key role in commemoration because they bring the individual into direct contact with sacred sites. Shrines are sacred because they are, by definition, 'containers' (Latin, scrinium – case, chest) of relics. In contrast to mundane things, relics are physical objects associated with extraordinary events or used by extraordinary people.

Naming patterns make past events ubiquitous by incorporating their memory into the identity of businesses, streets, cities, towns, counties, states, rivers, and mountains.

Observances (periodic performance of anniversary, centennial, and holiday rites) have been analyzed more extensively than other commemorative forms, but their functions are the same: to maintain the vitality of the past, to prevent it from being forgotten and to preserve what is essential to the collective consciousness.

Commemoration's observances, together with its other media, comprise a sign system recommending selected historical events for contemplation, promoting a common legacy by making them tangible and accessible to all.

Commemoration, History, and Collective Memory

Commemoration, history, and collective memory perform interrelated functions. Memory is a necessary property of the mind, a fundamental component of culture, and an essential aspect of tradition. Although individuals alone possess the capacity to remember the past, they never do so singly; they do so with and against others situated in different groups and through the knowledge and symbols that predecessors and contemporaries transmit to them.

A variant of collective opinion, collective memory refers to the *distribution* throughout society of what individuals believe, feel, and know about the past, how they judge the past morally, how closely they identify with it, and how much they are

inspired by it as a model for their present conduct. The word 'distribution' is emphasized because its key property is variation, which denies the possibility of consensus. That every distribution also has a central tendency makes total dissensus equally impossible. Because similar memories appear in comparable groups of individuals widely dispersed and unknown to one another, and because these memories replicate themselves across generations they must be treated as emergent, objective entities that stabilize and link the consciousness of present and past.

History, as French sociologist Maurice Halbwachs (1980 [1950]) defined it, seeks an objective standpoint to assess the causes and consequences of events. It is 'situated external to and above groups' and describes the past independently of contemporary opinions and conditions. Once established, Halbwachs believed, historical knowledge remains stable – its stream of facts and demarcations 'fixed once and for all.' In contrast, commemoration and collective memory remain significant only as long as they remain relevant (pp. 80–81). Pierre Nora, a French historian, explicitly adopts Halbwachs' belief in the fundamental opposition between history and memory, and he regards all commemorative forms as *lieux de memoire* (sites of memory), that is, feeble efforts to defend against the relentless disenchantment of analytic history.

History's goal is, thus, to rationalize the past; commemoration and its sites, to sanctify it. History makes the past an object of analysis; commemoration, an object of commitment. History is a system of 'referential symbols' representing known facts and their sequence; commemoration is a system of 'condensation symbols' (Sapir, 1930: pp. 492–493) that simplify events of the past and clarify the moral sentiments they inspire. History, like science, investigates the world by producing models of its permanence and change; commemoration, like ideology, seeks to promote commitment to the world by producing symbols of its values and aspirations (Geertz, 1973: pp. 193–233).

Contrasts among history, commemoration, and memory must not be exaggerated. James Dunn (2003), prominent among historical Jesus scholars, declares that 'the quest for the historical Jesus' can only be a quest for the 'Jesus remembered' by contemporaries and successors who heard or wrote about him (p. 335). Dunn's statement stands out in the context of successful efforts, beginning in the first decade of the twenty-first century, to integrate collective memory, commemoration, and biblical studies (Kirk and Thatcher, 2005).

Commemorative Networks

When commemorative symbols are combined, they convey meanings that differ from those borne by any one symbolic object. To activate the commemorative process through story, monument, song, icon, ceremonial observance, or other signs, therefore, is to activate memories that include but extend beyond the persons and events commemorated. The significance of historical figures and events also change as they are placed in commemorative networks composed of objects absent in previous networks. Thus, Jefferson Davis was sworn in as president of the new Confederacy in Richmond, Virginia beside a massive equestrian statue of George Washington, which was itself adopted as the Confederacy's emblem. During the Civil War, Abraham Lincoln ordered large-scale offensive operations on Washington's Birthday. Upon death, scores of mourning prints placed him in the company of Washington. During the Progressive Era, visual images of Abraham Lincoln were paired with contemporary figures, including President Theodore Roosevelt, identifying Lincoln as one who would have supported Progressive Era reforms. The pairing processes fanned out broadly not only within a network of people whose prestige enhanced and was enhanced by Lincoln's but also to networks of events and structures – notably, political and economic reforms and relentless condemnation of plutocracy. In such mnemonic webs, existing beliefs about Lincoln were reorganized, their institutional roots deepened, and a new and autonomous collection of beliefs about him accumulated. Common features of people and events with which Lincoln was symbolically connected were generalized to Lincoln himself. In this way, he changed through the late nineteenth century from a symbol of national unity to a symbol of social and economic reform.

Commemoration, as Abraham Lincoln's case suggests, is a structuring process that partially overrides the qualities of its objects and imposes upon them its own pattern. Commemoration is creative because it fills in what is lacking in its objects, whether historical figures or events, by attributing to them the features of people and events that a given community deems similar (and other communities deem offensively dissimilar).

Memory of Commemoration

Commemorative symbols represent both the events they commemorate and the concerns of those creating them. Because commemorative symbols refer to the memory of their own genre, they exhibit traces of the way events presently commemorated have been recognized previously. In other words, commemorations are 'path dependent' – shaped by their development as well as their referents; they express an ongoing process of meaning making. Jeffrey Olick's (2007) analyses of discourses marking historical regret, including the memory of National Socialism in Germany, demonstrate this point persuasively. Not only are present commemorations modeled on past commemorations; the *style* of these earlier commemorations, as Andrea Cossu's (2011) study of Italian antifascist representation shows, are themselves path dependent and therefore essential to reinforcing the memory of commemoration and the understanding of how it works.

The redundant, cyclical features of path dependency are exemplified in the holiday calendar. In the United States, the annual repetition of Martin Luther King Day, Lincoln's Birthday, Washington's Birthday, Memorial Day, July Fourth, Columbus Day, Veterans Day (formerly Armistice Day), Pearl Harbor Day, in that order, enables the single individual to experience, year after year, although with demonstrably decreasing enthusiasm and solemnity, the representation of major events in his nation's history.

Path dependency can negate rather than reaffirm earlier commemorative objects. Conceived in light of their generic predecessors, certain war memorials, in certain social circles, evoke more criticism than approval. The establishment of

a Korean War Veterans Memorial, supervised by a committee hostile to the Vietnam Veterans Memorial's antiwar inspiration, set down firm requirements in its invitation to design competitors (Schwartz and Bayma, 1999). The Korean Memorial had to be visible, uplifting, expressing honor, pride, and dignity, evoking admiration and gratitude. It had to be visible rather than set below ground level, as is the Vietnam Memorial, whose original plan expressly excluded patriotic symbolism. The Korean War Memorial's design is shaped as much by its negative Vietnam War predecessor as by the actual war it commemorates.

Commemoration in the Post-Heroic Era

Commemoration sites are viewed, witnessed, or attended with solemnity by some; ignored, regarded with indifference, or even ridiculed by others. During the Enlightenment, commemoration began to lose its capacity to instruct and inspire, but this trend assumed its sharpest downward turn during the last third of the twentieth century, strongly evidenced in the United States by multiple indicators, including site visitation statistics and national surveys of historical figures' diminishing renown. As the public's interest in its history fades, however, its commemoration ceases to be taken for granted and scholarly interest therefore escalates. In the *Dissertation Abstracts*, with commemoration the keyword and Masters and PhD theses included, two entries are found between 1970 and 1979, 6 and 17 entries during the 1980s and 1990s, respectively. During the first decade of the twenty-first century, 53 entries appeared. In *JSTOR*, an online journal archive, the very same trend appears. Academic interest in commemoration increases exponentially at the turn of the twenty-first century as the heroic past fades from the collective consciousness and becomes an object for dissection rather than adoration.

Until the last third of the twentieth century, references to the anniversary of significant events, such as George Washington's birth, filled local newspapers' front pages with editorials, commentaries, ink drawings, stories about different parts of his life, including schedules where civic organizations, from trade to professional organizations and churches, held their programs. Schools gave special lessons and ceremonies the day before Washington's Birthday, when all federal and most state and municipal activities were suspended. Presently, the commemoration of his birth consists mainly of brief remarks about his life in newspapers and television commentary. In contrast, the anniversary of Martin Luther King's January birth is widely recognized, with an increasing number of streets and organizations annually named for him and more printed space and television time devoted to him. These rites are consolidated during the entire month of February, federally recognized in 2008 as Black History Month, succeeding Negro History Week, begun in 1926.

The present 'Era of Commemoration' largely perpetuates the memory of victims of injustice, despised minorities, and oppressed people. 'The explosion of minority memoriesm,' according to Pierre Nora (2002), has changed commemoration's cultural function. "Swayed by the democratization of history," it has come to resemble a protest movement, a "revenge of the underdog… ."

The superimposition of sites of regret upon sites of triumph is now the keystone of many national commemorations. Writing about Grozny's monument to the Soviet Union's wartime deportation of Chechnyans, Brian Williams (2000) declares: "One must know a nation's tragedies and the way they commemorate them to know its soul" (p. 128). Tragedy, not triumph, he says, defines national memory. But the two are often indistinguishable. Villainy and heroism, the mundane and the epic, events worth forgetting and those worth preserving – it is the blurring of these lines that goes to the core of the postmodern condition. The transition from a pre–World War II culture of commemoration stressing heroic achievement to a postwar culture stressing oppression, trauma, victimhood, and injustice is widely evident. In his observations on the commemoration of French resistance against Nazi occupiers, Nathan Bracher (2007) declares that the general evolution of European sensibilities is driven by disinterest, if not disdain, for warrior virtues, and a newfound appreciation of warriors' victims. Europe as a whole is now united by rejection of war, intolerance, and inequality – all dramatized by its use of the Holocaust narrative as a more or less de-Judaized symbol of human rights. France, for example, commemorates often the victims of the occupation; rarely the valor of the resistance. The memorial to militant resisters, located at Mount Valerian, site of more than 1000 resister executions, bears no mention of their successes against the strong enemy with whom most Frenchmen passively cooperated. It implies reluctance to commemorate French men and women as slayers rather than victims of their Nazi rulers.

Commemorating historical wrongdoing became commonplace during the late twentieth century in connection with the emerging 'human rights movement.' The new 'politics of recognition' also bends the commemorative impulse to achievements and sufferings of communities previously marginalized and underrepresented in the historical record. These sites of regret are more likely to be found among 'cultures of dignity,' dedicated to the cultivation of democracy and individual rights, than among 'cultures of honor' dedicated to the cultivation of individual obligations to the nation, state, or community. The creation by the American government of the United States Holocaust Memorial, to take but one instance, justifies itself, as do European memorials, on the ground of a universal moral relevance transcending the suffering of European Jewry. Doing so, the Memorial erodes the Holocaust's particularistic relevance (Rabinbach, 1997) by ignoring the link between genocide and Jewish history. The same rationale applies to museums of Native American, African-American, Asian-American, and Hispanic-American experience. The commemoration of helplessness is even evident in the design of heroic monuments. Franklin D. Roosevelt was leader of the free world during its most severe economic and military challenges, but his memorial recognizes Americans' affection for handicap by placing on the site a statue of the paralyzed president in his wheelchair. Likewise, Dwight D. Eisenhower's Memorial, as presently planned, deliberately qualifies his military and presidential achievements by representing him as a barefoot boy in Kansas.

Diversity, another aspect of post-heroic commemoration, distinguishes the Korean War Veterans Memorial, which includes a wall divided into slanted vertical and straight horizontal lines, the resulting cells of which represent each of the

military services. Within each service, different branches of activity, and within each branch of activity different racial, ethnic, and gender groups, are recognized. For example, engineering, armor, infantry, artillery, and medical units are represented by men and women of all racial and visibly ethnic groups. The 19 statues of infantrymen, too, are meant to be recognizable as European-American, African-American, Asian-American, and Hispanic-American. A war memorial conceived to honor sacrifice for the sake of a transcendent nation ironically takes a pluralistic turn and, by so doing, commemorates equally the separate communities of which that nation is composed. Commemorative pluralism, in this precise respect, articulates the mundane segments, rather than the transcendence, of the nation-state.

Anomalous Commemoration

Diversification of the commemorative process admits of anomalous sites. Anomalous commemoration combines aspects of the past conventionally deemed morally incompatible, such as President Ronald Reagan's (1985) visit to Germany's Bitburg military cemetery, where the remains of Wehrmacht and SS personnel are buried, on the same day as his visit to the Bergen-Belsen concentration camp site. Tokyo's Yasukuni Shrine, likewise, commemorates the souls of convicted and executed war criminals as well as the innocent who died for their country, which is why Prime Minister Junichiro Koizumi's public visits to the shrine set off indignant protests abroad, especially in Korea and China, and, at home, a torrent of controversy that has yet to abate (Mochizuki, 2010). In the United States exists a similar incongruity: monuments to Confederate leaders and soldiers who defended slavery are accompanied in Richmond, Virginia, the Confederacy's capital, by a statue of The Great Emancipator, Abraham Lincoln, and his son overlooking the former Tredegar weapons factory. The Vicksburg, Mississippi statue of Lincoln in the friendly presence of Confederate President, Jefferson Davis, is another instance of anomalous commemorative pairing. In Bitburg, Tokyo, and Richmond, those considered virtuous and villainous are commemorated together.

Monuments to executed military deserters are equally striking. Once the embodiment of disloyalty and cowardice, desertion is now defined as a symptom of stress or battle fatigue. The Shot at Dawn Memorial in Staffordshire, England, dedicated in 2001, consists of the statue of an executed 17-year-old surrounded by 306 stakes to which British and Commonwealth soldiers were tied and executed during World War I. A single stake, also dedicated in 2001, commemorates English, French, and Belgian soldiers executed in Ypres, Belgium. French President Nicholas Sarkozy spoke at the Duaumont Fort, France, site of the Battle of Verdun, announcing that those executed were 'men like us' who deserved to be remembered because they had 'reached their extreme limits.' Germany commemorates more deserters than any other country. In Kassel, Hannover, Ulm, Cologne, Stuttgart, and other cities stand monuments recognizing the men executed during World War II by the German military.

Commemoration is no less anomalous when its intensity is disproportionate to the significance of its object. The apotheosis of Rosa Parks included public memorial services in three cities, including Washington, DC, where she laid in state in the US. Capitol Rotunda under a flag flying half-mast while a plan was inaugurated to place her statue in the Capitol Rotunda. Her celebrity occasioned the forgetting of others who risked and accomplished more, far more, than she (Schwartz, 2009).

Surface Commemoration

The 'death dip,' which refers to lowered mortality rates among the intensely religious on days leading up to their major holidays, followed by the steep increase in rates when the holiday is over, is a measure of commemoration's role in the life of the individual. This effect is variable. To find a death dip among orthodox Christians prior to Christmas and Easter, and among orthodox Jews prior to Rosh Hashana and Yom Kippur is unsurprising; to find it related to such American holidays as Presidents' Day, whose object is ambiguous, would be shocking.

The Monday Holiday Act presupposes the weakened state of the national consciousness by moving holidays that fall in the middle of the week to Mondays with an explicit view to promoting more tourism, shopping, travel (and fuel use) by automobiles and public carriers over long weekends. Presidents' Day is the new name for George Washington's Birthday, the only president on the federal holiday calendar. Absent from this calendar but now on most state calendars, Presidents' Day commemorates different people: here, George Washington; there Abraham Lincoln; elsewhere, Washington and Lincoln; somewhere else, all presidents; in another place, Abraham Lincoln and Daisy Bates. The casualness with which Presidents' Day is recognized embodies Amitai Etzioni's (2001) claim that as 'recommitment holidays' have declined in relevance, 'tension management' holidays, including New Years Day, Halloween, and Valentine's Day, have become more widely celebrated. Tension management holidays are said to perform a 'safety valve' function: they contribute to reducing tension accumulated through the normal difficulties of life. Presidents' Day is one of these holidays. Most importantly, certain recommitment holidays have been converted into holidays of tension release. Memorial Day and July Fourth, observed with almost religious passion a century ago, are still federal holidays but observed with relatively little ritual and even less fervor. They provide more opportunity for recreation than commemoration.

To be severed from the past is no longer to be identified with it. Michel-Rolph Trouillot (2000), for example, explains that national apologies and rituals of regret for past offenses are superficial because their participants, representatives of aggressors and victims, have no organic connection to the communities they represent. Commemoration is only skin-deep when it preserves but fails to instruct and inspire, when its participants recognize forebears as benefactors but no longer feel themselves to be beneficiaries, when earlier sacrifices, even those made explicitly for the benefit of posterity, fail to evoke gratitude, when the past is remembered but unfelt and unjudged.

The death of living faith is not contradicted by recent surges of national enthusiasm. Ritual displays accompanying the 1991 Gulf War, 9/11, and enthusiasm at the start of the Iraq War show that nationalist sentiment, which generally exists in

a latent state, is a powerful source of social energy during crises. Yet, the ends for which such energy is expended exist in the national present; they recognize but draw little inspiration from the national past.

Transitory Commemoration

Some commemorative structures, although inspired by deep sentiment, are impermanent. They consist of easily removable objects that express feelings (usually regret) over a (usually recent) loss or calamity. Immediately after 9/11, people across the country set up memorial candles in memory of the thousands killed. As the shock wore off, the sites were vacated. Often, religious markers are placed at the site of a fatal automobile accident, but most are recognized by only a few drivers before the elements destroy them. Some temporary sites assume conspicuous dimensions. After Princess Diana's death, an enormous mound of flowers appeared at the gate of Buckingham Palace, but on the first anniversary of her death the Palace gate was conspicuously vacant. The 2005 London Subway Bombing, too, led to a fleeting, media-driven commemoration but no physical monuments (Hoskins, 2011).

In some places, the permanent and transitory coexist. At the Vietnam Veterans Memorial, the aggrieved place personal objects beneath the name of the lost. Gathered up by the US. Park Service and stored, these items are selectively displayed in a remote museum, giving them at least a probability of permanence they would otherwise lack.

Aborted Commemoration

'To abort' is an infinitive denoting the act of preventing an activity from completing itself. Among the best known of aborted commemorations is that of the Smithonian Museum's (1995) Enola Gay exhibit that was intended to depict the consequences of the nuclear attack on Hiroshima and Nagasaki. The exhibit's undertone, critical of the bomb's use and deftly criticizing the United States for excessive force, despite the 1000 *daily* American casualties at sea during the weeks prior to Hiroshima, disturbed many influential persons and associations. Congress withheld funding, and the exhibit was canceled. So, too, organizers of the International Freedom Center planned to use the Ground Zero site to redefine 9/11 as one of a long line of atrocities throughout human history, including American slavery, the mistreatment of native Americans, Japanese Americans, and poverty. New York's political representatives perceived the Freedom Center a 'blame America first' project; others interpreted its plans as an effort to mute 9/11 by commemorating victims who had no relation to it. Despite its many supporters, including the influential *New York Times*, Governor George Pataki declared the museum an anti-American platform at worst; at best, an inappropriate site for ideological expression, and he refused to approve it.

The logic framing The Freedom Center's proposal and rejection is also manifest in efforts within the nation of Israel to commemorate the *Nakba* (Catastrophe), namely, the establishment of the state of Israel. Almost all Arab-Israelis, including their *Knesset* (Parliament) representatives, commemorate this day with expressions of grief, but the idea is not only debated but endorsed within leftist Jewish institutions. The most enthusiastic supporters are, in fact, urging their society to mourn its success at preventing its own destruction. In addition, editorial headlines in Israel's left-leaning *Ha'aretz* endorse efforts to change the lyrics of the national anthem and the design of its flag to make them acceptable to Israeli-Arabs. Although reenergized within many Jewish and Muslim organizations, the movement to commemorate the Nakba's anniversary is vigorously rejected by the Israeli mainstream.

Delayed Commemoration

Construction of some commemorative sites are unduly delayed rather than aborted. Throughout World War II, American and British magazines honored Air Chief Marshall Arthur (Bomber) Harris by placing his photograph on their front cover. He had avenged the Blitz many times over by attacks on German cities. Almost 50 years passed before a statue of Harris was officially approved. England's left had always viewed the ferocity of Harris's raids as disproportional to the damage and death caused by the Luftwaffe. Indeed, activists vandalized the statue shortly after its London unveiling. In this connection, 67 years passed before public subscriptions paid for an official memorial to the RAF Bomber Command, which lost 55 000 men in the air war against Germany. All the social and economic problems associated with recovery from war make delayed commemoration the rule rather than the exception, but opposition to Bomber Harris and his air crews could have only arisen on a mass scale in the late twentieth century, when obsession with human rights transformed German populations supporting Nazi war factories into helpless victims of overwhelming force; the British airmen, brutes deserving no place in posterity. Allies used that same 'disproportionate force' to drive the Nazis out of France – a heroic act at the time but one that now violates the moral sensibilities of French textbook writers as much as it disturbs the humanitarian conscience of many English and American intellectuals.

Theories of Commemoration

The common feature of current commemoration theories is their strong although not exclusive emphasis on minorities and victims, but each takes up a different facet of the problem. Hegemony theory, multiculturalism, postmodernism, and critical theory give insight into the political motives that commemorative objects embody. By contrast, the cultural lens focuses on why creators design such objects as they do and what they mean to viewers and listeners. The theories also differ on what commemoration is: whether a process of construction or retrieval, a mirror reflecting the concerns and interests within a society or a lamp of inspiration and moral precedent.

The Politics of Commemoration

In the first strand of the politics of commemoration, power is concentrated and expresses the tastes, values, and interests of

society's elite. Elite producers of commemorative symbolism bring the past to life as they manipulate the attitudes of their audience in their own interest (Gillis, 1994). Despite efforts to universalize this model, it is most applicable to authoritarian societies wherein 'memory laws' (Koposov, 2011) sustain the official version of history. The premise of the second strand of the politics of memory is that power is diffused and that commemorative objects emerge from a matrix of cross-cutting coalitions, networks, and enterprises rather than a ruling elite's interest in maintaining its privilege. The fate of artistic and presidential reputations, Holocaust memories, monument making, place naming, and the organization of museums exemplify commemoration as a pluralistic network of interest and power. John Bodnar's (1992) distinction between 'official' and 'vernacular memory' is an instance of the pluralistic mode, as is Lorraine Ryan's (2011) conception of resistance as a necessary component of memory. Monopolies on memory and commemoration are, thus, difficult to establish because competing interpretations are inevitable, and free societies allow, indeed, encourage, their expression.

The politics of commemoration are also evident in 'counter-monuments.' The Korean War Veterans Memorial was expressly conceived to counter the design of the nearby Vietnam Veterans Memorial. An even more radical instance of counter-monuments is based on James Young's conviction that commemoration is a substitute for memory and cuts off the people from a real appreciation of their past. Open-ended debate over the meaning of an event, Young asserts, is the ultimate counter-monument (374).

Shortly after the assassination of Israeli Prime Minister Yitzhak Rabin, admirers constructed a monument near the spot where a right-wing zealot assassinated him. For many years, Rabin admirers, mainly liberal, assembled at the monument every Friday to hear speeches and music in remembrance of his life. No such assembly and no such monument exist in the conservative city of Jerusalem (Vinitzky-Seroussi, 2009).

Political ideology also influences commemoration at the aesthetic level. Cultural conservatives tend to favor 'triumphal' commemorative sites, exemplified by the Washington Monument, Lincoln Memorial, Mount Rushmore, the Iwo Jima Memorial – sites that affirm patriotism and the virtues of the nation-state and its leaders. In contrast, liberals tend to reject traditional commemorative forms. Massive monuments, imposing statues, long and noisy parades, among other ostentatious displays are sources of disfavor. The political left, however, commemorates its favored groups by this very same aesthetic. Leftist critics consider Washington's World War II Memorial to be 'fascist' and condemn it artistically because its gothic design fails to conform to the existing Mall monuments. They raise no such objections against the Native American museum, located near the Capitol Building but colored yellow to give the appearance of a huge adobe structure. Nor do they criticize the massive Martin Luther King monument, designed by a sculptor from the People's Republic of China, representing the great peace leader in the same genre as that commemorating Kim Il Sung in North Korea and Mao Zedong in China. Visitors look up to see an enormous Martin Luther King stepping out of an enormous granite block.

Subordinating memory to political power, the politics of memory falls short in three respects. First, in the United States and Europe, no institution does more to commemorate its past wrongdoings than the state. Second, given that all commemoration requires political authority to make it effective, the question of why national and local elites choose to commemorate some persons and events and not others remains unanswered. Third, ideology often trumps politics in determining why the public accepts some elite efforts and rejects others.

Commemoration as a Cultural System

Most commemorative events involve political ideals, sentiments, and judgments, but they are rarely efforts to gain political advantage. The roots of commemorative conflict run too deep for such a simplistic account. In Michael Schudson's (1992) words, "Examining the past is motivated not only by a drive for legitimation … People seek to know what is right, what is true … They seek some kind of direction when they are aimless. They seek in the past some kind of anchor when they are adrift" (p. 213). In short, commemorative symbols articulate meaning, not power.

Conceived as a source of meaning, commemoration is less the construction of a manipulative interest group than a moral frame in terms of which people orient themselves. Commemorative agents, thus, perform more like priests than like manipulators and lobbyists; they seek to reaffirm beliefs they share with rather than impose on their audiences. Commemoration is always a model *of* society – a reflection of its needs, interests, fears, mentality, and aspirations – but it is more importantly a model *for* society – a program defining its values and goals and a frame for organizing the meaning of its members' experience.

As a cultural system, commemoration makes past and present meaningful in terms of one another, which is why the production of commemorative objects becomes most urgent when some significant event is in danger of being forgotten – a situation to which Jan Assmann's concept of *Traditionsbruch* (tradition break) refers. Assmann introduced the *Traditionsbruch* concept during the 1990s, but it was rapidly assimilated into the field of biblical studies during the first decade of the twenty-first century (Kirk and Thatcher, 2005; Thatcher, 2008). Thus situated, it demonstrates the limits of 'communicative memory' (including oral tradition) and explains why the first written gospel appeared when it did. As the generation that lives through a particularly significant event dies out, those intent on preserving knowledge of it begin to write, paint, sculpt, build monuments, and organize observances. In particular, the first Gospel was written 40 years after the death of Jesus, when the last of his contemporaries were still able to tell each other what they knew about him. Likewise, questions of guilt and identity on the British islands of Guernsey and Jersey, whose citizens cooperated with the Nazi occupiers, "became more urgent with the realization that the witnesses of these times would no longer be around" (Carr, 2012: p. 53). Just so, the Korean War and World War II Memorials were rushed to completion as those who deserved recognition for serving in these wars were on the verge of vanishing. Each case reflects society's fear of the *Traditionsbruch* – the rupture that erodes irretrievably the present generation's knowledge of its forebears.

Society's fear is justifiable, for symbol systems are mutable. The generation that ascribes the same meaning to commemorative objects as did its forebears is theoretically impossible, for doing so would require a society to be so fixed as to clone itself culturally. Commemorative transformations are always superimposed upon commemorative continuities, and they can be dramatic. The destruction of King George's statue by rebelling Americans colonists in 1776; the deterioration and defacing of countless Civil War monuments throughout the United States, and the late twentieth-century vandalizing of images of sacred figures are simple examples. More complex is the development in Dublin of the Irish National Memorial to honor 300 000 Irishmen who fought with the English during World War I (Rigney, 2008). The Memorial's plan was drafted in 1919, implemented during the 1930s, but interrupted by World War II and Irish neutrality. After the war, the eroded monument was restored and formally dedicated in 1988 as England and Southern Ireland reconciled. Such are the vicissitudes of commemorative objects and of the mentalities they articulate.

Commemorative symbols gain and lose relevance and resonance independently of the distribution of power, but power or, for that matter, any social or economic asset, performs significant cultural functions. Brian Conway's *Bloody Sunday* (2010) is one of the best examples of a multiple factor theory of commemoration, defining explicitly its cultural, economic, political, and situational contexts – a 'thick (contextual) description' that roots the concept of commemoration deeply in the social fiber. That Conway's topic consists of victims of oppression rather than agents of victory, however, is what makes it so apt an example of commemoration as a twenty-first century cultural system.

Conclusion

The past's relevance diminished during the late twentieth century in two ways. First, the individual's connection to his or her society and its past was weakened by the discrediting of its metanarratives – its great legends, myths, and histories. Since the key feature of metanarratives is embodied in a plot structure that interprets events in a way that can be widely understood and shared, people require them, collectively and individually, to make sense of experience. 'Of what great story is my life and death a part?' is the question they ask in their effort to define themselves and the meaning of their lives. This question would not be worth asking if the great stories that impressed our forebears still impressed us – and if there were no consequence, no cost, to the repudiation of the greatnesses of history.

As the positive aspects of societal narratives are ignored, the orienting past tense of the individual's life is negated. These positive narratives are not completely ignored, but few would deny that diminished engagement with the past distinguishes recent from remote generations. The adjective 'diminished,' however, must be underscored, for diminishings are more difficult to describe and comprehend than disappearances. At question is whether traditional reverence for the past is destined to disappear or whether it remains because it is not traditional at all but rather a fundamental requirement of societal continuity, because there is a level below which commemorative interest, if not passion, does not fall. Amy Corning and Howard Schuman's (2013) surveys of events as different as 11 September and Woodstock are relevant to this point. As the 10th and 40th anniversaries of these events approached, media coverage grew and more people named them as memorable events. The percentage mentioning these events fell after the mass media ceased coverage, but that percentage remained higher after the coverage than before.

The present situation, then, must be rightly understood: balkanized preferences have enervated the West less politically than culturally. Because the West's mentality is currently progressive, compassionate toward the underdog and skeptical of everyone else, because it is irreverent, pluralistic, relativistic, and less attached to tradition than ever before, it has begun to alienate individuals from whatever historical legacy they possess. In the deliberate minimizing of founders and their ideas, the substitution of shallow pity for the invocation of sacrifice, courage, and perseverance, and the anxious commemorative concessions to even the most outrageous claims of 'oppressed minorities,' there is something at once gained and lost. The West has never been freer of religious, ethnic, and racial hatred; yet, its historians – often justly, frequently unjustly – challenge official accounts of their nations' historical virtues. As democratic peoples adopt the *mea culpa* perspective on history, they lose sight of the great transformations wrought by their forebears; they lose sight of themselves as historical beings, forget that they have inherited their most valuable possessions from those whom they so ardently criticize. Merely to contemplate these issues is to affirm commemoration's ironic functions in twenty-first century society.

See also: Civil Religion; Collective Identity; Collective Memory, Anthropology of; Collective Memory, Psychology of; Durkheim, Emile (1858–1917); Names and Naming; Narrative Networks.

Bibliography

Bodnar, J., 1992. Remaking America: Public Memory, Commemoration, and Patriotism in the Twentieth Century. Princeton University Press, Princeton, NJ.
Bracher, N., 2007. Remembering the French resistance: ethics and poetics of the epic. History and Memory 19, 39–67.
Carr, G., 2012. Occupation, heritage, commemoration, and memory in Guernsey and Jersey. History and Memory 24, 87–117.
Corning, A., Schuman, H., 2013. Commemoration matters: the anniversaries of 9/11 and woodstock. Public Opinion Quarterly 77, 433–457.
Conway, B., 2010. Commemoration and Bloody Sunday: Pathways of Memory. Palgrave MacMillan, New York.
Cossu, A., 2011. Commemoration and processes of appropriation: the Italian Communist Party and the Italian resistance (1943–48). Memory Studies 4, 386–400.
Dunn, J.F., 2003. Jesus Remembered. William B. Eerdmans, Grand Rapids, MI.
Etzioni, A., 2001. Holidays: the neglected seedbeds of virtue. In: The Monochrome Society. Princeton University Press, Princeton, NJ, pp. 113–140.
Geertz, C., 1973. Ideology as a cultural system. In: Geertz, C. (Ed.), The Interpretation of Cultures. Basic Books, New York, pp. 193–233.
Gillis, J.R. (Ed.), 1994. Commemorations: The Politics of National Identity. Princeton University Press, Princeton, NJ.
Griswold, W., 1987. A methodological framework for the sociology of culture. In: Clogg, C. (Ed.), Sociological Methodology. American Sociological Association, Washington, DC, pp. 1–35.
Halbwachs, M., 1980[1950]. Collective Memory. Harper & Row, New York.
Hoskins, A., 2011. 7/7 and connective memory: interactional trajectories of remembering in post-scarcity culture. Memory Studies 4, 269–280.

Kirk, A., Thatcher, T. (Eds.), 2005. Memory, Tradition, and Text: Uses of the Past in Early Christianity. Society of Biblical Literature, Atlanta, GA.

Koposov, N., 2011. 'The Armored Train of History': The Politics of History in Post-Soviet Russia. Perspectives on History, American Historical Association. www.historians.org.

Mochizuki, M.M., 2010. The Yasukuni Shrine conundrum: Japan's contested identity and memory. In: Kim, M., Schwartz, B. (Eds.), Northeast Asia's Difficult Past: Essays in Collective Memory. Palgrave Macmillan, New York, pp. 31–52.

Nora, P., 2002. Reasons for the current upsurge in memory. Eurozine. www.eurozine.com.

Olick, J.K., 2007. The Politics of Regret: On Collective Memory and Historical Responsibility. Routledge, New York.

Rabinbach, A., 1997. From explosion to erosion: Holocaust memorialization in America since Bitburg. History and Memory 9, 225–255.

Rigney, A., 2008. Divided pasts: a premature memorial and the dynamics of collective remembrance. Memory Studies 1, 89–97.

Robinson, S., 2009. 'We were all there:' remembering America in the anniversary coverage of Hurricane Katrina. Memory Studies 2, 235–253.

Ryan, L., 2011. Memory, power, and resistance: the anatomy of a tripartite relationship. Memory Studies 4, 154–169.

Sapir, E., 1930. Symbols. In: Seligman, E.R.A., Johnson, A. (Eds.), Encyclopedia of the Social Sciences, vol. 14. McGraw-Hill, New York, pp. 492–495.

Schudson, M., 1992. Watergate in American Memory. Basic Books, New York.

Schwartz, B., Bayma, T., 1999. Commemoration and the politics of recognition: the Korean war veterans memorial. American Behavioral Scientist 42, 946–977.

Schwartz, B., 2008. Collective memory and abortive commemoration: presidents' day and the American holiday calendar. In: Mack, A., Hirst, W. (Eds.), Social Research, vol. 75, pp. 75–110.

Schwartz, B., 2009. Collective forgetting and the symbolic power of oneness: the strange apotheosis of Rosa Parks. Social Psychology Quarterly 72, 123–142.

Trouillot, M.-F., 2000. Abortive rituals: historical apologies in the global era. Interventions: International Journal of Postcolonial Studies 2, 171–186.

Vinitzky-Seroussi, V., 2009. Yitzhak Rabin's Assassination and the Dilemmas of Commemoration. SUNY Press, Albany, NY.

Williams, B.G., 2000. Commemorating 'the deportation' in post-soviet Chechnya: the role of memory in the 1994–1996 and 1999–2000 Russo-Chechen wars. History and Memory 12, 101–135.

Young, J.E., 2002. At Memory's Edge: After-Images of the Holocaust in Contemporary Art and Architecture. Yale University Press, New Haven, CT.

Commodity Chains

Elaine R Hartwick, Framingham State University, Framingham, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Geographically conceived, commodity chains consist of significant production, distribution, and consumption nodes, the spatial connections among these, and the economic, social, cultural, and natural conditions of production and consumption. This concept connects the generation of cultural (signifying and representational) effects at the consumption node of a commodity chain with the social and natural conditions at the production end of the chain, the purpose being to 'bring home' to consumers the consequences of consumption. Formulated in this way commodity chains are said to constitute an analytical basis for a transformative politics and a new platform for understanding global capitalism.

Commodity chains follow specific commodities as they are moved over space from points of production to points of consumption. Commodities are seen as originating in regions of production and moving through networks of economic agents, as with producers, distributors and transporters, to regions of consumption, where they are advertised, retailed and consumed. The commodity chain concept has potential for demonstrating the effects of consumption on labor and the environment. Armed with greater geographical knowledge, organized around the commodity chain concept, consumers can make choices that more fully consider the effects of their consumption on workers and the natural environment.

Conceptions of Commodity Chains

Since the middle 1970s there has been a resurgence of interest in space conceptualized not as a homogenous expanse of territory, but as channels, networks, nodes, and chains of connection. In economic geography the main interest resides in investigating the spatial organization of business networks (Yeung, 1994). While related to this interest, commodity chain analysis differs, in that analysis focuses on commodities, rather than industrial organizations or networks of interfirm relations. The geographical literature contains numerous references to concepts similar to commodity chains. For example, Harvey (1989) implies the notion of a commodity chain in arguing the importance of knowing the effects of consumption on the conditions of production: "We can take our daily breakfast without a thought for the myriad of people who engaged in its production. All traces of exploitation are obliterated in the object." However, the commodity chain concept only began to be used formally in the middle 1980s. There is now a substantial record of case studies (Hughes and Reimer, 2004). Three versions of commodity chain analysis can be discerned: global commodity chains; systems of provision; and commodity circuits. These versions have recently been criticized as limited in scope and applicability, and an expanded commodity chain concept proposed.

Global Commodity Chains

The notion of commodity chains has been developed in the world systems approach mainly in the discipline of sociology (Hopkins and Wallerstein, 1986; Gereffi, 1994; Bair, 2005, 2008). As characterized by the world systems approach, a commodity chain, or value chain, refers to a network of labor and production processes whose end result is a finished commodity (Hopkins and Wallerstein, 1986). Gereffi and his coworkers see interorganizational networks linking households, enterprises, and states to one another within the world economy (Gereffi et al., 1994). Global commodity chains stress the organizational scope of transnational production arrangements: that is, the linkages among economic agents, such as raw material suppliers, factories, traders, and retailers. In this view, commodity chains have three main dimensions: (1) an input–output structure, in the sense of a set of products and services linked together in a sequence of value-adding economic activities; (2) territoriality, as with spatial dispersion or concentration of production and marketing networks; and (3) a governance structure, in terms of authority and power relationships that determine how financial, material, and human resources are allocated and flow within a chain (Gereffi et al., 2005; Cattaneo et al., 2010).

Two distinct types of governance structure are distinguished: producer-driven and buyer-driven commodity chains. In producer-driven commodity chains, transnational corporations or other large integrated industrial enterprises control the production system. In buyer-driven commodity chains, large retailers, brand-named merchandisers, and trading companies set up decentralized production networks in exporting countries, typically in the Third World (Hassler, 2003). The latter pattern of trade-led industrialization is common in labor-intensive, consumer goods industries such as garments, footwear, toys, household goods, consumer electronics, and a wide range of handcrafted items, such as furniture or ornaments. International contracting is generally carried out by independent Third World factories making finished goods, rather than components or parts, under original equipment manufacturer arrangements, and with specifications supplied by the buyers and branded companies that design the goods (Gereffi, 1994; Gereffi and Memedovic, 2003). In these studies firms constitute the main unit of analysis and unequal distributions of wealth along the chain are explained as outcomes of competition and innovation at the level of firms. A third type of governance structure has also been proposed emphasizing the rise of

a financial sphere made up of institutional investors and executives of large corporations commanding the top of global commodity chains (Palpaceur, 2008).

Systems of Provision

A second approach to commodity chain analysis derives from Fine and Leopold's (1993) rejection of horizontal analyses of consumption that stress commodification and aestheticization connected with retailing, and their favoring, instead, of the vertical dimension of the system of provision – the commodity-specific chain connecting production, distribution, marketing, and consumption, and the material culture surrounding these elements. Different systems of provision result from distinct relationships between material and cultural practices spanning the production, distribution, and consumption of goods (Fine and Leopold, 1993). This perspective allows for a more dynamic consideration of consumer behavior in economic transformations within advanced capitalist economies.

Some accounts argue that production is now consumer-driven (Piore and Sabel, 1984), while others emphasize technological change or regulatory institutions in causing economic change (Harvey, 1989). It has been argued that both sets of literature inadequately conceptualize consumption, and that a system of provision approach allows for a more balanced treatment of the relationships between production and consumption (Leslie and Reimer, 1999). Fine and Leopold (1993) also recognize the changing relationships between the symbolic meanings of goods and their physical contents. Thus it is important to trace not only commodity movements over space, but also the discourses and representations operating in systems of provision.

Commodity Circuits

A third approach to commodity chain analysis, tracing commodity circuits, combines work in the political economy tradition with ideas derived from poststructural social theory. The notion of a circuit of culture has been used in examining the ways geographical knowledge of commodity systems is shaped and reshaped and in showing how particular cultural artifacts move through the circuit. Unlike the notion of a chain, circuits have no beginning or end, but instead Cook and Crang (1996) envision dense webs of interactions between all sites, with commodities interrelating with other goods as they travel. Rather than privileging one site, such as production, in interpreting the meaning of a commodity, explanation lies in combining a number of different processes. An appropriate politics of consumption recognizes the fragmentary and contradictory nature of the knowledges through which commodity systems are imagined (Cook and Crang, 1996). Rather than assuming that consumers have little awareness of the quality and geographical origins of commodities, this approach sees commodity circuits constructed and reconstructed through consumer knowledge. This view might be criticized for eventually abandoning the concept of a chain altogether and for neglecting ideas of consumer resistance and activism (Leslie and Reimer, 1999).

Recent Tendencies in Commodity Chain Analysis

Although useful in terms of conceptualizing the connections between firms and actors across space in the production of final outputs, these versions of the commodity chain concept fail to connect the sign of the commodity at the consumption end with labor and the natural environment at the production end. In not following through at the consumption end, world systems theory misses a vital opportunity to connect commodity chains with advertising, popular culture, and the geography of consumption. By not tracing firms' activities back to the conditions of workers, economic geography reduces labor to being yet another input in the overall production process, divorcing economic analysis from the politics of production. By not rooting commodity production in the natural environment, all these approaches fail to connect with political ecology, an important emerging area in critical theory (Peet and Watts, 1996). Significantly, these instances of neglect occur at the most politically sensitive sites along commodity chains, preventing the concept from uniting, in a comprehensive blend of politics, consumption, culture, labor, and nature.

Expanded Version of Commodity Chains

In a more complex model of commodity chains, the effects of the commodity at the consumption node, the social and natural conditions at the production node, and the complex intersections of commodity chains at various intermediating and terminal points, are added to a simple model of commodity movement. This expanded version integrates the generation of cultural (signifying and representational) effects at the consumption node with the social and natural conditions at the production end of the chain, the purpose being to show the effects of one on the other, bringing home to consumers the results of consumption. In contrast to earlier theories, this expanded formulation sees commodity movements (the vertical dimension) connecting nodes that are conceived as places (horizontal dimensions) located at different positions along commodity chains. This emphasis on place lends the analysis potential for bringing together several aspects of the production and consumption of commodities. In this vein, the commodity chain is posited as a way of tying together material and signified realities, consumption and production, and activities separated by space, providing a fuller interpretation of the material world and the real, human actors who live in it (Hartwick, 1998).

This version has underlain various works in which commodity chain analysis is tied with issues of development and global trade. The seminal work by Bardnt follows the tomato commodity chin from Mexican fields to Canadian supermarkets, exploring all the various nuances of the chain, temporally and spatially (Bardnt, 2002). Others have explored various commodity chains such as cocoa and cotton, moving from the local level to the points of where these goods are consumed miles away (Ryan, 2011; Moseley and Gray, 2008). A more recent aspect is to add speculation in commodity markets like the New York Mercantile Exchange as the controlling mechanism. Many commodities are traded as 'futures' divorced from real, material goods, and even further

removed from workers, environments, and even consumers who pay higher prices for the same commodity (Peet, 2011).

Consumption Nodes

At the terminal nodal point of consumption, the radiating image of a commodity produces a halo effect of socially constructed meanings. Such meaning systems can be analyzed in a number of ways, including the aesthetics of representation and consumer taste. The preferred theoretical tradition is the semiotics of Peirce (1931–1935). In such a materialist social semiotics, a commodity has two meanings: a first-order meaning at the level of function; and a second-order imputed meaning at the level of symbolic effect. Semiotic analysis tries to reveal the social prescription of the meaning of a commodity. In a postmodern sign economy, first-order meanings, based in physical needs, have been transformed into second-order representational meanings, emanating from social status and cultural allusion (Gottdiener, 1995). However, for a complete explanation, the conditions of production must be included within social semiotic analysis. Stark contrasts in material and cultural conditions so intensify second-order symbolic meanings that they mutate into a third, specifically geographical, meaning-form. Images are usually added to commodities at, or near, the organizational centers of global trading systems, where the producers of sign value share the cultural norms of the consumers. Underdevelopment, unequal exchange, profit flows and power differences, vastly different conditions of life, must be hidden behind the image of the commodity to give the impression that consumption is effortless and has no negative side effects. In general, adding social semiotics to commodity chain analysis expands the horizontal analysis of commodities at the consumption end and creates awareness of vertical linkages to producers (Hartwick, 1998).

Production Nodes

The conditions of production at the other terminus of the commodity chain include the social relations of production, relations with the state, relations between workers and families, hence the reproduction of the household. This, in turn, consists of social relations of reproduction (gender, class, ethnicity) and social relations with nature, including the worker-family's reproductive use of subsistence land and resources. The complete set of the conditions of existence of the producers, and the full range of natural conditions of production that enter into the materiality of the commodity, and pass thereby into the distant bodies of consumers, are explored. Of importance too is the relation between commodity production, inequality and development: many Third World states are trying to escape their roles as mere suppliers of raw materials and move into more advanced stages of the chain (Talbot, 2002; Heintz, 2003).

Commodity Bundles

Commodity chains articulate one with another at various nodal points of the intersection of their pathways. At the production end, commodities of prime significance form the conditions of existence for secondary commodity chains. At other significant nodes, processing points in particular, chains converge into commodity bundle that make up imaged complexes entering the consumption node – the entire complex combines to signify status, wealth, and power through radiating interconnected significations. As Glennie and Thrift (1992) note, any commodity chain is characterized by leakiness, in both material and symbolic ways, which influences other commodity chains – they catalog the intertwining between, advertising of linked products, consumer reflexivity, and consumption across several commodity systems. Nodes of consumption are thus points of convergence for series of imaged complexes of commodity chains. In this last sense, each purchase in the market is seen as culminating in an entire complex of commodity production and representation, linking together several lines of waged workers, unwaged reproducers of waged workers, and natural resources.

Commodity Chains and the Politics of Consumption

Geographical analysis centered on commodity chains connects the generation of cultural (signifying) effects within consumption to the real and material (social and natural) conditions involved in production. Deconstruction, understood in a geographical and materialist sense, uncovers these hidden dimensions of the commodity sign. Moving to a geographical focus on place broadens commodity analysis giving it greater political potential as a critical tool. Using this device, a series of local worlds are linked by connecting-chains into global systems. Commodity chains enable a greater awareness of the geographical implications of the political economy of consumption. Informed by commodity chain analysis, consumers learn to deconstruct, or critically analyze, the disconnected messages that bombard them by reconnecting images with material realities. However, companies also learn too in terms of adding a veneer of social responsibility to their normal business practices (Hughes, 2006). Also the focus on people-in-places, as key nodal points along a commodity chain, broadens economic analysis so that social, cultural, and natural relations can be incorporated. This gives commodity chain analysis greater political potential as a comprehensive political tool. Formulated in this way commodity chains are said to constitute an analytical basis for a transformative politics (Hartwick, 2000).

See also: Hierarchies and Markets; International Business; Multinational Corporations; Network Analysis; Supply Chain Management.

Bibliography

Bair, J., 2005. Global capitalism and commodity chains: looking back, going forward. Competition and Change 9, 153–180.
Bair, J., 2008. Frontiers of Commodity Chain Research. Stanford University Press.
Bardnt, D., 2002. Tangled Routes. Rowman and Littlefield, Lanham.
Cattaneo, O., Gereffi, G., Staritz Global, C., 2010. Value Chains in a Postcrisis World. A Development Perspective. World Bank.

Cook, I., Crang, P., 1996. The world on a plate: culinary culture, displacement and geographical knowledges. Journal of Material Culture 1, 131–153.
Fine, B., Leopold, E., 1993. The World of Consumption. Routledge, London.
Gereffi, G., 1994. The international economy and economic development. In: Smelser, N.J., Swedberg, R. (Eds.), The Handbook of Economic Sociology. Princeton University Press, Princeton, NJ, pp. 206–233.
Gereffi, G., Korzeniewicz, M., Korzeniewicz, P., 1994. Introduction: global commodity chains. In: Gereffi, G., Korzeniewicz, M. (Eds.), Commodity Chains and Global Capitalism. Praeger Press, Westport, CT, pp. 1–4.
Gereffi, G., Memedovic, O., 2003. The Global Apparel Value Chain: What Prospects for Upgrading by Developing Countries? UNIDO, Vienna.
Gereffi, G., Humphrey, J., Sturgeon, T., 2005. The governance of global value chains. Review of International Political Economy 12, 78–104.
Glennie, P.D., Thrift, N.J., 1992. Modern consumption: theorizing commodities and consumers. Environment and Planning, D: Society and Space 11, 603–606.
Gottdiener, M., 1995. Postmodern Semiotics: Material Culture and the Forms of Postmodern Life. Blackwell, Oxford, UK.
Hartwick, E., 1998. Geographies of consumption: a commodity chain approach. Environment and Planning, D 16, 423–437.
Hartwick, E., 2000. Towards a geographical politics of consumption. Environment and Planning, a 32, 1177–1192.
Hassler, M., 2003. The global clothing production system. Commodity Chains and Business Networks Global Networks 3, 513–531.
Harvey, D., 1989. The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change. Blackwell, Oxford, UK.
Heintz, J., 2003. The New Face of Unequal Exchange: Low-wage Manufacturing, Commodity Chains, and Global Inequality. PERI Working Paper Series Number 59.
Hopkins, T.K., Wallerstein, I., 1986. Commodity chains in the world economy prior to 1800. Review 10, 157–170.
Hughes, A., 2006. Learning to trade ethically: knowledgeable capitalism, retailers and contested commodity chains. Geoforum 37, 1008–1020.
Hughes, A., Reimer, S. (Eds.), 2004. Geographies of Commodity Chains. Routledge, London.
Leslie, D., Reimer, S., 1999. Spatializing commodity chains. Progress in Human Geography 23, 401–420.
Moseley, W.G., Gray, L.C., 2008. Hanging by a Thread. Ohio University Athens, Ohio.
Palpaceur, F., 2008. Bringing the social context back in: governance and wealth distribution in global commodity chains. Economy and Society 37, 393–419.
Peet, R., 2011. Contradictions of finance capitalism. Monthly Review 63, 18–32.
Peet, R., Watts, M., 1996. Liberation Ecologies: Environment, Development, Social Movements. Routledge, London.
Peirce, C., 1931–1935. Collected Papers of Charles Sanders Peirce. Harvard University Press, Cambridge, MA.
Piore, M., Sabel, C., 1984. The Second Industrial Divide: Possibilities for Prosperity. Basic Books, New York.
Ryan, Orla, 2011. Chocolate Nations. ZED Press, London.
Talbot, J.M., 2002. Tropical commodity chains, forward integration strategies and international inequality: coffee, cocoa and tea. Review of International Political Economy 9, 701–734.
Yeung, H.W., 1994. Critical reviews of geographical perspectives on business organizations and organizations of production: towards a network approach. Progress in Human Geography 18 (1994), 460–490.

Common Law

Simon Roberts, London School of Economics and Political Science, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article on the common law examines its nature, utilization, historical development, geographical diffusion, theorization, and likely future development.

Introduction

The term common law first came into use in England during the thirteenth century. It referred to the 'common' repertoire of norms and understandings that had begun to be recognized across the whole realm during the latter half of the previous century, as central government was becoming consolidated. These widespread common understandings, which were seen to contrast with surviving customs of a simply localized character, became institutionalized as law as they were invoked to underpin the decisions of judges in the royal courts. In short, this evolving common law claimed to represent the immemorial understandings prevailing in the social world and was at the same time put to use as an instrument of rule by the courts as they developed into a distinctive branch of government. This primary sense of the term common law has survived and continues in use, particularly as a means of contrasting this source from others that the English courts now draw on concurrently – 'statute' and 'equity' (q.v.). But the term has now acquired a general secondary sense in identifying the broader legal domain formed through the diffusion of English law under the process of colonial expansion. This common law world is contrasted with that occupied by the codified civil law systems of continental Europe directly inspired by Roman law and those originating in them.

Historical Development in England

While the common law is seen to originate in the everyday practices and understandings of the social world, its growth from the twelfth century took place in the context of dual, linked processes: the progressive consolidation of central government and the gradual emergence of a specialist legal profession. This growth was realized through the discursive formulation of 'law' in the pleadings of lawyers and the judgments given in the Royal courts. It was associated from an early stage, and remains associated today, with one primary assumption and a distinctive method. The assumption was that the common law sought to realize and secure the practices and understandings of everyday life (immemorial custom). The strength of this assumption can be illustrated through numerous examples, such as the continuing invocation of the reasonable man in arriving at the appropriate standard of care in respect of claims arising from civil wrongs (tort); the care taken to ascertain normal business practice in the area concerned when judging commercial disputes; and the deference accorded to the customs of conveyancers by courts dealing with property disputes. The method distinctive to the common law lay in the procedure under which lawyers recovered the rules to underpin their arguments, and judges to justify their decisions, from the records of earlier judgments. This procedure, the doctrine of precedent, was made possible by the maintenance from an early stage of semiofficial written records (the *Year Books*) and subsequently sustained by private series of reports before official law reporting became institutionalized. Consistency in decision making was reinforced by the understanding that courts should wherever possible follow their own previous decisions and were obliged to follow those of courts superior to them in the judicial hierarchy (the doctrine of *stare decisis*). This procedure can be contrasted with the mode of decision associated with the continental civil law world under which the judge seeks the rules to ground a decision in an authoritative, ultimately codified, text.

The working practices that fashioned the common law were facilitated by the close relationship prevailing from medieval times between the judges and the specialist groups that ultimately became the legal profession. Notwithstanding the strong historical link between adjudication and the consolidation of government, in England there has never been a career judiciary. Even in recent times, the higher ranks of the judiciary have been recruited exclusively from the legal profession, with appointment to the 'bench' representing the ultimate career stage of the successful lawyer. While some elaborate ritual, including the conferment of knighthood in the case of the higher judiciary, marks the transition from barrister to judge, and some formal distance is subsequently maintained between judges and their former colleagues, judges remain sociologically very much part of the professional group from which they emerge. They continue to belong to long-established networks of information exchange and support – often crossing generations – begun in the great private schools, continued at university, and subsequently prevailing in the collegiate groups to which barristers belong (the Inns of Court). So the courts can be just as easily seen as the apex of the legal profession as a specialized branch of government.

The common law that developed in the King's courts through resort to precedent has traditionally been contrasted with two other sources of law. The first of these has been described by legal historians as arising directly out of the formal working practices of the King's courts. These were characterized from an early stage by procedures requiring anyone who sought redress to bring the claim within one of a limited number of elaborate written formulae (known as

writs). If some mistakes were made in the pleadings, the claimant would be denied his remedy. Ensuing direct appeals for help to the King, which such apparent injustices gave rise to, led to the growth of a parallel remedial jurisdiction presided over by the Lord Chancellor. This jurisdiction, known as equity, had a distinct body of doctrines, rules, and practices, and was long administered in separate courts.

Both the common law and equity evolved alongside, and are contrasted with, statute law – rules promulgated by the sovereign. While progressively larger areas have become at least partially regulated by statute – notably in the fields of crime, commercial law, and property – a comprehensive legislative scheme of the kind introduced through the great continental codes of the eighteenth and nineteenth centuries has never been attempted in England. This more limited role and reach of legislation provides a further contrast with the civil law regimes of continental Europe.

The 'rediscovery' of Roman law in Italy during the eleventh century is generally presented by historians as having a limited effect on the growth of the common law in England. But England was in no sense isolated from this learning. Some of the clerics who accompanied William I to England were distinguished as scholars of Roman law, so Justinian's great *Digest* was well known there from an early stage. Holdsworth also makes it clear that Bracton, a senior judge writing in the mid-thirteenth century, used Roman law to fill the gaps left by his extensive repertoire of case histories. But while Roman law provided part of the intellectual climate within which a unified legal order evolved in England during the medieval period, its use as a direct source was limited. Even during the English Renaissance, the legal historian Maitland notes that the common law was still characterized by "an amazingly continuous persistence of medieval doctrine." He attributed this durability to the presence of the collegiate groups, the Inns of Court, within which English lawyers associated. He saw the medieval Inns as 'schools of law' that provided arenas of communication within which young lawyers were taught and their seniors formulated a coherent repertoire of legal doctrine. He wrote: "What is distinctive of mediaeval England is not parliament, for we may everywhere see in Europe assemblies of estates … but the Inns of Court and the Year Books that were read therein, and we shall hardly find their like elsewhere." By the time of the great continental codifications of the eighteenth and nineteenth centuries, which were closely informed by Roman law, the evolving understandings and practices of English judges and lawyers were already part of a largely autonomous, differentiated legal order. By then, the common law tradition was too well set to be disturbed.

Geographical Diffusion of the Common Law

While originating in England, the common law has undergone wide diffusion as English settlers carried it with them abroad, and the government at Westminster imposed it upon annexed or protected territories in the course of colonial expansion. Across India, North America, Australasia, and considerable areas of Africa and Asia, the common law remains today a significant component, linking together a range of independent legal orders. Following the now ancient precedent of the United States of America, most former colonies have retained significant elements of the common law, justifying the shared label of the common law world. But the sheer diversity of contemporary arrangements within this loose association inevitably makes any generalization hazardous.

English settlers establishing colonies overseas took with them the general law of England as it stood at the moment of first settlement. By this means, the common law was established as the foundational law of the United States and those regions of what became the British Commonwealth that were subject to settlement. In these cases, very little recognition was initially given to the existing laws and customs of indigenous peoples. In those further territories that were conquered, annexed, or declared protectorates as the British colonial empire subsequently expanded, the statutes of general application, common law, and doctrines of equity in force in England at the time were typically imposed by Order in Council from Westminster. In these latter cases, notably in India and Africa, early recognition was given to diverse local legal orders encapsulated in the course of colonial expansion. In the case of Africa, this recognition extended to indigenous dispute institutions. Elsewhere, notably in Canada and Australasia, recognition of the prior arrangements of indigenous peoples was postponed well into the twentieth century (*see* Indigenous Rights).

Once English law, including its major common law component, had been received in a colony, it remained in force until repealed or modified by the local colonial legislature or courts – or by the overarching authority represented by the Parliament at Westminster and the Privy Council (the ultimate appellate tribunal of the Empire). These initial linkages between the metropolis and the colonial world have had variable duration. In the case of the American colonies, they were severed early on by the Declaration of Independence. During the twentieth century, they progressively disappeared in the self-governing Commonwealth countries, but some former colonial territories have chosen to retain a final appeal to the Privy Council.

Beyond the impetus given by the immediate necessity of explicit constitutional provision, the 'fresh starts' generated by the colonial enterprise resulted in vigorous, eclectic localized growth for the common law. In these new local contexts, the energy to attempt rationalization, even toward codification, was perhaps greater than in the metropolis. Major initiatives included the codification of wide areas of private law and the criminal law undertaken in India. The Indian Penal Code was subsequently reused in modified form in several other colonial territories. In the United States of America, with the particular problem of its numerous private law jurisdictions, a general attempt at rationalization has been the American Law Institute's Restatement of the Law. This initiative, now in its third generation of restatements, has brought together judges, lawyers, and professors in the preparation of an agreed statement of common law across the state jurisdictions. In the same context, the great American law schools have on the whole self-consciously attempted to present an overall view of the common law in their teaching and research (*see* Legal Education). Elsewhere across the postcolonial world, the superior courts in different jurisdictions have become increasingly attentive to each other's formulations, indicating explicitly in

their judgments that they are working together on a shared overarching fabric that is in the widest sense a common law.

Theorization of the Common Law

Attempts to 'theorize' the common law, in the sense of drawing together rules and procedure for the purposes of exposition and commentary, were made from a very early stage. Holdsworth even proposed that Glanvil's compilation, prepared sometime toward the end of the twelfth century, "marks the true beginning of the common law." Judge Bracton's great treatise, compiled on the basis of 2000 cases that he had collected from the plea rolls, followed in the mid-thirteenth century. Both Glanvil and Bracton were deeply involved in what was becoming specialized, legal administration and for the remainder of the medieval period the common law continued to reside in practice, receiving discursive formulation only in the judgments of the courts and in authoritative commentaries. All the great legal texts of the period were prepared by practitioners whose learning was acquired through their work and in the context of their specialist associations, the Inns of Court.

Just as native legal theory in the medieval period concentrated on parochial commentary, that of the early modern period continued to focus largely on the immediate practicalities at hand. During the constitutional crisis early in the seventeenth century, Chief Justice Coke eloquently prioritized the relative authority of judges, King (he "hath no prerogative but that which the common law allows him"), and Parliament. But Coke's arguments were not distanced speculation; they rather reflected the workings of a practical mind, the holder of high office seeking to justify his position in the urgent context of struggles for power.

Prior to the mid-eighteenth century, the common law had never been the subject of exposition and research in English universities. English university teaching of the common law can be said to begin with the lectures that William Blackstone, the first Vinerian Professor, delivered at Oxford in 1758. These lectures were subsequently published in his famous *Commentaries* (1765–69). The *Commentaries* took more the form of popular exposition than scholarly treatise; but the account he gave provided a framework that would remain readily recognizable even 200 years later. In this work, the common law takes pride of place as the law by which "proceedings and determinations in the king's ordinary courts of justice are guided and directed". The validity of common law doctrines, "depending on immemorial usage for their support", is determined by the judges. In reaching judgment "it is an established rule to abide by former precedents where the same points come again in litigation."

It was only in the early nineteenth century that a body of specifically legal theory, focused on the common law tradition yet distanced from practice, began to develop in England. This body of theory links law firmly to government and maintains a consistent focus on adjudication. The link between government and law is established at the outset in the work of John Austin, the first Professor of Jurisprudence at University College, London. In *The Province of Jurisprudence Determined* (1832), Austin presents law as the command of a sovereign directed toward members of a society in a 'habit of obedience,' with failure to comply being met with 'sanctions.' That central association safely in place, English jurisprudence turned to focus – virtually exclusively over the next 175 years – on providing theoretical foundations for adjudication.

The first modern attempt to provide a general theory of adjudication was made by Henry Maine in *Ancient Law* (1861). In this work, taking a broad comparative, historical sweep, he insisted on locating the core of law in adjudication and in the emergence of rules underpinning judicial decision making. He began by mapping out a broad transition from small kin-based groups to larger, territorial units. The very origin of social life was identified in the steering role exercised within a group of kin by the senior male agnate. These old patriarchs made decisions on an ad hoc basis; no consistent rules underpinned the decisions they took, yet government was supposed to be by adjudication by the senior male, before whom all disputes were brought. In the form of society that followed, numbers of these small groups of agnates became clustered together under chiefs, but the (sometimes fictional) assumption of shared kinship remained the basic organizing principle. Then came the territorial stage, in which members identified themselves through their common occupation of a defined tract of land, rather than through kinship. Around the end of the second stage and the beginning of the third, 'law' developed as rulers began to pronounce the same judgments in similar situations, providing their decision making with an underlying set of rules. Later in the development of territorially based societies, the settlement of disputes fell into the hands of a specialized elite, who alone had access to the principles to be followed in their resolution. As Maine wrote: "What the juristical oligarchy now claims is to monopolize the knowledge of the laws, to have the exclusive possession of the principles by which quarrels are decided." There followed the 'era of codes,' and so on; but we can leave the developmental process at this stage.

Several important features are clear from this summary. First, there is the focus on dispute resolution and the foregrounding of a particular mode of intervention. For Maine, there were no structural changes in the dispute settlement process over the fundamental stages of societal development. From the senior male agnate onward, disputes were resolved by decision, handed down by a third party; there was no suggestion of negotiatory modes of settlement giving way to processes of third-party adjudication. Second, the presence of a normative basis for decision making was the key attribute of law for Maine, and the emergence of this feature heralded the transition from the prelegal to the legal world. Third, there was the later development of specialization as legal rules became separated off from other rules operating in society. So for Maine, social life is the product of 'government,' and the key variables lie in the identity of the third parties responsible for adjudication and in the evolving criteria underpinning decisions.

After Maine, the perfection of theories of law as adjudication occupied the jurisprudence of the common law world through the rest of the nineteenth century and much of the twentieth century. Much of this scholarship was generated in North America where thorough-going reexamination was made of the doctrines of precedent and *stare decisis*. The greater part of this work remained within the traditional commentary mode

of legal scholarship, exemplified by Goodhart's now classical attempts to refine the doctrine of precedent. More ambitiously, scholars like Frank and Llewellyn of the legal realist school drew on sociological perspectives in questioning vigorously whether these doctrines actually operated in the manner claimed in native legal theory. They indicated the relative freedom of judges to distinguish previous decisions they did not want to follow, suggesting that rules were drawn from the repertoire to justify desired outcomes as opposed to determining what those outcomes would be. This iconoclastic strand survived on to the end of the twentieth century in the work of the critical legal studies movement.

One notable attempt to escape from the grip of adjudication, and to think about law unconstrained by the institutional shapes of the modern state, was made by Hart (1961). Rejecting Austin's command theory as simplistic and in some respects inaccurate, Hart's elucidation of law lays central emphasis on rules. One essential category of these is made up of 'primary rules' imposing obligations. The distinguishing mark of such a rule is the seriousness of the pressure to conform to it: "Rules are conceived and spoken of as imposing obligations when the general demand for conformity is insistent and the social pressure brought to bear upon those who deviate or threaten to deviate is great."

While Hart concedes that a society might exist with primary rules only, they alone do not constitute a legal system. For the step from 'the prelegal into the legal world' to be made, these primary rules must be supplemented with 'secondary rules' of three kinds: rules of recognition making it possible to identify what the primary rules are; rules of change providing for the alteration of primary rules when the need arises; and rules of adjudication empowering certain people to determine whether on a given occasion the primary rules have been broken. Most systems will also contain other secondary rules authorizing the application of penalties where primary rules are violated, but this further refinement is not seen as essential to a legal system.

Through the device of these secondary rules, Hart appears to disassociate his model from any parochial institutional content and invest it with considerable value for cross-cultural purposes. However, the escape seems to be illusory, leaving Hart's account of law firmly grounded in the institutions of the state and state-sponsored adjudication. He himself recognizes that the three categories of secondary rules imply, and require, the presence of legislative and adjudicatory agencies. While Hart struggled to marginalize adjudication by identifying law as an affair of rules, subsequent exegeses of his work hastened to re-present his core idea as a theory of adjudication. It is significant too that his most influential successor, Ronald Dworkin, has consistently refocused the task of jurisprudence on understanding and justifying judicial decision making.

The Direction of the Common Law

Close links with government and a core role in providing trial and judgment have historically been central to the understandings of what a 'court' is in the West. Under the ideology of the rule of law, 'justice' has been conceptualized in terms of command realized through reasoned judgment, and judges have been thought of virtually exclusively in those terms. This dominant neoclassical image of the court persists in the cosmologies of the common law world, an apparently essential and enduring component of the ideology of legalism. Yet it does so in the face of widely recognized transformations. Over the last 30 years, it has become generally acknowledged that lawyers routinely use the procedural path of litigation for their settlement negotiations and that the number of civil trials has gradually decreased. Latterly also, the articulated priorities of government, and of the courts themselves, have shifted dramatically. Sponsorship of settlement is now explicitly identified as the courts' primary responsibility, with supervision of trial and the delivery of judgment becoming residual, the solution of last resort. So common law courts are today sites where the profoundly different rationalities that ground rule-based adjudication and negotiated agreement coexist and interact.

In England, for example, this 'culture of settlement' has been advocated by the higher judiciary, adopted as government policy, enshrined in a new regime of civil procedure, and increasingly realized in court practice. In broad, schematic terms, the procedural regime introduced through the Civil Procedure Rules of 1998 promoted settlement through two radical changes. First, through Pre-Action Protocols, a wholly new prelitigation field is marked out under which information exchange and negotiation is required prior even to the filing of a claim if the possibility of later cost sanctions is to be avoided. So the procedural framework is now designed to enforce and police negotiations in the prelitigation phase, with governmental intervention reaching out into what was previously the private sphere. Second, once a claim is filed, the dispute moves into a phase of 'case management' during which the parties are subject to judicial supervision as they move along the procedural route at the end of which lies the theoretical possibility of trial and judgment. So even once 'litigation' is formally commenced, settlement continues to be prioritized.

In the face of these profound transformations, a clearer, empirically based understanding of what happens in today's common law civil courts is required. This needs, as Martin Shapiro argued in his seminal book *Courts* (2005), to be supplemented with 'a more general theory of the nature of judicial institutions.' But the argument here demands going beyond Shapiro's call to expand the neoclassical model to allow for a mediatory element. The folk model of the court as agent of trial and judgment no longer provides a plausible analytic starting point. Instead, recognizing that civil courts no longer correspond to the neoclassical model, and that they today reconceptualize their primary role as the sponsorship of settlement, there is a need to begin again by examining empirically both how they are in practice used by their lay clientele and the professionals representing them, and how court personnel have both actively sponsored and responded to that changed use.

How then should the court be reconceptualized in the light of the new ideological priority attached to 'settlement' in the legislative and judicial cosmologies? Clearly, an idealized conceptualization of the court as agent of trial and judgment can no longer be maintained when this is no longer the way in which the judges represent themselves or an accurate reflection of the operational reality on the ground. The conclusion must be that it is time to reconceptualize the court as an arena for

structured bilateral negotiations, one in which the dominant mode of interaction is an often protracted conversation between the parties themselves. This is a drama in which both active judicial sponsorship of settlement and the theoretical availability of trial and judgment provide a background structure to the process. But beyond that they play a residual role.

Well before the procedural reforms introduced at the end of the twentieth century, it was widely acknowledged that most civil cases settle. But with the officially sanctioned move of courts to provide arenas for bilateral negotiation, the dislocation between form and substance increases. The courts – buildings with now often empty courtrooms – take on an increasingly symbolic character, just as the judges now appear as 'exemplary,' ceremonial figures, legitimating other peoples' decision making. Arguably, with this decline in the importance of command, the decay of an operational/practical role, the ceremonial, ritual dimension increases in importance. Symbolism and ritual are what the court provides. No longer required to mask and justify the power exercised through the court's own commands, these dimensions now legitimize the decision making of the parties themselves.

See also: Indigenous Rights; Legal Education; Norms; Professions and Professionalization, History of.

Bibliography

Austin, John, 1832. The Province of Jurisprudence Determined. J. Murray, London.
Blackstone, W., 1765–69. Commentaries. Clarendon Press, Oxford.
Dworkin, Ronald, 1977. Taking Rights Seriously. Duckworth, London.
Frank, Jerome, 1930. Law and the Modern Mind. Coward, New York.
Goodhart, Arthur L., 1931. Essays in Jurisprudence and the Common Law. CUP, Cambridge.
Hart, H.L.A., 1961. The Concept of Law. Clarendon Press, Oxford.
Holdsworth, William, 1909. A History of English Law. Methuen, London.
Holmes, Wendell, Oliver, 1881. The Common Law. Harvard, Cambridge, MA.
Llewellyn, Karl N., 1962. Jurisprudence: Realism in Theory and Practice. University of Chicago Press, Chicago, IL.
Maine, Henry, 1861. Ancient Law. J.Murray, London.
Maitland, F.W., 1907. Select Essays in Anglo-American Legal History. CUP, Cambridge.
Shapiro, Martin, 2005. Courts. University of Chicago Press, Chicago, IL.

Common Rule: US Regulatory Protections for Human Research Participants

Carl H Coleman, Seton Hall University School of Law, Newark, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Common Rule, more formally known as the Federal Policy for the Protection of Human Subjects, applies to human subjects research conducted or supported by 15 US federal agencies, as well as to privately funded research conducted at most institutions that receive federal research funding. Covered research must be reviewed and approved by an institutional review board, which must determine that the risks of the study have been minimized; the risks are reasonable in relation to the anticipated benefits; adequate plans exist for obtaining participants' informed consent; and other ethical issues have been appropriately addressed.

The US regulatory protections for human research participants emerged in the early 1970s, following a series of public revelations about unethical medical experiments. In a 1966 article in the *New England Journal of Medicine*, the physician Henry Beecher documented 22 'unethical or questionably ethical studies' in the United States, including experiments in which mentally retarded children were deliberately infected with the hepatitis virus and a study in which live cancer cells were injected into indigent, chronically ill, elderly patients (Beecher, 1966). A few years later, the *New York Times* reported a study that the US Public Health Service had been conducting in Tuskegee, Alabama, over a period of four decades, in which treatment was deliberately withheld from poor, African-American men with syphilis so that scientists could study the 'natural course' of the disease (Jones, 1993).

In response to these and other incidents, Congress enacted the National Research Act in 1974. The Act requires any entity that receives federal funds to conduct biomedical or behavioral research involving human subjects to establish oversight committees known as 'institutional review boards,' or IRBs. The Act also established the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research and charged it with identifying basic ethical principles for research and recommending guidelines for research supported with federal funds. The result of the Commission's work was the 1979 Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. The report identified three basic ethical principles relevant to all research with human participants: respect for persons, or the protection of individual autonomy; beneficence, or the maximization of benefits and minimization of risk; and justice, or the equitable distribution of the benefits and burdens of research (National Commission, 1979).

In 1981, the Department of Health and Human Services (DHHS) revised its regulations governing research with human subjects, using the ethical principles articulated in the Belmont Report as the overarching framework. Ten years later, DHHS, along with 14 other federal agencies that conduct, support, or otherwise regulate human subject research, promulgated a new set of regulations called the Federal Policy for the Protection of Human Subjects, or, more simply, the Common Rule. Each of these 15 agencies has published an identical version of the Common Rule in its own section of the Code of Federal Regulations (CFR). References to the Common Rule usually refer to the version published by DHHS, which is located at 45 CFR part 46.

The Common Rule includes only Subpart A of 45 CFR part 46, which contains the general requirements applicable to all research involving human participants. Part 46 contains three additional subparts dealing with special categories of research: Subpart B covers research with pregnant women, fetuses, and neonates; Subpart C covers research with prisoners; and Subpart D covers research with children. There is also a Subpart E, which sets forth requirements for IRB registration. These additional Subparts are not part of the Common Rule, and only some of the other Common Rule agencies have adopted them.

The Common Rule applies to any research that is conducted or supported by one of the agencies that has adopted it. For example, if the National Institutes of Health (NIH), which is part of DHHS, awards a grant or contract to a university to conduct a human research study, that study would be subject to the Common Rule requirements. In addition, any institution that receives federal research funding must enter into an 'assurance' with the government that explains the standards the institution will apply to nonfederally supported research, such as studies sponsored by industry or private foundations. At a minimum, the institution must agree that nonfederally supported research will comply with the general ethical principles articulated in the Belmont Report. Most institutions, however, go further and agree that all of their research will adhere to the full Common Rule requirements, even studies that are not supported by federal funds (Coleman et al., 2005).

The applicability of the Common Rule does not depend on where the research will be conducted; even studies conducted entirely outside the United States may fall within the Common Rule's reach. However, the Common Rule recognizes that "[w]hen research covered by this policy takes place in foreign countries, procedures normally followed in the foreign countries to protect human subjects may differ from those set forth in this policy." To accommodate these differences, it authorizes agencies to permit institutions to rely on foreign procedures if those procedures 'afford protections that are at least equivalent' to those contained in the Common Rule. Yet, because DHHS has never deemed any other country's human research protections to be 'equivalent' to those in the Common Rule, the option of substituting foreign procedures remains more theoretical than real (Dove et al., 2014).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.11024-4

Research covered by the Common Rule may also be subject to additional human participant protection requirements. For example, any study involving products regulated by the Food and Drug Administration (FDA) – i.e., drugs, medical devices, or biologics – must also comply with the FDA's human subject protection rules, regardless of the study's funding source. Health care organizations and their 'business associates' are also required to comply with privacy and security regulations enacted pursuant to the Health Insurance Portability and Accountability Act. Institutions that receive NIH funding are subject to additional rules regarding the disclosure and management of research-related conflicts of interest. In addition, the Common Rule does not displace state laws that provide additional protections to research participants, nor does it preclude research institutions from adopting stricter rules as a matter of their own institutional policies.

Activities Covered by the Common Rule

The Common Rule applies only to activities that meet the definition of 'research' with 'human subjects.' Research is defined as "a systematic investigation … designed to develop or contribute to generalizable knowledge." This definition is intended to exclude situations in which the information being collected will only be relevant to the particular individuals who are involved in the project, such as a single individual or small group of people in a particular setting (Coleman et al., 2005).

It is not always easy to determine when an activity is 'designed to develop or contribute to generalizable knowledge.' Consider, for example, a hospital that collects data from the medical charts of surgery patients to better understand hospital-specific circumstances leading to infection, such as the choice of nursing staff or the scheduling of procedures. On the one hand, because the data collected from this intervention will be specific to the particular hospital, one might conclude that the results are not capable of being generalized to other settings, where the staffing and scheduling issues are likely to be different. On the other hand, staffing and scheduling issues that cause problems at one hospital might also cause problems at other hospitals with a similar patient and staff makeup, potentially making it possible to generalize the results of the intervention to other settings (World Health Organization, 2013).

In situations like this, it has been suggested that the applicability of the Common Rule depends on the intent, or purpose, of those who designed the activity. As explained by the US Centers for Disease Control and Prevention (CDC), "If the purpose is to develop or contribute to generalizable knowledge, the project is research." The CDC has defined term 'generalizable knowledge' as "new knowledge that has relevance beyond the population or program from which it was collected, or information that is added to the scientific literature" (Centers for Disease Control and Prevention, 2010). Thus, if the purpose of those who designed the activity were simply to provide information for an internal process of quality improvement, the activity would not constitute research and would not be subject to the Common Rule.

Even if a particular project meets the definition of 'research,' it will not be subject to the Common Rule unless it also implicates the regulatory definition of 'human subject.' The Common Rule defines a human subject as "a living individual about whom an investigator … obtains (1) data through intervention or interaction with the individual; or (2) identifiable private information." Thus, studies involving the review of documents containing individually identifiable personal information would implicate 'human subjects' under the definition, even if the investigators never interact with the individuals. Similarly, studies in which investigators speak or otherwise interact with individuals would also involve human subjects, even if no individually identifiable information were ever obtained.

The Common Rule identifies some activities that – despite meeting the definition of 'research' with 'human subjects' – are, nonetheless, exempt from the Rule's usual requirements. The exemption that is most relevant to the social and behavioral sciences applies to research involving the use of educational tests, survey procedures, interview procedures, or observations of public behavior. Such studies are exempt from the Common Rule unless (1) the researcher will be recording identifiable information about the participants; and (2) disclosure of the information could place the participants "at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation." In light of this exemption, survey or interview studies in which no identifiable information about the participants will be recorded are exempt from the Common Rule, even if the questions being asked are sensitive and may cause the participants psychological distress.

Although research that falls into one of the exemption categories is not subject to the Common Rule, most institutions require investigators who engage in such research to submit their studies to the IRB, so that the IRB can confirm that the requirements for the exemption have actually been satisfied. Some institutions go further and subject exempt research to at least a minimal form of ethical review. By contrast, activities that do not meet the regulatory definition of 'research' or 'human subject' can be undertaken without any involvement of the IRB at all.

Institutional Review Boards

The Common Rule gives IRBs the primary responsibility for ensuring that studies subject to its jurisdiction conform to the Common Rule and the ethical principles articulated in the Belmont Report. Studies subject to IRB review may not be commenced until the IRB has affirmatively granted approval. In addition, IRBs must provide 'continuing review' of ongoing studies at least once per year.

IRBs must have at least five members with a diversity of experience and expertise. IRBs may not consist "entirely of members of one profession," and "every nondiscriminatory effort" must be made to ensure that IRBs include both men and women. In addition, each IRB must include "at least one member whose primary concerns are in scientific areas and at least one member whose primary concerns are in nonscientific areas," as well as "at least one member who is not otherwise affiliated with the institution and who is not part of the immediate family of a person who is affiliated with the institution." Certain categories of studies deemed to involve no more than minimal risks to participants may be reviewed by expedited procedures, in which a single IRB member has the

authority to grant approval without the involvement of the full board. A list of studies eligible for expedited review is available on the DHHS web site.

The 'I' in IRB refers to the fact that most of these committees operate as part of the institutions conducting the research. The National Commission considered this sort of 'local' review as an essential part of the system of human subject protection. The Commission reasoned that "local committees have the advantage of greater familiarity with the actual conditions surrounding the conduct of research" and can ensure that the process of ethical review incorporates local attitudes and values (Peckman, 2001).

In recent years, however, two important developments have challenged the locally based IRB system. The first is the emergence of independent IRBs that are not part of any research institution – in essence, 'noninstitutional review boards.' These boards are hired by research sponsors, including commercial entities such as pharmaceutical and medical device companies, and often operate on a for-profit basis. Some commentators believe that the fact that these IRBs are paid by research sponsors creates an inappropriate conflict of interest. However, others suggest that these conflicts of interest are less problematic than those that arise in traditional institutional-based IRBs, where reviews are conducted by researchers' professional colleagues (Heath, 2001).

The second development challenging the locally based nature of IRB oversight is the increasing use of multisite studies. While it is still common for multisite studies to undergo IRB review at each institution where the study will be conducted, this practice has been criticized as both ineffective and inefficient (Menikoff, 2010). Various mechanisms have been developed to reduce the use of duplicative IRB reviews, including procedures allowing institutions to delegate their review responsibilities to other institutions and the use of centralized or multi-institutional IRBs (Wagner et al., 2010).

Criteria for Approval of Research

Before an IRB approves a study, it must determine that the risks to participants are minimized; the risks are reasonable in relation to anticipated benefits; the selection of participants is equitable; the process for obtaining and documenting informed consent is appropriate; and, when appropriate, there are adequate provisions for data and safety monitoring and ensuring privacy and confidentiality. The regulations also provide that 'additional safeguards' must be provided if participants "are likely to be vulnerable to coercion or undue influence, such as children, prisoners, pregnant women, mentally disabled persons, or economically or educationally disadvantaged persons." The Common Rule does specify what these additional safeguards might entail, although, as noted above, DHHS and a few other federal agencies have adopted additional regulations applicable to research with pregnant women, prisoners, and children.

Risk-Benefit Assessment

Assessing the risks and potential benefits of research is one of the IRB's core responsibilities. First, the IRB must determine that the risks have been 'minimized' by using procedures that are 'consistent with sound research design' and 'do not unnecessarily expose subjects to risk.' The regulations provide that, 'whenever appropriate,' researchers should use procedures that are "already being performed on the subjects for diagnostic or treatment purposes."

Second, in addition to ensuring that the risks have been minimized, the IRB must determine that the risks are "reasonable in relation to anticipated benefits, if any, to subjects, and the importance of the knowledge that may reasonably be expected to result." This balancing process is inherently uneven. On the risk side, the IRB's focus is limited to the risks to the subjects in the study. The regulations specifically provide that IRBs "should not consider possible long-range effects of applying knowledge gained in the research (for example, the possible effects of the research on public policy) as among those research risks that fall within the purview of its responsibility." On the benefits side, however, the IRB considers not only potential benefits to subjects, but also the "importance of the knowledge that may reasonably be expected to result." In light of this language, it is clear that the IRB may approve research even if there are no anticipated benefits to the subjects in the study, provided the knowledge expected to be gained from the research is sufficiently important.

In assessing the risks of research, IRBs are expected to consider not only the potential for physical harm, but also the possibility of psychological, social, economic, and other types of injuries (Prentice and Gordon, 2001). For example, in some studies, the unauthorized release of sensitive personal information could cause embarrassment or possibly discrimination and harassment. The existence of these risks does not mean that the study cannot be conducted, but the IRB needs to assure itself that the risks have been minimized as much as reasonably possible, and the anticipated benefits of the study are sufficient to justify the potential harms.

According to most commentators, when IRBs consider potential benefits to subjects, their focus should be on the direct benefits resulting from the research intervention itself – that is, the possibility that the intervention will produce its intended beneficial outcomes. In some cases, subjects may be offered various incentives to participate, such as free meals, small gifts, or even payments. However, these benefits are generally not considered to be the kind of direct benefits that IRBs should use to offset research-related risks (King, 2000).

Informed Consent

General Requirements for Informed Consent

With a few exceptions described below, researchers subject to the Common Rule must obtain written informed consent from the subject or, if the subject is unable to provide consent directly, the subject's 'legally authorized representative.' The term 'legally authorized representative' is defined as "an individual or judicial or other body authorized under applicable law to consent on behalf of a prospective subject to the subject's participation in the procedure(s) involved in the research." Because state law usually governs the process of providing consent on behalf of individuals who lack decision-making capacity, the effect of this provision is to delegate the definition of 'legally authorized representative' to the law of the

state where the research is taking place (or, if the research is taking place abroad, the law of the relevant foreign jurisdiction).

The Common Rule provides that informed consent must be obtained "only under circumstances that provide the prospective subject or the representative sufficient opportunity to consider whether or not to participate and that minimize the possibility of coercion or undue influence." It also requires that information provided during the consent process must be "in language understandable to the subject or the representative." This requirement implies not only that English-language consent materials must be translated for non-English speakers, but also that the level of the language used (whether English or otherwise) is likely to be understandable by the target population.

The Common Rule prohibits the use of "exculpatory language through which the subject or the representative is made to waive or appear to waive any of the subject's legal rights, or releases or appears to release the investigator, the sponsor, the institution, or its agents from liability for negligence." While this provision protects subjects' right to seek compensation for negligently caused injuries, it does not require compensation for harms caused in research that are not due to negligence. In this respect, the Common Rule differs from most international guidelines on research ethics, which require compensation for all research-related injuries, whether or not they are attributable to negligence (Presidential Commission, 2011).

Office for Human Research Protections (OHRP) has emphasized that informed consent is a 'process' that begins with "the initial approach of an investigator to potential subjects." This means that any advertisements or solicitations for subjects must be submitted to the IRB for review. IRB review is not required, however, for listings on web sites that are limited to the title and purpose of a study, a summary of the protocol, basic eligibility criteria, study site locations, and contact information (Table 1).

Unless otherwise waived by the IRB, subjects or representatives must document their informed consent in writing. Researchers may either ask the subject or representative to sign a written document that embodies all the required information, or instead to sign a 'short-form written consent document' stating that the required elements have been provided orally. If the short-form option is used, a witness must be present during the oral presentation. In addition, the IRB must approve a written summary of what is to be said during the oral presentation, and this summary must be given to the subject or representative and signed by both the person obtaining consent and the witness.

Waivers of the General Requirements for Informed Consent

The Common Rule authorizes IRBs to waive or alter the usual informed requirement in two sets of circumstances. First, the regulations allow the waiver or alteration of consent in studies related to public benefit or service programs, if the studies are to be conducted by or subject to the approval of state or local governments, and the research "could not practicably be carried out without the waiver or alteration." OHRP has suggested that this waiver would be appropriate in "a study of identifiable private information about program benefit recipients using 20-year-old records … if current contact information for those recipients is not available."

The second, are far more frequently used, waiver provision applies to certain types of 'minimal risk' research. The term 'minimal risk' is defined in the regulations to mean "that the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance

Table 1 Elements of informed consent under the common rule

Unless the IRB has authorized a waiver or modification of the usual consent requirements, the following information must be provided to each prospective research participant:

1. A statement that the study involves research, an explanation of the purposes of the research and the expected duration of the subject's participation, a description of the procedures to be followed, and identification of any procedures that are experimental;
2. A description of any reasonably foreseeable risks or discomforts to the subject;
3. A description of any benefits to the subject or to others, which may reasonably be expected from the research;
4. A disclosure of appropriate alternative procedures or courses of treatment, if any, that might be advantageous to the subject;
5. A statement describing the extent, if any, to which confidentiality of records identifying the subject will be maintained;
6. For research involving more than minimal risk, an explanation as to whether any compensation and medical treatments are available if injury occurs and, if so, what they consist of, or where further information may be obtained;
7. An explanation of whom to contact for answers to pertinent questions about the research and research subjects' rights, and whom to contact in the event of a research-related injury to the subject; and
8. A statement that participation is voluntary, refusal to participate will involve no penalty or loss of benefits to which the subject is otherwise entitled, and the subject may discontinue participation at any time without penalty or loss of benefits to which the subject is otherwise entitled.

When appropriate, one or more of the following elements of information shall also be provided:

1. A statement that the particular treatment or procedure may involve risks to the subject (or to the embryo or fetus, if the subject is or may become pregnant), which are currently unforeseeable;
2. Anticipated circumstances under which the subject's participation may be terminated by the investigator without regard to the subject's consent;
3. Any additional costs to the subject that may result from participation in the research;
4. The consequences of a subject's decision to withdraw from the research and procedures for orderly termination of participation by the subject;
5. A statement that significant new findings developed during the course of the research, which may relate to the subject's willingness to continue participation will be provided to the subject; and
6. The approximate number of subjects involved in the study.

Source: 45 C.F.R. Sections 46.116(a) and (b).

of routine physical or psychological examinations or tests." If the IRB concludes that the research involves no more than minimal risk, then it is authorized to waive or alter the usual consent requirements provided it further determines that doing so "will not adversely affect the rights and welfare of the subjects" and the research "could not practicably be carried out without the waiver or alteration." In addition, it must consider whether it would be appropriate to provide subjects with 'additional pertinent information' about their participation after the intervention is over.

One of the most common areas in which IRBs are asked to waive informed consent is psychological research. Psychologists often seek permission to withhold information from subjects because, if the subjects knew what was going to happen to them and why, that knowledge could alter the very behavior that is being studied. A famous example of this so-called 'deception' research was Stanley Milgram's experiments in the 1960s designed to examine obedience to authority, which is discussed elsewhere in this volume. Deception research requires researchers to withhold one of the basic elements of informed consent – an explanation of the purposes of the research. As such, an IRB may not approve the use of deception unless it determines that the study involves no more than minimal risk; withholding information about the purpose of the study will not adversely affect the subjects' rights and welfare; the study could not practicably be carried out without deception, and that subjects will be debriefed about their experience after the intervention is over.

In addition to the provisions for waiving or altering the usual informed consent requirement, the Common Rule also allows IRBs to waive the usual requirement of obtaining a signed consent form. Such waivers are allowed if either the only record linking the subject and the research would be the consent document and the principal risk would be potential harm resulting from a breach of confidentiality; or the research presents no more than minimal risk of harm to subjects and involves no procedures for which written consent is normally required outside of the research context. An example would be a study in which a single blood sample will be taken from HIV-positive individuals, which will then be examined in a laboratory for purposes unrelated to the subjects' health care. In this case, the risk from the potential disclosure of the subjects' HIV status would likely be greater than the risk from the blood draw, which is minimal. Therefore, to protect the subjects' confidentiality, the IRB could decide to waive the requirement for obtaining a signed consent form, thereby enabling the study to proceed without any individually identifiable information ever being recorded.

Other Approval Criteria

The IRB's obligation to ensure the equitable selection of subjects stems from the ethical principle of justice, which requires a fair distribution of the benefits and burdens of research. The Common Rule directs IRBs to "take into account the purposes of the research and the setting in which the research will be conducted" and to consider "the special problems of research involving vulnerable populations, such as children, prisoners, pregnant women, mentally disabled persons, or economically or educationally disadvantaged persons." OHRP guidance emphasizes the importance of protecting vulnerable populations without depriving them of the benefits of research or the opportunity to participate.

The Common Rule also requires IRBs to ensure that studies provide adequate provisions for data monitoring in order to ensure the safety of subjects. The level of monitoring required will depend on the risks involved in the study. While formal data and safety monitoring plans are most often used in studies involving medical interventions, they also may be appropriate in social and behavioral research in which adverse events are foreseeable. Adverse events in social and behavioral research could include breaches of confidentiality, emotional distress, or impact on subjects' social relationships.

Finally, the IRB is required to ensure that adequate protections exist to protect the subjects' privacy and confidentiality. Commonly used protections include storing sensitive information in locked cabinets or password-protected computer files and limiting the number of people who have access to the information. In some situations, the best way to protect subjects' privacy may be to encourage the researchers not to collect identifiable information in the first place. As discussed above, if the only involvement of human subjects consists of educational tests, surveys, interviews, or observations of public behavior, and the researchers do not record any identifying information, the study would qualify for an exemption to the Common Rule.

Enforcement of the Common Rule

OHRP is the federal agency responsible for enforcing the Common Rule. As part of carrying out this mission, it enters into agreements, or 'assurances,' with institutions that conduct federally funded research. These assurances require institutions to affirm that their researchers will comply with all of the Common Rule requirements. Institutions are required to enter into assurances with OHRP if their employees or agents are 'engaged' in nonexempt human subject research. Engagement in research is defined as intervening or interacting with living individuals for research purposes or obtaining individually identifiable private information for research purposes.

OHRP also enforces the Common Rule through by conducting site visits, either randomly or in response to complaints. If OHRP determines that the Common Rule is not being followed, it may require the institution to engage in corrective action, including rereviews of protocols, implementation of new database management strategies, and education and training of investigators or IRB members. OHRP may also restrict, attach conditions to, or suspend an institution's assurance, or recommend the suspension or removal of an investigator or institution from a particular project. The most extreme sanction OHRP can recommend is debarment of an individual or institution. Debarment is a government-wide sanction that precludes the debarred person or entity from contracting with the federal government for any purpose.

The Common Rule does not provide a 'private right of action,' meaning that it does not give individuals the right to sue for violations of its provisions. However, some courts have held that the Common Rule's informed consent provisions create a standard of care "against which informed consent in

human subjects research will be compared where the case involved inadequate informed consent" (Mastroianni, 2006). As a result, violation of the Common Rule's informed consent provisions may be relevant if a research participant is injured and brings a state-law claim based on a theory of lack of informed consent.

See also: Anthropological Research, Ethics of; Ethical Guidelines and Codes of Conduct in Social and Behavioral Research; Ethics for Biomedical Research Involving Humans: International Guidelines; Fieldwork: Ethical Aspects; Informed Consent of Research Subjects; Pharmaceutical Industry: Political Economies of Drugs and Knowledge; Placebo-Controlled Trials, Ethics of; Privacy of Individuals in Social Research: Confidentiality; Research Ethics, Cross-Cultural Dimensions of; Research Funding: Ethical Aspects; Research Publication: Ethical Aspects ; Stem Cell Research and Cloning; Surveys and Polling: Ethical Aspects; Therapeutic Effectiveness Assessment: Approaches in Contemporary Research.

Bibliography

Bankert, E.A., Amdur, R.J., 2006. Institutional Review Board: Management and Function, second ed. Jones and Bartlett, London.

Beecher, H., 1966. Ethics and clinical research. New England Journal of Medicine 274, 1354–1360.

Centers for Disease Control and Prevention, 2010. Distinguishing Public Health Research and Public Health Nonresearch. http://www.cdc.gov/od/science/integrity/docs/cdc-policy-distinguishing-public-health-research-nonresearch.pdf.

Coleman, C.H., Menikoff, J.A., Goldner, J.A., Dubler, Nancy Neveloff, 2005. The Ethics and Regulation of Research with Human Subjects. LexisNexis, Newark, NJ.

Dove, E.S., Knoppers, B.M., Zawati, M.H., 2014. Towards an ethics safe harbor for global biomedical research. Journal of Law & the Biosciences 1, 3–51.

Grady, C., 2010. Do IRBs protect human research participants? Journal of the American Medical Association 304, 1122–1123.

Heath, E., 2001. The history, function, and future of independent institutional review boards. In: National Bioethics Advisory Commission, Ethical and Policy Issues in Research involving Human Participants, vol. 2. National Bioethics Advisory Commission, Bethesda, MD, pp. E1–E23.

Jones, J.H., 1993. Bad Blood. The Free Press, New York.

King, N.M.P., 2000. Defining and describing benefit appropriately in clinical trials. Journal of Law, Medicine & Ethics 28, 332–343.

Mastroianni, A.C., 2006. Liability, regulation and policy in surgical innovation: the cutting edge of research and therapy. Health Matrix 16, 351–442.

Menikoff, J., 2010. The paradoxical problem with multiple-IRB review. New England Journal of Medicine 363, 1591–1593.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979. The Belmont Report (accessed 30.11.13.). http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html.

Peckman, S., 2001. Local institutional review boards. In: National Bioethics Advisory Commission. Ethical and Policy Issues in Research Involving Human Participants, vol. 2. National Bioethics Advisory Commission, Bethesda, MD, pp. K1–K35.

Prentice, E.D., Gordon, B.G., 2001. Institutional review board assessment of risks and benefits associated with research. In: National Bioethics Advisory Commission, Ethical and Policy Issues in Research Involving Human Participants, vol. 2. National Bioethics Advisory Commission, Bethesda, MD, pp. L1–L16.

Presidential Commission for the Study of Bioethical Issues, 2011. Moral Science: Protecting Participants in Human Subjects Research. Presidential Commission for the Study of Bioethical Issues, Washington, DC.

Wagner, T.H., Murray, C., Goldberg, J., Adler, J.M., Abrams, J., 2010. Costs and benefits of the National Cancer Institute central institutional review board. Journal of Clinical Oncology 28 (4), 662–666.

World Health Organization, 2013. Ethical Issues in Patient Safety Research. http://www.who.int/patientsafety/research/ethical_issues/en/.

Relevant Websites

http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html – The Common Rule, 45 C.F.R. Part 46.

http://www.fda.gov/scienceresearch/specialtopics/runningclinicaltrials/default.htm – Food and Drug Administration Human Subject Protection.

http://www.hhs.gov/ocr/privacy/index.html – HIPAA Privacy Rule.

http://grants.nih.gov/grants/policy/coi/ – National Institutes of Health Conflict of Interest Policy.

http://www.hhs.gov/ohrp/index.html – Office for Human Research Protections.

Common Sense, Anthropology of

Michael Herzfeld, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Common sense is not uniform across all sociocultural groups. Anthropologists typically confront the differences among 'common senses' at the very start of ethnographic fieldwork; their informants treat as self-evident actions and other phenomena that to the anthropologists are opaque and sometimes even absurd. From anthropology's relativist and empirical perspective, claims to a universal rationality are themselves a parochial form of common sense. Western-inspired theoretical models such as rational choice theory thus actually resemble everyday habits of reasoning and rest on culturally parochial and historically shallow assumptions of universal validity. Still earlier claims to a higher form of reason, used to defend atheism and agnosticism against religious authority, are similarly revealed as local versions of common sense, and thus as culturally as well as historically contingent. By removing common sense from the realm of the banal, and treating it instead as worthy of critical dissection even as it appears in their own ordinary assumptions, anthropologists also bring their reflexive and critical understanding to bear on the relationship between common sense and power.

Frederic Harrison, in *The Philosophy of Common Sense* (1907: 9), seeks to understand the universal qualities of common sense as a reconciliation of science and religion. In so doing, he appears to take a position entirely consonant with modern relativism: "It is usual entirely to put aside the testimony of all the uncivilised or semi-civilised races. And thus countless myriads of intelligent human beings, as completely as our ancestors, as entirely links in the chain of progress as our own parents, are abstracted from the inquiry into the innate qualities of the human mind." On closer inspection, however, his argument rests on evolutionary assumptions that were central to Victorian epistemology and that constitute a core area of what is held to be common sense in much of the Western world today. This illustrates a problem that is at the heart of the present discussion: how to achieve a position of relative independence from the object of inquiry when that object is assumed to inform the analysis itself? Much as a critic of structuralism might suspect that its claims about human cognition make its analyses circular, so any investigation of common sense inevitably starts from common sense assumptions.

The solution, I suggest, lies in the comparative approach that remains the core of most theory in social anthropology. If we do not assume that common sense is universal but instead tackle the forms it takes in different cultural settings, we may achieve some purchase on this slippery topic. As Mary Douglas (1972: 28) succinctly expressed the matter, "We have to see that the categories and actual principles which we find in our own world present the same problems of rational justification that baffle us in the exotic worlds of foreigners." This is true – perhaps especially so – when we examine our own theory making. As the archaeologist Robert Dunnell (1982: 12), remarks, "In one way, common sense is culture – it determines the kinds of observations, the rules for assembling those observations into sense, and even what constitutes sense. As a sense-making system, common sense is functionally equivalent to theory in the sciences." The risk, as he points out, is that, once we recognize that all common sense is in fact culturally specific, we must also see that its role in theory building is inevitably ethnocentric. Perhaps this is why scholars have so often, as Jovchelovitch (2008: 432–436) has argued, often background the relationship between science and common sense as much as possible. For social and cultural anthropologists, there would be two main ways of addressing this weakness, in addition to keeping it in plain sight at all times: sustained awareness of the scholar's own role in the construction of the data and the arguments, marked, as I have already noted, by the principled use of the first-person pronoun; and a commitment to maintaining and increasing the comparative study of common sense claims in a context that also permits the critical inclusion of our own.

The latter requirement requires that we focus on the extraordinary variety of logics that constitute common sense for their exponents, and, very often, for them alone. We do not have to agree with Harrison's (1907: xiii) rather truculent observation that "the mass of what is called Metaphysics is the fruitless search after insoluble puzzles: a search which it is wise to understand as an intellectual gymnastic, but whereon nothing practical, real, or true can be built" – a curious anticipation of Arendt's (1978) recognition that philosophy and common sense are often at war with each other. His statement should stand, rather, for the attitude that we usually understand by the term common sense itself, an attitude that demands conformity with what seems to be obvious (Miceli, 1982: 583) – in his case, the evolutionary superiority of the Western intellect, impatience with its more pedantic manifestations, and an agnostic stance toward religion. Indeed, Harrison's argument represents only a slight retreat from the most extreme manifestation of *le bon sens* in the eighteenth century, when some writers, paradoxically and rather contemptuously countering what they perceived as the opinion of the masses, saw it as a validation of atheism (Rosenfeld, 2011: 118–128).

Reason, in those earlier phases of Western thought, was attributed an existence and a purity that raised it above the heedless assumptions of common belief; it was actively opposed to common sense and exempted from cultural specificity (Arendt, 1978). The idea that any form of reason must itself have been historically and culturally contingent – that it

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12039-2

was a form of common sense – would have scandalized both these writers and their Victorian successors such as Harrison. It may still offend many who begin their arguments with phrases like 'it stands to reason' – phrases that often turn out to be straightforward appeals to the idea of a universal common sense. In this respect, modern religious antirelativism (see Remotti, 2008) often comes remarkably close to the authoritarianism of the eighteenth-century atheists as well as to that of some currently popular models of 'rational choice' (on which see Tambiah, 1990: 119–120) – but the similarity is enabled by widely divergent views on what constitutes common sense itself. This example alone suffices to show why a universalist understanding of common sense simply does not work. No anthropologist today would accept the argument that reason is independent of context – the position that Arendt identified as a characteristic of most of the history of Western philosophy.

A comparative approach recognizes that the obvious, or, in Douglas's (1972) terms, self-evidence, is only obvious or self-evident from an insider's perspective; the obviousness itself is constructed and contingent. To the exponents of any claim of obviousness, in any given culture, it stands to reason that the self-evident requires no further inquiry – exactly the attitude expressed by Harrison. When two versions of common sense collide, however, as in Pope Benedict XVI's attack on relativism and the anthropologist Francesco Remotti's (2008) reply invoking the cultural specificity of the forms of common sense, that specificity – famously described by Antonio Gramsci (see Bates, 1975) as the conceptual realization of hegemony – may become manifest; but it will only do so if one is willing to take a relativistic position even toward relativism itself. In short, the approach requires particularistic methodologies; it is in the detailed examination of how common sense is actually used that its cross-cultural inconsistencies are revealed.

Historical research and ethnographic fieldwork thus offer the most promising paths to exploring these specificities and resisting the seductive generalizations that they generate. In the realm of politics, it is the polemical use of the term 'common sense' that shows us, if we can suspend our partiality in the actual dispute at hand in any given context, how its meanings respond to the ideological and political demands of specific situations. Anthropologists' preferred method of ethnographic fieldwork is especially well suited to decoding the underlying assumptions of local forms of common sense. This is because anthropologists are constantly surprised by unfamiliar assumptions of this kind; these often conflict with the scholars' own assumptions about what is 'natural.'

Most anthropologists tell stories about moments of sudden insight; they are almost always about revelations of different kinds of common sense. When I first asked the shepherds in a mountain village among whom I was doing research why they stole sheep from each other (Herzfeld, 1985), for example, they argued that they did so in order to gain each other's friendship. I was initially nonplussed but intrigued. For their part, the shepherds managed to convey the distinct impression that I was not grasping something self-evident, obvious, and, indeed, commonsensical. Their reaction was one of amused astonishment that I could not understand that a tough man had to display his toughness in this way if he wanted to create alliances with other shepherds (Herzfeld, 1985: xiii, 174–183).

Their response was reminiscent of Evans-Pritchard's (1937) Zande informants' reactions to his initial failure to grasp why the collapse of a granary on a man's head was not the result of termite damage alone but must have entailed witchcraft for it to have happened at that particular moment; the explanation was determined by the social circumstances of the event. That example has led to heated debates about the extent to which non-Western people can be said to possess a conceptual framework compatible with the production of scientific thought (Crick, 1976: 121; Tambiah, 1990: 130–139; Winch, 1972: 8–49). What did emerge with clarity from these debates is the realization that logic is not the prerogative of any one cultural tradition or even educational level – a point that was paralleled within the Western world by Labov's (1972) remarkable demonstration of some scholars' inability to recognize logic in statements by young urban African Americans in which the ideas were not structured according to scholarly criteria of syntax and precision.

The anthropologist's first task in the field, as these examples illustrate, is to understand why an apparently meaningless explanation 'works' for local people – why, in other words, it satisfies their criteria of common sense. The explanation may be a straightforward question of translation, but it more frequently resides in different perspectives on what constitutes common sense. In Crete, I was confronted by the question of how to decode a logic (we have to steal if we want to have friends) that seemed obvious to my respondents but was initially opaque to me. When it became apparent that in their tough environment these men needed assurances that those with whom they allied themselves were reliably fierce and agile, it began to make sense that they would introduce themselves by carrying out a raid. It also became clear to me that I was adjusting my own ideas about common sense. I had been trained to do this, to be sure; but the experience, especially in such a dramatic context, produced a momentary sense of dislocation from the safe moorings of the familiar.

That insight, moreover, was probably the starting-point of a much more general reflection that eventually led me to characterize social and cultural anthropology as the comparative study of common sense (Herzfeld, 2001: 1). That characterization – which also contributes to Remotti's defense of relativism – is admittedly somewhat tongue-in-cheek, but it has the advantage of laying out clearly the key challenge, that of convincing practitioners of the other social sciences that obviousness, like boredom (Herzfeld, 2004: 117–120), is inherently worthy of study – "precisely because," as Miceli (1982: 583) remarks, "all cultural objects actually presuppose a point of view about reality and impose it, tacitly, through the simple fact that one uses it and learns to understand its significance and usefulness." How this comes about, moreover, is a central issue for the analysis of power. Why are some things obvious, and to whom, and who has an interest in maintaining that state of affairs? (With boredom, the corresponding questions are those of why boredom should be interesting and what it conceals from other social actors.) How, in short, is obviousness used to maintain or subvert dominant ideologies?

The investigation of these questions requires a very personal commitment to the analysis, since the ethnographer's own version of common sense is a key element in the comparative

exercise that ensues. A relevant aspect of anthropological reportage that sets it apart from the conventions of most other social sciences is the use of the first person in accounts of research. Anthropologists thereby seek to restore agency to their own intellectual endeavors and so to undermine the claims of immunity to challenge that assertions of common sense typically entail. At the same time, this insistence on foregrounding their own presence has methodological consequences that are especially important in investigating common sense – a property they may ostensibly share with their informants while then discovering that in fact their respective understandings of it are quite divergent.

Thus, for example, Lilith Mahmud (2013: 6, 36–37, 43–45), in her fieldwork on female Freemasons in Italy, gained traction by addressing her own presuppositions in the same frame as those held both by her informants and by the larger Italian public. She confronted three sources of common sense. First, as an Italian, she discovered that her own African appearance conflicted with what her informants automatically assumed about how Italians should look. Second, she also discovered in herself significant traces of official attitudes, replete with conspiracy theories, toward the entire Masonic phenomenon. Third, her own surprise at discovering that there were female Masons eventually morphed into a vibrant field experience. Using her own positionality as a *comparandum* proved to be an invaluable tool for disinterring a range of assumptions, including some she had originally shared with many other Italians, and making them an object of analytic interest.

The particular value of ethnography and archival work in exploring the variability of common sense lies in the focus of both methods on well-contextualized detail. Precisely what we mean by context varies, however, so that, from an anthropological perspective, some ethnomethodologists' emphasis on the context of particular conversational situations (e.g., Garfinkel, 1973) may seem too narrow, while philosophers' analyses – even when historically grounded (e.g., Ledwig, 2007: 122) – make little use of context and often appear to dismiss it as trivial. Anthropological research on common sense, by contrast with both these tendencies, ideally seeks to link vertically differentiated contexts (e.g., national and community levels) and thereby show how the assumptions made in the name of common sense work in practice to create a sense of common identities based on shared ideas. It is by focusing on use, rather than only on form and definition, that we may be able to rescue the concept of common sense from its taken-for-granted relegation to the background of both research and everyday politics, and this is consistent with anthropologists' preference – itself the outcome of their commitment to fieldwork – for use or action theories of meaning over more static concepts of reference (see Crick, 1976, for an early exploration of this issue).

A useful starting-point for any investigation of how common sense is used may be found in the work of the sociologist Alfred Schutz. Schutz's insistence (1973: 18) that "the pre-scientific vernacular can be interpreted as a treasure-house of ready-made pre-constituted types and characteristics, all socially derived" describes precisely what should attract the attention of social scientists in the very idea of common sense. True, there is an unfortunate hint of evolutionism in Schutz's invocation of a 'prescientific vernacular;' Gupta (1998: 172–179), in particular, has exposed the commonsensical Western assumption that 'indigenous' and 'scientific' conceptual worlds are separate and unrelated, showing thereby that the forms of common sense we encounter in the field may have multiple (or hybrid) sources, although they tend today to be hierarchized in accordance with residual ideas about the superiority of the 'scientific' over the 'folk.' But Schutz's insight into the socially constructed nature of common sense does not require us to adopt a similarly evolutionist stance, and it does allow us to track the emergence of ideas of common sense and to trace their realizations as various kinds of classification back to their social roots in specific activities and experiences. In this way, Schutz especially paved the way for Mary Douglas's (1966, 1970, 1972) seminal contributions to the analysis of self-evidence and its relationship to taxonomy.

Schutz's was thus a vital move, since one of the more obstinate problems is the insistence of most forms of common sense that there is nothing to analyze. Common sense as a concept suffers from the paradox of parochial universalism. Every village holds that its version of common sense represents a universal truth (and anthropologists are often upbraided by their informants for asking questions about things that are deemed too obvious to merit discussion); some areas of Western scholarship have, *mutatis mutandis*, done much the same. But all claims of common sense, scholarly ones included, emerge, as Schutz argued, from ordinary practices and successful repetition. If a particular version of common sense seems to work, this is because it inevitably rests on a presumption of universal truth grounded in very particular local practices and experiences.

The utility of any version of common sense to the ends of power lies in the ease with which the reality of its context-dependency can be concealed – and often is concealed, for political reasons that animated Gramsci's widely acknowledged reading of common sense as hegemony. Indeed, while – for example – some bank economists are active in creating public understandings of what constitutes common sense, as Douglas Holmes (2013) argues, it is only when we see these economists as anthropologists trying to unravel a specific cultural problem that we realize how situation-specific their current version of common sense actually is. Their actions exactly parallel those of professors and bureaucrats, for example, who pronounce that something 'stands to reason.' The performative force of such declarations (Austin, 1975[1962]) occludes the fact that they have a history, one in which they have been brought into being to express a particular ideology or worldview. If Holmes is right about the economists, whose discipline certainly does have a highly variable history of engagement with ideas of common sense (Coates, 1996), in recent years they have worked to achieve general acceptance for a very particular and highly self-interested perspective, while making that perspective seem, like any nationalism, to be an eternal, universal, and *discovered* verity, rather than a contingent, local, and *invented* opinion. With Gramsci, but with more ethnographic specificity, Holmes shows exactly how people's complicity in their own cooptation is engineered in social practice.

If ethnography allows us to disabuse ourselves of the universal validity of any version of common sense, a diachronic perspective helps us to disinter the cultural and historical

accidents on which the assumption of a universal common sense usually rests. Even from a short-term perspective, we all know from personal experience that what seemed bizarre at one point in our lives has been naturalized as common sense; this is as true, moreover, for our own theories, methods, and writing practices as it is for ordinary actions and policy decisions (see Klein, 2006: 1158). Sophia Rosenfeld (2011) generously acknowledging that the decoding of common sense has been especially the focus of anthropologists such as Clifford Geertz (and therefore primarily synchronic in focus), nonetheless wisely argues that a historically grounded examination of the uses of the term and concept is no less necessary. As she remarks, when common sense is examined historically we see immediately that its tenets are "culturally and temporally variable in content" (Rosenfeld, 2011: 15). Comparison can be cross-epochal as well as cross-cultural; given the chronological depth of the concept of common sense, tracing its translation into various languages and the recorded shifts in meaning in each such language is no less necessary than a consideration of what makes sense to our contemporaries in different societies.

In fact, transposed onto a historical plane, Geertz's (1983: 75) perspective offers an enhanced insight into why common sense so often appears to have the authority of doctrine, sometimes explicitly so ('God-given truths'): "Religion rests its case on revelation, science on method, ideology on moral passion; but common sense rests its [case] on the assertion that it is not a case at all, just life in a nutshell. The world is its authority." Precisely so, but that world (Greek *kosmos*) is itself ground to the figure of cosmology or religion. And the world is not experienced identically, even in the most ordinary of sensory domains (such as sight), in exactly the same way in every social group. The apparent inability of the Abelam people of New Guinea to make sense of a two-dimensional photograph as the representation of a three-dimensional person (Forge, 1968) is no stranger than the similar inability of Western anthropologists to understand why a pattern of lines and squiggles provides an iconic representation of deity – or, for that matter, why one has to steal sheep in order to create new friendships. Geertz understood that common sense was always self-referential: one appeals to common sense in order to justify common sense. That understanding rests on a more radical proposition: that the circular source of the self-confidence we call common sense requires a blissful unawareness – willful or not – of the contingent nature of the ground itself.

Appeals to common sense are inherently political, in at least two important ways. First, they rest on a hegemonic form of understanding; this was Gramsci's great contribution to the way we conceptualize common sense today. Second, common sense is often viewed as the basis of notions of rationality in government; rational choice models rest on such an argument since they usually assume that what constitutes a rational choice is determined by economic and political factors rather than, in part, by cultural values.

The political aspect of common sense has not yet been a major focus of anthropological research; it does not get much of an airing in Geertz's work, for example. This may be about to change, particularly as a burgeoning anthropology of Western societies 'exoticizes' many forms of obviousness. Nonanthropological discussions of the role of common sense in politics, for example, tend to revolve around its relationship to the demarcation between democracy and populism (e.g., Arendt, 1958; see also Rosenfeld, 2011), whereas an anthropologist would typically ask *how* claims of democratic process, transparency, and social justice were being used in everyday practice. One interesting exception to the general pattern is the work of the political scientist Michael Connors (2003: 17) on what he calls "democrasubjection" – "the subjection of people to imaginary forms of their own rule." Connors's analysis, which sought to decode the tactics of former Thai Prime Minister Thaksin Shinawatr (but perhaps applies equally to his self-proclaimed democratic opponents) is nicely paralleled in anthropology by Rosalind Morris's (2004) dissection of the uses of 'transparency' to produce its exact opposite – especially in the form of the concealment of favor-peddling – during the same phase of Thai political history. While neither of these authors specifically addresses common sense as a concept, their analyses, although coming from radically different intellectual traditions, are striking for their convergence in addressing how political rhetoric can exploit commonsensical assumptions and so hide the operation of power.

These studies stand in stark contrast to Arendt's (1958) claims for common sense as the basis of true democracy. Her argument seems to assume that a freely sought consensus will always favor the interests of the majority. Her position, however, ignores the historical evidence that, in democratic systems, electoral majorities often demand policies that entail the persecution of minorities, and it assumes that the shared assumptions of the majority necessarily guarantee the happiness of even those who strongly disagree with the majority view (see also Rosenfeld, 2011: 254–255). The thinker who recognized the banality of evil seems to have overlooked the dangers lurking in the banal.

Moreover, the banal, as Billig (1995) has demonstrated, is the means through which ideologies such as nationalism *become* common sense. These ideologies assert that common sense will automatically show the way to social improvement: "What Georgia as a country needs is a culture of common sense. Nothing would develop this culture better than fostering a habit of basing judgments on facts, rather than on 'values' or political agendas" (Nodia, n.d.). But who is to direct such changes? How will their fairness be guaranteed, and who sets the agenda for, and definition of, fairness? Such positions entail an abdication of agency at exactly the moment at which they appear to embrace it.

Statements of this kind illustrate the process that Bourdieu and others have recognized as the naturalization of culture: "It is because subjects do not, strictly speaking, know what they are doing that what they do has more meaning than they know. The *habitus* is the universalizing mediation which causes an individual agent's practices, without either explicit reason or signifying intent, to be none the less 'sensible' and 'reasonable'" (1977: 79). What goes for individuals goes, a fortiori, for entire social groups, and it is inevitably a political process in the broadest sense. It is for this reason that anthropology's commitment to the analysis of the particular and especially of the details of the relationship between agency and social order

and change makes it an especially salutary basis for a critique of common sense.

But this is also a reason to be very cautious about giving it the status of theory. While our informants *do* theorize, and sometimes in terms that we can recognize as such (Herzfeld, 1985; Labov, 1972), and it is important to recognize that capacity, their theoretical observations are less a matter of what they regard as common sense than they are of carefully considered positions not unlike those of the scholars who study those societies. Unreflective assumptions offer no more analytical advantages in their social lives than they do in our scholarly pursuits. Dunnell's (1982: 12–13) treatment of common sense as an adaptive mechanism, however, suggests that for him it had an evolutionary function, and that this is what invests it with universal value even as we recognize its internal variability. But that argument is itself an appeal to a common sense that while it has remained popular in the media, is no longer universally entertained within anthropology.

Thus, I prefer to argue for a critical deployment of analyses of 'common senses' (a phrase already used by Dunnell) as a counterpoint to the ethnocentric assumptions he rightly identifies as underlying a great deal of commonsensical theorizing – what goes for archaeology is also true for other branches of anthropology. Some may wish to expand that use of common senses still further, arguing for a distinct anthropology of common sense. But anthropology has long been committed to the decoding and comparative analysis of everyday phenomena, and, as such, the discipline has necessarily been at its very core a critique of the claims – the many claims, to be sure – of whatever is locally or globally upheld as common sense. There thus seems little advantage in trying to create a special area or subfield; indeed, doing so would probably reduce anthropology's traction among the social sciences in general, and diminish its relevance to public debate as well. Especially now that some values, including common sense, appear to be gaining a new purchase on universal validity, the particularism of anthropology as a discipline offers a necessary and lively resistance to their sedimentation in some ethnocentric but allegedly global hierarchy of value.

See also: Comparative Method in Anthropology; Cultural Relativism, Anthropology of; Ethnocentrism; Ethnography; Everyday Life, Anthropology of; Hegemony: Anthropological Aspects; Hybridity; Intercultural Communication: An Anthropological Perspective; Knowledge, Anthropology of; Reflexivity in Anthropology; Values, Anthropology of; Worldview.

Bibliography

Arendt, Hannah, 1958. In: Canovan, Margaret (Ed.), The Human Condition, second ed. University of Chicago Press, Chicago.

Arendt, Hannah, 1978. Thinking. The Life of the Mind, vol 1. Harcourt Brace Jovanovich, New York.

Austin, J.L., 1975[1962]. In: Urmson, J.O., Sbisà, Marina (Eds.), How to Do Things with Words, second ed. Harvard University Press, Cambridge, MA.

Bates, Thomas R., 1975. Gramsci and the theory of hegemony. Journal of the History of Ideas 36, 351–366.

Billig, Michael, 1995. Banal Nationalism. Sage, London.

Bourdieu, Pierre, 1977. Outline of a Theory of Practice (R. Nice, Trans.). Cambridge University Press, Cambridge.

Coates, John, 1996. The Claims of Common Sense: Moore, Wittgenstein, Keynes and the Social Sciences. Cambridge University Press, Cambridge.

Connors, Michael, 2003. Democracy and National Identity in Thailand. Routledge-Curzon, London.

Crick, Malcolm, 1976. Explorations in Language and Meaning: Towards a Semantic Anthropology. John Wiley/Halsted, New York.

Douglas, Mary, 1966. Purity and Danger: An Analysis of Concepts of Pollution and Taboo. Routledge & Kegan Paul, London.

Douglas, Mary, 1970. Natural Symbols: Explorations in Cosmology. Barrie & Rockliff, the Cresset Press, London.

Douglas, Mary, 1972. Self-evidence. Proceedings of the Royal Anthropological Institute 1972, 27–43.

Dunnell, Robert C., 1982. Science, social science, and common sense: the agonizing dilemma of modern archaeology. Journal of Anthropological Research 38, 1–25.

Evans-Pritchard, E.E., 1937. Oracles, Witchcraft and Magic among the Azande. Clarendon Press, Oxford.

Forge, Anthony, 1968. Learning to see in new Guinea (A.S.A. Monographs, vol. 8). In: Mayer, Philip (Ed.), Socialization: The Approach from Social Anthropology. Tavistock, London, pp. 269–291.

Garfinkel, Harold, 1973. Background expectancies. In: Douglas, Mary (Ed.), Rules and Meanings: The Anthropology of Everyday Knowledge. Penguin, Harmndsworth, pp. 21–23.

Geertz, Clifford, 1983. Local Knowledge: Further Essays in Interpretive Anthropology. Basic Books, New York.

Gupta, Akhil, 1998. Postcolonial Developments: Agriculture in the Making of Modern India. Duke University Press, Durham.

Harrison, Frederic, 1907. The Philosophy of Common Sense. Macmillan, London.

Herzfeld, Michael, 1985. The Poetics of Manhood: Contest and Identity in a Cretan Mountain Village. Princeton University Press, Princeton.

Herzfeld, Michael, 2001. Anthropology: Theoretical Practice in Culture and Society. Blackwell, Oxford.

Herzfeld, Michael, 2004. The Body Impolitic: Artisans and Artifice in the Global Hierarchy of Value. University of Chicago Press, Chicago.

Holmes, Douglas, 2013. Economy of Words: Communicative Imperatives in Central Banks. University of Chicago Press, Chicago.

Jovchelovitch, Sandra, 2008. The rehabilitation of common sense: social representations, science and cognitive polyphasia. Journal for the Theory of Social Behavior 38, 431–448.

Klein, Lisl, 2006. Applied social Science: is it just common sense? Human Relations 59, 1155–1172.

Labov, William, 1972. Language in the Inner City: Studies in the Black English Vernacular. University of Pennsylvania Press, Philadelphia.

Ledwig, Marion, 2007. Common Sense: Its History, Method, and Applicability. Peter Lang, New York.

Mahmud, Lilith, 2013. The Brotherhood of Freemason Sisters: Gender, Secrecy, and Fraternity in Italian Masonic Lodges. University of Chicago Press, Chicago.

Miceli, Silvana, 1982. In nome del segno: Introduzione alla semiotica della cultura. Sellerio, Palermo.

Morris, Rosalind, 2004. Intimacy and corruption in Thailand's age of transparency. In: Shryock, Andrew (Ed.), Off Stage/On Display: Intimacy and Ethnography in the Age of Public Culture. Stanford University Press, Stanford, pp. 225–243.

Nodia, Ghia. Social sciences and the culture of common sense. In: Anon., (Ed.), How Social Sciences Can Contribute to Changing a Society. Academic Swiss Caucasus Net, Fribourg, n.d.

Remotti, Francesco, 2008. Contro Natura: Una Lettera Al Papa. Laterza, Bari.

Rosenfeld, Sophia, 2011. Common Sense: A Political History. Harvard University Press, Cambridge, MA.

Schutz, Alfred, 1973. The frame of unquestioned constructs. In: Douglas, Mary (Ed.), Rules and Meanings: The Anthropology of Everyday Knowledge. Penguin, Harmondsworth, pp. 18–20.

Tambiah, Stanley J., 1990. Magic, Science, Religion, and the Scope of Rationality. Cambridge University Press, Cambridge.

Winch, Peter, 1972. Ethics and Action. Routledge & Kegan Paul, London.

Communalism

Peter van der Veer, Max Planck Institute for the Study of Religious and Ethnic Diversity, Göttingen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Communalism refers to communities of belonging and especially to such communities in Indian society. More specifically, it refers to the articulation of religious communities into mutually antagonistic, social, political, and economic groups. It emerges within the context of the colonial modernization of Indian society, in which the Indian population was classified, counted, and measured in terms of community. Communalist ideologies emerged in India in the Hindu, Muslim, and Sikh communities. Today they are connected to forms of transnational community.

The concept of 'communalism' is based on that of 'community' and has to be understood in relation to the concepts of 'nation' and 'civil society.' Communalism signifies a politics of belonging that does not emphasize the nation in all its diversity, but the homogenous, religious community. It should be clear that the term communalism comes up as an alternative to nationalism, both in theories that explain the failure of the nation-state to emerge in certain parts of the world and in theories that explain the demise of the nation-state under conditions of globalization. Like the use of the concept of 'tribalism' in the study of African politics, 'communalism' is primarily used in the analysis of Indian society and politics and is seldom encountered in theoretical arguments of a more general nature. In the present era, theories of the decline of the nation-state emphasize the possibility of transnational, virtual communities that depend on new communication technologies. The extent to which these communities articulate a politics of communal belonging that is not territorially bound is striking.

The concept of 'communalism' derives its meaning from arguments about the place in modern society of communities of belonging. According to liberal, political theory, communalism is based on primordial attachments and stands in the way of the formation of a modern, civil society, in which individuals act as equals on the basis of their allegiance to the state in which they enjoy the rights and duties of citizenship. In modernization theories, communalism signifies a hybrid phase between traditional society and modern society, since it combines modern democracy with political mobilization of religious communities in a plural society. Such theories make a sharp distinction between secular nationalism on the one hand and communalism on the other. Inherent in all these theories is an opposition between religion, seen as part of traditional society, and secular modernity. The fact that religious communities form a strong political force in modern societies all over the world cannot be understood in terms of standard modernization and secularization theories.

Communalism then signifies a politics of belonging that does not emphasize the nation in all its diversity, but the homogeneous, religious community. It resembles the current notion of communitarianism in that it refers to a political agenda of communicative interaction within a particular community on the basis of shared 'habits of the heart,' but rather than emphasizing the defense of a 'common way of life,' it is much more aggressively antagonistic toward other, similar communities. It should be clear that the term communalism comes up as an alternative to nationalism, both in theories which explain the failure of the nation-state to emerge in certain parts of the world and in theories which explain the demise of the nation-state under conditions of globalization.

'Communalism' is primarily used in the analysis of Indian society and politics and is seldom encountered in theoretical arguments of a more general nature. It refers to the articulation of religious communities into mutually antagonistic, social, political, and economic groups in India. Again, like 'tribalism' in Africa, communalism in India has a history as colonial theory with great social and political consequences. It emerges in the context of the colonial modernization of Indian society. The Indian population was classified, counted, and measured in terms of community. Notions of caste and religion were crucial in the census operations which took place on a regular basis from the 1870s. British colonial policies were strongly influenced by essentialist assumptions about the nature of Hindus, Muslims, and Sikhs. In their system of governance, the British sought alliances with the 'natural leaders' of communities and, ultimately, this created a system of representation in which membership of a community took precedence over place of residence. Recruitment for jobs both in civil administration and in the military was also organized around notions of race, caste, and religious community. The notion of 'martial races' was specifically applied to the Sikhs, a religious community in the Punjab, North India. The British recruited only 'pure' Sikhs who were following the strict norms of what the British considered to be the proper Sikh faith. In this way, the Sikhs became a dominant section in the Indian military until today, and a particular 'orthodox' form of Sikhism was promoted with great social and political consequences for the region.

The extent to which language politics was influenced by colonial ideologies of communal difference was also important. The medium of communication was supposedly related to communal identity. A major issue was the question of whether Hindi or Urdu should be the language of instruction and the language of government besides English. In the debates about it, Hindi came to be conceptualized as the

language of the Hindu community while Urdu came to be conceptualized as the language of the Muslim community. In the later part of the nineteenth century, Hindu leaders increased their attacks on Urdu and Indo-Persian culture by emphasizing the communal basis of the differences in script and vocabulary between Hindi and Urdu. A new vernacular literature emerged which attacked the joint Hindu–Muslim culture, as expressed in Urdu or the spoken vernacular, Hindustani. The principal arenas for the Hindi–Urdu controversy were schools and entry exams for government service. After the Partition of 1947, Urdu became the national language of Pakistan and Hindi, besides English, the main administrative language of India. The identification of language with religious community was also the ground on which the Sikh community after Independence started to demand a special status for the language of their sacred scriptures.

The legal structure of both colonial and postcolonial India has been greatly influenced by the notion of communal difference. Colonial law was grafted on Hindu and Muslim legal traditions which were interpreted and codified in a joint effort of British legal scholars and native religious specialists. Ancient Sanskrit texts, for instance, became foundational for the creation of Hindu civil law. Personal law in India until the present day is based upon communal difference. The contradictions between the claim that India is a secular state and the existence of communal laws can be briefly illustrated by the Shah Banu conflict of 1985. Shah Banu had been divorced by her husband according to Muslim customary law in 1978, but obtained alimony on the basis of a general criminal law against vagrancy. Her husband appealed to the Supreme Court arguing that Muslim law did not require him to pay alimony, but the court rejected that argument in 1985 with the further comment that the government should draw up a uniform civil code. Muslim organizations quickly responded to this judgment with countrywide agitations, claiming that Muslim personal law was in danger. The then Prime Minister, Rajiv Gandhi (1944–91), gave in to their demands and overturned the Supreme Court decision by making a law which would no longer require a Muslim husband to pay alimony. This in turn was taken up by Hindu communal organizations which objected against the government's pro-Muslim stance, as well as by secular feminists who saw the government action as male chauvinism.

Communalist ideologies emerged in India in a variety of forms during the twentieth century. In the 1920s, V.D. Savarkar (1883–1966) published a book which connected Indian territory with Hindu identity and thus formulated a communalist ideology that made Muslims, Christians, and others into outsiders, at best second-rate citizens. His thoughts continue to play a significant role in the ideology of the Rastriya Swayamsevak Sangh (RSS), the leading Hindu communalist organization founded in 1925. This organization and its many branches are active in trade unions, political parties, and religious organizations, but it primarily tries to train young men in a military fashion to 'defend the Hindu nation.' The RSS has often been at the forefront of political violence against Muslims and Christians, before and after Partition. It was the driving force behind actions to demolish a sixteenth-century mosque in the Hindu pilgrimage center Ayodhya in 1992, an event which made the political party with which it is connected into the leading party of the country.

On the Muslim side, it is the Muslim League and the Jama'at-i-Islami which are, historically, the major communalist organizations. The Muslim League's demand for a separate homeland for Muslims in the early 1940s began the political process which led to the Partition in 1947. The poet Muhammad Iqbal had formulated the wish for separation in his writings, but his main influence had been on Muslim students both in India and abroad. It only received political significance in the 1940s when Jinnah (1875–1948), the leader of the Muslim League, took it up. The speed with which the separationist movement gained support demonstrates the extent to which the political and social system in India was conducive to the communalist articulation of religious difference in India, with immense political consequences until the twenty-first century. The complexity of Muslim communalism can be demonstrated by the fact that the founder of the most militant Muslim organization in South Asia, the Jama'at-i-Islami – and as such the counterpart of the Hindu RSS – Maulana Maududi (1903–79), was in principle against the formation of Pakistan. Partition, however, forced Maududi to move to Pakistan, where his organization became a redoubtable political force in campaigns for an Islamic constitution and laws against un-Islamic practices. His revolutionary ideology had a great influence on the Muslim brotherhoods in Egypt and the Sudan.

Another major communalist development in India is the emergence of the Khalistan movement among Sikhs. Sant Jarnail Singh Bhindranwale (1947–84), a young religious leader, created a large movement, popular largely among the urban youth, demanding a separate state for the Sikhs, Khalistan. It was repressed by the Indian army in 1984. The movement can be partly explained in terms of an identity politics in which Sikhs try to maintain the symbolic boundaries which divide them and the surrounding Hindu majority. These symbolic boundaries, such as the wearing of uncut hair and turbans for men, are only maintained if they are supported by the fact that Sikhs derive material gains from their membership, as they clearly did in the colonial period. The political and economic aspects of membership are intrinsic to the religious ones and it is therefore not useful to understand the Khalistan movement as either a religious or a political movement. It is like the Muslim and Hindu movements, a communalist movement which articulates the religious community in a political and economic way.

In the twenty-first century era, theories of the decline of the nation-state emphasize the possibility of transnational, virtual communities which depend on new communication technologies. The extent to which these communities articulate politics of communal belonging that is not territorially bound is striking. This form of transnational communalism enables a politics which simultaneously connects very diverse issues in very different politics. The best example of this has been the so-called Rushdie affair, which started with the banning of the novel *The Satanic Verses* by the Indian government in 1988 under pressure from Muslim communalist leaders. This was quickly followed up by demonstrations, organized by leaders of 'the' Muslim community (mostly Indian and Pakistani) in the United Kingdom, and then followed by a call for the

execution of the author by the self-proclaimed leader of the Muslims, the Sh'ite cleric, and leader of the Islamic Republic of Iran, Khomeiny. The use of the notion of community in these political actions is widely varied, but simultaneously it allows some sense of unity across the globe. It is strongly tied up with migration to the North by members of religious communities and thus with minority politics and secularism in the countries of immigration. At the same time, it is engaged in communalist projects in the countries of origin, where these communities might constitute the majority, as indeed Sri Lankan Buddhists, Indian Hindus, and Pakistani Muslims do. It is this kind of transnational communalism, strongly supported by new media, which will become more important in the current era of globalization. This should not be understood as the defense of primordial attachments in the face of modernization and globalization, but as completely new forms of community and identity politics.

See also: Asia, Sociocultural Overviews: South Asia; Colonialism, Anthropology of; Communitarianism, Sociology of; Communitarianism; Diaspora; Islam: Asia; Nationalism, Historical Aspects of: South Asia; Religions of India; Transnational Citizenship; Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns; Women in Buddhism.

Bibliography

Ahmad, I., 2009. Islamism and Democracy in India. The Transformation of Jamaat-e-Islami. Princeton University Press, Princeton, NJ.

Dumont, L., 1970. Homo Hierarchicus. The Caste System and Its Implications. Weidenfeld and Nicolson, London.

Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.

Hansen, T.B., 1999. The Saffron Wave. Democracy and Hindu Nationalism in Modern India. Princeton University Press, Princeton, NJ.

Jaffrelot, C., 1996. The Hindu Nationalist Movement. Hurst, London.

van der Veer, P., 1994. Religious Nationalism. Hindus and Muslims in India. University of California Press, Berkeley, CA.

Communes, Anthropology of

Terence MS Evens, University of North Carolina, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article treats, with few exceptions, ethnographical studies of communes. By 'communes' is meant relatively small collective communities that typically regard themselves as alternative to their ambient social orders. Although these communities lend themselves well to ethnographic research, anthropologists have not shown great interest in them. The most notable and substantial ethnographic studies focus on North American Anabaptist settlements (the Hutterian brethren), the Israeli kibbutz, and Japanese communes. Taking their cue from the self-proclaimed alternative character of communes, these studies tend to focus on the institutional and psychological difference collective social arrangements make, and on the ability of the communities to cope given an adverse social surround. The resulting analyses of the tension between collectivist theory and practice, and of the nature of collectivist commitment, are both informative and provocative. But the evident tendency to take as guiding sociological problems, the 'social problems,' in terms of which these communities worry and busy themselves seems to grant too much to the ideological presumption of 'alternative' and to overlook the anthropological advantage of a research perspective gained from the disciplined study of 'other cultures.'

For purposes of anthropological investigation, 'commune' may be taken to refer to relatively small communities, which have set themselves up, typically self-consciously, as alternative to the wider social order, and embrace collectivism in some substantial fashion. They tend to be small enough to operate principally by virtue of face-to-face relations, they are inclined to hold resources and property in common, and usually, they favor communal decision making for political purposes. This article treats the study of such communities by, especially, professional anthropologists or ethnographic researchers.

'Alternative' Communities

The great bulk of social and humanistic science research on communes may be found in the disciplines of history, sociology, and psychology. Anthropological studies are relatively few in number, and only one small book treats communes as a general topic (Hostetler, 1974b). Considering that communes have much in common with the sort of small-scale community that anthropologists have preferred to investigate, it is curious that these scholars have not shown more interest in them. Like village communities, communes lend themselves to the kind of intensive fieldwork diagnostic of ethnography, and are especially accessible to the anthropological aspiration to research communities in the whole.

Perhaps one reason for the relative lack of anthropological attention is the consideration that in the social science literature, communalism has been largely associated with Western social settings. Whereas anthropology has focused on 'other cultures,' most (but not all) communes are not *other* in the relevant sense. They characteristically present themselves as essentially *alter*native to the societal orders in which they arise. However, in the present connection, significantly, the otherness entailed by such alterity can be misleading. For, as inverse and, in part or whole, utopian reactions to the dominant social order, such communities share with the latter rootedness in the same body of social thought. In the West, this body of thought revolves around the question of the relationship between society and the individual, especially when that relationship is apprehended as a dualism. Seeing the dominant social order as primarily individualistic, communal societies propose to resolve the dualism to the contrary, by creating social and cultural conditions that ensure the valuational primacy of the collective over the individual. Typically, the realization of the primacy is keyed to institutional mechanisms that move the individual to grasp his or her self-identity primarily in terms of the collective.

Anthropological and Ethnographic Studies

The most notable and substantial ethnographic studies of communes pertain to three different communal social settings: North American Anabaptist settlements, the Israeli kibbutz, and Japanese communes. Typically these studies take their cue from the presumptively alternative social nature of their subject, and accordingly tend to feature the way in which collectivism affects 'normal' institutions and practices, and the capacity of the community to withstand the opposing pressures of the surrounding social order.

The Hutterian Brethren

Hostetler's work on the Hutterites is exemplarily anthropological, distinguished by its ethnographic comprehensiveness and depth of cultural knowledge. He studied this people over a period of several years, and his publications (see especially Hostetler (1974a) and Hostetler and Huntington (1996)) provide rich descriptions of the 500-year history of the Hutterites, their worldview and critical system of Christian belief and practice, the colonies' economic, political, and domestic organization, and the difficulties experienced by this people in sustaining their unique communal way of life. But the principal focus of his research is socialization. Given the long-term survival of this social order under presumptively

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12041-0

adverse external societal and cultural conditions, Hostetler supposed as the key to this evident success the inculcation of cultural values. Accordingly, he set out to examine the Hutterite child rearing system. He finds that the system of socialization is remarkably effective for reasons of its ideological certitude, utter consistency, and understanding attitude toward modest forms of individual deviancy.

The system is predicated on the Christian belief in the absolute authority of God and the presumption of the primacy of communal order as the key to ensure the submission of the will of the individual self to the will of God. Human nature is seen as dualistically divided between carnality and spirituality. The former quality is understood as in need of firm submission to the latter, a condition meant to be realized by means of communal existence. All the institutional practices of the Hutterites, including, of course, the education and training of children, are designed to function strictly on behalf of this spiritual principle of hierarchical authority. In result, Hostetler concludes, the Hutterite's sense of individual self is only weakly distinguished from his social identity and is predisposed to self-denial and obedience to authority. In view of this analysis, Hostetler (in line with Hutterite opinion, and perhaps tautologically) is inclined to explain defection as a product of aberrant circumstances in which individuality emerges as an unruly force. These circumstances are, he suggests, associated with undesirable competition among family lines and the failure of these families to truly subject their autonomy to the social whole.

In his substantial monograph on the Hutterites, John Bennett (1967), like Hostetler, is interested in how this order of communal life perdures and maintains its traditional integrity, given the contrary nature of the surrounding social environment. But instead of socialization, Bennett focuses on the agricultural economy of the colonies and the impact of technological development. In line with the general tenor of Hostetler's account, he argues that for reasons of their sure sense of identity, extremely efficient management, well-formed social order, and rational approach to economic matters, these communities are exceptionally adaptable. More particularly, because they are very open to technological change but not to consumeristic influence, they are enabled to thrive without corrupting their traditional way of life.

Zablocki's philosophically rich study of the Zablocki (1971), another deeply religious Anabaptist community, less traditionalist than the Hutterites, also constitutes an important contribution to the anthropology of communes. Zablocki is a sociologist by training. But by virtue of his intensive fieldwork in a single settlement, and his attention to the quotidian course of social events, his study is patently ethnographical. He, too, was interested in the relative success of the Bruderhof. He addresses this issue in terms of 'commitment,' a notion pertaining to the fusion of self-interest with the needs of the community and given currency in Kanter's *Commitment and Community*, a sociological study (contemporary with Zablocki's) of communes in general (Kanter, 1972).

In addressing the question of the manner in which the Bruderhof manages to create and sustain the commitment of its members, Zablocki takes, in addition to the standard functionalism of Hostetler's and Bennett's work, a strong psychodynamic turn. Unlike Hostetler, whose approach to the question of commitment typifies the 'culture and personality' school of interpretation, Zablocki draws on psychoanalysis in a way that bears directly on sociological dynamics, not personality. His main finding is that the Bruderhof is characterized by an institutional structure the critical function of which is to call on and employ the ordinarily repressed psychodynamic energy of collective behavior in order to sustain commitment. Hence, when the community suffers a social crisis that threatens commitment on the part of members (such crises are, he finds, inevitable), this structure of resolution – a joyful prayer meeting of the whole – is brought into play, revitalizing commitment. In effect, he depicts a community keyed to something like Rousseau's 'general will' (a concept of 'commitment' if ever there was one), the difference being that in the Bruderhof this will is regarded as accessible (it is thought to preexist its concrete determination) only by means of the realization of a spiritual rather than rational or political unity. By arguing that this spiritual unity is brought about through collective behavior and affective effervescence, Zablocki charges Durkheim's famous but vague sociological thesis, about the transcendent force of society, with a determinate psychoanalytical dynamic.

Underlying Zablocki's concern for commitment is a more general concern about the relation between community and freedom. This concern is of interest to him for reasons beyond the question of the Bruderhof, as he finds that American society, in general, suffers from a need for community. Hence, for him the examination of the Bruderhof carries lessons for the society at large, the most prominent of which is the importance of distinguishing a kind of freedom that is nonindividualistic.

The Israeli Kibbutz

Melford Spiro's first book on the kibbutz (1970) continues to serve as an anthropological introduction to this well-known communal society. It presents usefully and clearly the community's historical beginnings, ideological principles, and institutional organization and practices. It was originally intended to serve as a preliminary chapter to the book resulting from Spiro's basic research project on the kibbutz, namely, a study of socialization and personality development (Spiro, 1965). Like Hostetler on the Hutterites, Spiro was intrigued by this communal enterprise's goal of recreating man in society's image, making him truly fit to live in social harmony (the kibbutz ideology's hoped-for 'New Man'). And like Hostetler, Spiro approached this problem as a question of the causal relationship between culture and personality, presuming that the personality of the founders' children must take its shape from the social arrangements fashioned by the parents. In particular, seeing the goal of communalism as essentially incompatible with conflict and hostility, Spiro sought to determine the degree to which kibbutz socialization and its system of 'collective education' were successful in eliminating the 'aggression drive' in the children reared in this way. This is the research theme around which centers his psychologically expert study *Children of the Kibbutz* (the title is an expression of a kibbutz ideological shibboleth), the major product of his research on this community (Spiro, 1965). As a result of his inquiry, he finds what he already anticipated in the epilogue of his first book on the kibbutz and is consistent with his

penchant for basic psychoanalytical theory that aggression, being a universal drive, can be curbed but not eliminated.

Spiro's psychoanalytical naturalism receives even greater emphasis in his later study of women in the kibbutz (Spiro, 1979), a topic presenting itself in light of the ideological aim to establish equality between the sexes. Here he concludes, unequivocally, that gender role differentiation is due to 'precultural' factors and therefore not subject to cultural elimination. In point of fact, the direction of Spiro's explanatory exercises was signaled, even before the publication of his first volume on the kibbutz, in a provocative article (1954) the thesis of which is that the kibbutz had managed by cultural design to do away with the family in favor of the undiluted integrity of the collective whole. A few years later, however, retracting this thesis and foreshadowing his naturalistic universalism, he concluded that the family does constitute a distinct feature of kibbutz social structure (Spiro, 1960). The merits of his psychological orientation aside, Spiro's work on the kibbutz is astutely and invaluably attentive right from the start to the stream of social problems that bear on the place of the individual and run steadily beneath the ideological and even empirically minded self-presentation of this communal society.

In *Two Kinds of Rationality*, Evens (1995) treats the question of liberalizing change in a kibbutz in relation to democratic procedure and generational conflict. His phenomenological emphasis on practice and tacit knowledge sharply distinguishes his approach in the present connection. It leads him to look beneath the ideology for, rather than biological or psychological causes, contradictory understandings that are so taken for granted they predispose their hosts, whose self-identity cannot be told apart from these understandings, to certain behaviors. He traces the understandings in question to Rousseau's contractarian theory of the relationship between the individual and society, and then to the sense of man and generation given in the Hebrew myth of Genesis. Like other students of communal societies, Evens is moved to use his research as a basis on which to develop a theory of human conduct; in his case, one keyed to an ontology of ambiguity and a considered notion of ethics. In an earlier essay (Evens, 1975), he closely examines a case of stigmatization in the kibbutz, in order to bring into relief the deepest meanings of the most basic social position in this society, that of 'member' or *chaver kibbutz*. He found that the meanings in question betray implicit, unreflected ideological expectations of failure in the face of perfectibilism, and foster in respect of these expectations defensive but disturbing social acts.

Allison Bowes's (1989) *Kibbutz Goshen* not only serves as a sound ethnographic introduction to the kibbutz, but also analyzes prominent social problems characterizing the community in which she did her research. Problems of hired labor, competition over work roles, the reemergence of familism, the sexual division of labor, and conflict between the generations are all subjected by her to eclectic but useful and empirically substantive sociological analysis. More broadly, turning a remedial eye on the very idea of ideology, she contends that the definition of these apparently troubled social circumstances as problems indicative of failure derives in large part from a misguided tendency to see kibbutz ideology as a set of fixed and decisive principles rather than a pliable and adaptive system of values.

The problem of work roles is the focus of Israel Shepher's (1983) ethnography of a kibbutz (see also Shepher, 1980). He is concerned with the threat to the practice of equality posed by the development of permanent work roles, given increasing specialization in the kibbutz economy and the concomitant demand for expert workers. He argues that insofar as this trend is comprehensively applicable to every member, the creation of a structure of inequality based on work roles is precluded, since everyone is equally indispensable to his or her place in the labor sphere. In *Women in the Kibbutz*, Tiger and Joseph Shepher (1975) take up the question of sex role differentiation. Their aim is to use the example of the kibbutz to arrive at universal truth in respect of the place of women in society. Like Spiro, they propose a naturalism. But their argument is sociobiological rather than psychoanalytical, keyed to the concept of 'biogrammar' or the genetic transmission of behavioral characteristics.

In his *Transforming Kibbutz Research*, Reuven Shapira (2008) sets out to explain the relative decline, in recent decades, of the kibbutz movement and its democratic ideal. Taking his cue especially from Robert Michels' 'iron law of oligarchy,' in which democracy is undercut by bureaucracy and in turn the development of power elites and self-interest, Shapira locates the principal source of the decline not in the communities themselves but in their federative order. He finds that on the level of the kibbutz federations, leadership has always been autocratic rather than democratic, and that this condition has had the trickle-down effect of subverting governance, and therewith the idyllic democracy, in the individual communities themselves. Thus, for example, the kibbutz principle of rotation, whereby all positions of authority are supposed to circulate with general regularity among the members, is in practice corrupted by the influence of privileged relationships between certain members and federation leaders.

As a defense against oligarchic influence, Shapira, drawing on sociological literature pertaining to democratic work organizations, proposes an alternative system of succession to positions of leadership in the kibbutz. The principles of his system are designed to impede the development of bureaucracy, excessive conservative thinking, self-interest, and the like, by promoting trust, moral integrity, sincere involvement, and concern for the social whole. A key mechanism to implement such a system is, Shapira suggests, to decelerate succession to leadership in the hope of electing and retaining proven, good leaders. One way to do this would be, he argues, to allow reelection of creative, highly moral leaders, by requiring a higher democratic majority with each successive reelection.

While Shapira's proposals in this regard are worth entertaining, the most important contribution of his book is the argument about the perverse impact, on the kibbutz movement, of the oligarchic nature of leadership at the level of the kibbutz federations. As he emphasizes, this argument is innovative in kibbutz studies. Perhaps it could only have been realized by a qualified social scientist who is also a homegrown *kibbutznik*, a committed 'insider.'

Japanese Communes

David Plath's work on utopian communities is of special interest because it pertains to a non-Western social setting.

In a brief discussion, representative of the usual analytical focus on the struggle of such communities to flourish in the midst of contrary social and cultural conditions, Plath (1966) examines the tactics used by four different Japanese communal societies to adapt to the ambient milieu without compromising their reformatory goals. But in his perceptive epilogue to *Sensei and His People* (Sugihara and Plath, 1969), a first-person narrative of a Japanese commune authored by the wife of the commune's charismatic founder and leader, Plath identifies suggestive differences between this and Western communal ventures. He observes that, having evolved spontaneously, Shinkyo (the name of the commune) cannot be regarded as an 'intentional community,' and he brings into relief the fundamental way in which this community is continuous with, and not simply reactionary to, the ambient society. The commune got its beginnings in reaction to the formalism of Tenrikyo, a nineteenth-century Japanese religious movement related to Buddhism (Shinkyo's founder had served at one time as a Tenri priest), as well as to the 'bossism' of the local village community. But its defining emphases – moral responsibility, self-sacrifice, mutual reliance, social equality, and spiritual/bodily cleanliness – are manifestations of the traditional values of rural Japanese society. In view of the overtly complementary relationship between Shinkyo and the surrounding community, Plath's exercise suggests that this commune, though certainly radical by the actual standards of the state of things against which it reacted, practices a non-absolutist communalism. As a result, by contrast to Western communitarian ventures, the individual's identification with the social whole is not so much secured institutionally as left – not simply in the end but for all practical purposes – to the individual member's quotidian discretion.

The Russian Kolkhoz and Mao's Communes

Caroline Humphrey's (1983) social anthropological study of the Soviet collective farm or *kolkhoz* may be mentioned here but not really considered. For one thing, as a direct function of the state, the *kolkhoz* is neither alternative in the relevant sense nor self-generating. For another, although one can learn much about the social and politicoeconomic life of Soviet collectives from Humphrey's study, it is more directly concerned with problems of social change deriving from the Soviet Union's need to integrate into itself ethnic minorities (in this case, the Mongolian Buryat) than with collectivism itself. As a result, it is difficult to place Humphrey's book directly in the literature on communes, and apart from her extensive use of materials published by ethnographers and analysts native to the society in which she worked, she herself neither draws on nor mentions that literature.

In connection with state imposed and controlled communal life, it is edifying to cite the Great Leap Forward in twentieth-century China. There are now a number of works that document the dire results of Mao's authoritarian collectivization of rural China. Because it is based on two decades of face-to-face interviews with surviving residents of a village in Henan province, I cite here only Ralph Thaxton's (2008) work on the catastrophic famine brought about by Mao's utopian endeavor to accelerate China's communistic development. The resulting communes were large, each comprising as many as 5000 persons, and were characterized by forcible commitment to collectivist practice that was nothing less than comprehensive. The result of such compulsory, holistic, ideological rigidity was a monstrous unviability, by virtue of which tens of millions of people starved to death, while others were tortured and murdered by the powers that be, and still others were moved by their dreadful circumstances to cannibalize the dead.

Edited Collections and a General Sociological Study

Susan Love Brown's (2002) edited collection focuses thematically on Victor Turner's concepts of 'communitas' and 'liminality' as theoretical foundations for the study of so-called intentional communities. The collection comprises eight pieces (Brown's introduction included), by seven different contributors. The contributions vary, from an ethnographic study of an existing commune to studies of historical ones, all of which are American. The topics are mixed. To cite just two examples, there is a comparative study of identity formation in refugee camps and communes, as well as a study of a historical incident of mob violence against a Shaker community. Basically the various pieces are held together by the interest in Turner's conceptual apparatus, although one of the essays draws on Gregory Bateson's concept of schismogenesis, while another makes use of Anthony Wallace's notion of revitalization.

A more comprehensive edited collection on communal life is Kanter (1973). The volume comprises 45 articles, by sociologists and anthropologists. The articles bear on collectivist endeavors from all over the globe, and cover a great variety of problems characterizing these endeavors, such as organization, leadership, family, work, children, crises, and so on. So extensive, this collection constitutes, as Kanter suggests, an invaluable source for bringing into relief the diversity of communal enterprises as well as commonalities that serve to define these enterprises as communes.

It is provocative that in her Overview to the volume, Kanter chooses to cite Plato, in his *Republic*, as a benchmark for the utopianism she finds in communal societies. This is because, while the utopianism of the *Republic* is not in doubt, the work itself, as Kanter mentions, is remarkably authoritarian and elitist – anything but egalitarian. Plato grasped 'the people' as incapable of making good decisions and therefore in need of a ruling elite of philosopher kings to run things by deceiving the plebs. One might wonder if this picture does not conjure up an image of at least some communalists as dupes of their own ideological commitments and charismatic leaders.

For a substantial sociological study of American communes in general, Zablocki's (1980) book is impressive. Based on 10 years of sociological research, its analysis takes into account 120 different contemporary collectives, including both rural and urban ones. Its broad goal, though, is not simply to illuminate collective life but to throw light on the problem of decision making in any social order. Zablocki proposes that if reaching consensus is problematic even in a small society comprising ideologically likeminded individuals, then we can learn from the study of communes much about the difficulties of decision making in social orders in general. In effect, soberly critical of the research ethic that confounds the social relevance of a phenomenon with its sociological relevance, Zablocki focuses on the analytical

promise of sociological study of communes for societal life at large. Conceptually, he makes use of especially two notions, 'alienation' and 'charisma,' to plumb the many communes he visited. Redefining alienation as a sense of powerlessness in respect of societal decisions, and charisma as a phenomenon that serves to redress this felt impotence, he sees these two notions as reciprocal. As in Brown's and Kanter's studies, the basic issue is that of moving the individual, by means of intentional and structural conditions, to 'invest' (Zablocki's purposefully economistic usage) his or her self in the collective self. Bearing in mind here that, as Zablocki himself mentions, this study (unlike his book on the Bruderhof, discussed above) is not ethnographical, one key sociological goal of the work is to investigate the communal conditions under which a charismatic leader might arise.

Communes and the Significance of Anthropological Research

Communalism and Collective Social Movements

In light of the consideration that, given their relatively small size and considered social dynamics, communes would seem ideal sites for anthropological field research, it remains noteworthy that there has never been a major anthropological interest in them. Indeed, it might be said that today, as far as anthropology is concerned, communes have gone the way of the dinosaurs. But this is not to say that the discipline has no interest in so-called alternative social endeavors. To the contrary, driven by social activism bent on counteracting contemporary modern society, which is to say society that, whatever its formal presentation of self, effectively operates on the basis of domination and dehumanizing instrumental rationality, today's discipline of anthropology is marked by a somewhat general focus on 'social movements' (see Escobar (1992) for a formal manifesto of this disciplinary focus). To be sure, the rise of the Israeli kibbutz constituted an exemplary social movement, as did, according to Zablocki (1980), American communitarianism from the 1960s through the 1970s. But the current anthropological interest enjoys a compass that is at once much wider and for that very reason, in a significant sense, narrower than communalism per se.

Whether initiated by and for women, peasants, subalterns, gays, and other categories of persons who are made to suffer systematic injustice for reasons of their self-identity, the kind of movements in question have in today's world a global reach and relevance. By contrast to communes, which typically aim to seclude themselves from society at large, the social movements in question constitute collective endeavors to resist, dislodge, and creatively supplant domination, capitalism, and the State as we know it. When one considers, though, that nazism, fascism, totalitarianism, capitalism, religious fundamentalism (of one kind or another), and the like are also of great intrinsic interest from the perspective of doing social science, this current anthropological focus on social movements appears to be restricted to just one kind of movement, namely, the kind dedicated to the overall reconstruction of social orders as redemptively humane and egalitarian.

This consideration raises at least two critical questions bearing on the nature of anthropology as a social science. For one: *Is there any difference between doing good and doing good research?* In response to this question, it is impossible to avoid addressing the distinction between objectivism and subjectivism. And even should one hold, non-dualistically, that taking an objective stance is a subjective determination, the question retains its relevance. Thus, if it is the case that the activism of the current disciplinary interest in social movements entails the rethinking of anthropology as in itself a social movement, the primary impetus of which is to facilitate the goals and values of the movements under study, the question at issue here (one with which Weber notably wrestled) cannot but loom large. Indeed, how one answers this question has a serious consequence for the efficacy of social science research as research, even in respect of the desire to do good by aiding and abetting the movements at issue and their goals. Consider, for example, that under such a picture the findings and conclusions of social science risk repudiation as mere partisan politics. Hence, the question at issue, although at one time more or less trouble-free in light of anthropology's (erstwhile) taken-for-granted scientism, is now critically thorny.

The second consequential question I wish to pose is this: *In the sense relevant to anthropology as a social science, is the alterity that these movements are deemed to present a valid presumption?* The alternance of the movements must depend on how truly creative or original they are in their efforts to reconstruct society in protest against an inhumane social system, and on behalf of one imagined in terms of some truly liberating form of egalitarianism. For all the ingenuity this alterity exhibits, though, given the ultimately Marxian nature of the goals of these movements, it is hard to get away from the misgiving that the alterity remains a function of Western thought and understanding. In which case, as is true also of so many of the studied communitarian endeavors, the difference represented by the kind of movements at point – even should this difference register, for the time being, an actual, relative inversion of the sociopolitical status quo – continues to presuppose a fundamentally Western ontology. As a result, whatever the relative degree of their success, and however ethically estimable they are, one might still wonder if in reality these movements can transcend the ontological foundation of the social systems in which they arose, in particular the ideal of the self-contained individual. In this regard, Georg Lukács's (2010) skepticism may be excessive, but human experience and history suggest that it is nonetheless worth bearing in mind. Citing a romantic desire that arose among German thinkers at the very end of the eighteenth century, a desire to create "a new, harmonious, all-embracing culture," Lukács speaks of a "reckless naiveté," "a dance on a glowing volcano," and "a radiantly improbable dream."

Regardless of the gravitational pull of received Western ontology, anthropology, with its focused interest on so-called other cultures, demonstrates that there is a sense of alterity that supposes a 'reality' credibly different or other. It is a reality in which there is no need of *intentional* mechanisms to move the individual to invest his or her self into the social whole. This is because, innocently and in the first place, it has never occurred to the 'individual' peopling this kind of social whole that he or she is indeed primarily self-contained and self-transparent. Rather, experiencing themselves, ingenuously and predominantly, in terms of their intrinsic otherness to their selves, their

sense of individuality, although real, remains phenomenologically subject to, and basically truncated by, the sense of their overridingly social constitution.

It seems unlikely that modern socioculture could effect so profound an ontological and phenomenological transformation as to realize such a radically different sense of the individual, a sense in which the self is naively conceived of as primarily other to itself, and individuality is primarily determined by reference to, rather than rationality and moral autonomy, idiosyncratic differences between one individual and another. Given that societies characterized by a non-Cartesian and non-Kantian sense of the individual also exhibit inhuman practices of their own, one might wonder whether such a transformation is even desirable. It might seem, then, that in modern society, with its characteristic feature of intensive self-consciousness, significant change is a matter of raising that consciousness to entertain, sincerely, values that serve to drive home the significantly dehumanizing nature of prevailing practices. How to do this is of course the problem; and when one considers that in point of fact such salutary values are already widely broadcast in modern society, with the incongruous result of reinforcing the very inhumanity these values oppose, the enormity of the problem is brought into relief.

Possibly the desirable phenomenological and ontological change hinges on acts that by their very doing alter the social order and the consciousness of its inhabitants at the very same time. It so happens that a biblical insight provides a prototypical example that luminously captures this remarkable human dynamic. In the myth of Genesis the first man and woman, by 'choosing' (if only in effect, since their self-consciousness at the time of doing so was largely embryonic) to ignore a threateningly lethal commandment (by their maker), virtually create their very own *world* while, concomitantly, realizing their very own *selves*. In other words, through, what is best described as a childlike (the standard exegesis in terms of sin, being based on an anachronism, amounts to tendentious hyperbole) but death defying act, the couple raise their consciousness, becoming acutely aware of their selfhood, which is to say, their moral sovereignty to create their own ends and values. As a consequence of this development, the myth goes on to remind the first couple, emphatically, not to take too seriously their newfound selfhood, for if only because there is no denying death's dominion, the sovereignty won is ultimately illusory.

It would appear that both communalism and collective movements for social justice have the utopist potential to revise and elevate consciousness just so. In the present relevance, these social enterprises may well promote liberation from the naïve way in which, in today's modern world, the rule of self-interest is rationalized as natural and ultimate. In doing so, the enterprises nurture the contemplative understanding of the self as vitally a manifestation of otherness and others. Such a radical turn of consciousness implicitly portrays the Genesis myth as a registry (albeit religious) of an essential human experience, namely, the emergence of intentional reflexivity and moral consciousness from a state of naiveté. In this connection, the particular conversion projected by the kind of collective initiatives in question here entail a reversal of intentionality: the truth of the primacy of relational holism becomes, contrariwise to the biblical myth, a product of conscious enlightenment, whereas the acute self-awareness recorded, in the myth, as an ascension of self-consciousness gets reconfigured as a naïve prejudice, powerful but intrinsically deceptive and, in its rabid individualism, ironically dehumanizing. In effect, the self-awareness gained in the biblical tale becomes, with the turnaround of intentionality, hyper-reflexive, grasping itself as – paradoxically, and in view of the absolutely vital connection of the self to others and otherness – fundamentally limited and dependent.

Standing together with the perception of humankind as inherently social, this ethically and politically felicitous elevation of consciousness cannot be gainsaid. But the selfsame epistemic development – the awareness of self-awareness as given to self-aggrandizement and perfectibilism – recalls the underlying *anthropological* theme of the biblical story: the fact of, in its basic secularity, man's finitude. Hence, for all the relative success of such communal and collective social enterprises, should we not acknowledge and expect, from the perspective of social science, consequential imperfection? That is to say, aren't these enterprises best conceived as promissory events that at no time can conclusively fulfill their promise but only continue to promote it through deliberation, creative practice, and interminable struggle with the inevitable backsliding and impediments?

Given the intensive nature of ethnographic fieldwork, it is not surprising that the anthropological study of communes tends to be distinguished by sharp attention to the discrepancies between communal practice and ideological design. In staying acutely mindful of this question, the anthropologist is dwelling on a theme at least implicit in the very operation of such communities. The ideological presumption of dualistic opposition between the individual and society defines the communal endeavor as dependent on self-sacrifice for its success, thus projecting the endeavor as a trial of transcendence in the face of material adversity. The adversity is typically apprehended as a matter of the pressures exerted by the external social environment. The adversity is also associated, by communards and researchers alike, with a naturalistic presumption of individualism and self-interest. The research interests of socialization, adaptability, internal conflict, women and work roles, familism, crisis management, commitment, etc. revolve round the ideological picture of these communities as social experiments under hostile conditions, and amount implicitly to attempts at measuring the success or failure of the communitarian enterprise.

As informative as the anthropological focus on the gap between theory and practice is, though, it falls short of the unique advantage anthropology has to offer, for it tends to confine the anthropologist to the communal endeavor's ideological social compass. That compass misleads, in part but significantly, by the presentation of itself as alternative or 'other': it is after all an integral function of the social universe against which its alternance is defined. In principle, however, even if working 'at home,' the anthropologist takes his or her direction from a more radical otherness, one that cannot be reduced to what is already known, at least not without disrupting or altering the latter. Thus, Plath's work in Japan may well disclose a communalism that is not dualist in the received (Western) sense of the term, and Evens's on the kibbutz, digging deconstructively beneath the ideology to

a nondualist mythic core, begins to rethink the very idea of the difference between theory and practice, away from its implicit perfectibilism and toward a relativism that serves to reform 'ethics' as the very name of human existence.

Given the radical way in which they tend to predicate a dualism of the individual and society, communal orders constitute exceptionally fertile ground for anthropological study. Because they thus magnify a fact that ordinarily gets taken for granted, their study helps explain, as Durkheim asserted of exceptional ('pathological') social phenomena, the norm. By the same token (the conspicuous disclosure of individual-society dualism), since the relation between the individual and society is also a predicative question for social science inquiry, the study of these societies critically orients self-examination into anthropology's own tacit foundations. In their connection, it is important to bear in mind that the dualism of the individual and society presupposes – as in, for instance, the work of two imposing figures of Western thought, Descartes, and Kant – the Western notion of the autonomous individual. But the annals of traditional anthropology are filled with studies of peoples who do not entertain so radical a conception of individuality. As many ethnographies attest, this gives us no reason to conclude that such peoples already reside in the paradisial gardens imagined by communards. Provocatively, though, it strongly suggests that underlying both the establishment of communes and the problems that beset them – especially the problem of the individual's 'investment' in the social whole – are certain axial ontological preconceptions.

See also: Attempted Utopias and Intentional Communities; Collectivism and Individualism: Cultural and Psychological Concerns; Communalism; Community Sociology; Community Studies: Anthropological; Community and Society: History of the Concepts; Movements, Social.

Bibliography

Bennett, J.W., 1967. Hutterian Brethren. Stanford U P, Stanford, CA.
Bowes, A.M., 1989. Kibbutz Goshen. Waveland Press, Prospect Heights, IL.
Brown, S.L. (Ed.), 2002. Intentional Community. State University of New York Press, Albany.
Escobar, A., 1992. Culture, practice, and politics: anthropology and the study of social movements. Critique of Anthropology 12, 395–432.
Evens, T.M.S., 1975. Stigma, ostracism and expulsion in an Israeli kibbutz. In: Moore, S.F., Meyerhoff, B.G. (Eds.), Symbol and Politics in Communal Ideology. Cornell U P, Ithaca, New York (Revised 1980 as Stigma and morality in a kibbutz. In: Marx, E. (Ed.), A Composite Portrait of Israel. Academic Press, London.).
Evens, T.M.S., 1995. Two Kinds of Rationality: Kibbutz Democracy and Conflict, Generational Conflict. University of Minnesota Press, Minneapolis.
Hostetler, J.A., 1974a. Hutterite Society. Johns Hopkins U P, Baltimore.
Hostetler, J.A., 1974b. Communitarian Societies. Holt, Rinehart and Winston, Inc., New York.
Hostetler, J.A., Huntington, G.E., 1996. The Hutterites in North America. Harcourt Brace, Fort Worth, Texas.
Humphrey, C., 1983. Karl Marx Collective: Economy, Society and Religion in a Siberian Collective Farm. Cambridge U P, Cambridge.
Janzen, R., Stanton, M., 2010. The Hutterites in North America. Johns Hopkins U P, Baltimore.
Kanter, R.M., 1972. Commitment and Community: Communes and Utopias in Sociological Perspective. Harvard U P, Cambridge.
Kanter, R.M. (Ed.), 1973. Communes: Creating and Managing the Collective Life. Harper & Row, Publishers, New York.
Lukács, G., 2010. On the romantic philosophy of life. In: Sanders, J.T. (Ed.), Soul and Form. Columbia U P, New York.
Plath, D., 1966. The fate of utopia: adaptive tactics in four Japanese groups. American Anthropologist 68 (5), 1152–1162.
Shapira, R., 2008. Transforming Kibbutz Research: Trust and Moral Leadership in the Rise and Decline of Democratic Cultures. New World Publishing, Cleveland, OH.
Shepher, I., 1980. Social boundaries of the kibbutz. In: Marx, E. (Ed.), A Composite Portrait of Israel. Academic Press, London.
Shepher, I., 1983. The Kibbutz: An Anthropological Study. Norwood Editions, Norwood, PA.
Spiro, M.E., 1954. Is the family universal? American Anthropologist 56, 839–846.
Spiro, M.E., 1960. Addendum, 1958 (to 'Is the family universal?'). In: Bell, N.W., Vogel, E.F. (Eds.), A Modern Introduction to the Family. The Free Press, Glencoe, IL.
Spiro, M.E., 1965. Children of the Kibbutz. Schocken Books, New York.
Spiro, M.E., 1970. Kibbutz: Venture in Utopia, Augmented Edition. Schocken Books, New York.
Spiro, M.E., 1979. Gender and Culture: Kibbutz Women Revisited. Duke U P, Durham, North Carolina.
Sugihara, Y., Plath, D., 1969. Sensei and His People: The Building of a Japanese Commune. University of California Press, Berkeley.
Thaxton Jr., Ralph A., 2008. Catastrophe and Contention in Rural China: Mao's Great Leap Forward Famine and the Origins of Righteous Resistance in Da Fo Village. Cambridge U P, New York.
Tiger, L., Shepher, J., 1975. Women in the Kibbutz. Harcourt Brace Jovanovich, New York.
Zablocki, B., 1971. The Joyful Community. Penguin Books, Baltimore.
Zablocki, B., 1980. Alienation and Charisma: A Study of Contemporary American Communes. The Free Press, New York.

Communication and Transportation, History of

Franz Baltzarek, University of Vienna, Vienna, Austria

Published by Elsevier Ltd.

Abstract

The history of communication refers to a broad and loosely integrated field of historical research, which includes parts of the history of transport, traffic, communication, media, and information techniques. It is a relatively young field and a new point of view within historical science. Some aspects have been partly studied by historians – economic historians in particular – for a long time, especially the role of new technologies in transportation and traffic and the history of the railways, ships, and automobiles, and telephone and telegraph. The history of journalism and newspapers, as well as the history of photography and films, had been studied for a long time. Since the 1970s new questions and combinations have arisen in respect of the revolution in communication (including computation, multimedia, and the Internet) and these have given an impetus to broader research.

The history of communication is a broad and loosely integrated field of research that includes elements of the history of transport, traffic, communication, media, and information techniques. It is a relatively new field and has developed an innovative perspective in historical scholarship. Some aspects of communication history have been partly studied by historians – economic historians in particular – for a long time, especially the role of new technologies in transportation and traffic; the history of the railways, ships, and automobiles; and the development of the telephone and telegraph. But there are also numerous more recent developments, some of which will be dealt with here.

Overlapping Concepts

The term 'communication' was first used in England within the sectors of transport, commerce, and trade immediately before the Industrial Revolution. It was the time of the extension of the road system with turnpike roads, of new attempts at canal building, and of the emendation of the British postal system. The improvement of roads; the regulations governing the use of rivers; the construction of canals; and the improvement of coaches and cars, boats, and river- and ocean-going ships, were aimed at more transport capacity, the reduction of high costs of transportation, and the widening of local and regional markets to national and international parameters. The Industrial Revolution from the 1820s onwards opened up transportation with the construction of railroads and steamships; and communication was improved by the telegraph and telephone.

Today these forms of facilitating the services of commerce, trade, and tourism – including the later inventions of automobiles, progress in the building of roads, and the development of air transport – are described under infrastructure and its technical equipment. There is therefore an overlap between infrastructure and communication and the last term has become more and more subsumed under the name of 'media.' Communication nowadays is only partly concerned with matters of transport and traffic and there is danger of a one-sided contraction of the content of this term, because the media and media devices are only partly subsumed under infrastructure.

Economists, especially those in public finance, normally define communication and transport as a necessary prerequisite for development and growth, calling it 'infrastructure,' (Hirschman, 1958; Tinbergen, 1962; Jochimsen, 1962, 1966). Before the production of goods for a broader market, a well-functioning transport and communication system is essential. The term 'infrastructure' is a word that originally came from French railway idiom and was adopted by NATO after World War II. Later economists used it as a new term. Good 'material infrastructure' includes the communication possibilities of roads, canals, railways, ships, and, of course, all media equipment.

But 'infrastructure' includes not only 'material infrastructure,' it also refers to 'immaterial infrastructure,' or learning, education, schools at different levels, universities, research institutes, etc. The role of communication in the diffusion of knowledge and information is evident. According to Kuznets (1959, 1965, 1966) 'infrastructure' was one of the most important prerequisites for the European Industrial Revolution. It guaranteed stable economic growth, the international and regional division of labor, made an impact on specialization, and ensured better economic, social, and cultural integration.

Today 'communication' is used especially in connection with the media. The rapid development of the media since 1970 culminating in the Internet, necessitated a new classification of all items of media techniques and their former use: from the human voice and messengers to ballad singers or actors in a theater; from writing, handwritten codices, and letters to printing in books, pamphlets, newspapers, and posters or typewriting; from pictures, drawings, and paintings to photos and since the late 1990s scanning possibilities; from messages given through drums, written on stones, papyrus, parchment, and paper to transmission by telephone, telegraph, film, radio and TV, CD-ROM, DVD, the computer, and the Internet. As a result, the discipline of journalism has widened its scope to include all the mass media and even studies their history. More and more sociological and psychological aspects were introduced in defining the different stages of mankind's communication modes. In this context, M.T. Poe (2010) speaks of five stages of development of mankind from a 'homo

loquens' to a 'homo scriptor,' a 'homo lector,' 'homo videns,' and then to a 'homo somnian' (humanity in the age of the Internet).

There is also an overlap with the term 'information,' which is especially stressed in economics and business science as an important outcome of communication and defined as the ability to share information. Since the 1960s 'information' has been seen as the 'new immaterial resource' in physiology, sociology, psychology, pedagogy, linguistics and information science, ethnology, political sciences and law, and, last but not least, in a bundle of technical sciences.

In economics, for example, it was F.A. Hayek (1968) who described markets as spontaneous networks of information about goods and services. In a competitive market, the profits make it possible to have and to use more information than normal and provide information easier, faster, and cheaper. Market prices in normal (competitive) situations provide very precise information about shortages and abundances, qualities, monopolies, and so on. Business scientists study the organization of firms and the optimization of information within business entities and they stress 'information' as an important factor of production.

As Pool (1963) points out, communication is a fundamental necessity for human social action. Therefore we can distinguish between: (1) interpersonal communication = information exchange within an individual person; (2) intrapersonal communication = information exchange between two or more persons; and (3) media-bound communication = exchange of information between media producers and media consumers (in mass consumption). Since the 1980s communication has come to mean not only exchange of information between human beings but also between computers.

The development of communication and transportation possibilities shows, in the very long run, a dramatic widening of the abilities and capacities of human communication. From the stages of primarily personal experience of communication by ears and eyes, it shifted to a secondary system which is technically arranged and represents 'artificial' communication (S. Krämer, 1997). Today this 'artificial' communication has reached a peak in the new media possibilities of TV, the PC, and the Internet. McLuhan (1964) points out that individuals recognize their environment primarily through the media, which increasingly use the new techniques of electronically transmitting and storing information. His diagnosis is that 'the medium is the message.'

According to N. Luhmann (1981) social systems can be defined as the possibilities of human beings (in their time–space related contexts) to take part in the experiences, procedures, and activities of society. Their ability to participate is, however, dependent on the availability of communication and transport equipment and techniques. Religious and political barriers or facilitations; problems of economic and social inequalities; and the concept of 'property rights' play a role.

The historical sciences, i.e., political, social, and economic history, as well as the history of technology, study the causes and consequences of the communication process from antiquity to modern times. As communication and transport are important prerequisites of development and globalization, one has to examine very thoroughly the rise of the West to see how it obtained leadership over the East. To a large extent, this leadership originates from better and new infrastructure and the organization of transportation and the media. Every step of European development since the Middle Ages can be seen to have happened in conjunction with big spurts and crucial points in the improvement of each new media and infrastructural equipment.

The European case also seems to show that with the improvement of media and transportation, a growing displacement of the state and the old elite went hand in hand with the increasing participation of human beings in communication. N. Elias (1939) in his famous book on civilization has pointed out the strict limitations of monopoly power in western societies. This freed the way for more and more social groups and individuals to use communication and transportation techniques – from the *cursus publicus* of the Roman Empire (which was only open for the purposes of the State) to the new postal system at the beginning of Early Modern Times; from the grand tours of the aristocracy and the commercial journeys of single merchants to mass tourism; from illiteracy to dense school systems; from the handwritten codices of the monks to Gutenberg's first printed Bible and Jerry Cotton; from the transport of gold, silver, silk, and amber to the transport of bulk goods over the whole globe; from the limited diffusion of knowledge and messages by fire signals, letters, etc. to the world wide web in the 'global village' with a 'global information infrastructure.'

Phases of Development in Europe

Historians and media specialists think that within western culture one can find four revolutions and periods in communication and transport development: (1) the Middle Ages (especially from eighth to fourteenth century AD); (2) the Early Modern Times (fifteenth–seventeenth century AD); (3) the onset of and then the Industrial Revolution (eighteenth–first half of the nineteenth century AD); and (4) the Communication Revolution, beginning in the last decades of the twentieth century.

M. Weber (1896) in his famous article on the decay of the Roman world sees the communication and transportation deficit in the Roman Empire as the single, main cause of its decline. He argues that all ancient high civilizations from China, Mesopotamia, Egypt, and the Greek world were river- and shore/coastal-oriented civilizations. The failure of the Roman Empire was that it went deeper and deeper into the European continent. Though the quality of Roman road building was very good and a relay system of changing tired for fresh horses enabled people to cover distances as far as 300 km a day, the ancient *cursus publicus* had not enough impact on common trade, commerce, and communication, because it was not open to the common man.

The Middle Ages and the Early Modern Times in Europe found new answers and, in the long run, more stable solutions in communication, and thus prepared the way for the great spurt of European civilization which, from the sixteenth century AD onwards, left behind the older and richer civilizations of Asia. In The Middle Ages (from the tenth and eleventh century AD onwards) the feudalistic system weakened the state and centralized authority, but strengthened local and regional authorities. At this time, dense networks of more or less

autonomous cities were created, in which trade and handicraft were defined as a primary privilege of citizens and new, dynamic dimensions of the concept of property rights' were introduced to the feudal manors. New communication and information techniques and systems were tested. We can see this in regional and interregional trade with new types of merchants and new commercial institutions (such as fairs and marketplaces), trade guilds, commercial law, special organizational forms of storing, borrowing, and changing money – all these were the outcome of perpetual communication. The consequence was a very rapid diffusion of technical knowledge in mining, architecture, handicraft, and agriculture.

This new system was open for inventions from the Orient, the Arabs, and the Islamic world and accepted new insights from China and India. Technical inventions which revolutionized communication and transportation were the horse collar (tenth century, used in China since 450 AD) and the stirrup. From 1200 AD onwards the diffusion of the use of Arab numbers began and made trade much easier. The compass and the compass needle (used in China from 1075 AD onwards), adopted from about 1200 AD, improved navigation. Though the quality of the techniques of Roman road building could not be attained, the opening up of roads to everybody was a big improvement on the *cursus publicus* in the ancient Roman world. In about 1380 AD cities, religious orders, and guilds introduced the new system of sending messages regularly and over long distances through private postal systems.

Predominantly face-to-face communication, using primarily oral means of information, shifted more and more to a written media stage. The monasteries and the first universities with already-formed libraries had a great impact on learning and knowledge. From the middle of the fourteenth century AD onwards, the improvement of shipping on the rivers through the construction of canals and canal locks began. Measuring time took on new dimensions with the invention of more precise clocks with wheels at about 1330 AD. From the mid-fourteenth century AD there were clocks on the towers of churches. In 1510 AD, the first pocket watches appeared.

The Early Modern Times brought about the second revolution of communication in Europe. The adoption of Chinese navigation techniques was completed; the compass freed the way for discoveries and the exploration of the New World. Trade in mass goods began. This became easier when plantations began to produce 'fungible goods' (like tea, coffee, cotton, and tobacco), which enabled a new trade on commodity markets.

The production and the use of paper and the idea of the movable type for printing (Johannes Gutenberg, 1397–1468 – also of Chinese origin; the first printed book appeared in China in the sixth century AD) brought about the shift from a 'written' to a 'printed' culture. Books now could be printed at a relatively low price and in large numbers. The woodcut print (first used in 1400 AD) and the copperplate print (first used in 1446 AD) now could be combined with the new printing technique. Pamphlets were printed and could be easily distributed. Political and religious controversies, teaching, and learning took on new dimensions. Newspapers were introduced. The monopoly on science and information was broken. The book became important for the unlimited storage and diffusion of information. The pamphlet was used for political discourse and opened up possibilities for participation in political and religious debates; and the newspapers opened the way to inform a growing number of readers about recent events in society.

The sovereigns of territories reorganized their postal system and widened communication possibilities – for example, in 1486 AD the 'Kurfürstliche Botenanstalt' was introduced in Brandenburg. In 1490 AD the postal system in the German Reich was reorganized under Maximilian I. In 1516 AD the Vienna–Brussels postline was introduced, while the Brussels–Rome postline was introduced in 1522 AD by Thurn and Taxis of Vienna-Nurenberg. The new communication advantage cleared the way for an European expansion over the Atlantic and created a second West in the New World of America.

The third revolution began before the Industrial Revolution and influenced the period up to the present day in three stages. In the first stage, the mercantile system introduced more regular traffic, a regular postal system, and supported canal- and road-building. Milestones in France included the 'Canal du Center' between the Loire and the Saône (1555 AD) and the 'Canal du Midi' between Southern France and the Atlantic (1666–81 AD). Soon after, many canals in German territories, in the Dutch Lands, and in England were built. The introduction of railways from the 1820s and 1830s onwards, the intensive use of steamships from the 1850s and 1860s onwards, the common use of telegraph, telephone, and the overseas cable since the 1850s to 1870s, (and especially the spectacular opening of the Suez and the Panama Canals) marked the second step, which reached its peak in a highly integrated global world economy before World War I.

The automobile and new road-building – especially in the form of highways – and the victory of the street over the railway and the airplane, widened communication and transportation possibilities in the following interwar period. This brought greater reductions in the regional differences in prices, greater allocations of production and resources, more division of labor, mass production for wide national and foreign markets, standardization, and the unification of languages, morals, and tastes. The third step began in about 1970 with the 'post-industrial phase' or the 'age of information.' The multimedia revolution brought the technical inventions of ISDN, the mobile phone (cellular phone), the fax, the PC, and the Internet (www, online and e-mail), and digital forms of TV, broadcasting, and printing – the 'information superhighway'.

The economic consequences were e-commerce, telebanking, more foreign investment in many countries, and a very sharp decrease in the costs for information and transport, globalization, digital traffic engineering, and teleworking. The socio-cultural consequences were teleteaching, a networked scientific community, and telemedicine. The political consequences were e-government and smart cards for electronic administration systems.

Historiographical Trends, Further Research, and New Fields

The historiography of former times seldom studied communication and transportation. Since the turn of the nineteenth century, 'economic history' achieved an independent status and transportation problems were described under the aspect of commerce. In the period before World War I in nearly all

industrialized countries, the first technical museums were created. They had a great impact on the study of the history of technology and innovations, especially railways, cars, automobiles, ships, airplanes, and information devices like the telephone and telegraph. The history of technology became a new subdiscipline within economic history. In the last few decades, the science of journalism has widened to a science of media and examines all media, especially mass media, in the historical context. Today a modern history of communication has to combine all these fields in cooperation with other social and economic disciplines.

This broad approach is also represented by the International Communication Association (ICA). Its forerunners go back to 1950; since 2003, ICA is officially associated with the United Nations as an international nongovernmental association. ICA also has a section studying communication history.

There are, of course, many possibilities for further historical research in the field, e.g., studies on the regional, national, and worldwide diffusion of communication and transportation techniques; crowding out and/or supplementary processes between the different communication technologies; the social consequences of media revolutions; the impact of the media and the transport revolution on political situations and military conflicts; the access to communication and information techniques; religious and political control and inhibition of communication (censorship); the effects of the transport revolution within economic development; the role of official and nonofficial statistics in information about economies; and the consequences of changing communication and transportation in mass tourism, leisure, and hobbies. Other possibilities include the rise of special cultural forms, traditions, and customs concerning, for example, letter writing, traveling by coach, railroad, automobile, ship, and airplane, and studies of the intensification of communication in personally written as well as in digital communication.

Even completely new fields can be developed under the aegis of communication. One example may be the creation of European 'financial intermediation' in banking and in stock exchanges. It developed as an important new way of communication in the economic and business sphere of the seventeenth and eighteenth centuries.

How to finance firms and enterprises and how to finance states and their budgets – these are the important questions in every modern economy. In former times people who had large savings and lent to the state or to private entrepreneurs could not get back the money before the end of the agreed term. So, normally, the only possibility was to hold large stockpiles of money, without being able to lend it, in order to remain liquid at any one time. This problem was shared by wealthy individuals and institutions, such as insurance companies, managements of orphans' properties, etc. No civilization had managed these problems adequately until Europe found new ways of communication to bring together lenders and borrowers. There are four areas that highlight this information and communication revolution in banking since the seventeenth century AD.

1. Since the period between the fifteenth and seventeenth centuries AD, communication and information has improved through investment in new roads, cars, ships, canals, and new media techniques. The unification and standardization of goods also brought improvement, especially through the creation of 'fungible' goods, wherein one item stands as guarantee of quality for all the other pieces of this product. Markets could thus be transformed partly into markets with 'fungible' goods. The exchange in products for certain metals and for certain crops, especially tropical plantation crops like cotton, tea, coffee, sugar, etc., was born. This facilitated economic communication enormously. As the qualities of these goods were defined *ex ante* and were well known, merchants could deal in great volumes of these goods and articles without actually bringing them physically to the markets. Transaction costs were thus dramatically reduced.
2. In combination with stock exchanges the experiences with 'fungible' goods were transferred in the seventeenth century to the financial sector. The creation of fungible bonds for the state and public authorities (instead of individual obligations for single lenders in former times, which could not easily be transferred to other capitalists) pulled hidden private stockpiles of wealth into the public economy and brought the money into the hands of sovereigns and governments. This idea spread to fungible shares for enterprises. Now it became possible to erect big enterprises in the form of joint stock companies. The first joint stock companies were firms within the trading sector, followed by companies in infrastructure and communication; by companies in the banking system; and last of all by big industrial firms.
3. The next point in this development was the improvement of the financial mediation of banks. A new type of bank developed, which brought together people who had money but did not need it at the time and people who did not have enough money to make productive investments in business. The 'goldsmith bankers' of London were the first to succeed at this business, because they found out that the communicative role of a bank has to be the transformation of risk, of terms, and of size. Fungibility and the new communicative role of banks became the main facilitators for a 'financial revolution,' which was a precondition of the financing of the Industrial Revolution.
4. Since the beginning of the twenty-first century new steps in financial communication have been taken, especially in the context of the PC and Internet systems. New generations of financial products have arisen, including e-banking and new techniques of worldwide speculations (high-frequency trading).

See also: Advertising and Advertisements; Agenda Setting, Media Effects on; Ethnocentrism; Film History; Narrative, Sociology of; Political Advertising; Public Sphere and the Media; Tactile Maps in Geography; Visual Images in the Media.

Bibliography

Aldcroft, D.H., Freeman, M.J. (Eds.), 1983. Transport in the Industrial Revolution. Manchester, UK.

Bagwell, P.S., 1974. The Transport Revolution from 1770. London Historical Society, London.

Beninger, J.R., 1986. The Control Revolution: Technological and Economic Origins of the Information Society. Harvard University Press, Cambridge, MA.
Birk, A., 1934. Die Straße. Ihre verkehrs- und bautechnische Entwicklung im Rahmen der Menschheitsgeschichte. Adam Kraft Verlag, Prag.
Brown, D., 1997. Cybertrends. Chaos Power and Accountability in the Information Age. Penguin Harmondsworth, Middlesex, UK.
Dienel, H.L., 2007. Verkehrsgeschichte auf neuen Wegen. Jahrbuch für Wirtschaftsgeschichte 2007/I. Akademieverlag, Berlin.
Donsbach, W. (Ed.), 2008. International Encyclopedia of Communication. John Wiley & Sons, UK.
Dyos, H.J., Aldcroft, D.H., 1971. British Transport. An Economic Survey from the 17th Century to the 20th. Leicester University Press, Leicester.
Eisenstein, E.L., 1983. The Printing Press as an Agent of Change, vol. 2. Cambridge.
Fang, I., 1997. A History of Mass Communication. Focal Press Boston, Oxford, Johannesburg, Melbourne, New Delhi, Singapore.
Goody, J., 1987. The Interface between the Written and the Oral. Cambridge University Press, Cambridge, UK.
Grübler, A., 1999. The Rise and Fall of Infrastructures. Dynamics of Evolution and Technological Change in Transport. Franz Steiner Verlag, Stuttgart.
Habermas, J., 1981. Theorie des kommunikativen Handelns, second ed., 2 vols. Suhrkamp, Frankfurt/Main.
Jahrbuch für Kommunikationsgeschichte (since 1998). Franz Steiner Verlag, Stuttgart.
Lay, M.C., 1994. Die Geschichte der Straße. Vom Trampelpfad zur Autobahn. Campus, Frankfurt/Main.
Lepetit, B., 1984. Chemins de terre et voies d'eau. Reseaux de transports et organisation de l'espace en France 1740–1840. Paris.
Mattelart, A., 1996. The Invention of Communication. University of Minnesota Press, Minneapolis, MN.
McLuhan, M., 1964. Understanding Media the Extensions of Man. McGraw Hill, New York.
McLuhan, M., Powers, B.A., 1989. The Global Village. Transformations in World Life and Media in the 21st Century. Oxford, New York.
Poe, M.T., 2010. A History of Communications, Media and Society from the Evolution of Speech to the Internet. Iowa University.
Pool, I., 1963. The role of communication in the process of modernization and technological change. In: Industrialization and Society. UNESCO, pp. 279–295.
Rösener, W. (Ed.), 2000. Kommunikation in der ländlichen Gesellschaft vom Mittelalter bis zur Moderne. Veröffentlichungen des Max Planck Instituts für Geschichte 156. Vandenhoeck & Ruprecht, Göttingen.
Sösemann, B. (Ed.), 1994. Beiträge zur Kommunikationsgeschichte, vol. 1. Stuttgart.
Teuteberg, H.-J., 1994. Entwicklung Methoden und Aufgaben der Verkehrsgeschichte. Jahrbuch für Wirtschaftsgeschichte 1994/I. Akademieverlag, Berlin.
Teuteberg, H.-J., Neutsch, C. (Eds.), 1998. Vom Flügeltelegraphen zum Internet. Geschichte der modernen Telekommunikation. Vierteljahrsschrift für Sozial- und Wirtschaftsgeschichte, Beiheft147. Franz Steiner Verlag, Stuttgart.
Treue, W., 1965. Achse, Rad und Wagen. 5000 Jahre Kultur- und Technikgeschichte. Bruckmann, München.
Wilke, J., 2011. Between fragmentation and integration: european news markets from the 19th to the 21st century. In: Putnis, P., Ch, Kaul, Wilke, J. (Eds.), International Communication and Global News Networks. Historical Perspectives. Hampton Press, New York, pp. 245–261.
Williams, F., 1982. The Communications Revolution. Sage, Beverly Hills, CA.

Communication Research and Media Studies, History of

Michael Meyen, Institute for Communication and Media Research, University of Munich, München, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Communication research is both the latecomer of the social sciences and a product of political and economic necessities. The discipline's formation took the search for effective propaganda designs and a growing interest of both media industry and students. Furthermore, the field's development is inseparably linked to the history of US universities. First, the strong emphasis on quantitative research was invented precisely there. Second, communication research made in the US was exported worldwide.

Communication research is both the latecomer of the social sciences and a product of political and economic interests. In the first half of the twentieth century, other social science disciplines like sociology, economics, anthropology, psychology, and political science were born but not communication (Abbott, 2002). The discipline's formation took World War II with its interdisciplinary work on propaganda that was clearly conditioned by Rockefeller Foundation's support, military, CIA, and State Department funding (Pooley, 2008) and the search for effective propaganda designs in the early Cold War sponsored by the very same agents (Simpson, 1994; Glander, 2000). Last but not least, it was a growing interest of both media industry and students that "communication research began to be recognized as a distinct academic field" (Craig, 2008) or, as in the case of German *Publizistikwissenschaft*, solved its legitimacy crises by a shift from a humanistic to an empirical social science pleading for the use of quantitative methods (Löblich, 2007). However, Robert Craig (2008) associated the 'rapid growth and institutional consolidation as an academic discipline' that communication underwent since then with certain problems: strain on resources, an overemphasis on practical training of undergraduates, and the threat of cooptation by commercial interests. To put it in a nutshell, in terms of scientific capital, the latecomer communication research seems to be still at the bottom of the table rather than at the top.

As the article's title already suggests, both the field's name and status are contested. All over the world, communication research and media studies are torn between social sciences and humanities and highly diverse in methods, theories, and objects of study (Craig, 2008; Averbeck, 2008). Not even in the US, the field has a singular authoritative institution like the Modern Language Association or the American Anthropological Association. In addition to purely academic associations such as the International Communication Association (ICA, more than 4000 members in over 80 countries) and the International Association for Media and Communication Research (IAMCR, 1500 members from some 100 countries), there are both associations that also embrace educators, teachers, politicians, and practitioners (National Communication Association, Association for Education in Journalism and Mass Communication, World Association for Public Opinion Research) and those that focus on very specific topics in cross-disciplinary fields (Society for Cinema and Media Studies, Association of Internet Researchers). Furthermore, even some of the field's most prominent figures such as Byron Reeves (2012) from Stanford view communication 'very much as a subject and not as a discipline like psychology or biology' to this very day. For all these reasons, the question continues to be key today whether communication research and media studies is already (or still) a discipline rather than 'one of the great crossroads where many pass but few tarry' – as Wilbur Schramm (1959) puts it against Bernard Berelson's death pronouncement in the *Public Opinion Quarterly* debate some 50 years ago when this academic enterprise was just new born (Wahl-Jorgensen, 2004). Part of the field's legitimation and identity problem are the notions of an 'academic Taiwan' (claiming all of China as its own, Peters, 1986), or, similarly unfavorable, of a stepchild that is tolerated by university administrators only because communication research and media studies helps pay the bill for the departments they really care about (Meyen, 2012).

The Field's Historiography

From the early days on, the history and present day structures of communication research and media studies have been of particular interest to those who are within this field (Dennis and Wartella, 1996); in part as a tool for 'authentication and legitimation,' easy to understand against the just outlined background, and partially because there seems to be neither significance nor identity without historiography (Hardt, 2008). The rich and growing body of literature on the field's origins has a certain focus on the intellectual roots of the discipline (Peters, 1986, 1999; Rogers, 1994; Hardt, 2001; Park and Pooley, 2008). There are 'milestones' (Lowery and DeFleur, 1983), edited collections of key or canonic texts (Katz et al., 2003; Peters and Simonson, 2004), myths told over and over again about both the four founding fathers Carl Hovland (1912–61, a psychologist), Harold D. Lasswell (1902–78, a political scientist), Paul F. Lazarsfeld (1901–76, a sociologist), and Kurt Lewin (1890–1947, a psychologist again) and the institution builder Wilbur Schramm (1907–87) who started the programs and institutes at Iowa in 1943, Illinois in 1947, and at Stanford in 1955 (Rogers, 1994; Chaffee and Rogers, 1997). On the one hand, those biographical approaches have been criticized as 'great-men-make-history' tales (Löblich and Scheu, 2011; Simonson, 2008) but on the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03054-3

other hand, there are portrays of the most promising female communication scholars based on personal conversations (Signorielli, 1996) or archival work (Dorsten, 2012) as well as current interviews with 57 leading scholars (Meyen, 2012).

Looking at the importance of propaganda back then, it seems easy to name the 'wartime research effort' as "the single most important catalyst for the new field of communication research" (Wahl-Jorgensen, 2004) and to point both to the need for legitimization and to the scholars coming in from various disciplines to study public opinion, mass media effects and persuasion (Glander, 2000; Simpson, 1994). However, there are also sophisticated trials to locate the discipline's development and institutionalization within "the context of communication as an evolving category of cultural knowledge and practice in society" (Craig and Carlone, 1998; Craig, 1999) or 'within the system of American higher education' (Delia, 1987) and academic structures (Pooley and Katz, 2008). Besides those dominant narratives, there are alternative constructions of the research field's history available (Carey, 1992; Hardt, 1992). The very same is true for research on almost all the facets of an academic hierarchy – on the relative positions of the different areas in relation to others in the field (Donsbach, 2006), on masters and doctoral programs (Craig and Carlone, 1998; Barnett et al., 2010), on 'the most important theorists' (Beniger, 1990), on citation patterns (Park and Leydesdorff, 2009), on research productivity both of individual scholars and departments (Bunz, 2005), and on leading associations (Weaver, 1977; Nordenstreng, 2008).

The Field's Geography

The academic discipline of communication research is (still) a US centered enterprise with strong pillars at the Big 10 universities and the West coast. There are no communication research and media studies programs at most Ivy League schools and other top-tier US universities. Stanford and the two Annenberg Schools for Communication at the University of Pennsylvania and at the University of Southern California (USC) in Los Angeles are three exceptions to the rule that the field's heart beats not at private US schools but at the large public research universities such as Illinois, Indiana, Iowa, Michigan State, Ohio State, Penn State, Purdue, Texas, and Wisconsin. Contrastingly, 'the golden era at Stanford' seems to be over. According to eyewitness Everett Rogers (1994), it is not so long ago ('from roughly 1955 to 1970') when that university 'dominated the field' and everybody would have wanted to hire one of Stanford's PhDs in communication. From that period is not too much left. Today, Stanford only has a rather tiny number of people in human–computer interaction and mass media studies (Reeves, 2012).

The dominant historical narrative that explains the field's geography is, at least in this case, quite similar both in the journalism- and speech-derived traditions (Rogers, 1994; Cohen, 1994; Pooley, 2008). Looking at those histories, the development of the communication research field is inseparably linked to the history of US universities. The land-grant schools were established to help develop the new territories of the US. In these universities, practical skills such as speech, rhetoric, and journalism were more respected. The focus on practical activities and the idea of instrumental tasking attracted a certain type of students: kids of social climbers and, in particular, of European immigrants. Conversely, the US elite universities repositioned themselves along the lines of science in the early 1900s (Meyen, 2012).

This narrative fits into Klaus Krippendorff's construction of communication's early days at the University of Pennsylvania in Philadelphia. Krippendorff (2012) who became an assistant professor in 1964 at Penn: "Walter Annenberg was the owner of *The Philadelphia Inquirer* and wanted to pay for a school that would create professionals that could work in his newspaper. The University of Pennsylvania didn't want to have a narrow professional focus but liked the idea of a communication school." Consequentially, the Annenberg School focused almost exclusively on research. The strong position of the Annenberg Schools at Penn and USC within the communication research field is not just a heritage of the long-term dean George Gerbner (1919–2005, in charge from 1964 to 1989) or a function of the commercial context the two are settled in. "One of our strength was financial," said Klaus Krippendorff (2012). "Walter Annenberg provided ample resources."

In the case of Stanford, it is more about reputation than just dollars. Everett Rogers (1994) first listed the eight US schools with the highest prestige (they 'tend to be private, old, and resistant to radical educational innovations') and then celebrated, logically from his point of view, "Schramm's move to Stanford University in 1955 and the founding of the doctoral program there" as "the key event in gaining acceptance of communication study in American universities." However, more than half a century later, Byron Reeves (2012), Schramm's successor's successor, still has to fight for the reputation of the discipline: "Everybody is always asking what communication is and why they don't have it at Harvard, Princeton, Yale, or most of the other top ten or top 20 universities. That's where the peers for all the other departments are." Therefore, both the geography of the field and its roots in more practical skills such as journalism and speech have implications for communication's position in the larger scientific field.

Concentrating on communication research and media studies' beginnings in the US does not in any way mean a devaluation of other world regions. Of course, there are similar developments in Europe in the first third of the twentieth century. It is well known that Paul F. Lazarsfeld who became an important point of reference for US centered historiographies of the communication research field started out in Austria where he, to give just one example, collected data on the listening habits of radio audiences in the early 1930s (Fleck, 2007). In Germany, the *Zeitungswissenschaft* (newspaper science) was established at universities during World War I in order to raise the level of journalism in the fight for public opinion (Löblich, 2007). However, it is not just the common root in practical training and administrative research that legitimizes a certain focus on the United States. First, the strong emphasis on quantitative research, including statistical methods and sophisticated data analysis, that became one of communication research's key identity elements from the late 1950s on and will be described in the next paragraph, was invented precisely there. Second, closely linked to the shift from training offerings to academic

approaches, communication research made in the US was exported worldwide. To put it another way, many of the founding mothers or fathers of national scientific communities were not just heavily influenced by the North American research tradition but rather a part of it. The exceptions include France and southern Europe where it was not until the 1970s that the discipline began to spread, Eastern Europe with its roots in Marxist–Leninist traditions of journalism training, British Cultural Studies and parts of Latin America (Averbeck-Lietz, 2013).

It is easy to see that the dominance of US-originated approaches to communication research and media studies is a function of both the power of English as the global language and, probably even more important, the level of commercialization of media systems. Media markets create the needs for application knowledge, well-trained students, and a public debate on the individual and societal consequences of commercial media products. That is why it is no accident that US universities not just exported theories and methods but also scholars that became leading figures in other national academic environments. To start with English-speaking countries, the native American Jay Blumler (United Kingdom) graduated from Antioch College. Later, in addition to his position at Leeds where he became one of the most prominent researchers in the field of political communication, Blumler was professor at the University of Maryland. Elihu Katz (Israel) was born in New York, Akiba Cohen (a second Israeli) in Detroit, Cindy Gallois (Australia) in Washington, and Janet Bavelas (Canada) in Portland, Oregon. James Taylor who inspired the famous Montreal School of organizational communication research worked as a lecturer at the Annenberg School in Philadelphia in the 1960s. All five of them got their PhDs at communication or, in Katz' case, sociology schools in the US. Youichi Ito (Japan) who is renowned for his information society concept did a master's degree at Boston University and had a fellowship at Tufts University in the 1970s. Osmo Wiio (Finland) attended an American seminar in Salzburg, Austria in the late 1950s. From then on, Wiio developed strong contacts with the US universities. The very same is true for Kaarle Nordenstreng, the second famous Finn in the international community. In the late 1960s, Nordenstreng had a fellowship at Southern Illinois University mainly working for his doctorate. He got to know communication research by traveling the US and interviewing American colleagues (Nordenstreng, 1968). Elisabeth Noelle-Neumann (Germany), to give one last example, received a scholarship at Missouri's School of Journalism in 1937–38 and was a visiting professor at the University of Chicago on a regular basis between 1978 and 1991. One of her students is Wolfgang Donsbach, who served as a World Association for Public Opinion Research (WAPOR) president in 1995–96 and as an ICA president in 2004–05 and edited Wiley's 12 volumes *International Encyclopedia of Communication* (Donsbach, 2008).

The Shift from a Humanistic and Practical to an Empirical Social Scientific Discipline

Communication research and media studies literally began with the invention of the printed press. However, in the first two and a half centuries after the world's first newspaper appeared in the City of Strasburg (1605, German language), the field remained a sphere dominated by outsiders. In the German language area, novelists, theologians, editors in chief, other journalists, or private scholars wrote about the topic. Although the first doctoral thesis on newspaper reports has already been submitted in 1690 (at the University of Leipzig), although there were a number of lectures devoted to the organization and the usage of the press in the eighteenth century, and although historians, economists, legal and literary scholars started to explore press-related problems with the unfolding of the mass press in the second half of the nineteenth century, German universities were not interested in the foundation of an independent communication discipline until World War I. At this early stage, the impetus for founding the *Zeitungswissenschaft* (newspaper science) came from publishers' and journalists' associations (Meyen and Löblich, 2006).

The German foundation history is an epitome for the field's emergence. In all countries, first, media distribution and academic communication research are closely connected. Second, media innovations had and have wide implications for researching and theory-building in the field of public communication. And third, the demand for qualified journalists, PR specialists and, at least in the US case, debaters was a major driving force of the field's academic institutionalization. In Germany, both *Zeitungswissenschaft's* forefathers and founders started out in journalism. They mainly focused on journalism education as well as on the discipline's legitimation and, therefore, could not improve their status at university. Furthermore, National Socialism stopped initial steps toward theorizing communication and media phenomena by forcing promising Jewish scholars into exile (Averbeck, 2001; Fleck, 2007). Those German researchers who could carry on in that period were compromised after 1945. The discipline's reestablishment was, again, the work of trained journalists such as Walter Hagemann (1900–64) who coined the term *Publizistikwissenschaft* and reinvented communication research and media studies from scratch (Wiedemann et al., 2012). Although Hagemann already experimented with social scientific methods, German *Publizistikwissenschaft* was still dominated by humanistic and practical approaches.

However, all over the world, well-rounded communication-related doctoral programs were not established until after World War II. Whether implemented in former journalism schools or speech departments, the new born US graduate programs emphasized quantitative empirical work. Rogers (1994) described the not just epistemological conflict between the nonscientific 'Green-Eyeshades' ("oriented to the profession of journalism rather than to the new science of communication") and the 'Chi-Squares' who survived in the end because their scholarly approach fit with the norms of research universities. In addition to the media industry that became engaged in polling and market research (Rogers, 1994; Löblich, 2007), the students experienced a kind of epiphany that matched the habitus of social climbers who focused on the idea of contributing socially. At least in the 1950s and 1960s, young journalists as well as young debaters went out into

the world to make a difference. Involved into scientific research projects, the future communication research professors became convinced that the answers derived from Hollerith machines and supported by statistical techniques would influence journalists and others (Meyen, 2012).

Michigan State University played a special role in the shift from a humanistic and practical to an empirical social scientific discipline. That shift is especially linked to the name of David Berlo (1929–96) who was one of the first individuals to receive a PhD degree in the new field of communication (at Illinois). Berlo founded the department at Michigan State and continued as chair for the first 14 years (1957–71), starting out not even 30 years old. According to Everett Rogers (2001), Berlo saw "that training large numbers of doctoral students was the route to implementing a social science conception of communication." To Gordon Whiting having a speech and rhetoric training, Berlo would say, "we are making a scientist out of you." Berlo's closest allies Gerald Miller (Iowa), Hideya Kumata, and Erwin Bettinghaus (both Illinois) had been trained at schools founded by Wilbur Schramm too. For Rogers, who came 'from an established department at Ohio State (in sociology),' at Michigan State the 'the close-knit nature of faculty relationships' was kind of new. That as well as the quantitative orientation he traced back to "the high degree of uncertainty about the new field of communication study" later. As chair, Berlo would perceive that his department was at risk and "convince his faculty to pull together against a hostile environment." In Roger's memoir, there is also a brief glance at Berlo's habitus: a bodyweight of over 270 pounds, hidden in 'well-tailored dark suits' and so pretending establishment (Rogers, 2001).

There could be no doubt that the Michigan State habitus began to rule the field very quickly. Parts of it were the strong emphasis on quantitative research, including statistical methods and sophisticated data analysis, the 'high degree of male bonding,' the notion that communication research is an academic underdog and the feeling that the resultant methodological orientation was superior to any other approaches. That habitus promised "a defence against the possible suspicion by higher levels of academic review that communication was not rigorous enough" as well as personal security for the newcomers at university and is around to this very day (Meyen, 2012).

However, the path to more recognition mapped along the lines set by the natural sciences at the power pole of the academic field has been challenged again and again, too. In addition to the already mentioned British Cultural Studies with its emphasis on power, media content, and media usage, there are other supporters of qualitative methods as well as theoreticians available. For example, Hanno Hardt (2001) called for a paradigm shift to critical communication research. In lamenting the crisis of US-based quantitative approaches, Hardt pointed to the European tradition in analysing the relationship between mass media and society and edited communication-related works of famous German sociologists and economists such as Karl Marx, Max Weber, Karl Knies, Albert Schäffle, and Ferdinand Tönnies as well as early American trials by Albion W. Small, Edward A. Ross, and William G. Sumner.

Communication Research and Media Studies on the East–West Battlefield

It is not just the field's roots in propaganda research during World War II and the postwar era up to the 1960s that stand for the considerable political influence on research problems, theories, methods, and even appointments. A glance at communication's two leading international academic associations today shows how long the road to more autonomy took. Until the mid-1970s, ICA had members from government, from business, and from teachers who did not want to do scholarship (Meyen, 2012).

Unlike ICA, IAMCR could not really extract itself from the grip of politicians or practitioners. IAMCR is a child of the postwar area. Founded with UNESCO support in 1957 'on ecumenical soil crossing both East-West and North-South divides' and in its beginnings dominated 'by Europeans, particularly the French' (Nordenstreng, 2008), those origins dominated the association at least till Cold War's end. From the mid-1960s until the fall of the Berlin Wall the association's conferences were not just about academic results but also (and maybe even more) ideological battlefields. The East European countries tried to influence both the program including the respective publications as well as the recruitment of leadership positions. In essence, three objectives can be distinguished. First of all, the communists wanted to promote their very own concept of scientific work on journalism and mass media effects. This concept included formulas, precepts, and doctrines of Marxism–Leninism. Second, closely linked to the first aim, the East Europeans wished to win colleagues from the developing countries in Asia, Africa, and Latin America for the socialist idea on how to educate journalists and other mass media practitioners. This objective is part of the two systems' worldwide struggle for supremacy back then. And third, along with professors and doctoral students from the third world, so-called progressive Western academics, were also an important target group. Scientists working with materialist or other classic left theories and ideas should be encouraged and helped to present the results of their work to an international audience. Therefore, it was more than a battle of ideologies. Focusing on clearly academic issues such as theory and methodology, the definition of 'good science' was at stake (Meyen, 2013).

Driving force of all those efforts was the German Democratic Republic (GDR). For the communists in East Berlin, international organizations such as IAMCR were a tool in their fight for international recognition. Therefore, they helped to kick-start IAMCR. Formerly known as a rather small association with a permanent lack of money, IAMCR had a major conference in Leipzig in 1974 with about 250 participants from all over the world and two-volume proceedings. Emil Dusiska (1914–2002), leading professor in the communist journalism training center at the University of Leipzig, became IAMCR's secretary-general and did most of the association's paper work from 1972 to 1978. In return, IAMCR's policy was discussed in the communist party's inner circle in East Berlin.

East Berlin communists' support for IAMCR and their advertising for the well-equipped journalism department in Leipzig perfectly match communication research's general situation. Looking back to the late 1960s and early 1970s, we find a small discipline fighting for its existence. ICA 'was more like a family

affair' back then. Klaus Krippendorff (2012) describing the association in its beginnings: "The scholars who gathered at annual conventions knew each other, and the program was printed on one page folded three ways. For me, one of the most interesting conferences was in Phoenix, Arizona, in 1971. ICA rented a motel with rooms around a swimming pool, and there were few parallel sessions. The climate was completely different."

However, beginning in the early 1970s, in the field of communication research and media studies, the academic logic began to triumph over the political one. Apart from GDR's declining interest in IAMCR after joining the UN in 1973 that is the second significant reason of the decreasing influence of Eastern European representatives on the association's agenda. In 1974, Western scholars rejected a plan to launch an official IAMCR journal supported by Eastern European money with academic arguments. In the very same year, interesting enough, George Gerbner reinvented the *Journal of Communication*, which was originally set up in 1951. Within a short period of time that ICA publication became the field's flagship journal.

Summary and Current Situation

The field of communication research and media studies does not have the same aura like established disciplines such as history, literary studies, linguistics, psychology, or sociology. It is comparatively young, rather small and at the same time very broad. Moreover, at many places it is still easy to think of communication as primarily a technical, skills- or service-oriented discipline, not a social science. Citation patterns are an indicator for the field's position at large. More established subjects do not do a good job in searching literature from communication research. A typical political scientist might study the impact of the Internet for political processes, without ever reading a communication journal (Meyen, 2012). In contrast to Sisyphus, there is hope for the field. All over the world, communication has been growing vastly. There is a tremendous demand for undergraduate places. As a result, the discipline gets the best students and therefore, probably the best future faculty on the market. There is a second market force working for communication: a fast growing societal interest in media and communication issues and, linked with that interest, growing outside funding. The access to external money is changing the field right now and is going to shape its future structures, too. A prime example of the ongoing changes is the rise of health communication, but there are other winners as well – organizational communication, new media, and research on children and media.

See also: Human Sciences, History of; Lazarsfeld, Paul Felix (1901–76); Lewin, Kurt (1890–1947); Mass Media and Politics; National Socialism and Fascism; Social Science and Universities; Weber, Max (1864–1920).

Bibliography

Abbott, A., 2002. Chaos of Disciplines. University of Chicago Press, Chicago.
Averbeck, S., 2001. The post-1933 emigration of communication researchers from Germany. The lost works of the Weimar generation. European Journal of Communication 14, 451–475.
Averbeck, S., 2008. Comparative history of communication studies: France and Germany. The Open Communication Journal 2, 1–13.
Averbeck-Lietz, S., 2013. French and Latin American perspectives on mediation and mediatization: a lecture note from Germany. Empedocles 3 (2), 177–195.
Barnett, G.A., Danowski, J.A., Feeley, T.H., Stalker, J., 2010. Measuring quality in communication doctoral education using network analysis of faculty-hiring patterns. Journal of Communication 60, 388–411.
Beniger, J.R., 1990. Who are the most important theorists of communication? Communication Research 17, 698–715.
Bunz, U., 2005. Publish or perish: a limited author analysis of ICA and NCA journals. Journal of Communication 55, 703–720.
Carey, J., 1992. Communication and Culture. Essays on Media and Society. Routledge, New York.
Chaffee, S.H., Rogers, E.M. (Eds.), 1997. The Beginnings of Communication Study in America: A Personal Memoir by Wilbur Schramm. Sage, Thousand Oaks, CA.
Cohen, H., 1994. The History of Speech Communication: The Emergence of a Discipline, 1914–1945. SCA, Annandale, VA.
Craig, R.T., 1999. Communication theory as a field. Communication Theory 9, 119–161.
Craig, R.T., 2008. Communication as a field and discipline. In: Donsbach, W. (Ed.), The International Encyclopaedia of Communication. Blackwell, Malden, MA, pp. 675–688.
Craig, R.T., Carlone, D.A., 1998. Growth and transformation of communication studies in U.S. higher education: towards reinterpretation. Communication Education 47, 67–81.
Delia, J.G., 1987. Communication research: a history. In: Berger, C.R., Chaffee, S.H. (Eds.), Handbook of Communication Science. Sage, Newbury Park, CA, pp. 20–98.
Dennis, E., Wartella, E., 1996. Communication Research: The Remembered History. Lawrence Erlbaum, Hillsdale, NJ.
Donsbach, W., 2006. The identity of communication research. Journal of Communication 53, 437–448.
Donsbach, W., 2008. The International Encyclopaedia of Communication, 12 vols. Wiley, New York.
Dorsten, A., 2012. "Thinking dirty": digging up three founding "matriarchs" of communication studies. Communication Theory 22, 25–47.
Fleck, C., 2007. Transatlantische Bereicherungen. Zur Erfindung der Sozialforschung. Suhrkamp, Frankfurt.
Glander, T., 2000. Origins of Mass Communication Research during the American Cold War: Educational Effects and Contemporary Implications. Lawrence Erlbaum, Mahwah, NJ.
Hardt, H., 1992. Critical Communication Studies: Communication, History, and Theory in America. Routledge, Oxon, UK.
Hardt, H., 2001. Social Theories of the Press: Constituents if Communication Research, 1840s to 1920s. Rowman & Littlefield, Lanham.
Hardt, H., 2008. Forword. In: Park, D.W., Pooley, J. (Eds.), The History of Media and Communication Research. Peter Lang, New York, pp. xi–xvii.
Katz, E., Peters, J.D., Liebes, T., Orloff, A. (Eds.), 2003. Canonic Texts in Media Research. Polity, Cambridge.
Kuhn, T.S., 1962. The Structure of Scientific Revolutions. Chicago University Press, Chicago.
Krippendorff, K., 2012. I was never a real mass communication person. International Journal of Communication 6, 1699–1707. Feature.
Löblich, M., 2007. German "Publizistikwissenschaft" and its shift from a humanistic to an empirical social scientific discipline. European Journal of Communication 22, 69–88.
Löblich, M., Scheu, A., 2011. Writing the history of communication studies: a sociology of science approach. Communication Theory 21, 1–22.
Lowery, S.A., DeFleur, M.L., 1983. Milestones in Mass Communication Research: Media Effects. Longman, New York.
Meyen, M., 2012. International communication association fellows: a collective biography. International Journal of Communication 6, 2378–2396.
Meyen, M., 2013. IAMCR on the East-West Battlefield. A Study on GDR's Attempt to Use the Association for Diplomatic Purposes. "New Histories of Communication Study". ICA Preconference, London.
Meyen, M., Löblich, M., 2006. Klassiker der Kommunikationswissenschaft. Fach- und Theoriegeschichte in Deutschland. UVK, Konstanz.
Nordenstreng, K., 1968. Communication research in the United States: a critical perspective. Gazette 14, 207–216.
Nordenstreng, K., 2008. Institutional networking: the story of the international association for media and communication research. In: Park, D.W., Pooley, J. (Eds.), The History of Media and Communication Research. Peter Lang, New York, pp. 225–248.

Park, D.W., Pooley, J. (Eds.), 2008. The History of Media and Communication Research. Peter Lang, New York.
Park, H.W., Leydesdorff, L., 2009. Knowledge linkage structures in communication studies using citation analysis among communication journals. Scientometrics 81, 157–175.
Peters, J.D., 1986. Institutional sources for intellectual poverty in communication research. Communication Research 13, 527–559.
Peters, J.D., 1999. Speaking into the Air: A History of the Idea of Communication. University of Chicago Press, Chicago.
Peters, J.D., Simonson, P., 2004. Mass Communication and American Social Thought: Key Texts, 1919–1968. Rowman & Littlefield, Lanham, MD.
Pooley, J., 2008. The new history of mass communication research. In: Park, D.W., Pooley, J. (Eds.), The History of Media and Communication Research. Peter Lang, New York, pp. 43–69.
Pooley, J., Katz, E., 2008. Further notes on why American sociology abandoned mass communication research. Journal of Communication 58, 767–786.
Reeves, B., 2012. We were changing how we thought about media messages. International Journal of Communication 6, 1778–1784. Feature.
Rogers, E.M., 1994. A History of Communication Study: A Biographical Approach. Free Press, New York.
Rogers, E.M., 2001. The department of communication at Michigan state as a seed institution for communication study. Communication Studies 52, 234–248.
Schramm, W., 1959. Comments on "The state of communication research". Public Opinion Quarterly 23, 6–9.
Signorielli, N.D. (Ed.), 1996. Women in Communication: A Biographical Sourcebook. Greenwood Press, Westport, CT.
Simonson, P., 2008. Writing figures into the field: William McPhee and the parts played by people in our history of media research. In: Park, D.W., Pooley, J. (Eds.), The History of Media and Communication Research. Peter Lang, New York, pp. 291–320.
Simpson, C., 1994. Science of Coercion: Communication Research and Psychological Warfare, 1945–1960. Oxford University Press, New York.
Wahl-Jorgensen, K., 2004. How not to found a field: new evidence on the origins of mass communication research. Journal of Communication 54, 547–564.
Weaver, C.H., 1977. A history of the international communication association. Communication Yearbook, 607–618.
Wiedemann, T., Meyen, M., Löblich, M., 2012. Communication science at the centre of Cold War's communication battles: the case of Walter Hagemann (1900–1964). In: Jones, S. (Ed.), Communication@the Center. Hampton Press, New York, pp. 107–120.

Relevant Websites

http://historyofcommunicationresearch.org/ – Jeff Pooley: Bibliography for the History of Communication Research (over 1800 English-language works on the history of communication research and cognate fields).
http://www.helsinki.fi/crc/Julkaisut/Koivisto-ThomasPDF.pdf – Koivisto, Juha, Thomas, Peter, 2011. Mapping Communication and Media Research: Paradigms, Institutions, Challenges. Helsinki.
http://blexkom.halemverlag.de/ – Michael Meyen, Thomas Wiedemann (Eds.), Biographisches Lexikon der Kommunikationswissenschaft (German language).

Communication Skills in Social Work

Barbara Staniforth, School of Counselling, Human Services and Social Work, Faculty of Education, The University of Auckland, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The ability to communicate well with people forms the basis of all social work practice. Being aware of how different cultures communicate is essential in an increasingly global practice environment. Self-awareness is also critical so that social workers can be aware of how they and others filter incoming messages and make meaning. There are certain microskills that form part of a social worker's communication repertoire. These include attending and responding, reflecting thoughts and feelings, asking questions, summarizing and using silence. Social workers need to consider the specific contexts that communication skills occur within and equip themselves for all situations.

Introduction

Good communication skills form the basis of all social work practice. Without them, workers would be unable to form relationships, effect change, or support, advocate, or problem solve. Unlike certain skills that may be fairly straightforward to acquire, such as knot tying or typing, learning how to communicate well is a complex process where learning must occur on several levels simultaneously. Communication in itself is also a process whereby the social worker is only a part of the equation. This article looks at some of the critical issues that social workers need to consider in communicating well with others. It moves from broader areas such as culture and self-awareness to looking at particular skills that are commonly taught to beginning social work students.

Culture and Communication

From the time of conception, people begin their learning about communication. Families form the crucibles where people learn about how to communicate to get their needs met, about emotions, and also about the rules that help people make sense of the world. Families sit within particular cultural contexts that have formed over time. These contexts hold many of the stated, and often unstated, messages that teach people how to communicate 'acceptably' within their cultures. It is important for social workers to recognize not only the messages that they have been given within their own cultures about how to communicate, but also the messages that are inherent in other cultures.

There are various traits within cultures that are important to be aware of when considering communication. Some of these are the degree to which a culture values collectivism over individualism, the degree to which context plays a part in communication, the structure within a culture in terms of how power is recognized, and the degree to which uncertainty within communication is tolerated (Tubbs, 2010).

Collectivism

One of the central differences between cultures is the degree to which they value individual rights and freedoms over the greater good of the group as a whole. Countries such as the United States have evolved with strong messages about individual rights and freedoms being paramount. This is espoused in things such as the right to 'free' speech and the right for people to bear arms. In such situations, people are encouraged to express their own beliefs and feelings about things and to own them as their own. In other areas, such as Japan, China, many Pacific Island cultures, and other non-European cultures, the concept of 'self' is, at times, nonexistent, with each person existing only within the interconnectedness of their families or other cultural groupings. In such situations, people's identities are located within the greater group, and speaking for 'oneself' and having one's own opinions, thoughts, and feelings is downplayed or seen as undesirable. These differences have big implications for those whom social workers see as their 'clients.' Should they communicate initially with the individual or with the whole group? How should that communication occur? For instance, in Māori (indigenous people of Aotearoa New Zealand) culture, it would not be uncommon in a job interview or therapeutic assessment for the individuals themselves not to say very much but to have the family attend and speak on their behalf (Mead, 2003).

Power Distance

Related to the individualist/collectivist variable is how power and equality are seen within cultures. This is known as 'power distance.' Some cultures, such as Korean, Japanese, and many other cultures, value the knowledge and experience gained through age and apportion a strong sense of respect and differentiation in roles between the old and the young. Other factors such as gender may also play a role, with either men or women being assigned certain roles as spokespeople for the group (Tubbs, 2010). These factors are important to consider when thinking about how to make contact and build rapport with clients and their families.

High- and Low-Context Communication

The directness with which messages are given is one of the areas where communication can differ markedly between cultures. Some cultures (many Western ones) believe in the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28119-1

importance of stating things directly, while others believe that getting one's message across more indirectly is more appropriate. These differences are known as 'high'-context and 'low'-context situations, where direct communication (saying exactly what you mean) requires less reliance on the context to send and receive messages. 'High'-context cultures rely heavily on things such as tone and body language to send and receive messages, and may place less emphasis on the words being said (Tubbs, 2010). Listening and paying close attention to the situation is required and it can be difficult for people who have not grown up or lived within the culture to pick up the cues easily.

Uncertainty Avoidance

Related to directedness is the degree to which cultures are able to tolerate uncertainty in their communication. High-context cultures are often able to tolerate a higher degree of uncertainty, as much of what is being said may rest on subtle nuances. Low-context cultures, which are used to speaking and responding in a direct manner, may do so as they have a lower tolerance for uncertainty or ambiguity in communication (Tubbs, 2010).

The above cultural variables have at times been researched and numeric values have been attached to the degrees to which these variables have been demonstrated. These are known as 'cultural taxonomies.' Hofstede (2001) was one of the first sociologists to study and record such cultural traits. **Table 1** demonstrates some of the traits explored through another ongoing study from the Global Leadership and Organizational Effectiveness Research Program, which has explored the relationships between social culture, organizational culture, and leadership traits (House et al., 2004; Smith, 2006). While at times these taxonomies may be useful for bringing awareness of what different cultures value in communication, caution is advised as these taxonomies assume that a nation is composed of an amorphous culture, with subcultures not being taken into account. Assumptions should not be made about people solely based on these taxonomies.

While consideration of the differences in cultural communication styles is important, it is just one of the factors that need to be considered. An awareness of self is also an important skill for all social workers to possess.

Communication can be defined as a process whereby one person 'codes' a message in oral language or other cues and sends it to another person or people. This message is then filtered by the receiver(s), decoded, and then considered. A response then may occur, whereby the process is reversed, with the recipient becoming the sender, coding their message, and sending it back (see **Figure 1**). A critical variable in this step is the 'filtering' process, where culture may play a big role. Another variable at work in the filtering process is worker identity. Self-awareness is important in order to be able to consider how one's history and the multifaceted formation of one's identity contribute to how information is interpreted. In social work, some of this gets framed as transference and counter transference in interpersonal relationships, but it goes far deeper than this (Chang et al., 2009).

Identity and Self-Awareness

Many theories on development frame early childhood experiences as determining how views of self, 'other,' and the world are constructed. Attachment theory describes these as internal working models, cognitive behavioral theory frames them as core beliefs, and psychodynamic theories may consider these to be a person's 'object relations.' Whatever the theoretical orientation, the message being given is that the filters used to interpret information are likely to have been formed in a person's early years and then to have been further reinforced (or at times altered) by subsequent experiences. The quotation from Tennyson's poem *Ulysses*, 'I am part of all that I have met'

Table 1 GLOBE cultural taxonomy

Power distance	*Uncertainty avoidance*	*In-group collectivism*	*Institutional collectivism*	*Assertiveness*
Morocco Nigeria 152	Switzerland 200	Philippines 168	Sweden 229	Albania 202
South Korea 106	China 129	Turkey 103	New Zealand 133	Hong Kong 143
France 27	England 81	Nigeria 58	Poland 67	Australia 38
South Africa (white) −1	Kuwait 8	Ireland 1	Kazakhstan 10	Singapore 9
USA −68	Israel −25	Italy −26	Namibia −28	Iran −26
Netherlands −252	Venezuela −119	Canada (English) −119	Argentina −139	India −110
South Africa (black)	Russia −212	Denmark −219	Greece −235	Sweden −204
Denmark −305				

The Global Leadership and Organizational Effectiveness Project (Project GLOBE) (House, R.J., Hanges, P.J., Javiden, M., Dorfman, P.W., Gupta, V. (Eds.), 2004. Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies. Sage Publications, Thousand Oaks, CA.) explored national cultures of 61 nations, focusing on nine cultural dimensions. Questionnaires were used to collect data from middle managers in various industries around the globe over a 10-year period. The dimensions were measured on a continuum. Within the above chart a high positive score indicates that the country demonstrated a high degree of that trait, while a high negative score indicates a low demonstration. Zero is the average score. A brief summary of each trait is found below.
Power distance – the extent to which a community accepts and endorses authority, power differences, and status privileges.
Uncertainty avoidance – the extent to which a society, organization, or group relies on social norms, rules, and procedures to alleviate the unpredictability of future events.
In-group collectivism – the degree to which individuals express pride, loyalty, and cohesiveness in their organizations or families.
Institutional collectivism – the degree to which practices encourage and reward collective distribution of resources and collective action.
Assertiveness – the degree to which individuals are assertive, confrontational, and aggressive in their relationships with others (House, R.J., Hanges, P.J., Javiden, M., Dorfman, P.W., Gupta, V. (Eds.), 2004. Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies. Sage Publications, Thousand Oaks, CA, p. 30).
For further descriptions and critiques of the study see McCrae, R.R., Terracciano, A., Realo, A., Jüri, A., 2008. Interpreting GLOBE societal practices scales. Journal of Cross-Cultural Psychology 39, 805–810. doi:10.1177/0022022108323806 and Smith, P.B., 2006. When elephants fight, the grass gets trampled: the GLOBE and Hofstede projects. Journal of International Business Studies 37, 915–921.

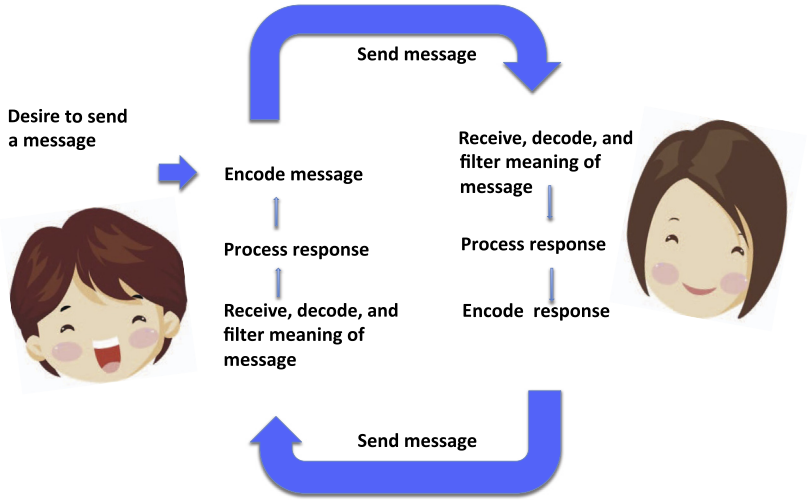

Figure 1 Process of communication.

(Tennyson, 1985: p. 57), demonstrates this well. It is therefore important for communicators to have a good knowledge of their experiences and how they have informed their filters as these are likely to inform how they both code and send messages, and also how they receive others' messages.

Some of the common areas that have been identified in the literature as impacting upon these communication filters are the following:

- the families that people have grown up within, and the rules they have imparted;
- the roles that people have played within their families;
- religious or spiritual experiences;
- the multiple cultures that people have been associated with;
- the connectedness or lack of connectedness that have been experienced with others;
- people's experiences of having their needs met or being responded to;
- gender;
- ethnicity;
- sexual orientation and experiences;
- experiences of safety, trauma, or abuse;
- experiences of being in a majority or minority culture;
- experiences of acceptance or oppression;
- genetic or physiological dispositions;
- experiences of health and illness; and
- the messages that have been given about equality and rights.

The list of factors is endless. Beginning social work students are often given assessments that encourage them to think about how some of these factors have impacted upon them and how they see themselves and the world. They are encouraged to consider how these things impact upon how they code and receive information and how all these same factors may have influenced how others code and receive information. Having this awareness enables students and social workers to see how people are all constructed differently and helps avoid hegemony or cultural relativism where it is difficult to see beyond the lenses or filters of the dominant group and the assumption that this is how all people see things (Chang et al., 2009).

Forms of Communication

People communicate in many different ways. Paintings, music, poetry, and various forms of behavior, such as violence, may all be intended to communicate ideas or emotions to others. While in social work some of these different forms of communication may be encouraged or observed, the most often taught forms of communication revolve around nonverbal communication, or body language, verbal communication, and written communication. Generally, all forms of communication are expected to occur within an environment whereby the social worker demonstrates the core interpersonal skills of respect, genuineness, and empathy. These skills form the foundation of trust and rapport that are essential in building good social work relationships.

Content and Process

Social workers are required to pay attention to many different aspects of communication simultaneously. The content of communication involves what is being said, planned, or decided upon in an interaction. Words exchanged between people form only a part of the communication process. Social workers must also be in tune with what is called 'process.' Process involves all that is being communicated between people that is not being expressed directly with words and often gives clues about what is occurring in the relationship between the people involved. For example, when working with someone, a social worker may find that the other person appears noncommunicative in response to questions asked of them. The content of the session may be the questions that are posed and answers that are stated. The process of the interview may be more complicated. A social worker may have to ponder upon the meaning of the apparent noncommunication. If someone has been made to attend the session, and does not

want to be there, the client may be angry for having to be there and feel resentful toward the worker. The client may be experiencing anxiety at meeting the worker for the first time and may not want to say 'the wrong thing.' Alternatively, the client may come from a culture where direct communication is not the norm and they may feel quite at ease in the situation. It is very important for social workers to be aware of the process and what is being communicated by it, and it is also important for them to check out their assumptions about what may be going on. A very important clue to exploring process may be communicated through nonverbal communication or body language.

Nonverbal Communication and Body Language

Much of what is communicated between people is not done through the use of words. Bull (cited in Harms, 2007) claimed that 85% of communication is transmitted nonverbally. Two aspects of nonverbal communication are 'proxemics' and 'kinesics.' Proxemics is concerned with distance and how close people like to be to one another, while kinesics involves things such as movements, gestures, expression, and eye contact (Lishman, 2009).

Proxemics

A common exercise for social work students is to get them to experiment with something that is sometimes known as 'space bubbles.' Students are asked to stand face to face and have one move toward the other. The nonmoving student must tell the other student to stop when they start to feel uncomfortable with the other's proximity. This exercise is done in different variations, with students standing, with one sitting, etc. It is a very valuable exercise as it demonstrates that all people have different comfort levels regarding their personal space. It is important for social workers to be mindful of the space bubbles of the people they are working with. This is especially important when working with people who might have experienced physical or sexual abuse, or in situations of high tension or anxiety that may be escalated by people feeling that someone is too close. Similarly, it may feel hard for someone to establish trust if their social worker sits too far away with a large table between them. Context also plays a big role in determining what a desirable or acceptable distance is. What may be acceptable on a rush-hour tube or subway in a big city will not be the same as on a remote and unpopulated beach.

Kinesics

How communication is delivered involves the use of kinesics. Movements, gestures, and facial expressions are all used to communicate. How eye contact is maintained or avoided may communicate different things in different cultures or situations. Gestures may be used to reinforce messages (nodding, shrugging shoulders, yawning) or symbols such as the 'thumbs up' signal or clapping after a performance may also be used to communicate. Gestures or facial expressions may also be employed to communicate emotions, such as smiling when someone is happy, rapid hand movements when a person is excited, or crying when the person is sad. In all these instances, social workers need to be mindful that the symbols or things being communicated by kinesics are not universal and can be open to misinterpretation. Culture and context always need to be considered (see Lustig and Koester, 2013). While a large amount of our communication may be done nonverbally, much of it is done through the spoken word.

Verbal Communication

There are many aspects of spoken communication that social workers need to be aware of when working with people. The volume, tone, speed, and words used all communicate messages to people. Social workers must be mindful of their audience when communicating and may need to communicate their message in very simple terms at times, or, at other times, in very technical language. It is crucial for the social worker to remain genuine and respectful no matter the context.

Congruence between verbal and nonverbal communication is encouraged for clear communication. Congruence in this context means that all aspects of a person's being are communicating the same message. When there is non-congruence between verbal and nonverbal communication (for example, smiling as very bad news is delivered), people will be confused about what message the sender wants to send. The sender may then appear nongenuine or untrustworthy. Matching verbal and nonverbal communication gives the best opportunity of ensuring that the message that is being sent is being received as intended.

Written Communication

Social workers must also be able to communicate well in written form. Good writing and documentation are important for many reasons. Many authors have written about writing and documentation (see, for example, Conroy, 2012; Healy and Mulholland, 2012; Musson, 2011; Staniforth and Larkin, 2006).

There are certain standards that are often cited as being critical in most written communication. Writing should be clear; free of grammar, spelling, and punctuation errors; and as concise as possible, while communicating all that it needs to communicate. Long winding sentences and use of slang or jargon should be avoided.

Like all forms of communication, written communication needs to be done with an awareness of its purpose and context. The language and style used for writing a court document will be quite different from what may be required in clinical notes, or in an advocacy letter to support a client.

Social Work in the Digital Age

Modes of communication have shifted dramatically over the past 10 years. Social workers are now able to communicate with their clients by e-mail, text, or instant messaging; via social networking sites; or by modes such as Skype. While these modes of communication have offered many advantages to

clients who may at times be quite isolated, these changes have happened quickly and social workers and agencies have been left scrambling to develop best practice policies and guidelines around use of these technologies. Issues such as boundaries and expectations are of paramount importance to explore, as these technologies offer access to all parties on an instant and continual basis. One of social work's core principles regarding confidentiality is also at threat as these forms of communication offer little protection from hacking, or the forwarding or transmitting of these messages all over the world within nanoseconds.

E-mail communications are often deemed to be legal documents, like clinical notes, and must be kept for certain periods of time. As such, the same consideration should be given to their writing style and content. Instant messaging has a language of its own (text-speak) and social workers should be mindful at all times about the context of the communication in terms of the style used. In both cases, social workers need to establish some form of documenting these communications within client files.

Facebook, Twitter, and other forms of social media communicate vast amounts of information about people, and social workers may unwittingly be disclosing very personal information about themselves that clients might easily be able to access. Similarly, there are ethical issues in social workers being able to monitor their clients via their Facebook pages or Twitter accounts. Communication via such sites is often done in a very public manner and needs to be very carefully considered. Healy and Mulholland (2012) and Thompson (2011) have explored these issues in useful depth. Many organizations, such as the National Association of Social Workers (2005) in the United States and the British Association of Social Workers (BASW) (2012) have put out policy documents regarding the use of technology and social media. BASW's position is that social media should be treated like other forms of communication whereby "social workers should ensure they maintain appropriate professional and personal boundaries and take responsibility for recognising ethical dilemmas presented by the use of different types of social media" (2012: p. 4). The term 'E professionalism' is used to describe how one manages information posted on the various types of social media. In the online world, there is little distinction about what someone may post in a personal and in a professional context, as it is all likely to be easily accessed. Social workers are cautioned against posting anything that may have the potential to breach client or agency privacy, or that may leave them open to the questioning of their professional conduct (BASW, 2012). Once something has been posted on the Internet it cannot be taken back. Social workers need to keep up to date with emerging technologies and recognize the importance of maintaining their critical skills in terms of using all forms communication.

Common Communication Skills

The social worker's main job in their initial interaction with people is to start to establish a relationship where they demonstrate that they are trustworthy and capable and that they care about the person or people with whom they are engaging. Being able to demonstrate the core skills of warmth, empathy, genuineness, and respect is critical in this process, known as 'forming rapport' with someone. There are other skills, often known as microskills or interview skills, which form the foundation of basic communication skills training for many helping professions (Koprowska, 2010; Thompson, 2009). As with all skills, careful consideration needs to be given to their appropriate use, depending on context, culture, and the individuals engaged in the process.

Observation, Attending, and Listening

People want to feel that they have been heard and understood in their interactions with others. A social worker who is intent on working well with their clients will take the time, at the beginning of a meeting, to carefully observe how the other person is presenting. The person's body language, appearance, emotional responses and how they speak and respond to questions can all reveal clues about what may be going on for the other person, and how the worker may need to respond or alter their way of being or presenting themselves to best enable the person to feel at ease. Social workers often develop a 'sixth-type sense' in being able quickly to interpret what may be occurring through careful observation. It is important, however, that social workers check out their assumptions with their clients. The use of reflection skills described further in this text may be a good way to do this.

Attending skills show people that the other person is really there with them. Ivey et al. (2012) describe attending as the building block of good listening and use the acronym 'three Vs + B' in good attending. 'Visual/eye contact' refers to looking at people when talking to them. 'Vocal qualities' refer to displaying warmth and interest. 'Verbal tracking' refers to helping the client tell their story by not interrupting and staying on track. 'Body language' refers to being genuine, being oneself, and demonstrating warmth by the use of body language and facial expressions (2012: p. 57). Being aware of how different cultures may interpret these elements, or what they may see as being ideal in attending, is essential in attempting to form relationships. The skills of responding and reflecting can also be used to demonstrate that someone is attending and listening well in a relationship.

Responding and Reflecting Thoughts and Feelings

Empathy forms one of the essential building blocks of social work and welfare practice (Egan, 2014). Students are taught the differences between sympathy and empathy in their first few classes and then spend the rest of their careers improving on their empathy skills as they expand their range of life experiences and working relationships. Empathy is often understood to mean trying to understand, as best as one can, how someone else is experiencing their life or a situation within it. While it is impossible for someone to truly know what it is like for someone else, when an attempt is made to try to get as close as possible to that experience, the other person is likely to recognize the effort, and feel more understood and that they have been listened to (Howe, 2013).

One of the ways that empathy may be demonstrated is through reflecting the thoughts and feelings of the other

person. Students often worry that this reflection will sound stilted, or that they may 'get it wrong,' thereby showing that they are not competent. Indeed, if what the social worker constantly reflects back is not what they are being told or shown, then the client may begin to feel misunderstood. More often than not, however, the social worker can demonstrate that they understand what the client thinks and feels, and, if slightly 'off the mark,' this gives the client the opportunity to expand or clarify their experience.

Asking Questions

There are many reasons why social workers ask questions. Questions can be used to obtain information about someone, gain rapport, ensure understanding of something, demonstrate interest in someone, or learn what someone thinks or feels about something. One of the most direct ways of obtaining information about people is to ask them questions. As discussed previously, different cultures value different levels of directness and the content of what is appropriate to ask someone will also vary according to context and culture. Within a social work assessment, social workers elicit information through asking open and closed questions (Sidell and Smiley, 2008).

Open questions are often used when trying to open up communication and obtain information in a general way. They are designed to allow the person to answer as they would like to, or to expand upon an idea or topic. Open questions often begin with words such as 'how' or 'what.' Open questions can also be phrased more as statements or requests, such as, 'Please tell me more about what you were thinking at that time.'

Closed questions are more directive and usually allow the responder to choose between limited options. Closed questions can be used in an interview or meeting if the client is finding it difficult to stay on track, or when the interviewer requires particular information and needs to get it in a timely way. Closed questions may ask for a 'yes' or 'no' answer, such as, 'Did you go skating yesterday?' or may ask for information such as age, birth date, or other personal data.

Once information has been elicited, it may be necessary to clarify certain points or the meaning held for the person. The skill of clarifying can be used in these situations. Clarifying may also be thought of as a 'drilling down' or funneling technique, as more specific information is sought about something. Asking for specific examples, such as, 'What does your child do when you think that she is acting out?' or 'What makes you think that nobody likes you?' can help in this process.

Summarizing

Students new to social work often struggle in role-play or interview situations. Many find that they are inclined to 'freeze' as they struggle to come up with the next question to ask, or are uncertain about what direction to proceed in. Summarizing can be a magical tool in such situations. Being able to review with the other person what has recently been said or occurred in a conversation can serve several purposes. It allows the interviewer the space and time to gather their thoughts and feelings together. It demonstrates to the client that they have been heard and understood, or alternatively provides a space for information and its meaning to be corrected or clarified. It can also be used as a skill for getting back on track when an interview has meandered too far from its desired path. While summarizing may be a useful skill to use when an interview is filled with long silent spaces, the use of silence is, in itself, an important skill to develop (Ivey et al., 2012).

Using Silence

Being able to sit comfortably with silence and being able to understand its potential meaning are two qualities that often demonstrate an advanced level of social worker 'use of self' and confidence in practice. It is also equally important, however, to be able to gauge a client's level of comfort with silence and to respond appropriately. Observation of body language may provide important cues in this regard.

Silence can mean many different things. Some possible things that may be occurring during silence could be that the other person requires time to think or process an emotion; that they may be anxious, overwhelmed, or angry; that they may not be comfortable sharing something; or they may have said all that they want to say (Lishman, 2009: p. 94).

Specialty Areas of Communication

Working Across Languages

Working across different languages presents rewards and challenges for people in relationships. While body gestures, facial expressions, and body language can all help with understanding, bridging a linguistic divide demands certain awareness and skills. Samovar et al. (2013) provide some considerations regarding communicating with someone whose first language is not the same as the worker's. Being mindful of the process for the nonnative speaker is important in terms of information needing to be taken in, translated, considered, retranslated, and then responded to. This can be a very tiring process. Speaking slowly and using language that is simple and free of technical terms or jargon is useful. Attention needs to be paid to nonverbal behaviors as cues to understanding, and workers should check in periodically to ensure that they are being understood.

Working with People Who Have Hearing Impairments

It is also important for social workers to be aware of potential things that can help if working with someone who experiences sight or hearing deficiencies. When working with people who are hearing impaired, it is important to learn about the culture of the hearing impaired world. In many countries, such as New Zealand, sign language is an official language. Learning the language, or at the very least common signs, can show respect and a strong desire to form a relationship. Creativity can help: perhaps using pen and paper or using of images in communication. Ensuring that the client can see the mouth of the social worker who speaks slowly and clearly may also be useful. Allowing more time for meetings is also a good idea (Koprowska, 2010).

Using Interpreters

Social workers will often make use of interpreters when working with people who speak a different language from theirs. Whether translating from oral language or sign language, there are some important ways to facilitate communication and demonstrate respect. Social workers should consider matching issues in terms of who may be appropriate to translate. Gender, linguistic variations, religion, and the community that the participants come from may all need to be considered to ensure that there is sufficient trust in the translator (Moss, 2012). Social workers should be very clear in their communication, and not express too many key ideas prior to the translator translating. The social worker should address the client and/or family members, not the interpreter. The relationship to build is with the client, not the translator. The worker should plan for meetings with translators to take more time than other meetings (Koprowska, 2010).

Skills in Advocacy Work

With social justice being at the center of much of social work practice, at times social workers need to advocate strongly with, or on behalf of, their clients. This may require more directive and assertive communication than many social workers may be familiar or comfortable with. Advocacy texts (Bateman, 2000; Moss, 2012), the Internet, assertiveness training courses, or public speaking courses may be very helpful in gaining the skills and confidence to speak out as needed. Consideration again of culture and context is extremely important in learning how to best help get client needs met and it is also important that the reasons for the advocacy and the manner in which it is to occur are clear to both worker and client. Social workers need to ensure that they do not disempower their clients through the process of 'doing for,' however well intentioned they may be.

Conclusion

Social workers need to be able to communicate well. They must be able to communicate across many different situations, in different agencies, and with people who may be quite different from themselves. Without the ability to communicate well, the job of social work cannot be done.

There are many issues to consider when attempting to communicate well with people. Knowledge of self and other, awareness of culture, and the development of a concrete set of skills are all essential. The development of empathy and a critical and reflective stance can also help social workers enhance their communication skills. This article has attempted to explore these basic aspects of communication while recognizing that there are many other skills and processes that go into communicating well. Social workers need to equip themselves with the general and specific communication skills that will enable them to work well with whomever they meet. It is an important ongoing learning process.

See also: Collectivism and Individualism: Cultural and Psychological Concerns; Cross-Cultural Training in Industrial and Organizational Psychology; Cultural Influences on Interpersonal Relationships; Cultural Intelligence and Competencies; Disabilism; Disability and Disabling Practices; Facial Emotion Expression, Individual Differences in; Facial Perception; First Language Acquisition, Developmental Psychology of; Māori Social Work; Managerial and Organizational Cognition; Social Work Ethics; Sociodigital Revolution: Digital Natives vs Digital Immigrants.

Bibliography

Bateman, N., 2000. Advocacy Skills for Health and Social Care Professionals. Jessica Kingsley Publishers, London, England.

British Association of Social Workers, 2012. BASW Social Media Policy. Downloaded from: http://www.basw.co.uk/resource/?id=954.

Chang, V., Scott, S., Decker, C., 2009. Developing Helping Skills: A Step by Step Approach. Brookes/Cole, Cengage Learning, Belmont, CA.

Conroy, K., 2012. Student writing in field education. In: Green, W., Levy Simon, B. (Eds.), The Columbia Guide to Social Work Writing. Columbia University Press, New York, pp. 85–113.

Egan, G., 2014. The Skilled Helper. A Problem-Management and Opportunity-Development Approach to Helping, tenth ed. Brooks/Cole, Belmont, CA.

Harms, L., 2007. Working with People. Communication Skills for Effective Practice. Oxford University Press, South Melbourne, Australia.

Healy, K., Mulholland, J., 2012. Writing Skills for Social Workers, second ed. Sage Publications, Thousand Oaks, CA.

Hofstede, G., 2001. Culture's Consequences: International Differences in Work-Related Values, second ed. Sage Publications, Beverly Hills, CA.

House, R.J., Hanges, P.J., Javiden, M., Dorfman, P.W., Gupta, V. (Eds.), 2004. Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies. Sage Publications, Thousand Oaks, CA.

Howe, D., 2013. Empathy. What It Is and Why It Matters. Palgrave MacMillan, Basingstoke, England.

Ivey, A., Bradford Ivy, M., Zalaquett, C., 2012. Essentials of Intentional Interviewing. Counseling in a Multicultural World. Brooks/Cole, Belmont, CA.

Koprowska, J., 2010. Communication and Interpersonal Skills in Social Work, third ed. Learning Matters, Exeter, England.

Lishman, J., 2009. Communication in Social Work, second ed. Palgrave Macmillan, Basingstoke, England.

Lustig, M.W., Koester, J., 2013. Interpersonal Communication across Cultures, seventh ed. Pearson, Upper Saddle River, NJ.

McCrae, R.R., Terracciano, A., Realo, A., Jüri, A., 2008. Interpreting GLOBE societal practices scales. Journal of Cross-Cultural Psychology 39, 805–810. http://dx.doi.org/10.1177/0022022108323806.

Mead, H., 2003. Tikanga Māori. Living by Māori Values. Huia Publishers, Wellington, New Zealand.

Moss, B., 2012. Communication Skills in Health & Social Care, second ed. Sage, Thousand Oaks, CA.

Musson, P., 2011. Effective Writing Skills for Social Work Students. Learning Matters Ltd., Exeter, England.

National Association of Social Workers & Association of Social Work Boards, 2005. NASW & ASWB Standards for Technology and Social Work Practice. Downloaded from: http://www.socialworkers.org/practice/.

Samovar, L.A., Porter, R.E., McDaniel, E.R., Roy, C.S., 2013. Communication between Cultures, eighth ed. Wadsworth, Cengage Learning, Boston, MA.

Sidell, N., Smiley, D., 2008. Professional Communication Skills in Social Work. Pearson Education Inc., New York.

Smith, P.B., 2006. When elephants fight, the grass gets trampled: the GLOBE and Hofstede projects. Journal of International Business Studies 37, 915–921.

Staniforth, B., Larkin, R., 2006. Documentation in social work: remembering our ABCs. Social Work Review 18 (3), 13–22.

Tennyson, A., 1985. Tennyson. Poems Selected by W.E. Williams, second ed. Penguin Books, Harmondsworth, England.

Thompson, N., 2011. Effective Communication. A Guide for the People Professions, second ed. Palgrave MacMillan, Basingstoke, England.

Thompson, N., 2009. People Skills, third ed. Palgrave MacMillan, Basingstoke, England.

Tubbs, S., 2010. Human Communication. Principles and Contexts, twelfth ed. McGraw Hill, New York.

Communication, Twostep Flow of

Gabriel Weimann, University of Haifa, Haifa, Israel

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 4, pp. 2316–2319, © 2001, Elsevier Ltd.

Abstract

The concept of the 'two-step flow of communication' suggests that the flow of information and influence from the mass media to their audiences involves two steps: from the media to certain individuals (i.e., the opinion leaders) and from them to the public. Since the introduction of the two-step flow model and the opinion leadership concept, numerous studies have sought to advance both the understanding and applicability of these ideas to various areas from marketing and consumer behavior, to fashion, politics, and scientific innovations. However, the concept has been a subject of growing criticism, leading to a decline in the popularity and attraction of the original concept and almost to its total collapse. Nevertheless, the paradigm's 'fight for survival,' through a dynamic exchange of arguments, criticism, response, and debate, resulted in several modifications of the original model.

The concept of the 'two-step flow of communication' emerged from *The People's Choice*, the 1948 American voting study conducted by Paul Lazarsfeld and his colleagues at the Bureau of Applied Social Research of Columbia University. They observed that: 'ideas often flow from radio and print to opinion leaders and from them to the less active sections of the population.' Thus, the flow of information and influence from the mass media to their audiences was suggested as taking place in two steps: from the media to the opinion leaders and from them to the public. This model certainly called into question most of the assumptions of the 'powerful media' notion by revealing the limits of media influence while highlighting the role played by personal influence and especially by certain individuals – the opinion leaders. These ideas caught the attention and imagination of a new generation of researchers who opened up a new theoretical and empirical vista.

The Emergence of the Concept

Beginning with *The People's Choice*, Lazarsfeld and his colleagues were occupied with a series of panel studies on the role of mass communication in the making of decisions in various areas (voting, consumption, public issues, and others). The researchers focused on the impact of campaigns in the media and were surprised to find that interpersonal communication had a stronger impact on individual attitudes than the mass media alone. Moreover, they found that certain individuals were more central and influential in their groups, often acting as intermediaries between the mass media and the public. In comparison with the rest of the population, these individuals were found to be more exposed to the mass media. Thus, if personal influence was so important and if it was executed mainly by certain individuals who were more exposed to the mass media than the people they influenced, then it could be postulated that there was a two-step flow process from the media to the opinion leaders and from them to their followers. These intermediaries, called Opinion Leaders, could allow messages to pass through (backing them with their own personal authority), strengthen or weaken them, or block their passage entirely. This became the 'two-step flow of communication' model.

The most thorough and systematic examination of the two-step flow model and the notion of opinion leaders was the Decatur Study (after the community in which it was conducted). The study's procedures and findings were published in *Personal Influence* and represented one of the most impressive early documentations of the opinion leadership concept (Katz and Lazarsfeld, 1955). The research focused on four areas of decision making in daily life: marketing, fashion, public affairs, and movie selection. By isolating specific changes of opinion and identifying the influences that allegedly produced them, the researchers located the individuals who were mentioned most frequently as the source of influence. As this was a pioneering study, Katz and Lazarsfeld applied various methods and measures to locate and identify the opinion leaders. Briefly, there are two ways to locate opinion leaders in specific areas of influence. One is through the testimony of those whom they have influenced. The other is through a self-designation procedure in which respondents themselves are asked about their influence on others. To verify the 'self-designating' answers, the researchers requested the confirmation of the persons mentioned as advice-seekers. Two-thirds confirmed the contact and the topic. Although the researchers were satisfied with this 'validation,' it is far from being an acceptable measure of validity and indeed attracted later criticism.

The early studies on personal influence and opinion leadership resulted in several attributes related to opinion leaders (Katz, 1957). Let us briefly review the main ones: (a) Opinion leaders are found at every social level and in most areas of decision making, influence people from the same social level. (b) Opinion leaders are found in both of the sexes, all professions, all social classes, and all age groups. (c) Opinion leaders tend to be more involved in various social activities and social organizations and occupy central positions in their personal networks. (d) Opinion leaders are considered to be experts in their field, but this is informal recognition by close friends, relatives, co-workers, colleagues, and acquaintances. (e) Opinion leaders are more exposed to the mass media than nonleaders. (f) Opinion leaders are more interested, involved, and updated in the field in which they are influential. (g) Opinion leaders tend to be monomorphous, that is, they are usually experts in one area but rarely in various areas (i.e.,

polymorphous). (h) Opinion leaders manifest a specific communication behavior insofar as they are more involved in formal and informal personal communication than nonleaders. (i) Opinion leaders are usually well aware that they serve as sources of information and influence for others.

Since the introduction of the two-step flow model and the opinion leadership concept, numerous studies have sought to advance both the understanding and applicability of these ideas. Already by the early 1970s several hundred of these studies were available and the number has continued to grow steadily (for a review of almost 1400 of these studies, see Weimann, 1994). Thus, the 1960s and 1970s may be considered as the 'Golden Age' of the research on opinion leaders, yielding convincing empirical evidence of the opinion leaders concept in various areas: marketing and consumer behavior, fashion, politics and voting, family planning, science and scientific innovations, agriculture, and health care.

Criticism of the Model

The model's theoretical and practical appeal crossed the boundaries of disciplines, societies, and cultures. It attracted researchers from medicine to politics, marketing to family planning, AIDS prevention to clothing fashion, and agriculture to birth control. However, the concept has been a subject of growing criticism. The weaknesses and shortcomings of the original conceptualization of opinion leadership, and especially that of the two-step flow model, were exposed by several studies which combined theoretical, ideological, and methodological aspects (for a review, see Katz, 1987).

Theoretical Ideological Criticism

The media, according to the Columbia research tradition, were not as powerful as was assumed, once the public was found to be more active, less lonely, and less vulnerable. Defended by their selective consumption of the media and by the selective mediating function of the opinion leaders, people could fend off media influences and protect their attitudes, preferences, and values. The limited effects of the media as a central element in the opinion leadership and two-step flow concept became the main source for the theoretical-ideological criticism directed at this paradigm. Gitlin (1978), for example, rejected the two-step flow model and the functions related to the opinion leaders. The opinion leader is celebrated, he argued, because administrative researchers study proximate causes which are accessible to their patrons, the media organizations themselves, and the dominant values of the capitalist society. The ideological-theoretical bias of the two-step model, according to this criticism, is manifested in a set of assumptions: (a) *Commensurability of modes of influence:* the impact of the mass media was presumed to be comparable to personal influence, with the two 'forms of influence' regarded as functionally equivalent. (b) *Power as distinct occasions:* the measurement of influence by studying discrete incidents in a 'behaviorization of power.' (c) *Commensurability of buying and politics:* the areas of marketing, politics, fashion, and moviegoing were perceived as comparable within a single theory and measurable by a single method. (d) '*Change*' *as the sole dependent variable:* influence was identified by the occurrence of 'attitude change,' disregarding the case of reinforcement as an effect or the prevention of change as an effect. The latent power of the media to prevent changes and maintain the status quo should not be regarded as a residual category, but as a major function and a powerful effect.

Empirical-Methodological Criticism

The empirical-methodological weaknesses cited (see Weimann, 1982) include the following aspects: (a) Ignoring the evidence of direct flow from the media to the public. (b) Ignoring horizontal flow among leaders or among followers. (c) Ignoring the more than two-step flow, or 'longer chains of influence.' (d) The use of a crude dichotomy between leaders and nonleaders instead of applying continuous measures of leadership. (e) Standardization of measurement for information flow and influence flow. (f) Standardization of measurement across domains, disregarding the variance across domains of decision making. (g) The use of short timeframes, far from reflecting the long-term, cumulative impact of media messages.

The Paradigm's Fight for Survival

The wave of criticism led to a decline in the popularity and attraction of the original concept and almost to its total collapse. However, the paradigm's 'fight for survival,' through a dynamic exchange of arguments, criticism, response, and debate, resulted in several modifications of the original model.

The first amendment to the two-step flow of communication was the introduction of multistep rather than two-step flow. This allowed the inclusion of other optional directions of flow, such as horizontal (among leaders, among followers) and vertical (leaders to followers, active to inactive); direct (linking two actors directly) and indirect (linking actors through a third party or a chain of intermediaries); and downward (leaders to followers, media to public) and upward (from opinion leaders to media, followers to influentials, marginals to centrals). Another important modification was the introduction of continuous measures of opinion leadership and their validation (e.g., Noelle-Neumann and Csikszentmihalyi, 1992; Weimann, 1991, 1994).

Finally, the most promising development is the merging of the model with modern network analysis (Weimann, 1989). The developing theory and methodology of network analysis, combined with modern technologies of data processing, namely high-speed computers, special programs for sociometric data analysis, and such sophisticated mathematical procedures as matrix algebra and sociometric topology, has facilitated the study of large communities. It has also contributed to 'modernizing' the concepts of personal flow of media messages, opinion leadership, and flow of influence. Sheingold (1973) called for the use of social networks data as a 'resurrection' of the Columbia research agenda, and indeed much research on social networks in recent decades has focused on the role of influentials or 'centrals,' applying various structural measures taken from the network analysis procedure. These new developments have linked the two-step flow model with other theories and concepts, such as the Agenda-Setting process

(e.g., is there a two-step flow of agenda setting?; see Weimann and Brosius, 1994; Brosius and Weimann, 1996) and have thus contributed to revitalizing some of the notions of the two-step flow model.

See also: Advertising: Effects; Advertising: General; Consumer Psychology; Consumption, Sociology of; Lazarsfeld, Paul Felix (1901–76); Media Effects; Social Influence, Sociology of.

Bibliography

Brosius, H.B., Weimann, G., 1996. Who sets he agenda? Agenda setting as a two-step flow. Communication Research 23, 561–580.

Gitlin, T., 1978. Media sociology: the dominant paradigm. Theory and Society 6, 205–253.

Katz, E., 1957. The two-step flow of communication: an up-to date report on an hypothesis. Public Opinion Quarterly 21, 61–78.

Katz, E., 1987. Communications-research since Lazarsfeld. Public Opinion Quarterly 51, S25–S45.

Katz, E., Lazarsfeld, P.F., 1955. Personal Influence. Free Press, Glencoe, IL.

Lazarsfeld, P.F., Berelson, B., Gaudet, H., 1944. The People's Choice. Duell, Sloan and Pearce, New York.

Noelle-Neumann, E., Csikszentmihalyi, M., 1992. Personality Strength: A New Variable for Opinion-Attitude Research. Paper presented at the WAPOR Conference, St. Petersburg Beach, Florida.

Rogers, E.M., Shoemaker, F.F., 1971. Communication of Innovation, second ed. Free Press, New York.

Sheingold, C.A., 1973. Social networks and voting: the resurrection of a research agenda. American Sociological Review 38, 712–720.

Weimann, G., 1982. On the importance of marginality: one more step into the two-step flow of communication. American Sociological Review 47, 764–773.

Weimann, G., 1989. Social networks and communication. In: Asante, M.K., Gudykunst, W.B. (Eds.), Handbook of International and Intercultural Communication. Sage, Newbury Park, CA.

Weimann, G., 1991. The influentials: back to the concept of opinion leaders? Public Opinion Quarterly 55, 267–279.

Weimann, G., 1994. The Influentials: People who Influence People. SUNY, Albany, NY.

Weimann, G., Brosius, H.B., 1994. Is there a two-step flow of agenda-setting? International Journal of Public Opinion Research 6, 323–341.

Communicative Competence: Linguistic Aspects

Constanze Vorwerg, Center for the Study of Language and Society, University of Bern, Bern, Switzerland; and Institut für Sprachwissenschaft, University of Bern, Bern, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S.H. Foster-Cohen, volume 4, pp. 2319–2323, © 2001, Elsevier Ltd.

Abstract

The notion of communicative competence refers to an individual's capability to communicate successfully in terms of both effectiveness (goal achievement) and appropriateness (acceptability in relation to context). The concept is related to debates about the nature of language and was developed by Hymes as a counter concept to Chomsky's proposal of linguistic competence. In contrast to this theoretical position, the concept of communicative competence sees language primarily as a means of communication, intertwined with other communicative media, coupled with and based upon other cognitive and social abilities, related to performance, and comprising all aspects of language including illocutionary and sociolinguistic knowledge, in contrast to being confined to grammar (in terms of syntactic, phonological, and semantic rules). The notion has also been developed within second-language teaching and interpersonal communication research. Depending on the specific focus of research, several constitutive components have been identified. It is a gradual concept, with communicative competence underlying performance and being built up with practice and experience. At the same time it may, depending on the communicative purpose, also be a function of more basic traits, states, and abilities (such as social relaxation, mind-reading abilities, and relevant kinds of knowledge). There is a vast amount of research on different aspects of communicative competence, including its implications for personal life; its components and measurement depending on setting and specific communicative function; the development of communicative competence, methods of improvement, and the processes involved in competent communicative behaviors.

The term 'communicative competence' refers to an individual's capability to communicate successfully and underlies actual communicative behavior in concrete situations. The notion of communicative competence comprises, thus, both knowledge of the communicative means of a community – as reflected in judgments of acceptability – and the ability for its use.

The concept emerged in the context of debates in the early 70s of the twentieth century, about the nature of language and the proper way to study it; and is thus closely related to fundamental theoretical issues, such as theory of language and the relationship between language and communication. Further developments in the scientific history of the notion of communicative competence occurred in conjunction with research on interpersonal communication and social interaction focusing on face-to-face settings and aiming at identifying the relevant dimensions of communicative competency in social encounters. Other models of communicative competence, in terms of components, stem from research in second language teaching and learning. The concept is thus of great practical relevance – also in respect of other areas, such as language development in children; specific communicative functions, such as persuasion, arguing, negotiations, emotional support, etc.; social settings, such as parenting, healthcare, educational settings, etc.; and intercultural encounters. There is now a substantial and diverse research literature on different aspects of communicative competence from different scientific fields, including communication sciences, sociolinguistics, psycholinguistics, psychology, language testing, speech disorders, education, management, and social work – even though definitions of what constitutes communicative competence differ widely.

The Concept of Communicative Competence in Historical Perspective

Diversity of Speech and Knowledge of Language

The term 'communicative competence' was introduced by the sociolinguist Dell Hymes (1972) and independently also by others, e.g., Sandra Savignon (1972) in the area of second language teaching, involving the recognition that mastery of a language comprises a wider range of abilities than grammatical knowledge (cf Hymes, 1992). While having referred to 'competence' in the context of communication previously, Hymes coined the expression 'communicative competence' specifically as a reaction to Noam Chomsky's definition of 'linguistic competence' and his fundamental distinction between 'competence' and 'performance.' "When 'Aspects of Syntax' appeared […] (1965), a first reaction was that Chomsky could not really be serious in dividing the world of language between 'competence' and 'performance', while defining the former in so limited a fashion" (Hymes, 1992: 32).

In his endeavor to lay the foundations for (the syntactic component of) a generative grammar – which is supposed to be able to generate, with a limited system of rules, all of a language's indefinitely many sentences out of structural descriptions – Chomsky (1965) maintains that "[l]inguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance" (p. 3). While a language's grammar in this theoretical approach is understood as description of the immanent

linguistic competence of the ideal speaker-listener, linguistic competence is also assumed to underlie a real speaker-listener's linguistic performance. However, according to Chomsky's conceptualization, linguistic use or performance in concrete situations is just a poor reflection of linguistic competence, as natural speech is full of wrong starts, rule deviations, midsentence changes, etc. Thus, for Chomsky, linguistic competence is the individual's knowledge of one's language, which is intuitive, tacit, and only in part accessible to consciousness, and which allows one, in principle, to understand any sentence of one's language and produce a sentence expressing this idea. For the sake of scientific inquiry it is assumed that it is homogeneously represented in the minds of the members of a speech community, abstracting from all variation such as styles or dialects (cf Chomsky, 1980).

This dichotomy of competence/performance is rejected by Hymes (1972) along with the idea of a perfect or even homogeneous competence in a language, its narrowing to aspects of grammar, and its isolation from sociocultural features (for arguments of other authors put forward in the controversy of the time, see Hymes, 1972, 1992). One concern of Hymes' approach is to explain the communicative differences between groups of children in order to provide support for children in need, for which a linguistic theory in terms of Chomsky's conceptualization seems inapplicable. Hymes argues that "[t]o cope with the realities of children as communicating beings requires a theory within which sociocultural factors have an explicit and constitutive role" (1972: 271). One starting point in his considerations is the situation of individual or societal bilingualism, in which fluent speakers often regard their languages as differing in communicative adequacy depending on situation and function. This observation can be extended to varieties of a language. More generally, there is a combination of diversity and differential competence within communities. In addition, competencies in production and reception may differ, and what is and is not acceptable and intuitively correct in linguistic grammar may depend on social and contextual constraints. Altogether, the term 'communicative competence' chosen by Hymes places an important role on the heterogeneity of speech communities, differential competencies, and sociocultural features in a linguistic theory.

Another important aspect of communicative competence, in this sense of the word, is that a child learning a language acquires not only knowledge of the grammar, but also of appropriateness. This involves competence of when to speak, what to say to whom in what situations, and in what manner, the mastery of a range of speech acts, and integration with other communicative means, as well as attitudes toward language. According to Hymes, there are several parts of communicative competence, of which grammatical competence is just one. There are several systems of rules underlying linguistic behavior, which are reflected in the judgments that language users are able to make. Instead of a dichotomy, Hymes proposes the following four types of judgments which must be recognized in a linguistic theory, formulated sufficiently generally to integrate language with other means of communication (p. 281):

1. Whether (and to what degree) something is formally *possible*;
2. Whether (and to what degree) something is *feasible* in virtue of the means of implementation available;
3. Whether (and to what degree) something is *appropriate* (adequate, happy, successful) in relation to a context in which it is used and evaluated; and
4. Whether (and to what degree) something is in fact done, actually *performed*, and what this entails.

The Nature of Language and Its Relation to Communication

Within language-focused debates, the controversy between accounts in terms of linguistic competence versus communicative competence is closely related to the fundamental question of what language is – forming the core of the theory of language. This, in turn, bears on issues such as: How is language acquired? How did language evolve and change? What is the function of language? How is it related to (other forms of) communication?

So, the description in terms of homogeneity versus heterogeneity; perfect competence which can only be impaired in performance versus differential competence reflected in language use; grammar as the core of human language capacity versus language as a range of abilities integrated with other modes of communication is not, in both approaches, simply a matter of different focus or abstracting away from certain aspects for simplification in scientific study. It is deeply rooted in fundamentally different ideas of what language is and how it should be studied.

'Linguistic competence' is seen as a characteristic of the human mind, whose acquisition by a child is based on an innate theory of language that allows them to construct or select a generative grammar out of an imperfect section of a language encountered, which they can use as primary linguistic data for testing hypotheses about the structure of the language. In this view, children acquire the infinite ability to create or understand sentences, (tantamount to linguistic competence) with ease and in a short time, because they have prior, innate knowledge of a universal grammar. Universal grammar, in this account, is assumed to express the profound regularities of all language capturing the creative aspect of language use (Chomsky, 1965) and to provide the innate predispositions or constraints underlying language development (Hauser et al., 2002). Language is mainly conceived of as an internal instrument for thought, of which only part reaches consciousness, and relative to which all usages that depend on externalization, including communicative functions, seem peripheral. A more recent way to put this is as a distinction between language as a communicative system and language as the internal component with underlying computation (Hauser et al., 2002). Accordingly, a distinction between the faculty of language in the broad sense (including a sensory-motor system, a conceptual-intentional system, and the computational mechanisms for recursion) and the faculty of language in the narrow sense (including only recursion) is made, and it is hypothesized that the latter is the only uniquely human component. According to this conception, the computational component is the core of the human language capacity; it links the sensory-motor and the conceptual-intentional systems, and it (i.e., language) may

have emerged without selection pressure, for reasons other than the adaptive evolution of communication.

'Communicative competence,' on the other hand, is seen as a complex range of linguistic abilities underlying the capacity to communicate successfully, not autonomous but intertwined with other cognitive (and maybe also motivational and emotional) processes, and being integrated with other means of communication. In this view, children acquire language in social interaction under important sociocultural conditions and in a social world, with true learning taking place and based on more general cognitive mechanisms instead of inborn grammatical principles (see, e.g., O'Grady, 2008). Variation is a central aspect of language, which not only adds to its structure but partly constitutes it. Language is intricately related to language users' cognitive constraints, such as learnability and processability. Both limitations of memory, articulation, attention etc., and social processes contribute to the actual form of linguistic structures (cf Hymes, 1992; for more recent approaches emphasizing the selectional pressure from generations of language users, the repeated construction and interpretation of utterances in social interaction, and choice from a pool of linguistic variation, see Christiansen and Chater, 2008; Hruschka et al., 2009). More generally, according to this account, language cannot be defined as linking meaning and sound. It is conceptualized as comprising a huge variety of speech acts and communicative functions and being related to communicative behavior and social life. The social dimension is constitutive for the linguistic level of speech acts, specifying whether a specific grammatical sentence is meant as, let's say, a statement, an offer, or a threat, as well as for the rules of use that determine what linguistic forms are appropriate under what conditions.

In addition to the different answers to the issue of the nature of language and other questions related to it, Chomsky's linguistic competence and Hymes' communicative competence take also different epistemological stances. Accordingly, the methodological positions and the ideas of how scientific inquiry should proceed within linguistic research differ considerably. The concept of linguistic competence belongs to a decidedly rationalist approach relying on intuition, introspection, and logic with an explicit rejection of objective data, leading to the situation that the intuitive, tacit, only partially accessible knowledge of language users is the only source for determining the adequacy of a linguistic theory in this concept. In contrast, the concept of communicative competence following rather an empiricist approach to scientific knowledge, attempts to draw on a variety of data collected with quantitative as well as qualitative methods, with convergence of different kinds of data used for complementation and the strengthening of evidence.

Another counter-model to Chomsky's 'linguistic competence' was the concept of 'communicative competence' developed by the social philosopher Jürgen Habermas (1970) in his endeavor to lay the foundations for a revision of sociological role theory and an analysis of the 'logic of explanation.' According to his theory, the ability for argumentation and discourse is a precondition for acting rationally and discursively reaching consensus and understanding, based on the communicative competence of interlocutors. Habermas criticizes the monological nature of the concept of linguistic competence constructed in the mind of speakers, prior to all communicative experience. His concerns relate mainly to the semantic component of generative grammar, specifically the conjecture of a priori universal, semantic meaning components constraining the possible concepts in mind, which is incompatible with the assumptions of (1) semantic universals being part of intersubjectively produced cultural systems (critique of monologism); (2) semantics reflecting specific realms of experience (critique of apriorism); and (3) more global representations of nature and society (critique of elementarism). Habermas holds that "[i]t is not enough, to understand communication as an application – limited by empirical conditions – of linguistic competence" and that "[o]n the contrary, producing a situation of potential ordinary language communication belongs by itself to the general competence of the ideal speaker" (p. 138). While Habermas, in contrast to Chomsky, emphasizes the role of intersubjectivity, regarding it as linguistic, his theory is similar to Chomsky's concept in referring to an 'ideal' speech situation: "[C]ommunicative competence relates to an ideal speech situation in the same way that linguistic competence relates to the abstract system of linguistic rules. … Communicative competence is defined by the ideal speaker's mastery of the dialogue constitutive universals irrespective of the actual restrictions under empirical conditions" (pp. 140–141).

Mastering a Second Language

The question of what it means to master a language is also of central importance to second language education. The concept of communicative competence entered this field independently of Hymes' proposal, in connection with communicative approaches to language teaching, but started soon to be discussed with reference to his theoretical approach and arguments. It generally recognizes the need for good communication skills in the language to be acquired and the necessity of using language (at least part of the time) for communicative purposes, in order to acquire this competence. The term may refer to the ability to participate in linguistic interaction and the importance of using communicative settings, or it may more specifically refer to the social rules and knowledge of appropriateness depending on context, situation, and purpose.

An influential model of communicative competence within the context of language teaching was developed by Michael Canale and Merril Swain (1980). It includes three main competencies: grammatical competence, sociolinguistic competence, and strategic competence. Grammatical competence includes the knowledge of lexical items and morphological, phonological, semantic and syntactical rules. Sociolinguistic competence refers to rules of appropriateness and of discourse, while strategic competence comprises the means to compensate for breakdowns in communication. In all three components of communicative competence, differential probabilities of occurrence are assumed to apply.

Based on this and other research, a refined model of the components of communicative language ability was constructed by Lyle Bachman (1990) as theoretical framework for language testing. Its main three components are: (1) the knowledge of the language in terms of the set of components needed for communication via language, called 'language

competence'; (2) the capacity for implementing those knowledge components in communicative use, called 'strategic competence'; and (3) the ability to actually execute language in terms of 'psychophysiological mechanisms.' The main components are, in turn, comprised of more specific components that can, again, hierarchically be subdivided into subcomponents. In language use, components interact with each other and also with features of the communication situation. Language competence includes both knowledge of how to relate and organize linguistic symbols (in terms of both grammatical and textual competence components) and knowledge of how to make utterances acceptable in certain communication situations (in terms of both illocutionary functions and sociolinguistic appropriateness). Strategic competence is related to the dynamic character of language use emphasized in this approach. Different from Canale and Swain's model, the term refers not only to compensatory communication strategies here, but more generally to the mechanisms involved in situation assessment (including the interlocutor's capabilities and achievement of communication goals) as well as language-planning and language-executing processes in productive or receptive language use. The execution itself depends also on psychophysiological mechanisms as the third main component of the model.

Communication and Interpersonal Interaction

A different approach to communicative competence has been taken within the literature on social interaction and interpersonal communication. In an attempt to analyze good or effective conversational behavior in everyday face-to-face interaction, John Wiemann (1977) constructed a model of communicative competence composed of the dimensions of (1) affiliation/support, (2) social relaxation, (3) empathy, (4) behavioral flexibility, and (5) interaction management skills (including conversational turn-taking, topic control, and attention.) These dimensions were tentatively proposed on the basis of three previous approaches to interpersonal competence: Goffman's self-presentation, Argyle's social-skill, and the training-in-social-relations approaches, and put to test in an experimental study. An important criterion for communicative competence in this conception is whether a conversational encounter is evaluated positively by other communicative participants. Results revealed a strong, positive, linear relationship between interaction management and communicative competence, and also positive correlations between perceived competence and other components of the model. According to these results, the competent communicator in face-to-face conversational encounters is empathic, supportive, and relaxed; as well as capable of adapting to situational changes and securing smooth speech turns.

The quality of interpersonal communication, with a focus on dyadic, face-to-face interaction is also the topic of Brian Spitzberg and William Cupach's (1984) model of 'interpersonal communication competence.' However, the scope of their theory is much broader than Wiemann's confinement to conversational encounters, with communication competence being defined as "the ability to engage in appropriate and effective interpersonal interaction" (p. 14) relevant in a multitude of interpersonal events, such as interviewing, persuasion, negotiation, conflict management, and social conversation. Accordingly, the theory draws on diverse historical precursors in rhetorics, psychiatric approaches to social functioning, and research on marital and other interpersonal relationships. It bases itself on a review of prominent theoretical approaches to competence in communication – including social-competence, interpersonal-competence, and relational-competence models in addition to the approaches of Chomsky, Habermas, and Hymes to linguistic versus communicative competence – and a discussion of key assumptions in research on communication competence. While all of these approaches deal with certain aspects of socially communicative behavior, some of them are concerned with the communicative behavior itself, focusing on appropriateness; some are concerned with the outcomes of communication, focusing on effectiveness; and some focus on the relation between communication processes and functional outcomes. Arguing that both appropriateness and effectiveness are integral to competent communication, Spitzberg and Cupach identify four global constructs of communicative behaviors involved: (1) 'Interaction management' is the ability to structure and maintain interpersonal interaction smoothly, brought about by behaviors such as topic-initiation, turn-yielding, and turn-taking signals, use of questions, sentence completion, head nods, and eye-contact initiation. (2) 'Altercentrism' concerns other-oriented behaviors, such as topic support, compliments, eye contact, head nods, body lean, smiling, and other manifestations of empathy, supportiveness, attentiveness, politeness, and approval. (3) 'Expressiveness' refers to the vividness of facial expression, vocal modulation, and gesturing, and the ability to articulate meaning using a variety of means, including clarifications, requests, humor, and expressions of opinions or feelings. (4) 'Social relaxation' is conceived of as the absence of indications of social anxiety, such as crossed arm or legs, eye contact avoidance, arm posture rigidity, speech nonfluencies, tense voice, and speech blockage.

An attempt to explicate communicative competence in terms of specific communication theories has been undertaken by Stephen Wilson and Christina Sabee (2003), dealing with five classes of communication theory: expectancy theories, attribution theories, goals-plans-action theories, hierarchical message production theories, and dialectic theories of relationships. While these theoretical approaches offer different perspectives on communicative competence and consider different aspects, a more general approach taken by a number of authors is to focus on core criteria that are most general and comprehensive and provide a common framework for various theories and empirical studies (see, e.g., Rickheit et al., 2008; Spitzberg, 2003).

The Essence of Communicative Competence

Putting Together Social and Linguistic Perspectives

Even though research on communicative competence has dealt with disparate aspects of able communication, there is a large overlap in concerns and basic ideas. A main distinction can be drawn between studies and theories that focus on social

encounters – both in terms of the type of communication situation studied and in terms of the processes involved in social perceptions and relationships – and between studies and theories that focus on language as a central and sophisticated means of communication, often intertwined with other communicative mediums.

The social perspective on communicative competence is closely related to concepts such as interpersonal, social, or relational competence, with terms chosen not always being indicative of the exact scope or issue of a study or theory. Some of the approaches undertaken from this perspective, address other-oriented and interaction-management characteristics related to social nonanxiety, emotional intelligence, and social judgment. Others aim at describing the essential elements of successful interpersonal communication, in terms of internal judgment and planning processes or mutual influences. The research on social interaction skills has been reviewed in several overview articles and comprehensive handbooks (e.g., Greene and Burleson, 2003; Knapp and Daly, 2011).

The linguistic perspective on communicative competence, also referred to as communicative language ability by Bachman, is devoted to an understanding of what is involved in the use of communicative language and what is needed in order to communicate in a given language, subscribing to an expanded view on language and language proficiency that takes into account the context of language use and the question of how language is used to achieve particular communicative goals and to do things. This includes all the kinds of knowledge required to use language, including linguistic form, textual as well as illocutionary linguistic functions, context in terms of both discourse and sociolinguistic situation, and sociocultural factors. An overview of the state of knowledge is given in the 'Handbook of Communication Competence' (Rickheit and Strohner, 2008), which appeared in the series 'Handbooks of Applied Linguistics.'

Both interpersonal and linguistic competencies contribute to communicative competence, with some factors enabling others to be brought to bear, some abilities making up for each other, and further aspects being indispensable depending on the specific communicative situation, such as empathy for providing emotional support or textual competence for giving a public speech. Most importantly, linguistic and social perspectives on communicative competence converge in similar views on the conceptual importance of communication in theorizing, in regarding the capacity to communicate skillfully as a kind of overarching goal of individual development, in seeing communicative competence as a graded concept, and in the dual character of communicative competence implying both effectiveness and appropriateness.

Effectiveness and Appropriateness as Core Criteria

While different approaches have put different emphases on effectiveness or appropriateness considerations, at least one of both definitional criteria plays a role in virtually all of them. Moreover, in contrast to early distinctions, many more recent approaches to communicative competence consider both elements as constitutive (see Rubin, 1990). Both norms and intentions guide our communicative behavior in production and interpretation; both appropriateness and effectiveness characterize communicative competence (see the reviews given in Rickheit et al., 2008; Spitzberg, 2003).

In production, effectiveness is goal accomplishment – the ability to achieve communicative goals, to get meaning across, to say what one means to say, to perform actions, to influence others, to form relationships, to express oneself, etc. Multiple goals may be pursued simultaneously; on the one hand they are often hierarchically embedded, on the other hand a main goal underlying a communicative act may be accompanied by secondary goals (such as maintaining a good relationship to the addressee). In some situations, efficiency – the ability to achieve communicative goals with less effort and resources – may be additionally relevant.

In reception, effectiveness concerns the ability to infer a speaker's meaning. This may, e.g., imply an understanding of whether an utterance is meant literally or ironically, what speech act a speaker wants to perform with an utterance, and whether, let's say, an ironic utterance is meant to be a critique or simply a joke. In both production and reception, effectiveness implies the knowledge of how language (or another communicative medium) is used to achieve particular communicative goals, and how to implement this knowledge in actual language use. In Bachman's model of communicative language ability, effectiveness is described in terms of both illocutionary functions and strategic competence (including the capacity to access what information is needed, and what language resources are available, for achieving a communicative goal, to retrieve the relevant items from language competence, and to form a plan to achieve this communicative goal).

Appropriateness concerns sensitivity to context – the ability to choose linguistic and other communicative means in accordance with norms of acceptability and legitimacy, which may depend on situation, function, the speaker's identity, and sociocultural factors, or to assess the communicative means chosen by another speaker accordingly. It involves sensitivity to differences in dialect, register, or other types of variety, judgments of probability of occurrence, stylistics, politeness, and a knowledge of the culturally, situationally, and textually specified rules of who can speak when, how to address other people, what topics are appropriate, what level of politeness is required, how to time conversations, how to end an encounter, how to regulate spatial proximity, touch, or direct gaze, etc.

In Bachman's model of communicative language ability, the component of sociolinguistic appropriateness captures several linguistic aspects of this criterion – focusing on varieties, registers, naturalness, and cultural meaning extensions. More broadly, appropriateness means the ability to relate to social and interpersonal norms or expectancies based on knowledge of these sociocultural factors. Usually this implies an avoidance of their violation; however, sometimes, overriding or altering existing rules may be appropriate communicative behavior, and some deviations from expectancies may have positive outcomes, as addressed by expectancy theory.

While neither effectiveness nor appropriateness are reducible to each other, they are also not independent from each other. On the one hand, the appropriateness of a communicative behavior might well depend on the communicative function performed; and on the other hand, the achievement of

a communicative goal might depend on the adherence to sociolinguistic rules. At times there may even be contradictions between both criteria, with one criterion constraining the possibilities of communicative behavior according to the other, and communicatively competent behavior being a function of a trade-off between both.

Adaptability

A fundamental aspect of communicative competence – related to both effectiveness and appropriateness – is adaptability, the ability to adapt to different conditions, functions, tasks, situations, communication partners, and contexts. It relates to appropriateness inasmuch as that is a function of social situation, communicative function, setting, and cultural context. It relates to effectiveness, as the means to achieve a communicative goal may depend on the interlocutors' abilities, interests, and other preconditions, and may need to be adapted to changing circumstances.

Mutual adaptation is one of the most pervasive phenomena in interactive communication (for reviews, see, e.g., Burgoon et al., 1995; Vorwerg, 2013). During interaction, people converge in their communicative behavior at many levels, including body posture and movements, aspects of linguistic content, gaze, proximity, voice features, pronunciation, linguistic elements and structures. Moreover, convergence in communicative behaviors is related to the quality of communication. It can also be assumed that the tendency to adapt to each other in communication contributes to the emergence of sociolinguistic patterns. Adaptation is also required to adapt in reception, in order to be able to understand communication partners who use communicative forms that do not belong to one's own active repertoire. More generally, the ability to adapt to a communication partner in production and reception depends on the variability of linguistic and other communicative means (Vorwerg, 2013), which may again be related to the dimension of expressiveness in Spitzberg and Cupach's model of interpersonal communication competence. The ability to choose aptly between different variants and to use them in novel, creative ways may add extra value to the quality of communication.

The Nature of the Concept of Communicative Competence

Most research treats communicative competence as a construct – an explanatory concept that has no direct observable referent but is indirectly observable in terms of manifestations or indicators (or, in other conceptions, is based on observables). While individual definitional proposals for communicative competence range at the extremes from treating the term as a theoretical concept whose meaning cannot be specified outside the context of specific theories (Wilson and Sabee, 2003) to describing it as the quality of interpersonal-skill enactment (Spitzberg, 2003); the most consistent and, arguably, the most fruitful conceptualization is to define communicative competence as the ability to communicate effectively and appropriately underlying actually observable behavior. This forms the basis for clarifying its relation to other concepts, such as performance, knowledge, and skill.

Communicative competence can thus be described as an individual's potential to perform communicative behaviors successfully in terms of goal accomplishment and acceptability. This capacity underlies (and also predicts) actual performance in communicative situations, which conversely can be regarded as the manifestation of this underlying competence. Accordingly, communicative competence comprises both the knowledge needed to engage in communication, and the ability to implement this knowledge. This knowledge includes – in addition to rules or regularities, factual knowledge, and individual memories – a range of skills, which can be considered as procedural knowledge, i.e., knowing how to, which in itself incorporates the ability for use. Thus, communicative competence is partly constituted by social, verbal, and other communicatively relevant skills, i.e., the capacities to perform complex social, verbal, or other patterns of behavior smoothly and adaptively. The specific use of the term 'skill' may vary between authors, from narrowing it to more elementary behaviors, such as asking a question or making eye contact, that may or may not be instrumental in a communicative context (see Spitzberg, 2003); to broader attempts to identify the key abilities involved in particular communication competencies in terms of skills, such as negotiation skills, arguing skills, or emotional support skills. However, the idea that a multitude of interrelated skills contributes to communicative competence seems to be generally agreed upon.

Communicative competence is related to other competence concepts. There is partial overlap with social, intersubjective, and interpersonal competence. Theory of Mind as a component of intersubjective competence, for example, is relevant in inferring a speaker's meaning. Sociolinguistic and metaphoric (or other figurative) competencies can be regarded as forming part of communicative competence, with pragmatic competence, as an intermediate level comprising both of them, in addition to knowledge employed in speech acts and cultural references. Finally, intercultural competence is a specific form of communicative competence required for communication across cultures.

Areas of Empirical Research

Communicative competence is a fruitful research field, as evidenced by the vast number of studies devoted to a variety of empirical questions. One area of research concerns its impact on different areas of an individual's life. Results show that communicative competence is related to the maintenance of personal relationships, including marriage, romance, and friendship; and to social embeddedness versus loneliness; and therefore also to health and happiness, which both depend on a person's social connections. Abilities in a range of communicative functions are also important for academic and occupational success.

Another area of research investigates the components of communicative competence – depending on settings and functions. This includes also research on the assessment and perceptions of communicative competence, the indicators of good communication, and the levels of particular communicative abilities present in certain populations. The relevance of communicative skills may vary between different settings,

such as private versus organizational settings, dyadic face-to-face situations, individual correspondence, social media, group communication, or public situations, symmetric versus asymmetric relations, classroom settings for language practice, health-care contexts, job interviews, etc. The involvement and importance of particular components, such as responsiveness, anticipation, means of remediation, immediacy, verbal clarity, listening skills, clear articulation, vocal variety, expressions of acceptance or concern, structural soundness of arguments, mind reading, etc., varies also between different communicative functions, such as persuasion, negotiation, arguing, instruction, emotional support, socializing, conversation, social problem solving, public speech, etc. Empirical questions in this context are also the generalizability and transferability of communicative skills acquired in a certain situation (such as the classroom setting), and the degree of consistency of communicative competence across communicative functions or settings, possibly related to communicative adaptability and behavioral flexibility (cf Spitzberg and Cupach, 1984) or very basic underlying abilities or traits.

Other research focuses on the development of communicative competence in childhood and adolescence, and beyond, as well as on the acquisition of communicative competence in a second language, and on interventional approaches. Developmental issues addressed include the emergence of communicative intentions, prelinguistic communication, the development of mind reading related abilities – such as intention reading, joint attention, understanding others' first- and second-order beliefs, and metarepresentations – and their relation to linguistic communication, the mastering of speech acts, the acquisition of linguistic variation, the understanding of figurative language, conversational abilities, the development of style and appropriacy judgments, the role of parenting skills, etc. (for introductory overviews, see Foster-Cohen, 1999; Hoff, 2009; Rickheit et al., 2008). Interventional approaches concern both training programs developed for specific communicative abilities and functions (such as arguing, negotiation, emotional support, public speaking, or marital communication), and diagnostic tools and therapeutic methods for the treatment of language and communication impairments (see, e.g., Greene and Burleson, 2003; Rickheit and Strohner, 2008).

A further area of research is focused around the processes involved in message production and message reception, which imply communicative competence. Research fields include the psycholinguistics of dialogue and mutual adaptation, cognitive pragmatics, social cognition, and psychological approaches to communication theory. Central questions of research are how interlocutors manage to coordinate their talk in interaction, what specific processes subserve the conveying of the intended meaning, conversely what is involved in inferring the speaker's meaning, also what mental representations and knowledge structures underlie competent communicative behaviors, and how nonverbal behaviors are generated and interpreted in face-to-face interaction (for a review, see Rickheit et al., 2008).

A novel research field is communicative competence in the electronic age, devoted to studying skills involved in communication in electronic environments, such as text messaging, mobile communication conversation, or social networking. As communication depends on the media used, the two fundamental issues involved are what constitutes media competence itself and how the available media influence communicative processes. A model of mobile communication competence proposed by Bakke (2010) and based on factor analyses includes six constructs: willingness to use, mobile preference, asynchronous communication, efficacy, affect, and appropriateness. Studies on the effects of mobile communication technology have shown that it has changed interpersonal coordination patterns, leading, among other things, to increased social face-to-face interaction; that it has brought about a growth in mobile work, rendering the term 'workplace' outdated for many; that it has changed relational communication, among other things, by adding a new layer of privacy through text messaging; and that it strengthens social network ties (see Campbell and Ling, 2008). The features, limits, and affordances of the media themselves (such as character length and screen sizes), as well as their social and communicative consequences (such as a lowered threshold for communicating, more permanent and closer connectedness, flexibility in communication, and microcoordination) have also lead to new forms of communicative means and conventions at all levels of communication (e.g., linguistic forms, nonverbal expressions of emotion, ways to perform speech acts, forms of address, turn-taking, rules of when to address whom about what), which concern both the efficiency and the appropriateness of communication.

See also: Communication Skills in Social Work; Critical Rationalism; Empiricism, History of; Intentionality During Infancy and Early Childhood, Development of; Intentionality in Language and Communication, Emergence of; Intercultural Communication: An Anthropological Perspective; Language Development: Emergentist Theories; Language and Society; Politeness and Language; Psycholinguistics: Overview; Social Intelligence and Competencies; Sociolinguistics; Theory of Mind During Infancy and Early Childhood Across Cultures, Development of.

Bibliography

Bachman, L.F., 1990. Fundamental Considerations in Language Testing. Oxford University Press, Oxford.

Bakke, E., 2010. A model and measure of mobile communication competence. Human Communication Research 36, 348–371.

Burgoon, J.K., Stern, L.A., Dillman, L., 1995. Interpersonal Adaptation. Dyadic Interaction Patterns. Cambridge University Press, Cambridge.

Campbell, S.W., Ling, R.S., 2008. Effects of mobile communication. In: Bryant, J., Oliver, M. (Eds.), Media Effects: Advances in Theory and Research, third ed. Lawrence Erlbaum, Mahwah, NJ, pp. 592–606.

Canale, M., Swain, M., 1980. Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics 1, 1–47.

Christiansen, M.H., Chater, N., 2008. Language as shaped by the brain. Behavioral & Brain Sciences 31, 489–509.

Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge MA.

Chomsky, N., 1980. Rules and Representations. Columbia University Press, New York.

Foster-Cohen, S., 1999. An Introduction to Child Language Development. Longman, London.

Greene, J.O., Burleson, B.R. (Eds.), 2003. Handbook of Communication and Social Interaction Skills. Lawrence Erlbaum, Mahwah, NJ.

Habermas, J., 1970. Toward a theory of communicative competence. In: Dreitzel, H.P. (Ed.), Patterns of Communicative Behavior. Macmillan (Recent Sociology No. 2), New York, pp. 115–148.

Hauser, M.D., Chomsky, N., Fitch, W.T., 2002. The faculty of language: what is it, who has it, and how did it evolve? Science 298, 1569–1579.

Hoff, E., 2009. Language Development, fourth ed. Wadsworth, Belmont, CA.

Hruschka, D.J., Christiansen, M.H., Richard, A., Blythe, R.A., Croft, W., Heggarty, P., Mufwene, S.S., Pierrehumbert, J.B., Poplack, S., 2009. Building social cognitive models of language change. Trends in Cognitive Sciences 13, 464–469.

Hymes, D., 1972. On communicative competence. In: Pride, J.B., Holmes, J. (Eds.), Sociolinguistics. Penguin, Harmondsworth, pp. 269–285.

Hymes, D., 1992. The concept of communicative competence revisited. In: Pütz, M. (Ed.), Thirty Years of Linguistic Evolution. John Benjamins, Philadelphia, pp. 32–57.

Knapp, M.L., Daly, J.A., 2011. Handbook of Interpersonal Communication, fourth ed. Sage, London.

O'Grady, W., 2008. The emergentist program. Lingua 118, 447–464.

Rickheit, G., Strohner, H., 2008. Handbook of communication competence. In: Knapp, K., Antos, G. (Eds.), Handbooks of Applied Linguistics, vol. 1. Mouton de Gruyter, New York.

Rickheit, G., Strohner, H., Vorwerg, C., 2008. The concept of communicative competence. In: Rickheit, G., Strohner, H. (Eds.), Handbook of Communication Competence, In: Knapp, K., Antos, G. (Eds.), Handbooks of Applied Linguistics, vol. 1. Mouton de Gruyter, New York, pp. 15–62.

Rubin, R.B., 1990. Communication competence. In: Phillips, G.M., Wood, J.T. (Eds.), Speech Communication: Essays to Commemorate the 75th Anniversary of the National Communication Association. Southern Illinois University Press, Carbondale, IL, pp. 94–129.

Savignon, S., 1972. Communicative Competence: An Experiment in Foreign Language Teaching. Center for Curriculum Development, Philadelphia.

Spitzberg, B.H., Cupach, W.R., 1984. Interpersonal Communication Competence. Sage, Beverly Hills, CA.

Spitzberg, B.H., 2003. Methods of interpersonal assessment. In: Greene, J.O., Burleson, B.R. (Eds.), Handbook of Communication and Social Interaction Skills. Lawrence Erlbaum, Mahwah, NJ, pp. 93–176.

Vorwerg, C., 2013. Language variation and mutual adaptation in interactive communication: putting together psycholinguistic and sociolinguistic perspectives. In: Wachsmuth, I., de Ruiter, J., Jaecks, P., Kopp, S. (Eds.), Alignment in Communication: Towards a New Theory of Communication, pp. 149–165. In: Dautenhahn, K., Cangelosi, A. (Eds.), Series Advances in Interaction Studies. Benjamins, Amsterdam, pp. 149–165.

Wiemann, J.M., 1977. Explication and test of a model of communicative competence. Human Communication Research 3, 195–213.

Wilson, S.R., Sabee, C.M., 2003. Explicating communicative competence as a theoretical term. In: Greene, J.O., Burleson, B. (Eds.), Handbook of Communication and Interaction Skills. Erlbaum, Mahwah, NJ, pp. 3–50.

Communism, History of

Jerzy Holzer, Polish Academy of Sciences, Polna, Warszawa, Poland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Following the different interpretations of the word communism, researchers concentrated their work on (1) the idea of social and political thought, (2) the political movements, or (3) state political systems. Research on communist political movements and systems of state has been seriously complicated due to changes in time and space since both have existed for almost 100 years and have manifested themselves in different zones of the world, most intensively, however, in Europe and East Asia. At present, communist parties exist only in a few countries and the communist systems exist exclusively in the Far East and in Cuba but with many changes to the previous form.

Introduction

The notion of communism has been used in three different ways. First, since antiquity communism has been referred to ideas relating to advanced forms of socialism advocating the need to organize societies as communities owning means of production and consumption. Second, the term communism signifies political movements with 'bolshevism', an extreme faction of Russian social democracy, as their protagonist since 1903. Third, communism denotes a political and socioeconomic system of state introduced for the first time in Russia in the year 1917.

Following the three different interpretations of the word communism, researchers concentrated their work on:

1. the idea of social and political thought,
2. political movements, or
3. state political systems with particular emphasis on dictatorship (totalitarian ideas), and economic systems, especially nonmarket economies, planned economies, and autarchies.

Research on communist ideology, political movements and systems of state has been complicated seriously due to changes in time and space since both have existed for almost 100 years and have manifested themselves in different cultural zones of the world, most intensively, however, in Europe and East Asia. Only in the last years of the Soviet Union were attempts made to approach the problems in all three spheres during the whole of the time and in all zones of the world. (Service, 2007; Priestland, 2009).

The Idea

The ideas of communism have surfaced since antiquity and have been present throughout the Middle Ages, the Renaissance, the Enlightenment, and the French Revolution. The proponents of communism resorted to moral or economic arguments advocating the need to organize societies as communities owning means of production and consumption. These ideas were adopted by some socialists in the nineteenth century. Marx and Engels published their '*Communist Manifesto*' (1848). To substantiate his ideas, Engels drew on the then studies on primitive societies as representative of classless communities (the natural state of society was disrupted by the division into social classes).

However, the notions of socialism and communism were sometimes used interchangeably in the nineteenth century. When social democracy developed in many countries, socialists hardly ever used the notion of communism, whereas anarchists resorted to it occasionally (Kropotkin's anarchocommunism).

Communist ideas have acquired a new meaning since 1918. They became equivalent to the ideas of Marxism–Leninism, that is, the interpretation of Marxism by Lenin and his successors. Endorsing the final objective, namely, the creation of a community owning means of production and providing each of its participants with consumption 'according to their needs', they put forward the recognition of the class struggle as a dominating principle of a social development.

In addition, workers (i.e., the proletariat) were to carry out the mission of reconstruction of the society. Conducting a socialist revolution headed by the avant-garde of the proletariat, that is, the party, was hailed to be a historical necessity. Moreover, the introduction of the proletariat dictatorship was advocated and hostile classes were to be liquidated.

Some researchers emphasized a quasireligious element of the communist (and earlier socialist) ideas (Muravczik, 2002). In the following decades, communist ideas have been subject to dogmatism, on the one hand, and, on the other, to current political pressures. Paramount importance was attributed to the 'classics', namely, to Marx, Engels, and Lenin, as well as Stalin during his rule. It was only the later 'classics' (i.e., Lenin and Stalin) who were believed to be entitled to correct and amend the views of the earlier 'classics'.

The communist ideology was expanded to include new components such as an exemplary role of the USSR and 'the principle of exacerbating the class struggle parallel to the building of socialism' (Stalin), which was called into question after 1953. In the years to follow, the ideas of a peaceful coexistence of socialism and capitalism as well as that of a peaceful transition to socialism were added.

Some differences have surfaced between communist ideas endorsed in the USSR and the Soviet bloc; in Yugoslavia, where social and economic aspects of self-government were regarded

as significant features of the communist ideology; and in China, where Maoism became a revolutionary ideology of the suppressed 'Third World' directed against capitalist metropolis (Harrison, 1972) as well as of the 'rural guerrilla revolution'. Some West European countries endorsed the ideology of Eurocommunism, that is, a transition to socialism, which complied with all rules of democracy and renounced the dictatorship of the proletariat.

On the whole, the ideology gradually lost its momentum in the countries governed by communists, particularly in the Soviet bloc. A much greater significance was attached to this ideology by groups opposing the Soviet-style communism. In the late 1960s, some milieus of 'the new left' in Europe and America formulated the ideas of communist spontaneity, which were in opposition to capitalism and bureaucratic communism.

Many researchers have treated communist ideas mostly as one of the interpretations of Marxism. Popper (1945) wrote about the division of Marxists into radical (communist) and moderate (social democratic) wings. Walicki (1995) claimed that Leninism and Stalinism 'were dominant forms of the Marxist ideology of the twentieth century'. In his work on Marxism, Kolakowski (1978) handled the ideas of Leninism as one of the interpretations of Marxism, yet in some areas, he saw them as a departure from Marx's original views. An even clearer division of communist ideas from Marxism was drawn by Furet (1999). Although he did not deny that these ideas had originated from Marxism, Furet pointed to other sources of origin such as Russian populist ideas.

After the collapse of communist rule in European countries, communist ideology lost its popularity. The communist party in Russia associated its ideological tradition of Marxism–Leninism with very intense national and sometimes religious Orthodox elements (March, 2002). In China, officially the ideas of Marx, Lenin, and Mao are maintained but they do not play any important role in the daily propaganda of the Communist Party. Political leaders are speaking sometimes about these ideas, but with an interpretation to adapt them to 'socialism with Chinese characteristics' and they have eliminated the formula of class struggle as an obstacle to China's economic development (Chan, 2003). In China as in Vietnam and Cuba, the communist ideas are deeply influenced by the national emotions, and in the Asiatic countries, by a reference to the Confucian tradition. Nationalism and to some degree racism are present in the self-reliance ideology of the so-called Juche in North Korea (Myers, 2010).

There are some intellectuals who support the thesis concerning the durability of communist ideas and their importance for the prospective form of human relations, an alternative to capitalism. As a key question, they are giving attention to the creation of an egalitarian society. As a rule, this rhetoric avoids an appreciation of communist ideas preached in the 'real socialism' countries.

The Political Movement

When in 1903 the Russian Social Democracy split into two groups, the majority faction was called 'Bolsheviks' and the minority 'Mensheviks'. The Bolsheviks in 1912 formed a separate party, an organization of active revolutionaries. Although the principle of democratic centralism was introduced, in practice, the system was centralist and devoid of any traces of democracy. Furthermore, violence was endorsed as a fundamental means of political activity. All these features could be accounted for by the situation in Russia, where the parliamentary system was in its infancy and despotism ruled unabated.

The Bolsheviks came to power in Russia in 1917, whereas in 1918, they transformed themselves into a communist party. Since 1918, communist parties also emerged in other countries, mostly as a result of a split within socialist parties. In 1919, the Communist International (Comintern) was set up and it soon became centralized and dependent, as a matter of fact, on the Russian (or later Soviet) party. The terms 'communism' and 'bolshevism' were still used interchangeably and 'bolshevization' came to denote the process of following suit and imitating the example of the Russian party.

Communist political movements manifested themselves most intensively in Europe and China. They kept changing their political profiles, adopting to the then current situation on the international arena (e.g., under the pressure of the 1929–33 crisis), but, above all, complying with interests of the Soviet Union. In the 1930s, this compliance was linked with Stalin assuming the role of the undisputed leader of the international communist movement. Opposition groups were excluded from the movement. One of the few groupings that remained active were the proponents of Leo Trotsky, a prominent party leader expelled from the Soviet Union.

However, in the late 1930s, when faced with the threat of fascism, communist parties started to endorse the slogans of parliamentary democracy and entered into cooperation with socialists and even nonsocialist democratic parties (the 'People's Front'). When the Soviet Union entered into an agreement with Germany in 1939, communist parties all over the world declared their neutrality toward both parties in World War II.

This policy was subject to a drastic change following the 1941 German invasion of the Soviet Union. Only then did communist parties come to regard the struggle against fascism as their priority and fostered the idea of forming the National Front. The idea, however, was rarely put into practice for several reasons including dependence of communists on the Soviet Union (although Comintern was dissolved in 1943) and their aspirations to play a dominant role in the resistance movement.

When World War II was drawing to an end, communist parties took over power in the countries in which the Soviet army was present, as well as in Yugoslavia and Albania. In the countries occupied by the armies of the Western allies, communists frequently entered into ruling coalitions.

With the escalation of conflict between two political blocs and an upsurge of the 'Cold War', communist parties in the West went into opposition. The Soviet control over the international communist movement was strengthened anew in 1947 when the Information Bureau of Communist and Workers' Parties (Cominform) was established comprising the largest European parties. A year later, the first serious conflict between parties came to the fore, when the Yugoslav party was condemned by Cominform.

The victory of the communist party in China in 1949 (and later in Vietnam and Cuba) laid foundations for subsequent decentralization of the communist movement. The process of decentralization took place gradually following the death of Stalin in 1953, who 3 years later was accused of great atrocities. The escalation of the crisis in 1956 brought about the dissolution of Cominform, and resulted in reforms in Poland and a revolution in Hungary. The 1960s witnessed a conflict between the Soviet and the Chinese parties during which some other parties attempted to take advantage of the situation and gain independence.

Yet another crisis, triggered by the intervention of the Soviet bloc in Czechoslovakia, resulted in the adoption of Eurocommunist ideas by some West European parties in the 1970s. At the same time, small communist groupings emerged in the West drawing on the experiences (or rather their own visions of the experiences) of China, Vietnam, Cuba, or Yugoslavia. Further crises in the international communist movement were linked to the Polish conflicts of 1980–81. In the 1980s, both ruling and opposition communist parties showed symptoms of decay.

In the countries ruled by communists, a considerable number of works on the communist movement was published, yet their content was subject to control. The most important publications had even an official character (*The History of the Communist Party* 1938). Little changed in this respect in the communist historiography after the 'de-Stalinization', although the party ceased to stress the official character of publications, relying heavily on the censorship and monitoring of publications. These publications rarely contained reliable information, featuring mainly interpretations devoid of any critical approach and containing hagiographic accents.

Many publications were written by representatives of the communist opposition, who embraced the fundamentals of communism yet criticized the political line of Stalin and his successors, as it was in the case of the Trotskyists. Some of these texts resembled political commentaries and contained different information not included in the official and semiofficial communist literature. This trend also comprised publications based on extensive studies (e.g., Deutscher, 1949, 1954, 1959, 1963). Works by noncommunist authors fall into yet another category of literature relating to the history of the working-class movement. Although this literature contained fewer political commentaries, it demonstrated a strong socialist bias. From a left-wing socialist perspective, communism, in spite of all its shortcomings, was perceived as a proletariat movement close to socialist ideas and as a potential ally in the socialists' struggle against the bourgeoisie. Right-wing socialists, on the other hand, saw it as a hostile force striving to exert influence over the same social class.

In recent decades, however, the history of the workers' movement encompassing the history of the communist movement has become subject to more academic research.

Among authors who dealt with the communist movement, there were some who came from within the party ranks as well as those who broke free from its legacy. Others were emigrants from countries governed by communists, who drew on their life experience and language skills. One should also consider here expert publications of academic institutions that dealt with communism as a sideline political movement within a larger scope of interests relating to the so-called Sovietology.

After the collapse of communist systems in Europe, nearly all communist parties rejected their former ideological and organizational principles and transformed into socialist or social democratic parties. The existing European communist parties are rather weak, with some exceptions (in Russia, where the communists are the strongest opposition party, and also in Czech Republic). In the non-European countries with still ruling communist parties (China, Vietnam, North Korea, Cuba), they have a central role in the political system. In China, the communist party changed its social reference to workers and peasants and decided to open the party's ranks for the 'new classes' including the representatives of big business (Brady, 2008).

The Political and Socioeconomic System

Communists themselves are more prone to call it a socialist system. For the first time, such a system was introduced in Russia after 1917. Its principles were outlined by Lenin before the victory of the revolution, yet his work on this subject was published only when the Bolsheviks came to power (*Gosudarstwo i rewoljucija* 1918). Two principles were of particular significance: (1) the state has to resort to violence in the political and social life and (2) elements of a centralized bureaucracy, that is, record keeping and monitoring, play a key role in the organization of the economy.

In the Soviet political system, the leading role was assigned to the centralized communist monoparty devoid of any traces of democracy within its ranks. All organizations were subordinated to the party and headed by its members. Moreover, the party exercised control over all publications and mass media directly or through state departments (e.g., censorship). The state authorities at all levels reported to party organs. Initially, the state authorities were organized in a hierarchical structure of councils with some chosen representatives of lower level sitting on the councils of a higher level. Following the establishment of the Soviet Union in 1922, the party and state authorities were to function according to the principles of a federal state. The 1936 constitution provided for general elections at all tiers of power. In practice, however, soon after communists came to power, they introduced the principle of nominating from above, a principle that applied to both the party and state authorities. The approval of such nominations in public elections was a mere formality. This political system was adopted after World War II in all communist countries.

A considerable autonomy was enjoyed by the police, and above all by security services. They gained a strong position in the period of widespread repression in the Soviet Union (collectivization and 'big cleansing' in the 1930s) as well as in the countries within the Soviet sphere of influence (in the late 1940s and early 1950s) and in China (during the cultural revolution). Furthermore, security services attempted to exert a still greater political influence also in other periods.

Another pivotal element of the system was the army. In the Soviet Union, it owed its position to the role played in introducing the system and in World War II. Later, it was a key means supporting the rivalry between superpowers and

political blocs as well as the Soviet expansion in the countries of the Third World. Similarly, in China and Vietnam, the army was the cocreator of the system. In Poland, the army won a particular importance in the face of the imposition of martial law in 1981. On the other hand, party officials feared an excessive degree of independence of the army, which had far-reaching repercussions in the Soviet Union during repressions directed toward army officials in the 1930s.

In the Soviet bloc, which included European communist countries (with the exception of Yugoslavia since 1948 and Albania since the 1960s) as well as Mongolia and North Korea (for a couple of years), there were initially informal methods of control from Moscow. System monitoring was exercised through the Cominform and Soviet advisers or directives sent to high-ranking officials. Since 1956, methods of control were restricted and formalized by means of political and military structures of the Warsaw Pact, as well as structures of the Council for Mutual Economic Assistance (Comecon).

Gradually, the centralized systems began to crumble on international and state levels alike. Retaining its leading role, the USSR had to take into account more and more frequently the positions of the countries within the Soviet bloc. Several countries of the Soviet Block and China managed to hand down some power to regional levels. Yugoslavia carried out some transformations in its political system, introducing elements of a limited federalism and a local government.

Since 1953, several communist countries, particularly those belonging to the Soviet bloc, witnessed recurring political crises. They were triggered by people's rising democratic aspirations influenced by examples emanating from Western democracies, by the lack of respect for national sovereignty and, to a large extent, by a growing economic crisis and a low standard of living. Another precipitating factor was the disillusionment as to the application of communist ideas in practice.

Following the crisis in Czechoslovakia and the invasion of the country by the Soviet bloc, the Brezhnev doctrine was proclaimed (1968) asserting the right of the communist community to intervene when any of its members demonstrates 'antisocialist degeneration'. Communist officials adopted the formula of 'real socialism' as a form of the existing system hindered by historical limitations on the road to the socialist and communist future.

The economy in communist countries was subordinated to politics. Nationalization was implemented in all spheres of economic life. Although agriculture and, to some degree, trade and small production were formally owned by cooperatives, they were supervised by the party and state administration. There were only negligible relics of private ownership with the exception of Poland and Yugoslavia where agriculture remained in private hands.

Centralized state planning functioned on the basis of 5-year plans, the allocation of raw materials and means of investments, regulated prices, and rationing commodities for sale. A top priority was the needs of the state, which resulted often from an international policy. Particular emphasis was placed on the development of heavy industry and reservoirs of raw materials, neglecting other branches of production. In the face of constant shortages and a crippled economy, many goods intended for production and consumption were subject to different forms of rationing. The state's monopoly on foreign trade along with slow technological progress, a limited exports potential and, consequently, import deficit led to a form of economic autarchy.

Yugoslavia's economy demonstrated a slight divergence from this general pattern after the introduction of self-government in enterprises. Moreover, Yugoslavia fostered more extensively mechanisms of the market economy and managed to avoid autarchy. Nonetheless, it put a limit on foreign investments and failed to develop technological innovations.

Communist systems that were created almost exclusively in poorly developed or underdeveloped countries demonstrated initially a high rate of economic growth. The rapid industrialization combined with urban development was based on extensive factors such as a cheap labor force. However, as these factors were gradually exhausted, the growth rate began to decline and the standard of living stabilized at a low level. Among other factors that aggravated the crisis were the growing costs of the arms race and political rivalry between superpowers and blocs.

From the late 1950s, attempts were made to overcome the crisis. On the one hand, restraints of the arms race were accepted (the Helsinki Agreement of 1975, treaties on nonproliferation of nuclear and conventional weapons) and, on the other hand, planned economies were transformed slightly by introducing in them elements of a market economy and stimuli of a technological progress. In spite of a successful arms race reduction, the costs of armaments and political expansion still exceeded the economic potential of communist countries, whereas the introduction of elements of market economy generated strong resistance of the bureaucracy in the state and the party administration.

The 1980s saw not only a growing opposition against the political system but also an ever-stronger conviction of an upcoming defeat shared by communist officials. The need to conduct economic reforms called for a more extensive popular support, which was possible to achieve thanks to political reforms. Political reforms were pursued to a limited extent in the Soviet Union (economic *perestroika* accompanied by political *glasnost* – transparency) and in Poland (the 'roundtable' talks, an attempt to achieve a compromise with the opposition). Finally, the attempts to carry out limited reforms of the economic system were doomed to failure. Coupled with a disintegrated political system, they led to the collapse of communism over the years 1989–91.

In China, however, the authorities resorted to extreme means of repression against the opposition (the massacre at the Tian'anmen square, 1989). They were then successful in effecting its economic reforms encompassing devolution; privatization of agriculture, small- and medium-size industry, and trade; as well as the introduction of elements of market economy. The country managed to restore a fast rate of economic growth and development creating openings for foreign investments linked with the import of new technologies.

Communist researchers into the political and economic system concentrated their work mostly on its premises rather than on reality. A distinct character was noticeable only in the discussions before the 1930s, which touched upon the 'Asiatic

system of production'. These works tackled directly the prospects of building of the communist system in China, and, indirectly, they reflected some ideas on the political and economic system in Russia. Among basic characteristics of oriental societies, they featured a dominating role of the state in the economy and, consequently, the durability of absolutist and bureaucratic tendencies. This concept was raised in discussions in the 1920s by Karl Wittfogel. He broke away later from communism and came to regard the Soviet Union as a follower of Asian despotic systems.

The communist authors emphasized the class conditioning of the system. Many works presented a mystified picture of common interests of the state and society. They referred also to the notions of 'People's Democracy' or 'Socialist Democracy' (as a state of equal opportunities). Communist publications are of little academic value, yet sometimes they can communicate significant information.

In the noncommunist literature as a whole, there were prominent publications that dealt with communist systems within a wider historical framework. In particular, they referred to the connections between the history of Russia and of the Soviet Union. The principle of a historical continuity was propagated (Kucharzewski, 1948; Pipes, 1993; Besançon, 1980). This conception was often accompanied by a tendency to view communist systems as something superimposed by the USSR upon other countries.

The notion of totalitarianism was used initially in Italian fascism, later by researchers whose works dealt with German national socialism. In the wake of experiences and the atmosphere of the 'Cold War', communist systems were regarded by some researchers as one of the forms of totalitarianism (Arendt, 1951; Friedrich and Brzezinski, 1956). Arendt saw the roots of totalitarian systems above all in the current demise of human values and the crisis of principles underlying the existence of societies and individuals. Friedrich and Brzezinski focused mainly on the techniques of wielding power and the totalitarian syndrome.

Concrete studies of the system were heavily influenced by the concept of totalitarianism. The concept of communism as a form of totalitarian systems was undermined to some extent as a result of growing tendencies toward decentralization and curbing repression. Researchers started to concentrate more on the differences between communist countries as well as on the fact of how inconsistent their respective systems were (Brzezinski, 1960). Another approach to communism stipulated that the system was an attempt to implement utopia invented by doctrinaires (Heller and Niekricz, 1986; Malia, 1994). There were also other interpretations of the communist system according to which it was a peculiar way to modernization concerning in particular the Third World (Schurmann, 1966) or to a failed modernization in the Eastern part of Europe (Berend, 1996).

Some authors treated communism above all as a system of genocide. A different interpretation claimed that Stalinism was only a criminal distortion of the communist system. Political commentaries on the subject emerged before World War II. Afterward, scientific studies were preceded by documentary literature (Herling-Grudzinski, 1951; Solzhenitsyn, 1974). In the 1960s, Robert Conquest published monographs on various aspects of genocide in the Soviet Union. An international team has tried to apply a comprehensive approach to this problem (Courtois et al., 1997). The question of genocide pointed to a comparison between communism and Nazism and sometimes to the claiming of a communist priority, based on the timing and dimension of crimes.

After the collapse of communist rule in European countries and opening of archives, many books and articles about different questions related to the communist systems were published. They explained particular mechanisms and instruments of communist rule, as well as the different aspects of crises in the communist countries and causes for collapse of their regimes. One emphasized the moral, ideological, social, and political delegitimization of European communism (Okey, 2004).

At present, communist systems exist exclusively in the Far East and in Cuba. All of them introduced elements of a market economy, in a more consistent way and as an example to other Asiatic countries. These included not only China (Gittings, 2005; Guo, 2006) but also Vietnam, and within a limited scope Korea and Cuba. In all these countries, the communist party exercises an unlimited power, but in China and Vietnam, with much more respect for the needs and feelings of their population, and in the Democratic People's Republic of Korea (North Korea), with an extremely repressive system. China and Vietnam are rather open in their relations with the outside world, whereas the Korean system makes every endeavor to prevent its citizens from contacts with foreign lands.

Another difference concerns the form of dictatorship. All existing communist countries perceive the danger of an individual rule without the principle of succession. These changes have occurred in a contradictory way. China and Vietnam opted to a collective leadership. Korea and Cuba introduced a succession in the ruling family. Korea implemented a quasi-monarchical order of succession from father to son, but to a younger son instead of to the eldest.

See also: Asia, Sociocultural Aspects: China; Cold War, The; Dictatorship in History; Dictatorship; Eastern European Studies: History; First World War, The; Genocide, Historical Aspects of; Japan: Sociocultural Aspects; Marx, Karl (1818–83); Political Parties: Communist; Revolutions of 1989–90 in Eastern Central Europe; Russian Revolution, The; Russian and Post-Soviet Studies: Politics; Russian and Post-Soviet Studies: Society; Socialism: Historical Aspects; Socialism; Socialist Societies: Anthropological Aspects; Totalitarianism.

Bibliography

Arendt, H., 1951. The Origins of Totalitarianism, first ed. Harcourt Brace, New York.
Berend, I.T., 1996. Central and Eastern Europe 1944–1993. Detour from the Periphery to the Periphery. Cambridge University Press, Cambridge.
Besançon, A., 1980. Present sovietique et passe_russe. Librairie. Generale Francaise, Paris.
Brady, A.M., 2008. Marketing Dictatorship. Propaganda and Thought; Work in Contemporary China. Rowman and Littlefields Publishers, Lanham.
Brzezinski, Z., 1960. The Soviet Block – Unity and Conflict. Harvard University Press, Cambridge, MA.
Chan, A., 2003. Chinese Marxism. Continuum, London-New York.

Courtois, S., Werth, N., Panne, J.-L., Paczkowski, A., Bartosek, K., Margolin, J.-L., 1997. Le livre noir du communisme. Crimes, terreur et repression. Editions. Robert Laffont, Paris.
Daniels, R.F., 2007. The Rise and Fall of Communism in Russia. Yale University Press, New Haven-London.
Deutscher, J., 1949. Stalin. A Political Biography. Oxford University Press, Oxford.
Deutscher, J., 1954. The Prophet Armed: Trotsky 1879–1921. Oxford University Press, Oxford.
Deutscher, J., 1959. The Prophet Unarmed: Trotsky 1921–1929. Oxford University Press, Oxford.
Deutscher, J., 1963. The Prophet Outcast: Trotsky 1929–1940. Oxford University Press, Oxford.
Friedrich, C., Brzezinski, Z.K., 1956. Totalitarian Dictatorship and Autocracy. Harvard University Press, Cambridge, MA.
Furet, F., 1999. The Passing of an Illusion. The Idea of Communism in the Twentieth Century. Chicago University Press, Chicago.
Gittings, J., 2005. The Changing Face of China. From Mao to Market. Oxford University Press, Oxford.
Guo, S., 2006. The Political Economy of Asian Transition from Communism. Ashgate Publishing, Hampshire.
Harrison, J.P., 1972. The Long March to Power. A History of the Chinese Communist Party, 1921–1972. Praeger, New York.
Heller, M., Niekricz, A., 1986. Utopia in Power. The History of the Soviet; Union from 1917 to the Present. Summit Books, New York.
Herling-Grudzinski, G., 1951. A World Apart. Heinemann, London.
Kolakowski, L., 1978. Main Currents of Marxism. Its Rise, Growth and Dissolution, vols. 1–3. Clarendon Press, Oxford.
Kucharzewski, J., 1948. The Origins of Modern Russia. Polish. Institute of Art and Sciences, New York.
Malia, M., 1994. The Soviet Tragedy. A History of Socialism in Russia 1917–1991. Free Press, New York.
March, L., 2002. The Communist Party in Post-Soviet Russia. Manchester University Press, Manchester.
Mc Hale, S.F., 2004. Print and Power. Confucianism, Communism, and Buddhism in the Making of Modern Vietnam. University of Hawaii Press, Honolulu.
Muravczik, J., 2002. Heaven on Earth. The Rise and Fall of Socialism. Encounter Books, San Francisco, CA.
Myers, B.R., 2010. The Cleanest Race. How North Koreans See Themselves and Why It Matters. Melville House Publishing, Holboken, NJ.
Okey, R., 2004. The Demise of Communism East Europe. 1989 in Context. Hodder Arnold, London.
Pipes, R., 1993. Russia under the Bolshevik Regime, first ed. Alfred A. Knopf, New York.
Popper, K.R., 1945. The Open Society and Its Enemies. Routledge and Sons, London.
Priestland, D., 2009. The Red Flag. A History of Communism. Grove Press, New York.
Schurmann, F., 1966. Ideology and Organization in Communist China. University of California Press, Berkeley, CA.
Service, R., 2007. Comrades! A World History of Communism. Macmillan, Basingstoke.
Solzhenitsyn, A.I., 1974. The Gulag Archipelago, 1918–1956: An Experiment in Literacy Investigation. Harper & Row, New York.
Walicki, A., 1995. Marxism and the Leap to the Kingdom of Freedom. The Rise and Fall of the Communist Utopia. Stanford University Press, Stanford, CA.

Communitarianism

Walter Reese-Schäfer, University of Goettingen, Goettingen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of communitarianism is developed by presenting the key elements and key theoreticians. The elements are a hermeneutic methodology, criticism of individualism, a concept of the good, the idea of community, and civic virtues. Among the key theoreticians, Charles Taylor, Michael Sandel, and Alasdair MacIntyre respond critically to the liberalism in the theory of John Rawls, emphasizing that it cannot explain loyalties to communities and the motivation to act morally. Michael Walzer relates common values and a shared idea of the just to common opinions in different Spheres of Justice that is directed against Rawlsian liberalism. Philip Selznick and Amitai Etzioni overcome the liberal-communitarian debate in pointing out that the development of common values presupposes freely consenting individuals. Robert Bellah points out the importance of civic engagement and value orientation for a Good Society. Etzioni criticizes on the one hand the negligence of the social context of individual integrity in liberal theory. On the other hand, he insists on internal democracy in communities as a criterion for the evaluation of value systems in order to avoid relativism and oppressive values. The concluding prediction is drawn that communitarian ideas will play a stable role in the self-reflection of modernizing societies.

Characteristic key elements of communitarian theory are the following five points: a hermeneutic methodology, criticism of what was called atomistic individualism, an Aristotelian concept of the good in opposition to rights, the renaissance of the idea of community, and a renewed insight into the relevant political function of civic virtues. Communities can be defined as webs of social relations that encompass shared meanings and shared values. Families may qualify as minicommunities, villages often are, although not necessarily. Communities need not be concentrated geographically, like the scientific community. Especially traditional communities are not necessarily places of virtue, but they can be oppressive, monolithic, and authoritarian. So a main problem for communitarian authors is to find overarching criteria to avoid accepting automatically any commonality of shared self-understanding as a good community (Etzioni, 1998, p. XIV).

In the 1980s, a group of political theoreticians – Charles Taylor, Michael Sandel, and Michael Walzer – challenged what they perceived as excessive individualism by the insight that individuality is based and dependent on social context. The radical liberal idea of an 'unencumbered self' was criticized as self-refuting, because it fostered antisocial behavior with the tendency to undermine its own social preconditions.

The academic and public impact of communitarianism has come about for two reasons. It seemed to give a theoretical answer to the most important political philosophy of the time, for example, the individualist liberalism of the renewed contract theory of John Rawls and others. The second reason is even more important: politically, communitarianism was a critical reaction against the 'liberalism of greed' of the Thatcher and Reagan years, so it became attractive to the New Democrats and New Labor as a rhetoric to win back the votes of the middle classes. In the early 1990s, Etzioni and others decided to take these ideas from the campus to the larger society. They founded the journal *The Responsive Community* and tried to establish communitarian ideas as a recognized third way of political thinking between liberalism and conservatism. Different from former communitarian ideas in the tradition of Ferdinand Tönnies (*Gemeinschaft und Gesellschaft*), Robert Nisbet, and Martin Buber, the new communitarians pointed out that a responsive community tries to avoid any authoritarianism and oppressiveness against the individual. The old communitarians privileged social bonds and traditional small communities over the individual. Communities in their definition were villages, small cities, some religious sects, and tribes. Asian communitarianism overemphasizes social harmony. The new American line of communitarian thinking concentrated much more on the question of balance between individual rights and social responsibilities. Individual autonomy and the common good are to be brought into a new equilibrium. It tried to add a sense of the social and communal preconditions of social integration to rights-centered liberalism and liberal neutrality.

The anthropologic key concept of modern communitarian thinking is Taylor's idea of humans as self-interpreting animals (Taylor, 1985b). Modern identity is based on inescapable frameworks of interpretation, which provides the background language for strong evaluations of right and wrong, good and evil, higher and lower. The quality of this framework is decisive for the quality of the society. Sandel radicalized this idea and transformed it into an argument against Rawlsian liberalism: the original position as a metaphor for justice, where nobody knows about his or her later social position before a decision is taken, is not only practically impossible (Rawls would agree), but also theoretically undesirable, because this conception of justice would lead to a world inhabited by morally neutral people without any character. Persons without loyalties to their family, their community, nation, or people are in Sandel's view persons without character. MacIntyre emphasized the missing foundation of liberal theories of justice. Even more, modern concepts cannot provide any motivation to act morally, because they are not based on fully responsible persons, but only on character masks such as the manager, the therapist, or the Kierkegaardian aesthete.

This anti-Rawlsian criticism was not conclusive, because his principles of justice were not meant to shape the whole person, but only their behavior in the political realm. Even more, Sandel did not take into account the possibility of a person

taking a distant stand from their community. So the foundations of Sandel's (and, as traditional virtues are concerned, MacIntyre's) criticism seemed to be rooted in old-fashioned communitarianism without the reflection of the fact that modern American society can be interpreted as the social basis for some kind of responsive individualism. Both Sandel and MacIntyre were uncomfortable with the label 'communitarian,' but their ideas belong to the present discussion of the concept.

A quite different kind of communitarian theory was developed by Walzer, who is also uncomfortable with the term, but refers to the same set of methods and ideas. In his *Spheres of Justice. A Defense of Pluralism and Equality* (1983) his social criticism and his ideas for reform lean on the hermeneutic interpretation of commonly shared values. The well-known tension between shared values and social practice is sufficient to legitimize social reform. Walzer's method differs on the one side from liberal Kantian or Rawlsian social constructionism, and on the other from conservative belief in the objectivity of traditional values (social ontology). Justice is only possible if it takes into account the differentiation of spheres. Different modes of distribution have to be applied concerning membership and citizenship, political power, security and welfare, money, office, education, recognition, kinship, and love. There is no general theory of justice, but a special justice for each of these spheres. What is considered as just depends on commonly shared and publicly discussed opinions about the criteria of justice.

The liberal-communitarian debate of the 1980s and early 1990s seems to be over at present. Those who decided to stick to the label 'communitarians' now call themselves liberal communitarians or communitarian liberals, such as Selznick (1992) and Etzioni (1996). The central idea is that, in a modern society, common values and virtues can only be developed by freely consenting individuals, not by the state or by pressure from religious movements of the fundamentalist denomination.

Selznick's 'communitarian liberalism' draws on the social philosophy of John Dewey. Morality is always community based. Character and integrity are relevant for the function of large organizations, including the state. A proper understanding of communities implies diversity and pluralism as well as social integration. It has to have a strong sense of a common good. Thus, Selznick combines sociological functionalism with the value-based traditions of American pragmatism. His interpretation of the term 'covenant' means a self-defining commitment, and an act of faith and resolve by the individual. It is prepolitical, foundational, and consensual. Liberal neutrality misses this aspect, and is based only on institutions for the purpose of social integration. Conservatism, on the other hand, tries to overload institutional processes with value orientations. The communitarian third way tries to restore values and virtues as functional prerequisites of the civil society.

Bellah et al. (1991) try to describe the state of American society and to make statements about its long-term viability through insights into its mores and the climate of opinion. They lean on Alexis de Tocqueville's *Democracy in America* (1835), which seems to be a key text for most communitarians. The results of the study are communicated in terms of eloquent sadness: the self-destruction of the liberation promises of modernity and the retreat from public engagement into the private sphere of the family and friends is stated by many of the interviews. Most of them articulated a profound yearning for the idealized small town, which can be interpreted as a yearning for meaning and coherence, for commitment, and community. The social ecology is not intact if there are no communal answers to the extreme fragmentation of modern everyday life. The ubiquitous language of individualism, in which American self-understanding normally is articulated, limits the ways in which people think and cuts off the opportunity of a broader orientation (once called 'civic virtues,' Bellah et al., 1985, p. 290). Because of their option for more substantive ethical identities and more active political participation Bellah and coworkers were labeled as communitarians. They preferred to describe their position in a slightly different way to avoid the risk of being misunderstood, privileging only small face-to-face groups such as families, congregations, and neighborhoods, and of being opposed to the state and all the other larger structures that dominate life at the start of the twenty-first century. They insist that freedoms and a Good Society depend on institutions, and they are critical that many individualistic Americans have trouble understanding and accepting social and political institutions (Bellah et al., 1991).

Etzioni's main criticism of liberal political theory is the negligence of the social preconditions that enable individuals to maintain their psychological integrity, civility, and ability to reason. On the other hand, he distances himself from those communitarians who hold that as long as a community has agreed on a set of values, these are the ultimate criteria as to what is morally appropriate (Etzioni, 1995, p. 17). The central problem is to establish criteria for the evaluation of the internal value system of different communities. The question is: how to avoid perspectivism and relativism, and to develop criteria to rule out authoritarian and oppressive community values. In the first place, internal democracy of the community is the criterion. But additional criteria are needed, because a majority in a specific community might violate individual and minority rights. The second criterion is a substantial process of consensus building within a community. Beyond any particular community, the values of a society as a whole are a framework within a higher order of legitimacy; that is, the application of the Constitution, especially the Bill of Rights. The third level is the cross-societal moral dialogues, in which the societal values have to be legitimized against others. The fourth criterion is the global community. But even the global community might act oppressively, as it did when women were excluded from voting worldwide until the end of World War I. So the last resort for the dissenting individual is the clearly communicable self-evidence of some core values.

All these criteria have in common that they are based on moral dialogues of different reach, not on ontological values or on artificially constructed value systems. The communitarian approach is the support of concrete societal and, if necessary, cross-cultural moral dialogues. Of course, they have to be limited to a set of core values (Etzioni, 1996, p. 237). Through these reflections Etzioni tries to overcome the restriction of traditional communitarian thinking to limited communities and to bring into discussion the notion of a global community. These are questions the worldwide discussion of communitarian ideas has made inevitable. Communitarian approaches

have to solve the question: Which community? What kind of community? Who belongs and who does not? This will be important especially for further discussions on the problem of multiculturalism (cf Taylor et al., 1992).

The discomfort with the label 'communitarian' seems to be a general characteristic of many authors in this field, with the exception of Etzioni. Communitarianism seems to be a permanent, accompanying phenomenon to liberalism, but can hardly stand on its own. The traditional social structures of community orientation have been dissolved by the processes of modernization, individualization, and fragmentation. Attempts to restore or invent dense forms of social integration as an answer to modernity such as communism, fascism (*Volksgemeinschaft*), or some forms of Islamic fundamentalism are discredited because of their atrocities and their contempt for individual rights. Thus, in modern, technology-based service economies, communitarianism is only acceptable as a reflexive admonition to care more about the common good, but not as a coherent and mobilizing social ideology. So the prediction follows that communitarian ideas in a moderate form of social self-reflection and social integration programs will be a permanent subdominant melody in the chorus of further modernizing societies, because it meets a fundamental human demand for social integration on a theoretical level.

The latest developments in communitarian political theory are marked by the publication of Michael Sandel's assessment of the moral limits of markets (Sandel, 2012), a criticism of the growing dominance of economic rationality over the other spheres of society. The core argument is in line with the communitarian tradition: Principles of Justice have to be developed from the respective pluralist contexts, not from only one narrow concept of rationality. Charles Taylor (2007) has presented a comprehensive diagnosis of our 'Secular Age' and of the conditions of religion and belief after secularization. Taylor's approach to the secular age offers a new turn in the sociology of religion. His concept of secularization is based on the idea that the United States as a whole society is secularized. It is no longer a society where it was virtually impossible not to believe in God to a society, where faith is one human possibility among others. His communitarian perspective concentrates on the context, the societal background of possible religious experiences. In this new form, the secularization thesis can resist challenges from authors like José Casanova, who argue that secularization only happened in Europe, but not in the United States and in Latin America.

See also: Civil Society and the Public Sphere: History of the Concept; Communitarianism, Sociology of; Community Sociology; Community and Society: History of the Concepts; Individualism versus Collectivism: Philosophical Aspects; Individuals in Society: History of the Concept; Liberalism: Historical Aspects; Liberalism: Impact on Social Science; Mass Society: History of the Concept; Tocqueville, Alexis de (1805–59).

Bibliography

Bellah, R.N., et al., 1985. Habits of the Heart. Individualism and Commitment in American Life, first perennial lib. ed. Harper & Row, New York.

Bellah, R.N., et al., 1991. The Good Society, first ed. Alfred Knopf, New York.

Etzioni, A., 1993. The Spirit of Community. Rights, Responsibilities, and the Communitarian Agenda, first ed. Crown, New York.

Etzioni, A. (Ed.), 1995. New Communitarian Thinking. Persons, Virtues, Institutions, and Communities. University Press of Virginia, Charlottesville, VA.

Etzioni, A., 1996. The New Golden Rule. Community and Morality in a Democratic Society, first ed. Basic Books, New York.

Etzioni, A. (Ed.), 1998. The Essential Communitarian Reader. Rowman & Littlefield, Oxford, UK.

MacIntyre, A., 1981/1984. After Virtue. A Study in Moral Theory. University of Notre Dame Press, Notre, Dame, IN.

Reese-Schäfer, W., 1994. Was ist Kommunitarismus? Campus Verlag, Frankfurt.

Sandel, M.J., 1982. Liberalism and the Limits of Justice. Cambridge University Press, Cambridge, MA.

Selznick, P., 1992. The Moral Commonwealth. Social Theory and the Promise of Community. University of California Press, Berkeley, CA.

Sandel, Michael, 2012. What Money Can't Buy: The Moral Limits of Markets. Straus and Giroux, Farrar.

Sandel, Michael, 2009. Justice. What's the Right Thing to Do? Farrar, Straus and Giroux, New York.

Taylor, C., 1985a. Atomism. In: Taylor, C. (Ed.), Philosophy and the Human Sciences, Philosophical Papers 2. Cambridge University Press, Cambridge, MA, pp. 187–211.

Taylor, C., 1985b. Self-interpreting animals. In: Taylor, C. (Ed.), Human Agency and Language. Philosophical Papers 1. Cambridge University Press, Cambridge, UK, pp. 45–76.

Taylor, C., 1989. Sources of the Self. The Making of the Modern Identity. Harvard University Press, Cambridge, MA.

Taylor, C., 2007. A Secular Age. Harvard University Press, Cambridge, MA.

Taylor, C., Gutmann, A., Rockefeller, S.C., Walzer, M., Wolf, S., 1992. Multiculturalism and the Politics of Recognition. Princeton University Press, Princeton, NJ.

Walzer, M., 1983. Spheres of Justice. A Defense of Pluralism and Equality. Basic Books, New York.

Walzer, M., 1990. The communitarian critique of liberalism. Political Theory 18 (1), 6–23.

Communitarianism, Sociology of

Henry Tam, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is based on new materials, following on from the previous article by A. Etzioni, volume 4, pp. 2336–2339, © 2001, Elsevier Ltd.

Abstract

'Communitarianism' is the name adopted by a group of American and British thinkers in late twentieth century to describe their shared sociological and philosophical concerns, and the ideas developed by them and their intellectual precursors in earlier times to address these concerns. One of its key themes is that society would be ultimately deficient and unstable, unless people learn to cultivate shared values and pursue them in cooperation with each other. Its significance is best understood as an evolving social philosophy that responds critically to threats to the development of inclusive community life.

What Is Communitarianism?

In the 1840s, the term 'communitarian' came into usage to describe the approach to community building through cooperative education and organization, championed by the British reformist, Robert Owen, and his followers on both sides of the Atlantic. In the 1980s, the same term was used by academic commentators as a generic label for an otherwise diverse group of thinkers (Alasdair MacIntyre, Charles Taylor, Michael Sandel, and Michael Walzer) who had one thing in common – they all penned criticisms of John Rawls' liberal philosophy because it treated people as unencumbered selves without reference to their relationships to their communities.

It was in the 1990s that a group of social and political theorists in Britain and America formally adopted 'communitarian' as the name for their shared philosophical outlook. Through the writings and collaborations of these thinkers, who included Philip Selznick, Robert Bellah, Jonathan Boswell, Amitai Etzioni, and Henry Tam, a body of work emerged to elucidate what communitarianism is about, provide a framework for understanding its intellectual heritage dating back to fifth century BC, and point to future challenges in the twenty-first century, particularly in relation to the twin threats of local community life being fragmented and global social development failing to generate sufficient shared values and solidarity.

Aristotle's insightful correction of Plato's reduction of all human beings into narrow categories assigning them in a single monolithic state is arguably one of the earliest examples of communitarian criticism. By pointing out that people are irreducibly social beings who are attached to a range of communal associations, Aristotle rejected the Platonic assumption that morally or intellectually people can be coherently defined in abstract terms without due reference to the relationships they have or seek in actual communities.

The disagreement between the two Greek philosophers, however, pales in comparison with the disputes that erupted in China from fifth century down to third century BC. The fierce contest between the traditionalist Confucians, the communitarian Mohists, and other rival schools of thought brought into sharp relief the dividing lines between conflicting visions of how society should be organized. Against those who just wanted to preserve rigid traditional hierarchies, those who would prefer to ignore all hierarchies and ties, and those who sought to impose new hierarchies at will, the communitarian followers of the philosopher, Mo Tze (*c*.479–399 BC), advocated the development of communities based on mutual values. They had to defend not only their doctrines but also their lives when their commitment to oppose aggressive wars put them repeatedly at odds with lords and princelings seeking to make territorial gains by force.

The Mohist and Owenite concerns with displacing oppressive social structures, not with any kind of free-for-all individualism, but with a deliberative and experimental reconstruction of community life based on the development of mutually respected values, point to philosophical continuities that ran through the late nineteenth/early twentieth century communitarian ideas of Émile Durkheim, L.T. Hobhouse, Jane Addams, and John Dewey, down to the contemporary advocates of communitarianism.

The key challenge facing communitarians in the twenty-first century is the resurgence of the 'self'-centred conception of human interactions. Academic relativism and popular culture converge on detaching the 'self' from any objective set of values with which it can be morally judged and socially restrained. The 'self' may be that of unaccountable political leader, a fundamentalist extremist, a corporate magnate, or simply an individual who cares nothing for the effects of his or her actions on others.

Communitarians reject both the relativist neutrality of supposing individual 'selves' should be left alone to act as they wish and authoritarian injunctions to impose the rigid order of hierarchical communities. Instead they put forward arguments and proposals to enable ordinary citizens and those in leadership positions to find ways to build and sustain a more inclusive form of community life, with the help of shared learning, deliberative dialogues, and participatory forms of collective decision making.

Although advocates for traditionalist hierarchical communities are sometimes labeled 'communitarians,' this is a term those advocates have never adopted, and their ideas have consistently been criticized by the writers who actually refer to themselves as communitarians. In this article, the term 'communitarian' is therefore applied only to the ideas that are philosophical antecedents of, intellectually complementary to, or directly expressed in, the works by thinkers who have used the term 'communitarian' to describe their own stance.

The Emergence of Communitarian Thinking and Its Critics

The core themes of communitarian debates first emerged in ancient China, during the late 'Spring–Autumn' and early 'Warring States' periods (500–250 BC), when rival social philosophers contested how stable and prosperous community life could be engendered against a backdrop of growing social fragmentation and armed conflicts. Four distinct positions were articulated. Their interrelationships provide a better illustration of what are in dispute than the more commonly invoked communitarian-liberal polarization.

The first position was formulated by Confucius (551–479 BC) and staunchly defended by his followers. In essence, it maintains that society functions best when it has a firm hierarchical structure, and everyone accepts the traditional role that has been assigned to them without seeking to usurp someone else's position. Its favored tool of socialization is the inculcation of rites and customs, with which people are routinely reminded that they ought to do what is expected of them.

Against this position, the Taoists advocated a romantic form of anarchic individualism. The writings attributed to Lao Tze (*c*.550–450 BC) suggested that the ruler should do as little as possible, leaving people to live in small communities in accordance with their own wishes. Chuang Tze (*c*.369–286 BC) went even further in proposing that individuals should withdraw from conventional ties altogether and follow whatever mode of behavior that might take their fancy. The Taoists believe that such liberated individuals would have a much better life without the pressures exerted on them by rigid social structures.

The Mohists (named after the founder of their school of thought, Mo Tze, *c*.479–399 BC) challenged both the hierarchical traditionalist thinking of the Confucians, and the anarchic individualism of the Taoists. The communitarian alternative they put forward consisted of three key elements. First, members of communities are more likely to attain an improved quality of life if they are prepared to support others as they would like to be supported themselves. Second, what kind of mutual support would be most effective should not be determined by reference to customs or revered texts alone, but with the help of a combination of historical records, contemporary observation, and future experimentation. Third, communities must devise ways of acting together in order to advance the kind of mutual support backed by a critical review of the evidence.

While the Confucians attacked the communitarian Mohists for their refusal to respect the total sanctity of customs, the Taoists frowned upon their determination to organize for collective action, the Legalists (flourished *c*.third-century BC) attacked all their rivals on the basis that they assumed human beings could be expected to serve what those in the ruling position might deem to be the greater good without being made to do so by forceful commands. The fourth position rejects moral traditions, critical reasoning, and benign individualism as unrealistic drivers for social goals set by a ruling elite. Instead, it posits the need for strict laws backed by stringent sanctions to direct social behavior.

This four-way contest provides a framework for understanding the roots of communitarianism and for mapping its core disagreement with rival social theories on how communities should be organized. These divisions have resurfaced throughout history in China and other parts of the world, and they represent underlying tensions in the diverse approaches for resolving disagreement within as well as between sovereign states.

Cooperative Ethos and Communitarian Social Experiments

Although the Enlightenment is often characterized as a movement for individualism, it contains important strands of communitarian thinking. As Peter Gay (1973) pointed out, it was no accident that the most quoted moral philosopher amongst pro-Enlightenment writers in the eighteenth century was Cicero who advocated the pursuit of the common good through civic solidarity and informed deliberations. It was Robert Owen (1771–1858), who synthesized a range of Enlightenment ideas into a cogent social philosophy which set a new mission for education; inspired the cooperative movement, and promoted social experiments that came to be known by the mid-nineteenth century as 'communitarian' (the earliest recorded use of the term dates back to 1841).

What Owen set out in *A New View of Society*, essays on the principle of the formation of human character (1991, originally published 1813/14), was bold in its analysis and radical in its prescription. Owen maintained that people's lives could be substantially improved if they interacted with one another in mutually supportive communities, rather than carrying on with exploitative arrangements, which benefited a powerful few at the expense of the majority. He further argued that experimental alternatives could be developed to test out how social arrangements could be reformed irrespective of what traditional beliefs might be invoked to defend the status quo. Most importantly, he claimed that education held the key to giving people the skills, confidence, and will to bring about new social arrangements.

Owen helped set up cooperative enterprise where workers and their families were treated with respect, given educational support, and empowered to share in the profits generated. His example inspired the widespread development of cooperatives, which gave their members an equal say in the running of their workplace communities.

Owenites in the United Kingdom and those who emigrated to the United States set up diverse communitarian experiments to enable people to take greater control of their lives, not as isolated individuals, but as fellow members of shared communities. What was sociologically significant about these experiments was that they opened the door to a new way to explore forms of human association. Instead of abstract models with their theoretical pros and cons, the Owenites pioneered the development and adaptation of community life through education and practical organization. As J.F.C. Harrison (2010) explained, while those early experiments produced a mix of successes and failures, they pointed later generations to the possibility of critical community development.

The Relationship between Liberalism and Communitarianism

Although in the 1980s there was a common trend in focusing on the debates between John Rawls and his communitarian

critics (e.g., Charles Taylor, Michael Sandel, and Michael Walzer) as a contest between liberalism and communitarianism, a close reading of these critics would reveal that they rejected any suggestion that they were putting forward an alternative social philosophy in the form of 'communitarianism' to displace liberalism as such. What they were interested in was to pick out particular weaknesses in liberal thinking, and explained why a more robust understanding of human relationships, in the context of what would constitute healthy and coherent community life, was necessary to supplement those liberal insights that were in fact valid. As such they echoed the intellectual debates that began a century earlier and which characterized the reorientation of liberal ideas in late nineteenth/early twentieth century.

Individualist liberalism had by the 1890s leaned toward an increasingly laissez faire outlook that would have heartened the most radical Taoists, if not for the fact that instead of an overbearing state, there were now powerful factory owners, merchants and bankers who were imposing the social arrangements they favored on people who, on their own, had no chance of making any alterations to those often oppressive arrangements.

Against this transformation of liberalism into a doctrine that privileged the commercial elite, turn-of-the-century thinkers such as Émile Durkheim, L.T. Hobhouse, Jane Addams, and John Dewey led a counter-movement in rethinking what a society that is both liberal and communitarian should be like.

For these thinkers, liberal concerns with freeing people from arbitrary intervention were not to be rejected wholesale for a communitarian alternative form of life. On the contrary, they believed those concerns needed to be properly addressed by recognizing that individuals would only develop freely as social beings if they were able to learn to live with and support each other as members of shared communities. Drawing on communitarian experiments in education, cooperative enterprise, and democratic collective action, they began to reorientate liberal thinking toward a more proactive commitment to community development.

Important studies have shown that communitarian thought grew in the context of augmenting liberalism, not in seeking to undermine it. Anyone wanting to know more about the symbiotic relationship between communitarian and liberal ideas in these thinkers should consult the works on Durkheim by Mark S. Cladis (1992); on L.T. Hobhouse and what came to be known as *New Liberalism* by Stephan Collini (1979); and on Dewey by Alan Ryan (1995), especially the section 'The Communitarian Turn' in Chapter 9. Re-acquaintance with their arguments to bolster liberalism with communitarian insights should go some way to displacing the misleading dichotomy of 'liberalism versus communitarianism' that became fashionable for a time in late twentieth century.

The Focus of Modern Communitarianism

In the 1990s, the long tradition of sociological insights and public advocacy relating to the critical development of community life outlined in the previous sections was further developed and applied to contemporary social problems when a group of American and British thinkers formally adopted the 'communitarian' name to describe this philosophy.

In the United States, Amitai Etzioni brought together a wide range of scholars and public figures, including William A. Galston, Philip Selznick, Robert Bellah, and Thomas Spragens Jr., under The Responsive Communitarian Platform: Rights and Responsibilities, a joint manifesto summarizing the core concerns of the group; produced the journal, The Responsive Community; and set up the Institute of Communitarian Policy Studies (http://icps.gwu.edu/).

In the United Kingdom, Henry Tam founded the Communitarian Forum with academics and civic leaders such as Tony Skillen, Chris Ormell, Lord Phillips of Sudbury (President, Citizenship Foundation), and Sheila Bloom (Chief Executive, Institute of Global Ethics, United Kingdom); devised a communitarian 'Working with Communities' program that was recognized with a national award by Prime Minister Tony Blair; and set up the Civil Renewal Unit for the UK Government in 2003 to run the 7-year nationwide 'Together We Can' campaign (http://www.hbtam.blogspot.co.uk/2013/06/find-our-more-about-together-we-can.html).

A good introduction to the core views of the modern communitarians can be found in the following key texts: Jonathan Boswell's (1990) *Community and the Economy*; Philip Selznick's (1992) *The Moral Commonwealth*; Amita Etzioni's (1993) *The Spirit of Community* and (1998) *The Essential Communitarian Reader* (edited collection); and Henry Tam's (1998) *Communitarianism*, and (2001a) *Progressive Politics in the Global Age* (edited collection).

These communitarian writers want to draw attention to five underlying problems engendered by the resurgence of a 'self'-centred social and political outlook since the 1980s. The growing emphasis on the individual self as owing nothing to anyone but one's own 'self' was exacerbating the corrosive effects of relentless marketization and the deterioration of social bonds. For communitarians, these problems would continue to worsen unless a better understanding of what would make for healthy relationships in the multiple layers of community life was attained to inform attempts at finding long-term solutions.

Of the five problems, the first and foremost is the failure to give adequate attention to both rights and responsibilities. If we only defend individuals' rights to claim what they in isolation seek to claim as their entitlement, and neglect to press people to recognize and fulfill responsibilities they owe others, those rights may become hollow in any case. As societies become ever more socioeconomically polarized, the need to find a proper balance between rights and responsibilities becomes even more urgent when those who have the most insist on their rights to do what they want with their wealth, while those with the least are repeatedly told to take responsibility for their own plight with little help from others.

The second problem is the masking of indifference as the face of neutrality. The imposition of values, especially when it would in effect exploit some to gratify others, is dubious at best, and repugnant at worst. But the rejection of objectionable values does not entail the rejection of all values. Radical relativism gives succor to irresponsible individuals who would like to pretend that everything they do is morally neutral. By reconnecting people to the attitudinal and behavioral expectations they actually have on others, communitarians believe the fabric of common values can be weaved together again.

Third, there is the problem of the neglect of communities by both state and market. In the nineteenth century, liberal constraints on the state were widely called for and applied, but that led to businesses becoming so powerful that they could dictate terms to people as workers and consumers. In the first half of the twentieth century, the state expanded to deal with the problems caused or neglected by the market. But throughout, the capacity of ordinary people to get involved, have a say, or exert some real influence over their lives was diminished. Both government institutions and businesses need to face up to the urgent need to rethink how they are to enable communities to play their part in a more balanced and inclusive society.

The fourth problem concerns the unfortunate severance between a sense of universal respect and a commitment to more local, communal attachments. In a globalized world, people increasingly look to their families, neighborhoods, and national borders for a sense of belonging and security. But at the same time, they need to adapt more than ever to interactions beyond settled boundaries. In place of the dichotomy of either a bland universalism or a blinkered localism, communitarians seek to help build the solidarity for a worldwide community of diverse communities.

Finally, human beings in the modern world need to be reconceptualized as people whose potential for a fulfilling life is developed and realized as members of overlapping communities. The tendency to see them as nothing more than cogs in a corporate machine, utility-maximizing consumers, or free-floating atoms indifferent to their proximity to others must be challenged.

The Key Communitarian Principles

In order to tackle the aforementioned problems, communitarians not only highlight the contradictions and distortions in the misguided conceptions of human relationship in contemporary society, they also put forward ideas for new models of interaction and power structure that would enable their participants to reach a more nuanced understanding to guide the future development of their social, economic, and political arrangements.

There are broadly three communitarian principles that are central to shaping appropriate models for human interaction. First, the principle of cooperative enquiry requires anyone making an assertion to be judged with reference to the extent to which informed participants deliberating under conditions of thoughtful and uncoerced exchanges would concur. Any provisional consensus reached by one group of individuals must in turn be open to possible revisions subject to examinations carried out with input from other groups. The ultimate strength of any truth claim rests with the likelihood of that claim surviving the critical deliberations of ever expanding circles of enquirers.

Second, the principle of mutual responsibility requires all members of any community to take responsibility for enabling each other to pursue those values which stand up to the test of reciprocity. What an individual may value cannot expect to command the respect from others if its pursuit is incompatible with the realization of goals valued by others. The range of mutual responsibilities may expand over time to cover, for example, direct and indirect care for dependents, help to those who would otherwise be neglected, safeguards for verifiable evidence and coherent reasoning, and cultivation of personal abilities not inimical to those of others. Omission to support, as well as action to harm, would be deemed a breach of the responsibility owed to each other.

Third, the principle of citizen participation requires that all those affected by any given power structure are to be able to participate as equal citizens in determining how the power in question is to be exercised. All those subject to potentially binding commands should be entitled to learn about, review, and determine how to reform decision-making processes. This applies to not only government institutions, but also businesses, schools, and community organizations. It follows that power relations should not retain structural or cultural barriers which hold people back from accessing information, putting forward their suggestions, questioning proposals, or sharing in decision-making processes.

The application of these principles to devising and evaluating modes of human interaction will help to show up barriers and identify specific improvements that should enable the people involved to relate to each other as members of communities with shared values. In practice, however, both established institutions and the mind-set of individuals will need to be changed if the communitarian principles are to be realized.

Communitarianism and Social Change: Key Questions

Communitarians want to transform communities so that they move closer to what community life should be like. At the most basic level, a community is no more than a group of people who have something in common which brings them and keeps them together. What is in common could be a geographical location, a historical identity, the fear of a common enemy, a place of worship, a shared interest, a common employer, or a bond that has grown out of solidarity against similar prejudices and discrimination.

At any given time, however, there are already social and institutional arrangements in place. Where these are entrenched and yet ill-disposed to embrace the improvements sought, communitarians face the problem their critics routinely present to them: how can they bring about social change without disrupting, and possibly undermining, prevailing community structures?

There are at least four lines of criticism derived from this problem. First, there is the suggestion that communitarian aspiration for social change is hemmed in by the limits set by many hierarchical communities. This applies to modern corporations as much as to traditional communities. Instead of complacently leaving those in charge to reform from within, communitarians have put forward substantial social, economic, and political reforms to open up debates on what changes ought to be considered (see, e.g., the works of Boswell, Etzioni, Derber, and Tam). Communities that have become accustomed to neglecting some of their members may well be reluctant to change, and that is where strong advocacy, subtle persuasion, and vivid presentation have their role to play in altering mind-sets and building alliances for social transformation.

Second, communitarians are often accused of demanding homogeneity as a precondition for effective community life. It is claimed that they want to secure social cohesion at the expense of cultural pluralism because they believe that the rejection of a common set of values and practices would lead to instability and conflicts. But the emphasis on the importance of common values and practices is not inherently incompatible with individual autonomy.

The distinction Durkheim and Hobhouse drew between mechanical and organic solidarity helps to explain precisely that whereas the narrow values which underpin mechanical solidarity demand total homogeneity, organic solidarity offers the opportunity for individuals to develop themselves and attain a stronger sense of social cohesion at the same time. Values grow out of an expanding mutual understanding between people. The fear of 'foreign' or 'alien' is often most pronounced in those areas where people are most isolated from others with different lifestyles and cultures. The community bonds in cosmopolitan cities show that people can have a strong sense of shared belonging and at the same time have attachments to diverse interests and customs within the overarching framework of a set of inclusive common values.

The third type of criticism concerns the issues of power and dominance in established communities. Advocates for women, gays, disabled people, minority ethnic groups, are rightly troubled by the sense of righteous behavior carried out by the members of certain communities. In those communities, some people define their identity in a disturbingly large part in terms of the superiority they allow themselves to feel toward those who are traditionally discriminated against.

The sociological starting point adopted by communitarians would maintain that there are different forms of community life. To develop those which are most likely to promote the widest opportunities for individual growth without fragmenting communal links into a state of indifference or even conflicts, we need to break down barriers to equitable power distribution so that through open communications and co-operation we can identify appropriate reform strategies. It follows that existing communities of family, neighborhood, and nation should be evaluated critically in relation to the key communitarian principles. The more they contain oppressive relations and practices which demand the unquestioned obedience of those in subordinate positions, the more they need to be exposed as authoritarian and subject to radical transformation.

The fourth and final area of criticism concerns the exclusivity of communities. This objection concerns the compatibility between strong community identity and the inclusive ideal of a community of communities. Miller (1990) and others have claimed that the two are incompatible because communities must have a high degree of exclusivity if they are to provide their members with a real sense of belonging.

From a communitarian point of view, community identity does not rest on exclusive allegiance. We can, and many of us do, feel a real sense of belonging toward a multiplicity of communities, our neighborhoods, towns, counties, regions, nations, clubs, unions, professional bodies, work teams, interest-based societies, lobbying groups, etc. Many of these enable their members to be involved in major decisions which affect their lives. Furthermore, these communities do associate with each other both horizontally and vertically. For example, the voluntary groups of a town coming together to form an umbrella group for the town; clubs across Europe forming a continental association; local businesses joining with regional businesses to set up a business support organization; and national and regional politicians working together on common interest issues. Some of these links are not strong, but others create their own communities where the sense of identity and belonging is as important as that within one of the constituent communities, and the nation state is only one of these communities.

People's attachment to a web of communities means that exclusivity is often an imagined desire rather than an actual barrier to a fulfilling community life. In reality, the task of developing healthy interactions within and between a multiplicity of communities is one becoming more urgent as the process of globalization takes out the option of communities existing in complete isolation from one another. As different communities impact more and more on each other, they will only be able to flourish if they learn to relate to each other by cultivating a set of common values and practices to which they will give their support. The alternative is a collusion course between rival communities, which will undermine the well-being of all.

Communitarianism, Education, and Community-Based Learning

Communitarians rely on deliberative education and reasoned persuasion to win over hearts and minds in changing communities. Like the Mohists, while they recognize that traditions and customs may have a role to play in ordering society, they insist that these must be supplemented, and where appropriate corrected, by facts we learn from experience and ideas that can be tested out through social experimentation.

Education therefore has a central role in communitarian philosophy. There are three aspects that merit particular attention. First, the education of children raises the question of what and how they are to be taught. Since from a communitarian perspective, dogmas and prejudices cannot be defended as just another set of customary beliefs or group values, the contents of education can be contested and arguments put forward to support teaching that upholds the communitarian principles. Schools need to move forward in terms of their structures and teaching methodologies so that pupils, teachers, and parents can function as partners in raising understanding. Fielding and Moss (2010) provides an example of how a more democratic and communitarian ethos can be developed in schools. In England, the cooperative movement in the 2010s began to give systematic support to the development of cooperative schools, which formally integrate the communitarian values of the movement into the running of each school. Outside the school, there is the issue of how the socioeconomic status of parents may affect the access their children have to effective education. Where individuals and schools cannot overcome these problems, the state at the local or national level would need to help parents and children – from support to enable them to have enough to eat or travel to school, to the provision of advice to parents and facilities for after-school care where necessary.

Second, there is the educative influence beyond teaching institutions for young people. Work-based learning, adult education, and media output generally can shape people's understanding and attitudes to a large degree. Not only is it necessary for schools to prepare for what other sources of influence may be at work on their pupils' minds, society also needs to be aware what people are led to think and feel after they have left school. Just as what parents or schools teach their children cannot be ringfenced as their private business that has nothing to do with the wider society, the activities of media, trainers, lifelong learning providers, research institutions, advertising and public relations agencies, all have social impact which cannot be shielded from critical scrutiny. Liberal safeguards against arbitrary censorship are not to be seen in this context as insurmountable barrier to public intervention, but as providing a reliable framework to counter the promotion of falsehoods and prejudices. Furthermore, the common good is not to be enhanced necessarily just by negative constraints, and it is likely that public support in the form of investment and coordination can considerably strengthen the quality and dissemination of research and information.

Finally, there is the need to improve education in citizenship and democratic collaboration. Communitarians have written about how the disposition and skills toward interacting with others with a mix of shared values and conflicting views should be cultivated. One of the most comprehensive initiatives to develop such an educational approach across all social spheres – in schools, community groups, local government, businesses, public policies – was taken forward by the British Government when it set up the Civil Renewal Unit in 2003 with the communitarian, Henry Tam, as its head. The initiative, under the title of 'Together We Can,' enabled neighborhoods to learn in partnership with public and private bodies how to improve their common well-being. It supported Civic Pioneer local authorities to help local residents and groups as well as their own staff to deepen their understanding of how more effective democratic decision making can be achieved. It helped academic institutions, community activists, and public service providers to develop Take Part projects to facilitate active citizenship learning in a wide range of communities. It also funded regional Community Empowerment Partnerships in all parts of England to bring together diverse groups and organizations to share ideas on how best to spread learning on what came to be known as 'cooperative problem-solving.'

Tam's (2013) paper on 'Cooperative Problem-Solving and Education' and the web resource generated by the 'Together We Can' program demonstrate how communitarian ideas on social transformation can be taken forward by coordinated efforts to educate citizens and the diverse communities to which they belong in applying cooperation and conflict resolution approaches in cultivating shared values, so as to attain better forms of community life.

See also: Attitudes, Political and Civic Culture; Citizenship, Historical Development of; Communitarianism; Community, Social Context of: The US Case; Cooperative Learning in Schools; Deliberation and Democracy; Dewey, John (1859–1952); Durkheim, Emile (1858–1917); Individualism versus Collectivism: Philosophical Aspects; Intergroup Relations; Power in Society; Republicanism: Philosophical Aspects; Social Enquiry and Action Research for Social Work; Social Identity in Social Psychology.

Bibliography

Annette, J., Mayo, M. (Eds.), 2010. Take Part? Active Learning for Active Citizenship. NIACE, Leicester.
Bell, D., 1993. Communitarianism and Its Critics. Clarendon Press, Oxford.
Bellah, R., et al., 1986. Habits of the Heart: Individualism and Commitment in American Life. University of California Press, Berkley, CA.
Boswell, J., 1990. Community and the Economy: The Theory of Public Co-operation. Routledge, London.
Boswell, J., Tam, H., 2013. Communitarianism Revisited. http://henry-tam.blogspot.co.uk/2013/02/communitarianism-revisited.html.
Brannan, T., et al., 2007. Re-energizing Citizenship: Strategies for Civil Renewal. Macmillan Palgrave, Basingstoke.
Cladis, M.S., 1992. A Communitarian Defense of Liberalism: Emile Durkheim and Contemporary Social Theory. Stanford University Press, Stanford.
Collini, S., 1979. Liberalism & Sociology: L. T. Hobhouse and Political Argument in England 1880–1914. Cambridge University Press, Cambridge.
Derber, C., 2000. Corporation Nation. St Martin's Griffin, New York.
Etzioni, A., 1993. The Spirit of Community. Touchstone, New York.
Etzioni, A., 1996. The New Golden Rule. Basic Books, New York.
Etzioni, A. (Ed.), 1998. The Essential Communitarian Reader. Roman & Littlefield, Lanham, MD.
Etzioni, A., 2004. The Common Good. Polity Press, Cambridge, MA.
Fielding, M., Moss, P., 2010. Radical Education and the Common Good: A Democratic Alternative. Routledge, London.
Gay, P., 1973. The Enlightenment: An Interpretation (The Rise of Modern Paganism). Wildwood House, London.
Hale, S., 2006. Blair's Community: Communitarian Thought and New Labour. Manchester University Press, Manchester.
Harrison, J.F.C. 2010 edition of 1969 original. Robert Owen and the Owenites in Britain and America: The Quest for the New Moral World. Routledge, Abingdon.
Lewis, M., Conaty, P., 2012. The Resilience Imperative: Cooperative Transitions to a Steady State Economy. New Society Publishers, Gabriola Island (Canada).
McGreal, I.P. (Ed.), 1995. Great Thinkers of the Eastern World. Harper Collins, New York.
Miller, D., 1990. Market, State and Community. Clarendon Press, Oxford.
Multhall, S., Swift, A., 1992. Liberals and Communitarians. Blackwell, Cambridge, MA.
Owen, R. 1991 edition of 1813/14 original. A New View of Society and Other Writings. Penguin Classics, London.
Ryan, A., 1995. John Dewey and the High Tide of American Liberalism. W.W. Norton & Co., New York.
Sandel, M., 1996. Democracy's Discontent. Harvard University Press, Massachusetts.
Selznick, P., 1992. The Moral Commonwealth. University of California Press, Berkeley, CA.
Selznick, P., 2002. The Communitarian Persuasion. Woodrow Wilson Center Press, Washington.
Sirianni, C., Friedland, L., 2001. Civic Innovation in America: Community Empowerment, Public Policy and the Movement for Civic Renewal. University of California Press, Berkeley and Los Angeles.
Tam, H., 1998. Communitarianism: A New Agenda for Politics and Citizenship. Macmillan, Basingstoke.
Tam, H. (Ed.), 2001a. Progressive Politics in the Global Age. Polity Press, Cambridge.
Tam, H., 2001b. The community roots of citizenship. In: Crick, B. (Ed.), Citizens: Towards a Citizenship Culture. Blackwell Publishers, Oxford, pp. 123–131.
Tam, H., 2005. Civil renewal & diversity. In: Lafleche, M. (Ed.), Social Capital, Civil Renewal & Ethnic Diversity. Runnymede Trust, London.
Tam, H., 2011. Rejuvenating democracy: lessons from a communitarian experiment. Forum for Promoting 3-19 Comprehensive Education 53 (3), 407–420.
Tam, H., 2013. Cooperative problem-solving and education. Forum for Promoting 3-19 Comprehensive Education 55 (2), 185–201.
Tze, Mo, 1929. Translated from the Original Chinese Text by Yi-Pao Mei. The Ethical and Political Works of Motse. Arthur Probsthain, London.

Communitarians

Matteo Bortolini, Università di Padova, Padova, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Communitarians' is the collective label used to identify a number of Anglo-American scholars who in the 1980s proposed a heterogeneous set of philosophical and sociological arguments as a critique of rights-based and universalistic versions of liberalism. After the philosophical debate faded in the mid-1990s, the label has been used to identify the members of an intellectual-political movement founded by Amitai Etzioni. The first section presents the main ideas of the four theorists who were widely considered as representative of communitarianism – Michael Sandel, Alasdair MacIntyre, Charles Taylor, and Michael Walzer – and explains their explicit rejection of the 'communitarian' label. The second section focuses on Etzioni's carefully crafted intellectual-political movement and investigates its imaginative relationship with the philosophical debate on liberalism and community.

Introduction

During the 1980s and the 1990s the label 'communitarians' was used to indicate a heterogeneous aggregate of intellectuals who shared a broad and articulated appreciation of community-centered philosophical arguments and political proposals. Thanks to its hermeneutical and narrative methodology, socially embedded conception of human beings, and the importance it ascribed to a collectively held conception of the good life, communitarianism was often associated to both republicanism and radical conceptions of democracy as the main alternative to the deontological and individualistic versions of liberalism championed by thinkers as John Rawls, Ronald Dworkin, and Robert Nozick. As the philosophical debate over liberalism and community waned in the mid-1990s, the label was used to define the members of an intellectual-political movement – the 'responsive communitarians' – founded by Amitai Etzioni in 1991. Lists of communitarian philosophers usually include Michael Sandel, Alasdair MacIntyre, Charles Taylor, and Michael Walzer, while Etzioni, Philip Selznick, and Robert Bellah are counted among the most representative communitarian sociologists.

Any reconstruction of the collective biography of the communitarians should take two apparent paradoxes into account. First, the four philosophers who were deemed as the major 'communitarians' have rejected such a label – ironically, one may even point to this denial as their only common trait. Second, while their ideas figure prominently in the literature associated with Etzioni's movement, they were not personally involved with the latter and some doubt could be cast about their substantial influence on the movement's 'Responsive Communitarian Platform.' In fact, the label 'communitarians' was projected from without on a diverse group of scholars to define them, first, in their relationship to another school of thought, mainstream procedural liberalism, and, second, as the intellectual forerunners of an organized intellectual-political movement. Most retrospective accounts have recognized and ignored both the uneasiness shown by Sandel, MacIntyre, Taylor, and Walzer toward the communitarian label and the lack of personal links with the responsive communitarians, and have discussed their work from a highly synthetic and abstract point of view – the only strategy that allows to group the four philosophers as 'communitarians.' In fact, as they presented arguments located on a host of different epistemic levels (methodological questions about the premises and the articulation of moral reasoning, ontological conceptions of the relationship between the self and society, and normative and advocacy-oriented positions), thinkers who might share the same ground on one level often parted ways on the others. In order to take the self-identification of the four scholars into account, the label 'communitarians' would have to be dropped in favor of other definitions – such as 'republican,' 'civic-humanist,' or, in Walzer's case, 'social democrat.' At that point, however, the scope of the story would have to be extended to include all non-individualistic versions of modern political theory and, with them, a diverse array of theorists.

What follows presents the collective biography of the 'communitarians' from a historical-sociological point of view which accounts for the emergence of the communitarian label by looking at late twentieth-century Anglo-American political and social theory as relational intellectual fields in which the position of each scholar was defined by the positions occupied, and the views promoted, by other actors. The first section supplements the summary of the main ideas of the four putative communitarian theorists – Sandel, MacIntyre, Taylor, and Walzer – with an investigation of their network of intellectual relationships aimed at explaining their differences and their response to the process which labeled them as 'communitarians.' The second section is dedicated to Etzioni's translation of the communitarian label from the philosophical debate to a carefully crafted intellectual-political movement of which no one of the four philosophers was a member. As it happens with any translation, this processes involved creative shifts in content, scope, and tone which are both intellectually and sociologically interesting.

'Communitarians' vs Liberals

Michael Sandel's *Liberalism and the Limits of Justice* (1982), the book which is widely considered as the debut of modern-day communitarianism, emerged from its author's experience as a student at Oxford in the late 1970s, where the thought of Aristotle, Kant, and Hegel was filtered by the philosophical

insights of Charles Taylor, the Canadian philosopher who occupied the Chichele chair of social and political theory, and Alan Montefiore, a British philosopher conversant with both the analytic and the continental traditions. The intense exchanges within a network which also included Stuart Hampshire, Ronald Dworkin, Leszek Kołakowski, Michael Oakeshott, and Sheldon Wolin, led Sandel (b. 1953) to develop a critique of what he saw as the thread joining apparently diverse versions of liberalism, the privilege accorded to 'the right' over 'the good' – the idea that a pluralist society is best arranged when it is governed by principles of justice that do not presuppose any particular conception of the good.

Sandel's criticism focused on the most influential contemporary rendition of Kantian-inspired liberalism – *A Theory of Justice*, the groundbreaking treatise published in 1971 by the Harvard philosopher, John Rawls, as a critique of Utilitarianism. Using an array of analytical arguments and thought experiments – among which the idea of an 'original position' where the actual distribution of natural talents, preferences, and social positions is hidden beyond a 'veil of ignorance' in order to reach a universalistic agreement on some basic principles of justice – Rawls had defined an articulated conception of 'justice as fairness' meant to replace the principle of utility as the main theoretical device for assessing the moral worth of the distribution of social and economic advantages in a democratic society. Besides receiving the critiques of major Utilitarians such as Richard M. Hare, Rawls had been attacked from a libertarian point of view by fellow Harvardian Robert Nozick, who grounded his case against redistributive justice on the absolute right of personal self-ownership (*Anarchy, State, and Utopia*, 1974).

Liberalism and the Limits of Justice criticized *A Theory of Justice* from the opposite pole of the philosophical spectrum: according to Sandel, Rawls's Kantian 'unencumbered' self, prior to its ends and independent from them, was too thin to justify both its adhesion to the principles of justice (the right) and its choice of ends (the good). In contrast, Sandel proposed an 'encumbered' idea of the self, whose identity was already constituted by a plurality of pre-existing communal narratives shared by different groupings of concrete human beings (family, city, tribe, class, nation, people); the encumbered self presupposed a constitutive understanding of the community which was at odds with the instrumental and sentimental conceptions of communal life acknowledged in *A Theory of Justice*.

Sandel first used the adjective 'communitarian' in the introduction of *Liberalism and Its Critics* (1984), a reader in which he collected, among the others, selections from Taylor, MacIntyre, Walzer, Oakeshott, Peter L. Berger, and Hannah Arendt. Sandel explicitly contrasted Utilitarianism and Kantian liberalism with the 'growing challenge' of 'communitarian critics' inspired by Aristotle, Hegel, and civic republicanism. In 1985, the American philosopher Amy Gutmann published an influential article, 'Communitarian Critics of Liberalism,' where Sandel, MacIntyre, Taylor, Walzer, and Roberto Unger were labeled as 'new communitarians.' Gutmann's intent was clearly polemical. Whereas 'old' – that is, Marxist – communitarians strived for equality and collective property ownership, modern communitarians favored 'traditions and established identities.' Gutmann defended Rawls against what she saw as an attack to individual rights and the very idea of justice, and criticized Sandel's suggestion of reviving the civic republican stream implicit in the American tradition on the ground that it would bring back majoritarianism and, eventually, 'Salem.'

A flow of articles – including well-known papers by Will Kymlicka, John Wallach, and Christopher Lasch – and special issues of journals quickly followed Gutmann's lead. The debate was mostly fed by liberals, who defended the deontological version of universalism from what they perceived as untimely, misguided, or even authoritarian critiques. The 'communitarian' label was often used in a dismissive and negative way which effectively, albeit not always openly, referred to Ferdinand Tönnies's discredited inclination for traditional *Gemeinschaft* over modern *Gesellschaft*. In fact, a rapid survey of the intellectual biographies of the alleged communitarians – all belonging to an older generation than Sandel's – reveals a broad overlapping of networks and interests (respectively, the British New Left and participatory democracy), as well as different scopes, arguments, and self-identifications – a heterogeneity that most of the commentators acknowledged but decided not to emphasize.

Educated at McGill University in Montreal and at Oxford under the supervision of Hampshire and Isaiah Berlin, Charles Taylor (b. 1931) developed an early interest in the thought of European thinkers such as Maurice Merleau-Ponty and Emmanuel Mounier, as well as a militant political activism in the British New Left (together with Stuart Hall, Gabriel Pearson, and Ralph Samuel, he was a founder of *Universities and Left Review*) and in the Canadian New Democratic Party. His first book, *The Explanation of Behaviour* (1964), focused on the continuing relevance of teleological explanations of human (and animal) agency, while his 1967 essay, 'Neutrality in Political Science,' argued for the convergence of empirical inquiry and normative theory within the framework of a wider hermeneutical view of the social and human sciences. In his book *Hegel* (1975), Taylor asserted from a non-Hegelian standpoint the significance of Hegel's work for the recurring modern problem of situating 'subjectivity by relating it to our life as embodied and social beings, without reducing it to a function of objectified nature.'

In the liberal-communitarian debate, Taylor was mostly cited for his 1979 essay, 'Atomism,' in which he attacked the Hobbes-Locke tradition of political theory and the priority it granted to individual rights in the justification of political action and institutions. Taylor argued that the primacy of individual rights was implicitly based on an anthropology of individual self-sufficiency, to which he opposed a theory of 'the social nature of man' according to which the very identity of independent individuals demands a 'social matrix' – a set of practices which recognized the individual's autonomy and his right to participate in deliberation about public action. In his conclusion, he underlined that any 'free individual' should be committed to sustaining the society which rendered his identity possible, reaffirming his principled support to the very idea of individual rights.

Born in New York City in 1935, Michael Walzer studied at Brandeis University with Lewis Coser and Irving Howe, whose journal, *Dissent*, published his first political articles and reportages in the late 1950s. A political activist in the civil rights movement and an affiliate with British New Left circles, Walzer studied in Cambridge, UK, with Geoffrey Elton and at Harvard

under the guidance of Samuel Beer, Barrington Moore, Jr., and Judith Shklar. In the mid-1960s, Walzer joined the Society for Ethical and Legal Philosophy – along with Thomas Nagel, Dworkin, Rawls, and Nozick – a company which would prove crucial for the development of his ideas on justice, community, and individual rights. In his early publications – e.g., *Obligations* (1970) and *Just and Unjust Wars* (1977) – he employed a contextualist method in which political theory was used to frame the study of historical processes, and historical cases were utilized to clarify theoretical arguments.

The focus of liberal critics of 'communitarianism' was Walzer's *Spheres of Justice* (1983), a book grown in a constant dialogue with Nozick (with whom Walzer taught a course on 'Capitalism and Socialism' in 1970–71), Gutmann, and Taylor, as well as with Clifford Geertz, Albert Hirschman, and Judith Jarvis Thomson. The book put forward a complex vision of distributive justice grounded in a 'thick' interpretive understanding of material and immaterial goods as deeply intertwined with the story of both communities and personal identities. Each social good, according to Walzer, constituted an autonomous distributive sphere within which only certain standards and rules were appropriate. Consequently, the biggest obstacle toward justice was the 'dominance' of one good over other goods. *Spheres of Justice* advanced a conception of 'complex equality' where arrangements of the different spheres would be defended by the political community, which should take the form of a participatory 'decentralized democratic socialism.' Walzer's work was from the very beginning closer to liberalism than Aristotelianism. While justice was always rooted in the particular understandings that constituted a 'community' as a shared way of life, the community he was speaking of was a democratic, differentiated, and ultimately liberal Western society.

In Alasdair MacIntyre's *After Virtue* (1981) the political focus that characterized, albeit in different manners, the works of Sandel, Taylor, and Walzer was subordinated to an analysis of the moral malaises of modernity. As a student in London, Manchester, and Oxford in the 1940s and 1950s, Scottish born MacIntyre (b. 1929) made his debut with a book in which he traced the intellectual relationships between Marxism and Christian theology (*Marxism: An Introduction*, 1953). Originally a member of the same New Left circles to which both Taylor and Walzer belonged, MacIntyre shifted his political allegiance to the Trotskyist *International Socialism* group around 1960. The publication of *A Short History of Ethics* (1966) was a step away from Marxism, a divorce that was consumed with the second edition of his first book, now titled *Marxism and Christianity* (1968). In 1969, MacIntyre moved to America and participated in the so-called 'Hegel Craze' of the mid-1970s editing an important collection of essays, *Hegel* (1972), to which Taylor contributed along with Walter Kaufmann and Shlomo Avineri.

In *After Virtue*, MacIntyre criticized the 'grave moral disorder' of contemporary society, and traced it back to the failure of the 'Enlightenment project.' Since morality is always rooted in a particular way of life, the universalistic quest of modernity could only fail in an unsettable clash of vocabularies, *genres*, and disembodied ideas – in this sense, Nietzsche had correctly understood that the Enlightenment would finally bring up the dessication of moral rationality and the advent of emotivism and nihilism. As an alternative, MacIntyre called for a neo-Aristotelian teleological ethic based on a composite understanding of human virtues. MacIntyre's conclusions restated that the contrast between modern and classical moral philosophy and suggested that the creation of a new kind of secluded community aimed at preserving the true moral life from the attack of modernity's 'barbarians.'

After Virtue undoubtedly resounded with both Taylor's and Sandel's works for its conception of morality as a set of norms which human actors 'find' as soon as they undertake their quest for the good life. At the same time, the differences were striking. Whereas Sandel's book was a thorough analysis of Rawls's *A Theory of Justice*, MacIntyre's critique of contemporary liberalism only occupied a short chapter at the end of a long history of Western ethics. Moreover, whereas none of the other alleged 'communitarians' opposed the general framework of contemporary political institutions, MacIntyre wrote that modern politics 'simply had to be rejected.' These substantial differences did not prevent liberal critics to consider MacIntyre as one of 'the communitarians' and to attribute an impoverished summary of some of his more radical positions to Sandel, Taylor, and Walzer.

Given these premises, it is no wonder that in the mid-1980s the four philosophers started to explicitly position themselves vis-à-vis the 'liberal-communitarian' debate – paradoxically, in distancing himself from 'communitarianism' each took for granted that such a position did exist and that at least some of the others would identify themselves with it. In 1989, Taylor published an article titled 'Cross-Purposes: The Liberal-Communitarian Debate,' in which he characteristically distinguished two sets of issues (ontological and advocacy-related) and pleaded for a more complex vision of the debate that would dissolve the two labels, for ontological questions (holism versus atomism) might influence, but never determine, the choice between political positions (individualistic versus collectivistic). In distinguishing the two 'teams' Taylor did not enroll himself neither among the communitarians (Sandel, MacIntyre, Walzer) nor among the liberals (Rawls, Dworkin, Nagel, Scanlon); rather, he filed his own work under the 'civic humanist' and republican tradition of 'holistic individualism.'

In a 1990 essay, Walzer saw 'the communitarian critique of liberalism' as a recurring feature of modern political thought and distinguished two versions of it. According to the first, liberal political theory accurately depicted liberal societies as aggregates of isolated individuals – Walzer ascribed this view to neo-Marxists, neo-conservatives, and neo-classical republicans, among which he included Sandel and MacIntyre. The second version of the communitarian critique maintained, on the contrary, that liberalism misrepresented social life. According to the 'vulgar idealism' of the authors of *Habits of the Heart* (of which more later), by negating the commonal texture of social life, the rhetoric of liberalism put modern society at risk. Walzer acknowledged the fragmented nature of modern society, but argued for the existence of a 'liberal tradition' as the foundation of a modern society. At the same time, he advanced the need to teach 'liberal selves' to consider themselves as the living embodiment of liberal values, a public goal that could only be furthered by a 'deliberately nonneutral' State – an idea that Walzer attributed to John Dewey. Siding with the liberals, Walzer closed his essay with two final critiques to 'communitarians': first, civic republicanism was too focused on the

political sphere to be viable; and second, since political theory regards the relations between constituted selves, the whole discussion on the self was purposeless.

Although they had been associated with communitarianism even by Taylor and Walzer, neither MacIntyre nor Sandel would accept the label. In the 'Postscript to the Second Edition' of *After Virtue* (1984), MacIntyre did not take on anti-'communitarian' critiques and only wrote about on the three relationships on which his forthcoming volume, *Whose Justice? Which Rationality?* (1988), was built: that between philosophy and history, that between his conception of virtues and ancient wisdom, and that between his work and Biblical theology. MacIntyre explicitly declared his distance from 'communitarianism' at various times and places – finally in the introduction of the 2007 edition of *After Virtue*. As for Sandel, soon after the publication of his book he wrote 'The Procedural Republic and the Unencumbered Self' (1984), a widely reprinted essay in which he summarized his thesis about the need for a deeper understanding of the communal foundations of liberalism, but somehow started to shift from analytical arguments to historical narrative. His 1994 book, *Democracy's Discontent*, presented a detailed, theoretically informed account of the emergence of the 'procedural republic,' defined as the practical embodiment of liberal neutralist ideals, from the Independence up to the mid-1990s. Ultimately, in the preface of the second edition of *Liberalism and the Limits of Justice* (1998), tellingly titled 'The Limits of Communitarianism,' Sandel declared 'some unease' with the 'communitarian' label and restated his case for a perfectionist and republican understanding of politics and law where rights rested on a substantive moral judgment of the ends they advanced, thus explicitly distinguishing his position from both liberal neutralism and communitarian majoritarianism.

The 'Responsive Community Platform' and Communitarianism

In the early 1990s, the debate between liberals and communitarians thrived and assumed such a central and attractive position that even those scholars belonging to other schools of thought – e.g., feminists or neo-Marxists – were almost compelled to take a position toward it, both in the United States and abroad. Will Kymlicka (1989), Robert Booth Fowler (1991), Stephen Mulhall and Adam Swift (1992), Stephen Holmes (1993), and Derek Phillips (1993) all dedicated detailed monographs to the critique of communitarianism, whereas readings and anthologies were edited by David Rasmussen (1990), Shlomo Avineri and Avner De-Shalit (1992), and Cornelius Delaney (1994). All these works took for granted that a small group of cultural objects – *Liberalism and the Limits of Justice*, *After Virtue*, *Spheres of Justice*, and some of Taylor's essays – constituted the nucleus of the 'communitarian critique of liberalism.' Since the dynamic of the production and exchange of new arguments and texts depended less on Sandel, MacIntyre, Taylor, and Walzer's intentions than on the ongoing structuration of the field of moral and political philosophy, the self-understanding of the four theorists, and their subsequent work, could do little to stop the process. Daniel A. Bell's *Communitarianism and Its Critics* (1993) was one of the few books which patently defended the communitarian approach. In the introduction, Bell suggested that the philosophies of Sandel, MacIntyre, Taylor, and Walzer might function as a sound scholarly basis for the intellectual and political project outlined in Amitai Etzioni's 'Responsive Communitarian Platform,' irrespective of the fact (which Bell recognized in a footnote) that the four philosophers had never declared to be 'communitarians' and were personally unrelated to Etzioni's movement.

German-born Etzioni (b. 1929) was a well-known sociologist, who had hitherto been extraneous to the philosophical debate. He studied with Martin Buber and Shmuel Eisenstadt in Jerusalem in the 1950s, before enrolling at the University of California at Berkeley, where he got his PhD in sociology under Seymour M. Lipset's guidance. In 1958, Etzioni was hired by Columbia University, where he spent most of his academic career writing on organizational sociology (*A Comparative Analysis of Complex Organizations*, 1961) and intervening as a public intellectual on topics as diverse as the prospects of a nuclear war, Project Apollo, the war in Vietnam, and biotechnologies (*The Hard Way to Peace*, 1962; *The Moondoggle*, 1964; *Genetic Fix*, 1973). An advocate of reindustrialization – his ideas for 'rebuilding America' were spelled in *An Immodest Agenda* (1983) – in the late 1970s Etzioni briefly served as an advisor to President Jimmy Carter. His 1990 book, *The Moral Dimension*, criticized neoclassical economics from a 'socio-economic' view according to which economic actors and processes should be understood as embedded in social relations and cultural values.

Profoundly concerned about the excesses of Reaganian individualism and the refusal to debate moral issues on the part of progressive liberals, in early 1990 Etzioni, now at George Washington University, and William Galston, a professor of public affairs at the University of Maryland, gathered a small number of scholars to discuss the creation of an intellectual and political agenda. The group included, among the others, Philip Selznick, a sociologist who had enthusiastically embraced the 'communitarian' label in his 1987 article 'The Idea of a Communitarian Morality'; political theorists Benjamin Barber, James Fishkin, Jane Mansbridge, and Thomas Spragens; and William Sullivan, a philosopher who had co-authored with Berkeley sociologist Robert Bellah and three young colleagues *Habits of the Heart* (1985), a brilliant critique of the excesses of American individualism that was commonly regarded as one of the landmarks of sociological communitarianism. It was decided that the group would take its name from the title of a short essay Etzioni had published in 1987, 'The Responsive Community (I & We),' and would elaborate a far-reaching moral and political program, which Etzioni drafted with the help of Galston and Mary Ann Glendon, the Harvard Law School professor who had just published *Rights Talk* (1991).

The 'Responsive Communitarian Platform: Rights and Responsibilities' was published in 1991 and was endorsed by a diverse group of intellectuals, politicians, and activists, among which were Bellah, Harvey Cox, Albert Hirschman, Richard John Neuhaus, Lester Thurow, Francis Fukuyama, David Riesman, Orlando Patterson, and Adlai Stevenson. The Platform advanced a deeply embedded vision of the human individual and stressed the need of balancing rights and responsibilities on both the side of the individual and that of

the many different communities to which each citizen of a pluralist, democratic society belongs. In order to strengthen the communitarian fabric of American society, the movement called for a stronger civil society founded on 'moral education and character formation,' and emphasized the role of both families and schools in rendering citizens sensitive to their duties toward each other and the wider political community. Etzioni stressed the importance of balanced, 'average' two parent families and refused the identification between communitarianism and majority rule, calling instead for a more participative and stronger democracy. The Platform also contained policy suggestions on domestic disarmament, parental leaves from the workplace, national and local service, and the reduction of campaign financing. In calling for more solidarity and less individualism, more grassroots initiatives and less bureaucracy, Etzioni's agenda looked for a third way between conservatism and liberalism that was nonetheless rooted in a modern understanding of politics.

These themes were restated and reworked in Etzioni's *The Spirit of Community* (1993) and *The New Golden Rule* (1996), two books aimed at reaching a wider, non-academic public. Along with Galston, Fishkin, and Glendon, Etzioni also launched *The Responsive Community*, a quarterly journal dedicated to communitarian concerns. The journal's editorial board included public intellectuals such as Martha Nussbaum, Jean Elshtain, Nathan Glazer, and Daniel Yankelovich. As Galston took a job as the President's deputy assistant for domestic policy, the Clinton administration sponsored a host of communitarian initiatives, among which were the annual White House Conferences on Character Building. In the 1990s, communitarianism and its main spokesperson, Etzioni, were widely praised in the United Kingdom, where Tony Blair subscribed to the Platform, and Germany, where Etzioni was able to gain support from social democrat, green, and conservative leaders.

The use of the noun 'communitarians' to refer to the members of Etzioni's movement suggested a fundamental continuity between the philosophical debate of the 1980s and the political movement of the 1990s, a relationship that was often taken for granted in critical literature, popular understanding, and encyclopedia entries. From the point of view of intellectual networks, the four 'original' communitarians maintained very different stances toward the movement. While Walzer and Sandel remained indifferent, in 1991 MacIntyre sent a letter to *The Responsive Community* in which he unquestionably rejected the label 'communitarian' on philosophical and political grounds. Even if he never endorsed the Platform, Taylor expressed his sympathy for the movement and in 1994 became a member of the editorial board of *The Responsive Community*, contributing four articles before the journal shut down in 2004 (Sandel contributed one article, Walzer none). In 1995 Etzioni edited an anthology, *New Communitarian Thinking*, aimed at establishing communitarianism as a robust social and political theory in its own right; interestingly enough, the reader included three essays by Walzer, Taylor, and Sandel, which did not belong to the key works of the liberals/communitarians debate.

In reclaiming the very label that no one of the four 'communitarian' philosophers wanted for himself, Etzioni succeeded in branding a viable intellectual-political position which creatively combined the community-oriented plea for the duties and the responsibilities of an educated citizenry with the liberal stress on individual rights. As an alternative to the cleavage between conservatives and liberals typical of contemporary American politics, Etzioni's communitarianism revamped a discourse of commonality and embeddedness that echoed with the communitarian *Zeitgeist* of the post-Reaganian era without embracing either the radical conclusions of Alasdair MacIntyre or the social democratic ideals of Michael Walzer. For this and other reasons, the continuity between the diverse works of Sandel, Walzer, Taylor, and MacIntyre, and Etzioni's sociological understanding of the relationship between the individual and society should not be taken for granted and deserves a more thorough examination from both an intellectual and a historical point of view.

See also: Atomism and Holism: Philosophical Aspects; Contractarianism; Cultural Rights and Culture Defense: Cultural Concerns; Ethics and Values; Hegel, Georg Wilhelm Friedrich (1770–1831); Individualism versus Collectivism: Philosophical Aspects; Justice: Philosophical Aspects; Kant, Immanuel (1724–1804); Kantian Ethics and Politics; Liberalism: Historical Aspects; Person and Self: Philosophical Aspects; Rawls, John (1921–2002); Republicanism: Philosophical Aspects; Sandel, Michael (1953–); Taylor, Charles (1931–); Virtue Ethics; Walzer, Michael (1935–).

Bibliography

Avineri, S., De-Shalit, A. (Eds.), 1992. Communitarianism and Individualism. Clarendon Press, Oxford.
Bell, D., 1993. Communitarianism and Its Critics. Clarendon Press, Oxford.
Bellah, R., Madsen, R., Sullivan, W., Swidler, A., Tipton, S., 1985. Habits of the Heart. University of California Press, Berkeley.
Delaney, C.F. (Ed.), 1994. The Liberalism-Communitarism Debate. Lanham Rowman & Littlefield.
Etzioni, A., 1993. The Spirit of Community. Crown Publishers, New York.
Etzioni, A. (Ed.), 1995. New Communitarian Thinking. University of Virginia Press, Charlottesville.
Fowler, R.B., 1991. The Dance with Community. University Press of Kansas, Lawrence, KS.
Gutmann, A., 1985. Communitarian critics of liberalism. Philosophy & Public Affairs 14 (3), 308–322.
Holmes, S., 1993. The Anatomy of Antiliberalism. Harvard University Press, Cambridge.
Kymlicka, W., 1989. Liberalism, Community and Culture. Clarendon Press, Oxford.
MacIntyre, A., 1984. After Virtue, second ed. University of Notre Dame Press, Notre Dame.
MacIntyre, A., 1988. Whose Justice? Which Rationality? University of Notre Dame Press, Notre Dame.
Mulhall, S., Swift, A. (Eds.), 1992. Liberals and Communitarians. Blackwell, Oxford.
Phillips, D., 1993. Looking Backward. Princeton University Press, Princeton.
Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge.
Rasmussen, D. (Ed.), 1990. Universalism vs. Communitarianism. MIT Press, Cambridge.
Sandel, M., 1996. Democracy's Discontent. Harvard University Press, Cambridge.
Sandel, M., 1998. Liberalism and the Limits of Justice, second ed. Cambridge University Press, Cambridge.
Taylor, C., 1975. Hegel. Cambridge University Press, Cambridge.
Taylor, C., 1985. Philosophy and the Human Sciences: Philosophical Papers 2. Cambridge University Press, Cambridge.
Taylor, C., 1989. Cross purposes: the liberal-communitarian debate. In: Rosenblum, N. (Ed.), Liberalism and the Moral Life. Harvard University Press, Cambridge.
Wallach, John R., 1987. Liberals, communitarians and the task of political theory. Political Theory 15 (4), 581–611.
Walzer, M., 1983. Spheres of Justice. Basil Books, New York.
Walzer, M., 1990. The communitarian critique of liberalism. Political Theory 1 (18), 6–23.

Communities and Crime

Lesley W Reid, University of Alabama, Tuscaloosa, AL, USA
Christopher Pell, University of Tennessee at Chattanooga, Chattanooga, TN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Crime clusters in communities, most often in those communities plagued by a host of other social problems, such as poverty, joblessness, and family dissolution. In the mid-1980s, many scholars dismissed the relevance of community characteristics on crime; today their study is a veritable cottage industry within criminology. This article (1) provides a synthesis of both the classical and contemporary research on the mechanisms through which community-level processes influence crime, (2) examines the areas of future exploration of these mechanisms, and (3) highlights research on how these mechanisms influence crime in specific types of communities.

Theoretical Understandings of Crime and Community

The classic and influential work of Shaw and McKay (1942) provides the foundation for contemporary work examining characteristics of communities and crime. According to Shaw and McKay, community distress created social disorganization, defined as the disruption of primary relationships, the weakening of norms, and the erosion of shared culture. Shaw and McKay's original formulation of the social disorganization theoretical model, presented in Figure 1, would have only included three cells as follows: the exogenous variable, concentrated disadvantage; the dependent variable, community crime rates; and an undifferentiated intervening variable, social disorganization (represented by the large gray box in Figure 1 labeled 'processes of social disorganization'). This foundational version of the theory proposed that concentrated disadvantage (e.g., poverty, joblessness, social isolation, residential instability, etc.) led to social disorganization, which in turn determined community differences in crime rates. While concentrated disadvantage could have had a direct effect on crime by generating instrumental crime, the primary mechanism through which disadvantage affected crime was through the conditions conducive to social disorganization, simply defined as the inability of a community to instill in its members a shared sense of cultural values and effectively sanction those who deviate from those values (Krohn, 1985; Sampson and Groves, 1989).

Up until the late 1980s, social disorganization theory was largely discounted for failing to articulate the precise mechanisms through which concentrated disadvantage influenced crime. In large part this was an unfair critique (Bursik, 1988). Early social disorganization theory was clear in its premise that concentrated disadvantage eroded shared norms and limited the capacity of neighborhood networks to socialize and sanction their members. However, Shaw and McKay's original version of the theory allowed for little variation in the

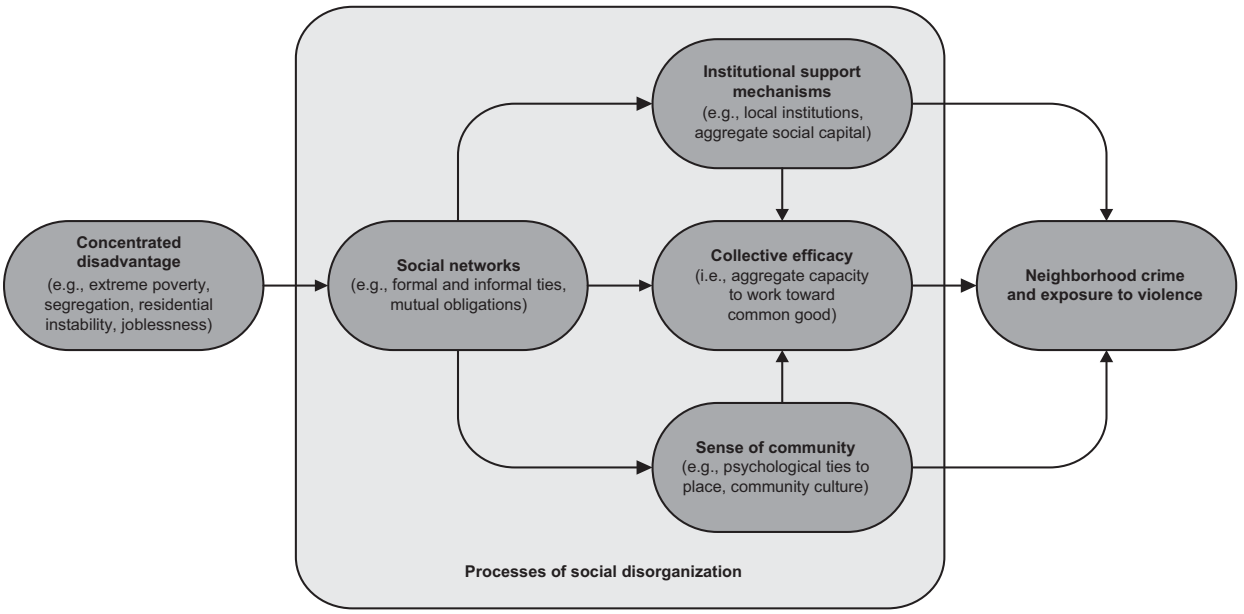

Figure 1 Heuristic model of social disorganization theory.

relationship between concentrated disadvantage and social disorganization. For the most part, the theory took a deficits-based approach (Almgren, 2005). If a community exhibited concentrated disadvantage, it would also exhibit social disorganization. The original theory did little to explain how some communities plagued by concentrated disadvantage were able to remain organized, efficacious places while others were not. Over the past 20 years, contemporary social disorganization research has addressed this weakness by exploring the role of a variety of factors in mediating the effects of concentrated disadvantage on crime (Sampson et al., 2002). These factors include institutional support mechanisms, social networks, a sense of community, and collective efficacy (represented by the dark gray cells embedded in the larger light gray box in **Figure 1**).

The contemporary social disorganization model (represented by the full model in **Figure 1**) defines communities as systems of interrelated networks. The role of a community network, according to social disorganization theory, is to maintain itself through the continued socialization of its members, including, but not limited to, the application of negative sanctions to deviant members. The maintenance of this system can be undermined by social disorganization manifested at the community level as weak social networks, limited participation in local organizations, and little social control. As the antithesis of social disorganization, each aspect of social organization (networks, organizations, and control) forms the social fabric of a community (Kasarda and Janowitz, 1974). This social fabric structures communities in the following ways. First, both formal social networks (e.g., school, church) and informal social networks (e.g., extended family, neighbors) enable community members to recognize nonmembers and thereby guard property of members against crime perpetrated by nonmembers. More importantly, these social networks sanction members for actions that are criminal (Krohn, 1985). Second, participation in local organizations serves to instill community solidarity through the support of those institutions that are the organizational base through which community members work toward common goals (Kornhauser, 1978). Finally, social disorganization is manifested as the inability of a community to exert either positive or negative social control on its members. In past research, this has been understood most often in terms of supervising juveniles in order to prevent delinquency.

Social Networks within Communities

The capacity of a community to exert control over its residents exists if for no other reason than neighbors know one another and interact on a regular basis (Sampson and Groves, 1989). Community networks, by definition place-based networks, control crime by reinforcing shared values; censuring criminal behavior, especially juvenile delinquency; and increasing communal guardianship of property (Bursik and Grasmick, 1993). However, neighborhood social networks need not be strong to exert social control. Building on the idea of the strength of weak ties (Granovetter, 1973), Bellair (1997) demonstrates that broad, loosely woven social networks often operate as well as narrow, tightly woven ones. Even in loosely organized networks, people look out for one another, acting as 'eyes on the street' in protecting each other's persons and property. Conversely, networks can be dense, but overwhelmed by the effects of concentrated disadvantage leaving residents with little control over their local environment (Wilson, 1996). Moreover, not all types of social networks lessen crime. Criminal networks, such as gangs and drug networks, generate significant crime and offset more efficacious networks (Venkatesh, 2000). Likewise, nonreciprocal, draining networks, can increase social isolation for those who eschew the burden of such relationships (Kleit, 2010).

Institutional Support Mechanisms

Social networks serve as the linchpin of the social disorganization model (Morenoff et al., 2001). However, social networks are 'necessary, but not sufficient' mechanisms of social control (Kubrin and Weitzer, 2003: 377). They work to affect crime through three key intervening mechanisms as follows: Institutional support, sense of community, and collective efficacy (see **Figure 1**). At what is often termed the parochial level (Bursik and Grasmick, 1993), social control is generated in a formalized way through engagement in local organizations (e.g., block watches and resident associations) and institutions (e.g., churches and schools). Local organizations and institutions link individuals and their social networks to broader economic and political structures (Peterson et al., 2000). Some serve as the physical places within which social networks operate. Others can provide employment thereby linking individuals to conventional values (Wilson, 1996). Moreover, neighborhoods can develop the capacity to generate social control externally at the public level by securing goods and services from larger political entities, such as police departments and city governments (Bursik and Grasmick, 1993). Differences in aggregate social capital across neighborhoods operate in this region of the model. Access to institutional support mechanisms, especially those embedded in the broader urban political economy, is determined in large part by the aggregate social capital a community can leverage (Putnam, 2005). All social capital is not, however, the same. Putnam (2005) differentiates between bonding, bridging, and linking capital. Bonding social capital exists within homogenous communities and serves to reinforce shared norms and values among like people. Bridging social capital pulls together people who may be sociodemographically diverse, but are hierarchically equivalent, creating horizontal networks that can effectively share information and leverage diverse resources for political lobbying. Finally, linking social capital connects people vertically across socioeconomic hierarchies, tying those with direct access to social and economic resources to those without. Bridging social capital may link the interests of higher and lower income community members, thereby increasing aggregate social capital and decreasing crime. In this context, higher income residents provide the social, economic, and political leverage that will lead to institutional involvement and that this connection to institutional support mechanisms will secure crime prevention for all of the community residents, regardless of socioeconomic background (Curley, 2010). This may not occur in all socioeconomically heterogeneous communities, however. Higher income residents may choose not to interact within their proximate environment and rather

invest their institutional resources outside the neighborhood (e.g., going to churches, sending their children to schools, and frequenting businesses outside the neighborhood) (Tach, 2009). And where institutional investments emerge within mixed-income neighborhoods, they may not benefit low-income residents. Low-income residents may be left out of major decision making in a community; once enforced, decisions on zoning, crime control, and other policies may work to ostracize, isolate, or criminalize lower income residents. This process can create institutions that support some community members and disenfranchise others, thereby undermining social organization.

Collective Sense of Community

While local institutions and social networks may serve as indicators of social organization in a community, research has also examined the role of less instrumental indicators of social organization. A sense of community can emanate not just from the people (social networks) and places (local institutions) in that community, but also from the feeling of belonging that individuals possess. Having a sense of community gives residents a social psychological incentive to work together to solve common problems (Chavis and Wandersman, 1990). Research on juvenile delinquency finds that youth from neighborhoods with a strong sense of community engage in more prosocial behaviors and suggests that emotional ties to place are the best indicators of whether someone is truly a part of a community (Cantillon, 2006). This shared sense of community can also encompass the idea of a neighborhood subculture. Kubrin and Weitzer (2003) point out that while subculture was a key aspect of Shaw and McKay's (1942) understanding of community organization, this focus has been omitted from recent research.

Social networks also facilitate the creation of a strong sense of community. Sense of community can be physical (connected to the place) or relational (connected to the people) (McMillan and Chavis, 1986). Emotional ties to a certain place and/or its people give community members a psychological incentive to solve problem together, separate from the incentives of resource attainment or the solidification of shared values (Chavis and Wandersman, 1990). Ties to community can develop place-based subcultures that are beneficial in attenuating the effects of concentrated disadvantage, even in very poor communities. Living in such communities necessitates the sharing of resources, often creating a shared place-based subculture within which members of the community view sharing of responsibilities as a basic survival mechanism – for example, residents helping with childcare so single mothers can get to work. This shared place-based subculture can also create the perception of shared security despite the realities of crime (i.e., the feeling that 'everyone has your back') (Clampett-Lundquist, 2010). Mutual trust and feelings of connectedness among neighbors can provide intangible benefits. These benefits include social support and information sharing (e.g., regarding childcare, jobs, and other resources), as well as informal social controls (e.g., discouraging crime or other undesirable behaviors such as smoking or alcohol use among youths, drunkenness, littering, and graffiti) (Sampson et al., 2002). The social support in this context can create feelings of safety because community members know and look out for their neighbors. Place-based subcultures are not necessarily beneficial, however. Some research suggests that ties to community may be based on a shared oppositional subculture that emphasizes street-code conditioning, negating actions like contacting the police and curtailing crime, especially among youth (Kubrin and Weitzer, 2003).

The Role of Collective Efficacy

At the forefront of the resurgence of social disorganization theory in the late twentieth century was the concept of collective efficacy. Collective efficacy is simply the aggregate capacity of a community to work toward common goals (Sampson et al., 1997). Socially organized communities, in other words those with a sense of community, intact networks, and vibrant community institutions, have a greater capacity for collective efficacy than disorganized communities. While sense of community and institutional support can each have a direct role in mediating the relationship between concentrated disadvantage and crime, they also have an indirect role by shaping the level of collective efficacy a community possesses. Social networks, however, influence crime solely through the mediating effects of collective efficacy (Morenoff et al., 2001). Collective efficacy translates the opportunities to enact social control into the actual expression of that control. It is the process by which a sense of community, social networks, and institutional resources are mobilized in such a way as to lessen crime.

At some level, the concept of collective efficacy calls to the forefront a key component of early social disorganization theory sometimes lost in the contemporary literature. Beyond economic disadvantage, Shaw and McKay (1942) were concerned with the effects of community heterogeneity. Their hypotheses regarding the effects of immigration on crime, for example, was not due to the assumed criminality of immigrants, but to the difficulty of establishing efficacious social networks (meaning those able to generate collective efficacy) among residents who did not share common cultures, traditions, or even languages.

Areas of Continued Research

Beyond finding support for the broad tenets of social disorganization theory, extant research indicates that rates of crime increase in tandem with many other specific characteristics that create variation among communities. These characteristics represent ongoing areas of continued research on community differences in aggregate crime. Three specific areas of continued research are highlighted here: (1) the role of employment as a stabilizing force in communities, (2) research using a nuanced understanding of the poverty effect, and (3) cross-disciplinary work incorporating crime as a predictor of other aggregate-level differences across communities.

The Role of Work as a Stabilizing Force in Communities

One vein of relatively recent social disorganization research has examined employment as a mechanism of social control. Marginal employment, that which is characterized by low

wages, variable hours, few benefits, limited requisite skills, and constrained mobility, has detrimental effects on both individuals and communities (Crutchfield, 1995). Marginal employment increases criminality by failing to provide disincentives to prevent crime. Individuals with 'good' jobs avoid situations that may jeopardize those jobs, while individuals with 'bad' jobs do not. Communities where large proportions of residents are marginally employed lack the aggregate structure and stake in conformity that communities with higher quality local labor markets provide. Moreover, such communities may promote involvement in a street culture that does not value present or future employment prospects (Jacobs and Wright, 1999).

Poverty and Community Differences in Crime

Poverty has been a central explanatory variable in community crime research since its inception. The conventional ideas then, as now, about the connection between concentrated poverty and crime held that deconcentrating poverty would deconcentrate crime. In more contemporary interpretations, researchers suggest that socioeconomically diverse neighborhoods will generate socioeconomically diverse social networks. These networks will then serve to distribute economic, social, and cultural capital across communities, benefiting those who are poor and those who are not poor alike. The combination of shared resources and efficacious social networks will yield communities with less crime. Recently, however, researchers have begun to study the veracity of these assumptions by examining the formation of social networks within the context of social disorganization theory. Socioeconomically diverse communities are, by very definition, heterogeneous communities. Those living in such communities may experience difficulty establishing efficacious social networks (meaning those able to generate collective efficacy) with others who share their same physical space, but who may not share similar life chances. The question then becomes whether or not heterogeneous social networks can be efficacious networks. Do the protective benefits of decreased, or diluted, disadvantage outweigh the criminogenic risks of heterogeneous social networks?

Crime as an Independent Variable

Finally, having established community context matters, research on neighborhood effects has recently taken a 'process turn' toward explaining *why* communities matter in explaining individual behaviors (Morenoff, 2003: 979). In such research, crime often becomes an independent variable instead of a dependent variable. This perspective has been widely adopted in the public health literature, where scholars suggest that one of the primary ways that aggregate community conditions influence numerous health disparities is by generating stress, citing criminal victimization and the fear of victimization as key stressors. But again, efficacious social networks within communities appear to work to temper the effects of stress (Diez Roux, 2004). While stress-inducing community characteristics, such as crime, can generate negative health outcomes, these effects can be offset by positive social relationships and active civic engagement among residents. This is but one of many examples where the basic precepts of social disorganization theory have been integrated into explanations of outcomes other than aggregate community crime, with criminal victimization and/or fear of criminal victimization operating as predictor variables.

Crime and Different Types of Communities

Just as the scope of the social disorganization theoretical framework has been expanded to include more broadly defined variables, its application has also been expanded to include analyses of specific types of communities. Three examples of research on such specific communities include research on crime in rural communities, public housing communities, and immigrant communities.

Crime in Rural Communities

Much research on communities and crime assumes that since crime is more prevalent in urban communities, the social disorganization framework does not hold for other types of communities, working from the assumption that individuals in urban communities are less likely to maintain social networks, join community organizations, and exert social control over peers. The result is the perspective that urban communities are by definition characterized by a greater degree of social disorganization than rural or suburban communities and hence exhibit greater rates of crime. Despite comparatively sparse research on the role of community correlates in shaping crime in rural communities, there are many reasons to expect some rural communities to mirror socially disorganized urban communities.

Recent demographic changes have created new stressors and strained existing rural resources. While rural areas in many industrialized nations have gained population in recent years, both those areas that experience population growth and those that experience population losses face significant challenges as a result. Areas gaining population have had to deal with an influx of new residents who differ from existing residents across numerous dimensions, not the least of which is ethnicity (Johnson, 2003). In particular, the growth of the minority population has changed the face of rural communities in many countries. Conversely, areas losing population have confronted the problem of maintaining community cohesion as the size of the community has shrunk (Johnson, 2003). Both of these contradictory trends have increased community stressors and strained community resources in the areas they have afflicted.

Economic changes have created even greater challenges for rural communities. Two primary economic changes confronted rural communities during the latter part of the twentieth century. The first was the farm crisis of the 1980s. Continuing the boom–bust cycles that have historically typified rural agriculture, the complex interplay of debt, inflation, and global commodities markets cost many rural families the farm, quite literally (Barnett, 2003). For some rural areas in the United States, however, the losses experienced as the result of the farm crisis were offset by employment gains due to the deindustrialization of urban areas. As manufacturing firms left the United States' industrial rust belt, rural areas, particularly in the south,

enticed firms to relocate with low taxes, inexpensive land, and right-to-work laws only to see these same firms leave the United States altogether in an effort to secure lower labor costs in other parts of the globe.

Both demographic changes and economic changes reshaped rural community institutions, including local governments, families, and religious organizations. The ability of local governments varied across rural areas in their ability to meet demand for services that may have been protective against crime (Duncan, 1999). The capacity of other social institutions to meet this demand also varied. While churches and high levels of religiosity have often been seen as protective factors in rural communities, declines in church attendance led to the decline of the church as a crime-protective institution in rural communities. Finally, the structure of rural families has changed in light of late-twentieth-century demographic and economic shifts. Matching their urban counterparts, rural families face high levels of divorce and single parent households (Lichter et al., 2003), both associated with less supervision of youth and more juvenile delinquency.

Crime in Public Housing Communities

Crime is a serious problem in public housing, with crime rates estimated to be as much as 80% greater in public housing projects than in surrounding neighborhoods (Holzman et al., 2001). Given the violence endemic to some public housing, ameliorating the crime problem in such places has been a central component of policy aimed at eliminating public housing projects or estates in order to break up concentrated pockets of poverty in socially disorganized communities and disrupt cycles of intergenerational poverty. Evidence from three large scale studies in the United States provides mixed support for this policy perspective. These studies are (1) the Gautreaux Assisted Housing Program; (2) the Moving to Opportunity (MTO) for Fair Housing Program; and (3) the Housing Opportunities for People Everywhere (HOPE VI) Program.

The Gautreaux Program moved thousands of project-based public housing residents to suburban and urban nonpoor neighborhoods. In terms of public safety, Gautreaux participants reported less violence and crime, safer streets, and better schools in their new neighborhoods; but residents that moved to the suburbs lost their previous social networks and institutional support systems (Rubinowitz and Rosenbaum, 2000). This became the primary reason for participants to leave the program, choosing to stay close to what and who they knew even if it was a neighborhood with higher crime (Pashup et al., 2005).

The Moving to Opportunity study used an experimental design to examine the causal effects of neighborhood disadvantage on a range of outcomes among public housing residents who relocated voluntarily to other types of communities. Families in a treatment group were given housing subsidy vouchers and required to move into low poverty neighborhoods, those in a second group were given vouchers but no stipulation on where they could move, and those in a control group remained in public housing. MTO findings have been mixed. Treatment group members reported less fear of crime, but no significant improvements for a host of other outcomes, including risk-taking behaviors (Ludwig et al., 2001; Orr et al., 2003).

Finally, the HOPE VI Program sought to reinvent public housing by demolishing large, spatially concentrated developments and replacing them with mixed-income housing, with the objective of deconcentrating poverty (Popkin et al., 2004; Smith, 2002). To assess the impact of this program, researchers tracked a sample of relocated public housing residents from developments that were demolished and rebuilt as mixed-income (Popkin et al., 2004). Like MTO, findings have been mixed, but overall relocation was beneficial for residents, particularly in terms of residing in safer neighborhoods (Popkin et al., 2004). Thus, in sum, research on public housing communities in the United States finds that moving into other types of communities leads to the experience of less criminal victimization and less fear of victimization, but comes at the expense of supportive social networks and engagement in bridging community ties.

Crime in Immigrant Communities

Finally, the origin of community studies of crime developed when the founders of the Chicago School of Sociology observed the social consequences of rapid immigration during the first half of the twentieth century (Park and Burgess, 1924; Shaw and McKay, 1942). Although, even during this period of rapid immigration and pervasive antiimmigrant sentiment, data did not indicate a clearly positive relationship between immigration and crime (Hourwich, 1912), the common assumption was often that immigration leads to more crime across communities.

Contrary to this assumption, contemporary scholarship on crime in immigrant communities repeatedly finds either no relationship or an inverse relationship between levels of immigration and levels of crime. Such relationships are observed in numerous United States cities, including Miami (Nielsen and Martinez, 2009), Chicago (Vélez, 2009), and New York City (Davies and Fagan, 2012). Immigrant settlement in inner-city areas has often occurred in those areas that experienced dramatic population declines and economic disinvestment of the 1970s and 1980s. Consequently, it appears that immigration reduces aggregate levels of crime, in part, by revitalizing urban neighborhoods (Reid et al., 2005).

Summary and Conclusions

Understanding why communities exhibit different levels of crime requires understanding the very fabric of communities. Poor communities are not necessarily violent communities. Nor are immigrant communities, urban communities, or any other broad type of community. To understand variation in crime rates across communities, one has to examine how the combined influence or social networks, institutional support mechanisms, sense of community, and collective efficacy operate within a community to create social environments that are either conducive to or protective against crime.

See also: Collective Efficacy and Crime; Crime and Class; Crime in the Rural Context; Crime, Geography of; Crime, Race and Ethnicity; Crime, Sociology of; Immigration and Crime in the United States.

Bibliography

Almgren, G., 2005. The ecological context of interpersonal violence: from culture to collective efficacy. Journal of Interpersonal Violence 20, 218–224.

Barnett, B.J., 2003. The U.S. farm financial crisis of the 1980s. In: Adams, J. (Ed.), Fighting for the Farm. University of Philadelphia Press, Philadelphia, PA, pp. 160–171.

Bellair, P.E., 1997. Social interaction and community crime: examining the importance of neighbor networks. Criminology 35, 677–704.

Bursik, R.J., 1988. Social disorganization and theories of crime and delinquency: problems and prospects. Criminology 26, 519–551.

Bursik Jr., R.J., Grasmick, H.G., 1993. Neighborhoods and crime: the dimensions of effective community control. Lexington Books, Lanham, MD.

Cantillon, D., 2006. Community social organization, parents, and peers as mediators of perceived neighborhood block characteristics on delinquent and prosocial activities. American Journal of Community Psychology 37, 111–127.

Chavis, D.M., Wandersman, A., 1990. Sense of community in the urban environment: a catalyst for participation and community development. American Journal of Community Psychology 18, 55–81.

Clampett-Lundquist, S., 2010. 'Everyone had your back': social ties, perceived safety, and public housing relocation. City & Community 9, 87–108.

Crutchfield, R.D., 1995. Ethnicity, labor markets and crime. In: Hawkins, D. (Ed.), Ethnicity, Race, and Crime. State University of New York Press, Albany, NY, pp. 194–211.

Curley, A.M., 2010. Relocating the poor: Social capital and neighborhood resources. Journal of Urban Affairs 32, 79–103.

Davies, G., Fagan, J., 2012. Crime and enforcement in immigrant neighborhoods: evidence from New York City. Annals of the American Academy of Political & Social Science 641, 99–124.

Diez Roux, A.V., 2004. The study of group-level factors in epidemiology: rethinking variables, study designs, and analytical approaches. Epidemiologic Reviews 26, 104–111.

Duncan, C.M., 1999. Worlds Apart: Why Poverty Persists in Rural America. Yale University Press, New Haven, CT.

Granovetter, M., 1973. The strength in weak ties. American Journal of Sociology 78, 1360–1380.

Holzman, H.R., Hyatt, R.A., Dempster, J.M., 2001. Patterns of aggravated assault in public housing: Mapping the nexus of offense, place, gender, and race. Violence Against Women 7, 662–684.

Hourwich, I.A., 1912. Immigration and crime. American Journal of Sociology 17, 478–490.

Jacobs, B.A., Wright, R., 1999. Stick-up, street culture, and offender motivation. Criminology 37, 149–174.

Johnson, K.M., 2003. Unpredictable directions of rural population growth and migration. In: Brown, D., Swanson, L. (Eds.), Challenges for Rural America in the Twenty-first Century. The Pennsylvania State University Press, University Park, PA, pp. 19–31.

Kasarda, J.D., Janowitz, M., 1974. Community attachment in mass society. American Sociological Review 39, 328–339.

Kleit, R.G., 2010. Draining ties: tie quality versus content in low-income women's social networks when displaced by redevelopment. Journal of Social and Personal Relationships 27, 573–588.

Kornhauser, R., 1978. Social Sources of Delinquency. University of Chicago Press, Chicago.

Krohn, M.D., 1985. The web of conformity: a network approach to the explanation of delinquent behavior. Social Problems 33, S81–S89.

Kubrin, C.E., Weitzer, R., 2003. New directions in social disorganization theory. Journal of Research in Crime and Delinquency 40, 374–402.

Lichter, D.T., Roscigno, V.J., Condron, D.J., 2003. Rural children and youth at risk. In: Brown, D., Swanson, L. (Eds.), Challenges for Rural America in the Twenty-first Century. The Pennsylvania State University Press, University Park, PA, pp. 97–108.

Ludwig, J., Duncan, G.J., Hirschfield, P., 2001. Urban poverty and juvenile crime: evidence from a randomized housing-mobility experiment. Quarterly Journal of Economics 116, 655–680.

McMillan, D.W., Chavis, D.M., 1986. Sense of community: a definition and theory. Journal of Community Psychology 14, 6–23.

Morenoff, J.D., 2003. Neighborhood mechanisms and the spatial dynamics of low birth weight. American Journal of Sociology 108, 976–1017.

Morenoff, J.D., Sampson, R.J., Raudenbush, S.W., 2001. Neighborhood inequality, collective efficacy, and the spatial dynamics of urban violence. Criminology 39, 517–558.

Nielsen, A.L., Martinez Jr., R., 2009. The role of immigration for violent deaths. Homicide Studies 13, 274–287.

Orr, L., Feins, J.D., Jacob, R., Beecroft, E., Sanbonmatsu, L., Katz, L.F., Liebman, J.B., Kling, J., 2003. Moving to Opportunity for Fair Housing Demonstration Program: Interim Impacts Evaluation. U.S. Department of Housing and Urban Development, Washington.

Park, R.E., Burgess, E.W., 1924. Introduction to the Science of Sociology. University of Chicago Press, Chicago, IL.

Pashup, J., Edin, K., Duncan, G., Burke, K., 2005. Participation in a residential mobility program from the client's perspective: findings from Gautreaux Two. Housing Policy Debate 16, 361–392.

Peterson, R.D., Krivo, L.J., Harris, M., 2000. Disadvantage and neighborhood violent crime: do local institutions matter? Journal Research in Crime and Delinquency 37, 31–63.

Popkin, S.J., Levy, D.K., Harris, L.E., Comey, J., Cunningham, M.K., Buron, L.F., 2004. The hope vi program: what about the residents? Housing Policy Debate 15, 385–414.

Putnam, R., 2005. Bowling Alone: The Collapse and Renewal of American Community. Simon and Schuster, New York.

Reid, L.W., Weiss, H.E., Adelman, R.M., Jaret, C., 2005. The immigration-crime relationship: evidence across us metropolitan areas. Social Science Research 34, 757–780.

Rubinowitz, L., Rosenbaum, J.E., 2000. Crossing the Class and Color Lines: From Public Housing to White Suburbia. University of Chicago Press, Chicago, IL.

Sampson, R.J., Raudenbush, S.W., Earls, F., 1997. Neighborhoods and violent crime: a multi-level study of collective efficacy. Science 227, 8–924.

Sampson, R.J., Groves, W.B., 1989. Community structure and crime: testing social-disorganization theory. American Journal of Sociology 94, 774–802.

Sampson, R.J., Morenoff, J.D., Gannon-Rowley, T., 2002. Assessing 'neighborhood effects': social processes and new directions in research. Annual Review of Sociology 28, 443–478.

Shaw, C.R., McKay, H.D., 1942. Juvenile Delinquency and Urban Areas: A Study of Rates of Delinquency in Relation to Differential Characteristics of Local Communities in American Cities. University of Chicago Press, Chicago, IL.

Smith, R., 2002. Housing Choice for HOPE VI Relocatees. The Urban Institute, Washington, DC.

Tach, L., 2009. More than bricks and mortar: neighborhood frames, social processes, and the mixed-income redevelopment of a public housing project. City & Community 8, 269–299.

Vélez, M.B., 2009. Contextualizing the immigration and crime effect: an analysis of homicide in Chicago neighborhoods. Homicide Studies 13, 325–335.

Venkatesh, S., 2000. American Project: The Rise and Fall of a Modern Ghetto. Harvard University Press, Cambridge, MA.

Wilson, W.J., 1996. When Work Disappears: The World of the New Urban Poor. Knopf, New York, NY.

Communities of Practice

Jacquelin McDonald, Learning and Teaching Services (LTS), The University of Southern Queensland, Toowoomba, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E. Wenger, volume 4, pp. 2339–2342, © 2001, Elsevier Ltd.

Abstract

Communities of practice (CoPs) have emerged as an influential concept to foster both individual and group knowledge within a social learning system. They have three fundamental elements: a 'domain' of activity or knowledge that creates a sense of common identity, the 'community' who are the members who care about that domain, and the shared 'practice' developed by community member. CoPs are established or emerging in many educational, government, and industry contexts. Leadership has been identified as an important success factor, and social media and mobile devices now are generating a range of opportunities for members to engage in their CoPs.

What Are Communities of Practice?

As social beings we are involved in many communities, such as professional associations, community or neighborhood watch, religious groups, or hobbies, such as a garden club. However, not all such communities are communities of practice (CoPs). A community of practice (CoP) is a unique combination of three fundamental elements: a 'domain' of activity or knowledge that creates a sense of common identity, the 'community' who are the members who care about the domain, and the shared 'practice' the community members develop to be effective in the domain (Wenger, 1998). The term 'communities of practice' emerged from Lave and Wenger's (1991) study that investigated the apprenticeship model learning, which showed that, rather than the one-on-one master/apprentice model, it was a complex set of social relationships within the whole practicing community that supported learning and membership of the knowledge domain and practice field. The identification of CoPs as a concept and social system (Wenger, 2010) that can be identified and cultivated has created a mechanism to study both individual and social learning behaviors across a range of contexts. The CoP concept is based on the study of the social nature of human learning inspired by anthropology and social theory (Lave and Wenger, 1991; Vygotsky, 1978; Wenger, 2010). Since 1991 Etienne Wenger's ongoing practice, research, and publications have provided the foundation for much of the growing investigation and implementation of CoPs in many contexts, including health care, education, local, national, and international government and business.

CoPs and Social Learning Systems

Thinking about CoPs as social learning systems enables us to reflect on both individual and group learning, and the relationships, interactions, and learnings that we forge or experience within these social systems. Wenger (2010: p. 179) suggests CoPs "exhibit many characteristics of systems more generally: emergent structure, complex relationships, self-organization, dynamic boundaries, ongoing negotiation of identity and cultural meaning, to mention a few." As CoP members explore and grow their domain knowledge over a sustained period of interaction, they develop trust, a sense of identity, community, and also resources that document a shared history of learning, thus creating a dynamic social structure that is a CoP. Individuals bring their knowledge, experience, and perspectives and negotiate the meaning of domain knowledge and practice with members, thereby building both their identity within the CoP and jointly creating a unique identity for the CoP. Wenger (2010: p. 182) notes that "learning is a social becoming."This individual and group identity reflects the complex relationship between the social and the personal and locates CoPs firmly in the social and behavioral sciences.

Stages of CoP Development and the Three Elements as an Organizing Structure

A CoP may emerge from a particular idea or issue, or be triggered by interactions among colleagues working together to explore an opportunity or solve a problem. This provides the impetus to start the CoP and often one or two enthusiasts emerge as CoP leaders, facilitators, or convenors. From start-up, the CoP usually moves through a number of stages or phases. Snyder and Briggs (2003) suggest that CoPs, like all living things, grow through distinct phases of development (discovery, coalescing, maturing, stewarding, and legacy), so each phase needs to be understood to address the issues and opportunities at each phase. Wenger et al.'s (2011) CoP phases of development include establishing the potential, coalescing, active, dispersed, and memorable, and these stages have been adapted and restructured to suit the context or social system of specific CoPs. McDonald (2012) used the CoP stages – initiation, creation, infancy, maturity/sustaining, and recreating – as an organizing structure to develop leadership resources, available at (http://www.usq.edu.au/cops/resources/altcfellowship). Choosing the most appropriate approach may be challenging as resources range from discussions of the theoretical foundations to practical guidelines and templates for starting a CoP. Identifying the type of CoP to be established is the first step and scoping the specific context and support available are important considerations for CoP leaders.

One approach that has proven effective in the author's experience is the adaption of the three elements of CoPs as the organizing structure for colocated CoP meetings (McDonald and Star, 2008). It is worth noting that these meetings or gatherings are quite different from the standard business format. In a CoP priority topics and agendas are established by members and time is allocated for each of the three CoP elements – knowledge domain, practice, and community building. At the initial meeting members identify and negotiate the domain-focused topics to be addressed by the CoP, location and times to meet, potential members and stakeholders, communication processes and alignment (or not) with institutional goals and mission. The elements of *domain* of knowledge, such as first year students, mobile learning or research supervision, establish the common focus and sense of common identity, the *community* consists of members involved in that domain and who meet over time and a sense of community, and the shared *practice* is what members bring to, and develop together, to be effective in the domain.

A common time allocation for physical CoP meetings is either 1.5 or 2 h, with approximately 30 min allocated for each element. The allocation of additional time to share emerging issues, such as a round of 2 min at the start for each member to share 'what's on top,' the current priority in their practice, allows time for items not set as a priority at CoP start-up. The unstructured community time, with quality refreshments, provides a time for informal networking and knowledge exchange. This is a commitment to build social learning and community bonding which is never usually factored into traditional meeting agendas. Community time provides the foundation for building the social learning system. Items for the domain and practice elements are usually identified by members at start-up, and may involve invited domain speakers or members presenting in their area of expertise, then the members share how they implement activities in their practice. For example, the meeting focus may be on implementing a student group learning process, so someone with a proven track record is invited to build the domain knowledge of members, and in the sharing practice time members share their experiences and plan applications of the new knowledge. This knowledge is captured and made available as a permanent record of CoP activity and as a resource for members. This approach has proven to be a powerful and sustainable application of CoP theory to higher education practice, and was noted as a 'unique approach' by Etiienne Wenger (2008), and is also being applied in digital communities (Reushle, 2011).

CoPs Established and Emerging in Many Contexts

CoPs are well established in business, government, healthcare, and international and national projects such as World Bank and World Health, with much of the early CoP literature (Wenger, 1998; Wenger et al., 2002, 2009) relating to CoPs in these contexts. Each context will influence the structure, operation, and impact of the CoP, depending on the specific cultural, political, and institutional environment. More recently, CoPs have emerged as an approach to foster learning communities in education, providing some dilemmas as CoPs, with their more informal, member-driven, social learning system, cut across established, hierarchal, and institutionalized educational systems. Issues of power, competency, and legitimacy can cause tension between the CoP and traditional, institutional structures, as knowledge emerging from the 'coal-face' practice of members challenges the top-down, institutionalized knowledge and systems.

In Australia higher education CoPs are gaining momentum as a way of creating social learning communities to support educators working to improve learning and teaching in, including a number of projects sponsored by the Australian Government Office of Learning and Teaching (OLT). An example is the OLT funded project that focused on clarifying and building the leadership capacity of facilitators of learning and teaching CoPs (McDonald et al., 2012).

Current Discourse around CoPs

The current interest in CoPs, and their implementation in various contexts, has generated debate about what does and does not constitute a CoP. The initial concept identified by Lave and Wenger (1991) described CoPs as spontaneous or organic, emerging from shared interest, practice, and ongoing interaction, and an organic CoP is held up by some (Stamps, 2000; Roberts, 2006) as the only 'true' CoP. However, CoPs as a simple social system are found in all contexts, and as their value in building individual and institutional knowledge becomes more widely recognized (Cox, 2006; Iaquinto et al., 2011; McDonald et al., 2012; Wenger et al., 2002; Wenger, 2010), there is increasing interest in supporting emerging or existing CoPs, and also cultivating 'intentional' CoPs. McDonald et al. (2012: p. 22) found that there are three types of higher education CoPs, as noted in **Table 1**.

The debate about what constitutes a 'real' CoP is a result of the shift from an analytical title used to define CoPs as a concept, to a more practical, or instrumental, application of

Table 1 Types of community of practice (CoP) and contextual issues

Type of CoP	*Organic*	*Nurtured/supported*	*Created/intentional*
Structure	Bottom-up	Modified bottom-up	Top-down
Support level	Minimal	Subsidized	Provided
Membership	Voluntary	Voluntary/suggested	Encouraged
Themes	Discipline-related	Discipline or issue related	Guided issues and cross discipline
Agenda	Self-determined	Self-determined/steered	Guided theme
Timing for outcomes	Self-determined	Self-determined and funding related	Short-term rather than long-term

the concept. Wenger (2010: p. 192) says "the concept of community of practice started out as an analytical concept, giving a name to a phenomenon that already existed. Now it is often used with an intention to create, cultivate, or capitalize on the process – almost as a technique." There is a danger that the uninformed, instrumental application of the concept may lead to a dilution of the concept (Hughes, 2007) where it is used to serve top-down corporate needs, rather than to create a social learning system. However, much of the debate seems to be about the different types of CoPs (Burch et al., 2012), reflecting the tension between CoP theory and how it has evolved in practice, rather than what actually constitutes a CoP. While there has been obvious advantages to cultivating CoPs to grow and support individual and institutional learning systems (McDonald and Star, 2008), in the instrumental application of the concept, there may also be a mismatch between the informal, peer-to-peer (horizontal) structure, with a member-initiated agenda and focus, and the top-down, hierarchical (vertical) and instrumental imperatives. Langelier (2005: pp. 16–17) discusses the subtle distinction between integration and institutionalization of CoPs, suggesting that a distinction needs to be made between the institutionalization of CoPs themselves, and the institutionalization of their existence in the institution. If alert to this dilemma, CoP members can ensure the integrity of the CoP by communicating the individual and institutional benefit of the CoP, and by educating management about the unique operation and value of CoPs. On a more positive note, even the instrumental introduction of CoPs will create dialogue around the way both individuals and the institutional managers view and foster learning, opening the door for the integration of horizontal (CoP) and vertical (institutional) social learning systems.

Role of Leadership of CoPs

Effective leadership of CoPs has been identified as an important contributor to the success and sustainability of higher education CoPs (Cox, 2006; McDonald and Star, 2008; McDonald and Palani, 2011; McDonald et al., 2012; Ortquist-Ahrens and Torosyan, 2008; Wenger, 1998). Wenger et al. (2009: p. 10) note that "leadership is an essential ingredient in a community of practice, whether formal or informal, concentrated in a few people or broadly distributed." Leadership of CoPs requires a combination of leadership attributes and skills, including the ability to support member goal and activities to build an effective social learning system, as well as positioning the CoP to succeed within the broad institution context. It will also require knowledge of the theory of CoPs so the leader can identify the type of CoP and particular skills and leadership attributes required for different stages as the community evolves within its broad context.

Where Are CoPs Heading in Our Digitally Connected World?

Social media and new mobile technologies provide opportunities to engage in national and international CoPs, any time and any place. This creates opportunities and challenges around the viability of learning communities. Also, there is an increasing interest in individual or personal learning journeys, supported by accesses to mega amounts of information and casual interactions with others via global networks. For example the increasing availability of Massive, Online, Open Courses (MOOCs), where community building and group activities are rarely components of the learning experience. On one hand, social media can support both group and one-to-one learning typical of CoP activities, as well as provide a rich range of media and opportunities to share and communicate the learning generated with both digital and colocated CoPs. On the other hand, access to such a vast range of knowledge and networks can be daunting and draining, with learning opportunities swamped by time consuming internet surfing and sense of aloneness in digital world. As individuals we decide how we will focus our participation in social learning systems; it may be by a leadership role, contributing as full or casual members or surfing through learning environment, picking up or sharing ideas as we travel our learning journey. Wenger (2010: p. 196) puts forward the idea of 'learning citizenship' and suggests that to contribute to the learning capability of social learning systems is 'a personal responsibility that comes with social participation,' which will have a positive impact on the learning capability of both personal and social systems.

See also: Leadership Education: Theory and Practice; Leadership in Organizations, Sociology of.

Bibliography

Burch, T., McDonald, J., Nagy, J., Star, C., Cox, M., Margetts, F., Collins, E., 2012. Identifying, building and sustaining leadership capacity for communities of practice in higher education, Literature review. http://www.cops.org.au/literature_review.pdf (viewed 14.05.13.).

Cox, M., 2006. Phases in the development of a change model: communities of practice as change agents in higher education. In: Hunt, L., Bromage, A., Tomkinson, B. (Eds.), The Realities of Change in Higher Education: Interventions to Promote Learning and Teaching. Routledge, London and New York, pp. 91–100.

Duguid, P., 2008. The art of knowing: social and tacit dimensions of knowledge and the limits of the community of practice. In: Amin, A., Roberts, J. (Eds.), Community, Economic Creativity and Organization. Oxford University Press, New York, pp. 69–89.

Hughes, J., 2007. Lost in translation: the journey from academic model to practitioner tool. In: Hughes, J., Jewson, N., Unwin, L. (Eds.), Communities of Practice: Critical Perspectives. Routledge, London and New York.

Iaquinto, B., Ison, R., Faggian, R., 2011. Creating communities of practice: scoping purposeful design. Journal of Knowledge Management 15 (1), 4–21.

Langelier, L., 2005. Work, Learning and Networked: Guide to the Implementation and Leadership of Intentional Communities of Practice. Centre francophone d'informatisation des organisations (CEFRIO), QC, Canada.

Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, Cambridge.

McDonald, J., 2010–11. Community, Domain, Practice: Facilitator Catch Cry for Revitalising Learning and Teaching through Communities of Practice. Australian Learning and Teaching Council (ALTC) Teaching Fellowship. http://www.usq.edu.au/cops/resources/altcfellowship/facilitator-resources.

McDonald, J., 2012. Australian Learning and Teaching Council Fellowship, Community of practice leadership resources. http://www.usq.edu.au/cops/resources/altcfellowship (accessed 14.11.12).

McDonald, J., Star, C., 2008. The challenges of building an academic community of practice: an Australian case study. In: Barrow, M., Sutherland, K. (Eds.), Engaging Communities. Proceedings of the 31st Annual HERDSA Conference, pp. 230–240. Rotorua, New Zealand. http://www.herdsa.org.au/?page_id=273.

McDonald, J., Palani, A., July 4–7, 2011. Building leadership capacity for community of practice facilitators: edgy professional development. In: Krause, K., Buckridge, M., Grimmer, C., Purbrick-Illek, S. (Eds.), Research and Development in Higher Education: Reshaping Higher Education, vol. 34, pp. 198–206. Gold Coast, Australia. http://www.herdsa.org.au/wp-content/uploads/conference/2011/papers/HERDSA_2011_McDonald.PDF.

McDonald, J., Star, C., Burch, T., Cox, M., Nagy, J., Margetts, F., 2012. Final report: identifying, building and sustaining leadership capacity for communities of practice in higher education, project web site. Office of Learning and Teaching. Australia. http://www.olt.gov.au/project-identifying-building-and-sustaining-leadership-capacity-communities-practice-higher-education.

Ortquist-Ahrens, L., Torosyan, R., 2008. The role of the facilitator in faculty learning communities: paving the way for growth, productivity, and collegiality. Learning Communities Journal 1 (1), 29–62.

Reushle, S.E., 2011. Collaborative Community Mentoring Program (C^2MP) – addressing the changing demands of a professional community of learners through engaging digital networks. In: Williams, G., Brown, N., Pittard, M., Cleland, B. (Eds.), Changing Demands, Changing Directions. Proceedings Ascilite Hobart 2011.

Roberts, J., 2006. Limits to communities of practice. Journal of Management Studies 43 (3), 623–639.

Stamps, D., 2000. Communities of practice: learning is social. Training is irrelevant? In: Lesser, E.L., Fontaine, M.A., Slusher, J.A. (Eds.), Knowledge and Communities. Butterworth-Heinemann, Woburn, MA, pp. 53–64.

Snyder, W., Briggs, X., 2003. Communities of practice: a new tool for government managers. Collaboration Series. IBM Center for the Business of Government, Arlington, VA.

Snyder, W., Wenger, E., 2010. Our world as a learning system: a communities-of-practice approach. In: Blackmore, E. (Ed.), Social Learning Systems and Communities of Practice. Springer, London, pp. 124–144.

Vygotsky, L., 1978. Mind in Society: Development of Higher Psychological Processes. Harvard University Press.

Wenger, E., 1998. Communities of Practice: Learning, Meaning, and Identity. Cambridge University Press, New York.

Wenger, E., McDermott, R., Snyder, W., 2002. Cultivating Communities of Practice. Harvard Business School Press, Boston.

Wenger, E., July 1–4, 2008. Keynote presentation at engaging communities. In: The 31st Annual HERDSA Conference. Rotorua, New Zealand.

Wenger, E., White, N., Smith, J., 2009. Digital Habitats: Stewarding Technology for Communities. CPSquare, Portland.

Wenger, E., 2010. Communities of practice and social learning systems: the career of a concept. In: Blackmore, C. (Ed.), Social Learning Systems and Communities of Practice. Springer Verlag and the Open University.

Wenger, E., Trayner, B., de Laart, M., 2011. Promoting and assessing value creation in communities and networks: a conceptual framework. Ruud de Moor Centrum, Open Universiteit Nederland Rapport 18.

Relevant Websites

McDonald, J., 2012. Australian Learning and Teaching Council Fellowship, Community of Practice Leadership Resources. http://www.usq.edu.au/cops/resources/altcfellowship (accessed 14.11.12.).

CPSquare, The Community of Practice on Communities of Practice. http://cpsquare.org/ (accessed 14.11.12.).

Etienne Wenger-Trayner Home Page http://wenger-trayner.com/ (accessed 14.01.13.).

Identifying, Building and Sustaining Leadership Capacity for Communities of Practice in Higher Education, OLT Leadership Project. www.cops.org.au/ (accessed 14.11.12.).

Community and Media

Gustavo S Mesch, University of Haifa, Haifa, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the role of media on the local community from three perspectives. One approach examines how different media outlets affect the identification and participation of residents in the local community. The second examines the characteristics of online communities and its main features. The third approach looks at the emerging role of online geographically based communities on the social integration of residents in the local community. Directions for future research are provided.

Concern about the role of the media in the formation and maintenance of the local community has been a central topic of interest in the social sciences since the early years of the twentieth century. Scholars associated with the Chicago School of Sociology were pioneers in the field. Generally speaking, the study of the association between the use of media and social integration in the community has focused on three different areas. One line of research has followed the tradition of urban studies, focusing on the role of the press and social media in the involvement in and attachment to the local community. A second line of study began in the early days of Internet adoption and transitioned to the study of online groups and virtual communities that exist regardless of physical propinquity. Finally, a more recent development has been the understanding of the role of locally based online communities for the process of social integration in the local community.

This article will provide a concise review of some of the theoretical frameworks and research findings based on these three distinct but complementary traditions.

The Media and the Local Community

In the early years of the twentieth century, studies in sociology and geography were interested in the characteristics of media audiences and the effects of media consumption on the local community. Robert Park (1922), studying the US as an immigrant society, investigated the role of the community press in identity formation among immigrant groups. He observed that newspaper reading was more characteristic of residents of urban areas than rural areas and that urban residents were interested in news from outside the area. Later, Robert Merton (1957) distinguished between localites and cosmopolitans. Localites were more oriented toward the local community, engaged in more locally based social activities and they saw the local newspaper as the source of local news. In contrast, cosmopolites were concerned with social activities on a broader scale and were more likely to consume media outside their local area. Janowitz (1952) then made an important contribution to the study of the role of media and community structure by investigating the association among media, a community's social organization and attachment to place. The systemic model of community attachment ascribes a central role to the process of socialization in the local community, participation in locally organized activities and involvement in local social networks. The model conceptualizes community as resulting from residents working together on collective problems. Participation and involvement in local issues vary depending on the stakes in local issues and involvement in local social ties. In this model, newspapers and the local media play a central role in integrating individuals into the community and contribute to the development of a shared belief system of values and norms supporting the process of socialization and social control in the local community. Janowitz found that community integration and involvement in local issues are associated with greater attention to the local press.

A serious problem in identifying the role of the press and media in a community's social organization is determining the direction of the effects. In other words, does local media consumption lead to social integration, or are the more active and involved residents also more likely to consume local media? More recent attempts have tried to clarify this issue. Stamm et al. (1997) developed a model that attempted to understand (1) the role of the media in social integration into the community and (2) the role of social integration into the community on media consumption. These relationships have become complex because of the effects of globalization, geographical mobility, suburbanization, and the increasing scale of human activities on local affairs. In their view, there are four mechanisms that link the media to the local community (Stamm et al., 1997). First, the power of local media lies in its ability to identify problems that need community attention and bring them to the attention of the public. Identifying local problems and expanding awareness helps community members become part of the instrumental solution to these problems. Second, the local media help identify local groups and those with influence, and bring them together. By identifying the key actors and issues, media stories support the efforts to mobilize local residents for action. Third, the local media can identify individuals with specific skills and specialties. Finally, by reading the local newspapers, residents have time to think about local problems. In his study, Stamm (1985) found that local media consumption was associated with community integration, but this correlation was modest. Furthermore, there was an association between interpersonal communication, community integration, and collective action. Local media consumption was associated with community

 http://dx.doi.org/10.1016/B978-0-08-097086-8.95074-8

integration through an increase in locally based interpersonal communication, a central feature of neighboring.

Regarding the fundamental issue of whether communities are a result of newspaper readership, or newspaper readership is a determinant of community ties, Stamm et al. (1997) concluded that most likely, the relationship is cyclical and works both ways. Those who read newspapers tend to be more connected to the local community, and those connected to the local community are likely to read their local newspapers.

A common assumption in the literature is that people are linked to their community through the local media that inform them about their environment and interpersonal networks that link them to neighbors, offering them the opportunity to share their political beliefs, and exposing them to the beliefs of others. Nevertheless, in the wake of the major changes that have occurred including media convergence and globalization, the local media has changed as well. In many cities in Western countries, the local press has retained a print edition and added an online edition. Despite the large-scale changes, these media are important communication channels for locals as well as national political actors. In their study, Skogerbø and Winsvold (2011) focused on the role of newspapers as they move to online platforms. Their findings show that print editions are still read more than online newspapers, confirming the continuing dominance of the print versions. Fewer people use the online newspapers as their primary source of information. However, both the general and elite audiences stated that online newspapers were a valuable addition for rapid updating and searching for news and historical information. Skogerbø and Winsvold also found socioeconomic differences between readers of the online and print editions. Print and online audiences read different types of content according to their age, social background, gender, and political activity. Regional and local newspapers were still viewed as mediators of identity and belongingness to the place, but there were differences between online and print audiences. Community attachment and reading the local newspaper correlate with one another. Thus, the local online audiences are different from the print audiences in many other respects too. The two platforms seem to evoke different uses and preferences among audiences, suggesting that people regard online and printed newspapers as different media.

Another major conceptual approach to the link between the media and the local community is the communication infrastructure theory (Ball-Rokeach et al., 2001). According to this approach, a 'media system' is an information system that individuals access in order to obtain the information that is necessary for achieving their goals. Communication infrastructure theory defines the media as including macro- and mesolevels. The macrolevel includes mainstream television and newspapers. The meso level includes local or ethnic media, including new communication technologies. The communication infrastructure is defined as a storytelling system set in a communication action context (Ball-Rokeach et al., 2001).

The storytelling system covers multiple levels, and each level is defined by the stories that are told by each actor. If a television station broadcasts stories that cover a nation, it is considered macrolevel media. On the other hand, if another television station focuses on stories about a local area, it is considered mesolevel media. Furthermore, this perspective includes not only macrofactors, but also local conditions such as the existence of public facilities, schools, and technological infrastructures that influence individuals and communications. The community infrastructure approach examines the communication infrastructure with an eye to two central factors. The first is the neighborhood storytelling network, which consists of three key storytellers: (1) the residents and their network of family, friends, and neighbors; (2) the community and nonprofit organizations that are located in the neighborhood and serve its residents; and (3) the media that target a particular geographic area. When these storytellers create a conversation about the neighborhood – its problems, opportunities, and events – people have a sense of belonging to a community. According to this view, storytelling is the basic way that communities are created.

The second component of the communication infrastructure is the communication action context, or the communication environment in which storytelling occurs. This environment may either constrain or facilitate the storytelling network. It consists of many elements, including the cultural diversity of the community, work conditions, the schools, libraries, parks, and other public spaces, the services available in a community such as health care services, retail stores and recreational facilities, the transportation system, and technological resources, such as Internet access. This last element is a key factor in this theoretical framework. Researchers have explored the nature of individuals' connections to Information communication and technology (ICT) and how those connections are embedded in and affected by the complex and multilevel communication infrastructure in which we live (Ball-Rokeach et al., 2001; Jung et al., 2001). According to the communication infrastructure theory, local communities differ in their communication opportunity structures. Some environments provide easy connections to necessary and useful communication channels, but others make it difficult to access such channels. Differences in communication opportunity structures are associated with differential access to social, political, cultural, and social capital resources (Kim et al., 2006). For example, Kim et al. (2006) found that the Internet was more integrated into the communication infrastructure of white neighborhoods than immigrant neighborhoods. Their study suggests the possibility that in ethnic neighborhoods, ICT creates a 'magnifying glass effect.' While it does not deter people from civic participation, ICT have a positive additive effect only on those people who already have connections to communication resources.

Virtual Communities

During the rapid expansion of access to and use of information and communication technologies, academic attention and interest has also focused on social groups that are not bounded by a specific geographic place. Members of such groups participate in computer-based social interactions with people they have often never met face-to-face. The participation of individuals in geographically dispersed groups through a mediated form of communication is not new. For example, at the beginning of the twentieth century, wireless radio was used to transmit long distance communications, and amateur radio

operators worldwide used this technology to chat and exchange information (Stephens, 1995). Nevertheless, the development of the Internet and World Wide Web is important because it allows reliable access across any distance in real time. A virtual community is a voluntary group of individuals with shared interests who interact with one another through computer-supported networks for a relative long period of time, exchanging sociability, social support, and resources. Through this social interaction, common ties and some social organizations develop (Driskell and Lyon, 2002).

The use of the term 'virtual community' created a controversy about whether online groups are communities and the degree to which a community can exist without a shared physical locality. In fact, conceptual models in the field of community studies emphasize shared needs and concerns over propinquity as the basis for a community. The idea of a "community without propinquity" developed by Melvin Webber (1964) maintains that individuals are members of an overlapping range of groups and create social networks that are not limited to physical location. According to this idea, a community should be one oriented toward social process, social interaction, and social networks. Rather than propinquity, the emphasis is on the access to social interactions and resources as the necessary condition for the formation of a community (Wellman and Whortley, 1990). Virtual communities have enjoyed a substantial amount of public and academic interest. Despite the nature of the communication, members of virtual communities seem able to develop intimate and personal relationships on the basis of common interests, not geographical location. Thus, an electronic space can be regarded as a shared space to which motivated actors are drawn in search of specific formal activities, such as the search for information, or informal activities such as social support and entertainment according to their particular interests.

An innovative aspect of the Internet is the proliferation of special interest groups devoted to almost any imaginable topic: music genres, sports and gardening, dating, parenting and pregnancy, chronic and life-threatening diseases, and professional interests. Online communities have become spaces for social interactions in which friendships and even close relationships are created and maintained (Mesch, 2007). Perceptions of shared interests and concerns are probably the most robust variables that the sociological literature has identified as conducive to the formation, development, and stability of interpersonal relationships. In fact, the chance of finding people in one's local community who share one's needs or interests may at times be slim. On the World Wide Web, the chances of forming a community of like-minded people are much better.

Although technology now allows interactions through audio (Internet phones) and video (Webcams), most communities rely on textual communication complemented by the use of pictures and movies. Such communication often consists of messages posted and stored on a server. Members can choose to receive the postings directly and read them at their convenience.

Virtual communities function as social spaces, providing two different broad types of resources: socioemotional support and the exchange of information and resources. These functions are not mutually exclusive, and in many communities both are performed. Participation in online communities can be extensive, and participants spend varying amounts of their daily lives involved in such activities.

Certain features of the new electronic communication have made participation in online communities attractive. The Internet can support various forms of communication including real time as well as delayed communication, person-to-person, person-to-people, and people-to-people communication. Thus, the access to information, entertainment, and social support is personalized, and provided round the clock whenever the individual needs it or seeks it. Furthermore, the communication of many with many in real time supports the kind of interaction required for the formation of a repetitive pattern of social interaction. It promotes the development of a common identity based on a shared interest, norms of behavior, and a social structure characteristic of the concept of community.

At the same time, it is important to acknowledge that community formation online does pose some problems, as we will see in the next section. First, face-to-face communication is a rich communication channel because it uses both verbal and nonverbal cues such as voice inflection, body gestures, nonverbal messages, immediate feedback, and bidirectional communication. In contrast, computer-mediated communication is devoid of personalization, making it difficult to convey personal feelings and emotions or to tailor messages to the needs of the receiver's current situation. The lack of cues can lead to misunderstandings. In order to overcome this problem, users must learn how to communicate using a shared system of symbols.

Another barrier to the formation and maintenance of virtual communities is the 'lurker' problem. While most activities in the community involve at least reading the participants' posts, not all participants contribute equally. Some lurk in the background, limiting their participation to the passive role of reading other messages but never posting their comments or original ideas. Nonactive participants in online communities become almost invisible to the other members because they leave no trace of their presence. While passive participants are part of the community in that they are often looking for information, when they constitute the overwhelming majority of the group, they endanger the group's existence. They leave the setting of the group's agenda to a few vocal people. In addition, passive participation maybe contagious, discouraging the participation of others. Some communities address this problem by restricting participation to only known and trusted individuals or creating a tiered system in which access to general postings is free to all, but more personal and sensitive material is by membership or invitation only.

In the past, another threat to virtual communities was identity deception. In any group, the purpose of the online conversation is the creation of a unique social system founded on a shared identity and norms of behavior. These features are problematic in online communities in which there are no clues about the participants' social status. Knowing the identity of others with whom one communicates is essential for understanding and evaluating the interactions. However, in the nongeographic world of virtual communities, identity is also ambiguous. Many of the basic clues about personality and social roles that we are accustomed to in the physical world are absent.

Finally, the extent to which norms are created in online communities is important both theoretically and practically. Early conceptualizations of computer-mediated communication argued that the absence of social and contextual cues about the communicators creates a serious problem when it comes to the formation of norms and even adherence to normative behavior. When indicators of social status and hierarchy are lacking, nonnormative and even flaming behavior is likely to occur. Nonnormative behavior, norm violation, and flaming are examples of negative interactions that undermine the sense of belonging to an online community. A viable community requires the formation of shared bonds that depend on messages sent to everyone simultaneously. Established norms that regulate the posting of messages, an accepted language of interaction, and even problem-solving techniques are needed. The sustainability of an online community requires sustained social interactions and the formation of shared collective norms and identity. For this reason, only a small number of online groups ultimately develop into online communities.

Thus, we need longitudinal studies about the creation, development, and decline of online communities in order to improve our understanding about the development of norms and collective identity. Longitudinal studies will provide important insights into the factors that explain why some communities survive over time while others do not. At the same time, we must acknowledge that the boundaries between online and off-line relationships are becoming blurred. As residential mobility increases, online communication helps us keep in touch with people we actually know. In the same vein, geographic communities tend to create shared electronic spaces in which to interact.

Geographically Based Online Communities

A third line of research has investigated how the use of new information and communication technologies by community members affects the local community. The use of ICT to support the local community is usually referred as community networks. A key distinction between online communities and community networks is that the former are based entirely on computer-mediated communication, while in the latter, the online communities are rooted in proximate geographic relations. In community networks, people are already part of one another's social networks and also have face-to-face interactions (Kavanaugh et al., 2005b).

Two of the major potential benefits of community networks for proximate communities are (1) increasing and promoting civic engagement and interactions among citizens; and (2) enhancing citizens' participation in community associations, political initiatives, and similar community activities (Putnam, 2000). Comparing the characteristics of online groups and community networks, Etzioni and Etzioni (1999) argued that the proper combination of both face-to-face communication and computer-mediated communication is more promising for creating a community than either of them separately. Indeed, according to communitarians, online and mobile communication may be a key element in revitalizing local communities and promoting democracy.

Unlike claims previously raised that the Internet has a negative effect on the quality of life in general, and on the social relations aspects of politics in particular, empirical evidence consistently indicates that community networks promote civic engagement and social interaction through residents' participation in community activism (Kavanaugh et al., 2005a). Nevertheless, scholars have yet to document the exact nature of the association between Internet use, social capital, and civic engagement systematically. It appears that Internet use is associated with 'participatory social capital.' Furthermore, there is a positive association between online and off-line voluntary participation in organizations. In a compelling study using cross-sectional and panel designs, Shah et al. (2005) demonstrate that online media complement traditional media in fostering political discussion and civic messaging. These two forms of political expression, in turn, promote civic participation. Similarly, in their Blacksburg Electronic Village study, Kavanaugh et al. (2005b) found a significant association between residents' community involvement and civic engagement, and the residents' effort to stay informed about community events and activities; membership in voluntary associations; level of personal activism in the community; collective efficacy in the community's ability to manage itself; and the use of the Internet for political and civic purposes.

From these studies, we learn that participation in local activities is not technologically determined. It is not Internet use in and of itself that increases participation in the local community. However, civic participation and involvement in local activities is more likely when Internet use is embedded in social relations and expectations. Individuals who are already active in the local community and interested in community life also tend to be members of and hold positions in community associations. Certainly, informed activists with memberships in multiple groups become more involved in local issues once going online, whereas nonactivists become less involved once online (Kavanaugh et al., 2005a). In other words, according to these studies, online communication is a facilitator, but not a necessary condition of civic involvement. Belonging to multiple groups or organizations facilitates the flow of information across a community and strengthens the sense of commitment to the local community. As a result, members of multiple organizations are more likely to use the Internet for local communication and to be active and involved in the community (Kavanaugh et al., 2005a).

Do community networks improve the quality of neighborhoods? The Netville study was one of the first studies to test the impact of new information and communication technologies on social relationships at the neighborhood level (Hampton, 2007). It was a natural experiment in which residents who moved to a suburban community in Canada were promised free broadband access and online services. The control group consisted of those who did not have an Internet connection. The results of the Netville study suggest that the Internet facilitates local community connections, not just long distance ones as was originally assumed. When wired residents were compared to nonwired neighbors, those with access to the Internet were more involved with their community. They recognized three times as many neighbors, talked to them more times, visited more in person, and called them on the telephone more often than those who were not connected to the Internet (Hampton and Wellman, 2003).

Another study, the e-neighbors study, investigated four neighborhoods in the Boston area. This was a longitudinal study designed to understand the extent to which the Internet facilitates interaction at the neighborhood level. The neighborhoods were similar in terms of socioeconomic status, but differed in their location and type of housing. Two were inner city neighborhoods and two were suburban, including one apartment building and one gated community. Three of the neighborhoods were given access to a neighborhood e-mail discussion list and a neighborhood Web site. The e-neighbors study found that Internet use became increasingly embedded in the neighbor networks. Furthermore, over time, those who were early adopters of the technology were more likely to know more of their neighbors. The study concluded that Internet use encourages the formation of local networks as well as participation in political discussions and engagement in political action. The study's main conclusion was that online activity increased local awareness and the rate of social interaction among neighbors (Hampton, 2007).

From these two studies, we can see that the use of media in the local community is associated with the formation of a sense of community, a positive emotional bond that develops between individuals and groups and their environment, and reflects identification with place. One very important question is what is the role of community networking in the formation and support of this sense of community? The Netville study (Hampton, 2007) shows that Internet connectivity supports a variety of social ties, both strong and weak, which can be instrumental, emotional, or social. Locally based ties are maintained via a combination of online and off-line interactions. The findings show that in addition to the ability of the Internet to serve as a global communication technology, a great deal of online activity takes place between people who live (or work) near one another (Hampton and Wellman, 1999). Wired residents are more likely than nonwired ones to know more neighbors and socialize with them more often (Hampton and Wellman, 1999, 2001). Other studies provide empirical evidence demonstrating that the use of e-mail lists can facilitate the formation of citizen associations at the local level and support their activities. Citizen associations use the Internet to locate and attract individuals and groups, provide a forum for the open discussion of plans of action, disseminate information about the events and activities of the group, and provide an opportunity for community participation and the expression of a wide range of ideas.

In sum, there is a long tradition of theoretical frameworks about the effect of media on community life. Local media promote the formation of local communities and communities without propinquity. For locally based communities, the media is an agent of socialization, promoting the identification of local problems, and providing an infrastructure for the formation and maintenance of a network of social ties that promote attachment to the community. Studies have demonstrated that the Internet is a technological infrastructure that promotes the formation of communities without propinquity, fosters the exchange of information, and provides resources for social support. It is often assumed that Internet use in and of itself increases communication among neighbors and participation in the local community. However, according to the participation hypothesis, Internet connectivity provides only the access to the opportunity for participation. Access to the Internet allows people to participate in locally based e-mail lists or forums that provide local information and that link residents to their neighbors. Thus, on a local level, the Internet serves as an intervening mechanism, adding to and amplifying participation and involvement in the local community. In addition, the findings consistently indicate that once residents participate in a geographically based online community, the likelihood of membership in locally based organizations increases and is followed by community attachment.

See also: Audiences, Media; Journalism; Mass Media, Representations in; Media Effects; Media and Social Movements.

Bibliography

Ball-Rokeach, S.J., Kim, Y.C., Matei, S., 2001. Storytelling neighborhood: paths to belonging in diverse urban environments. Communication Research 28 (4), 392–428.

Etzioni, A., Etzioni, O., 1999. Face-to-face and computer-mediated communities, A comparative analysis. The Information Society 15 (4), 241–248.

Driskell, R.B., Lyon, L., 2002. Are virtual communities true communities? Examining the environments and elements of community. City and Community 1, 373–390. Comparative analysis. The Information Society 15, 241–248.

Hampton, K.N., Wellman, B., 1999. Netville online and offline observing and surveying a wired suburb. American Behavioral Scientist 43 (3), 475–492.

Hampton, K.N., Wellman, B., 2001. Long distance community in the network society contact and support beyond netville. American Behavioral Scientist 45 (3), 476–495.

Hampton, K.N., Wellman, B., 2003. Neighboring in netville: How the Internet supports community and social capital in a wired suburb. City & Community 2 (4), 277–311.

Hampton, K., 2007. Neighborhoods in the network society: the e-neighbors study. Information, Communication and Society 10, 714–748.

Janowitz, M., 1952. The Community Press in an Urban Setting. Free Press, Glencoe, IL.

Jung, J.C., Qiu, J.L., Kim, Y.C., 2001. Internet connectedness and inequality: beyond the 'divide'. Communication Research 28, 507–535.

Kavanaugh, A.L., Carroll, J.M., Rosson, M.B., Reese, D.D., Zin, T.T., 2005a. Participating in civil society: the case of networked communities. Interacting with Computers 17, 9–33.

Kavanaugh, A., Carroll, J.M., Rosson, M.B., Zin, T.T., Reese, D.B., 2005b. Community networks: where offline communities meet online. Journal of Computer-Mediated Communication 10 (4). Retrieved from: http://jcmc.indiana.edu/vol10/issue4/kavanaugh.html.

Kim, Y.-C., Jung, J.-Y., Ball-Rokeach, S.J., 2006. 'Geo-ethnicity' and neighborhood engagement: a communication infrastructure perspective. Political Communication 23 (4), 421–441.

Merton, R.K., 1957. Social Theory and Social Structure. Free Press, Glencoe, IL.

Mesch, G.S., 2007. Online communities. In: Cnaan, R.A., Milofsky, C. (Eds.), Handbook of Community Movements and Local Organizations. Springer, New York, pp. 227–242.

Park, R.E., 1922. The Immigrant Press and Its Control. Harper, New York.

Putnam, R., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon & Schuster, New York.

Shah, D.V., Cho, J., Eveland Jr., W.P., Kwak, N., 2005. Information and expression in a digital age: modeling Internet effects on civic participation. Communication Research 32 (5), 531–565.

Skogerbø, E., Winsvold, M.S., 2011. Audiences on the move? Use and assessment of local print and online newspapers. European Journal of Communication 26, 214–229.

Stamm, K.R., 1985. Newspaper Use and Community Ties: Toward a Dynamic Theory. Ablex, Norwood, NJ.

Stamm, K.R., Emig, A.G., Hesse, M.B., 1997. The contribution of local media to community involvement. Journalism Quarterly 74 (1), 97–107.

Stephens, M., 20 November 1995. Radio: From dots and dashes to rock and Larry King. The New York Times.

Webber, M., 1964. Order in diversity: community without propinquity. In: Wingo Jr., L. (Ed.), Cities and Space: The Future Use of Urban Land. John Hopkins University Press, Baltimore, MD, pp. 23–54.

Wellman, B., Whortley, S., 1990. Different strokes from different folks: community ties and social support. American Journal of Sociology 96, 558–588.

Community and Society: History of the Concepts

Bo Stråth, University of Helsinki, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Community and society are key concepts in sociology, political theory, and philosophy of history. In ancient Greece, community and society were kept together in one concept. Translated into the political organization of the Roman Republic and Empire, the two concepts began to take on various meanings although without clear distinctions. During the nineteenth century community and society split up and had become counterconcepts expressing social and institutional alternatives. The roots of this separation were developed already in the eighteenth century. Together with another, closely connected, couple of counterconcepts, state and society, they played a crucial role in the reflection and the theory building in social sciences with clear ideological and political implications. This article discusses the contradictory and supplementary views on the two concepts expressed by philosophers like Hobbes, Spinoza, Locke, Kant, Smith, Hume, Ferguson, Fichte, Herder, Hegel, Durkheim, Marx, Engels, Tönnies, Weber, Simmel, Parsons, and König, and ends with a connection to the debate between the communitarians and the libertarians during the last decades of the twentieth century.

Community and society are key concepts in sociology, political theory, and philosophy of history. In antique Greece, community and society were kept together in one concept. Translated into the political organization of the Roman Republic and Empire, the two concepts began to take on various meanings although without clear distinctions. The synonymity – and the ambiguity – continued with the scholastic philosophy. In the eighteenth century the complexity of the conceptual field around the equivalent terms of community and society had increased and included communion, participation, joining, taking part, sharing, liaison, union, corporation, cooperative, friendship, brotherhood, fellowship, etc. Rooting in the eighteenth century but explicitly only during the nineteenth century community and society split up and had become 'counterconcepts' expressing social and institutional alternatives. Together with another, closely connected, couple of counterconcepts, state and society, they played a crucial role in the reflection and the theory-building in social sciences with clear ideological and political implications. This article discusses the contradictory and supplementary views on the two concepts expressed by philosophers like Hobbes, Spinoza, Locke, Kant, Smith, Hume, Ferguson, Fichte, Herder, Hegel, Durkheim, Marx, Engels, Tönnies, Weber, Simmel, Parsons, and König, and ends with a connection to the debate between the communitarians and the libertarians during the last decades of the twentieth century.

Community and society are key concepts in sociology and political and historical philosophy. In antique Greece, community and society were kept together in one concept: *χοινωνι΄α* (Choinōnia). The antique Greek political philosophy did not distinguish between the institutions of community and society. *χοινωνι΄α* was translated into Latin as sometimes *societas*, sometimes *communitas* without clear distinction. With Thomas Hobbes (1588–1679) and Baruch Spinoza (1632–77) in the seventeenth century nature and natural rights disappear as explanation of human collectivity. With the exception of the family all forms of socialization and communalization (collegium, corporation, craft, parish, municipality, partnership, trading company, etc.) were secondary formations within a state federation (*civitas*, *res publica*) constituted through a social convention. Hobbes talked about 'systems of people, numbers of men joined in one interest, one business.' Instead of the concept of society understood as synonymous with community a new concept of society emerged based on a conventional meaning identical to that of the state in the modern sense of the word.

From the middle of the eighteenth century and increasingly from the 1850s society and the social were seen as taking shape between individuals or households and the political order. Almost all social thought during this period was preoccupied with conceptualizing the collective entities that emerged within the 'social' in order to describe and analyze their nature, stability, and ties of cohesion. Besides community and society, concepts like nation, class, and culture were suggested. Societies were seen as organized by division of labor and social interests, by class struggle, or by cultural–linguistic identity. This was the more general framework of the debate on community and society.

John Locke (1632–1704) emphasized the transformation from the traditional view on social relationships, based on natural right, to the modern thought based on rationalism and positive law, and replacing Aristotle's teleological view with the functional idea of collective security. The emphasis on collective security should be seen in the context of the growing role of private property. This development, wherein political philosophy society had become the preferred concept, did not mean that community disappeared. With the exception for political economy, where 'community' was narrowed down to trade company, the definition of society as community remained in everyday language until the end of the eighteenth century. Immanuel Kant (1724–1804) viewed society, in its context of being based on reason and conventions among individuals, as a dynamic totality, the elements of which acted with and against one another. The construction of community for social exclusion and protection meant both inclusion and exclusion, the identification of both 'with whom' and 'against whom.' The 'unsocial sociability' (*ungesellige Geselligkeit*) of humans drove them to form societies. However, this drive toward convention-based collectivity provoked resistance, and the threat of the dissolution of the society was permanent. Kant's view was in this respect close to the school of English-Scottish moral philosophers (Adam Smith (1723–90), David Hume (1711–76), and Adam Ferguson (1723–1816)).

Kant's question about a perfect society based on the free interaction among individuals was taken up by Johann Gottlieb Fichte (1762–1814) at the beginning of the nineteenth century and put into a historical and linguistic-critical context. Although Fichte often used society and community as synonymous terms, he exposed the two concepts, and that of state, to linguistical–critical scrutiny observing the prevailing conceptual confusion.

As far as the German-speaking area is concerned, around 1800 Johann Gottfried Herder (1744–1803) suggested a distinction between community and society. His point of departure was the state–society connection, where 'state' emphasized the unity and closure of the collectivity, while 'society' had a focus on the multiple and open forms of collectivity including those based on the individual rights. In order to describe the relationship between state and church, Herder used society, based on Enlightenment reflection on the contextuality and historicity of religious forms. In order to emancipate the spirit and the mind, the *Geist*, from its entanglement with the state, and to rediscover the suprahistorical core of Christianity, Herder required the separation of society and community, *Gesellschaft* and *Gemeinschaft*, where the latter term was used to describe the community of spirits independent of the state. Herder's distinction, which again took up the old Greek and Biblical-Christian *χοινωνι΄α* concept and infused it with new meaning, failed however to go beyond a specific religious vocabulary.

Kant's moral philosophy provoked criticism from G.W.F. Hegel (1770–1831), who rejected the idea of the autonomous individual. In Hegel's view the individual is certainly rational, although rationality is historical and *transindividual*. The normative and legal order of a society could not have its point of departure in the idea of its individuals as a priori free, equal, and rational. Humans are free only within a morally strong community. Society is not based on convention but on morality. Such is Hegel's view. The human effort, to be recognized at individual levels, results in the emergence of various communities, which are kept together by morally strong virtues, *Sittlichkeit*. The social community was linked to an ethical dimension, which, according to Hegel, appeared in various shapes through history, which was nothing but the development toward higher forms of ethical realization. The highest stage in this historical development was the state, which gave all its citizens true freedom through the confirmation and recognition of them in a meaningful totality. The individual is emancipated through subordination in the collective realization of *Sittlichkeit*.

From this condensed rational–romantic version of community, inspiration could be picked up a century later in Germany for extreme versions of the concept, where it was linked to the *Volk* ('people') concept and political action orientation. The difference between the individualistically oriented Kant and the collectivistic view of Hegel is not restricted to German history but is however more general and has been discernible in two alternative world views on community and society ever since. In the years around 2000 these alternatives were still being expressed in the debate between liberals and communitarians.

Emile Durkheim (1858–1917) developed Hegel's key question about the moral basis of modern societies in new directions. 'Modern society' is henceforth understood in terms of industrialization, division of labor, and democratization, processes that only had just begun when Hegel reflected on society, but were much more evident for Durkheim. His key question dealt with social solidarity, and how to bridge the distance between the state and the individual. The point of departure for Durkheim when he posed this question was the debate about the nature of the moral and political sciences. He was unsatisfied with the speculative character of political philosophy. He was convinced neither by the individualistic and utilitarian approach developed by Kant or in the classical political economy, nor by Hegel's holistic enlightenment–romantic and mystical view on community, where the societies had goals, which were above and beyond those of the individuals. Durkheim looked in his positivist and empiricist approach for a position between Kant and Hegel for a society that was rather different than theirs. This debate was the very foundational moment of the social sciences in the contemporary sense.

The motor of the new turn of the debate was what in the 1830s was identified as the social question. The spread of industrial capitalism, with contract-based property rights and wage labor, provoked new forms not only of property but also of poverty and proletarianization, which went beyond the curing capacity of the conventional tools of philanthropy and Christian charity. A systemic problem inherent in industrial capitalism was identified and a debate on what was labeled the social issue emerged. With the ever stronger class language of the working class the social question was in the 1870s transformed into the class question.

Durkheim emphasizes the role of religion in premodern societies, in which, beyond the family, no other functionally cohesive institution other than religion existed. Religion provided 'mechanical solidarity,' which functioned as a social cement. This mechanical solidarity can be related to Hobbes's understanding of society (and nature) as clockwork divinely created but then going by itself. Post-Renaissance rationalism and the imagery of mechanical solidarity through religious cohesion, in the shape of Hobbes and others takes on new meaning when in the second half of the nineteenth century it was faced with a truly mechanical universe, i.e., the advances of physics, the division of labor, automated production, etc. In modern societies the division of labor changed the preconditions for social cohesion. A need for coordination of functional relationships emerged. The societies had become ever more complex with growing divergence among individuals in their social orientation. Their interpretations of the world did not necessarily converge toward one collective will. The state was the key instrument for collective action when societies had developed a certain degree of complexity. The fiction of the state as the tool of social security emerged with a capacity to promote social integration through political and social reforms suggested by the social and political sciences. This was the social state much more differentiated and with more tasks than the earlier state which basically was a war machine and a Hobbesian instrument for domestic social peace. The state as the instrumental–rational forms of belonging was connected to the imagery of the nation unified by organic, cultural, and social bonds integrating both civic and ethnic dimensions. With this kind of reflection in the

wake of the growing social division of labor organic solidarity arose alongside mechanical solidarity.

A growing emphasis on ethnicity and a stronger demarcation to other nations accompanied the increasing social dimension of nationalism. It was not by coincidence that Bismarck's imagery of state socialism (below) went hand in hand with the martial German unification, welfare with warfare. The social conservatively oriented nationalism from the 1870s/1880s onwards was different from the earlier liberal nationalism with an emphasis on the civic dimension and with less interest in the demarcation between nations. Giuseppe Mazzini (1805–72), a protagonist of liberal nationalism and the founder of both Young Italy and Young Europe, did not see any problem in coexisting national projects peacefully united in an international order. Nationalism and cosmopolitan internationalism constituted one another.

The state as the functional answer to social differentiation and the problem of social disintegration in the wake of economic integration, and the imagery of the nation as the locus of social cohesion and community reinforced potentiating one another, although the final merger between state and nation in the modern nation states occurred only after 1945. Tensions emerged in multiethnic empires. One way to solve such tensions was reinforced nationalism at the imperial level under demarcation to other empires and underpinning the colonial extension of empires. The empires in struggle for survival (Charles Darwin (1809–82) and Herbert Spencer (1820–1903)) with other empires promoted social cohesion, and social imperialism. The nineteenth-century location of nation varied under more or less tension between a specific state and within a multinational empire which accommodated several nationalities.

Durkheim problematizes the mediation between the individual and the society in different directions and from different points of departure. The political philosophy of Enlightenment focused on concepts like people's sovereignty and political representation and on the issue of the institutional distribution of power. With intensified industrialization, capital concentration, the rise of the labor movement, and ideas of universal suffrage, the theoretical framework provided by the Enlightenment philosophers had become insufficient. New forms of relationships between the individual and the state levels had to be developed. In this context Durkheim evokes the notion of secondary institutions and develops ideas about what kind of organizational forms would fit an industrial society.

Taking his theoretical point of departure in his idea of division of labor as the basic principle of modern societies, he suggests interest-based representation through corporations rather than territory-based political representation. The corporations would reflect modern professions rather than traditional crafts and estates. In his functionalist approach Durkheim developed a moral theory in which he paid relatively little attention to questions of power and authority structures and social inequality. However his view on morality based on corporate representation and related to his idea of solidarity indicated the social question. The political function of the secondary institutions was both to keep the state sufficiently free from influences from the masses, so that the state could fulfill its role as the shaper of the society, and to protect the individuals from oppression by the state.

Karl Marx (1818–83) and Friedrich Engels (1820–95) followed a different route. In their historical materialism, and in their emphasis on the political role of the proletariat, they depreciated the concept of society ('bourgeois'). They reevaluated the community concept with its derivative communism, where socialism, with the same etymological origin as society, was seen as a transitional phase toward communism, the final stage, where history ceased. Marx and Engels criticized the view of society in earlier French and British social history, where society was seen as a set of mutual interactions. They criticized this view of society as being a society of exchange in which every member was a merchant. The society Marx and Engels discerned was societies in plural, where each of them was determined by the specific ways in which the production was organized. In each type of society specific power relationships prevailed. They followed one another in historical phases: the original society (*Urgesellschaft*), slave society, feudal society, bourgeois or capitalist society, socialist society, and communist community, where the merger into the old Greek *χοινωνί'α* was fulfilled in a new action- and future-oriented utopia.

Hegel's problematization of the leading idea in moral philosophy and political economy, that society was based on the individuals and their agreement on conventions, constituted in the German-speaking area a step in the development from the traditional to the critical theory of society (cf Ruda (2011) and Honneth (2011)). The critical thought was a core dimension of the enlightenment philosophy (Koselleck, 1989 [1959]). It became more pronounced through the debate on the social question and from the 1870s onwards to the class question. Karl Marx and Friedrich Engels were protagonists in the development of critical thought. Also positivism in the wake of Auguste Comte (1798–1857) and French cultural radicalism in the 1880s promoted the attention to criticism. These developments culminated in the critical theory established in the 1930s by Max Horkheimer (1895–1973), Theodor Adorno (1903–69) and their circle with a Marxist subtext but influenced also by Hegel and Freud (the Frankfurt school). Their long chain of critical thought introduced new basic concepts and distinguished between society and community, between society and state, and between *bürgerliche* (the German term connotes in English translation both 'bourgeois' and 'civil') society and society. This critical social theory must be seen in the context of the social revolutionary movement, which emerged in the nineteenth century. This movement could be seen as a response to the crisis, which since the French Revolution, and with the industrial revolution, marked the everyday experiences in Europe and gradually also beyond Europe. Critical theory was an alternative reflection on society and community with respect to Durkheim's, based on ideas of functional differentiation and division of labor.

Politically the most long-term important element of the tradition of critical thought was the conceptualization of a social history, or history of society, *Gesellschaftsgeschichte*, with explicit connection of historical processes to social theories, in particular those designed by Max Weber but partly also Karl Marx. Long before the French revolution German scholars had propagated a historical approach to political, constitutional, and legal questions under demarcation to the unhistorical rationalism of the French *philosophes*. Legal scholars insisted on the value of empirical 'positive law.' The

historical school argued that economic performance and social organization could not be conceptualized in universal categories but had to be understood in the framework of customs, legal norms, political institutions and practices, culture and the economic organization and practices of the production of subsistence and wealth, which all had emerged and grown historically, specific for each people. An academic key discipline in this development was law. Friedrich Carl von Savigny (1779–1861) and Carl Friedrich Eichhorn (1781–1854) developed the historical school where they emphasized the historical tradition as opposed to speculation in natural right philosophy. The focus was on the sources of law as they had emerged from the spirit of the people. The efforts of the historical school aimed at the reconstruction of a national past paralleling and reinforcing the movement toward political and legal unity of a unified German state.

The historicists put the classical economic theory in economics in the wake of Adam Smith under pressure. Already the 'older historical school' (Gustav Schmoller) with names like Wilhelm Roscher (1817–94), Bruno Hildebrand (1812–78), and Karl Knies (1821–98) had begun to undermine the classical theory with arguments about the historicity and contextuality of economic and social processes. Not abstract models or universal laws of how the economies functioned, but the historical experiences laid down in political, social and economic institutions, and widely varying from country to country, explained the functioning and the development of the economy in each specific case. This was the view that Friedrich Carl von Savigny had applied to legal interpretations, and which from there evolved a methodology in other academic disciplines, not the least political economy. The center of this academic movement against universalism was Germany. Karl Marx and Friedrich Engels tried to develop a historical universalism with a certain degree of openness toward historical variety.

The younger historical school with Gustav Schmoller (1838–1917) as uncontested protagonist gave the historicist approach greater cohesion and more punch. He confronted the axiomatic–deductive approach of classical economics. He was critical to what the economists in the classical tradition argued to be regularities and conformity to law in economic life. Their theory about rationally acting economic individuals led astray. Only the historical view on society and humans could provide interpretative power to developments and behavior.

The economists did not accept the criticism of Schmoller and the other historicists, but hit back. They were later to go under the label of the neoclassical school with their neoclassical theory. Schmoller called them the Austrian school to suggest provincialism in a disliking review of a book by Carl Menger (1840–1921) which attacked the methods of the historical school. The debate that followed came to become known as the *Methodenstreit*.

In the 1870s social sciences and social politics had become ever more intertwined in Germany, social political and social scientific intentions merged. After the violent unification of a German nation through Bismarck's wars with the neighboring countries the turn came to domestic national integration. Social sciences and social politics were effective instruments here. In 1872, a group of professors of political economy, *Nationalökonomie*, met to discuss how to promote state intervention in social issues on the basis of historical–economic investigations. In 1873, they founded the *Verein für Sozialpolitik*, the Association for Social Politics. The establishment of the professors' association for social politics stood in close connection to the foundation of the German *Reich* in 1871.

Supported by this academic trend, German Chancellor Otto von Bismarck's great political achievement was the appropriation of the national question from the liberals and the social question from the socialists making nationalism and ideas of a social state (*Sozialstaat*) conservative virtues.

The political intention of the association of economists was fully focusing on the state. Despite all socialist reproaches from classical laissez-faire-oriented economists, who referred to the members of the association as *Kathedersozialisten*, alluding to their academic chairs, a label which soon broke through in the public debate in Germany and abroad, the members of the association saw themselves in a mediating position between socialists and market apologetics. The task of the association was to become a forum in which academic, economic and political–administrative elites could communicate about necessary measures to be undertaken on the basis of social scientific research results. An urgent requirement was the establishment of a public debate, a *bürgerliche Öffentlichkiet*. The idea of social politics was based on the conscious insight that state and society necessarily had to be activated in order to establish a balance between the social interests drifting apart.

In France the development was similar although there the academic discipline was sociology, in particular in the version developed by Émile Durkheim. The development was somewhat later in the framework of the Third Republic. However, also in political theory there was a movement from the emphasis on neoclassical theory toward historically-cum-analytically derived theories for state intervention. In the *sciences morales et politiques*, too, efforts emerged to develop a political science aiming at supporting national action capacity in the social field, not least in the teaching for the cadres of the administrative elite.

A few years after the death of Karl Marx and after ever stronger experiences of industrial change and economic concentration, Ferdinand Tönnies (1855–1936), analyzed quite differently, but still with a utopian embedding of the community concept, the connection between *Gemeinschaft* and *Gesellschaft*. In his view community took on a social–romantic meaning connoting an archaic past based on the family and providing feelings of holism. His utopia was projected toward the past instead of the future. Society was the term used to describe the breaking-up of this past through industrialization, social disintegration, and experiences of atomism. Tönnies saw communism as a cultural system of community and socialism as the cultural system of society, i.e., communism was rather referred to as a historical or even archaic category, a kind of prototype, while socialism reflected the industrial society.

Tönnies' subordination of society to community had long-term political impact. Community was, in his view, in all sociological conceptualizations the superior and in all historical analyses the anterior concept. Community was the dialectical womb from which society arose in historical processes. The

concept of society described the law-bound normal process of the decline of all community: the same process that Marx and Engels – in accordance with Hegel's idealistic philosophy and their critical–normative prognosis, where society had become community – described as emancipation and a development toward higher stages of social organization, Tönnies analyzed as decay, as inexorable as the coming communism in the view of Marx and Engels. The only future hope that Tönnies could discern was weak: that the force of the community also within the era of society would remain although it was declining. The community could in his view be promoted through the cooperative movement and through the trade union and solidarity idea in the labor movement.

Although Tönnies' concept of society includes the totality of the relationships among the individuals mediated through exchange, and in this respect can be historically situated, this historical location involves conceptual distinctions which cannot be justified either in genetic or in linguistic–critical terms. Both Max Weber (1864–1920) and Georg Simmel (1858–1918) observed this inconsistency and tried to solve it through a paradigmatic change of the sociological terminology. To consider social institutions like state, church, the cooperative (*Genossenschaft*) as independent entities preceding the individual and attach them to a general society concept was in Weber's view a metaphysic residual, which needed to be overcome. He replaced society as a basic sociological concept with the term *Vergesellschaftung*, 'societalization,' an extremely general concept containing all forms of rationally motivated interest relationships under one umbrella concept. The generalization of the sociological conceptual framework around the term *Vergesellschaftung* certainly described processes and social change, but was basically ahistorical, where the analysis of the development was deprived of historical contextualization. *Ahistorical* here does not mean a lack of a chronological dimension but refers to the fact that the social dynamics was deeply teleological losing the openness and contingency of historical processes out of sight. Historical contexts were understood as 'cultural tragedy' (Simmel) or 'destiny of the rationalization of society through science' with total disenchantment as the final outcome of the *Vergesellschaftung* (Weber). The process of dehistorization of the concept of society that Tönnies began with his outline of a universal process, and Weber continued with his criticism of Tönnies, had the effect of drawing growing attention to the community concept. *Gemeinschaft*, which can be seen as a kind of pseudohistorization, had become topical. With the introduction of the concept of community in this way the sociological theory in Germany unintentionally opened up for the reactionary, opposition against the industrial society. While in other European languages society and community remained synonymous concepts – in the wake of the historical school and Durkheim, who had narrowed down the gap between these two concepts that the enlightenment philosophers had opened up – *Gemeinschaft* had become in Germany an ideological key concept connected to another highly ideologized crucial concept: the *Volk*. The ideological entanglement of the two concepts had become a national–conservative instrument to mobilize the *völkische* movement after World War I against socialism, capitalism, and industrialism.

The cult and the dogmatization of the *Gemeinschaft* concept were also confirmed in German philosophy at this time. Max Scheler (1874–1928) identified community as a person *sui generis*, 'a person of persons' as opposed to society, which was nothing but the sum of its individuals. He thus reinforced Tönnies' distinction. Community was based on blood, tradition, and historicity, while society was the waste product of internal disintegrative processes in the community. Society was synonymous with public, which continuously removed – Marx would have said alienated – the individual. Community was synonymous with personal solidarity, which reintegrated the individual. In Durkheimian terms this was a move from organic toward mechanical solidarity.

Already before World War II, Talcott Parsons (1902–79) developed a view where community and society were seen as key concepts in structural–functional alternatives of value orientation, 'pattern variables': affectivity versus affective neutrality, particularity versus universalism, ascription versus achievement, diffuseness versus specificity. His approach was one, where, like in the German debate, the two concepts were separated, but for the rest his contribution was very different from the German debate.

In the 1950s René König (1906–92) was one of the leading figures in the reconstruction of German sociology. He argued vehemently against Tönnies' dichotomy and use of the *Gemeinschaft* concept. He argued from his basically Durkheimian perspective for an empirically oriented sociology with bridges to the American functionalism, where Talcott Parsons was a leading figure, as a source of inspiration, and warned against any attempt to conceptualize society as a totality, which only would lead to speculation and dogmatism. Community in the German form of *Gemeinschaft* was in the 1950s rejected from the political and sociological vocabulary and society emerged instead in the 1960s as the key concept to describe and analyze the social organization.

At the same time as König in Germany warned against the *Gemeinschaft*, concept community had become the name of the European integration project in the 1950s. One of the key problems when this project was formulated was how to make Germany a strong ally in the Cold War and at the same time guarantee that the strength was controllable for its allies. The main architects of this design came, in particular, from the USA, the victor, but also from France, where the concern about the key problem for historical reasons was particularly great. There, *communauté*, for historical reasons, in the 1950s had a much more positive value charge than *Gemeinschaft* in Germany integrating Durkheim's organic solidarity based on functional division of labor. Moreover, the Communauté Européenne was a concept that bypassed the conceptual bias on state and state sovereignty in international relations theory and international law. In the form of the European Community in German translation, *Gemeinschaft* came back in as a political concept and an intellectual tool to promote the German incorporation in the West European Alliance in the Cold War, and which, therefore, was disconnected from its historical heritage in the German speech area. When European integration in the 1970s was exposed to strains, the idea of a European identity emerged connected to ideas of a supranational Europe. This intensification of the European Community concept was emphasized in the 1990s when community was changed for union. This European development can be understood as part of a modernization theory which was resurfacing with

a communitarian twist in the 1990s as the reverse side of the liberal post-Soviet coin: the economist globalization rhetoric about profit-maximizing individuals on global markets untied from suffocating social ties. The imagery of social individuals freed from social bounds created its counterimagery of individuals tied together in a European Union with a European identity. The homogenizing discourse on a European identity attempted to provide the glue to bind a European community together. Such a theory of social integration, which took for granted that cultural cohesion existed for national societies, and then simply transferred the idea of state – formation onto a supranational European level, built around the concept of European identity, missed the point that the nineteenth-century construction of the nations in Europe did not start from preexisting units but emerged slowly in a tough and often violent conflict around the social and the class issues. The nations were the outcomes of these conflicts, not the point of departure. The final compromise around the social conflict was the merger of the state and the nation in the nation-states after 1945. The problem in translating the national experiences since the 1830s into a European future, revealed in today's European crisis, is the lack of a social dimension in the European construct. A process of renationalization accompanies the financial crisis. The general theoretical framework of the problem of the tension between economic integration and social disintegration is the debate between libertarians and communitarians in social sciences since the 1970s, which debate, in turn, in a long historical view could be seen as continuity under transformation of the bicentenary debate between Kantians and Hegelians. A point of departure for this theoretical debate was John Rawls's (1921–2002) *A Theory of Justice* (1971). Erected in the tradition of contract theories and Kantian liberalism his theory bypassed the questions of political obligation and the welfare state through the idea of distributive rights in individualist terms. Utilitarian philosophers objected but in the long term the most substantive opposition to the neo-Kantian approach emerged from the 'communitarians.' The debate between the liberals and the communitarians had become one of the most important issues in political philosophy in the 1980s. Methodologically the communitarians argued that the premise of individualism such as the idea of free and rational choice is wrong or biased and neglects the social, cultural, and historical contexts in which individuals act. In order to understand individuals one must look at their communities and communal relationships. In a normative sense the communitarians argued that the premises of individualism give rise to morally unsatisfactory consequences. Alasdair MacIntyre (b. 1929) in *After Virtue* (1984) questioned the fundamental philosophical assumptions of liberalism. He and others questioned the claim of the priority of the right over the good and the image of the freely choosing individual. Political arrangements cannot be justified without reference to common purposes. Freedom flourished within social interdependence, not among free-floating individuals. However, the kind of community the communitarians drafted was one that often focused on the family, the tribe, or the neighborhood rather than on the nation, the state, or the class. The communitarian community is more than a mere association; it is a unity in which the individuals are members not in artificial or instrumental terms, but within its own intrinsic value. This is a view with similarities to Hegel's *Sittlichkeit* idea.

In the debate, communitarianism has been dismissed both as little more than a corrective to the excesses of liberalism and as a form of old-fashioned collectivism. In the framework of the former view approaches have emerged, which can be described as a merger between liberalism and communitarianism. Michael Walzer (b. 1935) sees communitarian thought as a permanently recurring corrective to liberalism (1983). Charles Taylor (b. 1931) is another leading name in this approach. Taylor and Walzer are often referred to as liberal communitarians. The theoretical debate moved from the thick communitarian critique of thin liberalism to positions that combine elements of both. This trend was nurtured not the least from legal theory where a strong recent 'republican' turn aspires to hold both communitarian ideals and democratic participation together. The law was seen as sustaining community and promoting the common good to be negotiated through participation in a polity where liberal rights guaranteed that bargaining is uncoerced. Also, the globalization discourse on the emergence of a globally rather than nationally organized economic order promoted the trend of closure between liberalism and communitarianism. The attempts to reestablish community, experienced as being lost to transnational globally operating forces, resulted in liberal politics of recognition of cultural alternatives in the name of multiculturalism, which, in turn, has become a concept with negative connotation in the wake of the expanding global migration, in addition to the originally positive value charge. Multiculturalism does not only mean respect of cultural difference, but also acceptance of social inequalities. The historical search for a concept that unifies individual freedom and social solidarity continues.

See also: Communitarianism, Sociology of; Community Sociology; Community Studies: Anthropological; Community, Social Context of: The US Case; Durkheim, Emile (1858–1917); Hegel, Georg Wilhelm Friedrich (1770–1831); Individualism versus Collectivism: Philosophical Aspects; Individuals in Society: History of the Concept; Kant, Immanuel (1724–1804); Law and Society; Libertarianism; Society and People: History of the Concept; State and Society: History of the Concept.

Bibliography

Avineri, S., de Shalit, A. (Eds.), 1992. Communitarianism and Individualism. Oxford University Press, Oxford, UK.

Brunner, O., Conze, W., Koselleck, R., 1979. Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland. Band 2. E Klett-Cotta, Stuttgart, Germany.

Christodoulidis, E.A. (Ed.), 1998. Communitarianism and Citizenship. Ashgate, Aldershot, UK.

Durkheim, E., 1933[1893]. The Division of Labour in Society. MacMillan, London.

Etzioni, A. (Ed.), 1995. New Communitarian Thinking. Persons, Virtues, Institutions, and Communities. University Press of Virginia, Charlottesville, VA, and London.

Gutman, A., Taylor, C. (Eds.), 1994. Multiculturalism. Princeton University Press, Princeton, NJ.

Honneth, A., 2011. Das Recht der Freiheit. Grundriss einer demokratischen Sittlichkeit. Suhrkamp, Berlin.

Jaume, L., 2000. La liberté et la loi. Les origines philosophiques du liberalisme. Fayard, Paris.

Kelley, D.R., 2011. Historians and lawyers. In: Jones, Gareth Stedman, Claeys, Gregory (Eds.), The Cambridge History of Nineteenth Century Political Thought. Cambridge University Press, Cambridge, pp. 147–170.

Koselleck, R., 1989[1954]. Critique and Crisis, first ed. MIT Press, Cambridge, MA.

König, R., 1955. Die Begriffe Gemeinschaft und Gesellschaft bei Ferdinand Tönnies. Kölner Zeitschrift für Soziologie und Sozialpsychologie 7.
Lukes, S., 1985. Emile Durkheim. His Life and Work. Stanford University Press, Stanford, CA.
MacIntyre, A., 1984. After Virtue. University of Notre Dame Press, Notre Dame, IL.
Milward, A., 1992. The European Rescue of the Nation-State. Routlegde, London.
Parsons, T., 1949[1937]. The Structure of Social Action. Free Press, Glencoe, IL.
Poggi, W., 2001. Durkheim. Oxford University Press, Oxford, UK.
Powell, 1990. Neither market nor hierarchy: network forms of organization. In: Staw, P.B.M., Cummings, L. (Eds.), Research in Organizational Behaviour, vol. 12. JAI Press, Greenwich, CT.
Putnam, R.D. (Ed.), 2000. A Decline of Social Capital? Political Culture as a Precondition for Democracy. Princeton University Press, Princeton, NJ.
Rawls, J., 1971. A Theory of Justice. Oxford University Press, Oxford, UK.
Riedel, M., 1979a. Gesellschaft, gemeinschaft. In: Brunner, O., Conze, W., Koselleck, R. (Eds.), Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland. Klett-Cotta, Stuttgart, Germany. Band 2. E-G.
Riedel, M., 1979b. Gesellschaft, bürgerliche. In: Brunner, O., Conze, W., Koselleck, R. (Eds.), Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland. Klett-Cotta, Stuttgart, Germany. Band 2. E-G.
Rorty, R., 1989. Contingency, Irony and Solidarity. Cambridge University Press, Cambridge, UK.
Ruda, Frank, 2011. Hegel's Rabble: An Investigation into Hegel's Philosophy of Right. Continuum, London.
Skinner, Q., 1998. Liberty before Liberalism. Cambridge University Press, Cambridge, UK.
Taylor, C., 1979. Hegel and Modern Society. Cambridge University Press, Cambridge, UK.
Taylor, C., 1989. Sources of the Self: The Making of the Modern Identity. Harvard University Press, Cambridge, MA.
Tönnies, F., 1991[1887]. Gemeinschaft und Gesellschaft. Grundbegriffe der reinen Soziologie. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany.
Wagner, P., 1994. A Sociology of Modernity: Liberty and Discipline. Routledge, London.
Wagner, P., 1990. Sozialwissenschaften und Staat. Frankreich, Italien, Deutschland 1870–1980. Campus, Frankfurt/Main.
Walzer, M., 1983. Spheres of Justice: A Defence of Pluralism and Equality. Robertson, New York.

Community Care in England

Caroline Glendinning, University of York, Heslington, York, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Community care is a favored way of supporting disabled and older people in many countries. Arrangements for funding and delivering community – i.e., noninstitutional – services vary between countries and welfare regimes. This article focuses on the development of community care policies and practice over two decades in England. Nevertheless, the themes shaping community care in England – particularly the growing emphases on markets and service user choice – are very similar to priorities in many other countries. Over the past three decades, there have been marked shifts away from hospital-based and institutional care for disabled and older people, toward increased provision of intensive support for people in their own homes. These changes have been accompanied by the wholesale marketization of community care services, with first local authorities and later service users themselves increasingly purchasing publicly funded care from private provider organizations or privately employed care workers. Additionally, a range of policy and practice measures have aimed to improve the efficiency of community care services, especially at their interfaces with acute health care. However, the effectiveness of all these policies is seriously threatened by a severe lack of funding for community care services, relative to both current and projected future demand.

Introduction

This article documents the development of, and current debates in, community care services in England. Similar policy goals exist in many other countries, although the specific arrangements for funding and delivering community-based services vary between welfare regimes. While this article focuses on England, the measures described – particularly the introduction and development of quasi-markets, the increasing emphases on consumer choice and personalization, and measures to improve the efficiency and effectiveness of services – are of international relevance. Unlike some other countries, however, continuing failure to introduce radical reforms in the critically important area of funding may restrict the potential impacts of these measures.

The article first outlines the structure of English community care services. It then describes the simultaneous introduction of community care policies and quasi-markets and the subsequent developments in those quasi-markets aimed at increasing the consumer power of service users. The implications for the social care workforce and selected measures intended to improve the efficiency of community care services are summarized. The article concludes by highlighting the problem of securing adequate funding for community care services and the limitations of successive reform attempts.

Structure and Scope of Community Care

'Community care' denotes services and support for older and disabled people provided outside institutional settings such as residential homes and hospitals. It includes the very substantial unpaid support provided by close family members.

Responsibilities in England are fragmented between the separate budgets, structures, and cultures of the National Health Service (NHS), local authorities, and the Department for Work and Pensions (DWP), as follows (Glendinning, 2007):

- Help with personal and domestic care, day care, and other daytime activities are the responsibility of local authorities. Local authorities rarely provide these services themselves. Rather, their role is to fund and commission services from for-profit enterprises, charities, and social enterprises. Local authority funding comes from taxation and user charges. Local authorities also have statutory duties to assess the needs of family carers and provide help for carers, mostly as one-off grants or respite care. Local authorities are additionally responsible for funding housing-based support services, particularly for people with learning disabilities or mental health problems. There is currently interest in developing special housing with high-level support services on the premises as an alternative to residential care for older people.
- The NHS provides nursing and therapy services for people living in their own homes. NHS services are funded from general taxation and free of charge.
- Older and working-age disabled people needing regular or substantial help with personal care or supervision may be eligible for cash benefits to meet extra disability costs. Family carers may also claim Carers Allowance, which is intended to replace earnings lost because of care responsibilities. These benefits are the responsibility of the DWP.

Access to public resources for community care support can involve multiple assessments including income and assets/wealth tests; health and treatment needs; capacity for self-care and risks of harm; and other individual circumstances, especially the availability of family carers. These arrangements lack strategic coherence, are not easily understood, may be poorly targeted at those with highest needs (Wanless, 2006), and involve costly duplication. In addition, around 6 million family carers contribute care equivalent to £87 billion a year (Carers UK, 2007). These arrangements date from the late 1940s when responsibilities for the 'treatment' and 'cure' of 'sick' people were assigned to the newly established NHS, with local authorities providing personal and practical support for those needing 'care and attention' (Means and Smith, 1998). However, the boundaries between NHS and local authority responsibilities have shifted markedly over time (Lewis, 2001), so that people considered to need 'care and attention' now include those with severe cognitive

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28017-3

impairments, who are bedbound, or are terminally ill (Means and Smith, 1998).

During the past two decades, funding constraints have increasingly restricted local authority-funded support to those with the highest levels of need. Reflecting continuing policies to move care from residential to community settings, there was a 1% shift in expenditure from residential to community services over each of the years from 2002/3 to 2005/7. The numbers using short-term, intensive intermediate care services (see below) following hospital discharge rose from 98 000 to 225 000 during that period, but there was a 22% reduction in numbers of households receiving less intensive community-based support. In 2009, an estimated 6000 older people had high support needs but received no formal services or informal care and 1.5 million people had some shortfall in their care needs, if they lacked family care to make up the deficit (House of Commons Health Committee, 2010: p. 41).

Community Care Policies and the Introduction of Quasi-Markets

Until the 1980s, English local authorities were both funders and providers of domiciliary, residential, and day care services. However, reforms to means-tested benefits in 1980 made it easier for low-income older and disabled people to obtain social assistance funding for private residential and nursing home care. The impact was dramatic. Local authorities quickly saw the opportunity to shift responsibilities for funding institutional care onto the social assistance budget. At the same time, the NHS was closing long-stay learning disability, mental illness, and old age hospital facilities; again, social assistance funding for community-based residential care was an attractive option (Lewis and Glennerster, 1996).

The consequences were an exponential increase in public spending on institutional care; the rapid growth of privately owned residential and nursing homes; and major disincentives to local authorities to develop community-based services. Between 1986 and 1991, social assistance spending on residential and nursing homes rose from £10 million to over £2000 million , as a substantial private care market developed. However, this was contrary to the stated policy of community care. Following a critical Audit Commission (1986) report, a Committee of Enquiry (Department of Health and Social Security, 1988) recommended that responsibility for assessing needs and funding social care services should rest clearly with local authorities and that choice, flexibility, innovation, and quality should be stimulated through a 'mixed economy' in which public, private, and voluntary providers would compete on equal footings for public funding (Lewis and Glennerster, 1996; Means et al., 2002). Individual assessments and micropurchasing of flexible care 'packages' would be done by local authority 'care managers.'

These recommendations were implemented in 1993. Local authorities were encouraged increasingly to purchase services from charitable and for-profit providers, with whom their own in-house services now competed. Social workers became care managers, responsible for assessing needs and purchasing care from those providers with whom the local authority had contracts (Means et al., 2003). Market and consumer choices were effectively exercised by care managers on behalf of service users.

These reforms were largely successful in shifting funding from residential to community-based services and in stimulating new providers. In 1992, the private sector supplied only 2% of all domiciliary care (home help) contact hours; by 2001 this had increased to 60%. By 2001, 85% of all residential care places for adults were provided by private organizations, although this was no longer an expanding market as increasing demand for domiciliary services, falling occupancy rates, and other economic pressures squeezed the sector (Means et al., 2003). The numbers of homes and places within them subsequently dropped (although average home size has increased) (CSCI, 2009).

The Rise of Consumerism and Individualization in Community Care

During the 1980s, organizations of working-age physically disabled people campaigned to receive their community care support in the form of cash payments rather than services in kind, so they could employ their own care workers (personal assistants) to provide help with activities of their choosing, at times and in ways that best suited them (Morris, 2006). Legislation allowing local authority adult services departments to make cash direct payments instead of services in kind was introduced in 1997 for working-age disabled people and, from 2000, to people aged 65+ years, parents of disabled children, carers, and disabled young people (Glasby and Littlechild, 2006).

However, relatively few people chose direct payments, while take-up remained stubbornly low and highly variable between older and younger disabled people and between localities. Barriers to take-up included a lack of support services to help manage direct payments, anxieties among older and disabled people about managing direct payments, shortages of people willing to be employed as personal assistants, and professional resistance by care managers anxious about threats to traditional social work practice (Ellis, 2007; Fernández et al., 2007). Despite intensive policy pressures to encourage the direct payment option, by 2009 still only 9.5% of working-age adults and 3.6% of older people using community care services had direct payments (Care Quality Commission, 2011).

However, from around 2000, a new choice-based consumerist discourse began to shape community care policies. This intensified earlier trajectories of squeezing provider and professional interests, increasing user control and transforming community care markets:

> By putting users at the heart of services, enabling them to become participants in the design and delivery, services will be more effective by mobilizing millions of people as the co-producers of the public goods they value…(Leadbeater, 2004: pp. 19–20).

Consumerism aimed to expand the marketization of community care by devolving purchasing power to individual service users themselves, thereby giving them command over the public resources allocated to fund their community care

services. This also involved transfers of risk, from the state to the private sector and to service users. "Choice has been put forward in a wider range of different services and processes…many of which are potentially liberating for service users but some of which raise concerns about risk, capacity and equity" (Needham, 2007: pp. 74–75).

Leading the implementation of this new approach was a social enterprise organization *In Control* that worked initially with learning-disabled adults to give them a bigger role in assessing the support they need and deciding how resources for that support – termed personal budgets (PBs) – would be used. Direct payments had generally been used for employing personal assistants to provide personal and domestic care, whereas *In Control* encouraged much greater flexibility in the use of PBs, including purchasing mainstream services such as art classes or gym membership rather than attending a special day center. *In Control* connected closely with the principles underpinning direct payments, but had broader aims of 'redesigning' community care toward 'self-directed support' (Duffy, 2004).

In Control successfully promoted this approach in many English local authorities and the experience was picked up in three major policy statements (Prime Minister's Strategy Unit, 2005; HM Government, 2005; Department of Health, 2005), all of which proposed the piloting of individual budgets (IBs). Unlike *In Control*, IBs were to bring together funding from several government departments and services including community care, housing-related support, equipment, and adaptations. The intention was to avoid duplicate assessments and to increase choice, control, and flexibility for community care service users in using the full range of resources to which they were entitled, in line with agreed individual priorities and preferences. Like *In Control*, IBs could be used for conventional community care services, employing personal assistants, paying relatives and friends, or buying mainstream goods and services.

IBs were piloted in 13 English local authorities and subject to rigorous evaluation (Glendinning et al., 2008). IBs were generally welcomed by users as allowing more choice and control over daily activities. However, older IB users reported lower psychological well-being compared to those receiving conventional services, lower satisfaction than younger IB holders, and there was no evidence of cost-effectiveness for older people (Moran et al., 2012). Staff involved in piloting IBs encountered many challenges, including establishing processes for estimating the size of IBs and legitimate boundaries for what IBs could be spent on. Attempts to integrate resources from different government departments were largely unsuccessful, because of central legal, regulatory, and accountability barriers (Moran et al., 2011).

Before the IB pilots ended, a 3-year 'transformation' program was announced involving the extension of PBs to everyone eligible for community care (Department of Health, 2008). Significantly, PBs involved only community care resources – the ambitions of integrating funding streams and reducing multiple assessments had been abandoned. All adults eligible for community care were to be given PBs within 3 years. To support this, local authorities have had to transform methods for allocating funding to individuals and managing financial and other risks, help local provider markets adapt to demands from PB holders for new services, and provide information and support to help people plan and manage their budgets. Again progress has been slow; by March 2012, just over half of all adults potentially eligible for a PB had one in place (ADASS, 2012).

The Implications for the Social Care Workforce

The introduction of direct payments and extension of PBs has implications for the social care workforce, although these are contested. New opportunities for community care users to manage their own support have been argued to undermine the case for professional social work (Leece and Leece, 2011). Others have heralded a new social work role of 'care brokers,' helping people to identify, organize, and manage their community care support (Scourfield, 2008). New tensions in professional social work roles have also been identified, for example, between promoting autonomy and protection and between prevention and rationing (Lymbery and Postle, 2010; Lymbery, 2012).

At the front line, PBs may prompt a more diverse and casualized community care workforce. By 2010, an estimated 200 000 personal assistants were employed by older PB holders (Skills for Care, 2010). The requirement that PB holders be responsible for ensuring the national insurance contributions and taxes of their personal assistants are paid gives relationships between service users and personal assistants a quasi-contractual basis (Dominelli, 2009). However, there are no requirements for PB holders to provide formal contracts or conditions of employment, nor for personal assistants to have minimum qualifications (Skills for Care, 2008).

Surveys of PB holders report greater reliability and flexibility and fewer experiences of abuse by directly employed personal assistants than from workers from private home care agencies. Personal assistants also report high job satisfaction, albeit with complaints about long hours or low pay. Local authorities are developing systems to help PB holders manage their employment responsibilities, but these are highly variable between localities (CSCI, 2008).

Improving Efficiency in Community Care

Other reforms have aimed to improve the delivery, efficiency, and cost-effectiveness of community care services. Two initiatives are described here: intersectoral collaboration and reablement.

Intersectoral Collaboration

Barriers between health and social care services have long been identified as problematic, particularly at the interface with acute hospital care and in demedicalizing support for people with learning disabilities or mental health problems. Attempts to overcome these are equally long standing (Means and Smith, 1998; Means et al., 2002) but new incentives and other measures were introduced from the late 1990s, including legal obligations on NHS organizations and local authorities to work in partnership and national service frameworks that set benchmarks across both sectors (see Glendinning et al., 2005).

Many joint service developments have resulted, particularly in community care services for older people and those with mental health problems. Pooled budgets, joint strategic needs assessments, and joint service commissioning between local authorities and local NHS services have become widespread. Recently, some resources have been transferred from NHS budgets for community care services, with the requirement that these are used in partnership to promote integrated services.

Despite the less than compelling evidence of the cost-effectiveness of collaborative service developments (Cameron et al., 2012), there is apparently still a strong belief in the potential efficiency gains. However, other policies, such as the piloting of IBs (Glendinning et al., 2011), and structural reorganizations of English NHS services, continue to undermine collaboration between health and social care. Moreover, the focus on this particular service intersection may itself inhibit collaborative activities across a wider range of community-based services (Newman et al., 2008).

Reablement

Since 2000, a range of intermediate care and reablement services have been introduced with the aims of improving the efficiency of acute hospitals, promoting independence and avoiding longer term service use, particularly among older people. They include short-term, intensive services to prevent hospital admission and support early hospital discharge (Department of Health, 2000; Godfrey et al., 2005). These are complemented by short-term home care reablement services that help people (re)acquire confidence and daily living skills at home following illness or surgery. Reablement services are often jointly funded by the NHS and local authorities; occupational therapist involvement and rapid access to equipment and assistive technology are vital elements of reablement (Glendinning et al., 2010).

Home care reablement has been found to reduce significantly the subsequent use of community care services, compared to people using conventional home care services, although some of this reduction is offset by the higher cost of the reablement intervention. Nevertheless, it is likely to be cost-effective because of the marked improvements in outcomes for users (Glendinning et al., 2010).

Overall Impacts on Efficiency

There have been concerns that these developments – particularly the growth of reablement services – have been driven primarily by the imperative of increasing efficiency in the acute hospital sector, rather than improving community care services and outcomes. Of increasing concern, however, is the extensive underfunding of community care services relative to need, increasing demographic pressures and changing patterns of morbidity. For example, around a third of the NHS resources recently allocated for partnership working and integrated services are being used to offset cuts in local authority budgets (Humphries, 2013). Similarly, the efficiency of home care reablement services is reduced if users cannot be discharged promptly because of inadequate funding and capacity in the standard home care services that provide longer term support (Glendinning et al., 2010). Moreover, reablement services demonstrate the greatest benefits for people with primarily physical health problems and are likely to be less effective for the growing numbers of people with multiple chronic health problems and dementia.

Since the mid-1990s, funding constraints have led local authorities to restrict eligibility for publicly funded community care services to those with the highest level needs and lowest income levels. Since 2010, there has been a 14% overall drop in the numbers of people receiving publicly funded community care, a trend that began in 2008 and is even steeper among older people. Three main consequences are evident. First, there is growing concern about the lack of support for people with lower level needs – those needing help with shopping, social activities outside the home, gardening, and cleaning – amid arguments that such support can have important preventive functions and delay needs for more intensive, crisis intervention services (ADASS/LGA, 2003; Audit Commission, 2000). Second, rising eligibility thresholds have led to burgeoning numbers of people funding their own community care services – around 270 000 older people according to one estimate (IPC, 2012). There is no evidence on the quality or effectiveness of the service choices they make. Third, there has been a 26% increase in the numbers of people providing substantial amounts of unpaid care to relatives and friends over the past decade (Carers UK, 2012). Substantial unpaid care-giving responsibilities that force carers to leave paid work have wider economic effects. Other extrinsic factors, including changes in local governance and reforms in other welfare sectors (Newman et al., 2008) also affect the success of efficiency-promoting initiatives and the optimization of outcomes and cost-effectiveness of community care services.

Funding and Fairness

Fragmented policy and funding responsibilities for community care are partly to blame for the long history of attempts to reform funding. There is widespread dissatisfaction with current arrangements, although public understanding is poor:

> It often comes as an unwelcome surprise to older people to discover that social care is means-tested and they are expected to rely on their own savings and income until their assets have fallen to the threshold set for state-funded care. It is a common complaint that the existing system penalizes those who have saved for their own old age. (Wanless, 2006: p. xxi)

Since 1996, there have been numerous proposals, from within and outside government, for funding reforms that would safeguard eligibility, increase fairness, and reduce local variations in access to and levels of community care services (Humphries, 2013).

The most recent proposals (Department of Health, 2011) primarily address the very high costs that can be incurred by a minority of people who spend extended periods in residential care. Implementation of these proposals from 2016 will limit total individual spending on care costs to £72 000 (HM Treasury, 2013). In principle, this introduces protection from the catastrophic costs of community care services similar to the protection offered by the NHS in

relation to catastrophic health care costs (Humphries, 2013). Implementation should also be accompanied by widespread access to information, advice, and assessment for community care, for older people, disabled people, and family carers, again extending the principle of universality. An additional 100 000 older people are expected to receive some public funding toward their care costs. Disabled young adults will be entitled to free community care support; those who become disabled between age 18 and 65 years will have a lower cap on their total personal spending. The reforms will mainly benefit younger disabled adults, and older people with relatively high care needs and modest incomes/assets, but are nevertheless broadly progressive (Humphries, 2013).

Although the reforms have been widely welcomed, they expose some major limitations. First, proposals for calculating which costs should count toward the £72 000 cap are far from transparent (Department of Health, 2013). There is huge potential for misunderstanding, disagreement, and error. Even more significant is the continuing failure to address the wider lack of funding for community care services and the structural fragmentation of current arrangements. Between 2005/06 and 2010/11, demand for social care by older people outstripped expenditure by around 9%. Recent cuts in public spending will reduce English local authority budgets for community care services by almost £2 billion between 2011 and 2013 (Community Care, 2013). For the future, continuing demographic pressures will increase the numbers of older people with significant care needs by 600 000 by 2030. The numbers of younger adults with learning disabilities are also expected to rise by one-third and those with physical or sensory disabilities by 7.5% during the same period (Humphries, 2013). An additional £8.3 billion will be needed by 2025, just to maintain local authority spending on older people's community care support at current levels.

Conclusions

Since 1990, the coverage of publicly funded community care services in England has contracted significantly. Intensive domiciliary services are provided only to those with the highest levels of need; many people are excluded altogether because of high eligibility thresholds, modest assets and/or incomes, and/or high levels of family support. With the introduction of quasi-markets, community care services have become more fragmented and this has increased with the introduction of PBs.

Underpinning these trends is the continuing challenge of finding a politically acceptable way of driving more money into community care services:

> The problems, and the options, for solving funding reform, have long been known; and prime opportunities to initiate reform … have been squandered. The failure to grasp this nettle is sadly indicative of the low priority given to social care by successive administrations. (House of Commons Health Committee, 2010: p. 67)

This political challenge is heightened by a highly complex and fragmented system, with separate central government responsibilities, funding streams, and local service delivery arrangements. Radical reforms therefore raise challenging constitutional issues. Nevertheless, unless more funding is made available, fewer and fewer people will be eligible for publicly funded community care. More people will consequently be exposed to catastrophic costs as they spend down their assets on privately purchased residential or intensive community care; and family carers will carry increasingly unsupportable burdens.

Against this background, the priority given to introducing quasi-markets and their extension to individualized funding through PBs is remarkable. But this could be of relatively little importance if increasing numbers of disabled and older people are assessed as ineligible for publicly funded community care services by continuing high eligibility thresholds. The marketization reforms have not injected further resources into community care services but simply altered the ways in which existing, limited resources are allocated and used. Moreover, neither direct payments nor PBs may succeed in delivering improved outcomes for the majority of community care service users. Both initiatives take a method of deploying resources that appears to work well for some community care service users (working-age physically disabled people and those with learning disabilities, respectively) and universalize this to all users, including frail and cognitively impaired older people. There remain significant questions as to whether the benefits of individualized funding and consumer choice outweigh the potential increases in responsibilities and risks for community care service users.

See also: Elder Abuse; Social Work Theories.

Bibliography

ADASS, 2012. Personal Budgets Survey March 2012. http://www.adass.org.uk/index.php?option=com_content&view=article&id=816 (accessed 25.08.13.).

Association of Directors of Adult Social Services/Local Government Association (ADASS/LGA), 2003. All Our Tomorrows: Inverting the Triangle of Care. Association of Directors of Adult Social Services/Local Government Association, London.

Audit Commission, 1986. Making a Reality of Community Care. Audit Commission, London.

Audit Commission, 2000. The Way to Go Home: Rehabilitation and Remedial Services for Older People. Audit Commission, London.

Cameron, A., Lart, R., Bostock, L., Coomber, C., 2012. Factors that Promote and Hinder Joint and Integrated Working between Health and Social Care Services. SCIE Research Briefing 41. Social Care Institute for Excellence, London.

Care Quality Commission, 2011. The State of Health Care and Adult Social Care in England. An Overview of Key Themes in Care 2009. HC 841. Care Quality Commission, London.

Carers UK, 2007. Carers Save UK £87 Billion per Year. www.carersuk.org (accessed 05.02.08.).

Carers UK, 2012. 2011 Census: Health and Provision of Unpaid Care, Local Authorities in England and Wales. www.carersuk.org/media/k2/attachments/Copy_of_2011_Census_data_tables_by_LA_increase_in_nmbers_12_12_2012_AS.pdf (accessed 28.04.13.).

Clarke, J., Newman, J., Smith, N., Vidler, E., Westmarland, L., 2007. Creating Citizen-Consumers. Changing Publics and Changing Public Services. Sage, London.

Community Care, 2013. Adult Care Will Be Damaged if Osborne Maintains Current Level of Cuts, Directors Warn. http://www.communitycare.co.uk/articles/12/04/2013/119093/adult-care-will-be-damaged-if-osborne-maintains-current-level-of-cuts-directors-warn.htm (accessed 25.08.13.).

CSCI, 2008. The State of Social Care in England 2006–07. Commission for Social Care Inspection, London.

CSCI, 2009. The State of Social Care in England 2007–08. Commission for Social Care Inspection, London.
Department of Health, 2000. The NHS Plan. A Plan for Investment, a Plan for Reform. Department of Health, London.
Department of Health, 2005. Independence, Well-being and Choice. Our Vision for the Future of Adult Social Care. Department of Health, London.
Department of Health, 2008. Putting People First: A Shared Vision and Commitment to the Transformation of Adult Social Care. Department of Health, London.
Department of Health, 2011. Fairer Care Funding: The Report of the Commission on Funding of Care and Support. Department of Health, London.
Department of Health, 2013. Caring for Our Future: Consultation on Reforming What and How People Pay for Their Care and Support. Department of Health, London.
Department of Health and Social Security, 1988. Community Care: An Agenda for Action. Her Majesty's Stationary Office (HMSO), London.
Dominelli, L., 2009. Introducing Social Work. Polity Press, Cambridge.
Duffy, S., 2004. In control. Journal of Integrated Care 12 (6), 7–13.
Ellis, K., 2007. Direct payments and social work practice: the significance of street-level bureaucracy in determining eligibility. British Journal of Social Work 37 (3), 405–422.
Fernández, J.L., Kendall, J., Davey, V., Knapp, M., 2007. Direct payments in England: factors linked to variations in local provision. Journal of Social Policy 36 (1), 97–121.
Glasby, J., Littlechild, R., 2006. An overview of the implementation and development of direct payments. In: Leece, J., Bornat, J. (Eds.), Developments in Direct Payments. The Policy Press, Bristol.
Glendinning, C., 2007. Improving equity and sustainability in UK funding for long-term care: lessons from Germany. Social Policy and Society 6 (3), 411–422.
Glendinning, C., Hudson, B., Means, R., 2005. Under strain? Exploring the troubled relationship between health and social care. Public Money and Management 25 (4), 245–252.
Glendinning, C., Challis, D., Fernández, J.L., Jacobs, S., Jones, K., Knapp, M., Manthorpe, J., Moran, N., Stevens, M., Wilberforce, M., 2008. Evaluation of the Individual Budgets Pilot Programme: Final Report. Social Policy Research Unit, University of York, York.
Glendinning, C., Jones, K., Baxter, K., Rabiee, P., Curtis, L., Wilde, A., Arksey, H., Forder, J., 2010. Home Care Re-ablement Services: Investigating the Longer-term Impacts (Prospective Longitudinal Study). Working Paper No. DHR 2438. Social Policy Research Unit, University of York, York.
Glendinning, C., Moran, N., Challis, D., Fernández, J.L., Jacobs, S., Jones, K., Knapp, M., Manthorpe, J., Netten, A., Stevens, M., Wilberforce, M., 2011. Personalisation and partnership: competing objectives in English adult social care? The individual budget pilot projects and the NHS. Social Policy and Society 10 (2), 151–162.
Godfrey, M., Keen, J., Townsend, J., Moore, J., Ware, P., Hardy, B., West, R., Weatherly, H., Henderson, K., 2005. An Evaluation of Intermediate Care for Older People. Final Report. Institute of Health Sciences and Public Health Research, University of Leeds, Leeds.
HM Treasury, 2013. Budget 2013. HC1033. The Stationery Office, London.
HM Government, 2005. Opportunity Age. Meeting the Challenges of Ageing in the 21st Century. Department for Work and Pensions, London.
House of Commons Health Committee, 2010. Social Care (Third Report of Session 2009–10, vol. 1, HC 22–1). House of Commons, London.
Humphries, R., 2013. Paying for Social Care: Beyond Dilnot. The Kings Fund, London.
IPC, 2012. People Who Pay for Care: An Analysis of Self-funders in the Social Care Market. Oxford Brookes University: Institute of Public Care.
Leadbeater, C., 2004. Personalisation through Participation. A New Script for Public Services. Demos, London.
Leece, J., Bornat, J. (Eds.), 2006. Developments in Direct Payments. The Policy Press, Bristol.
Leece, J., Leece, D., 2011. Personalisation; perceptions of the role of social work in a world of brokers and budgets. British Journal of Social Work 41, 204–223. http://dx.doi.org/10.1093/bjsw/bcq087.
Lewis, J., 2001. Older people and the health-social care boundary in the UK: half a century of hidden policy conflict. Social Policy and Administration 35 (4), 343–359.
Lewis, J., Glennerster, H., 1996. Implementing the New Community Care. Open University Press, Buckingham.
Lymbery, M., 2012. Social work and personalization. British Journal of Social Work 42 (4), 783–792.
Lymbery, M., Postle, K., 2010. Social work in the context of adult social care in England and the resultant implications for social work education. British Journal of Social Work. http://dx.doi.org/10.1093/bjsw/bcq045. Advanced access published April 9 2010.
Means, R., Smith, R., 1998. From Poor Law to Community Care: The Development of Welfare Services for Elderly People. The Policy Press, Bristol.
Means, R., Morbey, H., Smith, R., 2002. From Community Care to Market Care? The Development of Welfare Services for Older People. The Policy Press, Bristol.
Means, R., Richards, S., Smith, R., 2003. Community Care: Policy and Practice, third ed. Macmillan, Basingstoke.
Moran, N., Glendinning, C., Stevens, M., Manthorpe, J., Jacobs, S., Wilberforce, M., Knapp, M., Challis, D., Fernández, J.L., Jones, K., Netten, A., 2011. Joining up government by integrating funding streams? The experiences of the individual budget pilot projects for older and disabled people in England. International Journal of Public Administration 34 (4), 232–243.
Moran, N., Glendinning, C., Wilberforce, M., Stevens, M., Netten, A., Jones, K., Manthorpe, J., Knapp, M., Fernández, J.-L., Challis, D., Jacobs, S., 2012. Older people's experiences of cash-for-care schemes: evidence from the English individual budget pilot projects. Ageing and Society 33 (5), 826–851.
Morris, J., 2006. Independent living: the role of the disability movement in the development of government policy. In: Glendinning, C., Kemp, P.A. (Eds.), Cash and Care: Policy Challenges in the Welfare State. The Policy Press, Bristol.
Needham, C., 2007. The Reform of Public Services Under New Labour: Narratives of Consumerism. Palgrave Macmillan, Basingstoke.
Needham, C., 2011. Personalising Public Services: Understanding the Personalisation Narrative. The Policy Press, Bristol.
Newman, J., Glendinning, C., Hughes, M., 2008. Beyond modernisation? Social care and the transformation of welfare governance. Journal of Social Policy 37 (4), 531–557.
Prime Minister's Strategy Unit, 2005. Improving the Life Chances of Disabled People. Cabinet Office, London.
Scourfield, P., 2008. Going for brokerage: a task of 'independent support' or social work? British Journal of Social Work 41 (3), 858–878.
Skills for Care, 2008. Employment Aspects and Workforce Implications of Direct Payments. Skills for Care, Leeds.
Skills for Care, 2010. The State of the Adult Social Care Workforce in England 2010. Skills for Care, London.
Wanless, D., 2006. Securing Good Care for Older People. Taking a Long-term View. Kings Fund, London.

Community Genetics

Leo Pieter ten Kate, VU University Medical Center, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Community genetics was defined by an international group of authors in 2010 as "the art and science of the responsible and realistic application of health and disease-related genetics and genomics knowledge and technologies in human populations and communities to the benefit of the individuals therein. Community genetics is multi-, inter- and transdisciplinary and aims at maximizing benefits while minimizing the risk of harm, respecting the autonomy of individuals and ensuring equity."

Activities and areas of interest include genetic screening, genetic literacy and education, access and quality of genetic services, genetics in primary care, genetics in low- and middle-income countries, genetics in disadvantaged subpopulations, registries of congenital and genetic disorders, genetics in preconception care, public consultation about genetic issues, and epidemiological, economic, psychosocial, ethical, legal, and policy issues.

Community genetics was first mentioned in the scientific literature in 1975. The international scientific journal *Community Genetics* was published by Karger in the period 1998–2008 and was succeeded by the *Journal of Community Genetics*, published by Springer, since 2010.

The international multidisciplinary community genetics network was founded in 2007. At the time of writing the network had 1058 members residing in 74 different countries.

Introduction

The term 'community genetics' has different meanings in medicine and biology. In biology, it refers to the study of evolutionary processes of interacting populations of organisms in a circumscribed location, while in the field of medicine, it refers to the application of genetics to the study and service of humans in a given community. Although communities may be defined by location (such as the inhabitants of a village, town, or country), they may also be defined by history (migrants from one country to another or specific ethnic groups within a country), or by a common problem (e.g., all carriers of a certain recessive gene in a certain population), and also on the basis of interaction with another organism (such as the malaria parasite).

Origin and Further Evolution

The earliest known use of the term 'community genetics' was found in a paper describing the cytogenetic observations in a genetic clinic servicing the community of Tulsa, Oklahoma (Coldwell et al., 1975). Apparently independent of this early use, the term 'community genetics services' was adopted by WHO in the 1980s, and a more comprehensive view on community genetics was given by Bernadette Modell in a paper in 1992 (ten Kate et al., 2010). As mentioned by Modell and Kuliev (1998), many individuals had already devoted a large part of their lives to applying genetic approaches at the community level long before the term was invented.

A dedicated journal, called *Community Genetics*, was published by the publisher Karger AG, from 1998 till 2008, after which its title was changed to *Public Health Genomics*. Another title, the *Journal of Community Genetics* was published by Springer from 2010 onward. A multidisciplinary international e-mail network with its own newsletter was started in 2007 (ten Kate and Plourde, 2011) and has 1057 members at the time of writing.

Definition

There have been several definitions of community genetics in the past. The most recent definition was drafted by a group of 14 authors from five continents (ten Kate et al., 2010).

Community genetics is the art and science of the responsible and realistic application of health and disease-related genetics and genomics knowledge and technologies in human populations and communities to the benefit of the individuals therein. Community genetics is multi-, inter-, and transdisciplinary and aims at maximizing benefits while minimizing the risk of harm, respecting the autonomy of individuals and ensuring equity.

Activities and Areas of Interest

While a definition may give some impression of what the field of community genetics is about, an overview of activities and other interests under this umbrella may help to make things more vivid.

Genetic Screening

Genetic screening is perhaps the most typical example of a community genetics activity. It is an offer of genetic testing to all persons within a community, or it is targeted to a subset of those people, such as newborns, or pregnant women. It is not restricted to those with a disease or a high risk of developing it as is the case with genetic testing in a clinical setting. A more extensive discussion of this issue can be found elsewhere in this encyclopedia. A recent paper by Kingsmore et al. (2012) on

genetic screening by next-generation sequencing in low- and middle-income (LMI) countries provides food for thought.

Genetic Literacy and Education

Genetic literacy of professionals, public, and policymakers facilitates the introduction of community genetic services and empowers the members of the community to make informed decisions on all matters of genetics arising. As scientific developments in genetics are moving fast and steadily, and as misconceptions about genetics are frequent, there is a need for permanent attention for genetics education for all three target groups.

Access and Quality of Genetic Services

Easy and equal access to high-quality genetic services for members of the population is a prerequisite for offering genetic screening and genetics education. Such services are commonly organized at clinical genetic centers where individuals and families can find genetic diagnosis and counseling for genetic disorders in affected persons in their families. Genetic diagnosis includes taking a thorough family history, syndromology, cytogenetics, DNA testing, prenatal diagnosis, and sometimes also preimplantation genetic diagnosis.

Genetics in Primary Care

Primary care is in a special position to act as a first observer, contact, and an advisor regarding genetic disorders and risks in the population under their surveillance. It is therefore important that primary care is empowered to perform these tasks.

Genetics in LMI Countries

There is a special need for community genetics services in countries where there are limited resources and availability of genetic services. WHO has repeatedly emphasized the importance of community genetics services in LMI countries (WHO, 2010).

Genetics in Disadvantaged Subpopulations

In every country, there are subpopulations that are disadvantaged by, for instance, socioeconomical status, infrastructure, cultural isolation, and educational gaps. Some of these subpopulations are also at special genetic risk. An example is the higher prevalence of sickle-cell anemia among immigrants of African descent in Western countries. Disadvantaged subpopulations need extra attention to ensure that they can benefit from genetic services as well as those who are in a better position.

Registries of Genetic/Congenital Disorders

Registries of genetic and/or congenital disorders contain data on large numbers of persons from defined communities and are of great value for needs assessment, surveillance, and research.

Genetics in Preconception Care

Preconception care offers the possibility to provide a better start of the embryo at conception or in the first weeks of prenatal life. Providing information on many factors like lifestyle, chronic diseases, medication, infectious diseases, and folic acid supplementation may prevent serious consequences for mother and child. Genetic counseling before conception increases the array of reproductive choices for the parents with genetic disorders or with genetic disorders in their families, and in case of consanguinity. 'Genetic aspects of preconception consultation in primary care' was the subject of a special issue of the *Journal of Community Genetics* in 2012 (Special issue, 2012).

Public Consultation about Genetic Issues

There are several examples of community genetics activities based on initiatives by people outside the medical or genetic community. Conversely, it is always advisable to consult the public, opinion makers, and authorities before embarking on a new initiative. There are many ways to do this, for example, by focus groups, surveys, or media discussions.

Epidemiologic Issues

Since gene frequencies and prevalences of genetic diseases differ between populations and also by population substructure, knowledge of these parameters is indispensable for planning and follow-up.

Psychosocial Issues

Psychosocial issues are very important, both when people are developing (or are at risk of developing) a severe disease in themselves or a close family member (e.g., hereditary breast cancer), and when they are confronted pre- or postnatally with a serious disorder (or a risk of such a disorder) in their child and have to choose between two evils like termination of a pregnancy and continuation of a pregnancy that will result in a life full of suffering for their child.

Ethical and Legal Issues

Ethical issues are manifold in this field. Their importance is illustrated by the amount of space devoted to ethical issues in the definition of community genetics: Community genetics aims at maximizing benefits while minimizing the risk of harm, respecting the autonomy of individuals and ensuring equity.

Next generation sequencing will increase the number of occasions in which ethical issues arise, as this procedure results in finding many more mutations in the individual than testing single genes does. While this problem has been discussed already in relation to research findings (Bredenoord et al., 2011), discussion will spread to clinical and community genetics as well (Kingsmore et al., 2012). For instance when offering next generation carrier screening to couples who want to have children, one question to be answered is what to do when for instance a mutation predisposing to a dominant disease is found. This should be discussed during prescreening counseling, but are we able to inform participants adequately about the many possible unsolicited findings?

Legal issues are of course also very important, but as legal situations tend to differ per country or larger region, they should be discussed at the country or regional level.

Economic Issues and Policy Issues

The first questions asked by health decision makers usually concern costs (Modell and Kuliev, 1998). If the costs of treatment of a large number of patients with for instance thalassemia greatly exceed national resources, a screening program to find and alert carrier couples may be justifiable as long as informed choice of carrier couples is guaranteed. In other cases, that is, where national resources are not overstretched, costs of services, as well as costs of absence of services still have to be presented as part of a rational decision-making process.

Delineation

In this section, we will make some remarks regarding differences between community genetics and some other related fields.

Clinical Genetics

Clinical genetics focuses on patients and families with genetic disorders. It is basically a reactive specialty, with some proactive elements. Establishing a diagnosis and providing genetic counseling may be followed by the offer of genetic testing in healthy family members if appropriate. Community genetics is basically proactive since members of the population are approached without prior knowledge of their personal or family history.

Population Genetics/Genomics and Genetic Epidemiology

Population genetics/genomics and genetic epidemiology study the fate of genes and genomes in populations and families, and so provide basic knowledge and understanding useful for community genetics. In the absence of a service component, they do not constitute a type of community genetics on their own.

Public Health Genetics and Genomics

The aim of public health is to prevent disease and improve health. Public health measures are judged by the amount of success achieved in reaching these goals. Preventing disease by eliminating the patient or the individual at risk is, however, a perverted view of prevention. Although employing genetic or genomic technology is acceptable in preventing common diseases occurring late in life in those already born, problems arise when public health approaches stressing prevention are used when it concerns the unborn. Although the majority of respondents in a study of public attitudes toward the introduction of noninvasive prenatal diagnosis described their first response as positive, especially with regard to the individual/medical rationale, they expressed unease concerning public health rationale and societal implications (Kelly and Farrimond, 2012). Community genetics, with its emphasis on the individual and on informed choice, does not carry the odium of public health prevention goal (ten Kate, 2005).

The Community Genetics Network

An international multidisciplinary e-mail network of researchers and others interested in community genetics was started in 2007. Members receive a newsletter with references to papers in the field of community genetics written by the members, a list of upcoming courses and conferences, and messages from members to the other members. The newsletter appeared originally once a month, but now appears twice a month. The appearance of the 50th newsletter in 2010 gave cause to compile a report on the first 3 years of the network, which connected 858 members in 70 countries at that time (ten Kate and Plourde, 2011).

At the time of writing, there are 1058 members in 74 countries. Four hundred and five members reside in the Americas (9 countries), 399 in Europe (30 countries), 147 in Asia (26 countries), 74 in Australia and New Zealand (2 countries), and 33 in Africa (7 countries). The five countries with the largest numbers of members are the United States (298 members), the United Kingdom (127 members), the Netherlands (105 members), Canada (81 members), and Australia (69 members).

The steady growth of the membership is based on invitations to authors of scientific papers on subjects in the field of community genetics on the one end and on spontaneous applications to commgennet@gmail.com by persons who express their interest in the network and newsletter on the other. Membership is free of charge. The multidisciplinary character of community genetics is evident from an alphabetical listing of 20 different disciplines in a small sample of the members: biochemistry, biostatistics, clinical genetics, epidemiology, ethics, genetic counseling, hematology, medical genetics, molecular genetics, nursing, obstetrics, ophthalmology, patient advocacy, pediatrics, pharmacogenetics, psychiatry, psychology, public health, and sociology.

When invited to indicate their interests from a list of possible issues, 51% of members highlighted genetic screening. Ethical and psychosocial issues were each mentioned by 29%, genetic education and legal issues by 28% and 27%, respectively. Interest in epidemiology scored 21%. Access to genetic services and genetics in primary care was mentioned by 18% and 17%, respectively. Lower scores were found for genetic registries, genetics of disadvantaged population, public consultation, and economic issues in decreasing order of adherence.

An analysis of the 1489 papers of members mentioned in the 50 newsletters in the first 3 years of existence of the network gives another indication of the diversity of its members. References to papers of the members are only included in the newsletter if the papers are written in English, are listed in PubMed, and appeared no longer ago than 3 months before publication of the newsletter. The 1489 papers were published in 485 different journals, of which 278 (57%) published only one of these papers (ten Kate and Plourde, 2011).

Membership of the community genetics network is free of charge. Those interested should send an e-mail to commgennet@gmail.com for information.

Discussion

Reporting on what has been achieved so far is an easy task compared to predicting future developments. Nevertheless, at this point in time we can be sure that new diagnostic methods like next-generation sequencing and noninvasive prenatal diagnosis will increase the application of genetic and genomic knowledge on a large scale in human populations and communities to the benefit of the individuals therein.

Acknowledgment

The support in preparation of the Community Genetics Network Newsletter by ECOGENE-21, the Canadian Institute of Health Research (CIHR team in community genetics (grant #CTP-82941)) is gratefully acknowledged.

See also: Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Screening for Disease; Genetics and Indigenous Communities: Ethical Issues; Genetics and Society; Genomics, Ethical Issues In; Incest, Inbreeding, and Their Consequences.

Bibliography

Bredenoord, A.L., Onland-Moret, N.C., Van Delden, J.J., 2011. Feedback of individual genetic results to research participants: in favor of a qualified disclosure policy. Human Mutation 32, 861–867.

Coldwell, J.G., Say, B., Jones, K., 1975. Community genetics I. Journal of the Oklahoma State Medical Association 68, 299–302.

Kelly, S.E., Farrimond, H.R., 2012. Non-invasive prenatal genetic testing: a study of public attitudes. Public Health Genomics 15, 73–81.

Kingsmore, S.F., Lantos, J.D., Dinwiddie, D.L., Miller, N.A., Soden, S.E., Farrow, E.G., Saunders, C.J., 2012. Next-generation community genetics for low- and middle-income countries. Genome Medicine 4 (3), 25 (Epub ahead of print).

Modell, B., Kuliev, A., 1998. The history of community genetics: the contribution of the haemoglobin disorders. Community Genetics 1, 3–11.

Special issue, 2012. Genetic aspects of preconception consultation in primary care. Journal of Community Genetics 3, 2.

ten Kate, L.P., 2005. Community genetics: a bridge between clinical genetics and public health. Community Genetics 8, 7–11.

ten Kate, L.P., Al-Gazali, L., Anand, S., et al., 2010. Community genetics. Its definition 2010. Journal of Community Genetics 1, 19–22.

ten Kate, L.P., Plourde, A., 2011. A short history of the first three years of the community genetics network and its newsletter. Journal of Community Genetics 2, 111–115.

WHO, 2010. Community Genetics Services. http://whqlibdoc.who.int/publications/2011/9789241501149_eng.pdf.

Community Resilience to Disasters

Ben Wisner, Aon-Benfield UCL Hazard Research Centre, University College London, London, UK
Ilan Kelman, Institute for Risk and Disaster Reduction (IRDR) and Institute for Global Health (IGH), University College London (UCL), London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Resilience has become a common word in aiming to articulate how communities should deal with disaster. Both 'resilience' and 'community' are contested through multiple definitions. Different understandings are explored through four artificial and overlapping hazard categories: (1) natural (or environment) hazards; (2) technological hazards; (3) violent social crisis such as terrorism and violent crime; and (4) nonviolent social crisis. The most encompassing and successful approaches for community resilience to disasters involve full participation of all community members in having access to resources and choices for helping themselves deal with day-to-day and longer term challenges.

Introduction

How do communities get through disaster? A common word these days is 'resilience,' so that much of the work in dealing with disasters aims for 'resilience building.' Definitions and conceptualizations of resilience vary, with different disciplines and traditions viewing the term and its application differently. Additionally, the answers to the questions 'resilience by what?' and 'resilience to what?' can be different. Psychologists might focus on the resilience of individuals to external influences. Engineers might highlight the resilience of a building to a force. Here, we focus on 'community resilience to disasters,' considering all forms of disaster across various scales. We start by considering how individuals react to personal calamity, which they might experience.

A loved one dies suddenly. You receive the call no one wants. What you experienced as 'normal' life then stops, pausing for a time. You are unfocused, depressed, and perhaps angry. You do not function as you usually do at work, in the home, on the street, or interacting with people. Your interpretation of the meaning of life is affected, perhaps profoundly through a tsunami of emotion or only weakened as a chimney after a small earthquake (Aloudat and Christensen, 2012: 570–572).

Most often you 'bounce back.' You recover from the disaster and the changes to your worldview. You begin to function as you did before. Your sense of meaning, hope, and expectations are restored. Friendship, love, and solidarity have helped you, as have rituals of remembrance. It takes time, but you celebrate the loved one's life and can turn that life, your contribution to it, and your memories around in your hands like a precious vase or crystalline geode. You miss this person, but you have accepted the passing along with the continuing present absence or absent presence. There may even be an enhanced understanding of life – openness to the losses and suffering of others. Spiritual emergency may have opened out into spiritual emergence (Ashley et al., 2008).

This is the way one generally experiences personal crisis or personal disaster, namely losses of many sorts that involve profound transitions – death, divorce, or loss of a job. The ability to 'bounce back' has been one popular interpretation of 'resilience' (e.g., Smith et al., 2008), which has also been advanced as being the ability to 'bounce forward' into a more positive state of mind and higher level of functioning (Manyena et al., 2011). These interpretations have been applied to individuals and collectives, such as ecosystems and human systems. How relevant are such interpretations for communities?

In order to discuss community resilience, it is necessary to define two slippery words: 'community' and 'resilience.'

What is Community?

The term 'community' emerged in the Middle Ages with its origins in 'common,' in the sense of fellowship, public, or joint ownership. Academically, the word might date from nineteenth-century German sociology, since when it has become popular within many diverse disciplines. Its meaning is contested. Bell and Newby (1973) identified 98 different uses of the word and these have expanded (Dominelli, 2006). Additionally, the term may be used to cover up exploitation and conflict (Agrawal and Gibson, 1999; Walmsley, 2006) with Walmsley (2006) particularly highlighting the lack of homogeneity in most communities, despite the term 'community' seeming to imply consistency. Usually in a community, there will be some individuals who are marginalized from access to decision making for reasons including stigma, past history, ethnicity, race, religion, gender, sex, age, ability, health status, immigration status, and sexual preference. There will be individuals with more power, wealth and access to information, financial and material resources – and some with less.

Here, we take 'community' to mean people living in a locality and their extended networks elsewhere in space and time. The latter is important where people believe that their deceased ancestors are an intimate part of their daily lives. The spatial aspect might not be continuous, but the community could involve different groups, which are a large distance apart, such as a diaspora.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28019-7

What Is Resilience?

Alexander (2013) gives a detailed etymology of the term 'resilience,' starting with its Latin meaning of 'bounce' and taking it through more than two millennia until it started to be associated with 'rebounding' rather than just bounce. Into the twentieth century, the term was increasingly used in medicine, engineering, and psychology, followed by natural ecology and then human ecology. In recent decades, economics and development studies among many other fields adopted their own views of the meaning of 'resilience.' For example, Manyena (2006) and Gaillard (2007) provide background and interpretations from a disaster risk reduction (DRR) perspective.

Consequently, 'resilience' and 'resiliency' today have numerous, disparate definitions and interpretations, as does 'community.' Resilience has, therefore, shifted away from simply meaning to 'bounce back' like materials under stress or a biological population in an ecosystem when many individuals are killed by predators or disease (Levine et al., 2012). A common contemporary interpretation based on ecology comes from Turnbull et al. (2013: 9) who define 'resilience' to be:

> … [T]he capacity of an individual, household, population group or system to anticipate, absorb, and recovery from hazards and/or effects of climate change and other shocks and stresses without compromising (and potentially enhancing) long-term prospects.

The basis for this definition has been significantly critiqued (e.g., Lewis and Kelman, 2010). Hewitt (1983) objects to the characterization of disasters as sudden shocks, considering the long-term vulnerability processes, which must exist for a disaster to manifest (see also Lewis, 1999). Certainly, the manifestation of climate change over generations can hardly be accepted as being a shock, which the definition above implies. Meanwhile, Dominelli (2012) suggests that resilience has stochastic properties, which do not always result in a linear movement forward. People could move backward and forward along a continuum that includes lack of extensive resilience, resilience sufficient for survival, or resilience maintaining the previous state (before the disaster); and through to resilience that is transformative and enables people to flourish, thrive, and do better.

Simultaneously, the notion of 'normality' is contested. Wisner (1993) argues that disasters are prefigured in the conditions of daily life. For subgroups within a community who have the least control of the wealth have access to few resources and who suffer discrimination or marginalization, 'normal' life is rife with everyday disasters such as seasonal flooding, shack fires, crime, and abuse by policy makers. Vulnerability to challenges of daily life are ever-present such as getting education and food alongside less usual experiences such as earthquakes and tornadoes (Wisner et al., 2004, 2014). Hewitt (1997: 143) considers vulnerability to be "embedded in the social geography of settlements and land uses, and the space of distribution of influence in communities and political organisation." That is normal, which shows itself to be remarkably resilient. There is no particular reason why, after a disaster, people would desire to 'bounce back' to that state of vulnerability since that normality, that state of vulnerability, caused the disaster in the first place (see also Glantz and Jamieson, 2000).

What Is Community Resilience?

Given the heterogeneity of subgroups within any community – with people stratified in terms of decision making, capacity and vulnerability, by wealth, income and sometimes ethnicity and religion, alongside commonplace differences such as gender, sexual orientation, age, and ability – there are both winners and losers while and after a disaster manifests. Often, humanitarian assistance, recovery, and reconstruction reestablish that status quo because resilience is seen as a return to the predisaster 'normal,' rather than trying to create a different community, which solves the problems that led to the disaster in the first place, frequently espoused by the tagline 'Build Back Better.' By returning to the status quo, vulnerability to disaster remains the same, or can even increase, while capacity and resilience can decrease, instead with those in control gaining resources, capacity, and resilience (Susman et al., 1983).

With risk being an intersection of hazard and vulnerability (Wisner et al., 2004), returning to the predisaster 'normal' can actually increase risk by doing nothing about the hazard while increasing vulnerability; that is, taking a narrow view of resilience can increase risk. Similarly, any lack of risk reduction, risk prevention, or preparedness will simply be reinstated, achieving little with regards to broader views of resilience. Instead, resilience depends on access to various kinds of resources that provide members of subgroups with a community with capacity to cope. The relationship between capacity and access to six kinds of resources is schematized in **Figure 1** (Wisner et al., 2012: 28).

Paton (2007) offers a schema that overviews resilience from the point of view of individuals, communities, and institutions. Under each of these headings, they list a series of characteristics or capacities providing the basis for 'bouncing forward' when a disaster causes disruption and loss. Of note is the recurrence of the word 'resources' in each list, reminding one of the unequal distribution of resources required for the capacities to recovery (and summarized in **Figure 2**) to materialize. The double-headed arrows suggest interaction and complex feedbacks, also noting that institutions upon which communities and individuals normally depend are likely to be disrupted by disaster. Other new institutions can be introduced, including those brought in by humanitarian actors, but their routines and cultures are different (Robinson, 2008).

Community resilience, therefore, involves interactions among individuals, groups, and institutions that usually result in collective action to enhance the capacities for recovering from a disaster. Acts of solidarity with those in need may emerge and transformative structural change could be effected if it is to transcend individual resilience, particularly, among those who are especially vulnerable in a given situation (Dominelli, 2012).

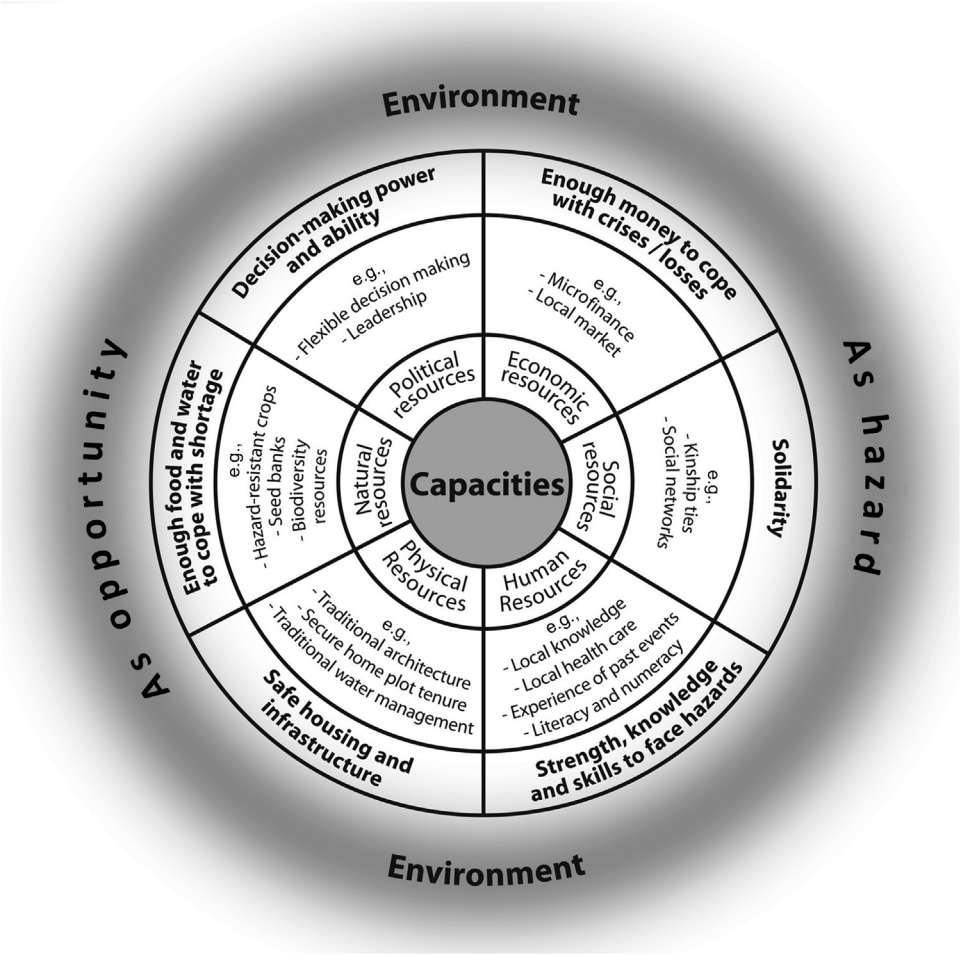

Figure 1 Resources and capacities for resilience. From Wisner, B., Gaillard, J.C., Kelman, I. (Eds.), 2012. The Routledge Handbook of Hazards and DRR. Routledge, London, p. 28.

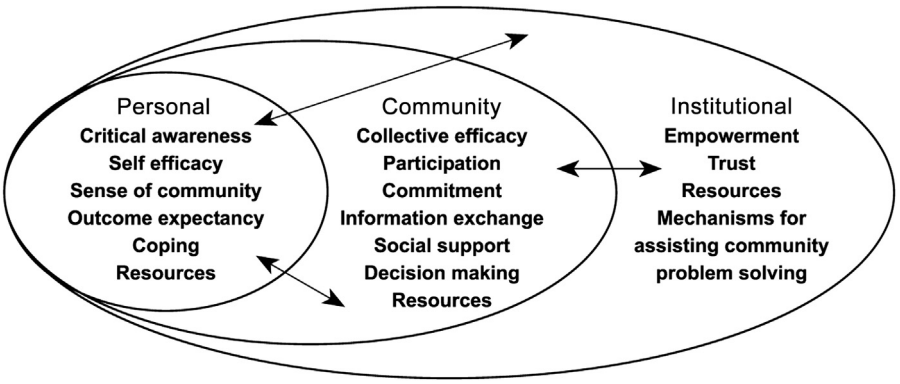

Figure 2 Resilience model. From Paton, D., 2007. Measuring and Monitoring Resilience in Auckland. GNS Science report 2007/18. GNS Science, Lower Hutt, New Zealand, p. 14.

Disasters and Community Resilience

When trying to understand community resilience, it is helpful to examine the range of hazards that may affect a community, stressing component subgroups within that community to differing degrees. Broadly speaking, these fall into four categories, using a mixture of vocabulary, which in itself, can be legitimately challenged: (1) natural (or environment) hazards; (2) technological hazards; (3) violent social crisis such as terrorism and violent crime; and (4) nonviolent social crisis

with examples being economic changes, corruption, and development-induced displacement. In addition to the vocabulary difficulties, this neat typography is troubled by the overlapping insults to minds, bodies, and livelihoods seen in reality – sometimes even perpetuated by the community rather than just happening to the community.

As an example of overlap, natural hazards may manifest themselves, realizing risk, in situations of violent conflict and displacement. On 26 December 2004, an earthquake and tsunami struck Aceh, Indonesia, and then the tsunami spread out across the Indian Ocean with Sri Lanka being the second worst country hit after Indonesia. Both locations at the time were in the grip of a long-running violent conflict. Sexual violence and trafficking in women and children hang over the best-run refugee and postdisaster encampments (Fordham, 2012). Natural hazards often trigger the catastrophic failure of pipelines, refineries, ships, and other technological facilities carrying or storing chemicals, or used for food and water distribution such as harbors and airports, causing additional burdens and losses. Despite humanitarian assistance, a large percentage of small businesses never recover from a disaster, especially when already being impacted by wider scale economic downturns (Sanderson, 2012).

That is, community resilience is not just about dealing with a specific disaster (or 'event' or 'shock') in space and time, but is also about capacities to deal with multiple vulnerabilities and intersecting, multiple hazards leading to multiple risks and crises, any of which might be conjunctive, cascading, or coincidental (simultaneously or not). In a sense, these characterizations are the disaster(s) to which communities seek resilience. Without forgetting these wider contexts, and recognizing that many of the points made in one section apply equally to the other three sections, understandings of community resilience for each of the four categories are now explored further.

Community Resilience and Natural Hazards

Communities may experience severe physical and social (including economic) harm as a result of natural events and processes such as earthquakes, tsunamis, floods, and volcanic eruptions. That is not saying that all such impacts are necessarily damaging or detrimental, since volcanic ash and floodwater, for example, bring nutrients to soil permitting people to live in locations affected by these environmental processes. But the difficulties and disasters should not be denied, even where there are gains as well as losses, and winners as well as losers.

People who lose infrastructure and livelihoods to natural hazards often wish to rebuild in the same place (Oliver-Smith, 1986; Vale and Campanella, 2005) irrespective of the chances of the same hazards recurring. To them, the familiarity of the place and living as they always have done, including owning their land and property if they did before the disaster, is part of their community resilience. While some success stories do exist, so do many failures, with one example being Simpson's (2014) account of the mixed results of humanitarian assistance 10 years after the 2001 earthquake in Gujarat, India.

A misnomer in the disasters field is that reconstruction means recovery and represents resilience. Merely putting up buildings and other infrastructure does not indicate how people, their families, their friends, and the wider social networks are doing. After a flood, are they afraid every time it rains? Were losses compensated for, or do the people need to start over financially? When loved ones are killed, what does 'recovery' mean for those left behind? The answers to such questions are usually different for different social classes, castes, genders, sexes, and other subgroups within a community. Simpson (2014), Boano (2013), and Sanderson et al. (2012), among others, have shown how humanitarian assistance is often skewed in ways that reinforce caste and class differences, wealth and income gaps, and spatial segregation.

As a case study, about one-third of the population of New Orleans, Louisiana never returned after Hurricane Katrina forced a massive evacuation in 2005. These nonreturnees were generally the poorest, many of them African-American, who were evacuated to Houston, Texas, and other cities (Dyson, 2007). Some more affluent residents also chose not to return, but the key word here is 'chose' – they have the power and resources to make their own decision and, as such, could be deemed to be individually resilient. But when large swathes of the population do not have similar choices, then the community can hardly be termed 'resilient' – even if infrastructure is renewed.

Community Resilience and Technological Hazards

The systematic study of the impact of large-scale technological hazards on communities might have begun in Canada with Henry Prince's PhD dissertation on the catastrophic 1917 explosion of a ship in Halifax, Nova Scotia's harbor, which killed approximately 2000 people (Prince, 1920). Canada also provides a contemporary example of a transportation-related technological disaster. In July 2013, a runaway train literally tore the heart out of the small town of Lac Mégantic, Quebec, when cars carrying crude oil derailed and exploded near the town's center, killing 47 people and displacing another 2000 (BBC, 2013). The survivors in Halifax rebuilt their city and lives, while the evacuees from Lac Mégantic are in the process of rebuilding their town, but they will never be able to replace the archives and historic buildings lost. These Canadian cities represent community resilience in terms of living with grief and moving forward from the catastrophe rather than letting the disaster consume them. Community resilience in the sense of returning to the predisaster state could never be achieved, since so much was lost.

Other technological hazards often preclude certain assumptions made about community resilience, such as the generations of Ukrainians and Japanese dealing with the legacy of nuclear power plant disasters. In 1986 at Chernobyl and in 2011 at Fukushima, radiation releases left settlements and the countryside uninhabitable around the plants. Personal stories indicate the challenges of community resilience in the face of permanent upheaval.

Two years after the Fukushima disaster, 150 000 people are still displaced from homes due to radiation. Despite more social services and housing assistance than most people can expect elsewhere in the world, these evacuees are unhappy. In the words of one, "I just want to get our life back to how it was before. We were 20 minutes from both [my wife's and my]

parents. We'd have barbeques on the weekends.… The hardest thing is that we can no longer do the things that seemed so natural before" (Bird, 2013: online). Smith (2013) calculates that the total number of deaths due to the effects of evacuation in Fukushima Prefecture is nearly equal to those killed in the original earthquake and tsunami (Smith, 2013: online):

> Causes of death in the aftermath have included 'fatigue' due to conditions in evacuation centers, exhaustion from relocating, and illness resulting from hospital closures. The survey also said a number of suicides had been attributed to the ordeal. Francis Markus, East Asia spokesperson for the International Federation of Red Cross and Red Crescent Societies (IFRC), said the conditions faced by those displaced is made worse by them not knowing when they can return.

One of the greatest challenges to community resilience, as seen by the effects on emotional well-being and other mental health aspects, is long-term or permanent displacement from one's home (Long, 2011).

Community Resilience and Violent Social Crisis

Communities experience many forms of disasters linked to violence, such as terrorist attacks, gun massacres, and assassinations, all of which illustrate aspects of community resilience. In these cases, as in all others, there is differential distribution of resources (**Figure 1**) that fuel the capacities required for resilience (listed in **Figure 2**). This holds within subgroups of communities as well as applying among whole populations and societies. Communities affected by violence, e.g., in Argentina, the USA, and Democratic Republic of Congo (DRC) do not have access to the same resources. Some other reasons include struggles over resources, but also, religious and nationalistic claims – how people see themselves in the world.

Activism in emergent groups working to prevent a recurrence of violence and to heal the emotional wounds may provide new, supplementary, spatially diffuse communities of experience and solidarity. These include those like the 'Mothers of the Disappeared' in Argentina (Hernandez, 2012) and 911 Families Association (http://911families.org/) in the USA, and may provide a way that gives the personal and collective loss meaning. Not all groups have such options, such as those who experienced rape used as a war weapon in DRC over past years. With some minor, externally funded exceptions, the women and girls affected have few options to create solidarity among themselves and to provide collective remembrance, because they have little or no political voice and resources (Women's Institute for Gender Justice, 2013).

Other populations have managed to set up collective processes, ostensibly for everyone, to try to build community resilience after a violent period. The process known as 'truth and reconciliation,' often associated with South Africa's post-apartheid coming to terms of that period but also replicated across many other postconflict situations, is controversial. Some assert that transparency and confession of violent acts by the perpetrators are part of the necessary honesty, grieving, and healing to ensure that the society does not return to 'normal,' but instead accepts what happened and moves on to do better. Others suggest that such a process leaves those who suffered violence unsatisfied when there is no punishment. It is easy to speak philosophically of restorative vs retributive justice when one has not been victimized or known someone who was tortured or killed (Kruks, 2012). In reality, it can be much harder to feel resilient or to support community resilience when those who caused suffering suffer no consequences.

Community Resilience and Nonviolent Social Crisis

Community resilience is being used as a tagline to deal with expectations of long-term, nonviolent crises, such as reduced availability of natural resources and creeping environmental changes including climate change. 'Localization' movements have emerged, seeking to make food, water, energy, livelihoods, and money much more local and hence less dependent on external factors or national support. Localization is seen as building community resilience by being able to deal with crises in the absence of external support.

An example is concerns about limited fossil fuel supplies, because the cost becomes prohibitive, supply is disrupted, or diminishing resources – all of which are linked. Consequences envisioned of limited fossil fuel supplies include food availability being reduced because vehicles have limited petrol to transport supplies to supermarkets; blackouts during hot and cold temperature extremes because electricity generating plants depending on fossil fuels must shut down; and emergency services being unable to respond promptly due to petrol restrictions. Making food, water, energy, and manufacturing more locally based reduces dependence on fossil fuels.

Localization also helps to build community resilience against other major disruptions, such as economic instability, a rapid decrease of water resources, and the sudden decentralizing of large sovereign states – with a prominent example being the sudden dissolution of the USSR in 1991. Some of the USSR's allies, such as Cuba, lost a major supplier and rapidly had to adjust to the lack of food and fuel being imported.

Building community resilience to nonviolent social crisis is much more than considering what happens externally to a community. It is also about the community itself. The construction and perpetuation of poverty can be due to deliberate exploitation of people and the construction of a social order favoring those with power, a process termed 'quiet violence' (Hartmann and Boyce, 1984), which effectively precludes a community resilient to anything. Vulnerability and lack of capacity to reduce it, and to build resilience, are aggravated by poverty, preventing any opportunity to 'bounce forward' after other hazards impact. Due to the persistent, ongoing crisis of poverty, it is hard enough just trying to 'bounce back' to the original state of poverty, so community resilience becomes a dream rather than a goal, blocked by the same powers, which create and perpetuate poverty in the first place.

Community-Based DRR

In the face of all these diverse challenges, with the inherent barriers and opportunities for community resilience,

what practical measures should communities implement? Numerous approaches have been developed and tested, inviting and supporting groups of people living in the same localities – communities or subgroups therein – to assess the specific hazards or crises they face; to identify and analyze the people, livelihoods, places, and assets in harms way (effectively vulnerability); and to appraise their own resources and capacities to reduce losses or even to prevent the hazard or crisis, thereby further indicating what external support they might need.

These approaches are embedded within the communities, but can be facilitated by external people so that no one in the community is left out and to ensure a balanced approach. This suite of methods has become known as Community-based Disaster Risk Reduction (CBDRR), which here, provides an operational process for supporting community resilience (e.g., **Box 1**). The importance of CBDRR is now recognized by most organizations working in DRR and is also promoted through the London-based Global Network of Civil Society Organisations for Disaster Reduction (GNDR, 2011).

As shown by **Box 1**, one of the keys for successful CBDRR for building community resilience is to ensure full and fair participation by all those in the community, but without necessarily handing over full control of the process to community leaders. As with everyone, community leaders have their biases and prejudices, which can inhibit resilience building. In fact, one interesting part of many participatory processes for CBDRR, including P3DM, is that neither the external participants nor community members are able to achieve the best results for community resilience on their own. Instead, the parties learn from each other, exchanging knowledge and wisdom, meaning that those external to community take plenty back with them to build resilience in their own community – if they choose to do so, preferably with external supporters for their own CBDRR.

Box 1 Participatory 3-Dimensional Mapping for enhancing community resilience to disasters

Participatory 3-Dimensional Mapping (P3DM) consists of building stand-alone large-scale relief maps made of locally available and cheap materials (e.g., carton, paper, cork) on which people place thematic layers of geographic information. P3DM enables the mapping of landforms and topographic landmarks, land cover and use, and anthropogenic features, which can be depicted by push-pins (points), yarn (lines), and paint (polygons). Participants plot land use and other geographical features and differentiate them according to their vulnerability. Pins of different shapes, sizes and colors can differentiate building materials and can locate the most vulnerable people in the community.

Community members can also delineate hazard-prone areas and locate local resources for dealing with hazards. It is then easy and quick to evaluate disaster risk based on hazards, threatened assets, vulnerabilities, and capacities. P3DM enables the planning of DRR measures to enhance community resilience based on multistakeholder group discussions, helping to find consensus among participants.

As such, P3DM fosters the participation and collaboration of a large range of stakeholders, including scientists, government officials, NGOs, and local communities, permitting the integration of bottom-up and top-down DRR measures. P3DM provides a tangible tool in which the most marginalized people, who may have a poor understanding of scientific concepts, can discuss DRR with scientists, who may have a poor understanding of the local context. All stakeholders can contribute their knowledge on the same tool and in the same forum. P3DM is credible to locals, who build the map and plot most of the information, and to scientists alongside government representatives who can easily overlap their own data and plans on scaled and geo-referenced maps. In the process, NGOs often serve as facilitators and moderators.

Despite all its advantages, P3DM is not a stand-alone tool and needs to be combined with other participatory tools. The reason is to capture dimensions of vulnerabilities and capacities, which are difficult to plot on a map, e.g., interpersonal relationships and temporal patterns. In addition, as for all participatory tools, the contribution and sustainability of a P3DM project depend upon the process of participation and the respect of local needs and aspirations. Having a pretty map is not an objective per se. Instead, it needs to be used to foster a dialogue among those involved for DRR, to facilitate disaster risk assessment, and to contribute to enhancing community resilience.

For more information, see http://www.preventionweb.net/english/professional/trainings-events/edu-materials/v.php?id=35534

Written by JC Gaillard (School of Environment, The University of Auckland, New Zealand) and Jake Rom D. Cadag (Institut de Recherche pour le Développement, France).

Conclusions: Resilience and Transformation

In the end, one must consider accumulating doubts about the uses of the word 'resilience' in the context of major challenges to the physical and mental well-being, assets and livelihoods of subgroups within communities – placing those doubts beside doubts about the word 'community.' Such doubts apply to all the categories described above. Too often, the use of the term 'resilience' is reductionist and technocratic, assuming that ecological concepts can be mapped directly onto society. Mechanistic approaches can masquerade as being all-encompassing, such as 'socioecological systems' dominating much community resilience work, when in reality, many (not all) such analyses emphasize paradigms from ecology, while obscuring the deeper politics of risk, power, equality, and access to resources.

Because such standard approaches are not achieving community resilience, as evidenced by increasing and expanding disaster losses, especially affecting those in society least able to cope with the losses (GNDR, 2011), many are now calling for community resilience to be achieved by transformation and transformative processes. Academic work has a long pedigree of analyzing societal transformation (e.g., Polanyi, 1944; and all the ensuing critiques, e.g., Lie, 1991). Major transformations have certainly been achieved in the past – such as banning slavery, outlawing smoking in indoor public places, and permitting women to vote – but what does it mean for community resilience to disasters?

The answer is that the contemporary discourse might be somewhat naïve in expecting communities to build resilience through transformation. As we have seen, the fundamental resilience challenges in communities emerge at the basic level: individuals and families dealing with reality day-to-day, on top which a rarer hazard might manifest, all of which can lead to disasters. Similarly, we have seen how the fundamental solutions to building wide concepts of resilience to disasters, including the day-to-day disasters,

come from the community at the basic level: individuals and families making better use of what they have already, including knowledge and assets.

Community resilience to disasters does not mean keeping the status quo nor does it mean sudden, overwhelming change. Instead, it means continuing to improve over the long term and seeking changes at all scales so that people have increased access to options and resources to help themselves do better in maintaining rewarding livelihoods and in living. That builds toward the long term, sustained process of resilience that communities need to reduce the risk of disaster.

See also: Afrocentric Approaches to Social Work; Bereavement; Citizen Mapping; Development Studies; Ecological Social Work: Shifting Paradigms in Environmental Practice; Environmental Cognition, Perception, and Attitudes; Environmental Risk; Environmental Sociology; Indigenous Social Work; Participatory Action Research in Environmental and Ecological Studies; Post-Traumatic Stress in Social Work; Risk Assessment in Social Work; Stress, Coping and Health.

Bibliography

Agrawal, A., Gibson, C., 1999. Enchantment and disenchantment: the role of community in natural resource conservation. World Development 27 (4), 629–649.

Alexander, D.E., 2013. Resilience and disaster risk reduction: an etymological journey. Natural Hazards and Earth System 13, 2707–2716.

Aloudat, T., Christensen, L., 2012. Psycho-social recovery. In: Wisner, B., Gaillard, J.C., Kelman, I. (Eds.), The Routledge Handbook of Hazards and Disaster Risk Reduction. Routledge, London, pp. 569–579.

Ashley, W., Samet, R., Tatim, M., 2008. From honeymoon to disillusionment to reconstruction – thinking outside the box. In: Roberts, S., Ashley, W. (Eds.), Disaster Spiritual Care. Skylight Paths Publishing, Woodstock, VT, pp. 139–161.

BBC (British Broadcasting Company), 27 July 2013. Canada Train Blast: Mourners Attend Lac Megantic Memorial. http://www.bbc.co.uk/news/world-us-canada-23480091.

Bell, C., Newby, H., 1973. Community Studies: An Introduction to the Sociology of the Local Community. Praeger, New York.

Bird, W., March 10, 2013. Two years on, Fukushima evacuees seek justice and a normal life. Japan Times. http://www.japantimes.co.jp/life/2013/03/10/people/two-years-on-fukushima-evacuees-seek-justice-and-a-normal-life/#.Up_4ISeBWSo.

Boano, C., 2013. Post-disaster recovery planning: introductory notes on its challenges and potentials. In: Lopez-Carresi, A., Fordham, M., Wisner, B., Kelman, I., Gaillard, J.C. (Eds.), Disaster Management: International Lessons in Risk Reduction, Response and Recovery. Earthscan, London, pp. 191–210.

Dominelli, L., 2006. Women and Community Action. Policy Press, Bristol.

Dominelli, L., 2012. Green Social Work. Polity Press, Cambridge.

Dyson, E., 2007. Great migrations? In: Dyson, E. (Ed.), Come Hell or High Water: Hurricane Katrina and the Color of Disaster. Basic Civitas Books, New York, pp. 213–222.

Fordham, M., 2012. Gender, sexuality and disaster. In: Wisner, B., Gaillard, J.C., Kelman, I. (Eds.), The Routledge Handbook of Hazards and Disaster Risk Reduction. Routledge, London, pp. 424–435.

Gaillard, J.C., 2007. Resilience of traditional societies in facing natural hazards. Disaster Prevention and Management 16, 522–544.

Glantz, M.H., Jamieson, D., 2000. Societal response to Hurricane Mitch and intra-versus intergenerational equity issues: whose norms should apply? Risk Analysis 20, 869–882.

GNDR, 2011. "If We Do Not Join Hands"…Views from the Frontline. GNDR (Global Network of Civil Society Organisations for Disaster Reduction), Teddington.

Hartmann, B., Boyce, J.K., 1984. A Quiet Violence: View from a Bangladesh Village. Zed Books, London.

Hernandez, V., 28 April 2012. Argentine Mothers Mark 35 Years Marching for Justice. BBC World News. http://www.bbc.co.uk/news/world-latin-america-17847134.

Hewitt, K. (Ed.), 1983. Interpretations of Calamity. Allen and Unwin, London.

Hewitt, K., 1997. Regions of Risk: A Geographical Introduction to Disasters. Longman, London.

Kruks, S., 2012. 'Eye for eye': the question of revenge. In: Kruks, S. (Ed.), Simone de Beauvoir and the Politics of Ambiguity. Oxford University Press, New York.

Levine, S., Pain, A., Bailey, S., Fan, L., 2012. The Relevance of 'resilience'? HPG Policy Briefs 49. Overseas Development Institute, London. http://www.odi.org.uk/sites/odi.org.uk/files/odi-assets/publications-opinion-files/7818.pdf.

Lewis, J., 1999. Development in Disaster-Prone Places: Studies of Vulnerability. Intermediate Technology Publications, London.

Lewis, J., Kelman, I., 2010. Places, people and perpetuity: community capacities in ecologies of catastrophe. ACME: An International E-Journal for Critical Geographies 9 (2), 191–220.

Lie, J., 1991. Embedding Polanyi's market society. Sociological Perspectives 34 (2), 219–235.

Long, K., 2011. Permanent Crises? Unlocking the Protracted Displacement of Refugees and Internally Displaced Persons. Policy Overview. Refugee Studies Centre, Oxford Department of International Development, University of Oxford. http://www.rsc.ox.ac.uk/publications/rsc-reports/Unlocking-protracted-displacement-policy-overview.pdf.

Manyena, S.B., 2006. The concept of resilience revisited. Disasters 30, 433–450.

Manyena, B., O'Brien, G., O'Keefe, P., Rose, J., 2011. Disaster resilience: a bounce back or bounce forward ability? Local Environment 16 (5), 417–424.

Oliver-Smith, A., 1986. The Martyred City: Death and Rebirth in the Andes. University of New Mexico Press, Albuquerque, NM.

Paton, D., 2007. Measuring and Monitoring Resilience in Auckland. GNS Science report 2007/18. GNS Science, Lower Hutt, New Zealand.

Polanyi, K., 1944. The Great Transformation. Rinehart, New York.

Prince, H., 1920. Catastrophe and Social Change. Based upon a Sociological Study of the Halifax Disaster (Ph.D. dissertation). Colombia University, New York.

Robinson, J., 2008. From honeymoon to disillusionment to reconstruction: working with outside groups to help rebuild a community. In: Roberts, S., Ashley, W. (Eds.), Disaster Spiritual Care. Skylight Paths Publishing, Woodstock, pp. 162–187.

Sanderson, D., 2012. Livelihood protection and support for disaster. In: Wisner, B., Gaillard, J.C., Kelman, I. (Eds.), The Routledge Handbook of Hazards and Disaster Risk Reduction. Routledge, London, pp. 697–710.

Sanderson, D., Sharma, A., Anderson, J., 2012. NGO permanent housing ten years after the Gujarat earthquake. Environment and Urbanisation 24 (1), 233–247.

Simpson, E., 2014. The Political Biography of an Earthquake: Aftermath and Amnesia in Gujarat, India. Oxford, New York.

Smith, A., 10 September 2013. Fukushima Evacuation Has Killed More than Earthquake and Tsunami. NBC (National Broadcasting Company) World News. http://worldnews.nbcnews.com/_news/2013/09/10/20420833-fukushima-evacuation-has-killed-more-than-earthquake-and-tsunami-survey-says.

Smith, B.W., Dalen, J., Wiggins, K., Tooley, E., Christopher, P., Bernard, J., 2008. The brief resilience scale: assessing the ability to bounce back. International Journal of Behavioral Medicine 15 (3), 194–200.

Susman, P., O'Keefe, P., Wisner, B., 1983. Global disasters: a radical perspective. In: Hewitt, K. (Ed.), Interpretations of Calamity. Allen and Unwin, London, pp. 263–283.

Turnbull, M., Sherrett, C., Hilleboe, A., 2013. Toward Resilience: A Guide to Disaster Risk Reduction and Climate Change Adaptation. Catholic Relief Services and Emergency Capacity Building Project, Baltimore, MD. http://www.ecbproject.org/new-practitioners-guide-to-disaster-risk-reduction-drr/practitioners-guide-to-drr–cca.

Vale, L., Campanella, T., 2005. The Resilient City: How Modern Cities Recover from Disaster. Oxford, New York.

Walmsley, J., 2006. The nature of community: putting community in place. Dialogue 25 (1), 5–12.

Wisner, B., 1993. Disaster vulnerability: scale, power, and daily life. GeoJournal 30 (2), 127–140.

Wisner, B., Blaikie, P., Cannon, T., Davis, I., 2004. At Risk: Natural Hazards, People's Vulnerability and Disasters, second ed. Routledge, London.

Wisner, B., Gaillard, J.C., Kelman, I. (Eds.), 2012. The Routledge Handbook of Hazards and Disaster Risk Reduction. Routledge, London.

Wisner, B., Kelman, I., Gaillard, J.C., 2014. Hazard, vulnerability, capacity, risk and participation. In: Lopez-Carresi, A., Fordham, M., Wisner, B., Kelman, I., Gaillard, J.C. (Eds.), Disaster Management: International Lessons in Risk Reduction, Response and Recovery. Earthscan, London, pp. 13–22.

Women's Institute for Gender Justice, 2013. Our Voices Matter: Congolese Women Demand Justice and Accountability. http://www.iccwomen.org/videos/index.php.

Community Sociology

Brian Jacobs, Staffordshire University, Stoke-on-Trent, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sociologists have focused in various ways on the structures, institutions, social relations, and economic changes affecting communities. Insights have been provided by community studies and through the broadening perspectives developed by urban sociology. With developments in social and cultural theory, researchers have increasingly emphasized the ordering of social life by meanings, beliefs, identities, and cultures. All this has produced multiple interpretations of the word *community* with contending approaches and interpretations.

Community Sociology

Sociologists have focused in various ways on the structures, institutions, social relations, and economic changes affecting communities. With developments in social and cultural theory, researchers have increasingly emphasized meanings, beliefs, identities, and cultures in ordering social life. All this has produced multiple interpretations of community, with contending approaches and interpretations.

The Classical Sociologists

The classical sociologists addressed changes taking place in modern society and how those changes affected social and communal relationships. The classical sociologists expressed a sense of something lost relating to community.

Auguste Comte, in *The Positive Polity* (1851–54), gave a utopian response to modern society. For Comte, the conditions existed for the creation of an ordered science- and technology-based society. He envisaged the restoration of morality and community with a new business class leading the march to social progress. Comte's moral community was hierarchical and corporatist, with positivist society subsuming the individual and the family in inculcating moral authority.

Frederic Le Play, in his 1855 six-volume work *The Workers of Europe*, employed social survey techniques to analyze differing patterns of kinship and community, with reference to types of family organization. With family life challenged by economic change, the prospect for communities was a future of insecurity and social disorientation. Le Play saw the family as being crucially important in relating individuals to the wider community, thereby providing a moral foundation for and a stabilizing influence on social intercourse.

In 1887, Ferdinand Tönnies, in *Gemeinschaft und Gesellschaft* (see Harris, 2001; Loomis, 2002) produced an ambitious and complex study of social change and individual and group relationships (Harris, 2001: xv). The essential union in Gemeinschaft was the family with its sense of blood, kinship, friendship, and place. Associations of Gemeinschaft were guilds, churches, fellowships, and so on, where people shared common beliefs and ways of working. In all these associations, ideas of family and hierarchy remained strong. The combination of Gemeinschaft unions and associations thus reinforced a strong notion of community and mutual cooperation. With the mechanical society of Gesellschaft, people associated on the basis of ends, means, and rationality, which were foundations of the modern commercial enterprise with its emphasis on regulation. Association was no longer based on kinship or friendship but on rationality and calculation, indicating a move away from community. In Gemeinschaft, "individuals developed their identities within the wider, co-existing, whole," whereas in modern civil and commercial society "individual identity was ontologically prior to that of the wider group, attachment to which was merely secondary and instrumental" (Harris, 2001: xviii). In Gesellschaft, individuality and production for profit had priority (see Nisbet, 1966: 75–75).

There was a similarity between Tönnies and Max Weber concerning the 'falling away' from 'patriarchalism and brotherhood' (see Nisbet, 1966: 39). For Tönnies, Gesellschaft provided the explanatory process for the decline of community whereas Weber confirmed the importance of the communal ethic and regarded the falling away as the result of rationalization. Despite the influence of Tönnies on Weber, Nisbet (1966) argues that Weber presented a more subtle and sophisticated treatment of social action and community, focusing on types of solidarity in social relationships and group interests that engendered collective identity. Weber pointed to distinctions between the ideal types of communal and associative solidarities. Communal relationships involved a subjective sense of togetherness as found, for example, in families, neighborhoods, and religious groups, and they involved people in voluntary association based on common interests or beliefs. Associative relationships arose from rational calculation of interests as, for example, in the economic market. The closure of either type of relationship protected them, whereas open communal and associative relationships were subject to the erosion of coherence and allegiance.

In 1893, in *The Division of Labor in Society*, Emile Durkheim referred to mechanical and organic social solidarity. The basis of mechanical solidarity was moral and social cohesion expressed through small communities where the individual was subject to the collective conscience. The basis of organic solidarity was the division of labor in which individuals became disengaged from traditional class, kinship, and collective conscience. Durkheim later modified his position by

recognizing social facts that reinforced tradition and constraint. Durkheim "transmuted the attributes of mechanical solidarity into the eternal characteristics of social facts in general" (Nisbet, 1966: 86). An underlying message was that excessive individualism would lead to 'anomie' (without law) and the consequent breakdown of moral community.

Georg Simmel (2002[1903]) also problematized the situation of the individual in modern society. He regarded the individual in the context of the growing modern metropolis where the division of labor and economic market exchange produced alienation and social isolation. This made the individual vulnerable to the danger of subordination to the 'mass.' Simmel, therefore, underlined the value of voluntary associations and the danger to them posed by impersonal state institutions.

Early Community Studies

New thinking in the United States around the turn of the twentieth century signaled a growing concern of sociologists with issues having to do with the specific experiences of real communities. William Du Bois documented the plight of African-Americans in his 1889 work *The Philadelphia Negro*, which represented growing interest in the problems of urban African-Americans (Green and Driver, 1978).

After the First World War, sociologists in the United States studied community experiences against a backdrop of social and economic uncertainty. Various sociological and social anthropological methods and approaches were employed, and the genre of community studies developed as a distinctive approach in which researchers selected a single community as a cultural sample, commonly with the researchers living within the community while conducting their investigations and producing noncumulative findings.

The classic social anthropological community study by Lynd and Lynd (1929), *Middletown: A Study in Modern American Culture*, viewed life in a small American city (actually Muncie, Indiana). This was a predominantly white community having a 'middle of the road quality.' Adopting a functional cultural approach, they viewed employment, homemaking, leisure activities, religious practices, 'training the young,' and community involvement over a 35-year period, and in a follow-up study (Lynd and Lynd, 1937), they showed how Middletown had adapted to changing economic fortunes during and after the Great Depression. The Lynds found Middletown divided, with a dominant business class and a working class that had become less unionized and less politically active.

Also within the genre of community studies was the work of anthropologist Lloyd Warner and his colleagues, who produced their Yankee City studies (actually Newburyport, New England) between 1941 and 1959. They adopted a variety of research techniques, including observation, interviews, and documentary analysis, resulting in a view of Yankee City in terms of social-class divisions. Influenced by Warner, social anthropologists Allison Davis et al. (2009[1941]) viewed class and caste distinctions in their study of American communities in the Deep South (in the 'old city,' or Natchez, MI), and psychologist John Dollard (1937) in *Class and Caste in a Southern Town*, studied class and caste divisions, arguing that caste differences affected community social structures and impacted both black and white ethnic groups.

Another notable community study of this era was William Whyte's (1993[1943]) *Street Corner Society*, based on the results of participant research within the Italian North End slum district of Boston. Whyte provided a vivid picture of life in the ethnic enclave. The study presented a snapshot of a community with a strong sense of cohesion and internal organization.

Colin Bell and Howard Newby (1971) pointed to serious flaws in the community studies 'method.' They argued that ethnographic community studies often produced overly sympathetic accounts of community life, particularly when the researchers became closely involved with the neighborhoods studied, and that the noncumulative nature of the studies restricted an appreciation of wider societal implications. Nevertheless, the community studies carried out between the 1920s and 1940s provided some graphic insights into the life experiences of people in small-town and rural America and are still of interest socially and historically.

The Chicago School

Apart from such community studies, sociological interest in the 1920s and 1930s in the United States focused increasingly on the experiences of groups in cities, and interest in urbanism provided some of the most innovative and challenging sociological research of the time.

Chicago epitomized the twentieth-century American big city with its diverse cultures, impressive skyscraper buildings, and rapid urban expansion. Sociologists at the University of Chicago concentrated their research on this experience, in which migrants from Europe engaged in establishing new communities in teaming modern metropolises and where people sought betterment rather than revolution.

Although it is possible to bracket the Chicago School under the heading of community studies (Bell and Newby, 1971: 55), it is appropriate today to regard the Chicago School as having made a distinctive contribution that provided a more theoretically sophisticated body of work than community studies and established an 'urban sociology.'

Robert Park inspired the Chicago School, and Peter Saunders (1986: 53) discerned in Park's urban ecology the influences of Auguste Comte, Georg Simmel, Emile Durkheim, Charles Darwin, and Herbert Spencer. Saunders maintained that Durkheim influenced Park's ontological assumptions regarding human nature, the relationship between the individual and society, and the conditions for social order, wherease the Darwinian influence showed in Park's regard for cities in terms of urban ecology, where humans constantly adapted to and negotiated with their environments. In this, Park referred to community as 'an object of analysis,' an 'ecological process,' and a 'biotic level of social life' (Saunders, 1986: 53, 55). Here, there was ongoing competition for survival and a tension between the flights of human nature and the need for moral constraint.

Robert Park et al. (1925) portrayed urban neighborhood organization with respect to the idea that cooperation and

competition were foundations of intergroup relationships. Underlying tensions undermined social control, but communities achieved order through social specialization and interdependence. Park (1952) linked the inherent competitiveness of urban life to the struggle to control scarce resources, and other Chicago School sociologists went on to emphasize human struggle within cities and the consequent adaptation of social organization to change.

The urban ecologists interpreted the relationship between urban and rural in the same terms. Louis Wirth (1938: 7) argued that cities were expanding outward into rural areas through a process of urbanization that blurred the distinction between urban and rural. Increasingly, rural life bore "the imprint of urbanism in the measure that through contact and communication it comes under the influence of cities." Cities were relatively large, densely populated, and socially heterogeneous. Variations in size, density, and heterogeneity produced varied patterns of human association. Density was a function of differentiation, with groups tending to congregate in specific locations. For Wirth, competitiveness among urban dwellers and the superficiality of urban social relations resulted in a social psychology of urban life in which class and ethnic affinities became less important. For Wirth, cities were impersonal, with divisions between people producing different ways of life that made consensus and stable community life difficult to achieve. As Robert Sampson (2012: 427f) notes, Wirth linked community 'diversity with decline' by focusing on social disorganization. This contrasted with Whyte's socially organized slum and with the findings of several later studies that pointed to the sustained cohesion of rural and urban communities.

The postwar period presented new challenges to which the Chicago School urban ecology was incapable of responding, and Sampson (2012) has pointed to major methodological challenges in ecological research in the 1950s. Ray Phal (1968) argued that Wirth's description of inner cities did not hold up in the postwar context. The characteristics of modern inner-city communities were different from those described by Chicago School sociologists, and Herbert Gans (1968) criticized Wirth for having inappropriately generalized lessons from the inner city to wider urban areas and suburban communities. Wirth's size, density, and heterogeneity criteria did not explain broader urban change, and causal relationships among the criteria were not established.

Although the Chicago sociologists inspired in various ways those who later emphasized the social construction of 'the phenomena of daily life' (see Sarbin and Kitsuse, 1994: 3 for a wider discussion of the influence of Chicago in this respect), the ecological approach nevertheless lost its appeal. The Chicago influence faded as sociologists turned away from ecological explanations of spatial variation that failed to provide an adequate basis for theorization about social relationships and perceptions of community.

Tradition, Power, and Later Community Studies

American conservative sociologist Robert Nisbet (1953) proposed a functionalist view of community. He advocated a reconsideration of the traditions and collective values of communities. For Nisbet, the concept of community "encompasses all forms of relationship which are characterized by a high degree of personal intimacy, emotional depth, moral commitment, social cohesion, and continuity in time" (Nisbet, 1966: 47). Community engendered "a fusion of feeling and thought, of tradition and commitment, of membership and volition" given symbolic expression by 'locality, religion, nation, race, occupation, or crusade.' Nisbet argued that historically people were on a 'quest for community' that sought to rediscover the foundations of moral order and cohesion. However, in modern society, it was difficult to realize true community, so people turned to substitutes to reignite a sense of collective togetherness. Totalitarian political movements or the state often fulfilled the role of providing comradeship, national identity, and welfare provision.

Despite its nostalgic appeal, the quest for traditional community neglected the crucial issue of local power. In the 1950s and 1960s in the United States, controversy thus arose over the nature and distribution of community power and the implications for democracy. Sociologist Floyd Hunter (1980 [1969]), in *The Community Power Structure: A Study of Decision Makers*, pointed to the influence of powerful elite decision makers in Atlanta (which he called Regional City). Hunter referred to racial segregation and social privilege in the city, and sociologist C. Wright Mills (1956) identified a Power Elite presiding over politics and society in the United States. By contrast, political scientist Robert Dahl (1961) provided a classic pluralist interpretation of local politics in a study of New Haven, CT. There he claimed, polyarchy represented the rule of the many where power was unequally distributed, but where groups could influence crucial local political decisions.

Internationally, some sociologists and others continued to work within the community-studies genre, but community studies failed to come to terms adequately with some of the broader influences bringing about community change. The genre still drew on its functionalist foundations, and Graham Day (2006) critically reviewed postwar community studies with reference to, for example, Margaret Stacy's (1960) study of English small-town Banbury. Day (2006) highlighted the limitations imposed by "treating communities as though they were free-standing entities" (Day, 2006: 152). Given controversies about community power, urban deprivation, and ethnic minorities, researchers therefore came increasingly to view communities in terms of wider structures and processes by placing greater emphasis on social class, ethnicity, gender, and social conflict.

Class, Ethnicity, and Gender

British sociologists Michael Young and Peter Willmot (2007 [1957]), in *Family and Kinship in East London*, exemplified the broadening of the research horizon. The East End had traditionally served as a point of entry for immigrant groups, such the Huguenots and Jews. However, Young and Willmot warned of the disruptive effects of urban renewal on traditional working-class communities. They pointed to the bonds and kinship affiliations that made the East End work,

but public planning policies were disrupting communities by displacing people to new towns beyond London and through the construction of mass housing projects that took no account of the informal networks so important to community life.

In the United States, urban communities were also undergoing change. Herbert Gans (1982[1962]: ix), in his sociological study of Boston's West End, sought to 'generalize beyond the community study' by viewing social class and community change. Gans lived in the West End between 1957 and 1958, shortly before the area's comprehensive redevelopment. For Gans, the West End was understandable in terms of class rather than white ethnicity. Before clearance, the West End was a cohesive white working-class Italian community in the heart of Boston that was "an ethnic goldfish bowl in which privacy and nonconformity were rare" (Gans, 1982[1962]: xi). However, like Young and Willmot, Gans witnessed the disruption of the community through urban renewal – a process that he criticized following extensive interviews with West End residents affected by relocation to other areas.

The West End study added to concern about the condition of American cities and problems of urban communities. Community activists in the United States sought to empower communities by operating within the political system, and policymakers and politicians prioritized urban policies to address post-industrial urban decline by engaging local communities through participatory initiatives. In the United States in the 1960s, studies viewed the 'ghetto' and experiences of racism, sexism, social exclusion, and poverty set against the backdrop of urban and industrial decline (for example Jacobs, 1966; Hannerz, 1969). These issues became more pressing in the wake of serious urban riots in the United States in the 1960s, with community-based activist groups such as the Black Panthers advocating black power and direct action. Feminist groups also become more influential as evidenced by American sociologist Nancy Naples (2008) who have viewed feminist activism from the 1960s onward and the links among feminism, race, and class.

Sociologists John Rex and Robert Moore (1967) addressed themes of exclusion confronting ethnic minorities in England. They studied ethnic minority access to housing in the Sparkbrook district in the English city of Birmingham. Adopting an approach influenced by Weber and Burgess's work on Chicago, they showed how the interaction of social and spatial conditions influenced peoples' goals. Sparkbrook had experienced economic decline while attracting black and Asian immigrants from overseas. Some managed to acquire housing in better areas while others could obtain access only to poorer housing in rundown districts. Rex and Moore viewed the differentiation and spatial segregation of communities. The middle class tended to move to suburban areas while inner city areas concentrated the new migrants, where they experienced racial discrimination and bureaucratic city council housing policies.

Pahl (1975), adopting a neo-Weberian approach, regarded urban state bureaucracy in Britain as a mechanism restricting the access of local people to urban services, with planners and officials shaping cities with limited regard of the interests of local people. Planned urban housing projects lacked social and leisure facilities and failed to provide solutions to the problems of cities that were experiencing urban decline.

New Attitudes, Neo-Liberalism, and the Market

Changes in social attitudes and the achievement of black civil rights in the United States during the 1960s and 1970s helped minorities and other excluded groups to gain access to the political system and articulate community-level demands more effectively. However, disillusion among policy makers and the public with state interventionism brought a paradigm shift favoring neoliberal market reforms, private enterprise, and consumerism. The election of Conservative Prime Minister Margaret Thatcher in the United Kingdom and Republican President Ronald Reagan in the United States reflected the mood of the times.

Sociologist Terry Nichols Clark and political scientist Vincent Hoffmann-Martinot (1998) pointed to the emergence of a post-1960s New Political Culture (NPC) that accompanied major changes in social attitudes and aspirations that underlay the growth of market individualism and the lessening of collectivism. The NPC increased the salience of issues such as the environment, citizen democracy, gay rights, abortion, and lifestyle in urban and community politics. Politics placed less emphasis on class and workplace and greater stress on single issues championed by diverse groups within a complex and changing sociopolitical environment. The NPC challenged traditional political parties and old state hierarchies and undermined the assumptions of the postwar welfare state. People were more skeptical about large public bureaucracies and likely to support human-focused policies, smaller and more localized government units, and more responsive forms of local democracy. The NPC emerged "more fully and forcefully in cities and countries with more highly educated citizens, higher incomes, and high-tech service occupations" (Clark and Hoffmann-Martinot, 1998: 3). This marked a change in the ways local communities articulated social and political demands. Local groups were more likely to organize across social class lines as they adopted new modes of social and political action. Information technology opened new horizons by providing opportunities for more effective networking and communication.

Despite the appeal of the market, Mark Gottdiener (1994) identified a radicalizing 'new urban sociology' that emerged during the 1970s, with its interest in social and class conflict in cities and communities. This included work by Marxists such as Manuel Castells and others who sought to develop critiques of neoliberalism and modern market capitalism through studies that broadened the critical research agenda, thereby regarding community within the wider contexts of social change (see below). For example, Castells (1977, 1996) regarded the political economy of urban and community change with regard to modes of capitalist production and the role of communities in the economic processes of accumulation. The neoliberal project, for Marxists and other radical researchers, marked a period of capitalism that consolidated social and economic inequalities and led to the emergence of new forms of local activism and community resistance.

Regime theorists focused on new configurations of economic and political power affecting cities and communities by viewing the impacts of the neoliberal market and the interests of different groups contending for scarce resources. Groups could influence the policy agenda in new ways in cities where conditions were conducive to cooperation. Urban regimes brought together public, private, and political interests that otherwise might conflict so as to boost city economies and market competitiveness. In his urban regime study of Atlanta, political scientist Clarence Stone (1989) assessed changes in the city since Floyd Hunter's (1980[1969]) classic study. Stone showed how a biracial coalition had become part of the city's urban regime. Coalition building helped black politicians access the political system from which they had for so long been excluded and in a city where communities had been racially segregated. Stone's Atlanta story was indicative of change in other cities where African-Americans and other minority groups began to exert political and cultural influence.

The persistence of social exclusion and inequalities demonstrated that market-oriented policies did not automatically guarantee social benefits. Charles Murray (1990) provided a popularized conservative view of the consequences in deprived urban communities of a welfare-dependent underclass the members of which lacked ambition and sense of being responsible for their own futures. Like Murray, sociologist and communitarian Amitai Etzioni (1995, 1996) called for greater moral responsibility and advocated a return to moral values and a greater sense of community and enterprise, while others pointed to social networks as being important in patterning relationships among members of local communities. The idea that cooperation and mutuality produced social benefits featured widely in discussions about viable communities. Public-policy specialist Robert Putnam (2000), in a popular study, found that 'social capital' had declined in communities in the United States. Communities once produced high levels of social capital, but they now had low levels of membership in local organizations and clubs and lower levels of local social support. Earlier, sociologists Harrison White (1963) and Claude Fischer (1982) had developed social-interaction models showing how networking influenced local organization, and sociologist John Scott (2004[1991]) showed how social networks of people influenced local relationships among actors in communities.

The Cultural Turn and Socially Constructed Communities

Sociologists, social theorists, cultural theorists, and those involved in the more broadly defined multidisciplinary field of cultural studies variously resituated socially constructed communities within wider theoretical projects related to social and cultural change. Social theorists came to view communities in the context of broader processes of change by challenging modernist structuralist assumptions in urban sociology. The 'cultural turn' brought about a radical change in the way in which many social researchers understood community in terms of postmodernity.

Geographer Edward Soja (2000, 2010) argued that the cultural turn created a new agenda for researchers – a move away from categorizations of community and social class toward concern with the "more cross-cutting and inclusive foundations of solidarity, collective consciousness, and coalition building" (Soja, 2000: 279). For Soja, this represented a shift away from concern with equality (race, gender, and class inequalities as evidenced in post-1960s urban sociology) toward new conceptions of justice relevant to groups forming new kinds of local-level coalitions.

Sociologists Scott Lash and John Urry (1987) integrated post-Fordism in their postmodernist work *The End of Organized Capitalism*. Henry Ford's modernist mass-production factories had given way to high-technology production located globally, including in countries such as China, South Korea, and India. The post-Fordist regime of accumulation replaced centralized mass production and mass consumption and ushered decentralized flexible specialization, customization, and the growth of service- and information-based activities. Lash (1990: 33) maintained, however, that postmodernism could actually produce 'a strong sense of locality' that, he argued, resembled the Greenwich Village-type of community spirit and diversity depicted by Jane Jacobs (1961). New social movements frequently had community-focused aspirations, along with "a heightened sense of the mix of ethnicity and diverse cultural groupings" that "foregrounds also a sort of metaphysics of 'the street' " (Lash, 1990: 33).

Lash argued that postmodernism could "foster radical-democratic and decentralized worker resistance" offering a vision favoring "community, street, and 'the labyrinth'" (Lash, 1990: 37). However, Lash (1990: 37) also referred to the 'fragmentation of opposition into decentralized social movements' and the resurgence of competitive market individualism at the expense of labor unions and traditional social organizations. In this context, governments attempted to foster market-oriented governance while 'oppositional postmodernism' expressed the desire for community. There was, then, a process of organization and disorganization whereby formally organized interests (such as trade unions and working-class organizations) were 'less coherently structured' and confronted an ideological crisis. This led to "the literal social disorganization of urban blight and violent rebellion by minority and marginal groups whether in Detroit, Liverpool's Toxteth, or Berlin's Kreuzberg" (Lash, 1990: 38).

Others, who we might regard as working within cultural studies (as distinct from specifically postmodernist cultural theory), provided insights into meanings and identities in society and local communities. For example, British social anthropologist Anthony Cohen (1985) portrayed community as a symbolic construct. For Cohen, although the physical geographical bases of community had diminished in importance, there had been a renewed assertion by communities of symbolic boundaries that demarcated one community from another. These boundaries came in many forms, such as rituals and customs that reinforced community identity, membership, and allegiance. The community could "make virtually anything grist to the symbolic mill of cultural distance, whether it be the effects

upon it of some centrally formulated government policy, or a matter of dialect, dress, drinking or dying." This reflected "the contrivance of distinctive meanings within the community's social discourse" (Cohen, 1985: 117). Cohen saw this as indicating the strength of communities because people could achieve collective identify based on strong ethnic, religious, and social symbolism. Benedict Anderson (1991) similarly detected strong identities in communities and culturally expressed through nations as imagined communities.

The Spatial Turn

Theorists concerned with social change variously came to recognize the need to address socially constructed spatiality in contextualizing urban and community change. Indeed, researchers from various disciplines contributed to a 'spatial turn' (Warf and Arias, 2009) that heightened the significance of socially constructed space in accounts of urban and community life and the emergence of new social movements.

Soja (2009: 24) advocated integrating spatial analysis with cultural theory. Soja refers to the 'defensive disciplinary response' to the spatial turn, notably in geography, but also in other disciplines that "often involved a vigorous rejection of the postmodernist critiques that were so closely associated with the spatial turn." Nevertheless, theorists across various disciplines witnessed a significant 'reconfigurative diffusion of spatial thinking' in the social sciences and humanities (Soja, 2009: 25).

Castells (1977), in *The Urban Question*, identified cityspace and activist 'communities of resistance.' For Castells, people in communities could bring about change if they organized themselves effectively. The conflicts among the state, capital, and new social mobilizations involved spatial contestations over cityspace. Capital and the state attempted to create cityspace in their own image, with state authorities and corporations parceling land and redefining the boundaries of neighborhoods. This constant process of reorganization frequently encountered locally organized opposition.

Soja's (2000) *Postmetropolis* depicted a fragmented postmodern urban-regional landscape in Los Angeles that produced 'unjust geographies' (Soja, 2011). Soja referred to the 'jumbled spatiality of the globalized post-Fordist exopolis' characterized by the restructured and fragmented city with its complex 'social mosaic' (Soja, 2000: 265). For Soja, cities juxtaposed meanings and experiences and generated multiple ambiguities. The post-Fordist exopolis was 'sociality' arranged "in complex patterns that are only beginning to be recognized, understood, and effectively studied" (Soja, 2000: 265). Although 'significant communities persist and cannot be ignored,' the new urban social order was no longer defined 'effectively by such conventional and familiar models of social stratification' as the class divided city; the 'neatly layered' hierarchical city or the racially divided city (Soja, 2000: 265). While such 'older polarities' still existed, the most developed cities in terms of their postmodern characteristics were fractal and productive of new social geometries.

Postmodernists frequently referred to the influence of Michel Foucault (1970[1966]), who conceived of heterotopias that in different ways immersed people in contexts that combined different spatial elements, changed over time, supported particular ways of behaving, and gave rise to different constructions of meaning. Heterotopias, such as gated communities, atria-enclosed shopping malls, private lifestyle and leisure spaces, and commercial enclaves generated new identities and ways of thinking about the self in different cultural contexts. Researchers alluded to the privatization of urban space and the social implications of segregated communities and inequalities. For example, Anna Minton (2009), in *Ground Control*, explored the privatization of urban space, social control, and surveillance. Mike Davis (1992) referred to the consequences of segregation and exclusion in communities in Los Angeles.

Marxist geographer David Harvey portrayed people and communities involved in defensive struggles against the destabilizing economic forces of crisis-ridden capitalism. Harvey (2011) has argued that capitalism spatially creates room for accumulation. Global capital is more assertive in creating new spaces of accumulation and consequently creates new sites of struggle and the potential of 'rebel cities' where activists mobilize around a variety of locally salient issues (Harvey, 2012).

Globalization

Sociologists, cultural theorists, and others commonly contextualized change with reference to globalization and the geographic and social mobility of people. The rise of mega-cities with populations of 20 million or more raised concerns about social cohesion and the implications of global shifts of migration, the creation of informal settlements, and poverty and social exclusion.

Castells, acknowledging Howard Rheingold (1993), argued that new information technologies integrated the world through global networks and a 'vast array of virtual communities' (Castells, 1996: 22) that networked groups and individuals in the 'network society.' The Internet and 'electronics grass-roots culture' (Castells, 1996: 354) enabled individuals to interact globally, but also privately. For Castells, the new globalized and customized mass media permeated all levels of society, directly affecting the local community and the individual while Roland Robertson (1995) argued that globalization reconstructed and produced new cultural conceptions of home, community, and locality. Glocalization combined a global and local perspective where local individuals, communities, and companies were oriented to the global.

Sociologist Saskia Sassen (2001) studied the global city, in which dramatic changes affected local communities. Migration and concentrations of global capital led to global cities' becoming vibrant centers of culture and economic activity. In developing nations, global cities acted as network hubs bringing together financial institutions, corporations, and people in a globalizing context. These attracted people from rural to urban areas. However, some cities missed out, and some poorer mega-cities and urban centers failed to achieve global status, thereby marginalizing local communities within the globalizing economy.

Michael Young, with colleagues Geoff Dench and Kate Gavron (Dench et al., 2006), in a follow-up to the original Young and Willmott (2007,[1957]) research already mentioned, regarded a globalized London by viewing the consequences of change for the 'historic working class' and the growing Bangladeshi communities in London's East End. The East End communities are seen in terms of ethnic identity and cross-cutting class cleavages with "cycles of conflict, competition and eventual accommodation between incoming groups and local residents" (Dench et al., 2006: 15). The research revealed greater social and geographical mobility and a marked 'loss of localism,' with many older white working-class people having a sense of having been 'abandoned by the flow of history' (Dench et al., 2006: 31).

Tim Cresswell (2006: 1) found that culture "no longer sits in places, but is hybrid, dynamic – more about routes than roots." Sociologists Anthony Elliot and John Urry provided an account of the global mobility of people and the impacts of mobility on the ways people lived and how they fashioned multiple identities and assumed 'portable personhood' (Elliot and Urry, 2010: ix). Elliot and Urry referred to the potential constraints on mobility in a post-modern and 'post-carbon' future.

The recent appeal of civil society has grown as a response to such processes of globalization and market change and the growth of state power and control. Civil society enables groups and individuals to organize themselves automatously – usually as free from state or corporate control as possible. According to Mary Kaldor (2003), civil-society groups can express group resistance to the impersonal globalizing forces of the market and state and potentially can build new communal identities. Kaldor et al. (2012) provide a comprehensive overview and compilation of the literature concerning civil society that demonstrates the interest of researchers across academic disciplines in the growth of civil activism in a globalizing world. Kaldor et al. (2012: 2) comment that not only in the West, "but also in growing parts of Asia and Latin America, Africa and the Middle East, today's generation are the children of the Internet, the mobile phone and cheap air travel – the 'globalization generation.' They know that the world is a singular fragile eco-system – and that while the national state does have a role to play, it is part of a broader global community."

Coda

The idea of the global community adds to the range of social conceptions of community, from the small-town community studies of the 1930s to the ghettoized communities of the 1960s, the postmodern fractile and mobile communities of globalization, and the global community itself. Sociologist Gerald Delanty (2003: 191–192) refers to the continued interest in community in the fields of sociology and other disciplines but notes that this reflects a 'radical pluralization' in which the 'forms of community are multiple' and the 'imaginary dimension of community' underlines 'the impossibility of community' as some universally agreed concept. Similarly, sociologist Zygmunt Bauman (2001: 3) has suggested that community "stands for the kind of world which is not, regrettably, available to us – but which we would dearly wish to inhabit and which we hope to repossess." This seems to confirm that community will continue as an enduring but ever-contested theme in sociology and in social and cultural theory.

See also: Capitalism: Global; Collective Beliefs, Sociological Explanation of; Communitarianism, Sociology of; Ethnic, Racial, and Nationalist Movements; Inner City; Minorities; Mobility, Social; Modernity; Modernization, Sociological Theories of; Postmodernism in Sociology; Race Identity; Rural Sociology; Social Identity in Sociology; Social Networks; Solidarity, Sociology of; Underclass; Urban Sociology.

Bibliography

Anderson, B., 1991. Imagined Communities: Reflections on the Origins and Spread of Nationalism. Verso, London.
Bauman, Z., 2001. Community: Seeking Safety in an Insecure World. Polity Press, Malden, MA.
Bell, C., Newby, H., 1971. Community Studies: An Introduction to the Sociology of the Local Community. George Allen and Unwin, London.
Castells, M., 1977. The Urban Question. Edward Arnold, London.
Castells, M., 1996. The Rise of the Network Society. Blackwells, Oxford, UK.
Clark, T.N., Hoffmann-Martinot, V., Gromala, M. (Eds.), 1998. The New Political Culture. Westview Press, Boulder, CO.
Cohen, A., 1985. The Symbolic Construction of Community. Routledge, New York.
Cresswell, T., 2006. On the Move: Mobility in the Modern Western World. Routledge, New York.
Davis, A., Gardner, B.B., Gardner, M.R., 2009[1941]. Deep South: A Social Anthropological Study of Caste and Class, with a New Introduction by Wallach, J.J. University of South Carolina Press, Columbia.
Davis, M., 1992. City of Quartz: Excavating the Future in Los Angeles. Vintage, New York.
Day, G., 2006. Community and Everyday Life. Routledge, New York.
Dahl, R.A., 1961. Who Governs? Democracy and Power in an American City. Yale University Press, New Haven.
Delanty, G., 2003. Community. Routledge, New York.
Dench, G., Gavron, K., Young, M., 2006. The New East End: Kinship, Race and Conflict. Profile Books, London.
Dollard, J., 1937. Caste and Class in a Southern Town. Doubleday, New York.
Elliot, A., Urry, J., 2010. Mobile Lives. Routledge, New York.
Etzioni, A., 1995. The Spirit of Community: The Reinvention of American Society. Simon and Schuster, New York.
Etzioni, A., 1996. The New Golden Rule: Community and Morality in a Democratic Society. Basic Books, New York.
Fischer, C.S., 1982. To Dwell among Friends: Personal Networks in Town and City. Chicago University Press, Chicago.
Foucault, M., 1970[1966]. The Order of Things. Tavistock, Andover.
Gans, H.J., 1982[1962]. The Urban Villagers: Group and Class in the Life of Italian Americans. The Free Press, New York.
Gottdiener, M., 1994. The New Urban Sociology. McGraw-Hill, New York.
Green, D.S., Driver, E.D. (Eds.), 1978. W.E.B. Du Bois on Sociology and the Black Community. Chicago University Press, Chicago.
Hannerz, U., 1969. Soulside: Inquiries into Ghetto Culture and Community. Columbia University Press, New York.
Harris, J. (Ed.), 2001. Tönnies: Community and Civil Society. Cambridge University Press, Cambridge.
Harvey, D., 2011. In: The Enigma of Capital and the Crises of Capitalism. Profile Books, Exmouth.
Harvey, D., 2012. Rebel Cities: From the Right to the City to the Urban Revolution. Verso, London.
Hunter, F., 1980[1969]. Community Power Structure: A Study of Decision Makers. The University of North Carolina Press, Chapel Hill.
Jacobs, J., 1961. The Death and Life of Great American Cities. Random House, New York.
Jacobs, P., 1966. Prelude to Riot. Random House, New York.

Kaldor, M., 2003. Global Civil Society: An Answer to War. Polity Press, Cambridge.
Kaldor, M., Moore, H.L., Selchow, S. (Eds.), 2012. Global Civil Society 2012: Ten Years of Critical Reflection. Palgrave Macmillan, Houndmills.
Lash, S., 1990. Sociology of Postmodernism. Routledge, London.
Lash, S., Urry, J., 1987. The End of Organized Capitalism. Polity Press, Cambridge.
Loomis, C.P. (Ed.), 2002. Community and Society: Ferdinand Tönnies. Dover, New York.
Lynd, R.S., Lynd, H.M., 1929. Middletown: A Study in Modern American Culture. Harcourt Brace, New York.
Lynd, R.S., Lynd, H.M., 1937. Middletown in Transition: A Study in Cultural Conflicts.
Mills, C.W., 1956. The Power Elite. Oxford University Press, New York.
Murray, C., 1990. The Emerging British Underclass. Institute of Economic Affairs, London.
Naples, N.A. (Ed.), 2008. Community Activism and Feminist Politics: Organizing across Race, Class, and Gender. Routledge, New York.
Nisbet, R.A., 1953. The Quest for Community. Oxford University Press, Oxford.
Nisbet, R.A., 1966. The Sociological Tradition. Basic Books, New York.
Phal, R.E. (Ed.), 1968. Readings in Urban Sociology. Pergamon Press, Oxford.
Pahl, R.E., 1975. Whose City?, second ed. Penguin, Harmondsworth, UK.
Park, R.E., Burgess, E.W., McKenzie, R.D., 1925. The City. University of Chicago Press, Chicago.
Park, R.E., 1952. Human Communities. Free Press, Glencoe, IL.
Putnam, R.D., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon and Schuster, New York.
Rex, J., Moore, R., 1967. Race, Community and Conflict: A Study of Sparkbrook. Oxford University Press for Institute of Race Relations, London.
Rheingold, H., 1993. The Virtual Community. Addison-Wesley, Reading, MA.
Robertson, R., 1995. Glocalization: Time-Space and Homogeneity-Heterogeneity. In: Featherstone, M., Lash, S., Robertson, R. (Eds.), Global Modernities. SAGE, Thousand Oaks, pp. 25–44.
Sampson, R.J., 2012. Great American City: Chicago and the Enduring Neighborhood Effect. The University Press of Chicago, Chicago.
Sarbin, T.R., Kitsuse, J.I. (Eds.), 1994. Constructing the Social. SAGE, Thousand Oaks, CA.
Sassen, S., 2001. The Global City: New York, London, Tokyo, second ed. Princeton University Press, NJ.
Saunders, P., 1986. Social Theory and the Urban Question, second ed. Hutchinson, London.
Scott, J., 2004[1991]. Social Network Analysis: A Handbook. SAGE, Thousand Oaks, CA.
Simmel, G., 2002[1903]. The Metropolis and Mental Life. In: Bridge, G., Watson, S. (Eds.), The Blackwell City Reader. Blackwell Publishers, Oxford, pp. 11–19.
Soja, E.W., 2000. Postmetropolis: Critical Studies of Cities and Regions. Blackwell Publishers, Malden, MA.
Soja, E.W., 2010. Seeking Spatial Justice. University of Minnesota Press, Minneapolis, MN.
Soja, E.W., 2009. Taking Space Personally. In: Warf, B., Arias, S. (Eds.), The Spatial Turn: Interdisciplinary Perspectives. Routledge, New York, pp. 11–35.
Soja, E.W., 1989. Postmodern Geographies: The Reassertion of Space in Critical Social Theory. Verso, New York.
Stone, C., 1989. Regime Politics: Governing Atlanta 1946–1988. University Press of Kansas, Lawrence, KS.
Stacy, M., 1960. Tradition and Change: A Study of Banbury. Oxford University Press, Oxford.
Warf, B., Arias, S. (Eds.), 2009. The Spatial Turn: Interdisciplinary Perspectives. Routledge, New York.
White, H.C., 1963. An Anatomy of Kinship. Prentice-Hall, Englewood Cliffs, NJ.
Whyte, W.F., 1993[1943]. Street Corner Society: Social Structure of an Italian Slum. University of Chicago Press, Chicago.
Wirth, L., 1938. Urbanism as a way of life. American Journal of Sociology 44, 1–24.
Young, M., Willmott, P., 2007[1957]. Family and Kinship in East London. Penguin, London.

Community Studies: Anthropological

Dan Rabinowitz, Tel Aviv University, Tel Aviv, Israel; and Central European University, Budapest, Hungary

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Rabinowitz, volume 4, pp. 2387–2389, © 2001, Elsevier Ltd.

Abstract

The term 'community' has been used by anthropology in a variety of ways and contexts. On a concrete ethnographic level, 'community' and 'community studies' have two main denotations. One is generic, echoing the interest anthropology and anthropologists have always had in ordinary, stable, small-scale localized collectives. The other pertains to studies of rural populations (e.g., villages, parishes, counties) and urban enclaves (e.g., neighborhoods and quarters) within industrialized, developed Western countries.

Analytically, the term tends to denote a syndrome rather than an accurately defined phenomenon. Used in various periods and within different theoretical orientations in reference to territorially based units, professional circles, solidarity groups and informal sociocultural amalgams, it could imply permanent collections with well-defined boundaries as well as loosely defined and transient formations.

The efficacy of the term 'community' as an interpretative tool was rather limited prior to the 1970s. New anthropological and ethnographic vistas since, however, have imbued it with more theoretical currency. It became, for example, a significant element in new anthropological tool-boxes developed for the analysis of rural groups in the northern, southern, and western margins of Europe. Later it found new uses as anthropologists began grappling with the need to conceptualize and historicize transnational social formations in a globalizing world.

The Concept of Community in the Social Sciences

Community, a rather ubiquitous and imprecise term in terms of social science theory, tends to be associated with an array of positive connotations, including solidarity, familiarity, common purpose, unity of values and a shared sense of identity. Earlier sociological debates of communities and their social significance stressed either a territorial base, i.e., groups whose members share a well-defined location (see Hawley, 1950), or, later, community as state of mind – a collective sharing of an underlying feeling of sameness and belonging regardless of geographical dispersion (see Bell and Newby, 1974). The latter option, while analytically promising and intellectually suggestive, produced a category which, while exceedingly popular, was also perplexing in its ambiguity. Since affiliation and identification with non-territorial communities can be unclear even to the actors themselves – let alone to social scientists attempting to characterize and study them – the concept can be (and has often been) used as an omnibus for a variety of related and unrelated meanings.

Not surprisingly, some sociological attention was given to the distinctions between the informal nature of communities and the more rigid structure of organizations (Gottshalk, 1975; Azarya, 1984) as well as to the problem of indeterminacy of community boundaries (Hillery, 1955). These not withstanding, community studies emerged since the 1920s and well into the 1980s as a solid branch of sociological inquiry, not least through the consistent preoccupation of the Chicago school with modernization and its impacts on urban culture, subcultures, and urban society.

Community (Re)Enters Anthropology

Anthropology, with its early emphasis on intensive field study focusing on small-scale groups whereby culture and society were assumed to be unproblematically linked to bounded space, had an inherent affinity with the view of communities as discernible sociocultural units. Habitually working with bounded communities, anthropologists tended to develop context-specific tools for their analysis. Early emphasis on customs and manners, the subsequent emergence of the concept of culture, the paradigms of structural functionalism, transactionalism, symbolic interaction, and structuralism *à la* Lévi-Strauss all offered, in their respective heydays, rich and convincing frameworks for studies of small-scale bounded groups. This may be the reason why the concept of community was not explored by anthropologists with any rigor, and why it was receiving limited critical attention in the discipline prior to the 1980s.

Thus, while the term community was always present in the anthropological lexicon, its syntactic status tended to be relegated to noun or synonym. It remained a finite idiom with limited interpretative power, its capacity to become a dynamic theoretical tenet largely arrested. Committed to depicting 'native' ways of explaining solidarity and social relations, and taken by the concept of an immutable culture occupying an ethnographic present, anthropologists had little use for the term *community*. And while Lynd and Lynd's classic study of Muncie (eternalized as *Middletown* in Lynd and Lynd, 1929), Lloyd Warner's *Yankee City* volumes (Warner and Lunt, 1941; Warner and Low, 1947) and Arensberg's work on the relationship between community and culture (see Arensberg and Kimball, 1965) undoubtedly inspired ethnographically oriented studies such as Robert Redfield's (1955) *The Little Community*, Marriott's edited volume on villages in India (Marriott, 1955) or Frankenberg's (1965) *Communities in Britain*, a strong and clear anthropological tradition of community studies was slow to materialize.

Two factors may have contributed to the emerging interest of anthropologists in the 1980s in the community as an analytical and theoretical category. One was the aforementioned

emergence in the 1960s and 1970s within sociology of the view of the community as a more abstract entity, whereby meaning and a sense of solidarity became more salient than tangible physical contact. The other was a growing need amongst anthropologists, by then more willing than ever before to study societies and cultures in the core and on the margins of Europe, to develop a theoretical toolbox for this new challenge.

An edited volume and an ethnography by Anthony Cohen (1982, 1985) indicated the efficacy of the concept of community for a more sophisticated understanding of the relationship between local rural and urban communities and wider affiliations and identity, not least the ones connected to the nation-state. The edited volume (Cohen, 1982) had six anthropologists writing two essays each about rural communities in Britain. One essay was incorporated into a section of the book entitled 'Belonging to the Part: Social Association within the Locality,' while the other was included in a section called 'Belonging to the Whole: The Community in Context.' The sections corresponded to the dual themes of the inquiry: how people develop a sense of belonging in a small community; and how they develop an identity and a set of practices that reflect their sense of place in the broader realm that lies beyond the local. The first theme naturally focuses on kinship, sect, regional histories, and affiliations. The second theme, pursued in Cohen's subsequent work, leads to a growing preoccupation with cultural boundaries and borders.

In terms of theory and analysis, the ethnographic attention paid by Cohen and his contributors to (mainly Celtic) rural Britain owes much to Clifford Geertz's emphasis on meaning and cultural interpretation. The community, according to Cohen's introduction to the volume, must be studied and interpreted not so much as a structured social organization but as an arena where experience is ordered, partly through a 'local' culture. This in turn produces a distinct local identity which is engaged in constant interplay with broader cultural formations, processes, and signals. Cohen's volume thus signposted a new field of inquiry, the form and theoretical approach of which were different from those of the adjoining anthropological arena of peasant studies, the latter focusing mainly on rural communities in semi-peripheral areas, including the margins of Europe and Latin America.

The importance attributed by Cohen to meaning and imagination, and the Wittgensteinian supremacy he advocates of context of use over lexical occurrence, become even clearer in *The Symbolic Construction of Community* (Cohen, 1985). Simple people too, he shows, are preoccupied with symbols, constantly using them to make comparisons (and sense) of appearance and reality, likeness and diversity, similarity and difference. Once more, social structure collapses at the feet of culture, subverting the monopoly of sociology over the concept of community. Significantly, more emphasis than ever before was now directed at the borders. Writers came to recognize that it is near boundaries and across them, that community culture and identity were being determined.

Community studies reduce methodological, analytical, and theoretical differences between anthropology and sociology. Conrad Arensberg's works of the 1950s and 1960s are by no means wholly sociological, and the same is true for Maurice Stein's *The Eclipse of Community* (Stein, 1960). By the same token, the strongly ethnographic turn of Warner's *Yankee City* volumes, West's *Plainville* (West, 1961) or the Lynds' *Middletown* (Lynd and Lynd, 1929) probably warrant counting them as anthropological, while the preoccupation displayed by Cohen (1982, 1985), Frankenberg (1965) and Hannerz (1980) with broad sociological issues merit labeling their works at least partly sociological.

Some writers have alluded to an obvious shortcoming of anthropological (and, for that matter, sociological) studies of communities – their propensity to be apolitical. Rich in analyses of inherent traits and characteristics, such studies lend themselves willingly to elaborate treatments of culture, cultural identity, and demarcation. This, however, can often serve as a cover for shying away from and even blindfolding themselves to issues of internal and external power disparities, political economy, and historicity (cf Aronoff, 1984; Cox, 1987).

A strength that anthropological studies of communities have had, at least since the 1980s, has been their emphasis on abstract experiences, not least imagination, as idioms shaping human identity and agency. Major inspiration for this project undoubtedly came from Benedict Anderson's work, not least from his creative employment of this notion in his seminal work on nationalism, *Imagined Communities* (Anderson, 1983). The focus on imagination resurfaced later in attempts to explore new ways in which anthropology could tackle the paradoxes and complexities of transnational experience and culture.

Critical Anthropology and the Concept of Community

Like 'ethnography,' 'the ethnographic present,' 'fieldwork,' 'informants,' and other key anthropological idioms, 'community' became, in the 1990s, a construct. Idealized as a composite, unproblematized pillar of old-guard anthropology, it could now serve as a hammering block against which critical ideas could be articulated.

Malkki (1997) looks at the connection between the concept of community and anthropology's propensity to look for the ordinary, durable, 'everyday' routines and practices. This approach, which gained for anthropology justified rewards and insights, gradually bred an expectation that all people studied by ethnographers are not just haphazard collections of humans somehow thrown together into the same location. Rather, they were expected to have always been part of a stable, permanent and – in particular – localized 'community.' But while such groups do of course exist, the emphasis on communities pushed anthropologists to overlook other possibilities, including those where life does create transient collections of humans randomly united. Such, Malkki (1997: p. 91) indicates, are what Barbara Myerhoff called *accidental communities* – Woodstock being a dramatic instance – and what Malkki herself found in refugee camps in Rwanda, where displaced people from a variety of places of origin huddled together, reinventing a sense of communal rootedness they may or may not have had in the lives they had before their dislocation.

Arjun Appadurai (1996), in his seminal work on the relationship between bounded place and culture in a world that is rapidly becoming dislocated, transnational, and diasporic, likewise uses an idealized icon of the term community in anthropology to develop a more sophisticated

concept of culture. Localities and communities, he asserts, must be historicized and contextualized, identified as processes of growth and shrinkage. Not least, anthropology should – and can – become more sensitive to what he calls the global production of locality.

Traditional concepts of community in the social sciences, as they had been applied to studies of rural Mexico, give Roger Rouse's (1991) study of migration into the USA a much needed structure against which to prop his concept of transnational migration. Earlier migration studies, he aptly indicates, had a bipolar picture, with culturally and demographically stable and homogeneous collectives on both ends. Immigrants, within this framework of analysis, simply moved – more or less successfully – between two 'communities.' His own ethnography, however, convinced him that what was taking place in fact was a more complex phenomenon, a translocal movement backward and forward. This productively subverts the wholesome view of the community – be it in rural Mexico or in the migrants' quarters of western USA – offering new tools for understanding of these complex and uneasy situations.

The new (twenty-first) century is yet to see a revival of the theoretical efficacy of the term 'community' in anthropology a renewal of its definitional powers as an ethnographic idiom. If anything, the vagueness of it, the silo capacity it has to denote an intuitive sense rather than index a specific phenomenon, is sometimes exploited for book or article titles (see for example Monterescu and Rabinowitz (2007)) which, while having an appeal for wider audiences, do not necessarily push the analytical prowess of the term to new terrains. Likewise, the choice of some departments in academic institutions, mainly in community college in the US, to change their names from 'Anthropology' to 'Community Studies' indicates a further resignation to the fact that rather than indicating theoretical precision, the term becomes suggestive of a fuzzier sense of solidarity and social cohesiveness which, for some, can represent what anthropology should stand for.

See also: Community Sociology; Community, Social Context of: The US Case; Practice: An Anthropological Approach.

Bibliography

Anderson, B., 1983. Imagined Communities. Verso, London.

Appadurai, A., 1996. Modernity at Large: Cultural Dimensions of Globalization. University of Minnesota Press, Minneapolis, MN.

Arensberg, C., Kimball, S., 1965. Culture and Community. Harcourt Brace, New York.

Aronoff, M., 1984. Review of Cohen 1982 belonging: identity and social organization in British rural cultures. Contemporary Sociology 13 (1), 46–47.

Azarya, V., 1984. The Armenian Quarter of Jerusalem. University of California Press, Berkeley, CA.

Bell, C., Newby, H. (Eds.), 1974. The Sociology of Community: A Selection of Reading. Frank Cass, London.

Cohen, A., 1982. Belonging: Identity and Social Organization in British Rural Cultures. Manchester University Press, Manchester, UK.

Cohen, A., 1985. The Symbolic Construction of Community. Methuen (Tavistock), New York.

Cox, G., 1987. Review of Cohen 1985 the symbolic construction of community. Sociologia Ruralis 27 (1), 76–77.

Frankenberg, R., 1965. Communities in Britain: Social Life in Town and Country. Penguin Books, London.

Gottshalk, S., 1975. Communities and Alternatives: an Exploration of the Limits of Planning. Shenkman, Cambridge, MA.

Hannerz, U., 1980. Exploring the City: Inquiries Toward an Urban Anthropology. Columbia University Press, New York.

Hawley, A., 1950. Human Ecology: A Theory of Community Structure. Roland Press, New York.

Hillery, G.A., 1955. Definitions of community: areas of agreement. Rural Sociology 20 (2), 111–123.

Lynd, R.S., Lynd, H.M., 1929. Middletown: A Study in American Culture. Harcourt Brace, New York.

Malkki, L., 1997. News and culture: transitory phenomena and the fieldwork tradition. In: Gupta, A., Ferguson, J. (Eds.), Anthropological Locations. University of California Press, Berkeley, CA, pp. 86–101.

Marriott, M. (Ed.), 1955. Village India: Studies in the Little Community. University of Chicago Press, Chicago.

Monterescu, D., Rabinowitz, D. (Eds.), 2007. Mixed Towns, Trapped Communities: Historical Narratives, Spatial Dynamics, Gender Relations and Cultural Encounters in Palestinian–Israeli Towns. Ashgate, London.

Redfield, R., 1955. The Little Community: Viewpoints for a Study of a Human Whole. University of Chicago Press, Chicago.

Rouse, R., 1991. Mexican migration and the social space of postmodernism. Diaspora 1 (1), 8–23.

Stein, M., 1960. The Eclipse of Community: An Interpretation of American Studies. Princeton University Press, Princeton, NJ.

Warner, L., Low, J.O., 1947. Social System of the Modern Factory. Yale University Press, New Haven, CT.

Warner, L., Lunt, P., 1941. The Social Life of a Modern Community. Yale University Press, New Haven, CT.

West, J., 1961. Plainville, USA. Columbia University Press, New York.

Community Supervision and Diversion in the United States

Lincoln B Sloas, Jennifer Lerch, and Faye S Taxman, Center for Advancing Correctional Excellence (ACE!), George Mason University, Fairfax, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

With over 5 million adults on some form of community supervision, this article examines the current practices related to community supervision and diversion in the United States. Specifically, four types of diversion are outlined, including (1) complete diversion; (2) justice leverage diversion; (3) probation diversion; and (4) enhanced diversion. The empirical evidence surrounding each type of leverage is provided. A research agenda for the future is provided.

The vast majority of the offender population in the United States is not housed in correctional facilities but rather supervised in community settings (Petersilia, 2011; Maruschak and Parks, 2012). A report produced by the Pew Center on the States (2009) found that of the slightly more than 7 million individuals involved in the criminal justice system, approximately 70% were supervised in the community rather than being incarcerated. While the number of offenders under correctional control has declined slightly in recent years, the proportion of individuals supervised in the community has remained about the same (Maruschak and Parks, 2012; Carson and Sabol, 2012). In this article, the authors describe community supervision in the United States, focusing on one supervision model known as diversion.

What Is Community Supervision?

Community supervision in the United States encompasses the modes of correctional control where individuals are monitored in the community. The primary types of community supervision include probation or parole. Both serve as alternative punishments to incarceration by allowing the individual to serve their complete or partial sentence in the community. The primary distinction between probation and parole is that parole follows release from prison while probation can follow a short incarceration or be the criminal sanction alone (Maruschak and Parks, 2012). Prisons are correctional facilities where individuals are sent for incarceration sentences greater than 1 year, whereas jails are facilities where individuals are incarcerated for sentences lasting less than 1 year. For many jurisdictions within the United States, probation and parole supervision are interchangeable and community supervision agents monitor both types of cases.

While under community supervision, individuals are monitored by criminal justice agents who enforce the terms of conditions placed on them through either the court or a parole board. A large part of community supervision requires offenders to abide by conditions of their release. For example, this may include having a curfew, frequency of contacts with their parole/probation officer, and not carry firearms. Put simply, offenders' behavior is controlled as a way to get them to comply with conditions of their supervision.

The primary philosophy and goal of community supervision in the United States has ebbed and flowed between rehabilitation and punishment orientations over the last century. Sociopolitical context within the United States has often determined the strength of support for either of these paradigms (Garland, 2001; MacKenzie, 2001). The original foundation of community supervision as alternatives to imprisonment, while providing rehabilitation, originated in England in the mid-1800s. In the 1840s, John Augustus, known as the contemporary father of probation, requested that an alcoholic be released into his custody under supervision for duration of 3 weeks. After this time period, the alcoholic returned and informed the judge of his reform (Petersilia, 2011; Taxman, 2014). Parole followed in the footsteps of probation with the early work of Alexander Maconochie and Sir Walter Crofton (Morris, 2002). Maconochie believed offenders should be rewarded for good conduct and, eventually, earn early release. Referred to as the marks system, offenders could gain their freedom by accumulating enough marks for completing tasks asked of them. This focus on rehabilitation was largely reflected in community supervision practices before the 1970s. Beginning in the 1970s, there was a shift toward control and punishment as paradigms in the criminal justice system, which was reflected in community supervision with a focus toward surveillance rather than rehabilitation (MacKenzie, 2001).

Recently, the pendulum has begun to swing back toward rehabilitation; however, the vying between paradigms has created and altered various models of community supervision over time. Diversion programs in community supervision came about in the 1960s with various other supervision models focused around rehabilitation and case management (Lawrence, 1991). The aims of these programs were to provide appropriate services to individuals who came into contact with the criminal justice system and to reduce the stigmatizing impact the criminal justice system can have on individuals. The role of diversion programs has traditionally been to lessen the amount of contact individuals have with the criminal justice system, therefore reducing any negative social labeling arising from the contact (Lawrence, 1991). Additionally, diversion programs can act as a mechanism to reduce the severity of a punishment an offender will serve while under community supervision.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45074-9

Community Supervision and Diversion

Originally, diversion programs fell into one of two types. The first type was complete or true diversion away from any formal processing by the criminal justice system. Through this model, individuals were least likely to be labeled or stigmatized through entering the criminal justice system and would instead receive necessary services to change their offending behavior (Lawrence, 1991). The second type was a leveraging stance where charges would be dropped if certain conditions were followed by the charged individual (Lawrence, 1991). This approach to diversion ensured the ability to enforce that required steps were taken by the offending individual to either receive treatment to change their behaviors or pay back the community or victim for wrong doing. This pay back could take forms such as community service or restitution.

As the philosophy of the criminal justice system moved away from a rehabilitation focus toward a punishment orientation, the types of diversion programs adapted with the changing goals. The original diversion types remained, but the punishment orientation of the criminal justice system helped develop two additional types of diversion models that focused more on control and punishment rather than on changing the individual's behaviors. One of these types can be called probation diversion. These are programs intended to be an alternative sanction to jailing individuals such as driving under the influence (DUI) programs. The second type is an enhanced probation diversion, which focuses on keeping higher risk individuals from entering prison.

All four types of diversion often involve legal pressures or coercion for individuals to participate in programming. Among the issues raised regarding legal coercion for treatment are ethical considerations of individual liberties and a lack of motivation among those coerced. Despite being of research interest for several decades, the use of legal pressure for treatment delivery, particularly for substance users, is still highly debated with mixed research findings (Klag et al., 2005). Some research finds that the severity of coercion or severity of the alternative punishment plays a part on the outcome (Hiller et al., 1998; Young et al., 2004). For instance, Young et al. (2004), assessing the Treatment Alternatives to Street Crime and Drug Treatment Alternative to Prison program as compared to other criminal justice referrals relying on less severe sentences, found individuals who were legally pressured into these two programs had lower levels of recidivism. We further discuss these four types of diversion programs in greater detail below.

Complete diversion, also referred to as true, prebooking, or precharge diversion, involves never allowing an individual to enter the court system to be processed by the criminal justice system. This type of diversion is typically initiated by police or medical/mental health professionals without the input of any community correction personnel. Complete or true diversion programs are typically reserved for juveniles, individuals with serious mental health disorders, substance use disorders, or co-occurring disorders who have committed a nonviolent, typically, misdemeanor offense. The goal of these programs is to increase access to needed treatment services that would prevent future, minor, criminal behavior from reoccurring.

Justice leverage diversion, also referred to as postbooking diversion, is when an individual is charged with an offense and goes to court, but offered to have the charges removed if they follow certain conditions. If the individual does not successfully complete the conditions, then they are processed through the criminal justice system with the charges remaining. Typically, the conditions to remove the charges are laid out by the court but overseen by the community supervision agency. In these instances, community supervision plays a vital role in determining whether an individual will have a criminal record or deemed successful at meeting the conditions for the charges to be lifted.

Probation diversion, or jail diversion, involves an individual who is adjudicated and sentenced by the court, but instead of being sentenced to jail time, the individual is ordered to be supervised in the community. Similar to the justice leverage diversion, conditions such as substance abuse treatment, mental health treatment, community service, or drug testing may be placed on the individual while in the community that they must follow or they will be sent to serve time in jail. However, in this type of diversion program, the charges remain even after the individual successfully completes the community supervision sentence. As with the other types of diversion, jail-based diversion programs are used with individuals with serious mental health diagnoses. The individual must be referred by a criminal justice actor (e.g., judge, lawyer) and then screened by a service provider (Lange et al., 2011; Lamberti and Weisman, 2004). After this process, individuals are either diverted into a treatment program or they can be completely diverted (Draine and Solomon, 1999).

Enhanced probation diversion, or prison diversion, includes community supervision programs that serve as an alternative either partially or completely to time in prison. These types of programs would typically be for individuals who are considered higher risk of reoffending and nonviolent who are in need of services that can be delivered in a community setting. Intensive supervision programs (ISPs) were intended to meet the need of this population by providing more tightly controlled community supervision than normal probation while reducing the number of individuals serving time in prison.

These four types of diversion programs, complete diversion, justice leverage diversion, probation diversion, and enhanced probation diversion, represent a framework for differentiating the various ways to manage offender populations in a community setting. Each of these types of diversion can be thought of as falling onto a continuum between the rehabilitation and punishment paradigms (Table 1). Complete

Table 1 Philosophical paradigms and diversion programs

Types of diversion	*Rehabilitation*	*Control*
Complete	+	−
Justice leverage	+	+
Probation diversion	−	+
Enhanced	−	+

+ = more emphasis; − = less emphasis.

diversion programs favor entirely a rehabilitative approach, whereas the probation and enhanced probation diversion have typically fallen completely into the paradigm of punishment. The justice leverage diversion program is typically seen as more of a balance between the paradigms. However, as we will discuss later, all of these diversion programs have the potential to find this balance between paradigms, which hopefully leads to improved individual outcomes and public safety. In the upcoming sections, we will discuss the prevalence of these types of diversion programs and the empirical evidence of each. The empirical evidence on the types of diversion programs is essential as to ascertain its relevance in the field of criminal justice.

Prevalence of Diversion Programs

Estimating the actual number of diversion programs used in the United States is presently beyond the scope of current research. Since most diversion programs are operated by local criminal justice agencies, determining a total number of programs overall for each of these four types would be an enormous undertaking. However, for some of the larger, more widely implemented programs, prevalence estimates are available.

Among complete diversion programs for adult populations, crisis intervention teams (CITs) are the most widely adopted across the country and world. As of 2008, there were an estimated 400 of these programs operating in the United States (Watson et al., 2008; Compton et al., 2008). Most recently, CIT International (2013) estimates that there are approximately 3000 of these programs operating internationally. A justice leverage diversion program that has been adopted across most of the country is that of problem-solving courts. As of 2009, there were 3648 total problem-solving courts operating in the United States (Huddleston and Marlowe, 2011). In 1992, Steadman et al. (1994) conducted a national survey of jails with a capacity of at least 50 inmates as identified through the American Jail Association. Of the 1263 jails surveyed, only 52 had a mental health jail diversion program that both screened for mental illness and used the program as either justice leverage or simply jail diversion.

These estimates provide a picture of the great variation in the number of these types of diversion programs operating in the United States. Recently, diversion programs, in the United States, have become more prevalent as a mechanism to divert offenders with mental health issues from the criminal justice system, including jail and prison (Lange et al., 2011). The growth of diversion programs comes in part from the support of organizations such as the Bureau of Justice Assistance and Substance Abuse and Mental Health Services Administration. Such agencies have recognized the need for diversion within criminal justice, particularly for the approximately 6–16% of offenders suffering from mental health-related issues (Ditton, 1999; Lamb and Weinberger, 1999). In the next section, we discuss the evidence as it pertains to complete diversion, justice-leverage diversion, probation diversion, and enhanced probation diversion programs.

What Do We Know About These Diversion Models?

Complete Diversion

Research on complete or true diversion programs with adult populations has largely been limited to programs focused on diverting individuals with serious mental illness or co-occurring disorders. The limited research does appear promising in that these diversion programs work to decrease criminal justice involvement and increase access to services within this specific population (Compton et al., 2008; Steadman and Naples, 2005; Steadman et al., 2000). However, the available research evidence is limited to CITs or other similar emergency/crisis response teams.

CITs are policing teams that receive specialized training to appropriately refer mentally ill individuals to mental health treatment services instead of involving them in the criminal justice system (Compton et al., 2008). This type of diversion program originated in Memphis, Tennessee, with the goals "to provide immediate response to and management of situations where the mentally ill are in a state of crisis; prevent, reduce, or eliminate injury to both the consumer and the responding police officer; find appropriate care for the consumer; and establish a treatment program that reduces recidivism" (Vickers, 2000: p. 2). There is some research indicating that CITs improve patient outcomes and increase the likelihood of receiving services sooner than those who enter the criminal justice system (Compton et al., 2008; Watson et al., 2010; Steadman et al., 2000). However, a recent study by Watson et al. (2010) found that the training did not impact arrest decisions made by officers. These authors also found the impact of training to be moderated by the officers' prior perceptions and knowledge of mental illnesses and services.

Despite the available research, there is still a lot unknown about complete diversion programs such as CIT. Most of the research is methodologically weak, examining only outcomes about officer perceptions rather than patient outcomes (Hartford et al., 2006; Compton et al., 2008). Lattimore et al. (2003) found that when comparing prebooking to postbooking diversion participants, prebooking participants were employed, had more education, less severe criminal history, mental health history, and substance use history, and overall more satisfied with life. More research needs to be done to understand the impact such differences among participants have on program effectiveness. Future research needs to explore how community supervision officers can impact these programs. For instance, these officers have a specialized knowledge regarding the availability of community resources for an underserved, criminal justice population that could potentially aid the effectiveness of these teams.

Justice Leverage Diversion

The most researched type of justice leverage diversion programs are problem-solving courts or specialty courts (Belenko, 2001). Specialty courts such as mental health courts and drug courts operate under the element of therapeutic jurisprudence. Rather than adhering to traditional forms of punishment, these courts seek to identify the etiology of offenders' behavior and treat it (Schneider, 2008). Since these courts are voluntary, offenders are provided with conditions to adhere to community

treatment that result in outcomes such as suspending their charges, sentencing, or probation (Griffin et al., 2002).

Much of the research on these courts is centered on the effectiveness of drug courts (Wilson et al., 2006; Rossman et al., 2011; Mitchell et al., 2012). In a meta-analysis of drug courts, Wilson et al. (2006) discovered, as a whole, drug courts reduced future offending for clients going through the program compared to those going through other correctional programs. In particular, they found that courts used as preplea or postplea models were most effective. Mitchell et al. (2012) supported this general finding, particularly among adult and driving while intoxicated (DWI) drug courts, and they also found that courts using either expungement or dismissal of charges greatly reduced drug recidivism. Recently, Rossman et al. (2011) conducted a multisite quasi-experimental design to examine the efficacy of drug courts as compared to control sites. Examining multiple outcome measures, the authors found that drug courts produced a significant impact in reducing both recidivism and substance use for clients. Additional benefits were detected for psychosocial outcomes, including employment and education as well as long-term impacts once the client had been out of drug court.

There is a growing body of research on other problem-solving courts such as mental health. Conducting a pre–post comparison of 386 misdemeanants in a mental health court, Henrinckx et al. (2005) found that these clients received more case management, management for their medication, and outpatient services after enrolling in the program. Mental health court clients also received fewer crisis services and inpatient treatment once enrolled in the program. In terms of recidivism, Henrinckx et al. (2005) discovered that more than half of the clients participating in the mental health court program experienced reductions in recidivism by more than half. Additionally, nearly two-thirds of clients experienced no new probation violations. In a similar study, Frailing (2010) used a quasi-experimental design on a sample of clients going through mental health court compared to sample of clients receiving treatment as usual. Frailing (2010) found for those clients participating in the mental health court significant reductions in both days incarcerated and substance used were experienced. In one randomized controlled trial (RCT) comparing mental health court clients receiving assertive community treatment, Cosden et al. (2005) found that clients participating in the mental health court with assertive community treatment had fewer new charges, reductions in substance abuse, fewer days in jail, and better improvements in mental health-related matters, including psychosocial functioning.

The evidence is very positive for the continued use of diversion programs in the form of problem-solving courts. The research shows through a variety of types of designs, including pre–post, quasi-experimental, and RCTs, problem-solving courts such as mental health courts and drug courts produce positive outcomes for clients who go through them. Despite these findings, though, more needs to be understood about problem-solving courts as well as other types of justice leverage diversion programs. The available research does not dissect the active ingredients within problem-solving courts (Wilson et al., 2006), thus making it not yet possible to conclude whether the nature of using them as a justice leverage diversion program drives the effectiveness.

Probation Diversion

Probation diversion programs include jail diversion (Shafer et al., 2004), postplea drug courts (i.e., after conviction) (Wilson et al., 2006; Mitchell et al., 2012), DUI diversion programs (Taxman and Piquero, 1998), and Proposition 36 (Farabee et al., 2004). Jail-based diversion programs have found some support in the empirical literature (Steadman et al., 1999). For instance, Lamberti et al. (2001) conducted a pre–post comparison of clients. The authors found that the average number of jail days as well as the average number of hospital days dropped and was statistically significant. Additionally, arrests, incarcerations, and hospitalizations were found to be reduced and statistically significant. However, other research has found no difference on outcomes between diverted individuals and nondiverted individuals with mental illness diagnoses beyond correctional cost savings (Steadman et al., 1999). Similarly, among postplea drug courts, meta-analysis finds them to be effective; however, recent results demonstrate that postplea drug courts are no more effective at reducing recidivism then preplea courts despite the threat of immediate severe sentencing available since conviction already occurred before the drug court experience (Wilson et al., 2006; Mitchell et al., 2012).

DUI diversion programs have been targeted at those individuals experiencing alcohol-related issues (Anson, 1987; Beck et al., 1999; Taxman and Piquero, 1998). For example, Anson (1987) examined the effectiveness of the Criminal Alcoholic Program using a longitudinal, pre–post design on a sample of client with criminal behavior, including alcohol-related issues. He found that the most statistically significant reductions in criminal behavior were for traffic offenses and alcohol offenses. Beck et al. (1999) tested the efficacy of an ignition interlock license reinstatement program on a sample of clients for 2 years. The study used an RCT and for those randomly assigned to the interlock group, significant reductions were experienced in terms of alcohol traffic violations in year one but no differences were detected in year two. These results led the authors to conclude the program ability to reduce alcohol-related offenses.

More recently, California has enacted Proposition 36 which is also known as the Substance Abuse and Crime Prevention Act (SACPA) (Farabee et al., 2004). This piece of legislation was voted on by Californians and seeks to act as a type of diversion for clients convicted of drug-related offenses. The purpose of the legislation is to use treatment as a diversion from incarceration. Comparing a sample of non-SACPA participants to SACPA participants, the researchers discovered that SACPA participants were more likely to experience being rearrested for a drug-related offense. The empirical research discussed in terms of jail-based and DUI diversion programs show some support in reducing future criminality of offenders but do not appear to be as strong as those discussed with justice leverage programs.

Enhanced Probation Diversion

Enhanced probation diversion programs come in the form of intermediate sanctions developed with the intent to reduce prison populations by supervising individuals more closely in

the community. The most prominent of this type of diversion are ISPs. These programs began in the 1980s as way to manage individuals deemed too dangerous for normal probation, but not serious enough for prison (Petersilia and Turner, 1993). ISPs vary greatly in what they include across the multitude of jurisdictions within the United States. Some of the common components are electronic monitoring (EM), drug testing, increased visits with community supervision agents, and home arrest (Petersilia and Turner, 1993). The findings regarding the effectiveness of ISPs has been mixed, with some research even finding that the closer supervision included in these programs actually increases recidivism (Turner et al., 1992; Petersilia, 1998).

Examining the use of intensive community supervision as a true prison diversion as compared to enhanced parole, Deschenes et al. (1995) looked at programs implemented in some Minnesota counties in the early 1990s. The programs included home detention, EM, required time on employment and education activities, drug testing, and frequent meetings with supervising agents. The authors found that, as a prison diversion, the intensive community supervision resulted in less time in prison, but more days on supervision than the control group. The prison diversion group had more technical violations than the control group, but there were no differences in new arrests. The intensive community supervision as an enhanced parole release did not have an impact.

ISPs have been criticized because while they were created to reduce prison intakes by serving as an intermediate sanction, some research has found that they ended up widening the net for less serious offenders who would have previously just been on probation (Petersilia and Turner, 1993; Byrne, 1990). In very few jurisdictions have ISPs actually been implemented as true prison diversion programs. Despite the overall negative findings, ISPs have been found to be effective when increased surveillance or punishment is used in tandem with treatment services to rehabilitate the offender (Petersilia, 1998). Similar support for this has been found in research conducted in Canada. Bonta et al. (2000) examined a program coupling EM and intensive evidence-based treatment. They found that EM alone had no impact, whereas the combined group was more likely to complete treatment at no greater risk to public.

In addition to ISPs, enhanced probation diversion includes some postplea drug courts. These drug courts provide similar levels of supervision as ISPs while addressing specific needs for substance users in the community.

Conclusions

Diversion programs are a widely used and supported form of community supervision, particularly for individuals with serious mental illness, substance use disorders, or co-occurring disorders. However, there is only limited research on the various types of diversion programs that are in operation. A key to furthering the research agenda across all four of these types is the investigation of legal pressure or coercion on individuals mandated to complete programming. All of these types of diversion exert coercive treatment upon individuals. Whether it is complete diversion where social workers and police directly order individuals to treatment or it is enhanced probation diversion where individuals must wear EM and go to drug court to stay out of prison, applied legal pressure is necessary for diversion to be successful.

The critiques offered by Klag et al. (2005) about the research conducted on the use of legal pressure with substance abusers offers a starting point for filling in the gap within the broader diversion program research base. These authors list five critiques of past coercion research as follows: (1) inadequate research designs, (2) poor measures for coercion, (3) no consistent definition of coercion, (4) unknown differences between voluntary and involuntary participants, and (5) lack of a continuum coercion measure. Research such as that conducted by Young et al. (2004) and Mitchell et al. (2012) begin to examine some of these inadequacies, but rigorous, clearly conceptualized research must examine differences between the amount/severity of coercion each of these program types uses. Understanding which, if any, of the legal pressures applied in these various diversion program types works best would provide an underlying basis for developing diversion programs aimed at best outcomes.

Furthermore, a common theme that runs throughout the available research is that diversion programs that strike a balance between the philosophical paradigms of rehabilitation and punishment tend to be the most effective. Such a form of community supervision has been termed behavioral management supervision (Taxman, 2014). Behavioral management supervision combines evidence-based practices (i.e., what research tells us works managing offenders) and treatment services. This type of supervision allows for community supervision agents to monitor and control individuals to protect public safety, but at the same time use evidence-based tools to promote individual change to prevent future criminal behavior. Advancing this balance between paradigms is important not only for community supervision practices overall, but also, as seen here, for specifically improving diversion programs.

See also: Re-entry and Reintegration After Incarceration; Rehabilitation of the Offender.

Bibliography

Anson, R.H., 1987. An examination of an intensive probation program for alcohol offenders. International Journal of Offender Therapy and Comparative Criminology 31 (3), 271–283.

Beck, K.H., Rauch, W.J., Baker, E.A., Williams, A.F., 1999. Effects of ignition interlock license restrictions on drivers with multiple alcohol offenses: a randomized trial in Maryland. American Journal of Public Health 89 (11), 1696–1700.

Belenko, S., 2001. Research on Drug Courts: A Critical Review 2001 Update.

Bonta, J., Wallace-Capretta, S., Rooney, J., 2000. A quasi-experimental evaluation of an intensive rehabilitation supervision program. Criminal Justice and Behavior 27 (3), 312–329.

Byrne, J., 1990. The future of intensive probation supervision and the new intermediate sanction. Crime and Delinquency 36 (1), 6–41.

Carson, E.A., Sabol, W.J., 2012. Prisoners in 2011. Bureau of Justice Statistics.

CIT International, 2013. http://www.citinternational.org/.

Compton, M.T., Bahora, M., Watson, A.C., Oliva, J.R., 2008. A comprehensive review of extant research on crisis intervention team (CIT) programs. Journal of the American Academy of Psychiatry and the Law 36 (1), 47–55.

Cosden, M., Ellens, J., Schnell, J., Yamini-Diouf, Y., 2005. Efficacy of a mental health treatment court with assertive community treatment. Behavioral Sciences and the Law 23 (2), 199–214.

Deschenes, E.P., Turner, S., Petersilia, J., 1995. A dual Experiment in intensive community supervision: Minnesota's prison diversion and enhanced supervised release programs. The Prison Journal 75 (3), 330–356.
Ditton, P.M., 1999. Special Report Mental Health and Treatment of Inmates and Probationers. US Department of Justice, Bureau of Justice Statistics, Washington, DC.
Draine, J., Solomon, P., 1999. Describing and evaluating jail diversion services for persons with serious mental illness. Psychiatric Services 50 (1), 56–61.
Farabee, D., HSER, Y.I., Anglin, M.D., Huang, D., 2004. Recidivism among an early cohort of California's proposition 36 offenders. Criminology & Public Policy 3 (4), 563–584.
Frailing, K., 2010. How mental health courts function: outcomes and observations. International Journal of Law and Psychiatry 33 (4), 207–213.
Garland, D., 2001. The Culture of Control: Crime and Social Order in Contemporary Society. University of Chicago Press, Chicago.
Griffin, P.A., Steadman, H.J., Petrila, J., 2002. The use of criminal charges and sanctions in mental health courts. Psychiatric Services 53 (10), 1285–1289.
Hartford, K., Carey, R., Mendonca, J., 2006. Pre-arrest diversion of people with mental illness: literature review and international survey. Behavioral Sciences and the Law 24, 845–856.
Henrinckx, H.A., Swart, S.C., Ama, S.M., Dolezal, C.D., King, S., 2005. Rearrest and linkage to mental health services among clients of the Clark County mental health court program. Psychiatric Services 56, 853–857.
Hiller, M.L., Knight, K., Broome, K.M., Simpson, D.D., 1998. Legal pressure and treatment retention in a national sample of long-term residential programs. Criminal Justice and Behavior 25, 463–481.
Huddleston, W., Marlowe, D.B., 2011. Painting the Current Picture: A National Report on Drug Courts and Other Problem-Solving Court Programs in the United States. National Drug Court Institute, Alexandria.
Klag, S., O'Callaghan, F., Creed, P., 2005. The use of legal coercion in the treatment of substance abusers: an overview and critical analysis of thirty years of research. Substance Use & Misuse 40, 1777–1795.
Lamb, H.R., Weinberger, L.E., 1999. Persons with severe mental illness in jails and prisons: a review. Issues in Community Treatment of Severe Mental Illness: A Compendium of Articles from Psychiatric Services 8.
Lamberti, J.S., Weisman, R.L., Schwarzkopf, S.B., Price, N., Ashton, R.M., Trompeter, J., 2001. The mentally ill in jails and prisons: towards an integrated model of prevention. Psychiatric Quarterly 72 (1), 63–77.
Lamberti, J.S., Weisman, R.L., 2004. Persons with severe mental disorders in the criminal justice system: challenges and opportunities. Psychiatric Quarterly 75 (2), 151–164.
Lange, S., Rehm, J., Popova, S., 2011. The effectiveness of criminal justice diversion initiatives in North America: a systematic literature review. International Journal of Forensic Mental Health 10 (3), 200–214.
Lattimore, P.K., Broner, N., Sherman, R., Frisman, L., Shafer, M.S., 2003. A comparison of prebooking and postbooking diversion programs for mentally ill substance-using individuals with justice involvement. Journal of Contemporary Criminal Justice 19 (1), 30–64.
Lawrence, R., 1991. Reexamining community corrections models. Crime and Delinquency 37 (4), 449–464.
MacKenzie, D.L., 2001. Corrections and sentencing in the 21st century: evidence-based corrections and sentencing. The Prison Journal 81 (3), 299–312.
Maruschak, L.M., Parks, E., 2012. Probation and Parole in the United States.
Mitchell, O.J., Wilson, D.B., Eggers, A., MacKenzie, D.L., 2012. Assessing the effectiveness of drug courts on recidivism: a meta-analytic review of traditional and non-traditional drug courts. Journal of Criminal Justice 40 (1), 60–71.
Morris, N., 2002. Maconochie's Gentleman: The Story of Norfolk Island, the Roots of Modern Prison Reform. Oxford University Press, New York.
Petersilia, J., 1998. A decade of experimenting with intermediate sanctions what have we learned? Federal Probation 62 (3), 3–9.
Petersilia, J., 2011. Community corrections: probation, parole, and prisoner reentry. In: Wilson, J.Q., Petersilia, J. (Eds.), Crime and Public Policy. Oxford University Press, New York, pp. 499–531.
Petersilia, J., Turner, S., 1993. Intensive probation and parole. Crime and Justice, 281–335.
Pew Center on the States, 2009. One in 31: The Long Reach of American Corrections.
Rossman, S.B., Rempel, M., Roman, J.K., Zweig, J.M., Lindquist, C.H., Green, M., et al., 2011. The Multi-site Adult Drug Court Evaluation: The Impact of Drug Courts, vol. 4. Urban Institute Justice Policy Center, Washington, DC.
Schneider, R.D., 2008. Mental Health Courts. Wiley Encyclopedia of Forensic Science.
Shafer, M.S., Arthur, B., Fanczak, M.J., 2004. An analysis of post-booking jail diversion programming for persons with co-occurring disorders. Behavioral Sciences and the Law 22, 771–785.
Steadman, H.J., Barbera, S.S., Dennis, D.L., 1994. A national survey of jail diversion programs for mentally ill detainees. Hospital and Community Psychiatry 45 (11), 1109–1113.
Steadman, H.J., Cocozza, J.J., Veysey, B.M., 1999. Comparing outcomes for diverted and nondiverted jail detainees with mental illnesses. Law and Human Behavior 23 (6), 615–627.
Steadman, H.J., Deane, M.W., Borum, R., Morrissey, J.P., 2000. Comparing outcomes of major models of police responses to mental health emergencies. Psychiatric Services 51 (5), 645–649.
Steadman, H.J., Naples, M., 2005. Assessing the effectiveness of jail diversion programs for persons with serious mental illness and co-occurring substance use disorders. Behavioral Sciences and the Law 23, 163–170.
Taxman, F.S., Piquero, A., 1998. On preventing drunk driving recidivism: an examination of rehabilitation and punishment approaches. Journal of Criminal Justice 26 (2), 129–143.
Taxman, F.S., 2014. Behavioral management probation. In: Bruinsma, G., Weisburd, D. (Eds.), Encyclopedia of Criminology and Criminal Justice. Springer, New York, pp. 134–145.
Turner, S., Petersilia, J., Deschenes, E.P., 1992. Evaluating intensive supervision probation/parole (ISP) for drug offenders. Crime Delinquency 38, 539–556.
Vickers, B., 2000. Memphis, Tennessee, Police Department's Crisis Intervention Team. U.S. Department of Justice: Bureau of Justice Assistance.
Watson, A.C., Morabito, M.S., Draine, J., Ottati, V., 2008. Improving police response to persons with mental illness: a multi-level conceptualization of CIT. International Journal of Law and Psychiatry 31 (4), 359–368.
Watson, A.C., Ottati, V.C., Morabito, M., Draine, J., Kerr, A.N., Angell, B., 2010. Outcomes of police contact with persons with mental illness: the impact of CIT. Administration and Policy in Mental Health 37, 302–317.
Wilson, D.B., Mitchell, O., MacKenzie, D.L., 2006. A systematic review of drug court effects on recidivism. Journal of Experimental Criminology 2 (4), 459–487.
Young, D., Fluellen, R., Belenko, S., 2004. Criminal recidivism in three models of mandatory drug treatment. Journal of Substance Abuse Treatment 27 (4), 313–323.

Community Work

Kieron Hatton, School of Health Sciences and Social Work, University of Portsmouth, Portsmouth, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Community work has been a dominant approach in delivering social welfare since the beginning of the twentieth century. This article considers the contested meaning of community, key theories in community work, and the close links between theory and practice. Focusing on three key thinkers, Alinsky, Gramsci, and Freire, developments in several countries' and NGOs' utilization of community development to further their agendas is examined. A short discussion of community youth development and social pedagogy is also provided with a conclusion of a reflection upon the potential efficacy of community work in achieving positive outcomes for marginalized communities across the globe.

Introduction

Notions of 'community,' a contested term, underpin community work. In this article, I examine definitions of 'community,' key theories in community work, and the links between theory and practice. I use three key thinkers, Alinsky, Gramsci, and Freire, to consider developments across Europe and how NGOs utilize community development to pursue their agendas. I end with a short discussion of community youth development and social pedagogy, and conclude by reflecting upon the potential usefulness of community work in achieving positive outcomes for marginalized communities everywhere.

Defining Community and Community Work

Community work as a paradigm within the work of human services or social professionals is a complex set of activities which can range from small, unfunded activities at a very local level (the street) to more complex activities at the level of communities, regions, or even across and between countries. At a political or philosophical level, it can encompass work which can be seen as unchallenging in that it meets needs which the state cannot or will not meet or as a radical and challenging approach which sees social justice as its primary focus and social change at the local, regional, national, or international levels as its overarching aim.

Part of the reason for this ubiquity is the concept of community itself. This is a highly contested concept which is appropriated by the right and left to support claims of accountability and relevance. As a major social theorist says:

> Community can be the warmly persuasive word used to describe an existing set of relationships. What is most important, perhaps, is that unlike all other terms of social organisation (state, society, nation) it seems never to be used unfavourably and given any opposing term.
>
> Williams, 1976: p. 77

'Community' is one of the most widely used and misused words in social science. It is often used indiscriminately to enhance the appeal of policies, which may themselves be unattractive – community policing, community safety, community partnership, community school, e.g., when the old Borstals were renamed in this manner in the UK. Day (2006) notes that it is often used in an ill-defined and vague way as "an excuse for not thinking hard enough about what exactly it is that people have in common … alluring in its promise but to be approached with extreme care" (Day, 2006: p. 2). What then does it mean?

Tonnies (1955) used the idea of community to describe the transformation in social relationships brought on by modernization (Breuggemann, 2006). Bauman (2001) sees community as a counterbalance to the atomization and individualization of a globalizing world and as a way through which control can be gained or regained. He argues that

> We all need to gain control over the conditions under which we struggle with the challenges of life – but for most of us such control can be gained only collectively.
>
> Bauman, 2001: p. 149

Delanty and Styrom (2003) sees community as a new way of understanding kinds of belonging with multiple and overlapping relationships at its center. The boundaries of community are contested and community in this sense is more a discourse with a focus on, "an imagined condition than a symbolically shaped reality based on fixed reference points" (p. 188). For the purpose of this article, I conceptualize community as an alternative to individualistic discourses which atomize people's experience and suggest instead ways in which people can regain control over their lives and build positive relationships which promote well-being. These dichotomies in the theorization of community reflect divergences in discourses around community work which I examine later.

Community Work – Key Theories and Theorists

Mayo (2008) argues that there were three stages in the development of community work. The first, she characterizes as the postcolonial legacy (during the 1960s) – the period when British community development reinvented itself; a period of optimism. The original colonial focus on winning the 'hearts and minds' of local indigenous peoples was replaced with a restructuring of Western economies, a fear of the consequences of the Cold War, and the growth of community

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28018-5

activism. The second stage revolved around community work's relationship with the state. Community work was originally seen by western/northern states as a means of resolving social problems, which were primarily caused by the inadequacies of poor and marginalized communities who were unable to manage their own affairs and needed 'experts' to assist them to do so (the 'social pathology' model of social change). Radical community workers soon recognized that the problems facing many of the areas in which they worked had similar causative factors (high unemployment and high levels of poverty, the restructuring of industry, the movement of finance capital outside of these areas). The social pathology paradigm was therefore replaced with a structural/Marxist analysis of the state and the role of community work within it.

A number of attempts have been made to conceptualize community work. The UK's *Gulbenkian Report* published in 1973 drew a distinction between community development and community action. The former is defined as based on consensus, is client-centred and an approach that has the community worker in an enabling role. It was based more on community consultation than community control. In community action, the Report's authors argued, the capacity to "appoint, dismiss and define the role of workers lies within the group itself" (p. 49). They conclude that, in contrast, "an outstanding characteristic of community action is the involvement of the socially deprived – the very ones usually excluded from other organizations – in order to improve their situation" (p. 50).

Stepney and Popple (2008) note a similar distinction. They characterize the community development approach as a reformist approach in which power is dispersed and exercised by the mass of people rather than a small elite or class. They note, "It encourages participation and prevents power being concentrated in the hands of the few" (Stepney and Popple, 2008: p. 139). This community action approach, they now characterize as transformative community action. Advocates of this approach argue that power is, 'concentrated in relatively few hands' and suggest: that, community work, "is more likely to adopt a radical approach that advocates far reaching and wide spread changes … it promotes autonomous ways of working through collective action to address moral and political issues."

Stepney and Popple, 2008: p. 140

It has been suggested that community work can be part of an oppositional politics which provides an alternative to neo-liberalism (De Filippis et al., 2009). De Filippis et al. suggest that "community organizations need to look beyond the local and incorporate a political and social analysis of the wider political economy into their work" (p. 50). Popple (1995) argues that community work has the potential to feed into debates surrounding social movements. The benefit for community work is an expanded understanding of macroforces which shape and impinge upon individuals and communities (p. 53). He draws a distinction between pluralist, radical/socialist, feminist, and antiracist community work. Briefly, these can be characterized as follows.

Pluralist Community Work

Pluralist community work, Popple argues, features a dispersal of power in which different stakeholders reach compromises rather than imposing their own solutions on communities. In a pluralist framework, power is equally shared and different stakeholders will be more influential at different times. The state's role becomes one of balancing competing interests and ensuring fairness. For the community worker, this means encouraging participation and the achievement of small-scale localized reforms. This approach lays considerable importance on the development of skills. Good practice becomes defined in relation to technical competence rather than values.

Radical and Socialist Community Work

Radical and socialist community work dates primarily from the 1960s and 1970s. It has involved community workers in advocating for large-scale change in the political, social, and economic system. Radical community work is often associated with anarchism and the belief that the state in particular is problematic and all existing structures need to be dismantled if real improvements are to be achieved. In some cases, echoes of this approach can be found in syndicalist and situationalist writings and activities at the beginning of the twentieth century.

However, it also incorporates the socialist tradition(s) from two primary sources: Fabianism and Marxism. Fabianism has its roots in the writings of GDH Cole and emphasizes the potential for achieving change within existing state structures. Unlike pluralism, with which it is sometimes associated, it does not confine itself to microlevel change and argues that the state can be appropriated for the benefit of people currently excluded from power. Fabianism does not, however, challenge the legitimacy of the state suggesting that social change should be gradual and not conflictual.

Marxist analysts see the state as reinforcing class inequality and argue that capitalist social relations and structures need to be overturned if lasting change is to be achieved (Purcell, 2012). This can be achieved by being 'in and against the state' and could happen when state workers, through their trade unions or other representative organizations work with local community activists to bring together a larger movement for change.

The London-Edinburgh Weekend Return Group (1980) and Bolger et al. (1981) argued that there was a need to take account of the potential of state employees to work within the state in a progressive way that could transform welfare capitalism into welfare socialism. They recognized the value of trying to find the oppositional space successfully to undertake work that sustains a struggle against the state (Popple, 1995: p. 36).

Feminist Community Work Theories

Feminist community work posed a challenge to the male-dominated community work theory in the late 1970s, arguing that there were specific factors in common between the growth of the second-wave feminism and the rise of community action by drawing links between personal and political action (Dominelli, 1990, 2006). Some feminist writers and activists, e.g., Morgan (1970), were criticized by black feminists like Lorde (1984) and Hill-Collins (1990) for concentrating primarily on the experience of white, middle-class women and playing down the marginalized experiences of black women.

Ledwith suggests that "white feminists stood accused of defining women from a White perspective, exposing a White power which has the arrogance to overlook other aspects of difference" (2009: p. 691) and argues for the incorporation of Gramsci into feminist debates. She says that Gramsci provided "the conceptual tools to make sense of my own life experience" (p. 695) and enables antioppressive activists to analyze and address, "the complex interconnectedness of a loci of oppressions" (p. 696).

Black and Antiracist Theories

Black and antiracist theories emphasize the link between race, class, and gender. They draw on black radicalism, socialism, Marxism, and black feminism. Central to this approach is the argument that state policy acted as a repressive force on black people (Ohri et al., 1982). Black and antiracist community workers developed a critique of society by focusing on how oppression/discrimination operate at personal, institutional, and structural levels. Importantly for community work it highlighted the need for white and black community workers to challenge racism in all its manifestations and the significant space that community work provides for communities and groups to develop their own cultural formations and organizations.

Dominelli (1998: p. 54) has suggested that:

> people from subordinated ethnic groups have had to develop their own services and to ensure that these services are capable of meeting their needs and can be used to challenge inadequate mainstream provisions.

Ife and Fiske (2006) argue that community development can contribute to tackling racism by promoting a human rights perspective predicated on a bottom–up approach to tackling structural oppression. In particular they point to how, "community based education, consciousness raising campaigns and ongoing movements at every level of society" (p. 300) can have a significant impact on public sentiments and contribute to developing an antiracist theory and practice.

Key Thinkers and Activists: Praxis

Central to understanding community work is the way that theoretical discourses are not standalone activities but provide understandings to guide action that leads to social change. Three theorists/activists who have had a significant impact on the development of community work, and who come from different traditions, are Saul Alinsky, Antonio Gramsci, and Paulo Freire. Each demonstrates in his own way as to how the synthesis of theory and action – *praxis* – can contribute to successful strategies of social change.

Alinsky

Alinsky was a radical community organizer in Chicago. He believed that if social and community change was to occur, an activity had to be underpinned with recognition of how power operates. In *Reveille for Radicals* (1969) he said:

> The organizer must be able to communicate and convince the people, that if they find a way to join together, they need not fatalistically accept their plights but will have the power to affect the shape of the world (p. 209).

> However his emphasis was on pragmatism and the development of achievable goals. He believed in the centrality of tactics and the need to constantly outthink the '*enemy*.' He argued that it was important to maintain constant pressure on the opposition. To achieve this he developed what he called 'power tactics' which he outlined in *Rules for Radicals* (1972). He suggested that "tactics is the art of how to take and how to give … our concern was with the tactic of taking" (p. 126).

Among the rules of power tactics he outlined were concerns with:

- acting outside of the opponents' experience and reacting quickly to the actions, and counteractions, of opponents;
- making sure that those involved in the campaign enjoyed it; and
- keeping the pressure on.

> Bunyan (2010) notes the centrality of power to Alinsky's work, in particular, "the importance of local institutions in the everyday lives of people … and the potential for such institutions to become the basis of a power organization that could act on the issues and concerns that mattered most to them" (p. 114). Alinsky argued the importance of education as a key element in the promotion of change and suggested that, education enabled people to make sense of the world in which they lived and consequently to "make informed and intelligent judgements."
>
> 1972: p. 124 about it

Alinsky's approach was influential within community work internationally, including the UK, Canada, Australia, and Hong Kong. While some people welcomed an approach that focused on organizing effectively and confronting real-world power relationships, others felt that his analysis lacked a broader framework for understanding how power operates. His writings and his campaigns highlighted tensions within the US and predicated on a pluralist analysis of society. He believed that people could achieve power within the locality or city, a belief that assumed that power was accessible if you had the tools to access it. As such, he ignored, or at least downplayed, structural inequalities such as class, race, and gender (Dominelli, 2006). Alinsky's ideas were supplemented by other writers who saw the potential for change from within a more sophisticated theoretical framework. Gramsci and Freire provide this theoretical framework.

Antonio Gramsci

Gramsci, an Italian Marxist active in establishing workers councils in Turin in the 1920s, rejected a simplistic version of Marxism which focused only on class relations. He developed a broader conception of politics which focused on the struggle for moral and political leadership through a process that he called *hegemony*. He argued that ruling classes dominate

through a mixture of force and consent where the consent is gained through their assuming the political, moral, and intellectual leadership in society. Gramsci saw civil society organizations as places where democratic struggles could be linked together to maintain hegemony.

The ruling class, he argued, retains control by making their rule and its legitimacy appear like 'common sense.' It is then the role of external agents such as intellectuals, union organizers, and community leaders to challenge common sense notions. Gramsci held that an oppressed group needs its own *organic intellectuals* (those that emerge from within their own ranks) to challenge these dominant ideas. He believed that social change could occur through the construction of alliances between industrial workers and civil society organizations but that organic intellectuals would have a key role in such alliances. Gramsci, talking about peoples' capacity to reflect upon their world says:

> 'everyone' is a philosopher … it is not a question of introducing from scratch a scientific form of thought into everyone's individual life, but of renovating and making 'critical' an already existing activity.
> Gramsci, 1971: pp. 330–331, cited in Allman, 2001: p. 114

Allman suggests that this passage should be read as Gramsci encouraging leaders or radical educators to help people question their already existing knowledge (common sense thinking) until it becomes 'critical activity' (dialectical thinking) (Allman, 2001: p. 11; Hatton, 2008: pp. 45–47).

Gramsci's ideas are important because they not only encapsulate the idea that people can take action to challenge unequal power relationships, but they also show how hegemonic or ruling ideas are transmitted. This then clarifies how ideas about how society is constituted, and then reconstituted. These are important considerations for community workers wishing to develop a more transformative practice.

Paulo Freire

Paulo Freire was a Brazilian educator, and at one time, Secretary of Education for Sao Paolo (Freire, 1993). Mayo (1999) describes Freire as possessing, "a vision of society characterised by relations of power and domination" (p. 59). His key ideas revolved around:

Agency – the ability of people to act in their own interests, but their capacity to 'act' can be impacted upon their psychological and social experiences.

Democratic Education – based on the idea of promoting education through *dialogue*, "the encounter of men [sic] in the world to transform the world" (1972: p. 100), as an alternative to conventional education. Central is the notion of praxis, the linking together of action and reflection.

Conscientizaçao (conscientization) is the key concept Freire used in enunciating how this will operate. Freire sees "education as an exercise in freedom, is an act of knowing, a critical approach to reality" (Mayo, 1999: p. 63). In Freire's words, "that lived experience be used as a point of departure so as to transcend it" (Freire, 1993: p. 110).

These are important processes if "human services practice is to become an emancipatory project, rather than a reinforcement of existing patterns of discrimination" (p. 215). Freire (1972) argues that a key way in which people without power are marginalized is through a process in which their behavior becomes pathologized and their human nature is distorted to construct what he describes as processes of indoctrination, manipulation, and 'dominated consciousness.' He argues that as a result, people lack the consciousness or understanding to decode their situations. Thus, he proposes a process of deindividualization, by which people are encouraged to see the commonality of their situation. Central to Freire's work is the idea of dialogue. Having established the importance of linking theory to practice, I look briefly at examples of the way in which this link has been expressed in community work.

Exemplifying Community Work/Community Action

Community Development Projects

The work of the Community Development Projects (CDP's) within the UK reflects earlier debates around the nature of community work. The introduction of the Community Development Projects in 1970 was initially seen as "a neighborhood based experiment" whose focus was on meeting the needs of people in "areas of high social deprivation" (CDP, 1974a: p. 1). The projects were originally based on a social pathology model of deprivation that assumed that the "problems of urban deprivation had their origins in the characteristics of local populations – in individual pathologies – and these could best be resolved by better field coordination of the social services, combined with the mobilization of self-help and mutual aid in the community" (1974a: p. 1).

By the time the CDP (1974b) *Inter-Projects Report* was published in 1974, many of the projects had changed emphasis and had moved toward a structural analysis of the urban deprivation that the communities that they were dealing with suffered from.

The arguments against social pathology were extended in the joint report, *Gilding the Ghetto* (CDP, 1977). The authors argued that, the major cause of poverty during the 1960s was the decline of the industrial base of the older areas. They pointed out how it was useful to the state to continue to define the problems of working class areas as due to as individual pathology. As they noted "the emphasis on tackling social needs in isolation inevitably distracts attention from the root causes of the problem by focusing attention upon personal deficiencies" (Craig et al., 2011: p. 95).

Assessing this Report's impact, Shaw (2003) suggested that "its title conveyed a powerful metaphor for the ideological value of 'community' in glossing over the structural causes of poverty and inequality" (p. 363). Shaw and Martin (2000: p. 404) suggested that the structuralist/Marxist analysis that the CDPs formulated represented "community work's potential to be both 'in and against' the state" (Shaw and Martin, 2000: p. 404).

Locally Based Community Development – Examples from Ireland and Wales

I have interviewed people working in community development projects in Wales and Ireland (Hatton, 2001). The project

leader in *We, The People*, a community development project in Cork, in Eire, argued that empowerment was a central issue for their organization:

> The empowerment aspect of it comes down to giving power back to people and making structures available to support them. But at the end of the day it's their power. By empowering someone you are giving them the power … the object of the organization is to empower people … you need to rebalance the power.
>
> Hatton, 2001: p. 209

A community worker in Cardiff, South Wales suggested that the challenge was to give people a belief in the efficacy of change, but suggested that:

> It's not realistic to believe that people on their own can achieve lasting change … we need a younger, more radical group to overcome the alienation people face. The focus should be on collective change at the political level. However, I am optimistic. I believe people are irrepressible, they will make demands.
>
> Hatton, 2001: p. 209

The Director of Community Action Network (CAN) also based in Cork suggested that this is one of the central elements of a community development approach:

> the vision that community development provides is one of promoting social change and that supporting communities to become more active and more empowered was a worthwhile exercise and if there was a gap in services it was worth pursuing.
>
> Hatton, 2001: p. 206

Community Development in Central Europe

Henderson and Vercseg (2010) have drawn attention to the way community development in Hungary grew out of community education and was influenced by ideas developed by Hungarian intellectuals influenced by Gramsci. They suggest that community development was influenced by two key approaches – social reform and foreign funding organizations or nongovernmental organizations (NGOs). There is also some evidence from the Czech Republic, of the growth of indigenous organizations who have developed what McLaughlin (1995) has called their own "ethnic intelligentsia" (a phrase he developed to describe the growth of an indigenous leadership within the Irish Traveling community). The development of Romdromas, a Roma-controlled organization negotiating directly with state and regional government bodies, 'Sduzeni Dzeno,' a Roma-run internet radio station, and the establishment of 'Evangelika Akademie Prague,' to train Roma as social pedagogues are examples of Roma peoples developing their own organizations and indigenous leadership (Hatton, 2008: pp. 133–138).

Community Youth Development

Perkins et al. (2001) apply ideas from community development in their discussion of community youth development (CYD). They suggest that CYD "shifts the emphasis from a dual focus of youth being problem free and fully prepared, to a triadic focus for youth being problem free, fully prepared and engaged partners … .[T]his focus recognises that there is an interdependent relationship between positive and healthy youth outcomes and positive and healthy communities" (p. 43). Perkins et al. (2001) suggest that community youth development at its core focuses on the young person's capacity to both understand and act upon the environment, a recognition that this involves the active support of people across the community. Central to such work is the engagement of youth in "constructive and challenging activities that build their competence and foster supported relationships with peers and with adults" (p. 48). Finally, Perkins et al. suggest that achieving these goals requires developments in youth participation and partnerships within the communities in which young people live.

NGOs and Community Development

Community development is an approach which is typical of many NGOs particularly those operating in African and Latin American contexts. Kamat (2004) suggests that:

> The presence of numerous NGOs at the grassroots that provide services to underprivileged populations, and at the same time mobilise them for their rights, presents a threat to the neo-liberal agenda of privatisation … In countries of the South, the Leftist origins of (NGOs) are reflected in their opposition to structural adjustment policies.
>
> Kamat, 2004: p. 167

Princen (1994) argues that NGO influence is:

> achieved by building expertise in areas diplomats tend to ignore and by revealing information economic interests tend to withhold … it is influence gained from speaking when others will not speak, from espousing something more than narrow self-interest, from sacrificing personal gain from broader goals, from giving voice to those who otherwise do not have it, from rejecting pessimism and looking for signs of hope.
>
> Princen, 1994 quoted in Betsill and Corell, 2001: p. 70

During the 1980s and much of the 1990s, the United Nations was highly responsive to the demands of the NGOs for participation and created important institutional space for NGOs to demand to participate in international policy forums on population, the status of women, and the environment (Kamat, 2004: 162).

Social Pedagogy and Community Development

Eriksson (2011) has pointed to the differences and similarities between community development and social pedagogy (Hatton, 2013: pp. 82–87). The former, she attributes to their differing origins (community development as a way of thinking, the latter as a way of doing); to the geographical spaces they occupy (European as against Anglo-Saxon); to the differing welfare regimes in which they operate; and to the way in which social pedagogy can be formulated as an

individual intervention while community development is by definition a collective response to the problems people face. However, Eriksson argues that there are clear similarities. Both provide a holistic understanding of education, development, and learning. She suggests that in both traditions, there is, "a tension between a radical and a conservative side" and that these raise the question of whether "the individual, society or both should be the focus of change" (Eriksson, 2011: pp. 414–417).

Lone (2010) raises similar issues when discussing the work of treatment collectives in residential care for people who misuse substances in Norway. She argues that because these agencies lack a hierarchy, there has been a "strong and radical political bias" (p. 62) which has allowed them to provide an alternative social pedagogy to the medicalized services available elsewhere. This, Lone argues, is partly to do with skepticism in Norway toward all kinds of residential care and quotes Edle Ravndal, who in a review of the working of collectively organized services suggests that:

> The collective model seems still to defend its place as an important component in a larger treatment system for the most challenged youths.
>
> Lone, 2010: p. 70

Social pedagogy and community development therefore appear to share the potential to transform people's experience and create new ways of delivering and experiencing welfare services. They can connect with traditions whose focus has been on promoting change not merely ameliorating the impact of inequality and disadvantage (Hatton, 2008, 2011).

Globalization and Community Development

Macionis and Plummer (2005: pp. 38–39) suggest that globalization had a number of elements that impact upon community work and community workers. These include:

- The development of new forms of international governance, e.g., the UN, ECHR, WHO.
- A growing awareness of common world problems.

Reviewing the welfare literature on globalization, Lyons et al. (2006: p. 37) suggest that the authors see the, "fragmentation of old arrangements; damaging effects upon general living standards and economic security; and pressure to reduce expenditure on welfare (in order to be competitive) … social problems are becoming increasingly globalized and … consequently social workers … must take account of events, institutions and policies beyond national and even regional borders."

The most effective way of achieving this goal is through engagement in forms of collective action which can help transform the meaning of everyday life and which can act as 'revealers' by exposing that which is hidden or excluded by the decision-making process. "Collective protest and mobilization bring to light those elements of silence, obscurity or arbitrariness buried within complex systems" (Melucci, 1989: p. 175). Craig (1998) argues that understanding the global context as it impacts on communities is important because, "throughout the world oppressed people have found a voice in part through the work of community development activists" (p. 15). Dominelli (2010: p. 172) makes a similar point when she argues that a focus on globalization can allow us to "bring marginalized areas of practice into the mainstream, particularly those linked to indigenous knowledges and new areas of endeavour."

Conclusion

Community work has a varied and complex history. As Mayo (2008) has suggested it has in many ways reflected the changing social, economic, and political conditions of the last century. She points to the tension within community work between those community workers:

> Who have preferred to fix their eyes on the gutters, rather than the stars, to concentrate on everyday issues rather than think strategically about how to effect change.
>
> Mayo, 2008: p. 16

This article has attempted to demonstrate how it is possible to fix your eyes both on the gutter and think strategically. It is precisely through this focus on action and theory that a meaningful community work can be sustained and contribute to resolving the wide range of social, economic, and political problems facing marginalized communities in the twenty-first century. As Henderson and Vercseg (2010) note:

> Community development can help civil society retain the dimension of challenge … the contribution that community development can bring are the experiences, learning and key messages of communities (p. 177).

The wide range of interventions described earlier indicate how this process of challenge can be maintained through the contributions that community work/development can make to these processes.

See also: Community Care in England; Community Resilience to Disasters; Green Social Work; Social Work Theory.

Bibliography

Alinsky, S.D., 1969. Reveille for Radicals. Vintage Books, New York.
Alinsky, S.D., 1972. Rules for Radicals. Vintage Books, New York.
Allman, P., 2001. Revolutionary Social Transformation: Democratic Hopes, Political Possibilities and Critical Education. Bergin and Garvey, USA.
Bauman, Z., 2001. Community: Seeking Safety in an Insecure World. Polity, Cambridge.
Betsill, M.M., Corell, E., 2001. NGO influence in international environmental negotiations: a framework for analysis. Global Environmental Politics 1 (4), 65–85.
Bolger, S., Corrigan, P., Dorking, J., Frost, N., 1981. Towards a Socialist Welfare Practice. Macmillan, London.
Breuggemann, W.G., 2006. The Practice of Macro Social Work, third ed. Thomson Higher Education, Belmont, CA.
Bunyan, P., 2010. Broad based organizing in the UK: reasserting the centrality of political activity in community development. Community Development Journal 45 (1), 111–127.

CDP (Community Development Project), 1974a. Interim Report. CDP Information and Intelligence Unit, London.
CDP (Community Development Project), 1974b. Inter Project Report. CDP Information and Intelligence Unit, London.
CDP (Community Development Project), 1977. Gilding the Ghetto, Inter-project Editorial Team. CDP Information and Intelligence Unit, London.
Craig, G., 1998. Community development in a global context. Community Development Journal 33 (1), 2–17.
Craig, G., Mayo, M., Popple, K., Shaw, M., Taylor, M., 2011. The Community Development Reader: History, Theories and Issues. Policy Press, Bristol.
Day, G., 2006. Community and Everyday Life. Routledge, London.
Delanty, G., Styrom, P. (Eds.), 2003. Philosophies of Social Science: The Classic and Contemporary Readings. Open University Press, Buckingham.
Dominelli, L., 1990. Women and Community Action. Venture Press, Birmingham.
Dominelli, L., 1998. Anti-oppressive practice in context. In: Adams, R., Dominelli, L., Payne, M. (Eds.), Social Work: Themes, Issues and Critical Debates. Macmillan, Basingstoke.
Dominelli, L., 2006. Women and Community Action. Policy Press, Bristol.
Dominelli, L., 2010. Social Work in a Globalizing World. Polity, Cambridge.
De Filippis, J., Fisher, R., Shragge, E., 2009. What's left in the community? Oppositional politics in contemporary practice. Community Development Journal 44 (1), 38–52.
Eriksson, L., 2011. Community development and social pedagogy: traditions for understanding mobilization for collective self-development. Community Development Journal 46 (4), 403–420.
Freire, P., 1972. Pedagogy of the Oppressed. Penguin, Harmondsworth.
Freire, P., 1993. Pedagogy of the City. Continuum Publishing Company, New York.
Gramsci, A., 1971. Selections from the Prison Notebooks. Lawrence and Wishart, London.
Gulbenkian Community Work Group, 1973. Current Issues in Community Work. Routledge and Kegan Paul, London.
Hatton, K., 2001. The Dialectics of Exclusion and Empowerment: An Examination of the Role of Social Professionals in Britain (unpublished thesis). University of Portsmouth, Ireland and Denmark.
Hatton, K., 2008. New Directions in Social Work Practice. Learning Matters, Exeter.
Hatton, K., 2011. Changing Professional Identities: towards a structural social pedagogy. In: Chytil, O., Friesenham, G.J., Lorenz, W., Seibel, F.W. (Eds.), European Developments and the Social Professions: Community, Education, Research. Professionalism. European Centre for Community Education, Kolenz.
Hatton, K., 2013. Social Pedagogy in the UK: Theory and Parctice. Russell House Publishing, Lyme Regis.
Henderson, P., Vercseg, I., 2010. Community Development and Civil Society: Making Connections in the European Context. Policy Press, Bristol.
Hill-Collins, P., 1990. Black Feminist Thought. Routledge, London.
Ife, J., Fiske, L., 2006. Human rights and community work: complementary theories and practices. International Social Work 49, 297–308.
Kamat, S., 2004. The privatization of public interest: theorising NGO discourse in a neoliberal era. Review of International Political Economy 11 (1), 155–176.
Ledwith, M., 1997. Participating in Transformation: Towards a Working Model of Community Empowerment. Venture Press, Birmingham.
Ledwith, M., 2009. Antonio Gramsci and feminism: the elusive nature of power. Education Philosophy and Theory 41 (6), 684–697.
London Edinburgh Weekend Return Group, 1980. In and Against the State.
Lone, A., 2010. Collectives – a Norwegian success story in residential care for drug addicts based on a social pedagogical approach. European Journal of Social Education 18/19, 60–72.
Lorde, A., 1984. Sister Outsider. Crossing Press, New York.
Lyons, K., Manion, K., Carlsen, M., 2006. International Perspectives on Social Work. Palgrave Macmillan, Basingstoke.
Macionis, J., Plummer, K., 2005. Sociology: A Global Introduction. Prentice Hall, Harlow.
Mayo, P., 1999. Gramsci, Freire and Adult Education. Zed Press, London.
Mayo, M., 2008. Community development, contestations, continuities and change. In: Craig, G., Popple, K., Shaw, M. (Eds.), Community Development in Theory and Practice: An International Reader. Spokesman Books, Nottingham.
McLaughlin, J., 1995. Travellers and Ireland, Undercurrents. Cork University Press, Cork.
Melucci, A., 1989. Nomads of the Present: Social Movements and Individual Needs in Contemporary Society. Vintage, New York.
Morgan, R. (Ed.), 1970. Sisterhood is Powerful. Random House, New York.
Ohri, A., Manning, B., Curno, P., 1982. Community Work and Racism. Routledge, London.
Perkins, D.F., Borden, L.M., Villarruel, F.A., 2001. Community youth development: a partnership for action. The School Community Journal 11 (2), 39–56.
Popple, K., 1995. Analysing Community Work. Open University Press, Buckingham.
Purcell, R., 2012. Community development and everyday life. Community Development Journal 47 (2), 268–281.
Shaw, M., 2003. Community Work: Policy, Politics and Practice, Working Papers in Social Sciences an Policy. University of Hull/University of Edinburgh, 200.
Shaw, M., Martin, I., 2000. Community work, citizenship and democracy: re-making the connections. Community Development Journal 35 (4), 401–404.
Stepney, P., Popple, K., 2008. Social Work and the Community: A Critical Context for Practice. Palgrave, Basingstoke.
Tonnies, F., 1955. Community and Association. Routledge and Kegan Paul, London.
Williams, R., 1976. Keywords. Fontana, Glasgow.

Community, Social Context of: The US Case

Rachel Garshick Kleit, The Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a replacement of the previous edition article by T.N. Clark, volume 4, pp. 2374–2378, © 2001, Elsevier Ltd.

Abstract

The social context of community encompasses varied geographies as well social relations embedded in a network of relationships. The context of social communities is not necessarily spatially bound, although the context may originate in particular localities; social networks, identity, and new technologies allow social communities to cross geographies to encompass the entire globe. Thus, the social context of community varies from communities where space defines the social community to those where affinity does. Varied contexts have differential impacts on individuals and their access to opportunity, social support, social capital, involvement in civic life, and engagement in social movements.

The social context of community encompasses varied geographies as well social relations embedded in a network of relationships. The context of social communities is not necessarily spatially bound, although the context may originate in particular localities; social networks, identity, and new technologies allow social communities to cross neighborhood and city boundaries to encompass the entire globe. Thus, the social context of community varies from communities where space defines the social community to those where affinity does. Community's social contexts can be found in local geographies (like neighborhoods and local governmental units), institutions (such as voluntary organizations and churches), in the spaces of everyday life (such as work, school, and home), through extended family relationships, and through national and global affiliations and concerns.

The Influence of Community Social Context

The social contexts in which community occurs are theorized to have varied impacts on the lives of individuals, ranging from being the source of social support to information about varied life opportunities to the context of social action. These resources flow through the social networks in which individuals are embedded, and that act as vital links to larger social systems. Networks affect the flow of resources and information to and from the individual (Lin, 2001); the relationships among individuals occur and vary in the context of this complexity. Additionally, the structure of social networks themselves can work to create a community social context that influences both helping relationships and economic advancement differentially.

The implication is that community social context can influence social capital – the flow of resources and information across people through social networks; relationships with others may facilitate action, provide information, or make connections between the microlevel and the macrolevel. Coleman (1988) outlines three types of social capital. First, he describes relationships of help, obligation, and trust. Second, he discusses social capital as the possibility for information that facilitates action. Third, norms also arise through social relations, providing a social context that can facilitate certain actions and constrain others. While social capital can variously provide social support or social leverage or be a link to social resources (Lin, 2001), the use of such ties varies with the social context in which they are embedded. As Alejandro Portes (1998) notes, ties that provide social support may facilitate social control but may not be the same ties that provide help in finding a job. Furthermore, ties that provide social support clash with those that enable social leverage, if local norms must be overturned to facilitate economic action.

Urban and Rural Contexts

In the face of rapid urbanization in the late nineteenth century, Tonnies suggested that social communities were different in urban and rural places. He characterized ties in rural areas as strong interdependent relations – *Gemeinschaft* – while in urban areas social ties are more distant – *Gesellschaft*. The notion was that people in urban areas are isolated from kin; ties are attenuated. Urban sociologists Wirth and Simmel suggested that there were fundamental differences in urban relationships that make them more distant, alienating, and instrumental than those in rural settings. Yet, the community studies of the mid-twentieth century demonstrated that neighborhood was a locale for social ties, especially in working class and poor areas. For lower income residents in urban areas, survival may depend on interactions with this core network, often comprising kin and friends who find themselves in similar situations.

In suburban environments in the United States, scholars have commented on the attenuation of community at the neighborhood level. In particular, the writers of the third quarter of the twentieth century viewed low-density suburbs as the negation of community and communal relationships. Sprawl has been viewed as weakening household connections to immediate neighbors because it reduces spontaneous social interactions. While those with lower incomes have more ties within their neighborhoods, as do women and married people, those who live in car-dependent neighborhoods tend to have fewer ties to neighbors.

Neighborhood as Community's Social Context and Access to Opportunity

These ideas about the limitations of communities based in neighborhoods for people living in poverty arise again in the late twentieth century, most famously highlighted in William Julius Wilson's (1987) *The Truly Disadvantaged*, which outlined the social and economic isolation of the urban poor in Chicago and other cities. The idea was that over time middle class families left inner city neighborhoods for the suburbs, leaving behind a neighborhood social context devoid of social ties and economic linkages that could provide access to social and economic opportunity. Over time, a large literature grew to investigate neighborhood effects, focusing on the isolation of these inner city neighborhoods and the neighborhood effects on the life chances of individuals (Jencks and Mayer, 1990).

The notion that community social context can influence social and economic opportunities has become the focus of affordable rental housing policy in the United States (Joseph et al., 2007). The idea is that community of residence influences access to beneficial social networks, social control at the community level, behaviors leading to work and responsibility, and working institutions. Of these, only the latter has both theoretical and empirical support.

Community of residence influences the type of nonkin that people associate with but not their number, and those who are involved with neighbors are those with few sources for relationships outside the neighborhoods. As Fischer (1982) comments, "Those who are most involved with neighbors tended to be those with few sources of extra-neighborhood (nonkin) ties elsewhere," meaning those not working or working part-time, those new to the city, and those committed to home by marriage or home ownership.

Along with proximity, homogeneity is a key condition for affiliation. Michelson (1976) suggests that spatial proximity is only effective in encouraging interaction when there are conditions of real or perceived homogeneity and where there is a need for mutual aid, like small house repairs. Homogeneity – as far as stage in the life cycle, home ownership (and thus proprietary interests), life styles, and values – is the basis for social relationships, and may be more important than propinquity. McPherson et al. (2001) comment that the most basic source of homophily is space – we are more likely to have contact with those who are closer in geographic location than those who are distant. Thus, neighborhood can be an important social context due to both propinquity and perceived homogeneity.

Social Networks and Economic Action within Communities

Generally, the more diverse the types of people to whom an individual has access, the more different types of information available to that person, and the more likely he or she will experience better outcomes when looking for a job. Yet, the capacity for varied and valuable information varies within communities and the ability to use such information also varies across demographic groups. Generally, the greater the range of an individual's social network – the more diversity, the larger it is, the more it contains people who are not close to the individual, and the less everyone in the network knows everyone else – the more access to different societal circles that social network may be able to provide. Since the social context that arises through social network relationships tends toward homogeneity, such a breadth of contact with different social spheres creates a more varied community social context for that individual and therefore is helpful for gathering information. Granovetter's (1995) concept of the strength of weak ties suggests that the ties that provide access to a more varied social context are that weak ties – those that are short in duration, not emotionally close, not mutually confiding, infrequent, and not necessarily engaged in reciprocal services – reach out of homogenous cliques to allow for diversity of information and contacts. Yet, in some social groups, finding jobs through contacts may be the best option, but the jobs will be of poor quality because those are all the group has to offer. Working-class ties may lead to jobs, but jobs from same-group contacts may lead to jobs that are no better than anyone else's in the group.

While Granovetter's hypothesis about the nature of helpful social resources – the strength of weak ties hypothesis – provoked a generation of research on job matching, unmeasured social and human capital and the information available to the searcher may be more important than weak ties. Furthermore, some argue the very social structure of personal communities is part of the matching process that influences posthire outcomes, and that network structure alone, and not the type of tie used, is important. That is, the context of the social community delineated by such ties may be influenced not only by the type of information available but the very structure of that community, although the relative importance may vary by race, gender, and immigrant status.

While whites in urban environments may tend to become more involved with subcultures of interest, research suggests that blacks do not appear to be involved in minority subcultures. Urban blacks have social networks that are more racially diverse and less dense than those of rural blacks. Furthermore, while ethnic relations may enable job finding, support, and opportunities, they may also constrain them if the ethnic ties are not connected to a realm of relations with useful job information. For African-Americans, the level of racial diversity within their networks may be an important factor in having a good job as the outcome of a job search. Furthermore, personal contacts may facilitate transitions to employment, and African-Americans may face more barriers to employment when using their personal tie. For blacks, the degree to which people in the network fulfill multiple roles – such as concurrently being neighbor, coworker, and kin – is associated with finding lower paying jobs than other blacks, while the opposite was true for whites. The information available within the network is what is important, not simply the structure (Green et al., 1999; Coverdill, 1998). The racial diversity of the network is a proxy for the quality of the information available. Even if ethnic networks contained quality information, it is not enough for quality information to be present in a network, but other conditions must also exist that are part of positive social capital, namely, trust, mutual obligations, and social norms functioning through social networks (Smith, 2003). That is, the collective efficacy of the community is the context for the use of social ties in job finding; without it, individuals have no motive to help each other.

For women, spatial and social constraints converge to create a social context that does not always support economic advancement. Socially, as Jeanne Hurbert's work shows, gender homophilous contacts can have a significant negative effect on the socioeconomic status of the job sought. Furthermore, women who have diverse and large networks are more likely to be working for pay, and network bridges to those of better education and people outside the neighborhood may enable women to use their social networks more effectively to find a job. These effects must also be considered in concert with space–time constraints due to household responsibilities – women face a limitation on the range of employment opportunities open to them as compared to those available to men. Very low-income women may be much more likely to experience a spatial constraint in the job search because of their dependence on social ties, direct application, and falling into opportunities without a search. Furthermore, African-American women and their social networks may be more spatially bounded. The spatial boundedness of poor women may lead to increased opportunities for social support while reinforcing economic constraints. Thus, the community bounded both by space and structure – tie homophily – may be limiting for economic gain.

Migrants' social networks often center on other migrants and expand outward (Massey et al., 1987). For some immigrants, these social relationships may be limiting, but for others, network characteristics and the individual's personal position within them may allow them to mobilize resources. The social community in ethnic enclaves, as Alejandro Portes (1998) outlines, may provide vital resources for firms – providing not only start-up capital but also tips about business opportunities. However, Portes notes that impoverished communities with fewer of the characteristics of thriving ethnic enclaves also tend to lack these types of social connections. The circumscribed nature of personal social ties within the enclave may cut off information.

Hagen (1998) and others note that these social network resources vary for different people or groups in ways that impact opportunities. In Hagen's study of Mayan immigrants in Houston, ethnic ties allowed both men and women to work, but in entirely different fields that transformed social networks in distinct ways. Men's social worlds were able to support bids for legalization, while women's social worlds, as domestic help, did not. This is a repeated pattern in the migration literature.

Social Context, Intergroup Contact, and Tolerance

Neighborhood as the social context for community also may have a role in bringing people together who might not otherwise meet each other. Gordon Allport's (1979) contact theory suggests that people of different backgrounds will be more likely to interact if the groups have equal status when interacting within a given situation, are able to work toward common goals, have intergroup cooperation, and have support from authority for positive contacts. That is, just living in the same neighborhood or next to someone is not enough, but rather equal status, supported, and goal-oriented contacts are vital to positive changes in attitudes. Furthermore, Thomas Pettigrew (1998) suggests a fifth condition – that the contact situation must allow participants the opportunity to become friends. Thus, while neighborhood may be one context for social community, such relationships only occur in the appropriate circumstance, such as through community building activities.

Communities of Affiliation and Social Support

Communities of affiliation arise within spatial communities – such as neighborhoods – or through membership, the perception of a common cause, or the perception of homogeneity. Fischer's (1982) subcultural theory of urbanism encompasses these ideas, and suggests that people form social ties based upon many other attributes than kin and locality. Relationships are formed based upon ethnicity, occupation, lifestyle, and other common traits. Within these communities, people find "fellowship, guidance, and meaning…" (Fischer, 1982). Barry Wellman's (Wellman, 1979; Wellman and Leighton, 1979) work delineates three perspectives on the social context within urban neighborhoods: the community lost perspective of early urban sociologists, the community saved perspective presented in community studies, and, lastly, his suggestion of a community liberated – where kinship ties are important, but not primary, and people create 'personal community networks', a set of active community ties. These relationships are socially diverse, spatially dispersed, sparsely knit, and not spatially bound by the neighborhood. In urban areas people are less involved with neighbors than are those in small towns. People in cities turn to neighbors only if they find them compatible in some way (Fischer, 1982).

Thus, social support also can arise out of community social context, most frequently associated with dense, strong, and homophilous ties rather than weak and heterophilous ties. Informal support arises from core, homogenous networks. Increased network density, strong ties, homogeneity combined with a sense of belonging, and a high degree of normative consensus associated with these high levels of social integration help to facilitate a routinization of support among people in the core network.

Furthermore, where weaker ties may provide information, a relatively small number of emotionally strong ties likely provide broader support in terms of emotional aid, services, and companionship. While physical access – like being a neighbor – can promote both large and small services among people, physical access is not related to receiving social support in the form of companionship, emotional aid, and financial aid because these can actually be offered over large distances. Interactions in group settings do not foster social support, and the involvement of kin in a social support system is based on the aid of a relatively small number of immediate kin. Women tend to provide more emotional support than men.

Yet, the network is more than the sum of the one-on-one relationships, because the network as a whole can exert group pressure to provide support. Furthermore, such network capital functions on many different levels, based on the individuals social status, the size of the network, the similarity of people within the network, the availability of resources within the network and their ability to be delivered, the history of support and reciprocity, the composition of the network, the structure of information flows through the

network, and indirect ties. Different levels of network are nested within others.

The extent to which the social context of a neighborhood matters for social support depends on the type of aid needed (Wireman, 1984). People tend to use friends when they need to agree about tastes or if longer term support is needed. Neighbors are best at everyday contact that does not require a long-term commitment. Research has found that those who are economically advantaged have larger networks because they are more integrated into the larger society. As such, they do not need strong links to neighbors for day-to-day support. Those of lower socioeconomic status, however, because of a relative lack of social integration and more tenuous financial status need nearby support and cultivate it within their neighborhoods. In short, those who are poorer, more tied to their neighborhoods, and are less mobile tend to be more tied to their neighborhoods for their day to day social support needs.

Community Institutions as Social Context

Alexis de Tocquville's admiration of the young United States' democracy and consistent involvement in community associations contrasted with the 1830's French state-dependency of its citizens. Local civic groups, linked to small autonomous governments were the lifeblood of the young American Democracy. Over a 150 years later, Robert Putnam (2000) outlines the decline of US civil society and community trust, suggesting a decline in community-level social capital – a decline in joining these civic organizations. Berry et al. (1993) found that citizen participation can be dismissed as troublesome; at the same time, voluntary organizations can make outstanding achievements, despite these attitudes of civic officials. The collectivity that arises within localities or voluntary institutions can be a source of political voice and action.

Citizenship, Nations, and Globalization

Citizenship confers more than membership in a nation – it confers rights. T.H. Marshall suggested that citizenship is a core concept linking citizens to the nation. Every citizen of a nation has rights that range from minimal to extensive and included the rise of egalitarianism. These rights are insured by a national welfare state, government agencies, and the justice system. However, national centralization varies across nations, as does the egalitarian nature of citizenship. And over time, social upheavals among European Eastern Bloc countries led to a movement away from centralized control and radical egalitarianism.

The Western notion of a social contract preserves notions of egalitarianism in the obligations that citizens of a state have to each other. Social connection, Iris Marion Young argues, is prior to the institutions of the state. In a Lockean view of the social contract, the need and desire for social institutions arise from socially connected individuals who recognize that their relationships may conflict. The obligation of justice is generated by the social connections, and political institutions are responsible for those obligations. Membership within a nation-state thus involves a set of rights and responsibilities.

In a period of globalization of economies and international migration, political upheavals leave people as refugees without a membership in a nation. In contrast to John Rawls' nation-state-based conception of the social contract, Seyla Benhabib argues that the world community should be viewed as a global civil society. People are organized into nation-states, and such states are major players – but not the only players. Those who are without national status – refugees in this globalized world – still have basic rights accrued to them as humans. Thus, the claim to membership is not limited only to those with a national identity.

The Networked World

Some feared that the growth of the Internet would bring with it a loss of personal relationships and a loss of local community. Barry Wellman and Caroline Haythornthwaite (2002) argue that virtual communities extend the social context for community. The Internet, rather than being separate, facilitates all sorts of community ties. It enhances the resources for social support. Rather than reducing interactions, it adds on to other interactions, increasing overall contact within a personal community of friends and relatives. Formal social groups also experience increased communication, with heavy Internet use being associated with voluntary organizations and politics. These benefits, however, do accrue to those already well situated and well connected in society, building upon preexisting relationships.

Civic engagement is also enhanced through such online communications. In the planning field, the context for community participation in local decision making extends beyond the local public hearing to Facebook pages, Twitter feeds, and crowd-sourcing citizen perspectives. Crowd sourcing is online problem solving by a large network of people. In local government terms, this allows a broader and sometimes deeper reach into a community for their perspective on a policy issues. The context of community has indeed become integrated into the networked world – consider the flow of Twitter feeds that helped Egypt's 2011 Arab Spring occur. The spread of communications over smart phones enabled the mobilization of resources throughout a protest community leading to a revolution.

See also: Citizen Participation; Economic Sociology; Ethnic, Racial, and Nationalist Movements; Migration and Social Capital; Nation, Sociological Aspects of; Neighborhood Effects; Organizations, Sociology of; Political Sociology; Prejudice in Society: Psychological Perspectives; Racial Relations; Social Network Analysis; Subculture, Sociology of; Urban Poverty in Neighborhoods; Urban Sociology; Voluntary Associations, Sociology of.

Bibliography

Allport, G.J., 1979. The Nature of Prejudice. Addison-Wesley, Reading, MA.

Berry, J.M., Portney, K., Thomson, K., 1993. The Rebirth of American Democracy. Brookings Insitution Press, Washington, DC.

Coleman, J.S., 1988. Social capital in the creation of human capital. American Journal of Sociology 94, S95–S120.

Coverdill, J.E., 1998. Personal contacts and post-hire job outcomes: theoretical and empirical notes on the significance of matching methods. In: Leicht, K.T. (Ed.), Research in Social Stratification and Mobility. JAI Press, Inc, Stamford.
Fischer, C.S., 1982. To Dwell Among Friends. University of Chicago Press, Chicago.
Granovetter, M.S., 1995. Getting a Job: A Study of Contacts and Careers. The University Of Chicago Press, Chicago.
Green, G.P., Tigges, L.M., Diaz, D., 1999. Racial and ethnic differences in job-search strategies in Atlanta, Boston, and Los Angeles. Social Science Quarterly 80, 263–278.
Hagen, J.M., 1998. Social networks, gender, and immigrant incorporation: opportunity and constraint. American Sociological Review 63, 55–67.
Haythornthwaite, C., Wellman, B. (Eds.), 2002. The Internet in Everyday Life. Blackwell Publishers, Malden, MA.
Hurlbert, J.S., Beggs, J.J., Haines, V.A., 2001. Social networks and social capital in extreme environments. In: Lin, N., Cook, K., Burt, R.S. (Eds.), Social Capital: Theory and Research. Aldine De Gruyter, New York.
Hurlbert, J.S., Haines, V.A., Beggs, J.J., 2000. Core networks and tie activation: what kinds of routine networks allocate resources in nonroutine situations? American Sociological Review 65, 598.
Jencks, C., Mayer, S.E., 1990. The social consequences of growing up in a poor neighborhood. In: Lynn Jr., L., Mcgeary, M.G.H. (Eds.), Inner-city Poverty in the United States. National Academy Press, Washington, DC.
Joseph, M., Chaskin, R., Webber, H., 2007. The theoretical basis for addressing poverty through mixed-income development. Urban Affairs Review 42, 369–409.
Lin, N., 2001. Social Capital: A Theory of Social Structure and Action. Cambridge University Press, Cambridge.
Massey, D.S., Alarcon, R., Durand, J., Gonzalez, H., 1987. Return to Aztlan: The Social Processes of International Migration from Western Mexico. University of California Press, Berkeley, CA.
McPherson, M., Smith-lovin, L., Cook, J.M., 2001. Birds of a feather: homophily in social networks. Annual Review of Sociology 27, 415–444.
Michelson, W., 1976. Man and His Urban Environment: A Sociological Approach, with Revisions. Addison-Wesley, Reading, MA.
Pettigrew, T.F., 1998. Intergroup contact theory. Annual Review of Psychology 49, 65–85.
Portes, A., 1998. Social capital: its origins and applications in contemporary sociology. Annual Review of Sociology 24, 1–23.
Putnam, R.D., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon and Schuster, Inc, New York.
Smith, S.S., 2003. Exploring the efficacy of African-Americans job referral networks: a study of the obligations of exchange around job information and influence. Ethnic and Racial Studies 26, 1029–1045.
Wellman, B., 1979. The community question: the intimate networks of East Yorkers. American Journal of Sociology 85, 1201–1231.
Wellman, B., Leighton, B., 1979. Networks, neighborhoods and communities. Urban Affairs Quarterly 14, 363–390.
Wilson, W.J., 1987. The Truly Disadvantaged. University of Chicago Press, Chicago.
Wireman, P., 1984. Urban Neighborhoods, Networks, and Families: New Forms for Old Values. Heath, Lexington, MA.

Relevant Websites

http://www.pewinternet.org/Reports/2004/The-Internet-and-Daily-Life.aspx – Pew Internet and American Life Project. 2004.
http://www.movements.org/.

Co-morbidity of Mental and Physical Conditions

Kate M Scott, University of Otago, Dunedin, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mental disorders including mood, anxiety, substance use, and psychotic disorders have been associated with a wide range of physical conditions in both clinical and general population studies. These mental–physical comorbidities can come about either through causal or noncausal mechanisms (shared determinants). Causal mechanisms include a variety of behavioral, biological, psychological, pharmacological, and service delivery pathways through which mental disorders could increase risk of subsequent physical conditions, or physical conditions could increase risk of mental disorders. Shared determinants may include environmental factors, such as childhood adversities, or genetic risk factors. The consequences of mental–physical comorbidity are severe and wide ranging.

Epidemiology of Mental–Physical Comorbidity

Introduction

Mental–physical comorbidity refers to the statistical association of a mental disorder with some kind of medical disease or chronic physical condition, either at a given point of time or within the lifetime of an individual. This association indicates that the two types of disorder (mental and physical) are co-occurring to a degree that is greater than would be expected by chance. Many studies measure mental–physical comorbidity in cross-sectional designs, detecting associations between mental disorder and physical condition prevalences within the same time period (such as the past 1–12 months). This approach documents the association but leaves it unclear whether one disorder influenced the onset of the other in a causal manner (and if so, which came first), or whether both developed at the same time, possibly caused by some other causal factor or factors common to both of them. Prospective studies examine sequential or longitudinal comorbidity, documenting a specific temporal sequence in the association of a disorder, such as depression, with the subsequent incidence of another, such as heart disease, years or decades later. At the population level, the burden of mental–physical comorbidity is considerable. At any age, the majority of those with a mental disorder experience some kind of comorbid chronic physical condition (Scott et al., 2008). In the following sections, the types of medical comorbidities that tend to co-occur with mental disorders are described, the various mechanisms underlying mental–physical comorbidity are discussed, and finally the impact of mental comorbidity is outlined.

Mental–Physical Comorbidities

Numerous studies, both cross-sectional and prospective, have established that depression is associated with a wide range of medical comorbidities such as diabetes, cardiovascular diseases, respiratory diseases, and a range of chronic pain conditions (Evans et al., 2005; Prince et al., 2007). Much of the research in this area involved clinical or primary care samples, but recent data from the World Mental Health surveys have confirmed these associations in general populations around the world (Scott et al., 2007).

Although more attention has been focused on depression, cross-sectional general population studies suggest that anxiety disorders show a similar pattern of concurrent comorbidity with a wide range of physical conditions (Von Korff et al., 2009). In prospective studies, anxiety is often studied with symptom screening scales that do not allow determination of which anxiety disorders might be implicated in which physical conditions. More research is required using clinical diagnoses of anxiety disorders to get a clearer picture of whether some anxiety disorders have specific links with particular physical conditions. The available research indicates that post-traumatic stress disorder (PTSD) is associated with a number of physical comorbidities, and that PTSD and panic disorder may have specific links with coronary heart disease. Panic disorder has also been linked in several studies to respiratory diseases such as asthma or chronic obstructive pulmonary disease (Vögele and Von Leupoldt, 2008).

Alcohol and other drugs have systemic biological effects (Brick, 2004). When these drugs are used with the frequency and duration that characterize substance use disorders, the medical consequences can be profound, ranging from the slow and insidious to acute and life threatening, depending on the drug in question and the dose.

Bipolar disorder and schizophrenia are associated with increased prevalence of a similar range of illnesses, in particular cardiovascular diseases, metabolic diseases, thyroid disease, respiratory disease, hepatitis C, and osteoporosis (Kilbourne et al., 2004; Lambert et al., 2003).

How Mental and Physical Disorders Come to Be Comorbid

There are many mechanisms that have been discussed that could contribute to the overlap of mental and physical conditions. These can be conceptually organized into two overarching models (Scott, 2009; Steptoe, 2007b). The first of these is the causal model. In this model, one condition increases the risk of the other, operating through mechanisms that lead from mental disorder to physical condition (Section From Mental Disorder to Physical Disease), or from physical condition to mental disorder (Section From Physical Disease to Mental

 http://dx.doi.org/10.1016/B978-0-08-097086-8.14147-9

Disorder). The other overarching model is the noncausal or indirect model. This refers to the possibility that there are shared determinants that increase the risk of both mental and physical disorders (Section Shared Determinants). In this model, the comorbidity comes about through these underlying shared risk factors, rather than through either the mental disorder causing the physical condition or vice versa.

From Mental Disorder to Physical Disease

There is now a great deal of evidence from more than 30 prospective studies that depression is associated with subsequent coronary heart disease incidence or occurrence of fatal or nonfatal heart attack (Goldston and Baillie, 2008). Anxiety, also, has been linked prospectively with subsequent heart disease and heart attacks. Several studies have also documented prospective associations of depression and/or anxiety with the onset of type 2 diabetes, stroke, peptic ulcer, chronic pain, and other physical conditions. Research has also shown that most of the medical comorbidities of bipolar disorder and schizophrenia arise following the mental disorder onset.

There have been many biological mechanisms proposed to account for these associations. A number of these mechanisms relate to the functioning of the physiological systems for responding to stress: the autonomic nervous system (through the sympathetic–adrenal–medullary (SAM) axis) and the neuroendocrine system (through the hypothalamic–pituitary–adrenocortical (HPA) axis). Dysregulation of both of these systems has been associated with depression and anxiety disorders (Stetler and Miller, 2011; Young et al., 2004). Dysregulation of the HPA and SAM axes has a range of adverse metabolic, cardiovascular, immune, and neurological effects. These include elevated blood glucose levels, insulin resistance and obesity, enhanced blood clotting, increased blood pressure and reduced heart rate variability, immunosuppression and increased inflammatory response, and a cascade of other effects that contribute to atherosclerosis (Miller et al., 2002; Chrousos and Kino, 2007). So, one main pathway through which depression and anxiety could be affecting subsequent physical health is through the chronic disturbances in stress physiology associated with these mental disorders.

Mental disorders can also affect physical disease risks through behavioral pathways. All mental disorders have a well-established relationship with behavioral risk factors for chronic diseases such as smoking, poor diet, sedentary behavior, and excessive alcohol consumption (Prince et al., 2007). Mental disorders also increase the risk of later onset of physical disease by affecting adherence to treatment of risk factors for specific diseases, whether that treatment involves behavior modification (e.g., smoking cessation, weight reduction, exercise, dietary change) or adherence to medication (e.g., cholesterol-lowering medication). Substance use disorders as well as depression can lead to reduced likelihood of adopting disease prevention measures, increasing risk of contracting as well as transmitting sexually transmitted and other infectious diseases.

An important contribution to the increased risk of physical comorbidities among people with mental disorders is the tendency for physical conditions to be less likely to be detected and treated in people with mental health problems. This may result from a mixture of factors, including the separate delivery of mental from other medical services in most countries. This is not optimal for the recognition and treatment of mental–physical comorbidity. Research among mental health professionals has identified barriers such as lack of explicit allocation of responsibility for medical treatment, lack of service delivery integration, and pessimistic attitudes among treatment providers as to whether improved physical health is possible, or a priority, among those with mental illness (Hyland et al., 2003). There are also treatment barriers that result from the mental disorder itself. The specific phobias involving medical fears, for example, can interfere with dental and medical treatment seeking and so elevate the risk of poor health subsequently. Another example is the cognitive impairment associated with schizophrenia that can interfere with the detection of physical health problems and communication of these to health care providers.

Finally, psychotropic medications underpin some medical comorbidities. This is particularly the case for bipolar and psychotic disorders where the antipsychotic medications prescribed for these disorders can dramatically increase the risk of obesity and metabolic syndrome. These medications can also increase risk for other medical diseases including osteoporosis and cardiovascular disease (Lambert et al., 2003).

From Physical Disease to Mental Disorder

It is intuitive that having a physical disease or chronic condition could, under some circumstances, become depressing or anxiety provoking, and hence increase the likelihood of the development of a comorbid mental disorder. Chronic physical conditions are associated with many difficulties such as disability and functional impairments, chronic pain, disfigurement, stressful medical procedures, challenging lifestyle changes, changes to social networks and intimate relationships, increasing dependency, and sometimes the prospect of a foreshortened future. Many of these difficult consequences of physical illness have been linked to the development of depression in prospective studies (Ormel et al., 1997; Steptoe, 2007a). This is not to suggest that the majority of people with chronic physical conditions develop depression or mental disorders – in fact only a minority do. Whether someone with a physical condition develops a comorbid mental disorder is related not only to the severity and consequences of the physical condition, but also to the individual's beliefs about the illness, its significance in their lives, and their perceived capacity to cope with it (Sharpe and Curran, 2006).

Physical diseases can also result in mental health symptoms or mental disorders mediated by biological disease processes. One example of this is poststroke depression, the likelihood of which has been found to vary depending on stroke laterality and location, suggesting the involvement of neuroanatomical and neurobiological processes (Gaete and Bogousslavsky, 2008). The inflammatory response in the brain following ischemic injury, for example, may lead to serotonin depletion in parts of the brain involved in emotional regulation. These biological changes following stroke may work alongside or interact with the psychological losses and disability associated with stroke to increase depression risk. A related area of research is on vascular depression as a specific subtype of depression in the elderly resulting from microvascular lesions

in the brain (Baldwin, 2005). The depression following Parkinson's disease may also be a biological consequence of underlying neurodegeneration; although as with stroke, the psychological effects of Parkinson's may be a contributing factor. Diseases such as cancer and HIV/AIDS are further examples of where the episodes of depression that frequently follow the disease diagnosis may result from an intertwining of biological and psychological factors.

A third pathway from physical disease to mental disorder is via medications prescribed for the physical condition. A number of cardiovascular medications have been suggested to increase depression risk (Shah et al., 2011; Licht et al., 2010), although this is methodologically challenging to determine and further research is needed. Treatment with interferon, which is used in a variety of diseases including some cancers, multiple sclerosis, and hepatitis C, is widely reputed to increase the risk of depression. Several other medications have also been implicated in depression and also other mental health problems. Efavirenz, for example, a commonly used nonnucleoside reverse transcriptase inhibitor frequently prescribed in combination with other antiretroviral regimens (e.g., protease inhibitors) for the treatment of HIV has been associated with behaviors such as suicidal ideation, aggressive/impulsive behavior, paranoid reactions, and manic reactions (Sütterlin et al., 2010) Nevertheless, reviews of this area have concluded that medically induced depressions are symptomatically different from typical major depressive disorder (Patten and Barbui, 2004).

Barriers also exist in detecting and treating mental disorders among those with physical conditions (Evans et al., 2005). Although this does not cause the comorbidity, it prolongs it, increasing the prevalence of mental–physical comorbidity observed in research studies (see also Section Consequences and Implications of Mental–Physical Comorbidity).

Shared Determinants

The shared determinants model refers to the fact that mental disorders and chronic physical conditions have some risk factors in common. This raises the possibility that their comorbidity could be a function of this shared etiology rather than a causal connection. One such shared etiological factor is poor fetal growth (manifesting as low birth weight for gestational age). This has been found to predict adult-onset cardiovascular disease, metabolic syndrome and type 2 diabetes, obesity, hypertension, and osteoporosis (Barker et al., 2002), as well as adolescent- and adult-onset depression (Thompson et al., 2001). One of the hypothesized mechanisms for these connections has been the influence of gestational stress on HPA axis settings.

Animal studies (Sanchez, 2006) and clinical studies in humans (Heim et al., 2000) indicate that reactivity to stress can be heightened not only prenatally, but also in early life by adverse maternal care or other early life adversities. As well as this laboratory research, there are also epidemiological studies linking childhood adversities such as abuse and neglect with later mental (Gilbert et al., 2009) and physical ill-health (Wegman and Stetler, 2009). This body of research has focused attention on childhood adversities as possible shared determinants of mental–physical comorbidity (Scott, 2009). Preliminary exploration of this possibility using retrospective data from the World Mental Health Surveys found that childhood adversities explained some of the association of mental disorders with subsequent onset of physical conditions (Scott et al., 2011), but not all. Some mental–physical comorbidities could have other environmental and/or genetic factors in common.

Consequences and Implications of Mental–Physical Comorbidity

The most severe consequence of mental–physical comorbidity is premature mortality. People with mental disorders, especially depression and schizophrenia, have increased mortality when compared with people without those conditions, even after statistically controlling for the effect of suicide (Harris and Barraclough, 1998) (Prince et al., 2007). Research also shows that comorbid mental disorders are associated with increased morbidity among people with a range of chronic physical conditions including arthritis, coronary heart disease, stroke, and diabetes (Von Korff, 2009). There have been many studies elucidating the reasons behind this effect of comorbid depression on worsening the prognosis of physical disease. Co-occurring mental disorders such as depression and anxiety increase the numbers of physical symptoms experienced in diseases such as diabetes and asthma, even after adjusting for medical severity. They are also known to amplify the experience of physical symptoms, whether medically explained or not. Mental disorders prolong and worsen the clinical course of physical conditions, in part by interfering with self-management of chronic conditions and treatment adherence. As one example, many studies have found that both depression and anxiety lead to poorer glycemic control in diabetes and reduced likelihood of maintaining dietary and physical activity changes. Similar findings have been observed for cardiovascular diseases. Quality of life is therefore greatly affected by co-occurring mental disorders, and functional disability is also amplified.

This body of research highlighting the deleterious effects of mental disorders on medical disease and chronic physical condition outcomes, together with the evidence of higher health care costs associated with comorbid depression, has motivated intervention trials of depression treatment in medical patients. Unfortunately the trials conducted to date in heart disease and diabetes have found few, if any, positive effects on medical outcomes (The ENRICHD Investigators, 2003; Katon et al., 2004). There could be a number of reasons for this, but one of them may be that the depression treatment in these trials is undertaken too late in the life course. Most people who have a depressive episode later in life (e.g., comorbid with a physical condition) have had a prior episode of depression, anxiety, or some other mental disorder earlier in life. As discussed above, depression and other mental disorders may be influencing physical health through biological, behavioral, and pharmacological pathways for many years prior to the clinical symptoms of the medical disease becoming manifest. This suggests that more rigorous and systematic detection and treatment of mental disorders in adolescence and early adulthood, coupled with careful monitoring of physical health in those with chronic mental disorders, has the potential

to greatly enhance mental and physical health throughout the life course.

See also: Cancer: Psychosocial Aspects; Cardiac Disease, Coping with; Chronic Illness, Psychosocial Coping with; Coronary Heart Disease: Psychosocial Aspects; Depression, Pessimism, and Health; Diabetes: Psychosocial Aspects; Epigenetics; Health and Illness: Mental Representations in Different Cultures; Mental Health; Patient Adherence to Health Care Regimens; Personality and Health; Personalized Medicine; Respiratory Disorders: Psychosocial Aspects; Stress, Coping and Health.

Bibliography

Baldwin, R.C., 2005. Is vascular depression a distinct sub-type of depressive disorder? A review of causal evidence. International Journal of Geriatric Psychiatry 20, 1–11.

Barker, D.J.P., Eriksson, J.G., Forsen, T., Osmond, C., 2002. Fetal origins of adult disease: strength of effects and biological basis. International Journal of Epidemiology 31, 1235–1239.

Brick, J., 2004. Handbook of the Medical Consequences of Alcohol and Drug Abuse. Haworth Press, New York.

Chrousos, G.P., Kino, T., 2007. Glucocorticoid action networks and complex psychiatric and/or somatic disorders. Stress 10, 213–219.

Evans, D.L., Charney, D.S., Lewis, L., Golden, J.M., Ranga Rama Krishnan, K., Nemeroff, C.B., Bremner, J.D., Carney, R.M., Coyne, J.C., Delong, M.R., Frasure-Smith, N., Glassman, A.H., Gold, P.W., Grant, I., Gwyther, L., Ironson, G., Johnson, R.L., Kanner, A.M., Katon, W.J., Kaufmann, P.G., Keefe, F.J., Ketter, T., Laughren, T.P., Leserman, J., Lyketsos, C.G., Mcdonald, W.M., Mcewan, B.S., Miller, A.H., Musselman, D., O'connor, C., Petitto, J.M., Pollock, B.G., Robinson, R.G., Roose, S.P., Rowland, J., Sheline, Y., Sheps, D.S., Simon, G., Spiegel, D., Stunkard, A., Sunderland, T., Tibbits, P., Valvo, W.J., 2005. Mood disorders in the medically ill: scientific review and recommendations. Biological Psychiatry 58, 175–189.

Gaete, J.M., Bogousslavsky, J., 2008. Post-stroke depression. Expert Review of Neurotherapeutics 8, 75–92.

Gilbert, R., Widom, C.S., Browne, K., Fergusson, D.M., Webb, E., Janson, S., 2009. Child maltreatment I. Burden and consequences of child maltreatment in high-income countries. The Lancet 373, 68–81.

Goldston, K., Baillie, A.J., 2008. Depression and coronary heart disease: a review of the epidemiological evidence, explanatory mechanisms and management approaches. Clinical Psychology Review 28, 289–307.

Harris, E.C., Barraclough, B., 1998. Excess mortality of mental disorder. British Journal of Psychiatry 173, 11–53.

Heim, C., Newport, D.J., Heit, S., Graham, Y.P., Wilcox, M., Bonsall, R., Miller, A.H., Nemeroff, C.B., 2000. Pituitary-adrenal and autonomic responses to stress in women after sexual and physical abuse in childhood. Journal of the American Medical Association 284, 592–597.

Hyland, B., Judd, F., Davidson, S., Jolley, D., Hocking, B., 2003. Case managers' attitudes to the physical health of their patients. Australian and New Zealand Journal of Psychiatry 37, 710–714.

Katon, W.J., Von Korff, M., Lin, E.H., Simon, G., Ludman, E., Russo, J., Ciechanowski, P., Walker, E., Bush, T., 2004. The pathways study: a randomized trial of collaborative care in patients with diabetes and depression. Archives of General Psychiatry 61, 1042.

Kilbourne, A.M., Cornelius, J.R., Han, X., Pincus, H.A., Shad, M., Salloum, I., Conigliaro, J., Haas, G.L., 2004. Burden of general medical conditions among individuals with bipolar disorder. Bipolar Disorder 6, 368–373.

Lambert, T.J., Velakoulis, D., Pantelis, C., 2003. Medical comorbidity in schizophrenia. Medical Journal of Australia 178, S67.

Licht, C.M.M., De Geus, E.J.C., Van Dyck, R., Penninx, B.W.J.H., 2010. Longitudinal evidence for unfavorable effects of antidepressants on heart rate variability. Biological Psychiatry 68, 861–868.

Miller, G.E., Cohen, S., Ritchey, A.K., 2002. Chronic psychological stress and the regulation of pro-inflammatory cytokines: a glucocorticoid resistance model. Health Psychology 21, 536–541.

Ormel, J., Kempen, G.I.J.M., Penninx, B.W., Brilman, E.I., Beekman, A.T.F., Van Sonderen, E., 1997. Chronic medical conditions and mental health in older people: disability and psychosocial resources mediate specific mental health effects. Psychological Medicine 27, 1065–1077.

Patten, S.B., Barbui, C., 2004. Drug-induced depression: a systematic review to inform clinical practice. Psychotherapy and Psychosomatics 73, 207–215.

Prince, M., Patel, V., Saxena, S., Maj, M., Maselko, J., Phillips, M.R., Rahman, A., 2007. Global mental health 1: no health without mental health. The Lancet 370, 859–877.

Sanchez, M.M., 2006. The impact of early adverse care on HPA axis development: nonhuman primate models. Hormones and Behaviour 50, 623–631.

Scott, K.M., 2009. The development of mental-physical comorbidity. In: Von Korff, M., Scott, K.M., Gureje, O. (Eds.), Global Perspectives on Mental-Physical Comorbidity in the WHO World Mental Health Surveys. Cambridge University Press, New York.

Scott, K.M., Bruffaerts, R., Tsang, A., Ormel, J., Alonso, J., Angermeyer, M.C., Benjet, C., Bromet, E., De Girolamo, G., De Graaf, R., Gasquet, I., Gureye, O., Haro, J.M., He, Y., Kessler, R.C., Levinson, D., Mneimneh, Z.N., Oakley Browne, M.A., Posada-Villa, J., Stein, D.J., Takeshima, T., Von Korff, M., 2007. Depression-anxiety relationships with chronic physical conditions: results from the World Mental Health surveys. Journal of Affective Disorders 103, 113–120.

Scott, K.M., Von Korff, M., Angermeyer, M., Bromet, E.J., Bruffaerts, R., De Girolamo, G., De Graaf, R., Fernandez, A., Gureje, O., He, Y., Kessler, R., Kovess, V., Levinson, D., Medina Mora, M., Mneimneh, Z.N., Oakley Browne, M.A., Posada-Villa, J., Tachimori, H., Williams, D., 2008. Age patterns in the prevalence of depressive/anxiety disorders with and without physical comorbidity. Psychological Medicine 38, 1659–1669.

Scott, K.M., Von Korff, M., Angermeyer, M.C., Benjet, C., Bruffaerts, R., De Girolamo, G., Haro, J.M., Lepine, J.-P., Ormel, J., Posada-Villa, J., Tachimori, H., Kessler, R.C., 2011. The association of childhood adversities and early onset mental disorders with adult onset chronic physical conditions. Archives of General Psychiatry 68, 838–844.

Shah, A.J., Veledar, E., Shallenberger, L., Murrah, N.V., Jawed, F., Bremner, D., Goldberg, J., Vaccarino, V., 2011. Association of antidepressant medications with carotid intima media thickness in middle aged veteran twins. Journal of the American College of Cardiology 57, E1588.

Sharpe, L., Curran, L., 2006. Understanding the process of adjustment to illness. Social Science & Medicine 62, 1153–1166.

Steptoe, A. (Ed.), 2007a. Depression and Physical Illness. Cambridge University Press, Cambridge, UK.

Steptoe, A., 2007b. Integrating clinical with biobehavioural studies of depression and physical illness. In: Steptoe, A. (Ed.), Depression and Physical Illness. Cambridge University Press, London.

Stetler, C., Miller, G.E., 2011. Depression and hypothalamic-pituitary-adrenal activation: a quantitative summary of four decades of research. Psychosomatic Medicine 73, 114–126.

Sütterlin, S., Vögele, C., Gauggel, S., 2010. Neuropsychiatric complications of efavirenz therapy: suggestions for a new research paradigm. The Journal of Neuropsychiatry and Clinical Neurosciences 22, 361–369.

The ENRICHD Investigators, 2003. Effects of treating depression and low perceived social support on clinical events after myocardial infarction. Journal of the American Medical Association 289, 3106–3116.

Thompson, C., Syddall, H., Rodin, I., Osmond, C., Barker, D.J.P., 2001. Birthweight and the risk of depressive disorder in late life. British Journal of Psychiatry 179, 450–455.

Vögele, C., Von Leupoldt, A., 2008. Mental disorders in chronic obstructive pulmonary disease (COPD). Respiratory Medicine 102, 764–773.

Von Korff, M., 2009. Understanding consequences of mental-physical comorbidity. In: Von Korff, M., Scott, K.M., Gureje, O. (Eds.), Global Perspectives on Mental-Physical Comorbidity in the WHO World Mental Health Surveys. Cambridge University Press, New York.

Von Korff, M., Scott, K.M., Gureje, O. (Eds.), 2009. Global Perspectives on Mental-Physical Comorbidity in the WHO World Mental Health Surveys. Cambridge University Press, New York.

Wegman, H.L., Stetler, C., 2009. A meta-analytic review of the effects of childhood abuse on medical outcomes in adulthood. Psychosomatic Medicine 71, 805–812.

Young, E.A., Abelson, J.L., Cameron, O.G., 2004. Effect of comorbid anxiety disorders on the hypothalamic-pituitary-adrenal axis response to a social stressor in major depression. Biological Psychiatry 56, 113–120.

Comparative Evolutionary Psychology: Current Status and a Proposal for a More Integrated Future

Jennifer Vonk and Todd K Shackelford, Oakland University, Rochester, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Taking into account the shortcomings and strengths of the related fields of comparative psychology and evolutionary psychology, we propose merging these subfields into Comparative Evolutionary Psychology. This new field will encompass the study of cognitive continuities and discontinuities across species, with the goal of understanding both ultimate and proximal causes for the emergence of various traits. Importantly, comparative evolutionary psychologists will embrace both field and laboratory approaches and shed the shackles of an exclusive focus on uniquely human traits. This new field will exemplify an appreciation for a diversity of traits in both closely related and distantly related species that an understanding of evolution demands.

Defining Terms

Every field of scientific inquiry has its problems. The subdisciplines within psychology are no exception. Perhaps one of the greatest stumbling blocks to progress is the perceived need of psychologists to self-identify as belonging to one particular area within psychology to the exclusion of others. This need is, at its foundation, antithetical to an integrative science and to interdisciplinary work. We see this problem expressed in the long-standing debate about what it is to be a comparative psychologist or an evolutionary psychologist, perhaps most exhaustively examined in a special issue of the *Journal of Comparative Psychology* published over two decades ago (Burghardt, in press). Papini (2003) suggested that the field of comparative psychology should focus on the phylogenetic history and adaptive significance as well as the ontogenetic origins of an organism's behavior. Cartwright (2000) suggested that comparative psychologists failed to bridge the divide between human and nonhuman, more successfully bridged by ecologists and sociobiologists. Evolutionary psychology has built upon these fields by focusing on the investigation of psychological mechanisms as adaptations produced by natural or sexual selection to solve the problems faced in our ancestral history. The focus of evolutionary psychology is thus on the ultimate cause of human behaviors and characteristics. Evolutionary psychology represents an integrated set of hypotheses focused on selection, modularity, and adaptation, and is particularly interested in applying evolutionary principles to the mind (Pinker, 1999; *see* Evolution of Cognition: An Adaptationist Perspective). Evolutionary psychology is linked more strongly to sociobiology and the biological sciences, in general, and to cognitive psychology and behavioral genetics, as well (*see* Sociobiology: Overview). Comparative psychology is tightly linked to primatology, behaviorism, ethology, and, often, cognitive science. There have been various proposals to unite the two fields (most recently, Vonk and Shackelford, 2012). Vonk and Shackelford (2012) focused on specific problems within each field and attempted to identify barriers to a more open foray into something akin to cognitive behavioral ethological psychology with a focus on studying evolved traits in a wide range of species. Our proposal, which we elaborate upon here, is to unite these areas of research into Comparative Evolutionary Psychology, with the goal of removing barriers between (1) those who work in the field and those who work in the laboratory, (2) those who work with different species with a nomothetic or ideographic approach, and, especially, (3) those with a human-centered perspective and those who study individual species that are of interest in their own right.

Goals and Limitations

Comparative Psychology: Focus on Continuity

We believe that both comparative psychologists and evolutionary psychologists have sometimes fallen into the trap of endorsing goals that are both narrowly focused and ultimately stand in defiance of evolutionary theory. For instance, we focused on the tendency of comparative psychologists to focus on continuity and similarity between species, to the neglect of equally interesting differences and species' unique adaptations (Vonk and Shackelford, 2012; *see* Comparative Method in Evolutionary Studies). Although we applaud recent efforts to expand the species of study and the questions and topics under investigation, we suggest that much would be gained by returning to the lessons of methodological behaviorists and logical positivists, who fought to place psychology among the sciences. We see the trend in animal research, particularly with regard to cognitive research, as highlighting similarities with human traits and abilities, but at the expense of objectivity. That is, the focus has been on cognitive continuity, and equally interesting discontinuities have sometimes been discounted in favor of arguing for contexts that may have interfered with the ability to express humanlike traits (Girndt et al., 2008; Vonk and Povinelli, 2006). Sometimes alternative outcomes are obtained and researchers still hold steadfast to 'preferred' hypotheses that center on similarities between humans and their closest relatives, rather than giving equal weight to equally plausible, or even more plausible, alternative explanations for their findings. As just one recent example, in a recent study of tool use in apes, Herrmann et al. (2008, p. 229) write: "Although our results show that apes

 http://dx.doi.org/10.1016/B978-0-08-097086-8.81015-6

succeeded in some problems spontaneously, their group performance never exceeded 70% in the initial six trials. It is true that with additional trials performance increased, but it still remained quite low in most conditions. One possible explanation for this outcome is that their performance is not based on causal knowledge about the task. Another possibility is that they possess some causal knowledge, but that certain task features make it hard to express it consistently." These authors insist that the apes have knowledge even when that knowledge is not clearly demonstrated in the tasks at hand rather than allowing the data to speak for itself and relinquishing their preferred hypothesis to accept a more conservative conclusion that perhaps the apes do not possess causal knowledge at all, and instead show the ability to learn by trial and error or reason about observable features.

This approach of seeking evidence for humanlike traits in nonhuman species has sometimes been deemed the 'Holy Grail' approach to comparative psychology (Povinelli and Vonk, 2004). We propose that, instead, investigators focus on traits that are expressed within a species' ecological and social environments and examine factors that may have led to the expression of these traits. Such an approach need not exclude laboratory studies that seek to explore the generalizability of behaviors to novel contexts under controlled conditions. Further, studying both captive and wild animals has the potential to illuminate the role of ontogenetic factors and the role of plasticity in development. This is not a novel argument, but when combined with a broadly comparative perspective and an evolutionary framework, this approach has the potential to unify the study of comparative psychology and evolutionary psychology in a manner that has clearly not been accomplished in recent years. The fact that this unification has not yet been achieved is evidenced by the recent number of proposals seeking greater integration. Along with our own volume (Vonk and Shackelford, 2012), see also special issues of *Philosophical Transactions of the Royal Proceedings B*, edited by Grodzinski, Clayton, and Thornton (2012) and Heyes and Frith (2012) focusing on the evolution of the animal and human mind, respectively.

Researchers should certainly be encouraged to consider whether existing capacities are similar to or different from those in both closely related and distantly related species. It is precisely such considerations that allow us to speculate about the evolutionary forces behind the emergence of these traits. However, we suggest that researchers should not be motivated by the desire to seek evidence for particular patterns of behavior simply because those behaviors exist in humans (Shettleworth, 2007), particularly if there is no ecologically relevant reason to find such a trait in the species in question. We should examine behaviors and abilities that the animal exhibits in its natural environment rather than design tasks with a predetermined goal or 'stack the deck' to find evidence for traits that have not yet appeared in natural environments. For instance, research should be motivated by an interest in adding to our understanding of capacities in species with particular behavioral ecologies, rather than attempting to demonstrate how 'intelligent' a particular species is. Thus, explorations of differences in spatial memory as a function of whether species cache food or must find patchily distributed food resources over large ranges are particularly fruitful (Shettleworth, 2010). Perdue et al.'s (2011) recent examination of sex differences in carnivore search strategies as a function of mating strategies and range size is another excellent example of such research. An unbiased approach is more likely to illuminate both parallel and convergent evolutionary processes.

The allure of Holy Grail pursuits (Povinelli and Vonk, 2004) has attracted many researchers. Rather than designing experimental tasks to uncover the cognitive capacities of a species with consideration of the ecological and social forces likely to have given rise to particular traits in particular populations, some researchers seem especially focused on seeking evidence for humanlike abilities in nonhumans, regardless of that species' ecological niche or phylogenetic distance from humans. This anthropocentric focus can be seen in book titles such as *Kanzi: The Ape at the Brink of the Human Mind* (Savage-Rumbaugh and Lewin, 1994), *Next of Kin* (Fouts and Mills, 1997), *Our Inner Ape* (de Waal, 2005), *Chimpanzee Politics* (de Waal, 1982), and *The Smartest Animals on the Planet* (Boysen & Custance, 2009). Although humans are naturally drawn to species that share certain humanlike characteristics such as the ability to read human social cues, the role of the scientist should be to laud other equally impressive abilities that various species have evolved to solve problems in their unique physical and social environments. Together, comparative evolutionary psychologists are best equipped to do so.

Evolutionary Psychology: Focus on Human Uniqueness

Some evolutionary psychologists have been similarly focused on the traits that define what it is to be human, such as cooperation and prosocial behavior (Fehr et al., 2002; Fehr and Fischbacher, 2004; Fehr and Gintis, 2007). Such a focus may result in a reversal of the Holy Grail pursuits by comparative psychologists in that researchers may focus on traits that they assume are not shared with other species, in the absence of evidence for these assumptions (Matsuzawa, 2012; Saxe, 2006; Tomasello, 1998). In actuality, evolutionary psychologists are poised to demonstrate continuities, with much work focused on sex differences in mating strategies that closely parallel sex differences in other species (Birkhead et al., 1987; Buss, 1995; Buunk et al., 1996; Schmitt, 2005; *see* Human Mate Choice, Evolution of). However, this work has been historically controversial within psychology. If psychologists were concerned with identifying both similarities and differences across broad taxonomic groups, with the goal of elucidating rather than assuming the cross-species links, comparative psychology and evolutionary psychology may find themselves less under attack and more often in the driver's seat when it comes to unifying the subfields within psychology.

Diverse Topics: The Need for Evolutionary Framework

On a positive note, examples of recent research abound with forays into previously discouraged topics for comparative research, such as consciousness and metacognition. Not only are such topics fervently studied nowadays, but their study is no longer restricted to primates; they have been extended to species as evolutionarily distant from humans as cephalopods (Mather, 2008), pigeons (Sutton and Shettleworth, 2008), and

rats (Foote and Crystal, 2007). However, seldom are such examples couched in terms of why such abilities might exist in such distantly related species. Even when such attempts are made, they seem like post hoc justifications for the research. For example, why should elephants exhibit mirror self-recognition (Plotnik et al., 2006)? Elephants live in relatively large and complex social groups and display extraordinary memory, but so do most canine species, and birds such as psittacines and corvids, and there is no evidence as of yet for mirror self-recognition in these latter species (although see Prior et al., 2008). One cannot simply suggest that a single factor, such as developing within a complex social environment, may have given rise to any particular 'higher order' cognitive trait, such as mirror self-recognition, without a logical exposition of why that factor would have led to the emergence of that particular trait, and why in this species or this population and not in others that share this factor. We would hope to encounter discussions of parallel or convergent evolutionary processes focused on life in complex social groups, or challenging physical environments necessitating the need for advanced causal reasoning to extract food, not as a post hoc suggestion, but as a driving impetus to conduct the research with multiple species and populations to begin with. In fact, one of us has recently been motivated to explore the cognitive abilities of a large-brained relatively nonsocial mammal with a nod to the importance of testing the 'social intelligence' hypothesis in nonsocial species (Vonk and Beran, 2012; Vonk et al., 2012). The social intelligence hypothesis (Humphrey, 1976; Jolly, 1966) states that species that have evolved to live in large, complex social groups will display superior social cognitive abilities (and perhaps more sophisticated cognition, in general) relative to less social species. Thus, we need to test this hypothesis by comparing social and nonsocial species that are otherwise similar (such as comparing lions with tigers or bears with wild dogs). We have also recently embarked upon a program of research to compare the cognitive abilities of various species of bats that differ both in sociality and in diet. All too often research seems instead to proceed more like a traditional Mardi Gras hunt – a search for all of the random ingredients to make a great gumbo. Once some piecemeal evidence for humanlike capacities in other species is uncovered, an exciting story is then woven together and submitted for publication in a high-profile journal where it will likely attract much publicity. This approach may tell us more about our own biases than about the minds or capacities of the species we so painstakingly study (Povinelli and Vonk, 2004; Shettleworth, 2007).

Biases: False Dichotomies

A biased approach is problematic for several reasons, not just because it flies in the face of the scientific process, but also because it may set the researcher up to fail. Creating scenarios in which the desired outcome must be an all-or-none result ignores the possibility of discovering many other results. Scientific reasoning should encourage researchers to acknowledge possible outcomes that they may not anticipate. When we do not leave ourselves open to such discoveries we may not discover them. If the only options we were open to interpreting consisted of 'humanlike'/'nonhumanlike,' we would miss out on 'chimpanzee-like' or 'elephant-like' behaviors and cognitions that we may not have imagined prior to our investigation. Before Povinelli (2012) began his extensive study of chimpanzees' understanding of weight, we may not have anticipated that the chimpanzee concept of weight was so tied to kinesthetic impressions on the body and so removed from the objects that they were lifting. The studies discussed by Povinelli (2012) demonstrate an attempt to understand how chimpanzees perceive weight – not just whether it maps onto the human concept of weight. Admittedly, it is more difficult to work from a nonanthropocentric frame as it may not be possible to experience the world through a sensory system we do not ourselves embody.

Approaching research with the idea that results will pivot on a false dichotomy such as humanlike/non humanlike, innate/learned, or conscious/automatic is bound to lead to pressure to interpret results in line with a particular strong viewpoint, when plausible alternatives might exist. The difficulty in publishing 'null' results has likely contributed to such a drive in our field. Scientists are rewarded by the public's enthusiasm for demonstrations of humanlike behaviors in even distantly related species. When rooks were discovered to drop rocks in water to raise the water levels so that they could retrieve previously inaccessible worms (Bird and Emery, 2009), the findings were touted as evidence for causal reasoning and insight, even 'tool use' – something that was once attributed only to humans. Thus, the results garnered much attention both in the scientific community and in the public. The same was true when Betty, a new Caledonian crow, another member of the corvid family, was seen to bend a piece of straight wire into a hook to retrieve a food reward from the bottom of a pipe (Weir et al., 2002), even though birds presumably bend wires and similar objects naturally when nest building. Betty may have arrived accidently at the solution to the task at hand, rather than by cognitive insight. If the birds were never to modify the objects, or use the tools at all, the results would likely not be published, or relegated to a lower impact journal, and would not garner any excitement either within or outside of the research community. Thus, one can see the motivation to seek empirical findings that stress continuity with human traits – particularly those that have been set as benchmarks of human uniqueness. Such traits are dangled like carrots tempting researchers to knock them from that status, just as 'tool use' and 'theory of mind (ToM)' have now forever been wrenched from the fists of Cartesians who would claim a strong divide between man and beast, with animals deemed incapable of any form of thought or feeling.

If there should be any lesson from the Cartesian dichotomy of the past, which allowed for decades of inhumane animal experimentation, it is that we should not create such strong divisions and set up false dichotomies when we design scientific studies. Although we have argued that researchers should not be blinded by desires to seek evidence for continuity in cognitive processes, abandoning equally strong evidence for cognitive discontinuities, it is important to stress that both continuity and discontinuity exist in nature; these complementary processes are together central tenets of evolutionary theory. In our recent volume on Comparative Evolutionary Psychology (Vonk and Shackelford, 2012), we stressed the value of casting aside such dichotomies and divisions between

those approaching research from different perspectives, frameworks, and methodologies. We advocated a release from the lack of objectivity that resonates in Holy Grail pursuits, and we advocated a return to the lessons of the logical positivists and the methodological behaviorists, who stressed the objectivity of the scientific method. We cannot move forward with credibility as a science if we do not do so. We cannot accept theories without extensive tests and we must revise hypotheses when the data do not fit, or are consistent with multiple hypotheses.

When investigators are overly invested in existing theory or preferred hypotheses, however, they sometimes manufacture middle ground hypotheses to satisfy their theoretical framework and fit the data. Thus, one might see researchers arguing that, first, chimpanzees do not share human ToM abilities (Tomasello et al., 1993), then that chimpanzees do share some ToM abilities (Tomasello et al., 2003), later that these abilities might be exhibited only within certain restricted contexts, such as in competitive contexts (Hare, 2001). These types of arguments, although interesting, do not address the very reason ToM is likely to have evolved in the first place – to enable an organism's behavioral flexibility in anticipating another organism's responses in a variety of novel circumstances. If one can reason about another's mental states in only a very constrained set of situations – those that have driven that species' fitness through imposing mating or feeding advantages (i.e., by allowing one to compete for food) – then it seems more likely a mechanism that is innate and inflexible, and precisely not what ToM is designed to do. By maintaining a clear evolutionary perspective, comparative psychologists will be better able to reject implausible hypotheses or explanations that are driven by biases to discover commonalities regardless of evidence for divergence.

The 'Problem' of 'Null' Results

Few people are excited by 'negative' findings or 'null' results, or the absence of humanlike traits in our nonhuman counterparts. The fact that we label the absence of evidence for such abilities as 'negative results' says more about our anthropocentric biases than it says about the value of the results. Such findings can be just as illuminating as 'positive' results. They tell us just as much about the animals' abilities and the way that they reason. Or at least they could, were the tests and experiments properly constructed in such a manner as to highlight what the subjects could do, and how they think, rather than being fashioned in a dichotomous manner: Do they do what humans do, or do they not? In one example, Vonk and Subiaul (2009) explored chimpanzees' ability to reason about humans' capability to perform a simple task – offering a food tray to the ape subjects. The humans' ability was constrained on the basis of whether their arms or legs were visible and available to perform the physical actions required. These results (and other similar findings) are often interpreted as failures on the part of the chimpanzees, because they did not evidence causal reasoning. However, what the researchers showed was that chimpanzees were reasoning – they just were not reasoning about the causally relevant variables in the task – even though those variables were directly observable. However, rather than showing what the chimpanzees could not do, Vonk and Subiaul, in a series of follow-up studies, attempted to show what features of the task the chimpanzees were attending to, and what features allowed them to perform successfully on some trials. And this is the bottom line in many similar studies – chimpanzees (and other animals) may or may not 'succeed' at the tasks we present to them; what matters for illuminating the cognitive processes underlying their performance is how they succeed (or fail) at the tasks. And the same should be said for determining how humans succeed in experimental tasks. We should not take for granted that humans use particular mechanisms for solving problems, until we have tested these hypotheses (Silva et al., 2005, 2008).

A closely related point is that researchers often suggest that chimpanzees 'fail' particular tasks when they do not perform in a manner consistent with how humans might be expected to perform (sometimes in the absence of having any available human data to bear on the issue). Such claims have led to suggestions that captive chimpanzees are unable to perform well in cognitive experiments because they have been cognitively deprived, removed from the socially rich lives they would have led in the wild, without the challenges of finding food, mates and shelter, and navigating social conflicts, forging alliances within large groups, and so on (Boesch, 2008). These arguments neglect the fact that experimentally experienced chimpanzees have a wealth of experiences that wild chimpanzees do not. They also live in social groups. But, in addition to that, they have extensive experience with human caretakers, human-made objects and artifacts, and contrived situations and contingencies that they would not face in the wild. Furthermore, it is precisely when you take an animal out of the environment in which evolution has sculpted them to survive that you can best determine their ability to generalize knowledge to unfamiliar and novel circumstances, as humans often do. Still, these critics also fail to acknowledge that chimpanzees that sometimes do not reach criterion or perfect performance on experimental tasks are not 'failing' our tasks. Rather, they are illuminating for us how the chimpanzee sees the world; and it is often quite different from our human-centered biases have led us to see. It is only through thoughtful probing of their responses in carefully constructed post hoc tests that we are able to learn more about their thoughts. The chimpanzees that have participated in dozens or hundreds of experiments designed by psychologists do not 'fail to be human.' Rather, they succeed at being chimpanzees, and as primatologists and objective scientists, it is our mission to learn what that means. Where results appear to differ between laboratories, it is often not the results that differ, but the interpretation of the results.

We should not claim continuity or discontinuity, or 'high' or 'low' order explanations, for either humans or nonhumans until we have examined alternative explanations for performance in our tests. Only when we fashion our experiments in an unbiased, nonhuman-centered way with an open mind as to how other species might approach problems with a consideration of the sorts of problems, they have evolved to face will we be enlightened as to the kinds of minds with which we share this planet. We will not find a human mind in another species, not even those species most closely related to humans. Those species have evolved their own traits, capacities, and thought

processes – some of which are equally impressive, astounding, and perhaps beyond our capacity to comprehend. We must embrace the nature of evolution – an appreciation for both continuity and discontinuity – even in closely related species such as humans and other great apes. In doing so, we may have to abandon Holy Grail pursuits and be open to finding evidence for abilities that our closest relatives have that we do not, and vice versa. We may also find evidence for humanlike abilities in distantly related species. We need not attempt to turn chimpanzees into some kind of half-human hybrid to appreciate their cognitive accomplishments. Our study of their minds should not be guided by such a missive.

Bridging Gaps

On that note, comparative evolutionary psychologists could take heed from the early ethologists who remained focused on the goals of determining mechanisms underlying both human and nonhuman behaviors at both the ultimate and proximate levels. To their credit, evolutionary psychologists have incorporated advances in cognitive science, behavioral economics, developmental psychology, linguistics, cultural anthropology, sociobiology, and ethology, and yet work from an overarching theoretical framework. The basic mechanisms of natural selection and sexual selection can explain both human and nonhuman mating practices (Buss, 1989, 1994), sex differences in spatial abilities (Voyer, Voyer & Bryden, 1995), attractiveness (Burke and Sulikowski, 2010), investment in offspring (Alvergne et al., 2009), and so on. Comparative psychologists, while unequivocally but sometimes only implicitly accepting evolutionary theory, have not always placed their work within the same overarching framework, and sometimes take esoteric side trips, investigating questions that, although interesting, do little to explain the forces giving rise to various traits in different populations.

This is not to say that comparative psychology is fundamentally flawed relative to evolutionary psychology, or that evolutionary psychology is beyond reproach. We have been critical of the focus of comparative psychologists to seek out commonalities between humans and other species at the expense of exploring many fascinating adaptations that humans do not share (echolocation in distantly related cetaceans and Chiroptera for instance). We have also been critical of evolutionary psychology for its tendency to focus exclusively on what other species can tell us about human evolution. Indeed, we recently attended a talk where an evolutionary psychologist defined evolutionary psychology as the study of human behavior and the human mind. This unitary focus on humans was surprising. There will be those who will staunchly defend their fields against such criticisms and point to the many counterexamples that defy our critiques. For example, there are researchers studying mate guarding in birds (Birkhead, 1988) that parallels work in humans. This work may be less prototypical of either comparative or evolutionary psychology. Should it be labeled as one or the other? Or should such work fall under the umbrella of Comparative Evolutionary Psychology, as it is both comparative in its nature and evolutionary in its focus? We would say yes – that is, exactly the kind of work that this new field would embrace. But the results of such investigations will shed light not only on human psychology but also on reasons for the emergence of this behavior in other species.

Although both comparative psychologists and evolutionary psychologists are interested in the mechanisms underlying behaviors exhibited by various species and specific human and nonhuman populations, as well as the evolutionary forces that gave rise to such traits and behaviors, these two fields have taken somewhat disparate trajectories, rather than culminating in a single field of study. We propose that much would be gained if the fields were merged into a new science of Comparative Evolutionary Psychology. The goals would remain – the new science would aim to explore the ultimate and proximate causes of behavior in humans and their close and distant relatives – and the relatively newfound appreciation for an eclectic approach to empirical questions and methodologies would be retained. We must adopt more open-minded approaches to our studies, in which we do not grasp firmly onto 'preferred hypotheses,' but are willing to formulate new theories and hypotheses that emerge from a consideration of the contexts that shape an animal's evolutionary history.

One of the solutions to removing these barriers is to increase communication and advocate for collaborations among researchers with different perspectives and specialized expertise. In order to best accomplish this open-minded perspective, we suggest dissembling the barriers that arose when creating false dichotomies between nature/nurture, behaviorism/nativism, field/laboratory, ideographic/nomothetic, and human/nonhuman research, and, most importantly, a willingness to let go of the idea that other species are most valuable when they most closely approximate humans (cf. Savage-Rumbaugh and Lewin, 1994). By doing so, we will allow ourselves to appreciate the diversity that exists on this planet, and to better understand the process of evolution – as it applies to both the behavior and psychology of organisms. Thus Comparative Evolutionary Psychology can be a single unified field for the study of evolved traits in all species.

See also: Cognition, Evolution of; Comparative Method in Evolutionary Studies; Human Mate Choice, Evolution of; Sociobiology: Overview.

Bibliography

Alvergne, Alexandra, Faurie, Charlotte, Raymond, Michel, 2009. Father-offspring resemblance predicts paternal investment in humans. Animal Behaviour 78 (1), 61–69. http://dx.doi.org/10.1016/j.anbehav.2009.03.019.

Bird, Christoper David, Emery, Nathan John, 2009. Rooks use stones to raise the water level to reach a floating worm. Current Biology 19, 1410–1414.

Birkhead, Timothy R., 1988. Behavioral aspects of sperm competition in birds. In: Birkhead, T.R. (Ed.), Advances in the Study of Behavior, vol. 18. Academic Press, San Diego, CA, pp. 35–72.

Birkhead, Timothy R., Atkin, L., Møller, A.P., 1987. Copulation behaviour of birds. Behaviour 101, 101–138. http://dx.doi.org/10.1163/156853987X00396.

Boesch, Christophe, 2008. Taking development and ecology seriously when comparing cognition: reply to Tomasello and Call (2008). Journal of Comparative Psychology 122, 453–455.

Burghardt, Gordon. The janus faced nature of comparative psychology – strength or weakness? Evolutionary Psychology, in press.

Burke, Darren, Sulikowski, Danielle, 2010. A new viewpoint on the evolution of sexually dimorphic human faces. Evolutionary Psychology 8 (4), 573–585.
Buunk, Bram P., Angleitner, Alous, Oubaid, Viktor, Buss, David M., 1996. Sex differences in jealousy in evolutionary and cultural perspective: tests from the Netherlands, Germany, and the United States. Psychological Science 7, 359–363. http://dx.doi.org/10.1111/j.1467-9280.1996.tb00389.x.
Buss, David M., 1989. Sex differences in human mate preferences: evolutionary hypotheses tested in 37 cultures. Behavioral and Brain Sciences 12 (1), 1–49.
Buss, David M., 1994. The Evolution of Desire: Strategies of Human Mating. Basic Books, New York.
Buss, David M., 1995. Psychological sex differences: origins through sexual selection. American Psychologist 50, 164–168. http://dx.doi.org/10.1037/0003-066X.50.3.164.
Cartwright, John, 2000. Evolution and Human Behavior: Darwinian Perspectives on Human Nature. MIT Press, Boston, MA.
de Waal, Frans B., 1982. Chimpanzee Politics Power and Sex among Apes. Jonathan Cape, London (other editions by Becht, Amsterdam, 1982; Harper & Row, New York, 1983).
de Waal, Frans B., 2005. Our Inner Ape a Leading Primatologist Explains Why We Are Who We Are. Riverhead, New York.
Fehr, Ernst, Fischbacher, Urs, 2004. Social norms and human cooperation. Trends in Cognitive Sciences 8, 187–190. http://dx.doi.org/10.1016/j.tics.2004.02.007.
Fehr, Ernst, Fischbacher, Urs, Gachter, Simon, 2002. Strong reciprocity, human cooperation, and the enforcement of social norms. Human Nature 13, 1–25.
Fehr, Ernst, Gintis, Herbert, 2007. Human motivation and social cooperation: experimental and analytical foundations. Annual Review of Sociology 33, 43–64. http://dx.doi.org/10.1146/annurev.soc.33.040406.131812.
Foote, Allison L., Crystal, Jonathan D., 2007. Metacognition in the rat. Current Biology 17, 551–555.
Fouts, Roger S., Mills, Stephen T., 1997. Next of Kin: My Conversations with Chimpanzees. William Morrow, New York.
Girndt, Antje, Meier, T., Call, Josep, 2008. Task constraints mask great apes' ability to solve the trap-table task. Journal of Experimental Psychology: Animal Behavior Processes 34, 54–62. http://dx.doi.org/10.1037/0097-7403.34.1.54.
Hare, Brian, 2001. Can competitive paradigms increase the validity of experiments on primate social cognition? Animal Cognition 4, 269–280.
Herrmann, Esther, Wobber, Victoria, Call, Josep, 2008. Great apes (*Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla*, *Pongo pygmaeus*) understanding of tool functional properties after limited experience. Journal of Comparative Psychology 122, 120–130.
Humphrey, Nikolaus K., 1976. The social function of intellect. In: Bateson, P.P.G., Hinde, R.A. (Eds.), Growing Points in Ethology. Cambridge University Press, Cambridge, UK, pp. 303–317.
Jolly, Alison, 1966. Lemur social behavior and primate intelligence. Science 153, 501–506.
Mather, Jennifer A., 2008. Cephalopod consciousness: behavioral evidence. Consciousness and Cognition: An International Journal 17, 37–48.
Matsuzawa, Tetsuro, 2012. What Is Uniquely Human? A View from Comparative Cognitive Development in Humans and Chimpanzees. Harvard University Press, Cambridge, MA, pp. 288–305, http://dx.doi.org/10.4159/harvard.9780674062917.c17.
Papini, Mauricio R., 2003. Comparative psychology. In: Davis, Stephen F. (Ed.), Handbook of Research Methods in Experimental Psychology. Blackwell, Malden, MA.
Perdue, Bonnie M., Synder, Rebecca J., Zhang, Zhihe, Marr, M. Jackson, Maple, Terry, 2011. Sex differences in spatial ability: a test of the range size hypothesis in the order Carnivora. Biology Letters 7, 380–383.
Pinker, Steven, 1999. How the Mind Works. WW Norton & Co., New York, pp. 386–389.
Plotnik, Joshua, de Waal, Frans B.M., Reiss, Diana, 2006. Self-recognition in an Asian elephant. Proceedings of the National Academy of Sciences of the United States of America 103, 17053–17057.
Povinelli, Daniel J., 2012. World without Weight: Perspectives on an Alien Mind. Oxford University Press, New York.
Povinelli, Daniel J., Vonk, Jennifer, 2004. We don't need a microscope to explore the chimpanzee's mind. Jointly published in Mind and Language 19, 1–28 (and Sue Hurley & Matthew Nudds (Eds.), *Rational Animals* (2006), Oxford University Press, New York.
Prior, Helmut, Schwarz, Ariane, Güntürkün, Onur, 2008. Mirror-induced behavior in the magpie (*Pica pica*): evidence of self-recognition. PLoS Biology 6, e202. http://dx.doi.org/10.1371/journal.pbio.0060202.
Savage-Rumbaugh, Sue, Lewin, Roger, 1994. Kanzi: The Ape at the Brink of the Human Mind. Wiley, New York.
Saxe, Rebecca, 2006. Uniquely human social cognition. Current Opinion in Neurobiology 16, 235–239. http://dx.doi.org/10.1016/j.conb.2006.03.001.
Schmitt, David P., 2005. Sociosexuality from Argentina to Zimbabwe: a 48-nation study of sex, culture, and strategies of human mating. Behavioral and Brain Sciences 28, 247–311. http://dx.doi.org/10.1017/S0140525X05000051.
Shettleworth, Sara J., 2007. Studying mental states is not a research program for comparative cognition. Behavioral and Brain Sciences 30 (3), 332–333. http://dx.doi.org/10.1017/S0140525X0700218X.
Shettleworth, Sara J., 2010. Cognition, Evolution, and Behavior, second ed. Oxford University Press, New York.
Silva, F.J., Page, D.M., Silva, K.M., 2005. Methodological-conceptual problems in the study of chimpanzees' folk physics: how studies with adult humans can help. Learning & Behavior 33, 47–58. http://dx.doi.org/10.3758/BF03196049.
Silva, F.J., Silva, K.M., Cover, K.R., Leslie, A.L., Rubalcaba, M.A., 2008. Humans' folk physics is sensitive to physical connection and contact between a tool and reward. Behavioural Processes 77 (3), 327–333. http://dx.doi.org/10.1016/j.beproc.2007.08.001.
Sutton, Jennifer E., Shettleworth, Sara J., 2008. Memory without awareness: pigeons do not show metamemory in delayed matching-to-sample. Journal of Experimental Psychology: Animal Behavior Processes 34, 266–282.
Tomasello, Michael, 1998. Uniquely primate, uniquely human. Developmental Science 1, 1–16. http://dx.doi.org/10.1111/1467-7687.00002.
Tomasello, Michael, Call, Josep, Hare, Brian, 2003. Chimpanzees understand psychological states – the question is which ones and to what extent. Trends in Cognitive Science 7, 153–156.
Tomasello, Michael, Kruger, A., Ratner, H., 1993. Cultural learning. Behavioral and Brain Sciences 16, 495–552.
Vonk, Jennifer, Beran, Michael J., 2012. Bears 'count' too: quantity estimation and comparison in black bears (*Ursus americanus*). Animal Behaviour 84, 231–238.
Vonk, Jennifer, Jett, Stephanie E., Mosteller, Kelly W., 2012. Concept formation in American black bears (*Ursus americanus*). Animal Behaviour 84, 953–964.
Vonk, Jennifer, Povinelli, Daniel J., 2006. Similarity and difference in the conceptual systems of primates: the unobservability hypothesis. In: Wasserman, E., Zentall, T. (Eds.), Comparative Cognition: Experimental Explorations of Animal Intelligence. Oxford University Press, New York, pp. 363–387.
Vonk, Jennifer, Shackelford, Todd K, 2012. The Oxford Handbook of Comparative evolutionary psychology. Oxford University Press, New York.
Vonk, Jennifer, Subiaul, Francys, 2009. Do chimpanzees know what others can and cannot do? Reasoning about 'capability'. Animal Cognition 12, 267–286.
Voyer, Daniel, Voyer, Susan, Bryden, M. Philip, 1995. Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. Psychological Bulletin 117 (2), 250–270. http://dx.doi.org/10.1037/0033-2909.117.2.250.
Weir, Alex A.S., Chappell, Jackie, Kacelnik, Alex, 2002. Shaping of hooks in New Caledonian crows. Science 297, 981.

Relevant Websites

http://www.hbes.com/.
www.jennifervonk.com.
http://www.toddkshackelford.com.
http://www.port.ac.uk/departments/academic/psychology/research/comparativeandevolutionary/.

Comparative Health Systems

Linda M Whiteford, University of South Florida, Tampa, FL, USA
Cecilia Vindrola-Padros, London South Bank University, London, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a replacement of the previous edition article by M. Ruggie, volume 10, pp. 6537–6544, © 2001, Elsevier Ltd.

Abstract

Comparative health systems research examines the similarities and differences between and among health policies, practices, institutions, and outcomes in different countries in order to examine the effects of health reforms, the ways in which international initiatives are adapted to suit local needs and interests, and the strategies different health systems use to deal with the same problems. This article provides a discussion of the comparative study of health systems – the limitations and benefits of this form of analysis, a review of some of the current issues researchers are engaged in, and two brief case studies.

Introduction

Efficiency? Cost-savings? Coverage? Participation? What can be more confusing than talking about health systems, except perhaps trying to compare them to each other? Which system is 'best', and what is 'best', and who defines 'best'? Should the comparison of different health systems focus on financing and spending, health system reform, the role of practitioners, the role of patients and the public, and/or technological innovation? This article provides a discussion of the comparative study of health systems – the limitations and benefits of this form of analysis, a review of some of the current issues researchers are engaged in, and then provides two brief case studies, and some concluding thoughts.

Cross-national comparative research explores the similarities and differences between/among health policies, programs, institutions, and outcomes in different countries with the aim of drawing conclusions about potential advantages/disadvantages (Johnson and Stoskopf, 2010). It often seeks to examine the effects of regional health reforms or the ways in which international initiatives are adapted to suit local needs and interests (Blank and Burau, 2007). Furthermore, it can lead to shared 'lessons learned' by pointing to the strategies different health systems use to deal with the same problems, the obstacles they faced, the solutions that were put in practice, and the kinds of outcomes generated (Gonzalez Block, 1997).

We begin with an overview of comparative health systems (CHS) research. That leads us to a discussion of current controversies and concerns in CHS research, for instance, the roles of privatization, decentralization, the participation of nongovernmental organizations (NGOs) in healthcare, and the function of patient and public involvement (PPI). In order to investigate one of these concerns – the role of PPI – we provide two case studies from very distinct health systems in Latin America (Cuba and Argentina). We conclude the article with new questions for the readers to consider.

Overview of CHS

The focus and methods used in CHS research have changed in relation to the approaches used during particular times. Initially methods relied on quantitative data derived from secondary sources and focused on macrolevel analysis that formed the preferred approach for carrying out CHS analyses (Gauld, 2005: p. 11). However, after several authors identified the limitations of this approach (Mechanic and Rochefort, 1996; Whiteford and Nixon, 1999), more recent studies have shown an interest in in-depth analyses using a mixed methods design or focusing solely on qualitative data to provide case studies of health systems in different locales (Atkinson et al., 2000; Bossert, 1998; Dawad and Veenstra, 2007; Gilson et al., 2001).

Contextualization has come to be recognized as an intrinsic component of comparative research as it provides an insight into knowledge absent from strictly quantitative analyses (Carmel, 1999) and is able to "reconcile complexity, detail and context" (Mangen, 1999: p. 110). This kind of context-sensitive approach has been proposed by Wrede et al. where comparisons are made at the mesolevel to "make connections between policy and practice in different welfare regimes" (2006: p. 2987). This approach can lead to an analysis capable of clarifying the practices of care (key actors involved, negotiation of institutional structures, historical processes, etc.) and not just the outcomes (Wrede et al., 2006: p. 2989). Even though debates about the best way to carry out CHS research continue, numerous publications concur that there are evident benefits to implementing this type of research despite its limitations (see **Table 1**).

Issues in CHS

Decentralization

Decentralization has been defined as a process "involving a variety of mechanisms to transfer fiscal, administrative, ownership and/or political authority for health service delivery from the central Ministry of Health (MoH) to alternative institutions" (Bossert and Beauvais, 2002: p. 14). This transfer of authority is envisioned as producing health systems capable of adapting to local contexts and the requirements of the local population, thus becoming more effective (in terms of outcomes and costs), accessible, and equitable (Bossert and Beauvais, 2002: p. 14). Previous studies using a CHS approach have shed light on the challenges associated with implementing decentralization strategies and the important role played by the local context in the

Table 1 Summary of the benefits and limitations of CHS research

Benefits
Identify cases of best practices and failures worldwide with the intent to improve healthcare delivery
Show global trends in health policies and systems and outline examples of convergence and social embeddedness
Investigate how healthcare systems impact health status
Shed light on the role the local context plays in the shaping of international policies and programs
Strengthen cross-country research capacity
Point to the processes behind the development of particular health systems
Inform future policy direction
Limitations
Ignores the variations and complexity within health systems and countries
The typologies of health systems are not always matched with the study objectives
Lack of a standard definition of 'health system' and its components
Difficulties applying the same analytical concepts cross-culturally
Difficulties obtaining or accessing the same type of data in different countries to make comparisons
Lack of funding for carrying out comparative research across numerous countries

Source: Blank, R., Burau, V., 2007. Comparative Health Policy. Palgrave Macmillan, New York, NY; Danad et al. (1999); Gilson, L., Kalyalya, D., Kuchler, F., Lake, S., Hezron, O., Ouendo, M., 2001. Strategies for equity: experience with community financing in three African countries. Health Policy 58, 37–67; Mechanic, D., Rochefort, D., 1996. Comparative medical systems. Annual Review of Sociology 22, 239–270; Rodwin, V. Gusmano, M., 2002. The world cities project: rationale, organization, and design for comparison of megacity health systems. Journal of Urban Health: Bulletin of the New York Academy of Medicine 79 (4), 445–463; Whiteford, L., Nixon, L., 1999. Comparative health systems: emerging convergences and globalization. In: Albrecht, G.A., Fitzpatrick, R., Scrimshaw, S.C. (Eds.), Handbook of Social Studies in Health and Medicine. SAGE Publications, pp. 440–452; Wrede, S., Benoit, C., Bourgeault, I., van Teijilingen, E., Sandall, J., De Vries, R., 2006. Decentred comparative research: context sensitive analysis of maternal health care. Social Science & Medicine 63, 2986–2997.

shaping of international policies (Bossert and Beauvais, 2002). They have concluded that one size fits all models seldom result in positive changes at the local level, and thus, decentralization strategies must be developed in accordance with the characteristics of each political, economic, and cultural context.

Health Financing and Expenditure

There has been a worldwide trend to increase the spending on health services (Mechanic and Rochefort, 1996: p. 244), but with economic crises, governments are finding it difficult to provide good-quality healthcare to a changing and expanding population in a context of scarce resources (Ruggie, 2001: p. 6543). As the percentage of GDP allocated to covering healthcare costs increases (particularly in developed countries) (Johnson and Stoskopf, 2010), researchers have explored issues linked to cost-containment and effectiveness (Whiteford and Nixon, 1999). CHS research has slowly abandoned studies with a narrow view of the economic performance of health systems which focused mainly on inputs and outputs and neglected the wide range of factors that affect the use of health resources. Instead, more complex research designs capable of considering multidimensional social, political, and cultural factors that affect economic efficiency are becoming more common (Whiteford and Nixon, 1999).

The Role of NGOs

NGOs in the form of grassroots movements, health NGOs (such as Partners in Health, and Medicines sans Frontieres), and transnational governmental organizations such as the World Health Organization (WHO) and Pan American Health Organization (PAHO) have greatly influenced the way healthcare is delivered worldwide (Walker et al., 2010). These organizations currently influence international policy decisions, run health educational and prevention campaigns, and provide health services in instances where the public health system does not have enough resources, cannot reach marginalized and isolated populations, or is dealing with a temporary crisis (e.g., an epidemic outbreak or a natural disaster) (Schuller, 2007, 2010; Pfeiffer, 2003). In many instances, they have become *de facto* health systems as aid organizations increasingly turn to NGOs to provide healthcare and funnel their resources through them (Pfeiffer, 2004).

Changing Demographics and Emerging and Re-Emerging Diseases

The effects of an aging population are being felt across many parts of the globe. Johnson and Stoskopf have identified these as: "(1) increasing healthcare costs/fewer young people to spread the risk, (2) an inability to replace the elderly in the workforce, (3) inadequate planning for long-term solutions/facilities, (4) how to finance health care and pensions, and (5) end-of-life costs" (2010: p. 408). This demographic change has direct effects on the types of services that must be made available to the population (mainly focusing on the management of chronic diseases and end-of-life care), as well as the programs of care outside of the health system that need to be put in place (Johnson and Stoskopf, 2010: p. 408).

Other recent challenges faced by health systems worldwide are emerging diseases such as obesity or re-emerging diseases such as multidrug-resistant tuberculosis. In 2008, it was estimated that over 1.4 billion adults older than 20 were overweight and over 40 million children younger than 5 were overweight in 2011 (WHO, 2013). Obesity in childhood raises the probability of developing chronic conditions such as type 2 diabetes, heart disease, hypertension, and some forms of cancer which require frequent medical attention and monitoring (Blank and Burau, 2007: p. 202). The increase in the international flows of people, the deterioration of living conditions and overcrowding, severe climate change, and the overuse of antibiotics have contributed to the emergence of new infectious diseases (SARS, avian influenza, and H1N1) and reemergence of previously eradicated ones (Johnson and Stoskopf, 2010: pp. 35–36). These diseases represent new challenges for already burdened health systems. Furthermore, as global epidemics become more common, health systems must become globalized in their conceptualization and treatment of disease (Armelagos and Harper, 2005).

Community Participation in Health

The process of enhancing the participation of the general public in healthcare is complex and has received multiple definitions. In some parts of the world, it is referred to as community

participation or involvement, while in other places, labels such as PPI, service user involvement, or users with active participation are more common (Belmartino and Bloch, 1984; Mockford et al., 2012; Reyes Álvarez et al., 1996; Solitario et al., 2008). Since the Declaration of Alma-Ata which established that "the people have the right and duty to participate individually and collectively in the planning and implementation of their health care", organizations such as the WHO, UNICEF, and PAHO have led international initiatives to promote community participation in health (WHO, 1999).

The main reasons for promoting public participation in healthcare, among others, are to increase accountability by creating more transparent decision-making processes, to challenge health inequities, to ensure the services provided are relevant to the population and adaptable to their changing needs, to maintain communication channels between health authorities and the general public, and to establish a sense of commitment and responsibility in the general public regarding health (Litva et al., 2002; Sang, 2004; Staniszewska and Denegri, 2013; Tritter, 2009). In theory, many countries around the world promote public participation in health, but in practice, the level of involvement of the general public ranges from direct involvement (where patients and the public actively participate in the making of decisions regarding health services and the use of resources) to indirect involvement (where health professionals and administrators elicit the views and experiences of the general public regarding health services to modify or develop health policies and programs tailored to their needs) (Tritter, 2009: pp. 276–277). Involvement can also be reactive (when it is shaped by a pre-existing agenda) or proactive (when the agenda is constructed by the general public) (Tritter, 2009: p. 277).

The type of involvement available is dependent on the local context, as it is influenced by previous attempts to ensure community participation, a history of success or failure, political engagement with an active public, the establishment of structural mechanisms for ensuring communication with the public, the designation of resources to promote public participation, and the ways in which strategies to guarantee participation are visualized by the population (Tritter, 2009). In order to provide a more in-depth discussion of community/patient/public involvement in health, we present two brief case studies from Cuba and Argentina.

Cuba

In 2006, Gallop Poll's affiliate organization in Costa Rica reported that 2/3 of all Cubans expressed a positive response to being asked if they had confidence in the Cuban healthcare system (Keeping Cuba healthy – BBC online 1 August 2006). While not perfect, the Cuban health system is doing many things right and is admired both within and outside of Cuba. What is right about the Cuban model are (1) the focus on prevention and (2) the role of public and patient participation (Whiteford and Branch, 2008). This model puts low-cost, universal interventions within the community and provides the personnel (family doctors and nurses, in particular) to carry out the interventions. 'Medicine-in-the-community' is a model developed in Cuba that puts a physician and nurse in every neighborhood. According to a recent BBC news report, Cuba has 1 physician for every 170 residents (BBC news). Neighborhood residents are obligated to be responsible for some portion of their healthcare while the State provides the rest. According to researchers, this reciprocal and symbiotic relationship between the patient and the state, and the active role of the Cuban patient and public in the primary healthcare system accounts for its success. Everyone is supposed to be visited at home at least once a year, often without warning, so the general practitioner can scrutinize a patient's lifestyle (Boseley, 2007). Some might consider this an intrusive prying into peoples' lives, however, as one Cuban physician noted when people became accustomed to his unannounced visits, they became more like visits among friends and communication was more open and honest. As a result, he felt he understood and communicated with his patients better as they came to understand him better (Whiteford and Branch, 2008).

While the State is responsible for the health of population, individual members of the population are also held accountable for their health. They are responsible for attending health fairs, doing exercise, and avoiding or reducing unhealthy practices such as smoking. Indeed, their participation is required and even mandated (Kath, 2006). In Cuba, healthcare policy development is very much a top-down process, but once in the neighborhood, the community plays a central role in the participatory playing out of those policies (Boseley, 2007).

What we can see from the Cuban primary healthcare system is the application of Kark's COPC model. It succeeds, in part, because of the effective involvement of the population and patients in accessing and improving their own healthcare. What we must also take into account with the Cuban primary health model is that it is also invasive and paternalistic, often challenging individual rights in its pursuit of controlling health outcomes. The relationship between the individual and the State in Cuba is complex, and it may be that in the healthcare arena it is most successful. That said, it must also be noted that the economic stress following the loss of the previous Soviet Union as a trading partner and supporter, has resulted in a loss of materials in most Cuban clinics. Cuba is rich in highly trained medical personnel and, reciprocally, in communities educated to access their services. In terms of public/community, patient involvement, the primary health system in Cuba is a successful model for other countries to follow.

Argentina

In the case of Argentina, the terms social participation (participación social) and the people's participations (participación popular) are used to describe a wide range of participatory strategies such as the monitoring of the distributions of medical resources, the approval of health programs, advocacy and the defense of health as a universal human right, consultation in the making of decisions, lobbying and the suggestion of changes in policies, and the involvement in health promotion activities (often at the level of disease prevention) (Zaldúa and Bottinelli, 2010).

In these strategies, NGOs, locally labeled as civil society organizations, are the main promoters of social participation in health. An example of this nongovernmental involvement is the creation of the Argentine Social Forum of Health and the Environment (Foro Social de Salud y Ambiente). This forum is

composed of "social, humanitarian, union-based, human rights, academic, territorial, and citizen organizations united to fight for the people's collective right to health" (Foro Social, 2002). In its aim to defend a universal health system, the Social Forum promotes social participation in health by prioritizing public involvement in the design, execution, management, and control of health policies (Foro Social, 2002). The Forum has made some notable advances such as arranging a meeting with representatives of local NGOs and the Vice Minister of Health in 2002 to discuss health-related issues and drafting a law on free health services and free transportation to health centers (Moyano and Escudero, 2007). Unfortunately, these initiatives have not yet been activated (Moyano and Escudero, 2007).

The Consulting Council (Consejo Consultivo) is a governmental group designated to promote social participation on a wide range of issues including healthcare. The purpose of this committee is to "generate a fluid exchange of information between government administrators and representatives of civil society" (CCSC, 2013). The council considers health to be a topic of collective interest and thus promotes open discussion with the public to establish the public health agenda and forward health policies (González García and Tobar, 2004: p. 18).

Social participation in health in Argentina has been promoted at both governmental and nongovernmental levels and has expanded from consultation of the local population about issues related to health services to their active engagement in the drafting of health policies and the raising of new issues. Much work still needs to be done to include sectors of society that are still being left out of conversations (such as young people or the elderly) to continue to foment the proactive direct involvement of the public and to establish long-term mechanisms of involvement that can bypass changes in political administrations (Anigstein, 2009; Belmartino and Bloch, 1984; Solitario et al., 2008). This last point has been the most notable obstacle in attempts to promote social participation in Argentina.

Conclusion

While the process of comparing health systems can be confusing and complex, it also can be an immensely useful activity. It allows the observer to become familiar with healthcare models other than his or her own, and equally importantly, it provides a template for assessing the various qualities of different models. As we review some of the lessons learned from the Cuba and Argentina examples we see that 'paper' participation as in Argentina is very different from the neighborhood popular participation in Cuba. Both Cuba and Argentina appear to have support for participatory healthcare from the very top of the political hierarchy, but only Cuba makes that support real and sustained in the community. But as CHS research clearly demonstrates, we must consider the differences is size, scale, economics, and political organization that Cuba and Argentina exemplify as we compare and contrast these realities. In sum, CHS is a critical element in our on-going analytical attempts to understand and improve healthcare systems globally.

See also: Culture as a Determinant of Mental Health; Globalization and Health; Health Care, Rationing of; Health Economics; Health Politics and Policy; Health Research, Qualitative; Health in Low-Income Countries; Health: Anthropological Aspects.

Bibliography

Anigstein, C., 2009. Investigación sobre participación ciudadana en salud en 4 países del Mercosur: Argentina, Bolivia, Paraguay y Uruguay. Medicina Y Sociedad 28 (4), 1–19.

Armelagos, G.J., Harper, K.N., 2005. Disease globalization in the third epidemiological transition. In: Guest, G. (Ed.), Globalization, Health and the Environment: An Integrated Perspective. AltaMira Press, Walnut Creek, CA, pp. 27–33.

Atkinson, S., Rolim Medeiros, R., Lima Oliveira, P., Dias de Almeida, R., 2000. Going down to the local: incorporating social organisation and political culture into assessments of decentralised health care. Social Science and Medicine 51, 619–636.

BBC, 2006. Keeping Cuba Healthy. http://news.bbc.co.uk/1/hi/programmes/newsnight/5232628.stm (accessed 18.08.13.).

BBC NEWS, 2006. Medical Know-How Boosts Cuba's Wealth. http://news.bbc.co.uk/2/hi/business/4583668.stm (accessed 18.08.13.).

Belmartino, S., Bloch, C., 1984. Reflexiones sobre participación popular en salud en Argentina. Cuadernos Médico Sociales 27, 1–14.

Blank, R., Burau, V., 2007. Comparative Health Policy. Palgrave Macmillan, New York, NY.

Boseley, S., 2007. Cuban Healthcare: First World Results on a Third World Budget. http://www.guardian.co.uk/world/2007/sep/12/film.health?INTCMP=SRCH (accessed 18.08.13.).

Bossert, T., 1998. Analyzing the decentralization of health systems in developing countries: decision space, innovation and performance. Social Science and Medicine 47 (10), 1513–1527.

Bossert, T., Beauvais, J., 2002. Decentralization of health systems in Ghana, Zambia, Uganda and the Philippines: a comparative analysis of decision space. Health Policy and Planning 17 (1), 14–31.

Carmel, E., 1999. Concepts, context and discourse in a comparative case study. International Journal of Social Research Methodology 2 (2), 141–150.

Consejo Consultivo de la Sociedad Civil (CCSC), 2013. ¿Qué es el CCSC? http://ccsc.mrecic.gov.ar/que-es-el-ccsc (accessed 07.07.13.).

Dawad, S., Veenstra, N., 2007. Comparative health systems research in a context of HIV/AIDS: lessons from a multi-country study in South Africa, Tanzania and Zambia. Health Research Policy and Systems 5, 13–19.

Foro Social de Salud y Ambiente Argentina, 2002. Quienes Somos. http://www.forosocialsaludargentina.org/es/template.php?file=quienes_somos/quienes_somos.html (accessed 05.07.13.).

Gauld, R., 2005. Comparative Health Policy in the Asia-Pacific. Open University Press, Berkshire, UK.

Gilson, L., Kalyalya, D., Kuchler, F., Lake, S., Hezron, O., Ouendo, M., 2001. Strategies for equity: experience with community financing in three African countries. Health Policy 58, 37–67.

Gonzalez Block, M., 1997. Comparative research and analysis methods for shared learning from health system reforms. Health Policy 42, 187–209.

González García, G., Tobar, F., 2004. Salud para los Argentinos. Ediciones Granica, Buenos Aires, Argentina.

Harkness, J., 1999. In pursuit of quality: issues for cross-national survey research. International Journal of Social Research Methodology 2 (2), 125–140.

Johnson, J., Stoskopf, C., 2010. Comparative Health Systems: Global Perspectives. Jones and Bartlett Publishers, Sudbury, MA.

Kath, E., 2006. Father Knows Best? Cuba's Proactive Approach to Healthcare Delivers Results but Paternalism Brings Some Compromises. ASCE, pp. 351–365.

Litva, A., Coast, J., Donovan, J., Eyles, J., Shepherd, M., Tacchi, J., Abelson, J., Morgan, K., 2002. 'The public is too subjective': public involvement at different levels of health-care decision making. Social Science and Medicine 54, 1825–1837.

Mangen, S., 1999. Qualitative research methods in cross-national settings. International Journal of Social Research Methodology 2 (2), 109–124.

Mechanic, D., Rochefort, D., 1996. Comparative medical systems. Annual Review of Sociology 22, 239–270.

Mockford, C., Staniszewska, S., Griffiths, F., Herron-Marx, S., 2012. The impact of patient and public involvement on UK NHS health care: a systematic review. International Journal for Quality in Health Care 24 (1), 28–38.

Moyano, G., Escudero, J.C., 2007. La salud en Argentina: ¿En manos de quién? Informe sobre la situación de la salud en Argentina. http://www.fmed.uba.ar/depto/ddhh/.../argentina.doc (accessed 07.07.13.).

Pfeiffer, J., 2003. International NGOs and primary health care in Mozambique: the need for a new model of collaboration. Social Science and Medicine 56, 725–738.

Pfeiffer, J., 2004. Civil society, NGOs, and the holy spirit in Mozambique. Human Organization 63 (3), 359–372.

Reyes Álvarez, I., Sanabria Ramos, G., Medina Gondres, Z., Báez Dueñas, R.M., 1996. Metodología para la caracterización de la participación comunitaria en salud. Revista Cubana de Salud Pública 22 (1), 1–11.

Rodwin, V., Gusmano, M., 2002. The world cities project: rationale, organization, and design for comparison of megacity health systems. Journal of Urban Health: Bulletin of the New York Academy of Medicine 79 (4), 445–463.

Ruggie, M., 2001. Health care systems: comparative. In: International Encyclopedia of Social and Behavioral Sciences. Elsevier, pp. 6537–6544.

Saltman, R., 1997. Convergence versus social embeddedness: debating the future direction of health care systems. European Journal of Public Health 7, 449–453.

Sang, B., 2004. Choice, participation and accountability: assessing the potential impact of legislation promoting patient and public involvement in health in the UK. Health Expectations 7, 187–190.

Schuller, M., 2007. Gluing globalization: NGOs as intermediaries in Haiti. PoLAR: Political and Legal Anthropology Review 32 (1), 84–104.

Schuller, M., 2010. Shattered and scattered: Haiti's quake through the lens of human rights. NACLA Report on the Americas 43 (4), 20.

Solitario, R., Garbus, P., Stolkiner, A., 2008. Derechos, ciudadanía y participación en salud: su relación con la accesibilidad simbólica a los servicios. Anuario de Investigaciones de la Facultad de Psicología 15, 263–269.

Staniszewska, S., Denegri, S., 2013. Patient and public involvement in research: future challenges. Evidence Based Nursing 16 (3), 69.

Tritter, J., 2009. Revolution or evolution: the challenges of conceptualizing patient and public involvement in a consumerist world. Health Expectations 12, 275–287.

Walker, M., Johnson, A., Ledlow, G., 2010. The role of nongovernmental organizations in global health. In: Johnson, J., Stoskopf, C. (Eds.), Comparative Health Systems: Global Perspectives. Jones & Bartlett Publishers, Sudbury, MA, pp. 387–402.

Whiteford, L., Branch, L., 2008. Primary Health Care in Cuba: The Other Revolution. Rowman and Littlefield Press, Plymouth, UK.

Whiteford, L., Nixon, L., 1999. Comparative health systems: emerging convergences and globalization. In: Albrecht, G.A., Fitzpatrick, R., Scrimshaw, S.C. (Eds.), Handbook of Social Studies in Health and Medicine. SAGE Publications, pp. 440–452.

World Health Organization (WHO), 1999. Declaration of Alma-Ata. International Conference on Primary Health Care, Alma-Ata, USSR, 6–12 September 1978. http://www.who.int/hpr2/archive/docs/almaata.html (accessed 02.07.13.).

World Health Organization, 2013. Obesity and Overweight Factsheet Number 311. http://www.who.int/mediacentre/factsheets/fs311/en/ (accessed 26.03.14.).

Wrede, S., Benoit, C., Bourgeault, I., van Teijilingen, E., Sandall, J., De Vries, R., 2006. Decentred comparative research: context sensitive analysis of maternal health care. Social Science & Medicine 63, 2986–2997.

Zaldúa, G., Bottinelli, M.M., 2010. Praxis psicosocial comunitaria en salud. EUDEBA, Buenos Aires, Argentina.

Relevant Websites

http://www.commonwealthfund.org/Resources/2008/Mar/Health-Care-System-Profiles.aspx – The Commonwealth Fund, 2008. Health Care System Profiles.

http://www.euro.who.int/en/who-we-are/partners/observatory – European Observatory on Health Systems and Policy Series.

http://www.insp.mx/ichsri/ – International Clearinghouse of Health System Reform Initiatives (ICHSRI).

http://www1.paho.org/english/sha/profiles.htm – Pan American Health Organization, 2013. Basic Country Health Profiles in the Americas.

http://www.hsph.harvard.edu/ihsg/publications.html – Publications of the International Health Systems Group (IHSG).

http://www.who.int/whosis/en/ – World Health Organization (WHO), 2013. WHO Statistical Information System (WHOSIS).

Comparative History

Heinz-Gerhard Haupt, European University Institute, Florence, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Comparative history as a way of writing and conceptualizing history was propagated against national history during the 1920s, but developed mainly during the 1960s and 1970s. This development was encouraged by historical sociologists in the United States and the reception of social science approaches and results in historical research. Historiographies that are near to analytical methods and social sciences are more open to comparative methods. Their application started in economic and demographic studies before moving to political and social history. From the 1990s onward comparative history, which remained in all national historiographies in a minority position, was contested by the cultural transfer approach and transnational history. Both objected to comparison to be too much linked to the nation-state and to neglect the entanglement of the compared units.

Comparative History

Comparative history differs from other historical methods in that it takes an explicit line of questioning to compare two or more cases stemming from different contexts. The aim of this operation is either to bring out the similarities and differences of the different cases or to determine the scope of social scientific theories or theoretical approaches. Seldom does comparison itself lead to the formulation of theories. These characteristics distinguish historical comparison monographic work concentrating on case studies, from historical syntheses with an international scope and from implicit comparison, which indeed belongs among the analytical instruments of all historical work.

The Story of 'Comparative History'

Although the comparative perspective has appeared repeatedly in the history of historiography, it has only presented itself as a methodical instrument of explicit theoretical comparison since the 1930s. Of course, the historiography of the Enlightenment made use of comparison in its endeavor to grasp the specific characteristics of the historical process by means of universal historical types; in this context, Montesquieu, Adam Smith, and the cameralist J.H.G. Justi should be mentioned. In Scottish, English, French, and German Enlightenment literature as well, cultural areas and development stages were compared. Thus, just as the demarcation of the European model against non-European civilizations in the eighteenth century belonged to the arsenal of universal history, so did the belief that developed societies can trace their own history in that of less developed societies (e.g., Karl Marx). In this attempt, development thinking became closely intertwined with a Eurocentric perspective. This universal historical tradition persisted, carried on in the twentieth century by such diverse authors as Oswald Spengler and Arnold Toynbee, Karl Polanyi, and Shmuel Eisenstadt. Modern comparative history, however, is not rooted in this tradition. Profoundly affected by the experience of World War I and the traumas suffered by parts of academia under nationalistic exclusionism, comparative history established its argument for comparison as a fruitful enterprise particularly based on the results produced in neighboring sciences. Henri Pirenne, a Belgian historian and one of the pioneers of internationally comparative historiography, did admittedly make reference to comparative history in his opening speech at the fifth international conference of historians in Brussels in 1923, but with the intention of combating the narrowness of purely national perspectives and the subjection of historical research to functional ends. However, Marc Bloch and Otto Hinzte, two of the most influential forerunners in the field of comparative study, were even more active in taking up approaches from neighboring disciplines where comparison had long been a commonly practiced method. Historiography, as a historically comparative discipline, was a latecomer compared with other sciences. In fact, as early as the late eighteenth and early nineteenth centuries, various sciences had practiced comparative study successfully. Such works were produced not only in anatomy and physiology, but also in religious studies, jurisprudence, geography, and philosophy. Among them was that of F. Max Muller, the first *Introduction to the Science of Religion*, published in 1873. Thus it was no coincidence that Marc Bloch, in his 1928 essay advocating the practice of comparative history, does not cite historians, but rather Antoine Meillet and the broad range of scholars who had been carrying out internationally comparative linguistic research since as early as the nineteenth century.

In the development of modern comparative history, contact with the social sciences played the central defining role. Marc Bloch, whose works on the 'royal touch,' on the agrarian history of England and France, and on the different forms of feudal society in Europe up to 1940 provided important examples of comparative study, had received theoretical stimuli, above all from Emile Durkheim. Durkheim saw comparison as the sociological method *par excellence*, enabling the delineation of the fundamental characteristics of social types. Bloch used this method, for example, in his analysis of feudal society when he concentrated on its dominant dependency relationships. Otto Hintze, who was among the pioneers of historical comparison, also took the work done in the social sciences as his model. He adopted the concept of the ideal type from Max Weber and went on to construct an ideal type concept of feudalism. He then applied this to the various manifestations of feudal domination that he found to be particularly pronounced in

Russia, Japan, and the Islamic countries. In the period following World War II in particular, the number of theoretically based works with an internationally comparative perspective increased. It was largely inspired by historical sociology. On the one hand, it profited from the broad scope of sociology's macro comparisons that were limited neither to a specific epoch nor to a specific continent. These comparisons studied different national cases, using stringent argumentation to examine theoretical questions. Thus Barrington Moore inquired as to the role of agrarian development in the emergence of either democracy or dictatorial fascistic regimes in a broad comparison of six national cases, while Jack Goldstone compared revolutionary crises of the early modern age in Europe and Asia. On the other hand, however, it was only the historical research in monographic form that made available to historical sociologists the theses and research results that would enable them to carry out extensive comparisons. The stimuli obtained from historical sociology were also accompanied by critique of its generalizations and its handling of the historical state of research, which it did not always improve upon extensively. Nevertheless, historical sociology provided an important stimulus to comparative history. Its works elicited an increasingly favorable response in the 1960s to the extent that history was open to theoretical questions, and gradually after 1970, these works were also emulated. The theory of modernization, which originated mainly in the United States in the 1950s, provided a powerful comparative impetus as well, and elaborated parameters of development against which it measured individual societies. The successful combination of market economy and political democracy was its criterion for modernity. In this sense, the vague but positively cast concept of 'modernity' was used widely in historiography on the nineteenth and twentieth centuries; it emphasized structures, functions, and interdependencies among the different parts of systems more strongly than their historical individuality. In individual historiographies – particularly in those countries catching up in industrial and political development – the comparison to the Anglo-Saxon 'leader' countries soon became one of the standard lines of argumentation. In the West German debate about the long- and medium-term causes of National Socialism, structural deficits were identified in comparison to Anglo-Saxon development, which itself was admittedly highly typological and idealized: these included the feudalization of the German middle classes, a general deficit in the middle-class way of life, the exceptional proximity of middle-class life to the state, and the persistence of privilege-based social orders in the old and new middle class. Especially in this debate about Germany's *Sonderweg* ('unique path') – but in Italian historiography as well – it was no longer just the linear and unceasing advance toward modernity that was at issue, but the ruptures, contradictions, and deficits in the development.

Disciplinary and National Patterns of Comparative History

Within historical writings, comparative methods by no means penetrated all disciplines to the same degree or simultaneously. They emerged earlier in economic history, population history, and political history than in social history, gender history, or cultural history. As early as 1957, the Economic History Association dedicated one day of its annual conference to comparative economic history, and it was no coincidence that a year later, economic historian Sylvia Trupp founded the journal *Comparative Studies in Society and History* with the expressed goal of publishing studies of an interdisciplinary and comparative nature. In Europe, the English model of industrialization and its spread was the central topic of the comparative studies that appeared even before 1960. The most broad-based and influential comparative analysis was produced by Alexander Gerschenkron, who studied the basic similarities and differences within the European process of industrialization, inquiring into not just the different answers to similar challenges, but also the functional equivalents among European societies. He explains the differences in development on the basis of both the specific position of the individual economy in the system as a whole ('relative backwardness') and the influences generated by developed societies. Walt Rostow's 'stage theory' of economic growth also provided an impetus to comparative economic history after 1960, as did Simon Kuznets's quantitative studies. Rostow's thesis that every society has to go through similar phases of economic growth and that therefore each society's position can be identified against this background was almost a direct invitation for checking by historical comparison. Kuznets, in contrast, instead emphasized the uniformity of modern economic growth, although without ignoring its national diversity. His three-sector thesis also led to comparative studies.

In historical demography, in 1963 at Princeton, Ansley Coale began comparing 600 European regions to answer the question of whether and why the birth rate was declining in Europe; this was called the European Fertility Project. Although the project did produce convincing demographic results, the explanation of demographic behavior on the basis of historical macro processes came under criticism. The reasons for the early comparative studies of political movements and institutions can be found not only in the field of 'comparative politics,' which gained importance in the USA in the 1960s, but also in a specific approach to research within political science. This approach applies to institutions that can be classified under the general type of democratic systems, but that nevertheless take on specific forms. Research on resistance movements and on social systems or constitutions paved the way for the advancement of comparison in social history.

Nevertheless, since the mid-1960s, sociohistorical comparison has taken on a leading role, showing a remarkable breadth of research and differentiation in approaches. In Europe, at least a handful of papers have been published every year since 1970 on comparative social history, none limited to a specialized area, but including the history of the family and social classes, and the development of the social state, the city, and education. These studies have been focused mainly on three central issues. First, they deal with the explanation of national peculiarities that often – because of political events – are seen as characteristic of the social development of the last few centuries. In the debate about the German *Sonderweg*, this orientation of comparative sociohistorical research has taken its most comprehensive and probably also most fruitful form as a scientific heuristic device. Second, they focus on the

different pathways of individual European societies into the modern age: their distinctive features are identified above all in comparisons between different societies, and between European and non-European cultures. A third focal point has been the European model and its social characteristics. All these studies have demonstrated methodical diversity in both quantity and in quality; have integrated social, political, and economic factors; and have focused attention on both linear and cyclical developments. Furthermore, they have compared societies as well as parts of societies.

European historiographies have not all made use of the comparative method to the same degree. Surprisingly, Marc Bloch's appeal in France went unheard for many years, as French historiography was and continues to be strongly oriented toward monographic work and highly skeptical of theory. In Great Britain as well, historical comparison did not catch on because of the dominance of national history and the distance from the social sciences. In Italy, comparative history remained on the periphery – despite the proximity of many historians to debates in the social sciences – because of the strong regional orientation. In West Germany, Scandinavia, Switzerland, and Austria comparative work was rather limited within total historical production, even though the number of comparative works was significantly higher than inside other European historiographies. Although sociohistorical comparison was particularly strong in Germany, this was mainly a result of both debate on the approach toward the German *Sonderweg*, which was dealt with in many comparative historical essays, and the explicit discussions of social theories and proposed explanations that characterized modern German social history. In general, one may assume that internationally comparative history had a difficult time gaining a foothold wherever the relation to theory was weak, and wherever attempts to explain national patterns of development in the philosophy of science remained only peripheral. Before 1989, this was the case in Iberian societies as well as in Eastern Europe. Historians in the United States, where, according to Raymond Grew, comparison is a weak point rather than a strong one, has not only brought forth a series of excellent comparative studies on European history, but also practices transepochal and transatlantic comparison more intensively than does European historiography. Among the comparative works on Europe, those of Charles Tilly, Peter Baldwin, and Susan Pederson are particularly worthy of mention. The study by Arno Mayer on the Holocaust, which he sees in the tradition of the Crusades and the Thirty-Years' War, takes a transepochal and internationally comparative approach and presents a comparative tableau of the situation of Jewish people in Eastern European societies in the 1930s. Transatlantic perspectives are opened up by those studies that place slavery in the South, the frontier problem, and the situation of migrants in a larger, multinational context. In gender history, the comparison of male and female roles was for a long time the constitutive element of analysis, and only recently have theoretical, internationally comparative studies been produced. These have concentrated on institutions or subareas, which, like social laws, have attained a special meaning for gender roles. Modern cultural history as well, which discusses the creation and relevance of stems of meaning, rituals, and symbols, has concentrated thus far on national configurations and ascribed little importance to international comparison. In historiography on classical antiquity, the Middle Ages, and the early modern age, comparative studies were carried out, comparing rulers, regimes, and individual institutions. Because of the international character of political systems, the national aspect decreased in importance as a unit of comparison. However, theoretically oriented comparisons remained scarcer in this area than in the historiography on modernity.

Functions of Historical Comparison

The comparative method pursues two goals alternately. On the one hand, it seeks to accentuate the distinctive feature of each individual case, and on the other, attempts to derive evidence on general developments from case studies. While historians tend toward the first approach, the second is more prevalent among social scientists. Comparative historical studies do, however, also deal with the question of commonalities. Four functions are granted to comparison in these studies: a heuristic function, a contrastive function, an analytical function, and a distancing function. Comparison plays a heuristic role when it alone can offer explanations and reveal phenomena that had been unknown or inadequately known up to that point. This function was already referred to by Marc Bloch, who, based on his knowledge of the English enclosure movement, sought an equivalent in French agrarian history. Founding his thesis on research in regional history, he dates the disappearance of collective rights in Provence as early as the fifteenth century, and concludes from this that similar movements took place in France not only earlier, but also under other conditions. Thanks to the comparative method, Bloch was able to discover a characteristic of French agrarian development through his studies of the agrarian history of other countries.

Historical comparison can be called contrastive when it serves to define more precisely the special features of a specific case. Above all in German–French comparison, this method was used to bring into bold relief the particularly xenophobic character of German nationalism in the early nineteenth century, the insurance character of Germany's social security system, and its unique characteristic of having a highly educated middle class. Comparison of Italian and Anglo-Saxon developments brought out the backward state of Italian industrial development, its lack of a modern party system, and its unique feature of having a patrician, noble class. Depending on the country selected for comparison and the logic behind this comparison, specific characteristics of the mostly isolated national case but also of regional patterns of development are illuminated and brought into clear contrast against the totality. At least two problems emerge in this perspective: the further the development stage in the reference country or region deviates from the individual case in question, the less suitable it is for defining the specifics of the case beyond mere identification of general deficits. If one measures the economic development in Italy in the nineteenth century using as a yardstick the conditions that promoted industrialization in England, one can only determine that these conditions were lacking in Italy, but cannot identify the specific conditions for the economic growth that took place there. The more comparison is used to highlight the contours of a specific case, the more typological and

reductive the resulting picture might be. International case studies written in comparison to the German case with the aim of presenting a general view of the livelihood and benefits of German white-collar employees have only rarely provided convincing analyses of the individual national developments in the employee milieu. This inherent bias of comparison should be borne in mind in each case. Comparison has an analytical character when it either tests a scientific hypothesis or identifies constellations of causes in a specific situation. The thesis that a causal relationship exists between capitalism and feudalism is relativized by the fact that strong fascist movements did not emerge in all capitalist societies; they were able to develop only under specific conditions. Jurgen Kocka qualified this assumption in his comparison of American and German society between the two World Wars. In a study of the French, Russian, and Chinese revolutions, Theda Skocpol inquired as to the results of collapsed administrative apparatuses, broad peasant rebellions, and movements among political elites, thereby demonstrating the extreme diversity in the constellations of causes that explain the outbreak of modern revolutions. In this analytical approach, comparison can also serve as an indirect experiment. When phenomenon *a* is ascribed to cause *b*, the historian can test their hypothetical connection by looking for constellations in other societies in which *a* appears without *b*, or where *b* exists without leading to *a*. This heuristically useful procedure is distinguished from experiments in the natural sciences in that here, the *ceteris paribus* conditions are seldom given.

Comparison has a distancing effect when it offers another perspective to observation and analysis. It can produce surprising discoveries as well as relativizing the tradition-based context of national historiographies. Especially for those historiographies that are deeply embedded in a national context, comparison can open up new and often broader vistas. When confronting comparisons with other reference cases – especially those from other cultures – not only does one gain experience with different types of question and method, but also fundamental assumptions of one's own historiography can be revealed and their problems expounded. The comparative view can also contribute new insights: Skocpol, for example, uses the comparative method to shed new light on the similarities between the French Revolution of the eighteenth century and the Chinese revolution of the twentieth century in her comparison of revolutions, and between the cities of France and Japan in the eighteenth century. Taking into consideration the variety of alternative pathways into the modern age prevents exclusive concentration on the European development path and offers insight into those conditions of European development that cannot be generalized.

Methodology of Comparison

Although comparative history does not possess its own methodology, it does present particular methodological problems in bold relief. Historians have used the proximity of their work to original sources as a special proof of its scientific nature. The goal of reconstructing history via a multitude of sources in as comprehensive a manner as possible is one of the fraternity of historians' standards. This fundamental methodological principle to a certain degree evades historical comparison. In so far as it compares more than two cases, it depends more heavily on secondary literature than on the evaluation of a consistent body of sources. Thus, it also faces the problem of having to assess adequately the historiographic context, where each individual work of secondary literature interprets the original sources in its own specific way. Another of the principles of historical study is that individual aspects of reality cannot be understood separately from their place within a totality or within a historical development. The full meaning of details comes into focus only when looked at in synchronic and diachronic perspective. The method of isolating variables – a common practice in empirical social research and political economy – can and should be used within the framework of research projects, but is inadequate for representing the results of research. Comparisons can trace diachronic development processes in their similarities or differences, but they can also examine transepochal cases along one specific line of questioning. For example, Raymond W. Goldsmith studies the premodern financial systems in ancient Mesopotamia, the Indian Mogul Empire, and England under the rule of Elizabeth I. For him, what is of interest is not how an epoch can be characterized, but the contribution of different states and ages to a central line of questioning. Comparison can also evade both chronology and the total context in question: one separates out the individual cases, subjects them to examination from a specific perspective, reduces their complexity, and studies them as exemplary cases belonging to a 'universal,' the *tertium comparationis*.

Because of this isolation and reduction, the methodology of comparison enters into a relationship of unresolvable tension with fundamental principles of historical study. What has proven to be a rule of thumb is that the loss of context in comparative study is less severe the smaller the number of reference cases. By choosing medium-range theories and argumentation on a moderate level of abstraction, one limits as far as possible both the loss of concreteness and the distance from the object of study. To do justice to the diverse kinds of relationship affecting a particular case, comparative historical studies include processes rather than constellations, and link the analysis of individual situations with questions derived from the historical study of relationships. The integration of individual cases into the social context can be intensified through close attention to the linguistic and conceptual specifics that define and influence these contexts. Despite all of these methodical problems, historical comparison also has a positive impact on historical research by raising its consciousness about the extent and limits of different approaches, the implicit premises of scientific study, and the coherence of the total context. It opens up historical study to theoretical reflection.

The unit of comparison changes according to the line of questioning and the information being sought. For a long time, a national framework of study was dominant in the science of history, both because this framework influenced the problems being studied and also because answers to questions of national history were sought in comparison. Research on nationalism and welfare states, democracy, and religious denominations has been given a privileged place in the national framework. This national approach has been called

into question by two developments: microhistory and intercultural history.

Microhistory has brought increased interest to local, often unique sources and problems that give a privileged role to the viewpoints of the people involved, and which – because of the emphasis on individuality – had been excluded previously from methodical comparison. In addition, this shows the importance of another area that calls into question the self-assuredness of past research on national historical developments. For the study of family constellations and famine unrest, local frameworks present themselves as the most appropriate, while for the analysis of linguistic dialects or industrialization, the regional level is preferable. As the study of history becomes increasingly global, the national level will decrease continuously in significance. This does not just apply to colonialism, the migration movements of the nineteenth and twentieth centuries, or the global economy that has been developing since 1880. Marc Bloch pointed out as early as 1928 that the development of manorial rule in the Middle Ages and early modern age in individual localities and regions cannot be understood independently of the decline of money rent throughout Europe.

In so far, then, as one seeks to identify the special features of Europe's development, it is necessary to include non-European cultures, which, however, should no longer serve merely as a foil as they did for Max Weber, but rather be examined with a differentiated view to their independence and specificity. In order to obtain a new perspective, Jurgen Osterhammel proposes a new approach: instead of analyzing non-European cultures from a European standpoint, one should take the opposite viewpoint and determine European specifics from the standpoint of non-European cultures. The methodical problems of such a perspective are as obvious as its advantages: these include on the one hand problems of languages, the knowledge of often very different historiographies, and aggregation of data that decrease in meaningfulness and are often statistically antiquated, and on the other, a new view of known developments, and the opening of a wider sphere of activity and relationships.

In particular, the history of transfer processes brought attention to the problem Galton is formulating for ethnology. He questioned whether two or more cases can really exist independently alongside one another, or if they are enmeshed in a visible or invisible web of relationships that influence them decisively. The question of how specific constellations develop through direct influence, appropriation or imitation of foreign models does indeed increase the complexity of comparison, but in no way replaces it. As is well known, Marc Bloch extolled the study of societies in the same epoch that are neighboring, mutually influential, and subjected to the same macro changes as one of the ideal methods of historical comparison.

Comparative History's Agenda and Unsolved Problems

Comparison poses methodical and theoretical problems for the science of history, but at the same time challenges it, and can even work as a factor to increase method-consciousness and support theoretical reflection. Currently, an expansion of previous approaches and fields of comparison in the following directions could be considered: Until now, comparative studies have limited themselves to defining the range and meaningfulness of theories. Such problems as whether Weber's class theory is appropriate for the history of the European middle class, whether Pierre Bourdieu's concept of cultural capital is useful for an analysis of intellectuals, or whether Marx's theory can be applied to modern revolutions – and whether all of these can explain more than just national cases – have been the focus of interest thus far. However, the question of what changes must be made in theoretical concepts if they are to explain more than one national development and encompass a multitude of empirical cases has seldom been asked. Therefore, comparative history has contributed little to the development of theory thus far, and has seldom taken the step to formulate theoretical alternatives by relativizing existing approaches.

During the last 20 years, transnational and global history approaches are challenging the practice of international comparative history that focussed on national case studies. The concept of transfer history promoted by the French cultural historians Michel Espagne and Michael Werner draws attention to the relationship between two cultures and highlights the importance of cultural mediators as well as the transformation of the transferred cultural product. It questions the image of culturally homogenous societies and stresses the importance of foreign contributions. Transnational history, started at the beginning of the 1990 in the United States is not a method, but a perspective of analyzing historical processes. It stresses the linkage between nations, regions, and towns in a global, European, or transnational perspective and is near to diplomatic and international history writing. Its main goal is to promote a history beyond the nation-states and to underline the dynamic character of movement and interpenetration between the analyzed units. The most important results have been obtained in analyzing migrations and migrants. But also studies on transnational networks as worker's international or international organizations as the World Bank or Amnesty International have been undertaken. Even if this approach tends to privilege an optimistic vision of transnational realities and should be more attentive to the effects of transnational relationships, it motivates comparative history to integrate the 'entanglement' between the compared units in their studies. As much as comparative history needs the knowledge of entanglements, as much research on entanglement must use comparisons.

Contemporary processes of globalization have left their mark above all on the capital and labor markets, but have also called into question the national contexts of thinking prevalent in the sciences. On the one hand, historical research in an international and global framework is becoming more important. Comparisons between civilizations and continents, Empires and regions are an important part of it. On the other hand, the individual societies themselves have become complex 'melting pots,' as reception and diffusion of cultural models, social modernization, and social and cultural coding have merged in them, accompanied by national-historical and global-social influences. With the increasing demand for an adequate survey of contemporary society, the demands placed on a theoretically based, methodically differentiated comparative history will

increase as well, expanding the spectrum of questions with which history will be interpreted.

See also: Civilization, Concept and History of; Comparative Method in Evolutionary Studies; Economic History; Global History: Universal and World; Historical Demography; History and the Social Sciences; Marx, Karl (1818–83); Modernization and Modernity in History; Modernization, Sociological Theories of; Polanyi, Karl (1886–1964); Social History; Weber, Max (1864–1920).

Bibliography

Atsma, H., Burguie re, A. (Eds.), 1990. Marc Bloch aujourd'hui. Histoire comparee et Sciences sociales. Paris.

Bloch, M., 1928. Pour une histoire comparee des societes européennes. In: Bloch, M. (Ed.), Melanges Historique. Paris, I, pp. 16–46.

Budde, G., et al. (Eds.), 2006. Transnationale Geschichte. Themen, Tendenyen und Theorien. Gottingen.

van den Braembussche, A.A., 1989. Historical explanation and comparative method: towards a theory of the history of society. History and Theory 28, 2–24.

Espagne, M., 1999. Sur les limites du comparatisme en histoire culturelle. Genese 17, 102–121.

Gerschenkron, A., 1962. Economic Backwardness in Historical Perspective. Cambridge, MA.

Goldstone, J.A., 1991. Revolution and Rebellion in the Early Modern World. Berkeley, CA.

Haupt, H.G., Kocka, J. (Eds.), 1996. Geschichte und Vergleich. Ansa tze und Ergebnisse international vergleichender Geschichtsschreibung. Frankfurt, Germany.

Haupt, H.G., Kocka, J. (Eds.), 2009. Comparative and Transnational History. Central European Approaches and New Perspectives. Oxford, New York.

Hintze, O., 1964. Soziologische und geschichtliche Staatsauffassung. In: Oestreich, V.G. (Ed.), Soziologie und Geschichte – Gesammelte Abhandlungen, vol. 2. Gottingen, pp. 239–305.

Kaelble, H., 2000. Der Historische Vergleich. Frankfurt, Germany.

Kocka, J., 1977. Angestellte zwischen Faschismus und Demokratie. Zur politischen Sozialgeschichte der Angestellten, USA 1890–1940 im international Vergleich. Gottingen, Germany.

Mayer, A., 1988. Why Did the Heavens Not Darken? The 'Final Solution' in History. Pantheon Books, New York.

Moore, B., 1966. Social Origins of Dictatorship and Democracy. Lord and Peasant in the Making of the Modern World. Boston Passato e presente 28 (1993).

Pederson, S., 1993. Family, Dependence and the Origins of the Welfare State: Britain and France 1914–1945. Cambridge, MA.

Rostow, W.W., 1960. The Stages of Economic Growth. Cambridge University Press, Cambridge.

Rossi, P. (Ed.), 1990. La storia comparata. Approcci e prospettive. Milan.

Skocpol, T., 1979. States and Social Revolutions: A Comparative Analysis of France, Russia and China. Cambridge University Press, Cambridge.

Tilly, C., 1984. Big Structures, Large Processes, Huge Comparisons. Russell Sage Foundation, New York.

Comparative Method in Anthropology

Andre Gingrich, Austrian Academy of Sciences, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents comparison in anthropology as the field's third main methodological inventory, together with ethnographic fieldwork and historical anthropology. A new pluralism of comparative procedures is discussed by examining binary, regional, temporal, distant, and fluid versions of comparison. Finally dissociated today from previously predefined connections to grand theories, anthropological comparison primarily is of a qualitative orientation, remaining closely connected to fieldwork insights while moving from it to an additional level of careful abstraction that may allow for better distinctions between particularities and commonalities in human sociocultural lives.

In its most elementary dimensions of meaning, comparison designates a basic activity of the human mind by the way of implicitly or explicitly identifying similarities and differences among two or more sets of experiences, impressions, and related phenomena. Any comparative perception thus would usually contain a simultaneous interplay of both, similarities as well as differences. The general concept of comparison therefore includes a continuous range of various versions of that interplay. One end of the continuum is represented by the dominance of identifying sameness as well as parallels and similarities. The other end of the same continuum can be seen in those instances when the perception of contrasts and difference is the most important. In fact, the identification of differences may be prevailing in the process to such an extent that some scholars prefer to classify 'contrasting' as a separate mental process. The present entry, however, shares the alternative thesis that contrasting is better understood as one version of comparison with a specific imbalance in the regular interrelation between finding similarities as well as differences. In such a perspective, comparison is closely related to, and intersecting with, other main mental activities and capacities such as recognition, memory, anticipation, intuition, or decision (Bloch, 1998).

Beyond their general relevance in human life, mental activities also are intrinsic to the specialized realm of research processes. In this sense, comparison also designates such elementary procedures of academic inquiry that raise systematic questions about similarities and differences among the phenomena under examination, and about their causes. This is why comparative questions about similarities and differences may be inspired by, or give at least rise to, considerations about the particular (or the unique), the specific (or a limited class of phenomena pertaining to the same set or class), and the more general. It follows that comparison is part of a variety of methodological procedures in academic research that basically are inspired and directed from the outset by the central research question(s), and in turn by the theories and empirical insights informing that central research question (Palmberger and Gingrich, 2014). Quite rarely is the central research question primarily or exclusively defined by theoretical deliberations alone ('deduction'), or by empirical insights alone ('induction'). Rather, in anthropology and related fields of the social sciences and humanities the most common instances of identifying a central research question directing comparative and other methodological procedures emerge out of the interplay between certain theoretical deliberations and empirical insights ('abduction') (Reichertz, 2004).

As an element in comparison at large, anthropological comparison shares the main features of the general type of mental operation discussed so far. In addition, it is characterized by a number of constellations that are more or less specific to this academic discipline and a few of its neighboring fields. For sociocultural anthropology during the early twenty-first century, comparison in this sense represents the third central element in the field's key methodological inventory (Gingrich, 2013). Ethnographic fieldwork in its many changing varieties comes first, ever since fieldwork was established as the field's primary mode of empirical research. Historical methods in anthropology gradually have come to be globally acknowledged during the final quarter of the last century as second to that. Comparative methods went through much more contested times in anthropology, before they became widely recognized again as the third key methodological element in the anthropology of the globalized present.

Comparative Approaches in Anthropology's History

Prior to the emergence of the current phase in anthropology, with its strong emphasis on transnational and global contexts, the field went through a series of crises and critical self-examinations during the last quarter of the twentieth century. Those crises and intellectual movements contributed to the dissolution of main national (and quasinational) traditions which had dominated much of anthropology's history throughout the twentieth century. Each of these main national traditions, and of the main phases of evolutionism that had preceded them, pursued such comparative procedures that served their respective theoretical purposes. In that sense, each and every 'meta-narrative' in anthropology's history before the 1980s had its own comparative priorities.

Evolutionism compared local societies and cultures by criteria such as differing socioeconomic, or organizational and ideological complexities, taking them as examples for past stages of human development (Sanderson, 1997). Diffusionism compared cultural features by criteria of differing aesthetic appearance and morphological form, taking them as evidence

of respective distance to previous external influence emanating from possible centers of origin (Schweitzer, 2011). Historical particularism, by contrast, remained thoroughly skeptical about most forms of comparison whatsoever. This was in line with the general assumption that particular cultures were so unique that they could hardly be compared against each other at all, except perhaps within one area of distribution (Darnell, 2008). Functionalism, on the other hand, compared local societies by criteria of institutional differentiation, which were assessed as primary indicators of social integration (Kuklick, 2008). Structuralism, finally, compared elementary relations in myths and in marriage alliances to elaborate its theories of the ways by which the human mind operates (Parkin, 2005) – and so on and so forth. In short, during the era of anthropology's grand theories and national traditions, comparison largely was reduced to a qualitative tool at the service of proving a grand theory's main assumptions and premises.

This instrumentalization of comparison by 'national' qualitative priorities then became radicalized, homogenized, and in a certain way, universalized by the elaboration of an *Atlas of World Cultures*. In this context, G.P. Murdock and his associates established the 'Human Relations Area Files' (HRAF). On the basis of the HRAF, comparison was developed into a generalized quantitative research tool to document and explore the global distribution of correlations between identified sociocultural criteria across cultures ('hologeistic sampled comparison' or cross-cultural method) (Sarana, 1996). In many ways, that quasiclimax in anthropology's comparative record of the twentieth century eventually turned out to be an anticlimax. For methodological reasons as much as for its contexts of emergence (during World War II Murdock's original project had been inspired and commissioned by US government interests (Price, 2008), to whom it maintained a certain affinity after 1945), comparative work with the HRAF fell out of fashion in anthropology since the 1960s and 1970s. Together with increasingly intense and more differentiated internal criticism of anthropology's grand theories and their underlying metanarratives, comparison as understood in the previous mainstream manner went into a period of demise that lasted until about the end of the twentieth century.

A New Pluralism of Procedures in Anthropological Comparison

Together with the emergence of transnational and global orientations in anthropology around the last turn of the century, comparison has returned back into the field's main methodological tool-kit, yet in more modest and in more pluralist forms than ever before (Holy, 1987; Gingrich, 2012). Key factors related to anthropological comparison's reemergence were the growing interest in the field's contemporary agenda, as well as questions related to anthropology's legitimacy as an academic discipline, and finally, developments in neighboring fields: In philosophy, sociology or history, the postmodern moment went by somewhat faster, and developments there indicated the sustained relevance of comparative factors (e.g., 'constant comparison' in sociology's grounded theory or wider comparative perspectives in postcolonial historical research). Inside anthropology, the rejection of old metanarratives at first had made versions of relativism and particularism somewhat more attractive for a while. Yet it soon became obvious that an academic field could not sufficiently address any larger issues, nor sustain any credibility and legitimacy if based on nothing else but particularities. At the same time, the growing impact of various dimensions of globalization also redirected researchers' attention increasingly toward inquiries about what humans share and where they differ in their interactions with that impact.

Anthropological comparison thereby gained new relevance in today's contexts. It now is largely dissociated from predefined connections to any metanarrative or general theoretical claims, and in that sense holds a status that is similar to current principles of participant observation and ethnographic fieldwork. Anthropological comparison primarily follows qualitative procedures, with quantitative elements at best playing a secondary role. The choice of examples to be compared usually follows the principle of 'controlled comparison' (Eggan, 1954): random choice is avoided, and the cases to be compared are thoroughly theorized and checked at first. This has led to today's new pluralism of comparative anthropological methods (Gingrich and Fox, 2002), among whom the most widely applied are these five:

1. *Binary comparison* relates phenomena in two different sociocultural settings with each other. Together with Laura Nader (1994), it can be argued that at least in an implicit way, almost every anthropologist pursues this form by comparing aspects of his or her fieldwork's host context with the researcher's home contexts (unless the two are identical). For reasons of transparency and communication, it is worthwhile to make these elementary forms of self-reflexivity explicit. This also helps to minimize the dangers of any stereotyping in terms of 'us vs them' that are in fact inherent to this approach. Carried out in a balanced way, however, binary comparison may provide profound cultural critique (Marcus and Fischer, 1986) and rich contrasts (Dumont, 1994; Lock, 1995; Sahlins, 2004).
2. *Regional comparison* is the best established version in the anthropological record (Fortes and Evans-Pritchard, 1940). While the temporal contexts of compared examples should be kept within fairly similar spans of ranges, three or more cases from the same – smaller or larger – regional sphere are compared against each other. This procedure requires careful reflection and theorizing about the region and the criteria to be compared in question. Without this, inherent potentials about envisioning some kind of 'cultural area,' 'cultural circle,' or 'style province' are difficult to avoid (Fillitz, 2002). Yet at its best, regional comparison may provide insights into the specificities of certain trajectories among multiple modernities or into specific yet not universal ways of the past (Godelier and Strathern, 1991; Kuper, 1982)
3. *Temporal comparison* relates examples from different instances along a time line with each other, while keeping the regional contexts fairly constant. If the earlier or earliest instance in temporal comparison is not thoroughly theorized, then misguided assumptions about 'origins' and 'archaic conditions' might misinform this procedure. In its appropriate and accurate versions, however, temporal comparison will elucidate further historical research, as

well as to synthesize some of its previous results (Comaroff and Comaroff, 1991/97; Tambiah, 1977). By definition, temporal comparison in anthropology has to rely more substantially than the other four versions on results from neighboring fields such as archaeology or history, yet it operates with anthropological concepts and methods that often also include insights derived from fieldwork.

4. *Distant comparison*, also called 'self-reflexive, controlled macro-comparison'. When three or more cases are compared with each other across widely different spatial and/or temporal contexts, then such a 'distant' comparison usually follows a well specified query: examining the conditions for the emergence of substantial urban cultural growth in Calcutta of the mid-nineteenth century, Vienna of the nineteenth/twentieth centuries' turn, and San Francisco of the mid-twentieth century was an example for such a query by Ulf Hannerz (1992). Tony Robben's (2010) exercise in applying a number of known insights from warfare elsewhere to the unknown situation in then Iraq represents another case in point. Both examples show that distant comparison in most cases has to rely on rich historical and anthropological work being already completed, so that the main comparison then can take place *a posteriori*. Perhaps the most obvious risk in this procedure – as in fact with all *a posteriori* comparison – is the likelihood that the kind of empirical evidence is not sufficiently available which would allow to substantiate the theoretical and comparative ambition. Once that is the case, however, distant comparison allows for fairly generalized results, and for theorizing with unexpected outcome (Gingrich, 2002).
5. *Fluid comparison* – also called 'shifting time–space comparison' – follows the flows of a phenomenon through time and space. This may be the production and usage of sugar cane in different eras and continents (Mintz, 1985), the rise and spread of Islam, or the expansion of capitalism around the globe (Wolf, 1982). By contrast to distant comparison, fluid comparison avoids jumping too much across temporal or spatial distances, but instead it traces lines of continuity and change. In its empirical dimension, this perhaps is the most demanding procedure, simultaneously also requiring substantial theoretical rigor with a potential of highly rewarding results if successfully carried out.

A number of examples have been indicated in this provisional overview of five main versions of current anthropological comparison. The examples illustrate that many among this field's main and pioneering works around and after the last turn of the century in one or the other way strongly relied on comparative procedures. These works range from classic works and milestones in economic and medical anthropology to transnational approaches, from cutting-edge work in historical anthropology to advances in anthropology's study of violence and conflict to the field's contributions in migration studies, cognitive research, and to the anthropology of art. Obviously, comparative approaches in anthropology have entered a new phase of respect and attraction. At the same time and because of this, it seems appropriate to emphasize that anthropological comparison is no self-serving endeavor. It is useful and in place to add an a priori or an *a posteriori* comparison to a research perspective whenever that promises additional insights that may not be delivered in sufficient ways by other forms of investigation. Identifying the *purpose of comparison* therefore is crucial – whether it is inspired by outlining the wider relevance or the specificity of the main examples under discussion, whether cultural critique is the driving motive or grasping certain aspects of wider developments such as the results of transnational flows or the potentials in multiple modernity. In the end, at their best the various versions of comparison therefore do serve conceptualization and theorizing at least by way of exclusion, yet sometimes also through substantiating an argument.

Current Debates

This final section discusses three main points of contemporary debates on comparison in anthropology:

1. The loose and possibly incomplete differentiation into the five versions of binary, regional, temporal, distant, and of fluid comparison already is providing elements of new answers to old questions about the choices and the limits of comparison. The kinds of choices that come along with each of these five main versions simultaneously depend on the guiding research question, as well as on the available empirical evidence and its interpretation. It is also obvious that the limits of comparison cannot be predetermined, neither for the individual cases nor for the wider comparative range into which they are integrated. It seems safe though to emphasize that these limits may remain fairly fluid and fuzzy in most cases, with possible exceptions in medical and ecological contexts (Bernard and Ryan, 2010: pp. 145–160). This also opens up these comparative methodological strategies to a plurality of epistemological orientations inside anthropology, which may range from a Wittgenstein-inspired focus on 'family resemblances' to more conventional versions of pragmatism, of speculative realism and to some among the more recent elements of rationalism.
2. If theory-inspired research questions and empirical research contexts do or do not attribute sense and meaning to comparative projects, then old-fashioned notions about a *tertium comparationis* become obsolete. Derived from the rhetorical and literary traditions of late antiquity, that principle necessarily required assumptions about an 'essence' that had to be envisioned inside all objects of comparison to make rhetorical or literary comparison meaningful. Now, most anthropologists would not feel comfortable with, and would not want to subscribe to a methodological point of view that speculates about any shared 'essence,' reducing comparison to a search for commonalities only while disregarding diversity, and being deeply rooted within the narrow constraints of Aristotelian logic at that. By contrast, qualitative anthropological comparison in all its main versions strives to remain open about the outcome of the comparative endeavor. While the Aristotelian assumption of a *tertium comparationis* reduces comparison to a function of objectivist essentials, anthropological comparison elaborates *criteria of comparison* in close dialogue with the ethnographic record. The point is that the researcher has to develop such criteria of comparison that are derived out of the main research question, while simultaneously having a capacity to actually address

the empirical features in question in each case example to be compared. In this manner, the criteria of comparison cross-cut all the cases used and investigated during a comparative project, eliciting similar or different results from each of them (Gingrich, 2012).

3. A final point returns to the relation between anthropological comparison and ethnographic fieldwork. Like all of sociocultural anthropology, its comparative methodologies rely on the results of ethnographic fieldwork and remain fairly close to them. Even those forms of comparison that rely on one or the other variety of historical anthropology usually do include some intersection with ethnographic fieldwork results. In that sense, anthropological comparison at large is not an independent methodological procedure. It gives priority to ethnographic fieldwork, and to a substantial extent it depends on its outcome. One could of course add that the reverse is also true: no ethnographic fieldwork could ever take place without some comparative exercises being carried out prior to fieldwork (including decisions about the choice of topics and of field sites), and during the course of fieldwork. After all, it is only through comparing again and again the different encounters and interactions during fieldwork that the ethnographer will distinguish situations in which certain elements follow some standard or repetitive order from others in which the particular or the accidental is prevailing. In that sense, the relation between ethnographic fieldwork and one or the other form of comparison is reciprocal. Yet within the tension between empirical findings and their interpretation at one end of the spectrum, and theorizing at the other end, fieldwork by definition is closer to the empirical side, and comparison is closer to theorizing. Due to a usually small number of examples to be compared, the amount of abstraction remains fairly modest in anthropological comparison. Writing or filming an ethnographic account already per se represents a first level of unavoidable abstraction and interpretation. Anthropological comparison is the next and second level in the same process – to the point that one may sometimes even speak of a 'thick comparison' that remains faithful to its contexts of emergence (Gingrich and Thelen, 2012; Niewöhner and Scheffer, 2010).

See also: Anthropological Writing; Anthropology and History; Anthropology at Home; Cognitive Anthropology; Cultural Critique, Anthropological; Cultural Relativism, Anthropology of; Diffusion: Anthropological Aspects; Ethnography; Ethnohistory; Fieldwork in Social and Cultural Anthropology; Functionalism in Anthropology; Globalization, Anthropology of; Interpretation in Anthropology; Modernity: Anthropological Aspects; Multisited Field Studies; Postcoloniality; Psychological Anthropology; Reflexivity in Anthropology; Structuralism.

Bibliography

Bernard, R.H., Ryan, G.W., 2010. Analyzing Qualitative Data: Systematic Approaches. Sage, Thousand Oaks, CA.

Bloch, M., 1998. How We Think They Think: Anthropological Approaches to Cognition, Memory, and Literacy. Westview Press, Boulder.

Comaroff, J., Comaroff, J., 1991/97. Of Revelation and Revolution. Christianity, Colonialism, and Consciousness in South Africa, vol. I. The Dialectics of Modernity on a South African Frontier, vol. II. University of Chicago Press, Chicago.

Darnell, R., 2008. North American traditions in anthropology: the historiographic baseline. In: Kuklick, H. (Ed.), A New History of Anthropology. Blackwell, Oxford and Malden, pp. 35–51.

Dumont, L., 1994. German Ideology: From France to Germany and Back. University of Chicago Press, Chicago and London.

Eggan, F., 1954. Social anthropology and the method of controlled comparison. American Anthropologist 56, 743–763.

Fillitz, T., 2002. The notion of art: from regional to distant comparison. In: Gingrich, A., Fox, R.G. (Eds.), Anthropology, by Comparison. Routledge, New York and London, pp. 204–224.

Fortes, M., Evans-Pritchard, E. (Eds.), 1940. African Political Systems. Oxford University Press, London.

Gingrich, A., 2002. When ethnic majorities are dethroned: towards a methodology of self-reflexive, controlled macro-comparison. In: Gingrich, A., Fox, R.G. (Eds.), Anthropology, by Comparison. Routledge, New York and London, pp. 225–248.

Gingrich, A., 2012. Comparative methods in socio-cultural anthropology today. In: Fardon, R., Harris, O., Marchand, T.H.J., Nutall, M., Shore, C., Strang, V., Wilson, R.A. (Eds.), Handbook of Social Anthropology, vol. 2. Sage, Los Angeles, London, New Delhi, Singapore, Washington, pp. 201–214.

Gingrich, A., 2013. Methodology. In: Carrier, J.G., Gewertz, D.B. (Eds.), Handbook of Sociocultural Anthropology. Berg-Bloomsbury, London and New York, pp. 107–124.

Gingrich, A., Fox, R.G. (Eds.), 2002. Anthropology by Comparison. Routledge, London and New York.

Gingrich, A., Thelen, T., 2012. Comparative anthropology: achievements, failures and future. Éthnologie Française 42 (2), 385–391.

Godelier, M., Strathern, M. (Eds.), 1991. Big Men and Great Men: Personifications of Power in Melanesia. Cambridge University Press, Cambridge.

Hannerz, U., 1992. Cultural Complexity: Studies in the Social Organization of Meaning. Columbia University Press, New York.

Holy, L. (Ed.), 1987. Comparative Anthropology. Blackwell, Oxford-New York.

Kuklick, H., 2008. The British tradition. In: Kuklick, H. (Ed.), A New History of Anthropology. Blackwell, Oxford and Malden, pp. 52–78.

Kuper, A., 1982. Wives for Cattle: Bridewealth and Marriage in Southern Africa. Routledge & Kegan Paul, London.

Lock, M., 1995. Encounters with Ageing: Mythologies of Menopause in Japan and North America. University of California Press, Berkeley, CA.

Marcus, G.E., Fischer, M.J. (Eds.), 1986. Anthropology as Cultural Critique: An Experimental Moment in the Human Sciences. University of Chicago Press, Chicago.

Mintz, S.W., 1985. Sweetness and Power: The Place of Sugar in Modern History. Viking, New York.

Nader, L., 1994. Comparative consciousness. In: Borofsky, R. (Ed.), Assessing Cultural Anthropology. McGraw-Hill, New York and Toronto, pp. 84–94.

Niewöhner, J., Scheffer, T. (Eds.), 2010. Thick Comparison: Reviving the Ethnographic Aspiration. Brill, Leiden.

Palmberger, M., Gingrich, A., 2014. Qualitative comparative practices: dimensions, cases, and strategies. In: Flick, U. (Ed.), Handbook of Analyzing Qualitative Data. Sage, Los Angeles, London, New Delhi, Singapore, Washington.

Parkin, R., 2005. The French-speaking countries. In: Barth, F., Gingrich, A., Parkin, R., Silverman, S. (Eds.), One Discipline, Four Ways: British, German, French, and American Anthropology. Chicago University Press, Chicago, pp. 157–253.

Price, D.H., 2008. Anthropological Intelligence: The Deployment and Neglect of American Anthropology in the Second World War. Duke University Press, Durham and London.

Reichertz, J., 2004. Abduction, deduction and induction in qualitative research. In: Flick, U., Kardorff, E. v., Steinke, I. (Eds.), A Companion to Qualitative Research. Sage, London- Thousand Oaks, CA and New Delhi, pp. 159–165.

Robben, A.C.G.M. (Ed.), 2010. Iraq from a Distance: What Anthropologists Can Teach Us about the War. University of Pennsylvania Press, Philadelphia, PA.

Sahlins, M.D., 2004. Apologies to Thucydides: Understanding History as Culture and Vice Versa. University of Chicago Press, Chicago.

Sanderson, S.K., 1997. Comparative method. In: Barfield, T. (Ed.), The Dictionary of Anthropology. Blackwell, Oxford, pp. 78–79.

Sarana, G., 1996. Comparative method. In: Barnard, A., Spencer, J. (Eds.), Encyclopedia of Social and Cultural Anthropology. Routledge, London and New York, pp. 118–121.

Schweitzer, P., 2011. Diffusion. In: Kreff, F., Knoll, E., Gingrich, A. (Eds.), Lexikon der Globalisierung, Transcript, Bielefeld, pp. 48–51.

Tambiah, S.J., 1977. The galactic polity: the structure of traditional kingdoms in Southeast Asia. Annals of the New York Academy of Sciences 293, 69–97.

Wolf, E.R., 1982. Europe and the People without History. University of California Press, Berkeley, CA.

Comparative Method in Evolutionary Studies

Chris Venditti and Mark Pagel, University of Reading, Reading, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Pagel, volume 4, pp. 2403–2411, © 2001, Elsevier Ltd.

Abstract

The 'comparative method' refers to a set of statistical methods and ideas for studying adaptive evolution across groups of biological species. Comparative methods are an important branch of evolutionary studies primarily because of their ability to reveal broad or general evolutionary trends, and to provide insight into adaptive evolution. In recent decades, it has also been appreciated that they can be used to study evolutionary trends in human societies and so have begun to be widely used by anthropologists as well as biologists.

Overview

The 'comparative method' refers to a set of statistical methods and ideas for studying adaptive evolution across groups of biological species (Figure 1(a)). We can illustrate their use by examining one of the classic comparative relationships – that between the size of an organism's body and the size of its brain (Figure 1(b)): brain size increases steadily and predictably with body size in mammals, and this broad relationship also holds within most other animal groups. What makes the relationship

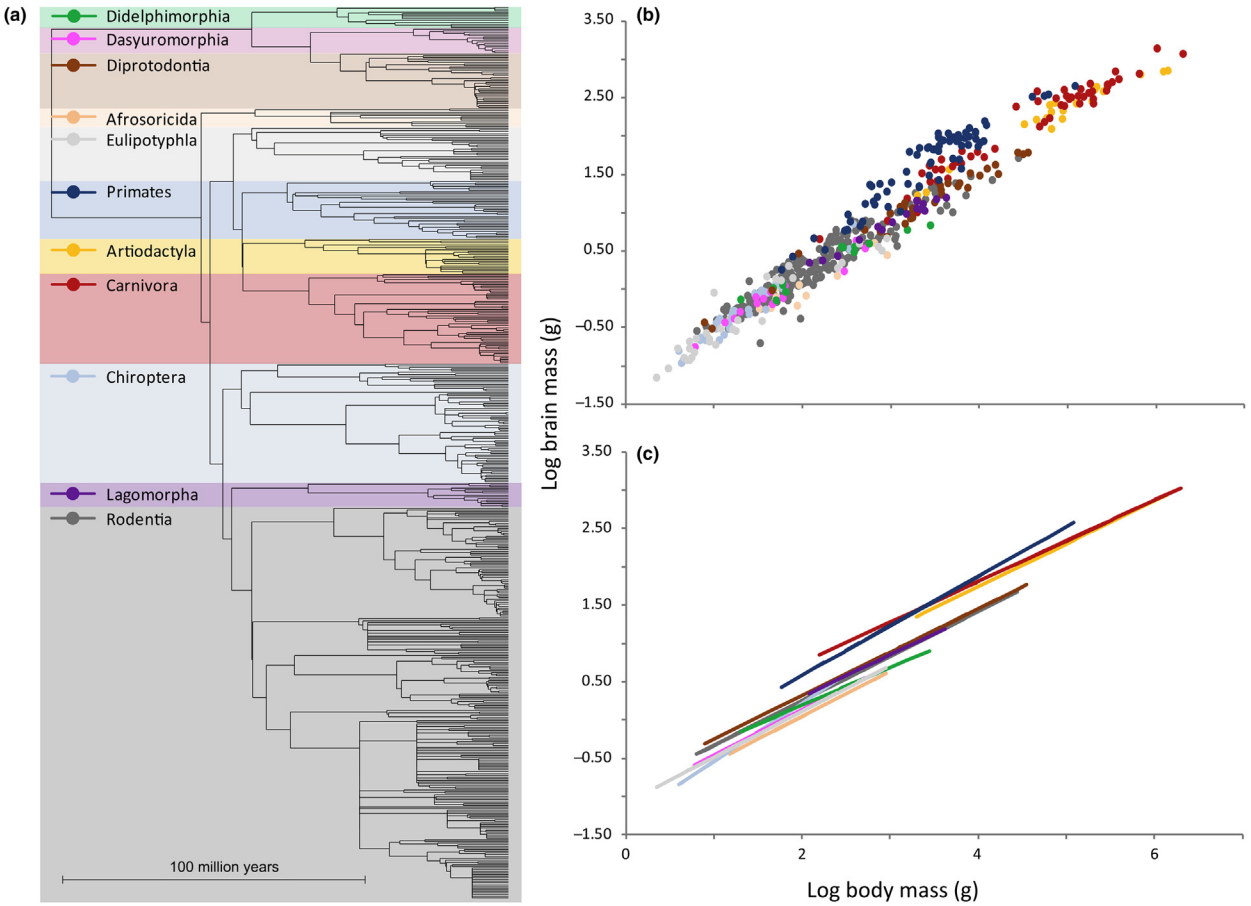

Figure 1 Mammalian brain and body size. (a) A phylogenetic tree of 538 mammals (the phylogeny is adapted from Bininda-Emonds, O.R.P., M. Cardillo, K.E., Jones, R.D.E., MacPhee, R.M.D., Beck, R., Grenyer, S.A., Price, R.A., Vos, J.L., Gittleman, and A. Purvis. 2007. The delayed rise of present-day mammals. Nature 446(7135): 507–512; (b) Plot of brain mass versus body mass for these 538 mammals. They are color coded according to order. Note the logarithmic scales; (c) The model predictions for a GLS model that always the slopes and y-axis intercepts vary by mammalian order. See text for further discussion.

striking is that across a range of animals from the size of mice to elephants, if we know the animal's size we can predict the size of its brain with surprising accuracy.

The natural interpretation of such comparative relationships is that they reveal general adaptive trends. This is to say that the organisms that individually evolved to conform to the predicted trend were somehow more fit or suited for survival than those that did not. Thus, as animals evolve to become larger, a larger brain is required for 'housekeeping' functions as well as simply to monitor the increased number of neuronal signals coming from the organism's body and internal organs.

Larger animals also tend to roam over wider areas, and a larger brain may be needed to, for example, keep track of spatial relationships or to remember where things are found. Small animals could have larger brains, but if the adaptive argument is correct, a relatively large brain in a small animal might be a disadvantage. Brains are costly to produce and maintain (they must be supplied with blood and nutrients) and so animals are unlikely to evolve 'spare capacity.'

One of the strongest pieces of evidence that a comparative relationship represents an adaptive trend is to show that it has arisen independently on more than one occasion. For instance, if we examine the relationship between brain size and body size separately among the different 'orders' of mammals, we get more or less the same result in each group (Figure 1(c)). These different orders represent groups of mammals such as the primates, the carnivores, the bats, or the artiodactyla (grazing animals), which in each case have evolved from a common ancestor that they do not share with species from the other groups. Knowing this, we can say that the basic relationship between brain and body size has arisen independently in each group, and this gives us confidence that it is a general feature of mammalian evolution.

This way of thinking about comparative relationships recognizes that evolution is a hierarchical process such that over evolutionary time, species evolve and give rise to daughter species, which in turn give rise to more species. Darwin called this 'decent with modification,' and the collection of such ancestral–descendant relationships it gives rise to is described by a phylogenetic tree. Figure 1(a) presents such a tree for the mammals, showing how all mammals derive from a common ancestor that might have lived about 165 Ma. Mammalian evolution then proceeded through a series of branching points that eventually gave rise to the modern day orders of mammals and eventually to the contemporary mammal species. At each branching point, all of the descendant species descend from a common ancestor that lived at that point. As we come up to the present, pairs of closely related species – such as two primates – will have shared nearly all of their evolutionary history, while pairs of distantly related species – such as a rodent and a primate – will have shared far less.

The statistical methods that together define the field of comparative methods seek to identify independent instances of comparative trends such as in Figure 1(c), and to do this must account for the shared evolutionary histories that the phylogenetic relationships imply. This is what distinguishes the phylogenetic comparative method from conventional statistics. To do this, comparative methods explicitly incorporate information about the phylogeny into their statistical tests.

Most comparative studies involve one or both of two broad classes of data, and these different kinds of data often require different statistical models. *Continuously varying traits*, as their name suggests, can adopt a range of values along a continuously varying scale. Brain size, height, lifespan, running speed, and gestation length are all continuously varying traits. *Discrete traits*, on the other hand, can adopt only a finite number of different states, and they need not be ordered in any way. Common discrete characters include such things as mating systems (monogamy, polygyny, polyandry) or perhaps nesting types (e.g., on the ground, in a tree, or in a hole).

Continuously Varying Traits

Species traits such as brain size, body size, generation time, and weaning age are 'continuous' or 'continuously varying traits.' Since the late 1990s, continuous data have been studied using a statistical technique known as *generalized least squares* or *GLS* (Garland, 2000; Schluter et al., 1997; Martins and Hansen, 1997; Pagel, 1997, 1999). The GLS approach for studying the evolution of continuous traits directly takes into account the phylogeny of a group of organisms and has superseded the earlier 'independent contrasts' approach (Harvey and Pagel, 1991; Pagel, 1993; Felsenstein, 1985). The latter is a computationally efficient method proposed to account for shared ancestry in comparative studies, but is not as flexible as GLS methods.

The Generalized Least Squares Approach

The GLS statistic models the evolution of continuous traits by assuming they have evolved by a process known as a *random walk*. The model assumes that traits evolve by some random amount in each of many short instants of time. At each instant of time, the change can be positive (e.g., the trait gets larger) or negative. The successive random changes to a trait are independent of each other and they are assumed to have a mean or average of zero and some degree of variability, known as the variance. If the observed differences across species are large, then this variance is also assumed to be large; if the variance is low (all species in the group have a similar trait value), then the variance is assumed to be low. In practice, the value of this variance is one of the parameters the GLS model estimates, and it can be viewed as the evolutionary rate of change for a trait (Schluter et al., 1997).

The GLS model uses this basic random walk process to describe the observed data on a phylogeny, such as the spread of brain sizes among a group of species. Because changes along the phylogeny are presumed to unfold independently at each instant of time and along each of the branches of the phylogeny, the phylogeny itself specifies how much a given species' trait value is expected to vary were evolution to be rerun, and how similar any pair of species is expected to be.

Thinking of evolution as ideally described by a random walk provides another and equivalent way of understanding why phylogenies are vital to comparative studies. Closely related species on a phylogenetic tree will have shared many of the same environments throughout their histories. If evolution is normally a slow and gradual process, then we expect closely related species on a phylogeny will tend to be similar. For example, lions

and tigers have shared all of their evolutionary history up to around 4 Ma – indeed they were the same species until that point.

Now, if we are trying to understand why some feature of an organism has evolved – such as its size, shape, or behaviors – then two closely related species will tend to share many of the potential explanations for that particular feature. For example, the plot of brain and body size data (Figure 1(b)) shows that the primates (dark blue) all tend to have large brains given the size of their bodies. If we want to understand why, the explanation could reside in just about any of the very many traits that all primates share by virtue of having all evolved relatively recently from a common ancestor. Comparative methods, by taking the underlying phylogeny into account, allow us to test ideas having removed the effects of shared evolutionary history.

Historically, the parameters of the GLS model were, and often still are, estimated by the method of maximum likelihood, but since around the beginning of this century Bayesian Markov chain Monte Carlo estimation procedures have become increasingly popular (e.g., Venditti et al., 2011). These methods do not just estimate the most probable value of some parameter, they also estimate the statistical uncertainty surrounding that estimate.

GLS Regression

Mammal Brain and Body Size

Researchers may be interested in studying the association between a continuous dependent (or response) variable and one or more independent (or explanatory) variables. In the case where there is one independent variable, this is called simple regression; where there is more than one it is called multiple regression. The evolution of mammalian brain size provides a good example of the use of the GLS regression approach to analyze comparative data. Theories of the evolution of mammalian brain size link brain volume to body mass either through body surface area or through basal metabolic rate. The 'body surface area' hypothesis supposes that brains must get larger in larger bodies to keep track of the increased number of inputs from the environment, and this predicts that the slope of the line relating brain volume to body mass is 2/3. The metabolic rate hypothesis predicts that the size of the brain is limited by the animals' metabolic turnover. This hypothesis can be shown to predict a slope of brain volume against body mass of 3/4.

Applied to the data of Figure 1(b) and the phylogeny of the mammals shown in Figure 1(a), the GLS model can be used to estimate the regression coefficient (or slope) relating brain size to body size, in addition to the variances of brain size that is explained by body size. Here the model returns a slope relating brain mass to body mass of 0.60, meaning that brain size does indeed increase more slowly than body size over evolutionary time as the theories predict. However, this slope is lower than the expected 2/3 to 3/4, despite a very strong correlation between the two traits. For this sample of mammals, the 95% confidence intervals around the estimated slope of 0.60 exclude both 3/4and 2/3.

Intuitively one might expect the brain–body size relationship to be more complex than a single slope implies, especially across a large ecologically diverse group of animals such as the mammals. Indeed using a different methodology, Pagel and Harvey (1989) showed that the nature (or slope) of the relationship changes between lower taxonomic groups in the mammals. If one constructs a GLS model that allows slopes (and y-axis intercepts) to vary between the orders of mammals (Figure 1(c)), we can see that the slopes vary from 0.50 in the Didelphimorphia (opossums) to 0.79 in the Chiroptera (bats), with primates being the second highest at 0.65.

Ancestral State Reconstruction and Phylogenetic Prediction

One of the more intriguing applications of comparative methods is to try to use the phylogeny in combination with information on the contemporary species to make guesses about what species were like in the past, or even to make predictions about the values of species' traits. Both approaches rely on the fact that the comparative method builds statistical descriptions or models of how evolution has progressed. It can then use that model to derive predictions.

In the case of ancestral states, in effect, what the methods do is rewind the statistical description of evolution to arrive at past or 'ancestral' states. For example, we can use the GLS model fitted to the mammals data to estimate that the ancestral brain and body size of mammals – the common ancestor to all of them that lived about 165 Ma – was 3.50 g and 0.31 kg, respectively. These are our best guesses as to what we think the ancestor was like, given our GLS model of evolution and the phylogenetic tree. One can use the same approach to reconstruct the size at any internal node of the phylogeny.

Phylogenetic Prediction: Dinosaur Genome Size

The genome size of birds is small in comparison with other amniote groups. Birds achieve this small genome by having a reduced number of repetitive genetic elements and less noncoding DNA – these genetic sequences do not code for proteins and are sometimes referred to as 'junk' DNA. It has been argued that large genomes are metabolically costly and so birds have evolved smaller genomes as a key innovation that made it easier to evolve or cope with the energetic demands of flight.

The difficulty with testing this idea is that birds trace their ancestry into the dinosaur lineage, but we have no way of knowing how big dinosaurs' genomes were because DNA does not fossilize. On the other hand, if we could find something correlated with genome size in modern animals, we might be able to use this relationship to predict the size of dinosaurs' genomes.

Organ et al. (2007) used the fact that animals with larger genomes tend also to have larger cells than animals whose genomes are smaller. To exploit this relationship, these authors first derived a GLS model relating cell size to genome size in a wide range of animals including the birds. They were then able to extract information on cell size in dinosaurs because, remarkably, fossilized bones preserve the size of individual cells. They then used the combination of the model relating cell size to genome size, along with the information on dinosaur cell size to make predictions about the unknown – and unknowable – dinosaur genome sizes.

The results of these analyses showed that the small genome size associated with present-day birds evolved millions of years prior to the evolution of birds themselves (see Figure 2).

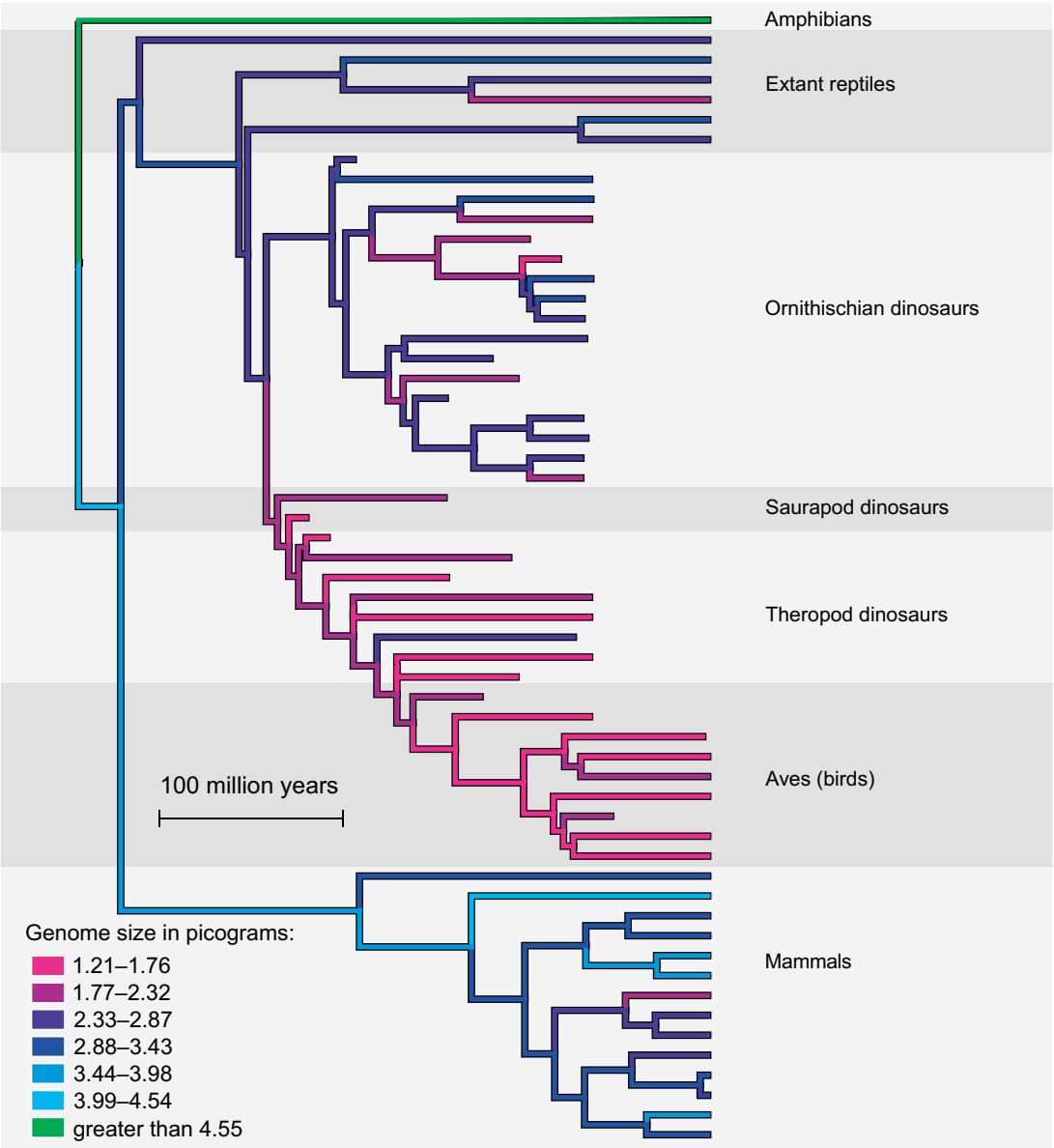

Figure 2 Broad scale amniote phylogeny that includes some extinct dinosaur species. The branches of the tree are colored according to the inferred genome size. This shows that small genomes arose millions of years before the origin of birds.

These results cast doubt on the idea that small genomes evolved to become smaller as an adaptation in response to the energetic costs of flight.

Phylogenetic Prediction: Feeding Time and the Evolution of Food Processing

Recently, phylogenetic predictions have been used in a new and interesting way – to help decide if a trait in a species is so different from what we expect as to be classified as an 'outlier.' Imagine if you believed that one species in a group was exceptional in the evolution of a particular trait. For instance, human brain size is very large – the largest for its size of any animal. To test whether we are truly outliers, though, one could construct a model of evolution (a GLS regression model in this case) for the relationship between brain and body size in mammals excluding humans. From this model it is possible to predict the brain size of humans for our body size. If the predictions are significantly smaller or larger (more likely smaller in this case) than the actual or observed trait value, then that species can be regarded as a phylogenetic outlier.

Organ et al. (2011) used this 'outlier' test logic to try to pinpoint when cooking first evolved in the hominins (*Homo*, and other members of the human clade after the split from the chimpanzees). They constructed a model relating the amount of time spent feeding each day to molar size, on the assumption that cooking, by making foods easier to process, may be associated with a small molar size and reduced time spent feeding. They then used their model to predict the amount of time spent feeding for

humans (Figure 3(a)), comparing that to the known value for our species. As expected, humans spend significantly less time feeding than predicted for a primate of their size, and we have small molars.

Hominin molars are well preserved in the fossil record and so it is possible to reconstruct where along the human lineage a significant reduction in molar size first occurred. Organ et al. (2011) found that molar size was significantly reduced in *Homo sapiens* and *Homo erectus* but not in any other hominin species. This might suggest that cooking or food processing in some form first arose in the common ancestor of *H. sapiens* and *H. erectus*, some 2 Ma (Figure 3(b)).

Phylogenetic Signal

In our description of the random walk, we stated that if evolution is normally a gradual homogeneous process, then the phylogeny predicts that closely related species will be similar.

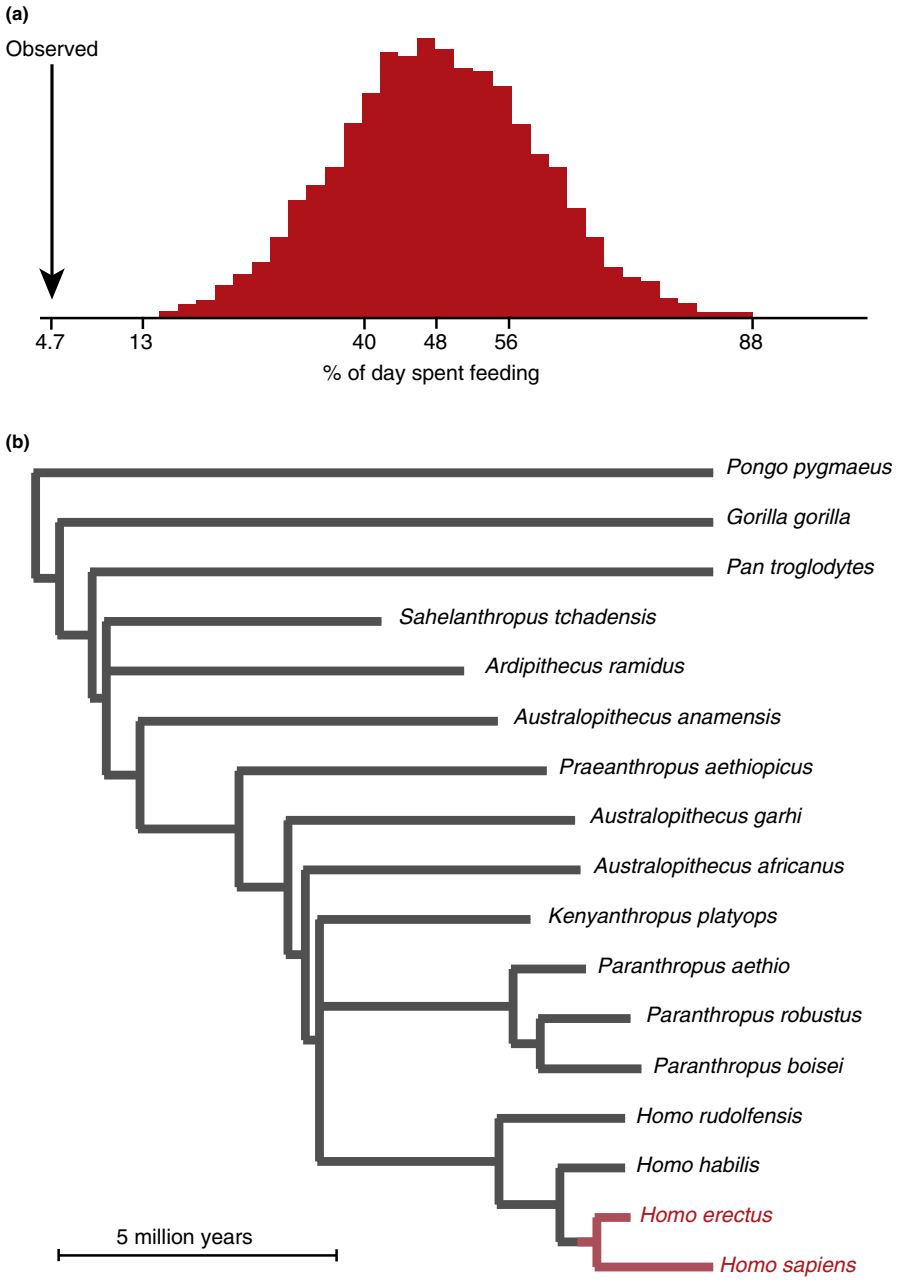

Figure 3 Feeding time in humans and the evolution of reduced molar size. (a) Shows a posterior distribution of predictions of percentage of the day spent feeding for humans. Note that the scale is logit transformed as feeding time is a percentage – this is a commonly used statistical transformation for percentages that ensures the data conform to the statistical assumption of the model. The arrow points to the actual or observed feeding time of humans – humans are a phylogenetic outlier; (b) A phylogenetic tree of hominins, including extinct species. The branches highlighted in red are those species that have undergone a significant reduction in molar size. This suggest that this reduction occurred in the common ancestor between *H. erectus* and *H. sapiens*.

But just how similar? We can imagine that for some traits even closely related species could appear quite different.

The extent to which species' similarities conform to what we expect given the phylogeny and the GLS random walk model is sometimes referred to as 'phylogenetic signal.' An important advantage of the GLS method is that it is possible to estimate the phylogenetic signal in the dataset under investigation, and then to adjust the analysis if this signal departs from expectation.

Pagel (1997, 1999) introduced a measure of phylogenetic signal, λ, which has become a routine part of comparative studies. The parameter λ varies between 0 and 1, where a 1 means the trait evolves as we would predict from the GLS model, but a zero means that the trait or traits evolve independently of the phylogenetic relationships among species. In this latter case, statistically accounting for the shared ancestry as implied by a phylogenetic tree may not be necessary.

The parameter λ should be simultaneously estimated along with other parameters in analyses to capture the historical trace of phylogenetic signal that is present in the data. This ensures that the phylogenetic contribution to the outcome is not under- or overemphasized. Some characteristics such as body size and metabolic rate have very strong phylogenetic signal. We expect other traits that might be ecological or behavioral, such as range size, feeding time, or parasite density, to have some phylogenetic signal, but that signal may not be as strong.

Other methods to detect phylogenetic signal have been proposed, including Blomberg et al.'s (2003) *K*, Abouheif's (1999) C_{mean}, and Moran's (Moran, 1950; Gittleman and Kot, 1990) *I*. A recent simulation study that examined these four most commonly used indices of phylogenetic signal (Münkemüller et al., 2012) showed that λ is the most robust measure.

Discrete Traits

The Continuous-Time Markov-Transition Process

The Markov-transition model provides a natural framework to analyze data that can adopt only a small number of distinct states (Pagel, 1994). Applied to a phylogenetic setting, the Markov approach estimates the rates at which a trait or character makes transitions among its possible states as it evolves over time along the branches of a phylogeny.

For a trait that can take only two values (e.g., 0 or 1), two rates must be estimated, one for transition from '0' to '1,' and the other for transitions from '1' to '0.' Conventionally, we assign 0 to the absence of the trait and 1 to its presence. Then, intuitively, if transitions from $0 \rightarrow 1$ proceed at a higher rate than transitions from $1 \rightarrow 0$, and if both processes are allowed to unfold along the branches of a phylogeny, then we expect to see more species with than without the trait. The continuous-time Markov model examines this distribution of 0's and 1's along with the phylogeny, and then estimates the transition rates that best explain the species' values.

The model easily generalizes to a trait with more than two values. Imagine a trait can take one of three states – 0,1, and 2 – such as might be used to categorize species into omnivores, herbivores, and carnivores. In this case, six transition rates are required to specify the 'forward' and 'backward' transitions among the three possible pairs of states. Figure 4 shows an example of a trait with four values, requiring 12 transition rates.

In addition to studying the evolution of single traits, the Markov-transition model can be used to study whether two discrete traits are coevolving. Consider two binary traits, each one of which can adopt a 0 or a 1. Then, if we consider all possible pairings of these two traits, there are four possible outcomes (0,0; 0,1; 1,0; 1,1). If we imagine we are interested in the hypothesis that when one of the states changes to a 1 the other one does as well (they change in a correlated fashion), then we expect a preponderance of species that are either in state (0,0) or (1,1).

The Markov transition model examines these states along with the phylogeny, and then constructs a model to describe their evolution. This model describes all possible transitions among the four sets of paired states. In practice, it can be shown that eight parameters are sufficient to calculate the probability of any kind of change in any branch of the tree. The method then compares the goodness of fit of this model allowing correlated evolution to the goodness of fit of a model in which the two traits evolve independently on the tree.

These principles can be applied to study the correlated evolution of traits with more than two (binary) values. However, in this situation the number of parameters or transition rates that need to be estimated usually makes this prohibitive.

The Evolution of Political Complexity

When applied to a single trait such as in these examples, the model not only estimates rates of change but can be used – as we saw with the GLS model – to estimate probable ancestral states. Political complexity varies among human cultures. Much of the Western world is dominated by states – defined by a centralized bureaucratic system that oversees specialized administrative offices – but acephalous societies (which lack political hierarchies) are widespread throughout the world. The evolutionary sequence from simple to complex societies is much debated. Do changes occur in small incremental steps or are there large nonsequential leaps? One can use the Markov approach described earlier to study this.

Currie et al. (2010) studied the evolution of political complexity in 84 Austronesian societies. A phylogenetic tree describing how these cultures are related to each other can be

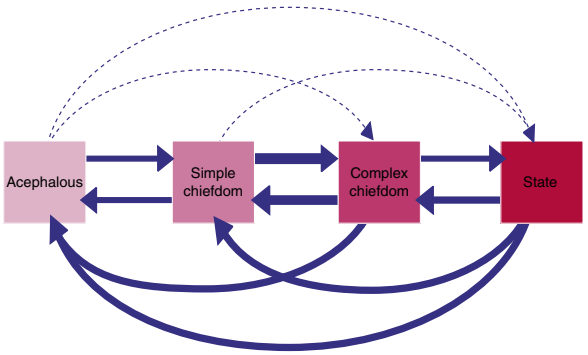

Figure 4 The estimated rate parameters for transitions between different forms of political structure. The thickness of the arrows is proportional to the estimated value of the associated transition rate. Thick lines mean high rates; dotted lines are rate that are indistinguishable from zero.

inferred from linguistic data (similarities and differences among languages or dialects); such trees are analogous to species phylogenies. The authors define levels of political complexity by the number of hierarchical decision-making levels; these are congruent with traditional social evolutionary classifications: societies lacking permanent leadership beyond the local community are called 'acephalous'; those with a single level beyond the local community are named 'simple chiefdoms'; those with two levels represent 'complex chiefdoms'; and societies with more than two levels are 'states.' They estimated the Markov transition rates between these states along the branches of the phylogenetic tree of cultures derived from language data.

Figure 4 provides a summary of the results. Political complexity always seems to accumulate from simple to more complex forms in a series of small sequential steps. By comparison, complexity can be lost in larger steps and even all at once. This study was the first to apply quantitative statistical methods to the study of the evolution of political complexity and it provide some interesting support for the idea that cultural evolution, like biological evolution, is most likely to occur via small incremental steps.

Genotypic Sex Determination and Adaptive Radiations in the Sea

The same logic that we applied to continuous traits in terms of phylogenetic prediction can be applied here for discrete traits that are correlated. If two discrete traits are associated, it is possible to predict the state of a species for which the value of some trait is not known. An example of this is the sex determination mechanism for extinct marine reptiles.

Sex determination mechanisms in reptiles can be broadly split into two categories. The first of these is genotypic sex determination, similar to what is seen in birds or mammal. The second is temperature-dependent sex determination – here the environmental temperature at some critical period of embryonic development determines the sex of an individual.

Sex determination mechanism of a species has been linked to life history and ecology but never to adaptive radiations, which are often associated with key innovations. Sex determination mechanism has the potential to be important in radiations of marine vertebrates, and the relatively stable temperature of the sea makes temperature determination problematic.

Organ et al. (2009) collected sex determination data (genetic or temperature determined) and data on whether a species bears live young (is viviparous) or lays eggs (is oviparous) for 94 vertebrate species. Using these data, they found a significant correlation between these two discrete traits. Figure 5 shows the evolutionary transition rates between sex determination and whether or not a species bears live young.

The fossil record shows that some marine reptiles (mosasaurs, sauropterygians, and ichthyosaurs) gave birth to live young (see Figure 3 in Organ et al., 2009). This information is used in conjunction with the model of correlated evolution in Figure 5 and the phylogenetic position of the species, to infer the mode of sex determination. Organ et al. (2009) showed that all these extinct lineages had a high likelihood of using genotypic sex determination.

These results mean that genotypic sex determination is present in all known fully marine vertebrate groups and support the idea that genotypic sex determination in a key innovation in producing adaptive radiations in marine environments.

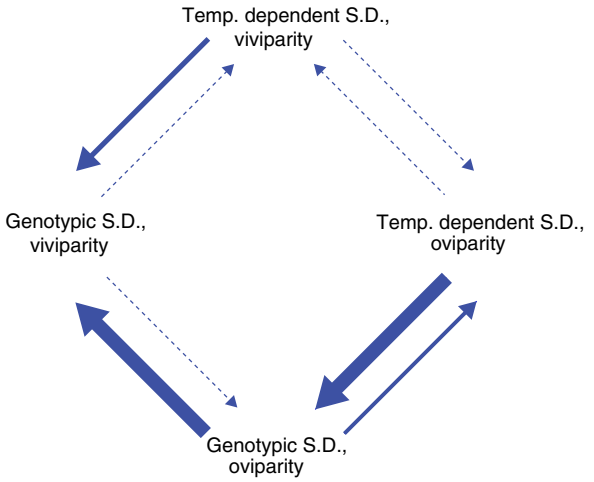

Figure 5 The evolutionary transition rates between sex-determining mechanism and reproductive mode in extant amniote species. The thickness of the arrows is proportional to the estimated value of the associated transition rate. Thick lines mean high rates; dotted lines are rate that are indistinguishable from zero. (S.D. = sex determination.)

New Developments

Researchers may be interested in investigating some evolutionary scenarios or hypothesis tests for which widely used and established comparative methods are not sufficiently developed. For example, an important distinction in the study of continuously varying traits is that between *measures* (like the traits discussed previously) and *counts*. Although these two categories may seem intuitively similar they have different statistical properties and should be treated differently in comparative studies. Many researchers attempt simple mathematical transformations of count data to make them behave more like measured data – the logarithmic transformation of count data is widely used – but this transformation can introduce biases into hypothesis tests. Rather, count data should be studied using models based on Poisson distributions (or negative binomial distributions) instead (see O'Hara and Kotze, 2010 for more details).

Another example is that researchers might be interested in studying the evolution of a binary discrete dependent variable in a regression, to ask whether the probability of being in one state or other is associated with other characteristics (continuous and discrete). For example, one might wish to investigate what ecological or life history variables explain the presence or absence of belly patches (sexual display) in phrynosomatid lizards. Such analyses can be done via logistic (Ives and Garland, 2010) or probit regression (de Villemereuil et al., 2012).

More recently, phylogenetic generalized linear models and phylogenetic generalized linear mixed models have been proposed (Ives and Helmus, 2011; Hadfield, 2010) for use when the variable to be predicted – the dependent variable – is based on counts or is a binary discrete trait in a phylogenetic context. These methods are in their infancy and are not currently widely used by the phylogenetic community. However, these will almost

certainly form the bases of a new generation of phylogenetic comparative methods for the future.

General Discussion

Comparative methods remain a popular set of techniques for investigating adaptation and the evolutionary process. New statistical approaches are easy to use and make use of more of the information in the data than previous methods. Coupled with increasingly accurate and informative phylogenies based upon genomic information and the fossil record, comparative methods can be used to test new hypotheses and to reexamine old ones.

Here we have discussed the application of comparative methods to study the evolution of traits or characteristics of species, as well as in one case the evolution of human societies. In addition to such applications, entire areas of comparative methods are devoted to using phylogenetic information to study speciation and extinction rates (see Nee, 2006; Ricklefs, 2007; Maddison et al., 2007; Venditti et al., 2010).

See also: Body, Evolution of; Brain, Evolution of; Comparative Studies: Method and Design; Cultural Evolution: Phylogeny versus Reticulation; Evolution and Language: Phylogenetic Analyses.

Bibliography

Abouheif, E., 1999. A method for testing the assumption of phylogenetic independence in comparative data. Evolutionary Ecology Research 1, 895–909.

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R.D.E., Beck, R.M.D., Grenyer, R., Price, S.A., Vos, R.A., Gittleman, J.L., Purvis, A., 2007. The delayed rise of present-day mammals. Nature 446 (7135), 507–512.

Blomberg, S.P., Garland Jr., T., Ives, A., 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57 (4), 717–745.

Currie, T.E., Greenhill, S.J., Gray, R.D., Hasegawa, T., Mace, Ruth, 2010. Rise and fall of political complexity in island South-East Asia and the Pacific. Nature 467, 801–804.

de Villemereuil, P., Gimenez, O., Doligez, B., 2012. Comparing parent–offspring regression with frequentist and Bayesian animal models to estimate heritability in wild populations: a simulation study for Gaussian and binary traits. Methods in Ecology and Evolution. http://dx.doi.org/10.1111/2041-210X.12011.

Felsenstein, J., 1985. Phylogenies and the comparative method. American Naturalist 125, 1–15.

Garland Jr., T., 2000. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. American Naturalist 155, 346–364.

Gittleman, J.L., Kot, M., 1990. Adaptation: statistics and a null model for estimating phylogenetic effects. Systematic Zoology 39, 227–241.

Hadfield, J., 2010. Mcmc methods for multi-response generalized linear mixed models: the MCMCglmm R package. Journal of Statistical Software 33, 1–22.

Harvey, P.H., Pagel, M., 1991. The Comparative Method in Evolutionary Biology. Oxford University Press, Oxford, UK.

Ives, A.R., Helmus, M.R., 2011. Generalized linear mixed models for phylogenetic analyses of community structure. Ecological Monographs 81, 511–525.

Ives, A.R., Garland, T., 2010. Phylogenetic logistic regression for binary dependent variables. Systematic Biology 59, 9–26.

Maddison, W.P., Midford, P.E., Otto, S.P., 2007. Estimating a binary character's effect on speciation and extinction. Systematic Biology 56, 701–710.

Martins, E.P., Hansen, T.F., 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. American Naturalist 149, 646–667.

Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. Biometrika 37 (1), 17–23.

Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K., Thuiller, W., 2012. How to measure and test phylogenetic signal. Methods in Ecology and Evolution 3, 743–756.

Nee, S., 2006. Birth-death models in macroevolution. Annual Review of Ecology Evolution & Systematics 37, 1–17.

O'Hara, R.B., Kotze, D.J., 2010. Do not log-transform count data. Methods in Ecology and Evolution 1, 118–122.

Organ, C., Janes, D., Meade, A., Pagel, M., 2009. Genotypic sex determination enabled adaptive radiations of extinct marine reptiles. Nature 461, 389–392.

Organ, C., Nunn, C.L., Machanda, Z., Wrangham, R.W., 2011. Phylogenetic rate shifts in feeding time during the evolution of Homo. Proceedings of the National Academy of Sciences of the USA 108 (35), 14555–14559.

Organ, C., Shedlock, A., Meade, A., Pagel, M., Edwards, E., 2007. Origin of avian genome size and structure in non-avian dinosaurs. Nature 446, 180–184.

Pagel, M., 1993. Seeking the evolutionary regression coefficient: an analysis of what comparative methods measure. Journal of Theoretical Biology 164, 191–205.

Pagel, M., 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proceedings of the Royal Society Bio 255, 37–45.

Pagel, M., 1997. Inferring evolutionary processes from phylogenies. Zoologica Scripta 26, 331–348.

Pagel, M., 1999. Inferring the historical patterns of biological evolution. Nature 401, 877–884.

Pagel, M., Harvey, P.H., 1989. Taxonomic differences in the scaling of brain on body weight among mammals. Science 244, 1589–1593.

Ricklefs, R.E., 2007. Estimating diversification rates from phylogenetic information. Trends in Ecology and Evolution 22, 601–610.

Schluter, D., Price, T., Mooers, A.Ø., Ludwig, D., 1997. Likelihood of ancestor states in adaptive radiation. Evolution 51, 1699–1711.

Venditti, C., Meade, A., Pagel, M., 2010. Phylogenies reveal new interpretation of speciation and the Red Queen. Nature 463, 349–352.

Venditti, C., Meade, A., Pagel, M., 2011. Multiple routes to mammalian diversity. Nature 479 (7373), 393–396.

Comparative Method in Linguistics

Lyle Campbell, University of Canterbury, Christchurch, New Zealand

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8929–8932, © 2001, Elsevier Ltd.

Abstract

The comparative method, the most important technique in historical linguistics, is explained, together with its basic assumptions and limitations. The comparative method compares related languages in order to reconstruct the ancestral language from which they descend; thus it serves to show that languages are related, the basis of linguistic classification. The questions of how realistic its reconstructions are and how far back in time they can reach are considered.

The comparative method is the most important of the various techniques used to recover linguistic history. It is a method (or a set of procedures) which compares related languages, descendants from a common ancestral language, in order to postulate, to 'reconstruct,' the ancestral language. It is commonly considered a major event in the history of ideas; inspired by the success of the comparative method in linguistics, numerous other disciplines in the nineteenth century adopted comparative methods of their own. In this article the comparative method is explained and its basic assumptions and limitations are considered.

The comparative method is the most important of the various techniques used to recover linguistic history. It is a method (or a set of procedures) which compares related languages, descendants from a common ancestral language, in order to postulate, to 'reconstruct,' the ancestral language. It is commonly considered a major event in the history of ideas; inspired by the success of the comparative method in linguistics, numerous other disciplines in the nineteenth century adopted comparative methods of their own. In this article the comparative method is explained and its basic assumptions and limitations are considered.

The Comparative Method and Language Families

Language which belong to the same family are said to be genetically related – this means that they derive from a single ancestral language, called a proto-language. Genetic relationships among languages are talked about in genealogical terms: 'sister languages,' 'daughter languages,' 'parent language,' and 'language families' (definitions below). Through linguistic changes, dialects develop in a language, and then as further changes accumulate in time, these dialects become distinct languages, related to one another ('sister languages') and descendants ('daughters') of the original language (the parent or proto-language). By applying the comparative method to related languages which descend from an original ancestor language, linguists postulate what that proto-language was like – that is, they reconstruct that language. For example, a comparison of English with its relatives, Dutch, German, Danish, Swedish, Icelandic, and so on, reveals what the proto-language, in this case called 'Proto-Germanic,' was like. Thus, English is, in effect, a much changed 'dialect' of Proto-Germanic, having undergone successive linguistic changes to make it what it is today, a different language from Swedish and German and its other sisters, each of which underwent different changes of their own. Therefore, every proto-language was once a real language regardless of whether it is successfully reconstructed or not, and all languages which have relatives have a history which classifies them in language families.

The aim ofthe comparative method is to reconstruct as much as possible of the proto-language by comparing its daughter languages (the related descendant languages), and to determine what changes they underwent as they diversified from their common ancestor. The comparative method serves two purposes: to show that languages are related, and to reconstruct as much as possible of the parent language. If reconstruction is successful, it shows that the initial assumption that the languages are related was warranted, and it reveals the nature ofthe proto-language.

Of course, success is dependent upon the extent to which evidence of the original traits is preserved in the descendant (daughter) languages and upon how astute the linguist is at applying the comparative method to them. For example, by comparing Romance languages (Spanish, Portuguese, French, Italian, etc.), the linguist reconstructs Proto-Romance; Proto-Romance is equivalent to the spoken language at the time Latin began to split up into its descendants. If successful, what is reconstructed for Proto-Romance by the comparative method should coincide with the Proto-Romance which was actually spoken at the time before it split up into its daughter languages. In this case, since Latin is abundantly documented, it is possible to check to see whether what is reconstructed by the comparative method accurately approximates the spoken Latin that is known from written sources at that time. However, the possibility of checking reconstructions in this way is not available for most language families, for whose proto-languages there are no written records. For example, for Proto-Germanic (from which English, German, Swedish, etc. descend), there are no written records, and this language is known only from comparative reconstruction.

Often attempts to show that languages are genetically related or to reconstruct a proto-language begin with recognition of similarities among the languages being compared. However, similarities among languages can be due to a number of things, for example to borrowing, accident, onomatopoeia, etc., in addition to inheritance from a common ancestor (see Campbell, 1998: pp. 311–326). Therefore, to show languages are related or to reconstruct successfully, similarities which are explained by factors other than common inheritance need to be

avoided. The comparative method, aided by regular sound correspondences, helps to eliminate these factors.

Application of the Comparative Method

To be able to apply the comparative method, one needs to know some technical terms

Proto-language. (1) The once spoken ancestral language from which daughter languages descend; (2) the language reconstructed by the comparative method which represents the ancestral language from which the compared languages descend. (To the extent that the reconstruction is accurate and complete, (1) and (2) should match.)

Sister language. Languages which are related to one another by virtue of having descended from the same common ancestor (proto-language) are sisters; that is, languages which belong to the same family are sisters to one another.

Cognate. A word which is related to a word in sister languages by reason of these forms having been inherited by these sister languages from a common word of the proto-language from which the sister languages descend.

Cognate set. The set of words which are related to one another in the sister languages because they are inherited and descend from a single word of the proto-language.

Sound correspondence. In effect, a set of 'cognate' sounds; the sounds found in the related words of cognate sets which correspond among related languages because they descend from a common ancestral sound. (A sound correspondence is assumed to recur in various cognate sets.)

Reflex. The descendant in a daughter language of a sound of the proto is said to be a reflex of that original sound; the original sound of the proto-language is said to be reflected by the sound which descends from it in a daughter language (cf Campbell, 1998: p. 112).

In the application of the comparative method, first potential cognates among related languages (or among languages for which there is reason to suspect related-ness) are sought. It is convenient to begin with cognates from 'basic vocabulary' (body parts, close kinship terms, low numbers, common geographical terms), since these tend not to be borrowed from other languages. Ultimately, it is the systematic sound correspondences discovered in the application of the comparative method which demonstrate which are true cognates. Words which are similar due to accident, borrowing, and other non-genetic factors usually do not exhibit the recurring sound correspondences that true cognates do. Some cognates from Greek, Latin, and English (members of the Indo-European family of languages) are illustrated in Table 1.

Next, the sound correspondences are sought. For example, in the cognates of Table 1, the first sound in each of the cognate sets exhibits the sound correspondence.

Greek *p*, Latin *p*, English *f*.

Table 1 Indo-European cognates.

Greek	*Latin*	*English*	
pod-	ped-	foot	'foot'
pater	pater	father	'father'

Since the sound correspondence recurs in these (and many other) cognate sets, it is unlikely that it could come about by sheer accident.

When the sound correspondences have been determined, the next step is to reconstruct the original sound. Linguists postulate what the original sound in the proto-language most likely was based on the phonetic properties of the sounds in the daughter languages as represented in the sound correspondence set. General guidelines help determine the most realistic reconstruction. For one, the known direction of certain sound changes is a valuable clue. For example, many languages have changed *p* to *f*, but change in the other direction, *f* to *p*, is extremely rare. In this case, it is said that the expected direction is *p* to *f* (and not *f* to *p*). This being the case, for the sound correspondence of Greek *p*-, Latin *p*-, English *f*-, **p* is reconstructed, since a change **p* to *f*, which would be necessary to derive the English reflex of *f*, is in an expected direction. (Note that the asterisk, as here with **p*, symbolizes reconstructed things.) If **f* were postulated for the proto-sound from which this sound correspondence descended, it would be necessary to assume the change of **f* to *p* in Greek and Latin, a change which goes against the expected directionality. Therefore, **p* is reconstructed in this case, together with the assumption that English underwent the common sound change of **p* to *f*.

Another guideline is the principle of 'majority wins': unless there is evidence to the contrary, for the proto-sound one reconstructs a sound coinciding with the particular sound in the correspondence set which is found in the greatest number of daughter languages. In this example, since *p* is found in both Greek and Latin, but *f* is only in English, 'majority wins' selects **p* for the reconstruction. Common sense is behind this; it is less likely that the same change (here **f* to *p*) would take place independently several times (here in Greek and in Latin) than that a single change would take place only once (**p* to *f*, only in English). The interpretation of the majority, however, can be complicated, requiring caution in the application of 'majority wins.'

This process of reconstructing sounds for the various correspondence sets is repeated until all the sounds of the proto-language have been reconstructed and the changes in the daughter languages worked out.

When all the proto-sounds have been reconstructed, it is possible to reconstruct words of the proto-language. For example, in the cognate set Greek *pod*, Latin *ped*, English *foot* 'foot,' **p* has already been reconstructed as the first sound of this word in the proto-language, based on thep-, *p*-,*f*- correspondence set just considered. For the second sound (*o*, *e*, *oo*), **e* is reconstructed on evidence not available in this short example; the correspondences for the vowel are complicated. For the last sound of the word, evidence from other cognate sets shows that the sound correspondence, Greek *d*, Latin *d*, English *t*, recurs and that the best reconstruction for this sound is *t. Putting these reconstructed sounds together, **p* for the first sound, **e* for the second, and **d* for the third, gives the reconstruction **ped* 'foot,' a word in the proto-language. In this way the vocabulary of the proto-language can be recovered.

The reconstruction of a sound, a word, or large portions of a proto-language is, in effect, a hypothesis (actually a set of interconnected hypotheses) concerning what those aspects of the proto-language must have been like. Hypothesized

reconstructions can be tested; sometimes they prove wrong; sometimes they must be modified, based on new insights. These insights may involve new interpretations of the data already on hand or new information that may come to light. The discovery of a previously unknown language may provide new evidence, a different testimony of the historical events which transpired between the proto-language and its descendants, which can change how parts of the proto-language are viewed. For example, with the discovery and decipherment of Hittite (and other languages of the Anatolian branch of Indo-European), the whole picture of Proto-Indo-European phonology changed.

There are numerous general treatments of the comparative method; for some of the clearer accounts, *see* Anttila (1989), Arlotto (1981), Campbell (1998), Crowley (1997), Hock and Joseph (1996), and Trask (1996).

Assumptions and Limitations of the Comparative Method

There are limitations on what can be recovered of a proto-language via the comparative method. Traditionally, some of these limitations are called 'assumptions,' since they are inherent consequences of how the comparative method is structured and how it is applied.

There are four main 'assumptions'

(1) The proto-language was uniform, with no dialects or variation.

Clearly this 'assumption' is unrealistic, since all known languages have regional or social variation, different styles, and so on. However, it is not so much that the comparative method 'assumes' no variation; rather, it is just that there is nothing built into the comparative method which would allow it to address variation directly. This means that what is reconstructed will not recover the once spoken proto-language in its entirety, that is, with all its variation. Still, rather than stressing what is missing, linguists are happy that the method provides the means for recovering as much of the original language as it does. This assumption of uniformity is a reasonable idealization; it does no more damage to understanding of the proto-language than, say, modern reference grammars do which concentrate on a language's general structure, typically leaving out consideration of regional, social, and stylistic variation. Moreover, dialectal variation is not always left out, since in some cases scholars do reconstruct dialect differences to the proto-language based on differences in daughter languages which are not easily reconciled with a single uniform starting point.

Assumptions (2) and (3) are related:

(2) Language splits are sudden.
(3) After the split up of the proto-language, there is no subsequent contact among the related languages.

These assumptions are a consequence of the fact that the comparative method addresses directly only material in the related languages which is inherited from the proto-language and has no means of its own for dealing with borrowings, the results of subsequent contact after diversification into related languages. Borrowing and the effects of language contact are, however, not neglected. Rather, it is necessary to call upon other techniques which are not formally part of the comparative method for dealing with borrowing and the results of language contact. It is true that the comparative method contains no means for addressing whether a language gradually diverged over a long period of time before ultimately distinct but related languages emerged, or whether a sudden division took place with no subsequent contact between the parts of the original language community. Usually, however, reconstruction by the comparative method is reasonably successful, regardless of how slow or sudden the split up of the proto-language may have been.

(4) Sound change is regular.

This assumption is true; it is not a limitation, but an asset, extremely valuable to the application of the comparative method. For example, in the example above, Proto-Indo-European *p regularly changed to f in English in all the words of this kind which formerly contained a *p. Knowing that a given sound in given circumstances changes in a regular fashion permits reconstruction of what the sound was like in the parent language. If a sound could change in arbitrary, unpredictable ways, it would not be possible to determine from a given sound in a daughter language what it may have been in the parent language, and, for a particular sound of the parent language, it would not be possible to determine what its reflexes in its daughter languages would be. If, for example, an original *p of the proto-language could arbitrarily for no predictable reason become f in some words, γ in others, k in still others, and so on, then it would not be possible to reconstruct it. In such a situation, comparing, say a p of one language with a p of another related language would be of no avail, if the p in each could have come in unpredictable ways from a number of different proto-sounds.

How Realistic Are Reconstructed Proto-Languages?

The success of any given reconstruction depends on the material at hand with which to work, and the ability of the comparative linguist to figure out what happened in the history of the languages being compared. In cases where the daughter languages preserve clear evidence of what the parent language had, a reconstruction can be very successful, matching closely the actual spoken ancestral language from which the compared daughters descend. However, there are many cases in which all the daughter languages lose formerly contrasting sounds and lose grammatical categories through changes of various sorts. It is not possible to recover via the comparative method things in the proto-language if the daughters simply do not preserve evidence of them. In cases where the evidence is severely limited or unclear, reconstruction often results in omissions or mistakes. The comparative linguist makes the best inferences possible, based on the evidence available and on everything known about the nature of human languages and linguistic change. It is necessary to make the most of what is available to work with. Often the results are very good; sometimes they are incomplete, occasionally wrong. In general, the longer in the past the

proto-language split up and the more linguistic changes that have taken place in the daughter languages, the more difficult it becomes to reconstruct with full success.

A comparison of reconstructed Proto-Romance with attested Latin provides a telling example. It is possible successfully to recover a great deal of formerly spoken Proto-Romance via the comparative method. However, the modern Romance languages for the most part preserve little of the former noun cases and complex verb inflections which Latin had. Subsequent changes have obscured these aspects of the grammar so much that large parts of Proto-Romance are not reconstructible by the comparative method. On the other hand, a great deal of Proto-Romance is recoverable.

An area of speculation and controversy concerns the outer limits of the comparative method. That is, is there a recognizable time depth beyond which the comparative method ceases to be effective? Some suggest that the comparative method cannot reach beyond the oldest currently reconstructed proto-languages; this cut-off date beyond which the comparative method is assumed not to be able to reach is variously speculated to be 6000, 8000, or 10 000 years. Since there is no reliable means of dating languages splits, these dates are rough guesses. The true issue is not really about dates at all, but rather about how much evidence related languages preserve upon which to base a reliable reconstruction. Typically, languages separated by many millennia have had longer in which to undergo changes, making them more different from one another than languages separated only by a few 100 years. Languages which have undergone many changes conceivably could preserve little of their original vocabulary or grammar; however, at the same time, it is also possible for a language to undergo relatively little change over relatively long periods. Thus, the upper limit of the comparative method is a function of whether accumulated changes have obliterated evidence in related languages, and not directly a matter of how long they have been separated.

See also: Areal Linguistics; Internal Reconstruction in Linguistics.

Bibliography

Anttila, R., 1989. Introduction to Historical and Comparative Linguistics. second, rev. ed. Benjamins, Amsterdam.

Arlotto, A., 1981. Introduction to Historical Linguistics. University Press of America, Washington, DC.

Campbell, L., 1998. Historical Linguistics. MIT, Cambridge, MA.

Crowley, T., 1997. An Introduction to Historical Linguistics, third ed. Oxford University Press, Auckland.

Hock, H.H., Joseph, B.D., 1996. Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics. Mouton, Berlin.

Trask, R.L., 1996. Historical Linguistics. Arnold, London.

Comparative Research in Education: IEA Studies

John P Keeves, The School of Education, University of Adelaide, Adelaide, SA, Australia
Petra Lietz, The Australian Council for Educational Research (ACER), Adelaide, SA, Australia

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.P. Keeves, volume 4, pp. 2421–2427, © 2001, Elsevier Ltd.

Abstract

During the second half of the twentieth century, the globalization movement transformed the nature of comparative research in education. The United Nations Organization and its agency UNESCO, promoting cooperation among nations through education, science, and culture, led this development through 'Education for All' and the integration of organizational, international, comparative, and global approaches to education. The conduct of educational research worldwide has been greatly influenced by the International Association for the Evaluation of Educational Achievement through the monitoring of educational outcomes and the introduction of systemic and modeling approaches to research in education leading to evidence-based making of educational policy.

Introduction

Comparative research in education is closely related to the field of comparative education. However, this field has been extended and greatly restructured by the movements that have operated during the second half of the twentieth century and the first decade of the twenty-first century involving initially internationalization and more recently globalization. The introduction of a worldwide program by the United Nations Educational Scientific and Cultural Organization (UNESCO) of 'Education for All' (EFA) has been concerned not only with the promulgation of basic education in developing countries (*see* Affordable Private Schools in Developing Countries), but also with the advancement of education at all levels in the more developed countries. These movements have been fed and greatly strengthened not only by the findings of research in the many fields of education, and by developments in information and communications technology (ICT) but also by the many agencies of the United Nations Organization. Each of these movements of comparative education, international education, and global education, as well as the organizational and developmental programs of UNESCO have employed their own strategies of research, that appear to be historically and structurally independent and different from each other. This article considers the four different components of a field of education that has experienced remarkable growth since the end of the Second World War when the United Nations Organization was established. In addition, this article considers the many different types of inquiry and research studies in education that have been undertaken with a cross-national, comparative, and developmental perspective. As the generative example of this research activity, scholarly inquiry, and developmental endeavor, the studies conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA) are examined from both theoretical and methodological perspectives, and the key findings of IEA studies are listed and briefly discussed. The concluding section of this article addresses the issues arising from the advancement of this field of research in education with a strongly systemic as well as a more traditional analytical approach, and considers the nature of comparative research and inquiry in the future.

Historically, comparative education has identified as its founding father Marc-Antoine Jullien (1817), whose seminal work was published in 1817, and in which he emphasized the collection of documents, objective observation, and thoroughness in recording and systematic analysis. At that time, the population of the world had barely exceeded 1 billion people. In the decade prior to the Second World War, the population of the world passed the level of 2 billion people. The world's population now exceeds the 7 billion mark. This explosion in the world's population led to a remarkable growth in formal education, that Connell (1980, p. 3) argued scarcely existed for the majority of the people in the world at the beginning of the twentieth century outside the more developed and industrialized countries of Western Europe, some British colonies, such as Australia, Canada, and New Zealand, as well as the United States. Subsequently, an expansion of tertiary education occurred and the ideas of lifelong and recurrent education were advanced, together with education for early childhood development.

The growth that has taken place in higher education during the period from 1960 onward has been reported and discussed on a global scale by Foskett and Maringe (2010, p. 307). It had been estimated that in 1960, the number of students in higher education was 13 million. It grew to 28 million by 1970, 68 million by 1991, and had exceeded 150 million by 2010. Moreover, by 2030, the student population in higher education could realistically be considered to rise to 250 million. This expansion is taking place as a result of the recognition by national governments of the need for economic and technological development in response to population growth and the necessary advancement of higher education to meet these needs.

The demand for university places in countries like China, India, Indonesia, and Malaysia now greatly exceeds the number of places available. As a consequence, there is a marked trend that involves international education for students to move from their homelands to countries where places can be found

in the institutions of higher education that are being expanded to meet the demand. Moreover, there is a similar movement at the upper secondary school level for students to move to countries where they can be educated, and can compete for entry to universities of choice. Kell and Vogl (2012) have documented the extent of transnational student mobility, particularly in the Asia-Pacific region, and have emphasized the global nature of the phenomenon.

The forces of globalization have thus reoriented the fields of comparative and international education from affording opportunities for a few privileged adults to travel in order to investigate and study how education is provided and conducted in other countries. International education now involves very large numbers of students, together with their parents, making comparative judgments about the countries in which students seeking school or higher education can best study.

Globalization and the Leadership of the United Nations

Following the end of the Second World War, the United Nations Organization was established in 1945, together with many specialist agencies, including the World Bank, the United Nations Development Program (UNDP), and the United Nations Children's Fund (UNICEF). In addition, a specialist agency was founded in 1946 to promote cooperation among nations through education, science, and culture and the promotion of peace between the countries and states of the world (Huxley, 1946) – the UNESCO. These different agencies all made the initial moves toward globalization in their specific areas of interest. While UNESCO encountered problems from time to time in the working together of its approximately 200 members, it is widely acknowledged that the forming of unity among nations is a major task that is dependent on educating the now 7 billion people living in the world. Moreover, only through programs for educating in literacy, particularly among women, can the world's population be contained, rather than distributing food and aid to those in need, or lending money to support economic advancement. In addition, the World Bank strengthens development for the reduction of poverty through microfinance programs to enable women and their families to become independent of aid. Both agencies operate within a worldwide context. These developments lead both directly and indirectly, to the advancement of globalization that is erroneously viewed by many people solely in economic terms. Giddens (1999), in discussing the *Runaway World: How Globalization is Reshaping Our Lives*, argued that globalization involved increasing social, political, and ideological interdependence among nations. Giddens (1999, p. 10) stated

> I would have no hesitation, therefore, in saying that globalisation, as we are experiencing it, is in many respects not only new, but also revolutionary. Yet, I don't believe that either the sceptics or the radicals have properly understood either what it is or its implications for us. Both groups see the phenomenon almost solely in economic terms. This is a mistake. Globalization is political, technological, and cultural as well as economic. It has been influenced above all by developments in systems of communication, dating back only to the late 1960s.

Thus, the phenomenon of globalization stands alongside comparative education and international education as a movement supported by the United Nations Organization and its agency UNESCO that has established 'global education' as a major force in today's world (*see* Globalization in Education).

A Model of Worldwide Activity in Comparative Education

Comparative education and international education now operate alongside global education and globalization at all levels of primary, secondary, higher, technological, and lifelong systems of education in all countries of the world. No longer is it appropriate to view only comparative and international education as the two sectors of a whole. The field under consideration appears to be best viewed as a tetrahedral model with four sides as is shown in Figure 1 that provides an unfolded 'net' of a three-dimensional structure.

This model is discussed in the four sections that follow with respect to the nature of the educational aspects of the worldwide activity in the traditional field of comparative education, together with the types of research that are undertaken in each specific sector. The four sides of the tetrahedral model are:

1. Organizational – with developmental and monitoring research;
2. International – with descriptive and structural research;
3. Comparative – with analytical and comparative research; and
4. Global – with systemic and modeling approaches to research.

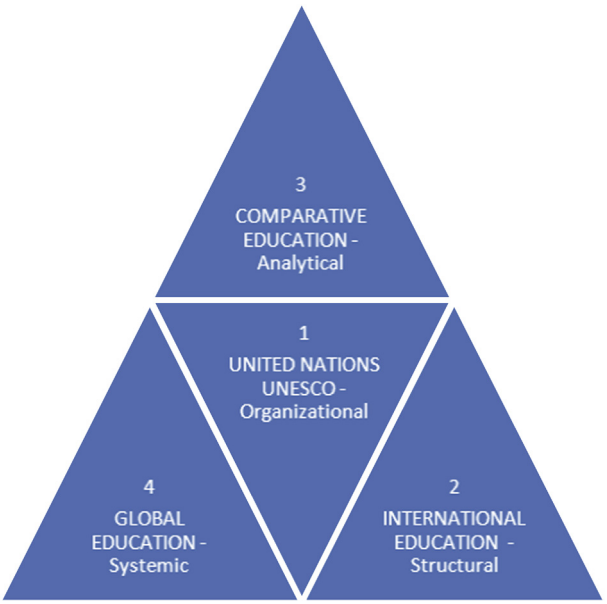

Figure 1 An unfolded model of worldwide activity in comparative education.

The IEA has now been in operation for more than 50 years, during which it has been engaged in all four approaches to research. Consequently, its research activities can be used to illustrate what is argued to be replacing the traditional field of comparative research.

Organizational and Developmental Education

UNESCO as a worldwide organization has collected information on the provision of educational services across the world through its institutes, namely, the International Institute for Educational Planning (IIEP) in Paris, the International Bureau of Education (IBE) in Geneva, and the UNESCO Institute for Education (UIE) in Hamburg. These three institutes have also maintained a scholarly and systematic orientation in their studies of education across both the developed and developing countries of the world. Two very important publications from UNESCO, *Learning to Be: The World of Education, Today and Tomorrow* (Faure, 1972) and *Learning: The Treasure within* (Delors, 1996), have provided the bases for informed debate both within and between nations on the future of education from a world perspective. These publications emphasized the idea of lifelong learning in a world context that included both developed and developing countries in which seven tensions were operating in a world where globalization was rapidly taking place, between

1. the global and the local;
2. the universal and the individual;
3. tradition and modernity;
4. long-term and short-term considerations;
5. the need for competition and concern for equality of opportunity;
6. the extraordinary expansion of knowledge and the capacity of human beings to assimilate it; and
7. the spiritual and the material.

The monitoring of the change caused by globalization in education across the world is currently a major challenge for research in education. Only with soundly based information on the magnitude and direction of change that is provided through the use of strong and well-recognized indicators can policy making and planning lead to avoiding the major catastrophes that could be caused by the use of nuclear weapons, the population explosion, degradation of the biosphere, and widespread death from disease and famine.

The World Conference on Education for All

In 1990, a World Conference on 'Education for All' was held at Jomtien in Thailand. This conference was a turning point in the globalization of the processes of education, and was a joint initiative of four world agencies, namely, UNESCO, UNDP, UNICEF, and the World Bank. The conference did not just consider initial education and the basic skills of literacy and numeracy. It sought to achieve, by the turn of the century in the year 2000, six important targets:

1. expansion of early childhood care and development;
2. universal access to completion of primary education;
3. improvement in learning achievement;
4. increase in adult literacy rates;
5. expansion of provision of basic education and training in essential skills required by youth and adults; and
6. increased acquisition by individuals and families of the knowledge, skills, and values necessary for better living and an improved quality of life (Maclean and Vine, 2003, p. 18).

In 2000, a World Education Forum was held at Dakar in Senegal to carry out a comprehensive review of what had been achieved in the different regions of the world where UNESCO centers operated to introduce and support these objectives during the previous decade. No region of the world could claim with the necessary evidence to have achieved its goals. Nevertheless, this program of purposeful and deliberate activities, operated on a worldwide front in all the many different regions through the governmental involvement of three-quarters of the countries of the world. This was a major educational revolution since it was the first occasion on which a great majority of the nations of the world had agreed through the introduction of educational policy that the provision and nature of education would be changed through specific programs. Education was clearly recognized as the only way through which a majority of the peoples of the world both as national groups and indirectly as individuals could agree to advance the quality of life of all people who were living on planet Earth.

Research within Organized Developmental Programs

Within organized programs undertaken on a worldwide basis, the type of research activity that is considered to be required involves the introduction and the monitoring of the changes in the levels of performance of the people for whom the developmental programs are provided. This strategy of the monitoring of educational outcomes, can be linked to the systematic and regular collection of information, or can involve conducting special assessment and evaluation studies in order to obtain the necessary information, before, during, and after a planned program is conducted. The monitoring of educational outcomes is considered in greater detail in a later section.

International Education

During the early decades of the sixteenth century, Erasmus, a scholar from Rotterdam traveled between the great universities of Europe – Rome, Florence, Venice, the Sorbonne, Padua, Cambridge, and Oxford – and wrote prolifically on his conversations and from his readings (Davis, 2010). He can be said to have laid the foundations for the internationalization of education that has since flourished as the institutions of higher education have spread across the world. Today, international education in universities, academies, and centers of research extends across a wide range of activities involving learning, teaching, and scholarly publication in books, journals, and the World Wide Web. These include

1. collaborative partnerships in research,
2. exchange of students and staff between institutions and countries,
3. recruitment of international students from other countries,

4. design of curricula together with changes in content, teaching methods, objectives, and assessment procedures,
5. preparation and production of encyclopedias, books, and articles,
6. conduct of offshore teaching and learning programs, and
7. widespread use of ICT for instruction, the dissemination of information and debate on ideas and policies (Maringe, 2012, pp. 17–34).

International education has led to the establishment of equivalent standards of performance in the award and mutual recognition of qualifications with a common degree framework and a common credit transfer system not only between institutions but also between countries as is provided in the Bologna Process across Europe. Moreover, this aspect of internationalization has had an impact on school systems with the establishment of the International Baccalaureate, and the use of language proficiency tests, particularly in English-speaking countries, as well as the development of aptitude tests for use across countries for university entry.

A further outcome of international education has been the employment of a variety of schemes for the ranking of institutions, particularly between countries, on many different criteria in order to identify the quality standards and reputations of the institutions involved. Inevitably, rankings, such as those produced by the Shanghai Jiao Tong University and by the *Times Higher Education* have been used for purposes of marketing. Such rankings have covered criteria such as international orientation, student support, graduate employability, third party funding, research output in terms of scholarly publications, and higher degree graduates. However, such rankings were highly contested as the information on which they were based was argued to be inaccurate, incomplete, or inappropriate to the disciplines involved (Usher and Savino, 2007).

Descriptive Research

While the impact of international education has been very visible at the level of higher education, it has led to the collection of information on the provision and conduct of education at all levels. The IBE in Geneva has since 1933 sought to obtain and publish statistical information on education from the now approximately 200 countries and states that were members of the UNESCO. The recording of such information with an adequate degree of accuracy, meaning, and comparability across these diverse members has been a very difficult task. Nevertheless, the *International Year-Book of Education* has served a useful purpose, particularly among the developing countries. The Organisation for Economic Cooperation and Development (OECD) has also undertaken the collection of similar statistical information from among its member countries that led to the production of *Education at a Glance*.

Comparative Education

The field of comparative education was established formally in the early decades of the nineteenth century and was advanced by the reports of Matthew Arnold and the writings of Emile Durkheim. It developed rapidly in the twentieth century with the greater ease of travel and communication and the rapid expansion of formal education at all levels. Courses on comparative education were introduced in universities, particularly in the United States and Great Britain. It was recognized that in this field of inquiry both descriptive and historical studies were inadequate. Furthermore, it was acknowledged that (1) attention solely to processes and not products; (2) the examination of qualities and the failure to consider quantities; (3) the restriction of investigation to single disciplinary areas, and the ignoring of an interdisciplinary strategy as well as policy-oriented issues; and (4) the use of a single method of inquiry, a so-called scientific method, were to overlook the variety and richness of the field. Thus, Kandel (1933) sought to ascribe the developments that occurred in education within a particular country to national character, and Mallinson (1957) contended that a so-called mental constitution influenced educational development within a country. However, Hans (1949) moved to a higher level of generality by identifying specific factors that influenced educational development, namely, (1) religious factors; (2) secular factors, such as humanism, socialism, nationalism, and democratic ideals; and (3) national factors, such as location, size, language, ethnicity, culture, economy and population. It could also be argued that the powerful ideas arising from the Darwinian concept of evolution influenced the drive for schooling in which children could learn to read, write, and count and contribute as adults to life in a modern industrial society (Connell, 1980) (*see* Evolutionary Approach to Education Science). Likewise, Karl Marx's ideas influenced educational policies and practices in a substantial section of the world, including, in particular, the former Soviet Union and China.

Analytical Research: Relationships, Processes, and Products

The complexity of the field of comparative education that includes (1) many disciplinary areas ranging across philosophy, history, sociology, demography, psychology, politics, economics, the sciences, and mathematics, to name only a few; (2) many levels of operation, involving the system, district, school, classroom, student, peer group, and home; and (3) many factors associated with the characteristics of the individual and the institution, as well as the country and region of the world present a major challenge to researchers and scholars. Moreover, education is primarily concerned with learning and development that necessarily involves the examination of change over time. Longitudinal studies operating at different levels are replacing cross-sectional studies that are associated with a specific point in time. Consequently, descriptive studies are being replaced by investigations that examine relationships, the operation of processes, and the generation over time of identifiable and measurable trends in outcomes and products. Fortunately, computer-based technology greatly extends what is possible in terms of collecting, recording, and storing data. In addition, the advances and ready access to such technology enables the systematic analysis of large bodies of data, with many variables operating at several levels that relate to substantial numbers of cases drawn from many countries of the world. Comparisons can now be made and rigorously

examined and analyzed with respect to their maximum likelihood and the results obtained in terms of probabilities using advanced statistical and stochastic procedures (Muthén and Muthén, 2012). While comparative education now has the possibility of undertaking investigations in highly complex situations in a field in which the operations and processes of interest are taking place, it currently lacks relevant theories that can be examined through cross-national empirical studies. This article in the sections that follow addresses this issue and discusses the contributions made over the past 50 years by the IEA.

Global Education

The global education movement is a very recent development in the rapid growth of education at all levels. Little has been written about it except at the level of higher education by Maringe (2010, pp. 17–34) who conducted a World Survey to identify thematical perspectives. It is argued in this present article that the global education movement has been proclaimed and supported by the United Nations Organization since it was established in 1945. This Organization and its agencies, particularly UNESCO, have established, since 1990, the major advancement of education in countries across the world. This development was made more effective by the removal of the barriers that separated the east from the west in Europe. This coincided with a World Conference on Education for All in Thailand, which initiated a program toward 'education for all' in those countries and states that formed the membership of UNESCO. Moreover, while the implementation of the program was allocated to the regional centers of UNESCO and the UNESCO Institutes, provision was made for the subsequent following up of the implementation of the program at a World Education Forum that was held a decade later, and on similar occasions that followed in later years. This program sees education not only from a systemic perspective across its component levels, but also from a perspective that involves all countries of the world. This is a major development in education across the world. Nevertheless, while Maringe's study is of considerable value, a global theory of education has, not as yet, been advanced. There appear to be three levels within a technological perspective into which countries can be placed, namely, the developed, the transitional, and the developing countries that are identified largely with respect to their levels of technological advancement. A second perspective is argued to involve political systems with respect to the operation of democracy and democratic governance. Giddens (1999, p. 70) noted that

> Since the mid-1970s, the number of democratic governments in the world has more than doubled. Democracy has spread to over 30 or more countries, while all existing democratic states have kept democratic institutions in place.

Nevertheless, the practice of democracy in the different parts of the world varies considerably.

A third theoretical perspective of globalization is the emergence of a world culture that has arisen from the rapid and widespread expansion of the use of ICT. This also appears to be linked to the growing use of English as a global language. The issue yet to be addressed is whether a world culture can be appropriately related in a meaningful way to the continued use of native languages and the rich variety of native cultures that currently exist.

A fourth perspective involves world economic integration with the gradual freeing of trade across countries and regions of the world. A controversial aspect of this perspective involves the free movement of people across borders, and thus to form a united body of people who occupy a world in which the population is continuing to grow rapidly. The issues of world peace and sustainable development can ultimately only be resolved through the spread of education across the world, particularly in the transitional and developing countries of Asia, Africa, and Latin America.

A reaction to the expansion of global mobility and development occurred, and Beck (1999) examined the processes of individualization and commodification of everyday life in the context of the 'risk society.' Subsequently, Kell and Vogl (2012, p. 181) have considered the need for "a shift in attitude to one where there is a global optimism that overcomes some of the ambivalence and anxieties about international alliances" in order to counter the consequences of the risk society. New associations or 'ensembles' involving global education are needed that are built around cooperation, reciprocity, and concord rather than merely commercial transactions. The IEA from its formative years has been just such a body in the field of comparative education research (*see* Cross-Cultural Study of Education).

The Global Monitoring of Educational Outcomes

UNESCO in the EFA program not only undertook an interim examination of the outcomes in 2000, but also initiated from the IIEP in Paris, the establishment in Sub-Saharan Africa of the Southern and Eastern Consortium for Monitoring Educational Quality (SACMEQ) in order to collect information at regular intervals in 16 countries on the effectiveness of the EFA program. Likewise, the World Bank, a partner in the EFA program, supported similar studies in countries such as Vietnam. In addition, Postlethwaite (2004) reported, in *Monitoring Educational Achievement* published by IIEP, on studies carried out by IEA in 71 countries of the World and by the Programme for International Student Achievement (PISA) set up by the OECD, in 62 countries in Europe and in other parts of the World, with some countries being involved in both IEA and PISA monitoring studies (*see* Cross-National Comparisons in Education: Findings from PISA). Furthermore, UNESCO commissioned for the EFA Global Monitoring Report in 2008 a study on *The Growth of National Learning Assessments in the World, 1995–2006* by Benavot and Tanner (2007) from IBE in Geneva. Subsequently, Kamens and McNeely (2009) discussed the relationships between *Globalization and the Growth of International Testing and National Assessment*. They drew attention to the programs funded by UNESCO in Latin American countries through the Latin

American Laboratory for the Evaluation of Quality in Education involving 15 countries and in African and the Caribbean Francophone countries through the *Conference des Ministères de l'Éducation des Pays Ayant le Français en Parlage* with 24 participating members. IEA laid the foundations for this worldwide movement and trained the leaders of the SACMEQ and PISA programs. Similar monitoring efforts have been undertaken through the Association of the Southeast Asian Nations and the Arab League (Kamens and McNeely, 2009, pp. 7, 10).

However, at a meeting in 1958 at the UIE in Hamburg, Germany, that led to the founding of IEA, the group of scholars included Georg Rasch who argued for the measurement of performance on scales that permitted absolute comparisons to be made over time, rather than relative comparisons between countries through cross-sectional surveys (Postlethwaite, 2009, p. 4). In this way, the absolute growth in educational achievement could be measured over time that resulted from the EFA program and other similar activities. This would add to the information already made available by the UNDP through the human development index and the gender-related development index.

Systemic Research and Modeling

The global education movement, which examines the high degree of complexity of the data available, demands a new approach. The data need to be considered not just with respect to their component parts, but as a whole. Traditionally, statistical analysis, as is implied in the name of the strategy employed, involves analytical thinking. However, such an approach is inadequate where a global view is required. Under these circumstances, it is necessary to employ a holistic or systemic approach to the examination of data, in which the interrelations between the components are estimated, following the identification and hypothesizing of causal effects. This involves the advancement of structural equation models that serve to explain the situation under investigation. Such models in comparative education are necessarily both multivariate and multilevel in nature, and within them, not only direct but also indirect effects operate, which indicate the mediation and moderation of relationships. Through the use of the hierarchical linear modeling computer program (Raudenbush et al., 2011), direct effects and moderating effects across levels can be examined and estimated. Using the Mplus computer program (Muthén and Muthén, 2012), not only can direct, as well as cross-level interaction or moderated effects be examined, but indirect effects that involve mediated relationships can also be hypothesized and the magnitudes of such relationships estimated and examined for significance. In addition, Mplus can take measurement error as well as latent variables into consideration.

A further aspect of systemic research involves the hypothesizing of relationships using data aggregated to the country level that come from different sources. Between country analyses of aggregated data have been examined by Hanushek and Woessmann (2011) who have reported strong relationships that not only have statistical significance, but also have highly practical significance for the making of national policy (*see* Education, Economics of).

The Research Program of IEA

In the late 1950s, a group of scholars who met at the UIE in Hamburg saw the need for comparative studies of the relationships and processes operating to influence the educational outcomes of students' achievement and attitudes. They recognized that there was generally insufficient variability within a particular country to detect the effects of specific factors on the outcomes of learning, using the existing procedures of statistical analysis. They then argued that if research studies could be undertaken between countries, as well as within countries, then the greater variability that existed across countries and the different school systems within countries would permit the identification of common factors of importance that influenced learning outcomes.

The feasibility of conducting research studies across countries into factors influencing educational achievement was tested in a pilot study in 1960. The success of this study led to the establishment of the IEA. Over a period of 50 years, IEA has carried out a sustained program of research studies, which have, in the main, been survey and cross-national investigations, into factors that have an effect on educational achievement, participation and attainment rates, and attitudes as well as views of school and school learning. The data collected have been examined using procedures of statistical analysis across countries, between schools within countries, and between students within countries, and where possible relationships have been tested after controlling for other factors that also affected the educational outcomes under consideration.

Key Findings of IEA Research Studies

Over the 50 years during which IEA studies have been in operation, when taken together, these studies have involved millions of students in testing programs, hundreds of thousands of schools and many thousands of research workers in many different fields of educational research from a very wide range of countries. Decisions regarding which studies to conduct are made by the IEA's General Assembly, which consists of member organizations from approximately 70 education systems. The financial support for IEA's program of research has varied greatly over the years, although a very strong source of financial support has been from agencies and foundations in the United States. The school subjects involved in IEA's testing programs have included mathematics, the sciences, civic education, reading and literacy, literature and writing, French and English as foreign languages, computing, as well as ICT. In addition, studies have been conducted in preprimary and primary schools as well as in secondary schools at all grade levels, covering 13-, 14-, 15-, and 16-year-old students and at the terminal secondary school stage. Furthermore, there has been a study of classroom environments, a study to investigate the teaching and learning processes involved in mathematics and science classrooms by way of video recording of classroom operations, as well as a teacher education study to examine the pedagogical knowledge and educational experiences of trainee teachers. It is difficult to identify and highlight the 'key findings' from so many studies

in so many different areas but 10 findings stand out, although some of the 10 may be considered controversial. Six of the key findings relate to the field of comparative education, and four of these key findings relate to policy issues.

Key Findings Relating to Comparative Education

1. There are marked differences in average levels of achievement between the students in school in the more developed countries and those in the developing (LDC) countries. This occurs in spite of the fact that in the LDCs less than 100% of the relevant age groups are enrolled in school.
2. The average level of achievement within a developed country at the terminal secondary school stage is inversely related to the proportion of the age group enrolled at school or participating in the study of the subject under survey.
3. In developed countries, at the terminal secondary school stage, when equal proportions of the age groups are compared, there are only small differences in levels of achievement, irrespective of the proportions of the age groups enrolled at school. Thus, the best students do not suffer through increasing retention rates.
4. Student achievement in mathematics, French as a foreign language, and the sciences, is positively related to the time given to the study of the subject at school (1) in comparisons across countries (comparative oriented), and (2) in comparisons between students within countries (policy oriented).
5. The level of reading resources in the home is positively related to student achievement, as are other indicators of language usage in the home, such as use of (1) dictionary and (2) whether or not the language of the home is the language of instruction.
6. Measures of socioeconomic status of the home are positively related to student achievement in all countries, at all age levels, and for all subjects.

Key Findings Relating to Policy Making

7. The achievement of students is related to the time spent on homework, after other factors influencing achievement are taken into account.
8. The average level of student achievement across countries is positively related to the opportunity that students have to learn the content and cognitive skills tested.
9. In developing countries, the availability and use of text books are found to have a significant effect on student learning, but the same effects have not been recorded from studies in the more developed countries.
10. Gender differences in achievement and attitudes vary in size and direction across countries, school subjects, and over time. These differences are argued to be societally and policy based.

For further information, the *World of School Learning* (Keeves, 1995) and research reports of individual studies can be consulted (e.g., Kozma, 2003; Martin et al., 2008; Montie et al., 2007; Mullis et al., 2007, 2008).

The Three Phases of the IEA Research Programs

The IEA programs of research evolved through successive stages that were related to the duration of each study. In addition, the specific research studies have been greatly influenced by the financial support obtained and by the availability of research centers that were willing to coordinate a study. While there were several stages in the program of research studies conducted by IEA, there would appear to be three distinct phases that have operated over a 50-year period. These phases have been formed by the nature, design, and purposes of the studies undertaken.

Phase 1 Analytical Research, 1960–75

The first phase may be considered to involve primarily comparative research, with an emphasis on cross-country comparisons in the initial pilot study, which tested in 1960. The First International Mathematics Study (referred to as FIMS) commenced in 1961, tested in schools in 1964, and released an international report in 1967 (Husén, 1967). In 1966, the Six-Subject Study was initiated. Testing occurred in 1970 and 1971 and the study was completed in 1975. In 1967, a conference was held at Lake Mohonk in the United States to discuss the findings of FIMS, and in 1973, a conference was held at Harvard University to discuss the implications of the findings of the first three subject areas in the Six-Subject Study.

The Lake Mohonk conference endorsed the comparative emphasis on inquiry and cross-national comparisons using analytical statistical procedures in the examination of the data collected. As a consequence, the Six-Subject Study, for which testing occurred in 1970–71, largely followed the strategy employed in the two earlier studies (Super, 1967).

Phase 2 Descriptive and Policy Research, 1975–92

From the Harvard Conference in 1973, a change in emphasis in IEA studies toward the making of educational policy emerged in the report *Educational Policy and International Assessment* (Purves and Levine, 1975). This resulted in the undertaking of studies in the areas of mathematics, for which testing occurred in 1982 (Robitaille and Garden, 1989); science, which tested in 1984, 14 years after the First IEA Science Study (Keeves, 1992), and reading, which was undertaken in 1990–91, 20 years after the first reading study (Elley, 1994). In addition, a study of the classroom environment was carried out with data collected from 1981 through to 1983 (Anderson et al., 1989).

Furthermore, a component of the Second International Mathematics Study involved a time series or longitudinal strategy in which the same sample of students was tested on two occasions in eight countries (Burstein, 1993). A similar study had been previously undertaken in Australia that was linked to the First IEA Science Study with students tested at the 13-year-old level (Keeves, 1972). Nevertheless, in the later reports and analyses of the science and reading studies, work was done on the examination of changes in levels of achievement at the system level over time, referred to as

a trend or a type of longitudinal strategy (Keeves, 1992; Lietz, 1996). These two studies used Rasch scaling procedures to examine absolute change at the system level over time.

Phase 3 Monitoring and Evaluation Research, 1992 and Ongoing

In the third phase, the direction of the major studies that were very strongly supported financially by the United States placed an emphasis on trends over time through the preparation and publication of so-called league tables. The statistical procedures employed were directed toward the very efficient estimation of national mean scores, with small estimates of error, so that the statistical significance of relative change of differences between countries could be made as part of an evaluation process. Little attempt would appear to have been directed toward monitoring and the examination of absolute change in performance between countries or on the absolute change in performance between grade levels within countries. However, this work resulted in a series of Trends in Mathematics and Science Studies (TIMSS) with testing taking place at intervals of 4 years. Likewise, a series of studies has been carried out at 5-year intervals arising from the study of reading literacy undertaken in 1990–91 (Elley, 1994) that are referred to as the Progress in International Reading Literacy Studies (PIRLS). Data collections for TIMSS and PIRLS coincided in 2011 with the international reports and databases that were released at the end of 2012 (Martin et al., 2012; Mullis et al., 2012a,b).

As an extension of TIMSS, IEA designed and implemented a video study of instruction in the mid-1990s with seven countries participating in the mathematics part and five countries participating in the science part (Jacobs et al., 2003; Stigler and Hiebert, 1999). National samples of mathematics and science lessons were videotaped in an attempt to obtain better information about teaching practices in these subject areas. These studies were highly innovative at that time.

In addition to TIMSS and PIRLS, IEA comparative research studies in other areas of education continued to be coordinated in other centers that were led by scholars in research institutes and universities who maintained a strong interest in comparative and analytical investigations into new fields that were emerging as a result of advances in technology and the drive toward democratic government that accompanied the globalization movement.

These IEA studies included (1) the International Civic and Citizenship Study in 2009, which incorporated novel regional assessment modules for Asia, Europe, and Latin America (Kerr et al., 2010; Schulz et al., 2010, 2011) and built upon the earlier Civic Education Study (Torney et al., 1976); (2) the Teacher Education Development Study conducted in 2008, which was IEA's first study of tertiary education, with a focus on policies and practices in mathematics teacher education (Tatto et al., 2008); and (3) International Computer Information Literacy Survey with a planned data collection in 2013, which builds on a tradition of IEA surveys carried out between 1989 and 2006 on ICT use in schools, but features a new computer-based student assessment of computer and information literacy.

Systemic Research and the Use of Models

In research studies that are primarily concerned with the making of decisions to guide policy and practice, specific hypotheses are advanced and tested. In a modeling approach to research, a model set within a system of operations is advanced and examined in the context in which the study is conducted. Initially, IEA's program of research was undertaken largely from the perspective of comparative education that was enriched by ideas drawn from psychology, economics, demography, sociology, and pedagogy. Before the systemic examination of educational processes, practices, and policies between students, between schools, and between countries can take place, progress must be made along three lines. First, it is necessary for theory to be advanced and models formulated or constructed of the specific processes involved in school learning. Second, it is necessary for data to be collected that are sufficiently comparable across countries, so that the examination of models in a comparative manner is meaningful. Third, statistical procedures for the examination of models must be developed in order to provide strong estimates of effects and the efficient estimates of errors for the examination of statistical significance and effect sizes.

Several models have been widely employed in comparative research studies both within and between countries in the 30 or more IEA studies carried out. From an IEA training workshop conducted at Gränna in Sweden in 1971, a model of curriculum implementation that operated at three levels was advanced. In this model shown in **Figure 2**, the curriculum can be considered to exist at three levels: (1) the intended curriculum, (2) the implemented curriculum, and (3) the achieved curriculum, which are influenced by antecedent and contextual factors as well as the learning conditions and the teaching practices operating in the schools. A second model of school learning was advanced by Carroll (1963) that has been the source of many theoretical discussions of the factors influencing educational achievement. The IEA studies provided opportunities for such discussions and for the derived models to be examined with empirical data (Carroll, 1975). Three variables were specified in the model of school learning in terms of time: (1) aptitude, which involved the amount of time a student would require to learn a task to a specified criterion level, given motivation, opportunity to learn, and optimal quality of instruction, (2) perseverance, which involved the amount of time that a student was willing to engage in active learning, and (3) opportunity to learn, which involved the amount of time provided for learning in a specific program. In addition, two variables were advocated that were not specified in terms of time, namely, (4) the ability to understand instruction, and (5) the quality of instruction provided.

A third model that has guided IEA's research has been a cross-national model of educational achievement in a national economy that was advanced by Dahllöf (1967). This model covered the (1) national economy, (2) demand for human resources, (3) objectives, (4) structural and frame factors, (5) process factors, and (6) levels of performance achieved and attainment. Walker (1967) advanced a fourth model for the effects of selection. In this model, each complete age group in a country had a similar distribution of ability and largely common curricula and common modes of

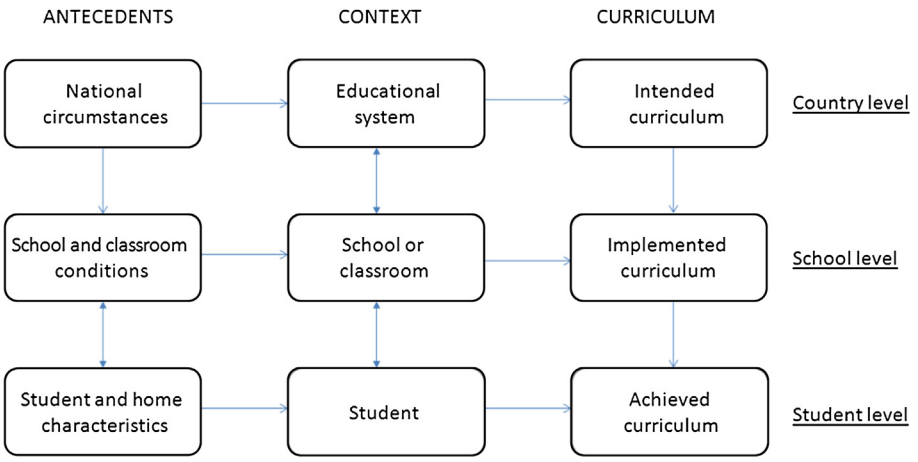

Figure 2 The context and components of the school curriculum. Source: Rosier, M.J. Keeves, J.P. (eds.), 1991. The IEA Study of Science I: Science Education and Curricula in Twenty-Three Countries. Pergamon Press, Oxford, p. 7, modified.

instruction were assumed to exist across countries. Differences in selection, retention, and participation influenced the level and spread of achievement test scores across countries as well as the yield of a school system.

In general, such models were examined with available data across the countries participating in the IEA studies. However, they were sufficiently complex that their examination was restricted by the statistical procedures available at the time. For example, in the model depicted in Figure 2, the model was examined at the student level, or at the school level or at the country level, but it could not be examined across all three levels simultaneously. However, it is now possible to examine such a model, in a single analysis across all three levels, with students nested within schools and schools nested within countries, so that the systemic structure of the model can be taken into consideration.

The IEA studies, both in sampling design and statistical analyses were initially strongly guided by Gilbert Peaker. He, having worked at Cambridge University in England, was greatly influenced by Sir Ronald Fisher, who advanced and developed the Fisherian tradition of statistical analysis. Although IEA sought assistance and advice from scholars working within a decision-making perspective in the Neyman – Pearson tradition and the Bayesian perspective, which took into consideration prior knowledge, Peaker's influence dominated the statistical analyses of the large bodies of IEA data. He adopted a stochastic view for the investigation and estimation of effects relating to educational policy and practice and the introduction of structural equation modeling into IEA studies during the first 30 years from 1960 to 1990. Only in the twenty-first century has the examination of hierarchical linear models and structural equation models started to be more widely used in the analyses of IEA data. Initially, partial least squares regression was advanced by Herman Wold, from Sweden, alongside the use of maximum likelihood procedures developed by Lawley from Scotland and by Jöreskog also from Sweden, where the IEA Headquarters were located, for the examination of models of educational processes. Still, while the estimation of the coefficients of structural models is relatively straightforward, the meaning and interpretation of these effects in both simple and complex models are more somewhat complicated as discussed by Pearl (2000) and more recently by Hoyle (2012).

Regression analysis was employed for the comparison of groups both within countries and between countries. While not feasible in the early IEA studies due to limitations in terms of modeling and computational power, more recent IEA studies have paid appropriate attention to examining the extent to which measures of both cognitive skills and contextual background constructs are comparable between countries and between subgroups within countries, as defined by, for example, gender, remoteness, or migration status. For cognitive skills test items, this tends to involve the examination of differential item functioning between countries and between male and female students. Such analyses were first undertaken for the IEA Reading Literacy Study in 1990/91 (Elley, 1994) and are continued to be applied in IEA's TIMSS and PIRLS assessment programs (see, for example, Foy et al., 2012). For background questionnaire constructs such as students' interest in the subject matter or school climate, this involves the examination of results from separate multigroup confirmatory factor analyses models where the groups can be defined, for example, by country or gender. Such analyses examine the extent to which the measurement model fits the data overall and in each particular subgroup. Four levels of measurement invariance are usually assessed, namely, configural, metric, scalar, and uniqueness invariance (see Meredith, 1993; Vandenberg and Lance, 2000). Each level of invariance increases in the restrictiveness of its assumptions from configural invariance – which holds if the same number of factors is found and the same items are associated with each factor across groups – to uniqueness invariance, which restricts the portion of item variance not attributable to variance in the underlying factor to be the same across groups.

However, it should be recognized that with data from large samples, which is the case in IEA studies, even smaller differences appear to be significant and that, therefore, using significance test to judge the level of invariance may tend to be problematic. Consequently, rather than examining the absolute

significance of the goodness-of-fit indices, relative changes in model fit across different multiple-group models with different equality constraints on parameters are reviewed (Schulz, 2009). More recently, however, Davidov et al. (2012) have shown how multilevel structural equation modeling can be applied to explain why invariance at the higher levels might be absent in large-scale cross-national research. In IEA studies, such invariance analyses have been applied, for instance, in the Civic and Citizenship Study (see Schulz and Friedman, 2011).

Comparative Research in Education in the Future

Very little work has been done by scholars in the field of comparative education to advance models for examination and the estimation of significant relationships and causal effects using the statistical programs that have been written during recent decades. However, Fägerlind and Saha (1989) advanced ideas and relationships toward the construction of a typology for education and development that was dialectical in nature with economic, political, and social, including cultural and ideological, dimensions. This model was then examined with respect to the Indonesian educational system, from a comparative perspective but without data being available for statistical analyses.

The report of the Lake Mohonk conference in 1967 considered the theoretical foundations that were seen to be required for the next stage in the IEA research program. The variability between countries in their economic circumstances was seen to be related to the differences in the provision of education in each country and to the nature and levels of achievement attained at successive levels of schooling. The comprehensive model of educational achievement advanced from the conference is presented in Figure 3 (Super, 1967).

Figure 3 was developed prior to an awareness among educational researchers of the possibilities of employing path analysis and using least squares regression procedures. The model sought to include the stages of input, process, output, and utilization in a multilevel framework with appropriate variables linked together by the flow of paths of influence. This model has given rise as well as the need to develop stochastic procedures for the analyses of large bodies of data and the estimation of models with many structural relationships. The development of Mplus (Muthén and Muthén, 2012) for the analyses of multivariate and multilevel structural equation models to investigate cross-level and within-level moderated effects together with mediated effects operating at three levels now enables the examination of complex models derived from Figure 3. Thus, this model and the model depicted in Figure 2 have guided, either explicitly or implicitly, the analyses of many IEA studies and have thus contributed extensively to the advancement of the body of knowledge in comparative education.

It is evident that the establishment of IEA has already given rise to a very substantial body of research, since over the past

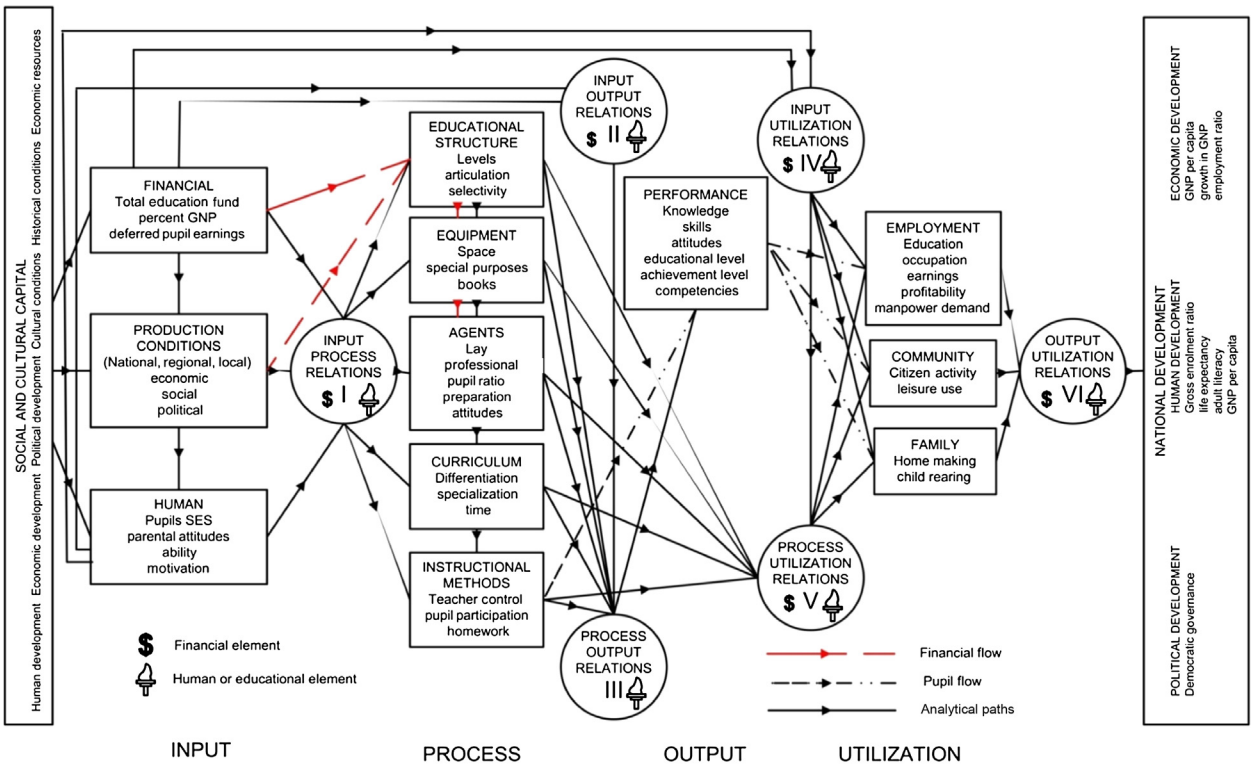

Figure 3 A comprehensive model of educational achievement. Source: Super, D. E. (Ed.), 1967. Towards a Cross-National Model of Educational Achievement in a National Economy. Teachers College, Columbia University, New York, p. xx (12) modified in Postlethwaite, T.N., 2009. Memories of IEA achievement surveys. In: Matthews, B., Gibbons, J.A. (Eds.), The Process of Research in Education. Shannon Research Press, Adelaide, pp. 1–20.

50 years IEA has conducted 31 major studies (Schwille, 2011, pp. 657–658). The fascinating story of the endeavors of thousands of people who have participated in IEA's journey in the area of comparative research has recently been documented (Papanastasiou et al., 2011). Nevertheless, there is clearly a need for rethinking the nature of research in comparative education in an international and global education context from new theoretical perspectives to serve researchers in the educational, behavioral, and social sciences as well as educators, policymakers, and teachers at all levels in the decades ahead.

Conclusion

While the Lake Mohonk model presented in **Figure 3**, together with the comparative, policy-oriented, and monitoring approaches, has guided either explicitly or implicitly the analyses of many IEA studies and thus contributed extensively to the advancement of the body of knowledge in comparative education, the advent of globalization requires a new approach.

Hanushek and Woessmann (2011, p. 192) have argued that with

> …sets of comparable achievement data for extensive samples of countries being linked to subsequent economic growth, and with the emerging long panels of regular achievement data for large samples of countries, the outlook for future research in the economics of international differences in educational achievement is clearly bright.

An interesting example that illustrates how use can be made of such data by employing procedures suggested by Hanushek and Woessmann can be found in the recent McKinsey Report (Mourshed et al., 2011), which examines how 20 school systems are able to sustain their rising levels of school achievement.

Finally, if information can be collected from other disciplinary perspectives in order to extend beyond the field of economics, and with the appropriate statistical procedures now available, the future of comparative educational research is moving into rich and fertile fields. These fields, as Hanushek and Woessmann have argued, are formed from the large monitoring studies that are currently in operation across the world. In addition, the systemic research studies that are envisaged and illustrated by the work that Hanushek, Woessmann, and their colleagues are already undertaking are comparative and systemic in nature, and not solely directed toward the making of policy or the evaluation and monitoring of existing programs of educational development in a global setting (*see* Education, Economics of).

See also: Cross-Cultural Study of Education; Cross-National Comparisons in Education: Findings from Pisa; Education, Economics of; Evolutionary Approach to Education Science; Globalization in Education; International Organisations and Education; Self-Efficacy: Education Aspects.

Bibliography

Anderson, L.W., Ryan, D.W., Shapiro, R.J. (Eds.), 1989. The IEA Classroom Environment Study. International Studies in Educational Achievement, vol. 2. Pergamon Press, Oxford.

Beck, V., 1999. What Is Globalisation? Polity Press, Cambridge.

Benavot, A., Tanner, E., 2007. The Growth of National Learning Assessment in the World, 1995–2006. Unesco, Paris.

Burstein, L. (Ed.), 1993. The IEA Study of Mathematics III: Student Growth and Classroom Processes. Pergamon Press, Oxford.

Carroll, J.B., 1963. A model of school learning. Teachers College Record 64 (8), 723–733.

Carroll, J.B., 1975. The Teaching of French as a Foreign Language in Eight Countries. Almqvist and Wiksell, Stockholm.

Connell, W.E., 1980. A History of Education in the Twentieth Century World. Teachers College Press, New York.

Dahllöf, U., 1967. Relevance and fitness analysis in comparative education. In: Super, D.E. (Ed.), Towards a Cross-National Model of Educational Achievement in a National Economy. Teachers College, Columbia University, New York, pp. 1–25.

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., Meuleman, B., 2012. Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. Journal of Cross-Cultural Psychology XX (X), 1–18.

Davis, G., 2010. The Republic of Learning. Harper Collins, Sydney.

Delors, J., 1996. Learning: The Treasure within. Report of the International Commission on Education for the Twenty-First Century. Unesco, Paris.

Elley, W.B., 1994. The IEA Study of Reading Achievement and Instruction in Thirty-Two School Systems. Pergamon Press, Oxford.

Fägerlind, I., Saha, L.J., 1989. Education and National Development: A Comparative Perspective, second ed. Pergamon Press, Oxford.

Faure, E., 1972. Learning to Be. Report of the International Commission on Education for the Twenty-First Century. Unesco, Paris.

Foskett, N., Maringe, F., 2010. The internationalization of higher education: A prospective view. In: Maringe, F., Foskett, N. (Eds.), Globalization and Internationalization in Higher Education. Continuum, London, pp. 305–317.

Foy, P., Martin, M.O., Mullis, I.V.S., Stanco, G., 2012. Reviewing the TIMSS and PIRLS 2011 Achievement Item Statistics. Boston College: TIMSS and PIRLS International Study Center.

Giddens, A., 1999. Runaway World: How Globalisation Is Reshaping Our Lives. Prolife Books, London.

Hans, N., 1949. Comparative Education: A Study of Educational Factors and Traditions. Routledge and Kegan Paul, London.

Hanushek, E.A., Woessmann, L.W., 2011. The economics of international differences in educational achievement. In: Hanushek, E.A., Machin, S., Woessmann, L.W. (Eds.), Handbooks in Economics, vol. 3. Elsevier, Amsterdam and Oxford, pp. 89–200.

Hoyle, R.H., 2012. Handbook of Structural Equation Modeling. Guilford Publication, New York.

Husén, T. (Ed.), 1967. International Study of Achievement in Mathematics, vols. 1 and 2. Almqvist and Wiksell, Stockholm. Wiley, New Work. pp. 166–135.

Huxley, J., 1946. UNESCO, Its Purpose and Its Philosophy. Unesco, London.

Jacobs, J., Garnier, H., Gallimore, R., Hollingsworth, H., Givvin, K.B., Rust, K., Kawanaka, T., Smith, M., Wearne, D., Manaster, A., Etterbeek, W., Hiebert, J., Stigler, J.W., 2003. TIMSS 1999 Video Study Technical Report, Volume 1: Mathematics Study (NCES 2003–012), US Department of Education. National Center for Education Statistics, Washington, DC.

Jullien, M.A., 1817. Esquisse d'un ouvrage sur l'éducation comparée. Colas, Paris.

Kamens, D.H., McNeely, C.L., 2009. Globalization and the growth of international education testing and national assessment. Comparative Education Review 54 (1), 5–25.

Kandel, I.L., 1933. Comparative Education. Houghton Mifflin, Boston.

Keeves, J.P., 1972. Educational Environment and Student Achievement. Almqvist and Wiksell, Stockholm.

Keeves, J.P. (Ed.), 1992. The IEA Study of Science III: Changes in Science Education and Achievement: 1970 to 1984. Pergamon Press, Oxford.

Keeves, J.P., 1995. A World of School Learning: Selected Key Findings from 35 Years of IEA Research. IEA, The Hague.

Kell, P., Vogl, G., 2012. International Students in the Asia Pacific: Mobility, Risks and Global Optimism. Springer, Dordrecht.

Kerr, D., Sturman, L., Schulz, W., Burge, B., 2010. ICCS 2009 European Report. Civic Knowledge and Attitudes in 24 European Countries. International Association for the Evaluation of Educational Achievement, Amsterdam.

Kozma, R.B. (Ed.), 2003. Technology, Innovation, and Educational Change: A Global Perspective. A Report of the Second Information Technology in Education Study Module 2. ISTE, Eugene, OR.

Lietz, P., 1996. Changes in Reading Comprehension across Cultures and over Time. Waxmann, Münster and New York.

Maclean, R., Vine, K., 2003. Achieving education for all in the Asia-Pacific region. In: Keeves, J.P., Watanabe, R. (Eds.), International Handbook of Educational Research in the Asia-Pacific Region. Kluwer, Dordrecht, pp. 17–27.

Mallinson, V., 1957. An Introduction to the Study of Comparative Education. Heinemann, Melbourne.

Maringe, F., 2010. The meanings of globalization and internationalization in higher education: findings from a world survey. In: Maringe, F., Foskett, N. (Eds.), Globalization and Internationalization in Higher Education. Continuum, London, pp. 17–34.

Martin, M.O., Mullis, I.V.S., Foy, P., Stanco, G.M., 2012. TIMSS 2011. International Results in Science. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.

Martin, M.O., Mullis, I.V.S., Foy, P., Olson, J.F., Erberber, E., Preuschoff, C., Galia, J., 2008. TIMSS 2007 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.

Meredith, W., 1993. Measurement invariance, factor analysis and factorial invariance. Psychometrika 58, 525.

Montie, J.E., Xiang, Z., Schweinhart, L.J. (Eds.), 2007. The Role of Preschool Experience in Children's Development: Longitudinal Findings from 10 Countries. The IEA Preprimary Project Phase 3. High/Scope Press, Ypsilanti, MI.

Mourshed, M., Chijioke, C., Barber, M., 2011. How the World's Most Improved School Systems Keep Getting Better. McKinsey and Company, London. Available at: http://www.mckinsey.com/Client_Service/Social_Sector/Latest_thinking/Worlds_most_improved_schools.aspx. last accessed 26 March 2012.

Mullis, I.V.S., Martin, M.O., Foy, P., Arora, A., 2012a. TIMSS 2011. International Results in Mathematics. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.

Mullis, I.V.S., Martin, M.O., Foy, P., Drucker, K.T., 2012b. PIRLS 2011. International Results in Reading. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., Foy, P., 2007. IEA's Progress in International Reading Literacy Study in Primary School in 40 Countries. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.

Mullis, I.V.S., Martin, M.O., Foy, P., Olson, J.F., Preuschoff, C., Erberber, E., Arora, A., Galia, J., 2008. TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.

Muthén, L.K., Muthén, B.O., 1998–2012a. Mplus User's Guide, seventh ed. Muthén and Muthén, Los Angeles, CA.

Muthén, L.K., Muthén, B.O., 1998–2012b. Mplus Version 7 (Computer software). Muthén & Muthén, Los Angeles, CA.

Papanastasiou, C., Plomp, T., Papanastasiou, E.C. (Eds.), 2011. IEA 1958–2008: 50 Years of Experiences and Memories, vols. 2. Cultural Center of the Kykkos Monastry, Nicosia.

Pearl, J., 2000. Causality: Models, Reasoning and Inference, vol. 29. MIT Press, Cambridge.

Postlethwaite, T.N., 2004. Monitoring Educational Achievement. Unesco, Paris.

Postlethwaite, T.N., 2009. Memories of IEA achievement surveys. In: Matthews, B., Gibbons, J.A. (Eds.), The Process of Research in Education. Shannon Research Press, Adelaide, pp. 1–20.

Purves, A.C., Levine, D.C., 1975. Educational Policy and International Assessment: Implications of the IEA Surveys of Achievement. McCutchan, Berkeley, CA.

Raudenbush, S.W., Bryk, A.S., Cheong, A.S., Fai, Y.F., Congdon, R.T., du Toit, M., 2011. HLM 7: Hierarchical Linear and Nonlinear Modeling (Computer software). Scientific Software International, Lincolnwood, IL.

Robitaille, D.F., Garden, R.A. (Eds.), 1989. The IEA Study of Mathematics II. Context and Outcomes of School Mathematics. Pergamon, Oxford.

Rosier, M.J., Keeves, J.P. (Eds.), 1991. The IEA Study of Science I: Science Education and Curricula in Twenty-Three Countries. Pergamon Press, Oxford.

Schulz, W., 2009. Questionnaire construct validation in the international civic and citizenship education study. IERI Monograph Series 2, 113–135.

Schulz, W., Ainley, J., Fraillon, J., Kerr, D., Losito, B., 2010. ICCS 2009 International Report. Civic Knowledge and Attitudes in 38 Countries. International Association for the Evaluation of Educational Achievement, Amsterdam.

Schulz, W., Ainley, J., Friedman, T., Lietz, P., 2011. ICCS 2009 Latin American Report. Civic Knowledge and Attitudes in Six Latin American Countries. International Association for the Evaluation of Educational Achievement, Amsterdam.

Schulz, W., Friedman, T., 2011. Scaling procedures for ICCS questionnaire items. In: Schulz, W., Ainley, J., Fraillon, J. (Eds.), ICCS 2009 Technical Report. International Association for the Evaluation of Educational Achievement (IEA), Amsterdam, The Netherlands, pp. 157–259.

Schwille, J., 2011. Experiencing innovation and capacity building in IEA research – a memoir 1963–2008. In: Papanastasiou, C., Plomp, T., Papanastasiou, E.C. (Eds.), IEA 1958–2008: 50 Years of Experiences and Memories, vols. 2. Cultural Center of the Kykkos Monastry, Nicosia, pp. 627–710.

Stigler, J.W., Hiebert, J., 1999. The Teaching Gap: Best Ideas from the World's Teachers for Improving Education in the Classroom. The Free Press, New York.

Super, D.E. (Ed.), 1967. Towards a Cross-National Model of Educational Achievement in a National Economy. Teachers College, Columbia University, New York.

Tatto, M.T., Schwille, J., Senk, S., Ingvarson, L., Peck, R., Rowley, G., 2008. Teacher Education and Development Study in Mathematics (TEDS-M): Policy, Practice, and Readiness to Teach Primary and Secondary Mathematics. Conceptual Framework. Teacher Education and Development International Study Center, College of Education, Michigan State University, East Lansing, MI.

Torney, J.V., Oppenheim, A.N., Farnen, R.N., 1976. Civic Education in Ten Countries. Almqvist and Wiksell, Stockholm. John Wiley, New York.

Usher, A., Savino, M., 2007. A global survey of university ranking and league tables. Higher Education in Europe 32 (1), 5–15.

Vandenberg, R.J., Lance, C.E., 2000. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. Organizational Research Methods 3, 4–69.

Walker, D.A., 1967. Mathematical achievement and school retentivity. In: Husén, T. (Ed.), International Study of Achievement in Mathematics, vol. 2. Almqvist and Wiksell, Stockholm. Wiley, New Work, pp. 116–135.

Comparative Studies: Method and Design

Dirk Berg-Schlosser, Philipps University, Institut fur Politikwissenschaft Universität Marburg, Marburg, Hessen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mill's canons are the comparative procedures described by John Stuart Mill in his famous work *A System of Logic* (1843). Among these procedures, in particular, are the following: (1) *the method of agreement*, comparison of two or more cases identifying their commonalities, ideally a single one as the common cause of a phenomenon; (2) *the method of difference*, comparison of two or more cases identifying their differences, ideally a single one (thus controlling for all others), as the cause of a phenomenon; (3) *the indirect method of difference*, double employment of the method of agreement once before and once after a certain event to identify an external effect (stimulus) that may have occurred in the meantime as in a 'quasi-experiment'; and (4) *the method of concomitant variation*, identification of factors that vary with others in a regular manner. Internal validity is the (approximate) causal relationship of two or more variables with each other. External validity is the generalization of a presumed relationship across other cases, settings, and time. Small *n* is the characteristic research situation for the 'comparative method' in the narrow sense of the term, ranging from single-case analysis ('configurative') to large *N* analysis in which statistical procedures, ideally based on random sampling from large entities, can be employed. Most similar systems–different outcomes is a systematic matching procedure to identify distinguishing factors between cases possibly responsible for the divergent outcome, while controlling for all others. Most different systems–same outcome is a systematic contrasting procedure to identify possible common factors for the same outcome, while eliminating all others. Multiple causation is the (necessary) combination of several variables to have a certain effect. Conjunctural causation is different combinations of variables having the same effect. Qualitative comparative analysis is a recent comparative technique used to analyze varying complex configurations of variables reducing them to their minimal expressions with the help of Boolean algebra.

Definition

Comparisons are essential in any science to establish systematic similarities and differences between observed phenomena and, possibly, to develop and test hypotheses and theories about their causal relationships. Whereas this broad definition of comparative inquiry includes several distinct methods, including case studies, statistical analysis, and experimental research, the comparative method in the narrower sense of the term characteristically involves a limited number of cases (small *n*) but usually complex phenomena and relationships often at the macro level of entire societies or political systems. In this latter sense, this article discusses comparative studies' method and respective designs.

Intellectual Context

Comparative procedures in the historical and political sciences date to antiquity, but their methodological backgrounds have been developed systematically in the natural sciences in the eighteenth and nineteenth centuries based on studies, for example, by Linné in botany, Cuvier in anatomy, or Bernard in experimental medicine. The logical foundations for this method were laid by Hume and, in particular, J. S. Mill's (1843) canons. The first of these refers to the method of agreement eliminating all similarities but one: "If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree is the cause (or effect) of the given phenomenon" (p. 390). By contrast, the method of difference establishes the absence of a common cause or effect, even if all other circumstances are identical: "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon" (p. 391).

Thus, both methods are concerned with the systematic matching and contrasting of cases to establish common causal relationships by eliminating all other possibilities. Both procedures are, however, somewhat extreme in the sense that they attempt to establish a single common cause, or its absence, by controlling all other possibilities and the entire environment. The method of difference, to some extent, can be approximated in pure laboratory situations in which such experiments may be conducted. The method of agreement is more applicable in field research situations, but then the question of how to control the environment becomes crucial (*see* Field Observational Research in Anthropology and Sociology).

Mill, therefore, devised a combination of the two methods, which he called the joint method of agreement and difference or the indirect method of difference, which consists of a double employment of the method of agreement, once before and once after a certain event: "If two or more instances in which the phenomenon occurs have only one circumstance in common, while two or more instances in which it does not occur have nothing in common save the absence of that circumstance, the circumstance in which alone the two sets of instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon" (Mill, 1974: p. 396). This quasi-experimental design, however, as Mill himself realized, is less cogent than the pure method of difference.

A fourth method, which is a derivation of the method of difference, is the method of residues. When cause and effect of other factors have been established by previous experiments, remaining variation in the outcome can be attributed to the residues in the original experiments: "Subduct from any phenomenon such part as is known by previous inductions to be the effect of certain antecedents, and the residue of the phenomenon is the effect of the remaining antecedents" (Mill, 1974: p. 398).

Finally, for instances in which a certain common external factor cannot be entirely eliminated by any of these procedures, its impact at least may be systematically varied from one situation to another (e.g., altitude above sea level, or heat or air pressure in certain experiments in physics) and then its varying influence assessed. This is the method of concomitant variation: "Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular manner, is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation" (Mill, 1974: p. 401).

Changes in Emphasis

Mill's canons imply rather rigid 'positivist' assumptions about the relationships of cause and effect and the state of valid theory in any given area of research. These are (1) contiguity between the presumed cause and effect; (2) temporal precedence, in that the cause had to precede the effect in time; and (3) constant conjunction, in that the cause had to be present whenever the effect was obtained (cf. Cook and Campbell, 1979: p. 10). On the whole, such relatively mechanical and deterministic relationships can be established only rarely even in the hard sciences.

By themselves, therefore, these methods do not produce any new *discoveries* unless some truly relevant factors have been included. Similarly, they may not *prove* any causal relationship unless a clear and complete (preconceived) model of such links has been tested and other factors have been controlled sufficiently. Even such a model eventually may be falsified and be integrated in a more comprehensive theory. So these methods may not produce any positive results. They constitute, however, a valuable step to eliminate many irrelevant factors and to approximate the causal conditions in the real world. In this sense they correspond to Popper's (1934) famous principle of falsification. Or as it was expressed in another classic of this period, Mill's methods are nevertheless "of undoubted value in the process of attaining truth. For in eliminating false hypotheses they narrow the field in which true ones may be found. And even where these methods may fail to eliminate all irrelevant circumstances, they enable us with some degree of approximation to so establish the conditions for the occurrence of a phenomenon, that we can say one hypothesis is logically preferable to its rivals" (Cohen and Nagel, 1934: p. 267).

If these qualifications have to be made from a general epistemological point of view, they apply even more to the social and behavioral sciences where objective facts and their subjective perceptions are closely interlinked and may change in the course of time. Mill himself was aware of the much higher complexity in the social sciences and the inapplicability of his methods in a strict sense: in "politics and history … Plurality of Causes exists in almost boundless excess, and effects are, for the most part, inextricably interwoven with one another. To add to the embarrassment, most of the inquiries in political science relate to the production of effects of a most comprehensive description, such as the public wealth, public security, public morality, and the like; results liable to be affected directly or indirectly either in *plus* or in *minus* by nearly every fact which exists, or event which occurs, in human society" (Mill, 1974: p. 452, emphases in the original). He, therefore, found the method of difference inapplicable and the methods of agreement and concomitant variation inconclusive in the social sciences (p. 881 ff.). Similarly, the method of residues requires a much more developed state of theory than is usually available in the social sciences (p. 884 ff.).

In the realm of the human and social sciences, however, Mill was willing to adopt a less deterministic and more probabilistic perspective:

> Inasmuch as many of those effects which it is of most importance to render amenable to human foresight and control are determined, like the tides, in an incomparably greater degree by general causes, than by all partial causes taken together; depending in the main on those circumstances and qualities which are common to all mankind, or at least to large bodies of them, and only in a small degree on the idiosyncrasies of organization or the peculiar History of individuals; it is evidently possible with regard to all such effects, to make predictions which will almost always be verified, and general propositions which are almost true. And whenever it is sufficient to know how the great majority of the human race, or of some nation or class of persons, will think, feel, and act, these propositions are equivalent to universal ones. For the purpose of political and social science this is sufficient (Mill, 1974: p. 847, emphases in the original).

Emphases in Current Theory and Research

Two basic problems remain in any research of this kind, namely, the questions as to what have been called the 'internal' and 'external' validity of any scientific findings (*see* Internal Validity; Causal Counterfactuals in Social Science Research). The first refers to the "approximate validity with which we infer a relationship between two variables is causal," the second refers to the "approximate validity with which we can infer that the presumed causal relationship can be generalized to and across alternate measures of the cause and effect as well as across different types of persons, settings and times" (Cook and Campbell, 1979: p. 37). Even though both kinds of validity can only be approximated and remain subject to the possibility of further falsification, they can be established with some kind of certainty with the help of procedures of randomization.

Regarding internal validity, Mill's original procedures can be enhanced by first randomly selecting two groups of subjects and then administering a certain stimulus or treatment to one of them. If a significant effect can be observed in the selected group, then the stimulus may be considered to have been the cause. This corresponds to Mill's method of difference, but where all kinds of external influences can now be controlled, within certain limits, by the equalizing procedure of random selection. This is the realm of experimental research proper.

Similarly, if a random sample has been drawn from a larger universe of persons, as for example in survey research, the relationships that can be observed in the sample may be generalized to the larger universe from which it was drawn. This is the realm of statistical analysis (*see* External Validity).

Unfortunately, however, both procedures of randomization are not feasible when we deal with the characteristic many variables–small *n* situation of many kinds of current social research at the macro level apart from the fact that many real experiments with groups of human beings are not possible for practical, political, or ethical reasons either. We are thus left with Mill's original predicaments and have to attempt to cope with them in some other ways. This depends, of course, on the kinds of cases and problems we are dealing with and the state of theory in any given area of investigation. In social research, cases often refer to larger distinct groups of people, for example, certain ethnic groups or entire societies and states that have a proper name and in which we are interested because of their intrinsic relevance and not just as an anonymously selected unit out of a multitude of cases.

The kinds of problems we are dealing with vary according to the respective emphasis of the different social and behavioral sciences. What they have in common is the multidimensionality (i.e., including objective, subjective, and often also normative dimensions) and the malleability (i.e., possible changes over time and probabilistic internal relationships) of their subject matter (Almond and Genco, 1977). The ways such relationships are perceived depend on the social (i.e., also theoretical) constructions of this reality. These, by necessity, can only be certain approximations as to their real (and changing) nature, but they have concrete referents and are not, in our view, entirely arbitrary whims of imagination as some of the more extreme postmodern authors pretend (*see* Postmodernism: Methodology). What to compare thus differs from one discipline to the other and depends on the respective state of theory.

By contrast, here we are concerned with *how* to compare under particularly difficult conditions problems of internal and external validity in the social sciences. From this perspective, a number of ways can be conceived to cope with the prevailing small *n*–complex cases dilemma. First, the number of cases may be increased, including for example, historical cases or investigating them at some lower level of analysis (e.g., certain districts or provinces rather than entire nation-states). But even then, neither randomly selected control groups nor random sampling for statistical purposes are possible. In any case, we usually are interested in concrete societies and their problems and not in any anonymous units of investigation. A second strategy concerns the purposeful selection of cases, and a third strategy concerns the reduction of complexity by means of combining several variables or employing better theory.

At the outset of any investigation, an area of homogeneity must be defined that establishes boundaries for the selection of cases. Cases must parallel each other sufficiently and be comparable along certain specified dimensions. This is meant by the common saying that apples and oranges should not be compared. In this regard, the subject matter and the problem in which we are interested must first be specified to make any sense. Thus, these fruits may well be compared concerning their sugar or water contents, their nutritional value, and so on when these dimensions as the *tertium comparationis* have been made explicit. The specification of relevant cases at the start of an investigation thus amounts to an explicit or implicit hypothesis that the cases selected are in fact alike enough to permit comparisons. The primary consideration in delimiting cases for a small *n* comparative study is the dependent variable. For example, the breakdown or survival of democratic regimes in interwar Europe presupposes the prior existence of some form of democracy in the selected cases. In addition, some limitations in time and space can enhance the homogeneity and thus the comparability of the cases examined. For example, certain kinds of colonial or other forms of external domination or religious-cultural influences may be useful criteria for selecting a specific group of cases. Area studies, for example, are based on such assumptions of regional, historical, or cultural homogeneity.

A second consideration concerns the extent of diversity within the selected universe. In this regard, a maximum of heterogeneity for a minimum number of cases should be achieved. Taking the previous example again, both survivors and breakdowns of democracy can be considered, and among the latter perhaps some more specific variants such as fascist versus more generally authoritarian outcomes.

Within this universe, two opposite strategies now become possible. One is the most similar system and the other is the most different system design. These have been explicitly formulated and discussed by Przeworski and Teune: "The most similar systems design is based on a belief that a number of theoretically significant differences will be found among similar systems and that these differences can be used in explanation" (1970: p. 39). By matching these similar cases as much as possible, most of the variables can be controlled. Mill's indirect method of difference, in which different outcomes may be attributed to the remaining factors that differentiate these cases, now becomes applicable. Even though only rarely just a single factor will remain to which the effect can be attributed, at least many others can be excluded, and the remaining ones can be examined more closely in a theoretically guided qualitative manner. The internal validity of the observed relationships thus can be greatly enhanced.

The opposite strategy is the most different systems design: this design, "which seeks maximal heterogeneity in the sample of systems, is based on a belief that in spite of intersystemic differentiation, the populations will differ with regard to only a limited number of variables or relationships" (Przeworski and Teune, 1970: p. 39). This contrasting of cases thus eliminates all factors across the observed range that are not linked to an identical outcome. In this way, more universal explanations are sought as far as the selected area of homogeneity is concerned. To some extent, the external validity of some hypothesized causal relationship can be extended and the range of its applicability, including certain limitations in time and space, can be established.

Until recently, however, these designs had not been fully operationalized. Several problems must be addressed that follow from the necessity of measuring the proximity or remoteness of pairs of cases in the heterogeneous, multidimensional space defined by the independent variables. These distance measures provide the basis for determining the most

different and most similar pairs or groups of cases with regard to the respective outcome. The two main issues are (1) choosing from among a variety of different ways of measuring the distance between pairs of cases in a multidimensional space; and (2) assigning relative weights to the variables that define this space. In this way, the complexity of the data can be retained as much as possible in the complexity of the proximity measure. These procedures then make it possible to identify most different systems–same outcome (MDSO) and most similar systems–different outcomes (MSDO). The causes of different outcomes may now be assumed to lie in the commonalities that remain among MDSO cases and the differences that distinguish MSDO cases (for such operationalizations see Berg-Schlosser and De Meur, 2009).

Another strategy for dealing with the many variables – few cases dilemma concerns the variables side. Here, for reasons concerning the internal validity of Mill's methods, one should take a rather comprehensive approach not to leave out any possibly relevant variables. This, of course, depends on the state of theory in any given area of investigation, but usually in the social sciences, several contending or supplementary hypotheses should be considered. Furthermore, not only bivariate relationships with an independent variable and the respective outcome but also the possibility of more complex causal problems – for example, multiple (several variables combined have a certain effect) or conjunctural ones (alternative combinations of different variables may lead to the same outcome, i.e., equifinality) – must be taken into account.

One technique is to look for constants across the observed cases. If one variable turns out to be consistently linked to a particular outcome, it becomes a strong candidate for inclusion in any explanation as a necessary but perhaps not sufficient condition. A related approach focuses on correlations between the independent variables and the dependent variable. Like the search for constants, the examination of correlations may provide strong hints. More sophisticated techniques include discriminant analysis and logistic regression (applied to dichotomous outcome variables). A relatively large number of variables can be simultaneously treated in this way. These techniques typically are used to assess the net, additive contribution of each independent variable to some outcome. Variables that are uncorrelated with other variables but strongly correlated with the outcome are favored by these methods. Thus, these techniques are biased toward causal factors that act independently, not conjuncturally.

These statistical approaches are useful to identify broad patterns, but they are of limited utility in situations of causal complexity. Most complex causal combinations are hidden from analyses based on correlations. For example, if a variable must be present in some conjunctures for an outcome but absent in others, the correlation between this cause and the outcome may be 0. Qualitative comparative analysis (QCA), a recent comparative technique based on Boolean algebra and set theory (cf. Ragin, 1987), can be used in situations of causal complexity. In contrast to the linear techniques just described, QCA focuses on configurations of variables. In QCA, an independent variable can be eliminated from an analysis if it does not uniquely distinguish any case that manifests an outcome from at least one case lacking the outcome.

Other techniques for reducing the number of variables focus on patterns among the variables. In light of relevant theoretical and empirical knowledge, investigators may select a more limited number of major variables or reconstruct them in different ways. Such reductions may proceed statistically. For example, researchers often use factor analysis to combine correlated variables into single indices.

Qualitative techniques for constructing more encompassing causal variables also are possible. For example, in the interwar study, the existence of a rural proletariat and large-scale landlords has been combined into the more encompassing concept of feudalism. Similarly, a Boolean addition of factors, constituting alternative constellations in combination with another variable, may be possible. For example, the presence of ethnic, religious, *or* regional social cleavages in the absence of any overarching (verzuiling) structures was combined to the variable of social heterogeneity in the study of democratic breakdowns and survivals (Berg-Schlosser and Mitchell, 2002). Use of ideal-typical constructs also can be conceived of as reductions of complexity emphasizing some characteristics and deemphasizing others. In this way, the overall number of variables may be reduced considerably while still retaining much of the original information. With the help of such strategies, both from the cases and the variables sides, the relationship between their respective numbers may be reduced to manageable proportions, thus avoiding the problem of overdetermination when the latter exceed the former.

QCA can be employed for other purposes. In its crisp-set version, it is based on dichotomized variables (which, however, can be constructed as dummies from multichotomous or interval ones) and reduces, by employing certain algorithms, the observed complexity of possible independent factors with regard to a particular outcome to a minimal logical formula (or several alternative ones) taking account of the possible interactions of these factors. The most important rule in QCA is as follows: "If two Boolean expressions differ in only one causal condition yet produce the same outcome, then the causal condition that distinguishes the two expressions can be considered irrelevant and can be removed to create a simpler, combined expression" (Ragin, 1987: p. 93). For example, if the combination of three factors A * B * C or A * B * c lead to the same outcome O (where upper case letters signify the presence and lower case letters the absence of a condition) then this expression can be reduced to A * B = O.

QCA can thus (1) examine a universe of cases with regard to a given output in an exhaustive manner; this description will be the most concise, including all possible alternate minimal expressions that allow for the most parsimonious interpretations; (2) test hypotheses by showing whether they are consistent with the combinations described; if many contradictions occur, the hypothesis fails; and (3) deduce the shortest possible formulas, describing the simplest subset of actual, potential, or counterfactual cases not contradicting the respective outcome; in this way, more general expressions can be deduced by using simplifying assumptions. QCA thus serves some of the most important aims of any science, namely, systematic description and the falsification and construction of theories.

QCA shares some of the problems and limitations concerning the loss of information and a certain amount of arbitrariness

in setting cutting-points for dichotomized variables. In contrast with common statistical procedures, however, it retains the full complexity of all cases and all variables considered. Nonunique formulas may emerge, but this must be considered a strength and not a weakness. It deals explicitly with outliers and covers them fully. It can arrive at several conjunctural constellations that explain the same outcome. When several different formulas cover a single case, the technique forces the researcher (as the MSDO technique) to take a closer look at the results and to *interpret* them in light of the historical knowledge of the case. In comparative inquiry, this approach is preferred to a purely mechanical procedure that, in many statistical analyses, entirely obscures the fate of particular cases. Here, in fact, begins the real *qualitative* work, depending very much on the training and quality of researchers, their in-depth knowledge of cases, as well as their sensitivity and understanding.

More recently, the possibility to include multivalue conditions (variables) for QCA has been incorporated in a new software (TOSMANA, i.e., tools for small *n* analysis) together with new features visualizing thresholds and Venn diagrams (Cronqvist and Berg-Schlosser, 2009). In this way, more information can be retained, but in a small *n* world, this procedure should be used sparingly not to individualize each case, that is, having distinct expressions for each of them and no longer broader patterns of cases.

Ragin (2008) has developed this approach further by employing fuzzy sets instead of crisp sets for a QCA analysis. This allows for the use of continuous (calibrated) scales on all conditions and includes measures of coverage (percentage of cases included) and consistency (degree of subset membership in a multidimensional vector space). It also can identify necessary and sufficient conditions in separate steps. Fuzzy set analysis, incorporating some probabilistic assumptions and parameters, requires, however, a somewhat larger (medium) *N* of 25 or so cases and above. On the whole, this approach constitutes a viable set-theoretical alternative in small and medium *N* settings to common statistical procedures and has found numerous applications across various social science fields.

Methodological Issues and Problems

Even though the small *n* comparative method and the statistical methods have distinct applications, each with its own advantages and problems, different schools of thought have formed around these approaches fraught with controversies. In part, these reflect remaining biases; to some extent, however, they are based on real problems that cannot be easily resolved. Some of these issues have been stated as quantitative versus qualitative, and case-oriented versus variable-oriented, holistic versus analytic, or deterministic versus probabilistic research. To some extent these distinctions pose false alternatives and represent differences in emphasis rather than in kind. Thus, statistical procedures are, by their very nature, based on larger quantities, but small *n* analyses similarly may deal with numbers and quantitative data. Qualitative aspects refer to the presence or the absence of a factor, as in QCA, but should not be confounded with qualitative methods as, for example, hermeneutics or participant observation (*see* Interpretive Methods: Macromethods; Participant Observation) in a different sense of the word.

In the same way, case-oriented means that emphasis is laid on a number of known and theoretically or historically particularly interesting and relevant cases, as opposed to, for example, the anonymous ones in random sampling, but all scientific inquiry is based on distinct units (cases) and variables and a number of relevant observations (King et al., 1994). *Holistic* is also a misleading term referring to the usual higher complexity of case-oriented research, but it does not mean in this context that cases can only meaningfully be understood in their entirety (e.g., in Confucian philosophy) as opposed to critical-rational analysis.

Finally, deterministic is a misnomer for procedures based on Millian methods or logical distinctions as in Boolean algebra. Only the presence or absence of a factor is referred to and not its probabilistic nature with respect to its effect on the internal or external validity of any particular finding. Logically minimized relationships found by Boolean procedures can be as spurious as correlational analysis. The revealed patterns indicate different more complex qualitative constellations, but they are not based on their relative frequency or probabilistic inferences.

As to the real problems, however, these persist and only certain approximations or trade-offs between certain emphases are possible. Millian methods and similar ones can serve only to eliminate false hypotheses and to delineate the conditions for the occurrence of a phenomenon (Cohen and Nagel, 1934). Within this realm, the truly qualitative and interpretative skills of knowledgeable experts are needed to provide meaningful assessments of the factors involved. These procedures may be more tedious and less universalistic, but they are often more meaningful and less superficial than many of the macroquantitative studies.

Most small *n* comparative studies, as well as the majority of statistical analyses, have been cross-sectional and static. There have been some advances in the statistical field in recent years as far as pooled time-series and event-history modeling are concerned (*see* Event Sequence Analysis: Methods), but these share some of the limitations and the relative superficiality of macroquantitative procedures. There are, as yet, no truly dynamic procedures incorporated in the MSDO or QCA techniques combining both cross-sectional and longitudinal aspects in a macroqualitative way. These still rely more on well-informed thick narratives (*see* Thick Description: Methodology), but some specific path-dependent techniques have been developed more recently.

A further limitation refers to the structure (macro-) and actor (micro-) dimensions of any social and, by implication, comparative analysis. Most comparative studies are concerned with the broader and presumably more enduring structures of social processes both in time and space. The final outcome, nevertheless, may very well depend on any particular (group or individual) actor in the analyzed chain of events. Again, only certain approximations can be arrived at in this regard and certain trade-offs between more universalizing, variation-finding, and individualizing approaches (Tilly, 1984) have to be made. No macro–macro shortcuts should be made in sociological explanations because concrete subjects with their respective perceptions and preferences always are involved, which only in a later step may be aggregated at a meso or macro level (Coleman,

1990). Whether these actors should be conceived of as rational do not concern us in this context. The broader structural aspects can be conceived as the opportunity set within the boundaries of which actual actors operate and respective choices are made.

A final point must again be mentioned, the problem of inference. The unified logic of inference of both quantitative and qualitative research has been emphasized once more by King et al. (1994). In spite of many important issues that are raised and clarified by the authors, their experiences and recommendations lean too strongly on the quantitative side of research. This relates both to the internal and external validity of the inferences to be drawn. In terms of the internal causal chains that should be established, King et al. look at them from the side of the independent variables in a forward sense toward their eventual effect on the outcome. Much of small *n* comparative research, however, is concerned with disentangling the different influences on an outcome from a backward perspective or through process tracing to identify the different factors in their complex interactions over time involving the major structures and actors, but also many contingent and unique aspects. This resembles more a detective's investigation of a crime or an accident than any more nomothetic generalizing scientific efforts and remains idiographic (Brady and Collier, 2010). The more complex pattern of causality that has been established in this way in terms of its internal validity can then serve to devise new more complex hypotheses that can be tested across more similar (comparable) cases and, possibly, establish a certain range of their external validity. A much stronger theory may thus emerge.

Probable Future Directions of Theory and Research

The strength of the comparative method lies, therefore, in its critical applications to theory testing and development. The state of affairs of theory building in the social sciences remains relatively weak and is subject to specific problems and restrictions and to the general malleability of its substance. Even if only certain approximations can be reached, the relevance of many macrosocial and political problems – for example, in policy research, institution-building, constitutional engineering, and so on – for potentially millions of people make it worth every effort. Better theory, in the longer run, then also may contribute to a better practice.

It is difficult to predict any more distinct developments in this regard, but at least some of the *desiderata* that reflect some of the existing deficiencies can be outlined. First of all, many more existing propositions across the various fields of social inquiry based on other methods of data collection and theory building should be subjected to systematic testing by comparative methods as far as they refer to small *n* environments. Counterfactual examples and most different cases should be sought to establish their respective range of external validity. Second, from this body of research, more coherent propositions may emerge delineating their respective range of application, both in space and time.

It seems likely that such medium-range theories increasingly will be based on regional and historical commonalities of certain countries and cultures. Some deep structures, such as the social cleavages in Western Europe, and deep cultures, such as certain characteristic submilieus in Germany or Italy, have persisted over centuries and can be linked to present developments (Putnam, 1993; Flora, 1999). What is needed are better qualitative and quantitative methods to assess their continuities and changes over time and dynamic concepts and tools to establish their interactions, not just in abstract theory but also in empirical reality.

See also: Comparative History; Comparative Method in Linguistics; Mill, John Stuart (1806–73).

Bibliography

Almond, G.A., Genco, S.J., 1977. Clocks, clouds and the study of politics. World Politics 29, 489–522.

Berg-Schlosser, D., De Meur, G., 2009. Comparative research design: case and variable selection. In: Rihoux, B., Ragin Ch, C. (Eds.), Configurational Comparative Methods – Qualitative Comparative Analysis (QCA) and Related Techniques. SAGE, Los Angeles, pp. 19–32.

Berg-Schlosser, D., Mitchell, J. (Eds.), 2002. Authoritarianism and Democracy in Europe 1919–39 – Comparative Analyses. Palgrave Macmillan, London.

Brady, H.E., Collier, D. (Eds.), 2010. Rethinking Social Inquiry: Diverse Tools, Shared Standards, second ed. Rowman & Littlefield, Lanham, MD.

Cohen, M., Nagel, E., 1934. An Introduction to Logic and Scientific Method. Harcourt, Brace, New York.

Coleman, J.S., 1990. Foundations of Social Theory. Belknap Press, Cambridge, MA.

Cook, Th D, Campbell, D.T., 1979. Quasi-Experimentation – Design and Analysis Issues for Field Settings. Houghton Mifflin, Boston.

Cronqvist, L., Berg-Schlosser, D., 2009. Multi-Value QCA. Rihoux & Ragin, op.cit., pp. 69–86.

Flora, P. (Ed.), 1999. State Formation, Nation-building and Mass Politics in Europe—the Theory of Stein Rokkan. Oxford University Press, Oxford, UK.

King, G., Keohane, R.O., Verba, S., 1994. Designing Social Inquiry: Scientific Inference in Qualitative Research. Princeton University Press, Princeton, NJ.

Mill, J.S., 1974/75. A System of Logic. In: Collected Works of J. S. Mill, vols. 7 and 8. Routledge and Kegan Paul, London (originally published in 1843).

Popper, K.R., 1934/1968. The Logic of Scientific Discovery. American ed. Harper and Row, New York.

Przeworski, A., Teune, H., 1970. The Logic of Comparative Social Inquiry. Wiley, New York.

Putnam, R.D., 1993. Making Democracy Work – Civic Traditions in Modern Italy. Princeton University Press, Princeton.

Ragin, Ch. C., 1987. The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies. University of California Press, Berkeley, CA.

Ragin, Ch. C, 2008. Redesigning Social Inquiry – Fuzzy Sets and Beyond. University of Chicago Press, Chicago.

Tilly, Ch., 1984. Big Structures, Large Processes, Huge Comparisons. Russell Sage Foundation, New York.

Compliance

Jing Cheng, University of California San Francisco, San Francisco, CA, USA
Dylan S Small, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Randomized studies are the gold-standard way to study the effect of a treatment. However, many randomized studies suffer from noncompliance – some subjects do not comply with their assigned protocol. When there is noncompliance, researchers are often interested in estimating the effect of receiving the treatment. The difficulty is that who receives the treatment is not random when there is noncompliance because subjects self-select whether they comply with their assigned protocol. We discuss approaches for unbiasedly estimating the effect of the treatment, focusing particularly on the instrumental variable approach.

Introduction

In many randomized trials, some subjects do not comply with their assigned treatment protocol (see **Figure 1**). In this article, we will discuss what quantities are of interest to estimate when there is noncompliance in a randomized trial and how to estimate those quantities. For most of the article, we will focus on two-arm trials and then we will discuss multiarm trials at the end of the article. As an example, we will consider the PROSPECT study, a randomized trial of an intervention to improve treatment of depression among depressed elderly patients in primary care practices (Bruce et al., 2004). The intervention was that a depression care specialist collaborated with the patient's primary care physician to facilitate adherence to a depression treatment strategy and provide education and assessment to the patient. The control was usual care. The study involved 539 depressed patients in 20 primary care practices at 3 sites. Each practice was randomized to either intervention (the treatment) or usual care (the control). Compliance with the intervention was categorized as a binary variable, whether or not a patient had seen a depression care specialist in the 4 months after the beginning of the trial. The outcome of interest is the change in a patient's Hamilton depression scores from baseline to 4 months after the beginning of the trial. The Hamilton score takes values between 0 and 50, with a higher value meaning more depression; so a negative change in Hamilton depression score means that depression has been reduced.

The as-received treatment effect is the effect of receiving the experimental treatment (hereafter referred to as the 'treatment') relative to receiving the control condition (hereafter referred to as the 'control'). By contrast, the intent-to-treat (ITT) effect (**Figure 1**) is the effect of being offered the treatment relative to being offered the control. Both the as-received treatment effect and the ITT effect are valuable to know. One situation when knowing the as-received treatment effect is particularly valuable is when the treatment nonadherence pattern is expected to differ between the study sample and the target population (Sommer and Zeger, 1991; Sheiner and Rubin, 1995; Small et al., 2006). For example, in studies of psychiatric interventions like the PROSPECT study, the study sample may consist of physically healthy patients highly adherent to their psychiatric care whereas the target population may contain many people receiving multiple treatments for medical comorbidities who are less likely to adhere to psychiatric treatments, or intensive efforts to maintain high levels of adherence in randomized trials may not be able to be maintained by the community of practitioners without similar resources to sustain adherence among their patients (Ten Have et al., 2008). In this situation, the ITT estimate may be biased for the effect of offering the treatment to the target population (Sommer and Zeger, 1991; Sheiner and Rubin, 1995). A key quantity that needs to be known to accurately predict the effect of offering the treatment to the target population is the effect of actually receiving the treatment (Sheiner and Rubin, 1995; Small et al., 2006). As an example, Sommer and Zeger (1991) discuss a trial of vitamin A supplementation to reduce child mortality. In the trial, the vitamin A supplementation was implemented by having children take pills and some children who were randomized to treatment did not take the pills. The ITT effect is the effect of making vitamin A pills available to children. However, if the trial showed that taking the pills was efficacious, vitamin A supplementation would not likely be implemented by providing pills but instead by fortifying a daily food item such as monosodium glutamate or salt (Sommer and Zeger, 1991). By knowing the effect of receiving vitamin A supplementation (the biologic efficacy) and the rate at which vitamin A supplementation would be successfully delivered under a fortification program, we can estimate the effectiveness

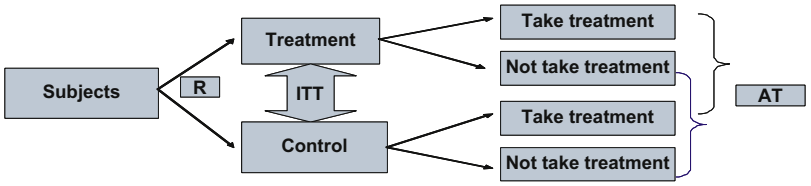

Figure 1 Noncompliance in randomized studies. ITT, intent-to-treat. R, Randomization; AT, Always Takers.

of a fortification program. A second situation in which knowing the effect of receiving treatment versus control is particularly valuable is when patients who are interested in fully adhering to a treatment are making decisions about whether to take the treatment (Sheiner and Rubin, 1995; Hernán and Hernández-Díaz, 2012). For example, Hernán and Hernández-Díaz (2012) mention the setting that to decide whether to use a certain contraception method, a couple may want to know the failure rate if they use the method as indicated, rather than the failure rate in a population that included a substantial proportion of nonadherers.

A commonly used estimator for the as-received treatment effect is the as-treated estimate (**Figure 1**). The as-treated estimate is the difference between the mean outcome among patients who received the treatment and the mean outcome among patients who did not receive the treatment. In a trial in which only the assigned-to-treatment arm can receive the treatment, the as-treated estimate will be biased if patients who would adhere to the treatment if randomized to it (would be adherers) differ in their characteristics from those who would not adhere to treatment if randomized to it (would be nonadherers) because the as-treated estimate compares would be adherers to a mixture of would-be adherers and would-be nonadherers. A vivid example of how would-be adherers can differ from would-be nonadherers was provided by May et al. (1981) in their discussion of the Coronary Drug Project, a randomized trial of lipid-lowering drugs for individuals who had survived myocardial infarction. Good adherers in the group assigned to the lipid-lowering drugs had a 5-year mortality rate of 15.0% whereas poor adherers had a mortality rate of 24.6%, making the lipid-lowering drugs seem promising. However, in the placebo group, good adherers had a mortality rate of 15.1% and poor adherers had a mortality rate of 28.2%, so placebo looks even more promising. For cardiac patients, the would-be adherers might have better diet and exercise more than the would-be nonadherers, leading to better outcomes if both groups received the same treatment. In contrast, in psychiatry, there are examples in which the would-be adherers are less healthy than would-be nonadherers (Small et al., 2006; Ten Have et al., 2008).

The as-treated estimate may be biased for estimating the as-received treatment effect because of unmeasured confounding, e.g., individuals with better diet may be more likely to adhere to treatment and to have better outcomes regardless of treatment. Another standard estimate of the as-received treatment effect is the per-protocol estimate, which compares the outcomes of subjects who were assigned to the treatment and followed the treatment protocol to subjects who were assigned the control and followed the control protocol; similar to the as-treated estimate, the per-protocol estimate may be biased for estimating the effect of receiving the treatment because of unmeasured confounding, e.g., individuals with better diet may be more likely to adhere to treatment but all subjects may adhere with the control if the control is usual care. Instrumental variable (IV) methods are methods for adjusting the unmeasured confounding. In a randomized trial with nonadherence, the random assignment of being offered the treatment may be an IV that can be used to control for unmeasured confounding in estimating the as-received treatment effect. In the next section, we will discuss the IV approach.

IV Approach

Framework and Notation

The Neyman–Rubin potential outcomes framework (Neyman, 1990; Rubin, 1974) will be used to describe causal effects and formulate IV assumptions. The classic econometric formulation of IVs is in terms of structural equations and assumptions about the IV being uncorrelated with structural error terms; the formulation in terms of potential outcomes that is described here provides clarity about what effects are being estimated when there are heterogeneous treatment effects and provides a firm foundation for nonlinear as well as linear outcome models (Holland, 1988; Angrist et al., 1996). Suppose there are N subjects. Let $\mathbf{Z}$ denote the N-dimensional vector of treatment assignments with individual elements Z_i, $i = 1, \ldots, N$, where $Z_i = 0$ or 1; the treatment assignment Z will be the IV in a randomized study. Let $\mathbf{D}^{\mathbf{z}}$ be the N-dimensional vector of potential treatment received under treatment assignment $\mathbf{z}$ with elements $D_i^{\mathbf{z}}$, $i = 1, \ldots, N$, $D_i^{\mathbf{z}} = 1$ or 0 according to whether person i would receive treatment level 1 or 0 under treatment assignment $\mathbf{z}$. Let $\mathbf{Y}^{\mathbf{z},\mathbf{d}}$ be the N-dimensional vector of potential outcomes under treatment assignment $\mathbf{z}$ and treatment received $\mathbf{d}$ where $\mathbf{Y}_i^{\mathbf{z},\mathbf{d}}$ is the outcome subject i would have under treatment assignment $\mathbf{z}$ and treatment received $\mathbf{d}$. The observed treatment received for subject i is $D_i \equiv D_i^{\mathbf{Z}}$ and the observed outcome for subject i is $Y_i \equiv Y_i^{\mathbf{Z},\mathbf{D}^{\mathbf{Z}}}$. Let $\mathbf{X}_i$ denote observed covariates for subject i. When we write expressions like $E(Y)$, we mean the expected value of Y for a randomly sampled subject from the population.

Angrist et al. (1996) considered treatment assignment to be an IV if it satisfies five assumptions – the stable unit treatment value assumption (SUTVA), the IV is positively correlated with treatment assumption, the IV is independent of unmeasured confounders assumption, the exclusion restriction (ER) assumption, and the monotonicity assumption. We will describe the first four of these assumptions and then describe the need for the fifth assumption or some substitute to obtain point identification. The first four assumptions are:

IV-A1 *SUTVA*. If $\mathbf{z}_i = \mathbf{z}_i'$, then $D_i^{\mathbf{z}} = D_i^{\mathbf{z}'}$ and if $\mathbf{z}_i = \mathbf{z}_i'$ and $\mathbf{d}_i = \mathbf{d}_i'$, then $Y_i^{\mathbf{z},\mathbf{d}} = Y_i^{\mathbf{z}',\mathbf{d}'}$. In words, this assumption says that the treatment assigned and treatment received by a subject only affects that subject and not other subjects, and that there are not different versions of the treatment which have different effects (see Angrist et al., 1996; Rubin, 1990 for details). The SUTVA allows us to write $D_i^{\mathbf{z}}$ as D_i^{z} where z here denotes subject i having treatment assignment z, and $Y_i^{\mathbf{z},\mathbf{d}}$ as $Y_i^{z,d}$ where z and d here denote subject i having treatment assignment z and treatment received d.

IV-A2 *Treatment assignment is positively correlated with treatment received*. $E(D^1) > E(D^0)$.

IV-A3 *Treatment assignment is independent of baseline variables and unmeasured confounders*.

$$Z \perp \left(D^1, D^0, Y^{1,1}, Y^{1,0}, Y^{0,1}, Y^{0,0}\right)$$

IV-A4 *ER*. This assumption says that the treatment assignment affects outcomes only through its effect on treatment received: $Y_i^{z,d} = Y_i^{z',d}$ for all i. Under the ER, we can write $Y_i^d \equiv Y_i^{z,d}$ for any z, i.e., Y_i^1 is the potential outcome for subject i if he or she were to receive level 1 of the

treatment (regardless of his or her treatment assignment) and Y_i^0 is the potential outcome if he or she were to receive level 0 of the treatment. The ER assumption is also called the no-direct-effect assumption.

Assumptions IV-A2, IV-A3, and IV-A4 are the 'core' IV assumptions that basically all IV approaches make; assumption IV-A1 is typically implicitly made as well. Figure 2 shows a graphical model view with random treatment assignment as an IV. Assumption IV-A2 will be satisfied in a randomized trial as long as the treatment assignment provides some encouragement to take the treatment. Assumption IV-A3 will be satisfied in a randomized trial as long as the treatment was in fact randomly assigned. Assumption IV-A4 could be violated if the treatment assignment has some effect beyond its effect on treatment assignment. In a placebo-controlled randomized trial, assumption IV-A4 would be violated if the placebo has some beneficial effect and never-takers have some chance of taking the placebo under the control; then, the never-takers would have a higher average outcome under assignment to the control (where they might take the placebo) than under assignment to the treatment (where they will not take the placebo nor the treatment).

The core IV assumptions IV-A2, IV-A3, and IV-A4 enable bounds on treatment effects to be identified but do not point identify a treatment effect (Balke and Pearl, 1997). To see why the core IV assumptions alone do not point identify a treatment effect and to understand what additional assumptions would identify a treatment effect, it is helpful to introduce the idea of compliance classes (Angrist et al., 1996). A subject in a study with binary IV and treatment can be classified into one of four latent compliance classes based on the joint values of potential treatment received (Angrist et al., 1996). The four compliance classes are referred to as never-takers, always-takers, defiers, and compliers. We denote subject i's compliance class as C_i, which are defined like so: C_i = never-taker (nt) if $(D_i^0, D_i^1) = (0, 0)$, i.e., subjects who would never take the treatment regardless of the assignment; complier (co) if $(D_i^0, D_i^1) = (0, 1)$, i.e., subjects who would take the treatment if assigned to treatment and would not take the treatment if assigned to control; always-taker (at) if $(D_i^0, D_i^1) = (1, 1)$, i.e., subjects who would always take the treatment regardless of the assignment; and defier (de) if $(D_i^0, D_i^1) = (1, 0)$, i.e., subjects who would always take the opposite treatment to the assignment. Note that there being four compliance classes is not an assumption but an exhaustive list of the possible types of compliance. Note also that a subject's compliance class is relative to a particular randomized trial, e.g., a subject might be a complier in a trial which provided a strong incentive to adhere to the treatment if assigned to the treatment (e.g., financial incentives for adherence) but a never-taker in a trial without such incentives. Table 1 shows the relationship between the latent compliance classes and the observed groups.

From Table 1, each observed group ($Z_i = z$; $D_i = d$; $z, d = 0, 1$) is a mixture of two latent compliance classes. Treatment effects for the latent compliance classes will often be of interest. However, even under assumptions IV-A1 to IV-A4 such that both never-takers and always-takers would have the same potential outcome under treatment and control, we are not able to directly observe the outcomes for the latent compliance classes under treatment and control, and therefore not able to identify the effects for the latent compliance classes. For example, suppose the outcome is binary so that the observed data (Y, D, Z) are multinomial random variables with $2 \times 2 \times 2 = 8$ categories. Under assumptions IV-A1 to IV-A4, there are 10 free unknown parameters: $P(Z = 1)$, $P(Y^1 = 1|C = at)$, $P(Y^1 = 1|C = co)$, $P(Y^0 = 1|C = co)$, $P(Y^0 = 1|C = nt)$, $P(Y^1 = 1|C = de)$, $P(Y^0 = 1|C = de)$, $P(C = at)$, $P(C = co)$, and $P(C = nt)$ (note that $P(C = de)$ is determined by $P(C = co) + P(C = at) + P(C = nt) + P(C = de) = 1$). Since there are 10 free parameters but the observed data are multinomial random variables that have only 8 categories (so 7 free probabilities), the model is not identified. Two types of additional assumptions have been considered that reduce the number of free parameters: (1) an assumption about the process of selecting a treatment to take based on the treatment assigned that restricts the number of latent compliance classes by ruling out defiers and (2) assumptions that restrict the heterogeneity of treatment effects among the different compliance classes.

We first consider the approach (1) to point identification of restricting the number of compliance classes. The assumption made by Angrist et al. (1996) rules out defiers:

IV-A5 *Monotonicity assumption*. This assumption says that there are no subjects who are 'defiers,' who would only take level 1 of the treatment if not encouraged to do so, i.e., there is no subject i with $D_i^1 = 0$, $D_i^0 = 1$.

Monotonicity is automatically satisfied for single-consent randomized encouragement designs in which only the subjects assigned to receive the treatment are able to receive it (Zelen, 1979) (for this design, there are only compliers and never-takers). Monotonicity is also plausible in many randomized trials in which assignment to the treatment ($Z = 1$) provides a clear incentive and no disincentive to take the treatment. For

Table 1 The relation between observed groups and latent compliance classes

Z_i	D_i	C_i
1	1	Complier or always-taker
1	0	Never-taker or defier
0	0	Never-taker or complier
0	1	Always-taker or defier

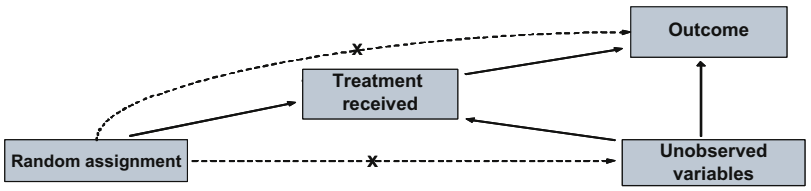

Figure 2 Graphical model view of randomization as an IV.

examples of trials where monotonicity may not hold, see Ten Have et al. (2004). For the setting of a binary outcome, assumption IV-A5 reduces the number of free parameters to seven, enabling identification since there are seven free probabilities for the observed data. However, the model only identifies the average treatment effect for compliers. The never-takers and always-takers do not change their treatment status when the instrument changes, so under the ER assumption, the potential treatment and potential outcome under either level of the IV ($Z_i = 1$ or 0) are the same. Consequently, the IV is not helpful for learning about the treatment effect for always-takers or never-takers. Compliers are subjects who change their treatment status with the IV, that is, the subjects would take the treatment if they were encouraged to take it by the IV but would not otherwise take the treatment. Because these subjects change their treatment with the level of the IV, the IV is helpful for learning about their treatment effects. The average causal effect for this subgroup, $E(Y_i^1 - Y_i^0 | C_i = co)$, is called the complier average causal effect (CACE) or the local average treatment effect. It provides the information on the average causal effect of receiving the treatment for compliers.

Approach (2) to making assumptions about IVs to enable point identification of a treatment effect keeps assumptions IV-A1 to IV-A4 but does not make assumption IV-A5 (monotonicity); instead, it makes an assumption that restricts the heterogeneity of treatment effects among the compliance classes. The strongest such assumption is that the average effect of the treatment is the same for all the compliance classes, $E(Y^1 - Y^0 | C = co) = E(Y^1 - Y^0 | C = at) = E(Y^1 - Y^0 | C = nt) = E(Y^1 - Y^0 | C = de)$. This assumption identifies the average treatment effect for the whole population (this can be derived by using the same reasoning as in the derivation of eqn [2]). A weaker restriction on the heterogeneity of treatment effects among the compliance classes is the no current treatment value interaction assumption (Robins, 1994):

$$E(Y^1 - Y^0 | D = 1, Z = 1, \mathbf{X}) = E(Y^1 - Y^0 | D = 1, Z = 0, \mathbf{X}) \quad [1]$$

Assumption [1] combined with IV-A1 to IV-A4 identifies the average effect of treatment among the treated, $E(Y^1 - Y^0 | D = 1)$. Under assumptions IV-A1 to IV-A5, assumption [1] says that the average treatment effect is the same among always-takers and compliers conditional on $\mathbf{X}$, since the left hand side of eqn [1] is the average effect of treatment among compliers and always-takers, and the right hand side is the average effect among always-takers (all conditional on $\mathbf{X}$). For trials in which subjects cannot take the treatment if assigned to the control (e.g., single-consent randomized encouragement designs), approaches (1) and (2) are equivalent since in this case, there are no always-takers and the monotonicity assumption and eqn [1] hold by the design of the trial, and the average effect of treatment among the treated is equal to the average treatment effect among compliers. For further information on approach (2), see Tan (2010) and Vansteelandt et al. (2011).

We are going to focus on approach (1) to the IV assumptions (i.e., assumptions IV-A1 to IV-A5) for the rest of this article. An attractive feature of this approach is that the monotonicity assumption IV-A5 is reasonable in many applications (e.g., when assignment to treatment provides a clear incentive to take treatment and no disincentive). Although the effect identified by IV-A1 to IV-A5, the CACE, is only the effect for the subpopulation of compliers, the data are in general not informative about average effects for other subpopulations without extrapolation, just as a randomized experiment conducted on men is not informative about average effects for women without extrapolation (Imbens and Wooldridge, 2007). By focusing on estimating the CACE, the researcher sharply separates exploration of the information in the data from extrapolation to the (sub)population of interest (Imbens and Wooldridge, 2007). See Section Extrapolating the CACE to the Full Population for discussion of extrapolation of the CACE to other (sub)populations.

Two-Stage Least Squares (Wald) Estimator

Angrist et al. (1996) show that under assumptions IV-A1 to IV-A5, the CACE is nonparametrically identified by

$$\begin{aligned}
E(Y^1 - Y^0 | C = co) &= \frac{P(C = co)\left[E(Y^1 - Y^0 | C = co)\right]}{P(C = co)} \\
&= \frac{\left[P(C = at)E(Y^1 | C = at) + P(C = co)E(Y^1 | C = co) + P(C = nt)E(Y^0 | C = nt)\right] - \left[P(C = at)E(Y^1 | C = at) + P(C = co)E(Y^0 | C = co) + P(C = nt)E(Y^0 | C = nt)\right]}{\left[P(C = at) + P(C = co)\right] - \left[P(C = at)\right]} \\
&= \frac{\left[P(C = at)E(Y^{1,1} | C = at) + P(C = co)E(Y^{1,1} | C = co) + P(C = nt)E(Y^{1,0} | C = nt)\right] - \left[P(C = at)E(Y^{0,1} | C = at) + P(C = co)E(Y^{0,0} | C = co) + P(C = nt)E(Y^{0,0} | C = nt)\right]}{\left[P(C = at) + P(C = co)\right] - \left[P(C = at)\right]} \\
&= \frac{\left[P(C = at | Z = 1)E(Y^{1,1} | Z = 1, C = at) + P(C = co | Z = 1)E(Y^{1,1} | Z = 1, C = co) + P(C = nt | Z = 1)E(Y^{1,0} | Z = 1, C = nt)\right] - \left[P(C = at | Z = 0)E(Y^{0,1} | Z = 0, C = at) + P(C = co | Z = 0)E(Y^{0,0} | Z = 1, C = co) + P(C = nt | Z = 0)E(Y^{0,0} | Z = 0, C = nt)\right]}{\left[P(C = at | Z = 1) + P(C = co | Z = 1)\right] - \left[P(C = at | Z = 0)\right]} \\
&= \frac{E(Y | Z = 1) - E(Y | Z = 0)}{E(D | Z = 1) - E(D | Z = 0)}
\end{aligned} \quad [2]$$

where the second equality follows from the IV that is correlated with treatment-received assumption (assumption IV-A2), the third equality follows from the ER assumption (assumption IV-A4), the fourth equality follows from the IV that is independent of baseline variables and unmeasured confounders assumption (assumption IV-A3), and the last equality follows from the monotonicity assumption (assumption IV-A5) that the set of defiers is empty.

The standard IV estimator for a binary IV and a binary treatment is the sample analogue of the last expression in eqn [2],

$$\widehat{\text{CACE}} = \frac{\widehat{E}(Y_i|Z_i = 1) - \widehat{E}(Y_i|Z_i = 0)}{\widehat{E}(D_i|Z_i = 1) - \widehat{E}(D_i|Z_i = 0)} \qquad [3]$$

where $\widehat{E}$ denotes the sample mean. The standard IV estimator is called the Wald estimator after Wald (1940).

The standard IV estimator is also called the two-stage least squares (2SLS) estimator because it can be obtained from the following 2SLS procedure: (1) regress D on Z by least square to obtain$\widehat{E}(D|Z)$ and (2) regress Y on $\widehat{E}(D|Z)$. The coefficient on $\widehat{E}(D|Z)$ from the regression (2) equals eqn [3]. Since eqn [3] can be obtained by the 2SLS procedures, we denote eqn [3] by $\widehat{\text{CACE}}_{\text{2SLS}}$,

$$\widehat{\text{CACE}}_{\text{2SLS}} = \frac{\widehat{E}(Y_i|Z_i = 1) - \widehat{E}(Y_i|Z_i = 0)}{\widehat{E}(D_i|Z_i = 1) - \widehat{E}(D_i|Z_i = 0)} \qquad [4]$$

To see why 2SLS provides a consistent estimate of the CACE, write $Y = \alpha + \text{CACE} \times D + u$, where α is chosen so that $E(u) = 0$. Then, under the monotonicity and ER assumption,

$$\alpha + u_i = \begin{cases} Y_i^1 - \text{CACE} & C_i = at \\ Y_i^0 + \left[\left(Y_i^1 - Y_i^0\right) - \text{CACE}\right]Z_i & C_i = co \\ Y_i^0 & C_i = nt \end{cases}$$

Thus, under the assumption that the IV is independent of baseline variables and unmeasured confounders (IV-A3), $E(Y|Z) = \alpha + \text{CACE} \times E(D|Z)$ and an unbiased estimate of the CACE can be obtained from regressing Y on $E(D|Z)$. We do not know $E(D|Z)$ but can replace it by the consistent estimate $\widehat{E}(D|Z)$ from regressing D on Z, and then regress Y on $\widehat{E}(D|Z)$; this is the 2SLS procedure. The standard error for $\widehat{\text{CACE}}_{\text{2SLS}}$ is not the standard error from the second-stage regression but needs to account for the sampling uncertainty in using $\widehat{E}(D|Z)$ as an estimate of $E(D|Z)$; see White (1984), Davidson and MacKinnon (1993), Imbens and Angrist (1994), and Freedman (2009: Chapter 9.8). Specifically, the asymptotic standard error for $\widehat{\text{CACE}}_{\text{2SLS}}$ is given in Imbens and Angrist (1994: Theorem 3). The 2SLS estimator $\widehat{\text{CACE}}_{\text{2SLS}}$ can also be written as $\frac{\widehat{\text{Cov}}(Y,Z)}{\widehat{\text{Cov}}(D,Z)}$ (Davidson and MacKinnon, 1993).

The 2SLS estimator does not take into full account the structure in **Table 1** that the observed outcomes are mixtures of outcomes from different compliance classes. Imbens and Rubin (1997a,b) and Cheng et al. (2009a,b) develop approaches to estimating the CACE that use the mixture structure to improve efficiency. For example, Cheng et al. (2009b) develop an empirical likelihood approach that is consistent under the same assumptions as 2SLS but that provides substantial efficiency gains in some finite sample settings. Sobel and Muthén (2012) develop an approach that assumes that the compliers are a mixture of subjects who are not affected by the treatment and subjects who are affected by the treatment, and estimates the proportion of subjects who are not affected by the treatment.

Robust Standard Errors for 2SLS

When there is cluster assignment in the randomized trial, standard errors that are robust to clustering should be computed. For 2SLS, this can be done by using robust Huber–White standard errors (White, 1982). Code for computing robust Huber–White standard errors for IV analysis with clustering in R is given in Section Example: the PROSPECT Study.

Even when there is no clustering we recommend always using the robust Huber–White standard errors for 2SLS, as the nonrobust standard error's correctness requires additional strong assumptions about the relationships between the different compliance classes' outcome distributions and homoskedasticity while the robust standard error's correctness does not require these assumptions; see Imbens and Angrist (1994: Theorem 3) and Angrist and Pischke (2009: Section 4.2.1). Code for computing robust Huber–White standard errors without clustering in R is available in the ivpack package in R, using the command robust.se.

Treatment Effect on the Whole Distribution of the Outcome

In addition to the average effect (CACE) discussed above, knowledge of the causal effect of a treatment on the outcome distribution and its general functions can often provide additional insights into the impact of the treatment and therefore be of significant interest in many situations (Poulson et al., 2012). For example, in a study of the effect of school-subsidized meal programs on children's weight, both low weight and high weight are adverse outcomes; therefore, knowing the effect of the program on the entire distribution of outcomes rather than just average weight is important for understanding the impact of the program. For an individual patient deciding which treatment to take, the patient must weigh the effects of the possible treatments on the distribution of outcomes, the costs of the treatments, and the potential side effects of the treatments (Hunink et al., 2001). Therefore, making the best decision requires information on the treatment's effect on the entire distribution of outcomes rather than just the average effect because a patient's utility over outcomes may be nonlinear over the outcome scale (Karni, 2009; Pliskin et al., 1980). Hogan and Lee (2004), Saigal et al. (1999), and Sommers et al. (2007) provide examples in HIV care, neonatal care, and cancer care respectively.

Abadie (2002) provided the standard IV estimate on the counterfactual cumulative distribution functions (cdfs) of the outcome of compliers with and without the treatment:

$$\widehat{H}_{c1}(y)^{\mathrm{SIV}} = \frac{\widehat{E}\{1(Y_i \le y)D_i|Z_i = 1\} - \widehat{E}\{1(Y_i \le y)D_i|Z_i = 0\}}{\widehat{E}(D_i|Z_i = 1) - \widehat{E}(D_i|Z_i = 0)}$$

$$\widehat{H}_{c0}(y)^{\mathrm{SIV}} = \frac{\widehat{E}\{1(Y_i \le y)(1-D_i)|Z_i = 1\} - \widehat{E}\{1(Y_i \le y)(1-D_i)|Z_i = 0\}}{\widehat{E}\{(1-D_i)|Z_i = 1\} - \widehat{E}\{(1-D_i)|Z_i = 0\}} \quad [5]$$

However, these standard IV estimates of the potential cdfs for compliers may not be nondecreasing functions as cdfs should be (Abadie, 2002; Imbens and Rubin, 1997b). Furthermore, the standard IV approach does not make full use of the mixture structure implied by the latent compliance class model (see **Table 1**) and hence could be inefficient (Imbens and Rubin, 1997b). Instead, Imbens and Rubin (1997b) proposed a normal approximation and two multinomial approximations to the outcome distributions. However, the estimator based on a normal approximation could be biased when the outcomes are not normal, and for the approach based on multinomial approximations, a systematic approach for choosing the multinomial approximations is needed. Cheng et al. (2009a) adopted a density ratio model (Anderson, 1979) to relate the potential outcome densities of latent compliance classes by an exponential tilt model and then maximized the empirical likelihood for semiparametric estimates of potential outcome densities and cdfs. Their approach makes full use of the mixture structure implied by the latent compliance class model without parametric assumptions on the outcome distributions, takes into account the nondecreasing property of cdfs and can be easily computed from the data. Their method can be applied to general outcomes and general functions of outcome distributions. Cheng et al. (2009a) showed that their estimator has good properties and is substantially more efficient than the standard IV estimator eqn [5].

Example: The PROSPECT Study

Table 2 shows the data from the PROSPECT study. This study was a single-consent randomized trial in which people could only receive the treatment if assigned to it. Among the 253 people assigned to treatment, 18 (7.4%) did not take the treatment. The ITT estimate for the effect of the treatment is $-7.46 - (-3.93) = -3.53$, i.e., a reduction in depression of 3.53 Hamilton points. The as-treated estimate is $-7.58 - (-4.07) = -3.51$. The per-protocol estimate is $-7.58 - (-3.93) = -3.65$. The as-treated and per-protocol estimates may be biased because the people who adhere to treatment when assigned to it (compliers) may be different types of people than never-takers; in contrast, the IV estimate is not biased if these groups of people are different. The IV estimate is $\frac{-7.46-(-3.93)}{235/253} = -3.80$ with a 95% confidence interval of $(-1.95, -5.65)$. The relationship between the as-treated and IV estimates suggests that compliers would have less of an improvement in depression than never-takers if both groups are given usual care.

Software for computing IV estimates is available in R, SAS, and STATA. The following R code using the library ivpack computes the IV estimate and cluster robust standard errors (the cluster is practice):

```
> library(ivpack)
> ivmodel=ivreg(hamiltonchange ~ treatreceivem4, ~interven)
> cluster.robust.se(ivmodel, practice)
[1] "Cluster Robust Standard Errors"

t test of coefficients:

               Estimate Std. Error t value  Pr(>|t|)
(Intercept)    -3.92496    0.61022 -6.4320 3.022e-10 ***
treatreceivem4 -3.80461    0.94187 -4.0394 6.228e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Cheng et al. (2009a) applied the method in Section Treatment Effect on the Whole Distribution of the Outcome to analyze the effect of the treatment in the PROSPECT study on the distribution of outcomes among compliers. The outcome distributions of compliers under treatment and control are significantly different ($P < 0.01$) and the treatment appears to have more of an effect on compliers with worse depression scores.

Censoring

For survival outcomes, a common complication is censoring. For example, one type of censoring is administrative censoring, which means that a subject survived until the end of the time window of the study. For estimating the effect of a treatment on a survival outcome using an IV, Robins and Tsiatis (1991) considered a structural accelerated failure time model and developed semiparametric estimators for this model. Joffe (2001) provided a good discussion of their approach and comparisons with other survival analysis methods. Loeys and Goetghebeur (2003) and Cuzick et al. (2007) considered a structural proportional hazards model in which the hazard of the potential failure time under treatment for a certain group of subjects is proportional to the hazard of the potential failure time under control for these same subjects. Both the structural accelerated failure time model and the structural proportional hazards model are semiparametric models, where the effect of

Table 2 Data from the PROSPECT study

Randomization assignment	*Treatment received*	*No. of People*	*Mean Hamilton change*
Control	Control	234	−3.93
Treatment	Control	18	−5.89
Treatment	Treatment	235	−7.58
Treatment	Control or treatment	18 + 235 = 253	−7.46
Control or treatment	Control	234 + 18 = 252	−4.07

the treatment on the distribution of failure times is modeled parametrically.

Baker (1998) extended the models and assumptions for discrete-time survival data and derived closed-form expressions for estimating the difference in the hazards at a specific time between compliers under treatment and control based on maximum likelihood. Baker's (1998) estimator is analogous to the standard IV estimator for a survival outcome. Nie et al. (2011) discussed this standard IV approach and parametric maximum-likelihood methods for the difference in survival at a specific time between compliers under treatment and control.

Let $S_{c1}(V)$, $S_{c0}(V)$, $S_{at}(V)$, and $S_{nt}(V)$ be the potential survival functions at time V of compliers in the treatment and control groups and of always-takers and never-takers respectively, $S_z(V)$ be the survival probabilities at time V for the group with assignment $Z = z$, and $S_{zd}(V)$ be the survival probabilities at time V for the group with assignment $Z = z$ and treatment received $D = d$. By **Table 1** and the monotonicity assumption IV-A5 that rules out defiers, the following holds:

$$\begin{aligned}
S_1(V) &= \pi_c S_{c1}(V) + \pi_{at} S_{at}(V) + \pi_{nt} S_{nt}(V),\\
S_{11}(V) &= \frac{\pi_c}{\pi_c + \pi_{at}} S_{c1}(V) + \frac{\pi_{at}}{\pi_c + \pi_{at}} S_{at}(V)\\
S_{10}(V) &= S_{nt}(V)\\
S_0(V) &= \pi_c S_{c0}(V) + \pi_{at} S_{at}(V) + \pi_{nt} S_{nt}(V),\\
S_{00}(V) &= \frac{\pi_c}{\pi_c + \pi_{nt}} S_{c0}(V) + \frac{\pi_{nt}}{\pi_c + \pi_{nt}} S_{nt}(V)\\
S_{01}(V) &= S_{at}(V)
\end{aligned}$$

Similar to the standard IV estimator for the CACE eqn [4], the standard IV estimator for the compliers difference in survival probabilities at time V under treatment versus control is the following, where $\widehat{S}_z(V)$ is the Kaplan–Meier estimator of the survival probability at time V for the group with assignment $Z = z$,

$$\widehat{S}_{c1}(V) - \widehat{S}_{c0}(V) = \frac{\widehat{S}_1(V) - \widehat{S}_0(V)}{\widehat{E}(D|Z = 1) - \widehat{E}(D|Z = 0)} \qquad [6]$$

In addition to IV-A1 to IV-A5, an additional assumption is needed to ensure that the estimator [6] is consistent for the compliers difference in survival probabilities at time V:

> Independence assumption of failure times and censoring times: *The distributions of potential failure times T and censoring times C are independent of each other. Type I censoring (i.e., censoring times are the same for all subjects) and random censoring are two special cases.*

Although the standard IV estimator [6] is very useful, it may give negative estimates for hazards and be inefficient because it does not make full use of the mixture structure implied by the latent compliance model. When the survival functions follow some parametric distributions, Nie et al. (2011) used the Expectation-Maximation (EM) algorithm to obtain the maximum likelihood estimate (MLE) on the difference in survival probabilities for compliers. However, the MLEs could be biased when the parametric assumptions are not valid. To address this concern, Nie et al. (2011) developed a nonparametric estimator based on empirical likelihood that makes use of the mixture structure to gain efficiency over the standard IV method while not depending on parametric assumptions to be consistent.

Missing Response Data

In addition to noncompliance, missing response data are a concern in many randomized studies. For example, Ferro et al. (1996) conducted a trial of methods for teaching breast self-examination (BSE) for breast cancer screening in which women were randomly assigned to either a standard treatment of receiving mailed information about BSE ($Z = 0$) or an enhanced treatment of attendance in a self-exam course in addition to receiving the mailed information ($Z = 1$). A woman was considered to have received the treatment if she attended the self-exam course. Of the women assigned the self-exam course, only 55% of the women actually attended the course. The primary outcome for the study was whether or not a woman practiced BSE 1 year after the beginning of the study, where the outcome was obtained from a mailed self-administered questionnaire. Only 65% of the women returned the questionnaire.

Frangakis and Rubin (1999) provide a method-of-moments estimate of the CACE when there are missing response data under two additional assumptions besides IV-A1 to IV-A5. Let $R_i^{z=1} = 1$ or 0 indicate whether or not subject i would respond if he or she were assigned to the treatment, and let $R_i^{z=0}$ indicate whether or not subject i would respond if he or she were assigned to the control. The two additional assumptions made by Frangakis and Rubin (1999) are (1) latent ignorability, which means that potential outcomes are independent of potential responses given a subject's compliance class, $(Y^1, Y^0) \perp\!\!\!\perp (R^{z=1}, R^{z=0})|C$ and (2) the response ER for never-takers and always-takers, which says that assignment to treatment versus control does not affect whether the subject will respond, i.e., $R^{z=1} = R^{z=0}$ if $C = at$ or nt. Frangakis and Rubin (1999) show that the usual 2SLS estimate of the CACE [4] is biased when there are missing response data but provide a consistent estimator under the additional assumptions (1) and (2). Mealli et al. (2004) discuss that sometimes, instead of assuming the response ER for always-takers and never-takers, it may be more reasonable to assume a response ER for compliers and always-takers, $R^{z=1} = R^{z=0}$ if $C = at$ or $C = co$, and provide a consistent estimator of the CACE under IV-A1 to IV-A5, latent ignorability, and the response ER for compliers and always-takers. Mealli et al. (2004) argue that in the breast cancer screening study mentioned above, the response ER for never-takers is not so plausible. Never-takers who were assigned to the new treatment and declined participation have explicitly refused to comply with their assigned (active) treatment which might induce them to refuse to respond to the posttest questionnaire in comparison to never-takers who receive the standard treatment and have not explicitly refused to comply with it. The response ER seems more plausible for compliers as compliers are willing to follow the protocol in their assigned treatment, making it more plausible that they would not be affected in their response behavior by that assignment. Chen et al. (2009) and Small and Cheng (2009) discuss other assumptions under which the CACE is identified when there are missing response data.

Sensitivity Analysis for Violations of Assumptions

Since most of the IV assumptions cannot be completely tested, in this section, we will discuss sensitivity analysis to examine how the estimate of treatment effect will change if a proposed IV violates the assumptions by certain magnitudes.

Sensitivity Analysis for the ER

There is sometimes uncertainty about whether the treatment assignment satisfies IV-A4, the ER. For example, in the PROSPECT study, even if the patient did not adhere to the treatment (i.e., never saw the depression specialist), the depression specialist's presence in a practice may have helped improve the treatment of depression in that practice by primary care physicians. Even in blinded studies, the ER may not hold. For example, among subjects considered not to have received the experimental treatment, those assigned to the intervention arm might have received a larger partial dose (although considered inadequate), while those assigned to the control arm might have received a smaller partial dose or no dose at all (Ten Have et al., 2008). To the extent that partial dose might lead to some benefits, not receiving treatment when assigned to treatment in the experimental arm might be better than not receiving treatment in the control arm. Conversely, if there is a placebo effect, then not receiving treatment in the experimental arm (which means one would not have benefited from a placebo effect) might be worse than not receiving treatment in the control arm (which means one might still have benefited from a placebo effect). In either case, the ER would be violated because assignment to the treatment impacts outcomes through a pathway other than treatment received.

We can carry out a sensitivity analysis for the impact of violations of the ER in the PROSPECT study as follows. A model for a direct effect of the treatment on never-takers is

$$\begin{aligned} Y_i^{z,d} &= \alpha + \beta d + \lambda z(1-d) + e_i \\ E(e_i|Z_i) &= 0 \end{aligned} \qquad [7]$$

The direct effect of being assigned to the treatment on never-takers is λ. Under model [7], a consistent estimate of and inferences for β can be obtained by carrying out a 2SLS analysis with $Y_i - \lambda(1 - Z_i)D_i$ as the outcome variable, D_i as the treatment variable, and Z_i as the IV. For the PROSPECT study, if $\lambda = -1$ so that a depression specialist's presence in a practice improves the treatment of depression by 1 Hamilton point for subjects who do not see the depression specialist, we estimate that the CACE is -3.73 with a 95% confidence interval of $(-1.87, -5.59)$; this compares to the estimate of -3.80 with a 95% confidence interval of $(-1.95, -5.65)$ when the ER holds. The IV results for the PROSPECT study are not highly sensitive to plausible departures from the ER.

Sensitivity Analysis for Violation of the Monotonicity Assumption

In the PROSPECT study, the monotonicity assumption IV-A5 cannot be violated because no one can take the treatment when assigned to the control. But in trials where it is possible to take the treatment if assigned to control or take the control if assigned to treatment, it is possible that the monotonicity assumption could be violated. Suppose that IV-A1 to IV-A4 hold but IV-A5, the monotonicity assumption, is violated. Then, the 2SLS estimate eqn [4] converges to the CACE plus a bias term (Angrist et al., 1996):

$$\begin{aligned} \widehat{\mathrm{CACE}}_{2\mathrm{SLS}} \xrightarrow{p} \mathrm{CACE} - \frac{P(C=de)}{P(C=co)-P(C=de)} \\ \times \left\{E\left[Y^1 - Y^0 \middle| C = de\right] - E\left[Y^1 - Y^0 \middle| C = co\right]\right\} \end{aligned} \qquad [8]$$

(this can be proved by considering the effect of violations of monotonicity in the derivation of eqn [2]). The bias due to violations of monotonicity is composed of two factors. The first factor, $-\frac{P(C=de)}{P(C=co)-P(C=de)}$, is related to the proportion of defiers and is equal to zero under the monotonicity assumption. The smaller the proportion of defiers, the smaller the bias will be from violations of the monotonicity assumption. However, if the IV is weak in the sense that the average effect of the IV Z on D is small, then even a few defiers could generate a large bias; the stronger the IV, the lesser sensitive the IV estimate is to violations of the monotonicity assumption (Angrist et al., 1996). The second factor, $E[Y^1 - Y^0|C = de] - E[Y^1 - Y^0|\ C = co]$, is the difference in the average treatment effect between defiers and compliers. If the average treatment effect is identical for defiers and compliers, violations of the monotonicity assumption generate no bias. The lesser variation there is in the average treatment effect among defiers and compliers, the smaller the bias from violations of the monotonicity assumption.

Extrapolating the CACE to the Full Population

As discussed in Section IV Approach, the IV method estimates the CACE, the average treatment effect for the compliers (i.e., $E[Y^1 - Y^0|C = co]$). The average treatment effect in the population is, under the monotonicity assumption, a weighted average of the average treatment effect for the compliers, the average treatment effect for the never-takers, and the average treatment effect for the always-takers:

$$\begin{aligned} E\left[Y^1 - Y^0\right] &= P(C = co)E\left[Y^1 - Y^0 \middle| C = co\right] + P(C = at) \\ &\times E\left[Y^1 - Y^0 \middle| C = at\right] + P(C = nt)E\left[Y^1 - Y^0 \middle| C = nt\right] \end{aligned}$$

The IV method provides no direct information on the average treatment effect for always-takers (i.e., $E[Y^1 - Y^0|\ C = at]$) or the average treatment effect for never-takers (i.e., $E[Y^1 - Y^0|C = nt]$). However, the IV method can provide useful bounds on the average treatment effect for the whole population if a researcher is able to put bounds on the difference between the average treatment effect for compliers and the average treatment effects for never-takers and always-takers based on subject matter knowledge. For example, suppose a researcher is willing to assume that this difference is no more than b. Then

$$\begin{aligned} E\left[Y^1 - Y^0 \middle| C = co\right] - b\left[1 - P(C = co)\right] \le E\left[Y^1 - Y^0\right] \\ \le E\left[Y^1 - Y^0 \middle| C = co\right] + b\left[1 - P(C = co)\right] \end{aligned} \qquad [9]$$

where the quantities on the left and right hand sides of eqn [9] other than b can be estimated as discussed in Section IV Approach and Imbens and Rubin (1997b), Hirano et al. (2000), and Cheng et al. (2009a). For binary or other bounded outcomes, the boundedness of the outcomes can be used to tighten bounds on the average treatment effect for the whole population or other treatment effects (Balke and Pearl, 1997; Cheng and Small, 2006). Qualitative assumptions, such as that the average treatment effect is larger for always-takers than compliers, can also be used to tighten the bounds, e.g., Cheng and Small (2006), Bhattacharya et al. (2008), and Siddique (2013).

In thinking about extrapolating the CACE to the full population, it is useful to think about how compliers' outcomes compare to always-takers' and never-takers' outcomes (Imbens and Wooldridge, 2007). The data provide some information about this. Specifically, we can compare

$$\begin{aligned} &E(Y^0|C = nt) \text{ vs } E(Y^0|C = co) \text{ and} \\ &E(Y^1|C = at) \text{ vs } E(Y^1|C = co) \end{aligned} \qquad [10]$$

Under IV-A1 to IV-A5, Abadie (2002) shows that the following are consistent estimates of the quantities in eqn [10] when there are no covariates: $\widehat{E}(Y^0|C = nt) = \widehat{E}(Y|D = 0, \ Z = 1)$, $\widehat{E}(Y^0|C = co) = \frac{\widehat{E}(Y(1-D)|Z=1) - \widehat{E}(Y(1-D)|Z=0)}{\widehat{E}(1-D|Z=1) - \widehat{E}(1-D|Z=0)}$, $\widehat{E}(Y^1|C = at) = \widehat{E}(Y|D = 1, Z = 0)$, and $\widehat{E}(Y^1|C = co) = \frac{\widehat{E}(YD|Z=1) - \widehat{E}(YD|Z=0)}{\widehat{E}(D|Z=1) - \widehat{E}(D|Z=0)}$; when there are covariates, the methods of Abadie (2003) or Hirano et al. (2000) can be used to estimate the quantities in eqn [10]. If compliers, never-takers, and always-takers are found to be substantially different in levels by evidence of a substantial difference between $E(Y^0|C = nt)$ and $E(Y^0|C = co)$ and/or between $E(Y^1|C = at)$ and $E(Y^1|C = co)$, then it appears much less plausible that the average effect for compliers is indicative of average effects for other compliance types. On the other hand, if one finds that potential outcomes given the control for never-takers and compliers are similar, and potential outcomes given the treatment are similar for compliers and always-takers, it is more plausible that average treatment effects for the groups are also comparable (Imbens and Wooldridge, 2007). For example, in the PROSPECT study described in **Table 2**, the mean change in Hamilton score for never-takers is estimated to be −5.89 and for compliers under control is −3.78. This difference in estimated changes in Hamilton scores between the compliance classes suggests that we should be cautious in extrapolating the CACE to the full population.

Predictors of Compliance

When there is noncompliance, it is often of interest to understand what the predictors of compliance or noncompliance are. We can compare the mean of a covariate for compliers to the mean for never-takers and always-takers. The mean of a covariate X_i among the compliers is the following under the IV assumptions IV-A1 to IV-A5, where f represents the probability mass function or probability density function,

$$\begin{aligned} E[X|C = co] &= \int x \frac{f(x|C = co)}{f(x)} f(x)\mathrm{d}x \\ &= E\left[X \frac{f(X|C = co)}{f(X)}\right] \end{aligned} \qquad [11]$$

where

$$\begin{aligned} \frac{f(x|C = co)}{f(x)} &= \frac{\frac{P(C=co|x)f(x)}{P(C=co)}}{f(x)} = \frac{P(C = co|x)}{P(C = co)} \\ &= \frac{E(D|Z = 1, \ X = x) - E(D|Z = 0, \ X = x)}{E(D|Z = 1) - E(D|Z = 0)} \end{aligned} \qquad [12]$$

We estimate $E(X|C = co)$ by estimating $\theta_i = \frac{f(X_i|C=co)}{f(X_i)}$ for $i = 1, \ldots, N$ based on eqn [12] (e.g., using logistic regression to estimate $E(D|Z, X)$ and then plugging into eqn [12]) and then taking the sample average of $X_i\theta_i$. For a binary characteristic X, eqn [11] simplifies to

$$P(X = 1)\left[\frac{E(D|Z = 1, X = 1) - E(D|Z = 0, X = 1)}{E(D|Z = 1) - E(D|Z = 0)}\right]$$

We can estimate the mean of X for always-takers and never-takers by $\widehat{E}(X|D = 1, \ Z = 0)$ and $\widehat{E}(X|D = 0, \ Z = 1)$ respectively.

For the PROSPECT study, there are only compliers and never-takers. **Table 3** shows that compliers are significantly more depressed at baseline than never-takers.

Multiarm Trials

Multiarm trials with noncompliance are common. Cheng and Small (2006) consider the setting of a treatment with three levels – control (0) and two active levels A and B, where A and B are not ordered by dose and some subjects may prefer A to B and some may prefer B to A. Subjects are randomly assigned to one of the three arms 0, A, and B, and then could either take the assigned treatment or not take it and receive the control (for the control arm, all subjects receive the control 0). The effect of treatment A versus control for subjects who would take treatment A if offered it (i.e., compliers with treatment A) is identified by analyzing only subjects who were either assigned to the control arm or the treatment A arm. But for this setting, Cheng and Small (2006) showed that the effect of treatment A for subjects who would take treatment A if assigned to it but not treatment B and the effect of treatment A for subjects who would take treatments A or B if assigned to A or B respectively are not point identified. However, the data provide information that can be used to narrow bounds on these treatment effects. These treatment effects are of interest for individuals making

Table 3 Comparison of the characteristics of compliers and never-takers. The *p*-value is for testing the null hypothesis that $E(X|C = co) = E(X|C = nt)$

X	$\widehat{E}(X\|C = co)$	$\widehat{E}(X\|C = nt)$	p-*value*
Baseline suicidal ideation	0.30	0.28	0.83
Age	70.15	74.11	0.10
White	0.74	0.83	0.34
Baseline Hamilton	19.08	13.83	<0.01

decisions about which treatment to take, e.g., for a very compliant subject who knows he or she would take either treatment A or B if offered, he or she would like to know whether treatment A or B is better among very compliant subjects like himself or herself; the treatment effects are also of interest for clinicians deciding which treatment to offer first and for health policy makers anticipating what would happen were the treatment(s) to be introduced into general practice in a setting in which compliance patterns are expected to differ from those of the trial (Cheng and Small, 2006).

Roy et al. (2008) consider the related setting of a randomized trial with two active treatments and noncompliance to each treatment. Roy et al. (2008) use compliance-predictive covariates to help obtain information on causal effects within compliance classes. Although these causal effects are not point identified without further assumptions, the effects are identified once a sensitivity parameter describing the correlation between a subject's compliance under the two treatment arms is specified.

Partial Compliance

We have considered compliance with a treatment to be all or none. In some studies, subjects may partially comply with the treatment. Jin and Rubin (2008), Bartolucci and Grilli (2011), and Ma et al. (2011) develop approaches for estimating the compliance class specific causal effects for these settings. These methods involve additional assumptions beyond IV-A1 to IV-A5 to deal with the more complicated setting. Another approach to deal with partial compliance is to consider assumptions about causal effects being homogeneous across different compliance classes. For example, consider the model that there is a constant effect of a unit of treatment β, e.g., $Y^d = Y^0 + \beta d$, where d represents the proportion of dose a subject takes, $0 \leq d \leq 1$. Then, under assumptions analogous to IV-A1 to IV-A4, 2SLS (i.e., regressing D on Z and then regressing Y on $\widehat{D}$) provides a consistent estimate of β (Holland, 1988). For approaches that make homogeneity assumptions that are weaker, see Robins (1994) and Tan (2010). Even if the treatment effect is not constant, the two-stage least estimator still converges to a weighted average of the effect of one unit changes in the treatment level as long as an extended monotonicity assumption holds (Angrist and Imbens, 1995). The extended monotonicity assumption is that $D_i^{z=1} \geq D_i^{z=0}$.

See also: Causal Inference: Overview; Instrumental Variables in Statistics and Econometrics; Longitudinal Causal Inference; Mediation, Statistical; Missing Data.

Bibliography

Abadie, A., 2002. Bootstrap tests for distributional treatment effects in instrumental variable models. Journal of the American Statistical Association 97, 284–292.

Abadie, A., 2003. Semiparametric instrumental variable estimation of treatment response models. Journal of Econometrics 113, 231–263.

Anderson, J.A., 1979. Multivariate logistic compounds. Biometrika 66, 17–26.

Angrist, J.D., Imbens, G.W., 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. Journal of the American Statistical Association 90, 430–442.

Angrist, J.D., Pischke, J.-S., 2009. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton.

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91 (434), 444–455.

Baker, S.G., 1998. Analysis of survival data from a randomized trial with all-or-none compliance: estimating the cost-effectiveness of a cancer screening program. Journal of the American Statistical Association 93, 929–934.

Balke, A., Pearl, J., 1997. Bounds on treatment effects for studies with imperfect compliance. Journal of the American Statistical Association 92 (439), 1171–1176.

Bartolucci, F., Grilli, L., 2011. Modeling partial compliance through copulas in a principal stratification framework. Journal of the American Statistical Association 106, 469–479.

Bhattacharya, J., Shaikh, A.M., Vytlacil, E., 2008. Treatment effect bounds under monotonicity assumptions: an application to swan-ganz catheterization. American Economic Review 98 (2), 351–356.

Bruce, M., Ten Have, T., Reynolds III, C., Katz, I., Schulberg, H., Mulsant, B., Brown, G., McAvay, G., Pearson, J., Alexopoulos, G., 2004. Reducing suicidal ideation and depressive symptoms in depressed older primary care patients: a randomized trial. Journal of the American Medical Association 291, 1081–1091.

Chen, H., Geng, Z., Zhou, X., 2009. Identifiability and estimation of causal effects in randomized trials with noncompliance and completely non-ignorable missing data (with discussion). Biometrics 65, 675–691.

Cheng, J., Small, D., 2006. Bounds on causal effects in three-arm trials with noncompliance. Journal of the Royal Statistical Society, Series B 68 (5), 815–836.

Cheng, J., Qin, J., Zhang, B., 2009a. Semiparametric estimation and inference for distributional and general treatment effects. Journal of the Royal Statistical Society, Series B (Statistical Methodology) 71 (4), 881–904.

Cheng, J., Small, D.S., Tan, Z., Ten Have, T.R., 2009b. Efficient nonparametric estimation of causal effects in randomized trials with noncompliance. Biometrika 96 (1), 19–36.

Cuzick, J., Sasieni, P., Myles, J., Tyler, J., 2007. Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination. Journal of the Royal Statistical Society, Series B (Methodological) 69, 565–588.

Davidson, R., MacKinnon, J., 1993. Estimation and Inference in Econometrics. Oxford University Press, New York.

Ferro, S., Caroli, A., Biggeri, A., Gambi, A., 1996. A randomized trial on breast self examination in Faenza (northern Italy). Tumori 82, 329–334.

Frangakis, C.E., Rubin, D.B., 1999. Addressing complications of intent-to-treat analysis in the presence of all-or-none treatment noncompliance and subsequent missing outcomes. Biometrika 86, 365–379.

Freedman, D.A., 2009. Statistical Models: Theory and Practice. Cambridge University Press, Cambridge.

Hernán, M.A., Hernández-Díaz, S., 2012. Beyond the intention to treat in comparative effectiveness research. Clinical Trials 9, 48–55.

Hirano, K., Imbens, G.W., Rubin, D.B., Zhou, X.H., 2000. Assessing the effect of an influenza vaccine in an encouragement design. Biostatistics 1 (1), 69–88.

Hogan, J.W., Lee, J.Y., 2004. Marginal structural quantile models for longitudinal observational studies with time-varying treatment. Statistica Sinica 14, 927–944.

Holland, P.W., 1988. Causal inference, path analysis, and recursive structural equations models. Sociological Methodology 18, 449–484.

Hunink, M., Glasziou, P., Siegel, J., Weeks, J., Pliskin, J., Elstein, A., Weinstein, M., 2001. Decision Making in Health and Medicine: Integrating Evidence and Values. Cambridge University Press, Cambridge.

Imbens, G.W., Angrist, J.D., 1994. Identification and estimation of local average treatment effects. Econometrica 62.

Imbens, G.W., Rubin, D.B., 1997a. Estimating outcome distributions for compliers in instrumental variables models. The Review of Economic Studies 64 (4), 555–574.

Imbens, G.W., Rubin, D.B., 1997b. Bayesian inference for causal effects in randomized experiments with noncompliance. The Annals of Statistics, 305–327.

Imbens, G.W., Wooldridge, J.M., Summer 2007. Lecture Notes 5, Instrumental Variables with Treatment Effect Heterogeneity: Local Average Treatment Effects. What's New in Econometrics. National Bureau of Economic Research.

Jin, H., Rubin, D.B., 2008. Principal stratification for causal inference with extended partial compliance. Journal of the American Statistical Association 103, 101–111.

Joffe, M.M., 2001. Administrative and artificial censoring in censored regression models. Statistics in Medicine 20, 2287–2304.

Karni, E., 2009. A theory of medical decision making under uncertainty. Journal of Risk and Uncertainty 39, 1–16.

Loeys, T., Goetghebeur, E., 2003. A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. Biometrics 59, 100–105.
Ma, Y., Roy, J., Marcus, B.H., 2011. Causal models for randomized trials with two active treatments and continuous compliance. Statistics in Medicine 30, 2349–2362.
May, G.S., Chir, B., DeMets, D.L., Friedman, L.M., Furberg, C., Passamani, E., 1981. The randomized clinical trial: bias in analysis. Circulation 64, 669–673.
Mealli, F., Imbens, G.W., Ferro, S., Biggeri, A., 2004. Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes. Biostatistics 5, 207–222.
Neyman, J., 1990. On the application of probability theory to agricultural experiments. Statistical Science 5, 463–480.
Nie, H., Cheng, J., Small, D.S., 2011. Inference for the effect of treatment on survival probability in randomized trials with noncompliance and administrative censoring. Biometrics 67, 1397–1405.
Pliskin, J.S., Shepard, D.S., Weinstein, M.C., 1980. Utility functions for life years and health status. Operations Research 28, 206–224.
Poulson, R.S., Gadbury, G.L., Allison, D.B., 2012. Treatment heterogeneity and individual qualitative interaction. The American Statistician 66, 16–24.
Robins, J.M., 1994. Correcting for non-compliance in randomized trials using structural nested mean models. Communications in Statistics, Theory and Methods 23, 2379–2412.
Robins, J.M., Tsiatis, A.A., 1991. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. Communications in Statistics, Theory and Methods 20, 2609–2631.
Roy, J., Hogan, J.W., Marcus, B.H., 2008. Principal stratification with predictors of compliance for randomized trials with 2 active treatments. Biostatistics 9, 277–289.
Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and non-randomized studies. Journal of Educational Psychology 66, 688–701.
Rubin, D.B., 1990. Formal modes of statistical inference for causal effects. Journal of Statistical Planning and Inference 25, 279–292.
Saigal, S., Stoskopf, B.L., Feeny, D., Furlong, W., Burrows, E., Rosenbaum, P.L., Hoult, L., 1999. Differences in preferences for neonatal outcomes among health care professionals, parents, and adolescents. Journal of the American Medical Association 281, 1991–1997.
Sheiner, L., Rubin, D., 1995. Intention-to-treat analysis and the goal of clinical trials. Clinical Pharmacology & Therapeutics 56, 6–10.
Siddique, Z., 2013. Partially identified treatment effects under imperfect compliance: the case of domestic violence. Journal of the American Statistical Association 108, 504–513.
Small, D., Cheng, J., 2009. Discussion of "Identifiability and estimation of causal effects in randomized trials with noncompliance and completely non-ignorable missing data". Biometrics 65, 682–686.
Small, D.S., Ten Have, T.R., Joffe, M.M., Cheng, J., 2006. Random effects logistic models for analysing efficacy of a longitudinal randomized treatment with non-adherence. Statistics in Medicine 25, 1981–2007.
Sobel, M., Muthén, B., 2012. Compliance mixture modelling with a zero-effect complier class and missing data. Biometrics 68, 1037–1045.
Sommer, A., Zeger, S., 1991. On estimating efficacy from clinical trials. Statistics in Medicine 10, 45–52.
Sommers, B.D., Beard, C.J., Dahl, D., D'Amico, A.V., Kaplan, I.P., , et al.Richie, J., 2007. Decision analysis using individual patient preferences to determine optimal treatment for localized prostate cancer. Cancer 110, 2210–2217.
Tan, Z., 2010. Marginal and nested structural models using instrumental variables. Journal of the American Statistical Association 105 (489), 157–169.
Ten Have, T., Elliott, M., Joffe, M., Zanutto, E., Datto, C., 2004. Causal models for randomized physician encouragement trials in treating primary care depression. Journal of the American Statistical Association 99, 16–25.
Ten Have, T.R., Normand, S.-L., Marcus, S., Brown, C., Lavori, P., Duan, N., 2008. Intent-to-treat vs. non-intent-to-treat analyses under treatment non-adherence in mental health randomized trials. Psychiatric Annals 38, 772–783.
Vansteelandt, S., Bowden, J., Babnezhad, M., Goetghebeur, E., 2011. On instrumental variables estimation of causal odds ratios. Statisitcal Science 26 (3), 403–422.
Wald, A., 1940. The fitting of straight lines if both variables are subject to error. Annals of Mathematical Statistics 11, 284–300.
White, H., 1982. Instrumental variables regression with independent observations. Econometrica 50, 483–499.
White, H., 1984. Asymptotic Theory for Econometricians. Academic Press, New York.
Zelen, M., 1979. A new design for randomized clinical trials. The New England Journal of Medicine 300, 1242–1245.

Compliance and Legal Authority

Jonathan Jackson, London School of Economics and Political Science, London, UK
Tom R Tyler, Yale University, New Haven, CT, USA
Mike Hough, University of London, London, UK
Ben Bradford, University of Oxford, Oxford, UK
Avital Mentovich, University of California Los Angeles, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by T.R. Tyler, volume 4, pp. 2440–2445, © 2001, Elsevier Ltd.

Abstract

This article addresses the question of how public compliance with the criminal law is cultivated and sustained. We first consider the empirical evidence for instrumental and normative modes of compliance and social regulation. After arguing that normative compliance with the criminal law is ethically and practically preferable to that secured by instrumental models of crime control, we outline some future directions of research into procedural justice and legitimacy.

According to Bottoms (2002) there are four types of explanation for compliance with authority in general and with the criminal law in particular: (1) prudential or self-interested calculations about the potential costs and benefits of punishment, which take into account the risks and costs of punishment; (2) normative considerations about the 'rights and wrongs' of non-compliance; (3) the impact of obstructive strategies, such as locking up offenders to prevent their reoffending, as well as locking up the targets of criminal attention, literally or metaphorically; and (4) habit.

This article addresses the question of how public compliance with the criminal law is cultivated and sustained. It is concerned neither with obstructive (incapacitative) explanations nor with habit – though the latter is arguably the best explanation for why so many of us break the law so infrequently, and one that is much ignored by criminologists (a notable exception being Wikström et al., 2012). Both these explanations are in a sense secondary, in that they presuppose (respectively) that something led to offending at such a level or rate that imprisonment was needed, or else created the habit of compliance with the law.

Our focus in this article is whether normative explanations for compliance – especially those that appeal to the legitimacy of institutions of justice (Tyler, 2006a,b, 2011a,b) – are better in explaining public attitudes toward law enforcement than those that simply invoke the rational calculations of *homo economicus*. Exploring the nature and impact of legal authority, we first consider the empirical evidence for instrumental and normative modes of compliance and social regulation. After describing procedural justice and legitimacy as one of the most effective ways of securing normative compliance with the criminal law – and after arguing that such compliance is ethically and practically preferable to that secured by instrumental models of crime control – we outline some future directions for procedural justice research in the context of compliance and legal authority.

Deterrence and Social Regulation

Legal systems generally seek to improve legal compliance by using threat and punishment aimed at deterring people from engaging in criminal behavior (Nagin, 1998). In public and political discourse answers to the question – why do people break the law? – typically revolve around the idea that crime occurs when the criminal justice system provides insufficient likelihood of punishment, or when insufficiently tough sentences are imposed. Mechanisms of coercive social control and credible risks of sanction hope to persuade the rational choice would-be offender that – while otherwise desirable – a criminal act is not worth the risk.

The argument driving deterrence strategies is threefold: that fear of possible future punishment leads to compliance with the law; that risk calculations are partly shaped by both the anticipated likelihood of punishment and by judgments about its severity; and that the focus is (and should be) on the power of legal authorities and institutions to shape behavior by threatening to deliver (or by actually delivering) negative sanctions for rule breaking. This way of viewing the relationship between legal authorities and citizens is referred to as the 'deterrence' or 'social control' model. It is this model of human behavior that (for better or for worse) currently dominates law and public policy.

At the heart of the deterrence model of compliance is rational choice theory. Derived from neoclassical economics (Blumstein et al., 1978), a simple deterrence model assumes that people calculate expected utilities by multiplying the probability of an outcome (e.g., getting caught for armed robbery or drunk driving) by its valence or impact and then balance the result against the benefits of crime (Paternoster, 2006) Rational self-interest is the motivational engine; if laws and sanctions are well calibrated then people will arrive at the conclusion that they should follow the law. It follows that to regulate behavior decision makers should adjust criminal sanctions to the required level; the expected losses associated with law breaking will then minimize the likelihood that people will break the law.

Yet, this is a costly and minimally effective system of social control. It is costly because it assumes the need to create and maintain a credible threat of punishment. People will only change their behavior when they feel that there is a reasonable risk of being caught and punished for wrongdoing. They will also try to hide their illegal behavior, so a system of surveillance

is needed to identify and deter wrongdoing. Such surveillance is often not possible, but even when it is possible, surveillance is costly and its magnitude of influence is at best weak. For instance, an authoritative recent review concluded that some studies have found that: "… punishment weakens compliance, some [have found] that sanctions have no effect on compliance, and some [have found] that the effect of sanctions depends on moderating factors" (Picquero et al., 2011: p. 1; see also Paternoster, 2006). When perceptions of the likelihood of being caught and punished do influence people's behavior, the effect seems to be relatively small. Social control strategies that are based exclusively on a deterrence model of human behavior have had at best limited success. Deterrence is a high-cost strategy that, at best, yields identifiable but weak results.

In fact, under the rational choice/deterrence model we would expect crime to be more appealing and more pervasive than it actually is. Legal systems have become increasingly punitive, but the likelihood of being caught and punished for most criminal behaviors remains very low, and often too low to offset the potential gains of law breaking. Despite this, most people continue to follow the law, most of the time, even if they would bear very little consequence for breaking it.

Normative Modes of Social Regulation

Asking why not people *break the law* but why people *obey the law* brings into sharp focus the fact that most individuals obey most laws most of the time because they think it is the right thing to do, or they have simply acquired the habit of doing so. Formal criminal justice is only one of many systems of social control, most of which have a significant normative dimension (Tyler, 2003; Tyler et al., 2007; Jackson et al., 2012a,b; Mazerolle et al., 2013). Moral norms are learnt in childhood, with greater or lesser success. Institutions like the family, schooling, and (in some countries) religion have an important part to play. Developments in later life may further support or erode these norms.

Importantly for the current article, a good deal of research supports the idea that perceptions of the legitimacy of the institutions of justice play a further role in normative compliance. A legitimate authority commands consent (a sense of obligation to obey) that is grounded in legality and moral validity (Tyler, 2006a,b; Hough et al., 2013a,b). Crucially, if legitimacy encourages people to self-regulate, this obviates the need for expensive and minimally effective deterrence strategies. Models of crime-control that recognize the importance of the legitimacy of justice institutions and the legal system may be more durable – and less costly to a society – than the coercive model that requires a credible deterrent threat. Legitimate institutions can avoid the cost, danger, and alienation that are associated with policies based on external rules underpinned by deterrent threat (Schulhofer et al., 2011).

How does legitimacy shape law-related behavior? The dominant account in the field focuses on consent and authorization. When citizens recognize the legitimacy of an institution, they believe that the institution has the right to prescribe and enforce appropriate behavior, and they feel a corresponding duty to bring their behavior in line with that which is expected. Kelman and Hamilton (1989) refer to legitimacy as 'authorization.' A person authorizes an authority to determine appropriate behavior within a given situation, and then feels obligated to follow the directives or rules that authority establishes. Importantly, the authorization of actions by authorities "seem[s] to carry automatic justification for them. Behaviorally, authorization removes the necessity of making judgments or choices. Not only do normal moral principles become inoperative but – particularly when the actions are explicitly ordered – a different type of morality, linked to the duty to obey superior orders, tends to take over" (Kelman and Hamilton, 1989: p. 16).

Numerous empirical studies have shown that legitimacy judgments predict compliance behavior even after adjusting for perceptions of the morality of a particular act and perceptions of the chances of getting caught (e.g., Sunshine and Tyler, 2003; Jackson et al., 2012a; Papachristos et al., 2012a; Tyler and Jackson, 2014). Perceptions of the legitimacy of the police and the law seem to lead to a respect for legal guidelines for action that dictates what behavior is appropriate and personally binding. These guidelines may not be perfectly aligned with everyone's moral system; we do not always agree with the moral force of each and every law. But legitimacy involves the public recognition that the social order needs a system of laws that generate compliance and respect above and beyond individual preferences. When people believe it is morally just to obey the law, then so long as they know that a particular act is illegal, it becomes by definition wrong to commit it. A different sort of morality has 'kicked in' – one that focuses on the morality of the law in a general, not specific, sense.

Procedural Justice

How do legal authorities generate and sustain legitimacy in the eyes of citizens? According to procedural justice theory, legal authorities build legitimacy by acting according to principles of procedural fairness (Tyler and Huo, 2002; Sunshine and Tyler, 2003; Tyler, 2006a,b; Hough et al., 2013a,b). On the one hand, fair and respectful treatment and neutral and objective decision making provides the moral validity that justifies their institutional position. People's judgments about the extent to which legal authority is just, fair, and valid are based in part on the degree to which individual justice agents wield their authority in just and fair ways.

On the other hand, the exercise of authority via the application of fair process – treating people in ways that are recognized to be fair, respectful, and legal, and making fair and neutral decisions – strengthens the social bonds between individuals and authorities. Procedural justice encourages not just the belief that institutions have "a just, fair, and valid basis of legal authority" (in the words of Papachristos et al., 2012: p. 417) but also identification with the group that the authority represents (typically assumed to be the state), as well as the internalization of the belief that one should follow the rules of the group (Sunshine and Tyler, 2003; Tyler and Huo, 2002; Tyler, 2006a, 2011a).

What is procedural justice? The two key issues affecting the generation of procedural justice are fairness of decision-making processes (i.e., processes that are neutral, transparent, and allow

voice or the active participation of all the involved parities) and fairness of interpersonal treatment, sometimes referred to as interactional justice (i.e., treatment with respect and dignity). A key element of procedural justice is the implementation of voice practices. Voice means providing opportunities for individuals to participate in decision-making processes. In situations of everyday disagreements and conflicts it is important to provide opportunities for individuals to state their case before decisions concerning them are made. Informal dispute resolution mechanisms are popular in part because participating in decision making allows people to voice their own personal concerns, stating what they think the issues involved are, making suggestions for how they should be handled. Such opportunities for voice need not involve a formal or elaborate mechanism. Studies of police street stops, for example, indicate that when officers provide people an opportunity to tell their side of the story before they take action, people are much more likely to feel fairly treated (Tyler and Huo, 2002).

The element of neutrality refers to making decisions based on the consistent application of rules based on proper procedure rather than on personal opinions or prejudices. The discretionary nature of the criminal justice system provides considerable opportunity for the capricious and arbitrary exercise of power, and for authorities to act based on personal prejudice and implicit bias. By being seen as acting according to rules and by applying those rules evenly across people and time, authorities are viewed as acting fairly. Moreover, treatment with respect and dignity is consistently one of the most important issues that concern people when they are dealing with authorities. When people feel demeaned or subjected to negative stereotypes, they view themselves as diminished as people and disrespected by authorities. Conversely, when authorities acknowledge people's rights and act with courtesy, they tend to feel fairly treated regardless of outcome. People are influenced by their inferences about the motivations of the authorities with whom they are dealing. If they feel that authorities are acting out of a sincere desire to do what is right, then they view the authorities as acting more fairly. Authorities communicate trust by giving people a chance to explain their concerns, showing that what they say is being considered, and explaining why and how decisions are made.

Future Directions of Research

The field of legitimacy and procedural justice has developed important insights about compliance with legal authority. But research continues, and below we outline what we believe to be four promising directions for future empirical inquiry.

Boundary Conditions for Procedural Justice and Legitimacy

As others have noted (e.g., Tankebe, 2009; Murphy and Cherney, 2012; Bradford et al., 2013; Sargeant et al., 2013), we need a better understanding of the boundary conditions of procedural justice and legitimacy in the context of criminal justice. Are there some contexts under which procedural justice and legitimacy become more or less important in compliance with criminal laws? (Note that there are parallels here to ongoing criminological work into the conditions under which deterrence and credible threat can be influential and the sorts of people who are motivated more by rational chance than by normative factors, see Picquero et al., 2011.)

Might legitimacy be more of a motivational force for deterring some types of offending than other types of offending? Legitimacy may be a more important influence on people's behavior with respect to the sort of 'everyday crimes' that straddle the line between more serious crimes and those that according to Karstedt and Farrall (2006: p. 1011) "… fall into a grey zone of legality and morality." Yet, a recent study linked perceived police legitimacy to certain types of self-reported violent behavior, i.e., carrying a gun and getting into a fight (Papachristos et al., 2012). More work is clearly needed on this issue.

Procedural justice and legitimacy may also be more important for some individuals. One possible moderating factor is one's identification with the superordinate group that the authority is representing (Huo et al., 1996; Smith et al., 1998). The identity relevant information contained in individuals' treatment by authorities may be more or less salient or meaningful to them depending on the nature and strength of preexisting group affiliations. For instance, procedural fairness may matter less to people with uncomplicated identification with the mainstream society that the police represent, since their status is not an issue (Sunshine and Tyler, 2003). Conversely, those on the margin of the group may pay relatively more attention to fairness since inclusion and status *is* an issue to them (De Cremer, 2002). People have varying psychological needs in relation to social groups, and people who have a greater need of 'belongingness information' (De Cremer and Tyler, 2005: p. 166) may place greater emphasis on fair treatment than those with less need of such information. Procedural fairness may also matter more in situations where group status is salient to actors. Greater salience increases their receptiveness to identify relevant messages (Prooijen et al., 2002), while lower salience (or lack of shared group membership) decreases receptiveness. Thus far, the extent to which group identity moderates and/or mediates the associations between procedural justice and legitimacy in real world setting is unclear (Bradford, 2012).

Another possible boundary condition is national context. Work on legitimacy originated in the US but has since successfully migrated to the UK, Australia, Israel, and some other countries. Studies have found that procedural justice is central to legitimacy and that legitimacy is a key predictor of various law-related behaviors. Round 5 of the European Social Survey – a comparative, cross-national survey that is particularly strong in its commitment to equivalence in terms of sampling, mode of interviewing, weighting, and measurement – also shows strong and consistent associations between procedural fairness and legitimacy (Hough et al., 2013a,b).

But in Ghana and South Africa, effectiveness seems to be at least as important a predictor of legitimacy as procedural justice. In these countries the basic social utility of police may be doubted more in the US or the UK. Under such conditions people may draw more heavily on their assessments of the effectiveness of the public police when forming their legitimacy judgments. In the US, UK, and Australia the essential social utility of the police is often taken as a given; people's response to

perceived crime problems is often not to blame the public police, nor to turn to alternative providers of policing services, but to call for a greater level of intervention *from* the police (Girling et al., 2000). It may be that a baseline assumption of police efficacy in the US, UK, and Australia opens up a greater space for procedural justice judgments, while at the same time dampening down variation in legitimacy linked to effectiveness judgments. In contexts like South Africa and Ghana, this baseline sense of effectiveness is still to be established, resulting in a comparatively greater emphasis on instrumental concerns about effectiveness. The broader historical backdrop is important. In the case of Ghana, Tankebe (2009: p. 1280) refers to "… the failure of successive postcolonial governments to embark on any genuine and meaningful democratic reforms of the Ghana police to build strong attachments between the police and the citizenry."

More Experimental Work

Because a good deal of the evidence on procedural justice and legitimacy is survey based, it is limited in its ability to isolate and test causal effects. Observational data of this kind allows one to model conditional correlations in the broader population, but the design has three main weaknesses: (1) there may be any number of confounding variables; (2) the arrow of causality can go both ways; and (3) the time ordering may be unclear.

There is a pressing need for more randomized controlled trials (RCTs) in the vein of Mazerolle et al. (2013). To estimate the causal effect of procedural justice, Mazerolle and colleagues randomized police officers conducting alcohol breath tests in Queensland, Australia into two test conditions. In the control group it was 'business as usual.' In the experimental group police officers followed a script based on principles of procedural justice. Compared to the control group, citizens in the experimental group were more satisfied with how they were treated, were more likely to believe that the police exercise legitimate authority, and reported being more likely to comply in the future.

While RCTs can estimate the causal impact of procedural justice on legitimacy and other accounts, field experiments are often limited in their ability to assess *exactly why* the treatment affects the outcome. There is thus a need for laboratory-based experiments to uncover the psychological mechanisms that link procedural justice, legitimacy, and law-related behavior. Traditionally, legitimacy has been connected to internalization, in which one willingly consents to the demands of legal authorities to obey the law because one accepts their right to dictate appropriate behavior and feels a corresponding duty to obey (Tyler, 2006a,b). As an influence on criminal behavior, internalization differs from compliance because of its normative rather that instrumental character.

But legitimacy may also influence law-related behavior through identification (especially if one includes 'normative alignment' in the definition of legitimacy, see Jackson et al., 2012a,b; Tyler and Jackson, 2014). Legal authorities make demands on citizens, most forcefully to obey the law, and according to Kelman's (2006) typology of social influence, legal authorities can draw upon people's motivations to avoid punishment, to internalize the moral value that it is right to obey the law (and to internalize other values, like it is wrong to steal), and to identify with certain societal roles and responsibilities. Identification refers to the idea that people are motivated to act in ways that satisfy a particular relationship because they draw value, worth, and status from that relationship. Conforming to the expectations of a social role shapes behavior because people want "to establish and maintain a satisfying self-defining relationship to another person or a group" (Kelman, 1958: p. 53). Conformity to the norms and values attached to the reciprocal–role relationship gives satisfaction not only because it agrees with the norms and values (one internalizes the values and act in ways that are intrinsically rewarding), but also because one gains value and worth from the self-defining relationship.

We recommend future lab-based studies to examine whether internalization and identification are two (mutually reinforcing) mechanisms by which procedural justice and legitimacy influence law-related behavior. How legal authorities treat individuals and make decisions conveys value, worth, and status to individuals within the wider group that the legal authorities represent (Tyler and Blader, 2003; Blader and Tyler, 2009). On this account, procedural justice leads not just to internalization of values (one of which is that it is right and proper to obey the law) but also to the merging of one's self-concept with the group and the adoption of reciprocal–role relationships. Drawing upon panel data from a national probability sample of Australians, for instance, Bradford et al. (2014a) found that social identity – identifying themselves as Australian and adopting one proper role of an Australian citizen, that is to be law abiding – mediated the association between procedural justice and perceptions of legitimacy. Procedural justice may encourage people to feel part of the system, to be normatively aligned with legal authorities, to be 'in it together,' and by extension, perhaps, to act accordingly. We need more work on identification as a mechanism of social influence linking procedural justice to law-related behavior.

Another possible mechanism links the effect of procedural justice to perceptions and experiences of power. Procedural justice operates in environments that are hierarchical by nature, and while the existence of hierarchy may be beneficial to authorities or social institutions, it is far riskier to subordinates. Consenting to authorities' power means that subordinates accept having less power, less control, and less choice over their circumstances. A new line of experimental work suggests that procedural justice can remove some of the threats inherent to having less power by changing subordinates' subjective perception of power in two important ways (Mentovich, 2012). First, procedural justice seems to *empower* subordinates – when treated fairly by an authority figure, subordinates experience themselves as having more autonomy and more power compared with when they are treated unfairly. Second, procedural justice seems to *equalize* power relations – the use of procedural justice causes the hierarchical power relations between subordinates and authorities to be seen more similarly to relations between equals. These effects have also been demonstrated in how citizens view their power in law enforcement settings and with regard to legal authorities (Mentovich, 2012). In describing their encounter with a police officer, for instance, citizens reported feeling more autonomous, more powerful, and more equal to the officer who treated them fairly compared with unfairly.

More experimental work in this vein may help our understanding of the role of procedural justice in fostering legitimacy and cooperation. Might power distance operate alongside internalization and identification to drive law-related behavior? One question is whether a sense of subjective empowerment is particularly conducive to public cooperation that, unlike obedience, is based on a sense of volition, choice, and the belief that one's action will make an impact. While disempowered subordinates are more likely to obey, those who are empowered are more likely to cooperate. And since empowerment is a positive psychological experience in and of itself, subjectively empowered citizens may continue to cooperate with the police simply since they view this experience as altogether positive. Conversely, while some degree of power inequality may be necessary in society, this can be difficult for the less powerful to accept, and it is easier for those who are in the lower side of the power ladder to accept arrangements that minimize power disparities to as large an extent as possible. We thus recommend experimental studies examine whether citizens – particularly those in democratic societies that are accustomed to relatively lower levels of power distance in relations to civic authorities – are more likely to legitimize systems and authorities that maintain a relatively minimal degree of power differentials over them.

Legitimacy as Dynamic and Interactive

Third, a recent theoretical paper expands upon the dynamic and interactive nature of legitimacy comprising claims to legitimacy from power-holders and audience reception of those claims (Bottoms and Tankebe, 2012). It certainly makes sense to study legitimacy over time as a dynamic interaction among power holders and subordinates. For example, one common comment about street stops is that a person who has been stopped repeatedly views an interaction differently from someone who has not. It may well be that, over time, the dynamic of such stops changes for both parties, in part based on the officer's changing sense of authority and power, and in part based on the citizen's reception of the officer's claims to power and authority. Understanding this dynamic requires longitudinal approaches of the type proposed by Bottoms and Tankebe. Longitudinal work would capture the citizen's sense of duty to authority, judgment of the morality of police action, and perceptions of the lawfulness of those actions. The same studies would also capture the officers' sense of power and authority, the moral values expressed in their actions, and the lawfulness of what they do and how they behave (cf Tyler et al., 2007; Bottoms and Tankebe, 2013; Jonathan-Zamir and Harpaz, 2014). Work is especially needed on the measurement of power-holders' beliefs about their own legitimacy and empirical insight into how and why these beliefs shape their subsequent behavior (see Bradford et al., 2014b, for a discussion of what might shape police officers attitudes toward different policing styles).

Integrating Procedural Justice Theory with Criminological Accounts of Compliance

Finally, drawing links to existing criminological frameworks into offending may advance our understanding of compliance and legal authority. Consider situational action theory (SAT). Locating individuals in the potentially criminogenic contexts within which they act, SAT helps us flesh out the way in which individual decisions about crime may be both structured and influenced by the actions of legal authorities (as well as many other factors). SAT sees intentional criminal acts as a subset of a wider universe of moral rule-breaking acts: "an act of crime is to intentionally break a prescription for behaviour stated in the law" (Wikström, 2006: p. 2). And importantly, for a person to commit a crime in a given situation, they must first perceive the criminal act to be an option; they then make a judgment about that option; and finally they must choose to act based on that judgment. Whether an actor sees a criminal act as an 'action alternative' is key. If – and only if – they perceive the possibility of crime in a given situation, then they must make a judgment is made based on this perception.

It is reasonable to speculate that people who perceive the police and other legal authorities to be legitimate are less likely to see crime as an action alternative. If they internalize the value that it is morally just to obey the law, an illegal act may not even cross their minds as an option, so deterrence may not even be a factor driving the action alternatives. Bringing insights from SAT into procedural justice theory may thus shed light on some quite complex relationships among legitimacy, morality, deterrence, and criminal behavior.

Another criminological account of compliance and offending centers upon the idea of collective efficacy. Sampson and colleagues (Sampson et al., 1997, 1999, 2002) have shown that neighborhoods characterized by deprived social and economic conditions are limited in their ability to control or supervise behavior (primarily that of young people), and that the statistical effects of concentrated disadvantage, residential stability, and population heterogeneity on violence (and perceived violence) are partly mediated by 'collective efficacy,' i.e., shared values and shared propensities for action. First, particular patterns of social and economic life shape the extent to which neighborhoods develop shared dispositions to feel and act, fostering social control, and cohesion. Second, significant variation in these shared propensities to act on behalf of the collective good, which is then related to levels of violence, as well as people's perceptions of violence in their locality.

A recent London-based study found that collective efficacy in a given neighborhood was strongly linked to residents' beliefs about police legitimacy – the police have the right to exercise legitimate authority in the eyes of local residents partly when the object and purpose of their power (to effect social order and control) is achieved in the everyday (Jackson et al., 2012b). Because of the 'fit' of the police to the activity of policing, the police organization garners legitimacy from the extent to which the establishment and reproduction of normative social order is strong (most palpable in the informal social control mechanisms that regulate most conduct, rather than the formal policing that steps in when informal controls have failed). People feel more obligated to obey officers when the local community seems well policed, and are more likely to feel that the police share their values when the neighborhood seems orderly and well regulated. Conversely, they question police power and authority – and begin to doubt the desirability of conferring police power and authority – when they perceive quotidian processes of social ordering to be failing.

Future work might examine whether people recognize and justify police power not only when the police wield this power in a fair way (i.e., procedural justice) but also when social order in their local neighborhood seems to be adequately maintained – that is, when the broader activity of 'policing' appears successful. It may be that collective efficacy exerts a downward pressure on crime not only through a direct influence of informal social control mechanisms, but also because it encourages people to legitimize the police (they seem to justify their power because policing is strong in their local neighborhood, cf Jackson and Sunshine, 2007; Jackson and Bradford, 2009). Legitimacy may then have a knock-on effect on compliance.

Final Words

In this article we have considered the research evidence for how legal authority can secure public compliance with the criminal law. Procedural justice theory is to date the best-evidenced theory explaining what generates legitimacy in criminal justice settings, and how legitimacy in turn influences law-related behavior. The lesson for policy is clear. A crime-control model driven by procedural justice in its structure, legislation, and day-to-day policing practice addresses dilemmas that are created by the tensions between crime-control models of the regulation of criminal behavior and due process ones, offering a resolution which privileges professional standards and the consolidation of institutional legitimacy over short-term goals of crime control.

But research continues and in this article we have recommended a broader integration with existing work on compliance and offending, more work on the boundary conditions of procedural justice and legitimacy, more experimental research into effects and mechanisms, and a new focus on the dialogues between power-holders and subordinates. These are exciting areas of research, with much to be done, and many important lines of inquiry to pursue.

See also: Collective Efficacy and Crime; Deterrence Theory: Crime; Deterrence: Sanction Perceptions; Deterrent Effect of Police and Prisons; Police Use of Force; Policing; Power; Trust, Political.

Bibliography

Blader, S., Tyler, T.R., 2009. Testing and expanding the group engagement model. Journal of Applied Psychology 94, 445–464.

Blumstein, A., Cohen, J., Nagin, N. (Eds.), 1978. Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates. National Academy of Sciences, Washington, DC.

Bottoms, A., 2002. Compliance and community penalties. In: Bottoms, A., Gelsthorpe, L., Rex, S. (Eds.), Community Penalties: Change and Challenges. Willan, Cullompton, pp. 87–116.

Bottoms, A., Tankebe, J., 2012. Beyond procedural justice: a dialogic approach to legitimacy in criminal justice. Journal of Criminal Law and Criminology 102, 119–170.

Bottoms, A., Tankebe, J., 2013. 'Voice within': power-holders' perspectives on authority and legitimacy. In: Tankebe, J., Liebling, A. (Eds.), Legitimacy and Criminal Justice: An International Exploration. Oxford University Press, Oxford.

Bradford, B., 2012. Policing and social identity: procedural justice, inclusion and cooperation between police and public. Policing and Society. http://dx.doi.org/10.1080/10439463.2012.724068 iFirst.

Bradford, B., Huq, A., Jackson, J., Roberts, B., 2013. What price fairness when security is at stake? Police legitimacy in South Africa. Regulation and Governance. http://dx.doi.org/10.1111/rego.12012.

Bradford, B., Murphy, K., Jackson, J., 2014a. Officers as mirrors: policing, procedural justice and the (re)production of social identity. British Journal of Criminology, in press.

Bradford, B., Quinton, P., Myhill, A., Porter, G., 2014b. Why do the law comply? Procedural justice, group identification and officer motivation in police organizations. European Journal of Criminology 11 (1), 110–131.

De Cremer, D., 2002. Respect and cooperation in social dilemmas: the importance of feeling included. Personality & Social Psychology Bulletin 28 (10), 1335–1341.

De Cremer, D., Tyler, T.R., 2005. Am I respected or not? inclusion and reputation as issues in group membership. Social Justice Research 18 (2), 121–153.

Girling, E., Loader, I., Sparks, R., 2000. Crime and Social Control in Middle England: Questions of Order in an English Town. Routledge, London.

Goudriaan, Heike, Wittebrood, Karin, Nieuwbeerta, Paul, 2006. Neighbourhood characteristics and reporting crime. British Journal of Criminology 46, 719–742.

Hough, M., Jackson, J., Bradford, B., 2013a. The governance of criminal justice, legitimacy and trust. In: Body-Gendrot, S., Lévy, R., Hough, M., Snacken, S., Kerezsi, K. (Eds.), The Routledge Handbook of European Criminology. Routledge, Oxon, pp. 243–265.

Hough, M., Jackson, J., Bradford, B., 2013b. Legitimacy, trust and compliance: an empirical test of procedural justice theory using the European social survey. In: Tankebe, J., Liebling, A. (Eds.), Legitimacy and Criminal Justice: An International Exploration. Oxford University Press, Oxford, pp. 326–352.

Huo, Y.J., Smith, H.J., Tyler, T.R., Lind, E.A., 1996. Superordinate identification, subgroup identification, and justice concerns: is separatism the problem, is assimilation the answer? Psychological Science 7, 40–45.

Jackson, J., Bradford, B., 2009. Crime, policing and social order: on the expressive nature of public confidence in policing. British Journal of Sociology 60 (3), 493–521.

Jackson, J., Bradford, B., Hough, M., Kuha, J., Stares, S.R., Widdop, S., Fitzgerald, R., Yordanova, M., Galev, T., 2011. Developing European indicators of trust in justice. European Journal of Criminology 8 (4), 267–285.

Jackson, J., Bradford, B., Hough, M., Myhill, A., Quinton, P., Tyler, T.R., 2012a. Why do people comply with the law? Legitimacy and the influence of legal institutions. British Journal of Criminology 52, 1051–1071.

Jackson, J., Bradford, B., Stanko, E.A., Hohl, K., 2012b. Just Authority? Trust in the Police in England and Wales. Routledge, Oxon.

Jackson, J., Huq, A., Bradford, B., Tyler, T.R., 2013. Monopolizing force? Police legitimacy and public attitudes towards the acceptability of violence. Psychology, Public Policy and Law 19 (4), 479–497.

Jackson, J., Sunshine, J., 2007. Public confidence in policing: a neo-Durkheimian perspective. British Journal of Criminology 47 (2), 214–233.

Jonathan-Zamir, T., Harpaz, A., 27 January 2014. Police understanding of the foundations of their legitimacy in the eyes of the public: the case of commanding officers in the Israel national police. British Journal of Criminology. http://dx.doi.org/10.1093/bjc/azu001. First published online.

Karstedt, S., Farrall, S., 2006. The moral economy of everyday crime: markets, consumers and citizens. British Journal of Criminology 46, 1011–1103.

Kelman, H.C., 1958. Compliance, identification, and internalization: three processes of attitude change. Journal of Conflict Resolution 2, 51–60.

Kelman, H.C., 2006. Interests, relationships, identities: three central issues for individuals and groups in negotiating their social environment. In: Fiske, S.T., Kazdin, A.E., Schacter, D.L. (Eds.), Annual Review of Psychology, vol. 57. Annual Reviews, Palo Alto, CA, pp. 1–26. http://scholar.harvard.edu/hckelman/publications/interests-relationships-identities-three-central-issues-individuals-and-groups.

Kelman, H.C., Hamilton, V.L., 1989. Crimes of Obedience. Yale, New Haven, CT.

Loader, I., Mulcahy, A., 2003. Policing and the Condition of England: Memory, Politics and Culture. Oxford University Press, Oxford.

Mazerolle, L., Antrobus, E., Bennett, S., Tyler, T.R., 2013. Shaping citizen perceptions of police legitimacy: a randomized field trial of procedural justice. Criminology 51 (1), 33–63. http://dx.doi.org/10.1111/j.1745-9125.2012.00289.x.

Mentovich, A., 2012. The Power of Fair Procedures: The Effect of Procedural Justice on Perceptions of Power and Hierarchy. Doctoral thesis. New York University.

Murphy, K., Cherney, A., 2012. Understanding cooperation with police in a diverse society. British Journal of Criminology 52, 181–201.

Nagin, D.S., 1998. Criminal deterrence research at the outset of the twenty-first century. Crime and Justice: A Review of Research 23, 1–42.

Pagliaro, S., Ellemers, N., Barreto, M., 2011. Sharing moral values: anticipated ingroup respect as a determinant of adherence to morality-based (but not competence-based) group norms. Personality & Social Psychology Bulletin 37 (8), 1117–1129.

Papachristos, A.V., Meares, T.L., Fagan, J., 2012. Why do criminals obey the law? the influence of legitimacy and social networks on active gun offenders. Journal of Criminal Law & Criminology 102 (2), 397–440.

Paternoster, R., 2006. How much do we really know about criminal deterrence? Journal of Criminal Law and Criminology 100, 765–824.

Picquero, A., Paternoster, R., Pogarsky, G., Loughran, T., 2011. Elaborating the individual difference component in deterrence theory. Annual Review of Law and Social Science 7, 335–360.

Prooijen, J.W. van, Bos, K van den, Wilke, H.A.M., 2002. Procedural justice and status: status salience as antecedent of procedural fairness effects. Journal of Personality and Social Psychology 83, 1353–1361.

Robinson, G., McNeill, F., 2008. Exploring the dynamics of compliance with community penalties. Theoretical Criminology 12, 431–449.

Sampson, R.J., Morenoff, J.D., Gannon-Rowley, T., 2002. Assessing neighbourhood effects: social processes and new directions in research. Annual Review of Sociology 28, 443–478.

Sampson, R.J., Morenoff, J., Earls, F., 1999. Beyond social capital: spatial dynamics of collective efficacy for children. American Sociological Review 64, 633–660.

Sampson, R.J., Raudenbush, S.W., Earls, F., 1997. Neighborhoods and violent crime: a multilevel study of collective efficacy. Science 277, 918–924.

Sargeant, E., Murphy, T., Chemey, A., 9 December 2013. Ethnicity, trust and cooperation with police: testing the dominance of the process-based mode. European Journal of Criminology. http://dx.doi.org/10.1177/1477370813511386. Published online before print.

Schulhofer, S., Tyler, T.R., Huq, A., 2011. American policing at a crossroads: unsustainable policies and the procedural justice alternative. Journal of Criminal Law and Criminology 101 (2), 335–375.

Smith, H.J., Tyler, T.R., 1997. Choosing the right pond: the influence of the status of one's group and one's status in that group on self-esteem and group-oriented behaviors. Journal of Experimental Social Psychology 33, 146–170.

Smith, H.J., Tyler, T.R., Huo, Y.J., Ortiz, D.J., Lind, E.A., 1998. The self-relevant implications of the group-value model: group membership, self-worth, and procedural justice. Journal of Experimental Social Psychology 34, 470–493.

Sunshine, J., Tyler, T.R., 2003. The role of procedural justice and legitimacy in shaping public support for policing. Law and Society Review 37, 513–548.

Tajfel, H., Turner, J.C., 1986. The social identity theory of intergroup relations. In: Worchel, S., Austin, W.G. (Eds.), Psychology of Intergroup Relations. Nelson-Hall, Chicago, IL, pp. 7–24.

Tankebe, J., 2009. Public cooperation with the police in Ghana: does procedural fairness matter? Criminology 47 (4), 1265–1293.

Turner, J.C., 1975. Social comparison and social identity: some prospects for intergroup behavior. European Journal of Social Psychology 5, 5–34.

Turner, J.C., Brown, R.J., Tajfel, H., 1979. Social comparison and group interest in ingroup favoritism. European Journal of Social Psychology 9, 187–204.

Tyler, T.R., Blader, S., 2003. Procedural justice, social identity, and cooperative behavior. Personality and Social Psychology Review 7, 349–361.

Tyler, T.R., 2003. Procedural justice, legitimacy, and the effective rule of law. In: Tonry, M. (Ed.), Crime and Justice 30, 431–505.

Tyler, T.R., 2006a. Psychological perspectives on legitimacy and legitimation. Annual Review of Psychology 57, 375–400.

Tyler, T.R., 2006b. Why People Obey the Law. Princeton University Press, Princeton, NJ.

Tyler, T.R., 2011a. Why People Cooperate. Princeton University Press, Princeton, NJ.

Tyler, T.R., 2011b. Trust and legitimacy: policing in the US and Europe. European Journal of Criminology 8 (4), 254–266.

Tyler, T.R., Boeckmann, R., 1997. Three strikes and you are out, but why? The psychology of public support for punishing rule breakers. Law and Society Review 31 (2), 237–265.

Tyler, T.R., Lind, E.A., 1992. A relational model of authority in groups. In: Zanna, Mark P. (Ed.), Advances in Experimental Social Psychology, vol. 25. Academic Press, Inc., San Diego, CA, pp. 115–191.

Tyler, T.R., Fagan, J., 2008. Legitimacy and cooperation: why do people help the police fight crime in their communities? Ohio State Journal of Criminal Law 6 (1,), 231–275.

Tyler, T.R., Huo, Y.J., 2002. Trust in the Law: Encouraging Public Cooperation with the Police and Courts. Russell-Sage Foundation, New York.

Tyler, T.R., Jackson, J., 2013. Future challenges in the study of legitimacy and criminal justice. In: Tankebe, J., Liebling, A. (Eds.), Legitimacy and Criminal Justice: An International Exploration. Oxford University Press, Oxford, pp. 83–104.

Tyler, T.R., Jackson, J., 2014. Popular legitimacy and the exercise of legal authority: motivating compliance, cooperation and engagement. Psychology, Public Policy and Law 20 (1), 78–95. http://dx.doi.org/10.1037/a0034514.

Tyler, T.R., Callahan, P., Frost, J., 2007. Armed, and dangerous(?): can self-regulatory approaches shape rule adherence among agents of social control. Law and Society Review 41 (2), 457–492.

Wikström, P.-H.O., 2006. Individuals, settings and acts of crime: situational mechanisms and the explanation of crime. In: Wikström, P.-H.O., Sampson, R.J. (Eds.), The Explanation of Crime. Cambridge University Press, Cambridge, pp. 31–60.

Wikström, P.-H.O., 2010. Situational action theory. In: Cullen, F.T., Wilcox, P. (Eds.), Encyclopedia of Criminology, vol. 1. Sage, Thousand Oaks, CA, pp. 1001–1007.

Wikström, P.-H.O., Oberwittler, D., Treiber, K., Hardie, B., 2012. Breaking Rules: The Social and Situational Dynamics of Young People's Urban Crime. Oxford University Press, Oxford.

Compression of Morbidity

Siegfried Geyer, Hannover Medical School, Hannover, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.F. Fries, volume 4, pp. 2449–2453, © 2001, Elsevier Ltd.

Abstract

The morbidity compression hypothesis as formulated by James Fries asserts that the time spent in states of chronic illness and disability prior to death will be reduced. It is driven by prevention, improved living conditions, and healthier lifestyles. It will take place within the context of increased life expectancy, but it may also occur in case of life expectancy remaining unchanged or being reduced. Studies on general health and cognitive impairment are pointing toward shorter periods of illness and disability. For more specific diseases, such as myocardial infarction, cancers, and diabetes, the available evidence is still insufficient. It remains to be determined whether morbidity compression occurs in the whole population or whether some subgroups are more affected than others. The degree of morbidity compression will determine the future development of health care and welfare systems, thus it is a topic for basic research as well as for public health practice.

Introduction

Morbidity compression is a key issue of active aging in Western countries and the US. The future development of morbidity and longevity will determine the allocation of health care and the way how the elderly will be cared for in the coming decades. This article gives an account of the current state of knowledge on morbidity compression. Morbidity expansion and the dynamic equilibrium, as alternative formulations, will be introduced, but they are not in the focus of this article, as each of them requires a separate discussion.

The different ways how health outcomes are being measured will be introduced. It will be shown that morbidity compression is a rather complex phenomenon, and that the available measures are depicting it in a rather incomplete way. The evidence on morbidity compression and its explanations will be presented by considering mental, physical, and general health separately. The relationships with social structural factors will be considered in a separate paragraph, and the contribution will conclude with a general evaluation of the present state of knowledge and with an outline of future directions of research.

The Morbidity Compression Hypothesis

Compression of morbidity was introduced in 1980 (Fries, 1980). James Fries assumed that prevention, changes of living conditions, and socioeconomic factors are contributing to shifting the first onset of disease, aging, and disability to the later years of life. The periods of time spent in states of disability and impaired health will become shorter. He did not only refer to onsets occurring in the ripe old age, but also to diseases starting earlier in life, e.g., myocardial infarction or states of minimal morbidity beginning around the age of 50 years (Fries et al., 2011). The hypothesis was formulated against the backdrop of the assumption of a maximum biological life span (Fries, 2012). Fries assumed that morbidity compression may be one option to improve health in spite of having reached the upper end of life expectancy. Whether a maximum biological age does exist is a matter of debate, and there is evidence that such an assumption is difficult to prove. Instead, increasing life expectancy may be explained by better health status in old age (Vaupel, 2010). However, for morbidity compression this is not necessarily of central importance, as healthy life expectancy may also be extended without total life expectancy to increase. It may also occur if human life span remains constant, and also in case of falling life expectancy (Fries et al., 2011). Morbidity compression has become an issue because in most Western countries life expectancy is increasing by 3 months every year (Doblhammer et al., 2012). The segments of the population where life expectancy has increased were changing as dependent on the time periods considered. From the early twentieth century onward, the reduction of neonatal and infant mortality was the main driving force, after 1960 it was the life expectancy above the age of 60 years. In the last decades this development continued into old age (Crimmins and Beltran-Sanchez, 2011), and a substantial proportion of the centenarians are surviving in good health (Andersen et al., 2012). The health of the elderly is of vital societal importance as it determines the future development of the welfare and the health care systems, the associated costs, and the decision about whether the age of retirement can be postponed beyond the age of 65 years.

The driving forces of morbidity compression are located in the areas of prevention that can contribute to the reduction of lifetime morbidity in different ways:

Primordial prevention refers, e.g., to the risk factors leading to smoking. The target groups of interventions are mainly adolescents where different behavioral (Thrul et al., 2013) as well as structural measures are most likely to be effective. This may entail a changing social connotation of smoking as a habit, to strengthen self-confidence and resistance against nicotine-promoting advertising. In the last decades, in Germany for instance, these efforts had been successful as the proportion of smokers in the younger age-groups has decreased, although the proportions are differing by social group (Schulze and Lampert, 2006). *Primary prevention* is directed toward reducing disease risks by quitting smoking, doing exercise, reducing stress levels at work, or changing work environments in ways that the risk of specific diseases is reduced. *Secondary*

prevention is directed toward the early detection of diseases with a popular example being breast cancer screening. It does not prevent disease onsets, but they will be detected in earlier stages. As a result, the burden of disease will be reduced or complete cure may become more likely. Finally, *tertiary prevention* is directed toward reducing risks of deterioration in case of an existing disease. This refers to lifestyle changes in patients, e.g., with myocardial infarction, in order to prevent the recurrence of a second infarction or related conditions and to support the primary therapy.

Besides these rather patient-centered measures, macrolevel and structural factors are important for improving health and life expectancy. The level of gross national income and the economic development have turned out as important determinants of life expectancy and risks of specific diseases like stroke or myocardial infarction (Mackenbach and Looman, 2013; Sposato and Saposnik, 2012). These macrolevel influences are representing the impact of so-called secondary factors such as the average educational level of a country, standard of living, or investments into infrastructure (e.g., sanitation, quality of health care, and public health efforts) or prosperity (Vaupel, 2010).

Alternative Directions of Morbidity Development

Morbidity compression posits an optimistic view of the direction in which health will be developing to. In case of its confirmation, population health will improve, the growth of health care costs will be smaller than expected, and the planning of infrastructure for frail people will have to be reconsidered. There are, however, two alternative assumptions about the development of population-based morbidity. In 1977, Ernest Gruenberg (Gruenberg, 1977) formulated a more pessimistic view that became to be known as 'expansion of morbidity.' He assumed that the development of medical treatments and drugs, e.g., antibiotics or techniques, for curing potentially lethal infections has positive outcomes, but there are also unintended effects of medical and technical progress. As a result, the number of chronically ill patients will increase, and cure will not be possible, instead they will continue to suffer from impaired health. In effect, life expectancy will improve, but it will take its toll on healthy life years, thus leading to increasing costs of health care and prolonged disability. As an example, Gruenberg described the case of patients with atherosclerosis who some decades ago died of associated complications such as pneumonia, coronary occlusion, and ruptured arteries. Medicine was successful in preventing them, but the consequence was an increase of the number of chronically ill patients. Another example is severe congenital malformations of the heart. In the last decades, the proportion of patients surviving after surgical treatment has increased (Moons et al., 2010), but the largest share is due to patients with complex malformations that can only be treated by palliative surgery. The risks of early cardiac insufficiency are increasing with disease severity, resulting in manifestations in the early years of life (Norozi et al., 2006). In spite of the operations of congenital heart disease being successful and with patients' life expectancy increasing, this sort of progress is creating a growing number of patients who are in constant need of medical supervision and treatment (Moons et al., 2010).

A position between these two scenarios was formulated by Manton (Manton, 1982). It became to be known as 'dynamic equilibrium,' as it combines elements of the compression and the expansion hypothesis. According to Manton, reduced mortality rates are the result of improved medical care and more advanced treatment methods that do not lead to reduced onset rates, but to disease and disability states of lower severity, thus leading to improved quality of life and improved abilities to master everyday life (Manton, 2008). Apart from medical care these developments may at least in part be explained by improvements of nutrition, sanitation, and education. Manton found evidence for a dynamic equilibrium when studying certain developments in the United States, e.g., for cardiovascular diseases, arthritis, and obesity (Manton, 2008). In a recent study from New Zealand, functional limitations and healthy life expectancy were examined longitudinally. It was concluded that the development in this country may rather be described as a dynamic equilibrium (Graham et al., 2004).

The Assessment of Morbidity Compression

Studies on morbidity compression are not based on a consistent definition of the term 'morbidity.' Only a small number refers to manifest diseases that can be diagnosed on the basis of well-defined organic manifestations such as acute myocardial infarction, stroke, or diabetes mellitus. A broad range of measures of health impairments can be used as further end points: The activities of daily living (ADL) (Katz et al., 1970) or instrumental activities of daily living (IADL) (Lawton and Brody, 1969) are respondent-based measures for assessing impairments of everyday functioning. They are covering a wide variety of indicators pointing to physical health. For dementia or cognitive impairment a multitude of measurement approaches is used, ranging from recall tests (Langa et al., 2008) to screening instruments such as the Mini-Mental State Examination (Folstein et al., 1975). Subjective health is often used in large population-based surveys. It consists of only one item asking respondents to rate their present state of health, and responses are depicted on a five-point scale (DeBruin et al., 1996). This item turned out to be correlated with a large variety of health-related outcomes including the prediction of mortality. More complex and more specific are symptom lists such as the Short Form-36 (SF-36) (Jenkins and Wright, 1993) or the Symptom Checklist-90R (SCL-90R) (Derogatis, 1992). They were developed to cover a broad range of complaints and symptoms.

A final group of indicators refers to life expectancy as summary measures of population health. They can be used to evaluate the sum of all developments occurring in defined time periods with respect to their effects on morbidity and mortality. Healthy life expectancy and healthy life years are expressing the number of years that a person at a given age may expect to live in good health. These measures take age-specific mortality, morbidity, and functional health status into account (Salomon et al., 2012).

Evidence from Studies

Fries was not focusing on particular diseases as his thinking was primarily directed toward morbidity compression within the context of disability, frailty, or general health (Crimmins and Beltran-Sanchez, 2011), concepts that may rather be subsumed under the heading of multimorbidity. The compression thesis, in connection with the effects of lifestyle patterns, was examined by surveying a sample of female and male alumni with a mean age of 68 years who had completed their school education between 1939 and 1944 (Chakravarty et al., 2012). Data collection took place between 1986 and 1998. The study examined whether lifestyle factors, in particular smoking, body weight, and physical exercise, were associated with functional status and age of death. An index was created by counting the number of risk factors, resulting in the definition of three risk groups. It turned out that the average age of death was the same in all three groups (M = 76 years), without pronounced gender differences. The relative risk of death in respondents with high risks as compared to the participants with low risks was 1.9 times higher, and if lifestyle factors were considered, the difference between the two extreme groups amounted to a relative risk of 2.8. A detailed examination revealed that in the group with the healthiest lifestyle patterns the overall disability rate was the lowest, and changes toward a deterioration of health status and disabilities were closer to the date of death than was the case in the comparison groups. The authors interpreted these findings as a confirmation of the morbidity compression thesis. They concluded that preventive efforts toward changing modifiable risk factors will reduce the burden of disease and the need for long-term care before death. In a follow-up assessment the observation period of this study population was extended beyond the age of 90 years with a mean age of M = 85 years (Chakravarty et al., 2012). The risk factors continued to predict progression of disability and changes of mortality. In the most favorable as compared to the most disadvantaged risk group, mortality was postponed by 3.6–3.9 years, and disability was postponed by 8.2 years. A preventive orientation in life has more effects on illness and disability than on the risk of death, and these findings are in line with the morbidity compression hypothesis. The study has nevertheless some limitations as the responses were based on self-reports and had not been cross-validated. Problems of selectivity should have been addressed in more detail, as individuals with an unfavorable health status are more likely to refrain from participating in surveys (Heilbrun et al., 1991; Panser et al., 1994). In addition, the sample was recruited from a particular population as their educational level was high, and it remained unclear whether the effects of health-related lifestyles vary by factors that had not been taken into account.

Fries and colleagues conducted a second study that covered an observation period of 21 years (Chakravarty et al., 2008). They collected data from female and male members of running clubs and healthy controls. The data were collected by means of self-administered questionnaires, and in addition data on health-related lifestyles, exercise habits, as well as illness and disabilities were assessed. Dependent variables were illness, disability, and death. Causes of death were drawn from the National Death Index.

After 21 years, disability levels in controls were higher than in the runners group, although the differences were not very pronounced at the beginning. Over the whole study period these differentials were widening, and this refers both to subjects who completed the study as well as to those who dropped out. Over the study period it turned out that disability levels were lower in the runners group, and among those with disability the progression was slower. Significant degrees of disability rates emerged with an average time lag between the initial measurement and follow-up of 2.3 years in controls and 8.7 years in runners. In addition, mortality rates were lower in the latter group. The reduced rate of deaths was primarily due to reductions of cardiovascular mortality, but to a smaller degree also to other causes. Again, the authors are interpreting the results as a confirmation of the morbidity compression hypothesis, and they are recommending regular exercise and a healthy lifestyle.

Surveys are overestimating the health status of target populations because the willingness to participate in studies is reduced in case of impaired health, thus leading to inappropriate conclusions about population health. The use of administrative data is a possible way out, and this had been done in a study with US Medicare data (Willis et al., 2012). A sample of women and men was used for examining effects of cardiorespiratory fitness in midlife on the onset of eight chronic conditions (congestive heart failure, ischemic heart disease, stroke, diabetes mellitus, chronic obstructive pulmonary disease, chronic kidney disease, Alzheimer's disease, and cancers of the colon and the lung). Outcomes were recorded after the age of 70 years. It turned out that higher fitness in midlife was associated with a lower incidence of chronic conditions in later life. Apart from lower risks, the average age of onset was postponed to later periods of life while survival rates were not changed significantly.

Manton et al. (2008) combined data of six surveys with health care utilization data. Respondents were at least 65 years old, and the study also included the oldest individuals of 95 years and beyond. This study is particularly interesting because it covered a period between 1982 and 2004, i.e., 22 years, and in addition to questionnaires, data on physical performance (ADL and IADL) were collected in order to assess impairments of daily activities and frailty. For women as well as for men it was found that the younger cohorts experienced longer lives in a nondisabled state as compared to individuals born earlier. Cohort effects were observed that were most pronounced at the upper end of life. Again, authors were interpreting these findings as support of the morbidity compression hypothesis.

The evidence on morbidity compression with respect to manifest diseases is scarce. Age-standardized incidences of cardiovascular diseases were decreasing in the last years, and there was a constant trend toward a shift into higher age-groups (Doblhammer et al., 2012). The same development was reported for the US (Crimmins and Beltran-Sanchez, 2011), but for Germany, age-standardized mortality rates only are available. In the male population they were decreasing since 1981, and in the female population the same development was found since 1985. In East Germany decline started with a delay after 1989. Similar findings were reported for stroke where the survival rates were increasing continuously since the 1980s

(Doblhammer et al., 2012). There are also developments in the opposite direction. Changes in lifestyles need not necessarily go hand in hand with improvements of health. In all European countries and in the US the obesity rates have increased in the last decades, and the age of onset shifted into younger age-groups (Astrup, 2012; Seidell, 2000). High rates of obesity in the population along with onsets in lower age-groups are associated with increased risks of secondary diseases, in particular type-2 diabetes (Brinks et al., 2012). At least in this case it has to be determined to what extent decreased morbidity rates and improvements of medical treatment are compensated by the concomitants of obesity.

The studies presented so far were focusing on physical health, functional impairments, or frailty, thus pointing to the effects of health-related lifestyle factors. Another line of research has concentrated on morbidity compression in terms of subjective health and cognitive functioning. Subjective health is frequently assessed as an unspecific measure of impaired health or functioning, as an indicator of distress and adverse experiences (Jylhä, 2009). The WHO has recommended it for use in health interview surveys (DeBruin et al., 1996) at a national, as well as at an international level, for purposes of health-related comparisons between countries.

In Germany, data of the Socio-Economic Panel and of the microcensus were used as basis for examining morbidity compression. No data on manifest diseases were available, but data on days spent in hospital and subjective health were collected (Kroll and Ziese, 2009). Comparing different age cohorts, evidence was found that subjective health in women and men in the higher age-groups was improving over the years. There was also a constant trend toward chronic diseases being associated with decreasing levels of disability. Both results are in line with the thesis of morbidity compression.

Self-reported health outcomes were also examined for the Austrian population between 60 and 89 years of age. Data of four waves of the national microcensus were used covering the period between 1978 and 1998 (Doblhammer and Kytir, 2001). As no data on specific diseases as end points were available, subjective health was used. As reported for other countries, life expectancy increased over the observation period. It was concluded that the health of elderly Austrians had improved over the years while the length of time spent in bad health became shorter. The findings support the conclusion that an absolute compression of morbidity had taken place. This development accelerated in the years from 1978 to 1983 and from 1991 to 1998, while from 1983 to 1991, a transitory stagnation occurred.

Social Differences in Morbidity Compression

In the studies described so far the study populations were stratified by gender and age, and both are essential, because life expectancies are strongly dependent on them. This is however neglecting effects of social factors as it is a well-established finding that large social differences in morbidity and mortality do exist (Mackenbach, 2006; Mackenbach et al., 2008), and their relative effects are varying by the disease and by the indicator chosen (Braveman et al., 2011; Geyer et al., 2006). For the most frequently occurring diseases the highest risks are found in individuals with the lowest education, incomes, and occupational positions. These social differences are emerging in midlife and persist into old age, although the effects are then weakened. According to Fries, morbidity compression is mainly triggered by preventive behavior and favorable living conditions, and both factors are unequally distributed across society. Thus, it has to be examined whether morbidity compression follows a social gradient, being more prevalent and starting earlier in the groups that are better off, while in the others this will occur with a time lag, as demonstrated in many other domains (Rogers, 1995).

Research on morbidity compression that includes indicators of social stratification is rare. Only a few studies included them at all, but usually as control variables (Qiu et al., 2013). Langa et al. (2008) used data of nationally representative samples from the US (N = 7406 and N = 7104) 70 years of age and older. They compared cohort effects of education on cognitive impairments. It turned out that the proportion of respondents with impairments dropped from 12.2% in 1993 to 8.7% in 2002, and higher mortality rates were found in respondents with higher degrees of cognitive impairment. The general level of education was higher in the cohort born later, and higher education was associated with lower morbidity levels.

Another study with respondents over 65 years of age covered a period of 16 years (Marioni et al., 2012). It examined whether cognitive activities, education, occupation, and social engagements were related to different levels of cognitive impairment. It turned out that cognitive, social, as well as intellectually demanding activities were related to better functioning. Expressed in quantitative terms, a man of 65 years of age with a low education, a low occupational position, and no marked social activities had an estimated residual life expectancy of 12.8 years with cognitive impairment and 8.5 years without cognitive impairment. At the opposite side are individuals with a high education, a high occupational position, and with social activities. Their life expectancy was 15.9 years, and they could expect 14.3 years without cognitive impairment. The figures of the female respondents are similar, thus the associations are holding both for women and men.

Taken together, educational level and favorable material conditions were associated with better cognitive functioning, and each indicator had independent effects. In both studies the authors explained the effects of education with reference to the cognitive reserve hypothesis (Whalley et al., 2004).

Concluding Remarks

The main goal of this article was to summarize the current evidence on morbidity compression. The empirical findings were structured along the three lines: cognitive impairments, specific diseases, and general measures of health. Studies have demonstrated that the time of life spent with cognitive impairments has decreased (Chakravarty et al., 2012; Doblhammer et al., 2013), and some found that this may also apply to dementia (Qiu et al., 2013). Studies dealing with general health (Chakravarty et al., 2008; Doblhammer and Kytir, 2001; Kroll and Ziese, 2009) reached the same conclusion. For myocardial infarction and other frequently occurring diagnoses there is evidence that morbidity compression may

occur, but the number of studies is still small. It is, however, not clear what will happen if the average age at the onset of chronic conditions is increasing, i.e., these disorders may then be replaced by other degenerative diseases, such as dementia. A more complete picture will emerge if the whole range of morbidity is taken into account.

Future studies should not only consider general concepts, such as cognitive impairment, general health, or disability, instead they should be differentiated by the type of disease. This refers to the most frequently occurring ones, notably myocardial infarction, stroke, malignant diseases, and diabetes. Disease-specific studies will also enable professionals to tailor preventive measures for higher age-groups. The next line of research refers to differentiated analyses of subpopulations, as in most studies conducted so far populations were differentiated only by gender and age. In studies on cognitive functioning, the samples were differentiated by indicators of socioeconomic position (Cagney and Lauderdale, 2002; Langa et al., 2008; Marioni et al., 2012) or by cultural background (Cagney and Lauderdale, 2002). It will be of interest to know whether morbidity compression is developing in the same way over the whole of a social structure, and whether some parts of the population are profiting more or less from it. Finally, the studies discussed so far were conducted mainly in Europe and in the US. Morbidity compression may however not be a universal finding as an international comparison of 187 countries concluded that healthy life expectancy was developing into different directions. In any case research on these topics is still at the beginning.

See also: Chronic Illness, Psychosocial Coping With; Coronary Heart Disease: Psychosocial Aspects; Disability: Social and Psychological Aspects; Environmental Stress and Health; Functioning, Disability and Health, International Classification of; Health Behavior, Psychosocial Theories of; Health Care, Rationing of; Health Education and Health Promotion; Health Outcomes, Assessment of; Life Course Epidemiology; Social Epidemiology; Socioeconomic Status and Health.

Bibliography

Andersen, S.L., Sebastiani, P., Dworkis, D.A., Feldman, L., Perls, T.T., 2012. Health span approximates life span among many supercentenarians: compression of morbidity at the approximate limit of life span. The Journals of Gerontology Series a: Biological Sciences and Medical Sciences 67A (4), 395–405.

Astrup, A., 2012. Healthy lifestyles in Europe: prevention of obesity and type II diabetes by diet and physical activity. Public Health Nutrition 4, 499–515.

Braveman, P., Egerter, S., Williams, D.R., 2011. The social determinants of health: coming of age. Annual Review of Public Health 32 (1), 381–398.

Brinks, R., Tamayo, T., Kowall, B., Rathmann, W., 2012. Prevalence of type 2 diabetes in Germany in 2040: estimates from an epidemiological model. European Journal of Epidemiology 27 (10), 791–797.

Cagney, K.A., Lauderdale, D.S., 2002. Education, wealth, and cognitive function in later life. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 57 (2), 163–172.

Chakravarty, E.F., Hubert, H.B., Lingala, V.B., Fries, J.F., 2008. Reduced disability and mortality among aging runners: a 21-year longitudinal study. Archives of Internal Medicine 168 (15), 1638–1646.

Chakravarty, E.F., Hubert, H.B., Krishnan, E., Bruce, B.B., Lingala, V.B., Fries, J.F., 2012. Lifestyle risk factors predict disability and death in healthy aging adults. The American Journal of Medicine 125 (2), 190–197.

Crimmins, E.M., Beltran-Sanchez, H., 2011. Mortality and morbidity trends: is there compression of morbidity? The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 66B (1), 75–86.

DeBruin, A., Picavet, H.S.V., Nossikov, A., 1996. Health Interview Surveys. Towards International Harmonization. WHO Regional Office for Europe, Kopenhagen.

Derogatis, L.R., 1992. SCL-90-R, Administration, Scoring & Procedures Manual-II for the R(evised) Version and Other Instruments of the Psychopathology Rating Scale Series. Clinical Psychometric Research Inc., Townson.

Doblhammer, G., Kreft, D., Dethloff, A., 2012. Gewonnene Lebensjahre. Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz 55 (4), 448–458.

Doblhammer, G., van den Berg, G.J., Fritze, T., 2013. Economic conditions at the time of birth and cognitive abilities late in life: evidence from ten European countries. PLOS ONE 8 (9), e74915.

Doblhammer, G., Kytir, J., 2001. Compression or expansion of morbidity? Trends in healthy-life expectancy in the elderly Austrian population between 1978 and 1998. Social Science & Medicine 52 (3), 385–391.

Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research 12 (3), 189–198.

Fries, J.F., 2012. The theory and practice of active aging. Current Gerontology and Geriatrics Research 420637. http://dx.doi.org/10.1155/2012/420637.

Fries, J.F., Bruce, B., Chakravarty, E., 2011. Compression of morbidity 1980–2011: a focused review of paradigms and progress. Journal of Aging Research 2011. http://dx.doi.org/10.4061/2011/261702.

Fries, J.F., 1980. Aging, natural death and the compression of morbidity. The New England Journal of Medicine 303, 130–135.

Geyer, S., Hemström, Ö., Peter, R., Vågerö, D., 2006. Education, income and occupational class cannot be used interchangeably in social epidemiology. Empirical evidence against an unquestioned practice. Journal of Epidemiology and Community Health 60, 804–810.

Graham, P., Blakely, T., Davis, P., Sporle, A., Pearce, N., 2004. Compression, expansion, or dynamic equilibrium? The evolution of health expectancy in New Zealand. Journal of Epidemiology and Community Health 58 (8), 659–666.

Gruenberg, E.M., 1977. The failure of success. Milbank Memorial Fund, Health and Society 55 (1), 3–24.

Heilbrun, L.K., Nomura, A., Stemmermann, G.N., 1991. The effects of non-response in a prospective study of cancer: 15-year follow-up. International Journal of Epidemiology 20, 328–338.

Jenkins, C., Wright, L., 1993. The SF-36 health survey questionnaire. Auditorium 2, 7–12.

Jylhä, M., 2009. What is self-rated health and why does it predict mortality? Towards a unified conceptual model. Social Science & Medicine 69 (3), 307–316.

Katz, S., Cash, H.R., Downs, T.D., Grotz, R.C., 1970. Progress in development of the index of ADL. The Gerontologist 10 (1), 20–30.

Kroll, L.-E., Ziese, T., 2009. Kompression oder Exppansion der Morbidität? In: Böhm, K., Tesch-Römer, C., Ziese, T. (Eds.), Gesundheit und Krankheit im Alter. Robert Koch-Institut, Berlin.

Langa, K.M., Larson, E.B., Karlawish, J.H., Cutler, D.M., Kabeto, M.U., Kim, S.Y., Rosen, A.B., 2008. Trends in the prevalence and mortality of cognitive impairment in the United States: is there evidence of a compression of cognitive morbidity? Alzheimer's & Dementia 4 (2), 134–144.

Lawton, M.P., Brody, E.M., 1969. Assessment of older people: self-maintaining and instrumental activities of daily living. The Gerontologist 9 (3), 179–186.

Mackenbach, J.P., 2006. Health Inequalities: Europe in Profile University Medical Center Rotterdam. Erasmus University Rotterdam, Rotterdam.

Mackenbach, J.P., Looman, C.W., 2013. Life expectancy and national income in Europe, 1900–2008: an update of prestons analysis. International Journal of Epidemiology 42 (4), 1100–1110.

Mackenbach, J.P., Stirbu, I., Roskam, A.J., Schaap, M.M., Menvielle, G., Leinsalu, M., Kunst, A.E., 2008. Socioeconomic inequalities in health in 22 European countries. New England Journal of Medicine 358 (23), 2468–2481.

Manton, K.G., 1982. Changing concepts of morbidity and mortality in the elderly population. Milbank Memorial Fund, Health and Society 60 (2), 183–244.

Manton, K.G., 2008. Recent declines in chronic disability in the elderly U.S. Population: risk factors and future dynamics. Annual Review of Public Health 29 (1), 91–113.

Manton, K.G., Gu, X., Lowrimore, G.R., 2008. Cohort changes in active life expectancy in the U.S. elderly population: experience from the 1982−2004 national long-term care survey. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 63 (5), S269–S281.

Marioni, R.E., Valenzuela, M.J., van den Hout, A., Brayne, C., Matthews, F., 2012. Active cognitive lifestyle is associated with positive cognitive health transitions and compression of morbidity from age sixty-five. PLOS ONE 7 (12), e50940.

Moons, P., Bovijn, L., Budts, W., Belmans, A., Gewillig, M., 2010. Temporal trends in survival to adulthood among patients born with congenital heart disease from 1970 to 1992 in Belgium. Circulation 122 (22), 2264–2272.

Norozi, K., Wessel, A., Alpers, V., Arnhold, J.O., Geyer, S., Zoege, M., Buchhorn, R., 2006. Incidence and risk distribution of heart failure in adolescents and adults with congenital heart disease after cardiac surgery. American Journal of Cardiology 97 (15.04.2006), 1238–1243.

Panser, L.A., Chute, C.G., Guess, H.A., Larsonkeller, J.J., Girman, C.J., Oesterling, J.E., Lieber, M.M., Jacobsen, S.J., 1994. The natural history of prostatism: the effects of non-response bias. International Journal of Epidemiology 23 (6), 1198–1205.

Qiu, C., von Strauss, E., Bäckman, L., Winblad, B., Fratiglioni, L., 2013. Twenty-year changes in dementia occurrence suggest decreasing incidence in central Stockholm, Sweden. Neurology. http://dx.doi.org/10.1212/WNL.0b013e318292a2f9.

Rogers, E., 1995. Diffusion of Innovations, fourth ed. The Free Press, New York.

Salomon, J.A., Wang, H., Freeman, M.K., Vos, T., Flaxman, A.D., Lopez, A.D., Murray, C.J., 2012. Healthy life expectancy for 187 countries, 1990−2010: a systematic analysis for the global burden disease study 2010. The Lancet 380 (9859), 2144–2162.

Schulze, A., Lampert, T., 2006. Bundes-Gesundheitssurvey: Soziale Unterschiede im Rauchverhalten und in der Passivrauchbelastung in Deutschland. Robert-Koch-Institut, Berlin.

Seidell, J.C., 2000. Obesity, insulin resistance and diabetes — a worldwide epidemic. British Journal of Nutrition 83 (S1), S5–S8.

Sposato, L.A., Saposnik, G., 2012. Gross domestic product and health expenditure associated with incidence, 30-day fatality, and age at stroke onset: a systematic review. Stroke 43 (1), 170–177.

Thrul, J., Stemmler, M., Bühler, A., Kuntsche, E., 2013. Adolescents' protection motivation and smoking behaviour. Health Education Research 28 (4), 683–691.

Vaupel, J.W., 2010. Biodemography of human ageing. Nature 464, 536–542.

Whalley, L.J., Deary, I.J., Appleton, C.L., Starr, J.M., 2004. Cognitive reserve and the neurobiology of cognitive aging. Ageing Research Reviews 3 (4), 369–382.

Willis, B.L., Gao, A., Leonard, D., DeFina, L.F., Berry, J.D., 2012. Midlife fitness and the development of chronic conditions in later life. Archives of Internal Medicine 172 (17), 1333–1340.

Computation: Expectation-Maximization Algorithm

Geoffrey J McLachlan, University of Queensland, Brisbane, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The EM (Expectation–Maximization) algorithm is a general-purpose algorithm for maximum likelihood estimation in a wide variety of situations best described as *incomplete-data* problems. On each iteration of the EM algorithm, there are two steps – called the *Expectation step* or the E-step and the *Maximization step* or the M-step. Because of this, the algorithm is called the EM algorithm. Often in practice the log likelihood function cannot be maximized analytically. In such cases, it may be possible to compute iteratively the maximum likelihood estimate using a Newton–Raphson maximization procedure or some variant, provided the total number of parameters in the model is not too large. Another possible alternative is to apply the EM algorithm, which is typically simpler to implement.

Introduction

The expectation–maximization (EM) algorithm is a broadly applicable approach to the iterative computation of maximum likelihood (ML) estimates. It is useful in situations where algorithms such as the Newton–Raphson method may turn out to be more complicated. On each iteration of the EM algorithm, there are two steps – called the *Expectation step* or the E-step and the *Maximization step* or the M-step. Because of this, the algorithm is called the EM algorithm. This name was given by Dempster et al. (1977) in their fundamental paper. Hereafter we refer to this paper as the DLR paper or simply DLR.

The basic idea of the EM algorithm is to view the problem of computing the ML estimates (MLEs) from the data at hand as being computationally simpler if data had been observed also on some other random variables. Thus with the EM algorithm the observed data are viewed as being incomplete, and the problem is reformulated in a complete-data framework in which the ML solution may exist in closed form or may be amenable to ML computation with a standard computer package.

In order to carry out estimation within the complete-data framework, there is the need to address the existence of the additional data (referred to as the 'missing' data). The EM algorithm does this in a way that is closely related to the ad hoc approach to estimation with missing data, where the parameters are estimated after filling in initial values for the missing data. The latter are then updated by their predicted values using these initial parameter estimates. The parameters are then reestimated, and so on, proceeding iteratively until convergence. With the EM algorithm, the 'missing data' are handled by taking the conditional expectation of the complete-data log likelihood given the observed data, using the current estimate of the unknown parameter vector. This is the so-called E-step and it yields the Q-function $Q\left(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}\right)$, where $\boldsymbol{\Psi}$ is the vector of unknown parameters and $\boldsymbol{\Psi}^{(k)}$ is its current estimate. The M-step then updates the estimate of $\boldsymbol{\Psi}$ to $\boldsymbol{\Psi}^{(k+1)}$ by choosing $\boldsymbol{\Psi}$ to globally maximize $Q\left(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}\right)$ over $\boldsymbol{\Psi}$.

As the above idea behind the EM algorithm is intuitive and natural, algorithms like the EM have been formulated and applied in a variety of problems even before the DLR paper. But as pointed out by McLachlan and Krishnan (2008), it was in the seminal DLR paper that the "ideas were synthesized, a general formulation of the EM algorithm established, its properties investigated, and a host of traditional and nontraditional applications indicated." The situations where the EM algorithm can be applied include not only evidently incomplete-data situations, where there are missing data, truncated distributions, censored or grouped observations, but also a whole variety of situations where the incompleteness of the data is not all that natural or evident. These include statistical models such as random effects, mixtures, convolutions, log linear models, and latent class and latent variable structures (McLachlan and Krishnan, 2008).

Early EM History

The earliest reference to literature on an EM-type of algorithm is Newcomb (1886), who considered estimation of parameters of a mixture of two univariate normals. McKendrick (1926) gave a medical application of a method in the spirit of the EM algorithm. Hartley (1958) gave a treatment of the general case of count data and enunciates the basic ideas of the EM algorithm. In a series of papers starting with Baum and Petrie (1966), Baum and colleagues dealt with an application of the EM algorithm in a Markov model. Orchard and Woodbury (1972) introduced the missing information principle, which is very much related to the spirit and basic ideas of the EM algorithm and note the general applicability of the principle. The relationship between the complete- and incomplete-data log likelihood functions established by Orchard and Woodbury (1972) leads to the fact that the MLE is a fixed point of a certain transform. Efron (1967) introduced the so-called self-consistency principle in a nonparametric setting for a wide class of incomplete-data problems as an intuitive analog of the ML principle and introduced the self-consistency algorithm for right-censored problems.

Prior to the appearance of the DLR paper, the resolution of mixtures of distributions for a variety of situations and distributional families gave rise to a number of algorithms that can be regarded as particular applications of the EM algorithm. These papers, which were surveyed in McLachlan (1982), McLachlan and Basford (1988), and McLachlan and Peel (2000), include

the seminal paper of Day (1969). Also, McLachlan (1975, 1977) proposed an iterative method for forming the (normal theory-based) linear discriminant function from partially classified training data.

Sundberg (1974, 1976) dealt with properties of the likelihood equation in the general context of incomplete-data problems from exponential families, and arrived at special forms for the likelihood equation and the information matrix, which have come to be known as Sundberg formulas. Sundberg (1976) acknowledged that his key 'iteration mapping,' which corresponds to the EM mapping of DLR, was suggested to him by A. Martin-Löf in a personal communication in 1967. Beale and Little (1975) developed an algorithm and the associated theory for the multivariate normal case with incomplete data.

For a more detailed account of the EM algorithm both before and after the DLR paper, the reader is referred to McLachlan and Krishnan (2008), the first edition of which appeared 20 years after DLR, and its second edition, which appeared just over 10 years later in 2008. In other publications devoted to the EM algorithm, there is the paper by Meng and van Dyk (1997), who read a paper to the Royal Statistical Society to celebrate the 20th anniversary of the presentation of the DLR paper. In their paper, after a brief historical account, strategies were investigated that aim to make the EM algorithm converge faster while maintaining its simplicity and stability (for example, automatic monotone convergence in likelihood). In another publication to celebrate the 20th anniversary of the DLR paper, there is the special issue of the journal *Statistical Methods in Medical Research* focused on the impact of the EM algorithm on medical statistics (McLachlan, 1997). In this issue, Meng (1997) provides a fascinating account of the work by McKendrick (1926) and its link with the EM algorithm. He also illustrates the central ideas underlying the EM algorithm and suggests ways for speeding up the algorithm. Meng (2007) in an article entitled "Thirty Years of EM and Much More" points out how EM and Markov chain Monte Carlo (MCMC) are intimately related, and that both have been 'workhorses for statistical computing.'

Before we proceed to give a formal definition of the EM algorithm, we introduce some notation and give a brief introduction to ML estimation.

Maximum Likelihood Estimation

We let $\boldsymbol{Y}$ be a p-dimensional random vector with probability density function (p.d.f.) $g(\boldsymbol{y}; \boldsymbol{\Psi})$ on $\mathbb{R}^p$, where $\boldsymbol{\Psi} = (\Psi_1, \ldots, \Psi_d)^T$ is the vector containing the unknown parameters in the postulated form for the p.d.f. of $\boldsymbol{Y}$. Here the superscript T denotes the transpose of a vector or a matrix. The parameter space is denoted by $\boldsymbol{\Omega}$. Although we are taking $\boldsymbol{Y}$ to be a continuous random vector, we can still view $g(\boldsymbol{y}; \boldsymbol{\Psi})$ as a p.d.f. in the case where $\boldsymbol{Y}$ is discrete by the adoption of counting measure.

For example, if $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ denotes an observed random sample of size n on some random vector W with p.d.f. $f(\boldsymbol{w}; \boldsymbol{\Psi})$, then

$$\boldsymbol{y} = \left(\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_n^T\right)^T$$

and

$$g(\boldsymbol{y}; \boldsymbol{\Psi}) = \prod_{j=1}^{n} f\left(\boldsymbol{w}_j; \boldsymbol{\Psi}\right)$$

The vector $\boldsymbol{\Psi}$ is to be estimated by maximum likelihood. The likelihood function for $\boldsymbol{\Psi}$ formed from the observed data $\boldsymbol{y}$ is given by

$$L(\boldsymbol{\Psi}) = g(\boldsymbol{y}; \boldsymbol{\Psi})$$

An estimate $\widehat{\boldsymbol{\Psi}}$ of $\boldsymbol{\Psi}$ can be obtained as a solution of the likelihood equation

$$\partial L(\boldsymbol{\Psi}/\partial \boldsymbol{\Psi}) = 0$$

or equivalently,

$$\partial \log\ L(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = 0 \qquad [1]$$

Briefly, the aim of ML estimation (Lehmann, 1983) is to determine an estimate for each n ($\widehat{\boldsymbol{\Psi}}$ in the present context), so that it defines a sequence of roots of the likelihood equation that is consistent and asymptotically efficient. Such a sequence is known to exist under suitable regularity conditions (Cramer, 1946). With probability tending to one, these roots correspond to local maxima in the interior of the parameter space. For estimation models in general, the likelihood usually has a global maximum in the interior of the parameter space. Then typically a sequence of roots of the likelihood equation with the desired asymptotic properties is provided by taking $\widehat{\boldsymbol{\Psi}}$ for each n to be the root that globally maximizes the likelihood; that is, $\widehat{\boldsymbol{\Psi}}$ is the MLE. We shall henceforth refer to $\widehat{\boldsymbol{\Psi}}$ as the MLE, even though it may not globally maximize the likelihood. Indeed, in some of the examples on mixture models to be presented, the likelihood is unbounded. However, for these models there may still exist under the usual regularity conditions a sequence of roots of the likelihood equation with the properties of consistency, efficiency, and asymptotic normality; see McLachlan and Basford (1988, Chapter 1).

Although we are focusing here on the EM algorithm for computing MLEs in a frequentist framework, it can be equally applied to find the mode of the posterior distribution in a Bayesian framework.

Formulation of EM Algorithm

We let $\boldsymbol{Y}$ be the random vector corresponding to the observed data $\boldsymbol{y}$, having p.d.f. postulated as $g(\boldsymbol{y}; \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = (\Psi_1, \ldots, \Psi_d)^T$ is a vector of unknown parameters with parameter space $\boldsymbol{\Omega}$. The EM algorithm provides an iterative procedure for computing MLEs in situations where, but for the absence of some additional data, ML estimation would be straightforward. Hence in this context, the observed data vector $\boldsymbol{y}$ is viewed as being incomplete and is regarded as an observable function of the so-called complete-data. The notion of 'incomplete data' includes the conventional sense of missing data, but it also applies to situations where the complete data represent what would be available from some hypothetical experiment. In the latter case, the complete data may contain some variables that are never observable in a data sense. Within this framework, we let $\mathbf{x}$ denote the vector containing the augmented or so-called complete data,

and we let $\mathbf{z}$ denote the vector containing the additional data, referred to as the unobservable or missing data.

Even when a problem does not at first appear to be an incomplete-data one, computation of MLE is often greatly facilitated by artificially formulating it to be as such. This is because the EM algorithm exploits the reduced complexity of ML estimation given the complete data. For many statistical problems the complete-data likelihood has a nice form.

We let $g_c(\mathbf{x}; \mathbf{\Psi})$ denote the p.d.f. of the random vector $\mathbf{X}$ corresponding to the complete-data vector $\mathbf{x}$. Then the complete-data log likelihood function that could be formed for $\mathbf{\Psi}$ if $\mathbf{x}$ were fully observable is given by

$$\log\ L_c(\mathbf{\Psi}) = \log\ g_c(\mathbf{x}; \mathbf{\Psi})$$

Formally, we have two sample spaces $\mathcal{X}$ and $\mathcal{Y}$ and a many-to-one mapping from $\mathcal{X}$ to $\mathcal{Y}$. Instead of observing the complete-data vector $\mathbf{x}$ in $\mathcal{X}$, we observe the incomplete-data vector $\mathbf{y} = \mathbf{y}(\mathbf{X})$ in $\mathcal{Y}$. It follows that

$$g(\mathbf{y}; \mathbf{\Psi}) = \int_{\mathcal{X}(\mathbf{y})} g_c(\mathbf{x}; \mathbf{\Psi})\mathrm{d}\mathbf{x}$$

where $\mathcal{X}(\mathbf{y})$ is the subset of $\mathcal{X}$ determined by the equation $\mathbf{y} = \mathbf{y}(\mathbf{X})$.

The EM algorithm approaches the problem of solving the incomplete-data likelihood eqn [1] indirectly by proceeding iteratively in terms of the complete-data log likelihood function, $\log L_c(\mathbf{\Psi})$. As it is unobservable, it is replaced by its conditional expectation given $\mathbf{y}$, using the current fit for $\mathbf{\Psi}$.

More specifically, let $\mathbf{\Psi}^{(0)}$ be some initial value for $\mathbf{\Psi}$. Then on the first iteration, the E-step requires the calculation of

$$Q\left(\mathbf{\Psi}; \mathbf{\Psi}^{(0)}\right) = E_{\mathbf{\Psi}^{(0)}}\{\log\ L_c(\mathbf{\Psi})|\mathbf{y}\}$$

where the expectation operator E has the subscript $\mathbf{\Psi}^{(0)}$ to explicitly convey that this (conditional) expectation is being effected using $\mathbf{\Psi}^{(0)}$ for $\mathbf{\Psi}$.

The M-step requires the maximization of $Q(\mathbf{\Psi}; \mathbf{\Psi}^{(0)})$ with respect to $\mathbf{\Psi}$ over the parameter space $\mathbf{\Omega}$. That is, we choose $\mathbf{\Psi}^{(1)}$ such that

$$Q\left(\mathbf{\Psi}^{(1)}; \mathbf{\Psi}^{(0)}\right) \geq Q\left(\mathbf{\Psi}; \mathbf{\Psi}^{(0)}\right)$$

for all $\mathbf{\Psi} \in \mathbf{\Omega}$. The E- and M-steps are then carried out again, but this time with $\mathbf{\Psi}^{(0)}$ replaced by the current fit $\mathbf{\Psi}^{(1)}$. On the $(k+1)$ th iteration, the E- and M-steps are defined as follows:

- E-Step. Calculate $Q(\mathbf{\Psi}; \mathbf{\Psi}^{(k)})$, where

$$Q\left(\mathbf{\Psi}; \mathbf{\Psi}^{(k)}\right) = E_{\mathbf{\Psi}^{(k)}}\{\log L_c(\mathbf{\Psi})|\mathbf{y}\} \qquad [2]$$

- M-Step. Choose $\mathbf{\Psi}^{(k+1)}$ to be any value of $\mathbf{\Psi} \in \mathbf{\Omega}$ that maximizes $Q(\mathbf{\Psi}; \mathbf{\Psi}^{(k)})$; that is,

$$Q\left(\mathbf{\Psi}^{(k+1)}; \mathbf{\Psi}^{(k)}\right) \geq Q\left(\mathbf{\Psi}; \mathbf{\Psi}^{(k)}\right) \qquad [3]$$

for all $\mathbf{\Psi} \in \mathbf{\Omega}$.

The E- and M-steps are alternated repeatedly until the difference

$$L\left(\mathbf{\Psi}^{(k+1)}\right) - L\left(\mathbf{\Psi}^{(k)}\right)$$

changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood values $\{L(\mathbf{\Psi}^{(k)})\}$. DLR showed that the (incomplete-data) likelihood function $L(\mathbf{\Psi})$ is not decreased after an EM iteration; that is,

$$L\left(\mathbf{\Psi}^{(k+1)}\right) \geq L\left(\mathbf{\Psi}^{(k)}\right) \qquad [4]$$

for $k = 0,1,2,\ldots$. Hence convergence must be obtained with a sequence of likelihood values that are bounded above.

It can be seen from the above that it is not necessary to specify the exact mapping from $\mathcal{X}$ to $\mathcal{Y}$, nor the corresponding representation of the incomplete-data density g in terms of the complete-data density g_c. All that is necessary is the specification of the complete-data vector $\mathbf{x}$ and the conditional density of $\mathbf{X}$ given the observed data vector $\mathbf{y}$. The choice of the complete-data vector $\mathbf{x}$ is not unique. It is chosen for computational convenience with respect to carrying out the E- and M-steps. The EM algorithm is really a generic device. As pointed out by a referee of the DLR paper, the use of the word 'algorithm' to describe this procedure can be criticized, "because it does not specify the sequence of steps actually required to carry out a single E- or M-step."

DLR established important fundamental properties of the EM algorithm. In particular, these properties imply that typically in practice the sequence of EM iterates will converge to a local maximizer of the log likelihood function $\log L(\mathbf{\Psi})$. If $L(\mathbf{\Psi})$ is unimodal in $\mathbf{\Omega}$ with $\mathbf{\Psi}^*$ being the only stationary point of $L(\mathbf{\Psi})$, then for any EM sequence $\{\mathbf{\Psi}^{(k)}\}$, $\mathbf{\Psi}^{(k)}$ converges to the unique maximizer $\mathbf{\Psi}^*$ of $L(\mathbf{\Psi})$. In general, if $\log L(\mathbf{\Psi})$ has several (local or global) maxima and stationary values, convergence of the EM sequence to either type depends on the choice of starting point. Furthermore, DLR shows that convergence is linear with the rate of convergence proportional to $\lambda_{\max}$, where $\lambda_{\max}$ is the maximal fraction of missing information. This implies that the EM algorithm can be very slow to converge, but then the intermediate values do provide very valuable statistical information. This also implies that the choice of the complete-data problem can influence the rate of convergence, since this choice will determine $\lambda_{\max}$; see the example given in Meng (1997) using the data of McKendrick (1926).

There have been quite a few developments in the methodology of the EM algorithm since DLR. In particular, Wu (1983) has given a detailed account of the convergence properties of the EM algorithm, addressing, in particular, the problem that the convergence of $L(\mathbf{\Psi}^{(k)})$ to L^* does not automatically imply the convergence of $\mathbf{\Psi}^{(k)}$ to a point $\mathbf{\Psi}^*$.

Often in practice, the solution to the M-step exists in closed form. In those instances where it does not, it may not be feasible to attempt to find the value of $\mathbf{\Psi}$ that globally maximizes the function $Q(\mathbf{\Psi}; \mathbf{\Psi}^{(k)})$. For such situations, DLR defined a generalized EM algorithm (GEM algorithm) for which the M-step requires $\mathbf{\Psi}^{(k+1)}$ to be chosen such that

$$Q\left(\mathbf{\Psi}^{(k+1)}; \mathbf{\Psi}^{(k)}\right) \geq Q\left(\mathbf{\Psi}^{(k)}; \mathbf{\Psi}^{(k)}\right) \qquad [5]$$

holds. The above condition on $\mathbf{\Psi}^{(k+1)}$ is sufficient to ensure that

$$L\left(\mathbf{\Psi}^{(k+1)}\right) \geq L\left(\mathbf{\Psi}^{(k)}\right)$$

Hence the likelihood $L\ (\mathbf{\Psi})$ is not decreased after a GEM iteration, and so a GEM sequence of likelihood values must converge if bounded above; see McLachlan and Krishnan (2008) for a discussion on what specifications are needed on the process of increasing the Q-function in order to ensure that the limit of $\{L\ (\mathbf{\Psi}^{(k)}\}$ is a stationary value and that the sequence of GEM iterates $\{\mathbf{\Psi}^{(k)}\}$ converges to a stationary point.

Example: Estimation of Mixing Proportions

The publication of the DLR paper greatly stimulated interest in the use of finite mixture distributions to model heterogeneous data. This is because the fitting of mixture models by ML is a classic example of a problem that is simplified considerably by the EM's conceptual unification of ML estimation from data that can be viewed as being incomplete. We consider here as an illustrative example the ML estimation of the proportions in which the components of the mixture occur, where the component densities are completely specified.

Suppose that the p.d.f. of a random vector W has a g-component mixture form

$$\left(f(\boldsymbol{w};\mathbf{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{w}),\right) \qquad [6]$$

where $\mathbf{\Psi} = (\pi_1, \ldots, \pi_{g-1})^T$ is the vector containing the unknown parameters, namely the $g-1$ mixing proportions $\pi_1, \ldots, \pi_{g-1}$, since

$$\pi_g = 1 - \sum_{i=1}^{g-1} \pi_i$$

The component p.d.f.'s $f_i\ (\boldsymbol{w})$ are completely specified.

This mixture model covers situations where the underlying population is modeled as consisting of g distinct groups $G_1,\ldots,G_g$ in some unknown proportions $\pi_1, \ldots, \pi_g$, and where the conditional p.d.f of W given membership of the ith group G_i is $f_i\ (\boldsymbol{w})$. We let

$$\boldsymbol{y} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)^T$$

denote the observed random sample obtained from the mixture density [6].

The log likelihood function for $\mathbf{\Psi}$ that can be formed from the observed data $\boldsymbol{y}$ is given by

$$\begin{aligned}\log\ L(\mathbf{\Psi}) &= \sum_{j=1}^{n} \log\ f(\boldsymbol{w}_j;\mathbf{\Psi}) \\ &= \sum_{j=1}^{n} \log\ \left\{\sum_{i=1}^{g} \pi_i f_i(\boldsymbol{w}_j)\right\}\end{aligned} \qquad [7]$$

On differentiating [7] with respect to π_i $(i = 1, \ldots, g-1)$ and equating the result to zero, we obtain

$$\sum_{j=1}^{n}\left\{\frac{f_i(\boldsymbol{w}_j)}{f(\boldsymbol{w}_j;\widehat{\mathbf{\Psi}})} - \frac{f_g(\boldsymbol{w}_j)}{f(\boldsymbol{w}_j;\widehat{\mathbf{\Psi}})}\right\} = 0 \quad (i = 1, \ldots, g-1) \qquad [8]$$

as the likelihood equation, which clearly does not yield an explicit solution for

$$\widehat{\mathbf{\Psi}} = \left(\widehat{\pi}_1, \ldots, \widehat{\pi}_{g-1}\right)^T$$

In order to pose this problem as an incomplete-data one, we now introduce as the unobservable or missing data the vector

$$\mathbf{z} = \left(\mathbf{z}_1^T, \ldots, \mathbf{z}_n^T\right)^T \qquad [9]$$

where $\mathbf{z}_j$ is a g-dimensional vector of zero-one indicator variables and where $z_{ij} = (\mathbf{z}_j)_i$ is one or zero according as whether $\boldsymbol{w}_j$ arose or did not arise from the ith component of the mixture $(i = 1, \ldots, g;\ j = 1, \ldots, n)$. Of course in some applications, the components of the mixture correspond to externally existing groups and so each realization $\boldsymbol{w}_j$ in the observed random sample from the mixture density does have a tangible component membership. But in other applications, component membership of the realizations is just a conceptual device to formulate the problem within the EM framework.

If these z_{ij} were observable, then the MLE of π_i is simply given by

$$\sum_{j=1}^{n} z_{ij}/n \quad (i = 1, \ldots, g) \qquad [10]$$

which is the proportion of the sample having arisen from the ith component of the mixture.

The EM algorithm handles the addition of the unobservable data to the problem by working with the current conditional expectation of the complete-data log likelihood given the observed data. On defining the complete-data vector $\mathbf{x}$ as

$$x = \left(y^T, z^T\right)^T \qquad [11]$$

the complete-data log likelihood for $\mathbf{\Psi}$ has the multinomial form

$$\log\ L_c(\mathbf{\Psi}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij} \log \pi_i + C \qquad [12]$$

where

$$C = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij} \log\ f_i(w_j)$$

does not depend on $\mathbf{\Psi}$.

As [12] is linear in the unobservable data z_{ij}, the E-step (on the $(k+1)$th iteration) simply requires the calculation of the current conditional expectation of Z_{ij} given the observed data $\boldsymbol{y}$, where Z_{ij} is the random variable corresponding to z_{ij}. Now

$$\begin{aligned}E_{\mathbf{\Psi}^{(k)}}\left(Z_{ij}|y\right) &= \mathrm{pr}_{\mathbf{\Psi}^{(k)}}\left\{Z_{ij} = 1|y\right\} \\ &= z_{ij}^{(k)}\end{aligned} \qquad [13]$$

where by Bayes' Theorem,

$$z_{ij}^{(k)} = \tau_i\left(w_j;\mathbf{\Psi}^{(k)}\right) \qquad [14]$$

and

$$\tau_i\left(w_j; \Psi^{(k)}\right) = \pi_i^{(k)} f_i(w_j)/f\left(w_j; \Psi^{(k)}\right) \quad [15]$$

for $i = 1,\ldots,g$; $j = 1,\ldots,n$. The quantity $\tau_i\ (w_j; \Psi)$ is the posterior probability that the jth member of the sample with observed value w_j belongs to the ith component of the mixture.

The M-step on the $(k+1)$th iteration simply requires replacing each z_{ij} by $z_{ij}^{(k)}$ in [10] to give

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} z_{ij}^{(k)}/n \quad [16]$$

for $i = 1,\ldots,g$. Thus in forming the estimate of π_i on the $(k+1)$th iteration, there is a contribution from each observation w_j equal to its (currently assessed) posterior probability of membership of the ith component of the mixture model. This EM solution therefore has an intuitively appealing interpretation.

The computation of the MLE of π_i by direct maximization of the incomplete-data log likelihood function [7] requires solving the likelihood equation [8]. The latter can be identified with the iterative solution [16] provided by the EM algorithm after some manipulation (McLachlan and Krishnan, 2008).

Even before the formulation of the EM algorithm by DLR, various researchers have carried out these manipulations in their efforts to solve the likelihood equation for mixture models with specific component densities; see, for example, Day (1969). As demonstrated above for the estimation of the mixing proportions, the application of the EM algorithm to the mixture problem automatically reveals the iterative scheme to be followed for the computation of the MLE. Furthermore, it ensures that the likelihood values increase monotonically. Prior to the DLR paper, various researchers did note the monotone convergence of the likelihood sequences produced in their particular applications, but were only able to speculate on this monotonicity holding in general. As the $z_{ij}^{(k)}$ are probabilities, the E-step of the EM algorithm effectively imputes fractional values for the unobservable zero-one indicator variables z_{ij}. In so doing it avoids the biases associated with ad hoc iterative procedures that impute only zero-one values; that is, that insist on outright component membership for each observation at each stage rather than fractional allocation as with the EM algorithm.

Some Advances since DLR

One initial criticism of the EM algorithm was that it does not automatically provide an estimate of the standard errors, or the full covariance matrix in multivariate situations, of the MLE, as do some other approaches such as Newton-type methods. Hence since DLR, a number of methods have been proposed for assessing the covariance matrix of the MLE $\widehat{\Psi}$ of the parameter vector Ψ, obtained via the EM algorithm (McLachlan and Krishnan, 2008, Chapter 4). The other common criticism that has been leveled at the EM algorithm is that its convergence can be quite slow. Thus proposals have been put forward for accelerating its convergence (McLachlan and Krishnan, 2008, Chapter 5).

One of the major reasons for the popularity of the EM algorithm is that the M-step involves only complete-data ML estimation, which is often computationally simple. But if the complete-data ML estimation is rather complicated, then the EM algorithm is less attractive because the M-step is computationally unattractive. In many cases, however, complete-data ML estimation is relatively simple if maximization is undertaken conditional on some of the parameters (or some functions of the parameters). An account of such approaches can be found in McLachlan and Krishnan (2008). For instance, Meng and Rubin (1993) introduced a class of generalized EM algorithms, which they call the Expectation–Conditional Maximization (ECM) algorithm. The ECM algorithm takes advantage of the simplicity of complete-data conditional maximization by replacing a complicated M-step of the EM algorithm with several computationally simpler CM-steps. Each of these CM-steps maximizes the expected complete-data log likelihood function found in the preceding E-step subject to constraints on Ψ, where the collection of all constraints is such that the maximization is over the full parameter space of Ψ.

There are also the variants of the EM and ECM algorithms known as the Expectation–Conditional Maximization Either algorithm, the Alternating ECM algorithm, and the Parameter-Expanded EM algorithm aimed at speeding up convergence. These modified and extended versions all share the simplicity and stability of the EM and ECM algorithms.

There is now a whole battery of EM-related algorithms and more are still being developed. The current developments are, however, in the direction of iterative simulation techniques or MCMC methods, many of which can be looked upon as simulation-based versions of various EM-type algorithms.

See also: Bayesian Statistics; Mixture Models in Statistics; Mixture Models in Statistics; Monte Carlo Methods and Bayesian Computation: Overview.

Bibliography

Baum, L.E., Petrie, T., 1966. Statistical inference for probabilistic functions of finite Markov chains. Annals of Mathematical Statistics 37, 1554–1563.

Beale, E.M.L., Little, R.J.A., 1975. Missing values in multivariate analysis. Journal of the Royal Statistical Society B 37, 129–145.

Cramer, H., 1946. Mathematical Methods of Statistics. Princeton University Press, New Jersey.

Day, N.E., 1969. Estimating the components of a mixture of two normal distributions. Biometrika 56, 463–474.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1997. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B 39, 1–38.

Efron, B., 1967. The two sample problem with censored data. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 4. University of California Press, Berkeley, CA, pp. 831–853.

Hartley, H.O., 1958. Maximum likelihood estimation from complete data. Biometrics 14, 174–194.

Lehmann, E.L., 1983. Theory of Point Estimation. Wiley, New York.

McKendrick, A.G., 1926. Applications of mathematics to medical problems. Proceedings of the Edinburgh Mathematical Society 44, 98–130.

McLachlan, G.J., 1975. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. Journal of the American Statistical Association 70, 365–369.

McLachlan, G.J., 1977. Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. Journal of the American Statistical Association 72, 403–406.

McLachlan, G.J., 1982. The classification and mixture maximum likelihood approaches to cluster analysis. In: Krishnaiah, P.R., Kanal, L.N. (Eds.), Handbook of Statistics, vol. 2. North–Holland, Amsterdam, pp. 199–208.

McLachlan, G.J. (Ed.), 1997. Special Issue on "The Impact of the EM Algorithm on Medical Statistics." Statistical Methods in Medical Research, vol. 6, pp. 1–98.

McLachlan, G.J., Basford, K.E., 1988. Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York.

McLachlan, G.J., Krishnan, T., 2008. The EM Algorithm and Extensions, second ed. Wiley, New York.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.

Meng, X.L., 1997. The EM algorithm and medical studies: a historical link. Statistical Methods in Medical Research 6, 3–23.

Meng, X.L., 2007. Thirty years of EM and much more. Statistica Sinica 17, 839–840.

Meng, X.L., Rubin, D., 1993. Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80, 267–278.

Meng, X.L., van Dyk, D., 1997. The EM algorithm – an old folk song sung to a fast new tune (with discussion). Journal of the Royal Statistical Society B 59, 511–567.

Newcomb, S., 1886. A generalized theory of the combination of observations so as to obtain the best result. American Journal of Mathematics 8, 343–366.

Orchard, T., Woodbury, M.A., 1972. A missing information principle: theory and applications. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, CA, pp. 697–715.

Sundberg, R., 1974. Maximum likelihood theory for incomplete data from an exponential family. Scandinavian Journal of Statistics: Theory and Applications 1, 49–58.

Sundberg, R., 1976. An iterative method for solution of the likelihood equations for incomplete data from exponential families. Communications in Statistics – Simulation and Computation 5, 55–64.

Wu, C.F.J., 1983. On the convergence properties of the EM algorithm. Annals of Statistics 11, 95–103.

Computational Approaches to Model Evaluation

Daniel R Cavagnaro, California State University Fullerton, Fullerton, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by I.J. Myung, volume 4, pp. 2453–2457, © 2001, Elsevier Ltd.

Abstract

Computational approaches to model evaluation utilize the latest innovations in statistical computing to explore complex relationships between models and data. A wide variety of approaches have been proposed, all of which are intended to help researchers to select among competing explanations of a phenomenon under investigation. This article reviews several of these approaches, with an emphasis on their theoretical foundations, ranges of applicability, and timing of application in the scientific process.

Introduction

Rapid improvements in statistical computing technology over the last few decades have helped to fuel an explosion of interest in mathematical modeling in the social and behavioral sciences. As a result, 'modeling is now sufficiently mainstream that one can get the impression that models themselves are replacing the theories from which they evolved' (Pitt and Myung, 2002: p. 421). But a model is not interchangeable with a theory; it is merely one of many possible quantitative instantiations of a theory. Therefore it is possible to have several different models of the same phenomenon under investigation, each having justifiable assumptions, and each providing an adequate fit to the available data. In such cases, it is essential to have robust methods for evaluating and selecting among the competing models.

A thorough evaluation of a model requires a method that is sensitive to its overall data-fitting capabilities, not just its ability to fit a single set of data. The preferred approach to model evaluation is therefore to assess how well a model's fit to one data sample generalizes to future data generated by that same process (Myung, 2000). However, because a model's generalizability is not directly observable, this can be a difficult standard to implement in practice. Several different methods have been developed, varying in both their theoretical frameworks and ranges of applicability. This article reviews several of those methods, including minimum description length (MDL), Bayesian model selection, accumulative prediction error (APE), and parameter space partitioning (PSP).

A common thread in the various methods for evaluating generalizability is that they are applied after data have already been collected. In that sense, they are limited by the quality of the data that they have to work with. Therefore, their efficacy can be improved by making effective decisions prior to data collection, to ensure that the data will be maximally informative about the models under consideration. This article reviews two analyses that can aid in making such decisions: design optimization (DO) and adaptive design optimization (ADO).

What Makes a Model 'Good'

The first challenge in model evaluation is deciding what makes a model 'good.' On the face of it, the ability of a model to fit data generated in an experiment might seem like the only necessary criterion for model evaluation. Indeed, measures of fit such as root-mean-squares-error and percent-variance-accounted-for (r^2) are frequently reported in empirical research, and can often be computed at the touch of a button, thanks to modern software packages. However, such measures are not reliable for selecting among models because of the fact that data are not error-free. Rather, data are presumed to contain both inherent regularity and idiosyncratic noise: regularity reflecting the phenomenon of interest and noise resulting from measurement error and sampling error. The goal of modeling is to capture only the regularity, not the noise, but goodness-of-fit cannot distinguish between the two; it provides a single measure of a model's fit to both the regularity and the noise. Thus a model can achieve a good fit for the wrong reason by fitting noise instead of regularity. Therefore model selection based solely on a model's ability to fit a particular data sample may result in selecting a powerful model with little scientific significance.

Because it is impossible to eliminate error from data, additional criteria besides goodness-of-fit are required for evaluating models. One such criterion is 'model complexity,' which refers to the inherent flexibility in a model that enables it to fit diverse patterns of data. All other things being equal, simple models are attractive because they are sufficiently constrained to make them easily falsifiable. In contrast, a complex model does not assume a single structure in the data but, 'rather like a chameleon,' it is able to assume multiple structures by finely adjusting its parameter values (Myung and Pitt, 2009: p. 448). By virtue of this flexibility, a highly complex model can fit a data set better than a simple model can, even if the simpler model generated the data. It can do this by fitting both the regularity (good) and idiosyncratic noise (bad), whereas the simpler model would only fit the regularity. This problematic phenomenon is called 'overfitting' (Pitt and Myung, 2002).

What is desired for model evaluation is a yardstick that measures a model's ability to capture the underlying regularity only, not idiosyncratic noise. This is the *generalizability* criterion (Pitt et al., 2002). Generalizability refers to a model's ability to fit not only the current data sample, but also future data samples from the same data-generating process. A model generalizes best when it is just complex enough to fit the underlying regularity, but not so complex that it overfits the data. Thus, the challenge of maximizing generalizability can be

viewed as a delicate balancing act between goodness-of-fit and complexity (*see* Generalizability Theory).

Model Evaluation Statistics

Because a model's generalizability is not directly observable, it must be estimated based on an analysis of the relationship between the model and the observed data sample. This can be done by computing model selection statistics, which are powerful and principled tools for evaluating models. Several different model selection statistics have been developed, and although they may differ in their theoretical underpinnings, each one implements the generalizability criterion by trading off goodness-of-fit and model complexity. This section discusses three model selection statistics: MDL, Bayes factor, and APE.

Minimum Description Length

The MDL principle is a method of inference based on information theory. It proposes that the basic goal of modeling should be to formalize regularities in the data, and that the extent to which a model can compress the data indicates how well it achieves that goal (*see* Algorithmic Complexity). Thus the goal of model selection by MDL is to identify the model that allows the greatest compression of the data. Conceptually, the intent is not to discover the truth about the world so much as it is to provide the most concise account of our observations about the world. This section offers a brief introduction to the conceptual foundations of MDL and its implementation via the normalized maximum likelihood (NML). For full details, the reader is encouraged to refer to Grünwald (2007).

The key measure for model evaluation under the MDL approach is the 'code length' of the data under the model. The code length can be thought of as the number of symbols needed to represent the data by way of the model; a shorter code length means better compression of the data. This code length is approximated by way of the NML (Rissanen, 2001). The NML measures how well a model fits observed data, relative to how well that model could fit any possible data. Formally, the NML is given by

$$\mathrm{NML} = \frac{f\left(x|\widehat{\theta}_x\right)}{\int f\left(y|\widehat{\theta}_y\right)} \qquad [1]$$

where $\widehat{\theta}_x$ denotes the maximum likelihood estimate for the observed data set x, and $\widehat{\theta}_y$ denotes the maximum likelihood estimate for a hypothetical data set y. Thus the NML of the data set x is proportional to the maximized likelihood value, normalized by the sum of maximum likelihoods of all potential data sets that could be observed in an experiment. When x is defined over a discrete sample space (e.g., binomial data), the integration symbol is replaced by a summation symbol.

The code length of the NML is referred to as the 'stochastic complexity' of the data set x (with respect to the model M), and is given by

$$\mathrm{MDL} = -\ln f\left(x|\widehat{\theta}_x\right) + \ln \int f\left(y|\widehat{\theta}_y\right)\mathrm{d}y \qquad [2]$$

The first term in eqn [2] is a measure of the lack of fit of the model to the data. The second term is a measure of model complexity. Therefore, model evaluation by MDL is carried out by trading off lack of fit for complexity. The complexity term measures how well the model fits across all possible data sets and can be interpreted conceptually in the following way: a model that fits almost every data pattern very well would be much more complex than a model that provides a relatively good fit to a small set of data patterns but does poorly otherwise.

The integral in eqn [2] is over all possible data sets that could be observed in the same experiment that yielded x. In all but the simplest cases, this integral is computationally intractable (although, for an illustrative calculation with binomial data, see Shiffrin et al., 2008). Therefore, the stochastic complexity must be approximated. A useful approximation is given by

$$\mathrm{FIA} = -\ln f\left(y|\widehat{\theta}_y\right) + \frac{k}{2}\ln\left(\frac{n}{2\pi}\right) + \ln \int \sqrt{\det I(\theta)}\mathrm{d}\theta \qquad [3]$$

where n denotes the sample size, k is the number of model parameters, and det $I(\theta)$ is the determinant of the Fisher information matrix. This is known as the Fisher information approximation (FIA) (Rissanen, 1996).

Although eqn [3] is often much easier to evaluate than eqn [2], it is important to note that the FIA may exhibit anomalous behavior if certain conditions are not satisfied. Rissanen (1996) remarks that the most important things to verify are that the maximum likelihood estimate lies 'sufficiently inside' the interior of the model, that all elements of the Fisher information matrix are continuous in θ, and that the integral term is finite.

Bayesian Model Selection

A related model selection statistic to MDL can be derived from statistical theory. In Bayesian statistics, models are evaluated based on their posterior probabilities, given the observed data (those not familiar with Bayesian statistics should *see* Bayesian Statistics, Theory of). The ratio of the posterior probabilities of two models is known as the Bayes factor, and it measures the evidence that the data provides for one model over another (Kass and Raftery, 1995). Numerous computational methods have been developed for estimating the Bayes factor, mostly based on a computer-driven sampling methodology known as *Markov chain Monte Carlo* (MCMC). A recent summary lists as many as 15 different methods (Gamerman and Lopes, 2006, Chapter 7).

Although the Bayesian approach to model evaluation is derived from an entirely different theoretical framework than the MDL approach, the two approaches are closely connected in a number of ways. Like MDL, Bayesian model selection also maximizes generalizability by trading off goodness-of-fit and model complexity. This can be seen most clearly in the Bayesian information criterion (BIC), which was derived by Schwarz (1978) as an asymptotic approximation to the negative log of the posterior probability of a candidate model. The BIC is given by

$$\mathrm{BIC} = -2\ln f\left(y|\widehat{\theta}_y\right) + k\ln\left(\frac{n}{2\pi}\right) \qquad [4]$$

where, again, y is the observed data sample, $\widehat{\theta}_y$ is the maximum likelihood estimate for y, k is the number of parameters, and n is the sample size. The model with the lowest BIC is preferred. Just as in the expression for MDL in eqn [2], the first term in eqn [4] can be seen as a measure of lack of fit and the second term can be seen as a measure of model complexity.

A stronger connection between the Bayesian and MDL approaches to model evaluation can be seen by examining the FIA expression in eqn [3] and noting that the sample size appears in the second term but not the third. It follows that the contribution of the third term relative to that of the second term becomes negligible as the sample size becomes large. Neglecting the third term and multiplying the result by 2 yields the BIC. Yet another derivation of MDL as an asymptotic approximation of the Bayesian posterior probability can be obtained by assuming a special form of the parameter prior density (Myung et al., 2006). These connections between Bayesian model selection and MDL reveal that choosing the model that gives the shortest description of the data is essentially equivalent to choosing the model that is most likely to be true in the sense of probability theory.

The Prequential Approach

A third model selection statistic, also related to MDL and the Bayes factor, can be derived from a framework of predictive inference. The prequential approach (the name is derived from 'sequential prediction'; Dawid, 1984) is based on minimizing prediction error for unseen data. Conceptually, this is a direct implementation of the generalizability criterion in the sense that the model that generalizes best to future data should also have the smallest prediction errors for future data coming from the same source.

To see how this approach can be implemented in practice, consider a case in which data points are observed sequentially in an experiment. Suppose that after each data point is observed, a model is used to make a prediction about the next data point based on all of the data that have been observed so far in the experiment. The sum of the errors in those predictions is called the 'accumulative one-step-ahead prediction error' (APE; Wagenmakers et al., 2006) of the model. The lower the APE for a model, the better it is at predicting future data from past data.

The prequential approach has strong theoretical ties to both Bayesian- and MDL-based model selection. Wagenmakers et al. (2006) show that APE, when implemented with a logarithmic loss function and Bayesian one-step-ahead predictions, is equivalent to Bayesian model selection. They further show that, under many other conditions, APE, MDL, and Bayesian model selection converge as many data are collected. Consequently, the three different approaches to maximizing generalizability that have been presented: (1) choosing the model that is best at predicting future data from past data, (2) choosing the model that is most likely to be true in the sense of probability theory, and (3) choosing the model that provides the shortest description of the data – are essentially equivalent. For a discussion of the relative advantages and disadvantages of these three approaches, the reader is encouraged to refer to the review by Shiffrin et al. (2008).

Evaluation of Other Types of Models

Model selection statistics are valuable tools for evaluating models, but their usefulness is limited when confronted with the diverse range of models in the social and behavioral sciences. The primary limitation is that their application requires each of the models under consideration to be formulated as a parametric family of probability distributions. This leaves out many important classes of models such as connectionist models (*see* Cognitive Modeling: Connectionist Approaches), simulation-based models (*see* Social Simulation, Computational Approaches), and cognitive architectures (*see* Information-Processing Architectures: Fundamental Issues). Another limitation is that they summarize the potentially intricate relationship between model and data into a single real number, whereas researchers are often interested in a more nuanced analysis of the model–data relationship. For example, one might want to know what range of data patterns a model is capable of fitting, beyond that of the current data sample, and whether those patterns resemble human behavior. Overcoming these limitations requires a different approach to model evaluation. One such approach, called PSP, is described next. For a more thorough presentation of this material, including technical and implementation details of the method, the reader should refer to Pitt et al. (2006).

Parameter Space Partitioning

To understand PSP, it is useful to view models as data-generating mechanisms. In this view, a model can be said to be consistent with an empirical pattern of data if it can generate patterns of data that are qualitatively similar to the empirical pattern (i.e., it can fit the data) for some value of its parameters. However, evaluating a model's ability to reproduce a given pattern of data provides only a glimpse of what the model can do. In many cases, it would be useful to know what other qualitative patterns the model could produce. For example, if the model can produce any logically possible pattern of data, then its ability to reproduce the empirical pattern would be much less impressive. In that sense, there is a problem of qualitative model complexity in this type of evaluation.

PSP is sensitive to this qualitative notion of model complexity, and rather than quantifying it, PSP generates a map of the model's parameter space indicating which regions correspond to qualitatively distinct predictions. Study of this map provides insights into the global performance of the model and provides several criteria on which to compare models. For example, one could compare models by counting the number of distinct, qualitative patterns of data that each model can produce. Alternatively, if the empirical data can only be produced by a tiny region within a model's parameter space, one might consider it to be irrelevant to the model's overall performance and favor a model for which empirical data are more representative.

Implementation of PSP begins with defining what constitutes a data pattern. This definition will vary by situation depending on the pertinent characteristics of the models and data under study. What is important is that this definition discretizes the prediction space into a finite set of distinct patterns. Once a data pattern is defined, the next task

in applying PSP is to map the model's parameter space according to the data patterns that it can generate. A simple solution to this problem would be to evaluate or simulate the model on a fine grid of points spanning the parameter space and keep track of the pattern produced at each point. However, for every parameter in the model, the search space increases by one dimension. Even with as few as six parameters, brute-force search methods such as grid search are too inefficient to finish in a reasonable amount of time. For example, if a model has six parameters, each of which can range from zero to one, and one wished to search the full parameter space by means of a grid of points spaced at increments of 0.01, then one trillion (100^6) sample points would need to be evaluated.

To facilitate the mapping of higher dimensional parameter spaces, Pitt et al. (2006) proposed an MCMC-based search algorithm. The idea of the algorithm is to build chains of parameter points corresponding to the same pattern. The algorithm starts at a single point, and spawns new chains as new patterns are found. The theory of MCMC guarantees that each chain generated through this process will be uniformly distributed over the region of the parameter space corresponding to the same data pattern. The volumes of these regions can then be estimated by multidimensional ellipsoids using the estimated mean and covariance matrix of the MCMC sample of the region. A Java illustration of the algorithm, and Matlab code for its implementation, are available at the PSP homepage on J. I. Myung's website.

Model Evaluation at the Front End of an Experiment

Model selection tools assist researchers in making inferences about models given data samples collected in experiments. However, because they are applied after data have been collected, their potential to yield definitive conclusions is limited by the quality of the empirical data that they have to work with; sometimes the data simply do not provide a clear differentiation between models. The problem is that although models aim to predict patterns of data across a wide range of possible stimuli, practical limitations force researchers to select only a handful of them for actual testing. If the selected stimuli do not permit a clear differentiation of model predictions, then model selection may be inconclusive. Therefore, the choice of stimuli is critical.

The choice of stimuli is just one aspect of experimental design that can affect the potential informativeness of the result. Other design parameters such as the sample size, the number of treatments, and the proportion of observations to be allotted to each treatment group must be chosen. These choices impact not only the statistical value of the results but also the cost of the experiment. In that sense, an optimal design is one that maximizes the informativeness of the experiment while being cost-effective for the experimenter. This section discusses computational approaches to finding optimal designs. For more details, the reader is encouraged to see the recent articles by Myung and Pitt (2009) and Cavagnaro et al. (2010), from which the material in this section is primarily drawn.

Design Optimization

DO applies statistical inference at the front end of the experiment (before data have been collected) in order to make model evaluation easer at the back end of the experiment (after data have been collected). Statistical inference is used to exploit design variables (e.g., treatment levels and values, presentation schedule) with the aim of identifying an optimal experimental design that will produce the most informative data. It is similar to a power analysis that is conducted in the planning stages of an experiment, but instead of identifying only the optimal number of participants to test, the optimal experimental design as a whole is identified.

The basic idea behind DO is as follows. To identify an optimal design, one must be able to define the design space, which refers to all possible designs that could be used in an experiment. In addition, the model being evaluated must be expressed in formal, mathematical terms, with a parametric likelihood function. An optimal design is chosen according to a Bayesian, decision-theoretic framework in which each possible experimental design is treated as a gamble, whose payoff is determined by the outcome of an experiment carried out with that design. The idea is to estimate the informativeness of hypothetical experiments carried out with a given design, so that an 'expected utility' of that design can be computed. This is done by considering every possible observation that could be obtained from an experiment with that design and then evaluating the relative likelihoods and utilities of those observations. The design with the highest expected utility is chosen as the optimal design. Formally, the expected utility of a design is given by

$$U(d) = \sum_{m=1}^{K} p(m) \iint u(d, \theta_m, y) p(y|\theta_m, d) p(\theta_m) \mathrm{d}y \mathrm{d}\theta_m \qquad [5]$$

where $m = 1, \ldots, K$ is a set of K models being considered, d is a design, y is the outcome of an experiment with design d under model m, and θ_m is a parameterization of model m. The function $u(d, \theta_m, y)$ inside the integral in the above equation is called the 'local utility' of the design d. It measures the utility of a hypothetical experiment carried out with design d when the true model is m with parameter value equal to θ_m and the outcome y is observed. Thus, the global utility function $U(d)$ represents the expected value of the local utility function, where the expectation is taken over prior distributions on models, $p(m)$, and model parameters, $p(\theta_m)$, and all possible observations given a particular model-parameter pair through the sampling distribution $p(y|\theta_m, d)$.

Selection of a local utility function that captures the goals of the experiment is critical to the success of DO. For example, if the goal of the experiment is to discriminate among models, then the local utility function should measure the extent to which the models are discriminated by the hypothetical experiment, through a measure such as model recovery rate or posterior model probability. See Myung and Pitt (2009) for more details.

Adaptive Design Optimization

Application of DO does not have to be a one-shot process performed prior to an experiment. The decision-theoretic

framework of DO can easily be extended to apply adaptively throughout the course of an experiment. In an adaptively designed experiment, testing proceeds over the course of multiple stages, each of which may have a different design. At each stage, the information gained from all preceding stages is used to improve the design at the current stage. Thus the problem to be solved in ADO is to identify the most informative design at each stage of the experiment, taking into account the results of all previous stages, so that one can infer the underlying model and its parameter values in as few steps as possible. This is done by applying eqn [5] after each stage, with the prior distributions $p(m)$ and $p(\theta_m)$ replaced by posterior distributions $p(m|D)$ and $p(\theta_m|D)$, obtained through Bayesian updating (Gelman et al., 2004) based on the data D that have been collected so far in the experiment. For additional details, see Cavagnaro et al. (2010). C++ code is available on J. I. Myung's website.

Conclusions

Investigating the relationship between models and data is at the heart of the scientific endeavor. Our ability to model complex phenomena has increased substantially in recent decades, thanks in part to technological innovations in scientific computing. Although an abundance of models is a sign of a productive area of research, it also increases the importance of having robust methods for evaluating and selecting among models. Fortunately, the same technological innovations have also facilitated the development of sophisticated model evaluation methods. These methods differ in their theoretical foundations, in their ranges of applicability, and in the timing of their application (before or after data collection), but they are united in their aim to enhance our understanding of the complex relationship between models and data.

See also: Algorithmic Complexity; Bayesian Theory, History, Applications, and Contemporary Directions; Cognitive Modeling: Connectionist Approaches; Generalizability Theory; Information Processing Architectures: Fundamental Issues; Model Testing and Selection, Theory of; Social Simulation: Computational Models.

Bibliography

Cavagnaro, D.R., Myung, J.I., Pitt, M.A., Kujala, J.V., 2010. Adaptive design optimization: a mutual information based approach to model discrimination in cognitive science. Neural Computation 22 (4), 887–905.

Dawid, A.P., 1984. Statistical theory: the prequential approach. Journal of the Royal Statistical Society A 147, 278–292.

Gamerman, D., Lopes, H.F., 2006. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Chapman & Hall, Boca Raton, FL.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis, second ed. Chapman & Hall/CRC, Boca Raton, FL.

Grünwald, P.D., 2007. The Minimum Description Length Principle. MIT Press, Cambridge, MA.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90 (430), 773–795.

Myung, I.J., 2000. The importance of complexity in model selection. Journal of Mathematical Psychology 44, 190–204.

Myung, I.J., Navarro, D.J., Pitt, M.A., 2006. Model selection by normalized maximum likelihood. Journal of Mathematical Psychology 50, 167–179.

Myung, J.I., Pitt, M.A., 2009. Optimal experimental design for model discrimination. Psychological Review 116 (3), 499.

Pitt, M.A., Kim, W., Navarro, D.J., Myung, J.I., 2006. Global model analysis by parameter space partitioning. Psychological Review 113, 57–83.

Pitt, M.A., Myung, I.J., 2002. When a good fit can be bad. Trends in Cognitive Sciences 6, 421–425.

Pitt, M.A., Myung, I.J., Zhang, S., 2002. Toward a method of selecting among computational models of cognition. Psychological Review 109, 472–491.

Rissanen, J., 1996. Fisher information and stochastic complexity. IEEE Transactions on Information Theory 42 (1), 40–47.

Rissanen, J., 2001. Strong optimality of the normalized ML models as universal codes and information in data. IEEE Transactions on Information Theory 47, 1712–1717.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

Shiffrin, R.M., Lee, M.D., Kim, W., Wagenmakers, E.J., 2008. A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. Cognitive Science 32 (8), 1248–1284.

Wagenmakers, E.-J., Grünwald, P., Steyvers, M., 2006. Accumulative prediction error and the selection of time series models. Journal of Mathematical Psychology 50, 149–166.

Relevant Websites

http://faculty.psy.ohio-state.edu/myung/personal/do.html – C++ Code for Design Optimization.

http://faculty.psy.ohio-state.edu/myung/personal/psp.html – Parameter Space Partitioning Homepage.

Computational Neuroscience

Terrence J Sejnowski, Howard Hughes Medical Institute, Chevy Chase, MD, USA; Salk Institute for Biological Studies, La Jolla, CA, USA; and University of California, San Diego, La Jolla, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The goal of computational neuroscience is to understand how brains generate behaviors using computational approaches. Computational models of the brain explore how populations of highly interconnected neurons are formed during development and how they represent, process, store, act upon, and become altered by information present in the body and the environment. Techniques from physics, computer science, and mathematics are used to simulate and analyze these computational models and provide links between the wide range of levels that brains are investigated, from molecular interactions to large-scale systems. Models are also used for interpreting experimental data and providing a conceptual framework for the dynamical properties of neural systems, which should lead to more comprehensive theories of brain function.

Many different types of physical systems can solve computational problems, including slide rules and optical analog analyzers as well as digital computers, which are analog at the level of transistors and must settle into a stable state on each clock cycle. What these have in common is an underlying correspondence between an abstract computational description of a problem, an algorithm that can solve it, and the states of the physical system that implement it. This is a broader approach to computation than one that is based purely on symbol processing (Marr, 1982; Schwartz, 1990; Churchland and Sejnowski, 1992; Dayan and Abbott, 2001; Arbib, et al., 2002; Eliasmith and Anderson, 2003; Sterratt et al., 2011). (Figures 1 and 2).

There is an important distinction between general-purpose computers, which can be programmed to solve many different types of algorithms, and special purpose computers, which are designed to solve only a limited range of problems. Most neural systems are specialized for particular tasks, such as the retina, which is dedicated to visual transduction and image processing. As a consequence of the close coupling between structure and function, the anatomy and physiology of a brain

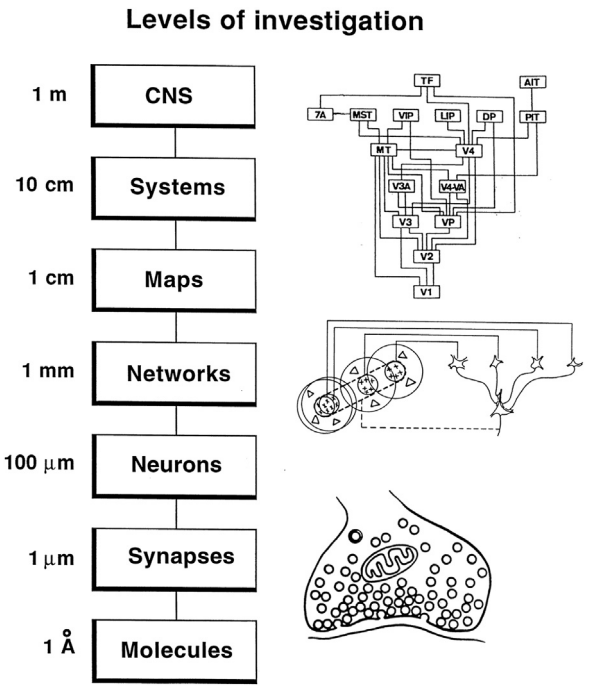

Figure 1 Multiscale levels of investigation for studying the brain. Each level requires a different type of model to explain the dynamical interactions between structures over a wide range of timescales. Integration across levels occurs by mapping the details of a lower level model into the parameters of a model at a higher level. Adapted from Churchland, P.S., Sejnowski T.J., 1992. The Computational Brain. MIT Press, Cambridge, MA.

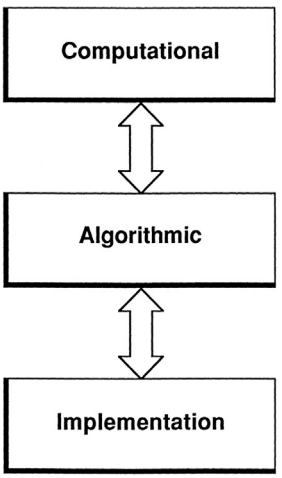

Figure 2 Levels of analysis. The computational level defines the problem to be solved; the algorithmic level provides procedures for solving the problem; the implementation level has devices that instantiate the algorithms. The two-way arrows indicate that constraints between levels can be used to gain insights in both directions. Theoretical approaches often start at the computational level and proceed downward toward the implementation level. Research in computational neuroscience often starts at the implementation level and proceeds up to the computational level. Adapted from Marr, D., 1982. Vision. Freeman, New York.

area can provide important clues to the algorithms that it implements and its computational function; in contrast, the hardware of a general-purpose computer may not reveal its function, which also depends on what software it is running.

Another major difference between brains and general-purpose digital computers is that computers are hardwired, but the connectivity between neurons and their properties are shaped by the environment during development and remain plastic even in adulthood. Thus, as the brain processes information, it changes its own structure in response to the information that is being processed. Adaptation and learning are important mechanisms that allow brains to respond flexibly as the world changes on a wide range of timescales, from milliseconds to years. The flexibility of the brain has survival advantage when the environment is nonstationary and the evolution of some cognitive skills may deeply depend on genetic processes that have extended the timescales for brain plasticity.

Brains are complex, nonlinear dynamical systems with feedback loops, and brain models provide intuition about the possible behaviors of such systems. The predictions of a model make explicit the consequences of the underlying assumptions, and comparison with experimental results can lead to new insights and discoveries. Emergent properties of neural systems, such as oscillatory behaviors, depend on both the intrinsic properties of the neurons and the pattern of connectivity between them. For example, the large-scale coherent brain rhythms that accompany different states of alertness and sleep arise from intrinsic properties of thalamic and cortical neurons that are reciprocally connected (Destexhe and Sejnowski, 2001).

Models at Different Levels of Detail

Realistic Models

Perhaps the most successful model at the level of the neuron has been the classic Hodgkin–Huxley model of the action potential in the giant axon of the squid (Koch, 1999; Koch and Segev, 2003). Data were first collected under a variety of conditions, and a model was later constructed to integrate the data into a unified framework. This type of model requires that most of the parameters in the model have been measured experimentally and only a few unknown parameters need to be fit to the data. Detailed models can be used to distinguish between different explanations of the data. Realistic models of thousands of interacting neurons can be explored with computer simulations.

In realistic models of neuron, synaptic signals are integrated on dendrites, leading to action potentials that are initiated near the soma and carried to other neurons through long axons. In compartmental models of neurons, dendrites and axons are modeled as cables with voltage-dependent ion channels, ligand-gate receptors, and ion transporters (Stuart et al., 1999; De Schutter, 2009). The great diversity of ion channels found in different combinations in different types of neurons gives rise to a wide range of properties. More than one set of parameters can reproduce a particular set of electrical recordings, which emphasizes the need for additional constraints on realistic models (Prinz et al., 2004). Active currents allow nonlinear integration to take place in dendritic branches and can support backpropagating action potentials that carry information in a retrograde direction from the cell body back to the distal synapses tree and influence synaptic plasticity through mechanisms that discussed below in the Section Learning and Memory.

The first model for the orientation specificity of neurons in the visual cortex was the feedforward model proposed by Hubel and Wiesel, which assumed that the orientation preference of cortical cells was determined primarily by converging inputs from thalamic relay neurons. Although there is strong experimental evidence for this model, local cortical circuits have been shown to be important in amplifying weak signals and suppressing noise as well as performing gain control through inhibitory feedback to normalize responses and extend the dynamic range. Models of competition between neurons in cortical circuits can explain many properties of single cortical neurons in awake, behaving monkeys during attention tasks (Reynolds and Heeger, 2009). These cortical models are governed by the type of attractor dynamics that was analyzed by John Hopfield (1982), who provided a conceptual framework for the dynamics of feedback networks.

Although the spike trains of cortical neurons are highly irregular, and are typically treated statistically, information may be contained in the timing of the spikes in addition to the average firing rate. This is well established for a variety of sensory systems in invertebrates and peripheral sensory systems in mammals (Rieke et al., 1996), but whether spike timing carries information in cortical neurons is an open research issue. In addition to representing information, spike timing could also be used to control synaptic plasticity through Hebbian mechanisms for synaptic plasticity, as discussed below in the Section Learning and Memory.

Small neural systems have been analyzed with dynamical systems theory (Izhikevich, 2006; Gerhard et al., 2013). This approach is feasible when the numbers of parameters and variables are small. Most models of neural systems involve a large number of variables, such as membrane potentials, firing rates, and concentrations of ions, with an even greater number of unknown parameters such as synaptic strengths, rate constants, and ionic conductances. In the limit that the number of neurons and parameters is very large, techniques from statistical physics become applicable in predicting the average behavior of large systems (van Vreeswijk and Sompolinsky, 1998). There is a midrange of system sizes where neither type of limiting analysis is possible, but where simulations can be performed (Bower and Beeman, 2007). One danger of relying solely on computer simulations is that they may be as complex and difficult to interpret as the biological system itself.

Signal Processing Models

Other types of models have been used to analyze experimental data in order to determine whether they are consistent with a particular computational assumption. For example, a 'vector averaging' technique has been used to compute the direction of arm motion from the responses of cortical neurons (Georgopoulos et al., 1986) and signal detection theory was used to analyze the information from cortical neurons responding to visual motion stimuli (Movshon and Newsome,

1996). The spatial tuning curves for neurons in the motor cortex could also be used to solve the equations of motion for multijoint arm movements (Tanaka and Sejnowski, 2013). Bayesian methods can effectively decode the position of a rat in a maze from place cell recorded in the hippocampus (Zhang et al., 1998). In these examples, the computational model was used to explore the information in the data but was not meant to be a model for the actual cortical mechanisms. Nonetheless, these models have been highly influential and have provided new ideas for how populations of neurons may represent sensory information and motor commands.

Cognitive Models

It is becoming possible to begin simulating cognitive functions such as perceiving, attending, deciding, and moving as computing continues to expand (van Hemmen and Sejnowski, 2006; Eliasmith et al., 2012). Neural network models that simplify the intrinsic properties of neurons can be helpful in understanding the computational consequences of information contained in large populations of neurons. An example of this approach is model of parietal cortex based on the response properties of cortical neurons (Pouget and Sejnowski, 2001). The parietal cortex is involved in representing spatial location of objects in the environment and computing transformations from sensory to motor coordinates. The model examined the issue of which reference frames are used in the cortex for performing these transformations. The model predicted the outcomes of experiments performed on patients with spatial neglect following lesions of the parietal cortex.

The oculomotor system has had a long tradition of using control theory models to guide experimental studies. Large-scale models of vertebrate brains are also becoming possible, especially in highly organized structures such as the retina, cerebellum, and cerebral cortex. However, despite the relatively stereotyped anatomical structure of the cerebellum, we still do not understand its computational functions. Early theories of the cerebellum as a motor controller have to be modified based on functional imaging studies that suggest that the cerebellum is also involved in higher cognitive functions. Modeling studies may help in exploring competing hypotheses.

Learning and Memory

One of the goals of computational neuroscience is to understand how long-term memories are formed through experience and learning. The strengths of interactions between neurons can be altered by activity. For example, high-frequency trains of stimuli at synapses in the hippocampus induce a form of long-term potentiation (LTP) that can last for days (Cooke and Bliss, 2006). Moreover, these synapses require simultaneous presynaptic activity and postsynaptic depolarization, as suggested by Hebb (1949). Hebbian synaptic plasticity can also be used to form maps and has been used to model the early development of the projection from the thalamus to the visual cortex (Toyoizumi and Miller, 2009). In particular, these models can explain why inputs from the right and left eyes form alternating stripes in primary visual cortex of cats and monkeys called ocular dominance columns. Specific mappings arise in the cortex because temporal contiguity in axonal firing is translated into spatial contiguity of synaptic contacts.

The change in the strength of synapses in the hippocampus and neocortex depends on the relative timing of spikes in the presynaptic neuron and the postsynaptic neuron. Reliable LTP occurs when the presynaptic stimulus precedes the postsynaptic spike, but there is long-term depression (LTD) when the presynaptic stimulus immediately follows the postsynaptic spike (Feldman, 2012). This temporally asymmetry in synaptic plasticity solves the problem of balancing LTD and LTP since chance coincidences should occur about equally with positive and negative relative time delays. When sequences of inputs are repeated in a network of neurons with recurrent excitatory connections, temporally asymmetric synaptic plasticity will learn the sequence and the pattern of activity in the network will tend to predict future inputs. There is evidence for this in the hippocampus where place cells representing nearby locations in a maze may be linked together (Blum and Abbott, 1996) and in visual cortex where simulations of cortical neurons can become directionally selective when exposed to moving visual stimuli (Shon et al., 2004).

The temporally asymmetric Hebbian learning rule can be used to implement the temporal difference learning algorithm in reinforcement learning and classical conditioning (Montague and Sejnowski, 1994; Sutton and Barto, 1998; Rao and Sejnowski, 2001). The unconditioned stimulus in a classical conditioning experiment must occur before the reward for the stimulus–reward association to occur. This is reflected in the temporal difference learning algorithm by a postsynaptic term that depends on the time derivative of the postsynaptic activity level. The goal is for the synaptic input to predict future reward: if the reward is greater than predicted the postsynaptic neuron is depolarized and the synapse strengthens, but if the reward is less than predicted, the postsynaptic neuron is hyperpolarized and the synapse decreases in strength. There is evidence in primates that the transient output from dopamine neurons in the ventral tegmental area carry information about the reward predicted from a sensory stimulus (Schultz et al., 1997).

The temporal window for classical conditioning is several seconds, much longer than the window for LTP/LTD observed at cortical and hippocampal synapses. A circuit of neurons in the basal ganglia and frontal cortex may be needed to extend the computation of temporal differences to these long time intervals (Sejnowski et al., 2014). It is surprising to find the same learning algorithm in different types of learning systems in different parts of the brain. This suggests that the temporal order of input stimuli is a useful source of information about causal dependence in many different learning contexts and over a range of time scales.

Future Directions

Molecular/Synaptic

Almost all of the proteins that make up synapses are now known and they sequencing of the human genome will make it possible to complete that list within the next 5 years. Coupled with high-voltage electron microscopy and methods for labeling these proteins it should be possible to develop a reasonably complete model for the neuromuscular junction and central synapses

within the next decade. This will allow us to understand the synapses as a molecular machine and to understand the mechanisms that are involved in synaptic plasticity (O'Donnell and Sejnowski, 2014). It is highly likely that advances in modeling and computational theory at this level will have direct impact on the pharmaceutical industry through the design of new drugs and new approaches to mental disorders.

Neuron/Network

Progress in this area has been slower because experimental technique for recording from many neurons simultaneously still lags behind. However, the advent of large electrode arrays and optical recording and stimulation techniques (Fenno et al., 2011) and reconstruction methods for circuits from electron microscope images (Denk et al., 2012) have made possible rapid progress. Computer models are being used to understand how the intrinsic properties of neurons and the synapses that join them produce complex spatiotemporal patterns of activity and how these networks are used to encode, store, and retrieve information in large neural systems.

Map/System

Functional magnetic resonance imaging has made it possible to localize activity within the human brain. New approaches, such as diffusion tensor imaging of long-range fiber tracts and magnetic resonance spectroscopy, should make it possible to extend current techniques to explore other structural and functions aspects of the brain (Smith et al., 2013). However, until the time resolution of these new techniques is improved, or they can be integrated with other techniques such as electroencephalography and magnetoencephalography, it will not be possible to uncover the mechanisms that underlie the activity that is observed with functional imaging. Understanding of brains at this integrated level will provide the ultimate insights into major philosophical questions about consciousness and autonomy (Koch, 2004).

Technology for Brain Modeling

Do new properties emerge as the number of neurons in a neural system becomes large? This question can be explored with large simulations of millions of neurons. Parallel computers have become available that permit massively parallel simulations, but the difficulty of programming these computers has limited their usefulness. An approach to massively parallel models introduced by Carver Mead is based on subthreshold cMOS VLSI (complementary metal–oxide–semiconductor very large-scale Integrated) circuits with components that mimic the analog computational operations in neurons. Several large silicon chips have been built that mimic the visual processing found in retinas. Analog VLSI cochleas have also been built that can analyze sound in real time. Analog VLSI chips have been built that mimic the detailed biophysical properties of neurons, including dendritic processing and synaptic conductances. These chips use analog voltages and currents to represent the signals, and are extremely efficient in their use of power compared to digital VLSI chips. A new branch of engineering called *neuromorphic engineering* has arisen to exploit this technology (Liu et al., 2015).

Conclusions

Although brain models are now routinely used as tools for interpreting data and generating hypotheses, we are still a long way from having explanatory theories of brain function. Digital computers have been increasing in computing power and data storage exponentially since 1950 and have transformed how experiments are performed, data are analyzed and the development of brain models and theories. Advance in brain imaging techniques, which has made it possible to study human brain activity noninvasively, would not have been feasible without fast digital computers. Similarly, improvements in optical imaging have been made by counting photons and computationally reconstructing subcellular structure with unprecedented resolution. The increasing power of computers is transforming our ability to analyze complex neural systems at many levels of investigation and will have far-reaching consequences.

At this stage in our understanding of the brain, a model should only be considered a provisional framework for organizing thinking and designing new experiments. Many partial models need to be explored at many different levels of investigation (**Figure 1**), each model focusing on a different scientific question. As computers become faster, and as software tools become more flexible, computational models will become increasingly important for guiding research. Close collaborations between modelers and experimentalists, facilitated by the Internet, should lead to an increasingly better understanding of the brain as a computational system.

See also: Cognitive Neuroscience; Eye Movement, Control of: Oculomotor Control; Information Processing Architectures: Fundamental Issues; Learning and Memory, Neural Basis of; Long-Term Potentiation and Long-Term Depression; Neural Synchrony as a Binding Mechanism; Neurons and Dendrites, Integration of Information in.

Bibliography

Arbib, M.A., Amari, S.-I., Arbib, P.H., 2002. The Handbook of Brain Theory and Neural Networks. The MIT Press, Cambridge, MA.

Blum, K.I., Abbott, L.F., 1996. A model of spatial map formation in the hippocampus of the rat. Neural Computation. 8, 85–93.

Bower, J.M., Beeman, D., 2007. Constructing realistic neural simulations with GENESIS. Methods in Molecular Biology 401, 103–125.

Cooke, S.F., Bliss, T.V., 2006. Plasticity in the human central nervous system. Brain 129 (Pt 7), 1659–1673.

Churchland, P.S., Sejnowski, T.J., 1992. The Computational Brain. MIT Press, Cambridge, MA.

Dayan, P., Abbott, L.F., 2001. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. MIT Press, Cambridge, MA.

Denk, W., Briggman, K.L., Helmstaedter, M., 2012. Structural neurobiology: missing link to a mechanistic understanding of neural computation. Nature Reviews Neuroscience 13 (5), 351–358.

Destexhe, A., Sejnowski, T.J., 2001. Thalamocortical Assemblies: How Ion Channels, Single Neurons and Large-Scale Networks Organize Sleep Oscillations. Oxford University Press, Oxford.

Eliasmith, C., Stewart, T.C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., Rasmussen, D., 2012. A large-scale model of the functioning brain. Science 338, 1202–1205.

Eliasmith, C., Anderson, C.H., 2003. Neural Engineering: Representation, Computation, and Dynamics in Neurobiological Systems. MIT Press, Cambridge, MA.

Feldman, D.E., 2012. The spike-timing dependence of plasticity. Neuron 75 (4), 556–571.
Fenno, L., Yizhar, O., Deisseroth, K., 2011. The development and application of optogenetics. Annual Review of Neuroscience 34, 389–412.
Georgopoulos, A.P., Schwartz, A.B., Kettner, R.E., 1986. Neuronal population coding of movement direction. Science 243, 1416–1419.
Gerhard, F., Kispersky, T., Gutierrez, G.J., Marder, E., Kramer, M., Eden, U., 2013. Successful reconstruction of a physiological circuit with known connectivity from spiking activity alone. PLoS Computational Biology 9 (7), e1003138.
Hebb, D.O., 1949. Organization of Behavior: A Neuropsychological Theory. John Wiley and Sons, New York.
van Hemmen, J.L., Sejnowski, T.J. (Eds.), 2006. 23 Problems in Systems Neuroscience. Oxford University Press, Oxford.
Hopfield, J.J., 1982. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences of the United States of America 79, 2554–2558.
Izhikevich, E.M., 2006. Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting. MIT Press, Cambridge, MA.
Koch, C., 1999. Biophysics of Computation: Information Processing in Single Neurons. Oxford University Press, Oxford.
Koch, C., 2004. The Quest for Consciousness: A Neurobiological Approach. Roberts and Company Publishers, Colorado.
Koch, C., Segev, I., 2003. Methods in Neuronal Modeling: From Synapses to Networks, second ed. MIT Press, Cambridge, MA.
Liu, S.-C., Delbruck, T., Indiveri, G., Whatley, A.M., Douglas, R. (Eds.), 2015. Event-Based Neuromorphic Systems. Wiley, New York.
Marr, D., 1982. Vision. Freeman, New York.
Montague, P.R., Sejnowski, T.J., 1994. The predictive brain: temporal coincidence and temporal order in synaptic learning mechanisms. Learning and Memory 1, 1–33.
Movshon, J.A., Newsome, W.T., 1996. Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. Journal of Neuroscience 16, 7733–7741.
O'Donnell, C., Sejnowski, T.J., 2014. Selective memory generalization by spatial patterning of protein synthesis. Neuron 82, 398–412.
Pouget, A., Sejnowski, T.J., 2001. Lesioning a basis function model of spatial representations in the parietal cortex: comparison with hemineglect. Psychological Review 108 (3), 653–673.
Prinz, A.A., Bucher, D., Marder, E., 2004. Similar network activity from disparate circuit parameters. Nature Neuroscience 7, 1345–1352.
Rao, R.P.N., Sejnowski, T.J., 2001. Self-organizing neural systems based on predictive learning. Philosophical Transactions of The Royal Society 361 (1807), 1149–1175.
Reynolds, J.H., Heeger, D.J., 2009. The normalization model of attention. Neuron 61 (2), 168–185.
Rieke, F., Warland, D., de Ruyter van Steveninck, R., Bialek, W., 1996. Spikes: Exploring the Neural Code. MIT Press, Cambridge, MA.
De Schutter, E. (Ed.), 2009. Computational Modeling Methods for Neuroscientists. MIT Press, Cambridge, MA.
Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science 275, 1593–1599.
Schwartz, E., 1990. Computational Neuroscience. MIT Press, Cambridge, MA.
Sejnowski, T.J., Poizner, H., Lynch, G., Gepshtein, S., Greenspan, R., 2014. Prospective Optimization. Proceedings of the IEEE 102, 799–811.
Shon, A.P., Rao, R.P.N., Sejnowski, T.J., 2004. Motion detection and prediction through spike-timing dependent plasticity. Network: Computation in Neural Systems 15, 179–198.
Smith, S.M., Vidaurre, D., Beckmann, C.F., Glasser, M.F., Jenkinson, M., Miller, K.L., Nichols, T.E., Robinson, E.C., Salimi-Khorshidi, G., Woolrich, M.W., Barch, D.M., Uğurbil, K., Van Essen, D.C., 2013. Functional connectomics from resting-state fMRI. Trends in Cognitive Sciences 17 (12), 666–682.
Sterratt, D., Graham, B., Gillies, A., Willshaw, D., 2011. Principles of Computational Modelling in Neuroscience. Cambridge University Press, Cambridge.
Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.
Stuart, G., Spruston, N., Hausser, M. (Eds.), 1999. Dendrites. Oxford University Press, Oxford.
Tanaka, H., Sejnowski, T.J., 2013. Computing reaching dynamics in motor cortex using Cartesian spatial coordinates. Journal of Neurophysiology 109, 1182–2013.
Toyoizumi, T., Miller, K.D., 2009. Equalization of ocular dominance columns induced by an activity-dependent learning rule and the maturation of inhibition. Journal of Neuroscience 29 (20), 6514–6525.
van Vreeswijk, C., Sompolinsky, H., August 15, 1998. Chaotic balanced state in a model of cortical circuits. Neural Computation 10 (6), 1321–1371.
Zhang, K., Ginzburg, I., McNaughton, B.L., Sejnowski, T.J., 1998. Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. Journal of Neurophysiology 79, 1017–1044.

Computational Organizational Theory

Geoffrey P Morgan and Kathleen M Carley, Carnegie Mellon University, Pittsburgh, PA, USA

Published by Elsevier Ltd.

Abstract

Research in computational organizational theory explores the complex interactions between organizations and their members. Organizations are legally autonomous entities that structure the efforts of individuals to achieve large goals. Interesting behavior often emerges from the interactions and demands of modeled primitives. This introduction describes the common ground and recent advances in multimodeling, multilevel modeling, and rapid model development. We conclude this summary with discussion of issues of model fidelity and model validation.

Introduction

Computational organization theory (COT) is a research discipline, which develops formal computational models to understand, extend, and examine organizational theory. Specific questions of interest to researchers in this field include organizational decision making, adaptation, and organizational learning. Often, these researchers are interested in the complex interdependencies between individuals and the organizations of which they are members. Organizations need individuals to make decisions, yet individuals benefit from membership in an organization: the organization is not merely the sum of its members.

Models developed by COT researchers tend to embody theory, represented as mathematical and computational structures and processes, of how an organization may behave. These models are intended to reproduce behavior analogous to the behavior of real-world organizations, but are implemented as computer programs. These models can be created and tested through low-cost simulation methodologies – although validation usually requires some form of comparison with either empirical data or with another model, which has been compared to empirical data (e.g., see Schreiber and Carley, 2013).

Organizational models have been implemented in many diverse ways – most often in the way that best suited the theoretical premises of the modeler and with which the modeler was fluent. These methods vary from formal mathematical models to boundedly rational agent models. Organizational models tend to focus on either processes (where the people are implicit) or people (where the processes are often implicit). This article focuses on organizational models that place the individual as an organizational decision maker, a people-oriented perspective (although subgroups may also be represented in such models). Thus, these models are multiagent simulations.

There are other methods for modeling organizations. For a thorough introduction to the art of modeling organizations as collections of intertwined processes, see Carlsen (1997) or Jørgensen (2004). In contrast, some researchers are primarily interested in organizational ecologies (Hannan and Freeman, 1977) – where the primary unit of analysis are organizations, rather than people within those organizations. Recent work in multilevel modeling (discussed below) is attempting to unite these various perspectives.

This article will frequently draw examples and grounding by presenting ideas from Construct (Carley, 1991a; Carley et al., 2009). Construct is a turn-based network-centric simulation of information and belief diffusion, based on the theory of Constructuralism (Carley, 1991b), and inheriting the information processing perspective of the Carnegie School (Simon, 1957; Cyert and March 1963).

Computational models, instantiated as simulations, are valuable to researchers in multiple ways. Computational methods allow theoreticians to explore the sources of complex and interesting organizational behavior, often finding that relatively simple rules when localized and in aggregate can produce these phenomena. This idea, that the interactions of simple processes produce complex and interesting behavior, is often called *emergence*. These methods also allow theoreticians to explore how large classes of organizations may behave in a particular environment of interest compared to other organizations from the same classification scheme. In general, the translation of theory into computational model tends to require that theory to be expressed with more clarity and precision than that required by textual description, which improves the resulting descriptions of the theory.

For applied research, computational models allow researchers to explore counterfactual scenarios for specific organizations of interest. These counterfactual scenarios allow researchers to answer 'what-if' questions, which it would be impossible to explore practically or ethically in the real world. Model results, of course, cannot be assumed to provide a specific and valid prediction of outcomes, but can often (if the theoretical model has been validated) be useful to examine from a trend analysis perspective.

COT makes these important assumptions about organizations and phenomena:

- Phenomena are modelable. If we cannot model it, we cannot examine it.
- Uncertainty is important. Individuals make mistakes. Environmental conditions can rapidly shift.
- Humans are boundedly rational. Because of cognitive and physical limitations, humans do not know everything about the world they live in, what they know may be wrong, and they may not be able to access what they know at all times.
- Organizations are organized to do work. Most metrics focus, explicitly or implicitly, on task performance.

- Outcomes are path dependent. Choices made in the past affect the present and future of an organization.
- Organizations may be designed and manipulated. Humans can create an initial plan for an organization, and these plans can be adjusted during an organization's life.

We will begin by defining the concept of an organization. From there, we will discuss the common elements of multiagent simulation, and the representational choices that inform these models. After we describe important elements of models, we will discuss methods used to both inform and test models. Finally, we will explore common issues that face all modelers and model consumers.

What is an Organization?

The definition of an organization, as opposed to other collections of people, can be difficult to clearly enumerate. We find it useful to characterize organizations as entities that have:

- Multiactor membership: actors may or may not be human.
- Task-orientation: organizations are structured to perform work.
- Goal-driven behavior: tasks support goals. Goals may change over time.
- Agency: organizations can alter themselves and their environment, and be affected by external changes.
- Independent knowledge: organizational knowledge is not directly tied to the knowledge of individuals.
- Legal autonomy: individuals are not directly responsible for the actions of the organization.

As such, organizations are multilevel entities such that the structure of the organization can be represented as a series of interconnected networks (Carley, 2005). One of these networks, and perhaps the most studied network, is the human-to-human network, which typically operates at a formal or authority level and at an informal or friendship, advice and collaboration level. Another network that is particularly important from a cyber perspective is the information technology (IT)-to-IT network, which is comprised of all such technology, associated databases, and the electronic modes of communication among these. Connecting the humans and the IT is another network often realized over multiple types of media. Then there are the knowledge network that links the agents to the data available and the activity network which links the agents to their tasks. And there is the traditional task network, familiar to operations research, showing what needs to be done before or in conjunction with what. A coherent collection of networks that describe a single organization is referred to as a meta-network (Carley, 2002). The analysis of these meta-networks for assessing organizations is a growing field, and there are analytical tools, such as ORA (Carley, 2014), optimized for this type of analysis. Using such tools organizational units can be designed and their performance assessed.

Organizations exist, from a functional perspective, to organize individuals and resources to achieve goals not within the capabilities of any single individual. As such, collections of individuals must be brought together. Individuals cooperate and perform tasks in service of the organization's goals. The goals of an organization are distinct from the goals of any of its members, although they may be well aligned. Goals are set by organizational processes, usually involving the input of multiple decision makers. In the pursuit of their goals, organizations change themselves and the environment around them. These changes may or may not be in their favor. Over time, as the organization achieves goals, unique organizational knowledge is formed. This organizational knowledge is not a description of the current knowledge of its individuals, but is instead distilled from individual contributions over time and is often stored in an organization's standard operating procedures. In order to organize individuals to do work toward large goals, particularly those that may have consequences for others, members of organizations must feel protected from the legal consequences of the organization's actions – the organization must have legal autonomy.

As stated earlier, organizations exist to organize the efforts of individuals toward large goals. How do organizations attempt this? They do this by structuring the responsibilities, interactions, and incentives of individuals to maximize the alignment of an individual's capabilities and goals with the organization's needs. This is sometimes referred to as the 'design' of an organization. Because there are complex interactions between the environment in which an organization operates, the capabilities of its members, and the goals of an organization, there is no one optimal design. One goal of COT's research is in understanding the interactions between all of these factors and an organization's design.

Now that we share a common understanding of an organization, we now discuss multiagent simulations and the representational choices required of developers of these simulations.

Defining a Multiagent Simulation

To define a working organizational computational model, the modeler/theoretician must establish representations of the environment.

The environment can be roughly characterized based on four questions.

- Is time continuous or discrete?
- Is space physical or virtual?
- What is in the environment?
- How could the environment change over time?

The following subsections will attempt to describe common modeling choices to answer each of these questions.

Representations of Time

Time, in simulation, can be either continuous (like the real world) or discrete, where it can be separated into equal-sized pieces. We call these pieces 'turns.' In discrete-time (or turn-based) simulation, everything that can act may have an opportunity to do so once per turn. In continuous-time simulations, agents may take new actions as computing resources allow. Some effort is usually made (in competitive environments in particular, see below) to ensure that agents have roughly equivalent computational resources, but not necessarily. Games like risk, chess, and monopoly can all be viewed as turn-based simulation.

Turn-based simulation may seem like a gross simplification of the real world, rendering such simulations suspect, but in practice it is a relatively harmless simplification that makes problems of agent concurrency and computational resource access trivial. Turn-based simulations are appropriate when (1) actors should be expected to have similar opportunities to participate, and (2) when the phenomenon of interest occurs over long periods of time: hours, days, or months, as opposed to seconds or minutes.

Representations of Space

Space in simulations may be physical or virtual. If space is physical, then agents are placed at points on a map. Physical space may, like time, be continuous or discrete. Discrete space involves dividing up the area of the map into equal-sized chunks, often called tiles or blocks. Cellular automata exist in such tile-worlds, and Schelling's Segregation Model (Schelling, 1971) also used such a world, see **Figure 1**. Tile-worlds may or may not be connected at their opposite edges – if both edges are connected, a tile-world forms a ring torus, which can remove problematic edge effects.

Physical space in simulations is useful if the theory requires it. A simulation examining, for example, how people interact as they move about an office building will require some idea of physical space and of entities moving through it and interacting with other entities. Typically, simulations with physical space use it to constrain agent perception and interaction.

Often in organizational models, space is virtual – entities are not placed in physical space. Instead, connections are drawn between entities that indicate who may affect and/or interact with who. These connections can be represented as networks, matrices of who is connected to whom, where '0' indicates no connection and '1' indicates a connection, as shown in **Figure 2**. Modern simulations often use multiple network semantics to enrich the virtual environment.

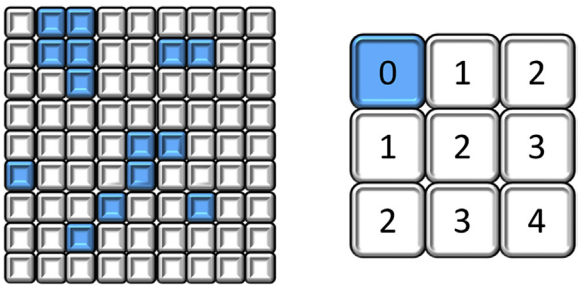

Figure 1 A tile-world, with agents in blue. On the right, Manhattan distance is shown relative to an agent on the top-left.

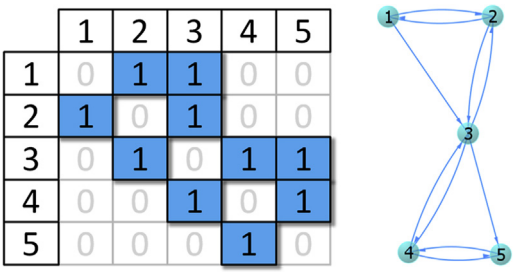

Figure 2 Networks are a convenient and useful representation of virtual space environments.

Virtual environments are useful when physical adjacency is not a primary driver of interaction in the simulation. For example, interactions between people doing knowledge work may be more driven by work responsibilities rather than physical adjacency. Virtual environments, like physical ones, are used to constrain agent perception and interaction possibilities.

Elements in the Environment

What exists within the simulation environment? This is a fundamental choice that should be driven by theory. Because organizations are goal-driven entities, organizational learning models usually include some representation of agents, knowledge, and tasks. Although an organization, as a whole, may be said to have some particular piece of information, that information may not be where it can be usefully applied (as defined by tasks). Thus, one common method of evaluating simulated organizational performance is through examining the agreement between three networks (although they may not be represented, explicitly, as networks by the modelers). These three networks are the knowledge network (who knows what), the assignment network (who should do what), and the requirements network (what knowledge is needed for what), as shown in **Figure 3**.

Not all organizational models include explicit representations of these three objects, but their presence is often implied in the theoretical descriptions of the model.

Agents

Usually, agents are representations of humans, but some simulations allow the set of entities to include nonhuman objects, such as IT systems, and may also include entities at different levels of granularity, an agent representing a single individual may interact with an agent representing a team of individuals.

Entities that can affect their own state or the state of others have agency. We refer to these entities as agents. Entities without agency will usually be referred to as objects. Agents may act on objects, but objects do not act directly on

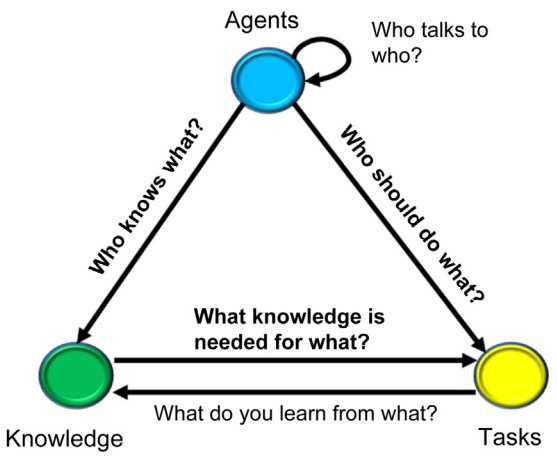

Figure 3 Three common objects within an organizational model are agents, knowledge, and tasks.

agents, although their presence may alter the actions chosen by an agent.

Agents have bounded rationality, limited access to the state of the larger world in which they inhabit. There are various limitations that can be prescribed to these agents, usually based on the larger theory of interest and drawn from human limitations. Agents may be only able to access state for things they can directly observe. They may forget knowledge they acquired earlier. They may have a limited number of refinement cycles before a decision must be made. Their perception of state may be error prone. Agents should have one or more of these limitations to be useful in a multiagent simulation.

Following directly from bounded rationality, agents also have limits to the possibilities available to them in their social interactions (Simon, 1991). Formal and informal mechanisms frequently prevent individuals from interacting with each other. Although it may be theoretically possible for the mailroom assistant to interact with the CEO of a Fortune 500 company, such interactions are unlikely to be routine and often may be safely discounted. Organizations often seek to impose structure to effectively distribute coercive, reward, and legitimate power (French and Raven, 1959). Individuals, in order to protect their own social capital, often seek to prune social ties for their own benefit (Burt, 1992, 1997). Thus, useful models of organizations will frequently limit interaction possibilities for their agents.

Agents must have actions available to them to be classified as agents. Again proceeding from theory, a modeler must define the actions available to agents and the decision heuristics used to select actions. Actions may be destructive (attacking another agent) or nondestructive (sharing knowledge with another agent), but all actions must have consequences that may alter future behavior of this or other agents.

Some models have entities that represent individuals but do not allow these representations to take independent action. Instead, a single model controller, which represents the organization, acts upon these agents. March's Mutual Learning Model (1991) is an influential example of such a model.

In Construct, agents may choose an interaction partner from all-available alters. They share knowledge with these partners, which may alter the partner's perception of the agent. The choice of who to interact with is weighted by two dynamic drives: (1) the desire to interact with alters similar to the agent, and (2) the desire to interact with alters who appear to have rare knowledge. Because of these two drives, the act of sharing knowledge will change the probabilities for both agent and alter of who they will interact with in the future.

Knowledge

Knowledge is information useful to the agents, either informing their choices of actions or helping them to complete tasks. Knowledge may be represented in a variety of ways; some methods include access to resources, rules that constrain an agent's actions, elements of the state of the environment that the agent can perceive, and as bits in a bit string. Organizational knowledge and individual knowledge may have different representations in the same simulation. In simulations where knowledge is represented as an explicit object, it is usually assumed to have some evidentiary basis – it is not merely an opinion. The transfer of knowledge, unlike other material resources, does not deplete the donor's store. Agents may not know, however, what they know.

In Construct, agents have two separate knowledge representations. The first is the model's ground state, what agents knows what, represented as multimode Agent x Knowledge matrix. Construct agents also have a perceptual matrix of what they think other people know, each agent thus has their own Agent x Knowledge matrix.

Tasks

Tasks are activities that are expected to have some useful larger purpose for the organization. They are not identical to goals, but may be instrumental in completing goals. Alternatively, they may be present in simulations where goals are never modeled or discussed.

Task performance may be dictated by agent actions, or may be inferred from agent characteristics. Task success is usually not guaranteed: it may be informed by agent knowledge, by exogenous probabilistic interference, or by third-party constraints.

In Construct, task performance is determined by access to knowledge relevant to the task. Tasks are assigned to agents at simulation run-time, and their performance is assessed each turn. Because tasks in Construct are assumed to be knowledge intensive, they are modeled as binary information tasks, if the agent has a majority of the required information for the task, the agent succeeds. If not, they fail at the task.

Current Research

In this section, we introduce some of the new trends in COT research. We discuss three trends: multimodeling, multilevel modeling, and rapid model development.

Multimodeling and Model Interaction

Multimodeling is an effort to leverage previous modeling work by incorporating one or more models into a single unified process. There are four ways models can be linked together on the basis of how they are connected by inputs and outputs, these are (1) integration, (2) docking, (3) interoperability, and (4) collaboration (Carley et al., 2012) as shown in **Figure 4**.

Integration, which involves refactoring models so that they work within a single common system envelope, is the most obvious but also most expensive method of integrating models. Representations of the phenomena must be aligned for all models, and this process often involves significant investment in new development work.

Docking (Axtell et al., 1996) is an approach to model validation where two models of the same phenomenon are directly compared on their features and their results. Docking two models require those models to have a common input and a common output. Model results are compared on three increasing levels of equivalence: relational, statistical, and numerical. Beginning with numerical equivalence, this is where one model produces the exact same numbers as the comparison partner, the two models can be, essentially, swapped interchangeably across tested values. Statistical equivalence is a characteristic of two models that they produce

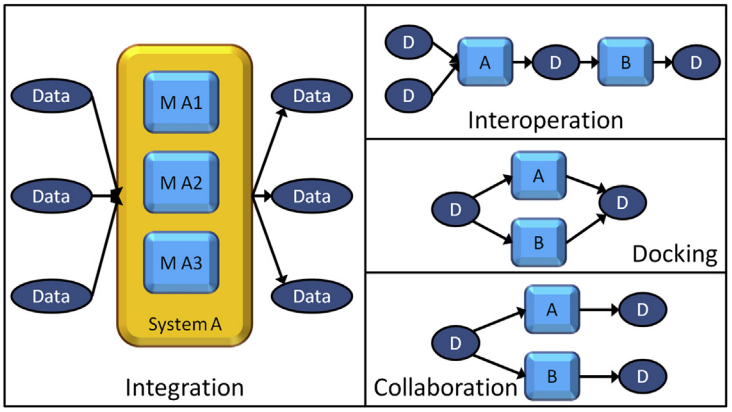

Figure 4 Multimodeling methods 'D' stands for data. A line is drawn from data to a model if it is input to that model, and a line is drawn from model to data if the data is an output of that model.

functionally identical results across tested values. Relational equivalence is where two models do not produce identical values, but instead produce similar patterns of output value across tested input values. Relational equivalence is usually considered sufficient to say that two models are docked to each other, because model inputs, although notionally similar, are rarely perfectly matched in the granularity and responsiveness of the input data.

Collaboration, like docking, involves two or more models that take in a common input, but do not produce similar outputs. Instead, both models take the input data, along with other data, and use that to predict different new data. Models can be easily made collaborative by matching the representations of their input data.

Interoperability is the idea that the outputs of one model can serve as inputs to another model. This is a useful and often inexpensive technique, particularly when building one model to extend and leverage the power of an older, more established model, because the data representations both required by and outputted by the older model are well understood. It is one approach that can lead toward Integration (where the models are in the same system envelope), but can be useful as part of a looser confederation of systems.

Multilevel Modeling

Multilevel modeling is an effort to account for both group and individual level effects concurrently within a model. Issues in multilevel representation have been central to organization science research (Rousseau, 1985). Statistical methods have found some success in grouping observations according to larger real-world structures (Snijders and Bosker, 2012). These methods help account for observations that, due to the inherent grouping of subjects within the frame of the research question, were not completely independent, such as students within a common classroom.

Within the modeling conducted in COT, multilevel modeling is used to indicate that there are multiple levels of dynamic description within a single model. For example, the model may have 42 individuals, 5 teams, and 2 multiteam organizations, but the number or characteristics of higher-order descriptions (the teams and organizations) should change over the length of the simulation. These descriptive characteristics are important to the model's mechanisms.

There may be multiple levels of agency within a single model. The individual may take actions within their purview, and the groups, teams, and companies of which they are members may also take actions separately as autonomous entities. The actions of individuals may influence the actions of the team, but they are not the same action. The actions allowed to each level of actor are usually distinct, although there may be parallels.

Adversarial models, where individuals are members of teams hostile to each other, could be described as a multilevel model, but only if the condition of the larger group unit has an effect on the actions of the individual, such as in Morgan et al., 2010. Adversarial models, where the state of the group has no direct impact on the individual's state should not be characterized as multilevel models.

In Construct, individuals use their knowledge of a not well-known actor's group affiliations to interpret what that actor is likely to know. As they interact with these actors, their perception of these groups will change, which will affect their likelihood of interacting with other not well-known alters. In this way, Construct is a multilevel model.

Rapid Model Development

Models take time to build. As models move from highly theoretical and intellective exercises to an interest in modeling specific scenarios of interest, the timeline for development shrinks even as the modeling requirements increase. Models can be made completely irrelevant by events if their results are predicated on assumptions no longer valid. For example, the organization has restructured, or a key leader has left the company. Thus, this research work is driven heavily by the needs of applied model users.

Rapid model development takes lesson from object-oriented and data-driven development paradigms. Rather than explicitly enumerating all state objects within a given system, the model software defines interface standards and file formats that can be used to provide this state. This decouples the

model, which specifies how agents will interact with their environment, from the model's initial state. Modelers must still define behavior for both agents and the environment.

This decoupling provides practical advantages when attempting to apply a theoretical model to a given applied context. Data used to inform the model can either be generated from theoretically useful test conditions, or created by translation of real-world data into a format digestible by the model. This translation process, although nontrivial, provides an opportunity for channeling the enormous data created by social media into models.

On a theoretical basis, it also provides an important distinction between the model as an environment and the initial state of the model. The validation of a model without a data-driven element requires evaluation not only of the model's behavior but also of the model's initial state. Separating these elements allows for concerns about the model to be more easily addressed.

For example, Construct models have been informed by the analysis of news articles (Pfeffer and Carley, 2012), by field studies of teams (Schreiber and Carley, 2007), by doctrinal texts on organizations (Lanham et al., 2011a), from combinations of doctrinal and text-analysis work (Lanham et al., 2014), and from subject matter expert estimates (Lanham et al., 2011b).

Model Fidelity and Model Validation: Common Concerns

Any discussion of COT is incomplete without discussing the highly related issues of model fidelity and model validation. The decisions modelers take in choosing to represent a phenomena are not guided, intrinsically, by deductive processes. Rather, the process is a creative one. Modelers may make poor decisions. Understanding how to assess models is important for consuming model results appropriately.

Model Fidelity

Model fidelity focuses on the issues of actor and state representation. Models exist on a fidelity axis, ranging from highly theoretical intellective models, which simplify and abstract the environment of phenomena to their essentials, to highly emulative models, which attempt to reproduce as realistic an environment as possible, as shown in **Figure 5**. All models, of course, must abstract at some point. A perfect reproduction of the modeled object is not a model, it is the thing itself.

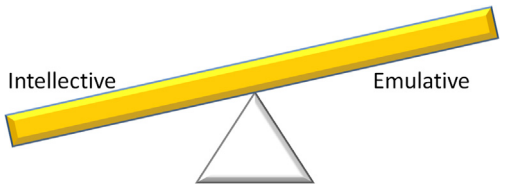

Figure 5 Balancing the fidelity needs of the simulation is an important modeling effort. Most computational organizational theory models are intellective.

High model fidelity (emulative simulation) is particularly important to simulations for use in training (Hays and Singer, 1989), although even intellective simulations can be useful for training in the business context (Van Ackere et al., 1993). The vast majority of COT models are intellective in nature.

Data-oriented design (as discussed in the rapid model development section above) blurs the line between emulative and intellective models. Models can be built with relatively simple behaviors available to both agents and the environment, but be instantiated with very complex and detailed initial states.

Questions of model fidelity that concern both a model developer and a model consumer are:

- Does this model include the factors I think are important to exploring this phenomena?
- Does this model include factors that I do not think are important to exploring this phenomena? Is it unnecessarily complex?
- Does this model ignore important interaction possibilities between agents and the environment?

Model Validation

Where model fidelity focuses on the questions of representation, model validation asks the question 'is this a good model?' By this, we mean 'does the model, given an initial state, produce reasonable behavior?' Of course, the answer is almost always 'it depends, how will the model's results be used?' Validation is thus tied, even more intrinsically than model fidelity, to the intended purposes of the model. Models expected to guide decision makers should be subjected to much more rigorous validation that models intended merely for discussion.

In the organizational theory domain, standard experimental assumptions do not hold. When making interventions on a real-world organization, it is rare for the experiment to be easily repeatable or for the results to be easily replicated. Experimental controls can be difficult to clarify, and gathering all necessary data is often infeasible or unethical. The difficulty of traditional experimentation makes a strong case for COT, but we should approach testing the resulting computational models with a similar level of rigor, evaluating them with 'virtual experiments.'

Virtual experiment methodology also makes some assumptions. It assumes the model's code is accurate, that the model has sufficient fidelity to the phenomena, and that standard statistical tests can be used to compare outcomes. Thus, questions of a model's fidelity should be addressed before a model's validation process even begins. Hopefully the modeler has considered the model's fidelity multiple times over the model's development period.

Validation, it should be noted, is hard. Model validation can require as much or more effort than the model's development. Human evaluators usually have some bias toward their products, and neutralizing this bias is difficult. Validation is knowledge intensive, requiring a deep understanding of the empirical phenomena being modeled (Louie and Carley, 2008). Validating agent-based COT simulations is also difficult because agent behavior is not predetermined and

path dependent – many simulation runs may be necessary to understand the full scope of possible system behaviors. Finally, the empirical data for validation may simply not be available at an appropriate granularity.

Validation can occur at multiple levels (Maxwell and Carley, 2009). The type of validity necessary for a model depends on its purpose (Burton and Obel, 1995). These levels are:

- **Internal validity**: the model's code is error free.
- **Parameter validity**: the model's parameters allow expression of all critical variables of the phenomenon.
- **Process validity**: the behavior of agents and the environment fit with current understanding of the phenomenon.
- **Face validity**: the results of the simulation, given the initial state, match the opinions of experts.
- **Pattern validity**: the outputs of the models show similar valence and magnitude of change across tested parameters to empirical data. It matches the pattern of the available empirical data. If abstracted, this is sometimes known as 'stylized fact validation.'
- **Value validity:** the outputs of the models show very similar values to those found in empirical data.
- **Theoretical validity:** the model correctly expresses a theory that explains the phenomena of interest.

Docking (Axtell et al., 1996; Burton and Obel, 1995) is also a model validation technique. A model that is docked to another model and has established relational equivalence (as discussed above), can make some claim to the pattern validity of the original model. Value equivalence, if achieved, would allow the docked model to inherit the value validity of the original model.

In multimodeling environments, validation of all the models at once can be difficult or impossible. Instead, inputs and outputs of these models are validated against real-world data individually, a process called *validation-in-parts*. This idea inherits from modern manufacturing process validation, which focuses on the outputs of the manufacturing effort, not on the mechanics of that effort.

Summary

In this article, we have introduced the area of COT. Research in COT explores the complex interactions between organizations and their members. Interesting behavior often emerges from the interactions and demands of modeled primitives, whether people or processes, although this introduction has focused on people-oriented simulation.

We have defined organizations as legally autonomous entities that attempt to achieve goals not within the capacity of any individual member. An organization's pursuit of its goals will change both the organization and the environment in which it operates, affecting other organizations (and themselves) either beneficially or adversely. Members of an organization do tasks, and both require and generate knowledge in the performance of these tasks.

To provide a useful introduction to COT research, we explored the requirements of defining and instantiating a multiagent simulation, using Construct (Carley, 1991a) as an illustrative example throughout. We discuss time, space, and elements within an environment, focusing on the importance of identifying agents and what actions those agents can take.

After identifying important elements to any single model, we discussed some important current trends in COT research. Multimodeling is an effort to leverage multiple models to provide a more complete picture of complex phenomena. Multilevel modeling allows actors at multiple levels of granularity to have important effects on actors at other levels. Rapid model development leverages data-driven approaches to software development: divorcing model behavior from a model's initial state.

Finally, we discuss important points every COT model consumer and developer should be aware of, issues of model fidelity and model validation. Model fidelity explores the issue of whether a model's representation of phenomena is appropriate, too simple, or unnecessarily complex for the phenomena of interest. Model validation asks whether the model is sufficient to the purpose for which it will be applied.

COT is a valuable and interesting area of research with important theoretical and applied areas. We hope this introduction has been useful.

See also: Behavioral Theories of Organization; Computational Approaches to Model Evaluation; Decision and Choice: Bounded Rationality; Dynamic Decision Making; Ecology: Organizations; Emergent Properties; Group Processes in Organizations; Interorganizational Relationships and Networks; Knowledge Representation; Network Analysis, History of; Organizational Control; Social Simulation: Computational Models; Stochastic Models.

Bibliography

Axtell, R., Axelrod, R., Epstein, J.M., Cohen, M.D., 1996. Aligning simulation models: a case study and results. Computational and Mathematical Organizational Theory 1, 123–141.

Burt, R.S., 1992. Structural Holes. Harvard University Press, Boston, MA.

Burt, R.S., 1997. The contingent value of social capital. Administrative Science Quarterly 42, 339–365.

Burton, R.M., Obel, B., 1995. The validity of computational models in organization science: from model realism to purpose of the model. Computational and Mathematical Organization Theory 1 (1), 57–71.

Carley, K.M., 1991a. Designing organizational structures to cope with communication breakdowns: a simulation model. Industrial Crisis Quarterly 5, 19–57.

Carley, K.M., 1991b. A theory of group stability. American Sociological Review 56 (3), 331–354.

Carley, K.M., 2002. Smart agents and organizations of the future. In: Lievrouw, L., Livingstone, S. (Eds.), The Handbook of New Media. Sage, Thousand Oaks, CA, pp. 206–220 (Chapter 12).

Carley, K.M., 2005. Organizational design and assessment in cyberspace. In: Rouse, W.B., Boff, K.R. (Eds.), Organizational Simulation. John Wiley Sons Inc., Hoboken, NJ.

Carley, K.M., 2014. ORA: a toolkit for dynamic network analysis and visualization. In: Alhajj, R., Rokne, J. (Eds.), Encyclopedia of Social Network Analysis and Mining, Springer.

Carley, K.M., Martin, M.K., Hirshman, B., 2009. The etiology of social change. Topics in Cognitive Science 1 (4), 621–650.

Carley, K.M., Morgan, G.P., Lanham, M.J., Pfeffer, J., 2012. Multi-modeling and socio-cultural complexity: Reuse and validation. In: Salvendy, G., Karwowski, W. (Eds.), Advances in Design for Cross-cultural Activities, vol. 2. CRC Press, Boca Raton, FL, pp. 128–137.

Carlsen, S., 1997. Conceptual modeling and composition of flexible workflow models (Doctoral Dissertation). www.idi.ntnu.no/grupper/su/publ/phd/carlsen-thesis.pdf.

Cyert, R., March, J.G., 1963. A Behavioral Theory of the Firm, second ed. Blackwell, Cambridge, MA.

French, J.R.P., Raven, B.H., 1959. The bases of social power. In: Cartwright, D. (Ed.), Studies in Social Power. Institute for Social Research, Ann Arbor, MI, pp. 150–167.

Hannan, M.T., Freeman, J., 1977. The population ecology of organizations. American Journal of Sociology 82 (5), 929–964.

Hays, R.T., Singer, M.J., 1989. Simulation Fidelity in Training System Design: Bridging the Gap between Reality and Training. Springer Verlag, New York.

Jørgensen, H.D., 2004. Interactive process models (Doctoral Dissertation). http://ntnu.diva-portal.org/smash/get/diva2:125137/FULLTEXT01.

Lanham, M.J., Morgan, G.P., Carley, K.M., 2011a. Data-Driven Diffusion Modeling to Examine Deterrence. IEEE Network Science Workshop, West Point, NY.

Lanham, M.J., Morgan, G.P., Carley, K.M., 2011b. Social network modeling and simulation of integrated resilient command and control in contested cyber environments. In: Levis, A., Carley, K.M., Karsai, G. (Eds.), Resilient Architectures for Integrated Command and Control in a Contested Cyber Environment. The Volgenau School of Engineering Department of Electrical and Computer Engineering System Architectures Laboratory, Technical Report SAL/FR-11-02, George Mason University, Fairfax, VA, pp. 63–94.

Lanham, M.J., Morgan, G.P., Carley, K.M., 2014. Social network modeling and agent-based simulation in support of Crisis De-escalation. IEEE Transactions on Systems, Man, and Cybernetics: Systems 44 (1), 103–110.

Louie, M.A., Carley, K.M., 2008. Balancing the criticisms: validating multi-agent models of social systems. In a Special Issue on "Simulating Organizational Processes" in Simulation Modeling Practice and Theory 16 (2), 242–256.

March, J.G., 1991. Exploration and exploitation in organizational learning. Organizational Science 2 (1), 71–87.

Maxwell, D., Carley, K.M., 2009. Principles for effectively representing heterogeneous populations in multi-agent simulations. In: Tolk, A. (Ed.), Complex Systems in Knowledge Based Environments. Springer-Verlag, pp. 199–228 (Chapter 8).

Morgan, J.H., Morgan, G.P., Ritter, F.E., 2010. A preliminary model of participation for small groups. Computational and Mathematical Organization Science 16, 246–270.

Pfeffer, J., Carley, K.M., 2012. Rapid modeling and analyzing networks extracted from pre-structured news articles. Computational and Mathematical Organizational Theory 18 (3), 280–299.

Rousseau, D.M., 1985. Issues of level in organizational research: multi-level and cross-level perspectives. Research in Organizational Behavior 7 (1), 1–37.

Schelling, T.C., 1971. Dynamic models of segregation. Journal of Mathematical Sociology 1 (2), 143–186.

Schreiber, C., Carley, K.M., 2007. Agent interactions in construct: an empirical validation using calibrated grounding. In: Proceedings of 2007 BRIMS, Norfolk, VA, 26–29 March, pp. 1–6.

Schreiber, C., Carley, K.M., 2013. Validating agent interactions in construct against empirical communication networks using the calibrated grounding technique. IEEE Transactions on Systems, Man and Cybernetics – Part A 43 (1), 208–214.

Simon, H.A., 1957. Models of Man: Social and Rational – Mathematical Essays on Rational Human Behavior in a Social Setting. Wiley, New York.

Simon, H.A., 1991. Bounded rationality and organizational learning. Organizational Science 2 (1), 125–134.

Snijders, T.A.B., Bosker, R., 2012. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, second ed. SAGE, Thousand Oaks, CA.

Van Ackere, A., Larsen, E.R., Morecroft, J.D.W., 1993. Systems thinking and business process redesign: an application to the beer game. European Management Journal 11 (4), 412–423.

Computer-Assisted Instruction

Joke Voogt, University of Amsterdam/Windesheim University, Amsterdam, The Netherlands
Petra Fisser, Netherlands Institute for Curriculum Development, Enschede, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since the early days of computer technology in education in the 1960s, it was claimed that computers can assist instructional practice and hence improve student learning. Since then computer technology has developed, and its potential for education has increased. In this article, we first discuss the changing terminology that accompanies the technological developments, followed by an overview of different types of applications of information and communication technology (ICT) for use in education. We then describe how different ICT applications align with different views on education. The potential of ICT for enhancing the curriculum as demonstrated in well-known studies is presented, followed by a picture of the actual use of ICT in education. Problems explaining the gap between potential and actual use are discussed.

The Need for New Conceptualizations of the Use of Computers in Education

Rapid changes in technology lead to changing conceptualizations of the use of computers in educational practice. In this respect, the term Computer Assisted Instruction (CAI) seems a bit obsolete, and closely connected to a specific type of computer use in instruction (Dede, 2008). CAI was adopted in the early days of computer use in education, indicating either a type of software program for education or a type of instructional process. However, the rapid developments in hardware and software ask for new terminologies and conceptualizations to study and discuss the potential of computers in education. The term computer technology, often used in the 1960s and 1970s, has been replaced by information and communication technology (ICT, mostly used in Europe) or information technology (IT) or technology (in North America). ICT refers to all technologies used for processing information and communication. Because of the integration of computers with communication systems including audio and video technology, terms such as multimedia or digital media are also being used. The development of computer technology from processing information to also supporting communication augmented its potential for education. In particular, the immersion of computer and other technologies in everyday life, as well as the different representations of technology (e.g., tablets, smart phones, robots), require educational researchers and professionals to rethink the potential of technology for education and the way it can be researched. Some therefore (e.g., van 't Hooft, 2008) prefer to speak about ubiquitous technology, to refer to the ever presence of technology in the environment but not typically seen or noticed. Although we realize that ICT has its limitations we will use this term in this article. We need to realize, however, that ICT is a broad concept. Its impact on curriculum, pedagogy, and student learning can only be discussed from the perspective of specific ICT applications, either hardware or software, that is being used in the teaching and learning process.

Software Applications for Educational Use

Introduction

Early software packages for education were mainly text based and used only simple graphics. This evolved in the 1990s and early 2000s to different combinations of modalities (pictures, moving images, sound, and text), often integrated into electronic environments. By then the increasing possibility to distribute information via the Internet was further explored and teachers were able to implement distance- or e-learning. As of 2003 the term "Web 2.0" emerged, indicating web applications that facilitate information sharing and collaboration on the World Wide Web. With Web 2.0, users are able to interact and collaborate with each other and generate their own content. Another development is cloud computing, which gained ground during the late 2000s. Cloud computing provides even more opportunities for teachers and students to work together online and it offers financial, administrative, and technical advantages. These new developments changed the way information is stored, distributed, and shared, but it has not made the more traditional software applications obsolete. The following types of applications are available for education.

Word Processor, Presentation Software, Spreadsheets, and Graphics Software

The software programs in this category are part of the basic application software of nearly every PC. They may support students in their learning process, because of their capability to structure information, and to easily present it in different modes. Research related to the use of word processors in education in particular shows the potential of this tool to stimulate students to write and to support the writing process. Presentation software is used to display information and it is therefore usable for organizing and presenting information, but also for (learning to) structuring information in such a way that people can learn from and work with it. Spreadsheets emerged as a potential tool for learning mathematics and science. Spreadsheets contain features for modeling situations, analyzing change, and linking key problem variables to tables and charts that can be presented visually. This provides teachers with tools to show science and mathematics concepts and processes, and gives students the opportunity to utilize these capabilities to gain a deeper understanding of the concepts and processes. Graphics software enables people to manipulate visual images. This can be incorporated in word processing, presentations, and the use of spreadsheets. The opportunities for manipulating images and using it for educational purposes have evolved rapidly in recent years. Not only the quality of the

images improved, but also the user-friendliness of the accompanying software. Even young children are able to work with graphics software, for instance while making mind maps about the topic that they have to learn about.

Drill and Practice and Tutorials

Drill and practice software is one of the earliest types of educational software and has as main purpose to let the learner memorize information (e.g., facts, language vocabulary). Most drill and practice programs will provide the correct answer after answering a question incorrect, so that the learner is not caught in a particular item for a too long period of time. The software often has facilities to provide feedback to the learners and to keep track of their performance. Tutorials can be used for learning new educational content, and the main purpose of tutorials is to instruct the learner. Tutorial software assesses the learner's knowledge and presents new instruction or gives a new practice based on the assessment of the learner's knowledge level. Newer forms of drill and practice and tutorials are adaptive tutoring software, intelligent computer assisted instruction (ICAI), and intelligent tutoring systems (ITSs). Traditional computer assisted instruction and tutoring systems operate according to an algorithm and provide sets of lessons, samples, exercises, and tests for the student. ICAI and ITSs diagnose the strengths and weaknesses of students, then adapt instruction to specific needs, and provide feedback through hints, examples, encouragement, and exercises that are appropriate for that specific student. ICAIs or ITSs reduce the time spent by teachers in grading and supervising. A main drawback of these systems is that the costs of producing them are high, and one of the main criticisms is that the systems do not really allow students to learn from their mistakes because the program guides them toward the right answer.

Simulations, Haptics, Games, and Virtual Worlds

Simulations are representations of a part of reality and provide learners the opportunity to engage in situations that would otherwise be too hazardous, costly, or lengthy to conduct in the classroom. Simulations could represent for instance certain behavior of a population or a set of processes. By manipulating variables in the representations and giving students the opportunity to visualize, model, explore, try out a variety of strategies, and study the effects, the students can gain an understanding of the reality. A review study about the use of simulations in science education showed that simulations enhanced with instructional support can have a positive effect on student learning (Rutten et al., 2012). A recent development is haptics, simulations in which also the sense of touch is incorporated. An example of this is hapTEL, a workstation that simulates clinical conditions for dental training (http://www.haptel.kcl.ac.uk/).

Games are competitive, situated, interactive learning environments based on a set of rules, and/or an underlying model, which with certain limitations and uncertain conditions, a challenging goal to be achieved. An educational or serious game is a game designed for educational purposes. Research has shown that the challenge to solve cognitive puzzles is the main motivation for playing games. The challenges in a game can be both visible and hidden. Visible challenges are for example the tasks that the player has to perform. Hidden challenges are discovering the unique rules, opportunities, constraints, and strategies of the game. To overcome the visible challenges, the player must first master the hidden challenges and in order to be able to do this the player must create a mental model of the obstacles built into the game world. This way of learning is very active, complex, and learner-centered. Educational or serious games have evolved vastly over the past decade. Examples of different kinds of games are computer games and massive multiplayer online games (MMOG). A well-known example of a multiuser virtual environment that fosters learning scientific inquiry and twenty-first century skills in middle grade students is The River City Project (http://muve.gse.harvard.edu/rivercityproject/). River City has the look and feel of a video game but contains educational content. While playing the students take on the identity of an avatar, communicate with other avatars, find hints and artifacts and have to define several hypotheses that have to be tested. The students learn principles and concepts of science, they acquire reasoning skills, and they carry out investigations that test their ideas. Such an environment is situated and authentic, which is important for fostering learning. Next to multiuser virtual environments other virtual worlds are emerging, such as MMOG, World of Warcraft and Metaverses, virtual worlds that are primarily social oriented. An example of such a world is Second Life. Second Life is a 3D online simulation where you walk around in the shape of an avatar, interact with others, buy property, etc. It is not intended as a game, but as a social space. This kind of virtual world offers many opportunities for education, from language to technical and marketing subjects, and there are currently hundreds of educational sites within Second Life.

Multimedia Cases

Multimedia cases offer the possibility to study parts of reality by presenting illustrations of real world practice that can be discussed and studied from several perspectives. Multimedia cases usually combine video segments, audio, and textual commentary together with random access facilities and hyperlinks, and so create an open-ended student-centered learning environment. Though used in many fields of study, they are increasingly used in teacher education for studying the complexity of the classroom from various viewpoints. In this context, multimedia cases provide the potential to incorporate contextual, theoretical, and pedagogical information into a rich and multilayered representation of classroom teaching (Voogt and Knezek, 2008).

Web 2.0

With Web 2.0 users are able to interact and collaborate with each other and to generate their own content. Examples of Web 2.0 include social networking sites, blogs, wikis, video sharing sites, mashups, and folksonomies. Social networking sites are online sites that focus on building social networks and relations among people, who share interests and/or activities with people having similar interests. This phenomenon gains interest of educators. They use social networks (Facebook, Ning, Wiki, etc.) as a communication tool, partly because they know that many students are already using a wide range of

these applications. The fact that many people use social networking as a means of communication brings concerns in relation to giving out too much personal information and the control of data; information can be altered, removed, or passed to third parties. The debate on how to deal with this in education has just begun.

Augmented Reality

Augmented reality (AR) is a view of a real-world physical environment or object, whose elements are augmented (supplemented) by computer-generated input such as sound, video, graphics, or GPS data. As a result, the technology enhances one's perception of reality. Typically, users have special equipment such as a heads up display, goggles, or a special computer. In education, this can be used to simplify complex tasks by inserting additional information into the field of view, or by for instance creating virtual objects in museums and exhibitions to explain certain objects that are physically present in the exhibition. An example of using AR in education is Handheld Augmented Reality Project (HARP) (http://isites.harvard.edu/icb/icb.do?keyword=harp), a project in which handheld computers and GPS receivers are used for middle school math and literacy curricula that use AR to deliver instruction in the form of games. These games are played in a real physical location, and on their handheld a map is displayed with a layer of digital objects and virtual people who exist in an AR world superimposed on real space. This creates a learning environment that stimulates the students to explore and to solve various mysteries and problems. Exploring and solving these problems impose active learning and new means of information gathering and communication.

Data Logging and Modeling

Data logging (formerly also called Microcomputer-Based Laboratories) is an application particularly useful for science education. Specific hardware added to a PC or a handheld computer in the form of probes allows for real-time data collection. It becomes much easier to repeat experiments, to measure different variables at the same time, to use a very short-time or a very long-time range, to analyze data, and to represent data graphically. Data logging can be used to collect and analyze data (temperatures, light intensity, motions, etc.) and may support students in their laboratory work. Instead of spending instructional time on data collection, it can be used for analyzing and interpreting data. However, to get an in-depth understanding in the relationship between variables, it is important to model these relationships. Modeling software supports learners in constructing their own models or adding to part-built models. Depending on the modeling environment this may involve: specifying formulae; writing a program in a simple programing language or manipulating a graphical or pictorial modeling language.

Views on the Use of ICT in Education

It is generally accepted that ICT as such does not support learning. The full potential of ICT for learning may only become realized when ICT is well integrated into a learning environment. In the early days of computer use in education these learning environments were narrowly defined, and referred to the computer software that supports certain types of learning. However, in more recent definitions, learning environments are considered dynamic complex systems, in which people (teachers, students, and others), media (including ICT), and the architecture of the (physical and/or virtual) environment interact (Voogt and Knezek, 2008). Dede (2008) provides an overview on how ICT was understood and used by different theories of learning. The 'narrow' view of an ICT-enhanced learning environment is represented in behaviorist and cognitive theories of learning. In the behaviorist perspective ICT applications, such as drill and practice and simple tutorials, were applied to master factual knowledge and (routine) skills at learner's own pace, and so allow for individual learning trajectories. Meta-analyses report small effects on student learning for these kind of applications. Cognitive theories of learning, also allow for individual learning trajectories but focus on cognitive understanding of complex concepts and skills. In this approach, ICT applications, such as ITS, based on artificial intelligence were developed to assist the learner in developing algorithms and heuristics needed to solve well-defined complex problems for specific subject domains. The effect sizes are somewhat higher than the ICT applications based on behaviorist theories of learning (Dede, 2008). Finally, constructivist theories of learning assume that the learner has to construct new knowledge and understanding through active participation in the learning process, eventually together with other learners or experts. In this approach, learning environments are more broadly defined and created in such a way that they allow for authentic learning and learner-centered education. Dede (2000) argues that ICT applications from a constructivist perspective of learning may contribute to

- The realization of a curriculum that is centered around real-world problems.
- Involve students in virtual communities of practice.
- Use advanced tools similar to those in today's high-tech workplaces.
- Facilitate guided, reflective inquiry through extended projects.
- Utilize modeling and visualization as powerful means of bridging between experience and abstraction.
- Enhance students' collaborative construction of meaning via different perspectives on shared experiences.
- Include pupils as partners in developing learning experiences and generating knowledge.
- Foster success for disabled and disenfranchised students.

A broad range of ICT applications can be incorporated in constructivist learning environments, such as general productivity tools (e.g., wikis, blogs, spreadsheets, the Internet), specific ICT applications to help individual learning (e.g., simulations, data logging, multimedia cases), as well as ICT applications that support collaborative learning (e.g., discussion forums, shared workspaces, virtual worlds). Because of the wide range of applications and settings there are no meta-analyses available that provide information about the effect of ICT-enhanced constructivist learning environments on student learning. However, several studies, such as the Jasper

Woodbury series (The Cognition and Technology Group at Vanderbilt, 1997), which was one of the first ICT-enhanced learning environments designed from a constructivist perspective, have shown effects on student learning.

ICT-Enhanced Learning Environments in the Curriculum

Several ICT-enhanced learning environments designed from a constructivist approach to learning have been studied for more than 15 years and have shown the potential of ICT to support student learning. They offer insights in the way ICT-enhanced learning environments need to be designed, in order to realize its goals. Knowledge Forum from Scardamalia and Bereiter (2006) is an environment to support students in collaborative knowledge creation in many domains. The work of Linn and colleagues (e.g., Linn and Eylon, 2011) in the domain of science education is an eminent example of an ICT-enhanced learning environment in the science domain. The Web Integrated Science Environment (https://wise.berkeley.edu/) focuses on concept learning through inquiry and collaboration. De Jong and colleagues (e.g., De Jong, 2006) studied the potential of simulations for inquiry learning in science education. They found that scaffolds and cognitive tools are needed to support the acquisition of the generation of hypothesis, prediction, and the exploration of models. These examples show what key characteristics for ICT-enhanced learning environments are: the provision of an infrastructure for collaboration between students and between students and their teacher, and scaffolds to facilitate collaboration, knowledge building, and metacognition (Voogt and Knezek, 2008). These classic studies are well documented and have demonstrated the added value of ICT for teaching and learning. However, despite their long history, they did not become part of main stream education.

Current Use of ICT in Schools

Recent large-scale international studies provide information about the current use of ICT in educational practices in different countries around the world. Data from the Second International technology in Education Study in 2006 from teachers teaching grade 8 (14 years of age) science classes showed that about 30% of the teachers did not use ICT, while about 20% use ICT on a weekly basis (Voogt, 2009). However, differences between countries are large. In Canada, Chile, and Hong Kong about 40% of the science teachers use ICT on a weekly basis, while in South Africa, Japan, and the Slovak Republic less than 10% use ICT so frequently. Results from the 2007 study of Trends in Mathematics and Science Study (TIMSS) (Mullis et al., 2008) showed that computers are hardly used in more than half of the lessons in eighth grade mathematics education (the international average is about 3% for the countries participating in TIMSS), despite the availability of computers. In fourth grade (9 years of age) mathematics, the use of ICT is also infrequent. Most use in fourth grade mathematics classes is for drill and practice (international average is 6% for use in half of the lessons or more),with exceptions for The Netherlands (30%) and Scotland (20%). The 2006 study of Progress in International Reading Literacy Study reported that about 30% of the fourth grade students had used computers to support reading instruction, and 39% of the students had read stories or other texts on the computer (Mullis et al., 2007). Also here the differences between countries are large. While in Austria, Singapore, Norway, and Hong Kong 60% or more of the students were supported by computers during reading instruction, in Georgia, Iran, Morocco, Moldova, and Macedonia 5% or less of the students were supported by computers. The picture that the results of these international monitoring studies tell us is that the immersion of computers in the curriculum is still scarce and shows a big gap between the expectations one might have about the added value of ICT for teaching and learning as found in research (see previous section) and the reality of the classroom.

Integration in Educational Practice

There is a variety of obstacles that hinder the use of ICT in educational practice. A major problem is the lack of communication and leadership between macro, meso, and micro levels in education. This lack of communication results in limited access to ICT in schools, problems with tuning ICT-enhanced learning environments with the curriculum, a lack of a shared vision at the school level and insufficient ICT integration competencies of teachers (Voogt and Knezek, 2008). Curriculum-related problems impeding the integration of ICT deal with the way ICT is present in national curriculum standards and student textbooks. Although many student textbooks nowadays have integrated software packages or links to Web sites, more advanced uses of ICT are often lacking. In addition advanced use of ICT requires more time than available in lessons periods of 45 min, in which many schools have organized their curriculum. These obstacles also refer to a broader discussion about curriculum goals, content, structure, and the role of ICT in a curriculum that matches requirements of the knowledge society (Voogt and Pareja Roblin, 2012). Research from Knezek and Christensen (2008) has shown that a teacher's use of ICT is affected by will (attitudes toward IT), skill (ICT competencies), and access to ICT tools. However, ICT competency cannot be not limited to basic ICT knowledge and skills. Meaningful use of ICT in education, however, requires teachers to develop knowledge and skills that enable them to integrate ICT with a suitable pedagogical approach for teaching specific subject matter in a certain context. Koehler and Mishra (2008) introduced the term Technological Pedagogical Content Knowledge to emphasize the importance of interaction between subject domain knowledge, pedagogical knowledge, and technological knowledge as a conceptual framework for describing the kind of competencies teachers need to integrate various ICT applications in their educational practice.

Determining Effects on Students

The ever-changing technology makes research into ICT in education difficult, complex, and challenging. This is particularly true for studying the impact for ICT on student learning.

The high expectations about the potential of ICT for student learning could not easily be confirmed by convincing evidence from research. The early meta-analyses from the 1980s reported moderate effects of computer-based treatments, compared to other instructional methods on student achievement, but many studies also did not find positive effects of ICT for students learning. Most convincing evidence for the effects of ICT is related to student learning in language arts, but the evidence with regard to student learning in math and science education seems less convincing (Voogt and Knezek, 2008). It appeared difficult to define the kinds of effects that were expected from ICT use, as well as ways to measure these effects. Initially, it was expected that ICT could enhance student achievement in traditional learning goals, but advanced uses of ICT aim to contribute to learning goals that are considered relevant in the knowledge society, such as higher order cognitive skills and so-called twenty-first century skills. These goals align with constructivist theories of learning. However, such learning goals require other assessment methods, such as formative and performance assessment (Voogt and Pareja Roblin, 2012).

Researching IT in Education

While early research on IT in education focused on the effect ICT applications may have on student learning, it was gradually realized that we also need to understand the way ICT is implemented in education to be able to assess whether and how ICT impacts student learning in the reality of the classroom. Compared to the early days of computing in education, research on ICT in education has become diversified and complex. While the teacher is considered to be a determining factor in the use of ICT in educational practice, much research is directed toward knowledge on factors that hinder or contribute to teachers' uptake of ICT, or to strategies that foster (pre- and in-service) teacher learning on ICT integration. To study the impact of ICT policy on the implementation of ICT in education, large-scale national and international monitoring studies are conducted, including international case study research to better understand the characteristics of innovative pedagogical uses of ICT. Next to research focusing on the implementation of ICT in education, technological changes result in new representations of knowledge (e.g., simulations, haptics, games) and new ways to construct knowledge (e.g., Web 2.0 environments). These technological developments require new and interdisciplinary approaches to study ICT in education.

See also: Computer Games and Academic Achievement; Lifelong Learning and Its Support with New Technologies; School Choice; Social Media.

Bibliography

Dede, Chris, 2000. Emerging influences of information technology on school curriculum. Journal of Curriculum Studies 32, 281–303.

Dede, Chris, 2008. Theoretical perspectives influencing the use of information technology in teaching and learning. In: Voogt, Joke, Knezek, Gerald (Eds.), International Handbook of Information Technology in Primary and Secondary Education. Springer, New York, pp. 43–62.

De Jong, Ton, 2006. Technological advances in inquiry learning. Science 312, 532–533.

Knezek, Gerald, Christensen, Rhonda, 2008. The importance of information technology attitudes and competencies in primary and secondary education. In: Voogt, Joke, Knezek, Gerald (Eds.), International Handbook of Information Technology in Primary and Secondary Education. Springer, New York, pp. 321–331.

Koehler, Matthew, Mishra, Punya, 2008. Introducing TPCK. In: AACTE Committee on Innovation and Technology (Ed.), Handbook of Technological Pedagogical Content Knowledge (TPCK) for Educators. Routledge, New York, pp. 3–29.

Linn, Marcia, Eylon, Bat-Sheva, 2011. Science Learning and Instruction: Taking Advantage of Technology to Promote Knowledge Integration. Routledge, New York.

Mullis, Ina, Martin, Michael, Foy, Pierre, 2008. TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.

Mullis, Ina, Martin, Michael O., Kennedy, Ann M., Foy, Pierre, 2007. IEA's Progress in International Reading Literacy Study in Primary School in 40 Countries. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.

Rutten, Nico, Van Joolingen, Wouter, Van der Veen, Jan, 2012. The learning effects of computer simulations in science education. Computers & Education 58, 136–153.

Scardamalia, Marlene, Bereiter, Carl, 2006. Knowledge building: theory, pedagogy, and technology. In: Sawyer, Keith (Ed.), Cambridge Handbook of the Learning Sciences. Cambridge University Press, New York, pp. 97–118.

The Cognition and Technology Group at Vanderbilt, 1997. The Jasper Project. Lessons in Curriculum, Instruction, Assessment and Professional Development. Lawrence Erlbaum, Mahwah, NJ.

van 't Hooft, Mark, 2008. Personal, mobile, connected: the future of learning. In: Voogt, Joke, Knezek, Gerald (Eds.), International Handbook of Information Technology in Primary and Secondary Education. Springer, New York, pp. 873–881.

Voogt, Joke, Knezek, Gerald (Eds.), 2008. International Handbook of Information Technology in Primary and Secondary Education. Springer, New York.

Voogt, Joke, 2009. How different are ICT-supported pedagogical practices from extensive and non-extensive ICT-using science teachers? Education and Information Technologies 14, 325–343.

Voogt, Joke, Roblin, Natalie Pareja, 2012. Teaching and learning in the 21st century. A comparative analysis of international frameworks. Journal of Curriculum Studies 44 (3), 299–321.

Relevant Websites

http://www.haptel.kcl.ac.uk – hapTEL.
http://isites.harvard.edu/icb/icb.do?keyword=harp – HARP.
http://muve.gse.harvard.edu/rivercityproject/ – River City.
https://wise.berkeley.edu/ – WISE.

Computer Games and Academic Achievement

Christopher J Ferguson, Stetson University, DeLand, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Research on the influence of video games on academic achievement has been controversial and inconsistent. Scholars have been examining the potential for both positive and negative influence of video games on academic achievement. As of yet, such research remains limited, controversial, and sometimes limited by methodological problems. Pathological gaming appears to be problematic for about 1–3% of children, although controversies remain on how to define and measure problematic gaming. Increasing attention has also been put into adapting video games for academic purposes. Initial evidence is limited but promising, although it is cautioned that scholars should remain conservative when discussing both positive and negative influences of video games on attention.

Since their advent in the 1970s, video games have been a source of considerable controversy and debate. Much of this debate has focused on the issue of violent content and whether such content increases aggressiveness (see Coyne et al., 2011; Ferguson, 2010, for competing views). Despite decades of research, debates on this issue have only intensified in recent years rather than being resolved. Although perhaps not as intense, debate has also focused on the issue of whether video games are associated with changes in academic achievement, either positive or negative (see Ferguson, 2011a; Gentile et al., 2012, for competing opinions). At present, debate continues regarding the potential of video games to positively, negatively, or not at all affect academic performance. The current article will begin by considering possible mechanisms for video game influence on attention, before turning to potential negative and positive effects in turn, including the issue of pathological gaming or gaming addiction.

Proposed Mechanisms by Which Video Game Might Influence Academic Achievement

One issue which has been recently raised in relation to the debate over video games and aggression is that of the *petition principia* or 'begging the question' logical fallacy in which a particular conclusion is assumed as the premise for conducting inquiry (Bennerstedt et al., 2012). This issue has also been defined in light of *moral panic theory* in which presumed harms about video games drove much of the research (which particularly conducted by investigators highly invested in the 'harm' view and sometimes funded by antimedia advocacy groups); such research claimed support for the initial premise, sometimes despite the actual data (Ferguson, 2010). Particularly in light of the recent Brown vs EMA decision (2011) as well as government reviews by Australia (2010) and Sweden (2011) which were critical of the violent game/aggression research, it is imperative that scholars remain aware of this 'tail wags dog' problem in other realms of video game research. In another way, the issue has emerged that scholars and activists have been quick to blame video games for all manner of social ills, often on the basis of weak and methodologically unsound data.

When it comes to video game effects on academic achievement the most obvious model for such effects is the "zero-end sum game." That is to say, any time spent on video games is time taken away from participation in other activities, particularly academic ones. Particularly given that the intrinsic appeal of academic work is certainly less than for video games, this approach makes a considerable amount of sense. Indeed most scholars would agree that moderation in the use of video games (as in all things) is a key feature of successful use of them. Research supports that moderate users of video games tend to be psychologically more healthier than either nonusers or extreme users (Allahverdipour et al., 2010; Kutner and Olson, 2008). Perhaps not surprisingly, this effect appears to hold for academic achievement as well (Ventura et al., 2012). This position can be taken to an extreme, however, in asserting that all or most of a child's free time should be focused on academics, and there is little 'value added' to allowing teens any time to engage in 'frivolous' pursuits.

A somewhat more extreme theoretical model is the 'brain damage' model in which the use of video games is thought to directly cause damage to the brains of players (e.g., Sax, 2007). Rather than merely asserting that video games distract from more worthy pursuits, this perspective asserts that the simple act of playing video games causes biochemical and physical alterations to brain structure, which are harmful. Proponents of this perspective tend to use the more extreme and frightening language of 'brain damage' or 'rewiring the brain' and this perspective is particularly popular among anti-video game advocates, who sometimes have funded some of this research (e.g., Center for Successful Parenting, 2011). Some scholars have claimed that video games, like other entertainment media, may induce attention problems by creating expectations for the brain to process smaller bits of information than are typical in traditional academic settings (e.g., Gentile et al., 2012), although this perspective has also been controversial (Ferguson, 2011a). Nonetheless this approach is less extreme than the general 'brain damage' perspective. The notion that video games cause damage to the brain was largely discredited by the US Supreme Court during the Brown vs EMA case (2011) and is arguably more the product of moral panic than good science (Ferguson, 2010).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92138-X

Potential Negative Influences on Academic Achievement

Despite the various concerns about video games on academic achievement, relatively little literature has investigated this issue (Blumberg, 2011). That what exists has generally been mixed.

Theories of how video games negatively influence attention take two forms. The first is the 'zero-end sum game' approach in which video game use simply is time taken away from cognitive tasks that may foster academic achievement. From this line of thought, some studies suggest that children who engage in problematic gaming behaviors have reduced academic functioning. For instance, Chiu et al. (2004) found that Chinese children with problematic gaming behaviors tended to have reduced academic performance. However, not all studies have found a relation between problematic gaming and academic performance (e.g., Desai et al., 2010). Nonetheless problematic gaming may remain an important risk factor for low academic achievement.

A second theoretical approach has focused on the possibility that the fast action and clip of video games (and television) may make it difficult for children accustomed to such stimulation to slow down and pay attention to less stimulating academic material (Swing et al., 2010). In essence this approach suggests that the brains of young people become accustomed to fast-paced stimuli and are thus quickly bored or turned off by slower paced stimuli. Swing et al. (2010) propose this in examining video game effects on academics, although they note that the research on television had been mixed. Although this model is generally used to imply that video games cause harm in this regard, it has been reversed by some advocates of video games (e.g., McGonigal, 2011) to imply that the problem lies with the poor and static quality of academic learning rather than video games. In other words video games should not be blamed for being stimulating; academics are at fault for not being stimulating enough. This approach is the 'gamification' approach which advocates changing real-life activities such as work or school to greater resemble the reward structure of video games so as to increase achievement.

In one of the first studies of this phenomenon, Swing et al. (2010) conducted a short-term prospective study of video game and television use and attention and academic problems in a sample of school-age children. The authors concluded that time spent playing video games contributed to attention and academic problems a year later. However, this study has been criticized on several fronts (Ferguson, 2011a). First, the authors did not use well-validated outcome measures. It has now been established that the use of poor and unstandardized measures in video game research is associated with spurious effects (Ferguson and Kilburn, 2009). Thus the potential for a type I error from this study is very high. Furthermore, the effects found, while statistically significant, were very weak, below effect size thresholds for 'trivial' effects.

Ferguson (2011a) in a correlational study of youth found that, with other factors well controlled and using well-validated outcomes, measures such as the child behavior checklist, video game exposure, either in times playing or in violent exposure, were not associated with attention problems or reduced academic performance. Gentile et al. (2012), in turn, suggested that these null results might have been due to inclusion of too many covariates, such as antisocial personality. Gentile et al. (2012), for instance, suggested that there was little theoretical reason to control for antisocial personality, neglecting to note that antisocial personality has long been linked with academic problems and that, indeed, it was a significant predictor in Ferguson (2011a). Furthermore Ferguson, 2011a, had none of the multicollinearity problems that would give credence to Gentile et al.'s comments and reducing the analysis to a small set of basic covariates did not change the outcome.

In a more recent study examining college students rather than children, Ventura et al. (2012) found a complex relationship between personality, video gaming, and academic performance. Students who tended to play moderate amounts of games highly enjoyed tended to fare best in terms of grade point average (GPA). As for specific genres, preference for social media games (i.e., Farmville) and first person shooters were weakly ($r = -0.13$ and -0.14, respectively) related to reduced GPA, but not for seven other genres of games. Also these last results are for *preference for genre* not time spent playing. Overall, results did not support a link between time spent playing video games and academic problems.

Thus debate on this issue remains. There is no clear answer to the question of whether video games contribute in any way to academic problems. However, it is worth noting that none of these studies suggests a moderate or strong effect on academic performance. At worst the influence of video games on academic performance is quite small.

One possibility is that by focusing on time spent gaming, researchers may have failed to distinguish between healthy gaming and pathological gaming, or gaming behaviors which interfere with other life responsibilities. Thus, it may be that although video game use in general is a poor predictor of negative outcomes, pathological gaming may nonetheless be a predictor of negative outcomes related to academic achievement (Ferguson et al., 2011).

Pathological Gaming

Although a direct link between video game use and academic achievement by and large difficulties has sometimes been investigated, as noted above, by large research, which finds such links are weak and inconsistent. This has lead to some scholars considering that it may be more fruitful to examine processes by which some gamers may develop academic problems whereas others may not. In this sense, video games may become involved with existing mental health problems, whether exacerbating them, or simply becoming symptomatic of these underlying problems. For some individuals, video game play may become obsessive, interfering with other life responsibilities and activities including academic work.

This phenomenon is sometimes called pathological gaming and sometimes video game addiction (pathological gaming will henceforth be used). Although pathological gaming has got the attention of scholars over the last decade, much debate and discussion remain regarding how to define it, how to measure it, how prevalent it is, and whether it is a unique phenomenon

to video gaming, or merely symptomatic of underlying mental health problems. As such, at present, pathological gaming is not slated for inclusion in the fifth edition of the *Diagnostic and Statistical Manual of the American Psychiatric Association*.

Because of difficulties in understanding how best to define and measure pathological gaming, prevalence estimates have ranged widely. Gentile (2009), for instance, suggested that just over 8% of youth exhibit signs of pathological gaming. However, as this sample was Internet based, it is likely to be spuriously high as heavy Internet users may have been more likely to both respond to the survey and exhibit signs of pathological gaming. Further Gentile uses a measurement approach which is based on symptoms of pathological gambling behaviors. A number of studies have used this approach in which it is assumed that symptoms of pathological gambling are easily transferable (with minor wording changes) to pathological gaming. These might include items such as 'I think about gaming when I am not playing' or 'I use gaming to relax' or 'I have gotten into arguments with others related to my gaming.' Thinking about gambling when not gambling, using gambling to relax, or getting into an argument with one's spouse over considerable financial losses due to gambling are obviously problematic. But it is less clear that these same questions transferred to gaming behaviors are problematic. Thinking about one's hobbies (including gaming) when not involved in them, using hobbies (including gaming) to relax, are certainly normative. Given that kids and parents routinely argue over time spent on hobbies or entertainment, it is also unclear that arguments in this case are indicative of problematic behavior. As such attempts to directly build pathological gaming measurements on items cribbed from pathological gambling have been criticized by some scholars as flawed (Ferguson et al., 2011). Specifically, in a meta-analysis, Ferguson et al. found that studies relying on such measurement tools tended to have spuriously high prevalence estimates and correspondingly weak associations with expected mental health and academic outcomes. In other words, attempts to use the 'pathological gambling analogy' vastly overidentify normal individuals as pathological.

Other scholars have focused on the degree to which symptoms interfere with other functions. Example items of this type might include, 'I have sometimes continued gaming so long I missed work or school' or 'I have difficulty balancing video gaming time with studying.' Such items tap into the degree to which gaming behaviors interfere with work, school, relationships, and other behaviors such as exercise. Generally these studies find much lower prevalence rates and these items tend to be better associated with mental health problems in the real world (Ferguson et al., 2011). For instance, Desai et al. (2010), examining children in the United States, found that video gaming in general was not particularly related to negative outcomes involving aggression, mental health, or poor academic performance. However, a subset of children, approximately 3–4%, exhibited problematic gaming behaviors and were prone to more severe behavioral and academic problems. Van Rooij et al. (2011), using a sample of Dutch children, found the prevalence rate to be closer to 1.5%.

From this it is clear that how one decides to measure and define pathological gaming can have dramatic influences on prevalence estimates. This is particularly an issue where video games have been subject to moral panics (Ferguson, 2010) in which the fear of video games may propel exaggerated claims of potential harm. This potential for moral panics has been raised as a potential concern for pathological gaming (Barnett and Coulson, 2010). Advocates (including some scholars) who have invested in anti-video game views can be expected to selectively report prevalence estimates at the higher end of the scale, even where these may be less valid. Given that scholars still actively debate the meaning of pathological gaming (Griffiths, 2008; Wood, 2008) it is important for scholars to tread carefully in this area.

Related to this issue is whether pathological gaming is a unique condition or merely symptomatic of underlying mental health conditions. For instance, some individuals may essentially have 'addictive personalities' in which any pleasurable activity may be overdone to the detriment of less pleasant but pragmatically useful activities such as work and school. On this issue, not unexpectedly, much debate remains. Desai et al. (2010) suggested that it is more likely that pathological gaming is a symptom of underlying mental health problems rather than a unique condition brought on by video games themselves. This view was supported by the meta-analysis of Ferguson et al. (2011). However, other scholars (Gentile et al., 2012) suggest that effects may be bidirectional. At this juncture relatively few scholars appear to be suggesting that pathological gaming is likely to result in individuals who are otherwise psychologically healthy. The debate appears now to focus on whether video games are *particularly* prone to addictiveness relative to other pleasurable activities. Some scholars claim, for instance, that video games use an intermittent reinforcement schedule which makes them prone to addictiveness (Strickland and Doan, 2012). However, it is important to note that when media are at the focus of moral panic, it is quite common for advocates to suggest reasons why new media might be *particularly* bad. For instance, with video games it was sometimes suggested that their interactive nature might make gamers particularly prone to imitating the violence. This suggestion continued to be repeated even after having been discredited, turning it into a form of scientific urban legend (Ferguson, 2011b). Thus, once again, we must be careful to distinguish plausible but advocacy-driven rationale from scientific fact. At present, although there is considerable research linking pathological gaming with negative outcomes related to academic performance, uncertainty remains regarding causal directions (Kuss and Griffiths, 2011).

As a further issue, some scholars have identified the issue that it is important to distinguish pathological gaming from healthy engagement (Charlton and Danforth, 2007). That is to say some players are *engaged* and may play for fairly long hours without their game play interfering with other life responsibilities. This distinction between engagement and pathological gaming may help to explain why hours spent gaming is typically a poor predictor of academic performance.

As such the issue of pathological gaming remains controversial on many fronts. Most scholars in the field agree that some form of pathological gaming probably exists. But considerable disagreement about how best to define and measure it means considerable uncertainty will persist for the near future.

Potential Positive Uses of Video Games in Academic Achievement

As might be expected, and certainly as predicted under moral panic theory, most of the attention on video games and academic performance has focused on the potential for negative effects. However, several avenues of research have examined whether video games may have a positive effect on academic achievement as well. Some scholars have investigated the use of both educational and commercial games in academic settings. Given the popularity of video games, teachers and scholars have expressed considerable interest regarding the potential use of video games in academic settings. Some scholars have suggested that video games may prove to be socially and academically revolutionary (McGonigal, 2011) although it should be cautioned that just as fears of video games were prone to hyperbole, the same may be true for hopes.

Considerable attention has focused on 'serious' games, that is, games which are specifically designed with a particular education, social, or behavioral goal in mind rather than being simply commercial and entertaining (Annetta, 2010). As Annetta notes, such games must strive to engage students given that they compete with commercially available games for student attention. Such games function by providing an immersive and interactive environment for students to enjoy in which useful educational information may also be embedded. Such games have the advantage over traditional education by holding student attention on enjoyable activities while providing education.

This is particularly the case where recent discussion among scholars have suggested that educational settings may have to shift how they teach in order to fully prepare students for an increasingly technological world (Anderson, 2012). In other words, the next generation of young workers will increasingly be expected to be proficient with technology and those who are prepared will have an advantage over those who are not. Much of the attention has focused on fields related to science, technology, engineering, and math which tend to be relatively low enrollment among US students. Some initial work has indicated that video games designed to teach science material provide advantages over traditional instruction while also preparing students for the use of modern technology (Annetta et al., 2010).

The game *Immune Attack* (Howell, 2005) provides one example of an educational serious game. Immune Attack provides simulations of the human immune system in which the player must use a variety of weapons based on immune functions and medicine to fight a host of invaders and threats. Produced by the Federation of American Scientists, Immune Attack is currently being evaluated for its successfulness as a learning tool. Other scholarship has suggested that serious games may be positive for educational purposes in fields such as history (Watson et al., 2011), urban planning (Thomas and Hollander, 2010), engineering (Coller and Scott, 2009), and health and physical education (Papastergiou, 2009). Nonetheless our understanding of how video games may contribute to or assist in a full paradigmatic and pedagogical shift to the use of new technologies in education remains poorly understood. Serious games face particular challenges in retaining student engagement when students may be expecting production values similar to those found in commercial games, despite that serious games typically must endure a slower and longer cycle of production and use (Boyan and Sherry, 2011).

Attention has focused as well on the applicability of commercial games as positive developers of academic achievement. Although commercial games are not typically intended as educational tools, some either directly or indirectly include content that may foster academic achievement. For instance, games that focus on historical strategy such as *Europa Universalis* or *Civilization* may either directly increase knowledge of history related topics, or may increase interest in these topics. For instance, Moshirnia and Israel (2010) found that a slightly modified version of *Civilization IV* was effective in imparting historical knowledge in students.

One scholar who has been the most active in this area is Constance Steinkuehler who, at the time of this writing, serves as the US White House's expert on video games. Her research has indicated the commercial games such as *World of Warcraft* promote literacy and verbal learning through participation in verbal activities in the game as well as part of discussion and message boards about the game (see Black and Steinkuehler, 2009). Thus, although probably never the intent of the video game producers, the interactivity and community built up around the game, in this case a massively multiplayer online game, fosters the kind of engagement and dialogue that promotes verbal learning.

Nonetheless, research on the applicability of video games, whether commercial or serious, remains relatively nascent. Just as we have learned we must be careful about exaggerating negative effects, so too we must be careful in overstating positive effects. For instance, in the realm of visuospatial cognition, a number of studies have suggested that video games might promote this type of learning (see Dye et al., 2009, for a review). However, other scholars have suggested that the visuospatial cognition literature may suffer from many of the same flaws as the negative effects literature (Boot et al., 2011) and not all empirical studies have supported the existence of positive effects (Valadez and Ferguson, 2012). Thus, as with the negative effects literature, scholars are well advised to remain conservative in their language when discussing the potential for positive video game effects.

Concluding Statements

The issue of whether video games contribute to academic success, either positively or negatively, remains a controversial one. There have been competing pools of research to suggest that video games may either enhance or impinge upon academic achievement. None of these research is beyond debate. It is possible, as often happens in the early stages of research in a new field, that biases tended to favor 'effects' hypotheses in both directions. That is to say that it is possible that both the benefits and risks of video games for academic achievement may have been exaggerated. This is not to say that no effect exists, rather that they may be more subtle and idiosyncratic than sometimes implied in the scholarly literature. Whether video games are beneficial or harmful may depend as much on the individual child. Thus, attempting to formulate

'one size fits all' recommendations to parents or policy makers may be doomed to failure. There is value in providing information to parents of course, but ultimately they may be the best arbiters to what is most beneficial to their individual children.

It is important that we recognize that during 'video game epoch,' when video games soared in popularity, standardized achievement test scores of American children have remained largely static. The video game epoch has ushered in neither a wave of intellectually deficient and academically challenged children, nor a wave of unusually proficient children. This does not mean that the impact of video games on academics is necessarily null, however. As the Pew Research Center has noted (2012), video games are part of an increasingly technological period in society, and mastering this technology is going to be necessary for future success. It remains unclear whether video games are going to be a panacea for society as some have suggested (McGonigal, 2011). It is nice to think that this might be the case. But given the degree to which new technology tends to elicit extreme predictions of both dire harm and great success, it may take time before we fully come to understand the influence of video games on our society. In the meantime, as emphasized earlier, the research on both positive and negative effects of video games on academic achievement remains controversial and scholars are recommended to remain as conservative as possible when discussing video game effects to policy makers and the general public.

See also: Evolutionary Approach to Education Science; Human–Computer Interfaces; Lifelong Learning and Its Support with New Technologies; Media Effects on Children; Media and Child Development; Tangible User Interfaces in Learning and Education.

Bibliography

Allahverdipour, H., Bazargan, M., Farhadinasab, A., Moeini, B., May 27, 2010. Correlates of video games playing among adolescents in an Islamic country. BMC Public Health 10, 286.

Anderson, J., 2012. Millennials Will Benefit and Suffer Due to Their Hyperconnected Lives. Pew Research Center. http://www.pewinternet.org/~/media//Files/Reports/2012/PIP_Future_of_Internet_2012_Young_brains_PDF.pdf (accessed 04.04.12.).

Annetta, L.A., 2010. The "I's" have it: a framework for serious educational game design. Review of General Psychology 14 (2), 105–112. http://dx.doi.org/10.1037/a0018985.

Annetta, L.A., Cheng, M., Holmes, S., 2010. Assessing twenty-first century skills through a teacher created video game for high school biology students. Research in Science & Technological Education 28 (2), 101–114. http://dx.doi.org/10.1080/02635141003748358.

Australian Government, Attorney General's Department, 2010. Literature Review on the Impact of Playing Violent Video Games on Aggression. Commonwealth of Australia.

Barnett, J., Coulson, M., 2010. Virtually real: a psychological perspective on massively multiplayer online games. Review of General Psychology 14 (2), 167–179. http://dx.doi.org/10.1037/a0019442.

Bennerstedt, U., Ivarsson, J., Linderoth, J., 2012. How gamers manage aggression: situating skills in collaborative computer games. Computer Supported Collaborative Learning 7, 43–61.

Black, R.W., Steinkuehler, C., 2009. Literacy in virtual worlds. In: Christenbury, L., Bomer, R., Smargorinsky, P., Christenbury, L., Bomer, R., Smargorinsky, P. (Eds.), Handbook of Adolescent Literacy Research. Guilford Publications, New York, pp. 271–286.

Blumberg, F.C., 2011. Ramifications of video game play for academic learning and cognitive skill acquisition: introduction. Child Development Perspectives 5 (2), 73–74. http://dx.doi.org/10.1111/j.1750-8606.2011.00158.x.

Boot, W., Blakely, D., Simons, D., 2011. Do action video games improve perception and cognition? Frontiers in Psychology 2, 226. http://dx.doi.org/10.3389/fpsyg.2011.00226.

Boyan, A., Sherry, J.L., 2011. The challenge in creating games for education: aligning mental models with game models. Child Development Perspectives 5 (2), 82–87. http://dx.doi.org/10.1111/j.1750-8606.2011.00160.x.

Brown v EMA, 2011. Retrieved 07.01.11 from: http://www.supremecourt.gov/opinions/10pdf/08-1448.pdf.

Center for Successful Parenting, 2011. The Center for Successful Parenting. http://www.sosparents.org/ (accessed 07.06.11.).

Charlton, J., Danforth, I., 2007. Distinguishing addiction and high engagement in the context of online game playing. Computers in Human Behavior. 23 (3), 1531–1548.

Chiu, S.I., Lee, J., Huang, D., 2004. Video game addiction in children and Teenagers in Taiwan. Cyberpsychology & Behavior 7 (5), 571–581.

Coller, B.D., Scott, M.J., 2009. Effectiveness of using a video game to teach a course in mechanical engineering. Computers & Education 53 (3), 900–912. http://dx.doi.org/10.1016/j.compedu.2009.05.012.

Coyne, S.M., Padilla-Walker, L.M., Stockdale, L., Day, R.D., 2011. Game on… girls: associations between co-playing video games and adolescent behavioral and family outcomes. Journal of Adolescent Health 49 (2), 160–165. http://dx.doi.org/10.1016/j.jadohealth.2010.11.249.

Desai, R., Krishnan-Sarin, S., Cavallo, D., Potenza, M., 2010. Video-gaming among high school students: health correlates, gender differences, and problematic gaming. Pediatrics 126 (6), e1414–e1424.

Dye, M.G., Green, C., Bavelier, D., 2009. Increasing speed of processing with action video games. Current directions. Psychological Science 18 (6), 321–326. http://dx.doi.org/10.1111/j.1467-8721.2009.01660.x.

Ferguson, C.J., 2010. Blazing angels or resident evil? Can violent video games be a force for good? Review of General Psychology 14 (2), 68–81.

Ferguson, C.J., 2011a. The influence of television and video game use on attention and school problems: a multivariate analysis with other risk factors controlled. Journal of Psychiatric Research 45 (6), 808–813.

Ferguson, C.J., 2011b. 'Video games are like smoking and lung cancer' and other urban legends in the violent video game debate. Interface 11 (1). http://bcis.pacificu.edu/journal/article.php?id=761.

Ferguson, C.J., Kilburn, J., 2009. The public health risks of media violence: a meta-analytic review. Journal of Pediatrics 154 (5)), 759–763.

Ferguson, C.J., Coulson, M., Barnett, J., 2011. A meta-analysis of pathological gaming prevalence and comorbidity with mental health, academic and social problems. Journal of Psychiatric Research 45 (12), 1573–1578.

Gentile, D., 2009. Pathological video-game use among youth ages 8 to 18: a national study. Psychological Science 20 (5), 594–602. http://dx.doi.org/10.1111/j.1467-9280.2009.02340.x.

Gentile, D.A., Swing, E.L., Lim, C., Khoo, A., 2012. Video game playing, attention problems, and impulsiveness: evidence of bidirectional causality. Psychology of Popular Media Culture 1 (1), 62–70. http://dx.doi.org/10.1037/a0026969.

Griffiths, M., 2008. Videogame addiction: further thoughts and observations. International Journal of Mental Health and Addiction 6 (2), 182–185.

Howell, K.K., 2005. Immune attack: building biologically based immune system simulations for education and training. Annual Review of Cybertherapy And Telemedicine 3, 59–63.

Kuss, J.D., Griffiths, M.D., 2011. Internet gaming addiction: a systematic review of empirical research. International Journal of Mental Health Addiction. http://dx.doi.org/10.1007/s11469-011-9318-5.

Kutner, L., Olson, C., 2008. Grand Theft Childhood: The Surprising Truth about Violent Video Games and What Parents Can Do. Simon & Schuster, New York.

McGonigal, J., 2011. Reality Is Broken: Why Games Make Us Better and How They Can Change the World. Penguin Press, New York, NY, US.

Moshirnia, A., Israel, M., 2010. The educational efficacy of distinct information delivery systems in modified video games. Journal of Interactive Learning Research 21 (3), 383–405.

Papastergiou, M., 2009. Exploring the potential of computer and video games for health and physical education: a literature review. Computers & Education 53 (3), 603–622. http://dx.doi.org/10.1016/j.compedu.2009.04.001.

Sax, L., 2007. Boys Adrift: The Five Factors Driving the Growing Epidemic of Unmotivated Boys and Underachieving Young Men. Basic Books, New York.

Strickland, B., Doan, A., 2012. Hooked on Games. FEP International, Coralville, IA.

Swedish Media Council, 2011. Våldsamma datorspel och aggression – en översikt av forskningen 2000–2011. http://www.statensmedierad.se/Publikationer/Produkter/Valdsamma-datorspel-och-aggression/ (accessed 01.14.11.).

Swing, E.L., Gentile, D.A., Anderson, C.A., Walsh, D.A., 2010. Television and video game exposure and the development of attention problems. Pediatrics 126 (2), 214–221. http://dx.doi.org/10.1542/peds.2009-1508.

Thomas, D., Hollander, J.B., 2010. The city at play: second life and the virtual urban planning studio. Learning, Media And Technology 35 (2), 227–242. http://dx.doi.org/10.1080/17439884.2010.494433.

Valadez, J.J., Ferguson, C.J., 2012. Just a game after all: violent video game exposure and time spent playing effects on hostile feelings, depression, and visuospatial cognition. Computers in Human Behavior 28, 608–616.

Van Rooij, A.J., Schoenmakers, T.M., Vermulst, A.A., van den Eijnden, R.M., van de Mheen, D., 2011. Online video game addiction: identification of addicted adolescent gamers. Addiction 106 (1), 205–212. http://dx.doi.org/10.1111/j.1360-0443.2010.03104.x.

Ventura, M., Shute, V., Kim, Y., 2012. Video gameplay, personality and academic performance. Computers & Education 58, 1260–1266.

Watson, W.R., Mong, C.J., Harris, C.A., 2011. A case study of the in-class use of a video game for teaching high school history. Computers & Education 56 (2), 466–474. http://dx.doi.org/10.1016/j.compedu.2010.09.007.

Wood, R., 2008. Problems with the concept of video game 'addiction': some case study examples. International Journal of Mental Health and Addiction 6 (2), 169–178.

Computer Mediated Communication

Yuhua (Jake) Liang, Chapman University, Orange, CA, USA
Joseph B Walther, Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Computer-mediated communication (CMC) involves sending messages through computer networks such as the Internet. Research on the social impacts of text-based CMC began with a cues-filtered-out perspective, highlighting the loss of social information due to the absence of face-to-face, nonverbal cues. More recent theories focus on how message senders exploit the CMC medium in order to create messages and relationships actively. A sender-receiver-message-channel-feedback model of communication reveals how CMC affects social dynamics.

Communication technologies have potentially altered the way human beings share information with one another. The ability to communicate by sending and receiving text information has been more readily available in recent history. Just as the Pony Express, the telegraph and the telephone enabled individuals to communicate across great distance, new communication technologies have again reinvented the available means of communication. Electronic communication technologies utilizing computers have refamiliarized our society with text-driven means of connecting with each other. Yet, despite the changes associated with historical text-based interpersonal media such as mail and phone, the rapid speed and other unique affordances of the Internet has sparked considerable interest in how *computer-mediated communication* (CMC) differs from traditional communication.

The earliest research on CMC focused on the effects of new technology related to task performance outcomes. In the 1980s and 1990s, CMC consisted of electronic email and computer conferencing, chats, and electronic bulletin board systems hosting online forums. Research addressed the effects of those text-based communication systems not only on tasks, but also how the social dynamics and effects of the technology on users affected tasks or group performance. Today, with a greater variety of social media and expanded roles for their application, communication technology now plays an integral role in how people interact with one another across myriad types of social relationships.

Communication technologies available today have increased the frequency, ubiquity, and impact of CMC. Individuals often work and interact with others via computers on a daily basis. People may access CMC through their mobile devices as well as computers. More complex and crowd-based systems exist today to provide new means of digital collaboration and online sharing. Social media technologies, such as Facebook and Twitter, provide highly accessible means of CMC by relating the content to the people we know. In addition to new enablement and accessibility, communication patterns today blend the use of offline and online channels across most of our personal and professional relationships (Ramirez and Zhang, 2007).

Despite the mass proliferation and integration of CMC into our lives today, the delineation between CMC and other forms of communication remains useful and necessary in order to understand how technology affects communication processes. Even though CMC now hosts more multimodal channels in terms of photos, videos, and immersive graphical environments, text-based communication remains as a mainstay of how individuals exchange messages. Research developed on the basis of older text-based forms of CMC should continue to apply. As we develop an understanding of how CMC can affect communicators we can posit new questions as newer communication technologies such as Web 2.0 continue to expand the range of CMC. With that in mind, this article reviews several theoretical perspectives on CMC by focusing on the effects of CMC on those who use it, following the perspective of a basic communication model.

CMC: A Communication Perspective

A basic model of communication aids in identifying the impact of CMC, the central components of which include senders, receivers, messages, feedback, and communication channel (see for review Walther, 1996). S*enders* initiate the exchange of information by articulating intentions into information to be conveyed to others. The *message* contains the information symbols representing the information to be exchanged. Senders deliver messages by using a *communication channel*, the vehicle that carries the message. Such vehicles include face-to-face communication, telephone, teleconference, Facebook posts, and electronic chats, among others. *Receivers* interpret the message and often provide a return message in the form of *feedback*. This feedback reverses the role of the senders and receivers as the cycle continues iteratively until the termination of the communication event.

In addition to the capabilities of the CMC channel, CMC also occurs across different contexts. For example, Facebook's social context alters how users may interpret and construct messages relative to email or other systems. When a Facebook user posts a status update, members of this person's social network may read the message in light of past interactions with that user. On the other hand, messages exchanged on a company or organization's office chat system likely result in a different interpretive framework. Employees may experience

 http://dx.doi.org/10.1016/B978-0-08-097086-8.95090-6

more reluctance in sharing social activities outside of the workplace on company websites as opposed to their friends and family via Facebook.

New communication technologies challenge existing theoretical perspectives. Early views of CMC focused on comparing the ability of the communication channel to those in face-to-face communication and/or other forms of CMC. Theoretical developments that followed offered a different view of CMC. Instead of focusing on the limiting constraints of the communication channel, these developments focused on how users can adapt to CMC to convey messages. We begin with an overview of the earlier theoretical view of CMC based on channel characteristics.

Cues-Filtered-Out Theories

Culnan and Markus (1987) first described a class of CMC theories as adopting a *cues-filtered-out* approach. The major contention in these theories maintains that CMC constraints the symbol systems that are available in the communication process, and thereby limits the meaning that users are able to exchange. Because CMC is primarily text-based, it lacks nonverbal cues. These theories assumed a relationship between the restricted cues that the channels offered and a commensurate restriction in senders' abilities to convey certain kinds of messages. There was less focus on the senders and receivers in terms of their potential adaptation of messages to alternative channels. Several particular theories typify the cues-filtered-out perspective.

Social Presence Theory

The first theory applied to CMC that reflected a cues-filtered-out perspective is social presence theory. Short et al. (1976) conceptualized that various communication systems differ in the availability of cue systems. Short et al. also claimed that the decrease of cues corresponded to a decrease in social presence between communicators. Although the theory originally focused on telecommunication systems, it was applied to CMC in the 1980s and 1990s.

The restriction of cue systems was suggested to impact the degree of social information a system could convey. Therefore, information about individuals' character and personality, and their interpersonal warmth and ease would be unlikely to be exchanged effectively without nonverbal cues, from a social presence perspective.

Media Richness Theory

Social information differs from more instrumental forms of information, for which certain other cues-filtered-out theories, media richness theory in particular (Daft and Lengel, 1984, 1986), suggest that "leaner" media (without nonverbal cues) may be more efficient. For instance, CMC may suffice when one person needs to tell another person the date for an event. On the other hand, if an individual wished to get to know another person's character or personality – characteristics which nonverbal cues might more efficiently convey – social presence and media richness theories suggest that CMC would denude the quality of information due to the lack of nonverbal cues to express socioemotional content.

Media richness theory focused on how organizations can utilize different types of communication channels to deal with decision-making contexts. Daft and Lengel (1984) attempted to provide a model for how different communication channels (e.g., memos, letters, chats, telephone conversations, and face-to-face interactions) rank in their abilities to reduce the equivocality managers experience during decision making. Equivocality refers to the ambiguity that arises from the presence of multiple, possibly conflicting interpretations. For example, a decision regarding whether a new policy should be adopted should evoke high equivocality. Such a context involves discussions among various managers and different interpretations regarding the consequences of the policy implementation.

These researchers contended that the ability of the communication channel to manage equivocality depended on information richness, which is comprised of (1) feedback speed or the rapidity of feedback, (2) multiple cues or the number of cues supported by the communication system, (3) language variety or the range of meaning that can be conveyed with language, and (4) personal focus or the ability of the communication to convey personal feelings and emotions (Daft et al., 1987). According to the media richness view, face-to-face communication would rank as the richest channel whereas other forms of mediated communication would be leaner (Daft et al., 1987). Face-to-face communication contains fast feedback, the greatest number of cues, the use of natural language, and the ability to adapt to individuals. On the other hand, CMC, such as chats and emails, would rank lower due to the lack of some richness criteria. It is regarded as a leaner group of channels.

Media richness theory argued that the best communication performance occurs as a function of matching the equivocality of the communication task to the richness of the communication channel. Specifically, high equivocality would require high information richness, and vice versa. For instance, managers attempting to make a policy decision require discussion and interpretation among multiple perspectives, thus face-to-face communication would be ideal. Alternatively, sharing accounting information or other objective data does not elicit a high degree of equivocality, meaning that a lean channel such as email would actually be more ideal in this context.

Empirical Findings

Both social presence theory and media richness theory provide a fairly intuitive understanding of CMC. The simplicity of treating different CMC categorically based on the absence of nonverbal cues is transparent and simple. Perhaps due to such an appeal, numerous investigations followed a cues-filtered-out approach in both organizational and interpersonal contexts. However, the empirical support has been generally mixed.

Social presence theory has received numerous critiques and challenges for the oversimplification explanation of CMC in a social context (Lea and Spears, 1992; Walther, 1992). In short, these critiques centered on challenging the assumptions and artifacts present in social presence work. Media richness

theory has also been challenged in from multiple angles. First, the original study did not directly test the theoretical proposition laid out by its authors. The empirical support used to substantiate the theory's formulation involved asking managers what communication channel they would select in a number of communication contexts that differed in equivocality. Daft et al. (1987) found a positive correlation between how well the managers selected the communication channel and the manager's actual performance evaluations. This correlation did not test the theoretical proposition that using various communication channels resulted in the best decision-making outcome, given situations that vary in equivocality. Dennis and Kinney (1998) experimentally tested media richness theory and failed to find supporting evidence. These researchers found that richness led to differences in the time that it would take to complete a task. However, richness did not affect decision quality. In short, the critiques and contradictory evidence limit the viability of the cues-filtered-out approach to understanding CMC.

Despite the critiques and problems associated with cues-filtered-out approaches, this view of CMC continues to perpetuate in recent studies of new media (see for review Walther, 2011). Adopting a cues-filtered-out approach often hinders the advances possible in the study of CMC. Fortunately, alternative theoretical views focus on how CMC integrates factors beyond those in the communication channel alone.

Social Identity Model of Deindividuation Effects

The social identity model of deindividuation effects, or *SIDE* model (Lea and Spears, 1992), shares some similarities with the other cues-filter-out theories. Specifically, SIDE also assumes that the absence of nonverbal cues limits the transmission of interpersonal information online. However, unlike other cues-filtered-out approaches, the SIDE offers a mechanism for online communication based on visual anonymity and social identification. Visual anonymity occurs when communicators utilize text-based communication where no visual information about each other is available. Under such a condition, communicators deindividuate and lose a sense of self-awareness. If they identify themselves with a salient social group, they will likely follow the behaviors of that social group.

The following scenario exemplifies a scenario for how SIDE may function in CMC. Assume that a college student reads and writes responses on a political forum. If the forum is text-based, the student should remain visually anonymous to others on the forum, as well as others to the student. The student is able to garner information about others on the forum and realizes that some of the forum participants are also students at the same university. If the student identifies with the in-group of students from the same university, he or she should orient toward the behavior of those from the same institution (e.g., evaluate them as more attractive, follow the recommendations advocated). On the other hand, if the student realizes that some of the other contributors on the same forum are from a rival institution, he or she may regard these people as out-group members. The student is likely to evaluate these out-group members as less attractive and be less likely to follow any recommendations advocated by those members.

The SIDE model's parsimonious explanation of visually anonymous CMC led to a wide following. The theory has even been applied to understand the effect of CMC in interpersonal contexts: In the case of visually anonymous romantic partners who may have met online, SIDE offers an interesting interpretation for their relational development. An application of the SIDE suggests that romantic communication and attraction toward partners result due to the social identification they feel toward salient group identities. According to SIDE, the attraction does not develop out of interpersonal processes but group identification processes alone (Lea and Spears, 1995).

Taken together, the cues-filtered out approaches to CMC provide a limited understanding of how people communicate. The primary assumption that text-based communication can only convey a certain quantity or form of meaning ignores any possibilities of relational development or transformative properties using CMC. In the case of SIDE, the prediction that social identification with group members would lead to adherence to group behavior also assumed that visual anonymity and text-based communication restrict or discourage users from conveying personal and interpersonal information. The next few of theoretical perspectives focus more on the interpersonal capabilities of CMC. In particular, these perspectives emphasize communicators' ability to utilize CMC in such a way as to provide meaning, to detect and foster personal impressions, and develop relational communication that approximates, or even exceeds, the level of relational regard that is typical in face-to-face communication.

Theories of Interpersonal Adaptation

A fundamental difference between cues-filtered-out approaches and theories described below can be understood from the communication model. Cues-filtered-out approaches focus on the channel primarily, with the exception of SIDE which also focuses on receivers' social identification. Instead, the theoretical frameworks of interpersonal adaptation and exploitation of media posit an understanding of CMC through all components of the communication model, raising in prominence the role of the sender and the sender's active creation of messages and channels. This approach to understanding CMC paints a picture of how communicators can take advantage of all communication elements to share meaning.

Social Information Processing Theory

The social information processing (SIP) theory (Walther, 1992) assumes that individuals aim to develop interpersonal impressions and relations with others. It proposes that when the communication channel limits the cues available to communicators, as is the case using text-based CMC, individuals will utilize the available cues in that channel to share their desired meaning. In particular, communicators primarily utilize the content of the written language (e.g., the choice and style of words) to encode and decode information from others.

In addition to CMC adaptation from a sender's standpoint, SIP supposes that the CMC channel alters the rate of social information exchange compared to face-to-face communication. In other words, to develop to the same level of relationship, communicators who use CMC would require more time than those in face-to-face communication. This rate difference originates from the diminished information available in text-based CMC as opposed to face-to-face communication. Even as senders encode messages by adapting to the cues and meaning available in CMC, the communication takes more time to accrue the same meaning as those in face-to-face communication. However, it is important to note that the rate difference does not equate to a superiority or inferiority of CMC when compared to the face-to-face counterpart. Instead, social information simply needs more time to accumulate and render effects on relationships in CMC.

The Hyperpersonal Model of CMC

The hyperpersonal model of CMC (Walther, 1996) takes a step beyond SIP theory. Specifically, the hyperpersonal model contends that CMC may aid in the development of interpersonal relationships that even surpass those in face-to-face communication. Such a process entails a series of theoretical propositions that relate to (1) how receivers may process CMC information, (2) how senders may encode messages, (3) characteristics of the communication channel, and (4) feedback. These propositions coalesce to describe a CMC process that involves all components of the communication model.

Receiver Effects

A receiver may experience exaggerated perceptions of the message sender in CMC. Since a receiver does not have the cues present in face-to-face communication, the receiver has to fill the void by creating impressions regarding the sender. When initial cues lead to positive first impression, impression development may then involve idealization, where the receiver bestows characteristics about the sender preferentially. This articulation originated from the notion of social identification to group identity in SIDE theory. Newer accounts of the hyperpersonal model suggests that individual stereotypes, not just those based on common social categories, may also activate heightened receiver impressions (Walther, 2006).

Sender Effects

With a text-based CMC in mind, communicators can selectively present aspects of themselves online. As a result, they can manage which aspects of themselves to present to others, with an underlying goal to create a favorable impression in the other communication partner. Of course, individuals can choose to disclose negative personal information; however, the lack of nonverbal cues in CMC actually minimizes the nonverbal behavior that might damage positive interpersonal impressions (e.g., poor eye contact, body postures) in face-to-face communication. As a result, CMC senders can more easily select aspect of themselves to present to their partners to elicit favorable responses.

Channel

Not only can senders selectively self-present, their ability to do so is enhanced by the communication channel. Several aspects of CMC facilitate message construction. For instance, in the case of asynchronous CMC, both the sender and the receiver can utilize time to construct the message and feedback. Unlike face-to-face communication where the feedback between communicators is more immediate, a CMC sender and a CMC receiver have time to write the message, edit that message, or even delete and rewrite the message. The use of time and editing leads to more positive and affectionate messages (see for review Walther, 2011).

Feedback

According to the former three components of the hyperpersonal model, the sender may selectively self-present to garner positive impressions, the receiver may idealize the sender, while both the sender and the receiver may capitalize on the CMC channel to construct favorable messages. Based on these enhancements, the feedback system created between the sender and the receiver should contribute to a continuous elevation of positive impression as they communicate with each other. In other words, the feedback system in CMC should enhance the interpersonal dynamics at play and intensify the effects iteratively over time.

A synthesis of all four hyperpersonal components illustrates a communication dynamic where the positive impressions of the CMC communicators intensify over time. The model points the possibility where such intensification may exceed those experienced in face-to-face communication. Both SIP theory and the hyperpersonal model challenge the limited views provided by the cues-filtered-out approaches to CMC. Moreover, the theories of adaptation of exploitation of media paint CMC as an interactional landscape where users adapt and appropriate all available means in the communication process to reach affinity goals. Although the domain of these theories resides primarily in the interpersonal context, the theories' application and interpretation suggest that communication should be understood from the communicator's standpoint rather than the technology or the channel.

Online Information Evaluation

Warranting

In addition to cyclical communication processes posited by theories of interpersonal adaptation and exploitation of media, some recent theoretical concepts examine how individuals may make sense of information they receive online. Online communication may involve the information individuals disclose to each other. Individuals often evaluate the validity of information about another person online, also known as warranting (Walther and Parks, 2002). The communication technology today provides quick access to information. Online users also have access to a lot of information about others. For example, online users may share a lot of personal information on their online dating profiles including interests, religious views, hobbies, among others. The warranting idea suggests that information presented online may lead to a suspicion of

possible distortions through selective self-presentation, such as the sender effect specified by the hyperpersonal model. For example, information on dating websites may show both distorted physical (weight/height) and psychological (hobbies) characteristics. The warranting concept argues that such type of distortion is less likely to occur when two communicators have access to each other's social network. Specifically, when one communication partner's social circle is available, one partner can hold the other responsible for the content of the self-disclosure and verify that content. Based on warranting, information that is more likely to relate to a communication partner's social network is less likely to be seen as distorted. Warranting is useful in predicting how communicators may judge the social information provided by others online.

Participatory Websites

The CMC available today has dramatically changed from the early text-based forms. Most online communication occurs via websites that feature some social information from other users. These types of participatory websites, known as Web 2.0, complicate the theoretical concepts for understanding the underlying communication dynamics at play. Moreover, these social websites juxtapose messages from various sources. An original author of some Web content may receive comments and evaluation from other users regarding that content. From a communication perspective, information regarding the sender, receiver, message, feedback, and context may blend together in a single webpage. The additional user-generated content may serve to modify how other readers may evaluate the content advocated by the original author, the original author as a source, others who comment on the source, as well as the content of the comments advocated by others.

This following example illustrates the complexity of understanding Web 2.0. Suppose that a product reviewer posts a positive product review on Amazon.com. Assuming that the review was meant to advocate to other readers for purchasing a particular product, other readers may compose comments on the review. If other readers write positive comments regarding the review, then those who read the original review and corresponding comments may regard the product as positive overall. However, if the review received negative comments that do not support what the review advocates, readers of the reviews and the comments that follow have to make sense of both positive and negative information. To complicate matter even more, Amazon.com reviewers and commenters may be rated in terms of how helpful they are. These types of information juxtapose against one another as the reader may review them on a single website.

Walther and Jang (2012) offered a conceptual foundation for defining the various sources available on Web 2.0 sites. They proposed four elements including proprietor content, user-generated content, deliberate aggregate user representations, and incidental aggregate user representations.

Proprietor Content

Proprietor content relates to the message composed by the primary author or the proprietor of a webpage. The proprietor has controls over message that he or she composes. The proprietor may edit that message (e.g., deleting a Facebook post). The message may include videos, text, and/or pictures. Furthermore, these forms of messages may operate in combination. A Facebook picture may accompany the written post in text. Examples of proprietor content include YouTube videos, Amazon.com reviews, Facebook profiles, among others.

User-Generated Content

User-generated content (UGC) relates to the messages that other users contribute. This content does not originate from the proprietor, but instead, from those who compose content in response to the proprietor content. UGC differentiates Web 2.0 from traditional websites. Unlike traditional websites (e.g., a webpage that does not allow users to write comments), Web 2.0 displays UGC alongside the proprietor content. Examples of UGC include comments responding to a Facebook post, comments on product reviews (comments written in response to an online news article).

Aggregated User Representations

Web 2.0 systems also allow users to leave evaluations of others' content. These evaluations may relate to proprietor content, as well as UGC. Web 2.0 systems often encourage users to rate or vote on the content produced by others. Then, the system aggregates the ratings for others to view. The aggregated ratings take the form of some statistics or figures. Examples of these representations include the number of stars, total votes, usefulness votes, helpfulness ratios, among others. To delineate among the different types of user representations, there are *deliberate and incidental* aggregate user representations.

Deliberate aggregate user representations include responses generated from directly requesting those responses from users. For instance, after displaying an opinion from a user, the Web 2.0 system may solicit readers to rate if that opinion was useful or not useful (e.g., 1 to 5 stars). An example of deliberate representation may include ratings of review helpfulness. Amazon.com displays the accumulated ratings of helpfulness in the form of a ratio (e.g., 40 out of 41 people found the review helpful).

In contrast, incidental aggregate user representations include information captured by Web 2.0 systems. In particular, these representations do not reflect behaviors that users would like to express, but instead, they are computed by the Web 2.0 system and displayed with other information. Examples of these representations include the number of Facebook friends, number of views on a YouTube video, the timestamp of a particular comment or review, among others. These representations offer a unique persuasive quality in that online users have no control over them. As a result, they may serve to aid certain aspects of self-presentations from a warranting perspective.

Participatory Websites and Juxtaposition among Sources

Participatory websites create unique theoretical challenges for CMC. Although the theoretical terms expressed previously aid in defining the various conceptual components involved in Web 2.0, they do not offer a set of theoretical

propositions regarding the relationships among the concepts. From a communication perspective, the model elements of the sender, receiver, message, context, and feedback also offer little explanatory utility as these components mix to serve multiple roles in Web 2.0. For instance, a message proprietor may respond to UGC and create a conversation between a proprietor and a commenter. At the same time, other users contribute to this dynamic through rating both the proprietor content and/or the UGC. When the Web 2.0 system displays all these contents simultaneously, how do other readers process that information and continue in the participatory process by adding or rating the existing content? What if more than one commenter is involved? This question typifies the theoretical challenge in the study of CMC as new technology enable online communicators to interact at such a convergent level across multiple information sources and ratings.

Summary and Future Direction

The current review began with an explanation of the communication perspective and elements in the communication model. With this perspective in mind, we reviewed a few of the theories involved in understanding CMC. The cues-filtered-out approaches assumed CMC's lack of nonverbal cues limited communication. On the other hand, theories of interpersonal adaption and exploitation of media posit a more interactive process that may match and/or exceed face to face. Finally, the conceptual framework for understanding participatory websites provides a glimpse of the complexity and challenge of theory development. As research incrementally increases our understanding of CMC, technology is also on the forefront of altering the landscape of digital interaction. Web 2.0 site continue to modify both interface as well as the type of information available to users.

See also: Audiences, Media; Information Society; Social Media.

Bibliography

Culnan, M.J., Markus, M.L., 1987. Information technologies. In: Jablin, F.M., Putnam, L.L., Roberts, K.H., Porter, L.W. (Eds.), Handbook of Organizational Communication: An Interdisciplinary Perspective. Sage, Newbury Park, CA, pp. 420–443.

Daft, R.L., Lengel, R.H., 1984. Information richness: a new approach to managerial behavior and organization design. In: Staw, B.M., Cummings, L.L. (Eds.), Research in Organizational Behavior, vol. 6. JAI Press, Greenwich, CT, pp. 191–233.

Daft, R.L., Lengel, R.H., 1986. Organizational information requirements, media richness and structural design. Management Science 32, 554–571.

Daft, R.L., Lengel, R.H., Trevino, L.K., 1987. Message equivocality, media selection, and manager performance: implications for information systems. Management Information Systems Quarterly 11, 355–368.

Dennis, A.R., Kinney, S.T., 1998. Testing media richness theory in the new media: the effects of cues, feedback, and task equivocality. Information Systems Research 9, 256–274.

Lea, M., Spears, R., 1992. Paralanguage and social perception in computer-mediated communication. Journal of Organizational Computing 2, 321–341.

Lea, M., Spears, R., 1995. Love at first byte? Building personal relationships over computer networks. In: Wood, J.T., Duck, S. (Eds.), Understudied Relationships: Off the Beaten Track. Sage, Thousand Oaks, CA, pp. 197–233.

Ramirez Jr., A., Zhang, S., 2007. When online meets offline: the effect of modality switching on relational communication. Communication Monographs 74, 287–310.

Short, J., Williams, E., Christie, B., 1976. The Social Psychology of Telecommunications. Wiley, London.

Walther, J.B., 1992. Interpersonal effects in computer-mediated interaction: a relational perceptive. Communication Research 19, 52–90.

Walther, J.B., 1996. Computer-mediated communication: impersonal, interpersonal, and hyperpersonal interaction. Communication Research 23, 3–43.

Walther, J.B., 2006. Nonverbal dynamics in computer-mediated communication or (and the net :('s with you:) and you :) alone. In: Manusov, V., Patterson, M.L. (Eds.), Handbook of Nonverbal Communication. Sage, Thousand Oaks, CA, pp. 461–479.

Walther, J.B., 2011. Theories of computer-mediated communication and interpersonal relations. In: Knapp, M.L., Daly, J.A. (Eds.), The Handbook of Interpersonal Communication, fourth ed. Sage, Thousand Oaks, CA, pp. 443–479.

Walther, J.B., Jang, J., 2012. Communication processes in participatory websites. Journal of Computer-Mediated Communication 18, 2–15. http://dx.doi.org/10.1111/j.1083-6101.2012.01592.x.

Walther, J.B., Parks, M.R., 2002. Cues filtered out, cues filtered in: computer-mediated communication and relationships. In: Knapp, M.L., Daly, J.A. (Eds.), The Handbook of Interpersonal Communication, third ed. Sage, Thousand Oaks, CA, pp. 529–563.

Walther, J.B., Van Der Heide, B., Ramirez Jr., A., Burgoon, J.K., Pena, J., 2015. Interpersonal and hyperpersonal aspects of computer-mediated communication. In: Sundar, S.S. (Ed.), The Handbook of Psychology and Communication Technology. Wiley-Blackwell, West Sussex, England.

Computerized Test Construction

Bernard P Veldkamp, University of Twente, Twente, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W.J. van der Linden, volume 4, pp. 2477–2480, © 2001, Elsevier Ltd.

Abstract

This article starts with a brief introduction to the area of computerized test construction. In the first section, the problem is formally stated as an optimization problem, with an objective function, for example, that maximizes the amount of information in the test or minimizes the amount of items in a test. In the second section, an overview is given of the methods that are available for computerized test construction. In sections computerized construction of linear test forms, construction of computerized adaptive tests, and construction of multistage tests, some specific issues related to the three most important test forms are discussed. The article ends with a discussion on one of the recent issues in computerized test construction, the problem of uncertainty in some of the parameters during optimization.

Test Construction as a Combinatorial Optimization Problem

Combinatorial optimization problems are those where mathematical techniques are applied to find optimal solutions within a finite set of possible solutions. The set of possible solutions is generally defined by a set of restrictions, and the set is too large for exhaustive search. A well-known example is the knapsack problem, where the value of the goods carried in the knapsack has to be maximized, while the weight of the goods that can be carried is limited. A second example is the traveling salesman problem, where the total traveling distance has to be minimized while each client is visited exactly once. The computerized test construction problem is also an example of a combinatorial optimization problem. From a finite item bank, a group of items has to be selected that is optimal with respect to the goal of testing, while the resulting test has to meet all specifications.

To formulate the test construction problem as a combinatorial optimization problem, a decision variable x_i is introduced that denotes whether item $i = 1,\ldots, I$, is selected ($x_i = 1$) or not ($x_i = 0$), an objective function has to be formulated to be optimized, and restrictions have to be identified.

Objective Functions in Computerized Test Construction

An objective function in test construction has to reflect the goal of testing that has been set in the first step of Birnbaum's approach. How to translate the objectives for test construction to psychometric properties depends on the measurement framework that is used. When the classical test theory (CTT) is applied to relate the responses of the candidate to a score representing the ability (Lord and Novick, 1968), the goals of testing are generally formulated in terms of reliability (approximated by Cronbach's alpha) or predictive validity of the test. Maximizing the reliability can be formulated in terms of the decision variable x_i as

$$\max \frac{n}{n-1}\left[1-\frac{\sum_{i=1}^{I}\sigma_i^2 x_i}{\left(\sum_{i=1}^{I}\sigma_i\rho_{iX}x_i\right)^2}\right], \qquad [1]$$

where n denotes the test length, σ_i the item variance, and ρ_{iX} the item discrimination. Maximizing the predictive validity can be formulated as

$$\max \frac{\sum_{i=1}^{I}\sigma_i\rho_{iY}x_i}{\sum_{i=1}^{I}\sigma_i\rho_{iX}x_i}, \qquad [2]$$

where ρ_{iY} is the item validity.

Within an item response theory (IRT) framework, the relationship between the responses of a candidate and the ability are modeled by an item response function, for example, the two-parameter logistic model (2PLM, Lord, 1980):

$$P_i(\theta) = \frac{e^{a_i(\theta-b_i)}}{1+e^{a_i(\theta-b_i)}}, \qquad [3]$$

where $P_i(\theta)$ denotes the probability that someone with ability θ will provide a correct answer to item i, and (a_i, b_i), also referred to as item parameters, denote the discrimination and the difficulty parameter of item i. Other examples of IRT models are the Rasch model, the three-parameter logistic model, or the normal ogive models (Lord, 1980). Within an IRT framework, the objective functions in test construction are generally formulated in terms of Fisher information. Fisher information has some favorable features when it comes to test construction. Fisher information for the whole test is equal to the sum of Fisher information of the items, and the inverse of Fisher information is asymptotically equal to the variance of the ability estimate. In other words, high information implies small uncertainty in the resulting scores. For the 2PLM, Fisher information for the whole test is given by

$$I(\theta) = \sum_{i=1}^{I}\frac{\partial P_i(\theta)/\partial\theta}{P_i(\theta)[1-P_i(\theta)]}x_i. \qquad [4]$$

This test information function (TIF) can be maximized for a single ability value or for a range of ability values, or different targets can be set for various values of the TIF.

For an overview of objective functions, both in a CTT and in an IRT framework, see van der Linden (2005, Chapter 5).

Restrictions in Computerized Test Construction

Restrictions in combinatorial optimization refer to specifications that have to be met for groups of items that are selected. In computerized test construction, three types of specifications can be distinguished: categorical, quantitative, and logical specification. Categorical specifications refer to item attributes, such as content, item type, or format, that categorize the items in different subsets. For each of these subsets, the number of items to be included in the test can be specified. Quantitative specifications refer to attributes like response times, word counts, or statistical attributes. The sum of the quantitative attributes of all items selected for the test, for example, the total word count, is restricted by these constraints. Logical specifications deal with relationships between items. Items that belong to an enemy set cannot be selected for the same test. Therefore, only one of the items from this subset can be selected. Van der Linden (2005, Chapter 3) presents an extensive overview of various types of constraints, and of its applications to test construction. All specifications can also be formulated in terms of the decision variables. A complete combinatorial optimization model for computerized test construction is formulated in the next section.

Computerized Test Construction Model

A very general model for computerized test construction can be formulated as

$$\max \sum_{i=1}^{I} I_i(\theta_j)x_i \quad \text{(objective function)} \qquad [5]$$

subject to

$$\sum_{i \in V_c} x_i \leq b_c \quad \forall c \quad \text{(categorical constraints)} \qquad [6]$$

$$\sum_{i=1}^{I} q_i x_i \leq b_q \quad \forall q \quad \text{(quantitative constraints)} \qquad [7]$$

$$\sum_{i \in V_l} x_i \leq b_l \quad \forall l \quad \text{(logical constraints)} \qquad [8]$$

$$\sum_{i=1}^{I} x_i = n, \qquad [9]$$

$$x_i \in \{0, 1\} \quad i = 1, \ldots, I \quad \text{(decision variables)}. \qquad [10]$$

Where I is the number of items in the bank, c, q, l are indices for the categorical, quantitative, and logical constraints, (b_c, b_q, b_l) denote the bounds for the constraints, and V_c and V_l denote sets of items affected by constraints c or l. The objective in this model is to maximize the TIF for ability level θ_j. This type of objective function is generally applied for mastery testing, where most information is needed around the cut-off point θ_j. Depending on the goals of testing, other objective functions can be applied.

Algorithms for Computerized Test Construction

Two classes of algorithms have been proposed for computerized test construction. First of all, mixed integer programming (MIP) solvers based on branch-and-bound algorithms can be applied. These algorithms guarantee that an optimal solution of the test construction problem is found, if one exists. If no solution to the test construction problem can be found, i.e., no test can be selected from the item bank that meets all the test specifications, these solvers can provide feedback on sets of constraints that together or by themselves cause the infeasibility problems (Huitzing et al., 2005). The commercial software package CPLEX (IBM, 2010) is one of the most powerful programs to solve test construction problems, but alternative solvers like the LpSolveAPI in the R-package are also available. Over the years, MIP solvers have proved to solve most computerized test construction problems within minutes or even seconds. Second, various heuristics for solving test construction problems have been proposed in the literature. These heuristics are often tailor-made for the test construction problem at hand. They are often easy to implement, but they approximate the optimal solution, so it cannot be guaranteed that the optimal test is constructed. Besides, finding the optimal settings of a heuristic might be rather time consuming. Some well-known heuristics for computerized test construction are the weighted deviation model (Stocking and Swanson, 1993), the normalized weighted average deviation heuristic (Luecht and Hirsch, 1992), network programming (Armstrong et al., 2005), simulated annealing (Veldkamp, 2002), genetic algorithms (Verschoor, 2007), and the MCMC (Markov Chain Monte Carlo)-based test assembly algorithm (Belov and Armstrong, 2005).

Computerized Construction of Linear Test Forms

Linear tests can be characterized as a fixed set of items presented in a fixed order. Paper-and-pencil tests are in this category, but they are also administered more and more often on a computer. The model in Eqns [5]–[10] can be applied to assemble a linear test form suitable for making mastery/nonmastery decisions around a cut-off point θ_j. If the goal of the test is to measure for a broader ability range, targets are generally formulated for Fisher information for these ability values, and the objective function for computerized test construction is to minimize the distance between the TIF and its target. Instead of minimizing the difference for all ability values, it often suffices to minimize the difference for a number of points θ_k, $k = 1, \ldots, K$, spread over the ability range of interest. This objective can be modeled as

$$\min_{k=1,\ldots,K} \left| \sum_{i=1}^{I} I_i(\theta_k)x_i - T(\theta_k) \right|, \qquad [11]$$

where $T(\theta_k)$ represents the target for Fisher information at θ_k. Boekkooi-Timminga and van der Linden (1989) developed a maximin model for this problem, in which a new variable γ is introduced, which serves as an upper bound for the

absolute difference between the TIF and the actual information function. In this way the optimization problem in Eqn [10] is rewritten as an MIP problem, and standard MIP solvers or heuristics can be applied:

$$\min y \qquad [12]$$

subject to

$$\sum_{i=1}^{I} I_i(\theta_k)x_i \leq T(\theta_k) + y \qquad [13]$$

$$\sum_{i=1}^{I} I_i(\theta_k)x_i \geq T(\theta_k) - y \qquad [14]$$

In many testing programs, more than one test has to be constructed. It could be that several parallel test forms have to be administered at the same time, that the test is administered periodically, or that the test is used in a quasiexperimental pre-posttest design (e.g., Cook and Campbell, 1979). To assemble multiple test forms concurrently, the test construction model in Eqns [5]–[10] is slightly modified by using decision variable x_{it} that indicate whether item i is selected ($x_{it} = 1$) for test t or not ($x_{it} = 0$), instead of decision variable x_i. When overlap between tests is prohibited or limited, the following logical constraint can be added to the model:

$$\sum_{t=1}^{T} x_{it} \leq n_i \quad i = 1, \ldots, I \qquad [15]$$

where n_i denotes the number of times item i can be selected concurrently. Since both the number of decision variables and the number of constraints grow linearly in the number of tests to be selected, 0–1 MIP algorithms might be very time consuming. Especially for the problem of concurrent construction of multiple partly overlapping tests, genetic algorithms and MCMC-based test assembly have demonstrated to be very efficient.

Construction of Computerized Adaptive Tests

Computerized adaptive tests (CATs) are individualized tests where the difficulty of the items is adapted to the ability of the candidate. In CAT, items are selected sequentially. After administration of each item, the ability of the candidate is estimated based on the responses he or she provided to the previous items. Then the item is selected and administered that maximizes Fisher information at the estimated ability level. The procedure is repeated until a stopping criterion, e.g., a fixed number of items or a minimum level of measurement precision, has been met. The first item can be selected randomly, or based on an initial guess of the ability level of the candidate. A major advantage of CAT is that test length is reduced up to 40%, and tests can be administered individually when a candidate prefers. For a thorough introduction to CAT, the reader is referred to Wainer (2000).

In case a set of specifications has to be met, the shadow test approach (van der Linden, 2005, Chapter 9) could be applied. This two-stage procedure for item selection constructs a 'shadow test' in every iteration of CAT from which the most informative item is chosen to be administered. The 'shadow test' consists of a group of most informative not-administered items which in combination with the items that have been administered already fulfill the requirements of the specifications. The following model can be applied to construct a 'shadow test' for selecting the g-th item:

$$\max \sum_{i=1}^{I} I_i(\hat{\theta}_{g-1})x_i \qquad [16]$$

subject to

$$\sum_{i=1}^{I} x_i = n \quad \text{(test length)} \qquad [17]$$

$$\sum_{i \in V_c} x_i \leq b_c \quad \forall c \quad \text{(categorical constraints)} \qquad [18]$$

$$\sum_{i=1}^{I} q_i x_i \leq b_q \quad \forall q \quad \text{(quantitative constraints)} \qquad [19]$$

$$\sum_{i \in V_l} x_i \leq b_l \quad \forall l \quad \text{(logical constraints)} \qquad [20]$$

$$\sum_{i \in S_{g-1}} x_i = g - 1 \quad \text{(previous items)} \qquad [21]$$

$$x_i \in \{0, 1\} \quad i = 1, \ldots, I \quad \text{(decision variables)}, \qquad [22]$$

where $\hat{\theta}_{g-1}$ denotes the ability to estimate after administering $(g-1)$ items and S_{g-1} denotes the set of items that has been administered so far. By application of this model, the same kind of specifications can be applied in CAT as in linear testing. An alternative procedure for dealing with test specification in CAT would be to apply the weighted deviation model (Stocking and Swanson, 1993).

In CAT, some items are selected more often than others, since they are more informative. As a result, their content might be compromised. To prevent these security issues and to increase item bank usage, exposure control methods can be applied. These methods prevent the selection of popular items either by conducting a chance experiment after selection of each item where the probabilities of selection are inversely related to the popularity (Sympson and Hetter, 1985) or by temporarily excluding them when they are administered too often (van der Linden and Veldkamp, 2004).

Construction of Multistage Tests

Multistage testing is a hybrid of linear testing and CAT. Instead of administering a single item, a small linear test, also referred to as a module (Luecht and Nungester, 1998), is administered before the ability estimate is updated. After estimating the ability, the candidate is routed to an easier or more difficult module. A multistage test design typically consists of several stages in which a module is administered. Each path a candidate might follow through the test has to meet all test specifications. An example of a multistage test design with three stages, six modules, and four paths is given in **Figure 1**.

To develop a multistage test, several decisions have to be made about the number of stages, routing rules, number of items in a module, and psychometric properties of the modules

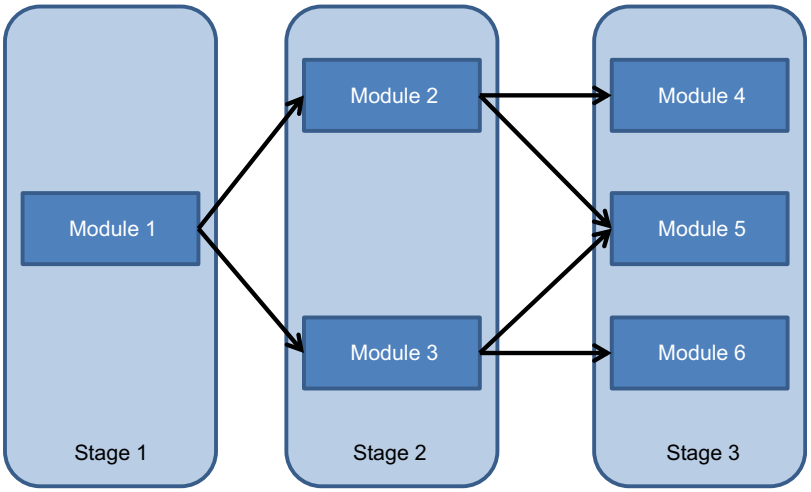

Figure 1 Multistage test design.

(Zenisky et al., 2010). The model for constructing the modules is equivalent to the model for the construction of a number of (non)overlapping small linear tests, while an additional set of specifications for each combination of tests that belong to a path has to be met:

$$\sum_{t \in P_p} \sum_{i \in V_c} x_{it} \leq b_c \quad \forall c, \forall p, \qquad [23]$$

$$\sum_{t \in P_p} \sum_{i=1}^{I} q_i x_{it} \leq b_q \quad \forall q, \forall p, \qquad [24]$$

$$\sum_{p \in P_p} \sum_{i \in V_l} x_{it} \leq b_l \quad \forall l, \forall p, \qquad [25]$$

where p is an index denoting various paths in the design. Zenisky et al. (2010) pointed out that it should be considered whether specifications should be achieved within stages or across the whole test. Meeting the constraints over the whole test provides greater flexibility, at the cost of imposing the additional constraints in Eqns [23]–[25] to the model.

Robust Computerized Test Construction

In all these test construction models, it is assumed that the item parameters and item attributes are known. Fixed values are used to compute the contribution of each item to both the objective function and the constraints. Unfortunately, many of them have been estimated. IRT parameters, for example, have been estimated when the items were pretested, and they do have uncertainty in them that is reflected by the standard error of estimation. This uncertainty is generally not taken into account. Hambleton and Jones (1994) already warned about the consequences. When the objective of test construction is to maximize the information in the test, and the uncertainty in the item parameters is not taken into account, the amount of information in the resulting test will be seriously overestimated, especially when the item parameters have been estimated with small sample sizes or when relatively few items are selected. Since the amount of information in the test is inversely related to the variance of the ability estimate, this implies that overestimation of the information in the test has serious consequences for the measurement precision and the reliability of the test.

De Jong et al. (2009) proposed a robust test construction algorithm to deal with these problems. Inspired by the work of Soyster (1973), they subtracted one time the standard of estimation from the parameter estimates, and formulated a robust test construction model as

$$\max \sum_{i=1}^{I} I_i(\theta_j, \zeta) x_i \quad \text{(objective function)} \qquad [26]$$

where ζ denotes the level of uncertainty, and

$$I_i(\theta_j, \zeta) = I_i(\theta_j) - SE(I_i(\theta_j)). \qquad [27]$$

For small values of the standard error of estimation, the results for the test construction problem in Eqns [5]–[10] and its robust counterpart are almost the same, but for large values, a considerable difference will be obtained.

See also: Classical (Psychometric) Test Theory; Psychometrics: Classical Test Theory; Psychometrics; Reliability: Measurement.

Bibliography

Armstrong, R.D., Jones, D.H., Wang, Z., 1995. Network optimization in constrained standardized test construction. In: Lawrence, K.D. (Ed.), Applications of Management Science: Network Optimization Applications, vol. 8. JAI, Greenwich, CT, pp. 189–212.

Belov, D.I., Armstrong, D.H., 2005. Monte Carlo test assembly for item pool analysis and extension. Applied Psychological Measurement 29, 239–261. http://dx.doi.org/10.1177/0146621605275413.

Birnbaum, A., 1968. Some latent trait models and their use in inferring an examinee's ability. In: Lord, F.M., Novick, M.R. (Eds.), Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA, pp. 397–479.

Boekkooi-Timminga, E., van der Linden, W.J., 1989. A maximin model for test design with practical constraints. Psychometrika 54, 237–247. http://dx.doi.org/10.1007/BF02294518.

Cook, T.D., Campbell, D.T., 1979. Quasi-Experimentation: Design & Analysis Issues for Field Settings. Houghton Mifflin Company, Boston.

De Jong, M.G., Steenkamp, J.-B.G.M., Veldkamp, B.P., 2009. A model for the construction of country-specific yet internationally comparable short-form marketing scales. Marketing Science 28, 674–689. http://dx.doi.org/10.1287/mksc.1080.0439.

Hambleton, R.H., Jones, R.W., 1994. Item parameter estimation errors and their influence on test information functions. Applied Measurement in Education 7, 171–186. http://dx.doi.org/10.1207/s15324818ame0703_1.

Huitzing, H.A., Veldkamp, B.P., Verschoor, A.J., 2005. Infeasibility in automated test assembly models: a comparison study of different methods. Journal of Educational Measurement 42, 223–243. http://dx.doi.org/10.1111/j.1745-3984.2005.00012.x.

IBM, 2010. CPLEX. IBM Corporation Software Group, Somers, NY.

Lord, F.M., 1980. Applications of Item Response Theory to Practical Testing Problems. Erlbaum, Hillsdale, NJ.

Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Welsley, Reading MA.

Luecht, R.M., Hirsch, T.M., 1992. Item selection using an average growth approximation of target information functions. Applied Psychological Measurement 16, 41–51. http://dx.doi.org/10.1177/014662169201600104.

Luecht, R.M., Nungester, R.J., 1998. Some practical examples of computer-adaptive sequential testing. Journal of Educational Measurement 35, 229–249. http://dx.doi.org/10.1111/j.1745–3984.1998.tb00537.x.

Nemhauser, G.L., Wolsey, L.A., 1988. Integer and Combinatorial Optimization. Wiley, New York.

Soyster, A.L., 1973. Convex programming with set-inclusive constraints and applications to inexact linear programming. Operations Research 21, 1154–1157.

Stocking, M.L., Swanson, L., 1993. A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement 17, 277–292. http://dx.doi.org/10.1177/014662169301700308.

Sympson, J.B., Hetter, R.D., 1985. Controlling item-exposure rates in computerized adaptive testing. In: Proceedings of the 27th Annual Meeting of the Military Testing Association. Navy Personnel Research and Development Center, San Diego, CA, pp. 973–977.

van der Linden, W.J., 2005. Linear Models of Optimal Test Design. Springer, New York.

van der Linden, W.J., Veldkamp, B.P., 2004. Constraining item exposure rates in computerized adaptive testing with shadow tests. Journal of Educational and Behavioral Statistics 29, 273–291. http://dx.doi.org/10.3102/10769986029003273.

Veldkamp, B.P., 2002. Multidimensional constrained test assembly. Applied Psychological Measurement 26, 133–146. http://dx.doi.org/10.1177/014216020026002002.

Verschoor, A.J., 2007. Genetic Algorithms for Automated Test Assembly, Unpublished Doctoral Dissertation. Enschede, The Netherlands: University of Twente.

Wainer, H., 2000. Computerized Adaptive Testing: A Primer, second ed. Lawrence Erlbaum Associates, Mahwah, NJ.

Zenisky, A., Hambleton, R.K., Luecht, R.M., 2010. Multistage testing: issues, design and research. In: van der Linden, W.J., Glas, C.A.W. (Eds.), Elements of Adaptive Testing. Springer, New York, pp. 355–372. http://dx.doi.org/10.1007/978-0-387-85461-8_18.

Computers: Impact on the Social Sciences

Gerd Gigerenzer, Max Planck Institute for Human Development, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The social organization of labor in the nineteenth century served as the model for Babbage's first computers. In the second half of the twentieth century, when working computers were finally constructed and invaded the offices of social scientists, they turned into theories of mind. This mutual inspiration first changed the meaning of calculation and then led to a new understanding of thought as computation based on hierarchically organized subroutines. The computer as a research tool has changed the social sciences in a fundamental way, from enabling large-scale simulations of cognitive and social systems to allowing mindless and mechanical use of statistics.

A new social organization of the workplace, the large-scale division of labor in manufacturing, served as the original model for Charles Babbage's computers. Today's computer metaphor of mind still reflects this origin: it pictures cognitive processes – from reasoning to creativity – as a hierarchical organization of subroutines. The two different uses of computers – for calculation (number processing) and computation (symbol processing, as in simulation) – vary markedly among scientific disciplines. The computer has revolutionized our conception of mind and society in the sense of a re-volution to the Enlightenment's combinatorial view of thought and Adam Smith's ideal of the division of labor.

Following the advent of writing and of the printing press, the invention of the computer has been described as the third information revolution. Computers transformed our world, in reality and fiction: from the World Wide Web to the vision of machines that will eventually surpass humans in intelligence and even replace us by self-replicating computers divorced from biological evolution. This revolution was difficult to foretell. Howard Aiken, the Harvard mathematician and builder of the Mark I calculator, recalled: "Originally one thought that if there were a half dozen large computers in this country, hidden away in research laboratories, this would take care of all requirements we had throughout the country." This prediction was based on the false assumption that computers are there to solve scientific problems only. Back in the 1960s, electrical engineer Douglas Carl Engelbart had already designed the first interactive computer tools, including the mouse, on-screen editing, screen windows, hypertext, and electronic mail. However, at this time, human–computer interaction still seemed a pipe dream – computers were for processing punched cards, not for interacting with humans. The impact computers had on society and science was difficult to imagine, and its future impact is equally impossible to envision.

What remains clear is that its impact is bidirectional: computers and humans coevolve. This is illustrated by the work of Charles Babbage (1791–1871), the English mathematician often credited with the invention of the digital computer.

The Computer as a Factory of Workers

Babbage's Analytical Engine, a mechanical computer, was inspired by and modeled on a new social organization of work: the large-scale division of labor, as evidenced in the English machine tool industry and in the French government's manufacturing of logarithmic and trigonometric tables for the new decimal system in the 1790s. Inspired by Adam Smith's praise of the division of labor, French engineer Gaspard de Prony organized the project in a hierarchy of tasks. At the top were a handful of first-rank mathematicians who devised the formulae, including Adrien Legendre and Lazare Carnot; in the middle, seven or eight persons trained in analysis; and at the bottom, 70 or 80 unskilled persons who performed millions of additions and subtractions.

Once it was shown that elaborate calculations could be carried out by an assemblage of unskilled workers – rather than by a genius such as Gauss – each knowing very little about the larger computation, Babbage was able to conceive of replacing these workers with machinery. Babbage, an enthusiastic 'factory tourist,' explicitly referred to this division of mental labor as the inspiration for his mechanical computer and adopted terms from the textile industry such as 'mill' and 'store' to talk about its parts. Similarly, he borrowed the use of punched cards from the Jacquard loom, the programmable weaving machines that used removable cards to weave different patterns. Thus, in the beginning was a new social system of work, and the computer was created in its image.

The Hierarchical Organization of the Mind

Babbage's Analytical Engine was never completed. Around 1890, the first 'digital calculators' were manufactured, so called because they represented numbers by cogged wheels whose positions corresponded to the *digits* in the numbers denoted. (An abacus was still cheaper and faster.) Digital computers competed with analog computers, in which continuous quantities, such as electrical potential or mechanical motion (slide rules) were used instead of discrete (e.g., on/off) states. Until the first third of the twentieth century, digital calculators remained essentially mechanical. In the late 1930s, electromechanical relay switches started replacing gears and cogs in calculators. The electronic computer based on vacuum tubes and containing no movable parts appeared during World War II and was invented in three separate locations: in Germany for airplane design, in the USA for calculating artillery tables, and in the UK for breaking German secret codes. Vacuum tubes

were eventually replaced by transistors and then by silicon chips, which allowed the transition from large mainframes to minicomputers. The first personal computers – complete with keyboard, screen, disk drive, and software – were marketed by Apple in 1977 and by IBM in 1981.

Through these dramatic improvements in hardware and speed, Smith's ideal of the division of labor became the basis for a fresh understanding of the human mind. Herbert Simon and Allan Newell proposed that human thought and problem solving were to be understood as a hierarchical organization of processes, with subroutines, stores, and intermediate goal states that decomposed a complex problem into simple tasks. In fact, a social system rather than a computer performed the trial run for the Logic Theorist, their first computer program. Simon's wife, children, and graduate students were assembled in a room where each of them became a subroutine of the program, handling and storing information. This social design resembled that of the nineteenth-century French *bureaux de calculs* and the Manhattan project, where calculations were performed by an unskilled, poorly paid workforce of mostly women. This vision of intelligence resulting from a crowd of ignorant elements made its way into cognitive theories. For instance, Marvin Minsky, one of the founders of artificial intelligence, regarded the mind as a society of dumb agents, collectively creating true intelligence. In the same spirit, anthropologists began to use computer analogies to understand how social groups make decisions, such as how the crew on a large ship solves the problem of navigation by storing, processing, and exchanging information. Thus, the direction of the analogy was eventually reversed: originally, the computer was modeled on a new social system of work; today, social systems of work are modeled on the computer.

From Calculation to Computation

A crucial insight occurred when the early generation of computer designers realized that a computer could be used not only for calculation but more generally for symbol processing. The switch positions inside the machine could take on other meanings than numbers; they could also stand for symbols representing concepts. This insight fueled the view of *cognition as computation*, that is, symbol manipulation. A computer that merely performed complex calculations would not have suggested itself as a model of human thought. During the Enlightenment, proficient calculation was indeed seen as the essence of intelligence and even moral sentiment, but after the large-scale division of labor introduced around 1800, prodigious mental reckoning had become the mark of the idiot savant rather than the genius.

The notion of computation led to two questions: Can computers think? Is human thought computation? To avoid endless philosophical debate on exactly how to define the first question, Alan Turing (1912–54) suggested the 'imitation game,' now known as the Turing test: A human interrogator must distinguish between a computer and a human subject based on their replies to the interrogator's questions. In a series of such experiments, a computer's ability to think could be quantified as the proportion of interrogators who misidentify the machine as the human subject. Turing asked whether the computer is like a mind: whether it thinks and has free will, and whether one can teach machines to become intelligent using the same psychological principles used to teach children. Those who focused on the second question – Is the mind like a computer? – saw the analogy point in the other direction. For instance, some tried to teach children using the same principles that had worked for computers. This analogy of the mind as computer triggered the cognitive revolution in psychology during the 1960s and 1970s.

The 'physical symbol system' hypothesis proposed by Simon and other researchers in artificial intelligence assumes that human cognition *is* computation in the sense of symbol manipulation. According to this view, symbols represent the external world; thought consists of expanding, disassembling, and reforming their structures, and intelligence is nothing more than the ability to process them. Information processing became the key to understanding human thought, while the tools of programming provided the mechanisms of thought: production systems, subroutines, recursion, iteration statements, local naming, interpreters, and so on. This view was not easily reconcilable with prevailing theories of thought, be it Freud's theory of mind as manipulating biological energies; the behaviorists' standpoint that thought is simply silent speech; or the Gestalt psychologists' emphasis on context dependency, interconnectedness, and emergent quality of thought. In essence, the vision of cognition as symbol manipulation revived the Enlightenment view of thought as a combinatorial calculus. For Condillac, d'Alembert, Condorcet, and other Enlightenment philosophers, the healthy mind worked by constantly breaking down ideas and sensations into their minimal elements, then comparing and rearranging these elements into novel combinations and permutations.

The view of thought as symbol manipulation also revived the old dream of a logic of scientific discovery, the existence of which Karl Popper had so vehemently disclaimed. Computer programs such as Bacon managed to infer scientific laws from data, and artistic creativity was modeled on the same combinatorial principles. Computer programs were written that composed music, wrote stories, and designed new Palladian villas and prairie homes. Leibniz's dream of a universal characteristic that would assign a number to each concept and reduce each problem to calculation seemed to have come true. Alongside Babbage and Pascal, Leibniz had designed one of the first computers.

Through the medium of the computer, Enlightenment psychology reentered modern psychology. This is, however, only half the story. Just as George Boole had set out in the mid-nineteenth century to derive the laws of logic and probability from the psychological laws of human thought, researchers such as Simon and Newell tried to enter the results of psychological research about human thought into their computer programs. For instance, the concept of heuristics that speeds up search in a problem space proved to be essential for intelligent programs. Heuristics are smart rules of thumb that humans apply in a world of unknown risks where optimization, such as Bayesian probability updating and maximizing expected utility, is impossible or would cost too much time (*see* Decision Making: Nonrational Theories).

The metaphor of the mind as computer must be distinguished from John von Neumann's (1903–57) metaphor of the

brain as computer. Von Neumann, known as the father of the modern computer, looked at similarities between the nervous system and computer, neuron and vacuum tube – but added cautionary notes on their differences. Turing, in contrast, believed that the observation that both the modern digital computer and the human nervous system are electrical was based on a superficial similarity. He pointed out that the first digital computer, Babbage's Analytical Engine, was mechanical rather than electrical and that the important similarities between computer and mind concern function rather than hardware.

How quickly did the metaphor of the mind as computer gain acceptance among cognitive psychologists? Consistent with the tools-to-theories heuristic – which says that new scientific tools suggest new metaphors of mind, society, or nature but are accepted by a scientific community only after its members have become users of the tool – the mind as computer metaphor was not broadly accepted until the 1970s and 1980s. Before the advent of personal computers, researchers had little direct contact with the large mainframe computers, and for those who did, computers were a source of constant frustration. For instance, in an average week in 1965–66, the PDP-4C computer of the Center for Cognitive Studies at Harvard saw 83 h of use, but 56 h of these were spent on debugging and maintenance. A 1966 technical report of the Center was entitled "Programmanship, or how to be one-up on a computer without actually ripping out its wires." The metaphor of the mind as computer became widely accepted only after the tool had become part of the daily routine in psychological laboratories – after the advent of personal computers.

From Thought Experiments to Simulation

Thought experiments in fields from physics to political science are research tools used in situations where no data can be obtained, such as counterfactual historical scenarios. The computer created a new species of thought experiments: computer simulation. Jay Forrester's Club of Rome world resource model – which predicted worldwide famine followed by war in the first third of the twenty-first century – is one of the best-known early social science simulations. Robert Axelrod's classic 'prisoner's dilemma' computer tournament demonstrated the success of a simple social heuristic called tit-for-tat and became the model for numerous simulations in game theory, reciprocal altruism, and political science. Simulations force researchers to specify their theoretical ideas precisely and have become a tool for discovering new implications of these ideas. In some fields, precise articulation of assumptions has become a requirement for what constitutes a theory: what cannot be simulated does not deserve to be called a theory.

Computer simulation allows for the modeling of dynamic systems, such as the evolution of social structures through repeated interaction between individuals. It is a truly new form of experiment, but social scientists are divided over its role in research. In artificial intelligence, organizational theory, and demography, for instance, simulation has a long tradition. In sociology, in contrast, simulation is still rare and frequently disparaged. Simulation tends to play a minor role in fields with a deeply entrenched distinction (or even hostility) between theoretical and empirical research, such as sociology, and where computers are heavily used for the analysis of empirical data. Perceived as the instrument of the 'other,' number-crunching camp, the computer's theoretical possibilities tend to be forgotten.

A World of Numbers and Statistics

When the statistician Karl Pearson assembled the famous *Biometrika* statistical tables between 1914 and 1934, he used a 50-lb Brunsviga calculator and numerous clerks. The 1946 ENIAC – commonly referred to as the first modern digital electronic computer using vacuum tubes – could calculate 5000 additions per second, that is, did the work of a whole factory of clerks. Fifty years later, a tiny and relatively inexpensive Pentium chip could conduct over 500 million operations per second. This explosion of calculating power together with decreasing cost and size has had a dramatic impact on the social sciences, both positively and negatively. In some fields, the emphasis has shifted from collecting qualitative observations to quantitative data, and from *a priori* theoretical thinking to the *post hoc* interpretation of significant correlations.

At the end of World War II, half of the articles in the *American Sociological Review* and *American Journal of Sociology* lacked any kind of quantitative mathematical analysis; 30 years later, this number had decreased to about 13%. In psychology, as in sociology, the predominant quantitative methods before the advent of electronic computers were descriptive statistics, such as means, variances, percentages, correlations, and cross tabulations. With the introduction of fast calculators, the social sciences have witnessed a rapid growth in the usage of computationally expensive statistical methods, including factor analysis, cluster analysis, analysis of variance, and multiple regression. The computer also provided tools that could make data structures visually transparent and help researchers see underlying structures – called exploratory data analysis. Sociologists, demographers, and anthropologists were finally able to analyze census data, voting polls, and other large samples in a short time.

Computerized statistical packages such as SPSS (Statistical Package for the Social Sciences) have not only helped but also hindered progress. Their ease of application (e.g., compute 'everything by all') has often replaced statistical thinking with statistical rituals. One widespread ritual is null hypothesis testing, in which the researcher mechanically tests whether data significantly deviate from a null hypothesis ('chance'), without ever specifying the predictions of the research hypothesis and of competing hypotheses, and without measuring effects and constructing substantive theories of some depth. As a consequence, researchers tend to skim through computer printouts looking not at the data but at the significance level of the correlations or chi-squares and engage in *post hoc* explanations for whatever comes out significant. The consequence of this grave misuse of computational power is an increasing number of false positive results and flawed research featuring significant effects that cannot be replicated.

While fast computers made routine statistical analysis feasible, they also became enlisted in the 'big data' program

that aims at dispensing with statistical sampling theory, and theory in general. Using huge amounts of data provided by Google and other sources, claims were made that sampling theory would be obsolete because '$N = \text{All}$,' that causes could be substituted by correlations, and that with enough data, the numbers would speak for themselves. For Simon, the computer promised a theory of mind, for proponents of big data, it meant the end of theory.

Research and Teaching Assistance

Next to data analysis, the most common function of computers in the social sciences is that of an efficient and patient research assistant. A computer can serve as a librarian who searches for relevant publications and historical records; as a data bank manager who stores and retrieves narrative information, qualitative and numerical data; as a secretary who edits and files texts, footnotes, and references; and as a fast electronic mail system that facilitates contacts with colleagues worldwide. In experimental psychology, computers run entire experiments with humans and animals; they present the experimental stimuli and record the participants' responses.

This interactive use of computers would not have been possible without the invention of the personal computer that replaced the large mainframes. Engelbart's invention of the mouse exemplifies how human abilities such as hand–eye coordination can be exploited to design appropriate tools with which humans can almost intuitively interact with computers. In social science teaching, human–computer interaction can increase the participation of all students, whereas in the traditional classroom small groups of students often tend to dominate the discussion. Facebook, chat rooms, and the World Wide Web in general have become new and cheap sources of 'big data' for analyzing social networks and consumer behavior. The Web in turn has generated legal and ethical problems concerning secrecy and data protection, motivating revisions of copyright, tax, and criminal law.

The Computer Revolution

Has the computer fundamentally changed the social sciences? As we have seen, its impact differs across the social sciences. The invention of the modern computer has increased interdisciplinarity through the creation of *cognitive science*, consisting of artificial intelligence, psychology, philosophy, linguistics, anthropology, and neuroscience. However, sociology is notably absent from that list. Electronic mail and the Internet have made international collaboration easier and faster than ever. Simulation, expert systems, neural networks, and genetic algorithms would be unthinkable without high-speed computers, as would statistical analyses of large bodies of data. In these respects, the social sciences today have a repertoire of powerful new tools and opportunities.

Has the computer revolutionized our conception of mind and society? The answer seems to be yes if we use the original meaning of the term – a re-volution, that is, a return to earlier conceptions. The metaphor of the mind as computer, for instance, has resurrected the combinatorial view of cognition of the Enlightenment and Adam Smith's ideal of a division of labor. Thinking, problem solving, creativity, and scientific discovery came to be explained by a hierarchically organized mind that decomposes complex problems into simpler subroutines. In this way, theories of mind and society that had existed before computers were expanded and perfected.

See also: Computational Approaches to Model Evaluation; Computer-Assisted Instruction; Computerized Test Construction; Metaphor, Role in Social Thought: History of the Concept; Pearson, Karl (1857–1936); Quantification in the History of the Social Sciences; Science and Technology, Social Study of: Computers and Information Technology; Social Simulation: Computational Models; Statistics, History of.

Bibliography

Ceruzzi, P.E., 1998. A History of Modern Computing. MIT Press, Cambridge, MA.
Collins, T.W., 1981. Social science research and the microcomputer. Sociological Methods & Research 9, 438–460.
Crevier, D., 1992. AI: The Tumultuous History of the Search for Artificial Intelligence. Basic Books, New York.
Cohen, I.B., 1998. Howard Aiken on the number of computers needed for the nation. IEEE Annals of the History of Computing 20, 27–32.
Daston, L., 1994. Enlightenment calculations. Critical Inquiry 21, 182–202.
Garson, G.D., 1994. Social science computer simulation: its history, design, and future. Social Science Computer Review 12, 55–82.
Gigerenzer, G., Goldstein, D.G., 1996. Mind as computer: the birth of a metaphor. Creativity Research Journal 9, 131–144.
Gigerenzer, G., 2000. Adaptive Thinking: Rationality in the Real World. OUP, New York.
Gigerenzer, G., 2004. Mindless statistics. Journal of Socio-Economics 33, 587–606.
Langley, P., Simon, H.A., Bradshaw, G.L., Zytkow, J.M., 1987. Scientific Discovery. MIT Press, Cambridge, MA.
McCorduck, P., 1979. Machines Who Think. Freeman, San Francisco.
Newell, A., Simon, H.A., 1972. Human Problem Solving. Prentice-Hall, Englewood Cliffs, NJ.
Schaffer, S., 1994. Babbage's intelligence: calculating engines and the factory system. Critical Inquiry 21, 203–227.
Simon, H.A., 1969. The Sciences of the Artificial. MIT Press, Cambridge, MA.
Turing, A.M., 1950. Computing machinery and intelligence. Mind 59, 433–460.

Comte, Auguste (1798–1857)

Mary Pickering, San Jose State University, San Jose, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Auguste Comte is the founder of the philosophy of positivism, the disciplines of sociology and the history of science, and the Religion of Humanity. These innovations were inspired by his desire to reorganize society after the upheaval of the French Revolution of 1789. Influenced by his education at the Ecole Polytechnique and his 7-year association with the social reformer Henri de Saint-Simon and disillusioned by the revolutions, monarchical regimes, republics, and empires that punctuated his existence, Comte rejected purely political and institutional reforms as ineffective. In the *Cours de philosophie positive* (1830–42) and the *Système de politique positive* (1851–54), he argued that social harmony could occur only if individuals were brought together by a consensus that was both intellectual and emotional in nature. Ideas as well as feelings had to be reordered. He first created positivism and its keystone, sociology, and then devised the secular 'Religion of Humanity' in order to construct a new cohesive society that would encompass the conservatives' call for order and the leftists' insistence on progress. Comte's doctrine had a large impact in Europe, the USA, and Latin America. It also influenced sociology, the history and philosophy of science, historiography, and anthropology.

Auguste Comte is the founder of the philosophy of positivism, the disciplines of sociology and the history of science, and the Religion of Humanity. These innovations were inspired by his desire to reorganize society during the period of upheaval following the French Revolution of 1789. Living through several revolutions, monarchical regimes, republics, and empires, Comte rejected purely political and institutional reforms as ineffective. Instead, he argued that social harmony could only occur if individuals were brought together by a consensus that was both intellectual and emotional in nature. Ideas as well as feelings had to be reordered. Comte's search for a consensus derived not only from the anarchy of the postrevolutionary period but from the instability in his own life, which led him to view inner harmony as a sign of health. He created positivism and its keystone, sociology, to construct a new cohesive society that would encompass the conservatives' call for order and the leftists' insistence on progress. This article will investigate Comte's personal and intellectual development as well as his contributions to knowledge and politics.

Life and Main Works

The son of a minor bureaucrat, Comte was born on 19 January 1798 in Montpellier, France. Growing up in this center of counterrevolutionary activity, he was deeply affected by the chaos stemming from the French Revolution. By age 14, he renounced both the monarchy and God, alienating his royalist, Catholic parents. Comte's rebelliousness was shared by many members of his generation, who longed for a sense of community and new principles of legitimacy that would stabilize moral and political life. His insubordinate behavior led to his expulsion from the Ecole Polytechnique, the prestigious Parisian engineering school, which had admitted him in 1814 because he had performed brilliantly on the entrance examination. (He was especially gifted in mathematics.) After his dismissal from the Ecole Polytechnique, in 1816, he committed himself to fulfilling the radical revolutionaries' dream of erecting a secular republic.

Living in Paris, he accepted in 1817 the position of secretary to Henri de Saint-Simon, who is often considered an early socialist. They became close collaborators in many journalistic enterprises. However, in 1824 they parted ways after Comte accused him of trying to take credit for his 'fundamental opuscule', the *Plan des travaux scientifiques nécessaires pour réorganiser la société*, which outlined his social ideas. Although Comte later denied Saint-Simon's influence, the older reformer showed him that they were in an era of transition that had to be completed by constructing a new society founded on industry and a new system of scientific knowledge, the 'positive philosophy'. This philosophy centered on the scientific study of society. Using his talent for synthesis and system building, Comte brought to fruition the seminal ideas strewn haphazardly throughout Saint-Simon's frequently incoherent works.

In 1826 Comte launched a course in positive philosophy in his apartment. Reflecting the high regard in which he was held, the course was attended by the famous economist Charles Dunoyer and many leading scientists, including Henri de Blainville, François Arago, Joseph Fourier, Alexander von Humboldt, and François Broussais. Yet after several lectures, Comte suffered a mental breakdown and spent 8 months in an asylum run by J.-E.D. Esquirol. He struggled against manic depression and paranoia for the rest of his life.

Moreover, he found it difficult to secure a career worthy of him, though he tried repeatedly to obtain a chair in analysis and mechanics at the Ecole Polytechnique and gain admission to the Collège de France. However, scientists and other scholars refused to consider him seriously because rumors about his mental instability continued to circulate, his prickly character made it hard for him to network, he never published a monograph on a particular science, and he never received an advanced degree. After working as a mathematics tutor and journalist, he obtained a lowly teaching and administrative position at the Ecole Polytechnique, where he remained from 1832 to 1851.

Between 1830 and 1842, Comte wrote his most important work, the *Cours de philosophie positive*. It covered the history of the sciences, demonstrating that it was time for the study of society to become a science as well. Once 'sociology' was

established, the system of scientific knowledge would be completed. Comte called this system 'positivism'. The *Cours* attracted the attention of several famous scholars and writers, including John Stuart Mill and Emile Littré, but they eventually became estranged from him.

Another person who experienced a rupture with Comte was his wife, Caroline Massin. A penniless director of a reading room, she married him in 1824 and helped him recover from his attack of madness. They had a tumultuous marriage, marked by his fits of jealousy, and they separated in 1842. To punish her, he later called her a prostitute.

Three years after their separation, Comte fell in love with Clotilde de Vaux, who was 17 years younger than him. Having been abandoned by her husband, this aspiring writer grew close to Comte but resisted his sexual advances. After she died of tuberculosis in 1846, Comte made her into his muse.

After her death, Comte became more of an activist, thanks in part to the reforming impulses let loose by the Revolution of 1848. He established his own club, the 'Positivist Society', to launch the positivist movement, and he wrote a manifesto, the *Discours sur l'ensemble du Positivisme*, which was a more extensive introduction to his philosophy than his *Discours sur l'esprit Positif* of 1844. Disappointed that scientists and other intellectuals neglected him, he targeted two new groups, workers and women, whose social roles were highlighted in the *Système de politique positive*. Published in four volumes between 1851 and 1854, this work expressed his political and social ideas. It promoted the importance of the emotions, established the Religion of Humanity, and made morality the seventh science. Many scholars blame his relationship with de Vaux for Comte's alleged decline into religious, sentimental thinking. Yet there was no sudden break in his career, for from the start he distrusted the purely scientific spirit, committed himself to moral reform, and sought to create a new belief system to replace traditional religion.

The last years of Comte's life were marked by increasing seclusion and conservatism. Much to the displeasure of many of his disciples, Comte supported the coup d'état of Louis Napoleon in 1851, partly out of fear of anarchy and partly out of hope of converting him to positivism. He also redoubled his efforts to gain the support of women, especially because he was disappointed that workers remained committed to socialism and would not consider his political system. To gain women's attention, he wrote the *Catéchisme positiviste* (1852), which featured a series of dialogues between a positivist priest of Humanity and a woman. Although not a feminist, he proclaimed that women would be given new respect in the positivist era because they would be essential moral agents, encouraging the development of sociability and ensuring a compassionate society. Later he published *Appel aux conservateurs* (1855) to persuade conservatives to join the positivist movement. Finally, in 1856 Comte wrote the *Synthèse subjective*, which explored the importance of images and the emotions in the context of mathematics and logic in order to bring together the three facets of our existence: intellectual speculation, affection, and activity. Dependent financially on a small group of followers, he died of stomach cancer in Paris on 5 September 1857.

Auguste Comte was a brilliant visionary. He sought to create in society the harmony and stability that he could not find in his private life or in postrevolutionary France. However, his paranoia, egoism, and dogmatism impaired his personal relations and his ability to launch a viable political and intellectual movement. He never gained the fame or power that he had hoped to achieve. Yet after his death, his ideas had a large influence on intellectual and political life both in Europe and on other continents.

Positivism and the History of Science

Comte's original version of positivism, which is far from the scientism with which it is often associated, had both a theoretical component and a practical, or political, aspect. It did not celebrate the sciences for their own sakes or seek to be value free as scientific systems often claim to be. Positivism sought to facilitate useful, constructive activity that would improve the human condition. In particular, it promoted the establishment of a new secular, republican society imbued not only with the certainty that came with scientific knowledge but with 'altruism', a word he coined around 1850 in order to counteract the sense of disconnectedness pervading modern society. Above all, Comte wanted members of society to be united by an intellectual consensus on fundamental aspects of knowledge and by emotional bonds in order to improve themselves and their natural and social environments.

Being an idealist, Comte believed that intellectual confusion was at the root of the postrevolutionary anarchy because ideas always determined the social structure. Rejecting the conventional abstractions and practices of traditional religion and politics as useless, he sought to create an intellectual revolution that would first lead to a moral revolution and then to a political and social revolution. An intellectual revolution would occur if one placed knowledge of society on a sure foundation to attain more control over it. To this end, he recommended extending the scientific, or 'positive', method from the investigation of nature to the study of politics and society.

The 'positive' designated the certain, precise, useful, relative (as opposed to the absolute), and constructive (as opposed to the negative). Positive laws must avoid references to God and to other first and final causes or substances, which were incomprehensible. These laws had to express the spatial and temporal relationships between phenomena and do so in such a way as to enable scientists to make predictions affecting the world of action. To arrive at these laws, scientists should try to replicate as much as possible the deductive, rational approach of mathematics and eschew pure empiricism because collecting isolated facts was a waste of time. To help them to construct abstract laws and theories, Comte sought to free scientists from the tedious task of direct observation. He insisted that individuals could not even make observations and retain facts without first formulating a theory to guide their perceptions. He also promoted provisional hypotheses as important ways of advancing scientific understanding when concrete, observed facts were missing. Furthermore, scientists should use their imagination to create 'scientific fictions', or hypothetical cases, to fill in poorly understood areas. Comte recognized the role of aesthetics and the subjective in formulating scientific hypotheses. To avoid giving reason excessive importance in scientific research, he refused to offer universal logical rules of scientific

procedure and scientific proof. Indeed, he insisted that one could never arrive at absolute truth, that is, an exact understanding of reality; everything was relative in his mind. Comte's only injunction was that every positive theory had to refer on some level to real, concrete phenomena that could be observed directly or indirectly.

Reflecting the impact of nineteenth-century historicism, Comte asserted, moreover, that one could not truly understand a phenomenon without grasping its history. To comprehend the positive method, one had to see its development in the major sciences, each of which had to be considered in relation to its own history, the history of the other sciences, and the history of society in general. The *Cours de philosophie positive* highlighted the history of science, which Comte believed was the most neglected and important part of the development of humanity.

The *Cours* revolved around Comte's renowned 'law of three stages'. This law demonstrated that as the mind went from one mode of thinking to another, it generated a different theoretical system, which in turn shaped the political and social system. There were three stages of development. In the theological stage, people used supernatural ideas, such as one or more gods, to link their observations. There were three phases of this stage: fetishism, which located gods in phenomena; polytheism, which promoted many gods; and monotheism, which held up one god. In a theological society, priests and military men ruled. The theory of divine right was the reigning political doctrine, while conquest was the principal material activity. In the transitional, metaphysical stage, God was replaced by a personified essence or abstraction, such as Nature or Reason. Lawyers and metaphysicians dominated society. The doctrines of popular sovereignty and natural rights were important in politics. More material activities were geared toward production. Finally, in the positive stage, supernatural and abstract forces would be replaced by scientific laws describing how, not why, phenomena functioned. Industrialists and positive philosophers would run a republic. Production would replace conquest as the goal of society.

The *Cours* also introduced Comte's famous 'classification of the sciences'. Each science went through the theological and metaphysical stages and finally became positive according to the simplicity of its phenomena and the distance of these phenomena from humans. Mathematics, which was not a 'direct' natural science, demonstrated the characteristics of the positive method. The positive method was then established in astronomy, physics, chemistry, and biology in that order. Each science depended on the simpler sciences that came before it in the hierarchy of sciences and could not advance to the positive stage until the preceding ones did. Comte argued against reducing one science to another, for each had its own distinctive subject matter and methodology. Moreover, Comte warned against the widespread use of mathematics in scientific investigation, especially in biology and sociology. He particularly disliked the use of statistics in sociology because it neglected the complexities of human nature.

The point of Comte's laws of three stages and classification of the sciences was to demonstrate the inevitable triumph of the positive method in every area of study because all aspects of knowledge were interrelated and the mind sought homogeneity. Now that mathematics, astronomy, physics, chemistry, and biology were positive, it was time for the positive method to be extended to the study of politics and society, which involved the most complex phenomena and the ones closest to humans. The study of politics and society was the last stronghold of theology and metaphysics. Its takeover by science would mark the beginning of a new era in history: the positive era.

In the *Système de politique positive*, written in the 1850s, Comte added a seventh science, morality, which covered the individual. It reflects his growing awareness of the complexity of the human condition and of human nature, especially the role of the emotions in shaping our ideas and actions. He wished to give sociologists more guidance when they had to tackle matters relating to the relationship between society and the individual.

He was also eager to respond to critics, who accused him of endorsing an authoritarian intellectual and political program. Comte admittedly did not see any criticism of his works thanks to his adoption in 1838 of a regime of 'cerebral hygiene', which forbade him from reading contemporary books, journals, and newspapers (except scientific bulletins). He would not brook any kind of debate on his ideas and insisted that dogmatism was the normal state of the human mind. Yet aware of the negative response to his own dogmatic pronouncements and to his support of Napoleon III, he recognized that he had a reputation as an authoritarian thinker. To counter the notion that he neglected the importance of the individual, he created this new science of morality and repeatedly insisted that he respected individuals, who in his mind, were interdependent parts of an organic whole, society. It was the excessive individualism of his period that he claimed was the object of his ire.

Sociology

Comte named his science of society 'sociology' in 1839, giving it a distinctiveness and importance in his system that it never had in Saint-Simon's work. Once it was created, positivism, the system that encompassed all scientific knowledge, would be unified and complete because all our concepts would be scientific and thus homogeneous. In addition, sociology would help unite knowledge by making all the sciences focus on improving humanity and its conditions of existence.

Comte's sociology had two parts, social statics and social dynamics. They stressed the interconnectedness of the members of the human species in space and time. Both had a moral mission, that of countering the egoism and selfishness of the modern age.

Social statics was devoted to studying the social order. It investigated the elements that kept society together. One crucial element was the family, which taught the importance of love as the key to moral self-improvement. This love was transferred to one's family and eventually to all of humanity. In short, social statics cultivated an individual's feelings of solidarity with other members of society. Although Comte's atheism was unusual in the early nineteenth century, his view of the sanctity of the family and his stance on other moral issues were conventional. He wanted to distance himself from contemporary reformers, such as the Saint-Simonians and Fourierists,

whose questioning of traditional social institutions, such as marriage, hurt their movement.

While social statics cultivated people's sense of solidarity with their contemporaries, social dynamics provided them with a feeling of connection with past and future generations of the human species. Social dynamics focused on social development, which Comte considered continuous, necessary, orderly, and limited. Each social state grew out of the preceding one and produced the next social configuration. Reflecting Comte's approval of different aspects of Lamarckism, the crucial feature of this development was that through exercise, the unique characteristics of the human species – intelligence and sociability – became more dominant within both the individual and society. The growth of reason and altruism ensured progress. Reflecting this progress, the law of three stages was the main principle of social dynamics.

In addition to sketching the two divisions of sociology, Comte delineated the new science's methods: observation, experimentation, and comparison. Because every aspect of society had many connections, Comte argued that observations should be varied. Sociologists should observe ordinary events, common customs, languages, different types of monuments, and other mundane social phenomena. Comte's grasp of the significance of everyday life has been buttressed by recent trends in social and cultural history. As for experimentation, Comte maintained that its use posed problems in sociology because of the impossibility of isolating any of the circumstances or consequences of a phenomenon's actions. He explained that a sociologist, like a biologist, should study pathological cases, which were forms of indirect experimentation. Because the pathological was simply a variation of the ordinary, examining periods of chaos, such as a revolution, provided clues to what was normal. In short, by studying social disorder, sociologists would gain insights into the laws of social harmony and history. As for comparison, Comte believed in the importance of comparing human and animal societies, different existing states of human society (i.e., savage and civilized peoples), and different consecutive social states.

The law of three stages of social dynamics and the comparative method of social statics point to the importance of history in Comte's system. He did not include history as one of the fundamental sciences, but he did think that the historical method was sociology's chief means of scientific investigation. History helped provide people with a sense of social solidarity and continuity, in short a feeling for humanity in the world and an understanding of their own role in its evolution.

Sociology was not simply a science; it had a practical, political mission, that of solving social problems. Because everyone would agree on the most significant intellectual and moral principles, the establishment of sociology would lead to the intellectual consensus required to build a stable industrial society. Like other scientific laws, sociological laws would facilitate predictions about social phenomena in the future. Armed with this knowledge, sociologists could construct policies to deal with social problems. In short, sociology would provide the rational principles necessary to reform society.

Seeking to go beyond the bitter partisan politics that marked the postrevolutionary era, Comte offered sociology as a way to create a society that would appeal to both the right and left. It would please the former by stressing order and appeal to the latter by promoting progress. Sociology could unite order and progress because social life, when dominated by universal love, would have order as its basis and progress as its objective.

Positive Politics and the Religion of Humanity

As mentioned above, Comte believed that the unique characteristics of the human species, that is, intelligence and sociability, developed through exercise within both the individual and society. This theory reflects Comte's multidimensional view of human nature, which was reinforced by romanticism. To him, the emotions were closely connected with the intellect, which they simulated. Because social needs were both intellectual and emotional, social reorganization had to involve the heart as much as the mind. In the first phase of his development, Comte emphasized the importance of creating a system of ideas that could be shared by all people. In this positivist system, sociology led everyone to focus their thoughts on the needs of society, or humanity. In his second phase, Comte stressed the necessity of reinforcing the intellectual consensus by encouraging emotional ties among individuals through the Religion of Humanity. Everyone should not simply think about Humanity but love it as well. In short, an increase in intellectual well roundedness would lead to a growth of altruism. Finally, moved by their love for Humanity, individuals would act to improve themselves, their society, and their material or natural environment.

Comte's promotion of universal love was inspired by his horror of the abuses of modern industrial and capitalist society. He believed that the division of labor, which was necessary for progress, increased specialization and made people egotistic, narrow-minded, and selfish. The lower classes suffered the most from the indifference of the rest of society. Their conditions of existence were deplorable. Existing political parties, tied to the old theological and metaphysical principles, could not solve social problems.

Comte created a secular religious culture to bring all individuals together, including the workers, whom he found superior to their employers. He took much criticism for reviving religion after he had called for the end of theology. However, considering its Latin root, '*religare*', Comte pointed out that religion did not have to relate to one or more deities but could be reduced to its essential function of regulating and bringing together with a view to improving society in some way. To him, religion regulated both private life and social existence, which were connected. Inner harmony was linked to social harmony, and both led to health and happiness.

Religion united the mind, heart, and activity, the three aspects of human existence. Stimulated by religion, the mind and heart could work together to vanquish the rampant egoism of society. Both shaped activity, for sentiments and opinions determined and influenced what we did, respectively. Given that the positivist regime cultivated altruism, activity would be directed toward social ends, and individuals would increase their cooperation to achieve those goals. Religion entailed this social cooperation as well as the cooperation between the mind and heart. Because intellectual and emotional capacities were developing and working together and individuals were increasingly cooperating, Comte announced his new 'law' that

human nature and society were becoming more and more religious. People were becoming more sympathetic emotionally, more synthetic intellectually, and more synergetic in terms of activity. The traditional battle between the intellect and the emotions was over.

Enthusiastic about his new Religion of Humanity, which had traces of the Catholicism of his youth, Comte designed new sacraments, temples of Humanity, a list of essential books ('The Positivist Library'), and a calendar of secular saints. Included in the Positivist Calendar were such luminaries as Aristotle, Caesar, Dante, Shakespeare, Joan of Arc, and Heloise. The cult of Humanity, including a cult of 'Woman', would honor individuals who had contributed to civilization with a continuous series of festivals. Comte deliberately constructed a collective memory. However, he also wanted to ensure that individuals who had not achieved public fame lived on in the memory of their families or others whom they had touched in their lives. In this way, they would gain immoratlity, despite the fact that the notion of heaven would disappear.

Comte also recognized that political action in the new age would consist of using education and the arts to shape collective sentiments, beliefs, and representations. Outlining the entire curriculum, he insisted that public schools offer boys and girls of all classes a broad education that cultivated feelings of sociability besides an understanding of the seven sciences. Having lived under Napoleon I and Napoleon III, both of whom used iconography to further their agendas. Comte also grasped the importance of visual culture in promulgating political and social messages. To that end, he had his portrait and bust done by Antoine Etex to immortalize his own image, he devised positivist flags with a young mother representing Humanity, he designed Temples of Humanity, and he made up simple charts of the human mind. In addition, he wanted the color of the positivist movement to be green, in contrast to the white of the royalists and the red of the radical revolutionaries. People might be brought together by common beliefs, but emotional bonds, developed with the aid of visual identifiers, were also important. Every means possible was to be harnessed to create a religious culture that could maintain social cohesion.

The Religion of Humanity also assumed features of fetishism, the most emotionally intense and concrete form of religion. Comte was one of the first thinkers to celebrate primitive religion and see its benefits. Unlike modern men, fetish worshipers, according to Comte, admired what was concrete and showed much regard for the natural world. Comte called the Earth, the 'Great Fetish'. Like the 'Great Being', that is, Humanity, the 'Great Fetish' ruled our lives and demanded respect. Indeed, Comte's attitude toward the Earth helps to explain why he chose green as the color of his movement. Positivism, the most advanced stage of history, would lead full circle to the beginning of civilization, reflecting his increasing doubts about the benefits of science and progress.

In his *Synthèse subjective*, he pointed out the dangers of studying mathematics, the science that he had taught since he was a young man. He argued that in order to counter its aridity, it should be injected with sentiments by adding graphic and phonic elements. Poetry, in particular, should be connected to the sciences. By the end of his life, he was hopeful that positive philosophers, generalists with knowledge of all the sciences, would become poets, creating verses that celebrated Humanity.

In short, although Comte's doctrine was scientific in nature, he insisted on the limits and dangers of the sciences in promoting specialization and narrowness, and he recognized the importance of practical activity to which most people were drawn. To prevent the emergence of a dictatorship of scientists, Comte recommended that there be two separate powers that would hold each other in check. The temporal power would be composed of industrialists, who would control the practical, material activities of society. The spiritual power would consist of positive philosophers, who would govern the world of ideas, opinions, and feelings. The philosophers' chief assistants would be workers, who would be finally incorporated into society, and women, who would make nurturing the goal of public life and would shape public opinion through salons. Women and workers represented the oppressed elements of society that were preserved from the shallow, materialist culture of the day. Both needed to be treated with respect for the advent of positivism to occur.

Comte was eager to address the oppressed not only to attract more adherents to the positivist movement, but to stress the importance of our common humanity. He took elements of the socialist and feminist agendas to appeal to the working class and women and hoped his privileging of fetishism would help bring him the support of 'primitive' races, including blacks. He frequently attacked slavery. Although positivism was often used later during the Third Republic to justify the French 'civilizing mission', Comte also condemned nationalism and imperialism, including France's takeover of Algeria. Anything that divided humanity instead of bringing it together incurred his disapproval. To display positivism's esteem for other faiths, he frequently commended Mohammed and Islam.

To create a more pacifist, more tolerant world, Comte argued in favor of breaking up large states into small republics, which were akin to city-states. Ideally, there would be 500 of these small republics throughout the globe. Because they would be about the same size, with 1–3 million people, and they would share the belief system of the Religion of Humanity, there would be less chance of war. The capital of this positivist world would be Constantinople, the city that Comte believed could best bring together the East and West.

Assessment and Influence

Comte's system was a grand synthesis designed to appeal to diverse philosophical and political positions. It incorporated ideas from many European thinkers, including Bacon, Descartes, Hume, Montesquieu, Condorcet, the Ideologues, the Scottish Enlightenment philosophers, Maistre, Gall, Kant, and Herder. In terms of politics, Comte supported the moral concerns associated with the counterrevolutionaries and remained committed to the leftist tradition that political authority must aid the people. In searching for a synthesis that was both objective and subjective in nature, Comte pursued the middle way between extremes in all realms. He sought to balance the needs of the heart and those of the mind, rationalism and empiricism, induction and deduction, materialism and idealism, religion and science, feminism and patriarchalism, democracy and authoritarianism, and individuality and sociability.

Yet on balance, Comte's philosophy tends to be authoritarian, dogmatic, and grandiose. He seemed eager to create a conformist, homogeneous society. Many of his so-called scientific assertions, including the law of three stages, are replete with assumptions. Paradoxically, several of the strengths of his system, that is, the diffuseness of his endeavor and his tendency to see the complexity of each problem, turned out to be weaknesses because they caused him to lose support in his own day. The left found him illiberal and religious, while the right considered him blasphemous and anarchical.

Nevertheless, Comte's thought was extremely influential. Because it was vast and complex, people took from it what they wanted. During his lifetime, it attracted numerous followers in France, the United Kingdom, the United States, and Latin America. His two most famous disciples, John Stuart Mill and Emile Littré, made positivism into a manifesto for the scientific age. George Henry Lewes popularized the positivist doctrine as did Harriet Martineau, who wrote an abridged translation of the *Cours* (1853), which Comte praised as superior to his original work.

Individuals on the left and right employed elements of positivism to bolster their diverse political causes. Left-leaning politicians of the Third Republic used positivism in their campaign against the Church. Enthusiasts included political leaders Jules Ferry and Léon Gambetta, and the writer Emile Zola. On the other side of the political spectrum, Charles Maurras employed positivist principles to enhance the proto-fascist doctrine of the *Action Française*. In Latin America, positivism proved very attractive to republicans, especially in Brazil, whose flag displays Comte's motto "Order and Progress". But Positivism was often used to bolster dictatorial regimes, such as that of Porfirio Diáz in Mexico. In the United States, reformers, such as David Croly, Thaddeus Wakeman, Edward Bellamy, Herbert Croly, and Lester Ward, embraced positivism as an alternative to religion and appreciated his emphasis on an intellectual elite who could push for social change.

Comte also influenced the humanities and social sciences. Philosophers of science have been interested in his approach to hypotheses and his explanation for the uniqueness of science. His work may have influenced Claude Bernard, Marcelin Berthelot, Paul Janet, Ernst Mach, Wilhelm Ostwald, and Pierre Duhem. Logical positivists, such as A.J. Ayer, Otto Neurath, Rudolf Carnap, Moritz Schlick, and Hans Reichenbach, shared Comte's concern with problems of verification and grappled with his thought. Positivism long dominated European historiography; Thomas Buckle is the usual example of a positivist historian. Lucien Lévy-Bruhl introduced many of Comte's ideas into anthropology. Through Emile Durkheim, positivism influenced sociology, which adopted Comte's concern with social solidarity and harmony. Other social scientists affected by Comte include Alfred Espinas, Herbert Spencer, Hippolyte Taine, Ernest Renan, Charles Booth, and Patrick Geddes. Comte's thought remains relevant today, when people still grapple with many problems that he highlighted, such as the role of affectivity in education and the sciences, the roots of moral crises, and the significance of collective memory. Particularly at a time when global interests are paramount, Comte's call for a commitment to humanity as a whole and the environment bears renewed examination.

See also: Altruism and Prosocial Behavior, Sociology of; Collective Memory, Anthropology of; Elites: Sociological Aspects; Emotions, Sociology of; Historical Sociology; Human Sciences, History of; Intellectuals, Sociology of; Modernization, Sociological Theories of; Positivism, Sociological; Science, History of; Secular Religions; Sociological Theory; Sociology, History of.

Bibliography

Bensaude-Vincent, B., 2000. L'Opinion publique et la science: A Chacun son ignorance. Institutd'édition Sanofi-Synthélabo, Paris.
Bourdeau, M., 2006. Les Trois Etats: Science, théologie et métaphysique chez Auguste Comte. Editions du Cerf, Paris.
Bourdeau, M., 2011. Auguste comte. In: Edward N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/archives/sum2011/entries/comte/.
Braunstein, J.-F., 2009. La Philosophie de la médecine d'Auguste Comte; Vaches carnivores, Vierge Mère et morts vivants. PUF, Paris.
Comte, A., 1929. Système de Politique Positive, 4 vols. Société positiviste, Paris.
Comte, A., 1998. Cours de Philosophie Positive, 2 vols. Hermann, Paris.
Fedi, L., 2004. Comte. Les Belles Lettres, Paris.
Gouhier, H., 1997. La Vie d'Auguste Comte (second rev. ed.). Vrin, Paris.
Grange, J., 1996. La Philosophie d'Auguste Comte: Science, Politique, Religion. Presses Universitaires de France, Paris.
Kremer-Marietti, A., 1982. Entre le Signe et l'histoire: L'Anthropologie d'Auguste Comte. Klincksieck, Paris.
Lenzer, G., 1998. Auguste Comte and Positivism: The Essential Writings. Transaction Publishers, New Brunswick, NJ.
Petit, A. (Ed.), 2003. Auguste Comte, Trajectoires Positivistes 1798–1998. L'Harmattan, Paris.
Pickering, M., 1993/2009. Auguste Comte: An Intellectual Biography, 3 vols. Cambridge University Press, Cambridge, UK.
Scharff, R.C., 1995. Comte after Positivism. Cambridge University Press, Cambridge, UK.
Wernick, A., 2001. Auguste Comte and the Religion of Humanity. Cambridge University Press, Cambridge, UK.

Relevant Websites

http://membres.multimania.fr/clotilde/index.xml – "The Auguste Comte Website", International Positivist Society.
http://www.augustecomte.org – Website of the Maison d'Auguste Comte, the museum in Paris devoted to his memory.

Concentrated Disadvantage

Paul A Jargowsky and Natasha O Tursi, Center for Urban Research and Education, Rutgers University – Camden, Camden, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Concentration of disadvantage' refers to the phenomenon of spatial clustering of economically or socially disadvantaged individuals within a set of neighborhoods and the resulting feedback effects that exacerbate the problems of poverty. The central concept is the disproportionate share of disadvantaged persons in geographic neighborhoods lacking critical resources, such as access to quality healthcare and quality education. Areas of concentrated disadvantage around the world are referred to by different terms such as ghettos, barrios, slums, favelas, and banlieues, each with a different set of connotations. Concentration and its effects are a critical component in the study of poverty and disadvantage.

Introduction

'Concentration of disadvantage' refers to the phenomenon of spatial clustering of economically or socially disadvantaged individuals within a set of neighborhoods and the resulting feedback effects that exacerbate the problems of poverty. The central concept is a disproportionate share of disadvantaged persons in geographic neighborhoods lacking critical resources, such as access to quality healthcare and quality education. Such neighborhoods typically have high crime rates, in part due to a flourishing underground economy (Anderson, 1990).

To characterize areas of concentrated disadvantage, social scientists use variables that are indicators of elevated levels of neighborhood socioeconomic distress. Depending on the availability of data, these variables include the poverty rate, median income, joblessness, housing quality, and family structure. These data points are individual-level characteristics that when aggregated to the neighborhood level indicate concentrated disadvantage. In the United States, which has an officially sanctioned poverty line, concentration of disadvantage is sometimes measured by examining the spatial concentration of poor persons. Jargowsky (Jargowsky, 1997) defines the 'concentration of poverty' as the proportion of an area's poor persons residing in census tracts where the official poverty rate is 40% or more. On average, in these high-poverty neighborhoods, fewer than half of adult males are in the labor force and more than half of children live in single-parent families (Jargowsky, 2013).

Exploiting the correlation of poverty with other indicators of socioeconomic disadvantage, the Project on Human Development in Chicago Neighborhoods created a multidimensional measure of concentrated disadvantage based on census tract levels of welfare receipt, poverty, unemployment, female-headed households, percentage black, and density of children (Sampson, 2012: 100). Neighborhoods with higher disadvantage are marked by lower collective efficacy and greater exposure of neighborhood residents to criminal victimization. Neighborhoods of concentrated disadvantage are highly durable over time, rarely experiencing fundamental socioeconomic progress (Sampson, 2012), outside of a few well-publicized cases of gentrification. Moreover, residence in high-poverty neighborhoods is highly persistent and is transmitted intergenerationally, especially for African-Americans (Sharkey, 2008).

Concentration of poverty and disadvantage must be distinguished from poverty and disadvantage per se. A given city or metropolitan area can have a greater or lesser amount of socioeconomically disadvantaged persons. Yet if those persons are spread evenly across the landscape, there would be no concentration. An extremely poor city, therefore, could have no concentration of poverty, while a relatively affluent city in which a small number of poor persons live alone in a specific set of neighborhoods could have total concentration of poverty (Jargowsky, 1997: 17). In practice, however, the level of poverty and the concentration of poverty are frequently correlated.

Concentration of disadvantage must also be distinguished from the concept of the 'underclass' or 'underclass neighborhoods' (Ricketts and Sawhill, 1988). While no one definition of the underclass ever achieved general acceptance, the term generally refers to people who are not just poor, but fundamentally unable to access the class structure due to structural unemployment, hostility to mainstream values, criminal tendencies, or addictions that render them incapable of holding on to a job (Myrdal, 1963; Glasgow, 1980; Auletta, 1982). Such neighborhoods can develop a 'social milieu' that fosters low labor force attachment, out-of-wedlock childbearing, drug and alcohol use, and other social pathologies (Wilson, 1987: 58). However, the residents of high-poverty neighborhoods are quite diverse, and many if not most do not conform to the underclass stereotype (Jargowsky, 1996).

Areas of concentrated disadvantage are referred to by different terms, each with a different set of connotations. In the United States, segregated black neighborhoods are often referred to as *ghettos*, which, according to Wacquant (2004: 2) is constituted of four elements: stigma, constraint, spatial confinement, and institutional encasement. However, the term is disputed and many areas commonly referred to as ghettos do not strictly meet Wacquant's definition (Small, 2008). Poor neighborhoods composed primarily of persons of Hispanic descent are called *barrios* (Moore and Pinderhughes, 1993). Another term used to describe concentrated disadvantage is *slum*, which is "characterized by overcrowding, poor or informal housing, inadequate access to safe water and sanitation, and insecurity of tenure" (Davis, 2006: 23; Forman, 1971). In other nations, a variety of different terms are used,

such as favelas, banlieues, shantytowns, etc., and these terms cannot be assumed to be equivalent in either form or function (Hutchison and Haynes, 2012).

Examples of Areas of Concentrated Disadvantage around the World

In 'Planet of Slums', Mike Davis (2006) discusses the history of slums internationally, where "classic [Nineteenth Century] slums were notoriously parochial and picturesquely local places…characterized by an amalgam of dilapidated housing, overcrowding, disease, poverty, and vice" (22). He argues that the British empire in the colonization period contributed greatly to slums in the Third World, calling the British "the greatest slum-builders of all time" (52). Ostensibly, their policies in Africa forced the local labor force to "live in precarious shantytowns on the fringes of segregated and restricted cities" (52). Similarly, they provided minimal infrastructure to native neighborhoods in India, Burma, and Ceylon, which increased death tolls related to disease outbreaks such as cholera and plague. Davis asserts that state-assisted housing in the Third World is for the most part detrimental to health and safety because public services are lacking (69).

Slums have been growing since the 1970s. In fact, most urban poor now live in slum communities on the periphery of cities in developing countries (Drakakis-Smith, 2012a: 99), in edge developments that are either squatter settlements or pirate subdivisions, both generating a shantytown landscape (Davis, 2006: 37–38). With "more than 200 000 slums on earth, ranging in population from a few hundred to more than a million people" (37), "Megaslums arise when shanty-towns and squatter communities merge in continuous belts of informal housing and poverty, usually on the urban periphery" (Davis, 2006: 26). Examples include barrios in Mexico City, Mexico; condos in Lima, Peru; kampungs in Jakarta, Indonesia, and Kuala Lumpur, Malaysia; and slums in Nairobi, Africa, Cairo, Egypt, and countless cities in developing countries across the world (27). The world's highest percentage (62% of the urban population) of slum dwellers is in sub-Saharan Africa, Sierra Leone, Sudan, and the Central African Republic (UN-Habitat, 2013), Ethiopia, Chad, Afghanistan, Nepal, Bombay, Mexico City, Dhaka, Lagos, Cairo, Karachi, Kinshasa-Brazzaville, Sao Paulo, Shanghai, and Delhi. In Bombay, for instance, there are an estimated 10–12 million squatters and tenement dwellers, (Davis, 2006: 23). Millions of people live in substandard settlements 'concentrated' in each of these cities. "The fastest growing slums are in the Russian Federation and the former Soviet republics" (Davis, 2006: 24, referencing e.g., 2003 UN-HABITAT case-studies).

Areas of concentrated disadvantage have taken various forms in different parts of the world such as primitive dwellings (e.g., Mexico City), favelas (e.g., Rio, Sao Paulo), illegal slums in Beijing where rural migrants live in crowded conditions, sprawling squatter settlements in South Asia, Indian slums in Mumbai and New Delhi, fetid slums in Nairobi and Mombasa, and areas in sub-Saharan Africa where hundreds of millions of slum dwellers live devoid of universal primary education and with high infant mortality rates (Davis, 2006: 16–19). Other examples are over-crowded rooms that the poor rent in inner-city tenements in Sao Paulo, mostly prior to the favela boom of the 1980s (Davis, 2006: 34).

There is great variation in the type of dwellings people inhabit in areas of concentrated disadvantage. The *chwals* of Mumbai, according to Davis (2006), are "mostly dilapidated, one-room rental dwelling that crams a household of six people into 15 square meters" (34). Clusters of the *jhuggi-jhonpri* (hutment) squatter settlements that are inhabited by an estimated 3 million (or one-third) of New Delhi's population are typically constructed out of mud, brick, and tin-thatched roofs and are often located right next to five-star hotels or modern high-rise buildings (Dupont et al., 2000). The *callejones* of Lima that are built to be rented by the poor are unstable, substandard dwellings. Thousands of squatters live in abandoned buildings and factories of Buenos Aires. Other examples include roof dwellers in Cairo and boat dwellers (floating slums) in Southeast Asia (Davis, 2006: 34–36). Squatters are the "pioneer settlers of swamps, floodplains, volcano slopes, unstable hillsides, rubbish mountains, chemical dumps, railroad sidings, and desert fringes" (Davis, 2006: 121). Over half of Johannesburg's nonwhite population dwells on "dangerous, unstable dolomitic soil contaminated by generations of mining" (Davis, 2006: 122). Similarly, the soil upon which the favelas of Sao Paulo and Rio de Janeiro, Caracas, or Manila are built is unsuitable for squatter settlements because of its flood plains.

Urban slums around the world have proliferated in conjunction with rapid urbanization in areas least able to cope with the rise in population. According to the 2009 UN Habitat Global Report on Human Settlements (UN Habitat, 2009), for the first time in history, most of the world's population lived in cities in 2008, and the majority and almost all of new global population growth is predicted to occur in cities and in the global South. With the rise of migrants in slum and squatter districts who often settle in cheap, substandard housing (Drakakis-Smith, 2012b: 16), a rise in urban concentrations of poverty and inequality is predicted and will likely exceed the 32% of the world's urban population (924 million people in 2001) that lives in slums on extremely low incomes (UN Habitat, 2009: 2263).

Marx et al. (2013) point out that contemporary slums of the developing world may constitute 'poverty traps' and 'policy traps' (Marx et al., 2013). Their argument is based on historical and contemporary facts such as surveys collected in slums of Bangladesh, India, Kenya, and Sierra Leone, as well as extensive fieldwork of the Kibera slum in Nairobi. Their findings suggest that slums have often been inhabited by several generations, and as such may constitute 'poverty traps', with it bringing a host of associated disadvantages. For example, slums lack public and private investment, and since housing is non-permanent and tenants have no ownership over the land, they are not incentivized to improve their living conditions. As a result, as more and more people settle into these substandard dwelling areas, overcrowding causes health and sanitation issues and potentially disease outbreaks.

Although slum dwelling has been characterized as intergenerational, slums continue to expand based on rural migration. Regardless, the economic and social mobility of slum dwellers is undermined by high rents to live in close proximity to the city. Ironically, urban slum dwellers often fare worse

than their rural counterparts, as the former have a higher rent burden than their rural counterparts who often do not pay rent. Failing governance, corruption, and otherwise inefficient allocation mechanisms frequently leave room for 'slumlords', local chiefs, gangs, or drug cartels to manage and control the area and financially exploit residents. To be sure, slums are areas of concentrated disadvantage and according to Marx et al. (2013) often pose a 'policy trap' (Marx et al., 2013). Indeed, slum neighborhoods exist 'in the margins' and typically do not receive public investment such as urban planning or economic development. In addition, slum populations are rarely captured by the census that leads to undersampling and other statistical problems that obfuscate reality and make even potential policy interventions difficult.

Concentration of Poverty in the United States

Neighborhoods in which more than 40% of the population have incomes below the federal poverty line are characterized by high rates of single parenthood, low levels of educational attainment, and detachment from the labor force (Jargowsky, 1996). Further, the lack of opportunities and role models in such neighborhoods contributes to a social milieu in which alcohol and drug abuse, gang membership, criminality, and other self-defeating behaviors become more prevalent (Wilson, 1987, 1996). Thus, a good measure of the concentration of the poverty is the percentage of the poor in a given area that lives in neighborhoods, defined as census tracts (Census tracts are small geographic units delineated by the US Census Bureau that contain 4000 persons on average (White, 1987).), in which 40% or more of the population is classified as poor according to the US federal poverty line (Jargowsky and Bane, 1991). (As of 2014, the federal poverty line is $24 000 per year for a family of four (Department of Health and Human Services, 2014).)

By this measure, the concentration of poverty rose dramatically during the 1970s and 1980s (Jargowsky and Bane, 1991; Jargowsky, 1997; Kasarda, 1993). The increases were fastest in the Midwest and Northeast, areas that suffered from rapid deindustrialization and rising poverty. In fact, concentration of poverty is closely correlated with overall economic conditions at the metropolitan level. Not surprisingly, therefore, the concentration of poverty declined substantially during the 1990s, a decade of rapid economic growth and low unemployment (Jargowsky, 2003; Kingsley and Pettit, 2003). The number of high-poverty census tracts – those with poverty rates of 40% or more – fell 26.5%, from 3417 in 1990 to 2510 in 2000 (Jargowsky, 2003).

Since 2000, the spatial concentration of poverty has surged again. (The results presented here draw heavily on Jargowsky (2013, 2014b).) The count of high-poverty census tracts increased by 800 (32%) between 2000 and the 2005–09 American Community Survey data to nearly the level of 1990, even though these data span more than 3 years before the financial crisis hit in late 2008. The latest data, spanning 2008–12, show that the number of high-poverty tracts has increased to 4054, a 62% since 2000 (Jargowsky, 2014b). The number of people living in high-poverty neighborhoods – census tracts with poverty rates of 40% or more – increased from 7.1 to 12.4 million, a 72% since 2000.

The racial/ethnic composition of the population of high-poverty neighborhoods has changed due to different rates of growth in among demographic subgroups. For example, while the population in high-poverty neighborhoods has risen for all racial and ethnic groups, it has grown fastest among non-Hispanic whites; the count of whites in high-poverty areas more than doubled, increasing 122% between 2000 and the most recent data. In comparison, the black and Hispanic population of those areas increased 51 and 74%, respectively. As a result, whites now make up 26% of the population of high-poverty tracts, compared to 20% in 2000. The black population share has declined from 42% in 2000 to 37% in 2008–12. The Hispanic share has remained constant at approximately 31%.

While the number of high-poverty neighborhoods has risen, it is also true that the number of poor persons has risen overall, both due to population growth and the general rise in the poverty rate since 2000. To ask whether poverty is more concentrated, we must examine the percentage of the poor living in high-poverty areas. In the average metropolitan area, the percentage of the poor living in high-poverty neighborhoods rose from 11.6 to 15% between the 2000 and 2008–12 period, although the concentration of poverty is still lower than that in 1990, when it was 16.8%. It rose for all groups, though the increase was fastest for whites – rising from 5.1 to 8.4%. Despite the recent increase in whites living in high-poverty areas, concentration of poverty still differentially affects members of minority groups. The black poor are nearly three times more likely than the white poor to live in high-poverty neighborhoods, 24.6% compared to 8.4%. The Hispanic poor are twice as likely as the white poor to be in neighborhoods of concentrated poverty.

Causes of the Concentration of Disadvantage

One of main causes of the concentration of poverty in urban centers in the United States is deindustrialization, which eroded the historic job base of many older industrial cities (Bluestone and Harrison, 1982; Gillette, 2011). Ironically, deindustrialization occurred just after a wave of building large-scale public housing projects, which anchored to the poor to areas where jobs for less well-educated workers were becoming nonexistent (Venkatesh, 2002).

Another important factor in the creation of high-poverty neighborhoods was the vast outmigration to the suburbs that began after World War II and accelerated in the 1960s when urban riots, violent crime, and racial tensions came to the fore (Mieskowski and Mills, 1993; Massey and Denton, 1988). Two factors are key to understanding the effect of the suburban explosion on central cities. First, suburban growth produced housing units much faster than needed to accommodate metropolitan population growth. Between 1970 and 1990, for example, most large central cities and some older inner-ring lost population rapidly, while successive new suburban rings gained population (Jargowsky, 2002). Second, zoning was used to prevent low-income persons and racial minorities from leaving the confines of the central city (Galster, 2012; Fischel, 2004). In effect, while discrimination on the basis of race is strictly illegal,

discrimination on the basis of income, through exclusionary zoning, is the law of the land. The result is selective outmigration, segregation by race and class, depopulation of the inner city and older suburbs, and concentration of poverty. In additions to limits on mobility, racial and class differences in neighborhood choice patterns tend to reinforce and sustain existing segregated patterns (Sampson and Sharkey, 2008).

In other nations, the causal sequence can be quite different. Some of the main contributing forces to slum expansion around the world are the rapid rate of rural to urban migration due to industrialization, the lack of neighborhood planning, and overpopulation in general (Henderson, 2010; UN Habitat, 2009). Most slum dwellers are rural residents who migrated to cities in search of employment opportunities, education for their children, and improved health and social services. The kinds of jobs that migrants seek are in manufacturing and the service sector. Although urbanization and economic development are intertwined, they are not causally related. Rapid urbanization does not necessarily lead to per capita income growth (Henderson, 2010). As a result, migrants looking for work and improvement in quality of life who move to the urban areas may wind up in similar or worse conditions of poverty then where they left.

Consequences of the Concentration of Disadvantage

Areas of concentrated disadvantage where people are subjected to adverse economic and physical conditions have negative effects on their residents, and there are also larger societal costs associated with slums. Nunes and Veloso (2012) write that the Brazilian favelas in Rio de Janeiro and other large cities "arose as a marginal space where economically and socially excluded individuals gathered – a function that has changed very little" (Hutchison and Haynes, 2012). Exclusion, it would seem, is dangerous when coupled with inequality. Gilbert (2012) points out that inequality has increased in Latin American cities where the institutions create a second-class citizenship status for the poor, and "the education that you get marks you for life, determining your occupational possibilities" (quoting Roberts, 2004: 196 in Hutchison and Haynes (2012: 212)).

Aside from the unsafe conditions of these settlements based on local geological and climatic hazards, these communities are undesirable places to live for a variety of reasons. As such, they are often marked by alcohol stores, litter, stagnant water, open sewage, etc. Such adversity often has detrimental effects. For instance, in refugee camps of Gaza and Zaire, the substandard sanitary conditions lead to cholera and other disease outbreaks. Not surprisingly, infant mortality rates are often much higher in unsanitary living areas.

Social segregation is accentuated through inequality, and rich areas are segregated from poor ones, for example, in Bogota, Rio de Janeiro, Caracas, Santiago, etc. creating spatial barriers also through gated communities. There has been a historical division into a 'two-class' society in which the well-off children attend private institutions and elite private recreational clubs. Gilbert (2012) argues that there is a different cultural, historical context in Latin America where the rich are spatially segregated (e.g., gated communities); however, settlements inhabited by the poor are not necessarily cut off from contextual educational and employment opportunities.

Relating to spatial segregation in France, Castañeda (2012) points out that the banlieues at the outskirts of Paris are comparable with the US ghettos in that they have produced a ghetto subculture inclusive of French rap and youth slang; are home to immigrants; and are marked by high unemployment rates, violence, lacking social integration and equality, and segregated, much like black ghettos in the United States (Castañeda, 2012: 177–179). He argues that relative deprivation is most important for people in ghettos and banlieues, and "long-lasting ethnic segregation is not voluntary but imposed from the outside" (Castañeda, 2012: 184).

The question of whether poor neighborhoods have independent effects on residents has been extensively studied in the United States, with mixed results (Duncan et al., 1997). Low-income neighborhoods have lower levels of education, employment, and earnings and higher rates of poverty, single-parent families, and other social problems. However, these characteristics may be compositional rather than causal; in other words, people with those characteristics may migrate into high-poverty areas to obtain cheap housing, or people who lack those characteristics may move out (Tienda, 1991). The question is whether these concentrations of poverty have negative effects on social and economic outcomes after controlling for individual and family traits. Increasing evidence suggests that it does have negative effects, particularly when studies take into account long-term and multigenerational exposure to disadvantage neighborhood contexts (Sampson et al., 2008; Sharkey and Elwert, 2011).

Concentration of disadvantage has implications for educational outcomes because schools are creatures of neighborhoods (Levy, 1995). Often, the relationship is legally encoded in school attendance zone boundaries. Even schools that draw on larger and less precise areas, such as magnet and charter schools, commuting time and transportation costs often restrict attendance to those who live relatively nearby. Schools usually closely reflect the racial and economic composition of the surrounding community. When they do differ, public schools will tend to have a greater proportion of minority and low-income children, due to life-cycle differences and differential selection into private schools and home schooling. Thus, schools are often even more segregated by race and income than is the surrounding community (Rivkin, 1994; Jargowsky, 2014a).

As a result, when poor families reside in different neighborhoods than middle- and upper-income families, their children will likely attend different schools than more affluent children. Over time, the schools themselves can become different. Schools in poorer neighborhoods have greater needs than schools with more advantaged children. Teachers and school administrators may develop lower academic expectations when they deal predominantly with poor children, many of whom who do not have resources or support in the home (Farkas, 1996). In some inner-city schools, working hard and getting good grades is derided as 'acting white' (Fordham and Ogbu, 1986; Fryer and Torelli, 2010). Even students who resist caving in to peer pressure

may still be impeded in learning if enough classmates are disruptive and slow the pace of instruction. These so-called peer effects on students have been documented in a number of carefully controlled studies (Hanushek et al., 2003; Summers and Wolfe, 1977; Zimmer and Toma, 2000). Over and above peer effects, neighborhood conditions have spillover effects on academic achievement (Crane, 1991; Crowder and South, 2003; Datcher, 1982; Jargowsky and El Komi, 2011).

A variety of studies have found that neighborhoods matter for child and adolescent development across a variety of developmental outcomes (Leventhal and Brooks-Gunn, 2000). For example, a child's IQ at 36 months of age is related to the presence of affluent families in the child's neighborhood after controlling for family income, mother's education, family structure, and race (Brooks-Gunn et al., 1993: Table 2). Children raised in families exposed to high-poverty neighborhoods for two consecutive generations score more than half a standard deviation lower on measures of cognitive ability, even after controlling for a wide range of family and individual characteristics (Sharkey and Elwert, 2011). Girls with fewer affluent neighbors initiated sexual activity earlier and were more likely to have out-of-wedlock birth, again controlling for family characteristics (Brooks-Gunn et al., 1993; Browning et al., 2004; Crane, 1991). Children with a high proportion of poor neighbors have more behavioral problems, lower self-esteem, and more symptoms of depression (Chase-Lansdale and Gordon, 1996; Chase-Lansdale et al., 1997). The Project on Human Development in Chicago Neighborhoods found that exposure to nearby homicides has an immediate and substantial effect on children's performance reading and vocabulary assessment (Sharkey, 2010).

The concentration of poverty often makes for an unhealthy environment with few parks and recreational resources, greater pollution, more alcohol outlets, more advertising for alcohol and tobacco, and less availability of healthy foods (Barbeau et al., 2005; Gordon-Larsen et al., 2006; Hackbarth et al., 2001; Lopez and Hynes, 2006; Wolch et al., 2005). Residents of disadvantaged neighborhoods suffer higher rates of communicable diseases such as tuberculosis (Acevedo-Garcia, 2000), premature birth (Osypuk and Acevedo-Garcia, 2008), self-report of poor health (Subramanian et al., 2005), diabetes (Ludwig et al., 2011), and obesity (Lopez and Hynes, 2006; Ludwig et al., 2011). Residence in economically and socially isolated census tracts increases the probability that adolescents will engage in health-risk behaviors (Frank and Bjornstrom, 2011; Frank et al., 2007).

Conclusion

Concentration of disadvantage, concentration of poverty, and related concepts are ways of thinking about the spatial dimension of disadvantage. The key insight is that the problems of poverty and other social disadvantages are magnified when disadvantage is clustered at the neighborhood level. Indeed, many of the problems associated with poverty – exposure to crime and violence, failing schools, dilapidated housing – are actually neighborhood characteristics. Hence, the study of poverty and disadvantage would be incomplete without consideration of concentration and its effects.

See also: Community Sociology; Human Ecology; Neighborhood Effects; Poverty, Culture of; Poverty, Geography of; Spatial Equity; Underclass; Urban Sociology.

Bibliography

Acevedo-Garcia, Dolores, 2000. Residential segregation and the epidemiology of infectious diseases. Social Science & Medicine 51 (8), 1143–1161.

Anderson, Elijah, 1990. Street Wise: Race, Class, and Change in an Urban Community. University of Chicago Press, Chicago.

Auletta, Ken, 1982. The Underclass. Random House, New York.

Barbeau, Elizabeth M., Wolin, Kathleen Y., Naumova, Elena N., Balbach, Edith, 2005. Tobacco advertising in communities: associations with race and class. Preventive Medicine 40 (1), 16–22.

Bluestone, Barry, Harrison, Bennett, 1982. The Deindustrialization of America: Plant Closings, Community Abandonment, and the Dismantling of Basic Industry. Basic Books, New York.

Brooks-Gunn, Jeanne, Duncan, Greg J., Klebanov, Pamela Kato, Sealand, Naomi, 1993. Do neighborhoods influence child and adolescent development? American Journal of Sociology 99 (2), 353–395.

Browning, Christopher R., Leventhal, Tama, Brooks-Gunn, Jeanne, 2004. Neighborhood context and racial differences in early adolescent sexual activity. Demography 41 (4), 697–720.

Castañeda, Ernesto, 2012. Places of stigma: ghettos, barrios, and banlieues. In: Hutchison, R., Haynes, B.D. (Eds.), The Ghetto: Contemporary Global Issues and Controversies. Westview Press, Boulder, CO., p. 159.

Chase-Lansdale, P. Lindsay, Gordon, Rachel A., 1996. Economic hardship and the development of five- and six-year-olds: neighborhood and regional perspectives. Child Development 67 (6), 3338–3367. http://dx.doi.org/10.1111/j.1467-8624.1996.tb01917.x.

Chase-Lansdale, P. Lindsay, Gordon, Rachel A., Brooks-Gunn, Jeanne, Klebanov, Patricia, 1997. Neighborhood and Family Influences on the Intellectual and Behavioral Competence of Preschool and Early School-age Children Context and Consequences for Children, vol. 1. Russell Sage Foundation, New York.

Crane, Jon, 1991. Effects of neighborhood on dropping out of school and teenage childbearing. In: The Urban Underclass. The Brookings Institution, Washington, DC, pp. 299–320.

Crowder, Kyle, South, Scott J., 2003. Neighborhood distress and school dropout: the variable significance of community context. Social Science Research 32 (4), 659–698.

Datcher, Linda, 1982. Effects of community and family background on achievement. Review of Economics and Statistics 64 (1), 32–41.

Davis, Mike, 2006. Planet of Slums. Verso, London; New York.

Department of Health and Human Services, 2014. Annual update of the HHS poverty guidelines. Federal Register 9 (14), 3593–3594.

Drakakis-Smith, David, 2012a. Urbanisation, Housing and the Development Process. Routledge.

Drakakis-Smith, David, 2012b. Urbanisation, Housing and the Development Process. Routledge.

Duncan, Greg J., Connell, James P., Klebanov, Pamela K., 1997. Conceptual and Methodological Issues in Estimating Causal Effects of Neighborhoods and Family Conditions on Individual Outcomes Context and Consequences for Children, vol. 1. Russell Sage Foundation, New York.

Dupont, Véronique, Tarlo, Emma, Vital, Denis, France Ambassade (India), Centre for Human Sciences, Institut de recherche pour le développement (France), 2000. Delhi: Urban Space and Human Destinies. Manohar Publishers & Distributors.

Farkas, George, 1996. Human Capital or Cultural Capital? Ethnicity and Poverty Groups in an Urban School District. Aldine de Gruyter, Hawthorn, NY.

Fischel, W.A., 2004. An economic history of Zoning and a cure for its exclusionary effects. Urban Studies 41 (2), 317–340.

Fordham, Signithia, Ogbu, John U., 1986. Black students' school success: coping with the 'Burden of "Acting White"'. The Urban Review 18 (3), 176–206.

Forman, Robert E., 1971. Black Ghettos, White Ghettos, and Slums. Prentice Hall, Englewood Cliffs, NJ.

Frank, Reanne, Bjornstrom, Eileen, 2011. A tale of two cities: residential context and risky behavior among adolescents in Los Angeles and Chicago. Health & Place 17 (1), 67–77.

Frank, Reanne, Cerdá, Magdalena, Rendón, Maria, 2007. Barrios and Burbs: residential context and health-risk behaviors among Angeleno adolescents. Journal of Health and Social Behavior 48 (3), 283–300.

Fryer Jr., Roland G., Torelli, Paul, 2010. An empirical analysis of 'acting White'. Journal of Public Economics 94 (5–6), 380–396.

Galster, George, 2012. Driving Detroit: The Quest for Respect in the Motor City. University of Pennsylvania Press.

Gilbert, A.G., 2012. On the Absence of Ghettos in Latin American Cities. http://discovery.ucl.ac.uk/1379449/.

Gillette Jr., Howard, 2011. Camden After the Fall: Decline and Renewal in a Post-Industrial City. University of Pennsylvania Press.

Glasgow, Douglas C., 1980. The Black Underclass: Poverty, Unemployment, and Entrapment of Ghetto Youth. Vintage Books, New York.

Gordon-Larsen, Penny, Nelson, Melissa C., Page, Phil, Popkin, Barry M., 2006. Inequality in the built environment underlies key health disparities in physical activity and obesity. Pediatrics 117 (2), 417–424. http://dx.doi.org/10.1542/peds.2005-0058.

Hackbarth, D.P., Schnopp-Wyatt, D., Katz, D., Williams, J., Silvestri, B., Pfleger, M., 2001. Collaborative research and action to control the geographic placement of outdoor advertising of alcohol and tobacco products in Chicago. Public Health Reports 116 (6), 558–567.

Hanushek, Eric A., Kain, John F., Markman, Jacob M., Rivkin, Steven G., 2003. Does peer ability affect student achievement? Journal of Applied Econometrics 18 (5), 527–544.

Henderson, J. Vernon, 2010. Cities and development. Journal of Regional Science 50 (1), 515–540. http://dx.doi.org/10.1111/j.1467-9787.2009.00636.x.

Hutchison, Ray, Haynes, Bruce D., 2012. The Ghetto: Contemporary Global Issues and Controversies. Westview Press.

Jargowsky, Paul A., 1996. Beyond the street corner: the hidden diversity of high-poverty neighborhoods. Urban Geography 17, 579–603.

Jargowsky, Paul A., 1997. Poverty and Place: Ghettos, Barrios, and the American City. Russell Sage Foundation, New York.

Jargowsky, Paul A., 2002. Sprawl, concentration of poverty, and urban inequality. In: Squires, Gregory (Ed.), Urban Sprawl: Causes, Consequences, and Policy Responses. Urban Institute Press, Washington, DC, pp. 39–72.

Jargowsky, Paul A., 2003. Stunning Progress, Hidden Problems: The Dramatic Decline of Concentrated Poverty in the 1990s. Living Cities Census Series. Brookings Center on Urban and Metropolitan Policy, Washington, DC.

Jargowsky, Paul A., 2013. Concentration of Poverty in the New Millennium: Changes in the Prevalence, Composition, and Location of High-poverty Neighborhoods. A Report of the Century Foundation and the Rutgers Center for Urban Research and Education. Century Foundation, New York, NY.

Jargowsky, Paul A., 2014a. Segregation, neighborhoods, and schools. In: Lareau, Annette, Goyette, Kimberly (Eds.), Choosing Homes, Choosing Schools: Residential Segregation and the Search for a Good School. Russell Sage Foundation, New York, pp. 97–136.

Jargowsky, Paul A., 2014b. Concentration of Poverty: An Update. The Century Foundation: Blog of the Century. http://tcf.org/blog/detail/concentration-of-poverty-an-update.

Jargowsky, Paul A., Bane, Mary Jo, 1991. Ghetto poverty in the United States: 1970 to 1980. In: Jencks, C., Peterson, P. (Eds.), The Urban Underclass. The Brookings Institution, Washington, DC, pp. 235–273.

Jargowsky, Paul A., El Komi, Mohamed, 2011. Before or after the bell? School context and neighborhood effects on student achievement. In: Newburger, Harriet B., Birch, Eugenie L., Wachter, Susan M. (Eds.), Neighborhood and Life Chances: How Place Matters in Modern America. University of Pennsylvania Press, pp. 50–72.

Kasarda, John D., 1993. Inner-city concentrated poverty and neighborhood distress, 1970 to 1990. Housing Policy Debate 4, 253–302.

Kingsley, G. Thomas, Pettit, Kathryn L.S., 2003. Concentrated Poverty: A Change in Course. Neighborhood Change in Urban America Brief 2. The Urban Institute, Washington, DC.

Leventhal, Tama, Brooks-Gunn, Jeanne, 2000. The neighborhoods they live in: the effects of neighborhood residence on child and adolescent outcomes. Psychological Bulletin 126 (2), 309–337. http://dx.doi.org/10.1037/0033-2909.126.2.309.

Levy, Frank, 1995. The future path and consequences of the U.S. earnings/education gap. Economic Policy Review 1 (1), 35–41.

Lopez, Russell P., Hynes, H. Patricia, 2006. Obesity, physical activity, and the urban environment: public health research needs. Environmental Health 5 (1), 25.

Ludwig, Jens, Sanbonmatsu, Lisa, Gennetian, Lisa, Adam, Emma, Duncan, Greg J., Katz, Lawrence F., Kessler, Ronald C., et al., 2011. Neighborhoods, obesity, and diabetes – a randomized social experiment. New England Journal of Medicine 365 (16), 1509–1519.

Marx, Benjamin, Stoker, Thomas, Suri, Tavneet, 2013. The economics of slums in the developing world. Journal of Economic Perspectives 27 (4), 187–210.

Massey, Douglas S., Denton, Nancy A., 1988. Suburbanization and segregation in U.S. metropolitan areas. American Journal of Sociology 94, 592–626.

Mieskowski, Peter, Mills, Edwin S., 1993. The causes of metropolitan suburbanization. Journal of Economic Literature 7, 135–147.

Moore, Joan, Pinderhughes, Raquel, 1993. In the Barrios: Latinos and the Underclass Debate. Russell Sage Foundation.

Myrdal, Gunnar, 1963. Challenge to Affluence. Pantheon Books, New York, NY.

Nunes, Brasilmar Ferreira, Veloso, Leticia, 2012. Divided Cities: Rethinking the Ghetto in Light of the Brazilian Favela. In: The Ghetto: Contemporary Global Issues and Controversies, vol. 225.

Osypuk, Theresa L., Acevedo-Garcia, Dolores, 2008. Are racial disparities in preterm birth larger in hypersegregated areas? American Journal of Epidemiology 167 (11), 1295–1304.

Ricketts, Erol R., Sawhill, Isabel V., 1988. Defining and measuring the underclass. Journal of Policy Analysis and Management 7, 316–325.

Rivkin, Steven G., 1994. Residential segregation and school integration. Sociology of Education 67 (4), 279–292.

Roberts, Bryan R., 2004. From marginality to social exclusion: from laissez-faire to pervasive engagement. Latin American Research Review 39, 195–197.

Sampson, Robert J., 2012. Great American City: Chicago and the Enduring Neighborhood Effect. The University of Chicago Press, Chicago.

Sampson, Robert J., Sharkey, Patrick, 2008. Neighborhood selection and the social reproduction of concentrated racial inequality. Demography 45 (1), 1–29.

Sampson, Robert J., Sharkey, Patrick, Raudenbush, Stephen W., 2008. Durable effects of concentrated disadvantage on verbal ability among African-American children. Proceedings of the National Academy of Sciences of the United States of America 105 (3), 845–852.

Sharkey, Patrick, 2008. The intergenerational transmission of context. American Journal of Sociology 113 (4), 931–969.

Sharkey, Patrick, 2010. The acute effect of local homicides on Children's cognitive performance. Proceedings of the National Academy of Sciences of the United States of America 107 (26), 11733–11738.

Sharkey, Patrick, Elwert, Felix, 2011. The legacy of disadvantage: multigenerational neighborhood effects on cognitive ability. American Journal of Sociology 116 (6), 1934–1981.

Small, Mario Luis, 2008. Four reasons to abandon the idea of 'The Ghetto'. City & Community 7 (4), 389–398.

Subramanian, S.V., Acevedo-Garcia, Dolores, Osypuk, Theresa L., 2005. Racial residential segregation and geographic heterogeneity in black/white disparity in poor self-rated health in the US: a multilevel statistical analysis. Social Science & Medicine 60 (8), 1667–1679.

Summers, Anita A., Wolfe, Barbara L., 1977. Do schools make a difference? American Economic Review 67 (4), 639–652.

Tienda, Marta, 1991. Poor people and poor places: deciphering neighborhood effects on poverty outcomes. In: Huber, J. (Ed.), Macro-Micro Linkages in Sociology. Sage Publications, Newberry, CA, pp. 244–262.

UN Habitat, 2009. Global Report on Human Settlements 2009: Planning Sustainable Cities. Earthscan, London.

UN Habitat, 2013. State of the World's Cities 2012/2013: Prosperity of Cities. Routledge, New York, NY.

Venkatesh, Sudhir A., 2002. American Project: The Rise and Fall of a Modern Ghetto. Harvard University Press.

Wacquant, Loic, 2004. Ghetto. In: International Encyclopedia of the Social and Behavioral Sciences, pp. 1–7.

White, Michael J., 1987. American Neighborhoods and Residential Differentiation. Russell Sage Foundation, New York.

Wilson, William Julius, 1987. The Truly Disadvantaged: The Inner-City, the Underclass and Public Policy. University of Chicago Press, Chicago.

Wilson, William Julius, 1996. When Work Disappears: The World of the New Urban Poor. Knopf, New York.

Wolch, Jennifer, Wilson, John, Fehrenbach, Jed, 2005. Parks and park funding in Los Angeles: an equity-mapping analysis. Urban Geography 26 (1), 4–35.

Zimmer, Ron W., Toma, Eugenia F., 2000. Peer effects in private and public schools across countries. Journal of Policy Analysis and Management 19 (1), 75–92.

Concept Formation in Childhood: Cross-Cultural Perspectives

Dušan Krnel and Darija Skubic, University in Ljubljana, Ljubljana, Slovenia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Described is the concept formation in early age according the latest accepted theories. Emphasis is on the constructivistic view and the role of classification ability in young children. Reviewed are different approaches toward classification and the properties or biases that attract children's attention. Exposed is the role of language and the relationship between thought and speech in concept formation. Cultural diversity is introduced with two examples: on the construction of the concept of time and on the construction of the concept of space.

Concept Formation

Alongside other knowledge related to 'disparate areas of the mind' are theories about concept formation that are slowly constructing a more coherent picture. As a result, even today one finds entirely contrary theories in the larger discourse, from nativism (Fodor, 1975, 1981) to models constructed upon the theory of cognitive development and constructivism. Nevertheless, there is some general agreement that concept formation is closely connected with cognitive growth; therefore, one of the stronger lines in developing a model of concept formation begins with J. Piaget. According to Piaget (Inhelder and Piaget, 1964), the process of classification or 'what things belong together' seems to play a role in every intellectual function – especially concept formation. After his first report in the 1930s, numerous studies followed regarding how children of different ages, genders, and social groups classified different objects and their argumentation for so doing (Reichard et al., 1944; Welch and Long, 1940; Sigel, 1953; Silverstein, 1960; Elkind, 1961, and many others). The classification process also proved intriguing for the behaviorist. While constructivists described these acts in logical-mathematical terms of extensivity and intensivity of the class of the object, and everything in the mind being derived from concrete actions, the behaviorist described classification and concept formation in terms of classes of stimuli or stimulus complexes (Lunzer, 1979). If one simplifies this role of classification in the process of concept formation – whatever the constructivist or behavioristic approach – the process can be described as follows: differentiation between the properties of objects in a given set, followed by discrimination between properties that are specific to an object and those that are common to the set; and finally, the identification of common properties, which then leads to the development of new concepts (Lovell, 1971). At the end of the process comes the label for this class, and this, according to Piaget, is the word (Labinowicz, 1989).

But for many psychologists, philosophers, and linguists, this view was over simplified and diminished the role of language. It was Lev Vygostky who, from the constructivist point of view, better elaborated the relation between thought and speech in the process of cognitive growth and concept formation.

The Role of Mental Operations

Let us first examine the Piagetian view of concept development as can be interpreted after all the modifications, criticism, and improvements his theory underwent. In new encounters with an object, the child classifies the object into a class that has already been generated. In this way, the new object can be assimilated into the existing structure. Alternatively, an encounter with a new object may not fit precisely with or into existing classes and may require a process of accommodation in which the characteristics of an existing class are modified, thus bringing about conceptual change (Labinowicz, 1989). Children use numerous ways to classify objects and ideas. Inhelder and Piaget (1964) identified 'graphic' and 'spatial' collections. Another route is the 'thematic' collection, whereby children group objects by telling stories about them; or the 'complex structure,' where objects are linked by an idea. Often, one also encounters collections of pairs – groups of two objects that match in some way; and 'chain connections' see new connections found between each new pair in the chain; as well as classification by association and still other means (Bruner et al., 1967). The research work of Nguyen and Murphy (2003) addresses multiple forms of conceptual organization. They have criticized the view that exposed only the taxonomic categorization in concept development and treated other types of categorization as primitive. They argue that adults too use thematic and script categories when certain relations are sufficiently strong. Their results showed that preschool children could selectively use different classification strategies.

Rosch (1980) proposed the 'prototype' model – a quick and effective method of classification made possible by comparing an object with a prototype. Prototypes are seen as an example of a group with typical characteristics. The consequence of this strategy is the differentiation of objects of the same category at different levels – those more central or closer to the stereotype. This differentiation method is central to many models in cognitive science. According to Rosch, so-called natural categories of objects are formed around prototypes. Those that belong to this category share not only certain perceptual similarities but also similar sensorimotor actions.

A characteristic of the natural category 'chair' is a flat surface – but also the 'action' of bending one's knees.

The constructivists in psychology hold that an infant's knowledge is contained in schemata and sensorimotor functions. A schema is defined as the representation of the patterned perceptual feature of an event, and need not possess a semantic component or an inference as to its origin (Kagan, 2008). The primary feature of an object is spatial extent, shape, color, contrast, motion, symmetry, angularity, circularity, and density of contour. The origin of classification and seriation are to be found in sensorimotor schemata; later these schemata are internalized and become one of the mental processes (Inhelder and Piaget, 1964).

The role of action in concept formation was studied by Ogborn and Bliss (1990) and Bliss and Ogborn (1994). Their model was used for the development of the concept of 'movement' (the psycho-logic of motion). A similar model was proposed by Krnel et al. (1998) with respect to the concept of matter.

In order to understand the concept of matter, primitive actions (e.g., to hold, to grasp, and to move) are necessary. Combinations of these actions are used to evolve prototypes. Primitive actions may also be integrated into schemata and are further upgraded into primitive laws (e.g., parts move together with the whole object, or movement can be triggered upon touch). Primitive actions, prototypes, schemata, and laws may be used by children to develop hypotheses about a number of concepts. Although derived from action, the function of an object or function of the category of objects (what they are used for) as criteria for classification appear later. The study of Horst et al. (2005) notes that from an early age, infants form appearance-based (where shape is one of the dominant properties) categories before the functional-based, and that this preference continues during the time they are learning the names of objects, between the ages of 2 and 4 years. Regarding the relationship between function and action, Perone and Oakes (2006) noted the important role of two features of infantile perception: infants can perceive a stimulus event that involves function as causal event; or they perceive, e.g., hands acting on objects as goal-directed agents.

Another model describing mental entities was proposed by di Sessa. He introduced the phenomenological primitives or p-prims. P-prims are small knowledge structures accepted intuitively as self-consistent explanatory facts. In some cases, they become the intuitive equivalent of physical laws, explaining other phenomena but themselves remaining unexplained (di Sessa, 1983). P-prims are based on experience, they do not represent a coherent system – and they may even stand in contradiction to one another. The most common example in literature for p-prims is heaviness. For children, heaviness is an intrinsic property of an object.

But these p-prims also have their own origin. Being based on experience means that some of them, e.g., heaviness, are developed from primitive actions like to hold, to move, or, more generally, from actions where children really feel the force of weight on various limbs or muscles.

In their work to determine how primitive concepts or natural kind concepts might be acquired, Laurence and Margolis (2002) introduced the kind syndrome. "The kind syndrome is a collection of properties that is highly indicative of a kind yet is accessible in perceptual encounters." Laurence and Margolis studied a number of different psycholinguistic experiments where researchers had discovered a number of important biases that enabled concept acquisition. For them, the central (research) question was how do children learn the meaning of a word when they are faced with an object displaying so many properties? How could a child know whether the word 'cup' referred to the color of the cup, the contents of the cup, to a particular situation, or any number of other possibilities? They found that one of the important biases is shape. Beginning with shape and other important biases together with an understanding of their relative importance enables children to represent kind syndrome.

But in many cases this is not enough. An object may appear just like the real or genuine object, but it can also be a fake. How are even young children able to recognize it as a fake? One of the answers to this question emerges from studies about the development of the concept of living (Carey, 1985). In these studies, it was found that children are able to recognize and verbalize some crucial processes and properties (feeding, growth, etc.); and recent studies support previous findings (Jipson and Gelman, 2007). Margett and Witherington (2011) discovered both competence and confusion in preschoolers' living kinds conceptualization, and suggested that this period marks the beginning of an abstract, biologically organized framework.

Another way of answering the question of young children's ability to distinguish between fake and real objects is essentialist tendency (Laurence and Margolis, 2002). When acquiring primitive concepts, children start out with a predisposition toward psychological essentialism. After an experience with a new object, new information is accumulated and linked to the existing mental presentation. In this way, this storing comes to compose a kind syndrome. Finally, the kind syndrome and the essentialist disposition together lead the inference of new representation.

Psychological essentialism is the idea that certain categories, such as 'living' or 'female,' have an underlying reality that cannot be observed directly. There is also evidence that psychological essentialism is an early cognitive bias. The ideas of essentialism see in young children the tendency to go or dig deeper, in search of a nonperceptive feature, which stands in opposition to the standard view of children as concrete thinkers (Taylor et al., 2009).

There are also interesting debates about the role and relation between perceptual and conceptual categorization. According to the theory of constructivism, concepts form a network, which as children mature become increasingly complex. Many constructivists see the process of learning as a process of differentiation between concepts and as a process of developing firm relations between concepts in this structure. Under the influence of the work of Piaget D.P. Ausubel (1968), assimilation theory and the concept of meaningful learning developed. Ausubel stressed the importance of prior knowledge or cognitive structures that should be prepared for bonding new concepts. Thus, if a child could not connect a new concept into an existing structure, the

assimilation failed and the new concept would remain without proper meaning. Through his theory of meaningful learning, Ausubel exposed the role of advance organizers. Advance organizers can be treated as devices used in the introduction of a new concept, which enable learners to orient themselves and to locate where a concept fits in and how it links with existing cognitive structure. Following D.P. Ausubel, J.D. Novak (1990) introduced the technique of concept mapping. Novak did impressive work with children as young as 6 years old making concept maps to represent their response to focus questions such as 'What is water?' and so on.

Using a hierarchically organized model of cognitive structure then raises the question of how the process of organization of concepts in a hierarchical order actually runs. Ausubel (1978) proposed three ways: subordinate learning (new information is linked under a superordinate concept), superordinate learning (existing concepts are recognized as more specific than the new one), and combinatorial learning (the new concept is on the same level as existing related concepts). Children spontaneously assimilate new concepts in all three ways.

In the process of learning, concepts move in their hierarchy not only vertically but in some cases also horizontally, from one bunch of concepts to another. This shift is well explained in the model of the ontological tree developed by Chi (1992). She suggested that there are fundamentally distinct conceptual trees for different ontological categories, for example, category of matter, category of processes, and category of mental states. Young children classify according to their psychological ontology; the concept of heat, for example, is classified in the same branch as material substances. In the process of adequate learning, it is possible to move 'heat' from the material branch to the processes branch, which is an appropriate classification. Chi argues that scientific concepts are difficult to learn if they have an ontological status different from their own initial classification.

The Role of Language

Vygotsky (1986) and Piaget (1962) place the developmental stage of thinking to entirely different developmental levels. Piaget considers this stage a transitional level from unconscious thinking to the logical, from the intimate-individual to the social. Vygotsky, however, offers that it is a transitional form, from the external to the internal, from the social to individual speech, including unconscious speech thinking.

According to Vigotsky's theory, a word is regarded as embodying a concept; therefore, the child uses a word for communication purposes before the child fully understands that word. In function, words are a means of focusing attention, selecting, analyzing, and synthesizing distinctive features in the process of concept formation. In Vigotsky's view, a child does not choose the meaning of his or her words, the meaning of the words is given to him in his conversation with adults; therefore, a child does not spontaneously develop concepts independent of the meaning that word has in a social environment. The meaning of a concept expressed by words is imposed upon the child and is not assimilated in ready-made, but is instead a process of constant development while the child is in communication with adults. Vygotsky claimed that two different preconceptual stages occur in early childhood: the syncretic heap stage and the complex stage. In the first stage, the child groups together objects or ideas that are objectively unrelated. In the next complex stage, the child begins to abstract different attributes about the ideas or objects, and starts to organize ideas into groups. These complexes are determined by the meaning of a given word from the language of adults.

With their ideas and their research in psychology and linguistics, Piaget and Vygotsky encouraged the development of a number of directions, looking for a theoretical explanation of the relationship between thought and speech. Some philosophers claim that it is conceptually necessary that all thought depends on language; other philosophers claim that all thought is, not only conceptually, but also metaphysically and causally, independent of natural language. Between these two poles is a multitude of plausible claims that most, some, or specific types of thought depend on natural language conceptually, metaphysically, or causally. There are many thoughts that one would never have thought in the absence of language.

To clarify these two theories, it is necessary to differentiate between schemata and semantic forms. A schematic prototype can exist without any support from semantic elements. A number of research projects demonstrate that infants under 8 months display reactions to events that are different from those expected, and that understanding these does not require any semantic representation (Kagan, 2008). Only semantic forms posses the semilogical relations that take the form of hierarchically ordered categories for objects and events. Therefore, speaking about concept structures, hierarchical organization, subsuming and superordinate learning, concept mapping, and other processes proposed by Ausubel and Novak imply semantic elements; and these relations emerge only after 5 or 6 years of age.

The relationship between semantic and biases (schemata) is well illustrated in the research of Yoshida and Smith (2003). They proved the hypothesis that names direct attention at two levels of abstraction: known names direct attraction toward the properties most relevant to the category and novel names direct attraction to the shape of objects. In this intercultural study, the object was presented to a child, its name was told, and asked which one of the three alternatives was called by the same name. In the second experiment, known names were replaced with novel, invented ones. The results showed that English- and Japanese-speaking children – in this specific situation – formed categories organized by shape.

In the same context of biases (shape bias in particular) and names, Diesendruck and Bloom (2003) argue that many tiny details influenced children's responses. They concluded that children use not only shape bias, but also appearance (schemata), as adults do. "Something which looks like a dog is likely to be called dog." They concluded that generally, children treat names as referring to kinds.

Two researchers who followed Vigotsky's idea were Diaz and Berk, who studied the self-directed verbalizations of young children during problem-solving activities. They found that children verbalized more when the tasks were

more difficult, and that children who verbalized more often were more successful in their problem solving.

Experts do allow for the possibility that certain types of thought depend on language, at least to the extent to which language is needed, or language is (*de facto*) the cause of their existence. The aim of many researches was to find a causal role for language in the acquisition of kind concepts. Perner et al. (1994) argued that, in the least instance, the use of language enhanced one's capacity for mind reading. Furthermore, Peterson and Siegal (1995) claimed that deaf children, who were late in their acquisition of language also showed a corresponding delayed capacity for mind reading. However, the existing data are consistent with the view of mental-state concepts as an innately structured mind-reading faculty.

Regardless of certain central theories about language and its relation to the thought process, relatively recent research supports the hypothesis that infants learn their first words differently than older children. The study of Pruden et al. (2006) suggested that the process of learning words undergoes a developmental shift. Infants at 10 months can learn new labels by relying on perceptual salient properties instead of social cues provided by a speaker. The latter process or approach is characteristic for older children of 12, 18, and 24 months.

A special case regarding the relationship between thought and language is the acquisition of the concept of number. There is considerable evidence that some numerical concepts exist independent of language (Xu and Spelke, 2000). Children in the preverbal (prior to forming words) period (6 months) can approximately assess a number. Similarly, they can roughly assess the size of the set/series and recognize that a particular set is bigger/smaller than the other. Older children are better at differentiating numbers than the younger, but this also depends on the size of the described lot, as differentiation becomes more difficult as the series increases in size. Moreover, some animals carry out simple calculations using approximate numerical representations. For example, they can calculate the average speed/distance (rate) of their return from the source of food. Other researchers have further argued that animals and children have the ability to accurately understand small numbers and to compare up to three or four numbers (Whynn, 1992; Deahaene, 1997). For example, children who see two objects behind a screen (canvas, curtain) will be surprised if either only a single object or three objects instead of the (anticipated) two are displayed when the screen is raised. Such understanding is clearly numerical, as it does not depend on the identity or properties of these objects.

As regards research with infants, it is worth noting certain methodological problems well explored by Kagan (2008). He demonstrates that many researchers ignore factors that can influence the duration of attention to an event, which is the main diagnosis indicator. These factors are the curvilinear relation between unfamiliarity and the duration of looking, the influence of brain maturation, and the significance of the physical features of the stimuli. Kagan therefore inferred that many facts are inconsistent with the hypothesis that infants' reactions to visual displays varying in number are a sign of competences that are similar to older children's concept of number.

When discussing exact numerical concepts for numbers that are greater than 4 (5, 6, 7, 64, 5 million, etc.), most researchers believe that their acquisition depends on language, especially on mastering the list of 'word count' (5, 6, 7, 8, etc.) and counting procedures. In this case, the whole class of concepts is language-dependent. But there is still an ongoing debate on the manner in which the acquisition process should appear. Some researchers have argued that the basic process includes planning the list/file of counting prior to the preexisting approximate numerical system (Whynn, 1992); others, however, have offered that children themselves find their way to the concept of bigger, exact numbers by underlining the first few items on their counting list with representations in the system of object types, and then making some kind of inductive skip or leap (Carey, 2004; Feigenson et al., 2004).

Thought, Language, and Cultural Diversity

The question – What is the relationship between language and thought? – has intrigued psychologists and philosophers for decades. The discussion essentially began with Worf and Sapir and their radical version of linguistic determinism. Whorf's view has long been abandoned on both theoretical and empirical grounds. Fodor (1975) and Pinker (1994) have argued that the position is conceptually incoherent. Empirical studies have shown that speakers of languages that mark different linguistic distinctions can nevertheless make the same conceptual distinctions. For example, color perception in the Dani (Dugum Dani is a stone-age agrarian population) language is fundamentally similar to ours, even though the Dani language has only two color terms (Rosch, 1975). Rosch and Olivier's (1972) investigations demonstrated that universal patterns of memory confusion over colors were independent of color labeling, and this could be interpreted as supporting to the theory of linguistic relativity. The way one understands language diversity – how languages differ in representing reality– affects one's approach to understanding linguistic relativity, how that diversity affects thought.

Recent research has focused on developing muted versions of the linguistic relativity view. Roberson et al. (2005) expanded on the work of Rosch and Oliver (1972) on the Dani people with a comparable group from Papua, New Guinea, who speak Berinmo, and which has five basic color terms. They found that categorical perception affected both English and Berinmo. It was concluded that possession of linguistic categories facilitated recognition and influenced perceptual judgments. Despite this and other investigations, color has continued to be regarded as a special case of universality, rather than of cultural or linguistic relativity.

Piaget already stated that variation in achieving different cognitive operations could be expected among different cultural groups, and that this could be explained by differences in educational and cultural emphases in the development of certain cognitive skills (Piaget, 1962). But there is scarce evidence related to cultural diversity in terms of obtaining basic cognitive operations. The reason for this lies in the research itself. The great majority of studies were carried out on populations of white, middle-class Western children. The second reason for the limited objectivity of results lies in the using of research instruments that were developed for children from Western cultures.

One of the studies from 2000 (Mwamwenda and Grewal, 2000) on conservation skills among preschool children (aged 5 years) in Umtata in South Africa, found that conservation skills had not been acquired. From a group of 70 children, only 5 were found to be operative on the conservation level. These results are not far from many Western studies. It is also well known – and this was one of the criticisms of Piaget – that results depend on the research instruments employed, and on the context in which they are performed. Kagan (2008) cited a study where an 11-year-old Mayan Indian child living in a small, isolated village in Guatemala, performed as well as Boston children when asked to recognize a large number of photographs. He argued that this is due to schemata they have acquired during their first encounter with pictures. The difference in cognitive ability between Australian Aboriginal and European children was studied by de Lacey (1970). In this study, a sample of very high-contact Aboriginals (children living in closer contact with Europeans and their technology) performed at least as well as white Australian children living in similar environments. Environmental differences (socioeconomic status and similar) were considered to have been a major influence in the performance differences found.

More diversity in concept acquisition that derived from cultural diversity can be observed in the acquisition of some ontological concepts such as time and space.

The Concept of Space and Language/Thought

Worf (1956), like many anthropologists both before and after him, was impressed by a variety of human cultures, whereas as a linguist he enthused over various grammatical forms and modes of conceptualization, demonstrated by different natural languages. Some languages, for instance, do not employ terms to describe 'to the left' and 'to the right,' 'to the south of' or with coordinates that relate to the object, e.g., 'between the river and the sea.' Worf then shows that such differences have significant effects on the cognitive processes for which reason (among others) the English and Dani understand the world very differently.

The results of the research carried out by Li et al. (2002) stand in contrast to some current versions of linguistic relativity. Language communities differ in their stock of reference frames (coordinate systems to reference locations and directions). English typically uses egocentrically defined axes ('left–right'). Other languages like Tseltal lack such a system but use geocentrically defined axes ('north–south'). They asked if the lexical resources available in one's language determined the availability or salience of certain spatial concepts. For example, does a 'left–right' lexical gap translate to a conceptual gap? In three experiments, they compared Tseltal speakers' ability to solve spatial problems requiring an egocentric frame of reference to those requiring a geocentric frame of reference. They found that the Tseltal speakers were above average in solving the egocentric problems, demonstrating that a lexical gap does not necessarily lead to a conceptual gap.

One could therefore expect that language affects the efficiency with which a set of relations is processed, and the easier it is memorized (Levinson, 1996). For example, English abounds in an extensive and frequently used vocabulary for expressing motion (walk, stride, saunter, hop, run, and sprint). English speakers should always be vigilant and encode such features. In Spanish and Greek, for example, movement is expressed by auxiliary phrases ('He went into the room at a run'). One could therefore assume that the speakers of these and similar languages are slower in learning and 'poorer' in memory retention (Slobin, 1996). This thesis was carefully examined in the experimental study by Li et al. (2002), in which they could neither detect nor confirm such effects.

Eric Pederson (1995) says that there are three basic types of distinction between languages in expressing spatial relationships.

The first difference lies in the use of some basic terms. Typically, in the European languages, for example, the terms expressing the front and the rear are used in the same way. The front of the tree is the side that faces the speaker. In contrast, some other languages (e.g., the Hausa language) use a different system, in which the front of the tree is the side that is turned away from the speaker.

The second difference lies in the use of different concepts for describing space. We distinguish between the absolute concept (it is characterized by the terms the North, the South, the East, the West) and the relative concept of space (defined by the terms on/to the left, on/to the right, in front of, at the back of). Most European languages use the relative concept.

The third difference lies in the functional load of the terms listed under the second point. Languages can comprise certain terms, but speakers use them only in different contexts. For spatial orientation (e.g., in the instructions where to go), a typical European speaker would use the terms of the relative system, i.e., to the left, to the right, forward, backward, while the speaker of some other language (e.g., Tamil) would use the terms of the absolute system, i.e., would use to the North, to the South, to the East, to the West. A cross-cultural study in the area of spatial language and cognition in Indonesia (Bali), India, Nepal, and Switzerland was made by Dasen and Mishra (2010). In their study, Dasen and Mishra examined a 'geocentric' system. Their general conclusion is that the development of geocentric spatial language and cognition occurs in a complex ecocultural system, which is adaptive and functional.

Temporal Concept and Language/Thought

The relationship between space and time has long been the subject of philosophical inquiry and psychological experimentation. Two sorts of proposals have emerged, one suggesting a symmetric and the other an asymmetric relationship between space and time in the mind. Observations from the first proposal were synthesized in *A Theory of Magnitude* (Walsh, 2003). According to this theory, space, time, and numbers are all represented in the brain and mind by an analog magnitude system and as such they are symmetrically interrelated. The second, alternative proposal also holds that space, time, and quantity are importantly related, but in a different way. For Lakoff and Johnson, representations (concepts) of time and quantity depend asymmetrically on representations of space. In general, young children use spatial terms earlier than their

temporal counterparts (Clark, 1973); for example, they use here and there to designate points in space before they use now and then for points in time, and produce where questions earlier than when questions.

Piaget (1955) stressed that time and space form an inseparable whole in the child's mind, which suggested a symmetric relationship. The results of Piaget's experiments (on time, motion, and speed) suggested that children might, mistakenly, use spatial information for time more than the other way round. In Piaget's theory, the formation of the concept of time starts from mere surface contact with immediate realities and gradually develops into objective time. In this view, time and space form first a unity and later, children's efforts are directed at differentiation between them (Lovell and Slater, 1960). Piaget argues that the notion of time is intellectual in construction, because duration cannot be perceived.

Cassanto et al. (2010) investigated the relationship between space and time in the developing mind. Native Greek-speaking children watched movies with two animals traveling along parallel paths for different distances or durations. The results showed that for the same stimuli, spatial information influenced temporal judgment more than temporal information influenced spatial judgment. This finding demonstrates a reliable cross-dimensional asymmetry. Kindergarten and elementary school-aged children can ignore irrelevant temporal information when making judgments about space, but they have difficulty ignoring spatial information when making judgments about time.

Like other concepts (e.g., quantity of matter), the concept of time is related to the cognitive operation of conservation. Murray's 2006 study found that the shift from nonconservation to conservation of time occurred between the ages of 8 and 10 years. The development of the concept of time is based on the perception of changes – and one of the omnipresent changes in human history has been weather and the cyclical changes in the moon and the seasons. Therefore, in some languages, the word for time and for weather remains the same (e.g., in Italian – tempo/tempo).

Few researchers, however, stress the existence of components or prerequisites for the acquiring of the concept of time (Kayra-Stuart, 1977). These components are order, simultaneity, and duration, as well as their coordination. The results support the following developmental sequence for the concept of time. First, the main components are understood separately (possibly in the above mentioned order), then they are integrated. Moreover, temporal relations are grasped first on a nonverbal level and later on a linguistic level. The study of Richie and Bickhard (1988) revealed that children have a perceptual experience of duration separate from one derived inferentially from nontemporal knowledge (e.g., start/stop relationships, speed, and distance) and that they can use this perceptual knowledge to judge duration problems. This suggests that the logical concept of time involves the capacity to experience duration perceptually. A similar hypothesis about the possession of a primitive capacity to compare duration was proposed by Berndt and Wood (1974). Levin (1977), in an attempt to construct a general model for the development of time concepts, used three 'types' of time: still time, rotational time, and linear time. Her hypothesis stated that a child's concept of time is basically temporal, but that the construction of the time concept is influenced by various factors, including spatial variables. The understanding of linear time is connected with temporally unidirectional events or 'arrows of time' (Friedman, 2003). In an experiment, children 3.5–6.5 years of age compared forward and reverse playings of events on video. He found that even children 3.5–4.5 years of age recognized the anomaly of the reverse versions of gravity and separation events.

A cross-cultural study on the formation of duration time concept in Chinese and Canadian children revealed that both the tendencies and the age at which concept formation develops are almost the same (Manshu et al., 1987).

Different from these particular studies of time concept acquisition based on concrete events and specific context is the general perception of time, which is culturally dependent.

The Aboriginal concept of time differs from the Judeo-Christian perception of time in that Aboriginal people do not perceive time as an exclusively linear category (i.e., past–present–future) and often place events in a 'circular' time pattern, according to which an individual stands at the center of 'time circles' and events are placed in time according to their relative importance for the individual and his or her respective community, i.e., more important events are perceived as being 'closer in time' (Janca and Bullen, 2003). The concept of time for various African tribes too is different from that in Western cultures. Akiiki (Msumange, 2012) describes time in African thought as phenomenological (existential), organic, cyclical, and sacred. Essentially, it is empirical and delineative, taking events as reference points. It is easily understood through the interaction of phenomena in the universe. Time is sacred and profane, and aims or serves to give meaning to events. It is cyclical, that is, spherically organized, and adds dimension to the line as it proceeds forward, giving past and future shape/dimension, which two are separated by the present. The researchers (Msumange, 2012) of this concept argue that this quite different perception of time wields significant influence over everyday life, even on the economic problems confronting Africa.

See also: Bruner's Theory of Cognitive Development; Cognitive Development During Infancy and Early Childhood across Cultures; Cognitive Development, Educational Theories of; Cross-Cultural Training in Industrial and Organizational Psychology; Mathematical Concepts During Early Childhood Across Cultures, Development of; Neo-Piagetian Theories of Cognitive Development; Piaget's Theory of Cognitive Development; Problem Solving During Infancy and Early Childhood, Development of; Scientific Concepts During Childhood, Development of; Spatial Cognition During Infancy and Early Childhood Across Cultures, Development of; Vygotsky's Theory of Cognitive Development.

Bibliography

Ausubel, D.P., 1968. Educational Psychology a Cognitive View. Holt Rinehart and Winston, New York.

Ausubel, D., 1978. In defense of advance organizers: a reply to the critics. Review of Educational Research 48, 251–257.

Berndt, T.J., Wood, D.J., 1974. The development of time concepts through conflicts based on primitice duration capacity. Child Development 45, 825–828.
Bliss, J., Ogborn, J., 1994. Force and motion from the beginning. Learning and Instruction 4, 7–25.
Bruner, J.S., Oliver, R.R., Greenfield, P.M., 1967. Studies in Cognitive Growth. John Wiley and Sons, New York.
Carey, S., 1985. Conceptual Changes in Childhood. MIT Press, Cambridge, MA.
Carey, S., 2004, Winter. Bootstrapping and the Origin of Concepts. Daedalus 133, 59–68.
Cassanto, D., Fotakopoulu, O., Boroditsky, L., 2010. Space and time in the child's mind: evidence for a cross-dimensional asymmetry. Cognitive Science 34 (3), 387–405.
Chi, M.T.H., 1992. Conceptual change within and cross ontological categories: examples from learning and discovery in science. In: Giere, R.N. (Ed.), Cognitive Models in Science. University of Minnesota Press, Minneapolis, pp. 129–186.
Clark, H.H., 1973. Space, time, semantics and the child. In: Moore, T.E. (Ed.), Cognitive Development and the Acquisition of Language. Academic Press, New York.
Dasen, P., Mishra, R.C., 2010. Spatial language and cognition: an ecocultural perspective. In: Cambridge Studies in Cognitive and Perceptual Development, vol. 12. Cambridge University Press, Cambridge, England.
Deahaene, S., 1997. The Number Sense. Oxford University Press, Penguin Press, New York, Cambridge, UK.
Diesendruck, G., Bloom, P., 2003. How specific is the shape bias? Child Development 74 (1), 168–178.
Elkind, D., 1961. The development of quantitative thinking: a systematic replication of Piaget's studies and many others. The Journal of Genetic Psychology and Theory of Human Development 98, 37–46.
Feigenson, L., Dehaene, S., Spelke, E., 2004. Core systems of number. Trends in Cognitive Sciences 8, 307–314.
Fodor, J.A., 1975. The Language of Thought. Harvard University Press, Cambridge, MA.
Fodor, J.A., 1981. The present status of innateness controversy. In: Representations: Philosophical Essays on the Foundations of Cognitive Science. MIT Press, Cambridge, MA.
Friedman, W.J., 2003. Arrows of time in early childhood. Child Development 74 (1), 155–167.
Gelman, S.A., 2004. Psychological essentialism in children. Trends in Cognitive Sciences 8 (9), 404–409.
Horst, J.S., Oakes, L.M., Madole, K.L., 2005. What does it look like and what can it do? Category structure influences how infants categorize. Child Development 76 (3), 614–631.
Inhelder, B., Piaget, J., 1964. The Early Growth of Logic in the Child, Classification and Seriation. Routledge & Kegan Paul, London.
Janca, A., Bullen, C., 2003. The aboriginal concept of time and its mental health implications. Australasian Psychiatry 11, S40–S44.
Jipson, J.L., Gelman, S.A., 2007. Robots and rodents: children's inferences about living and non-living kinds. Child Development 78 (6), 1675–1688.
Kagan, J., 2008. In defense of qualitative changes in development. Child Development 79 (6), 1606–1624.
Kayra-Stuart, F., 1977. The Development of the Concept of Time. ERIC database (136915).
Krnel, D., Glazar, S.A., Watson, R., 1998. Survey of research related to the development to the concept of "matter". International Journal of Science Education 20 (3), 257–289.
de Lacey, P.R., 1970. A cross-cultural study of classificatory ability in Australia. Journal of Cross-Cultural Psychology 1 (4), 293–304.
Labinowicz, E., 1989. The Piaget Primer. DZS, Ljubljana.
Laurence, S., Margolis, E., 2002. Radical concept nativism. Cognition 86, 25–55.
Leslie, A., Gallistel, C., Gelman, R., 2007. Where do the integers come from? In: Carruthers, P., Laurence, S., Stich, S. (Eds.), The Innate Mind, Foundations and the Future, vol. 3. Oxford University Press, New York.
Levin, I., 1977. The development of time concepts in young children: reasoning about duration. Child Development 48 (2), 435–444.
Levinson, S.C., 1996. In: Gumperz, J. (Ed.), Rethinking Linguistic Relativity. Cambridge University Press, Cambridge.
Li, P., Abarbanell, L., Papafragou, A., 2002. Spatial Reasoning Skills in Tenejapan Mayans. http://papafragou.psych.udel.edu/papers/lietal.pdf.
Lovell, K., 1971. The Growth of Basic Mathematical and Scientific Concepts in Children, fifth ed. Unibooks, University of London Press, London.
Lovell, K., Slater, A., 1960. The growth of the concept of time: a comparative study. Child Psychology and Psychiatry 1, 179–190.
Lunzer, E.A., 1979. Concept formation and abstraction (1). Educational Review 31 (2), 111–121.
Manshu, Z., Zhenguo, S., Harvey, C.B., 1987. Formation of duration time concept in Chinese and Canadian children. Psychological Science 2.
Margett, T.E., Witherington, D.C., 2011. The nature of preschoolers' concept of living and artificial objects. Child Development 82 (6), 2067–2082.
Msumange, G., 2012. Time - the Concept of Future in the Thought of the "Hehe" and Other African Tribes. http://iff.ovh.org/baca/Msumange.htm.
Murray, F.B., 2006. Conservation aspects of the concept of time in primary school children. Journal of Research in Science Teaching 6 (3), 257–264.
Mwamwenda, T.S., Grewal, A.S., 2000. Conservation skills among preschool children in Umtata, South Africa. Psychological Reports 87, 21–22.
Nguyen, S.P., Murphy, G.L., 2003. An apple is more than just a fruit: cross-classification in children's concepts. Child Development 74, 1783–1806.
Novak, J.D., 1990. Concept mapping: a useful tool for science education. Journal of Research in Science Teaching 27 (19), 937–949.
Ogborn, J., Bliss, J., 1990. A psycho-logic of motion. European Journal of Psychology of Education 5 (9), 379–390.
Pederson, E., 1995. Language as context, language as means: spatial cognition and habitual language use. Cognitive Linguistics 6, 33–62.
Perner, J., Ruffman, T., Leekham, S., 1994. Theory of mind is contagious: you catch it from your sibs. Child Development 65, 1228–1238.
Perone, S., Oakes, L.M., 2006. It clicks when it is rolled and it squeaks when it is squeezed: what 10-month infants learn about object function. Child Development 77 (6), 1608–1622.
Peterson, C., Siegal, M., 1995. Deafness, conversation, and theory of mind. Journal of Child Psychology and Psychiatry 36, 459–474.
Piaget, J., 1955. The Child's Construction of Reality. Routledge and Kegan Paul, London.
Piaget, J., 1962. The Language and Thought of the Child. Routledge & Kegan Paul, London.
Pinker, S., 1994. The Language Instinct. William Morrow, New York.
Pruden, S.M., Pasek, K.H., MichnickGolinkoff, R., Hennon, E.A., 2006. The birth of words: ten-month-olds learn words through perceptual salience. Child Development 77 (2), 266–280.
Reichard, S., Schneider, M., Rapaport, D., 1944. The development of concept formation in children. American Journal of Orthopsychiatry 14, 156–161.
Richie, D.M., Bickhard, M.H., 1988. The ability to perceive duration: its relation to the development of the logical concept of time. Developmental Psychology 24, 318–323.
Roberson, D., Davies, I.R., Davidoff, J., Shapiro, L.R., 2005. Color categories: confirmation of the relativity hypothesis. Cognitive Psychology 50, 381–411.
Rosch, E., 1975. Cognitive representations of semantic categories. Journal of Experimental Psychology: General 104, 192–233.
Rosch, E., 1980. Classification of real world objects: origins and representation in cognition. In: Johnson-Laird, P.N., Wason, P.C. (Eds.), Thinking, Readings in Cognitive Science. Cambridge University Press, Cambridge, MA.
Rosch Heider, E., Olivier, D.C., 1972. The structure of the color space in naming and memory for two languages. Cognitive Psychology 3, 337–354.
Sigel, I.E., 1953. Developmental trends in abstraction ability of children. Child Development 24 (2), 131–144.
Silverstein, A.B., 1960. A cluster analysis of object sorting behavior. Journal of Consulting Psychology 24 (1), 98.
di Sessa, A., 1983. Phenomenology and the evolution of intuition. In: Genter, D., Stevens, A. (Eds.), Mental Models. Lawrence Erlbaum, London.
Slobin, D.I., 1996. From "thought and language" to "thinking for speaking". In: Gumperz, J.J., Levinson, S.C. (Eds.), Rethinking Linguistic Relativity. Cambridge University Press, Cambridge, pp. 70–96.
Taylor, M.G., Rhodes, M., Gelman, S.A., 2009. Boys will be boys; cows will be cows: children's essentialist reasoning about gender categories and animal species. Child Development 80 (2), 461–481.
Vygotsky, L.S., 1986. Thought and Language. MIT, Massachusetts.
Walsh, W., 2003. A theory of magnitude: common cortical metrics of time, space and quantity. Trends in Cognitive Science 7 (11), 483–488.
Welch, L., Long, L., 1940. A further investigation of the higher structural phases of the concept formation. The Journal of Psychology 10, 211–220.
Whynn, K., 1992. Children's acquisition of the number words and the counting system. Cognitive Psychology 94, 220–251.
Worf, B.L., 1956. Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf. MIT Press, London, Cambridge.
Xu, F., Spelke, E., 2000. Large number discrimination in 6-month-old infants. Cognition 74, B1–B11.
Yoshida, H., Smith, L.B., 2003. Known and novel noun extensions: attention at two level of abstraction. Child Development 74 (2), 564–577.

Concept Learning and Representation: Models

David R Shanks, University College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Concept learning describes the process by which experience allows us to partition objects in the world into classes for the purpose of generalization, discrimination, and inference. Formal models of concept learning and categorization make divergent assumptions about knowledge representation (for instance, prototypes, exemplars, and rules). As reviewed here, empirical comparisons of such models have enhanced our understanding of key issues such as the role of selective attention in concept learning, the time course of categorization decisions, and the interpretation of neuroscientific findings.

Concept learning describes the process by which experience allows us to partition objects in the world into classes for the purpose of generalization, discrimination, and inference. When children learn that certain objects are apples and others oranges, their newly acquired knowledge allows them to make informed guesses about new objects they encounter (this is an apple but that is not…), to tell members of these categories apart, and to make decisions that support various actions (this is an apple so I can eat it).

Models of concept learning have adopted one of several contrasting views concerning category representation (see Knowledge Representation). In prototype theories, the concept learning process is assumed to yield an abstract representation corresponding to the central tendency of the category exemplars on each of the dimensions of variation. In exemplar models, the concept is simply the set of mental representations of all the category exemplars that have been previously observed, with each instance assumed to be stored as a separate trace. In decision rule models, the learner is assumed to construct a boundary or rule in psychological space, which partitions it into different category regions. These different perspectives, and some of the evidence supporting each of them, are considered in turn. Next, the role of selective attention in categorization, and the way in which the different models deal with selective attention, is discussed. Evidence that categorization may be controlled by multiple mechanisms is evaluated, and finally the fine-scale dynamics (i.e., time course) and some aspects of the neuropsychology of categorization are reviewed.

Prototype Theories

Prototype theories make a simple assumption about concept representation, namely, that a concept is represented as its 'ideal' or 'average' category exemplar. Thus, representations of specific category exemplars do not influence classification. Numerous forms have been suggested for the (probabilistic) response rule; here, we describe one common example (for other examples, see *Categorization and Similarity Models*). The probability that stimulus i is classified into category J, $P(R_J/S_i)$, is given by

$$P(R_J/S_i) = \frac{b_J s_{iP(J)}}{\sum_K b_K s_{iP(K)}} \quad [1]$$

where $b_J (0 < b_J < 1, \sum b_J = 1)$ represents a bias toward making category response J and $s_{iP(J)}$ is the similarity between exemplar i and the prototype of category J. The model states that classification depends on the similarity between exemplar i and the category J prototype relative to i's similarity to the prototypes of all other categories. Similarity is usually measured with multidimensional scaling methods (*see* Signal Detection Theory).

One piece of evidence supporting prototype models is that a previously unseen prototype may in some circumstances be classified with higher accuracy in the test phase than the actual training stimuli. Moreover, compared to the training stimuli, the prototype may be particularly resistant to forgetting (Homa et al., 1981). Notwithstanding their attractive simplicity, several lines of evidence have been marshaled against prototype models of concept learning, at least within the domain of perceptual classification. Prototype models have often been shown to provide poorer fits than other models to large sets of classification data; they make a number of predictions that have been falsified, and they fail to account for a number of well-established phenomena (see Nosofsky, 1992 for a thorough review).

Exemplar Theories

Although it may seem implausible that classification is mediated by a 'chorus' of category exemplars, there is very substantial body of evidence consistent with the idea that a trace is created in memory for each category member that is observed and that the effects of individual exemplars on later classification are long lasting. For example, judging whether a letter string is a word or a nonword is influenced by a single prior encounter with that string, and this influence can persist for many days and be specific to the exact surface form (e.g., font) of the string (Tenpenny, 1995). The best developed formal exemplar-based model of concept learning is Nosofsky's Generalized Context Model (GCM). In the basic form of the GCM, $P(R_J/S_i)$ is given by the equation

$$P(R_J/S_i) = \frac{b_J \sum_{j \in J} s_{ij}}{\sum_K (b_K \sum_{k \in K} s_{ik})} \quad [2]$$

where s_{ij} is the similarity between exemplars i and j and, as before, b_J represents the bias toward making category response J. The idea is that classification depends on the summed similarity between i and each of the members of category J, relative to i's similarity to the exemplars of all other categories. As in prototype models, similarity is usually measured using multidimensional scaling methods.

A sizable amount of evidence supports exemplar models such as the GCM. These models have been shown not only to provide excellent fits to parametric data sets but also to predict a number of qualitative phenomena. For example, they predict that a previously unseen prototype may in some circumstances be classified with higher accuracy in the test phase than the actual training stimuli, and that classification of test items may sometimes be almost completely unrelated to recognition of the same items.

There is also more direct evidence of preservation of the training exemplars in memory. Thomas (1998) first trained participants in a two-dimensional classification task in which values on the dimensions were highly correlated across exemplars. Next, participants were shown stimuli that possessed only one of the dimensions, together with the appropriate category label, and asked to predict the value of the stimulus on the other dimension. Consistent with exemplar theories, Thomas found that many participants were able to perform this task accurately, predicting values for the missing dimension consistent with the correlation observed across the exemplars during training. This implies that they retained quite detailed information about the conjunctions or correlations of feature values across exemplars. Many theories in which information about specific training exemplars is discarded (e.g., prototype theory) would be unable to account for this finding.

The major difficulties faced by exemplar models are that they do not fully reproduce individual differences in the distributions of responses across test stimuli (Nosofsky et al., 1994) and they do not account for certain base-rate effects in categorization (Nosofsky et al., 1992). We will also see below that some further data obtained by Thomas (1998) are very problematic for exemplar theories.

Recently, models have begun to be developed that view concept learning as a problem of probabilistic inference and that use Bayesian computations to determine the probability that a particular item belongs to a given category. Depending on specific assumptions, such models have been shown to have close connections to both prototype and exemplar models.

Decision-Bound Models

In decision-bound or rule-based models, concepts are represented as regions in psychological space. A rule or boundary demarcates the dividing line between one category and another. In their most general form, decision-bound models can assume any mathematical form for these boundaries, but in practice most researchers have assumed that people can only learn 'simple' boundaries, and have therefore attempted to study the plausibility of special cases in which they are constrained to be linear or quadratic. Another reason for studying these particular forms is that the optimal boundary for discriminating between two bivariate normal distributions is either linear or quadratic, and it is reasonable to speculate that many natural categories are multivariate normal. The theory then assumes that, in a two-category situation, if a stimulus (or, more accurately, the percept of the stimulus) falls on one side of the boundary, the participant assigns it to the corresponding category and if it falls on the other side, it is assigned to the other category. Importantly, the distance of the stimulus from the boundary is assumed to be irrelevant as far as $P(R_J/S_i)$ is concerned. In that sense, categorization is assumed to be deterministic.

Decision-bound models have proved difficult to discriminate empirically from exemplar-based models such as the GCM – indeed, under some circumstances they are formally equivalent (Ashby and Maddox, 1993). Decision-bound models are favored by data from nonnormally distributed categories (Maddox and Ashby, 1993) and by evidence that participants in some circumstances do not retain information about specific training exemplars. One compelling example was obtained by Thomas (1998) in the study described previously. Although many participants in Thomas' study were able to predict one dimension value on the basis of another in stimuli drawn from categories with correlated dimensional values, others were quite unable to do this and instead responded by essentially giving the mean value of the missing dimension, despite the fact that they had learned the initial classification. This strongly suggests that during the course of category learning they had extracted a decision bound and discarded information about the specific training exemplars.

On the other hand, there is rather little evidence, even in one-dimensional perceptual categorization tasks, that participants use all-or-none cutoffs as the decision-bound theory predicts (Kalish and Kruschke, 1997).

Selective Attention in Concept Learning

It has been recognized for many years that concept learning increases the amount of attention paid to stimulus dimensions that are relevant for solving a classification problem and decreases the amount paid to irrelevant ones. Goldstone (1994), for example, found that perceptual discrimination of highly confusable stimuli in a same–different judgment task increased as a result of category learning if the stimuli were assigned to different categories, whereas, at least in some cases, it decreased if the stimuli were assigned to the same category.

In the GCM, this process of selective attention is accommodated by assuming that the similarity between a test item and category exemplars stored in memory is a function of their separation in psychological space, and that this space can be stretched or shrunk (thus altering the similarities) along orthogonal dimensions as a consequence of selective attention. A similar process can be incorporated into prototype theories. Nosofsky and his colleagues (e.g., McKinley and Nosofsky, 1996) have now provided abundant evidence on the one hand that this process of selectively weighting the dimensions of stimulus variation in the generation of category decisions is essential in fitting categorization data within the exemplar-based framework.

On the other hand, decision-bound models have been paradoxically successful at fitting categorization data without

needing to assume an explicit selective attention process. In this theory, the structure of the classification problem affects the placement of the decision boundary, but no additional notion of selective attention to stimulus dimensions is necessary. It is perhaps unclear how decision-bound theories would explain results such as those of Goldstone, which provide direct evidence for changes in the perceived similarities of stimuli as a consequence of category learning, but selective attention does not seem to be required for predicting categorization responses per se.

One possible resolution is based on the distinction between selective attention as a perceptual or as a decisional process. The former sees attention as directly altering the perceptual representation of a stimulus, while the latter assumes that attention operates later in the processing stream, perhaps weighting certain features of the stimulus more heavily than others in making a classification decision. Goldstone's data provide strong support for selective attention affecting the perceptual representation of stimuli, whereas direct evidence for an influence on decisional processes – as assumed in the GCM – is largely lacking. Therefore, there is scope for decision-bound models to dispense with selective attention as a decisional process so long as some provision is made for it to operate perceptually, for example, via changes in the variances of stimulus percepts on the different dimensions. (For further discussion, see Maddox and Ashby, 1998.).

One implication of the idea that selective attention operates by stretching and shrinking orthogonal dimensions in psychological space is that all points on a dimension are affected equally – the stretching/shrinking is uniform across the dimension. Thus, equivalent enhancements or attenuations in discrimination should occur at all points along the dimension. A challenge to this, however, comes from evidence that attentional changes can apply nonuniformly; in other words, discrimination at just one part of a stimulus dimension may be affected by experience. Johansen et al. (2010) presented evidence that the GCM with this sort of feature-specific attention can account for certain patterns of categorization behavior better than the standard version with dimensional attention.

Hybrid Models

An obvious possibility suggested by the evaluation of prototype, exemplar, and decision-bound models is that people can use different categorization strategies in different circumstances, and indeed there is now a wealth of evidence in support of this conjecture. For example, Johnstone and Shanks (2001) asked participants to classify letter strings as either 'grammatical' or 'ungrammatical' after exposing them to strings, all of which were consistent with a complex rule. With incidental training conditions, classification was controlled purely by featural similarity of studied exemplars, but with explicit 'code-breaking' training conditions; classification of exactly the same items was controlled by the complex rule and was unrelated to similarity of studied exemplars. Erickson and Kruschke (1998) review the growing body of evidence for this 'multiple-mechanism' view of category learning and representation. Another hybrid model combines prototype and exemplar representations (Blair and Homa, 2003).

A question that arises is how one might computationally model a multiple-mechanism categorization system, and several attempts have recently been made. Erickson and Kruschke (1998), for example, had some success with a 'mixture-of-experts' model, which incorporates both an exemplar and a decision-bound module. In this model, an adaptive learning rule (see Connectionist Models of Concept Learning) determines the extent to which the final category decision is based on one module or the other, and the weight assigned to each module is an inverse function of the prediction error it generates for that particular classification problem.

Time Course of Classification

An important recent development in the study of category learning and representation is the application of formal models to the fine-scale dynamics – specifically, the time course – of categorization. Until quite recently, research concentrated almost exclusively on categorization probabilities, but new methodological techniques provide researchers with powerful tools for studying more microscopic aspects of the categorization process.

Two rather different proposals have been advanced to extend exemplar models to deal with this issue. One approach (Nosofsky and Palmeri, 1997) is to assume that when an item is presented for classification, a race begins to retrieve exemplars from memory. The time to complete the race depends on how similar the stored exemplars are to the test item. When one race is completed, and an exemplar retrieved, a new race commences, and a classification response is only emitted when a random-walk process is driven to one or another decision barrier. The alternative approach (Lamberts, 2000) is to assume that there is a gradual accumulation of perceptual information about the test stimulus and that similarity to stored exemplars varies during the construction of its representation. On this account, classification at any given moment is governed by eqn [2], but the summed similarity term in that equation is dependent on how much information about the test stimulus has been accumulated.

Proponents of the decision-bound theory have advanced a rather straightforward hypothesis about the time course of categorization, namely, that classification response times depend on distance to the decision bound, with greater distances leading to faster responses.

Attempts to distinguish between these accounts are in their infancy, but each appears to have some merits. Nosofsky and Stanton (2005) provide one attempt to contrast them empirically. Further study of the temporal dynamics of categorization is sure to be a major part of future research.

Neuropsychology of Category Learning and Representation

Neuropsychology raises two important issues concerning models of concept learning. First, there is evidence (albeit controversial) that the mental representations controlling categorization are functionally and neurally distinct from those

accessed in standard 'explicit' tests of memory such as recall and recognition. If this is true, it suggests that models of memory are unlikely to provide any useful input to our understanding of concepts and concept learning. The main piece of evidence is that amnesic patients suffering from anterograde amnesia seem to behave normally in category learning tasks while being profoundly impaired on tests of recall and recognition (e.g., Squire and Knowlton, 1995). For instance, one densely amnesic individual (EP) performed normally in learning to classify dot patterns generated from a prototype, while performing at chance in a recognition test for those patterns. It has been concluded from this that the hippocampus, which controls explicit memory and which is damaged in amnesia, plays no role in category learning.

However, the conclusion that categorization and memory are functionally and neurally disconnected has been questioned on the basis of demonstrations that exemplar models such as the GCM can, just like EP, show dissociations between classification and recognition. In Nosofsky and Zaki's (1998) modeling approach, only a single parameter, the sensitivity parameter, was varied in order to account for a general learning difference between amnesiacs and controls. This parameter represents the ability to discriminate among distinct exemplars stored in memory, an ability that is plausibly reduced in amnesiacs as a result of their deteriorated memory capacity. Nosofsky and Zaki showed that decreasing the value of this parameter reduced classification performance only slightly, but reduced recognition performance considerably. Thus, the GCM successfully accounts for the dissociation between classification and recognition observed by Squire and Knowlton (1995).

Second, there is now a wealth of evidence concerning the brain localization of conceptual knowledge. Brain imaging studies have suggested that knowledge is organized structurally at the neural level. In a pioneering study, Martin et al. (1996) explored the neural organization of conceptual knowledge by presenting participants with pictures of animals and tools to name while brain activity was recorded by positron emission tomography. As a control, participants saw nonsense objects. Compared to the control condition, brain activity for the meaningful objects was increased in the ventral temporal lobes bilaterally and in Broca's area in the left hemisphere. The latter is probably related to word generation. Relative to pictures of tools, pictures of animals produced increased activation in the left medial occipital lobe. Relative to pictures of animals, pictures of tools produced increased activation in the left dorsal temporal lobe and the left premotor region. This premotor area is also activated by imagined hand movements, suggesting that part of the conceptual representation of tools is a specification of the hand movements they afford. Overall, Martin et al.'s results – and many subsequent neuroimaging experiments – confirm that gross category distinctions are present in the way the brain organizes conceptual knowledge.

For a fuller review of recent psychological and neuroscience research on models of concept learning and representation, see Pothos and Wills (2011).

See also: Connectionist Models of Concept Learning; Knowledge Representation; Scientific Concepts: Development in Children.

Bibliography

Ashby, F.G., Maddox, W.T., 1993. Relations between prototype, exemplar, and decision bound models of categorization. Journal of Mathematical Psychology 37, 372–400.

Blair, M., Homa, D., 2003. As easy to memorize as they are to classify: The 5–4 categories and the category advantage. Memory & Cognition 31, 1293–1301.

Erickson, M.A., Kruschke, J.K., 1998. Rules and exemplars in category learning. Journal of Experimental Psychology: General 127, 107–140.

Goldstone, R., 1994. Influences of categorization on perceptual discrimination. Journal of Experimental Psychology: General 123, 178–200.

Homa, D., Sterling, S., Trepel, L., 1981. Limitations of exemplar-based generalization and the abstraction of categorical information. Journal of Experimental Psychology: Human Learning and Memory 7, 418–439.

Johansen, M.K., Fouquet, N., Shanks, D.R., 2010. Featural selective attention, exemplar representation, and the inverse base-rate effect. Psychonomic Bulletin and Review 17, 637–643.

Johnstone, T., Shanks, D.R., 2001. Abstractionist and processing accounts of implicit learning. Cognitive Psychology 42, 61–112.

Kalish, M.L., Kruschke, J.K., 1997. Decision boundaries in one-dimensional categorization. Journal of Experimental Psychology: Learning, Memory, and Cognition 23, 1362–1377.

Lamberts, K., 2000. Information-accumulation theory of speeded categorization. Psychological Review 107, 227–260.

Maddox, W.T., Ashby, F.G., 1993. Comparing decision bound and exemplar models of categorization. Perception and Psychophysics 53, 49–70.

Maddox, W.T., Ashby, F.G., 1998. Selective attention and the formation of linear decision boundaries: comment on McKinley and Nosofsky (1996). Journal of Experimental Psychology: Human Perception and Performance 24, 301–321.

Martin, A., Wiggs, C.L., Ungerleider, L.G., Haxby, J.V., 1996. Neural correlates of category-specific knowledge. Nature 379, 649–652.

McKinley, S.C., Nosofsky, R.M., 1996. Selective attention and the formation of linear decision boundaries. Journal of Experimental Psychology: Human Perception and Performance 22, 294–317.

Nosofsky, R.M., 1992. Exemplar-based approach to relating categorization, identification, and recognition. In: Ashby, F.G. (Ed.), Multidimensional Models of Perception and Cognition. Erlbaum, Hillsdale, NJ, pp. 363–393.

Nosofsky, R.M., Kruschke, J.K., McKinley, S.C., 1992. Combining exemplar-based category representations and connectionist learning rules. Journal of Experimental Psychology: Learning, Memory, and Cognition 18, 211–233.

Nosofsky, R.M., Palmeri, T.J., 1997. An exemplar-based random walk model of speeded classification. Psychological Review 104, 266–300.

Nosofsky, R.M., Palmeri, T.J., McKinley, S.C., 1994. Rule-plus-exception model of classification learning. Psychological Review 101, 53–79.

Nosofsky, R.M., Stanton, R.D., 2005. Speeded classification in a probabilistic category structure: Contrasting exemplar-retrieval, decision-boundary, and prototype models. Journal of Experimental Psychology: Human Perception and Performance 31, 608–629.

Nosofsky, R.M., Zaki, S.R., 1998. Dissociations between categorization and recognition in amnesic and normal individuals: an exemplar-based interpretation. Psychological Science 9, 247–255.

Pothos, E.M., Wills, A.J., 2011. Formal Approaches in Categorization. Cambridge University Press, Cambridge, UK.

Squire, L.R., Knowlton, B.J., 1995. Learning about categories in the absence of memory. Proceedings of the National Academy of Sciences 92, 12470–12474.

Tenpenny, P.L., 1995. Abstractionist versus episodic theories of repetition priming and word identification. Psychonomic Bulletin and Review 2, 339–363.

Thomas, R.D., 1998. Learning correlations in categorization tasks using large, ill-defined categories. Journal of Experimental Psychology: Learning, Memory, and Cognition 24, 119–143.

Concepts and Methods of Cognitive Therapies

Keith S Dobson and Liza Mastikhina, University of Calgary, Calgary, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The cognitive therapies are one of a broad school of psychotherapies, which have evolved since they first emerged in the 1970s. This article discusses the three 'waves' of cognitive therapies, and provides definitions for many of the major forms of therapy. Issues of broad interest in the field, including mechanisms of action, and a recent move toward transdiagnostic approaches to therapy are also discussed. The article concludes with directions for future research and theory development.

Cognitive-behavioral therapies (CBTs) are a broad class of treatments that focus on cognitive and behavioral change. At their core, CBTs share three fundamental propositions (Dobson and Dobson, 2009). These propositions state that (1) it is possible to consciously monitor one's cognitive activity (i.e., 'the access hypothesis'), (2) cognitions impact emotions and subsequent behaviors (i.e., 'the meditational hypothesis'), and (3) one can intentionally modify the manner in which he/she responds to events in the environment (i.e., 'the change hypothesis'). Based on a large number of outcome trials, CBT has been suggested as the treatment of choice for a variety of psychological disorders (for a review, see Epp and Dobson, 2010). Further, the evidence base suggests that CBTs are in some instances as effective as pharmaceutical treatments, and that they are beneficial for certain medical problems such as chronic pain (cf Sensky, 2004). Given the breadth of the literature on the CBTs, this article will focus mainly on the basic concepts and will present the latest developments in the field.

'Classic' Cognitive Therapy

Cognitive therapy (CT) was developed in the latter part of the twentieth century and is associated for the most part with its originator, Dr Aaron Beck. CT was originally developed through Beck's work with depressed patients and is sometimes now referred to as 'classic' CT. While CT shares the core propositions of all CBTs, it differs to some extent from other CBTs in the organization of its theoretical underpinnings. The CT model proposes that emotional and behavioral responses are consequences of the manner in which an individual views, appraises, or perceives events around him/her (Beck, 1970). The content of the situation-specific appraisals differs based on the individual's activities and the appraisals can be accurate or distorted, as well as positive or negative. Thus, a principal goal in CT is to recognize and modify the patient's distorted appraisals of life events and make them more realistic and adaptive (Dobson and Dozois, 2009). Treatment is collaborative in nature and heavy emphasis is placed on educating the patient about the cognitive model. The 'collaborative empiricism' that is encouraged in CT involves the therapist designing learning experiences for the patients that can help them monitor their automatic thoughts; recognize how cognitions, affect, and behaviors influence each other; test the validity of their automatic thoughts and, if necessary, replace the distorted thoughts with more realistic and adaptive cognitions; and, finally, to identify and alter the underlying beliefs that make them susceptible to flawed patterns of thinking (Kendall and Bemis, 1983).

Cognitive therapists assume that experiences, and the meaning that patients attach to them, are idiosyncratic (Dobson and Dozois, 2009). Therefore, patients must take an active role in treatment and question the inferences that they make about their experiences. The therapist's role is to teach the patient the cognitive model and the underlying principles of CT, as well as to help the patient plan behavioral experiments that can serve as testing grounds for patients' hypotheses (i.e., thoughts and beliefs). Cognitive therapists do not lecture patients on a 'correct way to think,' nor do they assume the reasons why certain thoughts may cause distress. Rather, patients are treated as the experts of their own experiences, are asked to discuss their thoughts and mental images, and are taught how their appraisal of those mental events influences their emotional state. The CT therapist uses Socratic questioning, which is a technique that is characterized by the use of inductive questions, which encourage patients to think through different situations and come up with answers on their own.

Key Concepts

Cognitive Distortions and Automatic Thoughts

The concept of cognitive distortions was originally identified by Beck through his work with depressed patients. Cognitive distortions are errors that people sometimes make in their thinking (DeRubeis et al., 2009), which represent inaccurate assumptions that individuals make about themselves, the future, and the world. Some examples of cognitive distortions are all-or-nothing thinking (i.e., categorizing experiences as either good or bad), overgeneralizing (i.e., drawing global inferences from one occurrence), mind reading (i.e., making assumptions of what another person is thinking, despite having little evidence to support them), and making 'should' statements (i.e., thinking that one should or should not be doing something, rather than wishing to do it). CT teaches patients to attend to negative thoughts, to recognize when they make cognitive errors, and to correct these distortions based on evidence, where possible.

Automatic thoughts are surface-level thoughts that arise as part of our everyday experience and, therefore, can be readily identified (DeRubeis et al., 2009). In CT, patients are encouraged to recognize these thoughts, to evaluate them, and to explore the implications and meanings of habitual thoughts. Often, thoughts are viewed as hypotheses that need to be tested through behavioral experiments. For example, to test a thought such as 'nobody likes me at work,' a patient might have a conversation with his or her work colleagues and to observe their reactions to him or her, in order to determine whether the automatic thought receives support. It should be noted that automatic thoughts are not always negative in nature and that they can sometimes be positive and adaptive. If that is the case, and the patient is still experiencing emotional distress, the therapist would move on to identifying more ingrained concepts, such as beliefs and schemas, that can be potentially fueling the patient's negative state.

Core Beliefs, Assumptions, and Schemas

Cognitive schemas are fundamental cognitive structures, which serve to organize one's experience (DeRubeis et al., 2009). Schemas "represent the organized thought patterns that are acquired early in an individual's development" (Dobson and Dozois, 2009: p. 15), and that develop throughout the life span. Also referred to as 'core beliefs,' schemas are at the nucleus of the cognitive disturbance, and result in distorted perceptions, erroneous problem solving, and various psychological disorders (Dozois and Beck, 2008). Schemas can influence the type of activities that the individual chooses to engage in, which may, in turn, reinforce these core beliefs. When elicited by triggers, core beliefs often appear in an 'if–then' fashion (e.g., 'if I cannot be perfect, I am a failure'), which are also sometimes referred to as assumptions.

The 'deeper' level of work that is involved in CT involves work to identify and potentially change the patient's schemas or core beliefs (DeRubeis et al., 2009). Schemas are assessed through an in-depth analysis of the themes and patterns of thinking in the patient's negative experiences. Once the schemas are identified, their utility or validity is examined. Patients may be asked to list pros and cons of holding the core beliefs, as well as whether they have supporting evidence for them. This type of work helps patients to recognize the situations in which the schemas operate and what alternative inferences they might consider. Throughout treatment, the strength of the negative schemas is expected to diminish, as they become replaced with newer and more adaptive schemas.

Waves of CBTs

First Wave CBT

There have been three generations or 'waves' of therapies that can be ultimately referred to as CBTs (Hayes, 2004). Each wave has approached the patients' problems from a different angle, and therefore can be conceptualized as a different dimension of CT. The first wave began in the 1950s and was characterized by therapies of strictly behavioral nature that appeared as a reaction to the lack of empirical and theoretical support for the clinical paradigms of the day. Specifically, early behavior therapies were based on a set of basic principles that were established through scientific inquiry. In general, the behavioral therapies were founded on learning theory (e.g., conditioning) and conformed to well-established experimental paradigms. As these approaches grew, they were expanded to involve social learning models (Mischel, 1981), and what became called cognitive behavior modification (e.g., Meichenbaum, 1977). One of the goals of these approaches was to tackle problematic behaviors through cognitive assessment and modification.

Second Wave CBT

The second wave, referred to today as classic CT, integrated the central themes of the first wave behavior therapy, such as 'first-order change' (i.e., alteration of the form, frequency, and/or content of abnormal behavior; Hayes, 2004). Second wave CBT also addressed cognition from a direct and clinically relevant standpoint. It incorporated cognitive concepts and posited that irrational thoughts, maladaptive cognitive schemas, or faulty information-processing styles can be reduced or eliminated if they are detected, corrected, tested, and proven to be dysfunctional. Behavioral principles were still evident; however, cognitive concepts were emphasized to a much greater extent, and the main focus of the second wave was on information processing. Classic CBT is now a form of psychotherapy that is actively used around the world and has garnered a substantial amount of empirical support for its effectiveness in treating a variety of psychological disorders.

Although Beck's CT model viewed behavioral change as a means to test out the hypotheses regarding the patients' cognitions, it is not until the emergence of CBT in the late 1980s that behavioral strategies became fused with cognitive processes, with the goal of achieving behavioral and cognitive change. CBTs evolved and expanded over time, though, and can now be categorized into three classes, based on their relative focus on cognitive versus behavioral change: coping skills therapies, problem-solving therapies, and cognitive restructuring method (Dobson and Dozois, 2009).

Coping skills therapies focus on problems that are external to the patient (Dobson and Dozois, 2009), and aim to identify and alter the manner in which a patient might exacerbate the influence of negative events (e.g., by avoidance) and to teach him/her strategies to lessen the impact of those negative events (e.g., relaxation). The goal is to provide the patient with a repertoire of skills to promote coping in stressful situations. In contrast, cognitive restructuring therapies target more internal problems such as automatic thoughts, as well as long-term beliefs that lead to negative outcomes. An assumption of this type of therapy is that maladaptive thoughts result in emotional distress; therefore, the goal is to examine and challenge those maladaptive thought patterns and to establish more adaptive ones. Finally, the problem-solving type of CBT can be viewed as a combination of the coping skills and cognitive restructuring-focused approaches. Problem-solving therapies emphasize the patients' development and use of general strategies that can be used with a variety of personal

problems. These therapies also highlight the importance of an active collaboration in the therapeutic context.

Third Wave CBT

The third, and most recent, wave of CBT psychotherapies is distinguished by newer concepts, including metacognition, cognitive fusion, acceptance, mindfulness, dialectics, and spirituality (Hayes, 2004). Third wave therapies employ a variety of different techniques, but are similar to each other in that they emphasize the use of skills deficit models to examine the central maintaining mechanisms of various disorders (Hayes, 2004). Third wave psychotherapies include the following: acceptance and commitment treatment (ACT; Hayes, 2012), behavioral activation therapy (Martell et al., 2010), cognitive behavioral analysis system of psychotherapy (McCullough, 2000), dialectical behavioral therapy (DBT; Linehan and Koerner, 2012), metacognitive therapy (MCT; Wells, 2008), mindfulness-based cognitive therapy (MBCT; Segal et al., 2012), and schema therapy (Arntz and van Genderen, 2009).

Compared to first and second wave psychotherapies, the focus of third wave therapies is on the *function* of different cognitions and behaviors, rather than the content or accuracy of one's perception of his/her experience. Consequently, the emphasis is on 'second-order change,' which can be described as a fundamental shift in one's perception of reality. As with second wave CBTs, third wave therapies emphasize psychoeducation, and the patients' active practice with tools and techniques that they can apply to their everyday experience. Although third wave therapies are largely considered to be empirically supported (Kahl et al., 2012), currently there is not enough evidence to suggest that they are superior to classic CT. Below is a description of a number of third wave therapies.

ACT (Hayes, 2012) is based on the theory that the way individuals respond to mental events (e.g., thoughts) and experiences stems from the relationships they draw between them (Kahl et al., 2012). Thus, it is not the content of thoughts that is as critical in ACT, so much as the function that different types of thoughts serve. Treatment methods include psychoeducation about issues, such as nonjudgmental acceptance and embracement of one's present experience, exercises in mindfulness, and cognitive defusion (i.e., not taking one's emotional reaction as reflection of reality). Similar to CT, distancing from thoughts is considered to be a core skill, and negative thoughts are treated as beliefs rather than facts. During treatment, patients' values are elicited and discussed; however, unlike in CT, no restructuring techniques are employed, and behavioral homework assignments are not used as a means to restructure thoughts and to regulate mood, but to increase their psychological flexibility. Finally, heavy emphasis is placed on engagement in behavior that is driven by patients' values, instead of behavior that is driven by emotional and experiential avoidance. ACT has been shown to be an effective treatment for a variety of psychological disorders such as depression (Karlin et al., 2013) and mixed anxiety disorders (Arch et al., 2012).

MBCT (Segal et al., 2012) is closely related to ACT in its stance on the importance of embracing the present moment. It was originally developed as a relapse prevention therapy for depression. The key concepts employed in MBCT are psychoeducation about the nature of automatic thoughts, as well as how thoughts affect emotions and vice versa, and the practice of mindfulness and acceptance. The core goal is for patients to develop metacognitive awareness, but unlike in CT, the aim is not to change dysfunctional thoughts but to learn to monitor them, recognize them as thoughts and not facts, and distance oneself from them. There is substantial evidence for MBCT as maintenance therapy for preventing relapse in depression (cf Godfrin and van Heeringen, 2010) and some evidence for treatment of current symptoms of chronic depression (Barnhofer et al., 2009).

Another therapy that involves mindfulness is DBT (Linehan and Koerner, 2012). DBT was originally developed for parasuicidal patients who suffer from borderline personality disorder (BPD) and has been shown to be an effective treatment for it (e.g., McMain et al., 2009), but it has also been modified for a broader range of psychopathology such as substance use and eating disorders (Kahl et al., 2012). The core assumption of DBT is that skill deficits in the area of emotion regulation lie at the center of these disorders. As a result, DBT employs a broad "spectrum of skills in the areas of mindfulness, distress tolerance, emotion regulation, and interpersonal effectiveness" (Kahl et al., 2012: p. 524).

MCT (Wells, 2008) evolved directly from classical CT. It is based on the assumption that metacognitions (i.e., aspects of cognitions that control mental processes and the direction of thinking) need to change in order for psychological treatment to be effective. Unlike CT, there are no interventions that focus on the content of the patient's thinking (e.g., beliefs and their validity). Instead, MCT "uses attention training techniques to develop skills in cognitive flexibility, teaches a special form of mindfulness (detached mindfulness) and guides cognitive and behavioural experiments to change metacognition" (Kahl et al., 2012: p. 525). In other words, it focuses on the process of thinking, and presumes that the problem can be attributed to "inflexible and recurrent styles of thinking in response to negative thoughts, feelings, and beliefs" (Kahl et al., 2012: p. 525). Thus, one of the goals of MCT is the removal of those unhelpful styles of processing the information. It has been shown to be effective in treatment of generalized anxiety disorder (van der Heiden et al., 2012) and depression (Normann et al., 2014).

Similar to MCT, schema therapy (Arntz and van Genderen, 2009) is characterized by an emphasis on the concepts of schemas and modes. It was originally developed to treat personality disorders and other mental disorders of chronic nature (Kahl et al., 2012) and has been shown to be effective in the treatment of BPD (Farrell et al., 2009). The techniques in schema therapy are based on emotion activation methods of Gestalt and Psychodrama origin, but the patient is taught the models from a behavioral perspective. The therapist teaches cognitive techniques that can help the patient analyze and alter his/her dysfunctional beliefs about current, as well as past events.

A recent focus in CBT research has been on the deconstruction of cognitive therapies with a goal to find the active ingredients. One product of these component treatment studies has been behavioral activation therapy (Martell et al., 2010). It was found to be a stand-alone component with a "similar or superior efficacy compared with cognitive therapy" (Kahl et al., 2012: p. 523), especially in the context of

treatment of major depression (Dobson et al., 2008). The central aim in behavior activation is to supply the patient with stable and valued, yet diverse sources of positive reinforcement and meaning, in order to produce behavioral change (Martell et al., 2010). The key techniques include psychoeducation, activity monitoring, functional analysis of behavior, scheduling of nonavoidance activities, and trouble-shooting. It differs from classical CT in its emphasis on behavior change instead of cognitive change. Behavior activation is an active therapy and requires the therapist to take a problem-solving stance (Kahl et al., 2012).

Common Methods in Cognitive Therapies

As the technology related to the delivery of CBTs has grown, and new models and methods have been developed, the science related to the mechanisms of CBT has also grown. A considerable base of understanding about the necessary ingredients for successful CBT, as well as common factors in CBT has evolved, as described in the following sections.

Necessary Components of CBT

The CBTs are forms of psychotherapy, and as such contain aspects that are often seen in models of psychotherapy. For example, therapeutic alliance is an essential ingredient in all psychotherapies, including CBT (Beck and Dozois, 2011), and was shown to be modestly related to the total variance in therapy outcome (Horvath et al., 2011). CBT therapists, like other psychotherapists, require good interpersonal skills, such as warmth and unconditional positive regard, to facilitate their rapport with the patient (cf Castonguay et al., 2006). It is important for the patient to be comfortable with the therapist, to disclose his/her innermost thoughts and beliefs. Thus, building a good relationship is paramount; however, it is viewed as a necessary, but not sufficient condition for successful therapy outcome. The unique aspect of the therapeutic relationship that distinguishes CBT from other psychotherapies is the collaborative working alliance that is characterized by a mutual understanding of treatment goals.

Furthermore, the concept of collaborative empiricism in CBT involves the patient and therapist working together to attain treatment goals and to investigate the patient's thoughts and beliefs (Beck and Dozois, 2011). The rationale behind collaborative empiricism is that the client evaluates his/her intrinsic data (e.g., beliefs) based on empirical evidence and with the help of the therapist (cf Tee and Kazantzis, 2011). The two techniques that are most commonly employed are Socratic questioning/dialogue and guided discovery (cf Overholser, 2011). Based on the view that patients are experts on their own condition, guided discovery uses personally directed experiments to examine alternative thoughts and behaviors (Beck and Dozois, 2011), rather than simple persuasion by the therapist as to what course of action to take. The goal of the therapist is to instill a sense of curiosity in the patients and get them to think about their everyday experience. Socratic questioning is a form of guided discovery, which involves asking a series of inductive questions to assist the patient arrive at the answers on their own, based on existing evidence. In some cases, it may be possible that during this process the patient arrives at normative thoughts and convictions that oppose general evidence, but are nevertheless preferable to him/her. Depending on the issue at hand, the therapist can further attempt to help guide the patient toward appraisals that are adaptive and in line with existing evidence; however, in some cases, if unsuccessful, the intervention may need to be postponed until the patient is sufficiently ready to make the changes necessary to attain his/her treatment goals.

Psychoeducation is another important technique in CBT. Therapists educate patients about the therapeutic model and familiarize them with the key concepts and methods as they work with the patient throughout treatment. Psychoeducation equips patients with the knowledge of what skills are being taught, and how to broadly apply these skills, so that patients can use these skills during therapy and into the future, if the problem recurs.

Behavioral strategies are a common component of effective CBT. In second wave CBT, behavioral exercises are used to test the validity of patients' hypotheses (i.e., thoughts and beliefs), and to gather objective evidence on whether to accept, reject, or assimilate these ideas. The behavioral techniques also include exercises, such as exposure, to habituate the patient to feared stimuli, and relaxation, with the goal to encourage adaptive coping and cognitive change (Beck and Dozois, 2011). Behavioral exercises are also employed in third wave CBT, for example, to anchor patients in the present moment. For instance, the body scan exercise is used in MBCT, which invites patients to hold parts of their body in awareness, and to help them to become conscious of the physical sensations that are characteristic of the present moment.

Another central technique in CBT is thought monitoring. The patient is taught to become aware of his/her thoughts and to notice how they affect emotional outcomes and behavioral choices. In second wave CBT, the therapist may then help the patient to recognize maladaptive thoughts, and to restructure or eliminate those thoughts. In contrast, the focus in third wave CBT may be to become aware of those thoughts and to disengage from them. This idea is exemplified in the differences between the concepts of distancing and decentering, which are also techniques that are considered to be essential in CBT. Distancing, which is emphasized more in second wave therapies, involves recognizing thoughts as beliefs and internal events rather than facts. On the other hand, decentering is more typical in third wave CBT, and involves viewing thoughts as beliefs, recognizing their impermanent nature, and learning to let go of them.

The assignment of homework is a component that is present in all CBTs and is integral to therapeutic success. Patients need to take an active role in their recovery, and the completion of homework assignments increases the sense of mastery about new skills, and provides increased control over interpersonal, environmental, and other conditions (Beck and Dozois, 2011). Homework also allows patients to practice the newly learned skills in the 'real world,' which aids with the generalizability of the skills. Finally, successful homework provides accomplishments, which can solidify the concepts that patients learned in session. For example, a patient who is afraid of public speaking, but who learns that the feeling of discomfort that is experienced when giving a presentation at work eventually decreases with

practice, is likely to be encouraged to try to continue giving other public presentations.

Barlow's Unified Protocol

As noted above, research has examined the utility of certain components of CBT and their utility for a range of psychological disorders. More recently, Barlow et al. (2010) have described a 'unified protocol,' which they proposed is a set of principles based on cognitive behavioral theory that can be used to treat many emotional disorders. Unlike the applications of CBT to specific disorders, the unified protocol does not target specific symptoms, but rather emphasizes the underlying mechanisms that exist along the full spectrum of disorders. The unified protocol presumes that individuals with emotional disorders use a set of maladaptive strategies to regulate their emotions, which are used to either avoid or diminish their unpleasant emotions, but that these strategies are counterproductive in that they perpetuate the very emotions they used to avoid. The goal of unified protocol is not to eliminate negative emotions, but, instead, to get patients to learn to tolerate them at a more functional level and learn that negative emotions can have adaptive functions. The protocol is essentially a skills building model, which can be conducted in much the same fashion as other forms of CBT.

The unified protocol lists five core skills that are combined into three modules that are introduced sequentially, albeit with flexibility, depending on the patients' individual needs and understanding of the treatment approach. The first core skill is called present-focused emotion awareness and presents patients with a model that emphasizes the interaction among thoughts, feelings, and behaviors. The assumption is that this model of emotion provides patients a more objective understanding of their emotional experience, which, in turn, fosters their emotional awareness. Emotional awareness is viewed as a necessary skill for understanding of the treatment concepts that are subsequently introduced.

The second core skill in the unified protocol is cognitive flexibility, which involves the challenge of the patients' negative evaluations of external and internal events (e.g., physical sensations and emotions) and an increase in cognitive flexibility. Adapted from Beck's concepts of cognitive distortions, a focus is placed on probability overestimation and catastrophizing. Specifically, patients are instructed to think about how likely the negative event is to occur and how negative can the consequences of that event be if it did occur.

Recognition of emotion avoidance and emotion-driven behaviors is the third core skill in the unified protocol. This skill involves identification and modification of maladaptive patterns of behaviors driven primarily by emotions (Barlow et al., 2010). The therapist teaches patients to identify patterns of their maladaptive thinking and change them, in order to regulate their emotions. Relatedly, the fourth core skill is awareness and tolerance of physical sensations, which is attainted through interoceptive exposure practices. The goal is to, again, bring patients' awareness to the physical experiences of emotion and to teach them the reciprocal nature of the relationships among sensations, thoughts, and feelings. Finally, the fifth core skill involves 'exposure exercises.' The goal of exposure exercises is to elicit emotional experiences by exposing patients to situations that evoke the emotion of interest (e.g., anxious emotions in social situations for someone with social anxiety). The emotional experience exercises typically range from least intense to most intense. Patients are encouraged to fully experience each situation, so that the therapist can identify potential avoidance or safety behaviors. The therapist and patient also work on reduction of safety behaviors and the acquisition of new learning experiences that do not result in negative emotions.

The 'unified protocol' is still an emergent concept in the CBT literature; therefore, more research is needed to determine its effectiveness in treatment of emotional disorders. However, there is preliminary evidence to suggest that it is efficacious in treatment of depression and anxiety disorders (Farchione et al., 2012; de Ornelas et al., 2013). Finally, it should be noted that other experts in the field of psychotherapy have suggested core skills that improve the efficacy and effectiveness of psychotherapy that are different from those listed in Barlow and colleagues' 'unified protocol' (cf Grawe, 2004); however, they are not discussed here because of their differing theoretical basis.

Future Research Directions

Despite the large amount of extant research on CBT therapies and mechanisms, and the existence of several journals that disseminate this work, several aspects of CBT require more research. Several of the key issues that currently exist include the relative effectiveness of the third wave therapies, long-term effects of CBT, optimizing delivery of CBT, understanding the mechanisms of change, and developing a better understanding of its suitability for, or limitations with various psychological problems.

First, there is a need for more research that compares the effectiveness of third wave therapies to classic CT, so as to determine whether or not they result in superior outcomes (Kahl et al., 2012). As well, more data are needed with respect to their suitability for certain psychological disorders and their limitations. Researchers have urged more meditational analyses of the effects of third wave therapies, specifically in relation to different emotion regulation strategies and whether they have an additional benefit to standard CBT. Third wave CBT also opens the possibility of tailored emotion regulation strategies, which might be selectively used with those patients who are most likely to benefit (Hoffman et al., 2010).

There is a need for more data regarding the most effective use of cognitive therapies. Research needs to examine the extent to which therapists are able to learn cognitive therapies and to follow treatment protocols with their patients in various contexts and regular clinical practice (DeRubeis et al., 2009). There is a further need for research on the mechanisms of change in cognitive therapies that result in the most stable benefits (Beck and Dozois, 2011). Particularly, there is a need for a better understanding of the interplay among "cognitive, affective, motivational, and behavioural" (Beck and Dozois, 2011: p. 406) factors in cognitive therapies. One limitation of process research is that it tends to be correlational in nature. Experimental studies are needed to supplement correlational research and to document the processes that lead to a positive therapeutic outcome (DeRubeis et al., 2009).

Continued research regarding the applicability of CBT to various disorders, as well as the long-term outcomes of treatment is needed. While research has primarily focused on relapse prevention in the months after the termination of treatment, studies need to be conducted to include even longer follow-up assessments, to provide estimates of both medium effects and long-term effects of cognitive therapies (DeRubeis et al., 2009). Comparative long-term analyses, including cost–benefit analyses with other modes of therapy, would also provide important information from the perspective of health-care policy and funding.

See also: Manual Based Treatments; Psychological Treatments, Empirically Supported; Psychological Treatments: Randomized Controlled Trials; Psychotherapy Integration; Psychotherapy Process Research; Therapist–Patient Relationship.

Bibliography

Arntz, A., van Genderen, H., 2009. Schema Therapy for Borderline Personality Disorder. Wiley-Blackwell, Chichester.

Arch, J.J., Eifert, G.H., Davies, C., Vilardaga, J.C., Rose, R.D., Craske, M.G., 2012. Randomized clinical trial of cognitive behavioral therapy (CBT) versus acceptance and commitment therapy (ACT) for mixed anxiety disorders. Journal of Consulting & Clinical Psychology 80, 750–765.

Barlow, D.H., Farchione, T.J., Fairholme, C.P., 2010. Unified Protocol for Trans-diagnostic Treatment of Emotional Disorder: Therapist Guide. Oxford University Press, USA, Cary, NC, USA.

Barnhofer, T., Crane, C., Hargus, E., Amarasinghe, M., Winder, R., Williams, J.M.G., 2009. Mindfulness-based cognitive therapy as a treatment for chronic depression: a preliminary study. Behaviour Research and Therapy 47, 366–373.

Beck, A.T., 1970. Cognitive therapy: nature and relation to behavior therapy. Behavior Therapy 1, 184–200.

Beck, A.T., Dozois, D.J., 2011. Cognitive therapy: current status and future directions. Annual Review of Medicine 62, 397–409.

Castonguay, L.G., Constantino, M.J., Holtforth, M.G., 2006. The working alliance: where are we and where should we go? Psychotherapy: Theory, Research, Practice, Training 43, 271–279.

DeRubeis, R.J., Webb, C.A., Tang, T.Z., Beck, A.T., 2009. Cognitive therapy. In: Dobson, K.S. (Ed.), Handbook of Cognitive-Behavioral Therapies, third ed. Guilford Press, New York.

Dobson, K.S., Hollon, S.D., Dimidjian, S., et al. 2008. Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the prevention of relapse and recurrence in major depression. Journal of Consulting and Clinical Psychology 76, 468–477.

Dobson, D., Dobson, K.S., 2009. Evidence-based Practice of Cognitive-Behavioral Therapy. Gilford, New York.

Dobson, K.S., Dozois, D.J.A., 2009. Historical and philosophical bases of the cognitive-behavioural therapies. In: Dobson, K.S. (Ed.), Handbook of Cognitive-Behavioral Therapies, third ed. Guilford Press, New York, pp. 3–38.

Dozois, D.J.A., Beck, A.T., 2008. Cognitive schemas, beliefs and assumptions. In: Dobson, K.S., Dozois, D.J.A. (Eds.), Risk Factors in Depression. Elsevier/Academic Press, Oxford, England, pp. 121–143.

Epp, A., Dobson, K.S., 2010. The evidence base for cognitive-behavioral therapy. In: Dobson, K.S. (Ed.), Handbook of Cognitive-Behavioral Therapies. Guilford, New York, pp. 39–73.

Farchione, T.F., Fairholme, C.P., Ellard, K.K., Boisseau, C.L., Thompson-Hollands, J., Carl, J.R., Gallagher, M.W., Barlow, D.H., 2012. Unified protocol for trans-diagnostic treatment of emotional disorders: a randomized controlled trial. Behavior Therapy 43, 666–678.

Farrell, J.M., Shaw, I.A., Webber, M.A., 2009. A schema-focused approach to group psychotherapy for outpatients with borderline personality disorder: a randomized controlled trial. Journal of Behavior Therapy and Experimental Psychiatry 40, 317–328.

Godfrin, K.A., van Heeringen, C., 2010. The effects of mindfulness-based cognitive therapy on recurrence of depressive episodes, mental health and quality of life: a randomized controlled study. Behaviour Research and Therapy 48, 738–746.

Grawe, K., 2004. Psychological Therapy. Hogrefe & Huber Publishers, Ashland, OH, US.

van der Heiden, C., Muris, P., van der Molen, H.T., 2012. Randomized controlled trial on the effectiveness of metacognitive therapy and intolerance-of-uncertainty therapy for generalized anxiety disorder. Behaviour Research and Therapy 50, 100–109.

Hayes, S.C., 2012. Acceptance and Commitment Therapy. American Psychological Association, Washington.

Hayes, S.C., 2004. Acceptance and commitment therapy, relational frame theory, and the third wave of behavioral and cognitive therapies. Behavior Therapy 35, 639–665.

Hofmann, S.G., Sawyer, A.T., Fang, A., 2010. The empirical status of the 'new wave' of cognitive behavioral therapy. Psychiatric Clinics of North America 33, 701–710.

Horvath, A.O., Del Re, A.C., Fluckiger, C., Symonds, D., 2011. Alliance in individual psychotherapy. Psychotherapy 48, 9–16.

Kahl, K.G., Winter, L., Schweiger, U., 2012. The third wave of cognitive behavioural therapies: what is new and what is effective? Current Opinion in Psychiatry 25, 522–528.

Karlin, B.E., Walser, R.D., Yesavage, J., Zhang, A., Trockel, M., Taylor, C.B., 2013. Effectiveness of acceptance and commitment therapy for depression: comparison among older and younger veterans. Aging & Mental Health. 17, 555–563.

Kendall, P.C., Bemis, K.M., 1983. Thought and action in psychotherapy: the cognitive-behavioral approaches. In: Hersen, M., Kazdin, A.E., Bellack, A.S. (Eds.), The Clinical Psychology Handbook. Pergamon, New York, pp. 565–592.

Linehan, M.M., Koerner, K., 2012. Doing Dialectical Behavior Therapy. Guilford, New York.

Martell, C.R., Dimidjian, S., Herman-Dunn, R., 2010. Behavioral Activation for Depression. Guilford, New York.

McCullough, J.P., 2000. Treatment for Chronic Depression. Cognitive Behavioral Analysis System of Psychotherapy (CBASP). Guilford, New York.

McMain, S.F., Links, P.S., Gnam, W.H., Guimond, T., Cardish, R.J., Korman, L., Streiner, D.L., 2009. A randomized trial of dialectical behavior therapy versus general psychiatric management for borderline personality disorder. The American Journal of Psychiatry 166, 1365–1374.

Meichenbaum, D.H., 1977. Cognitive Behavior Modification. Plenum, New York.

Mischel, W., 1981. A cognitive-social learning approach to assessment. In: Merluzzi, T., Glass, C., Genest, M. (Eds.), Cognitive Assessment. Guilford, New York, pp. 479–502.

Normann, N., van Emmerik, A.A.P., Morina, N., 2014. The efficacy of metacognitive therapy for anxiety and depression: a meta-analytic review. Depression and Anxiety 31, 402–411.

de Ornelas, M.A.C., Braga, A.A., Nunes, C.A., Nardi, A.E., Silva, A.C., 2013. Trans-diagnostic treatment using a unified protocol: application for patients with a range of comorbid mood and anxiety disorders. Trends in Psychiatry and Psychotherapy 35, 134–140.

Overholser, James C., 2011. Collaborative empiricism, guided discovery, and the socratic method: core processes for effective cognitive therapy. Clinical Psychology: Science and Practice 18, 62–66.

Segal, Z., Williams, J.M., Teasdale, J., 2012. Mindfulness-based Cognitive Therapy for Depression, second ed. Guilford Press, New York, NY, USA.

Sensky, T., 2004. Cognitive-behavior therapy for patients with physical illnesses. In: Wrights, J. (Ed.), Cognitive-Behavior Therapy. Review of Psychiatry. APA Publishing, Arlington, VA, pp. 83–121.

Tee, J., Kazantzis, N., 2011. Collaborative empiricism in cognitive therapy: a definition and theory for the relationship construct. Clinical Psychology: Science and Practice 18, 47–61.

Wells, A., 2008. Metacognitive Therapy for Anxiety and Depression. Guilford Press, New York.

Relevant websites

http://www.academyofct.org/ – Academy of Cognitive Therapy.
http://www.abct.org/ – Association of Behavioral and Cognitive Therapies.
http://www.eabct.com/ – European Association of Behavioral and Cognitive Therapies.
http://www.the-iacp.com/ – International Association of Cognitive Psychotherapy.

Conceptual History: *Begriffsgeschichte*

Mario Wimmer, University of California Berkeley, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Conceptual history is a mode of historical inquiry concerned with the problem of the change in the semantics of particular concepts. The article outlines its historical formation in both the disciplines of history (R. Koebner, O. Brunner, C. Schmitt, R. Koselleck) and of philosophy (R. Eucken, R. Eisler, E. Rothacker), its further institutional and intellectual development, as well as an overview of the major methodological issues. This article also puts conceptual history in the context of other fields with similar methodological concerns such as the history of ideas and mentalities, intellectual history, historical epistemology as well as Luhmann's sociology.

Conceptual history is a field of historical study concerned with the problem of the change in the semantics of particular concepts. It engages with a general problem in the writing of history: the difference between the language of the sources and the analytical categories of the historian, more specifically the change in semantics over time, as every historical-empirical approach has to deal with the problem of the changing semantics of historical sources, be they texts or other artifacts from the past.

This problem has a long history in historiography. Conceptual history is one specific epistemological formation in this long history of a more general problem, the study of the historicity of the conditions of the emergence of meaning and sense, experience and perspective. Ever since the formation of the historical disciplines in the late eighteenth century the criticohistorical reading of written documents has been one of their basic tools and tasks. These issues have been the subject of international debates e.g., as language and written matter as the necessary condition of historical thought (a widely discussed issue in historical methodology in classical textbooks from Johann G. Droysen to Charles Langlois/Charles-Victor Seignobos), the problem of symbolic (Ernst Cassirer) or intellectual (Johan Huizinga) forms, or under the label of the *linguistic turn* (Richard Rorty). Since the 1960s this universal challenge of historical thinking was reconsidered as the study of discourses and their various institutional structures and material conditions that contribute to the emergence of historical formations (epistemes) of cultural production (Michel Foucault), or more recent debates in labor history on how the experience of class consciousness is structured by language (Jones, 1983; Scott, 1974; Rancière, 1989).

Ever since conventional representation has become problematic at the beginning of the twentieth century, with a transformation for example witnessed by the fictional "Letter of Lord Chandos" (Hofmannsthal, 2005) and, anticipated by Friedrich Nietzsche, this crisis of language became thematic in such different contexts as philosophy of language, pragmatism, Gottlob Frege's logic, Edmund Husserl's phenomenology, Ferdinand de Saussure's linguistic, or Freudian psychoanalysis. In other words: the situation between history and its language had significantly changed and generated fresh scholarship on the problem. In the German context it has been conceptual history or history of concepts (*Begriffsgeschichte*) that dealt with the phenomenological problem of historical sources and their language in a particular way: Scholars deemed past events or a set of historical conditions that are incorporated as condensed experience in concepts or notions. The language of the sources contributes to their formation; thus, we need to engage with the historicity of the concepts of analysis with respect to the language of historical sources. As a consequence conceptual history studies the changing semantics of concepts or notions over time. Indeed, it is not self-evident that concepts *do*, or at least *can*, have a history. It is only if we acknowledge the event-character of discursive practices that concepts themselves can be read in historical context.

While *Begriffsgeschichte* is still a well established field in the history of philosophy and philosophical terminology, nowadays most historians along with political and social scientists, linguists or cultural theorists prefer an approach that allows for a broader engagement with semantic constellations and discursive practices. This is not the only sense in which the history of concepts is not a strictly formal discipline but a set of methodological tools that can follow different – historical, philosophical, or cultural – paths. Since the 1980s there has been a gradual shift from the study of individual concepts toward a more general interest in the systematic study of language, or as Willibald Steinmetz (born 1957) recently put it, semantic 'situations' as a basis for historical-social, cultural, and conceptual understanding. Conceptual history and, the larger field of so-called historical semantics, enjoyed their most formative phase in parallel with the development of historical metaphorology, i.e., the study of the change of metaphors over time – though, the latter is often ostracized. (Bödeker, 2002) All these approaches share specific goals and certain characteristics with the history of ideas, discourse analysis, the history and philosophy of science, as well as–in different ways – lexicology and the linguistic study of semantic change.

Initially, conceptual history was mostly considered a 'German' approach, which is commonly traced back to two monumental projects, the *Historisches Wörterbuch der Philosophie* (Dictionary of Philosophical Terminology) and the lexicon *Geschichtliche Grundbegriffe* (thereafter *Grundbegriffe*, Basic Concepts in History: An Historical Dictionary of Political and Social Language in Germany), as well as the work of the historian Reinhart Koselleck (1923–2006), one of the most influential German historians in the twentieth century and an

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03074-9

original theorist of history. The project of the *Grundbegriffe* was derived from comparative studies of the semantic changes in 122 basic sociopolitical concepts in different European languages (with a focus on German) around 1800, in some cases reaching back to antiquity. The understanding of basic concepts (*Grundbegriffe*) for the encyclopedic project of the *Grundbegriffe* was distinguished not only by its exclusion of concepts that define or categorize (*Leitbegriffe*), keywords (*Schlüsselwörter*), catchphrases or slogans (*Schlagwörter*) (for the history of buzz-words see Bauer, 1920; Bahner, 1963), but also by the attempt to at least try to consider multilinguism and thus reference texts not only in German (Old, Middle, and modern), but also classical and medieval Latin, classical Greek, French, English, and Italian.

Only a few publications in English provide a good starting point and introduction to the field: Melvin Richter's *The History of Political and Social Concepts* (1995); the more focused books by Niels Akerstrom Andersen (2003) and Niklas Olsen (2012); several introductions to the oeuvre of Koselleck give broad contextualizations (see also the 1989 special issue of *History Workshop Journal* on 'Language and History,' in particular Schöttler, 1989).

Conceptual History before and after Koselleck

To my knowledge, the use of the word *Begriffsgeschichte* goes back to G.W.F. Hegel's lectures on the Philosophy of History where he introduced it as a variation on 'reflexive history,' as a form of cultural and intellectual history that defines its objects close to philosophy. This rather particular understanding of the history of concepts has not prevailed. The German tradition of *Begriffsgeschichte* has instead other origins in the fields of philosophy and history. Rudolf Eucken's (1846–1926) *History and Critique of Basic Concept of the Present Day* (1878) along with his *History of Philosophical Terminology* (1879) were among the first publications in the history of concepts published in German language, though different from our current understanding in terms of its conception and approach. Another early example in the study of the historical formation of philosophical concepts was Rudolf Eisler's (1873–1926) single-authored *Dictionary of Philosophical Terminology and Notions* first published in 1899. Eisler was an Austrian philosopher, who remained a private scholar without a university career due to antisemitic prejudices in German-speaking academia. Yet, although between 1943 and 1948 four new philosophical dictionaries would appear in German, none with success, and until the first issue of the new periodical *Archiv für Begriffsgeschichte* (Archive for Conceptual History, edited by the German philosopher Erich Rothacker, 1888–1965) was published in 1955, the revised edition of Eisler's multivolume *Wörterbuch* remained the standard reference for academic philosophy besides the fact that it was not systematic in its philological approach to the history of philosophical terminology.

Rothacker's *Archiv* for its part was meant to be a collection of studies in the historical semantics of mostly philosophical concepts as a resource for a future encyclopedia of conceptual history. Indeed, its contributors were – and still are – mostly philosophers and intellectual historians. For decades it served as a platform for vivid discussions and intellectual exchange in both the fields of conceptual history and historical metaphorology. Most contributions gave crisp, sometimes comparably narrow historical overviews on certain concepts occasionally grouped around thematic issues, but it also allowed for long, almost monographic articles on important topics, among them the remarkable book-length essay by Hans Blumenberg, *Paradigms for a Metaphorology*. This essay not only earned him a university position but became an instant classic and was later published as a book in its own right (Blumenberg, 2010 (German 1960)).

It was only in 1972 when Joachim Ritter (1903–74), Karlfried Gründer (1928–2011), and Gottfried Gabriel (born 1943) initiated the multivolume Dictionary of Philosophical Terminology when Rothacker's vision of a complete historical survey of philosophical concepts began to become reality. Nonetheless the editors began to present their own project as a complete reedition of Eisler's *Wörterbuch*, even though the many volumes of the *Archiv* had done much of the spadework. By the time that the last volume of the *Wörterbuch* was published in 1997 it had become one of the most important enterprises of German *Geisteswissenschaften* after World War II; it contains about 6000 entries on 3670 philosophical concepts by more than 1500 authors. What distinguishes their various contributions is that they do not aim at fixed and dogmatic definitions of philosophical concepts, but every article discusses the origin, evolution, and transformation of the meaning of a concept or notion through systematic reference to source texts.

In the discipline of history, as opposed to philosophy, the situation was similar only in one respect: Again, conceptual history was pioneered by a German-Jewish scholar, Richard Koebner (1885–1958), whose work was again marginalized by his German colleagues even though his approach also followed, at least to some degree, in their then dominant German 'Research of the East' (*Ostforschung*), which since the 1920s contributed to the planning of ethnic cleansing of local indigenous populations and settlement of Germans aiming at "Germanizing" Central and Eastern Europe (Michael Burleigh, 1988). Koebner's work in the field started with his Berlin dissertation on the concept of marriage in late medieval Germany (Koebner, 1911). His Habilitation on the history of Cologne in 1919 involved discussions of several methodological problems in conceptual history; it was followed by a series of articles on various concepts. In 1934, after he was forced out of Nazi Germany and accepted a position at the Hebrew University in Jerusalem, he published his last article in German, a groundbreaking study on conceptual problems in cultural history. However his seminal article 'Semantics and Historiography' (1953) together with other important contributions to the field of conceptual history (on the history of the concepts of Western Civilization, Empire, or Imperialism) should have inspired those German historians, who during and following World War II established the seemingly new field of *Begriffsgeschichte*. Instead, they often referred to older studies in German literature and linguistics – as demonstrated in detail by Jan-Marco Sawilla (2004) – thus Koebner and his work were and mostly remain ignored – except for the homage paid him by Koselleck on the occasion of a talk on the history of concepts he delivered in Israel (Koselleck, 2006; the article was recently

also published in German translation in Joas and Vogt, 2011). Similar claims could be made for other exiled German scholars such as Paul Oskar Kristeller, Ernst Cassirer, or Karl Mannheim.

Concepts in the History of Ideas

Early conceptual history had also structural similarities with new disciplines emerging in the early twentieth century, such as the history of ideas and mentalities. In Germany it was most prominently the work of Friedrich Meinecke that was driven by a similar interest in the change of political ideas over time (e.g., Meinecke, 1924). The work of the American philosopher and intellectual historian Arthur O. Lovejoy (1873–1962) in the history of ideas shares various similarities with the study of the history of concepts. His attempt to identify what he called 'unit-ideas,' or individual concepts, as well as the study of their change over time suggests an approach rather concerned with intellectual and cultural than social and political history; precisely because of the reduction of his argumentation to ideas or concepts as basic 'units' of historical study. In his famous *The Great Chain of Being* (1933) he outlined and pursued an approach that in many respects shared interests with research in the history of concepts and German History of Ideas.

The so-called Cambridge School (Quentin Skinner, J.G.A. Pocock, among others) in intellectual history followed in some respects the model set by Lovejoy's history of ideas. However, there were few affinities with German *Begriffsgeschichte*. Indeed, Pocock admitted to not even being aware of the German studies in the history of political language until he was invited to contribute to a workshop on the occasion of the completion of the *Grundbegriffe*. (Lehmann and Richter, 1996) Nonetheless, different attempts have been made to find common ground between these different approaches, even though the main bone of contention always was – and still remains – the relation between the 'internal' logic of political language of the past and its 'external' reference to history proper (what Koselleck liked to call '*Sachgeschichte*'); in other words: a focus on intellectual and political language on the on side of the channel, and the attempt to relate semantic change to social and political history. While the Cambridge method was based on language philosophy (Ludwig Wittgenstein (1889–1951), John L. Austin (1911–60) among others) and exclusively focused on the analysis of political language and its relation to political theory, the German approach engaged emphatically with the dynamic between the history of concepts and their relation to experience of the historical past and not only in that respect to social history at large. (For a comparative approach of Anglo-Saxon and German conceptual history see Palonen, 2004; Leonhard, 2004.)

The French Approach: the History of Mentalities and Philosophy of Science

In France at the turn from nineteenth to twentieth century an interdisciplinary group of scholars formed around the philosopher Henri Berr (1863–1954). The group's intellectual gravity shifted toward historical study and Berr's ideal of synthesis. This collaborative spirit contributed to the founding of the journal *Annales* (in 1929) by the eminent historians Marc Bloch (1886–1944) and Lucien Febvre (1876–1956). They involved themselves in the historical study of mentalities, thus also paying attention to relation between language and history (see also the considerations of Barthes, 1981). In a rubric of the new journal Febvre published a series of articles (e.g., on 'frontier,' 'civilization,' or 'work') under a separate section entitled '*Les mots et les choses*' (*Words and Things*). Bloch and Febvre were both engrossed by historical-comparative linguistics as well as Saussurian structural linguistics. Their colleague Berr also edited a volume on the history of the concept of *civilization* with contributions by Febvre, Marcel Mauss, and Emile Tonnelat, among others. The volume was a signal of a different method of writing the history of a concept as it aimed at transnational comparisons and was critical of the idea of progress. Although it belonged to the tradition of the Enlightenment encyclopedia, even the *Encyclopédie française*, edited by Febvre and Anatole de Monzie, can be viewed as an early project in the history of concepts insofar as it followed a radical humanist, theoretical, and, only to some extent, historical conception of a total synthesis of diverse elements that allowed for a comprehensive thinking of our world (Schöttler, 2005).

Surprisingly, one can find hardly any connection between German conceptual history and the major studies in the history of scientific concepts in French and American History and Philosophy of Science. Discourse analysis and the work of Michel Foucault (1926–84) may be considered one of the few exceptions in this respect, even so the work of the French-American tradition of historical epistemology continued to be as ill-informed of the project of *Begriffsgeschichte* as German and other international scholars in this field are of the works of Georges Canguilhem (1904–95) or Ian Hacking (born 1936). Ever since the introduction of an 'historical apriori,' i.e., the historicity of the conditions of possibility of perception, by Edmund Husserl (1859–1938) to the philosophy of science, French, and later North American, historians and philosophers of science have become concerned with the history of scientific concepts. Thus, Gaston Bachelard (1884–1962) argued convincingly that knowledge could only be adequately understood if studied empirically through its constituent concepts in their historical context and development. The earliest examples of studies in the history of scientific concepts are Canguilhem's books on the emergence and development of the concepts of *reflex* (Canguilhem, 1955) and of *evolution* (Canguilhem, 1952). Both books make compelling cases for a new way of writing the history of science as a history of scientific concepts rather than a history of theories that employ those concepts. It was Foucault who steered this approach toward an analysis of institutionalized modes of speaking, its rules, and arrangement as well as the historical conditions for the possibility of statements; he was less interested in individual concepts so much as what he termed discursive practices and formations. The Canadian historian and sociologist Régine Robin (born 1939) in turn adopted this approach for the quantitative study of historical semantics, including microanalyses of individual concepts. The Canadian-French analytical philosopher Hacking engaged in this tradition when he published his study on *The Emergence of Probability* (1975). Ever since then Hacking has

made important contributions not only to the methodology but also to the actual history of scientific concepts inspiring studies by Lorraine Daston (born 1951), Peter Galison (born 1955), and others on concepts like *objectivity*, *probability*, or *empiricism*.

The Lexicon Geschichtliche Grundbegriffe

Returning to the eminent historian of conceptual history, Reinhart Koselleck, let us remember that although he considered himself an academic outsider, he would become not only the executive editor of one of the big collaborative projects in twentieth-century humanities, but also one of the leading scholars in the history of concepts; and certainly the strongest voice in attempting to develop a theoretical and methodological basis for *Begriffsgeschichte* in the larger context of a social history of politics and the public sphere. In Koselleck's understanding, *Begriffsgeschichte* was not only "concerned with political or social events and circumstances" (Koselleck, 1985(German 1979): p. 85) articulated in the language of their respective historical sources, but the basic concepts also indicated the "process of transformation into modernity"; only this nexus of approaches would allow a deeper understanding and reflexive use of our language today. In that sense, evolving concepts would embody novel experiences and relate to shifts and ruptures in semantics. The transformation of basic concepts of around 1800 (the so-called '*Sattelzeit*,' which translates to English as 'saddle-time' or threshold) introduced a temporal dimension to the formation of meaning; according to Koselleck they are 'Janus-faced': they capture both past and future perspectives of this period of transformation. Koselleck assumed that around 1800 key concepts changed and new concepts emerged, which embodied historical time, for example 'development,' 'revolution,' or the famous 'history as such' (history in the collective singular), and that this change in semantics would parallel a new understanding of European modernity.

This close connection between the language of the sources and our present understanding of past events was predominantly informed by the work of both the Austrian medievalist Otto Brunner (1898–1982) as well as by the political theorist Carl Schmitt (1888–1985). Brunner's book *Land and Lordship*, originally published as a study in German *Volksgeschichte*, was concerned with the language of sources in opposition to an approach that employs concepts from legal and constitutional theory in order to write the history of the State. Brunner's work became seminal for the methodology of German conceptual history and – in a reedited version that replaced the ideological Nazi terminology like race, people, etc. by 'neutral' notions like 'structure' (Gadi Algazi, 1997) – is to date considered a classic in medieval studies. Koselleck emphasized that without the work of Brunner the lexicon "would have been unthinkable," (Koselleck, 1972) Schmitt for his part had been Koselleck's mentor during his dissertation work on *Critique and Crisis* at Heidelberg; it was Schmitt, who suggested focusing on the changing of the two title concepts alongside others like 'war,' 'civil war,' and 'revolution.' The historical study of the change of meaning of concepts aided by crisp historical arguments and original theoretical points became distinctive features of his *oeuvre*. *Critique and Crisis* was received as a contribution to both history and sociopolitical theory, but criticized for its politically conservative attitude (most prominently by Habermas, 1977). Although often not considered part of a school, Koselleck and his early work was perceived as being part of a politically conservative group of German historians. (cf Olsen, 2012)

With his habilitation *Prussia between Reform and Revolution* (1965) Koselleck's work became reasonably independent from the early political and methodological circumstances mentioned above. He finally committed to the historical discipline. From now on he insisted on the necessity of theory for the historical discipline by coining a whole set of methodological concepts, which were based on his research in historical semantics and translate a set of philosophical ideas from the phenomenological tradition (like *horizon*, *life world*, or *historicity*) into methodological tools for historians, such as *Erfahrungsraum* (space of experience) and *Erwartungshorizont* (space of expectation), *Sattelzeit*, among many others; all of which became cornerstones of his methodology of conceptual history (Schinkel, 2005). However, the practice of conceptual history certainly departed to some extent from Koselleck's presumptions and guidelines, in some cases to the extent that sections or whole articles were rejected by the editors because of their methodological dissents (cf introduction to the *Grundbegriffe*).

From 1972 onward, Koselleck coedited, together with the historians Werner Conze (1910–86) and Otto Brunner (1898–1982), the eight-volume encyclopedia dedicating most of his research and intellectual engagement to this project. In accordance to the imperative that history is to be rewritten, Koselleck instead revised and adopted not only the methodological framework of the *Grundbegriffe* but started to rethink conceptual history from within the more general context of historical semantics and the need for theory in the historical discipline. Consequently, in his later work he returned to a concern that was already present in his early writings (Olsen, 2012): the dialectic between his reflections on historical semantics and an anthropological theory of history that reframes conceptual history (Hoffmann, 2010).

Critique, Transformation, and Resumption of the Encyclopedia of Basic Concepts in History

The project of the *Grundbegriffe* and Koselleck's early work in the history of concepts were both highly acclaimed and the subjects of severe critique. Probably the strongest criticism came from Koselleck's colleague at Bielefeld, the social historian Hans-Ulrich Wehler (born 1931), who, like others, claimed that not only did conceptual history not contribute to the urgent questions of the historical discipline but also that it was an intellectual dead-end and would have no impact. His colleague at Stanford, James Sheehan (born 1937), joined him in the critique of the *Grundbegriffe* (Sheehan, 1978) but mostly of its conception and methodology. For Sheehan, the corpus of sources was insufficiently representative and not diverse enough to plumb the depths of popular culture and everyday life (Lüdtke, 1995).

Besides all efforts to the contrary, the focus and perspective of the *Grundbegriffe* remained on German history. It was also

criticized for its attention to the origins and evolution of a concept rather than situational semantics and diachronic differences, as well as narrow range of reference in canonical texts, authors, and published sources. The attempt to include newspaper articles, journals, pamphlets and records was not particularly successful despite many extensive citations, which should document the language of the sources and give voice to their 'right of veto' (Koselleck). Substantial critique and enhancement came mostly from inside the field (e.g., Kittsteiner, 1991; Steinmetz, 1993; Busse et al., 1994; Lüsebrink and Reichardt, 1997; Palonen, 1997; Reichardt, 1998), yet, even by Koselleck himself, who kept revising the theoretical framework of conceptual history.

However, any critique should repay its subject in the same coin, and at least two major critical responses almost outweighed the *Grundbegriffe* in their ambition and conception: the *Handbuch politisch-sozialer Grundbegriffe in Frankreich: 1680–1820* (Companion to Political-Social Basic Concepts in France), directed by Rolf Reichardt (born 1940) and published in 20 volumes; and the *Aesthetic Basic Concepts. Historical Dictionary in Seven Volumes*, conceptualized under the institutional umbrella of the East German Academy of Sciences, and edited by the literary scholar Karlheinz Barck (1934–2012), *Martin Fontius* (*born 1934*), *Friedrich Wolfzettel* (*born 1941*), and *Burkhart Steinwachs* at the Berlin think-tank *Zentrum für Literaturforschung* (*ZfL*), finally released in 1992–2005. Currently, the ZfL continues research in the field with an emphasis on both interdisciplinary research and conceptual history of the sciences (Müller and Schmieder, 2008).

In contrast to the *Grundbegriffe*, the *Companion* had strong preference for diachronic differences, included a much broader variety of sources, and thus offered a survey of the most significant social and political concepts in France from the beginning of the Enlightenment to the Restoration. Its interest lay largely in the literary, social, and political conception of reality. The long introduction spells out the methodological rigor with which the whole enterprise was undertaken: it focused on the changes in language and meaning before, during, and after the French Revolution by using a methodological mix of the quantitative techniques of lexicometry together with the analysis of historical semantic of concepts including a standardized set of sources ranging from literary texts, newspapers, journals, tracts, and pamphlets to folk songs, popular images, and other products of mass print culture, all in order to pay attention to the social history of the illiterate majority. The handbook of aesthetic concepts, on the other hand, programmatically allowed its authors an expansive methodological and conceptual freedom. It aimed at balancing the history of aesthetic thought through the history of its somewhat 170 major concepts. Yet, it was by no means comparable in its theoretical ambition, and thus did not attempt in establishing an independent method – neither applicable for the project nor for conceptual history at large.

The problem of semantics was also addressed in social theory, mostly as part of a theory of modernity that sustained the assumption of a major transformation around 1800. Koselleck's colleague at Bielefeld, the German sociologist Niklas Luhmann (1927–98), introduced the concept of semantics to his social theory of communication in order to deal with the production of meaning or sensemaking within social systems (Luhmann, 1995). Meaning for him is the selective realization of possibilities within a given social system; semantics in that sense are structures within a concrete system that allow for the preservation of certain premises of the production of meaning. According to Luhmann, semantics evolve in retrospective adaptation to the autopoietic evolution of social structures (Stäheli, 1997).

International Trends in Conceptual History

Today, the practice of conceptual history belongs to a broader understanding of historical semantics as well as to political and intellectual history. It has become an international endeavor absorbing many different national traditions, most prominently those of Finland, the Netherlands, and the Spanish-speaking world, with the respective flagship-projects often modeled after the *Grundbegriffe*. Bibliographies for the growing scholarship are available at http://www.hpscg.org/bibliography (last updated 2008) and http://www.historyofconcepts.org/database (currently updated). Curiously, Russian conceptual history mostly remains the domain of German-speaking scholars. (Thiergen, 2006; Sperling, 2012) Recently, the import of German historical scholarship to China also includes historical semantics. (e.g., Wan-ting, 2012) All these national traditions of conceptual history aspire to international comparison but are mostly busy with the groundwork in their respective national contexts. On a methodological level historians like Willibald Steinmetz (born 1957) advocate a nonnominalistic approach that deals not only with concepts but the historical semantics of different kinds of condensed past experiences (including popular and visual culture as well as unpublished source materials) in a transnational context. Along these lines not only counterconcepts (Junge and Postoutenko, 2011) but also blind spots and eclipses can play a significant role. (Hölscher, 2009). Similar claims have been made for the necessity for a historical semantics that pays attention to performative and discursive practices, as well as for the historical-semantic inquiry into the imaginary (e.g., Koselleck, 1985; Wimmer, 2012) and the visual (Kohle and Reichardt, 2008).

For research in the historical semantics in the premodern era the situation is slightly different. As the concept-oriented approach reaches its limits for research in this period, the analysis of historical semantics is the most profitable route forward. The existing canonical corpora of medieval sources that have been edited since the nineteenth century (and which were digitized early on) have become the resource of quantitative scrutiny in the historical semantics of premodern Europe. This applies mostly to Latin medieval discourse; a similar scrutiny of vernacular languages remains a desideratum (Jussen et al., 2008). Whatever the case, the problem of multilingual communication in premodern times, necessarily suggests different research strategies.

In recent periods, new and easily accessible technologies like Google book search and the experimental Ngram toolbox (Schmidt, 2012) have become available, which allow for large-scale literary computing that triggered a revival of quantitative approaches to historical semantics. Digital

humanists like Franco Moretti (2013) have argued for a new form of macroanalysis (Jockers, 2013) of so-called big data in the humanities, which as the huge collaborative editorial projects of the nineteenth century have a tradition of dealing with large corpora of texts (Bloch, 1994) but have developed few methodological procedures for dealing with them. At this point it seems unclear whether the new technological possibilities will allow for hermeneutic innovation that can replace or revolutionize critical and close reading or will only change the ways we access, select, and connect texts. However, the data deluge will not put an end to the necessity for theory but will ultimately urge us to rethink the data (as well as the algorithms that process and read these digital materials) theoretically and historically.

Besides the *Archiv* mentioned above to date, several journals have appeared concerned with research into historical semantics and the history of concepts, i.e., the US *Journal for the History of Ideas* (since 1940); the French *Cahier de Lexicologie* (since 1959); Italian *Journal of Interdisciplinary History of Ide*as (since 2012); the German *Forum Interdisziplinäre Begriffsgeschichte* (since 2012); the Finnish periodical *Rediscriptions. Yearbook of Political Thought and Conceptual History* (since 1997); the international journal *Contributions to the History of Concepts* (since 2005); and the *Journal of the History of Ideas in East Asia* (since 2011). In particular the latter pair provides platforms for an international community to publish original empirical and theoretical research.

Today conceptual history is mostly understood and practiced within a broader constituency of the study of semantic change, to some extent taking issue with the methodology of discourse analysis. One of the major concerns seems to transgress national traditions and languages in order to provide transnational and transcultural analyses of semantic change (for the international debate see Hampsher-Monk et al., 1998; for Finland see Ahonen, 2006; for Italy Chignola, 2002), including regions that have so far not been subject of study, such as Russia (Sperling, 2012) and former colonies like India (Pernau, 2012). Future approaches might include the study of traveling concepts (Przyrembel, 2011) or even blind spots (Hölscher, 2009), which allow for transcultural comparison without perpetuating reigning models of power and meaning. Another trend will probably be a conceptual history of the last century, as there exists relatively little research on the history of concepts in the twentieth century.

Finally, the research experience in the field may open even more perspectives on social and cultural history and the history of everyday life as well as historical metaphorology and discourse analysis with the common goal of maintaining an as wide horizon as possible for fruitful future research in the field. The aim of conceptual history should, maybe paradoxically, become to do without a *telos*, in order to engage in new ways with contingency, singularity, and the very meaning of the event as to keep the writing of history open for a history that goes beyond the limits of rationality.

See also: Discourse Studies; Hermeneutics; Historical Thought and Historiography: Current Trends; History: Theories and Methods; Intellectual History; Linguistic Turn and Discourse Analysis in History; Linguistic Turn; Luhmann, Niklas (1927–98); Metaphor, Role in Social Thought: History of the Concept; Modernization; Political Thought, History of; Schmitt, Carl (1888–1985); Science, History of; Semantic Differential; Semantics; Semiotics.

Bibliography

Ahonen, P., 2006. The Finnish orientation in the history of political theory and the history of political concepts: an analysis with examples from political economy and public management. Halduskultuur – Administrative Culture 7, 24–48.

Algazi, G., 1997. Otto Brunner – 'Konkrete Ordnung' und Sprache der Zeit. In: Schöttler, P. (Ed.), Geschichtsschreibung als Legitimationswissenschaft, 1918–1945. Suhrkamp, Frankfurt am Main, pp. 166–203.

Andersen, N.A., 2003. Discursive Analytical Strategies: Understanding Foucault, Koselleck, Laclau, Luhmann. Policy Press, Bristol.

Bahner, W., 1963. Zum Charakter des Schlagworts in Sprache und Gesellschaft. Beträge zur romanistischen Philologie 2, 139–149.

Barthes, R., 1981. The discourse of history. Comparative Criticism 3, 7–20.

Bauer, W., 1920. Das Schlagwort als sozialpsychische und geistesgeschichtliche Erscheinung. Historische Zeitschrift 122, 189–240.

Bloch, R.H., 1994. God's plagiarist. Being an Account of the Fabulous Industry and Irregular Commerce of the Abbe Migne. The University of Chicago Press, Chicago, IL.

Blumenberg, H., 2010. Paradigms for a Metaphorology. Cornell University Press, Ithaca, NY.

Bödeker, H.E. (Ed.), 2002. Begriffsgeschichte, Diskursgeschichte, Metapherngeschichte. Wallstein, Göttingen.

Brunner, O., Conze, W., Koselleck, R. (Eds.), 1972. Geschichtliche Grundbegriffe: Historisches Lexikon zur politisch-sozialen Sprache in Deutschland. Klett-Cotta, Stuttgart.

Burleigh, M., 1988. Germany Turns Eastwards: A Study of Ostforschung in the Third Reich. Cambridge University Press, New York.

Busse, D., Hermanns, F., Teubert, W. (Eds.), 1994. Begriffsgeschichte und Diskursgeschichte: Methodenfragen und Forschungsergebnisse der historischen Semantik. Westdeutscher Verlag, Opladen.

Canguilhem, G., 1952. La connaissance de la vie. Vrin, Paris.

Canguilhem, G., 1955. La formation du concept de réflexe aux XVIIe et XVIIIe siècles. Presse universitaire du France, Paris.

Chignola, S., 2002. History of political thought and the history of political concepts: Kosellecks proposal and Italian research. History of Political Thought 23 (3), 517–541.

Habermas, J., 1977. Kultur und Kritik. Suhrkamp, Frankfurt am Main.

Hampsher-Monk, I., Tilmans, K., Vree, F. van (Eds.), 1998. History of Concepts: Comparative Perspectives. Amsterdam University Press, Amsterdam.

Hoffmann, S.-L., 2010. Koselleck, Arendt, and the anthropology of historical experience. History and Theory 49, 212–236.

Hofmannsthal, H.v., 2005. The Lord Chandos Letter: and Other Writings. The New York Review, New York.

Hölscher, L., 2009. Semantik der Leere. Grenzfragen der Geschichtswissenschaft. Wallstein, Göttingen.

Joas, H., Vogt, P., 2011. Begriffene Geschichte. Beiträge Zum Werk Reinhart Kosellecks. Suhrkamp, Berlin.

Jockers, M., 2013. Macroanalysis: Digital Methods and Literary History. University of Illinois Press, Champaign, IL.

Jones, G.S., 1983. Languages of Class: Studies in English Working Class History, 1832–1982. Cambridge University Press, Cambridge.

Junge, K., Postoutenko, K. (Eds.), 2011. Asymmetrical Concepts after Reinhart Koselleck: Historical Semantics and beyond. Transcript, Bielefeld.

Jussen, B., Mehler, A., Ernst, A., 2008. A corpus management system for historical semantics. Sprache und Datenverarbeitung. International Journal for Language Data Processing 7, 81–89.

Kittsteiner, H.-D., 1991. Die Entstehung des moderenen Gewissens. Insel Verlag, Frankfurt am Main and Munich.

Koebner, R., 1911. Die Eheauffassung des ausgehenden deutschen Mittelalters. Doctoral Disseratation. Friedrich-Wilhelms University, Berlin.

Kohle, H., Reichardt, R., 2008. Visualizing the Revolution: Politics and the Pictorial Arts in Late Eighteenth-century France, Picturing History. Reaktion, London.

Koselleck, R., 1972. In: Einleitung, Brunner, O., Conze, R., Koselleck, R. (Eds.), Geschichtliche Grundbegriffe, vol. 1. Klett-Cotta, Stuttgart, pp. XIII–XXVII.

Koselleck, R., 1979. Historische Semantik und Begriffsgeschichte. Klett-Cotta, Stuttgart.
Koselleck, R., 1985. Futures Past: On the Semantics of Historical Time. MIT Press, Cambridge, MA.
Koselleck, R., 1988. Critique and Crisis: Enlightenment and the Pathogenesis of Modern Society. Berg, Oxford.
Koselleck, R., 2003. Zeitschichten. Studien zur Historik. Suhrkamp, Frankfurt am Main.
Koselleck, R., 2006. Begriffsgeschichten. Studien zur Semantik und Pragmatik der politischen und sozialen Sprache. Suhrkamp, Frankfurt am Main.
Koselleck, R., 2010. Vom Sinn und Unsinn der Geschichte: Aufsätze und Vorträge aus vier Jahrzehnten. Suhrkamp, Berlin.
Lehmann, H., Richter, M. (Eds.), 1996. The Meaning of Historical Terms and Concepts. New Studies on 'Begriffsgeschichte'. German Historical Institute Occasional Papers 15, Washington, DC.
Leonhard, J., 2004. Grundbegriffe und Sattelzeiten – Languages and Discourse. In: Habermas, Rebekka (Ed.), Interkultureller Transfer und nationaler Eigensinn: Europäische und anglo-amerikanische Positionen der Kulturwissenschaften. Wallstein, Göttingen, pp. 71–86.
Lovejoy, A.O., 1936. The Great Chain of Being; a Study of the History of an Idea. Harvard University Press, Cambridge, MA.
Luhmann, N., 1995. Gesellschaftsstruktur und Semantik. Studien zur Wissenssoziologie der modernen Gesellschaft, 4 vols. Suhrkamp, Frankfurt.
Lüdtke, A., 1995. History of Everyday Life. Reconstructing Historical Experiences and Ways of Life. Princeton University Press, Princeton, NJ.
Lüsebrink, H.-J., Reichardt, R., 1997. Kulturtransfer im Epochenumbruch: Frankreich-Deutschland 1770 bis 1815. Leipziger Universitätsverlag, Leipzig.
Meinecke, F., 1924. Die Idee der Staatsräson in der neueren Geschichte. Oldenbourg, München.
Moretti, F., 2013. Distant Reading. Verso, London.
Müller, E., Schmieder, F. (Eds.), 2008. Begriffsgeschichte der Naturwissenschaften zur historischen und kulturellen dimension naturwissenschaftlicher Konzepte. W. de Gruyter, Berlin.
Olsen, N., 2012. History in the Plural: An Introduction to the Work of Reinhart Koselleck. Berghahn Books, New York, Oxford.
Palonen, K., 1997. An application of conceptual history to itself. From method to theory in Reinhart Koselleck's begriffsgeschichte. Finnish Yearbook of Political Thought 1, 39–69.
Palonen, K., 2004. Die Entzauberung der Begriffe. Das Umschreiben der politischen Begriffe bei Quentin Skinner und Reinhart Koselleck. LIT, Münster.
Pernau, M., 2012. Whither conceptual history? From national to entangled histories. Contributions to the History of Concepts 7, 1–11.
Przyrembel, A., 2011. Verbote und Geheimnisse: das Tabu und die Genese der europäischen Moderne. Campus Verlag, Frankfurt.
Rancière, J., 1989. The Nights of Labor: The Workers' Dream in Nineteenth-Century France. Temple University Press, Philadelphia.
Reichardt, R., 1998. Das Blut der Freiheit: Französische Revolution und politische Kultur. Fischer, Frankfurt am Main.
Reichardt, R., Herding, K. (Eds.), 1989. Die Bildpublizistik der französischen Revolution. Suhrkamp, Frankfurt am Main.
Reichardt, R., Schmitt, E., 1985. Handbuch politisch-sozialer Grundbegriffe in Frankreich 1680–1820. Oldenbourg, München.
Sawilla, J.M., 2004. "Geschichte:" Ein Produkt der deutschen Aufklärung? Eine Kritik an Reinhart Kosellecks Begriff des "Kollektivsingulars Geschichte". Zeitschrift für Historische Forschung 3, 281–428.
Schinkel, A., 2005. Imagination as a category of history: An essay concerning Koselleck's concepts of Erfahrungsraum and Erwartungshorizont. History and Theory 44, 42–54.
Schmidt, B., 2012. Theory First. Journal of Digital Humanities 1.
Scott, J.W., 1974. The Glassworkers of Carmaux: French Craftsmen and Political Action in a Nineteenth-Century City. Harvard University Press, Cambridge, MA.
Schöttler, P., 1989. Historians and discourse analysis. History Workshop Journal 27 (1), 37–65.
Schöttler, P., 2005. 13, rue du Four. Die "Encyclopédie Française" als Mittlerin französischer Wissenschaft in den 1930er Jahren. In: Nemeth, E., Roudier, N. (Eds.), Paris–Wien. Enzyklopädien im Vergleich. Springer, Vienna, pp. 179–204.
Sheehan, J.J., 1978. "Begriffsgeschichte" theory and practice. Journal of Modern History 50, 312–331.
Sperling, W., 2012. "Schlafende Schöne"? Vom Sinn und Unsinn der Begriffsgeschichte Russlands. Ein Diskussionsbeitrag. Jahrbücher für Geschichte Osteuropas 60 (3), 375–405.
Stäheli, U., 1997. Exorcising the "popular" seriously: Luhmann's concept of semantics. International Review of Sociology 7, 127–145.
Steinmetz, W., 1993. Das Sagbare und das Machbare. Zum Wandel politischer Handlungsspielräume: England 1780–1867. Klett-Cotta, Stuttgart.
Stempel, W.-D., Koselleck, R. (Eds.), 1973. Geschichte: Ereignis und Erzählung. W. Fink, München.
Thiergen, P. (Ed.), 2006. Russische Begriffsgeschichte der Neuzeit. Beiträge zu einem Forschungsdesiderat, Cologne et al. Böhlau. (multilingual).
Wan-ting, Hu, 2012. The emergence and evolution of the word 'nationalism' in modern China (1901–1935). Journal of the History of Ideas in East Asia 2, 219–262.
Wimmer, M., 2012. Archivkörper: Eine Geschichte Historischer Einbildungskraft. Konstanz University Press, Paderborn.

Conditional Cash Transfers: Influence on Marriage and Fertility

Sarah Baird, Milken Institute School of Public Health, George Washington University, Washington, DC, USA
Berk Özler, Development Research Group, The World Bank, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the theoretical foundations and investigates the existing evidence on the impact of conditional cash transfers (CCTs) on the marriage and fertility of the transfer recipient (typically the adult female in the household) as well as the school-age beneficiary. The relatively scant existing literature suggests small to no impact on the marriage and fertility of the transfer recipient, with suggestive evidence of small declines in both outcomes for female school-age beneficiaries. New research on CCTs that directly condition on marriage, as well as additional longer term evidence, will allow for more robust conclusions to be drawn.

Introduction

Conditional cash transfers (CCTs) have become an important policy tool in developing countries and largely aim to improve children's health and schooling outcomes. CCTs are typically targeted at poor households with infants and children and involve regular cash payments to the household conditional on a certain set of behaviors. These behaviors usually involve investments in children's schooling and health such as regular school enrollment or preventive health care visits. A large and empirically well-identified body of evidence has demonstrated the ability of CCTs to improve schooling and health outcomes in developing countries (see Fiszbein and Schady (2009) for an extensive review).

CCTs may also influence other outcomes that are not directly related to the condition including marriage and fertility. These effects may be seen both among the transfer recipients, who are frequently the adult women in the household (i.e., mothers of the target school-aged children or infants or pregnant women), the eligible children, or even among their ineligible siblings. This article first provides the theoretical foundations for why CCTs may affect marriage and fertility, and then discusses the existing evidence on the influence of CCTs on marriage and fertility, as well as a brief discussion of evidence on the impact of other types of pseudoconditional transfers (in-kind transfers, school uniforms, and scholarships) on these outcomes.

Theoretical Foundations for the Impact of CCTs on Marriage and Fertility

From a theoretical perspective, the impact of a CCT on both marriage and fertility is ambiguous. Focusing first on fertility, the increased income to the household makes it more affordable to have an additional child of the same schooling level, thus potentially leading to an increase in fertility (Schultz, 1997, 2004). Moreover, since many CCTs have certain benefits that are provided per child, households can increase benefits by having more children (Steklov et al., 2007). On the other hand, the classic quantity–quality trade-off (Becker, 1960; Becker and Lewis, 1973) suggests that if quantity and quality are substitutes in the parents' utility function then increased quality of children reduces demand for more children (Rosenzweig and Wolpin, 1980). If CCTs increase the quality of children (through increased schooling or improved health), this could lead to a corresponding reduction in fertility (Schultz, 1997). That said, the transition from high quantity/low quality to low quantity/high quality typically occurs over many generations, so it is unclear to what extent short-term influxes of cash will have, or in more established programs, whether enough time has passed since the onset of these programs to realistically expect to see any change (Steklov et al., 2007). There may also be effects on fertility through the supply side component of the CCT, if the CCT includes health-related conditions, through, for example, increased knowledge about contraception or breastfeeding (Steklov et al., 2007).

In terms of marriage of the household head, effects may differ depending on whether you are married or not at the onset of the program. For couples that are married, the increased income to the household may reduce stress and conflict over resources within the household, thus increasing the likelihood of a marriage staying together. On the other hand, the CCT might increase marital dissolution as a result of increased empowerment of the women (women are often the recipient of the transfer) or through increased conflict in the household over how to spend the transfers (Bobonis, 2011). For single individuals' eligible for the transfer, they may become more attractive in the marriage market due to the increased income (Bobonis, 2011). In addition, given the strong focus of these programs on children, CCTs might increase the desire of single individuals to form families, thereby increasing marriage rates (Steklov et al., 2007). Overall, the theoretical underpinnings suggest that the direction of the impact of a CCT on marriage and fertility of the household head is uncertain and will likely depend on a variety of factors related to the country and the program.

For school-age females of childbearing age, CCTs raise the opportunity cost of getting married and/or having a child. In the short run, because marriage and schooling are mutually exclusive in many settings and pregnancy usually leads to dropping out of school (at least missing school for some period), the occurrence of either would lead to a loss of income for the family. It should be noted that while it is the

young women who have to abide by the condition to attend school regularly, it is usually the parents who receive the cash transfers. Under certain circumstances, this may lead to no changes in the opportunity cost of marriage or childbearing for the school-age child and, in fact, may make some of them worse off, if their disutility from schooling is sufficiently high. We abstract from these intrahousehold issues in this article and assume a unitary model of the household. Schultz (2004) argues this point more formally by stating that the cross-price elasticity effects of CCT programs should reduce demand for substitutes for children's schooling. Therefore CCTs should, on average, delay marriage and childbearing, which can lead to reductions in total fertility. In the longer run, CCTs can cause further reductions in childbearing by either changing fertility preferences of young women and/or by changing the opportunity cost of the marginal child (due to increased earning potential through increased schooling). Hence, CCT effects on school-aged female beneficiaries are more straightforward: we should observe delays in teen pregnancies and marriages in the short run as well as reductions in total fertility in the longer run as a result of CCT interventions.

Empirical Evidence on the Impact of CCTs on Marriage and Fertility

The knowledge base on the impact of CCTs on marriage and fertility is fairly limited with only six peer reviewed articles looking at the marriage and fertility of the adult female in the household and six peer reviewed articles on the marriage and fertility of the school-aged children. The search strategy for this article utilized Scopus and Google Scholar and utilized the search terms cash transfer and Marriage or Fertility. References of the initially found manuscripts were also checked, and key researchers in the field were contacted to confirm that no key resources were missing. There are two main reasons for this limited body of evidence. First, the majority of evaluations of CCT programs focus on measuring impacts directly related to the condition(s), and do not collect data on indirect effects. Second, and particularly in the case of school-age children, the expected impacts on marriage and fertility will likely only manifest themselves in the longer term and many of the existing studies of CCTs are not set up to evaluate longer term impacts, or the programs are too new to evaluate these longer term impacts.

Marriage and Fertility of the Adult Female

As mentioned in the previous section, the direction of the impact of CCTs on marriage of the adult female is ambiguous: the additional income in the household may relieve stress and decrease the likelihood of divorce; while on the other hand, the transfer (particularly when transferred directly to the women) may increase her bargaining power within the marriage and increase the likelihood of the marriage dissolving. There are only two papers that we could find in the literature that examines this question. Bobonis (2011) investigates the impact of CCTs on marital transitions using data from the PROGRESA program in Mexico. PROGRESA (now called *Opportunidades*) was established by the Mexican government in 1997, and provides transfers to the adult female in poor households conditional on children's school attendance and preventive health checks. Utilizing data from the initial randomized phase-in of the program from 1997 to 1999, Bobonis (2011) finds that the overall share of women in a marital union did not change as a result of the intervention, but that there was increased turnover (increase of 0.32 percentage points in the intervention communities over a base of 0.47%). Steklov et al. (2007) mentions that they find no impact of Programa de Asignación Familiar (PRAF), a CCT in Honduras, on marriage, but do not show the results. Bobonis (2011) also finds increased formation of unions among program-eligible mothers who were not in a relationship at baseline. Overall, Bobonis (2011) concludes that in the short-term marriage transitions, as opposed to marriage rates, are affected by PROGRESA. However, ultimately the impacts are fairly small, and the author is restricted by the short time frame of the analysis. Perova and Reynolds (unpublished manuscript), examining Bolsa Família in Brazil, also finds no impact of the program on marriage rates; however, similar to Bobonis (2011), they find that the number of separations have risen as Bolsa Família has expanded.

As with marriage, the impact of CCTs on fertility of the adult female head is also ambiguous. CCTs may increase fertility due to increased resources for the household or if higher benefits are provided to larger families. On the other hand, if CCTs lead to increased bargaining power of the women or an emphasis on quality of children, over quantity, fertility may decrease. Once again evaluating PROGRESA, Schultz (2004) finds no differences in the probability of birth in the 6 months prior to November 2009 (about 2 years since the start of the program). Feldman et al. (2009) using a hazard model finds no impact of PROGRESA/Opportunidades on birth spacing (both after 2 years and after 6 years), although they do find short-term impacts on contraceptive use. One possible reason for this lack of effect is the phase-in experimental design that restricted treatment/control comparisons to a relatively short window – as the control groups received treatment with a delay of approximately 2 years.

Stecklov et al. (2007) also look at the impact of PROGRESA on fertility alongside analysis of two other programs in Latin America: PRAF in Honduras and Red de Protección (RPS) in Nicaragua. The authors utilize a measure of fertility that takes on a value of one if the adult woman in the beneficiary household gave birth in the 12 months preceding the survey. Exploiting the randomized implementation of all three of these programs, and utilizing difference in difference estimates, the authors find no short-term impacts (approximately 2 years after the start of each program) on fertility of PROGRESA or RPS, but do find that PRAF leads to an increase in fertility of 1.7 percentage points (or 45% of the decline of 3.8 percentage points in the control group between the pretreatment and posttreatment period), statistically significant only at the 10% level. The authors utilize an alternative specification that defines fertility (before and after the treatment) as having given birth during the past 12 months or being pregnant. The effects for PROGRESA and RPS are again zero, while the effect for PRAF is approximately 4 percentage points, significant at the 1% level. The authors attribute these differential effects to key

elements in the design of the programs related to eligibility and particular whether or not the roster was 'open.' They define a roster as open, if households can become eligible for a component of the program by changing household composition after the onset of the program through, for example, having an additional child. Both PROGRESA and RPS were for the most part closed programs, but the PRAF roster could change in initially targeted communities, thus potentially incentivizing childbearing. All three programs do, however, have limits on the total amount that can be transferred to a given household.

Todd et al. (2011) also look at the impact of RPS on fertility 4 years after the start of the program and find no impact of the program on whether or not the household had given birth in the past 30 months. Note that the control group started receiving treatment 3 years after the start of the program, so the impact estimates are measured off of the differential exposure of 3 years. Utilizing a hazard model, however, the authors do find that the program reduced the odds of a birth occurring by 32% and that the impact is greater for women with more children. The authors suggest that this finding is consistent with an increase in birth spacing and with CCT programs having subtle impacts on household fertility decisions.

Overall, the evidence suggests that the impact of CCTs on fertility and marriage of the household head is negligible; however, one should be careful in using this as conclusive evidence that CCTs have no impact on these outcomes due to a number of important limitations. First, the body of evidence consists of six papers covering three different programs. Second, the analysis is largely measuring short-term impacts, and one might expect impacts on fertility and marriage to only play out in the longer term. Further research from additional CCT programs as well as longer term analysis is needed to fully understand the impact of CCTs on marriage and fertility among adults in beneficiary households.

Marriage and Fertility of the School-Age Children

Through its effects on health and schooling, CCTs also have the potential to impact the marriage and fertility outcomes of school-age children in beneficiary households. Exploiting the randomized design, and using a probit estimation strategy, Schultz (2004) looks at the short-term impacts of PROGRESA on the likelihood of having given birth in the past 6 months. Focusing on 14- to 18-year-old girls prepared to enroll in secondary school, the author finds no significant impact of the program on fertility. The author does finds that increased school enrollment is associated with decreased fertility among this group, but the finding is based on enrollment being instrumented by program eligibility. It may not be entirely appropriate to assume that a CCT program affects the likelihood of births only through its effect on school enrollment for school-age females. Darney et al. (unpublished manuscript) use multiple rounds of a population-based survey and matching techniques to examine the effect of exposure to Opportunidades on pregnancy among 15- to 19-year-old females. The authors do not find any evidence that the program reduced the odds of becoming pregnant. However, the authors are focusing on the impacts of Opportunidades above and beyond the effect of the program on education. If the entire effect of Opportunidades on fertility was via improvements in schooling, it would not be captured in their estimation strategy. Using a difference in difference estimation strategy Behrman et al. (2009) look at the medium-term impacts of PROGRESA/Opportunidades on marriage. Using data from 2003, and comparing those who resided in a locality that became eligible for treatment in 1998 versus those in localities that became eligible in 2000, they find no overall impact of the program on the marriage of boys or girls who were 9–15 when the program started. The authors do, however, find suggestive evidence that youth with early exposure to the program are delaying marriage. (Gardner and Reynolds (unpublished manuscript) examine Bolsa Família and find no reductions in fertility among teens.

Using a randomized evaluation, and looking at a baseline sample of 13- to 22-year-old, never-married schoolgirls in Malawi, Baird et al. (2011) find no impact of CCTs on marriage and fertility. Despite the fact that school enrollment and marriage are practically mutually exclusive in this setting, the improvements in schooling among the CCT group were too small to lead to significant changes in marriage and pregnancy at the end of the two-year program. Moreover, girls who dropped out of school, and thus were not receiving the transfer, were just as likely to be married as the control group. Interestingly, the authors also find large and significant declines in marriage and pregnancy in the experimental study arm receiving unconditional cash transfers (UCTs). While there was little schooling improvement in this group compared to the CCT arm, the income effect among those who dropped out of school during the program seems to be responsible for the decline in marriage and pregnancy. Using data from the same experiment, but focusing on girls who were not enrolled in school at the start of the intervention, Baird et al. (2009) find large short-term impacts of the intervention on marriage (a reduction of more than 40%) and on fertility (a reduction of more than 30%). The large effects in this group are likely due to the sizable schooling effect among this group – baseline dropouts in the intervention group were 3–4 times more likely to be in school than their counterparts in the control group.

IEG (2011) evaluates the Punjab Female School Stipend Program in Pakistan, which started in 2003, and provided quarterly cash transfers of around $10 to families of girls enrolled in middle school with the condition that they attend at least 80% of classes. The program also includes a supply component and was extended to high-school girls in 2006. Utilizing quasiexperimental methods and looking at the impact of the program after 4 years, the authors find no impact on the probability of girls of age 15–19 years being legally married, but do find that after the program started girls in stipend areas were likely to marry about 1.2 years later. In terms of fertility, the authors focus on 17- to 19-year-old girls and find no impact of the program on the probability that the girl has given birth, but do find that girls exposed to the program have, on average, 0.3 fewer children. IEG (2011) also finds relatively large impacts on enrollment (between 14 and 28 percentage points), which may partially explain these impacts.

Overall, the empirical results suggest that CCTs do have the potential to reduce the marriage and fertility of the school-age participants; however, impacts of the program on schooling must be large enough to drive significant changes in these outcomes. Furthermore, the possibility exists that UCTs may also be effective in delaying marriage and pregnancy among adolescent girls and young women (Baird et al., 2011). Again, as with the effects of CCT programs on the marriage and fertility of adults in beneficiary households (or parents of eligible children), further research is needed on longer term effects of these programs to fully understand the impact.

While this section has focused on the impacts of traditional health and schooling CCTs on marriage and fertility, there are two ongoing evaluations of programs that offer cash to families conditional on girls not getting married until the age of 18 years. The International Center for Research on Women (ICRW) is in the process of evaluating the Apni Beti Apta Dhan (ABAD) program in India. Starting in 1994, the ABAD program dedicated bonds to newly born girls in the Indian state of Haryana that can be cashed out when the girl is in the age of 18 years conditional on the girl being unmarried (ICRW, 2013). The first girls turned 18 years in 2012, thus presenting the first opportunity to assess whether the program delayed marriage. Results are expected in 2015. Field and Glennerster (2013) are in the process of evaluating the Kinshoree Kontha project in Bangladesh. One of the intervention arms involves a direct incentive to delay marriage until the legal age of 18 years. Girls aged 15–17 years in the intervention arm are eligible to receive 16 l of cooking oil per year on the condition that they remain unmarried, an amount large enough to cover the financial cost associated with delaying marriage. Results from this study are also pending.

Impact of Other Forms of Conditional Transfers on Marriage and Fertility

The focus of this article is on the impact of CCTs on marriage and fertility; however, there are other conditional transfer programs that use instruments other than cash that are worth briefly discussing. Das et al. (2005) define any program requiring a specified course of action to receive a benefit (cash or in-kind) as a CCT. Their definition is broad enough to include workfare programs (cash conditional on working in a program for a given wage rate) and consumption transfers (in-kind transfers). They argue that these are conceptually identical to CCTs, with in-kind transfers being interpreted as cash conditional on the purchase and consumption of the relevant quantity. We discuss findings from three such studies below.

Duflo et al. (unpublished manuscript) evaluate the impact of the Education Subsidy program in Kenya, a program that provided two free school uniforms to upper primary students over the last 3 years of primary school, thus subsidizing the cost of education. The authors find that the likelihood of having a child declined by 2.7 percentage points (17%) after 3 years and 4.4 percentage points (13%) after 5 years among girls in program schools. The authors also find significant declines in marriage (of 20 and 11%, respectively). After 7 years, the impacts on fertility had disappeared, but there was still a 10% decline in marriage rates (significant at the 10% level).

Angrist et al. (2002) investigate the impact of a school voucher system in Colombia, where recipients are determined via a lottery, which partially covered the cost of private secondary school for students who maintained satisfactory grades. Three years after the lottery, winners were significantly less likely to be married or cohabiting, although since these individuals were only 15 marriage rates were low overall (at 1.6% in the control group). There were no impacts on fertility.

There are also noncash incentive programs that directly target fertility. For example, Ashraf et al. (unpublished manuscript) examine an intervention in Zambia, where married women received vouchers for easy access to modern contraceptives – either alone or with their husbands – compared to a control group who received no intervention. They find that while women who received vouchers alone increased their use of concealable forms of contraception by 38% and reduced unwanted births by 57%, there was no effect on unwanted births among women who received the 'couples' treatment. They argue that increasing the supply of contraceptives (or making access to contraceptives easier) may not have any impact on fertility unless accompanied by changes in *de facto* spousal consent requirements.

Conclusions

This article has investigated the existing evidence on the impact of CCTs on the marriage and fertility of the transfer recipient, typically the adult female in the household, as well as of the school-age beneficiary. For the adult female, from a theoretical standpoint, the direction of the impact on both marriage and fertility is ambiguous. For school-age beneficiaries, theory would suggest that both marriage and fertility would decrease as a result of the program, due to an increase in their opportunity costs, as well as through the effect of increased schooling in the longer run.

The extant empirical evidence suggests at most a limited impact of CCTs on the marriage and fertility of the adult female in the household. Evidence from one program, PRAF in Honduras, suggests that fertility may have increased by approximately 2 to 4 percentage points in program areas largely due to a program design that incentivized childbearing. Outside of this one case, there is no evidence that CCTs change the total number of children, but there may be slight effects on the timing of childbearing and birth spacing. In terms of marriage, there is no evidence that overall marriage rates change as a result of CCTs, but evidence from both Mexico and Brazil suggest that CCTs may increase marriage turnover.

Turning to the school-age female beneficiaries, the overall empirical evidence suggests that CCTs do lead to small declines in fertility and marriage. Short-term evidence from PROGRESA largely shows no impact of the program on fertility, but longer term estimates suggest that individuals with early exposure to the program are delaying marriage. Evidence from Malawi is mixed, with large impacts on marriage and fertility for young women who were out of school at baseline and no impacts for girls who were in school at baseline. The authors attribute this result to the large impact of the program on education for the former group. In Pakistan the authors also find impacts of the

program on marriage and fertility, once again likely due to the large impact of the program on schooling. There is also evidence that other types of transfer programs, such as education subsidies, may also influence marriage and fertility.

Given the general scarcity of evidence on this topic, the readers should be cautious in interpreting our summary. In particular, it is important to keep two things in mind: First, the existing evidence comes from the evaluation of a small number of studies compared to the full scope of CCT programs that exist across the world. One of the reasons for this, as mentioned earlier, is that the majority of evaluations of CCT programs focus on measuring impacts directly related to the condition(s), and do not collect data on indirect effects, which is where marriage and fertility fit in. Second, the expected impacts on marriage and fertility will likely only manifest themselves in the longer run and many of the existing studies of CCTs are not set up to evaluate longer term impacts, or the programs are too recent to evaluate these longer term impacts. Given that CCTs are one of the most rigorously evaluated social protection programs, the existing body of evidence is likely to grow substantially in coming years to allow for a more robust understanding of the impact of CCTs on marriage and fertility.

See also: Family Theory: Economics of Childbearing; Family Theory: Economics of Marriage and Divorce.

Bibliography

Angrist, J., Bettinger, E., Bloom, E., King, E., Kremer, M., 2002. Vouchers for private schooling in Colombia: evidence from a randomized natural experiment. The American Economic Review 92 (5), 1535–1558.

Ashraf, N., Field, E., Lee, J., Household Bargaining and Excess Fertility: An Experimental Study in Zambia, unpublished manuscript. http://ipl.econ.duke.edu/bread/papers/0910conf/Field.pdf.

Baird, S., Chirwa, E., McIntosh, C., Özler, B., 2009. The short-term impacts of a schooling conditional cash transfer program on the sexual behavior of young women. Health Economics 19 (Suppl. 1), S55–S68.

Baird, S., McIntosh, C., Özler, B., 2011. Cash or condition? Evidence from a cash transfer experiment. Quarterly Journal of Economics 126 (4), 1709–1753.

Becker, G.S., 1960. An economic analysis of fertility. NBER Chapters. In: Demographic and Economic Change in Developed Countries. National Bureau of Economic Research, Inc., pp. 209–240.

Becker, G.S., Lewis, H.G., 1973. On the interaction between quantity and quality of children. Journal of Political Economy 82 (2), S279–S288.

Behrman, J.R., Parker, S.W., Todd, P.E., 2009. Medium-term impacts of the Oportunidades conditional cash transfer program on rural youth in Mexico. In: Klasen, Stephan, Nowak-Lehmman, Felicitas (Eds.), Poverty, Inequality and Policy in Latin America. MIT Press, Cambridge, MA.

Bobonis, G.J., 2011. The impact of conditional cash transfers on marriage and divorce. Economic Development and Cultural Change 59 (2), 281–312.

Darney, B., Weaver, M., Sosa-Rubi, S., Walker, D., Servan-Mori, E., Prager, S., Gakidou, E., Oportunidades conditional cash transfer program: impact on adolescent pregnancy in rural Mexico, unpublished manuscript.

Das, J., Do, Q.T., Özler, B., 2005. Reassessing conditional cash transfer programs. World Bank Research Observer 20, 57–80.

Duflo, E., Dupas, P., Kremer, M., Education, HIV, and early fertility: experimental evidence from Kenya, unpublished manuscript.

Feldman, B.S., Zaslavsky, A.M., Ezzati, M., Peterson, K.E., Mitchell, M., 2009. Contraceptive use, birth spacing, and autonomy: an analysis of the Oportunidades program in rural Mexico. Studies in family planning 40 (1), 51–62.

Field, E., Glennerster, R., 2013. Empowering Girls in Rural Bangladesh. http://www.povertyactionlab.org/evaluation/empowering-girls-rural-bangladesh (accessed 06.05.13.).

Fiszbein, A., Schady, N., 2009. In: Ferreira, Francisco H.G., Grosh, Margaret, Kelleher, Nial, Olinto, Pedro, Skoufias, Emmanuel (Eds.), Conditional Cash Transfers: Reducing Present and Future Poverty. World Bank, Washington, DC.

Gardner, R., Reynolds, S., Brazil's Bolsa Família: can a conditonal cash transfer reduce teen fertility? unpublished manuscript. http://paa2012.princeton.edu/papers/121121

IEG (Independent Evaluation Group), 2011. Do Conditional Cash Transfers Lead to Medium-term Impacts?: Evidence from a Female School Stipend Program in Pakistan. World Bank, Washington, DC.

ICRW (International Center for Research on Women), 2013. Evaluating the Power of Conditional Cash Transfers (CCTs) to Delay Marriage in India. http://www.icrw.org/where-we-work/evaluating-power-conditional-cash-transfers-ccts-delay-marriage-india (accessed 14.04.13.).

Perova, E., Reynolds, S., Do economic gains for women reduce domestic violence in Brazil? Female homicide and Bolsa Família, unpublished manuscript.

Rosensweig, M.R., Wolpin, K.I., 1980. Testing the quantity-quality fertility model: the use of twins as a natural experiment. Econometrica 48 (1), 227–240.

Schultz, T.P., 1997. Demand for children in low income countries. In: Rosenzweig, M.R., Stark, O. (Eds.), Handbook of Population and Family Economics, Handbook of Population and Family Economics, first ed., vol. 1. Elsevier, pp. 349–430 (chapter 8).

Schultz, T.P., 2004. School subsidies for the poor: evaluating the Mexican PROGRESA poverty program. Journal of Development Economics 74 (1), 199–250.

Schultz, T.P., 2007. Population policies, fertility, women's human capital, and child quality. In: Paul Schultz, T., Strauss, John A. (Eds.), Handbook of Development Economics, vol. 4. Elsevier, pp. 3249–3303 (Chapter 52).

Stecklov, G., Winters, P., Todd, J., Regalia, F., 2007. Unintended effects of poverty programmes on childbearing in less developed countries: experimental evidence from Latin America. Population Studies 61 (2), 125–140.

Todd, J.E., Winters, P., Stecklov, G., 2011. Evaluating the impact of conditional cash transfer programs on fertility: the case of the Red de Protección Social in Nicaragua. Journal of Population Economics 25 (1), 267–290.

Todd, P., Wolpin, T., 2006. Assessing the impact of a school subsidy program in Mexico: using a social Experiment to validate a dynamic behavioral model of child schooling and fertility. American Economic Review 96 (5), 1384–1417.

Relevant Websites

http://go.worldbank.org/BWUC1CMXM0.
http://www.icrw.org/.
www.popcouncil.org.
http://www.poppov.org/.
http://www.povertyactionlab.org/.

Confabulation

Ana Bajo, Institute of Psychiatry, London, UK
Aikaterini (Katerina) Fotopoulou, University College London, London, UK
Simon Fleminger, Glenside Hospital, Salisbury, UK
Michael D Kopelman, King's College London, London, UK; and St Thomas' Hospital, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Confabulation, a form of false memory, has been puzzling clinicians and academics since the 1880s. To date, limited consensus has been reached regarding the definition, classification, differential diagnosis, and cognitive mechanisms underlying this disorder. Three main explanatory frameworks have been proposed, each putting forward and testing different predictions regarding the psychopathology and clinical course of confabulation. The present article reviews theoretical and empirical developments in the topic and considers their implications for the definition and classification of confabulation.

Description of Confabulation

Definition and Features

Confabulations are false memories, either because they are entirely or partially incorrect in themselves, or because they consist of real memories jumbled up and retrieved out of context (Kopelman, 2010). The content of confabulated memories is usually autobiographical and unintentional; the patient is usually unaware of the condition (Burgess and Shallice, 1996). Confabulations often compel the person to act upon them (Schnider, 2001). Confabulation is a debilitating condition: the person may self-discharge from hospital against medical advice, and social support networks are often further strained as relatives suffer unjustified or even paranoid accusations and struggle to appreciate their unintentional nature. Confabulating patients lack insight into their false memories, but may have limited awareness of other memory difficulties.

Classification

Several classifications have been suggested (Berlyne, 1972; Feinberg and Roane, 1997; Gilboa and Moscovitch, 2002). Early classifications distinguished between 'momentary' confabulations (plausible confabulations based on reality) and 'fantastic' confabulations (without basis in reality, often with a grandiose theme) (Berlyne, 1972). A widely used classification distinguishes between 'provoked' versus 'spontaneous' confabulations (Kopelman, 1999). Provoked confabulations arise in response to a challenge to memory (e.g., prompted by orientation or memory tests), whereas spontaneous confabulations arise in the absence of any external trigger and the person often acts upon them (e.g., a hospital in-patient who searched anxiously for his uniform thinking erroneously he was on army maneuvers).

More recently, Schnider distinguished behaviorally spontaneous confabulations from three other types of confabulation (Schnider, 2008):

- intrusions – described as 'simple provoked confabulations';
- momentary – plausible, semantic, or autobiographical false verbal statements which may arise spontaneously during conversation or be directly provoked by questions;
- fantastic – confabulations with no basis in reality; sometimes difficult to distinguish from elaborate momentary confabulations.

According to Schnider, behaviorally spontaneous confabulations appear without external trigger and in this way they differ both from false recognition during memory tests and from provoked confabulations, both of which arise in response to a cue. In addition, while provoked confabulations are associated with a variety of lesions and can happen even in healthy people, spontaneous confabulations are only associated with lesions in the orbitofrontal cortex (OFC) or anterior limbic areas connected with the OFC (Schnider, 2003). Behaviorally spontaneous confabulations are also different from fantastic and from momentary confabulations, in that neither momentary nor fantastic confabulations direct behavior. However, in a later article Schnider concluded that behaviorally spontaneous confabulations do share some features with both momentary and fantastic confabulations (Nahum et al., 2012). Other authors have argued that there is no clear distinction between spontaneous and provoked confabulations: moreover, these often co-occur in the same patient (Gilboa et al., 2006; DeLuca, 2009).

Features of Behaviorally Spontaneous Confabulation

According to Schnider, the main characteristic of behaviorally spontaneous confabulations is that people believe strongly in their confabulations and feel compelled to act on them. "For the patients, confabulations are the honest narrative of their perceived reality, rather than invented stories" (Schnider, 2003). In most cases, the person is simply re-enacting an old habit (Schnider, 2008; compare Dalla Barba and Boisse, 2010).

Confabulations can be traced back to true events (Schnider, 2003). Other themes, such as fantastic and semantic (or general knowledge), have been reported in patients with this syndrome, and reduplicative misidentification (such as thinking that one is in two places at once) is often present, but these are not consistent features of spontaneous confabulations (Schnider, 2008; Kapur and Coughlan, 1980).

The syndrome is also characterized by amnesia and disorientation, although some confabulators have demonstrated normal memory function. Orientation to person (i.e., one's

own identity) often resumes once the patient is out of the confusional state (Schnider, 2008).

Differentiation and Etiology

The precise specification of confabulation has led to intense controversy. Some authors distinguish between false recall and confabulation (Schacter et al., 1998; Gilboa and Moscovitch, 2002). False recall usually refers to nonpersonal information studied postinsult, whereas confabulations usually refer to preinjury autobiographical knowledge; although confabulations about semantic material or general knowledge have also been observed (Dalla Barba et al., 1999; Kopelman et al., 1997).

The differentiation between confabulations and delusions remains unresolved. Some authors think confabulations are best understood as disorders of memory, whereas delusions are viewed as disorders of thought (DeLuca, 2000, 2009; Kopelman, 2010). Other authors suggest they share underlying cognitive deficits (Johnson and Raye, 2000; Turner and Coltheart, 2010; Fotopoulou, 2010). Spontaneous confabulations are often held with strong belief, not amenable to change, and often of bizarre content; on this basis, it has been proposed recently that they are best viewed as delusional beliefs, and therefore only provoked confabulations qualify as genuine confabulations (Coltheart and Turner, 2009: p. 181; Turner and Coltheart, 2010). Conversely, some authors have proposed that only spontaneous confabulation constitutes a specific neurological syndrome observed in clinical populations; they have argued it differs from provoked memory lapses, which can occur in the general population (Kopelman, 1999; Schnider, 2008).

The etiology of confabulation is varied; the syndrome has been observed following anterior communicating artery aneurysm (ACoA), tumors, encephalitis, traumatic head injury, and Korsakoff syndrome (Schnider, 2008). In addition, confabulations have been described in the context of schizophrenia (Nathaniel-James and Frith, 1996), degenerative disease (Dalla Barba et al., 1999), and confusional states (Kopelman et al., 1997; Box et al., 1999). However, spontaneous confabulatory errors are often distinguished from lapses arising from confusional states (DeLuca, 2000; Schnider, 2008).

Brain Pathology

Most authors view the ventromedial prefrontal and OFC as the critical areas where pathology is associated with spontaneous confabulation (Gilboa and Moscovitch, 2002). Schnider (2008) has argued that the critical area is specifically the OFC or areas directly linked to it. Medial ventrocaudal (Broadman Area 25 (BA25)) and basal forebrain have also recently been associated with 'confabulatory-type errors' (Turner et al., 2008b). The medial rostral prefrontal cortex (BA 9/10) has been found to be critical for remembering self-generated material (Turner et al., 2008b). Evidence derived from positron emission tomography (PET) scan investigations pointed to BA13 and BA10 as crucial areas for monitoring memory outcomes, and an impairment of monitoring has been postulated as an essential mechanism underlying confabulation (Schnider, 2008; Schnider et al., 2005; compare Moscovitch and Winocur, 2002).

Early Explanations and Theoretical Models of Confabulation

Many models have attempted to explain the underlying psychological deficits that are critical in the appearance of confabulations. At present at least three types of hypotheses are being followed: (1) 'temporality' hypothesis; (2) 'strategic retrieval' hypothesis, and (3) 'motivational' accounts (Kopelman, 2010).

Current models of cognitive mechanisms underlying confabulations drew ideas from earlier models or explanations.

Temporality Accounts

Early models proposed that confabulations were a function of a basic amnesic deficit coupled with a distorted sense of time. This explained why confabulators often think past events are occurring in the present. For example, Korsakoff (1889, cited in Kopelman, 1999; Schnider, 2008) observed that confabulations were often based on the person's past events and prompted the patient to act upon them. Korsakoff's (Schnider, 2008; Kopelman, 2010) model was based purely on clinical observation, but the distorted temporality hypothesis was later developed by several authors, who have provided more detailed explanations based on well-controlled experiments (Schnider, 2008).

Strategic Accounts

Various early authors noted that in order for confabulations to appear there must be a monitoring deficit that allows incorrect memories to be outputted as correct (DeLuca, 2009; Schnider, 2008). These ideas were later developed into more detailed explanations (Burgess and Shallice, 1996; Gilboa and Moscovitch, 2002). These theoretical explanations were primarily concerned with deficits in memory specification, formation, and/or verification processes that give rise to confabulations.

Motivational Accounts

Early motivational theories of confabulation can be divided into at least three distinct types. The first type of early motivational account, the 'gap-filling' hypothesis, is perhaps the oldest 'compensation' theory of confabulation. According to it, confabulation occurs as a purposive act contrived by the patient to spare him from the embarrassment of not being able to remember the events of his life (Bonhoeffer, 1901; cited in Talland, 1961). This view has been criticized as patients seem genuinely unaware of their memory errors (anosognosic) and thus it seems clinically invalid to describe them as consciously attempting to conceal what they seem to ignore, i.e., their memory deficit. Also, many confabulating patients did not show a tendency to increase their confabulation rate when they were confronted with questions to which they did not know the answer (e.g., Dalla Barba et al., 1997; Kopelman et al., 1997; Mercer et al., 1977; Schnider et al., 1996).

Pick (1905, 1915 in Talland, 1961) and Korner (1935 in Talland, 1961) noticed that the statements of confabulating patients can be easily steered by leading questions. Thus, they

argued that 'suggestibility' has a causal role in the gap-filling forms of confabulation. Moreover, Johnson and colleagues (1997) reported a patient who among other deficits showed 'a propensity toward detailed imaginations' (Johnson et al., 1997: p. 203). In this sense, suggestibility may be seen as part of one's personality structure, or more generally an individual predisposition to external suggestion and compliance. This approach is similar to the second set of motivational theories, reviewed below.

According to the second kind of theories, certain personality traits may distinguish between patients who are likely to develop confabulation and those who do not. Berlyne (1972) for example, observed that five out of the seven Korsakoff patients who he studied were premorbidly extraverts. In a study on confabulation in the context of senile dementia, Gainotti (1975) conducted what remains until today the most systematic study of this hypothesis (see also Conway and Tacchi, 1996). He asked relatives to describe the premorbid personality of 8 confabulating patients and 22 nonconfabulating control patients. The results indicated opposite personality traits between the two groups.

Some psychoanalytically oriented clinicians recognized in the content of confabulation the mark of psychodynamic mechanisms (for reviews see Berlyne, 1972; Betlheim and Hartmann, 1951; Davidson, 1948). For example, some authors suggested that the individual's wishes and interests guided confabulation in the same way as they controlled dream fantasy (Berlyne, 1972; Betlheim and Hartmann, 1951; Van der Horst, 1932). In addition, Lidz (1942 in Talland, 1961) had discussed confabulation as a response to overwhelming anxiety and Zangwill (1953 in Berlyne, 1972) had conceptualized confabulation as a defense against 'a catastrophic reaction,' i.e., a depressive reaction to one's illness.

The hypothesis that confabulation is a form of 'psychological denial,' i.e., a compensatory coping mechanism instigated by excessive anxiety, was mostly advanced and crystallized in Weinstein and colleagues' (for a review see Weinstein, 1996) formulation of the 'defense' hypothesis. Weinstein and colleagues claimed that although confabulations frequently involve references to past events, they are to some degree, symbolic representations, dramatizations, or explanations of some current personal experience, preoccupation, or disability (Weinstein et al., 1956). These symbolic functions are based on kernels of awareness of deficit, i.e., confabulations implicitly express illness preoccupations and anxieties, which the patient is not capable of fully appreciating and explicitly expressing. Finally, the symbolic function of confabulation seems to have an ultimate adaptive role. Weinstein observed, following Bonhoeffer (1901, cited in Weinstein, 1996) that during the narration of confabulations, patients become engrossed in detail, lose their previous irritability, and then appear utterly relaxed.

Although these views have highlighted the positive aspects of confabulation, there has been little experimental evidence in their support. Moreover, these coping mechanisms are considered independent of the specific neuropathological features of the syndromes they aim to explain. Thus, their explanatory power with regards to the syndromes themselves is limited.

Nevertheless, it is important to emphasize that not all psychodynamic authors considered emotional mechanisms in isolation. Instead, some regarded the emotional manifestations of confabulation as the direct consequence of dysfunction of the normal mechanisms of remembering (Betlheim and Hartmann, 1951).

In the last two decades, as the study of emotions was included among the mainstream interests of cognitive neuropsychologists and neuroscientists (see Fotopoulou, 2010), several neomotivational accounts emerged (Conway and Tacchi, 1996; Fotopoulou et al., 2004). These are complementary to the temporality and strategic retrieval models outlined later: they do not make claims about the cognitive deficits that lead to the emergence of confabulations per se; instead they attempt to explain the emotional factors that are dysregulated by such cognitive deficits and hence influence the content and form of confabulations.

Executive Dysfunction

Several authors have suggested that executive dysfunction could explain confabulation independently of amnesia (Kopelman, 1987; Kapur and Coughlan, 1980; Benson et al., 1996). For example, in one study two patients underwent a neuropsychological assessment before and after they recovered from confabulation. At retest both patients showed improvement on executive tests in parallel with decreased confabulations, but their amnesia scores remained unchanged at the severe level of function. The authors concluded that frontal dysfunction, and not amnesia, was the critical deficit associated with confabulations (Kapur and Coughlan, 1980; Benson et al., 1996; but see DeLuca, 2009). However, behavioral experiments have indicated that executive dysfunction did not distinguish between spontaneous confabulators and other patients with amnesia, nor did it parallel recovery from confabulation (Schnider, 2003). In addition, confabulation has been described in the apparent absence of executive dysfunction (Burgess and Shallice, 1996; Johnson et al., 1997; Dalla Barba et al., 1999).

Current Main Theories of Confabulation: Temporality Models

Korsakoff, Tiling, Van der Horst, and Talland had noticed disturbed temporality or chronology as a salient feature of confabulation (Schnider, 2008; Kopelman, 2010; Talland, 1965). Two models have derived from early temporality explanations: a 'disturbed temporal consciousness' model and a 'temporal context confusion' (TCC) model.

Disturbed Temporal Consciousness

Disturbed temporal consciousness is one of the current explanations for the appearance of confabulations (Dalla Barba et al., 1999). This model proposes the following explanation for confabulations. Confabulators may be aware of a personal past, present, and future which is distorted. Temporal consciousness is strongest for well-established knowledge. Consequently, when memory traces are weak, temporal

consciousness is framed by more stable memories: semantic knowledge. Therefore, when prompted, confabulators are likely to answer with memories or plans corresponding to well-rehearsed memories or habits (Dalla Barba, 2009; Dalla Barba and Boisse, 2010).

Dalla Barba's model is unique in that it is the only explanation that proposes a conscious mechanism of confabulation. He has criticized other models for putting forward explanations based on unconscious mechanisms which are not scientifically verifiable (Dalla Barba, 2009).

At present, Dalla Barba's Confabulation Interview is the only established protocol for exploring confabulations across episodic, semantic, personal, and nonpersonal domains (Dalla Barba, 1993). Moreover, an updated version of this interview also explores confabulations regarding future plans (Dalla Barba et al., 1997).

Temporal Context Confusion

The distorted temporality characteristic of confabulators is the focus of one of the most influential models to date: the TCC model (Ptak and Schnider, 1999). Schnider believed confabulators often had a distorted sense of time independent of memory deficits (Schnider, 2008: p. 230). His model refers specifically to 'behaviorally spontaneous confabulations' (Schnider, 2008). These are false memories that occur in the absence of an external trigger and that direct behavior.

Clinical Evidence for TCC

Schnider's model was developed from a series of elegant experiments employing a TCC task. The TCC task (Schnider et al., 2000a) consists of two runs of a continuous recognition task (run 1 and run 2). The first run measures straightforward storage and recognition. In the second run, however, items that in the previous run were targets become distracters and vice versa. As a result, the targets in the second run suffer intense interference from former targets, and consequently this part of the task cannot be completed relying on familiarity alone. An example of the stimuli used in this task can be seen in Figure 1 (Schnider, 2001).

In a series of investigations, behaviorally spontaneous confabulators performed as well as amnesic controls in the first run of the TCC task, whereas they consistently failed the second part. Schnider and Ptak (1999) initially concluded from these results that spontaneous confabulators were unable to

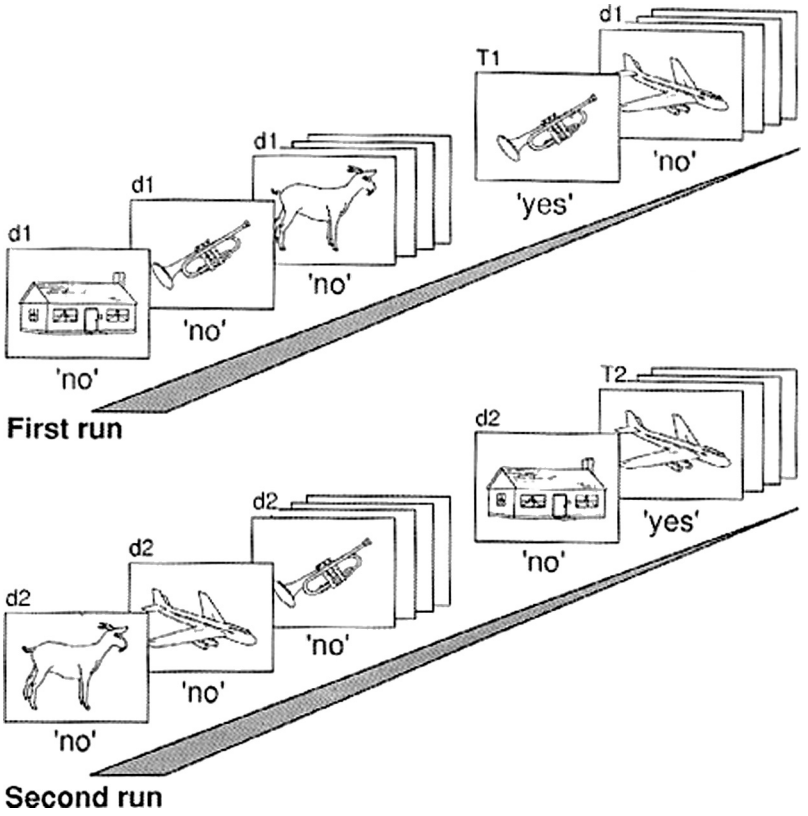

Figure 1 The experiment. In the first run, subjects see a long series of pictures, among which several pictures are repeatedly presented. Subjects have to indicate picture recurrences. This run measures learning and recognition of new information. The second run is composed of precisely the same picture series, arranged in a different order. Subjects are asked to forget that they have already seen all pictures and to indicate picture recurrences only within this current second run. d1, d2, first presentation of a picture in run 1 and run 2 (distracters); T1, T2, picture recurrences within run 1 and run 2 (targets). Variations of this task with different numbers of runs, items, and intervals were used in our studies. Schnider, A., 2001. Spontaneous confabulation, reality monitoring, and the limbic system–a review. [Review, 70 refs]. Brain Research - Brain Research Reviews 36 (2–3), 150–160. Reproduced with permission from Elsevier.

suppress memories that were currently incorrect. However, this explanation did not accommodate processes such as day-dreaming or creativity. Therefore, this deficit was later reformulated as an inability to 'filter' incoming information that pertains to ongoing reality (Schnider, 2008). Schnider's results with this task have been replicated either totally or partially (Schnider and Ptak, 1999; Schnider et al., 2000a; Turner et al., 2010; Gilboa et al., 2006; Bajo, 2011).

Schnider followed up eight confabulators who performed the TCC task again after 3 years (Schnider et al., 2000a). At follow-up seven out of the eight patients had stopped confabulating, and they performed normally on the TCC task. No other cognitive measures used (i.e., California Verbal Learning Task, a verbal and design fluency task, and the short Stroop test) paralleled the course of confabulation. Therefore, Schnider concluded that TCC was necessary and sufficient to explain the presence and clinical course of confabulation (Schnider et al., 2000a).

Over time significant limitations of the TCC task have been described. Schnider himself described a confirmed confabulator who performed normally on the TCC task. This was attributed to the patient having adopted a conservative response criterion as a strategy in order to cope with the confusion caused by the TCC task (Schnider, 2008: p. 251). Bajo (2011) also found seven confabulators who performed at the control level on this task. Another patient did not confabulate despite abnormal TCC performance. Schnider suggested that perhaps this person's preserved memory function had overridden the effects of TCC, thereby stopping confabulation (Schnider, 2008: p. 251). Gilboa et al. (2006) and Bajo (2011) also found that a group of patients who had suffered ACoA and a group of brain injured survivors with moderate to severe dysexecutive difficulties did poorly on this task despite being free from confabulations. This indicates that although TCC may be sensitive to confabulation, it is not specific to the syndrome. Instead, TCC may be specific to pathology that affects the ventromedial prefrontal cortex (VMFC) (Gilboa, 2010). Indeed, in a recent paper, Schnider and colleagues concluded that TCC alone cannot predict the presence of behavioral confabulation without reference to level of amnesia (Nahum et al., 2012).

Imaging Evidence for Temporality Models

Further reformulation of Schnider's model came from neuroimaging studies. Schnider et al. (2000b) scanned a healthy sample while carrying out a more difficult version of the TCC task. Results showed brain activation during the critical part of the task (run 2) corresponded to the posterior medial OFC (BA13).

Evoked potentials and cortical activation maps showed a striking difference in frontal brain activation between distracters in run 2 and other stimuli. At 200–300 ms after stimuli presentation, processing of distracters in run 2 skipped a stage of cortical synchronization. This early activity occurred even before encoding took place (at 400–600 ms) (Schnider et al., 2002).

One explanation put forward was that activated memories go through a preconscious 'filtering' stage. If memories are not currently relevant, the OFC inhibits cortical synchronization. In this way memories are tagged as either being currently relevant or not, and go on to influence behavior accordingly (Schnider, 2003, 2008: p. 269).

Why should confabulators be so convinced of their false memories even in the face of contradictory evidence? Schnider speculated that confabulations represent a behavioral extinction deficit. He noted that primates with ablations of the OFC had trouble extinguishing responses that are no longer required. Therefore he proposed that perhaps confabulators had difficulty extinguishing memories that are no longer relevant to their current situation.

PET experiments in which healthy participants performed decision-making tasks showed that extinction trials underwent different processing to other trials at 200–300 ms. The areas involved were BA10 and BA13. One explanation was that the OFC acted as an 'outcome monitoring system' signaling the nonoccurrence of a predicted event through cortical activation. Absence of this signal would lead to a false sense that predictions had been met. In the case of confabulators, it would provide certainty that the event recalled was correct (Schnider, 2008: p. 291).

Summary of Temporality Models

In summary, temporal models focus on a strikingly distorted sense of time which is characteristic of confabulations. Clinical experiments suggested TCC was the critical disorder that predicted both the appearance and recovery from confabulation. Imaging experiments suggested the OFC was responsible for early inhibition of cortical synchronization of memories that are not currently relevant. Schnider proposed that confabulations appear when this early 'filtering' or 'extinction' process is faulty.

Aspects of this model do not seem incompatible with some features of strategic retrieval models, which are reviewed next. The basic deficit in confabulation proposed by Schnider is one of faulty monitoring during memory reconstruction (Schnider, 2008: p. 270). Indeed Schnider's 'reality filtering' mechanism is not dissimilar to other filtering processes (Gilboa, 2010); compare Burgess' notion of 'gateway' sited in BA10 (Burgess et al., 2007a,b). In addition, the OFC as an 'outcome monitoring system' resembles the idea that the OFC is involved in 'matching events to internal predictions' (Turner et al., 2008b). However, despite these commonalities, the models have very different emphasis.

Current Main Theories of Confabulation: Strategic Retrieval Models

Early authors (e.g., Korsakoff, Kraepelin, and Tiling) explained confabulations as a combination of amnesia deficits and a monitoring deficit (DeLuca, 2009; Schnider, 2008). Kapur and Benson had highlighted the role of executive dysfunction on confabulation (Kapur and Coughlan, 1980; Benson et al., 1996). The following explanations of confabulation hinge on the dysfunction of specific control mechanisms that operate during normal recollection.

Memory Reconstruction Processes

Many authors view memory as a series of constructive processes, and as such, it is open to errors during memory reconstruction (Schacter et al., 1998; Moscovitch and Melo, 1997). Cognitive mechanisms proposed by these models apply to both normal and abnormal memory. Two key processes of successful memory retrieval have been proposed:

1. Different elements of the information to be retrieved must be bound together in order to produce a coherent memory. Dysfunction of this process leads to source memory deficits, where events are recollected without important details, for example, timing or location in which the event occurred.
2. Similar events must be distinguished from the specific memory to be retrieved. If this does not happen, only the gist of an event would be remembered.

Control Mechanisms Underlying Memory Reconstruction

According to this view, confabulation (of any type) arises as a consequence of dysfunction of one or more executive processes involved in memory reconstruction (Moscovitch and Winocur, 2002; Burgess and Shallice, 1996; Kopelman, 1999). These can be related either to the search strategy or to the monitoring of the retrieved information (Moscovitch and Melo, 1997). Three main strategic retrieval processes are described: 'descriptor,' 'editing,' and 'mediating' processes (Burgess and Shallice, 1996; Burgess and McNeil, 1999; Gilboa and Moscovitch, 2002). Figure 2 depicts this model from the viewpoint of Burgess and Shallice (1996).

1. Descriptor or cue specification processes are responsible for generating internal cues to guide information retrieval. Normally, generic memories are used as default 'starting value templates' to speed up the process of memory reconstruction (Burgess and Shallice, 1996; Moscovitch and Winocur, 2002). They are subsequently refined and modified as they undergo verification (Moscovitch and Melo, 1997). However, in the absence of adequate specifications to guide recollection, these generic 'input templates' are erroneously reported as the target memory (Burgess and Shallice, 1996). The absence of descriptor cues could also result in exaggerated external cueing effects in recollection; this would be the case, for example, of a patient who mistakenly reported she was a member of staff in the hospital in which she was a patient (Kopelman et al., 1997).
2. Editing or monitoring mechanisms are responsible for verification, accuracy checking, and comparison of retrieved memories (Burgess and Shallice, 1996; Moscovitch and Winocur, 2002). Malfunction of these verification processes would mean that retrieval of erroneous material would go unchecked.

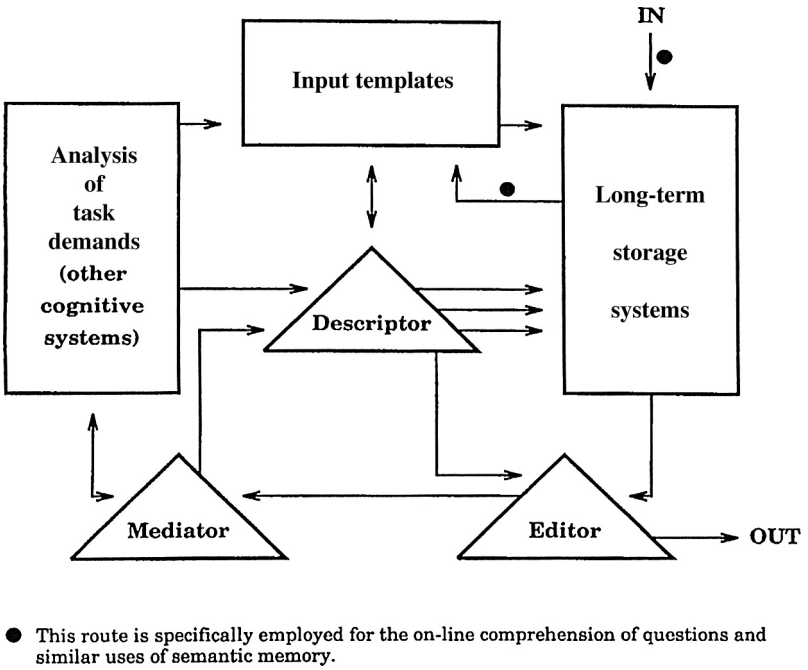

Figure 2 Relationship between memory control processes (in triangles) and other memory and cognitive systems. In this model, autobiographical recollection occurs through the action of two processes. An input template (i.e., a generic representation derived from perceptual and cognitive input and pre-existing (semantic) associations) activates key nodes within the long-term store. These nodes are linked by paths of differing associative strength. Recall of semantic information proceeds through activation of these nodes, where a node is a context-independent representation. Autobiographical recollection proceeds through activation of these nodes via a template concurrently with a descriptor process. The descriptor defines a precise set of paths in the net which corresponds to aspects of specific events. Recollection itself is the constant reiteration of descriptors in such a way as to 'explore' the net. The editor detects incompatibilities either within the recall structure or between the original description and the output representation. Resolution of these incompatibilities, impasses, or failures in activation proceeds through problem-solving routines initiated by the mediator, to which the other cognitive systems are enslaved. The products of these supervisory processes are used to form new descriptions. Burgess, P.W., Shallice, T., 1996. Confabulation and the control of recollection. Memory 4 (4), 359–411. Reproduced with permission from Elsevier.

3. Mediator processes are problem-solving processes, not necessarily specific to memory. If retrieval of a false memory is detected, mediating processes would prevent its erroneous output and redirect the search. For example, fantastic confabulations are said to be caused by malfunction of mediator processes which fail to check the general plausibility of the material remembered.

One criticism of the above model is that the monitoring processes proposed are not specific to confabulation – they apply to true memories too; furthermore, they do not distinguish between different types of confabulation (Schnider, 2008: pp. 227–228). A further critique argues that these models rely on nonobservable processes that make verification of these theories difficult (Dalla Barba, 2009). For example, there is a lack of evidence concerning the physiological basis of the monitoring processes described by these models (Schnider, 2008: p. 228). Some retrieval models have been developed without reference to clinical data (Schnider, 2001). Finally, they do not explain the temporal confusion that usually accompanies confabulation (Schnider, 2003).

Cognitive Mechanisms Underlying Confabulation

A recent model has structured the above processes in a different way and extended the above theoretical explanations. Gilboa and colleagues (Moscovitch and Melo, 1997; Gilboa, 2010) proposed that three types of cognitive mechanisms underlie confabulation: (1) 'core' processes, which are essential for the appearance and course of confabulations, and are specific to this syndrome; (2) 'constitutional' processes, which may be necessary but are not specific to confabulations; and (3) 'associated' features, which are not critical for the appearance of confabulations, but may influence their content. Three types of core processes are reviewed below, followed by a summary of constitutional processes and associated features, as proposed by Gilboa and colleagues.

Core Processes

Three core mechanisms are described; the first two are different types of monitoring processes: (1) a 'preconscious feeling of rightness (FOR)'; (2) conscious monitoring mechanisms; and (3) control processes.

1. 'Preconscious FOR' mechanisms appear to be heuristic decision-making processes that verify retrieved memory traces and tag them accordingly very early in the process of retrieval. When these processes break down, false memories may be accepted as correct. In addition, false memories that are autobiographical in nature may direct behavior, because they are held with great conviction. This may be because autobiographical information is both very well established and salient to each individual.
2. Conscious monitoring mechanisms are more thorough verification processes, which take place later in the process of retrieval (Gilboa, 2010; Burgess and Shallice, 1996). Memories are verified against retrieval cues and contrasted with available information. If these mechanisms function correctly, they may overhaul FOR processes and prevent confabulations. Breakdown of conscious monitoring mechanisms may result in fantastic or inconsistent confabulations.
3. Control processes allow us to consider the consequences of reporting the retrieved material and help us decide whether to report it or not. For example, if the cost of reporting a memory is too great should it turn out to be false, the person may decide to withhold it. Even if the other two core mechanisms are faulty, control processes may stop a confabulation from being reported. For this reason, all three core mechanisms must be faulty in order for confabulations to appear.

Constitutional Processes

Constitutional processes (Gilboa, 2010) are those that are usually associated with confabulation, but are not unique to confabulation. Deficits in these processes may be necessary for the appearance of confabulation, but they are not sufficient. For example, a memory deficit seems to always accompany confabulation but it is not specific to this syndrome. Search initiation and cue specification deficits may also contribute to the appearance of confabulation, but can also appear in its absence. In contrast with Schnider (2008), some authors (Gilboa, 2010; Bajo, 2011) claim that TCC may be sensitive but not specific to the presence of confabulation. Failure in this task has been associated with particular brain pathology in the VMFC, rather than with confabulation per se (Gilboa et al., 2006).

Associated Features

Associated features may be present in connection with confabulations but are often absent, therefore they are not considered constitutional to the syndrome. They typically determine the content of confabulation and form part of normal reconstructive function. For example, many confabulations have a self-serving bias (see Fotopoulou, 2010, for a review), but this may not be a universal bias (Metcalf et al., 2010; Bajo et al., 2010). Confabulations often represent 'generic' well-rehearsed memories (Burgess and McNeil, 1999; Dab et al., 1999; Metcalf et al., 2010). These seem to be exacerbations of normal recollection errors and appear in many contexts. Perseveration has also been observed in confabulators, especially in semantic memory (Kopelman et al., 1997), but this seems to be associated with frontal pathology in general.

Imaging Evidence in Support of Strategic Retrieval Models

Strategic retrieval models are based on imaging and behavioral experimental data in support of both monitoring mechanisms and control processes. Imaging data have demonstrated that patients with VMFC lesions skipped an early positive modulation that was present in controls at 240 ms poststimulus (Gilboa, 2010). This appears to be consistent with Schnider's imaging findings which showed an early deactivation (Schnider et al., 2002). Furthermore, Gilboa's patients also lacked a positive modulation that controls showed at a later stage: 350 ms poststimulus. Gilboa (2010) argued that these two results would support the theory that two distinct monitoring mechanisms are critical for the appearance of confabulation: (1) early 'FOR' and (2) late and more in-depth monitoring processes.

Clinical Evidence in Support of Strategic Retrieval Models

Behavioral experiments showed that confabulators seem to report false memories even when the need for cue specification and creating a search strategy are minimized (Gilboa et al., 2006). One confabulating patient made far more errors than controls, despite showing a general conservative response criterion on a free report memory task (she only reported information if she was at least 75% sure it was true, Gilboa, 2010). These results indicate that confabulation is not merely the consequence of a general bias or a response inhibition deficit, but a very specific monitoring deficit (compare Turner et al., 2008a). Confabulations were also associated with an indiscriminate sense of confidence in their veracity, even when they were highly implausible (Gilboa et al., 2006). These findings are consistent with a specific deficit in 'preconscious FOR' as a critical mechanism in the appearance of confabulation.

Evidence for the role of control processes in confabulation has come from a single-case experiment where the confabulating participant was free to withhold answers during the course of free report (rather than structured interview, Gilboa, 2010). This patient seemed unable to stop generating answers, even after she had reported the correct answer. This is consistent with the critical role of control mechanisms in the appearance of confabulations.

Furthermore, using tasks sensitive to confabulation, Gilboa (2010) found delusional patients presented a distinct pattern of response to that of confabulators. In contrast with confabulators, delusional patients showed poor control function, good monitoring ability, and a liberal response bias. They further differed from confabulators in that they did not show a primary memory deficit.

Summary of Strategic Retrieval Models

In summary, strategic retrieval models explain confabulation as a dysfunction of monitoring mechanisms that operate during normal memory reconstruction. They do not discriminate between different types of confabulation. Core cognitive processes that are critical and specific to the appearance of confabulations include: (1) an early FOR about a memory; (2) a more in-depth monitoring process; and (3) control mechanisms that allow the individual to decide whether to report a memory or not, irrespective of its perceived veracity. All three must be impaired for confabulations to appear. Constitutional processes may be necessary, but they are neither specific to nor sufficient for the appearance of confabulations. Memory deficits and TCC fall into this category. Associated features have been observed during confabulation, but they are often absent. These may affect the content of confabulation, for example, making it predominantly self-serving or generic (i.e., devoid of detail).

Strategic retrieval models provide the most comprehensive explanation of confabulations. They share a common base with motivational models. Both types of models explain confabulations as distortions of normal memory reconstructive mechanisms. However, motivational models make predictions specifically about the content of confabulations. They are reviewed in the next section.

Current Main Theories of Confabulation: Neomotivational Accounts

Modern motivational theories of confabulation do not regard confabulations as arising from compensatory, motivational tendencies or copying mechanisms per se, but rather from a dysregulation (due to neurocognitive impairment) of the organically predetermined influence of motivation on memory formation. Positive, self-serving biases in memory and cognition are now well established (Sharot, 2011; Walker et al., 2003). Neomotivational models suggest that when irrelevant memory representations are not inhibited and memories are not retrieved in appropriate manner (as the various neurocognitive theories support; Burgess and Shallice, 1996; Gilboa et al., 2006; Johnson and Raye, 2000), motivational factors may acquire a greater role in determining which memories are selected for retrieval and accepted as true (Conway and Tacchi, 1996; Solms, 2000; Fotopoulou et al., 2004; Fotopoulou, 2010; Turnbull et al., 2004, 2005). This lack of affective regulation in memory is attributed to the typically reported damage of the ventromedial frontal cortex in confabulation, which is more generally thought to be responsible for affective regulation (e.g., Bechara et al., 2000).

Fotopoulou and colleagues (2004) further put forward the prediction that the false recollections of confabulating patients should show a self-serving bias that is greater than that typically encountered in the memory distortions of healthy volunteers (see above). The content of spontaneous confabulation has been found to contain mostly positive and wishful descriptions in a number of single-case and group studies (e.g., Fotopoulou et al., 2004, 2008b; Turnbull et al., 2004). A positive bias has also been observed in the provoked confabulations of another patient, elicited during an autobiographical memory interview (Fotopoulou et al., 2007a). In the latter study, the patient's confabulations portrayed a self-representation that was more positive than that depicted in his real memories and those depicted in the memories of matched healthy controls. A further case series showed that confabulating patients were more likely to perform reality monitoring errors, accepting false or temporally irrelevant information as part of their recent past, when the information was pleasant rather than unpleasant (Fotopoulou et al., 2007b). For example, a patient was significantly more likely to falsely claim that he actually remembered winning the lottery recently than to falsely claim that he remembered losing his job. Finally, in a recent prose recall study, confabulating patients ($N = 15$) showed a strong selective bias in recalling negative self-referent stories, in that they recalled such information in a manner which portrayed a more positive image of themselves (Fotopoulou et al., 2008a). This positive bias was not present in stories that were not encoded in a self-referent manner. This study shows that confabulating patients do not have difficulty in processing negative emotions in general. Instead, they show a specific self-related motivational bias in their memory.

The self-related bias has been further explored in clinical studies on confabulation (e.g., Fotopoulou, 2008), as well as confabulation in patients with right-hemisphere syndrome affecting their visuospatial and motor awareness (Fotopoulou et al., 2010; Fotopoulou, 2012). This more general perspective

on illusion, delusions, and confabulations concluded that deficits in higher-order areas such as the VMFC and parietal cortex may lead to dynamic exaggerations of self-enhancement, self-deceit, and self-coherence tendencies in neurological patients (Fotopoulou, 2008, 2012, in press). The latter proposal is consistent with recently emerging, multifactorial models of confabulation that include personal goals and wishful ideation among the critical factors that may determine the retrieval of false memories in confabulating patients (Gilboa et al., 2006; Johnson and Raye, 2000; Kopelman, 1999; Metcalf et al., 2007). Acknowledging the role of emotion in confabulation has also pivotal clinical consequences (see Fotopoulou, 2008, for review). It thus seems that at least some authors have recognized that emotion may have a determining role in confabulation and that only its interaction with the neurocognitive factors yields the total psychological picture of the syndrome.

Summary

In summary, confabulation is a debilitating condition that may cause patients to act according to erroneous beliefs. Available evidence indicates confabulation is associated to pathology in the VMFC. At present, there seems to be little consensus among authors regarding its classification and underlying cognitive mechanisms. In addition to a basic amnesic syndrome, several cognitive and emotional factors that may be responsible for confabulation have been proposed: temporal dysfunction, monitoring dysfunction, and motivational factors. Their specific contribution remains unclear, although it is possible that they all contribute to the appearance of confabulation in some measure. Very few longitudinal investigations have been carried out, thus the course of confabulation is at present relatively unknown. Further investigation of confabulation along these lines would be paramount in order to discover the best way to treat this condition.

See also: Agnosia; Amnesia: General; Amnesia: Psychogenic; Autobiographical Memory; Brain Damage: Neuropsychological Rehabilitation; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Cognitive Neuroscience; Episodic Memory; False Memories; Hippocampal Amnesia; Hippocampus and Related Structures; Implicit Memory; Limbic System; Memory Binding; Prefrontal Cortex; Recognition; Reconstructive Memory, Psychology of; Retrieval; Self and Brain; Semantic Memory; Source Monitoring; Temporal Distinctiveness Models of Memory; Temporal Lobe.

Bibliography

Bajo, A., 2011. Confabulation and mood: a programme of studies. Thesis submitted to the Institute of Psychiatry, King's College London. March 2011, London.

Bajo, A., Fleminger, S., Kopelman, 2010. Confabulations are emotionally charged, but not always for the best. Journal of the International Neuropsychological Society 16 (06), 975–983.

Bechara, A., Damasio, H., Damasio, A.R., 2000. Emotion, decision making and the orbitofrontal cortex. Cerebral Cortex 10 (3), 295–307.

Benson, D.F., Djenderedjian, A., Miller, B.L., Pachana, N.A., 1996. Neural basis of confabulation. Neurology 46 (5), 1239–1243.

Berlyne, N., 1972. Confabulation. British Journal of Psychiatry 120 (554), 31–39.

Betlheim, S., Hartmann, H., 1951. On parapraxes in the Korsakow psychosis. In: Rapaport, D. (Ed.), Organisation and Pathology of Thought. Columbia University Press, New York, pp. 289–307 (Original work published 1924).

Box, O., Laing, H., Kopelman, M., 1999. The evolution of spontaneous confabulation, delusional misidentification and a related delusion in a case of severe head injury. Neurocase 5 (3), 251–262.

Burgess, P.W., Dumontheil, I., Gilbert, S.J., 2007a. The gateway hypothesis of rostral prefrontal cortex (area 10) function. Trends in Cognitive Sciences 11 (7), 290–298.

Burgess, P.W., McNeil, J.E., 1999. Content-specific confabulation. Cortex 35 (2), 163–182.

Burgess, P.W., Shallice, T., 1996. Confabulation and the control of recollection. Memory 4 (4), 359–411.

Burgess, P.W., Simons, J.S., Dumontheil, I., Gilbert, S.J., 2007b. The gateway hypothesis of rostral prefrontal cortex (area 10) function. In: Duncan, J., Phillips, L., McLeod, P. (Eds.), Measuring the Mind: Speed, Control, and Age. Oxford University Press, Oxford, pp. 217–248.

Coltheart, M., Turner, M., 2009. Confabulation and delusion. In: Hirstein, W. (Ed.), Confabulation: Views from Neuroscience, Psychiatry, Psychology, and Philosophy. Oxford University Press, NY, pp. 173–187.

Conway, M.A., Tacchi, P.C., 1996. Motivated confabulation. Neurocase 2 (4), 325–338.

Dab, S., Claes, T., Morais, J., Shallice, T., 1999. Confabulation with a selective descriptor process impairment. Cognitive Neuropsychology 16 (3–5), 215–242.

Dalla Barba, G., 1993. Confabulation: knowledge and recollective experience. Cognitive Neuropsychology 10 (1), 1–20.

Dalla Barba, G., 2009. Temporal consciousness and confabulation: escape from unconscious explanatory idols. In: Hirstein, W. (Ed.), Confabulation: Views from Neuroscience, Psychiatry, Psychology, and Philosophy. Oxford University Press, NY, pp. 223–262.

Dalla Barba, G., Boisse, M.-F., 2010. Temporal consciousness and confabulation: is the medial temporal lobe "temporal"? Cognitive Neuropsychiatry 15 (1), 95–117.

Dalla Barba, G., Cappelletti, J.Y., Signorini, M., Denes, G., 1997. Confabulation: remembering 'another' past, planning 'another' future. Neurocase 3 (6), 425–435.

Dalla Barba, G., Nedjam, Z., Dubois, B., 1999. Confabulation, executive functions, and source memory in Alzheimer's disease. Cognitive Neuropsychology 16 (3–5), 385–398.

Davidson, G., 1948. Psychosomatic aspects of the Korsakoff syndrome. Psychiatric Quarterly 22, 1–17.

DeLuca, J., 2000. A cognitive neuroscience perspective on confabulation. Neuro-Psychoanalysis 2 (2), 119–132.

DeLuca, J., 2009. Confabulation in anterior communicating artery syndrome. In: Hirstein, W. (Ed.), Confabulation: Views from Neuroscience, Psychiatry, Psychology, and Philosophy. Oxford University Press, NY, pp. 13–32.

Feinberg, T.E., Roane, D.M., 1997. Anosognosia, completion and confabulation: the neutral-personal dichotomy. Neurocase 3 (1), 73–85.

Fotopoulou, A., 2008. False selves in neuropsychological rehabilitation: the challenge of confabulation. Neuropsychological Rehabilitation 18 (5–6), 541–565.

Fotopoulou, A., 2010. The affective neuropsychology of confabulation and delusion. Cognitive Neuropsychiatry 15 (1), 38–63.

Fotopoulou, A., 2012. Illusions and delusions in anosognosia for hemiplegia: from motor predictions to prior beliefs. Commentary. Brain 135, 1344–1347.

Fotopoulou, A., in press. Time to get rid of the 'Modular' in neuropsychology: a unified theory of anosognosia as aberrant predictive coding. Journal of Neuropsychology.

Fotopoulou, A., Conway, M., Griffiths, P., Birchall, D., Tyrer, S., 2007a. Self-enhancing confabulation: revisiting the motivational hypothesis. Neurocase 13 (1), 6–15.

Fotopoulou, A., Conway, M.A., Solms, M., 2007b. Confabulation: motivated reality monitoring. Neuropsychologia 45 (10), 2180–2190.

Fotopoulou, A., Conway, M.A., Solms, M., Tyrer, S., Kopelman, M., 2008a. Self-serving confabulation in prose recall. Neuropsychologia 46 (5), 1429–1441.

Fotopoulou, A., Conway, M.A., Tyrer, S., Birchall, D., Griffiths, P., Solms, M., 2008b. Is the content of confabulation positive? An experimental study. Cortex 44 (7), 764–772.

Fotopoulou, A., Pergino, S., Maeda, R., Rudd, A., Kopelman, M., 2010. Implicit awareness in anosognosia for hemiplegia: unconscious interference without conscious re-representation. Brain 133, 3564–3577.

Fotopoulou, A., Solms, M., Turnbull, O., 2004. Wishful reality distortions in confabulation: a case report. Neuropsychologia 42 (6), 727–744.

Gainotti, G., 1975. Confabulation of denial in senile dementia. An experimental study. Psychiatria Clinica 8 (3), 99–108.

Gilboa, A., 2010. Strategic retrieval, confabulations, and delusions: theory and data. Cognitive Neuropsychiatry 15 (1), 145–180.

Gilboa, A., Alain, C., Stuss, D.T., Melo, B., Miller, S., Moscovitch, M., 2006. Mechanisms of spontaneous confabulations: a strategic retrieval account. Brain 129 (6), 1399–1414.

Gilboa, A., Moscovitch, M., 2002. The cognitive neuroscience of confabulation: a review and a model. In: Baddeley, A., Kopelman, M.D., Wilson, B.A. (Eds.), The Handbook of Memory Disorders, second ed. John Wiley and Sons Ltd, Chichester, UK, pp. 315–342.

Johnson, M.K., O'Connor, M., Cantor, J., 1997. Confabulation, memory deficits, and frontal dysfunction. Brain and Cognition 34 (2), 189–206.

Johnson, M.K., Raye, C.L., 2000. Cognitive and brain mechanisms of false memories and beliefs. In: Schacter, D.L., Scarry, E. (Eds.), Memory, Brain, and Belief. Harvard University Press, Cambridge, MA, US, pp. 35–86.

Kapur, N., Coughlan, A.K., 1980. Confabulation and frontal lobe dysfunction. Journal of Neurology, Neurosurgery and Psychiatry 43 (5), 461–463.

Kopelman, M.D., 1987. Two types of confabulation. Journal of Neurology Neurosurgery and Psychiatry 50 (11), 1482–1487.

Kopelman, M.D., 1999. Varieties of false memory. Cognitive Neuropsychology 16 (3–5), 197–214.

Kopelman, M.D., 2010. Varieties of confabulation and delusion. Cognitive Neuropsychiatry 15 (1), 14–37.

Kopelman, M.D., Ng, N., van den Brouke, O., 1997. Confabulation extending across episodic, personal, and general semantic memory. Cognitive Neuropsychology 14 (5), 683–712.

Mercer, B., Wapner, W., Gardner, H., Benson, D.F., 1977. A study of confabulation. Archives of Neurology 34 (7), 429–433.

Metcalf, K., Langdon, R., Coltheart, M., 2007. Models of confabulation: a critical review and a new framework. Cognitive Neuropsychology 24 (1), 23–47.

Metcalf, K., Langdon, R., Coltheart, M., 2010. The role of personal biases in the explanation of confabulation. Cognitive Neuropsychiatry 15 (1), 64–94.

Moscovitch, M., Melo, B., 1997. Strategic retrieval and the frontal lobes: evidence from confabulation and amnesia. Neuropsychologia 35 (7), 1017–1034.

Moscovitch, M., Winocur, G., 2002. The frontal cortex and working with memory. In: Stuss, D.T., Knight, R.T. (Eds.), Principles of Frontal Lobe Function. Oxford University Press, NY, pp. 188–209.

Nahum, L., Bouzerda-Wahlen, A., Guggisberg, A., Ptak, R., Schnider, A., 2012. Forms of confabulation: dissociations and associations. Neuropsychologia 50 (10), 2524–2534.

Nathaniel-James, D.A., Frith, C.D., 1996. Confabulation in schizophrenia: evidence of a new form? Psychological Medicine 26 (2), 391–399.

Ptak, R., Schnider, A., 1999. Spontaneous confabulations after orbitofrontal damage: the role of temporal context confusion and self-monitoring. Neurocase 5 (3), 243–250.

Schacter, D.L., Norman, K.A., Koustaal, W., 1998. The cognitive neuroscience of constructive memory [Review, 157 refs.]. Annual Review of Psychology 49, 289–318.

Schnider, A., 2001. Spontaneous confabulation, reality monitoring, and the limbic system–a review [Review, 70 refs]. Brain Research - Brain Research Reviews 36 (2–3), 150–160.

Schnider, A., 2003. Spontaneous confabulation and the adaptation of thought to ongoing reality [Review, 101 refs]. Nature Reviews Neuroscience 4 (8), 662–671.

Schnider, A., 2008. The Confabulating Mind: How the Brain Creates Reality. Oxford University Press, NY.

Schnider, A., Gutbrod, K., Hess, C.W., Schroth, G., 1996. Memory without context: amnesia with confabulations after infarction of the right capsular genu. Journal of Neurology, Neurosurgery and Psychiatry 61 (2), 186–193.

Schnider, A., Ptak, R., 1999. Spontaneous confabulators fail to suppress currently irrelevant memory traces [see comment]. Nature Neuroscience 2 (7), 677–681.

Schnider, A., Ptak, R., von Daniken, C., Remonda, L., 2000a. Recovery from spontaneous confabulations parallels recovery of temporal confusion in memory. Neurology 55 (1), 74–83.

Schnider, A., Treyer, V., Buck, A., 2000b. Selection of currently relevant memories by the human posterior medial orbitofrontal cortex. Journal of Neuroscience 20 (15), 5880–5884.

Schnider, A., Treyer, V., Buck, A., 2005. The human orbitofrontal cortex monitors outcomes even when no reward is at stake. Neuropsychologia 43 (3), 316–323.

Schnider, A., Valenza, N., Morand, S., Michel, C.M., 2002. Early cortical distinction between memories that pertain to ongoing reality and memories that don't. Cerebral Cortex 12 (1), 54–61.

Sharot, 2011. The optimism bias. Current Biology 21 (23), R941–R945.

Solms, M., 2000. A psychoanalytic perspective on confabulation. Neuro-psychoanalysis 2, 133–143.

Talland, G.A., 1961. Confabulation in the Wernicke-Korsakoff syndrome. Journal of Nervous and Mental Disease 132, 361–381.

Talland, G.A., 1965. Deranged Memory. Academic Press, London.

Turnbull, O.H., Berry, H., Evans, C.E., 2004. A positive emotional bias in confabulatory false beliefs about place [Review, 47 refs]. Brain and Cognition 55 (3), 490–494.

Turnbull, O.H., Evans, C.E.Y., Owen, V., 2005. Negative emotions and anosognosia. Cortex 41 (1), 67–75.

Turner, M., Coltheart, M., 2010. Confabulation and delusion: a common monitoring framework. Cognitive Neuropsychiatry 15 (1), 346–376.

Turner, M.S., Cipolotti, L., Shallice, T., 2010. Spontaneous confabulation, temporal context confusion and reality monitoring: a study of three patients with anterior communicating artery aneurysms. Journal of the International Neuropsychological Society 16 (06), 984–994.

Turner, M.S., Cipolotti, L., Yousry, T.A., Shallice, T., 2008a. Confabulation: damage to a specific inferior medial prefrontal system. Cortex 44 (6), 637–648.

Turner, M.S., Simons, J.S., Gilbert, S.J., Frith, C.D., Burgess, P.W., 2008b. Distinct roles for lateral and medial rostral prefrontal cortex in source monitoring of perceived and imagined events. Neuropsychologia 46 (5), 1442–1453.

Van der Horst, L., 1932. Ueber die Psychologie des Korsakowsyndromes. Psychiatry and Neurology 83, 65–84.

Walker, W.R., Skowronski, J.J., Thompson, C.P., 2003. Life is pleasant–and memory helps to keep it that way! Review of General Psychology 7 (2), 203–210.

Weinstein, E.A., 1996. Symbolic aspects of confabulation following brain injury: influence of premorbid personality. Bulletin of the Menninger Clinic 60 (3), 331–350.

Weinstein, E.A., Kahn, R.L., Malitz, S., 1956. Confabulation as a social process. Psychiatry: Journal for the Study of Interpersonal Processes 19, 383–396.

Confederations, History of

Lakatos Artur, George Bariţiu Institute for History, Romanian Academy of Sciences, Cluj-Napoca, Romania

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article addresses the term 'Confederation', its definition, and notable examples history. Usually, scholars in political science and international relations tend to ignore giving a comprehensive definition to this term, and prefer to interpret it in comparison to 'federation', a term providing a close, but at the same time, very different structure to confederation. The illustrative case studies highlight the specific cultural and political environments in which a confederate structure can arise.

Definition of Confederation and Primary Characteristics

A confederation is an entity in the system of international relations that first appeared as a political organization of independent actors, usually of states or their legal equivalents, created by their will to achieve certain common goals or to unite in certain aspects, without giving up their sovereignty.

Among the main characteristics of a confederation are the following: (1) It is a structure created by volunteer association of several sovereign states with the purpose of achieving common goals. (2) Compared with similar interstate systems, it has a transitory, hybrid character, being less centralized than a federation or a union, aside from imperial structures realized by conquest, colonialism, or hegemonic power; however, at the same time, it is more centralized than a league, alliance, or intergovernmental organization. (3) The confederate form of association is appreciated by its founders because it involves collective security and enhancing efforts for the common good of its members, without the loss of individual sovereignty of any of its components; however, these aspects often can be guaranteed by more loose systems, too, like alliance, custom union, or concert of powers. (4) Creating and joining a confederation depends on the will of the respective state or its equivalent, and the theoretical – and often the practical – possibility to leave the structure by secession. (5) The base of a confederation is represented by a founding act or constitution, adopted through consensual decision of its members. In preliterate societies – like the Native American tribes from before the conquest of America by European colonists – the founding treaty or a constitution is not written but rather is transmitted orally. This fact does not affect the respect for the content of the treaty, as it is based on mutual interests, trust, and common law. Changes made to the common legislation usually requires consensus, and if not consensus, at least a majority vote. (6) A confederation usually has central organs, like a common monarch or president, who represents the structure as a single entity toward other actors of international relations, and also has a parliamentary consultative – a legislative institution – with members serving as the delegates of the component political units of the confederation. Because a confederation is created with the respect of the principle of nonintervention of central organs in the internal affairs of its component units, the decisions taken by central organs have a consultative, recommending character rather than a compulsory one. Often, a central organ acts as judge or mediator in internal disputes, because common security and a reduced level of internal conflicts is one of the main goals of any kind of confederate structure. In this latter case, the decision of the organ of judgment is compulsory and is guaranteed by the common will of the members, or at least by common laws. (7) In the ceremonial plan, central institutions prevail to institutions of its component units – for example, the parliament of the confederation is granted higher publicity and respect than a regional or local decision-making institution. The coercive power of the central organs, however, is nonexistent or minimal, because the base of a confederate structure is the maintenance of sovereignty of its composing members. (8) Political institutions of composing units can delegate a part of their competences to the central organs. This happens usually in the plan of common defense, foreign affairs, and more rarely, certain economic plans –for example, like adopting a common currency or the instauration of a custom union. Generally, among the competences of the central institutions are the coordination and harmonization of policies of component units by providing support in this sense for all of the members. (9) In the case of a confederate structure, central organs act only upon their member units, not over the individual citizens of those member units, too, who owe allegiance first to the state where they live and not to the confederate structure as a whole. As a consequence, becoming a citizen in a confederate structure is possible by being granted a citizenship by one of its component states, a practice common in Switzerland, too, where naturalization is granted by the canton and not by the state of Switzerland. The latter body is only accepting the decision made by competent organs of the respective canton. In case of most of the confederate structures, common citizenship does not exist, but there are exceptions as well, like in case of the historical Canadian Confederation (1867–), the United States during its confederate period (1781–89), or during the Old Swiss Confederacy (1291–1798). In some case, coercion through force brought the end of the confederate structure, as in case of the Confederate States of America (1861–65) in the Civil War, or to the Deutscher Bund/German Confederation (1815–66) in 1867, by the victory of Prussia in the Seven Weeks' War. (10) The confederation as structure grants a common national banner under which its members can act in solidarity or individually, but it does not

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62159-1

create a new political nation. (11) Many thinkers consider that confederation as a structure can represent only a temporary solution and is unviable in the long term, because of its loose structure, and with time, it evolves to dissolution toward a more centralized unitary structure, most likely to a federal one. This was the case for Switzerland, the United States, Canada, and Argentina, whereas other historical confederations have ceased to exist. In proper conditions, a confederate structure can represent a viable long-term solution, and does not have to suffer a metamorphosis in a compulsory way. (12) A series of misnomers exist in case of using the term confederation, usually being confused with a 'federation', which designs a different kind of international structure of state association. The term 'confederation' is the perfect synonym for 'confederacy' or 'confederate states', but the term often is mistaken with other similar ones, having a slightly different meaning. Often, the term 'confederacy' or 'confederation' is used to name a federal structure. On the other hand, international structures that often are unions or federations by definition are acting in practice as a functional confederation. According to John T. Rourke, "the European Union provides a current example of what is, or at least, approaches a confederation" (Rourke, 2001: p. 224). (13) In a functional on long-term confederation, there is a high level of interdependency, or at least a permanent common challenge, which prevents its split or dissolution.

The confederation should not be confused with the federation, which is a structure functioning on similar principles, but in a more centralized way. It also should not be confused with the personal union. In the case of a personal union, the key figure is the common monarch, who is holding the function of head of the state of two or more monarchies, with respect to the internal legislation and cultural heritage of each one monarchy and not the will of these states to come together as a centralized new actor of international relations. Personal unions with a confederate character can exist, however, as was the case of the Kalmar Union (1397–1523), when the sovereign kingdoms of Denmark, Sweden, and Norway were unified through a series of personal unions under a common monarch.

In another approach, it is not incorrect to apply the term confederation to certain tribal confederations, which are manifesting themselves with a prestate structure. At the same time, the term should not be used by its modern definition; starting with the Modern Age, the main actors of international relations are the states themselves. In Ratzel's definition, the state is characterized by three major traits: it has territory, a population, and common institutions (Ratzel, 1897). Premodern tribal structures often had more power and prestige than centralized states did; for them, the state structures were far less important than the organic unity of the community of its members.

The following sections present a series of confederate projects of important thinkers, as well as some historical examples for functioning confederations.

The Issue of Tribal Confederations

Structures with a confederate character existed long before the birth of the notion of modern confederation, in different places and time, throughout all the history. Such examples include the possible – unproved yet – existence of the Thinite Confederacy in Ancient Egypt, lead by Menes, which met its end by the unification of Egypt; the coexistence of the 12 biblical Jewish tribes from Joshua until the appearance of the first king, Saul, which also bears elements of confederacy; and the Etruscan and the Samnite civilizations in Italy until their conquest by the Romans. Some classical examples for tribal confederation can be found in the history of European barbarian people of Antiquity, including the Gauls (Mandubii, Boii, Volcae) and Germanics (Bastarnic or Marcommanic tribal confederations). The best examples for so-called tribal confederations can be found in the case of horseback nomadic people of the steppes, which may have been Mongol, Turkic, or Indo-European by their origin, as well as the Xiongnu, Hun, Yuezhi, Rouran, Onogur, and Uyghur tribal confederations. The early principality of Hungary, until the appearance of the centralized Hungarian kingdom, also can be considered to be a tribal confederation. Most of these tribal cooperation can be called a 'tribal confederation' in certain periods of their existence, but only when they have the characteristics of a veritable confederation, concerning relations between central leadership and a tribal one, and when their decision-making process is similar to a confederate one. For example, even if at the beginning of its existence, the Hun tribal alliance resembled a confederate structure, under the rule of Attila the Hun, it cannot be considered a confederation , because the central authority – Attila as king – was the main decision maker. It is more appropriate to call these structures 'tribal alliances', as this term is more general and does not impose a certain kind of power structures.

Statehood and nationalism prevail even in those parts of the world where tribal identity is still strong. There are countries, however, in which a tribal alliance or confederate structures are very strong and determine internal politics. Such is the case in Yemen, where the Hashid, Bakil, and Madhaj tribal confederations are playing important roles in state politics.

Confederate Projects from the Prehistory of European Construction

The process of European Integration has deep roots within the thinking of European intellectuals, dating back for centuries. Most of the projects were designed as federation projects, with a few exceptions. This section describes those projects that envisaged a confederate structure.

One of the first coherent projects for unity of Christian Europe was elaborated by Pierre Dubois (1255–1321), a publicist and jurist who was in service of both Philip the Fair of France and Edward I (Longshank) of England. In his writings, especially in the *De recuperatione terrae sanctae*, he suggested the need for unity of all European Christian nations and monarchs for the success of the crusade against the Muslim world. He suggested the creation of a European confederation under the spiritual authority of the Papacy. The confederation was accompanied by monastic and military reforms, the constitutive states were preserving their sovereignty. The internal peace was granted by a court appointed by the Council of the Members of the Confederation, which could adopt

certain sanctions. As Ladislau Gyémánt wrote, "it was for the first time that such a project took into consideration the issue of the relationship between state sovereignty and supranational institutions – the core issue of the further debates on this topic" (Gyémánt, 1999).

Unlike Dubois, George Podiebrad (in his noble name George of Kunstat and Podiebrady; 1420–71) was a monarch, the king of Bohemia (1458–71). Because of his sympathy for the Hussites, he entered in conflict with Pius II. Conscious of the Ottoman threat, even though Bohemia shared no common border with the emerging Turkish empire, its southern neighbor, Hungary, did, and King Mathias of Hungary was the son-in-law of Podiebrad, and to improve his precarious position among European monarchs, King Mathia presented to King Louis XI of France and to the Republic of Venice a project of an anti–Ottoman European confederation. This project proposed the creation of a confederation of all of the states of Western Christianity, based on a treaty and lead by an assembly, residing in the first 5 years of its existence in Basel, and then changing its location on the basis of a rotation among different European cities. In the Assembly, one vote was granted for each of the sovereigns, with decisions being taken by majority of vote. Besides the coordinated anti-Ottoman efforts, the confederation was meant to grant the existing status quo by preventing war among its members, and with disputes being solved by a Court of Justice or general Consistory, which was appointed by the Assembly. Although Podiebrad's project was rejected, centuries later it inspired Herman Van Rompuy, president of the European Council, to say the following in Prague, on 25 April 2013: "One could think as far back as King Podiebrad, in fifteenth-century Bohemia, who launched an initiative for all European kings and princes to come together in a Council. In fact it was the first ever idea of the European Council!" (Van Rompuy, 2013).

Emeric Crucé (1590–1648), political thinker and one of the first pioneers for the theory of international relations and pacifist ideas, is the author of the first project including the known world of that time in a federal or confederate structure. In his work, Nouveau Cynée ou Discours d'Estat représentant les occasions et moyens d'establir une paix générale et la liberté de commerce pour tout le monde (1623), he projected the creation of an international structure including all of the European states, Russia, the Turkish empire, Persia, China, Japan, Ethiopia, the East and West Indies, Morocco, and the Tatars. The role of this new structure was to maintain permanent peace and to ensure freedom of commerce and development of economy and science. The main ruling body was meant to be a permanent Assembly or Senate of the member states, residing in Venice and presided by the Pope. Crucé's plan did not receive much publicity during his life and in the forthcoming centuries, and most of his ideas seems to be too idealist even for the twenty-first century. It also can be considered to be the early predecessor of the ideas on which the current largest international organization, the United Nations, is based.

Vincenzo Gioberti (1801–52), Italian philosopher, priest, and politician, was highly influenced in his intellectual formation by the religious education received in his youth and by the idealist revolutionary ideas of Giuseppe Mazzini. Gioberti contributed his Church-centered concepts to that time's debate on the future of Italy and the need for its unification. In his view, the unification of Italy was to be realized as a confederation of a series or sovereign regional states, under the authority of the Papacy. But this was meant to be only the first phase in Gioberti's mind, being followed by the unification of the whole Europe in similar terms, granting for each one of the European states and nationalities their own specificity. On the top of this system was meant to be the Pope, exercising the right of arbitration as a peaceful spiritual power.

Lajos Kossuth (1802–94), Hungarian lawyer and politician, was the most representative leader of the Hungarian Revolution and Freedom Fight from 1848 to 1849. He was author of the most coherent revolutionary plan from the period 1848–1914, which was meant to reorganize the state system of East-Central Europe. Following the defeat of the Hungarian Revolution by the Habsburg Monarchy using Tsarist support, Kossuth in emigration became a sworn enemy of the Habsburgs and was planning for the independence of the historical Hungarian kingdom in a nation-state form, which would maintain strict cooperation with its neighbors. As for the future of the entire region, Kossuth expressed his ideas in two successive works, in 1851 and in 1862. The basic ideas for both were the dismantle of the Habsburg empire and the reorganization of the region, using French or Italian backup, in a confederation of independent national states, with the respect of historical borders and tradition.

Postimperial and Postcolonial Transitory Confederate Structures

In the opinion of some scholars – among them the Hungarian professor emeritus Horváth Jenő – certain structures, like the commonwealth; the French Union as it appeared post–World War II with the dissolution process of the British and French colonial empires; or the Commonwealth of the Independent States, created after the fall of Communism by the great part of the states of the former Soviet Union (Horváth, 2005: pp. 30–31) were born because of a need for compromise. The compromise would grant the independence and national sovereignty for those states that had just obtained their independence, but it also would allow the former imperial center to maintain a part of its former influence, at least on a symbolic level. It also prevents violent splits, with disintegration of the structure taking place gradually. The structure is also an umbrella for new, territorially smaller, but more integrated structures like the Eurasian Economic Community.

The Confederation of Serbia and Montenegro

Serbia and Montenegro (2003–06) was a confederation, representing the final moment of the dissolution of the former federal state of Yugoslavia. The confederation had a common president (Svetozar Marovic) and limited competence in issues of defense and foreign affairs. The confederation met its end through the successive formal declarations of independence of its two components in 2006. The independence of Montenegro represented the last step of the dissolution of the federal Yugoslav Republic.

Notable Examples of Confederate Systems in Europe

The German Confederation (1815–66)

The German Confederation or the Deutscher Bund was the result of realistic great power politics of the Congress of Vienna, 1815. The main idea was the creation of a structure that could prevent any possible French future aggression and is in harmony with the principles of the Holly Alliance. The Napoleonic wars proved how vulnerable the divided German-speaking world could be in front of an external great power; however, at the same time, neither Austrian or Prussian hegemony was desired by others. As a consequence, the German Bund was created as a successor of the defunct Holy Roman Empire and of the Napoleonic structures of the Confederation of the Rhine. The peace of Paris (1814) established that the states of Germany should be independent, but also united in a federative structure. The Final Act of the Congress of Vienna (1815) established the constitutional environment of the Bund, which was completed by the Final Act of Vienna (1820). The Bund was meant to ensure internal security of peace of the German-speaking world and possible conflicts were meant to be brought in front of specialized commissions of the Confederate Assembly; if the mediation of the commission failed, the conflict would be submitted to a neutral court. Members of the Confederation had different ranks and rights in the Confederation, and non-German-speaking territories of German monarchs were not included in the Bund. The two chambers of the Congress, the Inner Council and the Plenum, adopted decisions in different ways. In the Plenum, each member had one vote, and in the Inner Council, important states had one vote each, whereas the smaller members – states having rulers that were not crowned heads – had collective voting rights. The Bund met its end with the Prussian victory against Austria in 1866 (Haldén, 2013).

The Swiss Confederation

Switzerland or the Swiss Confederation is an example of an integrated federal state that calls itself a confederation. It was constituted as a confederacy of three cantons from the Alps, as a consequence of common security politics, and with the expansion of the structure, it retained its original name. The confederacy facilitated common interests of its member cantons in common issues of security and defense, leaving a series of legal competence in the cantonal jurisdiction. Throughout history, these often lead to civil war, like the two Wars of Kappel in the sixteenth century or the Sonderbund War in 1847. Although Switzerland now functions as a federal democratic republic, it has retained its official name the Swiss Confederation (in Latin, *Confoederatio Helvetica*). From the point of view of constitutional evolution of Switzerland, three major periods are distinguished in history: the period of the Old Swiss Confederacy (1291–1798), the Napoleonic Era or the years of the Helvetic Republic (1798–1815), and the period of the federal state.

Other Examples

The United States of Belgium (1790) was a short-lived historical predecessor of modern Belgium in the form of a confederate republic of eight provinces from the Southern Netherlands, who rebelled against the rule of the Habsburg emperor Joseph II. The rebellion is known also as 'Brabantse revolutie' orthe 'Brabant revolution', and is one of the best historical examples of Flemish and Walloon coordinated efforts. During the short existence of this state, each one of the provinces had its own government, and the common government had its seat in Brussels, with representatives from all of the eight provinces. The revolt was defeated by the Habsburg Army of Leopold II, the successor of Joseph II, but it represented a model for the subsequent and more successful revolution of 1830, which created the modern state of Belgium. Currently, Belgium is a federal state that shares a few common aspects with a confederation.

Confederate Ireland (1642–49) was created by the special conjuncture of the English Civil War. Following the outbreak of hostilities between the King and the Parliament in England, the Irish Catholic elite remained loyal to the rule of Charles I. The initiative was formulated by the catholic bishop of Ferns, Nicholas French, and the Catholic Church played a major role in the confederation's politics. The central organs were the General Assembly, as legislature, and the Supreme Council, as executive, while the head of state was accepted to be the king of Britain. The Irish Catholic Confederation met its end by the invasion of the English Parliamentary Army led by Oliver Cromwell in 1649.

The European Union has no common classification, and the current structure resembles in some aspects a federation, in others, a confederation. The European Union operates common economic policies and has several common institutions, but its leadership lacks efficient centralization, and its competences in fields like issues of foreign policy and defense, despite all the efforts, are seriously limited. Member states of the European Union are preserving, in great part, their sovereignty.

Confederate Structures of North America

Native American Confederacies of North America

Native Americans had several confederate structures, perhaps the most well known is the Iroquois Confederation, a structure that currently is functioning. It is based on the Great Binding Law, an orally transmitted Constitution. The primary purpose of this confederate structure was to maintain peace among its members. Inequality of its members was institutionalized in the decision-making process. Originally, the confederation had 'five nations', with the Tuscarora becoming the sixth nation in 1722. The common institution for debates was the Confederation Council, which was hosted on the Onondaga territory and had 50 sachems, delegated by the participants. Originally, the Onondaga had 14 sachems, the Cayuga had 10, the Mohawk and Oneida had nine each, and the Seneca had eight. When the Tuscarora joined, they received six sachems, and total membership of the Council was increased to 56. The decision-making process required, at its end, unanimity. When a decision that implied the competence of the whole confederacy was meant to be taken, deliberations were held by each tribe following an established order. The first ones who deliberated the issue were the Mohawks; after they reached a decision, their

proposal was conveyed to the Seneca, and then to the Oneida and to the Cayuga, with the Onondaga being the last (Lutz, 1998). Because of its efficiency, the Confederation earned the respect of other Native American tribes and also of the colonists coming from Europe. The confederacy lost its real power following to the American War of Independence, but a common organization of the Iroquois still exists and sometimes acts as an independent actor of International Relations.

Confederate Experience of the United States of America

In the first phase of the existence of the independent United States of America was in form of a confederation. Its functioning was based on the Articles of Confederation, which first served as a provisional Constitution of the States, starting in 1777. It was replaced by the current US Constitution in 1789. The confederation included independent and sovereign states, which were preoccupied in protecting their freshly won freedom against active and possible threats.

A second phase of the American confederate experience took place in 1860–65, when political and social internal tensions caused the secession of several southern states of the Union. The Confederate States of America was a structure created in 1861 by several member states of the United States of America, as a consequence of their secession from the Union following the election as president of Abraham Lincoln. The United States rejected the secession and considered the confederacy to be illegal and defeated it in a bloody civil war (1861–65), which forced the secessionists to return in the Union. Originally, seven southern states declared their secession and later were joined by another four states following to the outbreak of the Civil War.

The first capital of the Confederacy was Montgomery, in the state of Alabama, from 4 February until 29 May 1861. The capital later moved to Richmond. Its first and only president was Jefferson Davis (1861–65) seconded by Alexander Stevens as vice president, and with the legislative power being represented by a bicameral legislature. The legislation of the confederacy was based on the Confederate States Constitution, which replicated in great part the US Constitution, with a few major differences.

Experience in Canada

Is one of the world's oldest liberal federations, an efficient solution was to keep together in one single unit Catholics and Protestants, French and English speakers, Native Americans and European colonists. The federation was born by a certain conjuncture: The British Crown held its control over these territories, but it was seriously threatened by their southern neighbor, the United States.

The 1867 confederation was designed under the auspice of equal rights of the two major communities. In fact, Canada is a federation and not a confederation. It was constituted in 1867 under the use of this term, as a self-governing Dominion of the British Empire. It consists of provinces and territories, with each one having extended competences in most major aspects of public administration and social life.

Canada has retained its bilingual character, somehow as a compensation for the French Canadians, whose existence within the state provided a cultural impediment on US expansion and a possible forced merging of Canada with the United States.

Tribal Confederacies of Pre-Columbian Mexico

The Chalco Altapetl or confederacy was a tribal alliance of the Chalca people in pre-Columbian Mexico, rivaling the more famous Aztec Empire, with which they waged wars for centuries. When Hernando Cortes landed in Mexico, Chalco was submitted to the Aztecs and paid tribute to them, but it quickly allied itself with the conquistadors and contributed in the defeat of Tenochtitlan. After the fall of the Aztecs, the Chalco accepted Spanish rule without major opposition.

The Aztec Empire was functioning as a confederation of city-states sharing common cultural heritage. Initially, it was created as an alliance of the states of Tenochtitlan, Acolhuacan (known after its capital more as Texcoco), and Tlacopan, coordinating common foreign policy and wars, with each one of them keeping its sovereignty in domestic affairs. With time, Tenochtitlan became the effective leader and the most representative state of the alliance; however, the other two allies kept their competencies and certain responsibilities divided. For example, Acolhuacan had most of its competencies in legislation and construction of public buildings, the elite of Tlacopan was responsible for the economy and handicraft, and Tenochtitlan kept the role of the most important religious center and was responsible for war and common foreign policy (Kulmar, 2010). As a result of its successful wars, the Triple Alliance formed an empire with a series of subsidiary provinces paying tribute to the Aztecs. The Aztec Empire was conquered by the Spaniards lead by Cortes in 1521.

Confederate Ideas and Structures in Latin America, Asia, and Africa

Confederate Experience in Postcolonial Latin America

The Argentine Confederation is one of the official names of Argentina, according to its Constitution. From 1831 to 1852, Argentina could be called an effective confederation when the provinces organized themselves without having a common president. Similar confederate structures in postcolonial Latin America included the Peru–Bolivian Confederation (1836–39) and the Granadine Confederation (1858–63).

Confederate Structures in Asia and the Far East

For most of Europeans, the great civilizations of the Far East are characterized by magnificent ancient culture, fatalist religions, and well-functioning, centralized states. In certain periods of their existence, however, certain cultures had confederate structures, special conjunctures, and a certain amount of historical time. These civilization can be included in the category of 'notable exception' rather than into a general trend, and even if some had lasted for centuries, they had a transitory character, because they were the result of a series of circumstances and not of an expressed need of the society.

During the history of the Korean peninsula, in the Proto-Three Kingdoms period, several independent tribes and

states created confederations. Such confederations were Mahan, Byeonhan, and Jinhan, which developed into the three kingdoms. A subsequent confederacy was the Gaya confederacy, which developed from the Byeonhan and led to a five-centuries-long independence, but was annexed to the kingdom of Silla.

The Confederation of the Madja, a state in the Philippines, was established by Bornean rebels who were forced by the monarch to leave the country. The new state was established in the Visayas Islands and was considered to be very efficient in wars. It fell under Spanish control in a series of successive colonizing actions between 1565 and 1716.

The Maratha Empire from North India was functioning as an aristocratic confederation in the last stage of its existence, starting with the death of Madhavrao I (1772) and ending with its conquest by the British Empire (1818). During this period, the common monarch, the Peshwa, had limited powers, and the Empire's politics were determined by the cooperation and rivalry between its subsidiary local dynasties.

African Confederate Experiences

Sub-Saharan Africa was and still is the home to several tribal confederacies. Some notable structures in African history include the Aro Confederacy (1690–1902) in the southern territories of modern-day Nigeria, with its center in the Arochukwu Kingdom. The confederacy united several city-states and kingdoms, and the binder of the confederacy was represented by a common economic interest of trading palm oil and slaves. The confederacy was destroyed by British conquest in 1902.

The Fante Confederacy and the Ashanti Confederacy are historical tribal confederations in Africa, from among which the Ashanti monarchy still exists in Ghana as a subnation-state. Generally, central institutions of postcolonial African states – where presidents, measured with European standards, often exercised dictatorial powers – are more comfortable accepting the existence and limited sovereignty of traditional tribal power structures than fighting against illegitimate warlords. As a consequence, centralization in a way that would abolish traditional tribal kingdoms and alliances is not very likely to happen in any state of Sub-Saharan Africa.

Pan-Africanism and Postcolonial Confederate Experiences of African States

George B.N. Ayittey, Ghanaian economist and supporter of Pan-Africanism, wrote the following: a "[m]odel would be an African Confederation, with extensive devolution of power to reflect Africa's own heritage: the ancient empires of Ghana, Mali Songhai and Great Zimbabwe were all confederacies, and they lasted centuries" (Ayittey, 2010). Although this sounds like an exaggeration, the pan-African movement is a serious debate among African intellectuals.

Pan-Africanism was developed over the late-ninetieth and early twentieth centuries to enhance emancipation of Black people of the colonies. Its founding belief is that people of Africa, regardless of their tribe, are sharing common problems and objectives. The ideology of Pan-Africanism and unification in an international cooperation of the continent was formed through a series of congresses between 1900 and 1927. During the decades of decolonization, the Organization of African Unity (OAU), in 1963, tried to promote idealistic ideals of democracy. This proved to stay only on a declarative, and as a consequence, the organization quickly lost its credibility because it lacked real power to improve the lives of Africans and to limit excesses of presidents and warlords. In 2001, the successor organization of the African Union was created in Libya.

In addition to the inefficient international organizations meant to enhance integration, two regional confederate attempts are worth mentioning. The Union of African States (or the Ghana-Guinea-Mali Union) was a regional organization formed from 1958 to 1963 in postcolonial Western Africa. It was based on the friendship of the three presidents, Kwame Nkrumah of Ghana (1958–66), Ahmed Sekou Toure of Guinea (1958–84), and Modibo Keita of Mali (1960–68). The charter of the Union was adopted in 1961, containing mutual assistance in case of war and cooperation in the plan of economic and cultural activities. The three leaders expressed high ambitions concerning integration processes, such as a common parliament, common flag, unification of diplomacy and economic policies, and so on. Most of these objectives proved to be unrealistic at that time, and the cooperation never passed through the level of occasional summits of the three presidents.

Senegambia (1982–89) was a confederation between the states of Senegal and Gambia, born because of common issues of internal security. During its existence, the integration process never went further than the military assistance of Senegalese Army against Gambian leftist rebels. After Gambia resisted with success to Senegalese pressure to move closer toward a union, Senegal dissolved the confederation in 1989.

In Africa, there is expressed need for regional or continental integration processes, and responsible politicians and intellectuals are clearly calling for this integration. Until most of the African states are wrestling with serious internal problems, however, the successful completion of an integration process, similar to the European one, is less likely.

State Confederations in the Arab World during the Cold War

From 1958 to 1961, the United Arab States was a confederation of Egypt, Syria, and North Yemen. This confederation was realized in two steps. First, the United Arab Republic (UAR) was formed by the union of Egypt and Syria in 1958, under the leadership of president Gamal Abdel Nasser. The structure was dominated by the Egyptian leading elite, which caused dissatisfaction in Syria, and ultimately led to its secession in 1961. From the region, Israel, Saudi Arabia, and Jordan were the most fervent opponents of the UAR.

In 1958, North Yemen, which previously signed a defense treaty with Nasser's Egypt, joined the UAR in a loose confederation. Unlike Syria, North Yemen remained an independent sovereign state under the rule of its own king, and was acting as an independent actor in global politics. The United Arab States was dissolved in 1961.

The Arab Federation of Iraq and Jordan was a confederate structure formed by the Hashemite kingdoms of Iraq and Jordan, and lasted from 14 February until 2 August 1958. The two kings, Faisal II of Iraq and Hussein, were cousins, and they

formed this union as a reaction to the rising of the menace represented by the ambitions of the Egyptian president Nasser. The Arab Union united the foreign policy and defense functions of the two countries, leaving the rest of the issues under national jurisdiction. Both of the kings kept their thrones, and Nuri as-Said was named as prime minister of the Union. The end of the Union was brought by a military coup in Baghdad, which resulted in the fall and death of King Faisal II.

See also: American Revolution, The; Contemporary History; Conventions and Norms: Philosophical Aspects; Cultural History; Global History: Universal and World; Historical Thought and Historiography: Current Trends; Intellectual History; Multiculturalism; National Liberation Movements; Nationalism, Historical Aspects of: Africa; Nationalism, Historical Aspects of: Arab World; Nationalism, Historical Aspects of: The West; Nations and Nation-States in History; Nomads and Nomadism in History; Political Thought, History of; State, History of.

Bibliography

Ayittey, George B.N., 2010. The United States of Africa: a revisit. The Annals of the American Academy of Political and Social Science 632, 86–102.

Gyémánt, Ladislau, 1999. Preistoria Constructiei Europene/Prehistory of the European Construction. Editura Fundatiei Pentru Studii Europene, Cluj-Napoca.

Haldén, Peter, 2013. Republican continuities in the vienna order and the German confederation 1815–1866. European Journal of International Relations 19 (2), 281–304.

Horváth, Jenő (Ed.), 2005. Világpolitikai Lexikon (1945–2005). Osiris Kiadó, Budapest.

Kulmar, Talmo, 2010. About the comparison of the state authority and social organization by Incas and Aztecs. Folklore: Electronic Journal of Folklore 45, 137–150.

Lutz, Donald S., 1998. The Iroquois confederation constitution: an analysis. Publius: The Journal of Federalism 28 (2), 99–127.

Ratzel, Friedrich, 1897. Politische Geographie. Verlang von R. Oldenburg, München und Leipzig.

Van Rompuy, Herman, 25 April, 2013. "At the Heart of the Union: The Czech Republic and the Future of Europe", Speech by Herman Van Rompuy, President of the European Council at the National Forum on the Future of the EU, Prague. http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ec/136940.pdf.

Rourke, John T., 2001. International Politics on the World Stage, eighth ed. McGrew-Hill/Dushkin.

Confidentiality and Statistical Disclosure Limitations

George T Duncan, Carnegie Mellon University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 4, pp. 2521–2525, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Under confidentiality, identifiable data provided for statistical purposes is protected from unauthorized disclosure. Organizations acting as brokers between respondents and data users seek to disseminate useful data products while keeping low the risk of confidentiality disclosure. Recognizing that deidentification of each data record is generally inadequate to protect its confidentiality against attack by a data snooper, agencies restrict the data they release for general use. Typically, these restricted data procedures have involved transformation or masking of the original, collected data through such devices as adding noise, topcoding, data swapping, and recoding. Another approach is to use the original data to determine a statistical model and use it to generate synthetic data. Generically, statistical disclosure limitation is a body of restricted data procedure that transforms data so that release of the transformed data adequately limits disclosure risk. Desirably, statistical analysis of the transformed data leads to inferences similar to that obtained by analysis of the original data. The technical procedures for implementation of disclosure limitation involve a range of mathematical and statistical tools.

Data Quality Rests on Usefulness and Confidentiality Protection

We capture enormous numbers of rich personal and proprietary records, store them in data warehouses measured in terabytes, analyze them using sophisticated computers and algorithms, and disseminate them worldwide via telecommunications channels that are cheap and fast. This explosive progress in information technology – look at the World Wide Web – heightens the tension between confidentiality and data access. Accepted principles of information ethics (Duncan et al., 1993) require that promises of confidentiality be preserved and that citizens have substantial access to information. Policies and procedures are needed to reconcile the data provider's interest in confidentiality and the data user's demand for data (Dalenius, 1988). Government statistical agencies, such as the US Census Bureau, play a particularly significant and delicate brokering role in this reconciliation of competing demands. They often have legal authority to compel individuals and firms to provide data, and the data are both sensitive – like personal income – and so comprehensive and accurate to be in substantial demand by a broad range of data users. Academic researchers, corporate planners, journalists, and government analysts all seek such data. Other organizations, such as health care providers, necessarily collect data in treating patients that is highly personal. These data are of substantial research value for issues spanning medical quality assurance, cost containment, and access to medical treatment by the poor (Duncan and Kaufman, 1996).

A compromise of the confidentiality pledge could harm the agency, the respondent, the public, or some other group. A statistical disclosure occurs when the data dissemination allows data snoopers to gain excessive new information about respondents whereby the snooper can isolate individual respondents and corresponding sensitive attribute values. Miller (1971: 176) illustrates the problem.

> Some deficiencies inevitably crop up even in the Census Bureau. In 1963, for example, it reportedly provided the American Medical Association with a statistical list of one hundred and eighty-eight doctors living in Illinois. The list was broken down into more than two dozen categories, and each category was further subdivided by medical specialty and area residence; as a result, identification of individual doctors was possible…

As shown in Figure 1, the process of assuring confidentiality through statistical disclosure limitation has the following components:

1. a data quality audit that, beginning with the original collected data, evaluates data utility and assesses disclosure risk;
2. a determination of adequacy of confidentiality protection;
3. if confidentiality protection is inadequate, the implementation of a disclosure limitation procedure; and
4. a return to the data quality audit.

Data Quality Audit: Data Utility and Disclosure Risk

A statistical agency audits the collected data, both to evaluate the utility of the data and to assess disclosure risk. Typically, with good survey design and implementation, the data utility is high. But, also, the risk of disclosure through the release of the original, collected data is too high, even when the data collected have been deidentified, i.e., apparent identifiers (name, e-mail address, phone number, etc.) have been removed. Reidentification techniques have become too sophisticated to assure confidentiality protection (Winkler, 1998). A confidentiality audit will include identification of

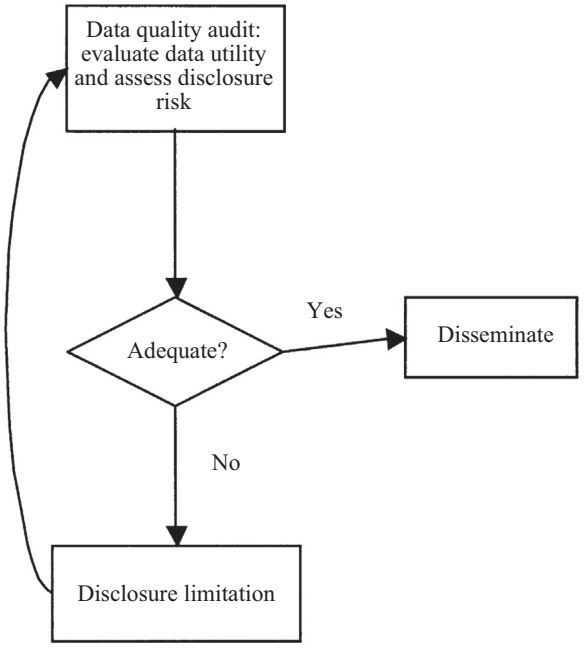

Figure 1 Statistical disclosure limitation process.

(1) sensitive objects and (2) characteristics of the data that make it susceptible to attack.

Based on its confidentiality pledges, a statistical agency seeks to protect certain sensitive objects. An agency seeks to protect the instantiation of sensitive objects in the data from a data snooper. Sensitive objects can be the instantiations of a variety of variables associated with a subject entity (person, household, enterprise, etc.). Examples include the values of numerical variables, such as household income, a realization of an x-ray of a patient's lung, and a specific patent application. A second class of sensitive objects is instances of relationships, whether social or mathematical. An example of the former is 'Who is friends with whom?' in a social network; an example of the latter is an Internal Revenue Service formula to trigger a tax audit.

Collected data having particular characteristics pose substantial risk of disclosure. Characteristics that suggest vulnerability include:

1. geographical detail (Greenberg and Zayatz, 1992);
2. longitudinal or panel structure;
3. outliers;
4. many attribute variables;
5. population data, as in a census, rather than a survey with small sampling fraction; and
6. existence of databases that are publicly available, identified, and share individual respondents and attribute variables with the subject data.

Data with geographical detail, such as census tract data, may be easily linked to known characteristics of respondents. Concern for this leads statistical agencies to place minimum population levels for geographical identifiers that they will release. Longitudinal data, which tracks entities over time, also poses substantial disclosure risk. Many electrical engineers lived in the Chicago area in 1998 and many electrical engineers lived in the Phoenix area in 1999, but few did both. Outliers, especially on variables like net worth, can easily lead to identifiable respondents. Data with many attribute variables allows easier linkage with known attributes of identified entities, and entities, which are unique in the sample, are more likely to be unique in the population. Further, population data, i.e., a census or near census, pose more disclosure risk than data that arise from a survey having a small sampling fraction. Finally, special concern must be shown when other databases are available to the data snooper and these databases are both identified and share with the subject data both individual respondents and certain attribute variables. Record linkage may then be possible between the subject data and the external database. The shared attribute variables provide the key.

Types of Statistical Disclosure

The legitimate objects of inquiry for statistical research are aggregates over individual records, for example, the median annual income of Hispanic social scientists in the United States. The statistical agency seeks to provide users with data that will allow accurate inference about such population characteristics. At the same time, because of confidentiality the statistical agency seeks to thwart the data snooper who might seek to use the disseminated data to draw accurate inferences about, say, the income of a particular Hispanic sociologist who now works in New York City. This capability by a data snooper represents a statistical disclosure.

There are two major types of disclosure – identity disclosure and attribute disclosure. *Identity disclosure* occurs with the association of a respondent's identity and a disseminated data record (Spruill, 1983). *Attribute disclosure* occurs with the association of either an attribute value in the disseminated data or an estimated attribute value based on the disseminated data with the respondent (Duncan and Lambert 1989; Lambert, 1993). In the case of identity disclosure, the association is assumed exact. In the case of attribute disclosure, the association can be approximate. Most statistical agencies place emphasis on limiting the risk of identity disclosure, perhaps because of its substantial equivalence to the inadvertent release of an identified record – a clear administrative slip up. On the other hand, an attribute disclosure, even though it invades the privacy of a respondent, may not be so easily traceable to actions of the agency.

Measures of Disclosure Risk

In the context of identity disclosure, disclosure risk can arise because a data snooper may be able to use the disseminated data product to re-identify some deidentified records. Spruill (1983) proposed a measure of disclosure risk for microdata: For each 'test' record in the masked file, compute the Euclidean distance between the test record and each record in the source file. Determine the percentage of test records that are closer to their parent source record than to any other source record. She defines the risk of disclosure to be the percentage of test records

that match the correct parent record multiplied by the sampling fraction (fraction of source records released).

More generally, and consistent with Duncan and Lambert (1986, 1989), the agency will have succeeded in protecting the confidentiality of a released data product if the data snooper remains sufficiently uncertain about a protected target value after data release. From this perspective, a measure of disclosure risk is built on measures of uncertainty. Further, the agency may model the decision making of the data snooper as a basis for using disclosure limitation to deter inferences about a target. Data snoopers are deterred from publicly making inferences about a target when their uncertainty is sufficiently high. Mathematically, uncertainty functions provide a workable framework for this analysis. Examples include Shannon entropy, which has found use in categorizing continuous microdata and coarsening of categorical data (Willenborg and de Waal, 1996: 138).

Generally, a data snooper has *a priori* knowledge about a target, often in the form of a database with identified records (Adam and Wortmann, 1989). Certain variables may be in common with the subject database. These variables are called *key* or *identifying*. When a single record matches on the key variables, the data snooper has a candidate record for identification. This candidacy is promoted to an actual identification if the data snooper is convinced that the individual is in the target database. This would be the case either if the data snooper has auxiliary information to that effect or if the data snooper is convinced that the individual is unique in the population. The data snooper may find that according to certain key variables that a sample record is unique. The question then is whether the individual is also unique on these key variables in the population. Bethlehem et al. (1990) have examined detection of records agreeing on simple combinations of keys based on discrete variables in the files. Record linkage methodologies have been examined by Fuller (1993) and Winkler (1998).

Restricted Data through Statistical Disclosure Limitation

Direct transformations of data for confidentiality purposes are called *disclosure limiting masks* (Jabine, 1993). With masked data sets, there is a specific functional relationship, possibly as a function of multiple records and possibly as a stochastic function, between masked values and the original data. Because of this relationship, the possibilities of both identity and attribute disclosures continue to exist, even though the risk of disclosure may be substantially reduced. The idea is to provide a response that, while useful for statistical analysis purposes, has sufficiently low disclosure risk. As a general classification, disclosure-limiting masks can be categorized as suppressions, recodings, or samplings.

Suppression

A suppression is a refusal to provide a data instance. For microdata, this can involve the deletion of all values of some particularly sensitive variable. In principle, certain record values could also be suppressed, but this is usually handled through recoding. For tabular data, the values of table cells that pose confidentiality problems are suppressed. These are the primary suppressions. Often, a cell is considered unsafe for publication according to the (n, p) dominance rule, i.e., if a few (n), say 3, contributing entities represent a percentage p, say 70%, or more of the total. Additionally, enough other cells are suppressed so that the values of the primary suppressions cannot be inferred from released table margins. These additional cells are called secondary suppressions. Tables of realistic dimensionality even with only a few primary suppressions present a multitude of possible configurations for the secondary cell suppressions. This raises computational difficulties that can be formulated as combinatorial optimization problems. Typical techniques that are employed include mathematical programming (especially integer programming), and graph theory (Chowdhury et al., 1999).

Recoding

A disclosure-limiting mask for recoding creates a set of data for which some or all of the attribute values have been altered. Recoding can be applied to microdata or to tabular data.

Examples of recoding as applied to microdata include data swapping, adding noise, and global recoding and local suppression. In data swapping (Dalenius and Reiss, 1982; Spruill, 1983), some fields of a record are swapped with the corresponding fields in another record. Concerns have been raised that while data swapping lowers disclosure risk it may excessively distort the statistical structure of the original data (Adam and Wortmann, 1989). A combination of data swapping with additive noise has been suggested by Fuller (1993). Masking through the introduction of additive or multiplicative noise has been investigated (e.g., Fuller, 1993). A disclosure limitation method for microdata that is used in the μ-Argus software is a combination of global recoding and local suppression. Global recoding combines several categories of a variable to form less-specific categories. Topcoding is a specific example of global recoding. Local suppression suppresses certain values of individual variables (Willenborg and de Waal, 1996). The aim is to reduce the set of records where only a few agree on particular combinations of key values. Both methods make the data less specific and so result in some information loss to the legitimate researcher.

Some common methods of recoding for tabular data are global recoding and rounding. Under global recoding, categories are combined. This represents a coarsening of the data through combining rows or combining columns of the table. Under rounding, every cell entry is rounded to some base b. The controlled rounding problem is to find some perturbation of the original entries that will satisfy (marginal, typically) constraints and that is 'close' to the original entries (Cox, 1987). Multidimensional tables present special difficulties. Methods for dealing with them are given by Kelley et al. (1990).

Markov perturbation (Duncan and Fienberg, 1999) makes use of stochastic perturbation through entity moves according to a Markov chain. Because of the cross-classified constraints imposed by the fixing of marginal totals, moves must be *coupled*. This coupling is consistent with a Gröbner basis

structure (Fienberg et al., 1998). In a graphical representation, it is consistent with data flows corresponding to an *alternating cycle*, as discussed by Cox (1987).

Sampling

Sampling as a disclosure-limiting mask creates an appropriate statistical sample of the original data. Alternatively, if the original data is itself a sample, the data may be considered self-masked. Just the fact that the data are a sample may not result in disclosure risk sufficiently low to permit data dissemination. In that case, subsampling may be required to obtain a data product with adequately low disclosure risk.

Whether for microdata or tabular data, many of these transformations can be represented as matrix masks (Duncan and Pearson, 1991), $\mathbf{M} = \mathbf{AXB} + \mathbf{C}$, where $\mathbf{X}$ is a data matrix, say $n \times p$. In general, the defining matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ can both depend on the values of $\mathbf{X}$ and be stochastic. The matrix $\mathbf{A}$ (since it operates on the rows of $\mathbf{X}$ is a record-transforming mask), the matrix $\mathbf{B}$ (since it operates on the columns of $\mathbf{X}$ is a variable-transforming mask), and the matrix $\mathbf{C}$ is a displacing mask (noise addition).

Synthetic or Model-Based Data

The methods described so far have involved perturbations or masking of the original data. These are called *data-conditioned* methods by Duncan and Fienberg (1999). Another approach, while, less studied, should be conceptually familiar to statisticians. Consider the original data to be a realization according to some statistical model. Replace the original data with samples (the synthetic data) according to the model. Synthetic data sets consist of records of individual synthetic units rather than records the agency holds for actual units. Rubin (1993) suggested synthetic data construction through a multiple imputation method. It remains an open research question as to what the impact of imputation of an entire microdata set is on data utility.

Rubin (1993) asserts that the risk of identity disclosure can be eliminated through the dissemination of synthetic data and proposes the release of synthetic microdata sets for public use. His reasoning is that the synthetic data carries no direct functional link between the original data and the disseminated data. So while there can be substantial identity disclosure risk with (inadequately) masked data, identity disclosure is, in a strict sense, impossible with the release of synthetic data. However, the release of synthetic data may still involve risk of attribute disclosure (Fienberg et al., 1998).

Rubin (1993) cogently argues that the release of synthetic data has advantages over other data dissemination strategies, because

1. masked data can require special software for its proper analysis for each combination of *analysis* × *masking method* × *database type* (Fuller, 1993);
2. release of aggregates, e.g., summary statistics or tables, is inadequate due of the difficulty in contemplating at the data-release stage what analysts might like to do with the data; and
3. mechanisms for the release of microdata under restricted access conditions, e.g., user-specific administrative controls, can never fully satisfy the demands for publicly available microdata.

The methodology for the release of synthetic data is simple in concept, but complex in implementation. Conceptually, the agency would use the original data to determine a model to generate the synthetic data. But the purpose of this model is not the usual prediction, control, or scientific understanding that argues for parsimony through Occam's Razor. Instead, its purpose is to generate synthetic data useful to a wide range of users. The agency must recognize uncertainty in both model form and the values of model parameters. This argues for the relevance of hierarchical and mixture models to generate the synthetic data.

Summary

A society based on democratic and free market principles cannot function without broad access to data, nor can it sustain those principles without affirming the individual autonomy assured by privacy and confidentiality. Statistical disclosure limitation methods work to ensure that the tension between data access and confidentiality can be resolved in ways that are favorable to both. This methodology draws on disparate fields of mathematics and is challenged to develop in response to the growing capability of information technology to capture, store and disseminate data. To learn more about this topic, see *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics* (Duncan et al., 1993), two special issues of the *Journal of Official Statistics* (Jabine, 1993 and Fienberg et al., 1998), the Federal Committee on Statistical Methodology, 1994 and Duncan and Fienberg (1999).

Bibliography

Adam, N.R., Wortmann, J.C., 1989. Security-control methods for statistical databases: a comparative study. ACM Computing Surveys 21, 515–556.

Bethlehem, J.G., Keller, W.J., Pannekoek, J., 1990. Disclosure control of microdata. Journal of the American Statistical Association 85, 38–45.

Chowdhury, S.D., Duncan, G.T., Krishnan, R., Roehrig, S.F., Mukherjee, S., 1999. Disclosure detection in multivariate categorical databases: auditing confidentiality protection through two new matrix operators. Management Science 45, 1710–1723.

Cox, L.H., 1980. Suppression methodology and statistical disclosure control. Journal of the American Statistical Association 75, 377–385.

Cox, L.H., 1987. A constructive procedure for unbiased controlled rounding. Journal of the American Statistical Association 82, 38–45.

Cox, Lawrence H., Karr, Alan F., Kinney, Satkartar K., 2011. Risk-utility paradigms for statistical disclosure limitation: how to think, but not how to act. International Statistical Review 79 (2), 160–183.

Dalenius, T., 1988. Controlling Invasion of Privacy in Surveys. Department of Development and Research Statistics, Sweden.

Dalenius, T., Reiss, S.P., 1982. Data-swapping: a technique for disclosure control. Journal of Statistical Planning and Inference 6, 73–85.

Duncan, G.T., Fienberg, S.E., 1999. Obtaining information while preserving privacy: a Markov perturbation method for tabular data. Eurostat. In: Statistical Data Protection '98, Lisbon 1999. Office for Official Publications of the European Communities, Luxembourg, pp. 351–362.

Duncan, G.T., Kaufman, S., 1996. Who should manage information and privacy conflicts?: institutional design for third-party mechanisms. The International Journal of Conflict Management 7, 21–44.

Duncan, G.T., Lambert, D., 1986. Disclosure-limited data dissemination (with discussion). Journal of the American Statistical Association 81, 10–28.

Duncan, G.T., Lambert, D., 1989. The risk of disclosure of microdata. Journal of Business and Economic Statistics 7, 207–217.

Duncan, G.T., Pearson, R., 1991. Enhancing access to microdata while protecting confidentiality: prospects for the future (with discussion). Statistical Science 6, 219–239.

Duncan, G.T., Jabine, T.B., de Wolf, V.A., 1993. Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics. Panel on Confidentiality and Data Access, Committee on National Statistics, National Academy Press, Washington, DC.

Federal Committee on Statistical Methodology, 1994. Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology. US Office of Management and Budget, Washington, DC.

Fienberg, S.E., Makov, U.E., Steele, R.J., 1998. Disclosure limitation using perturbation and related methods for categorical data. Journal of Official Statistics 14, 347–360.

Fuller, W.A., 1993. Masking procedures for microdata disclosure limitation. Journal of Official Statistics 9, 383–406.

Greenberg, B., Zayatz, L., 1992. Strategies for measuring risk in public use microdata files. Statistica Neerlandica 46, 33–48.

Jabine, T., 1993. Statistical disclosure limitation practices of United States statistical agencies. Journal of Official Statistics 9, 427–454.

Kelley, J., Golden, B., Assad, A., 1990. Controlled rounding of tabular data. Operations Research 38, 760–772.

Lambert, D., 1993. Measures of disclosure risk and harm. Journal of Official Statistics 9, 313–331.

Matthews, Gregory J., Harel, Ofer, 2011. Data confidentiality: a review of methods for statistical disclosure limitation and methods for assessing privacy. Statistics Surveys 5, 1–29.

Miller, A.R., 1971. The Assault on Privacy: Computers, Data Banks and Dossiers. University of Michigan, Ann Arbor, MI.

Raghunathan, Trivellore E., Reiter, Jerome P., Rubin, Donald B., 2003. Multiple imputation for statistical disclosure limitation. Journal of Official Statistics 19 (1), 1–16.

Rubin, D.B., 1993. Satisfying confidentiality constraints through the use of synthetic multiply-imputed microdata. Journal of Official Statistics 9, 461–468.

Spruill, N.L., 1983. The confidentiality and analytic usefulness of masked business microdata. In: Proceedings of the Section on Survey Research Methods. American Statistical Association, pp. 602–607.

Willenborg, L., de Waal, T., 1996. Statistical Disclosure Control in Practice. Lecture Notes in Statistics 111. Springer, New York.

Winkler, W.E., 1998. Re-identification methods for evaluating the confidentiality of analytically valid microdata. Research in Official Statistics 1, 87–104.

Conflict and Conflict Resolution, Social Psychology of

Ronald J Fisher, School of International Service, American University, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D.G. Pruitt, volume 4, pp. 2531–2534, © 2001, Elsevier Ltd.

Abstract

Social psychology has made important contributions to the study of social conflict and to methods for the resolution of conflict between parties, particularly at the intergroup level. This article outlines perceptual and cognitive processes along with group and intergroup dynamics that feed conflict escalation and intractability. While augmenting the study of negotiation and mediation, social psychology has led the development of interactive conflict resolution, which engages members of conflicting parties in analysis, dialogue, and problem solving.

Introduction

Social psychology is both a subfield of sociology and psychology, and as an 'interdiscipline' is concerned with how the behavior of individuals is influenced by and in turn influences others in various social contexts. It focuses on how individual motives, attitudes, and actions are intertwined with those of others in social relationships and interactions, such as social conflict. Conflict can be defined as "a social situation involving perceived incompatibilities in goals or values between two or more parties, attempts by the parties to control each other, and antagonistic feelings by the parties toward each other" (Fisher, 1990: 6). Social conflict thus has three main components: differences between individuals or groups, attempts to influence each other, and negative feelings that develop along with this interaction. Conflict resolution is seen as the management of differences in a manner that is constructive and cooperative, and yields outcomes that are mutually satisfactory and have a long-term commitment from the parties. Thus, conflict analysis and resolution is compatible with the nature of social psychology, and therefore, as this new multidisciplinary field of endeavor developed in the social and behavioral sciences, social psychologists were at the forefront, for example, in founding the *Journal of Conflict Resolution* in 1957 at the University of Michigan.

The social-psychological approach to conflict analysis and resolution is rooted in three key elements (Fisher, 1990). The first draws on the phenomenological nature of the discipline itself, which emphasizes the importance of subjective experience in determining human beings' representations of reality and our responses to it. Hence, perceptions, attitudes, and values held by individuals and shared by groups in conflict are an important determinant of their behavior in relation to the conflict and the other party. This element also holds promise for conflict resolution, in that cognitions can change and the valuing of different positions and outcomes is malleable. The second element emphasizes the critical role of behavioral interaction among the parties in terms of how they orientate, communicate, respond, and negotiate with each other over their differences. Conflict interaction is an emergent phenomenon, and how the parties treat each other is equally important as the nature of their interaction in determining the course and the outcomes of the conflict. The third element assumes that social conflict can only be understood by giving attention to multiple levels of analysis through a systems approach. The level of the conflict itself, whether it be interpersonal, intergroup, or international needs to be identified, and each of the levels involved needs to be analyzed for its contribution to the conflict interaction. For example, individual perceptions and cognitions are fundamentally important, but in intergroup conflict these need to be understood in the context of the group and intergroup environments in which they arise and operate. This is because members of groups do not operate freely as individuals, but in terms of their group identity and norms, and in response to past interactions between the groups. Social behavior in conflicts is thus multiply determined, and a wide range of variables at multiple levels must be considered in analyzing causation and in approaching resolution.

In addressing social conflict, this treatment does not assume that the phenomenon is predominantly good or bad for humankind – it is simply real and omnipresent. How situations of difference are approached and managed by the parties will determine whether the process and outcomes are constructive or destructive, or a combination of these two. Social conflict has a functional side in that it is often a source of social change toward greater justice and equality, among individuals, groups, and societies. If parties enter into conflict with open, complex communication, a cooperative orientation, a degree of flexibility in preferences with firmness on basic requirements, and sense that the conflict is a shared problem, then the probability of constructive processes and outcomes is increased. If parties enter the zone of the conflict with closed, manipulative communication, a competitive orientation, rigidity in perceptions and positions, and a sense that the conflict is a war to be won, then the norm of reciprocity determines that destructive escalation will occur to a point of mutual debilitation. Of course, parties often approach conflict in the latter manner with the confidence that their influence strategies will prevail, and they will win at reasonable cost and attain their desired outcomes. Unfortunately, the world is replete with instances where this has not occurred and parties become locked in intractable situations that demonstrate the clear need for the field of conflict analysis and resolution.

This article will outline most of the major contributions that social psychology has made to conflict analysis and resolution. It will begin by describing some of the perceptual and cognitive

factors, particularly distortions and biases, which can occur in conflict interaction and escalation, followed by some of the group-level dynamics, which tend to exacerbate the prosecution of the conflict. After identifying the common methods used to address conflicts, greater consideration will be given to interactive small group methods, primarily problem-solving workshops, which flow from a social-psychological analysis and principles.

Conflict Analysis

This section describes many of the conceptual contributions that social psychology has made to the understanding of social conflict (Bar-Tal, 2011; Fisher, 1990; Pruitt and Kim, 2004). The assumption is that one must first understand a social phenomenon in order to develop methods to address it, in this case, to manage conflict in a way that minimizes its destructive potential, especially the use of violence.

Social Perception and Cognition

Social psychology is in large part the study of how people see the world and how they think about it, because these processes in large part determine how they respond to it. Social psychologists have invested considerable effort in the study of attitudes, that is, an individual's tendency to evaluate and respond to social objects in a consistently favorable or unfavorable way. Attitudes are typically seen as composed of three components: a cognitive one consisting of the beliefs about the object, an affective one comprising the emotions connected with the object, and a behavioral one, which includes the action tendencies or orientations toward the object. In the study of social conflict, the cognitive component has predominated, particularly through the concept of stereotype, which is a simplified and rigid set of beliefs, usually negative, that are attached to a disliked or enemy group. Stereotypes are common in intergroup relations, and at a low level of hostility can be somewhat innocuous misperceptions of out-group attributes, while at the same time capturing some distinctive out-group characteristics. However, at higher levels of conflict intensity, stereotypes become exaggerated and inflexible, can result in selective and distorted perceptions, and can be used to justify destructive behaviors ranging from discrimination to dehumanization and to genocide.

The concept of image, popular in the study of international relations, builds on the idea of stereotypes, and is also congruent with the definition of attitude given above. Enemy images are very similar to negative stereotypes and see the other party in conflict as having undesirable characteristics (untrustworthy, irrational), deserving of hostile emotions (suspicion, hatred), and requiring coercive behavioral responses (deterrence, punishment). An additive concept is that of mirror image, in which each party in conflict holds a negative enemy image toward the other along with a positive image of itself, a cognitive construction which is of course impossible in reality. Thus, one's own group or country is regarded as good, peaceful, and well-intentioned, while the other side is seen as evil, aggressive, and malevolent, much like the modal perceptions held in the United States and the Soviet Union during the Cold War. The danger with such mirror images is not only that they influence ongoing perceptions in the direction of selectivity and distortion, but also that they drive escalation of the conflict to more intense levels, because countering actions by the other are interpreted as due to hostile and aggressive (rather than defensive) intentions and indeed to the very nature of the enemy as being malevolent, if not evil.

Causal Attribution

Social psychology studies the process of causal attribution by which individuals make judgments about the reasons for their own and other people's behavior, and about how they then make inferences about the characteristics of others, which of course affects how they may respond. A basic distinction is made on whether behaviors are attributed to the personal or dispositional characteristics of others and the group to which they belong, or to the situation or environment in which they are operating. This relates to the first common error relevant to social conflict, the so-called fundamental attribution error, by which observers of others' actions have a tendency to attribute causation to the inherent characteristics of the other, rather than to the situation. In contrast, individuals tend to attribute their own behavior to situational or contextual factors. Thus, in a social conflict, there is a bias for actors to attribute the other's behavior to their personality or culture, while seeing their own behavior more as responses to situational cues, including the other's actions.

At the level of intergroup relations, this bias feeds into the so-called ultimate attribution error, in which undesirable behaviors by an out-group member are more likely to be attributed to the negative characteristics of the out-group (and the individual group member) in line with the attendant stereotype. However, positive behaviors by an out-group member tend to be attributed to situational determinants, that is, rationalized and explained away. In contrast, desirable actions by an in-group member are seen as due to the positive in-group (and member) characteristics, captured in a self-serving in-group stereotype, while undesirable behavior is attributed to situational factors. In essence, out-group members cannot win in terms of attributions made, while in-group members cannot lose. Although these errors in attribution may not be universal, they likely feed conflict escalation and also may become more extreme as intensity and hostility between the groups moves to higher levels.

Confirmatory Mechanisms

Attribution processes in situations of intergroup conflict thus operate to confirm and strengthen group stereotypes, but they are not the only mechanisms that operate in this direction. In general, attitudes and images once formed tend to be resistant to modification, even in the light of contradictory experiences. Attitudes perform some important functions for individuals, including bolstering one's self-esteem, expressing one's values, and maximizing the benefits one receives from social behavior. However, once formed they can be blinders that restrict perceptions of social reality through selectivity and distortion of incoming information as well as the recall of received information. In all cases, information that is congruent with

existing cognitive structures, including attitudes, self-concept, and group identity, is more likely to survive and even thrive in how people see, interpret, and remember the social world.

One powerful confirmatory mechanism is the self-fulfilling prophecy in which the expectations defined by stereotypes can lead the holder of the stereotype to treat the object of the stereotype in ways that lead the object to then behave in ways that confirm the stereotype. This effect was first studied in the classroom, where teachers' expectations about the capabilities of their students (some of which were bogus ones incorrectly supplied to them) were found to be related to subsequent student performance, ostensibly because the teachers treated the students in line with their expectations. In social interaction in a context of intergroup conflict, this effect can be brought forward by stereotyped expectancies that one group holds of another group. For example, if an out-group is perceived as untrustworthy, members of that group will likely be treated with caution and suspicion, which may signal them to behave in nontrusting and noncooperative ways, thus confirming the original expectation.

A powerful force in the confirmatory power of existing attitudes and images is found in the concept of cognitive dissonance, which is an unpleasant psychological tension created by inconsistency between different cognitive elements, such as between our attitude and the attitudes of important others, or between our beliefs and our actions. Individuals can attempt to reduce the inconsistency and the dissonance through a variety of actions, such as changing their beliefs or behavior, or seeking more information to support their existing beliefs, or explaining away the contradictory information. With strongly held beliefs, such as negative stereotypes, it is more likely that the individual will work to maintain consistency by screening out or reinterpreting contradictory information or experiences, thus maintaining consistency and preserving the stereotype.

Social Judgment and Decision-Making

The numerous biases and errors in gathering, interpreting, and recalling information transfer readily from individual judgments to group decision-making in situations of intergroup and international conflict. Thus, social psychology along with the other social sciences has raised serious concerns about the rational actor model of decision-making, which initially predominated in both organizational and international studies (Janis and Mann, 1977). The critique contends that decision-making often involves biased information processing, the use of simple heuristics or short cuts, a lack of objective criteria, and a limited search for options. These deficiencies in judgment and decision-making are seen to increase under conditions of stress and fatigue, which are typically present in situations of crisis.

Much of the work on social judgment is now captured in so-called prospect theory, which sees decision-making as involving choices among prospects or options that will yield different outcomes, which have different probabilities of occurrence. As opposed to the rational actor model, prospect theory proposes that individuals evaluate choices not in terms of their absolute value, but in terms of their gains and losses in comparison to the status quo. It also proposes that losses are overvalued in comparison to gains and thus are seen as more important. In decision-making, this bias leads to loss aversion and the taking of greater risks to avoid losses and thereby a tendency to stick with the status quo. When problems are framed in terms of losses, parties are thus more likely to take risks than when they are framed in terms of gains (Kahneman and Tversky, 2000). In application to conflict resolution, these propositions mean that parties are biased toward the status quo, and will see the losses they might incur through negotiating concessions as more important than the gains they might achieve. Hence, mutual dissatisfaction with potential outcomes in a peace agreement results in stalemate rather than settlement.

Another contribution of social psychology to the study of decision-making is the model of groupthink that describes the process by which an insulated and highly cohesive group engages in concurrence seeking to the extent that it overrides the realistic appraisal of alternatives, thus leading to poor decisions (Janis, 1982). Such a group overestimates its own capability and morality, exhibits closed mindedness in which stereotypes and rationalizations are prominent, and censors itself in supporting the initial direction provided by a powerful leader. As a result, the group shows incomplete information search, biased information processing, inadequate risk analysis, and a lack of contingency plans. Although the model and the research on which it is based have been criticized, it has also been supported and elaborated to stand as an important counterpoint to the rational actor model.

As conflicts escalate, a number of fundamental transformations occur in the issues, tactics, and motives of the parties (Pruitt and Kim, 2004). At the highest level of escalation in protracted conflicts, an insidious process known as entrapment becomes a driver of intractability. Entrapment is seen as a cognitive trap in which the parties become increasingly and unwittingly committed to destructive and costly courses of action that would not be predicted by the rational actor model. Thus, the parties expend increasing resources to prosecute the conflict, in part to justify the investments they have already made and hopefully to compensate for the losses they have already incurred. This phenomenon can be partly explained through the functioning of postdecision dissonance reduction, in which a chosen alternative is subsequently evaluated more positively in order to reduce the cognitive dissonance created by its earlier choice. Thus, parties lock into courses of action which do not add up in terms of costs and benefits, but which are cognitively consistent and also privilege losses over gains as predicted by prospect theory.

Another line of work on the social psychology of decision-making focuses on the cognitive or integrative complexity that leaders and other policymakers demonstrate in their thinking and communication about a conflict (Conway et al., 2001). Integrative complexity is demonstrated by a number of indicators, including the extent of search for new information, the rigidity or conditionality of perceptions, and the number of options being considered. These indicators can be grouped into overall measures of cognitive differentiation or the number of dimensions used in interpreting information and cognitive integration or the degree and nature of connections among different elements. Research supporting the role of integrative complexity demonstrates that it tends to decrease during international crises that lead to war and tends to increase

during periods of agreement among leaders. While some of this effect may simply be due to the rhetoric used in situations of conflict, it is likely that it tells us something about the cognitive functioning of policymakers engaged in international and other crises.

Ethnocentrism, Social Identity, and Social Dominance

The concept of ethnocentrism, developed in the disciplines of sociology and anthropology, has been incorporated as an important element of the social psychology of conflict analysis. Referring initially to a tendency of social groups to be ethnically centered in that they see themselves to be superior while holding other groups in contempt, ethnocentrism has been largely cast as a form of generalized prejudice toward all manner of out-groups. Ethnocentrism took a lead position in the development of realistic group conflict theory, which posits that real conflicts of interest cause intergroup conflict. Incompatible goals and competition over scarce resources are seen to result in real threat, which increases in-group solidarity and awareness of identity, causes hostility to the source of the threat, and increases ethnocentric reactions (LeVine and Campbell, 1972; Sherif, 1966). As ethnocentrism moves to higher levels, it is predicted that the cognitive distortions noted above increase in the direction of more extreme stereotypes, negative enemy images, and influential self-serving biases.

A more recent line of theorizing, social identity theory, puts less emphasis on real conflicts of interest and more on the processes of social categorization and identity formation in causing intergroup conflict (Tajfel, 1982). It appears that the mere fact of categorizing individuals into groups gives rise to a perception of heightened similarity within groups along with exaggerated differences among groups. Identity groups are typically defined in ethnic, racial, religious, cultural, or national terms and are the object of identification by their members in ways that constitute an important element of their social identity. It is posited that individuals strive to enhance their self-esteem, which in part derives from the social groups to which they belong, resulting in motivation to achieve and maintain a positive social identity. This leads to making favorable comparisons with other groups, in which one's own group is seen as distinctive and held in higher regard, while other groups may be the objects of derogation and discrimination. In this way, a different explanation is offered for ethnocentrism in comparison to realistic group conflict theory. At the same time, the two theories can be seen as complementary, in that both help explain the insidious power of group and intergroup dynamics to exacerbate conflicts of interest.

Social dominance theory augments the above two theories by affirming group differences in power and status while also explaining individual differences in out-group discrimination (Sidanius and Pratto, 1999). Individuals vary in their social dominance orientation, and those who are high on social dominance support ideologies and policies that legitimize hierarchies among groups, and individual and institutional discrimination in favor of more powerful groups in the hierarchy. Social dominance theory thus complements both realistic group conflict theory and social identity theory by adding a dimension of individual differences, much like the syndrome of authoritarianism, which can help explain the causation and maintenance of conflict between different groups in society.

Escalation and Intractability

Conflicts of interest and identity when not handled constructively have a built-in tendency to escalate, that is, to become more intense and hostile, to involve more issues and parties, and to engage the parties in more contentious and coercive strategies (Fisher, 1990; Pruitt and Kim, 2004). Escalation tends to feed on fear, threat, defensiveness, and reciprocity. Out of fear of losses, parties tend to threaten or carry out controlling or coercive actions to advance their goals, moves which are then reciprocated by the other party following a general law of human behavior: you give what you get. As intensity increases, all of the perceptual and cognitive distortions and errors become more pronounced. Enemy and mirror images move from a dimension of good versus bad to one of virtuous versus evil, while expectations of intransigent or destructive behavior are fulfilled, thus legitimating further attempts to threaten and control. Patterns of communication and interaction move from words to actions and to direct attacks designed to injure. Issues shift from a concern about interests to matters of identity and ultimately to a question of survival. As conflict escalates, the intent becomes not simply to reach one's goals but to eliminate the other party. In this process, the other becomes demonized and dehumanized so that all manner of actions are justified, often leading to a situation of mutual victimization. However, when one side has a clear power advantage, the victimization is unilateral and can result in extreme destructive outcomes as in mass killing and genocide.

The protracted nature of destructive conflicts in many spheres has led to growing interest in the phenomenon of intractability, by which conflicts are enduring or long lasting, remain at a high level of intensity with sporadic outbreaks of violence or other destructive behavior, and are also extremely difficult to deescalate and resolve. These conflicts may be based in a history of domination and injustice by one group over another, may involve very significant issues or resources at stake, and may engage controversial moral or ideological issues, which in a sense are not resolvable. Intractable conflicts often threaten group identities, well-being, or survival, and thus engage a high level of emotional intensity, which further fuel cognitive distortions and aggressive behaviors. At the same time, these conflicts are inescapable in that they involve members of the same family, entities in the same organization or community, or identity groups interspersed in the same territory and political system. Such conflicts present the greatest challenge to methods of conflict resolution, which clearly have to grow in complexity and effectiveness in order to reduce the human costs of destructive, protracted conflict (Coleman, 2011).

Conflict Resolution

Once social science understands an undesirable phenomenon such as destructive conflict, the next question is what practices and policies can we provide to ameliorate the problem. This is the challenge for the interdisciplinary field of conflict

resolution, which has grown up over the past several decades, and to which social psychology has made unique and important contributions. Social psychologists have provided ideas to both improve traditional methods of conflict management such as negotiation and mediation, and have led the way in helping to develop unique methods to address protracted, identity-based conflicts under the rubric of interactive conflict resolution (Fisher, 1997).

Conflict Management Styles, Negotiation, and Mediation

Conflict resolution refers to both the process of addressing difficult conflicts and the outcomes of the process, which must be mutually acceptable and beneficial and are self-sustaining and adjustable to changing conditions. As such, conflict resolution deals with both the objective issues in dispute and the subjective or human elements in minds and in the relationship of the parties. Conflict management is used more to connote dealing with continuing disputes over tangible interests within the context of a generally satisfactory relationship between the parties. In either case, the field prescribes that a cooperative yet firm approach is the best way to successful management or resolution. In this regard, social psychologists were the first to postulate that two dimensions of behavior were required to adequately represent different orientations or styles for addressing conflict: assertiveness or concern for self and cooperativeness or concern for the other party (Blake et al., 1964). Placing these two orientations as the axes of a two-dimensional space yields five distinct conflict management styles: competing, avoiding, accommodating, compromising, and collaborating. While these styles were initially developed to help understand and deal more effectively with interpersonal conflict, they also have relevance to the intergroup and international levels. Each style is seen as having its advantages and disadvantages and each is more applicable in certain situations of conflict and relationship, thus leading to the idea that styles can be adapted to different conflict situations for maximal effectiveness.

Moving to the method of negotiation, by which parties interact to develop agreements that regulate their behavior and typically divide resources, we also see the imprint of the firm yet cooperative approach. Traditional hard bargaining exemplifies a firm approach to the point of relentless competition, whereas social psychologists have contributed to a cooperative and integrative approach in which a party works to meet some of its own interests but is also amenable to the other party meeting some of its interests (Pruitt and Kim, 2004). This results in mutual benefits as well as mutual concessions with outcomes that are more satisfactory and more sustainable over time. The term 'win–win' was coined by Blake et al. (1964) to capture this phenomenon. Social psychologists have been at the forefront of developing models, propositions, and techniques for negotiation that maximize the potential for integrative outcomes.

Mediation is a method of third-party intervention typically applied in conflicts where interaction or negotiation by the parties has not been able to achieve either a one-sided or mutually acceptable outcome. Mediation is a voluntary, nonbinding intervention designed to achieve a negotiated agreement on the objective issues in dispute, and is typically carried out by an impartial, credible, and trusted intermediary. Social psychologists have contributed to the study and practice of mediation in general (Kressel et al., 1989), and have also emphasized the distinction between pure mediation and power mediation, in which the basic functions of persuasion and the creation of options are augmented by the application of leverage (the use of rewards and punishments) by the third party in order to manipulate the parties toward agreement. The use of power mediation, particularly at the international level, continues to raise theoretical, practical, and ethical questions that require continuing attention.

Interactive Conflict Resolution

The failure of traditional methods of conflict management to achieve mutually acceptable agreements and peaceful, sustainable relationships between parties locked in destructive and intractable conflict led the field of conflict resolution to develop alternative and innovative methods. Many of these attempts can be captured under the rubric of interactive conflict resolution, which refers to third-party facilitated, small group discussions involving dialogue, analysis, and problem solving that bring influential participants from parties engaged in destructive conflict together in order to mutually diagnose their conflict and to develop jointly acceptable options for resolving it (Fisher, 1997, 2005). Interparty dialogue is an initial form of this approach, but the exemplar method is that of the so-called problem-solving workshop, which is convened and facilitated by a third-party team of skilled and knowledgeable consultants. As opposed to mediation, which seeks agreement on substantive matters and tries to work around emotional issues, the third-party-consultants in problem-solving workshops address both objective aspects (territory, form of government) and subjective elements (perceptions, attitudes, and relationship issues such as mistrust) in the conflict. Thus, the role requires human relations and small group skills for understanding, communication, and problem solving as well as knowledge from social science about destructive conflict and its resolution. Interactive conflict resolution can also include other methods, such as conflict-resolution training, which brings antagonists into the same educational setting to learn about concepts for analyzing conflict and methods for addressing it; cross-conflict teams, who work cooperatively on projects in peacebuilding; and reconciliation, which facilitates postconflict interactions to reestablish harmony and cooperation.

The rationale for interactive conflict resolution is based squarely in the social-psychological approach to and analysis of conflict provided above. Thus, it is assumed that conflicts are a mix of objective and subjective factors, both of which must be addressed for resolution, and that subjective elements play a larger role as conflict escalates, thus increasing the relevance and applicability of interactive methods. It is also posited that constructive and authentic face-to-face interaction between the parties or their representatives is necessary to confront and overcome the distorted and inaccurate perceptions and cognitions, and to change the competitive, adversarial orientations and interactions of the parties. Given the system orientation of the social-psychological approach, it is also clear that any individual changes in cognitions and orientations that occur among participants must be

transferred to both policymaking and the public discourse among the conflicting parties for deescalation and resolution.

The Problem-Solving Workshop: Characteristics and Goals

Problem-solving workshops are often appealing to parties locked in destructive conflict because it provides an informal, quiet, neutral, low risk, and noncommittal forum compared to official avenues such as negotiation, mediation, or arbitration. In this way, the method has similarities to back channel or secret negotiations between parties, but it must be emphasized that the workshop does not engage parties in negotiation. In contrast, it is designed to explore the issues in the conflict, which must eventually be negotiated and agreed upon; thus, such interventions have often served an important pre-negotiation function. Workshops can assist the parties in understanding their different and common interests, and in distinguishing their positions on issues from their deeper interests, values, and needs. On the subjective side, such encounters can help clear up misperceptions, increase understanding about the conflict dynamics, and build trust, reassurance, and cooperation between the parties. Alongside negotiations, workshops can be used to analyze particularly thorny issues or to provide exploration of anticipated future issues. In the postnegotiation phase, the method can be useful for dealing with implementation difficulties associated with agreements and for building reconciliation between the former antagonists. Thus, the potential complementarity of problem-solving workshops to official interactions is evident in all phases of conflict resolution.

In terms of implementation, the problem-solving workshop is a small group discussion method with only five or so participants from each party (usually only two) and from the third-party team, although larger workshops can be designed wherein a number of problem-solving groups engage in parallel discussions, the results of which are then brought back to plenary sessions. Workshops typically take place in neutral, secluded, and residential settings and last for approximately 4 to 6 days. Participants are unofficial yet influential opinion leaders or advisors who are well-connected to policymakers and/or the public discourse on the conflict through roles such as journalist, author, or academic. Often participants are at the midlevel of their societies, so that they can connect to the leadership above, but also to the wider society below in term of having access and influence. Typically, a rough agenda for discussion is provided, which moves from exchanging experiences and perspectives on the conflict with attention to underlying interests and concerns (particularly needs and fears), to a joint diagnosis of the conflict, and to the mutual creation of options for deescalation and resolution. The role of the third-party consultants is enacted through a number of strategic functions: increasing problem-solving motivation, improving communication, regulating the interaction, and helping to diagnose the dynamics of the conflict. To maintain the quality and respectfulness of the discussion, the participants are held to a number of ground rules or norms of interaction that include items such as no personal attacks, the sharing of air time, analytical discussion rather than debate or argument, and confidentiality in the sense of no attribution of statements to individuals following the workshop. Thus, the problem-solving workshop provides a rare and very different forum for discussion from that to which the parties and participants have become accustomed to in the hostile and adversarial field of the conflict.

As implied above, the problem-solving workshop has multiple goals at several levels of analysis. For individual participants, it is expected that they will experience increased understanding of the conflict and the other party, and that their attitudes to each other will undergo increased complexity and positiveness. They should also develop an increased sense of trust and cooperation toward the other party, and will then work to transfer these attitudinal and relationship changes back to their respective constituencies. This may include contributing new realizations and ideas to negotiations and policymaking along with attempting to positively affect the political discourse and atmosphere in each constituency in ways that support a sustainable resolution. Given that the large gap between a small number of participants and their groups or societies, it is clear that the transfer process is the greatest challenge facing the method; nonetheless, an increasing amount of evidence shows that it is not an insurmountable one (Fisher, 2005).

Problem-Solving Workshops: Rationale in the Intergroup Contact Hypothesis

Some of the rationale supporting the problem-solving workshop can be found at the individual- and group-levels of analysis, but the most powerful explanations for its utility in resolving intergroup and international conflict are to be found in the so-called contact hypothesis (Hewstone and Brown, 1986). At the individual level, the workshop interaction provides new and challenging information about the other party that contradicts the erroneous and simplistic misperceptions and stereotypes common in intergroup conflict. This produces a more differentiated and realistic image of the other group, and also creates cognitive dissonance in the participants that can be reduced by changing their attitudes and orientations in more positive and cooperative directions. In addition, common attribution errors may be identified and countered in the analysis, leading to a decrease in mutual blaming and victimization. In relation to group identity, participants may come to realize that accommodating to the other's need for identity does not mean denying one's own identity, and that recategorization of identities is possible in that overarching dual or multiple identities can be formed, such as different group identities being accommodated in the same national identity.

The intergroup contact hypothesis has been one of the primary contributions of social psychology to intergroup relations, and states that contact and interaction between members of different and conflicting groups can result in positive attitude and relationship change, if it occurs under specified facilitative conditions. These include a high degree of acquaintance potential so that participants get to know each other are persons, equal status in terms of educational and professional accomplishments, social norms supporting open and respectful interaction, a cooperative task and

reward structure that has participants working and achieving together, and a selection of competent and responsible individuals as participants. It is immediately apparent that the design of the problem-solving workshop is geared to implement these facilitative conditions. However, it is essential that the social identities of the participants remain salient in the interaction, so that as participants build bonds of interpersonal attraction they also regard members of the other group as typical rather than exceptional. Thus, any attitude changes should generalize to the other group at large.

Problem-Solving Workshops: Transfer and Effectiveness

The systems approach requires that the workshop method elucidate and evaluate how changes in individual participants' attitudes, understandings, and orientations might be translated into changes in policymaking at the elite level and in the political discourse and atmosphere at the societal level. Transfer is identified as the process by which individual effects and group products (principles, proposals) are moved from the unofficial intervention to the official and public domains (Fisher, 2005). At both the elite and societal levels, effective transfer would mean that a perceptual shift takes place in both parties that sees the other party as responsible and approachable, and thus as a valid partner for negotiations to address the conflict. This also involves a shift from a competitive to a cooperative orientation that supports nonviolent bilateral interaction to manage the differences that have led to the conflict and its escalation.

Most of the evidence supporting transfer effects is anecdotal and provided mainly by the consultants who carried out the interventions. In some cases, research studies of specific interventions have used interviews and other methods to collect information from participants and interveners on the perceived effects of workshops. Recently, the comparative analysis of multiple cases of intervention has been carried out to produce more valid and generalizable results. These studies taken together generally support the theory of practice that is spelled out above and that has been implemented over the past five decades in all regions of the world to address violent ethnopolitical conflicts. When focused mainly at the elite level, the research indicates that the targets of transfer have been decision-makers, policymakers and negotiators, and that the most common objective has been to provide a prenegotiation function to help initiate and inform official talks. Transfer has taken many forms from cognitive changes to creative ideas, and from proposals to increased connections across conflict lines and also between unofficial and official actors within the same party. In a number of difficult cases, it is apparent that problem-solving workshops have made important contributions to successful peace processes. These contributions have usually been in line with a contingency model of third-party intervention, in which methods are matched to the level of escalation and then sequenced to deescalate and resolve the conflict (Fisher, 1990). Regardless of some success, it is also clear that escalated, identity-based conflicts with a high degree of power asymmetry are extremely resistant to both unofficial and official attempts at resolution.

Conclusion

The social psychology of conflict analysis and resolution has provided insight into the causation and escalation of destructive conflict, particularly through the identification of perceptual, cognitive, and interactive processes that tend to exacerbate incompatibilities. A variety of structures and biases in people's minds attend conflict and its escalation in ways that distort social reality and drive adversarial and coercive strategies. Group-level factors amplify and augment individual cognitions leading to rigidity, ethnocentrism, and intractability. In addition to making contributions to the understanding and effectiveness of negotiation and mediation, social psychologists have led the development of small-group interactive methods for analysis and problem solving, which offer an innovative and unique possibility for humankind toward the resolution of complex and protracted conflicts. The problem-solving workshop is the prototype of interactive conflict resolution that has been applied to escalated identity conflicts with some degree of success in terms of contributions to peace processes.

See also: Conflict Mediation; Conflict and Consensus; Conflict and Socioemotional Development; Ethnic Conflicts; Peace; Territorial Conflicts; Urban Conflicts.

Bibliography

Bar-Tal, D. (Ed.), 2011. Intergroup Conflicts and Their Resolution: A Social-Psychological Perspective. Psychology Press, New York.
Blake, R.R., Shepard, H.A., Mouton, J.S., 1964. Managing Intergroup Conflict in Industry. Gulf, Houston, TX.
Coleman, P.T., 2011. The Five Percent: Finding Solutions to Seemingly Impossible Conflicts. Perseus Books Group, New York.
Conway, L.G., Suedfeld, P., Tetlock, P.E., 2001. Integrative complexity and political decisions that lead to war or peace. In: Christie, D.J., Wagner, R.V., Winter, D.D.N. (Eds.), Peace, Conflict, and Violence: Peace Psychology for the 21st Century. Prentice Hall, Upper Saddle River, NJ, pp. 66–75.
Deutsch, M., 1973. The Resolution of Conflict. Constructive and Destructive Processes. Yale University Press, New Haven, CT.
Deutsch, M., Coleman, P.T., Marcus, E.C. (Eds.), 2006. The Handbook of Conflict Resolution: Theory and Practice, second ed. Jossey-Bass, San Francisco, CA.
Fisher, R.J., 1990. The Social Psychology of Intergroup and International Conflict Resolution. Springer-Verlag, New York.
Fisher, R.J., 1997. Interactive Conflict Resolution. Syracuse University Press, Syracuse, NY.
Fisher, R.J., 2005. Paving the Way: Contributions of Interactive Conflict Resolution to Peacemaking. Lexington Books, Lanham, MD.
Hewstone, M., Brown, R. (Eds.), 1986. Contact and Conflict in Intergroup Encounters. Blackwell, Oxford.
Janis, I.L., 1982. Groupthink, second ed. Houghton-Mifflin, Boston, MA.
Janis, I.L., Mann, L., 1977. Decision Making. Free Press, New York.
Kahneman, D., Tversky, A. (Eds.), 2000. Choices, Values, and Frames. Cambridge University Press, New York.
Kelman, H.C., Fisher, R.J., 2003. Conflict analysis and resolution. In: Sears, D.O., Huddy, L., Jervis, R. (Eds.), Oxford Handbook of Political Psychology. Oxford University Press, Oxford, pp. 315–353.
Kressel, K., Pruitt, D., Associates (Eds.), 1989. Mediation Research: The Process and Effectiveness of Third-Party Intervention. Jossey-Bass, San Francisco, CA.
LeVine, R.A., Campbell, D.T., 1972. Ethnocentrism: Theories of Conflict, Ethnic Attitudes and Group Behavior. Wiley, New York.
Pruitt, D.G., Kim, S.H., 2004. Social Conflict: Escalation, Stalemate, and Settlement, third ed. McGraw-Hill, New York.
Sherif, M., 1966. In Common Predicament: Social Psychology of Intergroup Conflict and Cooperation. Houghton-Mifflin, Boston.
Sidanius, J., Pratto, F., 1999. Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression. Cambridge University Press, New York.
Tajfel, H. (Ed.), 1982. Social Identity and Intergroup Relations. Cambridge University Press, Cambridge, UK.

Conflict and Consensus

Katy Hayward, School of Sociology, Social Policy and Social Work, Queen's University Belfast, Belfast, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Citrin, volume 4, pp. 2547–2550, © 2001, Elsevier Ltd.

Abstract

There are widely differing conceptions as to whether healthy social relations are, in essence, conflictual or consensual; such differences give rise to different approaches to finding peace and managing power. This article outlines the two broad schools of thought from conflict theory (in which society functions through competition) and consensus theory (which sees society developing through cooperation). It outlines the middle ground between them, as found by pluralism and agonism, before considering the ways in which assumptions vis-à-vis conflict and consensus are reflected in different models of democratic system and, in particular, different priorities for post-conflict recovery.

Introduction

Social and behavioral science has traditionally been dominated by the presumption that conflict and consensus are oppositional: conflict arises from incompatibility and consensus arises from compatibility. This is reflected in two differing schools of thought as to the fundamental nature of society and collective decision-making within it. One can be broadly described as 'conflict theory,' whose proponents view all social action largely within the terms of some form of antagonism or imbalance. The other we can categorize here as 'consensus theory,' which sees progress in society as only being possible through negotiation and cooperation. This article outlines the key features of each approach and the insights they offer before turning to examine different interpretations as to the relationship between consensus and peace. The article's conclusion draws on a range of disciplines in its argument that, far from being polar opposites, conflict and consensus are two sides of the same coin: both conflict and consensus are equally essential to the functioning of society and group relations within it.

Nature of Conflict

Conflict is more than mere difference; it arises from a direct clash in the values or goals of protagonists. The presence of conflict thus infers none other than a win/lose outcome, in which one set of values or goals takes precedence over others. Power is at the heart of this dynamic; conflict is a struggle for power, i.e., the capacity to enact values or goals that annul those of others. And the process of articulating these values and goals among groups thus becomes, by necessity and definition, a political one. This may be seen in all forms of conflict in society. Class conflict, for example, centers on a fundamental inequality of power that spirals out from differential access to resources. Racial or ethnic conflict arises from a challenge to the values and goals of a dominant group that legitimates, and indeed benefits from, the subjugation of other racial or ethnic groups. Gender conflict also, in its simplest terms, is about an inequality of power as a consequence of the prevailing values and goals. Conflict in all these cases arises from group recognition of such a power imbalance and, in some cases, of collective movement to redress it. Into this mix come leadership, ideology, and mass opinion: all of which, again, indicate the centrality of power.

Conflict Theory

At the core of conflict theory – in its myriad of guises – is the belief that power (or, more specifically, an imbalance of power) is fundamental to social relations. Social interaction is thus, in essence, a struggle for control. Further to this, the decisions, identities, and perceptions arising from social exchange are imbued with conflict: 'this way, not that,' 'us not them,' 'mine not yours.' This merging of contradiction with power results in fundamental inequalities permeating throughout all society.

The major conflict theories center upon the most predominant forms of inequality. Marxism, for example, is founded on the premise that material inequality is the most significant of all power imbalances in society. Thus, struggle between classes (the laborers and the capitalists, as Marx (1867) would have it) is the overarching trope of western society, permeating into all forms of political, social, and economic behavior and outcomes. Where Marxian analysis differs from other conflict theories, however, is in its assertion that pervasive social conflict is primarily a by-product of the dominance of capitalism and thus will be eliminated when a more equitable form of exchange supplants the capitalist system.

The use of the notion of 'conflict' to explain the development of modern society reemerged in the 1950s, particularly in sociology departments in American universities. A sociologist, Wright Mills (1956), who applied conflict theory to advance understanding of the effects of social structures on individuals in the modern age, argued that the most influential public institutions are founded, not on agreement, but on conflict and competition. Their role ever after, he claimed, has been to perpetuate stark inequalities in resources and influence, with the result that social norms, values, interests, and decisions thus all come to be shaped by a powerful elite. Social order is thus ensured by coercion applied, through various means, by those with power. Society is consequently stratified, with access to advantage tightly restricted to groups already enjoying social privilege.

Whilst Mills shared Marx's negative view of conflict, other theorists have argued that the pervasiveness of conflict at every level of society makes it, simply, essential to its functioning. A contemporary of Mills, Lewis Coser (1956), claimed that conflict must be recognized as a fundamental element of human relationships. He argued that the struggle for resources and recognition behind conflict serves the vital function of stimulating social innovation and change. Coser built on the work of Simmel (see below), particularly in his focus on the role of conflict in the creation and maintenance of group identities. Conflict acts as a means of socialization, through which mutual antagonism can enhance group solidarity and preserve social stratification. From this perspective, Coser critiqued overtly stable social relationships, rigid social structures, and processes of centralization for making violence and dissent more likely and more intense. Instead, he argued, social conflict itself should be tolerated as naturally functioning to raise awareness of common norms, instigate conformity to laws, and, over time, to generate new social structures. Such positive analysis of the role of conflict was extended in the next generation of US-based sociologists, among whom Randall Collins (1975) made the contribution of further elaborating the links between the macro-context and the micro-level of personal interaction and behavior. The constant struggle for status of each individual – and the conflict this entails – only makes sense if one appreciates the stratified environment in which (s)he lives. The conclusion he drew from this was that, whilst inequality and conflict predominate in modern society, competition and control are the drivers of stability and progress within it.

Consensus Theory

A contrasting approach to understanding social and political relations is to emphasize the vital role of consensus in society. Durkheim (1893) argued that the most interesting aspect of society is not its various dysfunctional elements but, rather, the fact that it functions at all – something that he attributes to the primacy of consensus in successful, healthy social relations. Social cohesion is founded on the interdependence of peoples, forged by social ties that range from kinship to the exchange of services. This functionalist interpretation sees all parts of society as working to uphold its solidarity and to minimize instability. Societies survive, Durkheim argued, by commitment to a common set of beliefs and practices – a form of 'collective conscience' thus functions as the bedrock of social order. Durkheim claimed that this solidarity was created in a 'mechanical' way in traditional societies, using rules to emphasize homogeneity and conformity; in more advanced, differentiated societies, he argued, rules instead regulate differences, necessitating a more 'organic' form of solidarity, or common purpose, between citizens. Durkheim's thesis on solidarity was advanced by Finnish sociologist Erik Allardt (1970) in order to explain different reactions to modernization in the mid-twentieth century. He observed that, in order to function optimally, a modern society needed to have a highly developed division of labor together with a high degree of tolerance and pluralism. A weakness in Communist industrial societies, he therefore argued, lay not in the extent of the division of labor but in their imposition of conformity.

That said, theorists in this tradition have conceded that consensus does not naturally occur but is sometimes founded – and sustained – through coercion (Shils, 1972). Yet a process of socialization in modern societies modifies the behavior of individuals as they learn how to conform, or at least what is 'acceptable,' to the society in which they live. Interaction between individuals and groups can hence be premised on an understanding of how each can be reasonably expected to behave, i.e., the role each will perform (Parsons, 1951).

The nature of consensus changes with the nature of social communication and interaction. In the digital age, consensus may be achieved (or at least sought) through social media and online networks, for example, as much as through persuasive political rhetoric or interpersonal influence. Just as Simmel (1955[1908]) noted the effects of industrial capitalism and modernization on the nature of social organization, solidarity, and choice, so others, a century on, have sought to demonstrate the effects of the new information society on social bonds and relationships (Castells, 1996; Bauman, 2006). Even though the nature and style of communication between individuals has radically changed, however, their interdependence (or, to put it differently, their lack of freedom of choice) remains as strong as ever.

It is important to note at this point that consensus differs from consent in that the latter indicates that each person concerned offers his or her agreement, whereas there is an implicit assumption of majoritarianism in the concept of 'consensus,' given that only 'general' assent need be given. One might say that, rather than explicit and unanimous agreement as to the destination, consensus requires only shared sentiment as to the intended direction of travel.

Such consensual sentiment is maintained by a commonality that is underpinned by shared attachment to core sociopolitical institutions and acceptance of the norms and laws they promulgate. This is not to say that interests and identities do not diverge (this is inevitable in a plural society), but in cases of conflict between groups or individuals, adjudication takes place with reference to these 'common' interests and identities (as defined by laws and rules) and thus consensus ultimately prevails.

Given the close interdependence of society, social and individual needs are best addressed through cooperation rather than competition (which would merely risk generating new problems and divides). If shared norms, values, and institutions are the essence of a healthy society, consensus theorists conclude, conflict is almost pathological.

Pluralism

A tentative 'bridge' between conflict theory and consensus theory emerged in the mid-twentieth century in the form of pluralism. Pluralism recognizes that society is fragmented into diverse groups that compete for resources and have conflicting interests but acknowledges this does not automatically result in nontraversable gulfs crisscrossing society. Instead, Dahrendorf (1959) argued, advanced industrial society has found ways of

finding agreement between these social groups, despite their conflicting interests. Moreover, stability (and change) is realizable through institutionalizing this conflict, such as in the role of trade unions in processes of arbitration with employers. Dahrendorf thus allows for a functionalist emphasis on the role of social institutions whilst maintaining a focus on the differential levels of authority in any society and their effects on sustaining conflict.

Dahl (1967) brought pluralism into political analysis with his vision of society as a mass of competing interests. He argued that inequitable power and control was behind the construction of identities, capacities, and concerns of social groups. Social order is thus only achieved by finding a natural equilibrium among society's component parts, with the state acting as an 'honest broker' as groups battle it out for influence and resources. And the nature of this struggle is neither overt conflict nor consensus, but endless bargaining and compromise.

Indeed, the politics of compromise has been advocated (by Bellamy, 1999) as the logical response to the plurality of identities, values, cultures, practices etc. in contemporary society. He argues that the sheer complexity of the sources and ties of social plurality today makes compromise a necessity, in that the dialogue it allows between incommensurable demands is itself a process to be welcomed and facilitated in liberal democracy.

Consensus versus Conflict?

The matter of what should be set in place after violent conflict remains, ironically, contentious in both theory and practice. The perennial debate as to whether the eradication of conflict is either possible or desirable is to the fore here. On the one hand some claim that the resolution of conflict has to be at least the aspiration of peacemakers. Galtung (1969), for example, identifies the structural and cultural violence (such as poverty, oppression, or corruption) behind conflict as something to be challenged by the goal of 'positive peace.' Others argue that conflict is a necessary element of democracy and social exchange. For example, Foucault (2003) inverted Clausewitz' dictum (on war as politics by other means) to define politics as war waged nonviolently by the state. Peace, thus, entails some form of conflictual power struggle. This is, to some degree, in line with the tradition of agonism, which seeks to acknowledge and foster the positive benefits of conflict for democratic society (Schaap, 2009).

Yet there is common ground to be found between these standpoints because conflict and consensus are not mutually exclusive. First, the traditional 'criss-cross' theory of conflict prevention assumes that the intersection of group identities and lines of transmission reduce the risk of violence (Galtung, 1966); but for conflict to be present at all, there needs to be at least a minimal degree of communication and connection between the competing parties. Second, all but few agree that deep differences among groups should be expressed politically rather than violently. But this gives rise to the problems of determining how best to give political voice to the groups concerned and the basis on which the power is divided between them.

Power-Sharing

Different conceptions as to the nature of society result in different constructions of the democratic state. The assumption in conflict theory that democracy centers on a constant struggle for dominance between divisions in society views adversarial politics as the most effective mode of decision-making. On the other hand, democratic institutions founded on the principle of consensus seek to incorporate a wide spectrum of opinions and to generate agreement rather than mere decisions. Consensus democracies, therefore, generally need a different electoral system to that in majoritarian systems, one which allows for proportional representation of the groups that constitute the population. This in turn requires greater political mobilization among a wide range of social clusters than in other forms of democracy.

The type of democratic institutions put in place after violent conflict are often intended to build-in agreement to executive and legislative decision-making and can thus be modeled on a form of consensus democracy. Consociationalism, a system originally conceived from the case of Dutch democracy as one proven capable of creating consensus across deep social cleavages, has been advocated as a model for power-sharing in other divided societies (Lijphart, 1977). Proponents of the consociational template for democracy after conflict argue that it guarantees all cleavage groups a proportionate input in policy and executive decision-making and that it minimizes the potential for systemic discrimination (Noel, 2005). Key features of consociationalism that lend themselves to finding post-conflict stability include executive power-sharing, (segmental) autonomy, commitment to proportionality, and veto rights (McGarry and O'Leary, 2006). Cases of the application of consociationalism, or a version of it, after conflict include the system of government in Bosnia-Herzegovina established by the 1995 Dayton Agreement and Northern Ireland's devolved executive after the 1998 Good Friday (Belfast) Agreement.

Consociationalism has had many critics, however, many of whom point to the elite focus of the model, in particular its reliance on power-holding elites to bring cooperation to the heart of the democratic institutions and for their lead to be willingly followed. Problems can arise for example, if the political elites are unable to bring their constituent supporters with them (as in the first years of the Northern Ireland executive), if deep inequalities between social groups remain unameliorated (as in Lebanon), or if the consociational constitution is repressively prescriptive (as in Fiji).

An alternative model of democracy after conflict is centripetal, i.e., one which seeks to concentrate political momentum toward the center ground between parties rather than allowing it to fester at the extremes. This model follows Horowitz (1985), who argued that intergroup cooperation ought to be made a condition for electoral success rather than, as with consociationalism, seen as a hoped-for outcome of power-sharing. Centripetalism also seeks to avoid the reification of ethnic identities that is seen as a danger in consociationalism by manufacturing incentives for interethnic and centrist politics (Reilly, 2001).

Peacebuilding with Consensus

Strong criticisms of both consociationalism and centripetalism are put forward by integrationists, who argue that such theories underestimate both the unifying potential of the state and the peacebuilding capacity of nonelite actors. The importance of local ownership of peace processes requires decision-making mechanisms that reach through to the grassroots. Donais (2012), for example, goes so far as to say that peacebuilding should be pinned to the concept of consensus-building, given its insinuation of inclusivity. Only by seeking consensus at all levels of the conflict-ridden society and at all stages of a peace process, it is implied, can that process be truly 'owned' and embedded and minimize the risks of further alienation or disaffection from the political sphere (Richmond, 2012). This approach is posed as an alternative to the dominant one of liberal peacebuilding, which is primarily a political, top-down process in which international support can be seen as more important than local endorsement. Consensus is thus coming, in twenty-first century peacebuilding, to be predominantly interpreted as a requirement to recognize and engage with the needs – and the wisdom – of the local population.

Conflict, Consensus, and Crisis

In the globalized world of the twenty-first century, there are new modes of conflict and consensus in developed societies. These came together in the case of the Occupy movement, which attempted to address stark extremes of social and economic inequality through new mechanisms of consensus-building. According to the slogan of the movement, 99% of the mass population was counterpoised against the powerful elite of 1%. As prime locations in international city centers in 2011–12 became occupied, the movement embodied innovation in finding consensus. In small urban protest occupations it used ancient means of human voices and coordinated hand signals (see Figure 1); via the internet, it used new forms of social media to advertise activities, disseminate ideas, and galvanize support locally and internationally. Thus, in the context of global economic crisis, social conflict reached new extremes and social consensus was sought by novel means.

The Occupy movement itself embodied differing conceptions as to whether contemporary social relations are, in essence, conflictual or consensual. As this article has outlined, such varying interpretations of the nature of society give rise to different approaches to the greatest and most enduring collective challenges, such as those of managing power and minimizing violence. Such assumptions also inform how we understand democracy to function and society to develop. But this article has also noted that recent scholarly contributions to the debate have sought to identify a middle ground between these positions, acknowledging not just the necessity of 'both' competition and cooperation in healthy social relations but also the possible complementarity between them. Conflict and consensus, it must be concluded, remain equally essential to contemporary society.

Figure 1 Hand signals used by the occupy movement to facilitate consensus negotiation among protestors. These were widely disseminated via the internet. Source: http://occuprint.org/Posters/HandSignals.

See also: Agonism; Conflict Mediation; Conflict and Conflict Resolution, Social Psychology of; Conflict: Organizational; Peace Processes; Peace; Social Protest.

Bibliography

Allardt, E., 1970. Types of protests and alienation. In: Allardt, E., Rokkan, S. (Eds.), Mass Politics: Studies in Political Sociology. Free Press, New York, pp. 45–63.
Bauman, Z., 2006. Liquid Times: Living in an Age of Uncertainty. Polity, Cambridge.
Bellamy, R., 1999. Liberalism and Pluralism: Towards a Politics of Compromise. Routledge, London.
Castells, M., 1996. The Rise of the Network Society, the Information Age: Economy, Society and Culture, vol. I. Oxford, Blackwell.
Collins, R., 1975. Conflict Sociology. Academic Press, New York.
Coser, L., 1956. The Functions of Social Conflict. Free Press, New York.
Dahl, R., 1967. Pluralist Democracy in the United States: Conflict and Consent. Rand McNally, Chicago.
Dahrendorf, R., 1959. Class and Class Conflict in Industrial Society. Stanford University Press, Stanford, CA.
Donais, T., 2012. Peacebuilding and Local Ownership: Post-conflict Consensus-Building. Routledge, London.
Durkheim, E., 1893. De la division du travail social: étude sur l'organisation des sociétés supérieures. Ancienne librairie Germer Bailliére et Cie, Paris.
Foucault, M., 2003. In: Bertani, M., Fontana, A. (Eds.), 'Society Must Be Defended,' Lectures at the Collège de France, 1975–1976. (D. Macey, Trans.). Picador, New York.
Galtung, J., 1966. A structural theory of aggression. Journal of Peace Research 1 (2), 95–119.
Galtung, J., 1969. Violence, peace and peace research. Journal of Peace Research 6 (3), 167–191.
Horowitz, D.L., 1985. Ethnic Groups in Conflict. University of California Press, Berkeley, CA.
Lijphart, A., 1977. Democracy in Plural Societies: A Comparative Exploration. Yale University Press, New Haven, CT.
McGarry, J., O'Leary, B., 2006. Consociational theory, Northern Ireland's conflict, and its agreement. Government and Opposition 41 (1), 43–63.

Marx, K., 1990[1867]. Capital: A Critique of Political Economy, vol. 1. (Introduced by E. Mandel; B. Fowkes, Trans.). Penguin, London.
Mills, C.W., 1956. The Power Elite. Oxford University Press, New York.
Noel, S.J.R. (Ed.), 2005. From Power Sharing to Democracy: Post-conflict Institutions in Ethnically Divided Societies. McGill-Queen's University Press, Toronto.
Parsons, T., 1951. The Social System. Routledge and Kegan Paul, London.
Reilly, B., 2001. Democracy in Divided Societies: Electoral Engineering for Conflict Management. Cambridge University Press, Cambridge.
Richmond, O., 2012. A Post-liberal Peace. Routledge, London.
Schaap, A., 2009. Law and Agonistic Politics. Ashgate, Farnham.
Shils, E., 1972. The Constitution of Society. University of Chicago Press, Chicago.
Simmel, G., 1955[1908]. Conflict and the Web of Group Affiliations. Free Press, Glencoe, IL.

Conflict and Socioemotional Development

Brett Laursen and Amy C Hartl, Florida Atlantic University, Boca Raton, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Conflict is inherent in growth, and human development cannot proceed without conflict. Intrapersonal conflict reflects competing impulses and desires as well as incongruent information. The resolution of intrapersonal conflict shapes the individual's view of the self and the self's place in the world. Interpersonal conflict is a social episode marked by overt behavioral opposition. The characteristics of the participants and their relationship, the issue, the frequency, and the tactics employed determine the developmental significance of interpersonal conflicts. Adaptive, constructive conflicts are known to facilitate individuation, growth, and positive adjustment. Maladaptive, destructive conflicts hinder development and foster social, emotional, and cognitive problems.

Conflict Defined

Conflict is a broad term, widely invoked in the common vernacular and in the scientific literature to suggest a state of disagreement or opposition. The term carries different connotations depending upon whether the experiencer is singular or plural. Intrapersonal conflict describes a state of emotional turmoil or intellectual dissent that occurs within an individual. Interpersonal conflict describes a social interchange that is marked by opposing goals and that involves two or more individuals. The earliest English usage of each term dates to the fifteenth century; both have common roots in the Latin *conflictus*, meaning throw or strike together (Shantz and Hartup, 1992). This shared etiology tends to muddy distinctions between these two distinct forms of conflict.

A conflict has been likened to a play or novel. In the case of interpersonal conflict, there is a protagonist and an antagonist (conflict participants). In the case of intrapersonal conflict, the protagonist and antagonist come in the form of a thesis and an antithesis. For both interpersonal and intrapersonal conflict, there is a theme (conflict topic), a complication (initial opposition), rising action (conflict behaviors), climax or crisis (conflict resolution), and denouement (conflict outcome). These discrete components combine to form a coherent whole. Plots unfold according to a prescribed sequence and so do conflicts.

Intrapersonal Conflict

Intrapersonal conflict encompasses two broad domains. At the emotional level, conflict reflects competing impulses or desires. Freud (1965) attributed these tendencies to contradictory aggressive and sexual proclivities with origins in the id (the unregulated physical self) and to irreconcilable differences between these selfish urges and constraints imposed by the superego (the social mores conveyed by family and culture). At the cognitive level, conflict reflects dissonance produced by incongruent facts, opinions, or modes of thought. Piaget (1985) attributed these tendencies to equilibration, the innate tension between a preference for stability, and a drive to master the environment. Assimilation encourages stability by incorporating new information into existing mental schema. Should this prove insufficient for mastering the environment, accommodation is required to alter mental schema in a manner consistent with the information available.

Intrapersonal conflict is by definition a private experience. It cannot be observed and it has no reliable metric. Early scientific psychologists relied on introspection to glean insight into thoughts, a practice modified and popularized by psychodynamic psychologists to probe unconscious emotional states. Subjective experience as a topic of study fell out of favor with the behavioral revolution. Reaction times, physiological cues, and projective responses were employed as proxies for inner conflict, but poor reliability and the absence of face validity undermined these efforts. With the cognitive (counter)revolution, new introspective methods gained legitimacy as advances in measurement and instrumentation were applied to mental processes. Scholars began to focus on the diverse experiences of different cultural and ethnic groups, and in so doing it became apparent that scientific inquiry into intrapersonal conflict cannot proceed in the absence of valid and reliable self-report data.

Interpersonal Conflict

Interpersonal conflict may take place at the level of the dyad or the group. Dyadic conflict is a social episode marked by overt behavioral opposition. The function of such conflicts and the manner in which they unfold vary according to characteristics of the participants and their relationship, the issue and context of the disagreement, the tactics employed, and their consequences (Deutsch, 1973). Group conflict refers to negative or incompatible attitudes and behaviors directed by members or representatives of one group toward those of another group. The nature and significance of intergroup conflict differs according to leadership styles and prevailing norms, member cohesion, resource availability, and the individual and collective talents and experiences of each group (Sherif et al., 1961).

Interpersonal conflict is a quantifiable, sometimes public event. It consists of a sequence of easily recognized and well-defined components including the initial opposition and the tactics, resolution, and outcome. Three prevalent patterns have been identified (Adams and Laursen, 2001): (1) coercive conflicts, which include negative affect, coercive tactics, power assertive resolutions, and unequal or unfavorable outcomes; (2) constructive conflicts, which include neutral or positive

 http://dx.doi.org/10.1016/B978-0-08-097086-8.96007-0

affect, cooperative tactics, negotiated resolutions, and equal or favorable outcomes; and (3) unresolved conflicts, which vary in terms of affect and tactics but are similar in that they involve disengagement without a clear resolution or outcome.

Empirical studies of dyadic and group conflicts suffer from inconsistent operational definitions. Conflict measures differ in terms of whether they assess unilateral or mutual oppositions, and in terms of whether the frequency of an event is distinguished from its affective intensity. Distinct patterns of interpersonal conflict emerge when considering reciprocated as opposed to unreciprocated disagreements and when angry disputes are examined apart from the mundane.

Early work on interpersonal conflict focused on groups and dyads in experimenter-contrived conditions designed to elicit disagreements. Concerns about the ethical treatment of participants and a lack of generalizability prompted improvements in ecological validity; unobtrusive observations in natural settings have been integrated with running assessments of psychophysiological states and moment-by-moment accounts of participant cognitions and affect (Gottman, 1994). The utility of self-reports increased once scholars realized that participant reports of conflict differ not only from one another but also from observer reports, suggesting that what was once considered measurement error may in fact represent important individual differences in how conflict is perceived and experienced. It is not uncommon for two individuals to disagree as to whether they were involved in a disagreement. For instance, parents are more likely to describe an interaction as a conflict than are children. Because objective indices of child outcomes are more closely tied to self-reports of disagreement than parent or observer reports, scholars may consider this an instance where questionnaires proffer the best insight into behavior.

The Role of Conflict in Development

Conflict is inherent in growth, and human development cannot proceed without conflict. Wisdom, emotional maturity, and social skills may be tied to a series of developmentally defined tasks and the conflicts specific to each. Conflict outcomes do not exist along a single continuum that runs from the beneficial to the detrimental, with benign outcomes in the middle. Instead, costs and benefits probably represent independent dimensions, each ranging from low to high. Therefore, a conflict can have both positive and negative consequences. For example, stress and anxiety may be an unavoidable part of healthy individuation; conflict that promotes the latter may increase the former. The developmental effects of conflict appear to be limited to specific domains of influence circumscribed by individual timetables and characteristics of the disagreement such as the issue and participants.

Intrapersonal Conflict and Development

Theory concerning the ontogenetic significance of intrapersonal conflict has outpaced research on the topic. There is less research on emotional development than on cognitive and social–cognitive development, but the available evidence supports the assertion that the influence of intrapersonal conflict is defined by the topic of disagreement and the cognitive or affective domain invoked. Intrapersonal conflict arises when a situation offers benefits that make an individual initially approach the outcome rapidly, but also involves risks that make an individual avoid completing the outcome (Lewin, 1935). Indecision results when the risks become more salient than the benefits. Effects are qualified by individual timetables: development moderates the influence of conflict. The likelihood that some developmental periods are more susceptible to alterations than others leads to the provocative suggestion that intrapersonal conflict is neither a necessary nor a sufficient condition for development but rather a factor that may, at certain times, facilitate advances in the specific emotional or cognitive arena in which it arises (Shantz and Hartup, 1992).

Adaptive Correlates of Intrapersonal Conflict

Intrapersonal conflict is thought to be instrumental to the attainment of several important milestones in cognitive and socioemotional development, although it is important to reiterate that empirical research on the topic is scarce. Intrapersonal conflict provides the impetus for cognitive and social cognitive growth and, in so doing, fosters profound shifts in behavior. Put another way, intrapersonal conflict may change the filter through which the world is interpreted and thereby altering motivations for social action.

Mechanisms of Adaptive Intrapersonal Conflict

In the Piagetian theory (1985), accommodation implies the occurrence of intrapersonal conflict that leads to significant developmental changes. Specifically, some forms of intrapersonal conflict prompt changes in internal schema and mental structures, such that prior views are discarded after weighing newly encountered perspectives. Additionally, intrapersonal conflict may alter affect and impulse control. Resistance to temptation can be thought of as a conflict between an immediate reward and a later cost. The strength model of self-control (Baumeister et al., 2007) suggests a mental mechanism akin to a muscle that improves in strength with exercise. Intrapersonal conflict also fosters social cognitive change. With cognitive maturity, the child comes to realize that people hold different views of the world, even when they have access to the same information. The ability to simultaneously consider multiple perspectives prompts advances in moral reasoning commensurate with the abandonment of egocentric views.

Developmental Outcomes of Adaptive Intrapersonal Conflict

Four instances in which intrapersonal conflict may be adaptive deserve mention. Accommodation (Piaget, 1985) exemplifies the first process: a child confronted with a difficult moral dilemma may be forced to disregard views previously held dear in order to objectively consider different perspectives and reach a fair solution. Second, intrapersonal conflict can improve abilities to delay gratification. A child who opts to delay gratification in the hopes of improved rewards will learn to succeed by cultivating strategies that minimize emotional arousal and divert attention from proximate stimuli. Third, role-taking abilities may improve as a result of intrapersonal conflict. A child who contemplates how to obtain goals and overcome obstacles must consider the best means to an end.

A consideration of influence and tactics may foster advances in the perceptions of others. Finally, intrapersonal conflict may force an individual to evaluate goals and establish priorities. As youth consider how to invest their time, effort, and resources they gradually commit to an individual identity.

Maladaptive Correlates of Intrapersonal Conflict

Intrapersonal conflict may also be the cause of adjustment difficulties. It should be noted, however, that the mechanisms have proven hard to substantiate, suggesting that causal evidence for adverse outcomes must be considered tenuous.

Mechanisms of Maladaptive Intrapersonal Conflict

Intrapersonal conflict may arise when the child is confronted with equally attractive (or equally unattractive) alternatives. Alternatively, the child may be presented with information that is inconsistent with specific beliefs or attitudes. Cognitive dissonance (Festinger, 1957) describes the tension arising from the need to resolve competing claims. Positive features of one position or negative features of another will be emphasized and exaggerated until the alternatives are no longer equal, making it easier to select an alternative. Rumination describes an obsessive focus on a dilemma or problem, creating intrapersonal conflict as the individual struggles to shift attention to other matters. Maladaptive intrapersonal conflict also encompasses incongruous urges or impulses. Sigmund Freud (1965) argues that intrapersonal conflict was inevitable as adolescents struggle with emerging sexual impulses arising from the id and societal constraints internalized in the superego during the genital stage. G. Stanley Hall (1904) similarly claimed that inner conflict was an inherent developmental milestone, famously dubbed adolescent 'sturm und drang' or storm and stress. The origins of this inner turmoil come from the contradictory urges and opportunities that accompany imminent adulthood.

Developmental Outcomes of Maladaptive Intrapersonal Conflict

Maladaptive outcomes attributed to intrapersonal conflict are legion. Cognitive dissonance can be debilitating. Failure to resolve dissonance may give rise to physiological stress. Yet the need to resolve dissonance may force the individual to make an ill-timed, uninformed decision. In the process, inaccurate information may be encoded with unintended consequences. Forced to select a partner for an outing, the child may enhance the esteem of one friend in relation to others, unwittingly limiting future social experiences. Rumination can give rise to depression, a logical response to a negative world view. Similarly, normative adolescent intrapersonal conflict may have pathological consequences. Depression may come from perceived powerlessness to control impulses and overcome obstacles. Antisocial behavior and attitudes sometimes accompany perceptions of injustice that accompany the realization that youth often have no control over their opportunities or future.

Interpersonal Conflict and Development

There is considerable support for the premise that the developmental significance of conflict depends on the relationship in which it arises. Three relationship properties are germane: closeness, permanence, and power. Disagreements in close relationships have the potential to alter developmental trajectories profoundly; few disagreements in other relationships have such effects. Disagreements in voluntary relationships are more apt to bring about individual change than those in obligatory relationships; participants in the former must behave in a manner that affords mutually satisfactory outcomes if the affiliation is to continue, whereas participants in the latter worry little about the dangers of relationship dissolution. Disagreements in horizontal or mutual relationships provide especially fertile ground for improved social skills because when power is shared, negotiation is necessary for an amicable resolution; in vertical or unilateral relationships, resolutions are more often the product of power differentials than of negotiation abilities. These variations notwithstanding, developmental shifts in conflict management may be identified such that the practice of coercion is gradually replaced across childhood and adolescence by negotiation and withdrawal as advances in the ability to negotiate during adolescence are accompanied by a growing appreciation of the important role mitigation plays in preserving interpersonal bonds (Selman, 1980).

Frequency also helps determine the developmental significance of conflict (Laursen and Hafen, 2010). Regarding detriments, there are costs associated with the complete absence of conflict because individuals who avoid conflict must suppress their needs. This has deleterious effects on relationships and participants, because relationships that are not sensitive to participant needs are perceived as lacking in closeness and support, which are risk factors for adjustment difficulties. As the number of conflicts increases to moderate levels, the costs associated with conflict should decline. Although conflict is never a risk-free enterprise, moderate amounts of nonabusive discord are not known to predict adjustment problems. As the frequency of conflict increases further, however, the costs associated with conflict should increase sharply. Thus, a positive association between conflict frequency and poor outcomes may be nonlinear, with adverse outcomes increasing as an exponential function of increases in conflict. Regarding benefits, there are few benefits associated with completely avoiding conflict. Up to a point, a positive association between conflict and benefits demonstrates that additional conflicts should be linked to better outcomes. Eventually, a saturation point is reached and outcomes do not improve or decline for those with conflict beyond the optimal level.

The effect of conflict frequency on developmental outcomes depends on an additional factor: relationship quality. In poor-quality relationships, those who avoid conflict may actually be better adjusted than those who do not, because it is unlikely that participants are missing out on conflict experiences that might buffer against adjustment problems. Disputes in unsupportive relationships tend not to have positive consequences. Each successive increase in conflict in poor-quality relationships is likely to bring fewer positive outcomes to the point where conflict offers virtually no benefits for participants. Adjustment problems should increase as a direct linear function of the number of conflicts.

In contrast, there may be costs in avoiding conflict in high-quality relationships because supportive relationships attuned

to the needs of participants may well respond to disagreements in a manner that protects against maladjustment. Avoiding conflicts leads individuals to miss out on the benefits that typically arise from disagreements in supportive affiliations. For this reason, individuals with moderate amounts of conflict in high-quality relationships demonstrate fewer adverse outcomes and better adjustment than individuals with no conflict (Adams and Laursen, 2007). It is probably the case that most benefits can be realized from only a few disagreements, but it is probably also the case that a few more disagreements are unlikely to interfere with the realization of benefits. At some point, however, conflict will become counterproductive and positive outcomes will decline in response to additional conflict. Excessive conflict is expected to correspond with few or no beneficial outcomes, regardless of the quality of the relationship, but it probably takes more conflicts to reach this point in high-quality relationships than in poor-quality relationships.

Adaptive Correlates of Interpersonal Conflict

Social skills and social development are entwined with interpersonal conflict. Conflicts that promote intellectual dialogue and justifications of views are more apt to be beneficial than conflicts that are limited to simple expressions of opposition. The adaptive significance of conflict varies across relationships. Better outcomes accompany those in better relationships. For example, disagreements, especially those arising in close relationships, are one of the primary means whereby individuals define themselves. Constructive conflicts, characterized by productive dialogue and affirming exchanges, should predict more beneficial and fewer detrimental outcomes, particularly when arising in supportive relationships.

Mechanisms of Adaptive Conflict

Three influence pathways merit mention. First, conflict may prompt changes in self-understanding and self-evaluation. Consequential disagreements force participants to reevaluate goals and tactics, offering opportunities to hone social perspective taking, and interpersonal negotiation skills (Selman, 1980). Second, conflict helps to determine patterns of social interaction in close relationships that may impact developmental outcomes. Third, conflict may define and alter expectations about individuals and relationships, such that in some cases roles and responsibilities are negotiated directly, but in other cases behavior is evaluated for cues as to whether current expectations are consistent with the demands of the situation and the maturity of the participants. A prime example of adaptive conflict is when young adolescents seek to gain authority over domains that were once subordinated to parents, especially in areas that concern the regulation of personal behavior.

Developmental Outcomes of Adaptive Conflict

Conflict with peers is thought to be especially critical to cognitive and social cognitive development because differences between relative equals are not typically settled by power assertion but instead require negotiation and perspective-taking abilities. Peer interaction is important not because peers are a more credible source of information, but because conversations between peers involve a unique form of discourse that requires each participant to articulate and defend a position. Social encounters between peers encourage children to reconsider their own views in light of the alternatives proposed, increasing the self-reflection required to discover errors and abandon untenable positions. According to this view, it is not conflict per se that promotes intellectual growth but rather the challenging, constructive dialogue that is embedded within it. Greater cognitive development is more apt to occur following peer conflict than parent–child conflict because of the frequency of exchanges that require elaboration, justification, and critical thought. Conflict with siblings and friends is also associated with greater affective perspective taking and emotional sensitivity, and higher levels of moral development.

Conflict promotes psychosocial development and well-being. Disagreement can enhance mental health and social adjustment insofar as it provides opportunities for improving self-expression and refining interpersonal collaboration skills. Conflict is posited to be an essential part of the process of adolescent individuation. Constructive disagreements between children and parents are positively associated with greater adolescent identity development, role-taking skills, and self-esteem. Conflict with parents serves to promote adolescent individuation and identity development, facilitating the normative process of separation and differentiation from the family. Some disagreement is thought to be necessary for the adolescent to develop a unique identity and to establish autonomy from parents. In addition, conflict may be the impetus that forces parents and adolescent children to renegotiate family roles and responsibilities in a manner that is consistent with the child's heightened maturity. Adolescent ego and moral development have also been tied to constructive family problem solving.

Maladaptive Correlates of Interpersonal Conflict

While some disagreements offer individuals an avenue to define themselves, others delimit social interactions. This can be reconciled with indications that conflict is beneficial insofar as the association between conflict and individual well-being is nonlinear. In this view, some disagreement is required to address problems and facilitate change, but there is a point where most benefits have been realized and additional conflict is counterproductive. Eventually, the adverse consequences of chronic discord overtake the gains that accrue from moderate disagreement. It follows that adolescents who report some conflict should fare as well or better than adolescents who report no conflict because experience managing disagreement may promote competencies that buffer against maladjustment; however, adolescents who report high levels of conflict should fare worse than those who report some or no conflict because unremitting contentiousness undermines self-worth and overwhelms coping mechanisms. Unsupportive relationships and coercive conflicts characterized by hostile assertions and behavior heighten costs and reduce benefits from conflict.

Mechanisms of Maladaptive Conflict

Two influence pathways merit mention. Conflict may lead to patterns of social interaction in close relationships that

negatively impact developmental outcomes. Persistent conflict corrodes closeness. Coercive exchanges tend to be self-perpetuating and, once started, they interfere with the dividends that normally accrue from positive affiliations (Patterson, 1982). Those who rely on coercion to resolve disagreements live in different social worlds from those who use other means to resolve disagreements. This is true in a figural sense, because the same partner will respond to coercion differently than to collaboration. Overtime, chronic hostility breeds maladjustment and troubled youth behave in ways that tend to elicit even greater antagonism. There is also literal truth to the claim that high-conflict individuals live in a different social world than low-conflict individuals. Most people find conflict aversive and seek to avoid it. Those who do not must eventually find themselves in the company of similarly coercive interaction partners, which may heighten disputatiousness, furthering isolation from low-conflict counterparts.

Conflict may also heighten existing problems and pathologies. The consequences of stable personality traits, such as low agreeableness, may be exacerbated when coupled with coercive conflict management strategies. The same may be said for youth with a predisposition for aggression or social anxiety. In each case, conflict may amplify attributes that threaten well-being.

Developmental Outcomes of Maladaptive Conflict

Strong evidence supports the assertion that children who experience high rates of angry conflict suffer from low self-esteem, poor social skills, interpersonal rejection, and adjustment problems. Coercive conflict is especially indicative of later difficulties (Laursen and Pursell, 2009). Much the same can be said of conflict during adolescence and young adulthood. Elevated rates of interpersonal conflict have been linked to adverse outcomes in at least eight areas: aggression, depression, delinquency and deviancy, school grades, health, stress, withdrawal, and relationship difficulties.

Conflict, particularly coercive conflict, is stressful. Negative affect is the primary source of stress arising from conflict; anger activates a fight or flight response which, among other things, may result in an extended period of elevated blood pressure. Chronic conflict may also provoke apprehension and anxiety, which can trigger a host of potentially debilitating physiological reactions. Conflict may foster rumination; when combined with anger and anxiety, rumination elevates risks for depression. Affective arousal also interferes with information processing abilities. The mental and physiological states that anticipate or follow from conflict may adversely impact decision-making skills, with consequences that range from unpleasant to disastrous. Coercion tends to beget coercion; disagreements that spiral out of control can trigger hostility and violence. Poorly managed conflict may foster coercive interpersonal processes that interfere with normal socialization; poor social skills give rise to school difficulties, peer rejection, and affiliation with delinquent age-mates. Finally, conflict may interfere with the supportive functions of relationships, leaving individuals isolated and alienated.

In poor-quality relationships, each successive increase in conflict has been linked to a commensurate decline in outcomes. Increases in conflict from low to moderate levels are linked to greater delinquency and withdrawal for adolescents in poorer but not better quality relationships (Adams and Laursen, 2007). High levels of conflict are never beneficial and appear worse for individuals in better quality relationships, perhaps because expectations of interpersonal harmony are violated by frequent discord. Nonlinear trends are evident in most low-negativity relationships: as conflict increased from medium to high levels, delinquency increased, school grades either worsened or stopped improving, and withdrawal either increased or stopped declining. Thus, negativity exacerbates the difficulties that accompany chronic conflict and inhibits the beneficial consequences of modest conflict by creating conditions of mistrust and apprehension that are antithetical to supportive relationships. Under these circumstances, disengagement from parents and the accompanying renegotiation of adolescent roles and responsibilities may prove difficult, prompting a downward spiral of maladaptation and contentiousness.

Youth who have difficulty getting along with others are at considerable risk. Disagreeable youth have high levels of concurrent and prospective adjustment problems, with rates of depression that rival victimized youth, and rates of antisocial behavior that rival aggressive youth.

Conclusion: The Paradoxical Influence of Conflict

Balance theories of conflict argue that the consequences of conflict depend on the proportion of disagreements that are constructive relative to those that are destructive (Gottman, 1994). Constructive conflicts foster supportive relationships, which promote optimal development. Destructive or coercive conflicts breed antagonistic relationships, which can undermine successful adjustment. According to this view, individual outcomes are more a reflection of the ratio of positive to negative interactions than the absolute number of each. Beneficial conflict outcomes characterize relationships where pleasant interactions outweigh unpleasant ones; detrimental conflict outcomes characterize relationships where negativity outweighs positivity. The ratio of positive interactions to negative interactions may be a better predictor of outcomes than the absolute frequency of either. It follows that conflict may predict adverse outcomes only when it crosses a threshold in terms of the proportion of disagreements to the total number of social interactions or when the number of destructive conflicts exceeds the number of constructive conflicts. Not everyone subscribes to this view. According to some, debilitating outcomes accompany destructive conflicts, regardless of their number or ratio to positive exchanges.

Conflict is often unpleasant and aversive. Given the option, most people would choose to avoid it. For this reason, it is commonly assumed that conflict hinders development. This assumption has been challenged on the grounds that linear models overlook the beneficent role that conflict plays in maturation. A curvilinear model of influence captures this alternative perspective: moderate conflict provides experience necessary for optimal socioemotional development, whereas too little conflict limits the potential for growth and too much conflict overwhelms mechanisms for coping with it. Conflict is not necessarily detrimental, but neither is it intrinsically beneficial; the consequences depend on its frequency, the way

in which it is managed, and the quality of the relationship in which it arises.

See also: Cognitive Development: Child Education; Cognitive Development: Mathematics Learning and Instruction; Cognitive Styles and Learning Styles; Conflict and Conflict Resolution, Social Psychology of; Conflict and Consensus; Conflict: Organizational; Learning and Instruction: Social-Cognitive Perspectives; Personality Development and Temperament.

Bibliography

Adams, R., Laursen, B., 2001. The organization and dynamics of adolescent conflict with parents and friends. Journal of Marriage and Family 63 (1), 97–110.

Adams, R., Laursen, B., 2007. The correlates of conflict: disagreement is not necessarily detrimental. Journal of Family Psychology 21 (3), 445–458.

Baumeister, R.F., Vohs, K.D., Tice, D.M., 2007. The strength model of self-control. Current Directions in Psychological Science 16 (6), 351–355.

Deutsch, M., 1973. The Resolution of Conflict: Constructive and Destructive Processes. Yale University Press, New Haven, CT.

Festinger, L., 1957. A Theory of Cognitive Dissonance. Stanford University Press, Stanford, CA.

Freud, S., 1965. The Problem of Anxiety. Norton, New York (Original work published 1936).

Gottman, J.M., 1994. What Predicts Divorce? The Relationship between Marital Processes and Marital Outcomes. Erlbaum, Hillsdale, NJ.

Hall, G.S., 1904. Adolescence Its Psychology and Its Relations to Physiology, Anthropology, Sociology Sex, Crime, Religion and Education, vol. II. D Appleton & Company, New York.

Laursen, B., Hafen, C.A., 2010. Future directions in the study of close relationships: conflict is bad (except when it's not). Social Development 19 (4), 858–872.

Laursen, B., Pursell, G., 2009. Conflict in peer relationships. In: Rubin, K.H., Bukowski, W.M., Laursen, B. (Eds.), Handbook of Peer Interactions, Relationships, and Groups. Guilford, New York, pp. 267–286.

Lewin, K., 1935. A Dynamic Theory of Personality. McGraw-Hill, New York.

Patterson, G.R., 1982. Coercive Family Process. Castalia, Eugene, OR.

Piaget, J., 1985. The Equilibration of Cognitive Structures (Original work published 1975). University of Chicago Press, Chicago.

Selman, R., 1980. The Growth of Interpersonal Understanding: Developmental and Clinical Analyses. Academic Press, New York.

Shantz, C.U., Hartup, W.W., 1992. Conflict in Child and Adolescent Development. Cambridge University Press, Cambridge, UK.

Sherif, M., Harvey, O.J., White, F.J., Hood, W.R., Sherif, C.W., 1961. Intergroup Conflict and Cooperation: The Robbers' Cave Experiment. University of Oklahoma, Norman, OK.

Conflict and War, Archaeology of: Behavior and Social Organization

David H Dye, University of Memphis, Memphis, TN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Archaeologists have a long-standing interest in the evolution of human conflict and war. They identify three modes of aggression and violence based on the archaeological record: self-redress homicides, feuds, and war. Self-redress homicides predominated early in human history and continued among egalitarian hunter-gatherers. Feuds are the typical means of tribal justice for nonegalitarian hunter-gatherers, pastoralists, and small-scale farming communities. Warfare employs organized deadly force by agricultural chiefdoms and various state societies that seek to alter the balance of power between autonomous political communities. The ways in which violent practices are constituted, result in part from the nature and scale of human social organization.

Over the past two decades archaeologists have continued to develop robust models for understanding conflict and war in the past. New geophysical techniques and bioarchaeological methods, for example, are providing key physical evidence of defensive systems (Keeley et al., 2003) and skeletal trauma (Knüsel and Smith, 2014; Martin et al., 2012). Archaeologists are also positing hypotheses concerning the causes and consequences of violence, and working toward assessing how conflict and war have shaped our evolution as a species (Fry, 2013). These studies of past societies are critical for informing us of the human capacity for aggression and violence, as well as cooperation and peace, not only in the past and present, but also for the future.

Critical evaluation of ethnographic studies has sharpened interpretations of past human behavior, and has led to questions central to understanding changes in conflict and war throughout the world (Ferguson and Whitehead, 2000; Fry, 2006; Reyna, 1994). Several questions have traditionally been of interest to archaeologists. How is aggression and violence expressed under varying circumstances? What political and social factors precipitate violence and warfare? What role does conflict and war play in efforts to increase political power, as well as to undermine political centralization? And finally, How is the status quo maintained in the face of conflict and war? The answers to these questions frequently take into account variations in both the intensity and prevalence of conflict and war across space and through time. The evolution of conflict and war is perceived by archaeologists, not from a monocausal perspective, but rather as part of a complex, developing social and political fabric. Such an approach forces the recognition of the multiple processes that give rise to intercommunity conflict.

Egalitarian Hunter-Gatherers and Self-Redress Homicides

Some degree of violence is present in virtually all human societies – people everywhere and at all times possessed the potential to achieve and forward their goals and relations with one another through aggression, conflict, and violence, as well as through cooperative and peaceful behavior. It is important to ask under what conditions different modes of conflict and war vary through time. How is aggression and violence expressed in terms of the scale at which society is politically and socially integrated? To what extent can archaeology document the variability of aggression and cooperation? For example, hunter-gatherer conflict and violence exhibit great range in scale and variation (Kelly, 2013). Archaeological and ethnographic perspectives draw important distinctions in the varying nature of hunter-gatherer conflict and violence: egalitarian, family-level foragers generally lack organized intercommunity conflict, while nonegalitarian, socially hierarchical, and segmented hunter-gatherer violence ranges from feuds to war (Fry, 2006; Kelly, 2000).

Among egalitarian foragers conflicts often erupt over personal grudges, which may be resolved through alliances, avoidance, negotiation, and toleration. When affiliative actions fail, the injured parties might resort to brawls and self-redress homicides – actions undertaken by relatives who seek their own justice in cases of injured or murdered kin. Family-level hunter-gatherer violence is typically limited because they lack the overarching, cohesive kin organizations necessary for coordinating social groups larger than the family. When family-level, organized violence does occur it settles disputes, but it is ineffective for controlling people (Fry, 2006; Reyna, 1994).

Hunter-gatherers were once characterized as living a Hobbesian existence – one of "continual fear, and danger of violent death," in which human life was "solitary, poor, nasty, brutish, and short" (Oakeshott, 2008: 94). But with ethnographic field work in the post-World War II years, this perspective shifted to one where family-level foragers were viewed as 'simple, harmless' people who lived in a state of harmony (Thomas, 1958). Based on more recent research, neither view is entirely justified; a more accurate appraisal characterizes egalitarian hunter-gatherer conflict as violence that might become lethal. In general, conflicts are resolved through a variety of social mechanisms. For instance, South Indian foragers typically exhibit self-restraint to manage group conflicts, in addition to separation (walking away from those who threaten or offend), silence, and tolerance. The Kadar, for example, simply move from one place to another when any harsh words are spoken, believing that one forest is as good as another. The Paliyar upon feeling they have been disrespected, typically refrain from verbal rejoinders or aggressive gestures

 http://dx.doi.org/10.1016/B978-0-08-097086-8.13029-6

such as a shaken fist (Gardner, 2013). These models of hunter-gatherer behavior and lifestyle have had a profound impact on how archaeologists view the past and how theoretical models are constructed for investigating the evolution of conflict and violence, as well as cooperation and peace.

A reevaluation of hunter-gatherer society based on archaeological and ethnographic research now suggests that family-level, egalitarian foragers embraced the politics of nonconfrontation with a general ethos of nonviolence (Fry, 2006). However, when a group's social mechanisms for dispute resolution fail, opponents may come to blows. And when violence does occur, it typically takes place within the social group, being confined to the adjudication of murders that usually result from pursuits over women, reactions to insults, and revenge for prior homicides. Overall, egalitarian foragers tend to have low rates of violence, typically limited to brawls or fist-fights, and rarely feuds. Aggression is relatively harmless, and while it is considered inappropriate to show anger, violence may erupt into a melee with no particular objective other than expressing pent-up frustrations resulting from personal grudges. For example, among the Mbuti of the Democratic Republic of the Congo disputes might erupt every 3 to 4 days (Turnbull, 1965). Certainly, such anger may become lethal, but violence is a product of an individual's own rage and not the calculated risk of one group choosing to raid another group for justice (feud) or the pursuit of power and the promotion of political agendas (war).

In small, family-level, egalitarian societies, aggression is typically personal and may or may not escalate to revenge killing. When a killer is killed, self-redress homicide or capital punishment is the socially sanctioned means of achieving justice. Killing a recidivist murderer is usually limited to the perpetrator of a prior homicide, i.e., a person whose criminal responsibility has been well established. The object of such 'eye-for-an-eye' actions may also include deaths attributed to conjurers, shamans, sorcerers, witches, or anyone who is believed to have killed their victims by 'black magic' or other forms of supernatural power (Fry, 2006; Kelly, 2013).

Until humans became relatively sedentary and began to domesticate animals and plants, the world was populated exclusively by hunter-gatherers, and virtually all of them lived in small, egalitarian, family groups, that is, societies without social classes, hierarchies, ranking, or segmentation. In other words, they lacked clans, lineages, or sodalities. Skeletal data are rare prior to about 25 000 years ago, but the available bioarchaeological evidence suggests that foragers lived a hazardous lifestyle, one involving close encounters with large game, as well as interpersonal violence, but there is a general lack of evidence for organized feuding or endemic warfare (Kelly, 2013). Living in widely dispersed bands with few material possessions to fight over, cannibalism, hunting accidents, predation, and self-redress homicides, rather than revenge killings or warfare, may account for virtually all skeletal traumas during much of human prehistory prior to about 12 000 years ago. A case in point is Kennewick Man, a 9300-year-old Paleo-Indian hunter-gatherer, who suffered multiple violent traumas, including old wounds that had healed – a still-embedded stone spear point in the hip, depression fractures on the skull, a broken rib, and massive blunt trauma to the chest. He also had an infected head injury and a fractured shoulder blade, neither of which shows the effects of healing (Chatters, 2004). Healed wounds, such as these, may reflect interpersonal violence rather than feuding or warfare in which death is the objective.

In a recent overview of European and Near Eastern conflict and war (Ferguson, 2013), the European Paleolithic revealed little evidence of violent injury – it was not until the subsequent Mesolithic period that several instances of multiple deaths are documented, suggesting the beginning of recognizable intercommunity conflict, such as feuding. Ferguson notes that Near Eastern hunter-gatherers exhibited some personal violence, but the shift to widespread intercommunity violence was not evident throughout the Near East until the Early Bronze age when war arrived in an abrupt and dramatic transition. While modern, egalitarian hunter-gatherers are not our 'contemporary ancestors,' nor exemplars of our remote past, they do provide important material correlates of how cultures maintain peace and limit aggression, conflict, and violence, and as such guide archaeological thinking about the ways in which our ancestors engaged in conflict and violence and promoted cooperation and peace.

Tribes, Feuds, and Kin-Militias

Conflict and violence in small-scale nonegalitarian sedentary societies is based on a form of justice pursued through feuding based on social substitution – that any member of the murderer's social group may be attacked and killed as a substitute for the miscreant (Kelly, 2000). But generally it is the perpetrator or a close relative who must pay the price for the crime. Feuds are carried out by small, informal, and temporary armed groups organized as kin-militia who pursue justice and retribution. Feuding is characterized by "fighting in pursuit of individual or family ends rather than community or societal ends" (Carneiro, 1994: 6), and may result in revenge raids where kinship ties outweigh personal or political allegiance. In such situations of reciprocal violence, a "grudge precipitates retaliation, which in its turn becomes another grudge that ignites counter retaliation, and so on" (Reyna, 1994: 38). The social mechanisms for cooperation and peace may not be entirely effective to squash lingering or smoldering feelings of injustice, especially when the individual or kin group demands equity and revenge to assuage the corporate sense of honor.

Unlike warfare, feuds are characterized by outbursts of mutual hostility and violent assaults that are usually limited to one or two killings at a time, with fighting most often involving brawls, raids, and less often small-scale battles (Fry, 2006: 101). These vendettas, often fueled by the sense of a sacred familial responsibility or social duty, might entail a long-running cycle of resentment and hostilities between the aggrieved parties, as there are no clear or necessary political objectives beyond the maintenance of individual or family honor. The amount of pressure brought to bear on feuding parties by their own groups to end an ongoing dispute increases in proportion to the necessity of cooperation between the two factions (Boehm, 1987). For example, the Kapsiki/Higi, a tribal people who live in the rugged Mandara mountain range that separates Nigeria from Cameron, engage in duels, feuds, and staff fights brought about by accusations of theft, jealousy over

women, quarrels concerning land, and prior homicides. Battles and raids are led by a small militia consisting of a war leader, warriors, and distinguished warriors (Otterbein, 1994).

In feuding, related kin groups typically operate under weak chains of command, having only rudimentary powers for conquering or dispersing their enemies. Although organized violence is more effective than that found in bands, it is still relatively harmless when compared to the violence of centralized polities, consisting mainly of brawls, raids, and rarely battles. Thus, feuding can be seen as a judicial mechanism in which aggrieved kin groups seek justice through social substitution based on resolving personal grudges by "evanescent congeries of kin and allies thrown together for one or two operations" (Reyna, 1994: 42).

Raids operationalize the demands of feuding and are culturally sanctioned, often elevating the status of the successful individual to that of warrior, a status not found in egalitarian hunting and gathering societies. Feuding typically entails armed combat between localized kin groups within a political community (Otterbein, 1994: xix–xx). Unlike egalitarian foragers who are dispersed and mobile, and whose band membership is fluid, tribal-level, nonegalitarian hunter-gatherers and horticulturists are often sedentary peoples, characterized by well-defined territories and hierarchical social segments, such as clans, lineages, sodalities, and villages. Violence arises among nonegalitarian hunter-gatherers when mobility becomes less of an option due to high population density (Kelly, 2013: 165). With residential movement increasingly limited, available habitation locales become limited, scarce, and much sought-after, even to the point of being fought over, sometimes erupting into periodic, all-out confrontations. Among pre-Columbian southern California hunter-gatherers, for instance, conflict dramatically increased based on the frequency of projectile injuries from about 1500–600 years ago, perhaps precipitated by subsistence stress exacerbated by a growing population size and multiple droughts during the Medieval climatic warming (Lambert, 2007).

Key archaeological evidence for intergroup feuding, as opposed to self-redress homicides, includes human bone modification, mass burials of individuals with embedded projectile points, perimortem mutilation associated with trophy-taking, and skeletal trauma exhibiting broken and fractured bones. Modified human bones as grave accompaniments are frequently evidence of warrior prestige conferred by successful revenge attacks that become trophy-taking raids. Trophy-taking escalates among nonegalitarian hunter-gatherers and tribal societies as warrior status takes center stage and human body parts become a social currency in 'war' rituals and for social advancement. For example, Late Archaic hunter-gatherers of the lower Tennessee Valley of eastern North America adopted a feuding style of conflict resolution between about 7000 and 3000 years ago (Anderson, 2002; Dye, 2009). Graves contained modified human remains as trophies in the form of skull cap bowls and gorgets, carved fibulae and radius awls/pins, and modified and cut tibia and femur shafts (Smith, 1997).

The initial rise of tribal societies throughout the world witnessed increased violence based on kin-based raiding although many examples of skeletal trauma undoubtedly resulted from interpersonal violence due to murders resolved by self-redress homicide. In a feuding style of conflict resolution, victims of violence begin to appear with regularity in the archaeological record, as evidenced by embedded spear points, mass graves, overkill (inflicting greater injury than would be necessary to kill someone), pin-cushioning (leaving a large number of arrows or spears in a body), and trophy-taking behaviors. There is abundant evidence substantiating archaeological claims for feuding as opposed to self-redress homicide or warfare, especially in contexts where large samples of human skeletal remains are available, suggesting that blood feuds were the violent outcome of individual and kin group personal grudges.

Tribal alliance formation and conflict resolution would have been achieved through various ameliorative practices as recorded in the archaeological record, especially adoptions, feasting, and reparation payments that utilized symbolic weaponry and tokens of exchange and wealth that circulated as gift items. A case in point is the Hopewell exchange system, which consisted of many diverse communities in eastern North America that interacted with one another between approximately 2200 and 1600 years ago. In addition to a complex ritual landscape, Hopewell long-distance exchange networks dampened the pursuit of vengeance, enhanced individual status, and promoted alliances. Exotic materials such as copper from the Great Lakes, marine shell from the Atlantic and Gulf Coasts, mica from the southern Appalachians, and obsidian from Wyoming, served to establish territorial alliances and ritual mechanisms that mediated intergroup disputes. Such articles were widely dispersed among tribal groups, in part as blood payments employed in exchange systems whose underlying purpose resulted from conflict resolution associated with a feuding style of justice seeking. For example, in the American Midwest during the Colonial period, death compensation payments resulting from intertribal murders 'covered the dead' and reestablished peaceful relations and social harmony (White, 1991: 77–79).

The appearance in the archaeological record of combat iconography, symbolic weaponry, and trophy-taking may be early hall-marks of a feuding style of violence fomented by kin-militias. The revenge feud is so basic to human conflict and justice seeking, especially when efforts at peace fail, that they are still practiced in many areas of the world where local law enforcement is weak or ignored. For example, in modern Montenegro in the European Balkans, blood revenge continues to engender an unspoken social obligation to seek retaliatory revenge to salvage one's honor as a result of moral or social disrespect, a prior murder, or a sense of public humiliation (Boehm, 1987).

War and the Rise of Chiefly Militias, Aristocratic Armies, and Industrial Armed Forces

War is ultimately about authoritative power – authority legitimized by military force and political rules – and the individuals who through social organizations control and deploy it (Mann, 1986). Through war, individuals or governments seek to increase political power or to maintain the status quo while undermining the political power of an opponent. With the evolution of centralized polities and their emerging political

and social hierarchies, new ways of structuring violence gave rise to the organization of lethal military forces. War involves the entire social group "organized as political communities, and – intentionally or unintentionally – its outcomes affect the balance of power between such social groups and communities" (Wolf, 1987: 129). While warfare may temporarily exist among less centralized societies, it is among regional polities such as chiefdoms and states that warfare pursues 'national' policies by inflicting unprecedented lethal aggression against one's enemies. War thus entails relatively impersonal violence for political motives between autonomous communities for political power (Fry, 2006: 140). The scale of warfare runs the gamut from occasional expedient chiefly battles to the campaigns of industrial nation-states.

Chiefdoms and Chiefly Militias

Warfare among chiefdoms is based on a corporate organization that includes a chief's retainers who form a council and the militias they call up from within the polity and neighboring, allied polities (Reyna, 1994: 43–44). Chiefly militia units might include warrior contingents from the low hundreds to several thousands (Carneiro, 1990) and were capable of superior tactics over tribal kin-militias because of improvements in military structure, which enabled chiefs as leaders to issue orders down through the chain of command. Religious cults further strengthened a chief's military authority by aligning a culture hero's interests with the ruling elite's political agenda. War councils, held to discuss tactics and strategies, depended on advice from experienced warriors, counsel from priests, and intelligence from spies. Chiefly military hierarchies possessed rules that underwrote chains of command, which in turn provided the leadership necessary to maneuver formations that concentrated militias at strategic points of attack and created the element of surprise. Chiefs had the capability to commit their forces in restricted campaigns with clear political objectives that could result in great slaughter, and to construct durable, substantial fortifications that attracted followers by providing protection and allowed horticultural production to be intensified.

Chiefly warfare was chronic in frequency and regularity because violence became an instrument of chiefly policy (Earle, 1997), with political objectives often including grudge resolutions arising out of elite conduct concerning political posturing, public affairs, and resource acquisition and intensification. Through military organization, force could also be deployed to acquire material goods as forms of tribute and to appropriate social labor. With the rise of chiefly militias violence could be deployed to "create, maintain, and transform social relations" (Reyna, 1994: 46), thus providing chiefs with an organizational apparatus to aggregate polities for their benefit and to construct massive fortifications. For example, the Colla of southern Peru's Titicaca basin (*c* AD 1000–1450), comprised of a series of weakly centralized confederacies, controlled the plains of the northern basin until about 1450, when they were conquered by the expanding Inca Empire. The Colla moved away from their rich farmlands during a time of extended drought, political fragmentation, and intense inter- and intraethnic conflict. They settled on defensible ridges and hilltops, sometimes gathering in large fortified towns. The extraordinary investment in their hillforts indicates a high level of concern with warfare that resulted in a balkanized political landscape prone to repeated conflict. Warfare at this time was unmatched at any other time in the archaeological sequence of the Titicaca basin, intensifying in the fourteenth century as power-hungry warlords constructed durable fortifications and engaged in cycles of revenge that embroiled the Colla in periodic wars (Arkush, 2011: 212).

The archaeological and ethnohistoric documentation of New Zealand Maori warfare also presents an instructive case for the evolution of chiefly conflict and war. There, a pattern of revenge and status rivalry devolved to a state of endemic and entrenched war over land, other resources, and politics. The construction of pa (fort) was an economic and political task undertaken by leaders who organized and financed fortification systems through negotiations with their followers, and in the process furthered their own political ambitions. While Maori pa made conquest much more difficult, it also enabled ambitious leaders to attract followers through the promise of enhanced security and also undermined further political centralization (Allen, 2006).

Archaic States and Aristocratic Armies

As archaic states arose between 5 and 3000 years ago in many areas of the world, aristocracies emerged that garnered unprecedented military and political power (Sinopoli, 1994). The nobility, through the deployment of violent force, routinely battled their neighbors. Unlike chiefdoms, with their chiefly militias, early states organized professional soldiers into armies headed by nobles who controlled the command structure, dominated critical weaponry, and welded the hierarchical organization into a coordinated and efficient fighting force. Regional campaigning, organized as a coordinated series of battles, was a perpetual practice of aristocratic armies and often resulted in recurring cycles of empire building and destruction. Through vastly increased lethality and organization, aristocratic armies possessed the means of violent force, which conferred greater aggregative powers than had previously existed in the history of the world (Reyna, 1994).

Aristocratic armies, composed of specialized combat teams that mutually supported one another on the battlefield, coordinated their strategies and tactics based on various maneuvers and weapon types. Once the logistical problems of planning and carrying out the movement and maintenance of military forces that plagued chiefdoms were overcome, armies became permanent fixtures on the political landscape, attempting to annihilate their enemies and to hold conquered lands through force of arms to extract revenues and tribute for the state. Appropriated peasant labor served as a form of taxation and tribute that underwrote the military organization and shored up the stratified class structure.

Grudges among aristocratic elites often prompted motivations for war in early states, with antipathies and hatreds among rulers stemming from political machinations. For example, William I, Count of Normandy, considered his genealogy as privileging his claim to the English throne. William thought his claim stronger than that of the current monarch, the Anglo-Saxon Harold Godwinson, the Earl of Wessex, a member of the most powerful family in England.

William's first cousin, the childless Edward the Confessor, had named Harold as successor to the throne, even though Edward had earlier promised the kingship to William. To further complicate matters, Edward had previously sworn allegiance to William. Furious at the turn of events, William amassed a large army and fleet, invaded England, and defeated and killed Harold at the Battle of Hastings. William went on to sack the English shires and was crowned king in London on Christmas Day 1066, becoming William the Conqueror of all England. William's claim to the highest office in England, based on his aristocratic genealogy, is an example of a 'noble grudge-of-state.' Much of the conflict and war in archaic states can be traced to aristocratic machinations, with the regional campaigns conforming to the logics of aristocratic economic and political interests (Reyna, 1994).

Nation-States and Industrial Armed Forces

The armies of nation-states appeared around the beginning of the nineteenth century, creating industrial fighting forces that fomented previously unimaginable lethality, unparalleled violence, and unprecedented devastation. Replacing military leaders whose offices were based on claims to aristocratic status with leaders who possessed professional abilities in management and technology, armies assumed technical expertise and created sophisticated war machines. Modern wars are often fought for profit, as a component of the military–industrial complex, driven by escalating arms races and the needs of large, standing military forces that must be maintained through taxation. Military enemies of the nation-state are potential foes who stand in the way of national aims and strategic interests. While chiefdoms might field chiefly militias of several hundred warriors and archaic states could deploy thousands of soldiers, the industrial nation-state has the capacity to field armies of millions. In some battles, which might extend for months, massive causalities could be incurred. For example, in World War I two roughly contemporary battles in 1916 resulted in over 2.5 million causalities. The Brusilov Offensive from 21 February to 18 December, resulted in 1.6 million causalities (Dowling, 2008), while in the Battle of the Somme, from 1 July to 18 November, more than 1 million people were wounded or killed (Hart, 2010).

The practices of ancient and modern armies are being studied by battlefield archaeologists (Geier et al., 2010; Scott and McFeaters, 2011), who investigate the organization and management of war at sites that played important roles in military events, including not only battlefields and combat zones, but also arsenals; camps; forts; supply depots; and prisoner-of-war, internment, and concentration camps. Battlefield archaeology holds that conflict and military sites reveal the cultural behaviors and practices of combatants and these practices can be studied through archaeological methods, techniques, and theory.

The Archaeology of Conflict and War

The archaeology of conflict and war has matured over the past several decades, with greater attention being paid to the role of aggression, violence, and war, as well as cooperation and peace (i.e., Dye, 2009; Fry, 2013). The ability of archaeologists to unravel the nature of intersociety aggression and conflict through the study of past human behavior and its material culture has improved considerably due to new techniques, methodologies, and theories. For example, the use of metal detectors has increasingly become an important component of conflict archaeology, while geophysical prospection has made it possible to virtually 'see' beneath the soil, revealing crucial information about battle grounds and durable fortification systems (Scott and McFeaters, 2011). Greater attention is also being paid to military science (Allen, 2007), as well as the ways in which fortifications may constrain warfare, as well as being a reflection of it (Arkush, 2011). Recent perspectives on the role of trophy-taking in feuding and war place such activities within a cultural context, removing interpretations that are based on demeaning and ethnocentrism stereotypes (Chacon and Dye, 2007). New developments in the archaeological study of conflict and war in the coming decades will embrace the ethics of reporting on violence and warfare (Chacon and Mendoza, 2012), as well as calling greater attention to the investigation of aggression and conflict and how violence developed in tandem with cooperation in the evolution of human society.

Archaeological studies of intersocietal violence are based not only on material culture, but also on studies of ethnographically and historically documented ancient and modern cultures. Archaeologists contend that patterns of conflict are not only recoverable in the archaeological record, but that archaeology is critical for answering questions concerning patterns of conflict and war in the remote past as well as more recent times. The archaeological record is clear – when feuding and war appear among past societies, the archaeological evidence is also present in the form of combat iconography, durable fortifications, settlement patterns, skeletal trauma, and specialized weaponry.

In virtually every society some degree of fighting or violence takes place, in spite of institutions that maintain positive relations, promote cooperation, and quell violence. But as shown by archaeological examples from throughout the world, sociopolitical complexity and lethal aggression are intertwined in complex, but patterned ways; thus, there is a strong correlation between conflict and social complexity (Kelly, 2000). Based on archaeological research, what is clear is that the degree and nature of social integration profoundly influences how conflict and cooperation are structured, underscoring the idea that violence is largely a behavioral choice. People engage in war when it is in their interests to do so, but war is not chosen lightly, as it is costly, dangerous, and lethal (Ferguson, 2013).

A growing body of archaeological data testifies to the presence of prehistoric violence around the world, dispelling the myth of a peaceful human past (Arkush and Allen, 2006; Chacon and Mendoza, 2007a,b; Dye, 2009; Keeley, 1996; LeBlanc and Register, 2003). The evidence of violence, however, is varied in both nature and scale, suggesting that neither war nor peace has been the 'natural' state of humankind. Instead, both are products of specific social circumstances, histories, and practices (Fry, 2013). The archaeology of

conflict and war is part and parcel of human culture, and as such can only be understood from the perspective of the political and social contexts in which intersocietal violence has evolved. Conflict, violence, and war are therefore functions of particular forms of political and social organizations, and the specific modes of sociocultural conditions and historical development have specific expressions of aggression and conflict (Kelly, 2000).

In addition to ritualized violence such as competitive sports and rough and tumble play, and a rugged, outdoor lifestyle, three basic modes of culturally sanctioned conflict may be defined for human societies: self-redress homicide, feud, and war (Dye, 2009; Fry, 2006; Kelly, 2000; Reyna, 2006). Each mode results in a distinctive archaeological signature as a result of the sociopolitical contexts of aggression and conflict. First, self-redress or revenge homicide is characteristic of egalitarian hunting and gathering bands, where social segments are weakly developed or lacking altogether. Self-redress homicide takes place when a killer is killed – not by the social group as a whole, but by one's kin as a form of justice sanctioned by the social unit. Egalitarian forgers typically lack feuding or warfare. Second, feuds and kin-militia raids are characteristic of tribes; in many cases these are societies in which aggrieved parties seek personal retaliatory justice either by killing the accused or the close kin of the accused through a sense of socially authorized revenge. On occasion feuding behavior may escalate into warfare. Finally, chiefdoms and states engage in warfare, defined as armed conflict with the goal of upsetting the balance of power between autonomous political communities. Each of these modes is a reminder that as society becomes increasingly complex, so too does its mode of aggression and violence. Of course, self-redress homicides and feuds do not occur in lockstep with the scale of political organization. They are pursued in virtually all societies, as different forms of violence can occur in varying sociopolitical settings. The point here is to emphasize predominant modes or trends in intercommunity conflict and violence and their relationship with the way in which society is organized.

The key to understanding conflict and war is to ask whether cooperation is a losing option or whether violence is the only viable course of action (Fry, 2006, 2013). The important question is: What is the nature of violence, and under what conditions do levels of violence vary? The intensity and level of warfare increases with the scale of cultural complexity and the degree society is integrated. Organized violence throughout the world underwent a series of major transformations, from the absence of warfare prior to about 5000 years ago to the creation of state-level armies in the nineteenth century. Violence among simple, unsegmented hunter-gatherers would have been restricted to self-redress homicides. As simple foraging societies began to develop social segmentation, feuding became more prevalent throughout the world. In time, chiefly warfare incorporated formal, hierarchical institutions and inherited offices as part of elite aggrandizement that escalated organized violence. With its great time depth and global perspective, archaeology offers a distinct and unique contribution to the study of human conflict and warfare.

See also: Aggression, Social Psychology of; Chiefdoms, Archaeology of; Conflict and War: Anthropological Aspects; Conflict: Anthropological Aspects; Hunter-Gatherer Societies, Archaeology of; Peace and Nonviolence: Anthropological Aspects; States and Civilizations, Archaeology of; Tribe; Violence, History of; Warfare in History.

Bibliography

Allen, M.W., 2006. Transformations in Maori warfare: toa, pa, and pu. In: Arkush, E.N., Allen, M.W. (Eds.), The Archaeology of Warfare: Prehistories of Raiding and Conquest. University Press of Florida, Gainesville, pp. 184–213.

Anderson, D.G., 2002. The evolution of tribal social organization in the southeastern U.S. In: Parkinson, W.A. (Ed.), Archaeology of Tribal Societies, Archaeological Series, vol. 15. International Monographs in Prehistory, Ann Arbor, pp. 246–277.

Arkush, E.N., 2011. Hillforts of the Ancient Andes: Colla Warfare, Society, and Landscape. University Press of Florida, Gainesville.

Arkush, E.N., Allen, M.W. (Eds.), 2006. The Archaeology of Warfare: Prehistories of Raiding and Conquest. University of Florida Press, Gainesville.

Boehm, C., 1987. Blood Revenge: The Enactment and Management of Conflict in Montenegro and Other Tribal Societies, second ed. University of Pennsylvania Press, Philadelphia.

Carneiro, R., 1990. Chiefdom-level warfare as exemplified in Fiji and the Cauca Valley. In: Haas, J. (Ed.), The Anthropology of War. Cambridge University Press, New York, pp. 190–211.

Carneiro, R., 1994. War and peace: alternating realities in human history. In: Reyna, S.P., Downs, R.E. (Eds.), Studying War: Anthropological Perspectives. Gordon and Breach, New York, pp. 3–27.

Chacon, R.J., Dye, D.H. (Eds.), 2007. The Taking and Displaying of Human Body Parts as Trophies by Amerindians. Springer, New York.

Chacon, R.J., Mendoza, R.G., 2012. The Ethics of Anthropology and Amerindian Research: Reporting on Environmental Degradation and Warfare. Springer, New York.

Chacon, R.J., Mendoza, R.G. (Eds.), 2007a. North American Indigenous Warfare and Ritual. University of Arizona Press, Tucson.

Chacon, R.J., Mendoza, R.G. (Eds.), 2007b. Latin American Indigenous Warfare and Ritual. University of Arizona Press, Tucson.

Chatters, J.C., 2004. Kennewick Man: a Paleoamerican skeleton from the northwestern U.S. In: Lepper, B.T., Bonnichsen, R. (Eds.), New Perspectives on the First Americans. Center for the Study of First Americans, pp. 129–135. College Station.

Dowling, T.C., 2008. The Brusilov Offensive. Indiana University Press, Bloomington.

Dye, D.H., 2009. War Paths, Peace Paths: An Archaeology of Cooperation and Conflict in Native Eastern North America. AltaMira, New York.

Earle, Timothy, 1997. How Chiefs Come to Power: The Political Economy in Prehistory. Stanford University Press, Stanford.

Ferguson, R.B., 2013. The prehistory of war and peace in Europe and the Near East. In: Fry, D.P. (Ed.), War, Peace, and Human Nature: The Convergence of Evolutionary and Cultural View. Oxford University Press, New York, pp. 191–240.

Second printing. In: Ferguson, R.B., Whitehead, N.L. (Eds.), 2000. War in the Tribal Zone: Expanding States and Indigenous Warfare. School of American Research Press, Santa Fe.

Fry, D.P., 2006. The Human Potential for Peace: An Anthropological Challenge to Assumptions about War and Peace. Oxford University Press, New York.

Fry, D.P. (Ed.), 2013. War, Peace, and Human Nature: The Convergence of Evolutionary and Cultural Views. Oxford University Press, New York.

Gardner, P.M., 2013. South Indian forager's conflict management in comparative perspective. In: Fry, D.P. (Ed.), War, Peace, and Human Nature: The Convergence of Evolutionary and Cultural Views. Oxford University Press, New York, pp. 297–314.

Geier, C.R., Babits, L.E., Scott, L.E., Orr, D.G. (Eds.), 2010. The Historical Archaeology of Military Sites: Method and Topic. Texas A&M University Press, College Station.

Keeley, L.H., Fontana, M., Quick, R., 2003. Baffles and bastions: the universal features of fortifications. Journal of Archaeological Research 15, 55–95.

Keeley, L.H., 1996. War before Civilization: The Myth of the Peaceful Savage. Oxford University Press, New York.

Kelly, R.C., 2000. Warless Societies and the Origin of War. University of Michigan Press, Ann Arbor.

Kelly, R.L., 2013. From the peaceful to the warlike: ethnographic and archaeological insights into hunter-gatherer warfare and homicide. In: Fry, D.P. (Ed.), War, Peace, and Human Nature. Oxford University Press, New York, pp. 151–167.

Knüsel, C., Smith, M.J. (Eds.), 2014. The Routledge Handbook of the Bioarchaeology of Human Conflict. Routledge, New York.

Lambert, P.M., 2007. The osteological evidence for indigenous warfare in North America. In: Chacon, R.J., Mendoza, R. (Eds.), North American Indigenous and Ritualized Violence. University of Arizona Press, Tucson, pp. 202–221.

LeBlanc, S.A., Register, K.E., 2003. Constant Battles: The Myth of the Peaceful, Noble Savage. St. Martin's Press, New York.

Mann, M., 1986. The Sources of Social Power: A History of Power from the Beginning to A.D. 1760, vol. 1. Cambridge University Press, New York.

Martin, D.L., Harrod, R.P., Pérez, V.R. (Eds.), 2012. The Bioarchaeology of Violence. University of Press of Florida, Gainesville.

Oakeshott, Michael (Ed.), 2008. Leviathan, or the Matter, Forms and Power of a Commonwealth Ecclesiastical and Civil. Touchstone, New York.

Otterbein, K.F. (Ed.), 1994a. Feuding and Warfare: Selected Works of Keith F. Otterbein. Gordon and Breach, Amsterdam.

Otterbein, K.F., 1994b. Higi armed conflict. In: Otterbein, K.F. (Ed.), Feuding and Warfare: Selected Works of Keith F. Otterbein. Gordon and Breach, New York, pp. 75–96.

Reyna, S.P., 1994. A mode of domination approach to organized violence. In: Reyna, S.P., Downs, R.E. (Eds.), Studying War: Anthropological Perspectives. Gordon and Breach, New York, pp. 29–65.

Scott, D.D., McFeaters, A.P., 2011. The archaeology of historic battlefields: a history and theoretical development in conflict archaeology. Journal of Archaeological Research 19, 103–132.

Sinopoli, Carla M., 1994. The archaeology of empires. Annual Review of Anthropology 23, 159–180.

Smith, M.O., 1997. Osteological indications of warfare in the archaic period of the Western Tennessee Valley. In: Martin, D.L., Frayer, D.W. (Eds.), Troubled Times: Violence and Warfare in the Past. Gordon and Breach, New York, pp. 241–265.

Thomas, E.M., 1958. The Harmless People. Vintage Books, New York.

Turnbull, C., 1965. Wayward Servants: The Two Worlds of the African Pygmies. Natural History Press, New York.

White, Richard, 1991. The Middle Ground: Indians, Empires, and Republics in the Great Lakes Region, 1650–1815. Cambridge University Press, New York.

Wolf, E.R., 1987. Cycles of violence: the anthropology of war and peace. In: Moore, K. (Ed.), Waymarks: The Notre Dame Inaugural Lectures in Anthropology. University of Notre Dame Press, Notre Dame, pp. 127–150.

Conflict and War, Archaeology of: Weapons and Artifacts

Helle Vandkilde, Institute of Culture and Society, Aarhus University, Hoejbjerg, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The archaeology of war is a late 'discovery' of the mid-1990s, but since then advances have been made in understanding the scale and roles of warfare in the early history of mankind. Warfare understood as violent encounter with the other probably occurred from the earliest hominids and has accompanied human societies ever since. There is ample evidence of this in terms of skeletal traumata, material culture, weapons, war-related ritual finds, fighting technologies, fortifications, and martial iconographies. New inquiries into these data sets testify that war was often present but rarely grew endemic. Rather the scale of warfare varied with historical circumstances tending to become more comprehensive and even habitual in times of radical social change. Peace was usually part of the picture too. This overall rhythm may suggest that warfare is deeply embedded in a cultural nexus rather than being a question of biology, evolution, or genetics. If not precisely a driver in history, warfare seems nevertheless often situated within the web of causes and effects of social change. The emergence of weaponry made specifically for war is overall contemporaneous with the first institutionalizations of warfare. This is a late phenomenon in the history of mankind dating to between the fifth- and third-millennia BCE from whence the warrior was commemorated in rituals as a figure of glory, beauty, and valor. Besides, genuine weapons and institutions of warfare concur widely with the implementation of more cogent forms of sociopolitical order with stronger leadership and explicit social divisions.

> So the possession of costly bronze daggers, swords, and rapiers consolidated the positions of war-chiefs and conquering aristocracies as did the knights's armour in the Middle Ages.
>
> Childe, 1941.

Archaeology is a science that provides knowledge of the past while relying on the material record of human interaction from very remote periods of early hominids to the recent history of modern humans all over the world. This record can be processed analytically to form complex data sets capable of delivering detailed knowledge of past human life worlds. Besides, it permits the analytical spotlight to be directed toward sociocultural encounters between human actors and the material domain of objects, buildings, resources, and environments. The actual interactions and organizations can be reconstructed as well as the ideological constructs based upon them. Because of the material nature of its sources, archaeology has the capacity to scrutinize concrete remnants of interactions as well as to track long-term historical trends and turning points. The discipline of archaeology therefore also has a huge potential to situate war, warfare, and the adjacent identities within a wider field of historical situations and trajectories.

Nevertheless, warfare as subject has been absent, or marginal, to scholarly inquiries except for the past couple of decades. Albeit there are exceptions as vividly exemplified in the opening quote by one of archaeology's founding fathers V.G. Childe, the 'discovery' of prehistoric warfare came as late as the mid-1990s triggered by the publication of L. Keeley's seminal book *War before Civilization: The Myth of the Peaceful Savage*. This event marked the beginning of a new scholarly awareness of the violent side of prehistoric societies which is much more in accord with the archaeological data, as shall be proposed below. The archaeology of periods with written records notably has a longer tradition for studying matters of war taking inspiration from military philosophy and military history. Stimulated by especially social anthropology and sociology, prehistoric archaeology constitutes the core of the current genre of 'Archaeology of War,' and from this broad category new academic fields are presently developing, such as the archaeology of battlefields and war commemorative studies.

The past 25 years have seen a huge body of publications often adhering to the Hobbesian thought that human beginnings were violent and chaotic. One might say that while the past in preceding studies was thought to be overall peaceful (doves, noble peaceful savage) interpretations after 1996 quite often take the opposite stance of a chronic state of war (hawks, fierce bellicose savage). While current archaeology and anthropology can sustain that warfare was very often present and sometimes grew habitual, it also proposes that warfare was rarely endemic in the Hobbesian sense of everybody's war against everybody. Warfare is not usually a form of interaction people engage in because they cannot be without it. Rather people in nonstate settings wage war because it is a less risky strategy than is peace inasmuch as allies can never be fully trusted. The archaeology of most periods and places convey a variable picture of both peaceful moments and different varieties and scales of warfare, which sometimes display an unspeakable cruelty that fully matches atrocities committed in our own time. In recent publications, warfare furthermore tends to be subjected to evolutionary schemes with the violence either becoming more, or less, frequent and lethal with time.

This brings up the question of war as a potential driver in history. Following R.L. Carneiro's seminal anthropological article from 1970, it is at times taken for granted that warfare expedites social change, and even produces state formations, as in the recent work of F. Fukuyama. By comparison, archaeology seems to point in the direction that warfare was often placed somewhere in the web of causes and effects of radical social

change. War can in some situations perhaps be singled out as a driver behind new social realities; however, the pathway was probably most often more indirect in that prolonged periods of warfare made existing hierarchies weak and promoted new social platforms from which new identities could rise.

Archaeology clearly has the potential to provide knowledge about the evolutionary position of warfare in a deep historical perspective, warfare's possible role as primary mover, and other topical issues, but it is a scientific venture only just begun. Insights will be affected by how narrowly or broadly warfare is defined at the outset. War and warfare are sometimes narrowly defined as the violent activities within state societies, but will here be widened considerably to encompass all forms of armed violent conflict between social units. While the term 'war' describes recurring situations of armed conflict, 'warfare' refers to the concrete waging war, but a degree of synonymy between the two must be accepted.

What Is Warfare?

Warfare may be subdivided into a number of overlapping categories varying from feuding among fraternal interest groups and raiding into foreign territory to large-scale military actions and place-bound battles; its precise form, scale, frequency, and content dependent on the specific societal and cultural setting. It makes good sense to understand warfare as basically cultural encounter, both internally in a group (intracultural) as well as externally toward other groups (intercultural), and not reducible to being merely a particular form of violent collective social action. Rather, warfare is interaction, both violent and nonviolent by character and comprising phases of persuasion, agreement, planning, and cooperation in addition to the violent confrontations, and very often postwar rituals. If warfare turns inward in the group (~civil war; feuding) this has in effect already been fractioned since warfare always seems to emerge from bounded social units. Warfare is waged against outsiders and violent conflict as such tends to rise from the construction of exclusionist identities, basically a question of us against them. Sociologically and historically, armed conflict both reinforces existing collectivities and has a tendency to produce new ones.

As a suitable entrance to archaeological inquiries, warfare can furthermore be comprehended as a *durée* of intentional activities; a continued flow of interactions, which ultimately imply violence and which produce unintentional consequences that are again acted upon. As a form of social interaction warfare cannot be meaningfully separated from the rules, norms, and resources of society as such. Besides, warfare often connects to rituals and other social activities, which may in sum be regarded more as a continuum of interaction than as separate practices. Not all conflicts end up in violent confrontation, but are often solved by nonviolent means, which is an important point to make.

Violence is at the very core of warfare, but is also a comprehensive concept comprising, in addition to warfare, for example, interpersonal aggression like homicide, intra-family hostilities and wife beating, forceful kinds of sport, human sacrifice, and juridical or political punishments such as torture, maiming, and death penalty. It may be difficult to distinguish clearly between these forms of violence in every case based on purely archaeological grounds. Moreover, the boundary between them can be quite subtle, and ethnography suggests that a high occurrence of interpersonal and ritualized violence other than strictly warfare quite often coincides with periods of intensified warfare. The term 'war-related violence' can often serve as a preliminary classification or acceptable compromise when archaeological data do not allow a precise categorization of the kind of violence involved.

The fighters themselves, the warriors and their leaders, are the chief actors of warfare whose roles rely on a considerable amount of movements across landscapes and at the places of fighting. Warriors are the special agents, even in some cases literally professionals, who have been trained in the bodily material techniques of warfare. War and warfare invariably produce death, suffering, and trauma, and warriorhood can therefore be defined as an identity, or profession, that domesticates the associated fear through a learnt scheme of skills and techniques, stored in brain, body, and material culture in complex not divorceable ways.

The term 'warriorhood' is another key to understanding warfare in the past: it can be defined as an individualized social identity closely connected to military actions, but also motivated by myth and ideologies of men, glory, and war. Because it requires a physical risk and offers a chance to display individual courage and bravery – warfare is in many societies acknowledged as a route to social success in that it may generate social capital suitable to invest in social strategies of power. Males may from early boyhood be socialized into the warrior role, as in many parts of colonial Africa, Melanesia, Australia, and South America. Here warfare was not necessarily conducted from firm military units, but able males were momentarily committed to the task and engaged in fighting on ad hoc basis. When, however, warriors organize in institutions of war, so-called warbands, they do it in three different ways with access to the 'club' based on principles of age (age grades) or social prestige and personal status, and in a number of instances also distinctions of rank and class. In this latter case recruitment happens from the dominant societal level or followers are more broadly appointed by a warleader of high social standing. The aristocratic warband with special codes of conduct (*Gefolgschaft*) is described in numerous ethnohistorical cases of chiefdoms and archaic states as well as in the ancient written sources from Tacitus' *Germania* to the epic classics of the Rigveda, Homer, and Beowulf; these are all rooted in Bronze Age and Iron Age worlds and therefore important for the understanding of warfare from the third and second-millennia BCE onward.

Other agents and groups than strictly warriors may be influential in warfare, such as victors, collaborators, allies, enemies, victims, civilians, and ritual specialists. Such groups may interact with each other in a flow of interactions that also incorporate material objects and environments in ways that impact society. Furthermore, warfare implies targeted use of weapons, which may nevertheless have functions in other social domains than war.

In the following the main categories of archaeological sources to warfare will be centrepieces in discussing the essences, scales, and organizations of war, weighing its material components and main actors as well as the often considerable

ideological input. Emphasis will be on war before history. Warfare as defined and outlined above permits us to resolve that war-related violence has occurred occasionally from the first human beginnings and throughout history and has periodically taken a predominant position in everyday life. The (co)emergence of institutionalized warriorhood and weapons made solely for war constitutes important watersheds in the history of warfare and ancient society.

Weaponry

Weaponry divides arbitrarily into implements with a potential for war and real weapons intended for offensive and/or defensive purposes. Bows and arrows, points, axe blades, knives, and daggers, made of various organic or nonorganic materials, belong to the first category. In fact, almost any object can find uses as weapon if need be, and this was certainly the case from the earliest times and throughout the history of mankind. The second category comprises mainly swords, spearheads, shields, and body armor, but also maceheads and battle-axes. These objects were definitely produced with warfare in mind using selected materials to effectuate this purpose. Weaponry for war is a late phenomenon measured on the background of millennia of human history. In Eurasia battle-axes and maceheads began to appear only after the introduction of farming, and particularly in Europe more systematic production and uses of these weapons are in particular attached to the so-called Corded Ware period dating to the onset and first centuries of the third-millennium BCE and which seem to link up with spread of Indo-European languages, wheeled vehicles, and probably horses from a homeland in the north Kazakh steppes. The early third-millennium BCE also brings us the first clear evidence of warband institutions: Organized frames for warfare and warriorhood.

Weapons for war typically occur in burials and other ritual contexts and they constitute a large category of finds from later prehistory and onward, hence testifying to the presence of warfare within a broader social domain. The appearance of these first weapons – made unequivocally for waging war and often made in copper, bronze, or spectacularly colored stone – in many places coincides with the emergence of marked social distinctions with cogent forms of leadership and discrete organizations of war. Throughout, the distinction between would-be weapons made in any material at hand, weapons such as bow/arrow with other uses, and genuine weapons made in special materials seems concurrent with social distinctions in society as such and with distinctions in warfare between war leaders, high-ranking warriors, and common fighters. With the mature Bronze Age from 1700 BCE, or somewhat earlier in some regions, this picture becomes particularly distinct sustained by the innovations of spearheads and swords and adjacent novel techniques of fighting in the upper echelons of society. In the Final Neolithic (Copper Age) and particularly in the Bronze Age, the group of genuine weapons grew markedly in numbers and in technical and ornamental elaboration. This seems tied to the general inventiveness of these periods, but almost certainly also to reorganizations of institutions of war and new ideologies to support them. Although the presence of warrior clubs is rather strongly indicated from around 2800 BCE in temperate Europe, their more precise position, central or marginal, in the societal meshwork is unknown, but could have varied enormously in the period 3000–500 BCE. In the Bronze Age the warband seems to grow in importance and its internal organization to be based on hierarchical principles.

The above typology of weaponry roughly suggests that most weapons have potential uses outside warfare, notably hunting. Bone assemblages from Final Neolithic and Bronze Age settlements normally show little inclusion of wild animals, and it can thus be presumed that hunting in these periods, and most certainly in the Bronze Age, was more a matter of prestige than economic necessity for at least the privileged part of the population. The boundary between prestige hunter and warrior can be blurred, as famously illustrated by the lion-hunt dagger from Shaft Grave IV at Mycenae. Dating to *c.*1600–1500 BCE, it shows warfare against the lion. This is likely a metaphor for bravery in combat and for the princely warrior chieftain in early Mycenaean society. In the thirteenth-century BCE, this symbol was appropriated by the royal house of Mycenae as their 'coat of arms,' visualized so vividly by the Lion Gate to the citadel.

The multifunctionality, or ambiguity, of material culture in general implies that weapons cannot be reduced merely to tools of war. A series of other potential functions and meanings presents itself. Notably, weapons often maintain identities of gender, age, rank, or various other types of group identity like kinship, ethnicity, and profession. In European prehistory the cultural biographies of weaponry typically ended with a ritual deposition either in burials or in votive offerings. Such depositions may comprise one item or a combination of several items depending on cultural and social factors. From the Bronze Age of temperate Europe and in several similarly organized societies around the world numerous weapons exist; the sheer quantity being ample testimony to warfare as an option that could be acted upon and indeed to the presence of warriorhood as an identity feeding from bellicose interaction and from ideologies of war. Macro- and microwear analyses of swords and other weapons confirm that many have indeed been put to repeated and forceful use. It is often speculated that some of the most spectacular weaponry: shields, helmets, giant axes, and body armor – for reasons of construction – cannot have been very suitable for warfare, but again such splendid items likely had several social uses in addition to warfare. Such weapons may have had ritual functions in funerary games, religious performances, and/or were parade armor that should impress and scare. This may be supported by the pictorial slabs from the Kivik tomb in Scania from around 1500 BCE, rock-carving scenes from various places in the Eurasian Neolithic and metal ages, and by Patroklos' funeral games in Homer's *Iliad*. By contrast, a substantial series of other weapons exists that were surely meant to wound and kill other human beings: spears and arrows definitely belong here, as do the swords. Especially the swords are central to our understanding of the Bronze Age and later prehistory/history as they mediate tales of supreme, active, and symbolic warriorhood among privileged strata in society. However, albeit less prestigious, spears and in particular arrows were key weapons for attack and defense, as shown by their massive presence in the early battlefield of Tollense near Greifswald in north Germany dating to around 1200 BCE. Projectile points of flint, and later in metal, sitting

in situ in human bone constitute a frequent form of skeletal trauma and cause of death throughout later prehistory, especially from Neolithic farming communities in the ninth to eighth millennia BCE and onward.

The exquisite and highly specialized craftsmanship of many of these weapons of metal suggests organized warriorhood in warbands. It would also support the suggestion that such warrior institutions were recruited and organized internally according to hierarchical principles, and thus not unlike the system that can be deduced from the *Iliad*, from Tacitus' *Germania*, and from a large number of ethnographically studied cases. A material hierarchy of weapons arguably existed, which was differentially constituted in different periods; in the later Bronze Age from *c.*1200 to 700 BCE likely with metal-hilted swords and body armor in the top part, followed by organic-hilted swords, axes, spearheads and arching equipment lower down, and wooden clubs and similar items in the bottom part. Such hierarchies of weaponry may be suggested to reflect social divisions within warrior institutions.

Weapons can be highly individualized objects – swords usually are – and thus invested with personalized names and meanings while simultaneously mediating in variable measure gender, status, rank, kinship, ethnicity, profession, and so on. Weapons may have whole biographies and stories attached to them, which could have been invoked on special occasions. The famed boar's tusk helmet which Odysseus puts on for a nightly spying expedition is a case in point, which could very well generally apply when looking at the finest prehistoric weapons, such as the twin horned helmets of bronze (*c.*1100 BCE) found in a bog at Viksö in Denmark. Such particular weapons are perhaps likely to be inalienable possessions kept in the family as tokens of memory and inheritance or ritually exchanged with the gods in sacred places or with very particular ancestors in monumental tombs.

The potential multifunctionality of material culture thus implies that weapons cannot be reduced merely to being war implements, but became intertwined with other social uses, as also suggested by contextual evidence. Even the most frightening or deadly of weapons very often had alternative social functions outside the strict sphere of waging war. Albeit the spotlight has here been on weaponry and what it can tell about warfare and its organization it is equally true that countless archaeological data exist that inform about peaceful moments, domains, and times.

The cultural biographies of weapons typically ended with a ritual deposition either in burials or in votive offerings. Such depositions in later European prehistory comprised one item or a combination of several items depending on cultural and social factors. The archaeologist thus obtains intimate knowledge mostly of the concluding activities, the ritual death of things, which may have had several very active lives prior to the final deposition. These final destinations of weaponry, however, may still be able to deliver information about the living society, as outlined above and substantiated below.

Weaponry in Ritualized Contexts

Weapons and other objects were from early on systematically deposited in burials and at other ritual sites. This may have happened for a variety of overlapping reasons such as emotion, social ambition, social rivalry, and social identity, commemoration of persons and events of war, and to sustain cultural values and norms, including religion. Ritual actions such as funerals and offerings can be carried out routinely as the way things should be done or more deliberately as political announcement questioning or legitimizing existing relations of social power.

Numerous graves with weapons are known from different periods all over the world, but their direct or indirect relationship with warriorhood has only just begun to be studied. From the Final Neolithic in temperate Europe – the Corded Ware and Bell Beaker periods – good cases can be made of warrior identities and institutions although allowing for cultural variations. From the Bronze Age and Iron Age one can easily identify conspicuous individual graves with rich weaponry attached to mature males who may have been high-ranking warriors and warleaders, perhaps still actively engaged in war and/or this role was mainly effected on a symbolic level as heads of society. Females were not usually buried with weapons, albeit notable exceptions exist such as Scythian–Sarmatian interments, so-called Amazons, at the Don and Volga. Women's roles in warfare – to judge from written sources – probably were more active and motivating than often assumed.

It seems in many cases clear that the deceased were commemorated as warrior heroes by kin and community in the same way as warriors and soldiers have been celebrated and remembered in history and in our own time. Such presentations of the warrior's 'beauty,' glory, and valor are of course strongly ideologically colored, contrasting with the knowledge that the warrior is an identity also sustained by violence. Grooming items such as razors, combs, awls, and tweezers in addition to ornaments for dress and body very often accompany elaborate weaponry in so-called warrior burials; hence tallying with a cross-cultural inclination of warriors to be preoccupied with their bodily appearance. Presentations of warrior identities, as mediated in burials and other ritual context, are rooted in specific historical events but also in myths of war and its heroes. This also exemplifies ancient social strategies of distancing 'us' from 'them.'

The existence of elaborate weapons in metal dividing into several categories may suggest that warriors in later prehistory organized in warrior institutions. This is in general harmony with the funerary evidence, which is however also a very complex data source. Weapons and dress accessories in burials may be a direct statement of the roles played by the dead individual in life. Swords seem in the Bronze Age very rarely associated with child or female interments, and sword burials would then be generally suggestive of male warriorhood. Yet, weapons and dress accessories may additionally be used to make more general metaphoric statements and thus have a less direct bearing on the lived identities of the deceased. This may include symbolic warriorhood in the sense of an earlier warrior identity, heroic status or ambition, political authority, or high social rank; all of it possibly in combination. Likewise, weapons accompanying small male children and young adolescents, such as those found in some Corded Ware communities between 2800 and 2500 BCE are likely to be metaphors originating in warrior values rather than signs of active warriorhood.

The inhumation cemetery at Neckarsulm in Baden-Württemberg dating to *c.*1300 BCE, the early Urnfield period of the Late Bronze Age, is interesting in terms of active warriorhood. The majority of the 51 bodies turned out to be adult males of excellent health and physique, bow-legged as a consequence of horse-riding. Most had standard equipment of fine pottery and body accessories, but were surpassed materially by three sword graves situated close together at the eastern margin of the cemetery. Many burials contained several interments of males who apparently died on the same occasion. It is fairly clear from this case that the cemetery was devoted to warriors, that is, the buried males were active warriors who died and were buried companion wise, and finally that the warrior institution(s) had an internal ranking defined by those with swords and those without swords. A similar picture is conveyed by the numerous skeletons of young adult males at the above-mentioned site of Tollense, also containing horse skeletons, deposited under circumstances of battle and, fairly certainly also as a consequence of elaborate and violent postbattle rituals in the adjacent river.

As warfare is definable as a flow of social interactions, it needs to be somehow initiated and concluded, and rituals on individual and group level present themselves as obvious choices. Indeed, a large number of weapons are known from sacrificial deposits all over northern and central Europe, and far beyond these regions, and many of them have been retrieved from wetlands. This tradition of ritually depositing weapons and other valuables commenced in the first farming communities but grew over time in significance and likely became linked to beliefs in otherworldly existences. This specific ritual practice has been explained with reference to several overall purposes: rites of passage, gifts to the gods, and 'potlatches' to promote the importance of particular persons or social institutions and to legitimize or question power and authority. Among the Northwest Coast Indians, potlatch rituals only became truly elaborate when the Canadian government prohibited warfare, hence potlatch in this case became a substitute for waging war.

It seems that warfare in prehistoric Europe could have been systematically linked with postwar rituals finalizing armed conflict or they were somehow interwoven with ongoing warfare. This accords well with the above-mentioned Bronze Age site at Tollense from *c.*1200 BCE, which has yielded weapon offerings and numerous skeletons along the stretch of the river. The rich and elaborate war booty offerings of the north European Iron Age and the commemorative deposits of weapons from Greek shrines throughout Antiquity constitute similar cases of postwar rituals. Ancient authors from Tacitus to Pausanias amply describe such depository traditions relating to war, which even today is intertwined with rituals. While Neolithic depositions contain mainly implements with the potential for war, the more monofunctional weapons in depositions from the Bronze Age make an association with warfare more clear. War booty offerings, a generally accepted find category in Iron Age Europe sustained by written sources, could well be rooted in the Neolithic and especially the Bronze Age. While warfare cannot in general be understood as a ritual, it nearly always contains aspects of ritual. However, it is important to note that prehistory all over the world is replete with funerary and other ritual deposits containing mundane and extraordinary items with peaceful purposes.

Weapon Combinations and Technologies

Weapon combinations in wetland depositions and in cemeteries can, used with caution, inform about the internal organization of the warrior bands and of the equipment and fighting methods of the warrior. The warrior may not be interred with his full equipment but rather only those parts of the equipment that had a certain symbolic meaning or could be spared by the living. Particular weapons may have been curated, rather than deposited, for any number of reasons; perhaps kept to be used in future wars and in effect transmitted from father to son as insignia of rank and warrior status. Close combat armament predominates in burials and hoarded depositions, which should not let us conclude that archery was unimportant during fighting, as the continued occurrence of arrows of flint and metal shows. Special eye openers in this respect are constituted by settlement sites from the early Neolithic onward showing numerous arrowheads in the banks and ditches of the fortifications. Similar evidence comes from skeletal trauma showing marks from clubs as well as wounding and death by arrow shots.

The social presentation of the warrior in funerary contexts varied systematically through time indicating that fighting methods also changed. In the mid-third-millennium BCE, in temperate Europe and somewhat earlier toward south and east, the battle-axe-wielding warrior gave way to the dagger-carrying archer, who in the mid second millennium at the threshold to the Middle Bronze Age became a sword, axe, and spear fighter with several fine toilet articles to groom hair and body. Battle-axes and other axes of potential use in war were still available, as were less conspicuous and less prestigious weapons made of wood and other organic materials. It is quite possible that body armor and shields of organic materials were in use and forming part of the full panoply of the elite warrior. Wheeled horse-pulled vehicles were also present, perhaps from the onset of the third-millennium BCE but grew in significance in the metal ages. From around 1300 BCE, weapons look increasingly effective and standardized with elite warrior equipment consisting of the whole suite: sheet bronze armor, helmet, round shield, sword, and spear. This was in addition to toilet implements for grooming and now more often a four-wheeled cart or a two-wheeled chariot. The wheeled vehicle was to bring the war chieftain to and from battle and to present him as a superior being on social and religious occasions. This impressive assortment of weaponry shows developments over time with new highlights in the Iron Age with tribal confederations, archaic states, and finally the empire of Rome. All along this trajectory, there are regional variations in the techniques of fighting and the employed weapons, but the general development is similar.

Weapon technology is not unimportant for the outcome of war. Improvements have been known to escalate the degree and viciousness of warfare in numerous ethnographic cases. It is thus possible that innovations in the technology of war could have increased the speed of social change in prehistory. A comparison of weapons over time can give clues to significant

changes in weapon technology, understood as the quality of weapons and styles of fighting. The emergence of the sword around 1600 BCE in central and northern Europe is likely to have been such an important innovation with considerable social and symbolic effects in warlike as well as in more peaceful interactions inside society and between societies. Likewise, the introduction of horses for driving and riding, more systematically around 1200 BCE, would have changed conventional warfare as well as the whole social habitat, as documented, for instance, among the Grand Chaco Abipón in South America during the eighteenth century.

Fortifications

All over the world, regions with fortified settlements and villages can be singled out from the Neolithic onward and progressively more frequently, while tending to concentrate in periods of radical social transformation. This latter trend can be exemplified by the situation in central Europe and the Balkans between 1700 and 1500 BCE and again around and after 1200 BCE when large fortified strongholds with elaborate embankments were built at elevated points in the landscape. The Cyclopean strongholds at Mycenae, Tiryns, and other Mycenaean sites in Late Bronze Age Greece constitute very elaborate examples adhering to palace-organized archaic states. Likewise, hillforts occur throughout the Iron Age, as testified by the Celtic-tribal strongholds at Manching and at Alecia, which were conquered with difficulty by the Romans during the long military campaigns of Julius Caesar. In other regions, such as southern Scandinavia, unfortified settlements are the rule throughout prehistory. It seems that the presence or absence of fortified settlements cannot be used as evidence of war and peace, but need to be compared with other data. Absence of fortifications might be rooted in different sociocultural practices rather than peace as a prevalent state of affairs.

Fortifications are as multifunctional as other kinds of material culture. Fortified settlements are sometimes urban or protourban hubs for a limited region or even nodal points for superregional trade and crafts. These are peaceful activities denying daily life to have been intrinsically and chronically violent. At other times, fortifications are marginal phenomena placed at the boundaries of settled areas to scare off the enemy and to protect against intruders. In the latter case, they may not be permanently settled, or perhaps only with certain specialized personnel. Fortifications and territorial marking go well together, but distinct territoriality can exist without fortifications.

Skeletal Trauma

Traumata on skeletons can be caused by other forms of interpersonal violence than warfare, but are in many cases war related as suggested by ethnohistorical cases. Skeletal traumata are, in fact, relatively frequent in prehistory when it is taken into account that in some areas skeletons are not well preserved, that they are not routinely examined for marks of violence, and that much physical violence does not leave visible traces on the skeleton. A very long list of incidences of trauma exists, probably starting with early hominids and definitely continuing into the later periods of the Palaeolithic, Mesolithic, Neolithic, and so on, with a seemingly unbroken line to recent times. Bone projectile points embedded in the spine of a child (Homo Sapiens) occur as early as 30 000 BCE at Grimaldi in Italy and at *c.*18 000 BCE in the Nile Valley. Somewhat later, at the Mesolithic cemetery *c.*11 000 BCE at Gebel Sahaba, also in the Nile Valley, over 40% of the 50 individuals had died from arrow wounds.

Many cases cluster at the transition to the Neolithic when new forms of socioeconomies developed. These instances often occur in heavily ritualized contexts such as at Talheim, dating to *c.*5000 BCE, in Germany. Here 34 men, women, and children displaying perimortem trauma of exaggerated and highly brutal character had been buried with some form of structure in a mass grave. A similar and roughly contemporary case occurs at Herxheim, also in Germany, where an enclosured site with ditches contained the remains of numerous individuals. Some were disembodied in a ritualized manner as exemplified by decapitated skulls. Such abnormal mass graves with several injured individuals occur occasionally during later periods. Normal single interments with one injured individual also occur. In addition, multiple burials are known with a more normal pattern following prescribed 'normal' funerary rituals. A case in point is Eulau in central Germany, where a Corded Ware burial *c.*2600 BCE contained a man and a woman and two children, all with fractures in skulls, forearms, and hands and with arrowheads still sitting in the bones. These injuries had in all likelihood resulted from a hostile raid on a nearby settlement. Several similar cases are known distributed across time, as shown by, for example, the Crow Creek massacre in South Dakota with the remains of *c.*300 individuals dating to around 1325 CE. In general, injuries vary from projectile wounds to marks and cuts from weapons. Some wounds occur healed up, while other wounds were the cause of death.

Skeletal trauma is a valuable source of additional information when it comes to detecting warrior identities and to assess the frequency and scales of warfare. The considerable number of skeletons from *c.*1200 BCE at Tollense suggest that at this point in time large-scale warfare took place involving not only warbands but entire armies, perhaps consisting of individual warbands such as those described in Homer's *Iliad*. Postwar rituals involving further violence probably also took place at this and other Bronze Age sites.

Interestingly, looking particularly at European sources, there is a rather clear tendency for traumatized skeletons to cluster in turning points periods, such as at the transition to the Neolithic in many regions, the Corded Ware period commencing *c.*2800 BCE, and later on *c.*1700–1500 BCE and again *c.*1200 BCE. This pattern of clustering is interesting from the point of view of situating war within the causes and effects of radical sociocultural change. Moreover, it strongly sustains the cultural, rather than biological, essence of hostile encounters and war.

Iconographic Representations in Art and Rituals

Representations of martial arts, myth, and ideologies occur carved into rock, painted on walls and pottery, and printed into

vessels of ivory, stone, and metal. The date and social context vary enormously as do the geographical distribution and the representational media. Scandinavian Bronze Age rock carvings with scenes of ships, chariots, and warriors – among other themes – are clearly of mythical nature but surely also embedded in real historical events of long-distance maritime journeys, likely involving both trade and raids between 1600 and 500 BCE. Similar rock panels occur in Spain and northern Italy from the Neolithic and the Bronze Age, including fighting scenes with bows and arrows and tactical formations of advance and retreat among groups of warriors. Such carvings occasionally show prevalent weaponry of a given period.

Warlike presentations cluster in the representational art of Mycenaean Greece, where painted war scenes adorn the walls of central halls in the palaces while bellicose imagery, including siege scenes, warrior processions, and chariot driving are also depicted on stelae and vases. This is clearly the art of a martial elite. Scenes of warfare even occur in quite different social settings such as the fighting scenes carved on rock in southern Africa and attributable to hunter-gatherers of the San Bushmen. From quite another context and period, Classic Mayan wall paintings at Bonampak, Mexico, represent a battle between warriors as well as postwar rituals of beheading prisoners. Such representational art may be quite differently motivated, but is glued together by warfare as well as ideologies in support of these violent encounters. Martial representations from prehistory and onward should nevertheless be measured against the bulk of iconographies with peaceful scenes.

Concluding Reflections

A situation of reciprocity between data and theory has not yet emerged in archaeology because the study of warfare and violence is a late arrival of the past decades. Recent archaeological studies, in concert with recent trends in anthropology, permit the proposition that the level of warfare has varied with time and place. Even those more extreme situations where war has grown to become habitual are far removed from the Hobbes-inspired belief that human beginnings were intrinsically violent and chaotic. These new and emerging insights are likely connected to the combination of theoretical frameworks with cautiously produced data patterns in archaeology: so-called endemic, chronic, or total war has been exaggerated as a historical phenomenon. In addition, it is difficult to find empirical or theoretical support for the evolutionist hypothesis that warfare (or peace) gets more common through time. Similarly, it can be argued, based on present knowledge and hence with some weight, that war-related violence is mostly rooted in cultural logics rather than in biology, genetics, or evolution.

See also: Peace and Nonviolence: Anthropological Aspects; War, Social Causes and Consequences; Warfare in History.

Bibliography

Carneiro, R.L., 1970. A theory of the origin of the state. Science 169, 733–738.
Chacon, R.J., Mendoza, R.G. (Eds.), 2007. North American Indigenous Warfare and Ritual Violence. The University of Arizona Press, Tucson.
Childe, V.G., 1941. War in prehistoric societies. The Sociological Review 33, 133.
Fukuyama, F., 2011. The Origins of Political Order. From Prehuman Times to the French Revolution. Profile Books, London.
Helbling, J., 2006. War and peace in societies without central power: theories and perspectives. In: Otto, T., Thrane, H., Vandkilde, H. (Eds.), Warfare and Society. Archaeological and Social Anthropological Perspectives. Aarhus University Press, Aarhus, pp. 115–141.
Hobbes, T., 1651[1958]. Leviathan. Bobbs-Merrill, Indianapolis.
Jabri, V., 1996. Discourses on Violence. Conflict Analysis Reconsidered. Manchester University Press, Manchester and New York.
Janzen, D., Brinker, U., Orschiedt, J., Heinemeier, J., Piek, J., Hauenstein, K., Krüger, J., Lidke, G., Lübke, H., Lampe, R., Lorenz, S., Schult, M., Terberger, T., 2011. A Bronze Age battlefield? Weapons and trauma in the Tollense Valley, north-eastern Germany. Antiquity 85 (328), 417–433.
Keeley, L.H., 1996. War before Civilization. The Myth of the Peaceful Savage. Oxford University Press, New York and Oxford.
Komar, D., 2008. Patterns of mortuary practice associated with genocide. Implications for archaeological research. Current Anthropology 49 (1), 123–133.
Meller, H., Muhl, A., Heckenhahn, K., 2010. Tatort Eulau: Ein 4500 Jahre Altes Verbrechen Wird Aufgeklärt. Theiß, Stuttgart.
Neth, A., 2001. Ein aussergewhönlicher Friedhof der Urnenfelderzeit in Neckarsulm, Kreis Heilbronn. Archäologische Ausgrabungen in Baden-Württemberg, 51–55.
Otterbein, K.F., 2004. How War Began. Texas A&M University Press, College Station, United States.
Peter-Röcher, H., 2007. Gewalt und Krieg im prähistorischen Europa. Beiträge zur Konfliktforschung auf der Grundlage archäologischer, anthropologischer und ethnologischer Quellen. Universitätsforschungen zur prähistorischen Archäologie. Verlag Dr. Rudolf Habelt GMBH, Bonn.
Potter, J.M., Chuipka, P., 2010. Perimortem mutilation of human remains in an early village in the American Southwest: a case for ethnic violence. Journal of Anthropological Archaeology 29 (4), 507–523.
Schulting, R., 2006. Skeletal evidence and contexts of violence in the European Mesolithic and Neolithic. In: Gowland, R., Knüsel, C. (Eds.), The Social Archaeology of Funerary Remains. Oxbow, Oxford, pp. 224–237.
Schweizer, T., 1990. Männerbünde und ihr kultureller Kontext in weltweiten interkulturellen Vergleich. In: Völger, G., von Welck, K. (Eds.), Männerbande-Männerbünde. (Ethnologica, Neue Folge, 15). Joest-Museum der Stadt Köln, Cologne, pp. 23–30.
Vandkilde, H., 2006a. Warriors and Warrior institutions in the European Copper Age. In: Otto, T., Thrane, H., Vandkilde, H. (Eds.), Warfare and Society. Archaeological and Social Anthropological Perspectives. Aarhus University Press, Aarhus, pp. 393–422.
Vandkilde, H., 2006b. Warfare and pre-state societies – an introduction. In: Otto, T., Thrane, H., Vandkilde, H. (Eds.), Warfare and Society. Archaeological and Social Anthropological Perspectives. Aarhus University Press, Aarhus, pp. 105–111.
Vandkilde, H., 2007. Culture and Change in Central European Prehistory, 6th to 1st Millennium BC. Aarhus University Press, Århus.
Vandkilde, H., 2011. Bronze Age warfare in temperate Europe. In: Hansen, S., Müller, J. (Eds.), Sozialarchäologische Perspektiven: Gesellschaftlicher Wandel 5000–1500 v.Chr. zwischen Atlantik und Kaukasus. Internationale Tagung 15–18 Oktober 2007 in Kiel. Archäologie in Eurasien 24. Verlag Philipp von Zabern, DAI-Berlin & Darmstadt, pp. 365–380.

Conflict and War: Anthropological Aspects

Douglas P Fry, University of Alabama at Birmingham, Birmingham, AL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although exceptions exist, war is usually differentiated from feuding within anthropology. A useful definition of war is specific enough to differentiate warfare from nonlethal brawls, single homicides, and kin-based feuding. Cross-cultural surveys show that a majority of societies practice warfare or feuding. At the same time, nonwarring societies also exist, sometimes as members of a nonwarring peace system. The occurrence of war and the reasons for warfare are correlated with the degree of social complexity, both archaeologically and ethnographically. Warfare has originated rather recently, within the last 10 000 years or so, in parallel with a worldwide shift away from a nomadic forager existence toward more complex social forms.

Introduction

> Each [Blackfoot] raiding party was hastily organized before departure and disbanded immediately after its return home …. Its members were motivated much less by tribal patriotism than by hope of personal gain – the economic security and social prestige that possession of a goodly number of horses would bring them.
>
> Ewers, 1967: 329

> Indeed, what the Gilyak called wars in ancient times were in fact nothing more than sporadic clan skirmishes motivated by vengeance or, even more frequently, were over women.
>
> Shternberg, 1933: 247

> Warfare among the Fijians was all-out and bloody, with no respect shown for sex or age. Women and children were killed ruthlessly and indiscriminately.
>
> Carneiro, 1990: 199

This article considers central questions regarding the anthropology of war. How is war defined? How does war differ from feud? How frequent are warfare and feuding cross-culturally? When did war originate, or in other words, how old is the institution of war? Do some kinds of human societies engage in more warfare than do others? What does anthropology suggest are the reasons for war?

Defining War

The way in which war is defined has direct relevance for answering questions about its origin, frequency, and distribution. Definitions of war vary from broad to narrow. If war is defined broadly as 'lethal aggression' (Low, 1993) without any specification, for example, that it pits one group against another, then homicides within or between groups could be counted as acts of warfare. Most scholars do not consider such behaviors to constitute warfare (Ferguson, 2013a,b; Sponsel, 1998).

Otterbein, who has been studying war, feuding, and other forms of violence since the 1960s, makes a distinction between feuding and war (Otterbein, 1968, 1970, 2004). He defines *feuding* as blood revenge that follows a homicide, a pattern of reciprocal killings, and he defines *war* as "armed combat between political communities" (Otterbein and Otterbein, 1965: 1470; Otterbein, 1970: 3). Like Otterbein, Boehm (1987: 221) distinguishes between feud and war: "Feuding is quite different from all-out warfare for several reasons: feuding is limited to one or two killings at a time; only one side takes the offensive at a time; and there is no necessary political objective beyond the maintenance of honor." Correspondingly, van der Dennen (1995: 93) writes: "in war – as contrasted with a feud – violence is relatively 'promiscuous': Anyone of the opposing community may be defined as an 'enemy' and thus be a potential victim. In a feud, the concept of enemy – and potential victim – is more restricted and selective: A particular member or members of a particular kinship group, family or clan."

On the other hand, Ember and Ember (1992: 172) lump together feud and warfare – a decision that leans toward maximizing the cross-cultural tally of warring societies – defining war as:

> … Socially organized armed combat between members of different territorial units (communities or aggregates of communities). In the ethnographic record, such combat usually involves groups on both sides, but a warfare event could involve the ambush of a single person of an 'enemy' group. Thus the phrase 'socially organized' means that there is a group of combatants on at least one of the 'sides.' This definition of warfare focuses on the behavior of groups, not on their presumed or reported motives for attacking and killing. Thus, we do not distinguish between warfare and feuding ('blood revenge').

Bowles (2009: 1294; see also Bowles and Gintis, 2011) recently has defined war so broadly as to not require any fatalities: Wars are "events in which coalitions of members of a group seek to inflict bodily harm on one or more members of another group." One problematic feature of this definition is that it is so inclusive that it could be applied to soccer hooliganism, riots, or banditry perpetrated by several persons. Most people do not consider such types of aggression to meet the definition of war. In contrast to overly inclusive definitions, Prosterman (1972) proposes a definition of war that

 http://dx.doi.org/10.1016/B978-0-08-097086-8.96031-8

corresponds with common usage and that differentiates war from homicide and feud. For Prosterman (1972: 140), war is:

> A group activity, carried on by members of one community against members of another community, in which it is the primary purpose to inflict serious injury or death on multiple nonspecified members of that other community, or in which the primary purpose makes it highly likely that serious injury or death will be inflicted on multiple nonspecified members of that community in the accomplishment of that primary purpose.

Prosterman's definition clarifies that war is a group activity between communities directed against nonspecified members of the other community, which makes likely the killing of multiple people. Prosterman's (1972) definition of war is more detailed than Otterbein's (1970: 3) "armed combat between political communities," but these two definitions of war are in rough correspondence. Prosterman's definition of war also excludes individual homicides, even if committed by more than one person, in contrast to the definition used by Bowles (2009) or Ember and Ember (1992).

The Cross-Cultural Frequency of War

How prevalent are feuding and war cross-culturally? Ericksen and Horton (1992) report that just over half of the societies in a large cross-cultural sample engaged in feuding. They also found that feuding varied across societies from being perceived as a moral imperative to being viewed only as the recourse of last resort. In a smaller sample, Otterbein and Otterbein (1965) also found approximately half of the societies to engage in feuding.

Turning to warfare, Otterbein differentiates between *internal war*, between political communities of the same culture, and *external war*, between political communities of different cultures, and he reports that the vast majority of his sample engaged in either internal or external warfare. Eight percent of his sample of 50 societies did not engage in any type of war. For a sample of nonpacified societies, Ember and Ember (1992) found warfare to be absent or rare in 9% of the sample.

Focusing on North America, Driver (1969: 310) concludes that feuding sometimes existed, but "most of the peoples of the Arctic, Great Basin, Northeast Mexico, and probably Baja California lacked true warfare before European contact." Additionally, Driver (1969: 312) reports that whereas some societies did make war in the immense areas including the North American Sub-Arctic, the Northwest Coast, the Plateau Region of the Northwest, California, and the Southwest, "at the same time, all these areas included some peoples with little in the way of violence, raids, or feuds, and no hostilities pretentious enough to be labeled war." Jorgensen et al. (1980: 241) concluded on the basis of surveying 172 societies of western North America:

> One of the most obvious and interesting aspects of the cultures of western North America's Indians at the time of contact with Europeans was that so few societies actually engaged in persistent offensive warfare, or even raiding, yet the prospects of armed altercations deeply influenced the internal organizations and external relations of these aboriginal societies.

Based on the examination of 590 societies from around the world, Wright found that 64% of the sample either lacked warfare or engaged in only feuding and mild forms of warfare. Wright's findings thus show the cross-cultural spectrum of human societies to be less warlike than commonly assumed.

Fry (2006) lists over 70 cultures from around the world that do not practice any type of warfare. For example, from Malaysia, "there is no evidence of Semang warring with one another or with non-Semang" (LeBar et al., 1964: 185). The Saulteaux of North America, "have never engaged in war with the whites or with other Indian tribes" (Hallowell, 1974: 278). On the Hanunóo of Southeast Asia: "There are no classes, no servants, no officials, and no warfare" (Conklin, 1954: 49). For the Machiguenga of South America, "there is no warfare in their region" (Johnson, 1983: 61, 63). And Gardner (2004: 58) states "I can report a complete absence of feuding within Paliyan society and a corresponding total lack of warfare." Some clusters of neighboring societies do not make war with each other and sometimes not with any outsiders either. These groups of nonwarring societies exist in various quarters of the globe, from Canada to Malaysia, and have been labeled *peace systems* (Fry, 2006). Clearly war is not a human universal.

How Old Is War?

There are two schools of thought regarding the antiquity of war. One view holds that war is ancient, going back hundreds of thousands if not several million years (Bowles, 2009; Bowles and Gintis, 2011; Pinker, 2011; Wrangham and Peterson, 1996). The other perspective is that war is recent, originating within the last 10 000 years or so (Fry and Söderberg, 2013; Haas, 1996, 1999; Haas and Piscitelli, 2013; Kelly, 2000; Roper, 1969).

The worldwide archaeological evidence favors the latter perspective. With the exception of one possible case, the earliest evidence of warfare worldwide is within the last 10 000 years (see the range of dates in Pinker, 2011, Figure 2.2, for example). Furthermore various archaeological sequences show multiple regional origins of war from conditions of warlessness, all within the last 10 000 years (Dye, 2009, 2013; Ferguson, 2013b; Flannery and Marcus, 2003; Haas, 1999; Kelly, 2000; Maschner, 1997; Roper, 1975). Although warfare sometimes occurred among pre-state civilizations within the last 10 000 years, war became more common and intensified with the rise of the state between 4000 and 6000 years ago.

Warfare leaves its marks on the archaeological record (Ferguson, 1997, 2013b; Haas and Piscitelli, 2013). Daggers, clubs, and other specialized weapons not used for hunting; fortifications surrounding settlements; numerous interments with projectile points embedded in the bones or resting within ribcages; depictions of war in rock art; evidence that immediately following the burning of a site a new cultural style replaced the previous style of artifacts; and repetition of the forgoing diverse bodies of evidence across the archaeological sites of an area: All these types of evidence demonstrate that war was present. When multiple lines of data point in the same

direction, we can be highly certain that warfare was occurring. On the other hand, the fact that war leaves unambiguous indicators means that when we have a detailed archaeological record for an area that contains no evidence of warfare, then this shows that war was absent.

Roper (1969: 448) reviews the skeletal finds for early humans – or hominids – over the last 3 million years, and she reaches the conclusion that sporadic homicides occurred but there is no evidence of warfare in the deep past. In the absence of any corroborating evidence of warfare such as defensive sites or specialized weapons, a single victim of lethal aggression does not justify the assumption of warfare (Ferguson, 2013a). A single death could represent a murder, a group sanctioned execution, or even a hunting accident (Fry, 2006), not war. Pinker (2011, Figure 2.2), following Keeley (1996), counts several single deaths as indicative of warfare. This practice contributes to an inflated impression of war in antiquity.

Turning to the evidence, the prehistoric site with the earliest indications of possible war anywhere in the world is Jebel Sahaba in Nubia. In the Jebel Sahaba cemetery, dated between 12 000 and 14 000 years BP, 24 out of 59 skeletons display evidence of violence. Clearly this is a very high frequency of violence. Some writers attribute these violent deaths to warfare, whereas other researchers caution that homicides, executions, and sacrifices cannot be ruled out, especially in the absence of any other archaeological indicators of warfare. Furthermore, the methods used to date the site are based on comparative stratigraphy, a relative dating technique that is not as precise as absolute dating methods. Therefore, the Jebel Sahaba violence may not be in as old as has commonly been assumed (Ferguson, 2013b).

Dating from 9500 to 9000 years ago, the walls of Jericho have been taken to be the first unambiguous evidence of warfare anywhere in the world, and this may be true. However, Bar-Yosef (1986: 161) has reviewed climactic and geological evidence of flooding at Jericho and has proposed that "a plausible alternative interpretation for the Neolithic walls of Jericho is that they were built in stages as a defense system against floods and mudflows." Jericho's famous walls may or may not reflect warfare.

Haas (1996: 1360) summarizes the overall findings on the origins of war:

> Archaeologically, there is negligible evidence for any kind of warfare anywhere in the world before about 10 000 years ago …. It was only about 10 000 years ago that the niches of the world were filled in through gradual population growth, and people had to develop new settlement and subsistence strategies to extract adequate resources from decreased territory …. The archaeological record indicates that endemic warfare was much more the exception than the rule until the first appearance of state level societies between 4000 and 2000 BC in the centers of the world "civilization."

In sum, the view that war has occurred over hundreds of thousands if not millions of years into the past has no archaeological basis to support it. The oldest site worldwide, Jebel Sahaba, that might or might not reflect warfare, also might not be as old as has been assumed. But aside from the complete lack of evidence of warfare in the deep past, there are other problems with the contention that warfare is ancient. Archaeology reveals that warfare correlates with the development of social complexity in the shallow past.

Social Complexity and the Origins of War

In bands and tribes, leadership and political power are weak and dispersed, or *uncentralized*. By contrast, in chiefdoms and especially in state civilizations, political power is *centralized* at the top of a social hierarchy (Fry, 2006). In other words, social relations in bands and tribes are relatively egalitarian compared to those within chiefdoms and states that are structured according to ranks or social classes. Relatedly, foraging societies are of two general types: nomadic foragers have the band type of social organization and are egalitarian whereas complex foragers tend to be sedentary small-scale chiefdoms with social hierarchies.

Is there any relationship between social complexity and warfare? A number of studies suggest that the answer is yes. Van der Dennen (1995: 142) expresses that "one of the most consistent and robust findings is the correlation between 'primitivity' and absence of war or low-level warfare, or in other words, the correlation between war and civilization."

The nomadic band was the ancestral form of human social organizations for many thousands of years prior to the beginning of the agricultural revolution about 10 000 years ago. From diverse parts of the globe, archaeological sequences tell parallel stories about the associated development of social complexity and war. Complex foragers are societies that partially or totally gave up the nomadic lifestyle, a process beginning as early as 12 500 years ago. Comparative ethnography shows that warfare is *rare* among egalitarian nomadic foragers where most lethal disputes are interpersonal but *pervasive* among complex hierarchical foragers (Fry, 2006; Fry and Söderberg, 2013; Kelly, 1995).

The shift away from a nomadic forager lifestyle to a settled one entailed many changes. The tell tale archaeological signs of social complexity include higher population density than among nomadic foragers, larger settlements, the presence of ceremonial areas and permanent dwellings, the development of artistic styles, and variation in burials regarding amounts and kinds of grave goods deposited, burial location, and entombment preparation. Variability in burial features indicates that a social hierarchy existed wherein individuals of high social status were interred more elaborately than commoners.

Numerous archaeological examples exist that show the birth of war in association with other social changes. For instance, Maschner (1997) chronicles social changes and the appearance of war for one part of the Northwest Coast of North America. Beginning at about 5500 years ago and persisting over at least a couple millennia, skeletal trauma almost exclusively involved nonlethal injuries – and not many of these. The rarity and nonlethality of the injuries in conjunction with absolutely no other indicators of warfare strongly imply interpersonal aggression rather than war. Later in this prehistoric sequence, evidence of war appears in correspondence with certain social changes. Maschner (1997: 270) explains: "The first large villages appear, status differences become apparent, a heavy

emphasis on marine subsistence develops, and warfare becomes visible in the archaeological record."

Another archaeological sequence recounts a similar tale. In the Near East, between 12 000 and 10 000 years ago, foraging gave way to plant and animal domestication. There is absolutely no evidence of war in the archaeological record 12 000 years BP, sparse evidence for war by about 9500 years ago, and then clear evidence of spreading and intensifying warfare after that. Multiple kinds of data validate the regional origin of war emerging from a state of prior warlessness: The nature of living sites, human skeletal remains, and cultural sequence data show the development of warfare in this region.

Yet another example of the regional evolution of warfare comes from southern Mexico. Flannery and Marcus (2003) find no evidence of group conflict among the nomadic band societies that foraged in the Valley of Oaxaca for six millennia between 10 000 and 4000 years ago. Toward the end of this period, the transition from nomadic foraging to sedentary village life was underway. By 2800, three rival chiefly centers had developed, buffered from each other by unoccupied zones. One chiefdom, San José Mogote, was attacked and sacked. The survivors withdrew to a mountaintop called Monte Albán and constructed defensive walls that encircled the settlement. Eventually, Monte Albán emerged as the capital of the prehistoric Zapotec state. By about 2300 years ago, the Zapotecs had become an empire and conquered the neighboring peoples up to 150 km from Monte Albán. The evidence is crystal clear in showing, first, no war during a long nomadic forager period, which was followed by the developmental of settled, hierarchical societies and the appearance of war, which in turn was followed by the development of intensified state level warfare.

Dye (2009, 2013) considers the archaeological evidence for war and peace in eastern North America. Again we see the now familiar pattern as Dye (2009: 17) writes, "I propose that for nearly half the time that Native Americans lived in eastern North America – approximately 6000 years – peaceful relations characterized these family-level hunter-gatherers, and that violent conflict among social groups emerged only with increasingly complex forms of social and political organization."

These archaeological sequences illustrate how war has originated in particular areas and then developed along with sociopolitical complexity. These foregoing sequences demonstrate a repeating pattern. First, evidence for prehistoric warfare among nomadic forager societies is lacking. Next, as some forager societies underwent social changes toward greater social complexity, sometimes but not always war appeared in an archaeological sequence. In places where civilizations arose, the archaeological record often shows increases in the occurrence and severity of war. The archaeological evidence also suggests that warfare may be augmented by population pressures and environmental change.

In conclusion, the view that war arose recently, within the last 10 000 years, is consistent with the worldwide archaeological evidence. As Kelly (2000: 1, 2) summarizes,

> The archaeological record provides very little clear-cut evidence of warfare within hunting and gathering populations prior to the development of agriculture …. By 5000 to 4300 BC [7000 to 6300 BP], fortifications, garrisons, and site destruction at a number of locations in the Near East provide archaeological evidence for a more general prevalence of warfare …. In the relatively brief span of 4500 years, a global condition of warlessness that had persisted for several million years thus gives way to chronic warfare that arises initially in the Near East and subsequently in other regions where a similar sequence of transformative events is reduplicated.

The Reasons for War

From an anthropological perspective, the reasons for warfare or its absence are diverse. Several generalizations can frame the discussion. First, there is no unicausal explanation for warfare. Second, warfare is least developed in and sometimes absent from uncentralized societies (foraging bands and tribal peoples) in comparison to more complex and hierarchical forms of social organization (chiefdoms, kingdoms, ancient civilizations, and modern nation-states). Third, in hierarchical societies, some wars can be traced to the self-interested actions of the leaders.

Unicausal theories of warfare can be rejected as too simplistic. Some societies engage in defensive war, as was the case for the Tohono O'odham of the Southwestern United States when they defended themselves against raiding Apaches or, similarly, as was the case for 10 Upper Xingu River basin tribes of Brazil – which constitute a nonwarring peace system among themselves – as they united for defensive purposes to keep hostile neighbors at bay. Sometimes the individual status and gain of the warriors is at stake, as in the raiding for horses and scalps in the Great Plains of North America. In order to marry, for instance, a young Blackfoot brave needed horses to give to a potential father-in-law when proposing marriage. Wars are fought also for access to resources of various sorts, whether they be garden land in New Guinea, cattle in East Africa, hunting grounds in northeastern North America (following the introduction of the fur trade by Europeans), access to trade goods in Amazonia, or control of a trade route as in Near Eastern prehistory (Ferguson, 2011; Fry, 2006; Roper, 1975). Furthermore, interpersonal factors such as longstanding rivalries among chiefs or revenge-seeking for past losses come into play as reasons and rationalizations for war, factors which often are also bound up with more material or political aims. In short, the reasons for warfare apparent in ethnographic descriptions are diverse and numerous and vary from case to case.

The reasons for war also relate to social organization. There are numerous factors that dictate against warfare among nomadic forager band societies. These groups lack material goods to plunder, the populations necessary to mount much of a fighting force, the absence of anyone with enough authority to order men into battle, areas too large to defend by small bands and, in fact, little reason to do so as animals move from place to place and wild plants, tubers, or nuts are widely dispersed.

Nomadic foragers may commit homicide for personal reasons such as over a woman, an insult, or to avenge the death of a loved-one, but they tend not to go to war, group versus group, and if they do, the type of war tends to be of low intensity (Fry, 2006; Fry and Söderberg, 2013; Reyna, 1994). Meggitt (1965: 245) captures the essence of the

nomadic band circumstances that dictate against warfare as he explains that the Walbiri of Australia did not in any way emphasize militarism: "There were no military leaders, elected or hereditary, to plan tactics and ensure that others adopted the plans …. There was in any case little reason for all-out warfare between communities. Slavery was unknown; portable goods were few; and the territory seized in a battle was virtually an embarrassment to the victors, whose spiritual ties were with other localities."

Kelly (2000) has pointed out that once a society develops social segments, for example separate lineages, then feuding is facilitated. Much of what has been called tribal warfare might more accurately be called tribal feuding (see Boehm, 1987; van der Dennen, 1995; Fry, 2006; Reyna, 1994). Because there are corporate groups – social segments – and weak overarching figures of authority, an initial killing may lead the victim's kin group to seek retribution against any member of the opposing group, which then can result in ongoing cycles of violent retribution, back-and-forth. In the process, trophies may be taken, or women captured as wives. This type of fighting has been called social war by Wright (1942).

In fact, Wright (1942) developed a four-category warfare classificatory scheme and applied it to 590 societies from around the world. He used the terms *defensive war, social war, political war, and economic war*. Thirty societies (5% of the total) were found to have *no war* – that is, the literature revealed no evidence of warfare, no military organization, and no special weapons – but societies in this group were nonetheless labeled by Wright as practicing defensive war rather than nonwarring. Wright assumed that members of such societies would fight to defend themselves if attacked, but in fact ethnographic accounts show that this is not always the case; members of such societies typically flee from would-be or actual aggressors (Fry, 2006).

Another 346 societies (59% of the sample) were rated by Wright (1942: 546, Appendix IX) "to be unwarlike or to engage only in mild warfare," provided that "no indication was found of fighting for definite economic or political purposes in the more specialized literature." Social war consisted, for instance, of headhunting, revenge-seeking, and feuding. The inclusion of feuding as a type of warfare contributes to the social war category being the largest of the four classifications.

It is worth noting that adding together the defensive/non-warring group and the unwarlike/mild-warring social war group shows that nearly two-thirds of the total sample (64%) were nonwarring or mild-warring. The rest of these 590 societies were determined by Wright to engage in war for economic or political purposes (29 and 7% of the total, respectively). The expansion of the ancient Zapotec empire in order to dominate rivals and exact tribute from the hinterlands provides an example of political warfare (Flannery and Marcus, 2003). Often economic and political reasons for war go hand-in-hand.

On the basis of a career spent studying the anthropology of war, Ferguson (2011) emphasizes the importance of considering the self-interests of the leaders. Those who actually start wars tend to be the ones who benefit most in terms of acquiring wealth, resources, and power. "To understand war, then, one must focus on the decision makers. That directs attention to the sociopolitical structure of a society. What different kinds of people, groups, and institutions contribute in what ways to decisions, and what are their interests in a given situation?" (Ferguson, 2011: 265). Of course the leaders, notes Ferguson (2011: 265), tend not to state to their followers, "go risk your life because it is good for me," but rather emphasize the importance of courage, the necessity of defense, the satisfaction of revenge, or the spoils of war for the warriors and society. Once a society has developed a complex of warrior values and institutions such as bravery, obedience, loyalty, and fortitude, coupled with a standing army and a social chain of command where refusal to fight can be severely punished, the rationale for going to war also can be justified as moral and right.

Conclusion

Anthropology shows that numerous factors can contribute to warfare, for instance, a quest for safety and security, a scarcity of or competition for resources, a desire for revenge, and the political and economic self-interest of the decision-makers themselves. Anthropology also shows that living without war is possible as nonwarring societies and nonwarring peace systems do exist (Fry, 2006). It is clear that the frequency and intensity of warfare correlate with social complexity, as reflected in both the worldwide archaeological record and in comparative ethnography of nomadic bands, tribes, chiefdoms, kingdoms, and states. When humankind made a living as nomadic foragers, the species was not plagued by warfare. Warfare has arisen repeatedly within the last 10 000 years and spread in recent millennia along with certain changes such as population increase, settling down, the development of social hierarchies and authoritative leadership, or, in a word, with the development of more complex forms of social organization, especially the state. The key question for the twenty-first century is whether humankind will go down fighting or, alternatively, give up the practice of warfare and resolve their differences without bloodshed.

See also: Conflict and War, Archaeology of: Weapons and Artifacts; Peace and Nonviolence: Anthropological Aspects; Peace and Nonviolence: Overview of Anthropological Aspects; Violence in Anthropology; War: Anthropological Aspects, Historical Development of; Warfare in History.

Bibliography

Bar-Yosef, O., 1986. The walls of Jericho: an alternative interpretation. Current Anthropology 27, 157–162.

Boehm, C., 1987. Blood Revenge: The Enactment and Management of Conflict in Montenegro and Other Tribal Societies, second paperback ed. University of Pennsylvania Press, Philadelphia.

Bowles, S., 2009. Did warfare among ancestral hunter-gatherers affect the evolution of human social behaviors? Science 324, 1293–1298.

Bowles, S., Gintis, H., 2011. A Cooperative Species. Princeton University Press, Princeton.

Conklin, H.C., 1954. The Relation of Hanunóo Culture to the Plant World. Unpublished doctoral dissertation. Anthropology Department, Yale University.

Dennen, van der, J., 1995. The Origin of War. Origin Press, Groningen, The Netherlands.
Driver, H.E., 1969. Indians of North America, second ed. University of Chicago Press, Chicago.
Dye, D., 2009. War Paths, Peace Paths: An Archaeology of Cooperation and Conflict in Native Eastern North America. Alta Mira, Lanham, MD.
Dye, D., 2013. Trends in cooperation and conflict in eastern native North America. In: Fry, D.P. (Ed.), War, Peace, and Human Nature: The Convergence of Evolutionary and Cultural Views. Oxford University Press, New York.
Ember, C., Ember, M., 1992. Warfare, aggression, and resource problems: cross-cultural codes. Behavior Science Research 26, 169–226.
Ericksen, K.P., Horton, H., 1992. 'Blood feuds': cross-cultural variations in kin group vengeance. Behavior Science Research 26, 57–85.
Ewers, J., 1967. Blackfoot raiding for horses and scalps. In: Bohannan, P. (Ed.), Law and Warfare. University of Texas Press, Austin, pp. 327–344.
Ferguson, R.B., 1997. Violence and war in prehistory. In: Martin, D.L., Frayer, D. (Eds.), Troubled Times: Violence and Warfare in the Past. Gordon & Breach, Amsterdam, pp. 321–355.
Ferguson, R.B., 2011. Born to live: challenging killer myths. In: Sussman, R.W., Cloninger, C.R. (Eds.), Origins of Altruism and Cooperation. Springer, New York, pp. 249–270.
Ferguson, R.B., 2013a. Pinker's list: exaggerating prehistoric war mortality. In: Fry, D.P. (Ed.), War, Peace, and Human Nature: The Convergence of Evolutionary and Cultural Views. Oxford University Press, New York.
Ferguson, R.B., 2013b. The prehistory of war and peace in Europe and the Near East. In: Fry, D.P. (Ed.), War, Peace, and Human Nature: The Convergence of Evolutionary and Cultural Views. Oxford University Press, New York.
Flannery, K., Marcus, J., 2003. The origin of war: new ^{14}C dates from ancient Mexico. Proceedings of the National Academy of Sciences 100, 11801–11805.
Fry, D.P., 2006. The Human Potential for Peace: An Anthropological Challenge to Assumptions about War and Violence. Oxford University Press, New York.
Fry, D.P., Söderberg, P., 2013. Lethal aggression in mobile forager bands and implications for the origins of war. Science 341, 270–273.
Gardner, P., 2004. Respect for all: the Paliyans of South India. In: Kemp, G., Fry, D.P. (Eds.), Keeping the Peace: Conflict Resolution and Peaceful Societies around the World. Routledge, New York, pp. 53–71.
Haas, J., 1996. War. In: Levinson, D., Ember, M. (Eds.), Encyclopedia of Cultural Anthropology, vol. 4. Henry Holt and Co., New York, pp. 1357–1361.
Haas, J., 1999. The origin of war and ethnic violence. In: Carman, J., Harding, A. (Eds.), Ancient Warfare: Archaeological Perspectives. Sutton Publishing, Goucestershire, UK, pp. 11–24.
Haas, J., Piscitelli, M., 2013. The prehistory of war: misled by ethnography. In: Fry, D.P. (Ed.), War, Peace, and Human Nature: The Convergence of Evolutionary and Cultural Views. Oxford University Press, New York.
Hallowell, A.I., 1974. Aggression in Saulteaux society. In: Hallowell, A.I. (Ed.), Culture and Experience. University of Pennsylvania Press, Philadelphia, pp. 277–290.
Johnson, A., 1983. Machiguenga gardens. In: Hames, R.B., Vickers, W.T. (Eds.), Adaptive Responses of Native Amazonians. Academic Press, New York, pp. 29–63.
Keeley, L.H., 1996. War before Civilization: The Myth of the Peaceful Savage. Oxford University Press, New York.
Kelly, R.C., 2000. Warless Societies and the Origin of War. University of Michigan Press, Ann Arbor.
Kelly, R.L., 1995. The Foraging Spectrum: Diversity in Hunter-Gatherer Lifeways. Smithsonian Institution Press, Washington DC.
LeBar, F.M., Hickey, G.C., Musgrave, J.K., 1964. Semang. In: LeBar, F.M., Hickey, G.C., Musgrave, J.K. (Eds.), Ethnic Groups of Mainland Southeast Asia. Human Relations Area Files Press, New Haven, CT, pp. 181–186.
Low, B.S., 1993. An evolutionary perspective on war. In: Zimmerman, W., Jacobson, H. (Eds.), Behavior, Culture, and Conflict in World Politics. University of Michigan Press, Ann Arbor, pp. 13–55.
Meggitt, M., 1965. Desert People: A Study of the Walbiri Aborigines of Central Australia. University of Chicago Press, Chicago.
Maschner, H., 1997. The evolution of Northwest Coast warfare. In: Martin, D.L., Frayer, D.W. (Eds.), In Troubled Times: Violence and Warfare in the Past. Gordon & Breach, Amsterdam, pp. 267–302.
Otterbein, K.F., 1968. Internal war: a cross-cultural study. American Anthropologist 70, 277–289.
Otterbein, K.F., 1970. The Evolution of War: A Cross-Cultural Study. Human Relations Area Files Press, New Haven, CT.
Otterbein, K.F., 2004. How War Began. Texas A&M University Press, College Station.
Otterbein, K.F., Otterbein, C., 1965. An eye for an eye, a tooth for a tooth: a cross-cultural study of feuding. American Anthropologist 67, 1470–1482.
Pinker, S., 2011. The Better Angeles of Our Nature. Viking Books, New York.
Prosterman, R.L., 1972. Surviving to 3000: An Introduction to the Study of Lethal Conflict. Duxbury-Wadsworth, Belmont, CA.
Reyna, S.P., 1994. A mode of domination approach to organized violence. In: Reyna, S.P., Downs, R.E. (Eds.), Studying War: Anthropological Perspectives. Gordon & Breach, The Netherlands, pp. 29–65.
Roper, M., 1969. A survey of the evidence for intrahuman killing in the Pleistocene. Current Anthropology 10, 427–459.
Roper, M., 1975. Evidence of warfare in the Near East from 10,000-4,300 BC. In: Nettleship, M.A., Dalegivens, R., Nettleship, A. (Eds.), War, its Causes and Correlates. Mouton, The Hague, pp. 299–340.
Sponsel, L., 1998. Yanomami: an arena of conflict and aggression in the Amazon. Aggressive Behavior 24, 97–122.
Shternberg, L., 1933. Semya I Rod U Narodov Severo-vostochnoi Azii. English Translation in the Human Relations Area Files, Id Number RX2. Human Relations Area Files, New Haven, CT.
Wrangham, R., Peterson, D., 1996. Demonic Males: Apes and the Origin of Human Violence. Houghton Mifflin, Boston.
Wright, Q., 1942. A Study of War. University of Chicago Press, Chicago.

Conflict: Anthropological Aspects

Georg Elwert[†]
Günther Schlee, Free University of Berlin, Berlin, Germany; and Max Planck Institute for Social Anthropology, Halle (Saale), Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G. Elwert, volume 4, pp. 2542–2547, © 2001, Elsevier Ltd.

Abstract

Conflict is a regularly occurring and socially organized phenomenon. Organized conflict in its nondestructive forms is a condition for the differentiation and expression of interest as well as for endogenous social change and the processing of exogenous stimuli of transformation. The self-stabilization of social systems and social change could not be analyzed without giving central attention to conflicts and their organization. Conflicts can be classified in respect of their stronger or weaker social embedding and in respect of the higher or lower incidence of violence in a field with four extreme types: the strongly embedded forms of warring and procedure and the weakly embedded ones of destruction and avoidance.

Conflict is an action based upon the perception of partially incompatible interests or intentions between two or more persons. This perception need not be shared by both sides. What should be considered as an interest or intention requires interpretation in respect of the context. For analytical purposes, these concepts often remain projections; they should, however, be grounded in observations of statements or of chains of action that are open to operationalization and falsification. They are necessarily culture bound and may even be a matter of debate within a given culture. Furthermore, the perception of conflict does not need real intentions or real persons in order to generate action. The assumption of witchcraft, e.g., may generate virtual feuds.

Goals of an Anthropological Theory of Conflict

An anthropological theory of conflict can

1. Contribute to the explanation of violence. Conflict is, however, not the only element that will contribute to an explanation of violence. Violence may, for example, be needed as a strong symbolic marker for ritual purposes. And it has to be underlined against a popular perception of 'conflict' that not every conflict is violent.
2. It can contribute to a theory of the attraction or displacement of emotions. The hatred against foes, for example, is not something natural but is an outcome of specific constructions of conflict.
3. The theory of conflict competes with culturalist approaches as an explanation of social cohesion. Forms of conflict regulation can reduce the relative conflict intensity between persons and, thus, create their allegiance to the system or organization that provides for this conflict regulation.
4. It contributes essentially to a theory of social evolution. The ways by which a society regulates conflict operate also as means for the selection of alternatives for its future.

[†] Deceased

The History of Anthropological Research on Conflict

Anthropological research on conflict is a history of many hopes lost and few results kept. It was a classic conviction (going back to Thomas Hobbes) that a 'primitive' or 'natural' state of society would be characterized by omnipresent violence. Konrad Lorenz (1966) proposed a biological paradigm insisting on a genetic program for violence, which Ashley Montagu (1968) convincingly opposed. One finds a broad variance of behavior toward potentially conflictuous situations in all different forms of human social organization. If something is encoded, it is both a potential for aggression on the one hand, and a potential for flight and conflict avoidance on the other.

Sahlins' (1961) study of the Nigerian Tiv, of 'lineages of predatory expansion,' suggested to some that genealogical closeness generally made violent conflict less probable. Barth's (1959) study on Swat Pathans has shown, however, that this pattern cannot be generalized, not even to lineage societies. Confrontation may follow from alliances against one's closest competitors, even if they are close relatives.

Types of violent conflict had been associated with different 'cultural levels' even by Malinowski (1941). It was argued that on the first levels there would be less violence. Other authors (e.g., Chagnon, 1968) assumed for these societies a specific type of resource competition as a strong motive for conflict (cf against these Hallpike, 1973). Helbling (1999), however, has shown that the occurrence and intensity of conflict is not directly correlated with different types of social organizations, but rather to population density within a given type of social organization. Thus swidden agriculturalists or hunter-gatherers with a high population density may also show a high incidence of conflicts.

Marcel Mauss (1923/24) built his theory of exchange ('the gift') in order to provide a theory of peace. Relations of reciprocity that serve as a basis for exchange including market exchange are, however, compatible with relations of warring. Polanyi (1968) showed this for the seasonal warfare of the kingdom of Dahomey in West Africa.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12232-9

For a long time, one of the most prominent anthropological theories of conflict was Max Gluckman's (1956) theory of cross-cutting ties, which would hinder the outbreak of conflict because the potential foes were always also indirectly linked by alliances. Hallpike (1977), cf Schlee (2000) has shown, however, that cross-cutting ties may as well work the other way round, as an organization of competition, thus fostering conflict.

A stable achievement of anthropology is Sumner's (1911) ethnocentrism theory. The organization of violence requires the definition of boundaries. Leach (1965) added dryly that killing is the most powerful tool of classification – for the survivors.

Another result, not so readily identified with a single name, is the perception that factors of social organization have a higher impact upon successful strategies of violent conflict than pure technology. Although early historical work insisted strongly upon the effect of the arrow, the warrior mounted on horseback and firing arms, or even more so bomber aircraft, later research, however, stressed rather the organization of access to technology, the transmission of knowledge and training of skills and, last but not least, the social organization of risk perception. Violent conflict requires that risks are played down against potential benefits.

Methodological Problems of Conflict Research

That a human action is categorized as conflictive requires that it fulfills in its form some preestablished patterns. That pattern perception also means that contingent action without conflictive intention is likely to be interpreted as conflictive if it is harmful and if a meaning may be interpreted into it.

Societies tend to present conflict as unforeseeable or even chaotic. Sometimes conflict is even excluded from a society's self-representation. Conflict may be characterized as evil, or as a psychological and not as a social phenomenon. The less conflict is 'normalized' in institutional forms of low intensity, the more difficulties will fieldworkers face when trying to learn something about existing conflicts or the pattern of their organization and resolution. The existence of conflicts may just be denied, or existent conflicts within a group may be hidden behind other conflicts with outsiders.

Conflict as Embedded Social Action

Anthropology has challenged the popular perspective of conflict as chaotic and disorderly. The image of chaos is, however, based upon experience, Successful conflict management in the interests of the actors requires surprise, which may create the impression of unforeseeability. Conflicts follow, however, socially ordered paths. Rather like in the case of avalanches, the exact moment and the intensity are difficult to foresee, but one knows in a given environment where they may happen and which path they will follow. Anthropological research has shown that even violent conflict follows culturally encoded patterns, has institutionalized forms, and is controlled and directed in its appearance. This has been called social embedding (Elwert et al., 1999).

Embedding is the ensemble of moral values, proper norms, and institutional arrangements that set limits to a specific type of action and make simultaneously the outcome of these actions calculable. The fact that conflict has both controlled and foreseeable aspects and an element of surprise creates a hybrid structure. Therefore, one may speak of partial embedding.

Conflict situations are characterized by a relatively high autonomy of time use for one of the actors. The outcome in conflicts is unforeseeable in respect of the relative gains and losses. Society, on the other hand, creates foreseeability through routines and norms that channel the realization of conflict. Norms as a means of channeling conflict not only have a stronger effect than just routines, but are also much less flexible than routines. It is, therefore, inherent to the character of social norms that they will be 'underlived.' One should therefore not expect norms regulating social conflict to be followed 100%.

The concept of norms allows us to differentiate two types of conflict: normative conflict versus actor conflict. In the former, a person or a group of persons clashes with a norm when violating other persons' or groups' rights. In the latter, individuals or groups clash and try to harm each other in a field of action that has normatively defined boundaries but where neither side can claim norms protecting its goals.

Forms of Actor Conflict

The multitude of concepts used in order to describe conflicts (war, genocide, competition, feud, legal battle, etc.) can be ordered in a field with four poles (see **Figure 1**). To speak of poles means that the empirical cases and emic categories may be positioned between these abstract ordering categories. The criteria are more or less violence, and stronger or weaker embedding. The category 'warring' includes both feud and war. A war is a violent conflict that includes physical risks but is limited by some rules. A feud is a violent conflict with stronger embedding and in general also less violence than wars. It is defined as a violent conflict where the potential victims, the place of conflict, the time of conflict, or the weapons allowed may be the matter of a binding agreement. A necessary condition for the definition of feud is that the means to end the conflict is part of the agreement. Feuds and competitions cannot be clearly separated, as all these categories have a fluid transition. Competitions, however, should rather be seen as being part of procedures. Procedurally organized competition, measured in abstract parameters (e.g., money or speed), may allow for an almost continuous conflict without violence.

In legal studies, procedures are defined by a specific – ritual – form, whereas anthropology as a comparative science cannot only rely upon conceptualizations defined by the actors. An

	More violence	Less violence
Stronger embedding	1	Procedure
Weaker embedding	Destruction	Avoidance

Figure 1 Differentiations of conflict.

anthropological concept of procedure, therefore, has to take up suggestions formulated by Niklas Luhmann (1969). A procedure is distinct from daily interaction by its form and ordered sequence of action. Elections, court cases, and auctions have also been seen as conflict resolution procedures. During a procedure, some power differentials are suspended. It may, for example, be excluded that a stronger person physically harms a weaker person during the procedure. A procedure has an outcome with consequences in action. Who will profit from the outcome is, in principle, open. It should be called procedure only when it ends in a conclusion that has meaningful consequences for action.

The conclusive procedure has to be distinguished from the pending procedure: something that has the ritual form of the procedure but that never comes to conclusions that are consequential. Thus, the pending procedure is not a way of conflict resolution but a way of conflict perpetuation. Pending procedures can have a high annoyance potential and are rather forms of warring with reduced violence.

Destruction (e.g., genocide) is distinguished from warring as the goal is the total annihilation of the other. There is no agreement that binds both partners together. Nevertheless, the ways the other side is destroyed follow a culturally encoded pattern. There is no human society that allows killing that does not simultaneously have some rules for 'proper' killing.

Avoidance is a concept rarely included in theories of conflict (Radcliffe-Brown, 1952), but to have a complete conceptual structure we need to consider also this 'zero-case.' In some societies, avoidance is the dominant mode of dealing with conflict (see Elwert, 2001 on the Byalebe of West Africa). In a situation where avoidance is dominant, it is more probable that people switch to destruction than that they switch to procedure.

The Normative Conflict

A normative conflict is a conflict in which a person's behavior toward another person or a thing, which is meaningful to a person, conflicts with a norm. In this context, norm is meant in the strict sense; it requires both monitoring and sanction. Prescriptive or prohibitive formulas without a sanction probability should be considered moral values, not norms. This definition is different from those given within societies. There are, in general, more prescriptive or prohibitive formulas that are considered 'norms' than those whose infraction is really prosecuted.

The monitoring is necessary in order to create a sanction potential. Monitoring is a sensitive element of this process. In some societies, monitoring is restricted to specialists or to persons with a recognized claim to competence (police, priests, and soothsayers).

Reactions to norm infraction range over a wide spectrum. In state societies, the dominant mode of reaction is typically the sanction by a central specialized apparatus, the coercion instance. Another solution is the 'sanction-service.' In state societies, this may appear parallel to the centralized apparatus; it is then often considered to be illegitimate, for example, as a Mafia (Krauthausen, 1997). In stateless societies, there may be specialists for sanction services, organized as corporations, age grades, or secret societies. A third option is self-help, which may take a form similar to feuding (Spittler, 1980). Self-help requires some kind of broader agreement that a norm has been transgressed. This can be a court decision, which then legitimates the persons considered as victims of the norm infraction to take action or to hire specialists for this purpose (cf Würth, 2001 on shari'a courts in Yemen). A fourth option, which again is a zero-option, is ignoring the conflict. If there are no strong central instances or sanction services available and if the dividing line between the actor conflict and normative conflict is blurred, this is a common solution.

Sanctions are not only physical sanctions (as corporeal harm or mutilation, restriction of movement, or material and economic fines) but also loss of status or attribution of negative reputation. The reputation sanction is very common in those contexts where a person can hardly escape contact with those who share the negative evaluation of the culprit. Since societies are never homogeneous, subgroups may not share specific norms (Pospisil, 1958). The reputation sanction is then limited in its efficiency. The redundancy of overlapping reference groups, each with a sanction capacity, can cope with that (Elwert, 1980).

The normative conflict is a two-tier process. The first element is the infraction; the second is the sanction process. That the infraction is an infraction is, however, only definitely clear after the sanction process. The working of the sanction process may be expeditive (e.g., lynching) or procedural. The expeditive reaction is based upon a single decision or a turning point from when on a plurality of single decisions is valid enough for action. All human societies – not only the 'modern' ones – know, however, the advantages of procedural forms of sanctioning over the expeditive ones. Truth may be a matter of disagreement, and the loss or investment of time characteristic of procedures may allow information to be validated and emotions to cool down. Procedures may, however, also be used for the alternative purpose of steering the multiplicity of perceptions into just one established truth, and to stir up the emotions in such a way that sanctions become more probable or more violent.

The Attraction and Displacement of Emotions

To live with emotions and act upon them requires some societal agreement. The popular notion that conflict is motivated by emotions does not hold against anthropological evidence – at least insofar as recurring conflicts are studied. People hardly react to emotions, but to assumed or perceived goals of action. The reaction in a conflict situation has higher strategic qualities if it is not based upon emotion but upon cool calculation. To 'show' emotions may, however, be very important in order to legitimate one's action in a conflict situation, or in order to impress the adversary.

Actions in conflicts, however, can attract emotions. This is correlated with changes in the structure of relevance; it may help to concentrate resources to the conflict. Actor conflicts are strong attractors of emotion, even more so if violence is implied. In any procedural treatment of conflict, and especially in normative conflicts, emotions can be less intense. They may

appear comparatively detached or displaced. A society that gives preference to normative conflicts and that privileges the procedural treatment allows for the displacement of emotions. The literary representation of conflict or sport competitions – a form of conflict with lesser implications to one's personal life goals – may then attract stronger emotions than those linked to lived conflict.

Conflict Motives

Motives to enter conflicts can be grouped into three broad categories: honor, power, and material gain. Each of these three general motives has its own definition of who won in a conflict situation. What, for example, an outside observer may see as an economic loss, may appear to the insiders as winning prestige. In most cases, however, there are several motives overlapping, and this can make it difficult to define who won in a given situation. To distinguish between these three categories of motives is, however, helpful in order to perceive different consequences of their priority for the respective organization of a conflict. If honor is the major goal – and even more, if honor may be even gained by a dead person declared to be a hero – then the propensity to risk increases. This may create patterns of recurrent warring, since new generations need new prestige (cf Evans-Pritchard, 1957 on Azande warfare). In the case that power, the control over persons, is the goal, the end of the conflict has to be such that the winning side retains a coercion potential over the losing side. If material gain is the goal, then it may be reinvested, in a self-perpetuating process, to gain even more.

Systems of blood feud or vengeance are often dominated by a system of attribution of honor and shame. To kill a person may create honor. If this killing is not retaliated, this creates shame. Vengeance 'washes away the shame' and reestablishes honor (Schwandner, 1999).

Behind a smokescreen of ideological motives, warlords exhibit a very clear priority for motives of economic gain, whether the means are peaceful or violent. This entrepreneurial activity creates a market without the typical embedding for peaceful markets, a market of violence (Elwert, 1999). Whether warlords subjectively give priority to noneconomic goals or not, they need weapons and soldiers (even if they call them volunteers) in order to reproduce their activity. The gains from robbery, blackmail, hostage taking, customs, and trade (especially in valuable items such as precious metals or stones, drugs, and weapons) have to cover expenses for the violence and generate a profit.

Group conflicts may originate in divergences of interest between the groups or they may be produced by the construction of such contradictions. In the latter case, the purpose is a strategy that aims primarily upon the control of power within that group, an endostrategic mobilization. In such a process, the outside conflict is heated up in order to homogenize the group, to push forward new norms or structures of leadership. Boundaries to 'the others' are created or become accentuated. Endostrategic mobilization is typically the instrument of would-be elites who want to gain power and therefore portray the actual elite as persons who do not sufficiently defend collective interests, or as agents of outside forces.

Conflict Regulation and Cohesion

It seems at first sight paradoxical to associate the social organization of conflict positively with social cohesion. Since anthropologists no longer believe that there is a natural harmony among human beings, however, one cannot escape the notion that the normalization and institutionalization of conflict in specific forms is an essential condition of social cohesion. Divergence of interest is omnipresent among human beings. If there is no channeled way to express it, then random conflict and – even more so – avoidance reduce chances for cooperation. Cooperation requires not that conflict is repressed but that it finds predictive ways of expression. Less destructive forms of conflict and especially forms that maintain both sides' interests in form of compromise ease the readiness to enter conflicts. They do so because the lower risk threatens less the conditions of cooperation during or after the conflict. Thus, an institutionalization of conflict may increase conflict incidence if conflict becomes less violent within the respective entity.

A social group with a common reference is not necessarily a 'conflict group,' a group capable of acting collectively against one other. A conflict group requires an internal sanction capacity. Without it, collective action would be undermined by free riders who want to enjoy the benefits of a victory but avoid the risk of the conflict itself. The creation of such a conflict capacity through an internal sanction system is an important indicator of a mobilization potential for civil strife and civil war.

We can observe that we-groups, organizations, or societies tend to reduce the violence of conflict within their respective boundaries. But this tendency does not create a general pattern. As cooperation creates resources, material or immaterial (such as honor), these resources may motivate further conflict. Thus, the violence of conflict within groups may be stronger than the violence between them. As long as these conflicts do not turn into a systematic destruction of a given subgroup, the entity's cohesion may be maintained nevertheless.

The economist Albert Hirschman (1994) has observed something about modern societies, which probably can be generalized to other societies as well: Group cohesion is strongly correlated with the access to accepted modes of conflict regulation. Shared cultural values, even if they are as emotionally binding as language or religion, do less to explain social cohesion than the presence of institutions that appear to give a fair chance to realize one's interests with reduced destructive risk. From an anthropological point of view, one has to add that procedures create a higher cohesion than other channeled ways of living a conflict.

Hirschman has further shown that those procedures which favor more-or-less solutions (incremental conflict resolution in the authors' terminology) have a better chance to create cohesion than those with either–or structure (alternative conflict resolution). The compromise maintains elements of interest of both sides. This eases the identification with the outcome of the conflict regulation. This perspective has been applied in an anthropological perspective to modern nations. It could be shown in one case that irrespective of cultural parameters such as language or religion, it is the deprivation of access to normatively regulated institutions of conflict resolution that

furthers processes of fission and separatism (Elwert, 1997). For African states, it has been shown that the classical ignition spark for civil war is the arbitrary use of violence by state agents (Wirz, 1982).

Institutions of Conflict Regulation as Central Determinants for Processes of Social Transformation

Ralf Dahrendorf (1954) has argued that the forms of conflict regulation are central for the difference between paths of social transformation. The normative structure most often (but neither always nor exclusively) ensures a society's reproduction and its sameness over time ('identity'). Social change requires breaks in the reproduction of sameness. It can, however, not proceed by 'anything goes,' since the end of reproduction would undermine further change. Stimuli for change, whether they are endogenous or exogenous, have to be treated as alternatives to something, and not as unrelated events. Alternatives are never abstract, they imply persons' intentions. This gives the organization of social conflict the key role in social change to conceive and select alternatives.

Societies that give preference to the normative conflict increase the foreseeability of one's own actions. As long as one follows the norms, the risk of conflict is reduced. But this produces overall a higher conflict-processing capacity. There is less relatively randomized conflict and thus more possibility to engage into ordered conflict. A higher conflict-processing capacity means more variation. If the society uses these institutions in order to select not only winners and losers, but also elements of social order, forms of consumption, and new products of human activity, then innovation can gain acceptance. It is the higher selectivity that reduces the fear of new elements of social order.

It cannot be stated, however, that such a society would be 'more highly developed' than others. Depending upon the type of social environment, different types of conflict regulation appear as optimal adaptations (Hallpike, 1973). Avoidance-biased societies are among the most expansive types of social organizations in an environment characterized by raiding, especially slave raiding. Avoidance produces a high dispersion and thus hinders the population concentrations necessary to make slave raiding a profitable enterprise, e.g., in some regions of the Sahel and the Sudan. At the margins of aggressive imperialisms, feud-biased societies appear to be one of the most stable arrangements. These societies keep internally a high level of warring competence that acts as deterrent or repellent to these imperialist forces – as can be shown in the case of the Caucasus or some regions of the Balkans. If societies interpenetrate each other by means of a reciprocal competition on their markets, a high innovation capacity gives an edge to this competition. Procedure-biased societies appear to be especially strong in this respect, e.g., some medieval oriental societies and some modern Western ones.

Retaliation, Compensation, and Punishment

In Sections The Normative Conflict; The Attraction and Displacement of Emotions; Conflict Motives; Institutions of Conflict Regulation as Central Determinants for Processes of Social Transformation, although they were not typological but addressed general issues of conflict theory, a number of types of conflict and of conflict management have been described (normative conflicts, especially in state societies; actor conflicts; blood feud and vengeance, warlordism; feud-based societies, procedure-biased societies). What remains to be done is to put these types in a comparative framework. Beyond the comparison of pure types, one has to consider how they coexist and interpenetrate in the real world.

Both violent action, like declaring and waging a war, and legal procedures or peaceful consultations can follow the logics of retaliation or of punishment. Retaliation tends to be associated with compensation (because paying compensation for some damage one has done is the means to avert retaliation) and both are based on the idea of equality of the partners on the discursive level, whatever the real power differential may be (parties tend to have different negotiating powers and ways to make noncompliance unattractive for their counterparts). The idea of punishment differs from that. Punishment takes place in the name of a higher entity, be it God, the State, the Law. Those who inflict it might even pretend to regret it and depict themselves as merely having to obey some normative order, said to stem from 'above.' (These spatial metaphors, horizontal versus vertical relationships, higher and lower classes, and tops and bottoms of hierarchies, are inevitable not only in popular parlance but apparently also in social theory.) When talking about the logic of horizontal versus vertical forms of conflict management, the authors refer to ideal types that reality always mixes.

One example of such a mixture or interpenetration can be drawn from conflict management in a region characterized by weak state control (like an inaccessible region in the African savannah) or in a subculture or social environment from which the state has withdrawn its day-to-day interference (like the scene of gangs and gang warfare or illegal but uncontrollable sectors of the economy). Here, one finds regulation by non-state actors to predominate. Also, here deals are struck, conflicts are resolved or at least managed, revenge may be taken or averted by some act of mollification, and so on. In all this, such settings resemble stateless societies in precolonial Africa, or other such social configurations that have come to be known as acephalous ('headless') in the anthropological literature. But in these settings, the state is not completely absent. It does exist in the background of all conflict management processes taking place outside government institutions. One of the parties can always threaten to call the police or to take legal action if the deal achieved in the sphere of custom, informal arrangement, mediation, or arbitration by non-state actors is not satisfactory (and if other conditions, like the courage to oppose these actors, are given). So all these nonstate forms of regulation take place in the shadow of the state and those who strike a bargain there might have to take the state, although it remains passive most of the time, into consideration.

Another example of the interpenetration of these different modes of conflict management can be found at the

global level. The world is an acephalous society. There is no world government and the states that make up the global diplomatic system are formally equal. In such a setting, one would expect retaliation and compensation as the primary modes of conflict management, and indeed much of what one observes on the international scene can be reduced to that. Still, appeals to some more normative level, and claims to obey higher imperatives rather than simply to follow material interests are constantly made, no matter how thinly the latter may be disguised. Even very strong politicomilitary powers taking action against relatively weak ones would seek 'coalitions' with other powers or legitimation by some international organization without a military arm of its own to be able to depict their cause as just and their action as carrying out a punishment for a crime against a broader set of values than a specific national interest. So also here the language of 'higher' authority penetrates the world of formal equals (Schlee and Turner, 2008).

Changing Identifications and Alliances

For the sake of simplicity, so far the authors have written about actors and groups of actors as if they were well defined and about their perceived interests as if they were constant over a significant period of time. Important research questions arise from the fact that they are not. Especially, situations of intensified conflict are characterized by acceleration processes of redefinition of social identities (identification) and changing alliances (alliances are always made up of groups defined as different). So one way to acquire or shed supporters is to enter into wider or narrower alliances with 'others,' often on just one issue of shared interest or a limited set of issues. The other one is to let one's own we-group shrink or widen by adopting exclusive or inclusive forms of recruitment or by engaging in discrimination against certain sets or subsets of members.

Changes in the composition of a conflict party (be it a group or/and alliance) are likely to affect the perception of group interests, because different people have different values, preferences, and definitions of a resource, skills, strategies, habits, etc. Costs and benefits, including gains and losses in a violent conflict, are measured against such changing perceptions of groups and alliances with changing compositions.

A core issue of conflict theory therefore is the dynamics of the composition of conflict parties. Identification often takes place along certain lines that have become conventional in the modern world, like nation, ethnicity, language, or religion, but 'switching' between these is possible (Elwert, 2002). Ethnicity can be nonterritorial or it can follow the nation-state model and become territorialized (Schlee and Shongolo, 2012). To find out about the rationality of identification in conflict situations requires an exploration of the options, i.e., the forms of identification and alliance available to strategic actors and the degrees of freedom they have in choosing among them (Schlee, 2008). That identifications, goals, and strategies change in the course of a conflict, contributes to the general emphasis of a newer kind of conflict theory, which explains conflict configurations at any moment as a result of conflict dynamics rather than preexisting structural givens (Zitelmann, 2004: 51).

See also: Conflict and Conflict Resolution, Social Psychology of; Conflict and Consensus; Conflict and War, Archaeology of: Behavior and Social Organization; Malinowski, Bronislaw (1884–1942); Mauss, Marcel (1872–1950); Simmel, Georg (1858–1918); Violence in Anthropology; War, Sociology of.

Bibliography

Barth, F., 1959. Segmentary opposition and the theory of games. A study of Pathan organization. Journal of the Royal Anthropological Institute 89 (1), 5–21.
Chagnon, N., 1968. Yanomamö: The Fierce People. Holt, Rinehart & Winston, New York.
Dahrendorf, R., 1954. Out of Utopia. American Journal of Sociology 64, 115–127.
Elwert, G., 1980. Die Elemente der traditionellen Solidarität. Kölner Zeitschrift für Soziologie 4, 681–704.
Elwert, G., 1997. Deutsche nation. In: Schäfers, B., Zapf, W. (Eds.), Zur Gesellschaft Deutschlands. Leske & Budrich, Opladen, pp. 123–134.
Elwert, G., 1999. Markets of violence. In: Elwert, G., Feuchtwang, S., Neubert, D. (Eds.), Dynamics of Violence. Processes of Escalation and De-escalation in Violent Group Conflicts. Duncker & Humblot, Berlin, pp. 85–102.
Elwert, G., 2001. Herausforderung durch das Fremde. In: Fikentscher, W. (Ed.), Begegnungen und Konflikt. Beck, Munich, pp. 132–144.
Elwert, G., 2002. Switching identity discourses: primordial emotions and the social construction of we-groups. In: Schlee, G. (Ed.), Imagined Differences: Hatred and the Construction of Identity. Lit-verlag, Münster, pp. 33–54.
Elwert, G., Feuchtwang, S., Neubert, D., 1999. The dynamics of collective violence. In: Elwert, G., Feuchtwang, S., Neubert, D. (Eds.), Dynamics of Violence. PRocesses of Escalation and De-escalation in Violent Group Conflicts. Duncker & Humblot, Berlin, pp. 7–31.
Evans-Pritchard, E., 1957. Azande warfare. Anthropos 52, 239–262.
Gluckman, M., 1956. Custom and Conflict in Africa. Blackwell, Oxford, UK.
Hallpike, C., 1973. Functionalist interpretation of primitive warfare. Man 8, 451–470.
Hallpike, C., 1977. Bloodshed and Vengeance in the Papuan Mountains. Oxford University Press, Oxford.
Helbling, G., 1999. The dynamics of war and alliance among the Yanomami. In: Elwert, G., Feuchtwang, S., Neubert, D. (Eds.), Dynamics of Violence. Processes of Escalation and De-escalation in Violent Group Conflicts. Duncker & Humblot, Berlin, pp. 103–116.
Hirschman, A., 1994. Social conflicts as pillars of democratic market society. Political Theory 22 (2), 203–218.
Krauthausen, C., 1997. Moderne Gewalten. Campus, Frankfurt Main.
Leach, E., 1965. The nature of war. Disarmament and Arms Control 3, 165–183.
Lorenz, K., 1966. On Aggression. Harcourt, New York.
Luhmann, N., 1969. Legitimation durch Verfahren. Luchterhand, Neuwied am Rhein, Germany.
Malinowski, B., 1941. War – past, present and future. In: Clarksen, J., Cochrane, T. (Eds.), War as a Social Institution. Columbia University Press, New York, pp. 21–31.
Mauss, M., 1923/24. Essai sur le don. Année Sociologique 1, 30–186.
Montagu, A. (Ed.), 1968. Man and Aggression. Oxford University Press, Oxford.
Polanyi, K., 1968. Dahomey and the Slave Trade. University of Washington Press, Seattle.
Pospisil, L., 1958. Kapauku Papuans and Their Law. Yale University Press, New Haven, CT.
Radcliffe-Brown, A., 1952. Structure and Function in Primitive Society. Cohen & West, London.
Sahlins, M., 1961. The segmentary lineage: an organization of predatory expansion. American Anthropologist 63, 322–345.
Schlee, G., 2000. Identitätskonstruktionen und parteinahme: überlegungen zur konflikttheorie. Sociologus 1, 64–89.
Schlee, G., 2008. How Enemies Are Made: Towards a Theory of Ethnic and Religious Conflicts. Berghahn, New York, Oxford.

Schlee, G., Shongolo, A.A., 2012. Pastoralism and Politics in Northern Kenya and Southern Ethiopia. James Currey, Woodbridge.

Schlee, G., Turner, B., 2008. Vergeltung: Eine Interdisziplinäre Betrachtung der Rechtfertigung und Regulation von Gewalt. Campus Verlag, Frankfurt Main.

Schwandner, S., 1999. Humiliation and reconciliation in Northern Albania. In: Elwert, G., Feuchtwang, S., Neubert, D. (Eds.), Dynamics of Violence. Processes of Escalation and De-escalation in Violent Group Conflicts. Duncker & Humblot, Berlin, pp. 133–153.

Spittler, G., 1980. Konfliktaustragung in akephalen gesellschaften. Jahrbuch für Rechtssoziologie und Rechtstheorie 6, 142–164.

Sumner, W., 1911. War and Other Essays. Yale University Press, New Haven, CT.

Wirz, A., 1982. Krieg in Afrika – Die Nachkolonialen Konflikte in Nigeria, Sudan, Tschad und Kongo. Steiner, Wiesbaden, Germany.

Würth, A., 2001. As-Sari'a fi Bab al-Yaman. Recht, Richter und Rechtspraxis. Duncker & Humblot, Berlin.

Zitelmann, T., 2004. Wege zur konflikttheorie. Eine subjektive erinnerung. In: Eckert, J. (Ed.), Anthropologie der Konflikte: Georg Elwerts Konflikttheoretische Thesen in der Diskussion. Transcript, Bielefeld, pp. 39–57.

Conflict Mediation

Isak Svensson, Uppsala University, Uppsala, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mediation is a form of conflict management that is based on conflicting parties' ultimate acceptance. Conflict mediation is a growing field, both academically as well as in terms of policy and practice. This article discusses the concept of conflict mediation and suggests that mediation research has developed in three major waves of research. We also examine what mediators do when they mediate, the practice of mediation. This article ends by an exploration of two major research areas in mediation research: the question of timing and mediation strategies.

Introduction

Conflict mediation is a growing field, both empirically in terms of number of active third-party mediators and mediation-supporting institutions around the world, and academically, in terms of an expanding research agenda. (There are some recent reviews of the mediation literature, which complements what is written here (Greig and Diehl, 2012; Wallensteen and Svensson, 2014).) This article will summon the main insights from mediation research. It consists of five sections: (1) a discussion on the definition of mediation, (2) an exploration of development of mediation research, (3) a discussion about the practice of mediation, and (4) an examination of mediation timing and mediation strategies.

What Is Mediation?

Mediation in the context of armed conflicts – conflict mediation – represents an important set of diplomatic measures taken in order to resolve or manage conflicts. Mediation has been defined in various ways. Bercovitch defines it as a "process of conflict management, related to but distinct from the parties' own efforts, whereby the disputing parties or their representatives seek the assistance, or accept an offer of help from an individual, group, state or organization to change, affect or influence their perceptions or behavior, without resorting to physical force, or invoking the authority of the law" (Bercovitch, 1992: 8). A later definition defines international mediation as "a reactive process of conflict management whereby parties seek the assistance from, or accept an offer of help from, an individual, group, or organization to change their behavior, settle their conflict, resolve their problem without resorting to physical force or invoking the authority of the law" (Bercovitch and Houston, 1996: 13). Notably, there is a slight difference in regard to what the mediators actually do: whereas the previous definition emphasizes the changing of perceptions of the parties, the latter one instead focuses on resolution and settlement of the issues at stake. International mediation can also be conceptualized as "the intervention of a third party; it is a voluntary process in which the parties retain control over the outcome (pure mediation), although it may include positive and negative inducements (mediation with muscle)" (Miall et al., 1999: 21–22). Furthermore, mediation can be seen as "a mode of negotiation through which a third party helps the parties find a solution that they cannot find by themselves" (Zartman and Touval, 2007: 437). In other words, mediation is "a form of conflict management in which a third party assists two or more contending parties to find a solution without resorting to force" (Kleiboer, 1996: 360). The common elements of these definitions are three. First, mediation represents efforts to manage or resolve conflicts constructively. It is a group of diplomatic techniques that can be utilized by diplomats, but also by organizations or citizens. Although its essence is peaceful (most definitions exclude the use of force), not all definitions exclude the use of pressure, including *threats* of force. There is nothing in the definitions of mediation which preclude heavy-handed pressure from the third party upon the parties in conflict. In fact, the ability to influence the conflicting parties, the degree of leverage, has been seen as one of the most important explanations for why some mediators are more successful than others (Zartman, 1995). Constructive mediation is here understood as a process in the direction toward more peaceful interactions between the parties in the conflict. The aim with mediation (although not necessarily its only motive) is to seek a solution to a conflict or a de-escalation of the conflict behavior, and thereby it represents a constructive management or resolution of the conflict. Second, mediation is voluntary in the sense that the conflicting parties contain a final veto-power over the outcome of the mediation process. Put differently, mediators might impose themselves and pressure the parties to accept a particular peace deal, but at the end of the day it is up to the parties to accept the mediated outcome. Third, it is an activity performed by third parties. A mediator cannot therefore be a representative of one of the sides (primary actor). It can be an external actor that is emerged in the conflict on one side (secondary actor). A third party is however not necessarily external to the conflict. There are internal third parties that come from inside the conflicting society but still can act as third parties. In fact, research shows that such domestic resources for peace have some particular advantages to bring to the peacemaking processes (Wehr and Lederach, 1991; Svensson and Lindgren, 2013). Hence, conflict mediation is a way of managing or resolving armed conflicts that involves third party actively seeking to bridge the gaps between the main antagonists. It differs from other attempts to regulate conflicts in the sense that it does not require the parties to beforehand accept the final outcome of the mediation process. It therefore depends on the acceptance

of the parties, and any proposal of a mediator needs to be accepted by the parties in conflict. It does include the use of direct force, although various types of political and power pressure are not unusual in some forms of international mediation.

Mediation Research

Mediation research has expanded through three major waves. The traditional approach to mediation was drawn from social psychology. Typical of this style was the work of Herbert Kelman (1997), focusing on particular relations (notably Israel–Arab relations). The main research problem has often been how mediation processes can help overcome barriers of distrust and misperception as well as to communicate shared values of peace. The social-psychological approach highlights limitations in the cognitive, emotional, or organizational capacities and has produced a wealth of insights, also applied in real life situations, notably in the form of problem-solving workshops (Rubin et al., 1994; Fisher, 1999).

A second mediation research tradition has been anchored in rational choice theory, focusing on cost-benefit calculations of the parties in a conflict. A central research question has been whether mediators can exert leverage over the parties in a way that changes such calculations. If mediators can contribute to increase the perception of costs associated with continuous armed conflict (the mutually hurting stalemate), and/or increase the attractiveness of a potential solution (mutually enticing opportunities), then they stand in the best position to influence the basic dynamic of conflicts. Mediators who carry hard-hitting sticks or carrots that can wet the appetite of the parties, are most likely to be successful as peace-brokers (Zartman and Touval, 1985; Stedman, 1991; Touval, 1992). It differs from the social-psychological tradition in a basic sense: the underlying assumption of the theoretical frameworks is not one of limitations of rational decision making of conflict actors, but rather the rational decision derived from the costs and benefits of war and peace.

A third mediation research tradition that has dominated the discourse for the past decade has been the bargaining perspective (Beardsley, 2008). In its systematization it has built on the earlier but moved into a firmer vocabulary and stringent theoretical requirements, paralleling developments in other field (Fearon, 1995; Kydd, 2010). At the same time, it has relied more heavily on data that oftentimes have been collected for particular, even different, reasons. Within this field of thinking, information failures and commitment problems have been the main concern. Information failures are postulated to occur when parties in conflict simultaneously have incomplete information about each other, and reasons to mislead the other side (for instance, by bluffing about real strength to get a better bargain, Fearon, 1995). The mediators' contribution could be to overcome this problem. If they possess information that the parties do not have, and if they can credibly transfer such information between the parties, the information failure problem is managed, and the parties will not need the battlefield to correct their information about each other (Kydd, 2003). This rests on two assumptions, (1) that the mediators can access information that is private (basically information about the antagonists' strength and willingness to fight), and (2) that mediators credibly can convey such information. Empirically, it should be noted, it remains to examine the validity of these two assumptions.

The Practice of Mediation

What do mediators do when they mediate? Put differently, what is the practice and techniques of mediation? A set of mediation activities can be identified. We can understand them if we relate the notion of 'mediation' to the notion of 'conflict.' Conflict can be defined as "a social situation in which a minimum of two actors (parties) strive to acquire at the same moment in time an available set of scarce resources" (Wallensteen, 2011: 16). Hence, a conflict consists of contested issues – aspirations over limited resources, a behavior ('strive' in the definition above, in the context of armed conflicts this implies the use of violence), and it represents a social situation where parties or actors are active. Issues, behavior, and actors are the three basic elements in the conceptualization of what we mean with conflict. Hence, conflict mediation can be seen as third-party-based processes of handling all or some of these dimensions: resolving conflicting issues, de-escalating or terminating violent conflict behavior, and transforming the actors. It is possible to discern seven categories of activities that mediators pursue when mediating, through which they are pursuing activities aiming at influencing issues, behavior, or parties in conflict. Not all mediators engage in all activities, and some mediators may specialize in a particular conflict in only doing one of these activities. As the field of mediation has started to move toward professionalization, it has also led to a higher degree of specialization. Hence some type of mediators may, for instance, work primarily with getting the parties to de-escalate their violent interaction, whereas others may work primarily on reaching an agreement on the contested issues.

The seven categories of activities for mediators:

1. crafting a mediation mandate
2. creating relationships
3. resolving the contested issues
4. handling process issues
5. dealing with regional actors
6. increasing the cost of conflicts and benefits of a solution
7. implementing the agreement

First, mediators craft a mandate for the mediation intervention. The mandate is important in mediation, as it frames the possibilities for the mediator to act in the conflict (Svensson and Wallensteen, 2010). Surprisingly, relatively little research has been conducted on mandates, in particular in relation to the central role that mandates have been playing in research and policy-making of international peacekeeping. A mandate is the basis for the intervention, on which the mediator has its rationale. Mediators can derive the mandate from the parties themselves; from an outside actor, such as a state or an organization; or they can stipulate the rationale for themselves, without connecting it to the parties. The mediator then often enters without an explicit mandate (a self-imposed mandate). If the mediator has an organizational mandate, it gives him or her the freedom to intervene regardless of any activities on the side of the organization. Mandate is often something that is created before or at the

onset of a mediation intervention. It would be too simplistic to see this in terms of a one-term event occurring only at the initiation of the mediation intervention. In fact, mediators have to interpret their mandate in light of evolving developments, and these interpretations may shift over time. Moreover, mediators can renegotiate their mandate in terms of getting more or less room for engagement and bear more or less expectations of what they will do. Furthermore, mediators can renew their mandate, which will give them opportunities to pressure the parties to make an agreement if a current mediator is ending his or her assignment. In an environment where there are several mediators, they need to show their usefulness. For mediators, it is not uncommon that they have to explain the mandate for the parties – the reasons for their involvement and the legitimate basis for their intervention. Since conflict parties are not unitary actors, but consist of a set of individuals, subgroups, and fractions, and also exist in a society where there are many different actors, this means that mediators commonly have to explain the basis for their involvement to new segments of the parties or the society.

Second, mediators strive to create relationships between the parties as well as a sense of understanding and empathy for each other. This also involves creating a trust in the mediator and the mediation process – thereby getting the parties to open up for interactions without their own primary group. Building trust is important both to overcome misperceptions between the belligerents and to move the mediation process forward. Parties in conflict usually have created negative stereotypes of the other side and a great sense of mistrust is common between the parties. This could hinder the mediation process from even starting, as the belligerents might not want to talk to each other or meet face-to-face. An ongoing mediation process might also reach a stalemate due to a lack of trust. At this point the mediator can serve as a vehicle to build or regain trust between the conflicting parties. One way to decrease tensions is by finding common ground beyond the conflict (conflicting parties are also parents and may share love for particular sports, etc.) that the mediator uses to bridge the gap between the parties. It is therefore crucial that the mediator strives to detect when the opportunities occur to find common ground.

Third, mediators try to resolve the contested issues between the parties. Hence, there is a set of activities that are issue-related. One of the most important assignments for the mediator is to serve as a 'reality-check' for the sides – a soundboard on what can be realistically achieved. This can be particularly important for the mediator in conflicts where one side is a rebel-group that does not necessarily interact so much with the outside world. Another activity for the mediator is to transfer and transform messages, sometimes technically just handing over messages, but most often interpreting them and trying to frame them in a perspective that can open up the possibilities for more constructive processes. To a large extent, the mediator's assignment is also to chair meetings, thereby systematically walk through the process of what the parties can agree on, as well as what they cannot agree on. The mediator can also help to identify common interests: sometimes there is underlying common ground that the parties are not able to detect themselves (either because they are blinded by negative stereotypes and enemy-images, or because they simply have rational reasons to hide information from each other). In a mediation round the mediator try to detect in which matters the parties' positions do converge and if there are any common subinterest or overriding interests. Whenever similar positions or interests are detected the mediator can use them as a common basis to start building a solution on. In the same line of thought, a mediator might suggest that the contested issues should be handled in a bottom-up process, where the procedural and 'easier' matters (return of refugees, minority rights, etc.) are dealt with and preferably solved before getting to the more difficult matters (independence or autonomy in territorial conflicts, etc.).

Fourth, the mediator handles a set of process issues, which are not related to the issue at stake, but rather relates to how the process should be conducted. It is not unusual that mediators get frustrated by the fact that they get stuck on what can be seen as minor questions, such as in which order the issues of an agenda should be discussed, which points that should be on the agenda, the order through which parties should enter the room, and so forth. For an outside observer, it may seem ridiculous that processes drag on due to these procedural issues, in particular, in situations where delays imply more people killed at the battlefield in the meantime. It might however be a necessary step, dealing with these procedural issues before getting to the main questions. We know that the question of recognizing rebel groups is a difficult problem in internal armed conflicts. Rebels seek recognition through international mediation while governments are reluctant to engage in mediation due to the same reason. Recognition issues are not limited to civil wars, even if they can be particularly difficult in those situations. Even in interstate conflicts, as for instance, the mediation process in war between Iran and Iraq, procedural issues that were intimately related to recognition were extremely sensitive for the parties.

Fifth, the mediator has to take the regional actors into consideration when conducting mediation. In fact, it is not uncommon that the mediator itself is a regional actor, as research shows that in at least half of the mediation cases the mediator origins from the conflict's geographical region (Regan et al., 2009). Regional organizations have also started to engage more frequently in conflict management since the end of the Cold War (Frazier and Dixon, 2006). The reason why regional actors engage in mediation varies. One aspect that might explain this trend is that states risk of getting affected by regional conflicts, and hence will engage in mediation to find a quick solution. First, there is a risk that the conflict will spread to surrounding countries; second, it is usually the closest neighbors that get the largest flows of refugees. Apart from the fear that the conflict will spread, regional actor generally also has political or economic motives to influence the mediation outcome. In order to find a lasting solution, the mediator might therefore have to consult (powerful) regional actors, as it is possible that the regional actor supports one of the conflicting parties and will influence the conflict until it is satisfied with the outcome. Regional actors usually have a greater knowledge about the area and might gain greater legitimacy than mediators without any previous links to the region. Regional actors may thus serve as both an obstacle and an asset to the mediation process depending on how they choose to engage in the process.

Sixth, increasing the costs and benefits of a solution, by affecting the actual situation on the ground (threat of use of force, by themselves, or by others); and increasing the perceived cost of the conflict and the perceived benefits of a solution. Here there is plenty of room for persuasion. The crucial aspect is that the belligerents perceive the threat given by the mediator as a possible consequence or the benefit as an opportunity. As already mentioned, the ability that the mediator has to influence the parties – the degree of leverage – is an important aspect when explaining mediation success. A mediator with low leverage might therefore need assistance from a more powerful actor, which can use its leverage to increase the costs and benefits, by e.g., promising financial support or by threating to use sanctions or military force in order to influence the outcome on the mediation table. A final threat from the mediator can be to withdraw from the mediation process (Höglund and Svensson, 2011).

Finally, mediation does not end with the reaching of an agreement. In fact, different interpretations of how an agreement should be interpreted can cause severe crisis in a post-agreement period. Therefore, conflict mediation will also in many circumstances occur in relation to the implementation of an agreement reached between the parties. To identify the mechanisms through which agreement should be put into reality is an important aspect of the mediators' work. In order to avoid interpretation errors in the implementation phase, it can be beneficial that the reached agreement should be well structured and clearly state the responsibilities for all the parties. This sounds easier than it actually is, as a mediator might have to choose between reaching a somewhat vague settlement, written imprecisely in order to satisfy all the involved parties, or not reaching an agreement at all. It is not uncommon that mediation takes place also after the agreement has been reached, now in the form of implementation-mediation. A written agreement usually cannot capture all the characteristics or solve all the problems that a postconflict society faces – therefore a third party, in the form of peace-keepers, might be necessary in order to maintain the peace.

Mediation Strategies and the Question of Ripeness

This section will discuss two of the main issues in conflict mediation: the question of timing and the question of strategies, and the section ends with a discussion on how these two mediation issues are related to each other. First, let us address the question of timing. An important contextual condition for any serious mediation attempt to be initiated is that the conflict is ripe for resolution (Zartman, 1995, 2003). Hence, the timing of mediation has been identified as important prerequisite for setting the stage for serious attempts to find a resolution to the conflict. The concept of ripeness is not about explaining success in mediation – for a solution to occur there are other factors at play than merely that the situation is ripe. It does however describe situations where there are windows of opportunities: where both sides, at the same time (but not necessarily for the same reasons) are willing to reconsider the war-option as the primary strategy. Due to the inability to make progress on the battlefield, conflict parties are at these moments more open to pursue a negotiated track forward. There are two major aspects of a ripe moment. First, there needs to be a mutually hurting stalemate: the conflict parties are stuck and none of the sides can make any significant breakthrough through unilateral military means. Such as stalemate needs to be incurring costs for the conflict parties, so that they have incentives to move away from the status quo. The other aspect is the more positive one: there needs to be a way out from their predicament, an openness to, and possibility for, a negotiated settlement. One important point to make here is that ripeness is perceptional: it is about the perceptions of the parties (which is, of course, influenced by the objective situation on the ground, but not necessarily the same thing).

In mediation research, three main strategies have been identified, namely the use of facilitative, formulative, and manipulative strategies (Beardsley et al., 2006). The first strategy is the least intrusive – here the mediator merely helps the parties to bring the process forward by themselves, by, e.g., clarifying the situation if the parties reach a deadlock or by transmitting messages between the parties. A mediator engaged in the second strategy, formulation, helps to coordinate the process more actively. This strategy might involve making recommendations or suggestions to the conflict parties or to steer the parties back to the relevant topics, if they start discussing unrelated issues. Manipulative strategies, the most intrusive strategy, involve using carrots and sticks in order to get to a desirable outcome. The mediator might use threats to coerce the parties or promise them benefits if they make concessions. Resent research has shown that formulative strategies are used most frequently by mediators, while manipulative strategies are used the least often (DeRouen and Bercovitch, 2012). Although less intrusive strategies are used most frequently by mediators, powerful third parties have traditionally been seen as the most effective type of mediators (Touval, 1975). Other research points to a completely different finding – the most durable agreements are those that have been reached with little third-party intrusion in the mediation process (Beardsley, 2008). In order to reach a greater consensus on negotiation strategy and successful outcome, one way could be to distinguish between war types. It is reasonable to assume that the intensity and the type of conflict (civil war or interstate war, etc.), will determine which strategy should be used. This is in line with the findings from Bercovitch and Gartner (2006) that high-intensity conflicts are best settled through the use of manipulative and forcing strategies, while fostering strategies are best used on low-intensity conflicts. Another possible conclusion could be that different strategies should be used in different stages of a war. How do these strategies relate to timing? In **Table 1**, there is an attempt to synthesize the main insights from previous mediation research.

Mediation strategies can be thought of in the following way. Manipulation is needed in order to change the objective aspects

Table 1 Conditions of Ripeness and Mediation Strategies – an overview

	Objective ripeness	*No objective ripeness*
Subjective ripeness	Facilitation	Manipulation
No subjective ripeness	Formulation	Manipulation and formulation

of ripeness, for instance, by increasing the actual costs of the conflict. If the conflict parties have the objective but not the subjective aspects, then formulation is needed – there is a need for persuasion, trying to convince the parties that they are in fact stuck in a costly situation and that reaching a settlement is the best way forward. If both the objective and subjective elements of ripeness exists, then the mediator can use a low-key engagement strategy, merely facilitating the process. On the other hand, if neither the objective nor the subjective elements of ripeness is present, the mediator needs to both formulate the benefits of peace and reaching a settlement, as well as changing the objective aspects through the use of manipulation.

See also: Conflict and Conflict Resolution, Social Psychology of; Conflict and Consensus; Ethnic Conflicts; Peace Keeping; Peace Movements; Peace Processes; Peace-Making in History; Peace; Policing in Divided Societies; Territorial Conflicts.

Bibliography

Beardsley, Kyle, 2008. Agreement without peace? International mediation and time inconsistency problems. American Journal of Political Science 52 (4), 723–740.

Beardsley, Kyle, Quinn, David M., Biswas, Bidisha, Wilkenfeld, Jonathan, 2006. Mediation style and crisis outcomes. Journal of Conflict Resolution 50 (1), 58–86.

Bercovitch, Jacob, 1992. The structure and diversity of mediation in international relations. In: Bercovitch, Jacob, Rubin, J.Z. (Eds.), Mediation in International Relations: Multiple Approaches to Conflict Management. St Martin's Press, New York, pp. 1–29.

Bercovitch, Jacob, Gartner, Scott Sigmund, 2006. Is there method in the madness of mediation: some lessons for mediators from quantitative studies of mediation. International Interactions 32, 329–354.

Bercovitch, Jacob, Houston, Allison, 1996. The study of international mediation: theoretical issues and empirical evidence. In: Bercovitch, Jacob (Ed.), Resolving International Conflicts: The Theory and Practise of Mediation. Lynne Rienner Publishers, Boulder; London, pp. 11–35.

DeRouen, Karl R., Bercovitch, Jacob, 2012. Trends in civil war mediation. In: Hewitt, Joseph J., Wilkenfeld, Jonathan, Gurr, Ted R. (Eds.), Peace and Conflict 2012. Paradigm Publishers, Boulder, CO.

Fearon, James D., 1995. Rationalist explanations for war. International Organization 49 (3), 379–414.

Fisher, Ronald J., 1999. Social-psychological processes in interactive conflict analysis and reconciliation. In: Jeong, H.-W. (Ed.), Conflict Resolution: Dynamics, Process and Structure. Ashgate, Aldershot, pp. 81–104.

Frazier, Derrick V., Dixon, William J., 2006. Third-party intermediaries and negotiated settlements, 1946–2000. International Interactions 32 (4), 385–408.

Greig, Michael, Diehl, Paul F., 2012. International Mediation. Polity, Cambridge, UK.

Höglund, Kristine, Svensson, Isak, 2011. Should I stay or should I go? Termination as a tactic and Norwegian mediation in Sri Lanka Norwegian mediation in Sri Lanka and the termination tactic. Negotiation and Conflict Management Research 4 (1), 12–32.

Kelman, Herbert C., 1997. Social-psychological dimensions of international conflict. In: Zartman, I. William, Rasmussen, L.J. (Eds.), Peacemaking in International Conflict: Methods & Techniques. United States Institute of Peace, Washington, DC, pp. 191–238.

Kleiboer, Marieke, 1996. Understanding success and failure of international mediation. Journal of Conflict Resolution 40 (2), 360–389.

Kydd, Andrew H., 2010. Rationalist approaches to conflict prevention and resolution. The Annual Review of Political Science 13, 101–121.

Miall, Hugh, Ramsbotham, Oliver, Woodhouse, Tom, 1999. Contemporary Conflict Resolution: The Prevention, Management and Transformation of Deadly Conflicts. Polity Press.

Regan, Patrick M., Frank, Richard W., Aydin, Aysegul, 2009. Diplomatic interventions and civil war: a new dataset. Journal of Peace Research 46 (1), 135–146.

Rubin, Jeffrey Z., Pruitt, Dean G., Kim, Sung Hee, 1994. Social Conflict: Escalation, Stalemate, and Settlement. McGram-Hill, New York.

Stedman, John S., 1991. Peacemaking in Civil War: International Mediation in Zimbabwe, 1974–1980. Lynne Rienner Publishers, Boulder, CO.

Svensson, Isak, Mathilda, Lindgren, 2013. Peace from the inside: exploring the role of the insider-partial mediator. International Interactions 39 (5), 698–722.

Svensson, Isak, Wallensteen, Peter, 2010. The Go between: Jan Eliasson and the Styles of Mediation. United Institute of Peace Press, Washington DC.

Touval, Saadia, 1975. Biased intermediaries: theoretical and historical considerations. The Jerusalem Journal of International Relations 1 (1), 51–69.

Touval, Saadia, 1992. The superpowers as mediators. In: Bercovitch, Jacob, Rubin, J.Z. (Eds.), Mediation in International Relations. St. Martin's Press, New York, pp. 232–248.

Wallensteen, Peter, 2011. Understanding Conflict Resolution, third ed. SAGE Publications, London.

Wallensteen, Peter, Svensson, Isak, 2014. Talking peace: international mediation in armed conflicts. Journal of Peace Research 51 (2), 315–327.

Wehr, Paul, Lederach, John P., 1991. Mediating conflict in Central America. Journal of Peace Research 28 (1), 85–98.

Zartman, I. William, 2003. The timing of peace initiatives: hurting stalemates and ripe moments. In: Darby, John, Mac Ginty, Roger (Eds.), Contemporary Peacemaking: Conflict, Violence and Peace Processes. Palgrave Macmillan, pp. 19–29.

Zartman, I. William, Touval, Saadia, 1985. International mediation: conflict resolution and power politics. Journal of Social Issues 41 (2), 27–46.

Zartman, I. William, 1995. Dynamics and constraints in negotiations in internal conflicts. In: Zartman, I. William (Ed.), Elusive Peace: Negotiating an End to Civil Wars. The Brooking Institution, Washington, DC, pp. 3–29.

Zartman, I. William, Touval, Saadia, 2007. International mediation. In: Crocker, Chester A., Hampson, F.O., Aall, P. (Eds.), Leashing the Dogs of War: Conflict Management in a Divided World. USIP Press, Washington, DC, pp. 437–457.

Conflict: Education and Youth

Sarah Dryden-Peterson, Harvard Graduate School of Education, Cambridge, MA, USA
Michelle J Bellino, University of Michigan, Ann Arbor, MI, USA
Vidur Chopra, Harvard Graduate School of Education, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores the multiple and complex relationships between conflict, education, and youth. We begin by outlining conceptions of education, youth, and conflict, as they vary across legal, social, and physical spaces. Drawing on global examples, we explore the ways education plays a role in instigating, reinforcing, escalating, mitigating, resolving, and preventing conflict. We then examine the psychological, social, economic, and political dimensions of the youth experience and the ways in which they are influenced by living in a conflict setting. Finally, we analyze the role of education in understanding the links between youth and conflict, focusing on access to high quality and relevant education that connects to future prospects for individual youth and their communities.

Introduction

Globally, over 300 million young people below the age of 25 years live in settings of armed conflict (Rahim and Holland, 2006). Popular discourse would have it that these young people are both cause of conflicts and hope for their resolutions. They are described as 'explosive' (Urdal, 2012), as harbingers of violence and threats to stability (Weber, 2013) and 'homogeneous blocks of potential menace' (Talbot, 2011). At the same time, young people have also been at the forefront of leading community and social change (Ginwright and James, 2002) and see themselves as future leaders, preparing to rebuild their communities after conflict.

Despite both destructive and constructive potential, youth have often fallen outside of development priorities and programming, both within countries and on a global level (INEE and Refugee Education Trust, 2013; Robinson, 2011). Educational opportunities for young people have been particularly limited, with attention focused instead on younger children. For example, Millennium Development Goals (MDGs) have targeted educational access among primary-age children, and emergency education as a field is often linked to children's rights and protections (Kirk, 2011). Aid has followed, with only 8% of World Bank lending directed to secondary education compared with 43% to primary education in 2005 (Chaffin, 2010). In conflict settings, only 79% of youth are literate compared with 93% of youth in nonconflict countries (UNESCO, 2011).

Among donors, there is often great skepticism about investments in secondary education when primary enrollment is not universal and postprimary opportunities are more expensive and difficult to coordinate. Staff members of United Nations agencies and NGOs describe how they are required to 'play with the words' and advocate for 'early secondary' and 'basic education and life skills' *in lieu* of postprimary opportunities in order to draw in donors (Dryden-Peterson, 2011b). Funding for education in conflict settings is shockingly low, at around 2% of all humanitarian aid (UNESCO, 2011). Within an already small pool of funds, education for youth is funded at low levels. For refugees, for example, funding for secondary education has been increasing, but it remains minimal. In 2010, it made up only 7% of the education budget of the United Nations High Commissioner for Refugees, however the allocated resources doubled by 2012, to 14% of the budget (Dryden-Peterson, 2011b).

As we look toward the renewal of the MDGs, the education of young people is a high priority. The UN High Level Panel report on the post-2015 development agenda identifies five transformative shifts in order to end poverty and promote sustainable development. The first is to 'leave no one behind' (United Nations, 2013). Despite remarkable progress toward universal education at the primary level, 57 million children and young people remain out of school globally – and 42% of them live in conflict-affected settings (UNESCO, 2013). Compared with other low-income countries, youth living in conflict-affected countries are far less likely to access secondary education. Young people in conflict settings who are poor, female, and from ethnic and linguistic minority groups are multiply marginalized, and they access and persist in school at much lower rates than youth in countries not affected by conflict (UNESCO, 2011).

This article explores the relationships between conflict, education, and youth. First, we clarify conceptions of education, youth, and conflict. Second, we explore the multiple and complex relationships between education and conflict. Third, we examine the psychological, social, economic, and political dimensions of the youth experience and the ways in which they are influenced by living in a conflict setting. Finally, we analyze the role of education in understanding the links between youth and conflict, focusing on access to high quality and relevant education that connects to future prospects for individual youth and their communities.

Definitions

We deliberately draw on a broad definition of education as an intentional learning process. Education, by our definition, can be public or private and include primary, secondary, or post-secondary schooling that is accredited by a government or other body. It can also be nonformal, including catch-up classes, tutoring, or other structured but noncredential-bearing activities organized by individuals or communities.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64120-X

Legal definitions of youth vary across country contexts (Auerbach, 2010). International agencies such as the United Nations tend to adopt chronological definitions of young people as those between the ages of 15–24 (United Nations, 2013). Among researchers, there is increased consensus that youth is a social construct wherein the markers defining the beginning and the end of this stage vary greatly across culture and context (Auerbach, 2010; Clark-Kazak, 2009; Sommers, 2012). Recent analysis of Gallup and Afrobarometer survey data demonstrates that the proportion of youth within the same African nations varies based on the indicators used to define the end of youth and the beginning of adulthood (Mo Ibrahim Foundation, 2012). Programs and policies that adopt legally defined, chronological, and age-static definitions of youth risk ignoring social understandings of youth and remaining disconnected from the lived realities and social complexities of young people. For the purposes of this article, we draw on Clark-Kazak's definition of youth as "all individuals who have passed puberty but who have not yet married" (Clark-Kazak, 2009).

Disciplines and organizations define conflict in various ways. The Uppsala Conflict Data Program defines conflict by the use of armed force and the number of battle-related deaths, with more than 25 constituting a conflict and more than 1000 a war (Department of Peace and Conflict Research, 2014). By this definition, in 2012, there were 32 armed inter- and intrastate conflicts globally, in 26 countries. In this same year, there were six on-going wars (Themnér and Wallensteen, 2013). Here we take a broader definition of conflict, which encompasses acute conflict and war, postconflict settings, and situations of organized criminal violence (World Bank, 2011). We use the term 'conflict-affected' to mean countries that are impacted by violence and armed conflict, resulting in weak governance and inequality in resource allocation that negatively affect the lives and livelihoods of young people and the provision of education (see Mundy and Dryden-Peterson, 2011).

Other terms at use in this field include 'fragility' and 'resilience.' The concept of 'fragility' is often used interchangeably with conflict. The concept of fragility, however, is typically much broader such as in the OECD definition of those countries that are "failing to provide basic services to poor people because they are unwilling or unable to do so" (OECD, 2007). While the term fragility has conceptual value, we do not use it here; the causes and dimensions of conflict are multiple, often contested, and recent evidence suggests that conflict and fragility may have divergent impacts on education (Paulson and Shields, 2014). Largely in reaction to the negative framing of 'fragility,' the concept of 'resilience' is in the process of being widely adopted. This approach recognizes that young people face risks but that they also have assets. A resilience approach to education seeks to minimize the risks, while maximizing the assets for both individuals and systems (INEE, forthcoming; World Bank, 2013a,b).

Connections between Education and Conflict: Multiple Faces

The assumption that education serves as an unequivocally protective factor for youth experiencing the risks and consequences of conflict has dominated policy and programming, until recently. Bush and Saltarelli (2000) instigated a rethinking of the relationships between education and conflict, as they defined the 'two faces' of education: one that builds peace and the other that contributes to conflict. In Sierra Leone, for example, a public education system that was geographically unequal and divided by class may well have been a driver of conflict; redressing the inequalities in this system was one of the central demands of the rebel RUF army in building peace after the cessation of conflict (Novelli and Smith, 2011). Further, educational spaces are often destroyed during conflict and schools transformed into militarized zones, where students are forcibly recruited for participation in armed conflict. Based on the UN Secretary General's 2013 report on Children and Armed Conflict, Save the Children found 3643 incidents of reported attacks against education in 17 countries and 90 cases of military use and occupation of schools in 11 countries (Save the Children, 2013). The 'two faces' of conflict have recently evolved into the 'multiple faces' of conflict, with growing agreement that, more than a fundamental right and a basic state service, education can play a role in instigating, reinforcing, escalating, mitigating, resolving, and preventing conflict.

Several theories complicate the notion that education functions dichotomously toward either peace or conflict. Davies (2011) proposes a 'spectrum of impact' (p. 35) to account for the multiple and intersecting ways that educational actors and institutions can simultaneously do both good and harm. Along this spectrum, educational content and structure both play influential roles and can reinforce or undermine one another. The consequences of unequal access to education can be stark. In 2001 in Nepal, for example, the literacy rate among the ruling Brahmins was 70% compared with only 10% among the lower/excluded castes (Novelli and Smith, 2011). Segregated education systems, such as in Bosnia and Herzegovina, may override curricular messages aimed at promoting interethnic harmony (Torsti, 2009).

Unequal access to education, institutionalized segregation, and lack of postprimary education are overt indicators of structural inequality and often become drivers of conflict. The absence of learning opportunities and the gap between educational accreditation and employment are regularly cited as young people's justifications for joining armed groups (Brown, 2010; Collier and Hoeffler, 2004; World Bank, 2011). Curriculum and classroom practice may also be exploited as weapons of war, taking advantage of education's wide reach and the captive audience of youth at a formative stage of moral and socioemotional development (Pingel, 2010; Spink, 2005; Tawil et al., 2004). Everyday classroom interactions can reveal explicit or subtle discrimination toward particular ethnic, religious, or political groups, even long after conflict has ended (King, 2013). History curriculum is especially vulnerable to politicization. Narratives of historical injustice have been repeatedly manipulated by governments and educational actors in ways that increase social distance between diverse groups, while rationalizing or reproducing intergroup grievances (Bam and Visser, 1996; Freedman et al., 2008). For example, in the former Yugoslavia, history education played a critical role in consolidating citizens' collective identity around histories of ethnic and religious suffering (Torsti,

2009). Some studies of contemporary Israel and Palestine indicate that students are socialized into conflict, normalizing intergroup conflict for future generations (Bar-Tal and Rosen, 2009; Mahshi and Bush, 2006).

In addition to having the power to divide, education can work to create peace and prevent conflict. International organizations and donors are increasingly adopting policies that require attention to the 'conflict sensitivity' of educational programming (INEE, 2013; UNICEF, 2012; USAID, 2013), in order to uphold the principle of 'do not harm.' Further, education may play a role in conflict transformation. Along these lines, education interventions in conflict settings coalesce around the role of education in peacebuilding, violence reduction, conflict resolution, social cohesion, multiculturalism, sustainable development, democratic skills, and human rights awareness (see, for example, Novelli and Smith, 2011).

Recovery from conflict is often an opportunity for building anew, with possibilities for innovation and reform that are more challenging in entrenched systems (Nicolai, 2009). There is a connection between rebuilding the physical structures of education and a perception of trust in a new government, often referred to as 'peace dividends' (see, for example, Practical Action Consulting, Save the Children, & CfBT Education Trust, 2012; United Nations Peacebuilding Support Office, 2012). The content and pedagogy of postconflict education can also work toward peace and the prevention of the recurrence of conflict (Blattman et al., 2011).

Youth and Conflict: Challenging Assumptions

One of the defining developmental experiences of youth is the transition from childhood to adulthood and the identity formation processes that accompany this critical developmental stage (Johnson et al., 2011). Due to the relative lack of tailored services for youth, young people are caught between programming designed for children or for adults, leaving them either excluded and 'on the margins' (Sommers, 2007) or funneled into programs that are mismatched for their developmental stage. Further complicating youth as a target population, legal definitions and culturally specific roles for youth shift (see, Clark-Kazak, 2009), more so in conflict settings than otherwise. In times of conflict, youth do not experience typical childhoods but prematurely enter adult worlds; conversely, conflicts can delay traditional rites of passage and hinder access to markers of adulthood, so that youth cannot become socially recognized adults (Sommers, 2012).

What is the relationship between youth populations and propensity for conflict? There are competing theories. Noncausal empirical research demonstrates that countries with youth bulges and poor economic performance are more susceptible to domestic armed conflict (Urdal, 2006) and democratic instability (Weber, 2013). Two dominant theoretical perspectives have attempted to explain this hypothesis. The first is an economic or opportunity theory approach, where a sudden supply of youth labor in an economic system unable to generate enough jobs and absorb a surge of youth labor supply will render many unemployed. As a consequence, armed groups tend to fill this vacuum by providing alternate income generating opportunities (Brown, 2010; Collier, 2007; Collier and Hoeffler, 2004). The second is a sociobiological approach, where young men (15–29 years), given their developmental stage, are considered to have higher propensities for extremist attitudes and intergroup hate and violence (Weber, 2013). Collier states that "[b]eing young, being uneducated, and being without dependents" may make one more likely to engage in political violence (2007). For some time, these two theories have dominated much of the discourse on youth in conflict, even though researchers have attempted to nuance them, as we discuss below (Beber and Blattman, 2013; Cincotta, 2008; Sommers, 2012; Urdal, 2006).

These theories do not account for three important dimensions of youth in conflict settings. First, they do not account for the agency and resilience that youth exhibit in shaping their life choices even under severe structural constraints and stressful situations (Blum and Blum, 2009). Second, they do not sufficiently illuminate the various ways that education systems interact with youth experience and thus reproduce and/or interrupt conflict. Finally, these theories do not account adequately for the interactions of various social, political, and historical circumstances. For example, in both stable Botswana and conflict-affected Democratic Republic of Congo (DRC), 22% of the population is between the ages of 15 and 24 (Central Intelligence Agency, 2013). Yet these oversized youth populations have not erupted into conflict in both places; while DRC has been engulfed in conflict for more than two decades, Botswana has been stable. A 'youth bulge' combined, as in DRC, with well-established armed groups and a lack of education and employment possibilities can shift the opportunity structure in favor of participating in conflict, over prosocial choices. Similarly, the 'Arab Spring' cannot be considered only a function of a youth bulge but is also associated with stagnating economies as well as corruption and poor governance in many of the countries that experienced these widespread protests (Pollack, 2011). These contexts demonstrate how conflict is embedded in particular sociopolitical contexts and historical moments.

Further theories circulate around youth vulnerability and susceptibility to joining armed groups based on adolescent desires to consolidate their social identity during a critical identity formation stage. Identity threat is among the many drivers of conflict (Novelli and Lopes Cardozo, 2008). Relying on developmental explanations, some argue that in an attempt to explore and assert their identities, youth are easily influenced and conformist in their attitudes, thereby exhibiting greater tendencies toward extremism and recruitment by armed rebel groups (Blattman and Annan, 2010). However, mounting evidence from the field of adolescent and youth development indicates that identity development is progressive and continues even in adulthood. Conceptions of ourselves 'vary along a continuum' (Davies, 2008: p. 187) such that different aspects of identity emerge and change over time along this developmental continuum (Meeus, 2011). This calls into question the sociobiological mechanisms that underpin the 'youth bulge' hypothesis. If identity statuses progress over adolescence and well into adulthood, there is little support for the argument that identity-driven attitudes of inter- and intragroup hatred and violence are not malleable among youth.

Conflict, Youth, and Education: A Synthesis

In examining the role of education for youth in conflict settings, three recurring challenges emerge. First is the limited access to education opportunities. Second is the quality of education, which includes dimensions of content and pedagogy and overall relevance of learning to postschooling prospects. Third is the dichotomous ways in which education can include or exclude youth from community.

Access to education for youth remains a challenge, particularly postprimary opportunities. Seventy-four million adolescents of lower secondary school age are out of school (Robinson, 2011). Lack of access to education is further exacerbated in times of conflict, when children's learning needs are often prioritized over those of youth (INEE and Refugee Education Trust, 2013).

Despite these limited opportunities, the 2011 EFA Global Monitoring Report argued that secondary school is the "cornerstone of education for youth" and that "formal education is the most effective base for developing learning and life skills" (UNESCO, 2011: p. 54). Education for youth is critical, both for individuals and societies. The existence of secondary school opportunities acts as a motivation to enroll in, and complete, primary school (Chaffin, 2010; Robinson, 2011). Further, the economic returns to secondary education are large. At the individual level, each additional year of formal education on average adds about 10% to earnings, and secondary education adds 20% for low-income individuals. The rates of return for secondary education are particularly high in sub-Saharan Africa, at 25%, and in non-OECD Asia, at 16% (Psacharopoulos and Patrinos, 2002: pp. 2, 12). Related, while individual returns to education can be inequitably distributed, the social returns to secondary education are also high: 18.4% in sub-Saharan Africa and 11.1% in non-OECD Asia (Psacharopoulos and Patrinos, 2002: p. 12). In addition, without continued investment in education of youth, the cadres of high-quality future teachers for both primary and secondary education are limited, which poses a major challenge to postconflict reconstruction (Buckland, 2005; Shriberg, 2007).

Education also affords youth opportunities for civic participation and quality of life and, in these ways, it is protective both for individuals and for societies (IIEP, 2011; INEE Working Group on Education and Fragility, n.d.). Education provides youth the opportunity to think about the future (Winthrop and Kirk, 2008) and to imagine what is possible (Martone and Neighbor, 2006: p. 3; Waters and Leblanc, 2005), especially displaced populations such as refugees who have no legal citizenship status. Particularly in situations where entry into the labor market is limited for young people, "they need the stimulus and challenge of education to absorb their energies and lessen their frustrations and anxiety about the future" (IIEP, 2006: p. 2). A 17-year-old Congolese boy who arrived in 2008 in Kyangwali refugee settlement in Uganda expressed the situation this way: "When we reached here in Kyangwali, life became really hard. Because there is nothing to do: in the morning you wake up, you are moving up and down, like someone who is looking to go somewhere but actually you are not going anywhere" (Wettstein, 2011). This sentiment speaks to education both as a process that gives order to the everyday life of young people, and as a pathway that leads to future opportunity.

In seeking to address access to education for youth, recent research highlights the necessity of a focus on the quality and relevance of that education (Dryden-Peterson, 2011a; Robinson, 2011; Winthrop and Matsui, 2013). Indeed, access and quality learning cannot be viewed as a dichotomy. For example, conflict in Cambodia not only interrupted educational attainment levels for the older age groups, but also contributed to subsequently lowering educational attainment rates among their children (UNESCO, 2012). The 2013–14 EFA Global Monitoring Report notes the presence of a global learning crisis wherein 175 million young people globally are unable to read all or part of a sentence (UNESCO, 2014). Quality education is education that is relevant, in that it builds foundational and transferable skills that allow young people to successfully transition from school to work, to build self-reliance, and to contribute to society (UNESCO, 2012).

For youth, there are multiple dimensions of relevance. First, building on a core adolescent need to construct identity through self-reflection and in collaboration with peers, educational programs that recognize the complexity of self-identity as a vital dimension of youth's psychosocial well-being are more effective (Davies, 2008; Meeus, 2011). Second, and related, the content and pedagogy of education can be protective for youth, building physical, psychological, and cognitive health. Drawing from longitudinal research and causal studies, Reyes (2013) concludes that the presence of positive cognitive, social, and behavioral skills in contexts of adversity is particularly protective. In particular, the skills of meaning making, behavior recognition, increased connectedness, and social harmony can enable success in later life. Third, constructing schools in conflict or postconflict settings with the hope of bringing previously divided groups within a shared, common space will fail to achieve its purposes, if youth within these spaces are not working toward a shared goal or learning to negotiate differences among themselves (Davies, 2013). Physical integration is but the first step toward social integration, for which curriculum, pedagogy, and the accompanying teacher training are the central elements (Dryden-Peterson, 2011b; Dryden-Peterson and Hovil, 2004). Experience from Burundi demonstrates that despite constructing extra classrooms, providing school uniforms and scholarships, many youth who were in exile in Tanzania and returned to Burundi remained out of school. The physical infrastructure was in place, but the language of instruction was different from what they were taught in Tanzania, and, consequently, learning remained inaccessible. Additionally, there were social tensions between those who stayed in Burundi and those who returned from exile (Refugee Education Trust, 2012).

The fourth dimension of relevance is the connection between what is learned in schools and livelihoods. A mismatch between skills, job opportunities, and expectations can lead to false hope, breeding immense frustration among refugee young people, and to recruitment into armies and armed militias when that is seen as the more secure livelihood options (Brown, 2010; Sommers, 2006b). In conflict settings,

youth education – especially for young people who have missed out on school as a result of conflict – is often seen as a context best suited to technical and vocational skills development (TVSD). Ideally, TVSD includes both 'hard' and 'soft' skills, "developed within a 'joined-up,' integrated development and delivery framework that seeks to improve livelihoods, promote inclusion into the world of work and that supports community and individual agency" (Conflict and Education Research Group, 2007: p. 2; see also, Lyby, 2003). This kind of training is varied across conflict contexts: among refugees in the city of Kuala Lumpur, Malaysia, it includes baking and computer classes; and, in the context of postconflict Mauritania, there are short (3-week) baking, tailoring, hairdressing, and mechanics trainings (Rahman, 2011; Rezzonico, 2011).

TVSD programs have demonstrated mixed success, can be quite expensive, and in some cases have been dramatically mismatched to market needs. For example, the Norwegian Refugee Council's Youth Education Pack Programme (YEP) is a 1-year full time program focused on literacy/numeracy, life skills, and basic vocational skills (NRC, n.d.). An evaluation in Liberia found that few of the youth trained with YEP could sustain themselves on the income from the new skills they learned in this program; although the skills were relevant to the local economy, the market was oversaturated with YEP graduates and diversification of training was needed (Moberg and Johnson-Demen, 2009). 'Skill mismatch' results from the tendency to train more students every year in the same skills, as in the case where plumbers were trained in Liberia. A former child soldier trained as a plumber reflected: "it's not easy to find work in plumbing, you know? … because most of Liberia doesn't have plumbing" (Conflict and Education Research Group, 2007: p. 13).

The third recurring challenge for education of youth in conflict settings is related to community building. The youth in conflict discourse is dominated by meta-narratives of 'fight or flight' and bellicosity (Clark-Kazak, 2011; Sommers, 2006a) that steer the approaches in conceptualizing learning opportunities for young people. However, causal evidence in the fields of experimental economics and political science finds increased instances of community and civic participation in some postconflict contexts. For example, war victims in postconflict Sierra Leone and Uganda are more likely to engage in their communities through participation in school committees, registering to vote, and in attending community meetings than nonwar victims within the country (Bellows and Miguel, 2008; Blattman and Annan, 2010). Research has established that youth's social connection with their communities can reinforce civic participation, just as increased participation can deepen community engagement (Syversten, 2006). In fact, the drive to find one's niche within society is a fundamental developmental need (Evans, 2007). Yet, many youth across different contexts find a paucity of opportunities provided by their environments (Camino and Zeldin, 2002). In this regard, learning opportunities – formal and informal – serve the role of 'mediating spaces' where youth practice citizenship by exploring what it means to be part of a community, negotiate differences and practice the rights and obligations associated with being associated with a collective (Flanagan et al., 2011). These data suggest that educational institutions are an important space for young people to reimagine themselves, their futures, and their relationship to their local communities, the state, and the global society.

Education programs targeted to former child and youth soldiers often focus on this element of community building (Hansen et al., 2002; Rahim and Holland, 2006). Militias and rebel groups that forcibly recruit young people usually seek to dismantle existing community as a mechanism for compliance and retention, such as when young people are forced to kill members of their own families (Rahim and Holland, 2006). Disarmament, demobilization, and reintegration (DDR) programs seek to rebuild community through education focused on strong relationships, for example. In war-torn Sierra Leone schools created peer supports for youth who had lost their immediate and extended families. This program also played an important role in bringing together former child soldiers and victims (Betancourt et al., 2008). However, education alone cannot integrate former combatants into communities unless psychosocial supports are integrated within the education system to bolster teachers and learners' coping capacities. Some studies of DDR programs (see, for example, Rahim and Holland, 2006) have found that young people demonstrate greater benefits from support structures within their communities than from institutional programming. Demobilizing ex-combatants depends on their capacity to meaningfully reintegrate into society.

Conclusion

The dichotomy of education as a catalyst for or a mitigator of conflict mirrors an underlying set of beliefs about youth's dual role in conflict: that youth are either a problem to be resolved or a solution to be cultivated. Across various fields of study, youth are cast as destabilizers, troublemakers, idle, and criminal, or they are treated as passive victims with stolen innocence (Schwartz, 2010). When prospects for peace emerge, hopes are pinned on youth as the vanguard of the future, and educational recommendations follow suit. Research demonstrates, however, that youth are not confined to these roles, which are outgrowths of the age-old perpetrator–victim divide. Educational policies nevertheless often homogenize youth experience on these somewhat limited and simplistic tropes in order to design educational interventions.

Educational opportunities must consider youth's unique developmental stage and move beyond the victim–perpetrator dialectic in order to maximize the potential for education to mitigate and prevent conflict. Critical elements of education, designed with the needs of youth in mind, acknowledge the central identity-shaping role of coming of age in conflict while at the same time grant young people the agency to define themselves outside of these roles and contexts of violence. Beyond basic skills, schools can provide critical opportunities for youth from traditionally divided groups to make collective and individual meaning of their experiences during conflict, shape new identities, and envision future trajectories.

Challenges remain, even in exemplary cases where educational access, quality, and relevance cooccur. Structural barriers

remain influential, such as the capacities to finance education for youth, access schools, retain quality teachers, determine relevant curriculum, and guarantee basic safety. Those youth who reintegrate in their home communities and seek to change their life trajectories may confront rejection and stigma, even when they are reunited and possess strong social support systems. There are also pragmatic concerns that school learning is disconnected from postgraduation opportunities. Many of these challenges are cross-cultural and resonate across states, even across age cohorts. No matter what context, conflict changes the way that young people relate to themselves, their societies, and their futures. Likewise, conflict shifts how societies envision youth and their roles in creating a postconflict future.

See also: Child Soldiers as Victims and Actors: Lessons from Ethiopia; Child Soldiers; Community Resilience to Disasters; Development and Employment: A Mutual Relationship; Early Education and Care in the World Regions; Human Rights: Legal Aspects; Informal Work in Developing Countries; Juvenile Justice: International Law and Children's Rights; Neurosociology; Peace; Skills Training: Public Policy Aspects; War, Political Violence, and Effects on Children; Youth Culture, Anthropology of.

Bibliography

Auerbach, J., 2010. Flowing into the State: Returning Refugee Youth and Citizenship in Angola. RSC Working Paper. University of Oxford, Oxford.

Bam, June, Visser, Pippa, 1996. A New History for a New South Africa. Kagiso Publishers, Cape Town, Johannesburg. Distributed by Thorold's Africana Books.

Bar-Tal, Daniel, Rosen, Yigal, 2009. Peace education in societies involved in intractable conflicts: direct and indirect models. Review of Educational Research 79 (2), 557–575.

Beber, B., Blattman, C., 2013. The logic of child soldiering and coercion. International Organization 67, 65–104.

Bellows, J., Miguel, E., 2008. War and Local Collection Action in Sierra Leone National Bureau of Economic. Research Working Paper. National Bureau of Economic Research, Cambridge.

Betancourt, T., Simmons, S., Ivelina, B., Brewer, S., Iweala, U., Soudiere, M.D.L., 2008. High hope, grim reality: reintegration and the education of former child soldiers in Sierra Leone. Comparative Education Review 52 (4), 565–587.

Blattman, C., Annan, J., 2010. The consequences of child soldiering. Review of Economics and Statistics 42 (4), 882–898.

Blattman, C., Hartman, A., Blair, R., 2011. Can We Teach Peace and Conflict Resolution? Results from a Randomized Evaluation of the Community Empowerment Program (CEP) in Liberia: A Program to Build Peace, Human Rights, and Civic Participation Evidence from Randomized Evaluations of Peacebuilding in Liberia: Policy Report. Innovations for Poverty Action and Yale University, New Haven.

Blum, L., Blum, R., 2009. Resilience in adolescence. In: DiClemente, R.J., Santelli, J.S., Crosby, R.A. (Eds.), Adolescent health: Understanding and preventing risk behaviors. John Wiley & Sons Inc., San Francisco, CA.

Brown, G.K., 2010. Education and Violent Conflict: Background Paper for EFA Global Monitoring Report 2011. UNESCO, Paris.

Buckland, Peter, 2005. Reshaping the Future: Education and Post-conflict Reconstruction. World Bank, New York.

Bush, Kenneth D., Saltarelli, Diana, 2000. The Two Faces of Education in Ethnic Conflict: Towards a Peacebuilding Education for Children. UNICEF, Innocenti Research Centre, Florence.

Camino, L., Zeldin, S., 2002. From periphery to center: pathways for youth civic engagement in the day-to-day life of communities. Applied Developmental Science 6 (4), 213–220.

Central Intelligence Agency, 2013. The World Factbook. CIA, Washington, DC.

Chaffin, Joshua, 2010. Framing Paper 1: Education and Opportunity: Post-primary and Income Growth. INEE, New York.

Cincotta, R., 2008. Half a Chance: Youth Bulges and Transitions to Liberal Democracy Wilson Center's Environmental Change and Security Program. Princeton University, Princeton, NJ.

Clark-Kazak, Christina R., 2009. Towards a working definition and application of social age in international development studies. Journal of Development Studies 45 (8), 1307–1324.

Clark-Kazak, Christina R., 2011. Recounting Migration: Political Narratives of Congolese Young People in Uganda. McGill-Queen's University Press, Montreal.

Collier, P., 2007. The Bottom Billion: Why the Poorest Countries Are Failing and What Can Be Done about It. Oxford University Press, Oxford; New York.

Collier, P., Hoeffler, A., 2004. Greed and grievance in civil war. Oxford Economic Papers 56 (4), 563–595.

Conflict and Education Research Group, 2007. Education for Livelihood and Civic Participation in Post-conflict Countries: Conceptualizing a Holistic Approach to TVET Planning and Programming in Sub-Saharan Africa. UNESCO-UNEVOC: International Centre for Technical and Vocational Education and Training, Bonn.

Davies, Lynn, 2008. Educating against extremism: towards a critical politicisation of young people. International Review of Education 55, 183–203.

Davies, Lynn, 2011. Can education interrupt fragility? Towards the resilient and adaptable state. In: Mundy, K., Dryden-Peterson, S. (Eds.), Educating Children in Conflict Zones: Research, Policy, and Practice for Systemic Change (A Tribute to Jackie Kirk). Teachers College Press, New York, pp. 33–48.

Davies, Lynn, 2013. The strength of weak ties: networks, activism and critical idealism. In: Paper Presented at the 12th UKFIET International Conference on Education and Development: Education and Development Post 2015 – Reflecting, Reviewing, Re-visioning. University of Oxford.

Department of Peace and Conflict Research, 2014. Definitions. From: http://www.pcr.uu.se/research/ucdp/definitions/ (retrieved 14.03.14.).

Dryden-Peterson, Sarah, 2011a. Refugee children aspiring toward the future: linking education and livelihoods. In: Mundy, K., Dryden-Peterson, S. (Eds.), Educating Children in Conflict Zones: Research, Policy, and Practice for Systemic Change (A Tribute to Jackie Kirk). Teachers College Press, New York, pp. 85–99.

Dryden-Peterson, Sarah, 2011b. Refugee Education: A Global Review. UNHCR, Geneva.

Dryden-Peterson, Sarah, Hovil, Lucy, 2004. A remaining hope for durable solutions: local integration of refugees and their hosts in the case of Uganda. Refuge 22, 26–38.

Evans, S.D., 2007. Youth sense of community: voice and power in community contexts. Journal of Community Psychology 35 (6), 693–709.

Flanagan, C.A., Martinez, M.L., Cumsille, P., Ngomane, T., 2011. Youth civic development: theorizing a domain with evidence from different cultural contexts. New Directions for Child and Adolescent Development 134, 95–109.

Freedman, Sarah Warshauer, Weinstein, Harvey M., Murphy, Karen, Longman, Timothy, 2008. Teaching history after identity-based conflicts: the Rwanda experience. Comparative Education Review 52 (4), 663–690.

Ginwright, S., James, T., 2002. From assets to agents of change: social justice, organizing, and youth development. New Directions for Youth Development 96, 27–46. http://dx.doi.org/10.1002/yd.25.

Hansen, A., Nenon, J., Wolf, J., Sommers, M., 2002. Final Evaluation of the Office of Transition Initiatives' Program in Sierra Leone: Final Report. USAID, Washington, DC.

IIEP, 2006. Guidebook for Planning Education in Emergencies and Reconstruction. UNESCO, Paris.

IIEP, 2011. Understanding Education's Role in Fragility: Synthesis of Four Situational Analyses of Education and Fragility: Afghanistan, Bosnia and Herzegovina, Cambodia, Liberia. IIEP, Paris.

INEE, 2013. INEE Guidance Note on Conflict Sensitive Education. INEE, New York.

INEE. Education-for-Resilience. An INEE Concept Note New York, INEE, Unpublished working paper.

INEE, Refugee Education Trust, 2013. Key Advocacy Messages of the INEE-RET Round Table, Youth Education 2015 and Beyond – Setting Priorities. Refugee Education Trust, Geneva.

INEE Working Group on Education and Fragility, n.d. The Multiple Faces of Education in Conflict-Affected and Fragile Contexts. INEE, New York.

Johnson, M.K., Crosnoe, R., Elder, G.H., 2011. Insights on adolescence from a life course perspective. Journal of Research on Adolescence 21 (1), 273–280.

King, Elisabeth, 2013. From Classrooms to Conflict in Rwanda. Cambridge University Press, Cambridge.

Kirk, Jackie, 2011. Education and fragile states. In: Mundy, K., Dryden-Peterson, S. (Eds.), Educating Children in Conflict Zones: Research, Policy, and Practice for Systemic Change (A Tribute to Jackie Kirk). Teachers College Press, New York, pp. 15–31.

Lyby, Eric, 2003. Vocational training for refugees: a case study from Tanzania. In: Crisp, J., Talbot, C., Cipollone, D.B. (Eds.), Learning for a Future: Refugee Education in Developing Countries. UNESCO, Paris.

Mahshi, Khalil, Bush, Kim, 2006. The Palestinian uprising and education for the future. In: Piper, B., Dryden-Pererson, S., Kim, Y.-S. (Eds.), International Education for the Millennium: Toward Access, Equity, and Quality. Harvard Educational Review, Cambridge, MA, pp. 221–234.

Martone, Gerald, Neighbor, Hope, 2006. The emergency alibi. In: Bayefsky, A.F., Fitzpatrick, J., Helton, A.C. (Eds.), Human Rights and Refugees, Internally Displaced Persons, and Migrant Workers: Essays in Memory of Joan Fitzpatrick and Arthur Helton. M. Nijhoff, Leiden; Boston.

Meeus, W., 2011. The study of adolescent identity formation 2000 – 2010: a review of longitudinal research. Journal of Research on Adolescence 21 (1), 75–94.

Mo Ibrahim Foundation (Producer), 2012, August 2013. African Youth: Fulfilling the Potential. Retrieved from: http://www.moibrahimfoundation.org/downloads/2012-facts-and-figures.pdf.

Moberg, Liv, Johnson-Demen, Annie, 2009. "Small-Small Steps" to Rebuild Communities with YEP: NRC Youth Education Pack Project in Liberia: Post Graduates and Income Generating Activities. NRC, Oslo.

Mundy, Karen, Dryden-Peterson, Sarah, 2011. Educating children in conflict zones: an overview and introduction. In: Mundy, K., Dryden-Peterson, S. (Eds.), Educating Children in Conflict Zones: Research, Policy, and Practice for Systemic Change (A Tribute to Jackie Kirk). Teachers College Press, New York.

Nicolai, Susan, 2009. Opportunities for Change: Education Innovation and Reform during and after Conflict. IIEP, Paris.

Novelli, Mario, Cardozo, Lopes, Mieke, T.A., 2008. Conflict, education and the global south: new critical directions. International Journal of Educational Development 28 (4), 473–488.

Novelli, Mario, Smith, Alan, 2011. The Role of Education in Peacebuilding: A Synthesis Report of Findings from Lebanon, Nepal and Sierra Leone. UNICEF, New York.

NRC, n.d. NRC core activity education: youth education pack (YEP). In: Brief, Background, Concept, Evaluations and Development Plans. NRC, Oslo.

OECD, 2007. Glossary of Statistical Terms: Fragile States. From: http://stats.oecd.org/glossary/detail.asp?ID=7235 (retrieved 06.04.11.).

Pingel, Falk, 2010. The power of the curriculum. In: Cahill, K.M. (Ed.), Even in Chaos: Education in Times of Emergency, first ed. Fordham University Press; Center for International Humanitarian Cooperation, New York, pp. 109–135.

Pollack, K.M., 2011. The Arab Awakening: America and the Transformation of the Middle East. Brookings Institution Press, Washington, DC.

Practical Action Consulting, Save the Children, & CfBT Education Trust, 2012. Synthesis Research Report: State-Building, Peace-Building and Service Delivery in Fragile and Conflict-Affected States. Practical Action Consulting, Save the Children, CfBT Education Trust, London.

Psacharopoulos, G., Patrinos, H.A., 2002. Returns to Investment in Education: A Further Update World Bank Policy Research Working Paper. World Bank, Washington, DC.

Rahim, A., Holland, P., 2006. Facilitating Transitions for Children and Youth. In: Lessons from Four Post- Conflict Fund Countries World Bank Social Development Papers, vol. 34. World Bank, Washington, DC.

Rahman, Joanna, 2011. Access to Education for Children under the Protection of the United Nation High Commissioner for Refugees (UNHCR): Case Study in Kuala Lumpur, Malaysia. University of Neuchâtel, Neuchâtel.

Refugee Education Trust, 2012. La reintegration socio-scolaire des jeunes rapatriés au Burundi. RET, Bujumbura.

Reyes, Joel, 2013. What Matters Most for Students in Contexts of Adversity: A Framework Paper. World Bank, Washington, DC.

Rezzonico, Laura, 2011. Access to Education for Children under the Protection of the United Nation High Commissioner for Refugees (UNHCR): Case Study: Mauritanian Returnees from Senegal. University of Neuchâtel, Neuchâtel.

Robinson, Jenny Perlman, 2011. A Global Compact on Learning: Taking Action on Education in Developing Countries. Brookings Institution, Washington, DC.

Save the Children, 2013. Attacks on Education: The Impact of Conflict and Grave Violations on Children's Future. Save the Children, London.

Schwartz, S., 2010. Youth in Post-conflict Reconstruction: Agents of Change. United States Institute of Peace, Washington, DC.

Shields, Robin, Paulson, Julia, 2014. 'Development in reverse'? A longitudinal analysis of armed conflict, fragility and school enrolment. Comparative Education 1–19. http://dx.doi.org/10.1080/03050068.2014.953314.

Shriberg, Janet, 2007. Teaching Well? Educational Reconstruction Efforts and Support to Teachers in Postwar Liberia. International Rescue Committee, New York.

Sommers, Marc, 2006a. Embracing the Margins: Working with Youth Amidst War and Insecurity. Brookings Institution, Washington, DC.

Sommers, Marc, 2006b. Fearing Africa's Young Men: The Case of Rwanda Social Development. World Bank, Washington, DC (Conflict Prevention & Reconstruction).

Sommers, Marc, 2007. Creating programs for Africa's urban youth: the challenge of marginalization. Journal of International Cooperation in Education 10 (1), 19–31.

Sommers, Marc, 2012. Stuck: Rwandan Youth and the Struggle for Adulthood. University of Georgia Press, Published in Association with the United States Institute of Peace, Athens, GA; Washington, DC.

Spink, Jeaniene, 2005. Education and politics in Afghanistan: the importance of an education system in peacebuilding and reconstruction. Journal of Peace Education 2 (2), 195–207. http://dx.doi.org/10.1080/17400200500185794.

Syversten, A., 2006. Civic identity development. In: Sherrod, L., Flanagan, C.A., Kassimir, R. (Eds.), Youth Activism: An International Encyclopedia. Greenwood, Westport, CT.

Tawil, Sobhi, Harley, Alexandra, Braslavsky, Cecilia, 2004. Education, Conflict and Social Cohesion. UNESCO, International Bureau of Education, Geneva.

Talbot, Christopher, 2011. Education for Young People: Are Both Positive and Negative Views Highlighted? In: INEE (Ed.), Network News Blog, Vol. 2011. INEE, New York.

Themnér, Lotta, Wallensteen, Peter, 2013. Armed conflict, 1946–2012. Journal of Peace Research 50 (4).

Torsti, Pilvi, 2009. Segregated education and texts: a challenge to peace in Bosnia and Herzegovina. International Journal on World Peace XXVI (2), 65–82.

UNESCO, 2011. Education for All Global Monitoring Report 2011: The Hidden Crisis: Armed Conflict and Education. UNESCO, Paris.

UNESCO, 2012. EFA Global Monitoring Report 2012: Youth and Skills: Putting Education to Work. UNESCO, Paris.

UNESCO, 2013. EFA Global Monitoring Report 2013: Children still battling to go to school. UNESCO, Paris.

UNESCO, 2014. EFA Global Monitoring Report 2013–14, Teaching and Learning: Achieving Quality for All. UNESCO, Paris.

UNICEF, 2012. Conflict Sensitivity and Peacebuilding in UNICEF, Technical Note. UNICEF, New York.

United Nations, 2013. A New Global Partnership: Eradicate Poverty and Transforma Economies through Sustainable Development, the Report of the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda. United Nations, New York.

United Nations Peacebuilding Support Office, 2012. Peace Dividends and Beyond: Contributions of Administrative and Social Services to Peacebuilding. Peacebuilding Support Office, New York, United Nations.

Urdal, H., 2006. A clash of generations? Youth bulges and political violence. International Studies Quarterly 50 (3), 607–630.

Urdal, H., 2012. A Clash of Generations? Youth Bulges and Political Violence. United Nations Department of Economic and Social Affairs.

USAID, 2013. Diagnostic Tool for Assessing Extent of Conflict Sensitivity for Education Programs. USAID, Washington, DC.

Waters, Tony, Leblanc, Kim, 2005. Refugees and education: mass public schooling without a nation-state. Comparative Education Review 49 (2), 129–147.

Weber, H., 2013. Demography and democracy: the impact of youth cohort size on democratic stability in the world. Democratization 20 (2), 335–357.

Wettstein, Manon, 2011. Access to Education for Children under the Protection of the United Nation High Commissioner for Refugees (UNHCR): Case Study in Kyangwali Refugee Settlement, Uganda. University of Neuchâtel, Neuchâtel.

Winthrop, R., Kirk, J., 2008. Learning for a bright future: schooling, armed conflict, and children's well-being. Comparative Education Review 52 (4).

Winthrop, R., Matsui, E., 2013. A New Agenda for Education in Fragile States Center for Universal Education Working Paper. Brookings, Washington, DC.

World Bank, 2011. World Development Report: Conflict, Security, and Development. World Bank, Washington, DC.

World Bank, 2013a. Learning and Resilience: The Crucial Role of Social and Emotional Well-Being in Contexts of Adversity. Education Notes. World Bank, Washington, DC.

World Bank, 2013b. RES-360°: Resilience in Education Systems. World Bank, Washington, DC.

Conflict: Organizational

Stewart Clegg, University of Technology Business School, Sydney, NSW, Australia
Elisabeth Naima Mikkelsen, Copenhagen Business School, Frederiksberg, Denmark
Graham Sewell, University of Melbourne, Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines four contemporary treatments of the problem of organizational conflict: social psychological, anthropological, neo-Darwinian, and neo-Machiavellian. Social psychological treatments of organizational conflict focus on the dyadic relationship between individual disputants. In contrast, anthropological treatments take a more socially and historically embedded approach to organizational conflict, focusing on how organizational actors establish negotiated orders of understanding. In a break with the social psychological and anthropological approaches, neo-Darwinians explain the characteristics of organizational conflict by appealing to the concept of natural selection: all forms of organizational behavior, including conflictual relations, stem from the effects of heritable traits associated with a universal human nature. Finally, this article proposes a neo-Machiavellian view of organizational conflict where members of an organization are seen as politicized actors engaged in power struggles that continually ebb and flow.

We make two explicit assumptions in this article. First, a life conducted in, or mediated through, organizations is one of the defining aspects of being a modern citizen. Second, despite the best intentions of some theorists and practitioners, relations in organizations will always entail a degree of conflict and political maneuvering because "[r]easonable people can be expected to disagree on major, uncertain and ambiguously defined issues, and to argue strongly for their personal convictions" (Buchanan and Badham, 1999, p. 21). As a corollary of these assumptions it is likely that we have all been participants in, or observers of, organizational conflict at some time and, in this respect, we may have an intuitive feel for what creates and diffuses conflict in organizations. In this article, however, we adumbrate the key characteristics of the following perspective on organizational conflict: (1) the orthodox social psychology approach; (2) ethnographies of conflict that incorporate the social, cultural, and historical dimensions of organization; (3) neo-Darwinian studies that see conflict as an expression of our essential human nature; and (4) a neo-Machiavellian approach where conflict reflects the politicized character of organizational relations. A general theme of this article is that we should be wary of explaining conflict on the basis of universal or abstract laws of behavior at the expense of a subtle consideration of organizational history and context. (**Table 1** summarizes the four perspectives.)

Organizational Conflict and Social Psychology

Orthodox social psychological approaches explore conflict through a hierarchy of dyadic interactions that start with the interpersonal and stretch to relations between discrete social entities, such as interest groups, unions, and even nation-states. The ways in which these interactions are played out can be mapped using the dimensions of *Process* and *Structure*, thus delineating the two main social psychological models of organizational conflict (Thomas, 1976, 1994). The *Process* model is primarily empirical in approach, as it requires us simply to observe and record conflict as it unfolds. This involves categorizing various generic stages before relating these dynamic circumstances to particular outcomes or 'settlements.' These outcomes tend to be classified in economic terms so that resources – frequently the focus of the conflict in the first place under this approach – are distributed between parties, usually in a zero-sum arrangement. In contrast, the *Structural* model involves identifying the antecedents that shape the dynamics of a particular dyadic conflict. These include cognitive factors (for example, the actors' awareness of issues and possible settlements, their intentions, etc.) and affective states such as emotions and empathy. Underlying the operation of these structural features of conflict is the assumption of bounded rationality: on the basis of the information available to them,

Table 1 A summary of the four main perspectives on organizational conflict

Perspective on organizational conflict	*Basic assumptions*	*Organizational implications*
Social Psychological	Conflict occurs between actors (e.g., individuals, organizational units, organizations, etc.) that are pursuing heterogeneous goals	Negative and pathological if poorly managed Positive and desirable if well managed
Ethnographic	Conflict an epiphenomenon of a negotiated order associated with a specific organizational context	Organizational conflict, a manifestation of local patterns of culture, identity, norms, etc.
Neo-Darwinian	Conflict an unavoidable consequence of our genetically inherited human nature	An organizations should be designed by taking into account the impact of its structure on our inherited human nature
Neo-Machiavellian	Conflict an outcome of the political maneuvring of actors who have varying degrees of political competence	In almost any organizational setting, the outcome of conflict will be shaped, in large part, by the deployment of political skill

actors in a conflict will move toward satisfying outcomes. Thus, as Thomas (1994) suggests, conflicts are decision-making processes that evaluate possible settlements using a combination of rational–instrumental considerations (what in this for *me*?) with normative considerations (is this a 'good' outcome?).

In practical terms, the *Structural* and *Process* models are inextricably bound as the former often provides the theoretical basis for the latter. For example, this implied causal relationship is evident in approaches based on assumptions of bounded rationality such as the behavioral theory of the firm (Cyert and March, 1992) where dividing an organization along, let us say, functional or geographical (i.e., structural) lines creates units that then pursue their own subgoals. This can lead to a commonly observed form of process conflict where it is apparently rational for the units to compete for more of the organization's resources even at the expense of its overall performance (Argote and Greve, 2007). Thus, even at this aggregated level, conflict can be understood as a dyadic relationship involving:

> …a process which includes the perceptions, emotions, behaviors, and outcomes of the two parties. Specifically, in order to differentiate conflict processes from other processes…conflict is the process which begins when one party perceives the other has frustrated, or is about to frustrate, some concern of his
>
> Thomas, 1976, p. 891

Thomas' use of the third person singular pronoun, in generic terms, unwittingly reveals two major assumptions underpinning the psychological approach. First, that we can empirically identify an entity that, much like a self-determining individual, possesses a coherent set of singular, stable, and universal interests, intentions, or affective states. Second, we can also empirically and unambiguously determine these specific psychological characteristics. The effect of these assumptions is to anthropomorphize *all* conflict relations to the extent that, at any level of aggregation, they will tend to follow certain psychological principles premised on the simplest person-to-person dyad. Beyond this obvious conceptual reductionism there are also strong implicitly functionalist assumptions; notably that the normal state for any collective entity is a healthy coherence that will be encouraged by good husbandry of those conflicts that do occur. Such assumptions give rise to the view that, so long as conflict is managed effectively using rational principles, we can keep the lid on organizational conflict and even use it to the organization's advantage.

Thomas (1976) offers three reasons why conflict can be functional for organizations. First, there may be a lack of tension in an organization that can create boredom and complacency while, in contrast, a certain level of conflict may stimulate creativity. Second, the existence of conflicting priorities and, hence, ideas may make for better solutions to organizational problems. Third, aggressive behavior need not necessarily be destructive, especially when it initiates conflict-resolution procedures. Thomas also notes that, as beneficial side effects of conflict, organizations can identify systemic equity problems, address internal power imbalances, and, in situations where the whole organization is united in conflict with another, improve its internal cohesiveness.

Recent theoretical developments have qualified Thomas' proposition, adding further nuances, and today, many conflict researchers work with a conflict topology framework that distinguishes between task conflict and relationship conflict. Much research concerns how task conflict is more likely to be constructive due to its potential to trigger information processing and to lead participants to consider various problem solutions and multiple perspectives. Task conflict may therefore enhance decision-making quality, individual creativity, and work-team effectiveness in general (De Dreu and Beersma, 2005; De Dreu and Van de Vliert, 1997; Jehn, 1994, 1995, 1997; Jehn and Mannix, 2001; Tjosvold, 1991, 2008; Van de Vliert et al., 1999). Conversely, relationship conflict is thought to interfere with performing tasks, thereby lowering effectiveness and innovativeness (De Dreu and Weingart, 2003; Jehn, 1995, 1997; Murnighan and Conlon, 1991) and can have negative long-term consequences for individual health and well-being (De Dreu et al., 2004). Consequently, much conflict research addresses how to reduce those types of conflict that are bad for the organization and stimulate other types of conflict, which are deemed constructive for increasing performance, creativity, and innovation.

The social psychological approach has spawned a whole industry of expert theory and practice that seeks better methods of containing conflict within the limits of acceptability. Generally speaking, however, two main methods of conflict containment are open to us: the resolution of conflict through negotiation or the regulation and minimization of conflict through governance structures. Psychological studies have emphasized the former option to such an extent that the conceptual terrain of organizational conflict is now dominated by ideas of how to negotiate our way out of it! An indication of the importance of this shift of emphasis can be obtained by comparing the entry on 'Conflict' in the first edition of the *Handbook of Industrial and Organizational Psychology* (Thomas, 1976) and the revised entry in the second edition (Thomas, 1994). In this latter contribution the validity of the *Process* and *Structural* models is largely taken 'as read' and most of the space is devoted to outlining subsequent developments in negotiation theory and practice, especially the role played by third-party technocrats in bringing about conflict resolution. Social psychology perspectives typically provide little critical reflection on the adequacy of their models or on the validity of their behavioral assumptions. These shortcomings are particularly troubling: representing *all* organizational conflict as if it took place through the operation of a dyad fails adequately to capture the complexities of the social relations found in organizations. In the light of this, we shall now turn our attention to alternative treatments of conflict.

Ethnographic Approaches to Organizational Conflict

According to Barley (1991), one common misapprehension about the nature of organizational conflict is indicated by our tendency to 'bracket' it. By this, Barley means that we see conflict as somehow being outside the usual day-to-day events of organizational life. It is to be seen as a special case to be treated in special ways. Barley considers this bracketing to be, indirectly, an artifact of the psychological discipline of

conflict and negotiation studies: seeing the world through eyes used to studying dyadic negotiations in laboratory settings predisposes us to look for self-contained conflicts with well-defined antecedents and a limited life span. Barley argues that we need to bring the study of conflict back from the margins and place it at the heart of the quotidian experience of organizations. This means rejecting ideas of utility maximization and individual cognition in favor of anthropology of specific conflict processes; an approach that involves developing an 'ethnography of disputing.' Particularly noteworthy is the use of the term 'dispute' instead of 'conflict.' The change in language stems from a desire to be dissociated from the existing descriptive and normative topologies within the social psychological approaches to organizational conflict. By introducing a perspective on organizational conflict as disputatious, the ethnographic approaches focus on dispute(s) as the unit of analysis to capture the complex interaction of players, issues, context, and dispute processes (Kolb and Putnam, 1992). Accordingly, a disputatious perspective makes it possible to uncover public and private domains of conflict unfolding and conflict management in different settings.

Barley regards the advantages of this approach to be the historicizing and contextualizing of the eternal conflict between organizational actors trying to make sense of their position in a negotiated order. Barley's position relies heavily on the work of Anselm Strauss and his earlier studies of the unstable divisions of labor found in many organizations. Strauss and his collaborators found that, rather than following externally imposed standardized rules and procedures, individual tasks and duties were more the product of bargaining and negotiation between the staff themselves (*see* Organizations: Negotiated Orders).

In general, an anthropological approach to organizational conflict places a good deal of emphasis on the social/cultural context in which the disputants find themselves. These include socially constructed identities (especially 'insider' vs 'outsider' status), the size and complexity of networks of dependency relations, the historical nature of social norms (especially those that define what a 'dispute' is and the rules by which it is handled), and also the resources that actors who are seeking to influence the trajectory of a dispute have at their disposal (Barley, 1991). Overarching these dimensions is an appreciation of the semantic and material implications of the language of conflict; in this way, the anthropological approach shares a number of methodological similarities with Actor Network Theory (Law, 1994) and Domain Theory (Gray, 1989).

The anthropological approach, as depicted by Barley, serves the important purpose of showing us that we do not have to think of conflict exclusively in psychological terms. Breaking the tyranny of the anthropomorphized dyad allows us to think about conflict as a complex nexus of social relations and avoids the functionalist view that it must be doing some 'good' somewhere. Nevertheless, the anthropological study of conflict still remains undertheorized. Indeed, it remains largely an empirical exercise and little theory is at work, an observation that is particularly striking when considering the dimension of power in shaping the antagonistic relations of everyday social settings. For example, Barley's (1991) conception of power extends little beyond the distribution of 'resources' available to parties to influence outcomes in competitive relations, a view that is also evident in some explicitly 'sociological' treatments of conflict (e.g., Tolbert, 1991). Thus, it is little different from social psychological approaches in dealing with the complexities of the power relations mediating organizational conflict. Recently, several authors have addressed this shortcoming in anthropological studies of conflict: for example, Sarat and Felstiner's (1995) application of Michel Foucault to their studies of divorce negotiations. Nevertheless, the general lack of nuance in the treatment of power by theorists concerned with organizational conflict points to an area where studies can be greatly enriched.

The conceptual arena of power underwent profound changes in the late twentieth century, shifting from arid behaviorism to arcane abstractedness, before finally settling into fine-grained analysis (see Flyvbjerg, 1998; also *see* Organizations: Authority and Power; Organizational Control). In part, this shift was symptomatic of a return to a less moralistic and more Machiavellian view of power. We shall return to these themes in the last section of this article.

Organizational Conflict and Neo-Darwinism

The basic premise of neo-Darwinian approaches to human behavior, such as the Evolutionary Psychology movement (e.g., Buss, 2005: passim), is that the human mind is stuck in the prehistoric African Savannah of our hominid ancestors. Behavioral traits – for example, human aggression and a predisposition toward intersubjective conflict – survived because they conferred adaptive advantage on us as a species. Thus, we can establish *ex post facto* explanations for almost any empirically observable form of behavior – on the basis of natural selection then, if it was not adaptive, it would have been bred out. Nicholson (1997) suggests that common characteristics of organizations, such as a predilection for gossip, the love of ritual, and even the existence of hierarchy, can be traced to biologically determined human nature.

The increasing influence of the neo-Darwinian approach (incidentally, a viewpoint that is opposed by many evolutionary biologists who see it as a misappropriation of Darwin's ideas) is already being felt in studies of organizational conflict. For example, it has long been considered that a tension between conflict and cooperation is part of our biological destiny. For example, it has been claimed that, by studying the mutual dependence systems of our hunter–gatherer ancestors, we are able to define a set of desirable social principles for current organizations that are "… more congruent with human nature" (Ehin, 1995, p. 667). Adopting a dualistic model of innate characteristics, neo-Darwinists argue that humans display either *agonic* (aggressive, individually territorial, and hierarchical) or *hedonic* (mutually supportive, group-centered, and egalitarian) tendencies. The extent to which we display one or the other is determined by context and, if we can return to our 'natural' settings, then hedonic predisposition ought to prevail. It is only when we are removed from this setting – in this case, placed in a hierarchy of work organization – that we tend to display agonic characteristics. This position stands as the intellectual equivalent of Hobbes' primitive ruminations and the organizational implications are clear. First, conflict is an evolutionarily advantageous response to many 'unnatural'

forms of organization and, second, if only we can return to a more natural form of organization, then conflict will be, to all intents and purposes, banished as we revert to our hedonic state. This reflects an increasingly popular view that presence of cooperation and the absence of conflict are not just virtuous aspirations within the modern society; they are manifestations of evolutionary success of our species (Pinker, 2011). Not only is this view naive with respect to issues of power, but also a contestable interpretation of Darwin's idea of natural selection. In particular, two specific objections can be laid against it.

The first is fundamental; according to Darwinian adaptation, evolution occurs through random mutation rather than strategic intent and anything that involves intent cannot be explained by Darwinian ideas: in short, evolution is not teleological. Of course, axiomatically this rules out all social situations, including those where parties are in conflict over the relative moral status of competing norms or the distribution of resources. But, as if this misinterpretation of Darwin was not sufficient, such studies can never satisfy the requirements of evolutionary science. This point was made many years ago by de Jong and van der Steen (1998) and has still not been successfully countered. They demonstrate the inability of any psychological study to establish that anything so abstract and socially constructed as a behavioral trait (for example, a predisposition toward conflict) could be advantageous from an evolutionary perspective. Few would deny that heredity plays *no* part in behavior, but it is stretching it beyond the conceptual and methodological bounds of Darwinism to suggest that the dynamics of organizational conflict will be *determined* by human genetics.

A Neo-Machiavellian View of Organizational Conflict

By way of a departure from the previous approaches to organizational conflict, we wish to offer a fourth framework for thinking about the problem. Our starting point is that we must acknowledge that conflicts matter in any organization and we are all players – even in the face of denial or ignorance (being devious or stupid does not disengage one from conflictual processes). There are four basic ways in which the political maneuvring of organizational conflict proceeds. The first three of these invariably lead to protagonists being ousted in conflict. The organizational agent who is not politically skilled will fail. Returning to the Machiavellian theme, the key implication is that skill in the arts of political engagement, strategy, and discourse should be central to organizational competencies.

The first typical mode is to be *in denial* of conflict. We can characterize this as unrealistic, for few rational actors could be insensitive enough to their organizational surroundings to survive in this state for long. Second, there is *passive acceptance without resistance* of what one is in conflict with. This is naive: it accepts what it does not seek to understand or counter. Third, there is *active recognition without action*, where one realizes the nature of the conflict but proposes no strategic counterpoint for dealing with it. We may characterize this as incompetent. Finally, there is *strategic cognition and treatment*, where one recognizes the phenomenon as conflict and treats it appropriately through devising strategies to manage it, with and through the use of power. We may characterize this, after Buchanan and Badham (1999), as developing 'power-assisted steering.'

Essential to these competencies is the ability to be able to creatively improvise around organizational conflicts, through developing a diverse repertoire. This includes image building, selective information release and dissemination, scapegoating of conflictual others, building formal alliances with potential allies, networking with friends and enemies of enemies, striking up appropriate compromises where necessary, being able to manipulate rules, and being ruthlessly pragmatic. Such tactics require deployment in specific organizational contexts, characterized by diverse stakeholders and variable political entrepreneurship. The exercise of power, in ongoing action, will be accomplished through the production of plausible accounts, the creation of viable reputations for political operators, and the achievement of outcomes regarded as 'wins' (Buchanan and Badham, 1999). Essential to these capabilities is a capacity to analyze strategies of power discursively. Those who would be players need to be able to ask questions about the extent to which their power games – actual or projected – will be characterized as ethically acceptable, contextually reasonable, and plausibly accountable? What will the consequences be for those whose reputations are involved in the stakes: who will be weakened, strengthened, or unchanged?

Organizational life involves developing and deploying *political competence*. (The theoretical derivation of this view of organizations are elaborated in Clegg (1989) and Buchanan and Badham (1999).) Sometimes the skill will be in containing conflict or nipping it in the bud, other times ensuring where and when it flares, in what arenas, and with what participants. Conflicts represent opportunities for honing the skills of political competence in whatever normative order the organization constitutes. Clearly, these are highly variable – but even the most sacred may be as readily characterized in these terms as may be the most profane – as any reader of Trollope's '*Barchester*' novels could attest. Political strategy does not depend on the normative environment in which it is lodged so much as the strategic capabilities deployed. On this reading, Machiavellian skills are appropriate in any organizational context.

See also: Administration in Organizations; Conflict and Conflict Resolution, Social Psychology of; Conflict: Anthropological Aspects; Darwinism: Social; Organization: Overview; Organizational Behavior, Psychology of; Organizations: Authority and Power.

Bibliography

Argote, L., Greve, H.R., 2007. A behavioral theory of the firm – 40 years and counting: introduction and impact. Organization Science 18, 337–349.

Barley, S.R., 1991. Contextualizing conflict: notes on the anthropology of disputes and negotiations. Research on Negotiation in Organizations 3, 165–199.

Buchanan, D.A., Badham, R., 1999. Power, Politics and Organizational Change: Winning the Turf Game. Sage, London.

Buss, D.M. (Ed.), 2005. The Evolutionary Psychology Handbook. Wiley, Hoboken.

Clegg, S.R., 1989. Frameworks of Power. Sage, London.

Cyert, R., March, J.G., 1992. A Behavioral Theory of the Firm, second ed. Wiley-Blackwell, New York.

De Dreu, C.K.W., Beersma, B., 2005. Conflict in organizations: beyond effectiveness and performance. European Journal of Work and Organizational Psychology 14 (2), 105–117.

De Dreu, C.K.W., Van de Vliert, E., 1997. Using Conflict in Organizations. Sage Publications Ltd., London.

De Dreu, C.K.W., Weingart, L.R., 2003. Task versus relationship conflict, team performance, and team member satisfaction: a meta-analysis. The Journal of Applied Psychology 88 (4), 741–749.

De Dreu, C.K.W., van Dierendonck, D., Dijkstra, M.T.M., 2004. Conflict at work and individual well-being. International Journal of Conflict Management 15 (1), 6–26.

Ehin, C., 1995. The quest for empowering organizations: some lessons from our foraging past. Organization Science 6, 666–671.

Flyvbjerg, B., 1998. Rationality & Power: Democracy in Practice. University of Chicago Press, Chicago.

Fukuyama, F., 1999. The Great Disruption: Human Nature and the Reconstitution of the Social Order. Free Press, New York.

Gray, B., 1989. Collaborating: Finding Common Ground for Multiparty Problems. Jossey Bass, San Francisco.

Jehn, K.A., 1994. Enhancing effectiveness: an investigation of advantages and disadvantages of value-based intragroup conflict. International Journal of Conflict Management 5, 223–238.

Jehn, K.A., 1995. A multimethod examination of the benefits and detriments of intragroup conflict. Administrative Science Quarterly 40 (2), 256–282.

Jehn, K.A., 1997. A qualitative analysis of conflict types and dimensions in organizational groups. Administrative Science Quarterly 42 (3), 530–557.

Jehn, K.A., Mannix, E.A., 2001. The dynamic nature of conflict: a longitudinal study of intragroup conflict and group performance. Academy of Management Journal 44 (2), 238–251.

de Jong, H.L., van der Steen, W.J., 1998. Biological thinking in evolutionary psychology: rockbottom or quicksand. Philosophical Psychology 11, 183–205.

Kolb, D.M., Putnam, L.L., 1992. The multiple faces of conflict in organizations. Journal of Organizational Behavior 13 (3), 311–324.

Law, J., 1994. Organizing Modernity. Blackwell, Oxford, UK.

Murnighan, J.K., Conlon, D.E., 1991. The dynamics of intense work groups: a study of British string quartets. Administrative Science Quarterly 36, 165–186.

Nicholson, N., 1997. Evolutionary psychology: toward a new view of human nature and organizational society. Human Relations 50, 1053–1078.

Pinker, S., 2011. The Better Angels of Our Nature: The Decline of Violence in History and Its Causes. Penguin, London.

Sarat, A., Felstiner, W.L.F., 1995. Divorce Lawyers and Their Clients: Power and Meaning in the Legal Process. Oxford University Press, Oxford, UK.

Thomas, K.W., 1976. Conflict and conflict management. In: Dunnette, M. (Ed.), Handbook of Industrial and Organizational Psychology. Consulting Psychologists Press, Palo Alto, CA, pp. 889–936.

Thomas, K.W., 1994. Conflict and negotiation processes in organizations. In: Dunnette, M., Hough, L.M. (Eds.), Handbook of Industrial and Organizational Psychology, second ed., vol. 3. Consulting Psychologists Press, Palo Alto, CA, pp. 651–717.

Tjosvold, D., 1991. Rights and responsibilities of dissent: cooperative conflict. Employee Responsibilities and Rights Journal 4, 13–23.

Tjosvold, D., 2008. The conflict-positive organisation: it depends upon us. Journal of Organizational Behavior 29, 19–28.

Tolbert, P.S., 1991. Negotiations in organizations: a sociological perspective. Research on Negotiation in Organizations 3, 99–117.

Van de Vliert, E., Nauta, A., Giebels, E., Janssen, O., 1999. Constructive conflict at work. Journal of Organizational Behavior 20 (4), 475–491.

Conformity: Sociological Aspects

Pierre Demeulenaere, Université Paris-Sorbonne, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Conformity is a central feature of social life. There are three main approaches to conformity in the social sciences. One stresses the role of psychological motives that lead to similar behavior. The second insists on the existence of a social level that would exert pressure on individuals. And the third describes the role of reasons and justifications, in various situations, in order to interpret conformist attitudes. The article will explore the ways those various approaches interpret social conformity. It will do so however by distinguishing three main fields where social conformity arises: first factual beliefs; second, the moral, political, and religious rules and norms which strongly encourage or forbid certain attitudes; third, lifestyle or cultural attitudes, including food, language, aesthetics. The article's main point stresses that the very formal notion of conformity disguises various problems when the specific fields of action are taken into account.

The concept of conformity in sociology has a very large scope due to its sheer formality: it can be applied to any social behavior and belief when it is considered from the point of view of its frequency within a group. This article examines the typical situations, constraints, and motives that lead to similar or dissimilar behaviors, which are more or less strongly reinforced by different types of sanctions or rewards. There are three main approaches to conformity in the socials sciences. The first stresses the role of psychological motives that lead to similar behavior. The second insists on the existence of a social level that exerts pressure on individuals. And the third describes the role of reasons and justifications in order to interpret conformist attitudes. The reasons depend on beliefs and justifications on one side, and on circumstances on the other. These justifications and circumstances differ according to the variety of fields of action. Three main fields where social conformity arises can be distinguished: first the positive beliefs; second, the moral, political, and religious rules which strongly encourage or forbid certain attitudes; third, lifestyle or cultural attitudes, including food, language, aesthetics. The article's main point stresses that the very formal notion of conformity disguises very different problems when the specific fields of action are taken into account. For instance, conformity can be opposed to deviance and to innovation altogether. But deviance and innovation are certainly not equivalent.

Three Models of Conformity

Conformity depends directly on the degree of similarity between an individual action and another, those actions being more or less frequent in a group. But in addition to this frequency aspect, there is a sense of obligation linked to the idea of conformity: people feel more or less obliged to act in a certain way. There are therefore in any conformity phenomenon three dimensions: the object of conformity (a given attitude), the sense for a given individual that he or she should adopt this attitude, and a diffuse pressure among people to force each other to adopt this attitude. Whatever its origin, conformity is always strengthened by unequally strong positive or negative rewards and sanctions. From death penalty to intellectual dissatisfaction, from imprisonment to snobbery, conformity is most often reinforced by social sanctions in different ways. Sanctions are at the same time a cause and an effect of conformity, since the belief that social norms should be followed will lead to the setting of sanctions for deviants; and the sanctions tend to reinforce the respect of norms.

Yet it is also observable that conformity can occur when no sanction at all is observable, for instance in the case of people's breakfast habits, which largely differ from one society to another, this difference being related to a group attitude. However, even in this case, people tend to believe that this attitude is the 'best' one for them, inside their group. The link between the selected attitude, its choice by an individual, and the social pressure to conform to it is the existence of a norm that prescribes a certain type of behavior for the members of a group.

In this respect the size of the reference group in which the norm should prevail can have a variable extension (Mead, 1962): for a Catholic, or a believer in human rights, humanity as such should comply with some specific norms, whereas the follower of an elitist fashion will stop liking it when the number of followers has reached a certain size (Coleman, 1990). However, people can belong to several groups at the same time, each of them being submitted to norms, that can be consistent or opposed: for instance gay people can be catholic believers at the same time, the catholic church being globally opposed to homosexuality. This is not a pattern of modern societies alone: in traditional societies, people can belong simultaneously to various groups, depending on their religion, work, family, or language (Gellner, 1983).

There are basically three models in the social sciences literature to interpret such conformity. The first one is mainly psychological. It stresses unconscious trends toward social conformity. The second one, rooted in Durkheim's theorization of society, insists on the role of a specific social level that exerts coercion on individuals. Finally, the focus on rationality will specify the reasons for which people adhere to norms and therefore display conformity in adopting them. The rational approach has itself two sides: one is utilitarian and is based on cost benefit analysis. The other one stresses, more broadly, the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32027-X

reasons one has to adhere, in a given situation, to specific beliefs or attitudes. Since those reasons, in given circumstances, are shared by individuals situated in similar social situation, this raises conformity.

The first model can be historically related to Tarde (1962) who stresses the importance of imitation to explain similar behaviors in social life. Imitation appears to him or her as a constant attitude, which explains all social phenomena: all fields of action are involved, and innovation, or dissent, depends always on previous imitations. In contemporary psychology or anthropology such a trend (mainly unconscious) toward imitation is frequently described as a key element toward the diffusion of common patterns of behavior (Sperber and Hirschfeld, 2004). It leads sometimes to a decidedly irrational explanation of the adoption of group attitudes (Haidt, 2012) based on evolutionary selection of behavioral patterns. People do not choose conformity on the basis of an individual choice, but because there are unconscious selected motives that lead them to do so.

Durkheim (1982) does not accept this rather general idea of imitation based on psychological motives. There is indeed for him or her a similar strong social tendency toward the building of common patterns of behavior: moral, religious, linguistic, or esthetic rules express in a common way the fact that society is built up on the necessity of integration of individuals throughout shared values. Despite any subsequent difference of attitudes, conformity is an essential feature of any society. It stems from a sense of oughtness shared by the members of a group: they all feel that they should behave in a certain manner. But whereas Tarde originates this sense of conformity in a psychological trend, Durkheim locates it in a specific social level. This social level can be interpreted in two different dimensions. First, it would be a specific social reality that exerts pressure on individuals, but this leads to ontological problems, since it is not obvious that such a social entity exists as such. More modestly, the social level will refer to the fact that a norm is always social, and its respect is always the result of an interactive reinforcement: it is not a lonely attitude.

With the notion of rationality, a third approach is introduced, which is not strictly psychological, nor 'social' in the durkheimian sense. The notion of rationality is a key element to understand conformity in a social world. It stresses either the interests someone, in a situation, has to adopt a norm; or, more generally speaking, and beyond narrowly selfish interests, the reasons someone has to adopt a norm. For sure, it can be said that compliance with norms is not necessarily directly conscious: it can even be difficult for people who follow a norm to give an accurate reason for why they do so. This is true of any behavior, and already been noted by Weber. This is also a key element of the psychological literature about norms compliance (Haidt, 2012), which stresses in a more radical way that justifications are only attempts by actors, ex post, to find reasons for behaviors that are not at all in fact determined by reasons, but by instinct and emotion. However, norms acceptance is often linked to conscious reasons that justify it. Conformity is related to beliefs and to justifications: people are conformists or deviants because they believe they should be so, and others are asked to behave in the same manner when they are thought to belong to the relevant group to which the norm that introduces conformity should apply. Therefore, a study of the conformity problem must include a study of the typical motives for which a justification is commonly accepted or rejected.

People most often obey a norm because they believe the norm is valid. They can also obey the norm because they are forced to comply with it, without accepting its validity. It can happen also that the members of an interaction follow a norm because everyone believes that the others want him or her to respect the norm, although everyone would also prefer not to do so. The result is paradoxical: everyone follows a norm (for example, heavy drinking habits among college students) whereas everyone would prefer not to do so (Willer et al., 2009).

Symmetrically, actors can refuse to obey a norm not because they do not acknowledge the validity of the norm, but because they have occasionally a stronger interest against its following. On the other hand, people can refuse to obey a norm because they believe it is unfair, or useless, or unnecessary, or inadequate: people may think that a norm is illegitimate from the point of view of their position in society. For instance, those who cannot through their personal effort achieve a social promotion will tend to believe that, since they do not have such opportunity, they cannot have any faith in a social order that builds its legitimacy on this very self-promotion through effort (Merton, 1957). Those conditions can be related to the position of an individual within a group, or to the members of a group within a larger group.

In sum, it can be said that in order to interpret social conformity, there are mainly three dimensions that should be taken into account: the psychological motives that lead to conformity; the social pressure that exerts conformity; and the reasons people have collectively to display conformity. Sometimes these three approaches can be contrasted. However, they can be seen as well as complementary, since any conformity phenomenon involves at the same time psychological motives (the fear for instance of losing one's reputation), social pressure, exerted through interaction, and reasons, in given circumstances, to adopt a belief or a norm. This interconnection can be seen in the various fields where social conformity arises.

Positive Beliefs

A first domain for conformist attitudes is that of positive beliefs. The very idea of 'positive' beliefs, that is free of any normative content, is a creation of philosophical and sociological tradition, when it stresses the difference (and sometimes the opposition) between facts and values. Moreover, sociology has described a historical evolution tending to separate positive beliefs from normative ones: it is one of the dimensions of the long run social process conceptualized by Weber as rationalization. The distinction between fact and value can certainly be contested (Putnam, 2002). Moreover people in current social life most often do not make such a distinction, and would maybe reject it. However, it is clear that people tend to have positive beliefs regarding specific facts, and that those beliefs do not correspond to an approbation or

disapprobation of these facts. In this field there are strong claims to conformity based on epistemic norms that are supposed to have a universal validity.

These norms are related to the norms of rationality that should impose on rational people to hold accurate beliefs given the information they have access to. Conformity derives in this case from common cognitive features that enable people to have correct statements. Rationality can be attacked in this respect in a radical constructivist or culturalist way that would deny the existence of intrinsically valid universal norms. However, it seems hard to avoid a notion of correctness based on rationality to interpret at the same time common social attitudes and beliefs, and science itself (and sociology in particular).

However, in this positive domain, there are frequent errors. The very notion of error presupposes correctness, and therefore a notion of rationality. Psychology will tend to stress the importance of three kinds of factors in explaining such typical errors. First, people can make errors because they do not have access to the correct information, or because they try to solve difficult problems with limited cognitive effort or abilities: they reason too quickly (Kahneman, 2011). Second, people can have false beliefs because they have some motivation, interest, or desire to hold them (Gilovich, 1991). These two dimensions are not social as such, although they can be widely diffused in social life. But there is no direct role of social interaction in these cases. However social psychologists have also experimentally studied the conditions under which conformity arises from specific social settings: it is proved that it is difficult to go against commonly held opinions, even when the actor has the means to recognize that they are erroneous (Asch, 1956; Sherif, 1972).

Generally speaking, people can hold false beliefs because their social environment supports such beliefs, and therefore others influence people in their opinions. Conformity arises in this case from the fact that one's opinion is not forged in loneliness, but depends on others' opinions on the subject. Tocqueville (1966), for instance, explains that people need general ideas to achieve their ordinary actions, but they do not have the means to verify the validity of all the beliefs they will refer to. In a society based on the dismissal of all authority, the majority of people will think that the most acceptable belief is the one that is commonly shared by others. There will be a tendency toward conformity since the adoption of majority's voice is less costly, both from an intellectual point of view and from a social point of view. The role of ideology can thus be very strong, based on the mixture of false beliefs and social reinforcement of those beliefs: since dissent is costly in a situation where a belief is widely shared, it will be minored. Dissent is costly in two different ways: first if others want you to believe something, and will oppose you if you do not share this belief, someone can be criticized or rejected and suffer this isolation. Secondly he or she will have to justify his or her beliefs: it can be difficult for someone to defend his or her opinions against all others. This is why generally dissent does not occur between one individual and a group, but between groups, inside of each of them everyone feels he or she is supported by the belief that others share the same attitude.

Social Rules in Moral, Political, and Religious Fields

As previously mentioned, conformity often implies a problem of legitimacy, although people are not necessarily capable of giving the reasons for which they follow a norm. There are strong social norms regarding social practices in moral, political, and religious fields. The aim of those norms is precisely to impose conformity: they prescribe and proscribe various behaviors and attempt at homogenizing them. The content of these norms tend to vary in a significant way from one society to another, and there is also a historical change that must be taken into account. Sociologists often consider that this variety must lead to a complete relativity of values. That would be a culturalist point of view, derived from anthropology. Conformity is inherent to social systems, but the heterogeneity of norms is a fact that cannot be reduced to any deeper unity. This view can be supported in a positivist tradition by the fact–value dichotomy: it would express that values cannot be vindicated, and are therefore arbitrary (Pareto). There is a war of gods (Weber).

However, it seems that there are four relatively constant norms throughout all cultures (Cusson, 1983): incest between mother and son, father and daughter, or brother and sister; rape of married women; murder, meaning the voluntary act of killing someone who belongs to one's own group; theft. Of course, the last case is the less accurate since the definition of legitimate belongings can vary in such a way that property itself can be seen as a theft. But these common features are a very small part of all the social norms linked to conformity.

Independently of any kind of intrinsically justification of norms, and beyond their constancy or variety, conformity can be then interpreted in terms of power and interests. The common adoption of norms can derive from a common shared interest in adopting them, which leads to conformity; alternatively some people can impose norms to others in a power relationship, where the interests or the ones are realized at the expense of the interests of the others. More rarely, norms exist that do not correspond to the interests of anyone: many taboos are of this kind, where the norms prohibit some kind of actions, without any kind of advantage produced to anyone (Elster, 1989). It can happen also that, although everyone has an interest in changing a norm, the coordination costs to achieve this change are too high, and therefore the inefficient norms remain.

Conformity can derive from a situation where people have a common interest in adopting a norm. This is the case for instance when people must decide whether they should drive on the right or the left. Once a norm has been selected, everyone complies with it because no one finds an interest in deviation from the norm. A more complicated choice occurs in a prisoner's dilemma situation. In this case, everyone would like others to follow a norm, but has an interest in cheating regarding this common interest (Coleman, 1990). In this case, either the norm emerges, and conformity arises from a common observance of its requirements, or it does not emerge, and people loose the benefits of cooperation. Conformity can derive from the respect of the norm, or from its nonrespect (everyone decides he or she should not respect the norm because he or she believes that others will equally

not respect it). Processes of effective emergence of norms are complex ones, and imply various circumstances and actors.

But interests can lead equally to relations of power and domination. Economists and sociologists have explored these situations. For instance, in institutional economics, so-called 'natural orders' are described as places where elites impose specific rules that favor them to the detriment of others (North et al., 2009). The norms resolve at the same time the problem of violence: but they do so by favoring some individuals at the expense of others in stable hierarchies. This does not necessarily lead to a sense of conformity, since people tend often to challenge these norms, and the fact that they are imposed at one time does not make them legitimate in a permanent way. Therefore the evolution of norms depends on the evolution of the equilibrium of powers inside a society.

Sociologists have equally insisted on the existence of norms favoring some at the expense of others. This can lead sometimes to a sense of conformity. Many labeling theorists (Becker, 1963) interpret conformity as the result of a power relationship: the powerful impose on others conventions that are arbitrary. Arbitrary conventions have two effects. Firstly, some people are labeled as deviants whereas others are praised as conformists. This expresses the structure of power in society, not an intrinsic validity. Secondly, once people have been labeled as deviant, this label reinforces by stigmatization their deviant situation, since others can no more accept them as nondeviants. This narrows the range of actions that are available to them; eventually the stigmatized can themselves believe that they are deviants. In this case, we have a psychological mechanism, based on social power asymmetry, where people might accept an arbitrary norm that is exerted to their detriment.

When there is an asymmetry of power, some people tend to impose on others a norm that depends only on a power relationship. But at the same time, people tend to impose norms that they often believe to be legitimate. For instance, it is clear that racist norms appear in slavery situations. When people have an interest in having slaves, they also tend to believe that they have good reasons to be a racist, and will consider the inferior traits of the oppressed as 'natural' features that deserve subordination, instead of realizing that these traits are precisely an outcome of the subordination (Douglas, 1986). The victims themselves might then believe sometimes they are intrinsically inferior.

But this scenario is not always the prevalent one. The stigmatized can claim and vindicate a change of norm that is the replacement of a nonvalid norm for them by a valid one. Moreover, independent observers will tend to have a more independent point of view. For instance, critics of racism will support the idea that there is no legitimate ground for racism, and they will try to set up a norm against racism (Appiah, 2010): when they succeed, racists become deviants. If labeling theorists maintain the idea that there is never any intrinsic value to certain rules, they will have no basis for the replacement of a racist rule by a nonracist rule, and they will treat in the same manner the deviant who does not accept a prevailing racist norm and the deviant who does not accept a prevailing nonracist norm.

Therefore, historical change must be taken into account as it can be in part related to the relative strength of different arguments regarding norms. If we consider, for instance, norms regulating sexual behavior, Christian religious belief tends to condemn, on a metaphysical basis, premarital sex, divorce, homosexuality, and pedophilia altogether. This is not directly linked to the interest of a category of the population exerted at the expense of the others. It is a moral representation based on religious beliefs (and maybe independent sexual taboos). But in a society where people are submitted to such norms, when they do not share the religious beliefs, they might consider themselves as victims of arbitrary norms. In such an instance they might believe that they have a right to do whatever they wish to do, in spite of those forbidding norms. It can be noticed, therefore, that when there is no more consensus on religious norms for sexual behavior, the only path usually found to maintain a legitimate norm that will be more or less accepted by everyone will be to refer to the principle of personal choice. Religious people will act according to their choice, and the nonreligious to their other choice. Modern societies thus increase the principle of personal choice, and diminish the authority of traditional or religious norms. The change is significant: premarital sex becomes the norm, divorce tends to be widely practiced, and homosexuality is increasingly legalized. But one significant aspect of this evolution is that norms against pedophilia tend on the opposite to be strengthened in such a society. This is easy to explain since, on one side, divorce or homosexuality are related to a freedom of choice among consenting partners. If a marriage is a contract between two adult persons, the notion of agreement can imply that the contract should be dismissed if the two partners do not agree any more, even if this is considered to be a failure from another point of view. On the contrary, pedophilia is definitely rejected because it involves children, legitimately seen as victims because of their lack of choice in such circumstances of abuse.

Therefore it cannot be said that a labeling theory is the last word on the problem of conformity or deviance. Both depend on beliefs and possibilities of legitimacy. In other words, norms can be said to depend on reasons. In this case, if strong norms depend on religious beliefs, when they are imposed on others who are nonbelievers, there is a problem of legitimacy if those beliefs cannot be proved to be right or wrong. Therefore, rationalization will tend to lead to a situation where such norms will apply only to believers. In this case, the principle of choice tends to legalize some practices, and to reinforce the fact that others are forbidden.

According to Boudon for instance (Boudon, 2001), we can try to find the reasons for the acceptance or the rejection of a norm, given the social situation and information of actors. Dissatisfaction can occur when there is a rejection of the elements on which the norm is based, or when the norm can be seen as valid in general, but not in the situation where the actors find themselves. Once again, for instance, according to Merton's description, someone will tend to reject a social order based on the idea of self-promotion through effort when it cannot be really achieved. Here a problem of consistency is faced: the norm is valid under certain conditions, but ceases to be so when those conditions are not met.

Hence, conformity should not be seen as static. Many norms depend on beliefs that are submitted to change and rationalization. One of the aspects of this rationalization is that

some norms cannot be valid to certain people: for instance, if there is a norm supporting slavery, it will tend to be withdrawn in the long run since it is inconsistent with the source of any legitimacy for individuals, their ability to accept a norm, which implies a certain amount of freedom of choice. Some norms are seen as oppressive which were not before, because the belief on which they were based is no longer shared. From a rationalistic point of view, there can be a discussion that shows that conformity or deviance depend partly on the perception of the validity of norms. It can be said that norms tend to be rejected in four major cases: when they just express a power relationship that is disguised in some other value; when they are grounded on false beliefs; when they are grounded on beliefs that cannot be proved and which have no self-evidence – in this case, they cannot be imposed on someone who does not share such beliefs; finally, a norm is not valid when it cannot be applied to someone's situation.

Norms of Culture

In the two previous sections, I have tried to isolate norms that can be discussed among people from the standpoint of their validity. But there are other norms that are connected to strong conformity that do not seem to be negotiated in terms of validity. For instance, Japanese women tend to walk in a way significantly different from European women. On one hand, there is no obvious reason to consider one way as better than the other; on the other hand, such behaviors express conformity, since they are group behaviors. How should we give an account of such attitudes applied to food, language, dress, architectural habits, etc.?

Three major elements can be stressed regarding this type of conformity. First, it can be related to practical constraints. A language must be shared among people to be efficient for communication. Here, conformity is just the result of the nature of the practice. There cannot be a private language (Wittgenstein, 1967).

Second, similarity can express integration to a group. For instance, it is obvious that religious dress is linked to the sense of belonging to a community, which can be turned into strategic devices. People tend to adopt some features to show that they belong to certain groups (Simmel, 1923). This can be done because belonging to a group is seen as important, or because belonging to a group is a consequence of a common faith or a common goal, or because there is some sort of distinction that is looked for (Veblen, 1953).

But finally, it must be said that in these fields conformity tends to be linked to innovation. It is obvious in the esthetic field: on the one hand, there is a huge amount of conformity that is observable in the constituency of local schools, such as the general use of Romanesque architecture during certain periods in Europe. This conformity, however, does not exclude innovation; Gothic style sooner or later replaces the Romanesque. It would be too difficult in this article to assess all the elements that play a role in such an evolution. It is clear at least that the previous ones are important: interactive constraints, and the sense of belonging to a community. But these allow the evolution of behaviors and the place for innovation.

See also: Collective Behavior, Sociology of; Collective Beliefs, Sociological Explanation of; Compliance and Legal Authority; Control, Social; Empirical Legal Studies; Law: Anthropological Aspects; Legitimacy, Sociology of; Norms; Social Influence, Sociology of; Status and Role, Social Psychology of; Status and Role: Structural Aspects; Values, Sociology of.

Bibliography

Appiah, Kwame Anthony, 2010. The Honor Code. How Moral Revolutions Happen. W.W. Norton & Company, New York.
Asch, S.E., 1956. Studies of independence and conformity: a minority of one against a unanimous majority. Psychological Monographs 70 (9), 1–70.
Becker, H., 1963. Outsiders. The Free Press, New York.
Boudon, R., 2001. The Origin of Values. Transaction, New Brunswick (USA) and London (1999).
Coleman, J.S., 1990. Foundations of Social Theory. Harvard University Press, Cambridge, MA.
Cusson, M., 1983. Le contrôle social du crime. Presses Universitaires de France, Paris.
Douglas, M., 1986. How Institutions Think. Syracuse University Press, Syracuse, New York.
Durkheim, E., 1982(1893). The Rules of Sociological Method. Macmillan, London.
Elster, J., 1989. The Cement of Society. A Study of Social Order. Cambridge University Press, Cambridge.
Gellner, E., 1983. Nations and Nationalism. Basil Blackwell, Oxford.
Gilovich, T., 1991. How We Know What Isn't So. Free Press, New York.
Haidt, J., 2012. The Righteous Mind. Why Good People Are Divided by Politics and Religion. Vintage books, New York.
Kahneman, D., 2011. Thinking Fast and Slow. Farar, Straus and Giroux, New York.
Mead, G.H., 1962(1934). Mind, Self and Society. University of Chicago Press, Chicago.
Merton, R.K., 1957. Social Theory and Social Structure, revised ed. Free Press, Glencoe, IL.
North, D.C., Wallis, J.J., Weingast, B.R., 2009. Violence and Social Orders. Cambridge University Press, Cambridge.
Putnam, H., 2002. The Collapse of the Fact/Value Dichotomy and Other Essays. Harvard University Press, Cambridge, MA.
Sherif, M., 1972. Experiments on norm formation. In: Hollander, E.P., Hunt, R.G. (Eds.), Classic Contributions to Social Psychology. Oxford University Press, New York.
Simmel, G., 1923. Die mode. In: Philosophische Kultur. Gesammelte Essais, third ed. Kiepenheuer, Potsdam.
Sperber, D., Hirschfeld, L.A., 2004. The cultural foundations of cultural stability and diversity. Trends in Cognitive Science 8, 40–46.
Tarde, G., 1962(1903). The Laws of Imitation. Peter Smith, Gloucester, UK.
Tocqueville, A., 1966(1835–40). Democracy in America. Harper, New York.
Veblen, T., 1953(1899). The Theory of the Leisure Class: An Economic Study of Institutions, revised ed. New American Library, New York.
Willer, R., Kuwabara, K., Macy, M.W., 2009. The false enforcement of unpopular norms. American Journal of Sociology 115 (2), 451–490.
Wittgenstein, L., 1967(1953). Philosophical Investigations. Blackwell, Oxford, UK.

Congress: United States

John B Gilmour, College of William and Mary, Williamsburg, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The United States Congress is distinctive among national legislatures in the world in that it is extremely powerful in legislating and policymaking. It is characterized by a system of committees that structure much of the legislative process. Members of Congress enjoy substantial incumbency advantage, and enjoy significant independence from national party leaders. Both the House and the Senate are characterized by complex internal procedures that have an important impact on legislative outcomes. An important recent development in Congress is an increased polarization between the major political parties.

The United States Congress is distinctive among the world's legislatures in its independence from the executive, its complex internal organization and procedure, and a capacity to develop its own legislative proposals. Senators are elected at large from their states, while representatives are elected from single member districts. The use of primary elections to make party nominations gives members of Congress a substantial measure of freedom from the leadership of their party, but makes them beholden to their district. Because of a substantial 'incumbency advantage' – due to advantages in fund-raising, and other factors – members of the House and Senate are reelected at very high rates. Internal structure of both the House and Senate is dominated by two institutions: parties and committees. Committees have great influence in shaping legislation and in determining what legislation will reach the floor. The majority party has much responsibility for organizing the agenda of Congress. The rules of the House of Representatives are designed to allow the majority to prevail, so they place strict limits on debate and amendment. The smaller Senate allows far more freedom for debate and amendment. The president has influence over members of Congress, especially those from his own party, but even members of the president's party are strongly influenced by their constituencies.

For much of the twentieth century the political parties in Congress were not sharply distinguishable, exhibiting considerable overlap – often to the regret of partisans. In recent decades, the Democratic and Republican parties in the House and Senate have become much more polarized. By some measures Congress since 2000 is more polarized than ever before.

Among the world's national legislatures, the United States Congress stands out as far and away the most powerful. It is distinctive in its independence from the executive, its complex internal organization and procedure, and a capacity to develop its own legislative proposals.

Congress is composed of two chambers: a Senate, in which each state has two senators; and a House of Representatives, in which the number of representatives from each state is determined by population. There are 100 senators and 435 representatives. To become law, bills must be passed by both chambers and be signed by the president. Bills vetoed by the president can become law, provided two-thirds of both the House and Senate vote to override, but that is rare. Senators are elected for 6-year terms, with a third of senate seats being filled every 2 years. Representatives are elected to 2-year terms, and all come up for reelection simultaneously.

Congressional Elections

Senators are elected at large from their states, while representatives are elected from single member districts (*see* by contrast, Proportional Representation. Also *see* Elections: Apportionment). The 2-year terms of Representatives are the shortest term of any national legislature. The authors of the Constitution intended the House to be the 'people's' chamber, and meant to keep them on a short leash. The 2-year term does this most effectively. Members of the House are for practical purposes always running for reelection, and they must constantly pay attention to constituency concerns. Indeed, it is widely accepted that the desire of members to be reelected has a vast influence on congressional structure and policymaking (Mayhew, 1974; King, 1997).

Gerrymandering

Senate 'districts' are the entire state, but House districts are each a subset of a state, devised through a political process (except for state that have only one House seat). States legislatures pass laws every 10 years drawing new district lines such that each district within the state has an equal population. The lines demarcating districts are often drawn to accomplish political purposes, a practice known as 'gerrymandering.' The purposes of gerrymandering include protecting incumbents, advantaging one party at the expense of the other, and promoting the election, or defeat, of racial and ethnic minorities. Gerrymandering commonly results in bizarrely shaped, buglike districts.

Primary Elections

In nearly all places in the United States, party nominations of candidates for Congress are done by means of primary elections, not by party committees or conventions. In other words, the national party organizations have little influence over who runs for Congress under their banner. This system provides members of Congress with a degree of freedom from their

national party leaders not enjoyed by members of the European parliaments. They must, however, satisfy the party activists in their state or district who tend to dominate primary elections.

One culprit allegedly responsible for increased polarization is a rise in primary challenges to incumbents. Several prominent and successful challenges to incumbents since 2000 have left a widespread impression that greater polarization of the electoral base has prompted more ideologically based challenges to incumbents, and this threat is driving members of the House and Senate away from the political center. The Tea Party movement in the Republican Party is seen as particularly responsible for this development. Boatright (2013) finds little evidence to support this thesis and contends that in fact primary challenges are not more common than in prior decades.

Incumbency Advantage

Members of Congress enjoy a substantial 'incumbency advantage' that enables them to be reelected at high rates. Typically, 90% or more of House incumbents who seek reelection win. Senate incumbents are somewhat less successful, but still manage to be reelected with great regularity. Even in elections featuring higher levels of turnover such as 1994 and 2010, nearly 90% of incumbents seeking reelection to the House are victorious. Even including voluntary departures from the House by retirement, normally 80% of the previous House is reelected. Incumbency advantage stems from the greater name recognition that incumbents enjoy compared to challengers, from their ability to do favors for constituents, and, perhaps most importantly, from their superior ability to raise money to finance their campaign. Congressional campaigns have become extremely expensive in recent decades, with the victorious House incumbents spending on average $1.5 million in 2012, and victorious Senate incumbents spending on average $11 million (*see* Political Money and Party Finance). According to Jacobson, there is a self-fulfilling quality to the high reelection rate of congressional incumbents. The seeming invulnerability of incumbents, and the daunting challenge of raising the money necessary to mount an effective challenge, discourage many would-be challengers, leaving incumbents to face incompetent, underfunded challengers who have little or no chance of winning, thus further inflating reelection rates (Jacobson, 2012).

Internal Organization

The United States Congress has a complex internal organization that has a great influence on how it translates the preferences of its members into legislation. Congressional organization is dominated by two partially competing institutions – committees and parties. Congress emerged more or less in its current, modern form at the end of the nineteenth century by means of a process of institutionalization (Polsby, 1968). Boundaries were established as once rapid turnover declined; internal complexity developed as standing committees became more firmly established; and universalistic decision-making criteria developed as seniority took hold.

Committees

Both the House and the Senate have a set of standing committees with fixed legislative jurisdictions. Normally, all legislation, once introduced by a representative or senator, is referred to a committee with jurisdiction over its subject matter. All members of Congress are assigned to committees. Senators have three or four committee assignments, while representatives have one or two. The party organizations in Congress make assignments to committees, but generally members are placed on committees where they have some interest. Some committees have more important jurisdictions than others and members naturally seek to be assigned to those committees. Seats on committees with jurisdiction over spending and taxation are among the most sought after, and the hardest to get.

Committees have great influence in shaping legislation and in determining what legislation will reach the floor – a bill is seldom considered by the entire House or Senate chamber unless it has been recommended favorably by the committee with jurisdiction. Committees do not have the power to force Congress to adopt legislation that its majority does not prefer, but they can block legislation a majority would favor, and thus they serve as an important veto point in the legislative process.

In both the House and the Senate, the chairperson of each committee is normally chosen according to a norm of seniority, whereby the member of the majority party with the longest continuous service on that committee is selected as its chairperson. The seniority norm is violated occasionally, especially when the majority party has just experienced a large influx of new members who have little commitment to seniority. This was the case in both 1975 and 1995, occasions on which the seniority norm was violated in three instances. Since 1995, there has been a declining commitment to seniority as a means of choosing committee chairs. Seniority has been partially replaced as the criterion for selection as a committee chair by fund-raising efforts on behalf of the party and ideological agreement with the party (Cann, 2008).

There is substantial debate among scholars about the consequences of delegating power to committees. One view holds that, because the members of a committee may not reflect the views of the entire chamber, they write legislation that advances their own interests, and not those of Congress as a whole. Another view holds that the chamber as a whole would not allow itself to be cheated in this way, and has effective tools to keep its committees in line and working for the benefit of the whole institution (*see* Delegation of Power: Agency Theory).

The committee system of Congress is commonly believed to strengthen Congress vis-à-vis the executive branch by encouraging its members to specialize. As issue specialists, members of Congress can better resist the executive branch, develop their own legislative proposals, and transform proposals submitted by the executive (Polsby, 1975).

Party Organization

Nearly all members of Congress belong to the Democratic or Republican parties. Party is an extremely important aspect of

congressional action, but political parties do not command the same degree of loyalty in Congress that they do in most other national legislatures. Party leaders in the House and Senate have a great deal of influence over the agenda, helping to determine what legislation will come to the floor, and when.

Recent party leaders have been motivated by varying conceptions of the leader's role. Newt Gingrich, Speaker of the House from 1995 to 1998, saw it as his responsibility to articulate a vision for the Republican party and he expected his party to follow him. This was an unusually expansive notion of congressional party leadership that his party was not willing to support indefinitely. In 1998, unhappiness with Gingrich's leadership resulted in his ouster. His successors, Dennis Hastert and John Boehner, both Republicans, and Nancy Pelosi, a Democrat, have returned to a more traditional leadership posture of seeking to identify a set of positions that unite his party and endeavoring to pass bills that accomplish those purposes. Parties and their leaders sometimes try to enforce party discipline by threatening wayward members with loss of committee assignments, chairmanships, or loss of other perquisites, but this is unusual, and sometimes drives dissidents into the waiting arms of the other party. The leaders' job might be better described as discovering rather than creating an agenda for the party. In 2013, Speaker John Boehner found himself the subject of an unsuccessful attempt to defeat him for reelection as speaker. The campaign against him was orchestrated by members of the Republican Party who found Boehner too accommodating to President Barack Obama.

The power of party leaders is limited by the willingness of party members to grant them power. With the members of Congress always eager to appeal to their own unique states and districts, they are reluctant to grant party leaders much power, for that might limit their own ability to pursue initiatives beneficial to their constituents.

To a substantial degree, the power of committees and party leaders is contradictory. To the extent that party leaders are able to increase their power, they do so at the expense of committees and committee chairs. When party leaders seek, as Gingrich did, to determine themselves what the agenda of the House will be, and what legislation will look like, they arrogate to themselves powers that committees and their chairpersons cherish.

Legislative Procedure

Because of their very different sizes, the House and Senate have fundamentally different procedures. In both, however, the rules of procedure have a profound impact on the legislative product. The House of Representatives, with its large size, cannot tolerate as much individual expression of its members as the Senate can, and its procedure is designed to ensure that the majority prevails. Consequently, the minority party and individual members often feel that they have been crushed by a legislative steamroller. The smaller Senate can more readily accommodate the peculiarities of its members and accordingly it allows far more freedom for debate and amendment.

Legislation in principle follows a regular path – a bill is introduced in the House or Senate, it is referred to a committee, voted on by the committee, and then sent to the full chamber for amendment and adoption. Then the bill is taken up the other chamber where it follows a similar course. The bills from the two chambers are seldom identical, so a conference committee is convened to iron out the differences. A compromise version of the bill is then adopted by both House and Senate and then sent to the president for his signature.

In practice, major legislation in today's Congress seldom follows this idealized path. Bills skip committee occasionally, they may be adopted as amendments to minor bills, bills may be packaged together as omnibuses to speed them through Congress, and so on. The variations on the standard legislative process are endless (Sinclair, 2011).

House Procedure

After important bills are reported favorably from committee, they normally go to the Rules Committee. The Rules Committee drafts a resolution, called a 'special rule,' that, on adoption by the House, determines the conditions under which the bill will be debated, amended, and voted upon. The power and flexibility of special rules to alter, amend, and waive the rules of the House is staggering. Special rules can either make no amendments in order (a closed rule), or place no restrictions on amendments (an open rule). Alternatively, they can make a specified list of amendments in order, they can declare amendments adopted without a vote, or they can amend amendments without the permission of the amendment's author. Special rules can alter or waive the rules of the House for the purposes of considering a particular bill. In short, special rules are infinitely flexible, limited only by the creativity of the Rules Committee, allowing the majority to do nearly anything it wants. Rules written by the Rules Committee must be adopted by a majority vote of the House. In general, the majority party members work closely with their party leaders to ensure that the position preferred by the party is able to prevail, and that potentially troublesome amendments are either forbidden, or considered in such a way that they will do the least possible harm to majority party control.

Senate Procedure

While action in the House is structured closely by its extensive rules, the Senate, by contrast, has relatively few rules and frequently ignores what rules it has. The Senate gives individual senators an almost unbelievably large amount of discretion and power. This has the predictable consequence of sometimes of paralyzing the Senate. The House works by assembling majorities and voting, but the Senate must work by a process of accommodating the individual senators, not by forming majorities, but by working to overcome the objections of individual senators.

The central feature of Senate procedure, around which all else is organized, is the possibility of the filibuster. Filibusters are efforts to defeat bills, not by voting them down, but by means of delay. The goal of filibusterers is to convince supporters of legislation that it will take so much time to pass the bill that they would be better off dropping it and moving on to something else.

Filibusters are possible because of three glaring omissions from Senate rules. The Senate lacks any means of limiting debating time, it lacks a 'germaneness' rule for interventions

and amendments, and it lacks a previous question motion. Because the House rules have all of these key features, dilatory tactics are not possible there. The usual filibustering tactics include the so-called 'extended debate,' whereby senators talk endlessly, and not necessarily about the bill in question, offering numerous amendments, which need not be germane to the bill, and offering numerous motions, such as suggesting the absence of a quorum, intended only to delay action. Filibusters can be ended by means of the 'cloture' process. If 60 senators are willing to vote for cloture, debate can be ended, but only after all amendments have been disposed of, and all senators have had an additional opportunity to speak for an hour.

Because even a single determined senator can cause a lengthy delay, losing the Senate the opportunity to consider other important legislation, the focus of Senate activity is not on breaking filibusters through cloture, but rather on avoiding them altogether. In advance of bringing legislation to the floor, bill sponsors often engage in extensive discussions with their colleagues in an effort to anticipate and satisfy objections. Most legislation passes the Senate with only minimal objection, in part because of efforts to build consensus early.

There have been no major changes to the rules governing filibusters since 1975, but in recent congresses there have been new efforts to make it easier for the majority to rule. Since the 1990s, the filibuster has become largely a tool used by the minority party to obstruct the agenda of the majority. Presidential nominations to positions in the executive and judicial branches have come to be routinely filibustered, resulting in significant delays and prolonged vacancies in important positions. To pass, bills must now effectively win a 60 vote supermajority. The Senate majority party, frustrated by delay and obstruction, has increasingly threatened to change the rules to make filibusters easier to curb. According to a traditional understanding of Senate procedure, rule changes, including changes to rules governing filibusters, are themselves subject to a filibuster. When they have been in the majority, both Democrats and Republicans have threatened to employ a novel procedure and change filibuster rules by a simple majority vote. To prevent this from happening, the minority party has made concessions and, at least temporarily, moderated their filibustering to keep the majority from changing the rules. It seems improbable, however, that the filibuster can survive much longer, when it serves primarily the interests of the minority party.

Congress and the President

By dividing the legislative and executive powers into separate branches, the authors of the US Constitution sought to encourage friction and disagreement in government. This plan has worked well for over two centuries, as Congress remains determinedly independent. Through the power of the veto, every president gains significant power in the legislative process. Vetoes of bills can be overridden by a two-thirds vote of both chambers of Congress, but in practice this happens seldom. In the past four presidencies, from 1989 to the present, there have been 95 total vetoes but only seven overrides. Presidents constantly use the threat of a veto to force Congress to make changes in pending legislation. The veto is potent, but it is mostly negative: presidents cannot readily use the veto power to force Congress to pass legislation.

Congressional independence is nowhere better demonstrated than in its handling of the federal budget. By statute, the president is required to submit to Congress each year a detailed budget proposal. Nothing, however, requires Congress to pay much attention to presidential proposals. Commonly in recent years the president's budget has been declared 'dead on arrival' at Capitol Hill, and Congress has gone on to draft its own, different budget, relying on its budget process, its committees, and its own budget agency. The president has influence over members of Congress, especially those from the president's own party; however, even members of the president's party are influenced strongly by their constituencies.

Congress and Foreign Policy

Congress defers to the executive branch to a great extent in foreign policy, but with important exceptions. Even though the Constitution gives Congress the power to declare war, power over warmaking drifted away from Congress to the president over the twentieth century, in part because of a need for speed and secrecy, and in part because of an unwillingness in Congress to challenge the president over the war power. Congress passed resolutions authorizing major conflicts in Iraq and Kuwait, but many other conflicts were initiated by the president without any authorization from Congress. The Korean War is the most prominent of these undeclared wars, but many other less bloody conflicts were begun entirely on the unilateral authority of the president, including invasions of Panama, Grenada, peacekeeping operations in Somalia and the Balkans, and support for insurgencies such as Libya in 2011.

In a series of decisions the federal courts made it clear that they were uninterested in adjudicating war powers disputes between Congress and the president (Fisher, 2000).

Congress passed the War Powers Resolution in 1973 over President Nixon's veto. It was an effort to delineate congressional and presidential powers over war. It ceded the president the power to introduce U.S. forces into situations involving hostilities for a period up to 90 days without congressional authorization. Forty years later, it appears that the War Powers Resolution has done practically nothing to restrain presidents. The development of drone warfare seems likely to shift the balance of power over warmaking even more in the president's direction. Since drones do not put U.S. forces in danger, they skirt the restrictions of the War Powers Resolution.

Polarization

To say that Congress is more polarized means that the ideological centers of the two parties have moved apart. For most of the twentieth century there was a substantial overlap between the two major parties. The most conservative Democrats were more conservative than many Republicans and the most liberal Republicans were more liberal than many Democrats. The overlap has vanished.

Polarization can be assessed by means of different kinds of scores that aggregate roll call voting of individual members into

an index of their overall voting tendency. The most common such measures in use today are the NOMINATE scores developed by Poole, Rosenthal, and McCarty. The mean NOMINATE scores of Republicans and Democrats in both the Senate and House have moved steadily away from the center since the 1970s. As of 2012, the parties were more polarized than at any in more than a century, possibly more polarized than at any other time in American history (McCarty et al., 2008).

The fact of increased polarization is undisputed, but the causes are complicated. As with any profound social change, it is wise to look beyond any single cause. At least three major explanations of polarization have been advanced. One is that the American electorate is more polarized and they not surprisingly elect more extreme representatives and senators (Abramowitz). A second is that Democrats and Republicans have 'sorted' themselves more efficiently into the correct party. Thee are currently far fewer Democrats being elected from conservative districts and fewer Republicans being elected from liberal districts (Fiorina et al., 2010). A third is that increased gerrymandering has made districts safe for one party or the other, which encourages members to move from the center. If representatives and senators are more or less assured of victory in the general election, losing a primary election becomes a greater risk to their reelection, and representatives will rationally shift from the center to appeal more to the sharp partisans who tend to dominate primaries. There is strong evidence to support the first two of these explanations, but the gerrymandering explanation is disputed by McCarty et al. (2009), who find that gerrymandering contributes only modestly to increased polarization. Further, gerrymandering cannot explain increased polarization in the Senate.

A polarized Congress functions differently than a non-polarized Congress. With the parties farther apart ideologically, it is harder to forge cross-party coalitions to pass legislation. In previous decades, a typical strategy in passing legislation was to assemble a coalition of most of one party, joined by members of the other party who were ideologically in agreement. In a polarized Congress major legislation tends to pass with few votes from the opposition party.

Sharp disagreements across the parties have led to the rise of highly confrontational tactics. Under Speaker Newt Gingrich, the Republicans in 1995 refused to pass appropriations bills, shutting down portions of the federal government, as a means of forcing President Bill Clinton to sign legislation cutting the budget deficit. In 2011, Republicans in the House of Representatives employed a similar tactic, refusing for months to pass legislation raising the federal debt ceiling, seeking to force President Obama to accept their budget plan (Draper, 2012). These theatrics create an uncertainty about whether the government will be able to function or pay its bills that was mostly absent in the era prior to polarization.

Democratic Accountability

Ideally, individual members of Congress should be held accountable for the collective actions of Congress, but this does not necessarily occur (Arnold, 1990). While members of Congress sometimes campaign for reelection by proclaiming the major legislative accomplishments of Congress, often they detach themselves from the institution as a whole and seek reelection for particular services rendered to the district or state. Adopting pork barrel projects such as dams, roads, and parks can help to make individual members popular, but the appearance of waste and venality can make Congress less popular. Fenno argued that representatives and senators distance themselves from Congress as an institution and seek to promote themselves separately, often at the expense of Congress as a whole (Fenno, 1978). The accountability of Congress to voters is inherently difficult because Congress acts as an institution, but its members are judged individually.

See also: Electoral Geography; Electoral Systems; Majoritarianism and Majority Rule; Parliamentary Government; Parliaments, History of; Presidency, United States.

Bibliography

Abramowitz, A.I., 2010. The Disappearing Center: Engaged Citizens, Polarization, & American Democracy. Yale University Press, New Haven, CT.

Arnold, R.D., 1990. The Logic of Congressional Action. Yale University Press, New Haven, CT.

Binder, S., 1997. Minority Rights, Majority Rule: Partisanship and the Development of Congress. Cambridge University Press, Cambridge, UK.

Boatright, R.G., 2013. Getting Primaried: The Changing Politics of Congressional Primary Challenges. University of Michigan Press, Ann Arbor, MI.

Cann, D.M., 2008. Modeling committee chair selection in the U.S. House of Representatives. Political Analysis 16, 274–289.

Deering, C.J., Smith, S.S., 1997. Congressional Committees, third ed. Congressional Quarterly, Washington, DC.

Draper, R., 2012. Do Not Ask What Good We Do: Inside the U.S. House of Representatives. The Free Press, New York.

Edwards, G.C., 1989. At the Margins: Presidential Leadership of Congress. Yale University Press, New Haven, CT.

Fenno, R., 1978. Homestyle: House Members in Their Districts. Little, Brown, Boston, MA.

Fiorina, M.P., Abrams, S.J., Pope, J.C., 2010. Culture War? The Myth of a Polarized America, third ed. Longman, New York.

Fisher, Louis, 2000. Litigating the war power with Campbell v. Clinton. Presidential Studies Quarterly 30, 564–574.

Jacobson, G., 2012. The Politics of Congressional Elections, eighth ed. Pearson, New York.

King, A., 1997. Running Scared: Why America's Politicians Campaign Too Much and Govern Too Little. Free Press, New York.

Mayhew, D., 1974. Congress: The Electoral Connection. Yale University Press, New Haven, CT.

McCarty, N., Poole, K.T., Rosenthal, H., 2008. Polarized America: The Dance of Ideology and Unequal Riches. MIT Press, Cambridge, MA.

McCarty, N., Poole, K.T., Rosenthal, H., 2009. Does gerrymandering cause polarization? American Journal of Political Science 53, 666–680.

Polsby, N.W., 1968. The institutionalization of the US House of Representatives. American Political Science Review 62, 144–168.

Polsby, N.W., 1975. Legislatures. In: Greenstein, F.I., Polsby, N.W. (Eds.), The Handbook of Political Science, Governmental Institutions and Processes, vol. 5. Addison-Wesley, Reading, MA.

Sinclair, Barbara, 2011. Unorthodox Lawmaking: New Legislative Processes in the U.S. Congress, fourth ed. CQ Press, Washington, DC.

Tiefer, C., 1989. Congressional Practice and Procedure: A Reference, Research and Legislative Guide. Greenwood, New York.

Connectionist Models of Concept Learning

Koen Lamberts, University of Warwick, Coventry, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A category is a collection of objects or events, and a concept is the mental representation of a category. Connectionist models assume that concept learning involves the development of associative links between mental representations of objects or events and mental representations of categories. These links form the basis of categorization, which is the process of generating the category name that corresponds to a given object. In this article, the basic principles of connectionist models of concept learning are presented, and their empirical and theoretical merits are discussed. The discussion starts with a simple model of concept learning (the delta rule model), which has formed the basis of much research on connectionist models of concept learning. Next, the fundamental problems of the delta rule model are highlighted. Finally, an alternative to the delta rule model is presented.

A category is a collection of objects or events. Concepts can be seen as mental representations of categories, and concept learning can be defined as learning to assign objects or events to categories. The study of concept learning has a long history in psychology. Experimental research on concept learning has revealed several regularities, which seem to apply across a wide range of stimuli and learning conditions. These regularities must be explained by any valid model of concept learning. A detailed overview of basic findings on concept learning is provided in *Concept Learning and Representation: Models*. For some time, connectionist models of concept learning have been popular (e.g., Ashby et al., 1998; Gluck and Bower, 1988a, 1988b; Kruschke, 1992). In connectionist models, it is assumed that concept learning involves the development of associative links between objects and categories. These links form the basis of categorization, which is the process of assigning an object to a category. In this article, the basic principles of connectionist models of concept learning are presented, and their empirical and theoretical merits are discussed. There are so many different connectionist models of concept learning that it is impossible to cover them all here. Instead, the article will focus on a few models that are well established and have been particularly influential. The discussion starts with a simple model of concept learning (the delta rule model) that has formed the basis of much research on connectionist models of concept learning. Next, the fundamental limitations of the delta rule model are highlighted. Finally, an alternative to the delta rule model is presented.

The Delta Rule Model

One of the simplest and best understood learning rules in connectionist models is the so-called delta rule, which was proposed by Widrow and Hoff (1960) and is formally equivalent to the learning rule in Rescorla and Wagner's (1972) model of associative learning (see Sutton and Barto, 1981). The delta rule is an error-correcting rule. It assumes that feedback is available on the classification performance of a network, and that this feedback is available for each stimulus that must be classified. A learning scheme that fulfills this condition is called a supervised learning scheme.

To illustrate the delta rule, assume that we have the simple model of concept learning shown in Figure 1. The model takes the form of a connectionist network, with two connected layers of processing units. The input layer is used to represent or encode the stimuli that can be presented, whereas the output layer represents the response alternatives (in a categorization context, each response alternative typically corresponds to a single category). Input coding is achieved by a number of units, each of which corresponds to a single feature or dimension of the stimulus. Furthermore, each input unit is connected to each output unit. The connections in the network have associated strength values. The strength of a connection determines the effect that activation of its source or input unit has on its output unit. In a network of this form, category learning is assumed to involve the modification of the strengths of the connections between input and output units in such a way that presentation of a stimulus to the input units will give rise to the activation of the output unit that corresponds to the category to which the stimulus belongs.

The input to the network on a given trial is represented by the pattern of activation across the input units. Suppose that each input unit represents a stimulus feature that can be either present or absent (such as 'red,' or 'large'). The set of features that characterizes a stimulus can then be represented by setting

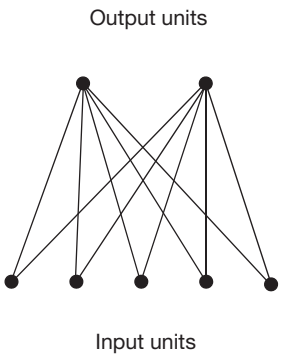

Figure 1 Delta rule network architecture.

the activation of each input unit that corresponds to a feature that is present to 1, whereas the activation of the other units can be set to 0. Of course, other representation schemes are possible, depending on the nature of the stimuli and the requirements of the learning task.

Upon presentation of a stimulus, the activation of the output units is computed as follows:

$$a_o = \sum_i w_{io} a_i \quad [1]$$

in which a_o is the activation of output unit o, a_i is the activation of input unit i, and w_{io} is the strength of the connection between input unit i and output unit o. Next, feedback about the network's output activation pattern is provided, and the discrepancy between the current output pattern and the desired output pattern is computed. In a concept-learning task, the desired output would typically be that one output unit (corresponding to the correct category) is activated, whereas the other units remain inactive. For each output unit, an error value e_o is computed:

$$e_o = d_o - a_o \quad [2]$$

in which d_o is the desired output value for unit o. Finally, the connection strengths in the network are modified as follows:

$$\delta w_{io} = \lambda a_i e_o \quad [3]$$

in which δw_{io} is the change in the connection strength and λ is a learning rate parameter.

After a number of presentations of the entire set of stimuli in a category-learning task, the delta rule is guaranteed to produce a set of connection strengths that is optimal for the category-learning task at hand, within the constraints imposed by the architecture of the network (Kohonen, 1977). For some category structures, the simple network in Figure 1 is sufficient to achieve perfect categorization of all stimuli. If perfect categorization is impossible (for reasons that will be explained in this article), the delta rule still converges upon a set of connection strengths that produces the best performance that can be achieved with the network to which it is applied.

Evaluation of the Delta Rule Model

Delta rule networks have been evaluated in a large number of empirical studies on concept learning (e.g., Estes et al., 1989; Friedman et al., 1995; Gluck and Bower, 1988a,b; Shanks, 1990, 1991) with considerable success. However, although the delta rule model can explain important aspects of human concept learning, it has a major weakness: It fails to account for people's ability to learn categories that are not linearly separable. In a network of the kind described here, the activation of any output unit is always a weighted sum of the activation of the input units. This implies that the network can learn only categories that can be separated by a linear function of the input values. In human concept learning, linear separability does not appear to be an important constraint. It has been shown repeatedly that people can learn nonlinearly separable category structures without difficulty (e.g., Medin and Schwanenflugel, 1981; Nosofsky, 1987).

There are several ways in which delta rule networks can be modified to handle nonlinearly separable categories. One way is to augment the input layer of the network with units that become active only when particular combinations of two or more features are present in the input (e.g., Gluck and Bower, 1988b; Gluck, 1991). If enough configural units are present (and if they encode the correct combinations of features), such networks can be made to learn any category structure using the delta rule. However, it is not clear that learning in such networks corresponds well to human learning, or that configural cue networks explain categorization after learning (Choi et al., 1993; Macho, 1997; Nosofsky et al., 1994; Palmeri, 1999). Moreover, the number of possible configural units grows exponentially with the number of stimulus dimensions. Configural cue models are therefore not particularly attractive as models of human concept learning.

A second way of modifying delta rule networks so that they can learn nonlinearly separable categories involves the use of a layer of 'hidden' units between the input and output units. Rumelhart et al. (1986) proposed a generalization of the delta rule for such networks. This backpropagation-of-error rule is used to determine how much the connection strengths between input and hidden units, and between hidden and output units, should be changed on a given learning trial to achieve the desired mapping between input and output. If the units in such multilayer networks have nonlinear activation functions, backpropagation networks can learn nonlinearly separable categories; in fact, they can learn arbitrary mappings between inputs and outputs, provided that they contain a sufficient number of hidden units (Hornik et al., 1989).

Despite their intuitive appeal and obvious computational power, backpropagation networks are not adequate as models of human concept learning (e.g., Kruschke, 1993). Their most important shortcoming is that they suffer from 'catastrophic forgetting' of previously learned concepts when new concepts are learned (McCloskey and Cohen, 1989). To overcome this difficulty, various alternative learning schemes have been proposed. One of the most successful alternatives was developed by Kruschke (1992), who proposed a hidden-unit network that retains some of the characteristics of backpropagation networks but does not inherit their problems.

Multilayer Networks

ALCOVE (Kruschke, 1992, 1993) is a connectionist network with three layers of units: an input layer, a hidden layer, and an output layer (see Figure 2). The model is based on an exemplar theory of concept learning and categorization, Nosofsky's (1986) Generalized Context Model (GCM; *see* Concept Learning and Representation: Models). Each input unit i of ALCOVE encodes a single stimulus dimension and is gated by a dimensional attention weight α_i, which reflects the relevance of the dimension for the learning task at hand.

Each hidden unit has a 'position value' on each stimulus dimension, which means that each hidden unit corresponds to a particular stimulus or exemplar. Whenever an input pattern is presented to the network, each hidden node is activated according to the similarity between the stimulus it represents

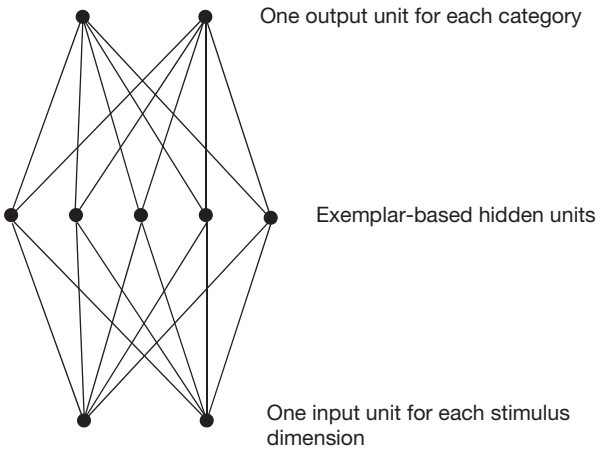

Figure 2 Architecture of the ALCOVE model of concept learning.

and the input pattern. In ALCOVE, similarity is defined as in Nosofsky's GCM:

$$a_j^{hid} = \exp\left[-c\left(\sum_i \alpha_i \left|h_{ij} - a_i^{in}\right|^r\right)^{q/r}\right] \quad [4]$$

in which a_j^{hid} is the activation of hidden unit j, h_{ji} is the position of hidden unit j on stimulus dimension i, c is a positive constant called the specificity of the hidden unit, a_i^{in} is the activation of input unit i, and r and q determine the similarity metric and similarity gradient, respectively.

Each hidden unit is connected to output units that correspond to response categories. The category units are activated according to the following rule:

$$a_k^{out} = \sum_j w_{kj} a_j^{hid} \quad [5]$$

in which w_{kj} is the association weight between hidden unit j and category unit k. This is the same activation rule used in the simple delta rule network discussed in this article (see eqn [1]). Finally, category unit activations are translated into response probabilities by the following rule:

$$P(K) = \exp(\phi a_j^{out}) / \sum_m \exp(\phi a_m^{out}) \quad [6]$$

where ϕ is a scaling constant. ALCOVE employs a variation of the backpropagation learning rule to adjust dimensional attention weights α_i and association weights w_{kj} in the course of learning (for details of the learning rule, see Kruschke, 1992). As a result of these adjustments, the network will eventually learn to classify each stimulus into the correct category.

ALCOVE has great advantages over the simple delta rule network for concept learning. The model is not affected by the linear separability constraint. Like standard backpropagation networks, ALCOVE can learn arbitrary mappings between stimuli and categories. Moreover, ALCOVE does not suffer from catastrophic forgetting (Kruschke, 1993). Most important, however, is the correspondence between ALCOVE's predictions about concept learning and human performance (e.g., Choi et al., 1993; Kruschke, 1992, 1993; Nosofsky et al., 1992; Palmeri, 1999). Since its original formulation, the model has undergone a number of refinements and modifications, which allow it to explain concept learning in a variety of conditions (e.g., Kruschke, 1996, 2008; Kruschke and Johansen, 1999; Denton and Kruschke, 2006). The model has also provided inspiration for some of the principles of other successful connectionist models of concept learning (e.g., Love et al., 2004). ALCOVE and its successors ultimately derive their strength from a combination of the principles of exemplar-based processing with those of associative learning.

Conclusions and Future Directions

Whereas connectionist models such as ALCOVE can explain many important aspects of human concept learning, it is becoming increasingly clear that they also have fundamental limitations. The models that were reviewed here all assume that concept learning is an associative process, in which links between stimulus and category representations are modified. There is little doubt that many concepts are learned in this way. However, the associative model does not apply to the learning of all concepts. Some concepts are learned by a process of rule discovery, which has characteristics that are very different from those of connectionist models of learning. An important challenge is to determine when associative models and rule-based models of concept learning apply. There have been a number of proposals for how associative and rule-based concept-learning principles can be combined (e.g., Ashby et al., 1998, 2010; Erickson and Kruschke, 1998; Love and Tomlinson, 2010), and it is likely that such models will become increasingly prominent.

See also: Cognitive Modeling: Connectionist Approaches; Concept Learning and Representation: Models.

Bibliography

Ashby, F.G., Alfonso-Reese, L.A., Turken, A.U., Waldron, E.M., 1998. A neuropsychological theory of multiple systems in category learning. Psychological Review 105, 442–481.

Ashby, F.G., Paul, E.J., Maddox, W.T., 2010. COVIS. In: Pothos, E.M., Wills, A.J. (Eds.), Formal Approaches to Categorization. Cambridge University Press, Cambridge, UK.

Choi, S., McDaniel, M.A., Busemeyer, J.R., 1993. Incorporating prior biases in network models of conceptual rule learning. Memory and Cognition 21, 413–423.

Denton, S.E., Kruschke, J.K., 2006. Attention and salience in associative blocking. Learning & Behavior 34, 285–304.

Erickson, M.A., Kruschke, J.K., 1998. Rules and exemplars in category learning. Journal of Experimental Psychology: General 127, 107–140.

Estes, W.K., Campbell, J.A., Hatsopoulos, N., Hurwitz, J.B., 1989. Base-rate effects in category learning: a comparison of parallel network and memory storage-retrieval models. Journal of Experimental Psychology: Learning, Memory, and Cognition 15, 556–571.

Friedman, D., Massaro, D.W., Kitzis, S.N., Cohen, M.M., 1995. A comparison of learning models. Journal of Mathematical Psychology 39, 164–178.

Gluck, M.A., 1991. Stimulus generalization and representation in adaptive network models of category learning. Psychological Science 2, 50–55.

Gluck, M.A., Bower, G.H., 1988a. Evaluating an adaptive network model of human learning. Journal of Memory and Language 27, 166–195.

Gluck, M.A., Bower, G.H., 1988b. From conditioning to category learning: an adaptive network model. Journal of Experimental Psychology: General 117, 227–247.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366.

Kohonen, T., 1977. Associative Memory: A System Theoretical Approach. Springer, New York.

Kruschke, J.K., 1992. ALCOVE: an exemplar-based connectionist model of category learning. Psychological Review 99, 22–44.

Kruschke, J.K., 1993. Human category learning: implications for backpropagation models. Connection Science 5, 3–36.

Kruschke, J.K., 1996. Base rates in category learning. Journal of Experimental Psychology: Learning, Memory and Cognition 22, 3–26.

Kruschke, J.K., 2008. Bayesian approaches to associative learning: from passive to active learning. Learning & Behavior 36, 210–226.

Kruschke, J.K., Johansen, M.K., 1999. A model of probabilistic category learning. Journal of Experimental Psychology: Learning, Memory and Cognition 25, 1083–1119.

Love, B.C., Medin, D.L., Gureckis, T.M., 2004. SUSTAIN: a network model of category learning. Psychological Review 111, 309–332.

Love, B.C., Tomlinson, M., 2010. Mechanistic models of associative and rule-based category learning. In: Mareschal, D., Quinn, P.C., Lea, S.E.G. (Eds.), The Making of Human Concepts. OUP, Oxford, UK.

Macho, S., 1997. Effect of relevance shifts in category acquisition: a test of neural networks. Journal of Experimental Psychology: Learning, Memory and Cognition 23, 30–53.

McCloskey, M., Cohen, N.J., 1989. Catastrophic interference in connectionist networks: the sequential learning problem. The Psychology of Learning and Motivation 24, 109–165.

Medin, D.L., Schwanenflugel, P.J., 1981. Linear separability in classification learning. Journal of Experimental Psychology: Human Learning and Memory 7, 355–368.

Nosofsky, R.M., 1986. Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General 115, 39–57.

Nosofsky, R.M., 1987. Attention and learning processes in the identification and categorization of integral stimuli. Journal of Experimental Psychology: Learning, Memory, and Cognition 13, 87–108.

Nosofsky, R.M., Gluck, M.A., Palmeri, T.J., McKinley, S.C., Glauthier, P., 1994. Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland, and Jenkins 1961. Memory and Cognition 22, 352–369.

Nosofsky, R.M., Kruschke, J.K., McKinley, S.C., 1992. Combining exemplar-based category representations and connectionist learning rules. Journal of Experimental Psychology: Learning, Memory and Cognition 18, 211–233.

Palmeri, T.J., 1999. Learning categories at different hierarchical levels: a comparison of category learning models. Psychonomic Bulletin and Review 6, 495–503.

Rescorla, R.A., Wagner, A.R., 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A.H., Prokasy, W.F. (Eds.), Classical Conditioning II: Current Research and Theory. Appleton-Century-Crofts, New York, pp. 64–99.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536.

Shanks, D.R., 1990. Connectionism and the learning of probabilistic concepts. Quarterly Journal of Experimental Psychology 42A, 209–237.

Shanks, D.R., 1991. Categorization by a connectionist network. Journal of Experimental Psychology: Learning, Memory, and Cognition 17, 433–443.

Sutton, R.S., Barto, A.G., 1981. Toward a modern theory of adaptive networks: expectation and prediction. Psychological Review 88, 135–170.

Widrow, C., Hoff, M.E., 1960. Adaptive switching circuits. Institute of Radio Engineers, Western Electronic Show and Convention. Convention Record 4, 96–104.

Conscientiousness

Anissa Mike and Kelci Harris, Washington University in St. Louis, St. Louis, MO, USA
Brent W Roberts, University of Illinois at Urbana-Champaign, Champaign, IL, USA
Joshua J Jackson, Washington University in St. Louis, St. Louis, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The personality trait of conscientiousness is related to a great number of important life outcomes. To better understand why the influence of conscientiousness is so pervasive, the current article provides a broad overview of the trait. First, the lower-order facet structure of conscientiousness is described, including a description of some of the most common ways to measure both the broad trait as well as the facets. Next, we review the influence of conscientiousness on several life domains and discuss pathways that relate the two. Finally, the development of conscientiousness is explored.

Conscientiousness is a personality trait that is best known for belonging to the taxonomy of the 'Big Five' personality traits (Goldberg, 1993), which also includes extraversion, agreeableness, neuroticism, and openness. Individuals high in conscientiousness tend to follow socially prescribed norms for impulse control, to be goal directed, planful, and to be able to delay gratification (Roberts et al., 2004). While typically thought of as belonging to the Big Five, constructs related to conscientiousness were discussed as far back as Aristotle and are included in classic personality taxonomies set forth by Cattell (1957), Eysenck (1947), Gough (1956), Tellegen and Waller (2008).

As seen in the definition, conscientiousness is not a singular dimension but instead encompasses a broad range of lower-order individual differences. In this review, we first unpack these lower-order facets of conscientiousness, and describe commonly used measures that assess conscientiousness and its lower-order facets. Much of the recent interest in conscientiousness stems from its associations with a number of important life outcomes. Thus, in the following section, we discuss the relationship between conscientiousness and important life outcomes, detailing how lower-order facets can help explain these associations. Finally, we describe the development of conscientiousness over the life span, focusing on the lower-order components.

The Lower-Order Structure of Conscientiousness

Personality traits describe the relatively enduring patterns of thoughts, feelings, and behaviors of individuals. While typically thought of as decontextualized constructs, traits can range from broad, overarching concepts to more specific, concrete manifestations. In other words, personality traits can be conceptualized as existing within a hierarchical structure, with broad constructs structured above more specific, lower-order facets of these traits, and behavioral, cognitive, and affective manifestations of these traits falling at even lower ends of the hierarchy (Jackson et al., 2009; Markon et al., 2005; Roberts et al., 2004). At the highest levels of the hierarchy, personality traits can be meaningfully organized into two (Digman, 1997), three (Tellegen and Waller, 2008), and even four factors (Markon et al., 2005), before settling into the well-known Big Five factors (Goldberg, 1993).

Similar to the hierarchical structure of personality traits, the trait of conscientiousness can be decomposed into narrower, lower-order facets, which can be broken down even further into more concrete behaviors, thoughts, feelings, and habits, like setting goals for oneself or turning work in on time. At the most basic two-factor division, conscientiousness can be thought of as consisting of two components, a proactive component and an inhibitory component (Costa et al., 1991; Jackson et al., 2010; Roberts et al., 2005a). The proactive component refers to being goal oriented and striving to do well in one's endeavors, while the inhibitory component refers to being responsible, delaying gratification, and controlling impulses. Together, these two overarching components provide a structure under which the rest of the facets can be organized. It should be noted that an alternative perspective finds that the two-facet structure of conscientiousness consisting industriousness and orderliness (DeYoung et al., 2007). While this two-factor solution may initially seem similar to the proactive and inhibitory components of conscientiousness, empirical examinations of the lower-order structure of conscientiousness show that industriousness and orderliness fall in the proactive component only (Roberts et al., 2005a).

More clearly defined facets emerge as one moves down the conscientiousness hierarchy. Three-factor solutions typically consist of factors associated with achievement, self-control, and responsibility (Jackson et al., 2010; Roberts et al., 2005a), though few studies have explicitly tested three-factor solutions. Perhaps the most agreed upon level of analysis for conscientiousness is the four-factor solution. Numerous studies identify four interpretable factors, namely orderliness, industriousness, responsibility, and self-control (Jackson et al., 2010; Peabody and De Raad, 2002; Roberts et al., 2004). While these four facets replicate across most studies, additional facets of conscientiousness, such as conventionality and formality, appear further down the hierarchy (Roberts et al., 2004, 2005a). Currently, however, these additional facets do not replicate across studies as consistently as the four-factor solution, making it difficult to identify the structure of five-factor solutions or greater.

Two facets of the four-factor solution, orderliness and industriousness, fall under the proactive domain of conscientiousness. Individuals high in orderliness are likely to be neat and clean, and situate their belongings in an organized

 http://dx.doi.org/10.1016/B978-0-08-097086-8.25047-2

fashion (Jackson et al., 2009). To a lesser extent, orderliness also includes one's proclivity toward making and adhering to plans and utilizing to-do lists and planners that aide in a more organized and orderly day. Industriousness refers to how hard-working one is, as well as one's ability to persevere and strive for success. Industrious individuals like to finish what they start, and will work long hours or even take on extra work in order to do a thorough job (Jackson et al., 2009, Mike et al., 2014). Individuals low in industriousness may avoid work, procrastinate, or give up easily when facing adversities. Naturally, high levels of industriousness are associated with succeeding in achievement domains such as academics and work.

Self-control falls on the inhibitory side of conscientiousness. Self-control may be labeled differently from measure to measure and is commonly called impulse control, impulsivity (reversed), constraint, or self-discipline. Within other fields, it may even be referred to as effortful control or executive functioning (DeYoung, 2011). Regardless of the name, self-control refers to the ability to inhibit impulses and avoid being reckless or out of control. Individuals high on self-control are able to put off immediate gratification in order to attain larger goals. Individuals low on self-control are more likely to make impulsive purchases, to put off or cancel plans at the last minute, and to be unable to control reactions when feeling angry or frustrated (Jackson et al., 2010). As indicated in the dissent in naming this construct, many disagree as to which Big Five trait self-control is associated with. Sometimes self-control is considered to be a facet of extraversion because of its relationship with sensation seeking; self-control can also be thought of as a component of neuroticism, where impulsive actions are likely to occur because of emotional instability (DeYoung, 2011). However, continuing evidence indicates that self-control is primarily related to conscientiousness (Paunonen, 1998; Roberts et al., 2004).

The final facet of the four-factor solution, responsibility, can be considered as both inhibitory and proactive. Sometimes referred to as reliability, this facet describes one's tendency to uphold obligations and follow rules. Those high in responsibility are less likely to break promises, miss appointments, or oversleep (Jackson et al., 2010). Those low on responsibility may be described as unreliable or flighty and are more likely to bend or break rules. The facet of responsibility often tends to be strongly associated with the trait of agreeableness, partly because the behaviors associated with the facet frequently involve an interpersonal context.

Conceivably, conscientiousness consists of more than four factors, and, indeed, solutions up to eight factors have been found (Roberts et al., 2004). Three additional facets are of particular note, as they have independently emerged across a number of different studies: conventionality, formality, and punctuality (MacCann et al., 2009; Roberts et al., 2004, 2005a). Conventionality is the inclination to support and follow the norms of society in order to maintain good social environments. Individuals high in conventionality are more likely to uphold family and cultural traditions, while individuals low in conventionality may be described as nonconforming or untraditional. Conventionality also extends to having more traditional or conservative beliefs, particularly within the political and spiritual spectrum, and is closely related to the trait of openness to experience (Johnson, 1994). Those high in formality can best be described as having a sense of propriety, which includes being polite and proper, and keeping one's physical appearance neat and clean. They may call others by formal titles, say please and thank you, and are less likely to swear. It should be noted that despite the different definitions for conventionality and formalness, in practice it is often difficult to meaningfully differentiate between the two as they are highly related (Jackson et al., 2010).

Finally, punctuality describes the practice of doing things in a timely manner and could be considered the narrowest facet of conscientiousness. Interestingly, however, punctuality tends to be equally related to each of the other facets of conscientiousness – unlike the other facets of conscientiousness – and thus could serve as a useful marker for overall levels of conscientiousness (Jackson et al., 2010).

In summary, conscientiousness consists of many distinct lower-order facets. Much like higher order personality traits, facets of conscientiousness can be ordered in a hierarchy. While the overarching concept of conscientiousness is generally agreed upon, a standard facet-level structure has not reached consensus. The most commonly identified facets are orderliness, industriousness, self-control, and responsibility (Roberts et al., 2012). However, five-factor solutions, which typically include conventionality or decisiveness, have been suggested (Jackson et al., 2009; Saucier and Ostendorf; Peabody and De Raad, 2002). Further down the hierarchy, there are 6-, 8-, and even 9-component solutions that may represent reasonable factor structures (MacCann et al., 2009; Madhavan, 2004; Roberts et al., 2004, 2005), though there has been less research to replicate these findings. As such, the measurement of conscientiousness reflects the discrepancy in facet solutions, as facets names and number of facets vary from scale to scale.

How to Measure Conscientiousness

Numerous measures of conscientiousness exist. Many of these measures belong to larger omnibus inventories that assess all of the Big Five factors, though some come from measures outside of the Big Five. The choice of which measure to use depends upon the researcher's needs and time constraints. Some measures aim to assess only the broad construct of conscientiousness while others break the assessment up into specific facets. Adding to the confusion, many of these facet measures have names that are specific to the scale, making comparisons across measures difficult. Measures also vary in length, ranging from two items to dozens of items.

One of the most common measures of conscientiousness is from the NEO-PI-R (Costa and McCrae, 1992). Though it was originally designed to measure only neuroticism, extraversion, and openness (hence NEO), it was later revised to include agreeableness and conscientiousness. The 48-item NEO measure of conscientiousness includes six facets, which are labeled: competence (closely related to the facet of industriousness), order, dutifulness (responsibility), achievement striving (industriousness), self-discipline (self-control), and deliberation (self-control). A related measure is the NEO-FFI, which is an abridged version of the NEO-PI-R, including only 12

items for conscientiousness. The NEO-FFI was not designed to assess particular facets but was intended to briefly measure the Big Five under time constraints. However, further research has indicated that the NEO-FFI can be used to assess a limited number of broader facets (Chapman, 2007; Saucier, 1998). For conscientiousness, this includes orderliness, goal-striving (industriousness), and dependability (responsibility).

Another common measure of conscientiousness is the Abridged Big Five Circumplex (AB5C; Hofstee et al., 1992). The AB5C has 486 items and includes nine facets of conscientiousness: organization, purposefulness, efficiency, rationality, dutifulness, orderliness, conscientiousness, perfectionism, and cautiousness. Unfortunately, these facets do not adequately cover the facets of conscientiousness discussed above, and appear to reduce to a three-factor solution (Roberts et al., 2005a). The purposefulness, efficiency, and rationality scales tap industriousness, while the perfectionism scale is mostly associated with orderliness; cautiousness being the only scale that reflects the self-control facet (Roberts et al., 2005a). Part of the reason for this lack of coverage is the way the AB5C derived its facets. The AB5C creates facets by crossing one Big Five trait with another under the idea that items do not neatly describe a single trait but instead are blends of multiple traits (e.g., being responsible may represent high levels of conscientiousness and agreeableness). Unfortunately, it appears that not all combinations of traits had an adequate number of items to meaningfully differentiate between these nine facets, resulting in a reduced number of facets captured.

The Big Five Inventory (BFI; John et al., 1991) is commonly used to measure the Big Five traits and includes nine items on conscientiousness. The BFI is a shorter length omnibus inventory and thus was not initially developed to measure facets of conscientiousness. However, two facets of conscientiousness may be captured using the BFI: self-discipline (self-control) and order (Soto and John, 2009).

The Ten Item Personality Inventory (TIPI, Gosling et al., 2003) was developed to be a very short and efficient measure of conscientiousness. Consisting of just two conscientiousness items, individuals rate how much they agree with the statements, "I see myself as dependable, self-disciplined," and "I see myself as disorganized, careless" (reverse-scored). Though longer measures are generally preferable (Credé et al., 2012), short measures such as the TIPI do provide adequate assessment of conscientiousness if one is under extreme time constraints.

The Multidimensional Personality Questionnaire (Tellegen and Waller, 2008) departs from the previous measures in that it does not explicitly measure the Big Five but instead focuses on the Big Three (positive emotionality, negative emotionality, and constraint). Here, conscientiousness is captured by constraint, which encompasses the lower-order facets of self-control, traditionalism (conventionality), and harm avoidance (which is more closely related to neuroticism and agreeableness). Another divergence from previous measures is that the 77-item constraint measure uses true–false items rather than likert scales.

The Hogan Personality Inventory (Hogan, 1992) has 206 items used to measure six personality trait dimensions called intellectance, adjustment, prudence, ambition, sociability, and likeability. Here, conscientiousness is referred to as prudence and is intended to identify individuals who are conscientious, conforming, and dependable. Facets of prudence are moralistic (conventional), mastery (industriousness), virtuous (formality), not autonomous (industriousness), not spontaneous (self-control), impulse control (self-control), and avoids trouble (responsibility).

A relatively new measure of conscientiousness is the Chernyshenko Conscientiousness Scales (Hill and Roberts, 2011). Six different facets are assessed across 60 items (orderliness, self-control, responsibility, traditionalism (conventionality), industriousness, and virtue). The virtuous facet is unique in that it assesses the tendency for conscientiousness individuals to be responsible members of the community who are honorable and rule abiding. It shares some overlap with the responsibility facet but is broader in scope and concerns issues of right and wrong rather than the tendency to follow through with personal obligations.

Conscientiousness can also be assessed using adjectives rather than statements (Saucier and Goldberg, 1996). The Conscientiousness Adjective Checklist (Jackson et al., 2009; Roberts et al., 2004) was designed to do exactly that. It consists of 123 personality adjectives that measure the facets of conscientiousness. This includes orderliness, industriousness, impulse control, responsibility, conventionality, and decisiveness (Roberts et al., 2004). Examples of adjectives included in the scale are organized, precise, thorough, responsible, reliable, and cautious.

Typical measures of personality traits such as the ones described above collapse items that tap thoughts, feelings, and behaviors to create an overall composite trait score. However, a scale was created in order to solely target the behavioral manifestations of conscientiousness, eschewing the assessment of thoughts and feelings (Jackson et al., 2010). Assessing only behaviors results in a more precise measure of an important component of conscientiousness, as such a scale will assess simply whether or not people behave conscientiously – regardless of their thoughts or feelings. The Behavioral Indicators of Conscientiousness scale (BIC; Jackson et al., 2010) includes a large pool of conscientiousness behaviors (185 behaviors ranging from cleaning a toilet, to canceling plans, to writing down important dates) that are rated on how often respondents engage in those behaviors. The BIC is ordered hierarchically and consists of a number of lower-order facets, depending on how broad or narrow the researcher's interest. Behavioral facets include: avoiding work, organization, impulsivity, antisocial behaviors, cleanliness, laziness, industrious, punctuality, formalness, responsibility, and appearance.

Other measures of conscientiousness that are less common, but just as valid, include the 20-item conscientiousness measure from the HEXACO inventory (Lee and Ashton, 2008). In this scale, conscientiousness is measured by assessing four facets: organization (orderliness), diligence (self-control), perfectionism (industriousness), and prudence (self-control). The Big Five Aspect Scale (BFAS; DeYoung et al., 2007) assesses two facets per Big Five trait (industriousness and orderliness for conscientiousness) using 10 items. Finally, the International Personality Item Pool (http://ipip.ori.org/) is a useful resource that offers a number of conscientiousness measures in addition to the AB5C and BFAS.

Conscientiousness and Important Life Outcomes

Conscientiousness plays an integral role in many important life outcomes, spanning success in school to living a longer and healthier life. Often, the associations between conscientiousness and these life outcomes are at levels equal to, if not better than, gold-standard predictor variables such as intelligence or socioeconomic status (Roberts et al., 2007b). This section discusses the various outcomes linked to conscientiousness and considers pathways responsible for each of the associations.

Conscientiousness is one of the best predictors of academic achievement, predicting GPA above and beyond even intelligence (Noftle and Robins, 2007). Conscientious students tend to have better grades, especially those who are high on the facet of impulse control (MacCann et al., 2009; Noftle and Robins, 2007; Paunonen and Ashton, 2001). These positive associations between conscientiousness and academic achievement are thought to be the result of how conscientious students think about school, their abilities, and the effort they put into achieving their academic goals. Conscientious students are more motivated to succeed in academic settings (Richardson and Abraham, 2009) and believe in their ability to be successful (Caprara et al., 2011). They are also more likely to couple their interests and self-concept in the domains in which they are successful (Denissen et al., 2007). For example, a conscientious student who makes good grades in geography is likely to both feel like she is good at geography and also be interested in geography. This interest flows back into one's achievement behaviors, which ultimately leads to higher grades. In general, conscientious students also devote more time to studying, completing projects, and turning assignments in on time. Together these characteristics make it more likely that conscientious students persevere in school, even through difficult or stressful situations.

As individuals leave school and enter into the work force, conscientiousness is again an important key to success. Conscientious people tend to earn more money (Ng et al., 2005; Spurk and Abele, 2011), are more satisfied with their jobs (Bowling and Burns, 2010; Ng et al., 2005), and are more likely to receive promotions (Ng et al., 2005) compared to those low in conscientiousness. Just as academic success was driven by the fact that conscientious individuals were engaged in effective academic behaviors, they are also, in effect, good workers. Conscientious individuals have greater career achievement related goals (Spurk and Abele, 2011), better attitudes toward their work (Bowling and Burns, 2010; Ng et al., 2005), they show more commitment to their jobs, and work well independently as well as in groups (Spurk and Abele, 2011; Sutin and Costa, 2010). Moreover, conscientious individuals tend to avoid detrimental activities such as stealing, drinking on the job, being tardy or missing work, and they are less likely to leave their jobs (Bowling and Burns, 2010; Roberts et al., 2007a).

Conscientiousness is also related to health and physical well-being. People who are conscientious report having fewer health problems than people who are not conscientious (Friedman et al., 1993; Lodi-Smith et al., 2009; Roberts et al., 2009) and conscientious individuals tend to live longer than their nonconscientious counterparts (Friedman et al., 1993; Hill and Roberts, 2011; Roberts et al., 2007b). Those low in conscientiousness are also more likely to develop specific diseases. such as metabolic syndrome, Alzheimer's disease, lung disease, and heart disease, and are also more likely to have a stroke (Sutin et al., 2011; Terracciano et al., 2013; Weston et al., 2014).

Most likely, conscientiousness individuals have more positive health outcomes due to engaging in more healthful behaviors, such as exercising and eating well (Lodi-Smith et al., 2009). Similarly, they are more likely to avoid unhealthful behaviors (Hampson et al., 2000), such as smoking, using drugs, alcohol, or engaging in risky driving or sexual behaviors (Jackson et al., in press; Weston and Jackson, in press; Roberts and Bogg, 2004; Friedman et al., 1995). If health issues do arise, conscientious individuals are more likely to adhere to their medical regimens (Hill and Roberts, 2011). By following their doctors' advice, and taking prescribed medications as needed, they are better able to overcome health ailments and prevent health declines. The association between conscientiousness and health emerges early in life – childhood levels of conscientiousness and related facets predict longevity and markers of health decades later (Friedman et al., 1993; Moffitt et al., 2011).

Conscientiousness also plays a role in the success and maintenance of romantic relationships. Conscientious people are less likely to get divorced (Solomon and Jackson, 2014), with some studies demonstrating conscientiousness' effect across decades (Roberts and Bogg, 2004). Likely this is a result of people high in conscientiousness having higher levels of relationship satisfaction and commitment levels (Dyrenforth et al., 2010; Solomon and Jackson, 2014b). In addition, highly conscientious people do fewer things to jeopardize their relationships (Hill et al., 2013). For example, infidelity is one of the main reasons for divorce, and conscientious individuals are less likely to cheat on their spouses (Buss and Shackelford, 1997). They are also less likely to abuse their spouses (Buss, 1991). Similarly, conscientious people are more likely to engage in relationship maintenance behaviors, such as correcting interpersonal mistakes and engaging in constructive problem solving (Baker and McNulty, 2011). Finally, conscientious individuals are more forgiving (Hill and Allemand, 2012), likely resulting in fewer disagreements and fights among couples.

Finally, it is noteworthy that the effects of conscientiousness on various life domains can extend beyond the conscientious individual and into the lives of close others. For example, the spouses of conscientious individuals tend to have better health outcomes (Roberts et al., 2009c) and work outcomes (Solomon and Jackson, 2014a).

While being high in conscientiousness has great benefits for school, work, health, and relationships, being low in conscientiousness can have serious negative consequences. First, levels of conscientiousness can have financial repercussions. Those low in conscientiousness are poor at managing their finances. They are less likely to save money and often have low credit scores (Bernerth et al., 2012; Webley and Nyhus, 2006). People low in conscientiousness are also more prone to gamble, particularly if they are highly impulsive (Myrseth et al., 2009). Moreover, individuals low in conscientiousness are more likely to engage in criminal behaviors (Jones et al., 2011; Wiebe,

2004), and low levels of conscientiousness predict getting arrested more often (Clower and Bothwell, 2001). Once arrested, prisoners with low conscientiousness levels spend a longer time in prison (Ireland and Ireland, 2011). Overall, it appears that conscientious individuals have positive life outcomes though engaging more often in constructive behaviors and avoiding harmful and counterproductive behaviors.

Development of Conscientiousness over the Life Span

While personality traits are relatively stable, they also demonstrate a degree of malleability and retain the capacity to change across the life. This means that conscientiousness is not a static construct, but can change over time – with individuals either increasing or decreasing with age. There are multiple ways to conceptualize this change. First, change can be thought of in terms or rank order stability – that is, how a person ranks on a trait compared to others. The rank order stability of conscientiousness increases steadily with age until plateauing between the ages of 50 and 70 (Caspi et al., 2005; Roberts and DelVecchio, 2000). This means that an individual who is above average on conscientiousness as a child will likely be above average on conscientiousness as an older adult.

However, these traits also remain open to outside influences and can change and develop over time. Mean level changes, or how a trait changes in absolute terms, also occur across the life span. In general, mean level changes in personality traits occur in the direction of greater psychological maturity (Roberts et al., 2008). Unsurprisingly, conscientiousness is part of this general trend toward maturity, and typically individuals increase in conscientiousness throughout the life span. Across a number of cross sectional and longitudinal studies conscientiousness was found to increase from young adulthood up to the age of 60 (Roberts et al., 2003b; Soto and John, 2012). This normative pattern of development replicates across a number of different cultures and countries, including Germany, Italy, Portugal, Croatia, South Korea (McCrae et al., 1999). Out of all of the Big Five traits, conscientiousness is the trait that evidences the largest increases, with increases as large as one full standard deviation across the life span (Roberts et al., 2005b).

While conscientiousness tends to increase across the life span, the rate and direction of change differs depending on age. Young adulthood has been identified as an especially important period for personality development and is possibly the period when personality traits change the most (Roberts et al., 2006b). In late childhood and adolescence, conscientiousness may demonstrate a negative trend before beginning to increase in young adulthood and continue to increase throughout adulthood (Soto et al., 2011). In old age, there is some recent evidence that conscientiousness may again demonstrate declines (Wortman et al., 2012). In general, findings regarding the trajectory of change in old adulthood are less cohesive than those for young and middle adulthood, as not all studies find older adults lower in conscientiousness (Allemand et al., 2008; Jackson et al., 2009).

The longitudinal studies of development that identify declines in conscientiousness during old age may differ from the studies that do not find declines in older adulthood in part because of the measures used to assess conscientiousness. Although conscientiousness as a whole generally increases across the life span, the facets of conscientiousness may each have unique developmental trajectories, demonstrating different trends from one another as well as from conscientiousness overall. Several studies examining the specific patterns of the development of conscientiousness facets find evidence for differences in development across the facets. For example, self-control increases across young adulthood and old age (Jackson et al., 2009; Soto et al., 2011). Industriousness demonstrates a similar pattern, increasing across the life span, though some evidence suggests that most of the changes occur earlier in adulthood (Jackson et al., 2009; Soto and John, 2012; Terracciano et al., 2005). Evidence regarding the development of responsibility suggests steady increases throughout the life span (Jackson et al., 2009). Conventionality does not show differences from young adulthood to middle adulthood, but older adults were found to be significantly more conventional than middle-aged adults (Jackson et al., 2009). Finally, the facet of orderliness differs from previous facets, as orderliness undergoes virtually no changes across the life span (Jackson et al., 2009; Soto et al., 2011, Soto and John, 2012). Overall, increases in conscientiousness earlier in life can be thought of as mainly resulting from increases in impulse control and industriousness, while increases in later adulthood are driven by changes in impulse control, reliability, and conventionality.

While the majority of people increase in conscientiousness as they age, there is a significant amount of people who do not change at all on conscientiousness, or perhaps even decrease (Lüdtke et al., 2011). These variations in development are thought to occur because of the unique experiences people go through and the roles that they take on throughout life (Lodi-Smith, 2007). While it is still unclear what experiences drive changes in conscientiousness, a number of studies have begun to associate life experiences with changes in conscientiousness.

Some of the most influential experiences relate to achievement, particularly achievement within academic settings. Students tend to increase in conscientiousness during their final year of high school, likely in preparation for the new responsibilities they anticipate taking on in college or in the workforce (Bleidorn, 2012). However, individual experiences are important as not everyone follows this pattern. Those who are not invested in school tend to change less compared to those that do invest in school. Similarly, failing an exam is associated with decreases in conscientiousness compared to those that did not fail an exam (Lüdtke et al., 2011).

Leaving school and obtaining one's first job is also associated with increases in conscientiousness (Specht et al., 2011). Similar to the findings in school settings, individuals who invest in their work also tend to increase in conscientiousness (Roberts et al., 2003a; Roberts et al., 2006a). Again, the opposite is also true; people who de-invest in work, such as committing counterproductive work behaviors, tend to become less conscientious, contrasting the typical developmental trend (Hudson et al., 2012; Roberts et al., 2006a).

In addition to one's level of conscientiousness predicting health and health behaviors, changes in conscientiousness also relate to health and changes in health behaviors. In one study, increases in conscientiousness were positively related

with changes in preventative health behaviors, such as healthy eating, exercising, and safe driving (Takahashi et al., 2013). These findings suggest that changes in conscientiousness are associated with changes in the health process. Likely, as people increase in conscientiousness, they increase in the ability and tendency to take better care of themselves, ultimately resulting in better overall health (Takahashi et al., 2013). Changes in health behaviors can also help explain changes in conscientiousness. For example, low levels of conscientiousness are associated with substance abuse in middle age, and continued substance abuse in turn is associated with nonnormative decreases in conscientiousness (Roberts and Bogg, 2004).

There is evidence that romantic relationships can impact the development of conscientiousness. For one, simply entering into a first romantic relationship can lead to increases in conscientiousness (Neyer and Asendorpf, 2001; Neyer and Lehnart, 2007). Relationship length may also have important implications for development. Evidence suggests that the longer women stay married, the more conscientious they become (Roberts and Bogg, 2004). Such increases may be a function of how conscientious behaviors are rewarded in the context of the relationship. Finally, divorce can impact conscientiousness, though evidence suggests effects differ depending on one's gender. One study found that conscientiousness increased after a divorce (Specht et al., 2011), while another study, consisting solely of women, found conscientiousness decreased after a divorce (Roberts and Bogg, 2004). Though further research is needed to determine the nature of these findings, one possible explanation is that divorced women may have fewer responsibilities because they no longer have to take care of their husbands, resulting in lower levels of reported conscientiousness, whereas men's responsibilities may increase after a divorce.

Criminality and externalizing behaviors, two correlates of conscientiousness, exhibit patterns of change over time that seem to complement the way conscientiousness develops (Roberts et al., 2009a). Criminality and externalizing behaviors are at their highest at ages when people are lowest in conscientiousness, and they decrease as conscientiousness increases. It is unclear whether changes in criminality and externalizing behaviors are associated with changes in conscientiousness, though some emerging studies suggest this is the case. For example, the development of conscientiousness is related to the development of drinking and drug use behaviors such that decreases in alcohol usage are related to increases in conscientiousness and impulse control (Littlefield et al., 2010). Similarly, decreases in tobacco consumption are related to increases in the facet of responsibility (Roberts and Bogg, 2004).

Conclusion

Conscientiousness is a hierarchical structure that, at its broadest level, describes individual differences in the tendency to be industrious, organized, responsible, and self-controlled, while, at its lowest level, describes specific behaviors related to these traits, such as keeping appointments and cleaning regularly. There are numerous ways to assess conscientiousness, and it is possible to do so at nearly every level of the hierarchy – from measures that look at conscientiousness as a whole to those that only examine behaviors. It is hoped that future researchers will use narrower levels of analysis to better understand the relationship between conscientiousness and life outcomes as well as to better understand how conscientiousness may development over the life span.

See also: Big Five Factor Model, Theory and Structure; Five Factor Model of Personality, Assessment of; Five Factor Model of Personality, Facets of; Personality Assessment: Overview; Personality and Educational Outcomes; Personality and Values at Work; Personality, Trait Models of.

Bibliography

Allemand, M., Zimprich, D., Hendriks, A.A.J., 2008. Age differences in five personality domains across the lifespan. Developmental Psychology 44, 758–770. http://dx.doi.org/10.1037/0012-1649.44.3.758.

Baker, L.R., McNulty, J.K., 2011. Self-compassion and relationship maintenance: the moderating roles of conscientiousness and gender. Journal of Personality and Social Psychology 100 (5), 853–873. http://dx.doi.org/10.1037/a0021884.

Bernerth, J.B., Taylor, S.G., Walker, H., Whitman, D.S., 2012. An empirical investigation of dispositional antecedents and performance-related outcomes of credit scores. Journal of Applied Psychology 97 (2), 469–478. http://dx.doi.org/10.1037/a0026055.

Bleidorn, W., 2012. Hitting the road to adulthood: short-term personality development during a major life transition. Personality and Social Psychology Bulletin 38 (12), 1594–1608. http://dx.doi.org/10.1177/0146167212456707.

Bogg, T., Roberts, B.W., 2004. Conscientiousness and health related behaviors: a meta-analysis of the leading behavioral contributors to mortality. Psychological Bulletin 130, 887–919. http://dx.doi.org/10.1037/0033-2909.130.6.887.

Bowling, N.A., Burns, G.N., 2010. A comparison of work-specific and general personality measures as predictors of work and non-work criteria. Personality and Individual Differences 49 (2), 95–101. http://dx.doi.org/10.1016/j.paid.2010.03.009.

Buss, D.M., 1991. Conflict in married couples: personality predictors of anger and upset. Journal of Personality 59 (4), 663–688. http://dx.doi.org/10.1111/j.1467-6494.1991.tb00926.x.

Buss, D.M., Shackelford, T.K., 1997. Susceptibility to infidelity in the first year of marriage. Journal of Research in Personality 31 (2), 193–221. http://dx.doi.org/10.1006/jrpe.1997.2175.

Caprara, G., Vecchione, M., Alessandri, G., Gerbino, M., Barbaranelli, C., 2011. The contribution of personality traits and self-efficacy beliefs to academic achievement: a longitudinal study. British Journal of Educational Psychology 81 (1), 78–96. http://dx.doi.org/10.1348/2044-8279.002004.

Caspi, A., Roberts, B.W., Shiner, R.L., 2005. Personality development: stability and change. Annual Review of Psychology 56, 453–484. http://dx.doi.org/10.1146/annurev.psych.55.090902.141913.

Cattell, R.B., 1957. Personality and Motivation Structure and Measurement. World Book, New York.

Chapman, B.P., 2007. Bandwidth and fidelity on the NEO-Five Factor Inventory: replicability and reliability of Saucier's (1998) item cluster subcomponents. Journal of Personality Assessment 88 (2), 220–234. http://dx.doi.org/10.1080/00223890701268082.

Clower, C.E., Bothwell, R.K., 2001. An explanatory study of the relationship between the Big Five and inmate recidivism. Journal of Research in Personality 35, 231–237. http://dx.doi.org/10.1006/jrpe.2000.2312.

Costa Jr., P.T., McCrae, R.R., 1992. NEO PI-R Professional Manual. Psychological Assessment Resources, Odessa, FL.

Costa, P.T., McCrae, R., Dye, D.A., 1991. Facet scales for agreeableness and conscientiousness: a revision of the NEO personality inventory. Personality and Individual Differences 12 (9), 887–898.

Credé, M., Harms, P., Niehorster, S., Gaye-Valentine, A., 2012. An evaluation of the consequences of using short measures of the Big Five personality traits. Journal of Personality and Social Psychology 102 (4), 874–888. http://dx.doi.org/10.1037/a0027403.

Denissen, J.A., Zarrett, N.R., Eccles, J.S., 2007. I like to do it, I'm able, and I know I am: longitudinal couplings between domain-specific achievement, self-concept, and interest. Child Development 78 (2), 430–447. http://dx.doi.org/10.1111/j.1467-8624.2007.01007.x.

DeYoung, C.G., 2011. Impulsivity as a personality trait. In: Vohs, K.D., Baumeister, R.F. (Eds.), Handbook of Self-regulation: Research, Theory, and Applications, second ed. Guilford Press, New York, NY, USA, pp. 485–502.

DeYoung, C.G., Quilty, L.C., Peterson, J.B., 2007. Between facets and domains: 10 aspects of the Big Five. Journal of Personality and Social Psychology 93 (5), 880–896. http://dx.doi.org/10.1037/0022-3514.93.5.880.

Digman, J.M., 1997. Higher-order factors of the Big Five. Journal of personality and social psychology 73 (6), 1246–1256.

Dyrenforth, P.S., Kashy, D.A., Donnellan, M.B., Lucas, R.E., 2010. Predicting relationship and life satisfaction from personality in nationally representative samples from three countries: the relative importance of actor, partner, and similarity effects. Journal of Personality and Social Psychology 99, 690–702.

Eysenck, H.J., 1947. Dimensions of Personality. Kegan Paul, Oxford England.

Friedman, H.S., Tucker, J.S., Tomlinson–Keasey, C., Schwartz, J.E., Wingard, D.L., Criqui, M.H., 1993. Does childhood personality predict longevity? Journal of Personality and Social Psychology 65, 176–185.

Friedman, H.S., Tucker, J.S., Schwartz, J.E., Tomlinson-Keasey, C., Martin, L.R., et al., 1995. Psychosocial and behavioral predictors of longevity. American Psychologist 50, 69–78.

Goldberg, L.R., 1993. The structure of phenotypic personality traits. American Psychologist 48 (1), 26.

Gosling, S.D., Rentfrow, P.J., Swann, W.B., 2003. A very brief measure of the Big-Five personality domains. Journal of Research in Personality 37 (6), 504–528. http://dx.doi.org/10.1016/S0092-6566(03)00046-1.

Gough, H.G., 1956. California Psychological Inventory. Consulting Psychologists Press, Palo Alto, CA; England.

Hampson, S.E., Andrews, J.A., Barckley, M., Lichtenstein, E., Lee, M.E., 2000. Conscientiousness, perceived risk and risk-reduction behaviors: a preliminary study. Health Psychology 19, 496–500.

Hill, P.L., Allemand, M., 2012. Explaining the link between conscientiousness and forgiveness. Journal of Research in Personality 46, 497–503. http://dx.doi.org/10.1016/m.mrp2012.05.007.

Hill, P.L., Allemand, M.A., Heffernan, M.A., 2013. Placing dispositional forgiveness within theories of adult personality development. European Psychologist 18, 273–285.

Hill, P.L., Nickel, L.B., Roberts, B.W., 2013. Are you in a healthy relationship?: linking conscientiousness to health via implementing and immunizing behaviours. Journal of Personality 82 (6), 485–492.

Hill, P.L., Roberts, B.W., 2011. The role of adherence in the relationship between conscientiousness and perceived health. Health Psychology 30 (6), 797–804. http://dx.doi.org/10.1037/a0023860.

Hofstee, W.K.B., De Raad, B., Goldberg, L.R., 1992. Integration of the Big Five and circumplex approaches to trait structure. Journal of Personality and Social Psychology 63, 146–163.

Hogan, R.R., 1992. Hogan personality inventory. Psychological Test Bulletin 5 (2), 130–136.

Hudson, N.W., Roberts, B.W., Lodi-Smith, J., 2012. Personality trait development and social investment in work. Journal of Research in Personality 46 (3), 334–344. http://dx.doi.org/10.1016/j.jrp.2012.03.002.

Ireland, J.L., Ireland, C.A., 2011. Personality structure among prisoners: how valid is the five-factor model, and can it offer support for Eysenck's theory of criminality. Criminality Behaviour and Mental Health 21, 35–50. http://dx.doi.org/10.1002/cbm.770.

Jackson, J.J., Bogg, T., Walton, K.E., Wood, D., Harms, P.D., Lodi-Smith, J., Edmonds, G.W., Roberts, B.W., 2009. Not all conscientiousness scales change alike: a multimethod, multisample study of age differences in the facets of conscientiousness. Journal of Personality and Social Psychology 96, 446–459. http://dx.doi.org/10.1037/a0014156.

Jackson, J.J., Wood, D., Bogg, T., Walton, K.E., Harms, P.D., Roberts, B.W., 2010. What do conscientious people do? Development and validation of the behavioral indicators of conscientiousness (BIC). Journal of Research in Personality 44 (4), 501–511. http://dx.doi.org/10.1016/j.jrp.2010.06.005.

Jackson, J.J. Connolly, J.J., Garrison, M, Levine, M., Connolly, S.L. Your friends know how long you will live: a 75 year study of peer-rated personality traits. Psychological Science, (in press).

John, O.P., Donahue, E.M., Kentle, R.L., 1991. The Big Five Inventory–Versions 4a and 54. University of California, Berkeley, Institute of Personality and Social Research, Berkeley, CA.

Johnson, J.A., 1994. Clarification of Factor Five with the help of the AB5C Model. European Journal of Personality 8, 311–334.

Jones, S.E., Miller, J.D., Lynam, D.R., 2011. Personality, antisocial behavior, and aggression: a meta-analytic review. Journal Of Criminal Justice 39 (4), 329–337. http://dx.doi.org/10.1016/j.jcrimjus.2011.03.004.

Lee, K., Ashton, M.C., 2008. The HEXACO personality factors in the indigenous personality lexicons of English and 11 other languages. Journal of Personality 76, 1001–1053.

Littlefield, A.K., Sher, K.J., Wood, P.K., 2010. A personality-based description of maturing out of alcohol problems: extension with a five-factor model and robustness to modeling challenges. Addictive Behaviors 35, 948–954. http://dx.doi.org/10.1016/j.addbeh.2010.06.008.

Lodi-Smith, J., 2007. Social investment and personality: a meta-analysis of the relationship of personality traits to investment in work, family, religion, and volunteerism. Personality and Social Psychology Review 11 (1), 68–86. http://dx.doi.org/10.1177/1088868306294590.

Lodi-Smith, J., Jackson, J., Bogg, T., Walton, K., Wood, D., Harms, P., Roberts, B.W., 2009. Mechanisms of health: educations and health-related behaviors partially mediate the relationship between conscientiousness and self-reported physical health. Psychology and Health, iFirst 1–15. http://dx.doi.org/10.1080/08870440902736964.

Lüdtke, O., Roberts, B.W., Trautwein, U., Nagy, G., 2011. A random walk down university avenue: life paths, life events, and personality trait change at the transition to university life. Journal of Personality and Social Psychology 101 (3), 620–637. http://dx.doi.org/10.1037/a0023743.

MacCann, C., Duckworth, A.L., Roberts, R.D., 2009. Empirical identification of the major facets of conscientiousness. Learning and Individual Differences 19 (4), 451–458. http://dx.doi.org/10.1016/j.lindif.2009.03.007.

Madhavan, P., 2004. Assessment of the psychometric properties of the facets of Conscientiousness. North American Journal of Psychology 6 (2), 309–326.

Markon, K.E., Krueger, R.F., Watson, D., 2005. Delineating the structure of normal and abnormal personality: an integrative hierarchical approach. Journal of Personality and Social Psychology 88 (1), 139–157. http://dx.doi.org/10.1037/0022-3514.88.1.139.

McCrae, R.R., Costa, P.T., Pedroso de Lima, M., Simões, a, Ostendorf, F., Angleitner, a, Piedmont, R.L., 1999. Age differences in personality across the adult life span: parallels in five cultures. Developmental Psychology 35 (2), 466–477.

Mike, A., Jackson, J.J., Oltmanns, T.F., 2014. The conscientious retiree: The relationship between conscientiousness, retirement, and volunteering. Journal of Research in Personality 52, 68–77. http://dx.doi.org/10.1016/j.jrp.2014.07.002.

Moffitt, T.E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R.J., Harrington, H.L., Caspi, A., 2011. A gradient of childhood self-control predicts health, wealth, and public safety. Proceedings of the National Academy of Sciences 108, 2693–2698.

Myrseth, H., Pallesen, S., Molde, H., Johnsen, B., Lorvik, I., 2009. Personality factors as predictors of pathological gambling. Personality and Individual Differences 47 (8), 933–937. http://dx.doi.org/10.1016/j.paid.2009.07.018.

Ng, T.H., Eby, L.T., Sorensen, K.L., Feldman, D.C., 2005. Predictors of objective and subjective career success. A meta-analysis. Personnel Psychology 58 (2), 367–408. http://dx.doi.org/10.1111/j.1744-6570.2005.00515.x.

Neyer, F.J., Asendorpf, J.B., 2001. Personality-relationship transaction in young adulthood. Journal of Personality and Social Psychology 81 (6), 1190–1204. http://dx.doi.org/10.1037//0022-3514.81.6.1190.

Neyer, F.J., Lehnart, J., 2007. Relationships matter in personality development. Evidence from an 8-year longitudinal study across young adulthood. Journal of Personality 75, 535–568.

Noftle, E.E., Robins, R.W., 2007. Personality predictors of academic outcomes: Big Five correlates of GPA and SAT scores. Journal of Personality and Social Psychology 93 (1), 116–130. http://dx.doi.org/10.1037/0022-3514.93.1.116.

Paunonen, S.V., 1998. Hierarchical organization of personality and prediction of behavior. Journal of Personality and Social Psychology 74 (2), 538–556. http://dx.doi.org/10.1037//0022-3514.74.2.538.

Paunonen, S.V., Ashton, M.C., 2001. Big five predictors of academic achievement. Journal of Research in Personality 35 (1), 78–90. http://dx.doi.org/10.1006/jrpe.2000.2309.

Peabody, D., De Raad, B., 2002. The substantive nature of psycholexical personality factors: a comparison across languages. Journal of Personality and Social Psychology 83 (4), 983–997. http://dx.doi.org/10.1037//0022-3514.83.4.983.

Richardson, M., Abraham, C., 2009. Conscientiousness and achievement motivation predict performance. European Journal of Personality 23 (7), 589–605. http://dx.doi.org/10.1002/per.732.

Roberts, B.W., Bogg, T., 2004. A longitudinal study of the relationships between conscientiousness and the social-environmental factors and substance-use behaviors that influence health. Journal of Personality 72 (2), 325–354.

Roberts, B.W., DelVecchio, W.F., 2000. The rank-order consistency of personality traits from childhood to old age: a quantitative review of longitudinal studies. Psychological Bulletin 126 (1), 3–25. http://dx.doi.org/10.1037//0033-2909.126.1.3.

Roberts, B.W., Bogg, T., Walton, K.E., Chernyshenko, O.S., Stark, S.E., 2004. A lexical investigation of the lower-order structure of conscientiousness. Journal of Research in Personality 38 (2), 164–178. http://dx.doi.org/10.1016/S0092-6566(03)00065-5.

Roberts, B.W., Caspi, A., Moffitt, T.E., 2003a. Work experiences and personality development in young adulthood. Journal of Personality and Social Psychology 84 (3), 582–593. http://dx.doi.org/10.1037/0022-3514.84.3.582.

Roberts, B.W., Robins, R.W., Caspi, A., Trzesniewski, K., 2003b. Personality trait development in adulthood. In: Mortimer, J., Shanahan, M. (Eds.), Handbook of the Life Course. Kluwer Academic, New York, NY, pp. 579–598.

Roberts, B.W., Chernyshenko, O.S., Stark, S., Goldberg, L.R., 2005a. The structure of conscientiousness: an empirical investigation based on seven major personality questionnaires. Personnel Psychology 58 (1), 103–139. http://dx.doi.org/10.1111/j.1744-6570.2005.00301.x.

Roberts, B.W., Walton, K.E., Bogg, T., 2005b. Conscientiousness and health across the life course. Review of General Psychology 9 (2), 156–168. http://dx.doi.org/10.1037/1089-2680.9.2.156.

Roberts, B.W., Jackson, J.J., Burger, J., Trautwein, U., 2009a. Conscientiousness and externalizing psychopathology: overlap, developmental patterns, and etiology of two related constructs. Development and Psychopathology 21 (3), 871–888. http://dx.doi.org/10.1017/S0954579409000479.

Roberts, B.W., Jackson, J.J., Fayard, J.V., Edmonds, G., Meints, J., 2009b. Conscientiousness. In: Leary, M.R., Hoyle, R.H. (Eds.), Handbook of Individual Differences in Social Behavior. Guilford Press, New York, NY, USA, pp. 369–381.

Roberts, B.W., Smith, J., Jackson, J.J., Edmonds, G., 2009c. Compensatory conscientiousness and health in older couples. Psychological Science 20, 6–10.

Roberts, B.W., Harms, P.D., Caspi, A., Moffitt, T.E., 2007a. Predicting the counterproductive employee in a child-to-adult prospective study. The Journal of Applied Psychology 92 (5), 1427–1436. http://dx.doi.org/10.1037/0021-9010.92.5.1427.

Roberts, B.W., Kuncel, N.R., Shiner, R., Caspi, A., Goldberg, L.R., 2007b. The power of personality: the comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. Perspectives on Psychological Science 2, 313–345.

Roberts, B.W., Lejuez, C., Krueger, R.F., Richards, J.M., Hill, P.L., 2012. What is conscientiousness and how can it be assessed? Developmental Psychology. http://dx.doi.org/10.1037/a0031109.

Roberts, B.W., Walton, K., Bogg, T., Caspi, A., August 2006a. De-investment in work and non-normative personality trait change in young adulthood. European Journal of Personality 474, 461–474. http://dx.doi.org/10.1002/.

Roberts, B.W., Walton, K., Viechtbauer, W., 2006b. Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. Psychological Bulletin 132, 3–27.

Roberts, B.W., Wood, D., Caspi, A., 2008. The development of personality traits in adulthood. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research, third ed. Guilford Press, New York, NY, pp. 375–398.

Saucier, G., 1998. Replicable item-cluster subcomponents in the NEO Five-Factor Inventory. Journal of Personality Assessment 70 (2), 263–276.

Saucier, G., Goldberg, L.R., 1996. Evidence for the Big Five in analyses of familiar English personality adjectives. European Journal of Personality 10, 61–77.

Saucier, G., Ostendorf, F., 1999. Hierarchical subcomponents of the Big Five personality factors: a cross-language replication. Journal of Personality and Social Psychology 76 (4), 613–627. Retrieved from. http://www.ncbi.nlm.nih.gov/pubmed/10234848.

Solomon, B.C., Jackson, J.J., 2014a. The long reach of one's spouse: spouses' personality influences occupational success. Psychological Science. http://dx.doi.org/10.1177/0956797614551370.

Solomon, B., Jackson, J.J., 2014b. Why do personality traits predict divorce? Multiple pathways through satisfaction. Journal of Personality and Social Psychology 106, 978–996.

Soto, C.J., John, O.P., 2009. Ten facet scales for the Big Five Inventory: convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. Journal of Research in Personality 43 (1), 84–90. http://dx.doi.org/10.1016/j.jrp.2008.10.002.

Soto, C.J., John, O.P., 2012. Development of Big Five domains and facets in adulthood: mean-level age trends and broadly versus narrowly acting mechanisms. Journal of personality 80 (4), 881–914. http://dx.doi.org/10.1111/j.1467-6494.2011.00752.x.

Soto, C.J., John, O.P., Gosling, S.D., Potter, J., 2011. Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. Journal of Personality and Social Psychology 100 (2), 330–348. http://dx.doi.org/10.1037/a0021717.

Specht, J., Egloff, B., Schmukle, S.C., 2011. Stability and change of personality across the life course: the impact of age and major life events on mean-level and rank-order stability of the Big Five. Journal of Personality and Social Psychology 101 (4), 862–882. http://dx.doi.org/10.1037/a0024950.

Spurk, D., Abele, A.E., 2011. The career satisfaction scale: longitudinal measurement invariance and latent growth analysis. Journal of Occupational and Organizational Psychology 84 (2), 315–326. http://dx.doi.org/10.1111/j.2044-8325.2011.02028.x.

Sutin, A.R., Costa, P.R., 2010. Reciprocal influences of personality and job characteristics across middle adulthood. Journal of Personality 78 (1), 257–288. http://dx.doi.org/10.1111/j.1467-6494.2009.00615.x.

Sutin, A.R., Ferrucci, L., Zonderman, A.B., Terracciano, A., 2011. Personality and obesity across the adult life span. Journal of Personality and Social Psychology 101, 579–592. http://dx.doi.org/10.1037/a0024286.

Takahashi, Y., Edmonds, G.W., Jackson, J.J., Roberts, B.W., 2013. Longitudinal correlated changes in conscientiousness, preventative health-related behaviors, and self-perceived physical health. Journal of Personality 81 (4), 417–427. http://dx.doi.org/10.1111/jopy.12007.

Tellegen, A., Waller, N.G., 2008. Exploring personality through test construction: development of the multidimensional personality questionnaire. In: Boyle, G.J., Matthews, G., Saklofske, D.H. (Eds.), Handbook of Personality Theory and Testing, Personality Measurement and Assessment, vol. II. JAI Press, Greenwich, CT, pp. 261–292.

Terracciano, A., Iancono, D., O'Brien, R., Troncoso, J.C., An, Y., Sutin, A.R., Ferrucci, L., Zonderman, A.B., Resnick, S.M., 2013. Personality and resilience to Alzheimer's disease neuropathology: a prospective autopsy study. Neurobiology of Aging 34, 1045–1050.

Terracciano, A., McCrae, R.R., Brant, L.J., Costa, P.T., 2005. Hierarchical linear modeling analyses of the NEO-PI-R scales in the Baltimore Longitudinal Study of Aging. Psychology and Aging 20 (3), 493–506. http://dx.doi.org/10.1037/0882-7974.20.3.493.

Webley, P., Nyhus, E.K., 2006. Parents' influence on children's future orientation and saving. Journal of Economic Psychology 27, 140–164. http://dx.doi.org/10.1016/j.joep.2005.06016.

Weston, S.J., Hill, P.L., Jackson, J.J., 2014. Personality traits predict the onset of disease. Social Psychological and Personality Science.

Weston, S., Jackson, J.J. (in press). Identification of the Healthy Neurotic: Personality traits predict smoking after disease onset. Journal of Research in Personality.

Wiebe, R.P., 2004. Delinquent behavior and the five-factor model: hiding in the adaptive landscape? Individual Differences Research 2, 38–62.

Wortman, J., Lucas, R.E., Donnellan, M.B., 2012. Stability and change in the Big Five personality domains: evidence from a longitudinal study of Australians. Psychology and Aging 27, 867–874. http://dx.doi.org/10.1037/a0029322.

Conscious Control during Childhood, Development of

Vladimir A Fomichov, National Research University Higher School of Economics, Moscow, Russia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reconstructs the panorama of studies on the development of conscious control during childhood that were carried out in the 1990s and 2000s. The principal attention is paid to the main stream in this direction: the achievements of the positive psychology movement (i.e., the branch of psychology aimed at supporting and developing the positive qualities of the person). With this aim, both broadly known and much less known but deep studies carried out in adjacent fields are analyzed.

The goal of studying the development of consciousness in childhood was posed as far back as in the end of the nineteenth century by J.M. Baldwin and in the first third of the twentieth century by L. Vygotsky. A large-scale research in this field was started only in the 1990s. However, by the beginning of the 2000s one was able to find in the scientific literature rather contradictory points of view concerning this problem.

The Birth of a New Paradigm in Psychology: The Enhancement of Positive Qualities

It seems that the best angle of look at the problem of development of conscious control (CC) during childhood can be found in the context of the main stream in modern psychology – the positive psychology movement. The reasons for this idea are as follows.

During the 1990s, it was possible to observe the steady growth of the number of children at school-age in the developed countries encountering various social, emotional, and behavioral problems. Numerous observations provide the possibility to conjecture that, to a large extent, it was a consequence of more intensive interaction with computers at lessons and at home and of the stormy Internet's expansion. Besides, the criminal films and horror films continued to negatively influence the mental state of very many children and adolescents, in particular, causing anxiety and aggression. These negative shifts became sufficiently noticeable by the beginning of the 2000s. According to the US Public Health Service (2000), approximately one-fifth of children and adolescents experienced problems showing their need for mental health services.

As a way out, many psychologists indicated the importance of promoting children's social and emotional experience in schools. As a consequence, a new paradigmatic shift was observed in psychology: a shift from the accent on repairing weakness to enhancing positive qualities and preventing the problems before they arise (Seligman and Csikszentimihalyi, 2000). As a result, the positive psychology movement appeared; the principal objective of this movement is studying the positive features of human development, in particular investigating such significant traits of the person as "subjective well-being, optimism, happiness, and self-determination" (Seligman and Csikszentimihalyi, 2000: p. 9).

As a logical consequence, the task of promoting positive emotions in children and adolescents was posed (Huebner and Gilman, 2003). The evidence obtained in the 2000s shows that a critical role in the success of children in school and in their social and emotional competence is played by self-regulation, in particular by controlling attention and inhibiting aggressive reactions.

The publications on positive psychology allow to distinguish the next factor that is beneficial to well-being; this factor is called 'mindfulness' (Kabat-Sinn, 2003). According to the definition given by Schonert-Reichl and Lawlor (2010), it is a way of directing attention. Generalizing a number of available definitions of this concept, mindfulness can be characterized as the ability to maximally proceed from the context while taking decisions in any situations. It is the ability to pay attention to many details while elaborating a decision, but not only 'mechanically' (e.g., by following a number of prescribed rules).

This explains why the ideas of the positive psychology movement have turned the attention of many researchers in the world to the models of development of CC during childhood.

The Levels of Consciousness Basic Model

It seems that the model proposed by Zelazo (2004) can be considered as a good working instrument for studying the development of CC during the first through fourth years of childhood. This model, called the levels of consciousness (LOC) model, emerged as a result of reflecting the experimentally discovered regularities of the development of CC of thought, action, and emotion. The model describes four transitions from one LOC to another, higher LOC; these transitions depend on age. Regarding the zero LOC, in the case of newborn babies and very young children aged up to 11–12 months, Zelazo (2004) characterizes the consciousness of this period as minimal consciousness; it is responsible for approach and avoidance behavior based on pleasure and pain and is present oriented, is unreflective, and doesn't operate with the self-concept.

The principal distinguished feature of LOC1 is the emergence of concepts and of the connections between the perceived objects and concepts (playing the role of labels of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23082-1

experienced objects). LOC1 is called by Zelazo (2004) the level of recursive consciousness. LOC2 emerges at the end of the second year; the essence of the transition from LOC1 to LOC2 consists in the emergence of symbolic thinking, in children's awareness of self. The signs of LOC2 are the first use of personal pronouns by children and their self-recognition in mirrors. Besides, children at this level feel the first self-conscious emotions, first of all shame.

LOC3 is called by Zelazo (2004) reflective consciousness 1; usually, this level characterizes the consciousness of 3-year-olds. The manifestation of this level is the ability of children to systematically use a pair of arbitrary rules (e.g., regarding objects of big size and of small size) for sorting pictures representing objects. However, the executive function of 3-year-olds is still limited, as was shown by experiments with the Dimensional Change Card Sort. For being successful in this game, children must integrate two incomparable pairs of rules into a single structure. This ability characterizes the LOC4, called by Zelazo (2004) reflective consciousness 2. Usually, LOC4 emerges by the end of the fourth year; this level is also characterized by a spectrum of metacognitive skills.

Expansion of the Levels of Consciousness Basic Model by the Theory of Dynamic Conceptual Mappings

The Necessity of Early Children's Socialization in Information Society

A number of phenomena caused by information and communication technologies (ICT) can be interpreted as negative shifts in the development of the personality in information society. A considerable part of information distributed via the Internet is false, but this false information is 'injected' into the net by concrete people. Very many high school and university students don't feel the value of a thought generated by another person and, as a consequence, use in their works the ideas, methods, and models belonging to other people. Acting in this way is very easy, because even by middle grades students possess well-developed skills in using e-mail, the Internet, and cell telephones. However, they are not socially mature and are not ready to suffer all the consequences of their deeds.

A new, tragic phenomenon of information society is cyber-bullying. Cell telephones, profile home pages, and social networking services are breeding grounds for cyber-bullying in some countries. Other negative deeds are attacks of young hackers against computer systems of socially very significant objects and military objects.

The common reason for all considered negative deeds is that children possess very high skills in using ICT but are very far from being socially mature and, as a consequence, far from being aware of their own social responsibility. Young children and adolescents usually can't clearly understand the severity of their actions when they are many times multiplied by the power of modern ICT.

That is why Fomichov and Fomichova (2012) believe that a possible way out is to elaborate methods of much earlier socialization of the child than are usually done in order, on one hand, to eliminate or considerably diminish children's aggressiveness, and, on the other hand, to contribute to the birth in children of feeling social responsibility and of understanding the severe consequences of their actions suffered by their peers and adults.

Shortly about Three Additional Levels of Consciousness Development

It seems in this connection that the broadly felt necessity of promoting children's emotional and social competence in schools and the lack in the scientific literature of rather simple solutions to this problem (see Schonert-Reichl and Lawlor, 2010) are the grounds for putting forward the following conjecture: the LOC model proposed by Zelazo (2004) indicates only some basic stages of consciousness development. The goal of creating appropriate theoretical foundations for promoting children's emotional and social competence will lead to discovering additional, higher stages of children's consciousness development corresponding to mature emotional and social competence of the child.

Realizing this idea, let's give a new interpretation of the methods of developing CC of thought, action, and emotion described in Fomichov and Fomichova (1997, 1998, 2006, 2012), Fomichova (2009), and Fomichova and Fomichov (2009). We'll suppose that these methods underpin the transition from LOC4 to LOC5, from LOC5 to LOC6, and from LOC6 to LOC7. The new levels LOC5, LOC6, and LOC7 will be called, respectively, the level of broad beauty appreciation, the level of appreciating the value of thought, and the level of enhanced awareness of social agreements and social responsibility.

A very short, preliminary description of these levels is as follows. Reaching LOC5 by the person means that this person possesses a well-developed feeling of beauty in various manifestations: the beauty of a thing, of an idea, of an expression, of a picture or sculpture, of interpersonal relationships, and so on (Fomichov and Fomichova, 1997, 1998; Fomichova, 2009; Fomichova and Fomichov, 2009). The successful transition from LOC5 to LOC6 means that (1) a child is aware of the fact that his or her ideas may be socially significant (i.e., the child may be appraised by peers or adults for the originality and beauty of his or her ideas), and (2) a child appreciates the value of the thoughts of other persons (Fomichov and Fomichova, 2006; Fomichova and Fomichov, 2000). Reaching LOC7 by a person means that this person is sufficiently mature in the social sense (i.e., possesses an enhanced awareness of social agreements and social responsibility) (Fomichov and Fomichova, 2012).

It should be underlined that modern preschool and school educational systems in various countries encourage only a rather small proportion of children to reach the fifth, sixth, and seventh levels of CC. However, to considerably increase this proportion is vitally important for the successful socialization of children in information society. Happily, at least one broadly applicable way of solving this problem has been available since the 1990s; it is outlined in this article.

About an Informational-Aesthetic Conception of Developing a Cognitive-Emotional Sphere in Learners

The methods underpinning the mindfulness-based educational interventions allowing for reaching the fifth, sixth, and seventh LOC development form an original approach to developing positive qualities in children and to early socialization of children. This approach is provided by the System of the Methods of Emotional-Imaginative Teaching (EIT-system). The core of the EIT-system was elaborated by O.S. Fomichova in the first half of the 1990s and has been expanded in the second half of the 1990s and in the 2000s. This system is underpinned by the Theory of Dynamic Conceptual Mappings (the DCM-theory), which was developed in the 1990s and 2000s by V.A. Fomichov and O.S. Fomichova. This theory is stated in numerous publications in both English and Russian, in particular in Fomichov and Fomichova (1997, 1998, 2006, 2012). The main component of the DCM-theory is an original informational-aesthetic conception of developing the cognitive-emotional sphere (in particular, emotional intelligence) of the learners: young children, adolescents, and university students. The central ideas of this conception are as follows:

1. It is important to actively develop a broad spectrum of information-processing skills of the child, starting at least at the age of 5. It applies, in particular, to associative abilities, the skill of integrating information from various sources, and the ability to establish time-causal relationships between events.
2. It is very important to combine the development of information-processing skills with inscribing, in a systemic way, the feeling of beauty into the world conceptual picture of the child.
3. Passing ahead the development of soul in comparison with the development of reasoning skills. A well-developed feeling of beauty plays an especially significant role in the realization of this idea. Besides, it is very important to be aware of the fact that children should have enough time for the development of soul: the time for contemplation, for imbibing the beauty of nature, and so on (i.e., children should have time for self-paced activity).
4. The principal cognitive precondition of successfully (as concerns a long-term perspective) acquainting children with computers is the realization of the thought-producing self of the child. It is a new concept expanding the collection of self-concepts considered in psychology. This concept means that the child should know that his or her thoughts may have a high social significance, that is, be appreciated by his or her peers, parents, grandparents, teachers, and so on (see Fomichova and Fomichov, 2000; Fomichov and Fomichova, 2006). The child should be aware of this fact before the time when adults start to systematically acquaint him or her with computers.
5. Due to mastering modern ICT (cell telephones, the internet, etc.), the consequences of children's negative actions may be very severe. That is why it is necessary to find ways of socializing children in the modern information society much earlier in order to eliminate or considerably diminish their aggressiveness and to contribute to the realization by children of the real scale of their misuse of ICT.

The Maps of Cognitive Transformations

Let us use a new graph notation allow to reflect the initial and achieved states of the cognitive-emotional sphere of the learner (Fomichov and Fomichova, 2012). This notation will be used in this article for describing the new methods of developing CC.

By definition, *a map of cognitive transformations* (MCT) is an oriented graph with the vertices of three classes (or types). The *A-vertices* are represented by the rectangles (or blocks) with the *single contour*; the texts inside these blocks describe the theoretical materials underpinning the teaching methods. The rectangles corresponding to *B-vertices* have the *double contour*; the text inside these rectangles describes the activity at a lesson (or lessons) either of a teacher or of the students. The *C-vertices* are represented *by the ovals*; the text inside these ovals describes the initial or achieved state of the cognitive-emotional sphere of the learner or the state of the spectrum of the learner's cognitive skills. **Figures 1–3** commented on in this article are examples of MCT.

The Level of Broad Beauty Appreciation

Initial Steps of Developing Conscious Control of Thought and Emotion

The EIT-system proposes a complex series of educational interventions aimed at reaching the consciousness level of broad beauty appreciation. An MCT realized at this stage is presented in **Figure 1**.

One of the distinguishing features of the approach to this problem is that it is realized in lessons for a foreign language (FL) – English – when the mother tongue of the children is Russian. The use of original, specially invented analogies (being the parts of fairy tales and thrilling stories) for teaching the English alphabet, the rules of reading, and the basic rules of English grammar contributes to developing associative abilities of children at the ages of 5–6.

Example. A difficult problem is to explain to very young children why the verbs in the third person of past simple tense have no 's' ending , but the same verbs in the third person of present simple tense do have such an ending ('reads' but 'read,' etc.). An interesting story from one of the previous lessons associates the 's' ending with a bow. The teacher explains that her young students were, in the past, babies and had no hair (were bald). Hence, it was impossible to tie a bow. That is why the verbs have no 's' ending in the third person of past simple tense. Five-year-old students accept this explanation with great joy and remember it very well. As a result of having heard stories of the kind, children become aware of the fact that symbolic objects have meanings pertaining to real or fairy-tale life.

These interesting stories about the life of verbs and other words establish in the consciousness of the young child

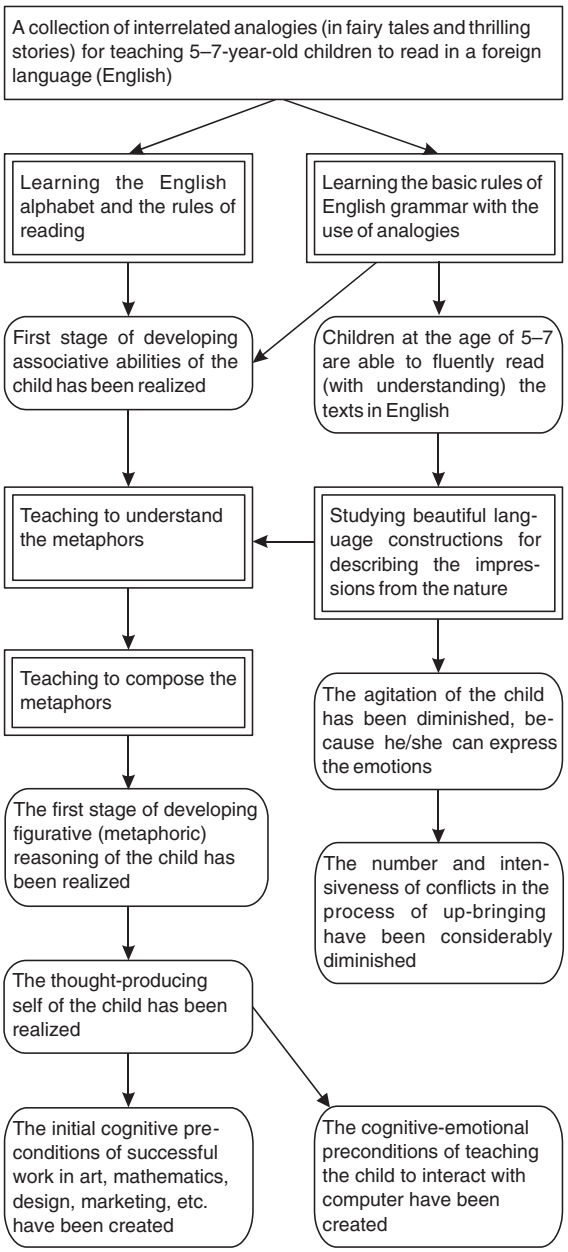

Figure 1 The map of cognitive transformations corresponding to the basic stage of developing creativity and broad beauty appreciation.

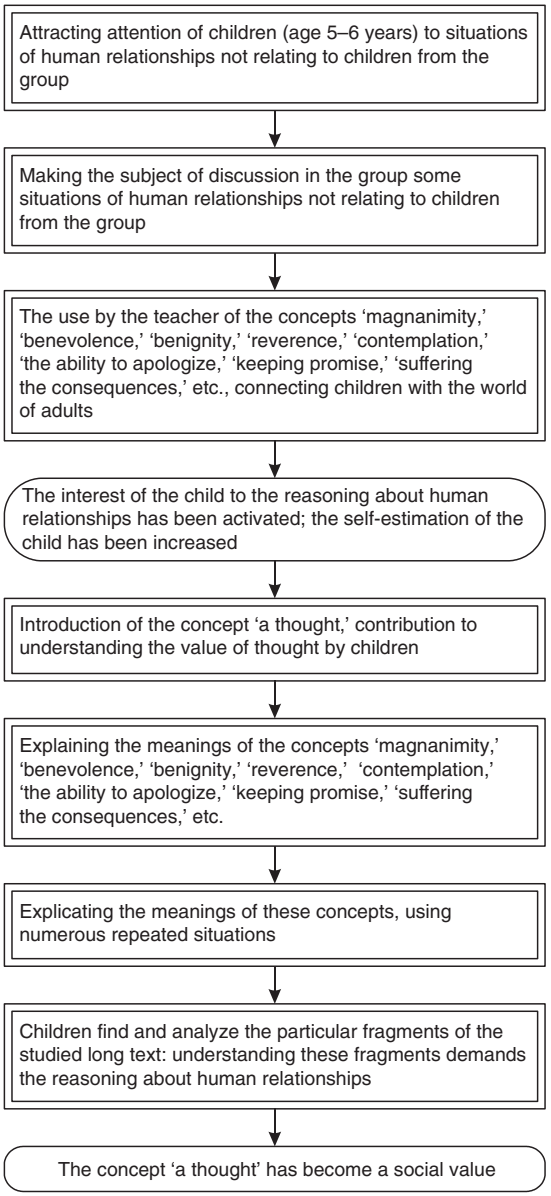

Figure 2 The map of cognitive transformations realized for achieving the level of appreciating the value of thought.

a mapping from the objects and situations of real life to the domain of language entities (verbs, nouns, pronouns, etc.). That is why the consciousness of the young child receives a considerable impulse to develop the ability to establish diverse analogies.

The next steps are as follows (see also Fomichova and Fomichov, 2009):

- reading and discussing complex texts in English as a foreign language (FL) at the ages of 5–6;
- mastering a rich sublanguage of FL for expressing the beauty of nature and the feelings evoked by nature;
- developing figurative reasoning + developing awareness of the social role of natural language;
- understanding poetical metaphors;
- creating metaphors;
- understanding the symbolic language of painting;
- developing the ability to decode the messages conveyed by the masterpieces;
- realizing the "Thought-Producing Self" (Fomichova and Fomichov, 2000) and the improvement of the feeling that a person is a link in the long chain of previous and future generations.

A Discovery: The Consciousness of 5–6-Year-Olds Needs a Very Beautiful Language

The principal reason for using the lessons of FL for positively developing CC is that (as a 23-year-long experience has shown)

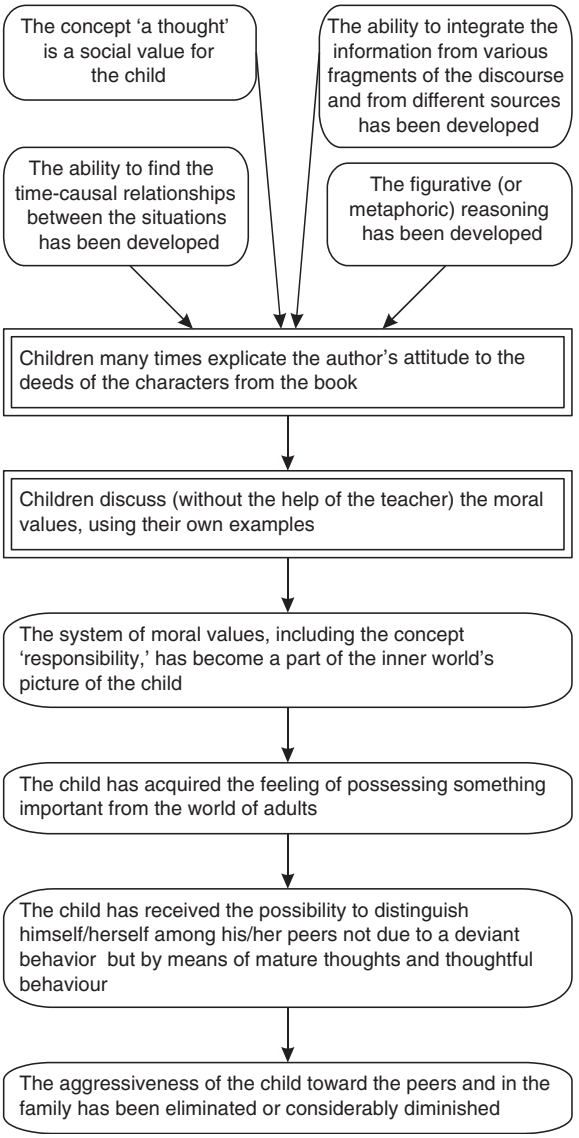

Figure 3 The map of cognitive transformations realized for achieving the level of enhanced awareness of social agreements and social responsibility.

young children learn beautiful language constructions for describing their impressions from nature easier in FL than with equivalent constructions in their mother tongue (Fomichov and Fomichova, 1997, 1998). The explanation of this phenomenon is that in this case, children don't feel any contradiction with the everyday use of language.

Proceeding from a 7-year experience of teaching more than 300 young children, Fomichov and Fomichova (1997, 1998) formulated the following conclusions. The deepness, brightness, and amplitude of the emotions of average 4–9-year-old children (i.e., children of usual, average abilities) evoked by pictures of nature are much stronger and higher than is generally accepted or believed in modern science and education. As a consequence, the consciousness of the 4–9-year-old child is open to mastering a very rich, very beautiful sublanguage of natural language enabling the child to adequately describe pictures of nature and the emotions evoked by such pictures. Average 5–9-year-old children master very easily, and with great joy, very rich language means for describing pictures of nature, landscapes, and seascapes. The reason for this ease and joy is that the learning of such rich language means deeply corresponds to the demands of the child's consciousness at that very important period of his or her development. These conclusions were confirmed by the experience of using this approach in the 2000s.

Example. Let's consider a fragment from the home composition "The Winter Day," which was written in English by a 6-year-old Russian student, Kseniya, who was in the second year of studies in the experimental groups (Fomichov and Fomichova, 2012: p. 124):

> In the picture I see a winter day. On the branches of fir-trees, pine-trees, and birches lies fluffy, white snow, glimmering in sunshine. It seems that snowdrops are covered with jewels. Near the wood there are fields with snow. On the edge of the snowfield the rill is dreaming. The snow is everywhere. Sunrays make one way through the grey, big, heavy clouds and run over top-trees. Pine-trees and fir-trees shine hoary green. The bare bushes of birches are covered with snow. It seems that the oak is with soft, white, and big leaves.

> Suddenly someone in heaven has dashed a big cup of sunlight upon the Earth, and the big old oak has turned into a fairy King in orange magnificent gown. And around him young birches in nice gowns are accompanying their beloved King.

The Level of Appreciating the Value of Thought

A system of values underpins a human being's conceptual picture of the world. Regarding thought as a value causes a more careful attitude toward language, because it is a tool for thinking; and arouses interest toward metaphors, because the construction of metaphors creates new images, makes the world brighter, and introduces beauty into it. One of the principal objectives of educational interventions at this stage is to enable the child to really hear one's partner in communication, because it is a significant precondition of successful socialization. Figure 2 illustrates the basic steps of a method proposed by the EIT-system.

The Level of Enhanced Awareness of Social Agreements and Social Responsibility

The method of reaching the seventh level of consciousness development proposed by Fomichov and Fomichova (2012) proceeds from the central idea of Searle (1995) about the natural language as the primary means of constructing social reality, and it considerably expands and works out in detail this idea.

The Steps of the Method

Taking into account the existence of the negative phenomena discussed in this article, the following new objectives for

influencing the development of the personality in the modern information society can be formulated. It is necessary to start earlier than is traditionally done to acquaint children with the very complex system of social agreements. Since this system is based on numerous symbols, scholars need to pay more attention to developing the symbolic information-processing skills of young children and teenagers. In addition, it is necessary to acquaint children early with the idea that different people may have considerably different inner-world pictures (i.e., conceptual systems), and it is very important to take this into account while interacting with people. It is important to explain to young children and teenagers that practically every person has various connections with many other people. That is why the suffering of a classmate or the like, in fact, causes the suffering of many other people: mother, father, brothers, sisters, grandmother, grandfather, and so on.

Figure 3 represents the basic components of this method. The central idea of the method is as follows. Rather often, a child tries to distinguish himself or herself among his or her peers by means of emulating a bad pattern of adult behavior: smoking, aggressiveness, following the formula of 'might is right,' and so on. This applies not only to teenagers but also to children at the ages of 10–11. It is important to underline that the bad patterns of adult behavior (drinking, etc.) most often are the consequence of misfortune or despair. Normally, children have no despair; they simply emulate the adults.

As a result, however, of employing the stated method at lessons under the framework of the EIT-system, the child acquires (by the end of the second year of studies, at the age of 6–7 years) the possibility to distinguish himself or herself not by means of a deviant behavior but with the help of mature thoughts and thoughtful behavior.

A New Way of Supporting a Balance of Conscious Control and Creative Imagination

Carlson (2010), in particular, formulates a paradox of CC: the enhancement of CC causes restrictions for the imagination and creativity of the learners. According to this logic, the approach to early socialization of children described here must diminish the creativity of the learners.

Happily, this is not true as concerns the EIT-system. This educational program manages to maintain a good balance of enhanced CC and creativity due to an original way of teaching young children to appreciate the value of thought and to realize their Thought-Producing Self. As was described in this article, it is the principal feature of LOC6, an advanced level of consciousness development. Due to this approach, children follow social agreements in standard situations of communication but are highly creative while solving posed tasks, in particular while inventing metaphors.

The System of Emotional-Imaginative Teaching as an Original and Well-Tested Approach to Constructing Mindfulness-Based Educational Programs

The Goals and Composition of the System

Retrospectively, it is clear that the first mindfulness-based educational program was created approximately 10 years before the birth of the term 'mindfulness.' This program was provided by the EIT-system. As was mentioned in this article, the core of the EIT-system was elaborated by O.S. Fomichova in the first half of the 1990s and was expanded by O.S. Fomichova and V.A. Fomichov in the second half of the 1990s and in the 2000s.

The principal goal of the EIT-system is to develop in young children and teenagers:

- skills in processing symbolic information and in reasoning abilities;
- a mature, rich, cognitive-emotional sphere, and the ability to perceive and appreciate beauty in all of its manifestations, in particular in the deeds and relationships of people;
- the ability to understand the peculiarities of the world conceptual picture of one's communication partner;
- an understanding of the complex system of social agreements;
- the skill of grounding one's own point of view and of participating in a dialogue;
- the feeling of belonging to a very long chain of previous and future generations as the principal cognitive precondition of sustainable development; and
- the state of being open to the achievements of both national and world culture.

For achieving these listed objectives, a collection of the interrelated educational methods and an original cross-disciplinary educational program have been developed by O.S. Fomichova. The program is intended for teaching children for 12 years, where the starting age is 5–6 years. The program has been personally tested in Moscow with great success by O.S. Fomichova over a period of 23 years. The total number of successfully taught students exceeds 900. The program includes the following series of lessons: (1) a 2-year course (the ages of the learners are 5–7 or 6–8 years) of studying the foundations of reading and speaking English as an FL, including learning basic elements of English grammar (present simple and past simple tenses); (2) a course on understanding the language (a part of FL) of describing nature and feelings evoked by nature; (3) a course on understanding the symbolic language of painting; (4) a course on understanding the language of poetry (with the accent on understanding metaphors and descriptions of nature); (5) a course aimed at first acquaintance with sciences and developing the abilities to argue one's own opinion, raise objections, and so on; (6) a course on improving knowledge of English grammar (mainly during the fifth year of studies); and (7) a course entitled "Foundations of Secure Living in Information Society." In fact, the lessons of courses (2) to (7) may interchange.

The developed program can be interpreted as an original model of a system of disciplines traditionally learned in school (elementary, middle, and high) and in the courses forming the humanitarian component of a university education.

A Broad Applicability of the EIT-System

The EIT-system has been mainly realized in lessons of English as an FL for Russian-speaking children, in lessons of poetry and literature in English, and in lessons devoted to explaining the symbolic language of painting, the culture of communication,

and the symbolic language of classical dance. These kinds of lessons are considered in numerous countries as highly appropriate for young children and teenagers. A carefully selected collection of texts used at lessons is provided by a number of classical, world-renowned fairy tales and novels, in particular 'Snow White,' 'Cinderella,' 'Sleeping Beauty,' 'Pinocchio,' 'Pollyanna,' 'The Life and Adventures of Santa Claus' by L. Frank Baum, 'Alice in Wonderland' by Lewis Carroll, 'The Wind in the Willows' by Kenneth Grahame, 'The Hundred and One Dalmatians' by Dodie Smith, and others. That is why the EIT-system may be used (after a certain adaptation requiring a small time) in English-speaking countries and in numerous countries where English is learned as a second language.

See also: Concept Formation in Childhood: Cross-Cultural Perspectives; First Language Acquisition, Developmental Psychology of; Life Management, Developmental Psychology of; Logical and Hypothetical Reasoning in Adolescence, Development of; Motivational Development, Systems Theory of; Personality Changes During Adolescence Across Cultures; Scientific Concepts During Childhood, Development of; Self-Development in Childhood and Adolescence; Self-Regulation During Adolescence: Variations Associated with Individual–Context Relations; Social Competence During Adolescence Across Cultures; Values Across Cultures, Development of.

Bibliography

Carlson, S.M., 2010. Development of conscious control and imagination. In: Baumeister, R.F., Mele, A.R., Vohs, K.D. (Eds.), Free Will and Consciousness: How Might They Work? Oxford University Press, Oxford, pp. 135–151.

Fomichov, V.A., Fomichova, O.S., 1997. An informational conception of developing the consciousness of the child. Informatica. An International Journal of Computing and Informatics (Slovenia) 21, 371–390.

Fomichov, V.A., Fomichova, O.S., 1998. A new theoretical and practical approach to early positive developing child's consciousness. In: Trappl, R. (Ed.), Cybernetics and Systems'98. Proceedings of the 14th European Meeting on Cybernetics and Systems Research, vol. 1. Austrian Society for Cybernetic Studies, Vienna, pp. 276–281.

Fomichov, V.A., Fomichova, O.S., 2006. Cognitonics as a new science and its significance for informatics and information society. Informatica. An International Journal of Computing and Informatics (Slovenia) 30, 387–398.

Fomichov, V.A., Fomichova, O.S., 2012. A contribution of cognitonics to secure living in information society. Informatica. An International Journal of Computing and Informatics (Slovenia) 36, 121–130.

Fomichova, O.S., 2009. Humanitarian Education – An Answer to the Challenge of Time. Moscow State University Press, Moscow, p. 368 (in Russian).

Fomichova, O.S., Fomichov, V.A., 2000. Computers and the thought-producing self of the young child. The British Journal of Educational Technology 31, 213–220.

Fomichova, O.S., Fomichov, V.A., 2009. Cognitonics as an answer to the challenge of time. In: Proceedings of the 12th International Multiconference Information Society, IS-2009, Slovenia, Ljubljana, October 12–16, 2009. The Conference Kognitonika/Cognitonics, Jozef Stefan Institute, 2009. 431–434. Available online at: http://is.ijs.si/is/is2009/zborniki.asp?lang=eng.

Huebner, E.S., Gilman, R., 2003. Toward a focus on positive psychology in school psychology. School Psychology Quarterly 18, 99–102.

Kabat-Sinn, J., 2003. Mindfullness-based interventions in the context: past, present, and future. Clinical Psychology: Science and Practice 10, 144–156.

Schonert-Reichl, K.A., Lawlor, M.S., 2010. The effects of a mindfulness-based education program on pre- and early adolescents' well-being and social and emotional competence. Mindfulness 1, 137–151.

Searle, J.R., 1995. The Construction of Social Reality. The Penguine Press, London.

Seligman, M.E.P., Csikszentimihalyi, M., 2000. Positive psychology: an introduction. American Psychologist 55, 5–14.

US Public Health Service, 2000. Report on the Surgeon's General's Conference on Children's Mental Health: A National Action Agenda. Department of Health and Human Services, Washington, DC.

Zelazo, P.D., 2004. The development of conscious control in childhood. Trends in Cognitive Sciences 8, 12–17.

Relevant Website

http://ssrn.com/author=1871348 – It is possible to access the selected papers on the subject of this entry by V.A. Fomichov and O.S. Fomichova on the Social Science Research Network (SSRN).

Consciousness, Neural Basis of

Tamar Weber, University of California Los Angeles, Los Angeles, CA, USA
Hakwan Lau, Columbia University, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C. Koch and F. Crick, volume 4, pp. 2600–2604, © 2001, Elsevier Ltd.

Abstract

In this article, we give an overview of the search for the neural correlates of consciousness (NCCs). We begin with a discussion of the conceptual complexities of defining the notion of an NCC. We then discuss some of the experimental approaches used to empirically investigate the NCCs. We then consider some competing views of NCCs. Finally, we consider how the competing views of NCCs bear on different theories of consciousness. We focus on the methodological and theoretical challenges facing this line of research.

Introduction

The study of consciousness is one of the biggest challenges for scientific explanation. Whatever such an explanation will be – to the extent that one is possible – it will no doubt require an empirical theory of the neural processes thought to give rise to the subjective qualitative nature of conscious experience. If you are reading this article you are no doubt acquainted with the phenomenon of consciousness. But though nothing could be more familiar than one's first-personal phenomenal experience of the world, we will see that the scientific study of consciousness faces some serious methodological challenges.

It will be useful to begin with the distinction between information processing and subjective experience. Consider the mundane example of looking at a glass of red wine sitting on a table before you. Light bounces off the surfaces of the glass and table and reaches your eyes. Invariably some information is lost since the light coming off the three-dimensional scene generates two distinct two-dimensional images each of which is recorded by one of your retinas. Your visual system is faced with the task of combining these two different images to recover the lost information needed to generate a representation of the three-dimensional scene before you. This is a tractable albeit complex information processing problem with a computational solution (Marr, 1982). Cognitive neuroscience has made great strides in studying the brain as a complex information processer. But the brain does more than just information processing. It also generates a subjective experience of the mundane scene. The qualitative aspect of this experience – the feeling of 'what it's like' (Nagel, 1974) to have a visual experience such as a visual experience as of a glass of red wine in front of you – raises a multitude of scientific and philosophical questions: why should there be something it is like to experience the scene? That is, why should the physically realized system that solves the information processing problem also give rise to a subjective experience? And how does it do so? These questions are at the core of the so-called hard problem of consciousness (Chalmers, 1996). What scientific progress can be made on these questions is controversial. There is, however, a related question that is well-suited for scientific inquiry. That is, what are the neural processes and states that are correlated with states of consciousness? While this is an important question in its own right, the hope is that studying the neural mechanisms that are correlated with such states and process might illuminate the hard problem as well (Crick and Koch, 2003).

In what follows, we begin by considering some of the conceptual intricacies of explicating the notion of a neural correlate of consciousness (NCC). We then consider some experimental approaches. We then turn to look at some competing views of NCCs, and the methodological problems that arise in formulating such theories. We then consider how the competing views of NCCs bear on different theories of consciousness.

What Is an NCC? Preliminary Conceptual Issues

The Neural Correlates of Conscious vs Unconscious States

On a first pass, the question of what are the neural correlates of consciousness could be interpreted as the following: (1) which neural states and processes must be active for a subject to be conscious at all? We can distinguish this question from the following: (2) which neural states and processes must be active for a subject to be in a conscious state with some particular content? (1) and (2) can be distinguished by their respective taxonomies of conscious states. (1) is thought to address the background or enabling conditions for consciousness. When these conditions fail to obtain we should expect an absence of consciousness. For example, Bogen's (1995) proposal regarding the function of the intralaminar nuclei of the thalamus is directed at (1). The basic distinction that (1) attempts to track is the difference between conscious and unconscious states. But as Baars (1988) points out there are two major components to consciousness: 'wakefulness' and 'awareness.' Wakefulness is thought to track the level of consciousness (also described as arousal). Awareness tracks the notion of the subjective experience of consciousness. In general one needs to be awake to be aware (though dreaming in rapid eye movement (REM)-sleep is thought to be an exception). Patients in pathological coma or under general anesthesia are considered unconscious because they cannot be awakened. An interesting and ethically relevant line of research concerns cases where wakefulness and awareness are dissociated. One such case is the vegetative state where patients seem to be awake but lack awareness. As Steven Laureys points out, since

we cannot use behavioral measures to differentiate vegetative from minimally conscious patients – that is, patients with some awareness – the use of functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) imaging studies are crucial, though much more work needs to be done before negative results from such studies can be taken as evidence for lack of awareness (Laureys, 2005).

The Neural Correlates of the Contents of Consciousness

Much of the current research on NCCs has been directed at (2), particularly in the visual modality. The goal of this research is to reveal the NCCs of particular types of conscious states, where conscious state types are individuated in terms of their contents. This provides a more fine-grained way of individuating conscious states. If some neural structure N is directly correlated with some conscious state C, where C is individuated in terms of particular contents of consciousness, then if N is active, we should expect a subject that realizes N to be in a state with those particular contents.

Research on the neural correlates of particular contents of consciousness suggests the need for a further refinement of our notion of an NCC. The focus of this research is on finding the minimal sufficient neural correlate for a content of consciousness. If some neural state N is minimally sufficient for a correlated conscious state C, then there will be no proper part of N that is sufficient for C. Without this restriction, we can trivially satisfy the search for the neural correlate of a conscious state by simply pointing to the total state of the brain at the time that the conscious state occurs. Chalmers (2010) points out that the notion of a minimally sufficient NCC requires making a distinction between a 'total' minimally sufficient NCC and a 'core' NCC. For example, the proposed neural correlates of the phenomenal experience as of a face (discussed below) are thought to lie in the 'fusiform face area' (FFA) (Kanwisher, 2001). But the FFA is not a minimally sufficient neural system whose activation guarantees the conscious experience as of a face. The notion of an NCC operative in this line of research is what Chalmers calls a 'core NCC.' The core NCC of a conscious state C with the phenomenal content as of a face (for example), is the part of the neural basis that distinguishes C from a conscious state C,' with different phenomenal content. The search for a core NCC requires distinguishing core processes from background process that make up a total minimally sufficient neural system. Chalmers suggests that drawing the distinction between total and core NCC requires specifying the relevant range of cases over which the correlation holds, given certain normal brain-functioning conditions.

Considering these conceptual distinctions in the abstract of specific views of NCCs can be difficult. We now turn to look at some specific examples of work on NCCs. Doing so will help bring these issues in contact with current research.

Experimental Approaches

Some considerable progress has been made in the search for the neural correlates of the contents of consciousness. One such line of research has pursued the core neural correlates of the visual experience as of motion. Both fMRI studies on human subjects, as well as electrophysiological recordings, and lesion studies in nonhuman primates suggest that the middle temporal visual area is correlated with the subjective experience as of motion (Heeger et al., 1999; Newsom and Paré, 1988). The empirically grounded assumption that humans and nonhuman primates have homologous sensory systems makes it possible to use single-cell recordings data from nonhuman primates to formulate more fine-grained hypotheses about the locations of the neural bases of particular contents of consciousness. The hope is that this will lead to a better understanding of how the activity of particular neuronal populations plays a role in generating the specific contents of a conscious experience.

Another example of research on the core neural correlates of particular contents of consciousness can be found in the work of Kanwisher and colleagues (mentioned earlier). Drawing on data from fMRI scans in a binocular rivalry stimulus condition and stimulus alternation condition, Kanwisher argues that the correlation of neural activity in the FFA and the subjective experience as of faces suggests that FFA plays a role in the awareness of faces in particular, as opposed to playing a role in perceptual awareness generally (Kanwisher, 2001).

The binocular rivalry paradigm is used to generate bistable percepts, in which constant retinal stimulus generates alternating percepts. fMRI scans of humans as well as single-cell recordings in nonhuman primates make it possible to monitor neural activity patterns during alternating percepts. Logothetis and his colleagues conducted a series of seminal experiments using a binocular rivalry paradigm along with single-unit recordings in macaques. The macaques were trained to pull a different lever in response to each of the alternating percepts in the binocular rivalry condition. Logothetis and colleagues repeated this paradigm for various contrasting stimuli pairs such as sunbursts and faces, and gratings with contrasting direction of orientation. The physiological recordings from such experiments showed that activity in primary visual cortex was not well correlated with a change in percept. However, in higher visual areas such as inferior temporal cortex, almost all neurons responded only to a preferred stimulus when that stimulus was perceptually dominant. (For an extensive review of a variety of such experiments see Logothetis, 1998.)

Competing Views of NCCs

We now turn to some competing views of NCCs and the methodological difficulties raised by such views. Due to space limitations, we focus mainly on the current major contenders for the NCCs of visual awareness, at the level of whole-brain functional anatomy. One advantage of focusing on this level of analysis is that the different views turn out to support different theories of consciousness, as we will discuss in the next section. Though we look specifically at the case of the NCCs of visual awareness, we expect that many of the points will generalize to the analogous debates over the NCCs of other sensory modalities such as audition. The analogous views in the case of audition would consider the role of early

auditory areas, where the views below discuss the role of early visual areas.

The Global View: Frontoparietal Network Plus Extrastriate Correlates

We begin with considering three claims, the conjunction of which make up what we'll call the 'Global' view of the NCCs of visual awareness. These are: (1) the primary visual cortex (V1) may be necessary for visual experience but the representations in V1 are not sufficient to give rise to specific contents of consciousness, conscious experience is thought to be generated further upstream; (2) the specific contents of consciousness are represented in the various relevant extrastriate areas, (such as the (FFA) for the conscious experience as of a face discussed above); and (3) activity in the prefrontal and parietal cortices is an enabling factor, which is essential for visually conscious states. The conjunction of (1)–(3) is a popular view that has garnered support from a number of high-profile imaging studies (Sergent and Dehaene, 2004).

We can start to see the contours of one aspect of the debate over the NCCs of visual awareness by comparing two alternative views that differ with respect to claim (3) above. The first alternative we look at rejects (3), and can naturally be extended to support a further distinction between phenomenal consciousness and cognitively accessible consciousness. The second view embraces (3), and holds that most of what has been claimed to be the neural correlates of consciousness in fact reflects signal strength or performance capacity, but not conscious awareness per se. This criticism suggests a distinction between stimulus awareness and subjective awareness.

Visual Activity Only View

Proponents of what we call a 'Visual Activity Only' view argue that activity in the prefrontal and parietal cortices are not essential to visual awareness. Rather these areas reflect associated processes such as attending or cognitive access for the purpose of reporting on the contents of visual experience (Block, 2005, 2007; Lamme, 2006; Macknik, 2006).

Block (2005, 2007) has suggested a distinction between what he originally termed phenomenal consciousness and access consciousness (Block, 1995). Phenomenal NCCs determine the contents of experience. But the phenomenal contents of consciousness are not necessarily cognitively accessible. Contents that are cognitively accessible are contents that are available to the brain's 'consumer' systems. Consumer systems are systems such as reasoning, planning, decision-making, voluntary direction of attention, and systems for rationally controlled action (Block, 2005). Here, we appeal to Block's refined view in which he claims that cognitive accessibility is not constitutively necessary for phenomenal consciousness (extensively defended in Block (2007)). The claim is germane to the current discussion since it highlights important methodological difficulties in the search for the neural correlates of consciousness.

Block defends the claim on both conceptual and empirical grounds. According to Block it is possible that there are modularly encapsulated representations that are inaccessible to report, but may very well be phenomenally conscious. The possibility raises a methodological puzzle for consciousness research. Ideally, we would pursue the question by identifying the neural basis of phenomenal consciousness in clear cases where we are confident that a subject is phenomenally conscious, and then look to see whether this neural machinery is active in cases of encapsulated representations. The methodological puzzle arises since the 'clear cases' will be ones where subjects report having a phenomenal experience and there are no empirically grounded reasons to doubt these reports (as there are in the case of Anton's blindness). But the neural correlates of these clear cases will invariably include the neural mechanisms underlying cognitive accessibility, since the total neural basis of a state with phenomenal consciousness C is itself sufficient for the instantiation of C. So the method would require already having a way to distinguish between the neural bases of cognitive accessibility and phenomenal consciousness. Block points out that it is cognitive accessibility not reportability that is at issue since we have reasons to think that subjects with locked-in syndrome or profound global aphasia (or both) still have phenomenal experience (Block, 2007).

Block's positive view could be considered a species of what we called the 'Visual Activity Only' view. On Block's view, cognitive accessibility is likely supported by prefrontal mechanisms, which are typically associated with higher-level cognition. He argues that sensory representations of the right kind in early visual areas are phenomenally conscious, but they are only cognitively accessible when they are propagated into the prefrontal circuitries. Part of the motivation for this view is that early visual processing is known to have higher informational capacity and spatial resolution than processes in the prefrontal cortex. Therefore, if cognitive access depends on prefrontal mechanisms, and phenomenal consciousness is supported by early visual activity, one can explain why phenomenal consciousness seems to 'overflow' access: that is, we may subjectively experience vivid details more than we can report or access. As evidence for the overflow of phenomenal experience, Block cites the results of Sperling's famous partial-report paradigm (Sperling, 1960) and similar studies (Sligte et al., 2008), as well as a study by Landman et al. (2003), which combines a change blindness paradigm, with the original Sperling paradigm. In the original Sperling paradigm, subjects are shown an array of alphanumeric characters for 50 s followed by a blank screen. Subjects report the subjective experience of having seen all or almost all the characters. However, when subjects were asked which characters they had seen they were typically able to report only three or four characters out of a three by four array. When trained with three different tones to cue the three different rows, subjects showed they were able to report all or almost all the characters of any given row, even though the cue appeared only after the array disappeared. The fact that subjects could do this for any one of the rows supports their subjective reports of having 'seen' all the characters. These studies suggest that only a small subset of what subjects experience is cognitively accessible to report.

Koch and Tsuchiya (2007b) have suggested phenomenology without cognitive access is an example of consciousness without top-down attention. According to Koch and Tsuchiya, empirical evidence that consciousness and top-down, volitionally

controlled endogenous attention are distinct neurobiological processes, supports the distinction between phenomenal consciousness and cognitive accessibility (Koch and Tsuchiya, 2007a,b; Tsuchiya and Koch, 2008). However Koch and Tsuchiya (2007b) seem to differ with Block methodologically in the role of subjects' denials of phenomenal experience. Their discussion suggests that a subject's introspective report (in the absence of compelling evidence to the contrary) should be 'accepted as brute fact.' While their work on the dissociation of top-down mechanisms and phenomenology supports Block, that work does not directly bear on the question of whether current fMRI studies can be used to show that phenomenal consciousness should be attributed to subjects in some of the cases where subjects deny having any phenomenal experience.

Prefrontal Only View

This view locates the essential neural basis of consciousness in later prefrontal and/or parietal areas. This is not to say that proponents of this view think that a state could be visually phenomenally consciousness without input from early visual areas. Rather the view claims that early visual activity alone is not sufficient, and that the subjective aspects of conscious experience are determined by prefrontal and parietal processes. In this sense, this view is similar to the Global View. The critical difference is that according to the Prefrontal Only view, much of early visual (including extrastriate) activity as well as prefrontal and parietal activity just reflects the perceptual signal, which is important for task performance capacity but does not reflect the essentially subjective aspects of conscious perception. If we control for task performance capacity (i.e., by keeping it constant), a difference in subjective conscious perceptual experience may be correlated with a difference in activity in relatively restricted areas in the prefrontal (and possible parietal) areas.

One motivation for this view is that clinical phenomena like blindsight demonstrate the dissociation between subjective reports and task performance. That is, such patients report having no phenomenal experience of a stimulus despite being able to perform well above chance in tasks such as visual discrimination (Weiskrantz, 1986, 1997, 2004). Neuroimaging findings (Persaud et al., 2011; Sahraie et al., 1997) as well as computational modeling (Ko and Lau, 2012; Pasquali et al., 2010; Timmermans et al., 2012) support the idea that when we control for task performance the subjectively reported conscious experience is associated with prefrontal activity.

Lau and Passingham (2006) have demonstrated that this dissociation occurs in normal-sighted subjects as well. They used a metacontrast paradigm to match task performance in subjects' discrimination of two targets in two conditions. Subjects were asked to make a forced-choice as to whether the target was a square or a diamond. Immediately after the forced-choice, subjects were asked to indicate whether they saw the target or simply guessed. Performance difficulty as measured by reaction times was also very similar across the two conditions. Though the two conditions were matched for discrimination performance, they yielded a difference in the frequency of subjects' reports of visual awareness. Comparison of the associated neural activity for the two conditions using fMRI showed a difference in activity in the dorsolateral prefrontal cortex (DLPFC). Notably, Lau and Passingham were unable to find a difference in the visual areas across the two conditions. In a further study, transcranial magnetic stimulation (TMS) application to DLPFC showed a similar finding. Task performance was matched in both TMS and non-TMS application trials, but subjective ratings of visibility differed between the two conditions (Rounis et al., 2010).

Theoretical Implications of the Different Views of NCCs

Global Theories of Consciousness

Many have taken the first view (i.e., the Global View) of the NCC as support for a certain class of theories consciousness. Most prominently, the Global Workspace Theory proposed by Baars (reviewed in 2005), and later developed in neuronal terms by Dehaene and colleagues (Sergent and Dehaene, 2004), would predict results similar to those of the Global View of NCCs. According to the global workspace theory, there exists a central broadcasting mechanism that allows the flexible and global exchange of information between peripheral modules (such as early vision, low level motor processing, and memory). When information enters this workspace, it becomes conscious, and is globally broadcasted to all modules and, therefore, all subsystems can share and edit such information. Dehaene and colleagues specifically hypothesize that the workspace depends on neurons in the prefrontal and parietal areas. It is not entirely clear whether all global theorists hold that early visual activity is necessary. That is, such an activity could be the normal cause that leads to global activity in the prefrontal and parietal cortex, which in turn is necessary for consciousness. Or, alternatively, the early visual activity as well as the activity in prefrontal and parietal cortex are both necessary.

Local Theories of Consciousness

If it turns out to be true that the visual NCC consists of only activity in the posterior visual areas in the brain, theories other than the global workspace would be supported. Block (2005, 2007) for instance, put forward what he calls a 'biological' theory of consciousness, according to which phenomenal visual awareness depends on specific biological substrates in the posterior visual areas in the brain. Note that the global workspace theory as characterized above does not make the distinction between phenomenally conscious contents and cognitively accessible conscious contents. Block's theory and the global workspace theory are often seen as in competition by those who claim that phenomenology that is not accessible should not be considered conscious at all.

Lamme (2006) similarly argues for the need to dissociate consciousness from other cognitive functions, such as attention, working memory, and reportability. On his proposal, visual awareness depends on local feedback loops within the visual areas. By contrast, processing in higher areas is related to the dissociable cognitive functions.

Another theory of visual consciousness which would be supported by the Visual Activity Only view of NCCs is Zeki's

(2003, 2007) 'microconsciousness' theory. The 'microconsciousness' theory suggests that activity in the early visual areas is correlated with specific conscious percepts. On Zeki's proposal, consciousness is not a single unity but consists instead of many components – the 'microconsciousness' – distributed in space and time and which constitute a hierarchy.

The common thread of these theories is that the neural substrate that supports the processing of visual information itself is sufficient to generate a conscious percept, independently of the neural substrate that supports further processing in higher cognitive areas.

Higher-Order Representational Theories of Consciousness

If the Prefrontal Only view of the visual NCC is right, that would lend support to at least some versions of higher-order representational theories of consciousness. According to higher-order theories, in order to have a conscious experience, we need more than just the first-order representation, which has a representation as of some object as its content. On top of that, we need to have a higher-order representation that is about the first-order representation. There is some debate as to whether the content of this higher-order representation is similar to thoughts (Rosenthal, 2005) or percepts (Lycan, 1996), or perhaps that the higher-order representation and the first-order representation should not be counted as separate, but rather as parts of a single self-referential system (Kriegel, 2009).

Lau (2008, 2010) has put forward a different version of a higher-order theory. According to Lau, the higher-order representation simply enables statistical inferences or judgments of perceptual certainty. That is, given that the first-order representations are often corrupted by noise in the brain, the higher-order representation contains some statistical information that allows a judgment of the reliability of the first-order representation. Subjective visual awareness occurs when a percept is judged to be sufficiently reliable.

Higher-order theories, in general, are congenial to the results that a change in activity in the prefrontal and parietal cortices are sufficient for a change in conscious percept as suggested by the prefrontal enabler view. One would expect the higher-order representations to be in these brain regions, as they seem to play roles in uncertainty monitoring, cognitive control, thinking and planning, etc. Lau (2010) predicts that if one keeps the first-order representation constant, a change in the higher-order representations alone would be enough to change the character of a conscious experience.

This prediction is important since it may serve to distinguish at least some higher-order theories from the Global Workspace Theory. It is unclear whether the global workspace theory is committed to such predictions. In most discussions, one typically thinks that a visual signal (roughly equivalent to a first-order representation) enters the global workspace by being stronger than other competing signals. Therefore, when a visual percept changes from consciously seen to unseen, one would expect the difference to be reflected not only in the higher global workspace areas, but in the lower visual signals too. Also, it may be possible that when the signal enters the workspace, the global broadcasting sends feedback to the earlier visual areas, which would in turn mean that the conscious signal would end up becoming stronger in the early areas.

Since the workspace is supposed to be functionally important even when the first-order visual signal is matched, and somehow in one condition the signal gets into the workspace and in another condition it does not, one would predict big behavioral differences, such as that in one condition the visual targets are better identified or reported. However, Lau's higher-order theory would expect that there can be few functional differences in terms of discrimination or identification of the target, if the first-order signal/representation is matched. A difference in higher-order representation leads to a difference in visual awareness, and the only accompanying behavioral difference is the judgment of perceptual certainty.

Concluding Remarks

The foregoing discussion is just a sample of some of the complexities involved in the search for the neural basis of consciousness. We've reviewed some of the conceptual issues in explicating the notion of an NCC. We saw that the search for the minimally sufficient core neural correlates of particular contents require a distinction between background or enabling conditions of those NCCs and the NCCs themselves. We also noted that there are some empirical grounds for dissociating the phenomenological aspects of conscious experience from cognitive functions such as attention and access for the purpose of report. But making such distinctions and dissociations also raises some methodological difficulties. These difficulties suggest that the scientific search for NCCs is quite complex. Clearly, there is much more work that needs to be done. That work includes empirical investigation into the neuronal correlates of the contents of consciousness, as well as reflection on the methodological and explanatory tools needed to theorize about empirical results.

See also: Cognitive Neuropsychology, Methods of; Cognitive Neuroscience; Face Recognition, Psychological and Neural Aspects; Self: Philosophical Aspects; Visual Perception, Neural Basis of.

Bibliography

Baars, B.J., 1988. A Cognitive Theory of Consciousness. Cambridge University Press, New York.

Baars, B.J., 2005. Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. Progress in Brain Research 150, 45–53.

Block, N., 1995. On a confusion about the function of consciousness. The Behavioral and Brain Sciences 18, 227–247.

Block, N., 2005. Two neural correlates of consciousness. Trends in Cognitive Sciences 9 (2), 46–52.

Block, N., 2007. Consciousness, accessibility, and the mesh between psychology and neuroscience. The Behavioral and Brain Sciences 30, 481–499.

Bogen, J.E., 1995. On the neurophysiology of consciousness, part 1: an overview. Consciousness and Cognition 4, 52–65.

Chalmers, D.J. (Ed.), 1996. The conscious mind Search of a Fundamental Theory. Oxford University Press, New York.

Chalmers, D.J., 2010. What is a Neural Correlate of Consciousness? The Character of Consciousness. Oxford University Press, New York, pp. 59–90.

Crick, F., Koch, C., 2003. A framework for consciousness. Nature Neuroscience 6, 119–126.
Heeger, D.J., Boynton, G.M., Demb, J.B., Seideman, E., Newsome, W.T., 1999. Motion opponency in visual cortex. The Journal of Neuroscience 19, 7162–7174.
Kanwisher, N., 2001. Neural events and perceptual awareness. Cognition 79, 89–113.
Ko, Y., Lau, H., 2012. A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences B 367 (1594), 1401–1411.
Koch, C., Tsuchiya, N., 2007a. Attention and consciousness: two distinct brain processes. Trends in Cognitive Sciences 11, 16–22.
Koch, C., Tsuchiya, N., 2007b. Phenomenology without conscious access is a form of consciousness without top-down attention. The Behavioral and Brain Sciences 30, 509–510.
Kriegel, U., 2009. Subjective Consciousness: A Self-Representational Theory. Oxford University Press, Oxford.
Lamme, V.A.F., 2006. Towards a true neural stance on consciousness. Trends in Cognitive Sciences 10 (11), 494–501.
Landman, R., Spekreijse, H., Lamme, V.A.F., 2003. Large capacity storage of integrated objects before change blindness. Vision Research 43, 2149–2164.
Lau, H.C., 2008. A higher order Bayesian decision theory of consciousness. Progress in Brain Research 168, 35–48.
Lau, H.C., 2010. Theoretical motivations for investigating the neural correlates of consciousness. Wiley Interdisciplinary Reviews: Cognitive Science 2 (1), 1–7.
Lau, H.C., Passingham, R.E., 2006. Relative blindsight in normal observers and the neural correlate of visual consciousness. Proceedings of the National Academy of Sciences of the United States of America 103, 18763–18768.
Laureys, S., 2005. The neural correlate of (un)awareness: lessons from the vegetative state. Trends in Cognitive Sciences 9 (12), 556–559.
Logothetis, N.K., 1998. Single units and conscious vision. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 353, 1801–1818.
Lycan, W., 1996. Consciousness and Experience. Bradford Books/MIT Press, Cambridge.
Macknik, S.L., 2006. Visual masking approaches to visual awareness. Progress in Brain Research 155, 177–215.
Marr, D., 1982. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. W.H. Freeman and Company, San Francisco.
Nagel, T., 1974. What is it like to be a bat? The Philosophical Review 83, 4435–4450.
Newsom, W.T., Paré, E.B., June 1988. A selective impairment of motion perception following lesions of the middle temporal visual area (MT). The Journal of Neuroscience 8 (6), 2201–2211.
Pasquali, A., Timmermans, B., Cleeremans, A., 2010. Know thyself: metacognitive networks and measures of consciousness. Cognition 117, 182–190.
Persaud, N., Davidson, M., Maniscalco, B., Mobbs, D., Passingham, R.E., Cowey, A., Lau, H., 2011. Awareness-related activity in prefrontal and parietal cortices in blindsight reflects more than superior visual performance. NeuroImage 58, 605–611.
Rosenthal, D., 2005. Consciousness and Mind. Oxford University Press, New York.
Rounis, E., Maniscalco, B., Rothwell, J., Passingham, R.E., Lau, H., 2010. Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. Journal of Cognitive Neuroscience 1, 3165–3175.
Sahraie, A., Weiskrantz, L., Barbur, J.L., Simmons, A., Williams, S.C., Brammer, M.J., 1997. Pattern of neuronal activity associated with conscious and unconscious processing of visual signals. Proceedings of the National Academy of Sciences of the United States of America 94 (17), 9406–9411.
Sergent, C., Dehaene, S., 2004. Neural processes underlying conscious perception: experimental findings and a global neuronal workspace framework. Journal of Physiology, Paris 98, 374–384.
Sligte, I.G., Scholte, H.S., Lamme, V.A.F., 2008. Are there multiple visual short-term memory stores? PLos One 3 (2), 1–9.
Sperling, G., 1960. The information available in brief visual presentations. Psychological Monographs: General and Applied 74 (11, Whole No. 498), 1–29.
Timmermans, B., Schilbach, L., Pasquali, A., Cleeremans, A., 2012. Higher-order thoughts in action: consciousness as an unconscious re-description process. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 367, 1412–1423.
Tsuchiya, N., Koch, C., 2008. The relationship between consciousness and attention. In: Laureys, S., Tononi, G. (Eds.), The Neurology of Consciousness. Elsevier, Amsterdam.
Weiskrantz, L., 1986. Blindsight: A Case Study and Implications. Clarendon Press, New York.
Weiskrantz, L., 1997. Consciousness Lost and Found: A Neuropsychological Exploration. Oxford University Press, New York.
Weiskrantz, L., 2004. Roots of blindsight. Progress in Brain Research 144, 229–241.
Zeki, S., 2003. The disunity of consciousness. Trends in Cognitive Sciences 7 (5), 214–218.
Zeki, S., 2007. The disunity of consciousness. Progress in Brain Research 168, 11–18.

Consequentialism (Including Utilitarianism)

Philip Pettit, Princeton University, Princeton, USA; and Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Consequentialism is the theory that the way to tell whether a particular choice is the right choice for an agent to make or to have made is to look at the relevant consequences of the decision: to look at the relevant effects of the decision on the world. It gives rise to many questions. Some are general questions as to how more precisely the doctrine can be defined and defended, and as to what the main forms of opposition are. Others are in-house questions that go to matters of detail in the specification of the theory. This article offers an overview of positions on those questions.

Consequentialism is the theory that the way to tell whether a particular choice is the right choice for an agent to make or to have made is to look at the relevant consequences of the decision: to look at the relevant effects of the decision on the world. It gives rise to many questions. Some are general questions as to how more precisely the doctrine can be defined and defended, and as to what the main forms of opposition are. Others are in-house questions that go to matters of detail in the specification of the theory. This article offers an overview of positions on those questions.

Roughly speaking, consequentialism is the theory that the way to tell whether a particular choice is the right choice for an agent to make or have made is to look at the relevant consequences of the decision: to look at the relevant effects of the decision on the world. In directing us to consequences, the theory is teleological in focus; the term comes from the Greek word *telos*, meaning end or goal. It is opposed to any theory that does not make the consequences into the ultimate touchstone of rightness and wrongness in choices. The classical antonym is the deontological approach – this term comes from the Greek work *deon*, meaning obligation or duty – but there are other antonyms too, as we shall see. Where consequentialism would assess a choice by looking at its consequences, a deontological approach would assess it by looking at how the choice measured up to the obligations incumbent on the agent.

The Definition of Consequentialism

Nonconsequentialism comes in many forms. Deontologists hold that agents ought to discharge certain duties: They ought to tell the truth, keep their promises, be nonviolent, etc. Kantians say that agents ought to act on the categorical imperative – act only on a maxim that they could accept as a general law of behavior – or ought to treat other people always as ends, never only as means. Virtue ethicists say that they ought to manifest certain virtues in their behavior. Rights theorists maintain that they ought to respect certain rights that others have against them. Contractualists assert that they ought to conform to principles that no one could reasonably object to as the bases of social life. Theorists of special obligation say that they ought to deal in a certain way with those who are bound to them, such as their children, spouses, friends, and so on.

What is the common thread in these nonconsequentialist positions? All nonconsequentialists speak, at whatever level of abstraction, about what any or every agent ought to do or be; in that sense they are universalists. All prescribe neutral patterns of behavior or psychology or relationships: that people act on the categorical imperative, manifest certain virtues, respect the rights of their dependants, nurture their friendships, and so on. The patterns are neutral in the sense that they can be understood in the same way by everyone. And all say that the right thing for an agent to do is to instantiate the prescribed patterns – so far as they are coinstantiable – in their own behavior or psychology or relationships. In particular, they say that that is the thing to do even if instantiating a pattern in their own life means, because of the perversity of the agent's circumstances, that the pattern will be less fully realized than otherwise in the world at large.

Consequentialism makes a dual assertion in response to such positions (Baron et al., 1997). First, it asserts that there are certain potentially shared values by means of which possible states of affairs can be ranked, although perhaps not completely. These may be neutral patterns in human behavior or psychology or relationships, of the kind that the nonconsequentialist prescribes, or they may be neutral outcome patterns that are more detached from agency: patterns such as the maximization of human happiness or the saving of uninhabited wildernesses. And second, consequentialism maintains that the right choice for an agent to take in any decision is one of those choices, assuming there is at least one – assuming incomplete ranking is not a problem – that promotes the overall realization of such values or patterns (more below on what it is to promote values).

The basic difference, then, between consequentialists and their opponents is that while each side privileges certain general patterns – treats them as values, as it is natural to say – the consequentialist side says that the important thing is to promote the realization of those patterns in the world at large, while the opposing side says that the important thing is rather to instantiate the relevant patterns in one's own life: in one's behavior or relationships or psychology, or whatever. Consequentialists say that privileged patterns should be treated as consequences to be promoted, nonconsequentialists that they should be treated as constraints to be satisfied.

The difference between the two approaches is particularly salient for simple cases where only one pattern is taken to be relevant, and where there is no issue about the relative

importance of different patterns. Consider those cases, then, where nonviolence is assumed to be the only privileged pattern – the only value – that is relevant. And consider the different ways in which that pattern may be espoused among pacifists. The nonconsequentialist pacifist says that nonviolence is to be instantiated by agents, even if this means that there is more violence overall: even if violence spreads as a result of being unopposed. The consequentialist counterpart says that nonviolence is to be promoted overall, even if this means instantiating violence: say, waging the war to end all wars. In practice, the two sorts of pacifist will often converge in their recommendations, as nonconsequentialist Quakers converged with Bertrand Russell in their opposition to World War I. But there is always a possibility of their coming apart, as indeed those Quakers and Russell came apart in their attitudes to World War II.

Making the distinction between consequentialism and nonconsequentialism by reference to whether the agent is required to promote certain privileged patterns or rather to instantiate them – to honor them – is only one among a number of ways of making it. Another common approach is to distinguish between agent-neutral values and agent-relative ones, and to say that consequentialists advocate the promotion only of agent-neutral values, and nonconsequentialists the promotion of a mixed bunch or of agent-relative values only (Nagel, 1986).

Agent-neutral values are patterns that are identified without reference to who is doing the identifying: patterns like those involved in peace being achieved, happiness maximized, a wilderness preserved, and so on. Agent-relative values are patterns that are identified by back reference to the identifying person: patterns, to take the case where I do the identifying, like *my* doing my duty, *my* conforming to the categorical imperative, *my* manifesting certain virtues, *my* looking after my children, and so on. Where consequentialists advocate the promotion of agent-neutral values, and in particular agent-neutral values of a universal cast – they do not involve benefits just to a particular person or place or whatever – nonconsequentialists can be represented as advocating the promotion of agent-relative values: the promotion by every agent of *their* doing their duty, *their* conforming to the categorical imperative, *their* manifesting certain virtues, *their* looking after their children, and so on.

A Theory of the Right, Not a Theory of the Good

To say that something is good is to hold that it has a certain value, in particular a certain positive value. To say that something is right is to hold that among certain relevant alternatives – say, among the options in a choice – it is the one to have. And so rightness and goodness may come apart. An alternative that is wrong may yet be an object of great value; it may be wrong simply because another option is an object of even greater value. An alternative that is right, on the other hand, may not be something of great value; it may be right, simply because it is the least bad option among a very poor set of alternatives. A theory of the good, a theory of value, would enable us to determine the values of different entities. A theory of the right would enable us to determine, for any set of alternatives, which alternative or subset of alternatives is the right one.

When I speak of what is good, and of what has value, I have in mind an impersonal or agent-neutral conception of goodness and value. If I value a prospect for the increase of happiness it promises, or even for a particular effect it will have, say on planet Earth, then I value it agent-neutrally. If I value it for the benefits it will have for me or mine, or for the fact that it will keep my hands clean, or for any reason of that self-referential kind, then I value it agent-relatively. The theory of the good, or the theory of the valuable, refers to the theory of what is agent-neutrally valuable.

Does consequentialism present itself as a theory of the good or as a theory of the right, or as something involving commitments in both areas? By the account given, the theory is concerned only with the right, not with the good. It says that the right alternative in any choice is that which promotes the good, without making any commitment on the nature of the good. While it makes the right a function of the good, it says nothing in itself on the matter of what is good.

This aspect of consequentialism is often found surprising, as many people tend to identify consequentialism with utilitarianism. But utilitarianism involves a double commitment. It says that the good should be equated with happiness, whether this is understood in the sense of pleasure or preference satisfaction, and then it adds that the right alternative in any set is that which best promotes such happiness (Sen, 1979). Thus when G. E. Moore (1903) rejected utilitarianism, arguing that the good was not reducible to happiness – it was a nonnatural property that allowed of no such reduction – he remained a consequentialist about the right. Many contemporary consequentialists have nonutilitarian theories of the good or the valuable. Indeed they often look like nonconsequentialists to the extent that they may think that the good encompasses reflexive patterns, such as those involved in people generally keeping their promises, respecting those with whom they deal, or giving their children special attention.

Robert Nozick (1974), himself an outstanding nonconsequentialist, supports this view when he distinguishes between a nonconsequentialist theory of rights – a theory of rights, proper – and a consequentialism or utilitarianism of rights. The consequentialist theory of rights would say that it is good that certain rights are enjoyed generally, and it would imply that the right option in any choice is that which promotes the enjoyment of such rights. This right option will itself instantiate the rights-observing pattern in most imaginable cases but it will not do so in those exceptional circumstances where the best way to promote the pattern is not to instantiate it oneself: where the best way to promote rights observance is to breach those rights in one's own case.

Consequentialism, Global, and Local

Consistently with being a theory of the right, consequentialism can be thought of as having a smaller or a larger range of application to human agents (Pettit and Smith, 2000). There are sets of alternative acts that we might perform in different choices, sets of alternative decision procedures that we might follow in choosing acts, sets of alternative motives out of which we might be led to act, sets of alternative rules that we might comply with in acting, and so on. Is consequentialism a local

doctrine that just tells us what the best act is in any choice, for example – or the best decision procedure, or motive, or rule, or whatever – and that then says that the best alternatives in other sets are to be defined in a derived way? If acts are privileged, the best decision procedure might be that which produces the best acts, the best rule that which is consistent with the choice of the best acts, and so on through other categories. Or should consequentialism be taken to apply globally, telling us that as the best act is that which produces the best consequences among the alternatives on offer, so the same lesson holds for items in the other sets of alternatives?

Most consequentialist positions in the literature tend to be local in character. Act consequentialism, as it is called, tends to say that the best act is the act with the best consequences, and that the best factor in any other set of alternatives involving an agent is to be defined in some way – a number of ways are possible – by reference to how it serves the choice of best acts. Rule consequentialism gives the priority to rules in a parallel way, and motive consequentialism gives the priority to motives.

The most natural position, however, would seem to be global consequentialism (Parfit, 1984). It would say that no matter what the set of alternatives under consideration, the one to be preferred is that which does best by the consequences. The same lesson will apply, of course, to that set of alternatives whose members are different sets of act–procedure–motive–rule combinations; the right combination will be that which does best by the consequences overall. The act or procedure or motive or rule selected for any context under the best combination may not be the best possible one relative to the apparent alternatives in that context. But it will be the best possible one that is available consistently with the best possible combination overall.

The Role of Consequentialism

There is a long tradition among the opponents of consequentialist doctrines, in particular the opponents of utilitarianism, of suggesting that consequentialists are committed to holding that every choice should be made in a highly calculative, actuarial mode. F. H. Bradley made the point in the last century, writing about the utilitarian approach: "So far as my lights go, this is to make possible, to justify, and even to encourage, an incessant practical casuistry; and that, it need scarcely be added, is the death of morality" (Bradley, 1876, p. 107; see too Williams in Smart and Williams, 1973). Consequentialists have almost always resisted this charge. Thus, Henry Sidgwick in the last century and J. J. C. Smart in this have argued forcibly that utilitarianism does not require that agents all make their decisions by explicit reference to how the options will do by the promotion of happiness (Smart and Williams, 1973). The point they have wanted to make was nicely summed up early in the nineteenth century by the jurisprude, John Austin, in defending the utilitarian thinker: "he approves of love because it accords with his principle, he is far from maintaining that the general good ought to be the motive of the lover. It was never contended or conceived by a sound, orthodox utilitarian, that the lover should kiss his mistress with an eye to a common weal" (Austin, 1954, p. 108).

As recent consequentialists have almost all taken the view that consequentialism is a theory of the right, not a theory of the good, so they have nearly unanimously argued that it is a theory for assessing the right alternative for an agent or an agency to have adopted, not necessarily a useful theory to be applied in choosing from among alternatives. The point is usually made with respect to acts but it can also be extended to alternatives in other categories. The claim is that making one's decision by reference to which option has the best consequences may be a way of making one's decision that does not itself have the best consequences. Suppose one is concerned with one's own pleasure, for example. It is notorious that agents who make their decisions by reference to which option will promote the most pleasure may actually enjoy less pleasure than those who make their decisions in the more spontaneous fashion, say by reference to certain rules of thumb (Railton, 1984; Pettit and Brennan, 1986).

This aspect of consequentialism has been described as 'the paradox of teleology' (Scanlon, 1998, p. 383, n. 15). It has important implications and these need to be appreciated, if the appeal of consequentialism is to be understood. It means that the consequentialist can look with approval on agents whose mode of decision making in everyday life involves a consideration of consequences only in periodic reviews of their performance or in times of crisis. These agents may be routinely available to their family and friends; they may be reliable truth tellers and promise keepers; and they may be dependably respectful of the claims of others upon them. The consequentialist will approve of such people to the extent that they are more likely to promote the good – whatever view of the good is adopted – than actuarial, calculative types.

The Positive Case for Consequentialism

Why be a consequentialist? Surprisingly, there has been very little discussion of this question in the literature. Consequentialists tend to assume that if one is to be serious about the enterprise of moral assessment, then there is no alternative to looking at the consequences – the consequences, neutrally, or impersonally characterized – of the choices assessed. How could it be rational, they ask, to neglect any such consequences? They assume that if nonconsequentialists think otherwise, that is because of an unquestioning commitment to a theological or a commonsense viewpoint. A theological viewpoint might suggest that human agents ought to fulfill prescribed obligations and let God look after the consequences. And a commonsense viewpoint might support the attitude that the important thing in moral decision making is to look after one's own moral standing, to keep one's own hands clean, and not to worry, except perhaps in exceptional circumstances, about any bad consequences that may follow on this.

There is also a second reason why consequentialists may not have concerned themselves much with providing a defense of their position. Consequentialists are a mixed bunch, as we saw, insofar as they differ widely among themselves on the question of which substantive sorts of consequences – which valuable patterns – are the ones by reference to which choices ought to be assessed. For example, they divide into utilitarians and nonutilitarians, depending on whether or not they accept the

view that it is only consequences that bear on the utility of sentient beings which matter: only consequences that affect the happiness or preference satisfaction of such beings. Being a mixed bunch, consequentialists tend each to be concerned more with arguing for their particular view of what consequences count than with arguing for the shared consequentialist credo that it is indeed consequences that matter and not the sort of thing on which deontologists and other nonconsequentialists focus.

My own belief is that those who are attracted to consequentialism and nonconsequentialism are moved by different views of the nature of justification (Pettit in Baron et al., 1997). According to consequentialism, people will be able to think of their choices as justifiable to one another only insofar as they think that there are agent-neutral values that others ought to accept and that serve to justify those choices. According to nonconsequentialist positions, people will be able to think of their choices as justifiable under a weaker condition: if they think that there are agent-relative values such that others can be made to understand why they should be moved by those values – others are moved by corresponding values in their own case – and the values justify their choices. Thus consequentialists think that justification must ultimately be based on potentially shared or convergent concerns: say, with overall happiness or peace or good relations between people. But nonconsequentialists think that it can be based, even at the most ultimate level, on concerns that are isomorphic or concordant rather than convergent: say, on my concern that I should not break my promises, your concern that you should not break yours, whatever the impact on promise keeping more generally, or my concern with my children, your concern with yours, whatever the repercussions for overall child care.

The Negative Case for Consequentialism

But if consequentialists have not done much in the way of arguing for their position, what have been the issues debated between them and opponents (Scheffler, 1988; Darwall, 2003)? These have mainly had to do with how far consequentialism can support firm commonsense intuitions about what is morally right. Opponents typically object to consequentialism on the grounds that in one way or another it would undermine otherwise compelling moral attitudes. Rather than providing positive reason for adopting their position, consequentialists have often been content to try to deal with those objections or to try to argue that while the points made are valid, they are not as important as deontologists make them seem.

There are three sorts of objection, in particular, that consequentialists have had to confront in this exchange. First, and most traditionally, that their doctrine would make many intuitively wrong choices right, many intuitively right choices wrong. Second, an objection associated in particular with Bernard Williams (Smart and Williams, 1973), that even where it represents many intuitively right choices as right it would provide intuitively the wrong reasons for thinking them right. And third, a much debated objection, that it would represent as obligatory various right choices that are supererogatory, that is intuitively beyond the call of duty (Scheffler, 1982; Kagan, 1989; Murphy, 2000; Mulgan, 2001; Singer, 1993). Consequentialists have often argued in this exchange that our intuitions on relevant matters need to be revised but the more general line has been to maintain that on reflection consequentialism proves less counterintuitive than its opponents maintain.

The Collective Face of Consequentialism

Suppose we had a society of utilitarians, each of whom acted in a way that utilitarians would approve. Is it possible that the collective result of their so acting might be worse than the collective result of actions that did not individually earn utilitarian approval? Is it possible, in particular, that the collective result might be worse in utilitarian terms: that it might actually lead to less human happiness or preference satisfaction than certain alternatives? Is it possible that, however satisfactory as a theory of rightness for individual agents, it is not in this sense a collectively satisfactory theory (Hodgson, 1967)?

Does it matter that a theory designed for individuals is collectively self-defeating? Does it matter that agents who try to do what is best by the theory, or succeed in doing what is best by the theory, generate a collective result that is worse in their own terms than it might have been (Parfit, 1984)? Some are happy to say that a theory can be self-defeating in this way and yet remain a compelling theory for individuals. Others will think that this is a matter of the first importance for the assessment of a theory.

Some may be willing to concede that consequentialism can be collectively self-defeating but most deny this, on a variety of grounds (Lewis, 1972; Singer, 1972). One well-known line is that where collective self-defeat appears to threaten, it usually transpires that one of the options before the individuals is not being taken into account: say, the option of exploring possibilities of cooperation with others, rather than acting unilaterally, or the option of seeking further information on the prospects facing the agent (Regan, 1980).

Collective problems are actually more troublesome for nonconsequentialists. The prisoner's dilemma is often taken to show that what is individually rational may not be collectively rational: each of two prisoners confesses to a crime because, whatever the other does, confessing promises a better result than refusing to confess; yet each would be better off if both refused to confess than they are by both confessing. It is worth noting that where agents have distinct ends, as they will have under nonconsequentialism, then a prisoner's dilemma is a constant threat. The distinct ends that agents have may consist in individually wanting to do their own duty, individually wanting to instantiate the categorical imperative in their own behavior, individually wanting to do the best by their own children, and so on. Such ends are potentially competitive and so it is a permanent possibility that by each pursuing their own ends as best they can, they will secure a result under which neither does as well as they might have done in their own terms. By each trying to do the best they can by their own children, for example, they may each do less well than if they pooled their resources and did their best for children as such. Whatever problems are raised for consequentialism on the

collective front, it is better protected in this respect against prisoner's dilemmas, for it assigns to each agent the same end of promoting the best consequences (Parfit, 1984).

Some In-House Issues

There are a number of more or less in-house questions that are debated among consequentialists. One bears on how we should identify the alternatives to be evaluated in any decision. Should an agent do A or B, we ask. A, we answer, because its consequences do better on the relevant theory of the good. But what if there is some other option, C, which the agent might have chosen instead and which would have itself been better than A? This sort of example alerts us to the fact that in the consequentialist assessment of any choice, we have to be clear about what are the relevant alternatives to bring into consideration.

The question about alternatives has not received as much attention as it deserves in the literature, but it is undoubtedly of great importance (Bergstrom, 1966; Jackson and Pargetter, 1986). According to one view, we should assess any action against the alternative (or disjunction of alternatives) that the agent would have taken, had the action not been performed. According to another, we should assess it against the alternatives that the agent could have performed instead: against the alternatives that were, in a relevant sense, within the power of the agent. From a third-person point of view, the first approach would seem to be the most pertinent, although from a first-person point of view, things are not so clear.

A second in-house question has been more widely debated in consequentialist circles. This is the issue as to whether in looking at the decision an agent or agency ought to have made, we consider the actual consequences of the option chosen and the consequences that would have followed on alternatives; or whether, rather, we look at the expected consequences at the time of choice (Lewis, 1969; Jackson, 1991). If the answer is that we should look at expected rather than actual consequences, then the question is whether we should look at the subjectively expected consequences – in the manner of decision theory – or whether we should look at the consequences that were in some sense objectively expected (Menzies and Oddie, 1992). Suppose that a doctor prescribes a drug for a nonfatal skin condition, which has the following features: there is a 10 percent chance that it will kill the patient, an 80 percent chance that it will make no difference, and a 10 percent chance that it will cure the complaint. Imagine that the drug works and the complaint is cured. Does consequentialism take its cue from actual consequences and say that the doctor made the right choice? Or does it look to expected consequences, whether subjectively or objectively expected, and say that the doctor made the wrong choice? That is the issue here.

The third more or less in-house issue is whether consequentialism should hold that the right option is that which does best by relevant consequences – that which is optimific – or whether it is sufficient for an option to be right that it does well enough, as we might say, by the relevant consequences. Some theories of rationality are maximizing theories, requiring rational agents to maximize some function, whereas others are satisficing theories: these require rational agents just to perform satisfactorily by some relevant criterion. The question here is whether consequentialism should not take the softer, satisficing line rather than the stern, maximizing one (Slote and Pettit, 1984; Bradley, 2006).

See also: Decision Making, Psychology of; Utilitarian Social Thought, History of.

Bibliography

Austin, J., 1954. The Province of Jurisprudence. Weidenfeld, London, UK.

Baron, M.W., Pettit, P., Slote, M., 1997. Three Methods of Ethics: A Debate. Blackwell, Oxford, UK.

Bergstrom, L., 1966. The Alternatives and Consequences of Actions. Almqvist and Wiksell, Stockholm.

Bradley, F.H., 1876. Ethical Studies, second ed., rev. Oxford University Press, Oxford, UK.

Bradley, B., 2006. Against satisficing consequentialism. Utilitas 18, 97–108.

Darwall, S. (Ed.), 2003. Consequentialism. Oxford University Press, Oxford, UK.

Hodgson, D.H., 1967. Consequences of Utilitarianism. Clarendon Press, Oxford, UK.

Jackson, F., 1991. Decision-theoretic consequentialism and the nearest and dearest objection. Ethics 101, 461–482.

Jackson, F., Pargetter, R., 1986. Oughts, options, and actualism. Philosophical Review 95, 233–255.

Kagan, S., 1989. The Limits of Morality. Oxford University Press, Oxford, UK.

Lewis, C.I., 1969. Values and Imperatives. Stanford University Press, Stanford, CA.

Lewis, D., 1972. Utilitarianism and truthfulness. Australasian Journal of Philosophy 50.

Menzies, P., Oddie, G., 1992. An objectivist's guide to subjective value. Ethics 102, S12–S33.

Moore, G.E., 1903. Principia Ethica. Cambridge University Press, Cambridge, UK.

Mulgan, T., 2001. The Demands of Morality. Oxford University Press, Oxford, UK.

Murphy, L., 2000. Moral Demands in Nonideal Theory. Oxford University Press, Oxford, UK.

Nagel, T., 1986. The View from Nowhere. Oxford University Press, Oxford, UK.

Nozick, R., 1974. Anarchy, State, and Utopia. Blackwell, Oxford, UK.

Parfit, D., 1984. Reasons and Persons. Clarendon Press, Oxford, UK.

Pettit, P., Brennan, G., 1986. Restrictive consequentialism. Australasian Journal of Philosophy 64, 438–455.

Pettit, P., Smith, M., 2000. Global consequentialism. In: Hooker, B., Mason, E., Miller, D.E. (Eds.), Morality, Rules and Consequences. Edinburgh University Press, Edinburgh, UK, pp. 121–133.

Railton, P., 1984. Alienation, consequentialism and the demands of morality. Philosophy and Public Affairs 13, 134–171.

Regan, D.H., 1980. Utilitarianism and Cooperation. Oxford University Press, Oxford, UK.

Scanlon, T.M., 1998. What We Owe to Each Other. Belkanap Press of Harvard University Press, Cambridge, MA.

Scheffler, S., 1982. The Rejection of Consequentialism. Oxford University Press, Oxford, UK.

Scheffler, S. (Ed.), 1988. Consequentialism and Its Critics. Oxford University Press, Oxford, UK.

Sen, A., 1979. Utilitarianism and welfarism. Journal of Philosophy 76, 463–489.

Singer, P., 1972. Is act-utilitarianism self-defeating? Philosophical Review 81, 94–104.

Singer, P., 1993. Practical Ethics, second ed. Cambridge University Press, Cambridge.

Slote, M., Pettit, P., 1984. Satisficing consequentialism. Proceedings of the Aristotelian Society 8 (Suppl.), 139–176.

Smart, J.J.C., Williams, B., 1973. Utilitarianism, For and Against. Cambridge University Press, Cambridge, UK.

Conservatism: Historical Aspects

JZ Muller, Catholic University of America, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 4, pp. 2624–2628, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

Conservatism refers to an ideology and to the political movements in which this ideology plays a role. For reasons explained below, conservatism has been more difficult to define than other major modern ideologies. This article begins by exploring definitional difficulties, discusses approaches to the study of conservatism, offers a definition and characterization of conservatism as a distinct mode of political and social analysis, and finally provides a brief look at the history of the term and of the changing emphases of conservative thought.

Approaches to the Problem

An intuitive procedure for defining conservatism is to begin by listing the institutions that conservatives have sought to conserve. This will not get us very far, for conservatives have, at one time and place or another, defended a wide range of social, political, and economic institutions.

The study of conservatism has proved difficult for modern social science. Compared with the other great ideological phenomena of the past two centuries, such as liberalism, socialism, communism, or fascism, the subject of conservatism is underilluminated by scholarship. One reason is because of the difficulty of arriving at meaningful generalizations about conservatism, which displays less obvious uniformity across national borders and tends to be more nationally particular than liberalism or socialism, which aspire to be universal in their reach. Moreover, since conservatism emphasizes the need for institutional and symbolic continuity with the particular past, its symbols and institutional ideals tend to be more tied to specific, usually national, contexts. For all these reasons, studies of conservatism that go beyond one particular national context are scarce; as a result, conceptions of conservatism have tended to be parochial.

One of the earliest social scientific approaches to the issue was formulated by the Hungarian–German sociologist, Karl Mannheim, in his essay *Conservative Thought* (1927). He introduced the important distinction between traditionalism and conservatism, the former a near-universal psychological propensity to do things as they have traditionally been done and the latter an articulated set of interrelated political ideas, an ideology that arose, Mannheim thought, in response to the new and dynamic historical processes associated with the Enlightenment (*see* Enlightenment) and capitalist modernization (*see* Capitalism) and the bourgeois class (*see* Bourgeoisie and Middle Classes, History of). A pioneer of the sociology of knowledge, Mannheim claimed that conservative thought was linked to a worldview, which could be correlated with class position, roughly that of the feudal nobility. The basic categories of each ideology, he maintained, were unconsciously determined by the experience of the world and the self-interest of the class that developed it. In the case of conservatism, it led to a focus on experience and the concrete, as opposed to universalistic, rationalistic theories of liberalism and the Enlightenment. Conservatives offered a critique of the rationalistic conceptions for reorganizing society proposed by advocates of radical reform, and focused on explaining the historical particularities of real, existing societies and the interconnectedness of their institutions.

Subsequent social scientific analysts of conservatism, even when they have not followed Mannheim's focus on class, have stressed the reactive nature of conservatism, the fact that it arises in response to a challenge (intellectual, political, or cultural) to existing institutions, on behalf of which conservative arguments are then developed. In one of the most perceptive scholarly analyses of the subject, Samuel Huntington (1957) argued that conservatism is best understood not as an inherent theory in defense of particular classes and institutions, but as a positional ideology. "When the foundations of society are threatened, the conservative ideology reminds men of the necessity of some institutions and desirability of the existing ones," Huntington suggested. He claimed that because "the articulation of conservatism is a response to a specific social situation … . The manifestation of conservatism at any one time and place has little connection with its manifestation at any other time and place."

The reactive nature of conservatism is a premise of one of the finest historical studies of the subject, Klaus Epstein's *The Genesis of German Conservatism* (1966), which pays close attention to political and intellectual context and demonstrates how conservatism arose in German-speaking Europe as a response to the Enlightenment, commercial capitalism, and bourgeois liberalism. Taking the reactive approach a step further, Martin Greiffenhagen (1971), suggested that because self-conscious conservatism only arises once the institutions it values have lost their hold, conservative thought seeks not a preservation of the status quo, but uses an imaginatively transfigured conception of the past with which to criticize the present. According to Greiffenhagen, the romanticization of the past is thus an intrinsic and recurrent element of conservatism.

Two thinkers presented nonreactive conceptions of conservatism during the immediate postwar decade. In *The Conservative Mind* (1953), Russell Kirk, a US conservative man of letters, offered a conception of conservatism as based on religion and 'belief in a transcendent order, or body of natural law,

which rules society as well as conscience' – a conception that assimilated conservatism to religious orthodoxy, and was of scant use as an analytic tool. Kirk's conception was rejected by Michael Oakeshott, the most important British conservative philosopher of the era, in *On Being Conservative* (written 1956 and published in *Rationalism in Politics and Other Essays*, 1962), which presented a brilliant phenomenological description of conservatism as a disposition centered "upon a propensity … to delight in what is present rather than what was or what may be". Oakeshott sought to conserve a liberal and pluralist conception of society, and doubted that 'a disposition to be conservative in politics' depended on the religious and metaphysical assumptions set forth by Kirk.

Definition and Characterization

To understand conservatism as a distinctive mode of social and political thought, it is useful to distinguish it, ideally typically at least, from reaction and orthodox. The orthodox theoretician defends existing institutions and practices because they are metaphysically true: the truth proclaimed may be based on particular revelation or on natural laws purportedly accessible to all rational men; it may be religious or secular in origin. The conservative defends existing institutions because their very existence creates a presumption that they have served some useful function; because eliminating them may lead to harmful, unintended consequences; or because the veneration that attaches to institutions that have existed over time makes them potentially usable for new purposes.

Conservatism can similarly be distinguished from reaction. The conservative seeks to conserve existing institutions, usually recognizing that the process of conservation may include the need for evolutionary reform. The reactionary, by contrast, is at odds with existing institutions, and seeks to return to some institutional status quo ante, often in a form transfigured by memory and ideology.

As misleading as the confusion between conservatism and orthodoxy is the false dichotomy of conservatism and the Enlightenment. Contrary to the frequent characterization of conservatism as the enemy of the Enlightenment, it is more historically accurate to say that there were many currents within the Enlightenment, and some of them were conservative. Indeed, conservatism as a distinct mode of thought is a product of the Enlightenment. What makes social and political arguments conservative as opposed to orthodox is that the critique of liberal or progressive arguments takes place on the enlightened grounds of the search for human happiness, based on the use of reason.

Characteristics of Conservative Thought

Conservative theorists repeatedly decry the application to society and politics of a mode of thought that they characterize as overly abstract, rationalistic, and cut off from experience. Whether termed 'the abuse of reason' (by Burke 1790), 'rationalism in politics' (by Oakeshott, 1962) or 'constructivism' (by Hayek), the recurrent conservative accusation against liberal and radical thought is that they are said to depend on a systematic, deductivist, universalistic form of reasoning that fails to account for the complexity and peculiarity of the actual institutions that they seek to transform.

For conservatives, the historical survival of an institution or practice – be it marriage, monarchy, or the market – creates a prima facie case that it has served some human need. The conservative emphasis on 'experience' is linked to the assumption that the historical survival of an institution or practice is evidence of its fitness in serving human needs. Since custom and habit are important features of human conduct, some of the usefulness of a practice comes from the fact that those engaged in it are already 'used' to it, and are apt to be discomfited by change.

Conservative thought has typically emphasized the imperfection of the individual, an imperfection at once biological, emotional, and cognitive. Conservatives typically contend that human moral imperfection leads men to act badly when they act upon their uncontrolled impulses, and that they require the restraints and constraints imposed by institutions as a limit on subjective impulse. Conservatives, thus, are skeptical of attempts at 'liberation.' Conservatives have also stressed the cognitive element of human imperfection, insisting on the limits of human knowledge, especially of the social and political world. They warn that society is too complex to lend itself to theoretical simplification, and that this fact must temper all plans for institutional innovation.

These assumptions explain the emphasis of conservative social and political thought on institutions, that is, patterned social formations with their own rules, norms, rewards, and sanctions. While liberals typically view with suspicion the restraints and penalties imposed on the individual by institutions, conservatives are disposed to protect the authority and legitimacy of existing institutions because they believe human society cannot flourish without them. Unlike liberals, who favor voluntary, contractual social relations, conservatives emphasize the ongoing importance of nonvoluntary duties, obligations, and allegiances.

Although there is no necessary link between conservatism and religious belief, conservatives have tended to affirm the social utility of religion. They make several arguments for the utility of religion: that it legitimates the state; that the hope of future reward offers men solace for the trials of their earthly existence and thus helps to diffuse current discontent, which might disrupt the social order; and that belief in ultimate reward and punishment leads men to act morally by giving them an incentive to do so.

Conservatives typically emphasize the unanticipated negative consequences of deliberate social action. Such negative consequences, they argue, occur because reformers are unaware of the latent functions of existing practices and institutions. Reformers are insufficiently cognizant, it is said, of the contribution of the practice to the preservation or adaptation of the larger social system in which it is implicated. That contribution may be unintended by those engaged in the practice, so that its function may be unrecognized, or recognized only retrospectively, once the reform of the practice has brought about negative unintended consequences.

Together, the concepts of unanticipated negative consequences, latent functions, and functional interdependence

serve as recurrent arguments against radical or wholesale reform, although not against reform as such.

Recurrent substantive themes of conservative social and political thought include:

1. a skepticism regarding the efficacy of written constitutions, as opposed to the informal, subpolitical, and inherited norms and mores of society; for conservatives, the real 'constitution' of society lies in its historical institutions and practices, which are inculcated primarily through custom and habit;
2. the central role of cultural manners and mores in shaping character and restraining the passions, and hence the political importance of the social institutions in which such manners and mores are conveyed;
3. an emphasis on the family as the most important institution of socialization, and despite considerable divergence among conservatives over the proper roles of men and women within the family, the assertion that some degree of sexual division of labor is both inevitable and desirable;
4. the legitimacy of inequality, and the need for elites, cultural, political, and economic; security of possession of property as a prime function of the political order; and
5. the importance of the state as the ultimate guarantor of property and the rule of law, and hence the need to maintain political authority.

Historical Overview

Terms related to 'conservative' first found their way into political discourse in the title of a French weekly journal, *Le Conservateur* founded in 1818 by François-René de Chateaubriand with the aid of Louis de Bonald. "The *Conservateur* upholds religion, the King, liberty, the Charter and respectable people (*les honnêtes gens*)," it proclaimed in its first issue. In England, the label 'Conservative Party' was first applied to the Tories by the publicist John Wilson Crocker in 1830. In Germany, the term 'conservative' came into use later in the same decade: it was first employed as an epithet against those designated as conservative, who were accused of seeking to preserve existing institutions at any price, but the term was eventually accepted by those it had been used to stigmatize. But as with many other historical developments, the phenomenon of conservatism long preceded the use of the term in political life.

The precursors of conservatism may be found in the Anglican critique of the Puritan contention that the elect or the inspired congregation, guided by their individual interpretations of the Bible, were entitled to exercise political authority. The thought of David Hume marks a watershed in the development of conservative social and political thought into a coherent, secular doctrine. Hume began by borrowing and expanding upon this critique of the politics of religious 'enthusiasm.' And he went on to criticize what he saw as its secular counterparts in the philosophically implausible and politically subversive doctrines of natural rights and of voluntary contract as the sole legitimate basis of political obligation. In 1757, Edmund Burke used similar terms to attack what he called the 'abuse of reason' among some enlightened thinkers. On the Continent, conservative social and political thought arose as a critique of the policies of enlightened absolutism. Among the foremost practitioners of this criticism was Justus Möser, himself a part of the German Enlightenment, and a reformer of a conservative and corporatist tinge. In Holland, another Enlightenment figure, Jean de Luzac, articulated politically conservative arguments opposing the claims of 'patriots' for a more democratic constitution. The contention that conservatism arose in opposition to the Enlightenment and to the French Revolution reflects Edmund Burke's polemical characterization of the Enlightenment in his critique of the French Revolution. Although frequently reiterated, the contention is historically untenable. Conservatism arose not against the Enlightenment but within it. Much of post-1789 intellectual conservatism is continuous with the analytic strategies, if not with the tone, of prerevolutionary conservative analysis.

It was the democratic radicalism of the French Revolution that evoked Burke's *Reflections on the Revolution in France* (1790), which became the most influential work in the history of conservative thought. Joseph de Maistre (1753–1821) and Louis Gabriel de Bonald (1754–1840) were the two leading figures in the tradition of French counterrevolutionary thought. The task with which they saw themselves confronted was rather different than the one that had motivated Burke's *Reflections*, and the nature of their conservatism varied accordingly. Burke's primary task was to conserve a British order that was substantially intact, although subject to intellectual attack. He could thus stress the historical utilitarian arguments that the existence of an institution creates a prima facie case that it has served some human need, and that historical continuity is intrinsically valuable because it increases the emotional hold of an institution on its members. Similarly, Burke could underscore the importance of habit and custom. By the time de Bonald and de Maistre came to write their major works, the forces of the French Revolution had already overturned or transformed many of the key institutions of the Old Regime. The counterrevolutionary conservatives, therefore, were forced by circumstances to offer a rationale other than that of historical continuity, for the institutions of the old regime had been disrupted by the Revolution. de Maistre responded to the revolution by carrying to new lengths previous conservative arguments of epistemological modesty, historical particularism, and the unanticipated consequences of deliberate action.

Burke's defense of monarchy, aristocracy, and the established church stamped European conservative thought at the end of the eighteenth century and in the early decades of the nineteenth century. Yet a similar critique of democratic radicalism appears, with a more republican but, nevertheless, recognizably conservative thrust, in the United States, first in the writings of the US Federalists, and later in the work of conservative Whig critics of Jacksonianism, such as Rufus Choate. Hume's combination of political conservatism with the championing of commerce was to remain characteristic of subsequent British and US conservatism. By contrast, the French, Catholic strains of conservatism of de Bonald and de Maistre, as well as the romantic German strains of conservatism remained less accepting of capitalist economic development than the British and US strains.

The development of conservative social and political thought after the mid-nineteenth century shows a movement from substance to function: from the defense of particular institutions to the defense of institutions in general, from the defense of the landed aristocracy to a defense of elites in general, from an emphasis on the role of the established Church to the function of culture in linking the individual to communal purposes, and from monarchical authority to the authority of the state in general. In the last decades of the nineteenth century, conservatism underwent an important shift, reflecting a change in opponents and in institutional substance. The increasing dominance of industrial capitalism, the expansion of the suffrage, the rise of socialist movements, and the rise of a new brand of liberalism that was more economically redistributionist and more culturally permissive, all led to a substantial transformation of conservatism. Many of the economic institutions and policies once associated with European liberalism (*see* Liberalism: Political Doctrine and Impact on Social Science) now became that which conservatives sought to conserve. This new brand of conservatism was pioneered in late nineteenth-century Britain, and is reflected in the works of James Fitzjames Stephen and W.H. Mallock and by the US theorist, William Graham Sumner.

The rhetoric by means of which institutions were defended also changed. It placed less emphasis on the veneration of tradition as such, while seeking to increase the emotional hold of existing institutions by showing that they were in accord with the needs of historical development. In England and the United States, the language of common law and of inherited tradition was increasingly abandoned in favor of the language of science. The justification of the social utility of landed elites was replaced by a more general defense of the need for elites in the realms of economics and politics, while the defense of the established Church was transformed into a defense of cultural elites. In Germany, too, the substance of conservatism was transformed, from a defense of aristocracy, a social order based on fixed estates, and paternalist monarchism, toward an emphasis on the functions of the state as such, a transformation exemplified in the work of the twentieth-century German theorist, Carl Schmitt.

One recurrent strand of thought that is related to conservatism, yet politically and analytically distinct from it, is radical conservatism, a strand particularly powerful in interwar Europe, where its adherents often came to favor fascist regimes (*see* National Socialism and Fascism). Radical conservatives share some of the concerns of more conventional conservatism, such as the need for institutional authority and continuity with the past. But they maintain that the processes characteristic of modernity have destroyed the valuable legacy of the past for the present, so that a restoration of the purported virtues of the past demands radical or revolutionary action. Hence the credo of one radical conservative, Moeller van den Bruck, that 'Conservative means creating things that are worth conserving.' Radical conservatism shares with conservatism an emphasis on the role of institutions in providing restraint and direction to the individual, but seeks to create new institutions that will exert a far stronger hold on the individual than do existing ones, which because of their relative tolerance are perceived by radical conservatives as 'decayed.'

After the defeat and discrediting of fascism, conservatism in the later half of the twentieth century was defined not only by its opposition to communism but also by antipathy to the spread of the welfare state and attempts to bring about economic redistribution, and by skepticism regarding the solution of social problems through massive governmental action.

The study of conservative movements and conservative ideology provide an ongoing challenge to social scientists. For they must provide analyses of a phenomenon that is dynamic and changes over time, that varies between nations and often between regions within nations, and that often appears in combination with a variety of other trends (religious, nationalist) and under a variety of nomenclatures. These difficulties – together with the political predilections of the majority of social scientists, who tend to be antipathetic to conservatism – have led to a gap between the historical significance of conservatism and the extent of its exploration by scholars. This is lamentable, for more-or-less conservative parties have dominated western European and US politics for substantial portions of the past two centuries, and conservative thinkers have played a leading role in the development of modern thought.

See also: Liberalism, Anthropology of; Liberalism: Historical Aspects; Mannheim, Karl (1893–1947); Tradition: Social.

Bibliography

Bogus, C.T., 2011. Buckley: William F. Buckley Jr. and the Rise of American Conservatism. Bloomsbury Press, New York.
Burke, E., 1790. Reflections on the Revolution in France. J. Dodsley, London.
Deutsch, K.L., Fishman, E.M., 2010. The Dilemmas of American Conservatism. University Press of Kentucky, Lexington, KY.
Dorey, P., 2011. British Conservatism: The Politics and Philosophy of Inequality. I.B. Tauris, London.
Epstein, K., 1966. The Genesis of German Conservatism. University of Princeton Press, Princeton, NJ.
Farber, D.R., 2010. The Rise and Fall of Modern American Conservatism: A Short History. Princeton University Press, Princeton.
Frohnen, B., Beer, J., Nelson, J.O., 2006. American Conservatism: An Encyclopedia. Del: ISI Books, Wilmington.
Greiffenhagen, M., 1971. Das Dilemma des Konservatismus in Deutschland. R. Piper, Munich.
Huntington, S., 1957. Conservatism as an ideology. American Political Science Review 51, 454–473.
Kirk, R., 1953. The Conservative Mind. Regnery, Chicago.
Mannheim, K., 1927. Das Konservative Denken reprinted. In: Wolff, K.H. (Ed.), 1964 Mannheim, Wissenssoziologie Neuwied, Luchterhard; Translated as Conservative Thought. Oxford University Press, New York, pp. 132–222, 1971. In: Wolff, K.H. (Ed.), From Karl Mannheim.
Muller, J.Z., 1987. The Other God that Failed: Hans Freyer and the Deradicalization of German Conservatism. University of Princeton Press, Princeton, NJ.
Muller, J.Z., 1997. Conservatism: An Anthology of Social and Political Thought from David Hume to the Present. University of Princeton Press, Princeton, NJ.
Nardin, T., 2001. The Philosophy of Michael Oakeshott. Pennsylvania State University Press, University Park, PA.
Neill, E., 2010. Michael Oakeshott. Continuum, New York.
Nisbet, R., 1986. Conservatism: Dream and Reality. University of Minnesota Press, Minneapolis, MN.
O'Sullivan, L., 2003. Oakeshott on History. Exeter. Imprint Academic, England.
Oakeshott, M., 1962. Rationalism in Politics and Other Essays. Methuen, London.

Oakeshott, M., 2011. The Vocabulary of a Modern European State: Essays and Reviews 1953–1988. Andrews UK, Luton.
Oakeshott, M., O'Sullivan, L., 2007. The Concept of a Philosophical Jurisprudence: Essays and Reviews 1926–51/Michael Oakeshott. In: O'Sullivan Exeter, Luke (Ed.). Imprint Academic, England; Charlottesville, VA.
Oakeshott, M., Nardin, T., O'Sullivan, L., 2011. Lectures in the History of Political Thought. Andrews UK, Luton.
Podoksik, E., 2012. The Cambridge Companion to Oakeshott. Cambridge University Press, Cambridge.
Quinton, A., 1978. The Politics of Imperfection: The Religious and Secular Traditions of Conservative Thought in England from Hooker to Oakeshott. Faber and Faber, London.
Quinton, A., 1993. Conservatism. In: Goodin, R.E., Pettit, P. (Eds.), A Companion to Contemporary Political Philosophy. Blackwell, Oxford, UK.
Robin, C., 2011. The Reactionary Mind: Conservatism from Edmund Burke to Sarah Palin. Oxford University Press, New York.
Scruton, R., 1991. Introduction: what is conservatism. In: Scruton, R. (Ed.), Conservative Texts: An Anthology. Macmillan, Houndmills, Basingstoke, UK.
Sirinelli, J.-F., 1992. In: Histoire des Droites en France, 3 Vols. Gallimard, Paris.
Skocpol, T., Williamson, V., 2012. The Tea Party and the Remaking of Republican Conservatism. Oxford University Press, Oxford.
Stern, F.R., 1961. The Politics of Cultural Despair. University of California Press, Berkeley, CA.
Thompson, M., 2007. Confronting the New Conservatism: The Rise of the Right in America. New York University Press, New York.
Vierhaus, R., 1968. Conservatism. In: Wiener, D.P. (Ed.), Dictionary of the History of Ideas, vol. 1. Scriber, New York, pp. 477–485.
Vierhaus, R., 1982. Konservativ, konservatismus. In: Brunner, O., Conze, W., Koselleck, R. (Eds.), Geschichtliche Grundbegriffe. Historisches Lexikon zur Politisch-sozialen Sprache in Deutschland, vol. 3. Klett-Cotta, Stuttgart, Germany, pp. 531–565.
Wolff, K.H. (Ed.), 1964. Wissenssoziologie. Luchterhand, Berlin.

Conservatism: Theory and Contemporary Political Ideology

Paul E Sigmund, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 4, pp. 2628–2631, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The origins and development of conservatism as political theory and political ideology are reviewed. While self-consciously conservative political parties could be found in England as early as 1832, a self-consciously conservative movement in the United States only emerged in the 1950s. A common hostility to Communism and the Soviet Union linked the advocates of religious and moral traditionalism with defenders of free enterprise and opponents of big government. Today, conservatism is identified mainly with resistance to change.

Edmund Burke, writing in his *Reflections on the Revolution in France* (1790) first expressed the basic principles of conservatism; he argued for authority and tradition over revolutionary transformation, for hierarchy and deference over equality, for the community over the individual, local control over state centralism, and religious and moral values as opposed to relativism and anticlericalism. His views were taken up by the British Tory Party when it renamed itself the Conservative Party following the 1832 Reform Act and were echoed by similar parties elsewhere in Europe. A self-consciously conservative movement in the United States only emerged in the 1950s. A common hostility to Communism and the Soviet Union linked the advocates of religious and moral traditionalism with defenders of free enterprise, and opponents of big government. They were joined in the 1970s by the neoconservatives, a group of ex-liberal intellectuals who opposed the radicalism of the 1960s and the anti-Americanism of some of the opponents of the Vietnam War. The Religious Right, reacting to Supreme Court decisions outlawing school prayer and upholding a constitutional right to abortion provided a broader populist basis for the movement. Under the leadership of Ronald Reagan conservatives won the presidency in 1980 and proceeded to enact their agenda in domestic and foreign policy. The end of the Cold War led to increasing tension between social conservatives who called for increased support for religious and moral values and economic conservatives who favored free enterprise, local control, and a diminished role for the federal government. However, the continuing human ambivalence about the pace and nature of change ensure that the liberal–conservative debate will continue for the foreseeable future.

Definition

An initial problem that arises in attempting to define conservatism is whether it is a theory or simply an attitude towards change. Most dictionary definitions incline toward the attitudinal definition, describing conservatism as a preference for existing institutions, and opposition to change. With this definition the content of conservatism will vary depending on whatever is the current institutional and ideological status quo. Yet the writers, parties, and movements that have described themselves as conservative share a number of common values and doctrines, which typically include theories of human nature, history, society, and politics.

The term 'conservatism' was first used in England after the 1833 Reform Act when the Tory Party, which had roots going back to the late seventeenth century, renamed itself the Conservative Party. The central doctrines of conservatism were expressed before the word was invented, most notably in Edmund Burke's *Reflections on the Revolution in France* (1790), where we can find most of the central doctrines of conservatism, developed in response to the liberal and radical thinking of the French revolutionaries. In opposition to their demand for liberty, Burke stressed the values of order, tradition, and gradual change. Against the revolutionaries' call for equality, Burke stressed society's need for hierarchy, deference, class differentials, and a special position for the church and the nobility. In opposition to the radical individualism of the French, he insisted on the importance of community, tradition, and history. He criticized the anti-institutional rationalism of the revolutionaries by stressing the importance of social institutions that have stood the test of time, and the weakness and fallibility of individual reason. Against their centralizing statism, he lauded the 'little platoons' of localism as defenders of liberty, and contrary to their anticlericalism, he praised the elevating effect of inherited religious beliefs.

In developing these ideas Burke was able to draw on the Western tradition, classical ideas of natural law, Christian views of hierarchy and man's inherited inclination to evil, as well as medieval practices of localism and organic corporate groups. Although himself an adherent of the Whig Party, his ideas were influential not only on the thinking of the British Conservative Party, but on conservative parties and movements elsewhere in Europe, and later in the United States. As originally expressed, their emphasis on order over freedom, hierarchy over equality, religion over anticlericalism, localism over centralism, and community over the individual stood in marked contrast to the liberal doctrines that they criticized.

European Conservatism

In the political struggles of nineteenth and twentieth centuries, European conservative parties drew on the conservative

tradition to defend the interests of property, the upper classes, the military, and the church, and appealed to nationalist and patriotic feelings in support of their programs. Their efforts were not restricted to the defense of class and privilege. In 1867, hoping to secure lower class support for 'tory democracy,' Benjamin Disraeli, the Conservative Prime Minister in England, extended the suffrage. Otto von Bismarck initiated a program of social insurance in Germany to counteract socialist appeals; and Pope Leo XIII in his 1891 encyclical, *Rerum Novarum*, defended the living wage for the worker, and promoted the organization of trade unions (preferably under church auspices) while denouncing the extremes of liberalism and socialism. Less laudably, conservative parties in Germany, Austria, and elsewhere, in the twentieth century, sometimes allied themselves with or supported extreme rightists and even fascists and Nazis, seeing them as a bulwark against Communism. Hitler, for example, was able to come to power legally because of the votes of German conservatives.

The conservative impulse continues to be important in European politics. Since the 1833 Reform Act the British Conservative Party has been in power much longer than any other party. On the continent Christian Democratic centrist parties often work together with conservative groupings, linked by their commitment to religion, and to a 'social market economy.' Twentieth-century conservatives, despite earlier hostility to market forces, have increasingly embraced free enterprise economics. They have been critical of European economic and political integration on nationalist grounds. Nationalist resistance to foreign immigration also accounts for some of the support for conservative parties and movements.

Conservatism in America

In *The Liberal Tradition in America*, published in 1955, Louis Hartz argued that with the exception of a few Southerners, notably John C. Calhoun, Americans were all Lockean liberals, and that European-style conservatism was absent from the American political scene (Hartz, 1955). Ironically at the very time that his book was published, a genuine and self-conscious conservative movement had begun to emerge on the American scene. Russell Kirk had published *The Conservative Mind* (Kirk, 1953) arguing that Burke rather than Locke should be the model for America. At the University of Chicago, the German émigré political theorist, Leo Strauss, published *Natural Right and History* (Strauss, 1953), which attacked Locke as a proto-Hobbesian and denigrated the natural-rights tradition in favor of classical models of politics, especially that of Plato. Another émigré, Erich Voegelin, appeared on the cover of *Time* magazine after he published *The New Science of Politics* – in which he associated communism, liberalism, and revolutionism with a Christian heresy, gnosticism, that attempted to establish heaven on earth (Voegelin, 1952). Whittaker Chambers, who had accused Alger Hiss of spying for the Soviet Union, also linked Communism and liberal utopianism in his book, *Witness* (Chambers, 1952). In 1955 William Buckley, the *enfant terrible* who had earlier published *God and Man at Yale*, an attack on the dominance of secular and leftist views at his *alma mater*, founded *The National Review* as a journal to promote conservative political and economic thought. Combining militant anti-Communism, the defense of traditional values, and the promotion of free enterprise economics, it became the principal organ of the conservative movement in America.

Together these publications formed the intellectual basis of the conservative movement in mid-twentieth-century America. As in the case of William Buckley, their conservatism was also linked to free-enterprise economics, as defended in the writings of two University of Chicago economists, Friedrich Hayek and Milton Friedman. In *The Road to Serfdom* (Hayek, 1944) and later works, Hayek argued that the growth of government threatened liberty and stifled the economy, while Friedman's 1955 lectures, published as *Capitalism and Freedom* (Friedman, 1962) developed similar themes, as well as making innovative proposals to increase competition and privatize government functions, many of which were adopted in later decades. Their arguments were based on classical economic liberalism, but in the American context the two economists were viewed as conservative since they defended the existing capitalist system. In addition, their common hostility to Communism linked them to the social and political conservatives in the newly burgeoning movement.

The conservative movement became a major factor in national politics with the presidential candidacy of Barry Goldwater in 1964. After Goldwater's defeat, Ronald Reagan became the movement's leader. Promising to revitalize national defense against the Soviet Union, restore moral values, and downsize the welfare state and reduce the scope of the federal government, he led the conservative wing of the Republican party to victory in the presidential election of 1980. In the late 1970s, the conservative movement received support from a new group, the Neo-Conservatives, former liberal intellectuals who criticized what they perceived to be liberal weakness in the face of the Communist threat, and rejected the excesses of the New Left counterculture that had emerged in the 1960s (Kristol, 1983). They provided key policy makers to the Reagan administration in the areas of defense and education, while the free market economists directed economic policy, cutting taxes and welfare programs.

The Religious Right

Conservatism in America had often been an elite phenomenon, but the Religious Right gave it a broader populist base. Building on the religious opposition to the Supreme Court decisions outlawing school prayer and recognizing a constitutional right to abortion, conservative churchmen, mostly but not exclusively evangelical Protestants, organized the Moral Majority, and later the Christian Coalition, to reverse what they saw as a liberal tendency to secularism and toleration of immorality. They also shared with other conservatives opposition to the Soviet Union, and a belief in free market economics. In the 1990s, they opposed the gay rights movements, and the increasing toleration of homosexuality in America. They became an important pressure group within the Republican Party, effectively vetoing candidates for national office who favored abortion, and pressing for the appointment of Supreme Court justices who were sympathetic to their views.

Tensions within Conservatism

American conservatism contains within it many different currents. Ronald Reagan and the Soviet threat could hold its various strands together, but with the end of the Cold War there were increasing tensions among them, in particular as to whether social and moral issues, or downsizing government and encouraging free markets, should be given priority in public policy. Concern about moral decline, a frequent conservative theme (Bennett, 1993) and issues such as abortion, gay rights, and public prayer became the subject of *Culture Wars* (Hunter, 1991) from local school districts to the Supreme Court.

The members of the Supreme Court, most of them appointed by Republican presidents, became deeply divided along conservative–liberal lines, both as to theories of constitutional interpretation – 'original position' versus 'the living constitution,' federal-state relations – and as to the rights of the individual against those of society. Decisions by a 5–4 majority were frequent, and the question of the ideological sympathies of future Supreme Court appointees was an issue in the 2000 presidential campaign.

The Future of Conservatism

A line in a Gilbert and Sullivan operetta asserts that 'every little boy and girl alive, is either a liberal or a conservative.' It seems that conservatism, in the sense of a reluctance to make rapid and wholesale change, is related to a continuing need in the human psyche. It also expresses a general set of preferences about politics and society that undergirds the ideology and program of many political parties and movements. While there may be a broad ideological consensus on the values of democracy and market economics, there will continue to be debate about the relation of the individual to society, the respective demands of freedom and order, the extent to which equality is necessary or desirable, the proper allocation of roles to state and non-state actors, and the need for mediating institutions and local self-government. As long as this is the case, a future role for conservativism as a theory and ideology seems assured.

Editor's Note

American conservatism did not die with Ronald Reagan, of course. For an update on American conservatism since the publication of Sigmund's entry, see Charles W. Dunn, 2014.

See also: Conservatism: Historical Aspects; Liberalism: Historical Aspects; Orthodox Churches; Religion and Politics in the United States; Religious Fundamentalism: Cultural Concerns; Sociological Traditions; The American Conservative Movement; Tradition, Anthropology of; Tradition: Social.

Bibliography

Bennett, W.J., 1993. The Book of Virtues. Simon and Schuster, New York.
Burke, E., 1984. Reflections on the Revolution in France. Viking Penguin (1790), New York.
Chambers, W., 1952. Witness. Random House, New York.
Dunn, Charles W. (Ed.), 2014. The Future of Conservatism: Conflict and Consensus in the Post-Reagan Era. Open Road Media.
Friedman, M., 1962. Capitalism and Freedom. University of Chicago Press, Chicago.
Hartz, L., 1955. The Liberal Tradition in America. Harcourt Brace, New York.
Hayek, F.A., 1944. The Road to Serfdom. Routledge, London.
Hunter, J.D., 1991. Culture Wars. Basic, New York.
Kirk, R., 1953. The Conservative Mind. Regnery, Chicago.
Kristol, I., 1983. Reflections of a Neo-conservative. Basic, New York.
Nash, G., 1976. The Conservative Intellectual Movement in America. Basic, New York.
O'Gorman, F., 1986. British Conservatism. Longman, New York.
Strauss, L., 1953. Natural Right and History. University of Chicago Press, Chicago.
Voegelin, E., 1952. The New Science of Politics. University of Chicago Press, Chicago.

Consociationalism

Rudy B Andeweg, Leiden University, Leiden, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Consociationalism or consociational democracy means that the destabilizing effects of deep social cleavages can be offset by cooperation between subcultures at the elite level. Consociationalism is characterized by grand coalition, segmental autonomy, proportionality, and minority veto. Consociational theory sometimes is equated with the publications of its most influential scholar, Arend Lijphart. The theory is used empirically to explain stable democracy in deeply divided societies (e.g., the Netherlands, Lebanon) and normatively to prescribe a solution when democracy is at risk because of social segmentation (e.g., Northern Ireland, South Africa). The theory sometimes is criticized for its classification of particular countries, the operationalization of its key concepts, and the identification of factors that are favorable to the emergence of consociationalism. A recent offshoot of consociationalism is consensus democracy, which is defined institutionally rather than behaviorally, and also applies to countries without deep social cleavages.

The term 'consociational' dates back at least to 1603, when the German scholar and statesman Althusius used it to denote a polity consisting of an association of several societies. In modern political science, "Consociational democracy means government by elite cartel to turn a democracy with a fragmented political culture into a stable democracy" (Lijphart, 1969: p. 216). Although this situation has been labeled in a variety of ways ('Politics of Accommodation,' 'Power-sharing,' '*Pacificatie-democratie*' (in Dutch), '*Proporzdemokratie*' or '*Konkordanzdemokratie*' (in German)), consociational democracy (or consociationalism) is the term that is most widely used. Consociationalism is used empirically to explain stable democracy in deeply divided societies and normatively to prescribe a solution to societies where democracy is at risk because of subcultural conflict.

The Consociational Explanation

Deep social cleavages, whether they are based on ideology, religion, class, language, or ethnicity, generally are regarded as unfavorable conditions for stable democracy. The risk is limited when cleavages are crosscutting (e.g., when members of a particular ethnic group belong to different social classes), because this is supposed to generate cross pressures which have a moderating effect on social conflicts. The risk is greatest when social cleavages coincide and reinforce each other (e.g., when all members of a particular ethnic group belong to one social class, and all members of another ethnic group to another). However, there are democracies that maintain political stability despite coinciding cleavages. In the late 1960s and early 1970s, case studies of several such countries revealed a common pattern: social conflict was not moderated by crosscutting cleavages at the mass level, but by cooperation between the social segments at the elite level (e.g., Lehmbruch, 1974; Lijphart, 1975; Steiner, 1974).

By looking at both the mass level and the elite level, a fourfold typology of democratic regimes was developed, as shown in Figure 1. The situation of a homogeneous society and competing elites is most conducive to stable democracy. The United States and the United Kingdom serve as examples of such 'centripetal democracy.' One might argue that the combination of social homogeneity and elite cooperation guarantees stability even more, but the lack of democratic choice in such a 'depoliticized democracy' is likely to lead to antisystem opposition. The situation that is most dangerous for stable democracy is where society is divided deeply and the leaders of the social segments compete. Classic examples of this 'centrifugal democracy' are Weimar Germany and the French Fourth Republic. Consociational democracy represents the situation in which deep social divisions are prevented from destabilizing democracy by elite cooperation.

The list of countries that are claimed to be or to have been consociational democracies gradually has expanded and, without claiming to be complete, now includes Austria (1945–66), Belgium (from 1918), Bosnia-Herzegovina (since 1995), Burundi (since 2005), Canada (1840–67), Colombia (1958–74), Cyprus (1960 and 1963), Czechoslovakia (1989–1993), India (1947 to the late 1960s), Lebanon (1943–75 and after 1989), Luxembourg, Malaysia (1955–1969 and since 1971), Macedonia (since 2000), the Netherlands (1917–67), Northern Ireland (after 1998), South Africa (1993–96), and Switzerland (since 1943), with contemporary Canada and Israel being classified as semiconsociational. The extent to which the political system of the European Union can be described as consociational is debated (e.g., Bogaards and Crepaz, 2002). As can be seen from the dates, consociationalism has disappeared in some cases, either because it failed (sometimes because of international interference: Lebanon, Cyprus) or because it was so successful in accommodating the social cleavages that it rendered

		Structure of society	
		Homogeneous	Plural
Elite behavior	Coalescent	Depoliticized democracy	Consociational democracy
	Adversarial	Centripetal democracy	Centrifugal democracy

Figure 1 Lijphart's typology of democratic regimes. Reproduced with permission of Yale University Press from Lijphart, A., 1977. Democracy in Plural Societies: A Comparative Explanation, p. 106, Figure 2.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.93025-3

itself superfluous (e.g., in Austria and the Netherlands). Because of the latter cases, consociationalism is sometimes seen as a transitional phase.

Although there have been deductive approaches to consociationalism (e.g., Nordlinger, 1972), the theory of consociational democracy has been developed primarily by generalizing from case studies (e.g., Lijphart, 1977; Taylor, 2009). Lijphart's contribution has been so influential that consociationalism often is equated with his work. He distinguishes four characteristics, each of which should be present for a country to be classified as consociational: (1) government by grand coalition (i.e., including representatives of all social segments); (2) segmental autonomy (federal arrangements in case of geographical concentration of subcultures (*see* Federalism), or subcultural autonomy with regard to schools, etc.); (3) proportionality in the electoral system (*see* Proportional Representation), but also with regard to civil service appointments, distribution of public funds for schools, etc.; and (4) minority veto (sometimes in the form of qualified majority requirements for certain types of legislation).

In a later offshoot of consociationalism, Lijphart elaborated a distinction between majoritarian democracy, which is characterized by 10 institutional features that concentrate power in the hands of the majority, and consensus democracy, which is characterized by an identical number of features that make decision making as inclusive of all relevant parties or groups as possible (Lijphart, 2012). Although the characteristics of consociational democracy are primarily behavioral whereas the characteristics of consensus democracy are largely institutional, and although consociationalism is also defined by the existence of deep social cleavages, both concepts clearly overlap and many consociational democracies are also classified as consensus democracies.

The Consociational Prescription

Consociational engineering is recommended as the only viable option for deeply divided countries that seek to achieve or maintain stable democracy. Two prominent cases in which this advice appears to have been influential are Northern Ireland and South Africa. The 1998 Good Friday Agreement in Northern Ireland, which was endorsed by referendums in Northern Ireland and in the Irish Republic, contains several clearly consociational characteristics (dual premiership, a proportional electoral system (STV), requirements for 'parallel consent' in the legislature, etc.). In South Africa, the transition from apartheid was facilitated by a consociational interim constitution (including a government of national unity and an electoral system of proportional representation), but in the constitution of 1996 few consociational elements remain.

The prescriptive use of consociationalism has sparked a debate on the conditions that are favorable and unfavorable to the successful adoption of a consociational solution (e.g., Andeweg, 2000; O'Leary, 2005). It is more likely that elites refrain from intersegmental competition if they perceive the benefits of competition to be outweighed by its costs. This is the case when no social segment has a majority or can entertain any hope of achieving a majority. However, even if one segment has a (near) majority, its leaders may still prefer to share power with the minority if that minority has inflicted casualties or economic damage on the majority or can credibly threaten to do so, or if the minority's support is indispensable to face an external threat. In addition, it is more likely that a consociational solution succeeds when followers agree with their leaders' choice of intersegmental cooperation, or when the elites are relatively immune from intrasegmental competition. The latter is the case when elites control the mass media, or when dense networks of subcultural organizations or patron–client networks make it difficult for competitors to challenge the incumbent elites.

Some other factors that are mentioned as favorable to the emergence of consociationalism are less self-evident (social segments of equal size, a small country, geographical concentration of segments, etc.). An important remaining question is whether some social cleavages are more amenable to elite accommodation than others. In view of consociationalism's potential for deeply divided countries, the question whether ethnic (as opposed to religious or class) cleavages constitute a favorable or an unfavorable condition is of particular relevance. Another open question is the extent to which consociational arrangements can be imposed successfully by outside actors (other countries, international organizations). Cases of coercive consociationalism, or at least outside involvement, include Northern Ireland, Macedonia, Bosnia-Herzegovina, and Burundi, but there are also examples of failure (Cyprus, Iraq).

Critiques of Consociationalism

The central tenet of consociationalism (i.e., that political stability need not be safeguarded by cross-cutting cleavages at the mass level, but can also be maintained by intersegmental elite cooperation) has gone relatively unchallenged. In addition to controversy over favorable and unfavorable factors, most of the criticisms that have been passed on consociationalism as an explanatory theory are directed at the classification of some countries as consociational democracies, and at the way in which key concepts have been defined and operationalized, in particular by Lijphart (e.g., Barry, 1975; Lustick, 1997). The country whose consociational credentials have been called into question most is Switzerland. The debate centers on the question whether Swiss society is deeply divided, as religious and class cleavages cross-cut linguistic divisions, and as the linguistic segments had very few ambitions that brought them into conflict with one another.

Such controversies, it is argued, arise from the vagueness and elasticity with which many of the key concepts of consociationalism such as segmentation are defined (Halpern, 1986). For example, one of consociationalism's four characteristics, grand coalition, includes not only all-party governments (as in Switzerland), but also oversized coalitions, all-party advisory commissions, 'diachronic coalitions' in which power rotates, as in Colombia between 1958 and 1974, or joining forces in a single 'consociational party' (Bogaards, 2014). These different mechanisms can still be seen as functional equivalents that all facilitate elite cooperation, however. A more important problem is that elites are unlikely to cooperate all the time, on all issues, and in

all arenas. Even in consociational democracies, elites tend to compete in elections, for example. Any classification of elite behavior as adversarial or coalescent therefore must rest on an often-impressionistic aggregation of elite behavior across issues and arenas.

The consociational prescription has also come under attack (see O'Leary, 2005), in particular with regard to the treatment of social cleavages. Consociationalism sees social cleavages as given obstacles to social peace, not as social constructs that might be changed (Dixon, 2012). Consociational engineering may even reinforce subcultural identities (Horowitz, 1985) and provide an incentive to incite subcultural conflict in order to obtain further concessions (Tull and Mehler, 2005). In response, a 'liberal consociationalism' that allows social identities to change is now distinguished from the original 'corporate consociationalism' (McCullogh, 2014). Another concern over the prescriptive use of consociationalism regards its democratic credentials. The democratic quality of consociationalism is sometimes impugned because of the predominance of elites and the absence of opposition. From the perspective of democracy as a device to keep elites accountable and responsive to the masses, this is a pertinent criticism. However, in the deeply divided societies for which consociationalism offers a solution, the most salient distinction is not between elites and masses, but between social segments of a more or less ascriptive nature. In that situation, democracy may be better served with a relative emphasis on inclusiveness, and in that respect consociational democracy (and consensus democracy) outperforms adversarial or majoritarian democracy.

See also: Conflict and Consensus; Democratic Theory; Majoritarianism and Majority Rule; Pluralism; Religion: Nationalism and Identity.

Bibliography

Andeweg, R.B., 2000. Consociational democracy. Annual Review of Political Science 3, 509–556.

Barry, B., 1975. Review article: political accommodation and consociational democracy. British Journal of Political Science 5, 477–505.

Bogaards, M., 2014. Democracy and Social Peace in Divided Societies. Palgrave Macmillan, Houndmills/New York.

Bogaards, M., Crepaz, M.M.L., 2002. Consociational interpretations of the European Union. European Union Politics 3, 357–381.

Dixon, P., 2012. The politics of conflict: a constructivist critique of consociational and civil society theories. Nations and Nationalism 8, 98–121.

Halpern, S.M., 1986. The disorderly universe of consociational democracy. West European Politics 9, 181–197.

Horowitz, D.L., 1985. Ethnic Groups in Conflict. University of California Press, Berkeley.

Lehmbruch, G., 1974. A non-competitive pattern of conflict management in liberal democracies: the case of Switzerland, Austria and Lebanon. In: McRae, K.D. (Ed.), Consociational Democracy: Political Accommodation in Segmented Societies. McClelland and Stewart, Toronto, pp. 70–89.

Lijphart, A., 1969. Consociational democracy. World Politics 21, 207–225.

Lijphart, A., 1975. The Politics of Accommodation: Pluralism and Democracy in The Netherlands, second ed. University of California Press, Berkeley, CA.

Lijphart, A., 1977. Democracy in Plural Societies: A Comparative Explanation. Yale University Press, New Haven, CT.

Lijphart, A., 2012. Patterns of Democracy: Government Forms and Performance in Thirty-six Countries, second ed. Yale University Press, New Haven, CT.

Lustick, I.S., 1997. Lijphart, Lakatos, and consociationalism. World Politics 50, 88–117.

McCullogh, A., 2014. Consociational settlements in deeply divided societies: the liberal-corporate distinction. Democratization 21, 501–518.

Nordlinger, E.A., 1972. Conflict Regulation in Divided Societies. Center for International Affairs, Harvard University, Cambridge, MA.

O'Leary, B., 2005. Debating consociational politics: normative and explanatory arguments. In: Noel, S. (Ed.), From Power-Sharing to Democracy. McGill University Press, Montreal, pp. 3–43.

Steiner, J., 1974. Amicable Agreement versus Majority Rule: Conflict Resolution in Switzerland. University of North Carolina Press, Chapel Hill, NC.

Taylor, R. (Ed.), 2009. Consociational Theory: McGarry and O'Leary and the Northern Ireland Conflict. Routledge, Abingdon.

Tull, D., Mehler, A., 2005. The hidden costs of power-sharing: reproducing insurgent violence in Africa. African Affairs 104, 375–398.

Conspicuous consumption and distinction, history of

Connor J Fitzmaurice, Boston University, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of 'conspicuous consumption' is the idea that consumers make explicit comparisons to others when consuming in an effort to portray actual and aspirational statuses. The concept of 'distinction' reduces the emphasis on explicit comparisons; consumer choices are seen as predicated on taken-for-granted, internalized norms of appropriate taste. These taken-for-granted ways of perceiving the world and consuming are means of marking the social boundaries of class. While status models of consumption were once dominant in the social sciences, these concepts have undergone considerable critique, particularly when applied outside the contexts in which they were initially developed.

Attention to the concepts of conspicuous consumption and distinction has played a critical role in understanding consumer behavior within the social sciences. However, the application of these concepts has not been uncontested. Of central concern is how taste shapes consumption, and the extent to which status-seeking behavior may or may not shape consumer preferences. Nevertheless, over the course of their intellectual history, these concepts – and their critiques – have provided analytic power for social scientists studying issues of consumption, status, and taste.

The role of taste in consumption has long been problematic for social scientists. For example, the models of mainstream economics rely upon the assumption of independent, fixed tastes or preferences. As George Stigler and Gary Becker explain, once a situation arises in which tastes appear not to meet these criteria of stability and independence, and ascertaining the underlying origins of taste becomes paramount, traditional economic inquiry is abandoned: for traditional economics, differences in taste are incontestable (1977). However, analytic insight is gained by paying attention to those situations in which consumer tastes change and, rather than operating independently, seem to be shaped by consumers' social positions and status-seeking behaviors. It is in such cases that social scientists have effectively appealed to the concepts of conspicuous consumption and distinction.

From Conspicuous Consumption to Distinction

The concept of 'conspicuous consumption' is the idea that consumers make explicit comparisons to others when consuming goods, services, and leisure in an effort to portray actual and aspirational statuses. Specifically, consumers are conceptualized as attempting to emulate the purchasing decisions of those above them on the status hierarchy or, at the very least, as attempting to defensively keep up with similar others. Developed later, the concept of 'distinction' reduces the emphasis on explicit comparisons. In fact, through Pierre Bourdieu's concepts of 'habitus' and 'cultural capital,' consumer choices are seen as predicated on taken-for-granted assumptions and internalized norms of appropriate taste ([1979]1984). Nevertheless, these taken-for-granted ways of perceiving the world and consuming are means of marking the social boundaries of class. Consumer choices, and the tastes that produce them, are a means of socially reproducing status differences. The following sections will trace the history of these concepts, their criticisms, and refinements.

Conspicuous Consumption

Interest in – and, often, social and moral concern with – what social scientists would now categorize as conspicuous consumption certainly predates the development of the concept. Roger Mason, for example, emphasizes early mercantilist concerns with ostentation (1998). As an academic discourse, Harvey Leibenstein notes that the Scottish economist John Rae was considering such purchasing behaviors in the first half of the nineteenth century (Leibenstein, 1950). However, it was Thorstein Veblen who popularized the theorization of such ostentation as 'conspicuous consumption' in *The Theory of the Leisure Class: An Economic Study of Institutions* ([1899]1912). Published toward the close of the American 'Gilded Age,' Veblen was writing during a period of both tremendous economic growth and growing inequality in the United States. Veblen saw waste, luxury, and pretentious displays of wealth as overarching themes of consumption under such conditions.

Andrew Trigg has noted that Veblen's theory was predicated upon this period of capitalist economic growth, pointing to Veblen's assertion that wealth generation and conspicuous consumption are linked. Once societies begin accumulating wealth, status displays through property gain significance for the maintenance of social position (Trigg, 2001). Wealth, while necessary for social status and esteem, was not enough: without displays of consumer goods and purchasing, there is no socially accepted means of proving wealth. Such public displays are necessary for two primary reasons. The first is what contemporary theorists call a moral hazard problem – there is a strong motivation to lie about how much wealth one actually possesses. The second is the existence of social norms that sanction boasting about ones wealth in an effort to gain status. As a result, transforming wealth into status goods provides a means of socially acceptable display that solves the moral hazard by providing verifiable evidence of wealth. Because these problems are enduring aspects of status-based societies, conspicuous consumption – both as a consumer practice and

a concept – has outlived the Veblenian era and context of America's Gilded Age.

The concept of conspicuous consumption fits within Veblen's larger stylized account of class society. Tracing the shifting meanings of class, status, and wealth across history, Veblen periodizes societies as 'savage,' 'barbarian,' and 'civilized.' For Veblen, distinct classes emerged during the period of 'barbarism' in feudal Europe and Asia, where wealth and status were linked to the spoils of war. However, with time, wealth came to be valued as a good in and of itself – such that inherited wealth could provide status above and beyond that attained by individual accomplishments. As a result, Veblen argued that in 'civilized' societies, displays of wealth aimed at attaining social esteem came to be the driving motivation governing individuals' practices surrounding work and consumption. In fact, he saw the desire for status as a more important determinant of consumption than even material needs. For Veblen, consumption was a thoroughly social phenomenon.

Of course, Veblen recognized that there were other potential solutions to the moral hazard problem surrounding wealth, and avoiding the social risks of boasting about wealth. Indeed, Veblen argued that where currency replaced more visible forms of wealth, the conspicuous display of a life of leisure emerged as an early solution. However, Veblen believed that humans possessed a productive instinct, such that idleness came to be seen as disreputable and was replaced by the acquisition of status goods. Gender features prominently in Veblen's account, with male heads of households seeking to maintain leisured wives who vicariously consumed lavish goods on their behalf. However, in the modern era, Veblen saw the endless cycle of competitive pressures toward conspicuous consumption as producing a societal strain. While conspicuousness began as an attempt to attain status, it increasingly took the form of a defensive effort to simply keep up.

In *The Theory of the Leisure Class*, Veblen saw consumption as 'conspicuous' in that it becomes an outward marker for economic wealth and social worth, and as a behavior driven by emulation and social comparison of taste. Of course, of central importance is the visibility of status goods. For Veblen, utilizing wealth for active status attainment only worked if the goods could become an outwardly verifiable marker one could compare with others. Emulation and comparison, while related, are two distinct arguments about the nature of consumer desire and behavior. By paying attention to emulation, Veblen suggests that consumer tastes are shaped by conscientious monitoring of similar others, rather than by stable individual preferences and rational concerns about the utility of consumer goods. For Veblen, no article will be worthy of consumption based upon its material utility alone ([1899] 1912: 157). Second, Veblen suggested that consumption was based on 'invidious' social comparisons: consumers enviously attempt to outdo similar others by looking upward to those of higher status ([1899]1912: 104). In essence, Veblen proposes a trickle-down theory of consumer tastes, in which the preferences of elites are reflected in the one-upmanship of those lower in socioeconomic status.

After Veblen's introduction of the concept of conspicuous consumption into the social science lexicon, the theory – and status-based models of consumption more generally – rose to become the predominant model for understanding consumer behavior in American sociology. For example, in the 1920s and 1930s, F. Stuart Chapin analyzed the living rooms of families' homes as measure of status. Strong reliability and validity were associated with using living rooms as a proxy measure of status, based on the Veblenian assumptions that living rooms are one of the most socially visible sections of a private residence, and that possessions serve as a reliable indicator of less visible – and, therefore, verifiable, socioeconomic status (Chapin, 1932). Even in the post-War era, Veblenian concepts of conspicuous consumption remained popularized in the work of economists such as John Kenneth Galbraith and Vance Packard.

The formulation of conspicuous consumption popularized by Veblen has undergone considerable refinement. For example, Harvey Leibenstein distinguishes between 'bandwagon,' 'snob,' and 'Veblen' effects in consumption (Leibenstein, 1950). As Leibenstein explains, the bandwagon effect refers to purchasing behavior in which consumer demand for a particular product is increased on the basis of other individuals' consumption of it. The snob effect, in turn, refers to situations in which demand for a good is decreased by others' consumption of it, as consumers seek to differentiate themselves from the masses. Finally, Leibenstein argues that Veblen effects are truly what qualifies as conspicuous consumption: such consumption occurs when individuals choose products on the basis of their higher relative price (Leibenstein, 1950: 189).

Even with such refinements, in more recent decades, the theory of conspicuous consumption has been the subject of considerable critical scholarship. Some, like Colin Campbell, have gone further by suggesting that the theory is not even testable, arising from the difficulties of matching consumer intentions with consumption choices (1995). For example, consumers may achieve the goal of status through other practices of consumption, or may make purchases for reasons unrelated to outdoing others but still end up doing so. Apart from criticisms surrounding the ability to empirically verify the concept of conspicuous consumption, Trigg has noted three distinct lines of criticism throughout its history: trends in consumption may begin at the bottom of the social hierarchy when those higher in socioeconomic status appropriate the styles of those below them; that consumption has become less conspicuous and status is now presented through avoidance of ostentation, and that models of conspicuous consumption based upon social status fail to reflect the identity projects that consumers undertake with their purchases (Trigg, 2001). While the first two critiques may very well be addressed by the acknowledgment of what Jeffrey Frankel has called the 'Bluejeans Effect' – the idea that the low price of certain goods, in and of itself, can increase demand as consumers seek to appropriate lower-class styles or demonstrate restraint (Frankel, 1975) – the postmodern critique has been more total. Douglas Holt, for example, has argued that postmodern consumption no longer rests upon acquiring consensual markers of status, but upon consumers' attempts to create lifestyles and identities (Holt, 2000).

Despite such critiques, recently some scholars have argued theoretically for the continuing analytic power of the theory of conspicuous consumption. Juliet Schor, for example, has

argued that the postmodern critique is not a sufficient basis for discarding status-based models of consumption (2007). She argues that researchers have yet to determine how independent of class position consumers' identity projects truly are, and that the presentation of a particular self-image through consumption is compatible with status models so long as consumers can read what symbolic meanings their choices will signal to others (Schor, 2007). Moreover, contemporary empirical studies continue to show Veblen effects in consumption choices. For example, Angela Chao and Juliet Schor have shown that found that women are more likely to buy expensive lipstick – since it is publicly applied – than face wash, despite the fact that price is not associated with higher material quality in lipsticks but is associated with superior face washes (1998). Moreover, Francis Bloch et al. have recently applied the concept of conspicuous consumption to the case of weddings in India, showing that a status-signaling model based on wedding size applied outside of the American context in which the concept was formulated (2004).

Despite empirical evidence that Veblen effects are still observable in consumer purchasing behaviors, Veblenian accounts of consumption remain contested. However, in many ways, the development of the concept of 'distinction' – with its inclusion of cultural capital and habitus – maintained a status-based model of consumption while addressing the critiques of intentionality.

Distinction

Amid the declining salience of Veblen's status-based account of consumer behavior in the United States, French sociologist Pierre Bourdieu introduced the concept of 'distinction' as a means of explaining the role of class in consumer and aesthetic preferences ([1979]1984). Taken alone, the idea of distinction is not all that different from the ideas behind the concept of conspicuous consumption. Both are predicated upon the idea that individuals mark their different statuses through their preferences and consumer choices. However, Bourdieu introduced the concepts of cultural capital and habitus as mechanisms through which class becomes operative in individual preferences. Fundamentally, Bourdieu asserts that cultural capital is the reserve of cultural and aesthetic knowledge an individual attains through his or her social position. Further, the habitus serves as the means through which cultural capital operates; it is the product of the internalization of such tastes and dispositions, such that preferences appear to the individual as natural and undirected, and often go unreflected upon.

These processes were of fundamental concern for Bourdieu since the ways individuals mark aesthetic differences through the classification of taste also mark differences between individuals: distinction is a means through which inequality is reproduced (Bourdieu, [1979]1984). Unlike Veblen, who was mostly concerned with conscious consumer choices made as a result of explicit comparisons, Michele Lamont and Annette Lareau explain Bourdieu's contribution of the concept of distinction as structural. In their survey of both Bourdieu's work and the earliest applications of the concept, Lamont and Lareau explain that Bourdieu's concern was not individual choices but the institutionalization of differentially valued tastes that function to reproduce social status and class (1988). Seemingly individual preferences, tastes, and dispositions are reframed as reflecting the differences between classes.

The early use of Bourdieu's concept of distinction and the related concepts of cultural capital and habitus revealed a dizzying array of applications and redefinitions, such that by 1988 – a mere 4 years after *Distinction* was published in English – Lamont and Lareau suggested the concept risked becoming meaningless (1988). Specifically, Lamont and Lareau argued that the concept of cultural capital – critical to explaining how distinction is attained through seemingly unintentional and habitual choices – had been employed to mean everything from human capital to high culture (1988). However, they argue that despite slippage in its application – and problems within the theory itself – Bourdieu's conceptualization was worth further theorization for its fundamental assertion that taste, as a distinguishing practice, provides a basis for exclusion and class reproduction (Lamont and Lareau, 1988).

As Schor (2007) has noted, Bourdieu's concept of distinction received an initial negative reception when it was published in English in 1984 in light of the declining use of status-based theories of consumption in the United States, such as Veblen's theory of conspicuous consumption. For example, Lamont and Lareau point out the frequency with which, outside of the French context in which the concept was developed, scholars were willing to dispense with the theory's fundamental concern with class reproduction in favor of a focus on status (1988). Despite recognizing value in the concept, specifically its focus on class reproduction, Lamont and Lareau themselves articulated strong critiques of distinction. For example, they argue that in complex societies, what counts as valued cannot possibly exist in a 'zero-sum' relationship to the dominated culture of lower-status individuals as a result of competing standards of evaluation. Further, they contend that lower-class standards of value can in fact operate autonomously and according to evaluative criteria opposing those of dominant, high-status groups (Lamont and Lareau, 1988). Such criticisms have allowed for more nuanced applications of the concept of distinction in recent years.

Recent refinements of the concept of distinction have come from applications of the concept outside of the French context in which the theory was developed – where scholars have argued consensually agreed upon hierarchies of status goods and tastes have become blurred. These scholars argue that there is no longer a definitive boundary between highbrow and lowbrow tastes. Instead, by the 1990s Richard Peterson and Roger Kern argued that taste and consumption were best understood as 'omnivorous' (Peterson and Kern, 1996). Such lines of research question the presumed homology between social status and taste present in the work of Bourdieu. Contrary to Bourdieu's assertion of discreet sets of classed tastes and dispositions (Bourdieu, [1979]1984), elites are recognized to consume across a much wider spectrum than in the past (Peterson and Kern, 1996). In a way, the conception of the one-dimensional consumer with tastes inextricably tied to class defies common sense; modern elite consumption is hardly as simple as a continuous stream of foie gras and Mozart. The elite no longer are conceived of as restricting their consumption to merely high culture, but are recognized as consuming across status lines.

The claims of omnivorous cultural consumption are supported by empirical work spanning a range of consumptive activities, demonstrating that the consumer seeking to portray cultural distinction is free to 'sample from a much broader cultural repertoire' (Johnston and Baumann, 2007: 200). Josée Johnston and Shyon Baumann, for example, trace the movement away from a hierarchical system of consensual status goods within the discourse of elite food consumption (Johnston and Baumann, 2007). However, the critique of the easily recognized, consensual hierarchies of status goods identified by Bourdieu as patterning consumption practices does not mean that the concept of distinction has been abandoned. Even in accounts of omnivorous consumption, distinction plays a prominent role. For Johnston and Baumann, for example, the ability of elites to consume widely across formerly discrete categories of high- and lowbrow foods and food experiences, and discern status value in a more open field of consumption, has become the new basis for status and distinction.

More recent critiques and refinements of the concept of distinction have benefited from the application of the concept outside of the European context in which it was initially developed and the American cultural milieu in which early applications occurred. Tuba Üstüner and Douglas Holt (2010), for example, suggest that studies of status consumption have been limited based on the assumption that empirical studies conducted in the United States and Europe could produce universally applicable, nomothetic models of causality in consumption. They argue that such a model merely extends the simplistic trickle-down model of status consumption advanced by Veblen, failing to integrate the more nuanced understandings Bourdieu's concept of distinction employed. While they recognize the global trickle-down model corrects for Bourdieu's assumption that status consumption occurred in cultural fields bounded by the nation-state, they seek to actively theorize how global efforts for distinction through status consumption play out in less industrialized countries. Specifically, their work suggests that some consumers must actively work to cultivate the type of habitus that will produce globally recognized forms of distinction, that cultural capital must be de-territorialized, and that the boundaries of consumption fields in a global context are inherently contested (Üstüner and Holt, 2010).

The Future of Conspicuous Consumption and Distinction

Since Veblen introduced the concept of conspicuous consumption into the social scientific lexicon at the close of the nineteenth century, theories of status consumption have maintained a prominent role in the ways social scientists conceptualize consumer behavior. The concepts of conspicuous consumption and distinction have many benefits over other models of consumption. Principally, they recognize the inherently social nature of consumption, and enable scholars to conceptualize the ways in which consumer culture helps to reproduce social inequality. However, as the histories of these theories indicate, these concepts are not without liabilities. For example, critics have rightly noted that fields of status consumption are not always based upon static, consensual hierarchies of consumer goods, that taste can operate at odds with the unidirectional trickle-down such concepts can imply, and that the concepts of conspicuous consumption and distinction often fail to account for the identity projects consumers undertake through their purchases. Likewise, the application of these concepts outside of the American and European contexts in which they were initially developed has further problematized them. While empirical studies have shown the simple applicability of status-based models of consumption in other contexts (Bloch et al., 2004), more detailed analyses have shown the very real limitations of conceptualizing particular models of distinction as universally generalizable (Üstüner and Holt, 2010).

Despite the many critiques of these status-based models of consumption across their intellectual histories, the current economic context suggests continued applicability and a chance for further scholarly refinement. The increasing stratification of society – both in the United States and globally – that resulted from the intense concentration of wealth beginning in the 1980s allowed for an expansion of markets in high-end, luxury items. While such changes have been noted by academics, they have also fueled a public discourse about consumer excess, particularly in the wake of the 2007 financial collapse. Many social commentators have begun to draw parallels to the Veblenian era in which the term conspicuous consumption first gained popularity. As in the Gilded Age, increasing economic stratification – a social process unlikely to be curbed with continued neo-liberalization and globalization of markets – may be fueling status consumption.

Whether status-based models of consumption, such as the concepts of conspicuous consumption and distinction, will regain salience in academic discourse remains to be seen. However, such conditions undoubtedly present a chance for refining such theories. Regardless, the critiques of these theories over the course of their intellectual trajectories have led to more nuanced understandings of consumer behavior – if only by serving as a foil for other approaches seeking to better understand the role of consumption in modern life.

See also: Consumption, Geography of; Consumption, History of; Consumption, Sociology of; Habitus: History of a Concept; Leisure and Cultural Consumption: The European Perspective; Leisure and Cultural Consumption: US Perspective; Social Inequality in Cultural Consumption Patterns; Veblen, Thorstein (1857–1929).

Bibliography

Bloch, F., Rao, V., Desai, S., 2004. Wedding celebrations as conspicuous consumption: signaling social status in rural India. Journal of Human Resources 48 (3), 675–695.

Bourdieu, P., [1979]1984. Distinction: A Social Critique of the Judgment of Taste. Harvard University Press, Cambridge, MA.

Chapin, F.S., 1932. Socio-economic status: some preliminary results of measurement. American Journal of Sociology 37 (4), 581–587.

Chao, A., Schor, J., 1998. Empirical tests of status consumption. Journal of Economic Psychology 19 (1), 107–131.

Campbell, C., 1995. Conspicuous confusion? A critique of Veblen's theory of conspicuous consumption. Sociological Theory 13 (1), 37–47.
Frankel, J.A., 1975. The bluejeans effect. American Economist 19 (1), 60.
Holt, D., 2000. Post-modern markets. In: Schor, J. (Ed.), Do Americans Shop Too Much? Beacon, Boston.
Johnston, J., Baumann, S., 2007. Democracy versus distinction: a study of omnivorousness in gourmet food writing. American Journal of Sociology 113 (1), 165–204.
Lamont, M., Lareau, A., 1988. Cultural capital: allusions, gaps, and glissandos in recent theoretical developments. Sociological Theory 6 (2), 153–168.
Leibenstein, H., 1950. Bandwagon, snob, and Veblen effects in the theory of consumers' demand. Quarterly Journal of Economics 64 (2), 183–207.
Mason, R., 1998. The Economics of Conspicuous Consumption: Theory and Thought since 1700. Edward Elgar, Cheltenham, UK.
Peterson, R.A., Kern, R.M., 1996. Changing highbrow taste: from snob to omnivore. American Sociological Review 61 (5), 900–907.
Schor, J., 2007. In defense of consumer critique: revisiting the consumption debates of the twentieth century. Annals of the American Academy of Political and Social Sciences 611, 16–30.
Stigler, G.J., Becker, G.S., 1977. De gustibus non est disputandum. American Economic Review 67 (2), 76–90.
Trigg, A.B., 2001. Veblen, Bourdieu, and conspicuous consumption. Journal of Economic Issues 35 (1), 99–115.
Üstüner, T., Holt, D.B., 2010. Toward a theory of status consumption in less industrialized countries. Journal of Consumer Research 37 (1), 37–56.
Veblen, T., [1899]1912. The Theory of the Leisure Class. MacMillan, New York.

Constitutional Courts

Brun-Otto Bryde, Justus Liebig University Giessen, Giessen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In an increasing number of countries, specialized courts, rather than ordinary courts, are trusted to guarantee the binding force of the constitution. These specialized courts are commonly known as Constitutional Courts. This article describes their history with special emphasis on situations of constitutional transition, compares them to other organizations of constitutional control (American model, mixed systems, and international courts) and gives an overview of their composition and jurisdiction. In addition, it discusses their role in the political system and maps their place in sociolegal studies.

Constitutional courts (CCs) are specialized courts with jurisdiction over constitutional matters. In constitutionalist systems the constitution is considered to be the highest law with normatively binding force over political actors, including the lawmaker. While one may detect antecedents in antiquity and in English common law, the modern history of the 'higher law' nature of the constitution starts with the constitutions of the American colonies and the American and French revolutions.

On a theoretical level, it is the distinction between *pouvoir constituant* (lit. constituting power) and *pouvoir constitué* (lit. constituted power) (Sièyés) that establishes the constitution's binding force over the lawmaker. In more recent constitutional theory, the constitution is also regarded as providing supreme protection for the democratic process and for human rights against parliamentary majorities. Today, most constitutionalist democracies use courts to guarantee the binding force of the constitution. While many countries follow the example invented by Marshall in *Marbury v. Madison* and give this function to the ordinary courts, an increasing number use specialized CCs.

Constitutional Courts in Historical and Comparative Perspective

Democracy without Judicial Review

Judicial control of the political system is not a necessary feature of a democratic society. In systems with 'flexible constitutions' like the United Kingdom and New Zealand, the constitution is at the disposal of the lawmaker. In Israel, however, which has a flexible constitution in this sense, the Supreme Court has developed general principles, especially the principle of human dignity, which it uses to control the lawmaker. But even in countries with 'rigid' constitutions that are not at the disposal of simple majorities, the binding force can be conceptualized as a sanctionless imperative for the lawmaker. This was, for example, the case in France until the constitution of the fifth Republic introduced a (moderate) form of constitutional control.

In other constitutionalist democratic systems, judicial review is expressly excluded (e.g., The Netherlands, Switzerland on the federal level); still others know only an organized self-control of the lawmaker during the lawmaking process (e.g., Finland until 2000). However, the prominent examples of constitutionalist democracies without any form of judicial control are disappearing rapidly. Even the bulwark of parliamentarianism, Great Britain, has introduced a (very modest) form of judicial control of Parliament with the 1998 Human Rights Act. Increasingly, some form of judicial control of the political system has become a defining element of modern democracies.

Constitutional Control as Function of Ordinary Courts: The American or Diffuse Model

The introduction of judicial control by the US Supreme Court in *Marbury v. Madison* (1803) is not based on a constitutional authorization but rather on a logical conclusion drawn from the higher law status of the constitution. This power is, therefore, shared by all courts; there are no special procedures or remedies in constitutional matters, and decisions have no *erga omnes* effect. This system (called 'diffuse' in comparative constitutional law) has been adopted widely in legal systems whose constitutions are silent on judicial review, such as the common law countries and Scandinavia. Even though this system is regularly called the 'American' system, the US Supreme Court is in many ways an atypical case: it concentrates (through the *certiorari* system) on cases with constitutional implications, its judges are selected with their constitutional role in mind, and the *stare decisis* rule – the principle that relevant precedents are to be followed – gives its judgments a *quasi-erga-omnes* effect. In a sociological sense it acts very much like specialized CCs and is, therefore, included in many comparative works on constitutional courts, although it is not one in a technical sense. This is atypical for the diffuse system because for ordinary courts (including supreme courts) in such systems (e.g., in Scandinavia) constitutional questions are only a minor part of their function and judicial review is rarely exercised. Outside the United States only a few Supreme Courts of this type (most notably the Canadian Supreme Court, the Indian Supreme Court, and the Israeli Supreme Court) have assumed an active constitutional role.

Constitutional Courts: The Austrian or Concentrated Model

In contrast to the American model, CCs are the result of conscious constitutional design. Their history is usually said

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86012-2

to start in 1920 with the Austrian Constitution, based on a draft by Hans Kelsen (even though it was preceded by the Czechoslovakian constitution by some months). The power to find laws nonconstitutional was reserved to a special constitutional organ (therefore, this system is called 'concentrated'), and in its original version only federal and state governments could appeal to the court.

It is no coincidence that CCs were invented by Kelsen. They are, on the one hand, a logical correlate to a hierarchical view of the legal system. On the other hand, Kelsen's pure theory of law tries to shield law from politics; therefore, Kelsen was an opponent of the American model. In this perspective the creation of a constitutional court was not only an institutionalization of constitutional control but at the same time was a barrier to the development of judicial review by the general judiciary, as it could only be mobilized by a limited number of state organs for a constitution that did not contain a bill of rights. Here again the name-giving case of the 'Austrian model' is not representative: the vast majority of CCs created later were given much broader powers in constitutional systems with extensive substantive provisions. From a beginning with a much more restricted role than in the American system, today most CCs are more active political players than most courts in diffuse systems.

The rise of CCs began after World War II: CCs were important elements in the democratic reconstruction of Germany and Italy. Their model was influential in the third wave of democratization, beginning with Spain and Portugal, followed by the transformation in Eastern Europe in which CCs played an important role.

Constitutional reforms in other parts of the world similarly introduced CCs (South Korea, Thailand). South Africa gave its CC the unique function of certifying that the new constitution complied with the principles agreed to in the peaceful abolition of apartheid. In addition, other countries without a tradition of judicial review introduced CCs in constitutional reforms (*Cour de Arbitrage* in Belgium 1980 – since 2007: *cour constitutionnelle*). France created a restrictive form of preventive judicial control through a *Conseil Constitutionnel* in 1958. This model was adopted by some former French colonies in Africa (e.g., Burkina Faso, Ivory Coast, Senegal), but others have changed to a constitutional court system (e.g., Benin, Guinea, Mali, Niger). It is no coincidence that the development of CCs occurs during times of constitutional transition. In this circumstance, values and interests protected by old law come sharply into conflict with the legitimacy of the new constitutional order.

In this situation a CC has obvious advantages over the American model because the ordinary judiciary will often be compromised by its involvement in the old order. While it is extremely difficult to completely transform legal cultures just by introducing a new constitution, a specialized court composed specifically for this task, one that possesses an institutional commitment to the new constitution, can play a decisive role in bringing the inherited legal culture into conformity with the new model.

Mixed Systems

Some countries have left constitutional cases to the ordinary courts but have recognized the special nature of constitutional litigation by concentrating it in a supreme court or special chambers, by creating special constitutional remedies, and by decreeing the *erga omnes* effects of judicial review. Examples can be found especially in Latin America (because of the competing influences of civil law inheritance and North American constitutionalism). But in Latin America many countries have also established constitutional courts (Bolivia, Chile, Colombia, Guatemala.) There are also examples of mixed systems in Europe (Ireland, Estonia).

International Constitutional Jurisprudence

Since the end of World War II, judicial control of the political system has gained an international dimension. With a decisive innovative approach, the European Convention on Human Rights (ECHR) allowed individuals to sue their governments for violation of human rights in an international forum (a model later followed by the American and, recently, the African human rights systems). For countries without a bill of rights, the ECHR organs served as a kind of substitute CC (and the desire to create a local forum to precede international proceedings has influenced the move toward judicial review, e.g., in Great Britain). The European Convention is also important because in some countries without judicial control of constitutionality (e.g., The Netherlands, Switzerland) the conformity of laws with international human rights can be examined by the courts.

With the European community for the first time, an international organization created mandatory and binding court control over community organs and member states. The main function of the European Court of Justice is to safeguard the uniform application of European law, but it also can be considered a CC of the European Union.

Organization and Powers of Constitutional Courts: An Overview

CCs across the world employ a variety of mechanisms to ensure constitutional control.

Composition of the Judiciary

In general, the rules for the selection of judges tend to address their political function more openly than those of ordinary courts. They try to square the circle between providing for judicial independence and democratic legitimacy. A majoritarian solution (election by the majority of the day) creates the fewest theoretical problems for democratic legitimacy but endangers the control function. Attempts to reconcile the two aims include election with supermajorities (Germany) leading *de facto* to equal appointment rights between government and opposition, the distribution of nominating rights among different constitutional organs (Austria, France), and the inclusion of professional bodies in the nominating process (Italy). As the political element is stronger, appointment to CCs tends not to simply depend on advancement in the judicial hierarchy. Law professors and former politicians are especially highly represented. A relatively strong political influence on the appointment process is counterbalanced regularly by strong

guarantees for judicial independence. Appointment or election is either for life or long term.

Jurisdiction

In contrast to the American system in which the constitutional question can only be raised incidentally, during a case, in a CC system the constitution and constitutional court acts circumscribe the jurisdiction of the Court. Systems differ greatly in the extent to which they grant jurisdiction to CCs. As a means of objective control outside concrete cases, the power of abstract judicial review on application from privileged applicants (state organs, entities of federal systems, quorums of members of parliament) is perhaps the one most typical for CCs (and most distinctive compared to the diffuse system). In its most restrictive, purely preventive, form, the French *Conseil Constitutionnel* must be engaged in the short interval between parliamentary vote and promulgation, thus reconciling constitutional control with the maxim *la loi est la loi*. Much more common is repressive abstract review after the law has come into force.

Concrete judicial review allows other courts to refer constitutional questions raised in their proceedings to the CC, which holds a monopoly on finding laws unconstitutional. Access to the CC is most open where it can be mobilized by individuals (constitutional complaint, *amparo*) alleging violations of human rights (usually only after exhaustion of other remedies). In addition to handling human rights complaints, CCs adjudicate disputes between state organs or between the components of federal systems, have quasi penal functions (impeachment, prohibition of parties), and supervise the election process.

Constitutional Courts and Political System

In political theory the role of CCs remains controversial. Their power to overrule the decisions of elected representatives of the people is difficult to reconcile with the theoretical premises of majoritarian democracy. In countries with a CC the debate tends to be less heated than in systems where judicial review has been invented by judges, because the constitution makes express provision for jurisdiction. With the extension of judicial control in the last decades, rigorous opposition against it in the name of democracy has declined. Despite the theoretical merits of the majoritarian argument against judicial control over parliament, practical experience has shown it to be a workable reconciliation of majority rule with safeguards against the dangers of majority rule. The international human rights discourse has provided additional legitimacy to mechanisms that impose control over elected governments in the name of a higher legal order.

There remains, however, a permanent tension between the power of CCs to interfere in the political process, and the tenets of democracy. But the debate has shifted from the merits of judicial control in the abstract sense to its proper boundaries. In jurisprudence this debate centers on concepts like 'judicial restraint' and 'political questions', while social scientists study the political role of CCs and their impacts. In this debate, one of the oldest contributions, Justice Stone's footnote 4 in *Carolene Products* (304 US 144 [1938]), provides a valuable guide: the protection of the democratic process itself and that of 'insular' minorities not able to protect themselves in this process is easier to justify than the enforcement of competing value judgments.

The Sociology of Constitutional Courts

CCs (and constitutionally activist Supreme Courts) hold a special place in sociology of law. Their distinguishing feature is almost banal: they are very small. Different from other sociolegal court studies, with CCs one is not dealing with a social group but individually known actors. CCs apply a body of law that typically is more open textured and politically more controversial than normal law. These contrasts with other courts both facilitate and impede their sociological analysis. On the one hand, CCs offer case studies of the relationship between law and politics that make the relationship between judicial values and decisions impressively obvious. Therefore, sociology of law treatises often use constitutional jurisprudence to illustrate this relationship. On the other hand, they are so distinct from other courts that the application of sociolegal findings about judicial behavior to them is difficult, and generalizations from constitutional courts to other courts are dangerous.

Actors

Because of the political role of CCs, the politics of elections and appointments is an interesting subject of research that is in this case more directed to concrete selection processes than to generalized study of recruitment patterns. While there are correlations between known judicial background (including party politics, religion, status) and decision, there is no simple one-to-one relationship. The guaranteed independence (cf 3a *supra*) of CC judges has often resulted in disappointment on the part of selecting politicians. This can be explained by the self-image of judges and the institutional interests of CCs. For their impact in a political system, CCs rely heavily on their perception as neutral arbiters, and they would lose legitimacy if they appeared to be mere instruments of those who selected them. The high prestige CCs enjoy in many systems (especially compared with politicians) very much depends on this aura of independence.

Mobilization

One of the major differences between courts and elected policy makers is the fact that courts are reactive, not active, participants in the public arena. While this fact is often used as an argument by defenders of judicial review against majoritarians, it appears to be of limited significance as a barrier for judicial involvement in politics. In systems with very restricted access to the court (like in Kelsen's original model), it holds true. But access to most CCs tends to be much broader (3b supra) than the original model provided, and the different avenues combine to bring most politically controversial matters before, the courts. In light of the German Bundesverfassungsgericht accepting less than 2% of

the *c.* 6000 complaints received annually, one may assume that the court does not so much react but rather actively selects those issues it wants to address.

In addition, CCs are inventive in finding ways to involve themselves in politically sensitive subjects without regard to restrictions under procedural rules. In a related development, the more active courts do not restrict themselves to decide the case before the court but use *obiter dicta* to make long exhortations on future political action, thus departing from the judge's role of reactively deciding on the past and invoking instead the politician's role of actively influencing the future.

Decision Making

Similarly, the binding force of law cited (especially by lawyers) as a restriction of judicial power is less important for CCs than for other courts. Judicial decision making always has a political dimension, but it is subject to constraints intended to promote a specific juridical rationality and to hinder idiosyncratic decision making by judges. These constraints appear to be less effective with CCs. Their decisions are less strictly programed by law. While some constitutional law provisions (organizational norms, e.g., on elections) can be very specific, those texts that are relevant today in most constitutional proceedings – fundamental rights or constitutional principles – are very open textured (and traditionally in constitutional jurisprudence have been further expanded, rather than narrowed). CCs have to balance broadly worded rights against similarly broadly worded infringement clauses; they then must find the solution by balancing private rights against the public interest, using extremely open concepts (like the principle of proportionality), which give widest possible room for interpretation.

CCs have transcended the classical boundaries of judicial politics in another way. In principle, courts are different from political actors because they are restricted to yes/no decisions (lawful/unlawful) over past behavior, which is a powerful incentive for judicial restraint (e.g., if a CC had only the alternative of either, declaring a tax law void and risking public bankruptcy or accepting the law as constitutional). CCs have found ways around this problem. For example, they refrain from declaring laws void but order the lawmaker to change them in a given time. Or, they use the interpretation of laws or *obiter dicta* (i.e., nonbinding commentaries of a judge with only incidental bearing of a case) to enforce political compromises between proponents and opponents of a law under scrutiny. Such flexible forms of judicial decision free the court from the stark yes/no alternative and broaden its potential to actively involve itself in policy making by influencing future policies, building coalitions, and instigating political compromises. A promising strategy for measuring judicial activism appears, therefore, to be the study of the extent to which CCs have given up the ideal typical judicial style of decision making for a political one.

See also: Civil Liberties and Human Rights; Civil Rights; Constitutionalism, Comparative; Courts and Adjudication; Democratic Transitions; Supreme Courts.

Bibliography

Beatty, D. (Ed.), 1994. Human Rights and Judicial Review: A Comparative Prospective. Nijhoff, Dordrecht, NL.

Bryde, B.-O., 1999. Constitutional Courts in Constitutional Transition. Liber Amicorum van Houtte, Acco, Leuven, Belgium.

Cappeletti, M., 1971. Judicial Review in the Contemporary World. Bobbs-Merrill, Indianapolis, MI.

Diallo, I., 2004. A la recherché d'un model africain de justice constitutionnelle. Annuaire international de justice constitutionnelle 20, 93–120.

Favoreu, L., 1986. Les Cours Constitutionnelles. Presses Universitaires de France, Paris.

Grewe, C., Ruiz Fabri, H., 1995. Droits Constitutionnels Européens. Presses Universitaires, Paris.

Kelsen, H., 1928. La Garantie juridictionnelle de la Constitution. Revue du Droit Public et de la Science Politique 45, 197–257.

Helmke, G., Rios-Figueroa, J. (Eds.), 2011. Courts in Latin America. Cambridge University Press, Cambridge, UK.

Klug, H., 2000. Constituting Democracy. Law, Globalism and South Africa's Political Reconstruction. Cambridge University Press, Cambridge, UK.

Kommers, D.P., 1989. The Constitutional Jurisprudence of the Federal Republic of Germany. Duke University Press, Durham, NC.

Landfried, C. (Ed.), 1988. Constitutional Review of Legislation. Nomos, Baden-Baden, Germany.

McWhinney, E., 1986. Supreme Courts and Judicial Law-Making. Nijhoff, Dordrecht, the Netherlands.

Sajó, A., 1995. Reading the invisible constitution: judicial review in Hungary. Oxford Journal of Legal Studies 15, 253–267.

Sindjoun, L., 2009. Les grandes décisions de la justice constitutionnelle africaine: Droit constitutionnel jurisprudenciel et politiques constitutionnelles au prisme des systèmes politiques africains. Bruylant, Bruxelles, Belgium.

Smith, E. (Ed.), 1988. Constitutional Justice under Old Constitutions. Kluwer, The Hague.

Yang, K., 1993. Judicial review and social change in the Korean democratizing process. American Journal of Comparative Law 41, 1–8.

Constitutionalism, Comparative

Heinz Klug, University of Wisconsin Law School, Madison, WI, USA; and University of Witwatersrand, Johannesburg, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Distinguishing the idea and practice of comparative constitutionalism from comparative constitutional law, this entry focuses on the historical roots and reemergence of comparative constitutionalism, which focuses on how constitutions are made, adopted, and used to structure society. It also discusses their legal existence and application within the political and legal practices of different countries.

Introduction

The idea and practice of comparative constitutionalism is a new field propelled by the political and technological changes of the late twentieth century (Jackson and Tushnet, 1999), yet it is rooted in antiquity, in the methodology of Aristotle and his students at the Lyceum. While not a well-established area of academic study (ACLS, 1990), comparative constitutionalism exists as a normative and legal practice as well as a form of comparative political and legal analysis. In contrast to comparative constitutionalism, comparative constitutional law has become a well-established and thriving discipline over the last decade (see Rosenfeld and Sajo, 2012; Ginsburg and Dixon, 2011). Nevertheless, comparative constitutionalism, which embraces not only the practice of comparative constitutional law but also the conceptualization and analysis of systems of governance that are based on the systematic allocation and control of power, remains a growing field.

Comparative constitutionalism is now widely practiced by both constitution makers and constitutional interpreters. Comparative constitutionalism became prominent in the late twentieth century, expanding exponentially in practice with the wave of constitution making that accompanied the end of the cold war. Constitutional reconstruction and amendment has continued to be a marked feature of national and international politics and law into the twenty-first century. Originating in the process of constitution making, in which participants sought to apply models and experiences from other contexts, comparative constitutionalism is perpetuated in the debates over the meaning of each new constitutional provision from the moment of its adoption. This practice is institutionalized in the jurisprudence of many constitutional and supreme courts who through their decisions effectively engage in a transnational comparative constitutional discourse on constitutional meanings and functions.

Academically, a single definition of comparative constitutionalism is difficult to construct. In philosophy, political science, and law, comparative constitutionalism is considered both a method of normative evaluation and an empirical and social scientific task of describing, contrasting, and reflecting upon the consequences of the different rights and institutional provisions of constitutions around the globe. As a normative exercise, the survey and comparison of constitutions remains close to the original Aristotelian method of comparing different constitutional provisions and institutions in order to define a better form of government. This approach has been supplemented with historical accounts that explore the development of constitutionalism over time (Pennington, 1993) and across national borders (van Caenegem, 1995). A normative emphasis is also prevalent in a range of studies that focus on some of the foundational questions underlying the adoption and implementation of constitutions: the relationship between constitutionalism and democracy (Elster and Slagstad, 1988; Habermas, 1996); the origin of the notion of fundamental rights and the role of civil society (Mamdani, 1991); and even the idea of a diversity of forms of constitutionalism that might be achieved through demands for cultural recognition in a postcolonial age (Tully, 1995). While increasingly of an interdisciplinary nature and central to the idea of constitutionalism, these studies tend to either focus on the conceptual and theoretical issues or adopt a country-by-country approach rather than a comparative perspective (Gordon, 1999; Pennock and Chapman, 1979).

However, with the emergence of globalization as both an empirical phenomenon and an object of study, the debate on constitutionalism has taken a comparative turn (Singh, 1989). Comparative constitutionalism in this latter sense reflects (or incorporates) two distinct approaches. On the one hand, there is a focus on comparing the different constitutional institutions that are employed to secure democratic forms of government (Sartori, 1994) or their role in shaping particular social issues such as identity and diversity (Rosenfeld, 1994; Habermas, 1998). On the other hand, there is an emerging focus on constitutional borrowing, adaption, and hybridization (Badi, 2000; Klug, 2000). The publication of two major works rooted in a comparative approach to constitutionalism has helped to further define this fledgling field (Hirschl, 2010; Jacobsohn, 2010).

Ran Hirschl argues for the recognition of a new legal form – constitutional theocracy – that he argues helps us to understand the coexistence of constitutional systems of government and a rising tide of theocratic authority that has undermined traditional notions of the separation of church and state that lies at the heart of liberal constitutionalism. Despite his fascinating analysis about the role of constitutional law and courts in bolstering secular forms of rule and his argument that there is "an intriguing conceptual affinity between constitutional law and religio[us] law, each with its own constitutive texts, interpretative hierarchies, and righteous morality alongside earthly interest" (Hirschl, 2010: p. 249), the most intriguing question

he posits seeks the source of conflict and change "within and between the religious and constitutional domains" (Hirschl, 2010: p. 17). Challenging the tendency in academia and popular culture to resist a political understanding of either law or religion, Hirschl points to the power of interpretation in both religion and law as a source of change; he also makes the argument that understanding the 'striking similarities' between these different symbolic systems may help us to understand how they evolve or "mutate to address social needs, economic interests, or political aspirations" (Hirschl, 2010: p. 239). While individual and collective identities might seem central to such a symbolic view of law and religion, Hirschl argues against the assumption of "firm identities and fixed group affiliations" (Hirschl, 2010: p. 45); instead he uses the facts of multiplicity in individual identities (as well as intra-group or intrafaith divisions) to emphasize the importance of politics, arguing that these markers of identity "may be brought to the fore or relegated to lesser status as coalitions shift, elites transform, and interests change" (Hirschl, 2010: p. 45).

This decentering of identity is in marked contrast to earlier works that either assumed that constitutions and their interpretation play a key role in the creation and evolution of national identities (Karst, 1989) or challenged the notion of a single national or constitutional identity and instead called for the recognition of a multiplicity of constitutional identities within particular constitutional orders (Tully, 1995). James Tully's argument for the recognition of a multiplicity of identities within modern constitutionalism resonates with Gary Jacobsohn's conception of the 'disharmonic constitution,' in which there exist both 'conflicting, even radically inconsistent ideas' (Jacobsohn, 2010: p. 3) and a broader constitutional identity – consisting of "identifiable continuities of meaning within which dissonance and contradiction play out" (Jacobsohn, 2010: p. 4). For Jacobsohn, the identity of any constitutional order fundamentally depends on the existence of constitutional disharmony, which serves as the catalyst of change.

This new scholarship on constitutionalism is premised on the idea that the constitution is not a particular historic document or even a legal structure but rather a broader regime that is constituted by a combination of legal commitments, institutions, and continuing political contestation over ideas and resources. In its comparative method it offers extraordinary insights into the structure and functioning of these varied constitutional orders. It also struggles to elucidate the mechanisms and pathways that have led to either gradual or dramatic change in different constitutional regimes. Instead of locating constitutionalism in a purely philosophical or normative discussion of appropriate forms of governance, these studies recognize its historical and institutional dimensions as key aspects of our conceptualization and debates over the place and role of constitutions. Instead of focusing our attention solely within the four corners of each constitutional document and the particular history and culture of the legal system in which it functions, this scholarship locates the discussion within the broader conception of constitutional orders – "a reasonably stable set of institutions through which a nation's fundamental decisions are made over a sustained period, and the principles that guide those decisions" – (Tushnet, 2003: p. 1) and the political realms these orders constitute and reflect.

The dispersal of sovereignty that is inherent in the process of globalization provides further difficulties for the definition of constitutionalism, as may be witnessed in the debate over citizenship and democracy in the European Union, or in the ideas of shared sovereignty that have been introduced into the debate over the future of Jerusalem. The post–twentieth century emphasis on the negotiated settlement of disputes over governance and power, whether at an international or local level, provides fertile ground for debating constitutional alternatives. At the same time, Stephen Holmes reminds us that "[c]onstitutions emerge and survive because, with a little help from their judicial friends, they serve the perceived interests of the best organized and therefore most powerful social forces. When the powerful discover the advantages they can reap from making their own behavior predictable, they voluntarily submit to constitutional constraints" (Holmes, 2012). From this insight he concludes that "[i]f you wish a constitutional norm to govern the way politicians behave, you need to organize politically to give ruling groups an incentive to pay attention and accept restraints on their own discretion for their benefit and yours" (Holmes, 2012). In other words, comparative constitutionalism can become a pragmatic alternative to continuing conflict.

The Global Context

Alarm over the Americanization of the world became a central theme for many critics of globalism in the 1990s. At the same time, many lawyers participated in and celebrated the new emphasis on the 'rule of law,' presuming the global adoption of law and legal processes as a precondition for successful transitions from authoritarian regimes and command economies into a new world order. The adoption of bills of rights and the empowerment of courts, which characterized many of these political transitions, led many to believe these developments reflected the embrace of American values in the post–cold war era. Bruce Ackerman heralded the emergence of a 'world constitutionalism,' arguing that the "enlightenment faith in written constitutions is sweeping the world" (Ackerman, 1997). Another prominent United States constitutional scholar, Cass Sunstein, argued that the "right kind of constitution could play an important role in fueling economic development and democratic reform; indeed, under current conditions, it may be indispensable to them"(Sunstein, 1993).

With the widespread adoption of constitutional forms genetically tied to the innovations of the American Revolution, it is not surprising that both advocates and detractors of the American experience assume the United States to be the hegemonic model. However, American lawyers and scholars participating in what was, in the 1990s, a more than billion-dollar program to promote the rule of law around the world found themselves in competition with alternative constitutional models. Although the bulk of activity in this rule of law movement focused on judicial and legislative reform designed to promote market economies, it is important to recognize that a central product of this dual process of political transformation and rule of law advocacy was fundamental constitutional reform in half the world – approximately 104 of the 188 members of the United Nations. Of these, a staggering

78%, or arguably 81 countries, adopted completely new constitutions.

Given a common belief in the United States Constitution as a "source of inspiration and ideas … studied and appraised by scholars, invoked by publicists, adopted or adapted by revolutionaries, nation builders and statesmen" (Henkin, 1989), many of those who joined the multitude of legal expeditions to the coal face of constitution making in Eastern Europe, Africa, and elsewhere were surprised to find themselves confronted with alternative constitutional models. Advocates of the American model found that instead of enjoying an unassailable status, they faced open competition from advocates of the German, Canadian, Indian, or other constitutional experiences. Given that these alternative constitutional experiences are to a large extent themselves based on the American model of constitutionalism, it is ironic that instead of serving as the dominant model, the American experience has become simply one of a number of alternative models. Yet this form of pluralism may be explained by the emergence of a practice of comparative constitutionalism. Constitution makers, interpreters, and litigants actively compared their own conditions and experiences with the current and historical experiences of other countries in order to advocate for one or another constitutional solution: whether a particular constitutional institution, the specific form of a constitutional clause, or the meaning of particular constitutional language. These practices became increasingly common through the 1990s as states around the globe engaged in one of the most profound processes of constitutional and political reconstruction ever. It is this context that has framed the new intellectual interest in comparative constitutionalism.

Constitutions, Constitutionalism, and the Politics of Comparison

The idea of comparative constitutionalism has shifted emphasis from a philosophical debate over the nature and form of good government to a more legal focus on the application and meaning of particular constitutional structures and rights, as well as the role of the judiciary in the project of limited government (Sajo, 1999). While the question of judicial review, and particularly the debate over the relationship between constitutionalism and democracy, dominated much of the academic discourse in the twentieth century, the changing social and political context has brought greater attention to the interaction of constitutional ideas and institutions across different legal systems and national boundaries (Schedler et al., 1999; Hanafin and Williams 1999; Teubner, 1997). This new trend is beginning to define a new field of comparative constitutionalism in theory and practice.

Comparative constitutionalism originated in the west with Aristotle, who in the *Nicomachean Ethics* made a case for the comparative study of constitutions, arguing that by studying the "constitutions we have collected … we shall perhaps be more likely to see which constitution is best, and how each must be ordered, and what laws and customs it must use" (Aristotle, 1996: pp. 1181b13-22). Although only the *Constitution of Athens* remains from this early venture, the product of this analytical exercise is presented by Aristotle in the *Politics* and remains a principal source for the study of comparative government, which still largely aims at the construction of a normative definition of institutional and political practices that are necessary in order to secure the good life.

The Aristotelian notion of constitutionalism (combining both a distinctive political order and the idea of a political regime in general) is, according to Carl Friedrich, the source of the "modern idea of a constitution as the organization of a government, its offices, and the relation of the offices" (Friedrich, 1968). For Friedrich, however, this understanding – which recognizes all forms of government as exhibiting one or other form of constitutionalism – was inadequate, for it incorporated dictatorial forms of governance within the idea of constitutionalism. Instead, Friedrich (1937), along with Charles McIlwain (1947) and Francis Wormuth (1949), emphasized the relationship between constitutionalism and individual liberty. For them, the essence of constitutionalism is not the organization of power but rather the idea of limited power. From the idea of restraints on power to the separation and division of power, the focus among Anglo-American theorists through the middle decades of the twentieth century was on the guarantee of individual liberty, which they took to be the product of various means of limiting state power (Pound, 1957).

Writing first in the shadow of Nazi Germany and then in the aftermath of World War II, and during the cold war, Friedrich shifted the focus away from the comparative perspective inherent in the Aristotelian approach by posing the question: What is the political function of the constitution? From this starting point, Friedrich identified the protection of the individual member of the political community as the primary object of constitutional government, thus redefining constitutionalism as a 'modern phenomena' whose central concern is the practice and theory of limited government (Friedrich, 1937, fourth ed., 1968). This focus empowered Western theorists to reject constitutional forms, particularly the 'symbolic' state socialist constitutions that failed to limit state power, as aberrations of the constitutionalist tradition. Thus constitutionalism was once again given a normative form but at the cost of limiting the concept to only those constitutional arrangements that successfully limited and controlled state power.

The failure of postcolonial constitutions in Africa (Nwabueze, 1973; Okoth-Ogendo, 1993) and other parts of the third world, despite the inclusion of clauses formally limiting, separating, and controlling state power, produced an alternative, sociolegal approach to constitutionalism. Legal academics and political scientists argued that it was important to understand the economic and social context that determined the effective constitution of power in these newly independent states (Ghai, 1991). Taken further, this analysis seeks to demonstrate how the underlying colonial and neocolonial conditions of social organization, particularly the dual legal system and lack of land tenure security, restrict the capacity of civil society in Africa (Mamdani, 1996; Shivji, 1998). Without changes in these conditions, it is argued, the notion of legal rights against the state and restraints on state power are ephemeral. Despite this disillusionment with formal constitutions and constitutionalism, struggles for democracy and interethnic strife in many developing countries has fueled

a continual search for understanding the place constitutionalism may play in the reconstruction of postcolonial states. Yash Ghai, for example, has sought constitutional solutions to problems of ethnic strife, both in his practice as a constitutional advisor and in his academic work (Ghai, 2000). He has also applied this sociolegal perspective to analyzing the role of constitutional formulations produced to define the relationship between the central government in China and a reintegrated Hong Kong (Ghai, 1997).

With the protection of the individual as its central goal and the restraint of political power as its essential feature, the study of constitutionalism lost steam in the Anglo-American world, splitting in two distinct directions, one legal and the other political. On the legal side, there developed an interest in judicial review as the central concern of legal constitutionalists. Here the idea of comparative constitutionalism became, and remains to some extent, equated with comparative constitutional law (Cappelletti and Cohen, 1979). In this context the comparative approach focuses either on the specific problem of judicial review, engaging in a comparison of the process of judicial review exercised by courts around the world (Cappelletti, 1989), or compares the case law of different jurisdictions in order to explore different means of constitutionally resolving particular issues from the protection of property rights (Allen, 2000) to abortion and hate speech (Jackson and Tushnet, 1999). On the political side, the fiscal crisis of the state and political conservatism in the late 1970s and 1980s led to a direct political attack on the role of the state in society. Despite challenges of cultural imperialism by those advocating an 'Asian-model' of autocratic paternalism or those who argued that "constitutionalism is largely a cultural phenomena and not simply the product of properly designed institutions and structures of government" (Franklin and Baun, 1995), the emphasis upon the need to restrict state power created an intellectual and political environment in which the constitutional protection of individual rights and the institution of judicial review became the dominant features of a post–cold war international political culture.

The emphasis on the legal form of constitutionalism, and judicial review in particular, shifted the focus onto the role of the judiciary, encouraging both a practical emphasis on judicial reform as well as an interest in the global expansion of judicial power (Tate and Vallinder, 1995). However, at the same time, the failure, and even collapse, of governments in Africa as well as the Asian economic crisis in the 1990s brought a renewed focus on the question of power and the need for effective governance. This focus raised the question of how the constitutional allocation of power should in fact empower the state to achieve the necessary capacity to secure a level of social coordination necessary to the achievement of the already acknowledged constitutionalist goals of individual security and freedom.

The Problem of Governance and the Role of Comparative Constitutionalism

If constitutionalism is historically the struggle to regulate political competition, then it is unsurprising to discover that different constitutional options in fact reflect continuing ideological and political alternatives. The politics of constitution making (Elster, 1995) is thus festooned with the claims of competing political groups who project their ideological perspectives and political goals into constitutionalist forms (Klug 2000). While the post–World War II response to the violation of human rights produced a movement that could insist on the recognition and promotion of human rights through both international and domestic fora, the reemergence of a nineteenth-century liberalism, in the guise of 1980s neoliberalism, had profound implications for the politics of constitution making in the post–cold war era. In this context Friedrich von Hayek's conception of politics as threatening a liberal order that is simultaneously complex and delicate (Hayek, 1960) provided a central argument for those who sought to shrink the state. Although democratic theorists and constitutional scholars had long highlighted the anti-democratic implications of constitutional review, the neoliberal project emphasized the rule of law – requiring an independent judiciary with powers of judicial review – and demanded a constitutionalization of many other institutions and areas of governance as a means to limit the destabilizing impact of politics.

Despite the gains made by neoliberalism at the end of the cold war, the politics of constitution making remained eclectic. Although the vast majority of states undergoing reconstruction through processes of constitutional change in this period seemed to accept the liberal paradigm of individual human rights and multiparty democracy, this did not preclude the simultaneous inclusion of a range of alternative constitutional elements, including socioeconomic and cultural rights, reflecting alternative traditions. Significantly, the dominant liberal democratic tradition itself contains different trends, sometimes complementary, sometimes contradictory, advanced by different social forces and reflecting a plurality of interests. Of these, the two trends with the most direct impact on processes of post–cold war state reconstruction were those emphasizing liberty and equality. While there is a significant overlap in liberalism's ideological commitment to liberty and equality, emphasis on one or the other provides a range of alternatives within the liberal democratic tradition. These extend from classic nineteenth-century liberalism with its emphasis on individual freedom and property rights to the claims of the democratic collectivity inherent in the social democratic liberalism of the post–World War II era. While both aspects of this tradition emphasize electoral democracy and the protection of individual rights, they also contain contradictions with important consequences for the shape and role of the state.

While the struggle for equality, whether aimed at economic or racial inequalities, elicited the power of the state to address entrenched private power, the struggle for individual freedom has eschewed the state, often characterizing governmental power as the very source of oppression. These different responses to the state have obvious consequences for the structure of government, the most dramatic being the emphasis on the need to downsize or limit the role and capacity of the state. Combined with a reaction against the commanding role held by the state in state socialist societies

and strengthened by the fiscal crisis of the state in western democracies, this tendency achieved a significant degree of influence in shaping the international political culture that framed the post–cold war process of state reconstruction. By contrast the classic elements of a social democratic state, including the state's social welfare orientation and regulatory role in the relationship between capital and labor, including the creation of corporatist institutions that gave organized labor an important voice in social organization, fell into political and ideological disrepute, and began to be dismantled.

It is in this context that comparative constitutionalism provides an opportunity to rethink many of the longstanding assumptions about traditional constitutional forms so as to understand their role in the construction and maintenance of different constitutional orders. There are three sets of values that frame traditional notions of constitutionalism: (1) federalism, or the spatial division of power; (2) the separation of powers between different branches of government; and (3) the notion of constitutional rights. Each of these addresses different aspects of the 'problem' of power – its allocation, application, and restriction – within the nation state. While these constitutional values, with their diverse historical origins, have become increasingly universalized, their application in increasingly varied historical, cultural, and political contexts has produced a diversity of approaches and understandings that provide a diverse global text for political actors, constitution makers, constitutional interpreters, and litigants.

The Practice and Theory of Comparative Constitutionalism

As a field of empirical academic research and legal practice, comparative constitutionalism suffers many of the methodological challenges that have encumbered the fields of comparative law and politics. Apart from the more obvious difficulties of language and the different legal traditions that plague the comparative law project, the sociological and economic contexts as well as each country's particular political history require attention when one is studying the origins and meanings of particular constitutional forms. More recently, the increasing comparative jurisprudence of constitutional and supreme courts in a number of countries has broadened the scope of comparative constitutional law and raised questions about the theory and practice of constitutionalism in the context of a globalizing world. On the one hand, there is a growing attempt to formulate constitutional theories to throw light on the "increasingly cosmopolitan character of constitutional interpretation" (Choudhry, 1999); on the other hand, there is a call for Anglo-American lawyers to recognize the need for a broader legal theory of the state or civil law type *Staatslehre* (Dyzenhaus, 1998) as an essential step toward the formulation of an adequate theory of the relationship between law and political power. It is this relationship that is at the heart of the idea of constitutionalism and the object of research and practice in the growing field of comparative constitutionalism.

See also: Courts and Adjudication; Human Rights and Social Work; Judicial Review; Law and Democracy; Law and Development; Postcolonial Law; Rights: Legal Aspects; Rule of Law (and Rechtsstaat).

Bibliography

Ackerman, Bruce, 1997. The rise of world constitutionalism. Virginia Law Review 83, 771.
Allen, Tom, 2000. The Right to Property in Commonwealth Constitutions. Cambridge University Press.
American Council of Learned Societies, 1990. The ACLS Comparative Constitutionalism Project: Final Report, ACLS Occasional Paper, No. 13.
Aristotle, 1996. In: Everson, Stephen (Ed.), The Politics and the Constitution of Athens. Cambridge University Press.
Badi, Bertrand, 2000. The Imported State: The Westernization of the Political Order. Stanford University Press.
van Caenegem, R.C., 1995. An Historical Introduction to Western Constitutional Law. Cambridge University Press, Cambridge, UK.
Cappelletti, Mauro, 1989. The Judicial Process in Comparative Perspective. Clarendon Press, Oxford.
Cappelletti, Mauro, Cohen, William, 1979. Comparative Constitutional Law. Virginia, Michie, Charlottesville.
Choudhry, Sujit, 1999. Globalization in search of justification: toward a theory of comparative constitutional interpretation. Indiana Law Journal 74, 819–892.
Dyzenhaus, David (Ed.), 1998. Law as Politics. Duke University Press, Durham and London.
Elster, Jon, 1995. Forces and mechanisms in the constitution-making process. Duke Law Journal 45, 364.
Elster, Jon, Slagstad, Rune (Eds.), 1988. Constitutionalism and Democracy. Cambridge University Press.
Franklin, Danial P., Baun, Michael J. (Eds.), 1995. Political Culture and Constitutionalism. M.E. Sharpe, Armonk, New York and London.
Friedrich, Carl J., 1937. Constitutional Government and Democracy. Blaisdell, Waltham, MA (fourth ed., 1968).
Friedrich, Carl J., 1968. In: Sills, David (Ed.), International Encyclopedia of the Social Sciences. MacMillan, New York.
Ghai, Yash, 1991. The theory of the state in the third world and the problem of constitutionalism. Connecticut Journal of International Law 6, 411.
Ghai, Yash, 1997. Hong Kong's New Constitutional Order. Hong Kong University Press.
Ghai, Yash (Ed.), 2000. Autonomy and Ethnicity. Cambridge University Press, Cambridge, MA.
Ginsburg, Tom, Dixon, Rosalind, 2011. Comparative Constitutional Law. Edward Elgar Publishing, Cheltenham, UK.
Gordon, Scott, 1999. Controlling the State. Harvard University Press, Cambridge, M.A.
Habermas, J., 1996. Between Facts and Norms (William Rehg, Trans.). MIT, Cambridge, MA.
Habermas, J., 1998. In: Cronin, C., de Greiff, P. (Eds.), The Inclusion of the Other. MIT, Cambridge, MA.
Hanafin, Patrick, Williams, Melissa, S. (Eds.), 1999. Identity, Rights and Constitutional Transformation. Ashgate.
Hayek, F. A von, 1960. The Constitution of Liberty. University of Chicago Press, Chicago, IL.
Henkin, Louis, 1989. In: Henkin, Louis, Rosenthal, Albert J. (Eds.), Introduction in Constitutionalism and Rights. Columbia University Press, New York.
Hirschl, Ran, 2010. Constitutional Theocracy. Harvard University Press, Cambridge, MA.
Holmes, Stephen, 2012. Constitutions and constitutionalism. In: The Oxford Handbook of Comparative Constitutional Law. Oxford University Press.
Jackson, Vicki C., Tushnet, Mark, 1999. Comparative Constitutional Law. Foundation Press, New York.
Jacobsohn, Gary J., 2010. Constitutional Identity. Harvard University Press, Cambridge, MA.
Karst, Kenneth L., 1989. Belonging to America: Equal Citizenship and the Constitution. Yale University Press.
Klug, Heinz, 2000. Constituting Democracy. Cambridge University Press, Cambridge, MA.
Mamdani, Mahmood, 1991. Social movements and constitutionalism: the African context. In: Shivji, Issa G. (Ed.), State and Constitutionalism: An African Debate on Democracy. Sapes Books, Harare, Zimbabwe.
Mamdani, Mahmood, 1996. Citizen and Subject. Princeton University Press.
McIlwain, Charles H., 1947. Constitutionalism: Ancient and Modern. Cornell University Press, Ithaca, NY.

Nwabueze, B.O., 1973. Constitutionalism in the Emergent States. C. Hurst, London.
Okoth-Ogendo, H.W.O., 1993. Constitutions without constitutionalism: reflections on an African Paradox. In: Greenberg, et al. (Eds.), Constitutionalism and Democracy. Oxford University Press, New York.
Pennington, Kenneth, 1993. The Prince and the Law 1200–1600. University of California Press, Berkeley, CA.
Pennock, Roland J., Chapman, John W., 1979. Nomos XX: Constitutionalism. New York University Press, New York.
Poggi, Gianfranco, 1978. The Development of the Modern State. Stanford University Press, Stanford, CA.
Pound, Roscoe, 1957. The Development of Constitutional Guarantees of Liberty. Yale University Press, New Haven, CT.
Rosenfeld, Michel, 1994. Constitutionalism, Identity, Difference, and Legitimacy. Duke University Press, Durham, NC.
Rosenfeld, Michel, Sajo, Andras (Eds.), 2012. The Oxford Handbook of Comparative Constitutional Law. Oxford University Press, Oxford, UK.
Sajo, Andras, 1999. Limited Government. Central European University Press, Budapest, NY.
Sartori, Giovanni, 1994. Comparative Constitutional Engineering. New York University Press, New York.
Schedler, Andreas, Diamond, Larry, J., Plattner, Marc, F., 1999. The Self Restraining State: Power and Accountability in New Democracies. Lynne Rienner Publishers.
Shivji, Issa G. (Ed.), 1991. State and Constitutionalism: An African Debate on Democracy. Sapes Books, Harare, Zimbabwe.
Shivji, Issa G., 1998. Not yet Democracy: Reforming Land Tenure in Tanzania. International Institute for Environment and Development and HAKIARDHI, University of Dar es Salaam, Tanzania.
Singh, Mahendra P., 1989. Comparative Constitutional Law. Eastern Book Company, Lucknow, India.
Sunstein, Cass R., 1993. On property and constitutionalism. Cardozo Law Review 14, 907–935.
Tate, C. Neal, Vallinder, Torbjorn (Eds.), 1995. The Global Expansion of Judicial Power. New York University Press, New York.
Teubner, Gunther (Ed.), 1997. Global Law Without a State. Ashgate.
Tully, James, 1995. Strange Multiplicity: Constitutionalism in an Age of Diversity. Cambridge University Press, Cambridge, UK.
Tushnet, Mark, 2003. The New Constitutional Order. Princeton University Press, Princeton, NJ.
Wormuth, Francis D., 1949. The Origins of Modern Constitutionalism. Harper, New York.

Constitutions

Kevin L Cope, Department of Political Science, University of Michigan, Ann Arbor, MI, USA
Mila Versteeg, School of Law, University of Virginia, Charlottesville, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A constitution is a set of fundamental principles, norms, and laws that govern a political or other organization. The term can refer either to a body of rules and practices that establish the elements of government and allocate government power, or to a foundational written document designed to do the same. Modern national written constitutions generally contain several key elements, including a bill of rights, a division of powers between government entities, and a claim to supremacy over other sources of internal law. The primary functions of constitutions may include intra-government coordination, rule pre-commitment, solidifying power, articulating national identity, and signaling to international audiences.

Introduction

A *constitution* is a set of fundamental principles, norms, and laws that govern a political or other organizational entity. The term 'constitution' can refer either to a body of rules and practices that constitute, allocate, and limit government power, on one hand, or to a written document designed to do the same, on the other. Modern national written constitutions generally contain several key elements, including a bill of rights, a division of powers between government entities, and a claim to supremacy over other sources of internal law. Commonly, either the constitution itself or historical practice mandates that one or more governmental bodies interpret and enforce the constitution's provisions. The primary functions of constitutions may include intra-government coordination, rule pre-commitment, solidifying elite power, articulating national identity, and signaling to international audiences. Though constitutions are most commonly associated with national governments, many sub-national entities and social organizations, such as country clubs and pirate ships (Leeson, 2007), are also governed by constitutions. This article focuses on national constitutions.

Definitions and Conceptual Approaches

Two Meanings

The meaning of the term *constitution* has evolved over its roughly two-thousand-year history. Aristotle (350 BC) termed a constitution 'the way of life' of a polity's citizens. Under modern parlance, the term commonly denotes one of two related concepts, to which scholars often apply the respective terms 'small-c' constitution and 'large-C' constitution.

A *large-C* constitution denotes one or more adopted written documents, often specifically designed, drafted, and ratified over a short period. The term is often associated with a *written constitutional text*. In some cases, the large-C constitution comprises multiple documents, such as with Sweden's and Israel's multi-part constitutions. The large-C constitution usually sets out to define the nature of the government, to identify branches and other actors, to allocate powers, and to limit government power generally. This last objective can be accomplished by stipulating policies, by enumerating rights guaranteed to its citizens or members (a list often referred to as a *bill of rights*), or both. Typically, large-C constitutions assert supremacy over other forms of domestic law. To operationalize this supremacy, written constitutions often, but not always, stipulate that amendments or replacement requires approval from supermajorities of the legislature or electorate, and/or a consensus of sub-national entities.

In contrast, the *small-c* constitution (so-named because it is normally used as a common noun, not as a document title) is generally broader and less formal. It denotes "all rules which directly or indirectly affect the distribution or the exercise of the sovereign power" of the state (Dicey, 1915: 22). Put another way, it includes the "body of rules, practices, and understandings" within an organization that "determines who holds what kind of power, under what conditions, and subject to what limits" (Law, 2010: 377). The small-c constitution therefore encompasses constitutional text, constitutional interpretations by judicial and other bodies, informal conventions, and other precedents. The term is sometimes used synonymously with the term *constitutional law*, as both denote constitutional norms distinct from a strict reading of a written text. Notably, constitutional law can change without formally amending the constitutional text. As a result, a small-c constitution often contains provisions not present in the large-C constitution, and sometimes, the converse is true (Schauer, 2010).

After the American and French revolutions of the late eighteenth century, a global trend began in which states increasingly codified their bodies of constitutional norms with a single written document (Lane, 1996), thereby giving those states both small- and large-C constitutional forms. Despite this trend, notable exceptions exist to this day. For instance, Great Britain has never merged its core laws into a single writing. Instead, the British characterize a collection of unwritten conventions and fundamental documents – including, for example, the Magna Carta and the Human Rights Act 1998 – as their 'constitution.' Notwithstanding Great Britain and a few other countries, most countries today have both forms of constitutions, as do many sub-national entities, such as states, provinces, and municipalities.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.93026-5

Conceptual Approaches to the Written Constitution

Beyond these broad definitions, scholars have struggled to reach consensus on the conceptual parameters of the constitutional text: that is, how to determine which of an entity's written legal provisions comprise its 'constitution,' and which are mere ordinary statutes, aspirational declarations, or the manifestos of autocrats, for instance. Given the important symbolic and identity functions of constitutions, whether or not a set of laws is 'constitutional' is more than a semantic question; among other consequences, it can have meaningful policy implications for a polity's citizens. Especially where corpora have non-constitutional labels, where they lack key common elements, or where they can or are frequently amended or replaced, their statuses as constitutions may be doubtable, and may turn on the conceptual approach applied.

There exist at least two such approaches to defining the parameters of a constitutional corpus: a *nominal* approach and a *functionalist/essentialist* approach. Under a nominal approach, a constitution consists of whatever elements a polity's citizens and policy-makers understand to be its 'constitution' (or the closest native-language translation of the word) (Versteeg and Zackin, 2014). Relative to other approaches, this meaning is both theoretically and practically inclusive; nearly every nation-state now self-identifies some set of its laws as 'constitutional.' To comparative scholars, a nominal approach is attractive because it defers to those arguably best positioned to assess the contours of a constitution: its own creators and subjects. On the other hand, a nominal approach faces practical challenges. Because it places semantics over form, applying a nominal approach can award similar legal corpora constitutional status in one country but not another, possibly frustrating meaningful attempts at comparison. For example, an over fifty-year-old American statute, the Civil Rights Act of 1964, protects fundamental minority rights from both public and private actors. It is considered a deeply entrenched element of American human rights law, making it effectively immune to all-out repeal (Eskridge and Ferejohn, 2001). In those respects, the British Human Rights Act 1998 is similar. But because Britons consider their human rights statute to be constitutional but Americans do not consider theirs as such, a nominal approach would give the two substantively similar statutes different constitutional statuses. This approach could lead the unwary to the false conclusion that only Britons enjoy entrenched legal protection from certain common forms of discrimination.

An alternative, a *functionalist/essentialist* approach, seeks objective, universal standards by which to identify a constitutional corpus. Under this approach, only bodies of law that accomplish certain traditional constitutional functions, or that include certain higher values, are constitutional.

This functionalist/essentialist approach can be divided into functional and normative variants. The functional variant requires a constitution to contain certain traditional functions to be recognized as constitutional. While not all agree on what those traditional functions are, they normally include at least two: empowering principle organs of government, and limiting government power vis-à-vis the citizenry. Other traditional functions include constitutional supremacy and constitutional entrenchment (that is, elevated procedural hurdles to amendment or replacement). Because this approach foists a set of global values on diverse legal traditions, it is vulnerable to the critique that it constitutes a kind of ideational constitutional imposition. The fact that one notion of constitutionalism has attained greater acceptance, critics may say, cannot justify imposing it on societies who prefer alternative versions. In response, defenders might insist that without a universal working definition, it is impossible to make useful comparisons between governmental systems.

The normative variant of the functionalist/essentialist approach applies a form of natural law or some other higher normative criteria to define constitutionality. Under this perspective, there exist certain liberal values, such as equality, personal dignity, and the protection of political minorities, which are fundamental to any constitutional order. This approach takes the view that only constitutions that embody those values are valid constitutions (Dworkin, 1997). While this approach is championed by some noted political philosophers, it has not gained wide acceptance outside of that circle; most comparativists consider even constitutions that oppress minority rights and that recognize few personal liberties (e.g., South Africa's pre-1994 constitution) to be valid constitutions, despite their lacking liberal values (Hardin, 2013).

Common Components

Even under a nominal definition, modern national constitutions display considerable similarity in both their substance and their structural characteristics. First, constitutions tend to share several key substantive similarities. At an abstract level, almost every national constitution contains a bill of rights that recognizes a set of protected higher values. These provisions constrain government by restricting the choices available to government actors. Even so-called positive rights limit the range of choices available to policy-makers by forcing government to undertake certain actions (Zackin, 2013). Nearly every constitution also organizes government functions, usually by establishing a judicial, legislative, and executive branch. In this sense, a constitution enables a government in conducting its political affairs (see Section Constitutional Functions below).

Similarities are also found at a more detailed level. For example, 83% of all countries today task a judicial branch with the power of judicial review (Ginsburg and Versteeg, 2015). Likewise, certain 'generic' constitutional rights, such as the freedom of religion or a right to private property, are adopted by at least 90% of all modern constitutions (Law and Versteeg, 2011: 1200). In some cases, moreover, entire bills of rights are copied from a common source. Most notably, all of the former British colonies in Africa and the Caribbean adopted bills of rights that were largely verbatim copies of the European Convention on Human Rights (Parkinson, 2007). Indeed, empirical research has shown that bills of rights are typically standardized documents (Law and Versteeg, 2011).

Many modern constitutions also share important organizational attributes. First, national constitutions tend to exhibit similar levels of stability. Globally, national constitutions

adopted since 1781 have, on average, been replaced altogether every 19 years (Elkins et al., 2009). Coincidentally, this period is precisely that which Thomas Jefferson believed that constitutions should last. Jefferson argued that "the earth belongs in usufruct to the living," and therefore, that each generation (which he considered to last 19 years) should create its own constitution (Jefferson, 1958: 392). In addition to being overhauled roughly every generation, constitutions are even more frequently amended: every 5 years on average. Another common organizational feature of modern constitutions is that they typically contain a large number of detailed provisions, and they have been growing longer over time. National constitutions comprise 21 960 words on average; the longest national constitution is that of India, with over 170 000 words (Versteeg & Zackin 2014, 113). Many sub-national constitutions also contain numerous detailed provisions. The U.S. state of Alabama, for example, is over 340 000 words long. Yet not all constitutions fit this model. One of the shortest national constitutions is that of the United States, which contains just 7762 words. Notably a constitution's stability and length are related: one study has found that among national constitutions, longer constitutions in OECD countries are amended more frequently than shorter ones (Nardi and Tsebelis, forthcoming 2015).

Constitutional theorists have commonly treated the U.S. Constitution of 1789 as the prevailing constitutional model. Often, they have effectively fused the constitutional tradition of the United States with notions of 'constitutionalism' generally, meaning that constitutional theory has centered largely on American constitutional law. But as the figures above reveal, the U.S. Constitution is in many ways an outlier among national constitutions. With its mere 27 amendments, its over two-hundred-year lifespan, its brevity, and its general omission of positive rights, the U.S. Constitution differs markedly from those of most other countries in several respects. This disparity suggests that constructing constitutional theory chiefly around the U.S. Constitution is an overly parochial approach, which may lead to a distorted view of the global constitutional picture (Versteeg & Zackin 2014).

Constitutional Functions

In light of the now-widespread popularity of written national constitutions, scholars and policy-makers have grappled with the related questions of what functions constitutions can serve and why government actors are motivated to adopt them. Comparative legal scholars and social scientists generally accept that most modern constitutions serve several primary functions. The most common functions likely include intra-government coordination, rule pre-commitment, solidifying elite power, articulating national identity, and signaling to international audiences. Each of these functions may serve the interests of one or more groups, including the governed citizenry, domestic government officials, domestic interest groups, foreign nations, international organizations, and the drafters themselves. Most constitutions are a product of multiple functions, and these functions supplement each other. Together, these functions partially explain the modern universalism of written national constitutions.

Pre-Commitment

One function of constitutions is as a *pre-commitment* device. That is, by promoting legal rules to constitutional status, written constitutions encourage a state to uphold those rules, even as the cost of doing so increases (Elster, 1984). The central premise of pre-commitment theory is that, although governments later may recognize the theoretical benefits of protecting civil rights and other higher values, many governments succumb to opportunism, abandoning rules that become inconvenient. According to pre-commitment theory, constitutions can address this time-inconsistency problem. They do so by creating procedural barriers that raise the cost of changing the rules. Unlike ordinary legislation, changing constitutional rules in most systems requires supermajorities from multiple political entities. Moreover, in systems with judicial review and an independent judiciary, courts nullify new non-constitutional rules that conflict with constitutional ones. Where such mechanisms succeed, constitutional commitments become more credible, thereby preventing a government from reneging on its constitutional promises. This can yield various economic and political benefits for the country (North and Weingast, 1989).

In this sense, constitutions can be likened to the metaphorical ropes with which Homer's Ulysses bound himself to the ship's mast, thereby preventing him from wreaking havoc when the sirens' song drove him insane (Elster, 1984; Homer, 1967). Likewise, a constitution enshrines counter-majoritarian values and creates mechanisms to uphold them, binding society to those values even when popular passions tempt leaders to abrogate them (Elster, 1984). As Stephen Holmes has famously put it, the constitution is "Peter sober," while ordinary legislation is "Peter drunk" (Holmes, 1995: 135).

A notable weakness of constitutional pre-commitment theory is its difficulty explaining why constitutionalizing legal rules have a stabilizing effect, in other words, to explain the nature of the mechanism that renders constitutions self-enforcing (Levinson, 2011). After all, procedural constraints on constitutional change are fallible; supermajorities can amend constitutions, and officials looking to undermine constitutional rules can appoint judges sympathetic to that cause. Indeed, it was exactly this concern that was on James Madison's mind when he noted that the bill of rights provisions in the U.S. Constitution would amount to mere 'parchment barriers' (Federalist No. 48, 1787).

Coordination

The *coordination function* of constitutions helps government actors to synchronize on a single fixed set of rules in order to facilitate more effective government (Hardin, 2013). This function is sometimes analogized to competitive rowers, who must row in-sync to move the boat efficiently, or to automobile drivers, who must all drive on the right (or left) to avoid logjam or collision. To operate effectively, government organizers must likewise agree on how many branches the government will have, how many legislative houses to establish, and how the polity will choose its representatives. Even autocrats and officials with little respect for constitutional rules benefit from the efficiency gains produced by a single set of rules. As Montesquieu

observed, if monarchies were to govern without rules of succession, "the vacancy of the throne [would] always [be] attended with a horrid civil war" (Montesquieu, 1748: 146).

According to some constitutional theorists, the coordination perspective explains leaders' inclination not only to adopt constitutions, but also to comply with them. In this sense, it addresses the 'why constitutions restrain' question with which pre-commitment theory struggles. Coordination theory holds that the constitutional constraints endure in part because they establish conventions that, once created, are costly to change. After a nation has successfully coordinated upon a set of conventions to conduct its political affairs, the cost of deviating from those conventions rises. As Russell Hardin notes, constitutional conventions "raise the costs of doing things some other way" and form "an obstacle to re-coordination" (Hardin, 2013: 60). Though political actors may dislike a given constitutional rule, they nonetheless recognize that having the rules facilitates better government coordination and allows them to exercise power more effectively. The existence of these conventions thus benefits power-holders, rendering the constitution partially self-enforcing.

Citizens are likewise more likely to acquiesce in an imperfect existing set of rules, in part because retaining the status quo is less costly than attempting to re-coordinate on a different set. Citizens recognize that internal conflict, sectarian violence, disruption of government services, and other insecurity can be the price of abandoning an existing legal order. Indeed, according to Hardin, this logic explains why even after the U.S. presidential election of 2000 (in which George W. Bush was elected with a minority of the popular vote), the country never abandoned the unpopular Electoral College method for selecting the president (Hardin, 2013: 60). Thus, unless the costs of maintaining the current order are higher than transitioning to a new one, citizens will acquiesce in existing constitutional arrangements, despite their flaws.

Power-Solidifying

Another potential function of constitutions is to serve constitution-making elites and dominant coalitions by helping them to solidify and consolidate power. Comparative theorists have increasingly applied a strategic-realist approach to explain constitution-making. Under this view, constitution-making is "politics by other means" (Hirschl, 2013: 163–170), in that it reflects the bargains, interests, and self-dealing of domestic elites and interest groups. As Ran Hirschl observes, "strategic behavior by politicians, elites and courts plays a key role in explaining the tremendous variance in the scope, nature and timing of constitutional reform" (Hirschl, 2013: 168). Given that one of the primary goals of a constitution is to create and channel power, it is unsurprising that political actors often take a sharp interest in constitution-making (Cope, 2013a).

In particular, the strategic-realist constitution-making approach posits that constitution-makers push for constitutional arrangements that maximize their personal interests or their institutions' post-ratification utility (Ginsburg et al., 2008: 211). Ran Hirschl has described constitutions as a product of "hegemonic self-preservation." (Hirschl, 2013: 11–12). According to Hirschl's theory, constitution-makers who anticipate losing power may attempt to lock-in values that favor their interests in the constitution and protect them with judicial review, fearing their ouster from office will prevent them from obtaining those benefits through ordinary politics. In a related theory, Ginsburg has posited that constitutional reform is like purchasing "political insurance" that protects ruling coalitions from losing out in the political process after the constitution has been adopted (Ginsburg, 2003, 55–57). This phenomenon may have materialized in South Africa in the mid-1990s, where white elites consented to relinquishing some of their power, but insisted on constitutional protection of property rights from public taking by an African National Congress–controlled majority (Ginsburg, 2003: 55–57).

Symbolism

In most cases, constitutions not only reflect the values and interests of elites, but also proclaim the highest values of the nation generally. In that sense, constitutions can serve an important symbolic function and articulate the nation's identity (Jacobsohn, 2010). By setting out a nation's "core, constitutive political commitments" and with aspirational and historical statements (King, 2013: 81), written constitutions can signal a national identity to audiences in and outside the country.

Both preambles and substantive provisions can accomplish these objectives. Preambles, in particular, present an opportunity to express the distinctive aspects of a country and its people (Cope, 2013a; Galligan and Versteeg, 2013; Tushnet, 2006). According to one commentator, preambles are akin to "mission statements"; they set out the nation's future goals in light of its past experiences (King, 2013). National history can also inspire constitutions' more substantive provisions. For example, as Galligan and Versteeg (2013) note the 1996 South African constitution aims to overcome its apartheid past by emphasizing international rights and authorizing its constitutional court to take account of foreign and international law. Likewise, the 1949 German constitution, enacted in the wake of the Third Reich and the Holocaust, proclaims "human dignity" as one of its basic principles (1949 Basic Law of Germany art. 1). In this sense, constitutions often tell "the stories of the nation's development through many centuries" (Holmes, 1881: 1). (Versteeg, 2014).

International-Signaling

Written constitutions can also serve to signal a willingness to engage with foreign states and the international community. For new countries, for instance, the very act of writing a constitution signals the country's independence. New countries promulgate constitutions for some of the same reasons that they "design[] a new flag, national anthem and postage stamps," that is, to take their distinct place in the international community (Brown, 2002: 10; Duchacek, 1968).

Written constitutions can also serve more strategic international goals. Many constitutions contain provisions written strategically to attract international benefits, such as foreign direct investment, foreign aid, or reputational benefits. For example, some drafters attempt to attract foreign investment by including provisions that protect rights and respect the rule-of-law, thereby using their constitution as a means to

signal good faith intentions (Farber, 2002; Law, 2008). Other constitutional provisions are the product of coercion by foreign donors, who condition aid on the provisions' inclusion (Goderis and Versteeg, 2015). Indeed, both foreign aid donors and international financial institutions often make their assistance conditional upon respect for human rights. To appease donors, then, a country might enshrine such rights in their constitution. Aid donors have also offered technical assistance for rule-of-law reforms that affect the constitution (Hirschl, 2004: 47).

Finally, some studies suggest that written constitutions are used to bolster a country's international legitimacy. A body of sociological literature has demonstrated that the international community is increasingly characterized by standardized models of statehood, including standardized constitutional templates (Meyer et al., 1997; Law and Versteeg 2014, 104; Goodman and Jinks, 2013). A formal constitution that embodies certain values and principles has become a statement of conformity to the values of "world society" (Meyer et al., 1997: 159). To win the acceptance of the international community, states must replicate a standardized script in their national constitutions. Thus, it is possible that states strategically adopt international norms in part to reap reputational benefits.

As a result of such processes, constitutions have increasingly come to resemble one another. For example, there is evidence that for much of the nineteenth century and the early twentieth century, the U.S. Constitution of 1789 served as a model for national drafters, particularly in Latin America (Billias, 2010). Since World War II, however, national constitutions, particularly bills of rights, have increasingly moved toward alternative templates (Law and Versteeg, 2012). Specifically, some evidence suggests that by acting as paradigms of universally accepted rights, post–World War II international conventions were partly responsible for this trend (Elkins et al., 2013). Some authors have observed that the constitutions of Canada and South Africa may also have served as global examples.

Design Efficacy

Though an important constitutional function is binding governments to their promises, it is self-evident and empirically well-documented that constitutional rules are often opportunistically modified or interpreted, or simply not upheld. This concern traces back to at least James Madison, who, as noted above, postulated that constitutions would offer mere "parchment barriers." A constitution can be amended, courts can neglect to enforce constitutional provisions, and officials may ignore those rulings the courts do produce. Some countries simply possess what some comparativists have called 'sham' constitutions (Law and Versteeg, 2013), that is, constitutions whose rules are upheld so infrequently that the constitutions lose any pretext of legitimacy as binding law. Conversely, a few countries, such as Australia, provide relatively robust rights protections despite having no constitutional bill of rights. Such examples of wide disparities between constitutional law and reality raise the question of whether constitutions really 'matter,' that is, to what extent they meaningfully affect the behavior of nations and governments. To date, very few studies have addressed this question empirically (see, e.g., Chilton and Versteeg, work in progress), and those that have, struggle to distinguish correlation from causation. Thus, the question of to what extent constitutional design truly impacts government performance remains largely unanswered.

Despite the lack of evidence that constitutional design matters, there exists a long tradition of constitutional functionalism that involves a searching for the 'best' constitutional principles (Tushnet, 1999; Lutz, 2006). This tradition dates back to the enlightenment era, and its optimism about man's people's ability to use the constitutionalism to shape a better society. Thomas Paine's famous statement that "a constitution is a thing antecedent to government; and a government is only the creature of a constitution" characterizes the belief that constitutions can create government and a more desirable social order. (Paine, 1791). From a functionalist perspective, the constitution is like the nation's operating system for a nation: to perform optimally, a country needs the most up-to-date, technically superior operating system available (Kreimer, 1999). Under this perspective, constitution-makers are engineers with technical expertise to design optimal constitutional solutions. Through constitutional design, constitutions can promote desirable goals such as economic welfare, a lasting respect for rights and liberty, stability, democracy, and conflict mitigation in divided societies.

In recent decades, however, the optimism of the functionalist tradition has waned. Among those involved in constitution-making in different countries, there appears to be a growing consensus that a general constitutional 'recipe' does not exist, and that what works well for one country and time may not work well in others. Indeed, numerous constitutional designs with good track records have failed to constrain governments when implemented elsewhere. Many scholars now agree that "[m]odern constitutionalism cannot be reduced to any particular form or any specific identity or difference" and that "[c]ross-cultural transplantation of constitutional provisions is always dangerous, as an unreflected generalization from experience in a single culture is always likely to be wrong" (Rosenfeld, 1994: 35).

As a result, recent years have seen a growing constitutional focus away from substance and toward procedure. For lack of clear substantive recipes, popular participation has become the new mantra in constitutional design (Aucoin, 2010: xiii). Emboldened by the reputed success of the 1996 South African Constitution, which was drafted with broad public input, there is a growing consensus in the international rule-of-law community that constitutions must be written with widespread popular input, and the people who know what works best for a nation are its citizens. As U.S. Institute of Peace president Richard Solomon summarized the new prevailing wisdom, "there are no one-size-fits-all formulates or models," but "well-conducted processes can … contribute to building stable, peaceful states, whereas poorly conducted processes most certainly undercut such efforts" (Solomon, 2010: xi). Indeed, the last two decades have seen a surge in popular participation in constitution-making, starting with the South African Constitution in 1994, followed by the Ugandan in 1995, the Eritrean constitution in 1997, Thailand's constitution in 1997, the radical populist constitutions in Latin America, and the 2011 crowd-sourced constitution of Iceland.

Acknowledgement

We thank Tom Ginsburg, Dominic Nardi, and George Tsebelis for helpful comments and suggestions.

See also: Delegation of Power: Agency Theory; Democratic Theory; Efficacy, Political; Governments; Institutionalism; State Formation, Theory of.

Bibliography

Aristotle, 1952. In: Barker, E. (Ed. and Trans.), Politics, bk. IV, fourth ed. Oxford University Press, New York (c. 350 B.C.E.).
Aucoin, L., 2010. Introduction. In: Miller, L. (Ed.), Framing the State in Times of Transition: Case Studies in Constitution Making. USIP Press, Washington, DC, pp. xiii–xviii.
Billias, G.A. American constitutionalism heard around the world 1776–1989: a global perspective.
Brown, N., 2002. Constitutions in a Nonconstitutional World: Arab Basic Laws and the Prospect for Accountable Government. SUNY, Albany, New York.
The Basic Law of the Federal Government of Germany, 1949, art. 1.
Chilton, A., Versteeg, M., 2014. Do constitutional rights make a difference. Work in progress.
Cope, K.L., 2013a. The intermestic constitution: lessons from the world's newest nation. Virginia Journal of International Law 53, 667.
Cope, K.L., 2013b. South Sudan's dualistic constitution. In: Galligan, D.J., Versteeg, M. (Eds.), Social and Political Foundations of Constitutions. Cambridge University Press, New York, pp. 295–321.
Dicey, A.V., 1915. Introduction to the Study of the Law of the Constitution, eighth ed. Macmillian, London.
Duchacek, I.D., 1968. National constitutions: a functional approach. Comparative Politics 1, 91–102.
Dworkin, R., 1997. Freedom's Law: The Moral Reading of the American Constitution. Oxford University Press, New York.
Elkins, Z., Ginsburg, T., Melton, J., 2009. The Endurance of National Constitutions. Cambridge University Press, New York.
Elkins, Z., Ginsburg, T., Simmons, B., 2013. Getting to Rights: treaty ratification, constitutional convergence, and human rights practice. Harvard International Law Journal 54 (1), 61–96.
Elster, J., 1984. Ulysses and the Sirens: Studies in Rationality and Irrationality. Cambridge University Press, Cambridge, UK.
Eskridge, W.N., Ferejohn, J., 2001. Super-statutes. Duke Law Journal 50, 1215–1276.
Farber, D.A., 2002. Rights as signals. Journal of Legal Studies 31, 83–98.
Galligan, D.J., Versteeg, M., 2013. Theoretical perspectives on the social and political foundations of constitutions. In: Galligan, D.J., Versteeg, M. (Eds.), Social and Political Foundations of Constitutions. Cambridge University Press, New York, pp. 3–48.
Ginsburg, T., 2003. Judicial Review in New Democracies: Constitutional Courts in Asian Cases. Cambridge University Press, Cambridge, UK.
Ginsburg, T., Chernykh, S., Elkins, Z., 2008. Commitment and diffusion: how and why national constitutions incorporate international law. University of Illinois Law Review 2008, 201–237.
Ginsburg, T., Versteeg, M., 2015. Why do countries adopt constitutional review? Journal of Law, Economics, & Organization 30 (39), 1–9.
Goderis, B., Versteeg, M., 2015. The Diffusion of constitutional rights. International Review of Law and Economics.
Goodman, R., Jinks, D., 2013. Socializing States. Oxford University Press, New York.
Hamilton, A., et al., 1999. The Federalist Papers. Penguin (1787–88).
Hardin, R., 2013. Why a constitution? In: Galligan, D.J., Versteeg, M. (Eds.), Social and Political Foundations of Constitutions. Cambridge University Press, New York, pp. 51–72.
Hirschl, R., 2013. The strategic foundations of constitutions. In: Galligan, D.J., Versteeg, M. (Eds.), Social and Political Foundations of Constitutions. Cambridge University Press, New York, pp. 157–181.
Hirschl, R., 2004. Towards Juristocracy: The Origins and Consequences of the New Constitutionalism. Harvard University Press, Cambridge, MA.
Holmes, O.W., 1881. The Common Law. Little, Brown, Boston, MA.
Holmes, S., 1995. Passion and Constraint: On the Theory of Liberal Government. University of Chicago Press, Chicago, IL.
Homer, 1967. The Odyssey of Homer (R. Lattimore, Trans.). Harper & Row, New York.
Jacobsohn, G.J., 2010. Constitutional Identity. Harvard University Press, Cambridge, MA.
Jefferson, T., 1958. Letter from Thomas Jefferson to James Madison (6 September 1789). In: Boyd, Julian P. (Ed.), The Papers of Thomas Jefferson. Princeton University Press, Princeton, NJ, p. 392.
King, J., 2013. Constitutions as mission statements. In: Galligan, D.J., Versteeg, M. (Eds.), Social and Political Foundations of Constitutions. Cambridge University Press, New York, pp. 73–102.
Kreimer, S.F., 1999. Invidious comparisons: some cautionary remarks on the process of constitutional borrowing. University of Pennsylvania Journal of Constitutional Law 648–650.
Lane, J.E., 1996. Constitutions and Political Theory. Manchester University Press, New York.
Law, D.S., 2010. Constitutions. In: Cane, P., Kritzer, H. (Eds.), The Oxford Handbook of Empirical Legal Research. Oxford University Press, New York, pp. 376–398.
Law, D.S., 2008. Globalization and the future of constitutional rights. Northwestern University Law Review 102, 1207–1349.
Law, D.S., Versteeg, M., 2012. The Declining Influence of the United States Constitution. New York University Law Review 87, 762.
Law, D.S., Versteeg, M., 2014. Constitutional Variation Among Strains of Authoritarianism. In: Ginsburg, T., Simpser, A. (Eds.), The Constitutions of Authoritarian Regimes. Cambridge University Press, New York, pp. 165–217.
Law, D.S., Versteeg, M., 2011. The evolution and ideology of global constitutionalism. California Law Review 99, 1163–1257.
Levinson, D., 2011. Parchment and politics: the positive puzzle of constitutional commitment. Harvard Law Review 124, 657–746.
Lutz, D.S., 2006. Principles of Constitutional Design. Cambridge University Press, Cambridge, UK.
Meyer, J.W., et al., 1997. World society and the nation-state. American Journal of Sociology 103, 144–181.
Montesquieu, 1748. The Spirit of Laws (T. Nugent, Trans.). University of California Press, Berkeley, CA.
North, D.C., Weingast, B., 1989. Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century England. Journal of Economic History 49, 803.
Nardi, D.J., Tsebelis, G., 2015. A long constitution is a (positively) bad constitution: evidence from OECD countries. British Journal of Political Science, forthcoming.
Paine, Thomas, 1791. In: Philip, Mark (Ed.), Rights of Man, Common Sense and Other Political Writings, vol. 122.
Parkinson, C.O.H., 2007. Bills of Rights and Decolonization: The Emergence of Domestic Human Rights Instruments in Britain's Overseas Territories. Oxford University Press, New York.
Rosenfeld, Michel, 1994. Modern constitutionalism as interplay between identity and diversity. In: Rosenfeld, Michel (Ed.), Constitutionalism, Identity, Difference and Legitimacy: Theoretical Perspectives, vol. 1, p. 35.
Schauer, F., 2010. Amending the presuppositions of a Constitution. In: Levinson, S. (Ed.), Responding to Imperfection: The Theory and Practice of Constitutional Amendment. Princeton University Press, Princeton, NJ, pp. 145–161.
Solomon, R.H., 2010. Foreword. In: Miller, L. (Ed.), Framing the State in Times of Transition: Case Studies in Constitution Making. United States Institute of Peace Press, Washington, DC, pp. xi–xii.
Tushnet, M., 2006. Some reflections on method in comparative constitutional law. In: Choudhry, S. (Ed.), The Migration of Constitutional Ideas, pp. 67–83.
Tushnet, M., 1999. The possibilities of comparative constitutional law. Yale Law Journal 108, 1269–1274.
Versteeg, M., 2014. Unpopular constitutionalism. Indiana Law Journal 89, 1133.
Versteeg, M., Zackin, E., 2014. American constitutional exceptionalism revisited. University of Chicago Law Review 81, 101–165.
Zackin, E., 2013. Looking for rights in all the wrong places: why state constitutions contain America's positive rights. Princeton, New York.

Constraint-Satisfaction Models

László Egri and Thomas R Shultz, McGill University, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Constraint satisfaction is the art of finding values for variables in a way that satisfies various constraints. Many everyday problems can be usefully couched in these terms. This survey of constraint-satisfaction models covers both symbolic approaches from artificial intelligence (AI) and neural network approaches from psychology and cognitive science. Initial symbolic AI approaches attempted to solve constraint-satisfaction problems in a hard fashion – either the constraints were perfectly satisfied or there was no solution. Due to its limited modeling capacity, this approach was not particularly appealing to computational modelers in psychology who intuited that the problems people face often have no perfect solutions that satisfy all constraints. Hence, psychological approaches more often turned to the soft constraint satisfaction of artificial neural networks. Newer AI work on constraint satisfaction has made considerable progress in building systems that optimize solutions to hard problems even when it is impossible to satisfy all constraints. The most dominant trends in constraint satisfaction are reviewed and integrated, and forecasts are made for future research and modeling applications.

Introduction

In daily life, regulations, restrictions, and requirements are all constraints that humans must take into account in certain decision processes. For example, staff rosters in hospitals must satisfy restrictions on team composition, personnel regulations, and perhaps preferences. In schools, timetables must be generated in such a way that no teacher will teach two different classes at the same time in two different places. Loosely speaking, a constraint-satisfaction problem (CSP) is a problem in which a search space is given and the objective is to find a solution subject to certain constraints.

Constraint satisfaction has been addressed within two quite different traditions – the symbolic and the neural. The field of artificial intelligence (AI) has mainly adopted a symbolic approach, devising algorithms that manipulate symbols and analyzing the computational properties of these algorithms. The field of psychology has mainly adopted artificial neural network techniques to simulate a wide variety of psychological phenomena. We review both traditions here, identifying their similarities and differences and their respective strengths and weaknesses. We end by suggesting the advantages of greater integration of these two traditions, both for understanding the nature of constraint satisfaction and for implementing successful computer simulations of constraint satisfaction in humans.

Symbolic Algorithms for Constraint Satisfaction

CSPs have been intensively studied in computer science since the 1970s, especially in the subfield of AI. One of the most popular formal versions of the CSP, classical CSP, has the following formulation. A CSP instance is given as a set of variables, a finite domain for each variable, and a set of constraints on the assignment of values to variables. The domain of a variable contains all values that the variable can take. A constraint is composed of a constraint scope and a constraint relation. The constraint scope is a list of variables, for example x_2, x_5, x_7. The constraint relation specifies those value combinations that the variables in the scope can take on at the same time. For example, the relation $\{(a, a, b), (d, c, b)\}$ indicates that the variables x_2, x_5, and x_7 can be assigned either $x_2 = a$, $x_5 = a$, $x_7 = b$, or $x_2 = d$, $x_5 = c$, $x_7 = b$.

A Natural Way to Solve CSPs: Local Consistency Checking

Arguably, the most natural method to solve CSPs is to employ some version of local consistency checking (Lecoutre, 2009). We demonstrate this method in the context of solving a type of crossword puzzle (see Figure 1). Each cell of the puzzle must be assigned a letter from the English alphabet, and therefore we associate a variable with each cell as indicated in the figure.

The domain of each variable x_i is D_i, which in this case is the English alphabet. There are four constraints. The first one is $C_1 = [(x_1, x_2, x_3, x_4), \{(D, R, U, G), (D, O, V, E), (F, O, U, R)\}]$, where (x_1, x_2, x_3, x_4) is the constraint scope and $\{(D, R, U, G), (D, O, V, E), (F, O, U, R)\}$ is the constraint relation. The remaining constraints are

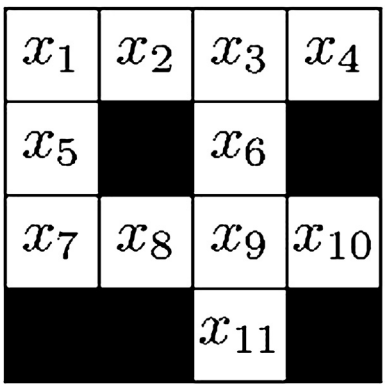

Figure 1 A crossword puzzle. Each column and row corresponds to a constraint scope.

$$C_2 = [(x_7, x_8, x_9, x_{10}), \{(G, O, A, T), (T, O, A, D)\}],$$
$$C_3 = [(x_1, x_5, x_7), \{(C, A, T), (D, O, G)\}],$$
$$C_4 = [(x_3, x_6, x_9, x_{11}), \{(V, E, A, L), (D, E, A, L)\}].$$

The goal is to find a solution, that is, to reduce the domain of each variable to a single letter in such a way that each constraint is satisfied. One strategy is to repeatedly detect and eliminate values in the variable domains that cannot participate in a solution because of some 'local inconsistency.' For example, initially all values of D_1 can be assigned to x_1. Looking at the constraint C_1, we see that if we set $x_1 = A$, then there are no possible values for x_2, x_3, and x_4 such that C_1 is satisfied. Therefore, we remove A from the domain of x_1. Repeating this procedure using the constraint C_1, we can eliminate all values from the domain of x_1 except for D and F. We next look at another constraint that involves x_1, constraint C_3. If we assign $x_1 = D$, we can find values for x_5 and x_7 such that C_3 is satisfied, so we cannot eliminate D. However, when we set $x_1 = F$, there are no values for x_5 and x_7 such that C_3 is satisfied, and therefore we can eliminate F. At this point we know that, in any solution of the puzzle, the only value x_1 could take on is D. Repeatedly reducing the possible values of the variables in this fashion results in the solution shown in Figure 2.

The algorithm we applied belongs to the class of local consistency-checking algorithms. There are many specific algorithms in this class but they all follow the same general principle: they forbid assigning certain values or combinations of values to a bounded number of variables because a certain subset of the constraints cannot be satisfied otherwise, and they repeat this until no more variable assignments can be discarded. A popular local consistency-checking algorithm, the arc-consistency algorithm, is discussed in the next section.

Arc Consistency

One of the most well-known consistency-checking algorithms is the arc-consistency algorithm developed in 1975. The origins of arc consistency go back to Waltz's filtering algorithm designed to automatically generate three-dimensional representations from line drawings of scenes. For simplicity, we restrict our attention to binary constraints, that is, we assume that the scope of all constraints consists of precisely two variables, and that every pair of variables is the scope of precisely one constraint. We say that a variable x is arc consistent with another variable y, if for every value a in the domain of x, there is a value b in the domain of y such that (a, b) is in the constraint relation of (x, y). The arc-consistency algorithm removes values from the domains of variables until all pairs of variables are consistent or the domain of a variable becomes empty.

In Figure 3, we demonstrate how this algorithm works. The CSP instance is shown to the left (stage 1). The variables are x, y, and z, each having domain $\{a, b\}$. The three constraints, one for each pair of variables, are indicated by the edges. For example, the constraint relation for (x, y) is $R_{xy} = \{(a, a), (b, a)\}$. We start by checking if variable y is consistent with x. It is not because if we set $y = b$, there is no value v for x such that (v, b) belongs to R_{xy}. Therefore, we remove b from the domain of y and observe that this is sufficient to make y consistent with x. At stage 2, we make z consistent with x by removing b from the domain of z, resulting in stage 3. At stage 3, we make z consistent with y by removing a from the domain of z. Because the domain of z is now empty at stage 4, we can conclude that there is no solution. This process of removing inconsistent values is often referred to as constraint propagation.

In general, arc consistency is established if the algorithm reaches a point where no domain reduction is possible. At that point, there are a few possibilities. If there is a variable for which the domain is empty, then there is no solution. If each domain has a single value, then there is a unique solution. Otherwise, there might or might not be a solution, and other techniques could be used to make further progress (e.g., search). Other well-known local consistency-checking algorithms are node consistency, path consistency, hyper-arc consistency, and so on.

Complexity and Applicability

CSPs are NP-complete in general. For example, the NP-complete *three-satisfiability problem* can be written as a CSP. In a nutshell, a problem is NP-complete if, given a solution, it can be verified efficiently (i.e., in polynomial time) whether that solution is correct, and the problem is among the 'computationally hardest' ones with this property. It is widely believed that NP-complete problems, and CSPs in particular, cannot be solved in polynomial time.

Local consistency-checking algorithms run in polynomial time, and therefore, as the previous paragraph suggests, they cannot solve every CSP. In fact, these algorithms suffer from a particular one-sided error. If a local consistency-checking algorithm discovers an inconsistency in a problem, then there is obviously no solution. The reverse is not true, however. Once local consistency is established, it is still possible that the given CSP has no solution. For example, consider the CSP in

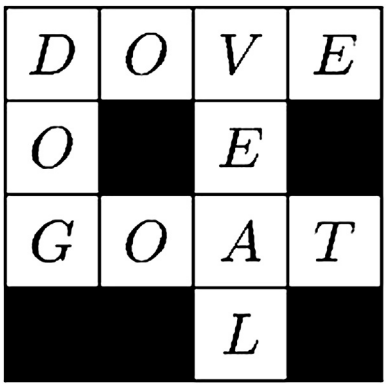

Figure 2 A crossword puzzle solution, that is, a variable assignment such that variables are assigned values that satisfy each constraint.

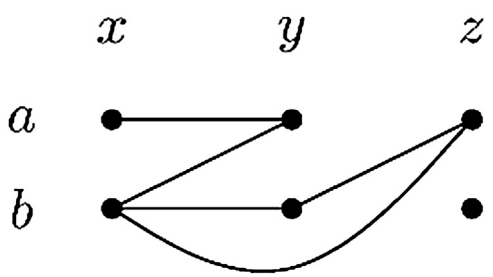

Figure 3 Constraint satisfaction by checking binary arcs.

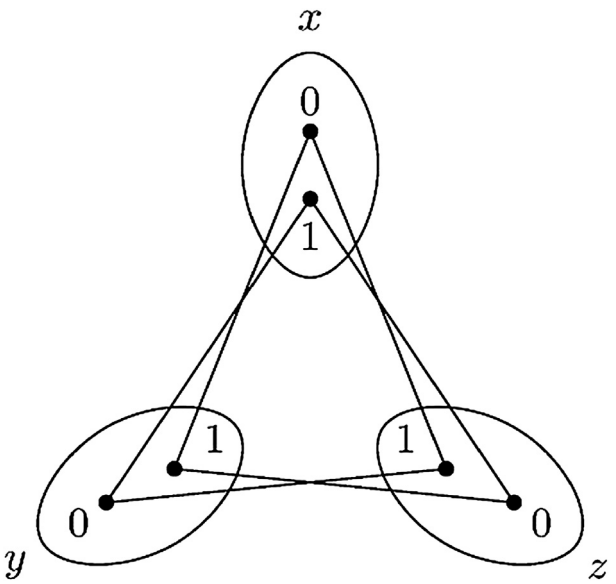

Figure 4 A CSP with local arc consistency, but no solution.

Figure 4, where the variables are x, y, and z, each with domain $\{0, 1\}$, and the constraints are indicated with the edges. This CSP is arc consistent (i.e., every variable is arc consistent with every other variable), and yet it has no solution. If we set $x = 0$, we must set $y = 1$, $z = 0$, and then we are forced to set $x = 1$, but we already set $x = 0$. If instead we set $x = 1$, we can obtain a contradiction in a similar way. It is not difficult to connect this example to the Möbius strip.

From a theoretical point of view, the power of local consistency-checking methods is relatively well understood. A popular variant of the CSP is when a list Γ of possible constraint relations is fixed in advance, and any given CSP instance must use relations only from Γ. Based on some technical properties of Γ, it is known whether CSP(Γ) can be solved using only local consistency-checking methods.

Backtracking Search

Backtracking search is an exhaustive search algorithm (depth-first search) that systematically assigns all possible combinations of values to the variables and checks if these assignments constitute a solution. Unlike the previously discussed algorithms, backtracking search is complete in the sense that it is guaranteed to find a solution if one exists. On the downside, the worst-case running time of backtracking search is exponential. From a practical point of view, there are numerous techniques for improving the efficiency of backtracking algorithms: nogood recording, backjumping, heuristics for variable and value ordering, constraint propagation (enforcing some sort of local consistency during exhaustive search to reduce the search space), and so on. The best combinations of these techniques often result in algorithms that can routinely solve large, difficult instances that are of practical importance in a reasonable amount of time. An excellent survey on backtracking algorithms is Chapter 4 of Rossi et al. (2006).

Local Search Methods for Classical CSPs

Local search is a metaheuristic method for solving hard optimization problems including CSPs. The intuition behind a local search algorithm is starting with an imperfect solution and then iteratively improving it until progress halts. For example, assume that we have to schedule exams subject to various space and time constraints. One strategy to obtain a valid schedule would be to start with a reasonable but conflicting schedule, and then to repeatedly change the time and location of those exams that cause the most conflicts.

More generally, a local search algorithm for CSPs initially assigns values to all variables, and then iteratively improves the solution by changing the value of a (few) variable(s) at a time. A natural heuristic to change the value of a variable in an iteration is the min-conflicts heuristic (Russell and Norvig, 2009). Two variables conflict if they violate a constraint (for simplicity, we are assuming that constraints are binary). The min-conflicts heuristic chooses a variable that conflicts with some other variable(s), and assigns it a value that minimizes the number of conflicts. We demonstrate the algorithm with the well-known n-queens problem, which can be easily stated as a CSP. The n-queens problem requires placement of n queens on an n by n chessboard such that no queens attack each other. In Figure 5, a two-step solution for an 8-queens problem using the min-conflicts heuristic is illustrated. For simplicity, we assume that each column contains a single queen and a queen can be moved only within its column. To get from state a to b, we observe that the chosen queen conflicts with the queen in the first and second rows, so it participates in two conflicts. The numbers in the squares indicate the number of conflicts that would result from moving the queen to the given square. There are three squares

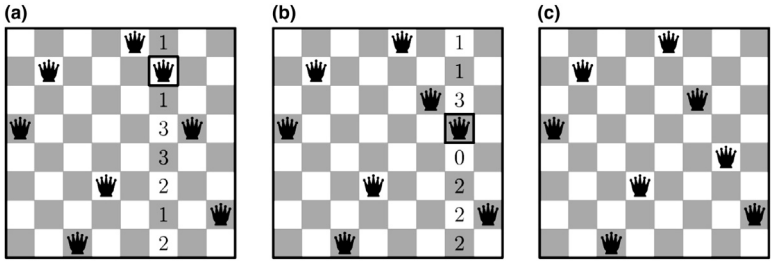

Figure 5 Solution of an 8-queens problem in two steps from left to right, using the min-conflicts heuristic. The queen selected for moving is outlined in black.

in which the queen would cause only one conflict. We move the queen to one of those three squares, arbitrarily the square just below its present location. In state c, we obtain a perfect solution.

Local search is very efficient for some classes of problems. For example, even if n is a million, this method solves the problem in about 50 steps. Nevertheless, just like local consistency methods, local search methods also exhibit a one-sided error: they may find a solution, but they may fail even if the problem is solvable due to getting stuck in a local minimum. To remedy this, local search algorithms often involve some sort of randomization to escape local minima. For example, if a state is reached where no local improvement is possible, the algorithm could be restarted with a fresh, randomly chosen initial state in the hope of obtaining a better solution. Simulated annealing is a more sophisticated way of locating a good approximation to the global optimum, briefly discussed in the Neural Constraint-Satisfaction section.

More Flexible CSP Variants

Classical CSPs treat constraints as hard, meaning that any constraint is either completely satisfied or completely violated, and in any solution, all constraints must be satisfied. In certain situations, however, it is desirable to have more flexibility. For example, given an unsolvable CSP instance, we could ask for a solution in which as many constraints are satisfied as possible. This variant of the CSP is called MaxCSP (Freuder and Wallace, 1992). Some other generalizations of the classical CSP that use soft constraints instead of hard ones are, for example, weighted CSP, probabilistic CSP (Fargier and Lang, 1993), valued CSP (Schiex et al., 1995), and semiring-based CSP (Bistarelli et al., 1997). The weighted CSP framework is useful when the satisfaction of some constraints is deemed to be more important than the satisfaction of some other constraints. Here, constraints are associated with weights, and the goal is to find a solution that maximizes the sum of the weights of the satisfied constraints. Note that the MaxCSP is a special case of the weighted CSP.

The probabilistic CSP framework is applicable when the presence of some constraints is uncertain. For example, suppose that a dinner party is to be organized, and some of the invited guests might be willing to eat red meat only with red wine, some of them might be vegetarian, and so on. In this way, the guests impose certain constraints on the dinner menu. However, some guests are unsure whether they will come, and therefore it is uncertain whether they should be taken into account in planning the menu. To model a situation in which the presence of some of the constraints is uncertain, we assign a probability to each constraint (these probabilities are assumed to be independent). The probability that a given variable assignment is a solution to the (unknown) real-life problem can then be computed by multiplying the probabilities that all the violated constraints are not present in the problem. The objective is to find a solution that maximizes this probability.

A powerful generalization of the classical CSP is the valued CSP (VCSP). In fact, all the above CSP variants can be thought of as special cases of the VCSP. For simplicity, we present a restricted version of the VCSP (see Chapter 9 of Rossi et al., 2006 for the general version). A restricted VCSP instance is similar to a CSP instance except that constraint relations are replaced by more general cost functions. A cost function f assigns a nonnegative numerical value or infinity to each tuple. The cost of a solution is

$$\sum_{[(v_1,\ldots,v_m),f]\in C} f(\phi(v_1),\ldots,\phi(v_m)) \qquad [1]$$

where $\phi(v)$ specifies the value of variable v, C is the set of constraints, and $[(v_1, \ldots, v_m), f]$ denotes the constraint with scope $(v_1, \ldots, v_m)$ and cost function f. A solution is an assignment with minimum cost. A model that is even more general than the VCSP is the semiring-based CSP (Rossi et al., 2006).

One approach to solve soft CSPs is to perform a type of exhaustive search referred to as depth-first branch-and-bound (DFBB) search. Other methods are based on inference (e.g., bucket elimination, soft local consistency, etc.) or on a combination of search and inference. An excellent survey on soft CSPs and how to solve them is Chapter 9 of Rossi et al. (2006).

Neural Constraint-Satisfaction

Each unit in a constraint-satisfaction neural network can represent a hypothesis that is interpreted as true when the unit is active and false when the unit is inactive. Connection weights between the units can represent implicational constraints between the hypotheses, whether positive or negative. These connections can be exhaustive, bidirectional, and symmetrical. There can also be external input to some units, reflecting external evidence, and a bias input, reflecting a priori tendencies. The relative importance of these various constraints is implemented by the sizes of weight, input, and bias values. The degree to which the network satisfies all the constraints can be computed as goodness.

In a generic constraint-satisfaction algorithm, the goodness contributed by a particular unit i is defined as

$$\text{goodness}_i = \sum_j w_{ji}a_ia_j + \text{bias}_ia_i + \text{input}_ia_i \qquad [2]$$

where a_i is the activation of receiving-unit i, a_j is the activation of sending-unit j, w_{ji} is the weight from unit j to unit i, bias_i is the bias of unit i, and input_i is the external input to unit i (Rumelhart et al., 1986). Goodness increases when units connected by positive weights achieve similar activations and units connected by negative weights achieve dissimilar activations, and as unit activations come to resemble their biases and their external inputs. Goodness over the entire network is defined as the sum of individual goodness values across all receiving-units i in the network:

$$\text{goodness} = \sum_{i=1}^{N}\left(\sum_{j=1}^{N} w_{ji}a_ia_j + \text{bias}_ia_i + \text{input}_ia_i\right) \qquad [3]$$

Such functions go by different names in different fields: Hopfield networks in cognitive neuroscience, Lyapunov functions in dynamic systems, Hamiltonian functions in statistical mechanics, cost functions in optimization theory, and fitness functions in evolutionary biology (Hertz et al., 1991).

Unit activations are updated so that overall goodness increases or eventually stays the same. The update rules employ net input to a unit i, defined as

$$\text{net}_i = t(\text{int}(\sum_j w_{ij}a_j + \text{bias}_i) + \text{ext}(\text{input}_i)) \quad [4]$$

where int and ext modulate the relative influence of internal and external factors, respectively, and net_i is a function of t. To ensure that net input stays within bounds $[-1, 1]$, t can be further defined as

$$t(x) = \begin{cases} x, & \text{if } -1 \le x \le 1 \\ -1, & \text{if } x < -1 \\ 1, & \text{if } x > 1 \end{cases} \quad [5]$$

If $0 \le \text{net}_i$, the activation of a unit i is computed as

$$a_i(t+1) = a_i(t) + \text{net}_i(\text{ceiling} - a_i(t)) \quad [6]$$

where $t+1$ indexes the updated activation, t indexes the previous activation, and *ceiling* is the highest activation that units can achieve. Equation [6] specifies that, before being added to the previous activation, net input is scaled by distance to the ceiling, thus preventing unit activations from rising above the ceiling. If $\text{net}_i < 0$, unit activation is updated by

$$a_i(t+1) = a_i(t) + \text{net}_i(a_i(t) - \text{floor}) \quad [7]$$

where floor is the lowest activation value. In eqn [7], activations are kept from falling through the floor by scaling net input by distance from the floor before adding to the previous activation. Overall goodness in the network continues to increase until units reach their extreme values or net inputs fall to 0. Keeping net_i within the bounds of $[-1, 1]$ was not in the original formulation of activation updates (Rumelhart et al., 1986), and its absence may not always pose a problem, but we add it here as a precaution. Note that this algorithm is a hill-climbing algorithm and is therefore a form of local search (see section Local Search Methods for Classical CSPs).

Algorithmic Variations

There are many possible variations on this basic algorithm. Activation values can be continuous or binary. Activation updates can be synchronous or asynchronous, deterministic or stochastic, and can minimize energy or maximize goodness. Energy is the negative of goodness.

The dynamics of network behavior can be conceptualized within a multidimensional space in which the activation of each unit is represented as a particular direction. Each point in this space, an activation vector, represents a possible state of the network. Over time, these networks tend to settle into a stable state, and such stable states are referred to as attractors. Attractors can be points, lines, rings, or planes, and their behavior can be stable, cyclic, or chaotic (Eliasmith, 2007). The hypothetical energy landscape of a hypothetical network with multiple fixed-point attractors is shown in Figure 6. Any input pattern in a particular basin of attraction will tend to make the network settle into the attractor state for that basin, similar to a marble rolling downhill. Such fixed-point attractors can simulate memory retrieval, generalization (or assimilation), pattern completion, categorization, and noise reduction.

Initial activations can start at 0 or be preset to other values. Unit biases, inputs, decay, self-connections, and internal and external modulators are optional. Weights can be symmetric or asymmetric. They can be determined by conventional learning algorithms, such as the Hebb rule or the delta rule (Hertz et al., 1991), or be set by a programmer, perhaps on the basis of psychological principles. Recurrent activation updating can also be incorporated into multilayered feed-forward networks that learn with back-propagation of error, in a technique called recurrent back-propagation. This allows for the inclusion of hidden units, enabling the learning of nonlinear functions. Boltzmann machines can also use hidden units, to the same advantage.

Constraint-satisfaction neural networks are not guaranteed to maximize goodness or, equivalently, minimize energy; they are only guaranteed to settle into a local optimum. In networks with stochastic binary units, search for a global goodness maximum can be facilitated by simulated annealing, a process of gradually lowering the temperature (or gain) of the activation update function in order to move networks out of local optima (Hinton, 2007). Stochastic binary units turn on with a given probability. Such networks are called Boltzmann machines because the probabilities of the states are characterized by the Boltzmann distribution in statistical mechanics. If the annealing schedule decreases the

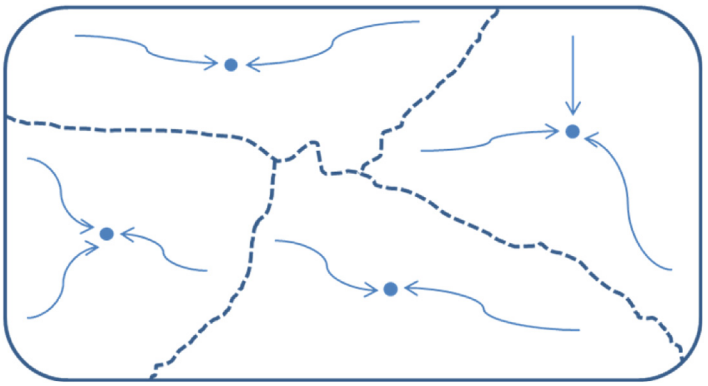

Figure 6 Hypothetical energy contour map with four attractors. An input anywhere in a basin moves toward that basin's attractor.

temperature in proportion to the inverse log of time, then simulated annealing is guaranteed to almost always converge to the globally optimal solution (Geman and Geman, 1984). Unfortunately, following this slow schedule usually exceeds the time required for a complete search of the solution space. In practice, faster cooling schedules may be employed with varying degree of success.

Modeling with Neural Constraint-Satisfaction

A variety of psychological phenomena have been simulated as fixed-point attractors with constraint-satisfaction neural networks. Content-addressable memories are stored and retrieved from cues containing part of the memory content (Hopfield, 2007). For example, provided with cues about a Japanese car company whose name begins with *M* and ends with *i*, many people can retrieve the Mitsubishi brand and recall details of some of its cars.

Values for particular slots in schemas are completed in a way that satisfies constraints imposed by external and internal information (Rumelhart et al., 1986). For example, provided with the cues of stove and refrigerator, other features of a generic kitchen can be accurately filled in.

Two important steps in analogical reasoning concern the retrieval and the mapping of an analogy. Analogies can be retrieved from memory by the parallel satisfaction of semantic, structural, and pragmatic constraints. For example, in trying to find a way to use an X-ray to destroy a tumor, an analogy about using a laser to fuse a filament in a light bulb is more likely to be retrieved than an analogy about using ultrasound to fuse a filament. This is because X-rays are more like lasers than like ultrasound devices. In addition, analogies are mapped onto new problems by satisfying constraints imposed by similarity, structure, and purpose (Holyoak and Thagard, 1989). For an example of purpose, if one is asked to compare Nicaragua with Cuba, any aspect of the two-memory representations may be relevant. However, if one is asked to assess likely political trends in Nicaragua in comparison to Cuba, then only political aspects, such as the characteristics of political leaders or the form of government, need to be mapped.

Favored explanations are those that maximize coherence among hypotheses and evidence. Thus, Darwin's theory of evolution of species is favored by scientists over a creationist account because of superior coherence. The two main evolutionary hypotheses (that species are subject to natural selection and thus evolve) explain many facts that creationism cannot, such as the geographical distribution of similar species and presence of vestigial organs. Moreover, the hypothesis of natural selection explains that of evolution. Further still, natural selection is explained by an exponential increase in population growth coupled with a slower increase in food supply, and there is a close analogy between natural and artificial selection, wherein farmers breed plants and animals to favor certain characteristics. A constraint-satisfaction network simulation confirms that this large difference in explanatory coherence strongly favors evolutionary theory over creationism.

Philosophical analyses can be treated in the same fashion. For example, belief in other minds is supported by providing the best explanation of behavior and by analogies between self and others.

Decisions and plans maximize coherence among goals, subgoals, and actions. In considering legal cases, for example, not only does evidence influence a verdict but a verdict also influences evaluation of the evidence. In both networks and people, this increases the coherence of understanding a case by resolving ambiguities and discrepancies.

Postdecisional cognitive dissonance is reduced as constraints imposed by attitudes, beliefs, and their implications are employed to increase consonance among these mental representations (Shultz and Lepper, 1996). For example, after making a difficult decision, the dissonance created by discovered shortcomings of the choice is reduced by increasing the valuation of the chosen item.

Forming impressions of people involves reconciling information about their behavior and characteristics and the stereotypes held by the perceiver. For example, in a story about a Mafia nun, there is pressure to reconcile the apparently conflicting information.

Attitudes change in a way that increases consistency across attitudes and beliefs. In addition, attitudes become more extreme merely by thinking about them, because neuronal activation change takes time. The more time spent updating activations, the more extreme they become. Moreover, stronger and more supported attitudes better resist persuasive attempts to change the attitude. Persuasion is an input constraint, but its power can be countered by other, interior constraints.

Objects are recognized by matching visual scene information with model features in a consistent way. An input string of letters is transformed within a recurrent feed-forward network into an attractor representing a pronunciation of the corresponding word. Damaging the units or connections of such recurrent feed-forward networks alters the basins of attraction, simulating several neuropsychological deficits such as dyslexia and prosopagnosia.

Oculomotor control has been modeled with line-attractor constraint-satisfaction networks. Ring attractors have been used to model cyclic control of head direction.

Constraint satisfaction also influences the understanding of sentences. Sentence parsing is faster when a sentence is more consistent with a verb's preferred subcategory structure. For example, the verb *claimed* can be used transitively as in sentence 1 or with a complement clause as in sentence 2. Because the latter use is more common than the former, sentence 2 is quicker to parse.

1. The boy claimed the prize.
2. The boy claimed the prize was big.

The constraint-satisfaction neural-network models of such phenomena typically involve a programmer-constructed network in which units can represent beliefs, evaluations, hypotheses, or evidential facts, and the weights between units represent constraints. Positive constraints, such as implication, are implemented by excitatory weights while negative constraints, such as contradiction, are implemented by inhibitory weights. After repeated updating of unit activations with equations such as 2–7, a network settles into a stable state, for which highly active units are interpreted as being believed or wanted, while those with low or negative activations are interpreted as disbelieved or unwanted.

For example, the belief that physical exercise combats obesity, coupled with a negative evaluation of obesity, could lead to a positive attitude toward exercise. Goodness increases with a positive triple product, in this case a minus obesity activation times a minus connection weight times a plus exercise activation. Despite the lack of guaranteed success, computational models using neural constraint-satisfaction routinely produce sensible and useful results and accurate simulations of human performance on simple CSPs.

In some cases, model predictions have subsequently been confirmed by further psychological research. For example, traditional dissonance theory predicted a spreading apart of the evaluation of chosen and rejected alternatives. Interestingly, the constraint-satisfaction model of dissonance reduction predicted a more refined version: a difficult choice between less desirable alternatives would produce a large increase in the evaluation of the chosen alternative, whereas a difficult choice between desirable alternatives would yield a large decrease in the evaluation of the rejected alternative. These more precise predictions were confirmed in a subsequent psychology experiment.

Future Trends in Constraint-Satisfaction Research

Current activities can be extrapolated to predict some near-future trends in research with constraint-satisfaction models. With a few early exceptions (notably, Waltz line labeling and crypt-arithmetic), computational models of constraint satisfaction have focused almost entirely on neural, rather than symbolic, algorithms. This may have been partly due to the inability of early (hard) symbolic algorithms to optimize constraint satisfaction when it was impossible to satisfy all the specified constraints. In contrast, the neural network approach optimizes naturally as an inherent part of processing. However, with the emergence of optimizing symbolic algorithms, perhaps it is time for cognitive modelers to revisit symbolic constraint-satisfaction algorithms. Some of these symbolic frameworks may have more expressive power than neural constraint-satisfaction techniques, and it could be worthwhile to explore their ability to simulate human problem solving.

As of this writing, research on finding efficient algorithms to solve classical CSPs is in a very active phase. In particular, tremendous progress has been made during the last 20 years on identifying all constraint relations Γ such that CSP(Γ) is solvable in polynomial time, and this quest is ongoing. (We note here that polynomial-time solvability of CSPs can also be guaranteed by imposing certain structural restrictions on the inputs and allowing arbitrary constraint relations.) For soft CSPs, current experimental studies indicate that certain algorithmic techniques did not yet reach their full potential, and furthermore, the study of polynomial-time-solvable soft CSPs is still in its infancy. On a different note, an important issue is how soft constraints are obtained, and in particular, there is a demand for appropriate learning or elicitation tools to transform user-specified preferences into soft constraints. Finally, preferences are more variegated than those that can be expressed with current soft constraint formalisms, and therefore further extensions of these frameworks are likely to emerge.

It is quite probable that neural constraint-satisfaction models will be extended to other psychological phenomena. These might include cognitive-balance theory, belief maintenance and perseverance, problem solving, and additional neuropsychological deficits that have not yet been explored.

For both new and current domains, it can be expected that constraint-satisfaction models will be run, not only to cover existing psychological phenomena but also increasingly to predict new phenomena that could guide empirical research.

Efforts might continue to expand the limited memory capacity of current neural constraint-satisfaction networks. For example, the number of patterns that Hopfield networks can store and accurately retrieve is only 13.8% of their number of units (Hertz et al., 1991). Even an entirely local representation scheme, with one unit per item, is more efficient than that. In addition, most multilayered feed-forward neural networks, using distributed representation schemes, are far more efficient than local representation schemes.

Neural network modelers might use some of the heuristics discovered in symbolic constraint-satisfaction research to enhance the power of neural models when faced with more difficult problems. There may also be some deeper relations between symbolic and neural constraint satisfaction that could be explored.

As the complexity of computational models increases, constraint-satisfaction neural networks could form important modules in these more complex models. Their robust ability to memorize, categorize, reduce noise, integrate, and so on makes them promising candidates as building blocks in complex models. An important challenge is to understand how such networks can be controlled in larger models. Such control could involve shifting the network state to another point on an attractor, or possibly require changing from one attractor type to another (Eliasmith, 2007).

The neurophysiology of constraint satisfaction could be more fully explored to the mutual benefit of both modeling and brain research (Wills et al., 2005). Evidence has been found for attractors in the olfactory system of rabbits and cats and in the visual cortex of monkeys. Brain imaging of humans suggests that the anterior cingulate cortex is involved in the monitoring of cognitive conflict and interference, and this can be modeled by neural-network energy. If the human brain is capable of soft, symbolic constraint satisfaction, it should be possible for realistically powerful neural networks to implement such processes. This could require an interesting three-way collaboration of AI, neuroscience, and constraint-satisfaction neural networks.

We conjecture that constraint satisfaction is more common than most cognitive modelers realize. Many of the issues that people struggle with on a daily basis can be usefully construed as CSPs. In this context, we recommend that researchers consider using powerful constraint-satisfaction algorithms as tools for humans. Successful models could possibly become useful adjuncts to human CSP solving by extending, rather than modeling, human abilities. Automating memory-intensive aspects of constraint satisfaction could be particularly helpful.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Cognitive Modeling: Connectionist Approaches; Problem Solving and Reasoning, Psychology of; Problem Solving and Reasoning: Case-based.

Bibliography

Bistarelli, S., Montanari, U., Rossi, F., 1997. Semiring-based constraint solving and optimization. Journal of the ACM 44, 201–236.

Eliasmith, C., 2007. Attractor network. Scholarpedia 2, 1380.

Fargier, H., Lang, J., 1993. Uncertainity in constraint satisfaction problems: a probabilistic approach. Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty, 97–104.

Freuder, E.C., Wallace, R.J., 1992. Partial constraint satisfaction. Artificial Intelligence 58, 21–70.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 72 I–74 I.

Hertz, J., Krogh, A., Palmer, R.G., 1991. Introduction to the Theory of Neural Computation. Addison Wesley, Reading, MA.

Hinton, G.E., 2007. Boltzmann machine. Scholarpedia 2, 1688.

Holyoak, K.J., Thagard, P., 1989. Analogical mapping by constraint satisfaction. Cognitive Science 13, 295–355.

Hopfield, J.J., 2007. Hopfield network. Scholarpedia 2, 1977.

Lecoutre, C., 2009. Constraint Networks: Techniques and Algorithms. Wiley, Hoboken, NJ.

Rosenfeld, A., Hummel, R.A., Zucker, S.W., 1976. Scene labeling by relaxation operations. IEEE Transactions On Systems, Man, and Cybernetics 6, 420–433.

Rossi, F., Beek, P.V., Walsh, T., 2006. Handbook of Constraint Programming: Foundations of Artificial Intelligence. Elsevier, New York, NY.

Rumelhart, D.E., Smolensky, P., Mcclelland, J.L., Hinton, G.E., 1986. Schemata and sequential thought processes in PDP models. In: Rumelhart, D.E., Mcclelland, J.L. (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge, MA.

Russell, S., Norvig, P., 2009. Artificial Intelligence: A Modern Approach. Prentice Hall, Upper Saddle River, NJ.

Schiex, T., Fargier, H., Verfaillie, G., 1995. Valued constraint satisfaction problems: hard and easy problems. International Joint Conference on Artificial Intelligence 1, 631–637.

Shultz, T.R., Lepper, M.R., 1996. Cognitive dissonance reduction as constraint satisfaction. Psychological Review 103, 219–240.

Wills, C.L., Cacucci, F., Burgess, N., O'keefe, J., 2005. Attractor dynamics in the hippocampal representation of the local environment. Science 308, 873–876.

Constructivism and Constructionism: Methodology

Robert A Neimeyer and Carlos Torres, University of Memphis, Memphis, TN, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.A. Neimeyer and H. Levitt, vol. 4, pp. 2651–2654, © 2001, Elsevier Ltd.

Abstract

Despite their diversity, constructivist and social constructionist methods share a common concern with revealing the personal or 'local' meanings that characterize an individual or group, and strive for pragmatic utility rather than objective veracity in the usual sense. To exemplify this approach, this article reviews representative constructivist methods, including repertory grid technique, the coding of narrative processes, discourse analysis of public rhetoric, and focus group methodology. Each is shown to have broad applicability and special relevance for research and practice in psychotherapy.

What makes a particular research method 'constructivist' or 'social constructionist' in orientation? In a sense, the answer lies less in the method or technique itself than in the philosophy with which the technique is used. For example, an interview can be conducted to count the number of positive self-statements in the interviewee's verbal output, or to classify him or her according to some extant diagnostic system. However different these aims might be from one another, both could be considered to express traditional research goals associated with applying or creating universally 'true' or valid categories for human experience. In contrast, other interviews might be conducted to tease out a network of personal meanings unique to a particular respondent, to reveal his or her shifting reliance upon different narratives over the course of therapy, to raise questions about the apparent coherence of a particular line of political discourse, or to grasp more clearly processes of consensus or dissent in a group focused on a particular issue. Such interview methods might well be constructivist, as they focus on distinctive patterns or processes of constructing (or deconstructing) meanings in a given personal or social context, without the presumption of universality that attends traditional knowledge claims in psychological science.

More specifically, a research method might be classified as constructivist or social constructionist to the extent that it (1) elucidates 'local' as opposed to 'universal' meanings and practices, (2) focuses upon provisional rather than 'essential' patterns of meaning construction, (3) considers knowledge to be the production of social and personal processes of meaning making, and (4) is more concerned with the viability or pragmatic utility of its application than with its validity per se. This emphasis on local, provisional, and pragmatic assessment of (inter)personal meanings can be illustrated by a closer consideration of four core techniques associated with a constructivist or social constructionist approach, each of which encompasses many different variations. These include repertory grid technique, analysis of narrative processes in spoken or written 'text', discourse analysis, and focus group methodology.

Repertory Grid Technique

Developed within personal construct theory, the repertory grid technique represents a widely used method for studying personal and interpersonal systems of meaning. Because of their flexibility, repertory grids (or repgrids) have been used in literally thousands of studies of a broad variety of topics, ranging from children's understandings of physical science principles and consumer preferences to formal structures of self-reflection within cognitive science and the mutual validation of belief systems between friends. However, their most consistent area of application has probably been in the clinical domain, where grids have been used to assess the meaning systems of clients in psychotherapy, and how these change over the course of treatment. The 'reptest' was initially designed by George Kelly, the author of personal construct psychology, as a means of assessing the content and structure of an individual's repertory of role constructs – the unique system of interconnected meanings that define his or her perceived relationships to others. In its simplest form, the reptest requires the respondent to compare and contrast successive sets of three significant people (e.g., my mother, my father, and myself), and formulate some important way in which two of the figures are alike, and different from the third. For example, if prompted with the above triad, a person might respond, 'Well, my father and I are very self-reliant, whereas my mother is very dependent on others.' This basic dimension, self-reliant versus dependent, would then be considered one of the significant themes or constructs that the person uses to organize, interpret, and approach the social world, and to define his or her role in it. By presenting the respondent with a large number of triads of varying 'elements' (e.g., a previous romantic partner, best friend, a disliked person, one's ideal self), the reptest elicits a broad sampling of the personal constructs that constitute the person's outlook on life and perceived alternatives. These constructs can then be interpreted impressionistically, used as the basis for further interviewing of the respondent, or coded using any of a number of reliable systems of content analysis, conducted either manually or using available computer programs.

Although the analysis of construct content is often revealing, most contemporary users prefer to extend the method beyond the simple elicitation of constructs by prompting the respondent subsequently to rate or rank each of the elements (e.g., people) on the resulting construct dimensions. For example, a respondent might generate a set of 15 constructs (e.g., self-reliant vs. dependent, hardworking vs. lazy, emotionally controlled vs. out of control), on which 10

important elements (e.g., my mother, father, self, partner) could be rated, yielding a matrix of 150 specific ratings that would then be amenable to a wide range of analyses. Although the repgrid was originally devised as an interview-based or paper-and-pencil measure, most contemporary users rely on any of a number of computer programs for their elicitation and analysis, such as the popular WebGrid 5 program available via the Internet.

Although analysis of the grid can focus on simple element ratings (e.g., observing that the respondent views himself or herself as self-reliant, hardworking, emotionally controlled, etc., on a sample of his or her personal constructs), it is typically more helpful to conduct a comprehensive analysis of the grid to discern larger patterns. This might involve correlating and factor analyzing the matrix of ratings to see at a glance which constructs 'go together' for the respondent (e.g., self-reliance is associated with being psychologically healthy, whereas dependence implies weakness), or to learn what people are most and least alike in the respondent's view. These linkages among constructs often suggest why people remain 'stuck' in symptomatic patterns, as when a client resists viewing himself as happy instead of depressed, because the former is associated with being superficial as opposed to deep, or selfish as opposed to selfless. Similarly, patterns of identification among elements in a grid can be clinically informative, with some of these (e.g., degree of correlation between actual self and ideal self) providing useful indices of progress in psychotherapy, while others suggest the person's psychological closeness or similarity to others.

At the most general level, grids can be useful for identifying how 'tightly' or 'loosely' a given person's meaning system is organized, as revealed, for example, by the average degree of correlation among the constructs that comprise it. Like most constructivist methods, repgrids focus on personal or local (e.g., family or organizational) systems of meaning, with no assumption that these generalize to other people. Moreover, repgrids measure construct systems at a particular moment in time, demonstrating meaningful sensitivity to variation across education, psychotherapy, or industrial organizational consultation, rather than presuming to reflect invariant psychological traits or essences. Finally, most users of repertory grids consider the ultimate hallmark of their value to be the degree to which they prompt relevant insights for the respondents who complete them, although some investigators have also demonstrated their validity and reliability in more conventional terms. Assessment methods having some affinity to the grid technique include the laddering technique, which elicits the higher order value implications of the respondent's constructs of behaviors and social roles, and the self-confrontation method, which evaluates an individual's sense of agency and communion with others through an analysis of important life events. A common feature of these methods is their focus on personal structure, viewing meaning in systemic terms that can be depicted visually to enhance its comprehension and discussion by clients or research participants.

Narrative Analysis

Narrative psychology considers thinking processes to be guided by the structures and elements of which stories are composed. Instead of viewing logic as the guiding force behind information processing and meaning making, narrative psychologists view people as organizing life events into plot structures and creating themes that give them significance. This model of organization can be classified as constructivist due to the agentic, creative processes involved in the formation of narratives.

Individuals create stories to help them comprehend and communicate their experiences and interactions in the world. Narratives are generated through the use of language, common scripts, and symbols derived from different cultures and traditions, and therefore are necessarily context dependent. Stories produced are seen as reflecting truths unique to their individual authors and contexts and as rooted within distinctive lived histories.

The narrative perspective has become popular over the last 30 years, and has grown rapidly into a movement in psychology, partly due to its broad relevance across cognitive, developmental, and clinical areas of research. In the cognitive domain, investigators have demonstrated that people inherently organize ambiguous experiences using story schemas and have studied the way that narrative formation structures memory as well as perception of events. Developmental researchers have traced the emergence of children's storytelling competencies, and their progressive shift toward accounting for the perspective of listener versus narrator. Clinical psychologists study how people construct stories to account for difficult life events, such as traumatic life experiences and the death of a loved one, and in so doing find meaning in them. In the domain of psychotherapy, researchers have developed methods for mapping the topical and thematic shifts that characterize client accounts of their experience, relating these to good and poor outcomes. The Narrative Process Coding and System illustrates the use of constructivist methods to identify and characterize narratives from transcribed psychotherapy sessions.

First, an investigator using the system segments the session dialogue into topic units that are identified through shifts in protagonists and themes. For instance, a client might begin a session discussing a conflict he or she is having with a colleague and then shift to examining a marital issue. The identification of topic segments can allow researchers to consider how themes are maintained or changed through discussion and how the interpersonal psychotherapeutic process can act to facilitate narrative development. Second, the researcher codes these topic segments into one of three narrative processes. External narrative sequences are dominated by event description. The recounting of the steps entailed in registering for university would be an example of this narrative sequence type. Internal narrative sequences focus upon emotional and experiential states. For instance, a portrayal of the different feelings present on the first day of classes would exemplify this narrative process. Finally, reflexive narrative sequences entail analysis and interpretation of events and internal reactions in order to understand their significance. An exploration of the meaning for the student of attaining a higher education would be classified as reflexive. Within different psychotherapy orientations, research has suggested specific patterns of narrative process associated with positive outcome. For instance, a shift between internal and reflexive exploration has been associated with successful humanistic psychotherapy.

Like other methods, whether or not the narrative process coding system is 'constructivist' depends on the epistemic intent that guides its use. For example, the system can be used within an objective, experimental model of science to compare the distribution of narrative sequences within different transcripts. Within such a model, results can be tabulated and statistical arguments for generalization of findings can be developed that aspire toward more universal implications. Although useful in some respects, such an application departs from the emphasis on individuality that characterizes most constructivist research. Alternately, this narrative tool can be used in a discovery-oriented framework, and themes identified that represent local, provisional truths. For instance, dominant narrative themes could be identified within the speech or writing of women with eating disorders on the topics of food and body image, and these themes could be explored in relation to the lives of these individual women and their social contexts.

Other narrative coding methods include the identification of core conflictual relationship themes implicit in client accounts of problems with other people; the innovative moments coding system, which tracks moments of change in a client's self-narrative across treatment; and the coding of interviews to establish the degree to which the narrator experiences himself or herself as an origin or pawn in connection with important events. In each case, the epistemic framework within which findings are presented and interpreted will dictate whether or not a method is being used in a constructivist spirit.

Discourse Analysis

Constructivist methods tend to focus on the meaning-making efforts of individuals, families, and small groups, whereas social constructionist methods shift attention to the broader systems of 'languaging' that characterize public speech, disciplinary discourse, or cultural contexts. Thus, the repertory grid method and its ability to uncover an individual's thematic constellation of meanings cogently follows the premises proposed by George Kelly's personal construct psychology, whereas social constructionist methods use social interactions as their unit of analysis and typically follow theoretical work initially presented by Berger and Luckmann. These two sociologists proposed that the study of 'everyday' or 'commonsense' knowledge should be the primary focus of any attempt at studying the social structures that guide individual as well as collective thought and action. Their proposal that knowledge is created and transmitted through social interaction argued for an ontological stance (that is, a claim about the nature of our world) that required an interpretivist epistemology that probed for the idiosyncratic interpretations and negotiations of meaning resulting from everyday practices using common language. Even when applied to individual language users, such methods tend to have a critical thrust, such as when careful 'deconstructive listening' is used to bring to light the hidden assumptions that constrain a speaker's view of what positions are permissible for self or others. Social constructionist methods are in this sense more politically radical than constructivist methods, attempting to uproot oppressive and marginalizing forms of thinking and speaking that have dehumanizing consequences.

An example of social constructionist methods is provided by discourse analysis, the critical evaluation of social 'texts' to reveal their concealed constraints and contradictions. For example, analyses of published debates in the South African parliament in the early 1980s have elucidated the shared assumptions about the superiority of whites held by both liberal and conservative Members of Parliament. Likewise, feminist discourse of the 1970s has been 'deconstructed' to expose the ways in which 'butch/femme' lesbians were marginalized to support a politic that advocated lesbian androgyny. Interestingly, discourse analysts have been especially active in undermining traditional assumptions in psychology, such as the largely unexamined convictions that people ideally function as rational 'selves' and that psychotherapy involves a move toward greater 'reality contact.' As a method, discourse analysis appears to be more an art than a science, borrowing from deconstructive and critical literary theory in order to understand and resist 'dominant narratives' that limit the terms of debate or the permissible range of human variation that is considered 'normal.' It is therefore sometimes used by psychotherapists concerned with helping clients free themselves from the oppressive narratives of their lives, and recover a sense of themselves as the authors of their experience.

Its social constructionist underpinning situates discourse analysis as a potent methodological choice for sociologists as well. By examining the threads of speech acts among individuals, sociologists uncover the metanarrative language patterns that weave together to negotiate societal meanings and group identities as well as maintain institutional power. Through a social constructionist lens, identity is viewed as relationally formulated and coordinated via discourse. It is therefore seen as more fluid than stable and best recognized as a contextualized performance on the social stage rather than the manifestation of an internal and idiosyncratic psyche. Through identity performance, individuals enact social values and communicate membership. Discourse analysis has thus aided sociological researchers in examining how groups such as certain Native American tribes construct norms for being a 'recognizable Indian' or how subcultures such as 'nerd girls' in high school negotiate their gender and societal identity through particularly defining speech acts. Social theorist Michel Foucault's view of discourse as an organizing structure that ultimately imposes rules and limits on what can and cannot be said has influenced many sociologists to examine power and status relations via language. Medical and mental health institutions have been widely studied in this regard.

The social constructionist effort to undermine traditional assumptions within psychology in a seemingly more 'artistic' than 'scientific' manner culminates in the use of the ethnodrama, a dramatic performance scripted from field data such as qualitative interviews, participant observation, and artifact collection. Embracing the critical spirit of deconstruction, ethnodrama aims to disrupt the 'researcher–subject' dichotomy. As performance, immediate and transitory, ethnodrama challenges the latent notions of power and privilege afforded to the researcher, who is otherwise left unchallenged to interpret the lives of those under study. In a performative space, meaning is synergistically spoken and embodied. For example, the

ethnodramatist might coach a group of Hispanic young people to enact their own story of life in their barrio, preserving and amplifying their voices, frustrations and aspirations, rather than dominant stereotypes of their lives, in a way that makes these more audible to observers. Scripts are prepared to be performed as written or left open so that performers can improvise. Actual text from interviews creates dialogue; thematic analysis of field setting data creates plot and structure. Although it is the researcher who writes the script, it is the performers who give it life. Thus, ethnodramatists attempt to bring individuals from both sides of the divide, researcher and participant alike, closer to the multiple meanings contained within the data. In this way, the methodological rationale is informed by the conceptualization of 'self' as proposed by more radical social constructionists: although an individual's ability to feel emotion or think a thought is not questioned, it is argued that the comprehensibility, meaning, and thus 'existence' of any affective or cognitive response are directed by a larger social discourse. Notions of 'self' materialize through repetitive acts sanctioned by cultural norms. We are all performers navigating meaning through our social roles. The researcher is only distinguished from the participant by title, and there is no privileged voice. The ultimate goal of the ethnodrama is to create a space that will empower and promote self-reflection and possibly subversion of oppressive metanarratives.

Something of an exception to this trend toward literary/dramaturgical rather than conventionally scientific methodology is the refinement of conversational analysis, which involves detailed transcription of dialogue between two or more speakers to capture not only the content of the exchange but also points of overlap in speech turns, pauses, lengthening or abbreviation of utterances, and other verbal and coverbal features of oral exchange. Such analyses have been applied, for example, to the negotiation of novel prompts in therapy, as when therapists invite their clients to consider how life would be different if a miracle occurred and their problem disappeared overnight, yielding a clearer understanding of how the two collaborate to open new possibilities for client change. Related social constructionist methods include the analysis of rhetoric, the tropes or moves used to advance a persuasive argument, and textual analysis, which reveals the way in which authors skillfully position themselves and others within the moral discourse of an account, so as to present themselves to best advantage.

Focus Groups

Focus groups first appeared in the 1930s and were used to investigate public reaction to propaganda messages. The methodology was subsequently picked up and propagated by market researchers. It was not until many years later that such groups were used by social scientists. Through its long history the methodology has remained essentially the same: a focus group brings together a small group of people, all of whom share a particular characteristic (e.g., all are registered Democrats or bereaved parents), and engages them in a focused discussion on a particular topic. For example, a researcher might decide to put together a focus group comprising middle-aged adults whose parents have Alzheimer's dementia to better understand the emotional coping strategies used by grown children when taking care of their parents.

Many researchers use focus groups as a cost-effective methodology capable of collecting a wide scope of responses in a single session. When used for this purpose, the focus group is akin to something of a group interview and the resulting data set is regarded as a compilation of discrete responses. However, the methodology's inherently social nature and its ability to engage individuals in deep, sustained discussion make focus groups particularly useful to researchers adhering to a social constructionist view of the meaning-making process, which is seen as occurring through the relational processes of language and culture. Through agreement and disagreement, affirmation and challenge, question and comment, and verbal and nonverbal performance, participants create a social dynamic through which meaning is personally negotiated and socially agreed upon. Thus, from a social constructionist stance, group interaction, in its totality, is the relevant unit of analysis, and not just individual, isolated commentary. The patterns of discourse and interaction shed light on larger sociocultural process of knowledge acquisition for a particular group.

A successful focus group engages participants, encouraging them to share their views and consider the perspective of the others. To this end, careful pregroup planning is crucial. Researchers must consider group size and membership. The recommended group size is between 4 and 12 participants, large enough to include diverse voices but small enough to allow each participant to speak. Group membership considers heterogeneity or homogeneity for personal characteristics such as age, status, etc. Although group membership is usually shaped by the research question, some researchers consider generalizability in their selection process and maximize demographic diversity. Because constructionist researchers embrace the notion of multiple, localized truths and eschew representing their findings as an exhaustive and definite presentation of the phenomena, they are more willing to work with 'biased' groups. Working with an existing group like a gang, for example, allows the researcher access to their slang and mannerisms. Similarly, working with a group others would consider too heterogeneous and thus likely to decline full disclosure, such as a group comprised of teachers and students, allows the researcher to examine nonverbal gestures, pauses, silences, etc., in order to understand how authority negotiates what can and cannot be discussed openly.

Although a rich data set relies on participant interaction, the group facilitator's presence influences group outcome as well. First, much like group size and membership, the script must be thoughtfully prepared and rehearsed in advance. Open-ended questions are the typical preference. Once participants begin to respond to the questions and to each other, the facilitator's role turns to that of a traffic controller, directing discussion and maintaining a sense of order so that all comments and exchanges remain audible and intelligible. Social constructionist researchers acknowledge that the facilitator's presence and style affects group interaction. Rather than attempting to control 'noise' or 'bias,' researchers of this ilk engage in reflexivity, a self-reflective process by which they clarify their own biases and subjectivities and their effect on the data collection process. Reflexivity allows qualitative researchers to properly situate themselves when interpreting their findings.

There are various strategies at a researcher's disposal for analyzing focus group data. The most common strategies employ text analysis of focus group transcripts, while others examine nonverbal data. Analysis may be time consuming, as it can involve transcribing recordings and coding data. Nonetheless, use of focus group methodology is increasing in the social sciences because of the rich data set it offers.

In summary, constructivist and social constructionist methods in psychology are multifaceted in form and diverse in application, but tend to express a shared commitment to revealing the structure of personal and collective meanings and the processes by which they are constructed. Given their convergence with postmodern trends in the human sciences, it is likely that their influence will continue to grow in the future.

See also: Discourse and Indentity; Discourse, Anthropology of; Narrative, Sociology of; Sexual Orientation: Historical and Social Construction; Social Constructionism.

Bibliography

Angus, L., McLeod, J., 2004. The Handbook of Narrative and Psychotherapy. Sage, Thousand Oaks, CA.

Barbour, R.S., Kitzingen, J. (Eds.), 1999. Developing Focus Group Research. Sage, London.

Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality. Doubleday, Garden City, NY.

Bucholtz, M., 1999. "Why be normal?": language and identity practices in a community of nerd girls. Language and Society 28, 203–223.

Caputi, P., Viney, L.L., Walker, B.M., Crittenden, N., 2012. Personal Construct Methodology. Wiley-Blackwell, Chichester, UK.

Lock, A., Strong, T., 2010. Social Constructionism. Cambridge, New York.

Lock, A., Strong, T., 2012. Discursive Perspectives in Therapeutic Practice. Oxford, New York.

Luborsky, L., Crits-Christoph, P., 1998. Understanding Transference: The Core Conflictual Relationship Theme Method. American Psychological Association, Washington, DC.

Neimeyer, R.A., 2009. Constructivist Psychotherapy. Routledge, New York.

Turner, B.S., 1995. Medical Power and Social Knowledge, second ed. Sage, London.

Wieder, D.L., Pratt, S., 1990. On being a recognizable Indian among Indians. In Carbaugh, D. (Ed.), Cultural Communication and Intercultural Contact. Erlbaum, Hillsdale, pp. 45–64.

Constructivist Theory and Psychotherapy

Robert A Neimeyer, University of Memphis, Memphis, TN, USA
David A Winter, University of Hertfordshire, Hatfield, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Constructivist psychotherapy is not so much a single model of therapy as it is a metaperspective that informs many contemporary approaches to psychological treatment. This article reviews several traditions of clinical practice that are informed by a constructivist emphasis on meaning-making in personal, family, social and linguistic domains, and notes their extension in the work of numerous international contributors to this perspective.

Introduction

Joined by a shared epistemology emphasizing the active, form-giving role of the mind and language in constructing reality, constructivist psychotherapies encompass a broad range of contemporary clinical perspectives. Beginning with personal construct theory, (Kelly, 1955/1991) which focuses on the idiosyncratic ways in which individuals construe themselves and others, constructivism has grown to influence therapy traditions ranging from the psychodynamic to the cognitive behavioral, and from the humanistic to the systemic. Among the chief contributions of constructivism are its attention to the dialectical processes of identity development, procedures for illuminating the tacit or nonconscious dimensions of lived experience, and methods for studying and facilitating the human penchant toward meaning-making through the construction of narratives. Although external critics express concerns about constructivism's erosion of the idea of objective truth and internal critics dispute the individual versus social basis of the construction of meaning, the growing coalition of constructivist approaches has contributed a fund of novel methods to psychotherapy research and practice.

Constructivist Philosophy

As a philosophical position that emphasizes both personal and social processes of meaning-making, constructivism has influenced several contemporary traditions of psychotherapy (Neimeyer, 2009). For this reason, it is more accurate to consider constructivism as a general approach to understanding people, conceptualizing psychological distress, and fostering human change than to view it as a distinctive 'school' of psychotherapy associated with a particular theorist, preferred method, or specific set of problems requiring treatment. Thus, constructivism is best viewed as a 'metatheory' that has influenced many late twentieth century and early twenty-first century developments in clinical theories as diverse as psychoanalysis, existential-humanistic psychotherapy, cognitive-behavioral therapy, and family systems approaches. In addition, a number of novel psychotherapies have been devised along avowedly constructivist lines, from personal construct psychotherapy, derived from a theory that was formally set forth in the 1950s, to narrative and dialectical therapy approaches that have become prominent only in the last 15–20 years. Appreciating the contributions of these diverse models, therefore, requires a consideration of their core philosophic similarities and the range of concepts and strategies that shape their expression at the level of clinical practice.

If there is a unifying theme that links constructivist forms of psychotherapy, it is at the level of their epistemology, or theory of knowledge. As might be expected, there are alternative forms of constructivism, which vary in the extent to which they acknowledge that a 'real world' exists outside of human consciousness or language. Nevertheless, they are all much more interested in the nuances in people's construction of the world than they are in evaluating the extent to which such constructions are 'true' in representing a presumably external reality (Neimeyer, 2009). This emphasis on the active, form-giving nature of the mind dates back at least to the Italian rhetorician and historian Giambattista Vico (1688–1744), who traced the development of thought to the attempt to understand the world by projecting upon it human motives, myths, fables, and eventually linguistic abstractions. The German philosopher Immanuel Kant (1724–1804) likewise emphasized the transformative character of the mind, which necessarily imposes spatial, temporal, and causal order on the phenomena of experience. From these philosophers, constructivists borrowed a model of knowledge as an active structuring of experience, rather than a passive or receptive assimilation of a reality of 'things in themselves,' uncontaminated by human knowing.

At the threshold to the twentieth century, these themes were elaborated by the German analytic philosopher Hans Vaihinger (1852–1933), whose *Philosophy of 'As If'* asserted that people develop 'workable fictions' (e.g., of mathematical infinity or God) to order and transcend the hard data of experience, and establish distinctively human goals. A similar emphasis on the distinction between our linguistic 'map' of experience and the 'territory' of the world was made by the Polish intellectual Alfred Korzybski (1879–1950), whose system of general semantics focused on the role of the speaker in assigning meanings to events. From these thinkers, constructivists drew the implication that human beings operate on the basis of symbolic or linguistic constructs that help them navigate in the world without contacting it in any simple, direct way. Moreover, they suggested that such constructions are viable to the extent that they help us live our lives meaningfully and find

validation in the shared understandings of others in our families, communities, and societies. 'Postmodern' thinkers who follow in this constructivist vein further stress the extent to which we live in a world constituted by multiple social 'realities,' no one of which can claim to be 'objectively' true across persons, cultures, or historical epochs. Instead, the constructions on the basis of which we live are at best provisional ways of organizing our 'selves' and our activities, which could under other circumstances be constituted quite differently.

Constructivist Psychotherapies

The first person to develop a thoroughgoing theory of psychotherapy along constructivist lines was the American clinical psychologist, (George Kelly 1955/1991). Working in the relative isolation of rural Kansas in the 1930s and 1940s, Kelly confronted the overwhelming psychological needs of farming communities that had been devastated by the twin crises of the Dust Bowl and the Great Depression. This prompted Kelly to design efficient psychotherapeutic procedures in which clients were coached to enact fictional identities in their daily lives for a fixed period of time (usually only 2 or 3 weeks). Working from a 'self-characterization' written by the client, the therapist and a consulting team first drafted an 'enactment sketch' of an imaginary person whose way of construing life differed from, but did not necessarily contradict, the client's own. The client and therapist then practiced this novel role in a series of therapeutic enactments with the therapist playing various acquaintances, authority figures, and family members, while the client simultaneously generalized the role to the actual relationships in his or her life. At the end of this period, the role was consciously set aside, and the therapist and client discussed the extent to which the client's sense of self and social world were permeable and provisional constructions that might be lived out quite differently if viewed in alternative ways. Kelly's 'fixed role therapy' was therefore the first form of brief therapy, and it foreshadowed the use of dramatic and narrative strategies of change incorporated in many contemporary constructivist therapies (Winter and Viney, 2005).

Eventually, Kelly drafted a comprehensive psychology of personal constructs that placed these procedures in a rigorous theoretical context, and suggested diagnostic, therapeutic, and research methods targeting the unique meaning systems that individuals devised to structure and anticipate the themes of their lives (Kelly, 1955/1991). He was the first clinician to make explicit use of a process of case formulation, in which the therapist uses 'professional constructs' in an attempt to understand the client's view of the world, and as a basis for planning therapeutic intervention. Moreover, he originated repertory grid technique to map the client's network of personal constructs, helping both therapist and client identify areas of the client's meaning-making that warranted closer therapeutic attention. Subsequently, repertory grid methods have been greatly extended, implemented in user-friendly computer programs, and applied to a vast range of topics in clinical, personality, social and organizational psychology, through the efforts of an international network of contributors such as Richard Bell, Devi Jankowicz, Finn Tschudi, and David Winter (Winter, 2003).

Interest in Kelly's work remains strong throughout North America, Europe, and Australasia, with contemporary personal construct theorists challenging, broadening, and applying his initial theory to a range of current problems and disorders. The methods that they have developed include brief therapies that aim to facilitate specific changes in an individual's construing, such as the resolution of dilemmas, as developed by Guillem Feixas and his colleagues in Spain; experiential approaches that pay greater attention to the therapeutic relationship, championed particularly by the American psychologist Larry Leitner; group therapy techniques, for example as developed by Kelly's student, Al Landfield; procedures that focus on the shared construct systems of couples, families, or broader social systems, as pioneered by the British personal construct psychologist Harry Procter; narrative elaborations of the personal construct approach, such as that developed by the Italian psychotherapists Gabriele Chiari and Maria Laura Nuzzo; and nonverbal techniques, many inspired by the British educational psychologist Tom Ravenette, but as applicable with adults as with children (Caputi et al., 2012; Chiari and Nuzzo, 2010). Finally, several contemporary scholars have mined the implications of a personal construct view for specific counseling contexts, such as the American psychologists Mark Savickas's (2011) approach to career counseling and Robert Neimeyer's (2012) meaning reconstruction model of grief therapy. A significant feature of this newer work is that it characteristically links classical Kellian themes of personal meaning systems, human agency and a creative, experimental approach to therapy to other trends in humanistic, systemic and social constructionist thought in a way that has fostered the emergence of a more integrative constructivist approach to therapy.

A second tradition of constructivist therapy began to gain momentum in the late 1970s, as an outgrowth of cognitive-behavioral therapies emphasizing the role of the person's interpretations of events in a range of disorders. Michael Mahoney, an American pioneer in the cognitive trend, began to critique its reliance on rationalistic epistemologies that presumed that emotional adjustment was a straightforward matter of making one's cognitions realistic and in line with an observable world. In particular, Mahoney (1991), along with colleagues such as the Italian theorist Vittorio Guidano (1991), began to turn attention to the 'core ordering processes' by which persons construct and maintain a sense of self in a social field. This gave rise to more developmental forms of constructivist psychotherapy, which focused on the periodic perturbation of an individual's conscious identity through confrontation with difficult or traumatic live events. With the creative linguistic contributions of American psychologist Donald Meichenbaum, these therapies began to develop a variety of procedures for enhancing clients' self awareness, empowerment, and adaptability, while helping them reconstruct an emotionally coherent sense of their identities across time. Significantly, this emphasis on identity construction redirected attention to the often tacit, intuitive, and nonconscious processes by which people organize their experience of self and others, extending the more common focus of cognitive-behavioral therapies on people's conscious 'inner dialogue' or 'automatic thoughts.' Finally, the incorporation of Buddhist mindfulness techniques into some 'third

wave' cognitive-behavioral approaches dovetails with the constructivist turn since, as American psychologist Spencer McWilliams points out, there are metatheoretical compatibilities between Buddhism and constructivism in the notion that what we take to be the 'self' and 'reality' are human constructions that are best held in a flexible and provisional fashion.

As the philosophy of constructivism became widespread, it began to permeate even long-standing traditions of psychotherapy, such as psychoanalysis. One harbinger of this trend was the American analyst Donald Spence, who argued that analysis could not unearth the historical truth of the patient's life, but only constitute a procedure for disclosing its narrative truth in the eyes of its author – the client. In keeping with the psychoanalytic tradition, however, constructivist analysts still tend to emphasize the usefulness of accessing and reviewing emotionally significant memories of one's early life, but view these less as veridical 'insights' than as inventions subject to the demand for narrative 'smoothing,' that is, the tendency to rewrite our remembered history in a way that corresponds to our dominant model of who we are.

Although it would not properly be considered a psychodynamic therapy, the coherence therapy developed by American psychologist Bruce Ecker shares with psychodynamic approaches a focus on unconscious meanings that can subvert a client's conscious intentions. Therapy in this vein uses a procedure of *radical questioning*, probing to reveal the deep purposes served by a client's symptoms or difficulties, to reveal the hidden *prosymptom position* that makes the problems important to perpetuate, despite the very real pain associated with them (Ecker et al., 2012). Like other contemporary constructivist approaches, this integrative model of therapy relies more heavily on experiential exploration of the client's meanings (e.g., using visualization exercises and direct statements of the client's 'emotional truth,' those meanings of the problem that account for his or her resistance to change), rather than on the interpretive stance preferred by classical psychoanalysts.

Another long-standing tradition to be influenced by a constructivist position is humanistic-existential psychotherapy, whose emphases on human choice and phenomenology or the study of experience converged with core themes in personal construct theory and related forms of constructivist theory. One outgrowth of this convergence is the *self-confrontation method* devised by the Dutch personality theorist, Hubert Hermans, which uses open-ended questions to elicit those 'valuations,' or important units of meaning that a person uses to structure a sense of her or his past, present, and future (Hermans and Hermans-Jansen, 1995). The client then rates critical life experiences on a standard list of affect terms, which yields measures expressing the degree of self-enhancement, union with others, negative emotions, and positivity associated with each, providing a basis for self-reflection or a method for mapping therapeutic changes in meanings over time. This approach ultimately has grown to integrate the postmodern notion of a 'dialogical self' (Hermans and Dimaggio, 2004) consisting of multiple voices, as in an earlier metaphor of 'community of self' used by the British personal construct psychologist Miller Mair. Therapy aims for the development of coalitions between these different voices or selves, and is a more general example of approaches that facilitate greater coherence in clients' narratives, the stories that they tell about themselves.

A second example of the infusion of constructivist themes into humanistic therapy is the emotion-focused approach of the Canadian psychologist Leslie Greenberg, which emphasizes the role of personal experience and internal conflict in the creation of subjective meaning. For example, therapy might attempt to synthesize two strongly felt and mutually contradictory internal experiences, such as self-contempt and self-prizing. This could involve helping the client vividly experience both stances through the use of a 'two chair' technique, in which the client successively speaks from each position and fosters their integration into a new structure (e.g., self-acceptance). In keeping with constructivist themes, the 'selfhood processes' that are the focus of these approaches are viewed as continually evolving and provisional, rather than a static 'essence' to be simply discovered or actualized.

In a similar way, a social constructionist perspective has revolutionized family systems therapies, as they have adopted a view of therapy as conversation whose goal is to alter the family's 'problem determined system.' In the work of American contributors Harlene Anderson and Harry Goolishian, the therapist functions less as an expert dispensing answers, than as a conversation manager who promotes the sort of exchanges among family members that 'dissolve' old problems by 'languaging' about them in a new way. That social constructionist approaches such as this may be usefully extended to people diagnosed as psychotic, if their psychotic symptoms are regarded as meaningful, has been demonstrated by Jaako Seikkula and his colleagues in Finland. The Italian clinical psychologist Valeria Ugazio has a somewhat different focus, emphasizing the semantic polarities, or polar opposite positions that characterize family conversations, including those with the therapist, and which she considers to be distinctive in different disorders (Ugazio, 2013).

Other family therapists have adopted the constructivist metaphor of lives as stories, and have devised novel means of helping clients free themselves from the 'dominant narratives' that originate in particular families and cultures, and that keep them from feeling like the authors of their own lives. For therapists such as Michael White in Australia and David Epston (1990) in New Zealand, externalizing the problem by reconstruing it as something separate from the client's 'self' provides a useful first step toward recognizing its destructive impact on clients and the relationships in which they are engaged. Therapy can then turn toward recognizing and validating those exceptional moments when a person begins to resist the dominant narrative, and rewrite his or her life story along more hopeful lines. A common theme in these systemic variants of constructivism is their emphasis on the resourcefulness of individuals, families, and communities in resolving problems in living. Accordingly, clinical work is cast as a search for solutions, rather than a diagnosis of dysfunction.

Finally, constructivist influences are apparent in some integrative approaches, such as cognitive analytic therapy, developed by British psychiatrist Anthony Ryle, who drew particularly upon personal construct theory. Indeed, constructivism can be

the basis for a 'theoretically progressive integration,' combining different therapeutic approaches that share constructivist meta-theoretical assumptions.

Critical Perspectives

Despite the creative contributions of constructivism to contemporary clinical practice, challenges to this perspective have been raised by philosophic realists who question what they view as its 'anything goes' relativism. From the vantage point of these external critics, the constructivist emphasis on multiple realities risks undermining conventional understandings of 'truth' as well as 'objective' procedures of research into therapy process and outcome. However, constructivists like American Jonathan Raskin have responded that a respect for the ways in which various people and cultures construe life differently accords well with the diversity and multiculturalism of contemporary life, and actually provides a psychological basis for twenty-first century ethics. Likewise, constructivists have carried out literally thousands of empirical studies, on topics as diverse as the structure of personal meaning systems and the intricacies of change in the process of psychotherapy. As a result, there is a growing evidence base for some forms of constructivist therapy, such as personal construct psychotherapy and emotion-focused therapy, with studies demonstrating comparable effect sizes to those in nonconstructivist therapies, a therapeutic process that is distinctively different from that in more rationalist therapies, and major differences in philosophical beliefs between practitioners of these contrasting approaches to treatment (Metcalfe et al., 2007). Significantly, these efforts have often required the development of new tools for revealing individual and conversational meaning-making, such as the role construct repertory grids and other innovative methods devised by personal construct theorists (Caputi et al., 2012; Neimeyer, 1993), inventories for measuring meaning-making in relation to troubling life experiences, and the narrative process (Angus et al., 1999) and 'innovative moments' coding schemes validated by psychotherapy process researchers (Gonçalves et al., 2010). Thus, it seems unlikely that constructivism will erode the disciplined study of psychological phenomena in an uncritical attempt to embrace all knowledge claims as equally valid.

Tensions also arise *within* the constructivist camp, particularly between personal construct theorists who champion the individual's role in meaning-making and social constructionists like American psychologist Ken Gergen and British theorist John Shotter, who emphasize the extent to which meanings reside in systems of culture and language that precede any given individual. Indeed, such debates are likely to continue, as they challenge the straightforward depiction of the self-as-agent endorsed by constructivists with more humanistic leanings. This same emphasis on the social and discursive construction of reality, however, has sensitized constructivism to issues of familial, ethnic, and cultural meanings, thereby enhancing its relevance and inclusiveness. Ultimately, the challenges faced by constructivist theories and therapies can be viewed as the predictable 'growing pains' encountered by a significant new perspective that promises to help shape the future, as well as the present, of psychotherapy.

See also: Experiential Psychotherapy; Feminist Psychotherapy; Narrative Networks; Psychotherapy Process Research; Psychotherapy and Hermeneutics.

Bibliography

Angus, Lynn, Levitt, Heidi, Hardke, Karen, 1999. The narrative processes coding system: research applications and implications for psychotherapeutic practice. Journal of Clinical Psychology 55, 1255–1270.
Caputi, Peter, Viney, Linda L., Walker, Beverly M., Crittenden, Nadia (Eds.), 2012. Personal Construct Methodology. Wiley, Chichester, UK.
Chiari, Gabriele, Nuzzo, Maria L., 2010. Constructivist Psychotherapy: A Narrative Hermeneutic Approach. Routledge, London.
Ecker, Bruce, Ticic, Robin, Hulley, Laura, 2012. Unlocking the Emotional Brain. Routledge, New York.
Gonçalves, Miguel M., Matos, M., Santos, A., 2010. Narrative therapy and the nature of "innovative moments" in the construction of change. In: Raskin, Jonathan D., Bridges, Sara K., Neimeyer, Robert A. (Eds.), Studies in Meaning, vol. 4. Pace University Press, New York.
Guidano, Vittorio F., 1991. The Self in Process. Guilford, New York.
Hermans, Hubert J.M., Hermans-Jansen, Elsa, 1995. Self-narratives. Guilford, New York.
Hermans, Hubert J.M., Dimaggio, Giancarlo (Eds.), 2004. The Dialogical Self in Psychotherapy. Brunner-Routledge, Hove, UK.
Kelly, George A., 1955/1991. The Psychology of Personal Constructs. Norton, New York. Reprinted by Routledge, London, 1991.
Mahoney, Michael J., 1991. Human Change Processes. Basic Books, New York.
Metcalfe, Christopher, Winter, David A., Viney, Linda L., 2007. The effectiveness of personal construct psychotherapy in clinical practice: a systematic review and meta-analysis. Psychotherapy Research 17, 431–442.
Neimeyer, Greg J. (Ed.), 1993. Constructivist Assessment. Sage, Newbury Park, CA.
Neimeyer, Robert A., 2009. Constructivist Psychotherapy. Routledge, London & New York.
Neimeyer, Robert A. (Ed.), 2012. Techniques of Grief Therapy: Creative Practices for Counseling the Bereaved. Routledge, New York.
Neimeyer, Robert A., Mahoney, Micahel J. (Eds.), 1995. Constructivism in Psychotherapy. American Psychological Association, Washington, DC.
Neimeyer, Robert, A., Meichenbaum, Donald, Stanley, Caroline, 2014. Developing a dialogue: constructivist convergence in psychotherapy and beyond. In: Raskin, Jonathan, Bridges, Sara (Eds.), Studies in Meaning, vol. 5. Pace, New York.
Neimeyer, Robert, A., Raskin, Jonathan (Eds.), 2000. Constructions of Disorder: Meaning-Making Frameworks for Psychotherapy, first ed. American Psychological Association, Washington, DC.
Savickas, Mark L., 2011. Career Counseling. American Psychological Association, Washington, DC.
Ugazio, Valeria, 2013. Semantic Polarities and Psychopathologies in the Family. Routledge, New York.
White, Michael, Epston, David, 1990. Narrative Means to Therapeutic Ends. Norton, New York.
Winter, David A., 1994. Personal Construct Psychology in Clinical Practice: Theory, Research and Applications. Routledge, London.
Winter, David A., 2003. Repertory grid technique as a psychotherapy research measure. Psychotherapy Research 13, 25–42.
Winter, David A., Viney, Linda L. (Eds.), 2005. Personal Construct Psychotherapy: Advances in Theory, Practice and Research. Whurr, London.

Relevant Websites

http://www.tandfonline.com/toc/upcy20/current#.Ugalhhbti0s – Official website for the Journal of Constructivist Psychology.
http://www.constructivistpsych.org – Site of the major professional association for constructivists in North America.
http://www.coherencetherapy.org – Website for Coherence Therapy training resources.
http://ijds.lemoyne.edu/index.html – Website for the International Journal of Dialogical Science.
http://www.pcp-net.org/journal/ – Website for the Journal Personal Construct Theory & Practice.
http://www.taosinstitute.net – Website for training resources and conferences in social constructionism.

Consumer Policy

Christoph Strünck, School of Arts and Humanities, University of Siegen, Siegen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Consumer policy aims at protecting and empowering consumers. Consumers are a very large group but they suffer problems of collective actions. This is why governments, consumer organizations, and companies apply tools such as bans, regulations, advice, litigation, or self-regulatory standards. Relations between companies and consumers are asymmetrical, with consumers lacking vital information. Additionally, behavioral economics reveal striking biases: habits, norms, and irrational attitudes shape everyday consumption. Governance of consumer policy is different, however, resting on different ideas and ideologies, as comparative politics shows. Complex digital markets, the service industry, and the quest for sustainable consumption are core challenges to consumer policy.

Introduction

Companies need consumers to sell. Consumers need companies to deliver. The relation between companies and consumers is not as adversarial as between employers and employees. Neoclassic economics holds that markets will provide whatever individual consumers long for: good quality, low prices, and choice at every level. Why consumer policy, then?

Consumers stand on the weak side of an asymmetric relationship. Struggling with information overload and short on time, comparing products is messy. Lots of information biases as well as behavioral biases shape consumption. This is why collective action of consumers is necessary to maintain effective markets. Supposing, mass of consumers buy products with minor flaws. No single consumer will take up efforts to complain or even to sue. This collective behavior leads to economic inefficiency. More fundamentally, people expect governments to safeguard against dangerous products and hazards. Last but not least, consumers also bear political responsibility because buying comes with price tags for others. This is why collective action is necessary.

Consumers are among the largest social groups in society, spanning all categories of social status, gender, age, and cultural attitudes. Consumption is both the basic need and symbol of expression. To consume is not only to buy a good, but also the way consumers use and recycle those goods is at least as important as purchase. Consumption is both individual choice and social behavior, it is deeply rooted in group thinking, norms, and tradition, and it is mostly embedded in households. Yet, consumers are not as well organized as employees, professions, or business groups. Their collective action problems exceed those of most other groups.

Consumers are a large and heterogeneous group that suffers a fundamental dilemma of free riding. Consumers neatly fit in what Mancur Olson called 'diffuse and weak interests' (Olson, 2003): nobody is willing to join collective action because everyone will benefit from it. This is why governments often support and subsidize consumer organizations, especially in Europe. In the United States, foundations and public interest law firms act as political entrepreneurs.

Additionally, governments act on behalf of consumers: they ban dangerous products, they enact safety and quality standards, and they regulate markets to protect consumers. Consumer policy cuts across different realms. As for the United States, agencies such as the Federal Trade Commission, the Environmental Protection Agency, or the Food and Drug Administration are in charge.

Consumer policy is not just government action and collective action of consumer groups. Companies pursue consumer policy, too. A classic example is self-regulation. In most advanced economies, there are self-regulatory bodies of industry. They monitor markets and blame competitors for wrongful practices.

Summed up, consumer policy aims at protecting consumers and empowering them to make informed choices. Governments, associations, and companies carry out consumer policy. Policy tools include bans, standards, obligations, education, information, and litigation. Consumer policy is mainly regulatory policy; it is less distributive or redistributive. Thus, costs and benefits are usually spread quite evenly. You might expect consumer policy to be a cherished subject for politicians, then.

However, consumer policy has to cope with a classic dilemma of cross-cutting policies (Stone, 2012). Most policies have clear boundaries in which government, business, and interest groups are separated from other constituencies. This is the case with social policy, environmental policy, or agricultural policy. With consumer policy, things are different. Consumer issues cut across a whole range of policies. On the one hand, regulatory capture is less likely because there is no 'iron triangle' of vested interests. On the other hand, comprehensive policies are less likely, too. Additionally, when costs are spread evenly it might provoke resistance from all sectors of business. Thus, similar to redistributive policies, consumer policy is shaped by conflicts and ideologies.

Are Consumers Rational: Ideas and Ideologies in Consumer Policy

Every policy faces cleavages that affect collective interests. As for consumer policy, it comes with a price tag for companies. They might have to disclose information on products and yield certain rights to consumers. Lawmakers argue that rules apply to everybody in order to maintain fair competition. Business

groups argue that consumers are sovereign to decide on their own. However, business groups are split over goals and instruments of consumer policy. Companies with high quality standards even support stronger rules to single out lemons.

Yet, the core conflict is about rational consumption. Do consumers act rationally; do they bear full responsibility for what they are doing? The paradigm of rational consumers is rooted in neoclassic economics. Equivalent to perfect markets, consumers are well informed and sovereign. However, institutional economics has hinted that transaction costs for consumers can be high. Additionally, institutional economics highlight asymmetrical relations between companies and consumers, as the famous 'market for lemons' by Akerlof has shown (Akerlof, 1970): buyers of used cars cannot correctly asses the quality of what they are buying. This problem systematically undermines effectiveness of markets. Only institutional rules can guarantee that consumers benefit from markets whose products they cannot compare as easily as textbook wisdom suggests.

Institutional economics paved the way for the 'information approach': consumer policy has to provide special information to make up for disadvantages of consumers. More recently, behavioral economics has cast doubt on the 'information approach.' Experiments show that framing, anchor effects, and habits shape consumption predominantly. Most consumers lack time and knowledge to compare, being stuck in routines and habits. Rational 'choice' is rare; often it is impossible (Albanese, 2006; Hantula, 2012). There are even effects of information overload that confuse consumers.

Additionally, there is no such thing as the average consumer. Social science has detected different types: confident consumers, responsible consumers, and vulnerable consumers (Gronow and Warde, 2001). Instruments and strategies of consumer policy need to be tailored to the needs of different forms of behavior.

These findings undermine conventional arguments on consumption. They also fuel ideological struggles in consumer policy. Should consumer policy just provide information or should it guide consumers? Does soft paternalism clash with individual choice and free markets? Should consumer policy focus on individual consumers or should it support group action? Is the general notion of 'consumers' in line with complex and heterogeneous societies?

Conflicts about ideas and ideologies have unfolded in different countries, embedding different approaches toward consumer policy. Ideas and ideologies not only include assumptions about consumers but also stretch to regulatory devices like the much debated precautionary principle.

Convergence or Divergence: Governance of Consumer Policy in Advanced Economies

For a long time, policies in the United States seemed to be more risk averse when it came to protecting consumers. Research reveals that the tide turned in 1990: since then it was rather the European Union and its member states that invented strict regulatory standards (Vogel, 2012). When it came to pharmaceuticals, chemicals, or food safety, US standards were strictest before 1990. Business groups complained about 'drug lags' and red tape. Deregulation and Reaganomics in the United States boosted cost–benefit analysis as a tool for risk regulation during the 1980s. Lobbying by business groups scaled back the precautionary principle further.

In the European Union, the common market brought about change. Following the Single European Act of 1986, regulatory competition unfolded in Europe. It often came with a 'California effect' (Vogel, 1997): influential countries with high standards of consumer protection served as blueprints. Subsequently, a race to the top went off. It also gave legitimacy to policies of deregulation. The European Commission championed these rules worldwide in order to achieve comparative advantages.

Theoretically and empirically, the precautionary principle is ambiguous. The rule of thumb 'better safe than sorry' can avert risks yet it can also deprive consumers of innovation (Sunstein, 2008). Anyway, the precautionary principle is not just meant to spare consumers risks. It might also serve as an equivalent for old school protectionism.

The history of the precautionary principle proves that motives and hidden agendas of consumer policy are manifold. Consumer policy can also boost the market position of national business. Yet governance is never rationally constructed. It results from a mix of culture, political institutions, lobbyism, and economic structure.

Existing modes of governance treat consumers as citizens, as buyers, or as social groups (Trumbull, 2006). Citizens with their rights shape the 'protection model,' whereas buyers are linked to the 'information model.' Consumers as social and political group nourish the 'negotiation model.' The United States came close to the 'protection model,' Germany largely represents the 'information model,' and most Scandinavian countries have enshrined the 'negotiation model.'

Historically, governance of consumer policy has been shaped by regulatory tradition, legal systems, and other institutional factors. The protection model in the United States emerged because regulatory agencies, common law, and judicial activism paved the way for mobilizing consumers (Vogel, 2012). Vice versa, public interest groups achieved new rights that could be enforced through litigation.

In Germany, ordoliberal tradition of economic policy considered competition to be the key for consumer policy (Kelemen, 2011). Government created a unique endowment, the *Stiftung Warentest* (product testing). It is supposed to provide neutral testing and information that serves all consumers.

In contrast, Scandinavian countries like Sweden have invented institutions like the ombudsman that consumer groups and business groups can turn to. This model does not treat consumers as anonymous buyers but as an organized group that can deal with conflicts. However, the relation is not considered adversarial – as in the case of the United States. Instead, procedures are supposed to mitigate cooperative relations.

In the meantime, there are signs of convergence. Sweden has invented class action for consumer issues, adding some elements of the 'protection model.' Regulatory agencies in Germany will get a mandate of consumer protection, resembling patterns in the United States and the United Kingdom. In addition, the European Union is still planning to bolster

public interest litigation. Consumer rights seem to be a tool of creeping convergence in consumer policy.

Fighting for Rights: Public Interest Litigation and Consumer Policy

Litigation has shaped consumer policy for a long time. Institutional differences between Europe and the United States are still striking. Class action is widely unknown in Europe. Instead, government agencies or associations take the lead in public interest litigation. Contingency fees and tort damages are still much more a feature of US adversarial legalism (Strünck, 2008).

As for the United States, consumer policy got a boost by judicial activism in the wake of the civil rights movement (Epp, 1998). Ralph Nader figured prominently as an early consumer advocate, running 'Public Citizen' as a powerful public interest group (Eisner, 2000). He picked General Motors as his first prominent adversary: his paper "Unsafe at Any Speed" depicted the popular car Corvette as a most dangerous vehicle.

Public interest litigation helped to raise safety standards, to ban hazardous products, and to protect vulnerable groups against risks such as smoking. On the other hand, public interest litigation in the United States averted more comprehensive legislation (Kagan, 2007). There are signs that the European Union officials jump on the bandwagon of class action (Kelemen, 2011). However, national legal systems still restrict the scope of public interest litigation in Europe.

Currently, there is no clear picture in European class actions on the opt-in and opt-out arrangements. Sweden has famously switched to an opt-in system, whereas countries as diverse as Portugal, Spain, Denmark, Norway, the United Kingdom, and Germany have introduced opt-out clauses, at least in consumer protection law (Cafaggi and Micklitz, 2009). Traditionally, European countries pursue opt-in strategies due to their legal tradition of individually authorized lawyers.

The US tradition of opt-out not only raises questions of accountability but also increases the pressure to settle cases prematurely. Contingency fees show the same effect. Interestingly enough, opt-in clauses have the effect of producing lead plaintiffs who are committed and credible toward their members. Most public interest groups can handle opt-in cases better than private attorneys.

Another institutional device that affects the activities of public interest groups is the role of government agencies (Strünck, 2008). In Sweden, for instance, government agencies lead public interest litigation despite its corporatist tradition. In Austria, the federal consumer organization dominates public interest litigation. The role of government touches on the issue of funding of the public interest litigation, as well. Third-party financing is an underestimated factor whenever the scope of representation is at stake. Above all, 'loser pays' rule is predominant in Europe and for public interest groups, it means that cases they are likely to win will not pose any financial burden to them (Hodges et al., 2009).

The 'loser pays' rule also results in high predictability, compared to punitive damages in the United States. It renders legal expenses insurance a predictable business, too. Thus, conditions for external funding are good. This does not necessarily mean that public interest groups can rely on funding from third parties. Some European governments delegate litigation to public interest groups and provide sufficient funding, as well. In Austria, the Ministry of Consumer Protection regularly assigns litigation to the federal consumer organization to carry out (Micklitz and Stadler, 2006).

In other countries like France, membership fees are a vital source of funding of litigation. This, however, creates negative incentives to sue. The same is true for the German *Act Against Unfair Practices*. It allows for skimming of the profits if corporate misconduct can be proven. Consumer organizations are entitled to act on behalf of consumers before courts. Injunctions are almost a standard routine operation for consumer organizations and self-regulatory bodies of business in Germany. The business-based center for competition issues 7000 injunctions every year.

To consumer groups litigation is an ambiguous tool. Depending on funding, they face a time-consuming and very sophisticated legal process. Whether litigation brings in new opportunities to frame public discourses depends on subjects and claimants that are affected. Some types of public interest groups are small, understaffed organizations that run campaigns. They are just not experienced enough to undertake ambitious legal efforts. In addition, campaigning gets quicker media attention for less effort. Thus, it is a strategic choice for public interest groups whether to embark on legal venues at large scale.

Certainly, outcomes for diffuse and weak interests are different when private parties or lawyers act instead of public interest groups. In the European setting, effects are mostly restricted to parties involved in the litigation. It might lead to policy changes if the case is highly salient and expensive and the fear of future litigation may deter the market participants from certain behavior. However, even smaller damages to consumers might harm public interests. In this case, third-party funding and standing rights are crucial factors in whether public interest groups pick up litigation as a strategy.

Litigation aims at enforcing law. Additionally, law firms and consumer groups can call attention to loopholes in legislation, mobilizing political clout to change the law. Historically, litigation has brought about social change, especially in the United States. Attitudes have changed (smoking), and power relations have been affected (abortion). Social change is the main topic of activists who champion 'political consumerism.' This trend revives concepts of collective action in consumer policy.

Fighting for a Better World: New Patterns of Political Consumerism

Empowering consumers is a crucial goal of consumer policy. However, some consumer groups need not to be empowered; they would rather empower other people and the public. Political consumerism is on the rise, not only in the industrialized world. The range of ideas is broad. There are movements that are fundamentally critical on mass consumption and the business of branding (Galvagno, 2011). Others target companies and monitor their activities. They claim that companies are not just businesses but bear political responsibility. Boycotts

and *buy*cotts are launched to change corporate behavior. At the same time, activists want consumers to see the broad picture. In their perspective, consumption is not just to buy and use something; consumption affects life of people everywhere.

Some of those activities have resulted in policy change: governments enacted new standards, a whole range of labels have emerged, and some companies have revamped their codes of conduct (Stolle and Micheletti, 2013). Does this trend increase political participation; does it bring consumers to act as political group?

New forms of political consumerism are 'individualized forms of collective action' (Holzer, 2006). Consumers do not engage as group but act individually. This kind of market-based activity makes up for problems of free riding. If consumers are difficult to organize it is another option to nudge them individually. This is why political consumerism is often equated with voting. In globalized economies with trust in traditional politics waning, 'political shopping' might be an alternative tool of participation.

The figures are striking: people in Sweden and Switzerland seem to be extraordinarily sensitive toward consumption, whereas people in South European and East European countries seem to care least about political consumption. In Sweden and Switzerland, 50% and more people asked had used positive buying in the last 12 months, and more than 30% had boycotted products. In Greece, Portugal, and Hungary figures did not exceed 10% (Koos, 2012).

It is not just postmaterialist attitudes that drive political consumerism. Political and economic variables matter, too, which is why countries show different forms of activism. Several variables account for these differences: The more affluent a society, the more likely are people to use consumption to express political values. Highly concentrated markets in retail services seem to trigger backlashes, too. Additionally, more decentralized political institutions ease political consumption. However, nongovernmental organizations and social movements do not significantly increase the level of political consumption. Apparently, their campaigns still much more focus on other issues that do not translate into critical consumption.

Technology has fundamentally changed opportunity structures. Social media make companies vulnerable to attacks on their reputation. Codes of Corporate Social Responsibility (CSR) are an answer to new waves of political consumerism (Stolle and Micheletti, 2013). Consumer policy indirectly contributes too. The more labeling, the better the opportunity structures (Koos, 2011). Vice versa, labeling results from campaigning.

Social scientists fret over whether critical consumption really resembles political participation. Some stress that people act as 'consumer citizens,' using economic tools for political reasons. Others decry the volatile behavior of consumers. Whatever the motivation for political consumption, environmental risks and labor standards in the developing world are often targets. Thus, consumer policy in advanced economies might help to improve those conditions. Thus, critical consumers primarily care about employees in developing countries.

Political consumerism calls attention to side effects of global production. This also sheds light on risks that consumers in developing countries are facing. Small producers might be exploited to serve the needs of Western consumers. On the other hand, consumers in those countries are exposed to fundamental risks of products, too.

In most developing countries, consumer policy is still in the wake. However, it has to tackle most elementary health risks to consumers, mainly connected with food and pollution. Aside from pollution, it is food safety that is most critical to China's ruling communist party (Jiang and Zhu, 2013).

Consumer policy figures prominently in advanced economies whose markets are regulated. Consumer organizations, the media, and lawyers push governments to act on behalf of consumers. It is not like that in developing countries. Consumer policy in the less developed world is just in the making. This leaves ample room for governments to imitate policies or to invent new forms of governance. Quite generally, changes in society and markets challenge consumer policy.

Challenges to Consumer Policy

Traditionally, safety of products has been one of the core issues of consumer policy. The rise of the service sector has highlighted the issue of quality control. More recently, the financial crisis has shown that even the fiercest competition does not keep consumer from being ripped off. Financial services, old-age provision, and other services are hard to assess for individual consumers. At the same time, those investments are supposed to last for a long time. Often, vulnerable consumers fall victim to services with long-term impact. Thus, creating transparency and setting enforceable quality standards are the main tasks of consumer policy in the future.

Additionally, digital markets pose new challenges. Do users of social media need to be protected against data transfer? In a strict sense, people are 'users' not 'consumers,' of social media. Does traditional consumer policy apply? Transnational shopping through the internet provides lots of advantages and also problems to consumers. Can they enforce their rights if the provider runs its business in a remote place? Can consumers trust providers that run consumer advice online? Regulating digital markets will remain a huge challenge to consumer policy.

Another aspect is sustainable consumption. Here the debate on soft paternalism pops in. Is it OK to nudge consumers in order to ponder sustainable choices? What to do about waste and reuse of products? Do companies primarily bear responsibility or is it consumers, too, who have to act?

A contentious issue in international politics is free trade agreements. The pending Transatlantic Trade and Investment Partnership between the United States and the European Union is a test case. Consumer advocates stir up fears of genetically modified organism entering European markets against the will of consumers. Which level of regulation is up for negotiations?

As for governance, the role of government is likely to change. Consumer policy has mostly been protection by law. However, are governments capable of enforcing the law in ever more complex markets? Alternatively, will litigation and tort damages step in as the more effective precautionary principle?

In fragmented political systems like the United States or the European Union, third sector organizations increasingly

implement policies. Will consumer groups and campaigns take on more responsibility? Will companies engage more in cooperative relations with consumer groups to bolster CSR policies?

The situation in developing countries is different. Grassroots protests against contaminated food or polluted drinking water urge governments to act. Yet, civil society might still be too weak to craft and implement policies themselves. Given the imminent threat of food safety issues and other risks, civil society organizations might pick up consumer policy as a crucial issue. Handling basic risks for consumers is a litmus test for governments' legitimacy; this is even truer in countries where everyday life comes with lots of risks.

Consumer policy has often been scandal driven. Nondemocratic governments might just ride out things. In democratic societies, scandals often punctuate a long time policy equilibrium, opening up space for outsiders to lobby (Baumgartner and Jones, 2002). Future politics and research will show who is able to shape and deliver consumer policy, for which groups, for which goals, and for how long.

See also: Antitrust Policy: Lessons from the US; Behavioral Economics; Ethical Codes, Professional: Business Codes; Governance; International Trade: Commercial Policy and Trade Negotiations; Internet and Privacy; Issue Networks: Iron Triangles, Subgovernments, Policy Communities, Policy Networks; Law and Economics; Public Choice Economics; Public Policy; Regulation, Empirical Analysis of; Regulation, Political Economy of; Regulatory Agencies.

Bibliography

Akerlof, George A., 1970. The market for "lemons": quality uncertainty and the market mechanism. The Quarterly Journal of Economics 84 (3), 488–500.

Albanese, Paul, 2006. Inside economic man: behavioral economics and consumer behavior. In: Handbook of Contemporary Behavioral Economics : Foundations and Developments. Sharpe, Armonk, NY [u.a.], pp. 3–23.

Baumgartner, Frank R., Jones, Bryan D., 2002. Policy Dynamics. University of Chicago Press, Chicago.

Cafaggi, Fabrizio, Micklitz, Hans-W., 2009. New Frontiers of Consumer Protection: The Interplay between Private and Public Enforcement. Intersentia, Antwerp; and Portland, OR.

Eisner, Marc A., 2000. Regulatory Politics in Transition, second ed. Johns Hopkins University Press, Baltimore.

Epp, Charles R., 1998. The Rights Revolution: Lawyers, Activists, and Supreme Courts in Comparative Perspective. University of Chicago Press, Chicago.

Galvagno, Marco, 2011. The intellectual structure of the anti-consumption and consumer resistance field. European Journal of Marketing (11–12), 1688–1701.

Gronow, Jukka, Warde, Alan, 2001. Ordinary Consumption. Routledge, London.

Hantula, Donald A., 2012. Consumers are foragers, not rational actors: towards a behavioral ecology of consumer choice. In: Handbook of Developments in Consumer Behaviour. E. Elgar, Cheltenham [u.a.], pp. 549–577.

Hodges, Christopher, Vogenauer, Stefan, Tulibacka, Magdalena, 2009. Costs and funding of civil litigation: a comparative study. Legal Research Paper Series 2009, 1–118.

Holzer, Boris, 2006. Political consumerism between individual choice and collective action: social movements, role mobilization and signalling. International Journal of Consumer Studies 30 (5), 405–415.

Jiang, Qijun, Zhu, Ying, 2013. Confronting the crisis of food safety and revitalizing companies' social responsibility in the People's Republic of China. Asia Pacific Business Review 19 (4), 600–616.

Kagan, Robert A., 2007. Globalization and legal change: the "Americanization" of European law? Regulation & Governance 1 (2), 99–120.

Kelemen, R. Daniel, 2011. Eurolegalism: The Transformation of Law and Regulation in the European Union. Harvard University Press, Cambridge, MA.

Koos, Sebastian, 2012. What drives political consumption in Europe? A multi-level analysis on individual characteristics, opportunity structures and globalization. Acta Sociologica 55 (1), 37–57.

Koos, Sebastian, 2011. Varieties of environmental labelling, market structures, and sustainable consumption across Europe: a comparative analysis of organizational and market supply determinants of environmental-labelled goods. Journal of Consumer Policy: Consumer Issues in Law, Economics and Behavioural Sciences 34 (1), 127–151.

Micklitz, H.-W., Stadler, A., 2006. The Development of collective legal actions in Europe, especially in German civil procedure. European Business Law Review [online], 17 (5), 1473–1503.

Olson, Mancur, 2003. The Logic of Collective Action: Public Goods and the Theory of Groups, 21. Printing. Harvard University Press, Cambridge, MA.

Stolle, Dietlind, Micheletti, Michele, 2013. Political Consumerism: Responsible Action in the Global Age. Cambridge University Press, New York.

Stone, Deborah, 2012. Policy Paradox: The Art of Political Decision Making. W.W. Norton & Co., New York, 3. Aufl.

Strünck, Christoph, 2008. Claiming consumers' rights.: patterns and limits of adversarial legalism in European consumer protection. German Policy Studies 2008, 167–192.

Sunstein, Cass R., 2008. Laws of Fear: Beyond the Precautionary Principle, 4. Printing. Cambridge University Press, Cambridge.

Trumbull, Gunnar, 2006. Consumer Capitalism: Politics, Product Markets, and Firm Strategy in France and Germany. Cornell University Press, Ithaca, NY.

Vogel, David, 1997. Trading Up: Consumer and Environmental Regulation in a Global Economy. Harvard University Press, Cambridge, MA.

Vogel, David, 2012. The Politics of Precaution: Regulating Health, Safety, and Environmental Risks in Europe and the United States. University Press, Princeton, NJ.

Consumer Psychology

Jacob Jacoby, New York University, New York, NY, USA
Maureen Morrin, Temple University, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The history, development, and future trajectory of the field of consumer behavior are discussed. The field's theoretical grounding in the basic disciplines of economics and psychology is reviewed, and a brief history of the field's major conceptual foci over time is provided. Also addressed are the areas of research generating increased attention among the current generation of consumer scholars.

Consumer behavior refers to the acquisition, consumption, and disposal of time and offerings (goods, services, and ideas) by decision-making units. Consumer behavior is pervasive and universal involving, as it does, choices made during the waking moments of virtually all human beings in all societies and cultures. In addition to overt observable behaviors (e.g., purchasing, word-of-mouth communications regarding products and services), consumer behavior also includes the covert psychological states and processes (e.g., motives, learning, attitudes, decision making, past experience) that underlie and influence the overt behaviors. Last, consumer behavior also includes decisions regarding the cessation or lessening of certain behaviors (e.g., smoking; unsafe sex; eating fatty foods).

Though also referring to the covert psychological processes, when used to describe a scholarly endeavor, consumer psychology refers to using distinctively psychological concepts and methods to study consumer behavior. These days, scholars employing clearly psychological concepts and methods to study consumer behavior are more likely to do so from venues other than Departments of Psychology.

The Complexity of Consumer Behavior

Consumer behavior includes often-overlooked layers of complexity such as the following. Offerings may be acquired in ways other than via purchasing, including via renting, borrowing, trading, finding, stealing, as a gift, etc. Similarly, consumption – which occurs over time periods of varying duration (from the almost instantaneous munching and swallowing of a potato chip, through the wearing of a shirt for several years, to buying and living in a home for decades) – generally involves distinct facets, including using, maintaining (e.g., washing the shirt), storing, and retrieving, all or some of which may be repeated over time. Once consumption ends, one may be faced with disposing of the item (or its container). One can decide to throw it away, convert it to serve some other purpose, trade or exchange it for something else, sell it, give it away, etc. Consumer behavior also reflects cyclical and contingent qualities. Having purchased, consumed, and disposed of a tube of toothpaste, we acquire a replacement and repeat the cycle. Car insurance and gasoline are generally acquired contingent upon having acquired (or gained access to) a car.

Time is finite; how we choose to spend it has tremendous implications for virtually all other aspects of our lives. In some respects, one can actually buy and trade time, e.g., paying someone to mow our lawn and free our time for something else. On the other hand, not being limited to the products and services provided by commercial, for-profit entities, offerings come in many forms. They include political candidates, television programs, newspaper articles, leisure time activities, different sections and articles within the same newspaper, etc.

Engaging in consumer behavior involves a number of roles, including behaving as the decision maker, purchaser, and/or user. Often, these roles are confined to a single individual (as when one chooses, buys, and then consumes a soft drink). Alternatively, they may involve two or more individuals (e.g., wife decides, husband purchases, child uses).

Last, consumer behaviors occur not in a vacuum, but in the context of everyday life where they tend to be influenced by numerous external factors, including cultural, social, political, environmental (e.g., not shopping during a snowstorm), and physical (e.g., not wanting to stoop for a product on a low shelf) considerations.

Why Study Consumer Behavior?

As understanding consumer behavior is often the key to business success, many who study consumer behavior come from disciplines (e.g., economics, marketing, advertising, retailing) that support business activities. But why do psychologists find it worthy of study? Discussed in greater detail elsewhere (Jacoby, 1975: pp. 981–985), a few of the reasons are as follows.

First, consumer behavior is a domain providing innumerable opportunities for important, socially meaningful research. Learning how to influence the underprivileged in this country and elsewhere to adopt and increase consumption of products and services good for them and to cease consumer behaviors that are less appropriate or even harmful is one example. Studying why there is so high a rate of noncompliance with physician prescription dosage instructions is another. Second, consumer theory and research have the potential to contribute to the development and extension of psychology itself. Examples of such contributions abound and include the conceptual (e.g., richer, more dynamic ways of conceptualizing

 http://dx.doi.org/10.1016/B978-0-08-097086-8.22004-7

Stimulus-Organism-Response (S-O-R); cf Jacoby, 2002) as well as the empirical (e.g., dynamic, online) methods that enable psychologists to better test dynamic theories regarding how information self-selected by an individual affects that person's higher order cognitive processing (e.g., attitudes, impressions; cf Jacoby et al., 2002).

Third, because of the universal and pervasive nature of consumer behavior, this context provides excellent opportunities for examining the validity and limits of psychological theory. Fourth, consumer research provides numerous opportunities for working with dependent variables that, from the perspective of test subjects, are meaningful and consequential. Fifth, in attempting to understand consumer behavior, one is forced to consider a broader array of factors than is typical in much psychological research. This occasionally leads to seeing relationships that might not have been noted had one not been forced to approach the subject using a broad perspective.

The Evolution of Scholarly Research on Consumer Behavior

Scholarly work on consumer behavior might be said to have begun with economists more than a century ago (e.g., Adam Smith, Jeremy Bentham, Alfred Marshall). These economists held that a person's desire for goods and services generally exceeded the person's ability to pay. Therefore, buying decisions were theorized to result from a rational process during which a value was assigned to each desired offering based upon the person's assessment of that offering's ability to satisfy his or her needs and desires. This want-satisfying ability was termed 'utility' and it was held that consumers would seek to maximize utility. According to these economists, the price of the offering was the key element in the maximizing utility equation. Raising an offering's price was predicted to lower its sales; lowering its price would increase sales. Though seemingly commonsensical, this proposition suggests that, given changes in price, consumers respond like billiard balls, buying when prices are lowered and refraining when prices rise. Nothing else matters.

In terms of the S-O-R model familiar to psychologists, it can be seen that the Economic Man conception is severely limited suggesting, as it does, a relatively empty organism whose behavior is determined by economic stimuli and little else. According to the classic S-O-R formulation, external stimuli (e.g., price changes) first need to be perceived, understood, evaluated, and integrated with other psychological phenomena (e.g., past experience, current needs, attitudes, expectations, etc.) before having impact, if any, on the individual's behavior. Thus, while one can see various reasons why a price decrease would lead to increased purchase (as predicted by economic theory), one can also see various reasons why it would have either no effect (e.g., the price change does not represent a 'just noticeable difference'; the consumer would purchase the same amount of the items regardless), or the reverse effect (e.g., sales decrease because the consumer interprets the price decrease as reflecting a decrease in quality, or because the consumer views the item as no longer having the desired snob appeal; etc.). To best understand and predict consumer behavior, we would have to know something about the consumer's motives, a priori expectations, how the incoming information is perceived and interpreted, and so forth – namely, all the factors traditionally studied by psychologists and then some. Although many key assumptions underlying the 'Economic Man' model have since been shown to be untenable, via a reincarnation termed Rational Choice Theory, this model remains popular among some in law, economics, and sociology (e.g., Posner, 1998).

Responding to limitations in the Economic Man model, both theorists and practitioners began adding variables to the stimulus and especially organism portions of the S-O-R model. The stream of research referred to as behavioral decision theory might be seen as efforts investigating the stimulus portion of the S-O-R model, especially in terms of how stimulus framing results in decisions deviating from normative outcomes. Prospect theory (Kahneman and Tversky, 1979) clearly showed how decision makers felt losses more than equivalent sized gains. During the 1930s through the 1960s, a number of European émigrés trained in Freudian psychoanalysis, collectively known as motivational researchers (e.g., Dichter, 1960), became influential among the business community. Other psychological constructs (including attitude and personality) also were introduced during this period. By the mid-1960s, the emerging models (developed not by psychologists, but by professors of marketing) reflected primarily psychological and/or sociological factors (e.g., Andreasen, 1965; Engel et al., 1968; Howard and Sheth, 1969). The work by Engel and his colleagues, now in its tenth printing, was the most influential introductory text for perhaps 15 years and remains popular. The 1960s also represented the time when an increasing number of psychologists with a scholarly bent began studying consumer behavior. Reflecting evolution, a number of today's leading texts (Hoyer et al., 2012; Peter and Olson, 2012; Solomon, 2012) are authored or coauthored by PhD psychologists.

Two developments did much to shape the emerging field *c.*1970. First came formation of the Association for Consumer Research (ACR, www.acrwebsite.org), today considered to be the principal scholarly organization in the field. Starting with a few hundred members in 1969–70, ACR's membership consists of researchers from a wide variety of disciplines and now numbers nearly 2000, including several hundred psychologists. Though its members focus on different issues (e.g., marketing researchers tend to be more interested in understanding purchase decisions, communication researchers more interested in interpersonal or mass media communication, home economists more interested in problems that may surface during usage, etc.), their shared frame of reference is that, influenced by elements in a complex environment, consumers are information-processing, problem-solving, decision-making, choice-making beings. Its annual series, *Advances in Consumer Research*, best illustrates the diversity of the field.

The second major development was the introduction of the *Journal of Consumer Research* (ejcr.org), the first journal devoted specifically to consumer research. Prior to the early 1970s, scholarly work on consumer behavior appeared sporadically in journals spanning many disciplines. Work having a behavioral science (including psychological) perspective tended to appear in the *Journal of Applied Psychology* and, to a lesser extent, in the *Journal of Marketing* and *Journal of Marketing Research* (www.marketingpower.com). While these journals continue to

publish some of the better consumer research, their principal foci lie elsewhere. The first issue of the *Journal of Consumer Research*, a collaborative venture by 10 national societies appeared in 1974 and quickly became the field's flagship journal. The cosponsoring societies of the inaugural issue were: American Association for Public Opinion Research, American Council on Consumer Interests, American Economic Association, American Home Economics Association, American Marketing Association, The Division of Consumer Psychology (Division 23) of the American Psychological Association since renamed the Society for Consumer Psychology, American Sociological Association, American Statistical Association, ACR, and The Institute of Management Sciences. Societies that have since become cosponsors include: American Anthropological Association, American Association of Family and Consumer Sciences, Institute for Operations Research and the Management Sciences, International Communication Association and the Society for Personality and Social Psychology (Division 8) of the American Psychological Association. Using the first 15 years of *Journal of Consumer Research* publications, Hoffman and Holbrook (1993) provide interesting perspective on the intellectual structure of the field.

The Introduction of Psychology to Studying Consumer Behavior

Interest by psychologists in consumer behavior and the factors that influence such behavior has a 110-year history. In 1900, H. Gale published work on the position of an ad on the printed page in 'The University of Minnesota Studies on Psychology'. In his 1903 book, *The Theory of Advertising*, Walter Dill Scott, whose dissertation on attention and persuasion was with Wundt, introduced the concepts of needs and motives to advertising. Scott's 1908 book (*The Psychology of Advertising in Theory and Practice*) included the word psychology it its title. However, except for occasional articles in the *Journal of Applied Psychology*, psychologists devoted little scholarly attention to consumer behavior during the first half of the twentieth century. When studied at all, consumer behavior was often viewed as a convenient place to test more fundamental psychological propositions. For example, when the United States was experiencing shortages in traditional cuts of beef during World War II, Kurt Lewin's classic studies on public vs private commitment assisted both government and industry in reorienting food consumption away from foods in short supply and making internal meats (e.g., brain, liver, kidney) more palatable to the American consumer. With the exception of Katona (1951, 1953), essentially a voice in the wilderness, few psychologists studied consumer behavior for its intrinsic interest. Situated in industry, most members of American Psychological Association's (APA) Division of Consumer Psychology had an applied orientation and were interested in research that assisted business.

This emphasis began changing in the mid-1960s. The first chapter on Consumer Psychology appeared in *Annual Review of Psychology* (Twedt, 1965). The second appeared (Perloff, 1968), 3 years later, as did a chapter in the *International Encyclopedia of the Social Sciences* (Krugman, 1968). In addition to the traditional focus on the consumer *qua* purchaser (i.e., as the object of business interests), these contributors devoted greater attention to the consumer *qua* consumer. This emphasis became entrenched during the decade that followed (Jacoby, 1975, 1976), gaining even greater strength in the intervening quarter century (Jacoby et al., 1998). By the early 1980s, the dominant focus of APA's Division of Consumer Psychology (which had been renamed The Society for Consumer Psychology) had shifted to studying the consumer *qua* consumer. In 1992, it began publishing the *Journal of Consumer Psychology* (www.journals.elsevier.com/journal-of-consumer-psychology) through Lawrence Erlbaum Associates, a prestigious publisher of works in psychology.

Concurrent with increasing numbers of psychologists finding that consumer behavior merited their scholarly attention, scholars in other disciplines eagerly embraced psychological theory and methods for their own consumer behavior research, with many advising their doctoral students to minor in psychology (particularly social, cognitive, or measurement psychology). This cross-fertilization was enhanced as increasing numbers of newly minted PhD psychologists assumed faculty positions in Departments of Marketing and other disciplines where they studied consumer behavior. Many later became highly regarded scholars and leaders in the field. Adding to this trend, prominent psychologists began participating in Society for Consumer Psychology and ACR conferences, occasionally publishing articles in consumer behavior journals; examples include William McGuire (1976), Robert Zajonc (Zajonc and Markus, 1982), and Daniel Kahneman (Ratner et al., 1999).

Current Content and Foci of Consumer Psychological Research

A psychologist reading today's introductory texts would feel right at home. Even those not (co)authored by psychologists devote chapters to familiar subjects, including motives, attention, perception, categorization, comprehending, learning, memory, attitudes, persuasion, and decision making. Large portions of other chapters are devoted to emotions, personality, values, cognitive dissonance, the self, affect, symbolism, and nonconscious processes. Much of the remaining content tends to be sociocultural in nature, with chapters typically covering cultural and cross-cultural influences, subcultures, social class, reference and other group (e.g., age, gender, race, religion) influences, and family and household considerations.

Though consumer researchers study a wide variety of psychological constructs, phenomena, and topics, some have dominated. Attitude theory has enjoyed sustained popularity over several decades. During the early 1970s, hardly an issue of the *Journal of Consumer Research*, *Journal of Marketing*, or *Journal of Marketing Research* appeared without at least one (and often more than one) article making heavy use of Fishbein's seminal work (Fishbein and Ajzen, 1975). In the late 1980s, attention shifted to other attitude theories (Chaiken, 1980; Petty and Cacioppo, 1986), and since then the dominance of attitude theory appears to have waned a bit.

In the early 1970s, research on various aspects of cognitive processes (including judgment and decision making) surfaced, quickly becoming a dominant intellectual focus in the field. In recent years, the focus has centered on understanding the effects of rapid and automatic or System I processing vs more

deliberate and careful, or System II processing on decision-making behavior. For a review see Evans (2008). This framework posits that the default approach to cognitive processing is one based on minimal cognitive effort in order to conserve limited cognitive resources. As such, System 1 processing tends to be nonconscious, automatic, and more affectively charged. When sufficiently motivated and able, a consumer will instead engage in more effortful cognitive processing efforts, using reason and conscious awareness of attitudes to make decisions. Such System 2 processing is conceived of as being slower, more reason based, and less emotion driven (Kahneman, 2003). Importantly, decision outcomes can vary as a function of which type of processing is engaged. Similar dual processing models were earlier evident in theories of attitude formation and change (Chaiken, 1980; Petty and Cacioppo, 1986). The dual processing approach is now much more broadly applied to various aspects of decision making however.

In recent years, a greater focus has shifted to the effect of nonconscious and noncognitive processes, moving the discipline further away from the Economic Man model. For example, the impact that bodily movements and sensory experiences have on consumer responses (Krishna, 2012) has attracted significant research attention, reflecting the influence of theories of embodied cognition and other 'feelings as information' approaches (e.g., Hong and Sun, 2012). Studies have shown, for example, that motor activity, muscle contractions, and other physical sensations can significantly influence consumer inferences and response patterns (Labroo and Nielsen, 2010). Firming one's muscles can enhance a consumer's willpower; increase their ability to withstand pain and to avoid tempting foods (Hung and Labroo, 2011). Feeling physically cold increases the liking of romance movies due to the activation of the need for social warmth (Hong and Sun, 2012). Researchers have begun to investigate the effects of each of the five senses on consumer behavior, especially those of olfaction, touch, and taste that had previously been largely ignored. It has been shown, for example, that the quality of the cup from which water is tasted can influence the perceived taste of the water (Krishna and Morrin, 2008), and that consumers differ in the degree to which they enjoy and need to touch products before purchasing them (Peck and Childers, 2003). Recognizing the growing interest in these areas, a special issue of the *Journal of Consumer Psychology* was devoted to sensory perception, embodiment, and grounded cognition (forthcoming 2014 issue edited by Aradhna Krishna and Norbert Schwarz).

Another thrust has been increasing attention devoted to how consumer research may be used to develop, evaluate, and influence public policy. While research of this sort appears in the journals noted above, much is now published in the *Journal of Public Policy and Marketing* (www.marketingpower.com), a journal established for this purpose *c*.1980.

The field has experienced shifts in research methodology. Early nonpsychological research placed great reliance on surveys. As psychological factors assumed greater prominence in the field's models, experimentation became the dominant methodological approach. Since the 1990s, there has been a gradual acceptance of qualitative research methods of various sorts (ethnographic, interpretivistic, introspective, etc.), but these approaches account for a relatively small proportion of published works. We estimate that less than 15–20% of the currently published work utilizes such methods. Recently, prominent consumer researchers have advocated for a renewed appreciation of research that reports novel empirical findings, in addition to the traditional focus on incremental, theory-based approaches to science (Alba, 2012; Park, 2012). Park (2012) refers to this distinction as one of 'cute' vs 'beautiful' research, arguing that the field of consumer research should embrace both. Such calls may be partly in response to the ever increasing number of studies typically required within a single article for publication at a top tier journal as well as the increased emphasis on the use of sophisticated statistical techniques to demonstrate process mechanisms such as mediation analysis via bootstrapping approaches and more complex variants such as moderated mediation (e.g., Preacher et al., 2007).

Emerging Foci in Consumer Research

In addition to the emphases mentioned above, others that can be anticipated include the following. First, though we do not expect many to forsake reliance on static verbal report methods, it is expected that consumer researchers will gradually display greater use of qualitative and quantitative methods that capture the dynamic nature of the behavior and underlying processes. Such approaches are aided by the availability of web-based surveys, consumer panels, mobile devices such as smartphones, computer software, and other technologies that can capture geographic location, vocal responses, track eye gaze paths, and monitor brain activation patterns in response to marketing stimuli – also known as neuromarketing (Plassman et al., 2012; Yoon et al., 2012b). The integration of social media into study designs as an effective way to track and possibly even alter consumer behavior will likely be seen in coming years.

Second, while some work has always focused on socially important issues and improving social welfare, we expect to see an increase in such research. Related to this trend is an increasing focus on consumer research that has the capacity to change consumers' lives for the better, often referred to as Transformative Consumer Research. This stream of research includes issues related to enhancing consumer well-being (Molinger et al., 2012), corporate social responsibility (Torelli et al., 2012), and environmental sustainability. This stream is also devoted to exploring issues associated with 'at risk' consumer populations such as the poor, the elderly, the disabled, the illiterate, and other relatively underresearched demographic groups. Hill and Stamey (1990) looked at the ways the homeless acquire possessions and how these are used to aid their survival, for example.

Researchers have turned their attention to globally pressing issues such as the obesity epidemic, for example, by examining goal theory, self-control, and approach-avoidance response (e.g., Yoon et al., 2012a). The way that consumers resolve goal conflicts, such as the desire to enjoy chocolate cake in the immediate future vs the desire to maintain a healthy body weight in the long term, has captured the interest of several consumer researchers. Construal level theory has been used to help explain inconsistencies in preferences as a function of how abstractly or concretely events are construed. Events

thought of as psychologically distant and more abstract, such as those occurring in the distant future, have been found to be more driven by long-term goals, whereas those thought of as psychologically near and concrete, such as those occurring in the near future, will be driven by short-term goals (Eyal et al., 2009). Ego-depletion theory (Baumeister, 2002), which suggests that the consumer's ability to exert self-control is limited much like a muscle that is depleted with use, has dominated much of the effort to understand issues related to self-regulatory behavior as well. Theories regarding how the exercise of self-control subsequently licenses the lack of self-regulation are also relevant in this domain (Khan and Dhar, 2006). These theories have also been used to examine issues such as financial investing, use of credit, and gambling.

Third, we expect consumer researchers to be drawn into applying their expertise in a number of other realms. Such a trend has been emerging in the legal arena, particularly in regard to deceptive advertising, trademark confusion, and product liability, where psychological theory and research are playing an increasing role (Jacoby, 2001). Such extensions can be expected to occur in regard to social media, consumer privacy, and in other unanticipated arenas as well.

Last, as noted, the ACR is widely acknowledged to be the central organization for consumer researchers of any stripe. However, unlike the Society for Consumer Psychology (which subscribes to the American Psychological Association's Code of Ethics), the governance of ACR has resisted most efforts to develop and implement any such code. For the good of that organization, one hopes this will change, preferably sooner rather than later. See Alberts and Shine, 1994, "Because questionable research practices are generally not appropriate targets for governmental or legal investigations, the scientific community itself must take responsibility for determining which practices are serious enough to warrant institutional or professional responses and what forms these responses should take…. The responsibility for scientific conduct falls on all parts of the research community, including…the leaders of scientific societies" Alberts and Shine (1994: p. 1660). At the time they wrote this article, Alberts and Shine were, respectively, President of the National Academy of Sciences and President of the National Institutes of Medicine. The need for a code of conduct seems especially acute in light of recent article retractions, such as that by the *Journal of Consumer Research* due to the use of fraudulent data (Retraction, *JCR* 2012: p. 444).

See also: Advertising and Advertisements; Advertising: Effects; Advertising: General; Conspicuous consumption and distinction, history of; Consumer Policy; Consumption, Geography of; Consumption, History of; Consumption, Sociology of; Ethical Consumers; Networks and Cultural Consumption; Time Use and Resource Consumption.

Bibliography

Alba, J., 2012. In defense of bumbling. Journal of Consumer Research 38 (6), 981–987.

Alberts, B., Shine, K., 1994. Scientists and the integrity of research. Science 266 (9), 1660.

Andreasen, A.R., 1965. Attitudes and consumer behavior: a decision model. In: Preston (Ed.), New Research in Marketing. Institute of Business and Economic Research, University of California, Berkeley, CA.

Baumeister, R.F., 2002. Yielding to temptation: self-control failure, impulsive purchasing, and consumer behavior. Journal of Consumer Research 28 (4), 670–676.

Blackwell, R.D., Miniard, P.W., Engel, J.F., 2005. Consumer Behavior, tenth ed. Dryden Press, New York.

Chaiken, S., 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. Journal of Personality and Social Psychology 39, 752–766.

Cote, J.A., Leong, Siew M., Cote, J., 1991. Assessing the influence of Journal of Consumer Research: a citation analysis. Journal of Consumer Research 18 (3), 402–410.

Dichter, E., 1960. The Strategy of Desire. Doubleday & Co., New York.

Elke, L.H., Smeesters, D., Trampe, D., June 2012. Retraction of 'the effects of messiness on preferences for simplicity'. Journal of Consumer Research 39 (1 and 2), 444.

Engel, J.F., Kollat, D.T., Blackwell, R.D., 1968. Consumer Behavior. Holt, Rinehart, Winston, New York.

Evans, J., 2008. Dual-processing accounts of reasoning, judgment and social cognition. Annual Review of Psychology 59, 255–278.

Eyal, T., Liberman, N., Trope, Y., 2009. Psychological distance and consumer behavior: a construal level theory perspective. In: Wänke, M. (Ed.), Social Psychology of Consumer Behavior (Frontiers of Social Psychology). Psychology Press, New York.

Fishbein, M., Ajzen, I., 1975. Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research. Addison-Wesley, Reading MA.

Hill, R.P., Stamey, M., 1990. The homeless in America: an examination of possessions and consumption behaviors. Journal of Consumer Research 17 (3), 303–321.

Hoffman, D.L., Holbrook, M.B., 1993. The intellectual structure of consumer research. Journal of Consumer Research 19, 505–517.

Hong, J., Sun, Y., 2012. Warm it up with love: the effect of physical coldness on liking of romance movies. Journal of Consumer Research 39 (2), 293–306.

Howard, J.A., Sheth, J.N., 1969. The Theory of Buyer Behavior. Wiley, New York.

Hoyer, W.D., MacInnis, D.J., Pieters, R., 2012. Consumer Behavior, sixth ed. Houghton-Mifflin, Boston MA.

Hung, I.W., Labroo, A.A., 2011. From firm muscles to firm willpower: understanding the role of embodied cognition in self-regulation. Journal of Consumer Research 37 (6), 1046–1064.

Jacoby, J., 1975. Consumer psychology as a psychological sphere of action. American Psychologist 30 (10), 977–987.

Jacoby, J., 1976. Consumer psychology: an octennium. Annual Review of Psychology 27, 331–358.

Jacoby, J., 2001. The psychological foundations of trademark law: secondary meaning, acquired distinctiveness, genericism, fame, confusion and dilution. The Trademark Reporter 91 (5), 1013–1071.

Jacoby, J., 2002. Stimulus-organism-response reconsidered: an evolutionary step in modeling (consumer) behavior. Journal of Consumer Psychology 12 (1), 51–57.

Jacoby, J., Johar, G.V., Morrin, M., 1998. Consumer psychology: a quadrennium. Annual Review of Psychology 49, 319–344.

Jacoby, J., Morrin, M., Jaccard, J., Gurhan, Z., Kuss, A., Maheswaran, D., 2002. Mapping attitude formation as a function of information input: studying on-line processing models of attitude formation. Journal of Consumer Psychology 12 (1), 21–34.

Kahneman, D., 2003. A perspective on judgment and choice. American Psychologist 58, 697–720.

Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47 (2), 263–292.

Katona, G., 1951. Psychological Analysis of Economic Behavior. McGraw-Hill, New York.

Katona, G., 1953. Rational behavior and economic behavior. Psychological Review 60, 307–318.

Katona, G., 1960. The Powerful Consumer. McGraw-Hill, New York.

Khan, U., Dhar, R., May 2006. Licensing effect in consumer choice. Journal of Marketing Research 43, 259–266.

Krishna, A., 2012. An integrative review of sensory marketing: engaging the senses to affect perception, judgment and behavior. Journal of Consumer Psychology 22 (3), 332–351.

Krishna, A., Morrin, M., 2008. Does touch affect taste? The perceptual transfer of product container haptic cues. Journal of Consumer Research 34 (6), 807–818.

Krugman, H.E., 1968. 'Consumer Behavior', International Encyclopedia of the Social Sciences. Macmillan and the Free Press, New York.

Labroo, A.A., Nielsen, J.H., 2010. Half the thrill is in the chase: twisted inferences from embodied cognitions and brand evaluation. Journal of Consumer Research 37, 143–158.

McGuire, W.J., 1976. Some internal psychological factors influencing consumer choice. Journal of Consumer Research 2 (4), 320–339.
Molinger, C., Aaker, J., Kamvar, S., 2012. How happiness affects choice. Journal of Consumer Research 39 (2), 429–443.
Park, C.W., 2012. Two types of attractive research: cute research and beautiful research. Journal of Consumer Psychology 22, 299–302.
Peck, J., Childers, T.L., 2003. Individual differences in haptic information processing: on the development, validation, and use of the 'Need for Touch' scale. Journal of Consumer Research 30 (3), 430–442.
Perloff, R., 1968. Consumer analysis. Annual Review of Psychology 19, 437–466.
Peter, J.P., Olson, J.C., 2012. Consumer Behavior and Marketing Strategy, ninth ed. Richard D. Irwin, Chicago.
Petty, R.E., Cacioppo, J.T., 1981. Attitudes and Persuasion: Classic and Contemporary Approaches. Wm. C. Brown, Dubuque, Iowa.
Petty, R.E., Cacioppo, J.T., 1986. Communication and Persuasion: Central and Peripheral Routes to Attitude Change. Springer/Verlag, New York.
Plassman, H., Ramsoy, T.Z., Milosavljevic, M., 2012. Branding the brain: a critical review and outlook. Journal of Consumer Psychology 22 (1), 18–36.
Posner, R.A., 1998. Rational choice, behavioral economics, and the law. Stanford Law Review 50, 1551–1575.
Preacher, K.J., Rucker, D., Hayes, A.F., 2007. Addressing moderated mediation hypotheses: theory, methods and prescriptions. Multivariate Behavioral Research 42 (1), 185–227.
Ratner, R.K., Kahn, B.E., Kahneman, D., 1999. Choosing less preferred experiences for the sake of variety. Journal of Consumer Research 26 (1), 1–15.
Robertson, T.S., Kassarjian, H.H. (Eds.), 1991. Handbook of Consumer Behavior. Prentice Hall, Englewood Cliffs, NJ.
Scott, W.D., 1903. The Theory of Advertising. Small, Maynard & Co., Boston.
Scott, W.D., 1908. The Psychology of Advertising in Theory and Practice. Small, Maynard & Co., Boston.
Solomon, M.R., 2012. Consumer Behavior, tenth ed. Prentice Hall, Upper Saddle River, NJ.
Torelli, C.J., Monga, A.B., Kaikati, A.M., 2012. Doing poorly by doing good: corporate social responsibility and brand concepts. Journal of Consumer Research 38 (5), 948–963.
Twedt, D.W., 1965. Consumer psychology. Annual Review of Psychology 16, 265–294.
Yoon, Y., Sarial-Abi, G., Gurhan-Canli, Z., 2012a. Effect of regulatory focus on selective information processing. Journal of Consumer Research 39 (1), 93–110.
Yoon, C., Gonzalez, R., Bechara, A., Berns, G.S., Dagher, A.A., Dube, L., Huettel, S.A., Kable, J.W., Liberzon, I., Plassman, H., Smidts, A., Spence, C., 2012b. Decision neuroscience and consumer decision making. Marketing Letters 23 (2), 473–485.
Zajonc, R.B., Markus, H., 1982. Affective and cognitive factors in Preferences. Journal of Consumer Research 9, 123–131.

Consumption, Geography of

Charles Mather, Memorial University, St John's, NL, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Scholarship on the geography of consumption is strongly focused on the consequences of consumption. These consequences can be mapped through three spatial registers: the body, consuming connections, and the globalization of consumption. This article reviews academic scholarship on the consequences of consumption through these three spatial registers and draws connections between them.

Introduction

It is almost 20 years since the publication of Nicky Gregson's (1995) paper *And now it's all consumption*? Gregson was writing about a growing interest within geography and the broader social sciences on consumption. Consumption, she argued, had become 'a major concern' for social scientists and a subject worthy of study in its own right. If Gregson were to write a follow up article now, she might be tempted to use the title *It's still consumption!*, for research and writing on consumption within geography and the social sciences continues unabated. The interest in consumption geographies justifies the publication of many overview papers of the field and a growing number of book-length contributions on the issue of consumption. There is also no limit to the public interest in reading about the 'dark side' of consumption with new popular books like *Excess*, *Crass Consumption*, and *Wasted World* providing readers with new stories about ills of over-consumption. It seems there is no limit to the academic and popular appetite for writing on consumption.

In her review of consumption in the mid-1990s, Gregson (1995) examined two central themes in the field: the mall as a key site for consumption and the promotion of consumer goods through advertising. In contemporary writing on consumption, the spaces and sites of consumption remain important areas of debate and analysis. Similarly, there has been ongoing work on the role of advertising. Yet, the key concern for social scientists writing about the geographies of consumption has shifted considerably since the mid-1990s. And this shift is perhaps best captured by the title of Daniel Miller's (2012) recent book *Consumption and its Consequences.* More than any other issue, it is the *effect* of our consumption practices that has become the central question in the debates on the geographies of consumption. Writing on consumption has focused on what our consumption practices do to our bodies, what effect they have on the natural environment, and how they are implicated in global inequality. And we are being urged to change our consumption practices in ways that improves personal health, protects the environment, and addresses the challenge of global uneven development and poverty.

The problem, of course, is that it is very difficult to assess the consequences of individual acts of consumption. Even when it comes to the body, it is not altogether clear what effect the consumption of food and other materials has on the body. Take, for example, Becky Mansfield's (2011) extraordinary analysis of farmed and wild fish, which are simultaneously nutritious food and harmful poison. When it comes to larger scale questions like the relationship between consumption habits and climate change, the links are even more opaque. Paradoxically, perhaps, the difficulty in tracing the connections between consumption practices and their effects has not dampened popular and academic interest in the consequences of consumption.

In this article of research on consumption geographies, I explore the various ways in which geographers and other social scientists are writing about consumption and its consequences. I do this by examining the effects of consumption in three spatial registers: the consuming body, consuming connections, and the global spread of consumption. While the issue of consuming connections – in the form of ethical consumption – is a frequent theme in reviews of the field of consumption geographies, the themes of the consuming body and the globalization of consumption are less often considered in overviews of the field. In this way, this article on the geography of consumption presents some novel material and potentially new and interesting directions for research. The article begins with the globalization of consumption, the consequences of which are receiving increasing attention.

Globalizing Consumption

Research on consumption focuses almost exclusively on North America and the United Kingdom (Trentmann, 2009). Yet, the situation is changing rapidly in the context of the rapid spread of consumerism beyond Anglo-America. The globalization of consumption is linked to the rise of a new 'global middle class,' a class defined not by its relation to production, but by how it consumes. This is a class of consumers that has enough disposable income to afford cars, homes, televisions, and other commodities normally associated with a middle-class lifestyle in the United States and parts of Europe. Perhaps, not surprisingly, the consequences of the globalization of consumption are being interpreted in dramatically different ways.

For economic organizations concerned with global economic growth, like the Organisation for Economic Co-operation and Development (OECD), this new middle class has the potential to reinvigorate the global economy in the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.72110-6

wake of the global economic meltdown after 2008. Significantly, the new middle class has the potential to fill the "void in global demand brought on by the retrenchment of the American consumer" (Kharas, 2010: 8). The new driver of the global economy, it seems, is moving east: "the underlying structural forces for global growth may be able to reassert themselves and usher in a new era of rapid progress, this time based on an Asian middle class" (Kharas, 2010: 39). The OECD is not the only organization hopeful that the new global middle class will resuscitate global economic growth. In January 2013, the *Economist* reported on several new studies emphasizing the significance of the new middle class for global growth. One of these reports has estimated the enormous size of this new market – 'the $10 trillion prize' – and the challenges of marketing products for 'nontraditional' consumers. In China, for instance, marketers are encouraged to pay heed to the importance of aspiration in families that often have only one child. Ronald McDonald in China is not a clown, but is instead 'Professor McDonald' who offers courses in writing and mathematics to consuming children.

For environmentalists, the spread of 'Western' consumer habits to countries like India and China represents an environmental calamity. And they have, for some time, warned about the potential environmental impact of spreading consumer culture to Asian countries. The World Watch Institute, for example, warned in 2004 that the planet could not afford so many people wanting access to household appliances, consumer goods, and increased mobility. To be fair to the World Watch Institute, it is equally concerned about the unsustainability of consumption in North America and Europe. The problem in countries like China and India, however, has to do with the size of the population and the argument that relatively small incomes in these countries leads to environmentally damaging consumption levels: "even at income levels that most observers would think of as subsistence – about $5000 – $6000 per person per year – people are already consuming at unsustainable levels" (Assadourian, 2010: 7).

Academic writing on consumption in countries where the global middle class is growing is beginning to provide more complexity to the stark difference between consumption as 'savior of the global economy' and consumption as 'environmental calamity.' This writing has explored the rise of a new class of consumer citizens, and the meanings and politics associated with new consumption practices. In China, the rise of a new class of consumers emerged after the death of Chairman Mao, and the period of market reforms that began in the 1980s. Consumption in China since then has been spurred by rapid economic growth as well as state policies that have, at times, encouraged spending rather than saving to stimulate economic growth and reduce the country's trade surplus.

Although there is some research that sees the rise of personal consumption in places like China and India as hedonistic and dystopic, most scholars have focused on the practices of consumption in these countries (Latham, 2006). An important theme in this work is to avoid seeing consumption as an unproblematic process of mapping global (or US practices) into new geographical contexts. Instead, consumption is viewed as a process that is shaped by situated historical, social, and political processes. The result is a growing body of work that is sensitive to the social and moral continuities and discontinuities between past and contemporary consumption patterns.

There are a number of interesting case studies that have explored the complexity of consumption outside Anglo-America. In her research on the culture of consumption in Korea, for example, Nelson (2006) found a high level of concern over excessive or inappropriate consumption practices. This is a culture that is highly concerned with "what kinds of consumption are appropriate – in terms of individual morality, national character, and mutual responsibility" (Nelson, 2006: 190). Pop singer Psy's wildly popular 'Gangnam style' music video and dance might be interpreted as a cultural reflection of Korea's attitude toward consumption: although it is not widely know, the hit-single is a biting parody of excessive consumption practices in Seoul's Gangnam district.

There are several other studies that have explored consumption in Asia through the analysis of specific commodities. Gong and Jackson (2012), for example, have recently examined the impact of the infant-feeding crisis in China in 2008. Although in one sense, the crisis represents the globalization of 'consumption anxieties,' which are well established in the West, the process in China is shaped by the specific context. Gong and Jackson (2012: 572) stress that consumption anxieties in China must be understood through the practices of an emerging middle class, the prevalence of one-child families, changes in interfamily relations, and the 'discourses of science and tradition' related to food consumption in general, and infant formula, in particular.

Talukdar's (2012) study of how women in India are consuming the 'ideal' skinny body shape imported from the West and promoted through advertisements for clothes and food provides an additional case of how globalized consumption practices are reshaped in local context. She argues that urban women in India have been able to "create individualised, personal experiences with their bodies and food while not disrupting familial and communal beliefs of eating and the body" (Talukdar's, 2012: 116).

A final example is Wiley's (2011) study of milk consumption in three countries – India, China, and the United States – which provides a fascinating example of how a single commodity means different things in different places. Although the globalization of milk consumption is associated with its positive relationship to growth (both personal and economic), the meaning of milk in China and India is shaped by cultural practices and political histories. In China, milk is seen as a "quintessential modern Western drink, identified with politically and economically strong nations," but in India, the meaning of milk must be negotiated in a context where the commodity is associated with both sacred cows and impure water buffalo (Wiley, 2010: 28).

What are the consequences of rising consumption of the global middle class? While there is a growing body of work on the impact of consumption in China and India on marine and terrestrial environments in other parts of the world (especially fish and timber), research on changing global value chains (GVC) is providing some unique insights on the impact of the changing geography of global consumption (Kaplinksy and Farooki, 2010). The GVC framework has been used since the early 1990s to analyze south to north GVC, and there is now

a very large body of work on this topic. A key finding of this research has been the role that Western markets have played in shaping production in southern countries through chains that are described as 'buyer driven.' Buyers in the north have shaped production in the south through their specific demands for quality and product standards. This has meant that in order to participate in global markets, southern producers have had to increase their capacity – or 'upgrade' – to meet the demands of northern buyers. In addition, southern producers have been required to adhere to a growing set of ethical and environmental standards. The key point is that export production in the south has been shaped by the specific demands of Western buyers.

The consequences of the shift of global supply chains from west to east are potentially profound for southern producers. It may lead to increased demand for commodities from southern exporters. But the nature of this demand is likely to change in a context where consumers are more interested in lower prices than they are in quality and product standards. In this way, it would reverse the trend of product and process upgrading that has characterized GVC in the recent past. The distinctive nature of 'Asian consumption' may also mean that these producers will be under less pressure to meet ethical and environmental standards (Kaplinksky et al., 2011). This research suggests that the changing global map of consumption will have a significant impact on producers, particularly those in the south used to supplying northern markets.

Consuming Connections

The growth of ethical consumption in the last 30 years has been extraordinary. Barnett et al. (2011) found that the consumption of ethical products in the UK increased from £13.5 billion in 1999 to £36 billion in 2009. Although these volumes are small relative to total sales of consumer goods, the sale of ethically traded products has been increasing at a rapid and consistent rate for several decades. The ethical label that continues to draw the most attention is Fairtrade. It is by far the longest-lived ethical consumption initiative and consumption figures been growing rapidly and consistently over more than three decades. Fairtrade products sold globally reached £4.3 billion in 2010, up by 27% from the previous year.

The production of academic writing on ethical trade seems to have kept pace with its growth in consumer markets. A key focus has been on the Fairtrade label, but research has also traced the emergence of a wide range of ethical labels that promise fair treatment of producers and workers, and environmental stewardship. Research has also explored new corporate-driven initiatives including 'consumer related marketing,' where the purchase of a commodity leads to a social investment somewhere in the developing world.

There are two central themes in the literature on ethical consumption. The first is the question of *mainstreaming*, which refers to the process through which corporate interests are appropriating alternative trading networks. Although corporate interests were until recently opposed to fair trade markets, the situation has changed considerably since the 1990s with large multinationals now selling fair trade certified products. Laura Raynolds (2009) has provided key insights into the process of mainstreaming through her work on coffee. She has argued that the Fairtrade market for coffee has fragmented into subchains that are more or less committed to the original goals of the movement. 'Mission-driven' companies maintain a strong commitment to the original principles of fair trade, and use the certification systems associated with Fairtrade to promote the sale of their products. The organizations identified as 'mission driven' do little retailing and most only sell products that carry the Fairtrade label. 'Quality-driven' firms, on the other hand, have become more institutionalized in the sense that their adherence to the fair trade certification system has more to do with how coffee and other commodities are sourced rather than on the values associated with the label. In other words, these companies have become involved with Fairtrade because of how these chains are coordinated, rather than by the goals of Fairtrade. 'Market-driven' companies represent the third group of firms associated with Fairtrade coffee. As Raynolds (2009: 1088) argues, this is the "largest and fastest growing share of Fair Trade certified coffee," and these companies have "little if any allegiance to Fair Trade's mission." The companies grouped under this category include multinationals like Nestle and Proctor and Gamble, as well as well-known retail and sales outlets like Starbucks.

Raynolds' analysis points to a fragmentation of Fair Trade value chains, with different levels of commitment to the original goals of the movement. Dolan (2010) and Goodman (2010) have provided a starker view of the effect of mainstreaming on the Fair Trade market, with equally bleak prospects for consumers wanting to make a difference. Dolan's findings are based on research conducted in Kenya's fair trade tea industry. There she found that the goals of fair trade are being undermined through practices of labeling and certification, which have become "surrogates for trust, dialogue and social ties" (Dolan, 2010: 38). In other words, the mechanisms associated with Fairtrade certification, including auditing and monitoring, are at odds with the social and developmental goals associated with fair trade. Dolan's argument, then, is that fair trade has become thoroughly institutionalized: besides "the social premium, there is absolutely no difference between the structures of the conventional tea trade and that of Fairtrade" (Dolan, 2010: 37).

Michael Goodman (2010) has provided an insightful analysis of the problem facing fair trade. Multinational corporations and global retail chains have played an increasingly important role in fair trade markets and have been largely responsible for recent increases in the volume of fairly traded goods. In one sense, this should be seen in a positive light: the more goods that are traded fairly, the greater the impact in the developing world. These increases might also be understood as supporting the positive role that ethical-minded consumers are having over capitalist markets. Goodman's argument, however, is that while the 'ends' of fair trade are being met, they are increasingly being achieved through 'means' that are not consistent with the goals and principles of fair trade. He writes: "in the desire to develop more collective politicized subjects in the consumers of fair trade and Northern consumers more generally, it really is the *means that matter* just as much as the ends when considering the specific practices and processes of how, why and by whom fair trade gets done" (Goodman, 2010: 115).

A second theme in the literature on ethical consumption concerns the politics and morals associated with a system that uses personal consumption choices as a driver of social development goals. Goodman is again an important critical voice in this theme: he argues that consumer choices in 'developmental consumption' commodity chains have become critical in determining whether development happens, and who gets receives this development. In this new era, "sustainable development can now and truly be store bought" (Goodman, 2010: 105). Goodman is not alone in raising concerns about how individual acts of consumption have become central to poverty alleviation. Rebecca Hawkins (2012) has written about several cause-related marketing (CRM) programs, which encourage consumers to purchase commodities that lead to different forms of social investment in the developing world. She is also concerned with how individual acts of consumption are marketed to consumers as potential solutions to complex development challenges in the south. The significance of Hawkins' work is that it is not the anonymous individual consumer that is important in her cases of CRM. Instead, it is women consumers who are targeted as key players in promoting a 'global motherhood' that connects caring mothers in the north with their needy counterparts in the south. The problem with this framing is that it limits "the possibilities for meaningful and respectful connections and political collaborations between women and more generally between distant people and places" (Hawkins, 2012: 757).

The politics of personal consumption are also evident in research on Bono's Product (RED) campaign, a program that Richey and Ponte (2008) have explored in considerable detail. The campaign brings together multinational corporations, celebrities, and fashion-conscious consumers to support Africans living with HIV/AIDS. The consumption of high-profile brands (e.g., Apple, Dell, Converse, Microsoft) by 'consumer citizens' in the north is used to support the campaign against HIV/AIDS in Africa. In this way, consumers can be both 'chic' and caring in their consumption choices. The problem for Richey and Ponte (2008) is that these campaigns focus attention on solving the manifestations of poverty and inequality – which can be solved through individual acts of 'low-cost heroism' – rather than the underlying causes of the problem located in unequal trade relations and exploitative labor practices. In this way, the attention is on "the welfare of Africans with AIDS in general, not of workers in factories producing RED products" (Richey and Ponte, 2008: 313).

An important theme in the writing on ethical trade is how it frames consumers as saviors of development and producers in the south as needy beneficiaries. This is a central theme in Le Billon's (2006) research on a campaign that has aimed at reforming the trade in conflict diamonds. While he is able to show that the efforts to reform the diamond trade had a positive effect, consumer action was mobilized through narratives of extreme violence and greed in diamond-producing countries. The effect of these narratives was to entrench the idea that "violence was a purely criminal enterprise driven by 'greed,' or a senseless attack on 'freedom,' rather than the outcome of exploitative and exclusionary histories of and uneven development" (Le Billon, 2006: 778).

Barnett and his colleagues have provided an important counterargument to the negative views of ethical consumption through a series of papers and, most recently, through a monograph (Barnett et al., 2011). They argue against the common view that "consumption activities are a sign of the substitution of privatized acts of consumer choice for properly political forms of collective action" (Barnett et al., 2011: one). Rather than seeing consumer choice as linked to the entrenchment of a neoliberal consuming subject, they argue that ethical consumption reflects new forms of political mobilization and civic involvement. Barnett et al.'s (2011) view is an important one in that it attempts to decenter the consuming subject in debates on ethical consumption, but it seems to be overshadowed by a growing literature on fair trade that is increasingly pessimistic about its potential to reform global trade and address underdevelopment and poverty (Moberg and Lyon, 2010).

The Consuming Body

The consequence of food consumption on the human body is the third spatial register considered in this article. It has become a key issue of policy debate and media attention, and it is most often articulated as a crisis associated with weight gain and, more seriously, obesity. The body mass index (BMI) is the ostensibly objective measure used to map and monitor over time the changing cartography of obesity. Predictably, these maps show that the problem is most severe in the United States where BMI rates have increased dramatically in the last 30 years. A recent report on the situation in the United States entitled *F is for Fat: How Obesity Threatens America's Future* found that adult obesity rates had doubled between 1980 and 2010. The report warns that unless something is done to reduce BMI trends, obesity rates could be as high as 44% in all US states, and as high as 60% in 13 states. The consequences of overconsumption and weight gain in the United States and elsewhere are usually framed in terms of personal health problems. Obesity is frequently cited as a cause of other serious health conditions including diabetes and heart disease. However, reports on obesity also warn of broader societal consequences including the impact on military preparedness and its effect on worker productivity (Guthman and DuPuis, 2006).

In the United Kingdom, obesity is viewed as an epidemic that is spreading from the United States with potentially dire consequences. Mepham (2010: 17), for example, has warned that the "impending epidemic of obesity in the UK is predicted to present a challenge comparable to that of global warming." His argument may have been based on a study conducted by the London School of Hygiene and Tropical medicine suggesting that fat people contribute more to global warming than thin people (cited in Guthman, 2012). As is the case in the United States, the impact of obesity in the United Kingdom is more than a personal problem. Mepham (2010) predicts that the costs of obesity to the UK's National Health Service will be as high as £10 billion by 2050. One of the notable responses to this 'impending crisis' in the UK has been to focus efforts on changing consumption habits of children. To this end, the UK has been conducting controversial programs to monitor the BMI of children at school.

Geographers have been relatively slow in developing a critical tradition of work on weight and obesity. Writing in the

mid-1990s, Rebecca Longhurst (1995: 252) argued that the widespread belief that 'people are voluntarily fat' was one of the reasons why geographers were not yet contributing to critical social science writing on obesity. She suggested that it "was time to write body shape *geographically*," and called for critical work on the diet industry, eating disorders, fat oppression, and the relationship between size and class (Longhurst, 2005: 247). Since then, a small but vibrant tradition of research has emerged on obesity that has explored a range of issues related to the discourse of obesity, social identity, and the spaces of consumption. The effort has included two special issues in the radical geography journal *Antipode*, and a new landmark book-length contribution by Julie Guthman (2012).

It is possible to identify several key themes in this article on the consuming body. The first has questioned the discourse of obesity and the problem of the 'obese subject.' Fundamental to contemporary debates on obesity is the Cartesian separation of body and mind. The issue is how, but more importantly, whether the mind is able to control the irrational urges of the body. In this way, bodies are "conceptualised as unruly, dangerous and in need of control" (Evans, 2006: 262). The emphasis on control and rationality has meant that individuals who become fat or obese, because they are unable to control their 'unruly bodies,' are seen as weak willed. Obesity and weight gain is therefore regarded as both a personal responsibility and a moral failing, and has led to marginalization of people who are considered to be overweight. Rawlins (2008: 146) shows that the problem of an obese subject goes beyond the personal: in the UK the "conception of a good citizen is an individual who looks after their own body not merely for their own sake, but for the good of the nation as a whole." As is the case in the United States, the consequences of the obesity crisis is in the United Kingdom are at the scale of the nation-state.

Social science writing on weight and obesity has been highly critical of the discourse of the 'obese subject' linking it other forms of social marginalization (e.g., racism). But it has done more: for example, researchers have raised serious questions about the taken-for-granted links between weight gain and heart disease or diabetes. The relationship between weight and health is complex, and public health policy often confuses cause with effect (Colls and Evans, 2009; Guthman, 2012).

Geographers have provided an important contribution to research on the 'obese subject' by exploring how individuals who are anxious about their weight negotiate consumption spaces. Hopkins' (2012) survey of a group of people who self-identify as fat or obese provides insights into how these individuals negotiate the discourse of obesity in everyday-consumption practices. Space and the relations between his informants and other people are critical to his analysis. In consumption spaces where Hopkins' informants felt uncomfortable, they would actively modify their behavior by limiting the amount of food or drink they consumed to demonstrate control over their 'bodily urges' and to show that they were actively trying to lose weight. In addition, they reported choosing clothes that they hoped would make them look thinner, and were careful to sit in places that they hoped would conceal their bodies from judgment. Hopkins's research provides an important example of how the discourse of obesity is negotiated in different consumption spaces and in the relations between individuals and groups.

Longhurst's (2012) extraordinarily candid and personal autobiography of weight loss provides a unique insight into the complex relationship between the consuming body, identity, and space. As a feminist scholar who has been at the forefront of critical debates on obesity, she reflects on her experience of losing weight and becoming thinner. In some ways, her experience is not dissimilar to what has been reported by individuals who have been the subject of research on obesity. As an overweight individual, she felt judged and uncomfortable in consumption spaces like restaurants, clothing stores, and galleries. And when she did lose weight, and others commented on her success, she felt that her "fierce determination, body vigilance and will power had been worthwhile" (Longhurst, 2012: 881). With a smaller body, she found it easier to negotiate in and through spaces of consumption. Yet, her efforts went against feminist writing on obesity that problematizes the issue of weight and the obese subject. In losing weight, Longhurst had to negotiate the challenge of dealing with colleagues and students who had felt she had somehow 'sold out' by becoming thin. As Longhurst (2012: 879) writes, the "spaces of weight loss are undoubtedly complex, fluid, contradictory and paradoxical."

A second theme within on the consuming body has focused on how the problem of obesity is known and measured. Given the significance of the BMI index in assessing weight gain, research has focused on how it is implicated in the performance of an obesity epidemic. The popularity and widespread use of the BMI is linked in part to its simplicity: it is a straightforward ratio of height and weight. This simplicity also means that individuals can self-report their BMI 'number' and assess it against the scale of thinness and obesity. In this way, BMI has become a means for individuals to self-regulate their consumption practices against an ostensibly objective scale.

The problem with BMI is that it is not an accurate measure of fat on the body, as it cannot distinguish between muscle, bones, and fatty tissue. The result is that individuals with large and highly muscled bodies are likely to have high BMI scores even though their body fat is relatively low. An additional problem with the BMI measure is that very small changes in BMI in a population can lead to significant changes in the proportion of the population measured as overweight or obese. This is because the BMI scale has hard and fast divisions between weight categories. The arbitrariness of the scale was revealed in 1999 when the scale for overweight was reduced from 29 to 25, which effectively increased the number of people in this category overnight. Finally, the BMI is not sensitive to gender and ethnic differences (Guthman, 2012), which skew the results.

Evans and Colls (2009) have contributed significantly to a critical understanding of BMI through their analysis of its use in the United Kingdom. In response to a perceived future obesity epidemic, the UK's Department of Health initiated a survey of children's BMI in schools in 2006. Evans and Colls explore how the measurement of BMI is implicated in processes of biopower with children's bodies becoming the object of regulation and discipline. Their analysis raises questions about the BMI as an "abstracted representation of the body" and reveals how the index has been used as a "biopolitical strategy which both regularises and disciplines bodies at a population and individual level" (Evans and Colls, 2009: 1053).

Julie Guthman's (2012) most recent book on obesity – *Weighin In* – represents a key contribution to research on the consuming body. It is difficult to do justice to this outstanding book in a short review, but it is, nonetheless, possible to outline the broad contours of her argument. Guthman draws on a rich tradition of research in critical political ecology to challenge existing approaches to obesity that focus on personal choice, and which medicalize and pathologize obesity. Central to her argument is the idea that the body has become a 'spatial fix' for the contradictions of neoliberal capitalism in the United States. In this way, obesity is not the result of poor personal choices, but is instead a consequence of neoliberal policies that have led to growing inequality and a larger class of people subsisting on marginal incomes. The food system, largely unregulated yet supported by a heavily subsidized agricultural economy, provides easy access to cheap food that is more fattening than it is nutritious. Neoliberal policies have also shaped the built environment and retail spaces in ways that encourage weight gain and obesity.

If the cause of obesity in the United States is the neoliberal food economy, then the solution cannot involve changing personal choices or developing alternative food markets. Instead, Guthman urges a focus on the political policies that have fixed the contradictions of neoliberalism in the (US) body. Guthman's important intervention provides an overarching analysis of the problem of obesity especially in the United States, with potential relevance to other countries affected by neoliberalism. In this way, her analysis decenters personal consumption habits, and the issue of choice and control, to suggest that weight gain and obesity should be understood in the broader political ecology.

Conclusion

This article of academic scholarship and popular writing on geographies of consumption has focused on the consequences of consumption at three spatial registers: the consuming body, consuming connections, and the globalization of consumption. Academic scholarship in these fields has provided critical insights into the consequences of consumption especially as they relate to questions of obesity, ethical and fair trade, and global consumption patterns. Although three spatial registers were discussed separately, there are important links between them. For example, the globalization of consumption is leading to what health experts are calling a 'nutrition transition.' This nutrition transition is regarded as the key reasons behind the spread of a 'pandemic of obesity' around the planet. In this way, the globalization of consumption and obesity are not separate processes, and demand the attention of geographers and other social scientists who have provided critical analysis of obesity in Western contexts. Similarly, there are links between consuming connections and the globalization of consumption. Academic writing on fair trade is increasingly concerned with how it is becoming mainstream, but a potentially more pressing question relates to whether the new global middle class in Asia and elsewhere will be attracted by ethical forms of consumption. This connection begs the question: what are the consequences for fair trade as global consumption moves from west to east?

There is a second way in which the three spatial registers are linked together. Scholarship on the geographies of consumption seems to be distancing itself from issues of personal choice in consumption practices. Instead, the emphasis is on balancing research on consumption practices with an analysis of the broader social and political processes that shape consumption and its consequences. This is especially evident in geographical writing on obesity, but it is also evident in writing on ethical consumption practices that attempts to displace the consumer from the central focus of study (e.g., Barnett et al., 2011). In shifting away from consumer choice, writing on the consequences of consumption also leads us inevitably, and appropriately, to questions of production, in the analysis of the consuming body, consuming connections, and the globalization of consumption.

See also: Consumption, History of; Consumption, Sociology of; Geographies of Food: Politics, Place, Power Reprised; Globalization: Geographical Aspects; Networks and Cultural Consumption.

Bibliography

Assadourian, E., 2010. The Rise and Fall of Consumer Cultures, in State of the World 2010: Transforming Cultures – from Consumerism to Sustainability. World Watch Institute, Washington, DC.

Barnett, C., Cloke, P., Clarke, N., Malpass, A., 2011. Globalizing Responsibility: The Political Rationalities of Ethical Consumption. Wiley, London.

Colls, R., Evans, B., 2009. Critical geographies of fat/bigness/corpulence. Introduction: questioning obesity politics. Antipode 41 (5), 1011–1020.

Dolan, C., 2010. Virtual moralities: the mainstreaming of fairtrade in Kenyan tea fields. Geoforum 41, 33–43.

Evans, B., Colls, R., 2009. Measuring fatness, governing bodies: the spatialities of the body mass index (BMI) in anti-obesity politics. Antipode 41 (5), 1051–1083.

Evans, B., 2006. 'Gluttony or sloth': critical geographies of bodies and morality in (anti) obesity policy. Area 38 (3), 259–267.

Gong, Q., Jackson, P., 2012. Consuming anxiety? Parenting practices in China after the infant formula scandal, food. Culture and Society 15 (4), 557–578.

Goodman, M.K., 2010. The mirror of consumption: celebritization, developmental consumption and the shifting cultural politics of fair trade. Geoforum 41, 104–118.

Gregson, N., 1995. And now it's all consumption? Progress in Human Geography 19 (1), 135–141.

Guthman, J., DuPuis, M., 2006. Embodying neoliberalism: economy, culture and the politics of fat. Environment and Planning D 24, 427–448.

Guthman, J., 2012. Weighing. In: Obesity, Food Justice and the Limits of Capitalism. University of California, Berkeley.

Hawkins, R., 2012. Shopping to save lives: gender and environment theories meet ethical consumption. Geoforum 43, 750–759.

Hopkins, P., 2012. Everyday politics of fat. Antipode 44 (2), 1227–1246.

Jarosz, J., 2009. Energy, climate change, meat and markets: mapping the coordinates of the current world food crisis. Geography Compass 3 (6), 2065–2083.

Kaplinksky, R., Terheggen, A., Tjaja, J., 2011. China as a final market: the Gabon timber and Thai cassava value chains. World Development 39 (7), 1177–1190.

Kaplinksy, R., Farooki, M., 2010. Global value chains, the crisis and the shift of markets from north to south. In: Cattaneo, O., Gereffi, G., Staritz, C. (Eds.), Global Global Value Chains in a Postcrisis World. A Development Perspective, World Bank, Washington.

Kharas, H., 2010. The Emerging Middle Class in Developing Countries, Working Paper No. 285, OECD Development Centre, Global Development Outlook. OECD, Geneva.

Latham, K., 2006. Introduction: consumption and cultural change in contemporary China. In: Latham, K., Thompson, S., Klein, J. (Eds.), Consuming China: Approaches to Cultural Change in Contemporary China. Routledge, London, pp. 1–21.

Le Billon, P., 2006. Fatal transactions: conflict diamonds and the (anti) terrorist consumer. Antipode 778–801.

Longhurst, R., 2005. Fat bodies: developing geographical research agendas. Progress in Human Geography 29 (3), 247–259.

Longhurst, R., 2012. Becoming smaller: autobiographical spaces of weight loss. Antipode 44 (3), 871–888.
Mansfield, B., 2011. Is fish health food or poison? Farmed fish and the material production of un/healthy nature. Antipode 43, 413–434.
Mepham, B., 2010. The ethical matrix as a tool in policy interventions: the obesity crisis. In: Gottwald, F.-T., Ingensiep, H.W., Meinhardt, M. (Eds.), Food Ethics. Springer, New York, pp. 17–29.
Miller, D., 2012. Consumption and Its Consequences. Polity, London.
Moberg, M., Lyon, S., 2010. What's fair? The paradox of seeking justice through markets. In: Lyon, S., Moberg, M. (Eds.), Fair Trade and Social Justice. New York University Press, New York, pp. 1–31.
Nelson, L.C., 2006. South Korean consumer nationalism: women, children, credit and other perils. In: Garon, S., Maclaclan, P.L. (Eds.), The Ambivalent Consumer: Questioning Consumption in East Asia and the West. Cornell University Press, Ithaca.
Rawlins, E., 2008. Citizenship, health education and the obesity 'crisis'. ACME 7 (2), 135–151.
Raynolds, L., 2009. Mainstreaming fair trade coffee: from partnership to traceability. World Development 37 (6), 1083–1093.
Richey, L.A., Ponte, S., 2008. Better (Red) than dead? Celebrities, consumption and international aid. Third World Quarterly 29 (4), 711–729.
Talukdar, J., 2012. Think but not skinny: women negotiating the 'never too thin' body ideal in urban India. Women's Studies International Forum 35, 109–118.
Trentmann, F., 2009. Crossing divides: consumption and globalization in history. Journal of Consumer Culture 9 (2), 187–220.
Wiley, A.S., 2011. Milk for 'growth': global and local meanings of milk consumption in China, India and the United States. Food and Foodways 19, 11–33.

Consumption, History of

Emanuela Scarpellini, Università degli Studi di Milano (University of Milan), Milano, Italy

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by V. de Grazia, volume 4, pp. 2682–2687, © 2001, Elsevier Ltd.

Abstract

The history of consumption has flourished over the last 30 years, interweaving traditional historical methodologies with features from a range of scientific disciplines. Its interests have developed along a chronological arc stretching from the Renaissance to contemporary phenomena of globalization, in search of the deep social, political, and cultural meanings that various forms of society have over time assigned to the material culture.

Definition

Consumption is an integral part of human behavior, so forms a key piece in the overall historical jigsaw. However, it was not until the 1980s that the culture and practice of consumption and the consumer as an important player in the drama became specific subjects of analysis. This relatively late arrival on the scene reflects a concentration on production, and is now both the strength and weakness of this research field, which has proved to be exceptionally dynamic, yielding outstanding results. On the one hand, research has exploited cross-cultural results, adopted interdisciplinary approaches, and differentiated its sources of inquiry. On the other hand, though, it has shown a degree of heterogeneity and sometimes a lack of theoretical background. Over the last 30 years work on the story of consumption has undergone exponential growth, often in the form of a proliferation of specialized subfields (food, material culture, fashion and design, the home, mass media and communications, gender, etc.) rather than a unified narrative covering a particular period – although several works have attempted a synthesis. The element that links the many research initiatives is an awareness that we are dealing with a primary aspect of historical reconstruction, a cultural process embedded in the social, economic, and political background. This process has its own dynamics and involves a large number of actors. The word 'consumption' derives from the Latin *consumere*, to waste, to eat, and its history is basically the story of the relationships between the members of a historically defined society and their material culture, in all meanings of the word.

Intellectual Background and Theories

As stated above, the history of consumption has encouraged the adoption of an interdisciplinary viewpoint, involving a variety of theories and intellectual backgrounds. These are the driving force behind research into sources, interpretations, and historical writings.

The theories of consumption that developed in modern times are rooted in the crucial period at the end of the ancien régime around the seventeenth century, a period that marked the end of a consumer pattern that directly reflected the hierarchical division of classes through sumptuary laws. These laws regulated the quality and quantity of usable goods, especially clothing, on the basis of class, profession, and gender, indirectly reinforcing the social status quo, as Alan Hunt has shown. In fact, what was at issue here was 'luxury' rather than 'consumption', and luxury was usually condemned in religious and moral terms.

It was Bernard Mandeville with *The Fable of Bees* (1714) who overtly contested the positions held at the time, suggesting that private vices – aspirations to luxury – were a seamless complement to public virtues – the development of commerce and the growth of communal wealth. This cast doubt on the fundamental link between parsimony, morality, and socio-economic growth in nations throughout history. The eighteenth century saw protagonists becoming entrenched in one of two positions. One group continued to stigmatize luxury as a form of inequality – in *The Spirit of the Laws* (1748) Montesquieu condemned luxury in the political structure of a republic, where all citizens are called to work toward the common good, but accepted it under a monarchy, where it would symbolize the inevitable social gulf and encourage the development of trade. Rousseau wrote that only austerity was suitable for a virtuous republic, on the pattern of ancient Sparta. On the contrary, David Hume in his essay *On Refinement of the Arts* (1742) justified luxury in modern society as an engine for the economy, recognizing that its significance was very different in ancient societies, as shown by Christopher J. Berry in his book on the concept of luxury. In his work *The Theory of Moral Sentiments* (1759) Adam Smith assigns a particularly important role to the theory of consumption: the author restates the need to separate moral from economic judgments, emphasizing that individual egoism and the new market mechanisms could coexist with luxury and social inequality, damping down the 'revolutionary' effect that the spread of luxury seemed to be having in the societies of the time.

Throughout the nineteenth century and on into the early twentieth century, the process of industrialization in Western societies produced many important effects, above all a focus on the economic and statistical aspects signaling a change in consumer patterns at all social levels. Alongside the references to consumption in the works of the most famous economists – David Ricardo, Alfred Marshall, Léon Walras, and finally Karl Marx, with his distinction between value of use and value of exchange of goods – we should spotlight two great scholars, Le Play and Halbwachs. Frédéric Le Play was a pioneer in research

into family budgets, showing the contribution of statistical instruments to the growth of knowledge in this field, shifting the center of gravity of research from single individuals to families as a fundamental unit of consumption. In 1913 Maurice Halbwachs developed an influential theory that contested 'Engel's law', where family consumption increased in direct proportion to wealth (from basic consumption of goods like food to more sophisticated outlets for wealth like culture, well-being, and travel). By introducing elements of social and cultural analysis he showed that at equal income levels, cultural elements could radically alter consumption patterns – for example, a worker's family and a middle-class family could have a similar income but diverse consumption.

At the same time, Werner Sombart's works, with their background in economy studies, exerted a great deal of influence. He saw luxury as one of the basic differences between the modern capitalist society and societies of the past. Thorstein Veblen was extremely influential in the United States: in his work *The Theory of the Leisure Class* (1899) in particular he analyzed overt displays of consumption by the new affluent classes as a symbolic banner deployed by the new rich to make a public statement of their wealth in a society where old distinctions of birth and class had definitively broken down. 'Conspicuous consumption' became one of the main methods of displaying social status, and came to form a specific feature in the profound economic and social transformation that was taking place in modern capitalist societies.

Many of these theories, especially those with a critical intent, were taken up in the years to come by the analyses of the Frankfurt School, which played a crucial role in examining consumption throughout the mid-twentieth century, from the late 1930s to the 1960s and 1970s. Many works treated the new consumer world, with contributions from high-profile figures like Walter Benjamin, Theodor Adorno, Max Horkheimer, and, later, Herbert Marcuse. However, the work that perhaps had the greatest impact was *Dialectic of Enlightenment* (1944), where Adorno and Horckheimer, taking a typically Marxian analysis as their starting point, developed a hard-hitting critique of consumerism, seeing it as the tool of a new form of oppression that acted through the manipulation of desires at every level of society, exercising power in a new way. Everything became a commodity; everything was bought and sold. Culture no longer had a libertarian, critical function, it had been transformed into an industry, becoming a consumable in its turn. In his work *The Society of the Spectacle* (1967), Guy Debord produced a definition that met with great success, identifying the entertainment-consumption tandem as a key to reading contemporary society – from a negative stance, of course.

Although they adopted a more concrete approach with a less negative thrust, theorists from the United States like John Kenneth Galbraith (*The Affluent Society*, 1958), Vance Packard, and David Riesman also highlighted the manipulative mechanisms involved in consumption in contemporary society, in this case linked above all to the activities of large corporations and the mass-media world. However, the late 1970s saw the development of a field of study on consumption that would lay the foundations for a different viewpoint. In particular the cultural studies tradition nurtured at Birmingham by Richard Hoggart and Stuart Hall began to analyze the new youth subcultures that were often expressed through new forms of consumption. These highlighted broad areas of independence and choice among such consumers, regarded for the first time as autonomous subjects, not passive victims of political and commercial manipulation.

Cultural developments in the 1980s led historians of consumption to cast their net even wider, taking in theoretical contributions emerging from the fields of semiotics, sociology, anthropology, and philosophy. One of the most influential thinkers was Roland Barthes, who spoke explicitly of 'mythologies' underlying even the most everyday acts of consumption. These acts became significant in themselves, and were rooted in complex historical contexts. Every object is made up not only of a material form, it is also charged with a 'universe of meanings' and deep cultural connotations that must be decoded. Every object is a sign of language. Jean Baudrillard studied above all the unreal dimension of the world of consumption, underscoring the proliferation of 'simulacra' created for commercial purposes. Drawing on reality as well as contents of the imagination, they eventually replace reality itself. The glamorous world of shopping and entertainment become our existential reference points. The work emanating from sociology departments also had a great impact, with Michel de Certeau and his analysis of the creation of social practices through everyday consumption, while anthropology also made important contributions. In this connection, the two main figures were Mary Douglas and Arjun Appadurai, who assigned a crucial role in communicating our identity to objects and material culture, given their ability to incorporate meanings. Pierre Bourdieu's notion of 'habitus', developed in relation to strategies of social differentiation based on consumption, had a great impact. Finally, work on social dynamics in contemporary society analyzed the creation of collective aggregations through shared consumption patterns, partly through the use of technical software.

Current Knowledge and the Main Achievements

Against this complex intellectual background, in the 1980s the field of consumption eventually emerged as a key paradigm for any understanding of modern society. It is no surprise that the variety and complexity of the intellectual stimuli underpinning this movement encouraged large numbers of heterogeneous contributions.

One of the first problems facing historians was to identify the appropriate periodization. Some identified the Italian Renaissance as the dawn of the consumer society, when a new value was given to the individual and material culture by the court and intellectual circles, creating an unprecedented market for luxury consumer goods and goods for personal use (Goldthwaite, 1993). Many others, however, have seen the seventeenth and eighteenth centuries as marking the birth of consumption in the modern sense. Norbert Elias regarded the development of a court society along the lines of the court of Versailles, which ranged from 'good manners' to appropriate styles of consumption, as a crucial cultural stage in the formation of modern centralized states. Simon Schama, on the other hand, sees the Dutch Golden Age as the most striking example of the spread of a new culture of consumption embodied in everyday life, at the basis of a new mercantile

culture where patterns of consumption interweave with ways of life inside and outside the home, and extend to interpreting gender roles, looking at art and experiencing wealth (Schama, 1987).

For many, though, eighteenth century Britain, and in part colonial America, was the crucible where a profound transformation was forged. Neil McKendrick, John Brewer, and J.H. Plumb have maintained that at this time a large market for goods was created at a national level, and people from various social classes took part in this exchange as 'consumers' (McKendrick et al., 1982). Many other works followed, such as those by John Brewer and Roy Porter, providing a more in-depth analysis of the various aspects of a consumer culture associated with the mercantilization of public life, the increasing availability of consumer goods, the influence of consumption in the structure of the social hierarchy and family and gender roles. This field of study was joined by two other important sectors that grew chronologically at least throughout the nineteenth century, identified by experts as another crucial node because the industrial revolution was now in full swing, a phenomenon underpinning the creation of a mass market. The first sector is associated with the development of new places and spaces for consumption, especially the department stores that began to appear around the mid-nineteenth century in Europe and the United States. Michael B. Miller showed how the ideals of an ascendant bourgeoisie – solidity, size, wealth, and morality – were reflected in the sales and communications practices, and even physical appearance, of a Parisian department store, 'Au Bon Marché' (Miller, 1981). Later works presented the experience of other countries – Heinz Gerhard Haupt, Wolfang König, and Christian Kleinschmidt for Germany for example – and described how these places became public spaces that were also open to females, now constructing their identity as 'modern' bourgeois women, as shown in works by Erika D. Rappaport, Geoffrey Crossick and Serge Jaumain, Elaine S. Abelson, and others. The second field of study, which made perhaps the most important contribution to the history of consumption, is linked to gender and women's history. Often ignored by 'official' social and political history, consumption proved to be central to an understanding of cultural transformations in the role of women inside and outside the home. They could appear outside the home in the places of trade and commerce, but not without moral qualms regarding the consequences of this desire for luxury, which could lead to kleptomania and promiscuity, endangering the virtue of the young department-store assistants. Inside the home this transformation took the form of constructing a form of domesticity embodied in the interiors of a bourgeois home – its furnishings, kitchen, and decor – as shown, for example, by Clive Edwards, Deborah Cohen, and Maxine Berg. The examination of domesticity continued with interesting results throughout the twentieth century, giving rise to pioneering works like those by Ruth Schwartz Cowan on the transformations – or failed transformations – of the role of the housewife, especially after the gradual introduction of domestic appliances (Schwartz Cowan, 1983).

The seminal work carried out by Sidney W. Mintz cast light on another important area (Mintz, 1985). Using a variety of sources, with an emphasis on anthropological works, he showed how the commercialization of sugar in modern times took on a range of meanings. On the one hand it became a turning point in the organization of domestic life and a sign of the rise of the working classes, to whom it was a cheap source of calories and a symbol of social prestige, driven by commercial advertising. On the other hand it was an integral element in the international trade and politics of the British Empire, inclining owners to respond to growing demand for the product by increasing the size of sugar plantations and the numbers of slaves employed on them. In this way consumption became linked to important international policy and trade decisions, and showed how it could drive change by overturning the traditional view that saw consumption as simply a consequence of production. Attracting agreement and criticism alike, Mintz's work was followed by many analyses examining the complex relationship between Europe and its colonies through the circulation of consumer goods. A recent example is David Ciarlo's work on imperial Germany.

In this way the history of consumption shows how goods and their cultural, political, and economic significance lay behind the appearance of little-known phenomena. These included the construction of domesticity, the role of women and sociocultural transformations of the family, as well as the creation of transnational public spaces for social interaction and political transactions. They also fostered the growth of a hedonistic culture with romantic origins within a range of social classes that would exist side by side with the work ethic that Max Weber regarded as the basis of capitalism, as suggested by Colin Campbell. What emerges is a dual significance linked to the development of consumption studies. On the one hand it clarifies aspects of material and everyday life that are often ignored, making an important contribution to knowledge within the field as a whole, while on the other it creates a prism through which we can observe the social dynamics we have already mentioned or particular historical moments, discovering elements that shed new light and provide comprehensive new interpretations.

From this viewpoint the studies on the twentieth-century dictatorships – especially Nazism and Fascism – are extremely interesting. They have emphasized how the propaganda pumped out by these regimes included a degree of attention to the politics of consumption, despite the strident emphasis on war and aggression. The creation of after-work organizations for holidays, sports, or games, intended to integrate the general population within the party structure, developed into a search for consensus and had the indirect effect of institutionalizing an attitude toward consumption and forms of collective amusement, organized by party or government, as described by Hartmut Berghoff, Jonathan S. Wiesen, and others.

This statist model was successfully contested by a model based on individual and family consumption, reaching its purest form in the United States and the 'American dream', the dominant concept in the post–World War II years. There appeared a quantity of studies examining this phenomenon in the US. An example is the work of Gary Cross, who mapped out a historical profile of the crucial importance of consumption, until it finally became an ideology in itself, one that permeated the national culture. This had a two-edged effect: it drove economic development and created unprecedented levels of well-being, but also often led to social fragmentation and decreasing political and social participation (Cross, 2000.

See also the contribution by Regina Lee Blaszczyk on the period 1865–2005). Lisbeth Cohen also traced a political arc from the New Deal to the post–World War II period, describing the gradual transformation of the citizen–consumer into the consumer–citizen. The state gained importance as a provider of basic goods and services, and the search for consensus increasingly involved the guarantee of a certain standard of living (Cohen, 2003). Susan Strasser, Charles McGovern and Matthias Judt, Martin Daunton and Matthew Hilton, Juliet B. Schor and Douglas B. Holt also produced some interesting work along these lines. From a perspective that emphasized the political fallout, many other works identified consumption as an important factor encouraging immigrants to the US to integrate into the society of their new home, highlighting how certain lifestyles, and especially food consumption, were a key to understanding the social dynamics of the country's various ethnic groups, as shown, for example, by Donna R. Gabaccia. Equally interesting studies analyzed the formation of specific ethnic markets, like the Afro-American community, which moved from widespread poverty to a strong economic growth, especially in the second half of the twentieth century, eventually making up a large section of the consumer goods market, and gradually being integrated into the country's overall development (see Robert E. Weems).

Historians have identified the exporting of the American consumer model to Europe in the postwar years as an essential historical watershed for the twentieth century. In an important contribution, Victoria de Grazia maintained that one of the strengths of the American model was its 'soft power' (with reference to the theories of Joseph Nye), promoted through new arenas for consumption like cut-rate stores and supermarkets, the construction of the figure of the modern housewife, the increasing popularity of fast food, the export of a dynamic entertainment industry, and so on. The book was part of a significant field of study that focused on the effect of exporting goods and consumption models between different places, assessing the results of this cultural hybridization with reference to the historical context. Authors like Richard Pells and Rob Kroes made important contributions in this connection. A number of works examine the effects of this phenomenon in Europe – Richard Kuisel on the ambivalent relationship of the French to Americanization, Volker Berghahn on Germany, Reinhold Wagnleitner on Austria, and Emanuela Scarpellini on Italy. Jonathan Zeitlin and Gary Herrigel have also produced interesting work in the field of economic history, examining the impact of American technology and new management techniques on the development of the European economies, while Ruth Oldenziel and Karin Zachmann look at the influence of the American model on the European domestic scenario.

The overall picture emerging from these studies is that in the movement from one cultural reality to another, the significance of objects and attendant social practices undergoes a series of transformations by which they are hybridized and creolized; this can lead to total rejection by or enhancement of indigenous traditions – or even newly invented versions of them – in response to the new models being offered. Research into these cultural transformations can be seen as one of the main contributions of historical research into consumption in recent decades. It spearheads the historical debate on the meaning and scope of the globalization process from the viewpoint of consumption, with all the ensuing positive and negative effects on contemporary societies.

Changes in Focus

There have been a number of new developments in the field over recent years. Generally speaking, there has been an attempt to deal with the 'cultural turn', returning, for example, to 'materiality', as Frank Trentmann suggests, so the focus on cultural aspects has widened to include social, economic, and political upheavals. However, rather than tracing a single line of development, the result is a much broader range of inquiry.

One important aspect to consider is the broader geographical viewpoint and the history of consumption in non-Western countries. Two trajectories emerge – on the one hand there have been shown to exist flourishing traditions in consumption, luxury, trade, and consumption in places that were totally free from Western influence, in both ancient and modern times. On the other, it has been pointed out how frequently contact with the West has stimulated specific positive or negative reactions. A pioneering example is the contribution made by the historical and anthropological work of Jack Goody. In the context of basic food consumption he emphasized the wealth of the traditions of ancient Egypt, imperial Rome and medieval China, examining in particular the indigenous traditions of various African regions (Goody, 1982). Karl Gerth showed the importance of consumption in the development of modern China. Above all, he studied the close link between the appearance of a consumer culture and the spread of nationalism, especially in the early twentieth century, when it marked a form of opposition to foreign invasion (Gerth, 2003). Japan is another interesting case that has attracted increasing attention. Japan, too, is an ancient civilization that experienced quite advanced forms of urbanization in its main centers, Edo, Osaka, and Kyoto, then had to deal with the impact of contact with the West in various stages up to the twentieth century. From the initial assumption that the modern forms of consumer culture essentially emerged following the encounter with Western culture, the emphasis moved on to the analysis of distinctive local features and a reassessment of indigenous characteristics (see Penelope Francks in Berghoff and Spiekermann, 2012). Africa is different for a variety of reasons. Although the trade links between Africa and the West and Asia are rooted deep in the past, few researchers have examined trade and consumption, as the poverty endemic in so many African states, which prevents standards of living reaching the levels enjoyed by the more developed countries, has been seen as militating against the appearance of a consumer culture. For this reason Africa has long been regarded as being dominated by a subsistence economy, becoming simply a source of raw materials to be exported to the rich countries. Recent studies, however, have highlighted the limitations of this viewpoint. First of all, there are huge economic and social disparities within Africa, so it is impossible to make generalizations. Secondly, the relative scarcity of available goods does not prevent the development of a specific 'culture of consumption', with hierarchies of values and original forms. Take, for example, the traditional communal property giving

way more and more to private ownership, or the deep political implications of the everyday objects of consumption associated with the colonizers. Finally, the scarcity of resources itself highlights the role assumed by governments and international agencies in influencing the direction of developments in consumption (Hahn, 2008). Equal interest has recently been shown in other geopolitical areas like Russia and the Eastern European countries, which are moving rapidly toward Western-style consumption patterns, or the Islamic countries with their complex interweaving of politics and religion, a reason why consumption becomes an inherent aspect of the globalization process (Stearns, 2001).

Ecology and the environment are also important themes for historians of consumption. The development of consumer societies across the world has led to pollution and the depletion of available resources, with a rapid increase in the size of humanity's environmental footprint. Consumption is one of the causes of this – as Susan Strasser reminds us, all we have to do is to look at the problem of urban waste and its recycling – even though it is often only a contributory element in studies that focus more closely on ecological problems. It is interesting to note that there appear to be two solutions to ecological crises – measures taken by the authorities, be they governments or NGOs, and changes in consumer behavior, especially following the appearance of a responsible 'green consumer' movement. This leads us to another crucial theme in recent studies on consumption, consumer organizations. Lawrence B. Glickmann tells the full story of consumer movements in the USA, from the first loosely organized boycotts to the development of fully fledged associations comparing the advantages and problems created by mass production. In the late twentieth century, the problem began to be expressed in terms of consumer protection. Matthew Hilton compares the development of consumer movements in the West with the movements emerging in less-developed countries. In the richer countries, attention has always focused on protecting consumers from abuse by producers by pushing for a suitable, effective legal framework, although more recently the spotlight has been turned onto environmental issues. Ralph Nader is perhaps the most famous leader among these movements, and he showed just how much the legal aspects of consumption have their political side, too. Another iconic figure is Anwar Fazal, president from 1978 to 1984 of the International Organization of Consumers Union, then Consumers International, a movement that now has a presence in 115 countries. He took a different approach – the priority in many less-developed countries is not consumer protection, but the attempt to develop and spread as far as possible a reasonable level of consumer standards. What is more, the values underpinning Western consumerism, from individual ownership and pluralism, run up against a well-established tradition of common use of goods and leadership. As a result many Asian countries are seeking for an alternative to consumerism, one inspired by 'Asian values' that can repel the invasion from the West. Once again, consumption becomes politically charged and NGOs and international organizations become very important, overturning the accepted patterns of national political power (Hilton, 2009).

Many new insights have been offered by studies on consumption. As well as providing international viewpoints, they also turn their attention to areas like visual culture, take an increasing interest in the impact of new technologies on consumption models, make a more in-depth analysis of formal and informal political players, and examine fractures in the social framework – with great emphasis on gender, taking masculinity into account, age, ethnic factors, the influence of old and new immigration, etc. The promotion of important international initiatives for the study of consumption by a number of stakeholders is a sign of growing maturity in this field. An example is the 'Cultures of Consumption' program led by Frank Trentmann from 2002 to 2007 that produced many original studies and gave a great stimulus for examination of the sector. The history of consumption continues to provide an original mirror of the development of historical studies.

Methodology

In the light of what has been seen above, it is very difficult to identify a specific methodology for the history of consumption. In fact, it makes more sense to speak of different methodologies existing side by side. On the one hand historians make frequent use of traditional primary sources like private and public archives, printed materials (magazines, newspapers, and collections), collections of objects, visual sources, and advertising; while on the other there is increasing use of audiovisual sources linked to new media, the Internet, interviews, and oral histories. Of course, the use of such varied sources creates very different methodological problems, sometimes concerning different fields of specialization, and it is not always easy to integrate these into an overarching historical narrative. It goes without saying that the reference period in the history of consumption extends over centuries, so the tools of study and methodologies chosen must adapt to suit the characteristics of the subject with a certain degree of flexibility.

We should also point out that the reverse side to this great wealth is a persisting lack of theoretical reflection in the structure of the studies and methodologies adopted, something all too often implicit in the narrative. In the same way, complex studies of chronological and geographic synthesis exist beside microstudies, and the knot in the relationship between a national and transnational perspective does not yet appear to have been untied. Although a great deal can be done in this direction, the quantity and quality of the results obtained over 30 years remind us that this 'multiple' perspective is not the product of chance, but may be a characteristic of scientific studies that wittingly stand on the frontier.

See also: Colonization and Colonialism, History of; Consumption, Sociology of; Environmental History; Gender History; Globalization and World Culture.

Bibliography

Berghoff, H., Spiekermann, U. (Eds.), 2012. Decoding Modern Consumer Societies. Palgrave Macmillan, New York.

Cohen, L., 2003. A Consumers' Republic: The Politics of Mass Consumption in Postwar America. Knopf, New York.

Cowan, R.S., 1983. More Work for Mother. Basic Books, New York.

Cross, G., 2000. An All-Consuming Century: Why Commercialism Won in Modern America. Columbia University Press, New York.
De Grazia, V., 2005. Irresistible Empire: America's Advance through Twentieth-Century Europe. Belknap Press, Cambridge.
Gerth, K., 2003. China Made: Consumer Culture and the Creation of the Nation. Harvard University Press, Cambridge.
Goldthwaite, R.A., 1993. Wealth and the Demand for Art in Italy, 1300–1600. Johns Hopkins University Press, Baltimore.
Goody, J., 1982. Cooking, Cuisine and Class: A Study in Comparative Sociology. Cambridge University Press, Cambridge.
Hahn, H.P. (Ed.), 2008. Consumption in Africa: Anthropological Approaches. Lit, Berlin.
Hilton, M., 2009. Prosperity for All: Consumer Activism in an Era of Globalization. Cornell University Press, Ithaca and London.
McKendrick, N., Brewer, J., Plumb, J.H., 1982. The Birth of a Consumer Society: The Commercialization of Eighteenth-Century England. Indiana University Press, Bloomington.
Miller, M.B., 1981. The Bon Marché: Bourgeois Culture and the Department Store, 1869–1920. Princeton University Press, Princeton.
Mintz, S.W., 1985. Sweetness and Power. Penguin, New York.
Schama, S., 1987. The Embarrassment of Riches: An Interpretation of Dutch Culture in the Golden Age. Knopf, New York.
Stearns, P.N., 2001. Consumerism in World History: The Global Transformation of Desire. Routledge, London and New York.

Relevant Websites

http://www.consume.bbk.ac.uk/publications.html#bibliography – Bibliography.
http://retailhistory.wordpress.com/about/ – Chord – The Centre for the History of Retailing and Distribution.
http://www.consume.bbk.ac.uk – Cultures of Consumption Program.
http://csdi.rca.ac.uk/didb – Domestic Interiors Project.

Consumption, Sociology of

Miki Nakai, Ritsumeikan University, Kyoto, Japan

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by K. Ilmonen, volume 4, pp. 2687–2690, © 2001, Elsevier Ltd.

Abstract

With the advent of consumer society, sociologists have investigated and developed theories on a variety of aspects of consumption to explain its relation to social structure and its cultural and social meaning. Although the focus has been in concert with social and economic changes, scholarly work has long established a link between social class/stratification position and cultural consumption, one of the classical perspectives that has a long history.

Introduction: Significance of Sociological Approaches to Consumption

Consumption has been studied in many fields of social and behavioral sciences such as economics, public policy, psychology, marketing, home economics, anthropology, history, as well as sociology. In early works in social sciences, consumption has often been conceptualized through analysis and interpretation of the sphere of production in the course of exchange of commodities in the market. Accordingly, one might think that economics has been in a central position in studying a variety of aspects of consumption. It is true that economists have spent considerable efforts trying to understand individual and household needs and expenditure, inherent qualities of goods and services, and how goods and services meet practical needs of consumers. However, most of the issues that are addressed in both disciplinary and interdisciplinary literature on consumption are absent in economics.

Mainstream economists' theories assume that consumers or households, as rational and selfish human beings, will attempt to maximize their overall levels of utility so as to satisfy their needs and wants. The levels of utility, or consumer preferences, are in principle independent of one's relationship with others. Economics has focused on how a society provides for consumer demand, but has not focused on the cultural qualities of goods, or what we call social and symbolic meanings of goods. Consumption in modern society is not what classical economic theories premised. Consumers often buy products for reasons other than the product's functional performance. They rather consume products for the cultural and symbolic meanings of the product. Consumption is not only a phenomenon that depends on the analysis of production; products are not just objects of use corresponding to requirements in the process of consumption. In fact, the essence of consumption lies in the irrational nature of human being, for example, insatiable appetite, luxury, out of ostentation, out of consideration of style, a demand for goods and services motivated by social stratification, etc. It is also an absolutely essential perspective for analyzing and theorizing consumption that it is influenced by real-world economic and social environment and the nature of consumption changes fundamentally with the time. It is important to take into account not only interaction between the agents communicating with each other but also analysis of the societies, transition in the nature of consumption, and interaction between consumers and their social settings.

With the advent of consumer society, sociological study of consumption has considered people's social relationship with others in the process of theorizing consumption behavior such as what motivates people to consume, what is the cultural meaning of consumption, how structures of inequality intersect with consumption, how people decide what products to consume, how advertising and marketing perform, what can result from them, etc. The arguments and concepts include, for example, conspicuous consumption by Veblen, pursuit of fashion, standard package by Riesman, and distinction. Sociologists have investigated and developed theories of consumption to explain the issues involved taking particular note of its relation to social structure and its cultural and social meaning.

Theoretical Development and Transition: Social Changes and Transition in Focus of Sociology of Consumption

The focus of sociology of consumption is in concert with social and economic changes. Sociological research on consumption is thought to have begun to emerge as early as the nineteenth century in response to the birth of capitalism and consumption revolution accompanying the modern Western market economy. When industrial revolution brought about many economic and social changes, the emergence and spread of the innovations of communication technology, the huge development of urban commercial space and mass press, and commercial advertising affected the consumption behavior of people.

Until the early stage of development of capitalism, only the highest ruling class people were able to exercise considerable discretion in consuming. Scarcity of necessities had been a central concern for most other social classes. From the late nineteenth century onward, with the growth in production, in Europe, England in particular, and America, the consuming class was no longer confined to the selected highest class and extended to the inferior classes. This brought about the rise of the middle classes in the nineteenth and especially twentieth centuries. Profound changes in the sphere of consumption caused by development of capitalism and the transformation in

the economy have drawn the interest of some sociologists. They include, for example, Tarde, Simmel, and Veblen, who theorized consumption and fashion. People increasingly tended to decide what to consume in attempting to differentiate themselves from others, and therefore many sociologists focused on ostentatious display of wealth or status-seeking consumption and the relationship between social class, consumption, and inequality. Also, the elites indulged in particular ways of consumption in order to differentiate themselves from the working class and kept changing their taste once their style was imitated by the people in the working class, which resulted in the emergence of fashion.

However, the early twentieth century, at the same time, was the period that the American market reached overproduction, which means production of goods exceeded consumers' demand. Especially in the United States around the 1920s, corporations started advertising to the working class in an effort to encourage new desire, embracing the idea that there were no limits on consumer wants and that such wants could be endlessly created and shaped. The social phenomena such that manufacturers create more demand by commercial advertising and other promotional devices and that consumption was controlled by manufacturers was critically analyzed by some sociologists. Critical theory, which emerged under the leadership of Adorno and Horkheimer, condemned consumer society. In the United States, from the post–World War era onward, the focus of a lot of sociological research such as that of Galbraith and Riesman has been on consumption pattern motivated by and fueled by the media, the effect of the emergence of mass, lifestyle and urbanization, and consumption as differentiation.

Since the 1970s, with highly developed consumer society as a backdrop, major theories of consumption have developed from varied perspectives. One of the influential arguments is Bourdieu's theory of cultural consumption and lifestyle. Another one is the work by Baudrillard, who is regarded as a priest of postmodernism, from symbolic approach to consumption. Several other postmodern theorists in the late twentieth century (e.g., Bauman, Beck, Giddens) also maintain, in contrast to the position by Bourdieu, that social stratification position has declining influence on the formation of lifestyles and the objects and patterns of consumption. They argued that in the postmodern era, where there is an increasing number of available commodities, lifestyles and consumption are critical to identity formation; at the same time, individuals are compelled to choose their identities. What this stress of the freedom implied by consumer choice means is that individuals have an equivalent degree of personal responsibility. Although postmodernists have emphasized the lack of correspondence between socioeconomic status and patterns of consumption, mainstream sociological research of consumption still focuses on the relationship between social stratification position and consumption. A lot of authors mostly in the United States, inspired by Bourdieu, have proposed a new relationship and explored applications of Bourdieu's theory of taste from around 1990 on the assumption that there are significant differences between France and other cultural settings such as America (Lamont and Lareau, 1988; Lamont, 1992; Peterson and Kern, 1996).

A more modern example of the authors who maintain a critical stance toward our consumer culture is Schor who coined the term competitive consumption to signify competition for conspicuous, status goods so as to keep up with the norms of the social group. The focus is on jeopardizing the quality of American life due to a growing aspirational gap, which means that desire and need exceeded consumers income, while at the same time pointed out the emergence of a society of simple livers, or downshifter.

Early Sociological Theories on Consumption: The Advent of Consumer Society and Consumption

One of the first systematic and influential studies of consumption and commodities in capitalist societies is by Thorstein Veblen (1857–1929). The ideas and concepts in his pioneering work *The Theory of the Leisure Class* published in 1899 made an important contribution to the field and following theorists. He emphasized that being able to be extravagant is a sign of privilege and he analyzed the feature of the upper classes: conspicuous consumption and conspicuous leisure fulfill the role of demonstrating their status to other social classes, the former displays the luxury material commodities and the latter, freedom from productive labor. He maintained that conspicuous consumption has an advantage over conspicuous leisure in a situation that the Gemeinschaft-type society, where people tend to settle in a certain region and know their neighbors, gave way to the Gesellschaft-type society, where one is anonymous and there is greater freedom from neighbors, that is, free time is comparatively invisible in impersonal urban societies, which was going on in the world at that time in the midst of rapid industrialization and liftoff.

As with Veblen, a lot of sociologists have theorized the notion that consumption plays the role of a method for achieving or expressing social status and prestige. Simmel and Tarde see consumption as a way to mark one's social belonging and class status. From the late nineteenth century through the beginning of the twentieth century, the emergence and the spread of the innovations of communication technology, the huge development of urban commercial space and of the mass press, and commercial advertising changed many aspects of the society such as rapid change of consumption behavior and behavior of masses.

In his work, *Les lois de l'imitation* (1890), French sociologist Gabriel de Tarde (1843–1904) theorized the principle of imitation focusing on the characteristics of fashion. According to Tarde, consumption has been viewed as an identity resource of the elite group and as a form of collective behavior at that time. People, in particular social class, imitate fashion, manners, as well as the ideas, wishes, and beliefs of their reference group and that is how fashion spread from the superior to the inferior and from the exterior to the interior.

This idea of imitation by Tarde and the idea of trickle-down theory by German sociologist Georg Simmel (1858–1918) have in common in the sense that consumption behavior transmits and diffuses from one social group, which often is superior, to other groups, which is lower level in many cases. Simmel argues that there are two engines of fashion change, differentiation and imitation, the opposing tendencies toward

conformity and individuality (1904). He theorized fashion as follows: fashion change is based on the tendency of the elite class to differentiate itself from the lower classes, reflecting their conformity; people emulate the consumption behavior of their superiors, that is, more specifically, middle class emulate the upper middle class, who emulate the elite class, in their attempt to achieve upward socioeconomic mobility. Accordingly, trends and innovations form at the top of the social and economic pyramid and trickle downward. At the same time, the elite class moved to adopt a new fashion in attempting to the differentiation. Simmel's analysis stresses the class-based differentiation and downward or 'trickle-down' spread of fashion along the class structure on the presupposition that most consumers of various social groups want the signs of the social elite for themselves.

Both Veblen and Simmel anticipated the birth of a mass consumption society and the escalation of its unfavorable outcome.

Criticism Over Consumer Society and Consumerism during the Postwar Period

Since around the 1950s critical sociologists provided some insight into consumer society and consumerism. *The Affluent Society* (1958) by John Kenneth Galbraith (1908–2006) is perhaps the most influential discussion of consumption in the post–World War II era. Galbraith argues that consumers' desires for goods and services are not intrinsic but that the production side creates consumer desire. This new desire is created by advertising, which adorns new commodities with more expressive and evocative meanings. He called this process the dependence effect, because wants depend on the process by which they are satisfied. He also claimed that resources are overproportionately invested in private production and consumption, and therefore this leads to a deficiency in the provision of public goods in the affluent society.

The Frankfurt School also developed a critical approach to culture and consumption in research into mass society. Max Horkheimer (1895–1973) and Theodor L. Adorno (1903–69) systematically analyze and criticize the industrialization of mass-produced culture around the 1920s and the 1930s in the United States. According to Adorno's and Horkheimer's arguments, in the article The Culture Industry: Enlightenment as Mass Deception (1944), culture industries, coined by them, played an important role in controlling over the whole society and individuals' personality structure became integrated into the needs of mass consumption economy. They presented their insight into the rise of consumer society and American capitalism well before the extensive critical arguments by postmodernists such as Baudrillard.

Meanwhile, in *The Lonely Crowd* (1950), David Riesman (1909–2002), who identified the social character types, claimed in times when consumption overtook production that members of a society became more other-directed. Under those circumstances people sought to become accepted into the mainstream by conforming to the expectations and preferences of peer groups. As a result, consumers are inclined to make relatively minor variations in the items to be consumed. Riesman coined the term standard package to signify the process of compulsive acquisition of commodities. He observed that there was a considerable degree of uniformity throughout American society by the effects of other-directed values on American culture. This is contrary to the viewpoint by Baudrillard, whose propositions are characterized as competition and personalization through consumption in the system of signs that differentiate people.

Postmodern Sociological Approach to Consumption Since the Late Twentieth Century Onward

Some sociologists claim that in postmodern consumer societies, consumption patterns no longer act to structure social classes. Consumption has been approached as the semiotic code constituting postmodern itself. French social theorist Jean Baudrillard (1929–2007), one of the leading postmodern social theorists of consumer society, centered his critical analysis on change in consumption in postindustrial societies. According to Baudrillard, "consumption is defined not any longer (1) as a functional practice of objects – possession, etc., or (2) as a mere individual or group prestige function, but (3) as a system of communication and exchange, as a code of signs continually being sent, received, and reinvented – as language" (Baudrillard, 1970/1998). This means that signs are consumed, not the objects. He also sees consumption undertake the role of integrating entire society, just like hierarchical or religious rituals did in primitive societies by training them in the unconscious discipline of a code, and competitive cooperation at the level of that code.

More recently, other postmodernists (e.g., Giddens, Bauman, Beck, and Featherstone) have conceptualized consumption as exemplifying the individualizing tendencies of a modern materialistic society. From this standpoint of individualization argument, they see consumption as central to identity formation by means of the cultivation of lifestyle (Bauman, 1988; Beck, 1992; Giddens, 1991). Consumption represents a principal mechanism through which individualization is expressed.

Theories of Social Distinction: Social Patterning of Cultural Consumption and Social Class

Scholarly work in the field of consumption has long established a link between social class/stratification position and cultural consumption, one of the classical perspectives with a long history. This relationship has been theorized most famously by French sociologist Pierre Bourdieu (1930–2002) in his work *Distinction* (1979/1984), whose theories were developed in the context of concrete empirical analyses of the French society during the 1960s and the 1970s. His analysis and the ones of several subsequent authors (e.g., Peterson, 1992) center on a variety of practices in different domains of consumption, in particular cultural consumption. Bourdieu, whose theoretical framework is identified as conflict theory, asserted that consumption patterns are not based only on individual desires, but on class location and habitus. According to Bourdieu, lifestyle is an expression of class position and that is a weapon in strategies of distinction and hence legitimizes and reproduces the class structure. There is a structural homology

between the space of social positions and the space of lifestyles. Specifically, Bourdieu differentiated highbrow, or legitimate culture, and lowbrow, or popular culture. Lifestyle of the higher status group is characterized by cultural refinement and snobbery, that is, the highbrow taste is a feature of the well-educated elite in the society; they, at the same time, disdain simple working class popular culture. Accordingly, people in higher social position tend to patronize the fine arts and avoid contact with popular entertainment, whereas people in lower position tend to enjoy 'brutish' popular entertainment. This is called the homology hypothesis. These distinctions in cultural preferences and lifestyles maintain social boundaries. In this manner, these differences of cultural practices establish symbolic boundaries between individuals occupying different locations in the class structure. The ruling class uses consumption styles in order to maintain its dominant position.

Such a standpoint for lifestyle as a means of differentiation and social reproduction is normally linked to Bourdieu, but Max Weber already discussed the importance of lifestyle for differentiation. Status, one of the principal stratification dimensions, is important in establishing class distinction; status groups are sustained by the collectivity of individuals who share similar consumption pattern and lifestyle.

Bourdieu theorized that family socialization process and educational experiences are the primary determinants of cultural capital, which in turn shapes people's tastes and preferences in combination with other forms of capital, or economic capital. Cultural practices and consumption such as patronizing the fine arts, preferences in literature or music, and manners and speech are closely linked to education and cultural capital inherited from one's social origin. Thus dispositions, to which he refers as habitus, that are acquired through the socialization process and produce individual practice and perceptions can be seen as a major predictor of cultural consumption patterns. He claimed "taste classifies, and it classifies the classifier. Social subjects, classified by their classifications, distinguish themselves by the distinctions they make, between the beautiful and the ugly, the distinguished and the vulgar, in which their position in the objective classifications is expressed or betrayed" (1979/1984).

Although Bourdieu's works have been highly praised, his theories attracted criticisms and alternative theoretical approaches emerged, devoting efforts to examining the validity of theories by Bourdieu in different time and space.

For example, since the 1980s some sociologists proclaimed 'death of class', which means that the works of Bourdieu are no longer applicable to modern society and new theories, for example, the individualization thesis, have been proposed instead. For all these arguments, however, the empirical validity of the arguments from postmodern perspective has been questioned due to the lack of compelling empirical data.

Some others claim that it is not necessarily appropriate to apply Bourdieu's theories to the situation in societies other than France. For example, Richard A. Peterson (1932–2010) and his collaborators (e.g., Peterson and Simkus, 1992; Peterson and Kern, 1996) emphasize that in modern societies the cultural taste of the higher status group is inclusive and broad, not snobbish or exclusive so that other nonelite classes are excluded as Bourdieu postulated. This is called the omnivore hypothesis. According to Peterson and Simkus (1992), high-status American people indeed are more likely to consume highbrow culture such as fine arts than lower status people, but only a minority of the higher status people consume any particular highbrow genre. Moreover, high-status people are more likely to enjoy many sorts of cultural activities. They also suggest that a historical shift from highbrow snob to omnivore is taking place. Empirical research that tested the relationship between cultural consumption and social status position in various countries shows that the cultural omnivore hypothesis rather than homology hypothesis by Bourdieu is currently an increasingly influential theory on the relationship between cultural consumption and social stratification.

Whichever hypotheses are taken, it is assumed that differential patterns in cultural consumption among individuals are related to their social stratification position. Substantial empirical studies have evidenced the existence of differences in cultural consumption between the higher status group and the lower status group in many societies in Europe, North America, and Australia.

Michèle Lamont, one of the sociologists who has been influenced by and builds upon the work of Bourdieu, also raised questions about applying Bourdieu's theories to the context of a different class structure on the grounds that these theories ignore national repertoires. She refers to the difference in lifestyle between France and the United States and shows that American upper middle class men make less use of high culture than their French counterparts do when constructing social boundaries and that both sets of people use several valued forms of culture (1992). She argued that highly educated people in the United States have cultural sophistication and also include mainstream culture, while in Europe there is a preference to remain culturally exclusive. There are more works on sociology of consumption from the point of view that difference in the mode of consumption reflects the difference in social groups such as gender, ethnicity, and culture in contemporary societies.

Downshifting, Sustainable Consumption, and Green Consumption

In theoretical discussion on sociology of consumption since Baudrillard, mainstream perspective has been the one similar to Bourdieu's, which understands consumption styles as related to the positions in social hierarchy and class dynamics. These ideas are succeeded by Juliet Schor who has developed the consideration on consumption in relation to both individuals' social position and working style, and the ideas about the people who have voluntarily changed their consumption behavior styles. Schor, whose research has been based on the analysis of the 1980s and the 1990s, argued (1992, 1998) that American people are increasingly involved in a vicious circle of overwork and overspend. This phenomenon has occurred because they keep upscaling their lifestyle norms, or competitive consumption. As has frequently been mentioned, identity is deeply bound with consumption and therefore individuals try to keep up with the group that they identify a reference group. Reference group in America today, however, is not neighbors or social group similar to themselves but the

lifestyles of the rich and famous have become a reference group for all people, due to, for example, decline of neighborhood contacts, less time with friends and family and instead more at work and in front of the television, as well as more married women entering the workforce particularly in white-collar jobs and being exposed to a more economically diverse group of people. At the same time, however, she introduces recent emergence of people who downshift: They downshift from their currently unsustainable high levels of consumption to simplified lifestyles, which means consuming less stuff voluntarily so as to put an end to the vicious circle of work and consumption.

As various strains brought on by consumption come to the fore, especially since the late twentieth and in early twenty-first centuries, an increasing number of concerns on the social consequences of consumption, such as global resource depletion and deterioration of environment, have been expressed in various researches on social and behavioral sciences. Their interests are why and how we are able to think about consumption behavior and lifestyles in relation to the issue of sustainability. From this standpoint, researchers raise awareness of the consequences of one's action, or consumption behavior, and advocate purchasing more green products and adopting alternative and more sustainable practices. Such concepts as sustainable consumption, green consumption, and ethical consumption become increasingly widespread in research on consumption in recent years mostly in Europe.

See also: Bourdieu, Pierre (1930–2002); Class: Social; Conspicuous consumption and distinction, history of; Cultural Capital and Education; Ethical Consumers; Fashion, Sociology of; Frankfurt School: Institute for Social Research; Leisure and Cultural Consumption: US Perspective; Riesman, David (1909–2002); Simmel, Georg (1858–1918); Social Inequality in Cultural Consumption Patterns; Sustainable Lifestyles; Sustainable Production and Consumption; Symbolic Boundaries; Veblen, Thorstein (1857–1929).

Bibliography

Baudrillard, J., 1970. La Societe de Consommation: Ses Mythes, Ses Structures, Editions Denoel. (The Consumer Society: Myths and Structures. Sage, London, 1998).
Bauman, Z., 1988. Freedom. Open University Press, Milton Keynes.
Beck, U., 1992. Risk Society: Towards a new modernity. Sage, London.
Bourdieu, P., 1979. La Distinction: Critique Sociale du Jugement. Minuit, Paris (Distinction: A social Critique of the Judgement of Taste. Harvard University Press, Cambridge, MA, 1984).
Cohen, M.J., Murphy, J., 2001. Exploring Sustainable Consumption: Environmental Policy and the Social Sciences. Pergamon, Amsterdam.
DiMaggio, P., 1987. Classification in art. American Sociological Review 52 (4), 440–455.
Featherstone, M., 1991. Consumer Culture and Postmodernism. Sage, London.
Galbraith, J.K., 1958. The Affluent Society. Houghton Mifflin, Boston.
Giddens, A., 1991. Modernity and Self-Identity: Self and Society in the Late Modern Age. Polity Press, Cambridge.
Holt, D.B., 1997. Poststructuralist lifestyle analysis: conceptualizing the social patterning of consumption in postmodernity. Journal of Consumer Research 23 (4), 326–350.
Ilmonen, K., 2011. A Social and Economic Theory of Consumption. Palgrave Macmillan, Basingstoke.
Lamont, M., 1992. Money, Morals, and Manners: The Culture of the French and American Upper-Middle Class. University Chicago Press, Chicago.
Lamont, M., Lareau, A., 1988. Cultural capital: allusions, gaps and glissandos in recent theoretical developments. Sociological Theory 6, 153–168.
Lamont, M., Fournier, M., 1992. Cultivating Differences. University of Chicago Press, Chicago.
Lamont, M., Molnar, V., 2001. How Blacks use consumption to shape their collective identity: evidence from marketing specialists. Journal of Consumer Culture 1 (1), 31–45.
McCracken, G., 1988. Culture and Consumption: New Approaches to the Symbolic Character of Consumer Goods and Activities. Indiana University Press, Bloomington.
Peterson, R.A., Kern, R.M., 1996. Changing highbrow taste: from snob to omnivore. American Sociological Review 61, 900–907.
Peterson, R.A., Simkus, A., 1992. How musical taste groups mark occupational status groups. In: Lamont, M., Fournier, M. (Eds.), Cultivating Differences. University of Chicago Press, Chicago.
Riesman, D., 1950. The Lonely Crowd: A Study of the Changing American Character. Yale University Press, New Haven.
Schor, J.B., 1992. The Overworked American: The Unexpected Decline of Leisure. Basic Books, New York.
Schor, J.B., 1998. The Overspent American: Why We Want What We Don't Need. Basic Books, New York.
Simmel, G., 1904. Fashion. International Quarterly 10, 130–155.
Simmel, G., 1930. Philosophie des Geldes, Funfte, unveranderte Auflage. Verlag von Duncker & Humblot, Munich and Leipzig.
Tarde, Jean-Gabriel, 1890. Les lois de l'imitation: Etude sociologique. Fe'lix Alcan, Paris.
Veblen, T., 1899. The Theory of the Leisure Class. Macmillan, New York.
Warde, A., 1997. Consumption, Food and Taste. Sage, London.

Contemporary History

Pat Thane, King's College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C.S. Maier, volume 4, pp. 2690–2697, © 2001, Elsevier Ltd.

Abstract

Contemporary history ends with the present. Its starting point is contentious and depends upon the topic of study. It has been practised since ancient times but has become most popular since the 1960s. It has been criticized by conventional historians for the difficulty of examining the recent past objectively and the absence of adequate documentation. These arguments are rejected since they apply equally to all periods. Like the study of other periods, it has been influenced by changes in the contemporary world and by the changing interests of historians such as the growing influence of social and cultural history since the 1960s.

What and When Is Contemporary History?

Periodization in history is always problematic and contentious, no less for contemporary history than for any time period. Unavoidably, historians of all times and places find it hard to agree on precisely when the period they identify with begins and ends. This is partly because historians have diverse preoccupations, necessarily if we are to gain a reasonably comprehensive understanding of the past, and because nations and societies do not experience all-pervasive breaks, even following a serious crisis such as war, revolution, occupation, or plague. Some features of the society may continue, such as religious faith or family structure, which will preoccupy some historians, while others focus on the experience of rupture.

However, one essential element of periodization is agreed among contemporary historians: their period ends with the present. It includes the history of the very recent past, so the end point of contemporary history is agreed but is constantly shifting. Where it begins is more contentious. Some have argued, in the past, that contemporary history is the history of the generation now living, or events within the lifetime of the historian. This will not do because generations overlap and some lives are very long. This definition leaves us with "ever-changing boundaries and an ever-changing content, with a subject matter that is in constant flux" as Geoffrey Barraclough wrote in his *Introduction to Contemporary History* (1966: pp. 13–14), the first modern work in English to tackle the subject. As he pointed out, at the time he was writing (the book was first published in 1964), it was still possible to meet people who had conversed with Bismarck, and others "for whom Hitler was as much an historical figure as Napoleon or Julius Caesar" (p. 13).

Barraclough proposed two complementary definitions of contemporary history. Firstly, it was the history of a time period: he argued that contemporary history should be dated from the late 1950s, when an array of major historical changes were clearly established, including the emergence of the United States and Russia as 'superpowers' and the weakening of Western Europe; the breakdown of empires; 'the readjustment of relations between white and colored peoples'; and the thermonuclear revolution and other technological transformations. 'Taken together' he argued "they give contemporary history a distinctive quality which marks it off from the preceding period" (p. 17). But, Barraclough argued, to understand the nature and significance of these changes the study of contemporary history should begin around 1890, when 'the forces took shape that have molded the contemporary world.' At this time, among other things, economic exchange became truly global; empires reached a peak; labor movements grew and strengthened and socialism became a strong force; more nations moved toward modern democracy; and European nations adopted alignments that would lead to two world wars.

But still, warned Barraclough, "we shall do well to beware of precise dates. *Contemporary History begins when the problems in the world today first take visible shape*" (my italics: p. 20). He recognized that, in order to understand certain contemporary phenomena, a much longer timescale was needed than for others. We might today give the examples of wars in Afghanistan, or of financial crises, understanding both can benefit from historical analysis but the timescales are very different. The second definition of contemporary history, he believed, was its capacity "to clarify the basic structural changes which have shaped the modern world," which "may take us far back into the past." (pp. 16–17). Hence "we cannot say that contemporary history 'begins' comprehensively in 1945, or 1939 or 1917 or 1898 or any other specific date we may choose"(p. 20). The beginning depends upon the aspect of the contemporary world under scrutiny. However, for practical purposes, such as teaching courses in contemporary history, a time frame must be chosen. When the aim is to teach the history of the recent past, or to convey a sense of it to a wide audience, rather than to explain a particular event or set of events, the choice may be somewhat arbitrary, determined by what is manageable in the time available.

Dating key features of the contemporary world to the late 1950s made sense in 1964. Fifty years later we might think differently. Since 1989, Communism has almost vanished from Europe and socialist ideas are more fiercely challenged everywhere than before. The Cold War is over and the Russia is much less influential, and so, more gradually, the influence of the United States has begun to dwindle, while China (still officially Communist) has become increasingly powerful in the world economy and politics. Versions of Islam influence world affairs as they have not for centuries, as do the politics of countries in which it is powerful. The invention of the World Wide Web has revolutionized communications, pervasively influencing everyday life including in entertainment, shopping, and politics. Though these changes need to be seen in context. As David Edgerton has argued, old technology has never ceased

to be important. The world consumed more coal and steam power was more widespread in 2000 than in 1950 or 1900 (Edgerton, 2006). The bicycle is still an important, and in Europe a growing, means of transport, much encouraged as a source of exercise and good health. Similarly, old ideas, sometimes very old such as religious faith, and practices, political, social, and cultural as well as technological, have not disappeared but profoundly influence day-to-day life.

The Long History of Contemporary History

Contemporary history became a notably active preoccupation of historians in the 1960s, hence Barraclough's book, but it was by no means a new concept. As with many contemporary themes we need to probe far back into the past to understand its origins and development. Benedetto Croce claimed in 1917 that "every true history is contemporary history" (Croce, 1917). He meant something different from Barraclough and most contemporary practitioners in the field, that the work of the historian of any period is influenced by present preoccupations and by the historian's personal experience. This is widely agreed. The term has come into common use and into university syllabuses relatively recently, since the 1960s, especially since the 1980s, but it is older than Croce. It made an early English-language appearance in the *Edinburgh Review* in 1808 (vol 12: p. 480) with the (now contentious) comment:

> There is this general distinction between Contemporary History and all other history- that the former is a witness, the latter a judge.

Yet the practice of studying the immediate past, up to the present, is a great deal older still. Thucydides, recognized as one of the greatest ancient historians, writing about the Peloponnesian wars as they happened (431–404 BC) and in which he had fought, rested his narrative "partly on what I saw myself, partly on what others saw for me," aiming to record the war accurately, in view of its intrinsic interest and importance, and to provide political lessons for the future. He hoped that his History would be useful to 'those who desire an exact knowledge of the past as a key to the future, which in all probability will repeat or resemble the past. The work is meant to be a possession forever, not the rhetorical triumph of an hour.' His ruling principle was strict adherence to carefully verified facts:

> As to the deeds done in the war, I have not thought myself at liberty to record them on hearsay from the first informant or on arbitrary conjecture. My account rests either on personal knowledge or on the closest possible scrutiny of each statement made by others. The process of research was laborious because conflicting accounts were given by those who had witnessed the several events, as partiality swayed or memory served them.

Problems that present day contemporary historians will recognize and which have been analyzed extensively by practitioners of oral history (Perks and Thomson, 1998).

Through the next 2000 years others wrote the history of their own times, designed to convey lessons as much as to expand knowledge. In England, Edward Hyde, Earl of Clarendon, adviser to Charles I, and Lord Chancellor to Charles II, late in life wrote *The History of the Rebellion and Civil Wars in England* (1702–1704), which he had experienced and with which he had no sympathy, a work designed to justify and praise the restoration of the monarchy. Thomas Babington Macaulay, politician and imperial administrator, part of a newly influential professional middle class, wrote his widely read four volume *History of England* (published 1848–1861, the final volume after his death), as uncritical celebration of the period since the Glorious Revolution of 1688. He described it as a time of brilliant progress, unique to England at the time, culminating in the Great Reform Act of 1832, which brought the middle classes into political power, and the repeal of the Corn Laws in 1846, which established free trade and further economic expansion, making England the dominant world economy. Democracy, he argued, had then gone far enough and acceding to the growing demands of the lower orders for inclusion could only undermine England's greatness.

Interest in studying the immediate past has often been sparked by a specific crisis – in Macaulay's case, the stirrings of working class discontent in the early nineteenth century – and the desire to understand its origins and impact. For some European historians in the nineteenth century, the French Revolution of 1789 was such an event, though historians of different political persuasions appropriated it in different ways. For conservatives, the French defeat at the hands of Prussia in 1870 demonstrated that the Revolution of 1789 had undermined national cohesion. Hippolyte Taine's *Origines de la France Contemporaine* (1875–93) explained what he believed to be the debility of contemporary French society as a legacy of the Revolution. The *Société d' Histoire Contemporaine*, established in 1890 by Catholic conservatives, espoused such views, while the Left developed its own version of recent history, celebrating the revolution. Emile Bourgeois (1886) rejected Taine's interpretation while accepting that his work was no less valid than writing the history of earlier periods. The Republicans created a Chair in Modern and Contemporary History in Paris in 1884 and, once established in the University, historians of the Left founded the *Revue d'Histoire Moderne et Contemporaine*, also in 1890. The two factions fought for control of the official history texts taught in schools, which included, from the Left, Ernest Lavisse's 10-volume *Histoire de la France Contemporaine de la Révolution à 1919*, one of several texts he produced for schools in the 1880s and 1890s (Noiriel, 1998).

As History became professionalized as a university discipline from the later nineteenth century it defined itself strictly as a study of the 'past,' of times beyond the memory of the living. It sought clearly to differentiate itself, its subject matter and its methods from the other emerging disciplines of the day, as universities themselves emerged in their modern forms and grew in number and diversity of areas of study. The past was the domain of the historian, the present and recent past that of the emerging disciplines of politics and economics, later also of the newer fields of anthropology and sociology. Professional history's source materials were, above all, documentary archives, the remnants of the past. This was despite the fact that one of the dominant figures in the nineteenth-century

professionalization of history, Leopold von Ranke, himself lectured on events in German history occurring during his lifetime, including the revolution of 1848 and the founding of the German Empire in 1871, as did a number of his contemporaries.

The highly committed, politicized nature of much that calls itself contemporary history, as in nineteenth-century France, probably hardened resistance to it among the majority of the earlier generations of professional historians. They believed that the practice of history required total emotional detachment from the subject in hand, that it was necessary to establish 'the facts' wholly objectively, regarding this as an essential dimension of their professional identity, which they were convinced was unattainable when studying the times the historian had personally experienced. Ranke urged the need to extinguish oneself (*sein Selbst ausloeschen*), all of one's values, in order to engage in reliable historical research.

For even more historians in the twentieth century, it was the Nazi regime and the horrors that resulted in an apparently civilized continent that required understanding. In the early 1920s, Dr Alfred Wiener, a German Jew, was horrified by the rise of antisemitism, campaigned against it and collected documents about it. These were lost when he fled Germany in 1933, first for the Netherlands, then for London, but he carried on collecting. After the war, the collection became the Wiener Library in London, still a preeminent archive of Nazism and antisemitism, which assisted the prosecutors at the Nuremberg Trials, amassed early survivor testimony and helped shape the academic study of the Holocaust which emerged after the war. It later also took the name Institute for Contemporary History (www.wienerlibrary.co.uk).

Also in 1949, the Federal Republic of Germany, soon after its own establishment, set up the Deutsches Institut für Geschichte der nationalsozialistischen Zeit in Munich for, as its name suggests, study of the Nazi period. In 1952, it was renamed Institut für Zeitgeschichte and gradually broadened its interests in recent German history (www.ifz-muenchen.de). Meanwhile in France, in 1944, shortly after the Liberation, de Gaulle's provisional government established a commission on the history of the occupation, resistance, and liberation, which, in 1951, became the Committee on the History of World War II. In 1978, this became, and remains, L'Institut d'Histoire du Temps Présent, with a wider historical remit (www.ihtp.cnrs.fr). After the downfall of communism in Europe, institutes were established in postcommunist countries to examine their recent histories, including the Czech Republic, Romania, and Slovenia. The Slovenian Institute, in Ljubljana, was established in 1959 as the Institute for the History of the Workers' Movement. It included an archive of the wartime occupation and resistance movement. By the 1970s, it was studying the history of Slovenia throughout the twentieth century. After the end of communism in 1989, it persuaded the authorities to rename it the Institute of Contemporary History and it gradually broadened its interests to cover economic and social history (www.inz.si/history.php). The histories of these institutions in themselves tell us something about contemporary history.

The origins of Nazism and the war, and later of Communism and its effects, had deep and long roots and the problem of where to begin to study such major themes remained, as Barraclough realized when writing his first book, *The Origins of Modern Germany*, published in 1946. This covered the period 800–1939, with a concluding chapter, "Germany yesterday, today and tomorrow." He started to write it during the war, while serving in the UK Royal Air Force, following 15 years studying German history before the war. He brought it as close to the present as he could at that time.

As the practice of contemporary history grew, most strongly from the 1960s, 1945 became a favored starting point, probably because contemporary history, like most professional history, tended at this time overwhelmingly to focus on political history, national and international. And post-1945 politics – the world of the Cold War and decolonization – seemed very different from what had gone before and to require new explanations and forms of analysis. But many historians recognized that analysis of the post-1945 world could not begin in 1945 if it was truly to be understood. When what became the leading international journal in the field, the *Journal of Contemporary History* – its founding itself a sign of the substantial new interest – was established in 1966, shortly after Barraclough published his book, it set out to study mainly European history throughout the twentieth century. Its founding editors were George Mosse, a German-born, Jewish American who had published widely on Nazism, and Walter Laqueur, also American and Jewish, who published on the history of fascism and communism and was then director of the Wiener Library and Institute of Contemporary History in London and professor at Georgetown University in the United States. Their own professional preoccupations did much to shape the journal in its early days. They acknowledged in their opening editorial that their definition of the field 'is bound to be to a certain extent arbitrary both in time and space.' They were 'aware that the first world war (not to mention more recent events) cannot really be discussed without reference to trends that go well back into the nineteenth century.' They also recognized that Europe did not exist in isolation from the rest of the world. They pledged to 'try to be sufficiently elastic in their editorial policy to accommodate contributions that at first sight seem to transcend the Journal's frame of reference' both in time and space. So the journal has done and its subject matter has reflected changing preoccupations in the broad field of contemporary history.

Contemporary History since the 1960s

The longer time span has become more commonplace in European university courses and journals since the 1960s, above all because of the evident difficulty of explaining important features of the world since 1945 without looking at least back to the later nineteenth century. Also the generation was emerging for whom 'Hitler was as much an historical figure as Julius Caesar' and they had wider interests, influenced by the changing world around them. It was impossible, for example, to understand the then rapid process of decolonization, for example, without a sense of how Empires had come about and their effects; or why West European countries felt a need to develop 'welfare states' after the war without understanding the extent of poverty and inequality past and present; or why, from the 1960s, women, excluded racial groups and homosexuals,

among others, were demanding equality so fiercely in so many countries, without awareness of the long history of the inequalities they had experienced.

Also, from the 1960s, the range of professional history writing widened and subspecialisms grew, challenging the dominance of political history. This was less wholly new than some thought at the time. In reality, political history had not held exclusive sway for some time. Economic history had begun to develop in the interwar years, closely linked with what came to be called social history. In 1926, R.H. Tawney, in *Religion and the Rise of Capitalism*, traced the origins of modern capitalism (of whose unequal outcomes he was a critic, actively supporting and advising the British Labor Party) back to early modern times and stressed the formative influence of protestant religion, combining cultural and economic history. Eileen Power, one of the first female professors of Economic History, at the London School of Economics, and a feminist, wrote and broadcast in the 1920s on the new medium of radio, about medieval women and other aspects of medieval social and economic history (Berg, 1996).

In France, the journal *Annales d'histoire économique et sociale*, was founded in 1929 as a direct challenge to the dominance of political history in France. It aimed to integrate the study of society, economy, politics, intellectual life, geography, and demography, ideally over long time periods up to the present, in order to understand the complex interactions that constitute a society, regarding the present as history as well as the distant past and skeptical about the existence of radical breaks in history. The *annalistes*, a self-conscious 'school,' drew on the growing insights and methods of the social sciences to enrich historical research.

In Britain, G.M. Trevelyan's *British Social History: a survey of six centuries, Chaucer to Queen Victoria*, written before the second world war in Britain, published in the United States in 1942 (perhaps with the original intention of informing a nation hesitating to enter the war on Britain's side, before it was impelled to do so by the Japanese attack on Pearl Harbor in December 1941). Trevelyan argued for the importance of social history not, as he is constantly misquoted, as 'history with the politics left out,' but as "the required link between economic and political history … without social history, economic history is barren and political history is unintelligible" (1942/1967: pp. 9–10). Trevelyan was concerned with a very contemporary question, the characteristics of the British culture that the nation was fighting to defend in the second world war, recognizing the need to trace its construction over many centuries, examining as many dimensions as possible: "the daily life of the inhabitants … the human as well as the economic relation of different classes to one another, the character of family and household life, the conditions of labour and of leisure, the attitude of man to nature, the culture of each age … took ever-changing forms in religion, literature and music, architecture, learning and thought" (1942/1967: pp. 9–10). He aimed originally to bring the book up to the time of writing, but "the war has rendered it impossible" (p. 9) though he hoped it would still serve its contemporary purpose.

These historians, and others in Britain and elsewhere, challenged the dominance of political history before the 1960s (Floud and Thane, 2005), but social history became much more prominent from the late 1960s. It was partly impelled by very contemporary concerns, of a newly active Radical Left, often influenced by Marxism, challenging continuing social and economic inequalities, allied with emerging antiracist, feminist, and gay campaigning groups. These movements influenced and extended the work of professional historians, some of whom were activists themselves, into topics, such as the histories of slavery and colonialism and their legacies, of gender inequalities, and of repression of homosexuality, for which the conventional political boundaries, such as the dates of wars, were largely irrelevant. It was also influenced by the gradual entry of rather more women and members of other excluded groups into what had previously been an overwhelmingly white, male historical, indeed wider university, profession.

Such histories challenged and extended conventional narratives about the contemporary world. For example, for some time, influential Anglo-American sociologists had explained the characteristics of the family in the contemporary 'developed' world in terms of the increasing isolation of the 'nuclear' core of parents and children from wider kin, as industrialization and migration disrupted 'traditional' agrarian society, which was supposedly static and composed of coresiding, coworking family groups, whereas, they believed, fast-moving modern society divided the generations geographically and emotionally, creating, it was said, unprecedented problems of isolation for older people and less grandparental care for the young. This tidy scheme owed nothing to historical research. From 1964, the demographic, previously intellectual, historian, Peter Laslett, in Cambridge, drew on the methods of the *Annales* school and the novel, if still cumbersome, aid of computers to explore the history of the family in England using parish records of births, marriages, and deaths back to their origins at the Reformation and the formation of the Church of England in the sixteenth century. New technology assisted new approaches to the history of the recent and the more distant past. Laslett and his colleagues in the Cambridge Group for the History of Population and Social Structure, which he founded, examined the actual history of English family structures and asked whether change had indeed taken the form that the sociologists claimed. They discovered a preindustrial society in England that was far from static, in which migration over long distances in search of work was common, large coresiding kin groups uncommon and nuclear families as normal as in the twentieth century, while higher death rates and lower life expectancy at all ages often left children without grandparents and older people without surviving children. Family structures had changed much less than had been believed. Findings that were later replicated for other parts of Northwestern Europe (Laslett and Wall, 1972; Wrigley and Schofield, 1989). Study of a key social institution over several centuries deepened understanding of its contemporary characteristics, leading to recognition that intergenerational ties were at least as strong, and possibly stronger, than in the preindustrial past in the contemporary world where life expectancy was longer and modern technology assisted communication between kin over long distances in unprecedented ways. They demonstrated that studying the history of social as well as political institutions could extend knowledge and understanding of contemporary issues.

Flexibility over the appropriate time period to study continues to characterize contemporary history. The journal

Contemporary British History, founded in 1986, focuses on the period since World War I, yet in a special issue on health inequalities in the later twentieth century the editors thought it helpful to publish an article on dimensions of inequality, 1700–2000 (*Contemporary British History*, Autumn 2002). Similarly, for reasons of manageability in the time available, contemporary history courses in European universities mainly focus on the period from the early twentieth century, examining still earlier periods when the topic requires it.

Approaches, Sources, and Methods in Contemporary History

Contemporary history may be pursued out of simple curiosity to explore and understand the recent past in all its dimensions, or in order to understand the origins and effects of a particular event or set of events or processes, sometimes for overtly political reasons such as to influence policy decisions, though often not. As the practice of contemporary history grew after 1945, it was much criticized on the grounds familiar since the nineteenth century, that the objectivity required of the professional historian was impossible toward events in the historian's own lifetime, which might have touched her or his own existence. This implied that distance in time necessarily created an objectivity that was impossible in relation to recent events. This overlooked the passionate partisanship professional historians could display over many other themes, such as long past events like the English Civil War of the seventeenth century; the deep repugnance of many of them toward slavery and long past repressive acts by imperial powers; and partisan conflict over the impact of industrialization on working-class living standards such as that waged between Marxists, who took a doom-laded view, and their more optimistic opponents in the late 1950s and early 1960s. (Hobsbawm, 1964; Hartwell, 1961).

The first editors of the *Journal of Contemporary History* countered with the argument that 'Distance in time … involves remoteness, lack of immediacy, difficulty in understanding the quality of life of a period that is hard to describe and define, but which may become as important as all the documents in the archive.' Indeed there were those who argued, in relation to World War II, that only those who had been closely involved, as victims of perpetrators of the horrors of the Nazi concentration camps or Japanese prisoner of war camps, could truly understand all dimensions of these experiences. These were not new observations. Alexis de Tocqueville, reflecting on the French Revolution commented that what contemporary writers:

> know better than does posterity are the movements of opinion, the popular inclinations of their times, the vibrations of which they can still sense in their minds and hearts. The true traits of the principal persons and of their relationships, of the movements of the masses are often better described by witnesses than recorded by posterity. These are necessary details. Those close to them are better placed to trace the general history, the general causes, the grand movements of events, the spiritual currents which men who are further removed may no longer find since these things cannot be perceived from the memoirs. (Quoted in Schlesinger, 1967)

Or as George Santayana later put it:

> It is not true that contemporaries misjudge a man. Competent contemporaries judge him … much better than posterity, which is composed of critics no less egotistical, and obliged to rely exclusively on documents easily misinterpreted. (Quoted in Schlesinger, 1967)

Also, for historians to believe that objective study of the contemporary world was impossible was, somewhat arrogantly, to dismiss the work of social scientists, who engaged in just that, and to refuse to recognize that historians could learn from their methods and experience. Indeed, the growth of the social sciences after World War II encouraged the spread of contemporary history, by making more widely accessible techniques for analysis of the recent past, previously neglected or scorned by historians, such as interviews, the use of contemporary documents such as letters, sources such as newspapers and such new resources as film and sound recordings. Contemporary history was frequently criticized by university historians on account of the limited availability for the recent past of the historical documents conventionally used by political historians. Political history relied predominantly upon government archives, which in most countries were closed for at least 40 or 50 years (in the case of the most sensitive for much longer or forever), as were many personal archives of prominent people, including monarchs and presidents. Since such sources were not accessible for the period of closure, it followed that, for conventional historians, contemporary history was impossible.

Nonetheless, the first editors of the *Journal of Contemporary History* were optimistic that

> there are few major secrets in modern democracies that can be kept for longer than a year or two; the time lag in dictatorships is unfortunately considerably longer, but it seems doubtful whether anything really important can be kept secret for very long even there. In any case, more often than not, the contemporary historian of Europe or America in the sixties is more likely to be confronted with a surfeit of source material rather than to suffer from a lack of documentary evidence. (JCH: p. iii.)

Indeed, extensive government archives were themselves a relatively recent phenomenon, following the growth of large and effective bureaucracies in more developed countries in the later nineteenth century. For previous periods, documentary sources were often patchy, selective, and biased toward an elite, male minority whose papers survived and therefore might dictate possibly selective readings of the past. This view also overlooked the selective nature of government archives even in the more recent past. There was, and is, no means of knowing which sensitive papers might have been destroyed or withheld. Only in 2011 was it discovered that the British government withheld from public access, and stored secretly in the British countryside, 8000 shelves of documents concerning the British army's maltreatment of insurgents during the independence struggles in Kenya in the 1950s. Under the then-prevailing 30-year rule, they should have been released to the National Archives in the 1980s. They came to light only when some

aging survivors of British brutality sued the British government, successfully, in 2011, and because UK Freedom of Information (FOI) legislation, passed only in 2000, permits requests for access to hidden documents (though these are not always agreed readily or at all). The furore that resulted when the existence of the documents was reluctantly revealed, was supported by many historians. This led the UK Foreign and Commonwealth Office, equally reluctantly, to reveal, in 2013, that it had a further 1.2 million files hidden in another country store, some going back at least to the Treaty of Paris of 1856 which ended the Crimean War, but too sensitive to release for over 150 years. It later became known that substantial numbers of documents had been destroyed, on the instruction of government ministers, to prevent material that 'might embarrass government … police, military forces, public servants or others' being revealed to postindependence governments (The National Archives, November 2013; *The Guardian*, 29 November 2013).

Such suppression by governments of shameful episodes in history, leading to pressure from historians for greater openness has not occurred only in Britain. To give just one example, Japanese governments have been very reluctant to respond to historians' demands to inform schoolchildren about Japanese atrocities during World War II, especially in China. Furthermore, the museum commemorating the atomic devastation of Nagasaki in 1945, providing what appears to be the official view, gives notably little indication of Japanese war guilt, in contrast to the museum in Nanjing, China, which commemorates the well-attested Japanese massacre of that city in 1937.

All such revelations justify the suspicions of skeptical historians that we cannot rely on official archives to reveal all the truth about history, that historians of all periods must be wide ranging and inventive in their search for sources and be well aware that all historical findings, about any period, up to the present, are provisional and can be overturned by discovery of new sources and new research at any moment. The belief that historical research, in any time period, seeks objective, factual truth rather than being part of a process of gradually, partially, piecing together an understanding of the past, can no longer be sustained and very few historians would try. This applies as much to historians of the contemporary world as to all others.

Despite closures, over time more official documents have been opened at earlier stages than before, facilitating contemporary history. Franklin D. Roosevelt set an example in the United States by decreeing that his personal papers should be donated to the national archive after his death and be made quickly accessible. His successors followed his example, establishing Presidential archives that are no longer closed for a generation. Since the death of John F. Kennedy, Presidential archives have included interviews with prominent figures and observers of the period of office of the President in question.

FOI, the right of historians and others (often journalists or campaigners) to request access to documents falling within the designated period of closure has spread over a similar period. It was introduced by law in the United States in 1966. But it is only a partial blessing to historians. Requests can be refused and government departments have various means of evasion: in Britain, a request can be refused if it is deemed too costly to make the required search, as it frequently is. In 2013, the Ministry of Justice, which administers FOI, further refused to publish statistics of successful appeals against refusal on the grounds that this also would be too costly (Hansard, *House of Lords*, 6 November 2013). Furthermore, there is the very real danger that, as historians in Sweden have put it, FOI leads to 'the emptying of the archive,' to politicians and civil servants committing fewer sensitive issues to paper for fear of exposure (Kandiah, 2008). Sweden was the first country in the world to guarantee public access to government documents in its constitution of 1766. But, even in Sweden, legislators can deny access to specific documents. There is a right of appeal, but appeals are not invariably successful. Of course historians have never had access to past face-to-face conversations about government business, which might be crucially important, or, since the invention of the telephone, to communication by that means. Newer technology is creating further obstacles. Students of history since the 1990s may find surviving official sources dwindling ever faster as communications by e-mail and other forms of new technology are preserved imperfectly, if at all.

Such problems have made contemporary historians inventive in their use of sources. They have been assisted by the availability of a wider range of sources and helped as well as hindered by new technology. Oral history came widely into use as a direct response to the growth of social history in the 1970s, with its growing interest in the history of the everyday lives of 'ordinary people' and awareness that their experiences and opinions rarely surfaced in official documents. A pioneer was the American Studs Terkel (a popular, not a university, historian, among his many other talents) who used oral history to explore social inequalities in the United States, published as *Division Street America* (1966), *Hard Times. An Oral History of the Great Depression* (1970) and *Working* (1974). The journal *Oral History* started in Britain in 1972. Oral history drew on the experience of sociologists and anthropologists for interviewing techniques, greatly assisted by new technology, as increasingly portable recording machines became available from the 1960s. Again, it faced the skepticism of conventional historians that individuals were less likely to tell 'the truth' in an interview than in a written document. This overstated the reliability of many documents as well as the gullibility of historians, who learned to interpret interviews as skeptically as any document and, as with documents, to seek corroborating sources for any assertion, while recognizing that the story presented by an interviewee represented one version of the events in question, which required careful interpretation. (Perks and Thomson, 1998).

The recent past has created ever newer media representing aspects of the contemporary world – film, television, sound recording – which demand new skills for their interpretation, as do the study of the press, memoirs, letters, diaries, and other personal documents, which it has become more common to use by political as well as social historians. New media will present new challenges to future contemporary historians, and others as they recede into a new past. Contemporary history must continue to be inventive, as it has always been while keeping up with an ever-changing present.

See also: Data Bases and Statistical Systems: Archives and Historical Databases; Economic History; Family and Kinship, History of; Gender History; Historical Thought and Historiography: Christian Tradition; Historical Thought and Historiography: Current Trends; Historical Thought and Historiography: Modern History (Since the 18th Century) – The west; History and the Social Sciences; History from Below, the History of Everyday Life, and Microhistory; History, Professionalization of; History: Overview; History: Theories and Methods; Oral History; Political History: History of Politics; Social History; Time, Chronology, and Periodization in History.

Bibliography

Barraclough, G., 1966. An Introduction to Contemporary History. Penguin, Harmondsworth, England.

Barraclough, G., 1946. Origins of Modern Germany. Blackwell, Oxford.

Berg, M., 1996. A Woman in History. Eileen Power 1889–1940. Cambridge University Press, Cambridge.

Bourgeois, E., 1886. L'histoire contemporaine et la science de l'histoire. Armand Colin, Paris.

Croce, B., 1917. Teoria e Storia della Storiografia. Laterza, Bari, Italy.

Croce, B., 1808. Edinburgh Review 12.

Edgerton, D., 2006. The Shock of the Old. Technology and Global History since 1900. Profile, London.

Floud, R., Thane, P., 2005. Sociology and history: partnership, rivalry or mutual incomprehension? In: Halsey, A.H., Runciman, W.G. (Eds.), British Sociology. Seen from without and within. Oxford University Press, Oxford, pp. 57–69.

Hansard, 6 November 2013. Parliamentary Debates, House of Lords. www.publications.parliament.uk/pa/ld201314/Since the death of.

Hartwell, R.M., 1961. The rising standard of living in England, 1800–1859. Economic History Review XIII, 400–422.

Hobsbawm, E.J., 1964. 'The British Standard of Living, 1790–1850' Labouring Men. Weidenfeld and Nicolson, London, pp. 64–104.

Hyde, Edward, 1702–1704. The History of the Rebellion and Civil Wars in England.

Kandiah, M.D., 2008. 'Contemporary History' in 'Making History' series published on website of the Institute of Historical Research. University of London. http://www.history.ac.uk/makinghistory/resources/artciles/contemporary_history.html (accessed 28.10.2013.).

Laslett, P., Wall, R. (Eds.), 1972. Household and Family in Past Time. Cambridge University Press, Cambridge.

Macaulay, Thomas Babington, 1848–1861. History of England, 4 vols.

Noiriel, G., 1998. Qu'est-ce Que L'histoire Contemporaine? Hachette, Paris.

Perks, R., Thomson, A. (Eds.), 1998. The Oral History Reader. Routledge, London.

Schlesinger, A., 1967. 'On the Writing of Contemporary History' the Atlantic Magazine. www.theatlantic.com/magazine/archive/1967/03.

Terkel, Studs, 1966. Division Street America. The New Press, New York.

Terkel, Studs, 1970. Hard Times. An Oral History of the Great Depression. The New Press, New York.

Terkel, Studs, 1974. Working. The New Press, New York.

The Guardian, 29 November 2013. Revealed: The Bonfire of Papers at the End of Empire. www.theguardian.com/uk-news.

The National Archives, 2013. www.nationalarchives.gov.uk/archive-records#colonial-administration-files.

Trevelyan, G.M., 1942/1967. English Social History. A Survey of Six Centuries, Chaucer to Queen Victoria. Longmans, Green; Harmondsworth: Penguin, US and Canada.

Wrigley, E.W., Schofield, R.W., 1989. The Population History of England 1541–1871: A Reconstruction. Cambridge University Press, Cambridge.

Content Analysis

Carl W Roberts, Iowa State University, Ames, IA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article's emphasis is on methods that content analysts have developed to draw inferences about large corpuses, or populations, of texts. After having acknowledged that language is not a neutral medium through which 'content' is unambiguously transmitted, researchers must make four key decisions: Are words only to be counted or are word relations to be encoded? Are word relations to be depicted as network characteristics or as variants of a semantic grammar? Do researchers presume they know more or less than the texts' sources regarding the latter's words? Are the texts under analysis being viewed as windows into sources' perspectives or, alternatively, into events experienced by sources? Only once these decisions are made, can one identify the types of research questions afforded by content analysis.

Introduction

Content analysis is a class of techniques for mapping symbolic data into a data matrix suitable for statistical analysis. These techniques may be applied to any representative sample of cultural artifacts (e.g., books, paintings, technological innovations, etc.), whereby various nonnumeric attributes of these artifacts are mapped into a matrix of statistically manipulable symbols. Thus, content analysis involves measurement, not 'analysis' in the usual sense of the word. Whereas one might perform a content analysis of campaign slogans, one uses the data generated in such a content analysis to analyze (i.e., not 'content' analyze) the slogans' effects on public opinion. Texts and transcripts have been by far the most common objects of content analysis. For this reason, this article will focus on the content analysis of texts (i.e., on quantitative text analysis). Discussions of qualitative text analysis techniques are available elsewhere (*see* Archival Methods; Ethnography; Event History Analysis: Applications; Linguistic Turn and Discourse Analysis in History; Qualitative Methods, History of).

Classical (or Thematic) Content Analysis

The first large-scale applications of content analysis came with World War II and the US government commissioned propaganda analyses performed by Lasswell, Berelson, George, and other content analysis pioneers. During the first decade of the postwar era, a classical approach to content analysis emerged that involved the generation of data matrices of a particular type. These matrices typically exhibit a row for each sampled text block and a column for each theme (or meaning category) that may or may not have occurred in each block. Cells in the matrix are counts of the number of times the theme associated with its column occurred in the text block associated with its row. Thus, classical content analysis – later called *thematic text analysis* – typically produces matrices of word or phrase counts, albeit sometimes ones embellished with value and salience weights (cf. Stone et al., 1966; Holsti, 1969; Smith, 2008).

During a conference held in 1955 at the University of Illinois-Monticello's Allerton House, many of these text analysis pioneers gathered to develop solutions to the methodological problems of the day. Pool (1959) was the scholarly legacy of this conference. The most influential, if not the largest, faction among the participants was a group of Harvard researchers who made extensive use of a thematic text analysis technique that they called 'contingency analysis.' After generating a matrix of theme occurrences, these researchers would compute a matrix of pairwise associations among the themes. Contingency analysis then proceeded as the researcher developed (usually post hoc) explanations of why some theme pairs co-occurred (i.e., were positively associated) and why others were dis-associated (i.e., were negatively associated).

In his contribution to the conference's proceedings, Osgood (1959) raised critical questions about this burgeoning contingency analysis technique. Surely one may legitimately draw inferences from thematic text analyses about 'what' texts' sources were writing or speaking about. Yet contingency analysts were going beyond this by making inferences about word relations without having encoded these relations in the first place – a flagrant commission of the ecological fallacy (*see* Ecological Fallacy, Statistics of) by inferring words' clause-level relations from their aggregated associations among text blocks. In Osgood's words, "The final stage, in which the analyst interprets the contingency structure is entirely subjective, of course" (1959, p. 76). Although there may be relatively little ambiguity in *what* words are used in texts, even honest researchers are likely to disagree in their subjective interpretations of *how* these words were used. Moreover, such disagreements may stem from some researchers' presumption that a text's source knows better than they do what the source's words mean versus others' presumption that their interpretation "is beyond the voluntary control of the source and hence yields valid inferences despite the strategies of the source" (Osgood, 1959, p. 75). In sum, the content analyst should consider not only how texts' words are related (preferably by eliminating post hoc subjective inferences and encoding both words and their relations before statistical analysis), but also whether the text's source is an unwitting novice or the ultimate authority regarding its meaning. If only for these reasons, it is meaningless to allude to 'the content' of a text without having first specified which 'word usages' and whose understanding – researcher's or source's – of these usages are deemed meaningful.

The next two sections of this article review recent developments in text analysts' methods for addressing each of these problems (i.e., the problem of measuring relations among

words and the problem of presuming researcher's versus source's interpretive expertise). The following section introduces the ongoing issue of whether a text population is being used to draw inferences either about sources' perspectives or about historical facts. The essay's conclusion then sketches out the domain of research questions that content analysts can address.

Theme-Relations within Networks versus Grammars

Every statistical analysis is based on a representative sample of phenomena. In text analyses, these phenomena are blocks of text (books, editorials, responses to open-ended questions, and so on) that together make up a well-defined text population (e.g., all health-related editorials published in the *New York Times* between 1965 and 2012). In a thematic text analysis, the words in these text blocks are merely counted. Now imagine that you wish to analyze not only 'what' but also 'how' words are used in a specific text population. Although impractical, one such relational approach to content analysis would be to map all grammatical relations among all words in each text block. Thereafter, one might even map relations among sentences – an activity referred to in the structuralist literature as the identification of components in a 'story grammar.' (In a sample of debates, e.g., such components might consist of problem, argument, rebuttal, reformulation, and solution.) Yet even if one could overcome the logistic challenges required in accomplishing a total mapping of all words and relations in a text block, there would still remain the *statistical* challenge of 'flattening' this network of word- and sentence-relations into a matrix of rows and columns.

Relational approaches to text analysis virtually always involve defining which word-relations are to be encoded. For example, one might only consider subject–verb–object (SVO) relations in the texts, thereby eliminating 'extraneous' parts (e.g., prepositional phrases) of each text block in the analysis. Thus, one already discerns an important difference between relational and thematic text analysis methods in that words eliminated in the former would have been counted in the latter.

After transforming text blocks from collections of words to collections of tuples (here, of SVO-relations), we need to rethink the form of our data matrix. Recall that in the absence of such transformation, thematic text analysts use 'flat' two-dimensional data matrices in which each row (or unit of analysis) corresponds to a text block and each column (or variable) corresponds to a theme. Relational text analysts have developed two ways to flatten a collection of tuples into a matrix of rows and columns: Instead of having each column represent a type of theme, each of the columns represents a possible tuple (e.g., a unique SVO theme-combination); and instead of having each row correspond to a specific text block, each of the rows references an empirical tuple in the text. The former yields a network analysis in which tuple frequencies are variables, and the latter yields a semantic text analysis in which tuples are a unit of analysis.

A *network text analysis* can now be defined as a content analysis method in which each text block is characterized in terms of sets of relations among its themes. For example, let us imagine that we encode only causal relations in a sample of text blocks, ignoring all SVO tuples in which the verb (V) conveys something other than 'facilitates,' 'causes,' 'influences,' and the like. From each text block, we can then construct a network of themes in which all linkages indicate causal relations (*see* Models of Influence in Networks). We might then decide to measure a text block's 'causal salience' of theme A on theme B as the proportion of all sequences of causal linkages in the text block that are ones in which the former is the cause and the latter is the effect.

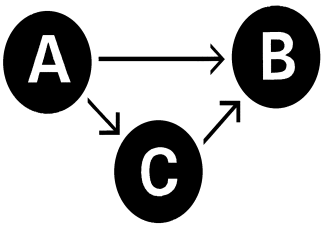

Figure 1 A network of causal relations among themes.

The simple three-theme network in Figure 1 is the graphic representation of causal relations in a text block that contains causal tuples consisting exhaustively of the same number of references to A causing B, A causing C, and C causing B. Note that the figure depicts four sequences of causal linkages, namely, A→B, A→C, C→B, and A→C→B. The causal salience of theme A on theme B thus equals .5, which is the proportion of the four sequences in which A is the cause and B is the effect. Rather than merely affording information on isolated tuples in the text block, it is worth noting that this measure incorporates information on all themes and links within the text block's network representation.

Table 1 illustrates a data matrix generated from networks – each corresponding to a randomly sampled text block – that exclusively contains causal salience measures on each pair of the texts' three themes. Many other measures of network characteristics have been developed. A theme's 'conductivity' is one example, referring to the number of linkages that the theme provides between pairs of other themes (Carley, 1997). Another measure is of theme linkages that are logically implied, but not explicitly stated in each block of text (Kleinnijenhuis et al., 1997). In each case, entire text blocks are assigned global measures that summarily reference all instances of a particular type of word relation.

A *semantic text analysis* is one in which every empirical tuple throughout the text that exhibits a specific type of semantic relation is characterized in terms of the words (or, more generally, themes) used in the respective parts of the relation. Semantics, not syntax, is referred to in this context because word relations with purely grammatical (i.e., syntactic) similarities may have quite disparate meanings (i.e., semantics). For instance, "Mary left an impression on Joe" may – depending on

Table 1 A data matrix from a network text analysis

	Causal salience measures					
ID number	*A on B*	*A on C*	*C on B*	*B on A*	*C on A*	*B on C*
1	.50	.25	.25	.00	.00	.00
2	.25	.00	.50	.00	.25	.00
3	.00	.00	.25	.25	.50	.00
4	.00	.00	.00	.50	.25	.25
5	.00	.25	.00	.50	.00	.25
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2 A data matrix from a semantic text analysis

ID number	*Subject*	*Action*	*Object*
1	10	34	53
2	8	21	78
3	13	34	61
4	10	35	61
5	16	29	44
⋮	⋮	⋮	⋮

the context – semantically reference a lingering thought or a painful indentation. Because content analysts generally are interested in words' meanings and not their superficial grammatical relations, each semantic text analysis begins with the construction of a template of semantic relations (also known as a semantic grammar). In the encoding phase, themes from sampled texts are mapped as components within this template. An example of this is Shapiro's and Markoff (1998) two-place semantic grammar for 'grievances,' which contains one syntactic component for the object of the grievance (i.e., what is being grieved about) and another for the action that should be taken toward this grievance.

Table 2 illustrates a data matrix that might have been generated in a semantic text analysis. Note that the cells in the matrix do not contain indicators of theme or tuple occurrences, but rather they contain discrete codes for the themes. The column in which a specific theme's code appears indicates the theme's role within the researcher's semantic grammar. In generating Table 2, clauses from a text population were encoded as sequences of subject–action–object triplets. Thus, for example, if 10 were the code for 'management official(s),' 34 the code for 'organize or arrange,' and 53 the code for 'conference or meeting,' the matrix's first row might be an encoding of the sentence, "Corporate leaders organized a meeting with union representatives for next week." Inferences from such a data matrix might be made to a population of newspaper accounts of labor disputes regarding the odds that representatives of management versus representatives of labor initiated collective bargaining. On the basis of semantically encoded transcriptions of speech from a random sample of minutes of prime-time television, inferences might be drawn from a similar matrix regarding the odds that blacks versus whites refer to themselves as objects of aggressive actions. More generally, semantic text analyses yield information on the frequency with which specific theme relations occur within the context of an a priori specified semantic grammar.

Researcher as Novice versus Expert

The American National Election Studies periodically includes a set of open-ended questions like the following:

> Now we have a set of questions concerning various public figures. We want to see how much information about them gets out to the public from television, newspapers and the like…. What about… John Roberts – What job or political office does he NOW hold?

As one of a series of 'political knowledge questions,' the correct (i.e., knowledgeable) answer to this question is "Chief Justice of the U.S. Supreme Court." Yet imagine that during an interview the response to this question is, "a job that he has performed worse at than any of his predecessors." At this point, a trained interviewer would recognize that the respondent has given not an incorrect but an inappropriate response to the question. In particular, the respondent mistakenly understood his role relative to the interviewer as one of expert regarding his own ideas rather than as novice regarding publicly available political knowledge. The point here is that interviewer and respondent must hold a consensus on who is expert and who novice regarding each survey question, leaving it the responsibility of the former to correct the latter when responses show these roles to have been reversed inappropriately.

Such 'correction activity' is impossible during secondary analyses of large (unresponsive) text populations, of course. If content analysts are to be consistent in their application of codes for themes and relations, they must self-consciously apply them in accordance with their status as novice or expert. For instance, when studying cultures, anthropologists typically adopt the role of 'humble novice,' thereby approaching documents and other materials with disinterested (i.e., non-prejudicial) interest in the meanings they imagine their sources to have intended. Archival texts themselves then become the ultimate touchstone in understanding why their authors thought and acted as they did. Yet the sorts of conclusions drawn by such a content analyst would undoubtedly differ from one who argued that the texts' authors "are unfamiliar with many fundamental properties of their own culture and prove unable to specify its structural rules…. (T)o recover culture's properties and rules, we cannot ask culture's participants to answer these questions. Instead, we must rely on outsiders as investigators and use their methods, however unreliable these may prove to be" (Namenwirth and Weber, 1987, p. 237). In sum, content analysts' stance as expert or novice respectively corresponds to their use of text analysis methods either to identify individual and societal characteristics about which society's members may be unaware, or to characterize texts in ways that their sources intended them to be understood (Shapiro, 1997).

To illustrate the disparate conclusions made by as-novice versus as-expert researchers, consider a small subpopulation of male novelists who write books only with elderly female heroines. In seeking to understand their choices of leading characters, the researcher-as-novice might scour the novelists' prefaces, possibly discovering references there to a need for elderly female heroines as a corrective to an overuse of young male heroes in contemporary literature. In contrast, a Freudian researcher-as-expert likely would find the choice of matronly heroines to be symptomatic of Oedipus complexes among the novelists. In the former case, the novelists' own meanings are assigned to their writings, whereas in the latter case, the researcher's (Freudian) perspective is applied.

Regardless of researchers' stance as novice or expert, they are always responsible for developing an explicit method to systematically apply the perspectives of sources or researcher in interpreting the texts under analysis. Methodological challenges related to as-novice text analysis are those associated with the researcher's ability to sympathetically understand the sources of his texts (*see* Ethnography; Qualitative Methods, History of). When researchers are capable of clearly articulating

their expertise, this latter approach has the advantage that it may be possible to fully automate the encoding process with text analysis software (e.g., Gottschalk et al., 2002). Software aides for the researcher-as-novice typically afford users with key-word-in-context (KWIC, or concordance) information so that themes and relations can be assigned with broader contextual information on sources' intended meanings (cf. the freeware, textual content analysis or TCA, referenced at the end of this article).

Inferences to Events versus Perspectives

In his grammatology, Jacques Derrida used the metaphor of a hinge to emphasize the 'dual life' of linguistic expressions. On the one hand, things people say or write are propositions that in varying degrees have a trace in the audience's sensory experiences. For example, a mother may find no trace of a 'freedom fighter' in the soldier who refers to her dead child as collateral damage. On the other hand, peoples' words are linked into a grammar. Thus, "gay marriage is wrong because (according to my sacred language's uncontestable grammar) marriage is between a man and a woman." Texts' dual life creates a methodological dilemma for the content analyst: If one is interested in using sources' narratives as windows into the *historical events* they experienced, then their perspectives are likely to distort these sources' rendering of the events. For example, members of conflicting groups may have noticed only their own but not other groups' attempts at reconciliation, making it difficult to discern from their writings which attempts had in fact been made. Yet if one wishes to investigate differences among *sources' perspectives*, then their responses to historical events may be mistaken for changes in perspective. Accordingly, autonomous competitors within an impersonal marketplace may temporarily 'become' patriotic members of a national civil society in reaction to a terrorist event. In brief, inferences from texts to historical events will be compromised to the extent that sources' perspectives are confounded with their renderings of facts, and inferences from texts to sources' perspectives will be compromised to the extent that historical events are confounded with the perspectives they adopt (succinctly put, perspectives constitute 'noise' when inferring events and events constitute 'noise' when inferring perspectives).

One way of managing this 'noise' is to include it as a measured component of the research design. When establishing facticity, researchers commonly attenuate the effects of sources' perspectives by comparing (triangulating [*see* Triangulation: Methodology]) their inferences with other event-related records. Events themselves may be used as contextual variables between which perspectives may be compared. Yet despite such strategies, the methodological challenge remains that content analysts must decide not only whether they are researching events or perspectives, but also how they are to attenuate the noise introduced by the 'other life' of the linguistic expressions being analyzed. The next two paragraphs illustrate relational text analyses in which inferences are respectively (and exclusively) drawn about events and perspectives.

Conflict resolution between nations is most effective when 'sticks' (i.e., conflictual actions) are directed toward both antagonistic parties at the same time that 'carrots' (i.e., cooperative actions) are directed toward the weaker antagonist. Schrodt and Gerner (2004) provide evidence in support of this proposition based on event data gleaned from Reuters News Service lead sentences in coverage of the Israel–Lebanon and Israel–Palestinian (1979–99), plus Serbia–Croatia and Serbia–Bosnia (1991–99) conflicts. In their analysis, frequencies of cooperative behavior between a mediator (United States, United Nations, or European Union) and either antagonist, and conflictual behavior either between antagonists or between mediator and antagonist were compared over time. In a time-series analysis, decreased frequency in reports of inter-antagonist conflict is associated with prior instances of the sticks and carrots. In this way, a content analysis of texts from a news service affords data on the timing and substance of political events at the close of the twentieth century.

When pressed for clarity on what is meant by 'socially constructed reality,' sociologists typically assert that such construction is accomplished rhetorically (i.e., via persuasive communications). Yet surely an argument that is persuasive in one culture might not make sense (i.e., lack rhetorical purchase) in another. For example, Roberts et al. (2010) provided evidence of statistically significant differences between Arab and Hindi cultures in how 'possibilities' are depicted in Saudi and Indian editorialists' rhetorical use of modal auxiliary verbs, such as 'want' and 'can.' Whereas Arabs tend to justify their wants, hopes, and attempts in normative ways, Hindi authors write about what they 'can' do in accordance with their natures – rather than their desires – as required by the circumstances at hand. This content analysis affords not only data on perspectives (not events) but also findings consistent with two distinct ways of socially constructing personhood, namely, by rhetorically characterizing people either as willfully striving to emulate quranic norms or as unwilled performances enabled by their respective natures.

What Questions Can Content Analyses Answer about Text Populations?

No discussion of content analysis would be complete without an emphasis on the fundamental importance of contextual variables (i.e., variables not obtained directly from the texts but variables indicating other characteristics uniquely associated with each text block under analysis). Notice how text-related variables become interesting only when compared between historical periods (of mediation vs. conflict events), cultures (Arab vs. Hindi), media (television vs. Internet), sources (male vs. female), and the like. Contextual variables allow the researcher to test hypotheses (or speculate exploratively) about why words were used and related as they are in a text population.

Beyond which contextual variables to consider, content analysts must make equally important decisions about the texts themselves. Because every content analysis of texts involves a researcher in the 'interrogation' of a text population, it is imperative that researchers spell out precisely what questions their encoding process will answer about the texts. First, is the analysis only to report which themes were discussed, or is information on the texts' word relations to be reported as well? Second, if word relations (or tuples) are of interest, are the tuples to be

treated jointly as networks of text block characteristics or as discrete instances of a semantic grammar? Third, are researcher's or sources' meanings to be assigned to the texts? And, finally, are the texts to be used as 'windows' either into events experienced by the texts' sources or (exclusively) into the perspectives sources brought to these experiences? Responses to these questions yield 12 distinct approaches to text analysis (thematic/network/semantic × as-expert/as-novice × events/perspectives), resulting in a vast array of possible content analysis techniques and applications (Roberts, 1997; Krippendorff and Bock, 2008). Yet the appropriate method for individual research is greatly simplified once one stops asking, "How are texts analyzed?" and instead addresses the four questions discussed in this article. After setting the parameters of content analysis in this way, guiding the analysis's substantive direction is where the researcher's own imagination takes hold.

See also: Archival Methods; Ecological Fallacy, Statistics of; Ethnography; Event History Analysis: Applications; Interpretive Methods: Macromethods; Interpretive Methods: Micromethods; Linguistic Turn and Discourse Analysis in History; Models of Influence in Networks; Qualitative Methods, History of; Symbolic Interaction: Methodology; Triangulation: Methodology.

Bibliography

Carley, K.M., 1997. Network text analysis: the network position of concepts. In: Roberts, C.W. (Ed.), Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts. Lawrence Erlbaum, Mahwah, NJ, pp. 79–100.

Gottschalk, L.A., Defrancisco, D., Bechtel, R.J., 2002. Computerized content analysis of some adolescent writings of Napoleon Bonaparte: a test of the validity of the method. Journal of Nervous and Mental Disease 190, 542–548.

Holsti, Ole R., 1969. Content Analysis for the Social Sciences and Humanities. Addison-Wesley, Reading, MA.

Kleinnijenhuis, J., de Ridder, J.A., Rietberg, E.M., 1997. Reasoning in economic discourse: an application of the network approach to the Dutch press. In: Roberts, C.W. (Ed.), Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts. Lawrence Erlbaum, Mahwah, NJ, pp. 191–207.

Krippendorff, K., Bock, M.A. (Eds.), 2008. The Content Analysis Reader. Sage Publications, Thousand Oaks, CA.

Namenwirth, J.Z., Weber, R.P., 1987. Dynamics of Culture. Allen & Unwin, Winchester, MA.

Osgood, C.E., 1959. The representational model and relevant research methods. In: Pool, Ithiel de Sola (Ed.), Trends in Content Analysis. University of Illinois Press, Champaign, IL, pp. 33–88.

Pool, Ithiel de Sola (Ed.), 1959. Trends in Content Analysis. University of Illinois Press, Champaign, IL.

Roberts, C.W. (Ed.), 1997. Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts. Lawrence Erlbaum, Mahwah, NJ.

Roberts, C.W., Zuell, C., Landmann, J., Wang, Y., 2010. Modality analysis: a semantic grammar for imputations of intentionality in texts. Quality & Quantity 44, 239–257.

Schrodt, P.A., Gerner, D.J., 2004. An event data analysis of third-party mediation in the Middle East and Balkans. The Journal of Conflict Resolution 48, 310–330.

Shapiro, G., 1997. The future of coders: human judgments in a world of sophisticated software. In: Roberts, C.W. (Ed.), Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts. Lawrence Erlbaum, Mahwah, NJ, pp. 225–238.

Shapiro, G., Markoff, J., 1998. Revolutionary Demands: A Content Analysis of the Cahiers de Doléances of 1789. Stanford University Press, Stanford, CA.

Smith, C.P., 2008. Motivation and Personality: Handbook of Thematic Content Analysis. Cambridge University Press, New York.

Stone, P.J., Dunphy, D.C., Smith, M.S., Ogilvie, D.M., 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press, Cambridge, MA.

Relevant Website

http://www.stat.iastate.edu/tca/ – Textual Content Analysis.

Content Validity

Linda Crocker, University of Florida, Gainesville, FL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 4, pp. 2702–2705, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article reviews the process of obtaining logical and empirical evidence by evaluating test content in relation to a defined domain of knowledge or performance. After considering some historical background, it explains the steps required for content validation: domain definition, judgment process, selecting and training judges, and data analysis from content judgment studies. It concludes by asserting the evolving standards for content validation.

Content validation is the process of obtaining evidence to support test use in the form of expert judgments about test content in relation to a defined domain of knowledge or performance (Krippendorff, 2013). Developers of educational achievement tests, personnel tests, and tests of licensure and certification often rely on evidence from content judgment studies. Definitions of content validity have appeared in measurement literature since the 1950s, but some theorists have questioned whether it qualified as a form of validation. Content judgment studies involve defining domain, identifying features of items to be reviewed, structuring the review process, selecting and training qualified experts, and collecting and summarizing the data from their judgments. The judgment process focuses typically upon relevance of item content to the domain, balance of coverage or representativeness, and technical quality of items. Ratings of frequency and criticality are often used to operationalize the concept of relevance. A variety of methods have been used for analyzing data from content judgment studies. Standards for evaluating the psychometric quality of content judgment studies are traced through five editions of the *Standards for Educational and Psychological Tests*, 1954–99.

Validation is a process of disciplined inquiry. In the social sciences, it entails the systematic assembly of logical and empirical evidence to support the intended use of the test scores. Content validation is the process of obtaining such evidence in the form of expert judges' evaluation of test content in relation to a defined domain of knowledge or universe of behavior. Suppose an examinee takes a test of foreign language proficiency and must translate sentences and identify the meaning of vocabulary words. The test user wants to draw inference from performance on the test items at hand to a broader universe of behavior represented by these tasks. Practical constraints of testing time, examinee fatigue, examinee motivation, and cost of item development limit the number of items on the test. Thus, these items and response formats must represent a hypothetical pool of tasks that could have been presented covering the same content. Validity arguments for this inference hinge largely on whether a clear domain of language proficiency was defined and how adequately this domain was sampled by the exercises.

Using expert judges to review the domain description and evaluate specific features of the items in relation to this domain, as well as collecting and summarizing their judgments, is known as content validation. Content validation evidence is useful in (1) assessment of educational achievement, (2) assessment of job knowledge or skills by employers, and (3) professional certification and licensure testing. It may also play a role in establishing construct validity of instruments for assessing affective traits.

History and Terminology

Attempts to define content validity were prominent in the measurement literature in the mid-twentieth century (Ebel, 1956; Gulliksen, 1950; Lennon, 1956). From the 1950s to 1980s, measurement experts (e.g., Cronbach, 1971) identified three distinct approaches to validation: (1) content validation, (2) criterion-related validation, both predictive and concurrent, and (3) construct validation. Throughout this period, however, a few theorists continued to challenge whether content judgment studies fit within the definition of validation (e.g., Fitzpatrick, 1983; Messick, 1975), but this view was not dominant.

Originally, content validation was applied primarily in educational achievement testing. Clearly, the usefulness of test scores as indicators of student learning depended upon the linkage between items presented in the test and the body of knowledge that the items sampled. The 1970s witnessed a growing trend for personnel psychologists to employ content validation to justify the use of test results in personnel selection and advancement decisions, despite some notable reservations (Ebel, 1977; Fitzpatrick, 1983; Guion, 1977). This application grew in inverse proportion to waning confidence in the adequacy and legal defensibility of other validation procedures for many jobs and employment settings. Licensing and certification testing also relied on content validation throughout this period (Shimburg, 1982).

Approaching the twenty-first century, a more unitary view of validity prevailed (Messick, 1989). One theme of this perspective is that proficiency in an academic subject or job performance domain should be viewed as a construct. Although not new, the idea gained mainstream acceptance. From this viewpoint, all validation efforts are designed to provide evidence of construct validity, but there are multiple aspects to the process of construct validation (and multiple types of evidence). Content is consistently the first aspect and

involves determination of the construct relevance, representativeness, and technical quality of the test using expert judgment (Messick, 1995; García-Valderrama and Mulero-Mendigorri, 2005).

Other terms are associated with content validity, but not synonymous. 'Face validity' is the extent to which items appear relevant to a trait in the eyes of laymen or examinees who are not necessarily experts in the subject tested (Mosier, 1947). 'Instructional validity' and 'curricular validity' were spawned in the mid-1970s during landmark litigation over the use of high-school competency examinations in Florida (see Yalow and Popham, 1983). Curricular validity refers to congruence between item content on an achievement and curricular objectives and textual materials. Instructional validity refers to congruence between item content and instruction from teachers.

Design of a Content Judgment Study

Content judgment studies typically are undertaken by test developers as part of their initial validation effort for a new test of knowledge or proficiency. Test users may also conduct local content validation prior to the decision to adopt an existing, published test. These studies may require replication over time to demonstrate that changes in curricula or job requirements have not undermined the meaning of the test scores. A content judgment study involves systematic collection of observations. Steps in the process may include:

1. Specifying the questions to be addressed
2. Defining the domain of interest
3. Identifying the features of the exercises to be examined in the judgment process
4. Developing a structured framework for the review process
5. Selecting qualified experts for that domain
6. Instructing the experts in the review process
7. Collecting and summarizing the data

Domain Definition

With achievement tests, domain definition typically is extracted from instructional objectives commonly used by those who teach the subjects; review of curricular materials and published textbooks also may play a central role in domain definition. Interviews or surveys of instructors have also been used in this process (see Yalow and Popham, 1983). Test specifications (blueprints), used by test developers to maintain the desired proportional coverage of various content areas and cognitive performance levels, may also be used in domain definition. In personnel or licensure testing, domains, or required or prerequisite knowledge, may be defined through a detailed job description provided by experts or a job analysis based on detailed observations of individuals performing the job (Kane, 1997); however, in licensure testing the domain is restricted typically to skills and knowledge that are essential to public safety (Smith and Hambleton, 1990). A 'critical incidents' technique may also be employed to capture important knowledge or skills that spell the difference between successful and failing job performance. Thorough domain definition also includes the format(s) in which individuals may be expected to display their knowledge: that is, it covers both a stimulus and a response domain (Guion, 1977).

The Judgment Process

The judgment process focuses upon at least three features of the test items: (1) relevance of item content to the domain; (2) balance of coverage or the extent to which the collection of items adequately sample the content domain; and (3) technical quality of the items, response formats, and scoring procedures (Messick, 1989, 1995).

Relevance of item content is assessed by evaluating the importance of the knowledge or skills required to answer the test item in relation to successful performance in the domain. Judges may be asked to rate frequency and criticality of the skill or knowledge tapped by the exercise. When the domain is complex, the study may involve having a sample of experts (1) evaluate the appropriateness of the objectives or list of behaviors that comprise the domain and (2) match exercises to those objectives.

Representativeness (balance of coverage) is ascertained by having the judges review the entire test to identify areas of the domain that may be over- or underrepresented. It is possible for all items on a test to be relevant when reviewed individually but yet for the coverage of the domain to be disproportionately limited to only a few objectives or skills. A mathematics test, for example, might stress computational skills but place little emphasis on problem-solving, even though both skills are stressed in the curriculum. The proportions of items matched to each objective or job behavior are examined to assess appropriateness of coverage. The format for examinee response to exercises may also be considered in evaluating representativeness. In judging technical quality of exercises, experts attend to features of the question, response format, scoring rubrics or keys, and administration and scoring instructions that might interfere with examinees' ability to display their knowledge of the correct response.

Selecting and Training Judges

Judges' qualifications and the overall representative nature of the panel of judges in terms of demographic characteristics and educational/professional backgrounds are important issues in the design of a credible content validation study. The instructions given to the judges and training they receive before undertaking their review should also be planned and documented. Finally, in large-scale content validations, level of agreement across the panel is useful information.

Analyzing Data from Content Judgment Studies

Thorough content validation studies yield substantial amounts of data. Each item should have a record containing multiple judges' ratings for each feature of the item that was reviewed.

Marshalling these data for analyses to summarize results at the individual item level and for the total test requires forethought. At the simplest level, indices such as percentage of items classified to objectives may be computed and reported. Judges' mean ratings for each item and over the test as a whole on frequency or criticality may be computed. More sophisticated treatments of data from content validation studies also have been suggested. Crocker et al. (1989) described various quantitative indices used in content judgment studies, including measures of (1) overall fit between test and curriculum, (2) degree of fit for individual items to content domain, (3) fit between test specifications and examinee performance, and (4) error estimation in judges' ratings. In addition, new technical approaches to analyzing job analysis data are emerging (e.g., Kane, 1997; Sireci and Geisinger, 1995; Penfield and Miller, 2004; Rourke and Anderson, 2004).

Evolving Standards for Content Validation

Standards for evaluating test development and use have been articulated since the 1950s by a series of joint committees of the American Association of Educational Research, the American Psychological Association, and the National Council on Measurement in Education. These standards have received global recognition. In the first published committee report, each recommendation was labeled as ESSENTIAL, VERY DESIRABLE, or DESIRABLE (APA, 1954: p. 20). The primary recommendation for content validation was as follows:

> If a test performance is to be interpreted as a sample of performance in some university of situations, the manual should indicate clearly what universe is represented and how adequate the sampling is.
> ESSENTIAL

Additional recommendations included that the descriptions be provided for sources of items, criteria for inclusion or exclusion of items, and method of sampling items from the domain. A decade later, the *Standards* (APA, 1966) specified that content validity information should include the following:

1. The universe that the test items represented (ESSENTIAL)
2. The adequacy of the procedure for selecting a sample of items from the universe (ESSENTIAL)
3. The dates of publication of any textbooks from which items' topics were selected (ESSENTIAL)
4. The professional qualifications of the expert judges (VERY DESIRABLE)
5. The directions that they received (VERY DESIRABLE)
6. The extent of agreement among judges (DESIRABLE)
7. Any classification or taxonomy used in item selection (DESIRABLE)

The next edition of the joint standards (APA, 1974) generally upheld the previous standards for content validity but recognized increasing applications of content validation in personnel selection:

> When a test is represented as having content validity for a job or class of jobs, the evidence of validity should include a complete description of job duties including relative frequency, importance, and skill level of such duties.
> ESSENTIAL. (p. 46)

In 1985, the joint *Standards* downplayed the notion of 'content validity' as a separate test quality but recognized the uniqueness of its form of evidence (AERA et al., 1985). In addition, these standards covered domain definition and content validation for licensure testing. The final edition published in the twentieth century (AERA et al., 1999) includes two standards for validity evidence from content judgment studies. First,

> When the validation rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified in reference to the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should be clearly explained and justified.
> (p. 18)

Second,

> When a validation rests in part on the opinions or decision of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications, and experience, of the judges should be presented. The description of the procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth.
> (p. 19)

Shifts in wording and focus of criteria for evaluating content judgment studies reflect trends in the measurement community's view of validity per se. Yet, given advances in the theory, quantitative models, and technology of modern psychometrics, it is remarkable that the underlying logic, methods, and criteria for content judgment studies have survived with so little change. Evidently, the concept possesses enduring utility for a broad range of test developers and test users.

See also: Content Analysis; Educational Evaluation: Overview.

Bibliography

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), 1985. Standards for Educational and Psychological Testing. American Psychological Association, Washington, DC.

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), 1999. Standards for Educational and Psychological Testing. American Psychological Association, Washington, DC.

American Psychological Association (APA), 1954. Technical recommendations for psychological tests and diagnostic techniques. Psychological Bulletin 5 (2).
American Psychological Association (APA), 1966. Standards for Educational and Psychological Tests and Manuals. American Psychological Association, Washington, DC.
American Psychological Association (APA), American Educational Research Association (AERA), National Council on Measurement in Education (NCME), 1974. Standards for Educational and Psychological Tests. American Psychological Association, Washington, DC.
Crocker, L.M., Miller, D., Franks, E.A., 1989. Quantitative methods for assessing the fit between test and curriculum. Applied Measurement in Education 2, 179–194.
Cronbach, L.J., 1971. Test validation. In: Thorndike, R.L. (Ed.), Educational Measurement, second ed. American Council on Education, Washington, DC, pp. 443–507.
Ebel, R.L., 1956. Obtaining and reporting evidence for content validity. Educational and Psychological Measurement 16, 269–282.
Ebel, R.L., 1977. Comments on some problems of employment testing. Personnel Psychology 30, 55–63.
Fitzpatrick, A.R., 1983. The meaning of content validity. Applied Psychological Measurement 7, 3–13.
García-Valderrama, T., Mulero-Mendigorri, E., 2005. Content validation of a measure of R&D effectiveness. R&D Management 35, 311–331.
Guion, R.M., 1977. Content validity: the source of my discontent. Applied Psychological Measurement 1, 1–10.
Gulliksen, H., 1950. Intrinsic validity. American Psychologist 5, 511–517.
Kane, M., 1997. Model-based practice analysis and test specifications. Applied Measurement in Education 10, 5–18.
Krippendorff, K., 2013. Content Analysis: An Introduction to Its Methodology. SAGE Publications, CA.
Lennon, R.T., 1956. Assumptions underlying the use of content validity. Educational Psychological Measurement 16, 294–304.
Messick, S., 1975. The standard problem: meaning and values in measurement and evaluation. American Psychologist 30, 955–966.
Messick, S., 1989. Validity. In: Linn, R.L. (Ed.), Educational Measurement, third ed. American Council on Education, Washington, DC, pp. 13–103.
Messick, S., 1994. The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher 23 (2), 13–23.
Messick, S., 1995. Standards of validity and the validity of standards in performance assessment. Educational Measurement: Issues and Practice 14 (5), 5–8.
Mosier, C.I., 1947. A critical examination of the concept of face validity. Educational and Psychological Measurement 7, 191–206.
Penfield, R.D., Miller, J.M., 2004. Improving content validation studies using an asymmetric confidence interval for the mean of expert ratings. Applied Measurement in Education 17, 359–370.
Rourke, L., Anderson, T., 2004. Validity in quantitative content analysis. Educational Technology Research and Development 52, 5–18.
Shimburg, B., 1982. Licensing and certification. In: Encyclopedia of Educational Research, fifth ed. American Educational Research Association, Washington, DC, pp. 1084–1092.
Sireci, S.G., Geisinger, K.F., 1995. Using subject matter experts to assess content representation: a MDS analysis. Applied Psychological Measurement 19, 241–255.
Smith, I.L., Hambleton, R.K., 1990. Content validity studies of licensing examinations. Educational Measurement: Issues & Practice 9 (4), 11–14.
Yalow, E.S., Popham, W.J., 1983. Content validity at the crossroads. Educational Researcher 12, 10–14.

Context Dependent Memory

William E Hockley and Tyler D Bancroft, Wilfrid Laurier University, Waterloo, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of the effects of environmental context for animal learning and human memory has a long history with mixed results. This research is reviewed to determine the conditions in which environmental context influences memory performance and the nature of these effects. Both interference and context reinstatement paradigms are considered. Global environmental context has a more reliable effect on free recall performance, whereas more local manipulations of context have more consistent effects on recognition memory. Context-dependent recognition is reliably observed, whereas context-dependent discrimination depends on the encoding of specific item–context associations. The neural bases of context effects are also reviewed.

> An experience can be recalled most readily in those environmental situations with which it has the most direct, the strongest, and most numerous associations. (Carr, 1925: p. 251)

In episodic memory, both the initial encoding and later retrieval of an event occur in contexts. As the above quote aptly illustrates, it has long been assumed that retrieval will be optimal when it occurs in the same context that was present at encoding and can be adversely affected when the context is changed. Assumptions regarding the effects of context on memory are consistent with Tulving's (1983) (Tulving and Thomson, 1973) encoding specificity principle, which states that memory improves to the degree that the information in the retrieval environment matches the information that was encoded at the time of study. Smith (1988) provides illustrative anecdotal examples of past memories that are brought to mind when one returns to a previously familiar place after an absence and notes that context can serve as an organizational cue for information associated with different contextual referents.

The effects of context on memory have been shown in both studies of animal learning (e.g., Carr, 1917; see Spear, 1978 for a review) and human memory (e.g., Dulsky, 1935; for reviews see Smith, 1988 and Smith and Vela, 2001). Context has been broadly defined and has included the physical environment (Godden and Baddeley, 1975; Smith et al., 1978), including factors such as odors (Cann and Ross, 1989; Schab, 1990), background music (Smith, 1985a), body posture (Rand and Wapner, 1967), drug-induced (Eich, 1980) and mood-induced (Bower et al., 1978; Eich and Metcalfe, 1989) states, influences on the semantic interpretation of events, (such as the response 'jam' rather than 'traffic' in response to 'strawberry') (Light and Carter-Sobell, 1970), and even the specific script in which words are typed (Reder et al., 2002).

Considering context in such broad terms has made defining context and distinguishing context from the to-be-remembered target event, somewhat problematic. Researchers have proposed several distinctions concerning the relation between the context and the target information that can influence or moderate the effects of context on memory: (1) context can be global or general and associated with a number of items or be local or specific to one or a few target items (Dalton, 1993; Glenberg, 1979); (2) context can be encoded incidentally or without intention to do so or be encoded intentionally with the target event (Smith and Vela, 2001); and (3) context can be independent from the target items or can be associated or integrated with the to-be-remembered information (Baddeley, 1982; Eich, 1985; Murnane et al., 1999).

This article will focus on the effects of physical or environmental context (EC) on memory because this type of context has played an important role in different theories of memory (e.g., Gillund and Shiffrin, 1984; Humphreys et al., 1989; Tulving, 1983) and in the applied aspects of memory research such as eyewitness memory (e.g., Dando et al., 2009), and because much more research has investigated this aspect of context. Historically, researchers have used rooms with different physical features to evaluate the effects of EC, but more exotic physical environments have also been contrasted, such as underwater versus on land (Godden and Baddeley, 1975) or on a boat (Emmerson, 1986). Early laboratory-based experiments used such improvised manipulations as background color (Dulsky, 1935; Weiss and Margolius, 1954), and having participants stand with their heads in oddly shaped boxes with flashing red and green lights (Dallett and Wilcox, 1968). More recent manipulations on computer screens have involved different colors and locations (Murnane and Phelps, 1993), background pictures of scenes (Hockley, 2008a), scenes viewed through LCD glasses (Hockley et al., 2012), and video recordings of real-life environments (Smith and Manzano, 2010). Somewhat surprisingly, the nature of the background or incidental context manipulation has not greatly influenced the qualitative pattern of the observed effects. The magnitude of the effect of context, however, has tended to be greater for more detailed or engaging manipulations (Hockley et al., 2012). For example, Murnane et al. (1999) showed that context effects for recognition memory were much greater with backgrounds that were perceptually and semantically richer compared with much simpler manipulations of color and location, and Smith and Manzano (2010)

obtained very large effects of context on recall performance using video backgrounds.

Two different experimental procedures have been used to examine context-dependent memory: interference paradigms and what Smith and Vela (2001) termed the reinstatement paradigm.

Interference Paradigms

Early studies of context-dependent memory in humans were carried out to test assumptions of the interference theory of forgetting (Postman and Underwood, 1972). This theory proposes that forgetting is due to proactive interference from prior learning and retroactive interference from later learning. In the interference paradigm, the to-be-remembered material is learned in one context and interfering stimuli are learned in either the same or a different context. As was typical of the period, experiments involved learning and relearning of paired-associate or serial lists of words or nonsense syllables. Although there were some exceptions (Nagge, 1935; Strand, 1970), several studies showed evidence of context-dependent memory: The effects of both proactive interference (Dallett and Wilcox, 1968) and retroactive interference (Bilodeau and Schlosberg, 1951; Eckert et al., 1984; Greenspoon and Ranyard, 1957) were significantly reduced when the interfering material was learned in a different context than the target information. Similar results have been found in nonhuman animal studies. For example, Zental (1970) observed reduced retroactive interference in spatial discrimination in rats by providing different contextual cues during the learning of two tasks.

Another phenomenon related to interference effects in associative learning which demonstrates context-dependent memory and that has been shown for both nonhumans and humans is the 'renewal effect.' This effect occurs when a change in context between extinction and test restores (or renews) the original conditioned response (Bouton and Bolles, 1979; Bouton et al., 1999). In the basic AAA/ABA context procedure demonstrating the renewal effect, subjects first learn a cue–outcome relationship in one physical context (A). This cue–outcome relationship is then extinguished by presenting the cue without the outcome in either the same context (A) or a different context (B). A final test is typically conducted in the original context (A). The renewal effect is seen as a recovery of the cue–outcome association, when the extinction was in a different context (B) compared with the situation when acquisition, learning, and test all took place in the same context (A).

The renewal effect is of particular interest because research has shown that the original learning is not context dependent. That is, the original cue–outcome association is quite robust and can be seen when tested in different contexts prior to extinction. Rather, the renewal effect is due to the context-specific nature of extinction. By way of explanation, it has been suggested that the meaning of the cue becomes ambiguous during extinction, leading subjects to pay more attention to the context (Bouton, 1997). Rosas et al. (2006) further suggested that when subjects pay attention to the context, all learning in that situation is associated with that context.

The Reinstatement Paradigm

The reinstatement paradigm has been, in recent years, the most commonly used method to examine context effects. In this procedure, participants study material in one environment, and evidence for context dependence is observed when memory is greater when the context is reinstated at test, as compared with when it is changed. Undoubtedly, the most cited study using this paradigm in natural environments is Godden and Baddeley (1975). They had divers study and recall lists of words either underwater or on land. Recall was greater when study and test were in the same physical environment and suffered when the contexts were different. The pattern of the interaction demonstrating the global EC context effect for recall found by Godden and Baddeley (1975) and many others is exemplified in **Figure 1**.

In a later study, Godden and Baddeley (1980) failed to find such an EC effect when they tested recognition memory rather than recall. Emmerson (1986), however, did find a modest EC effect for recognition in a replication of Godden and Baddeley's (1980) study. While several studies using different rooms as the context manipulation have found an EC effect for recognition (Dalton, 1993; Geiselman and Bjork, 1980; Geiselman and Glenny, 1977; Smith, 1986; Smith and Vela, 1992), a number of others have not (Fernandez and Glenberg, 1985; Jacoby, 1983; Smith, 1986; Smith et al., 1978). The literature appears to suggest that it is easier to observe EC effects for recall than for recognition, but there have been failures to find EC effects in tests of recall as well (e.g., Eich, 1985; Saufley et al., 1985). Fernandez and Glenberg (1985) reported eight failures to find an EC effect and questioned the reliability and robustness of this effect using standard manipulations of EC. Why would EC effects, particularly in tests of recognition, show such inconsistency? Several explanations have been proposed.

Factors Affecting Context Effects

Overshadowing, outshining, and mental reinstatement have all been suggested as possible reasons to explain failures to find EC effects.

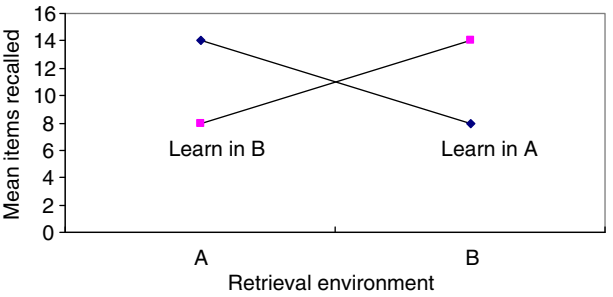

Figure 1 A representation of the interaction between the learning environment and the retrieval environment showing that recall is greater when the study and test environments are the same.

Overshadowing

If the encoding of context information was inhibited or otherwise interfered with at study, no effects of context would be observed. For example, if participants were given the task of making associations between pairs of study items, it might inhibit or greatly reduce the incidental encoding of contextual information or associations between study items with contextual features. As Smith and Vela (2001) noted, associative processing between study items was involved in seven of the eight experiments that did not show CE effects, as reported by Fernandez and Glenberg (1985).

Outshining

According to the outshining hypothesis (Smith, 1988), context effects will be attenuated or eliminated to the extent that effective noncontextual cues are available at test to guide retrieval. In this view, CE effects should be greatest for free recall, reduced for cued recall, and reduced further for recognition, as the cues provided in cued recall and recognition tests will typically be more effective than contextually based cues.

Mental Reinstatement

Research has shown that instructions to imagine or bring back to mind a previous context can improve recall (Dando et al.,2009; Smith, 1979, 1984) and recognition (Krafka and Penrod, 1985; Malpass and Devine, 1981). Such instructions are now part of the 'cognitive interview' that was developed by Fisher and Geiselman (1992) to improve eyewitness memory. If participants being tested in a changed context mentally reinstated the study context, the context manipulation would be compromised and the chances of observing the effects of EC would be substantially reduced.

Smith and Vela's (2001) Meta-Analysis

In a meta-analysis based on 75 experiments from 41 articles that examined incidentally encoded global EC effects in human memory, Smith and Vela (2001) showed that overall EC effects in the reinstatement paradigm were reliable. This analysis also provided evidence that overshadowing and mental reinstatement can attenuate context-dependent memory. The evidence in support of the outshining hypothesis was mixed, however, as the predicted difference in EC effects for free recall and recognition was not found. Averaged over studies, the effect sizes were similar in magnitude and greater than zero for free recall and recognition, with effect sizes falling within Cohen's (1988) small effect range, whereas the effect size associated with cued recall did not differ significantly from zero. Dougal and Rotello (1999) also found no support for another prediction of the outshining hypothesis, as the EC effect in recognition was not reduced when the target item was strengthened.

In response to the apparent inconsistency of EC effects in tests of recognition, there was a shift from using what Fernandez and Glenberg (1985) termed 'standard methods' that were predominantly global in nature to more local-based manipulations of context. The influential work of Murnane and Phelps (1993, 1994, 1995) exemplified this approach and led to the distinction between context-dependent recognition and context-dependent discrimination (Murnane et al., 1999).

Context-Dependent Recognition and Context-Dependent Discrimination

To minimize the possibility (and utility) of mental reinstatement of the study context, researchers used multiple contexts at study and test (Bjork and Richardson-Klavehn, 1989; Smith, 1979, 1985b). As Smith has shown, mental reinstatement becomes more difficult when multiple contexts are used both in tests of recall (Smith, 1979) and recognition (Smith, 1985b). Note that in using multiple study contexts rather than one, the context manipulation changes from global (many items in one context) to local (fewer items in each context) in terms of Dalton's (1993) dichotomy.

Murnane and Phelps (1993, 1994, 1995) presented lists of words in different computer screen locations and in different foreground and background colors. For the recognition tests, items were presented in their original study context or a new context consisting of contextual elements not seen at study. They found that both the hit rate (proportion of correct old responses) and the false alarm rate (proportion of incorrect old responses) were higher for items tested in their old context compared with a new context. The increase in the hit and false alarm rates was similar or concordant, and thus there was no overall difference in discrimination (the ability to distinguish old from new test items) between the reinstated and changed context conditions. Murnane et al. (1999) termed this concordant effect context-dependent recognition. Context-dependent recognition was found when test items were presented in an old context configuration that was paired with a different item at study, that is, a 'rearranged old context' (Murnane and Phelps, 1994), and also when a test item was presented in a new configuration of old context elements (Hockley, 2008a: Exp. 1). These results indicate that the elements of these simple contexts were not bound together and were not specifically associated with the study item. The results of Hockley (2008a: Exp. 2) depicting context-dependent recognition are shown in the top panel of **Figure 2**.

In a different manipulation of context on a computer screen, Murnane et al. (1999) presented the study words on a chalkboard in a classroom scene, on a television in a living room, on a sign on a desert road, and on a banner trailing from an airplane. A different fifth scene was used as the new context condition at test. In contrast to context-dependent recognition found by Murnane and Phelps (1993, 1994, 1995) for simple context backgrounds, Murnane et al. showed that more semantically meaningful contexts resulted in better recognition accuracy for items presented in their original study context at test, a result they termed context-dependent discrimination. The results of Hockley (2008a: Exp. 5B) illustrating the pattern of context-dependent discrimination are shown in the lower panel of **Figure 2**.

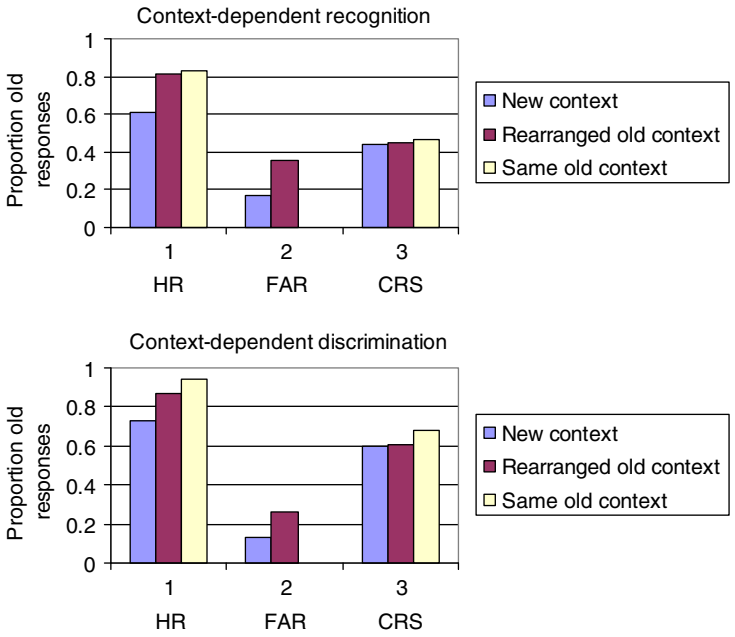

Figure 2 Mean hit rates (HR; correct old responses), false alarm rates (FAR; incorrect old responses), and corrected recognition scores (CRS; CRS equals HR minus FAR and provides a measure of discrimination) for context-dependent recognition (top panel; Hockley, 2008a: Exp.2) and context-dependent discrimination (bottom panel; Hockley, 2008a: Exp. 5B).

The ICE Model

Murnane et al. (1999) developed a global matching model, ICE (item, context, and ensemble information), to account for context-dependent recognition and discrimination. In global matching models, the information in the retrieval environment is matched with all information in memory and recognition is based on the overall strength or familiarity value associated with the degree of this match. Murnane et al. proposed that context information encoded at study and reinstated at test would increase the familiarity value associated with the test item. A target item presented in the same context at study and test would have a high degree of familiarity, because both item and context information would match the information stored in memory. A target presented in a new context at test would have a lower degree of activation because only the item itself would match the information stored in memory. Thus, context-dependent recognition would be observed when a test item is presented in an old context compared with a new context, because the old context will increase the familiarity associated with an old or a new test item. Context-dependent discrimination, however, is the result of an additional source of information. If item and context information are associated or integrated at encoding, this ensemble information provides an additional source of information that can further boost the familiarity of old items, but not new items, at test. Ensemble information, when it is encoded, produces a discrimination advantage for old items tested in their original study context.

The ICE model of Murnane et al. (1999) is based on the assumption that recognition is familiarity based. More recently, a number of researchers have argued for a dual-process view of recognition memory (see Yonelinas, 2002, for a review). In this view, recognition can be based on recollection of specific details of a prior event or on familiarity in the absence of recollection. Macken (2002) proposed a dual-process account of context-dependent discrimination.

Familiarity versus Recollection

Macken (2002) supplemented Murnane and Phelps' (1993) background and foreground color context manipulation with Tulving's (1985) subjective remember–know response procedure. In this procedure, participants classify their old recognition decisions as being based either on the recollection of specific details of the prior presentation (remember) or on a strong feeling of familiarity in the absence of recollection of any specific details (know). Macken found that context-dependent discrimination was associated with 'remember' but not with 'know' decisions and argued that EC effects are due to the encoding and retrieval of specific item–context associations. This view is also consistent with research showing that while not all recollection is based on the retrieval of context information, participants are much better at making contextual judgments about items they claimed to recognize on the basis of recollection than they are for items deemed familiar (Perfect et al., 1996).

Not all studies, however, have found a clear relationship between EC effects and remember–know responses. Hockley (2008a), for example, did not find context-dependent discrimination for remember responses using different background pictures of scenes as the context manipulation when context was encoded incidentally. In contrast, Gruppuso et al. (2007) found a large EC discrimination effect

for 'remember' responses but not for 'know' responses, following a study task in which participants rated the degree to which the target faces matched their picture contexts. As Macken (2002) noted, the dual-process view of context-dependent discrimination provides an explanation as to why EC effects are not always observed in tests of recognition. As EC effects are dependent on conscious recollection of item–context associations, they will only be found in tasks or conditions that support the encoding and retrieval of these associations.

Research has shown that the retrieval of specific associations becomes more difficult as the number of associations between different targets and the same cue increases, due to the interference between the competing associations (i.e., 'the fan effect'; Anderson and Reder, 1999). It follows that increasing the degree of fan of the context, or the number of associations between target items and the same context, will diminish the memorial benefits of context reinstatement. This is exactly what has been found in tests of recognition (Reder et al., 2002, 2013; Rutherford, 2004) and free recall (Hockley et al., 2012; Exp. 3; Smith and Manzano, 2010), providing further support for the view that the memorial benefits of EC are the result of specific item–context associations. In contrast, no effects of contextual fan have been observed for context-dependent recognition (Hockley, 2008a; Murnane and Phelps, 1993), which is to be expected if specific item–context associations have not been encoded. A variety of factors can influence the encoding of such associations including the nature of the context and the to-be remembered stimuli and whether or not the context is part of the focus of attention.

Nature of the Context and the Target Stimuli

Memorable context manipulations such as random photographs of scenes may produce large context-dependent recognition effects, but might not be very amenable to the encoding of item–context associations that give rise to context-dependent discrimination under incidental encoding conditions (e.g., Hockley, 2008a). In contrast, context-relevant scenes would be more likely to promote the integration of the target items and their context (e.g., Murnane et al., 1999). It is also the case that EC effects vary with the nature of the to-be-remembered stimuli. Many studies, particularly early in the study of EC effects, used words as the target information. More recently, other types of stimuli, such as photographs of unfamiliar faces, have been used with mixed success (e.g., Dalton, 1993; Gruppuso et al., 2007; Reder et al., 2013; Russo et al., 1999). The ease of encoding item–context associations, and the susceptibility to interference of these associations, will vary depending on the nature of the context and the target stimuli involved.

Incidental versus Intentional Encoding of Context

Context-dependent recognition is a robust phenomenon that has been observed with a variety of stimulus manipulations under incidental encoding conditions. It has been found with intact and rearranged combinations of background and foreground screen colors (Hockley, 2008a: Exp. 1; Macken, 2002; Murnane and Phelps, 1993, 1994), background pictures of scenes (Hockley, 2008a: Exp. 2; Hockley et al., 2012: Exp. 1) and background pictures at test that are physically and categorically similar to the background pictures at study, such as different pictures of sunsets and the Grand Canyon (Hockley, 2008a: Exp. 4). The fact that context-dependent recognition has also been found in two-alternative, forced-choice recognition (Hockley et al., 2012: Exp. 2) indicates that it is largely a familiarity-based phenomenon and not solely due to a response bias that classifies stimuli as old when they are shown in an old context, although response bias may also play a role (Feenan and Snodgrass, 1990). Target stimuli presented in an old context, or in a context that is similar to an old context, are associated with a greater sense of familiarity than tests presented in a noticeably different context.

Context-dependent discrimination, however, has been more elusive when encoding is incidental. The influence of context on recognition discrimination can be assessed by comparing performance for old items tested in their original or intact study context with old items tested in a rearranged, old context. A rearranged, old context condition controls the familiarity of an old context, but the rearranged context is not associated with the target item. Under incidental encoding conditions, this comparison typically yields trends towards greater accuracy in the intact context condition that often fall short of statistical significance (e.g., Hockley, 2008a; Hockley et al., 2012; Murnane and Phelps, 1994).

Much larger EC effects can be observed when attention is directed to the context as well as the target item. As examples, context-dependent discrimination was observed by Winograd and Rivers-Bulkeley (1977) when participants rated pairs of male and female faces in terms of their compatibility as a couple, by Gruppuso et al. (2007) who instructed participants to rate how well target faces matched their picture contexts and by Hockley (2008a: Exp. 5B) when participants rated how easy it was to associate the target word with the background scene. Such study tasks involve the processing and encoding of associative information that links the context and the target item. Similarly, when the context information is meaningfully related to the target item, it will also be processed with the target information. This was the case when Murnane et al. (1999) presented their target words in such relevant contexts as on a blackboard in a classroom, and when Vakil et al. (2007) examined recognition of faces that were presented in the context of hats that were characteristic of different occupations, such as a firefighter's helmet or a cook's cap.

Lucke et al. (2013) have recently demonstrated that the magnitude of the influence of context in the renewal effect also depends on whether the context was or was not relevant in a task where human participants were required to predict the occurrence of specific outcomes on the basis of different cues. Extinction was faster and response recovery was greater in the condition where context was task relevant. An analysis of eye movements further showed that the dwell times of eye gazes were longer on context stimuli when they were task relevant.

As in the case of the dwell times of eye gazes, mean response time has been shown to be a sensitive measure of cognitive performance generally, and of memory in particular (see Hockley, 2008b). Rutherford (2004) measured response time of recognition decisions when context was manipulated by background screen color. Hit rates were greater in the same than in the changed context condition, and in one of two experiments, response was also faster for hits in the same context condition. Hockley et al. (2012) analyzed response time in a recognition experiment where context was manipulated by presenting study words against different picture backgrounds. The new context condition at test was a blank background. In addition to replicating the context-dependent recognition effect, they found that response time was slower when old or new tests were shown in an old study context compared to the new (blank) context condition. Thus, even though context was incidental to the task of remembering the words, participants did spend appreciable time viewing the background context.

Smith and Vela (2001) noted the importance of examining the effects of the incidental encoding of context for theoretical, empirical, and applied reasons. It seems obvious that the incidental encoding of context information would have much more modest effects on memory performance than when contextual associations are encoded in a more directed or intended fashion. The intentional encoding of context, however, may provide a better laboratory analogue of the representation of contextual information that accrues incidentally with repeated experiences in real-life situations. This is what Gruppuso et al. (2007) attempted to demonstrate in their experimental analogy of Mandler's (1980) well-cited example of having a strong sense of familiarity but failing to recognize one's butcher when seen out of context on a bus.

Neuroscience Studies of Context Effects

Whereas the effects of context on memory have been examined primarily using behavioral methods, efforts to understand the neural systems underlying the processing of context information have proven illuminating. Memory tasks requiring the encoding and/or retrieval of context information have commonly identified the parahippocampal cortex as a critical system; the parietal cortex and the prefrontal cortex have also been implicated.

Bar and Aminoff (2003) performed four functional MRI (fMRI) experiments examining the processing of contextual information. In one experiment, subjects were presented with images of objects that had either weak preexisting associations with a specific context (e.g., a camera or a rope) or a strong preexisting contextual association (e.g., a roulette wheel), on the assumption that objects with strong contextual associations would be more likely to activate stored information about the context. For example, a subject presented with a roulette wheel would be more likely to activate stored information about the obvious context (a casino) than would subjects presented with objects that are not strongly associated with any particular context. Objects with strong contextual associations produced greater activity in parahippocampal areas, which have previously been shown to encode environmental and spatial information (Epstein and Kanwisher, 1998).

Hayes et al. (2007) presented subjects with visual objects superimposed over either a white background or a visual scene. During a recognition test, objects were presented either in their original context or over a white background, and participants displayed decreased performance when context was changed between study and test. Greater activity in the right parahippocampal cortex was found when participants successfully recognized objects previously presented in a scene. Intriguingly, when context changed between study and test, successful recognition of objects was associated with greater parahippocampal activation during encoding, suggesting that context reinstatement may be used as a retrieval/decision-making strategy.

Hayes et al. (2010) extended their previous work by examining the neural correlates of the 'butcher on the bus' phenomenon. At study, faces were presented either alone or superimposed over a background scene. At test, faces were always presented alone. Although subjects were instructed to focus only on the faces at study, recognition of faces was better for participants in the face study condition than the face–scene study condition. Faces from the face–scene study condition induced greater parahippocampal activation upon successful retrieval at test than the faces from the face-only study condition. Activation of neural systems involved in the storage of environmental and spatial information when presented with an associated face suggests that participants do form associations between a target item and its context and is also consistent with the notion that participants attempt to recall an associated context as a retrieval strategy.

Ecker et al. (2007) suggested that context affects primarily recollective processes in recognition, based on EEG/ERP evidence. Further support for the recollective effects of context comes from Balderas et al. (2008), who found that long-term (~24 h) memory for objects in contexts was impaired by pharmacological inhibition of protein synthesis in the hippocampus, but not in the perirhinal cortex. A number of investigators have proposed that recollection is primarily mediated by the hippocampus, and familiarity is primarily mediated by perirhinal (and associated) cortex (O'Reilly and Norman, 2002; Norman, 2010).

Combining fMRI with a virtual reality paradigm, Burgess et al. (2001) placed the parahippocampus in a network including parietal and various prefrontal regions. Activation of the prefrontal cortex is consistent with other studies of episodic memory (e.g., Köhler et al., 1998).

Skinner et al. (2010) used faces and scrambled faces as the context for target words rather than contexts containing spatial information, such as pictures of landscapes. Skinner et al. found increased activation in the fusiform face area for words studied with faces, relative to words studied with scrambled faces. Interestingly, this difference in activation was found although there was no difference in the overall performance between the context conditions. Similar to studies discussed above, activation in the prefrontal cortex was identified. Skinner et al.'s results suggest that rather than the parahippocampus being a 'context' area in the brain, neural recruitment of the cortical regions is likely to rely on

the type of contextual information associated with the to-be-remembered information.

Summary

In almost 100 years of study, behavioral research has clearly established that EC can influence memory performance in both human and nonhuman animals in a variety of ways. With regard to human memory, global EC has been shown to have a reasonably consistent effect on free recall performance. In contrast, local manipulations of context have produced more reliable effects for recognition memory. These effects are seen as context-dependent recognition or as context-dependent discrimination depending on whether specific item–context associations are encoded, and have been accounted for in both global matching and dual process theories of recognition memory. Neuroscience investigations of EC effects are delineating the neural systems responsible for these effects.

See also: Episodic Memory; Eyewitness Testimony; Proactive Interference; Recognition; Retroactive Interference.

Bibliography

Anderson, J.R., Reder, L.M., 1999. The fan effect: new results and new theories. Journal of Experimental Psychology: General 128, 186–197.

Baddeley, A.D., 1982. Domains of recollection. Psychological Review 89, 708–729.

Balderas, I., Rodriguez-Ortiz, C.J., Salgado-Tonda, P., Chavez-Hurtado, J., McGaugh, J.L., Bermudez-Rattoni, F., 2008. The consolidation of object and context recognition memory involve different regions of the temporal lobe. Learning & Memory 15, 618–624.

Bar, M., Aminoff, E., 2003. Cortical analysis of visual context. Neuron 38, 347–358.

Bilodeau, I.M., Schlosberg, H., 1951. Similarity in stimulating conditions as a variable in retroactive inhibition. Journal of Experimental Psychology 41, 199–204.

Bjork, R.A., Richardson-Klavehn, A., 1989. On the puzzling relationship between environmental context and human memory. In: Izawa, C. (Ed.), Current Issues in Cognitive Processes: The Tulane Floweree Symposium on Cognition. Erlbaum, Hillsdale, NJ, pp. 313–344.

Bouton, M.E., 1997. Signals for whether versus when an event will occur. In: Bouton, M.E., Fanselow, M.S. (Eds.), Learning, Motivation, and Cognition: The Functional Behaviourism of Robert C. Bolles. American Psychological Association, Washington, DC, pp. 385–409.

Bouton, M.E., Bolles, R.C., 1979. Contextual control of extinction of conditioned fear. Learning and Motivation 10, 445–466.

Bouton, M.E., Nelson, J.B., Rosas, J.M., 1999. Stimulus generalization, context change, and forgetting. Psychological Bulletin 125, 171–186.

Bower, G.H., Monteiro, K.P., Gilligan, S.G., 1978. Emotional mood as a context for learning and recall. Journal of Verbal Learning and Verbal Behavior 17, 573–585.

Burgess, N., Maguire, E.A., Spiers, H.J., O'Keefe, J., 2001. A temporoparietal and prefrontal network for retrieving the spatial context of lifelike events. NeuroImage 14, 439–453.

Cann, A., Ross, D.A., 1989. Olfactory stimuli as context cues in human memory. American Journal of Psychology 2, 91–102.

Carr, H.A., 1917. Maze studies with the white rat: I. Normal animals. Journal of Animal Behavior 7, 259–275.

Carr, H.A., 1925. Psychology: A Study of Mental Activity. Longmans Green, New York.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, second ed. Lawrence Erlbaum Associations, Inc, Hillsdale, NJ.

Dallett, K., Wilcox, S.G., 1968. Contextual stimuli and proactive inhibition. Journal of Experimental Psychology 78, 475–480.

Dalton, P., 1993. The role of stimulus familiarity in context-dependent recognition. Memory & Cognition 21, 223–234.

Dando, C., Wilcock, R., Milne, R., 2009. The cognitive interview: the efficacy of a modified mental reinstatement of context procedure for frontline police investigators. Applied Cognitive Psychology 23, 138–147.

Dougal, S., Rotello, C.M., 1999. Context effects in recognition memory. American Journal of Psychology 112, 277–295.

Dulsky, S.G., 1935. The effect of a change in background on recall and relearning. Journal of Experimental Psychology 18, 725–740.

Eckert, E., Kanak, N.J., Stevens, R., 1984. Memory for frequency as a function of environmental frequency. Bulletin of the Psychonomic Society 22, 507–510.

Ecker, U.K.H., Zimmer, H.D., Groh-Bordin, C., 2007. Color and context: an ERP study on intrinsic and extrinsic feature binding in episodic memory. Memory & Cognition 35, 1483–1501.

Eich, J.E., 1980. The cue-dependent nature of state-dependent retrieval. Memory & Cognition 8 (2), 157–173.

Eich, E., 1985. Context, memory, and integrated item/context imagery. Journal of Experimental Psychology: Learning, Memory, and Cognition 11, 764–770.

Eich, E., Metcalfe, J., 1989. Mood dependent memory for interval versus external events. Journal of Experimental Psychology: Learning, Memory, and Cognition 15, 443–455.

Emmerson, P.G., 1986. Effects of environmental context on recognition memory in an unusual environment. Perceptual & Motor Skills 63, 1047–1050.

Epstein, R., Kanwisher, N., 1998. A cortical representation of the local visual environment. Nature 392, 598–601.

Feenan, K., Snodgrass, J.G., 1990. The effect of context on discrimination and bias in recognition memory for pictures and words. Memory & Cognition 18, 515–527.

Fernandez, A., Glenberg, A.M., 1985. Changing environmental context does not reliably affect memory. Memory & Cognition 13, 333–345.

Fisher, R.P., Geiselman, R.E., 1992. Memory-Enhancing Techniques for Investigative Interviewing: The Cognitive Interview. Thomas, Springfield, IL.

Geiselman, R.E., Bjork, R.A., 1980. Primary versus secondary rehearsal in imagined voices: different effects on recognition. Cognitive Psychology 12, 188–205.

Geiselman, R.E., Glenny, J., 1977. Effects of imagining speakers' voices on the retention of words presented visually. Memory & Cognition 5, 499–504.

Gillund, G., Shiffrin, R.M., 1984. A retrieval model for both recognition and recall. Psychological Review 91 (1), 1–67.

Glenberg, A.M., 1979. What memory is for? Behavioral & Brain Sciences 20, 1–55.

Godden, D.R., Baddeley, A.D., 1975. Context-dependent memory in two natural environments: on land and underwater. British Journal of Psychology 66, 325–331.

Godden, D.R., Baddeley, A.D., 1980. When does context influence recognition memory? British Journal of Psychology 71, 99–104.

Greenspoon, J., Ranyard, R., 1957. Stimulus conditions and retroactive inhibition. Journal of Experimental Psychology 53, 55–59.

Gruppuso, V., Lindsay, S., Masson, M.E.J., 2007. I'd know that face anywhere! Psychonomic Bulletin & Review 14, 1085–1089.

Hayes, S.M., Baena, E., Truong, T.K., Cabeza, R., 2010. Neural mechanisms of context effects on face recognition: automatic binding and context shift decrements. Journal of Cognitive Neuroscience 22, 2541–2554.

Hayes, S.M., Nadel, L., Ryan, L., 2007. The effect of scene context on episodic object recognition: parahippocampal cortex mediates memory encoding and retrieval success. Hippocampus 17, 873–889.

Hockley, W.E., 2008a. The effects of environmental context on recognition memory and claims of remembering. Journal of Experimental Psychology: Learning, Memory, and Cognition 34, 1412–1429.

Hockley, W.E., 2008b. Memory search: a matter of time. In: Roediger III, H.L. (Ed.), Cognitive Psychology of Memory. In: Byrne, J. (Ed.), Learning and Memory: A Comprehensive Reference, vols. 4, vol. 2. Elsevier, Oxford, pp. 417–444.

Hockley, W.E., Bancroft, T.D., Bryant, E., 2012. Associative and familiarity-based effects of environmental context on memory. Canadian Journal of Experimental Psychology 66, 81–89.

Humphreys, M.S., Bain, J.D., Pike, R., 1989. Different ways to cue a coherent memory system: a theory for episodic, semantic, and procedural tasks. Psychological Review 96, 208–233.

Jacoby, L.L., 1983. Perceptual enhancement: persistent effects of an experience. Journal of Experimental Psychology: Learning, Memory, & Cognition 9, 21–38.

Köhler, S., Moscovitch, M., Winocur, G., Houle, S., McIntosh, A.R., 1998. Networks of domain-specific and general regions involved in episodic memory for spatial location and object identity. Neuropsychologia 46, 129–142.

Krafka, C., Penrod, S., 1985. Reinstatement of context in a field experiment in eyewitness identification. Journal of Personality and Social Psychology 49, 58–69.

Light, L.L., Carter-Sobell, L., 1970. Effects of changed semantic context on recognition memory. Journal of Verbal Learning and Verbal Behavior 9 (1), 1–11.

Lucke, S., Lachnit, H., Koenig, S., 2013. The informational value of contexts affects context-dependent learning. Learning & Behavior 41, 285–297.

Macken, W.J., 2002. Environmental context and recognition: the role of recollection and familiarity. Journal of Experimental Psychology: Learning, Memory, and Cognition 28, 153–161.

Malpass, R.S., Devine, P.G., 1981. Guided memory in eyewitness identification. Journal of Applied Psychology 66, 343–350.
Mandler, G., 1980. Recognizing: the judgment of previous occurrence. Psychological Review 87, 252–271.
Murnane, K., Phelps, M.P., 1993. A global activation approach to the effect of changes in environmental context on recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition 19, 882–894.
Murnane, K., Phelps, M.P., 1994. When does a different environmental context make a difference in recognition? A global activation model. Memory & Cognition 22, 584–590.
Murnane, K., Phelps, M.P., 1995. Effects of changes in relative cue strength on context-dependent recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition 21, 158–172.
Murnane, K., Phelps, M.P., Malmberg, K., 1999. Context-dependent recognition memory: the ICE theory. Journal of Experimental Psychology: General 4, 403–415.
Nagge, J.W., 1935. An experimental test of the theory of associative interference. Journal of Experimental Psychology 18, 663–682.
Norman, K.A., 2010. How hippocampus and cortex contribute to recognition memory: revisiting the complementary learning systems model. Hippocampus 20, 1217–1227.
O'Reilly, R.C., Norman, K.A., 2002. Hippocampal and neocortical contributions to memory: advances in the complementary learning systems framework. Trends in Cognitive Sciences 6, 505–510.
Perfect, T.J., Mayes, A.R., Downes, J.J., Van Eijk, R., 1996. Does context discriminate recollection from familiarity in recognition memory? Quarterly Journal of Experimental Psychology: Human Experimental Psychology 49 (A), 797–813.
Postman, L., Underwood, B.J., 1972. Critical issues in interference theory. Memory & Cognition 1, 19–40.
Rand, G., Wapner, S., 1967. Postural status as a factor in memory. Journal of Verbal Learning & Verbal Behavior 6, 268–271.
Reder, L.M., Donavos, D.K., Erickson, M.A., 2002. Perceptual match effects in direct tests of memory: the role of contextual fan. Memory & Cognition 30, 312–323.
Reder, L.M., Lindsay, W.V., Mannelis, A., Oates, J.M., Dutcher, J.M., Bates, J.T., Cook, S., Aizenstein, H.J., Quinlan, J., Gyulai, F., 2013. Why it's easier to remember seeing a face we already know than one we don't: preexisting memory representations facilitate memory formation. Psychological Science 24 (3), 363–372.
Rosas, J.M., Callejas-Aguiler, J.E., Ramos-Alvarez, M.M., Abad, M.J.F., 2006. Revision of retrieval theory of forgetting: what does make information context-specific? International Journal of Psychology and Psychological Therapy 6, 147–166.
Russo, R., Ward, G., Geurts, H., Scheres, A., 1999. When unfamiliarity matters: changing environmental context between study and test affects recognition for unfamiliar stimuli. Journal of Experimental Psychology: Learning, Memory, and Cognition 25 (2), 488–499.
Rutherford, A., 2004. Environmental context-dependent recognition memory effects: an examination of ICE model and cue-overload hypotheses. Quarterly Journal of Experimental Psychology 57A, 107–127.
Saufley Jr., W.H., Otaka, S.R., Bavaresco, J.L., 1985. Context effects: classroom tests and contextual independence. Memory & Cognition 13, 522–528.
Schab, F.R., 1990. Odors and the remembrance of things past. Journal of Experimental Psychology: Learning, Memory, & Cognition 16, 648–655.
Skinner, E.I., Grady, C.L., Fernandes, M.A., 2010. Reactivation of context-specific brain regions during retrieval. Neuropsychologia 48, 156–164.
Smith, S.M., 1979. Remembering in and out of context. Journal of Experimental Psychology: Human Learning and Memory 5, 460–471.
Smith, S.M., 1984. A comparison of two techniques for reducing context-dependent forgetting. Memory & Cognition 12, 477–482.
Smith, S.M., 1985a. Background music and context-dependent memory. American Journal of Psychology 98, 591–603.
Smith, S.M., 1985b. Effects of number of study environments and learning instructions on free recall clustering and accuracy. Bulletin of the Psychonomic Society 23, 440–442.
Smith, S.M., 1986. Environmental context-dependent recognition memory using a short-term memory task for input. Memory & Cognition 14, 347–354.
Smith, S.M., 1988. Environmental context-dependent memory. In: Davies, G.M., Thompson, D.M. (Eds.), Memory in Context: Context in Memory. Wiley, New York, pp. 13–34.
Smith, S.M., Glenberg, A., Bjork, R.A., 1978. Environmental context and human memory. Memory & Cognition 6, 342–353.
Smith, S.M., Vela, E., 1992. Environmental context-dependent eye-witness recognition. Applied Cognitive Psychology 6, 125–139.
Smith, S.M., Vela, E., 2001. Environmental context-dependent memory: a review and meta-analysis. Psychonomic Bulletin & Review 8, 203–220.
Smith, S.M., Manzano, I., 2010. Video context dependent-recall. Behavior Research Methods 42, 292–301.
Spear, N.E., 1978. The Processing of Memories: Forgetting and Retention. Lawrence Erlbaum, Hillsdale, NJ.
Strand, B.Z., 1970. Change of context and retroactive inhibition. Journal of Verbal Learning & Verbal Behavior 9, 202–206.
Tulving, E., 1983. Elements of Episodic Memory. Oxford University Press, New York.
Tulving, E., 1985. Memory and consciousness. Canadian Psychologist 26, 1–12.
Tulving, E., Thomson, D.M., 1973. Encoding specificity and retrieval processes in episodic memory. Psychological Review 80, 359–380.
Vakil, E., Raz, T., Levy, D.A., 2007. Multifactorial context effects on visual recognition memory. The Quarterly Journal of Experimental Psychology 60, 916–923.
Weiss, W., Margolius, G., 1954. The effect of context stimuli on learning and retention. Journal of Experimental Psychology 48, 318–322.
Winograd, E., Rivers-Bulkeley, N.T., 1977. Effects of changing context on remembering faces. Journal of Experimental Psychology: Human Learning and Memory 3, 397–405.
Yonelinas, A.P., 2002. The nature of recollection and familiarity: a review of 30 years of research. Journal of Memory and Language 46 (3), 441–517.
Zental, T.R., 1970. Effects of context change on forgetting rates. Journal of Experimental Psychology 86, 440–448.

Contextual Studies: Methodology

Laurence Steinberg, Temple University, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 4, pp. 2705–2709, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article examines the ways in which context is taken into account in contemporary social science research. The ecological perspective provides a conceptual framework within which to investigate complex interactions between persons and environments. Proximal settings are seen as connected to each other, contained within broader institutional and community contexts that shape their structure and influence what takes place within them. The scientific study of context presents a number of formidable methodological challenges. Among the most important of these challenges discussed in this article are deciding upon the most appropriate way to assess context, developing models of the transactional nature of contextual influence, and disentangling the effects of hierarchically organized environments. The study of context in the social sciences, once the province of disciplines such as sociology and anthropology, has now become an accepted focus of inquiry within perspectives that formerly had been concerned exclusively with the study of individuals in controlled, laboratory settings, such as psychology. This article examines the ways in which context is taken into account in contemporary social science research.

Historical Underpinnings

The contemporary study of context has its origins in a number of different disciplines, not only sociology and anthropology, but also *ethology* (the study of animal behavior in natural habitats), *ecology* (the study of biological organisms and their environment), and urban geography (the study of the development and functioning of neighborhoods in cities). All of these disciplines emphasize the fact that any human or animal behavior cannot be understood outside of the setting in which it occurs. In this sense, the organism and context are seen as inextricably linked.

Psychology, as a discipline, was slow to embrace a contextual point of view. Although numerous writers had attempted to move the field in this direction – most notably, Kurt Lewin (1951) – the study of psychology was dominated for much of the twentieth century by experimental scientists, who studied human behavior in the laboratory. Context was seen as a nuisance variable – something that got in the way of pure science – and the goal of experimental research was to study behavior under conditions that permitted the investigator to control any variable that was seen as extraneous to the research focus. If one wanted to study, for example, how individuals make decisions, the researcher would need to control the setting in which the decision making took place, so that each subject in the study could be observed under identical circumstances. The simplest way to do this was to perform research in a laboratory, where each and every element of the context could be controlled.

The benefit of controlled laboratory research is that it permits the scientist to isolate pure behavioral phenomena, uncontaminated by the outside environment. But this benefit must be weighed against the fact that phenomena observed under controlled experimental conditions may not be the same as phenomena observed in their natural context. The way in which individuals make decisions in a psychology laboratory, for instance, may not resemble the way they make decisions at home, at work, or in a store. Just as ethologists would not want to generalize observations of caged animals to the behavior of that same species in the wild, social scientists must be careful about generalizing findings derived from controlled experiments to human behavior outside the laboratory. This is not to question the importance of laboratory-based research but, rather, to note that the laboratory is itself a context, and that features of this context, like any other context, necessarily influence the behavior that occurs within it.

During the 1970s, concerns about ecological validity – the extent to which an observed phenomenon in a research investigation is a valid reflection of that phenomenon as it occurs in natural contexts – were voiced by numerous investigators working in a variety of areas. In the study of intelligence, for example, scholars raised questions about whether standardized intelligence tests adequately captured intelligent behavior as it occurs in the real world. Several researchers found that individuals who performed poorly on standardized tests of mathematics frequently demonstrated the ability to use very sophisticated mathematical reasoning when observed during day-to-day situations, such as shopping. Similarly, scholars of social behavior discovered that individuals' behavior was considerably influenced by the context in which it took place, and that conclusions drawn from studies of behavior in one context (e.g., observations of young children at home) could not always be extended to other settings (e.g., observations of these same youngsters in the preschool classroom). Ecological psychologists drew attention to the fact that individual behavior was profoundly shaped by the structure of the physical environment in which people lived. Several classic studies of schools indicated, for example, that students' behavior in small schools was markedly different from students' behavior in large schools (Barker and Gump, 1964).

In the late 1970s, psychologist Urie Bronfenbrenner drew widespread attention to the role of context in human development and behavior by noting that the scientific literature on child development was based almost entirely on the study of

children interacting with strangers in strange situations for the briefest amounts of time. How, Bronfenbrenner wondered, could social scientists hope to contribute to the solution of social problems, if they insisted on studying human behavior in a contextual vacuum? Although his comments were aimed at students of child development, they applied equally well to other disciplines. In his book, *The Ecology of Human Development*, Bronfenbrenner (1979) called on the field to think in more sophisticated fashion about the settings in which development took place, and he outlined a model of context that researchers could use to guide their thinking and empirical investigation. This work, and the thinking it would subsequently inspire, led to the development of the ecological perspective on human development, a perspective that grew more and more influential over the next several decades (Bronfenbrenner and Crouter, 1983).

The Ecological Perspective

Within the ecological perspective, the context is construed as a series of nested environments, with each level of contextual influence embedded within a larger level, not unlike a set of Russian dolls. At the innermost, or most proximal level of contextual influence, are the immediate settings in which individuals spend time – the family, the friendship group, the classroom, the workplace, the laboratory, and so forth. Most studies of the role of context in shaping human behavior and development are studies of single contexts examined in isolation from one another (e.g., the impact of the family context on child development, the impact of teaching practices on student achievement).

The ecological perspective provides a conceptual framework within which to investigate more complex interactions between persons and environments, however. Three notions are especially important in this regard. First, proximal settings are seen as connected to each other, because events that take place in one setting often have ramifications for individual behavior and development in another. Socialization in the family context influences how adolescents behave in the peer group, peer influence processes affect how adolescents behave in the classroom, experiences in the classroom affect students' performance in extracurricular settings, and so on. Within the ecological perspective, therefore, an important focus of inquiry is on the ways in which, and the extent to which, different settings are linked, and the impact of these links on individual behavior.

Second, within the ecological perspective, these proximal settings are seen as contained within broader institutional and community contexts that shape the structure of settings and influence what takes place within them (Duncan and Raudenbush, 1998; Jencks and Mayer, 1990). Families and friendship networks reside within neighborhoods, for example, and qualities of neighborhoods influence how families and friendship networks function. Parents interact with their children differently in poor neighborhoods than they do in more affluent ones, for example, and, as a consequence, child development may vary across neighborhood contexts. Similarly, schools exist within communities, and characteristics of communities – urban versus rural, tightly knit versus impersonal – necessarily influence the ways in which schools are organized and operate. Within the ecological perspective, therefore, one needs to examine relations among nested contexts as well as between individuals and settings.

Finally, proximal settings and the broader environments that contain them are located within particular historical, social, economic, political, geographical, and cultural contexts – 'macrosystems' – that influence the nature, structure, function, organization, and influence of all levels of the environment. Schools in Japan function differently than those in Canada, and differences between Japanese and Canadian schools have implications for the ways that teachers and students interact with each other, which in turn have implications for understanding student achievement in each context. A city's neighborhoods today may bear little resemblance to these very same neighborhoods 50 years ago, and the ways in which neighborhoods have changed over historical time may have important implications for understanding transformations in family life and parent–child relationships. Important social and economic events – wars, economic depressions, natural disasters – shape proximal and distal settings in important ways that have ramifications for individuals' behavior and development. Within the ecological perspective, researchers are interested in the influence of macrosystems on the settings located within them, and, consequently, in the rippling effects of these influences on individual behavior and development.

Methodological Challenges in Contextual Research

As more and more social scientists began to incorporate an increasingly sophisticated model of context into their research designs, it became clear that the scientific study of context presented a number of new and formidable methodological challenges. The discussion shifted from *whether* one needed to study context (the consensus was that one did) to how best to do so. Among the most important of these concerns involved deciding upon the most appropriate way to assess context, developing models of the transactional nature of contextual influence, and disentangling the effects of hierarchically organized, or nested, environments.

Objective versus Subject Assessment of Context

One of the longstanding debates in the study of context concerned the relative merits of objective versus subjective assessments of the environment. Objective assessments attempt to separate the measurement of context from the assessment of the individuals whose behavior is the target of investigation. A researcher interested in the influence of the parent–child relationship on child development might conduct impartial observations of family interaction in the home and draw his or her own conclusions about the relationship on the basis of what was observed. Similarly, a student of neighborhood stress and its impact on mental health might draw upon official statistics on crime, poverty, unemployment, and housing density and look at these in relation to residents' mental health problems.

Subjective assessments of context, in contrast, rely on the research participants' perceptions of their environment. These

perceptions can be assessed by asking individuals to complete questionnaires or by interviewing them about the settings in which they live. Thus, a researcher interested in parent–child relationships might interview parents about their socialization practices or ask children to complete questionnaires about their parents' behavior. Similarly, someone interested in neighborhood stress and mental health might collect information from neighborhood residents on their perceptions of the community's dangerousness, economic health, or housing quality and examine the links between these perceptions and one or more indices of mental health.

These different approaches to the assessment of context have particular strengths and weaknesses. The advantage of using objective assessments inheres in the fact that the assessment of context is not contaminated by the characteristics of the individuals whose behavior is being studied. Suppose, for instance, a researcher interested in neighborhood stress and mental health gathered information about each by interviewing a sample of residents, and found that depression was more prevalent in dangerous areas. Because depressed individuals may describe their neighborhoods in more negative terms than nondepressed individuals, however, it would be difficult to know whether the observed link between dangerousness and depression was genuine or merely an artifact of biased reporting. If, however, the relationship between dangerousness and depression were confirmed in a study using official crime statistics, one would have more confidence in the validity of the finding.

The advantage of using subjective assessments of context is that such assessments capture the context as it is experienced by the participant. Some social scientists argue that it is the subjectively experienced context, and not the objective context, that has the most profound impact on behavior and development. Individuals' perceptions of their neighborhood as safe versus dangerous, for example, may well influence how they behave far more than the actual level of crime as measured by official statistics. As the sociologist W.I. Thomas wrote early in the twentieth century, if individuals "perceive situations as real, they are real in their consequences" (Thomas and Thomas, 1928). Of course, one problem that arises when researchers rely on subjective assessments of context concerns the reconciliation of different subjects' perceptions of the same context. In family research, for example, it is common to find that different family members characterize their family context in different ways.

Ultimately, the most thorough assessment of context is one that employs multiple sources of information and uses multiple methods of assessment. By combining information from objective and subjective sources, from multiple vantage points, and gained through observational, self-report, and archival data, researchers are often able to provide richer and more accurate characterizations of a setting than are achieved through one approach alone.

The Transactional Nature of Contextual Influence

A second methodological challenge involves capturing the transactional nature of contextual influence. Although researchers often claim to be studying the influence of context on the individual, in reality the situation is far more complicated than this, in at least three important ways. First, individuals are not the passive recipients of contextual influence, but are active agents who shape the contexts in which they live. At the same time, contexts are not static entities, but settings that change and develop over time. Thus, while children are no doubt influenced by their family environment, family environments are themselves influenced by the children who inhabit them. Similarly, friendship groups both affect, and are affected by, their members. Finding ways to model the reciprocal and dynamic relations between individuals and their environments has been an especially difficult challenge for researchers.

Second, it is often difficult to separate the effects of contextual selection from those of contextual influence (Jencks and Mayer, 1990). In laboratory experiments, individuals can be randomly assigned by the researcher to different groups and then compared to see whether group membership affects individual behavior. Outside the laboratory, however, people often pick the contexts in which they spend time. Consider, for example, the often reported finding that delinquent teenagers often are members of peer groups composed of other delinquent adolescents. Is this because delinquent peers influence their friends to engage in antisocial behavior, or is it due, instead, to the propensity of delinquent individuals to choose to affiliate with others who share similar attitudes and values? Obviously, it is not possible to answer this question without randomly assigning individuals to different peer groups (not a feasible option in most research) or observing individuals and contexts over time. Similar problems exist in attempts to document the influence of neighborhoods on their residents. Because individuals have some control over where they live, it is difficult to disentangle the effects that neighborhoods have on individuals from the factors that lead some people to select particular neighborhoods in which to settle.

Third, different individuals are often affected by the same contextual factors in very different ways, making it extremely difficult to generalize about the effects of various contextual factors on individual behavior and development. A highly active and impulsive child may fare poorly in a relatively unstructured classroom, whereas this same context may be perfectly suitable for a more passive, more placid youngster. Similarly, an employee who is self-reliant and self-directed may thrive in a work environment that permits a great deal of autonomy, but this same work environment may prove problematic for an employee who needs a great deal of guidance and supervision. The recognition that contextual influence varies as a function of differences between individuals has stimulated a good deal of interest in the study of 'person–environment fit' – the extent to which a given context meshes well with the individual's personal characteristics.

Hierarchical Nature of Contextual Influence

The third methodological challenge concerns the hierarchical and interactive nature of contextual influence. The recognition that individuals are affected by multiple contexts simultaneously has necessitated the development of more sophisticated conceptual and statistical models in order to best describe the complicated ways through which individuals are influenced by the contexts in which they live (Duncan and Raudenbush, 1998).

Some contexts, for example, affect individuals directly, as when the behavior of a parent directly affects the development of a child. Other contextual effects occur through their impact on some other setting, however; stress in the parent's workplace may affect the child indirectly, through its impact on the family. In this instance, the second setting (in this case, the family) is said to 'mediate' the impact of the first setting (i.e., the workplace). And still other contextual effects occur by transforming, or 'moderating,' the impact of other contexts. Neighborhoods, for example, often exert their strongest impact on individuals not directly, but by attenuating or enhancing the impact of other contexts, such as the family or school. Within the context of an affluent and safe neighborhood, close parental monitoring may have deleterious effects on the child (because such monitoring may be experienced by the child as overprotective or intrusive), but the very same level of parental monitoring may have beneficial effects within the context of a dangerous, economically stressed community. In order to analyze direct, mediating, and moderating effects, researchers have had to develop new classes of multivariate statistical procedures, such as path analysis, structural equation modeling, and hierarchical linear modeling (Gelman and Hill, 2006).

Concluding Note

The growing appreciation of the importance of context in the study of human behavior and development stimulated a wave of new research in the 1980s and 1990s that sought to understand how individuals are affected by different levels of the environment, both independently and in interaction with each other. Multidisciplinary and interdisciplinary research that brought together social scientists from different disciplines became increasingly common, as scholars realized that understanding the complex interrelations among individuals and their environments necessitated drawing on the methods and theories of multiple disciplines. Over time, the study of individuals in natural settings has come to dominate research in fields in which controlled laboratory experiments had once been the dominant paradigm. Once considered a nuisance that interfered with the pursuit of 'pure knowledge,' the context of human behavior is now recognized as a necessary and enormously important focus of contemporary social science inquiry.

See also: Causal Inference: Overview; Comparative Studies: Method and Design; Ecological Fallacy, Statistics of; Ecological Inference; Endogeneity: Methodology; External Validity; Internal Validity; Mechanism-Based Causal Analysis; Moderator Variable: Methodology; Nonequivalent Group Designs; Statistical Analysis: Multilevel Methods.

Bibliography

Barker, R.G., Gump, P.V., 1964. Big School, Small School: High School Size and Student Behavior. Stanford University Press, Stanford, CA.

Bronfenbrenner, U., 1979. The Ecology of Human Development: Experiments by Nature and Design. Harvard University Press, Cambridge, MA.

Bronfenbrenner, U., Crouter, A., 1983. The evolution of environmental models in developmental research. In: Kessen, W., Mussen, P.H. (Eds.), Handbook of Child Psychology, History, Theory, and Methods, fourth ed., vol. 1. Wiley, New York, pp. 357–414.

Duncan, G., Raudenbush, S., 1998. Neighborhoods and Adolescent Development: How Can We Determine the Links? Northwestern University, Evanston, IL. Working Paper Series, Institute for Policy Research.

Gelman, A., Hill, J., 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Jencks, C., Mayer, S., 1990. The social consequences of growing up in a poor neighborhood. In: Lynn, L.E., McGeary, M.G. (Eds.), Inner-City Poverty in the United States. National Academy Press, Washington, DC, pp. 111–186.

Lewin, K., 1951. Field Theory and Social Science. Harper, New York.

Thomas, W.I., Thomas, D.S., 1928. The Child in America. Knopf, New York.

Continental Philosophy of Language

Cristina Lafont, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Continental philosophy of language is a tradition of reflexion on language initiated in the eighteenth century by authors such as J. G. Hamann and J. G. Herder, and later developed by W. von Humboldt. Through the work of M. Heidegger it has extended its influence to contemporary hermeneutic philosophy (H.-G. Gadamer, K.-O. Apel, J. Habermas). All the authors of this tradition are primarily interested in the world-disclosing function of language, that is, in the question of how language shapes our experience of the world through the net of concepts that it entails. Thus, they oppose the traditional view of language as a mere tool for the designation of entities given independently of language or the expression of prelinguistic thoughts. In contradistinction to this view, they consider language as constitutive of our experience of the world. But given the plurality of natural languages, this view seems immediately confronted by the problem of linguistic relativism: if speakers of different languages experience the world in genuinely different ways, how can they communicate with each other or reach shared knowledge of the world? Due to this difficulty, one of the main concerns in this tradition is to explain how understanding across different languages is possible.

Although philosophical interest in language is as old as philosophy itself, during the twentieth century this interest flourished to such an extent that the analysis of language has now become one of the most productive philosophical fields. Besides the general interest in language present in virtually all different philosophical traditions of the past century, it is possible to distinguish at least two very different ways of approaching linguistic analysis, namely the so-called Anglo-American and continental philosophies of language. The first (also called analytic philosophy) usually refers to the kind of analysis of language initiated at the end of the nineteenth and the beginning of the twentieth century by authors such as G. Frege, B. Russell, L. Wittgenstein, the logical positivists, etc. The second, continental philosophy of language, refers to the conception of language that finds its roots in the eighteenth century, in the works of authors such as J. G. Hamann and J. G. Herder, and was further developed in the nineteenth century by W. von Humboldt. Through the work of M. Heidegger it has extended its influence to contemporary hermeneutic philosophy (H.-G. Gadamer, K.-O. Apel, J. Habermas, etc.). In this article the specific features of the continental conception of language are examined.

In general terms, one of the crucial differences between the Anglo-American and the continental philosophies of language is the explicit attempt found in all the authors of the continental tradition to break with what Humboldt termed the "primacy of logic over grammar," that is, the predominance given from the very beginnings of Greek philosophy to the cognitive function of language at the expense of its communicative function. As a consequence of conceiving language less as a vehicle of knowledge and more as a means of understanding, this tradition analyzes the role that language plays not only in our relationship with the objective world (by allowing us to have propositional knowledge about it) but also in our relationship with the social world (which is constituted through communicative interaction) and with our own subjective worlds (which are expressible through linguistic articulation). In this way, language is considered in its multidimensional world-disclosing function. In keeping with this focus, the continental philosophy of language has always concentrated on the analysis of natural languages, developed through contingent historical processes, and has considered these as constitutive of our experience of the world.

To understand the roots of the distinctive features of the continental conception of language, it is important to keep in mind that this conception was developed against the background of an explicit confrontation with Kant's philosophy. This is important because, despite all the criticisms of Kant, the continental philosophy of language remains to this day under the decisive influence of transcendental philosophy. This influence explains two of its most characteristic features: its idealist tendencies (which in some cases lead to a clear hypostatization of language, most remarkably in the work of the late Heidegger) and one of the crucial problems that have concerned the authors of this tradition from the very beginning, namely the problem of linguistic relativism. The connection between these features can be understood best through an analysis of the different historical phases of development within this tradition.

The Hamann–Herder–Humboldt Tradition

As already suggested, most of the reflexions on language found in the works of the authors of this tradition take place in a more or less explicit confrontation with Kant's philosophy. In the cases of Hamann and Herder, this confrontation takes the specific form of a metacritique of Kant's critique of pure reason (see Hamann (1784) and Herder (1799)). One important target of both metacritiques is Kant's attempt to purify reason from its internal connection with language. The constant reference that Hamann and Herder make to the common origin of both expressions language and reason in the Greek word *logos* is by no means a merely etymological remark. With the explicit identification of language and reason (i.e., the claim that "reason is language, *logos*" or "without the word, neither reason nor world"), these authors initiate a way of thinking about both language and reason that moves constantly

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63012-X

between the extremes of universalism and relativism, rationalism, and irrationalism. On the one hand, the identification of language and reason supports the attempt to transform Kant's idea of a critique of reason into the romantic idea of the essential limitations of reason (due to the limits of the natural languages that contingently developed in history). This is a powerful motive behind the critique of the Enlightenment in which many authors of this tradition participate (from Hamann to Heidegger and Gadamer). On the other hand, however, the identification of language and reason also supports the attempt to preserve the universalism of Kant's conception of reason in an equally universalist conception of communicative rationality (from Humboldt to Apel and Habermas). But beyond the remarkable differences in the philosophical programs among the authors of this tradition, the common presupposition behind their identification of language and reason is the view of language as constitutive of our experience in the strict sense given by transcendental philosophy. Accordingly, language is conceived as having the transcendental power to constitute experience that Kant erroneously ascribed to an abstract, transcendental ego.

Hamann's critique of Kant can be seen in retrospect as the initiator of such a view. Hamann identifies language as the common root of sensibility and understanding, which was sought by Kant without success. In this way, he considers language as both empirical and transcendental at the same time. By conceding to language such a quasitranscendental status, Hamann certainly disagrees with Kant's view of language. However, he does not disagree with the most basic assumptions of transcendental philosophy. Hamann is actually taking for granted Kant's transcendental idealism, namely the idea that the conditions of possibility of experience are at the same time the conditions of possibility of the objects of experience. As a consequence of Hamann's critique, this idea is now understood to mean that the conditions of possibility of our linguistically mediated experience of the world are, at the same time, the conditions of possibility of that which is experienced in the world. Now, the claim that language does not merely mirror but actually shapes our experience of the world faces a problem completely foreign to transcendental philosophy. For the various historical languages do not seem to offer an equivalent for the Kantian transcendental ego. First of all, these languages are plural. And second, they do not allow for a sharp distinction between the empirical and the transcendental, between that which is supposed to count as valid in them a priori (knowledge of the language) and that which is valid a posteriori (knowledge of the world). Thus, if languages shape our experience of the world, they do so in merely contingent and plural ways. As Hamann points out, the meaning and use of words of different languages are "a priori arbitrary and indifferent, but a posteriori necessary and indispensable." Seen in this light, Hamann's critique of Kant seems to force the relativist conclusion that Humboldt will express later in one of his most famous theses, namely, that "every language contains a peculiar world-view."

Hamann never addressed the problem of linguistic relativism entailed in his transformation of transcendental philosophy. Moreover, given the decisively antisystematic impulse behind his critique of Kant's philosophical system, he never attempted to work this new view of language into a systematic conception. His reflections on language are limited to more or less isolated remarks spread throughout his writings. A much more elaborate account of this new conception of language can be found in Humboldt's work. Although Humboldt's general reflections about language do not amount to a philosophical theory of language – they are mostly confined to the introductions to empirical works in comparative linguistics – one of their specific aims is to bring about a radical break with the traditional conception of language. Humboldt is convinced that without such a break a science of language is not possible at all. Thus, his writings contain a very articulated account of the essential features that distinguish the new conception of language from the received view.

In the context of criticizing the traditional conception of language in his work *Ueber den Nationalcharakter der Sprachen*, Humboldt enumerates the basic features of his own conception of language when he remarks "that the diversity of languages is more than a mere diversity of signs, that words and their syntax simultaneously shape and determine our concepts and that, given their systematicity and their influence on knowledge and sensibility, different languages are in fact different world-views" (Humboldt, 1903–1936, IV, p. 420). In this brief summary, we can see the three fundamental claims that separate Humboldt's position from the received view: (1) the rejection of the view of language as a mere system of signs, as a tool for the transmission of prelinguistic thoughts or the designation of objects that exist independently of language; (2) the thesis of the identity of thought and language that justifies the critique of the instrumental conception of language; and as a consequence, (3) the claim that different languages are in fact different world-views. The problem of linguistic relativism implicit in this consequence can be traced back to the very core of Humboldt's critique of the instrumental conception of language.

The traditional view of language as a mere instrument for the designation of entities independent of language is based on an assimilation of two relations that are actually irreducible to one another: the relation of designation between a name and an object and the relation of satisfaction between a predicate and the objects to which the predicate applies. From this assimilation there arises the problem of determining what the predicates actually designate, a problem encountered already in the Middle Ages; the ensuing dispute over the universals offers clear testimony as to its depth. Nonetheless, this highly problematic conception of language can be considered as the ruling view from Aristotle to Kant. Humboldt seeks a decisive break with this reduction of the powers of language to their designating function, advocating a more stratified analysis of language. In contemporary terms, such an analysis would have to deal not only with the name/object distinction but also with the distinction between concept (or class) and object. By the same token, concerning the interpretation of signs, this analysis would have to differentiate between their meaning and reference. Humboldt states this last distinction as follows: "The word acts in such a way that the object given in it is represented by the soul. This representation must be distinguished from the object; … along with the objective part *that relates to the object*, it has a subjective part that *lies in its manner of conceiving*. … In turn, we scarcely need to mention that this division rests upon [an] abstraction; [further,] that the word can have no place

outside of thought, and neither can its object, if it is non-physical … This remains so even in the case of material objects, *since it is never the objects directly, but always only the representation of them provided by the word, that is made present in the soul*" (Humboldt, 1903–1936, V, p. 418).

It is important to note that by introducing the useful distinction between the referent of the word and its meaning, Humboldt also excludes every possibility of a pure designating-relation between a name and its object, which was the paradigm for the traditional explanation of the functioning of language. He generalizes the meaning/reference distinction as applicable to all linguistic signs. Accordingly, Humboldt extends this distinction even into the realm of names, including proper names. He continues: "The word conceives of *every concept as general*, always designating, strictly speaking, classes of reality, *even if it is a proper name*; for thus it comprehends … all the various states, with respect to time and space, of that which it designates" (Humboldt, 1903–1936, V, p. 419).

Thus, Humboldt counteracts the problematic reduction of predicates to mere names in the traditional view with the inverse reduction, namely the assimilation of names to predicates. This move can be considered the systematic point from which the problem of linguistic relativism arises. If this is not immediately evident, it soon becomes clear by analyzing its implications. According to this assumption, names are actually general concepts. Instead of designating an object, they too, by strict analogy with predicates, can relate only in a mediated fashion to their referent, as if to an object falling under a concept. Humboldt explains: "the picture of language as merely designating *objects*, already perceived in themselves, is also not sustained by examination of what language engenders as its product … Just as no concept is possible *without language*, so also *there can be no object for the mind*, since it is only through the concept, of course, that anything external acquires full being for consciousness. But the whole *mode of perceiving* things *subjectively* necessarily passes over into cultivation and the use of language. For the word arises from this very perceiving; it is a copy, not of the object in itself, but of the image thereof produced in consciousness" (Humboldt, 1903–1936, VII, p. 60). Seen in this light, the idealistic conclusion that Humboldt draws from this assumption seems entirely consistent: namely that "man [lives] with the objects … *exclusively* in the way that they are conveyed to him by language" (ibid.). This in turn explains why different languages are in fact different world-views. As Humboldt remarks in a similar context: "In this way, in words of identical meaning that are found in a number of languages, there arise diverse representations of the same object, and this characteristic of the word contributes chiefly to the fact that *every language yields a peculiar world-view*" (Humboldt, 1903–1936, V, p. 420).

However, once the diversity of conceptual systems or world-views contained in the different languages is recognized, it is no longer clear how speakers of different linguistic communities can talk about the same things and thus communicate successfully. Humboldt has to admit that "every language places *definite boundaries* upon the spirits of those who speak it, and insofar as it provides a determinate orientation, excludes others" (Humboldt, 1903–1936, VII, p. 621). The commensurability of different linguistic world-views does not become a problem for Humboldt because he has any skeptical doubts about the possibility of translation between languages. Actually, Humboldt remarks repeatedly that "experience with translation of the most diverse languages … shows that *it is possible to express every concatenation of ideas in every language*" (Humboldt, 1903–1936, IV, p. 17). Instead, the problem of linguistic relativism originates in the fact that through his criticism of the traditional conception of language Humboldt has destroyed the plausibility of the traditional answer to this problem. His conception of language as world-disclosing precludes the possibility of explaining interlinguistic understanding by appealing to a world in itself accessible to all human beings in the same way and to which all speakers refer through the different conventions of their languages.

Humboldt is aware of this challenge and indeed tries to articulate an alternative answer to the question of interlinguistic understanding. His answer is also based on a criticism of the traditional view of language, but this time from a different angle. Humboldt opposes the reduction of language to a mere system of signs entailed in the traditional view. In the introduction to his work on the Kawi language, Humboldt expresses this opposition with the help of one of his most famous claims, namely, that language is not a product (*ergon*) but an activity (*energeia*). To take this claim seriously implies transforming the traditional conception in two important respects.

First, in the sense already mentioned, Humboldt's view of language as an activity underscores the fact that language does not passively mirror the world but it is actually world-disclosing. Language is what makes the world appear as an ordered whole. For it is through the web of linguistic concepts that what would otherwise be a disconnected flux of perceptions is transformed into an order of objects properly so called. In this sense, Humboldt claims that "the generation of language is, from the beginning, a *synthetic process* in the most genuine sense of the term, *where the synthesis creates something that did not exist before in any constituent part*" (Humboldt, 1903–1936, VII, p. 95). Humboldt terms this genuine synthesis "articulation" and regards it as "the authentic essence of language, the lever *through which language and thought are produced*, the culmination of their intimate mutual connection" (Humboldt, 1903–1936, VI, p. 153).

However, there is a second, more radical sense in which Humboldt sees language as an activity. According to Humboldt, the synthetic process through which both language and thought are produced is not something given once and for all, that is, fixed through the grammar and lexicon of a language. It is rather the result of praxis, namely the process of communication between speakers. As he remarks repeatedly: "Language only exists in ongoing speech; grammar and lexicon are hardly comparable to their dead skeleton" (Humboldt, 1903–1936, VI, p. 148). Humboldt opposes the reduction of language to a mere vehicle of thought that underlies the analysis of language in logic. In this context, language is considered in abstraction from the use of language in speech for the purposes of analyzing the general structure of thoughts. But Humboldt's opposition to this abstraction is not based on the obvious fact that language, in addition to being a vehicle of thought, is also a means of communication. His view is rather that language can only be a vehicle of (intersubjective) thought to the extent that it makes communication and understanding between speakers possible. Thus, a proper analysis of the cognitive

function of language is impossible if language is considered in isolation from its communicative function. Behind Humboldt's claim is the idea that the different communicative perspectives, generated through the system of personal pronouns present in all languages, are the condition of possibility of any intersubjectively valid thought. For these different perspectives make it possible that speakers both express their views in their irreducible singularity and achieve a common understanding through conversation as well. Thus, the commensurability of linguistic world-views is not guaranteed in advance by a unique syntax or lexicon; it is actually the result of a different feature of all languages, namely its dialogical structure. It is in virtue of the pragmatic structure of speech that a free synthesis of perspectives among speakers can come about in conversation.

Humboldt is fascinated by the double nature of language as the *source* of both *generality* and *individuality*. Due to its general character, language makes it possible that different speakers share identical concepts (and in this way fulfills a socializing function). But due to its expressive potential, it also allows speakers to maintain their genuine individuality (i.e., it makes individualization possible). Humboldt explains that "in language the *individualization* within a *general conformity* is so wonderful, that we may say with equal correctness that the whole of mankind has but one language, and that every man has one of his own" (Humboldt, 1903–1936, VII, p. 51). However, there is an obvious tension in this view. On the one hand, Humboldt claims that in virtue of its intrinsic generality (the fact that we can share it), language entails the condition of possibility of understanding among speakers. It guarantees, in Humboldt's words, that "nobody may speak differently to another from the way in which the latter, under similar circumstances, would have spoken to him" (Humboldt, 1903–1936, VII, p. 47). But on the other hand, the possibility of understanding among speakers is not a mere function of a shared grammar or lexicon. Given that the only reality of language is speech, genuine understanding can only be the lucky outcome of the process of communication itself, in which different individuals express their individuality or particular world-view. For "only in the individual does language get its last determination … A nation as a whole has indeed the same language, but not all the individuals in it have … exactly the same one, and if one looks more carefully, every person actually has his own" (Humboldt, 1903–1936, VI, p. 183). But given that speakers' linguistic differences are actually differences in world-view, it is not clear within Humboldt's account how speakers can talk about the same things despite their genuinely different linguistic world-views. Without such an explanation the possibility of understanding in dialogue remains puzzling. Humboldt has to recognize that "nobody means by a word precisely the same as his neighbor does, and the difference, be it ever so small, vibrates, like a ripple in water, throughout the entire language. Thus, all understanding is always at the same time a not-understanding, all agreement in thought and feeling at the same time a divergence" (Humboldt, 1903–1936, VI, p. 183). The tension present in Humboldt's writings between the roles ascribed to language in its world-disclosing and in its communicative dimensions remains a characteristic feature of the works of most of the authors of this tradition.

Further Developments in the Twentieth Century

M. Heidegger is undoubtedly the most influential defender of Humboldt's conception of language as world-disclosing in twentieth-century philosophy. However, despite the deep similarities between the conceptions of language developed by both authors, Heidegger was rather critical of Humboldt's approach to language at different times during his philosophical development and for different reasons (see Heidegger, 1927 and 1986). The view of language as responsible for world-disclosure is not articulated explicitly in Heidegger's early writings, although some features of what would be his later conception of language already appear in them, including in *Being and Time* (see Lafont, 2000). Most of Heidegger's reflections on language are contained in a collection of articles that appeared in English under the title *On the Way to Language*.

As is characteristic of all the authors in this tradition, Heidegger articulates his view of language as world-disclosing in direct opposition to the traditional conception of language as a mere tool for the designation of objects that exist independently of language. He argues that the linguistic activity of naming something with a word cannot be understood as a pure relation of designation between a name and an entity. It does not consist in the fact "that something already familiar to us is provided with a name, rather … through this naming, the entity is first nominated as that which it is. It is in this way that it is known as an entity" (Heidegger, 1944). In *The Origin of the Work of Art* Heidegger explains: "When language names the entity for the first time, such naming alone brings the entities to word and to appearance. This naming nominates the entity to its being from out of this being. Such saying is a projecting of the clearing, in which it is announced as what the entity comes into the open" (pp. 59–60). In *On the Way to Language* he expresses the same idea in a much more provocative way using a line of Stefan George's poem, namely, that "there is no thing where the word is lacking." Heidegger explains: "The thing is a thing only where the word is found for the thing. Only in this way *is* it … The word alone supplies being to the thing, [for] … something only *is*, where the appropriate word names something as existing and in this way *institutes* the particular entity *as such* … The being of that which is resides in the word. For this reason, the following phrase holds good: language is the house of being" (pp. 164–166).

As Humboldt had already argued entities can only be identified as such once they are linguistically categorized. It is in this sense that language is responsible for world-disclosure: language determines our possible experience with entities and thus determines in advance what kind of world is disclosed to us. With regard to the world-disclosing power of language, Heidegger basically agrees with Humboldt. However, on this very same basis he criticizes Humboldt's view of language as an activity. As mentioned previously, Humboldt tried to conceive the production of a linguistic world-view as a result of the human activity of speech. This would explain how speakers can synthesize their different world-views and revise them through dialogue. However, as Heidegger argues, in order for a dialogue to be possible at all, it is first necessary that one and the same world be disclosed to all speakers. As he had already claimed in *Being and Time*, "only he who already understands can listen." Heidegger interprets Humboldt's

emphasis on the communicative dimension of language as just a reminiscence of the instrumental conception of language. Heidegger directs one of his most provocative claims against any attempt to transfer language's power of world-disclosure to the human activity of speech. He writes: "Man acts as though he were the shaper and master of language, while in fact language remains the mistress of man … For strictly, it is language that speaks. Man first speaks when, and only when, he responds to language by listening to its appeal" (Heidegger, 1954, p. 184).

As a result of this extremely idealistic conception of language, Heidegger sees the world-disclosure, which is anchored in the language of a specific people or culture, as the most significant event of their historical development (see Heidegger, 1980). Based on this idea Heidegger interprets the history of Western culture as the fateful development of the understanding of being as presence, which is disclosed in Western languages and may bring this culture toward its own destruction. From a Humboldtian point of view, the reification of language implicit in Heidegger's view seems just a consequence of ignoring that language is not a product but an activity, that is, that language is nothing other than the communicative activity between human beings. However, it is hard to question Heidegger's criticism of this Humboldtian claim, given that for both authors language has the power to constitute our experience in the strict sense set by transcendental philosophy. So long as this presupposition is not questioned, achieving the balance sought by Humboldt between the world-disclosing and the communicative dimensions of language remains an unresolved task in this tradition.

H.-G. Gadamer, Heidegger's most influential student in contemporary German philosophy, tries to achieve this balance in his main work *Truth and Method*. In this work, Gadamer articulates the essential features of the approach to philosophy that he calls "philosophical hermeneutics." At the core of this approach is an analysis of hermeneutic experience, that is, of the interpretative relationship that human beings entertain with each other and with their symbolic products (e.g., texts, works of art, and the like). To support his account of hermeneutic experience, he appeals to the conception of language developed in the Hamann–Herder–Humboldt tradition, to which he wishes to contribute with his own reflections on language. He uses the Humboldtian model of dialogue between speakers who try to reach agreement about something in the world and generalizes it as the appropriate model for describing any hermeneutic experience whatsoever (e.g., the understanding of canonical texts, works of art, and any other symbolic products by an interpreter, most significantly in the context of the human sciences). In this respect, Gadamer explicitly focuses on the communicative dimension of language. However, he looks at the Humboldtian model of dialogue with Heideggerian eyes.

Gadamer emphasizes that a genuine dialogue by its very nature is not just the product of the speakers' activity. First of all, as Gadamer remarks, "no one knows in advance what will come out of a conversation. Understanding or its failure is like an event that happens to us" (Gadamer, 1960, p. 383). But more important, a conversation is only possible if the participants share a linguistically disclosed world. Thus, a shared world-disclosure, that is, a common tradition, is the precondition for any understanding or agreement that speakers try to bring about in conversation. Although Gadamer opposes, as much as did Humboldt, the thesis of the incommensurability of different linguistic world-disclosures, he has never seriously confronted the challenge of explaining how intercultural understanding is possible, that is, how a dialogue between participants who do not belong to the same tradition and thus do not share the same linguistic world-disclosure is possible at all.

More recently K.-O. Apel and J. Habermas have tried to work Humboldt's and Gadamer's model of dialogue into a systematic theory of communication (see Apel, 1994 and Habermas, 1981). Both authors identify in the communicative function of language a specific kind of rationality (communicative rationality) that allows participants in communication to reach a rational agreement about something in the world. Although they share the view of language as world-disclosing that characterizes the continental tradition, their approaches introduce a strong dichotomy between form and content that aims to conceive rational understanding as a universal possibility (and thus to avoid the problem of incommensurability between different linguistic world-disclosures). In clear opposition to Heidegger and Gadamer's claim that understanding is only possible on the basis of a factual agreement among speakers with a shared linguistic world-disclosure, Apel and Habermas claim that understanding depends only on a "counterfactual agreement" that all speakers share just in virtue of their communicative competence. This agreement is based on formal presuppositions and thus it does not depend on a shared content or shared world-disclosure among the participants in conversation. A rational reconstruction of these formal presuppositions of communication is the core of the analysis of language that these authors offer in their respective approaches. Apel understand this analysis as a transformation of transcendental philosophy within the linguistic paradigm and thus calls it "transcendental semiotics." Habermas disagrees with Apel's claim that the formal presuppositions of communication have a transcendental status in the strict Kantian sense; thus he prefers the designation of "formal pragmatics" for his approach. But beyond these methodological differences, their reconstructions of the formal presuppositions of communication are very similar.

According to both approaches, speakers who want to reach an agreement about something in the world have to presuppose the truth of what they are saying, the normative rightness of the interaction they are establishing with the hearer through their speech acts, and the sincerity or truthfulness of their speech acts. Complementary to these three validity claims (truth, normative rightness, and truthfulness), speakers share the notion of a single objective world that is identical for all possible observers. This is just a formal notion that is freed of all specific content. It is important to notice that this presupposition does not correspond to the naïve belief in a world in itself accessible to all speakers that characterizes the traditional conception of language. As Habermas argues, if speakers had such a naïve belief they would identify the *contents* of their respective linguistic world-views with the assumed world-order itself. But with such a dogmatic attitude they would not admit the possibility that in case of disagreement the claims raised by other speakers may turn out to be right about the way the world

is. In order for a rational discussion to be possible at all, speakers have to be able to adopt a hypothetical attitude vis-à-vis their own beliefs about the world. And this is only possible if they distinguish, however counterfactually, between the assumed world-order itself and their respective claims about how the world is. Seen in this light, the formal presupposition of a single objective world is just a consequence of the universal claim to validity built into the speakers' speech acts. It is just an expression of the communicative constraint that makes rational criticism and mutual learning possible, namely that from two opposed claims only one can be right. Thus, the formal notion of world and the three universal validity claims build a system of coordinates that guides the interpretative efforts of the participants in communication toward a common understanding, despite their differences in beliefs or world-views.

It is an open question whether this proposal is actually successful in explaining the possibility of interlinguistic understanding, as much as it is open whether any current theory within contemporary philosophy of language is so successful. However, Apel's and Habermas's approaches are intentionally designed to meet this challenge in a way unprecedented in the prior continental philosophy of language. Equally unprecedented is the scope and sophistication of the philosophical analysis of language and communication that these authors offer. This is in part due to the reception and creative integration in their own work of most of the approaches developed within contemporary philosophy of language across the borders of the different philosophical traditions.

See also: Generative Grammar; Hermeneutics; Kant, Immanuel (1724–1804); von Humboldt, Wilhelm (1767–1835).

Bibliography

Apel, K.O., 1963. Die Idee der Sprache in der Tradition des Humanismus von Dante bis Vico. Bouvier Verlag, Bonn.

Apel, K.O., 1973. Transformation der Philosophie, 2 vols. Suhrkamp Verlag, Frankfurt (English trans.: Towards a Transformation of Philosophy, Routledge & Kegan Paul, London, 1980).

Apel, K.O., 1994. Selected essays. In: Towards a Transcendental Semiotics, vol. 1. Humanities Press International, New Jersey.

Apel, K.O., 1998. From a Transcendental-Semiotic Point of View. Manchester University Press, Manchester/New York.

Gadamer, H.G., 1960. Wahrheit und Methode. J C B Mohr, Tübingen (English trans.: Truth and Method, Continuum, New York, 1994).

Gadamer, H.G., 1986. Hermeneutik II, Gesammelte Werke, vol. 2. Möhr, Tübingen.

Habermas, J., 1981. Theorie des kommunikativen Handelns. Suhrkamp Verlag, Frankfurt (English trans.: Theory of Communicative Action, 2 vols, Beacon Press, Boston, 1984/87).

Habermas, J., 1984. Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns. Suhrkamp Verlag, Frankfurt.

Habermas, J., 1998. On the Pragmatics of Communication. MIT Press, Cambridge, MA.

Habermas, J., 1999. Hermeneutic and analytic philosophy: two complementary versions of the linguistic turn? In: O' Hear, A. (Ed.), German Philosophy since Kant. Cambridge University Press, Cambridge.

Habermas, J., 2003. Truth and Justification. MIT Press, Cambridge, MA.

Hamann, J.G., 1784. Metakritik über den Purismus der Vernunft. In: Hamann 1988, pp. 205–212.

Hamann, J.G., 1949–1957. Sämmtliche Werke. In: Nadler, J. (Ed.), 6 vols. Vienna.

Hamann, J.G., 1988. Vom Magus im Norden und der Verwegenheit des Geistes. In: Majetschak, J. (Ed.), Ein Hamann Brevier. Deutscher Taschenbuch Verlag, Munich.

Heidegger, M., 1927. Sein und Zeit. Max Niemeyer Verlag, Tübingen (English trans.: Being and Time, Harper & Row, New York, 1962).

Heidegger, M., 1944. Hölderlin und das Wesen der Dichtung. In: Erläuterungen zu Hölderlins Dichtung. Klostermann, Frankfurt.

Heidegger, M., 1954. Vorträge und Aufsätze. G. Neske, Pfullingen.

Heidegger, M., 1980. Der Ursprung des Kunstwerkes. In: Holzwege. Klostermann, Frankfurt.

Heidegger, M., 1986. Unterwegs zur Sprache. G. Neske, Pfullingen (English trans.: On the Way to Language, Harper & Row, New York, 1971).

Herder, J.G., 1960. In: Heintel, E. (Ed.), Sprachphilosophische Schriften. Felix Meiner Verlag, Hamburg.

Herder, J.G., 1799. Eine Metakritik zur Kritik der reinen Vernunft. In: Herder 1960, pp. 183–227.

Humboldt, W.V., 1903–1936. Gesammelte Schriften. In: die königlich Preussischen Akademie der Wissenschaften, 17 vols. B. Behr's Verlag, Berlin.

Humboldt, W.V., 1963. Schriften zur Sprachphilosophie. In: Flitner, A., Giel, K. (Eds.), Of Werke in Fünf Bänden, vol. 3. Klett-Cotta Verlag, Darmstadt.

Humboldt, W.V., 1988. On Language. The Diversity of Human Language-Structure and Its Influence on the Mental Development of Mankind. Cambridge University Press, Cambridge.

Lafont, C., 1999. The Linguistic Turn in Hermeneutic Philosophy. MIT Press, Cambridge, MA.

Lafont, C., 2000. Heidegger, Language, and World-Disclosure. Cambridge University Press, Cambridge.

Taylor, C., 1985. Theories of meaning. In: Human Agency and Language: Philosophical Papers. Cambridge University Press, Cambridge, pp. 248–292.

Trabant, J., 1990. Traditionen Humboldts. Suhrkamp Verlag, Frankfurt.

Continuing Professional Development

Helen S Timperley, Faculty of Education, School of Learning, Development and Professional Practice, The University of Auckland, Epsom, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Challenges in education are rising with expectations from policy makers about what teachers can do to meet them constantly increasing. Continuing professional development for teachers has been promoted as the multimillion-dollar solution, but unfortunately most approaches have little demonstrated its impact on teachers' ability to meet these demands. This article draws on the international literature that does provide evidence of effectiveness in terms of having a positive impact on student learners. The article outlines some of the shifts in thinking that underpin the effective enactment of a systematic inquiry and knowledge-building approach to continuing professional learning together with evidence of improvement in students' engagement and learning.

Introduction

The 'face' of education is changing across many parts of the world: new curricula are appearing with accompanying new approaches to assessment. New technologies are pervading teaching and learning spaces with student learners often outstripping the skill levels of their teachers. In many educational jurisdictions, the composition of the student body is changing along with expectations that all students will learn, not just those who respond readily to the teaching offered. Raising the bar and closing the gap has become a major concern in many countries (OECD, 2010) with continuing teacher professional learning as the multimillion-dollar solution.

The rationale underpinning this solution is that teaching matters, and therefore improving the quality of teaching through professional development will solve the problems. A consistent finding of meta-analyses investigating the effects of education on student outcomes indicate that teacher effects are much larger than school effects whatever the educational jurisdiction (e.g., Bransford et al., 2005; Muijs and Reynolds, 2001; Nye et al., 2004; Scheerens et al., 1989). The case for focusing on teachers and teaching is clearly articulated into a study by Wright et al. (1997):

> The results of this study will document that the most important factor affecting student learning is the teacher. In addition, the results show wide variation in effectiveness among teachers. The immediate and clear implication of this finding is that seemingly more can be done to improve education by improving the effectiveness of teachers than by any other single factor.
>
> p. 63

By 2005, the growing international agreement about the new evidence base informed the landmark OECD (2005a) report 'Teaching Matters.' In the accompanying influential OECD Indicators Report (2005b), the policy importance of both pedagogy and professional development were highlighted: "at the level of the education system, professional development of teachers is a key policy lever" (p. 20).

The hope generated by this research into the importance of teaching has rarely been matched by the impact of continuing professional development to improve it. After outlining the importance of teacher quality, Hanushek (2005), for example, rejected continuing teacher education as the key policy lever because "despite some success in general they [continuing professional development programs] have been disappointing" (p. 19), and "existing evidence on in-service programs gives insufficient means for selecting a program that is likely to yield significant gains in teaching performance" (p. 19).

This issue is illustrated by the contrast between the international enthusiasm for, versus the practice of, assessment for learning. Since the Assessment Reform Group in England coined the term 'Assessment for Learning' in 1998, it has become part of many educational systems around the world. A quick Google search yields over 11 million hits; countless books have been written about it; and it has become a stalwart of assessment policy statements and professional development in many educational jurisdictions. Despite this wide acceptance, there is research evidence that assessment for learning is either absent or superficial in most classrooms with little of the desired change evident in many learning environments. The Learning How to Learn Project in England (James et al., 2007), for example, found that the participating teachers' implementation often reflected what they called the 'letter' of assessment for learning, focusing on the surface techniques, rather than the 'spirit,' based on a deep understanding of the principles underlying the practices. Only about 20% of the teachers in the study were using assessment for learning in ways that were designed to help students develop as learners.

The impact of continuing professional development on reducing the gaps between those who traditionally succeed in our education systems and those who form the core of the rapidly rising school populations is particularly concerning. The recent report on equity by the OECD (2010) reinforces what has become a consistent finding over time that many schooling systems and the teaching that occurs within them serve to maintain rather than reduce these inequalities despite all the efforts to change them. Many students are not learning what they need to learn despite interventions to help their teachers to teach more effectively. On the other hand, there are many documented instances where teachers and schools are making this difference through their involvement in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92134-2

continuing professional development. What is needed, therefore, is to draw on the evidence of the approaches that have demonstrated effectiveness to rethink the typical approaches to professional development for teachers.

The Research Base

The primary evidence base for this article is a synthesis of the international literature on approaches to continuing professional development that have resulted in changes in practice and improved outcomes for student learners (Timperley et al., 2008). The synthesis began with the development of a theoretical framework to identify the possible attributes of continuing professional development likely to result in improved outcomes for student learners. The framework included attributes of the wider educational context (e.g., formal educational policies), the professional learning context (e.g., people and practicalities), the content of the professional learning opportunities (e.g., a focus on curriculum or pedagogies), the activities constructed to promote learning of the content (e.g., coaching), the professional learning processes promoted (e.g., creating dissonance with current understanding), the responses of diverse teacher learners (e.g., high/low levels of teacher engagement), and the impact on diverse learners including social, personal, and academic outcomes.

Each attribute was treated as neutral with regard to its impact on teaching practice or student learners until it was tested against the empirical evidence provided through the published literature which documented relationships between continuing professional learning, changes in teaching practice, and improved outcomes for learners. The application of standard methodological criteria identified 72 individuals and groups of related studies of independent interventions with a total of 227 effects (Cohen, 1988). The reported attributes of the professional development in these studies were then mapped onto the theoretical framework to identify the kinds of teacher knowledge and the circumstances under which it was acquired that were associated with benefits to students. Identification of the attributes in each category was then combined into a model that captured the main characteristics of continuing professional development that resulted in positive outcomes for student learners.

The model was further developed, assessed, and refined through its application in a professional learning project involving 300 schools. It was found to be highly effective in improving outcomes for learning, particularly those not well-served by our education systems. This project focused on improving literacy in separate cohorts of approximately 100 schools per cohort over 6 years. Repeated patterns of improvement beyond expected gains in reading and writing, particularly for the lowest achieving students, were evident in each cohort (Timperley et al., 2010). After taking into account the average expected gain, the average effect sizes for the final cohort as a whole was 0.44 for reading and 0.88 for writing on a nationally standardized test. This equates to a rate of progress 1.85 times greater than expected nationally for students in schools with a reading focus, and 3.2 times the expected improvement for those in schools with a writing focus. The rate of progress for those students beginning in the lowest 20% was even larger, with an effect size of 1.13 for reading and 2.07 for writing. These gains equate to progress of 3.2 times nationally expected gains for the lowest 20% of students for reading and 6.2 times expectation of students for writing. The effect sizes were calculated using Cohen's d (1988) with Hedge's correction.

The Model

Most empirical studies analyzed in the synthesis that reported significant differences to student outcomes began by involving teachers in cycles of inquiry that asked them to identify what students needed to know and do to meet curricula and other goals valued by the communities in which students live and are educated (see **Figure 1**). When profiles of students' achievement were well understood, teachers then inquired into the knowledge and skills they needed if they were to become more effective in addressing the learning needs of individuals and groups of students, particularly those not achieving as well as others. From there, teachers engaged in new professional learning that deepened their knowledge and refined their skills so they could engage students in new learning experiences in classrooms. Given that the effectiveness of all teaching practice is influenced by context and no particular practices can be guaranteed to result in particular outcomes, the final stage of the inquiry involved examining the impact of changed actions on the outcomes for students that were the focus of the inquiry. Through this process teachers came to understand what had been effective and what had not. Those studies demonstrating sustainability of improvement in teaching and outcomes for students used the findings from this examination of impact to inform ongoing cycles of inquiry (Fennema et al., 1993; Lai et al., 2009; McNaughton et al., 2004; O'Connell, 2009; Schmitt et al., 2005).

Each of these dimensions of the cycle will be described in greater detail in the following section, but before doing so, some assumptions underpinning the model and shifts in thinking usually required for continuing professional development to make the difference are outlined.

Shifts in Thinking to Make the Difference

The first assumption underpinning the model involves a shift in thinking from continuing professional development to continuing professional learning. Both are intentional, ongoing, and systematic processes (Guskey, 2000). Over time, however, the term 'professional development' has taken on connotations of delivery of some kind of information to teachers so as to influence their practice, whereas 'professional learning' implies an internal process in which individuals create professional knowledge through interaction with this information in a way that challenges previous assumptions and creates new meanings.

The second shift in thinking highlights the need for professional learning to be underpinned by an explicit and defensible theory of professionalism. Those who require or offer teachers discrete opportunities to participate in professional

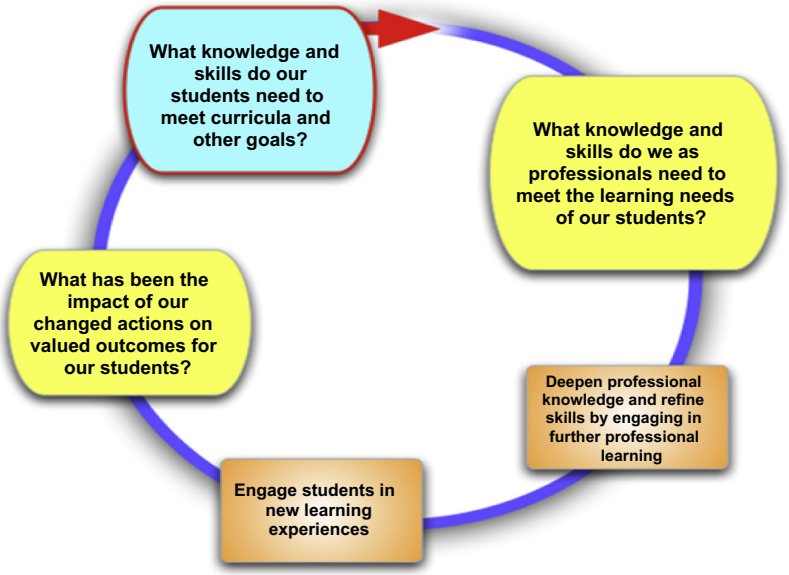

Figure 1 Teacher inquiry and knowledge-building cycle to promote valued student outcomes.

development workshops, courses, or conferences, typically come from a vision of teaching as one of routine expertise. Learning for routine experts involves finding out how to apply a core set of skills and routines with fluency and efficiency. Notions of routine expertise are based on assumptions that novice teachers become experts through supported practice (Dall'Alba and Sandberg, 2006). This perspective is best summed up as 'practice makes perfect,' where skill development takes place in a stepwise, cumulative manner. Becoming a highly competent professional, therefore, involves progressively developing a set of knowledge and skills relevant to that profession (e.g., Dreyfus and Dreyfus, 1986) in which problems are identified intuitively and holistically with appropriate strategies enacted to solve them.

Professional learning that impacts positively on student engagement and learning, described in this article, is underpinned by an alternative vision of teachers being adaptive experts (Hammerness et al., 2005; Feiman-Nemser, 2008; Hatano and Inagaki, 1986). Timperley (2012) defined adaptive experts as those focused on the moral imperative of promoting the engagement, learning, and well-being of each student for whom they have responsibility. To achieve this outcome they are aware of the assumptions underpinning their practice, including their cultural positioning, and know when these assumptions are helpful for their learners and when to question them, and if necessary, to let them go. Fundamental to doing this is the active seeking of deep knowledge about both the content of what is taught and how to teach it effectively for their learners in particular contexts. In their practice they work with others, including their student learners, to:

- Retrieve, organize, and apply professional knowledge in light of the challenges and needs presented by the learners they are teaching, particularly those not engaged.
- Seek evidence about the impact of their teaching on learners' engagement, learning, and well-being.
- Seek to develop innovative approaches when regular routines are not working for their learners and know when they need to seek help.
- Engage in ongoing inquiry for building knowledge as core to their professionalism.

Rather than continuing professional development being seen as a series of events in which teachers participate, it is seen as an active process of systematic inquiry into the effectiveness of practice to improve student engagement, learning, and well-being. From this perspective, the 'continuing' part of professional learning means 'all the time' throughout a teacher's career.

The third shift in thinking involves seeing student learners at the center of the process. Improvements in outcomes for students are not a by-product of professional learning but rather its central purpose. Student learners must be the touchstone and the reason for teachers to engage, the basis for understanding what needs to be changed and whether those changes have been effective. Part of this shift involves creating mindsets that have at their core a belief that schooling is about ensuring deep learning for all.

The fourth shift in thinking involves foregrounding the knowledge and skills of focus rather than forms or delivery methods for continuing professional learning. Many studies analyzed in the best evidence synthesis on professional learning and development (Timperley et al., 2008) found that a great deal of emphasis was given to descriptions of the details of the activities in which teachers engaged rather than what they learned. No activity or process, whether facilitated by others (e.g., coaching, modeling, and engaging with professional readings) or self-directed (e.g., discussing mutually identified problems, reflection, and inquiry), was consistently associated with improved student success. What was important was the depth of knowledge and skills learned as a result of engagement in the activity or process.

The final shift directs attention to those who support teacher learning within schools or outside of them. Teachers alone cannot solve entrenched problems within our education system; so everyone who has a place in the chain of influence from policy to practice needs to engage in inquiry and knowledge-building cycles to ensure their efforts are effective in developing the kind of professional and student learning that makes a difference. Just like teachers, those responsible for promoting professional learning need to engage in ongoing inquiry into the impact of their policies and practices. This impact is not always positive (Coburn et al., 2009; Datnow, 2005). Fullan (2005) suggests that successful problem solving requires the whole system to be involved in codependent partnerships. Nowhere is this more important than in the area of professional learning. Successful problem-solving involves a process of learning both up and down the system layers.

The remainder of this article unpacks the model developed from the synthesis of the research literature (see **Figure 1**) together with its application in practice and highlights where the shifts described above are particularly important. The whole cycle encapsulates the shifts from professional development to professional learning that involves an active process of systematic inquiry to develop adaptive expertise. Other shifts are particularly relevant to specific dimensions.

Identifying Students' Knowledge and Skills

In order to have an impact on student outcomes, the inquiry and knowledge-building cycle (**Figure 1**) begins and ends with students as described in the third shift in thinking outlined above. Teaching is a highly contextualized activity with teachers constantly responding to the students in front of them. Continuing professional learning that focuses on new practices decontextualized from the immediate demands of student learning in a given teacher's class is not likely to be translated into that environment. Teachers might find the information interesting but rarely apply it given the competing and immediate demands of their students and the curriculum to be taught when they return.

The process of unpacking the question in **Figure 1** "What knowledge and skills do our students need to meet curricula and other goals?" involves asking some specific additional questions that might include:

What do they already know?
What sources of evidence have we used, and how adequate are they?
What do they need to learn and do?
How do we build on what they know?

To answer these questions, leaders and teachers need deep knowledge of how to assess students for the purposes of diagnosis, particularly for those students who do not respond readily to existing instructional practices. For many leaders and teachers this is the first phase of their professional learning as to how to undertake such assessments and how to interpret the information to identify teaching/learning challenges. In many cases, it involves taking a whole new perspective on the purposes of assessment as one of professional inquiry to identify professional learning needs rather than one of grouping, labeling, or credentialing students.

Identifying Teachers' Knowledge and Skills

This second part of the cycle asks teachers to identify what it is they need to know and do to be more effective, particularly with those students achieving less well than others and is central to the aspect of adaptive expertise that asks teachers to become aware of the assumptions underpinning their practice and know when these assumptions are helpful for their learners and when to question them, and if necessary, to let them go. Actively inquiring into and identifying their own learning needs means their further engagement is motivated by the need to know. Much professional development about effective teaching practices is motivated by someone else's desire to tell teachers what they should learn and do.

In the desire to tell approach, policy makers, researchers, or professional development providers believe or have evidence that some kinds of teaching practices are more effective than others. They then create professional development opportunities to inform teachers about these practices without creating the need to know beyond compliance or teacher interest. Some leaders address issues of motivation by calling for volunteers. The synthesis of the evidence on professional learning and development found that student outcomes did not depend on whether teachers volunteered or were expected/directed to be involved (Timperley et al., 2008). In some situations the expectations of volunteers were not met, particularly when the extent of change was greater than they had anticipated. In some other situations the expectations of those who were directed were exceeded. What was more important than initial volunteering was that teachers became engaged in the process at some point. Most engaging was the opportunity to solve puzzling teaching and learning problems related to individuals or groups or learners in the teachers' classes.

The process of identifying one's own professional learning needs is not easy and usually involves the assistance of others because it is difficult to step outside one's own frame of reference. The 'others' may be a school leader or outside facilitator. Specific questions for teachers to answer with a professional learning facilitator might involve:

How have we contributed to existing student outcomes?
With whom are we most effective?
With whom are we not connecting and why?
What do we already know that we can use to promote better outcomes?
What do we need to learn and do to promote better outcomes?
What sources of evidence/knowledge can we utilize?

To answer these questions, teachers need to link their findings from the first dimension of the cycle into students' profiles of achievement to specific teaching practices. It involves taking teachers through a process of unpacking their preconceptions about how students learn and how best to teach them. Why are particular aspects of practice valued? Is what is valued most effective? In order to move on, some key assumptions underpinning existing practice and how they fit with new practice may need to be challenged. Spillane (1999) found, when investigating a mathematics reform in the United States, for example, that if teachers "… are to get to the core reform ideas, [they] have to question, unlearn and discard much of their current, deeply rooted

understandings of teaching, learning and subject matter" (p. 154). Without this kind of examination, suggestions for new practice are likely to be forgotten if they do not fit existing beliefs, or to be assimilated into existing knowledge and practice without a deep understanding what it is that needs to change and why (Hammerness et al., 2005). Hammerness and colleagues refer to this process as one of over-assimilations whereby teachers make superficial changes believing they are making appropriate adjustments to their practice, but do not really understand what this involves, and hence are unable to do so.

This process of examining and critiquing beliefs underpinning particular practices shifts the spotlight from student learners to the teachers. Here, leadership is central because the process is as much about emotions as the acquisition of new knowledge (Hargreaves, 1998). Expectations to change may touch raw nerves because they are likely to impinge on teachers' sense of professional identity and competence while they ask themselves, "Has what I have done before been ineffective?" Stoll et al. (2003) describe how "neglecting emotions can close people to learning, and lead teachers to behave defensively to protect themselves from situations that they feel might expose their 'inadequacies'" (p. 85). Elmore (2004) notes one of the strongest social norms in schools is that everyone is expected to pretend that they are equally effective in what they do even when they feel they are unable to do it. Having to pretend is antithetical to professional learning.

While **Figure 1** implies a sequential identification of student' and teacher' learning needs, in practice this process and all others in the cycle may occur alongside one another. Engagement of teachers' initial understandings occurs throughout the professional learning experience as information is understood more deeply and challenges arise. Learning is iterative rather than sequential.

Deepening Professional Knowledge

This dimension of the cycle foregrounds new knowledge and skills as outlined in the fourth shift in thinking above. It is referred to as "*deepening* professional knowledge and *refining* skills by engaging in further professional learning" because it is assumed that the teachers involved have some experience rather than starting as novices. Professional development events do little more than raise teachers' awareness of new information and rarely lead to changes in practice because much deeper learning is required to change practice sufficiently to make a difference to learner outcomes.

The knowledge identified as important in the synthesis (Timperley et al., 2008) typically involved the integration of knowledge of assessment, the curriculum and relevant aspects of the underlying discipline, and how to teach it. It is only when this knowledge is understood within an overarching conceptual framework of teaching and learning that teachers are able to retrieve and apply it in the moment-by-moment decisions they need to make every day in classrooms.

As with other professions, teaching is practice based; hence, new knowledge is usually insufficient on its own to create change. Ways to translate knowledge into practice needs also to be attended to so that the theoretical understandings become the basis for making ongoing principled adjustments to practice in response to students' learning needs.

When substantial change is required, as is usually the case to make a difference to outcomes for student learners, teachers need multiple opportunities to learn. Through this process, partially understood ideas are revisited as they are tried out in classroom contexts. Over time, understandings are deepened and knowledge becomes more easily retrieved, and skills enacted in the face of daily classroom challenges. Such opportunities to learn require extended time for changes to become a part of the routine of practice.

The specific approaches developed to acquire the relevant knowledge and skills depend partly on the level and kind of professional learning required. Teachers are as diverse in their knowledge and beliefs as their students. If the changes advocated are consistent with existing beliefs and can be readily integrated into existing practice, then less time and engagement is needed than when existing beliefs are dissonant with the ideas promoted through the change. In the latter case, much closer engagement is needed.

Engaging Students in New Learning Experiences

Little is likely to change for students if classroom teaching and learning activities do not change as a result of the inquiry and knowledge development. Changes within the instructional core are those that matter (Elmore, 2004).

Part of the process of continuing professional learning through a systematic inquiry process, therefore, must be to apply new learning in practice and to view efforts to implement new ideas as part of learning. Professional learning cannot be seen as a process of acquiring, then applying new knowledge and skills. Thus, the analysis of classroom practice must be integral to the professional learning process (Timperley, 2011). Leaders play a central role in this process as they observe and support the learning of their teachers to problem-solve together new puzzles that arise.

Assessing Impact

The final question in the inquiry and knowledge-building cycles asks teachers and leaders to assess the impact of any changes on the outcomes valued for the students. It places students at the center once more and contributes to developing adaptive expertise by asking:

How effective has what we have learned and done been in promoting our students' learning?

Impact should be judged on both a lesson-by-lesson basis and more long-term. The contextualized nature of teaching practice – this teacher with this group of student learners in this school – means there can be no guarantee that any specific activity will have the anticipated result. The lesson-by-lesson check assesses immediate understandings of the lesson. Longer-term assessment ensures that the progress made is adequate against agreed benchmarks, and identifies which areas of the curriculum need further work.

The development of this kind of expertise focuses on developing self-regulated learning processes for teachers and allows

them to take "control of their own learning by defining learning goals and measuring their progress in achieving them" (Donovan et al., 1999, p. 23). In the inquiry and knowledge-building cycles, this approach encourages teachers not only to define learning goals and to progress toward them, but also to judge their impact on the outcomes for students.

Assessing impact is not the end of the cycle. As **Figure 1** shows, the arrows keep cycling and continue to redefine what is meant by 'continuing' in 'continuing professional development.' If assessing impact shows previously identified problems persisting, then different approaches to professional learning need to be taken. On the other hand, if the desired improvement is achieved, new challenges are usually identified because the demands of teaching are rarely static.

The Leadership and System Challenge

The research summarized at the beginning of this article that focused on the higher effects of teachers than schools on student learning has led to a focus on teachers and their learning, rather than that of others, such as school and system leaders. Yet, it is difficult for teachers to learn in the ways described above if they work in environments that do not provide conditions that actively promote systematic evidence-informed inquiry and building of knowledge.

While school and system effects are typically much smaller than teacher effects (e.g., Hattie, 2009), recent research on the impact of leadership is moving away from generic school effects to unpacking what makes a difference in leadership practice to outcomes for students. This research has contributed to a growing consensus that particular leadership practices have greater effects on outcomes for students than others (Supovitz et al., 2010; Robinson et al., 2008). Robinson et al., for example, identified five dimensions that were particularly important in this regard. They found that leadership practices involving promoting and participating in professional learning and development had twice the effect sizes of other practices (ES = 0.84). The others (with effect sizes noted in brackets) included establishing goals and expectations (ES = 0.42); resourcing strategically (ES = 0.31); planning, coordinating and evaluating teaching and the curriculum (ES = 0.42); and ensuring an orderly and supportive environment (ES = 0.27). This research provides a strong justification that continuing professional development for teachers is more likely to be effective for student learners if it is promoted by leaders. This role is variously described as instructional leadership, pedagogical leadership, or leadership for learning (Timperley and Robertson, 2011).

Leadership involvement in continuing professional learning is particularly important for sustainability of improved outcomes for student learners by making cycles of inquiry and knowledge-building core business for those responsible for student learners (O'Connell, 2009; Timperley, 2012). Although the research base is relatively recent and still developing, the ongoing gains in outcomes are more evident in continuing professional learning initiatives that develop adaptive expertise than those that do not (e.g., Lai et al., 2009; O'Connell, 2009). Challenges that arise are treated as problems to be solved and addressing them through an inquiry framing becomes part of each school's organizational routines (Newmann et al., 2001).

The research findings about the impact and practices of school leadership cited above can be extended logically to include network, district or local authority leadership. I began this article with the limitations of continuing professional development in realizing policy goals of raising the bar and closing the gap. A difficulty that often arises is the limited understanding at different system layers about approaches that are likely to make a difference. Many persist with one-off workshops and conferences with little follow-through to ensure that what is happening is making a difference. Even more of a problem are those policy environments that create conflicting policy demands and withdraw funding just as the professional learning is gaining traction for student learners (Coburn et al., 2009; Datnow, 2005). On the other hand, policies and practices that promote an inquiry and knowledge-building approach to continuing professional learning, focused on developing adaptive expertise, are more likely to have greater impact than approaches that are inconsistent with the principles and practices outlined above (Lai et al., 2009; McNaughton et al., 2004).

Conclusions

Schools, and those who teach in them, increasingly face a myriad of challenges. Learner demographics, curriculum, and technologies are all changing. Job and life opportunities for learners are also changing, so educating only those who readily engage with existing approaches to teaching and learning is no longer acceptable. While continuous professional learning for teachers has the potential to help them to make the changes in approaches to teaching and learning to meet these demands, teachers cannot do this in an environment where others decide what is good for them and tell them what to do through workshops, conferences, or one-off sessions. Nor is there much evidence to support the idea that teachers can identify, by themselves, what they need to learn independently of a supported analysis of what their students already know and need to learn, accompanied by a systematic assessment of their own capabilities, to meet the challenges that are identified.

In this article I have outlined some of the shifts in thinking that underpin the effective enactment of a systematic inquiry and knowledge-building approach to continuing professional learning with a proven record of improvements in student engagement and learning. These shifts included moving from professional development to professional learning; from defining professionalism as developing routine expertise to developing adaptive expertise; from learner engagement and learning being a 'hoped for by-product' to having these issues central to the process; and finally from a focus on the teachers to a focus on the system that supports their learning. The biggest shift in thinking, however, is that professional learning must become seriously continuous in the sense of becoming core business for every teacher in every school and educational jurisdiction, if we are to make the difference that the student learners who are required to attend our schools deserve.

See also: Communities of Practice; Networking for School Improvement; School Effectiveness Research; School Improvement; School Management; Teacher Education; Teacher Expertise; Teaching as a Profession: United States.

Bibliography

Assessment Reform Group, 2002. Assessment for Learning: 10 Principles. Author, London.

Bransford, J., Brown, A., Cocking, R. (Eds.), 2000. How People Learn: Brain, Mind, Experience and School. National Academy Press, Washington, DC.

Bransford, J., Darling-Hammond, L., LePage, L., 2005. Introduction. In: Darling-Hammond, L., Bransford, J. (Eds.), Preparing Teachers for a Changing World: What Teachers Should Learn and Be Able to Do. Jossey Bass, San Francisco, pp. 1–39.

Coburn, C., Touré, J., Yamashita, M., 2009. Evidence, interpretation and persuasion: instructional decision making at the district central office. Teachers College Record 111 (4), 1115–1116.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, second ed. Erlbaum, Hillsdale NJ.

Dall'Alba, G., Sandberg, J., 2006. Unveiling professional development: a critical review of stage models. Review of Educational Research 76 (3), 383–403.

Datnow, A., 2005. The sustainability of comprehensive reform models in changing district and state contexts. Educational Administration Quarterly 41 (1), 121–153.

Donovan, M., Bransford, J., Pellegrino, J., 1999. How People Learn: Bridging Research and Practice. National Academy Press, Washington, DC.

Dreyfus, H., Dreyfus, S., 1986. Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer. Free Press, New York.

Elmore, R., 2004. School Reform from the Inside Out: Policy, Practice and Performance. Harvard Education Press, Cambridge, MA.

Feiman-Nemser, S., 2008. Teacher learning: how do teachers learn to teach? In: Cochran-Smith, M., Feiman-Nemser, S., McIntyre, D., Demers, K. (Eds.), Handbook of Research on Teacher Education, third ed. Routledge, New York, pp. 697–705.

Fennema, E., Franke, M., Carpenter, T.P., Carey, D., 1993. Using children's mathematical knowledge in instruction. American Educational Research Journal 30 (3), 555–583.

Fullan, M., 2005. Leadership and Sustainability, System Thinkers in Action. Corwin Press, Thousand Oaks, CA.

Guskey, T.R., 2000. Evaluating Professional Development. Corwin Press, Thousand Oaks, CA.

Hammerness, K., Darling-Hammond, L., Bransford, J., Berliner, D., Cochran-Smith, M., McDonald, M., Zeichner, K., 2005. How teachers learn and develop. In: Darling-Hammond, L. (Ed.), Preparing Teachers for a Changing World: What Teachers Should Learn and Be Able to Do. John Wiley & Sons, San Francisco, pp. 358–389.

Hanushek, E., 2005. Economic Outcomes and School Quality. Education Policy Series. International Academy of Education & International Institute for Educational Planning, UNESCO. http://www.smec.curtin.edu.au/iae/.

Hargreaves, A., 1998. The emotional practice of teaching. Teaching and Teacher Education 4, 835–854.

Hatano, G., Inagaki, K., 1986. Two courses of expertise. In: Stevenson, H., Azama, H., Hakuta, K. (Eds.), Child Development and Education in Japan. Freeman, New York, pp. 262–272.

Hattie, J., 2009. Visible Learning. Routledge, London and New York.

James, M., McCormick, R., Black, P., et al. 2007. Improving Learning How to Learn – Classrooms, Schools and Networks. Routledge, Abingdon.

Lai, M., McNaughton, S., Timperley, H., Hsiao, S., 2009. Sustaining continued acceleration in reading comprehension achievement following an intervention. Educational Assessment, Evaluation and Accountability 1 (1), 81–100.

McNaughton, S., Lai, M.K., MacDonald, S., Farry, S., 2004. Designing more effective teaching of comprehension in culturally and linguistically diverse classrooms in New Zealand. Australian Journal of Language and Literacy 27 (3), 184–197.

Muijs, D., Reynolds, D., 2001. Effective Teaching: Evidence and Practice. Paul Chapman Publishing, London.

Newmann, F.M., Smith, B.A., Allensworth, E., Bryk, T., 2001. Instructional program coherence: what it is and why it should guide school improvement policy. Educational Evaluation and Policy Analysis 23 (4), 297–321.

Nye, B., Konstantanopoulos, S., Hedges, L.V., 2004. How large are teacher effects? Educational Evaluation and Policy Analysis 26 (3), 237–257.

O'Connell, P., 2009. Is Sustainability of Schooling Improvement an Article of Faith or Can It Be Deliberately Crafted? (Unpublished Ph.D. thesis) The University of Auckland.

OECD, 2005a. Teachers Matter: Attracting, Developing and Retaining Effective Teachers. Organisation for Economic Cooperation and Development, Paris.

OECD, 2005b. Education at a Glance: OECD Indicators 2005. OECD Publications, Paris.

OECD, 2010. PISA 2009 Results: Overcoming Social Background – Equity in Learning Opportunities and Outcomes, vol. II. Retrieved from: http://dx.doi.org/10.1787/9789264091504-en.

Robinson, V.M.J., Lloyd, C., Rowe, K.J., 2008. The impact of leadership on student outcomes: an analysis of the differential effects of leadership type. Educational Administration Quarterly 44 (5), 635–674.

Scheerens, J., Vermeulen, C., Pelgrum, W.J., 1989. Generalizability of instructional and school effectiveness indicators across nations. International Journal of Educational Research 13 (7), 789–799.

Schmitt, M.C., Askew, B.J., Fountas, I.C., Lyons, C.A., Pinnell, G.S., 2005. Changing Futures: The Influence of Reading Recovery in the United States. Reading Recovery Council of North America, Worthington, OH.

Spillane, J., 1999. External reform initiatives and teachers' efforts to reconstruct their practice: the mediating role of teaches' zones of enactment. Journal of Curriculum Studies 31, 143–175.

Stoll, L., Fink, D., Earl, L., 2003. It's about Learning [and It's about Time]: What Is in It for Schools? Routledge Falmer, London.

Supovitz, J., Sirinidies, P., May, H., 2010. How principals and peers influence teaching and learning. Educational Administration Quarterly 46 (1), 31–56.

Timperley, H., 2011. Realizing the Power of Professional Learning. Open University Press, Berkshire, England.

Timperley, H., 2012. Learning to Practice in Initial Teacher Education. Report to the New Zealand Ministry of Education, Wellington.

Timperley, H., Parr, J., Meissel, K., 2010. Making a Difference to Student Achievement in Literacy: Final Research Report on the Literacy Professional Development Project. Learning Media Ltd., Wellington.

Timperley, H., Robertson, J., 2011. Establishing platforms for leadership and learning. In: Robertson, J., Timperley, H. (Eds.), Leadership and Learning. Sage Publications, Los Angeles.

Timperley, H., Wilson, A., Barrar, H., Fung, I., 2008. Teacher Professional Learning and Development: Best Evidence Synthesis Iteration. Ministry of Education, Wellington. www.minedu.govt.nz/goto/bestevidencesynthesis.

Wright, S.P., Horn, S.P., Sanders, W.L., 1997. Teacher and classroom context effects on student achievement: implications for teacher evaluation. Journal of Personnel Evaluation in Education 11, 57–67.

Contractarianism

Alan Hamlin, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Contractarianism and contractualism identify a range of positions in moral and political theory that is distinctive in terms of its emphasis on considerations of individualistic eligibility: social, political, or moral rules, norms, or institutions are to be explicated and justified by appeal to the extent to which they would be chosen, at least by certain types of agents in certain situations. Classical contractarianism was developed by Hobbes and Locke, and by Rousseau and Kant, and a range of contemporary contractarian and contractualist positions emerged in the second half of the twentieth century to challenge utilitarian and other idealist approaches to political and moral philosophy. Classical contractarianism is often identified with an apparently false belief in a real historical contract, or with a deeply problematic relationship between a hypothetical contract and the real world. Contemporary contractarianism and contractualism are more heuristic in their use of the idea of agreement, using the notions of contract and agreement as a means of investigating social, political, and moral questions. Contemporary contractarianism sits in the broadly Hobbesian tradition and takes individual interests as basic, seeking to explore the prospects for mutual advantage among rational individuals. Contemporary contractualism sits in the broadly Kantian tradition and emphasizes the rational contemplation of social, political, or moral questions from an impartial perspective as the appropriate means of investigation. Both traditions face issues of analyzing the nature of *ex ante* agreement; the basis for *ex post* compliance; the relationships between the individual, the social, the rational, and the moral; and the connections with ideas such as individualism, democracy, and pluralism.

Introduction

The voluntary act of entering into a contract or agreement may be understood as the act of choosing to be involved with others in a specified joint venture or commitment. Any argument that places an idea of contract, agreement, or consent in a central role in explaining or justifying a particular joint venture or commitment might then be seen as broadly 'contractarian' in nature. Most prominent among the joint ventures and commitments that might be explained or justified are political, social, and legal institutions including government, and moral values. Broadly contractarian arguments are distinctive in terms of their emphasis on considerations of eligibility, so that demonstrating that a particular arrangement or commitment was, or might have been, chosen, at least by certain types of individuals in certain circumstances, carries weight in explicating and justifying that arrangement or commitment. Contemporary contractarian and contractualist arguments differ from each other along several dimensions, most obviously according to the basis of agreement or consent, the domain of the argument, the role of the idea of a contract or agreement in the argument, and the specific nature of the relevant contract. These dimensions map out a space that includes many positions. Not all points in this space are tenable, but a number of them are occupied by contemporary contractarians and contractualists.

In the dimension associated with the basis of agreement or consent, the most fundamental distinction is that between those who, following Hobbes, locate agreement in individual interests and mutual advantage, and those who, following Kant, locate agreement in an impartial assessment. It is this distinction that is tracked by the contemporary use of the terms 'contractarianism' and 'contractualism' respectively.

In the domain dimension, a modest contractarianism or contractualism would seek to explain or justify some particular activity, norm, or other aspect of social life against a background that already includes a rich variety of social institutions and norms. By contrast, a moderately ambitious or *political* contractarianism or contractualism seeks to explain and justify basic social and political institutions themselves, rather than merely the activities that occur within them. Finally, a highly ambitious or *moral* contractarianism or contractualism seeks to provide a theory of morality itself.

In the dimension that identifies the role of agreement in any contractarian or contractualist argument, we note two major possibilities: agreement might play either a shallow, indicative role or a deep, constitutive role. An indicative role might be relevant in cases where the agreement serves as a useful but inessential signal of some further and more basic property (the signal may be sufficient but not necessary for the realization of the underlying property). By contrast, agreement could play a deeper, constitutive role where the idea of contract is essential to the argument (so that the contract is both necessary and sufficient).

In the dimension that identifies the specific nature of the relevant contract, the most obvious range extends from the case of the fully explicit, actual, or historical agreement, to various specifications of tacit, implicit, or hypothetical agreements. But there is also scope for qualifying the nature of agreement in other ways. Thus, some contractarian or contractualist arguments will emphasize the informational context of agreement, the extent to which agreement is grounded in fully rational deliberation, or the extent to which agreement could be reasonably rejected.

The basic idea or vision that unites all contractarian and contractualist thought is the conception of society (in the case of political contractarianism) or morality (in the more ambitious case) as human constructions, so that, for example, society is 'a co-operative venture for mutual advantage' (Rawls, 1971, p. 4). This places contractarianism and contractualism in

a broadly individualist setting in which society or morality is to be explained and justified by appeal to the extent to which it responds to human needs, supports cooperation, and suppresses various forms of conflict.

In the Hobbesian contractarian tradition the interests of individuals are taken as basic so that agreement is seen as identifying mutuality of interest. The most basic version of such a theory would begin in a presocial state of nature, allow individual interests to be whatever the individuals concerned perceive them to be, and require real, explicit, and unanimous agreement between individuals. Such a theory is open to a range of obvious criticisms: that agreements based on mistaken or misperceived interests cannot carry the normative weight required by the argument; that since explicit, historical agreements of the required type are not a common feature of the real world, the contractarian argument cannot explain or justify the actual political or moral features of the real world; and so on. The point to be emphasized here is that, notwithstanding the success of these criticisms against the most basic form of Hobbesian contractarianism, more sophisticated versions of the idea of contract or agreement are important as a means by which to investigate the prospects for mutual advantage for rational individuals.

In the Kantian contractualist tradition, attention shifts from mutual advantage to impartiality. The basic idea here is that the rational contemplation of the world from an impartial standpoint will reveal what political arrangements are just or what is moral. There is no fundamental attempt to link the contractualist idea to real, explicit agreements. The essential nature of the Kantian contractualist method is to construct a hypothetical, impartial position from which to consider what is just by asking what might be rationally chosen or agreed from that position.

Contract and Compliance

Although the possibility of cooperation is a hallmark of the vision of society in both the Hobbesian contractarian and Kantian contractualist approaches, it is the nature of the conflict that threatens cooperation that sets the more detailed agenda. This agenda may be illustrated by a sequence of simple games that have become staples of the literature – the pure coordination game, the mixed-motive game, and the prisoners' dilemma. In developing these illustrations, I will use the language of a simple Hobbesian, interest-based contractarian argument, where the contract is explicitly entered into.

In the pure coordination game, each player has an interest in coordinating with others, and no other relevant interest. This is illustrated in the simple two-person, simultaneous-play game of Figure 1, where the payoffs from each combination of actions are shown in the form (payoff to row, payoff to column).

All that matters in this game is that the two players should take the same action – both should X, or both should Y – since the interests of the two players are perfectly aligned. In this setting, it is easy to see how preplay communication between the players might allow an agreement that identifies a *convention* in the sense of Lewis (1969). So there is little difficulty in identifying the potential grounds for agreement (although there might be a minor difficulty in identifying which of the actions is to be coordinated on). It is equally easy to see that such an agreed convention will be self-enforcing in that neither player has any reason to depart unilaterally from the convention once it is established. There is no individual compliance problem here. This, then, is a model in which a simple contractarian argument promises to work well – the agreement seems likely to pick out social or political conventions that are self-enforcing in the relevant population. However, the obvious criticism is that a model of pure coordination is an inappropriate starting point for even moderately ambitious political contractarians since, although it may capture the essential features of some social and political issues, it can hardly claim to capture the fundamental nature of society where both competition and conflict must be recognized.

A move toward incorporating conflict is made in Figure 2, which represents the mixed-motive game. The benefits of cooperation are still clear – each player still has an interest in doing what the other does – but now the two players have divergent interests concerning the action to coordinate on. Row has an interest in coordinating on X, while Column has an interest in coordinating on Y. This introduces a difficulty into the contractarian argument: that any preplay agreement must now overcome an essential conflict of interests. If we are willing to assume that some agreement is reached, we might still argue that such an agreement forms a convention and is self-enforcing. Consider the case in which Row and Column agree that both should X; Column now has good reason to suppose

		Column	
		X	Y
Row	X	(1,1)	(0,0)
	Y	(0,0)	(1,1)

Figure 1 Pure coordination.

		Column	
		X	Y
Row	X	(2,1)	(0,0)
	Y	(0,0)	(1,2)

Figure 2 Mixed motives.

		Column	
		X	Y
Row	X	(2,2)	(0,3)
	Y	(3,0)	(1,1)

Figure 3 Prisoners' dilemma.

that Row will X and, given this belief, Column will also have good reason to X. There may be no individual compliance problem here despite the *ex ante* conflict between players.

A further step in the direction of conflict is illustrated in the standard prisoners' dilemma game of Figure 3. Here each individual has a private reason to Y, since this is the dominant strategy. The potential and mutual gains from cooperation in this case come from shifting the behavior of *both* players from Y to X. In a sense, the problem here is the opposite of the problem in the mixed-motive game. The nature of any agreement seems clear enough, precisely because the coordinated shift from Y to X does offer mutual benefits, while no other route to mutual benefits exists. But now the existence of a severe compliance problem may undermine any agreement to X. If both agree to X in preplay negotiations, it is not clear that either will comply with the agreement, since it will be in their individual interests to renege on the agreement regardless of whether they believe the other will keep to the agreement (this is just what it means to say that Y is the dominant strategy). The preplay agreement may be dismissed as 'cheap talk' that does not actually affect individual decision making. If the compliance problem is so severe as to render prior agreement ineffective, there will be no reason for individuals to waste their time in reaching ineffective contracts.

These simple illustrative games serve to identify two of the basic issues facing contractarians and the relationship between these issues: on the one hand, the issue of the nature and content of any agreement given the likely presence of *ex ante* conflict between individuals and, on the other, the issue of compliance with any agreement given the *ex post* conflicting private incentives facing individuals.

Contemporary Contractarian Positions

We have identified two broad traditions – the Hobbesian interest-based contractarian tradition and the Kantian impartiality-based contractualist tradition – and two broad issues: the *ex ante* agreement issue and the *ex post* compliance issue. The remainder of this article will briefly consider contemporary contractarian positions – representing both traditions and offering different approaches to the basic issues. The intention is both to sketch some salient features of contemporary debate and to reveal some of the strengths and limitations of contractarian and contractualist argument.

Hobbesian Contractarianism – Buchanan and Gauthier

Both Buchanan's and Gauthier's contractarianism are clearly and explicitly Hobbesian in nature, but an initial difference between these two contemporary contractarians lies in the scale of their ambition. Buchanan ties his contractarian argument to political constitutionalism, so we might view Buchanan as an essentially *political* contractarian. Buchanan takes contractarianism as providing a justification for political and constitutional structures, rather than merely an explanation, so that he is committed to contractarianism as a means of moral enquiry; but he does not set out to derive any specific or substantive moral principles from the idea of agreement. Buchanan is certainly a moral individualist, but he may be interpreted as allowing individuals to hold whatever moral views they may happen to have (including none at all), insisting only that agreement or contract is the means by which individual moral values, as well as individual desires or preferences, must be brought together in any exercise in normative social or political theory. By contrast, Gauthier's ambition is that of a full-scale moral contractarian. *Morals by Agreement* sets out an argument (with many of the details revised and extended in later work) for the contractarian construction of morality. More specifically, "Moral principles are introduced as the objects of fully voluntary *ex ante* agreement among rational persons. Such agreement is hypothetical … But the parties to the agreement are real, determinate individuals, distinguished by their capacities, situations and concerns … As rational persons understanding the structure of their interaction, they recognize a place for mutual constraint, and so for a moral dimension in their affairs" (Gauthier, 1986, p. 9). These differences in ambition between Buchanan and Gauthier relate to further differences in the treatment of both *ex ante* agreement and *ex post* compliance.

A basic distinction in Buchanan's argument is that between 'constitutional' and 'in-period' choices. Constitutional choices are described as choices of rule systems, while in-period choices are described as choices within rules. The *ex ante* agreement issue, for Buchanan, is then the question of why individuals who may differ markedly in their interests (and values), and in their in-period motivations and behavior, may nevertheless agree on rules. Buchanan's approach to this issue focuses on considerations of the nature of the objects of choice rather than the choosing agents. As already noted, Buchanan does not impose on his individuals any restriction in the set of interests or values that they might have. On the contrary, Buchanan frequently insists that 'we start from here' and that individuals must be taken 'as they are,' recognizing their diversity. Rather, Buchanan argues that the nature of constitutional choice itself will present alternatives in such a way as to reduce conflict between individuals, and so enhance the prospects for agreement. One key element of this argument is that constitutional provisions are 'general' in that they are expected to apply both across many particular in-period choice situations and across time, so that any individual contemplating the choice between rule systems will have good reason to judge the alternative rule systems in terms of their more general properties, rather than in terms of the implications for any particular in-period decision.

Thus, for example, in choosing a voting rule to be applied to a broad class of in-period political decisions, any individual at the constitutional stage will have good reason to balance the implications of each potential voting rule across a wide range of issues, and this suggests that different individuals may judge voting rules similarly, even if their specific preferences or values are rather different. This argument seems secure; surely the move from the level of the particular in-period decision to the more abstract constitutional level will tend to suppress at least some of the potential conflict between individuals and so allow potential agreement. However, this apparent success in at least diluting the *ex ante* agreement problem may simply transfer the problem to the arena of compliance. The *ex ante* agreement problem is still more severe for Gauthier. To the extent that Gauthier simply argues for the 'recognition of mutual constraint' at the *ex ante* level, there is no substantial difference between Gauthier and Buchanan – 'mutual constraints' may be identified with constitutional rules without much loss. However, Gauthier needs to go a step further to argue that individuals will come to agree on moral principles rather than just particular, institutionalized constraints or rules. To take this extra step, Gauthier relies on a particular model of bargaining (although the specific theory of bargaining is an area in which Gauthier has revised his views) and a version of the Lockean proviso. The proviso is of particular interest since it seems to be a moral principle, but is not argued to arise out of agreement. It is, rather, a precondition for agreement. Essentially, the proviso states that initial bargaining between individuals should be constrained so that no individual should be worse off in the initial bargaining situation than they would be in a nonsocial context of no interaction. This effectively introduces a minimal set of rights into the original bargaining situation. Gauthier argues that such a proviso has to be accepted by each individual if agreement is to be possible. In one sense, this proviso might be thought of as a technical device to ensure that the bargaining outcome actually realizes mutual advantages, and therefore stays true to the underlying Hobbesian idea; but in another sense, it reveals a potential weakness of any bargaining model in that the outcome will reflect bargaining power relative to some baseline 'rights.' If such power is to be viewed as illicit in some cases – and ruled out by the proviso – why should the outcome of power be granted moral status in other cases?

The basic compliance question for any Hobbesian contractarian is simply why any individual should feel him- or herself to be bound or obligated by a prior agreement (whether on political rules or moral principles), even if that agreement was explicit and real. This question may be put in terms of the *ex post* motivational impact of the *ex ante* agreement, and whether the motivational impact is sufficient for the agreement to be self-enforcing. If interests are the currency of both *ex post* decision making and *ex ante* agreement, it would seem to place these two activities on all fours with each other so that the claimed priority of the *ex ante* contract over the *ex post* decision is threatened.

The prospects for wholly self-enforcing contracts seem to be hopeless if either we are to take individual interests as being entirely unspecified, so that any individual may have any interests at all, or we insist on interests being modeled in terms of narrow self-interest along the lines of the standard depiction of *Homo economicus*. In either of these cases, the fact of the contract (if it is a fact) and the content of the contract (whatever it might be) would seem to have no independent or reliable role in determining in-period behavior in situations such as the prisoners' dilemma where in-period incentives are clear. In order to generate reliable compliance in such cases, we seem to need to assume at least some motivational relevance or status for the contract itself or the rules or principles that derive from that contract. Without such a move, the range of political and moral principles that might be explained or justified by a Hobbesian contractarian would seem to be limited to those that can be analyzed as self-enforcing conventions (see Hardin, 2003). Gauthier attempts such a move by arguing for what he terms 'constrained maximization' rather than straightforward maximization as the appropriate interpretation of rational decision making, where "A constrained maximizer has a conditional disposition to base her actions on a joint strategy, without considering whether some individual strategy would yield her greater expected utility … in other words, a constrained maximizer is ready to co-operate in ways that, if followed by all, would yield outcomes that she would find beneficial and not unfair, and she does co-operate should she expect an actual practice or activity to be beneficial" (Gauthier, 1986, p. 167). The basic point here is that Gauthier is attempting to ground his moral contractarianism in a specific interpretation of what rationality requires. The detailed claim about rationality is clearly contentious, but without some similar claim, there would be little prospect for contractarianism that goes much beyond the securing of conventions.

A second aspect of the question of compliance relates to the possibility of enforcement. Clearly this is of much more relevance to a political contractarian than a moral contractarian since a moral contractarian must by necessity be concerned with the acceptance and internalization of moral principles, but, even for a moral contractarian, it is clear that enforcement may play some role. So, can enforcement resolve the compliance problem? As usual, the answer is both yes and no. Of course, the possibility of more complex contracts that include enforcement mechanisms will extend the range of political and social situations that can be treated. Such complex contracts need raise no new issues at the *ex ante* stage. For example, if a group of individuals contract to engage in a joint venture that has the characteristics of providing a public good (the draining of a swamp, perhaps), it will be appropriate for them to recognize that a simple contract in which each person agrees to contribute to the in-period effort and cost may founder since each individual will face strong in-period incentives to free-ride. In such a case, a form of enforcement, perhaps raising taxes against the credible threat of some further punishment if taxes are not paid, seems to both solve the *ex post* compliance problem and be something that offers appropriate mutual advantage so that it will be eligible at the *ex ante* stage. However, this is too fast. Enforcement of the type required cannot simply be assumed to be available. Typically, enforcement requires enforcers, and their motivations and decisions must be explicitly considered. Assuming that judges will virtuously apply the law, or that police officers will act regardless of their own individual interests, is not to solve the compliance problem but simply to relocate it. This is not to

deny that enforcement is sometimes possible and useful, but merely to suggest that while a political or social arrangement may involve enforcement within its structure, the structure as a whole must be seen to be self-enforcing in the sense that the behavior of all individuals – including those who act as enforcers – are seen to be consistent with their motivational characters and the incentives facing them.

Kantian Contractualism – Rawls and Scanlon

Some of the discussion of the Hobbesian positions of Buchanan and Gauthier applies with only minor modification to the Kantian positions of Rawls and Scanlon. In this Section, I will concentrate on rather different concerns to avoid repetition.

The Kantian nature of Rawls's contractualism is most obvious in the formulation and discussion of the original position. This position, with its relatively thick veil of ignorance, is the basic means by which Rawls attempts to overcome the potentially conflictual nature of the *ex ante* agreement and generate the impartial conditions under which to explore what justice requires. To (over)simplify, the original position attempts to strip away all of the distinctive and partial characteristics of specific individuals, their particular circumstances and interests, and so on, thus reducing the extent of conflict between individuals. In the limit, this process reduces the idea of rational agreement among individuals in the original position to the idea of the rational choice of a single hypothetical and representative individual. This then places almost the entire burden of identifying the institutions and principles that should govern society on the idea of rationality. The danger here is that the original position must satisfy two rather conflicting conditions. On the one hand, it must be sufficiently removed from the real world to satisfy the requirement of impartiality, that is, it must strip away almost all the recognizably human aspects of the individual, while on the other hand it must retain sufficient structure to allow the idea of rationality and rational choice to be meaningful and recognizable. This is no trivial task. The standard thin theory of rationality favored by economists and seemingly endorsed by Rawls is one in which rational choice is defined in terms of the desires and beliefs of the particular individual; in the absence of specific desires and beliefs, it is not clear what the thin theory of rationality entails.

Scanlon's contractualism differs from that of Rawls in several key respects – first, in terms of its ambition, where Scanlon is explicitly more ambitious than Rawls in arguing for a moral rather than a merely political contractarianism; and, second, in his formulation of the idea of *ex ante* agreement, which makes no major use of the construction of an original position. Scanlon offers the following definition of his account of contractarianism: "An act is wrong if its performance under the circumstances would be disallowed by any system of rules for the general regulation of behaviour which no one could reasonably reject as a basis for informed, unforced general agreement. This is intended as a characterization of the kind of property which moral wrongness is" (Scanlon, 1982, p. 110). Contractarianism, in this view, is concerned with providing a distinctive basis for moral reasoning by individuals. Thus, for example, "A satisfactory moral philosophy will not leave concern with morality as a simple special preference, like a fetish or a special taste, which some people just happen to have. It must make it understandable why moral reasons are ones that people can take seriously" (Scanlon, 1982, p. 106); and, again, "According to contractualism, the source of motivation that is directly triggered by the belief that an action is wrong is the desire to be able to justify one's actions to others on grounds they could not reasonably reject" (Scanlon, 1982, p. 116). This formulation ties together the *ex ante* agreement and *ex post* compliance issues in a particular way. By formulating the idea of agreement in terms of principles 'which no one could reasonably reject', Scanlon allows a degree of flexibility – there is no argument to suggest that there exists a unique set of principles that could not be reasonably rejected – and so weakens the requirement of agreement in Rawls' formulation. But, at the same time, the specification that the hypothetical 'agreement' (or nonrejection) is something that occurs in the minds of real individuals (rather than individuals in an original position) who are also motivated to justify their actions to others provides a route to compliance. It is the desire to justify action to others that drives the motivational side of the problem, and the requirement of 'reasonableness' in the nature of justification that bears the strain in terms of identifying particular justifying arguments. But the dangers here are clear: that the desire to justify my action to others is simply smuggled in as a 'simple special preference like a fetish or a special taste,' which is not itself grounded in any contractarian logic, and that the idea of 'reasonableness' is insufficiently detailed to bear the strain placed on it.

There is a final concern that is shared by moral contractarians and contractualists, whether Hobbesian or Kantian. Even if we grant that it may be possible to parlay rational agreement, in whatever interpretation of that phrase, into moral principles, will the moral principles that emerge be recognizable as our moral principles? In particular, will the resources of contractarianism or contractualism be sufficient to ground the range of moral views that we actually have, and, if not, what would be the implications of that fact? Would it point to the possibility that some of our 'moral' views are not really moral after all, but mere tastes or preferences? Or would it point to the possibility that there may be other means by which moral principles may be grounded, so that agreement might not be the only means of approaching moral principles? Of course, this issue arises only once substantive moral principles have been derived from the contractarian method, and there is certainly no consensus as to precisely which substantive moral principles can be so derived.

Finale

Contemporary contractarianism and contractualism offer approaches to political and moral theory that share as their basic commitments a form of individualism and the recognition of the central role of rationality in human affairs. In a simple sense, contractarianism is dedicated to the question of how rational people can live together. This simple and appealing structure goes a long way to explaining the continuing appeal of contractarianism. Further refinements in contractarian and contractualist thought are available. For

example, Southwood (2010) offers a discussion of 'deliberative contractualism' which, he argues, overcomes at least some of the limitations of both Hobbesian contractarianism and Kantian contractualism by focusing on an account of rationality that embraces deliberation rather than either individual interests or the impartial perspective.

However, not all accept the basic vision underpinning contractarian and contractualist thought. Communitarians and others who take society as prior to the individual, for example, must object to contract thinking in all of its forms by reason of its underlying individualism. Equally, not everyone accepts that the basic idea of eligibility is compelling or even interesting. A question to all such critics of contractarianism and contractualism is how they might expect to persuade individuals (even beyond the range of philosophers and political theorists) that their preferred alternative is valid without falling into a form of contractarianism themselves. For if a particular substantive moral theory or political constitution can be justified to others, it might be presumed that it could be derived by a contractarian argument that emphasizes that mode of justification. This is simply a way of pointing to the fact that debate between contractarian and contractualist accounts and others is sometimes guilty of a mismatch between the contractarian method and particular substantive positions that might, at least in principle, be derived from that method.

Contemporary contractarianism and contractualism are engaged in the study of the relationships between the idea of rationality, the nature of individual interests and dispositions, self-interest and morality, political and social rules, and moral psychology. While there are many areas in which these relationships are still obscure, and while there are surely noncontract approaches that may prove rewarding, it seems difficult to deny the importance of these relationships, or the potential of the contractarian and contractualist enterprise to illuminate them.

See also: Contracts: Legal Perspectives; Justice and Law; Kant, Immanuel (1724–1804); Kantian Ethics and Politics.

Bibliography

Barry, B., 1989. Theories of Justice. University of California Press, Berkeley, CA.
Barry, B., 1995. Justice as Impartiality. Clarendon Press, Oxford, UK.
Buchanan, J.M., 1975. The Limits of Liberty. Chicago University Press, Chicago, IL.
Dworkin, G., 2002. Contractualism and the normativity of principles. Ethics 112, 471–482.
Gauthier, D., 1986. Morals by Agreement. Clarendon Press, Oxford, UK.
Hardin, R., 2003. Liberalism, Constitutionalism and Democracy. Oxford University Press, Oxford, UK.
Hampton, J., 1986. Hobbes and the Social Contract Tradition. Cambridge University Press, Cambridge, UK.
Kraus, J., 1993. The Limits of Hobbesian Contractarianism. Cambridge University Press, Cambridge, UK.
Kumar, R., 2003. Reasonable reasons in contractualist moral argument. Ethics 114, 6–37.
Lewis, D., 1969. Convention. Harvard University Press, Cambridge, MA.
Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge, MA.
Scanlon, T.M., 1982. Contractualism and utilitarianism. In: Sen, A., Williams, B. (Eds.), Utilitarianism and Beyond. Cambridge University Press, Cambridge, UK.
Scanlon, T.M., 1998. What We Owe Each Other. Harvard University Press, Cambridge, MA.
Southwood, N., 2010. Contractualism and the Foundations of Morality. Oxford University Press, Oxford, UK.

Contracts: Legal Perspectives

Leon E Trakman, University of New South Wales, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article analyzes the common law of contracts from both legal and social behavioral perspectives. Section Contracts in Common and Civil Law Systems distinguishes between contract law in civil and common law systems, giving primary focus to the common law – or 'judge made' law – of contracts. Section Freedom to and Freedom from Contracts identifies the often seemingly irreconcilable ideological tension between the freedom of parties to and the freedom of parties from contracts. It illustrates that this tension is manifest differently according to the type of contract, such as business and consumer contracts. It also suggests that this ideological tension over contractual freedom often transcends the kind of contract in issue, focusing instead on competing conceptions of fairness and efficiency in judicial practice. Section The Judicial 'Management' of Common Law Contracts evaluates how judges manage contracts. It identifies different methods by which they do so, such as by interpreting contracts according to the contract text, or by implying terms into those contracts. It evaluates the perceived strengths and limitations of these methods, including the critical view that contract law can perpetuate social and economic advantage more than redress them. The conclusion is that competing conceptions of contract law, and the role of common law judges in construing them, will continue. However, these tensions are preferable to complacency or resigned acceptance of the status quo.

Introduction

A contract is a legally recognized agreement between two or more parties, which is binding and enforceable in law, the default or 'breach' of which gives rise to legal liability. The usual remedy in common law contracts for breach is expectation damages for a loss of profit or other benefits arising from the breach (Fuller and Perdue, 1936; Macneil, 1982). However, 'equitable remedies' such as injunctions and specific performance are sometimes also available.

A contract is formed if the parties consent to contract. Consent in the common law system entails one party making an offer to the other party, which that other party accepts, and the parties exchange valid consideration for their mutual promises. In addition, the material terms of the contract must be certain. Nor may they be illusory in a material respect, as when one party retains an unfettered discretion to decide whether or not to perform an obligation assumed by contract.

Some contracts are required to comply with legal formalities, such as the registration of the sale of land and marriage contracts, or the requirement that a contract be in writing and signed by the parties or their agents. However, contracts in general are not required to be formally executed. Informal agreements, such as entered into by business parties over lunch, may well be binding if the parties actually or reasonably intend to be bound by them (Macaulay, 1963).

A contract can be set aside due to a 'defect' in consent. For example, a common law court may nullify a contract arising from a fraudulent misrepresentation, such as when a seller knowingly misleads the buyer into buying a valuable work attributed to Picasso, which is actually a forgery. A court may set a contract aside because of a common mistake, such as where both parties are under the mistaken belief that the painting is a Picasso.

Parties to contracts ordinarily can enforce them. However, in limited circumstances a court will hold that a contract is not enforceable because it offends good morals, is against public policy, or violates a statute or administrative regulation.

Parties to contracts are required to perform their contractual obligations. A party can terminate a contract if the other party is in default, as when a seller delivers materially defective goods under a contract of sale. A contract can also be rendered unenforceable for frustration, as when an unforeseen circumstance reasonably beyond a party's control, such as a war, renders that party's performance impossible or unduly burdensome.

Social arrangements, such as an agreement to attend a family or social gathering, are ordinarily not binding in law and have no legal consequences. However, they may have moral, social, or economic significance (Coleman, 1980). In addition, so-called socially responsive courts may be influenced by such arrangements, such as when they are interpreting contracts in light of their social and economic context (Macneil, 1980; Pound, 1908).

Contracts in Common and Civil Law Systems

Contract law in civil law jurisdictions (across Europe, Latin America, Japan, China, and parts of Africa) is largely codified and is part of the civil law of 'obligations.' Courts in civil law jurisdictions, in turn, apply the articles of applicable civil law codes in deciding particular contract cases. Commentary on those provisions by jurists, the so-called *opinio juris*, further elucidates upon their meaning and application in particular disputes. As a result, the civil law system of obligations is conceived as being primarily deductive in nature: the law applied in particular cases is deduced from general principles of law, which are contained in codes of law and elucidated upon by jurists (von Mehren).

In contradistinction to civil law codes, common law contracts (in the British Commonwealth and the United States) are primarily judge made. Common law judges develop principles, standards, and rules of law deductively to govern the formation, performance, and termination of contracts. Unlike in civil law jurisdictions, common law

legislatures ordinarily play only a secondary role in regulating contractual relations. However, there are increasingly significant codes in common law jurisdictions that regulate particular kinds of contracts, such as the Uniform Commercial Code in the United States regulating, among others, contracts of sale and secured transactions (Llewellyn, 1930). There is also legislation across all common law jurisdictions regulating particular kinds of contracts such as related to banks and insurance companies, or particular kinds of parties such as employees and consumers.

Civil and common law systems predominate globally. Not atypically, both Japan and China adopted the German civil law governing the law of obligations. However, the Chinese law of obligations is also significantly influenced by Taiwanese law. In many jurisdictions, too, customary laws regulate family contracts and further augment the civil or common law.

It is important in the understanding of common law contracts to also appreciate the differences from civil law obligations, and in particular cases, customary laws of contracting. Given that we occupy a 'global village,' many transactions are conducted between parties conducting their dealings across national and legal boundaries. Traditionally, such transactions were limited to merchants who in bygone times developed a distinct transnational law merchant to govern their dealings. Today, contracts are conducted across national boundaries by hosts of nonmerchants, such as by purchasers of goods and services over the Internet. These technological developments establish the importance of having an international body of contract law or obligations that is capable of transcending national boundaries and regulating global commerce in a comprehensible, efficient, and fair manner (Trakman, 2009).

For practical purposes, this article concentrates upon the role of courts in common law system of contracting, unless it provides expressly otherwise.

Freedom to and Freedom from Contracts

There is ongoing debate, in both common and civil law systems, over the extent to which contracts ought to be subject to regulation. The traditional nineteenth-century view was that the law should protect the freedom of parties to contract by enforcing their subjective 'wills' or intentions (Williston, 1921). According to this view, legislative, administrative, and judicial regulations that interfered with the 'wills' of the parties undermined their consent to contract and led to courts improperly 'making' or 'unmaking' contracts. As a result, common law courts in the nineteenth and early twentieth centuries enforced agreements according to the subjective intention of the parties. For example, they determined whether the parties had actually exchanged promises, rather than whether it was reasonable for a court to assume that they had done so. In the absence of clearly expressed wills of one or both parties, common law courts of that era determined that there was no *consensus ad idem* between the parties, no 'meeting of the minds,' and no contract (Williston, 1921).

This preoccupation with the freedom of parties to contract continues to influence the common law of contracts and to a lesser extent, the civil law of obligations in the latter half of the twentieth century and today. However, the so-called 'wills theory' by which parties concluded contracts has receded in importance. Courts are less likely to insist that the parties must express their subjective wills clearly and completely in order for agreements to be enforced. They have sought instead to determine whether the parties intended to enter into contracts based on the reasonable intention of the parties, in the absence of clearly or completely articulated wills. They also increasingly consider the social context surrounding contracts, such as the family or commercial setting in which parties concluded family or commercial contacts (Pound, 1908).

A judicial tension in common law jurisprudence relates to the parameters of an objective theory of contracting. Some courts construe the objective theory restrictively, requiring objective evidence of an intention to contract within the text of the contract, thereby limiting themselves to the contractual language used. Others construe contract terms more expansively in light of the practices of the parties, the usages of their trade, and expansive customs. Other judges are more audacious still, constructing contract terms according to whether, in the court's opinion, it is expedient or fair to do so in the circumstances. This tension between narrow textual and broad contextual interpretation of contracts is reflected in judicial attempts to construe – and sometimes construct – contract terms to redress structural and functional bargaining inequalities between contracting parties. For example, a court may interpret an onerous term in a contract literally according to its 'plain word meaning,' such as by interpreting literally, in favor of the employer, a term entitling an employer to dismiss its employee without cause, or by interpreting literally, in favor of the landlord, a term permitting a landlord summarily to evict a tenant. Alternatively, a court may construe such onerous terms in favor of the employee or tenant, respectively, according to standards of due process and fair notice that the court implies by law or fact into employment and tenancy contracts. The judicial object, often reinforced by statute, is to protect vulnerable sectors of society from being unduly subjugated to 'adhesion' contracts drafted by dominant parties (Kessler, 1943).

Such judicial protection of vulnerable contracting parties is by no means new in common law jurisprudence. Already in the nineteenth century, the Court of Chancery in England, responsible for equity, sought to protect mine and factory workers from being forced to accept unreasonable work conditions.

This concern to protect disadvantaged sectors of the economy from contractual exploitation was reinforced in the mid-twentieth century, with the growth of the modern welfare state and the commitment of democratic states to protect consumers in general. Notable among such developments were codified efforts, such as expressed in the American Uniform Commercial Code, to ensure that contracts for the sale of goods were not unconscionable (Llewellyn, 1930). Statutory developments in other common law jurisdictions sought to address such unfairness through specific legislation such as the Unfair Contracts Act in the United Kingdom and the Consumer Protection Act in Australia in 2011. Further legislation protects employees, among other vulnerable groups.

The result is the development of significant bodies of law that are 'made' by judges in interpreting or construing statutes,

and sometimes, in constructing terms beyond those statutes. For example, judges sometimes determine when an employer has unfairly limited by contract its liability for work-related injuries suffered by employees, and less readily, when an insurance company has imposed unduly onerous duties of disclosure by contract upon an insured in order in denying subsequent insurance claims.

Perhaps most frequently, judges have developed principles, rules, and standards to regulate the conditions under which mass producers of consumer goods can exclude liability by contract, such as for defects in computers sold *en masse* to consumers in mass markets (Trakman, 2009).

One result of statutory and common law developments governing different sectors of the economy, such as employment, insurance, and consumer affairs, is the perception that there are different laws of contract applicable to different kinds of contracts or parties. For example, the law of contracts that judges apply to merchants is sometimes conceived as being different in kind and degree from the law of contracts judges apply in family transactions. This distinction is overstated, given that the principles of contract law developed by courts, and glossed by legislation, are ordinarily applied to contracts in general, regardless of the discrete kind of contract in issue. However, it is arguable that courts are more reluctant to 'interfere' with a bargain struck between comparatively equal parties (such as experienced merchant traders) than with one between parties with unequal bargaining power, particularly when the consent of the inferior bargaining party is in question (Bernstein, 1966).

Disparities in bargaining power between parties are nevertheless not necessarily grounds for courts 'interfering' with bargains. Common law rules of construction, such as the rule of construction that contracts ought to be construed against the drafter, usually the dominant party, are mere guidelines; rules of construction are not binding rules of law. Conversely, courts sometimes impose rules of fairness in so-called merchant contracts to prevent one party from being subjected to misrepresentation or unconscionable conduct by the other party, even though both parties are sophisticated traders.

More contentious perhaps is the debate over whether courts, as a matter of policy, should decide contracts cases primarily according to their business expediency (Kronman, 1980), or primarily according to their fairness (Dworkin, 1980). The sale of an end-user good at a discounted price under a contract that excludes warranty protection is conceivably efficient in promoting sales at discounted prices; however, it may be unfair to the purchaser who buys a defective good without warranty protection.

The Judicial 'Management' of Common Law Contracts

A particular concern for courts is in not interpreting contract terms too restrictively on the one hand, or constructing them too liberally on the other. As discussed in Section Freedom to and Freedom from Contracts, common law courts may use an objective test restrictively to determine the intention of contracting parties based on the literal meaning the court ascribes to the actual contractual language, such as the actual words contained in an employment or tenancy contract. A court that adopts a more liberal view may examine the context surrounding the contract as a whole, including the practices or habits of parties to particular kinds of employment or tenancy contracts and the laws, customs, and usages that govern them. An even more liberal judicial approach is to construct contract terms in employment and tenancy contracts according to standards of reasonableness that are based on judicial conceptions of equity that transcend both the literal text and, arguably, the practices and usages of the parties to such contracts (Devlin, 1951).

However, the liberal construction of contract terms is not altogether novel in common law jurisprudence. Courts have long used an objective test to establish the meaning of contract terms within the 'four corners' of the contract. Typically, they have implied terms into contracts that supersede the express terms of the contract (Costigan, 1920). These expansive methods of constructing contract terms are less controversial when courts imply terms mandated by statute, such as implied warranties of fitness under Sale of Goods legislation. Courts act somewhat more controversially when they imply terms into contracts based on their own conceptions of fairness, as distinct from standards prescribed by legislation, which expressly empower courts to imply such terms by operation of law (Farnsworth, 1968).

Such judicial management of contracts has led to disquiet over the perceived risk of courts overmanaging contracts and operating outside the boundaries of reasonable judicial interpretation. A principled criticism of such practices is that the 'bargain' should be determined, first and foremost, by the contracting parties themselves. Judges should neither 'make' nor 'remake' contracts. Nor should they imply terms into contracts based on amorphous and multifarious judicial conceptions of fairness or commercial expediency (Trakman, 1983). They should rather interpret contracts textually in order to regulate contracts efficiently (Schwartz and Scott, 2003).

A countervailing argument is that courts should manage contracts liberally precisely in order to preserve the freedom of the parties to contract, to ensure that parties freely and voluntarily consent to contract, and to prevent parties from being induced or deceived into contracting. It is arguable that courts should adopt such a liberal approach in construing, or even constructing, contract terms in their delivery of justice to the parties, that their resort to principles of equity and norms of social policy is not only appropriate but also necessary. Imputing fair terms into employment contracts is to ensure equitable treatment of employee, not only as a private matter between employer and employee but also as a social policy directed at the reasonable protection of employees in the workplace.

The prospect of judges diverging over the literal or liberal interpretation of contract terms in managing contracts is unavoidable in a common law system in which judges conceive of unfair surprise between contracting parties differently and adhere to disparate notions of social justice. It is also simplistic to insist that the late nineteenth and early twentieth centuries heralded the 'rise' of freedom of common law contracts while the period since then has given rise to the 'fall' of that freedom (Atiyah, 1979). Indeed, there are abundant illustrations of courts in the twenty-first century declining to construe

contracts liberally on diffuse grounds, including in order not to overextend their judicial mandate, and to not unjustifiably interfere with 'private' bargains.

A cautious compromise between freedom to and freedom from contract is for common law courts to maintain that they should manage contracts on predetermined principled grounds only. For example, they should reconcile the freedom to and the freedom from contract only when doing so complies with principled conceptions of consent in contracting. For example, courts should rescind a contract for misrepresentation in finding material evidence of a 'vice' or 'defect' in consent. Similarly, a court should render a contract unenforceable for frustration on principled grounds that the nonperforming party did not reasonably foresee and was not reasonably able to avert or mitigate the frustrating event. In each case, the court's power to manage the contract is restricted by the requirement that it apply the law of misrepresentation or frustration to contracts in a principled and not an arbitrary manner.

A critical response to this supposedly principled judicial approach toward contract construction is that courts do not arrive at the 'true' objective meaning of the contract. They reach neoliberal results out of neoliberal principles of contract law that are devised by empowered classes to perpetuate their privileges. In effect, the interacting principles of freedom to and freedom from contract idealize individual autonomy that empower wealthy white men who historically devised the liberal foundations of modern contract law and who have used contract law to perpetuate their dominance over working classes, ethnic minorities, and women (Dalton, 1985; Kennedy, 1982; Unger, 1986). Alternatively phrased, principles of law that are rationalized according to liberal conceptions of consent are more truly irrational according to any other measure of consent (Singer, 1984).

The tension between liberal and critical conceptions of the common law of contracts is likely to be played out on ever new judicial playing fields, not least of all in cyberspace. A sobering response is that, however innovative cyberspace may be technologically, the Internet has not changed the substance of contractual practice, only its form. However seemingly complex online e-consumer sales contracts may be, the fact is that incomprehensible fine print existed in consumer markets long before contracting in cyberspace was conceived. A welfare view is that legislatures and common law courts should energetically manage e-consumer markets to redress the opportunistic behavior of e-sellers that induce and deceive e-consumer end users into ill-advised e-contracts. A postwelfare – or neoliberal – view is that e-consumers should look after their own contractual interests, not expect legislatures and courts to do so for them. After all, e-consumers are better educated and informed than were traditional consumers. They also have access to the Internet, along with consumer blogs and Web sites, to assist them in reaching informed e-purchasing decisions.

Conclusion

It is too early to declare a victory for proponents of cautious, tempered, or aggressive legislative or judicial management of common law contracts. What is evident is that law makers in common law jurisdictions struggle to agree on why, how, and when to regulate e-markets. This struggle is demonstrated by the unwillingness of the vast majority of American states to enact the Uniform Computer Information Transactions Act developed by Uniform Law Commissioners to regulate e-markets. What is also apparent is that common law courts lack the capacity to regulate divergent e-sales contracts in a systematic and sustained manner (Trakman, 2011).

A salutary conclusion is not to expect common law courts to provide right or wrong answers in regulating often complex contractual practices. The common law is not a pure science. Nor are judges physicists. However principled common law courts purport to be, their decisions are not scientific truths. Nor should they be so conceived, in fact or in law.

Bibliography

Atiyah, P.S., 1979. The Rise and Fall of Freedom of Contract. Oxford University Press, Oxford, UK.

Bernstein, 1996. Merchant Law in a Merchant Court: Rethinking the Code's Search for Immanent Business Norms, vol. 144. University of Pennsylvania Law Review. pp. 1765–1821.

Coleman, J.L., 1980. Markets, Morals and the Law. Cambridge University Press, Cambridge, UK.

Costigan, G.P., 1920. Implied in fact contracts and mutual assent. Harvard Law Review 33, 376–400.

Dalton, C., 1985. An essay on the deconstruction of contract doctrine. Yale Law Journal 94, 997–1114.

Devlin, P., 1951. The relation between commercial law and commercial practice. Modern Law Review 14, 249–266.

Dworkin, R., 1980. Wealth as value. Journal of Legal Studies 9, 191–226.

Farnsworth, E.A., 1968. Disputes over omissions in contracts. Columbia Law Review 68, 860–891.

Fuller, L., Perdue, W.R.A., 1936. The reliance interest in contract damages. Yale Law Journal 46, 52–96.

Kennedy, D., 1982. Distributive and paternalistic motives in contract and tort. Maryland Law Review 41, 563–658.

Kessler, F., 1943. Contracts of adhesion—some thoughts about freedom of contract. Columbia Law Review 43, 629–642.

Kronman, A., 1980. Wealth maximization as a normative principle. Journal of Legal Studies 9, 227–242.

Llewellyn, K.N., 1930. What price contract? An essay in perspective. Yale Law Review 11, 704.

Macaulay, S., 1963. Non-contractual relations in business: a preliminary study. American Sociological Review 28, 55.

Macneil, I., 1980. The New Social Contract. Yale University Press, New Haven, CT.

Macneil, I., 1982. The efficient breach of contract: circles in the sky. Virginia Law Review 68, 947–969.

Patterson, E.W., 1924. The apportionment of business risks through legal devices. Columbia Law Review 24, 335–359.

Pound, R., 1908. Liberty of contract. Yale Law Journal 18, 454–487.

Schwartz, A., Scott, R.E., 2003. Contract theory and the limits of law. Yale Law Journal 113, 541–619.

Singer, J.W., 1984. The player and the cards: nihilism and legal theory. Yale Law Journal 94, 1–70.

Trakman, L.E., 1983. Frustrated contracts and legal fictions. Modern Law Review 46, 39–45.

Trakman, L.E., 2009. The boundaries of contract law in cyberspace. International Business Law Journal, 161–194. Revue de Droit des Affaires Internationales.

Trakman, L.E., 2011. From the medieval to the twenty first century law merchant. American Business Law Journal 48, 775–834.

Unger, R., 1986. The Critical Legal Studies Moment. Harvard University Press, Cambridge, MA.

Williston, S., 1921. Freedom of contract. Cornell Law Quarterly 6, 365–380.

Control Behavior: Psychological Perspectives

Jutta Heckhausen, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J. Heckhausen, volume 4, pp. 2719–2724, © 2001, Elsevier Ltd.

Abstract

The scope of control-related phenomena is extensive. Control behavior should be distinguished from perceptions or beliefs about control. The evolutionary origin of control-related behavior is the universal strive to achieve outcomes in the environment by one's own activity. Human agency unfolds its potential across developmental change throughout the life span. The motivational theory of life span development (MTD) proposes that primary control across the life span indicates a fundamental indicator of adaptive development. Primary control striving is organized into cycles of goal engagement and disengagement utilizing specific control strategies. Goal engagement and disengagement are not adaptive as such, but only when congruent with control opportunities given in the concurrent social and developmental ecology.

Psychological Perspectives on Control Behavior

The scope of control-related phenomena is extensive. It encompasses numerous concepts such as perceived control, personal control, cognitive control, vicarious control, illusory control, proxy control, outcome control, primary and secondary control, decisional control, and predictive control (Skinner, 1996). In her review of a decade of control-related research, Skinner (1996) identified some 100 different control-related constructs and cautioned that "the large number of terms has produced some theoretical confusion about the boundaries of the topic of control, about the interrelationships among constructs, and even about which constructs can be appropriately included in the study of control" (Skinner, 1996: p. 550). In order to avoid the widespread conceptual confusion, I begin with a set of definitions for the major control constructs are given. *Control-related behavior* is directed at producing contingencies between own behavior and events. When the events targeted are located in the outside world, the control-related behavior is *primary control behavior*. In contrast, when internal events are targeted it is *secondary control behavior*. *Control* is realized contingency between behavior and event. *Perceived control*, by contrast, refers to the extent to which an individual believes to control behavior-event contingencies. It thus becomes clear that perceived control does not reflect behavior, but internal representations, and may thus be involved in secondary control.

A Functional Analysis of Human Control Behavior: Evolutionary Origin and Human Characteristics of Control Behavior

A common characteristic of all motivated behavior is that the organism attempts to achieve outcomes in the environment by its own activity. In all activities relevant for survival and procreation, such as foraging, attracting a mate, or competing with a rival, the organism strives for control in terms of bringing about desired outcomes and preventing undesired ones. Thus, the most fundamental and universal motivational tendencies relate to this basic strive to control the environment referred to as *primary control striving* (Heckhausen, 2000; Heckhausen and Schulz, 1995). The striving for control is shared with a broad range of species and goes back far into the phylogenetic past; at least as far back as to those species, which first acquired a notable flexibility in their behavior programs (see review in Heckhausen, 2000). This assumption converges with White's (1959) classical article on the motivation for competence, effectance, and mastery as a universal striving of humans and mammals in general.

The fundamental status of control striving in behavior regulation is supported by findings about neonatal behavioral preferences (see review in Heckhausen and Schulz, 1995). Neonatal humans are able to detect behavior-event contingencies, learn head movements to control milk reinforcement, and most importantly, continue to show the control behavior even when after complete satiation, the milk reinforcement has lost its consummatory value and merely provides a contingency event to the infants' head-turning behavior.

Humans hold a unique potential for control in terms of pursuing goals, which can involve multiple subgoals and cover extensive time spans. This capacity is based on the mental representation of goals, which renders the organism independent from immediate external response elicitation and feedback, and sets it free for self-regulation.

Another important characteristic of humans' control-related behavior is their sense of self, and their quest for learning about and feeling good about the characteristics of their own self (e.g., about competence). The sense of self is acquired in early childhood and in itself becomes a motivational incentive for control striving (see review in Heckhausen and Schulz, 1995). Control goals are pursued not only for the sake of their outcomes, and the behavior-event contingency they represent, but also in view of the anticipated enhancement or confirmation of self-ascribed capacities to control outcomes. Mental representations of one's own control capacity and of the control potential inherent in a situational challenge are referred to as *perceived control* (Skinner, 1996).

Particularly for challenging control pursuits, the anticipation of positive information about one's own competence implied in a successful completion of the control striving, becomes a major resource for motivation. Anticipatory positive

self-evaluation motivates the individual to persistently invest effort into acquisitional processes. Only this way can time-extended acquisitional processes of expertise development be sustained motivationally. However, this ability to develop conceptions about one's own competence and control also renders the human motivational system vulnerable to experiences of failure. From about 4 years of age, humans perceive failure as negative feedback about their own competence and react with embarrassment and shame (see Heckhausen and Schulz, 1995). For humans older than 4 years of age, failure may thus undermine self-ascribed competence and perceived control, both important motivational resources needed for persistent primary control striving.

The system of control behavior in humans therefore needs to involve processes that protect and restore conceptions of own competence and mastery. Heckhausen and Schulz in their motivational theory of life span development (formerly "life span theory of control; " Heckhausen and Schulz, 1995; Heckhausen et al., 2010; Schulz and Heckhausen, 1996) have proposed to subsume diverse, mostly cognitive processes, which serve to protect motivational resources and focus them on a given goal (i.e., volitional commitment) of primary control striving, under the construct of *secondary control striving*. Among such secondary control processes are those that help the individual to disengage from a futile goal (e.g., devalue goal, increase value of alternative goals) and protect the self against unfavorable implications of failure and loss (e.g., downward social comparison, egotistic causal attributions). In addition, secondary control strategies are also needed for generating and maintaining motivational commitment to a chosen goal (e.g., enhance value of chosen goal, devalue alternatives, overestimate perceived control for goal attainment; Gollwitzer and Kinney, 1989; Heckhausen, 1999; Wrosch and Heckhausen, 1999).

Motivation and Life Span Development: The Role of Individual Agency

Human ontogenetic potential is vast and that has not just beneficial implications but also gives rise to a need for regulation so that organized and reasonably predictable developmental change and transitions will occur. Regulation can happen via a variety of influencing systems, the biology of maturation and aging, the societal timing of transitions (e.g., school entry, retirement) and canalization of life course paths, and last but not least the individual agent with his/her goals for their own development and life achievements. A motivational approach to life span development focuses on the individual agent who operates in a time-extended action field of their life course. This action field is of course influenced in terms of opening opportunities and setting constraints, by the individual's biological characteristics and by their social context (micro, meso, macro; Bronfenbrenner and Morris, 2006).

Adaptiveness of development can be estimated by determining the degree to which the individual has realized an optimal degree of control in different domains of life across different times of the life span. Criteria for adaptive development should take into account the unique constraints of the individual (e.g., disability), be comprehensive covering multiple important domains of functioning and the grand total of mastery across the individual's life span (Heckhausen et al., 2010).

Most developmental researchers would agree that the individual plays a critical role in shaping their development and life course path (Baltes et al., 1998; Brandtstädter, 2006; Heckhausen, 1999; Lerner and Busch-Rossnagel, 1981). Individual agency has been conceptually and empirically elaborated in action models of developmental regulation, most notably the dual-process model by Brandtstädter and his colleagues (see summary in Brandtstädter, 2006) and the motivational theory of life span development (Heckhausen and Schulz, 1995; Heckhausen et al., 2010; Schulz and Heckhausen, 1996). This article presents the motivational theory of life span development. Discussions about the differences and similarities with the dual-process model can be found elsewhere (Haase et al., 2013; Poulin et al., 2005).

The MTD proposes that the desire to exert control over ones environment and thus realize primary control rules the system of control behavior. MTD elaborates the distinction between primary and secondary control, originally proposed by Rothbaum et al. (1982) and applies it to the human life span. *Primary control striving* is expected to remain the dominant motivator of behavior across the life span, while the potential to realize primary control undergoes radical changes across age (see also Heckhausen, 1999). The capacity for primary control, in contrast, increases during childhood, adolescence, and young adulthood, then reaches a maximum and plateau during midlife, and declines with the loss of social roles and physical fitness associated with old age. The increasing discrepancy between primary control striving and primary control capacity at older ages provides a challenge for the individual, in that it can only be managed by disengagement from age-inappropriate goals and engaging with new more age-adapted goals. Such cycles of goal disengagement and engagement can only be mastered when the individual employs specific strategies addressing the internal world of the individual in terms of goal hierarchies and motivational commitment. Such strategies are conceptualized as 'secondary control strategies.'

Heckhausen and Schulz (1995) and Heckhausen (1999) distinguish between four types of control strategies: *Selective primary control* strategies refer to the investment of internal behavioral resources such as effort, time, and skills in the pursuit of a goal. *Compensatory primary control* is required when internally available behavioral resources of the individual are insufficient to attain the goal, and external resources such as help or advice from others or technical aids (e.g., wheelchair) have to be recruited. *Selective secondary control* serves to enhance and maintain motivational commitment to a chosen action goal, particularly when it is challenged by unexpected obstacles or attractive alternatives. Selective secondary control strategies include enhanced valuation of the chosen goal, and devaluation of nonchosen alternatives, as well as positive illusions about one's control, potential for the chosen goal. *Compensatory secondary control* is required when the individual experiences a loss of control and when the goal becomes unattainable or excessively costly. Compensatory secondary control can be attained by disengaging from the obsolete goal. In addition, compensatory secondary control involves specific self-protective strategies, such as self-protective causal attribution

(avoid self-blame) and downward social comparisons, which deflect the potential negative effects of failure experiences on important motivational resources of affective balance and self-esteem.

The Action-Phase Model of Developmental Regulation

A key proposition of the MTD is that effective developmental agency needs to focus on goals for which the opportunities in the biological and social setting are favorable. Once opportunities fade away (e.g., biological clock for childbearing goal) the individual needs to disengage from the formerly productive but now futile goal and reinvest their motivational and behavioral resources in goals that are more feasible. This is the *principle of congruence* between control opportunities and control striving (Heckhausen et al., 2010).

MTD has integrated models of action regulation, which view goal-directed behavior as organized in action cycles of phases of goal selection, goal engagement, and goal deactivation and reevaluation (Gollwitzer et al., 1990; Heckhausen, 1991). Specifically, the action-phase model of developmental regulation in MTD is shown in **Figure 1**. The action cycle comprises the phases of choosing a goal, engaging motivationally and behaviorally with it, encountering depleted opportunities for goal attainment (developmental deadline), and disengaging from the goal when it is no longer attainable.

During the first phase of optimized goal choice, individuals should use three general heuristics. First, they chose goals that are *congruent* with their control capacity and that varies greatly across age. Second, they select those goals that minimize unfavorable, and maximize favorable, *consequences for other domains* of functioning. And third, individuals try to maintain some *cross-goal diversity* in their goal pursuits to maintain developmental options in case of failure. Once the individual has made a decision about which goal to pursue and thus crossed the decisional Rubicon (see Heckhausen, 1991; also Achtziger and Gollwitzer, 2010), he/she moves into the phase of goal engagement. An important aspect of this phase is that perceptions and information processing become biased toward the goal being pursued. For example, when women are engaged with the goal of bearing a child, their information processing (e.g., what they recall in an incidental memory task) is attuned to child-relevant topics (Heckhausen et al., 2001). For many goal pursuits the opportunities vary over time and across the life span, and can be characterized by an inverted U-shaped trajectory of age-graded changes from steep increase during childhood and adolescence to peak in midlife and decrease in old age. As opportunities (e.g., for childbearing, a final career advancement) fade away, the pursuit of these goals becomes increasingly urgent and thus individuals experience an urgency to intensify their goal engagement. Finally, once the opportunities for goal attainment have slipped away to a degree that makes further pursuit futile (i.e., developmental deadline), individuals need to switch from goal engagement to goal disengagement.

An important proposition of the action-phase model is that the shifts from goal choice to goal engagement and from goal engagement to goal disengagement are conceptualized as discrete and radical, not gradual or continuous. Behavior and thoughts should not oscillate between engagement and disengagement, but be coordinated with regard to the given function of the action phase (either goal choice, goal engagement, or goal disengagement). This idea of radical and discrete switching between action phases has not yet been addressed in empirical research, but is a promising area for future investigation in many domains of functioning, including education, career achievement, and managing health problems.

Note that there is considerable variability in individuals' ability to negotiate important life transitions. As a consequence, different individuals attain different long-term outcomes of developmental regulation, both subjectively and objectively (Heckhausen et al., 2010). Those who have major difficulties with choosing appropriate goals, engaging in attainable goals, or disengaging from unattainable goals waste their control efforts and deplete their resources for future primary control striving. On the other hand, those individuals who manage these transitions in accordance with their capacity

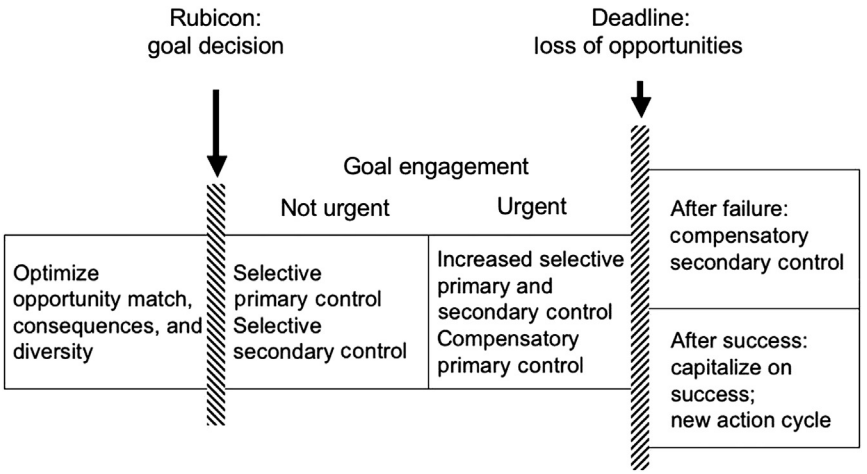

Figure 1 Action-phase model of developmental regulation. Adapted from Heckhausen, J., 1999. Developmental Regulation in Adulthood: Age-normative and Sociostructural Constraints as Adaptive Challenges. Cambridge University Press, New York.

for control maximize their developmental potential and the likelihood of successful life span development.

Developmental Regulation Around Developmental Deadlines

One way to study the responsiveness of control behavior to different control potential is to study developmental regulation across distinct life course transitions in control potential (Heckhausen, 1999). Developmental regulation across the life span is organized along a timetable of waxing and waning opportunities to attain important developmental goals, such as graduating from school, building a career, finding a permanent partner, and having a child. The age-graded opportunity structures for various developmental goals involve final time constraints, conceptualized as developmental deadlines.

An example is the so-called 'biological clock' for childbearing (Heckhausen et al., 2001). Developmental deadlines were shown to be the watersheds between urgent predeadline goal engagement and postdeadline goal disengagement and self-protection (Wrosch and Heckhausen, 1999). Individuals approaching a developmental deadline, such as the early 40s in the case of childbearing, activated goal engagement strategies of control (i.e., selective primary, selective secondary, compensatory primary control) fervently striving to attain the goal before the time runs out. In contrast, once a deadline has been passed (e.g., women in their early 40s or 50s with regard to childbearing), control strategies of goal disengagement, such as devaluing the goal and self-protective downward comparison with others were preferred. Moreover, it was shown that the employment of phase-congruent control strategies (i.e., goal engagement in the predeadline phase and goal disengagement in the postdeadline phase) was associated with greater psychological well-being and mental health (e.g., less depressive symptoms, more positive affect) than phase-incongruent control strategies. In addition, effects of congruence in goal engagement and disengagement were also found for nonintentional biases in information processing. Postdeadline women who recalled items pertaining in topic to childrearing and childbearing reported enhanced negative affect.

Another deadline-related study addressed the congruence of partnership-related goal engagement and disengagement of recently separated young or late-middle-aged men and women with census-based estimates of their remarriage probability (Wrosch and Heckhausen, 1999). Younger adults suffered a deterioration of positive affect if they disengaged from partnership goals after a separation, whereas adults in their 50s benefitted from such disengagement.

The deadline idea was extended to the area of health management. Here the notion is that certain health problems are controllable while others are not and individuals should adapt their goal engagement/disengagement accordingly. In a study of elderly caregivers, health engagement control strategies were associated with decline in depressive symptoms over time, but only for those caregivers who had acute and controllable health problems (Wrosch et al., 2002). Most striking evidence comes from a study of older Canadian adults with serious health problems that were either acute and controllable (post-stroke or post-heart attack rehabilitation) or chronic and uncontrollable (arthritis, heart disease) (Hall et al., 2010). Patients were asked about their most serious health-related restriction in an activity of daily living and whether or not they try to overcome it (goal engagement) or have accepted it (goal disengagement). Goal engagement strategies were associated with greater longevity and better health status for the patients with acute compared to chronic health problems. In contrast, goal disengagement predicted poorer physical health among those who suffered from acute conditions, but better health for those who suffered from chronic conditions.

Goal Adjustment and Goal Engagement in the Transition to Adulthood

In addition to the research addressing the suitability of the action-phase model to reflect the process of goal engagement and disengagement around developmental deadlines and serious health problems, we also investigated processes of goal engagement and managing goal adjustment in the transition to adulthood (e.g., Haase et al., 2008; Heckhausen and Chang, 2009; Heckhausen et al., 2013; Heckhausen and Tomasik, 2002; Shane et al., 2012). The transition to adulthood is a time where more is at stake than at any other time in the life course. This transition decides over whether or not youth maintain or exceed their parents' social status in society or whether they lose status. In societies with much social mobility a lot depends on what the individual youth makes of their opportunities. Different societies, however, may more or less scaffold young people to optimize outcomes in this transition and such scaffolding may pertain to educational paths and/or to paths into vocational careers (Heckhausen, 2002; Heckhausen and Chang, 2009).

We conducted two longitudinal studies of youth's transition to adulthood, one set in the German system of entering vocational training (i.e., applying for apprenticeships after finishing the 10th and final grade of the middle tier of the German school system), and one set in the California multi-tiered college system of transitions after high school graduation. The two systems contrast on the important dimensions of permeability and transparency (Hamilton, 1994). In the German vocational training system (apprenticeship combined with 2 days of school) employers largely rely on school grade information when making decisions about who to hire as an apprentice in more (bank clerk, insurance broker) or less (construction worker) prestigious vocations. The training as an apprentice qualifies for a career in a highly skilled (nonacademic) vocation, but not for entry into college (except via some elaborate educational detour). This means there is high predictability (what does it take to get into certain vocational tracks), but low permeability between vocational and professional career tracks. In contrast, in the United States and particularly in California with its elaborate system of Community Colleges, State universities, and research universities predictability into careers is low, but permeability between jobs, vocational training, college education, and respective careers is high.

We were able to show that in both of these societal contexts the extent of goal engagement was critical for predicting long-term success and well-being (Haase et al., 2008; Heckhausen et al., 2013). In the German study the critical motivational

engagement was with goals of attaining an apprenticeship, whereas in the U.S. context it was an engagement with educational but not with career goals that predicted educational attainment, progress toward career goals, and well-being 4 years after high school graduation (Heckhausen et al., 2013).

However, the adjustment of goals differed in the two contexts. In the more predictable and less permeable context of the German vocational training system, youths closely calibrated their goals in terms of vocational prestige to their relative school grades, and thereby maximized their career prospects while making sure that they did not overshoot in their ambition and ended up empty-handed (Heckhausen and Tomasik, 2002; Tomasik et al., 2009). In contrast, in the less predictable but more permeable California system of higher education, the youth with the highest educational ambitions were most likely to succeed in the long run (4 years after college) (Heckhausen and Chang, 2009).

In sum, research on goal selection and goal engagement in the transition to adulthood shows that goal-focused motivational engagement is essential across different societal conditions (at least in Western industrialized societies). However, different societal conditions (educational and vocational training systems) render different goal adaptations in terms of ambitiousness and calibration more adaptive and this is reflected both in the overall patterns of goal selections and their associations with desired outcomes.

Engagement and Disengagement from Career Goals

MTD was also applied in the domain of adult developmental regulation of career and work trajectories. Using the Midlife in the U.S. study (Shane and Heckhausen, 2013), we found that adults who report high congruence between high control perceptions and high goal engagement regarding one's career goals attain the best developmental outcomes a decade later, both in terms of career outcomes and with regard to subjective well-being. Those adults who report relatively low control over their career prospects yet high engagement with career goals tend to achieve greater career improvements a decade later than those who gave up ambitious career goals. However, they do this at the cost of undermining their long-term primary control striving and their mental and physical health.

A study conducted in the context of post-reunification Germany showed the benefits of disengagement from work-related goals for those who live in regions with extremely impoverished prospects for employment and career progress (Tomasik et al., 2010). Thus, under conditions of severely constrained primary control, giving up even such central goals as finding gainful employment can be adaptive.

Further Research and Future Directions

Further research using the MTD has investigated the implications of opportunity-congruent engagement and particularly disengagement for physical functioning and the physiological stress system (e.g., Wrosch et al., 2009; Wrosch and Schulz, 2008).

Future research about control behavior and developmental regulation using the MTD should investigate the processes involved in goal engagement and disengagement in a micro-sequential manner using experimental paradigms as well as concurrent assessment of affect and physiological indicators such as heart rate, skin conductance, hormonal and brain-activation patterns as correlates of engagement and disengagement phases in the action cycle.

Another important area of future research is the development of individual differences in motivational predispositions for engagement, disengagement, and sensitivity to control opportunities and constraints. In this context, gene–environment interactions need to be explored particularly with regard to sensitive developmental periods in early childhood and adolescence.

See also: Academic Engagement; Affect-Regulation Motivation; Apprenticeship and School Learning: Lessons from Germany; Burnout, Psychology of; Coping across the Lifespan; Epistemological Beliefs and Motivation; Grit; Health Self-Regulation, Motivational and Volitional Aspects of; Motivation and Actions, Psychology of; Passion and Motivation; Personal Projects; Schooling: Impact on Cognitive and Motivational Development; Self and Emotional Development in Adulthood and Later Life; Self-Regulation in Adulthood; Social and Economic Change: Psychological Challenges for Individuals; Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model.

Bibliography

Achtziger, A., Gollwitzer, P.M., 2010. Motivation and volition in the course of action. In: Heckhausen, J., Heckhausen, H. (Eds.), Motivation and Action. Cambridge University Press, New York, NJ, pp. 272–295.

Baltes, P.B., Lindenberger, U., Staudinger, U.M., 1998. Life-span theory in developmental psychology. In: Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology, vol. 5. Wiley, New York, pp. 1029–1143.

Bandura, A., 1977. Self-efficacy: toward an unifying theory of behavioral change. Psychological Review 84, 191–215.

Brandtstädter, J., 2006. Action perspectives on human development. In: Lerner, R.M., Damon, W. (Eds.), Handbook of Child Psychology, sixth ed., vol. 1. Wiley, Hoboken, NJ, pp. 516–568.

Bronfenbrenner, U., Morris, P.A., 2006. The bioecological model of human development. In: Lerner, R.M., Damon, W. (Eds.), Handbook of Child Psychology, , Theoretical Models of Human Development, sixth ed., vol. 1. John Wiley, Hoboken, NJ, US, pp. 793–828.

Gollwitzer, P.M., Heckhausen, H., Steller, B., 1990. Deliberative and implemental mind-sets: cognitive tuning toward congruous thoughts and information. Journal of Personality and Social Psychology 59, 1119–1127.

Gollwitzer, P.M., Kinney, R.F., 1989. Effects of deliberative and implemental mind-set on illusion of control. Journal of Personality and Social Psychology 56, 531–542.

Haase, C.M., Heckhausen, J., Köller, O., 2008. Goal engagement during the school-to-work transition: beneficial for all, particularly for girls. Journal of Research on Adolescence 18, 671–698.

Haase, C., Heckhausen, J., Wrosch, C., 2013. Developmental regulation across the life span: towards a new synthesis. Developmental Psychology 49, 964–972.

Hall, N.C., Chipperfield, J.G., Heckhausen, J., Perry, R.P., 2010. Control striving in older adults with serious health problems: a 9-year longitudinal study of survival, health, and well-being. Psychology and Aging 25, 432–445.

Hamilton, S.F., 1994. Employment prospects as motivation for school achievement: links and gaps between school and work in seven countries. In: Silbereisen, R.K., Todt, E. (Eds.), Adolescence in Context: The Interplay of Family, School, Peers, and Work in Adjustment. Cambridge University Press, Cambridge, UK, pp. 267–303.

Heckhausen, H., 1991. Motivation and Action. Springer, New York.
Heckhausen, J., 1999. Developmental Regulation in Adulthood: Age-normative and Sociostructural Constraints as Adaptive Challenges. Cambridge University Press, New York.
Heckhausen, J., 2000. Evolutionary perspectives on human motivation. American Behavioral Scientist 43, 1015–1029.
Heckhausen, J., 2002. Transition from school to work: societal opportunities and the potential for individual agency. Journal of Vocational Behavior 60, 173–177.
Heckhausen, J., Chang, E.S., 2009. Can ambition help overcome social inequality in the transition to adulthood? Individual agency and societal opportunities in Germany and the United States. Research in Human Development 6, 1–17.
Heckhausen, J., Chang, E., Greenberger, E., Chen, C., 2013. Striving for educational and career goals during the transition after high school: what is beneficial? Journal of Youth and Adolescence 42, 1385–1398.
Heckhausen, J., Schulz, R., 1995. A life-span theory of control. Psychological Review 102, 284–304.
Heckhausen, J., Tomasik, M.J., 2002. Get an apprenticeship before school is out: how German adolescents adjust vocational aspirations when getting close to a developmental deadline. Journal of Vocational Behavior 60, 199–219.
Heckhausen, J., Wrosch, C., Fleeson, W., 2001. Developmental regulation before and after a developmental deadline: the sample case of "biological clock" for childbearing. Psychology and Aging 16, 400–413.
Heckhausen, J., Wrosch, C., Schulz, R., 2010. A motivational theory of life-span development. Psychological Review 117, 32–60.
Lerner, R.M., Busch-Rossnagel, N.A. (Eds.), 1981. Individuals as Producers of Their Development: A Life-span Perspective. Academic Press, New York, NY.
Poulin, M., Haase, C., Heckhausen, J., 2005. Engagement with and disengagement from goals across the life span: a comparison of two-process models of developmental regulation. In: Greve, W., Rothermund, K., Wentura, D. (Eds.), The Adaptive Self: Personal Continuity and Intentional Self-development. Hogrefe/Huber Publisher, Göttingen/New York, pp. 1–19.
Rothbaum, F., Weisz, J.R., Snyder, S.S., 1982. Changing the world and changing the self: a two process model of perceived control. Journal of Personality and Social Psychology 42, 5–37.
Schulz, R., Heckhausen, J., 1996. A life-span model of successful aging. American Psychologist 51, 702–714.
Schulz, R., Heckhausen, J., 1999. Aging, culture and control: setting a new research agenda. Journal of Gerontology: Psychological Sciences 54B, P139–P145.
Shane, J., Heckhausen, J., 2013. Motivational self-regulation in the work domain: congruence of individuals' control striving and the control potential in their developmental ecologies. Research in Human Development 9, 337–357.
Shane, J., Heckhausen, J., Lessard, J., Chen, C., Greenberger, E., 2012. Career-related goal pursuit among post-high school youth: relations between personal control beliefs and control strivings. Motivation and Emotion 36, 159–169.
Skinner, E.A., 1996. A guide to constructs of control. Journal of Personality and Social Psychology 71, 549–570.
Tomasik, M.J., Hardy, S., Haase, C.M., Heckhausen, J., 2009. Adaptive adjustment of vocational aspirations among German youths during the transition from school to work. Journal of Vocational Behavior 74, 38–46.
Tomasik, M., Silbereisen, R.K., Heckhausen, J., 2010. Is it adaptive to disengage from demands of social change? Adjustment to developmental barriers in opportunity-deprived regions. Motivation and Emotion 34, 384–398.
Weisz, J.R., 1983. Can I control it? the pursuit of veridical answers across the life span. In: Baltes, P.B., Brim, O.G. (Eds.), Life-span development and behavior, vol. 3. Academic Press, New York, pp. 233–300.
White, R.W., 1959. Motivation reconsidered: the concept of competence. Psychological Review 66, 297–333.
Wrosch, C., Heckhausen, J., 1999. Control processes before and after passing a developmental deadline: activation and deactivation of intimate relationship goals. Journal of Personality and Social Psychology 77, 415–427.
Wrosch, C., Miller, G.E., Schulz, R., 2009. Cortisol secretion and functional disabilities in old age: Importance of using adaptive control strategies. Psychosomatic Medicine 71, 996–1003.
Wrosch, C., Schulz, R., 2008. Health-engagement control strategies and 2-year changes in older adults' physical health. Psychological Science 19, 537–541.
Wrosch, C., Schulz, R., Heckhausen, J., 2002. Health stresses and depressive symptomatology in the elderly: the importance of health engagement control strategies. Health Psychology 21, 340–348.

Control Beliefs: Health Perspectives

Kenneth A Wallston, Vanderbilt University, Nashville, TN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article is about the various ways in which beliefs about control have been defined and operationalized in the context of health. Constructs covered include health locus of control, health self-efficacy (along with related constructs of mastery and competence), perceived behavioral and situational control, and helplessness. The last two sections discuss methods for assessing health control beliefs and altering those beliefs, respectively.

Introduction

Beliefs about control have been related to many aspects of health, including health behavior, health outcomes, and health care. In general, the greater one's sense of control, the more likely one will engage in appropriate behavior and the better one's outcomes. However, the relationship between control beliefs, behavior, and outcomes is not always straightforward. This article covers the following topics: locus of control; self-efficacy, mastery, and competence; perceived behavioral control; perceived situational control; helplessness; assessing control beliefs; and altering control beliefs.

Control refers to the ability to influence what is happening or what will happen. *Beliefs* about control in the context of health refer to the thoughts (or cognitions) that an individual has regarding the ability to influence health behavior, health status (or other health outcomes), or health care. Control beliefs are one of a number of determinants of health behavior and health outcomes. A person's health status can influence and can be influenced by his or her control beliefs. Individuals who believe that they have control over their health behavior are more likely to engage in healthy behavior; thus, they are more likely to have better health outcomes (*see* Health Behaviors). Better health outcomes are associated with more positive control beliefs. Also relevant are beliefs about control over health care delivery. Believing that one can influence one's own health care is a primary determinant of satisfaction with health care. Perceived control over one's health behavior, health outcomes, and/or health care is different from *desire* for control over these aspects. It is assumed that most individuals desire control over their health behavior and health outcomes; no such assumption is made regarding control over how health care is delivered.

Locus of Control

'Locus' refers to the *place* where control (over outcomes) resides – either 'internal' to the individual (i.e., based on his or her own behavior or enduring characteristics) or 'external' to the individual (e.g., due to fate, luck, or chance; other people's behavior; or the influence of God or 'a higher power'). Although it was once thought that *internal* and *external* beliefs were at opposite ends of a continuum (Rotter, 1966), it is now understood that these two belief orientations are independent of one another. That is, a person might simultaneously hold internal and external beliefs about the locus of control of a given phenomenon (e.g., his or her health status; Wallston, 1989, 2001). For example, a person newly diagnosed with lung cancer might attribute the diagnosis to internal factors such as a lifetime of smoking cigarettes in addition to poor dietary and exercise habits. At the same time, the person might attribute the lung cancer to external causes such as 'bad genes' (inherited from one's parents), punishment from God, pollution in the environment, and/or poor luck. Furthermore, one's locus of control orientation about the cause of a health problem (its etiology) is unrelated to one's beliefs about responsibility for the course of the problem (i.e., whether it will get better or worse). It is possible to attribute the etiology of a disease to external causes (e.g., a virus) and to adopt an internal orientation to dealing with the condition once diagnosed (e.g., learn everything one can about the condition and its treatment). Many patients with a medical condition believe that health care professionals either fully or partially control what happens to their health status.

Better health outcomes are typically associated with more internal beliefs about control of one's health. However, an internal health locus of control orientation does not necessarily signify that one attributes only good health outcomes to one's own behavior. For some individuals, having 'internal' beliefs may mean blaming oneself for one's poor health outcomes. Being 'responsible for' one's outcomes does not always translate into being in control of those outcomes. Furthermore, internal beliefs about one's health are only weakly predictive of engaging in healthy behavior or avoiding health risks. Internal beliefs are somewhat more predictive of healthy behavior for people who highly value having good health compared with those for whom good health is less important, but locus of control explains only a relatively small amount of variance in health behavior even for people who highly value good health (Wallston, 1992). Valuing good health and believing that one's health is contingent on one's own actions are necessary, but not sufficient, determinants of health behavior.

Self-Efficacy, Mastery, and Competence

Locus of control is only one type of control belief. Other psychological constructs that are similar to locus of control are *self-efficacy* (Bandura, 1977), *mastery* (Pearlin and Schooler,

1978), and *competence* (White, 1959). Self-efficacy, or the belief that one *can do* a specific behavior in a specific situation, is much more predictive of actually engaging in that behavior in that situation than is an internal locus of control orientation (*see* Self-Efficacy and Health). Both mastery and competence are more generalized constructs than self-efficacy, and they encompass control over the situation and outcomes as well as over behavior. Individuals with a sense of self-efficacy, mastery, or personal competence (and also those with an internal locus of control orientation) generally feel very good about themselves (i.e., have high self-esteem and psychological well-being) and are receptive to learning about and engaging in new behavior. An internal locus of control orientation coupled with a high degree of self-efficacy, mastery, or personal competence is a potent resource for helping individuals cope with health-related stressors.

Perceived Behavioral Control

Ajzen (1988) introduced the construct 'perceived behavioral control' into his theory of planned behavior as a determinant of both behavioral intention and the behavior itself. On a conceptual basis, perceived behavioral control is similar to self-efficacy – both constructs refer to the person's belief that the behavior in question is under his or her control – but, operationally, perceived behavioral control is often assessed by the ease or difficulty of the behavior (e.g., "I find it difficult to exercise three times a week"), while self-efficacy is operationalized by the individual's confidence in being able to carry out the behavior in the face of extenuating circumstances (e.g., "I am confident that I can exercise three times a week even when I am away on vacation").

Like Ajzen's theory of planned behavior, most psychosocial theories of health behavior (e.g., the health belief model (Rosenstock, 1990), protection motivation theory (Prentice-Dunn and Rogers, 1986), and the health action process approach (Schwarzer, 1999)) have incorporated perceived behavioral control as a major determinant of intention to engage in a health behavior, and/or as a determinant of actually engaging in the behavior, but these last three theories explicitly label this construct as 'self-efficacy.' Regardless of the label attached to the construct, believing that one has control over a behavior is highly associated with actually doing that behavior.

Perceived Situational Control

Beliefs about control encompass control over one's situation, as well as control over one's behavior and outcomes. One example of a situation is being a patient in a clinical facility. Does the patient know why he or she is there? Does the patient know what will happen to him or her, and why? What it will feel like? Does the patient have any choice over what will happen? The answers to these questions partially determine the patient's degree of perceived situational control. In turn, the more control the patient believes he or she has in that situation, the less distressed and more satisfied the patient will be, and the more likely the patient will be able to participate in his or her own care, including adhering to situational demands. This is true regardless of how much control the patient wants in that situation (cf Wallston, 1989).

Helplessness

Helplessness is the belief that there is nothing that anyone can do to improve a bad situation (such as being diagnosed with an illness). In many ways, then, helplessness is a belief that control over the situation or its outcomes is impossible. Like all beliefs, helplessness is learned (Seligman, 1975). Learned helplessness is analogous to a chance locus of control orientation (i.e., the belief that one's outcomes are due to fate, luck, or chance). Helplessness beliefs can be either *universal* (i.e., there is nothing that *anyone* can do) or *personal* (i.e., there is nothing that *I* can do). Either type of helplessness belief is associated with motivational, behavioral, and/or affective deficits. People with learned helplessness are not inclined to learn or engage in new, potentially effective behavior, and they exhibit higher-than-normal levels of anxiety and depression.

Assessing Control Beliefs

A number of well-developed, standardized instruments exist for the purpose of assessing individual differences in control beliefs applicable to health and health care settings (Wallston, 1989). Some of these measures, such as the Perceived Health Competence Scale (Smith et al., 1995) are unidimensional. Others, such as the health locus of control scales (Wallston, 2005), are multidimensional and can be adapted to assess control over one's health status in general or control over specified health conditions (e.g., cancer or diabetes). Researchers have also had some success in asking one or two pointed questions about subjects' perceptions of control (see Wallston, 2001, for examples), but the psychometric soundness of this approach is questionable. Measures of desire for control are only weakly correlated with measures of perceived control.

Altering Control Beliefs

Because control beliefs are learned, they can easily change as a function of life events (e.g., receiving a new medical diagnosis), or they can be systematically modified through the application of some form of cognitive behavioral therapy. Shapiro and Astin (1998) developed an integrated approach to psychotherapy, health, and healing that they term 'control therapy.' Unfortunately, that work has stimulated little empirical research by health researchers. However, there has been a great deal of health research having to do with control beliefs based on the theory of self-regulation (cf Boekaerts et al., 2000; Vohs and Baumeister, 2011), and a growing number of investigators are basing their work on control theory, an interdisciplinary branch of engineering and mathematics that deals with the behavior of dynamical systems with inputs (such as beliefs about control over health status).

See also: Explanatory Style and Health; Health Behavior, Psychosocial Theories of; Self-Efficacy and Health; Self-Efficacy.

Bibliography

Ajzen, I., 1988. Attitudes, Personality, and Behavior. Dorsey Press, Chicago.

Bandura, A., 1977. Self-efficacy: toward a unifying theory of behavior change. Psychological Bulletin 84, 191–215.

Boekaerts, M., Pintrich, P.R., Zeidner, M. (Eds.), 2000. Handbook of Self-Regulation. Academic Press, San Diego.

Pearlin, L.I., Schooler, C., 1978. The structure of coping. Journal of Health and Social Behavior 19, 2–21.

Prentice-Dunn, S., Rogers, R.W., 1986. Protection motivation theory and preventive health: beyond the health belief model. Health Education Research, Theory and Practice 1, 153–161.

Rosenstock, I.M., 1990. The health belief model: explaining health behavior through expectancies. In: Glanz, K., Lewis, F.M., Rimer, B.K. (Eds.), Health Behavior and Health Education: Theory, Research, and Practice. Jossey-Bass, San Francisco, pp. 39–62.

Rotter, J.B., 1966. Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs 80, 1–28.

Schwarzer, R., 1999. Optimism, goals, and threats: how to conceptualize self-regulatory processes in the adoption and maintenance of health behaviors. Psychology and Health 3, 759–766.

Seligman, M.E.P., 1975. Helplessness. W.H. Freeman, San Francisco.

Shapiro Jr., D.H., Astin, J.A., 1998. Control Therapy: An Integrated Approach to Psychotherapy, Health, and Healing. Wiley, New York.

Smith, M.S., Wallston, K.A., Smith, C.A., 1995. The development and validation of the perceived health competence scale. Health Education Research 10, 51–64.

Vohs, K.D., Baumeister, R.F. (Eds.), 2011. Handbook of Self-Regulation: Research, Theory, and Applications, second ed. Guilford Press, New York.

Wallston, K.A., 1989. Assessment of control in health care settings. In: Steptoe, A., Appels, A. (Eds.), Stress, Personal Control and Health. Wiley, Chichester, UK, pp. 85–105.

Wallston, K.A., 1992. Hocus-pocus, the focus isn't strictly on locus: Rotter's social learning theory modified for health. Cognitive Therapy and Research 16, 183–199.

Wallston, K.A., 2001. Conceptualization and operationalization of perceived control. In: Baum, A., Revenson, T.A., Singer, J.E. (Eds.), The Handbook of Health Psychology. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 49–58.

Wallston, K.A., 2005. The validity of the multidimensional health locus of control scales. Journal of Health Psychology 10, 623–632.

White, R.W., 1959. Motivation reconsidered: the concept of competence. Psychological Review 66, 297–333.

Control, Social

Maurice Cusson, Université de Montréal, Montréal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

At the beginning of the twentieth century, social control referred to those aspects of social life contributing to the social order, in fact making social order possible. Later, in the writings of sociologists and historians studying prisons, mental institutions, and the like, social control came to mean either a label putting in motion some amplification of deviance process or the organized way in which society reacts to deviance. The term was given a pejorative overtone, meaning coercion, manipulation, and stigmatization. In contemporary criminology, the term is often used in a theory trying to explain juvenile delinquency by a failure of informal controls and to explain the level of crime and violence in neighborhoods. The concept of social control is used by anthropologists to describe the methods by which different cultures resolve conflicts and enforce norms. The social control of crime also refers to all the means specifically aimed at reducing the probability or severity of crime. In the latter case, social control is not so much conceived as the influence of society on individuals, but as the intentional influence by social actors.

An important concept of US sociology at the beginning of the twentieth century, social control referred to those aspects of social life contributing to the social order. It was used to answer a perennial question in sociology: How is it that social order is possible? In the 1970s, a shift in meaning became noticeable. Sociologists and historians studying state institutions designed to deal with deviants, like prisons and mental health facilities, gave it a pejorative connotation: social control came to mean coercion, manipulation, labeling, and stigmatization. At the beginning of the twenty-first century, the term is more often used in criminology than in sociology. Two different meanings can be found. First, in control theories, it means informal social controls imposed on juveniles within the family and the school. Second, it is sometimes used as a synthetic term to include all the actions and institutions intended to curb crime.

The Original Meaning in Sociology

In the eyes of the early sociologists, the question of how society curbs individual passions and interests so as to perpetuate itself in peace was of paramount importance. Social control "referred to the capacity of a society to regulate itself according to desired principles and values" (Janowitz, 1975, p. 82). In 1901, E. Ross published a book called *Social Control. A Survey of the Foundation of Order*, which was influential in putting the concept at the center of US sociology. The question he undertook to answer was: "How is it that people come to live together and cooperate with that degree of harmony we see about us?" A kind of natural order emerges first from people living together with their capacities for sympathy, sociability, sense of justice, and resentment. Second, the society develops a certain measure of control over the individual. The means of control surveyed by Ross (1901) cover a whole range of social practices and institutions: public opinion, law, belief, social suggestion, education, religion, personal ideals, ceremony, art, social valuation, etc. According to Ross, social control includes all those aspects of social life inducing conformity among individuals and contributing naturally to the social order.

Ross shared the preoccupations of Emile Durkheim, who was also trying to discover the roots of social harmony. The French sociologist used the term 'social constraint' to refer to the influence of the 'collective conscience' on individuals. The pressures of society on the individual does not take only the form of external constraints, it also creates a sense of moral obligation to accept duties and to obey the rules (Durkheim, 1960 (1893), 1895, 1923).

Following Ross, sociologists of the Chicago School, like Park, Burgess, and Thomas, continued to place social control at the center of sociology. "All social problems turn out to be problem of social control" (Park and Burgess, 1921, p. 785).

While developing the concept of social control, the early sociologists had another agenda besides explaining social order. They wanted to demonstrate the capacity of sociology to explain social facts more completely than economics or plain common sense. They first wanted to show the limitations of Adam Smith's analysis of the natural harmony of interests fostered by the famous invisible hand of the market. Economic self-interest cannot fully account for the existence of the social order (Pitts, 1968; Janowitz, 1975). They also wanted to debunk the common sense assumption that order resides in the hands of the police, judges, prison guards, and the like. Cooperation, harmony, and organization do not rest predominantly on coercion or on utilitarian calculations but on subtle pressures, expectations, and informal sanctions in the family, in the peer group, the workplace, the local community, etc. However, if social control is neither market regulation nor state control or coercion, one is left with a notion referring essentially to informal social control. But if one sticks to the ambition of taking into account most of the social order, how is it possible to ignore the utilitarian and coercive ways of achieving it?

Another problem emerges: by its very nature, social control is a static concept. When it occupies a central place, sociology becomes the science of social order, unable to analyze social disorder or social change. During the 1920s and 1930s, "sociologists became so caught up with the source of cohesion that

they neglected to examine the source of change" (Rothman, 1983, p. 108).

Those problems cooled the enthusiasm of sociologists for the concept. Some restricted it to the control of deviance. In many contemporary textbooks and handbooks of sociology, one finds a chapter on deviance and social control (e.g., Scull, 1988). After the 1970s, the few sociologists attempting to write on general theory do not make more than a passing mention of the term. The search for the foundation of social order is no longer conceived in terms of the control of society on individuals. A major sociologist like Coleman (1990) explains the social order with the concepts of interdependence, transaction (including exchange, bribes, threats, promises, and resource investment), authority relations, rights, and norms.

The Skeptic Reaction

In the 1960s and 1970s, sociologists critical of the rhetoric and reality of the institutions dealing with deviants (mainly prisons and mental hospitals) gave a whole different tone to the term. Social control came to mean first those stigmatizing actions putting in motion an amplification of the deviance process, and second those bureaucracies pretending to reform and treat deviants but, in fact, indulging in manipulation and coercion.

Labeling

In labeling theory, social control refers mainly to those reactions of police officers, judges, psychiatrists, etc., which consist of arresting, condemning, classifying, and incarcerating. Somebody comes to be classified as criminal, mentally disturbed, dangerous, or deviant in one way or another, and is set apart in a prison, an asylum, or any other place of exclusion. Those reactions mark such individuals, they stigmatize them, and they attach a label to their personality. Doing so, they modify not only the social identity of the individuals but also their self-perception, their way of life, and their behavior. "The experience of being caught and publicly labeled as a deviant … has important consequences for one's further social participation and self-image" (Becker, 1963, p. 31). "Treating a person as though he were generally rather than specifically deviant produces a self-fulfilling prophecy. It sets in motion several mechanisms which conspire to shape the person in the image people have of him" (Becker, 1963, p. 34). Instead of producing order, the attempt at social control amplifies the deviant tendencies of the labeled person (Wilkins, 1964). The arrest, the sentence, the psychiatric diagnostic, the imprisonment, or the institutionalization in a mental hospital puts individuals subjected to one of those treatments at risk of internalizing the infamous label while putting them at the margin of society. They might feel obliged to become part of a deviant group and act as one of the gang. Accordingly, they might become more deviant after being labeled than before.

The labeling theory has been criticized mainly on two counts. The labeling process is not totally arbitrary or completely unrelated to the preexisting deviant behavior and characteristics of the labeled individual. More often than not, the court condemns accused people who really committed a crime, and generally, the more serious the crime, the more severe the sentence. The number of crimes committed in the past by an individual is also related to the severity of the punishment. If such is the case, the question to be asked is: Does the label create a new social identity or is it just a disclosure of a deviant identity already there?

Empirical tests of the labeling hypothesis provide an inconsistent picture, with some researchers finding a weak effect of social reaction on subsequent offending, others failing to find one. The overall impact is small. "The preponderance of research finds no or very weak evidence of labeling effects. The more carefully the research keeps other factors constant, the less likely it is to find evidence that labeling has significant effect on criminal or deviant behavior" (Akers, 1994, p. 134, see also Braithwaite, 1989).

The History and Sociology of State Dealing with Deviants

A related trend of research focuses not so much on the effects of social control but on the characteristics and evolution of state institutions dealing with offenders, mental patients, and other deviants. In this context, social control is defined as "the organized ways in which society responds to behavior and people it regards as deviant, problematic … or undesirable" (Cohen, 1985, p. 1). As with the labeling theory, the concept carries strong pejorative overtones. Gone are the benevolent informal controls to be replaced by coercion, the repressive state, and the manipulation used by the powerful against the lower classes. The negative overtones come also from a disenchantment with the rehabilitative ideal and from skepticism concerning the so-called reform movement with its self-serving ideologies (Rothman, 1983). The historians and sociologists working in this perspective embark upon a work of demystification, confronting the official discourses with the real functioning of the social control apparatus. Psychiatrists and correctional officials pretend to treat and rehabilitate; in fact, they cannot help but keep on putting mental patients in total institutions and incarcerating offenders. Reformers speak of decarceration, decriminalization, and diversion, but in fact the old prison system remains, more crimes are created by the law, and new deviants are caught in the social control net. "Overall, the system enlarges itself and becomes more intrusive, subjecting more and newer groups of deviants to the power of the state and increasing the intensity of control directed at former deviants" (Cohen, 1985, p. 38).

This discrepancy between discourse and reality, between what officials say they do and what they do in fact, is an enlightening exercise provided that it goes beyond plain conspiracy theory.

The main contribution of authors working in this perspective is in the history of changes in punishment in Europe and the United States beginning at the end of the eighteenth century. Foucault (1975) and Ignatieff (1978, 1983) documented the following trends: (1) a sharp decline in physical punishments inflicted in public in order to inspire terror in the spectators, (2) the emergence of the prison as the dominant punishment for serious crimes, and (3) the creation of the

penitentiary, an institution pretending to reform, deter, and isolate prisoners not only from the outside world but also from the criminogenic influence of their fellow prisoners.

In a related vein are sociological studies of the internal functioning of mental hospitals and prisons, like *Asylum* by Goffman (1961). The thesis of the book is that institutions like prisons, mental hospitals, concentration camps, and even monasteries share some basic characteristics. They are all total institutions in which guards, nurses, doctors, etc. are given the mission and the power to control the inmates and isolate them from the outside world. In those total institutions, the dominating group organizes everyday life in minute details for maximizing their control over inmates. They impose detailed regulations so as to make the inmate's behavior most predictable. Swift punishment and degrading measures are visited upon the rebellious.

Social Control in Criminology

Nowadays, it is mostly in criminology that the term social control is found. It is often used in the same skeptical spirit in which the term was used by sociologists criticizing the state's apparatus of control. Some criminologists, however, remain faithful to the orientation of the early sociologists preoccupied with the foundation of social order. One aspect of this social order, which, for obvious reasons, is the focus of criminologists, is a society having a tolerable level of crime. Accordingly, social control is conceived as the means to achieve it.

Control Theory

Instead of asking the usual question, 'Why do people commit crimes?' criminologists who developed the control theory turned the problem around and asked 'Why do people obey the rules of society?' "Deviance is taken for granted; conformity must be explained" (Hirschi, 1969, p. 10). Basically, this is the kind of questioning that was at the center of the works of Hobbes, Durkheim, and Ross.

Control theory was mostly used to explain juvenile delinquency or, more precisely, its absence. Such was the intention of Reiss (1951). He was able to show that delinquency is a result of "a failure of personal and social control." Of interest is his definition of social control: "the ability of social groups to make norms or rules effective."

It is Hirschi's (1969) formulation of control theory that is the most influential in criminology. His basic proposition is that "delinquent acts result when an individual's bond to society is weak or broken" (p. 16). For adolescents, the bond materializes in the family, at school, and with their peers. This bond is made of four elements: attachment, commitment, involvement, and belief. Attachment means the close emotional ties to parents, school teachers, and peers, and caring about their expectations. Commitment is the investment in conformity, in education, and in a future career that would be jeopardized by violating the law. Involvement is the time spent in legitimate activities, like studying, taking part in sports, or leisure activities in the family. Belief refers to the adhesion to the rules of society, to the conventional values, to the idea that law has a moral value and should be obeyed. Hirschi provided the measures for those elements of the bond and a persuasive demonstration that the stronger young people's ties to society (in fact, the family and the school) are, the fewer delinquent acts they report.

Later research found many confirmations of the theory, especially with the attachment and involvement constructs. It is beyond contest that juveniles attached to their family and school do not tend to indulge in delinquency. Also, strong involvement, measured by educational and occupational aspirations, is related to a low level of delinquency (Akers, 1994; pp. 116–20, Braithwaite, 1989; pp. 28–9).

Collective Efficacy and Social Control

In *Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy*, Sampson et al., (1997; see also Sampson, 2012) place social control at the center of a major research study that examines 343 neighborhoods in Chicago. Their main hypothesis is that the differential ability of neighborhoods to maintain social control, cohesion, and trust is a major source of neighborhood variation in violence, including homicide. The construct 'collective efficacy' includes two elements: the social cohesion among neighbors and their willingness to intervene on behalf of the common good. By social control, the authors refer to the capacity of a group to regulate its members according to their goals, a primary goal being the desire of residents to live in environments that are free of crime, especially violence. Informal social control is hypothesized to be closely related to social cohesion and trust. These concepts were measured in a survey distributed to neighborhood residents. To measure informal social control, residents were asked about the likelihood that their neighbors could be counted on to intervene in a variety of contexts including if children were spray-painting graffiti on a local building, if they were showing disrespect to an adult, if a fight broke out in front of their house, etc. To grasp social cohesion and trust, respondents were asked how strongly they agreed that 'people around here are willing to help their neighbours,' 'this is a close-knit neighborhood,' 'people in this neighborhood can be trusted,' etc. The results of a sophisticated analysis show that the combined measure of informal social control, cohesion, and trust is a robust predictor of lower rates of violence. Three other dimensions are also relevant: concentrated disadvantage, immigration concentration, and residential stability. These dimensions explain a high percentage of collective efficacy and that their association with violence is largely mediated by collective efficacy.

Cultural Anthropology

The concept of social control is widely used by anthropologists to describe the methods by which different cultures resolve conflicts and enforce norms. The *Oceania* volume of the *Encyclopedia of World Cultures*, published by Human Relation Area Files (Hays ed. 1991), provides information on the social organization and culture of approximately 100 different societies, prior to their being submitted to the influence of the

'white man.' For each culture, one finds an entry about social control defined as "The sources of conflict within the culture and informal and formal social control mechanisms" (p. XX). The following classification of social controls in archaic societies in *Oceania* relies on a compilation of what is written about social control in all cultures where sufficient information is provided.

1. "Gossip": talk, discussion, public opinion, approval and disapproval, shaming, ridicule, mocking, public scolding, ostracism (53 occurrences).
2. "Big-men": chiefs or senior members of a clan have the authority to arbitrate conflicts; some can order confiscation, restitution, fine, flogging, beating, banishment, execution (52 occurrences).
3. "Sorcery": threats and fear of witchcraft, accusations of using sorcery to bring sickness or death, ostracism, assassination, or execution of suspects of sorcery (52 occurrences).
4. "Self-help": physical retaliation by the offended party, revenges, fights, duels, blood feuds (42 occurrences).
5. "Moots": assemblies, ceremonial meetings, feasts aimed at conflict resolution: discussion, negotiation, apologies, compensation, and mediation (29 occurrences).
6. "Taboo": fear of divine retribution by spirits, ancestors, and ghosts (24 occurrences).
7. "Reciprocity": in case of nonreciprocity in an exchange, the aggrieved party withdraws cooperation or demands compensation (14 occurrences).

This list provides a framework of the diversity of sanctions, natural and supernatural, and the degree of severity that exists in different cultures. It also demonstrates the importance of informal social control and the role of third parties. Finally, it illustrates that what is social control from one point of view may be perceived as a crime from another, a point stressed by Black (1983).

Black: Controlling to Right a Wrong

In *The Social Structure of Right and Wrong*, Donald Black (1993; see also Black, 1976; 1984) defines social control as "any process by which people define and respond to deviant behaviour." This definition includes law, arrest by police, lawsuit, punishment, compensation, and intervention in conflict by third parties. Black insists that social control should be conceived as a dependent variable to be explained and predicted by the characteristics of the setting in which it exists. The social stratification and the other dimensions of the social organization are important to explain and predict many manifestations of social control. Paradoxically, crime itself can be and in fact is social control, as exemplified by vengeance. Black distinguishes four styles of social control: compensatory, penal, therapeutic, and conciliatory. Conflict management is a major element in the social control of violence and takes on five forms: self-help (including vengeance, discipline, and rebellion), avoidance, negotiation, settlement, and toleration.

Often, third parties intervene in conflicts. Sometimes, they play support roles, including advisor, advocate, and ally. More importantly, these third parties may act as peacemakers and mediators, who refuse to take sides and seek a solution, typically through compromise. The arbitrators provide opinions, often designating who is right and wrong. The judge has the power to impose and enforce his decision. The importance of third parties for understanding violence, including homicide, lies in the fact that, in some social contexts, third parties take sides and encourage the fighters, which may contribute to the escalation of violence. In contrast, in other contexts, third parties strive to end the conflict and work to restore peace. In principle, the homicide rates should be low in societies where third parties have a tendency to act as peacemakers.

The Social Control of Crime and Rational Choice Theory

The scientific evaluation of the effectiveness of social controls is becoming an important field of research. The revival of deterrence research is a case in point. In its classic formulation, the deterrence hypothesis predicts that the frequency of crime will vary inversely with the certainty, celerity, and severity of punishment handed by the state. Recent researchers cannot find much deterrent effect of severity on crime, but crime rates tend to be low when the certainty of punishment (often measured by the police clearance rate) is high. It is also the case that there is much more evidence of a general effect of certainty on crime rate (general deterrence) than on the level of recidivism of punished offenders (specific deterrence) (Andenaes, 1974; Gibbs, 1975; Blumstein et al., 1978; Cook, 1980; Nagin, 1998; Apel and Nagin, 2011).

However, there are other ways to limit crime than deterrence and informal controls. The state uses nonpunitive ways to keep crime at bay, like therapies to rehabilitate or reform offenders and crime prevention measures such as street lighting, surveillance of public areas, gun control, and regulations on car antitheft systems. In civil society, crime control activities are very common. Everybody devotes time, money, and effort to protect their goods against theft and to safeguard their person against aggression. People put their money in banks, lock the doors of their houses, avoid dangerous places; in some cases, they even buy guns for self-protection. Businesses and other organizations spend considerable amounts of money to protect their assets from loss or crime: they hire private guards, control access to their premises, use safes, install closed-circuit television, etc. All those actions and precautions – public and private, formal and informal, repressive and preventive – are clearly aimed at reducing the probability of a crime occurring. As such, their common goal is crime control. This leads us to a last definition: the social control of crime refers to all the means specifically aimed at reducing the probability or severity of crime.

The reader will note that social control is now defined by its intention or aim, not by its results. By doing this, one follows Gibbs (1989, pp. 23–24) who criticizes the sociological concept for belittling the intentional quality of social control. In its common use, the term 'control' conveys intention: one tries deliberately to control, to direct, and to influence another. Actions having the unintended effect of preventing crime no doubt exist. For example, Felson (1998) explains the sharp decline in crime rates starting in 1994 in the United States by the coming of a cashless society. People use more credit cards

and the like, so they have less cash in their pockets and their houses. Having less cash to steal, the offenders become less active. In this instance, one should not speak of social control but of an unintended preventive effect of an economic evolution.

Results (more or less crimes) are important matters but should not be included in the definition of social control. Attempts at social control, including failures, are social control. The latter's impact is not a matter of definition but of evaluation. Because of their exclusive focus on scientific evaluation, Sherman et al. (1998, p. 2) chose another path. They define crime prevention not by its intention but by its conse-quences. It is "any practice shown to result in less crime than would occur without the practice."

If social control is made of intentional actions and choices, is it possible to conceive its impact on offenders in terms of actions and choices? The potential offender – that is, the individual having the intention to commit an offence – might choose to do it despite social controls or not to do it because of them. Such an individual is a decision maker acting under the constraints of social control.

The impact one tries to have on offenders when one attempts to control crime is essentially to (1) increase the effort of committing crime (e.g., by target hardening or gun control), (2) increase the risks (by surveillance, punishment, burglar alarms, and the like), (3) reduce the anticipated rewards of crime (by target removal, identifying property, etc.), and (4) remove excuses used by offenders to minimize the moral opprobrium cast on crime (e.g., by rule setting and public condemnation of crime) (Clarke, 1997). If potential offenders live in a well-ordered society where those impacts are achieved, they will find themselves in a radically different choice situation than in a disorganized society where social controls are erratic. Most of the time they will find committing crime difficult, risky, unrewarding, and reprehensible. If they are minimally rational (Cornish and Clarke, 1986), they will tend to look for noncriminal alternatives. This means that where and when the social controls operate reasonably well, they shape the alternatives of choice for social actors. They close most of the criminal options for all. They attach quite negative utilities (in the economic sense) to the criminal options. To commit crime in such a situation, one needs a fair amount of greed, temerity, disregard for long-term consequences, or plain foolishness.

Logically, the more serious a type of crime is, the more it should be worthwhile controlling. In fact, one finds that guardianship is lax when minor valuables are to be protected and increased when major valuables or life must be protected. Police detectives work harder investigating murders than burglaries. The positive relation between the gravity of crime and the severity of punishment, as well as between the gravity of crime and the certainty of punishment are basic facts of research on penal decision making (Gottfredson and Gottfredson, 1980). This heightened pressure of social controls on the most serious crimes will give incentive to offenders to choose the least criminal option, the least severe offense in the eventuality that they persist in crime. That should explain the inverse relation one finds between the frequency of a type of crime and its severity (there are fewer murders than robberies, and fewer robberies than burglaries). Those pressures on criminal choices can be called the structuring effects of social control (Cusson, 1993).

However, the lesson learned by sociologists and historians, showing that social controls often operate in an erratic manner, should not be forgotten. Formal and informal controls are not in place where they should be for a number of reasons: groups are too disorganized, resources are lacking, and acts do not follow the rhetoric. This is to mean that the quality and intensity of social control have every reason to be highly variable in space and time. In turn, this uneven quality and intensity of social control should be related to the uneven distribution of crime rates in space and time.

See also: Crime, Sociology of; Criminal Justice, Sociology of; Delinquency, Sociology of; Justice and Law; Law and Everyday Life; Law: Economics of Its Public Enforcement; Norms; Punishment: Comparative Politics of; Punishment: Social and Legal Aspects; Sanctions in Political Science; Stigma, Social Psychology of; Totalitarianism.

Bibliography

Akers, R.L., 1994. Criminological Theories: Introduction and Evaluation. Roxbury, Los Angeles, CA.
Andenaes, J., 1974. Punishment and Deterrence. University of Michigan Press, Ann Arbor, MI.
Apel, R., Nagin, D.S., 2011. General deterrence: a review of recent evidence. In: Wilson, J.Q., Petersilia (Eds.), Crime and Public Policy. Oxford University Press, Oxford.
Becker, H., 1963. Outsiders. The Free Press, New York.
Black, D., 1976. The Behavior of Law. Academic Press, New York.
Black, D., 1983. Crime as social control. American Sociological Review 48, 35–45.
Black, D., 1993. The Social Structure of Right and Wrong. Academic Press, San Diego.
Blumstein, A., Cohen, J., Nagin, D. (Eds.), 1978. Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates. National Academy of Science, Washington, DC.
Braithwaite, J., 1989. Crime, Shame and Reintegration. Cambridge University Press, Cambridge, UK.
Clarke, R.V., 1997. Introduction. In: Clarke, R.V. (Ed.), Situational Crime Prevention. Successful Case Studies. Guilderland, New York, pp. 1–44.
Cohen, S., 1985. Visions of Social Control. Polity Press, Cambridge, UK.
Coleman, J.S., 1990. Foundations of Social Theory. Belknap Press of Harvard University Press, Cambridge, MA.
Cook, P., 1980. Research in criminal deterrence: laying the groundwork for the second decade. In: Morris, N., Tonry, M. (Eds.), Crime and Justice, vol. 2. University of Chicago Press, Chicago, IL, pp. 211–268.
Cornish, D.B., Clarke, R.V. (Eds.), 1986. The Reasoning Criminal. Springer-Verlag, New York.
Cusson, M., 1993. The structuring effects of social control. International Annals of Criminology 31 (1–2), 45–57.
Durkheim, E., 1895. Les règles de la méthode sociologique. Presses Universitaires de France, Paris.
Durkheim, E., 1923. L'éducation Morale. Presses Universitaires de France, Paris.
Durkheim, E., 1960. De la division du travail social, seventh ed. Presses Universitaires de France, Paris (1893).
Felson, M., 1998. Crime and Everyday Life, second ed. Pine Forge Press, Thousand Oaks, CA.
Foucault, M., 1975. Surveiller et punir. Gallimard, Paris.
Gibbs, J.P., 1975. Crime, Punishment, and Deterrence. Elsevier, New York.
Gibbs, J.P., 1989. Control: Sociology's Central Notion. University of Illinois Press, Urbana, IL.
Goffman, E., 1961. Asylum. Doubleday/Anchor, Garden City, NY.
Gottfredson, M., Gottfredson, D.M., 1980. Decision-making in Criminal Justice: Toward the Rational Exercise of Discretion. Ballinger, Cambridge, MA.
Hays, T.E. (Ed.), 1991, Encyclopedia of Word Cultures, vol. II. Hall and Co. and Human Relation Area Files, Oceania. Boston.
Hirschi, T., 1969. Causes of Delinquency. University of California Press, Berkeley, CA.

Ignatieff, M., 1978. A Just Measure of Pain: The Penitentiary in the Industrial Revolution, 1750–1850. Pantheon, New York.
Ignatieff, M., 1983. State civil society and total institutions: a critique of recent social histories of punishment. In: Cohen, S., Scull, A. (Eds.), Social Control and the State. Basil Blackwell and Martin Robertson, Oxford, pp. 75–105.
Janowitz, M., 1975. Sociological theory and social control. American Journal of Sociology 81 (1), 82–108.
Nagin, D.S., 1998. Criminal deterrence research at the outset of the twenty-first century. In: Tonry, M. (Ed.), Crime and Justice. A Review of Research. University of Chicago Press, Chicago, pp. 1–42.
Park, R.E., Burgess, E.W., 1921. Introduction to the Science of Sociology. University of Chicago Press, Chicago, IL.
Pitts, J.R., 1968. Social control: the concept. In: International Encyclopedia of the Social Science. Macmillan and The Free Press, New York.
Reiss Jr., A., 1951. Delinquency as the failure of personal and social control. American Sociological Review 16 (2), 196–206.
Ross, E.A., 1901. Social Control: A Survey of the Foundation of Order. Macmillan, New York.
Rothman, D., 1983. Social control: the uses and abuses of the concept in the history of incarceration. In: Cohen, S., Scull, A. (Eds.), Social Control and the State. Basil Blackwell and Martin Robertson, Oxford, UK, pp. 106–117.
Sampson, Robert, 2012. Great American City. University of Chicago Press, Chicago, IL.
Sampson, R.J., Raudenbush, S.W., Earls, F., 1997. Neighborhoods and violent crime: a multilevel study of collective efficacy. Science 277 (5328), 918–924.
Scull, A., 1988. Deviance and social control. In: Smelser, N.J. (Ed.), Handbook of Sociology. Sage, Newbury Park, CA, pp. 667–695.
Sherman, L.W., Gottfredson, D.C., Mackenzie, D.L., Eck, J., Reuter, P., Bushway, S.D., 1998. Preventing Crime: What Works, What Doesn't, What's Promising. National Institute of Justice. Research in Brief. US Department of Justice. Office of Justice Programs, Washington, DC.
Wilkins, L., 1964. Social Deviance. Prentice-Hall, Englewood Cliffs, NJ.

Control, Syntactic

Klaus-Uwe Panther, Nanjing Normal University, Nanjing, China; and University of Hamburg, Hamburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by K.-U. Panther, volume 23, pp. 15397–15401, © 2001, Elsevier Ltd.

Abstract

The term 'control' denotes a relation of coreference between one (or occasionally several) nominal constituents in a main clause and the covert subject of a nonfinite subordinate clause, where the former (controller) determines the reference of the latter (controllee). This article traces the history of control theory from English-based syntax-oriented approaches in generative grammar to accounts that acknowledge cross-linguistic differences and the relevance of semantic and discourse-pragmatic factors in control relations.

Introduction

The term 'control' is attributed to the generative linguist Paul Postal (1970). It metaphorically expresses the observation that the reference of an understood grammatical subject in a subordinate clause, the controllee, may depend on the reference of another constituent, the controller, in the main clause. An example illustrating the control relation is given below in (1a), with its relevant syntactic structure represented in (1b). In (1b) constituents with the same reference are marked by identical subscripts; the bracketed part of the sentence contains a nonfinite subordinate clause with an implicit subject symbolized as 'Ø' (for 'empty subject'). These conventions are followed throughout this article.

1. a. Mary convinced Bill to do her homework.
   b. Mary_i convinced Bill_j [$Ø_j$ to do her homework].

In (1), the noun phrase *Bill* controls the reference of the empty subject of the infinitival complement clause. Every native speaker of English will understand (1) as meaning that Bill, and not Mary herself, will do her homework. The challenging linguistic problem is to find a principled account for why in (1) the main clause object is the controller, why the main clause subject cannot have this function, and why the empty subject cannot just refer to a third person not mentioned in the sentence at all.

Sentences such as (1) exhibit the notion of obligatory control, that is, a case in which the empty subject must be coreferential with a main clause constituent. Obligatory control contrasts with free (also called arbitrary or nonobligatory) control as in:

2. [Ø to wait in the crowded airport] would be a mistake.

In the subject clause in (2), the reference of the understood subject is not determined by any other noun phrase in the sentence but depends on the discourse context.

A comprehensive state-of-the art report on control, especially in the framework of generative grammar, has been compiled by Landau (2013).

Some Control Phenomena

In this article, the term 'control' will be understood as meaning 'obligatory control.' The following nonexhaustive list gives some sentences that exemplify obligatory control.

Infinitival Purpose Clauses:

3. John_i bought a book_j [$Ø_i$ to amuse himself with] (adapted from Nishigauchi, 1984, p. 229).
4. Mary_i bought the little boy_j a toy [$Ø_j$ to amuse himself with] (adapted from Nishigauchi, 1984, p. 229).

Infinitival Complement Clauses:

5. John_i begged Mary_j [$Ø_j$ to leave].
6. Mary_i begged [$Ø_i$ to leave].

Gerundial Complement Clauses:

7. Paul_i thanked Mary_j for [$Ø_j$ writing the report].
8. Paul_i thanked Mary_j for [$Ø_i$ having been admitted to the tennis club].

Most work on control has been done on nonfinite complement clauses, that is, on clauses exemplified by sentences (5) through (8). The remainder of the article, therefore, is concerned mostly with control in nonfinite complement clauses.

Types of Control

Control relations can be classified in terms of the kind of coreference that links controller and controllee. Control can be 'exhaustive,' 'partial,' 'variable,' and 'split.' In an exhaustive control relation, controller and controllee 'completely overlap' (Stiebels, 2007, p. 5) or are 'strictly identical' in reference (Landau, 2010, p. 364). The following examples from the Corpus of Contemporary American English (COCA: Davies, 2008–) illustrate this property:

9. Obama_i has managed [$Ø_i$ to use effective diplomacy to enhance American influence] (COCA, 2012: Davies, 2008–).
10. […] President Obama_i promised [$Ø_i$ to cut the deficit in half […]] (COCA, 2012: Davies, 2008–).

In (9) and (10) the controllers *Obama* and *President Obama*, respectively, are strictly identical to the denotatum of the covert subject in the complement clause.

A controller may also determine only a subset of the possible referents of the controllee. This phenomenon, exemplified by (11), is called 'partial control' (see Hornstein, 2003; Landau, 2000; Stiebels, 2007 for more examples).

11. She$_i$ preferred [Ø$_{i+}$ to meet in person […]] (COCA, 2012: Davies, 2008–).

The interpretation of (11) depends crucially on the meaning of *meet*, which requires at least two participants. The subscript 'i+' in (11) – a notation borrowed from Hornstein (2003) – symbolizes reference to at least one additional participant. This plurality interpretation follows from the semantics of *meet*. The additional referent (or referents) has no overt antecedent in the main clause of (11). In contrast to mental state verbs like *prefer* and *want*, implicative verbs like *manage* and *force* do not allow partial control. For example, the following contrasts can be observed (see also Landau, 2000; Hornstein, 2003; Stiebels, 2007):

12. After about 3 h, all our guests$_i$ managed [Ø$_i$ to gather in the apartment, tired and starving] (COCA, 2010: Davies, 2008–).
13. *The chair$_i$ managed [Ø$_i$ to meet/gather at 6] (adapted from Hornstein, 2003, p. 37).
14. The chair$_i$ preferred [Ø$_{i+}$ to meet/gather at 6] (adapted from Hornstein, 2003, p. 37).

Sentence (12) is well formed since the empty subject in the infinitival clause is exhaustively controlled by the matrix subject. The set of guests is strictly identical to the set of people that gather in the apartment. In contrast, in (13) the singularity of *the chair* clashes with the plurality of participants entailed by the meanings of *meet* and *gather*, respectively. Finally, sentence (14) is fully acceptable because the verb *prefer* allows partial control and is, therefore, compatible with the singular controller *the chair* despite the plurality of referents evoked in the complement clause.

Variable control can be illustrated with the following German example (adopted from Stiebels, 2007, p. 3):

15. Maria$_i$ schlug Peter$_j$ vor [Ø$_{i/j/i+j}$ einen Tisch im Restaurant zu reservieren]. "Maria suggested to Peter that she/he/she and he reserve a table in the restaurant."

The German verb *vorschlagen* 'suggest, propose' allows control by the matrix subject, the matrix object, and even split control (see next paragraph), that is, the interpretation that the referents of the subject and the object in (15) reserve a table. In addition to the variable control reading, a partial control reading is also possible: a third person not explicitly identified in the main clause could also undertake the action expressed in the complement clause.

Split control typically occurs in the case of agreements about joint or collaborative efforts (cf Stiebels, 2007, p. 5).

16. This summer, Fiat$_i$ agreed with Mitsubishi$_j$ Ø$_{i+j}$ to develop a sport utility vehicle for sale in Europe (COCA, 1999: Davies, 2008–).

Sentence (16) conveys that only *Fiat* and *Mitsubishi* cooperate in developing a new sport utility vehicle. From a syntactic perspective one can call (16) an example of split control since there is not a unique controller, but a division of labor between two controllers (subject and prepositional complement). From a semantic–pragmatic point of view, one could also speak of 'joint control' because both controllers are collaboratively involved in the task specified in the complement clause.

A Short History of Control

In the 1960s and 1970s, the early days of generative grammar, control was generally believed to be a purely a syntactic phenomenon. More recently, it has become increasingly evident that semantic and pragmatic factors play an equally, if not more important role in control.

Syntactic Approaches

The syntactic approach to control assumes that control relations can be determined on a purely structural basis. The most influential proposal along these lines was made by Rosenbaum (1967, 1970). He postulates a Minimal Distance Principle (MDP), which predicts that the reference of the (empty) subject of a nonfinite complement clause depends on the closest available controller in the main clause. In sentence (1) above, the main clause noun phrase closest to the empty subject Ø is *Bill*. Indeed, this noun phrase controls the reference of the empty subject, and the MDP thus makes the right prediction. Hornstein (1999) assumes a Minimal Link Condition to account for the same facts. The Minimal Link Condition, as formulated within the minimalist framework of generative grammar, stipulates a preference for minimal derivations.

There are, however, many sentences for which a minimality condition such as the MDP is falsified. Difficulties emerge with constructions that display more than one potential controller in the main clause, as in (17) and (18):

17. Mary$_i$ persuaded Bill$_j$ [Ø$_j$ to clear her$_i$ desk].
18. Mary$_i$ promised Bill$_j$ [Ø$_i$ to clear her$_i$ desk].

For sentence (17) the MDP correctly selects the closest noun phrase in the main clause, *Bill*, as controller. *Persuade* belongs to the class of control verbs that normally select the main clause object as controller. However, for sentence (18) the MDP wrongly predicts *Bill* as controller; in fact, the main clause subject *Mary* is the controller. Subject control verbs of the *promise* class constitute a problem for the MDP and have to be marked as exceptional in the lexicon. In this connection, it is however worth noting that in ordinary language there is a strong tendency to avoid the use of *promise* with a matrix object. Thus, sentence (19) is compatible with the MDP – at least in the sense that the closest explicit argument is selected as controller.

19. Mary$_i$ promised [Ø$_i$ to clear her desk].

Although instances of *promise* with a matrix object are relatively infrequent in English-language corpora, in other languages, for example, German, the corresponding verb *versprechen* often occurs with an overt matrix object and falsifies the Rosenbaum's MDP because *versprechen* like *promise* exhibits subject control in the default case. The main objection to the MDP is, therefore, that it is too language specific, that is, tailored to account for control phenomena in English.

Another fact that challenges the validity of the MDP is that one and the same verb may admit either subject or object control:

20. Bill$_i$ thanked the coach$_j$ for [Ø$_j$ selecting him$_i$].
21. Bill$_i$ thanked the coach$_j$ for [Ø$_i$ having been admitted to the tennis club].

In (20), which can be regarded as the default case, the controller is *the coach* (object control), whereas in (21) a control switch or shift takes place: the most likely controller is the matrix subject *Bill*. The MDP wrongly predicts that in both sentences *the coach* should be the controller.

Other syntactic control theories have been proposed by Bresnan (1982) and by Bach (1979) in the frameworks of Lexical–Functional Grammar and Montague Grammar, respectively. What these theories have in common is that they regard syntactic structure as the decisive factor for determining the control relationship. However, there is a substantial body of evidence that strongly suggests the importance of semantic and pragmatic information in the determination of control relations.

Since the 1980s, two theoretical problems have dominated the discussion about obligatory control into complement clauses. The first of these concerns the syntactic status of the complement clause, more specifically the question whether the empty placeholder PRO (a 'silent' pronoun, symbolized as 'Ø' in this article) is a syntactic constituent, that is, a genuine subject. Generative grammarians have generally been divided between a propositional and a predicational analysis of control. The propositional approach assumes that the complement clause expresses a complete (closed) proposition; hence PRO is believed to be a syntactic – albeit silent – constituent. This position is especially advocated by Chomsky and his followers (see e.g., Chomsky, 1980). In contrast, the predicational theory of control (see e.g., Williams, 1980) holds that nonfinite complement clauses are predicates, that is, open propositions or unsaturated predicates that must be predicated of a nominal referent in the matrix clause (see e.g., Landau, 2004 for a brief description of this controversy and further bibliographical references).

The second theoretical issue concerns the relationship between raising (a constituent out of a subordinate clause into the main clause) and control (see e.g., Hornstein, 2003; Davies and Dubinsky, 2006 for discussion). In classical generative grammar, raising and control have been regarded as distinct syntactic mechanisms:

22. John$_i$ seems/appears/is likely [t$_i$ to like Mary] (adapted from Hornstein, 2003, p. 7).
23. John$_i$ tried/hoped/expected [PRO$_I$ to like Mary] (adapted from Hornstein, 2003, p. 7).

Sentence (22) is traditionally regarded as involving movement of the subject of the embedded clause, *John*, into the subject position of the matrix clause, leaving a trace of the moved constituent (t) in the embedded clause. In contrast, in (23) the covert subject of the embedded clause (PRO) remains *in situ* and is coindexed with the subject of the matrix clause, *John*; in other words, the matrix clause subject controls the reference of the understood subject in the infinitival complement clause (see Hornstein, 2003; Hornstein and Polinsky, 2010 for an in-depth discussion of this issue within the framework of generative grammar).

Semantic Approaches

A comparison of sentences (17) and (18), which look alike structurally, invites the conclusion that their differing control properties are caused by the meanings of the verbs *persuade* and *promise*, respectively. To persuade somebody to do something means that the addressee of an act of persuasion is led to believe that he or she should do something. This prospective action is expressed in the infinitival complement clause. Since in an active sentence such as (9) the addressee is the grammatical object of the main clause, it is not surprising that this constituent functions as controller. In contrast, a promise involves a commitment of the speaker to a future course of action. In an active sentence such as (18), the 'promiser' is the grammatical subject and one would, therefore, expect the subject to control the reference of Ø.

An account that acknowledges the importance of the meaning of control verbs is Sag and Pollard's (1991) control theory formulated within the framework of Head-Driven Phrase Structure Grammar. These authors use semantic criteria to distinguish three classes of control verbs: (1) the *order/permit* class, which typically induces object control; (2) the *promise* class, which typically induces subject control; and (3) the *want/expect* class, which again induces subject control. The *order/permit* and the *promise* classes conceptually imply actions, whereas verbs of the *want/expect* class express what Sag and Pollard call an 'orientation' in their complement clause.

A touchstone for any control theory is the phenomenon of control shift (from subject to object control, and vice versa). Control shift tends to occur in specific syntactic contexts, for example, passivized and modalized complement clauses (especially with *be allowed to*), or complement predicates that are recipient oriented (see Panther and Köpcke, 1993; Panther, 1994; Köpcke and Panther, 2002; Stiebels, 2007). Sag and Pollard (1991), and subsequently Jackendoff and Culicover (2003) and Culicover and Jackendoff (2005), propose that control switch can be explained in terms of a conceptual enrichment of the literal meaning of the complement clause. Consider an example with a passivized complement clause such as (24):

24. Mary$_i$ asked Peter$_j$ [Ø$_i$ to be invited to the party] (Stiebels, 2007, p. 4).

The requestive verb '*ask*' usually induces object control, as in (25):

25. Mary$_i$ asked Peter$_j$ [Ø$_j$ to leave her alone].

However, in (24) the most plausible interpretation is that the matrix subject *Mary* corefers with the understood subject of the complement clause. The meaning of (24) can be paraphrased as (26):

26. Mary asked Peter to bring about the event 'Mary is invited to the party.'

The predicate of the complement clause in (24) has thus been coerced into a causative meaning. The causer that brings about the state of affairs 'Mary is invited to the party' is *Peter*. Stiebels (2007, p. 4) observes that the syntactic controller in (24) is the matrix subject *Mary*, but the semantic controller remains, as required by the semantics of *ask*, the referent of the matrix object.

Pragmatic Approaches

Since the mid-1980s, the insight has grown that, in addition to the meaning of control verbs, the context in which a sentence is

uttered and world knowledge about the participants' status play a significant role in assigning the coreferential links between the controller and the controllee (e.g., Comrie, 1984, 1985; Růžicka, 1999; Duffley, 2014).

These additional pragmatic factors are illustrated by the following German sentences (adapted from Wegener, 1989):

27. Der Anwalt$_i$ versprach dem Häftling$_j$ [Ø$_j$ bald aus dem Gefängnis entlassen zu werden]."The lawyer promised the prisoner that he (=the prisoner) would be released from jail soon."
28. Der Häftling$_i$ versprach seiner Frauj [Ø$_i$ bald aus dem Gefängnis entlassen zu werden]."The prisoner promised his wife that he (=the prisoner) would be released from jail soon."

The most plausible interpretation for both (27) and (28) is that the prisoner will be released from jail. These interpretations follow from the speech act verb *versprechen* 'promise,' the status of the referents of the noun phrases *der Anwalt*, *dem Häftling*, and *seiner Frau*, and the meaning of the passive complement clause. A promise implies a promiser who is committed to a future action and a promisee who benefits from the action of the promiser. This semantic–pragmatic scenario can be applied to the interpretation of sentences (27) and (28) and the most plausible controller can be inferred (cf Panther and Köpcke, 1993; Panther, 1994; Köpcke and Panther, 2002).

In (27) the lawyer is the promiser; hence the lawyer is committed to some future action. However, the complement clause does not express an action explicitly. A step toward a reasonable interpretation is to assume that the complement clause expresses the result of an action performed by the promiser. The action itself has to be inferred. Furthermore, the result of the action is beneficial to the prisoner. Therefore, the most plausible controller is the noun phrase *dem Häftling*. On this account, the assignment of the controller is not a matter of finding the noun phrase that is syntactically closest to Ø in the matrix clause, but a matter of common-sense knowledge about the roles of lawyers and prisoners.

For (28), a coherent interpretation results if it is assumed that the prisoner behaves in such a way that he causes his release from jail. Again, the complement clause is interpreted as the result of an action – an action carried out by the prisoner to the effect that he will be released from prison. His wife can be regarded as the beneficiary of this action. The interpretations sketched above are not the only ones possible. On the assumption that the lawyer is in jail (27) could be interpreted as conveying that the lawyer will be released from jail; and (28) could have the interpretation, 'The prisoner promised his wife that she would be released from jail soon' given the background knowledge that the prisoner's wife is in prison. It is in the nature of the pragmatic approach to control that different background assumptions have an impact on what is the most plausible controller of Ø.

Similar observations have been made for Chinese by Huang (1994, p. 153), who contrasts the following sentences:

29. Xuesheng$_i$ daying laoshi$_j$ [Ø$_i$ mingtian jiao zuowen].
'The pupil promises the teacher that he (=the pupil) will hand in the essay tomorrow.'
30. Laoshi$_i$ daying xuesheng$_j$ [Ø$_j$ mingtian fang yi tian jia].
'The teacher promises the pupil that he (=the pupil) will have a day off tomorrow.'

The control verb *daying* 'promise' can be used either in the 'unmarked,' that is, normal sense 'commitment-to-action,' as in (29), or in the 'marked' sense 'commitment-to-permission,' as in (30). These senses are derived pragmatically; they constitute 'preferred' interpretations that can be inferred on the basis of the meaning of the speech act verb *daying* and world knowledge about teacher–pupil interactions in the institutional framework of a school.

To summarize, the interpretation of control sentences relies not only on syntactic and semantic information but also, to a substantial degree, on world knowledge and contextual clues about the utterance situation.

Language-Specific Coding Differences of Control Relations

One might surmise that the mechanisms of control are very similar across languages, but on closer inspection it turns out that there are language-specific differences in the way that control structures are coded. A selective list of these coding differences is given below (for a detailed account, see Růžicka, 1999).

Modal versus Nonmodal Coding

For many speakers of English the following sentence is ambiguous:

31. Johnny$_i$ asked his mother$_j$ [Ø$_{i/j}$ to go to the movies].

Sentence (31) can mean either that Johnny wants his mother to go to the movies or that Johnny asks for permission to go to the movies himself. This latter interpretation seems to be the most likely one, assuming that Johnny's mother has authority over Johnny (see the section 'Pragmatic Approaches').

As Růžicka (1999, p. 45) points out, the 'ask-for-permission' (subject control) reading is also available alongside the 'ask-for-action' interpretation (object control) in Spanish, although the concept of permission is not coded explicitly.

32. María$_i$ le$_j$ pidió a Juan$_j$ [Ø$_i$ hablar con los muchachos].
'María asked Juan that she be allowed to talk to the boys.'

The 'ask-for-permission' interpretation is not possible, however, if the same sentence is translated literally into German:

33. Hänschen$_i$ bat seine Mutter$_j$ [Ø$_j$ ins Kino zu gehen].
'Johnny asked his mother to go to the movies.'

Sentence (33) can only mean that Hänschen's mother is supposed to go to the movies. In order to express the 'ask-for-permission' sense, with Hänschen as the controller, German speakers have to resort to an utterance that codes permission by means of the modal auxiliary *dürfen*:

34. Hänschen$_i$ bat seine Mutter$_j$ [Ø$_i$ ins Kino gehen zu dürfen].
'Johnny asked his mother to be allowed to go to the movies.'

Causative versus Non-Causative Coding

Another important coding difference between English and languages like German, Russian, and Czech is exemplified by

35. Peter$_i$ persuaded his father$_j$ [Ø$_j$ to be examined by a doctor].

Sentence (35) means that Peter's father should undergo a medical examination. This sense is, however, not explicitly coded in sentence (35), but is nevertheless conventionally conveyed by it. What is coded in the subordinate clause is the result of the action, not the action itself. If sentence (35) is translated literally into German, for many speakers, unacceptability results and, for those who accept the sentence, an interpretation with Peter as controller is more plausible:

36. Peter$_i$ überredete seinen Vater$_j$ [Ø$_i$ von einem Arzt untersucht zu werden].
'Peter persuaded his father that he (=Peter) should be examined by a doctor.'

The only way of rendering the intended message of (35) in German and Russian is to code *seinen Vater* and *otca*, respectively, as an active participant who engages in an action that results in a medical examination by a doctor:

37. Peter$_i$ überredete seinen Vater$_j$ [Ø$_j$ sich von einem Arzt untersuchen zu lassen].
'Peter persuaded his father to have himself (=his father) examined by a doctor.'
38. Peter$_i$ ugovoril otca$_j$ [Ø$_j$ podvergnut' sebja osmotru vrač] (Růžicka, 1999, p. 58).
'Peter persuaded father that he (=father) should undergo a medical examination.'

Both the German and Russian languages have to code the agent role of 'father' explicitly. This is done by the causative verb *lassen* in German and the action verb *podvergnut' sebja* 'submit oneself' in Russian.

Implicit Control versus Explicit Control

Another important difference between English and other languages such as German and Russian concerns the ability for a conceptually present controller to remain unexpressed. In this section the focus is on contrasts between German and English. In German it is possible to say:

39. Der Professor$_i$ bat Ø$_j$ [Ø$_j$ den Zeitschriftenartikel zu kopieren].
'The professor requested (of some person) that this person photocopy the journal article.'

In (39) the controller of the understood subject remains unexpressed. The person or group of persons supposed to comply with the request is indeterminate. This phenomenon may be called implicit control (cf Panther, 1997; Landau, 2010 for more detailed discussion of explicit and implicit control).

However, a word-for-word translation of (39) into English would not have the same effect:

40. The professor$_i$ asked Ø$_j$ [Ø$_i$ to photocopy the journal article].

The only possible interpretation of (40) is that the professor himself wants to photocopy the journal article and asks for permission to do so. In German, however, the translation of (40) requires modal coding with *dürfen* 'be allowed to' (see section 'Modal versus Nonmodal Coding'), as in (41):

41. Der Professor$_i$ bat Ø$_j$ [Ø$_i$ den Zeitschriftenartikel kopieren zu dürfen].
'The professor asked to be allowed to photocopy the journal article.'

The first point to be made is that in English examples like (40) either the control verb *ask* is 'coerced' into reading as 'ask for permission' or, alternatively, that the complement clause predicate *photocopy the journal article* is shifted to the meaning 'be allowed to photocopy the journal article.' Second, since the direct object in the matrix clause is covert, the controller function in (40) is automatically switched to the overtly expressed matrix subject. In contrast, in German the covert matrix object is the obligatory controller in (39). Third, control shift to subject control is possible in German with directive speech act verbs like *bitten* 'ask, request' but requires additional coding in the complement clause (here: *dürfen* 'may, be allowed to').

Another example of implicit control in German is (42):

42. Der Arzt$_i$ riet Ø$_j$ [Ø$_j$ den Fettanteil in meinem Essen zu verringern].
'The doctor advised (someone) that this person should reduce the amount of fat in my diet.'

The literal translation of (42) into English leads to ungrammaticality (marked by an asterisk):

43. *The doctor$_i$ advised Ø$_j$ [Ø$_j$ to reduce the amount of fat in my diet].

Sometimes the ban on implicit object control in English can be circumvented by resorting to a gerund construction. A possible rendering of (42) in English is (44):

44. The doctor$_i$ advised Ø$_j$ [Ø$_j$ reducing the amount of fat in my diet].

There exist a few verbs in English that allow implicit object control. Among them are *signal* and *motion*, exemplified in (42):

45. a. The clerk$_i$ signaled Ø$_j$ [Ø$_j$ to approach the teller].
b. Bill$_i$ motioned Ø$_j$ [Ø$_j$ to come in].

In general, German permits implicit complement control more freely than English. There are however control verbs in both English and German whose matrix object has to be coded explicitly and which, hence, do not license implicit control. These verbs can be characterized as 'manipulative' and comprise some directive illocutionary verbs and implicative verbs. Examples for the subclass of directive speech act verbs are given in (43) and (44). These verbs convey a high degree of imposition and the intensity of their directive force is indicated in dictionaries by manner adverbs such as *earnestly*, *urgently*, *fervently*, *anxiously*, and *desperately*. They include for example, *implore* and *beseech*, and their German counterparts *beknien* 'entreat,' and *beschwören* 'beseech' (see Panther, 1994, 1997; Köpcke and Panther, 2002 for a more detailed analysis):

46. a. Peter$_i$ beschwor Paul$_j$ [Ø$_j$ nicht zu rauchen].
b. Peter$_i$ beseeched Paul$_j$ [Ø$_j$ not to smoke].

c. *Peter$_i$ beschwor Ø$_j$ [Ø$_j$ nicht zu rauchen].
d. *Peter$_i$ beseeched Ø$_j$ [Ø$_j$ not to smoke].

47. a. Hans$_i$ bekniete Maria$_j$ [Ø$_j$ den Computer zu kaufen].
b. Hans$_i$ entreated Maria$_j$ [Ø$_j$ to buy the computer].
c. *Hans$_i$ bekniete Ø$_j$ [Ø$_j$ den Computer zu kaufen].
d. *Hans entreated Ø$_j$ [Ø$_j$ to buy the computer].

The second class of control verbs requiring an explicit object controller comprises verbs that entail the propositional content expressed in the complement clause. Such implicative verbs are for example, *induce* and *force* and their German equivalents *veranlassen* and *zwingen*, respectively, exemplified in (48) and (49):

48. a. Sie$_i$ veranlasste ihn$_j$ [Ø$_j$ ein neues Projekt zu beginnen].
b. She$_i$ induced him$_j$ [Ø$_j$ to start a new project].
c. *Sie$_i$ veranlasste Ø$_j$ [Ø$_j$ ein neues Projekt zu beginnen].
d. *She$_i$ induced Ø$_j$ [Ø$_j$ to start a new project].

49. a. Helga$_i$ zwang Otto$_j$ [Ø$_j$ das Projekt aufzugeben].
b. Helga$_i$ forced Otto$_j$ [Ø$_j$ to abandon the project].
c. *Helga$_i$ zwang Ø$_j$ [Ø$_j$ das Projekt aufzugeben].
d. *Helga forced Ø$_j$ [Ø$_j$ to abandon the project].

In sum, the data in (46)–(49) suggest the following generalization: the more manipulative the meaning of a control verb, the more likely the controller is explicitly coded (Panther, 1997).

Conclusion

Human languages employ various means to track referents on the sentence and discourse levels. Control relations can be seen in this larger context. As with other coreference relations, it turns out that the interpretation of control structures involves a complex interplay of syntactic, semantic, and pragmatic parameters.

See also: Anaphora; Deixis; Generative Grammar; Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects; Reference and Representation: Philosophical Aspects; Syntax (General).

Bibliography

Bach, E., 1979. Control in Montague grammar. Linguistic Inquiry 10, 515–531.
Bresnan, J., 1982. Control and complementation. Linguistic Inquiry 13, 343–434.
Chomsky, N., 1980. On binding. Linguistic Inquiry 11, 1–46.
Comrie, B., 1984. Subject and object control: syntax, semantics, pragmatics. In: Berkeley Linguistics Society: Proceedings of the Tenth Annual Meeting. Berkeley Linguistics Society, Berkeley, CA.
Comrie, B., 1985. Reflections on subject and object control. Journal of Semantics 4, 47–65.
Culicover, P.W., Jackendoff, R., 2005. Simpler Syntax. Oxford University Press, Oxford.
Davies, William D., Dubinsky, Stanley, 2006. The place, range, and taxonomy of control and raising. Syntax 9, 111–117.
Davies, Mark, 2008–. The Corpus of Contemporary American English: 450 million words, 1990-present. Available online at http://corpus.byu.edu/coca.
Duffley, P., 2014. Reclaiming Control as a Semantic and Pragmatic Phenomenon. Benjamins, Amsterdam.
Hornstein, N., 1999. Movement and control. Linguistic Inquiry 30, 69–96.
Hornstein, N., 2003. On control. In: Hendriks, R. (Ed.), Contemporary Grammatical Theory. Blackwell, Oxford, pp. 6–81.
Hornstein, N., Polinsky, M., 2010. Control as movement: across languages and constructions. In: Hornstein, N., Polinsky, B. (Eds.), Movement Theory of Control. Benjamins, Amsterdam, pp. 1–41.
Huang, Y., 1994. The Syntax and Pragmatics of Anaphora: A Study with Special Reference to Chinese. Cambridge University Press, Cambridge.
Jackendoff, R., Culicover, P.W., 2003. The semantic basis of control. Language 79, 517–556.
Köpcke, K.-M., Panther, K.-U., 2002. Zur Identifikation leerer Subjekte in infinitivischen Komplementsätzen – ein semantisch-pragmatisches Modell. Folia Linguistica 36, 191–218.
Landau, I., 2000. Elements of Control: Structure and Meaning in Infinitival Constructions. Kluwer, Dordrecht.
Landau, I., 2004. The scale of finiteness and the calculus of control. Natural Language & Linguistic Theory 22, 811–877.
Landau, I., 2010. The explicit syntax of implicit arguments. Linguistic Inquiry 41, 357–388.
Landau, I., 2013. Control in Generative Grammar: A Research Companion. Cambridge University Press, Cambridge.
Nishigauchi, T., 1984. Control and the thematic domain. Language 60, 215–250.
Panther, K.-U., 1994. Kontrollphänomene im Englischen und Deutschen aus semantisch-pragmatischer Perspektive. Narr, Tübingen.
Panther, K.-U., 1997. An account of implicit complement control in English and German. In: Verspoor, M., Lee, K.D., Sweetser, E. (Eds.), Lexical and Syntactical Constructions and the Construction of Meaning. Benjamins, Amsterdam.
Panther, K.-U., Köpcke, K.-M., 1993. A cognitive approach to obligatory control phenomena in English and German. Folia Linguistica 27, 57–105.
Postal, P.M., 1970. On coreferential complement subject deletion. Linguistic Inquiry 1, 439–500.
Rosenbaum, P.S., 1967. The Grammar of English Predicate Complement Constructions. MIT Press, Cambridge, MA.
Rosenbaum, P.S., 1970. A principle governing deletion in English sentential complementation. In: Jacobs, R., Rosenbaum, P.S. (Eds.), Readings in English Transformational Grammar. Ginn, Waltham, MA.
Růžicka, R., 1999. Control in Grammar and Pragmatics. Benjamins, Amsterdam.
Sag, I.A., Pollard, C., 1991. An integrated theory of complement control. Language 67, 63–113.
Stiebels, B., 2007. Towards a typology of complement control. In: Stiebels, B. (Ed.), Studies in Complement Control (ZAS Papers in Linguistics 47). Zentrum für Allgemeine Sprachwissenschaft, Berlin, pp. 1–80.
Wegener, H., 1989. 'Kontrolle'–semantisch gesehen: Zur Interpretation von Infinitivkomplementen im Deutschen. Deutsche Sprache 17, 206–228.
Williams, E., 1980. Predication. Linguistic Inquiry 11, 203–238.

Control Theories and Crime

Chester L Britt, School of Criminology and Criminal Justice, Northeastern University, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Control theories of crime focus on the inhibiting effects of conventional social institutions on criminal behavior. In general, control theories of crime emphasize how strong social ties to institutions, such as one's family (e.g., parents, spouses, and children), peer group, school, church, community, and workplace, among others, are expected to reduce the likelihood of crime by highlighting the negative consequences of criminal acts for those social ties. The cumulative body of contemporary criminological research is largely supportive of the key expectations of control theories of crime.

What Are Control Theories of Crime?

Control theories of crime represent a wide range of explanations that have their roots in the classical school of thought (Britt and Gottfredson, 2003; Roshier, 1989). The fundamental component to any control theory of crime is the expectation that individuals will engage in criminal activity if it is perceived to be in their self-interest. What this means is that it is the responsibility of the society to monitor individuals so that it reduces the chances of crime being perceived as beneficial in some way and the negative consequences are more clearly understood. To be clear, no assumption of rationality is required here – all individuals will recognize the immediate benefits of a wide range of criminal acts, but will differ in their perceptions of the consequences associated with those specific criminal acts. If the consequences do not outweigh the apparent benefits of the criminal act, it is more likely to occur. A flawed or short-term oriented decision-making process is not important for control theory. The theory allows for the likelihood that decisions may be made very quickly, often under the influence of alcohol or drugs. Regardless of how impaired (or not) the decision-making process is, it is still decision-making made by individuals which is reflective of their perceived self-interest at the moment of some act.

The individual-level assumption crucial to all control theories is the idea that individuals are, by nature, self-interested. The ability or degree to which individuals are able to think through the various long- and short-term consequences of all their actions conditions whether the criminal act is perceived to hold some benefit – be it economic, psychological, or social. At the societal level, the assumption critical to control theories is the expectation that society can introduce mechanisms that will either prevent or reduce the chances where an individual will perceive a criminal act as beneficial. As described below, these mechanisms may take a variety of different forms, but all have as the goal the prevention of criminal acts by enhancing the negative consequences of crime. At a much more general level, what makes the intervention of society possible is the functioning of a social order that reflects a broad set of common values and norms. For example, Ruth Kornhauser noted, "Rules governing the use of force and fraud are a prerequisite of group life" (1978, p.250). Social order is sustained by a wide range of social institutions – both formal and informal – that reflects these values and norms.

The combination of assumptions about individual self-interest and social control of behavior leads to the fundamental claim of all control theories: crime is caused by the weak control of individuals. To the extent that individuals are highly controlled, the likelihood of their committing criminal acts should be relatively low. The personal and social costs of being caught are high enough such that individuals would typically not perceive the commission of a crime to be in their self-interest. Alternatively, for individuals who are weakly controlled, the likelihood of crime commission should be considerably higher, since the consequences at either the individual or the social level are perceived as small relative to the pleasure or satisfaction gained from committing some crime.

Sources of Control

Control theories of crime locate the primary cause of crime in the weak control of individuals. What are the sources of control? There are several different dimensions of control, but the following discussion is limited to the two primary dimensions of control: (1) a distinction between internal and external sources of control and (2) a distinction between formal and informal control.

Internal versus External Control

The location of control – whether internal or external to the individual – is an important differentiator of control theories of crime. Internal control refers to mechanisms that function within the individual and include such characteristics as self-control (Gottfredson and Hirschi, 1990), morality (Wikström et al., 2012), and the like. These kinds of individual characteristics are expected to influence individual behavioral choices in ways that may alternatively decrease or increase the chances of crime commission, but not as motivations for crime. No special motivation is needed for a control theory of crime, since the benefits of many types of crime (e.g., financial) will be obvious to nearly everyone. Rather, from the perspective of a control theory, these kinds of individual characteristics have the potential to prevent crimes from occurring by operating in ways that highlight the negative consequences of crime for an individual. For example,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45047-6

according to Gottfredson and Hirschi (1990), individuals with low levels of self-control will have short-term orientations and make behavioral choices without thinking through the long-term consequences. An implication is that criminal acts – something that involves the use of force or fraud – will be more attractive because of the immediate benefits provided to the offender. In a similar way, Wikström and colleagues (e.g., 2012) explain how an individual's level of morality – how strongly the person holds common values of right and wrong – will influence behavioral choices. For an individual with a high level of morality, criminal and delinquent acts will be avoided, because they conflict with the individual's personal morality – that sense of whether an act is right or wrong. In both of these cases, the individual characteristic – self-control or morality – is viewed as the mechanism that inhibits criminal acts from being committed by emphasizing the negative personal and social consequences of crime.

External control refers to mechanisms of control located outside the individual. A strictly external source of control would be direct supervision or surveillance by another person or group. Most individuals will avoid committing criminal acts when they know they are being watched – by a parent or spouse, by a camera, by a supervisor at work, or by a police officer with a radar gun alongside a highway. There is no social meaning attached to these types of situations, just that the individual is inhibited from committing some acts out of a concern of being caught and the personal, social, and/or legal consequences that would likely follow.

External controls may also reflect the ties that individuals have to a variety of social institutions. These kinds of controls fall somewhere between strictly internal and external controls and have been referred to as a form of 'virtual supervision.' This is control that comes from both internal and external sources. Most commonly, the external source of control is assumed to be family, peers, school, the workplace, and other social institutions to which individuals are connected. The parent, the teacher, or the supervisor may not be physically present to offer supervision over the individual, but they are psychologically present within the individual, so that they have the power to affect behavioral choices – hence the notion of virtual supervision.

How is this kind of supervision possible? For control theory, the key theoretical mechanism is some notion of a 'social bond' (Hirschi, 1969) that represents how well connected an individual is to any number of conventional social institutions. The stronger the social bond, the stronger the ties of the individual to conventional society and a reduced likelihood of crime, because the social costs of crime outweigh whatever immediate benefits may be provided by the crime. This does not mean that individuals are engaged in some kind of rational calculus about how or whether a criminal act would have social costs and how to quantify these costs. Rather, it is the idea that engaging in some act, should it be found out by one's parents, peers, or spouse, would create difficult interpersonal situations that most people will prefer to avoid, which consequently inhibits a great many acts.

Two other examples are useful for highlighting the notion that internal controls are inextricably linked to external controls and that separating the two types of controls is often difficult. Hirschi's (2004) reconceptualization of the social bond and linking it to self-control provides one example, while Wikström's (Wikström et al., 2012) situational action theory and its emphasis on morality provide another example. In both approaches, an individual-level characteristic is conceptualized through an individual's links to conventional society. For example, Hirschi (2004) suggests that self-control can be measured, in part, by the level of attachment to a parent, which has traditionally been viewed as a measure of the social bond (Hirschi, 1969). This assertion reflects the notion that internal controls condition external controls – the ability of a spouse or a parent to influence behavioral choices will be contingent on the degree of internal control an individual has, and whether the social consequences of an action are given any consideration. Put another way: Does the social tie matter for the individual? If yes, to what degree? Alternatively, Wikström and colleagues' (Wikström et al., 2012) conceptualization of individual morality as how strongly the person holds common values about right and wrong grounds an internal control within an external control, again blurring the lines between internal and external sources of control, but with a theoretical mechanism that works in much the same way as that described by Hirschi (1969, 2004).

Formal versus Informal Control

The second primary dimension on which to consider sources of control is a dimension of formal and informal sources of control. Formal sources of social control are typically represented by the criminal justice system and the role that legal sanctions – both threat and application – may play in affecting behavioral choices. Perhaps the most common form where this concern appears in the criminological literature is in the application and test of 'deterrence theory,' where the key idea being assessed is the expectation that individuals will consider the costs and benefits of potential criminal acts. When the potential costs of crime – indicated by legal sanctions – outweigh the potential benefits, deterrence theory suggests that individuals will refrain from committing crimes. It is enough to note here that it is the expected role of the state, through the application of laws and sanctions, by the criminal justice system that is the primary source of control over the behavior of individuals.

In contrast to the emphasis on formal institutions of social control, informal controls refer to a wide range of social institutions, such as, the family, the school, the workplace, as well as the friends and peers of the individuals. The mechanism that links any informal social institution to behavioral choices is comparable to that linking a formal institution of social control to crime: Behavioral choices have consequences for the individual and their social ties. Considering the extent to which a criminal act would make an important social relationship vulnerable, individuals would less likely engage in that behavior. Hirschi's (1969) description of the social bond and its potential to affect behavioral choices helped to clarify the role that family, school, and peers play as sources of control. Specifically, the stronger the emotional attachment and commitment to any conventional social institution, the less likely criminal choices become, due to restrictive impact of the social institution on the individual's behavior. For example, youth with close relationships with at least one parent are less

likely to commit criminal acts out of concern for maintaining that relationship and not wanting to disappoint that parent. This is the 'virtual supervision' noted above that helps to regulate behavioral choices, as the youth thinks about how a parent might react is present to see what was going on (Hirschi, 1969).

Varieties of Control Theory

A focus on specific examples of control theory reveals that there are often two primary ways of distinguishing the control theory mechanism within each explanation. One distinction pertains to the level of explanation: Does the theory try to explain crime at the level of the individual (a micro-level approach) or at the level of the neighborhood, community, or some larger geospatial unit of analysis (a macro-level approach)? In addition, control theories can be distinguished by whether they are seemingly static and attempt to explain crime at some fixed point in time or dynamic and attempt to explain crime over time. These two ways of distinguishing control theory mechanisms are complementary, leading to the possibility of control theories that are micro-static, micro-dynamic, macro-static, and macro-dynamic.

The following discussion highlights some of the exemplars of different varieties of control theory. Although each theory to be discussed may not be viewed as a control theory by their primary proponents, the key mechanism for each one is that weak control of individuals results in a higher likelihood of crime. The specific sources of control vary across the different explanations, consistent with the internal–external and formal–informal dimensions noted above, but when pushed to explain why crime becomes more or less likely, each of them relies on a notion of weak control.

Macro- versus Micro-Control Theories

One of the more common misconceptions that appears in the research literature is for many authors to assume that control theories are micro-level theories of crime, such as, those described by Hirschi (1969), Gottfredson and Hirschi (1990), Sampson and Laub (1993), and Laub and Sampson (2003). In contrast, other authors have taken a more ecumenical approach to control theory and begin with the view that control theories are macro-level explanations of crime first and micro-level explanations, second. For example, Kornhauser (1978) used the control theory label on social disorganization theory as well as a number of other macro-level theories of crime in a way that allowed her to distill the common themes of many different versions of control theory. Her approach makes a great deal of sense: Control theory is based on a conception of social order that is separate from cultural values, that is, norms that prohibit nearly all forms of assault, murder, theft which are by necessity universal (Kornhauser, 1978, p. 40). This makes control theory inherently a macro-level theory of crime, even if the focus turns to individual characteristics, such as, morality, self-control, or the social bond. The following discussion of control theories is brief, but intended to highlight some of the key works that have been the basis for numerous other tests in the research literature and ongoing theoretical development in criminology.

At the macro-level, social disorganization theory, whether it is the classic statement by Shaw and McKay (1942) or the more recent reframing of disorganization theory by Bursik (1988), locates the cause of crime in the breakdown of social institutions. There are numerous forces – economic, cultural, social – that lead to the deterioration of social ties to and within these institutions; but at the end of any sort of 'causal explanation' of crime is the notion that it is weak social institutions, and the inability of those institutions to regulate (control) the behavior of individuals that leads to more crime. Sampson's many contributions to the development of the notion of 'collective efficacy' as an explanation of crime follow in this tradition (see, e.g., Sampson, 2012; Sampson et al., 1997) by describing how the dynamics of social institutions within a community (however defined) have the effect of inhibiting (or not) criminal behavior. Collective efficacy is usually defined as the shared expectations about social control (e.g., Sampson, 2012, p. 152). More efficacious communities are expected to have lower crime rates, because residents of those communities exhibit a greater level of social cohesion, reflected in the strength of social ties to each other and social institutions, as well as higher expectations of community action aimed at preventing problems from occurring. Initial statements of routine activities focused on macro-level explanations for patterns of crime (Cohen and Felson, 1979; Felson and Cohen, 1980, 1981) that highlighted the importance of social ties, but in the form of supervision. One of the important claims made within the routine activities approach was the impact of economic conditions on the weakening of social institutions and how weaker social institutions were less capable of providing adequate supervision of both individuals and property in neighborhoods, resulting in higher rates of both property and violent crime.

Micro-level control explanations are more commonly thought of as control theories of crime. This is likely a consequence of the impact of Travis Hirschi's *Causes of Delinquency* (1969), which focused on explaining delinquency among adolescents and has become one of the classic statements of a control theory. For Hirschi, it was the strength of the social bond – the level/degree of attachment, of commitment, of involvement, and of belief to conventional others – that was key to understanding crime. Individuals with weak social ties were less likely to be influenced by conventional others and more vulnerable to commit crime. More recently, Gottfredson and Hirschi's (1990) linking individual propensity (self-control) to crime with social situational characteristics (opportunities) provides a further elaboration on why social ties may be more or less effective at controlling the criminal behavior of individuals. Recent research on the routine activities of youth and adults has helped to more directly identify the mechanisms through which individual characteristics make some persons more or less likely to commit crimes (Osgood et al., 1996; Wilcox et al., 2003).

More recently, there has been an expansion of what might be called multi-level theories of crime that link the macro- and micro-level explanations of crime. Over the past two decades, this has been one of the key areas of research attention, with the theoretical mechanism driving much of the work

attributed to Sampson (e.g., 2012; Sampson et al., 1997). In particular, Sampson's (2012) *Great American City* is a comprehensive study of social conditions and crime and attempts to integrate years of work that have been more strictly focused on macro- or micro-level explanations. Of note, from a control theory perspective, Sampson's (2012) approach shows time and again how the characteristics of a neighborhood, or community, affect social ties to a range of other social institutions at the micro-level. Ultimately, Sampson's theoretical approach locates the primary 'cause' of crime in the weakened control of individuals.

Static versus Dynamic Control Theories

Static theories of crime typically focus on trying to explain patterns of crime at a single point in time, while dynamic theories of crime focus on trying to explain how changes within individuals over time affect criminal behavior. Much of the more recent effort in micro-level control theories has tended to focus on developing dynamic theories of crime, relying on concepts in developmental psychology and life course sociology. Unfortunately, some authors use the descriptor 'static' for a theory in a way that often has a pejorative connotation, yet many of the 'dynamic' theories of crime use a static version as the foundation for the explanation and become dynamic when the author incorporates elements of time into the explanation.

Two of the most frequently cited and tested control theories of crime are typically described as static. Hirschi's (1969) description of the social bond and its link to crime took the strength of the social bond at a point in time as a given parameter. How the social bond had been strengthened or weakened over some period of time was viewed as less important. Rather, if one was trying to explain recent patterns of criminal behavior, the expectation was that the strength of social ties at that point in time the crime had been committed was of greatest importance. The factors that led to the bond being weakened or strengthened did not alter the proximity of the current social ties to crime. In a similar way, Gottfredson and Hirschi's (1990) linking of low self-control to crime dismissed the idea that the factors affecting the individual's level of self-control were necessarily predictors of criminal behavior. Again, the proximal cause of crime was self-control, and to a lesser degree opportunity. How or whether an individual's self-control was increasing or decreasing over time did not matter as an explanation that at a fixed point in time, level of self-control affects the likelihood of crime.

More recently, Laub and Sampson's (Laub and Sampson, 2003; Sampson and Laub, 1993) age-graded theory of informal social control, links changes in social ties to changes in criminal behavior across the life course. In general, they argue that as social ties to conventional others increase across the life course, for example, through marriage and/or becoming a parent, criminal behavior becomes less likely due to the increasing social costs of crime. From a theoretical perspective, there is no incompatibility between their work and that of either Hirschi's (1969) or Gottfredson and Hirschi's (1990). For all of these perspectives, the key theoretical mechanism is the strength of social ties – crime is higher among those individuals with weaker social ties to conventional others. Laub and Sampson's (2003; Sampson and Laub, 1993) theoretical perspective is a straightforward and important extension of Hirschi's (1969) social control theory that focuses attention on explaining changes in criminal behavior over time, rather than criminal behavior at some fixed point in time.

Empirical Support for Control Theories of Crime

Control theories of crime are perhaps the most frequently tested theories in criminology and have been for several decades. In large part, this stems from the theory being stated in ways that facilitate measurement and empirical tests. The following discussion does not attempt to provide a comprehensive summary of the research literature and its support for control theories of crime. Instead, the intent is to highlight some of the key findings on the correlations of crime most relevant to control theories.

At the micro-level, the key findings focus on the individual propensity to commit crime, the effects of families, the effects of peers, the effects of the routine activities of youth and young adults, and patterns of versatility and stability in offending. In an early meta-analysis of the effect of self-control on crime, Pratt and Cullen (2000) found moderately strong evidence for a relationship between self-control and crime. In the years since their analysis was published, numerous studies have continued to document a persistent relationship, regardless of how self-control or some other measure of individual propensity may have been measured. In what represents an important shift from initial reactions to Gottfredson and Hirschi's description of self-control theory (e.g., Grasmick et al., 1993), it is more typical now for an individual propensity to crime to be assumed as a given parameter and included as a 'statistical control' variable in a multivariate analysis, where the researcher's attention is focused on some other measure's relationship with crime.

Much of the history of the study of crime has found evidence for family effects. Of particular note, the structure of the family matters less than the relationships within the family (see, e.g., Farrington et al., 2012). Parental separation, perhaps through death or divorce, has often been correlated with increased delinquency, but it appears to be largely a function of changes in effective parental supervision (Demuth and Brown, 2004). What seems to matter most is the quality of the parenting (see, e.g., Farrington et al., 2012; Piquero and Jennings, 2012; Sampson and Laub, 1993) – ineffective parenting skills result in children and youth who have trouble outside home and consequently a greater likelihood of criminal activity.

Another longstanding correlate of crime is the effect of peers. Until recently, much of the research on peer effects relied on the self-reported behavior of peers, which was rightfully viewed with suspicion: How well would an individual know about the criminal behavior of his or her peers unless being present for it? This has led some researchers to assume that reports of peer crime are simply another measure of self-reported crime (e.g., Gottfredson and Hirschi, 1990). More recently, however, the more frequent collection of peer network data, where individuals and their friends are surveyed separately, has demonstrated that peer effects observed in prior self-report

delinquency surveys are not spurious. In analyses using network data, a consistent pattern of comparable levels of self-reported crime across peers has been reported (e.g., Haynie et al., 2005). What these data have been unable to answer, thus far, is why peers have comparable levels of crime (e.g., Warr, 2002).

The adaptation of the routine activities approach to individual-level data has confirmed a number of expectations in the field, showing that youth who have more free time in unstructured activities have greater levels of crime and victimization (Osgood et al., 1996; Wilcox et al., 2003). This research has been key to understanding how individual-level characteristics that make individuals more or less crime-prone interact in important ways with various social conditions. In some cases, the social conditions will inhibit crime, regardless of the individual's propensity to commit an act, while other social conditions will enhance the possibility that crime occurs (Felson, 2006; Wikström et al., 2012). Importantly, too, this research has shown that an individual's propensity to commit crimes affects choices of social settings – more crime-prone individuals will tend to seek out crime-prone social situations, and vice versa (Wikström et al., 2012).

Gottfredson and Hirschi's (1990) self-control theory of crime was based on the empirical findings of versatility and stability in offending patterns – two facts of crime that Hirschi (1969) was not aware of at the time he wrote *Causes of Delinquency*. The many findings on the versatility and stability of offending (e.g., Britt, 1996; Farrington et al., 1988) have been some of the more important new empirical regularities uncovered in the field, since Hirschi's initial test of social control theory. It is difficult to apply many of the other theories of crime to explain why the primary correlates of criminal behavior also tend to be the strongest correlates of other behaviors, such as, accidents, injuries, failure of personal relationships, and a poor employment history (see, e.g., Laub and Sampson, 2003; Piquero et al., 2007).

At the macro-level, the key findings focus on routine activities, collective efficacy, institutional anomie, and concentrated disadvantage. Cohen and Felson's (1979; Felson and Cohen, 1980, 1981) research on a macro-level routine activities approach to understanding post-World War II crime rates marked an important shift from traditional notions of social disorganization, but relied on many of the same indicators of social problems to try to explain patterns of crime. The key findings from this research – demographic, economic, and other indicators of social change affected rates of crime – reinforced some of the basic claims of social disorganization theory as highlighted in Kornhauser (1978), but explained the links in different ways. More importantly, perhaps, the macro-level research on routine activities in the 1970s and 1980s helped to set the stage for a new wave of theoretical development and research on macro-level causes of crime.

Collective efficacy refers to social cohesion and trust among residents of a neighborhood that interacts with expectations for social control (Kirk, 2010; Sampson et al., 1997; Sampson, 2012). From a more traditional social disorganization perspective, collective efficacy represents the strength of social ties among individuals in a neighborhood that affect the likelihood of crime: the greater the collective efficacy of a neighborhood, the greater the inhibitor to crime. The argument holds both cross-sectionally and longitudinally – whether comparing neighborhoods at a single point in time or a single neighborhood over time, the efficacy of a neighborhood results in a greater ability to monitor and regulate the individuals within that neighborhood, reducing the chances of crime.

A closely related set of findings focuses on what has been labeled institutional anomie (Messner and Rosenfeld, 2007). Instead of thinking about social cohesiveness in terms of the accumulation of individual social ties, institutional anomie considers the cohesiveness of specific social institutions, such as the economy, the family, and the education system. Much of the research aimed at testing institutional anomie theory has found that when traditional institutions of social control – families, churches, schools – become less effective, perhaps due to changes in the economy, then these institutions become less effective at regulating populations, which in turn leads to more crime. The mechanism is similar to that of collective efficacy, but the conceptualization of social cohesiveness and what it represents differs.

Concentrated disadvantage represents a variety of socioeconomic conditions, including poverty, unemployment, and public assistance, that tend to be positively correlated with each other and more prevalent in those neighborhoods with low levels of collective efficacy and social cohesion. Those neighborhoods with high unemployment rates will also tend to have higher rates of poverty (overall and child), as well as higher levels of public assistance, combining to indicate the level of economic difficulty/disadvantage across neighborhoods. Again, a consistent finding has been the positive correlation of concentrated disadvantage with crime rates – neighborhoods with greater socioeconomic problems have higher rates of crime (e.g., Morenoff et al., 2001), but it is not concentrated disadvantage that directly causes crime. Rather, concentrated disadvantage weakens social ties of the residents in those neighborhoods, which in turn frees individuals to commit crimes.

Overall, some of the most consistent findings in criminology over the past 30+ years have focused on the individual and the situational/social characteristics best explained by control theories of crime because of the effect they have on traditional forms of informal social control, which in turn affect the likelihood that individuals commit criminal acts. Interestingly, many of the publications that have contributed to our knowledge base did not set out to test or to add to the evidence in support of control theory. Yet, when the most consistent predictors are individual propensity, family ties, versatility and stability in offending, routine activities (micro and macro), concentrated disadvantage, and various forms of social cohesion at the neighborhood or community level, the empirical evidence points to a solid body of research that supports the central claims of control theories of crime.

See also: Collective Efficacy and Crime; Crime: Knowledge about and Prevalence; Criminology: Development and Current Status of the Discipline; Delinquency, Sociology of; Routine Activities, Criminal Opportunities, Crime and Crime Prevention; Self-Control Theory and Crime.

Bibiliography

Britt, C.L., 1996. The measurement of specialization and escalation in the criminal career: an alternative modeling strategy. Journal of Quantitative Criminology 12, 193–222.

Britt, C.L., Gottfredson, M.R., 2003. Control Theories of Crime and Delinquency: Advances in Criminological Theory, vol. 12. Transaction Publishers, New Burnswick, NJ.

Bursik Jr., R.J., 1988. Social disorganization and theories of crime and delinquency: problems and prospects. Criminology 26, 519–552.

Cohen, L.E., Felson, M., 1979. Social change and crime rate trends: a routine activity approach. American Sociological Review 44, 588–608.

Demuth, S., Brown, S.L., 2004. Family structure, family processes, and adolescent delinquency: the significance of parental absence versus parental gender. Journal of Research in Crime and Delinquency 41, 58–81.

Farrington, D.P., Loeber, R., Ttofi, M., 2012. Risk and protective factors for offending. In: Welsh, B.C., Farrington, D.P. (Eds.), The Oxford Handbook of Crime Prevention. Oxford University Press, New York, pp. 46–69.

Farrington, D.P., Snyder, H.N., Finnegan, T.A., 1988. Specialization in juvenile court careers. Criminology 26, 461–488.

Felson, M., 2006. Crime and Nature. Sage, Thousand Oaks, CA.

Felson, M., Cohen, L.E., 1980. Human ecology and crime: a routine activity approach. Human Ecology 8, 389–406.

Felson, M., Cohen, L.E., 1981. Modeling crime rate trends – a criminal opportunity perspective. Journal of Research in Crime and Delinquency 18, 138–164.

Gottfredson, M.R., Hirschi, T., 1990. A General Theory of Crime. Stanford University Press, Palo Alto, CA.

Grasmick, H.G., Tittle, C.R., Bursik Jr., R.J., Arneklev, B.J., 1993. Testing the core empirical implications of Gottfredson and Hirschi's general theory of crime. Journal of Research in Crime and Delinquency 30, 5–29.

Haynie, D., Giordano, P.C., Manning, W.D., Longmore, M.A., 2005. Adolescent romantic relationships and delinquency involvement. Criminology 43, 177–210.

Hirschi, T., 1969. Causes of Delinquency. University of California Press, Berkeley, CA.

Hirschi, T., 2004. Self-control and crime. In: Baumeister, R.F., Vohs, K.D. (Eds.), Handbook of Self-Regulation: Research, Theory, and Applications. The Guilford Press, New York, pp. 537–552.

Kirk, D., 2010. Sampson, Robert J.: collective efficacy theory. In: Cullen, F.T., Wilcox, P. (Eds.), Encyclopedia of Criminological Theory. Sage Publications, Thousand Oaks, CA, pp. 802–806.

Kornhauser, R.R., 1978. Social Sources of Delinquency: An Appraisal of Analytic Models. University of Chicago Press, Chicago.

Laub, J.H., Sampson, R.J., 2003. Shared Beginnings, Divergent Lives: Delinquent Boys to Age 70. Harvard University Press, Cambridge, MA.

Messner, S.F., Rosenfeld, R., 2007. Crime and the American Dream, fourth ed. Wadsworth, Monterey.

Morenoff, J.D., Sampson, R.J., Raudenbush, S.W., 2001. Neighborhood inequality, collective efficacy, and the spatial dynamics of urban violence. Criminology 39, 517–558.

Osgood, D.W., Wilson, J.K., Bachman, J.G., O'Malley, P.M., Johnston, L.D., 1996. Routine activities and individual deviant behavior. American Sociological Review 61, 635–655.

Piquero, A.R., Farrington, D.P., Blumstein, A., 2007. Key Issues in Criminal Career Research: New Analyses of the Cambridge Study in Delinquent Development. Cambridge University Press, New York.

Piquero, A.R., Jennings, W.G., 2012. Parent training and the prevention of crime. In: Welsh, B.C., Farrington, D.P. (Eds.), The Oxford Handbook of Crime Prevention. Oxford University Press, New York, pp. 89–101.

Pratt, T., Cullen, F.T., 2000. The empirical status of Gottfredson and Hirschi's general theory of crime: a meta-analysis. Criminology 38, 931–964.

Roshier, B., 1989. Controlling Crime: The Classical Perspective in Criminology. Lyceum Books, Chicago.

Sampson, R.J., 2012. Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago Press, Chicago.

Sampson, R.J., Laub, J.H., 1993. Crime in the Making: Pathways and Turning Points through Life. Harvard University Press, Cambridge, MA.

Sampson, R.J., Raudenbush, S.W., Earls, F., 1997. Neighborhood and violent crime: a multilevel study of collective efficacy. Science 277, 918–924.

Shaw, C.R., McKay, H.D., 1942. Juvenile Delinquency and Urban Areas. University of Chicago Press, Chicago.

Warr, M., 2002. Companions in Crime: The Social Aspects of Criminal Conduct. Cambridge University Press, New York.

Wikström, P.-O.H., Oberwittler, D., Treiber, K., Hardie, B., 2012. Breaking Rules: The Social and Situational Dynamics of Young People's Urban Crime. Oxford University Press, Oxford, UK.

Wilcox, P., Land, K.C., Hunt, S.A., 2003. Criminal Circumstance: A Dynamic Multi-contextual Criminal Opportunity Theory. Transaction, New Brunswick, NJ.

Control Variables in Research

Paras D Mehta, University of Illinois, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 4, pp. 2726–2730, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Control variables refer to variables whose effects on an outcome variable are statistically adjusted in order to estimate causal effects of an explanatory variable. This article discusses methods of controlling for both observed and unobserved sources of confounding. Regression and propensity score are presented as methods in the context of various research designs. Methods that rely on conditioning on observed variables are effective if all sources of confounding are measured and included in the analysis. These methods produce biased causal estimates in the presence of unobserved confounding. Instrumental and mediator variables are presented as methods for estimating effects of the primary explanatory variables in the presence of such confounding.

An important goal of empirical research is to support causal assertions about the effects of explanatory variables on outcome variables. Extraneous variables that influence the outcome represent potential confounds whose effects must be eliminated before causal assertions regarding the explanatory variable can be made. In its broad sense, the term 'control' refers to a researcher's systematic attempts to eliminate the confounding effects of extraneous variables associated with sampling, measurement, experimental manipulation, randomization, and statistical adjustment. The term 'control variable' is used most commonly in the narrow context of statistical adjustment of potential confounds.

The current presentation of 'control variables' is guided by three broad tenets. (1) Strategies of statistical control and its interpretation depend on an understanding of the nature of the relationship between extraneous variables and the explanatory and outcome variables. The term control variable is best understood within this broader explanatory framework. (2) The nature of potential confounds and therefore the need for their statistical control depends on the specific research design. The need for statistical control increases as the inherent controls imposed by the research design diminish. (3) The potential confounding variable may be observed or unobserved. Although most methodological and empirical literature on statistical control focuses on the control of observed variables, this article is organized around prototypical methods of controlling for both observed and unobserved sources of extraneous variability. This approach highlights the importance of a causal framework for understanding statistical control and also pinpoints the most common cause of the failure of observed variables methods, i.e., the presence of unobserved sources of confounding (Morgan and Winship, 2007).

Controlling for Observed Variables

Several methods of control are available when it can be assumed that the potential confounding variables are measured including regression, matching, and propensity score. Regression (Pedhazur and Pedhazur, 1991) or analysis of covariance (ANCOVA; Hutiema, 1980) based approaches are the most commonly used methods of statistical control. The purpose and interpretation of regression adjustment depends on the specific research design employed.

Regression and ANCOVA-Based Approaches

Regression of a dependent variable Y on an independent variable X provides an estimate of the causal effect of X on Y under the critical assumption that all omitted variables C that influence Y are uncorrelated with X. If an omitted extraneous variable C that influences Y is also correlated with the independent variable X, then the regression parameter is biased as a measure of direct effect of X on Y. In this case, the effect of explanatory variable X is confounded with the extraneous variable C. The true direct effect of X on Y (a in Figure 1) may be obtained by measuring and including C in the model. The key assumption made in regression-based approaches is that all extraneous variables are measured without error and included in the analysis.

Regression Adjustment in Randomized Experiments

Randomized experiments (Cook and Campbell, 1979) offer the most robust method for making causal inferences and imply two powerful means of control. These are: (1) the independent or explanatory variable is controlled by direct

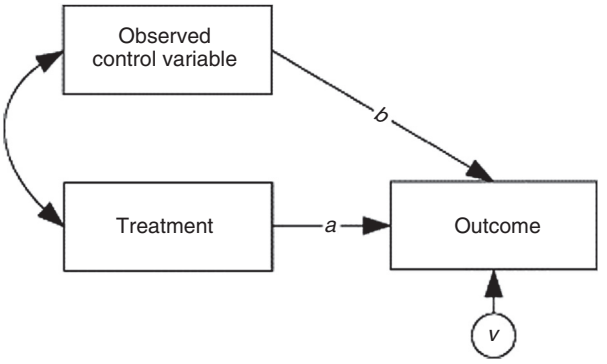

Figure 1 Path "*a*" shows the direct effect of the treatment on outcome after controlling for the effect '*b*' of the observed covariate.

manipulation. (2) All other sources of extraneous variability are controlled by randomization or chance. Random assignment, *if successful*, makes experimental and control groups probabilistically equal with respect to all potential confounding variables and hence minimizes bias due to initial differences among treatment conditions. As a result, extraneous variables are expected to be uncorrelated with the manipulated variable or the treatment assignment and their exclusion does not bias the estimates of treatment effects. On the other hand, including such variables in analysis reduces the unexplained within-group variability in the outcome, thus increasing the power for detecting treatment effects. Pretest measures of the dependent variable are frequently used for this purpose. It is important to note that randomization may fail to make the experimental conditions equal with respect to various confounders. In addition, noncompliance and attrition may render various treatment conditions noncomparable.

Regression Adjustment in Quasiexperimental and Nonexperimental Designs

In quasiexperimental designs, the independent variable is manipulated in the absence of random assignment. In these designs, subjects may be nonrandomly assigned to or may self-select themselves into different conditions. For example, in a study investigating the effects of remedial education on achievement, students who are expected to benefit the most may be assigned by the researcher into the treatment condition. As a result, the groups may not necessarily be equal at baseline with respect to various observed and unobserved characteristics that may influence the outcome of the treatment. In nonexperimental settings, the researcher may wish to infer causation from the observed association between a hypothesized independent variable and an outcome, e.g., the effect of smoking on lung cancer.

The purpose of regression adjustment in nonexperimental designs is to make the experimental conditions statistically equal with respect to various observed characteristics that may also influence the outcome. The success of regression-based approaches in quasiexperimental and observational research depends on the plausibility of the assumption that all potential confounding variables are measured without error and included in the analysis. This assumption may not be tenable in most quasiexperimental settings where self-selection or assignment to treatment may be based on unmeasured expectations about outcome under treatment and control conditions.

The observed characteristics that are typically controlled to include the pretest measure of the outcome variable and background characteristics such as gender, age, and socioeconomic status. The control variable frequently serves as a proxy for some other unmeasured theoretical variable of interest. This method of controlling for theoretical variables is a common practice in applied settings and leads to a number of additional problems. Proxy variables, unlike the measured indicators of a construct, contain *both* random error as well as unspecified sources of systematic errors. The presence of random measurement error in regressors leads to biased estimates of the regression parameters. On the other hand, the presence of systematic variability other than the hypothesized construct introduces new and unspecified sources of bias. As a result, the use of background characteristics as proxies in evaluation research may result in contradictory findings regarding the effects of the explanatory variable in question depending on the specific set of covariates included in the analysis.

Control Variable: Validity and Differential Treatment Effectiveness

The use of control variables for statistical adjustment is motivated primarily by a desire to increase the internal validity of the study. An alternative way of eliminating confounding due to extraneous variables is to include only those individuals at a specific level of the confounding variable. For example, if ethnicity and gender are related to the treatment assignment and to outcome, the researcher may choose to include only white males in the study. Such control by exclusion limits the generalizability of the findings to the population actually included in the study. In contrast, the regression-based approach allows generalization across all levels of the controlled covariate present in the sample if the treatment effects are the same within each level of the covariate. In other words, the possibility of differential treatment effects across levels of the covariate must be ruled out before the results can be generalized. For example, a specific intervention for treating childhood aggression may be more effective at higher levels of initial aggression. In this case, the initial level of aggression could be included in the analysis along with an interaction term with the manipulated variable to test if the magnitude of the treatment effect depends on the level of the observed covariate. When the interaction between the independent variable and a covariate is significant, the covariate is said to moderate the effect of the independent variable on the outcome. In this situation, the arbitrary distinction between an explanatory and a control variable begins to blur. The researcher must now explore the mechanism of differential treatment effectiveness and ascribe proper causal status to both the covariate and the explanatory variable.

Propensity Score

In contrast to the regression approaches, the propensity score approach explicitly models the treatment selection/assignment process using the *observed* background characteristics. The *propensity score* $P(T|C)$ is the probability that an individual with a given set of background characteristics C will be assigned to a particular treatment group (Rosenbaum, 1995; Rubin, 2006). If C also has an effect on the outcome variable, then conditional on the propensity score, the treatment assignment will be random with respect to the outcome. The propensity score captures the collective influence of all observed covariates on treatment assignment. In other words, by controlling for the propensity score, it is possible to control for all observed characteristics that determine treatment assignment. This can be achieved by including the propensity score in the regression equation (Figure 2) or by matching on propensity score. The propensity score may be estimated using logit or probit regression by regressing observed treatment assignment on the set of covariates and their interactions.

Controlling for Unobserved Variables

Both the regression and propensity score methods make a strong assumption that all potential sources of confound are

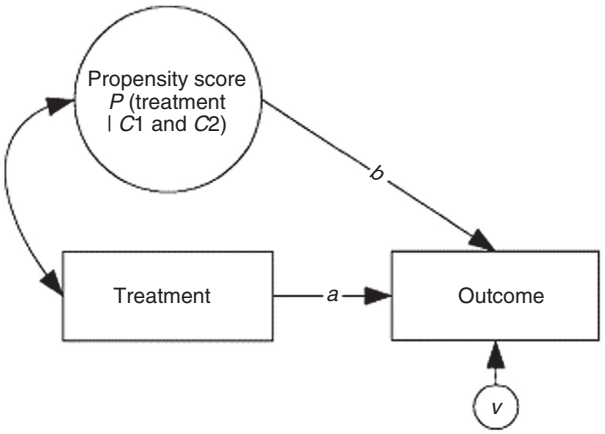

Figure 2 Path "*a*" shows the direct effect of the treatment on outcome after controlling for the effect '*b*' of the propensity score estimated using two observed variables *C*1 and *C*2 (not shown).

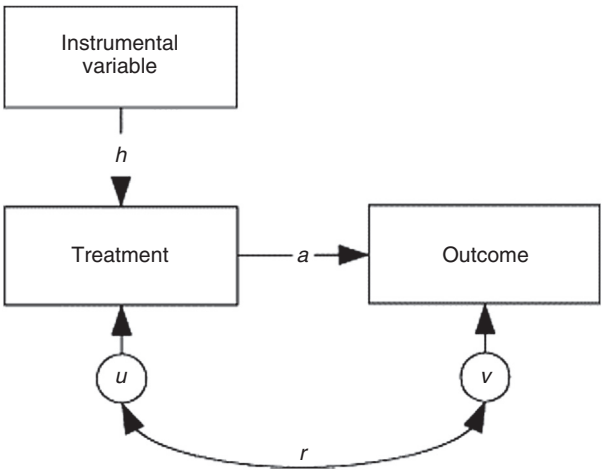

Figure 3 Path "*a*" shows the direct effect of the treatment on the outcome independent of the correlation (*r*) between the residuals for the treatment (*u*) and the outcome (*v*). This effect is identified by the presence of an instrumental variable that is presumed to influence the outcome exclusively via its effect on the treatment assignment (path *b*).

measured and included in the analysis. It is likely that selection into or exposure to a treatment depends on a subject's expectations regarding the outcome. For example, an individual's implicit beliefs regarding expected outcomes in the treatment and control groups may influence their decision to participate in a particular treatment as well as the eventual outcome. Even in randomized experiments, treatment compliance and attrition may be related to such unobserved variables. In this situation, observed background characteristics may not contain relevant information regarding a subject's decision process. As a result, methods of control based on observed variables are likely to produce biased estimates of treatment effects. Rosenbaum (1995) provides methods of assessing the sensitivity of the results to potential hidden biases in the context of observational studies. The problem of isolating true causal effects from observed association in the presence of latent confounds can be thought of as an identification problem. If one is willing to make certain assumptions about the nature of treatment effects under various conditions, it is possible to express treatment effects using inequalities or bounds (Manski, 1995).

Alternatively, the identification issue can be resolved by measuring additional variables that are causally related to the explanatory and outcome variables but are *assumed* to be independent of the unobserved confounds. In other words, the presence of such variables allows the true causal effects of an independent variable to be isolated from the unobserved sources of confounds. Such variables can help identify the true causal effects under two circumstances: (1) if the effects of these variables on the outcome are entirely mediated by the independent variable or (2) if the variable mediates the effect of the independent variable on the outcome.

Instrumental Variable Method

The instrumental variable approach for controlling unobserved sources of variability is the *mirror opposite* of the propensity score method for controlling observed variables (Angrist et al., 1996). Unlike an observed control variable, an instrumental variable is *assumed not* to have any *direct effect* on the outcome. Instead, the instrumental variable is thought to influence only the selection into the treatment condition. In other words, the effect of the instrumental variable on the dependent measure is entirely *mediated* via its effect on treatment assignment (*b* in Figure 3). This condition is also known as *exclusion restriction*. If the independent variable were regressed on the instrumental variable, the residual would contain all unobserved sources of variability that determine treatment assignment and also influence the outcome variable (represented by the correlation *r* in Figure 3). As a result, the existence of an instrumental variable *identifies* or *isolates* the average direct effect (*a* in Figure 3) of the treatment on the outcome independent of the unobserved sources of variability. The success of this strategy rests on the reasonableness of the assumption of exclusion restriction (Heckman et al., 2006).

Mediator Variable Method

The effects of unobserved confounds may be isolated in a radically different fashion by modeling the causal mechanism by which the treatment influences the outcome (Pearl, 2009). In this approach, the goal is to find an observed variable that *mediates* the effect of the independent variable on the outcome *but is unrelated to the unobserved common cause* (Figure 4). For example, in the past cigarette companies have claimed that the relationship between smoking and lung cancer is spurious: a result of a gene (common cause) that leads to craving for tobacco and also causes lung cancer. If such a common cause in fact existed, it is still possible to isolate the independent effect of smoking on lung cancer by identifying a variable that mediates the effects of smoking on cancer but is *not influenced by the hypothesized common-cause*. For example, tar in lungs is one such variable known to cause cancer which accumulates in lungs as a result of smoking, passive smoke inhalation, and environmental pollution. It is reasonable to argue that the level of tar in lungs is not caused by the common gene, especially among the passive smokers. The existence of such a mediator

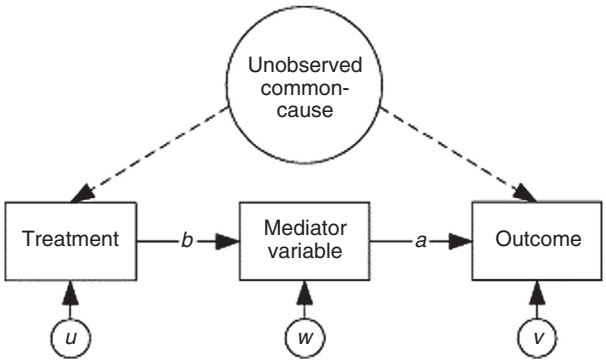

Figure 4 The direct effect of the mediator on outcome is represented by "*a*" and that of the treatment assignment on the mediator is represented by "*b*." The unobserved common-cause of treatment assignment and outcome is assumed not to influence the mediator. As a result, the mediated effect of the treatment on the outcome ($a \times b$) is independent of the confounding due to plausible common cause.

variable can statistically identify the effect of the independent variable on the dependent variable ($a \times b$ in Figure 4) even in the presence of unobserved sources of confounding. It is to be noted that the instrumental and mediator variables are not control variables themselves. Instead, the use of these variables identifies true treatment effects by isolating the confounding effects of other unobserved sources of variability.

See also: Causal Inference: Overview; Comparative Studies: Method and Design; Endogeneity: Methodology; Experimental Design: Bayesian Designs; Experimental Design: Overview; Experimental Design: Randomization and Social Experiments; External Validity; Internal Validity; Latent Structure and Causal Variables; Mechanism-Based Causal Analysis; Moderator Variable: Methodology; Nonequivalent Group Designs; Observational Studies: Overview; Quasi-Experimental Designs; Regression Discontinuity Design; Selection Bias, Statistics of; Social Experiments.

Bibliography

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91, 444–472.

Cook, T.D., Campbell, D.T., 1979. Quasi-Experimentation: Design & Analysis Issues for Field Settings. Rand McNally College Pub. Co., Chicago.

Heckman, J.J., Urzua, S., Vytlacil, E., 2006. Understanding instrumental variables in models with essential heterogeneity. The Review of Economics and Statistics 88, 389–432.

Hutiema, B.E., 1980. The Analysis of Covariance and Alternatives. Wiley, New York.

Manski, C.F., 1995. Identification Problem in the Social Sciences. Harvard University Press, Cambridge.

Morgan, S.L., Winship, C., 2007. Counterfactuals and Causal Inference: Methods and Principles for Social Research. Cambridge University Press, Cambridge.

Pearl, J., 2009. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge.

Pedhazur, E.J., Pedhazur, S.L., 1991. Measurement, Design and Analysis: An Integrated Approach. Lawrence Erlbaum Associates, Hillsdale, NJ.

Rosenbaum, P.R., 2002. Observational Studies, second ed. Springer-Verlag, New York.

Rubin, D.B., 2006. Matched Sampling for Causal Effects. Cambridge University Press, Cambridge.

Conventions and Norms: Philosophical Aspects

Richard H McAdams, University of Chicago Law School, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Conventions and norms are distinct but overlapping social regularities. 'Convention' has been variously defined as a type of agreement, a jointly accepted principle of action, and a set of coordinated expectations that solve a recurrent coordination problem. The related meaning of 'norm' is a regularity of behavior among a population of individuals, where the regularity is at least partly sustained by the fact that at least most individuals approve conformity to the regularity and/or disapprove nonconformity. Game theory addresses how conventions arise without agreement, but the process turns out to be difficult to explain based on rationality alone. Some theorists contend that norms are based on emotion rather than rationality, but others contend that the rational pursuit of interest can explain norm origin.

Conventions and norms have various meanings. Of importance to the social sciences, they refer to particular types of behavioral regularities and/or the beliefs or reasons that sustain the regularities. A predominant view of 'convention' is that it describes a kind of equilibrium behavior among individuals in a recurrent coordination problem, a basic feature of which is the existence of at least one potential alternative equilibrium. Others view conventions as a type of agreement or a jointly accepted principle of action. 'Norm' has several meanings. It can refer to a mere statistical regularity or, in moral theory, a prescriptive rule for human behavior. In the sense pursued here, however, a norm describes a regularity of behavior among a population of individuals, the central feature of which is that most or all the individuals approve conformity to the regularity and/or disapprove nonconformity. Thus, conventions and norms are distinct but overlapping social regularities.

Introduction

A few (putative) conventions will identify the topic and illustrate its importance. A common example is the regularity of driving on only one side of the road (left or right). Another is monetary currency (see Hume 1978[1740]: p. 490). The convention here is to treat a particular form of metal or paper as having value sufficient to exchange it for goods and services whose value is otherwise (without the convention) far greater. In each case, an individual wishes to follow the regularity once it arises – to avoid the danger of driving into oncoming traffic or the inefficiency of barter. Yet these conventions are arbitrary: everyone driving on the left works like everyone driving on the right; everyone accepting one paper form as currency works much like everyone accepting another.

Language is similarly thought to be conventional because different sounds or symbols could be used to refer a given thing, but to communicate successfully, an individual is better off using the sounds or symbols that others use. One might object to the idea that all of language is a convention if a convention is thought to require a convening, i.e., an actual agreement. An agreement to use certain sounds as language must be expressed in language, so some language must exist prior to any agreement-based convention (see Quine, 1967 [1936]). Lewis (1969) and others have since argued that a convention can occur without any agreement, suggesting that language may indeed be conventional.

Hume (1978[1740]: pp. 477–513) contended that justice, particularly the institution of property, was a convention (see Vanderschraaf, 1998a). Some theorists give conventions a central role in the theory of morality or law (see Law and Morality: An Analytical Perspective; Law and Morality: A Continental European Perspective; Social Work Ethics). Conventions and norms (discussed below) also explain 'spontaneous order' that arises without a centralized force such as the state. If so, conventions (or norms) might sustain some forms of cooperation in the 'state of nature' and thereby complicate or refute social contract theories that justify the state by the disorder that exists in its absence (see Skryms, 1996; Taylor, 1987). Once the state exists, spontaneous order may limit the domains in which state power is necessary (see Ellickson, 1991), or contrarily, state action may be necessary to restrict or unsettle undesirable conventions (or norms) (see McAdams, 1997: pp. 409–432).

Defining 'Convention'

Convention as Agreement or Jointly Accepted Principle of Action

On one view, a convention is a 'jointly accepted principle of action' (Gilbert, 1989: p. 377). A principle of action is a rule prescribing (at least presumptively) how to behave in certain situations. An individual can adopt a principle of action unilaterally or jointly. Gilbert (1990: p. 16) gives an example of the latter: *A* chastises *B* for failing to adhere to a previous telephone regularity they had observed (about who calls back when they are accidentally disconnected) without prior agreement. *B* acquiesces in the criticism and apologizes. By this communicative exchange, Gilbert says the parties have created a convention by 'jointly accepting' a principle of action requiring behavior they previously engaged in without convention. More precisely, "A population P has a convention of conformity to some regularity in behavior R in situations of type S if and only if the members of P are jointly committed to accept as a body, with respect to themselves, the fiat: *R is to be conformed to*" (Gilbert, 2008: p. 12). She describes her

account as 'holistic,' in contrast to the individualistic game theory approach described below (p. 5).

Joint acceptance or commitment does not obviously explain the common intuition that conventions are in some sense arbitrary. Marmor (1996) criticized Gilbert for misunderstanding and omitting this 'defining feature' of conventions from her account. More recently, Gilbert (2008: p. 12) explains that her definition captures a sense of arbitrariness because 'no particular rationale' for the commitment 'is presupposed' in the definition, unlike a rule sounding in morality, which has a moral rationale, or tradition, which has a historical rationale. Yet the agnosticism of Gilbert's definition on the rationale of joint commitment does not mean there is not usually or always, for each convention, some rationale. So it is not clear that her approach captures arbitrariness. In any event, if Marmor (1996) and others are right, the kind of arbitrariness required is the existence of an alternative regularity that could serve the rationale of the convention in the manner (although not necessarily as well as) the conventional regularity does.

In addition, Gilbert's telephone example reveals how close this understanding of convention is to the idea that conventions are simply agreements. Arguably *A* and *B* have implicitly agreed because each believes that *A*'s criticism impliedly asserted the existence of a mutual obligation and that *B*'s acquiescence revealed *B*'s assent to this assertion. The joint acceptance idea clarifies how agreements produce conventions, but not how conventions arise without agreement.

Conventions as Solutions to Recurrent Coordination Games

David Lewis' Definition

Lewis' *Convention* (1969) is the origin of a game theory view of convention, one that explicates the way in which they are arbitrary and contends that no explicit or implicit agreement is necessary. Lewis (1969: p. 42) initially modeled conventions as solutions to recurrent coordination games. A *game* is a situation in which the outcome of anyone's action depends on the decisions of other individuals as well as one's own. In game theory, *players* (usually) rationally choose among alternative *strategies*. Each combination of the players' strategies produces an *outcome*, which consists of a utility index for each player (see Rational Choice Explanation: Philosophical Aspects; Game Theory: Noncooperative Games).

Pure coordination games involve purely common interests among the players. Consider people who want to find each other but who have failed to communicate in advance where to meet. Lewis (1969: p. 9) uses the matrix in Figure 1 to describe such a coordination game involving two players and three possible meeting places. One player – 'R' – can play strategies R1, R2, or R3, which consist of going to places P1, P2, or P3, respectively. The other player – 'C' – can play strategies C1, C2, and C3, which consists of going to the same three places, respectively. The players choose their strategies simultaneously, not being able to observe what the other has done. Each receives more utility from meeting (1) than from not meeting (0). (The number in the lower left corner of a cell represents the utility of row chooser; the upper right number is the column-chooser utility.) Each player therefore has an interest (equal in this example) in coordinating his or her action with that of the other player.

	C1	C2	C3
R1	1 1	0 0	0 0
R2	0 0	1 1	0 0
R3	0 0	0 0	1 1

Figure 1 A pure coordination game.

As a result, there are three pure-strategy Nash equilibria, outcomes in which no one player can gain by changing his or her strategy (because each is playing a best reply to the strategies of the other players): R1/C1, R2/C2, and R3/C3. For example, R1 is the best response R can make to C1 and that same strategy – C1 – is the best response C can make to R1. Thus, given R1/C1, neither player would want to switch strategies unilaterally. The same is true for R2/C2 and R3/C3 too. (There are also mixed-strategy Nash equilibria, as where R and C each go to each of the three places with a probability of one-third.)

This analysis reveals the need for *coordinated expectations*. Except by accident, players 'solve' a coordination game only if they have accurate expectations about what the other(s) will do. In the example, each player needs to *expect* the other to go to the same place, for example, P1. But if each player knows the other wishes to meet, then, except by accident, it seems that R will expect to meet C at P1 only if R believes that C expects to meet R there as well. But C will expect to meet R at P1 only if C believes that R expects C to go there. Lewis (1969: pp. 52–60) claims the need for *common knowledge*. Common knowledge of a fact *x* means that everyone in the relevant population believes *x*, believes that everyone believes *x*, believes that everyone believes that everyone believes *x*, and so on, *ad infinitum*. Except by accident, coordination seems to require that each player have common knowledge of what the other(s) will do.

Finally, consider how a coordination game can recur. Such a situation would exist if, on repeated occasions, R and C find themselves involuntarily separated in a crowd and in need of a meeting place. More generally, the question of which side of the road to drive on, what paper to use as currency, and what sounds to use to communicate, all arise repeatedly.

Based on these points, Lewis (1969: p. 78) proposes this definition:

> A regularity *R* in the behavior of members of a population *P* when they are agents in a recurrent situation *S* is a *convention* if and only if it is true that, and it is common knowledge in *P* that, in almost any instance of *S* among members of *P*,

1. almost everyone conforms to *R*;
2. almost everyone expects almost everyone else to conform to *R*;
3. almost everyone has approximately the same preferences regarding all possible combinations of actions;

4. almost everyone prefers that any one more conform to R, on condition that almost everyone conform to R; and
5. almost everyone would prefer that any one more conform to R' on condition that almost everyone conform to R';

where R' is some possible regularity in the behavior of members of P in S, such that almost no one in almost any instance of S among members of P could conform both to R' and to R.

Notice that the convention is not merely the behavioral regularity, but also the set of expectations and preferences that sustain the regularity.

Modifications in Lewis' Definition

In several ways, Lewis' approach is narrow and theorists have proposed modifications. First, Lewis (1969: p. 78) requires common knowledge of all the necessary elements of a convention, even the fact that there is an alternative regularity (condition 5). Burge (1975) observes that requiring common knowledge of the alternative excludes an arbitrary regularity from being a convention merely because most individuals never discover, or later forget, what the alternative is. On this view, a population's language would not be conventional if the population were unaware of the possibility of other languages. Where Burge advocates against requiring common knowledge of condition 5, Skyrms (1998: p. 84) and Binmore (2008: p. 23) more generally doubt that any of the required common knowledge can arise in a large population. They propose instead the tools of evolutionary game theory, which do not require common knowledge or even rational agents. Cubitt and Sugden (2003) and Vanderschraaft (1998b) attempt to reconstruct the common knowledge concept to make it less demanding so as to answer this concern.

Moreover, Lewis's definition insists on two unnecessary features of pure coordination games, requiring (3) 'almost everyone has approximately the same preferences regarding all possible combinations of actions,' and (4) 'almost everyone prefers that any one more conform to R, on condition that almost everyone conform to R'. The latter point means that most individuals conditionally prefer both their own conformity and the conformity of others. Each requirement arguably excludes regularities that are plausibly conventions.

First, there is an element of coordination in many games that also involve conflict. As long as there is some common interest, one may wonder why the equilibrium outcome, when others are possible, is not a convention. For example, in Figure 2 (sometimes called 'the battle of the sexes' based on an early illustration), each player prefers coordinating at Nash equilibria R1/C1 or R2/C2 to either other outcome. But each player also strongly prefers one equilibrium to the other and these preferences conflict. Thus, the players do not have 'approximately' the same preferences for outcomes (which they would if one replaced each 5 with, say, 1.05). Lewis would not count a solution to this game as a convention even though it were one of two equilibria and the players each prefer coordinating at an equilibrium than reaching a nonequilibrium outcome.

	C1	C2
R1	5, 1	0, 0
R2	0, 0	1, 5

Figure 2 A simple mixed motive game.

Another weakness in Lewis' account is his requirement that the players conditionally prefer *other* players' conformity even though the stability of a regularity depends only on players preferring their own conformity. With a mix of common and conflicting interests, players may prefer that other players deviate from any of the multiple equilibria. In Figure 3, R1/C1 and R3/C3 are the only Nash equilibria. As in Figure 1, the players are indifferent between these equilibria and prefer to coordinate on either than to reach other outcomes that are worse for each (R1/C3, R2/C2, and R3/C1). But unlike Figure 1, in each of these equilibria, the players do *not* prefer that the other conform to the equilibrium because each would gain by the other playing a different strategy. At R1/C1, R prefers that C plays C2 and C prefers that R plays R2. The same is true at R3/C3.

Lewis (1969: pp. 45–46) acknowledges the point with this example: suppose those planning to attend a party wish to wear the same style of clothing that others wear. But suppose many 'nasty' people prefer that many others fail to conform to the regularity and thereby suffer embarrassment. Because these nasty people do not conditionally prefer that most others conform, Lewis would not consider it a convention even if, based on coordinated expectations, everyone arrives at the party dressed in the same style. The more common intuition is that, regardless of nasty preferences, a stable and arbitrary clothing regularity is a convention (see Miller, 1986: p. 123).

These examples suggest dispensing with the requirements (3) that players have approximately the same preferences for all outcomes and (4) conditionally prefer that others conform. A convention may then be more broadly but simply defined as the coordinated expectations that sustain a pure-strategy Nash equilibrium, in recurrent circumstances where multiple pure-strategy equilibria are possible, and the behavioral regularity that the equilibrium represents. Some would add stability conditions that narrow the definition. Others would expand the equilibrium concept (see Vanderschraaf, 1995) and possibly include mixed strategy equilibria where individuals play different pure strategies, each with some probability.

	C1	C2	C3
R1	5, 4	7, 2	1, 3
R2	3, 5	0, 0	2, 5
R3	4, 2	7, 3	5, 4

Figure 3 Adapted from Kornhauser, L.A., 1996. Conceptions of social rule. In: Draybrooke, D. (Ed.), Social Rules: Origin; Character; Logic; Change. Westview Press, Boulder, CO, p. 209.

On such a broad view, more regularities are conventions. For example, in an iterated prisoner dilemma game, there can be multiple equilibria, some of which involve strategies of contingent cooperation. Contingent cooperation can then be conventional even though the players do not have approximately the same preferences for all outcomes (see Verbeek, 2008: pp. 77–80). Iterated 'hawk/dove' games may produce conventions even though players do not always prefer others to conform (see Bicchieri, 1993; Sugden, 1986). The definition still excludes behavioral regularities (1) that do not arise from situations of strategic interaction (e.g., breathing and sleeping); and (2) that arise in games with only one equilibrium (e.g., a one-shot prisoner's dilemma).

Additions to Lewis' Definition

Marmor (1996) agrees that Lewis has correctly identified solutions to recurrent coordination problems as conventions, but denies that only these solutions are conventions. Marmor (p. 366) would add a second category consisting of rules 'constituting autonomous practices,' such as the rules of etiquette; the rules of chess, football, or patience; or the rules determining artistic genres and styles. The rules of games like chess and football are not apt examples; however, as they did evolve slowly in what plausibly was an effort of individuals to coordinate on creating a interesting game, rewarding to players and in some cases spectators, and to avoid the coordination failure where teams or individuals from different regions play 'the game' by different rules and therefore cannot fairly play each other. Some sociological explanations of etiquette are also consistent with Lewis as they involve a class of individuals coordinating on complex but arbitrary behavioral rules in order to distinguish themselves from a 'lower' class that fails to master the rules. The question remains open whether all of Marmor's rules constituting autonomous practices may be assimilated into Lewis' framework.

How Conventions Emerge

The simplest way to produce the coordinated expectations necessary for a convention is by *agreement*. When the parties have sufficiently (not necessarily purely) common interests, there is no need to enforce the agreement. If C and R desire to meet, the agreement to meet at P1 may give each a reason to expect the other to go to P1. Neither has a reason to defect from the agreement. But the game theory account of convention purports to explain their emergence without agreement. Some conventions – language and other forms of spontaneous order – seem to require such an origin. How could people form coordinated expectations without agreement?

Much of game theory addresses how players reach equilibrium (*see* Rational Choice Explanation: Philosophical Aspects; Game Theory: Noncooperative Games). The issue is complex and unsettled. It is often quite difficult to demonstrate that players will reach any equilibrium using the common assumption that players are perfectly rational. In a single instance of a coordination game, for example, there is no purely rational basis for expecting a player to select one particular equilibrium outcome. Game theorists believe, however, that players might coordinate on a 'salient' equilibrium or, if the game is recurrent, that players' strategies might evolve to an equilibrium. Consider each idea in turn.

First, empirical data suggests that people can solve coordination problems without agreement. Schelling (1960: pp. 54–58) experimented with such problems and found that the psychological 'prominence' of certain solutions provided the players 'focal points' for coordinating. For example, when the game was selecting a single time of day to meet another, most players selected noon; when asked to select the same positive number, two-fifths selected 1; and when asked to pick a meeting spot on a map, most players selected the only unique feature on the map. Apparently, people sometimes solve coordination problems by selecting what each believes the other is most likely to perceive as unique. Lewis (1969: pp. 35–36) calls this psychological feature salience. As he observes, what is salient need not be the most desirable, but may even be the least desirable equilibrium (but is still more desirable for each than noncoordination).

The empirical effectiveness of salience seems sufficient to demonstrate that conventions can arise without explicit or implicit agreement. But salience itself is difficult to analyze. Salience does not seem to work by rationality alone. Knowing that one equilibrium strategy 'sticks out' from the rest does not justify expecting other players to select it unless one knows that those players favor salient outcomes, which is not required by rationality. Perhaps there is a common metaconvention to pick the salient option, given the frequency of coordination problems. In addition, rationality alone does not explain what is salient, that is, it does not explain how to select among various dimensions by which to evaluate uniqueness. There are many dimensions from which to evaluate whether or which past incident 'sticks out' as being like the present one (see Sugden, 1998a; Gilbert, 1989).

For recurrent games, theorists increasingly offer evolutionary accounts to explain the emergence of an equilibrium. The evolution in these models typically does not occur through greater sexual reproduction of players who enjoy greater success in the recurrent game. Rather, evolution occurs by players, often imperfectly rational, *learning* from experience in the game and switching to strategies that provide greater expected utility (see Skryms, 1996). Evaluating the success of these models is beyond the scope of this article. Some argue, however, that the same problem of induction that affects salience also infects evolutionary models, to the extent they require players to detect the patterns that constitute a game and other players' strategies in that game (see Sugden, 1998a).

Defining 'Norm'

As noted above, a norm can be a mere behavioral regularity among individuals, as the fact that most people in a population drink coffee or tea nearly every morning. In moral theory, a norm can be a prescriptive rule or reason for action. Thus, one could sensibly assert the existence of a moral norm obligating certain behavior even though no one actually adheres to the norm. In a third sense, a norm is a behavioral regularity, but not 'merely' one. It is a regularity sustained in part by the fact that individuals generally approve

(and otherwise reward) conformity and/or disapprove (and otherwise punish) nonconformity. Norms are, therefore, perceived by individuals in the relevant population as obligatory, a regularity one 'ought' to follow, whether or not they are justified in moral theory. The rest of this article discusses this third sense of the term, given its importance to the social sciences.

The definition of norm (in the sense indicated) entails at least three elements (see Pettit, 1990). First, it is a behavioral regularity generally followed by the individuals in a population. This requirement distinguishes a society's mere aspirations for behavior from the norms that its members actually follow. One might want to expand the definition, however, to include cases where most do not follow the regularity but most *either* follow the regularity *or* bear significant costs to conceal their nonconformity. For example, one could coherently say there is a norm against adultery in a community even though only a minority of its members refrain from adultery, because the remaining members incur nontrivial costs to prevent others from discovering their adultery.

Second, a norm is a regularity in which conformity is approved and/or nonconformity is disapproved. On a broad view, the approval and disapproval need not be morally justified or even based on principle but may include the simple attitudes of being positively or negatively inclined toward – liking or disliking – the behavior. Pettit (1990: p. 731) suggests a third requirement: that the pattern of approval and disapproval 'helps to ensure' the existence of the regularity. The causal contribution may be unconscious and contingent, in that other considerations may usually suffice to sustain the regularity but the (dis)approval pattern fills in when those considerations fail. One may also follow Pettit (1990: p. 751) in requiring some kind of common knowledge of the existence of the above three elements, although, for the reasons discussed above concerning conventions, one might imagine that regularities recognizable as norms arise across societies despite the failure to create common knowledge of all three elements.

Some theorists conceive of conventions as a subset of norms (see Marmor, 2006: p. 349), but on the view stated here, the two concepts are distinct (although in practice there is overlap). Conventions do not require approval of conformity and/or disapproval of nonconformity, while norms do not need to be arbitrary in the sense of there existing an alternative means of coordinating.

The Origin of Informal Norms

Legal norms are formal; they are promulgated and enforced by the centralized authority of the state. Nongovernmental organizations also formally promulgate and enforce norms. But some norms arise and are enforced informally and without centralized authority. Informal negative sanctions include censure, gossip, ostracism, and violence. Positive sanctions include praise and the provision of social and economic opportunities. Beyond external sanctions, one may approve or disapprove one's own conformity or nonconformity, respectively, and experience pride or guilt. Informal norms are important because they contribute to spontaneous order, which has implications for, among other things, political theory.

A key question is how informal norms first arise. What causes the behavioral regularity and what causes people to embrace it as obligatory? Some theorists emphasize nonrational forces and, indeed, view norm-based explanations of behavior as an alternative to rational choice explanations. As implied above, norms can be 'internalized' so that one feels guilt or anxiety from violating a norm, and pride or self-satisfaction from complying, even if no one else is aware of one's behavior. Elster (1989: pp. 97–151) emphasizes this emotional feature of norms. Internalization, however, seems a poor candidate for explaining norm origin. However it occurs, it seems more likely to follow than precede the approval of conformity and/or disapproval of nonconformity.

By contrast, rational choice theorists have offered two nonexclusive ways of explaining norm origin (see Pettit, 1990). They differ by whether one explains the behavioral regularity as arising before or after the pattern of approval and disapproval.

Deriving Norms from Regularities

The common way to derive a norm from interest is to explain first how a behavioral regularity arises and second how people come to approve conformity to the regularity and/or disapprove nonconformity. Game theory addresses the first issue – regularities – by its study of the mechanisms by which equilibrium is achieved. The (dis)approval patterns may subsequently arise for various reasons. Lewis (1969: pp. 97–100), for example, notes that conditions of a convention may easily give rise to approval of conformity. Even if conventions need not include his requirement that most individuals conditionally prefer that *others* conform (discussed above), it will often be the case that the condition holds. For example, one driving on the conventional side of the road desires that others follow the convention. In this circumstance, it is easy to imagine that self-interested individuals begin egoistically to *approve* of others conforming to the convention, and to resent nonconformity, given the danger they each face from nonconformity (see McAdams, 2010; Sugden, 1998b). Similarly, if contingent cooperation is a convention in an iterated prisoner's dilemma game, one will not only expect but also prefer that another engages in the conventional cooperation (see Bicchieri, 1993).

A consequence of this analysis is that a regularity might first be a convention, become both a convention and a norm, and end up as a norm only. Before the convention is a norm, people expect others to behave in conformity to the regularity, but do not yet disapprove of them for nonconformity. After the regularity is a norm for some time, it might cease to be arbitrary in the sense described above. The incentives created by the norm – particularly if it is ultimately internalized – might change the payoffs so that the only equilibrium outcome was conformity with the norm regularity.

Deriving Norms from Patterns of Approval and Disapproval

A less common means of deriving norms from interest is to explain first how a pattern of approval and disapproval arises and second how this produces a behavioral regularity. Various

mechanisms might produce the attitudinal pattern, the most obvious being the egoistic effect just discussed. Despite there being no regularity at this stage, it is easy to imagine that one person disapproves (approves) of another for conduct that harms (benefits) the former. Thus, prior to a regularity emerging, an individual may negatively regard a neighbor who makes loud and disturbing noises and positively regard those who clean up common areas. The same result could occur through moralistic rather than egoistic reasoning, if individuals each tend to disapprove (approve) of behavior they believe generally harms (benefits) others (see McAdams, 2010).

These patterns of (dis)approval can create behavioral patterns if individuals sufficiently value *esteem*, i.e., the approval and good opinion of others, as an end. In some circumstances, individuals may costlessly acquire information about the behavior of others, from which they reflexively form an esteem judgment of approval or disapproval. When the judgments are automatic, like being unthinkingly repulsed or attracted, the costs of making such judgments fail to prevent them. If so, the allocation of esteem creates an effective means of punishing or rewarding others. The desire for approval and the fear of disapproval may cause individuals to conform to a known pattern of esteem judgments (see Pettit, 1990; McAdams, 1997).

Elster (1989: p. 133) observes that *expressing* disapproval is always costly, as it takes time and risks the retaliation of the target. One might add that selectively expressing approval risks retaliation by those not approved. But these costs need not prevent esteem-judgments from operating as a sanction and producing a norm. First, judgments of disapproval need not be expressed to be perceived. Introspection may suffice: if *A* hates having his or her sleep disturbed by loud noise, *A* may simply guess that others will disapprove of him or her for waking them with a loud noise (see Pettit, 1990: p. 740). Second, there may be a cost to concealing one's esteem judgment that offsets any costs of expression. If *B* prefers to socialize only with people he or she esteems, then inviting to a social gathering people *B* does not esteem (or forgoing the gathering altogether) may be more costly than revealing the negative esteem judgments. *B* therefore rationally invites only those whom he or she esteems and, despite the costs, *indirectly* expresses negative judgments about others. As to both points, if individuals value esteem, they will seek evidence of how to get it, so even a trivial clue may cause people to form beliefs about the judgments being made. Once individuals perceive a pattern of approval and disapproval, they may gain (net) approval by directly expressing views consistent with the pattern.

Because individuals will tend to approve commonly beneficial behavior more than commonly harmful behavior, the discussion implies that norms derived from (dis)approval patterns will tend to be socially beneficial. A more detailed analysis would reveal several offsetting effects. However, norms are derived, they can cause harm in a variety of ways, particularly by allowing one group to benefit itself at the expense of another (see Hardin, 1995; Ullman-Margalit, 1977: pp. 134–197).

See also: Economics and Ethics; Evolution, Natural and Social: Philosophical Aspects.

Bibliography

Bicchieri, C., 1993. Rationality and Coordination. Cambridge University Press, Cambridge, UK.

Burge, T., 1975. On knowledge and convention. Philosophical Review 84, 249–255.

Ellickson, R.C., 1991. Order without Law: How Neighbors Settle Disputes. Harvard University Press, Cambridge, MA.

Elster, J., 1989. The Cement of Society: A Study of Social Order. Cambridge University Press, Cambridge, UK.

Gilbert, M., 1989. On Social Facts. Routledge, London.

Gilbert, M., 1990. Rationality, coordination, and convention. Synthese 84, 1–21.

Gilbert, M., 2008. Social convention revisited. Topoi 27, 5–16.

Hardin, R., 1995. One for All: the Logic of Group Conflict. Princeton University Press, Princeton, NJ.

Hume, D., 1978[1740]. A Treatise of Human Nature. Oxford University Press, Oxford, UK.

Kornhauser, L.A., 1996. Conceptions of social rule. In: Draybrooke, D. (Ed.), Social Rules: Origin; Character; Logic; Change. Westview Press, Boulder, CO.

Lewis, D.K., 1969. Convention: A Philosophical Study. Harvard University Press, Cambridge, MA.

Marmor, A., 1996. On convention. Synthese 107, 349–371.

McAdams, R.H., 1997. The origin, development, and regulations of norms. Michigan Law Review 96, 338–433.

McAdams, R.H., 2010. Resentment, excuse, and norms. In: Cane, P. (Ed.), The Hart-Fuller Debate, 50 Years on. Hart Publishing, Oxford, UK.

Miller, S.R., 1986. Conventions, interdependence of action, and collective ends. Nous 20, 177–240.

Pettit, P., 1990. Virtus normativa: rational choice perspectives. Ethics 100, 725–755.

Quine, W.V., 1967[1936]. Truth by convention. In: Lee, O.H. (Ed.), Philosophical Essays for Alfred North Whitehead. Russel & Russel, New York.

Schelling, T.C., 1960. The Strategy of Conflict. Harvard University Press, Cambridge, MA.

Skryms, B., 1996. Evolution of the Social Contract. Cambridge University Press, Cambridge, UK.

Sugden, R., 1986. The Economics of Rights, Co-Operation and Welfare. B. Blackwell, Oxford, UK.

Sugden, R., 1998a. The role of inductive reasoning in the evolution of conventions. Law and Philosophy 17, 377–410.

Sugden, R., 1998b. Normative expectations: the simultaneous evolution of institutions and norms. In: Ben-Ner, A., Putterman, L. (Eds.), Economics, Values, and Organization. Cambridge University Press, Cambridge, UK.

Taylor, M., 1987. The Possibility of Cooperation. Cambridge University Press, Cambridge, UK.

Ullmann-Margalit, E., 1977. The Emergence of Norms. Clarendon Press, Oxford, UK.

Vanderschraaf, P., 1995. Convention as correlated equilibrium. Erkenntnis 42 (1), 65–87.

Vanderschraaf, P., 1998a. The informal game theory in Hume's account of convention. Economy and Philosophy 14, 215–247.

Vanderschraaf, P., 1998b. Knowledge, equilibrium and convention. Erkenntnis 49, 337–369.

Vanderschraaf, P., 2001. Learning and Coordination: Inductive Deliberation, Equilibrium and Convention. Routledge, New York, NY.

Verbeek, B., 2008. Conventions and moral norms: the legacy of Lewis. Topoi 27, 73–86.

Conversation Analysis: Sociological

Jack Sidnell, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Conversation analysis (CA) is an approach to social interaction, which describes the stable structures and organizations of practice through which persons accomplish ordinary activities. Through a consideration of a single fragment of conversation, this article introduces some of the main methods, findings, and theoretical assumptions of CA focusing on the problems of turn-taking, action ascription, and repair.

Introduction

In the late 1950s and early 1960s, two great pioneers of the social and behavioral sciences – Erving Goffman and Harold Garfinkel – identified significant gaps in the study of human social life. Goffman, who had been trained by sociologists and anthropologists in the methods of ethnography, wrote of the 'neglected situation' – that substratum of copresent interaction that runs underneath and provides the foundation for all social life (see Goffman, 1964, 1967). Garfinkel, a student of Talcott Parsons who had been inspired by the writings of Alfred Schutz and other phenomenologists, noted that social order is the product of members' methodic procedures for producing and recognizing such order and thus that social life is fundamentally a matter of practical reasoning (see Garfinkel, 1967).

These insights provided the basis for an approach to social interaction known as conversation analysis (CA). The founders of CA, Harvey Sacks and Emanuel Schegloff, studied with Goffman at Berkeley and Sacks later worked with Garfinkel on a research project that examined calls to a suicide prevention hotline. Together, Sacks and Schegloff, along with Gail Jefferson, developed the methods of CA in such a way as to allow for the systematic study of the practical reasoning and endogenous methods that members of society use in order to engage in interaction with one another. Although CA is concerned with interaction in general, language figures centrally in human social life and as a result CA is often primarily concerned with talk-in-interaction.

In this article, the author briefly sketches some of the main ideas that guide current research in the field. A reader who wants to go further with study in this area is encouraged to seek out Sidnell and Stivers *The Handbook of Conversation Analysis* (Sidnell and Stivers, 2012) or Sidnell, *Conversation Analysis: An Introduction* (Sidnell, 2010).

Constructing Turns and Identifying Actions

In speaking, participants are doing things – they are, for example, greeting, requesting, inviting, offering, complaining, and so on. A basic task for the recipient of some bit of talk-in-interaction is, then, to identify what some speaker is doing – what action they mean to accomplish – in speaking in the way that they do. For instance, if one person says to another, "How great is that?" are they requesting information or complimenting the recipient? (See Sidnell, forthcoming.) Or, if they say, "it's raining again" are they suggesting that the recipient should take an umbrella if going for a walk, complaining about the state of the weather, or requesting that someone bring in the laundry? (The issues here are actually much more complex than is suggested by this example and well beyond the scope of the present article. For a more developed argument about the ontology of action in interaction, see Sidnell and Enfield, forthcoming.) Even simple examples such as these make it clear that, in interaction, participants engage in complex inferential process in order to get from what is said to what is meant.

This, then, is just the kind of practical reasoning to which Garfinkel drew attention in his studies. But an arguably more basic aspect of practical reasoning is involved in identifying the particular bits of talk – the utterances – to which actions are linked, whether by speakers in producing them or recipients in recognizing them, in the first place. That is, a recipient in interaction must determine where a given utterance begins and, perhaps more problematically, where it ends. In a groundbreaking study, Sacks et al. (1974) described a 'simplest systematics' that would allow not only for the identification of utterances – or what they refer to as 'turns-at-talk' – but also for the orderly transition between them. An important feature of the system is that it operates independently of whatever particular action(s) the participants are engaged in – that is, whether the participants are greeting, requesting, inviting, offering, complaining, or whatever else they do so in turns-at-talk that have beginnings, middles and possible completions. For this reason, Sacks et al. (1974) argue that the turn-taking system for conversation is 'context independent' since it operates in the same way regardless of what actions are being performed, how the participants are related to one another, where the interaction takes place, and so on. At the same time, they go on to suggest that it is capable of extreme context sensitivity since the details of any particular turn-at-talk reflect in a multitude of ways just these and other contextual parameters. A basic but crucial idea is, then, that there are different levels or orders of organization. There is, for instance, one order pertaining to the organization of talk as action. There is another order related to the organization of talk as turns or, put another way, opportunities to participate in interaction and so on.

Let us consider a single example now for what it can reveal about members' methods for identifying utterances (turns at talk) and actions. The following is an utterance drawn from

a telephone conversation between two friends, Debbie (the speaker) and Shelley (the recipient).

"It's not 'cause Mark's not going."

Here the utterance is written out using (broadly) standard English orthography. We can see that this is a declaratively formatted sentence in which Debbie is speaking about the causal relationship between two events. Specifically 'it' (event 1) is apparently not caused by Mark not going (event 2). This describes (however vaguely) the propositional content of Debbie's utterance. Now, if we are to understand how it is that Shelly recognizes this as a turn-at-talk and as an action-in-interaction we will need to consider some of the ways in which it is produced (its composition) and the particular place in an unfolding conversation that it occurs (its position). Here then is the same utterance shown as part of the conversation in which it was produced.

Examples are presented using the transcription conventions originally developed by Gail Jefferson. For present purposes, the most important symbols are the period ('.'), which indicates falling and final intonation, the question mark ('?') indicating rising intonation, and brackets ('[' and ']') marking the onset and resolution of overlapping talk between two speakers. Equal signs, which come in pairs – one at the end of a line and another at the start of the next line or one shortly thereafter – are used to indicate that the second line followed the first with no discernable silence between them, that is, it was 'latched' to it. Numbers in parentheses (e.g., (0.5)) indicate silence, represented in tenths of a second. Finally, colons are used to indicate prolongation or stretching of the sound preceding them, the more colons, the longer the stretching. (For an explanation of other symbols, see Sacks et al., 1974; Sidnell, 2010.)

1. Debbie and Shelley

```
12  Shelley:  distric attorneys office.
13   Debbie:  Shelley:¿
14  Shelley:  Debbie¿=
15   Debbie:  ↑what is tha dea::l.
16  Shelley:  whadayou ↑mean.
17   Debbie:  yuh not gonna go:.?
18            (0.2)
19  Shelley:  well -hh now: my boss wants
20            me to go: an: uhm finish this
21            >stupid< trial thing,u[hm
22   Debbie:->                      [<its not
23            causeuh:m (0.5) Mark's not going.
24  Shelley:  no- well that wuz initially and
25            then I'm like no: I'll just go
26            and then uhm yaknow this- this
27            tow bandit (·) thing that I have,
28            that were doing, [he w]a:nts
```

We can now see that the utterance in question is actually produced with significant delay between 'cause' and 'Mark's,' This delay is first manifested as the vocalization 'uhm' and then as silence. What Shelley initially hears, then, is something like:

"It's not cause uhm"

Consider this in relation to the question of how it is that participants are able to recognize turns at talk in conversation. One possible solution to that problem might be that they simply wait until the speaker is finished speaking. But if that were all there was to it, Shelley might have heard "It's not cause uhm" as Debbie's complete utterance. This silence is pretty substantial in conversational terms and thus if Shelly were simply waiting for a break in Debbie's talk as an indication that she was finished speaking she might have identified this as the completion of the turn. But of course it is not – we know that and Shelley apparently does too. The question is how do we know that? Perhaps the 'uhm' tells Shelley that, although Debbie will pause the production of her talk, she has not actually completed the turn. There is something to that suggestion and we will, below, consider in a bit more detail just what 'uhm' is doing here. But for now, let us note that it cannot be the whole story. After all, in the prior talk (lines 19–20) Shelley was herself in the course of producing an 'uhm' at just the point when Debbie began speaking. So 'uhm' cannot be an unambiguous signal that there is more to come, or at least, if it is, it is not one that prevents a next speaker from starting their own turn. Rather, there must be something about the talk here (perhaps in combination with the 'uhm') that tells Shelley that Debbie has not completed her turn when she says, "It's not cause uhm." We as observers and as competent conversationalists can hear, I think, that this is not complete – it is not a complete sentence and it is not a complete turn in this context. Specifying how we are able to hear that it is not complete was one of the goals of the paper by Sacks et al. (1974) on turn-taking.

Essentially, these authors proposed that turns-at-talk are composed of what they called 'turn constructional units' or TCUs and that a current speaker is initially entitled (and obligated) to produce one such unit. They further suggested that, in English, there are four types of TCUs and that these can be described in grammatical terms as lexical, phrasal, clausal, and sentential units. It is not that any such unit will work in any context (obviously this is not true). Rather, the proposal appears to be that speakers select some particular unit type and that recipients listen to the talk in order to locate such units (using the analogy of linguistics and computing we could say that they 'parse' the talk into such units). According to these authors, these units have a feature of projectability that essentially means that, having heard the start of a unit, a recipient can anticipate – project – what it will take to reach a first point of possible completion. Participants in interaction apparently do this more or less automatically and, although they are not infallible – sometimes misprojecting the completion of a unit – the degree of accuracy and precision they are able to achieve in ordinary conversation is remarkable. Any strip of conversation will reveal split-second timing of transitions between speakers, suggesting that recipients are able to accurately predict where a unit will reach completion and are prepared to begin their own talk at just such points. Thus, if we consider the first few lines of the conversation between Debbie and Shelley, we can see that the transitions between speakers are achieved with minimal gap and overlap:

2. Debbie and Shelley

```
12  Shelley:  distric attorneys office.
13   Debbie:  Shelley:¿
14  Shelley:  Debbie¿=
15   Debbie:  ↑what is tha dea::l.
16  Shelley:  whadayou ↑mean.
17   Debbie:  yuh not gonna go:.?
```

We can now perhaps understand a bit better what is happening in lines 22 and 23. Here, Shelley can presumably hear from "it's not cause" that Debbie's turn is in progress and that it has not reached completion. Thus, the silence that emerges between 'uhm' and 'Mark's' is not a place for Shelley to take a turn. Rather, something else is presumably happening here, just what we will consider below.

We have briefly considered how it is that Shelley might be able to identify the boundaries of the utterance here – how she might, that is, recognize in this talk a TCU and how that has implications for the timing of her own response. Let us now consider what is involved in recognizing the action here. We can start by noting as we did earlier that this is a declaratively formatted sentence in which Debbie is speaking about the causal relationship between two events. Looking back a bit, we can see that Debbie and Shelley are talking about why Shelley is not going (specifically not going on a trip that the two had planned as is revealed later on in the conversation). So Debbie offers "Mark's not going" as not the reason Shelley is not going. If the logic of this seems somewhat complex, even opaque, that is probably relevant to how it is understood by Shelley and thus how she recognizes what action is being done. But let us start with some rather more simple observations.

First, we can see that, although the sentence is declaratively formatted, with it, Debbie appears to be asking a question rather than making an assertion. In order to understand how it does this, we have to think about this utterance in relation to a presumed distribution of knowledge between the participants. So here Debbie is talking about Shelley's reasons for not going on a trip. Shelley's reasons for not going on the trip (her reasons for anything) are something that she knows about better than Debbie, indeed, better than anyone. Her reasons for doing things fall, we can say, squarely within her 'domain of knowledge.' Thus, when Debbie says this, she is not heard as asserting but rather as questioning or requesting confirmation. So we can see from this example that one of the basic issues confronting the recipient of an utterance such as this is whether the action being done is asserting or questioning, telling or asking. And you can see from the examples below that when a speaker uses a declaratively formatted utterance to talk about something that the recipient knows better than does the speaker, the speaker is routinely and unproblematically heard as asking rather than telling (see Heritage, 2012).

3. Coffee chat – 11, 8:10

```
25 Rich:    You been in Colvina:,
26          (0.6)
27 Betty:   Ye:ahs, that's right. That's right.
```

4. Coffee chat – 18, 12:25

```
01 Tom:     =Oh yer [on the first floor again?
02 Hank:            [Yeh.
03          YEAH.
```

5. TC_II_(b):_#28

```
16 Bush:    He's down in Mexico or some'in?,
17 Pyatt:   I don't know,
```

6. SBL-1-10

```
10 Rose:    Ah'it left'er (0.4) quite
11          permanently damaged °I suppose°
12 Bea:     Uh:pparently,
```

In examples 3 and 4, the speaker is talking about the recipient. "You been in Colvina" and "You're on the first floor again?" are heard as questions that request confirmation and in both cases confirmation is what they get in the next turn. In examples 5 and 6, the speaker is talking about a third party (neither speaker nor recipient) but, again in both cases, the declaratively formatted utterance is treated as a question by the response. In example 5, the recipient, Pyatt, responds with "I don't know" thereby accounting for not answering the question as put (i.e., he cannot answer because he does not have adequate knowledge to do so). In example 6, two nurses are talking about an accident suffered by a patient. Bea has more recent and detailed knowledge here and when Rose says "it left her quite permanently damaged" she calls attention to her lack of information by appending "I suppose" (which marks what she has said here as an inference). Bea then tentatively confirms what Rose has suggested with 'apparently.'

So we can see a robust phenomenon at work here: where the speaker's declaratively formatted utterance pertains to matters about which the recipient knows better (or about which the recipient has greater 'epistemic' rights to talk about), that utterance is heard as a question, specifically a request for confirmation. Labov (1970) described this in terms of A events, B events, and AB events, where A events are those known by A, B events those known by B, and AB events those known by both. He writes, "If A makes a statement about a B-event, it is heard as a request for confirmation."

The discussion so far has drawn on some of the basic methods of CA. Let us try to make these explicit. Notice from the outset that we are concerned with members' methods of practical reasoning – here, practical reasoning about interaction and specifically what a given utterance is doing. In order to study this, we need some access if not to their methods directly then to the results of those methods. And this is how we use subsequent talk. That is, if we want to see how an utterance in any of these examples was understood – whether in terms of what action it was doing, or in terms of where it reached completion – we can examine the response that it elicits. Participants try to make their contributions fitted to what has come before – and most importantly immediately before – and thus any contribution will, at least tacitly, convey an understanding of prior talk. In the case of a question and its response, this is particularly obvious. Thus, when Betty says "Ye:ahs, that's right. that's right," she can be seen to be answering a (yes–no) question. This response thus reveals to both us and to the other participants how the prior talk was understood – it reveals, that is, an analysis of that talk. It provides us, analysts, with data-internal evidence to support the argument that these utterances are hearable as questions despite the fact that they are formatted as declaratives.

Now we can take this a bit further and doing so will reveal something crucially important about the organization of action. So if we go back to the example we began with, we can

see that Shelley does answer the question but that is not all she does:

7. Debbie and Shelley

```
24  Shelley:  no- well that wuz initially and
25            then I'm like no: I'll just go
26            and then uhm yaknow this- this
27            tow bandit (·) thing that I have,
28            that were doing, [he w]a:nts
```

With 'no,' Shelley has apparently answered the question by confirming what Debbie has proposed (the reason she is not going is not because Mark is not going). But notice that before this 'no' is even brought to completion, it is cut off and Shelley continues. In what she says after the 'no,' she goes beyond simply answering the question. Here, she tells Debbie in some detail about her reasons for not going. According to Shelley, the fact that Mark was not going did initially influence her decision. However, according to what she says here, although she decided not to go at first, she then changed her mind ("I'll just go"). It was only after that initial decision had been made and changed that she again changed her mind because of something happening at work. So here Shelley admits, in response to Debbie's question, that, yes, Mark's not going did affect her own decision but goes on to propose that ultimately this was not the reason why she decided not to go.

The point here then is that there are other layers of action besides simply 'answering.' Here, for instance, Shelley can appropriately be described as first admitting and subsequently explaining or perhaps excusing herself. She excuses herself by offering an account (or reason) for her decision in which she is characterized as not-at-fault simply because she had no choice. Rather, her boss required that she work that weekend thereby making it impossible for her to go on the trip she had planned with Debbie.

An admission or an excuse is a kind of responsive action that is produced in response to a particular kind of initiating action. If, as we have proposed, responses reveal their speaker's analysis or understanding of the turn to which they respond, Shelley must have heard "it's not cause Mark's not going" as an accusation or complaint or some other such action worthy of an excuse in response.

8. Debbie and Shelly

```
12  Shelley:  distric attorneys office.
13   Debbie:  Shelley:¿
14  Shelley:  Debbie¿=
15   Debbie:  ↑what is tha dea::l.
16  Shelley:  whadayou ↑mean.
17   Debbie:  yuh not gonna go::?
18            (0.2)
19  Shelley:  well -hh now: my boss wants
20            me to go: an: uhm finish this
21            >stupid< trial thing,u[hm
22   Debbie:->                      [<its not
23            causeuh:m (0.5) Mark's not going.
24  Shelley:  no- well that wuz initially and
25            then I'm like no: I'll just go
26            and then uhm yaknow this- this
27            tow bandit (·) thing that I have,
28            that were doing, [he w]a:nts
```

In our discussion of how it is that participants in interaction come to recognize a particular utterance as some specific action, we have focused primarily on the composition of the talk. For instance, we noted that declaratively formatted utterances that pertain to matters within the recipient's domain of knowledge (or epistemic domain) are routinely heard as questions. In order to understand how Shelley can find in "It's not cause Mark's not going" an accusation or complaint, we need to consider not only the composition of the turn but also its position within an unfolding course of talk and sequence of turns.

Notice then that this turn comes very near the beginning of the conversation between Debbie and Shelley. Debbie has called Shelley at work and after speaking to a receptionist has reached her at line 12 with Shelley self-identifying with "District attorney's office." With this, Shelley shows she is expecting a work-related call and is prepared to interact according to her work-related identity. However, the turns that follow – a reciprocal exchange of names – immediately resets the call as conversational rather than institutional. At line 15, the conversationalists have reached anchor position – a place for the caller to announce or to otherwise indicate what it is she is calling about. Debbie does this in a very particular way. With "what is the deal," Debbie is clearly not only asking a question but also more than that. The positioning is abrupt and the question itself obviously presumes that Shelley knows already what it is Debbie might be calling about. More specifically by referring to 'the deal,' Debbie suggests that there is a known-in-common matter that might constitute 'a deal,' which is to say a problem.

After Shelley initiates repair with "what do you mean?," thereby suggesting she does not know what it is Debbie is talking about, Debbie asks a revised version of the question, "you're not going to go?" Compared to an interrogative format such as "Are you going to go?," this one suggests that Debbie is relatively certain of the answer. With "you're not going to go?," Debbie reveals the problem that she indexed with 'the deal.' Specifically, Debbie shows that she already knows Shelley plans not to go on the trip. Now, in response to the question "you're not going to go?," Shelley does not answer specifically with yes or no, but rather provides an account for why she will not go. This account assigns responsibility for Shelley's decision to her boss. Notice also that Shelley refers to the work-related matter as this "stupid trial thing." In this way, Shelly assesses the work-related matter negatively ('stupid' and 'thing') and characterizes it as an inconvenience to her.

It is in this context that Debbie produces the question, "it's not 'cause Mark's not going?" We can see quite plainly that Debbie's question here asks about something that Shelley's already talked about. Specifically, Debbie suggests an alternate and competing explanation for why it is that Shelley is not going to go. This explanation, the one that she offers, assigns responsibility for the decision squarely to Shelley.

So we can see that Shelley uses both the composition and position of this utterance to find in it complaint or even accusation. It is to this action that she responds with an excuse. So importantly, in interaction, actions are multiple and layered. A single utterance can be understood in a number of different ways. Here Debbie's question "it's not because Mark's not going?" is a question, a complaint, and

perhaps also hearable as an accusation that Shelley is not being completely honest in the immediately prior talk and so on.

In this section, we have also seen something of the method of CA. Specifically, we have seen that if we want to propose that a given utterance functions as a question we must look to see how a recipient responded to it. There is obviously more to it than that but this is nevertheless crucial. We can use the participants' own conduct to ground an analysis of how they understood prior talk. Thus, the evidence we use in CA is internal to the data itself and is evidence that the participants themselves have access to. Now, of course, as analysts we draw on evidence afforded by multiple occasions involving different participants in order to see how the structures of conversation are generic, cohort independent, and useable in different contexts. But our first line of argument draws upon the data internal evidence that participants themselves produce in the course of face-to-face interaction (for a more detailed account of the method, see Sidnell, 2012).

Word Searching and Repair

So far, we have seen that the utterance "it's not because Mark's not going" can be seen as a question and as a possible complaint or even accusation. If we now look at the utterance in greater detail, we will see that it can also be understood in a quite different way.

<its not causeuh:m (0.5) Mark's not going.

As we have already noted, Debbie produces this utterance with a substantial pause between 'causeuh:m' and 'Mark's.' What might she be doing here? Consider the following examples:

9. G 86:490

```
19 Mike:            [This guy's name was Blake.
20        (0.4) he was in the movie
21         uh:, (0.6) In Cold Blood
```

10. Auto discussion P.11

```
28  CUR:  Keegan usetuh race uhr uh- er ih
29        was um, (0.4) usetuh run um,
30        (1.0)
31  CUR:  Oh::shit.
32        (0.4)
33  CUR:  Uhm,
34        (0.4)
35  CUR:  car.
36  MIK:  Three en[n a[quarter?
37  CUR:          [Thr[ee enna quarter.
```

11. Berkeley (1:19)

```
18  RIT:   it was ca:lled uh::(hh)
19         (.) the Wayside Inn.
20         (0.9)
21  LES:   the Wayside Inn¿hh=
22  RIT:   =yeah.
```

Across these examples we see a recurrent pattern. This involves the speaker producing a turn at talk up to the point at which they are going to say a name. In each case, the speaker encounters trouble remembering the name and engages in a search for it. So in example 9 Mike says, "he was in the movie" and here begins a search for the name "in cold blood," In 10, Curt is looking for the name 'three and a quarter.' He first encounters trouble at line 28. The trouble continues until Mike provides the name at line 37. And in 11, Rita apparently cannot remember the name 'the Wayside Inn.' Although there are significant differences across these examples, there are also similarities. Specifically in each case, we see a speaker marking the onset of trouble with 'uh:' (9), 'uhr' (10), 'uh::' (11) then delaying, then either producing the name or, as in 10, extending the search by various means until a point at which one of the other participants produces the name. (In these examples, it is a name that causes the problem but this is by no means always the case.)

A basic assumption that participants in interaction appear to make is that next items (e.g., a sound, a word, a sentence, and so on) will forward the action initiated by previous items. In the cases we are examining here, however, the progress of the talk is suspended such that that assumption will not hold. Word searches then are a subtype of repair (see Schegloff et al., 1977; Goodwin and Goodwin, 1986). More specifically, they are a subtype of self-initiated repair. Humans are obviously imperfect. As such, in interaction they encounter various kinds of trouble. Repair refers to an organized set of practices by which participants may identify, locate, and fix those troubles. Other subtypes of self-initiated repair deal with other types of trouble. For instance, often a speaker will say one thing and subsequently replace it with something else. So in the extract from Debbie and Shelley below we see Shelley replacing 'that I have' with 'that were doing:'

```
"this tow bandit (·) thing that I have,
that were doing,"
```

And in the extract from example 10 below, we see Curt replace 'used to race' with 'used to run:'

```
"Keegan usetuh race uhr uh- er ih was um,
(0.4) usetuh ru[n"
```

Such cases of self-initiated self-repair are extremely common in ordinary conversation and other forms of talk (see Hayashi et al., 2013).

Episodes of repair are composed of parts (Schegloff et al., 1977). A repair initiation marks a "possible disjunction with the immediately preceding talk" while a repair outcome results either in a "solution or abandonment of the problem" (Schegloff, 2000: p. 207). The particular segment of talk to which the repair is addressed is called the trouble source. Repair can be initiated either by the speaker of the repairable item or by some other participant. With these distinctions, we can describe any instance of repair in terms of who does what and where. Thus, we have already proposed that the cases above are instances of self-initiated repair. We can now further specify that they are instances of same turn, self-initiated, self-repair. We can compare such examples of same turn, self-initiated, self-repair with examples such as that from Debbie and Shelley below in

which repair is other-initiated (with "whadayou mean.") in the next term and completed in the turn after that.

12. Debbie and Shelly

```
15   Debbie:  ↑what is tha dea::l.
16  Shelley:  whadayou ↑mean.
17   Debbie:  yuh not gonna go:.?
```

Here then, the trouble source is the question "what is the deal." The repair initiation takes the form "what do you mean?" in the next turn and the repair is provided by the speaker of the trouble source, Debbie, with "yuh not gonna go?"

In a classic paper, Schegloff et al. (1977) proposed that all these forms of repair are organized into a single system. That system has a bias, according to these authors, such that self-correction or repair is preferred to any other kind. 'Preference' here refers to a bias in the organization of interaction evidenced by the fact that other-correction and even other-initiation tends to be delayed, hedged, modulated, and only done after the speaker of the trouble source has had an opportunity to repair a problem in their own talk.

Conclusion

In this brief article, the author has attempted to describe some of the main methods and findings of the field of CA. Focusing on a single example, the conversation between Debbie and Shelley, but simultaneously looking across various occasions of interaction, we have seen some of the more robust, stable practices and structures of social interaction such as those involved in taking turns, recognizing actions, and repairing troubles. These practices and structures are the very phenomena that Goffman proposed as an object of sociological study and underlying them are those modes of practical reasoning to which Garfinkel was at pains to our draw attention.

See also: Action, Theories of Social; Garfinkel, Harold (1917–2011); Goffman, Erving (1922–1982); Joint Action.

Bibliography

Garfinkel, H., 1967. Studies in Ethnomethodology. Prentice-Hall, Englewood Cliffs, NJ.

Goffman, E., 1964. The neglected situation. American Anthropologist 66 (6, pt. 2), 133–136.

Goffman, E., 1967. Interaction Ritual: Essays in Face to Face Behavior. Doubleday, Garden City, NY.

Goodwin, M.H., Goodwin, C., 1986. Gesture and coparticipation in the activity of searching for a word. Semiotica 62 (1/2), 51–75.

Hayashi, M., Raymond, G., Sidnell, J. (Eds.), 2013. Conversational Repair and Human Understanding. Cambridge University Press, Cambridge.

Heritage, J., 2012. Epistemics in action: action formation and territories of knowledge. Research on Language and Social Interaction 45, 1–25.

Labov, William, 1970. The study of English in its social context. Stadium Generale 23, 30–87.

Sacks, H., Schegloff, E.A., Jefferson, G., 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50 (4), 696–735.

Schegloff, E.A., Jefferson, G., Sacks, H., 1977. The preference for self-correction in the organization of repair in conversation. Language 53 (2), 361–382.

Sidnell, J., 2010. Conversation Analysis: an Introduction. Wiley-Blackwell, Oxford.

Sidnell, J., 2012. Basic conversation analytic methods. In: Sidnell, J., Stivers, T. (Eds.), Handbook of Conversation Analysis. Wiley-Blackwell, Oxford, pp. 77–100.

Sidnell, J. The architecture of intersubjectivity revisited. In: Enfield, N.J., Kockelman, P., Sidnell, J. (Eds.), The Cambridge Handbook of Linguistic Anthropology. Cambridge University Press, Cambridge, forthcoming.

Sidnell, J., Enfield, N.J. The ontology of action in interaction. In: Enfield, N.J., Kockelman, P., Sidnell, J. (Eds.), The Cambridge Handbook of Linguistic Anthropology. Cambridge University Press, Cambridge, forthcoming.

Sidnell, J., Stivers, T. (Eds.), 2012. Handbook of Conversation Analysis. Wiley-Blackwell, Oxford.

Cooperation and Competition

David W Johnson and Roger T Johnson, University of Minnesota, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The four major theoretical orientations to cooperation and competition are cognitive development, behavioral learning, social cognitive, and social interdependence. Of the four, social interdependence theory has generated the most research and applications. The basic premise of social interdependence theory is that the type of interdependence structured in a situation determines how individuals interact with each other which, in turn, determines outcomes. The hundreds of studies that have been conducted indicate that cooperation, compared with competitive and individualistic efforts, tends to result in higher achievement and productivity, more positive interpersonal relationships, and greater psychological health.

Introduction

During the twentieth century, there have been a series of social forces that have highlighted the importance of cooperation and competition and often created a creative tension between them. In the early 1900s, the advent of the industrial revolution and the ending of colonialism focused attention on the importance of well functioning, cooperative groups. But it also promoted the belief that competition was the secret to increase economic success. During the 1920s and 1930s, with the worldwide depression and the rise of dictatorship in Europe, the importance of cooperation was again highlighted. But at the same time, business and industry groups, such as the Liberty League in the USA, increased their emphasis on competition in business and industry, education, and other areas of life. In the aftermath of the Second World War, there was great concern about the future of democracy and the need to better understand how democratic organizations could be made to function more effectively. The health of a democratic society, it was thought, depended on the effectiveness of its component groups. Strengthening the quality of cooperation in the family, the community, and the multitude of groups within society was viewed as the primary means of ensuring the vitality of democracy and the successful solving of the existing social problems. At the same time, the rise of social Darwinism promoted competition as the natural order of life. In the early 1960s, anthropology and ethnology proposed that competition over weapon systems sparked human evolution. The civil rights movement and concerns about the possibility of nuclear war, on the other hand, created great interest in cooperation among diverse individuals. The tension between and interest in cooperation and competition sparked discussions in three major theoretical orientations: cognitive developmental, behavioral, and social interdependence.

Cognitive Developmental Theory

The 'cognitive developmental perspective' is largely based on the theories of Piaget (1950), Vygotsky (1978), cognitive science, and academic controversy (Johnson and Johnson, 2007, 2009a). To Jean Piaget (1950), cooperation is the striving to attain common goals while coordinating one's own feelings and perspective with a consciousness of others' feelings and perspective. From Piaget and related theories comes the premise that when individuals cooperate on the environment, sociocognitive conflict occurs that creates cognitive disequilibrium, which in turn stimulates perspective-taking ability and cognitive development. Cooperation in the Piagetian tradition is aimed at accelerating a person's intellectual development by forcing him or her to reach consensus with others who hold opposing points of view about the answer to the problem. Lev Semenovich Vygotsky (1978) and related social constructionist theorists claim that our distinctively human mental functions and accomplishments have their origins in our social relationships. Mental functioning is the internalized and transformed version of the accomplishments of a group. Knowledge is social, constructed from cooperative efforts to learn, understand, and solve problems. A central concept is the 'zone of proximal development,' which is the zone between what a person can do on his or her own and what the person can achieve while working under the guidance of older individuals or in collaboration with more capable peers. Unless persons work cooperatively, they will not grow intellectually and the time in which persons work alone should therefore be minimized.

The central aspect of cooperation according to cognitive-developmental theory is conflict among ideas or controversy. 'Constructive controversy theory' (Johnson and Johnson, 2007, 2009a) posits that being confronted with opposing points of view creates disequilibrium, uncertainty, or conceptual conflict, which creates a search for more information and a reconceptualization of the issue, which results in a more refined and thoughtful conclusion. The key steps of constructive controversy are organizing what is known into a position, advocating that position to someone who is advocating an opposing position, attempting to refute the opposing position while rebutting the attacks on one's own position, reversing perspectives so that the issue may be seen from both points of view simultaneously, and creating a synthesis to which all sides can agree.

Behavioral Learning Theory

The 'behavioral learning perspective' assumes that individuals will work hard on those tasks, for which they secure a reward

of some sort and will fail to work on tasks that yield no reward or yield punishment (Bandura, 1977; Skinner, 1968). Cooperative efforts are designed to provide incentives for the members of the group to participate in a group effort since it is assumed that individuals will not intrinsically help their classmates or work toward a common goal. Skinner focused on group contingencies, Bandura focused on imitation, and Homans as well as Thibaut and Kelley focused on the balance of rewards and costs in social exchange among interdependent individuals.

The most developed behavioral theory of cooperation and competition was developed by Kelley and Thibaut (1978). They assumed that individuals act to maximize their self-interests by behaving in ways to maximize their rewards and minimize their punishments or costs. 'Rewards' are defined as positive interaction experiences that yield outcomes that promote physical or emotional well being, self-esteem, or personal growth such as enjoyment, fulfillment, or contentment. 'Costs' are defined as negative interaction experiences that yield physical or emotional pain or failure. People attempt to recreate experienced interactions that yield pleasure and avoid interactions that yield pain. 'Cooperation' is thus defined as acting in ways perceived to maximize joint rewards and minimize joint costs, 'competition' as acting in ways to maximize one's own rewards and minimize one's own costs relative to others, and 'individualistic efforts' as acting in ways to maximize one's own rewards and minimize one's own costs with little or no regard of the outcome for others. In evaluating outcomes, individuals have a 'comparison level' about what constitutes an acceptable outcome and a 'comparison level for alternatives' about how those outcomes compare to other possible outcomes. Cooperation tends to occur when outcomes are correspondent and competition tends to occur when outcomes are noncorrespondent.

Social Cognitive Theory

Social cognitive theory views cooperation as collective agency (Bandura, 2000), the shared belief in the collective power to produce desired results. In collective agency, individuals have to work together to secure what they cannot accomplish on their own. Perceived collective efficacy influences the goals the group strives to achieve, how well they use their resources, how much effort they put into their group endeavor, their staying power when collective efforts fail to produce quick results or meet forcible opposition, and their vulnerability to the discouragement that can beset people taking on tough problems.

From the social cognitive perspective, cooperation involves modeling, coaching, and scaffolding (i.e., conceptual frameworks that provide understanding of what is being learned) (Lave and Wenger, 1991). Ideally, the learner will cognitively rehearse and restructure information for it to be retained in memory and incorporated into existing cognitive structures (Wittrock, 1990). An effective way of doing so is explaining the material being learned to a collaborator. Finally, social cognitive theory places cooperation at the center of a community of practice, a group of people who share a craft or a profession.

Social Interdependence Theory

The roots of social interdependence theory lie in the early 1900s when Kurt Koffka (one of the founders of the Gestalt School of Psychology) proposed that groups were dynamic wholes, in which the interdependence among members could vary. In the 1920s and 1930s, Kurt Lewin refined Koffka's notion, proposing that the essence of a group is the interdependence among members created by common goals, and that goal interdependence results in the group being a 'dynamic whole,' so that a change in the state of any member or subgroup changes the state of any other member or subgroup. Continuing Lewin's work, in 1949, Morton Deutsch (1949, 1962) formulated social interdependence theory.

'Social interdependence' exists when individuals share common goals and each individual's outcomes are affected by the actions of the others (Deutsch, 1962; Johnson and Johnson, 1989). It may be differentiated from social dependence (i.e., the outcomes of one person are affected by the actions of a second person but not vice versa) and 'social independence' (i.e., individuals' outcomes are unaffected by each other's actions). There are two types of social interdependence: positive (cooperative) and negative (competitive).

'Cooperation' exists when individuals work together to accomplish shared goals (Johnson and Johnson, 2009b). When a situation is cooperatively structured, individuals' goal achievements are positively correlated; individuals perceive that they can reach their goals if, and only if, the others in the group also reach their goals (Deutsch, 1962; Johnson and Johnson, 1989). Thus, individuals seek outcomes that are beneficial to all those with whom they are cooperatively linked. 'Competition' exists when individuals work against each other to achieve a goal that only one or a few can attain (Johnson and Johnson, 2009b). When a situation is structured competitively, individuals' goal achievements are negatively correlated; each individual perceives that when one person achieves his or her goal, all others with whom he or she is competitively linked fail to achieve their goals (Deutsch, 1962; Johnson and Johnson, 1989). Thus, individuals seek an outcome that is personally beneficial but detrimental to all others in the situation. 'Individualistic efforts' exist when individuals work by themselves to accomplish goals unrelated to the goals of others (Johnson and Johnson, 2009b). When a situation is structured individualistically, there is no correlation among participants' goal attainments. Each individual perceives that he or she can reach his or her goal regardless of whether other individuals attain or do not attain their goals (Deutsch, 1962, Johnson and Johnson, 1989). Thus, individuals seek an outcome that is personally beneficial without concern for the outcomes of others.

Of the theories of cooperation discussed, social interdependence theory is by far the most important in terms of research generated and practice applications derived. It will, therefore, be discussed in some depth.

Interaction Patterns

The basic premise of social interdependence theory is that the type of interdependence structured in a situation determines how individuals interact with each other which, in turn,

determines outcomes (Deutsch, 1949, 1962; Johnson and Johnson, 1989, 2009b). Positive interdependence tends to result in promotive interaction; negative interdependence tends to result in opposition or contrient interaction, and no interdependence results in an absence of interaction. 'Promotive interaction' occurs when members help and assist each other, exchange resources, give and receive feedback, challenge each other's reasoning, and encourage increased effort. Group members focus on maximizing the success of their groupmates as well as their own. 'Oppositional interaction' occurs as individuals discourage and obstruct each other's efforts to achieve. Individuals focus both on increasing their own success and on preventing anyone else from being more successful than they are. 'No interaction' exists when individuals work independently without any interaction or interchange with each other. Individuals only focus on increasing their own success and ignore, as irrelevant, the efforts of others. The relationship between the type of social interdependence and the interaction pattern it elicits is assumed to be bidirectional. Each may cause the other. Each of these interaction patterns, furthermore, creates different outcomes.

Psychological States

Deutsch (1949, 1962) noted that depending on whether individuals promote or obstruct each other's goal accomplishments, there is 'substitutability' (i.e., the actions of one person substitute for the actions of another), 'cathexis' (i.e., the investment of psychological energy in objects and events outside of oneself), and 'inducibility' (i.e., openness to influence). Essentially, in cooperative situations, the actions of participants substitute for each other, participants positively cathect to each other's effective actions, and there is high inducibility among participants. In competitive situations, the actions of participants do not substitute for each other, participants negatively cathect to each other's effective actions, and inducibility is low. When there is no interaction, there is no substitutability, cathexis, or inducibility.

Outcomes

The investigation of cooperation and competition is the longest standing research tradition within American social psychology. From 1898 to 2014, researchers have conducted over 1200 studies from which effect sizes can be computed on social interdependence and hundreds of more studies that do not provide enough information to compute effect sizes (Johnson and Johnson, 1989, 2009b). Since research participants have varied widely as to gender, economic class, age, and cultural background, since a wide variety of research tasks and measures of the dependent variables have been used, and since the research has been conducted by many different researchers with markedly different orientations working in different settings and in different decades, the overall body of research on social interdependence has considerable generalizability. The numerous outcomes studied may be subsumed within three broad categories (Johnson and Johnson, 1989, 2009b): (1) effort to achieve; (2) positive relationships; and (3) psychological health. In terms of productivity and achievement, a meta-analysis of all available studies (Johnson and Johnson, 1989) found that the average person in a cooperative situation performed at about two-third a standard deviation above the average person learning within a competitive (effect size = 0.67) or individualistic situation (effect size = 0.64). These results held constant when group measures of productivity were included as well as individual measures, for short-term as well as long-term studies, and when symbolic as well as tangible rewards were used.

In terms of quality of relationships, cooperative efforts, compared with competitive and individualistic experiences, promoted considerable more liking among individuals (effect sizes = 0.66 and 0.62, respectively) even when individuals were from different ethnic and historical backgrounds, or were handicapped (Johnson and Johnson, 1989). In addition, cooperative experience promoted greater task-oriented and personal social support than did competitive (effect size = 0.62) or individualistic (effect size = 0.70) efforts.

Finally, the research indicates that (1) working cooperatively with peers and valuing cooperation result in greater psychological health than does competing with peers or working independently and (2) cooperative attitudes are highly correlated with a wide variety of indices of psychological health (Johnson and Johnson, 1989). Competitiveness was in some cases positively and in some cases negatively related to psychological health, and individualistic attitudes were negatively related to a wide variety of indices of psychological health. In addition, cooperative experiences promoted higher self-esteem than did competitive (effect size = 0.58) or individualistic (effect-size = 0.44) experiences. Cooperative experiences also tended to increase perspective-taking ability while competitive and individualistic experiences tended to promote egocentrism (being unaware of other perspectives other than your own) (effect sizes of 0.61 and 0.44, respectively). It is through cooperative efforts that many of the attitudes and values essential to psychological health (such as self-efficacy) are learned and adopted. Social skills and competencies tend to increase more within cooperative than in competitive or individualistic situations (Johnson and Johnson, 1989, 2009b). Thus, there seems to be a strong relationship between cooperativeness and psychological health, a mixed relationship between competitiveness and psychological health, and a strong relationship between an individualism and psychological pathology.

These outcomes of cooperative efforts (effort to achieve, quality of relationships, and psychological health) form a gestalt where each influences the others and they are therefore likely to be found together.

Two meta-analyses were conducted to determine the impact of quality of interpersonal relationships and motivation on achievement. Over eight decades of research on over 17 000 early adolescents from 11 countries found that higher achievement and more positive peer relationships were associated with cooperative rather than competitive or individualistic goal structures (Roseth et al., 2008). More specifically, 33% of the variation in achievement was accounted for by positive peer relationships (40% when the low-quality studies were removed). Johnson et al. (2014) found that the results of 629 independent studies representing 26 different countries indicated that cooperative

goal structures resulted in greater achievement and motivation than did competitive or individualistic goal structures. Compared to competitive and individualistic goal structures, the relative effects of cooperative goal structures on motivation and achievement were positively correlated, accounting for 14% (24% when low-quality studies were removed) of the variance in achievement (and vice versa).

A concern in schools is bullying. Choi et al. (2011) examined the relationship between (1) cooperative, competitive, and individualistic predispositions and experiences and (2) harm-intended aggression (i.e., bullying) and prosocial behavior. The more cooperative the classroom, the greater the students' cooperativeness. Cooperativeness predicted prosocial behaviors and the absence of harm-intended aggression, while competitiveness predicted harm-intended aggression. In other words, engaging in competition and being competitive tends to increase bullying and engaging in cooperative experiences and being cooperative tends to decrease bullying and increase prosocial behavior.

Basic Elements

In order for cooperative efforts to be effective, five elements must be structured in the situation: positive interdependence, individual accountability, promotive interaction, social skills, and group processing. 'Positive interdependence' exists when one perceives that one is linked with others in a way, so that one cannot succeed unless they do (and vice versa) and that one must coordinate one's efforts with the efforts of others to complete a task (Deutsch, 1962; Johnson and Johnson, 1989). Cooperative efforts may be built on outcome interdependence (i.e., goals and rewards), means interdependence (i.e., resource, role, and task interdependence or group members being tied together by complementary resources, interlocking roles, or a division of labor), and environmental interdependence (i.e., interdependence based on shared space, proximity, and boundaries) (Johnson and Johnson, 1989, 1999, 2009b).

'Individual accountability' exists when the performance of each individual person is assessed and the results are given back to the group and the individual. It is important that the group knows who needs more assistance, support, and encouragement in completing the task and group members know that they are responsible for doing their share of the work. 'Promotive interaction' may be defined as individuals encouraging and facilitating each other's efforts to complete tasks and achieve in order to reach the group's goals. Group members help and assist each other, exchange information and resources, and encourage each other to be productive.

'Social skills' are the actions needed to effectively interact with other people. Leadership, decision making, trust building, communication, and conflict management skills are examples. Social skills form the basic nexus among individuals, and if individuals are to work together productively, they must have a modicum of these skills. 'Group processing' may be defined as reflecting on group efforts to (1) describe what member actions were helpful and unhelpful in achieving the group's goals and maintaining effective working relationships among members, and (2) make decisions about what actions to continue or change. The purpose of group processing is to clarify and improve the effectiveness of members in contributing to the joint efforts to achieve the group's goals.

Enhancing Factors

Deutsch (1962, 1973, 1985) has posited that trust, conflict, and distributive justice are mediators of the effectiveness of cooperation. The greater the trust, the more effective the cooperative efforts tend to be. Effective cooperation requires that conflicts of interest are resolved constructively. The effectiveness of cooperation is increased when benefits are distributed based on equality and need and when the distribution of benefits according to equity (i.e., high performers receive more than low performers) is avoided.

Competitive and Individualistic Efforts

Competition tends to result in constructive consequences when it occurs within a broader cooperative context, clear and fair rules, and criteria for winning are present, the task is appropriate, the task may be completed individually, competitors have an equal chance of winning, and winning is of low importance (Johnson and Johnson, 1999; Stanne et al., 1999). The conditions under which individualistic efforts have an advantage over cooperative efforts are unclear.

Applications

Cooperation underlies almost all aspects of human experience. Interpersonal interaction, family life, community life, organizational life, and even life within a society are all cooperative endeavors. Competition can occur and end, but the underlying structure of human existence is cooperation with others. Social interdependence theory has been applied to education, business, individual and group therapy, family therapy, mediation, organizational and community development, international conflict and peace building, and many other areas. The two arenas in which cooperation and competition have been especially applied, however, are education and business and industry.

The application of social interdependence theory to education is one of the most successful and widespread applications of social psychology to practice. From social interdependence theory, practical procedures have been created for structuring cooperative, competitive, and individualistic efforts in education at both the classroom and the school levels (Johnson et al., 2013). *Cooperative learning* is the instructional use of small groups, so that students work together to maximize their own and each other's learning. From being relatively unknown and unused in the 1960s, cooperative learning is now utilized in schools and universities throughout most of the world in every subject area and from preschool through graduate school and adult training programs. Its use so pervades education that it is difficult to find a textbook on instructional methods, teachers' journal, or instructional materials that do not discuss cooperative learning. There are three types of cooperative learning: formal cooperative learning that is used for assignments that

last from one class period to several weeks, informal cooperative learning that is used with direct teaching for quick discussions that last from a few minutes to one class period, and cooperative base groups that last for a semester or year provide each member with the support, help, encouragement, and assistance that he or she needs to progress academically. The three types of cooperative learning may be used in an integrated way to form an overall instructional program.

The second area in which social interdependence theory is applied is in business and industry (Tjosvold, 1989). While there is a long history of discussing the importance of cooperation within organizations, it is the advent of the team-based organization that most recently highlighted the importance of social interdependence theory for organizational functioning. Teams are organized into interdependent units (such as departments), which are then organized into an even higher level of interdependent units (such as divisions), which in turn are organized into one interdependent unit (the organization as a whole). Teams, rather than individuals, then become the basic organizational building block. In many cases, however, teams perform at less than an optimal level. A number of ways to improve team performance have been developed based on social interdependence theory (Tjosvold, 1989).

Breath of Social Interdependence Theory

A strength of social interdependence theory is that it may be used to understand the dynamics of cooperation on multiple levels. Not only does it illuminate interpersonal relations, but it also illuminates the dynamics of intergroup, organizational, societal, and world interdependence. Just as individuals are interdependent through their common goals, groups, organizations, and societies are also united through mutual goals. The dynamics of cooperation and competition at each level are the same. In addition, cooperation simultaneously has powerful effects on numerous outcomes, from productivity, motivation, quality of relationships, social support, psychological health, self-esteem, and many, many others.

Relationship among Theory, Research, and Practice

Social interdependence theory provides an example of how psychological theory, research, and practice interact in ways that enhance all three. Theory identifies and defines the phenomena of interest and their relationships with each other. Social interdependence theory is well formulated, clearly defines cooperation and competition, summarizes the existing research, and has generated considerable new research. Well over a thousand research studies have been conducted to validate or disconfirm social interdependence theory. Thus, there is sufficient research, with both high internal and external validity, to test the theory. The amount, quality, and generalizability of the research provide strong confirmation of the basic propositions of the theory and the effectiveness of cooperative relative to competitive and individualistic efforts. Finally, social interdependence theory has been applied in many diverse areas (most notably in education and business) and the applications have had profound effects on the theory and related research, revealing inadequacies in the theory and pointing to further research that needs to be conducted.

Future Directions

Despite the numerous studies that have been conducted, considerably more research is needed on the nature of cooperation and the variables that mediate its effectiveness. The nature of competitive and individualistic efforts and the conditions under which they are effective need to be further investigated. The values inherently taught by participating in cooperative and competitive situations are just beginning to be directly studied (Johnson and Johnson, 1999). Cooperation, for example, may promote a concern for the well-being of others and a desire to contribute to the common good, while competition may promote a desire to be better than others and a concern for self-gain at the expense of others. Individualistic experiences may promote a focus on self-benefit with little or no concern about others. There is a need to investigate the differences in the use of power in cooperation and competitive situations. The importance of studying cooperation and competition may continue to increase as the importance of world interdependence (such as increasing economic interdependence among countries) and the need for peace increase.

See also: Collectivism and Individualism: Cultural and Psychological Concerns; Cooperation: Sociological Aspects; Helping Intergroup Relations; Intergroup Relations; Social Cognition.

Bibliography

Bandura, A., 1977. Principles of Behavioral Modification. Holt, Rinehart, and Winston, New York.

Bandura, A., 2000. Exercise of human agency through collective efficacy. Current Directions in Psychological Science 9 (3), 75–78.

Choi, J., Johnson, D.W., Johnson, R., 2011. Relationship among cooperative learning experiences, social interdependence, children's aggression, victimization, and prosocial behaviors. Journal of Applied Social Psychology 41 (4), 976–1003.

Deutsch, M., 1949. A theory of cooperation and competition. Human Relations 2, 129–152.

Deutsch, M., 1962. Cooperation and trust: some theoretical notes. In: Jones, M. (Ed.), Nebraska Symposium on Motivation. University of Nebraska Press, Lincoln, NE, pp. 275–319.

Deutsch, M., 1973. The Resolution of Conflict. Yale University Press, New Haven, CT.

Deutsch, M., 1985. Distributive Justice: A Social Psychological Perspective. Yale University Press, New Haven, CT.

Homans, G., 1974. Social Behavior: Its Elementary Forms, second ed. Harcourt Brace Jovanovich, New York.

Johnson, D.W., Johnson, R., 1989. Cooperation and Competition: Theory and Research. Interaction Book Company, Edina, MN.

Johnson, D.W., Johnson, R., 2007. Creative Controversy: Intellectual Challenge in the Classroom, fourth ed. Interaction Book Company, Edina, MN.

Johnson, D.W., Johnson, R., 1999. Learning Together and Alone: Cooperative, Competitive, and Individualistic Learning. Allyn and Bacon, Boston.

Johnson, D.W., Johnson, R.T., 2009a. Energizing learning: the instructional power of conflict. Educational Researcher 38 (1), 37–52.

Johnson, D.W., Johnson, R.T., 2009b. An educational psychology success story: social interdependence theory and cooperative learning. Educational Researcher 38 (5), 365–379.

Johnson, D.W., Johynson, R.T., Holubec, E.J., 2013. Cooperation in the Classroom, eighth ed. Interaction Book Company, Edina, MN.

Johnson, D.W., Johnson, R.T., Roseth, C.J., Shin, T.S., 2014. The relationship between motivation and achievement in interdependent situations. Journal of Applied Social Psychology 44, 622–633.
Kelley, H., Thibaut, J., 1978. Interpersonal Relations: A Theory of Interdependence. Wiley, New York.
Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, Cambridge.
Piaget, J., 1950. The Psychology of Intelligence. Harcourt, New York.
Roseth, C.J., Johnson, D.W., Johnson, R.T., 2008. The relationship between interpersonal relationships and achievement within cooperative, competitive, and individualistic conditions: a meta-analysis. Psychological Bulletin 134 (2), 223–246.
Skinner, B.F., 1968. The Technology of Teaching. Appleton-Century-Crofts, New York.
Stanne, M., Johnson, D.W., Johnson, R., 1999. Social interdependence and motor performance: a meta-analysis. Psychological Bulletin 125 (1), 133–154.
Tjosvold, D., 1989. Team Organization: An Enduring Competitive Advantage. Wiley, Chichester.
Vygotsky, L., 1978. Mind and Society. Harvard University Press, Cambridge, MA.
Wittrock, M., 1990. Generative processes of comprehension. Educational Psychologist 24, 345–376.

Cooperation: Sociological Aspects

Andreas Diekmann, Swiss Federal Institute of Technology, ETH-Zurich, Chair of Sociology, Zurich, Switzerland
Siegwart Lindenberg, University of Groningen, Groningen, The Netherlands; and Tilburg University, Tilburg, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The most prominent studies of cooperation in sociology proceed from some kind of rationality assumption. There are basically two groups of studies, one starting from the assumption of rational egoism and the other starting from the assumption of social rationality, in which other regard has a prominent place. The studies based on the assumption of rational egoists are strongly inspired by (mostly game-theoretical) developments in economics and political science, but add important sociological aspects to the analysis. Most important among these contributions are the dynamic analysis of cooperation, the link between formal and informal institutions, and the importance of social networks for embedding problematic transactions. Studies based on social rationality introduce assumptions on social preferences, learning, and framing. Important questions concern the circumstances that strengthen or weaken social preferences, the evolution of signals and conventions that govern trust, and the effects of goals on the definition of the situation. Studies using social rationality are on the rise.

The most prominent studies of cooperation in sociology proceed from some kind of rationality assumption. There are basically two groups of studies, one starting from the assumption of rational egoism and the other starting from the assumption of social rationality, in which other regard has a prominent place. The studies based on the assumption of rational egoism are strongly inspired by (mostly game-theoretical) developments in economics and political science, but add important sociological aspects to the analysis. Most important among these contributions are the dynamic analysis of cooperation, the link between formal and informal institutions, and the importance of social networks for embedding problematic transactions. Studies based on social rationality introduce assumptions on social preferences, learning, and framing. Important questions are those about the circumstances that strengthen or weaken social preferences, the evolution of signals and conventions that govern trust, and the effects of goals on the definition of the situation. Studies using social rationality are on the rise.

Individuals cooperate if each willingly acts in a manner that contributes to the other's welfare. Cooperation is one of the oldest and most revered topics in sociology, even though it has appeared under a variety of headings, such as 'social order' and 'solidarity.' The classical sociological answers to the question of how cooperation among individuals comes about differ in detail but share a common core: Cooperation is the result of shared values and norms and norm-conforming behavior.

A problem with this 'obligation-centered' solution is that there is no explanation of where norms come from and why the degree of norm conformity varies even for the same individual over time. Theories that might solve these problems were not developed until the 1970s, when a sea change in the basic assumptions on human action took place within sociology. At that time, analyses of cooperation among rational egoists (mostly rooted in game theory) began to shed new light on the way one might analyze cooperation, norms, and norm-conforming behavior. For some time, game-theoretic analyses of cooperation had become the most serious attempt to explain cooperation. Big advances were made but many problems remained, especially the problem that individuals do not seem to be as self-centered as they are assumed to be in game theory. Still, the game-theoretic analyses were powerful enough to suggest that there was no way back to explanations of cooperation that ignored human rationality all together. More recently, analyses based on some form of social rationality have arisen in order to address some of the unsolved problems. The following sections first present the most important aspects of sociological research on cooperation among rational egoists, and then discusses some contributions on cooperation with social rationality.

Cooperation Among Rational Egoists

It may not be quite realistic to assume that individuals are always looking out for themselves and that they do so by weighing the costs and benefits of their behavior. However, this assumption of 'rational egoism' seems to have four strong points in its favor. First, it allows one to pinpoint why cooperation is an interesting problem. Second, it allows rigorous theory formation with tractable structures of argumentation. Third, everyday experience and research give this assumption enough support to make it worthwhile to use it in the light of its analytic advantages, especially when analyzing interaction among interdependent individuals. Fourth, at later stages of theory formation, this assumption can be relaxed in favor of more realistic assumptions.

Game theory has furnished the most important instruments for such analyses, and since the late 1970s – inspired by developments in political science and economics – sociologists have begun to use it to study questions concerning cooperation, typically bringing sociological aspects into the analyses, such as network embedding. The crucial first step in the 'new' analysis was to identify typical situations in which cooperation is problematic: social dilemmas. A social dilemma is – roughly speaking – a situation in which the rational decisions of

individual actors are likely to lead to a collectively suboptimal outcome. An outcome is suboptimal (i.e., not 'Pareto-optimal') if one or more actors could improve their position without worsening the position of any other actors. In a social dilemma, individual and collective rationalities are in conflict. Cooperation can now be succinctly defined: An actor cooperates if and only if he or she chooses a course of action that will lead to a collectively rational outcome when other actors behave cooperatively as well. Social dilemmas occur when Pareto-superior behavior cannot be achieved by enforceable contracts. This may have many different reasons. For example, it may be impossible to communicate with others in order to come to an agreement; or, if communication is possible, transaction costs may be too high to actually reach an agreement. If agreements are within reach, it may be practically or epistemically impossible to specify all relevant future contingencies in the contract. Even if a contract can be drawn up, there may be no court with sanctioning power; or if there is such a court, it may not be possible to prove uncooperative behavior in court.

There are structurally different variants of social dilemmas, and game theory helps to distinguish them. For example, there is the common's dilemma (Hardin, 1968; Ostrom, 1990) in which various actors have free access to a resource (such as fishers from different nations in international waters); there are step-level public goods (Taylor, 1987), for which only a limited number of actors have to cooperate in order to produce the public good for a much larger group. A special case of this is the volunteer's dilemma (Diekmann, 1985), for which only one volunteer is necessary to produce the public good for all, such as rescuing a drowning child with a crowd of onlookers; there is the trust game (Dasgupta, 1988) in which the trustee can honor or opportunistically exploit the cooperation of the other (as when you lend money to an acquaintance who promises to pay it back with interest).

The prisoner's dilemma (PD) is by far the best-known variant of a social dilemma. It involves an arbitrarily large number of people. The special case of a two-person PD has been explored the best so far. In a two-person PD, both players have the option to cooperate or to defect (Figure 1). If both cooperate, each gets a payoff of value R. If one defects while the other cooperates, the defector gets T and the cooperator gets S. If both defect, both will get P, with $T > R > P > S$. There is a dominant strategy for each player (i.e., an individually rational choice, no matter what the other does) leading to a 'Nash equilibrium' (in which no player can unilaterally improve his or her position by deviating from the equilibrium strategy) and

	C	D
C	R, R	S, T
D	T, S	P, P

Figure 1 Prisoner's dilemma game. In the Axelrod simulation is T (temptation) = 5, R (reward) = 3, P (punishment) = 1, and S (sucker's payoff) = 0.

to a Pareto-inferior result. The assumption of rational egoism leads to two motives in a PD: the defensive motive to shield against exploitation (if the other defects you are better off defecting yourself), and the aggressive motive to exploit a cooperative player (if the other cooperates, you get even more by defecting). Together, these motive result in the conflict between individual and collective rationality. Mutual cooperation would be the Pareto-optimal result, but the two motives lead to mutual defection.

The Embedding of Cooperation

Raub and Weesie (2000) have suggested that game-theoretic work on solutions to social dilemmas can be divided into three kinds of embeddings of potential cooperation: embedding in time, in institutions, and in social networks. We follow their suggestion for this overview.

Time

When actors find themselves in a social dilemma that is iterated indefinitely over time, new sources for cooperation arise. First of all, defection can be sanctioned by the other players in later moves, and the resumption of cooperation by the defector can be rewarded by a similar response of the other players. Friedman (1977) and Taylor (1987) have shown that if the value of future outcomes is high enough, this opportunity to make one's cooperation conditional on that of the other leads to cooperative strategies with Nash equilibrium and Pareto optimality. Axelrod (1984) built on this result to ask the question of which strategy would be most conducive to lead to cooperation in an iterated PD. In order to answer this question, he conducted computer simulations in which various strategies played against each other for about 200 rounds. Axelrod had invited suggestions for the strategies that would create the most cooperation. The winner was the conditional strategy 'tit for tat' (TFT) suggested by the mathematician and psychologist Anatol Rapoport who gained his insight from early experimental studies with the repeated PD (Rapoport and Chammah, 1965).

TFT is very simple: (1) always begin with cooperation; (2) for the present round, copy the move of your partner in the previous round. TFT has a number of distinct features. It is 'nice,' because it starts with cooperation; it cannot be provoked, because it sanctions exploitation; and it is forgiving, because it returns to cooperation after the other has done so. Contrast this, for example, with 'permanent retaliation' in which defection is answered by unforgiving defection, a strategy that did much worse than TFT. It does not do well against all sorts of strategies and it does not respond optimally to mistakes that might occur with small probability (Nowak, 2012).

The heart of Axelrod's results, however, is thoroughly in line with a classical sociological insight: that the norm of reciprocity is probably the most essential norm generating cooperation in time-embedded interactions. As sociologists have argued for quite some time, homogeneity of a group in terms of important characteristics (including the use of 'nice' strategies) will increase the likelihood of reciprocity, as does smallness of the group because it helps identification of defection. Similarly, a low rate of fluctuation of membership helps reciprocity because it increases the shadow of the future.

A related argument has been made in terms of 'social capital' (Coleman, 1990).

Axelrod's tournament was based on the opportunity for 'direct reciprocity' among two actors. In larger groups, a system of cooperation may emerge because of 'indirect reciprocity.' In such a system A's cooperative decision in favor of B is not rewarded directly by B but by some other actor observing A's helpful act. Provided that cooperative acts increase an actor's reputation score, the possibility of indirect reciprocity leads, under certain conditions, to a stable state of cooperation (Nowak, 2012). Not formally proved systems of direct and indirect reciprocity had already been explored by classical sociologists and anthropologists decades ago (Blau, 1964), but modern game theory and agent-based simulation methods help to find out the conditions for the evolution and stability of cooperative systems.

Institutions

When reciprocity norms work, they are self-enforcing because no player can do better by unilaterally deviating from the cooperative strategy (Nash equilibrium). However, under many conditions, norms of reciprocity will not work properly. For example, when future contingencies are so complex that it is difficult to determine a priori what constitutes defection or difficult to detect defection, a reciprocity norm is not enough to secure cooperation. Similarly, such self-enforcing norms do not work well for the production of collective goods in larger groups, the less so the more inhomogeneous and fluctuating the group membership. In such situations, expectations concerning the relative importance of one's own and others' contribution are likely to be low and individuals tend to free ride rather than to contribute to the production of the collective good.

When self-enforcing norms do not work, cooperation may be achieved by embedding interactions in formal or informal rules that create 'selective incentives'; that is, added rewards for cooperating and punishments for not cooperating. This argument was first forcefully made by Olson (1965) and has been refined by sociologists and other social scientists (e.g., Hechter, 1987; Marwell and Oliver, 1993; Ostrom, 1990). For example, the state punishes tax evasion and provides some standard instruments for contracting; in many countries, labor unions offer extra benefits (such as cheaper insurance) for joining the labor union. The formation of institutional solutions has been explained by Hechter (1987) as a two-stage process. First, groups for the joint production of private goods are established (such as insurance). They are small enough that noncooperators can be detected and excluded (hence no free rider problem). The profits made from the joint production can then be used to establish selective incentives for the joint production of collective goods. In this way, labor unions may emerge from mutual insurance groups. Abell (1996) adds to this kind of analysis that a proper institutional explanation of cooperation should include the explanation of how cooperation can be combined with rivalry and competition. More work in this direction is needed.

Williamson (1985) has pointed to the importance of 'private orderings' in which, in the shadow of the law, contract partners may voluntarily restrict their own options, thereby creating selective incentives for their own cooperative action. For the partner, such measures work if there are 'credible commitments,' for example, by posted hostages and granted warranties. Sociologists have used this idea successfully to study contractual behavior (e.g., Weesie et al., 1998).

Often, credible commitments are difficult to establish (due to the size of the group) and formal institutional arrangements (with enforced sanctions) may not be available. In such groups, informal rules (social norms) that govern selective incentives for cooperation are needed. Like all rules, informal rules work only if they are enforced, and this creates the second-order free rider problem. Sanctioning noncooperators is itself a (step-level) collective good. There seems to be an infinite regress. How do groups get out of this problem? There is no handy solution to this problem within the rational egoistic approach, although suggestions were put forward recently (Raihani and Bshary, 2011; Przepiorka and Diekmann, 2013). However, as Heckathorn (1996), following Marwell and Oliver, has shown, public good production is not always subject to the first- and second-order free rider problems. In fact, the production function (i.e., the function that relates the proportion of cooperators to the level of the public good production) is likely to be S-shaped and only for certain values of the function will the situation have the structure of a PD. The reason for this is that an S-shaped production function implies a changing marginal utility from cooperation as the proportion of cooperators increases. Interestingly, Heckathorn also found that in certain phases 'hypocritical cooperation,' in which people criticize others for not cooperating even though they do not cooperate themselves, is quite effective in eliciting contributions to the collective good (Yamagishi, 1986). The main lesson to be learned from Heckathorn's analysis is that collective action should be analyzed dynamically, a conclusion one can also draw from Hechter's analysis of institutional controls mentioned above.

Social Networks

It has long been known in sociology that it often matters to people what other people say about them. Cooperation is helped by the effects of reputation within networks. For example, DiMaggio and Louch (1998) found in a national sample that consumers, when they are uncertain about products and performance, prefer to buy goods and services from suppliers with whom they have social ties outside the transaction itself. If the suppliers in such a network are clearly unreliable, the consumers are likely to have heard about it. If these suppliers begin to be unreliable, the network relations can be turned against them. Thus, networks have an information function (people learn about the cooperativeness of certain others) and a control function (people can use the network to sanction noncooperators).

Both aspects have been worked out game-theoretically by Buskens and Raub (2013). For dyadic transactions, characteristics of the network (such as density, 'indegree,' and 'outdegree'; that is, the number of others with whom people are connected) play an important role but it turns out that reputation effects are quite complex and there are indications that they might be less important than the temporal embeddedness. For economic transactions, Uzzi (1996) has found interesting additional effects of networks on tacit knowledge and joint problem solving. Network embeddedness for collective action has apparently not been studied yet in any rigorous way.

Yet another use of networks for the explanation of cooperation has been suggested by Coleman (1990). He points to the

importance of intermediaries for the creation of trust and thus for cooperation. Intermediaries can play a number of roles: as advisors who introduce interested parties and facilitate the interaction necessary to create trust; as guarantors who absorb the risk if trust turns out to be misplaced; and as entrepreneurs who combine the resources of several actors to place them into the hands of others who are expected to realize gains. For future research, a sociologically interesting question is under what conditions will structures with certain kinds of intermediaries emerge.

Cooperation Among Socially Rational Individuals

There is a growing sociological literature on cooperation that holds on to the assumption of rational action but adds specific 'bridge theories' on preferences, learning, and framing, the effect of which is to render regard for others quite prominent in the analyses. Rationality is seen as having evolved together with the ability to derive adaptive advantages from living in groups (Dunbar, 2003). Hence the term social rationality. Preferences are likely to have been influenced by this process, and the social influence on cognitive and motivational processes may even create situationally different kinds of rationality. In this literature the trade-off between empirical richness and analytical tractability is different than in most game-theoretic analyses. In the latter, added analytical power and tractability is often achieved by the rational egoist assumption at the expense of some loss in closeness to reality, whereas in the former, some analytical power and tractability is sacrificed in favor of added closeness to reality. Clearly, each kind of approach profits from the existence of the other. In this review, there remains room for the discussion of only a few developments.

Goals (Preferences)

By far the largest amount of literature on cooperation in the 'social rationality' category is concerned with 'social preferences.' Although the term preference is often used in this context, a better term would be 'social goals.' One variant of this approach, found in social-psychological and sociological research, is 'social orientation' research, in which individuals are differentiated according to their orientation toward others. This orientation is often assumed to be more or less stable and the joint result of nature and nurture. Different types of orientation are distinguished, most frequently the following three: 'cooperative' (the goal is to maximize joint payoffs); 'individualistic' (the goal is to maximize individual payoff); and 'competitive' (the goal is to maximize the positive difference between own and other's payoff). Such 'types' can be combined with game-theoretic analyses or with analyses of evolutionary learning at the population level.

A related but different approach emphasizes that human beings have evolved to have (pro)social goals that make them dependent on others and thus willing to cooperate for a variety of reasons. This concerns mainly the value of status (leading to cooperation by the strategic adaptation to authority relations, or to lack of cooperation due to status competition), the value of behavioral confirmation (leading to cooperation on the basis of conformity to peers' expectations), and the value of affection (leading to cooperation on the basis of liking) (Lindenberg and Frey, 1993).

Importantly, cooperation also depends on negative goals, especially the desire to punish noncooperators. An influential experiment by Fehr and Gächter (2002) demonstrates that subjects punish free riders in collective good situations even if sanctioning is costly. This type of behavior is called 'altruistic reciprocity.' Actors pay a price for negatively reciprocating to noncooperative behavior, thereby sustaining cooperation in a public good situation. Having the choice between a regime with sanction-free and sanctioning institutions in a collective good situation, most subjects chose the former in the beginning. However, after the breakdown of cooperation under the sanction-free regime, subjects increasingly change to the alternative regime. This experiment by Gürerk et al. (2006) exhibits that sanctioning institutions solving the problem of cooperation emerge in a dynamic, evolutionary process. Yet, on a societal level, the willingness to engage in costly punishing varies considerably among different cultures and countries (Henrich et al., 2006).

Sociologists and economists who focus on different types of people (in terms of orientation) are especially concerned with questions of signaling and the spread of strategies over time, some of them using genetic algorithms taken from evolutionary biology. This makes trust a central concept in the analysis of cooperation. For example: "Are the tendencies to trust and to honor trust correlated?" (Snijders, 1996); "How can cooperative types convincingly signal their type to others so that they will be trusted and cooperation comes about?" (Posner, 2000). Or "Under what conditions will effective rules for trusting (i.e., conventions for detecting cooperators) evolve in neighborhood interactions and spread across the population through contact with strangers?" (Macy and Skvoretz, 1998). One problem with this concept of social orientation is that it does not allow for situational influences on preferences and cognitions. Could not there be situations in which one is prosocial and others in which one would be quite egoistic?

Framing Effects

The willingness to cooperate seems also to depend on how people define a situation. Of particular importance is the difference between gains and losses. Kahneman and Tversky have shown that individuals are keener on avoiding losses than on achieving gains, *ceteris paribus*. This leads to the hypothesis that it is easier to bring about cooperation concerning the defense of a common pool resource than the achievement of a (new) public good. For example, it should be easier to bring about a campaign to preserve a rainforest than a campaign to create a wildlife preserve. The evidence for this is somewhat mixed (Van Assen, 1999), but when the disturbing effects are carefully controlled, the framing prediction finds support, especially for initial stages of cooperation (Sell and Son, 1997). There is also a claim that people often cognitively transform a PD into a game with conditional cooperation (assurance game). The interesting question then is how the beliefs that others will cooperate are influenced by social circumstances or by cultural beliefs (Hayashi et al., 1999), and what influence certain thresholds of cooperation have for triggering conditional cooperation in a larger group (Yamaguchi, 2000).

A different framing effect comes from 'goal framing' (Lindenberg and Steg, 2007). Overarching goals have a strong impact on the way individuals define a situation and thereby create a difference between 'foreground' and 'background' information and preferences. What is in the foreground will have a much stronger impact on behavior than what is in the background, *ceteris paribus*. For example, if the overarching goal (the goal frame) 'to act appropriately' is salient, then information on normative expectations and ways to meet them are in the foreground and in the center of attention, whereas the overarching gain goal 'to increase and guard one's resources' is in the background. As a result, when the goal-frame is normative (i.e., the goal to act appropriately is salient), orientation toward common goals and norms is focal and costs of cooperation are much less prominently perceived. Costs of cooperation must then be much larger in order to reduce people's willingness to cooperate than when gain itself is the goal frame. Sociologically, an important feature of overarching goals is that people are likely to signal their goal-frame to others through their behavior (e.g., through 'relational signals'), creating contagion effects that make (non)cooperation spread. The cooperative orientation linked to the normative goal-frame seems to be triggered especially when others show concern for norms and also when the situation is clearly one of reaching a common goal ('joint production,' Lindenberg and Foss, 2011). Group identity processes also play an important role (Simpson and Macy, 2004). All this puts a high premium on including cues that signal goal-frames (prominently including symbolic behavior) in the analyses of cooperation, even in game-theoretic analyses, and in the analysis of how institutions work to foster cooperation.

All in all, in sociology, studies on cooperation are both based on the assumption of rational egoists and on some kind of social rationality. The latter are clearly on the rise in number and sophistication.

See also: Affect-Regulation Motivation; Cooperation and Competition; Decision and Choice: Bounded Rationality; Dispute Resolutions in Economics; Emotions, Sociology of; Game Theory: Noncooperative Games; Game Theory; Groups, Sociology of; Institutions; Network Analysis; Norms; Organizational Decision Making; Organizations, Sociology of; Prisoner's Dilemma, One Shot and Iterated; Psychoanalysis in Sociology; Rationality in Society; Social Relationships in Adulthood; Solidarity, Sociology of; Trust, Sociology of.

Bibliography

Abell, P., 1996. A model of the informal structure (culture) of organizations: help, trust, rivalry and team spirit. Rationality and Society 8, 433–452.
Axelrod, R., 1984. The Evolution of Human Cooperation. Basic Books, New York.
Blau, P.M., 1964. Exchange and Power in Social Life. Wiley, New York.
Buskens, V., Raub, W., 2013. Rational Choice Research on Social Dilemmas: Embeddedness Effects on Trust. In Wittek, R., Snijders, T.A.B., Nee, V (Eds.), Handbook of Rational Choice Social Research. Stanford University Press, Stanford, CA, pp. 113–150.
Coleman, J.S., 1990. Foundations of Social Theory. Harvard University Press, Cambridge, MA.
Dasgupta, P., 1988. Trust as a commodity. In Gambetta, D. (Ed.), Trust: Making and Breaking Cooperative Relations. Blackwell, New York, pp. 49–72.
Diekmann, A., 1985. Volunteer's dilemma. Journal of Conflict Resolution 29, 605–610.
DiMaggio, P., Louch, H., 1998. Socially embedded consumer transactions. American Sociological Review 63, 619–637.
Dunbar, R.I.M., 2003. The social brain: mind, language, and society in evolutionary perspective. Annual Review of Psychology 32, 163–181.
Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 425, 137–140.
Friedman, J.W., 1977. Oligopoly and the Theory of Games. North-Holland, Amsterdam.
Gürerk, G., Irlenbusch, B., Rockenbach, B., 2006. The competitive advantage of sanctioning institutions. Science 312, 108–111.
Hardin, G., 1968. The tragedy of the commons. Science 162, 1243–1248.
Hayashi, N., Ostrom, E., Walter, J., Yamagishi, T., 1999. Reciprocity, trust, and the sense of control. Rationality and Society 11, 27–46.
Hechter, M., 1987. Principles of Group Solidarity. University of California Press, Berkeley, CA.
Heckathorn, D., 1996. The dynamics and dilemmas of collective action. American Sociological Review 61, 250–277.
Henrich, J. et al., 2006. Costly punishment across human societies. Science 312, 1767–1770.
Lindenberg, S., Frey, B., 1993. Alternatives, frames, and relative prices: a broader view of rational choice. Acta Sociologica 36, 191–205.
Lindenberg, S., Foss, N., 2011. Managing joint production motivation: the role of goal-framing and governance mechanisms. Academy of Management Review 36, 500–525.
Lindenberg, S., Steg, L., 2007. Normative, gain and hedonic goal frames guiding environmental behavior. Journal of Social Issues 65, 117–137.
Macy, M., Skvoretz, J., 1998. The evolution of trust and cooperation between strangers. American Sociological Review 63, 638–660.
Marwell, G., Oliver, P., 1993. The Critical Mass in Collective Action. Cambridge University Press, Cambridge, MA.
Nowak, M.A., 2012. Evolving cooperation. Journal of Theoretical Biology 299, 1–8.
Olson, M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.
Ostrom, E., 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, Cambridge, UK.
Posner, E.A., 2000. Law and Social Norms. Harvard University Press, Cambridge, MA.
Przepiorka, W., Diekmann, A., 2013. Individual heterogeneity and costly punishment: A volunteer's dilemma. Proceedings of the Royal Society B 280, 20130247.
Raihani, N.J., Bshary, R., 2011. The evolution of punishment in n-player public goods games: A volunteer's dilemma. Evolution 65, 2725–2728.
Rapoport, A., Chammah, A., 1965. Prisoner's Dilemma. A Study in Conflict and Cooperation. University of Michigan Press, Ann Arbor, MI.
Raub, W., Weesie, J., 2000. The management of matches: a research program on solidarity in durable relations. Netherlands' Journal of Social Sciences 36, 71–88.
Sell, J., Son, Y., 1997. Comparing public goods with common pool resources: three experiments. Social Psychology Quarterly 60, 118–137.
Simpson, B., Macy, M.W., 2004. Power, identity, and collective action in social exchange. Social Forces 82, 1373–1409.
Snijders, C., 1996. Trust and Commitment. Thela Thesis, Amsterdam.
Taylor, M., 1987. The Possibility of Cooperation. Cambridge University Press, Cambridge, UK.
Uzzi, B., 1996. The source and consequences of embeddedness for the economic performance of organizations. The network effect. American Sociological Review 61, 674–698.
Van Assen, M.F., 1999. Effects of individual decision theory assumptions on predictions of cooperation in social dilemmas. Journal of Mathematical Sociology 24, 1–11.
Weesie, J., Buskens, V., Raub, W., 1998. The management of trust relations via institutional and structural embeddedness. In Doreian, P., Fararo, T. (Eds.), The Problem of Solidarity: Theories and Models. Gordon and Breach, Amsterdam, pp. 113–138.
Williamson, O.E., 1985. The Economic Institutions of Capitalism. Free Press, New York.
Yamagishi, T., 1986. The provision of a sanctioning system as a public good. Journal of Personality and Social Psychology 51, 110–116.
Yamaguchi, K., 2000. Subjective rationality of initiators and of threshold-theoretical behavior of followers in collective action. Rationality and Society 12, 185–225.

Cooperative Game Theory

Toru Hokari, Department of Economics, Keio University, Keio, Japan
William Thomson, University of Rochester, Rochester, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

We survey the literature on coalitional games, focusing on recent developments and on the axiomatic approach to the subject.

Introduction

Consider a group of agents, or 'players,' to whom is available a set of payoff vectors, when a choice has to be made from this set. Included in the description of the situation is a set of payoff vectors available to each subgroup of players. Altogether, we have a 'coalitional game.' The question is: What should be assigned to each player? In answering this question, proper account should be taken of the opportunities open to the various groups to which he belongs; these opportunities should be compared, and what they mean for players should be assessed. But how? What formula or procedure should we use? On what principles should we base their evaluation? The objective of the 'theory of cooperative games' is to answer these questions. What follows is a presentation of the main concepts that have been elaborated in this theory, and some central results. The literature is vast, consisting of several thousand items, and we have had to be very selective. Hopefully, our survey still conveys a sense of the achievements of this research program.

The analysis starts with the specification of domains of games. Two main ones are traditionally considered. First, what a coalition can achieve is given as a single number, interpreted as 'utility,' which the coalition can distribute among its members in any way it wishes, the sum of their assignments being equal to that number. Then, utility is interpersonally comparable across players and transferable at a one-to-one rate from one to the other. These are 'transferable utility' games. (It is simplest to think that what is achievable by each coalition is expressed in dollar terms. For that reason, these games are sometimes called 'games with side-payments.') Second, what a coalition can achieve is specified as a subset of its 'utility space' satisfying some regularity conditions. Utility may still be transferable but at a rate that depends on the starting point. These are 'nontransferable utility' games. (They are also sometimes called 'games without side-payments.' The unfortunate way in which the terminology has solidified does not reflect well what distinguishes the first situation from the second one; it is not so much whether utility or payoff is transferable that is at issue, but rather whether the rate at which transfers can be made is constant or not. A game in which utility is not transferable at all would consist of the comprehensive hull of a single point for each coalition.)

At the risk of oversimplifying, one can say that the early literature on coalitional games was mainly concerned with defining solutions and with existence and structural issues. In the past 30 years, most work has been axiomatic: desirable properties of solutions are formulated, as axioms, and the search is for solutions satisfying the properties in various combinations.

What are the main directions that the recent literature has taken? First, a number of new axioms and new solutions have been defined; new characterizations of standard solutions have been developed and these new solutions have also been characterized. Second, multiple enrichments of the base models have been proposed and analyzed in detail. Third, applications have multiplied.

Several specialized surveys have appeared in the *Handbook of Game Theory* (1994). General treatments of the subject are Weber (1994), Peleg and Sudhölter (2007), and Peters (2008).

General Definitions: Games with Transferable Utility

There is a group N of *players* who can gather in *coalitions*. (A coalition is a *nonempty* subset of N; we apply the term even to subsets consisting of a single player.) The group N itself is the *grand coalition*. For each coalition $S \subseteq N$, a number $v(S)$ is specified, the *worth of S*, which is understood as the maximal amount of 'utility' that S can achieve somehow, and that can be distributed among its members, any vector of payoffs in $\mathbb{R}^S$ whose components do not add up to more than $v(S)$ being available to them. Summarizing, a *transferable utility coalitional game*, simply a *TU game*, is a vector $v \equiv (v(S))_{S \subset N} \in \mathbb{R}^{2^{|N|}-1}$. Let $\mathcal{V}^N$ be the domain of all games.

Games can be given as primitives, but they can also be derived from other data, for instance, resources that players control individually or within coalitions. Production processes may be available to a coalition. When some goods are public goods, or when other forms of externalities are present, the calculation of the worth of a coalition is delicate because how the complementary coalition behaves is relevant to the calculation, and assumptions have to be made about what that coalition has chosen to do. One should not expect a simple and unique answer to these type of questions, and extreme scenarios have been used as benchmarks. According to the most pessimistic scenario, the easiest to define, the complementary coalition contributes nothing. Alternatively, one can imagine a strategic game played by coalitions: each coalition chooses some action, and to each profile of actions is associated outcome; the strategic equilibria are identified, and the resulting payoffs to the coalitions define their worths. (This is what von Neumann and Morgenstern suggested in general to derive cooperative games from strategic situations.) Absent such a game, one can simply think of $v(S)$ as some index of opportunities open to S. What is important is that the index be calculated consistently across coalitions. It is usually straightforward to calculate the worth of the grand

coalition, however, because the complementary coalition is empty.

Various subclasses of games have been considered. Some have arisen naturally in applications to economics and political science; others have emerged out of mathematical considerations. It will be convenient to collect all the definitions here. Most of them have to do with quantifying the synergies that may exist between players. The game v is *inessential* if there is $x \in \mathbb{R}^N$ such that for each $S \subseteq N$, $v(S) = \sum_S x_i$. It is *essential* if, for each $S \subseteq N$, $\sum_S v(\{i\}) \leq v(S)$. It is *simple* if all its coordinates are either 0 or 1 (in political science applications, these numbers represent 'losing' or 'winning' an election). It is *monotonic* if, given two coalitions related by inclusion, the worth of the larger is at least that of the smaller one: for each pair S, $T \subseteq N$ with $S \subset T$, we have $v(S) \leq v(T)$. It is *super-additive* if the worth of a coalition is at least the sum of the worths of the components of each of its partitions: for each $S \subseteq N$ and each partition $(S_k)_{k \in K}$ of S, we have $v(S) \geq \sum_{k \in K} v(S_k)$. This is the case if these subcoalitions do not get in each other's way. The property would not be met, for example, if each subcoalition would need a particular facility controlled by S, but that only one subcoalition could use at a given time. The game is *convex* if there are strong synergies between players: for each pair S, $S' \subset N$ such that $S \subset S'$ and each $i \notin S'$, we have $v(S \cup \{i\}) - v(S) \leq v(S' \cup \{i\}) - v(S')$. Consider, for instance, workers who are supplying homogeneous labor to a convex technology; then, plotting the worth of coalitions as a function of their size gives us a convex function; alternatively, a game is convex if for each pair S, $S' \subset N$, we have $v(S \cup S') + v(S \cap S') \geq v(S) + v(S')$. A game is *symmetric* if the worth of a coalition only depends on its size.

The *dual of* v, v^d, is the game defined by setting, for each $S \subseteq N$, $v^d(S) \equiv v(N) - v(N \backslash S)$. Duality is useful in structuring the theory. The notion can be applied to solutions as well, and to properties of solutions. Properties of games that are preserved under duality are identified by Charnes et al. (1978). Other studies of duality are by Funaki (1995), Klijn et al. (2001), and Oishi et al. (2013).

A (feasible) *payoff vector* for v is a vector $x \in \mathbb{R}^N$ such that $\sum_N x_i \leq v(N)$. Let $\mathcal{G}^N \subseteq \mathcal{V}^N$ be a domain of games with player set N. A *solution on* $\mathcal{G}^N$ is a correspondence that associates with each $v \in \mathcal{G}^N$ a nonempty subset of its set of payoff vectors. Each x selected by the solution for $v \in \mathcal{G}^N$ is interpreted as a list of amounts that the players can 'expect' from v, either from some process of bargaining that is left unspecified or as the recommendation that an arbitrator could make. A *single-valued* solution is a *value*. Several of the central solutions are *single valued*, but several important others are not.

We also consider an enriched model in which the population of players is not fixed. There is now a set of 'potential' players, which may be infinite, but any game involves finitely many of them. Let $\mathcal{N}$ be the class of its finite subsets. Here, a solution is a mapping from $\cup_{N \in \mathcal{N}} \mathcal{G}^N$ to $\cup_{N \in \mathcal{N}} \mathbb{R}^N$. It associates with each $N \in \mathcal{N}$ and each $v \in \mathcal{G}^N$ a nonempty subset of the set of feasible payoff vectors of v.

The *subgame of a game* $v \in \mathcal{G}^N$ *with respect to* $N' \subset N$ is the game $(v(S))_{S \subseteq N'} \in \mathcal{G}^{N'}$. Many of the properties of games listed above are inherited by their subgames (e.g., super-additivity and convexity), as is easily checked.

It will also be useful to evaluate a payoff vector selected for a game $v \in \mathcal{G}^N$ in relation to the list of payoff vectors selected for its subgames. A (feasible) *payoff configuration for* v is a list of payoff vectors, one for v and one for each subgame of v. A *payoff-configuration solution on* $\mathcal{G}^N$ is a correspondence that associates with each $v \in \mathcal{G}^N$ a list, indexed by the subsets of N, of nonempty subsets of the sets of payoff vectors for the corresponding subgames of v. A *payoff-configuration value* is defined in a similar way, but it is *single valued*. Note that two games v and $v' \in \mathcal{G}^N$ may have the same subgames relative to some coalition S (i.e., $v_S = v'_S$), but the components relative to S of the payoff configurations chosen for v and v' by a payoff-configuration value differ.

An Inventory of Solutions

In presenting the various solutions and values that have been discussed in the literature, we proceed in steps, starting with naive and imperfect definitions, and progressively attempting to come up with more satisfactory ones. The definitions can be organized in several ways. A meaningful partition is based on whether they are *single valued* or not. *Single-valued* solutions make precise recommendations and are more likely to be useful. Another partition, which is how we chose to organize the material, is into *coalition-centric* and *player-centric* mappings. Indeed, some solutions can be seen as attempting to satisfy, or at least to 'mollify' coalitions, taking into account what they can achieve. The difficulty in doing so is that the number of instruments, namely the coordinates of the payoff vector chosen for a game, is very small in relation to the number of coalitions (n versus $2^n - 1$). The focus of the player-centric solutions is on assessing what individual players contribute to the various coalitions to which they belong, and on basing the calculation of payoffs on this assessment. What makes this approach delicate is that the members of a coalition contribute jointly to creating its worth (at least, the data of a game do not include information that can help us in disentangling their contributions), and that coalitions overlap. How much of the worth of a coalition should be imputed to each of its members, and how should the fact that each player belongs to multiple coalitions, be handled?

One should not expect to find a unique way of solving these conceptual difficulties, but the axiomatic approach helps us evaluate the relative merits of solutions and values that have been proposed as answers.

For several of the solutions introduced in the subsequent sections, nonemptiness is not guaranteed, and for some, identifying those games for which nonemptiness holds has been a central research question.

Let N and $v \in \mathcal{G}^N$ be fixed.

- Our first solution embodies the common and basic requirement that there should be no waste: the worth of the grand coalition should be fully allocated.

A *pre-imputation of* v is a vector $x \in \mathbb{R}^N$ whose coordinates add up to $v(N)$: $\sum x_i = v(N)$. The *pre-imputation solution*, *preI*, selects all pre-imputations of v.

- An *imputation* is a pre-imputation at which each player is assigned at least his own worth. The *imputation solution*, *I*,

selects all of these pre-imputations of v: $I(v) \equiv \{x \in preI(v)$: for each $i \in N$, $x_i \geq v(\{i\})\}$. For imputations to exist, it is obviously necessary and sufficient that $v(N) \geq \sum v(\{i\})$. This is the case for what we call an essential game.

- What underlies the definition of the *core*, C (Gillies, 1953) is the assumption that a coalition has access to all the possible vectors whose coordinates add up to its own worth. The core selects all the imputations of v over which no coalition S can improve by distributing $v(S)$ among its members: $C(v) \equiv \{x \in \mathbb{R}^N : \sum x_i = v(N), \quad \text{and for each } S \subset N, \quad \sum x_i \geq v(S)\}$.
- The validity of an objection may be called into question if it has a compelling counter-objection. Several definitions of *bargaining sets* as sets of payoff vectors such that each objection has a counter-objection have been proposed, implementing the idea. Peleg and Sudhölter (2007) are particularly detailed on the subject and we can do no better than to recommend that they be consulted. We limit ourselves to the central ideas. The original definition is due to Aumann and Maschler (1964). Given a pre-imputation x of v and two players, i and j, an *objection* of i against j is a pair of a coalition S that contains i but not j, and a payoff vector y feasible for S such that its members' payoffs are at least as large at y as at x, strict inequality holding for at least one i. A counter-objection is raised by a coalition T that contains j but not i; it is a payoff vector z feasible for T such that each of its members's payoff is at least as large at z as at x or as at y if he belongs also to T.

Related definitions are due to Mas-Colell (1989) and Zhou (1994). Shimomura (1997)'s definition differs from that of Mas-Colell's in two ways: he works with imputations instead of pre-imputations, and he insists that to be valid, objections and counter-objections should bring about an increase in the payoffs of all of the players who raise them. Another definition, proposed by Granot (1994, 2010), is studied by Granot and Maschler (1997).

- A' *von Neumann–Morgenstern solution of* v (note the indefinite article) is a subset K of the set of feasible vectors satisfying two conditions: *internal stability* says that no point of K should dominate any other point of K; *external stability* says that each point not in K should be dominated by some point of K.
- The definitions just given do not distinguish between the objections that a coalition can make to a payoff vector in terms of the magnitude of the difference between the sum of the payoffs to its members and its worth; only the sign of this difference matter. The next definition takes that magnitude into account; indeed, one would expect that it should play a role in motivating coalitions to identify and register objections. Given x, a pre-imputation of v, define the *dissatisfaction of coalition S at x in v* as the difference $v(S) - \sum_S x_i$.

The *prenucleolus*, *PreNuc* (Schmeidler, 1969) selects the pre-imputation of v at which the dissatisfactions of the various coalitions are minimized in the lexicographic order, starting with the coalition with the largest dissatisfaction. We write in singular ('the pre-imputation') because it turns out that there is a unique point solving this sequence of minimizations. The *nucleolus*, *Nuc*, is defined in a similar way but the minimization is carried out within the set of imputations.

- The next definition is based on an attempt to reach a 'balance' between how each player is treated in relation to each other player. First, we propose a measure of what player i can achieve without the help of player j. Given x, a pre-imputation of v, and $i, j \in N$, let $s_{ij}(v, x) \equiv \max\{v(S) - x(S) | S \subseteq N,\ i \in S,\ j \notin S\}$. The *prekernel*, *preKer*, selects the pre-imputation at which these measures are as equal as possible across players: $preKer(v) \equiv \{x \in preI(v)$: for each pair $i, j \in N$, $s_{ij}(v, x) = s_{ji}(v, x)\}$.

For the *kernel*, equality of the measures is insisted upon, but payoff vectors are required to be imputations.

- The next definition is concerned with overall distributions of payoffs. Equality is the goal, but the core constraints have to be respected, so we approximate equality subject to them.

The *Dutta–Ray value*, *DR* (sometimes called the Fushishige–Dutta–Ray solution; Dutta and Ray, 1989) selects, for each convex game v, the element in the core of v – the unique element – that is closest to the payoff vector at which each player is assigned an equal share of $v(N)$ (see the plain egalitarian value in the following). It turns out that the Dutta–Ray payoff vector 'Lorenz-dominates' each other element in the core of v. (Given two vectors $x, y \in \mathbb{R}^N$ with $\sum x_i = \sum y_i$, x 'Lorenz-dominates' y if rewriting their coordinates in increasing order, and denoting the results $\tilde{x}$ and $\tilde{y}$, we have $\tilde{x}_1 \geq \tilde{y}_1$, $\tilde{x}_1 + \tilde{x}_2 \geq \tilde{y}_1 + \tilde{y}_2$, and so on, with at least one strict inequality holding.)

The mappings defined next, whose focus is on rewarding individual players, what we called 'player centric,' are *single valued*. Most of them are based on the concept of a *player's contribution to a coalition*. It will be convenient to use the expression in two different ways depending upon whether the player is a member of the coalition or not. Let $i \in N$ and $S \subseteq N$. If $i \in S$, it is the difference $v(S) - v(S \backslash \{i\})$; if not, it is the difference $v(S \cup \{i\}) - v(S)$. *Player i's principal contribution* is his contribution to the grand coalition, $M_i(v) \equiv v(N) - v(N \backslash \{i\})$.

- Equality and proportionality are two polar notions in almost all branches of the theory of economic justice. Implementing the first notion requires that a baseline be chosen and implementing the second one that a direction also be chosen. As we see, several options are possible. We start with simple egalitarian ideas – one can almost say 'simplistic' – and make successive adjustments so as to better achieve our goal of assigning to each player what he deserves. At first, the baseline for each player is a payoff of 0.
- The *plain egalitarian value* (sometimes called the 'egalitarian value'; Radzik (2013) calls it the 'equal split value' and van den Brink (2001) calls it the 'equal division solution') divides the worth of the grand coalition equally among all players: for each $i \in N$, $x_i = \frac{v(N)}{|N|}$. Because it ignores the worths of other coalitions, this value seems far from achieving our goal.

In the next two definitions, baselines are derived from the game.

The *egalitarian-over-individual-worths value* (sometimes called the CIS value; the name the 'egalitarian value' is also used, for

instance, by Chun and Park (2011), while van den Brink (2001) calls it the 'equal surplus division solution') first assigns to each player his individual worth, then divides what remains of $v(N)$ equally among all players: for each $i \in N$,

$$x_i = v(\{i\}) + \frac{v(N) - \sum_N v(\{j\})}{|N|}$$

This value is a little more sensitive to the data defining the game but not much; it depends on only $n-1$ of its $2^n - 1$ coordinates.

The *egalitarian-over-principal-contributions value* (sometimes called the 'egalitarian value'; when applied to cost allocation, it is called 'equal division over non-separable costs') first assigns to each player his principal contribution, then divides what remains of $v(N)$ equally among all players: for each $i \in N$,

$$x_i = v(N) - v(N \setminus \{i\}) + \frac{v(N) - \sum [v(N) - v(N \setminus \{j\})]}{|N|}$$

In economics, rewarding factors on the basis of their marginal product is a well-established idea. However, its application here delivers a solution that, once again, does not seem responsive enough to the data defining a game; it also ignores most of its coordinates.

- Imagining that players enter the game one at a time, in some order, and assigning to each his contribution when he arrives generates a plausible value provided one can justify the order. No such justification is given in the description of a game but the idea will nevertheless be a central ingredient in several important definitions below, so a formal introduction is needed.

Let $\mathcal{O}^N$ be the set of strict orders on N. Given $\prec \in \mathcal{O}^N$, the *sequential contribution value associated with* $\prec$, $SC^{\prec}$ (sometimes called 'marginal values') assigns to each $i \in N$ his contribution to the coalition of all the players who are listed before him in the order: $SC_i^{\prec}(v) \equiv v(\{j : j \prec i\} \cup \{i\}) - v(\{j : j \prec i\})$.

How much a player is assigned by a sequential contribution value obviously depends on where he stands in the order and on the game: in general, there is no relation between a player's position and his payoff. (For special classes of games, such as convex games, this is not the case however.) To introduce some fairness into the process, averaging over all orders is natural. The *Shapley value Sh* (Shapley, 1953) which occupies a central place in the literature, can be interpreted in this way, as the simple average of the sequential contributions values calculated under the assumption that all orders occur with equal probabilities, namely $Sh(v) \equiv \frac{1}{2^{|N|}} [\sum_{\prec \in O^N} SC^{\prec}(v)]$. Collecting terms (all of the orders in which the same players precede player i produce the same contribution for him), we obtain the better-known formula:

$$Sh_i(v) \equiv \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)]$$

.

Other formulas for the Shapley value are due to Kleinberg and Weiss (1985), Rothblum (1985), and Béal et al. (2012b).

The proposal has also been made to work with the convex hull of the sequential contributions values. The result is the *Weber solution* (Weber, 1988).

- We now turn to notions of proportionality. Here, when distributing payoffs, in addition to the question 'Where from' we should ask 'To where?' A baseline should still be specified for each player from which to measure his gain, but also an amount to which he may 'aspire.' The values defined next provide simultaneous answers to these questions. We could imagine others.

The inspiration for the *τ-value* (Tijs, 1981) is a solution for bargaining games known as the Kalai–Smorodinsky solution (Kalai and Smorodinsky, 1975). Recall that $M_i(v)$ designates player i's principal contribution in v, and let $m_i(v) \equiv \max_{S \subset N \setminus \{i\}} [v(S) - \sum_S M_j(v)]$ and $m(v) \equiv (m_i(v))_{i \in N}$. Then, on the domain of games for which $m(v) \leq M(v)$ and $\sum m_i(v) \leq v(N) \leq \sum M_i(v)$ (Tijs calls these games 'quasi-balanced') $\tau(v)$ is the efficient vector on the segment connecting $m(v)$ and $M(v)$. A generalization of the definition to the domain of all games is due to Driessen and Tijs (1981).

Questions to Be Addressed

What issues can be raised when studying solutions and values?

Nonemptiness

First is nonemptiness. Many values are given by an explicit formula, so the issue of existence does not arise for them. Examples are the Shapley value, its variants, and all of the other player-centric values that are based on the random arrival scenario that underlies the Shapley value but involve alternative distribution schemes.

Several important solutions are only defined indirectly. Proving that they are well defined, or identifying classes of games for which they are, has filled the agenda of researchers for many years.

This is the case for the core, for the various notions of bargaining sets, and for the von-Neumann–Morgenstern solution. For the core, a necessary and sufficient condition has been identified that guarantee nonemptiness. Keeping fixed the worths of all the coalitions $S \subset N$, an explicit lower bound on $v(N)$ can be given as a function of the former. This is the well-known Bondareva–Shapley theorem. First define a *balanced pair* as a list $\mathcal{S}$ of coalitions and a list of weights $(\lambda_s)_{S \in \mathcal{S}}$ for them, such that for each $i \in N$, the sum of the weights assigned to those coalitions in $\mathcal{S}$ that contain player i is equal to one: $\sum_{S \in \mathcal{S}: i \in S} \lambda_S = 1$. Let $\mathcal{B}$ denote the class of all such pairs. The lower bound on $v(N)$ for $C(v)$ to be nonempty is the maximum over all $(\mathcal{S}, (\lambda_s)_{S \in \mathcal{S}}) \in \mathcal{B}$ of the weighted sum of the worths of the coalitions appearing in the list: $\max_{(\mathcal{S}, (\lambda_s)_{S \in \mathcal{S}}) \in \mathcal{B}} \sum_{S \in \mathcal{S}} \lambda_S v(S)$. Let us call it the *Bondareva–Shapley lower bound*. The inequality turns out to be sufficient. A game for which it is met is *balanced*.

Theorem 1 (Bondareva, 1963; Shapley, 1967) *A game has a nonempty core if and only if the worth of the grand coalition meets the Bondareva–Shapley lower bound.*

For the von-Neumann–Morgenstern, it took many years for the issue of existence to be understood. In fact, it is not generally guaranteed (Lucas and Rabie, 1982). For the various notions of bargaining sets, for the prekernel, an important proof technique is based on the facts that the prenucleolus belongs to it, and that the prenucleolus is well defined (Schmeidler, 1969).

Some solutions are defined by means of maximization exercises. Showing the existence of solutions may be straightforward, but solving these exercises may be computationally complex. This is true for the nucleolus, which is in general notoriously difficult to calculate. Simpler algorithms may be available for certain classes of games, however. An explicit formula can be derived for the nucleolus for the simple cost allocation problem known as the airport problem (Littlechild, 1974). Convex games (Arin and Inarra, 1988) and assignment games (Solymosi and Raghavan, 1994) also lend themselves to explicit calculations. There are also domains of games on which a particular solution or value may coincide with one that is usually easy to calculate. We identify below a domain for which the nucleolus and the Shapley value coincide.

Accessibility

For a solution whose nonemptiness has been established, the follow-up question arises of how to reach at least one of the payoff vectors it recommends, for each game, is important in practice. We speak of the *accessibility* of the solution then. Much is known about the core in this regard. When the core is nonempty, it can be reached through a step-by-step process that starts at an arbitrary payoff vector and involves satisfying blocking coalitions as they emerge. Much is also known about how many steps are needed (Sengupta, A. and Sengupta, K., 1996; Kóczy, 2006; Béal et al., 2012a; Yang, 2010, 2011).

Extensions

When a solution is empty for a class of games of interest, the question arises of extending its definition to the class so as to recover existence. For an extension to be meaningful, it should preserve the spirit of the original definition, it should coincide with the solution whenever the solution is nonempty, and preferably, the extension should be minimal. For the core, several extensions have been proposed. Given $\epsilon > 0$, the *ϵ-core* of a game consists of all its efficient vectors at which none of the blocking constraints is violated by more than ϵ. Its *least core* is defined to be its ϵ-core for the smallest value of ϵ that guarantees nonemptiness. Recent definitions are due to Bejan and Gómez (2009) and Derks et al. (2013).

When the core is empty, at least one coalition has to be assigned in total less than its worth. One could argue that fairness across coalitions requires that all coalitions be so treated. An imputation belongs to the *anticore* of v if for each $S \subset N$, $\sum x_i \leq v(S)$. A game can be such that both its core and its anticore are empty however.

The *positive core* (Orshan and Sudhölter, 2010) is a combination of the core and the prenucleolus. It coincides with the core on the domain of TU games whose core is nonempty.

The *Lorenz stable set*, the *leximin stable allocation*, and the *egalitarian core* (Arin and Inarra, 2001) are extensions of the Dutta–Ray value to the entire domain of games whose core is nonempty. (Recall that this value is defined on the class of convex games.) The *equal split-off set* (Branzei et al., 2006) is a further extension, to the domain of all TU games.

Selections

When a solution is nonempty, the question of selection arises. To the extent that eventually, only one payoff vector will have to be chosen from the set of vectors that the solution recommends, identifying a most desirable one in the set is important.

The *core center* (González-Díaz and Sánchez-Rodríguez, 2007) selects the center of gravity of the core.

Given a game whose core is nonempty and an order of players, consider the following procedure: give the first player the most he can get within the core; give the second player the most he can get within the subset of the core of payoff vectors at which the first player gets the amount just calculated; give the third player the most he can get within the subset of the core of payoff vectors at which the first and second players get the amounts calculated above, and so on. The *average lexicograhic value*, *Alex* (Tijs, 2005; Tijs et al., 2011) picks the average of the payoff vectors obtained this way for each order of players.

Solution-Specific Structural Issues

For each particular solution and value, specific issues may arise concerning the algebraic and geometric properties of the set of payoff vectors it selects, and how this set responds to changes in games. An example of such a result, which concerns the core, is a decomposition: a positive game with a nonempty core can be written as a positive linear combination of veto-controlled simple games (Derks, 1987). These components have a clear economic interpretation.

For a particular solution, special games may come up in its study. A game is *exact* (Schmeidler, 1972) if for each $S \subset N$, there is x in the core of v such that $\sum_S x_i = v(S)$.

The *upper core* of a game is the set of payoff vectors that violate all of the core constraints. The core of a game is *large* (Sharkey, 1982) if for each vector y in its upper core, there is a core vector that is dominated by y (Moulin, 1990; Azrieli and Lehrer, 2007; Getán and Montes, 2010; Estévez-Fernández, 2012).

Certain operations on games preserve the nonemptiness of a particular solution. An example is increasing the worth of the grand coalition. Van Gellekom et al. (1999) refer to a property that is preserved by such an operation as a *prosperity property*. Super-additivity and convexity are examples. Having a nonempty core is obviously a prosperity property. It is unknown whether the core being a von-Neumann–Morgenstern solution is a prosperity property (Van Gellekom et al., 1999).

The space of games being a convex space, games can be added. An interesting question is under what conditions applying a solution to a sum of games is the same thing as applying the solution to each of them and adding the results. This question has been asked and answered about the core (Bloch and de Clippel, 2011, elaborating on Tijs and Branzei, 2002). Essentially, the extreme points should be determined

by the same sets of coalitions. The condition is satisfied in particular on the domain of convex games.

Variants of the Central Solutions

Variants of several of the main solutions have been defined.

The *equal division core*, C_{ed} (Selten, 1972) is obtained by restricting the possible divisions of a coalitional worth: objections by a coalition take the form of an equal division of its worth. (Given $w \in \Delta^N$, the *equity core relative to w* (Selten, 1978) is the set of vectors whose coordinates add up to the worth of the grand coalition to which no coalition can successfully object by means of a division of its worth among its members proportional to the components of w pertaining to them.) The equal-division core contains the core and its nonemptiness that is guaranteed under much weaker conditions. It suffices that the worth of the grand coalition be at least as large as the sum of the worths of the components of any partition (this is a weak form of super-additivity). This condition is not necessary however (Llerena and Vilella, 2013).

The *per capita nucleolus* and *per capita prenucleolus* (Grotte, 1970) are defined similar to the nucleolus and the prenucleolus, but using the per capita dissatisfactions instead of the dissatisfactions themselves.

The *modiclus* (Sudhölter, 1997; Raghavan and Sudhölter, 2005) is defined as the prenucleolus but using the differences of excesses instead of the excesses themselves.

The *least-square prenucleolus* (Ruiz et al., 1996) selects the pre-imputation at which the function $x \in \mathbb{R}^N \to \sum_{S \subset N} [v(S) - \sum_{j \in S} x_j]^2$ is minimized.

For each payoff configuration $x = (x^S)_{S \subseteq N}$, each pair of coalitions (S, T) such that $S \subset T$ and $|T| = |S| + 1$, and each $i \in S$, consider the difference $x_i^T - x_i^S$. The *monoclus* (Slikker and Norde, 2011) is a payoff configuration value that selects the payoff configuration at which the list of these differences is lexicographically minimized over the set of those payoff configurations that associate each coalition with an imputation for the corresponding subgame. The *premonoclus* is defined in a similar way but the minimization is carried out within the set of those payoff configurations that associate each coalition with a pre-imputation for the corresponding subgame.

Relating Concepts

How two solutions are related is an important issue. The Shapley value is always well defined, but a game may have a nonempty core and yet not contain its Shapley value. Hence, an interesting question is for what kinds of games does the Shapley value belongs to the core. One answer is the domain of convex games. In fact, the Shapley value has a simple expression then. We have already noted that the core of a convex game is a polytope whose corners are the vectors of sequential contributions. Thus, for convex games, the Shapley value is the geometric average of the corners of the core.

There are other games with the property that the Shapley value belongs to the core. Examples are the average convex games. This is shown by Sprumont (1990), Inarra and Usategui (1993), and Driessen (1998). Izawa and Takahashi (1998) and Marin-Solano and Rafels (1999) are other studies of this class of games.

Given the completely different ways in which the Shapley value and the nucleolus are defined, one would not expect them to often make the same recommendations. However, interesting classes of games have been identified for which they do (Deng and Papadimitriou, 1994; Kar et al., 2009; Chun and Hokari, 2007; Chang and Tseng, 2011). In fact, for convex games, many solution concepts are related that are not related in general.

When the core is nonempty, the nucleolus is a selection from it but two games may have the same core and yet have different nucleoli. For certain classes of games, this *core-only* property is satisfied however.

The Weber solution – recall that it is defined as the convex hull of the sequential-contributions values – contains the core, and for a convex game, it is equal to the core (Shapley, 1971).

When is the core the unique von-Neumann–Morgenstern stable set? An extensive and difficult literature has been devoted to the understanding of this question. A sufficient condition is that the core be large (Sharkey, 1982). This condition is also necessary for symmetric games (Biswas et al., 2000). Further work on this issue is by Van Gellekom et al. (1999) and Shellshear and Sudhölter (2009).

Our final result relates the core and the classical bargaining set. Keeping fixed the worths of all coalitions except that of the grand coalition, there is a number such that, as soon as the worth of that coalition is at least as large as that, the Davis–Maschler bargaining set of the resulting game equals the core (Izquierdo and Rafels, 2012).

Axioms

An axiom is the mathematical expression of the intuition we have about how a solution should behave in certain situations. A *punctual axiom* applies to each game separately, and a *relational axiom* relates choices made for games that are related in a certain way. We start with punctual axioms. For such an axiom, the only issue is existence: for each game, are there payoff vectors satisfying the axiom?

Basic Axioms

Let φ be a solution, $v, w \in \mathcal{G}^N$.

- *Efficiency* says that the sum of the payoffs is equal to the worth of the grand coalition.
- Some games have a trivial structure, and only one outcome seems natural for them. The following axioms pertain to such situations:

The *trivial-game axiom* says that if $v = 0$, then $\varphi(v) = 0$. Generalizing the idea slightly, we obtain the *inessential game axiom*, which says that if there is $x \in \mathbb{R}^N$ such that for each $S \subseteq N$, $v(S) = \sum_S x_i$, then $\varphi(v) = x$.

- There are other situations in which how much should be assigned to a player seems unequivocal.

The *dummy player axiom* says that if a player's contribution to each coalition is 0 – he is a *dummy player in the game* – he

gets 0. The idea expressed next is similar. The *null player axiom* says that if a player's contribution to each coalition is equal to his own worth – he is a *null player in the game* – he gets his own worth.

Similar requirements can be formulated for coalitions: if a coalition does not add anything to the worth of any other disjoint coalition by joining it beyond its own worth, its members do not get in total more than that amount.

The *nullifying player axiom* (van den Brink, 2007) says that if a player is such that the worth of each coalition containing him is 0 – he is a *nullifying player in the game* – then he gets 0.

- Next are bounds on assignments. Lower bounds may come to mind first. The *minimal-contribution lower bound* (Kikuta, 1976; Funaki, 1986; Gérard-Varet and Zamir, 1987) for a player is the smallest of his contributions to any coalition.

Upper bounds can be defined in a symmetric way. The *maximal-contribution upper bound* (Milnor, 1952) for a player is the largest of his contributions to any coalition.

Gérard-Varet and Zamir (1987) argue that unless a game is super-additive, the *maximal-contribution upper bound* may not be that relevant. They suggest that instead, the bound be calculated for the smallest super-additive game that dominates v, its *super-additive cover* (this concept is well defined).

- Next is a minimal punctual fairness axiom. First, *two players are symmetric in a game* if they contribute the same amount to each coalition to which neither belongs: for each $S \in N \backslash \{i, j\}$, $v(S \cup \{i\}) - v(S) = v(S \cup \{j\}) - v(S)$. Then, *symmetry* says that if two players are symmetric in a game, their payoffs are equal.

All of the solutions that we have defined are *symmetric* except of course for the sequential-contributions values. *Symmetry* appears to be a basic requirement but it is not always desirable. Two players who enter symmetrically in a game may differ in dimensions that are not explicitly incorporated in the model, and solutions and values that do not necessarily treat symmetric players symmetrically provide us with useful flexibility. Such mappings can often be obtained by modifying a standard *symmetric* definition through the introduction of weights assigned to players. A calibration step allows then a user of the theory to identify the weights that are the most desirable in his application. For some values, the parametrization is more complex. These extensions are for the Shapley value (Shapley, 1953; Kalai and Samet, 1987; Haeringer, 2006), the prenucleolus (Derks and Haller, 1999; Orshan and Sudhölter, 2012), the Dutta–Ray value (Hokari, 2000, 2002), the prekernel (Orshan and Sudhölter, 2012), and the solidarity value (Calvo and Gutiérrez, 2012).

- Next, we associate with each game an order relation on the player set, reflecting his 'desirability as a partner,' and require that the value should respect it. *Player i is at least as desirable as player j in v* if for each $S \in N \backslash \{i, j\}$, $v(S \cup \{i\}) - v(S) \geq v(S \cup \{j\}) - v(S)$; he is *more desirable* if in addition, one of these inequalities is strict. A solution *respects the desirability relation* (Maschler and Peleg, 1966), if whenever player i is at least as desirable as player j in v, his payoff is at least as large as player j's.

Similarly, "a coalition S can be defined to be at least as desirable as a coalition S' in a game" (Lapidot, 1968), and "a value can be said to respect the desirability relation for coalitions" if under these circumstances, it assigns to S in total at least as much as what it assigns to S' in total. The desirability relation for individual players is transitive but incomplete. The desirability relation for coalitions is neither transitive nor complete. Circumstances under which it is complete for simple games are studied in Einy (1985).

A related notion is the following: *player i is at least as desirable on average as player j in v* if $\sum_{S \subset N \backslash \{i,j\}} [v(S \cup \{i\}) - v(S)] \geq \sum_{S \subset N \backslash \{i,j\}} [v(S \cup \{j\}) - v(S)]$. This relation is both complete and transitive. The value φ *respects the average desirability relation* if in the circumstances just described, player i's payoff is at least as large as player j's: $\varphi_i(v) \geq \varphi_j(v)$ (Kultti and Salonen, 2007).

- When utilities are interpreted as von Neumann–Morgenstern utilities, the solution should be independent of the choice of origin of the chosen representative in the family. *Strategic equivalence* says that adding, for each player $i \in N$, an arbitrary constant $a_i \in \mathbb{R} > 0$ to his utility, and correspondingly adjusting the worths of all coalitions (so that for each $S \subseteq N$, $v(S)$ is replaced by $v(S) + \sum_S a_i$), results in a new game whose solution outcome is obtained from the solution outcome of the original game by adding to each player i's payoff the amount a_i.

A version of the axiom can be written for coalitions. *Coalitional strategic equivalence* (Chun, 1989) says that adding to the worth of each coalition containing some coalition $T \subseteq N$ the same number should not affect the payoffs to the players not in T. (*Strategic equivalence* is the special case obtained by repeated application for $|T| = 1$.)

- A game is a point in a Euclidean space. Thus, one can talk about the convergence of a sequence of games. *Continuity* says that the solution outcome of the limit game is the limit of the sequence of solution outcomes of the games in the sequence.

Almost all of the values that we have defined are *continuous*. These facts are easy to prove for those that are given by explicit combinatorial expressions (the Shapley value, for example). For some other values, proofs of *continuity* require work. Continuity notions can also be defined for solutions.

Monotonicity Axioms

A monotonicity axiom expresses the requirement that a player's payoff, or the payoffs of the members of some group, should move in a particular direction, as one parameter of the game to be solved moves in a particular way. (The hypotheses may involve several parameters changing together.) It may be felt that a player is entitled to at least as much as what he was assigned initially; or that each player in a particular group is entitled to at least as much as he was assigned initially, or that in the aggregate, the players in some group are entitled to at least as much as they were assigned initially. Due to the richness of the description of a game, these axioms come in a great variety of forms.

- The simplest parameter change we consider is an increase in the worth of the grand coalition. *Grand-coalition-worth monotonicity* (Megiddo, 1974) says that if the worth of the grand coalition increases, the payoff of no player decreases.

Many values satisfy this property. However, as soon as there are four players, the nucleolus violates it even on the domain of convex games (Hokari, 2000). Characterizations of selections from the core that satisfy *grand-coalition monotonicity* and other basic axioms are obtained by Calleja et al. (2013).

Individual-worth monotonicity says that if the worth of player i increases, his payoff does not decrease.

One can go further: *general individual–worth monotonicity* (Young, 1985) says that if the worths of all the coalitions containing player i weakly increase, then player i's payoff does not decrease.

Next, we consider changes in the worth of an arbitrary coalition and formulate two parallel axioms. *Coalitional-worth monotonicity* (Shubik, 1962) says that if the worth of a coalition increases, then the payoff of no member of that coalition decreases.

General coalitional-worth monotonicity says that if v and v' differ only in that for some $S \subseteq N$, and for each $\tilde{S} \supset S$, $v'(\tilde{S}) \geq v(\tilde{S})$, then the payoff of no member of S decreases.

Contribution monotonicity (Young, 1985) says that if the contribution vector of a particular player weakly increases, then his payoff should not decrease. *Weak contribution monotonicity* (Hokari and van Gellekom, 2002) says that if the contribution vectors of each player weakly increases, then the payoff of no player should decrease.

We have the following impossibility:

Theorem 2 (Young, 1985) *For games with at least five players, no value that selects from the core is coalitional-worth monotonic.*

Theorem 2 holds on the domain of super-additive games. However, on the domain of convex games, the weighted Shapley values and the weighted Dutta–Ray values are *coalitional-worth monotonic* (Hokari, 2002).

The following weaker property can be met more generally. The hypotheses are the same, but the conclusion pertains to the aggregate payoff of the members of the coalition whose worth increases. *Weak coalitional-worth monotonicity* (Zhou, 1991) says that under the hypotheses of *coalitional-worth monotonicity*, the sum of the payoffs of the members of S should not decrease.

The nucleolus satisfies *weak coalitional-worth monotonicity*. (As it is a selection from the core, by Theorem 2, it violates *coalitional-worth monotonicity*.)

When the worth of a coalition, which is not the grand coalition, increases, we may be interested in the impact of the change on the payoffs of the members of the complementary coalition, and demand that none of them should increase.

Further studies of monotonicity properties are by Arin and Feltkamp (2012). These authors formulate various strengthenings and conditional strengthenings of certain of these properties and they study their implications on the domain of games with nonempty cores and that have veto players.

- *Fair ranking* (Chun, 1989) says that if the worth of a coalition changes, there should be no reversal in the relative payoffs of two members of that coalition.

Independence and Invariance Axioms

Next are independence and invariance properties. The first one asserts the primacy of the concept of a contribution. As such, it is of particular significance to economists. *Own-contributions–only* (Young, 1985) says that if the contribution vectors of a given player to two games are the same, then his payoffs should be the same in both games. This axiom is implied by *contribution monotonicity*.

- Consider now a group of players who are involved in two separate games. The games may be solved separately, or they may be consolidated into one (by addition). Because it may be difficult to make the case than one way to proceed is better than another, if a solution is such that one player benefits from the consolidation whereas another player is hurt by it, it is less likely that a consensus will emerge as to what to do. The next axiom says that it should not make any difference to any player. *Additivity* (Shapley, 1953) says that the payoff vector selected for the sum of two games should be equal to the sum of the payoff vectors selected for each of them.

The next requirement applies to solutions: it is that players should have more options when they consolidate two games into one. *Super-additivity* (Peleg, 1986) says that the payoff vectors selected for the sum of two games should be obtained by adding payoff vectors chosen for each of the games.

- A *partnership in* v is a set of players S that affect the worth of a coalition only if all of them are included: if a coalition R does not contain all the members of S, its worth is equal to the worth of the coalition consisting of the players in R that do not belong to S. The *partnership axiom* (Kalai and Samet, 1987) says that if S is a partnership in v, the payoff to each member of S in v can be calculated by applying the value to the game in which the worth of each coalition containing S is equal to the aggregate payoff to the members of S, and the worth of each other coalition is equal to 0 (this is the multiple of the unanimity game associated with S scaled by $\sum_S \varphi_i(v)$).
- The *reciprocal impact axiom* (Myerson, 1977) says that the difference between player i's payoffs when everyone is present and when player j is not (then the game that is solved is the projection of the initial game onto the set $N \backslash \{j\}$) is equal to the difference between player j's payoffs when everyone is present and when player i is not.
- *Rationalizability by a potential* (Hart and Mas-Colell, 1989) says that there is a real-valued function P defined on $\mathcal{V}^N$ such that for each $v \in \mathcal{V}^N$ and each $i \in N$, $\varphi_i(v) = P(v) - P(v_{N \backslash \{i\}})$. This function is the *potential for the value*.

Some Characterizations

One of the foundation stones of the theory is the following:

Theorem 3 (Shapley, 1953) *The Shapley value is the only value satisfying efficiency, the null player axiom, symmetry, and additivity.*

Theorem 3 holds on the domain of super-additive games, and on the domain of simple games. The Shapley value also emerges on the domain of monotonic simple games and on the domain of simple super-additive games if *additivity* is

replaced by a variant (Dubey, 1975). This result is generalized by Einy (1988).

The weighted Shapley values have been characterized along the same lines but without *symmetry* (Kalai and Samet, 1987). The implications of dropping *efficiency* in Theorem 3 are studied by Dubey et al. (1981).

A related characterization is obtained by replacing the last two axioms of Theorem 3 by the requirement that adding to a game v a game v' in which two players are symmetric results in a new game $v + v'$ in which these two players' payoffs change by equal amounts (van den Brink, 2001). Although Theorem 3 does not extend to the domain of simple games (Dubey, 1975), this variant does.

A characterization of the values satisfying the axioms of Theorem 3 except for the *null player axiom* is due to Chameni Nembua (2012). Included are the plain egalitarian value and the solidarity value.

The critical relational axiom in the following result is *own-contributions-only*. Compared to Theorem 3, note that here the *null player axiom* is not listed.

Theorem 4 (Young, 1985) *The Shapley value is the only value satisfying efficiency, symmetry, and own-contributions-only*.

Theorem 4 holds on the domain of super-additive games (Young, 1985), on the domain of convex games (Hokari and van Gellekom, 2002), and on the domain of constant-sum games (Khmemlnistkaya, 2003). Both Theorems 3 and 4 hold on any domain consisting of a game and all the sums of integer multiples of its subgames extended to the whole player set by making the players who are not directly involved null players (Neyman, 1989).

Symmetry is not listed in the next result, but some symmetry is embedded in *fair ranking*.

Theorem 5 (Chun, 1989) *The Shapley value is the only value satisfying efficiency, the trivial-game axiom, coalitional strategic equivalence, and fair ranking*.

Theorem 5 also holds on the domain of super-additive games and on the domain of convex games.

The next characterization shows the strength of the *reciprocal impact axiom*, as *efficiency* is the only other axiom it involves.

Theorem 6 (Myerson, 1977) *The Shapley value is the only value satisfying efficiency and the reciprocal impact axiom*.

A parallel result holds when the *reciprocal impact axiom* is replaced by an axiom of *reciprocal impact axiom in subgames*. For player i, we calculate the difference between his payoff when everyone is present and his payoff when player j is not, and the game that is played is defined by setting the worth of each coalition equal to the difference between the worth of S together with the complement of S. We require equality of 'action and reaction.'

Our final characterization, of the Shapley value, is based on the concept of a potential.

Theorem 7 (Hart and Mas-Colell, 1989) *The Shapley value is the only value to admit a potential*.

All of the characterizations we presented are of the Shapley value and variants but more recently, other values have emerged.

Characterizations of the τ-value are due to Tijs (1987) and Calvo et al. (1995).

Variable-Population Axioms

We now turn to axioms pertaining to possible variations in population size. Accommodating such variations requires a different formalism. We imagine that there is a set of potential players, which we take to be the natural numbers. In each game, only finitely many of these players are present however. Let $\mathcal{N}$ be the class of finite subsets of $\mathbb{N}$. A value is then defined on $\cup_{N \in \mathcal{N}} \mathcal{G}^N$ and it takes its values in $\cup_{N \in \mathcal{N}} \mathbb{R}^N$.

- The first axiom is the most straightforward:

The *null player out* axiom says that the departure of a null player should leave unaffected the payoffs of the remaining players.

Notions of replications have been important in general equilibrium theory as a first step toward modeling situations with a large number of agents none of whom holds significant market power. To apply the idea to coalitional games, the first question is how to replicate a game. Several definitions are possible. Here is one that models minimal compatibility of interests between the original players and newcomers. In a *conflictual replicated game* (Thomson, 1986; Kalai and Samet; 1987), the worth of a coalition is equal to the worth of the set of players (or 'types'), in the original game, whose clones are all included in the coalition. The following axiom is not a property of a single solution but a property of a family of solutions consisting of a symmetric solution and its asymmetric generalizations.

- *Replication invariance* says that in a conflictual replicated game, the sum of what the players of each type obtain, under an application of a given symmetric solution, should be equal to what the player they are representing would obtain in the game that is replicated under an application of the asymmetric generalization of the solution with weights proportional to the orders of replication of the different types.
- To define the next property, we need to define the idea of two players merging into one and the converse operation, the idea of one player appearing under two identities. Suppose that players i and j merge into one player. Let S be a coalition that contains neither i nor j. Then, there are several ways to define the worth of the coalition consisting of S and the single player that represents i and j. It could be just the worth of $S \cup \{i, j\}$ in the original game, or it could be the maximum of the worths of $S \cup \{i\}$, $S \cup \{j\}$, and $S \cup \{i, j\}$.

Splitting proofness and *merging proofness* (Lehrer, 1988; Derks and Haller, 1999; Derks and Tijs, 2000; Casajus, 2012) says that when two players merge into one, the sum of what they obtain in the original game is equal to what the single player representing them would obtain in the new game.

- The next property pertains to situations in which some players may leave or some players may arrive. It was first considered in situations in which a social endowment of resources has to be distributed and each agent has monotonic preferences with respect to how much he consumes. In that context, a natural solidarity property is that, for example if new agents arrive, all agents initially present should be affected negatively. Given the generality of the type of situations covered by the model under study here, it

is not necessarily the case that when a game is extended to accommodate new players, opportunities would evolve in a way that would require sacrifices from the existing players. In fact, if there are sufficient synergies between them and the new players, the opposite may be true. This is the case that has mainly be studied. How much of a synergy is needed for it to be possible for all of the initial players to benefit? Convexity is enough.

Population monotonicity (Thomson, 1983) says that when a game is extended so as to accommodate new players, the payoff of none of the players initially present should decrease.

On the class of convex games, the Shapley value, the Dutta–Ray value, their generalizations called the monotone-path Dutta–Ray values are *population monotonic* (Hokari, 2002).

Instead of looking for solutions that respond monotonically to changes in the player set, one may inquire about the existence of payoff vectors for a given game that are related by inequalities to payoff vectors that could be chosen for subgames. A *population-monotonic payoff configuration for* $v \in \mathcal{G}^N$ (Sprumont, 1990) is a list of payoff vectors indexed by the subsets of N, $(x_S)_{S \subseteq N}$, such that for each $S \subseteq N$, $\sum_{i \in S} x_{iS} = v(S)$.

One may think of the list as the values taken by a *population-monotonic* solution defined on a very limited domain, namely the game v itself and all of its subgames. It is obvious that for a game to admit a *population-monotonic* payoff configuration, it should have a nonempty core and so should all of its subgames. Recall that such games are said to be totally balanced. This is not enough. Here is the complete characterization of these games:

Theorem 8 (Sprumont, 1990) *A game admits a population-monotonic payoff configuration if and only if it is totally balanced and it is the sum of an inessential game and a positive linear combination of monotonic simple games each of which has at least one veto player.*

Other relevant contributions here are by Moulin (1990) and Getán and Montes (2010). The monoclus (Silikker and Norde, 2011) is an example of payoff configuration value that picks a *population-monotonic* payoff configuration whenever the latter exists.

- Informally speaking, the 'consistency principle' says the following. Consider a game and identify a payoff vector chosen for it by a solution. Imagine now that some players leave with their payoffs and reassess the situation at this point; here, this means assess the opportunities open to each subgroup of the remaining players. The requirement is that in the resulting 'reduced game,' the solution should select a payoff vector at which each of the remaining players is assigned the same payoff as initially. Many scenarios have been imagined when defining the 'reduction operation.' We present the three central ones. A complete list is given in Thomson (2013).

For our first definition, given $N \in \mathcal{N}$, $v \in \mathcal{G}^N$, $x \in \varphi(v)$, and finally $N' \subset N$, we imagine that each coalition $S \subset N'$ is *required to get together with all the members of* $N \backslash N'$ *and pay them their components of* x. Then, what is left for S is the difference $v(S \cup (N \backslash N')) - \sum_{N \backslash N'} x_i$. We define this difference to be its revised worth.

Complement consistency: For each pair N, N' of elements of $\mathcal{N}$ with $N' \subset N$, each $v \in \mathcal{G}^N$, and each $x \in \varphi(v)$, we have $r^x_{N'}(v) \in \mathcal{G}^{N'}$ and $x_{N'} \in \varphi(r^x_{N'}(v))$, where $r^x_{N'}(v)$, the 'complement reduced game of v relative to N' and x,' is defined by setting, for each $S \subseteq N'$ (Moulin actually defines $r^x_{N'}(v)(N') \equiv \sum_{N'} x_i$; however, in the presence of *Pareto-optimality*, stated below, this definition is equivalent to that of ours), $r^x_{N'}(v)(S) \equiv v(S \cup (N \backslash N')) - \sum_{N \backslash N'} x_i$.

Second, given a proposed payoff vector $x \in \mathbb{R}^N$, in the reduced game of v relative to $N' \subset N$ and x, the worth of a coalition $S \subset N'$ is computed under the assumption that S can secure the cooperation of any subgroup S' of $N \backslash N'$, provided each member of S' receives his component of x. After these payments are made, what remains for S is the difference $v(S \cup S') - \sum_{S'} x_i$. Maximizing behavior on the part of S involves finding $S' \subseteq N \backslash N'$ for which this difference is maximal.

Max consistency: The definition differs from that of *complement consistency* in that the 'max reduced game of v relative to N' and x,' is defined by setting $r^x_{N'}(v)(N') \equiv v(N) - \sum_{N/N'} x_i$, and for each $S \subset N'$, $r^x_{N'}(v)(S) \equiv \max_{S' \subseteq N \backslash N'} \{v(S \cup S') - \sum_{S'} x_i\}$.

Our third main definition of a consistency notion is as follows. Let φ be a *single-valued* solution and let $v \in \mathcal{G}^N$. In calculating the worth of a coalition $S \subseteq N'$ in the reduced game of v relative to $N' \subset N$ *and the solution* φ, we require S to cooperate with the complement of N' (as in the definition of *complement consistency*); this cooperation is assumed to be feasible provided the members of $N \backslash N'$ are paid what φ would recommend for them in the subgame of v faced by the group $S \cup (N \backslash N')$. That is, we imagine the disappearance pure and simple of the group $N' \backslash S$ and we ask what should S receive in total in the subgame $v_{S \cup (N \backslash N')}$? The suggested answer is the difference $v(S \cup (N \backslash N')) - \sum_{N \backslash N'} \varphi_i(v|_{S \cup (N \backslash N')})$. For this calculation to be meaningful, φ should be *single valued*, which is why we limited our attention to that case.

Self-consistency: For each pair N, N' of elements of $\mathcal{N}$ with $N' \subset N$, and each $v \in \mathcal{G}^N$, we have $r^\varphi_{N'}(v) \in \mathcal{G}^N$ and $x_{N'} = \varphi(r^\varphi_{N'}(v))$, where $r^\varphi_{N'}(v)$, the 'self-reduced game of v relative to N' and φ,' is defined by setting, for each $S \subseteq N'$, $r^\varphi_{N'}(v)(S) \equiv v(S \cup (N \backslash N')) - \sum_{N \backslash N'} \varphi_i(v|_{S \cup (N \backslash N')})$.

On the domain of convex games, a number of interesting logical relations hold between these properties and others (Hokari and van Gellekom, 2002).

Theorem 9 (Tadenuma, 1992) *On the domain of TU games whose core is nonempty, the core is the only solution satisfying individual rationality and complement consistency.* (It may be puzzling at *Pareto-optimality* does not appear in this theorem. This is because *individual rationality* can be seen as a 'one-player' optimality property. This property is lifted (the concept of lifting is described in the section?) by the solution from one-player games (which are in the domain). Lifting also occurs with some of the alternative definitions of *consistency* discussed later.)

Theorem 10 (Sobolev, 1975) *On the domain of all TU games, the prenucleolus is the only single-valued solution satisfying Pareto-optimality, anonymity, invariance under common rescaling of utilities, and max consistency.*

(Sobolev does not have an optimality property as he uses a slightly different definition of the reduced game.)

Orshan (1994) and Hokari (2005) investigate the implications of dropping *anonymity*.

Theorem 11 (Peleg, 1986) *On the domain of TU games whose core is nonempty, the core is the only solution satisfying individual rationality, super-additivity, and max consistency.*

Theorem 12 (Hart and Mas-Colell, 1989) *The Shapley value is the only value satisfying Pareto-optimality, symmetry, invariance under rescaling of utilities, and self-consistency.*

Non-transferable Utility Games

A *non-transferable utility game*, an *NTU game* in short, is a list of sets $V \equiv (V(S))_{S \subseteq N}$, where for each $S \subseteq N$, $V(S) \subset \mathbb{R}^S$ represents the set of payoff vectors achievable by S. A TU game can be represented as an NTU game in which each $V(S)$ is a hyperplane normal to a vector of ones. A *hyperplane game* is an NTU game such that each $V(S)$ is defined by a hyperplane. Here, the normal vector can be any positive vector. The notion of a *convex game* has been generalized in several different ways. In a *bargaining game*, there exists a 'disagreement point' $d \in V(N)$ such that for each $S \neq N$, $V(S)$ is the comprehensive hull of d_S.

Most of the solutions and values that were first defined for TU games have been extended to domains of NTU games, but that has often caused complications, both conceptual and technical. A victim has often been *single valuedness*. Also, characterizations obtained for the TU case have not always been extendable to the NTU case. Let $\mathcal{W}^N$ be the domain of all NTU games. Let $V \in \mathcal{W}^N$.

- The *core* selects the set of vectors $x \in \mathbb{R}^N$ such that $x \in V(N)$ and there is no $S \subseteq N$ and $y \in V(S)$ such that $y \geq x_S$. A counterpart of the condition that was shown to be necessary and sufficient for the core of a TU game to be nonempty can be formulated, but here it is only sufficient (Scarf, 1967). A generalization is due to Billera (1970). A domain of convex games is covered by Greenberg (1985). A necessary and sufficient condition for nonemptiness is given by Predtetchinski and Herings (2004).

Several characterizations of the core have been developed. Keiding's (1986) is based on *anti-monotonicity*: given two games V and V' such that for each $S \subset N$, $V'(S) \subseteq V(S)$, and $V'(N) = V(N)$, then $\varphi(V') \subseteq \varphi(V)$. Other characterizations by Peleg (1985) and Tadenuma (1992) are based on NTU versions of *max consistency* and *complement consistency*, respectively.

Recall that a 'core extension' is a solution that coincides with the core when the core is nonempty. Such an extension is proposed and characterized by Keiding and Pankratova (2010).

- The *bargaining set* is defined in the same way as in the TU case. A sufficient condition for nonemptiness of the bargaining set of NTU games is due to Vohra (1991). This condition is not necessary (Chang and Lee, 1993).
- Shapley's objective in his fundamental paper (1969) was to extend his TU value to the NTU case. In order to solve an NTU game, his strategy was as follows: (1) to associate with it a TU game by supposing utility transfers to be feasible at some rates, (2) to solve this TU game by applying the Shapley value, and (3) to verify that the solution outcome of the TU game is feasible in the original game without in fact performing any utility transfers. Here is the formal definition of the *NTU Shapley solution*. Let $\lambda \in \mathbb{R}^N_{++}$ and v_λ be the TU game defined by setting for each $S \subseteq N$, $v_\lambda(S) := \max\{\lambda \cdot x | x \in V(S)\}$. Then the solution selects for V those payoff vectors x for which there is $\lambda \in \mathbb{R}^N_{++}$ such that for each $i \in N$, $\lambda_i x_i = Sh_i(v_\lambda)$.

A characterization of the NTU Shapley solution is developed by Aumann (1985) based on a list of axioms that includes the next two:

Conditional additivity says that if x and y are two efficient vectors chosen for each of two NTU games, and their sum $x + y$ is efficient for the sum of the two games, then it should be chosen for that game.

Contraction independence says that if x is chosen for a game and the feasible set for the grand coalition shrinks the feasible sets for other coalitions remaining unchanged, and x is still feasible for the grand coalition, then x should be chosen for the new game.

The question had been left open whether *contraction independence* was independent of the other axioms. The answer is negative (Peleg et al., 2012). A family of weighted NTU Shapley solutions is defined and characterized by Levy and McLean (1991).

- The *Kalai–Samet payoff-configuration value* (Kalai and Samet, 1985) selects a payoff configuration $(x^S)_{S \in 2^N}$ that is constructed recursively as follows: For each $i \in N$, let $t^{\{i\}}$ be the maximal payoff that he can obtain in $V(\{i\})$ and $x^{\{i\}} \equiv t^{\{i\}}$. For each two-player coalition $\{i, j\}$, let $x^{\{i,j\}} \equiv (t^{\{i\}} + t^{\{i,j\}}, t^{\{j\}} + t^{\{i,j\}})$, where $t^{\{i,j\}}$ is chosen so as to achieve efficiency in $V(\{i, j\})$. For each three-player coalition $\{i, j, k\}$, let $x^{\{i,j,k\}} \equiv (t^{\{i\}} + t^{\{i,j\}} + t^{\{i,k\}} + t^{\{i,j,k\}}, t^{\{j\}} + t^{\{i,j\}} + t^{\{j,k\}} + t^{\{i,j,k\}}, t^{\{k\}} + t^{\{i,k\}} + t^{\{j,k\}} + t^{\{i,j,k\}})$, where $t^{\{i,j,k\}}$ is chosen so as to achieve efficiency in $V(\{i, j, k\})$. For larger coalitions S, $x^S \in \mathbb{R}^S$ is defined in a similar manner.

Weighted versions of the solution can be defined: Given a weights vector $\lambda \in \mathbb{R}^N_{++}$, the associated payoff-configuration solution picks the payoff-configuration constructed as follows: For each $i \in N$, let $x^{\{i\}}$ be the maximal payoff that he can obtain in $V(\{i\})$ and $t^{\{i\}} \equiv \lambda_i x^{\{i\}}$. For each two-player coalition $\{i, j\}$, let

$$x^{\{i,j\}} \equiv \left(\frac{t^{\{i\}} + t^{\{i,j\}}}{\lambda_i}, \frac{t^{\{j\}} + t^{\{i,j\}}}{\lambda_j} \right)$$

where $t^{\{i,j\}}$ is chosen so as to achieve efficiency in $V(\{i, j\})$. For each three-player coalition $\{i, j, k\}$, let

$$x^{\{i,j,k\}} \equiv \left(\frac{t^{\{i\}} + t^{\{i,j\}} + t^{\{i,k\}} + t^{\{i,j,k\}}}{\lambda_i}, \frac{t^{\{j\}} + t^{\{i,j\}} + t^{\{j,k\}} + t^{\{i,j,k\}}}{\lambda_j}, \frac{t^{\{k\}} + t^{\{i,k\}} + t^{\{j,k\}} + t^{\{i,j,k\}}}{\lambda_k} \right)$$

where $t^{\{i,j,k\}}$ is chosen so as to achieve efficiency in $V(\{i, j, k\})$. For larger coalitions S, $x^S \in \mathbb{R}^S$ is chosen in a similar manner.

Characterizations of the Kalai–Samet value and of their weighted versions are available (Kalai and Samet, 1985; Hart and Mas-Colell, 1989; Chun, 2004; de Clippel et al., 2004). Kalai and Samet's (1985) characterization is based on the

following direct generalization of *coalitional-worth-monotonicity* for the TU case:

Coalitional monotonicity says that if $V(T) \subset V'(T)$ and for all $S \neq T$, $V(S) = V'(S)$, then $\varphi_T(V) \leq \varphi_T(V')$.

Hart and Mas-Colell's (1989) characterization is based on an NTU version of *self-consistency*.

- The *Harsanyi payoff-configuration solution* selects those payoff configurations $(\gamma^S)_{S \in 2^N}$ for which there is $\lambda \in \mathbb{R}^N_{++}$ such that γ^N maximizes $\sum_{i \in N} \lambda_i x_i$ over $V(N)$, and $(\gamma^S)_{S \in 2^N}$ is selected by the corresponding weighted Kalai–Samet payoff-configuration value.

A characterization of the Harsanyi payoff-configuration solution is also available (Hart, 1985). Remarkably, it only involves rewriting the axioms imposed by Aumann (1985) in his characterization of the NTU Shapley solution so that they be applicable to payoff-configuration solutions.

The *Harsanyi–Shapley solution* is defined as the intersection of the Harnanyi solution and the NTU Shapley solution (Chang and Hwang, 2001, 2003).

- The *Maschler–Owen value for hyperplane games.* Given a hyperplane game H, the Maschler–Owen value selects the average of the vectors of contributions corresponding to the various orders in which players may arrive. Since $H(N)$ is a hyperplane, this average is feasible and efficient.

Results on hyperplane games are due to Maschler and Owen (1989), Owen (2004), and Bhattacharya and Ziad (2006).

The *Maschler–Owen solution* is defined recursively by constructing hyperplane games and applying the Maschler–Owen value to them:

1. For each $i \in N$, let $H(\{i\}) := V(\{i\})$.
2. Let $S \subseteq N$ and suppose that for each $T \subset S$, $H(T)$ has been defined. Then $H(S)$ is chosen so that the Maschler–Owen value for the resulting hyperplane game for S is feasible and efficient for $V(S)$.

Characterizations of the Maschler–Owen solution are also available (de Clippel et al., 2004; Hart, 2005).

- The *compromise value* selects the maximal point on the line passing through two points derived from the game, a 'utopia' vector and a derived vector of lower bounds (Borm et al., 1992). It is defined on the subdomain of games for which the latter vector is dominated by the former, the latter belongs to the feasible set of the grand coalition and the former is not dominated by any point in this set.

It can be seen as an extension of the τ-value to NTU games as well as an extension of the Kalai–Smorodinsky (KS) solution for bargaining games.

Characterizations of the compromise value can be obtained by extending the characterizations of the τ-value and KS solution (Borm et al., 1992; Otten et al., 1996).

- The *MC-value for monotonic NTU games* (Otten et al., 1998) selects the efficient point on the line through the origin (games are 0-normalized), and the sum of the vectors of contributions corresponding to the various orders in which players may arrive.

The value coincides with the Shapley value on the subdomain of TU games, with the KS solution on the subdomain of bargaining games, and with the Owen–Maschler value on the subdomain of hyperplane games.

Two axiomatic characterizations of the MC value are available (Otten et al., 1998).

- The *procedural value* (de Clippel, 2007) is defined recursively by constructing bargaining games and applying the Raiffa solution to them. For each two-player coalition, the Raiffa solution is applied. For each three-player coalition, let a_i be the maximal payoff for player i in $V(\{i, j, k\})$ given that other players j and k get what the Raiffa solution selects for $V(\{j, k\})$. Then the Raiffa solution is applied to select a payoff vector from $V(\{i, j, k\})$ assuming that the disagreement point is the simple average of a_i, a_j, and a_k. The bargaining games for larger coalitions are defined in a similar manner.

The resulting value coincides with the Shapley value on the subdomain of TU games, with the Raiffa solution on the subdomain of bargaining games, and with the Owen–Maschler value on the subdomain of hyperplane games.

Extensions of the Base TU Model

Various enrichments of the base TU model have been formulated and for many of these extensions, a number of the solutions and values developed for the base model have been extended and often characterized along the lines of characterizations they had been given for the base model.

- In *games with a coalition structure* a partition of the player set – the coalition structure – is added to the model (Aumann and Drèze, 1975). Payoffs can be obtained by means of a two-step process, by first inducing a game among the components of the partition, and then dividing what each component has been assigned among its members. Four proposals have been made, based on the Shapley value and the Banzhaff value. They differ in the order in which the two values are applied. A characterization of a value due to Owen has been developed (Hamiache, 1999, 2006).

Studies of this class of games are numerous. A generalization of the model is when an ordered list of coalition structures is added, each element in the list refining the previous one, with the finest possible partition (which consists of all one-player coalitions) and the coarsest partition (which consists of the grand coalition only) being included. These are *games with levels of participation* (Winter, 1989).

- In a *game in partition function form*, the worth of a coalition depends on how the complementary coalition is partitioned (Lucas and Thrall, 1977). Indeed, what it can achieve may depend on the activities undertaken by the other groups that have formed. Such a game is sometimes called a *game with externalities* (de Clippel and Serrano, 2008; Dutta et al., 2010). Characterizations of a version of the Shapley value for this domain of games along the lines of Theorem 3 are

due to Myerson (1977) and Pham Do and Norde (2007). The consensus value has also been adapted to this setting and a characterization developed by Y. Ju (2007).

- A *game with a graph structure* includes a graph structure on the player set. A coalition can obtain its worth as specified in the game only if its members form a connected component of the graph. A characterization of a variant of the Shapley value is due to Myerson (1977). A *tree game* is a special case (Maschler et al., 2010; Mishra and Talman, 2010).
- For a *game with precedence constraints* (Faigle and Kern, 1992; Grabisch and Sudhölter, 2012), there is a partial order on the set of coalitions. A coalition is feasible only if it contains all players preceding a member of it according to the partial order.
- In a *game with a permission structure*, the player set is equipped with an additional structure meant to capture restrictions on who can associate with whom (Gilles et al., 1992). It takes the form of a correspondence from the set of players into itself that is required to satisfy certain properties: it associates with each player a set of 'successors.' This information can be used in several ways. One possibility is that a coalition forms only if each of its members (except for the 'top player') has at least one predecessor who belongs to it. Alternatively, a coalition forms only if each of the predecessors of each of its members belong to it. van den Brink (2010) adapts the Banzhaf value to this setting and develops an axiomatization of it.
- In a *game with restricted cooperation* (the expression 'partially defined games' has also been used) not all coalitions can form, so the specification of a game includes only the worths of a subclass of subsets of the player set. Characterizations of versions of the Shapley value and of the core on the basis of, mainly, a linearity axiom in the first case and a consistency axiom in the second case, have been developed (Faigle, 1989).
- In a *game with upper bounds*, an upper bound on how much certain coalitions can be assigned in total is specified. A notion of 'truncated core' has been defined for such games and characterized along the lines of Theorem 11 (Carpente et al., 2010).
- In a *multi-choice game*, each player can participate at a finite number of different levels (Hsiao and Raghavan, 1993). The notion of a *bi-cooperative game* is closely related: each player has two options, to participate in a positive way or to participate in a negative way (Bilbao et al., 2008). In a *fuzzy game*, each player's participation level is a real number between 0 (non-cooperation) and 1 (full cooperation). A fuzzy coalition is a profile of players' participation levels. A fuzzy game specifies for each fuzzy coalition its worth (Hwang, 2007; J. Liu, and X. Liu, 2012).
- In a *dynamic TU games* (Kranich et al., 2005), players face a finite sequence of exogenously specified TU games. Each player has a time-separable utility function defined over sequences of payoffs. There are several ways to define a core-like solution for this model.

The final class of games is defined so as to incorporate uncertainty.

- For an *interval games*, the worth of a coalition, instead of being a well-defined number, is specified as an interval in the real line, no probabilistic information being given on the relative likelihood of each point in the interval (Alparslan-Gök et al., 2009).

Acknowledgments

This work was supported by NSF under grant SBR-9731431 and SES-0214691.

Bibliography

Aumann, R.J., 1985. An axiomatization of the non-transferable utility value. Econometrica 48, 599–612.
Aumann, R.J., Drèze, J., 1975. Cooperative games with coalition structures. International Journal of Game Theory 3, 217–237.
Aumann, R.J., Maschler, M., 1964. The bargaining set for cooperative games. In: Dresher, M., Shapley, L.S., Tucker, A.W. (Eds.), Advances in Game Theory. Princeton University Press, pp. 443–477.
Bejan, C., Gómez, J., 2009. Core extensions for non-balanced TU-games. International Journal of Game Theory 38, 3–16.
Bondareva, O.N., 1963. Some applications of linear programming methods to the theory of cooperative games (in Russian). Problemy kibernetiki 10, 119–139.
Chun, Y., 1989. A new axiomatization of the Shapley value. Games and Economic Behavior 1, 119–130.
Chun, Y., 2004. On weighted Kalai-Samet solutions for non-transferable utility coalitional games. Games and Economic Behavior 47, 257–267.
Chun, Y., Park, B., 2011. Population solidarity, population fair-ranking, and the egalitarian value. International Journal of Game Theory 41, 255–270.
de Clippel, G., Peters, H., Zank, H., 2004. Axiomatizing the Harsanyi solution, the symmetric egalitarian solution and the consistent solution for NTU-games. International Journal of Game Theory 33, 145–158.
Dubey, P., 1975. On the uniqueness of the Shapley value. International Journal of Game Theory 4, 131–139.
Gillies, D.B., 1953. Some Theorems on n-Person Games. Ph.D. Dissertation. Princeton University.
Hart, S., 1985. An axiomatization of Harsanyi's nontransferable utility solution. Econometrica 53, 1295–1450.
Hart, S., 2005. An axiomatization of the consistent non-transferable utility value. International Journal of Game Theory 33, 355–366.
Hart, S., Mas-Colell, A., 1989. Potential, value and consistency. Econometrica 57, 589–614.
Hokari, T., 2000a. The nucleolus is not aggregate-monotonic on the class of convex games. International Journal of Game Theory 29, 133–137.
Hokari, T., 2000b. Population monotonic solutions on convex games. International Journal of Game Theory 29, 327–338.
Hokari, T., 2005. Consistency implies equal treatment in TU-games. Games and Economic Behavior 51, 63–82.
Hokari, T., van Gellekom, A., 2002. Population monotonicity and consistency in convex games: some logical relations. International Journal of Game Theory 31, 593–607.
Kalai, E., Samet, D., 1985. Monotonic solutions to general cooperative games. Econometrica 53, 307–327.
Kleinberg, N., Weiss, J., 1985. A new formula for the Shapley value. Economics Letters 17, 311–315.
Maschler, M., Owen, G., 1989. The consistent Shapley-value for hyperplane games. International Journal of Game Theory 18, 390–407.
Maschler, M., Owen, G., 1992. The consistent Shapley value for games without side-payments. In: Selten, R. (Ed.), Rational Interactions. Springer-Verlag, pp. 5–12.
Megiddo, N., 1974. On the non monotonicity of the bargaining set, the kernel and the nucleolus of a game. SIAM Journal of Applied Mathematics 27, 355–358.
Peleg, B., 1985. An axiomatization of the core of cooperative games without side-payments. Journal of Mathematical Economics 14, 203–214.
Peleg, B., 1986. On the reduced game property and its converse. International Journal of Game Theory 15, 187–200.
Peleg, B., Sudhölter, P., 2007. Introduction to the Theory of Cooperative Games, second ed. Springer.
Peters, H., 2008. Game Theory: A Multi-Leveled Approach. Springer.
Schmeidler, D., 1969. The nucleolus of a characteristic function games. SIAM Journal on Applied Mathematics 17, 1163–1170.

Shapley, L.S., 1953. A value for n-person games. In: Kuhn, H., Tucker, A.W. (Eds.), Contributions to the Theory of Games, vol. 2, pp. 307–317.

Shapley, L.S., 1967. On balanced sets and cores. Naval Research Logistics Quarterly 14, 453–460.

Shapley, L.S., 1969. Utility comparisons and the theory of games. In: La Decision. Editions du CNRS, Paris, pp. 251–263. Reprinted in A.E. Roth, 1988. The Shapley Value, Cambridge University Press, pp. 307–319.

Shapley, L.S., 1971. Cores of convex games. International Journal of Game Theory 1, 11–26.

Sobolev, A.I., 1975. Characterization of the principle of optimality for cooperative games through functional equations (in Russian). Mathematical Methods in the Social Sciences 6, 150–165.

Sprumont, Y., 1990. Population-monotonic allocation schemes for cooperative games with transferable utility. Games and Economic Behavior 2, 378–394.

Tadenuma, K., 1992. Reduced games, consistency, and the core. International Journal of Game Theory 20, 325–334.

Thomson, W., 1983. The fair division of a fixed supply among a growing population. Mathematics of Operation Research 8, 319–326.

Tijs, S., 1981. Bounds for the core and the τ-value. In: Moeschlin, O., Pallaschke, D. (Eds.), Game Theory and Mathematical Economics. North-Holland, pp. 123–132.

Tijs, S., 1987. An axiomatization of the τ-value. Mathematical Social Sciences 13, 177–181.

Weber, R., 1994. Games in coalitional form. In: Aumann, R.J., Hart, S. (Eds.), Handbook of Game Theory, vol. II. Elsevier, Amsterdam.

Young, P., 1985. Monotonic solutions for cooperative games. International Journal of Game Theory 14, 65–72.

Zhou, L., 1991. A weak monotonicity property of the nucleolus. International Journal of Game Theory 19, 407–411.

Cooperative Learning in Schools

Robert E Slavin, Johns Hopkins University, Baltimore, MD, USA; and University of York, York, England

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.E. Slavin, E.A. Hurley and A.M. Chamberlain, volume 4, pp. 2756–2761, © 2001, Elsevier Ltd.

Abstract

In cooperative learning, students work in small groups to help one another learn academic objectives. This article describes the theoretical basis for cooperative learning, and the evidence base for common pragmatic approaches. It concludes that cooperative learning works best for increasing achievement when student teams have group goals they can only achieve if all group members are gaining in learning, so that they focus on teaching and learning from each other rather than merely completing a group task.

Cooperative learning refers to instructional methods teachers use to organize students into small groups, in which students work together to help one another learn academic content. Cooperative learning methods are extensively researched, and under certain well-specified conditions they are known to substantially improve student achievement in most subjects and grade levels. Yet the structured forms of cooperative learning that have proven to be effective are not used as often as more informal forms. Further, there remains considerable debate about the theoretical basis for achievement outcomes of cooperative learning. This article reviews and describes widely used, practical forms of cooperative learning, and presents evidence on their effects on academic achievement.

Cooperative learning methods vary widely in their details. Group sizes may be from two to several. Group members may have individual roles or tasks, or they may all have the same task. Groups may be evaluated or rewarded based on group performance or the average of individual performances, or they may simply be asked to work together.

In one form or another, cooperative learning has been used and studied in every major subject, with students from preschool to college, and in all types of schools. There have been hundreds of studies of cooperative learning focusing on a wide variety of outcomes, including academic achievement in many subjects, second language learning, attendance, behavior, intergroup relations, social cohesion, acceptance of classmates with handicaps, attitudes toward subjects, and more (see Johnson and Johnson, 1998; Rohrbeck et al., 2003; Slavin, 1995; Webb, 2008).

Theoretical Perspectives on Cooperative Learning

Although there is a fair consensus among researchers about the positive effects of cooperative learning on student achievement, discussed in this article, there remains a controversy about why and how cooperative learning methods affect achievement and, most importantly, under what conditions cooperative learning has these effects. Different groups of researchers investigating cooperative learning effects on achievement begin with different assumptions and conclude by explaining the achievement effects of cooperative learning in quite different theoretical terms. In earlier work, Slavin (1995) identified motivationalist, social cohesion, cognitive-developmental, and cognitive-elaboration as the four major theoretical perspectives on the achievement effects of cooperative learning.

The motivationalist perspective presumes that task motivation is the single most impactful part of the learning process, asserting that the other processes such as planning and helping are driven by individuals' motivated self-interest. Motivationalist-oriented scholars focus more on the reward or goal structure under which students operate (Slavin, 1995, 2009). Methods derived from this perspective emphasize the use of group goals and individual accountability, meaning that group success depends on the individual learning of all group members (see below). By contrast, the social cohesion perspective (also called social interdependence theory) suggests that the effects of cooperative learning are largely dependent on the cohesiveness of the group. This perspective holds that students help each other learn because they care about the group and its members and come to derive self-identity benefits from group membership (Johnson and Johnson, 1998).

The two cognitive perspectives on cooperative learning focus on the interactions among groups of students, holding that, in themselves, these interactions lead to better learning and thus better achievement.

The alternative perspectives on cooperative learning may be seen as complementary, not contradictory. For example, motivational theorists would not argue that the cognitive theories are unnecessary. Instead, they assert that motivation drives cognitive process, which in turn produces learning (Slavin, 1995). They would argue that it is unlikely over the long haul that students would engage in the kind of explanations to each other found by Webb (2008) and others to be essential to profiting from cooperative activity unless the learning of their teammates is important to them. Similarly, social cohesion theorists might hold that the utility of extrinsic incentives must lie in their contribution to group cohesiveness, caring, and prosocial norms among group members, which could in turn affect cognitive processes.

A simple path model of cooperative learning processes, adapted from Slavin (1995), is diagrammed below. It depicts the main functional relationships among the major theoretical approaches to cooperative learning.

Figure 1 begins with a focus on group goals or incentives based on the individual learning of all group members. That

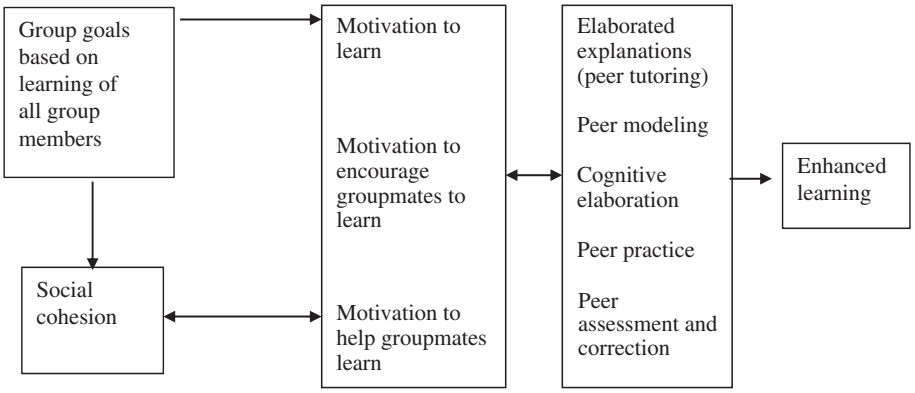

Figure 1 Integration of theoretical perspectives on cooperative learning effects on learning.

is, the model assumes that motivation to learn and to encourage and help others to learn activates cooperative behaviors that will result in learning. This would include both task motivation and motivation to interact in the group. In this model, motivation to succeed leads to learning directly, and also drives the behaviors and attitudes that lead to group cohesion, which in turn facilitates the types of group interactions that yield enhanced learning and academic achievement. The relationships are conceived to be reciprocal, such that task motivation leads to the development of group cohesion, which may reinforce and enhance task motivation. By the same token, the cognitive processes may become intrinsically rewarding and lead to increased task motivation and group cohesion.

Group Goals and Individual Accountability

Considerable evidence from practical applications of cooperative learning in elementary and secondary schools supports the position that group rewards are essential to the effectiveness of cooperative learning, with one critical qualification. Use of group goals or group rewards enhances the achievement outcomes of cooperative learning, if and only if the group rewards are based on the individual learning of all group members (Slavin, 1995). Most often, this means that team scores are computed based on average scores on quizzes, which all teammates take individually without teammate help. For example, in Student Teams-Achievement Divisions, or STAD (Slavin, 1994), students work in mixed-ability teams to master material initially presented by the teacher. Following this, students take individual quizzes on the material, and the teams may earn certificates based on the degree to which team members have improved over their own past records. The only way the team can succeed is to ensure that all team members have learned, so the team members' activities focus on explaining concepts to one another, helping one another practice, and encouraging one another to achieve. In contrast, if group rewards are given based on a single group product (e.g., the team completes one worksheet or solves one problem), there is little incentive for group members to explain concepts to one another, and one or two group members may do all the work (see Slavin, 1995, 2009).

A review of 99 studies of cooperative learning in elementary and secondary schools that involved durations of at least 4 weeks compared achievement gains in cooperative learning and control groups (Slavin, 1995). Of 64 studies of cooperative learning methods that provided group rewards based on the sum of group members' individual learning, 50 studies (78%) found significantly positive effects on achievement, and none found negative effects (Slavin, 1995). The median effect size for these 64 studies was $d = +0.32$ (32% of a standard deviation separated cooperative learning and control treatments).

In contrast, studies of methods that used group goals based on a single group product or provided no group rewards found few positive effects, with a median effect size of only $d = +0.07$. Comparisons of forms of cooperative learning with and without group rewards within the same studies found similar patterns; group goals based on the sum of individual learning performances were necessary to the instructional effectiveness of the cooperative learning models (e.g., Fantuzzo et al., 1990; Fantuzzo et al., 1989).

Why are group goals and individual accountability so important? To understand this, consider the alternatives. In some forms of cooperative learning, students work together to complete a single worksheet or to solve one problem. In such methods, there is little reason for more able students to take the time to explain what is going on to their less able groupmates or to ask their opinions. When the group task is to *do* something, rather than to *learn* something, the participation of less able students may be seen as interference rather than help. It may be easier in this circumstance for students to give each other answers than to explain concepts or skills to one another. More aggressive students may dominate the group, and others may avoid participating, letting others do the work (and the learning).

When the group's task is to ensure that every group member *learns* something, it is in the interest of every group member to spend time explaining concepts to his or her groupmates, and to ask groupmates for explanations and help in understanding the topic of study. Studies of student behavior within cooperative groups have found that the students who gain most from cooperative work are those who give and receive elaborated explanations (Webb, 1985, 2008). In contrast, giving and receiving answers without explanations were *negatively*

related to achievement gain. Group goals and individual accountability motivate students to give elaborated explanations and to take one another's learning seriously, instead of simply giving answers.

Research on Alternative Pragmatic Approaches to Cooperative Learning

Research and development over the years have led to the creation and evaluation of several practical approaches to cooperative learning. The most widely used and extensively researched of these programs are described in the following sections. These sections discuss cooperative learning studies that met the standards of the Best Evidence Encyclopedia (BEE; www.bestevidence.org) in reading and math. Inclusion in the BEE reviews requires a well-matched control group, a duration of at least 12 weeks, and measures not inherent to the treatment (see Slavin, 2008). Mean effect sizes are computed weighting by sample size. Full reports of the BEE reviews are published for elementary reading by Slavin et al. (2009b), for secondary reading by Slavin et al. (2008), for elementary math by Slavin and Lake (2008), and for secondary math by Slavin et al. (2009a).

Cooperative learning methods fall into two main categories. One set, *Structured Team Learning*, involves rewards to teams based on the learning progress of their members, and individual accountability, which means that team success depends on individual learning, not group products. A second set, *Informal Group Learning Methods*, includes methods more focused on social dynamics, projects, and discussion than on mastery of well-specified content.

Structured Team Learning Methods

Student Team Learning

Student Team Learning (STL) techniques were developed and researched at Johns Hopkins University (see Slavin, 1994, 1995). More than half of all experimental studies of practical cooperative learning methods involve STL methods.

All cooperative learning methods share the idea that students work together to learn and are responsible for one another's learning as well as their own. STL methods also emphasize the use of team goals and team success, which can only be achieved if all members of the team learn the objectives being taught.

Two concepts are central to all STL methods: *team rewards* and *individual accountability*. Using STL techniques, teams earn certificates or other team rewards if they achieve above a designated criterion. *Individual accountability* means that the team's success depends on the individual learning of all team members. This focuses the activity of the team members on explaining concepts to one another and making sure that everyone on the team is ready for a quiz or other assessment that they will take without teammate help.

Four principal Student Team Learning methods have been extensively developed and researched. Two are general cooperative learning methods adaptable to most subjects and grade levels: STAD and Teams–Games–Tournament (TGT). The remaining two are comprehensive curriculums designed for use in particular subjects at particular grade levels: Team-Assisted Individualization (TAI) for mathematics in years 3–6 and Cooperative Integrated Reading and Composition (CIRC) for reading and writing instruction in grades 3–5. Middle school adaptations of CIRC are called Student Team Reading and The Reading Edge.

Student Teams-Achievement Divisions

In STAD (Slavin, 1994), students are assigned to four-member learning teams mixed in performance level, sex and ethnicity. The teacher presents a lesson, and the students work within their teams to make sure that all team members have mastered the lesson. Finally, all students take individual quizzes on the material, at which time they may *not* help one another.

Students' quiz scores are compared to their own past averages, and points are awarded based on the degree to which students meet or exceed their own earlier performances. These points are then summed to form team scores, and teams that meet certain criteria earn certificates or other rewards. The whole cycle of activities, from teacher presentation to team practice to quiz, usually takes three to five class periods.

STAD has been used in a wide variety of subjects, including mathematics, language arts, and social studies. It has been used from grade 2 through college. STAD is most appropriate for teaching well-defined objectives, such as mathematical computations and applications, language usage and mechanics, geography and map skills, and science facts and concepts.

Numerous studies of STAD have found positive effects of the program on traditional learning outcomes in math, language arts, science, and other subjects (Barbato, 2000; Mevarech, 1985; Reid, 1992; Slavin, 1995; Slavin and Karweit, 1984). Across 11 comparisons that met BEE inclusion criteria, nine of which used random assignment to conditions, the sample-size-weighted effect size was +0.14. These studies involved a total of more than 4000 students in grades 3–12.

Teams–Games–Tournament

TGT (Slavin, 1994) uses the same teacher presentations and teamwork as in STAD, but replaces the quizzes with weekly tournaments. In these, students compete with members of other teams to contribute points to their team score. Students compete at three-person tournament tables against others with a similar past record in mathematics. Table assignments rotate to keep the competition fair. The winner at each tournament table brings the same number of points to his or her team, regardless of which table it is; this means that low achievers (competing with other low achievers) and high achievers (competing with other high achievers) have equal opportunity for success. As in STAD, high-performing teams earn certificates or other forms of team rewards. TGT is appropriate for the same types of objectives as STAD. Several studies of TGT have found positive effects on achievement in math, science, and language arts (Slavin, 1995).

Team-Assisted Individualization

TAI (Slavin et al., 1986) shares with STAD and TGT the use of four-member mixed-ability learning teams and certificates for high-performing teams. However, where STAD and TGT use a single pace of instruction for the class, TAI combines cooperative learning with individualized instruction. Also, where STAD and TGT apply to most subjects at grade levels, TAI is specifically designed to teach mathematics to students in grades 3–6 (or older students not ready for a full algebra course). Across five comparisons (2 randomized) involving almost 3000 students, the sample-size-weighted effect size for TAI was +0.19.

Cooperative Integrated Reading and Composition

A comprehensive program for teaching reading and writing in the upper-elementary grades is called CIRC (Stevens et al., 1987). In CIRC, teachers use reading texts and reading groups, much as in traditional reading programs. However, all students are assigned to teams composed of two pairs from two different reading groups. While the teacher is working with one reading group, the paired students in the other groups are working on a series of cognitively engaging activities, including reading to one another, making predictions about how narrative stories will come out, summarizing stories to one another, writing responses to stories, and practicing spelling, decoding, and vocabulary. Students work as a total team to master main idea and other comprehension skills. During language arts periods, students engage in writing drafts, revising and editing one another's work, and preparing for publications of team books.

In most CIRC activities, students follow a sequence of teacher instruction, team practice, team preassessments, and quizzes. That is, students do not take the quiz until their teammates have determined that they are ready. Certificates are given to teams based on the average performance of all team members on all reading and writing activities.

Research on CIRC and similar approaches has found positive effects in upper-elementary and middle school reading (Stevens and Durkin, 1992; Stevens et al., 1987). CIRC has been adapted as the upper-elementary and middle school component of the Success for All comprehensive reform model and is currently disseminated under the name *Reading Wings* by the Success for All Foundation. Across 13 studies (one of which was randomized) involving more than 14 000 students in grades 1–8, the weighted mean effect size was +0.27.

Peer-Assisted Learning Strategies

Peer-Assisted Learning Strategies (PALS) is a dyadic learning approach in which pairs of children take turns as teacher and learner. The children are taught simple strategies for helping each other, and are rewarded based on the learning of both members of the pair. Research on PALS in elementary and middle school math and reading has found positive effects of this approach on student achievement outcomes (e.g., Calhoon, 2005; Calhoon et al., 2006; Fuchs et al., 2001; Fuchs et al., 1999; Mathes and Babyak, 2001). Positive effects of a similar program called Classwide Peer Tutoring (Greenwood et al., 1989) have also been found, and another similar approach has been found to be effective in two Belgian studies (Van Keer and Verhenge, 2005, 2008). Across 11 reading studies, five of which used random assignment, the weighted effect size was +0.30. There were a total of about 1700 students across the studies. Three randomized studies of PALS in math, involving more than 600 students, found a weighted mean effect size of +0.09.

IMPROVE

IMPROVE (Mevarech, 1985) is an Israeli mathematics program that uses cooperative learning strategies similar to those used in *STAD* but also emphasizes teaching of metacognitive skills and regular assessments of mastery of key concepts and reteaching of skills missed by many students. Studies of *IMPROVE* have found positive effects on the mathematics achievement of elementary and middle school students in Israel (Mevarech and Kramarski, 1997; Kramarski et al., 2001). Across three studies (one randomized) involving almost 700 Israeli middle schoolers, the weighted mean effect size was +0.52.

Informal Group Learning Methods

Jigsaw

Jigsaw was originally designed by Elliot Aronson et al. (1978). Slavin (1994) developed a modification of Jigsaw at Johns Hopkins University and then incorporated it in the STL program. In this method, called Jigsaw II, students work in four- or five-member teams as in TGT and STAD. Instead of each student being assigned a particular section of text, all students read a common narrative, such as a book chapter, a short story, or a biography. However, each student receives a topic (such as 'climate' in a unit on France) on which to become an 'expert.' Students with the same topics meet in expert groups to discuss them, after which they return to their teams to teach what they have learned to their teammates. Then students take individual quizzes, which result in team scores based on the improvement score system of STAD. Teams that meet preset standards earn certificates. Jigsaw is primarily used in social studies and other subjects where learning from text is important (Mattingly and Van Sickle, 1991).

Learning Together

David Johnson and Roger Johnson at the University of Minnesota developed the Learning Together models of cooperative learning (Johnson and Johnson, 1998). The methods they have researched involve students working on assignment sheets in four- or five-member heterogeneous groups. The groups hand in a single sheet and receive praise and rewards based on the group product. Their methods emphasize team-building activities before students begin working together and regular discussions within groups about how well they are working together. Numerous relatively brief experiments have shown positive effects of these approaches (see Johnson and Johnson, 1998).

Group Investigation

Group Investigation, developed by Shlomo Sharan and Yael Sharan (1992) at the University of Tel-Aviv, is a general classroom organization plan in which students work in small groups using cooperative inquiry, group discussion, and cooperative planning and projects. In this method, students form their own two- to six-member groups. After choosing subtopics from a unit being studied by the entire class, the groups further break their subtopics into individual tasks and carry out the activities necessary to prepare group reports. Each group then makes a presentation or display to communicate its findings to the entire class. A study in Israel by Sharan and Shachar (1988) found positive effects of Group Investigation on achievement in language and literature.

Directions for Further Research and Development

Cooperative learning occupies a strange place in educational research and practice. On one hand, it is universally known and almost universally admired. Most researchers and educators have positive attitudes toward cooperative learning and believe it to be effective for many outcomes (see Antil et al., 1998). Yet the forms of cooperative learning that have been found to be effective, especially for academic achievement outcomes, remain at the edge of practice, perhaps in a state of permanent innovation. That is, cooperative learning has never disappeared but has never become common practice.

More research is needed on how and why cooperative learning works, and how it may be made to work better. Studies comparing conditions under which dyads may be more or less effective than groups of four might be interesting, and much more needs to be known about specific means of helping students learn and use effective groupwork strategies and metacognitive learning strategies. Further explorations are needed to understand how and under what conditions group goals and individual accountability affect learning outcomes of cooperative learning. There is a need to develop approaches integrating cooperative learning activities with various types of technology, both classroom technologies such as interactive whiteboards and traditional computers.

Although there is already a great deal of research on cooperative learning, there is still much room for theoretical and practical advances. Cooperative learning is so different from ordinary teaching that it opens a vast set of questions and possibilities that are far from being exhausted.

Conclusion

Research on cooperative learning over a 40-year period has found that under a set of well-defined circumstances, students working in structured small groups can learn significantly better than can students working in traditional classrooms. Positive learning outcomes depend on the use of programs in which students have group goals and are individually accountable for learning the content the group is engaged with. Outcomes are generally enhanced if students are taught specific ways of working in groups dealing with both metacognitive and social strategies for making best use of the group learning setting. Providing sufficient training and follow-up to ensure high-quality implementation is also essential.

Although important research continues to appear, the basic principles of cooperative learning have been established for many years, and there are many pragmatic training programs available. Yet cooperative learning remains an innovative approach familiar to most teachers but not used as a standard part of instruction. Most school principals can lead a visitor to a teacher enthusiastically using cooperative learning programs that are demonstrably working for the students, yet the visitor will note on the way to see that teacher the many fellow-teachers in the same school who are teaching students in rows, or using informal forms of groupwork without group goals or individual accountability, which research has rarely supported. Studies of actual use of cooperative learning (e.g., Antil et al., 1998) find that most use of cooperative learning is informal, and does not incorporate the elements that research has repeatedly found to be essential.

There remains a need for development and evaluation of cooperative learning programs that solve key problems of teaching and learning in all subjects and grade levels, and for continued research to identify the conditions under which cooperative learning is most likely to be effective. The greatest need at this point, however, is to develop and evaluate forms of cooperative learning that can be readily and successfully adopted by schools on a large scale, and to study the impediments to successful adoption of cooperative strategies. After 40 years of research and application, cooperative learning still has much more to contribute to students' learning.

Acknowledgments

Portions of this article are adapted from Slavin (1995). It was written under funding from the Institute of Education Sciences, US Department of Education. However, any opinions expressed are those of the author and do not necessarily represent Department of Education positions or policies.

See also: Cognitive Styles and Learning Styles; Educational Productivity; Educational Sociology; Group Processes in the Classroom; Instructional Psychology; Learning Theories and Educational Paradigms; Learning and Instruction: Social-Cognitive Perspectives.

Bibliography

Antil, L.R., Jenkins, J.R., Wayne, S., Vadasy, P.F., 1998. Cooperative learning: prevalence, conceptualizations, and the relation between research and practice. American Educational Research Journal 35 (3), 419–454.

Aronson, E., Blaney, N., Stephan, C., Sikes, J., Snapp, M., 1978. The Jigsaw Classroom. Sage, Beverly Hills, CA.

Barbato, R., 2000. Policy implications of cooperative learning on the achievement and attitudes of secondary school mathematics students (Unpublished Doctoral Dissertation). Fordham University, New York, NY.

Calhoon, M., 2005. Effects of a peer-mediated phonological skill and reading comprehension program on reading skill acquisition for middle school students with reading disabilities. Journal of Learning Disabilities 38 (5), 424–433.

Calhoon, M., Otaiba, S., Greenberg, D., King, A., Avalos, A., 2006. Improving reading skills in predominately Hispanic title I first grade classrooms: the promise of peer-assisted learning strategies. Learning Disabilities Research and Practice 21 (4), 261–272.

Fantuzzo, J.W., Riggio, R.E., Connelly, S., Dimeff, L.A., 1989. Effects of reciprocal peer tutoring on academic achievement and psychological adjustment: a component analysis. Journal of Educational Psychology 81 (2), 173–177.

Fantuzzo, J.W., Polite, K., Grayson, N., 1990. An evaluation of reciprocal peer tutoring across elementary school settings. Journal of School Psychology 28 (4), 309–323.

Fuchs, L.S., Fuchs, D., Kazden, S., Allen, S., 1999. Effects of peer-assisted learning strategies in reading with and without training in elaborated help giving. The Elementary School Journal 99 (3), 201–221.

Fuchs, L.S., Fuchs, D., Karns, K., 2001. Enhancing kindergartners' mathematical development: effects of peer-assisted learning strategies. Elementary School Journal 101 (5), 495–510.

Greenwood, C.R., Delquadri, J.C., Hall, R.V., 1989. Longitudinal effects of classwide peer tutoring. Journal of Educational Psychology 81 (3), 371–383.

Johnson, D.W., Johnson, R.T., 1998. Learning Together and Alone: Cooperative, Competitive, and Individualistic Learning, fifth ed. Allyn & Bacon, Boston, MA.

Kramarski, B., Mevarech, Z.R., Lieberman, A., 2001. Effects of multilevel versus unilevel metacognitive training on mathematical reasoning. Journal of Educational Research 54 (5), 292–300.

Mathes, P.G., Babyak, A.E., 2001. The effects of peer-assisted literacy strategies for first-grade readers with and without additional mini-skills lessons. Learning Disabilities Research & Practice 16 (1), 28–44.

Mattingly, R.M., Van Sickle, R.L., 1991. Cooperative learning and achievement in social studies: Jigsaw II. Social Education 55 (6), 392–395.

Mevarech, Z.R., 1985. The effects of cooperative mastery learning strategies on mathematics achievement. Journal of Educational Research 78 (6), 372–377.

Mevarech, Z.R., Kramarski, B., 1997. IMPROVE: a multidimensional method for teaching mathematics in heterogeneous classrooms. American Educational Research Journal 34 (2), 365–394.

Reid, J., 1992. The effects of cooperative learning with intergroup competition on the math achievement of seventh grade students. (ERIC Document Reproduction Service No. ED355106.)

Rohrbeck, C.A., Ginsburg-Block, M.D., Fantuzzo, J.W., Miller, T.R., 2003. Peer-assisted learning interventions with elementary school students: a meta-analytic review. Journal of Educational Psychology 94 (20), 240–257.

Sharan, S., Shachar, C., 1988. Language and Learning in the Cooperative Classroom. Springer-Verlag, New York.

Sharan, Y., Sharan, S., 1992. Expanding Cooperative Learning through Group Investigation. Teachers College Press, New York.

Slavin, R.E., 1994. Using Student Team Learning, second ed. Johns Hopkins University, Center for Social Organization of Schools, Baltimore, MD.

Slavin, R.E., 1995. Cooperative Learning: Theory, Research, and Practice, second ed. Allyn & Bacon, Boston.

Slavin, R.E., 2008. What works? Issues in synthesizing educational program evaluations. Educational Researcher 37 (1), 5–14.

Slavin, R.E., 2009. Cooperative learning. In: McCulloch, G., Crook, D. (Eds.), International Encyclopedia of Education. Routledge, Abington, UK.

Slavin, R.E., Karweit, N., 1984. Mastery learning and student teams: a factorial experiment in urban general mathematics classes. American Educational Research Journal 21 (4), 725–736.

Slavin, R.E., Karweit, R., 1985. Effects of whole-class and ability group instruction on math achievement. American Educational Research Journal 22 (3), 351–367.

Slavin, R.E., Lake, C., 2008. Effective programs in elementary mathematics: a best-evidence synthesis. Review of Educational Research 78 (3), 427–515.

Slavin, R.E., Madden, N.A. (Eds.), 2009. Two Million Children: Success for All. Corwin, Thousand Oaks, CA.

Slavin, R.E., Leavey, M.B., Madden, N.A., 1986. Team Accelerated Instruction Mathematics. Mastery Education Corporation, Watertown, MA.

Slavin, R.E., Cheung, A., Groff, C., Lake, C., 2008. Effective reading programs for middle and high schools: a best-evidence synthesis. Reading Research Quarterly 43 (3), 290–322.

Slavin, R.E., Lake, C., Groff, C., 2009a. Effective programs in middle and high school mathematics: a best-evidence synthesis. Review of Educational Research 79 (2), 839–911.

Slavin, R.E., Lake, C., Chambers, B., Cheung, A., Davis, S., 2009b. Effective reading programs for the elementary grades: a best-evidence synthesis. Review of Educational Research 79 (4), 1391–1466.

Stevens, R.J., Durkin, S., 1992. Using Student Team Reading and Student Team Writing in Middle Schools: Two Evaluations. Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students, Baltimore, MD. Report No. 36.

Stevens, R.J., Slavin, R.E., 1995a. Effects of a cooperative learning approach in reading and writing on handicapped and nonhandicapped students' achievement, attitudes, and metacognition in reading and writing. Elementary School Journal 95 (3), 241–262.

Stevens, R.J., Slavin, R.E., 1995b. The cooperative elementary school: effects on student achievement and social relations. American Educational Research Journal 32 (2), 321–351.

Stevens, R.J., Madden, N.A., Slavin, R.E., Farnish, A.M., 1987. Cooperative integrated reading and composition: two field experiments. Reading Research Quarterly 22 (4), 433–454.

Van Keer, H., Verhaeghe, J., 2005. Comparing two teacher development programs for innovating reading comprehension instruction with regard to teachers' experiences and student outcomes. Teaching and Teacher Education 21 (5), 543–562.

Van Keer, H., Verhaeghe, J. Strategic Reading in Peer Tutoring Dyads in Second and Fifth-grade Classrooms. Ghent University, Belgium. Unpublished report.

Webb, N.M., 1985. Student interaction and learning in small groups: a research summary. In: Slavin, R., Sharan, S., Kagan, S., Hertz-Lazarowitz, R., Webb, C., Schmuck, R. (Eds.), Learning to Cooperate, Cooperating to Learn. Plenum, New York.

Webb, N.M., 2008. Learning in small groups. In: Good, T.L. (Ed.), 21st Century Education: A Reference Handbook. Sage, Los Angeles, CA, pp. 203–211.

Coping across the Lifespan

Ellen A Skinner, Portland State University, Portland, OR, USA
Melanie Zimmer-Gembeck, School of Applied Psychology and Griffith Health Institute, Behavioural Basis of Health, Griffith University, Gold Coast, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E.A. Skinner, volume 4, pp. 2761–2766, © 2001, Elsevier Ltd.

Abstract

Coping, a basic process integral to adaptation and survival, depicts how people detect, appraise, deal with, and learn from stressful encounters. Decades of research in the social and medical sciences have examined coping in many domains across the life span. Mainstream research, focusing on measurement of individual differences and correlates of coping, suggests that coping can buffer or exacerbate the effects of stress on mental and physical health and functioning, as well as directly shape the development of psychopathology and resilience. New directions include construction of developmental theories, measurement schemes that augment self-report questionnaires, and designs that capture coping processes.

Introduction

Coping is an adaptive process, integral to functioning and development. At its core, coping depicts how people detect, appraise, deal with, and learn from actual and anticipated stressful encounters. The processes used to accomplish these tasks arise from many levels, and so coping is an organizational construct that includes a tightly integrated system of neurophysiological, attentional, emotional, motivational, behavioral, cognitive, social, and interpersonal processes. Because coping refers to how individuals deal with actual stressors on the ground in real time, a major focus of the area has been the identification of different 'ways of coping.' A wide variety of potentially adaptive ways have been studied, including instrumental action, problem-solving, support-seeking, reappraisal, accommodation, information-seeking, emotion expression and regulation, and negotiation; an even greater number of potentially maladaptive ways have also been identified, including avoidance, escape, rumination, denial, helplessness, passivity, confusion, concealment, isolation, self-pity, dependency, anxiety amplification, self-blame, other-blame, confrontation, venting, and aggression.

Considering the building blocks of the area, individual differences in these ways of coping have been the target of tens of thousands of studies, examining their links to physical and mental functioning, health, and psychopathology across the life span and in a wide variety of domains, including work, family, parenting, medicine, school, peers, bereavement, unemployment, aging, as well as nonnormative life events, like accidents, crime victimization, cultural upheaval, economic downturns, war, and relocation. Because these kinds of stressful life events comprise a host of changing challenges, threats, and losses, people typically employ many different coping strategies to deal with them, as both events and coping transactions unfold over time. People's coping strategies are shaped by their appraisals of the stressful situation, their personal resources (such as intelligence or optimism), and the social resources they can access (such as instrumental help or emotional support). Although findings are not always clear or conclusive, in general, research suggests that the ways that people cope can make a material difference to how stressful encounters are resolved and their short- and long-term consequences. For this reason, coping has been the target of intervention efforts across a wide spectrum of life stresses and events. This work suggests that coping is malleable and responsive to direct instruction, reappraisal of events, promotion of personal resources, and the provision of social supports.

Mainstream research, focusing largely on the measurement and correlates of coping, has been criticized for lack of consensus about core ways of coping, for overreliance on self-reports, and for use of cross-sectional designs. Developmental conceptualizations that link coping to regulation as well as process-oriented research in both naturalistic and lab settings, using a variety of assessments (interviews, narratives, daily reports, observations) and examining intraindividual relations and change over time, promises to reinvigorate the area.

Definitions

Given the current prominence of transactional views, coping is often defined as "constantly changing cognitive and behavioral efforts to manage specific external and/or internal demands that are appraised as taxing or exceeding the resources of the person" (Lazarus and Folkman, 1984: 141). Coping incorporates processes from many levels. It reflects evolution; humans come prepared to recognize and react to environmental demands in ways that promote survival. Coping has a tightly integrated physiological, psychological, and social basis. Many features of psychological functioning, such as emotions, motivation, attention, volition, cognition, and communication, are organized in ways that allow them to contribute to coping. For example, emotions signal and organize interactions of adaptive significance. Many aspects of social relationships and cultural systems also contribute to coping. For example, attachment relationships provide a safe haven of protection during times of distress.

Coping has both normative and individual difference features. Normatively, humans are prepared to cope adaptively. In fact, stress is likely necessary for the kinds of sustained constructive engagement that facilitate development. However, human responsiveness to stress can also represent

a vulnerability; too much stress can overwhelm and damage people. In addition, there are enormous differences in how physiological, psychological, social, and cultural resources for coping are distributed across individuals and sociocultural groups.

Finally, coping is a process that unfolds over time. Coping takes real time; any specific stressful transaction involves sequences of moment-to-moment interactions. Coping also takes place over episodic time; dealing with demands encompasses a protracted series of specific real-time transactions. As noted by White, "Described not inappropriately in military metaphors, adaptation often calls for delay, strategic retreat, regrouping of forces, abandoning of untenable positions, seeking fresh intelligence, and deploying new weapons" (White, 1974: 50). Coping also changes over developmental time: the means of coping as well as the personal and resources brought to bear during coping efforts show radical changes and qualitative shifts from birth and across the life span (Aldwin, 2007; Skinner and Zimmer-Gembeck, 2007).

Given its complexity and centrality to adaptation, it is not surprising that coping has been approached from a wide range of theoretical perspectives. It has been defined as a specific person–context transaction, personality in action under stress, a repertoire of strategies, a hierarchically organized set of ego processes, an indicator of competence, a function of emotion, an outcome of temperament, an expression of stress physiology, and a quality of action regulation. Although it has been suggested that coping is primarily a manifestation of other phenomena, most researchers vigorously argue that coping processes have an independent causal status. More specifically, they maintain that coping makes a material difference to how stressful interactions are resolved and is a key factor influencing long-term mental and physical heath, well-being, and development in the face of adversity.

Historical Contexts

Coping did not appear as a term in *Psychological Abstracts* until 1967. Since that time, research has been voluminous. A search of *PsycINFO* on 'coping' from 1967 to 2013 yields about 61 000 entries, marking it as a topic of significant interest to psychologists. Historically, work on coping started well before the 1960s. It has early forerunners in psychoanalytic work on defenses and has been heavily influenced by the study of stress. Work with children and the elderly incorporates a focus on development. Theories have been informed by a host of interventions designed to help people of all ages cope more adaptively with trauma and stressful life events. (For historical overviews, see Lazarus, 1993; Lazarus and Folkman, 1984; Murphy, 1974; Parker and Endler, 1996; Snyder, 1999.)

Coping and Defense

The concept of defense, as it first appeared in psychology in the early writings of Freud (1894–1962), referred to psychological mechanisms (such as repression or projection) used to protect the psyche from disturbing thoughts and emotions, mainly by distorting reality or removing the distressing material from conscious awareness. These ideas were expanded by Anna Freud (1936/1946) who argued that individuals have particular styles of defense, which are linked to specific forms of psychopathology. This formulation influenced several generations of ego-psychologists (e.g., Haan, 1977; Valliant, 1986; see Cramer, 1998). These theorists argued explicitly for a hierarchical model of ego processes, in which some defenses are more mature than others. For example, Haan (1977) posited a three-tiered taxonomy of ego processes: coping, considered more forward-looking, flexible, largely conscious, and attentive to reality; defense, considered to be inherently organized around issues from the past, rigid in operation, unconscious, and distorting of reality; and ego failure or 'fragmentation.'

Many notions from this approach appear in current theorizing. They include the idea that coping occurs in response not only to environmental demands, but also to intrapsychic pressures; that some ways of coping are inherently superior; and that people consistently show characteristic styles of coping with threat that reflect personality or ego development. Moreover, this work introduced several themes that are resurfacing in current discussions of coping, including the idea that some modes of adaptation are unconscious and unintentional, that distortions of reality can be adaptive, and that the self (or ego) and its regulatory functions are central to processes of coping.

Stress and Coping

A second strand of work emerged from research on stress, a concept prominent in the social and health sciences since the 1930s (Selye, 1950). Systematic attempts to link stress exposure to specific psychological or somatic outcomes revealed marked individual differences in the effects of stress. Explicit recognition that living organisms 'fight back,' as captured in terms like 'host resistance,' galvanized interest in the social and individual factors that determine vulnerability or resistance to the effects of stress. An important branch of this work contributed to the study of coping (Coelho et al., 1974; Lazarus and Folkman, 1984; Moos and Billings, 1982; Pearlin and Schooler, 1978). The transactional perspective, as articulated by Lazarus and his colleagues, held that the experience of stress is shaped by interpretations or appraisals of the meaning of environmental demands. Because these conceptualizations arose partly as a reaction to definitions of coping as an outcome of personality processes ordered along a hierarchy of ego maturity, transactional perspectives emphasize the importance of the actual demands a person is facing, and the impossibility of judging a priori which coping strategies are adaptive, without knowing the context and the social and personal resources available. They focus especially on the individual's *appraisal* of the significance and meaning of the stressful encounter. Coping itself is seen as a transactional process, iterating with (re) appraisals and environmental changes, and as having two primary functions: to deal directly with the environmental stressor (referred to as problem-focused coping) and to deal with the emotional consequences of the stressful transaction (emotion-focused coping).

The basic tenets of a transactional perspective, reflected in much of the current writing on coping, include the notion that coping does not directly reflect personality but varies in accordance with the situations in which it occurs; that no ways

of coping are inherently better; that any stressful event (such as illness or job loss) gives rise to a complex host of specific and changing demands; and that, in order to cope with these, any individual can and does use a wide variety of strategies. Perhaps most critical is the emphasis on coping as a process that unfolds over time.

Several methodological shifts in research accompanied this new approach. The measurement of coping shifted from clinical assessment to self-report and efforts were made to unconfound measures of coping from assessment of outcomes. In addition to a focus on coping with traumatic events (e.g., life-threatening illness or victimization), research also came to include the study of chronic stressful conditions and relatively minor everyday stressors or 'daily hassles.' Moreover, research became more specialized by type of stressor (e.g., health problems, work stress, or bereavement). Even within domains, research has become more finely differentiated. For example, different medical conditions are often studied separately.

Coping during Childhood, Adolescence, and Aging

Child psychologists have had a long-standing interest in the impact of adversity on children and, since the early part of the twentieth century, have attempted to document the effects, for example, of maternal deprivation, serious illness, hospitalization, and wartime conditions, as well as more recent attention to the effects of parental mental or physical illness, unemployment, poverty, and divorce. The first formal study of children's coping is often traced to Lois Murphy and her colleagues at the Menninger Clinic, who used intensive observations and interviews to conduct an 18-year longitudinal study of the ways in which normal children cope (Murphy and Moriarity, 1976).

The 1980s witnessed an explosion of quantitative work on children and adolescents, sparked by the publication of two seminal works on stress, coping, and development (Compas, 1987; Garmezy and Rutter, 1983). Although the current wave of research on children and youth closely resembles work on adults (Compas et al., 2001), it nevertheless highlights several important themes. It makes evident that coping is inherently shaped by social partners, close relationships, families, and communities (Hetherington and Blechman, 1996; McCubbin et al., 1996), and that many cognitive coping strategies, so prominent in adults, are not commonly used by children until late childhood (Zimmer-Gembeck and Skinner, 2011). Research on children also encourages a renewed consideration of the neurophysiological (Compas, 2006; Gunnar and Quevedo, 2007; Taylor and Stanton, 2007) and temperamental (Derryberry et al., 2003) underpinnings of coping, as well as attention to coping with normative developmental demands and tasks (Wolchik and Sandler, 1997).

Finally, a shift in focus to development across the life span spurred interest in how adults successively deal with life transitions, social changes, and the challenges of aging (Aldwin, 2007; Baltes and Baltes, 1990; Brandtstädter and Rothermund, 2002; Brandtstädter et al., 1999). Research on successful aging, in documenting that most people maintain a high level of psychological functioning despite objective biological declines and social losses, focuses attention on positive coping strategies, like selection, compensation, planning, and proactive coping. It also highlights the importance of 'accommodative' modes of coping, which allow people to acknowledge limitations, accept constraints, relinquish goals, withdraw from goal pursuit, and focus on gratitude and joy in current conditions.

Interventions to Improve Coping

A relatively distinct strand of work involves psychological and behavioral medicine interventions, which have succeeded in helping people deal adaptively with traumatic events and their associated ongoing stressors. In contrast to traditional therapy, these interventions focus on provision of information and social support, and on helping people build coping resources and skills for dealing more effectively with particular stressors, such as illness, chronic pain, divorce, or bereavement (Coyne and Racioppo, 2000; Sandler et al., 1997; Taylor and Stanton, 2007).

Transactional Perspectives on Coping

Transactional perspectives, which view coping as an interactive process that unfolds in several recursive steps, have guided much of the research on coping over the last 30 years (Folkman and Moskowitz, 2004; Lazarus and Folkman, 1984). According to this perspective, as pictured in **Figure 1**, coping transactions are initiated by encounters with *stress*, defined as internal and external events that individuals appraise as important to their well-being and as taxing or exceeding their resources (Lazarus and Folkman, 1984). Cognitive appraisals, focusing on the extent to which the stressor is personally relevant and amenable to personal control, result in views of the encounter as constituting a *threat* (i.e., impending harm), a *loss* (i.e., irreversible harm that has already been incurred), or a *challenge* (i.e., a stressor the individual is confident about mastering).

These appraisals trigger bouts of coping, defined as "cognitive and behavioral efforts to master, tolerate, or reduce external and internal demands and conflicts among them" (Folkman and Lazarus, 1980: 223), which utilize personal and social resources to solve the stressful problem or manage the individual's negative emotional reactions to it. These efforts produce coping outcomes, which, by feeding back to both the stressful event and individuals' reappraisal processes, can terminate or prolong the stressful transaction. According to this perspective, coping can be seen as a process that involves a wide variety of ways of reacting to and dealing with stressors that are organized sequentially, forming an interconnected action sequence or coping episode (Folkman and Lazarus, 1985).

Transactional Coping Research

Since the advent of the transactional approach, mainstream research has focused on *individual differences* in each of the links in the coping process. Typical studies examine a single population at a single time, and assess self-reported ways of coping (e.g., problem-solving, help-seeking, avoidance) with either a variety of unidentified stressors or a prespecified narrow class of stressors. Studies examine how these different ways of

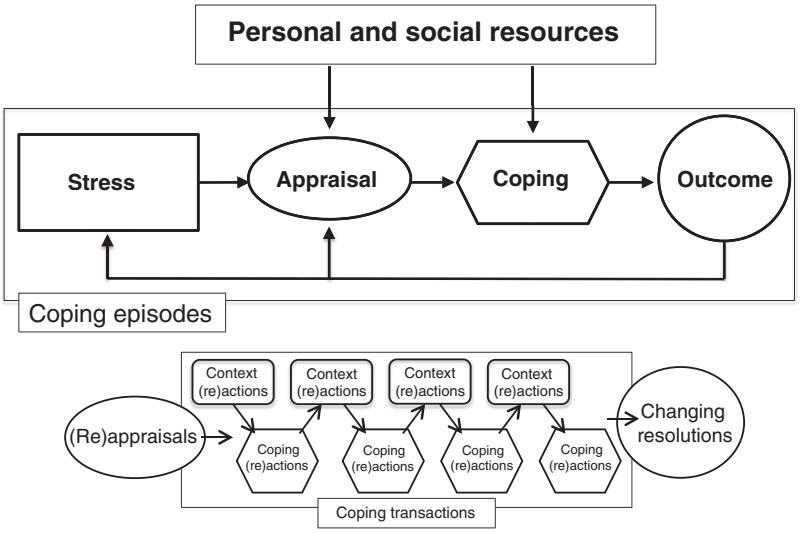

Figure 1 A transactional model of coping as an episodic process and a transactional process.

coping are connected to a variety of outcomes, usually indicators of mental or somatic functioning (such as emotional adjustment or physical symptoms) in an attempt to identify adaptive and maladaptive coping strategies. Complementary studies examine the associations of different ways of coping with self-reports of proposed personal and social antecedents, focusing both on individual characteristics (such as self-efficacy, optimism, or perceptions of the availability of social support) and on characteristics of the social context (such as provision of instrumental aid or emotional comfort).

Much has been learned from these decades of research on individual differences and correlates of coping. Certain ways of coping seem to be 'adaptive' in that they are linked with indicators of mental health and functioning, particularly (1) problem-focused coping and the set of control-related factors that support it; (2) constructive thinking and the factors that support it (e.g., optimism); and (3) perceptions of support from trusted others. In contrast, certain ways of dealing with stress, such as escape, avoidance, rumination, or venting, seem to be maladaptive in that they are associated with mental distress and disorder. The status of other ways of coping, such as help-seeking, support-seeking, secondary control, and emotion-focused coping, is less clear because they are inconsistent in their connections to antecedents and outcomes. A number of individual and social resources for coping have also been identified, chief among them intelligence, optimism, mastery, self-esteem, sociability, perceived control, and social support (Taylor and Stanton, 2007).

As a body, however, this work has come under attack on both conceptual and methodological grounds. Studies have been criticized for being light on theory, for inconsistent use of coping categories, for relying heavily on self-report questionnaires, for using designs that cannot examine change, and for including only a narrow range of short-term outcomes. As noted by Sommerfield and McCrae, "The seemingly boundless enthusiasm for coping research seen in the 1980s has been replaced by widespread disaffection, intense scrutiny, and corresponding calls for change" (2000: 620). Cutting-edge research has already moved away from this empirical format and reflects the use of several innovative approaches.

Coping as a Process

Despite consensus that coping unfolds over time, research methodologies able to capture processes have just begun to (re) appear during the last decade. These include case studies, ethnographic narratives, interviews, and observations, which were popular decades ago (Block and Block, 1980; Folkman, 1997; Murphy and Moriarity, 1976; Valliant, 1986), as well as more recent quantitative methods, such as intraindividual time series analysis (Tennen et al., 2000). Designs are longitudinal, repeating measures over short times (e.g., daily) or over several years, often including markers of the progress of stressful events (e.g., diagnosis and surgery, or caretaking, deterioration, and death of a partner).

For microgenetic approaches, patterns of intraindividual relations among variables (e.g., problem- and emotion-focused coping) across time are determined. Then interindividual difference variables (e.g., depression or optimism) are used to form groups that are compared on their patterns of intraindividual relationships (Tennen et al., 2000). Observations of coping interactions with social partners (e.g., mothers and their children) can also be used to examine the sequential effects of individual coping and partner reactions over time (Skinner and Edge, 2002). Narratives and interviews capture a wider range of experiences and can track their progress over longer intervals (Folkman, 1997). The yield from this labor-intensive research seems promising. For example, such studies already suggest that both generation of and focus on positive experiences are critical in dealing constructively with chronic stress; that depression influences day-to-day reactions to chronic pain; and that maternal interpretations of negative events shape children's subsequent appraisals and coping. Many consider the study of coping as a process to be the most important goal of the next generation of coping researchers (Lazarus, 1993).

Two productive empirical strategies for capturing coping processes have been (1) to make particular categories of coping, such as accommodation or rumination, the focus of detailed programs of laboratory and naturalistic research (e.g., Nolen-Hoeksema et al., 2008) and (2) to study in fine-grained detail how children and adults deal with certain classes of environmental demands, such as noncontingency, failure, unpredictability, separation, loss, interpersonal conflict, competition, and coercion (e.g., Maier and Watkins, 2005). Relevant studies come from areas as diverse as learned helplessness, self-efficacy, perceived control, reactance, self-determination, compliance, attachment, interpersonal problem-solving, attributional biases, Type-A personality, and self-regulated learning.

These programs of research have produced relatively well-accepted conclusions about positive ways of coping, such as mastery-oriented thinking, disclosure and discussion, constructive thinking, proximity-seeking, accommodation, and optimism, and about the generally maladaptive consequences of rumination, helplessness, blaming others, catastrophizing, and social isolation. Programs of research on other ways of coping, such as help- or support-seeking, positive illusions, avoidance, denial, and relinquishing control, have produced less conclusive results, including evidence that their effects depend on characteristics of the situations in which they are employed. It should be noted that almost every activity that is considered a way of coping also has its own well-developed body of research. Most of these areas of research are largely independent of work on coping, although some examine the functioning of these processes under stress, and future research could combine them into a profile or repertoire of coping responses.

Developmental Perspectives on Coping

'Stress' is not inherently damaging to people, and the primary function of coping is not simply to shield people from stressful experiences. Coping is a process intrinsic to psychological and relational growth and one that individuals and groups are predisposed to do well. Interest in positive coping (Snyder, 1999) and the study of growth from adversity, transformation in the face of trauma, and benefit finding can be viewed as attempts to conceptualize and study the growth-facilitating functions of coping. Analysis of the role of coping in resilience in children (Haggerty et al., 1994) and the relationship of coping and development across the life span (Aldwin, 2007; Skinner and Zimmer-Gembeck, 2007) may help put these issues in a larger perspective. In research on children and adolescents, developmental conceptualizations are emerging that, building on transactional perspectives, tie coping back to its roots as an adaptive process, focus on the role of coping as inherently regulatory, consider ways of coping as forms of action regulation, and make a place for learning and development.

Coping as Regulation

An explicit theme resurfacing in discussions and research is the notion that coping is an inherently regulatory process (Aspinwall and Taylor, 1997; Barrett and Campos, 1991; Block and Block, 1980; Eisenberg et al., 1997; Rossman, 1992; Skinner and Zimmer-Gembeck, 2009). A focus on adaptation suggests that the cognitions, appraisals, motivation, volition, behaviors, and emotions, which are clearly parts of coping, must be assembled into actions, which are sequentially coordinated with internal and external conditions. Labels under which such processes are studied include ego functioning (e.g., ego-control and ego-resilience), conscience, functionalist theories of emotion, theories of self-regulation (and regulation of behavior, emotion, and attention), and theories of volition and action regulation.

As with regulation, conceptualizations of coping have converged on dual-process models. Within regulation, the distinction is made between the target to be regulated (typically an emotion or impulse) and the processes used to modulate it. In work on coping, theorists differentiate (1) stress reactions, which refer to immediate involuntary physiological, psychological, and behavioral responses to stressful situations, from (2) action regulation, which describes how people mobilize, guide, manage, energize, and direct behavior, emotion, and orientation (or how they fail to do so) under stress. There is active debate about how these two processes are coordinated, but researchers agree that they mutually influence each other (Compas et al., 1999). Manifest coping responses likely reflect the balance (or more precisely, the imbalance) between reactivity and regulation: Involuntary stress reactions are likely the product of extreme stress reactivity combined with weak or disabled regulatory systems, whereas volitional coping responses are the result of weak stress reactions and/or well-developed action regulation systems (Metcalfe and Mischel, 1999).

Regulation and coping differ in at least two important ways (Compas et al., 2014). First, coping focuses only on "regulatory processes in a subset of contexts – those involving stress" (Eisenberg et al., 1997: 42). Second, research on regulation, as reflected in its labels, typically examines efforts to manage a single facet of action (e.g., emotion, attention, or behavior), whereas coping, in contrast, is an organizational construct, encompassing the simultaneous regulation of all these processes. When facing stressful events, individuals attempt not only to handle emotions, but also to manage thoughts, behaviors, attention, and even physiological reactions. Coping is essentially concerned with the integration and coordination among facets of action, for example, how a focus on one may have an (often unintended) impact on others, as when rumination undermines problem-solving or when seeking help bolsters persistence. Conceptions of coping as a dynamic regulatory process may contribute to the next phase of coping work (Carver and Scheier, 1998; Neufeld, 1999). Features of the self involved in recognizing and creating 'relational meaning' (Lazarus, 1993) may be studied as integral organizers of regulation.

Ways of Coping as Strategies of Action Regulation

Of central concern to coping researchers are the ways people deal with actual problems. The study of actual coping categories, such as problem-solving, support-seeking, rumination, or escape, distinguishes research on coping from closely related work on stress, adaptation, risk, resilience, and competence. The examination of a variety of responses distinguishes the

study of coping from the disparate programs of research focusing on each of the individual ways of coping. Hence, constructing category systems to conceptualize and measure coping has been a central endeavor of the field. However, this task is made difficult by the complexity of coping. Coping responses, because they are tailored to particular demands and resources, and attuned to the specific contexts in which they are enacted, are practically limitless in their variety; a recent review collected over 400 different category labels (Skinner et al., 2003). Research has been slowed by widespread disagreement about taxonomies of coping. No consensus exists about core categories. Years of exploratory factor analyses have failed to converge on a set of higher-order categories, nor do theorists agree on systems for rational classification. Given the seemingly infinite variety of specific coping responses, even the possibility of identifying a core subset of ways of coping has been questioned.

One promising approach uses theory and confirmatory factor analysis to identify multiple core categories. Over the last two decades, researchers have put enormous effort into conceptualizing and assessing hierarchical models that use higher-order categories or families to organize multiple lower-order ways of coping. Researchers have attempted to conceptually identify a finite set of basic adaptive processes and then to classify many different ways of coping as belonging to a 'family' of ways of coping that serve each function. Examples of theories used in these attempts include functionalist theories of emotion, theories of motivation, regulation, primary and secondary control, and ego processes.

Despite differences in theoretical approaches and dimensions, conceptual and empirical analyses have converged on a small number of families of coping, perhaps a dozen or so, that can be used to classify, most if not all, the ways of coping identified in previous research (Skinner et al., 2003). These include problem-solving, support-seeking, escape, distraction, cognitive restructuring, rumination, helplessness, social withdrawal, emotional regulation, information-seeking, negotiation, opposition, and delegation. Families, however, include not only the lower-order ways of coping depicted in their labels, but also all the other ways of coping that serve that same set of functions For example, 'escape' includes not only physically leaving a stressful situation, but also mental escape, withdrawal of effort, denial, avoidance, and other actions that remove the person from contact with distressing interactions. These 12 families of coping include the most common ways of coping utilized by people and studied by researchers (Skinner et al., 2003; Zimmer-Gembeck and Skinner, 2011). The identification of these higher-order families helps clarify the complex structure of coping and encourages renewed discussions of its adaptive functions (Coelho et al., 1974; Lazarus and Folkman, 1984). Such discussions expand on the distinction between problem-focused versus emotion-focused coping and on coping styles (e.g., repression vs sensitization) as well as reopening discussions of the functions of coping in adaptation.

The Development of Coping

When coping is returned to its core meaning as a fundamental adaptive process, the transactional steps of coping can be divided into basic tasks: (1) to detect and interpret information about internal and external demands (threat detection and appraisal); (2) to prepare a response based on internal and external guides and capacities (action readiness); and (3) to execute a response by coordinating action tendencies with internal and external demands and resources (action regulation). Moreover, in order to develop, the coping system also needs (4) to recover and learn from stressful encounters. These tasks are depicted in Figure 2.

Hence, at the most global level, one way of outlining the age-graded progression of coping, as well as its qualitative

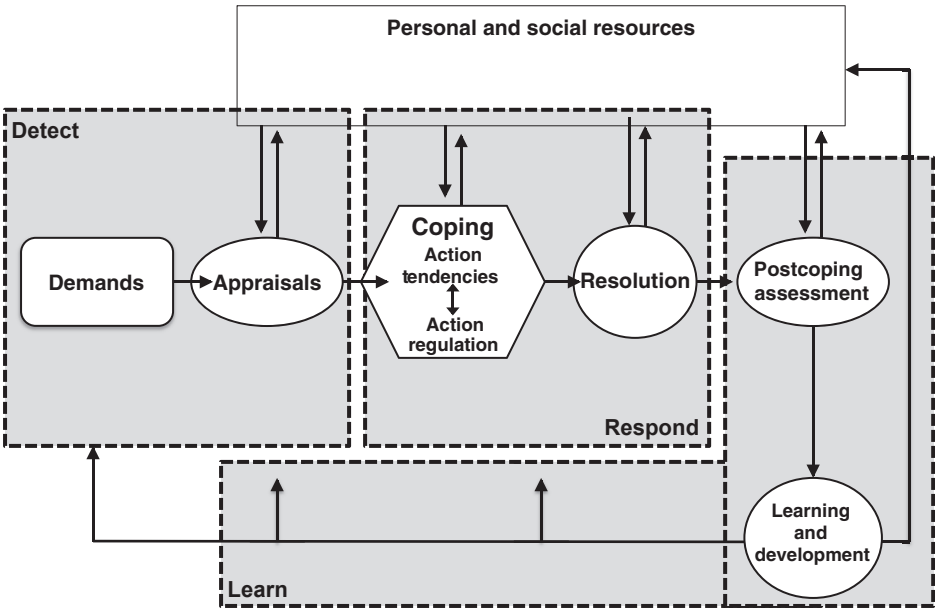

Figure 2 Coping as a process that allows individuals to detect, deal with, and learn from stressful encounters.

shifts, is to consider several broad developmental phases that are characterized by different mechanisms of detection, appraisal, reactivity, regulation, and learning, as well as different kinds of participation by social partners. As we described in our earlier work:

> Infancy would begin with stress reactions governed by reflexes, soon to be supplemented by coordinated *action schema*; during this period, caregivers would carry out coping actions based on the expressed intentions of their infants (interpersonal co-regulation). During toddlerhood and preschool age, coping would increasingly be carried out using *direct actions*, including those to enlist the participation of social partners; this would be the age at which voluntary coping actions would first appear (intrapersonal self-regulation). During middle childhood, coping through *cognitive means* would solidify, as described in work on distraction, delay, and problem-solving; children would be increasingly able to coordinate their coping efforts with those of others. By adolescence, coping through *meta-cognitive means* would be added, in which adolescents are capable of regulating their coping actions based on future concerns, including long-term goals and effects on others.
>
> Skinner and Zimmer-Gembeck, 2007: 128

Age-graded shifts in the basic tasks of coping (i.e., detection and appraisal, action tendencies and regulation, and learning) are shaped, on the one hand, by the development of neurophysiological subsystems and, on the other hand, by changes in the demands and resources provided by social partners, and especially caregivers. Research on the development of regulatory subsystems (including attentional, emotional, and behavioral regulation) can be used to trace the course of age-related changes in how these subsystems are triggered and coordinated in the face of stress. Social partners, especially caregivers, play a crucial role in the development of all these subsystems, both in the emergence and consolidation of these regulatory resources and capacities. Moreover, the caregiver's role in coping *changes* over development, from one in which they are doing most of the coping for newborns based on their infants' expressed preferences, to one of direct participation, then cooperation, and finally acting as a resource and backup system to the relatively independent coping of which adolescents are capable by the time they reach emerging adulthood.

Taken together, these developments should suggest normative patterns of change and normative transitions that characterize the healthy development of adaptive systems for constructive coping. They should also suggest differential pathways of development, based on factors (such as temperament, quality of attachment and parenting, and exposure to stressful events) that contribute to ways of negotiating these normative transitions that put individuals at risk for the development of maladaptive patterns of coping and psychopathology.

Conclusion

As the construct of coping reaches the end of its fifth decade as an official term in the psychological lexicon, it continues to inspire researchers and interventionists with its potential to capture important features of 'host resistance,' that still elusive quality that allows individuals to actively wrestle with intrapsychic and environmental challenges and problems, sometimes besting them, sometimes accepting them, or just plain losing, but still with the possibility to extract from these difficult and painful interactions lessons about how to avoid or approach the next stressful encounter. The promise of the area of coping, and its challenge to researchers, is to examine coping as an engine of development or a catalyst for change, to determine if conceptions of coping can provide a process mechanism that explains how people (individually and collectively), in the face of adversity, are able to find and create opportunities for development.

See also: Depression; Post-Traumatic Stress Disorder; Resilience; Self-Efficacy; Self-Esteem.

Bibliography

Aldwin, C.M., 2007. Stress, Coping, and Development: An Integrative Perspective, second ed. Guilford, New York.

Aspinwall, L.G., Taylor, S.E., 1997. A stitch in time: self-regulation and proactive coping. Psychological Bulletin 121, 417–436.

Baltes, P.B., Baltes, M.B. (Eds.), 1990. Successful Aging: Perspectives from the Behavioral Sciences. Cambridge University Press, Cambridge, UK.

Barrett, K.C., Campos, J.J., 1991. A diacritical function approach to emotions and coping. In: Cummings, E.M., Greene, A.L., Karraker, K.H. (Eds.), Life-Span Developmental Psychology: Perspectives on Stress and Coping. Erlbaum, Hillsdale, NJ, pp. 21–41.

Block, J.H., Block, J., 1980. The role of ego-control and ego-resiliency in the organization of behavior. In: Collins, W.A. (Ed.), Minnesota Symposium on Child Psychology, Development of Cognition, Affect, and Social Relations, vol. 13. Erlbaum, Hillsdale, NJ, pp. 39–101.

Brandtstädter, J., Wentura, D., Rothermund, K., 1999. Intentional self-development through adulthood and later life: tenacious pursuit and flexible adjustment of goals. In: Brandtstädter, J., Lerner, R.M. (Eds.), Action and Self-Development: Theory and Research Through the Lifespan. Sage, Thousand Oaks, CA, pp. 373–400.

Brandtstädter, J., Rothermund, K., 2002. The life-course dynamics of goal pursuit and goal adjustment: a two-process framework. Developmental Review 22, 117–150.

Carver, C.S., Scheier, M.F., 1998. On the Self-Regulation of Behavior. Cambridge University Press, Cambridge, UK.

Coelho, G.V., Hamburg, D.A., Adams, J.E. (Eds.), 1974. Coping and Adaptation. Basic Books, New York.

Compas, B.E., 1987. Coping with stress during childhood and adolescence. Psychological Bulletin 101, 393–403.

Compas, B.E., 2006. Psychobiological processes of stress and coping: implications for resilience in children and adolescents. Annals of the New York Academy of Science 1094, 226–234.

Compas, B.E., Jaser, S.S., Dunbar, J.P., Watson, K.H., Bettis, A.H., Gruhn, M.A., Williams, E.K., 2014. Coping and emotion regulation from childhood to early adulthood: points of convergence and divergence. Australian Journal of Psychology 66 (2), 71–81.

Compas, B.E., Connor, J.K., Saltzman, H., Thomsen, A.H., Wadsworth, M., 1999. Getting specific about coping: effortful and involuntary responses to stress in development. In: Lewis, M., Ramsay, D. (Eds.), Soothing and Stress. Erlbaum, Mahwah, NJ, pp. 229–256.

Compas, B.E., Connor-Smith, J., Saltzman, H., Thomsen, A., Wadsworth, M., 2001. Coping with stress during childhood and adolescence: problems, progress and potential in theory and research. Psychological Bulletin 127, 87–127.

Coyne, J.C., Racioppo, M.W., 2000. Never the twain shall meet? Closing the gap between coping research and clinical intervention research. American Psychologist 55, 655–664.

Cramer, P., 1998. Coping and defense mechanisms: what's the difference? Journal of Personality 66, 895–918.

Derryberry, D., Reed, M.A., Pilkenton-Taylor, C., 2003. Temperament and coping: advantages of an individual differences perspective. Development and Psychopathology 15, 1049–1066.

Eisenberg, N., Fabes, R., Guthrie, I., 1997. Coping with stress: the roles of regulation and development. In: Wolchik, S., Sandler, I. (Eds.), Handbook of Children's Coping: Linking Theory and Intervention. Plenum Press, New York, pp. 41–70.

Folkman, S., 1997. Using bereavement narratives to predict well-being in gay men whose partners died of AIDS: four theoretical perspectives. Journal of Personality and Social Psychology 72, 851–854.

Folkman, S., Lazarus, R.S., September 1980. An analysis of coping in a middle-aged community sample. Journal of Health and Social Behavior 21, 219–239.

Folkman, S., Lazarus, R.S., 1985. If it changes it must be a process: study of emotion and coping during three stages of a college examination. Journal of Personality and Social Psychology 48, 150–170.

Folkman, S., Moskowitz, J., 2004. Coping: pitfalls and promise. Annual Review of Psychology 55, 745–774.

Freud, A., 1936/1946. The Ego and the Mechanisms of Defense. International Universities Press, New York (Original work published in 1936).

Freud, S., 1984/1962. The neuro-psychoses of defense. In: Strachey, J. (Ed. and Trans.), The Standard edition of the Complete Works of Sigmund Freud, vol. 3. Hogarth Press, London, pp. 161–185 (Original work published in 1894).

Garmezy, N., Rutter, M. (Eds.), 1983. Stress, Coping and Development in Children. McGraw-Hill, New York.

Gunnar, M., Quevedo, K., 2007. The neurobiology of stress and development. Annual Review of Psychology 58, 11.1–11.29.

Haan, N., 1977. Coping and Defending: Processes of Self-Environment Organization. Academic Press, New York.

Haggerty, R., Sherrod, L., Garmezy, N., Rutter, M., 1994. Stress, Risk, and Resilience in Children and Adolescents. Cambridge University Press, New York.

Hetherington, E.M., Blechman, E.A. (Eds.), 1996. Stress, Coping, and Resiliency in Children and Families. Erlbaum, Mahwah, NJ.

Lazarus, R., 1993. Coping theory and research: past, present, and future. Psychosomatic Medicine 55, 234–247.

Lazarus, R.S., Folkman, S., 1984. Stress, Appraisal, and Coping. Springer, New York.

Maier, S.F., Watkins, L.R., 2005. Stressor controllability and learned helplessness: the roles of the dorsal raphe nucleus, serotonin, and corticotropin-releasing factor. Neuroscience and Behavioral Reviews 29, 829–841.

McCubbin, H.I., Thompson, A.I., McCubbin, M.A., 1996. Family Assessment: Resiliency, Coping, and Adaptation. University of Wisconsin Publishers, Madison.

Metcalfe, J., Mischel, W., 1999. A hot/cool-system analysis of delay of gratification: dynamics of willpower. Psychological Review 106, 3–19.

Moos, R.H., Billings, A.G., 1982. Conceptualizing and measuring coping resources and coping processes. In: Goldberger, L., Breznitz, S. (Eds.), Handbook of Stress: Theoretical and Clinical Aspects. Free Press, New York, pp. 212–230.

Murphy, L.B., 1974. Coping, vulnerability, and resilience in childhood. In: Coelho, G.V., Hamburg, D.A., Adams, J.E. (Eds.), Coping and Adaptation. Basic Books, New York, pp. 47–68.

Murphy, L.B., Moriarity, A., 1976. Vulnerability, Coping, and Growth: From Infancy to Adolescence. Yale University Press, New Haven, CT.

Nolen-Hoeksema, S., Wisco, B., Lyubomirsky, S., 2008. Rethinking rumination. Perspectives on Psychological Science 3, 400–424.

Neufeld, R.W.J., 1999. Dynamic differentials of stress and coping. Psychological Review 106, 385–397.

Parker, J.A., Endler, N.S., 1996. Coping and defense: a historical overview. In: Zeidner, M., Endler, N.S. (Eds.), Handbook of Coping. Wiley, New York, pp. 3–23.

Pearlin, L.I., Schooler, C., 1978. The structure of coping. Journal of Health and Social Behavior 19, 2–21.

Rossman, B.B.R., 1992. School-aged children's perceptions of coping with distress: strategies for emotion regulation and the moderation of adjustment. Journal of Child Psychiatry 33, 1373–1397.

Sandler, I.N., Wolchik, S.A., MacKinnon, D., Ayers, T.S., Roosa, M.W., 1997. Developing linkages between theory and intervention in stress and coping processes. In: Wolchik, S.A., Sandler, I.N. (Eds.), Handbook of Children's Coping: Linking Theory, Research, and Intervention. Plenum Press, New York, pp. 3–40.

Selye, H., 1950. The Physiology and Pathology of Exposure to Stress. Acta, Montreal, Canada.

Skinner, E.A., Edge, K., 2002. Parenting, motivation, and the development of children's coping. In: Dienstbier, R.A., Crockett, L.J. (Eds.), Agency, Motivation, and the Life Course: vol. 48 of the Nebraska Symposium on Motivation. Nebraska University Press, Lincoln, NE, pp. 77–143.

Skinner, E.A., Zimmer-Gembeck, M.J., 2007. The development of coping. Annual Review of Psychology 58, 119–144.

Skinner, E.A., Zimmer-Gembeck, M.J. (Eds.), 2009. Coping and the Development of Regulation. In: Larson, R., Jensen, L. (Eds.-in-Chief), New Directions in Child and Adolescent Development. Jossey-Bass, San Francisco.

Skinner, E., Edge, K., Altman, J., Sherwood, H., 2003. Searching for the structure of coping: a review and critique of category systems for classifying ways of coping. Psychological Bulletin 129, 216–269.

Snyder, C.R., 1999. Coping: The Psychology of What Works. Oxford University Press, New York.

Sommerfield, M.R., McCrae, R.R., 2000. Stress and coping research. Methodological challenges, theoretical advances, and clinical applications. American Psychologist 55, 620–625.

Taylor, S., Stanton, A., 2007. Coping resources, coping processes, and mental health. Annual Review of Clinical Psychology 3, 377–401.

Tennen, H., Affleck, G., Armeli, S., Carney, M.A., 2000. A daily process approach to coping: linking theory, research, and practice. American Psychologist 55, 626–636.

Valliant, G.E., 1986. Empirical Studies of Ego Mechanisms of Defense. American Psychiatric Association, Washington, DC.

White, R.W., 1974. Strategies of adaptation: an attempt at systematic description. In: Coelho, G.V., Hamburg, D.A., Adams, J.E. (Eds.), Coping and Adaptation. Basic Books, New York, pp. 47–68.

Wolchik, S.A., Sandler, I.N. (Eds.), 1997. Handbook of Children's Coping: Linking Theory, Research, and Intervention. Plenum Press, New York.

Zimmer-Gembeck, M.J., Skinner, E.A., 2011. The development of coping across childhood and adolescence: an integrative review and critique of research. International Journal of Behavioral Development 35, 1–17.

Coping, Personality and

Gerald Matthews, Institute for Simulation & Training, University of Central Florida, Orlando, FL, USA
Moshe Zeidner, Laboratory for Research in Personality, Emotions, and Individual Differences, University of Haifa, Haifa, Israel
Richard D Roberts, Center for Innovative Assessments, Professional Examination Service, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Personality traits are associated with stable affective dispositions, including vulnerability to stress. Coping is conceptualized as a process that mediates interactions between people and sources of environmental demand. Thus, research on coping may help to explain associations between personality and affective response. Various traits correlate with coping preferences. Traits associated with stress vulnerability, such as neuroticism, may predispose maladaptive coping with challenging events, although coping is also strongly influenced by situational factors. Adaptive personality characteristics including resilience and emotional intelligence may be associated with use of effective coping strategies. Research in this field may be applied to enhancing individual resilience to stress.

Personality and coping are two key constructs in psychology that lack obvious compatibility (Parker and Wood, 2008). The dominant paradigm in contemporary personality research emphasizes the relative stability of personality traits, such as the Five Factor Model (FFM or 'Big Five'). That is, individuals display consistent patterns of behavior across a range of different situations. By contrast, theories of coping with stress (Lazarus, 1999) tend to emphasize the variability of coping as an expression of the person's idiosyncratic evaluations of specific situations. Such theory lends itself to intraindividual investigations, focusing on situational determinants of coping, rather than research on interindividual differences displayed across multiple situations (Parker and Wood, 2008). Fortunately, a changing theoretical landscape increasingly permits integration of these different perspectives within an interactionist perspective that acknowledges the importance of both person and situation factors in shaping response to challenging events (Matthews et al., 2009). It is also acknowledged that coping may be viewed both as a stable preference for using certain strategies (a trait or style) and as a specific set of strategies and behaviors applied to an immediate situation (a state).

In this article, we consider first the differing theoretical perspectives traditionally associated with personality and coping (section 'Theoretical Frameworks'), followed by a brief survey of the measurement models used in research (section 'Measurement Models'). In reviewing research in the area (section 'Research On Personality Traits And Coping'), we differentiate correlational studies (section 'Descriptive Correlational Studies') from more theory-driven research that aims to use coping models to understanding personality processes (section 'Coping as a Personality Process'). In the section 'Individual Differences in Coping Effectiveness', we turn to the issue of coping effectiveness and adaptation to stress, with subsections on the definition of effectiveness (section 'Defining Coping Effectiveness') and the relation between personality and resilience to stress (section 'Personality and Resilience'). We then discuss the popular construct of emotional intelligence (EI) (section 'Emotional Intelligence and Coping'). We consider how EI may be conceptualized in relation to effective coping and emotion regulation (section 'Conceptualizing Coping as EI in Action') and relevant empirical studies (section 'Empirical Studies of EI and Coping'). In the section 'Applications of Research' we cover applications of personality and coping research to clinical, educational, and industrial-organizational psychology, prior to the conclusion (section 'Conclusion').

Theoretical Frameworks

The predominant personality theory applied to studies of coping is the FFM (McCrae, 2009), which discriminates Extraversion, Neuroticism, Conscientiousness, Agreeableness, and Openness (E, N, C, A, and O, respectively) as the preeminent stable personality traits. Numerous other traits have also been investigated as potential correlates of coping, but typically such traits can be aligned with one or more FFM factors. Definitions of the FFM and other traits imply that they may be linked to stable interindividual differences in coping style. For example, we might expect that (1) extraverts would be more likely to seek social support than introverts and (2) persons high in N would engage in ineffective or counterproductive efforts at coping.

Traditionally, variation in traits is attributed to individual differences in brain systems for arousal, including structural and functional facets or reinforcement sensitivity. Psychobiological personality theories do not include coping as an explanatory construct capable of mediating the emotional and behavioral expressions of traits. More recently, the role of cognitive processes in personality has been increasingly accepted. Emotional expressions of traits may in part reflect individual differences in cognitive appraisal and the choice and regulation of coping strategies (Matthews et al., 2009). Thus, the negative emotionality associated with high N may reflect not only a highly sensitive brain punishment system, but also systematic cognitive biases, such as appraisal of events as threatening, and a preference for coping with events through worrying about them or self-blaming. It is then hypothesized that measures of traits will predict how the individual copes when exposed to a threatening or challenging situation.

Development of the cognitive theory of personality requires the core concepts of contemporary psychological stress theory.

According to Lazarus (1999), these are *appraisal*, the evaluation of the personal significance of events, and *coping*, the selection of a strategy to effectively manage or control the event, for example, through mitigating or circumventing harm or finding a way to benefit from it. Coping here is conceptualized as an extended process that typically changes over time, in response to both changing external events and the person's own reappraisals of the success or otherwise of their attempts to cope. Ineffective or maladaptive coping will promote psychological and physiological symptoms of stress, especially if the threat is severe, prolonged, or involves multiple stressors. The additional insight from personality theory is that although situational factors, such as the controllability of the event, exert a powerful influence on coping, there are also systematic biases in choice of strategy that are related to personality traits. For example, a person high on N may prefer to cope via self-blame, whereas a person high on E may prefer seeking social support.

Testing the cognitive theory of personality in relation to coping requires accommodation of situational factors, consistent with an interactionist perspective (Matthews et al., 2009). That is, a stressful encounter may interact with a person's personality to impact choice of coping strategy, which in turn may determine outcomes of the stress process. One approach is to conduct experimental studies in which individual differences in coping with a laboratory stressor may be observed. Such studies have the advantage of experimental control and precise measurement of the time course of response, but are lacking in other respects. Stressors must perforce be relatively mild, and perhaps unrepresentative of real-life events. This issue raises the perennial question of the ecological validity of lab experiments. Furthermore, it is difficult to observe the continual flow of person–situation interaction stated by Lazarus (1999) to be central to the stress process. The role of personality in life stress may be studied through longitudinal studies in which key variables are tracked over time, allowing causal modeling of the data. Such models may also reveal influences of stress factors on personality, as well as the influence of personality on stress (Ormel et al., 2012). Various statistical techniques for conducting such longitudinal analyses (latent change modeling, latent growth curve analyses, etc.) are now available and can be readily employed to test hypotheses related to the interplay between coping and personality over time.

Measurement Models

Personality research has benefited from the rigorous psychometric models developed over decades of quantitative studies of the FFM and other structural schemes (McCrae, 2009). By contrast, early research on coping focused on how individual persons cope qualitatively (the 'idiographic' approach), leading to a neglect of quantitative assessment. Nevertheless, it became clear that there are different modes of coping, which may be distinguished operationally. A review of coping measures is beyond the present scope (see Schwarzer and Schwarzer, 1996), but some of the main distinctions between facets of coping will be presented. Lazarus and his colleagues made a theoretical distinction between problem-focused coping that is directed toward systematic efforts to improve the external situation, and emotion-focused coping referring to efforts at changing one's feelings and thoughts about the event. Endler and Parker's (1990) Coping Inventory for Stressful Situations (CISS) included scales for these two constructs, as well as items that defined a third dimension of avoidance. Furthermore, Endler and Parker (1990) found that E was associated with a preference for problem-focus, and N with greater use of emotion-focus and avoidance, at least according to the respondent's self-report.

Subsequently, research has employed a wide variety of instruments. Recently, Carver and O'Connor-Smith (2010) highlighted the following constructs incorporated into coping scales, in addition to problem- versus emotion-focus.

Engagement versus Disengagement

Engagement refers to coping through approaching or directly confronting the stressor, not only through problem-focused action, but also by forms of emotion-focus including seeking support and constructive thinking. Disengagement refers to avoidance, denial, and wishful thinking. E and N are often linked to approach and avoidance tendencies, respectively.

Accommodative and Meaning-Focused Coping

Rather than attempt to control a stressor directly, a person may try to accept changes in life circumstances, to restructure their personal goals, and to find meaning in otherwise adverse events. Such 'positive-thinking' strategies are often seen as an attribute of EI, as further discussed below in the section EI and Coping.

Proactive Coping

People may try to anticipate and forestall adverse events, typically through problem-focused efforts to prepare contingency plans, and build up resources to defend against threat. We might expect conscientious individuals to favor such strategies.

Research on Personality Traits and Coping

Descriptive Correlational Studies

Numerous investigations have reported cross-sectional associations between a wide range of personality traits and coping scales. **Table 1** shows some of the major associations between the FFM and coping dimensions determined in a meta-analysis of 165 samples (O'Connor and Flachsbart, 2007). In terms of the descriptive scheme set out by Carver and O'Connor-Smith (2010), both E and C were linked to use of engagement strategies, whereas N was associated with several disengagement strategies as well as venting negative emotion. Numerous small-magnitude associations have been omitted from the table; both A and O were weakly associated with several engagement strategies. The meta-analysis also confirmed the psychometric distinctiveness of personality and coping measures. None of the associations listed exceeded 0.5, and there were several coping strategies including acceptance and emotion regulation that were minimally related to the FFM.

Table 1 Major relationships between Five Factor Model (FFM) traits and coping in a meta-analysis

FFM trait	*Preferred coping strategies*
Extraversion	Problem-solving
	Use of social support
	Cognitive restructuring
Neuroticism	Expression of negative emotion
	Withdrawal
	Wishful thinking
	Substance use
Conscientiousness	Problem-solving
	Cognitive restructuring
Agreeableness	–
Openness	–

Note: The table lists associations for which the mean estimated personality–coping correlation was 0.20 or higher.
Connor-Smith, J.K., Flachsbart, C., 2007. Relations between personality and coping: a meta-analysis. Journal of Personality and Social Psychology 93, 1080–1107.

Cross-sectional studies are valuable for elaborating the network of relationships between the various constructs, providing evidence on the validity of personality and coping measures alike and suggesting hypotheses for further research. However, they are limited in their neglect of the dynamic relationships emphasized by Lazarus (1999) and frequent failure to accommodate the role of contextual moderators. For example, O might be more strongly related to coping in intellectual or creative contexts, and the impact of Aon coping might be different in cooperative and competitive social situations. There is also a concern that the association between personality and coping may be inflated by method factors. Both types of scale depend on retrospective self-reports of general personal tendencies, which in the case of coping may not be a veridical representation of how the person actually behaves when exposed to some specific stressor. Trait measures do indeed tend to correlate with dispositional styles of coping more strongly than with immediate situational coping (Carver and O'Connor-Smith, 2010).

Coping as a Personality Process

Coping may be important for understanding how personality traits influence behavior and subjective experience. Are high N individuals typically unhappy *because* they cope ineffectively with the demands of life? That is, are some of the various expressions of personality mediated by individual differences in coping? Mediation hypotheses may be tested in both experimental and longitudinal survey designs. Experimental studies of performance stressors such as cognitive overload and time pressure are suitable for investigating the role of individual differences in situational coping in acute stress response. The stress response is multivariate. Depending on task characteristics, the person may show more or less task engagement, distress, or worry. For example, cognitive overload tends to produce distress responses, characterized by anxiety, tension, unhappiness, and loss of self-confidence, whereas monotonous, fatiguing tasks induce loss of task engagement (Matthews et al., 2013).

Performance studies that draw on the Lazarus (1999) model of stress and coping variously show that individual differences in stress response are systematically linked to both situational coping and appraisal of task demands. In terms of the Endler and Parker (1990) model, task- (or problem-) focus tends to elicit high task engagement, self-critical emotion-focus produces distress and worry, and avoidance also adds to worry. Appraisal may also influence stress response; e.g., challenge elevates task engagement (Matthews et al., 2013). Appraisal may also affect response indirectly; e.g., lack of perceived control elicits emotion-focused and avoidant coping (Matthews et al., 2013). Coping also has behavioral consequences. Task-focus also appears to be effective in promoting superior objective performance, whereas emotion-focus and avoidance are negatively related to performance (Matthews et al., in press).

Personality traits are modestly associated with both cognitive stress processes and stress outcomes, in experimental studies of performance and other stressors. Effect sizes (*r*s) rarely exceed 0.3–0.4. The most reliable associations are found for N and trait anxiety, which are associated with heightened appraisals of threat and uncontrollability and greater situational use of emotion-focus and avoidance, consistent with studies of dispositional coping style (O'Connor and Flachsbart, 2007). For example, Matthews et al. (2006) investigated three task stressors: overload induced by high time pressure, failure on an impossible anagrams task, and monotony resulting from a demanding vigilance task. N was associated with elevated distress response, and the personality effect was mediated by emotion-focused coping. In addition, C was associated with both situational task-focus and higher task engagement.

Experimental studies are essential for testing potential-mediating mechanisms in personality research, but they also have limitations. Assessments of acute stress response and performance provide only limited insights into coping effectiveness, as previously discussed. For example, while problem-focused coping is normally adaptive in work environments, in a dead-end job there may be advantages to emotion-focused coping directed toward deciding whether it is worthwhile to consider a different line of employment. Simple mediation models also miss some of the subtlety of the interplay between personality, coping, and stress response. Personality may also have a moderator effect in determining how well a strategy works for the individual (Bolger and Zuckerman, 1995). For example, task-focused coping may only be effective if the person has the skills to implement such a strategy successfully, such as computer skills when seeking an alternative job in the hi-tech industry. Similarly, seeking social support may be more effective if the person is high in A or E and is embedded in a supportive social network. Traits may also influence exposure to stressors, necessitating higher frequency of coping (Matthews et al., 2009). Someone high in N but low in A may commonly have to deal with interpersonal conflicts.

Individual Differences in Coping Effectiveness

Defining Coping Effectiveness

Research on personality and coping begins with a descriptive account of how major personality traits are associated with

measures of the various coping styles, such as those differentiated by Carver and O'Connor-Smith (2010). Identification of the coping styles typical of the different traits leads to various additional research questions. A common research preoccupation is the extent to which individual differences in coping translate into variation in outcomes and adaptive success or failure. In particular, people high in N and in related characteristics such as trait anxiety tend to experience lower well-being and greater vulnerability to stress and even emotional disorder. Individual differences in coping may contribute to such outcomes. Building an argument of this kind requires that we can differentiate coping strategies according to their effectiveness. However, it is important to note at this juncture that coping strategies are not uniformly adaptive and it is often difficult to predict the effectiveness of a given coping strategy a priori (Zeidner and Saklofske, 1996). As stressed by Zeidner and Saklofske (1996), the consequences of a given coping style or behavior are determined to a major extent by the interaction of personal traits, needs, and preferences, in dynamic interaction with the constraints of the parameters of the particular stressful encounter. Thus, specific coping strategies tend to be more or less effective depending on the type of stress encountered by a specific person.

Coping processes are of pivotal importance in that they may affect adaptive outcomes, for better or worse. Effective coping will help an individual adapt to a stressful situation by reducing distress and anxiety, but misplaced, counterproductive coping attempts may make the situation worse and eventually lead to elevated anxiety. People typically develop a repertoire of coping strategies for dealing with the difficulties that arise in relationships, at work, and in financial matters. Most people attain reasonable competence in coping with stress and anxiety through learning (direct instruction, modeling behavior of relevant others, reinforcements) and experience (positive or negative consequences of coping behaviors). Coping strategies may protect us by eliminating or modifying the conditions that produce anxiety or by keeping the emotional consequences within manageable bounds. Coping may also affect outcomes through its impact on the frequency, intensity, duration, and patterning of physiological stress reactions and the resultant affective and somatic outcomes. Often, coping strategies may impede – rather than promote – health-related behaviors. For example, a person's health may be negatively affected when coping involves risk taking (e.g., high-speed car racing or substance abuse).

The psychology of stress gives us a rich descriptive account of different types of coping. The practical question though is 'what works?' Unfortunately, the simple question of what works has no simple answer: a coping response might be judged successful relative to one outcome criterion but not another (Zeidner and Saklofske, 1996). Indeed, the resolution of one coping task might even come at the expense of another (e.g., working long hours to complete a PhD thesis might contribute to problems at work or a marriage breakdown). Coping effectiveness must be examined in the social context in which problems occur. Preferred coping methods and perceived effectiveness must be appraised relative to a social or cultural group's values, norms, worldview, symbols, and orientation. Consider the case of a female student who devotes herself to her children or ailing parents at the expense of personal achievement goals at her academic studies. The evaluation of this coping approach is not merely a scientific but a moral matter and may differ in traditional versus achievement-oriented societies. Thus, any statement about coping effectiveness is at best a broad generalization. With this caution in mind, there are a number of specific techniques that have been typically judged by researchers as adaptive, others that have been judged as maladaptive, whereas other techniques present dilemmas to researchers.

Theorists have frequently emphasized the positive effects of problem-focused coping on psychological outcomes (Lazarus and Folkman, 1984). In fact, active coping is preferred by most persons and is generally more effective in stress reduction in Western industrialized and individualistic societies. Other strategies are increasingly used when the source of stress is unclear, when there is a lack of knowledge about stress modification, or when there is little one can do to eliminate stress. By contrast, theorists have frequently emphasized the negative effects of avoidance and emotion-focused coping on psychological outcomes, especially when the threatening situation can be ameliorated by the subject's responses (Lazarus and Folkman, 1984). Behavioral or mental disengagement, ventilation of emotions, and short-term tension reduction strategies (e.g., use of drug, alcohol, smoking, gambling) are generally candidates for dysfunctional coping tactics for the amelioration of anxiety. By contrast, there may be psychological benefits to more constructive forms of emotion-focus, such as gaining understanding of one's emotional reactions to an uncontrollable event, such as bereavement. Similarly, avoidance coping has both its adaptive and maladaptive aspects. On the one hand, avoidance coping that leads to delay in taking necessary action (e.g., going to see the dentist for that regular checkup) is positively tied to concurrent distress. On the other hand, avoidance may be a useful tactic at times if it gives the person a psychological breather and an opportunity to escape from the constant pressures of the stressful situation. Also, the effectiveness of a particular coping strategy may change across the phases of a stressful encounter. While problem-focused coping may be effective in preparing for a final exam, emotion-focused coping may be effective during the anticipatory phase of awaiting the test results.

Personality and Resilience

In tune with current interest in 'positive psychology,' personality researchers are increasingly interested in the qualities that support personal growth and success in challenging or adverse environments. If we can delineate generally effective coping strategies, perhaps we can also identify effective copers. Resilience is seen as an important, general aspect of personality that moderates the impact of stressors (Skodol, 2010). In a sense, resilience is close to being tautological with coping effectiveness. Personality scales for qualities such as hardiness, toughness, grit, and sense of coherence may capture some general coping ability distinct from any specific coping style or preference. For example, the US military aims to train resilience to enhance combat effectiveness and to reduce vulnerability to traumatic stress. Skodol (2010) cites several longitudinal studies that suggest that traits for resilience are associated with greater long-term well-being following various negative or

challenging life events, including childhood abuse, abortion, and cancer surgery. For example, sense of coherence may impact upon the traumatic memories of holocaust survivors and is related to well-being in late life. However, the extent to which personality effects on outcomes are mediated by coping varies from study to study.

There are several barriers to conceptualizing personality in terms of dispositional resilience. First, it is unclear whether traits for resilience are psychometrically distinct from broader constructs such as emotional stability (low N, and the related concept of core self-evaluations; see Elliott et al., 2013). In fact, some have perceived hardiness as a combination of low N and high E. As such, hardiness may simply be a different label for the same underlying traits. Second, it is often difficult to establish reliable moderator effects in research (Matthews et al., 2009). The assumption is that resilience becomes evident only in demanding, potentially stressful environments, in line with the saying that "when the going gets tough, the tough get going." Thus, stress outcomes should reflect an interaction between trait resilience and environmental stress. In some studies, though, individuals high in resilience exhibit higher well-being regardless of external pressures, so that personality and external stressors have additive effects, contrary to the moderator hypothesis. Third, there remains uncertainty over the extent to which resilience should be conceptualized as a preexisting personality characteristic that emerges when the going gets tough, or whether resilience is better characterized as an ongoing process of adjustment.

In sum, resilience is a promising new construct in stress research that may be associated with personality. However, substantial questions remain about the conceptualization of resilience and the extent to which the construct offers anything novel to personality and coping researchers. Another construct that is related to general emotional adaptation is EI, which we will cover in more detail subsequently.

EI and Coping

This section discusses the role of EI in coping with psychological stress – an area of research in which EI has been claimed to play a pivotal role. Specifically, we critically examine the conceptual underpinnings and empirical evidence for the utility of EI in determining or predicting styles and strategies of coping.

Conceptualizing Coping as EI in Action

EI has been one of the most popular, but contentious, constructs in modern day psychology (Zeidner et al., 2009). Broadly defined, EI represents a set of hierarchically organized core competencies for identifying, processing, and regulating emotions – in both self and others (Salovey et al., 1999). There are currently two different and conflicting models for conceptualizing and researching EI: (1) EI as a cognitive ability, best measured via performance-type tests, and (2) EI as a noncognitive factor or trait, best measured via self-report inventories. In this section we focus mainly on the latter model, although both aspects of EI are linked to greater well-being and other positive social–emotional outcomes (Zeidner et al., 2009).

Recent theorizing and research suggest that EI may indeed be a pivotal factor determining how a person handles stress in daily life and confronts: what Shakespeare so aptly termed "the slings and arrows of outrageous fortune." Various researchers have proposed theoretical links between EI and coping. For example, Zeidner et al. (2006) suggested that "adaptive coping might be conceptualized as EI in action, supporting mastery of emotions, emotional growth, and both cognitive and emotional differentiation, allowing us to evolve in an ever changing world" (p. 460). Current thinking among proponents of EI suggests that "heeding the wisdom" provided by the emotional system determines effective coping behaviors and shapes adaptive outcomes. Salovey et al. (1999) claimed that EI helps individuals cope successfully because they "accurately perceive and appraise their emotional states, know how and when to express their feelings, and can effectively regulate their mood states" (p. 161).

Conceptually, EI should indeed effectively mobilize coping processes to promote adaptation, plausibly operating as a personal resource in determining choice and implementation of an appropriate coping style, which in turn affects adaptive outcomes. Emotionally intelligent individuals, who are skilled at expressing, understanding, and managing their emotions and who are capable of adaptively coping with the stressors and hassles of everyday life, should also experience more positive outcomes than their low-EI counterparts. In fact, EI may be part of one's affective arsenal and 'bag-of-tricks' to manage emotion-laden encounters and coping with stressful encounters.

Empirical Studies of EI and Coping

Given the proposed conceptual links between EI and coping, a body of research has examined the empirical associations between EI and adaptive coping styles. Research devoted to uncovering relations between EI and effective coping strategies has generally touched on two related issues. The first and simpler issue has involved determining how EI measures correlate with established coping measures, including, for example, dispositional or situational coping strategy (EI → coping disposition → situational coping).

Overall, studies that have examined the EI–coping relationship have found that questionnaires for EI related significantly to more use of problem-oriented, active, and rational coping, and less use of emotion-focused, passive, and irrational coping styles. Zeidner et al. (2006) summarized much of the research linking EI and coping, noting that correlations among these constructs range between 0.20 and 0.60. Furthermore, the strength and direction of the relationships varies according to the ways in which EI is operationalized (trait EI versus ability EI) as well as the type of coping strategy measured (problem-focused, emotion-focused, or avoidant). Trait EI shows a consistent positive relationship to problem-focused coping, with correlations mainly at the $r = 0.30$–0.40 levels. Trait EI also shows a consistent negative association with emotion-focused coping at around the $r = -0.30$ level. Relationships of trait EI with avoidance coping tend to be trivially small but consistently negative.

Furthermore, trait EI was shown to promote the choice of adaptive strategies not only in the case of stress, but also in

other emotions such as anger, sadness, fear, jealousy, and shame, as well as promoting the use of adaptive strategies to maintain joy. An alternative research strategy is to assess situational coping with a specific challenge, as in the performance studies reviewed previously. A body of research suggests that participants who were better at repairing their emotions and could clearly distinguish their emotions also reported less negative emotional responses and personal distress, which enables individuals to adapt to stressful encounters.

The second, more subtle, issue has involved ascertaining whether coping mediates (EI → coping → adaptive outcomes) or moderates associations between EI and adaptive outcomes (EI × coping → adaptive outcomes). **Table 2** lists several mediating mechanisms that might be responsible for superior social–emotional adaptation in those high in EI; some but limited evidence supports these proposals. In any case, as a broad-based characteristic, it is likely that EI supports multiple, complementary competencies in handling adverse events. A small number of studies also suggest the EI may have moderator effects, for example, by amplifying the beneficial effects of active or engagement coping (Zeidner et al., 2006). However, research on these issues is still rather scant.

Overall, the scope of individual differences in coping linked to EI is very broad indeed. Some of the mediating mechanisms (e.g., adaptive regulation) refer directly to coping with emotion. Other mechanisms, such as managing exposure to stressors, are more likely to influence emotion indirectly, depending on the outcome of the encounter. Most self-report measures of EI have proved to be substantially confounded with existing personality and stress symptom measures, with relatively few studies statistically controlling for personality factors. At present, there is a dearth of research demonstrating how EI impacts outcomes via multiple coping strategies and how the respective influences of EI and standard personality traits should be differentiated.

Table 2 Possible mediators for effects of emotional intelligence on adaptive coping and well-being

Mediator	*Mechanism: emotionally intelligent individuals...*
Circumventing stressful encounters	Identify, anticipate, and forestall socially threatening encounters
Constructive thought patterns	Identify and reconstruct faulty or unproductive appraisals and future expectancies
Richer personal coping resources	Access greater reserves of cognitive, emotional, social, physical, and spiritual resources in coping with unfamiliar challenges
Mood-regulation and repair	Understand sources of unwanted emotions and regulate them accordingly
Disclosure of past personal traumas	Restructure emotionally disturbing experiences to add meaning and coherence to life
Use of social support	Enlist others to support personal coping, assisted by high social skills
Flexibility of coping	Identify most appropriate strategy for a specific stressor, and switch strategy when required

Applications of Research

Understanding interrelationships between personality and coping may also contribute to tackling various applied issues. The role of dysfunctional coping is especially evident in *clinical psychology*. Use of strategies such as rumination and perseverative worry is commonly observed in patients diagnosed with mood and anxiety disorders. Coping strategies that work to increase awareness of symptoms and to block paths toward constructive management of the disorder may contribute to the onset and/or maintenance of emotional disorders, although there is debate over the causal role of coping in disorder. The Self-Referent Executive Function theory (S-REF: Wells and Matthews, 1994) sees dysfunctional coping as dynamically linked to intrusions of symptoms and harmful self-beliefs. For example, the socially anxious individual may cope by avoiding social encounters, which in turn reinforces beliefs in personal incompetence and reduces opportunities to acquire social skills. Training improved coping skills may contribute to psychotherapy.

Misdirected coping efforts are also a target for interventions in *educational* and *school psychology*. Students may cope poorly with academic demands, for example, through procrastinating. Success in school is also bound up with more general personal development. Children whose personalities incline them to cope with conflicts with peers and teachers through withdrawal and worry (internalizing behaviors) or through aggression and confrontation (externalizing behaviors) may be at risk of academic failure as well as stress symptoms. Recent interest in social–emotional learning (SEL) programs includes efforts at training effective coping strategies for dealing with various academic and interpersonal demands. Subsequently, we consider test anxiety in more detail as an example of an educational issue for which understanding the roles of personality and coping is requisite.

Finally, the roles of stress and emotion in the workplace are increasingly acknowledged in *industrial-organizational psychology*. Personality traits are at least modestly associated with a variety of affective outcomes. For example, employees high in N experience lower job satisfaction and may be more prone to absenteeism, ill-health, and high job turnover (Matthews et al., 2009). Failure to cope effectively may contribute to such stress outcomes, and studies suggest that maladaptive coping strategies such as self-criticism may mediate associations between N and stress outcomes (Matthews et al., 2009). Understanding the role of personality may help workplace stress management programs to address the specific vulnerabilities in coping of the individual.

Conclusion

Personality is often expressed through emotional functioning and stress response. Theories of coping provide an important path toward understanding individual differences in emotion regulation. Standard personality traits are commonly correlated with measures of coping, sometimes substantially so. It remains challenging to specify the role of personality in the dynamic person–situation interaction that characterizes the stress process. However, coping processes potentially mediate

and moderate the impacts of personality on emotional and behavioral response to stress. The nature of adaptive patterns of coping is also somewhat elusive, but resilience in personality may be associated with coping effectiveness. Indeed, effective coping and emotion regulation may be a central feature of EI. At a practical level, understanding the role of coping will allow support and training geared to the individual's personality across a range of applied domains.

See also: Affect and Emotion, Anthropology of; Agreeableness; Conscientiousness; Emotional Intelligence and Competencies; Extraversion; Neuroticism; Openness to Experience.

Bibliography

Bolger, N., Zuckerman, A., 1995. A framework for studying personality in the stress process. Journal of Personality and Social Psychology 69, 890–902.

Carver, C.S., Connor-Smith, J., 2010. Personality and coping. Annual Review of Psychology 61, 679–704.

Connor-Smith, J.K., Flachsbart, C., 2007. Relations between personality and coping: a meta-analysis. Journal of Personality and Social Psychology 93, 1080–1107.

Elliott, D.C., Kaliski, P., Burrus, J., Roberts, R.D., 2013. Exploring adolescent resilience through the lens of core self-evaluations. In: Prince-Embury, S., Saklofske, D. (Eds.), Resilience in children, adolescents, and adults: Translating research into practice. Springer, New York, pp. 199–212.

Endler, N., Parker, J., 1990. Multidimensional assessment of coping: a critical review. Journal of Personality and Social Psychology 58, 844–854.

Lazarus, R.S., 1999. Stress and Emotion: A New Synthesis. Springer, New York.

Lazarus, R.S., Folkman, S., 1984. Stress, Appraisal and Coping. Springer, New York.

Matthews, G., Deary, I.J., Whiteman, M.C., 2009. Personality Traits, third ed. Cambridge University Press, Cambridge.

Matthews, G., Emo, A.K., Funke, G., Zeidner, M., Roberts, R.D., Costa Jr., P.T., Schulze, R., 2006. Emotional intelligence, personality, and task-induced stress. Journal of Experimental Psychology, Application 12, 96–107.

Matthews, G., Szalma, J., Panganiban, A.R., Neubauer, C., Warm, J.S., 2013. Profiling task stress with the Dundee stress state questionnaire. In: Cavalcanti, L., Azevedo, S. (Eds.), Psychology of Stress: New Research. Nova Science, Hauppage, NY, pp. 49–90.

Matthews, G., Warm, J.S., Shaw, T.H., Finomore, V.S., 2014. Predicting battlefield vigilance: a multivariate approach to assessment of attentional resources. Ergonomics 57, 856–875.

McCrae, R.R., 2009. The five-factor model of personality traits: consensus and controversy. In: Corr, P.J., Matthews, G. (Eds.), The Cambridge Handbook of Personality Psychology. Cambridge University Press, Cambridge, pp. 148–161.

Ormel, J., Riese, H., Rosmalen, J.G., 2012. Interpreting neuroticism scores across the adult life course: immutable or experience-dependent set points of negative affect? Clinical Psychology Review 32, 71–79.

Parker, J.D.A., Wood, L.M., 2008. Personality and the coping process. In: Boyle, G.J., Matthews, G., Saklofske, D.H. (Eds.), The SAGE Handbook of Personality Theory and Assessment. SAGE, Los Angeles, pp. 506–519.

Salovey, P., Bedell, B.T., Detweiler, J.B., Mayer, J.D., 1999. Coping intelligently: emotional intelligence and the coping process. In: Snyder, C.R. (Ed.), Coping: The Psychology of What Works. Oxford, New York, pp. 141–164.

Schwarzer, R., Schwarzer, C., 1996. A critical survey of coping instruments. In: Zeidner, M., Endler, N. (Eds.), Handbook of Coping: Theory, Research, Applications. Wiley, New York, pp. 107–133.

Skodol, A.E., 2010. The resilient personality. In: Reich, J.W., Zautra, A.J., Hall, J.S. (Eds.), Handbook of Adult Resilience. Guilford Press, New York, pp. 112–125.

Wells, A., Matthews, G., 1994. Attention and Emotion: A Clinical Perspective. Lawrence Erlbaum, Hove, UK.

Zeidner, M., Matthews, G., Roberts, R.D., 2006. Emotional intelligence, coping with stress, and adaptation. In: Ciarrochi, J., Forgas, J., Mayer, J.D. (Eds.), Emotional Intelligence in Everyday Life: A Scientific Inquiry, second ed. Psychology Press, Philadelphia, PA, pp. 100–125.

Zeidner, M., Roberts, R., Matthews, G., 2009. What We Know about Emotional Intelligence: How It Affects Learning, Work, Relationships and Our Mental Health. MIT Press, Cambridge, MA.

Zeidner, M., Saklofske, D., 1996. Adaptive and maladaptive coping. In: Zeidner, M., Endler, N.S. (Eds.), Handbook of Coping: Theory, Research, Applications. Wiley, New York, pp. 505–531.

Coping with Job Stress

Wilmar B Schaufeli, Utrecht University, Utrecht, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The psychological process of dealing with job stressors in order to reduce strain or negative health effects is called coping. More specially, coping with job stress refers to mental and behavioral strategies of employees to handle the stressors they encounter at work. Coping with job stress involves the dynamic interplay of the employee and his or her job environment or work setting that is appraised as stressful in terms of harm, threat, or loss. More recent conceptualizations of coping broaden the concept by emphasizing its proactive nature, that is, the mastery of positively appraised challenging demands.

Coping with job stress is a key concept in understanding people's adaptation to their work roles. At work, employees are confronted with various kinds of demands, which may become 'stressors' when they tax or exceed the employee's adaptive capabilities. Examples of common job stressors include work overload, role problems, poor job control, lack of support from supervisors and coworkers, and interpersonal conflicts. These stressors may lead to negative psychological (e.g., depression, irritability, and burnout), physical (e.g., headaches, heart palpations, and hyperventilation), and behavioral (e.g., absenteeism, turnover, and violence) symptoms or 'strains.' The psychological process that is involved in dealing with these job stressors in order to reduce strain is called 'coping.' More specially, coping with job stress refers to the mental and behavioral strategies that employees use to handle the stressors they encounter at work.

The Concept of Coping

Rather than a homogeneous concept, coping represents a diffuse umbrella term. Nevertheless, a widely used definition of coping exists that refers to the constantly changing cognitive and behavioral efforts a person makes to manage demands that tax or exceed his or her personal resources (Lazarus and Folkman, 1984). Coping refers to a dynamic interplay between person and environment that occurs when an individual appraises a situation or an event as stressful. Applied to the work situation, this interplay involves the employee and the work setting that is appraised as stressful. It is important to note that coping refers merely to the effort made and not to the quality of the outcome. That means that coping may be either effective or successful in preventing, avoiding, or controlling individual distress or it may be ineffective or unsuccessful. For instance, compensatory or comfort behaviors such as excessive eating, smoking, drinking, or working commonly are maladaptive coping strategies. Furthermore, coping may refer to overt actions or observable behaviors as well as to covert, unobservable thoughts and events, and it may occur as a response to an event or in anticipation of upcoming demands.

The Historical Development of Coping

Initially, coping was studied from a psychodynamic point of view in the context of general human adaptation. In the first half of the twentieth century, leading psychoanalysts like Freud, Adler, and Jung described adjustment and defense mechanisms by which individuals were supposed to deal with intrapsychic conflict. Building upon this earlier work, in the 1960s and 1970s, ego psychologists distinguished between 'coping' as an advanced or mature ego process leading to successful adaptation and 'defense' as an unsuccessful neurotic mode of adaptation.

The second approach emerged from the 1960s and conceived coping in terms of personality traits. Relatively stable dispositions or coping styles (i.e., denial, repression, and vigilance) were expected to determine the individual's coping behavior across various situations. Although empirical research has shown that people possess stable and consistent coping preferences, it is likewise true that individuals are able to vary their responses to stressful situations in accordance with situational constraints and possibilities. Personality traits have made a comeback in the 1970s and 1980s as so-called personal coping resources like 'hardiness,' 'dispositional optimism,' 'self-efficacy,' and 'sense of coherence.' Such coping resources were expected to foster successful coping since the individual can draw upon them when under stress.

The currently dominant transactional approach originates from the work that Lazarus and his colleagues have performed since the 1950s. In contrast to the psychoanalytic approach, coping is defined in this approach by an effort that is independent from its success and includes proactive as well as defensive strategies. Furthermore, coping is considered to be a dynamic and transactional process involving person and environment rather than a personality disposition.

Following the emergence of positive psychology, the concept of coping has been broadened since the turn of the century and includes self-regulated goal attainment strategies and personal growth as well. In other words, coping is expanded beyond dealing with adverse events and extends to mastering challenges.

Coping has been applied to various aspects of human life. Initially, it was used predominantly in clinical and health psychology to understand how patients deal with chronic disease, substance use, loss, chronic pain, and trauma. Also, minor coping efforts involving so-called daily hassles have been identified and investigated. Coping with job stress (including unemployment) has been studied since the end of the 1970s and involves long-term chronic job stressors (e.g., future job insecurity), short-term daily hassles at the job

 http://dx.doi.org/10.1016/B978-0-08-097086-8.14010-3

(e.g., interpersonal conflict), as well as acute severe stressors (e.g., armed robbery). Most recently, ambitious goal setting and tenacious goal pursuit (e.g., learning new job skills) have been studied by way of 'positive coping,' which promotes personal growth and development (Schwarzer and Knoll, 2003).

General Coping Theory

The leading cognitive-behavioral theory of stress and coping of Richard Lazarus and his coworkers (*see* Stress, Coping, and Health) may be applied to work situations as well. It assumes that a particular situation or event is appraised negatively as being stressful in terms of threat, loss, or harm. Alternatively, the situation or event may also be appraised neutral or positive as a challenge. This so-called primary appraisal refers to the extent to which something is at stake for the person. For instance, a conflict with a colleague may threaten the employee's self-esteem, whereas unemployment is likely to be perceived as a loss of status and income. 'Secondary appraisal' refers to options the person has to cope with the harm, threat, or loss. These depend on personal characteristics, such as motivation, beliefs about oneself and the world, as well as on resources for coping such as financial means, social and problem-solving skills, particular personality traits, and a favorable work situation with social support from one's colleagues and supervisors. It is important to note that both appraisal processes, primary (demand) appraisal and secondary (resource) appraisal, are two components of one cognitive process that takes place simultaneously instead of successively.

Generally speaking, a main distinction is made between two targets: (1) instrumental, problem-focused coping, which involves attempts to deal with stress by directly altering the situation and (2) palliative, emotion-focused coping, which involves attempts to regulate the emotional responses to a stressful situation. For instance, employees might consult their supervisors after a conflict with a colleague (problem focused) or they may try to relax or divert themselves (emotion focused). In addition, various methods of coping are distinguished such as active cognitive (i.e., intrapsychic attempts to manage the appraisal of the stressor through, for instance, constructive dialogue or positive comparison), active behavioral (i.e., overt attempts to deal directly with the stressor), and avoidance (i.e., avoiding or denying the stressor). Although targets and methods of coping can be distinguished theoretically, this distinction is usually quite difficult to make at the practical level.

After the coping effort has been completed, 'reappraisal' of the initial stressful event takes place along the same lines as the original primary appraisal. The conflict with one's colleague may be solved or one may have found a new job. In that case, the reappraisal is positive and there is no need for further coping efforts. However, in case initial coping efforts failed and the event is reappraised as stressful, coping attempts continue.

Originally, Lazarus' stress and coping theory was neither applied to future demands nor to positive, challenging demands. Recently, anticipatory coping has been defined as an effort to deal with pending harm, threat, or loss; the person faces a negative event that is fairly certain to occur in the near future. In a somewhat similar vein, preventive coping is an effort to prepare for uncertain events in the longer run by building up the necessary resources. Such future-directed coping efforts seem to be particularly relevant in today's continuously changing world of work. Finally, proactive coping reflects efforts to build up general resources that facilitate promotion toward challenging goals and personal growth. Rather than by negative appraisals, proactive coping is sparked by positive challenges.

The Individual and Organizational Perspective

In Table 1, a taxonomy of individual ways of coping with job stress is displayed, including some examples of coping behaviors. Typically, in occupational settings, workplace strain and coping are individual as well as organizational issues. This means that interventions to reduce workplace strain may be targeted at the individual and/or the organizational level (Quick et al., 1997).

Interventions that focus on the individual include job stress education activities, relaxation, cognitive-behavioral programs, and employee skills training in such areas as time management and assertiveness. Interventions that focus on the organization seek to reduce strain by changing or improving the work setting (e.g., job redesign, selection and training policies, and organizational development). Although organization-based interventions offer considerable potential for reducing or eliminating

Table 1 Ways of coping with job stress

Method		Problem/task	Emotions/reactions
	Cognitive	Control (e.g., planning, organizing, and prioritizing work assignments)	Control (e.g., try to think of oneself as a winner, as someone who always comes through)
		Escape (e.g., try to pay attention only to your duties in order to overlook difficulties)	Escape (e.g., to tell oneself that difficulties are unimportant)
	Behavioral	Social (e.g., sit down with somebody and talk things over)	Social (e.g., to express one's irritation to other colleagues or to let off steam)
		Solitary (e.g., to do what has to be done, one step a time)	Solitary (e.g., spend time on a hobby)
		Control (e.g., delay or leave some of one's normal job responsibilities)	Control (e.g., show no emotional reaction and do not communicate distress to anyone)
		Escape (e.g., get busy with other things to keep one's mind off the problem)	Escape (e.g., change to nonwork activities, take pills, smoke more)

Adapted from Latack, L.C., Havlovic, S.J., 1992. Coping with job stress: a conceptual evaluation framework for coping measures. Journal of Organizational Behavior 13, 497–508.

job stressors, most workplace initiatives focus on the individual employee. That is, these interventions improve the employee's resilience by changing their maladaptive beliefs, improving their lifestyle, exercising stress management skills, and augmenting coping resources. Different individual programs are available including onsite fitness facilities, health screening, dietary control, smoking cessation, cardiovascular fitness programs, relaxation classes, stress management training, anti-burnout workshops, stress and health education, and psychological counseling. Although these activities are aimed at the individual, this does not exclude collective coping efforts (Pieró, 2008). Collective coping occurs when a group (work team, department) faces a common perceived threat or noxious situation that collectively initiates actions, for instance, by preventing, eliminating, or reducing the stressful situation, by positively reinterpreting the situation, or by alleviating its negative consequences.

Research

Earlier studies on coping with job stress often found that employees use fewer problem-focused than emotion-focused strategies directed at changing the problems encountered in the context of work. This was interpreted by noting that the very nature of problems at work is perhaps not ameliorated easily by attempts to change them. It was reasoned that employees are perhaps better off by developing a renewed cognitive perspective and by regulating their negative emotions. More recent studies, however, point to the opposite; currently employees seem to favor problem-focused coping instead of emotion-focused coping. More detailed investigations have shown that the level of job control plays a crucial role. When a high level of job control exists, problem-focused strategies prevail, but when job control is poor employees seem to favor emotion-focused strategies. Unfortunately, the latter are generally less effective compared to the former. For instance, studies on organizational changes, such as corporate restructuring and downsizing, suggest that employees cope rather poorly with these events, which are largely beyond their control. They use predominantly emotion-focused and avoidant coping strategies and suffer long-term adverse mental health.

As a general rule, a more active, problem-focused, and control-oriented coping strategy goes along with less (mental) health symptoms and higher employee well-being. In contrast, emotion-focused strategies as well as escape-oriented or avoidant strategies are usually associated with poor (mental) health and reduced well-being. However, two factors seem to be important for successful coping: time frame and availability of coping resources. It seems that emotion-focused or escape-avoidant-oriented strategies are effective when dealing with short-term stressors, whereas problem-focused and control-oriented strategies are more effective when dealing with chronic or recurrent job stressors. Furthermore, employees who have more coping resources at their disposal appear to use more active coping efforts and less avoidant coping. Consequently, employees who can draw upon coping resources are more likely to deal successfully with job stress.

Generally speaking, individual-based interventions, particularly relaxation programs and cognitive-behavioral programs (e.g., rational emotive behavior education), seem to be effective in reducing employee strain (i.e., burnout, distress, depressed mood, anxiety, and psychosomatic symptoms), whereas the effectiveness of organization-based strategies is still to be demonstrated in a convincing way.

As far as the relationship between employee coping and sociobiographical factors (e.g., age, gender, and education) and job factors (e.g., level and type of job) is concerned, no conclusive empirical evidence is available (Dewe et al., 2010). For instance, among employees no consistent gender-specific or job-specific coping patterns have been observed. Most likely, this lack of conclusive evidence has to be attributed to the complex interaction patterns that exist among the factors involved and the idiosyncratic nature of coping with job stress.

Methodological Issues and Future Directions for Research

The measurement of coping is fraught with problems that are partly due to the fact that almost exclusively self-report questionnaires are used. More specifically, empirically developed coping scales yield dimensions or coping behaviors that differ from theory-based coping scales. Furthermore, the number of coping dimensions included in various questionnaires varies (e.g., from 2 to 28), and there is no agreement either on the hierarchy of these dimensions or on the fact whether coping should be assessed as dispositional (trait) or situational (state) characteristic. Although various measures exist to assess coping with job stress, there is no 'gold standard.' Given the complexity and the idiosyncrasy of coping, a standard way of assisting it is unlikely to be ever achieved. The future research challenge is to expand coping with job stress beyond its traditional meaning of an individual reaction to adversity by exploring how employees come to grapple with (1) future job demands (anticipatory and preventive coping), (2) positive challenges (proactive coping), and (3) common job stressors (collective coping).

See also: Control Beliefs: Health Perspectives; Occupational Health; Resilience; Self-Efficacy and Health; Stress at Work.

Bibliography

Dewe, P., O'Driscoll, M.P., Cooper, C.L., 2010. Coping with Work Stress: A Review and Critique. Wiley-Blackwell, Chichester.

Latack, L.C., Havlovic, S.J., 1992. Coping with job stress: a conceptual evaluation framework for coping measures. Journal of Organizational Behavior 13, 497–508.

Lazarus, R.S., Folkman, S., 1984. Stress, Appraisal, and Coping. Springer, New York.

Pieró, J.M., 2008. Stress and coping at work: new research trends and their implications for practice. In: Näswall, K., Hellgren, J., Sverke, M. (Eds.), The Individual in the Changing Working Life. Cambridge University Press, Cambridge, pp. 284–310.

Quick, J.C., Quick, J.D., Nelson, D.L., Hurrell, J.J., 1997. Preventive Stress Management in Organizations. American Psychological Association, Washington, DC.

Schwarzer, R., Knoll, N., 2003. Positive coping: mastering demands and searching for meaning. In: Lopez, S., Snyder, C.R. (Eds.), Positive Psychological Assessment: A Handbook of Models and Measures. American Psychological Association, Washington, DC, pp. 393–409.

Core Self-Evaluations

Russell E Johnson and Chu-Hsiang (Daisy) Chang, Michigan State University, East Lansing, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article we review core self-evaluations (CSE), which refers to deep-seated evaluations that people hold of themselves. CSE is a higher order personality construct comprised of fundamental traits that have an evaluation focus and are broad in scope, such as self-esteem and generalized self-efficacy. People with high CSE have a heightened sensitivity to positive information coupled with a reduced sensitivity to negative information. CSE is an important dispositional predictor of satisfaction with one's job and life, and appraisals of one's job characteristics and social support. CSE also contributes to effective job performance behaviors at work.

A great deal of research by organizational scholars has examined how employees' personality and individual differences shape their thoughts and actions at work, especially job satisfaction and performance. Although satisfaction and performance are influenced by a myriad of external factors, such as characteristics of the job (e.g., the amount of autonomy inherent in work tasks) and social relations (e.g., fairness from supervisors), internal factors like self-esteem and emotional stability also contribute to satisfaction and performance at work (Judge and Bono, 2001). Although research on employee personality proliferated beginning in the early 1990s, a criticism of this research is that various dispositions and traits were examined in a piecemeal fashion with little attempt to consolidate and explain why they function as antecedents of job satisfaction and performance. In response to this criticism, the construct of core self-evaluations (CSE) was proposed by Judge et al. (1997).

CSE is a deep-seated evaluation that people hold of themselves, which ranges in valence from positive to negative. Evaluations of the self are thought to be the most basic appraisals that people make, which spillover to influence reactions to other people, things, and situations. Appraisals of one's job and coworkers, for example, mirror how people evaluate themselves. An employee who sees herself in a positive light will also view her work circumstances more favorably. Given its deep-seated nature, CSE is believed to underlie other dispositions and traits, which are manifestations of the latent CSE construct (Judge et al., 1997). Thus, CSE theory avoids the criticism of piecemeal research by providing an integrative framework that identifies the key criteria shared by dispositional and personality sources of job satisfaction and performance.

Three criteria are shared by traits that reflect CSE: evaluation focused, fundamental, and broad in scope. *Evaluation focused* is the extent to which traits reflect evaluations of the self as opposed to description of the self. For example, self-esteem is evaluative because it is an appraisal of self-worth (Rosenberg, 1965) whereas openness to experience is descriptive because it reflects preferences for being imaginative and intellectually curious (Goldberg, 1993). *Fundamental* is the extent to which traits are central to one's self-concept rather than peripheral. Central traits have a high degree of interconnectivity to other traits, attitudes, and perceptions. As an example, neuroticism (or its converse, emotional stability) is associated with the tendency to make negative or threatening evaluations and to experience anxious emotions (Eysenck, 1967), and it has a dense web of connections to other traits that similarly reflect sensitivities to threats and anxiety (e.g., avoidance motives; Lanaj et al., 2012). Finally, *broad in scope* refers to the breadth of the trait. Broad or global traits reflect what a person is generally like across time and different situations, whereas narrow traits are manifested only under specific circumstances.

Trait Indicators of CSE

Based on the aforementioned criteria, four CSE traits were identified: self-esteem, generalized self-efficacy, emotional stability, and locus of control (Judge et al., 1997, 1998b). *Self-esteem* refers to a person's overall appraisal of their personal and social worth (Rosenberg, 1965). Self-esteem is a fundamental evaluation that predicts a variety of attitudes and behaviors at work (Pierce and Gardner, 2004). *Generalized self-efficacy* is an evaluation of one's ability to perform well and cope successfully across a range of situations (Chen et al., 2004). In contrast to self-esteem, generalized self-efficacy is an appraisal of one's capabilities rather than self-worth. *Emotional stability* is the tendency to avoid making negative evaluations and experiencing anxious emotions (Eysenck, 1967). People who score high on emotional stability view themselves and their environment in a positive light and are not overly sensitive to perceiving threats. *Locus of control* is an evaluation of whether or not one's behavior is capable of altering the environment in desired ways (Rotter, 1966). Employees with an internal locus of control believe they can shape work events, which elicits positive reactions toward their job (Spector, 1982).

Empirical evidence appears to support the status of these four traits as indicators of an underlying CSE construct. For example, meta-analytic results indicate that the four traits are highly correlated with one another (Judge et al., 2002). The average meta-analytic estimate is $\rho = +0.60$, with the strongest relationship between self-esteem and generalized self-efficacy ($\rho = +0.85$) and the weakest relationship between emotional stability and locus of control ($\rho = +0.40$). Factor analytic results also suggest that the four traits load on a single higher order latent construct (e.g., Erez and Judge, 2001). Self-esteem and generalized self-efficacy have the highest loadings on the construct, whereas locus of control

has the lowest. That the four traits share a common underlying core is also hinted at by research examining relationships of the traits with outcomes. Self-esteem, generalized self-efficacy, emotional stability, and locus of control all have comparable correlations with job satisfaction and performance (Judge and Bono, 2001), and the individual traits do not appear to be more predictive of outcomes than the shared variance between them (Judge et al., 2003).

Despite this supportive evidence, some issues regarding the structural validity of the higher order CSE construct have been identified (Chang et al., 2012; Johnson et al., 2008). One issue pertains to possible errors of omission. That is, besides the four traits mentioned above, other traits also appear to be evaluation focused, fundamental, and broad in scope. For example, trait optimism causes people to evaluate experiences in a positive light via the favorable attributions they make (e.g., optimistic employees attribute positive events like promotions to internal and stable causes; Gillham et al., 2001). Similarly, traits that reflect an approach temperament, such as positive affectivity and extraversion, elicit positive evaluations of one's self and the environment (Johnson et al., 2008). It has therefore been proposed that optimism and positive affectivity, as well as other candidate traits (e.g., conscientiousness, negative affectivity), also warrant consideration as indicators of CSE (Bono and Judge, 2003; Johnson et al., 2008).

A second issue pertains to possible errors of commission. In particular, the status of locus of control as an indicator of CSE has been questioned for both conceptual and empirical reasons. Conceptually, locus of control is more reflective of beliefs about the controllability of the environment rather than evaluations about one's capabilities. People with an internal locus of control believe the environment is responsive to their actions, whereas those with an external locus attribute events and outcomes to chance (Rotter, 1966). Although locus of control is a core evaluation, it is not necessarily a core evaluation about the self. Empirically, locus of control has the weakest correlations with the other CSE traits and the lowest loading on the latent CSE construct (Johnson et al., 2008; Judge et al., 2002). In fact, when the traits are measured using techniques that reduce biased responding due to social desirability and mood, locus of control no longer correlates with the other traits nor does it load on the CSE construct (Johnson et al., 2011). There is therefore need for additional research that evaluates the status of the aforementioned traits as indicators of the CSE construct.

A final issue is the need to establish the nature of the relationship between the latent CSE construct and its trait indicators, of which there are two possibilities. One possibility is that the underlying CSE construct is responsible for people's standing on each trait, where the direction of causality flows from CSE to the traits. According to this view, CSE is a superordinate construct with effects indicators, and the traits are manifestations of CSE and interchangeable with one another (Johnson et al., 2012). An example of a superordinate construct is general cognitive ability, which fuels people's lower level verbal, quantitative, and analytic abilities. Another possibility is that CSE is a formative construct such that the traits are causal indicators and the direction of causality flows from the traits to CSE. An example of a formative construct is socioeconomic status, which is the sum of earners' household income, education, and occupation. Although the contemporary view is that CSE is a superordinate construct, this view has received criticism because the traits are not interchangeable nor do they have identical antecedents (Johnson et al., 2008). For example, past success is a key driver of self-esteem and self-efficacy, whereas emotional stability is rooted more in neurological systems.

In sum, the significant overlap among various self-evaluative traits suggests the existence of a latent CSE construct, but further research is needed to confirm the full set of trait indicators and the nature of the relationship between the CSE construct and its traits (Chang et al., 2012). Identifying the correct traits is necessary in order to properly model the latent CSE construct, and verifying whether CSE is a superordinate or formative construct is necessary because empirical findings are biased when latent constructs are misspecified (Johnson et al., 2012).

Measuring CSE

Given that CSE is a latent construct comprised of multiple trait indicators, there are different approaches for measuring CSE (Chang et al., 2012). Initial research utilized indirect approaches in which the traits are measured separately and then used to estimate people's CSE. One way is to aggregate all of the item-level data from the trait scales into an overall score. Alternatively, trait scores can first be computed and then aggregated into an overall score. These indirect approaches have two major drawbacks. First, CSE reflects the shared variance among the traits, but aggregating item- or trait-level data into an overall score mixes together the shared variance among the traits and the unique variance associated with each trait. The unique variance does not reflect CSE, thus it should not be included. Second, computing a simple sum or average of the data is problematic because it gives equal weight to the traits. As mentioned above, this assumption is incorrect: self-esteem and generalized self-efficacy have the highest loadings on the construct, followed by emotional stability and then locus of control. Thus, aggregating the data biases the CSE construct by overestimating the contribution of locus of control and emotional stability.

A better indirect approach is to model CSE as a higher-order latent factor via factor analysis and structural equation modeling. This approach is advantageous because it distinguishes between shared and unique variance, and only the shared variance is used to generate the CSE factor. Thus, traits with weaker relations to the CSE construct (locus of control and emotional stability) do not unduly bias the CSE factor. Another advantage of modeling CSE as a higher order latent factor is that the strength of the relationship that CSE has with other variables (e.g., job satisfaction) can be compared to the strength of the relationship that each trait has with the same variable. Doing so enables researchers to evaluate the utility of the higher order CSE construct vis-à-vis the lower order traits. When individual traits have greater predictive validity, it is more appropriate (and parsimonious) to examine the traits rather than the latent construct (Johnson et al., 2012).

CSE can also be measured via a direct approach, which involves creating survey items that specifically tap the underlying CSE construct. A popular direct measure is the Core Self-Evaluations Scale (CSES; Judge et al., 2003) that includes items spanning multiple traits. The primary advantage of direct measures is their short length (e.g., the CSES has 12 items whereas measuring the individual traits separately has anywhere from 20 to 40 items). Direct measure are, however, limited in that they do not preserve the trait structure of CSE, making it impossible to assess the utility of the higher order CSE construct against the lower order traits. This tradeoff can be balanced by developing indirect measures with fewer items per trait (e.g., measuring each trait with 4 items would only require 16 items). Indirect measures are preferable because they allow researchers to tease apart variance that is shared versus unique to each trait (Johnson et al., 2012).

Core Self-Evaluation and Its Outcomes

CSE has proved to be a useful predictor of job satisfaction and performance, among other affective, cognitive, and behavioral outcomes. It was proposed that CSE may influence outcomes through four processes (Judge et al., 1997). First, the positive self-views associated with high CSE may directly influence outcomes through positive emotional generalization. Next, CSE may influence outcomes by altering the cognitive appraisals that individuals make about the characteristics of their environment. Third, CSE may influence outcomes by influencing the actions that individuals engage in. Finally, CSE may moderate relationships between events and individuals' reactions (e.g., negative events are less detrimental to well-being when CSE is high).

More recently, scholars (Chang et al., 2012; Ferris et al., 2011, 2013) have integrated CSE with the approach/avoidance framework to provide a comprehensive theoretical framework for explaining the effects of CSE. This framework suggests that individual traits can be characterized by the extent to which they represent tendencies to approach positive stimuli versus avoid negative stimuli (Elliot and Thrash, 2002). These differences in sensitivities toward positive (approach orientation) and negative (avoidance orientation) information have implications for individuals' affect, cognition, and behavior. An approach orientation leads individuals to experience emotions ranging from elation to depression, focus more on positive information when evaluating a given situation, and adopt goals that are directed toward pursuing desirable outcomes (Elliot and Thrash, 2002; Johnson et al., 2013). In contrast, individuals with an avoidance orientation tend to experience emotions ranging from relief to anxiety, pay more attention to negative information, and adopt goals for avoiding undesirable outcomes. Consistent with an approach/avoidance framework, meta-analytic findings verify that CSE has positive relationships with approach-oriented traits like positive affectivity ($\rho = +0.61$) and approach motives ($\rho = +0.32$), whereas it has negative relations with avoidance-oriented traits like negative affectivity ($\rho = -0.60$) and avoidance motives ($\rho = -0.54$). Below we summarize research evidence relating CSE to various work-related outcomes, organized by the affective, cognitive, behavioral, and moderation processes discussed by Judge et al. (1997).

Affective Outcomes

Individuals with high CSE are likely to experience positive spillover effects from self-views to evaluations of their environment. Moreover, owing to high approach and low avoidance orientations, high-CSE individuals pay closer attention to positive aspects of their daily encounters, thus experiencing positive emotions such as excitement. CSE is therefore positively related to job satisfaction (i.e., a positive emotional state resulting from one's job and experiences at work) and affective organizational commitment (i.e., accepting organizational goals, exerting effort on behalf of the organization, and maintaining membership in the organization; Meyer and Allen, 1997). Meta-analytic findings confirm that CSE is positive related to these outcomes ($\rho = +0.36$ for job satisfaction; $\rho = +0.30$ for affective commitment; Chang et al., 2012).

In addition to positive feelings about their jobs and organizations, high-CSE individuals also report high life satisfaction ($\rho = +0.54$). This may reflect their general sensitivity toward positive events, as well as their higher optimism and resilience. Meta-analytic results indicate that CSE is negatively associated with stress-based reactions like psychological strain (e.g., negative emotions: $\rho = -0.49$) and physical strain (e.g., psychosomatic symptoms: $\rho = -0.42$).

Cognitive Outcomes

Beyond emotional generalization, CSE may also influence outcomes via cognitive appraisals. Specifically, when evaluating environmental characteristics, high-CSE individuals tend to hold positive appraisals of their environment and rarely view circumstances as threatening or overwhelming. For example, CSE has significant, positive relationships with perceived job characteristics ($\rho = +0.39$), such that jobs are viewed as having high autonomy and opportunities to demonstrate one's knowledge and skills (Chang et al., 2012). High CSE is also related to more favorable appraisals of fairness ($\rho = +0.15$) and perceived support from companies ($\rho = +0.32$), supervisors ($\rho = +0.23$), and coworkers ($\rho = +0.28$; Chang et al., 2012). Moreover, meta-analytic evidence also suggests that CSE is negatively related to perceptions of job stressors (Chang et al., 2012), including role ambiguity ($\rho = -0.45$), interpersonal mistreatment ($\rho = -0.23$), perceptions of organizational politics ($\rho = -0.25$), and work–family conflict ($\rho = -0.28$). These results are consistent with the theoretical linkage between CSE and low avoidance orientation, such that high-CSE individuals are less sensitive to negative aspects of their environment.

Behavioral Outcomes

CSE also influences the types of actions that employees engage in. In employment settings, three types of behaviors receive the most research attention: task performance, organizational

citizenship behavior (OCB), and counterproductive work behavior (CWB). Task performance refers to work behaviors that contribute to the technical core of the organization, such as the production of a product or delivery of a service (Borman and Motowidlo, 1997). OCB includes actions that contribute positively to the general social and psychological environment at work, such as helping coworkers, participating in voluntary committees, and defending the company's reputation to outsiders (Borman and Motowidlo, 1997). CWB refers to actions that harm the organization or its members, such as sabotage, theft, and interpersonal mistreatment (Spector, 1997).

Task performance and OCB are approach-oriented behaviors that contribute positively to the organization and its members. As such, high-CSE employees are expected to exhibit higher task performance and OCB. On the contrary, CWB is often the result of experiencing negative emotions and perceiving stressors, which are associated with an avoidance orientation. Thus, high-CSE employees should be less likely to engage in counterproductive behaviors. Meta-analytic results are consistent with this line of reasoning (Chang et al., 2012). Specifically, CSE has positive relations with task performance ($\rho = +0.19$), and with OCB directed at coworkers ($\rho = +0.20$) and the organization ($\rho = +0.30$). In contrast, CSE has negative relations with CWB directed at coworkers ($\rho = -0.16$) and the organization ($\rho = -0.28$). Taken together, high-CSE employees contribute positively to their organizations by fulfilling their job responsibilities and improving the social and psychological environment, and by refraining from harmful and counterproductive acts.

Mediating Effects

In addition to direct effects of CSE on outcomes, a number of empirical studies have examined how cognitive appraisals and goal-related actions mediate the effects of CSE. For example, high-CSE employees tend to make more positive evaluations of their job characteristics and environment, which in turn result in higher job and life satisfaction (Judge et al., 1998a). As another example, CSE also has positive effects on job satisfaction and affective commitment via the upward and downward comparisons that employees make with coworkers (Brown et al., 2007). In addition to cognitive appraisals, goal-based variables have been examined as mediators of the effects of CSE. One such mediator is setting performance goals, such that high CSE leads to stronger goal setting motivation which in turn leads to higher task performance (Erez and Judge, 2001). Job engagement (i.e., the devotion of physical, attentional, and affective resources toward one's work-related goals) has also been found to mediate the effects of CSE on task performance as well as OCB (Rich et al., 2010). Finally, Ferris et al. (2011) examined the work-based approach and avoidance motivations as mediators of the effects of CSE on employee performance. Consistent with the integrated theoretical framework discussed above, CSE had positive, indirect effects on task performance and OCB, and negative indirect effects on CWB via approach and avoidance motivations.

Moderating Effects

A fourth process through which CSE may influence outcomes is by moderating employees' reactions toward environment stimuli, including stressors and job characteristics. For example, scholars have argued that CSE is an important personal resource that enables employees to effectively cope with stressors. High CSE may provide individuals with a strong sense of control and a positive outlook for their coping efforts, thus high-CSE individuals experience less detrimental effects when faced with potential stressors. Empirical studies has provided support for the buffering effects of CSE, such that social stressors (Harris et al., 2009) and organizational constraints (Best et al., 2005) have weaker negative effects on employees' well-being when CSE is high (vs low). Unfortunately, the buffering effects of CSE appear inconsistent, such that some scholars failed to find any buffering effects (e.g., Kammeyer-Meuller et al., 2009), whereas others found that high CSE actually enhanced the negative effects of stressors on employee performance (Kacmar et al., 2009).

In addition to moderating the effects of stressors, CSE also shapes the effects of perceived job characteristics on outcomes. In support of the view that the positive aspects of one's job are more salient for high-CSE employees, it was found that high CSE strengthened the relationship between perceived job characteristics and organization-based self-esteem (Pierce and Gardner, 2009). However, similar to above, these moderating effects are also inconsistent. While some studies failed to find a moderation effect (Judge et al., 1998a), others have observed that low CSE strengthened the beneficial effects of job characteristics (Ng and Feldman, 2010). Taken together, the moderating role of CSE remains unclear.

Additional Issues Concerning CSE

As highlighted in this review, some questions about CSE remain. Some of them pertain to the structural validity of the construct, such as whether additional traits (e.g., optimism) should be included as indicators, whether existing indicators should be excluded (e.g., locus of control), and whether CSE is a superordinate construct with effects indicators or a formative construct with causal indicators. There is also room to improve how CSE is measured, such as by developing shorter indirect measures with specific items that assess each trait. Besides these issues, which were already covered, we discuss two additional ones below.

The first issue is in regards to possible drawbacks of having high levels of CSE. The meta-analytic review of the CSE literature by Chang et al. (2012) clearly indicates that CSE contributes in favorable ways to numerous outcomes, including employee satisfaction, well-being, and performance. It has been suggested, however, that especially high levels of CSE may be too much of a good thing (Hiller and Hambrick, 2005). For example, excessive CSE may give rise to unrealistically high levels of overconfidence and personal control over the environment. Thus, high-CSE people may be prone to suboptimal performance and decision-making by not exercising due diligence when gathering and analyzing available data. Excessive CSE may also cause individuals to disregard the

advice of other people as well as information from other external sources (e.g., market reports, financial indicators). If they ignore external sources of information, especially when it is inconsistent with their personal views, then high-CSE individuals will fail to adapt to changing circumstances and instead persist in less effective (or even destructive) courses of action. This logic suggests that CSE may have curvilinear relationships with performance, such that performance improves as CSE increases from low to high levels, but performance will begin to deteriorate at especially high levels of CSE (an inverted-U pattern). Research is needed that explores possible links of excessive CSE with the behavioral consequences of overconfidence and hubris (Hiller and Hambrick, 2005).

A second issue is to explore the utility of CSE for human resource management. Although research on CSE originated in the organizational and behavioral sciences, little work has explored the potential of CSE to inform practice. Given the significant relationships of CSE with multiple aspects of job performance (e.g., task performance, OCB, and CWB), assessing CSE as part of the selection process may improve the accuracy of hiring decisions. What first needs to be determined, however, is whether CSE predicts variance in job performance incremental to other established predictors, such as cognitive ability and conscientiousness. An initial evaluation of CSE's utility for selection is unclear at this time because its relationships with other established predictors are small in some cases ($\rho = +0.12$ with cognitive ability) but large in others ($\rho = +0.49$ with conscientiousness; Chang et al., 2012). Thus, CSE may not explain much incremental variance in job performance beyond predictors like conscientiousness.

The appropriateness of CSE for selection purposes also depends on other factors. For example, hiring people based on CSE scores would not be recommended if doing so results in high adverse impact (i.e., disproportionately fewer people of a particular race, sex, or ethnicity group are selected). Hiring based on CSE scores would also be problematic if applicants are prone to fake responses on measures of CSE. This is a real concern for CSE because people are especially motivated to paint a desirable picture of themselves when responding to self-evaluative items (Crampton and Wagner, 1994). Consistent with this idea, preliminary empirical evidence suggests that scores on CSE measures are biased to some extent by social desirable responding (Johnson et al., 2011). Thus, before CSE can be incorporated into selection practices, more work is needed concerning its incremental validity vis-à-vis other predictors, the extent to which it creates adverse impact, and the extent to which applicants are motivated and capable of faking on CSE measures.

Another human resource function that CSE may impact is performance management. Key ingredients that drive effective performance appraisal systems are goal setting and feedback. First, employees with high CSE may be prone to setting more difficult goals for themselves and view challenging tasks as opportunities to learn. Indeed, Chang et al.'s (2012) review concluded that CSE is positively related to goal difficulty ($\rho = +0.22$), goal commitment ($\rho = +0.42$), and learning goal orientation ($\rho = +0.33$). Thus, high levels of CSE start employees off on the right foot when it comes to goal setting and striving. Second, high levels of CSE may also facilitate the feedback process that follows goal setting and striving. For example, owing to their positive self-regard, employees with high CSE may respond more constructively to negative performance feedback because they believe themselves capable of altering their behavior and/or the work environment to make the necessary improvements. Their positive self-regard may also give employees the confidence to seek feedback from others, especially negative feedback, and then use that feedback as a launching point for self-development. Employees with low CSE, in contrast, tend to shy away from negative feedback and avoid difficult tasks that draw attention to deficiencies in their knowledge and skills (the very same tasks that, ironically, would help them improve their knowledge and skills). Taken together, more attention ought to be paid to the practical implications of CSE for human resources management.

Conclusion

CSE, which refers to deep-seated evaluations that people hold of themselves, is a higher order latent factor comprised of fundamental traits that have an evaluative focus and are broad in scope (e.g., self-esteem and generalized self-efficacy; Judge et al., 1997). People with high CSE have a heightened sensitivity to positive information (i.e., strong approach motivation) coupled with a reduced sensitivity to negative information (i.e., weak avoidance motivation). Thus, it is not surprising that CSE is related to higher satisfaction with one's job and life, more favorable appraisals of job characteristics and social support, and less strain and psychosomatic symptoms (Chang et al., 2012). CSE also contributes to effective behaviors at work and reduces destructive behaviors. Although there are some issues concerning the nature and measurement of CSE that require attention (Chang et al., 2012; Johnson et al., 2008), CSE is a useful higher order personality construct that sheds light on the dispositional sources of work attitudes, well-being, and behavior.

See also: Job Satisfaction; Neuroticism; Organizational Citizenship Behavior; Organizational Commitment; Personnel Selection, Psychology of; Workplace Deviance.

Bibliography

Best, R.G., Stapleton, L.M., Downey, R.G., 2005. Core self-evaluations and job burnout: the test of alternative models. Journal of Occupational Health Psychology 10, 441–451.

Bono, J.E., Judge, T.A., 2003. Core self-evaluations: a review of the trait and its role in job satisfaction and job performance. European Journal of Personality 17, S5–S18.

Borman, W.C., Motowidlo, S.J., 1997. Task performance and contextual performance: the meaning for personnel selection research. Human Performance 10, 99–109.

Brown, D.J., Ferris, D.L., Heller, D., Keeping, L.M., 2007. Antecedents and consequences of the frequency of upward and downward social comparisons at work. Organizational Behavior and Human Decision Processes 102, 59–75.

Chang, C.-H., Ferris, D.L., Johnson, R.E., Rosen, C.C., Tan, J.A., 2012. Core self-evaluations: a review and evaluation of the literature. Journal of Management 38, 81–128.

Chen, G., Gully, S.M., Eden, D., 2004. General self-efficacy and self-esteem: toward theoretical and empirical distinction between correlated self-evaluations. Journal of Organizational Behavior 25, 375–395.

Crampton, S.M., Wagner III, J.A., 1994. Percept-percept inflation in micro-organizational research: an investigation of prevalence and effect. Journal of Applied Psychology 79, 67–76.

Elliot, A.J., Thrash, T.M., 2002. Approach-avoidance motivation in personality: approach and avoidance temperaments and goals. Journal of Personality and Social Psychology 82, 804–818.

Erez, A., Judge, T.A., 2001. Relationship of core self-evaluations to goal setting, motivation, and performance. Journal of Applied Psychology 86, 1270–1279.

Eysenck, H.J., 1967. The Biological Basis of Personality. Thomas, Springfield, IL.

Ferris, D.L., Johnson, R.E., Rosen, C.C., Djurdjevic, E., Chang, C.-H., Tan, J.A., 2013. When is success not satisfying? A moderated mediation model of the relation between core self-evaluation and job satisfaction. Journal of Applied Psychology 98, 342–353.

Ferris, D.L., Rosen, C.C., Johnson, R.E., Brown, D.J., Risavy, S., Heller, D., 2011. Approach or avoidance (or both?): integrating core self-evaluations within an approach/avoidance framework. Personnel Psychology 64, 137–161.

Gillham, J.E., Shatté, A.J., Reivich, K.J., Seligman, M.E.P., 2001. Optimism, pessimism, and explanatory style. In: Chang, E.C. (Ed.), Optimism & Pessimism: Implications for Theory, Research, and Practice. APA, Washington, DC, pp. 53–75.

Goldberg, L.R., 1993. The structure of phenotypic personality traits. American Psychologist 48, 26–34.

Harris, K.J., Harvey, P., Kacmar, K.M., 2009. Do social stressors impact everyone equally? An examination of the moderating impact of core self-evaluations. Journal of Business and Psychology 24, 153–164.

Hiller, N.J., Hambrick, D.C., 2005. Conceptualizing executive hubris: the role of (hyper-) core self-evaluations in strategic decision making. Strategic Management Journal 26, 297–319.

Johnson, R.E., Chang, C.-H., Meyer, T., Lanaj, K., Way, J., 2013. Approaching success or avoiding failure? Approach and avoidance motives in the work domain. European Journal of Personality 27, 424–441.

Johnson, R.E., Rosen, C.C., Chang, C.-H., Djurdjevic, E., Taing, M.U., 2012. Recommendations for improving the construct clarity of higher-order multidimensional constructs. Human Resource Management Review 22, 62–72.

Johnson, R.E., Rosen, C.C., Djurdjevic, E., 2011. Assessing the impact of common method variance on higher-order multidimensional constructs. Journal of Applied Psychology 96, 744–761.

Johnson, R.E., Rosen, C.C., Levy, P.E., 2008. Getting to the core of core self-evaluations: a review and recommendations. Journal of Organizational Behavior 29, 391–413.

Judge, T.A., Bono, J.E., 2001. Relationship of core self-evaluations traits – self esteem, generalized self-efficacy, locus of control, and emotional stability – with job satisfaction and job performance: a meta-analysis. Journal of Applied Psychology 86, 80–92.

Judge, T.A., Erez, A., Bono, J.E., 1998a. The power of being positive: the relationship between positive self-concept and job performance. Human Performance 11, 167–187.

Judge, T.A., Locke, E.A., Durham, C.C., Kluger, A.N., 1998b. Dispositional effects on job and life satisfaction: the role of core evaluations. Journal of Applied Psychology 83, 17–34.

Judge, T.A., Erez, A., Bono, J.E., Thoresen, C.J., 2002. Do the traits self-esteem, neuroticism, locus of control, and generalized self-efficacy indicate a common core construct? Journal of Personality and Social Psychology 83, 693–710.

Judge, T.A., Erez, A., Bono, J.E., Thoresen, C.J., 2003. The core self-evaluations scale: development of a measure. Personnel Psychology 56, 303–331.

Judge, T.A., Locke, E.A., Durham, C.C., 1997. The dispositional causes of job satisfaction: a core evaluations approach. Research in Organizational Behavior 19, 151–188.

Kacmar, K.M., Collins, B.J., Harris, K.J., Judge, T.A., 2009. Core self-evaluations and job performance: the role of the perceived work environment. Journal of Applied Psychology 94, 1572–1580.

Kammeyer-Meuller, J.D., Judge, T.A., Scott, B.A., 2009. The role of core self-evaluations in the coping process. Journal of Applied Psychology 94, 177–195.

Lanaj, K., Chang, C.-H., Johnson, R.E., 2012. Regulatory focus and work-related outcomes: a meta-analysis. Psychological Bulletin 138, 998–1034.

Locke, E.A., 1976. The nature and causes of job satisfaction. In: Dunnette, M.D. (Ed.), Handbook of Industrial and Organizational Psychology, first ed. Rand McNally, Chicago, IL, pp. 1297–1349.

Meyer, J.P., Allen, J.J., 1997. Commitment in the Workplace: Theory, Research, and Application. Sage, Thousand Oaks, CA.

Ng, T.W.H., Feldman, D.C., 2010. Idiosyncratic deals and organizational commitment. Journal of Vocational Behavior 76, 419–427.

Pierce, J.L., Gardner, D.G., 2004. Self-esteem within the work and organizational context: a review of the organization-based self-esteem literature. Journal of Management 30, 591–622.

Pierce, J.L., Gardner, D.G., 2009. Relationships of personality and job characteristics with organization-based self-esteem. Journal of Managerial Psychology 24, 392–409.

Rich, B.L., LePine, J.A., Crawford, E.R., 2010. Job engagement: antecedents and effects on job performance. Academy of Management Journal 53, 617–635.

Rosenberg, M., 1965. Society and the Adolescent Self-image. Princeton University Press, Princeton, NJ.

Rotter, J.B., 1966. Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs 80. Whole No. 609.

Spector, P.E., 1982. Behavior in organizations as a function of employees' locus of control. Psychological Bulletin 91, 482–497.

Spector, P.E., 1997. The role of frustration in antisocial behavior at work. In: Giacalone, R.A., Greenberg, J. (Eds.), Antisocial Behavior in Organizations. Sage, Thousand Oaks, CA, pp. 1–17.

Core–Periphery Patterns in Networks

Martin G Everett, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A network has a core–periphery structure if it consists of two groups of actors, core actors who interact mainly with other core actors and peripheral actors who connect to the core but do not interact among themselves. We examine techniques designed to uncover core–periphery structures in networks. There are two distinct approaches, the first is to try and find a bipartition of the network, which optimizes a fit function designed to capture the ideal structure. The second is to derive a coreness score for each actor that captures the extent to which they are a core or a peripheral player.

Introduction

Core–periphery patterns in the context of social networks were initially discussed without any recourse to a formal definition. They were informally discussed in a variety of studies such as national elites and collective action (Laumann and Pappi, 1976; Alba and Moore, 1978), citation networks (Mullins et al., 1977; Doreian, 1985), interlocking directorates (Mintz and Schwartz, 1981), and primates (Corradino, 1990). All of these informal ideas shared the intuition of dividing the nodes into two groups namely the core and the periphery. The concept was examined in more detail by Borgatti and Everett (1999) and although they suggested their paper should be the starting point of a methodological debate on what constitutes a core–periphery their formulation has come to be the accepted standard. The simplest way to envisage their definition is in terms of a block model. Given a network with an adjacency matrix in which the rows and columns have been arranged so that the core nodes are followed by the peripheral nodes then the core–periphery partition gives rise to four blocks as depicted in **Figure 1**. The core block contains the core to core interactions, the peripheral block contains the periphery to periphery interactions with the two off-diagonal blocks containing the core to periphery and the periphery to core interactions. In an undirected network the two off-diagonal blocks are simply the transpose of each other.

In a core–periphery structure, we expect core nodes to be well connected to other core nodes and we would also expect peripheral nodes to not be connected to other peripheral nodes. Given this the perfect core–periphery structure for a binary network would be a one block as core, that is a block consisting entirely of ones (except possibly the diagonal) and a zero block as peripheral, that is a block consisting entirely of zeros. Here we are using the terminology of block models where the one and zero blocks are the only permissible building blocks of structural equivalence. Suppose that our network is undirected, if we maintain structural equivalence as our underlying model then the core–periphery block can only be a zero block or a one block. If it were a zero block then the ideal structure would have a clique as a core with isolates as a periphery. It is true that some researchers are quite content to have isolates in the periphery but the majority feel that this does not really capture the core–periphery concept, since we expect the periphery to be related in some way to the core. That is the nodes are peripheral to a defined core and are not simply just peripheral, this provides a distinction between a peripheral node and an isolated node. If we retain the structural equivalence model then the core–periphery blocks must be one block. This implies the rather strict condition that every peripheral member is connected to every core member. The network shown in **Figure 2** has this structure with nodes 1, 2, 3, and 4 in the core and the remainder in the periphery.

	Core nodes	Peripheral nodes
Core nodes	Core	Core–periphery
Peripheral nodes	Periphery–core	Peripheral

Figure 1 Core–periphery block model structure.

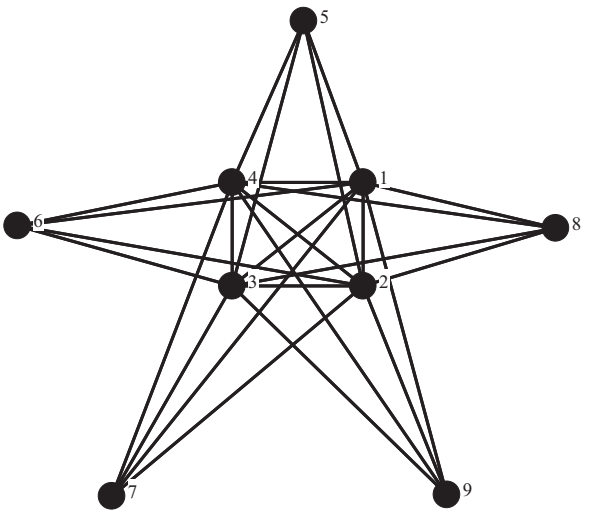

Figure 2 A simple core–periphery structure.

Clearly such idealized patterns are not likely in real data and so we usually look to relax the conditions. One approach is to only consider the core and periphery blocks and in essence ignore the off-diagonal blocks. In so doing, it would make sense to at least insist on some core–periphery interaction, and this can be achieved by making sure that the network is connected if it is undirected or weakly connected if it is directed. In examining relaxations from the ideal Borgatti and Everett proposed two models, the discrete model and the continuous model, which we shall now consider these in more detail.

Discrete Model

A simple description of the discrete approach is to assign each node to either the core or the periphery so that the resultant matrix is as close as possible to the block model described in **Figure 1**, which we shall now refer to as the structure matrix. The first issue is to find a measure that captures how close two matrices are to each other. There are a number of possibilities but two have been extensively used in the development of methods and in applications. The first is to use Pearson correlation that is to correlate the structure matrix with the data matrix. In this case we use a modified structure matrix and treat the core–periphery and periphery–core blocks as missing values. The second method is to simply count the number of errors, that is, the number of times the data matrix has a zero when the structure matrix has a one and the number of times the data matrix has a one when the structure matrix has a zero, sometimes referred to as the Hamming distance. Again we do this just for the core and the peripheral blocks and we ignore the two off-diagonal blocks. Both measures try to assess the extent to which the core is well connected and the periphery is sparsely connected. Once we have a measure of how good a partition is we can try and optimize this over all possible partitions. It is simply not possible except for very small networks to look at every possible assignment into core and periphery and so we need to use combinatorial optimization techniques or other heuristic methods. There are a number of possible methods designed for solving such problems and a comparison of the techniques was undertaken by Boyd et al. (2006). They found some subtle differences on the quality of solutions but with any major differences probably attributable to different implementations rather than any fundamental advantages. This is in keeping with most findings in this area. They did find that a Kernighan–Lin type algorithm developed by them was the fastest but this is an area of continuous development and most users would not need to concern themselves with the underlying solution method.

As an example of this approach, we take the citation data among the top social work journals in a 1-year period (1985–86) reported by Baker (1992). These data record if one journal cites another journal during the period. The results of the core–periphery using the correlation method in UCINET (Borgatti et al., 2002) are given in **Figure 3**.

We note that the journals do cite articles within themselves and this produces the ones down the diagonal but the algorithm ignores these. The correlation between the data matrix given in **Figure 3** and the idealized block structure is 0.81, which indicates a strong core–periphery structure. The core block has a density of 0.86 and the peripheral block has a density of 0.05 indicating even more strongly the core–periphery structure of this data in which core journals cite other core journals and peripheral journals do not cite other peripheral journals. This is further reinforced by examining the two off-diagonal blocks. The top right core–periphery block has a density of 0.15, whereas the bottom left periphery to core block has a density of twice that at 0.32. This shows that

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
		C	S	J	S	S	S	A	C	F	C	A	J	B	P	C	J	C	S	S	I
1	CW	1	1	1	1	1			1						1			1			
2	SW	1	1	1	1	1	1	1							1				1	1	
3	JSWE		1	1	1	1	1	1												1	
4	SSR	1	1	1	1	1	1	1													
5	SCW	1	1	1	1	1	1	1	1	1	1						1		1	1	
6	SWRA	1	1	1	1	1	1														
7	ASW		1	1	1		1	1							1						
8	CAN	1	1			1			1												
9	FR		1			1				1											
10	CSWJ		1		1	1					1										
11	AMH		1									1									
12	JSP				1								1								
13	BJSW		1	1										1							
14	PW	1							1						1						
15	CCQ	1				1										1					
16	JGSW		1			1											1				
17	CYSR	1	1		1	1	1		1	1					1	1		1			
18	SWG		1	1	1	1								1					1	1	
19	SWHC		1			1														1	
20	IJSW		1																		1

Figure 3 Discrete core–periphery analysis of citation data.

peripheral journals cite twice as many core journals as core journals cite peripheral journals providing further evidence of the hierarchical structure of the journals.

As already discussed, an alternative measure to correlation is to use the Hamming distance. This was suggested by Brusco (2011) although it had already been implemented in UCINET many years earlier using the same genetic algorithm as is used for the correlation method. Brusco proposed an exact algorithm based on branch and bound techniques using this measure. As it happens, if we use the Hamming distance then the problem can be solved very efficiently in polynomial time. This result is due to a theorem by Hammer and Simeone (1981) that implies for undirected data that all we need to do to find the best fit is to order the vertices in decreasing order of degree. We then just need to find a threshold value for the degree and every node with degree above the threshold will be a core node and the rest will be peripheral and this can be done efficiently. This result requires a trivial modification to deal with directed data.

This technique is straightforward and would seem to be the best approach. If we apply this method to the citation data given in **Figure 3**, for example, we obtain exactly the same results. But this approach does have a few problems. First it cannot be directly applied to valued data whereas the correlation method can. Second, and more importantly, it does not give information about the core–periphery structure of the network. It is quite possible for the best assignment of nodes to a core and periphery structure measured in this way to result in a core block that is a zero block or has very few entries. To see this, consider the following example by Brandes (personal communication). Start with a path of 11 nodes numbered in sequence so that 1 and 11 are the end nodes. To each even numbered node attach 9 pendants so that each even node has degree 11. The result will be a graph with 56 nodes, 5 with degree 11, 4 with degree 2, and 47 with degree 1. Consider the partition in which the five nodes with degree 11 are in the core and the rest are in the periphery. Since the only connections are between core and periphery both the core block and the peripheral block are zero. As a consequence the Hamming distance will be 20. Moving a core node into the periphery will create a Hamming distance of 22, moving a degree 2 vertex into the core will create a Hamming distance of 26. The best solution is therefore our original split in case we are left with a core, which is a zero block.

This example does not mean the Hamming distance should not be used but it should be used with care and only be used when computationally the correlation method is not feasible. When it is used the user needs to make sure that they have a structure, which intuitively fits the core–periphery criteria.

Continuous Model

One further issue with the discrete model is that it places nodes either in the core or in the periphery and this may be too simplistic. Some authors have suggested the idea of a semi-periphery, which is equivalent to introducing more blocks into the block structure. The problem with this is that it increases the complexity of the problem and it becomes even more difficult to decide on the nature and structure of the additional blocks. An alternative formulation is to try and find a node level measure of the amount of core–periphery. We have already seen that degree centrality does precisely this for the Hamming distance. High degree actors must be in the core, low degree actors are in the periphery, and actors who are close to the boundary value would be in the semiperiphery or on the edge of the core. But as we have seen degree is not always a good measure and does not extend to valued data. To overcome this, Borgatti and Everett suggest using ideal structure matrix Δ constructed as the outer-product of a vector $\mathbf{c}$ of coreness values as shown in eqn [1].

$$\Delta = \mathbf{c}\mathbf{c}^{\mathrm{T}} \qquad [1]$$

In this equation the i, jth entry of the matrix Δ is simply the product of $\mathbf{c}_i$ with $\mathbf{c}_j$. If two nodes have a high coreness score then their product and hence the corresponding value in Δ will be high. If both entries are low then the product will be low and so the corresponding value in Δ will be low. If one is high and one is low the corresponding Δ value will be intermediate. In the extreme case in which we constrain the entries of $\mathbf{c}$ to be either 0 or 1, then Δ would correspond exactly to a block model structure of core–periphery as in the discrete case and $\mathbf{c}$ would be an indicator vector where a value 1 would indicate a core node and 0 a peripheral node. It follows that our discrete method can be formulated as finding a binary vector $\mathbf{c}$ in which the correlation between our data matrix A and our product matrix Δ is as large as possible. If we now relax the condition that $\mathbf{c}$ is binary and allow $\mathbf{c}$ to take on any positive values we obtain the continuous core–periphery model. The values of $\mathbf{c}$ now give the coreness of each node within the network. This generalization from the discrete to the continuous is remarkably similar to the extension of degree centrality to eigenvector centrality. As already mentioned, the degree centrality can be thought of as a coreness measure when we use the number of changes measure of fit for core–periphery and eigenvector serves a similar purpose for the continuous approach. (In fact if we use Euclidean distance between matrices instead of correlation then this similarity becomes explicit providing we include the diagonals.) Applying the continuous method to the citation data used for the discrete method in **Figure 3** yields the results given in **Figure 4**.

The two columns are the split between core and periphery recommended by the software. This recommendation is based on correlating the scores above with an ideal set of scores where core members score a value of one and peripheral members score a value of zero. This is then examined for all possible core sizes from 1 to $n - 1$ and the recommended split is the highest correlation from these $n - 1$ values. In this case a vector of eight ones is followed by 12 zeros. This measure is called the concentration. If we compare this output to the one in **Figure 3**, we see that they have six journals in common. We note that ASW is the highest scoring peripheral journal and that the discrete method has this journal in the core. Two journals that were peripheral for the discrete method, namely SWG and CYSR, are placed in the core. Looking at the discrete output we see that these are the only journals with multiple peripheral block entries. There is also one further important characteristic of the continuous method, the matrix Δ must be by definition symmetric. It follows that the continuous method, which must take account of the off-diagonal blocks, cannot differentiate between core–periphery and periphery–core interaction. As

5	SCW	0.473	7	ASW	0.180
2	SW	0.358	8	CAN	0.167
17	CYSR	0.349	10	CSWJ	0.164
1	CW	0.308	9	FR	0.120
4	SSR	0.284	16	JGSW	0.120
6	SWRA	0.256	19	SWHC	0.120
3	JSWE	0.255	15	CCQ	0.113
18	SWG	0.239	13	BJSW	0.088
			14	PW	0.068
			11	AMH	0.051
			20	IJSW	0.051
			12	JSP	0.041

Figure 4 Continuous core–periphery scores for citation data.

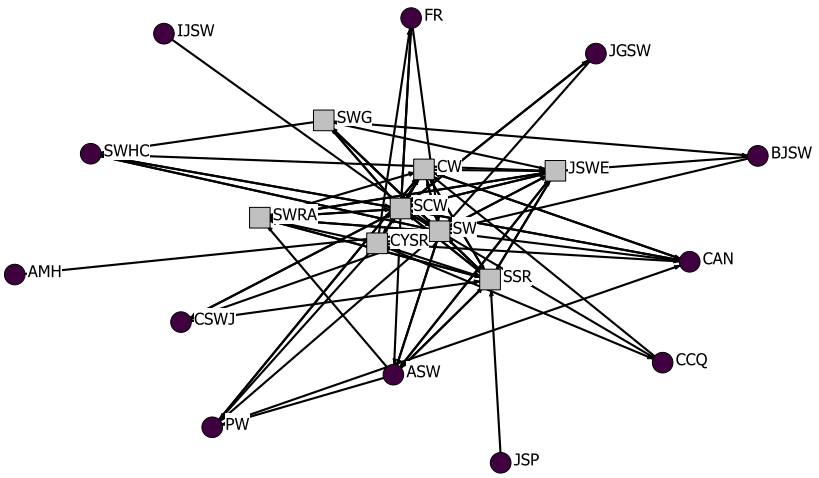

Figure 5 Core–periphery layout of citation data.

a consequence it is far better suited to symmetric data (or at least data that is very nearly symmetric). In fact if the two methods are compared on symmetrized versions of the citation data they are in complete agreement.

The above example uses correlation but we could use the sums of square differences as our measure of association. If we include the diagonal as zeros then this is the same as finding the principal eigenvector. If we ignore the diagonals then this is the same as the MINRES (MINimal RESiduals) procedure developed by Harman (1967) and Comrey (1962) in a different context. Boyd et al. (2010) have extended this approach and suggest using an MINRES Singular Valued Decomposition (SVD). This has the advantage of being able to deal with asymmetric data and at the same time introducing meaningful (imputed) values on the diagonal. In this case they produce both in and out coreness scores, which can be used to define cores and peripheries.

The continuous model has one other advantage over the discrete model in that we can use the coreness values to order the nodes to obtain good visualizations. If we submit Δ to a metric mds then use this output to position the nodes we will obtain a network in which the nodes in the core will be placed at the center and the nodes in the periphery will be around the outside. We can then immediately see which nodes are on the edge of the core or part of the semiperiphery particularly if we shape or color the nodes that are in the recommended core. We have done this for the citation data and the results are shown in Figure 5 with core nodes as gray squares and peripheral nodes as black circles.

The position of the core journals in the center and the strong connections between them are apparent. In addition, we see that ASW is placed near to the core and so must be close to being considered a member of the core, in addition SWG, JSWE, and SWRA are clearly on the edge of the core and are close to being in the periphery.

Boyd et al. use the **c** values to order the nodes in an adjacency matrix and use a density plot to display the values in the data.

Extensions

One issue we have not considered is trying to assess statistically whether the data actually has a core–periphery structure. Boyd et al. (2006) describe a generic permutation test that can be applied to the discrete case. Their approach is computational intensive since they propose permuting the data and a running a discrete core–periphery optimization routine after each

permutation. This is repeated thousands of times to obtain a distribution of the fitness function. It was for this reason that they were keen to develop a heuristic that was accurate and fast. Rather than permuting the rows and columns they propose a permutation that preserves the cell distribution, which also preserves the total number of ties. This means that row and column marginals are not preserved. They also propose a similar approach to assess the result of their MINRES/SVD approach (Boyd et al., 2010). Again it is necessary to run the continuous optimization routine after each permutation. In this instance, they propose permutations that preserve both transpose pairs and the ordering implied by **c**.

In the previous section, we discussed a number of possible measures for assessing the similarity of the data to ideal structure matrices. It is quite possible to substitute any number of different functions such as chi-square measure or a log-likelihood. Muñiz and Carvajal (2006) suggest using an entropy type measure.

So far we have only considered single mode data. The ideal structure matrix is simply a nonsquare version of the structure matrix for one mode data. The only difference is that the blocks themselves are not square and so the partitioning of the columns is independent of the partitioning of the rows. Extending the discrete model is therefore straightforward but the fact that the rows and columns are partitioned independently means that this is computationally much more difficult. Everett and Borgatti (2013) suggest that it is better to apply the discrete method using correlation (or any other method that works on valued data) to the two projections AA^T and A^TA and then use these results to partition the rows and columns of A. They refer to this as the dual projection method. **Figure 6** gives the results of doing this using the discrete correlation method in UCINET on the Southern Women data (Davis et al., 1941).

We see in these data that the core group of women attend most events and that the core events are attended by most women but that the peripheral events are sparsely attended by the peripheral women.

The same approach can be used for the continuous method. That is we can apply the continuous method to the two projections and so have two sets of coreness scores one for each mode. If we used minimizing the sums of squares as our measure (including the diagonals), then the coreness scores would be the same as the **u** and **v** scores of a singular valued decomposition. For this reason it would be just as appropriate to use the MINRES/SVD approach discussed above on two mode data as these gives two separate coreness scores.

Software, Applications, and Scalability

The only software package that contains both the discrete and continuous methods is UCINET (Borgatti et al., 2002). The discrete method uses a generic algorithm and not the Kernighan and Lin type algorithm. At the time of writing UCINET does not include the permutation tests or the MINRES/SVD extensions suggested by Boyd. Boyd uses mathematica (Wolfram, 2003) in both his extensions and gives the commands required for those familiar with the software. Pajek (Batagelj and Mrvar, 1998) has the facility to have prespecified block models and hence the core–periphery structures can be specified and this provides an alternative algorithm for the discrete model. Since eigenvectors or singular vectors can be viewed as coreness type measures (although the fact they include the diagonal could be seen as an issue), these can be used and are available in a large variety of mathematical and statistical software packages.

Examples of applications can be seen in a variety of disciplines and the benefits and disadvantages of core–periphery structures have been extensively examined. In addition to the earlier work already mentioned Cattani and Ferriani (2008)

		8 E	9 E	6 E	7 E	5 E	3 E	4 E	1 E	2 E	10 E	11 E	12 E	13 E	14 E
1	EVELYN	1	1	1		1	1	1	1	1					
2	LAURA	1		1	1	1	1		1	1					
3	THERESA	1	1	1	1	1	1	1		1					
4	BRENDA	1		1	1	1	1	1	1						
14	NORA		1	1	1						1	1	1	1	1
13	SYLVIA	1	1		1						1		1	1	1
6	FRANCES	1		1		1	1								
8	PEARL	1	1	1											
7	ELEANOR	1		1	1	1									
10	VERNE	1	1		1								1		
11	MYRNA	1	1								1		1		
12	KATHERINE	1	1								1		1	1	1
9	RUTH	1	1		1	1									
5	CHARLOTTE				1	1	1	1							
15	HELEN	1			1						1	1	1		
16	DOROTHY	1	1												
17	OLIVIA		1									1			
18	FLORA		1									1			

Figure 6 Discrete core–periphery of the Southern Women data using the dual projection method.

look at the core–periphery structures in the Film Industry, Bodin and Crona (2009) examine natural resource governance and Crowston et al. (2006) analyze open source software team communication. These are just a few examples chosen from an increasing and diverse literature on the subject.

The fact that we are now able to collect and analyze very large data sets presents issues of scalability. Most of the work on core–periphery structures is premised by the idea that the network is a social network with a cost to maintaining links. It is also assumed that the core is such that it is well integrated and two core actors are at most one or two steps apart. If we wish to consider networks such as the World Wide Web or even those generated by social networking sites such as Facebook, then this concept would need to be reconsidered. One approach is to look for multiple core–periphery structures; this was first suggested by Everett and Borgatti (1999) and has been developed in the context of larger networks by a number of authors see Holme (2005), for example. The basic idea is to define the cores in terms of cohesive subgraphs and then trace the peripheries associated with the cores. The use of k-cores (a group of vertices in which every vertex is connected to k or more other members of the group), which can be quickly and easily detected in large complex networks are a useful starting point.

See also: Clustering and Cohesion in Networks: Concepts and Measures; Network Centrality, Measures of; Social Network Algorithms and Software; Structural Equivalence: Meaning and Measures.

Bibliography

Alba, R.D., Moore, G., 1978. Elite social circles. Sociological Methods and Research 7, 167–188.

Baker, D.R., 1992. A structural analysis of the social work journal network: 1985–1986. Journal of Social Service Research 15, 153–168.

Batagelj, V., Mrvar, A., 1998. Pajek – program for large network analysis. Connections 21 (2), 47–57. http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

Bodin, Ö., Crona, B.I., 2009. The role of social networks in natural resource governance: what relational patterns make a difference? Global Environmental Change 19, 366–374.

Borgatti, S.P., Everett, M.G., 1999. Models of core/periphery structures. Social Networks 21, 375–395.

Borgatti, S.P., Everett, M.G., Freeman, L.C., 2002. Ucinet for Windows: Software for Social Network Analysis. Analytic Technologies, Harvard, MA.

Boyd, J.P., Fitzgerald, W.J., Beck, R.J., 2006. Computing core/periphery structures and permutation tests for social relations data. Social Networks 28, 165–178.

Boyd, J.P., Fitzgerald, W.J., Mahutga, M.C., Smith, D.A., 2010. Computing continuous core/periphery structures for social relations data with MINRES/SVD. Social Networks 32, 125–137.

Brusco, M.J., 2011. An exact algorithm for a core/periphery bipartitioning problem. Social Networks 33, 12–19.

Cattani, G., Ferriani, S., 2008. A core/periphery perspective on individual creative performance: social networks and cinematic achievements in the hollywood film Industry. Organization Science 19, 824–844.

Comrey, A.L., 1962. The minimum residual method of factor analysis. Psychological Reports 11, 15–18.

Corradino, C., 1990. Proximity structure in a captive colony of Japanese monkeys *(Macaca fuscata fuscata)*: an application of multidimensional scaling. Primates 31 (3), 351–362.

Crowston, K., Kangning, W., Qing, L., Howison, J., 2006. Core and periphery in free/libre and open source software team communications. In: Proceedings of the 39th Hawaii International Conference on System Sciences – 2006, vol. 39. IEEE, New York, pp. 118–124.

Doreian, P., 1985. Structural equivalence in a psychology journal network. American Society for Information Science 36 (6), 411–417.

Davis, A., Gardner, B.B., Gardner, M.R., 1941. Deep South. University of Chicago Press, Chicago.

Everett, M.G., Borgatti, S.P., 1999. Peripheries of cohesive subsets. Social Networks 21, 396–407.

Everett, M.G., Borgatti, S.P., 2013. The dual projection approach for two-mode networks. Social Networks 35, 204–210.

Harman, H., 1967. Modern Factor Analysis, second ed. University of Chicago Press, Chicago.

Hammer, P.L., Simeone, B., 1981. The splittance of a graph. Combinatorica 1 (3), 275–284.

Laumann, E.O., Pappi, F.U., 1976. Networks of Collective Action: A Perspective on Community Influence Systems. Academic Press, New York.

Mintz, B., Schwartz, M., 1981. Interlocking directorates and interest group formation. American Sociological Review 46, 851–868.

Mullins, N.C., Hargens, L.L., Hecht, P.K., Kick, E.L., 1977. The group structure of cocitation clusters: a comparative study. American Sociological Review 42, 552–562.

Muñiz, A.S.G., Carvajal, C.R., 2006. Core/periphery structure models: an alternative methodological proposal. Social Networks 28, 442–448.

Wolfram, S., 2003. The Mathematica Book. Wolfram Media/Cambridge University Press, Cambridge.

Coronary Heart Disease: Psychosocial Aspects

Andrew Steptoe, University College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Davey Smith and A. Ness, volume 4, pp. 2773–2782, © 2001, Elsevier Ltd.

Abstract

Coronary heart disease (CHD) is a major cause of death and disability in the modern world. It is a multifactorial condition, and there is growing evidence that psychosocial factors are relevant. The factors involved include exposure to adversities such as chronic stress, impoverished social relationships leading to isolation and loneliness, and emotional conditions including depression and hostility. Psychosocial factors are relevant at several stages of the evolution of CHD, from the development of coronary atherosclerosis, through triggering of acute cardiac events, to the experience of patients who survive a myocardial infarction. The article outlines the evidence for the importance of psychosocial factors in CHD, highlighting methodological issues in this field of research, and outlines implications for prevention and clinical management.

Coronary Heart Disease across the World

Coronary heart disease (CHD) is a leading cause of death and disability throughout the world. Statistics from the American Heart Association indicate that in 2010, CHD caused around one in six deaths in the USA, with more than 600 000 having a heart attack each year. Across Europe, CHD accounts for 20% of deaths among men and 22% among women, many of which occur in people less than 75 years old. Fortunately, there have been substantial reductions in deaths from CHD in recent decades. The death rates from CHD in the UK have declined by more than 70% since 1968, and over the last 15 years alone, CHD in high income Western countries has fallen by around a third. This pattern is due in roughly equal measure to two processes. First, changes in risk factors such as smoking and better treatment for high cholesterol and high blood pressure (hypertension) mean that fewer people are developing advanced CHD. Second, treatments of people with advanced disease or those who have suffered a heart attack (myocardial infarction) have improved, so survival is greater. One consequence is that there is a growing population of older people living with CHD. In the UK, there are thought to be over 1.6 million men and 1 million women with CHD, while estimates in the USA total more than 15 million. It should also be said that the increase in obesity and diabetes over recent decades may reverse the positive trends in risk factors.

The underlying cause of CHD is coronary atherosclerosis, a progressive condition involving the infiltration of cholesterol, inflammatory products, and smooth muscle cells into the walls of the coronary arteries, leading to the development of coronary plaque. This process starts early in life, but typically comes to clinical attention at advanced stages in the form of angina pectoris (chest pain on exertion), an acute myocardial infarction (MI), or sudden cardiac death (a death resulting from abrupt loss of heart function occurring within a short time of symptom onset). The long-term nature of CHD is important from the psychosocial perspective, since different factors may contribute to different stages.

There is no single cause of coronary atherosclerosis and CHD, so the focus is on risk factors that increase the likelihood of developing the condition. These can be divided into three categories: *nonmodifiable factors* such as age, gender, ethnicity, and heredity; *treatable conditions* such as hypertension, high cholesterol, and diabetes; and *modifiable risk factors*. The latter are primarily behavioral, and include smoking, physical inactivity, obesity, dietary choice, and excessive alcohol intake. Psychosocial factors fall into this third group, since they are potentially modifiable aspects of people's experience.

Psychosocial Factors

There is a tradition stretching back more than 100 years of speculation about the role of stress and personality in CHD. More recently, the concept of psychosocial factors has emerged to describe the range of psychological characteristics and social experiences that may be relevant to CHD. Table 1 summarizes the factors that have been investigated in most detail. They are divided into two categories: risk factors that increase the likelihood of developing CHD, and protective factors that may reduce risk. Within each category, there are three types of factors: first, exposures to various types of potentially stressful experience such as low socioeconomic status (SES), chronic work stress, and other types of life stress; second, experiences of the social environment that might be either negative (social isolation) or protective (social support); third, psychological characteristics and personality traits that may again be negative (e.g., depression) or positive (e.g., positive well-being). The evidence relating to these

Table 1 Psychosocial factors in coronary heart disease

Risk factors	*Protective factors*
• Low SES	• Social connectedness
• Work stress	• Social support
• Life events and chronic stress	• Positive well-being
• Social isolation	• Optimism
• Loneliness	
• Depression	
• Anger and hostility	

different aspects varies in strength and quality (Steptoe and Kivimaki, 2012, 2013).

Methods of Investigation

Several different methods have been used to investigate the role of psychosocial factors in CHD. First, animal studies can evaluate the impact of different social conditions on CHD in a highly controlled fashion, involving experimental designs that permit definite conclusions about causality to be drawn. For example, an influential set of studies of cynomolgus macaques by Kaplan and Manuck (1999) established the role of social stress in the development of CHD. Animal studies are also central to the investigation of behavioral and pharmacological interventions that may ameliorate the impact of psychosocial adversity. However, they do not in themselves prove that similar processes operate in humans, and the subtlety of human experience is difficult to capture in animal behavioral studies.

A second method of investigation is to carry out psychophysiological or mental stress testing with human participants. This technique involves monitoring biological processes while people are exposed to different conditions (*see* Cardiovascular Reactivity and Recovery). The method allows the impact of psychological and social stimuli on the biological processes involved in CHD to be examined under controlled conditions, and provides important mechanistic data about the biological pathways through which psychosocial factors increase cardiovascular risk. Patients and healthy controls or people at higher and lower risk can be compared, as can conditions relevant to stress processes such as controllability and social support.

The conditions in which psychophysiological stress testing is carried out are artificial, and the stimuli used (such as difficult cognitive tasks) are seldom encountered in real life. There has therefore been a growth in naturalistic or ambulatory monitoring in which biological processes relevant to CHD are measured in everyday life. Ambulatory blood pressure, heart rate, and heart rate variability are commonly monitored in these studies, as is salivary cortisol. The paradigm allows the impact of psychosocial stimuli on biology to be assessed in everyday conditions where ecological validity is high. Measures are limited to biomarkers that can be repeatedly assessed, and sophisticated statistical modeling is needed to take account of covariates such as time of day, smoking, concurrent physical activity, and other potential confounders.

These methods are primarily focused on mechanisms linking psychosocial factors with CHD, rather than the factors themselves. Cross-sectional comparisons between people with CHD and healthy controls are difficult to interpret since the presence of illness may change people's appraisals of their past experience. Longitudinal observational population studies are therefore needed. These studies have the following features: first, the recruitment of a large, preferably representative, population sample that is screened to ensure that participants do not already suffer from CHD; second, the measurement of exposure to the psychosocial factors being tested (e.g., chronic stress, low social support), along with other factors known to influence the outcome; third, the monitoring of the population over time, measuring the development of CHD and other health outcomes; finally, multivariate analysis to test whether exposure to the psychosocial factor is associated with CHD after the covariates have been controlled.

Such studies inevitably take many years to complete, and since assessments have to be made on hundreds or even thousands of participants, the detail in measures is often limited. Investigations of this kind are correlational rather than experimental, so cannot definitively establish causality; there may be unmeasured factors that influence both exposure to the psychosocial factor and outcome. In the absence of large trials where people are randomized to high and low risk conditions for many years, the longitudinal observational study provides the strongest data about the role of psychosocial factors. However, a true understanding of the role of psychosocial factors in CHD requires the convergence of evidence from all these methods of investigation.

Acute versus Chronic Psychosocial Factors

As noted earlier, different psychosocial factors may be relevant at different stages of CHD, as shown in Figure 1. The long-term progression of coronary atherosclerosis and development of CHD may involve exposure over many years to chronic psychosocial adversity (pathways 1 and 2). The longitudinal observational study is most appropriate here. On the other hand, triggering of acute cardiac events in people with advanced CHD may involve acute stressors (pathway 3), and require a different methodology. If people have survived a myocardial infarction or another acute

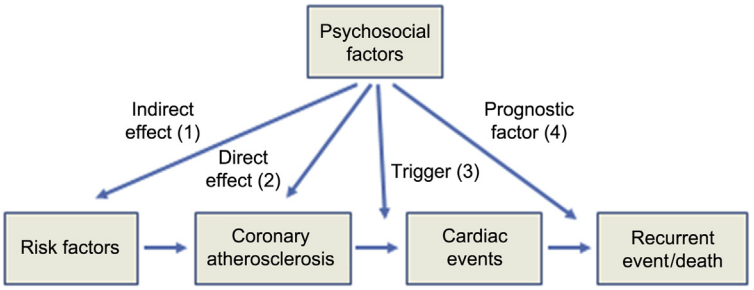

Figure 1 Psychosocial factors and the development of CHD. After Steptoe, A., Kivimaki, M., 2012. Stress and cardiovascular disease. Nature Reviews Cardiology 9, 360–370.

coronary event, they are at increased risk for future recurrent events. This provides a further stage for psychosocial factors, potentially affecting whether or not the person has another acute cardiac event (pathway 4). This article summarizes the impact of psychosocial factors at these three stages.

Chronic Psychosocial Adversity and CHD Risk

Chronic adversity takes many forms, as shown in **Table 1**. Perhaps the broadest concept relevant to CHD is SES defined by income, occupation, education, or wealth. There is a marked SES gradient in CHD in high income countries, with higher rates of CHD incidence and mortality in lower SES groups (*see* Socioeconomic Status and Health). The pattern is stronger in some countries than others, due in part to the social patterning of risk factors such as smoking and excessive alcohol consumption (Mackenbach et al., 2008).

Work Stress

The relationship between work stress and CHD has been studied extensively, and numerous large-scale prospective observational studies have been completed. The most common method of conceptualizing work stress is the 'job strain' or demand/control model, in which high work stress is thought to arise when high demands at work are coupled with low control over how the work is done, and how much flexibility the worker has. A recent *meta*-analysis of more than 15 prospective studies showed an age and sex-adjusted relative risk for job strain of 1.34 (95% confidence intervals (CI) 1.18–1.51); that is, people experiencing job strain have a 34% increase in risk of developing CHD than those without job strain (Steptoe and Kivimaki, 2013). Somewhat similar estimates emerge from studies of the second model of work stress used in this field, the effort/reward imbalance model in which stress is thought to arise from an imbalance between the amount of effort the person puts into the job compared with the rewards of work (in terms of salary, prestige, advancement, etc). Other aspects of work stress that appear to contribute to CHD risk include long working hours, perceived unfairness at work, and worry about whether future employment is secure.

There are several difficulties in interpreting this evidence. One important issue is the possibility of 'reverse causality': people with subclinical CHD might inflate their ratings of work stress or take on stressful work commitments, leading to an apparent prospective association between work stress and CHD. This issue can be addressed by removing from the analysis any new cases of CHD that occur within a few years of the measurement of work stress. For example, in one large European study the risk associated with job strain was the same after excluding CHD cases that arose in the first 5 years, arguing against reverse causation (Kivimaki et al., 2012). A second problem is confounding of the association between work stress and CHD by other factors that are linked both with stress and disease (such as low education). Although social epidemiologists take other factors into account in their statistical modeling, it is not possible to be completely confident that all possible confounders have been ruled out.

Other Forms of Chronic Stress

Other types of chronic stress have been studied less extensively in relation to CHD. However, a recent *meta*-analysis of prospective studies showed that self-reports of stress are associated with an increased risk of CHD, with a risk ratio of 1.27 (95% CI 1.12–1.45) (Richardson et al., 2012). Specific forms of life stress that have been linked with CHD include social isolation, loneliness, caring for a disabled or ill spouse at home, and death of a child (Steptoe and Kivimaki, 2013). Early life deprivation is also related to the development of CHD in middle age, though there is controversy about whether early life has a direct effect, or contributes through its relationship with adult social circumstances.

Emotional Responses

An important set of psychosocial influences on CHD are psychological characteristics, notably emotional distress and some personality features. Large-scale epidemiological studies have been used to study depression, with many studies showing that both clinical depression and lower levels of depressive symptoms predict future CHD in previously healthy populations (Nicholson et al., 2006). The relative risk after adjustment for controls was 1.90 (95% CI 1.49–2.42) in one *meta*-analysis, compared with no depression. Serious mental illness in general appears to be a risk factor, while raised levels of anxiety in the general population increase CHD risk as well (Roest et al., 2010). Another psychological feature that predicts future CHD is hostility and anger. A *meta*-analysis of 21 prospective studies indicated that hostility and anger were associated with a 19% increase in risk for future CHD events (Chida and Steptoe, 2009). Hostility was a core element of the concept of Type A behavior which became popular in the 1960s and 1970s. Although the Type A concept fell from favor because of failures to replicate earlier associations with CHD, some aspects of the syndrome still appear to be relevant. There is also intriguing evidence that basic personality constructs such as conscientiousness may be important, with greater longevity and reduced CHD risk in more conscientious individuals.

Protective Psychosocial Factors

The psychosocial factors described earlier increase risk of CHD. There is also a series of factors that may reduce risk (**Table 1**). Chief among these are positive social relationships, including social integration and social support. A *meta*-analysis in 2010 reviewed 148 studies, and showed an average 50% increase in likelihood of survival for people with stronger social relationships (Holt-Lunstad et al., 2010). Of course, social relationships can have negative effects as well, since interpersonal and family conflict are among the most prominent of chronic stressors in everyday life. But having supportive relationships and a social network that provides a sense of belonging and connectedness is health protective.

Another potentially protective factor that has been increasingly studied over recent years is positive well-being.

Positive well-being includes positive hedonic states (happiness, joy), sense of meaning and purpose in life, and positive evaluations of life (life satisfaction). Current evidence suggests that hedonic well-being and characteristics such as optimism may reduce risk of future CHD (Boehm and Kubzansky, 2012). Positive well-being is not merely the absence of emotional distress, but reflects positive states that rise above the neutral.

Mediating Mechanisms

Two broad sets of mechanisms likely mediate between these psychosocial factors and the long-term development of CHD. The first involves the physiological processes contributing to CHD development (**Figure 1**, pathway 2). Thus population-level studies of work stress are corroborated by mechanistic research on the biological processes relevant to CHD. For example, studies involving repeated measurements of blood pressure over the working day show that levels are slightly elevated in people reporting low control at work, and similar patterns have been reported for cortisol. Heart rate variability, an indicator of the balance between the sympathetic and parasympathetic branches of the autonomic nervous system, is reduced over the day in people experiencing work stress or social isolation, indicating a sympathetic nervous system dominance that may be cardiotoxic. Chronic work stress also predicts the development of the metabolic syndrome, a constellation of central obesity, high blood pressure and disturbances of metabolism that presages future CHD (Chandola et al., 2008). More broadly, stress induces increases in blood pressure, neuroendocrine responses, inflammation, and hemostatic (blood clotting) processes implicated in atherosclerosis. Heightened blood pressure responses to mental stress and impaired poststress recovery are characteristic of chronically stressed individuals, and predict future cardiovascular risk (Chida and Steptoe, 2010). Heightened cortisol responses to stress and altered cortisol profiles over the day are risk factors for the development of CHD. The protective psychosocial factors described in Section Protective Psychosocial Factors appear to have the reverse effect, reducing physiological stress responsivity in cardiovascular, neuroendocrine, and inflammatory pathways.

The second set of mechanisms relate to lifestyle (**Figure 1**, pathway 1). A study of more than 100 000 working people found that job strain is related to an increased likelihood of being physically inactive, smoking, and not drinking in moderation (Heikkila et al., 2013). The elevated risk of CHD in people with serious mental illness is mediated in part through its relationship with smoking and alcohol misuse. Psychological characteristics such as depression are associated with physical inactivity and smoking. It is likely that the protective effect of conscientiousness is because more conscientious individuals have healthier lifestyles, thereby reducing their risk (Bogg and Roberts, 2004). Social network studies have found that obesity, food choice, alcohol consumption, and other behaviors relevant to cardiovascular disease are transmitted through social networks, probably through a combination of shared environments, changing norms for behavior, and direct modeling of behavior.

Acute Psychosocial Triggers of Cardiac Events

Triggers are defined as activities or stimuli that exacerbate the acute physiological and pathophysiological processes that initiate cardiac events such as acute MI or sudden cardiac death (Nawrot et al., 2011; Steptoe and Brydon, 2009). They can include physical exertion, exposure to air pollution, and respiratory infection. An important feature of triggers is that they typically act in people who have advanced coronary atherosclerosis. Secondly, the psychosocial stimuli are not long-term stressors or social conditions, but short-term experiences. Triggers are not therefore involved in the development of CHD, but rather in causing clinical problems in people with preexisting disease.

The idea that psychological stress can trigger a heart attack has a long anecdotal history, but first came to wider attention following the study of people exposed to major earthquakes. Such individuals show an excess of cardiac deaths and hospital admissions for CHD over the days following these events. An increase in rates of clinical CHD has now been reported in response to other types of disaster such as major industrial accidents, wars, and terrorist attacks. The events need not be life-threatening either, since there is evidence that watching important sports matches is also associated with peaks in cardiovascular events. For example, an evaluation of cardiac emergencies in the Munich region of Germany during the 2006 World Cup identified peaks on days on which the German team played, particularly in the 2 h after the start of matches (Wilbert-Lampen et al., 2008).

It is not completely certain in these situations that the emotional stimulus caused the cardiac event. The circumstances of each case are difficult to reconstruct in the turmoil surrounding such situations, and factors such as physical exertion, heat, excessive alcohol consumption, or exposure to pollutants may contribute. It is also true that most acute cardiovascular events do not take place in response to major events affecting entire populations, but in individuals going about their everyday lives.

Studies of emotional triggers must overcome a number of methodological difficulties including survival effects (only people who do not die can report their experience), and the retrospective collection of information which may lead to failures of recall, and to biases in recall as the person searches for meaning in their experience (Steptoe and Kivimaki, 2012). New statistical techniques (such as case-crossover models) and systematic data collection methods have enabled more precise conclusions to be drawn. A *meta*-analysis of trigger studies showed that risk of acute coronary syndromes is increased around 2.5-fold when preceded by acute anger, stress, and depressed mood in the 2 h before onset of symptoms (Steptoe and Kivimaki, 2013). It is likely that very intense stress has stronger effects that last longer; one study found that the incidence of acute MI increased more than 20-fold in the 24 h following the death of a significant person, while a large study in Sweden found that the relative risk of death due to cardiovascular disease was 5.6-fold in people within 1 week of a diagnosis of cancer.

Two aspects of this evidence should be highlighted. First, the stimuli involved predominantly involve emotional stress, not the other types of psychosocial factor as listed in **Table 1**.

Second, although effects are striking, the majority of people do not experience heightened acute emotion in the hours before their cardiac event. For example, in the study showing associations with the death of a significant person, the number of patients exposed to such an event was very small. An analysis of the population attributable risk has estimated that emotional stress is involved in around 3.9% of acute cardiac events (Nawrot et al., 2011).

Physiological Mechanisms Underlying Triggering

Mechanistic studies have identified a number of psychophysiological processes that could underlie the triggering of acute cardiac events by emotional stimuli. One possibility is stress-induced transient myocardial ischemia (MISMI), a phenomenon discovered through detailed imaging of cardiac function during acute stress (Strike and Steptoe, 2003). It has been found that in a substantial proportion of people with advanced CHD, acute stress induces reversible reductions in blood flow through the coronary arteries (ischemia). These reductions in flow are due to constriction of the coronary arteries and may lead to abnormalities of the contraction of the heart that could potentially contribute to clinical coronary events. A second mechanism is cardiac arrhythmia which may again result in failures of effective contraction of the heart muscle (Ziegelstein, 2007). Additionally, heightened inflammatory responses to stress and alterations in hemostatic (clotting) processes may be involved. What this suggests is that coronary plaque rupture following periods of acute emotional stress may coincide with an inflammatory and prothrombotic milieu inside the coronary arteries that favors the development of coronary thrombus, leading to blockage of the vessels and severe tissue damage (Steptoe and Kivimaki, 2013).

Implications for Prevention

Knowledge about the role of psychosocial stimuli in acute triggering has not yet been translated into effective preventive or management strategies. In principle, it might be possible to identify individuals who are prone to responding to emotional stress with heightened physiological responses that put them at risk for acute cardiac events, but this is beyond current technology or resources. It is probably not feasible to ensure that patients with CHD are not exposed to natural disasters or to intensive emotions. However, increased awareness of triggers both among clinicians and the public would be helpful, as would increasing the availability and effective use of emergency equipment. It is also conceivable that cardiac patients who are particularly hostile or anger prone be provided with anger management training, thereby reducing the range of situations that are likely to elicit potentially damaging psychophysiological responses (Tofler and Muller, 2006).

Stress and Recovery Following Acute Cardiac Illness

The impact of psychosocial factors does not end when a person has an MI or is diagnosed with CHD. Survivors of MI are at increased risk for future fatal and nonfatal cardiac events, and the same psychosocial factors that are associated with the long-term development of coronary atherosclerosis also predict recurrent disease. Thus chronic life stress, continued job strain, social isolation, marital conflict, and financial strain following an MI have all been linked with increased risk of recurrent cardiac events (Arnold et al., 2012; Lett et al., 2007). But the factor that has been studied most intensively is depression. Severe depression is present in around 20% of patients after an MI, with a further 25% experiencing marked depressive symptoms (Thombs et al., 2006). Cardiac patients who are depressed are at increased risk for recurrent cardiac events and for death from CHD (Meijer et al., 2011). It might be thought that depressed mood is merely a marker of more severe coronary disease, but many studies have carefully evaluated the clinical cardiological condition of patients, and have found that even when this is taken into account there is still an increased risk to future health. This is a field of intensive research, with much interest in whether certain subtypes of depression are particularly cardiotoxic. The processes responsible for the links between depression and cardiac morbidity are likely to involve both behavioral pathways – such as poor adherence to medication, physical inactivity, and other activities that are prejudicial to health – and psychobiological pathways, including disturbances of autonomic nervous system activity, neuroendocrine responses, and vascular inflammation.

Implications for Patient Care

Research on psychosocial factors in CHD patients has stimulated major efforts to modify the link with recurrent cardiac disease. The two main targets of these efforts have been the treatment of depression and the management of stress. Large-scale studies focused on treating depression with pharmacological or cognitive–behavioral methods have had disappointing results with respect to cardiac outcomes, and this has prompted a re-evaluation of the most appropriate strategies both of identifying and managing patients (Thombs et al., 2013). More sophisticated approaches are now being tested, and may prove beneficial (Davidson et al., 2013). Stress management has also been used for a number of years with cardiac patients with mixed success (Linden et al., 2007); these methods are effective in reducing self-reported stress, but not necessarily with any impact on cardiovascular health. However, a new generation of studies appears to be having more positive effects on cardiac outcomes (Gulliksson et al., 2011).

Conclusions

Psychosocial processes are emerging as important factors in CHD, opening at several stages of the disease process. There are many methodological pitfalls in this field, and it is difficult to draw definitive conclusions about the causal role of psychosocial adversity in CHD. The development of effective interventions and management strategies will help establish the importance of psychosocial factors in clinical science.

See also: Cardiac Disease, Coping With; Co-morbidity of Mental and Physical Conditions; Coping with Job Stress; Hypertension: Psychosocial Aspects; Justice and Health; Life Course Epidemiology; Occupational Health; Rehabilitation (Cardiac): Psychosocial Aspects; Social Epidemiology; Socioeconomic Status and Health; Stress and Cardiac Response.

Bibliography

Arnold, S.V., Smolderen, K.G., Buchanan, D.M., Li, Y., Spertus, J.A., 2012. Perceived stress in myocardial infarction: long-term mortality and health status outcomes. Journal of the American College of Cardiology 60, 1756–1763.

Boehm, J.K., Kubzansky, L.D., 2012. The heart's content: the association between positive psychological well-being and cardiovascular health. Psychological Bulletin 138, 655–691.

Bogg, T., Roberts, B.W., 2004. Conscientiousness and health-related behaviors: a meta-analysis of the leading behavioral contributors to mortality. Psychological Bulletin 130, 887–919.

Chandola, T., Britton, A., Brunner, E., Hemingway, H., Malik, M., Kumari, M., Marmot, M., 2008. Work stress and coronary heart disease: what are the mechanisms? European Heart Journal 29, 640–648.

Chida, Y., Steptoe, A., 2009. The association of anger and hostility with future coronary heart disease: a meta-analytic review of prospective evidence. Journal of the American College of Cardiology 53, 936–946.

Chida, Y., Steptoe, A., 2010. Greater cardiovascular responses to laboratory mental stress are associated with poor subsequent cardiovascular risk status: a meta-analysis of prospective evidence. Hypertension 55, 1026–1032.

Davidson, K.W., Bigger, J.T., Burg, M.M., Carney, R.M., Chaplin, W.F., Czajkowski, S., Ye, S., 2013. Centralized, stepped, patient preference-based treatment for patients with post-acute coronary syndrome depression: CODIACS vanguard randomized controlled trial. JAMA Internal Medicine 173, 997–1004.

Gulliksson, M., Burell, G., Vessby, B., Lundin, L., Toss, H., Svardsudd, K., 2011. Randomized controlled trial of cognitive behavioral therapy vs standard treatment to prevent recurrent cardiovascular events in patients with coronary heart disease: secondary prevention in Uppsala Primary Health Care Project (SUPRIM). Archives of Internal Medicine 171, 134–140.

Heikkila, K., Fransson, E.I., Nyberg, S.T., Zins, M., Westerlund, H., Westerholm, P., Consortium, I.P.-W, 2013. Job strain and health-related lifestyle: findings from an individual-participant meta-analysis of 118,000 working adults. American Journal of Public Health 103, 2090–2097.

Holt-Lunstad, J., Smith, T.B., Layton, J.B., 2010. Social relationships and mortality risk: a meta-analytic review. PLoS Medicine 7, e1000316.

Kaplan, J.R., Manuck, S.B., 1999. Status, stress, and atherosclerosis: the role of environment and individual behavior. Annals of the New York Academy of Sciences 896, 145–161.

Kivimaki, M., Nyberg, S.T., Batty, G.D., Fransson, E.I., Heikkila, K., Alfredsson, L., Theorell, T., 2012. Job strain as a risk factor for coronary heart disease: a collaborative meta-analysis of individual participant data. The Lancet 380, 1491–1497.

Lett, H.S., Blumenthal, J.A., Babyak, M.A., Catellier, D.J., Carney, R.M., Berkman, L.F., Schneiderman, N., 2007. Social support and prognosis in patients at increased psychosocial risk recovering from myocardial infarction. Health Psychology 26, 418–427.

Linden, W., Phillips, M.J., Leclerc, J., 2007. Psychological treatment of cardiac patients: a meta-analysis. European Heart Journal 28, 2972–2984.

Mackenbach, J.P., Stirbu, I., Roskam, A.J., Schaap, M.M., Menvielle, G., Leinsalu, M., Kunst, A.E., 2008. Socioeconomic inequalities in health in 22 European countries. New England Journal of Medicine 358, 2468–2481.

Meijer, A., Conradi, H.J., Bos, E.H., Thombs, B.D., van Melle, J.P., de Jonge, P., 2011. Prognostic association of depression following myocardial infarction with mortality and cardiovascular events: a meta-analysis of 25 years of research. General Hospital Psychiatry 33, 203–216.

Nawrot, T.S., Perez, L., Kunzli, N., Munters, E., Nemery, B., 2011. Public health importance of triggers of myocardial infarction: a comparative risk assessment. The Lancet 377, 732–740.

Nicholson, A., Kuper, H., Hemingway, H., 2006. Depression as an aetiologic and prognostic factor in coronary heart disease: a meta-analysis of 6362 events among 146 538 participants in 54 observational studies. European Heart Journal 27, 2763–2774.

Richardson, S., Shaffer, J.A., Falzon, L., Krupka, D., Davidson, K.W., Edmondson, D., 2012. Meta-analysis of perceived stress and its association with incident coronary heart disease. American Journal of Cardiology 110, 1711–1716.

Roest, A.M., Martens, E.J., de Jonge, P., Denollet, J., 2010. Anxiety and risk of incident coronary heart disease: a meta-analysis. Journal of the American College of Cardiology 56, 38–46.

Steptoe, A., Brydon, L., 2009. Emotional triggering of cardiac events. Neuroscience Biobehavioural Review 33, 63–70.

Steptoe, A., Kivimaki, M., 2012. Stress and cardiovascular disease. Nature Reviews Cardiology 9, 360–370.

Steptoe, A., Kivimaki, M., 2013. Stress and cardiovascular disease: an update on current knowledge. Annual Review of Public Health 34, 337–354.

Strike, P.C., Steptoe, A., 2003. Systematic review of mental stress-induced myocardial ischaemia. European Heart Journal 24, 690–703.

Thombs, B.D., Bass, E.B., Ford, D.E., Stewart, K.J., Tsilidis, K.K., Patel, U., Ziegelstein, R.C., 2006. Prevalence of depression in survivors of acute myocardial infarction. Journal of General Internal Medicine 21, 30–38.

Thombs, B.D., Roseman, M., Coyne, J.C., de Jonge, P., Delisle, V.C., Arthurs, E., Ziegelstein, R.C., 2013. Does evidence support the American Heart Association's recommendation to screen patients for depression in cardiovascular care? an updated systematic review. PLoS One 8, e52654.

Tofler, G.H., Muller, J.E., 2006. Triggering of acute cardiovascular disease and potential preventive strategies. Circulation 114, 1863–1872.

Wilbert-Lampen, U., Leistner, D., Greven, S., Pohl, T., Sper, S., Volker, C., Steinbeck, G., 2008. Cardiovascular events during World Cup soccer. New England Journal of Medicine 358, 475–483.

Ziegelstein, R.C., 2007. Acute emotional stress and cardiac arrhythmias. Journal of the American Medical Association 298, 324–329.

Corporate Culture

Edgar H Schein, MIT Sloan School of Management, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Corporate culture is described and defined as an extension of the general concept of culture. Culture is both a state or property of a group and a perpetual process of reenactment among the members of the group. Corporate culture can be defined as the learned outcome of a group's solving its problems of external survival and internal integration. A model of culture is described and a number of approaches to describing the content of cultures are reviewed and illustrated. The degree to which corporate culture relates to economic performance is briefly reviewed and the concept of culture creation and change is discussed.

How to Think about and Define 'Culture'

To understand *corporate* culture one must first understand the concept of *culture*. A chronic issue in conceptualizing culture is whether to think of culture as a static *property* of a given organization – its shared customs, beliefs, norms, values, and tacit assumptions – or as a *dynamic human process* of constructing shared meaning (Frost et al., 1985). Culture *creation* is one of the unique characteristics of humans, being based on our capacity to be self-conscious and able to see ourselves and others from each others' points of view. It is this reflexive capacity of humans that makes culture possible. At the same time, it is the human need for finding meaning that creates the motivation for culture stabilization. Without some predictability, social intercourse becomes too anxiety provoking. Developing shared meanings of how to perceive, categorize, and think about what goes on around us is necessary to avoid the catastrophic anxiety that would result from reacting to everything as if it were a new phenomenon (Weick, 1995; Weick and Sutcliffe, 2007).

Given these human characteristics, it then becomes clear that culture is both a process and a state (Schein, 2010). In new situations, shared meanings must be constructed through a social learning process. As these meanings help the participants to make sense of their world they become stabilized and can be viewed as 'states.' At the same time, as the members of a group interact, they not only recreate and ratify prior meanings but also construct new meanings as new situations arise.

A useful way to think about this issue is to take a cue from the anthropologist Marshall Sahlins (1985) who argues that one cannot really understand certain social phenomena without understanding both the historical events and the cultural meanings attributed by the actors to those events. While it is undeniably true that we produce and reinforce culture through perpetual enactment and sense making, it is equally true that the actors in those same social events bring to them some prior meanings, stereotypes, and expectations that can only be understood in a historical context. Culture production in the enactment sense, then, is either the perpetuation of or a change of some prior state, but that prior state can be thought of as 'the culture' up to that point. One can then describe that culture as if it were a 'state' of the existing system, even though one knows that the system is dynamic and perpetually evolving. The direction of that evolution will be a product of several forces: (1) technological and physical changes in the external environment; (2) changes in the internal dynamics of the social system; and (3) historical circumstances that are fortuitous or serendipitous.

A good example is Sahlins' analysis of the death of Captain Cook at the hand of the Hawaiians. Because Captain Cook was viewed as a God (as predicted in the Hawaiian mythology), the sexual favors offered to his sailors by the Hawaiian women were viewed as gifts and as opportunities to relate to the divine. The sailors' cultural background defined this as a version of prostitution, however, for which they felt they should pay. When they offered the women something in exchange for the sex, the women asked for something that was scarce in the society, namely metal. Once the loose metal on board the ships had been used up, the sailors began to pull nails from the ship itself, weakening it structurally. Hence, when Captain Cook set sail he discovered that the ships needed repair and ordered a return to harbor. In Hawaiian mythology, a God returning under these circumstances had to be ritually killed. At the same time, Hawaiian social structure was undergoing change and became permanently altered because the subordinate role of women in the society was altered by their ability to acquire metal, a scarce resource that gave them social power.

Clearly, there was a culture in Hawaii and a different culture on board the British ship, and clearly the interaction of these two cultures produced events that changed both of these cultures. Culture in this sense evolves in all social systems that have any kind of stability of membership and a history of experiences. Thus, nations, religious groups, and occupational groups based on professional education and shared experience all have cultures and increasingly bring those cultures into organizations.

Implications for 'Organizational' Culture Analysis

The major lesson of Sahlins' analysis is that *when we have access to historical data we should use it*. Organizations have defined histories. Therefore, if we want to analyze organizational cultures, we should reconstruct their histories, find out about

the values of their founders and early leaders, look for the critical defining events in their evolution as organizations, and be confident that when we have done this we can indeed describe sets of shared assumptions that derive from common experiences of success and/or shared traumas. We can then legitimately think of these sets of assumptions as 'the corporate culture' at a given time. But the organization will also include subcultures based on occupational units or shared experiences. Some of these may be in conflict with each other and there may be subunits that have not yet had enough shared experience to have formed shared common assumptions (Martin, 2002).

We can also study the day-to-day interactions of the members of an organization with each other and with members of other organizations to determine how given cultural assumptions are reinforced and confirmed, or challenged and disconfirmed (Kunda, 2002). We can analyze the impact of these perceptual interactive events in order to understand how cultures evolve and change. This process could be especially productive in mergers, acquisitions, and joint ventures of various sorts. Whether one chooses to focus one's cultural research on building typologies of cultural 'states,' categories that freeze a given organization at a given point in time, or on analyzing the moment-to-moment interactions in which members of a given social system attempt to make sense of their experience and, in that process, reinforce and evolve cultural elements, becomes a matter of choice. Both are valid methodologies and in practice they should probably be combined.

A Formal Definition of Culture

Culture can now be defined as:

> "A pattern of shared basic assumptions that the group has learned as it solved its problems of external adaptation and internal integration, that has worked well enough to be considered valid and, therefore, to be taught to new members as the correct way to perceive, think, and feel in relation to those problems" (Schein, 2010).

Once these patterns have been learned, they function to reduce anxiety and provide moment-to-moment meaning and predictability to daily events. Or, as Trice and Beyer (1993) put it: "Human cultures emerge from people's struggles to manage uncertainties and create some degree of order in social life" (p. 1). Note that patterns of overt *behavior* are not part of the formal *definition* because such patterns can be elicited by other causes such as common instinctive patterns (e.g., ducking when we hear a loud noise) or by common reactions to a common stimulus (e.g., everyone running in the same direction to avoid some threat). If one understands the shared tacit assumptions, one can determine which regularities of behavior are 'cultural' and which ones are not. One cannot, however, infer the assumptions just by observing behavior.

Culture defined in this way can apply to an entire organization, group, or occupational unit provided there is a common history and member stability. If a group has no history or its members change all the time, it does not develop a culture.

The main problem in defining 'corporate culture' derives from the fact that any organization in today's multicultural world is a conglomerate of cultures and subcultures and is therefore also influenced by the deeper assumptions of nations, ethnic groups, and occupational units.

A Conceptual Model for Analyzing Culture

Whether one chooses to analyze culture as a state or as a process, it is helpful to differentiate the levels at which culture as a shared phenomenon manifests itself. Any group, organization, or larger social system can be analyzed in terms of (1) its visible and feelable 'artifacts,' (2) its espoused beliefs and values, and/or (3) its less visible, taken for granted shared tacit assumptions (see Figure 1).

This model is useful in two ways: (1) It helps a novice observer of organizational culture to differentiate the superficial manifestations and espoused values that most organizations display from the cultural substrate or essence, the tacit, shared assumptions that drive the day-to-day behavior that one observes; and (2) The differentiation of levels is a necessary conceptual tool in helping an organization to decipher its own corporate culture and subcultures.

To further understand these distinctions, especially the distinction between espoused values and shared tacit

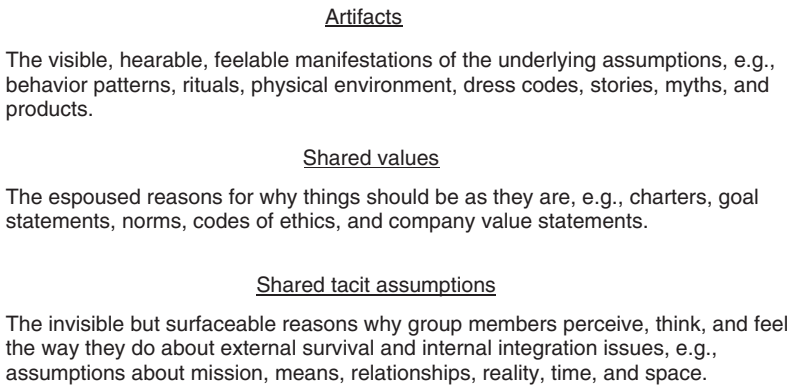

Figure 1 The levels of culture.

assumptions, it is necessary to take a historical evolutionary point of view toward culture formation. If we take a typical business organization, the process starts with one or more entrepreneurs who found a company based on some *personal* beliefs and values. As they hire others to work with them, they will either choose those others on the basis of their compatibility or will socialize newcomers to the beliefs and values that they regard as core to running the new business. The founders' beliefs and values will cause the organization to make decisions in its environment and, if those decisions are successful, the newcomers will begin to entertain collectively the idea that the founder's beliefs and values must be 'correct.' If the environment continues to reinforce the behavior of the organization, what was originally the founder's *personal* beliefs and values gradually come to be *shared* and *imposed on new members* as the organization grows. If this success cycle continues, these shared beliefs and values gradually come to be taken for granted, drop out of awareness, and can, therefore, be thought of as 'taken for granted assumptions' that become increasingly nonnegotiable.

This process of culture formation will occur not only in reference to the organization's primary task in its various external environments but also with respect to its internal organization as well. Those founder beliefs and values, which provide meaning, make life livable and reduce interpersonal anxiety, will gradually come to be taken for granted and come to be tacit assumptions about the 'correct' way to organize. But in both domains, the driving force is what actually works. Culture is the *result* of successful action. If things do not work out, the group and, therefore, the culture as a culture will disappear but its assumptions may be carried on in individual survivors of the organization as was the case with Digital Equipment Corporation (Schein, 2003).

A basic need in human groups is to justify what they do. Paradoxically, the justifications often are not the same as the tacit assumptions that actually determine the behavioral regularities. Thus groups create ideologies, aspirations, visions, and various other kinds of 'espoused values,' which may or may not correspond isomorphically with the tacit assumptions. The most difficult aspect of deciphering organizational cultures, then, is to determine to what extent the claimed espoused values actually correspond to the behavioral regularities observed and, if not, to determine what the shared tacit assumptions are.

How to Describe Cultures

There are three approaches to describing cultures: (1) profiling them on various preselected dimensions; (2) creating conceptual typologies into which to fit given cultures; and/or (3) clinical or ethnographic descriptions that highlight unique aspects of a given culture.

Among the profiling approaches, one of the most widely used is Hofstede's four dimensions based on factor analyzing questionnaire responses: individualism (vs groupism), masculinity (the perceived gap between male and female roles), tolerance for ambiguity, and power distance (the perceived distance between the most and least powerful in the society). These dimensions were originally used to describe national cultures but have also been extended to descriptions of organizational cultures (Hofstede, 2001).

A different approach has been to use sociological dimensions developed by Parson (1951) and elaborated by Kluckhohn and Strodtbeck (1961) as a way of creating profiles of a given organization and identifying critical value issues – universalism vs particularism, individualism vs collectivism, affectively charged vs neutral, specific vs diffuse, time orientation, and degree of control over nature (Hampden-Turner and Trompenars, 1993; Trompenars and Hampden-Turner, 2011). A third version of this approach is to take one or two dimensions, consider them to be conceptually central, and create a typology based on a two-by-two table with those dimensions. Cameron and Quinn (2011) use flexibility vs stability and control and internal integration vs external differentiation to create four types of cultures: clan, adhocracy, hierarchy, and market. Goffee and Jones (1998) use the dimensions of degree of solidarity and degree of sociability to create four types: networked, communal, fragmented, and mercenary. Among the typologies we have Likert's Systems 1 to 4, Harrison and later Handy's 'Gods of Management,' McGregor's Theory X and Theory Y, Ouchi's Theory Z, and the more popular distinctions between 'command and control' vs 'committed and empowered' (Likert, 1967; Handy, 1978; McGregor, 1960; Ouchi, 1981).

The clinical or ethnographic approach attempts to deal with any given culture as a unique pattern of shared assumptions around issues that are not known before the clinician/researcher is on the scene as an observer and active inquirer (Schein, 1987). The prime objection to questionnaires as research tools for the study of *culture* is that they force us to cast the theoretical net too narrowly. The advantage of the ethnographic or clinical research method is that we can consciously train ourselves to minimize the impact of our own models and to maximize staying open to new experiences and concepts we may encounter. In the end, we may well sort those experiences into the existing categories we already hold. But at least we will have given ourselves the opportunity to discover new dimensions and, more importantly, will have a better sense of the relative salience and importance of certain dimensions within the culture we are studying. The issue of salience is very important because not all the elements of a culture are equally potent in the degree to which they determine behavior. The more open group–oriented inquiry not only reveals how the group views the elements of the culture, but, more importantly, also tells us immediately which things are more salient and, therefore, more important as determinants (Schein, 1996, 2003).

As to general dimensions that would be useful for studying any organization, I have found it empirically useful to start with a broad list of (1) 'survival functions' (what any group must do to survive in its various environments and fulfill its primary task) and (2) 'internal integration functions' (what any group must do to maintain itself as a functioning system). This distinction draws on a long tradition of empirical research in group dynamics that always turns up two critical factors in what groups do: (1) task functions and (2) group building and maintenance functions. Ancona (1988) pointed out that we must add a third set to these two, boundary maintenance functions. Task and boundary maintenance functions are

external survival issues, and group building and maintenance functions are internal integration issues. We may then construct different lists of what specific dimensions of behavior, attitude, and belief, we will look for in each domain but at least we have a model that forces us to cast the net widely, and a reminder that culture is for the group the learned solution to all of its external and internal problems.

To the extent that national and occupational cultures increasingly influence organizations one must also consider the influence of deeper assumptions about time and space, the human nature and human relationships, how the self is defined in relation to the collective, how the group views itself in relation to nature, and the nature of reality itself and how one arrives at truth (Schein, 2009, 2010).

Does Corporate Culture Matter?

Several claims based on various different kinds of research have tried to show a connection between the strength and/or type of culture and economic performance (Davis, 1984; Deal and Kennedy, 1982; Denison et al., 2012; Kotter and Heskett, 1992; Collins and Porras, 1994). The problem is that different kinds of cultural dimensions relate to different kinds of environments in ways that are not entirely predictable. The best way to summarize, then, is to say that culture certainly influences economic performance, but the manner in which this occurs remains highly variable. A culture that can be very functional in one environment or at one stage in a company's evolution can become dysfunctional and cause that same company to fail. Culture needs to be analyzed and understood but until much more research has been done, one cannot make generalizations about its impact.

Can Culture Be Created or Changed?

As indicated earlier, culture is created when a group is created by its founders. It then evolves as a function of actual experience. That evolution can be influenced by leaders setting new examples and mandating new kinds of behavior, but the culture will not change until the new behavior produces more desirable results and, thereby, comes to be accepted and eventually taken for granted. This is usually a slow process that can take many years (Schein, 2009, 2010).

In most mature organizations in today's more complex multinational world, one sees cultural evolution occurring in many different ways in the subunits of the organization. The question of 'culture change' then becomes the more difficult question of which elements of the culture are to be changed and, most importantly, why? When organizations face these questions it is best to focus on why the organization needs any change at all. Once a case is made for change in relation to organizational performance, one should then specify how that change would look behaviorally in the future, and only then consider how various elements of the culture or various subcultures would aid or hinder the proposed change (Schein, 2010).

See also: Bureaucracy and Bureaucratization; Bureaucracy, Sociology of; Corporate Governance; Entrepreneurship and Economic Growth; Ideology, Sociology of; Leadership in Organizations, Sociology of; Management: General; Myths and Symbols: Organizational; Organizational Climate; Organizational Culture, Anthropology of; Organizational Ecology; Organizations, Sociology of.

Bibliography

Ancona, D.G., 1988. Groups in organizations: extending laboratory models. In: Hendrick, C. (Ed.), Annual Review of Personality and Social Psychology: Group and Intergroup Processes. Sage, Beverly Hills, CA.
Cameron, K.S., Quinn, R.E., 2011. Diagnosing and Changing Organizational Culture, third ed. Wiley, New York.
Collins, J.C., Porras, J.I., 1994. Built to Last. Harper, NY.
Davis, S., 1984. Managing Corporate Culture. Ballinger, Cambridge, MA.
Deal, T.E., Kennedy, A.A., 1982. Corporate Cultures. Addison-Wesley, Reading, MA.
Denison, D., Hooijberg, R., Lane, N., Leif, C., 2012. Leading Culture Change in Global Organizations: Aligning Culture and Strategy. Jossey-Bass, San Francisco, CA.
Frost, P.J., Moore, L.F., Louis, M.R., Lundberg, C.C., Martin, J. (Eds.), 1985. Organizational Culture. Sage, Thousand Oaks, CA.
Goffee, R., Jones, G., 1998. The Character of a Corporation. Harper, NY.
Hampden-Turner, C., Trompenars, A., 1993. The Seven Cultures of Capitalism. Doubleday-Currency, NY.
Handy, C., 1978. Gods of Management. Pan Books, London.
Hofstede, G., 2001. Culture's Consequences, second ed. Sage, Thousand Oaks, CA.
Kluckhohn, F.R., Strodtbeck, F.L., 1961. Variations in Value Orientations. Harper, NY.
Kotter, J.P., Heskett, J.L., 1992. Corporate Culture and Performance. Free Press, NY.
Kunda, G., 2002. Engineering Culture, Rev. ed. Temple University Press, Philadelphia.
Likert, R., 1967. The Human Organization. McGraw-Hill, NY.
Martin, J., 2002. Organizational Culture. Sage, London.
McGregor, D., 1960. The Human Side of Enterprise. McGraw-Hill, NY.
Ouchi, W.G., 1981. Theory Z. Addison-Wesley, Reading, MA.
Parson, T., 1951. The Social System. Free Press, Glencoe, IL.
Sahlins, M., 1985. Islands of History. University of Chicago Press, Chicago, IL.
Schein, E.H., 1987. The Clinical Perspective in Fieldwork. Sage, Newbury Park, CA.
Schein, E.H., 1996. Strategic Pragmatism: The Culture of Singapore's Economic Development Board. MIT Press, Cambridge, MA.
Schein, E.H., 2003. DEC Is Dead: Long Live DEC: The Lasting Legacy of Digital Equipment Corporation. Berrett/Kohler, San Francisco, CA.
Schein, E.H., 2009. The Corporate Culture Survival Guide, second ed. Jossey-Bass, San Francisco, CA.
Schein, E.H., 2010. Organizational Culture and Leadership, fourth ed. Jossey-Bass, San Francisco, CA.
Trice, H.M., Beyer, J.M., 1993. The Cultures of Work Organizations. Prentice-Hall, Englewood Cliffs, NJ.
Trompenars, A., Hampden-Turner, C., 2011. Riding the Waves of Culture. McGraw-Hill, New York.
Weick, K.E., 1995. Sensemaking in Organizations. Sage, Beverly Hills, CA.
Weick, K.E., Sutcliffe, K.M., 2007. Managing the Unexpected, second ed. Wiley, New York.

Corporate Finance: Financial Control

Marta Renzetti, SDA Bocconi-Corporate Finance, Milan, Italy

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 4, pp. 2792–2797, © 2001, Elsevier Ltd.

Abstract

Corporate finance is concerned with financial decisions made by corporations. These decisions can be grouped easily into two major categories: investment decisions and financing decisions. Investment decisions and financing decisions must contribute together to create value for the company's shareholders. In order to make optimal financial decisions in keeping with the major objective of the firm (i.e., maximizing its shareholders' wealth), the financial manager adopts some analytical tools in the analysis, planning, and control activities of the firm. In fact, financial analysis is a necessary condition for making appropriate financial decisions. Accordingly, financial control can be defined as the whole of activities and tools that enable financial managers to carry out financial analyses for purposes of internal and external control. Since they need to keep under control both the profitability of capital invested and the cost of resources raised from investors, financial managers make use of an integrated system of analyses that includes financial statements, ratios, financial planning, investment valuation, and capital structure analysis. Each of these tools plays a key role with reference to the financial control activity, providing the financial manager with a wide range of information that helps the company's investment and financing decisions.

Corporate finance is concerned with financial decisions made by corporations. These decisions can be grouped easily into two major categories: investment decisions and financing decisions.

Regarding the first major decisions a firm must take, each company needs real assets (tangible and intangible) to carry on its business and keep its operations going. How much a firm should invest and what specific assets a firm should invest in are the typical questions related to investment decisions. Investment decisions, then, determine the total amount of assets held by a company and the composition of these assets. The economic evaluation of investment projects aims at orienting corporations' choices connected to their asset allocation process. This process that leads a company to allocate its capital to investment proposals whose benefits are to be realized in the future is generally referred to as *capital budgeting*.

All the assets we mentioned above are to be paid for. To raise the money necessary to their investments policy, corporations issue securities on financial markets. These financial assets have value since they represent claims on companies' real assets. Securities issued by corporations include not only shares of stock but also bonds, bank loans, lease obligations, and any other form of financing. How the cash required for investments should be obtained is the answer that financing decisions must give. Financing decision, in fact, are concerned with determining the best mix of financial sources in the form of equity and debts. In other words, financing decisions define the total amount and the composition of a company's capital structure. The definition of capital structure referred to is that which takes into account the explicitly costly forms of financing raised from the company's investors. For sake of simplicity, they are divided into financial debts and equity without dwelling on the different form of financing.

According to what has been highlighted about corporate finance decisions, financial management attempts to find specific answers to needs arising from both investment and financing areas.

After defining what corporate finance decisions are concerned with, the next step in this analysis will be to reread investment policies and capital structure decisions in the light of the theory of shareholders' value. Since the principle of maximizing shareholders' wealth largely is accepted as the rational guide for running a business and for the efficient allocation in an economy, it is assumed to be the objective in considering how financial decisions should be made.

Investment decisions and financing decisions must contribute together to create value for the company's shareholders. From one hand, a financial manager that acts in its shareholders' interest should invest in those projects that increase the overall firm's value and, then, its shares' value. From the other hand, the mix of financial resources that maximizes the company's value should be adopted. From this point of view, it becomes crucial to understand how corporate borrowing affects its shareholders' wealth according to Modigliani and Miller proposition II. In their article, Modigliani and Miller (1963) state that, in a world with taxes, the value of a levered firm increases in proportion to the debt–equity ratio (Debt/Equity). As we will show in Section Capital Structure Analysis this is due to the fact that financial charges are tax deductible whereas dividends and retained earnings are not; an increase in debts will reduce the company's cost of capital and raise the firms' value.

In addition to that, it must be kept in mind that investment decisions and financing decisions interact, sometimes producing different consequences on the company's value. In fact, any change in the firm's capital structure influences the cost of capital that a company applies to discount investments cash flows. Consequently, any rise in the cost of capital could make any investment less profitable, reducing its contribution to the overall company's value.

From what has been said, it can be concluded that both investment and financial decisions have to be analyzed in relation to their impact on the firm's major objective: to make the firm's shareholders as well off as possible. As a result, an

optimal combination of the two areas of decision will maximize the value of the firm for its shareholders.

Consistent with that, the financial manager must act as an intermediary between the firm's operations and capital markets where the firm's securities are traded. The purpose of capital markets is allocating savings efficiently in an economy, from ultimate savers to ultimate users of funds who invest in real assets (companies). In fact, companies issue securities and raise cash by selling those financial assets to investors. Then the cash is invested in the firm's operations and used to purchase real assets. Later, if the firm does well, the real assets generate cash inflows that more than repay the initial investment. Finally, once the firm's operations generate new cash, this is either reinvested in the company or returned to investors who hold the securities originally issued. The last decision is not completely free for the financial manager, but it depends on several factors such as the characteristics of the specific financing, the dividend policy, the financial needs of the company.

Corporate Finance Decisions and Financial Control

In order to make optimal financial decisions in keeping with the major objective of the firm (i.e., maximizing its shareholders' wealth), the financial manager must implement inside the firm an effective process of analysis, planning, and control and make use of some analytical tools. In fact, financial analyses are a necessary condition for making appropriate financial decisions in the meaning as stated before. From a very simplistic point of view, in order to create value for all the company's investors (debtholders and stakeholders), the asset allocation process must yield more than the company's cost of capital.

Since the financial manager has to grant that the cash flow ability of the firm will properly remunerate its financial sources over the long run, they must evaluate the size and timing of cash flows and monitor the company's capital structure, its major sources and uses of funds, and its profitability over time. As a measure of the company's profitability ROI (return on investments = operating income/capital invested) is taken into consideration. According to that, the financial manager is concerned on one hand with the profitability of the company's investments and with the efficiency of the implemented asset management decisions. It means that they have to make use of rational tools to evaluate capital budgeting decisions. On the other hand, the financial manager is interested in keeping constantly under control the changes in the company's capital structure that can affect both the firm's cost of capital and the yield required by capital suppliers. If the company's cost of capital is defined as the weighted average cost of capital (WACC) and represents the net cost of debts and equity weighted by their percentage on the total capital structure, an increase in the percentage of debts, for instance, can reduce the WACC since they are less costly than equity, due to their fiscal deductibility; at the same time, an excessive rise in debts can make the company be perceived as a riskier investment and thus induce investors to require higher yields on debt and equity. Both these effects have to be taken into account once the financial manager is approaching any financing decision.

In the light of what financial managers must do to maximize shareholders' value, *financial control* can be defined as the whole of the activities and tools that enable financial managers to carry out financial analyses for the purposes of internal and external control. By internal control, we mean evaluating the investments profitability and the efficiency/effectiveness of the capital budgeting process. In addition to that, a form of internal control is considered as any analysis turned to keep under observation the composition of the company's capital structure. External control is any aspect of financial analysis that outside suppliers of capital take into consideration in evaluating the firm and that the financial manager should be interested in the endeavor of bargaining more effectively for outside funds.

So far it has been stated that the main corporate goal is to increase the shareholders' wealth. According to that, a company creates value if its shareholders' wealth rises. It is worth mentioning that, recently, financial markets and investors have taken into account new parameters in the evaluation of financial investments. These parameters add a social and ethical appraisal of financial investments to a pure financial valuation. Such parameters aim at reminding companies of their social responsibilities and spur them to give constructive contributions to society and bring positive energy to the solutions of its problems.

As a consequence of the new approach to interpreting the corporate role within the society, investors choose to invest their money only in those companies that behave ethically. On the other hand, financial products suppliers are tailoring new financial products which are labeled as ethical, and that respond to the new evaluation requirements.

The Dow Jones Sustainability Index (DJSGI) can be seen as an example of this new trend inside financial markets. That index, in fact, consists of more than 200 companies that represent the top 10% of leading sustainability companies in 64 industry groups in the 33 countries covered by the DJSGI.

The companies included in the DJSGI have shown over time a strong commitment on the achievement of the five corporate sustainability performance principles: innovation, high standard corporate governance, shareholders' requirements, best practices, and social behaviors.

From what has just been highlighted, managers must be aware that new principles beyond those suggested by financial theories have to be pursued in running their business.

An Integrated System of Financial Analyses

An effective financial analysis, planning, and control process is critical to enhance enterprise value and, then, shareholders' wealth. To make rational decisions in the shareholders' interest, financial managers need to be supported by an integrated system of financial analyses and reports. These tools enable financial managers constantly to control both the profitability of the company (in the meaning of the firm's ability to generate cash flows from its operating activities) and the cost of resources applied to undertake operating investments that determine the company's cost of capital.

Since they need to keep under control both the profitability of capital invested and the cost of resources raised from

investors, financial managers make use of an integrated system of analyses that includes financial statements, ratios, financial planning, investment valuation, and capital structure analysis. Each of these tools plays a key role with reference to the financial control activity, providing the financial manager with a wide range of information that helps in orienting the company's investment and financing decisions. Even if financial analysis, planning, and control activities are often considered as different tools, they have to be viewed as a whole since they together aim at accomplishing the maximization of shareholders' value.

In large companies it often happens that different managers carry out the above mentioned analyses in different departments. It becomes important, then, to construct an integrated system of data exchange that allows the different tools to work jointly. To assess an investment's profitability, for instance, it is necessary to forecast the company's market share in order to determine its potential incremental revenues and cash flows. In addition to that, the composition of the expected capital structure will be also needed to calculate the new cost of capital and discount future cash flows at that rate.

Financial Accounting

Financial accounting (financial statements, ratios, and cash flow statements) gives the financial manager a picture of the firm's past performance and its current standing. These analyses help the financial manager to understand the company's current economic and financial situation and make forecasts for the future. It also allows the manager to represent the firm's position according to its competitors and its evolution over time.

The reorganization of financial statements (assets and liabilities statement and income statement) according to some predefined schemes is the first step of a process that, starting from accounting information, ends up with the appraisal of a company's performance from both income and cash flow perspective.

One of the criteria applied to reorganize the company's financial statements is, for instance, the activity-related method. With relation to this approach, we group the balance sheet entries in current and noncurrent assets and liabilities according to their pertinence or otherwise to the operating cycle of the company. The same criterion applied to the items contained in the income statement allows us to separate revenues and costs in relation with the four activity areas that can be identified in an ongoing business: current, financial, subsidiary, and extraordinary activities.

Once financial statements have been reorganized, it is possible for the financial manager to combine income statement and balance sheet entries in order to obtain significative ratios. Second, from reorganized financial statements, a cash flow statement can be drawn up as well.

In particular, a complete system of ratios describes the current state of health of the company according to different points of view: liquidity, profitability, leverage. This analysis offers a static diagnosis of the firm that can be helpful for-managerial control and for a better understanding of what outside suppliers of capital may expect in terms of a company's financial condition and performance. In fact, both the financial manager and the company's investors are interested in measuring the return on investments of the company, its capability of repaying debts in the short run and in the long run, and the relation between debt and equity. Consistent to that aim, ratios are mostly employed for *ex-post* analysis, to verify the effect of investment and financing decisions on the company's profitability, and for the composition of its assets and liabilities.

The final step in the above mentioned process of financial analysis is represented by the construction of the cash flow statement. Cash flow analysis can be seen as a dynamic analysis that highlights the company's ability of generating cash flow. Using cash flow, the financial manager identifies the sources and uses of monetary resources during a given period. The cash flow statement groups all the inflows and outflows according to their pertinence to the current activity, the financing and reimbursement activity, the investment activity (operating and nonoperating), and to the remuneration of debt holders and stakeholders. In other words, this tool leads to the quantification of the cash inflows that the company is currently able to generate from its operating activity, and the comparison of them to the flows absorbed by the investment policy, the debt repayment, and the dividend policy. In fact, it is one thing if a company can provide the monetary resources autonomously to undertake investments and adequately remunerate the capital invested, while it is quite another if the same company regularly needs to find outside sources of capital in order to avoid financial distress caused by an insufficient production of cash inflow.

Consistent with that, the cash flow analysis can be interpreted as an *ex-post* analysis that helps the financial manager to appraise the impact of investment and financing decisions in terms of financial flows.

Financial Planning

Financial planning enables financial managers to assess the effects of financing and investment decisions *ex-ante*. Financial planning is a necessary analysis as investment and financing decisions interact and they should not be made independently. Financial planning is concerned with analyzing financial flows of the firm as a whole, forecasting the consequences of investment and financing decisions, and weighing the effects of different alternatives. In fact, any financial decision that affects the investment policy of the firm or its mix of financial sources raised from investors influences the company's future profitability as well as its capital structure (and its cost of capital). The financial manager has to define both a target financing mix and a target return on investments and undertake only the investments/financings consistent with those targets. In order to keep under control any unexpected change that can modify the company's return on investments or its cost of capital, the financial manager has to analyze any financial strategy very carefully and to take into consideration any economic and financial impact that it may have. Moreover, financial planning makes possible the comparison of different investment and financing strategies in order to evaluate which one maximizes shareholders' wealth.

The tangible tool applied in the planning process is the financial plan describing the firm's financial strategy and projecting its future consequences in a proforma balance sheet, income statement, and cash flow statement. In particular, the forecasted cash flow statement provides an effective method for the financial manager to assess the growth of the company and its resulting financial needs as well as to determine the best alternative to finance those needs.

In fact, the plan sets financial goals and it represents a benchmark itself for evaluating subsequent performance. Helping the company to achieve these goals is one of the chief responsibilities of financial planning.

Finally, the planning horizon depends on the company and its business. High-tech companies, for instance, produce much shorter financial plans than most companies do as a consequence of technologies life cycle, whereas public utilities and energy companies make longer financial plans due to the long lead time of their capital projects.

Investments Valuation

Capital budgeting involves the entire process of planning capital expenditures whose returns are normally expected to extend beyond 1 year. Since we stated that investment decisions must be made so that they maximize shareholders' value, capital budgeting decisions forcedly must be related to the firm's overall strategic planning. Capital budgeting must be integrated with strategic planning as excessive investments or inadequate investments could cause serious consequences for the future of the firm. If the company invested too much in fixed assets overestimating its potential growth, it would incur unnecessarily heavy expenses that would reduce its return on investments. On the other hand, if a company has not invested enough in new production capacity it may lose a portion of its customers to rival firms. A related problem is how to properly phase the availability of capital assets in order to have them 'working' at the correct time. An effective process of capital budgeting aims at improving the timing of asset acquisitions and the quality of assets purchased. Another reason that highlights the relevance of the capital budgeting process is that asset expansion typically involves substantial expenditures. Before a company undertakes an expensive investment such as the purchase of a new plant, the launch of a new product, or the signature of a commercial agreement, it needs to plan how to raise funds required for that specific project. A number of factors make capital budgeting one of the major financial management decisions. In fact, not only does any investment decision contribute to determine the company's profitability (and its ROI) and, then, the shareholders' value, but also the capital budgeting process vitally affects all departments of a firm (such as production, marketing, finance, to mention a few). Since financial managers need to keep under control the asset allocation process so as to grant the maximum value to their shareholders, it can be said confidently that one of the most important corporate finance functions is the economic valuation of investment projects. In fact, the first step in undertaking any allocation project is the analysis of its economic and financial profile. The latter aspect is taken into account in drawing up the financial plan, where the financial manager verifies the compatibility of the cash flows related to the project under consideration with those that happen at the overall company's level.

The financial manager must carry out those projects that maximize the incremental value for the company's shareholders. In order to increase shareholders' wealth, a project must yield more than the cost of funds used to undertake the capital expenditure. It is, then, possible to conclude that the financial manager must judge a certain investment project according to its expected return. If this is higher than the company's cost of capital, the specific investment will create value for the company's shareholders.

As the asset allocation process is led on a continual basis, the financial manager must measure constantly the profitability of the firm's investment activity. Such a form of control makes use of appropriate methods of analysis like the NPV and the IRR. The NPV (net present value) is the present value of the cash lows stemming from the investment project; the IRR (internal rate of return) is that rate that makes the NPV equal to zero and it can be seen as the gross return of an investment project.

The implementation of the above mentioned tools of financial control on investments prevents the funds being misallocated. In fact, when many alternatives of investment are evaluated, the most profitable will be chosen first, then the others one by one according to their expected rate of return. This *modus operandi* becomes a necessary approach if the company has a limited amount of funds available to finance investment projects (*capital rationing*).

Capital Structure Analysis

Capital structure refers to how the firm's assets are financed. If the monetary resources that the current activity is able to produce are not enough to offset the cash-outs due to the company's investment activity, the firm needs to raise new funds from investors. Accordingly, capital structure is concerned with explicitly costly forms of financing just as financial debts (bonds, loans, overdrafts) and equity are. In fact, both debt holders and shareholders want to be remunerated for the money they invested into the company. It is then possible to determine the cost of capital as the average of the net cost of debt and equity (WACC). To calculate the average, the impact of debts and equity on the total capital structure has to be taken into account.

The composition of capital structure influences the company's cost of capital from two different points of view. First, any change in the amount of debts and equity will make their weights vary so that we come up with a new WACC, other things being equal. Second, the cost of capital will differently change according to which form of financing is chosen. Since financial charges are tax deductible whereas dividends and retained earnings are not, an increase in debts will reduce the WACC. However, it must not be concluded that a company should be 100% debt financed. In fact, the more debts rise the higher the risk of financial distress will be. This higher risk perceived by investors leads debt holders and stakeholder to ask for being granted a higher rate of return.

The composition of capital structure impacts on the shareholders' value since it influences the cost of capital invested in

the company. If the return on investments remains equal, a rise in the cost of capital will cause a decrease in the company's profitability and in shareholders' wealth. The financial manager has to take into consideration the effect on the capital structure when any financing decision is evaluated. Once a financial need arises from the planning activity, the financial manager should simulate what impact a debt or equity issue may have on the overall company. This activity enables the financial manager to quantify the effects of financing alternatives in terms of the firm's cost of capital and financial riskiness.

We are assuming that there are no capital constraints that affect the company so that the firm can count on an unlimited access to both the equity market and the debt market. According to this hypothesis, the financial manager can choose the financing alternative that maximizes the shareholders' value and minimizes the cost of capital. Sometimes, however, the firm has to face with a limited availability of financing alternatives to raise money from investors. In fact, many factors such as capital markets features or aspects related to companies' ownership and control could make some financing decisions unfeasible. At the same time, other aspects such as corporate taxation can make some sources of financing preferable in comparison to others.

See also: Corporate Governance; Finance, Geography of; Organizational Decision Making.

Bibliography

Bernstein, L.A., 1978. The Analysis of Financial Statements. Dow Jones-Irwin, Homewood, IL.

Brealey, R., Myers, S. Principles of Corporate Finance, fifth ed. McGraw-Hill Irwin, New York.

Broke, M.Z., 1990. Handbook of International Financial Management. Macmillan.

Cheng, F.L., 1985. Financial Analysis and Planning. Addison-Wesley, Reading, MA.

Dallocchio, M., 1996. Corporate Finance. Managing Corporate Decisions. Egea, Milan, Italy.

Modigliani, F., Miller, M.H., 1958. The cost of capital, corporate finance and the theory of investment. American Economic Review 48 (3), 261–297.

Modigliani, F., Miller, M.H., 1963. Corporate income taxes and the cost of capital: a correction. American Economic Review 53 (3), 433–443.

Ross, S.A., Westerfield, R.W., Jaffe, J., 1996. Corporate Finance. Irwin.

Solomon, E., 1963. The Theory of Financial Management. Columbia University Press, New York.

Van Horne, J.C., 1989. Financial Management and Policy. Prentice Hall, Englewood Cliffs, NJ.

Weston, J.F., Thomas, T.E. Managerial Finance, ninth ed. Dryden Publisher, Orlando, FL.

Weston, J.F., Thomas, T.E., 1992. Financial Theory and Corporate Policy. Addison-Wesley, Reading, MA.

Corporate Governance

Brett H McDonnell, University of Minnesota Law School, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.M. Blair, volume 4, pp. 2797–2803, © 2001, Elsevier Ltd.

Abstract

'Corporate governance' rules define decisionmaking roles among various corporate stakeholder groups. There is disagreement as to the goals of corporate governance, with some advocating an exclusive focus on shareholder value while others advocate including the interests of other stakeholders such as creditors and employees. Major means used to advance the interests of shareholders and/or other stakeholders include the board of directors, shareholder voting, monitoring by various gatekeepers, fiduciary duty rules, and incentives created by various markets.

'Corporate governance' refers to the legal rules and institutional framework that structure decisionmaking within business corporations. Major groups, which may play roles within corporate governance include directors, officers, shareholders, creditors, and employees. Several bodies of law are implicated in corporate governance, above all corporate or company law, but also securities law as well as elements of tax, antitrust, contract, commercial, employment, and bankruptcy law. Several types of markets are also implicated, above all stock markets, but also other financial markets, managerial labor markets, and general labor markets.

Following a brief outline of the basic governance problem, this analysis of corporate governance begins by considering the various goals corporate governance may serve, seen through the light of a series of conflicts in the interests of different groups. It proceeds to consider the means of corporate governance, that is, various legal and nonlegal mechanisms, which help structure how decisions within corporations are made within the context of these competing interests. Corporate governance frameworks have not only major similarities across countries, but also significant differences. After considering these similarities and differences and a variety of recent developments, the analysis concludes with an overview of a longstanding debate over the extent to which different national frameworks are and should be converging over time.

The Governance Problem

In all but the simplest of businesses, the actions of a variety of types of persons must be coordinated: equity investors, employees, managers, creditors, suppliers, and customers, to name some of the most important. These persons could all coordinate their actions through simple spot markets, with constantly shifting prices for their products and services. However, in many cases people have found it useful to form a firm in which their actions are coordinated through the directions of managers (Coase, 1937).

The corporation has been the leading legal form used by business firms. It is not the only such form – partnerships also have a long history, and more recently a number of new legal forms have been developed, most importantly the limited liability company (Ribstein, 2009). Here, though, the focus is only on corporations. The equity investors in a corporation are called shareholders or stockholders. A corporation may have just one shareholder, or a small number of shareholders with no active market for their shares, or thousands of shareholders whose shares may be readily traded upon a public stock market. The practical governance problems a corporation is likely to face vary greatly with their ownership structure. In the late nineteenth century there began to emerge a type of corporation in which there were many shareholders who owned only a small fraction of the total shares, and no one person or group owned a controlling block of shares. In a seminal work, Berle and Means (1932) identified such corporations as creating a separation of ownership and control, and the problem of that separation has dominated American corporate governance ever since. However, such corporations remain rare in most countries, where the role of controlling shareholders is critical.

Corporate (or company) law provides the central legal framework for corporations. Most legal authority for running a corporation is vested in the board of directors. However, the board may, and in larger corporations almost certainly will, delegate that authority to a series of officers. Between them, the directors and officers are charged with management of a corporation. In doing their jobs, they must supervise the work done by the corporation's employees. They must also raise money through the issuance of shares or through various forms of debt, and then must manage relationships with the resulting shareholders and creditors. They must obtain needed goods and services from outside suppliers, and must find and satisfy customers who buy the goods or services, which the corporation produces. Thus, the actions of many different types of persons must be coordinated, and the interests of those different persons will often differ. The corporate governance framework provides both legal and nonlegal mechanisms by which these different groups are coordinated and their conflicts of interest managed. The traditional literature in corporate law and economics, following Berle and Means (1932), has conceived of this as an agency cost problem in which the temptation of managers to make decisions that benefit

themselves rather than the corporation must be constrained (Jensen and Meckling, 1976).

Goals

In whose interests is and should a corporation be run? Given the many decisions that must be made within a corporation, and the many groups involved, that question could have quite a complex set of answers. A long-running debate tends to oversimplify the question into two alternatives: shareholders or stakeholders. That is, either corporations should be run to maximize the return to its shareholders, or else it should balance the interest of a variety of the groups mentioned above. The shareholder primacy view has come to dominate in the United States and to some extent other English-speaking, common law countries, while the stakeholder view dominates elsewhere. However, a more fine-grained understanding of the conflicting possible goals underlying corporate governance flows from considering a number of significant potential conflicts between the interests of various groups involved in corporations (Kraakman et al., 2009).

Shareholders vs Managers

Shareholders elect the board of directors, but then have little other formal authority in running a corporation, although shareholders do vote on a few major decisions such as amendments to the corporate charter, mergers, and dissolution. The board may in turn delegate much of its authority to various officers. By managers, I refer to both directors and officers. As noted in the discussion of Berle and Means above, where there are many dispersed shareholders, it will often be the case that no one shareholder has the incentive and the means to monitor the behavior of the managers. The managers may then be able to make decisions that benefit themselves at the expense of shareholders generally. This is the problem of the separation of ownership and control that is the central concern of American law for public corporations.

Controlling Shareholders vs Minority Shareholders

In corporations where one shareholder or a small group of shareholders owns enough shares that it can effectively choose on its own who will be on the board, the controlling shareholder has effective control over the corporation through its domination of the board. Moreover, if the control block is large enough, even on those matters in which shareholders do have a vote, the controlling shareholder may control enough shares to gain approval of measures without the support of other shareholders. In such corporations, the controlling shareholder may have the means and the incentive to dictate decisions that benefit it at the expense of other shareholders. In most countries, this is the dominant problem of corporate law, as in most countries almost all corporations have a controlling shareholder. Even in the United States, this conflict plays a major role in some parts of corporate law, most notably the law governing squeezeouts of minority shareholders.

Note that for most companies, one or the other of these first two conflicts will be a major concern, but not both. If there is no controlling shareholder, then this second conflict is not a problem. If there is a controlling shareholder, then in most instances that shareholder will have both the incentive and the means to monitor managers enough to significantly reduce the conflict between shareholders and managers. That conflict will not disappear – monitoring is costly and hence imperfect – but in such corporations the conflict is likely to be limited and manageable enough that it will not be a major focus of legal concern. In some unfortunate types of companies, though, both conflicts may be simultaneously severe. An important instance may be many state-owned enterprises (SOEs) in China (Yang et al., 2011). Most SOEs today have minority shareholders, but the dominance of a branch of the government makes those shareholders highly vulnerable to actions against their interest. And yet, agency problems such as corruption within the state itself will often leave the managers of SOE with much leeway to pursue their personal interests.

Shareholders vs Creditors

The legal recognition of limited liability for the shareholders of corporations creates a conflict of interest between shareholders and creditors. If a corporation does not fare well and does not have enough income or assets to repay its debts, in general its shareholders will not be held personally liable for those debts. This gives shareholders an incentive to induce the corporation to take on too much risk. Should highly risky actions lead to high profits, the shareholders, as residual claimants, are the main beneficiaries, while should those risky actions cause the corporation to go bankrupt, shareholders' losses are limited and much of the burden is borne by creditors. Creditors, in contrast, want the corporation to take on quite small degrees of risk – they care only that the business earn enough to pay them off, amounts earned beyond that do not benefit them.

Many mechanisms exist to address this conflict. Many of these are nonlegal self-help by creditors, such as loan covenants, monitoring before extending credit, staged credit, higher interest rates to compensate for credit risk, and so on. There are also many legal rules that exist to help protect creditors. Some of these are part of corporate law, including legal capital rules and the doctrine of piercing the corporate veil. Others are part of other areas of the law, including much of bankruptcy law, the law of secured transactions, and fraudulent conveyance statutes.

Shareholders vs Employees

Employees are the persons who do most of the actual work to carry out the actions of a corporation. Insofar as they are supposed to be advancing the ends of shareholders or of the other constituents, there is thus a classic agency problem, as employees may be able to act in ways that benefit themselves but not shareholders or the corporation. The risk of opportunism runs both ways, though. Especially insofar as employees have acquired firm-specific human capital, they will be interested in the long-run success of the business, as their alternative sources of employment will often be less attractive. Risks that look worthwhile to diversified shareholders with

limited liability may look much less so to employees whose livelihood is closely tied to the continuation of a particular business.

As noted above, American corporate law does little to either constrain or protect employees, outside of basic agency law. Other areas of law, especially labor, employment, and contract law, serve those purposes. However, employees are a significant part of corporate law in some countries, most notably Germany.

Other Stakeholders

The above four sets of conflicting interests set forth the constituency groups that play the biggest role in corporate governance. Yet, other groups may in some circumstances be relevant as well. A majority of states have enacted corporate constituency statutes, which allow boards to take into account the interests of groups other than just shareholders. In addition to creditors and employees as mentioned above, the other groups most frequently cited in constituency statutes include suppliers, customers, and the communities in which corporations are located.

Note that noncorporate business forms, particularly various forms of cooperatives, may make some constituency group other than shareholders, in effect the owners of the business. Constituent groups so empowered in cooperatives may include employees, suppliers, and customers.

Means

The previous section set out the main groups whose interests may be protected or balanced as a part of the system of corporate governance. This section considers major legal and nonlegal mechanisms by which those interests are protected or balanced.

The Board of Directors

The board of directors is the core locus of authority within corporations as a matter of law. Some justify this as the best way of protecting the interests of shareholders, who elect the board (Bainbridge, 2003), while others justify it as a way of balancing the interests of various constituencies (Blair and Stout, 1999). In practice, the traditional role of boards in public corporations was to provide strategic guidance and advice. In recent decades, there has been an increased focus on boards as monitors of the behavior of officers.

Along with the contemporary focus on the monitoring role of boards, have come the new best practices and then legal requirements that a majority of directors be independent of the corporation, aside from their role as directors. Exchange and securities rules specify who counts as independent, and also require committee to carry out various functions, including audit, compensation, and nomination committees. Controversy surrounds this move to a focus on monitoring and independence, with some arguing that boards have largely shown themselves unable to deliver on these new expectations (Fairfax, 2010).

Voting

The power of shareholders to direct corporate affairs through voting is limited. Shareholders elect directors (exclusively, outside of the codetermination context), and as noted above they vote on certain fundamental matters. But at least in the United States, shareholders vote on matters such as charter amendments, mergers, and dissolution only after the board has first voted in favor of a proposal. Thus, shareholders cannot initiate changes (although shareholders can act on some of these items in some countries without board approval). Moreover, traditionally the power to elect directors has been of limited use in public corporations, since most shareholders do not attend shareholder meetings but instead vote by proxy, and generally only the corporation itself finds it financially worthwhile to distribute proxy material. Thus, shareholders have traditionally only been able to vote on the slate presented to them by the board. The main context in which the shareholder's power to vote for the board has had value (in public corporations) is when there is a hostile takeover, so that bidders can acquire a controlling share and thereby take control. However, boards have been able to find highly effective defenses against takeovers, and for the most part American law has allowed those defenses (British rules have been less accommodating of takeover defenses).

More recently, certain institutional shareholders with an interest in active corporate governance have attempted to expand their ability to influence corporate behavior through voting. Their efforts are discussed below in the Recent Developments section.

Fiduciary Duty, Antifraud Rules, and Shareholder Suits

Shareholder suits against directors, and more rarely officers, are one of the major legal tools used to prevent and discipline misbehavior. Such suits are brought under both state and federal law. State law suits are brought to enforce the fiduciary duties that corporate law imposes upon officers and directors. Traditionally there were two main duties, the duty of loyalty and the duty of care. The duty of loyalty forbids managers from taking actions in which their personal interests differ from those of the corporation. The duty of care requires managers to gather adequate information before making decisions. The business judgment rule protects managers in care suits and makes liability extremely unlikely for such cases (Black et al., 2006).

In recent decades, state fiduciary duty cases have expanded beyond these traditional categories. Specialized standards of review have been developed where boards adopt antitakeover defenses, where they sell control of the corporation, and where they take action to limit the shareholder franchise. Delaware courts have also expounded upon the requirement to act in good faith as a way to review types of behavior that fall in between the categories of care and loyalty, notably including the decisions of boards to establish monitoring systems and the setting of executive compensation (Hill and McDonnell, 2012). Good faith is discussed further in the Recent Developments section below.

Federal law shareholder suits are brought under various antifraud rules of securities law, most commonly Rule 10b-5.

Federal courts made such suits easier to bring in the 1960s and early 1970s, but then started cutting back on this trend through a variety of opinions. Congress further tightened liability standards with the Private Securities Litigation Reform Act of 1995. However, shareholder litigation under both state and federal law still remains robust (Thomas and Thompson, 2012).

Gatekeepers

A variety of professionals function as corporate 'gatekeepers.' Gatekeepers monitor and collect information about a corporation and certify that information to investors and other outsiders. As such, they help reduce the problems of asymmetric information, which lie at the heart of the agency problems that are at the heart of the corporate governance problem (Coffee, 2006). Significant gatekeepers include auditors and accountants, lawyers, credit rating agencies, directors and officers' insurers, research analysts, and investment banks.

A variety of legal and nonlegal mechanisms work to give gatekeepers incentives to diligently gather and honestly disseminate information, although those mechanisms do not work perfectly. A leading nonlegal mechanism is reputation: being associated with a particular gatekeeper is more valuable if that gatekeeper has a good reputation for providing accurate information. Such a reputation allows gatekeepers to charge higher fees, thereby giving them incentive to continue to be honest and diligent. However, both internal agency costs within gatekeepers and potential one-time big payouts from dishonesty may undermine reputational incentives. Reputation and other nonlegal mechanisms are supplemented by various legal rules, which impose regulations on gatekeepers. These legal rules have evolved significantly in recent years, as discussed below in the Recent Developments section.

Markets

A variety of markets help shape patterns of corporate governance. Most crucial are stock markets. A classic element of corporate governance is 'the Wall Street Rule.' If shareholders are unhappy with the performance of a company with an active market for its shares, they can always sell their shares. This Wall Street Rule undermines some other governance mechanisms, notably shareholder voting – shareholders have less reason to engage in costly voting and monitoring if they can simply sell their shares when unhappy. However, the Wall Street Rule itself helps to discipline managers. If shareholder's selling leads to a lower stock price, a corporation will find it more costly to raise funds through new issuances. Perhaps more significantly, lowered share prices hurt managers personally insofar as a large part of their compensation is received through shares and options, as is true for most corporations today. A low share price also makes a corporation more vulnerable to a hostile takeover.

Other markets also help discipline managers. Bond markets are a more important source of funding than stocks for public corporations, so that variations in the price of a corporation's bonds can crucially affect its costs of capital. Various contractual and market features in credit markets help discipline managers (Whitehead, 2012). The managerial labor market also affects corporate governance: for managers who seek more prestigious and higher paying future jobs, earning a good reputation is important. Labor markets for lower level employees also play a role: a corporation with a reputation as a good employer will be able to attract new employees at lower wages. Product markets also play a role: a well-run company will be able to offer quality good or services at lower prices, and as a result improve its profitability, while poorly run companies ultimately face the threat of bankruptcy as customers move to competitors.

A key underlying element in debates over corporate governance concerns how well these various markets tend to function as disciplinary devices for corporate management. Those who argue for relatively light legal regulation of corporate governance believe these markets do a good job, and hence regulation should be enabling and limited (Easterbrook and Fischel, 1991; Romano, 1993). Those who believe in more extensive and mandatory governance regulation see greater imperfections in these markets, although they do not deny the value of them as elements of corporate governance (Bebchuk, 1989).

International Differences

Corporate governance structures and practices as well as laws differ systematically between countries. A major factor driving these differences is variations in typical shareholding patterns, although these differences in shareholding patterns in turn are reinforced by various practices and laws.

Anglo-American

As noted above, large public corporations with dispersed shareholders and no controlling shareholder became more widespread and dominant in the United States than in any other country. Such corporations are also common in the United Kingdom, and to a lesser extent also exist in Australia and Canada. Such corporations face the conflict between managers and shareholders generally to a more intense degree than other corporations, but do not face a conflict between controlling and minority shareholders. Corporate governance practices and laws in Anglo-American jurisdictions are thus more adapted to the former conflict and less to the latter than the laws of other jurisdictions.

Some believe that there is also a systematic difference between corporate governance in countries with a common law tradition as opposed to those with a civil law tradition. A large literature on law and finance has grown up exploring this hypothesis. This literature hypothesizes that countries with a common law tradition tend to provide stronger protections for investors. This protection leads to deeper and more developed financial markets, which in turn encourages stronger economic growth (La Porta et al., 2008). The theory is controversial (Armour et al., 2009).

Europe and Japan

Corporate governance in continental Europe and Japan is the flip side of the Anglo-American countries. Controlling shareholders exist in most companies, and most countries have

a civil law background. Two other characteristics are particularly noteworthy for the two leading countries in this category, Germany and Japan. First, banks play an important role in corporate governance, through share ownership, lending, and board membership. Second, employees have greater influence in corporate governance, either formally through codetermination in Germany or informally through the influence of senior management in Japan (Aoki, 1987).

Emerging Economies

With strong economic growth in emerging economies, especially China and India, corporate governance in those economies is of growing importance. The basic legal rules for the most part resemble those in more developed economies, although in many emerging economies the quality of court and agency enforcement is questionable. The general tendency is for controlling shareholders to be able to exploit their positions more fully in such economies, as both legal and nonlegal mechanisms provide weaker protections (Dyck and Zingales, 2004).

Corporate governance in China is of particular interest, both because of China's great and growing economic power and because of the still large role of the state (Yang et al., 2011). Many of China's largest corporations are SOEs, with a branch of the state as the controlling shareholder. Corporate governance challenges are heightened in such SOEs, as noted above. Even in nominally private corporations, informal state influence is strong, as the rule of law remains weak and patronage from state officials is important to protect companies from potential predation.

Recent Developments

In the last decade or two there have been a number of major developments in corporate governance, both in the United States and internationally. These developments affect all of the major means of corporate governance discussed above in Means section.

Institutional Shareholder Activism

A growing percentage of the shares of American public corporations are owned by institutions rather than individuals. Some types of institutions have engaged in forms of shareholder activism aimed at influencing corporate governance without trying to gain actual control of a corporation as in the older phenomenon of hostile takeovers. Union and public employee pension funds have been the traditional sources of such shareholder activism, although more recently some hedge funds have also become participants.

Activist shareholders have engaged in two basic kinds of formal campaigns, along with more informal attempts to communicate their opinions to managers. One kind of activism uses shareholder proposals under Rule 14a-8 to attempt to get shareholders collectively to express their opinions on a topic. Traditionally 14a-8 proposals have been nonbinding, although more recently shareholders have proposed bylaw amendments which if passed would be binding upon management. Shareholder proposals have covered a range of topics, from staggered boards to majority voting for directors to executive compensation to proxy access, among others. The other kind of activism attempts to elect shareholder nominees to the board. A major recent legal debate has concerned whether corporations must include shareholder nominees within the corporation's proxy material (so-called proxy access). The US law has not traditionally required proxy access. In 2010, the Securities and Exchange Commission (SEC) enacted a rule requiring proxy access, but it was struck down by the D.C. Circuit Court of Appeals. Currently proxy access is not required, although shareholders may propose bylaws that would require proxy access within a particular corporation. The Dodd-Frank Act added a new limited shareholder voting power: shareholders now have a nonbinding vote on the compensation of the top officers of a corporation. There is much debate over the effect and desirability of these forms of shareholder activism (Ferri, 2012).

Independent Directors

As noted in Section The Board of Directors, both exchange rules and securities laws impose independence requirement on directors of American public corporations. These requirements are recent developments. The exact definition of independence varies somewhat for each rule, but all require that a director not have been employed by the corporation within a certain period of time, and limit other financial ties to the corporation. These rules have been enacted in response to recent financial crises. The first sets of independence requirements were enacted in response to the dotcom crash of 2000. In 2002, the Sarbanes-Oxley Act required that boards have an audit committee which oversees the audit process, and that all members of the audit committee must be independent. In 2003, the New York Stock Exchange and Nasdaq imposed independence requirements for the directors of listed companies. These require that a majority of the board must be independent, and that companies must have compensation, audit, and nomination/governance committees composed entirely of independent directors. The last enactment requiring board independence is part of financial reform regulation enacted following the financial crisis of 2008. In 2010, the Dodd-Frank Act required that compensation committee members all must be independent.

The Rise and Fall of Good Faith

As mentioned in Section Fiduciary Duty, Antifraud Rules, and Shareholder Suits, Delaware courts over the past decade or so have paid increasing attention to the good faith requirement as an aspect of fiduciary duty. Several earlier developments encouraged plaintiffs to plead that defendants had acted in bad faith. If a plaintiff shows a decision or judgment was not in good faith, that decision loses the protection of the business judgment rule, which normally protects directors from liability. Moreover, behavior not in good faith is not covered by the exemption from personal liability that most corporations have adopted as allowed by section 102(b)(7) of the Delaware corporate law. Thus, if plaintiffs succeed in showing that behavior was not in good faith, they avoid two of the main obstacles that usually block them from holding directors

personally liable. Courts have thus had to define what is covered by 'good faith.'

In the *Disney* case concerning the compensation of Michael Ovitz, the court stated that "failure to act in good faith may be shown. . . where the fiduciary intentionally fails to act in the face of a known duty to act, demonstrating a conscious disregard for his duties" (*In re Walt Disney Co. Deriv. Litig.*, 960 A.2d 27, 67 (Del. 2006)). Courts have also used the good faith rubric to analyze a board's duty to monitor behavior by lower level employees. The intentionality requirement in the *Disney* standard has proven to impose a very high barrier to finding directors liable (Hill and McDonnell, 2012).

Gatekeepers

Section Gatekeepers noted a variety of gatekeepers who play an important role in corporate governance. American legal regulation of these gatekeepers evolved significantly in recent years. Regulation of gatekeepers played a major role in the main statutory responses to the two financial crises of the 2000s.

The Sarbanes-Oxley Act, the legal response to the dotcom crash of 2000, focuses above all on gatekeepers. Outside auditors receive the greatest attention. The Act created a new independent agency to regulate auditors, the Public Company Accounting Oversight Board. It also required audit partners to rotate after five years, and limited the kinds of nonaudit services auditors could provide their audit clients. Sarbanes-Oxley required the SEC to promulgate new rules requiring lawyers to report securities law violations up the ladder within the hierarchy of their clients. Sarbanes-Oxley also led the SEC to promulgate Regulation AC, under which research analysts must disclose compensation received in connection with their recommendations.

The other, greater crisis of the 2000s was the 2008 financial crisis, and the legislative response was the Dodd-Frank Act. The main new gatekeeper regulation in Dodd-Frank concerned credit rating agencies, containing many relevant provisions. Administrative agencies are required to remove references to credit ratings from a variety of regulations. Rating agencies are now potentially more exposed to liability for mistakes in ratings. There are new internal control requirements, and a variety of new disclosure requirements.

Federalization

Rules affecting corporate governance in the United States have long been made at both the federal and the state law. However, the core laws creating and governing the corporate law, the basic corporation laws, are set at the state level. Corporations can choose to incorporate in any state, no matter how limited their economic involvement in that state. This has led to a long debate as to whether the competition for corporate charters has led to a race to the bottom (Cary, 1974) or to the top (Winter, 1977).

Until recently, federal rules affecting corporate governance were mostly limited to disclosure and antifraud rules in certain contexts surrounding the issuance and trading of securities. One longstanding exception to this limitation is the proxy voting rules, which have for decades played a major role in structuring shareholder voting procedure for public corporations. However, federal involvement in corporate governance became more broad and widespread in the 2000s with the Sarbanes-Oxley and Dodd-Frank Acts. The previous subsections detail how those acts have affected board independence, shareholder voting, and the regulation of gatekeepers.

OECD Principles

Internationally, one of the most salient recent developments has been the promulgation in 1999 and subsequent spread of the Organization for Economic Cooperation and Development (OECD) Principles of Corporate Governance (OECD, 2004). These principles contain broad standards which both companies and countries are expected to strive to meet. These standards are divided into six categories: ensuring the basic for an effective corporate governance framework, the rights of shareholders, the equitable treatment of shareholders, the role of stakeholders, disclosure and transparency, and the responsibilities of the board.

Convergence

A longstanding debate asks whether the corporate governance systems of different countries are tending to converge toward each other, and whether they should converge. Over time, different countries have been seen as providing the best model to which others do or should aspire. In the 1980s it was Japan; in the 1990s, it was the United States. Some have seen signs of significant convergence, either in legal form or in underlying functional reality (Kraakman et al., 2009). Others are more skeptical of the reality of convergence, or whether the question is even well-posed in the literature (Clarke, 2010). There is much debate as to how to measure the performance of different corporate governance systems. The debate reveals much normative disagreement as to what one should be trying to accomplish through corporate governance, as well as empirical disagreement as to what different systems actually do accomplish. It remains a mystery as to where the corporate governance systems of different companies and countries are headed.

See also: Agency Theory; Corporate Culture; Corporate Finance: Financial Control; Management: General; Stockholders' Ownership and Control.

Bibliography

Aoki, M., 1987. Information, Incentives, and Bargaining in the Japanese Economy. Cambridge, New York.

Armour, J., Deakin, S., Mollica, V., Siems, M., 2009. Law and Financial Development: What We Are Learning from Time-Series Evidence. Brigham Young University Law Review, pp. 1435–1500.

Bainbridge, S.M., 2003. Director Primacy: The Means and Ends of Corporate Governance. Northwestern University Law Review 97, 547–606.

Bebchuk, L.A., 1989. Limiting contractual freedom in corporate law: the desirable constraints on charter amendments. Harvard Law Review 102, 1820–1860.

Black, B., Cheffins, B., Klausner, M., 2006. Outside director liability. Stanford Law Review 58, 1055–1159.

Blair, M.M., Stout, L.A., 1999. A team production theory of corporate law. Virginia Law Review 85, 247–328.
Berle Jr., A.A., Means, G.C., 1932. The Modern Corporation and Private Property. Commerce Clearing House, New York.
Cary, W.L., 1974. Federalism and corporate law: reflections upon Delaware. Yale Law Journal 83, 663–705.
Clarke, D.C., 2010. Nothing but wind? the past and future of comparative corporate governance. American Journal of Comparative Law 59, 75–110.
Coase, R.H., 1937. The nature of the firm. Economica 4, 386–405.
Coffee Jr, J.C., 2006. Gatekeepers: The Role of the Professions and Corporate Governance. Oxford, New York.
Dyck, A., Zingales, L., 2004. Private benefits of control: an international comparison. Journal of Finance 59, 537.
Easterbrook, F., Fischel, D., 1991. The Economic Structure of Corporate Law. Harvard, Boston, MA.
Fairfax, L.M., 2010. The uneasy case for the inside director. Iowa Law Review 96, 127–193.
Ferri, F., 2012. 'Low-cost' shareholder activism: a review of the evidence. In: Hill, C.A., McDonnell, B.H. (Eds.), Research Handbook on the Economics of Corporate Law. Edward Elgar, Northampton, MA, pp. 192–215.
Hill, C.A., McDonnell, B.H., 2012. Fiduciary duties: the emerging jurisprudence. In: Hill, C.A., McDonnell, B.H. (Eds.), Research Handbook on the Economics of Corporate Law. Edward Elgar, Northampton, MA, pp. 133–151.
Jensen, M.C., Meckling, M.H., 1976. Theory of the firm: managerial behavior, agency costs and ownership structure. Journal of Financial Economics 3, 305–359.
Kraakman, R., Armour, J., Davies, P., Enriques, L., Hansmann, H.B., Hertig, G., Hopt, K.J., Kanda, H., Rock, E.B., 2009. The Anatomy of Corporate Law: A Comparative and Functional Approach, second ed. New York, Oxford.
La Porta, R., Lopez-de-Silanes, F., Shleifer, A., 2008. The economic consequences of legal origins. Journal of Economic Literature 46, 285–332.
OECD, 2004. OECD Principles of Corporate Governance.
Ribstein, L.E., 2009. The Rise of the Uncorporation. Oxford, New York.
Romano, R., 1993. The Genius of American Corporate Law. AEI Press, New York.
Thomas, R.S., Thompson, R.B., 2012. Empirical studies of representative litigation. In: Hill, C.A., McDonnell, B.H. (Eds.), Research Handbook on the Economics of Corporate Law. Edward Elgar, Northampton, MA, pp. 152–169.
Whitehead, C.K., 2012. Creditors and debt governance. In: Hill, C.A., McDonnell, B.H. (Eds.), Research Handbook on the Economics of Corporate Law. Edward Elgar, Northampton, MA, pp. 68–84.
Winter Jr, R.K., 1977. State law, shareholder protection and the theory of the corporation. Journal of Legal Studies 6, 251–292.
Yang, J., Chi, J., Young, M., 2011. A review of corporate governance in China. Asian-Pacific Economic Literature 25, 15–28.

Corporate Law: Publicly Traded Corporations

Arthur R Pinto, Brooklyn Law School, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Marchetti and M. Ventoruzzo, volume 4, pp. 2803–2810, © 2001, Elsevier Ltd.

Abstract

The article focuses on the importance of the corporate business form and publicly traded corporations where shares of stock representing its ownership interests can be traded in stock markets. Corporate law primarily deals with protection of shareholders and the need for those who manage the corporation to be able to do so. The share ownership of publicly traded corporations in the United States and in other countries may be widely dispersed without any controlling group of shareholders or concentrated with control groups and the public shareholders in the minority. Each ownership structure had advantages and disadvantages, and there are different explanations of why these differing ownership structures developed around the world. In order to protect shareholders, different approaches have been suggested. In addition, there are differing views on the extent to which corporate law should also protect stakeholders other than shareholders.

Introduction

There is no standard definition for the term 'corporation'; the United States Supreme Court in *The Trustees of Dartmouth College v. Woodward*, 17 U.S. (4 Wheat.) 518, 636 (1819), described it as follows:

> A corporation is an artificial being, invisible, intangible, and existing only in contemplation of law. Being a mere creature of law, it possesses only those properties which the charter of its creation confers upon it, either expressly, or as incidental to its very existence.

Thus, a corporation is a business form that has a separate legal entity from those who own it ('shareholders'), and owes its existence to the state which allows for its formation and provides its attributes. As a separate legal entity there is a separate group called the board of directors elected by the shareholders to oversee the management of the business and select officers to run it (sometimes referred to as 'managers'). There are different types of corporations which have particular issues, rules, and legal issues depending upon the ownership structure. Two significant classifications are closely held and publicly traded corporations. In closely held corporations, there are few shareholders and the distinction between the shareholders and managers may be insignificant because they may substantially overlap. A major issue in corporate law for closely held corporations is to protect minority shareholders from controlling shareholders using their control to benefit them at the expense or exclusion of the minority. There are also publicly traded corporations with large numbers of shareholders whose shares therefore trade in stock markets. Even though the corporation is just one of several ways to structure a business (partnerships, limited liability companies, and sole proprietorships are other forms of business which far outnumber corporations in the United States), the economic impact of the corporate format is significant, since it is the form chosen by most large enterprises that are publicly traded corporations. In addition, the personal savings and retirement investments of many people are invested in these corporations.

In order to organize and create a corporation in the United States, the incorporator must choose the state in which to file the articles of incorporation. Corporate law can be viewed narrowly, as the state law governing corporations that are incorporated in a given state. State corporate law deals with the formation of the corporation, some financing issues, and its internal structure and governance. State corporate statutes are an important source of the corporate rules, but states also have a judicially created common law which creates corporate law that governs various aspects of corporate life. In a broader sense, corporate law can also include other sources of law and rules that are designed to protect investors in the corporation. For example, important sources of investor protection are federal securities law and stock market rules for publicly traded corporations (Pinto and Branson, 2013). Corporate law plays a significant role in dealing with corporate governance which has been described as "the system by which companies are directed and controlled" (OECD).

The article primarily deals with how corporate law allows for the corporate form, publicly traded corporations, and different ownership structures worldwide, the corporate law focus of shareholders and stakeholders, different theories of the firm and allocation of power, and how different ownership systems developed throughout the world.

Corporate Structure

Corporate law allows for the creation of a corporate legal entity separate from its owners and provides for significant corporate attributes. In addition to creation of a separate legal entity, other important attributes include limited liability, free transferability of interests, continuity of existence, and centralized management. As a legal entity, each of these other attributes flow from that entity. These attributes are common in almost all countries (Kraakman et al., 2009).

As a separate legal entity, there is limited liability since the corporation is able to incur its own liabilities, shielding the shareholders as owners from personal liability. Thus, the corporation's creditors cannot usually go after the personal

assets of the shareholders, and the creditors of the shareholders have no right to directly go after the assets of the corporation. This limited liability means that, in most cases, the shareholders will at most lose their investment in the corporation. The shielding advantages are the primary reason why the corporate form is chosen for many businesses.

There are times when a court will pierce the corporate veil removing the shield and allow creditors to pursue remedies directly against the shareholders. Those cases are not common and require some inequitable conduct by those in control who neglect the separate corporate existence. An example would be, if the corporation is unable to pay its debts, the controlling shareholder put insufficient ownership funds into the corporation (thin capitalization), and used corporate assets for personal gain.

Free transferability follows since shareholders are separate from the corporation and they can freely transfer shares to different owners without having an effect on the separate corporate entity. There is continuity because any change in the share ownership of the corporation has no legal effect on the corporation's existence, providing corporations with a perpetual existence. For example, if a shareholder dies the corporation continues.

As a legal entity, the governance can be centralized and separated from the owners. The governing structure of a corporation is composed of the shareholders as the owners of the company, and the board of directors who oversee the management of the company. This legal model allocates to directors and officers the authority to manage while it provides the shareholders, as owners, with some ability to monitor the managers' performance. The board of directors is usually elected by the shareholders at an annual shareholder meeting. In the corporate context, the corporation, as a separate legal entity from its owners, has a board of directors that need not be owners who manage the business.

Corporate law establishes these characteristics but generally each may also be modified by contract among the interested parties and that usually happens in a closely held corporate context. For example, although the corporate form provides limited liability, creditors may insist on personal guarantees from the shareholders to induce the creditors to loan to the business, thus, eliminating by contract shareholders' limited liability.

All of these above attributes are necessary in order to raise large amounts of ownership capital from many shareholders which facilitates the publicly trading of shares and creation of publicly traded corporations. Limited liability attracts passive investors because the most they have at stake is their investment. Free transferability of shares, perpetual corporate existence, and centralized management in a board of directors also allow for the easy trading of shares in stock markets (Easterbrook and Fischel, 1985).

Publicly Traded Corporations and Share Ownership

As closely held corporations grow, they often require substantial capital. To attract capital, a business may have several options. It can attract capital by finding more private investors or having another established business invest in the business. Another alternative is for the business to sell its shares more widely to the public, in an initial public offering, known as an 'IPO,' which is regulated by federal securities laws. The money raised in going public can be used to fund expansion of the business, to pay back the initial investors or its employees who have been given shares. With shares trading the corporation can use shares to buy other businesses by paying with shares or offer stock options (the right to buy shares at a fixed price) to compensate employees and managers. Thus, the option of going public provides flexibility to businesses to raise capital and use its shares for business purposes, and allows founders, employees, and private investors the ability to cash in on their investments.

Once a corporation completes a public offering, the shares are publicly traded in a stock market where investors can continually trade their shares of the corporation. These corporations are subject to stock market rules and federal securities regulation with its emphasis on providing full disclosure to the markets and investors. Most of the well-known large businesses in the United States are publicly traded corporations, i.e., the shares are held by a large number of shareholders and the shares trade in a stock market. The public stock markets provide investors with liquidity because it offers them the ability to sell their investment more easily in the market and thus lessens their risk. Without liquidity, corporations would have to compensate investors for that risk. This liquidity encourages investment and may be beneficial to corporations by lowering their costs of raising capital.

Public markets also provide numerous investment choices, allowing the public to purchase shares in different companies. This makes diversification of investment possible which is beneficial because investors reduce their risks by buying a portfolio of different investments and not owning a large percentage of shares of any one corporation. In most large publicly traded corporations in the United States, there are a large number of widely dispersed shareholders which results in no large or controlling shareholder(s). While directors are elected by the shareholders, the shareholders of many publicly controlled corporations are passive. These shareholders are unlikely to want to incur expenses related to monitoring managers if they do not own a significant amount of shares in their corporation because the cost of activism may be greater than the benefits. But even if there are benefits from activism, a rational profit maximizer would rather let someone else enforce rights on its behalf and thus 'free ride.' Free riding is a problem because it raises an issue of fairness. This problem – where it would be beneficial to act but costs limit such actions – has been described as a collective action issue (Clark, 1986).

Without a concentration of ownership in shares and shareholder passivity, managers (the board and officers) who control corporate assets, information, and the voting mechanisms are in *de facto* control of the corporation with little oversight by the owners, i.e., the shareholders. Thus, these publicly traded corporations are described as experiencing a 'separation of ownership from control.' Because of the vacuum created by the separation of ownership from control, management of many large corporations has become self-perpetuating. This separation of ownership and control can be beneficial when the managers operate the business in ways

that benefit the shareholders. Managers serve as specialists who use their expertise to increase the value of the firm while the shareholders are passive investors who are diversified and supply large amounts of capital, seeking the gains from increases in the value of the business. But separation of ownership from control with dispersed shareholders can also be harmful. Managers can mismanage or act to benefit themselves at the expense of shareholders (Berle and Means, 1932).

While dispersed ownership and the separation of ownership from control has been a significant influence on corporate law development, the US shareholder story is more complex. There are a number of publicly traded corporations in which there is concentrated ownership with a control group, and the public are minority shareholders. These corporations range from traditional, family-founded businesses (e.g., The New York Times) to some high-tech companies (e.g., Google) where control remains with the original owners. It may be more accurate to describe the US publicly traded corporation as a mixture of both dispersed and concentrated owners.

The original story of separation of ownership from control involved individual investors as shareholders. Individual shareholders have historically played a role in the development of corporate law. However, the rise in the number and ownership of large institutional investors has also been important over the last several decades. These institutions consist of private and public pension funds, insurance companies, foundations, hedge funds, universities, brokerage firms, bank trust companies, and mutual funds. Thus, dispersed ownership now includes the presence of institutional ownership, owners that now hold a larger number of shares representing a big dollar investment (albeit still a small ownership percentage overall) compared to individuals. When these institutions act collectively, they can influence their corporations. In the past, shareholders disappointed with their investment would follow the 'Wall Street Rule' and sell their shares rather than try to influence corporate behavior, i.e., choosing exit over voice. Some institutional investors (particularly public pension funds, labor union pension funds, and corporate governance hedge funds) have become more active in the corporations whose shares they own. In addition, institutional investors particularly from the United States have also been playing a significant role in other countries (Pinto, 2009).

Outside of the United States and a few other countries such as the United Kingdom, Canada, and Australia, there are publicly traded corporations where the predominant ownership story is different. Generally, most of those publicly traded corporations have concentrated ownership where there are controlling shareholders while the publicly traded shareholders are in the minority. These controlling shareholders can be a family (many of which were the founding members), another corporation acting as parent company to a publicly traded subsidiary, or large financial institutions like banks or insurance companies. The presence of a large controlling shareholder raises concerns about favoring their interest over the public shareholders who are in the minority. In addition, some publicly traded corporations may have significant state ownership or government links which can mean political influences in business decisions. Just as dispersed ownership greatly influenced the corporate law of the United States, concentrated ownership influenced the corporate law of those countries.

Corporate Law and Other Stakeholders

A significant role of corporate law is the protections of investors in the corporation. Long-term financing of the business can be supplied by its owners or by creditors who lend their money to the business. If capital is lent, then the relationship between the lenders and the business creates a debtor–creditor relationship where the creditor is looking for eventual repayment plus some current return usually in the form of interest on the loan. If the capital is provided for an ownership stake, then the investors are willing to forgo a promise to be paid interest or a set return on the investment for the potential for sharing in the success of the business as owners. Their shareholder investment provides what is called equity for the business as opposed to debt. Because creditors generally have priority in payments for interest and principal, shareholders are at a greater risk but willing to undertake that risk because of the greater return if the business is successful. Those employed by the business can also be viewed as suppliers of human capital and also have a long-term interest in the success of the corporation. Large publicly traded corporations also have other important stakeholders including customers, suppliers, and the community (Licht, 2004).

Because of the importance of publicly traded corporations in society, there is a debate over the focus of corporate governance and the extent to which stakeholders other than shareholders should have a role in governance. The publicly held corporation can be viewed in purely economic terms as a means by which capital is raised from a large number of public savers and used by businesses. Under that focus, corporate governance may focus on the suppliers of capital (creditors and shareholders). While creditors to some extent are protected in corporate law – through disclosure and legal capital rules that focus on dividends and other distributions to shareholders that can harm them – creditors primarily are protected by their contract and bankruptcy law. Since shareholders are the owners and face greater risk than creditors, they are the primary focus of corporate law.

Given the economic significance of publicly traded corporations and their impact on society and other interests, a broader view has long been advocated to include protection of other stakeholders who do not supply capital to the business, e.g., labor (who invests human capital) (Blair and Stout, 1999). An internal perspective on such protection would place stakeholders directly into corporate governance. This could include representation on the board of directors, a voice in corporate decisions or requiring those who manage corporations to have legally enforced duties to protect the stakeholders. In the US, stakeholders are not usually represented on boards, which generally owe their primary duty to the shareholders who select the board of directors. Many countries with concentrated ownership lean toward a stakeholder model. In these countries, stakeholders are protected to a greater extent than in the United States and they have a direct role to play in corporate governance. Germany remains a prime example because of its dual board and co-determination where labor

serves on the supervisory board of publicly traded corporations. In Japan, maintaining long-term employment had been a major goal of corporations.

An external perspective sees stakeholders as outside the internal corporate governance. This may suggest that stakeholders' interests can be protected by the concept of corporate social responsibility which encourages consideration and a positive impact through its activities on the environment, consumers, employees, communities, and other stakeholders. But advocates of shareholder-gain as the sole motive for corporations believe that corporations are formed to run a business for a profit and other activities should not be within their focus. Allowing managers to pick and choose philanthropic or socially beneficial activities permits use of corporate funds for management's personal charities and enables the managers to act as a sort of unelected civil servant. In addition, profit is the yardstick used to measure the success of a business and ultimately, profit maximization is the most socially responsible activity (Friedman, 1970).

Advocates for corporate social responsibility recognize that large publicly traded corporations have considerable power and resources that play a significant role in the economy. Corporations, as part of the larger community, gain from philanthropic activities through enhanced reputation and helping to preserve the advantages that the law provides for businesses. Philanthropy should be encouraged and managers should look to protecting the whole enterprise and not just the shareholders. Reasonable amounts of corporate philanthropy are not only specifically authorized by state statutes and tax deductible but are common and significant.

Corporate social responsibility is fostered by self-regulation as opposed to mandatory rules and now appears to be a goal of many large publicly traded corporations throughout the world. The difficulty is determining the extent to which such policies and goals are real or a form of public relations.

Shareholder Protection and Theories of the Firm

A major focus of corporate law and corporate governance is the protection of shareholders but there are different views of how best to do so. With a shareholder perspective, financial economists and law and economic scholars have often viewed corporate governance and corporate law as an issue of agency costs. In any such relationship, the agent who acts on behalf of a principal does not always act in the best interests of the principal. Thus there are certain costs associated with the relationship called agency costs. Agency costs include monitoring expenditures (costs of supervising the agent), bonding expenditures (the agent attempts to assure the principal that the costs will be minimized), and residual loss (there will inevitably be some cost to the principal). From this analysis, it is generally beneficial to have a system that will reduce these agency costs to the lowest possible level. Corporate law and governance rules are often analyzed with this agency cost perspective using a cost/benefit analysis.

The issue of agency costs in the corporate context and protection of shareholders differs depending upon the ownership structure. With widely dispersed ownership, the shareholders can be conceptualized from a financial perspective as being the principal while the managers (directors and officers) are their agents acting on their behalf. Since shareholders in corporate law do not control managers, this is not a legal concept. The concerns are that the managers will act in ways to harm the shareholders either by mismanaging the business or by self-dealing, favoring their interests over the shareholders. The dispersion of ownership and shareholder passivity allows the managers to control the corporation even though they might own relatively few shares. In this context, issues arise as to how to lower the agency costs and potential losses as well as what bonding and monitoring mechanisms are best. In determining legal protection, there also needs to be a balance so that those who manage are able to take the risks necessary to promote the business (Jensen and Meckling, 1976).

With concentrated ownership, there is a group of shareholders constituting the majority who control the business while the minority are the public shareholders. Under the agency cost perspective, the controlling shareholders are viewed as the agents and public minority shareholders as principal. In this context, issues may arise concerning the control group's relationships if it engages in conflicts of interest transactions or policies that favor their interest over the minority. There are different views as how to best protect shareholders. In the US, this debate has produced different theories of the firm for publicly traded corporations and different views on how corporate law should allocate power.

The law and economics approach or contractual model recognizes the need for shareholder protection and favors shareholder wealth maximization. Under this approach, managers should run the business and abuse by managers would be controlled with market-based solutions as the optimal way of looking at the firm and for lowering agency costs. Markets present a more efficient (cost less and benefit more) means to monitor the agency relationship. For example, if managers are poorly managing the business the stock price would decline. The low share price may encourage shareholders to challenge directors through shareholder voting (called a proxy fight because in publicly traded corporations shareholders vote by executing proxies in advance of a meeting rather than attend) to replace directors. A low price could also induce a bidder to try to buy the corporation which if opposed by the managers is a hostile takeover. Therefore, management fearful of losing their positions will attempt to manage efficiently and in favor of shareholders to keep the share price high. This is an example of the market for corporate control providing a check on managers. In light of the monitoring of the markets, the law and economics model argues that corporate law should only facilitate market mechanisms or provide rules when markets fail and are unable to constrain managerial behavior (Easterbrook and Fischel, 1991).

Some law and economics theorists also conceptualize the corporation in terms of contract law. A corporation can be viewed as a nexus of contracts through which various claimants such as creditors, workers, shareholders, and consumers enter into agreements. Private contracts are viewed as an efficient means to lower costs in the agency relationship between the shareholders and managers. One can view the articles of incorporation and the bylaws as a contract between the shareholders and the managers setting out the rules governing their relationship. This private ordering through

contracts allows the parties to provide rules to maximize value and minimize costs. Under this view, corporate law should provide the basic terms of these contracts (i.e., default rules), but the shareholders and the managers should be allowed to change the terms, thus providing an optimal and mutually agreeable system (Easterbrook and Fischel, 1989). While many corporate law rules can be waived or modified because the statutes explicitly allow for it, critics of the contractual approach argue that some minimal protection and regulation is needed, and some of the rules should not be changed so easily. One of the most convincing arguments against all rules being subject to contractual modification is that in a publicly traded company there is no real bargaining over the terms of the contract. Management controls the structure and timing of the process while shareholders are widely dispersed, unable to act collectively (Clark, 1985). Although the contract theory has not been adopted for all corporate rules, it remains influential and does provide a framework for analyzing those rules. Courts and legislatures have taken the theory into account in fashioning rules. Generally speaking, in civil law countries one finds more mandatory rules where common law countries allow for more opting out of statutory legal rules (Pistor, 2002).

A regulatory approach seeks stronger state law protection of shareholders and an increased federal presence through an expansion of the reach of federal securities law. Given the failure of markets and passivity of shareholders, certain corporate rules must be made mandatory to assure minimal standards to protect shareholders. These rules can be enforced by government actors and private litigation by shareholders. There are also different approaches on how power should be allocated within the corporation. One is a management approach favoring laws that give managers (officers and inside directors) tremendous latitude in their activities. This model is premised on the idea that managers will protect the interests of shareholders because they have a mutual interest in protecting the corporation. Although managers do not own large numbers of shares, their livelihood depends upon a successful company. Thus, the shareholders and managers share the similar goal of corporate advancement. Directorial primacy and centralized decision making is the essential attribute of efficient corporate governance and discourages intershareholder opportunism. In addition, managers are often in the best position to consider the different demands on the corporation and are thus well suited to advance corporate interests and consider other constituencies such as labor and the community (Bainbridge, 2005). However, there are often conflicting ways in which managers seek corporate advancement. For example, some managers might see expansion or enlargement of the corporation as being in the company's best interest, but that may not be beneficial to the shareholders.

Concerns about the lack of management accountability have led some commentators to argue for increased shareholder powers and easier mechanisms to replace or influence managers. Shareholders as owners should be able to protect their investment and need to be given the power to do so especially to effectuate significant decisions (Bebchuk, 2005). But there are concerns that increased shareholder power will undermine the efficient management of corporations and even harm other stakeholders.

In a comparative context, the US has a managerial model where shareholders have less direct power over the directors compared to other countries where shareholders can implement and vote on more issues (Ruggiero and Bruno, 2011).

Dispersed versus Concentrated Ownership

As discussed, publicly traded corporations are often viewed through the lens of ownership, with the widely dispersed shareholder model used in the US and concentrated ownership generally found elsewhere. Important comparative corporate governance scholarship has focused on these different ownership structures with explanations of why these different systems developed and whether one system is more optimal. If one system is viewed as more advantageous, then the lesson on how and why each developed may be instructive for future developments of corporate law and economic development. A number of different explanations have tried to explicate the reasons for the different ownership patterns. These explanations can be broadly described as law matters, history and politics, private ordering, interest groups, and deep causation.

Law matters was based upon leading financial economists' empirical research comparing countries with dispersed and concentrated ownership. Their empirical evidence found significant differences in the legal protection of investors between countries. They initially told a story that the absence of law protecting shareholders in many countries explained the emergence of the different systems. Without strong legal protection, shareholders had little choice but to maintain control, which was one way to monitor managers and protect their investment. The more protective the law was to investors, the more likely the emergence of widely dispersed ownership. Thus concluding that the law determined the ownership structure, system of corporate finance, and governance. Generally it was common law legal systems that allows for judicial law creation and flexibility was more protective than the civil law. While some have questioned the studies as incomplete and limited, the *law matters* thesis set the agenda for much of the significant comparative corporate governance scholarship (La Porta et al., 1998, 1997).

The other explanations were responses to the *law matters* story. They all recognize a role for law but see other explanations for why dispersed ownership did or did not emerge. The *history and politics* explanation views politics as the important determinant of ownership structure. In the United States, political factors such as hostility to bigness in government and in business meant legal limits on large financial institutions which helped prevent the rise of strong owners (Roe, 1994). In Europe, social democracy in postwar Europe was a political outlook that promoted welfare of all citizens and prevented wide disparities, resulting in considerable government control in addressing social inequalities, such as unemployment. Given the ever present possibilities of government interference with an enterprise, owners enter into pacts with labor and government to maintain favorable social policies (often in return for protection of management interests) and limit market-based capitalism with dispersed owners despite the existence of market-friendly laws. Thus, *history and politics* was determinative in the development of particular ownership systems (Roe, 2003).

A *private ordering* explanation suggested that *history and politics* painted too broad a picture. Some countries, like the United Kingdom with a social democratic background, developed dispersed capitalism. What mattered was private ordering or contracting (e.g., stock market rules or role of private bankers and investors) which allowed optimal substitutes from formal law, and some countries, like the United Kingdom, developed systems of private ordering which provided investor protection that law followed to solidify or formalize activities created by the local interests. Under this view, law matters if it allows for private ordering and the common law is more adaptive and accepting of such ordering. Thus, private ordering and self-regulation can have the effect of constraining society (Coffee, 2001).

An *interest group* explanation suggested a different political story which related the issue of widely dispersed shareholders to the creation of liquid securities markets and to how open a country is to outside investment. While politics may be relevant, it is not social democracy but the protectionism of local interests that keeps markets closed from competition. Thus, the key to dispersed ownership, raising capital, and improved corporate governance is a market-based system and open competitive trade. This is because certain countries cut back on openness to finance because new financial entrants threaten established, local interests and can cause short-term disruption by supporting new firms that threaten the established order. The established companies and financial providers use the popular political opposition generated by this disruption strategically, e.g., to cut down on the access to financial capital by local companies. The country becomes worse off and declines, as local (particularly new) businesses lose competitively priced capital, but the established elites guard their privileged position (Rajan and Zingales, 2003).

In looking at explanations for the development of ownership structures, some have explored whether *deep causation* can help explain different ownership systems. Legal rules themselves are not a complete story and a societal explanation, which includes cultural factors and studies that correlate between these cultural factors and the law, is a fuller story. For example, while shareholder rights are higher in countries belonging to the English-speaking cultural region, these countries are no better than others in protecting creditors. This casts doubt on the alleged general superiority of statutes in common law countries for protecting investors. The implication for a deep causation approach is that legal change may not have a significant effect without cultural norms that promote investor protections (Licht, 2001).

Both dispersed and concentrated ownership systems have advantages and disadvantages. The widely dispersed model relies more on markets for financing and the importance of liquidity for investors. A deeper and more liquid capital market may be beneficial to corporations by lowering their cost of raising money. Further, publicly traded corporations provide disclosure, allow for investors to cash in on their investments which encourage the creation of new businesses, and allow for greater societal participation in ownership. But dispersed ownership allows for managerial abuse. In addition, publicly traded markets have the disadvantage of being volatile and at times may be subject to speculation and bubbles. Such concerns may lead to short termism by corporations. Concentrated ownership systems relies more on financial institutions such as banks for both finance and monitoring and can be viewed as more patient than markets for long-term planning. Concentrated ownership deals with the agency problem of the managers by providing close monitoring of management but there is a different agency problem of controlling shareholders taking advantage of the minority shareholders in publicly traded corporations (Maher and Andersson, 2000).

Later studies using the *law matters* thesis elaborated further on the relationship between ownership and corporate performance. Research suggested that increased legal protection and dispersed ownership is in fact beneficial compared to the concentrated ownership system. Increased legal protection encourages not only dispersed ownership but the related depth of the capital markets which can lower the cost of capital for corporations and foster economic development. Thus *law matters* not only in creating dispersed shareholders but for economic development. If concentrated ownership is a poor substitute for markets, the solution for both developed and developing countries may be a legal one, i.e., changing law to protect investors to encourage greater use of markets (La Porta et al., 2002). The other explanations such as politics and deep causation suggest changes may be more difficult. Corporate governance and law may be strongly 'path dependent': i.e., certain political forces push a country down a particular path, and it then becomes difficult for the country to pull itself out of the path when local interests and institutions develop connections to and affinities for the path (Bebchuk and Roe, 1999).

Globalization in trade, capital markets, and multinational corporations have had an impact on the different corporate governance systems. Increased trade increases competition which has the effect of breaking down local barriers resulting in increased competition in the products market. This open competition directly affects the performance of a country's firms. This raises the issue of whether a particular corporate governance model is a factor in a firm's performance. Another economic aspect of globalization is the internationalization of the capital markets. With the expanded markets resulting from trade, firms raise capital and investors invest globally. Competition for capital means that corporations seeking capital would need to provide good governance practices to attract investment. In addition, investors will demand such practices in order to invest. If certain corporate law or practices can attract capital more cheaply then this may provide an economic advantage in the firm's performance and a country's overall economic well-being (Pinto, 2005). In addition, there may be some convergence in the 'shareholders and stakeholders models' described earlier (Gordon and Roe, 2004).

See also: Business Law; Common Law; Contracts: Legal Perspectives; Corporate Governance; Law and Development; Law and Economics; Regulatory Agencies; Stockholders' Ownership and Control; Venture Capital.

Bibliography

Bainbridge, S., 2005. Director primacy and shareholder disempowerment. Harvard Law Review 119, 1735.

Bebchuk, L., Roe, M.J., 1999. A theory of path dependence in corporate ownership and governance. Stanford Law Review 52, 127.

Bebchuk, L., 2005. The case for increasing shareholder power. Harvard Law Review 118, 833.

Berle Jr., A.A., Means, G.C., 1932. The Modern Corporation and Private Property. Commerce Clearing House, Inc., New York.

Blair, M.M., Stout, L.A., 1999. A team production theory of corporate law. Virginia Law Review 85, 247–328.

Clark, R., 1985. Agency costs versus fiduciary duties. In: Pratt, Zeckhauser (Eds.), Principles and Agents: The Structure of Business. Harvard Business School Press, Boston.

Clark, R., 1986. Corporate Law. Little Brown, Boston.

Coffee, J.C., 2001. The rise of dispersed ownership: the role of law and the state in the separation of ownership from control. The Yale Law Journal 111, 1.

Easterbrook, F.H., Fischel, D.R., 1985. Limited liability and the corporation. The University of Chicago Law Review 89, 52.

Easterbrook, F.H., Fischel, D.R., 1989. The corporate contract. Columbia Law Review 89, 1416.

Easterbrook, F.H., Fischel, D.R., 1991. The Economic Structure of Corporate Law. Harvard University Press, Cambridge, MA.

Friedman, M., 1970. The Social Responsibility of Business is to Increase its Profits. The New York Times Company, New York.

Gordon, J., Roe, M., 2004. Convergence and Persistence in Corporate Governance. Cambridge University Press, Cambridge.

Jensen, M.C., Meckling, M.H., 1976. Theory of the firm: managerial behavior, agency cost and ownership structure. Journal of Financial Economics 3, 305–359.

Kraakman, R., et al., 2009. The Anatomy of Corporate Law: A Comparative and Functional Approach. Oxford University Press, UK.

La Porta, R., et al., 1998. Law and finance. Journal of Political Economy 106, 1113.

La Porta, R., et al., 1997. Legal determinants of external finance. Journal of Finance 52, 1131.

La Porta, R., et al., 2002. Investor protection and corporate valuation. Journal of Finance 57, 1147.

Licht, A., 2004. The maximands of corporate governance: a theory of values and cognitive style. Delaware Journal of Corporate Law 29, 649.

Licht, A., 2001. The mother of all path dependencies toward a cross cultural theory of corporate governance systems. Delaware Journal of Corporate Law 26, 147.

Maher, M., Andersson, T., 2000. Corporate governance: effects on firm performance and economic growth. In: Renneboog, L.J. (Ed.), Convergence and Diversity of Corporate Governance Regimes and Capital Markets at 411–412. Oxford University Press, UK.

Organisation for Economic Co-Operation and Development, 2004. Principles of Corporate Governance.

Pinto, A.R., Branson, D., 2013. Understanding Corporate Law. Lexis.

Pinto, A.R., 2005. Globalization and the study of comparative corporate governance. Wisconsin International Law Journal 23, 477.

Pinto, A.R., 2009. The European Union shareholder voting rights directive from an American perspective: some comparisons and observations. Fordham International Law Journal 32, 587.

Pistor, K., 2002. The evolution of corporate law: a cross-country comparison. University of Pennsylvania, Journal of International Economic Law 23, 79.

Rajan, R.G., Zingales, L., 2003. The great reversals: the politics of financial development in the twentieth century. Journal of Financial Economics 69, 5.

Roe, M.J., 1994. Strong Managers, Weak Owners: The Political Roots of American Corporate Finance. Princeton University Press, Princeton, NJ.

Roe, M.J., 2003. Political Determinants of Corporate Governance. Oxford University Press, UK.

Ruggiero, E., Bruno, S., 2011. The Public Companies and the Role of Shareholders: National Models towards Global Integration, co-ed. Kluwer Law International, The Netherlands.

Corporate Taxation

James R Hines, University of Michigan, Ann Arbor, MI, USA; and NBER, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Corporate taxation is an important source of government revenue around the world and a major consideration in planning business activities. This article reviews the economic considerations raised by corporate taxation and evidence of the behavioral impact of taxation.

Scope and History

Corporate tax obligations consist of both direct and indirect taxes. Direct taxes are typically expressed as fractions of corporate income, whereas indirect taxes include sales taxes, value added taxes, excise taxes, property taxes, payroll taxes, and import and export duties. Corporate indirect tax payments exclusive of payroll taxes commonly exceed their direct tax payments in overall magnitude (Desai et al., 2004a). Much of the interest in corporate taxation nevertheless turns on the direct tax component, since the indirect taxes such as value added taxes are imposed on all business activities, whether or not undertaken by corporations.

The corporate income tax rate structure is usually progressive, meaning that average tax rates rise with income, typically reaching a maximum rate rapidly enough that almost all of the income of large corporations is subject to tax at the highest rate. In the United States in 2014, corporations earning net income up to $50 000 paid a 15% tax; this tax rate rose to 25% on income above $50 000, 34% on income above $75 000, and was 35% on income above $10 million. Other countries tax corporate income at similar rates that can vary according to the size, location, and industry of the corporate taxpayer.

Much popular and public policy attention is devoted to the tax planning activities and corporate tax obligations of big businesses – and with good reason, since corporate income tends to be concentrated in a relatively small number of large companies. For example, in 2011 there were 5.8 million active corporations in the United States, of which 2831 had assets exceeding $2.5 billion. These 2831 large corporations accounted for 81% of total US corporate assets, 68% of net corporate income, and 68% of total corporate tax payments.

The American federal corporate income tax was introduced in 1894 but was found unconstitutional the following year; it reappeared as a gross receipts tax in 1909, and was modified to become a genuine income tax following ratification of the 16th amendment to the US constitution in 1913. The US corporate tax rate in 1913 was 1%. The corporate tax rate rose over time, though it remained below 15% until World War II, at which point it rose sharply. The corporate tax was a major revenue source for the US federal government between World War II and the late 1960s, when corporate tax collections consistently represented more than 20% of federal government revenue. Corporate tax collections have since fallen as a fraction of total government revenue, though federal corporate tax collections exceeded $220 billion for tax year 2011, when total corporate tax collections by all levels of US government were 9.4% of total government revenues, or 2.3% of US gross domestic product (GDP). In the same year, the 34 OECD countries collected corporate tax revenue that averaged 8.7% of total government revenues, or 3.0% of GDP. The US (federal plus average state) statutory corporate tax rate in 2013 was 39.1%, making it the highest of all the OECD countries, which averaged 25%, and which included five countries (Ireland, Slovenia, Poland, Hungary, and the Czech Republic) with statutory corporate tax rates below 20%.

Impact on Economic Behavior

The taxation of corporate income encourages entrepreneurs and managers to structure and conduct their business operations in ways designed to avoid taxes. Corporations generally reduce their tax obligations, and those of their shareholders, by using debt rather than equity finance, investing in assets that can be rapidly depreciated for tax purposes and those for which generous tax credits are available, and avoiding dividend payments or other tax-disadvantaged distributions to investors (Auerbach, 2002).

A major tax consideration in corporate finance is that interest payments to bondholders are deductible from taxable income, whereas dividend payments to corporate shareholders are not deductible. As a result, corporations generally have tax incentives to issue more debt than they would otherwise, effectively replacing equity finance with debt finance in order to reduce taxable corporate income. The magnitude of the incentive to use debt finance is an increasing function of interest rates and tax rates, and there is ample evidence (Graham, 1996; Desai et al., 2004b) that firms facing higher corporate tax rates (because they are located in countries with high corporate tax rates or because they have very low probabilities of incurring tax losses) have higher debt/asset ratios than other firms. Since dividend payments typically generate tax obligations for recipients, corporations also have incentives to buy back shares or use other nondividend methods of distributing income to shareholders, and while they persist in paying dividends, there is evidence that taxation of dividend receipts at the shareholder level discourages dividend payouts (Chetty and Saez, 2005).

Tax incentives to issue debt are mitigated in situations in which firms anticipate nonrefundable tax losses, or in which

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71013-0

there is strong investor preference for returns in the form of capital gains and dividends rather than interest income. The strength of investor preference for stock ownership depends on circumstances. Individual investors most strongly prefer owning a firm's stock to owning its bonds if they reside in countries that permit dividend recipients to claim credits for corporate taxes paid on income generating dividends, though most tax systems – primarily those in European countries – that once featured this type of corporate integration no longer do so. In countries without such personal and corporate tax integration, the favorable tax treatment of accruing capital gains means that some investors, typically those in the highest tax brackets, prefer stock ownership (despite the associated corporate tax obligations) to holding interest-bearing debt. With sufficient numbers of such investors, it follows that asset markets will price stocks and bonds so that the economy's debt/equity ratio increases with the corporate tax rate, while any individual corporation is indifferent between debt and equity finance (Miller, 1977).

Taxation influences the timing, magnitude, and composition of corporate investment in plant and equipment, inventories, research and development, and other business assets. Higher tax rates generally are associated with lower investment levels (Djankov et al., 2010), though this depends on tax treatment of investment expenditures. Investors generally do not reduce taxable income immediately by the full amount of spending on business investment, instead amortizing the cost of investment spending over a period of years in which 'depreciation allowances' are permitted. These depreciation allowances are most valuable if claimed soon after an investment is undertaken, since they are seldom adjusted for inflation and the time value of money. Hence governments that seek to encourage business capital formation often do so by permitting investors to depreciate their investments for tax purposes over short periods of time. Governments also often provide special tax credits for investments in selected asset categories such as research and development (available in many countries, including Australia, Brazil, Canada, China, France, India, Japan, Spain, the United Kingdom, and the United States). Depreciation allowances and investment credits together have the potential to stimulate investment by reducing required rates of return (Hall and Jorgenson, 1967). These tax considerations are succinctly captured by "effective tax rates" (Auerbach and Jorgenson, 1980) on business investment, which can be usefully compared over time and between countries (King and Fullerton, 1984). The available evidence indicates that rates of business investment are inversely related to effective tax rates (Chirinko et al., 1999; Hassett and Hubbard, 2002), though the effect of tax changes on investment levels is dampened by adjustment costs that firms incur when abruptly changing their capital stocks (Goolsbee, 1998; House and Shapiro, 2008).

Corporate taxation also affects business organization by discouraging the incorporation of profitable businesses that can be organized in noncorporate form. In the United States, business organizations whose income is not subject to the corporate income tax include small ('S') corporations, partnerships, sole proprietorships, and limited liability companies, and establishing a company with one of these organizational forms is an alternative to setting up a new corporation. There is evidence (Da Rin et al., 2011) that among European countries higher corporate tax rates are associated with reduced rates of forming new corporations.

The Taxation of Multinational Corporations

Multinational corporations – those with active business operations in more than one country – pose special problems for tax systems, since it is necessary to determine the location and character of taxable income, as well as the means by which double tax relief is to be provided. Double tax relief is essential because a corporation's home country claims the right to tax all of its income, including the income earned by its foreign affiliates, while host countries in which foreign affiliates are located insist on their rights to tax affiliates' incomes. In the absence of tax relief, income earned by foreign affiliates would be subject to taxation both by host countries and by home countries at cumulative rates that would often be prohibitive to business operations.

There are two practical systems of international double tax relief, both of which permit countries to tax fully any income earned by economic activity undertaken within their borders. Double taxation is avoided whenever home countries forego taxing multinational corporations on income earned by their foreign affiliates, or else when home countries tax the incomes of foreign affiliates but permit taxpayers to claim credits for foreign taxes paid. Use of the credit method effectively subjects income earned by foreign affiliates to home-country taxation at a rate equal to the difference between home and foreign tax rates. Actual methods of double tax relief often resemble a combination of these two methods. A majority of the world's countries exempt from tax most of the income earned by foreign affiliates of domestic multinational corporations, though 7 of the 34 OECD countries, including Ireland, Mexico, Poland, and the United States, subject such income to taxation but permit credits to be claimed for foreign taxes.

There is ample evidence that international tax rate differences influence the timing and location of investment by multinational corporations (Hines, 1996, 1999; Gordon and Hines, 2002). Countries and subnational jurisdictions that tax corporate income at low rates receive unusually large volumes of foreign direct investment, particularly from countries that do not tax foreign income. Furthermore, multinational firms arrange their business transactions to avoid international taxes. Thus, foreign affiliates located in high-tax countries tend to be financed with higher debt/equity ratios than are investments in low-tax countries (Desai et al., 2004b; Huizinga et al., 2008), and are more likely to remit profits in the form of (deductible) royalty payments than (nondeductible) dividends. Only a small fraction (16% in 1984) of the foreign affiliates of American companies pay dividends to their parent companies each year, those that do tending to have particularly tax-advantaged situations (Hines and Hubbard, 1990; Desai et al., 2001; Dharmapala et al., 2011). And the pretax profitability of foreign affiliates is negatively correlated with host-country tax rates (Hines and Rice, 1994; Huizinga and Laeven, 2008; Clausing, 2009), which is suggestive of

tax-motivated income reallocation, though these simple correlations may overstate the extent to which taxpayers are able to adjust their own tax liabilities this way (Dharmapala and Riedel, 2013; Hines, 2014).

Incidence of the Corporate Income Tax

The requirement that corporations pay taxes does not mean that owners of corporations necessarily bear the burden of the corporate tax, since this burden might be partially or entirely shifted to consumers in the form of higher prices, or to workers in the form of lower wages. Indeed, there are reasonable situations in which the existence of the corporate tax actually enhances returns to capital owners, including owners of corporations (Harberger, 1962).

Corporate tax obligations contribute to the cost of investment and thereby encourage substitution of other productive factors (such as labor) for capital used by corporations. Labor expenses are deductible against taxable income, so the corporate tax does not affect the marginal condition characterizing a firm's decision of whether or not to employ additional labor. While the substitution of labor for capital depresses the demand for capital and thereby reduces its after-tax market return, there exists a separate effect of corporate taxation that stems from its induced intersectoral reallocation of resources.

Corporate taxation increases the cost of producing corporate output, thereby raising output prices, depressing demand, and shifting output from the corporate sector of the economy to the noncorporate sector. This reallocation affects factor demands to the extent that factor input ratios differ between the corporate and noncorporate sectors of the economy. If the corporate sector of the economy has a lower capital/labor ratio than the noncorporate sector, then the introduction of a corporate tax shifts resources into the noncorporate sector and thereby raises the demand for capital. If this effect is large enough, then it has the potential to exceed in magnitude the countervailing impact of factor substitution, thereby implying that higher rates of corporate tax are associated with greater after-tax returns to capital – including capital invested in corporations. It would then follow that labor bears the burden of the corporate tax in the form of lower wages.

The distribution of the corporate tax burden is an important area for future empirical research. The limited available statistical evidence is largely inconclusive and mostly quite old (Krzyzaniak and Musgrave, 1963), making it difficult to know what groups in the population currently bear the burden of corporate taxes. The globalization of business activity may add an important dimension to the incidence of corporate taxes, since international capital flows tend to reduce the impact of corporate taxes on the returns to owners of corporations, and instead shift the burdens to owners of other productive factors (Harberger, 2006).

See also: Corporate Finance: Financial Control; Financial Regulation; Income Taxes; Multinational Corporations; State and Local Taxation.

Bibliography

Auerbach, Alan J., 2002. Taxation and corporate financial policy. In: Auerbach, Alan J., Feldstein, Martin (Eds.), Handbook of Public Economics, vol. 3. North-Holland, Amsterdam.

Auerbach, Alan J., Jorgenson, Dale W., 1980. Inflation-proof depreciation of assets. Harvard Business Review 58, 113–118.

Chetty, Raj, Saez, Emmanuel, 2005. Dividend taxes and corporate behavior: evidence from the 2003 dividend tax cut. Quarterly Journal of Economics 120, 791–833.

Chirinko, Robert S., Fazzari, Seven M., Meyer, Andrew P., 1999. How responsive is business capital formation to its user cost? An exploration with micro data. Journal of Public Economics 74, 53–80.

Clausing, Kimberly A., 2009. Multinational firm tax avoidance and tax policy. National Tax Journal 62, 703–725.

Da Rin, Marco, Di Giacomo, Marina, Sembenelli, Alessandro, 2011. Entrepreneurship, firm entry, and the taxation of corporate income: evidence from Europe. Journal of Public Economics 95, 1048–1066.

Desai, Mihir A., Foley, C. Fritz, Hines Jr., James R., 2001. Repatriation taxes and dividend distortions. National Tax Journal 54, 829–851.

Desai, Mihir A., Foley, C. Fritz, Hines Jr., James R., 2004a. Foreign direct investment in a world of multiple taxes. Journal of Public Economics 88, 2727–2744.

Desai, Mihir A., Foley, C. Fritz, Hines Jr., James R., 2004b. A multinational perspective on capital structure choice and internal capital markets. Journal of Finance 59, 2451–2487.

Dharmapala, Dhammika, Foley, C. Fritz, Forbes, Kristin J., 2011. Watch what I do, not what I say: the unintended consequences of the homeland investment Act. Journal of Finance 66, 753–787.

Dharmapala, Dhammika, Riedel, Nadine, 2013. Earnings shocks and tax-motivated income-shifting: evidence from European multinationals. Journal of Public Economics 97, 95–107.

Djankov, Simeon, Ganser, Tim, McLiesh, Caralee, Ramalho, Rita, Shleifer, Andrei, 2010. The effect of corporate taxes on investment and entrepreneurship. American Economic Journal: Macroeconomics 2, 31–64.

Goolsbee, Austan D., 1998. Investment tax incentives, prices, and the supply of capital goods. Quarterly Journal of Economics 113, 121–148.

Gordon, Roger H., Hines Jr., James R., 2002. International taxation. In: Auerbach, Alan J., Feldstein, Martin (Eds.), Handbook of Public Economics, vol. 4. North-Holland, Amsterdam.

Graham, John R., 1996. Debt and the marginal tax rate. Journal of Financial Economics 41, 41–73.

Hall, Robert E., Jorgenson, Dale W., 1967. Tax policy and investment behavior. American Economic Review 57, 391–414.

Harberger, Arnold C., 1962. The incidence of the corporation income tax. Journal of Political Economy 70, 215–240.

Harberger, Arnold C., 2006. Corporate tax incidence: reflections on what is known, unknown, and unknowable. In: Diamond, John W., Zodrow, George R. (Eds.), Fundamental Tax Reform: Issues, Choices and Implications. MIT Press, Cambridge, MA.

Hassett, Kevin A., Hubbard, R. Glenn, 2002. Tax policy and business investment. In: Auerbach, Alan J., Feldstein, Martin (Eds.), Handbook of Public Economics, vol. 3. North-Holland, Amsterdam.

Hines Jr., James R., 1996. Altered states: taxes and the location of foreign direct investment in America. American Economic Review 86, 1076–1094.

Hines Jr., James R., 1999. Lessons from behavioral responses to international taxation. National Tax Journal 52, 305–322.

Hines Jr., James R., 2014. How serious is the problem of base erosion and profit shifting? Canadian Tax Journal 62, 443–453.

Hines Jr., James R., Hubbard, R. Glenn, 1990. Coming home to America: dividend repatriations by US multinationals. In: Razin, Assaf, Slemrod, Joel (Eds.), Taxation in the Global Economy. University of Chicago Press, Chicago, IL.

Hines Jr., James R., Rice, Eric M., 1994. Fiscal paradise: foreign tax havens and American business. Quarterly Journal of Economics 109, 149–182.

House, Christopher L., Shapiro, Matthew D., 2008. Temporary investment tax incentives: theory with evidence from bonus depreciation. American Economic Review 98, 737–768.

Huizinga, Harry, Laeven, Luc, 2008. International profit shifting within European multinationals. Journal of Public Economics 92, 1164–1182.

Huizinga, Harry, Laeven, Luc, Nicodeme, Gaetan, 2008. Capital structure and international debt shifting. Journal of Financial Economics 88, 80–118.

King, Mervyn A., Fullerton, Don (Eds.), 1984. The Taxation of Income from Capital. University of Chicago Press, Chicago, IL.

Krzyzaniak, Marian, Musgrave, Richard A., 1963. The Shifting of the Corporation Income Tax. Johns Hopkins Press, Baltimore, MD.

Miller, Merton H., 1977. Debt and taxes. Journal of Finance 32, 261–275.

Corpus Linguistics

Chu-Ren Huang and Yao Yao, Hong Kong Polytechnic University, Hong Kong

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Kennedy, volume 4, pp. 2816–2820, © 2001, Elsevier Ltd.

Abstract

This article introduces basic concepts of a modern linguistic corpus and corpus linguistics. A corpus is defined as a collection of examples of language in use that are selected and compiled in a principled way and corpus linguistics as linguistic studies of such corpora. We explicate classification, basic procedures of data collection, construction, and annotation of corpora. Representative research areas and applications where corpus and corpus-based analysis play crucial roles are also introduced. Finally, trends and future directions of development of corpus linguistics are discussed.

What Is a Corpus?

A corpus is a collection of examples of language in use that are selected and compiled in a principled way. The term corpus refers to the intention for it to be a representative body of evidence for the study of language and language use. In the most general terms, the purpose of a corpus is to document a language. Hence a corpus is often an essential component of *language documentation* or *language archives*. Corpora are often used in language sciences, including linguistics, computational linguistics (CL), language education psycholinguistics, sociolinguistics, and translation studies, as well as other studies where a collection of texts or other language uses are crucial, such as anthropology, communication studies, literary studies, or political science.

What Is Corpus Linguistics?

The term *corpus linguistics* refers to corpus-based linguistic studies in general (Biber et al., 1998; Tognini-Bonelli, 2001, among others). Archetypical corpus work existed well before the modern digital era, as exemplified by the early attempts of word indexing and concordancing of the Christian Bible in the thirteenth century. However, the emergence of corpus linguistics as an academic discipline is closely linked to the availability of digital tools to record, store, and examine corpora, heralded by the rapid development of computational technology as well as the increasing availability of digital content since the second half of the twentieth century. Although the use of computational and statistical tools plays crucial roles in corpus linguistics now, it is important to note that the two essential elements in corpus linguistics are the study of linguistic issues and the use of corpora. Thus 'computer-aided armchair linguistics,' the term Charles Fillmore fondly coined in 1992 (Fillmore, 1992), still applies to corpus linguistics to the extent that both linguistic theories and argumentation are required to lay the foundation of corpus-based processing and analysis.

The first modern corpus is the Brown University Standard Corpus of Present-day American English ('the Brown corpus'; Francis and Kučera, 1979; Kučera and Francis, 1967), compiled at Brown University in the 1960s. With about one million words of text collected from a balanced, wide variety of sources, the Brown corpus represented a breakthrough in terms of both corpus size and corpus design. In subsequent years, however, the standard of corpus size increased rapidly. Thanks to the enhanced storage and processing power of modern computers, it is now not uncommon to encounter a corpus that contains hundreds or thousands of millions of words. What came hand in hand with the increase in scale was the expansion in genre. In addition to the so-called 'balanced' corpora (see below for the definition of *balanced*) such as the Brown corpus, corpora have also been developed for highly specialized domains and with unconventional data types (e.g., multimodal language data, text collected from the World Wide Web (WWW)).

Building a Corpus

Design Criteria

The construction of a corpus starts with decisions on design criteria. Corpus design criteria are mainly driven by the purpose of the corpus, but may also be affected by meta-theoretical concerns such as evaluation methods, reusability, and interoperability. Pragmatic concerns such as data availability, sharability, as well as budgetary and technical constraints may also play a role. In setting up a set of design criteria, it is important to bear in mind that a corpus is meant to be a representative body of linguistic data as evidence; hence the question of whether the design criteria will yield the right sample of data to address the questions posed must be asked. The design criteria will determine what kind of data to collect, how much and how to collect, what kind of information will be annotated on the data, and by what method.

Data Collection

Corpus construction begins with the collection of raw language data. As stated earlier, corpus developers need to first decide on the purpose of the corpus (i.e., *Language use in which domain(s) of human life will be represented in this corpus?*) based on which they may decide further where to collect data. Data collection in the earlier stage of corpus linguistics involves both selection and digitization of text. With the advance of the WWW and the accessibility of digital content, data collection of modern corpora typically involves manual or automatic selection or extraction of electronic texts. Both methods, however, require an elaborated and challenging effort to clear copyrights to the

acquired texts, unless a specific text is already in the public domain. The recent emergency of the Web as Corpus (WAC, see Baroni et al., 2009) approach, which treats the WWW as a source of data, has the potential to side-step some, but not all, copyright issues. Under the WAC approach, computer programs (which are called *robots*) are designed to crawl the web and collect data according to a set of predetermined criteria. Robots are usually designed to extract only a small portion of text from each website in order to meet the fair-use requirement. Another WAC method to circumvent the rights issue is to collect as much data as needed but only share the extracted generalization, and never the original data in pure text form, to users. Two best known corpora taking this approach are Google's n-gram corpus, which gives users access only to the collection and statistics of n-grams (n consecutive words) but not the original text; and Liverpool's WebCorp, where concordance lines are dynamically extracted and not saved. Discussion on corpus and corpus construction will be focused on the typical corpus as a collection of text in what follows.

Balanced versus Specialized

A *balanced* corpus (also referred to as a 'general corpus') is designed to represent the use of a certain language in all domains with the consideration of relative frequency of language use in each domain. Therefore, the data of a balanced corpus are representative of all topics, genres, styles, and registers, coming from a potpourri of sources which often include (but not limited to) books, magazines, newspapers, academic journals, and TV programs, with a controlled proportion of each data source and data type. The proportions are decided by the corpus developers in some principled way, e.g., according to the number of publications in each source/type in a certain time period. Most existing balanced corpora are of the English language, with the Brown corpus setting the standard for early corpora, and the British National Corpus (BNC) setting the standard for more recent work. Balanced corpora of other languages are also available, such as the Academia Sinica Balanced Corpus of Modern Chinese (Sinica Corpus) and the Balanced Corpus of Contemporary Written Japanese (KOTONOHA or BCCWJ). Balanced corpus can also be constructed to represent variations of the same language, such as the International Corpus of English.

A *specialized* corpus, as the name suggests, only represents the use of language in a specific domain or for a specific topic. The range of specialized corpus varies widely from single author corpus (e.g., a Shakespeare corpus), single genre corpus (e.g., a Twitter corpus) to single domain corpus (often a corpus of language for specific purpose, e.g., a corpus of engineering English) and so on. A specialized corpus can also be complied for the purpose of a specific research topic area, such as a learner's corpus for the study of second language learning, an acquisition corpus for the study of first language education (e.g., CHILDES), an emotion corpus, a language disorder corpus, or a metaphor corpus.

Synchronic versus Diachronic

A *synchronic* corpus contains language data that are produced in roughly the same time period, whereas a *diachronic* corpus contains data from different time periods, which may or may not be consecutive. For instance, a corpus of Shakespeare's works is synchronic, but a corpus of English literature in the seventeenth–twentieth centuries would be diachronic. Diachronic corpora are often used in historical linguistics and the study of language change, thus the time span must be long enough for changes to be observed. The emergent field of evolutionary linguistics also relied heavily on diachronic corpora for modeling and simulation. Most balanced corpora are synchronic, simply because compiling a balanced corpus with enough data to cover an extensive time span would be too overwhelming of a task.

Spoken, Written, or Mixed

A *spoken* corpus contains transcribed speech, whereas a *written* corpus contains only written text. Of course it is possible to have both spoken data and written data in a corpus, resulting in a *mixed* corpus. An example of written corpora is the Brown corpus, which only included published written works. Some more recent balanced corpora, such as BNC, included both written works and transcribed spoken data, in recognition of the vast differences between written and spoken language (Chafe and Tannen, 1987; Halliday, 1994). The past three decades have witnessed the emergence of a number of corpora with only spoken data, such as the SWITCHBOARD corpus of telephone speech and the Buckeye corpus of conversational speech. Since these corpora usually contain speech recordings and phonetic transcription alongside word transcription, they are often used for studying phonetics and phonology, as well as speech recognition and processing.

Single Language versus Multilanguage

As the names suggest, a *single-language* (or *monolingual*) corpus contains data from only one language, whereas a *multilanguage* (or *multilingual*) corpus contains data from more than one language. A multilanguage corpus can be either a parallel corpus or a comparable corpus (see below for the differences between the two). Earlier multilanguage corpora are mostly parallel corpora, while comparable corpora became more popular recently with the availability of original electronic data from more languages on the web. Multilanguage corpora are widely used in translation studies, computer-assisted translation, machine translation, and other language-related computer technologies. The first widely used multilanguage corpus is the Hansard Corpus of English and Canadian French from Canadian parliamentary proceedings, which contain texts in both English and Canadian French.

Parallel versus Comparable

A *parallel* corpus contains texts of the same content in different languages (e.g., an original text and its translation(s) in one or more other languages). Hence directionality is crucial for a parallel corpus. For example, a parallel English–French corpus will be different from a parallel French–English corpus, even if they are both based on the same text. The parallel relation between the source and target also varies according to the nature and purpose of translation. Parallel corpora containing texts with legal or documentary functions, such as parliamentary proceedings, tend to maintain lexico-syntactic correspondences, while those containing texts that are informative or expressive tend to maintain semantic-pragmatic correspondences. One of the most important steps in

processing a parallel corpus is alignment, which can be performed at lexical, phrasal, sentential, or paragraph level, depending on the nature of the corpus as well as the purpose of the study. Comparable corpora, on the other hand, are multilanguage corpus containing a collection of texts from two or more languages collected under the same set of criteria. Comparable corpora are less costly to construct since they do not rely on translators to create the target texts. They also allow comparative study of two languages in similar native language production context, in contrast to parallel corpora, where target language text is more restricted and may contain unnatural texts. It is also possible to construct a large comparable corpus without having to translate all the texts from source language. Hence the comparable corpus approach opens up the possibility of constructing a multilanguage corpus in language pairs without qualified translators or sufficient pre-existing translated texts.

Annotation and Storage

As mentioned earlier, modern corpora are digitally stored. In particular, to facilitate future processing by computer programs, corpora must be stored in a 'machine-readable' electronic format. To this end, a popular approach is to use the Extensible Markup Language (XML) for encoding and storing corpus files. Raw language data may require some preprocessing, e.g., converting text on paper to electronic text, transcribing speech, and adjusting electronic text format, before they can be stored in an XML format.

An XML file contains two types of elements: content and markup. In a corpus XML file, content mainly consists of the linguistic data collected for the corpus. Markup, on the other hand, is a major place for adding *annotation* to the linguistic data, which may be any extra information about the text.

Corpus annotation has become a critical step in corpus construction (see Stede and Huang, 2012) in spite of the fact that annotation is not a mandatory component of a corpus and that corpus annotation is typically costly and labor-intensive. A simple reason for the convention is that annotation enriches a corpus and facilitates discovery of generalizations and knowledge which is difficult, if not impossible, to obtain from a raw corpus. For example, the annotation of author gender will be crucial for analyzing any gender-related phenomenon in language use. Take another example. Many English words may be used with more than one part of speech (POS) (e.g., 'table,' 'convict'). If someone is to research the use of 'table' as a verb, it will be much more convenient if he or she uses a corpus with POS tags, which would allow word search to be restricted by POS and therefore prevent search results to be overwhelmed by instances of 'table' used as a noun. In addition, POS tagging also facilitates various natural language processing (NLP) applications such as information retrieval and machine translation.

A corpus can be annotated at three different levels. Most corpora are at least encoded with some basic metadata (such as author, publication date, genre) which are attributes of the text as a whole. There are also corpora in which the texts are tagged with grammatical information, phonetic transcription, or other linguistic or extralinguistic information for a specific purpose of study. Overall, the text encoding initiative (TEI) provided the XML schema for encoding of all levels of corpus annotation. TEI aims to ensure that members of its community of practice will be able to share their annotated digital output. Even though TEI is designed for the wider community for all digitalized texts, it has been adapted as the standard format for corpus annotation. First, for the encoding of metadata, Dublin Core is the *de facto* standard of the open archives community, and is adapted by the Open Language Archives Community for language resources including corpora. Second, the tagging of linguistic information, which is the most common and most versatile level of annotation, can be applied to all kinds of corpora and can easily be reused. POS tagging is the most commonly applied annotation. When literatures in corpus linguistics refer to a tagset, it means the set of POS tags used in a corpus by default. All currently available balanced corpora are POS-tagged, and a corpus that is annotated with other information typically has a name that explicitly refers to that annotation level, e.g., prosody corpus (a syntactically annotated corpus, however, is conventionally called a *treebank*). It is also important to note that although the most commonly assumed unit in a corpus is a word, it is possible to annotate units smaller than words. For instance, a corpus of speech data can be segmented into phonemic or phonetic units, which are labeled with segmental phonetic transcriptions, although the same corpus could also be annotated with suprasegmental features that span across multiple segments (or even words). The General Ontology of Linguistic Description is a community of practice effort to define and standardize all linguistic features that can be annotated on a language resource. A parallel effort to provide a uniform and sharable linguistic annotation scheme is the Lexical Markup Framework group (Francopoulo, 2013), which works in conjunction with ISO TC37 SC4 to produce an ISO standard for linguistic annotation: ISO 24613:2008. Finally, a corpus can also be annotated with extralinguistic features to enable it to support specific research. For instance, in a Learner's Corpus, the common annotation scheme, on top of POS tagging, identifies the errors and error types made by a language learner. An emotion corpus can be annotated with both the emotion type and emotion-related event structure.

Standard procedure of corpus annotation starts with drafting of annotation guidelines based on linguistics or other theoretical criteria. Annotators are trained and their performance is evaluated before given the actual task. Annotation tasks are usually carried out in a double-blind manner, meaning a pair (or more) of annotators annotating the same text without input from each other. After the annotation is completed, the annotations where both (or all) annotators agree on are assumed to be correct, and the annotations where there are disagreements, they are presented to an adjudicator, often the person who drafted the guidelines, to decide upon. Hence, the measurement of interannotator agreement is often taken to be a good indicator of the quality of corpus annotation. Although the earliest corpus started with manual annotation, modern corpus annotation is typically carried out electronically with an annotation tool. The tool may also have the capacity to provisionally predict the annotation based on either statistic or heuristic rules. The annotator's main role is then to correct wrong predictions or to supply annotation when the program fails to provide any annotation. Fully

automatic annotation can be employed when the corpus scale is too big for human annotation, such as with the WAC approach. In this case, a tolerance of error rate is predetermined and the program is tested on a sample corpus with gold standard (i.e., correct) annotation to make sure that it does not make excessive errors before the program is deployed. Such automatic annotation programs typically employ either machine learning algorithms or stochastic models; they are sometimes also supplemented by heuristic or hand-crafted linguistic rules.

Crowdsourcing is a recent innovation in corpus annotation which has the potential to generate a paradigm shift. Crowdsourcing relies on web users to perform tasks that require human intelligence, typically the kind of tasks that a computer program has limitations in. Corpus annotation falls nicely to the category of human intelligence tasks. The typical procedure for crowdsourced annotation also starts with the design of annotation guidelines, especially with presenting the guidelines with examples to ensure least ambiguity for workers who take on the task. The corpus owner then publishes this task on the web through a crowdsourcing platform, with optional monetary compensation. The workers who accept the task must typically pass a pre-test to show that they have the knowledge or skill to perform the task. Since the compensation is low and since the number of potential works on the web is big, crowdsourcing allows a corpus to be annotated by multiple annotators with a fraction of the cost of dedicated annotators. The larger number of annotators allows corpus owner to have a wider range and more reliable data to determine which annotation is more reliable. The most popular crowdsourcing platforms for language-related studies are Amazon Mechanic Turks and Crowdflower.

Corpus Linguistics: Research and Applications

The topics in corpus linguistics research are not different from (computational) linguistic research. The only differences are in the approaches to how data are collected and to how generalizations are arrived. Hence, we will focus on research topics generated by and solved with corpus linguistics.

Corpus Processing Tools

The design and improvement of corpus processing tools is an ongoing issue in corpus linguistics. The Natural Language Toolkit gives a good overview of what kind of tools and resources are available or required for corpus processing. First of all, tokenizer or lemmatizer is often required as the first step of corpus processing. Since not all languages conventionally mark the boundaries of words, and even with those that do, not all boundaries are consistently marked (e.g., 'the White House' is a single word), hence tokenization is an important first step for languages without conventionalized word boundary markers. For example, the tokenization task for Mandarin Chinese, which is more specifically referred to as Chinese word segmentation, remains an active research topic (see Huang and Xue, 2012). Lemmatizer, on the other hand, can play the same function as tokenizer in languages with rich morphology. At word level, it serves the purpose of enabling words which undergo regular or irregular morphology to be recognized as tokens of the same type of words (e.g., 'buy' and 'bought'). It also enables generalizations of the behaviors of affixes. Second, a POS tagger is dependent on a well-designed tagset, a comprehensive lexicon as well as strategic information from distributional properties of each POS. Finally, a parser can assign syntactic structure to a corpus, which is not only crucial for construction of treebanks but also allows event structure to be defined and annotated on a corpus. All these processing tools lay the foundation for more advanced corpus analysis such as word frequency counting or automatic identification of grammatical relations (see below for more discussion on corpus analysis).

Applications of Corpus

The first and best known application of corpus is in lexicography. The COBUILD team demonstrated that high-quality dictionaries and grammars can be produced based on corpus data (Sinclair, 1987). Today, all major dictionaries/lexica in English are corpus based and some with a very powerful user interface are able to provide linguistic generalization automatically. Since the first application in lexicography, corpus has played an increasingly important role in general linguistic research. By definition, a corpus contains a large amount of naturally occurring language data and therefore becomes an ideal data source for investigating language and language use. It is not uncommon now for a study of syntax or semantics to cite example sentences collected from natural corpora. For this purpose, the most often used corpus analyses are word frequency counting, concordance, and keyword in context, all of which are standard functions available in most corpus websites and corpus analysis software. A more advanced tool is Sketch Engine (Kilgarriff et al., 2004), which automatically extracts grammatical relations based on statistical patterns in the corpus. In addition to syntactic and semantic research, the use of corpus is especially important in discourse analysis and studies of language variation. While discourse analysis usually employs qualitative analysis aided by the commonly used corpus analysis tools listed earlier, a corpus-based study of language variation typically uses quantitative analysis with statistical models built on large datasets extracted from the corpus for the examination of variation patterns.

Today, corpus has also become an indispensable component in CL research. Currently almost all major CL applications (e.g., information retrieval, sentiment and emotion analysis, machine translation, speech recognition and synthesis) heavily rely on corpus data and the use of corpus-based approaches. Since the core of a CL application often involves recognizing some patterns or features of the language data, the application typically makes use of two sets of corpus data: a training set and a testing set. The training set contains data that have already been tagged (often manually) with gold standard annotation, which are then used to train the CL program to calculate the heuristic and statistical rules for automatic annotation. Goodness of the program may be evaluated by comparing the annotation results with the gold standard annotation on the same dataset. Later, the trained CL program will be applied to the testing dataset in order to obtain automatic evaluation. Obviously, the more similar the training set and the testing set

are, the more likely it is for the trained CL program to gain success with the unseen data in the testing set. In reality, it is often recommended that the two datasets are sampled from the same corpus (or corpora of a similar type) in order to ensure high data similarity. The earlier description also indicates a close relationship between corpus linguistics and CL. While corpora provide the essential resources for CL research, the latter also contributes novel and efficient ways to process and annotate corpus data.

Trends and Future Directions of Corpus Linguistics

There are a number of noteworthy trends in the development of corpus and corpus linguistics. First of all, as mentioned earlier, a general trend is that corpora are becoming increasingly larger than before, thanks to the development of computers and corpus technology. Second, the establishment of various corpus construction and annotation standards promises the emergence of more 'standardized' corpora with high reusability and interoperability. In the past decade, we have also noted the fast development of multimodal corpora, which contain not only textual data but also nontextual audio or video data, making it possible to study language use in a broader cognitive and communicative context. Such corpora have been proved highly useful in the study of child language development and nonverbal communication. Finally, there are also some emerging trends related to the development of the WWW. As noted before, novel technology such as crowdsourcing has been employed in the process of corpus construction, making annotation more efficient and economic. Another trend is the use of language data crawled from the web, i.e., the WAC movement, which easily produces gigantic corpora of naturally produced text although questions about data quality also arise.

Overall we are highly optimistic about the future of corpus linguistics. Given the current achievement and the emerging trends, we are looking forward to a future where corpora make even greater contributions to linguistic research and applications, language professions, and the broader fields of social and behavioral sciences.

See also: Areal Linguistics; Dialectology; Etymology; Indo-European Languages; Information Structure in Linguistics; Language and Society; Linguistic Typology; Pidgin and Creole Language Varieties; Sapir–Whorf Hypothesis.

Bibliography

Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E., 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation 43 (3), 209–226.

Biber, D., Conrad, S., Reppen, R., 1998. Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press, Cambridge.

Chafe, W., Tannen, D., 1987. The relation between written and spoken language. Annual Review of Anthropology 16, 383–407.

Fillmore, C., 1992. 'Corpus linguistics' or 'computer-aided armchair linguistics'. In: Svartvik, J. (Ed.), Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 (Trends in Linguistics Studies and Monographs 65). Mouton de Gruyter, Berlin, pp. 35–60.

Francis, W.N., Kučera, H., 1979. Brown Corpus Manual (Revised and Amplified), Department of Linguistics. Brown University, Providence, RI.

Francopoulo, G. (Ed.), 2013. LMF: Lexical Markup Framework, Theory and Practice. ISTE, London.

Halliday, M.A.K., 1994. Spoken and written modes of meaning. In: Graddol, D., Barret, O.B. (Eds.), Media Texts: Authors and Readers: A Reader. Multilingual Matters, Clevedon/Philadelphia, pp. 51–73.

Huang, C.-R., Xue, N., 2012. Words without boundaries: computational approaches to Chinese word segmentation. Language and Linguistics Compass 6 (8), 494–505.

Kilgarriff, A., Rychlý, P., Smrz, P., Tugwell, D., 2004. The sketch engine. In: Proceedings of the 11th EURALEX Congress, Lorient, France.

Kučera, H., Francis, W.N., 1967. Computational Analysis of Present-Day American English. Brown University Press, Providence, RI.

Stede, M., Huang, C.-R. (Eds.), 2012. Special Issue: Linguistic Annotation, Language Resources and Evaluation, vol. 46 (1).

Sinclair, J.M. (Ed.), 1987. Looking up: An Account of the COBUILD Project in Lexical Computing. Collins ELT, London and Glasgow.

Tognini-Bonelli, E., 2001. Corpus Linguistics at Work. John Benjamins, Amsterdam and Philadelphia.

Relevant Websites

http://icame.uib.no/brown/bcm.html – Brown Corpus Manual.
http://www.natcorp.ox.ac.uk/ – British National Corpus (BNC).
http://crowdflower.com/ – Crowdflower Crowdsourcing Platform.
http://www.elda.org/ – Evaluations and Language Resources Distribution Agency (ELDA).
http://linguistics-ontology.org/ – General Ontology for Linguistic Description (GOLD).
http://www.ldc.upenn.edu/ – Lexical Data Consortium (LDC).
http://www.lexicalmarkupframework.org/ – Lexical Markup Framework (LMF).
http://nltk.org/ – Natural Language Toolkit (NLTK).
http://www.sinica.edu.tw/SinicaCorpus/ – Sinica Corpus.
http://wacky.sslmit.unibo.it/ – Web-As-Corpus Kool Yinitiative (WaCky).

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 5

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser—Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 5

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 5 *xliii*

VOLUME 5

Correspondence Analysis 1
Michael Greenacre

Correspondence Analysis, History of 6
Jörg Blasius and Andreas Schmitz

Corruption: Political and Public Aspects 12
Alina Mungiu-Pippidi

Cortical Areas Engaged in Movement: Neuroimaging Methods 21
Jody Culham

Cortical Development: Transplantation and Rewiring Studies 30
Mriganka Sur

Coser, Lewis A. (1913–2003) 35
Thomas Crosbie

Cosmopolitanism 41
Huon Wardle

Cost–Benefit Analysis 47
Anthony E Boardman

Cost-Benefit and Cost-Effectiveness Analyses in Evaluation Research 55
Brian T Yates

Counseling and Psychotherapy: Ethnic and Cultural Differences 63
Vivian Ota Wang

Counseling Psychology 69
Charles J Gelso

Counterculture: 1960s and Beyond 73
Ralph W Larkin

Counterculture: The Classical View 80
Sheila Whiteley

Counterfactual Reasoning, Qualitative: Philosophical Aspects 87
Stathis Psillos

Counterfactual Reasoning, Quantitative: Philosophical Aspects 95
Alan Hájek

Counter-Insurgencies and the Social Sciences 99
Juan J Navarro and Juan J Morales

Courts and Adjudication 103
Malcolm M Feeley and Carlo Guarnieri

Craft Production, Anthropology of 108
Esther N Goody

Creativity Training 113
Mark A Runco and Sarah E Sumners

Creolization: Sociocultural Aspects 117
Christine Jourdan

Crime and Class 122
Roland Chilton and Ruth Triplett

Crime and Criminal Justice: Applied Research from Routine Monitoring to Evidence-Based Practices 127
Roberto Hugh Potter and Gail S Humiston

Crime and Gender 132
Samantha R Cumley, Karen Heimer, and Stacy De Coster

Crime, Desistance, and Persistence 137
Michael E Ezell

Crime, Geography of 145
David Herbert

Crime, Immigration, and Ethnicity 151
Sandra M Bucerius and Katherine Hancock

Crime in the Rural Context 158
Joseph F Donnermeyer

Crime: Knowledge about and Prevalence 164
Rosemary Gartner

Crime: Organized 170
Andrew V Papachristos and Sandy Y Zhao

Crime Policy in the United States 176
Candace McCoy

Crime, Race and Ethnicity 182
Malcolm D Holmes

Crime, Sociology of 189
Lorine A Hughes and James F Short, Jr

Crime Trends and Debates 194
Patricia L McCall and Joshua A Hendrix

Crime: Whitecollar 203
Susan P Shapiro

Criminal Justice and Social Work Practice 208
Roger Smith

Criminal Justice, Sociology of 214
Myron Orleans

Criminal Liability and Criminal Defenses 219
Gerry Ferguson

Criminal Procedure: Adversarial and Inquisitorial Legal Systems 227
Stephen C Thaman

Criminology: Development and Current Status of the Discipline 232
Charles F Wellford

Criminology, History of 238
Joachim J Savelsberg, Lorine A Hughes, Janne Kivivuori, James F Short Jr, Maximo Sozzo, and Richard Sparks

Critical Criminology 244
Bruce A Arrigo and Heather Y Bersot

Critical Legal Studies 251
Robert W Gordon

Critical Planning Theory 255
Michael Gunder

Critical Psychology 261
Horst Gundlach

Critical Race Theory 266
Angela P Harris

Critical Rationalism 271
Volker Gadenne

Critical Realism in Geography 277
Andrew Sayer

Critical Reflexivity and Social Work Practice 281
Christine Morley

Critical Social Work Practice 287
Jan Fook

Critical Theory 293
Darrell P Arnold

Critical Theory: Contemporary 304
Larry Ray

Cross-Cultural Psychology 311
Ype H Poortinga

Cross-Cultural Research Methods in Psychology 318
Fons van de Vijver

Cross-Cultural Research Methods in Sociology 322
Fons van de Vijver

Cross-Cultural Study of Education 327
Akane Zusho and Karen E Clayton

Cross-Cultural Training in Industrial and Organizational Psychology 334
Ariel Lelchook and Mary Sully de Luque

Cross-National Comparisons in Education: Findings from PISA 342
Ming M Chiu and Sung W Joh

Crossover Designs 349
Michael G Kenward

Cross-Sectional Methodologies in Developmental Psychology 354
Victoria CP Knowland, Harry Purser, and Michael SC Thomas

Crowds in History 361
M Gailus

Cuban Revolution, The 367
Manuel R Moreno Fraginals and Teresita Pedraza Moreno

Cultural Capital and Education 375
Susan A Dumais

Cultural Critique, Anthropological 382
Keith Hart

Cultural Evolution: Overview 388
Alex Mesoudi

Cultural Evolution: Phylogeny versus Reticulation 394
Peter Bellwood

Cultural Evolution: Theory and Models 401
Laurel Fogarty and Marcus W Feldman

Cultural Geography 409
Kevin Hannam

Cultural Heritage 414
Christoph Brumann

Cultural History 420
Roger Chartier

Cultural Influences on Interpersonal Relationships 426
Patricia MR Mosquera

Cultural Intelligence and Competencies 433
Soon Ang, Thomas Rockstuhl, and Mei Ling Tan

Cultural Mediators and Gatekeepers 440
Susanne Janssen and Marc Verboord

Cultural Participation, Trends in 447
Henk Roose and Stijn Daenekindt

Cultural Policies in East Asia 453
Nobuko Kawashima

Cultural Policy Regimes in Western Europe 460
Vincent Dubois

Cultural Production in Networks 466
Katherine Giuffre

Cultural Psychiatry 471
Ahmed Hankir and Dinesh Bhugra

Cultural Psychology 478
Richard A Shweder

Cultural Relativism, Anthropology of 484
James W Fernandez

Cultural Resource Management: Conservation of Cultural Heritage 488
Thomas J Green

Cultural Rights and Culture Defense: Cultural Concerns 491
Alison Dundes Renteln

Cultural Sociology, History of 498
Matthias Revers

Cultural Studies of Science 504
Joseph Rouse

Cultural Views of Life Phases 507
M Annette Grove and David F Lancy

Culturally Appropriate Interventions in Social Work 516
Kish Bhatti-Sinclair

Culture and Actor Network Theory 523
Ignacio Farías and Sophie Mützel

Culture and Economic Development 528
Eelke de Jong

Culture and Economy 535
Nina Bandelj and Paul J Morgan

Culture and Emotion 542
Batja Mesquita, Nathalie Vissers, and Jozefien De Leersnyder

Culture and Institutional Logics 550
Patricia H Thornton

Culture and Networks 557
Ronald L Breiger and Kyle Puetz

Culture and the Self: Implications for Psychological Theory 563
Joan G Miller and Malin Källberg-Shroff

Culture as a Determinant of Mental Health 570
Carolyn B Murray

Culture, Cognition and Embodiment 576
Omar Lizardo

Culture: Contemporary Views 582
Richard A Shweder and Les Beldo

Culture of Poverty: Critique 590
Hervé Varenne and Michael Scroggins

Culture Policy Regimes in China 596
Yue Zhang

Culture, Production of 602
Richard A Peterson

Culture, Production of: Prospects for the Twenty-First Century 608
Jennifer C Lena

Culture Shock 614
Waud Kracke

Culture, Sociology of 617
Mabel Berezin

Culture-Specific Psychiatric Disorders 622
Oyedeji Ayonrinde

Customer Satisfaction 630
Michael D Johnson

Cyber Crime: Identity Theft 633
Katie A Farina

Cyberbullying 638
Elizabeth Whittaker and Robin M Kowalski

Cybernetics 645
Andrew Pickering

Cyclical Industrial Dynamics 651
Hao Tan

Dahl, Robert A (1915–2014) 655
Charles HG Crothers

Dahrendorf, Ralf (1929–2009) 661
Hans-Peter Müller

Dance, Anthropology of 666
Helena Wulff

Dark Triad: The "Dark Side" of Human Personality 671
Peter K Jonason and James P Middleton

Darwin, Charles Robert (1809–82) 676
Michael T Ghiselin

Darwinism 681
Michael Ruse

Darwinism: Social 686
Alain Drouard

Data Bases and Statistical Systems: Applied Social Research 692
Joseph M Gasper and Kathryn A Henderson

Data Bases and Statistical Systems: Archives and Historical Databases 700
O Dascher

Data Bases and Statistical Systems: Communications and Media 704
Wolfgang Seufert

Data Bases and Statistical Systems: Crime 712
Clayton Mosher and Terrance Miethe

Data Bases and Statistical Systems: Crime and Delinquency 717
Mark Warr and David S Kirk

Data Bases and Statistical Systems: Crime Measurement 720
Ronet Bachman

Data Bases and Statistical Systems: Culture 727
Gunnar Otte and David Binder

Data Bases and Statistical Systems: Demography 735
Michaela Kreyenfeld and Frans Willekens

Data Bases and Statistical Systems: Economics (General) 742
F Thomas Juster

Data Bases and Statistical Systems: Economics (Macroeconomics) 751
Thomas P Hill

Data Bases and Statistical Systems: Economics (Statistical Systems) 758
Jacob Ryten

Data Bases and Statistical Systems: Education, General 763
Andreas Schröer and Richard B Händel

Data Bases and Statistical Systems: Education, Statistical Systems 769
Tom Snyder

Data Bases and Statistical Systems: Environment and Urban Studies 775
Erika Nóra Regös

Data Bases and Statistical Systems: Health 781
Ilona Kickbusch and Claudia Stein

Data Bases and Statistical Systems: International Comparative Research 789
Markus Quandt and Ruud Luijkx

Data Bases and Statistical Systems: Labor and Work 798
Constance Sorrentino

Data Bases and Statistical Systems: Law 803
Ryan Harrington

Databases and Statistical Systems: Linguistics 811
Christopher Cieri

Data Bases and Statistical Systems: Overview 817
Shawna N Smith and Aaron Ponce

Data Bases and Statistical Systems: Political Behavior and Elections 824
Rüdiger Schmitt-Beck

Data Bases and Statistical Systems: Political Science (General) 830
Max Kaase

Data Bases and Statistical Systems: Security and Conflict 836
Thomas Wencker, Christoph Trinn, and Aurel Croissant

Data Bases and Statistical Systems: Sociology 844
Jürgen HP Hoffmeyer-Zlotnik and Uwe Warner

Data Collection: Interviewing 851
Albert Hunter

de Finetti, Bruno (1906–85) 857
Bruce Hill

Death and Dying, Psychology of 860
Mary A Varga

Death and Dying, Sociology of 865
Tony Walter

Death, Anthropology of 870
Henry Abramovitch

Death Penalty 874
Benjamin Fleury-Steiner and Ross Kleinstuber

Decision and Choice: Bounded Rationality 879
Roy Radner

Decision and Choice: Economic Psychology 886
Stephen EG Lea

Decision and Choice: Heuristics 892
Henrik Olsson and Mirta Galesic

Decision and Choice: Luce's Choice Axiom 895
Timothy J Pleskac

Decision and Choice: Random Utility Models of Choice and Response Time 901
Hans Colonius and Anthony AJ Marley

Decision and Choice: Sequential Decision Making 906
Adele Diederich

Decision Making: Nonrational Theories 911
Gerd Gigerenzer and Wolfgang Gaissmaier

Decision Making, Psychology of 917
Joop van der Pligt

Declarative Memory, Neural Basis of 923
Elizabeth A Buffalo and Larry R Squire

Deconstruction 927
Jonathan Culler

Deductive Reasoning Systems 933
Wolfgang Bibel and Christoph Kreitz

CONTRIBUTORS TO VOLUME 5

Henry Abramovitch
Tel Aviv University, Tel Aviv, Israel

Soon Ang
Nanyang Technological University, Singapore

Darrell P. Arnold
St. Thomas University, Miami Gardens, FL, USA

Bruce A. Arrigo
University of North Carolina at Charlotte, Charlotte, NC, USA

Oyedeji Ayonrinde
Bethlem Royal Hospital, South London and Maudsley NHS Foundation Trust, Beckenham, UK

Ronet Bachman
University of Delaware, Newark, DE, USA

Nina Bandelj
University of California, Irvine, CA, USA

Les Beldo
University of Chicago, Chicago, IL, USA

Peter Bellwood
Australian National University, Canberra, ACT, Australia

Mabel Berezin
Cornell University, Ithaca, NY, USA

Heather Y. Bersot
University of North Carolina at Charlotte, Charlotte, NC, USA

Kish Bhatti-Sinclair
University of Chichester, Chichester, UK

Dinesh Bhugra
Institute of Psychiatry, King's College London, London, UK

Wolfgang Bibel
Technische Universität Darmstadt, Darmstadt, Germany

David Binder
Institute of Sociology, Johannes Gutenberg University Mainz, Mainz, Germany

Jörg Blasius
Institute for Political Science and Sociology, University of Bonn, Bonn, Germany

Anthony E. Boardman
University of British Columbia, Vancouver, BC, Canada

Ronald L. Breiger
University of Arizona, Tucson, AZ, USA

Christoph Brumann
Max Planck Institute for Social Anthropology, Halle, Germany

Sandra M. Bucerius
University of Alberta, Edmonton, AB, Canada

Elizabeth A. Buffalo
Washington National Primate Research Center, Seattle, WA, USA

Roger Chartier
Ecole des Hautes Etudes en Sciences Sociales, Paris, France

Roland Chilton
University of Massachusetts Amherst, Amherst, MA, USA

Ming M. Chiu
Purdue University, West Lafayette, IN, USA

Christopher Cieri
Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA

Karen E. Clayton
Fordham University, New York, NY, USA

Hans Colonius
Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

Aurel Croissant
University of Heidelberg, Heidelberg, Germany

Thomas Crosbie
Yale University, New Haven, CT, USA

Charles H.G. Crothers
School of Social Sciences and Public Policy, AUT University, Auckland, New Zealand

Jody Culham
Mind Institute and Brain, University of Western Ontario, London, ON, Canada

Jonathan Culler
Cornell University, Ithaca, NY, USA

Samantha R. Cumley
University of Michigan-Flint, Flint, MI, USA

Stijn Daenekindt
Ghent University, Ghent, Belgium

O. Dascher
Nordrhein—Westfalen State Archives, Düsseldorf, Germany

Stacy De Coster
North Carolina State University, Raleigh, NC, USA

Eelke de Jong
Radboud University, Nijmegen, The Netherlands

Mary Sully de Luque
Thunderbird School of Global Management, Glendale, AZ, USA

Adele Diederich
Jacobs University Bremen, Bremen, Germany

Joseph F. Donnermeyer
School of Environment and Natural Resources, The Ohio State University, Columbus, OH, USA

Alain Drouard
National Center for Scientific Research, Paris, France

Vincent Dubois
Université de Strasbourg, Strasbourg cedex, France

Susan A. Dumais
Lehman College, City University of New York, Bronx, NY, USA

Michael E. Ezell
Northern Illinois University, DeKalb, IL, USA

Ignacio Farías
WZB Berlin Social Science Center, Berlin, Germany

Katie A. Farina
Cabrini College, Radnor, PA, USA

Malcolm M. Feeley
School of Law, University of California at Berkeley, Berkeley, CA, USA

Marcus W. Feldman
Stanford University, Stanford, CA, USA

Gerry Ferguson
University of Victoria, Victoria, BC, Canada

James W Fernandez
University of Chicago, Chicago, IL, USA

Benjamin Fleury-Steiner
University of Delaware, Newark, DE, USA

Laurel Fogarty
Stanford University, Stanford, CA, USA

Jan Fook
Kingston University, Surrey, UK; and University of London, London, UK

Volker Gadenne
Johannes Kepler University Linz, Linz-Auhof, Austria

M. Gailus
Technical University of Berlin, Berlin, Germany

Wolfgang Gaissmaier
University of Konstanz, Konstanz, Germany

Mirta Galesic
Max Planck Institute for Human Development, Berlin, Germany

Rosemary Gartner
University of Toronto, Toronto, ON, Canada

Joseph M. Gasper
Abt Associates, Bethesda, MD, USA

Charles J. Gelso
University of Maryland, College Park, MD, USA

Michael T. Ghiselin
California Academy of Sciences, San Francisco, CA, USA

Gerd Gigerenzer
Max Planck Institute for Human Development, Berlin, Germany

Katherine Giuffre
Colorado College, Colorado Springs, CO, USA

Esther N. Goody
Cambridge University, Cambridge, UK

Robert W. Gordon
Yale University Law School, New Haven, CT, USA

Thomas J. Green
Arkansas Archeological Survey, Fayetteville, AR, USA

Michael Greenacre
Universitat Pompeu Fabra, Barcelona, Spain; and Barcelona Graduate School of Economics, Barcelona, Spain

M. Annette Grove
Utah State University, Logan, UT, USA

Carlo Guarnieri
University of Bologna, Bologna, Italy

Michael Gunder
University of Auckland, Auckland, New Zealand

Horst Gundlach
Adolf-Würth-Zentrum für Geschichte der Psychologie, Würzburg, Germany

Alan Hájek
Australian National University, Canberra, ACT, Australia

Katherine Hancock
University of Alberta, Edmonton, AB, Canada

Richard B. Händel
Evangelische Hochschule Darmstadt, Darmstadt, Germany

Ahmed Hankir
National Institute for Health Research Academic Clinical Fellow in Psychiatry, Manchester University, Manchester, UK

Kevin Hannam
International Centre for Research in Events, Tourism and Hospitality, Carnegie Faculty, Leeds Beckett University, Leeds, UK

Ryan Harrington
Yale Law School, New Haven, CT, USA

Angela P. Harris
University of California, Davis, CA, USA

Keith Hart
London School of Economics, London, UK; and University of Pretoria, Pretoria, South Africa

Karen Heimer
University of Iowa, Iowa City, IA, USA

Kathryn A. Henderson
Westat, Rockville, MD, USA

Joshua A. Hendrix
North Carolina State University, Raleigh, NC, USA

David Herbert
Swansea University, Swansea, UK

Bruce Hill
University of Michigan, Ann Arbor, MI, USA

Thomas P. Hill
University of East Anglia, Norwich, UK; and National Accounts and Economic Statistics, OECD, Paris

Jürgen H.P. Hoffmeyer-Zlotnik
University of Giessen, Germany

Malcolm D. Holmes
University of Wyoming, Laramie, WY, USA

Lorine A. Hughes
University of Nebraska at Omaha, Lincoln, NE, USA

Gail S. Humiston
University of Central Florida, Orlando, FL, USA

Albert Hunter
Northwestern University, Evanston, IL, USA

Susanne Janssen
Erasmus University Rotterdam, DR Rotterdam, The Netherlands

Sung W. Joh
Seoul National University, Seoul, Korea

Michael D. Johnson
University of Michigan, Ann Arbor, MI, USA

Peter K. Jonason
University of Western Sydney, Milperra, NSW, Australia

Christine Jourdan
Concordia University, Montreal, QC, Canada

F. Thomas Juster
University of Michigan, Ann Arbor, MI, USA

Max Kaase
University of Mannheim, Mannheim, Germany

Malin Källberg-Shroff
New School for Social Research, New York, NY, USA

Nobuko Kawashima
Doshisha University, Kyoto, Japan

Michael G. Kenward
London School of Hygiene and Tropical Medicine, London, UK

Ilona Kickbusch
Graduate Institute of International and Development Studies, Genève, Switzerland

David S. Kirk
University of Texas at Austin, Austin, TX, USA

Janne Kivivuori
National Research Institute of Legal Policy, Helsinki, Finland

Ross Kleinstuber
University of Pittsburgh — Johnstown, Johnstown, PA, USA

Victoria C.P. Knowland
Division of Language and Communication Science, City University London, London, UK

Robin M. Kowalski
Clemson University, Clemson, SC, USA

Waud Kracke
University of Illinois, Chicago, IL, USA

Christoph Kreitz
Institut für Informatik, Universität Potsdam, Potsdam, Germany

Michaela Kreyenfeld
Hertie School of Governance, Berlin, Germany

David F. Lancy
Department of Sociology, Utah State University, Logan, UT, USA

Ralph W. Larkin
The City University of New York, New York, NY, USA

Stephen E.G. Lea
University of Exeter, Exeter, UK

Jozefien De Leersnyder
Center for Social and Cultural Psychology, University of Leuven, Leuven, Belgium

Ariel Lelchook
Gettysburg College, Gettysburg, PA, USA

Jennifer C. Lena
Columbia University, New York, NY, USA

Omar Lizardo
University of Notre Dame, Notre Dame, IN, USA

Ruud Luijkx
Tilburg University, Tilburg, the Netherlands; and University of Trento, Trento, Italy

Anthony A.J. Marley
University of Victoria, Victoria, BC, Canada

Patricia L. McCall
North Carolina State University, Raleigh, NC, USA

Candace McCoy
John Jay College and the Graduate Center, City University of New York, New York, NY, USA

Alex Mesoudi
Durham University, Durham, UK

Batja Mesquita
Center for Social and Cultural Psychology, University of Leuven, Leuven, Belgium

James P. Middleton
University of Western Sydney, Milperra, NSW, Australia

Terrance Miethe
University of Nevada Las Vegas, Las Vegas, NV, USA

Joan G. Miller
New School for Social Research, New York, NY, USA

Juan J. Morales
CONICET, Argentina

Manuel R. Moreno Fraginals[†]

Paul J. Morgan
University of California, Irvine, CA, USA

Christine Morley
University of the Sunshine Coast, Maroochydore DC, QLD, Australia

Clayton Mosher
Washington State University Vancouver, Vancouver, WA, USA

Patricia M.R. Mosquera
Wesleyan University, Middletown, CT, USA

Hans-Peter Müller
Humboldt-University Berlin, Berlin, Germany

Alina Mungiu-Pippidi
Hertie School of Governance, Berlin, Germany

Carolyn B. Murray
University of California Riverside, Riverside, CA, USA

Sophie Mützel
University of Lucerne, Lucerne, Switzerland

Juan J. Navarro
UNCUYO, Argentina

Henrik Olsson
Max Planck Institute for Human Development, Berlin, Germany

[†] Deceased.

Myron Orleans
California State University, Fullerton, CA, USA

Gunnar Otte
Institute of Sociology, Johannes Gutenberg University Mainz, Mainz, Germany

Andrew V. Papachristos
Yale University, New Haven, CT, USA

Teresita Pedraza Moreno
Miami Dade College, Miami, FL, USA

Richard A. Peterson
Vanderbilt University, Nashville, TN, USA

Andrew Pickering
University of Exeter, Exeter, UK

Timothy J. Pleskac
Michigan State University, East Lansing, MI, USA

Aaron Ponce
Indiana University, Bloomington, IN, USA

Ype H. Poortinga
Tilburg University, Tilburg, Netherlands

Roberto Hugh Potter
University of Central Florida, Orlando, FL, USA

Stathis Psillos
University of Athens, Athens, Greece

Kyle Puetz
University of Arizona, Tucson, AZ, USA

Harry Purser
Kingston University, London, UK

Markus Quandt
GESIS Leibniz-Institute for the Social Sciences, Cologne, Germany

Roy Radner
Stern School of Business, New York University, NY, USA

Larry Ray
University of Kent, Canterbury, UK

Erika Nóra Regös
Max-Weber-Institute for Sociology, Heidelberg, Germany

Alison Dundes Renteln
University of Southern California, Los Angeles, CA, USA

Matthias Revers
Department of Sociology, University of Graz, Graz, Austria

Thomas Rockstuhl
Nanyang Technological University, Singapore

Henk Roose
Ghent University, Ghent, Belgium

Joseph Rouse
Wesleyan University, Middletown, CT, USA

Mark A. Runco
University of Georgia, Athens, GA, USA

Michael Ruse
Florida State University, Tallahassee, FL, USA

Jacob Ryten
Federal Government of Canada, London, ON, Canada

Joachim J. Savelsberg
University of Minnesota, Minneapolis, MN, USA

Andrew Sayer
Lancaster University, Lancaster, UK

Rüdiger Schmitt-Beck
University of Mannheim, Mannheim, Germany

Andreas Schmitz
Institute for Political Science and Sociology, University of Bonn, Bonn, Germany

Andreas Schröer
Evangelische Hochschule Darmstadt, Darmstadt, Germany

Michael Scroggins
Teachers College, Columbia University, Oakland, CA, USA

Wolfgang Seufert
Friedrich-Schiller-University Jena, Jena, Germany

Susan P. Shapiro
American Bar Foundation, Chicago, IL, USA

James F. Short
Washington State University, Pullman, WA, USA

Richard A. Shweder
University of Chicago, Chicago, IL, USA

Roger Smith
Durham University, Durham, UK

Shawna N. Smith
Institute for Social Research, University of Michigan, Ann Arbor, MI, USA

Tom Snyder
Chantilly, VA, USA

Constance Sorrentino
US Bureau of Labor Statistics, Washington, DC, USA

Maximo Sozzo
Universidad Nacional del Litoral, Santa Fe Province, Argentina

Richard Sparks
University of Edinburgh, Scotland, UK

Larry R. Squire
Veterans Affairs San Diego Healthcare System, San Diego, CA, USA and University of California, La Jolla, CA, USA

Claudia Stein
World Health Organization Regional Office for Europe, Copenhagen, Denmark

Sarah E. Sumners
University of Georgia, Athens, GA, USA

Mriganka Sur
Massachusetts Institute of Technology, Cambridge, MA, USA

Hao Tan
University of Newcastle, Callaghan, NSW, Australia

Mei Ling Tan
Nanyang Technological University, Singapore

Stephen C. Thaman
Saint Louis University School of Law, St. Louis, MO, USA

Michael S.C. Thomas
Birkbeck College, London, UK

Patricia H. Thornton
Duke University, Durham, NC, USA

Christoph Trinn
University of Heidelberg, Heidelberg, Germany

Ruth Triplett
Old Dominion University, Norfolk, VA, USA

Joop van der Pligt[†]
University of Amsterdam, Amsterdam, The Netherlands

Fons van de Vijver
Tilburg University, Tilburg, The Netherlands; North-West University, Potchefstroom, South Africa; and University of Queensland, Brisbane, QLD, Australia

Fons van de Vijver
Tilburg University, Tilburg, The Netherlands

Hervé Varenne
Teachers College, Columbia University, New York, NY, USA

Mary A. Varga
University of West Georgia, Carrollton, GA, USA

Marc Verboord
Erasmus University Rotterdam, DR Rotterdam, The Netherlands

Nathalie Vissers
Center for Social and Cultural Psychology, University of Leuven, Leuven, Belgium

Tony Walter
University of Bath, Bath, UK

Vivian Ota Wang
National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

Huon Wardle
University of St Andrews, St Andrews, UK

Uwe Warner
Centre d'Etudes de Populations, de Pauvreté et de Politiques Socio-Economiques, Luxembourg

Mark Warr
University of Texas at Austin, Austin, TX, USA

Charles F. Wellford
University of Maryland, College Park, MD, USA

Thomas Wencker
University of Heidelberg, Heidelberg, Germany

Sheila Whiteley
University of Salford, Greater Manchester, UK; Bader International Study Centre, Queen's University, Kingston, ON, Canada; and Hailsham, Sussex, UK

Elizabeth Whittaker
Clemson University, Clemson, SC, USA

Frans Willekens
Max Planck Institute for Demographic Research, Rostock, Germany

Helena Wulff
Stockholm University, Stockholm, Sweden

Brian T. Yates
Department of Psychology, American University, Washington, DC, USA

Yue Zhang
University of Illinois at Chicago, Chicago, IL, USA

Sandy Y. Zhao
Yale University, New Haven, CT, USA

Akane Zusho
Fordham University, New York, NY, USA

[†] Deceased.

Correspondence Analysis

Michael Greenacre, Universitat Pompeu Fabra, Barcelona, Spain and Barcelona Graduate School of Economics, Barcelona, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Correspondence analysis (CA) is applicable to data in the form of rectangular tables, where the entries are nonnegative measures of association between the row and column entities. The primary example of a table suitable for CA is a cross-tabulation, or contingency table, although the method extends smoothly to the analysis of almost any table of nonnegative numbers measured on the same scale, where the relative values in each row and in each column are of interest. The results of CA are one or more sets of scale values for the rows and columns, values that have a geometric interpretation leading to visualizations of the similarities between rows and between columns, as well as the row–column associations. Important variants of CA are multiple correspondence analysis, applicable to multivariate respondent-level categorical data, and canonical correspondence analysis, where the CA solution is linearly constrained by external explanatory variables.

Historical Background

In the early 1960s, a dedicated group of French researchers, led by the extraordinary scientist and philosopher Jean-Paul Benzécri, developed methods for structuring and interpreting large sets of complex data. This group's method of choice was Correspondence analysis (CA), a method for transforming a rectangular table of data, usually frequency counts, into a visual map that displays rows and columns of the table with respect to continuous underlying dimensions.

Benzécri's contribution to data analysis in general and to CA in particular was not so much in the mathematical theory underlying the methodology, known for some time already, as in the strong attention paid to the graphical interpretation of the results and in the broad applicability of the methods to data in many contexts. His initial interest was in analyzing large sparse matrices of word counts in linguistics, but he soon realized the power of the method in fields as diverse as biology, archaeology, physics, and music. The fact that his approach paid so much attention to the visualization of data, to be interpreted and verbalized with ingenuity and insight into the substantive problem, fitted perfectly the *esprit géometrique* of the French and their tradition of visual abstraction and creativity. Originally working in Rennes in Western France, this group consolidated in Paris in the 1970s to become an influential and controversial movement in post-1968 France. In 1973, they published the two fundamental volumes of *L'Analyse des Données* (Data Analysis), the first on *La Classification*, that is, unsupervised classification or cluster analysis, and the second on *L'Analyse des Correspondances*, or CA (Benzécri, 1973). From 1976 until 1997 they published the journal *Les Cahiers de l'Analyse des Données* (available at www.numdam.org), which further reflected the depth and diversity of Benzécri's work. The first books in English explaining this approach to multivariate data analysis were by Lebart et al. (1984) and Greenacre (1984). For a more complete historical account of the origins of CA, see Greenacre (1984), Gifi (1990), and the first part of the book edited by Blasius and Greenacre (2014).

Correspondence Analysis

CA is a variant of principal component analysis (PCA) applicable to categorical data rather than interval-level measurement data (*see* Factor Analysis and Latent Variable Models in Personality Psychology). For example, **Table 1** is a contingency table obtained from the 2009 International Social Survey Program (ISSP) survey on social inequality (ISSP Research Group, 2012), tabulating responses from 41 countries on the question: 'What description describes the society in your country best?' (1) A small elite at the top, very few people in the middle, and the great mass of people at the bottom; (2) A society like a pyramid with a small elite at the top, more people in the middle, and the most at the bottom; (3) A pyramid except that just a few people are at the bottom; (4) A society with most people in the middle; (5) Many people near the top, and only a few near the bottom.

Of interest in CA are the *profiles* of the rows and columns of the table, defined as the relative frequencies in the rows (i.e., row entries divided by their respective row sums) and the relative frequencies in the columns (i.e., column entries divided by their respective row sums). For example, the profile of Argentina, which has row total (i.e., sample size) 1133, is $506/1133 = 0.447$, $400/1133 = 0.353$, etc., while that of Austria, with row total 1019, is $160/1019 = 0.157$, $243/1019 = 0.238$, etc. Similarly, the column profile of Type A, with column total 14 151, is $506/14\,151 = 0.0358$, $160/14\,151 = 0.0113$, etc.

The theory and computing algorithm of CA can be summarized by the following steps:

1. Let $\mathbf{N}$ be the $I \times J$ table with grand total n and let $\mathbf{P} = (1/n)\mathbf{N}$ be the *correspondence matrix*, with grand total equal to 1. If $\mathbf{N}$ is a contingency table, then $\mathbf{P}$ is an observed discrete bivariate distribution.
2. Let $\mathbf{r}$ and $\mathbf{c}$ be the vectors of row and column sums of $\mathbf{P}$, respectively, and $\mathbf{D}_r$ and $\mathbf{D}_c$ diagonal matrices with $\mathbf{r}$ and $\mathbf{c}$ on the diagonal. The elements of $\mathbf{r}$ and $\mathbf{c}$ are called row and column *masses*, and are also the average, or expected, column and row profiles, respectively. For example, $\mathbf{r}$ can be equivalently defined as the column sums divided by the

Table 1 Response frequencies from 41 countries to the same question

		Type A	*Type B*	*Type C*	*Type D*	*Type E*	*Cannot choose*
AR	Argentina	506	400	104	78	21	24
AT	Austria	160	243	284	207	23	102
AU	Australia	89	431	318	595	30	15
BE	Belgium	71	360	245	335	31	49
BG	Bulgaria	598	256	52	31	5	55
CH	Switzerland	78	290	293	466	43	59
CL	Chile	354	698	191	170	42	45
CN	China	644	1509	359	357	66	75
CY	Cyprus	43	227	544	123	10	53
CZ	Czech Republic	360	408	215	157	24	41
DE-E	Former E. Germany	92	158	78	63	18	30
DE-W	Former W. Germany	144	286	211	170	35	110
DK	Denmark	23	154	367	847	51	48
EE	Estonia	313	447	92	94	14	45
ES	Spain	185	454	237	188	39	57
FI	Finland	58	198	273	298	10	23
FR	France	444	1454	443	329	44	36
GB	Great Britain	137	385	173	192	32	27
HR	Croatia	617	310	69	59	20	119
HU	Hungary	537	307	57	35	13	42
IL	Israel	204	613	168	102	12	74
IS	Iceland	90	173	181	442	42	13
IT	Italy	338	428	133	122	22	29
JP	Japan	130	452	310	235	47	115
KR	South Korea	303	556	409	240	79	12
LV	Latvia	723	214	57	25	39	11
NO	Norway	29	152	333	795	100	28
NZ	New Zealand	59	300	232	303	15	13
PH	The Philippines	371	476	132	118	80	23
PL	Poland	437	389	160	149	44	83
PT	Portugal	339	301	104	54	33	160
RU	Russia	565	486	174	116	47	215
SE	Sweden	77	251	321	409	20	36
SI	Slovenia	245	292	252	114	25	101
SK	Slovakia	468	420	91	79	15	72
TR	Turkey	565	507	169	126	45	45
TW	Taiwan	364	733	522	318	47	42
UA	Ukraine	1264	404	88	43	30	178
US	The United States	200	454	175	304	34	372
VE	Venezuela	301	346	161	101	49	26
ZA	South Africa	1626	1017	278	212	69	83

grand total: in this example, the mass of the first column (Type A) is $c_1 = 14\,151/54\,597 = 0.259$; hence Argentina has higher than average frequency of Type A, whereas Austria has lower than average, and so on for the other columns.

3. Compute the *singular value decomposition* (SVD) of the centered and normalized matrix, with general element $s_{ij} = (p_{ij} - r_i c_j)/\sqrt{r_i c_j}$:

$$\mathbf{S} = \mathbf{D}_r^{-1/2}\left(\mathbf{P} - \mathbf{r}\mathbf{c}^{\mathrm{T}}\right)\mathbf{D}_c^{-1/2} = \mathbf{U}\mathbf{D}_\sigma\mathbf{V}^{\mathrm{T}} \qquad [1]$$

where the left and right singular vectors are orthonormal: $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$, and the singular values in the diagonal matrix $\mathbf{D}_\sigma$ are in descending order.

4. Compute the *standard coordinates* $\mathbf{X}$ and $\mathbf{Y}$ of the rows and columns:

$$\mathbf{X} = \mathbf{D}_r^{-1/2}\mathbf{U} \quad \mathbf{Y} = \mathbf{D}_c^{-1/2}\mathbf{V} \qquad [2]$$

and respective *principal coordinates* $\mathbf{F}$ and $\mathbf{G}$:

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\sigma = \mathbf{X}\mathbf{D}_\sigma \quad \mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_\sigma = \mathbf{Y}\mathbf{D}_\sigma \qquad [3]$$

The respective row and column *contribution coordinates* are simply the matrices of left and right singular vectors $\mathbf{U}$ and $\mathbf{V}$, i.e., in terms of the standard coordinates:

$$\mathbf{U} = \mathbf{D}_r^{1/2}\mathbf{X} \quad \mathbf{V} = \mathbf{D}_c^{1/2}\mathbf{Y} \qquad [4]$$

The total variance, called *inertia*, is equal to the sum of squares of the matrix decomposed in [1]:

$$\text{inertia} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij} - r_i c_j\right)^2}{r_i c_j} = \sum_{i=1}^{I}\sum_{j=1}^{J} r_i c_j\left(\frac{p_{ij}}{r_i c_j} - 1\right)^2 \qquad [5]$$

which is equal to 0.370 in this example. The inertia is identical to the Pearson chi-squar statistic calculated on the original table divided by n (*see* Multivariate Analysis: Discrete Variables (Overview)). The squared singular values $\sigma_1^2, \sigma_2^2, \cdots$, called the *principal inertias*, decompose the inertia into parts attributable to the respective principal axes, in descending order of importance, just as in PCA the total variance is decomposed along principal axes. In this example, the principal inertias have values (and percentages of the total) equal to 0.238 (64.4%), 0.0557 (15.0%), 0.0461 (12.5%), 0.0241 (6.5%), and 0.0060 (1.6%) – the number of principal inertias is one less than the number of rows or columns, whichever is smaller.

Apart from their several algebraic properties, the coordinates are used to represent the rows and columns in joint plots, where the interpretation differs depending on the choice of coordinates. The principal coordinates are the projections of the profiles onto principal axes that explain maximum inertia. These projected profile points are equivalently interpreted as approximating *chi-square distances* between the profiles, for example, between the profiles of rows i and i':

$$d(i, i') = \sqrt{\sum_{j=1}^{J}\frac{\left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}}\right)^2}{c_j}} \qquad [6]$$

The total inertia of the table can then be written as the weighted average of the squared chi-square distances of the profiles to their average, or centroid, for example, for the row profiles:

$$\text{inertia} = \sum_{i=1}^{I} r_i[d(i, c)]^2 = \sum_{i=1}^{I} r_i \sum_{j=1}^{J}\frac{\left(\frac{p_{ij}}{r_i} - c_j\right)^2}{c_j}$$

The standard coordinates are convenient reference points for the interpretation, being the projections onto the principal axes of the unit profiles $[1 \quad 0 \quad 0 \quad \cdots \quad 0]$, $[0 \quad 1 \quad 0 \quad \cdots \quad 0]$, etc. The contribution coordinates, which are a rescaled version of the standard coordinates, are related to the contributions of each row or column to the principal axes (see below). When the substantive variable, in this case the column classification,

has few categories, one of the most popular displays is the *symmetric map*, where both rows and columns are displayed in principal coordinates, i.e., using the first two columns of **F** and **G** for the respective row and column coordinates in a two-dimensional display, as shown in Figure 1.

Types A to D form a curved pattern from left to right, with Latvia and Ukraine the most extreme on the left, i.e., the highest proportions of respondents saying their societies contain a small elite, few in the middle and most at the bottom. Denmark and Norway are at extreme right, i.e., the highest proportions saying their societies have most people in the middle and less at the extremes. Thus the horizontal position of the countries, accounting for almost two-thirds (64.4%) of the inertia, scales them in terms of how egalitarian the respondents view their societies, while the vertical position, accounting for 15%, contrasts Types B and C against the others. For example, France and the United States are near average (the center) horizontally, but for different reasons: 52.9% of French respondents opt for Type B (pyramid structure of society), while only 29.5% of US respondents do, the average across all countries being 32.9%. Not obvious from Figure 1 is that 24.2% of US respondents 'cannot choose,' whereas in France it is only 1.3% – the 'cannot choose' category dominates the third dimension of the CA, explaining an additional 12.5% of the inertia, which can be seen in the three-dimensional solution shown as a *video* in the online supplementary material.

Alternative choices of coordinates give slightly different interpretations of the results. In the *asymmetric map* of the rows, for example, also called the *row-principal biplot*, rows are shown in principal coordinates and columns in standard coordinates (vice versa for the column-principal biplot). The choice between a row-principal or column-principal biplot is governed by whether the original table is considered as a set of rows or a set of columns, when expressed in percentage form – usually, as in this example, the substantive categories are the columns, and the row-principal option would be the natural choice. Choosing contribution coordinates instead of standard coordinates in a row-principal biplot is a convenient alternative when there are many column categories, as is the case in community ecological applications, where samples (rows) are described by very many species (columns). The resultant *contribution biplot* shows the researcher exactly which columns are determinant for interpreting the differences between samples. Note that in CA the row and column profiles are weighted by their corresponding masses, which affect the solution as well as the interpretation – thus the low-frequency Type E category receives low weight in determining the optimal solution. In this example, the row (country) masses are proportional to the sample sizes, so that South Africa, for example, with a sample size of 3285, obtains three times the weight as Argentina. If it is preferred to give each country the same weight, this is achieved simply by transforming the table to row proportions, or row percentages, before applying the CA algorithm, to force all row totals to be the same while not affecting the row profiles.

CA can be equivalently considered a *bilinear model* for the observed frequencies, since the SVD in [1] may be written in terms of the standard coordinates in the following equivalent form, for the (i, j)-th element $p_{ij} = n_{ij}/n$ of the correspondence matrix:

$$p_{ij} = r_i c_j \left(1 + \sigma_1 x_{i1} y_{j1} + \sigma_2 x_{i2} y_{j2} + \cdots\right) \qquad [7]$$

(see Multivariate Analysis: Discrete Variables (Correspondence Models)). There are as many terms $\sigma_k x_{ik} y_{jk}$ in [7] as there are dimensions, the maximum number of dimensions. For a model with two dimensions, for example, retaining $k = 1$ and 2, the residual elements have been minimized by weighted least-squares.

Contributions to Inertia

The overall quality of the CA solution is judged by the percentages of inertia, similar to a regression analysis: for example, in Figure 1, 79.4% of the inertia (variance) is explained by the two-dimensional solution, and in the three-dimensional solution in the supplementary material an additional 12.5% is

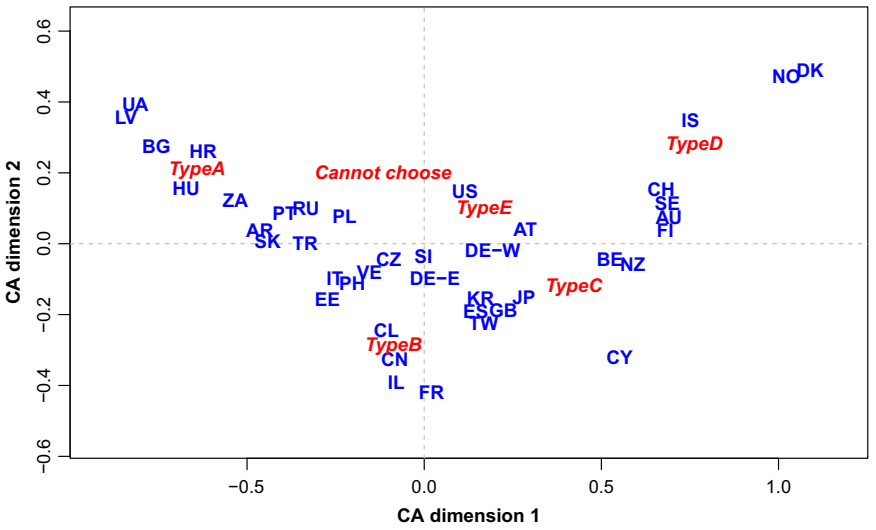

Figure 1 Correspondence analysis of Table 1: two-dimensional symmetric map showing rows and columns in principal coordinates; 79.4% of the inertia explained.

explained, bringing the overall percentage to 91.9%, with only 8.1% unexplained inertia. More detailed diagnostics in CA can be obtained in the form of the so-called *contributions to inertia*, based on the decompositions of the total inertia in [5] according to individual rows or individual columns, dimension by dimension:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij}-r_ic_j\right)^2}{r_ic_j}=\sum_{k=1}^{K}\sigma_k^2=\sum_{i=1}^{I}\sum_{k=1}^{K}r_if_{ik}^2=\sum_{j=1}^{J}\sum_{k=1}^{K}c_jg_{jk}^2 \quad [8]$$

For example, every column component $c_jg_{jk}^2$ can be expressed relative to the principal inertia $\sigma_k^2=\sum_j c_jg_{jk}^2$ of the corresponding dimension k. These relative values provide a diagnostic for deciding which columns (or in a similar way, which rows) are important in the determination of the kth principal axis. In fact, these relative values, summing to 1, are exactly the squared contribution coordinates defined previously. An alternative diagnostic is to express these elemental components $c_jg_{jk}^2$ relative to their sum over the dimensions, $\sum_k c_jg_{jk}^2$ which is just $g_{jk}^2/\sum_k g_{jk}^2$ – these values are analogous to the squared factor loadings in factor analysis, that is, squared correlations between the column category and the corresponding principal axis or factor (*see* Factor Analysis and Latent Variable Models in Personality Psychology). For the above application, these values for Type E are 0.108, 0.043, and 0.016 on dimensions 1–3, which means that only 0.166 (16.6%) of the variance of this column is explained by the first three dimensions, considered as 'predictors' of Type E. For the category 'cannot choose,' the respective values are 0.026, 0.046, and 0.918, showing that more than 90% of its variance is explained by dimension 3.

Extensions

CA is regularly applied to analyze multiway tables, tables of preferences, ratings, as well as measurement data on ratio- or interval-level scales. Such extensions of the method conform to Benzécri's conception of CA as a universal technique for exploring many different types of data through judicious transformations or coding.

Concatenated Tables

For multiway tables, there are two approaches. The first approach is to convert the table to a flat two-way table, which is appropriate to the problem at hand. Thus, if a third variable is introduced into the example above, say 'gender of respondent,' then an appropriate way to flatten the three-way table would be to interactively code 'country' and 'gender' as a new row variable, with $41 \times 2 = 82$ categories. The table for males is thus contented beneath the table for females. For each country there would now be a female and a male point and one could compare genders and countries in this richer map. This process of *interactive coding* of the variables can continue as long as the data do not become too fragmented into category combinations with very few counts. A less informative way of concatenating tables is to juxtapose several cross-tabulations of a set of row variables and a set of column variables into a supermatrix of tables. For example, given a set of demographic variables such as age, education, and marital status, and another set of substantive categorical variables, all cross-tabulations of variable pairs composed of a demographic and a substantive variable are stacked row- and column-wise, and the whole matrix of tables subjected to a CA. It should be remembered that only pairwise relationships are displayed in the eventual visualizations.

Multiple Correspondence Analysis

Another approach to multiway data, called *multiple correspondence analysis* – also called *homogeneity analysis* (Gifi, 1990; Greenacre, 2007: Chapter 18), applies when there are several categorical variables skirting the same issue, often called 'items.' In the same ISSP survey, for example, there are 11 questions concerning the issue of 'getting ahead,' from "How important is coming from a wealthy family" to "How important is being born a man or a woman," to which respondents have to respond on a five-point scale from 'Essential' to 'Not important at all.' MCA is usually defined as the CA algorithm applied to an indicator matrix $\mathbf{Z}$, with the rows being the respondents and the columns being dummy variables for each of the categories of all the variables. The data are zeros and ones, with the ones indicating the chosen categories for each respondent. The resultant map shows each category as a point and, in principle, the position of each respondent as well. Alternatively, one can set up what is called the *Burt matrix*, $\mathbf{B}=\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$, the square symmetric table of all two-way cross-tabulations of the variables, including the cross-tabulations of each variable with itself. This matrix, named after the psychologist Sir Cyril Burt, is reminiscent of a covariance matrix and the CA of the Burt matrix can be likened to the PCA of a covariance matrix. The analysis of the indicator matrix $\mathbf{Z}$ and the Burt matrix $\mathbf{B}$ give identical standard coordinates of the category points, but slightly different scalings in the principal coordinates because the principal inertias of $\mathbf{B}$ are the squares of those of $\mathbf{Z}$. A variant of MCA called *joint correspondence analysis* (JCA) avoids the fitting of the tables on the diagonal of the Burt matrix, an approach that is analogous to least-squares factor analysis – see Greenacre (2007: Chapter 19). Notice that the Burt matrix is a concatenated matrix of pairwise relationships within a set of variables, whereas the concatenated matrix described at the end of Section Concatenated Tables contains pairwise relationships between two sets of variables.

Ratings, Rankings, and Paired Comparisons

As far as other types of data are concerned, namely rankings, ratings, paired comparisons, ratio-scale, and interval-scale measurements, the key idea is to recode the data in a form, which justifies the basic constructs of CA, namely profile, mass, and chi-squared distance. For example, in the analysis of rankings, or preferences, applying the CA algorithm to the original rankings of a set of objects by a sample of subjects is difficult to justify, because there is no reason why weight should be accorded to an object in proportion to its average ranking. A practice called *doubling* resolves the issue by adding either an 'anti-object' for each ranked object or an 'anti-subject' for each responding subject, in both cases with rankings in the reverse order – see Greenacre (2007: Chapter 23). This addition

of apparently redundant data leads to CA effectively performing different variants of PCA on the original rankings. The same idea can be applied to ratings data, for example five-point ordinal scales, usually coded 1–5. A rating of 2, for example, would be recoded as a pair of ratings with values 1 and 3, since there is one scale point on the scale to the left of the rating 2, and three scale points to the right of it. This doubled pair of variables would code opposite poles of the original scale, and the CA of the doubled table based on several rating scales is again quite similar to the PCA of the original undoubled table, the only difference being the normalization inherent in the chi-square distance.

Fuzzy Coding

There are several ways for CA to handle continuous data – one of the most versatile is to recode each continuous variable into a set of fuzzy categories. This is a generalization of the dummy variable coding of MCA, which has only zeros or ones, also called 'crisp coding.' Coding a variable into three categories such as 'low,' 'medium,' and 'high' leads to fuzzy-coded values such as $[0 \quad 0.23 \quad 0.77]$, where 0.77 indicates that the datum in question is 'mostly' in the high category but 'somewhat' in the middle category. This is a more accurate coding than crisply coding the datum as $[0 \quad 0 \quad 1]$. Fuzzy coding is achieved through the use of membership functions, which transform the original variable into a set of fuzzy categories, with the fuzzy values being nonnegative and summing to 1. Once the data have been recoded, CA is applied in the usual way – see Aşan and Greenacre (2011).

CA Can Perform MDS on a Distance Matrix

CA can also be shown to be a limiting case of multidimensional scaling (MDS), by applying it to an appropriately transformed square symmetric matrix of distances. The distances are squared and subtracted from a constant C, which is at least as large as the largest squared distance in the table (so the diagonal of the transformed matrix would consist of C's). The CA of this transformed matrix yields a solution, which, after an appropriate rescaling that depends on C, approximates the classical scaling solution of the distance matrix and tends to the exact solution as C is increased (Carroll et al., 1997).

Subset Correspondence Analysis

In some cases we would like to analyze certain subsets of rows or columns, for example, a specific set of response categories in survey research, that is, all the response categories excluding the 'missing' ones (refused to answer/not applicable/etc.). Because CA uses the table margins as weights, the simple use of a table with such categories simply omitted leads to problematic situations, which can be easily solved by maintaining the original margins of the full table whenever a subtable is analyzed – see Greenacre and Pardo (2006) and Greenacre (2007: Chapter 21).

Canonical Correspondence Analysis

CA finds the optimal dimensions of a data table, but in certain situations additional variables serving as explanatory variables are available. For example, in a social survey like the one described in Section Multiple Correspondence Analysis, the age, number of years of education, and marital status might be available for each respondent. The optimal CA dimensions can be regressed on the variables, to assess whether and how much the dimensions can be explained by these variables. A more specific analysis can be performed when the dimensions are forced to be related, usually linearly, to the explanatory variables. In other words, the researcher is only interested in the variance that can be explained by age, education, and marital status. This is achieved by first projecting the primary response data onto the space defined by the explanatory variables, and then performing the dimension reduction in this reduced space. In the CA context this is called *canonical correspondence analysis*, and is used extensively in ecology when biological community data, usually abundance counts of many species, are related to several environmental variables (ter Braak, 1986; Greenacre, 2007: Chapter 24).

The latest developments on the subject, including discussions of sampling properties of CA solutions and comprehensive reference lists, may be found in the volumes edited by Greenacre and Blasius (1994, 2008) and Blasius and Greenacre (1998, 2014).

See also: Factor Analysis and Latent Structure Analysis: Overview; Multidimensional Scaling I; Multidimensional Scaling II; Multivariate Analysis: Discrete Variables (Correspondence Models); Multivariate Analysis: Discrete Variables (Overview).

Bibliography

Aşan, Z., Greenacre, M.J., 2011. Biplots of fuzzy coded data. Fuzzy Sets and Systems 183, 57–71.

Benzécri, J-P., 1973. L'Analyse des Données. La Classification, vol. I. L'Analyse des Correspondances, vol. II. Dunod, Paris.

ter Braak, C.J.F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67, 1167–1179.

Blasius, J., Greenacre, M.J., 1998. Visualization of Categorical Data. Academic Press, San Diego, CA.

Blasius, J., Greenacre, M.J., 2014. Visualization and Verbalization of Data. Chapman & Hall/CRC Press, Boca Raton, Florida.

Carroll, J.D., Kumbasar, E., Romney, A.K., 1997. An equivalence relation between correspondence analysis and classical metric multidimensional scaling for the recovery of Euclidean distances. British Journal of Mathematical and Statistical Psychology 50, 81–92.

Gifi, A., 1990. Nonlinear Multivariate Analysis. Wiley, Chichester, UK.

Greenacre, M.J., 1984. Theory and Applications of Correspondence Analysis. Academic Press, London. Freely downloadable from: www.carme-n.org.

Greenacre, M.J., 2007. Correspondence Analysis in Practice, second ed. Chapman & Hall/CRC Press, Boca Raton, Florida. Published in Spanish as La Práctica del Análisis de Correspondencias by the BBVA Foundation, Madrid, and freely downloadable from: www.multivariatestatistics.org.

Greenacre, M.J., Blasius, J., 1994. Correspondence Analysis in the Social Sciences. Academic Press, London.

Greenacre, M.J., Blasius, J., 2008. Multiple Correspondence Analysis and Related Methods. Chapman & Hall/CRC Press, Boca Raton, Florida.

Greenacre, M.J., Pardo, R., 2006. Subset correspondence analysis: visualization of selected response categories in a questionnaire survey. Sociological Methods and Research 35, 193–218.

ISSP Research Group, 2012. International Social Survey Programme: Social Inequality IV – ISSP 2009. GESIS Data Archive, Cologne. ZA5400 Data file Version 3.0.0. http://dx.doi.org/10.4232/1.11506.

Lebart, L., Morineau, A., Warwick, K., 1984. Multivariate Descriptive Statistical Analysis. Wiley, Chichester, UK.

Correspondence Analysis, History of

Jörg Blasius and Andreas Schmitz, Institute for Political Science and Sociology, University of Bonn, Bonn, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents an overview of the basic principles and statistical foundations of correspondence analysis with a special emphasis on its historical and further developments. With respect to the social sciences, we concentrate on the work of Pierre Bourdieu and his scholars and their empirical construction of 'social spaces' and social fields.'

Introduction

Correspondence analysis (CA) is a multivariate method for analyzing categorical data, its main objective is to visualize rows and columns of a data table in a low-dimensional (usually two-dimensional) space, called *map*. CA was developed by Jean-Paul Benzécri and his collaborators in France in the 1960s. Together with multiple correspondence analysis (MCA), it is the central part of the French "Analyse des Données" (cf Benzecri et al., 1973; Greenacre, 1984, 2007; Lebart et al., 1984; Le Roux and Rouanet, 2004). In contrast to traditional statistical methods such as Ordinary Least Square (OLS) regression, CA, and MCA do not make any assumptions on the structure of data, they rather identify the structure of relations between the variables and between the objects (e.g., individuals), in survey research usually the individuals.

There is a particularly large number of applications of this method in the social sciences, the work of Pierre Bourdieu and his scholars is probably the best-known example. Whenever the discussion turns to 'social space,' the empirical foundation is usually based either on CA, or on MCA.

In this article we first discuss the basic principles of the methods, their historical development, and the underlying statistical background. In the next step we make CA accessible discussing examples of Pierre Bourdieu's work on *social spaces* and *fields*. In the final section 'Further Developments and Related Techniques' we discuss some recent developments of the method, and we conclude with a short discussion and outlook.

Principles of CA and MCA

The basic form of CA is to analyze a single ($I \times J$) cross-table (with $I =$ number of rows, $J =$ number of columns), with nonnegative entries. Using the two variables 'occupation groups' and 'education level,' input of CA could be the frequencies of the cells from the respective cross-table. The solution would be (1) a graphical display of row and column categories and (2) a numerical solution which provides, among other aspects, the correlations between manifest and latent variables, i.e., factor scores equivalent to the factor loadings, which are well known from principal component analysis (PCA).

In the example, the CA solution would allow studying the associations within the set of occupation groups with respect to education, the associations *within* the levels of education with respect to occupation and the relations between 'educational level' and 'occupation groups.' There are only a few real-life applications that are limited to only two variables, the main objective of CA (and MCA) is to visualize more complex structures of data. A simple extension of the given example is to add other variables by row- and/or column-wise concatenating of frequency tables, for example, age groups and party preferences.

To give an example, assume that we have eight social groups (g1 to g8), $N = 750$ respondents, 10 indicators on some kind of abilities (A1 to A10) with three categories each ("yes, I am able to do it" (yes), "yes, I am able to do it but just a bit" (jab), "no, I am not able to do it" (no)), the dimension of the respective matrix is 30×8.

CA and MCA allow to distinguish between *active* and *passive* (or *supplementary*) variables, While the *active* variables, for example, the 10 abilities mentioned above, are used to create the geometric orientation of the axes in the projection space, the *passive* variables do not have any influence on it, but they can be interpreted together with the *active* variables. In the given example the 10 abilities with a total of 30 variable categories are treated as *active*, leading to a *space of abilities*, some sociodemographic characteristics such as sex, marital status, age groups, and income groups can be added as *passive* variables, permitting for the joint representation together with the abilities. In this context of applying CA there is usually one variable to be described (here the eight groups) and one or more describing variables.

For reasons of simplification, we assume that there are no missing values in the data set. The structure of the respective matrix of input data is shown in Table 1.

The first 30 rows and the first eight columns (upper left block) contain the frequencies of the cross-tabulations between the 10 abilities (with categories 'yes,' 'jab,' and 'no') and the eight social groups. This 'active' part of the table will be used for constructing the space when applying CA. Since there are no missing data, the number of valid cases for each of the 10 abilities is the same: $n_{1+} + n_{2+} + n_{3+} = n_{4+} + n_{5+} + n_{6+} = \ldots = n_{28+} + n_{29+} + n_{30+} = N$.

In the French tradition of interpreting CA solutions, researchers usually use the so-called 'symmetric map' for visualizing the solution, which is the joint representation of row and column profiles. Thereby, the distances within the set of rows and within the set of columns are Euclidean ones, the distances between rows and columns are not defined; for

Table 1 Matrix of input data, stacked table, with supplementary rows and columns

		Active				*Passive*				
	Ability	*g1*	*g2*	…	*g8*	*1*	*2*	…	*750*	*Sum (g1–g8)*
Active	A1_yes	$n_{1,1}$	$n_{1,2}$	…	$n_{1,8}$	1	0	…	0	n_{1+}
	A1_jab	$n_{2,1}$	$n_{2,2}$	…	$n_{2,8}$	0	0	…	0	n_{2+}
	A1_no	$n_{3,1}$	$n_{3,2}$	…	$n_{3,8}$	0	1	…	1	n_{3+}
	A2_yes	$n_{4,1}$	$n_{4,2}$	…	$n_{4,8}$	0	0	…	1	n_{4+}
	A2_jab	$n_{5,1}$	$n_{5,2}$	…	$n_{5,8}$	1	0	…	0	n_{5+}
	A2_no	$n_{6,1}$	$n_{6,2}$	…	$n_{6,8}$	0	1	…	0	n_{6+}
	…	…	…	…	…	…	…	…	…	…
	A10_yes	$n_{28,1}$	$n_{28,2}$	…	$n_{28,8}$	0	0	…	1	n_{28+}
	A10_jab	$n_{29,1}$	$n_{29,2}$	…	$n_{29,8}$	1	0	…	0	n_{29+}
	A10_no	$n_{30,1}$	$n_{30,2}$	…	$n_{30,8}$	0	1	…	0	n_{30+}
Passive	Male	$n_{81,1}$	$n_{81,2}$	…	$n_{81,8}$	1	0	…	1	$(n_{(m+)})$
	Female	$n_{83,1}$	$n_{83,2}$	…	$n_{83,8}$	0	1	…	0	$(n_{(f+)})$
	Single	$n_{83,1}$	$n_{83,2}$	…	$n_{83,8}$	0	0	…	0	$(n_{(s+)})$
	…	…	…	…	…	…	…	…	…	$(n_{(\ldots+)})$
	Sum	10 × g1	10 × g2	…	10 × g8	(10)	(10)	(10)	(10)	10 × *N*

further details we refer to Greenacre (1984, 2007). To avoid this possible misinterpretation, rows and columns can be displayed in separate maps.

In case CA is applied to an indicator matrix (=MCA), the columns contain the variable categories (active as well as passive), and the rows enclose the objects (e.g., individuals). The cells of this matrix consist of 'ones,' which are used for the true options, and 'zeros,' which are used for the false ones. In the given example with 10 abilities containing three categories each, for each ability the ones indicate if a respondent is able to perform it (yes), if he/she is somewhat able (just a bit), or if he/she is not able to do it (no). With respect to passive variables the 'ones' indicate, among other things, if a respondent is a male or a female, if he/she is single, married, divorced, or widowed. It is evident that every variable has only one true possibility, all others are false. For example, a respondent either has a specific ability, or he/she has 'just a bit' of this ability, or not. Furthermore, a person can be either a male or a female but not both or none of them – and if there would be a third and meaningful category of gender, one could easily add another category.

The new example contains of $N = 750$ cases, the matrix to be analyzed has the dimension 750 × 30. When starting the work on CA and especially on MCA, the decomposition of such a large matrix was a sophisticated computational problem (Benzécri, 2006), so there would have been a simple reason to restrict the problem to a 30 × 8 matrix. Meanwhile processing time and storage capacity have undergone substantial improvements; nowadays every standard PC with standard statistical software is able to decompose a 750 × 30 matrix in a few seconds. In other words, today there are no computational limitations for social science applications in the field of MCA. However, as it applies to all statistical approaches, it is meaningless to include as many variables as one may like to consider, one should have a substantive theory (or a frame model, Le Roux and Rouanet, 2010, p. 10), which guides the process of data analysis. The example of the indicator matrix discussed above is given in Table 2.

Table 2 Indicator matrix, with supplementary columns

	Active (abilities)										*Passive (sociodemographic characteristics)*				
	A1			*A2*			…	*A10*			*Sex*		…		
I	*Yes*	*Jab*	*No*	*Yes*	*Jab*	*No*	…	*Yes*	*Jab*	*No*	*m*	*f*	*sin*	…	*Sum*
1	1	0	0	0	1	0	…	0	1	0	1	0	0	…	10
2	0	0	1	0	0	1	…	0	0	1	0	1	0	…	10
3	1	0	0	1	0	0	…	1	0	0	1	0	1	…	10
4	0	1	0	0	1	0	…	0	1	0	0	1	0	…	10
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	10
750	0	0	1	0	0	0	…	1	0	0	1	0	0	…	10
Sum	n_{1+}	n_{2+}	n_{3+}	n_{4+}	n_{5+}	n_{6+}	$n_{\ldots+}$	n_{28+}	n_{29+}	n_{30+}	$n_{(m+)}$	$n_{(f+)}$	$n_{(s+)}$	$n_{(\ldots+)}$	7500

Table 2 contains 10 abilities with three categories each ('yes,' 'jab,' and 'no'), and a set of supplementary variables, the first one is 'sex' with two categories, the second is 'family status.' In total there are Q (=10) active variables with a total of $J = 30$ categories and $N = 750$ individuals. The first person is male and at least not single, he mentioned that he possesses the first ability, for the second and the 10th ability he marked 'just a bit.' The second person is female, she exhibits none of the three abilities displayed in Table 2, the third person is male and single, he stated to have all three abilities. The first category (first ability, response category 'yes') has been mentioned n_{1+} times, the second category (first ability, response category 'jab') has been mentioned n_{2+} times, and so on. Since there are no missing values in Table 2, it yields: $n_{1+} + n_{2+} + n_{3+} = n_{4+} + n_{5+} + n_{6+} = \ldots = n_{28+} + n_{29+} + n_{30+} = 750$, and since there are 10 active variables, *total N* is 7500.

When applying MCA to the indicator matrix, both individuals and variable categories can be visualized in a map that can be read off like a geographic map. High positive associations between variable categories (as well as between individuals) are reflected by short distances, high negative associations by large distances. In general, the closer two categories (or two individuals) are located to each other in the projection space, the more similar they are. Note again, the distances between individuals and variable categories are not defined.

Methodological Background

Although there is a numerical representation of CA and MCA similar to PCA, the focus of interpretation in CA is traditionally based on the graphical solution that allows detecting the relationships between the variables (variable categories) of interest in a straightforward way, for example, between a set of lifestyle attributes and sociodemographic characteristics. In contrast to PCA, there is no requirement for the CA at the scale level of the data, they are typically ordered or unordered categorical.

Whereas CA is applied to a single contingency table or to a stacked table, MCA uses the same algorithm to an indicator matrix or a Burt matrix. Applying the CA algorithm to the indicator matrix ($\mathbf{Z}$), the table of input data has as many rows as cases in the data set, and the number of columns is equal to the sum of variable categories over all indicators included in the model. Thereby, a '1' indicates that the respondent chose a specific response category while a '0' indicates that he/she did not choose the specific response category. If there are no missing values in the data or if nonsubstantive responses such as "I don't know" are treated as valid responses or if listwise deletion has been applied to the variables included in the model, the row sums are equal to the number of variables while the column sums reflect the marginal of the variables.

Instead of using the indicator matrix as input to MCA it is also possible to apply the Burt matrix ($\mathbf{B}$). This matrix can be computed by multiplying the transposed indicator matrix by itself, i.e., $\mathbf{B} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}$. Alternatively, $\mathbf{B}$ can be generated by cross-tabulating all variables by all variables, including the cross-tabulations of the variables by themselves, and creating a row- and column-wise stacked table. $\mathbf{B}$ is a symmetric matrix in which the main diagonal elements contain the marginal of the respective variables, while all other elements of the main diagonal matrices are zero. The off-diagonal matrices contain the frequencies of the cross-tabulations between the respective variables. The numerical solutions of the columns from $\mathbf{Z}$ (=MCA solutions) can be converted to those of $\mathbf{B}$ by simple rescaling, the squared eigenvalues of $\mathbf{Z}$ are equal to those of $\mathbf{B}$. As it is true for PCA, MCA contains all first order interaction effects, the method can be seen as a generalization of PCA applied to unordered (or ordered) categorical data (Greenacre, 1984). The algorithm is given in Blasius and Greenacre (1994) and in Nenadić and Greenacre (2006).

Principal Investors and Early Work

CA can be traced back to the early twentieth century, especially to the works of Karl Pearson, Hermann Otto Hirschfeld (probably better known under his anglicized name H.O. Hartley), and R.A. Fisher; all three were interested in conceptualizing the association between categorical variables (Greenacre, 1984; Beh and Lombardo, 2012; Lebart and Saporta, 2014). Going back to the early beginning, Pearson (1906) already searched for 'optimal scores' that he could assign to the rows and columns of a contingency table that linearize regressions. Almost 30 years later, Hirschfeld (1935) could show for the two variable cases with I and J categories that $\min(I, J) - 1$ solutions exist for the problem mentioned by Pearson. The respective scores he found via an eigen equation are nothing more than the row and column coordinates in CA up to some standardization (Lebart and Saporta, 2014). Further, he could show that the sum of eigenvalues is equal to the well-known chi-square coefficient for two-way contingency tables divided by the total of the table, a value that is called *total inertia* in the CA framework. Some years later, Fisher (1940) obtained CA while generalizing discriminant analysis to a single categorical predictor (for more details see Lebart and Saporta, 2014).

A similar development for describing the association between continuous variables is the case of PCA, which can be traced back to Hotelling (1933) who was searching for principal axes of a multivariate normal distribution in a set of metric data. Guttman (1941) is usually mentioned as the first author on MCA, since his paper contains all the formulas that characterize MCA (cf Greenacre, 1984; Lebart and Saporta, 2014). In this framework, MCA can be understood as PCA applied to categorical data.

A different statistical approach was introduced in the late 1970s by a group of statisticians from the University of Leiden (Gifi, 1981, 1990). What they called *homogeneity analysis* (and what is implemented in SPSS), is the same as MCA. Thereby homogeneity analysis belongs to a set of scaling techniques applied to metric, ordered, and unordered categorical variables (for the historical background and some methodological details, see De Leeuw, 2014).

Pierre Bourdieu's Use of CA and MCA

In statistics, outside of France the discussion started with the textbooks by Greenacre (1984) and Lebart et al. (1984), in the

social sciences with the translation of Bourdieu's *La Distinction* (1979, in English, 1984). Thereby, visualizing data structures via CA and MCA reflects a way of thinking that is typical for the social sciences in France, which are especially associated with the name of Pierre Bourdieu. Many statisticians who worked in France in the 1970s and 1980s and published mainly in French – among others Ludovic Lebart, Brigitte Le Roux, and Henry Rouanet – were closely associated with Pierre Bourdieu and his work (Blasius and Schmitz, 2014).

The philosophy behind CA and MCA, from both the statistical and the social space point of view, can be expressed with the *second principle of data analysis* from Jean-Paul Benzécri, who pointed out that "The model should follow the data, not the inverse" (Blasius and Greenacre, 2006, p. 6). Instead of limiting the data to restrictive and subjectively formulated statistical models, they discussed relational structures in the data, following Benzécri's *first principle of data analysis* "statistics is not probability" (Greenacre, 1984, p. 10).

Bourdieu's methodology consisted in the identification of *social fields*, which are parts of the *social space*, and relationships within and between these fields. Although he was opposed to a rigid methodological fetishism, his empirical research was based on the relational construction of *social fields* and *social spaces* (Rouanet et al., 2000; Blasius and Schmitz, 2014). His research can be seen as a methodological alternative to the practice of statistical modeling – instead of carrying strong assumptions such as linear constraints on the data, he visualized and interpreted the nonlinear structure of the data. His research was guided by the following questions: (1) Does the distribution of various indicators follow a certain pattern? (2) Is there an empirical structure of positions and relationships that accommodate all objects/individuals in the data set? (3) Are there any empirically observable practices (statements, classifications, self-presentations, etc.) that correspond to the positions of the individuals in the *social space* or in parts of them? (4) How can the dynamics of the objects (individuals) and the respective classes be described? In his *social space* approach, Bourdieu usually used the lifestyle attributes as active variables, and the sociodemographic characteristics as passive ones; the latter ones were used to corroborate the interpretation of the axes estimated on the basis of the lifestyle indicators.

In *La Distinction* (Bourdieu, 1979, in English 1984) and in *Anatomie du Goût* (Bourdieu and de Saint Martin, 1976), the *social space* was constructed on the basis of the row-wise stacked frequency tables of 'occupational groups' (the column variable) times a set of (active) dichotomized lifestyle attributes as well as (passive) sociodemographic characteristics (the row variables). The first two dimensions of the solution can be interpreted as in traditional factor analytic approaches; Bourdieu (1979) labeled them *composition of economic and cultural capital* and *capital volume*. When rotating the axes by 45°, the new dimensions are called *economic capital* and *cultural capital* (for more details see Blasius and Friedrichs, 2008). By definition, 'income' (in groups) should be positively associated with *economic capital*, and the 'level of education' with *cultural capital*. Furthermore, both variables should be positively related to *capital volume*. In this approach the positions of single individuals are not directly available, since the input data contain the frequencies of the cross-tabulations between the respective variables. If one would like to include the individuals in this solution, one could add the information of each respondent as supplementary information by adding them as column variables. Bourdieu (1979, 1984) did not explain how he computed the geometric figures that cover the 'occupations,' it could be that these visualizations are approximations from the results that he obtained when he included the individuals as passive indicators.

In *Homo Academicus*, (Bourdieu, 1984, in English 1988) and in *La Patronat* (Bourdieu and Saint-Martin, 1978) Bourdieu applied MCA. In this case the columns contain the variable categories, and the rows include the single individuals. In this application he distinguished professors by their disciplines, faculty memberships, and sociodemographic characteristics within the academic field. The cells of the input data table contain 'ones' and 'zeros' only; thereby, the 'ones' are used for the true options and the 'zeros' for the false options. For example, in case of faculty membership, the ones reflect the existing membership(s), and the 'zeros' reflect the nonexisting memberships. This is the case of MCA where Bourdieu analyzed the data on the individual level instead of the aggregate level as he did in *La Distinction*.

Further Developments and Related Techniques

For more than 25 years, CA and MCA have been important methods in the social sciences, and they have been increasingly used over time (Beh and Lombardo, 2012). However, some parts of the techniques and some further developments are rarely used in the framework of *Analyse des Données*. In the following we briefly outline some helpful techniques inherent to CA and MCA as well as techniques that complement them.

First, as it has been mentioned before, *passive variables* can be represented in the projection space previously constructed with *active variables*, and jointly interpreted with them without influencing the geometric orientation of the axes. This is of particular interest for all situations where (the interrelations of) some variables are theoretically interpreted as the reference system for variables, which are treated as functions of the geometric space. Bourdieu and his scholars used this technique to illustrate their argument of the correspondence between sociodemographic characteristics and lifestyles.

Second, a widespread task in statistical analysis is the classification of objects into categories. Although the factor scores of MCA solutions can be used as input for clustering techniques (Lebart, 1994), this is seldom used. As a kind of *heuristic clustering*, Bourdieu (1979, 1984), for example, highlighted social classes in his *social space* by simply adding circles and rectangles without statistical justification. Apart from a subsequent clustering of factor scores (Lebart, 1994), it is also possible to derive and map *concentration ellipses* (Le Roux and Rouanet, 2010, p. 69; Rouanet, 2006, p. 146). Concentration ellipses are geometrical representations of the concentration around the mean points of a priori specified subclouds (such as the concentration around the two centroids of gender). This technique is useful when a certain variable is assigned causal primacy and is thus interpreted as a 'structuring factor' (Rouanet, 2006; Le Roux and Rouanet, 2010). In contrast to the *space of individuals* (objects), concentration ellipses facilitate

a differentiating inspection of the particular spatial distribution for all categories of a theoretically relevant variable.

Third, there are good reasons that one might be interested only in a subset of the data, for example, one is interested in the valid responses only. A typical technique is to apply listwise deletion, which has the disadvantage that one may lose many cases and that the remaining sample is biased. In the CA/MCA framework one also could add the missing responses as valid categories since the method deals with categorical data. The disadvantage of this possibility is that the missing cases are usually highly intercorrelated, which has the effect that the first dimension might just reflect *substantive versus nonsubstantive responses*, which is of little relevance in substantive research. An alternative is to apply *subset MCA* (Greenacre and Pardo, 2006; Greenacre, 2007; Blasius and Thiessen, 2012), which allows to exclude the 'not interesting' parts of information without influencing the remaining structure of data. Subset MCA also allows to concentrate on parts of the data, for example, on the 'fully agree' responses in a set of ordered categorical variables.

Fourth, traditional statistical tests of significance are seldom discussed and almost never applied in practical applications of CA and MCA (at least outside of France). Thus, CA and MCA are sometimes criticized because of their purportedly explorative nature and their lack of statistical significance procedures. However, there are several possibilities that can make CA, MCA, and related techniques more attractive for researchers from the world of 'traditional statistical' modeling and inference. Among other developments, *confidence ellipses* indicate the certainty of the *true* position of a variable category by specifying an area around the mean point (see Le Roux and Rouanet, 2010). Similar to the logic of parameter variance in classical inference statistic, the size of the ellipse expresses the statistical (un)certainty of the mean point's position. This approach can be used for distinguishing significant from nonsignificant category positions as well as for discussing the statistical relevance of two distances between points (see also Le Roux, 2014).

For more than 20 years, it has been possible to perform CA and MCA with all major statistical software packages such as Stata, SPSS, and SAS. Furthermore, in R there are specialized software solutions for CA, allowing to apply the advanced techniques mentioned above, for example, the packages CA (Nenadić and Greenacre, 2006) and FactoMineR (Lê et al., 2008).

Conclusion

As Beh and Lombardo (2012, p. 137) state "Over the past half a century correspondence analysis has grown from a little known statistical technique designed to graphically depict the association structure of categorical variables that form a contingency table to a very popular tool used in a wide variety of disciplines."

The main advantage of CA is that it can handle categorical data, ordered as well as unordered ones. As discussed, the only restriction is that CA and MCA can handle nonnegative data only. Since there are no negative values in survey research, there is no methodologically induced limitation in using these methods. However, although CA and MCA are explorative methods, one always should have a substantive theory about the structure of the variables to be included in the model. Bourdieu and his scholars, for example, use the idea of *social spaces* and *social fields* to relate social groups, e.g., occupational groups, to lifestyle indicators and sociodemographic characteristics. The axes (dimensions or factors) of their *social spaces* are often called 'composition of economic and cultural capital' and 'capital volume,' or 'cultural capital' and 'economic capital.' This is very similar to what is well known from PCA and factor analysis.

A particular advantage of these methods is the possibility of visualizing the dimensional structure of data and to read them off like a geographic map. In a geographic map it is very easy to catch which cities are close to each other, which are far away from each other, and how they are related by distance to each other. The same holds when reading a CA or MCA map, it is easy to catch the relations between a large set of variable (categories) and between a large number of individuals. Reading off the structure of the data also can help to see methodologically induced similarities, for example, all kinds of response sets (Blasius and Thiessen, 2012). There are many other techniques closely related to CA and MCA, for example, the *biplot methodology* and *multiple factor analysis*; an overview of these techniques can be found in the readers by Greenacre and Blasius (2006) and by Blasius and Greenacre (2014).

See also: Bourdieu, Pierre (1930–2002); Classification, Understandings of in the Social Sciences; Correspondence Analysis; Cultural Sociology, History of; Culture, Sociology of; Exploratory Data Analysis: Univariate Methods; Field Theory; Graphical Markov Models: Overview; Graphical Methods, Analytic; Graphical Methods: Presentation; Habitus: History of a Concept; Life-style, History of the Concept; Multivariate Analysis: Discrete Variables (Correspondence Models); Sociology, History of; Statistics, History of; Visual Methods in the Social Sciences.

Bibliography

Beh, E.J., Lombardo, R., 2012. A genealogy of correspondence analysis. Australian and New Zealand Journal of Statistics 54, 137–168.

Benzecri, J.P., et al., 1973. L'analyse des Données. Tôme 1: La Taxinomie. Tôme 2: L'analyse des Correspondances, second ed. 1976. Dunod, Paris.

Benzécri, J.P., 2006. L'@nalyse des données: Histoire, bilan, projects, …, perspective. In memoriam: Pierre Bourdieu. Revue MODULAD. Numéro 35, 1–5.

Blasius, J., Thiessen, V., 2012. Assessing the Quality of Survey Data. Sage, London.

Blasius, J., Greenacre, M.J., 1994. Computation of correspondence analysis. In: Greenacre, M.J., Blasius, J. (Eds.), Correspondence Analysis in the Social Sciences. Academic Press, London, pp. 53–78.

Blasius, J., Greenacre, M., 2006. Correspondence analysis and related methods in practice. In: Greenacre, M., Blasius, J. (Eds.), Multiple Correspondence Analysis and Related Methods. Chapman and Hall, Boca Raton, pp. 3–40.

Blasius, J., Greenacre, M. (Eds.), 2014. Visualization and Verbalization of Data. Chapman and Hall, Boca Raton, FL.

Blasius, J., Friedrichs, J., 2008. Lifestyles in distressed neighborhoods – a test of Bourdieu's "taste of necessity" hypothesis. Poetics 36, 24–44.

Blasius, J., Schmitz, A., 2014. The empirical construction of Bourdieu's social space. In: Blasius, J., Greenacre, M. (Eds.), Visualization and Verbalization of Data. Chapman and Hall, Boca Raton, FL, pp. 205–222.

Bourdieu, P., Saint-Martin, M., 1976. Anatomie du Goût. Actes de la Recherche en Sciences Sociales 5, 1–110.

Bourdieu, P., Saint-Martin, M., 1978. La Patronat. Actes de la Recherche en Sciences Sociales 20–21, 3–82.

Bourdieu, P., 1979. La distinction. Critique Sociale du Jugement. Minuit, Paris (Distinction, 1984. Harvard University Press, Cambridge, MA).

Bourdieu, P., 1984. Homo Academicus. Minuit, Paris (Homo Academicus, 1988. Stanford University Press, Stanford, CA).

De Leeuw, J., 2014. History of nonlinear principal component analysis. In: Blasius, J., Greenacre, M. (Eds.), Visualization and Verbalization of Data. Chapman and Hall, Boca Raton, FL, pp. 45–60.

Fisher, R.A., 1940. The precision of discriminant functions. Annals of Eugenics 10, 422–429.

Greenacre, M.J., 1984. Theory and Applications of Correspondence Analysis. Academic Press, London.

Greenacre, M.J., Pardo, R., 2006. Subset multiple correspondence analysis. In: Greenacre, M.J., Blasius, J. (Eds.), Multiple Correspondence Analysis and Related Methods. Chapman and Hall, Boca Raton, FL, pp. 197–217.

Greenacre, M.J., 2007. Correspondence Analysis in Practice, second ed. Chapman and Hall/CRC, London.

Greenacre, M., 2010. Biplots in Practice. BBVA Foundation, Madrid.

Greenacre, M.J., Blasius, J. (Eds.), 2006. Multiple Correspondence Analysis and Related Methods. Chapman and Hall, Boca Raton, FL.

Gifi, A., 1981. Nonlinear Multivariate Analysis. DSWO-Press, University of Leiden.

Gifi, A., 1990. Nonlinear Multivariate Analysis. Wiley, Chichester, UK.

Guttman, L., 1941. The quantification of a class of attributes: a theory and method of scale construction. In: Horst, P. (Ed.), The Prediction of Personal Adjustment. Social Science Research, New York, pp. 318–348.

Hirschfeld, H.O., 1935. A connection between correlation and contingency. Proceedings of the Cambridge Philosophical Society 31, 520–524.

Hotelling, G., 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24, pp. 417–441, 498–520.

Lebart, L., Morineau, A.K., Warwick, K.M., 1984. Multivariate Descriptive Statistical Analysis. Wiley, New York.

Lebart, L., 1994. Complementary use of correspondence analysis and cluster analysis. In: Greenacre, M., Blasius, J. (Eds.), Correspondence Analysis in the Social Sciences. Recent Developments and Applications. Academic Press, London, pp. 162–178.

Lê, S., Josse, J., Husson, F., 2008. FactoMineR: an R package for multivariate analysis. Journal of Statistical Software 25 (1), 1–18.

Lebart, L., Saporta, G., 2014. Historical elements of correspondence analysis and multiple correspondence analysis. In: Blasius, J., Greenacre, M. (Eds.), Visualization and Verbalization of Data. Chapman and Hall, Boca Raton, FL, pp. 31–44.

Le Roux, B., Rouanet, H., 2004. Geometric Data Analysis, from Correspondence Analysis to Structured Data Analysis. Kluwer, Dordrecht.

Le Roux, B., Rouanet, H., 2010. Multiple Correspondence Analysis. Sage, Thousand Oaks, CA.

Le Roux, B., 2014. Structured data analyses. In: Blasius, J., Greenacre, M. (Eds.), Visualization and Verbalization of Data. Chapman and Hall, Boca Raton, FL, pp. 185–203.

Nenadić, O., Greenacre, M.J., 2006. Correspondence analysis in R, with two- and three-dimensional graphics: the CA package. Journal of Statistical Software 20 (3), 1–13.

Pearson, K., 1906. On certain points connected with scale order in the case of a correlation of two characters which for some arrangement give a linear regression line. Biometrika 5, 176–178.

Rouanet, H., Ackermann, W., Le Roux, B., 2000. The geometric analysis of questionnaires: the lesson of Bourdieu's La Distinction. Bulletin de Méthodologie Statistique 65, 5–18.

Rouanet, H., 2006. The geometric analysis of structured individuals $\times$ variables tables. In: Greenacre, M.J., Blasius, J. (Eds.), Multiple Correspondence Analysis and Related Methods. Chapman and Hall, Boca Raton, FL, pp. 138–159.

Corruption: Political and Public Aspects

Alina Mungiu-Pippidi, Hertie School of Governance, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by Y. Mény, L. de Sousa, Volume 4, pp. 2824–2830, © 2001, Elsevier Ltd.

Abstract

The interest of social science, media, and policymakers for corruption has exploded after 1990 due to increasing evidence that corruption is accountable for poor growth and chronic underdevelopment. This had impressive policy consequences: by 2012, no less than 161 state parties had adopted the United Nations Convention against Corruption (UNCAC), which was put into force on 14 December 2005. This article discusses the various conceptualizations of corruption as an individual behavior as well as a governance regime (particularism), the evolution of Western thought until the adoption of ethical universalism as universal governance norm (with UNCAC) and the performance of countries in establishing control of corruption in historical and contemporary times.

Etymology and Definition

Every language knows the word corruption, but its definition is ambiguous in all of them, so most have developed a whole set of words to back it up with more concrete terms. Corruption is departure from what should be (the norm) in ethics, morality, tradition, law, and civic virtue and it was used in all such contexts; by mid-fourteenth century, Europeans used it in parallel to indicate decay of material things, especially dead bodies, but also spiritual things, such as soul and morals. Its ambiguity derives from the relativity of norms across cultures, historical times, and position of social groups and individuals in various such contexts. The use of the term 'corruption' in relation to public office dates from the early fifteenth century in Europe. The Romans, however, knew already the essentials of corruption, in Latin *corruptionem* (nom. *corruptio*). One has as proof two excellent maxims that have survived to this day: 'The corruption of the best is the worst' (*corruptio optimi pessima*) and Terence's line "The most corrupt state has the most laws" (*corruptisima re publica plurimae leges*).

Although historians of government have always devoted attention to corruption, its rise to prominence in the attention of social scientists is of recent date. The last 20 years have produced over 16 000 entries on Google Scholar with 'corruption' in the title, while the previous 20 return only a modest 1740. (Search run on 8 April 2012.) This exponential growth is shown in the tracing over the last two decades of the Internet, the last 10 years returning nearly four times more articles than the previous 10. The explanation is obvious when considering that the most quoted paper in the lot (available from http://qje.oxfordjournals.org/content/110/3/681.abstract) is 'Corruption and Growth,' by economist Paulo Mauro from International Monetary Fund. This association between corruption and growth raised the interest of social science, media, and policymakers to the current heights. By 2012, no less than 161 state parties had ratified the United Nations Convention against Corruption (UNCAC), which was put into force on 14 December 2005. In articles 7 (public sector) and 9 (procurement), the treaty spells out the modern principles of efficiency, transparency, merit, equity, and objectivity as the only accepted governance norms. It also goes far beyond the criminalization of bribing or influence trading, stating in article 1.c that it will "promote integrity, accountability and proper management of public affairs and public property." In other words, UNCAC sets universal benchmarks of good governance.

Governance can be defined as *the set of formal and informal rules determining who gets what in a given polity*. (Developed from the definition of politics according to Laswell (1951) and the definition of governance by World Bank (Huther and Shah, 2000).) A *governance regime is any* stable configuration of governance rules, norms, and practices. *'Good governance' is both the process and the outcome of a governance regime based on ethical universalism*, which the author describes as the situation when the state rules impartially and impersonally in favor of public interest (being autonomous from narrow private one, individual, group, or elite capture). Ethical universalism is a norm and a practice of individualistic societies, where equal treatment applies to everyone regardless of the group to which one belongs (Parsons, 1997: pp. 80–82). Its opposite is particularism, the governance norm allowing allocation of public resources in a nonuniversal way, according to particular ties. For most countries and much of history, ethical universalism was not the dominant governance regime: access to resources was rather strictly controlled by violence and corruption (North et al., 2009). UNCAC demands that its signatories deliver in the same time certain outcomes of governance (accountability, integrity, equity, and control of corruption) and arrive at this through a certain process (transparency, consultation, and 'proper' management).

The United Nations Development Program defines good governance as "The exercise of economic, political, and administrative authority to manage a country's affairs at all levels. It comprises mechanisms, processes, and institutions through which citizens and groups articulate their interests, exercise their legal rights, meet their obligations, and mediate their differences" (UNDP, 1997). The World Bank (1994) describes what it should be and look like: "… epitomized by predictable, open and enlightened policy making; a bureaucracy imbued with a professional ethos; an executive arm of government accountable for its actions; and a strong civil society participating in public affairs; and all behaving under the rule of law."

Such definitions obviously describe *national* institutions and practices. This causes much confusion, since the legal

definition, meant to circumscribe 'corruption' as a crime, obviously depicts *individual* behavior. The most frequent definition of corruption in current literature is 'the abuse of public office for private gain' (Tanzi and Davoodi, 1997; Kaufmann, 2007; Rose-Ackerman, 1999), with variants such as 'abuse of power' or 'abuse of entrusted authority.' Corruption is nearly always defined as *deviation from the norm* (Scott, 1972) because it presumes that authority or office is entrusted to someone not to promote private gain (for self or others) but to promote the public interest, in fairness and impartiality. This was justly criticized by Kaufmann and Vicente (2011), and enlarged by Joseph Nye to "behaviour which deviates from the formal duties of a public role (elective or appointive) because of private-regarding (personal, close family, private clique) wealth or status, gains, or (which) violates rules against the exercise of certain types of private-regarding influence" (Nye 1967(1966)). By including 'status' next to 'wealth' among gains, Nye practically defined as corrupt all types of nonuniversal exchanges within the government realm. This is confirmed by UNCAC, which implies that any authority that is not exercised in a universal way is corrupt.

The current definition of corruption rests therefore on the presumption that it is wrong to use public office for private gain. But public office has no longer been considered a source of exploitable income only in the modern state; prior to modernity, it has been the norm for many centuries (Weber 1922(1968)). In *The Gift*, French sociologist Marcel Mauss argued that the act of giving gifts creates a social bond with an obligation to reciprocate on part of the recipient, since "the objects are never completely separated from the men who exchange them" (Mauss 1925(1990): p. 31); in the absence of reciprocation, a loss of 'status' and 'honor' would occur (Mauss 1925(1990)). Roland Mousnier, a French historian, has furthermore argued that until the eighteenth century, honor, status, and social prestige were far more important in certain European societies (such as France and Spain) than wealth (Mousnier, 1969). In their 1984 book, sociologists Eisenstadt and Roniger (1984) also described nonmodern patron–client relations that shape many contemporary societies in different cultures (Catholic, Buddhist, and Islamist) and decisively determine both interpersonal exchanges as well as the type of connection between a people and its government. Neither historical sociology nor anthropology have ever made the claim that the kind of impersonal and objective bureaucratic relations that Max Weber described as underpinning modern states is some sort of final unavoidable stage in a long chain of evolution (Weber 1922(1968): p. 959).

Taking modernity for granted includes not just the presumption of ethical universalism as the main norm informing exchanges, but a second, equally flawed presumption of the existence of a clear distinction between public and private in every society. As Alena Ledeneva (2009) has shown most formulas of corruption can be understood as a 'twist' of something public into something private and assume not only that the public and private spheres operate according to corresponding sets of rules and norms, but that it is also wrong to mix them. Even today, however, whole societies exist that rely on collectivistic, not individualistic, arrangements and do not value and articulate in practice such distinctions. "It is wrong to presume that the default design of government is of an impersonal state on one side and the individual citizen on the other." Many societies are instead organized as groups of 'clients' with varying degrees of influence and whose interests compete or overlap. Multiethnic societies tend to use ethnicity as the chief marker, but also clan, religion, city of birth, political party, club, family, former school, etc. have all been reported to provide personal ties that inform exchanges and transactions between an individual and the state, crossing the private–public boundary permanently. Connections of every kind are used to personalize transactions with the state and a multitude of particular transactions take place daily, some illegal, others not. We find words for it in many languages. We criminalize only particular exchanges, which include material gifts (bribes) and which often occur in the absence of a bond/connection of another nature with the goal of establishing one. But why is bribing worse than being hired simply because one has the same skin color as the supervisor or is the son of a friend?

Grounded in the delusion that modernity is the norm – with everything this presumes: impersonal bureaucrats, objective knowledge, etc. – one discovers the third major untested presumption of current economic theory of corruption, 'that a well-meaning principal always exists whose trust is abused by some agent' (Rothstein, 2011). Generally, the ruler is seen as the principal and the bureaucrat as the agent (Becker and Stigler, 1974; Van Rijckeghem and Weder, 2001). The rulers cannot exactly observe which agents behave honestly, since they do not possess all the relevant information that the agents have; this provides the agent with opportunities for corrupt exchanges. The presumption that a principal is above corrupt exchanges is taken for granted, and most assistance for good governance programs is directed to such principals (government and anticorruption agencies). But those who have the highest discretionary power also have more opportunities to be corrupt, which makes high-level government officials – represented by legislators or elected public officials – as the best placed to institute or manipulate policy and legislation in favor of particular interest groups in exchange for benefits. Thus, the principal risks being more often than not a patron, if not a *capo di tutti capi* (*the top 'Godfather'*) and such strategies risk running like the old Soviet joke of the man who left the factory every day at the end of program with an empty wheelbarrow that security would scrupulously check to prevent him from stealing, only to turn out after some months that he was actually stealing wheelbarrows.

A less popular version of the principal agent model in good governance literature presumes the vigilant and honest citizens (voters) acting as principals to keep officials accountable (Myerson, 1993; Persson and Tabellini, 2000; Besley, 2006). But what if at the grassroots one finds only clients or disempowered citizens? Insofar as a sufficient number of actors play the game instead of challenging it, this structure of incentives leads to a collective action problem of the second order (Ostrom, 1998). The overwhelming evidence stemming from developing countries is that rulers are rather inclined to treat the state as their own patrimony, and that personal autonomy and political participation, two indicators of political modernity informing a society's capacity to

constrain rulers, are generally very low. In collectivistic societies with closed access, actors strive to accede to the privileged status groups rather than challenge the rules of the game and make them more equitable to everyone (Mungiu-Pippidi, 2006).

The fact that most definitions of corruption refer to the individual level leads to another problem: how can the deviation-based definition be applied in a system where 'particularism (treating a person not as an indistinct individual, but according to particular ties or group affiliations)' is actually the norm? Consider a government that is captured by an ethnic group or a clan of some kind that distributes benefits mostly to its own members and forces others to engage in bribery to get some share of those benefits. What is the deviation here and what should be criminalized? Fortunately, governments adopt constitutions and treaties – it is a part of the great and often content-empty modernization effort. Most of these documents make claims to ethical universalism as a governing principle, creating important contradictions between the informal and the formal norm in such societies (van de Walle, 2001).

Some scholars struggled to find differences between various manifestations of particularism, for instance, clientelism and corruption, in the observance of legal norms (Eisenstadt and Roniger, 1984; della Porta and Vanucci, 1999: p. 273; Johnston, 2005). Other scholars struggle with depictions of various corruption categories (grand, petty, administrative, political, etc.). These distinctions are not very useful, as societies can display a great variety of particularistic practices. What is important is if particularism, as a general category indicating the deviation from the norm of ethical universalism, is the exception or actually the norm. *Is particularism the rule of the game or not (are the majority of transactions carried by the state particular or universal)? If it is the rule of the game (because control of corruption does not exist), one will find a cluster of such practices: vertically structured favoritism (clientelism/patronage), horizontally structured particularism (negative social capital networks, old boys), kinship-motivated particularism (nepotism, ethnic favoritism), bribing, extortion, etc.*

'Corruption' in a developing context is therefore often used only as a catch word for all such nonuniversalistic practices, in other words for the nonmodernity of a society. Public *corruption emerges from many development studies as an indicator of the lack of the institutionalization of market economy and democracy in transition societies* (Haerpfer and Wallace, 2000). To paraphrase St Augustine, who famously considered evil to be the absence of good, *public corruption can be seen as the absence of a rational legal order translated into everyday behavior.*

Typology of Governance Regimes

While ethical universalism and particularism are the two opposites of a continuum defining relations between individuals on the one hand and individuals and the state on the other (with trust and social trust, respectively, defined as *expectation shaped by the experience of such relationships*), patrimonialism was originally described by Weber more as a form of political domination (Weber 1922(1968)). Thus, another variable power was introduced. Here, authority rests on the personal and bureaucratic power exercised by a royal household, whose power is formally arbitrary and under the direct control of the ruler. Domination is secured by means of a political apparatus staffed by slaves, mercenaries, conscripts, or some other group (not a traditional land-owning aristocracy) that has no independent power base. By controlling the instruments of power in this way, the patrimonial ruler can extend personal grace and favors at the expense of traditional limitations on the exercise of authority. In his 1973 book, *Traditional Patrimonialism and Modern Neopatrimonialism*, Shmuel N. Eisenstadt used the derived term 'neopatrimonialism' to describe a mixed system in which formal institutions of modernity from the West are imitated, with most informal institutions remaining of a patrimonial nature.

Neopatrimonial regimes have single rulers who treat the state as their 'own' patrimony. However, many new democracies, which no longer have such rulers or families, but instead competing political parties, have similar nonuniversal allocation of public resources systems, including patronage, nepotism, and favors (O'Donnell, 1996: p. 40). In many new democracies, informal particularistic structures exist alongside and tend to undermine formal universalistic institutions (O'Donnell, 1996: p. 41). How sustainable are such regimes? Some countries have operated for decades under such arrangements, from the North African neopatrimonial dictatorships that collapsed – to everyone's surprise in early 2011 (e.g., Tunisia) – to electoral democracies in Latin America, Southeast Asia, and the Balkans. The claim is that these countries are *not developing precisely due to the persistence of this state of affairs, termed as limited access order* (North et al., 2009).

Table 1 proposes a typology of governance regimes attempting to integrate current theories. 'A governance regime is defined by a stable normative configuration around a dominant norm.' The first governance regime is defined by the *open access order* (North et al., 2009), which corresponds to Dahl's *polyarchy* or Karl Popper's *open society*. This type of regime relies on an individualistic society, with political equality, high personal autonomy, and high civic participation; the state is autonomous from private interest and allocation and policy formulation are achieved on a basis of ethical universalism and transparency. There is very little contradiction between formal and informal institutions, and corruption, when it occurs, is indeed a deviation from the norm of ethical universalism and public impartiality. There is currently an unprecedented large number of countries that aspire to be perceived as having such regimes – Transparency International's ratings has spurred the competition among states to prove themselves closer to the standards published in their annual Corruption Perception Index.

The majority of countries – nearly all with modern constitutions and various kinds of institutional transplants imported from Western democracies – fall under the category of limited access order. Here, the author introduces an important distinction within limited access order societies and differentiates between the patrimonial type described by Weber (where power is a monopoly) and those that have managed to introduce some form of pluralism with the institution of regular elections (several groups compete for the spoils, but spoiling the state and state capture remain the rule of the game). The latter category is called 'competitive particularism,'

Table 1 Governance regimes and their main features

Governance regimes	*Limited access order*			*Open access order (Universalism)*
	Patrimonialism	*Competitive particularism*	*Borderline*	
Power distribution	Hierarchical with monopoly of central power	Stratified with power disputed competitively	Competitive with less stratification	Citizenship Equality
State autonomy	State captured by ruler	State captured in turn by winners of elections	Archipelago of autonomy and captured 'islands'	State autonomous from private interest (legal lobby, etc.)
Public resources	Particular and predicable	Particular but unpredictable	Particular and universal	Ethical universalism
Separation private–public	No	No	Poor	Sharp
Relation formal/informal institutions	Informal institutions substitutive of formal ones	Informal institutions substitutive of formal ones	Competitive and substitutive	Complementary
Mentality	Collectivistic	Collectivistic	Mixed	Individualistic
Government accountability	No	Only when no longer in power	Occasional	Permanent
Rule of law	No; sometimes 'thin'	No	Elites only	General; 'thick'

a governance regime that has replaced violent power grabbing with corrupt elections (sometimes free, although not fair), which is a step forward from patrimonial regimes. Allocation is, however, particular and unfair, rent-seeking practically common behavior, rule of law is poor (those in power are above the law), and the state is perceived as an instrument of spoliation of the many and enrichment of the few, which greatly subverts its legitimacy and capacity. A similar concept was used in connection with postwar Italy, described as 'competitive corruption' (Golden and Chang, 2001). *A culture of privilege reigns in such societies based on closed access order, making unequal treatment the accepted norm in society.* Social acceptance of particularism varies across regimes, but elites can buy social peace if resources exist (for instance, by subsidizing basic foods, as in some neopatrimonial Middle East and North African countries). Due to poor civil society and low personal autonomy, collective action to render governments more accountable and transparent is costly and infrequent. Free elections bring about occasional changes of government, but hardly better governance, as political parties are the main profiteers from the system and form cartels in favor of the status quo. In the lower echelons, frequent corruption might be regarded as a perk of the inferior position, tolerated because it represents an unofficial form of compensation granted to subordinate officials for their lack of influence and power (Huntington, 2007: pp. 259f). These categories (already described in Mungiu-Pippidi, 2006) overlap in part with Kaufmann and Vicente's (2011) (1) 'illegal corruption equilibrium,' where the political elite does not face binding incentives; (2) a 'legal corruption equilibrium,' where the political elite is obliged to incur on a cost to 'deceive' the population; and (3) a 'no-corruption equilibrium,' where the population cannot be deceived.

The ideal types of governance regimes in Table 1 present important differences: the 'borderline' category is not a regime type but a transitional regime with fuzzy borders, which corresponds to the 'doorstep' category of North et al. (2009). In other words, this type captures those societies that have fulfilled some basic and necessary conditions of progress toward open access order and the two normative orders (the old and the new) coexist confrontationally without one managing to become dominant. The important differences between governance types cut across state and society: power distribution is uneven, from the classic one group (family/person) as main power holder and owner of the state with chief control over rents allocation (in patrimonialism) to a larger strata (political elites) holding the same privilege (and disputing it across status groups) in competitive particularism and finally to the equal power regime with open access. Pluralism does not bring about better governance, because patronage subverts democracy, informal client–patron relations structuring politics on a vertical and particularistic relation. Voters in such regimes act as clients 'selling' their vote against favors; occasionally, unaffiliated voters manage to elect an anticorruption president, but then the new president would develop a new clientele, as between elections there are no normative constraints. Under patrimonialism, rule of law can at best evolve to 'thin' (applied predictably, even if not 'just' in its essence); under competitive particularism it will always be interpreted in favor of the group in power (hence the danger of unleashing anticorruption campaigns, which can be used only against political opponents). Power holders are accountable in competitive particularism only when they fall from power, because only rulers in open access order regimes are not above the law and can be prosecuted at any time. Today, one finds that most democratized patrimonial regimes remain in the realm of competitive particularism, with a few Southeast Asian and more East European cases evolving to borderline situations.

The three ideal types differ in the essential elements explaining governance: informality, power distribution in society, and autonomy of the state toward private interest, which result in a certain type of resource allocation. The borders between categories are not well defined and historical thresholds are difficult to identify. Historians present the evolution from the first regime type to the last type as

a sequence, rather than a big bang. Furthermore, this involves such a broad institutional transformation it is highly difficult to distinguish it from modernization in general. In his book, *The Origins of English Individualism*, historian Alan Macfarlane (1978) argues that many countries in the developing world have not yet achieved the first transformation from rural collectivist society to basic market relations based on individual property. However, they are expected to succeed in the second transformation from governance based on particularism to one based on ethical universalism.

Development of Control of Corruption

The explanation for ethical universalism becoming the current dogma (although still on paper for many countries in the world) lies in its association with highly successful economies. Its intellectual heritage seems to run from the Stoics (doctrine of natural law) to Cicero (106–43 BC), Saint Isidore of Seville (*c.*560–636), John of Salisbury (*c.*1120–80), Aquinas (1225–74), Brunetto Latini (1220–94), and Marsilio of Padua (*c.*1275–1342). From then on, it spread in several directions with the Renaissance to become inbuilt in modern state building doctrine in the works of Montesquieu and the American Federalists. The legal philosophy of Cicero seems to have been the most influential early source (Carlyle and Carlyle, 1903; Neumann, 1986; Skinner, 1989). In the first book of *De Re Publica*, Cicero argued for equal treatment of citizens under the law and called the Republic 'an association of rights' (Cicero, 1877: p. XXXII). In *De Legibus*, he further argued that both justice and law derive their origin from Nature, while 'wicked and unjust statutes' are 'anything but laws' (Cicero, 1877). His ideas had an impressive following in the Middle Ages, from Isidore (Isidore of Seville: The Etymologies (or Origins) quoted after the critical edition by W. M. Lindsay, published by Oxford University Press, Oxford, 1911, available at: http://penelope.uchicago.edu/Thayer/E/Roman/Texts/Isidore/home.html.) of Seville who passed them to Aquinas (Aquinas quoted after Aquinas, Thomas: Summa Theologica, Part I–II (Pars Prima Secundae), from the Complete American Edition, translated by Fathers of the English Dominican Province, (release date 2006), available at: http://www.gutenberg.org/cache/epub/17897/pg17897.txt) and then to Brunetto Latini who reformulated and transmitted the republican ethic to other thinkers like Remigio and Dante, and he also used it as the basis for his formal treatise on city-state government (Jones, 1997: 461–462). On the basis of Cicero and Isidore, Aquinas argued in the first article (Aquinas: I–II, Q. 96, Art. 1) of *Summa* "Whether human law should be framed for the community rather than for the individual?", "human laws should be proportionate to the common good," and further, "laws are enacted for no private profit, but for the common benefit of the citizens" (Aquinas: P(2a)- Q(90)-A(1, 2)). In *Defensor Pacis*, Marsilio of Padua further stipulated that the wisdom of a minority does not entitle them to enact legislation on behalf of the majority. Recognizable within Aquinas' and Marsilio's ideas are seeds of ideas of government accountability, 'thick' rule of law and public participation, the three key foundational concepts of what today is called 'good governance.' While these ideas (the thought, not the practices) were extremely widespread, particularly in Italy's communes, where they informed the special tracts of city government that erupted in the early thirteenth century and infused them with republican ideology, such arguments were also made on behalf of monarchs. John of Salisbury argued in the *Policraticus*, Book Four, Chapter 1, that "Between a tyrant and a prince there is this single or chief difference, that the latter obeys the law and rules the people by its dictates, accounting himself as but their servant. […]"

But what is the sense of establishing ethical universalism as the main norm of treatment of citizens or subjects if they are undeserving? In Chapter XVII of his *Discussion on Livy*, Machiavelli soberly suggests that a corrupt people coming onto their liberty can maintain its freedom only with the greatest difficulty. The foundation of control of corruption is in the mentality of the people: "where the people are not corrupted, tumults and other troubles do no harm; but where corruption exists, well ordered laws are of no benefit." He also discussed the possible inconsistency between legal mores and real mores, or, between formal and informal institutions, when he stated "For as good customs have need of laws for maintaining themselves, so the laws, to be observed, have need of good customs."

Mainly due to Aquinas, Cicero's influence survived the Middle Ages. From the Renaissance to the Enlightenment and up to the eighteenth century, his work remained mandatory reading for elites and spread far beyond the circle of legal scholars. Authors such as Pocock (1975) argued that America's founding fathers obtained many ideas from Machiavelli, but Cicero's direct influence can also be traced to Montesquieu and the Federalists (Botein, 1978: p. 150; Reinhold, 1984: p. 315).

The foundations of the Western vision of 'good governance' are therefore Latin and pagan, but they were reinforced by mainstream Catholic doctrine in the twelfth century, long before the Renaissance and Reformation, and their later rediscovery by the Enlightenment. Ethical universalism thus became strong in the European legal philosophy early on and was exported to British colonies; later, political and religious emancipation opened the road to broader interpretations of fair and virtuous governance. But nearly two millennia of a Western tradition of upholding ethical universalism as a basic governance principle passed until its becoming universal doctrine.

European control of corruption must be understood as a gradual process of state building: the doctrine and the practice both took a very long time to become established. One can hardly speak of a 'state' under European feudalism, which was a transterritorial and hierarchical connection between individuals. The result of such a network was not a 'public' state the way it is understood it today. Power offices were completely patrimonial. In the same way that vassals equipped their own troops and came to fight for their overlord when summoned, judges and tax farmers (Tax collectors were private individuals who covered their expenses from what they collected. The tax collector who receives a salary from the government is mostly a modern phenomenon.) were entrusted a task to fulfill as a duty using their own material and human resources. They were allowed to reimburse themselves from plunder of the enemy, fees from claimants, and other means. This in itself was not dishonest. The 'corruption' encountered in the Middle

Ages, the Roman Curia, or the Stuarts Court is improperly labeled, as no public interest, objectively defined, existed, and few 'constitutions' rested upon ethical universalism. It was mostly embezzlement and fraud from the masters. This was different from Italian Republics, where public property and interest were clearly defined. Only when central power lost control over tax farming and office sale regulations, bad governance and corruption can be referenced. However, initially, the sale of offices was a progressive act, praised by Montesquieu, Burke, and Bentham, a way to open offices, previously held by privilege, to new classes. It was in fact a way to democratize access to power. As long as demand surpassed the offer, conditions could be put on bidders to fulfill certain criteria of competence, and the funds they collected afterward could be regulated. When demand fell, however, criteria were relaxed and incompetents started to fill in these positions and govern poorly and arbitrarily, resorting to extortion to cover their investment (Swart, 1949: pp. 92–94). West European sovereigns who used sale of offices to finance their wars – nearly all of them – took time to realize that this was also a way to lose wars. In the eighteenth century, incompetent French and British officers who had bought their way into the army lost many battles in the colonies before the practice of selling commissions was amended by some form of examination of competence. By and large, the sale of offices and the financing of government by fees rather than taxes were stages in the development of the state and it is questionable that they could have been altogether skipped. The process of private–public separation was lengthy because the 'public' needed to be created from scratch. It was first created in the law, even in feudal monarchies, and far later in other realms such as finance and army. The medieval kings of France financed war from their own coffers or loans from Italian financiers. The first act to render the budget transparent attempted by Louis XVI's Minister Necker was rewarded by public outrage, because the Court's expenses, part of public expenditure, seemed to the public both extravagant and exaggerated. The French Revolution introduced the separation between public and private. It took many private bankruptcies to promote public finance sheltered from a monarch's personal extravagance, just as it took heavy military defeat or threat of defeat to promote merit-based competitions in the army. Military threat or serious defeats prompted military reform and merit-based systems for officers in the monarchies of Sweden, Denmark, Britain, Hapsburg Empire, and Prussia. Of course, there was always some ideology behind the promoters of reform: Empress Marie-Therese and her son Joseph II had enlightened ideas, Burke advocated merit-based systems for the British abroad, the Federalists had read Montesquieu and Cicero, and so on. But it seems fair to say that triumph of reformist ideas and the building of ethical universalism as a norm was brought about by necessity rather than virtue. Plainer even, the party of virtue would have remained marginal or in opposition had striking necessity not shown that the old system was leading to disasters of existential proportions.

When European countries like Denmark and Britain started to decisively implement political modernization in the second half of the nineteenth century, some fundamental elements of control of corruption had already been put in place, despite still widespread particularism. Enfranchising new groups and removing the privileges of the old (titles, sinecures, and immunities) was a gradual process that strengthened impartiality and objectivity of government but did not put it into question. This was the gradualist path, resulting from power equilibrium between the new and the old, which did not grant any group sufficient power to become a predatory elite. It evolved through a succession of equilibria that took at least a century. In Britain, the reform of sinecures and 'old corruption' started in 1780, and the building of a more impartial civil service was as nearly over by 1840, followed periodically by new waves of improvements (Cohen, 1965). Electoral regulation reform also evolved starting with the beginning of the nineteenth century and culminating between 1868 and 1883. The main goal of these reforms was to prevent the vote from becoming a commodity and being traded, as it had been in early British democracy. The secrecy of an extended ballot and very transparent and carefully audited election expenses (with ceilings for spending) worked in the end. While in Denmark the monarch was the main principal of reforms, in Britain it was the Parliament that debated, investigated, and gradually adopted reforms to foster the integrity of the voting process. In France, the path was revolutionary. The French Revolution was to a large extent an antiparticularism revolution, directed more against privilege than against property. The 1791 French Constitution stated clearly that sovereignty 'belongs to the nation. No segment of the people and no individual can appropriate it.' However, the succession of regimes induced by the Revolution brought anything but impartiality. In the nineteenth century, administrative 'cleansing' (épuration), initiated by the Revolution became the rule of the game in the relation between political power and administration. Each change of regime was followed by a total sweep of the previous administration in 1815, 1830, 1838, 1852, the great political turmoil of 1977–79, and 1883 (Rosanvallon, 1990: pp. 77–79). Appointment of partisans to administration assured that administrators would be loyal to the policies of leaders, elected or not. Civil servants did not swear an oath to public interest, but to the party in power, a habit that persisted long into the twentieth century. In parallel, the nineteenth century saw the development of a strong trend in the public opinion against favoritism and particularism in general. However, the revolutionary path based on power and administrative overhauls proved less conducive to the development of ethical universalism: patronage and political clientelism became resilient features.

In the era of small government, in the United States, party machines provided services for the poor, the unemployed, and new immigrants and patronage allowed the construction of mass-based parties (Arnold, 2003). The passage of the US Pendleton Act of 1883, which introduced the merit-based system, was triggered by the assassination of President Garfield by one of his electoral campaign workers, disappointed not to be appointed ambassador as reward. The Act, influenced by the earlier Northcote–Trevelyan reforms in Britain (themselves originated in scandals related to the lack of preparedness of the army), specified for the first time that "no person in the public service is … under any obligation to contribute to any political fund," and "no person in said service has any right to use his official authority or influence to coerce the political action of any person…." Nevertheless, it took

decades for the Americans to arrive at a merit-based non-politicized civil service, which they managed before the French.

Thus, there is a variety of paths to control of corruption. In Europe, the Scandinavian countries, Austria, and Germany managed to develop an impartial state prior to democratization. Britain combined some limited democracy with control of corruption, and advanced steadily through gradual reforms. The French and the American ways to ethical universalism, although fairly specific, are similarly tied into the history of their respective republics, with intense politicization, spoiling, and favoritism. Every country, sometimes every region, seems to have evolved on a path of its own to control of corruption and geography is not enough to provide clear and distinct paths. The work remains to some degree unfinished everywhere.

Achieving Control of Corruption in Contemporary Times

The problem with assessing corruption at country level has always been the difficulty of measuring it. Seeing its nature as described in this article, it seems quite a challenge to measure corruption across time and countries. Kaufmann et al. (2006) have nevertheless forcefully argued that corruption can, and is being, measured in three broad ways. (1) *By gathering the informed views of relevant stakeholders*. These include surveys of firms, public officials, and individuals. (2) *By tracking countries' institutional features*, which provide indirect information, e.g., procurement practices, budget transparency, etc. These do not measure actual corruption, but can provide useful indications of the possibility of corruption (…). (3) *By careful audits of specific projects*. These can be purely financial audits or more detailed comparisons of spending with the physical output of projects.

To measure corruption, one therefore disposes of expert scores (criticized for their lack of transparency); of public opinion surveys, which ask about both perception and experience; of business or experts' surveys, which aggregate more individual experiences and beliefs (for instance World Economic Forum); and finally of statistical aggregates of all the above, either in the form of simpler averages, such as the Corruption Perception Index by Transparency International, or of more refined unobserved components extracted by the World Bank Institute (Control of Corruption, Rule of Law). While these indicators have been severely criticized for their lack of validity and reliability because an industry of criticizing indicators developed against the industry of producing them (Andrews, 2008; Thomas, 2009), the fact remains, as argued by Kaufmann et al. (1999) and Knack (2006), that most sources are highly correlated: citizens, businesses, and experts tend to agree on which countries are more or less corrupt. The corruption index developed by Daniel Kaufmann for the World Bank Institute has been published every two years since 1996, covers almost 200 countries, and is computed on the basis of some 100 individual variables on perception of corruption, drawn from about 40 data sources produced by more than 30 different organizations (Kaufmann et al., 2005). To combine the various corruption indicators into a single index, an unobserved component is extracted. The estimate of corruption is given therefore by the weighted average of (rescaled) scores of each of the component indicators. The variance of this conditional distribution provides an estimate of the precision of the corruption indicator for each country. The point estimate of corruption is the mean of the conditional distribution given the observed data and ranges between −2.5 (most corrupt) and +2.5 (least corrupt).

Figure 1 shows regional averages of Worldwide Governance Indicators Control of Corruption. (For the full composition of sources used for this aggregate indicator see http://info.worldbank.org/governance/wgi/pdf/cc.pdf.) The Organisation for Economic Co-operation and Development (OECD) countries' average is separated by a gap of over 30% of the scale from the next region, Eastern Europe (including the Baltics and the Balkans). The latter has evolved more than any other in the world since 1989 (roughly up 15% of the scale), although this applies only to the successful Western part, since what is left of the former Soviet Union actually rates at the very bottom, below sub-Saharan Africa. Not only do the rest of the continents lag far behind the OECD area but – with the exception of Eastern Europe – the trend is toward stagnation, not improvement. The exception is the Caribbean, not represented in the graph on account of their small population.

Recent studies have therefore progressed from asking what causes corruption as deviation to how can control of corruption as norm be explained and in particular how can the transition from one governance regime to another be made (Collier and Vicente, 2011; Mungiu-Pippidi, 2006; Mungiu-

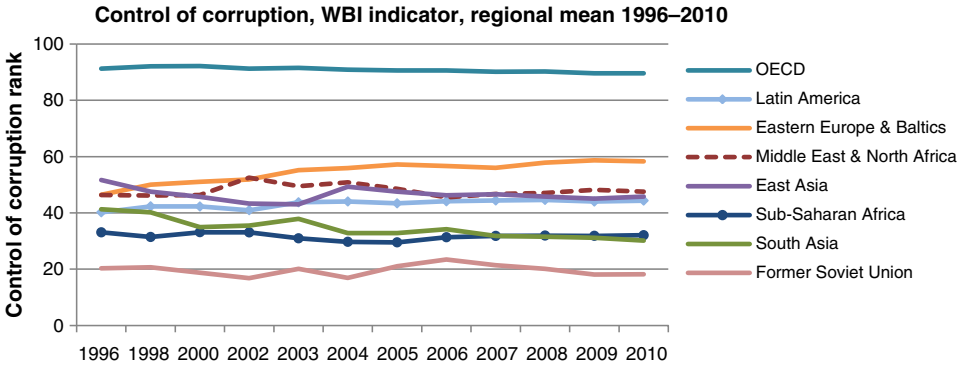

Figure 1 Evolution of control of corruption by global regions (on average).

Pippidi et al., 2011; North et al., 2009; Vicente and Wantchekon, 2009; Birch, 2012). In the absence of serious counterweights, the spoiling of the state by those in power should be presumed to exist by default, since most human societies have limited resources to share, and people tend share them with their closest relatives and not with everyone else. A change in such a governance regime would imply a change in the power equilibrium suggested by Robert Klitgaard (1988: p. 75) in this formula:

$$\text{Corruption} = \text{Monopoly} + \text{Discretion} - \text{Accountability}$$
$$(C = M + D - A)$$

The formula has been developed by Mungiu-Pippidi et al. (2011) to the more explicit:

$$\text{Control of corruption} = \text{Resources (Power discretion} + \text{Material resources)} - \text{Constraints (Legal} + \text{Normative)}$$

This equilibrium formula can be empirically tested and offers a more complex picture, not only of the individual causes of corruption (or even categories of factors), but also of their interaction, which allows for a better understanding of why certain policy combinations work and others do not. All elements of the formula can be affected by human agency. Resources, for example, are not an absolute given; they can be manipulated by policy. Power resources can be increased by discretionary regulation and red tape and decreased by transparency; many anticorruption policies focus on that area. Material resources generally include four basic categories:

- Public jobs, as the public sector can be politicized and each winning party fills not only political offices but many civil service positions with its own people.
- Public spending, for example, the commissioning of public works, but also preferential bailouts, subsidies, loans from state banks, and any form of monetary rents.
- Preferential concessions and privatizations from the state property.
- Market advantages in the form of preferential regulation.

This equilibrium cannot be changed on behalf of legal constraints alone, as evidence shows (Huther and Shah, 2000; Mungiu-Pippidi et al., 2011). But effective strategies do exist, as long as credible principals can be found to lead the change process.

See also: Bureaucracy and Bureaucratization; Collective Behavior, Social Psychology of; Crime: Organized; Equality of Opportunity; Governance; Institutions; Modernity; Modernization, Sociological Theories of; Modernization; Power in Society; Quality of Government; Rule of Law (and Rechtsstaat); Trust, Sociology of.

Bibliography

Andrews, M., 2008. The good governance agenda: beyond indicators without theory. Oxford Development Studies 36 (4), 379–407.

Arnold, P.E., 2003. Democracy and corruption in the 19th century United States: parties, "spoils" and political participation. In: Tiihonen, S. (Ed.), The History of Corruption in Central Government. IOS Press, Brussels.

Becker, G., Stigler, G., 1974. Law enforcement, malfeasance, and the compensation of enforcers. Journal of Legal Studies 3, 1–19.

Besley, T., 2006. Principled Agents: Motivation and Incentives in Politics. Oxford University Press, Oxford.

Birch, S., 2012. Electoral Malpractice. Oxford University Press, Oxford.

Botein, S., 1978. Cicero as role model for early American lawyers: a case study in classical 'influence'. The Classical Journal 73 (4), 315.

Carlyle, R.W., Carlyle, A.J., 1903. A History of Medieval Political Theory in the West, vol. 1. Blackwood, Edinburgh.

Cicero, M.T., 1877. Cicero's Tusculan Disputations; also, Treatises on the Nature of the Gods, and on the Commonwealth. Project Gutenberg (2005). Available at: http://www.gutenberg.org/dirs/1/4/9/8/14988/14988-h/14988-h.htm#FN-309.

Cohen, E.W., 1965. The Growth of the British Civil Service 1780–1939. Frank Cass, London.

Collier, P., Vicente, P., 2011. Violence, bribery, and fraud: the political economy of elections in Sub-Saharan Africa. Public Choice Available at: http://www.springerlink.com/content/55537kw65446t581/export-citation/.

della Porta, D., Vanucci, A., 1999. Corrupt Exchanges: Actors, Resources, and Mechanisms of Political Corruption. Walter de Gruyter, Inc., New York.

Eisenstadt, S.N., 1973. Traditional Patrimonialism and Modern Neopatrimonialism. Sage Publications, Beverly Hills.

Eisenstadt, S.N., Roniger, L., 1984. Patrons, Clients and Friends: Interpersonal Relations and the Structure of Trust in Society. Cambridge University Press, Cambridge.

Golden, M.A., Chang, C.C., July 2001. Competitive corruption: factional conflict and political malfeasance in postwar Italian Christian democracy. World Politics 53, 588–622.

Haerpfer, W., Wallace, C., 2000. Democratization, economic development and corruption in East-Central Europe: a 11-nation-study. In: Sociological Series, vol. 44. HIS, Vienna Available at. http://www.ihs.ac.at/publications/soc/rs44.pdf.

Huntington, S.P., 2007. The Clash of Civilizations and the Remaking of World Order. Simon & Schuster.

Huther, Jeff, Shah, Anwar, 2000. Anticorruption Policies and Programs: A Framework for Evaluation. Policy Research Working Paper 2501. World Bank, Washington, DC.

Johnston, M., 2005. Syndromes of Corruption: Wealth, Power, and Democracy. Cambridge University Press, Cambridge.

Jones, P.J., 1997. The Italian City-State: From Commune to Signoria. Clarendon Press.

Kaufmann, D., Kraay, A., Zoido-Lobatón, P., 1999. Aggregating Governance Indicators. World Bank. Policy Research Working Paper No. 2195.

Kaufmann, D., Kraay, A., Mastruzzi, M., 2005. Governance Matters IV: Governance Indicators for 1996–2004. World Bank Policy Research Working Paper No. 3630.

Kaufmann, D., Kraay, A., Mastruzzi, M., 2006. Measuring corruption: myths and realities. Development Outreach September 2006, 124–137.

Kaufmann, D., Kraay, A., Mastruzzi, M., 2007. Aggregate governance indicators 1996–2006. http://info.worldbank.org/governance/wgi200.

Kaufmann, D., Vicente, P.C., 2011. Legal corruption. Economics & Politics 23, 195–219.

Klitgaard, Robert, 1988. Controlling Corruption. University of California Press, Berkley, CA.

Knack, S., 2006. Measuring Corruption in Eastern Europe and Central Asia: A Critique of the Cross-Country Indicators. World Bank. Policy Research Working Paper No. 3968.

Ledeneva, A., 2009. Corruption in postcommunist societies in Europe: a re-examination. Perspectives on European Politics and Society 10 (1), 69–86.

Macfarlane, A., 1978. The Origins of English Individualism: The Family, Property and Social Transition. Blackwell, Oxford.

Mauss, M., 1925(1990). The Gift: The Form and Reason for Exchange in Archaic Societies (Halls, W.D., Trans.). Routlege, London.

Mousnier, R., 1969. Les Hiérarchies sociales de 1450 à nos jours. Presses Universitaires de France, Paris.

Mungiu-Pippidi, A., 2006. Corruption: diagnosis and treatment. Journal of Democracy 17 (3), 86–99.

Mungiu-Pippidi, A., et al., 2011. Contextual Choices in Fighting Corruption: Lessons Learned. NORAD, Report 4/2011. Available at: http://www.norad.no/en/tools-and-publications/publications/publication?key=383808.

Myerson, R.B., 1993. Effectiveness of electoral systems for reducing government corruption: a game-theoretic analysis. Games and Economic Behavior 5 (1), 118–132.

Neumann, F., 1986. The Rule of Law: Political Theory and the Legal System in Modern Society. Berg, Heidelberg.

North, D., Wallis, J.J., Weingast, B.R., 2009. Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History. Cambridge University Press, New York.

Nye, J.S., 1967(1966). Corruption and political development: a cost-benefit analysis. The American Political Science Review 61 (2), 417–427.
O'Donnell, G., 1996. Illusions about consolidation. Journal of Democracy 7 (2), 34–51.
Ostrom, E., 1998. A behavioral approach to the rational-choice theory of collective action. American Political Science Review 92 (1), 1–22.
Parsons, T., 1997. Introduction to Max Weber. The Theory of Social and Economic Organization. The Free Press, New York.
Persson, T., Tabellini, G., 2000. Political Economics: Explaining Economic Policy. MIT Press, Cambridge and London.
Pocock, J.G.A., 1975. The Machiavellian Moment: Political Thought and the Atlantic Republican Tradition. Princeton.
Reinhold, M., 1984. Classica Americana: The Greek and Roman Heritage in the United States. Wayne State University Press, Detroit, Mich.
Rosanvallon, P., 1990. L'État en France de 1789 à nos jours. Le Seuil, L'Univers historique, coll. Points Histoire, 1993 et 1998.
Rose-Ackerman, S., 1999. Corruption and Government: Causes, Consequences, and Reform. Cambridge University Press, London.
Rothstein, B., 2011. The Quality of Government: Corruption, Social Trust, and Inequality in International Perspective. The University of Chicago Press, Chicago.
Scott, J.C., 1972. Political Corruption. Prentice-Hall, Englewood Cliffs.
Skinner, Q., 1989. Ambrogio Lorenzetti: the artist as political philosopher. In: Belting, H., Blume, D. (Eds.), Malerei und Stadtkultur in der Dantezeit: die Argumentation der Bilder. Hirmer, Munich, pp. 85–103.
Swart, K.W., 1949. Sale of Offices in the Seventeenth Century. Martinus Nijhoff, The Hague.
Tanzi, V., Davoodi, H.R., 1997. Corruption, Public Investment, and Growth. IMF Working Paper 97/139.
Thomas, M.A., 2009. What do the worldwide governance indicators measure? European Journal of Development Research 22 (1), 31–54.
UNDP, 1997. Reconceptualizing Governance. Discussion Paper 2. New York.
Van Rijckeghem, C., Weder, B., 2001. Bureaucratic corruption and the rate of temptation: do wages in the civil service affect corruption, and by how much? Journal of Development Economics, Elsevier 65 (2), 307–331.
van de Walle, N., 2001. African Economies and the Politics of Permanent Crisis, 1979–1999. Cambridge University Press, Cambridge.
Vicente, P.C., Wantchekon, L., 2009. Clientelism and vote buying: lessons from field experiments in African elections. Oxford Review of Economic Policy 25 (2), 292–305.
Weber, M., 1922(1968). In: Roth, G., Wittich, C. (Eds.), Economy and Society. An Outline of Interpretive Sociology, vols. 1–3. Bedminster Press, New York.
World Bank, 1994. Governance: The World Bank's Experience. World Bank, Washington, D.C.

Cortical Areas Engaged in Movement: Neuroimaging Methods

Jody Culham, Mind Institute and Brain, University of Western Ontario, London, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Nomenclature

aIPS anterior intraparietal sulcus (human brain area; putative homolog of AIP)
AIP anterior intraparietal area (macaque brain area involved in grasping)
aSPOC anterior division of SPOC (human brain area; putative homolog of V6A)
cIPS caudal intraparietal sulcus (macaque brain area involved in processing visual orientation)
CZ cingulate zone of secondary motor cortex
DTI diffusion tensor imaging
FEF frontal eye fields
fMRA functional magnetic resonance adaptation
fMRI functional magnetic resonance imaging
LIP lateral intraparietal area (macaque brain area involved in eye movements)
M1 primary motor cortex
mIPS medial intraparietal sulcus (human brain region)
MIP medial intraparietal sulcus (macaque brain area involved in reaching)
MVPA multivoxel pattern analysis
NHP nonhuman primate
PEF parietal eye fields (human brain region; possible homolog of macaque LIP)
PET positron emission tomography
PFC prefrontal cortex (region of frontal lobe anterior to premotor cortex)
PMC premotor cortex
PMd dorsal premotor cortex
PMv ventral premotor cortex
PPC posterior parietal cortex (all of parietal cortex behind S1)
pre-PMd region in front of PMd-proper
pre-SMA region in front of SMA-proper
PRR parietal reach region (macaque brain area involved in reaching, overlaps with MIP and perhaps V6A)
pSPOC posterior division of SPOC (human brain area; putative homolog of V6)
RSC resting state connectivity
S1 primary somatosensory cortex
SEF supplementary eye fields
SMA supplementary motor area
SMC secondary motor cortex
SPOC superior parieto-occipital cortex (human brain area involved in reaching; comprised of aSPOC and pSPOC)
TMS transcranial magnetic stimulation
vIPS ventral intraparietal sulcus (human brain area; putative homolog of VIP)
VIP ventral intraparietal area (macaque brain area involved in head movements)

Abstract

Neuroimaging, particularly functional magnetic resonance imaging, has identified and characterized many areas of the human cerebral cortex involved in planning and executing body movements. Motor commands are sent to the spinal cord from primary motor cortex, which receives input and modulation from secondary motor cortex, parietal cortex, and prefrontal cortex.

Some have argued that all mental functions ultimately have bodily movements as the end goal; after all, evolution selects on actions rather than perceptions and thoughts (except insofar as they affect movements). Thus it could be argued that the entire brain is involved in body movements. Nevertheless, a more useful framework would be to consider how different brain regions are involved at different levels in producing movements. This article will address primate, especially human, brain regions – particularly within the cerebral cortex – that are involved at intermediate levels of planning and controlling movements, with an emphasis on evidence from human neuroimaging.

Hierarchies of Movement Control

One could think of four hierarchical levels in the neural control of movements (Figure 1). At the lowest level are the final output stages of a movement. Muscle movements are driven by input from motor neurons that originate in the spinal cord. These motor neurons receive sensory input about touch and body position from the body that enable fast reflexes (such as when the finger gets burned on a hot stove and reflexively retracts). They also receive input from the brain. Structures in the brainstem and midbrain send output to the spinal cord to control balance and body posture. However, the main source of input

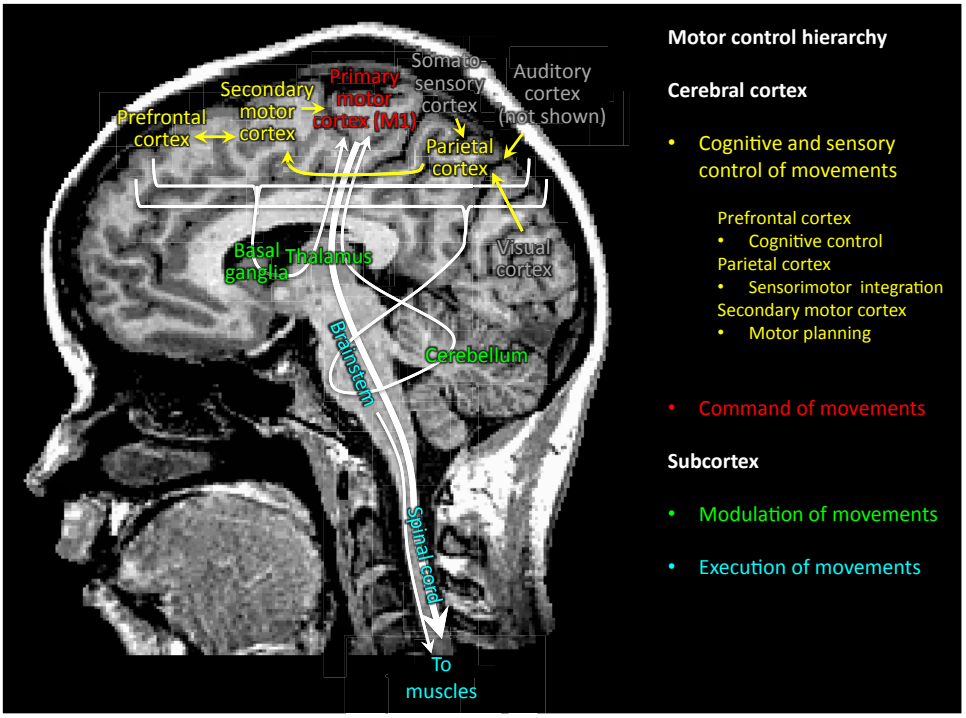

Figure 1 Schematic of key cortical and subcortical regions in the motor control hierarchy and their interconnections. Motor output is sent to the muscles via the spinal cord which receives output via the corticospinal tract (thick white arrow) from primary motor cortex (M1). In addition, the spinal cord receives input from various subcortical tracts, many of which pass through the brainstem. M1 receives input from parietal cortex and secondary motor cortex. Parietal cortex receives input from brain regions that process sensory input from the three spatial senses: visual cortex, somatosensory cortex, and auditory cortex. Secondary motor cortex has crosstalk with regions in prefrontal cortex. In addition, processing in M1 is modulated by two subcortical loops that receive input from many cortical areas and feed back to M1. One loop goes though the basal ganglia and thalamus. Another loop goes through the brainstem (pons), cerebellum, and thalamus. The hierarchy provides one way to think about the motor functions of the regions shown. The schematic is simplified, not all connections (nor bidirectionality of connections) are shown, and not all brain regions lie on the parasagittal plane depicted.

to the spinal cord comes from the primary motor cortex (M1) via the corticospinal tract. M1 receives direct and indirect input from secondary motor cortex (SMC), posterior parietal cortex (PPC), and prefrontal cortex (PFC). SMC includes a host of various brain regions that serve as an interface between sensory inputs (via the PPC) and cognitive inputs (via the PFC). The PPC (behind primary somatosensory cortex, S1) receives multisensory input from visual cortex, somatosensory regions (which code touch, pain and body position sense) and auditory cortex. PPC combines sensory information to plan, execute and control actions. PFC is involved in selecting and planning the appropriate response (and inhibiting inappropriate responses) based on expectations and the situational context. M1 is modulated by two major feedback loops. One modulatory loop involves the basal ganglia and plays a role in selecting and initiating movements based on cortical input. A second modulatory loop involves the cerebellum and plays a role in motor predictions and feedback, motor coordination, and motor learning based on cortical input.

In sum, the cerebral cortex – including primary motor cortex, premotor cortex, parietal cortex and prefrontal cortex – plays a key role in to top-down control of voluntary movements. This control is modulated by subcortical feedback loops and executed though spinal cord output to the muscles.

The Evolution of Cortical Areas Involved in Movement Control

Although most animals other than primates certainly have the ability to move, we see that the complexity and sophistication of movement control is particularly striking in mammals, especially primates. Species without backbones (invertebrates such as flatworms) have simple symmetric nervous systems that convey sensory information to ganglia (clusters of neurons) that control movements for simple locomotion. In relatively simple species with backbones (vertebrates such as frogs), a brain and spinal cord are present. Moreover the nervous system is crossed: the left side of the brain controls the right side of the body and vice versa. These species have both a cerebrum and a cerebellum that enable complex motions. The cerebrum is divided into two hemispheres that are each covered with a 'bark' or cortex comprised of the dendrites and cell bodies of neurons, where information is exchanged. The axons of the neurons form the white matter beneath the cortex that provides the 'wiring' between areas of the cerebral cortex and to subcortical brain structures. The cerebral cortex of vertebrates has shown dramatic expansion in more complex species. In simpler mammals, such as rodents, the cerebral cortex is largely composed of primary motor and sensory

areas. However, higher mammals have additional zones of cortical 'association cortex.' Areas within association cortex are not unequivocally dedicated to motor control or to processing sensory input from a single sense. Rather they are involved in integrating multiple senses, enabling the senses to guide motor actions, and performing advanced cognitive functions (such as decision making). For many decades, association cortex was poorly understood and not well differentiated, however, over the past 30 years, newer techniques have led to the discovery of a host of specialized areas within association cortex. Functional neuroimaging has played a key role in deciphering the functional regions within human association cortex.

Techniques to Study Cortical Areas Involved in Movement Control

Historically, the richest source of understanding the neural basis of movement control came from human neuropsychology, which revealed many clinical disorders of motor control arising from brain damage and disorders. For example, patients with damage to parietal and/or frontal cortex may show apraxia, a disorder in performing purposeful movements.

One additional source for the early understanding of cortical areas engaged in movement came from studies using animal models. Researchers in the nineteenth century found that stimulating different parts of the primary motor cortex in the dog brain led to muscle movements on the opposite side of the body. Strikingly, stimulating different locations in M1 led to movements of different body parts. This work, along with Wilder Penfield's seminal brain stimulation studies in humans in the 1930s (Penfield and Boldrey, 1937), showed that M1 contains a 'map' of the body in which adjacent body parts are represented in adjacent parts of M1. Notably, this map, like that in S1 does not represent all body parts equally. Rather the body parts that are most important for actions in everyday life (notably the hands and lips, for manipulation and speech) take up most of the motor map in M1, while those that are less important (namely the torso and legs) take up relatively little.

Beginning in the 1970s, neurophysiology of awake, behaving nonhuman primates (NHPs) led to the discovery of many specific cortical areas and their neural coding. For example, when electrodes were used to record single neurons in M1, the responses showed tuning for direction. For example, if the animal were pointing to locations on a large clock, one neuron might respond vigorously to the 11 o'clock position (and partially to the 10 and 12 o'clock positions) while another might respond vigorously to the 3 o'clock position (and partially to the 2 and 4 o'clock positions).

Recordings from single neurons in NHPs also revealed a mosaic of brain regions in parietal and frontal cortex with specialized functions tuned for example to the body part to be moved. For example, one area of parietal cortex (the lateral intraparietal area, LIP) was found to use visual information to guide eye movements while another parietal area (the anterior intraparietal area, AIP) was found to use visual information to guide hand movements like grasping and manipulation.

The advent of neuroimaging, particularly functional magnetic resonance imaging (fMRI) in 1992, enabled the emerging field of cognitive neuroscience to develop a much finer understanding of the human brain regions involved in movements. When a particular region of the brain is active, blood flow increases, causing increases in blood oxygenation that can be measured with fMRI. fMRI can sample small regions (called voxels, typically about 3 mm in each direction) rapidly every couple of seconds while the participant receives different stimuli or performs different tasks. Statistical comparisons can then reveal the brain regions that are involved during one condition relative to another. Another neuroimaging technique, positron emission tomography (PET), can also measure brain activation but is less commonly used than fMRI in present day because it involves low doses of radiation and provides less spatial detail and slower timing.

Much of the early neuroimaging work led to the discovery of human brain regions that likely correspond to established areas within association cortex of NHPs. However, increasingly, human neuroimaging is addressing new questions such as which brain areas are involved in advanced human movements like tool use. In addition newer fMRI methods now enable researchers to infer the type of information that is coded in specific brain regions. In one approach (multivoxel pattern analysis, MVPA), researchers can use the spatial pattern of brain activation within brain regions to predict upcoming actions. In another approach (fMRI adaptation or fMRA), the fact that activation decreases when the same stimulus is repeated can be used to investigate what variations are considered 'the same' by an area.

Neuroimaging studies of bodily movements have been less common and slower to develop than studies of other sensory and cognitive functions because of the technical challenges of having people move inside MRI scanners. In fact, many studies, particularly early neuroimaging studies, investigated imagined, pantomimed or observed actions rather than actual action performance. Nevertheless, various approaches have been developed to enable the study of some real actions with fMRI, at least for limited movements with some body parts, especially the eyes and the hands. The movements of other body parts such as the head or legs are still largely problematic and scarcely studied with these techniques. Although most of the neuroimaging research on motor control has been done on humans, increasingly researchers are using fMRI with NHPs, an approach that holds considerable promise for understanding similarities and differences between humans and other primates.

Neurostimulation (particularly transcranial magnetic stimulation, TMS) is increasingly being used to temporarily perturb the function of specific human brain areas to see the effect on movements. While neuroimaging shows that a brain area is activated during a certain condition, it is a correlational technique that cannot demonstrate whether or not an area is essential for a particular cognitive function. Neuropsychology and neurostimulation can demonstrate that damage to a particular region affects the function, enabling stronger arguments about causality. In addition, neuroimaging may be combined with neurostimulation (with neurostimulation during or before fMRI scans) or with neuropsychology (to examine which brain areas are damaged and spared in

patients); however, these approaches have been relatively rare in the study of motor control.

More recently, newer neuroimaging techniques have enabled researchers to study not just the cortical areas engaged in movement but their interconnections. Diffusion tensor imaging (DTI) can use measurements of the diffusion of water molecules along various directions to reconstruct likely white matter tracts that connect areas. Resting state connectivity studies (RSC) can be performed while the subject is lying in the brain scanner without performing a task. This technique identifies brain areas that show synchronized activity, which can be used to infer functional connections (though not necessarily direct anatomical connections).

Specific Human Cortical Areas Engaged in Movement

The following review will focus on cortical regions involved in motor control (Figure 2). We will begin our discussion with primary motor cortex, the main source of commands to lower levels of the motor hierarchy. We will then examine the role of other regions of the cerebral cortex: the role of secondary motor cortex in motor planning, the role of posterior parietal cortex in sensorimotor control of various body parts, and the role of prefrontal cortex in the cognitive control of actions. The emphasis will be on human cortex based on results from human neuroimaging. Nevertheless, given that much human neuroimaging work has been motivated by physiological results from NHPs – particularly the macaque (rhesus) monkey brain – proposed equivalents (homologs or perhaps analogs) between the two species will be noted.

Primary Motor Cortex (M1)

M1 lies along the precentral gyrus of the frontal lobe. Neuroimaging has corroborated findings from other species and other techniques that human M1 contains a topographic map of the body representation. This map corresponds to the opposite (contralateral) side of the body (left M1 controls the right side of the body and vice versa), is inverted (with the feet at the top of the motor strip and the head at the bottom) and has an overrepresentation of the hands and face. The hand area of motor cortex is found midway up the precentral gyrus, where the gyrus curves backward. Within the hand area, various groups have found that moving different fingers yields activation that overlaps but has foci that follow a map. Above the hand area lie the representations of the arms and torso. M1 wraps over the top of the brain to the medial side, where the foot area representation is found. Below the hand area lies the representation of the face and head, particularly lips and tongue.

Secondary Motor Cortex

Numerous secondary motor regions lie in front of motor cortex, along the lateral convexity and midline of the frontal cortex.

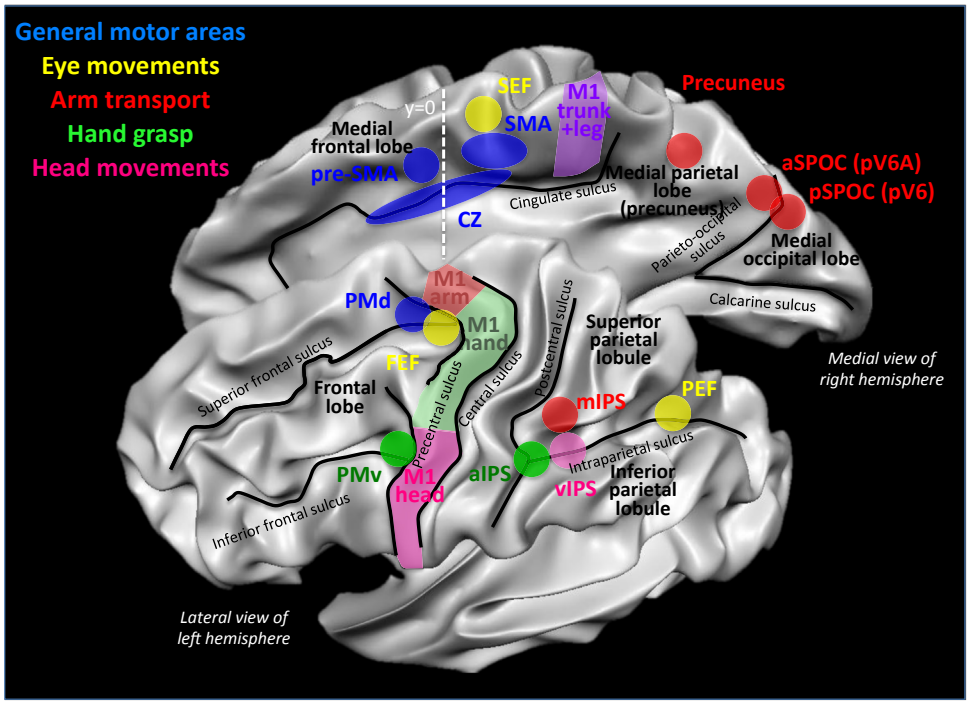

Figure 2 View of the key regions of the human cerebral cortex involved in various types of movements. A lateral view of the left hemisphere is shown in front of a medial view of the right hemisphere. The renderings of the cerebral hemispheres have had the gray matter surface removed to enable regions within the sulci (concavities) to be seen. The fundus (deepest point) of each sulcus is indicated by a black line. The approximate location of cortical brain areas is shown in colors corresponding to their role in general motor control or the control of specific body parts.

Premotor Cortex

Premotor cortex lies just in front of M1 and is typically divided into dorsal (PMd) and ventral (PMv) premotor regions. Both regions have close connections with M1 but also have direct projections to the spinal cord that bypass M1 and are thought to modulate the control of movements. Numerous neuroimaging studies have implicated the dorsal subdivision, PMd, in motor control tasks that require selecting an appropriate response based on past experience. An everyday task that would involve these components is when a driver selects an appropriate response based on the color of a traffic light (green, yellow, or red). The motor decision involves selecting an appropriate action in response to an arbitrary cue based on past experience. Much of the work on PMd has focused on reaching movements and certainly PMd plays a key role here; however, recent work has shown it is also involved in controlling actions with other body parts such as the feet. Two subdivisions of PMd have been proposed: a posterior portion (PMd-proper) involved in hand movements and a more anterior portion (pre-PMd) involved in more cognitively demanding movements (such as conditional actions)

TMS to PMd affects participants' ability to select the appropriate response and to control arm movements during reaching and grip forces during grasping and lifting an object.

PMv lies beneath PMd. Given that PMv seems to be clearly tied to hand actions, it will be discussed in a later section.

Motor Regions along the Medial Wall of the Frontal Lobe

Numerous motor regions along the medial wall of the frontal lobe are also implicated in the modulation of motor cortex. The supplementary motor area (SMA, sometimes called SMA-proper), like premotor areas, has direct connections to both M1 and the spinal cord. The pre-SMA lies immediately anterior to the SMA (in fact, the anterior commissure, $y = 0$ in the Talairach system, is a useful reference point as the SMA lies posterior to it while the pre-SMA lies anterior to it). In contrast to SMA, pre-SMA has no direct connections to M1 and the spinal cord but has connections with prefrontal cortex. While SMA is more of a motor area, pre-SMA plays more of a cognitive role. Thus the medial areas show a similar organization as the lateral (premotor) regions of secondary motor cortex; namely more posterior divisions have more simple motor responses and direct connections with M1 while the more anterior divisions have more cognitive functions. Superior to the SMA in humans lies the supplementary eye fields (SEF; discussed in more detail below). Together these three regions can be thought of as the supplementary motor complex (SMC).

Whereas premotor cortex is involved in actions driven by external stimuli, SMA is thought to be more involved in initiating movement sequences. Patients who have SMA damage can make reflexive movements but cannot initiate voluntary movements. Intriguingly, some patients with lesions to the SMC may show alien-limb (or 'Dr Strangelove') syndrome in which it seems like "the hand has a mind of its own." These patients are compelled to perform actions such as grasping a tool in a functionally appropriate way; they cannot inhibit the action. Thus the SMC seems to be involved not just in initiating voluntary movements but also inhibiting them based on the situation.

Regions in the SMC are activated in anticipation of upcoming movements. PET studies in the 1980s showed that performing a simple hand action activated both M1 and SMA while rehearsing the movement without performing it activated only SMA. Activation in pre-SMA is stronger for actions initiated voluntarily than those initiated by an external stimulus. In addition, SMA is proposed to play a role in movement sequencing and coordinating actions between the two hands. TMS to SMA affects bimanual coordination. TMS to pre-SMA can disrupt participants' ability to do the right task when the instructions change or to veto an action that has been initiated.

Beneath the SMC, a complex of motor-related areas lies along the cingulate sulcus on the medial side of frontal cortex. Though subdivisions have been proposed, they have rarely been differentiated in human neuroimaging studies. Suggested functions for this cingulate zone (CZ) include roles in selecting actions and monitoring conflicts. As in premotor cortex, the cingulate zone is activated by tasks that involve conditional responses (such as the example of reactions to a traffic light).

Areas Involved in Hand Actions

Ventral Premotor Cortex

Ventral premotor cortex (PMv) is involved in controlling hand actions during object manipulation. For example, in a task such as grasping a coffee cup, the actor must use visual information about the cup to guide the hand and open the fingers to an appropriate size and posture. Once the hand contacts the cup, the actor must use the appropriate grip force depending on expectations regarding factors like the fullness of the cup. In neuroimaging studies, PMv is sometimes (but not always) activated in tasks such as grasping, object manipulation, and control of grip forces.

In the macaque brain, PMv contains 'mirror neurons,' which respond not only when the monkey itself performs a hand action (such as picking up a raisin), but also when it observes another individual performing that same hand action. Mirror neurons have been postulated to play a role in understanding others' actions, imitation and even empathy. Although neuroimaging studies cannot detect individual neurons, ample research has reported activation around PMv when participants watch videos of other individuals performing actions. Such research suggests that humans also have a 'mirror system.' In both humans and macaques, responses are stronger for goal-directed actions directed toward target objects (compared to actions without target objects). Despite many similarities between the human and macaque mirror systems, the existence of human mirror neurons per se – populations that respond to both execution and observation of the same action – has been somewhat controversial. Techniques (MVPA and fMRA) that have aimed to find common coding of specific actions during execution and observation have yielded equivocal results.

TMS to PMv affects participants' ability to grasp objects correctly. Consistent with the role of mirror neurons in PMv,

neurostimulation also affects discrimination of the action performed by an actor.

Anterior Intraparietal Sulcus

An area of the human anterior intraparietal sulcus (aIPS; thought to be equivalent to macaque area AIP) is reliably activated during fMRI studies of grasping. Typically aIPS is identified by a contrast between grasping (which requires both transporting the arm and forming a grip) and reaching (which requires only transporting the arm). Notably however, aIPS is activated even if the arm remains still but the hand grips (vs touches) an immediately adjacent object. aIPS responses are typically bilateral but stronger for grasping with the contralateral hand (e.g., left aIPS is activated more when the right hand is used than the left).

During delayed grasping, human aIPS (like macaque AIP) is activated during the visual presentation of a graspable object, the delay period and the period during which the grasp is performed (even in the dark). Within aIPS, there appears to be a gradient in which action execution and somatosensory feedback yield activation in anterior aIPS; whereas, the visual presentation of an object to be grasped yields activation in posterior aIPS. Posterior but not anterior aIPS is activated by viewing of tools (such as screwdrivers) to a greater degree than objects that are merely graspable (such as an apple). Moreover, aIPS is a multimodal area, responding to both vision and touch. In addition to grasp-selective responses, aIPS responds to the amount of grip force used by a participant during squeezing or lifting. It also responds during haptic manipulation of shape, more so for complex than simple objects.

Human fMRI has revealed many additional properties of human aIPS. aIPS is more strongly activated by precision grips (such as grasping a penny with the index finger and thumb) than whole-hand grasps (such as grasping a baseball). During action observation, aIPS appears to code the goal of an action (e.g., grasp the apple, regardless of whether it is on the left or the right) rather than the trajectory (e.g., grasp the object on the left, regardless of whether it is an apple). However, during action execution, aIPS codes both the object shape and the placement of the digits on the object. aIPS is also activated when subjects 'grasp' with a pair of tongs; however, in this case it appears to code the specifics of the grasp rather than the general goal of acquiring the target object. Processing of object shape alone appears to be insufficient to activate aIPS; specifically, judging the size of an object or pretending to grasp it yields little or no aIPS activation. Like PMv, human aIPS (and regions near macaque AIP) are activated during the observation of hand actions.

TMS to aIPS shows that it plays a key role in visually guided grasping performance. Specifically, TMS to aIPS affects participants' ability to use visual information to form an accurate grip and to adjust that grip if the object changes.

In the macaque brain, and presumably also in the human brain, aIPS gets input from a region in the caudal intraparietal sulcus (cIPS) and sends output to PMv. Neurons in both macaque and human cIPS are tuned for the orientation of an object. For example, one cIPS neuron may show the most vigorous response for a flat surface (like a dinner plate) with a particular orientation in 3D space (e.g., with the top receding in depth) while another neuron might show the most vigorous response for an elongated object with a different orientation in 3D space (e.g., a cylinder oriented horizontally). Thus it is thought that cIPS may do preliminary visual processing of object orientation that can be used to select the orientation of the hand to grip the object. Although a number of human neuroimaging studies have found regions in the intraparietal sulcus behind aIPS that show some of the properties of macaque cIPS, the anatomical location across different studies is so variable that there remains no consensus about a human homolog or its location.

Areas Involved in Arm Actions

Superior Parieto-Occipital Cortex, Precuneus, and mIPS

Several human brain areas have been implicated in reaching movements such as pushing an elevator button. One well-studied reach-related area is in superior parieto-occipital cortex (SPOC), although reach-related activation is also sometimes observed in the precuneus and medial intraparietal sulcus.

SPOC responds more to actions that require transporting the arm to a new location compared to actions that don't. However, arm transport is not strictly necessary as it can be activated even when the arm is used to indicate a distant location without reaching (such as pointing to an object beyond reach such as a star). As such, one likely account is that SPOC is particularly involved in coding target locations for the purpose of guiding hand actions. This area is also activated when fMRI participants plan a reaching action (vs an eye movement). It is only activated during reaching when the eyes are not looking directly at the reach target. As such, SPOC is thought to be involved in hand-eye coordination. Indeed, in one intriguing DTI study, participants who were trained to juggle showed training-related enhancements in white matter around SPOC.

SPOC has visual responses, particularly to large moving stimuli and responds more when the hand can be seen during reaching. SPOC also shows preferential responses to near space. That is, SPOC activity increases when participants are looking at near targets or when graspable objects are placed in near regions of space where grasping typically occurs. In addition, SPOC responses are stronger for reaches in the lower visual field, where hand actions typically occur.

In some studies, SPOC has been divided into anterior and posterior subdivisions (aSPOC and pSPOC). Based on a variety of functional similarities, aSPOC and pSPOC are thought to be equivalent to macaque regions V6A and V6, respectively. If that is indeed the case, pSPOC (V6) should serve more as a visual area that processes information in the visual periphery; whereas, aSPOC (V6A) should be more of a visuomotor area that uses visual information for reaching.

Activation for reaching and pointing tasks is also sometimes observed in the precuneus (regardless of whether the hand is seen, unlike SPOC) and medial intraparietal sulcus (mIPS; regardless of whether or not the person is looking directly at the target, unlike SPOC). Although many researchers have proposed homologies between human and macaque reaching areas, these proposals are still quite speculative. An area called

the parietal reach region (PRR) has been reported in the macaque brain and largely overlaps with a medial intraparietal area. However, different neuroimaging groups have used the PRR label to describe SPOC, precuneus and mIPS. Until stricter criteria for homologies are established, it is preferable to use labels that describe the anatomical location instead of proposed labels based on macaque functional areas.

Areas Involved in Eye Movements

Frontal Eye Fields

One of the most well-studied human brain areas involved in eye movements is the frontal eye fields (FEF). In humans, the FEF are much further back (at the junction of the superior frontal and precentral sulci) than in the macaque (in the arcuate sulcus). Although careful fMRI studies show that FEF is located below PMd, the areas are very close to one another and the vast majority of neuroimaging studies base area labels on expectations more than data. When fMRI uses spatial resolutions (~1 mm) finer than is typical (~3 mm), separate subdivisions of human FEF, like macaque FEF, are activated by saccades (abrupt shifts in gaze) and smooth pursuit of moving targets.

Parietal Eye Fields

Parietal activation is frequently observed for eye movements. Saccades are the most well-studied type of eye movement. The most common focus of activation for saccades falls along, and sometimes medial to, the intraparietal sulcus. Activation in this region is sometimes but not always seen to be stronger during the planning of eye movements compared to arm movements. This region contains a 'saccadotopic' map of locations for planned eye movements. For example, part of the parietal eye field (PEF) in the left hemisphere will be activated when a saccade is planned to the 2 o'clock position and this part will be between those activated when saccades are planned to the 1 and 3 o'clock positions. Originally the first parietal map that was discovered was thought to be the human equivalent of the macaque LIP (though multiple topographic maps have now been identified along the intraparietal sulcus). Given that in humans, the activation tends to be more medial than lateral and that homologies are quite speculative, here the label PEF is preferred.

Supplementary Eye Fields

The SEFs lie on the medial wall of the cortical surface, just above the SMA. The function of the SEF is not as well understood as other cortical regions involved in eye movements. However, some evidence suggests the SEF are involved in resolving conflicts between saccade targets.

Areas Involved in Head Movements

Ventral Intraparietal Sulcus

Macaque neurophysiology has revealed an area in the ventral intraparietal sulcus (VIP) that appears to be involved in moving the head. Stimulation to macaque VIP evokes defensive movements such as ducking the head. VIP also responds to the sense of motion (through vision, touch or hearing), particularly near the head.

Investigations of the human ventral intraparietal sulcus (vIPS) have been relatively rare due to the difficulty of allowing head movements with fMRI. Nevertheless human vIPS does display some similarities to macaque VIP, including activation to motion from different senses.

Neuroimaging Studies of Connectivity

Neuroimaging studies suggest that the anatomical and functional connections between brain regions in humans are similar to those of NHPs. Anatomically, DTI tractography can identify the corticospinal tracts descending from M1 and can even resolve fibers from M1 zones that control different body parts. RSC studies have also revealed other functional connections. In particular, even at rest, left and right M1 show strongly correlated activity, suggesting strong interhemispheric connections.

DTI tractography and RSC have also revealed expected connections between cortical association areas in humans. These include connections between aIPS and PMv and between mIPS and PMd. One study used TMS to stimulate human M1 or PMd while recording PET activation throughout the brain. TMS to M1 affected PET activation in secondary motor areas and subcortical areas. In contrast, TMS to PMd affected PET activation in other secondary motor areas (including PMv), parietal cortex (including aIPS and mIPS) and prefrontal cortex.

TMS studies have also revealed connections among the key cortical areas involved in actions. One useful approach is dual-site TMS in which a premotor or parietal area is stimulated several milliseconds before motor cortex to evaluate the effects of interactions between areas upon muscle responses (motor-evoked potentials). Dual-site studies have corroborated the existence of a PMd-M1 network involved selecting an appropriate response. They have also shown that aIPS-PMv-M1 network is involved in preparing the hand to grasp a visible object and that this effect depends on the type of grasp planned (precision vs whole-hand grasp). Furthermore, they have supported the hypothesis that SPOC-M1 interactions are implicated in arm transport while aIPS-M1 interactions are implemented in grip.

Summary and General Organizational Principles

Given the plethora of specific cortical regions implicated in movement control, it may be helpful to consider the general organizational principles that govern their anatomical and functional arrangement.

Although premotor and parietal cortices have been viewed as a mosaic of specialized areas, another view suggests that they have a general organization that accounts for why specific functions are performed at specific locations on the cortical sheet. Brain organization can be simplified by understanding a few core principles. One principle is that "neurons that fire together wire together"; put another way, connections form between regions that share information processing. A second principle is that the wires (axons of neurons) should be as short

as possible. One outcome of these principles is that the cortex often contains maps that enable, in this case, adjacent parts of the body to be represented in adjacent parts of the cortex. Another outcome is that adjacent maps (in adjacent areas) are often mirror images. These organizational principles have been well established in sensory systems; however, more recently, they are being applied to understand motor systems as well.

Of course, the most obvious motor-related map is the body map seen in M1 and S1; however, this body map also accounts for some organization of premotor and parietal cortex, albeit more coarsely. In one fMRI experiment, participants watched movies of people performing actions with the foot, hand or mouth. Activation for hand actions was found around aIPS, while activation for observing foot actions was higher up and activation for mouth actions was lower down. A similar organization of activation was found in premotor cortex.

According to one view, premotor cortex may be part of a two-dimensional sheet (if one imagines unfolding the cortex and laying it flat) that represents not just the body map but also aspects of the most common real-world behaviors. Work by Michael Graziano (2008) and his colleagues has shown that strong electrical stimulation to various zones of macaque premotor cortex (PMC) evokes highly specific movement patterns. Stimulation at the top of PMC evokes climbing behaviors (in the PMC area adjacent to the M1 representation of the limbs). Stimulation lower down evokes grasping behaviors (in the PMC area adjacent to the M1 representation of the hands). Even lower stimulation evokes defensive movements to protect the head and feeding behavior (in the PMC area adjacent to the M1 representation of the head and mouth). Thus it seems there is a topography in PMC that reflects the fundamental motor behaviors associated with various body parts, paralleling the arrangement in M1. However, the mapping of body parts may not be the only feature that affects premotor organization. For example within the zone of PMC that evoked grasping actions, stimulation to adjacent regions led the monkey to reach and grasp to specific regions of space.

Taken together, this data and model suggest that cortical regions surrounding M1 – PMC and PPC – are organized along an up-down (superior–inferior) dimension that reflects body parts and the most common behaviors they subserve in natural settings.

In addition, the cortical areas subserving actions also have a front-back (anterior-posterior) organization with motor cortex at the crux. The parietal lobes are positioned between motor cortex and sensory cortex that processes vision, touch and hearing. Indeed there seems to be a gradient within PPC, with more anterior zones having strong somatosensory and motor responses and more posterior zones having strong visual responses. The PMC is positioned between prefrontal cortex and motor cortex and acts as a 'liaison' between executive decisions and their execution. Some have suggested a broader organization for a wide range of parietal and frontal regions involved not just in motor control but other cognitive functions like attention, calculation and language. Specifically, the proposal is that frontal and parietal regions form a mirror symmetric arrangement reflected across M1/S1. For example, in frontal cortex, areas involved in language tasks are in front of areas recruited by mathematical calculation tasks whereas in parietal cortex, they are behind forming a 'palindromic' arrangement indicative of mirror reflections that can yield efficient wiring in the folded brain.

In sum, motor cortex is thus optimally located to receive bottom-up sensory input (from parietal cortex), top-down executive input (from premotor and prefrontal cortex), and send the output to the two modulatory loops (through the basal ganglia and cerebellum) for feedback before directing the motor commands to the muscles via the spinal cord. There may be at least two fundamental organizational principles for parietofrontal cortex: a top-bottom (dorsal–ventral) arrangement related to body parts and the ecological functions they subserve and a back–middle–front (posterior–central sulcus–anterior) arrangement related to sensory–motor–cognitive processing (and perhaps including other cognitive functions beyond motor control).

See also: Apraxia; Cerebellum: Cognitive Functions; Cerebral Cortex; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Embodiment Theory; Eye Movement, Control of: Oculomotor Control; Motor Cortex; Multisensory Integration, Principles of; Neural Plasticity; Occipital Lobe (Visual Cortex): Functional Aspects; Parietal Lobe; Prefrontal Cortex; Premotor Cortex; Topographic Maps; Visual Perception, Neural Basis of; Visual Streams: Dorsal and Ventral; Williams Syndrome.

Bibliography

Castiello, U., 2005. The neuroscience of grasping. Nature Reviews Neuroscience 6, 726–736.

Chouinard, P.A., Paus, T., 2006. The primary motor and premotor areas of the human cerebral cortex. Neuroscience 12, 143–152.

Chouinard, P.A., Paus, T., 2010. What have we learned from "perturbing" the human cortical motor system with transcranial magnetic stimulation? Frontiers in Human Neuroscience 4, 173.

Clavagnier, S., Prado, J., Kennedy, H., Perenin, M.-T., 2007. How humans reach: distinct cortical systems for central and peripheral vision. Neuroscientist 13, 22–27.

Culham, J.C., Cavina-Pratesi, C., Singhal, A., 2006. The role of parietal cortex in visuomotor control: what have we learned from neuroimaging? Neuropsychologia 44, 2668–2684.

Culham, J.C., Valyear, K.F., 2006. Human parietal cortex in action. Current Opinion in Neurobiology 16, 205–212.

Filimon, F., 2010. Human cortical control of hand movements: parietofrontal networks for reaching, grasping, and pointing. Neuroscience 16, 388–407.

Goodale, M.A., Milner, D., 2013. Sight Unseen, second ed. Oxford University Press, Oxford, UK.

Graziano, M., 2008. The Intelligent Movement Machine: An Ethological Perspective on the Primate Motor System. Oxford University Press, Oxford, UK.

Grefkes, C., Fink, G.R., 2005. The functional organization of the intraparietal sulcus in humans and monkeys. Journal of Anatomy 207, 3–17.

Keysers, C., Fadiga, L., 2008. The mirror neuron system: new frontiers. Social and Neuroscience 3, 193–198.

Konen, C.S., Mruczek, R.E.B., Montora, J.L., Kastner, S., 2013. Functional organization of human posterior parietal cortex: grasping- and reaching-related activations relative to topographically organized cortex. Journal of Neurophysiology, 2897–2908.

Morin, O., Grèzes, J., 2008. What is "mirror" in the premotor cortex? A review. Neurophysiologie Clinique 38, 189–195.

Mountcastle, V.B., 1995. The parietal system and some higher brain functions. Cerebral Cortex 5, 377–390.

Nachev, P., Kennard, C., Husain, M., 2008. Functional role of the supplementary and pre-supplementary motor areas. Nature Reviews Neuroscience 9, 856–869.

Penfield, W., Boldrey, E., 1937. Somatic motor and sensory representation in the cerebral cortex of man studied by electrical stimulation. Brain 60, 389–443.

Picard, N., Strick, P.L., 2001. Imaging the premotor areas. Current Opinion in Neurobiology 11, 663–672.
Sakata, H., Taira, M., Kusunoki, M., Murata, A., Tanaka, Y., 1997. The TINS Lecture. The parietal association cortex in depth perception and visual control of hand action. Trends in Neurosciences 20, 350–357.
Simon, O., Kherif, F., Flandin, G., et al., 2004. Automatized clustering and functional geometry of human parietofrontal networks for language, space, and number. Neuroimage 23, 1192–1202.
Sweeney, J.A., Luna, B., Keedy, S.K., McDowell, J.E., Clementz, B.A., 2007. fMRI studies of eye movement control: investigating the interaction of cognitive and sensorimotor brain systems. Neuroimage 36 (Suppl. 2), T54–T60.
Tunik, E., Rice, N.J., Hamilton, A., Grafton, S.T., 2007. Beyond grasping: representation of action in human anterior intraparietal sulcus. Neuroimage 36 (Suppl. 2), T77–T86.
Turella, L., Pierno, A.C., Tubaldi, F., Castiello, U., 2009. Mirror neurons in humans: consisting or confounding evidence? Brain and Language 108, 10–21.
Vesia, M., Crawford, J.D., 2012. Specialization of reach function in human posterior parietal cortex. Experimental Brain Research 221, 1–18.

Relevant Websites

http://thebrain.mcgill.ca/flash/d/d_06/d_06_cr/d_06_cr_mou/d_06_cr_mou.html – The Brain from Top to Bottom: Body Movement and the Brain.
http://brainconnection.positscience.com/the-anatomy-of-movement/ – Brain Connection: The Anatomy of Movement.
http://www.fmri4newbies.com – fMRI for Newbies.
http://www.ted.com/talks/daniel_wolpert_the_real_reason_for_brains.html – TED Talks: Daniel Wolpert, "The Real Reason for Brains."

Cortical Development: Transplantation and Rewiring Studies

Mriganka Sur, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The cerebral cortex is formed by discrete functional subregions that have distinct patterns of connections, cell types, and molecular organization. These distinct properties allow the subregions to perform distinct functions. During development, the unique projections and neural activity act to shape fundamental properties of areal organization. Perturbation of these projects can profoundly alter both the organization and function of cortical regions. Such changes can be induced by transplanting one developing cortical patch to another region, surgically induced rewiring of the developing cortex, or more naturally by auditory or visual deprivation in development.

The development of the mammalian neocortex involves the formation of many discrete areas that process different kinds of information uniquely. Individual areas are characterized by specific sets of inputs, processing networks, and outputs. The development of these features involves factors that are intrinsic to the tissue, or are regulated by extrinsic or environmental influences. Transplanting one region of cortex to another at particular times in development, or rewiring inputs to a part of the cortex, provides fundamental information on developmental mechanisms of the cortex. Such studies demonstrate that the features that make a cortical area unique are expressed progressively as development proceeds. Importantly, they show that electrical activity can have an instructive role during the development of cortical networks, so that specific spatial and temporal patterns of activity regulate specific patterns of connections between cortical neurons. The mammalian neocortex, a folded sheet that in humans contains over 10 billion neurons, is the seat of the most complex sensory, motor, and cognitive abilities. Cortical development involves the formation of many discrete areas that uniquely process different kinds of information. Individual areas are characterized by specific sets of inputs, processing networks, and outputs. The development of these features involves factors that are intrinsic to the tissue or are regulated by extrinsic or environmental influences. This article examines the relative roles of such influences, employing in particular evidence from experiments that have probed developmental mechanisms by transplanting cortical regions or by rewiring inputs to a part of the cortex.

Development of the Cortex

The principal method for studying mechanisms of development is to manipulate developmental processes. In general, it is not possible to separate intrinsic cleanly from extrinsic aspects of developmental programs of an organism: No gene acts independently of an environment, and the environment always acts on a scaffold. A classic paradigm in developmental biology for clarifying the role of the local environment in determining the features of a tissue is to translocate the tissue to a novel environment. Transplanting one part of the mammalian cortex to another at different stages of development has been a valuable tool for determining whether or not unique features of a cortical area derive from intrinsic properties of the cortex, or are induced by new inputs in the host environment. A powerful recent method that complements transplantation studies is to route sensory inputs that normally drive one part of the cortex to another part. Thus, instead of altering the location of a target and preserving host inputs, the inputs to existing targets are altered. Together, these studies demonstrate that the features that make a cortical area unique are expressed progressively as development proceeds. Importantly, they show that electrical activity can have an instructive role during the development of cortical networks, so that specific spatial and temporal patterns of activity regulate specific patterns of connections between cortical neurons.

Formation of Cortical Layers

The adult mammalian neocortex consists of six layers. These are generated by a heterogeneous population of precursor cells that lie along the walls of the lateral ventricles very early in development. The time of neurogenesis in the ventricular zone regulates the final disposition of cells in individual layers. Cells that are born earlier form the deeper cortical layers, whereas later-born cells form the progressively more superficial layers. Transplantation experiments have shown that there is a progressive restriction of cell fate in the precursor population: Early progenitors are able to give rise to neurons which come to reside in any cortical layer, whereas later progenitors give rise only to neurons of more superficial layers.

One key set of experiments (McConnell and Kaznowski, 1991) demonstrates that transplanting neurons from the ventricular zone of a donor animal at an age when layer 6 is being born, into a host in which layer 2/3 is being born, causes the donor cells to settle into either host-specific layer 2/3 or donor-specific layer 6, depending on the stage of the cell cycle at which donor cells are extracted. If donor cells are extracted prior to or during the phase of DNA synthesis, they are able to switch their fate and migrate to the layer that is being generated in the host (layer 2/3), due presumably to environmental cues in the host. If donor cells are extracted later in the cell cycle, their fate is sealed prior to transplantation and they settle into the layer determined in the donor (layer six). The precise

 http://dx.doi.org/10.1016/B978-0-08-097086-8.55014-4

nature of extrinsic signals that determines or alters the laminar fate of newborn cells during a certain part of the cell cycle is unknown.

Events occurring at the ventricular zone can contribute in other ways to regional differences in cortical architecture. The dynamics of the cell cycle show regional variation. Most notably, the germinal zone of the primary visual cortex (V1) of primates, a region with twice the neuronal density of other areas, has a notably higher rate of cell production which is associated with changes in the parameters of the cell cycle. Thalamic afferents can exert a mitogenic effect on cortical progenitors, resulting in an increase in the number of cell divisions. Thus, thalamic inputs are one important source of extrinsic influences on cortical development, starting with the cell cycle and lamina formation.

Formation of Cortical Areas

There is considerable evidence suggesting that at least the initial broad parcellation of the cortex into discrete areas is regulated by molecular determinants that are intrinsic to the proliferative zone of the developing cortex. For example, evidence from transplantation and coculture studies suggests that the initial expression of region-specific markers or phenotypes is not dependent on specific thalamic innervation. A number of genes are expressed differentially between cortical areas before thalamic afferents are believed to have invaded the cortical plate. Similarly, the expression of several genes can occur normally even in the absence of thalamocortical innervation. Finally, gradients of gene expression in the early cortex can regulate initial cortical parcellation, and such genes can also serve to attract appropriate sets of inputs from the thalamus, or direct appropriate outputs from the cortex (Rubenstein and Rakic, 1999).

While the expression of genes and molecules may delineate broad areas of cortex at very early stages of development, subsequently thalamic afferents can influence the size and even the identity of specific areas (O'Leary et al., 1994). Transplants of extremely immature cortex – at embryonic day (E)12 in rats – from the limbic to the somatosensory cortex may take on inputs and express molecules characteristic of the host region rather than the region of origin. This ability is lost if the transplants are derived at a later developmental stage. Transplanting visual cortex into somatosensory cortex in rats causes the graft to accept inputs from the somatosensory thalamus and form 'barrels,' a characteristic of the rodent somatosensory cortex. Other work shows that the transplant needs to occur early, before E16, for the graft to form substantial connections with the somatosensory thalamus, and later grafts are already committed to form connections with the visual thalamus. With respect to output connections, occipital to frontal transplants at E12 can form projections to the spinal cord, a characteristic of the host cortex rather than the donor; however, by E14 that capacity has been lost and projections to the tectum, a characteristic of the region of origin, are formed instead. Interestingly, a similar age-dependent switch between E12 and E14 influences the corticocortical connections made by transplanted perirhinal cortex into the parietal region. These findings indicate that there is an early time window during which the input and output connections of a cortical area may be sculpted by thalamic and local or regional signals to produce an area-specific phenotype. Thus, cortical areas arise by progressive specification of their region-specific phenotype from a multipotent phenotype (Levitt et al., 1997).

Formation of Cortical Networks: Ocular Dominance and Orientation Columns in Visual Cortex

The cortex differentiates progressively into layers, and as layer 4 appears in the cortical plate, thalamic innervation specifies the principal sensory areas. Concurrently, descending projections from the cortex innervate specific thalamic nuclei, primarily by the targeting of axons of layer 5 cells to principal and/or association nuclei, followed by the development of layer 6 projections to principal relay nuclei. Thus, a thalamocortical loop is set up very early in development, and the initial specification of cortical areas is fundamentally a specification of unique feedforward and feedback connections between thalamus and cortex. The development of intracortical circuitry follows thalamic innervation; considerable recent evidence, primarily from the visual cortex, demonstrates that a combination of intrinsic and extrinsic factors (such as electrical activity in thalamic inputs) is responsible for the formation and maintenance of specific thalamocortical and intracortical connections.

Ocular dominance columns – regions within layer 4 of V1 in higher mammals that receive input exclusively from one eye or the other via the lateral geniculate nucleus (LGN) of the visual thalamus – are paradigmatic of thalamocortical patterns that specify individual cortical areas. In primates they are present by birth, suggesting that visual experience is not required for their formation. Recent work suggests that they may also be set up before eye opening in ferrets, soon after geniculocortical axons innervate layer 4 (Crowley and Katz, 2000), and similarly in cats. Surprisingly, they may not even require the presence of the eyes for their initial establishment. Monocular enucleation does not degrade these early columns immediately, and binocular lid suture does not reduce their development in cats for the first 3 weeks. One possibility is that ocular dominance columns are set up initially by the molecular matching of inputs from different layers of the LGN (each layer representing a given eye) with appropriate target regions in V1.

An alternative view is that ocular dominance columns are set up by activity-based rules of development. Thus, if inputs from one eye are correlated with each other and uncorrelated with inputs from the other eye, application of a Hebbian developmental rule for strengthening connections (and of a complementary rule for weakening connections), together with local excitation and long-range inhibition in the cortex, can lead to cortical stripes that resemble ocular dominance columns. Correlated inputs in the form of spontaneous waves of activity exist in the retinae prior to eye opening. In the LGN, the firing of neurons within a given eye-specific layer is well correlated, and uncorrelated with the firing of neurons in an adjacent layer which represents the other eye. Interestingly, ablation of the cortex abolishes these correlations, whereas ablation of inputs from the eyes has little effect. The pattern of electrical activity

transferred to the cortex, therefore, appears to be regulated strongly by interactions between the thalamus and the cortex, and activity in the thalamocortical loop may be sufficient for eye-specific patterning. Significant support for the hypothesis that ocular dominance columns can arise, at least in principle, by activity-dependent sorting of inputs comes from the finding that eye-specific stripes form in the optic tectum of 'three-eyed' frogs forced to receive input from two eyes (Law and Constantine-Paton, 1981). Whether or not ocular dominance columns in V1 require electrical activity to instruct their formation, there is convincing evidence that electrical activity is required to maintain them (Sherman and Spear, 1982).

Orientation selectivity is a major emergent property in visual cortex that is created by aligned thalamic inputs whose activity is amplified by local intracortical connections. Orientation selectivity is present in V1 of primates at birth, and to a degree in cats and ferrets at or before the time of natural eye opening, although selectivity increases with visual experience (Chapman et al., 1999). Orientation-selective cells in visual cortex are linked by long-range horizontal intracortical connections in the superficial layers to form an orientation map. The adult pattern of clustered horizontal connections is present at birth in primates. In cats and ferrets, crude clusters appear just prior to eye opening, and are refined after the onset of vision; binocular deprivation prevents this refinement. The orientation map, revealed by optical imaging of intrinsic signals, develops in parallel with the development of orientation selectivity in single cells.

Monocular lid suture after orientation maps have already formed disrupts the map from the closed eye; reverse lid suture restores the map precisely. In addition, matching orientation maps for the two eyes develop even in cats raised under an alternating-suture paradigm, so that the two eyes never have common visual experience. Together, these studies indicate that the emergence of orientation-selective responses in single cells and the overall layout of the orientation map do not require patterned vision, although visual experience is necessary for orientation selectivity and maps to fully mature. However, blocking electrical activity in V1 reduces both the orientation tuning of single cells and the clustering of horizontal connections. Thus, spontaneous electrical activity in cortex or in the thalamocortical loop is required for the initial establishment of local and long-range intracortical connections.

Patterned Electrical Activity and the Development of Cortical Networks

The vast majority of experiments on the influence of activity on visual cortical networks have involved manipulations such as lid suture or activity blockade that alter the amount of activity in the visual pathway. Relatively few experiments have examined the influence of the pattern of activity on cortical network development and maintenance; such manipulations include artificial strabismus, specific rearing paradigms, and rewiring of visual projections to the auditory pathway.

Strabismus refers to misalignment of the two eyes' optical axes, so that activity from corresponding retinal loci in the two eyes are no longer temporally correlated. Artificially induced strabismus causes neurons in V1 to become almost exclusively monocular, and ocular dominance columns to be more sharply delineated and have altered spacing compared to normal animals. Long-range horizontal connections within the superficial layers of V1 are also affected. In normal cats these connections cluster to link regions with similar orientation preference but do not align with ocular dominance columns, whereas in strabismic cats the connections come to link columns with similar eye dominance and orientation preference (Schmidt et al., 1999). Thus, correlations in input activity are used to sharpen and organize ocular dominance columns, and to shape intracortical horizontal connections.

Rearing kittens in a visual environment consisting of alternating black and white stripes at specific orientations restricts pattern vision and appears to shift the orientation preference of cortical cells toward the experienced orientation. While a shift could be caused by a passive loss of responses to nonexperienced orientations, there is an expansion of cortical columns devoted to the experienced orientation (Sengpiel et al., 1999), indicating an instructive effect of patterned visual experience on orientation selectivity and the orientation map.

Rewiring the Cortex

A unique paradigm for investigating the role of patterned afferent activity in the development of cortical circuitry and function is experimentally to redirect afferents carrying information about one sensory modality to central targets and pathways that normally process a different sensory modality (Schneider, 1973). Retinal axons can be induced to innervate the auditory thalamus in the ferret (Sur et al., 1988), a carnivore with a highly organized visual pathway and which is born at a very early stage in development. The routing of retinal projections to the auditory pathway can be induced at birth, allowing the probing of the role of patterned activity in the establishment (and not simply the maintenance) of cortical circuitry. Importantly, visual input is relayed from the auditory thalamus to the primary auditory cortex (A1) via thalamocortical projections whose physical identity is unchanged but which provide spatiotemporal patterns of electrical activity to the auditory cortex that are very different from normal (**Figure 1**).

Visual driving leads to emergent responses in 'rewired' A1 that demonstrate the role of patterned activity in shaping cortical networks. An orderly two-dimensional retinotopic map develops in A1 of rewired ferrets, in cortex that normally maps a one-dimensional surface, the cochlea. The retinotopic map in rewired A1 arises despite widely dispersed and overlapping thalamocortical projections, suggesting that well-localized receptive fields and their orderly progression are created by correlation-based mechanisms that operate intracortically to select and strengthen a specific subset of synapses from a broader set available anatomically.

Visual cells in rewired A1 have orientation-tuning, direction-tuning, and velocity-tuning properties that are quantitatively indistinguishable from V1 cells, suggesting that similar mechanisms operate in the generation of receptive field properties in the two cortices. Orientation-tuned neurons in A1 are organized into an orientation map that resembles in important respects the map in V1 (Sharma et al., 2000). The differences between orientation maps in V1 and rewired A1 reflect

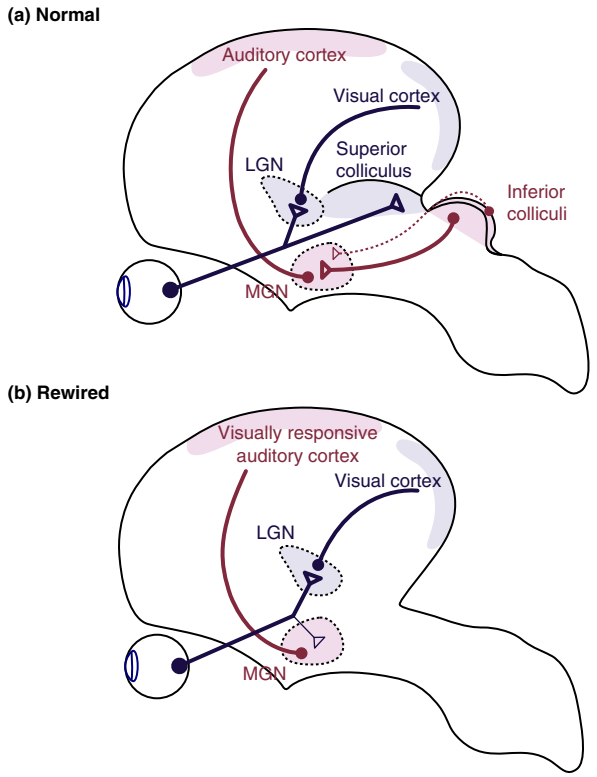

Figure 1 Retinal inputs routed to the auditory thalamus in rewired ferrets drive the auditory cortex with visual activity without altering thalamocortical projections. (a) Visual and auditory pathways in normal ferrets. The retina projects predominantly to the lateral geniculate nucleus (LGN) and superior colliculus. The LGN projects to the primary visual cortex. The medial geniculate nucleus (MGN) receives most of its subcortical afferents from the ipsilateral inferior colliculus, although afferent projections also arise from the contralateral inferior colliculus. The MGN projects to the primary auditory cortex. (b) Visual pathways in rewired ferrets. Subcortical inputs to the MGN in one hemisphere are removed in early postnatal ferrets. This induces retinal axons to innervate the deafferented MGN. The MGN still projects to the primary auditory cortex, as in normal ferrets, but in rewired ferrets it relays visual rather than auditory inputs. This change in spatiotemporal patterns of input activity early in development has a profound effect on networks in auditory cortex, and on its function. Adapted from Angelucci, A., Sharma, J., Sur, M., 2000. The modifiability of neocortical connections and function during development. In: Kaas, J.H. (Ed.), The Mutable Brain, Harwood, Amsterdam.

differences in the underlying organization of superficial-layer long-range horizontal connections in these two cortices. In V1, horizontal connections are patchy, spatially periodic, and extend mediolaterally. In contrast, horizontal connections in normal A1 are more band-like in organization, show little spatial periodicity, and are anisotropic anteroposteriorly. Horizontal connections in rewired A1 show features that are intermediate between V1 and normal A1: The connections in rewired A1 form much smaller and more regular patches than in normal A1, though the patches are less tightly clustered and are larger in size than in V1. Thus, horizontal connections within the rewired cortex are significantly altered by visual input, but in a manner that appears to be constrained by intrinsic features of the auditory cortex (see also Majewska and Sur, 2006).

Rewired ferrets also provide an opportunity for examining whether the behavioral role of a cortical area is set by intrinsic determinants or by the pattern of afferent activity during development. Behavioral experiments, supported by several kinds of control experiments, indicate that rewired ferrets interpret visual stimuli which activate the rewired projection as visual rather than as auditory (von Melchner et al., 2000). Thus, the function of a cortical area is dependent fundamentally on the spatiotemporal pattern of activity it receives during development. It is possible that all 'auditory' pathways central to the thalamus in the rewired hemisphere are turned 'visual,' including the cortex and downstream structures, with a concomitant respecification of their perceptual identity.

The rewiring studies are not unique to ferrets as similar results have been shown in hamsters (Frost et al., 2000), although the neurophysiological properties of the A1 neurons appear more similar to V1 cells in the hamsters (Ptito et al., 2001). In addition, some of the visually responsive cells in A1 are bimodal – they can respond to both visual and auditory stimuli.

Rewiring does not require surgical intervention as in the ferret and hamster studies but can also occur naturally in conditions such as congenital deafness (e.g., Barone et al., 2013; Butler and Lomber, 2013). In addition, it has been shown that the visual cortex can respond to nonvisual input in individuals with visual deprivation (e.g., Bedny et al., 2012).

Cortical development involves a progressive shaping of the fate of the cortical epithelium into discrete cortical areas with specific inputs, outputs, and local networks. Developmental manipulations, which include, importantly, transplantation and rewiring studies, demonstrate a continual interplay of intrinsic and extrinsic factors at all stages of development – at the ventricular zone during the cell cycle, in the cortical plate during the parcellation of cortical areas, and within the cortex during the formation and maintenance of cortical networks. The nature of extrinsic signals varies with developmental time, and likely includes intercellular signals in the ventricular zone that influence the formation of layers, trophic or permissive electrical signals in early area formation, and instructive electrical signals in late network development, which persist into adulthood as substrates for learning and memory.

See also: Brain Development: Ontogenetic Neurobiology of; Brain: Response to Enrichment; Neural Plasticity; Topographic Maps.

Bibliography

Angelucci, A., Sharma, J., Sur, M., 2000. The modifiability of neocortical connections and function during development. In: Kaas, J.H. (Ed.), The Mutable Brain. Harwood, Amsterdam.

Barone, P., Lacassagne, L., Kral, A., 2013. Reorganization of the connectivity of cortical field DZ in congenitally deaf cat. PLoS One 8. http://dx.doi.org/10.1371/journal.pone.0060093.

Bedny, M., Pascual-Leone, A., Dravida, S., Saxe, R., Brain, L., September 2012. A sensitive period for language in the visual cortex: distinct patterns of plasticity in congenitally versus late blind adults. Brain and Language 122 (3), 162–170. http://dx.doi.org/10.1016/j.bandl.2011.10.005.

Butler, B.E., Lomber, S.G., 2013. Functional and structural changes throughout the auditory system following congenital and early-onset deafness: implications for hearing restoration. Frontiers in Systems Neuroscience 7. http://dx.doi.org/10.3389/fnsys.2013.00092.

Chapman, B., Godëcke, I., Bonhoeffer, T., 1999. Development of orientation preference in the mammalian visual cortex. Journal of Neurobiology 41, 18–24.

Crowley, J.C., Katz, L.C., 2000. Early development of ocular dominance columns. Science 290, 1321–1324.

Frost, D.O., Boirfe, D., Gingras, G., Ptito, M., 2000. Surgically created neural pathways mediate visual pattern discrimination. Proceedings of the National Academy of Sciences of the United States of America 97, 11068–11073.

Law, M.I., Constantine-Paton, M., 1981. Anatomy and physiology of experimentally produced striped tecta. Journal of Neuroscience 1, 741–759.

Levitt, P., Barbe, M., Eagleson, K., 1997. Patterning and specification of the cerebral cortex. Annual Review of Neuroscience 1–24.

Majewska, A.K., Sur, M., 2006. Plasticity and specificity of cortical processing networks. Trends in Neurosciences 29, 323–329.

McConnell, S.K., Kaznowski, C.E., 1991. Cell cycle dependence of laminar determination in developing neocortex. Science 254, 282–285.

von Melchner, L., Pallas, S.L., Sur, M., 2000. Visual behaviour mediated by retinal projections directed to the auditory pathway. Nature 404, 871–876.

O'Leary, D., Schlagger, B., Tuttle, R., 1994. Specification of neocortical areas and thalamocortical connections. Annual Review of Neuroscience 17, 419–440.

Ptito, M., Giguer, J.F., Boire, D., Frost, D.O., Casanova, C., 2001. When the auditory cortex turns visual. Progress in Brain Research 134, 447–458.

Rubenstein, J., Rakic, P., 1999. Genetic control of cortical development. Cerebral Cortex 9, 521–523.

Sanes, D.H., Reh, T.A., Harris, W.A., 2012. Development of the Nervous System, third ed. Academic Press, Oxford, UK.

Schmidt, K., Galuske, R., Singer, W., 1999. Matching the modules: cortical maps and long-range intrinsic connections in visual cortex during development. Journal of Neurobiology 41, 10–17.

Schneider, G.E., 1973. Early lesions of superior colliculus: factors affecting the formation of abnormal retinal projections. Brain, Behavior and Evolution 8, 73–109.

Sengpiel, F., Stawinski, P., Bonhoeffer, T., 1999. Influence of experience on orientation maps in cat visual cortex. Nature Neuroscience 2, 727–732.

Sharma, J., Angelucci, A., Sur, M., 2000. Induction of visual orientation modules in auditory cortex. Nature 404, 841–847.

Sherman, S.M., Spear, P.D., 1982. Organization of visual pathways in normal and visually deprived cats. Physiology Review 62, 738–755.

Sur, M., Garraghty, P.E., Roe, A.W., 1988. Experimentally induced visual projections into auditory thalamus and cortex. Science 242, 1437–1441.

Coser, Lewis A. (1913–2003)

Thomas Crosbie, Yale University, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Lewis A. Coser (1913–2003) was a leading figure in American sociology in the second half of the twentieth century. Born in Germany, he fled to Paris in 1933 and later to the United States in 1941, where he met his future wife and fellow sociologist, Rose Laub Coser. He gained recognition with his first book, *The Functions of Social Conflict* (1956), as well as for later work on the political role of the intellectual and his influential textbooks. Coser also maintained a career as a socialist intellectual, primarily through *Dissent*, the journal he cofounded with longtime collaborator Irving Howe.

Ludwig Cohn, 1913–33

The man who would become Lewis Coser was born in Berlin in 1913. His father was a banker and broker with the surname Cohn. He was Jewish but as Coser later recalled, "blue-eyed and blond … more German than most Germans" (Rosenberg, 1984). The elder Cohn was partly assimilated, quite patriotic, and though his closest friends were fellow Jews, he had married a Protestant woman from a lower-class family. A son was born 27 November 1913 and named Ludwig Cohn; and at some point, a daughter was added to the family. The family fortunes continually improved and while still a boy Ludwig Cohn moved with parents and sister to a 150-acre estate in Charlottenburg, where his father "play[ed] the highly respected lord" (Rosenberg, 1984) (Biographical details of Coser's life are taken from the following sources, which are drawn upon throughout this article: Coser's (1993) autobiographic essay for ARS; Powell and Robbins's (1984) essay in the *Festschrift*; Rosenberg's (1984) interview with Coser in the *Festschrift*; Coser's (1988) introduction to a collection of his papers; his New York Times obituary (Martin, 2003); an obituary by Stephen Kalberg (2004); and an obituary in *Dissent* (Rule, 2003). Some additional details are drawn from the reminiscences of Coser published in *Footnotes*, including those of Judith Blau, Donna Gaines, Bill Gamson, Stephen Kalberg, Donald N. Levine, Aldon Morris, Steven Nock, Walter W. Powell, Philip Selznick, Barrie Thorne, and Eviatar Zerubavel. Invaluable assistance was offered by Coser's students Walter W. Powell and Stephen Kalberg. Additionally, it should be noted that Coser's son Stephen compiled an excellent bibliography of his father's works in the *Festschrift*, although this extends only to 1983.)

The banker Cohn ultimately made the decision to push his children further along the road of assimilation, and changed their surname to Coser. A stern disciplinarian with bourgeois attitudes and aristocratic pretentions, the father disappointed the son. Coser later recalled, "whatever I became was partly a reaction to that" (Rosenberg, 1984). The reaction took the form of political activism: Ludwig Coser became intensely involved in German socialist circles and soon found confirmation for his rejection of paternal values in the excesses of the Weimar Republic. Politics continued to drive father and son apart, while Ludwig's sister moved to Shanghai with a new husband. The parents soon separated.

In 1933, Hindenburg appointed Hitler as the Reich Chancellor and Ludwig Coser was warned that his socialist activism had made him a target. He fled to Paris after a brief period studying in England but his father, disdaining his son's socialist politics, soon cut off the stream of money. Coser was broke, and "lived miserably, starving for a while" (Rosenberg, 1984). Down and out, nursing drinks at cafes, and secretly reading through the wares at bookstores, Coser found himself in a liminal class of refugee intellectuals simmering in the poor places of Paris.

Refugee in Paris, 1933–41

A work permit did eventually arrive and with a stroke of luck led to an interview to work as the assistant to the Swiss author Ludwig Bauer, who was then beginning research for a biography of King Leopold. Coser landed the job and although it paid well, Coser soon ran through his savings and was once again down and out in Paris. Later, he found work at an American brokerage house and literary work with André Gide (Additional information for this section was drawn from Jacoby (1984).)

Throughout this period, Coser's work had little to do with his true passion, which was to live as an organic intellectual, contributing his full mental efforts to establishing socialism as a global political project. Coser worked his way through the feuding socialist sects, subjecting again and again their schismatic doctrines to systematic interrogation, only to find them wanting. He flirted with Communist orthodoxy and was at least acquainted with the apologia for Stalin's dictatorship, but found himself with a deepening suspicion of empowered bureaucracies. A chance discovery of a library of French Revolutionary material with his friend Henry Jacoby, further set him along a path of critical, democratic socialism, and away from the Comintern.

Through free classes at the Sorbonne, Coser was exposed to an academic base for his intellectual pursuits. He achieved enough success that he could consider an advanced degree in comparative literature, looking to novels as a means of understanding English and German social structures. His advisor's reaction led him to switch to sociology and Coser would remain committed to this discipline (in addition, however, to his socialist work) for the rest of his life.

Sociology at the Sorbonne was then dominated by Célestin Bouglé, who hewed closely to a doctrinaire form of Durkheimianism which included rote memorization of the master's texts. Although in close proximity to many leading European sociologists, Coser's engagement with academic sociology remained constrained at this point; the Sorbonne sociologists were little interested in Weber, Simmel, or even Marx. While the academic arena offered structure and the hope of employment, it was the local discussion groups and tiny sects (many with their own journals) that offered Coser the bulk of his intellectual stimulation.

For some time after the outbreak of war, Coser seems to have been unaffected by the suffering felt throughout Europe. Coser knew that his father had left Germany to join his sister in Shanghai (his mother's fate is uncertain). Coser himself had adopted a pacifist position and remained well behind the lines. By the mid 1930s, he spoke perfect French, was deeply engaged in antifascist political movements, and so may well have had reason to think that he was something other than a 'Boche' to his neighbors. This notion was dispelled when French authorities rounded him up with the other German émigrés.

Various detainments and humiliations would follow. In his seventies, he could recall vividly a night spent in a sports arena with other detainees, expecting to be gassed: "one guy went mad and screamed throughout the night. I can still hear that" (Rosenberg, 1984). And yet, the worst did not happen. Through resourcefulness, through reliance on an extensive network of friends, and of course through good fortune, Coser received a coveted exit visa, secured passage across the Atlantic, and found himself in New York City.

A Second Refuge, 1941–51

The 28-year-old Ludwig Coser who arrived in the United States was a committed democratic socialist, had extensive experience as partisan and organic intellectual, and had trained for several years in Durkheimian sociology. Over the next 10 years, he was to rerun each of these aspects of his life in this second refuge, a refuge where, on the advice of an immigration official, he took on a new name, Lewis; or, to his friends, Lew. (Additional information for this section is drawn from Jumonville (1990) and Rodden (2004).)

Lewis Coser began his life in the United States on a promising note. His caseworker with the International Relief Association was the very intelligent and politically engaged Rose Laub (herself an émigrée), who would connect him to young New York Trotskyists, including Dwight Macdonald, Seymour Lipset, Daniel Bell, and Irving Howe, with whom he began to work through the socialist factions in New York. He also soon found steady employment, first as a freight handler and later for the Department of War in a variety of menial roles. Wary of causing problems with his government employer, Coser took on the pen name Louis Clair and began publishing in literary journals (*The Call, Challenge, Forum, Partisan Review, Politics, and Socialist Review*). As Louis Clair, Coser expounded a political vision that reflected his preferred Schachtmanite concepts and contributed to Max Schachtman's notion of a Third Camp (a variation of Trostky's Common Front that seemed to address Coser's concerns from his Paris days).

From a chance reacquaintance at a Schachtmanite party, Coser and Rose Laub were soon living together and they married. The birth of their first child eventually lead Rose Coser to study under Robert Merton at Columbia University. In 1948, Lewis Coser was 35, had a wife and child, and had defined a political position within the Schachtmanite camp. Later in life, Coser recalled: "Rose and I talked one day about one of our friends who had gotten a job teaching German at a little college. We said to each other, 'Do you think we might possibly achieve something like that?'" (Rosenberg, 1984). With this modest goal in mind, he followed Rose as a PhD student of Merton and Robert Lynd in the sociology department at Columbia.

Coser enrolled at Columbia and began to commit himself to a life both as an independent intellectual and an academic sociologist. Before he could begin, however, he received the surprising offer to teach at the University of Chicago, recruited by David Riesman as part of an expansion effort. In Chicago the Cosers found themselves in an exceedingly rich intellectual environment, which Coser would later describe as a "surfeit of honey" (Coser, 1993). Rose began to assist Riesman in his research for *The Lonely Crowd* while Lewis continued his studies at Columbia by distance, taught at Chicago, and learned some hard lessons in departmental politics.

The Frontiers of Structural Functionalism, 1951–54

The Chicago sojourn came to an end and the Cosers returned to New York, where they found that Columbia was also undergoing an intellectual rejuvenation. Finally dedicating himself full time to his studies, Coser was struck by the feeling that he was "working at the frontiers of sociology," noting later that it was "a joy to be alive in the Columbia atmosphere in the late 1940s and early 1950s" (Coser, 1993).

Coser is perhaps best known today as a precursor of the broad rejection of Parsonian functionalism's dominant place in the academy. The experience in Chicago had awakened both Cosers to the breadth of American sociology and their positive view of sociology was further confirmed by Merton, whom both quickly befriended. Rose found a mentor and intellectual interlocutor in Merton whom she would follow for much of her career. Lewis's connection to Merton was also deep; but his response to Merton's influence would be one of creative synthesis rather than intellectual indebtedness.

At the time, Merton was given a free hand by his colleagues in guiding the department and he set his sights on turning it into a home for structural functionalism. His approach was deeply indebted to his mentor Talcott Parsons but stressed new methods, a wide reading in the history of sociology, and dedicated engagement with theoretical issues. Thus, the functionalism Coser encountered at Columbia was already a reform version, one generation removed from Parsons.

While Parsons's star had already begun to dip, Coser retained a great deal from this first generation of structural functionalist thought. Coser noted later in life, "I had enormous respect for [Parsons] and had no sympathy for many of his critics who attempted to picture him as a reactionary servant of the American power elite" (Coser, 1993). He instead viewed Parsons as a classic reform liberal,

blinded perhaps by too much faith in stability yet brilliant nonetheless.

It was nevertheless a contentious intellectual kinship. One of Coser's first publications under his own (American) name was a review of Parsons's *Essays in Sociological Theory*. This collection was significant for Parsons since it represented the first attempt to integrate psychology into his framework, a move that would come to dominate his thinking. Coser criticized the move soundly by noting that "functional analysis becomes ideology – the ideology of Candide. … [This] leads to a basically conservative interpretation of functionalism which is not necessarily inherent in the theory as such" (Coser, 1950). In a section where Parsons defends the use of propaganda, Coser accuses him of "advocating manipulation and a sort of intellectual rape" (Coser, 1950). This review in the *American Journal of Sociology* is a remarkable case of a graduate student savaging the foremost sociologist in the country – an appropriate representation of Coser's intellectual independence.

Functions and Academic Success, *c.*1954

Coser's initial idea for his dissertation was an intellectual biography of Georg Simmel whose work had been pushed to the side in the wake of the rising Parsonian orthodoxy (Jaworsky, 1997). Merton disliked the dissertation idea, leading Coser to shift the project to become an extended engagement with Simmel's *Conflict*. This plan was accepted by Merton and the result would become Coser's breakthrough work, *The Functions of Social Conflict*.

At an early point Coser's dissertation began to spread beyond Simmel to encompass the concept of conflict in early American sociology. He found that American sociologists stand in a long tradition of liberal advocacy and reform, much more directly so than German and French sociologists. Indeed, the original audience for American sociological research came from the reformers at the University of Chicago, which allowed early reformer-sociologists like Robert Park the resources needed for systematic work in the city of Chicago. Coser argued that Park and his colleagues placed conflict at the forefront of both sociological theory and sociological work, working from an understanding that conflict had both positive and negative qualities. Even partial reformers such as Sumner or Giddings had a clear understanding of the importance and duality of conflict, although they were more inclined to stress the negative aspects than the positive.

Coser's review of Parsons in the *American Journal of Sociology* anticipated the first part of his dissertation where he argued that the institutionalization of sociology by figures such as Parsons at Harvard had diminished the place of conflict and reform in sociology. This loss led to an unnecessary tendency among functionalists to focus on stability and continuity rather than change and disruption. For Coser, perhaps the single greatest intellectual cost of this conservative drift was the loss of the recognition of the possibility of things improving through conflict and the tendency of conflicts to generate coordination, solidarity, and integration.

The first part of Coser's dissertation armed him with considerable resources in his critique of existing structural functionalist thought. He would draw on the classics to reform the orthodoxy of his day. Coser considered Simmel's work to have most clearly articulated the importance of conflict in social life and so the second part of his dissertation focused entirely on its explication and has come to define Coser's theory of conflict.

Although at times adopting Simmel's style, Coser departs significantly from Simmel's propositions by twisting the very word 'conflict' to cover a broader range of social acts than was generally the case, particularly to include ideational states such as 'dissent' (Cuber, 1957). Indeed, Coser, following Weber more than any American precursor, was eager to capture a richer sense of the culture-infusing nature of conflict than the more formal model offered by Simmel.

Coser's major contributions to the theory of conflict are threefold. First, he introduces the distinction between realist and nonrealist conflict. Realistic conflicts use conflict as a means to achieve some goal. Nonrealistic conflicts posit conflict itself as an end, sometimes as the result of losing sight of original objectives or often through interested parties, eager to benefit from the side effects of conflict.

A second major insight is his refinement of the place of political ideology in conflict. Intellectuals enter here as key figures, since they are uniquely capable of defining objects of contestation. His insight is that collective action is easier to manage when there are objective elements of contestation.

A final major insight is again a sort of Weberian supplement to Simmel's model. While Simmel attempted to show continuity across levels of conflict, Coser stresses the important differences inherent in modern warfare, which involve both national ideologies and vastly more complex role differentiation and structural configurations. By doing so, Coser folded the meso-level of organization into the study of conflict, in a way that Simmel had resisted.

Coser's dissertation allowed him to synthesize his long-term commitment to radical politics with his own growing investment in the United States as a relatively stable political entity, where he would raise his children and find security as a semiautonomous intellectual. He could now look at liberals like Parsons not as enemies rejecting social change but as potential allies able to reach back to a long tradition of reform. He could look to socialism as a project ever in the making, an endless campaign against injustice and ignorance waged by intellectuals within the very system they sought to change.

Dissent and Intellectual Success, *c.*1954

Coser's first book, so important for his intellectual development, received mostly positive reviews but at first did little for his academic career. This was beside the point, however, since his secondary Columbia advisor, Robert Lynd, recommended Coser for a position at Brandeis University. Rose found work at Northeastern University, so the move made sense. Brandeis was then a new experiment without traditional academic divisions. At first Coser was alone, free to teach what he wanted and burdened with few administrative duties; but he soon discovered that the lack of structure created difficulties in recruiting other sociologists. The environment at Brandeis was "stimulating," Coser recalled later, but in the words of Irving

Howe, it was a "lively, irreverent, not-very-academic atmosphere" (Howe). The outcome was that Coser was again drifting more toward the role of engaged intellectual than to the academic side of his twin careers (Additional information for this section is taken from Howe (1984).)

For many years, Coser and Irving Howe had shared an intellectual network without becoming close friends. Both were reformed Schachtmanites and both contributed extensively to literary journals; but while Coser began finding moderate success in his secondary career as an academic, Howe had remained an unattached professional writer and intellectual. Shortly after the Cosers moved to the Boston area, Howe decided to pay them a visit in the hope of entangling Lewis in a literary venture. Their conversation convinced both men that they were kindred spirits. They were right: they remained writing partners until Howe's death 40 years later.

Coser and Howe cemented their friendship with the decision to launch a journal of ideas, for which Coser supplied the name *Dissent*. Their idea was to present socialist pieces to the literary public without falling into what they viewed as the three principal dangers: advancing a party line, publishing inferior work by friends, and publishing work that picked over Marxist minutiae (Howe, 1984). Coser was then still at work on his dissertation in addition to teaching classes at Brandeis. Despite this heavy workload, he and Howe quickly launched the journal in 1954 with enough funding to see it through a full year's run. They expected it would fold after a year, which they considered would still be a "noble effort" (Howe, 1984). The first issue generated some interest but was considered by both the editorial team and by many of the reviewers to be theoretically slight. Accordingly, for the second issue, Coser and Howe published 'Images of Socialism,' an extended meditation on unsuccessful socialist experiments characterized by failures of intellectual effort and the dangers of bureaucratic empowerment. Despite these failures, they affirmed socialism as Utopia, a 'no-place' that is nevertheless "the name of their desire," – a project ever in the making. The piece was partly adapted from an earlier article by Coser but it was the reworked version with Howe that would become a classic among American socialist writings. Howe would later note, "I imagine Lew would agree with my conviction that it is one of the best things either of us has done" (Howe, 1984).

The year that 'Images of Socialism' was published was something of an *annus mirabilis* for Coser. With Howe, he had both launched a successful journal and provided a socialist credo to guide it. He finally finished his dissertation and, at the age of 41, was awarded his PhD. At Brandeis, he had found an ideal environment to foster his political writing. Perhaps more importantly, both he and Rose were successfully employed as scholars and so had already achieved the modest dream they shared some ten years earlier.

Conflict Coser, 1954–68

At this stage Coser's immediate plans were to revise his dissertation into a book and start work with Howe on a history of *The American Communist Party*. In 1956, Coser published *The Functions of Social Conflict*, and in the next year published *The American Communist Party: A Critical History, 1919–1957*. The latter was both an intellectual history and scholarly analysis of sectarianism, in many ways an extension of 'Images of Socialism' that allowed Coser and Howe to sharpen their claims that dysfunction at the organizational level is a major obstacle to achieving the utopian vision shared by socialists and communists.

That same year, Coser also published a textbook *Sociological Theory* (1957) with Bernard Rosenberg. The textbook (much of which was excised from *The Functions of Social Conflict*) allowed Coser to make use of his research on the history of both European and American sociology. Meanwhile, *Dissent* was still going strong, although as Howe later noted, the journal quickly began "shedding more and more ideology; by about 1960 most of us no longer thought of ourselves as Marxist" (p. 66).

While the journal drifted away from Marx, the broader discontent at the heart of the journal was being echoed in dissent movements and acts of civil disobedience that seemed to be filled with the possibility of Utopia. The new tensions and conflicts rising to public awareness sorely tested the consensus-orientation of Parsons and his followers, including Merton. Instructors were faced with teaching material that seemed very out-of-date and they flocked to the three available books about conflict: Ralf Dahrendorf's (1959) *Class and Class Conflict in Industrial Society*, Max Gluckman's (1955) *Custom and Conflict in Africa*, and Coser's own *The Functions of Social Conflict*. The sudden and sustained popularity of the book boosted Coser's academic standing considerably. He was suddenly 'Conflict Coser,' which seemed apt enough, given the relative emphasis on conflict in his textbook and the considerable place of conflict in 'The American Communist Party,' as well as a stream of articles bearing on the subject. By 1964, his professional profile was such that he became President of the Eastern Sociological Society.

Coser next set to work on his most Weberian and Mertonian project to date, an ambitious work on the sociology of knowledge and culture, dedicated to understanding the place of intellectuals in political life, published as *Men of Ideas* in 1965. *Men of Ideas* resembles in several ways Jürgen Habermas's (1991[1962]) *The Structural Transformation of the Public Sphere*, which Coser had not read. (From personal correspondence with Walter W. Powell.) The main commonality rests in the fact that both authors are attentive to the relationship between spatial configurations and deliberative cultures as the proper starting point for understanding democratic political life. *Men of Ideas* reveals a probing engagement with the relationship between a mass public and those intellectuals who are both expert in understanding the public, yet unwilling to bend to serve it. This for Coser is the *sacrificium intellectus* an inevitable trade-off between gaining an audience and retaining intellectual honesty and independence.

Following *Men of Ideas*, which was well received, Coser expanded on his first book in the *Continuities in the Study of Social Conflict* (1967). This was primarily a compilation of Coser's articles and political writings on conflict, although it included some new pieces. The result was more tepidly received, with one reviewer expressing frustration that Coser did not attempt to expand the theoretical edifice of his first book. While his 'Conflict Coser' persona had already grown wearisome, the discipline continued to read *The Functions of Social Conflict*.

"Inside the Whale," 1969–75

The publication of *Continuities in the Study of Conflict* came at a time of growing unease for both Cosers. Powell and Robbins note that, "[b]y the late 1960s … a good deal of Brandeis's vitality had been sapped by internal conflicts, severe political polarization, and perhaps simply the sheer liabilities of aging" (Powell and Robbins, 1984). Coser himself typically viewed it through a Weberian lens, noting that Brandeis "became routinized like all charismatic institutions. By the time we left, 18 years later, I was glad to leave" (Rosenberg, 1984). Rose too had grown unhappy at Northeastern. Coser was approached by both Amherst College and the State University of New York at Stony Brook but neither considered Rose's future position. Their friend at Stony Brook, Hanan Selvin, understood that the Cosers were a team, and made sure that the next offer included both. The offer also included the chance to work with Selvin in reconstructing the department. They accepted and moved to Long Island in 1969 (Additional information for this section was drawn from the recollections of Coser's student Walter W. Powell.)

In his autobiographical writings, Coser would often point to feelings of marginality, of being "a stranger within the gate" (Coser, 1988). By contrast, he would describe his arrival with Rose at Stony Brook, he as a distinguished professor, she as a full professor, as final confirmation that they were "inside the whale" (Coser, 1988). The Cosers invested themselves deeply in aiding Selvin in building the department, to the point where their home became a regular setting for faculty events, PhD seminars, and other functions.

The first year at Stony Brook was spent on what Merton described as a classic. Coser's *Masters of Sociological Thought* (1970), is a productive synthesis of his popular textbook, *Sociological Theory*, his conflict research, and his work in the sociology of knowledge. *Masters of Sociological Thought* unfolds detailed biographical and sociological portraits of 12 classical figures (including the European fathers of sociology, Comte, Marx, Spencer, Durkheim, Simmel, and Weber, as well as the early American reformer sociologists Veblen, Cooley, Mead, and Park). As with his textbook, this is written with a light hand and aimed in part at students and nonspecialist readers. But there is a depth to the work, drawing both from Coser's sense of productive conflicts and from his theoretical interests in intellectuals as key political actors. The outcome is a set of prismatic encounters with the great figures of the discipline, providing new insights into the connection between audience, political context, and classical sociological texts.

Coser experienced a second year of remarkable success in 1974. His administrative efforts with Selvin were paying off in a rapidly growing department. He also published *Greedy Institutions: Patterns of Undivided Commitment* (1974), which marked a turning point in Coser's decades-long engagement with the notion of sects. The book included several previously published pieces, including a study of the modern housewife written with Rose, but deeper connections emerge in his new material. The idea of a "greedy institution" that is oriented to taking more and more from the individual allowed Coser to reconsider his experiences among socialist sects, viewing them within the context of broader societal patterns of conflict. As the sect attempts to take more and more resources from its members in its conflict with the rest of society, it repeats the familiar pattern of strengthening internal ties by breaking down existing relationships to the broader society and triggering nonrealist conflict dynamics. The result may be personally frustrating (as in Coser's experience with socialist sects) or profoundly devastating (as with the cultists at Jonestown). A corollary would be to once again reaffirm the importance of unattached intellectuals capable of speaking the truth to power and of realigning groups to real goals.

In addition to his productive work at Stony Brook and the publication of *Greedy Institutions*, Coser realized a considerable professional achievement in 1973, when he was named the incoming President of the American Sociological Association. Typically, he saw that as an opportunity for productive conflict and political reform, in this case by means of a highly provocative presidential address titled 'Two Methods in Search of a Substance.' The 'methods' in question were ethnomethodology and the status-attainment research associated with the University of Wisconsin, which Coser viewed as communities of methodologists rather than as coherent theoretical traditions. Coser also sees here the logic of sectarianism. His address ends with the evocative warning that the discipline will "degenerate into congeries of rival sects and specialized researchers who will learn more and more about less and less" (Coser and Larsen, 1976). The address would give rise to a lively series of responses that would be published the following year as *The Uses of Controversy* (1976).

Around the same time Coser once more joined with Howe in a collaboration designed to stir controversy in the political arena. Their highly critical collection *The New Conservatives* (1973) advanced a very similar thesis to his critique of sociology. In both his presidential address and in this piece, Coser spoke to the new generation of writers (academic and intellectual), who he saw as hopelessly mired in internal and ultimately nonrealist conflicts. He warned them that their efforts served only to harden positions and deepen divisions. He told them that little true intellectual work was being done, little attention was being paid to social rather than sectarian concerns.

Sociologist and Socialist, 1976–2003

Coser was 62 when he was President of the ASA, a considerable accomplishment given that his career then only spanned two decades. Coser retired in 1987. The 12 years in between, saw him maintain his considerable productivity and professional standing. He undertook two more major projects, *Books: The Culture and Commerce of Publishing* (1982) and *Refugee Scholars in America* (1984). He also wrote a library of textbooks and edited volumes, political essays, and reviews. He had, however, established his core ideas and these projects were elaborations rather than revisions or redirections in his political and academic vision.

In *Books*, Coser (aided by his coauthors Charles Kadushin and Walter W. Powell) for the first and only time in his career, followed the traditional process of securing a major grant, collecting and analyzing data with the help of research assistants, and synthesizing the data in a series of articles and a manuscript. This reorientation to the empirical rather than

conceptual realm was both a professional success and another reflection of his enduring fascination with the very social character of knowledge.

The publishing project and the refugee scholars project further deepened his exploration of the intellectual as political actor, offering new insights into the institutional structures that intervene in the work of organic intellectuals. Of the two projects, the latter seemed particularly significant to both Cosers. The idea of 'coming to America,' treated for the first time at length in Lewis's last book, was a topic of pressing interest to Rose as well. Both approached the topic with rich biographical understanding but also, both treated it without the romantic or tragic grandeur often associated with the theme.

Coser prided himself on taking on the challenge of integration with good will and sincerity, never scorning what was American as many other intellectual refugees did. This second refuge was, after all, a permanent one. The Coser children were born and educated in the United States. In addition, Coser's father and sister eventually left Shanghai to join Lewis and Rose in America. In a revealing incident, the father and son were reunited after years of estrangement in tears and in laughter, as the father joked, "I lost all my money. Now you can be a socialist all you want. It doesn't make any difference to me" (Rosenberg, 1984).

Coser's Legacy

Lewis Coser's legacy must be connected to Rose's legacy. Both Cosers received the honor of Festschrifts and Rose was the first female sociologist so honored. Lewis Coser repeatedly stressed she was a major contributor to all of his works and said of her: "[s]he has had so intimate a part in my intellectual and emotional life that I find it almost impossible to sort out ideas of hers that later cropped up as mine. Ours continued to be an ideal collaboration of mind and a common sensibility that nothing could shake throughout the common joys and trials of half a century" (Coser, 1993). Rose Laub Coser died in 1994, a year after Irving Howe's death. Despite the loss of these great figures in his life, Coser remained intellectually active for some years to come, writing his last book review in 1999. He died in 2003 at the age of 89.

Coser, like his father, had the ability to face quite considerable obstacles with humor and optimism. This optimism is clearly discernible in the brand of democratic socialism that he continued to evolve. By the time he wrote 'Images of Socialism' with Howe, Coser no longer conceived his political project in terms of a social movement or grand ideology. Instead, his mature political writings reflect a contingent, contextual form of socialist agitation. It is here that his two careers can be seen to merge. Although strenuously adhering to Weber's ideal of 'value-neutrality,' this form of agitation is visible throughout his scholarly work. In his two books on conflict, Coser outlined how political dissent could foster richer democratic outcomes. In his many books in the sociology of knowledge and intellectuals, he stressed repeatedly the duty of the scholar to 'whole-of-society' concerns, the risks of sectarianism, and the great danger of reifying the bureaucratic structures of governance.

Coser pointed out with pride that his legacy was carried on by his many students, among whom he singled out George Becker, Bob Hoffman, Arthur Mitzman, Stephen Kalberg, and Walter W. Powell. His political legacy also endures in the form of the journal *Dissent*, now almost 60 years old; and here Coser would often point particularly to his successor at that journal, Michael Walzer, and others among this younger generation who sought to balance the roles of both scholar and intellectual. More diffusely, however, Coser's legacy is reflected in all forms of academic and independent critical sociology, those efforts dedicated to principled critique – whether in challenging conflicts that divide or equilibria that repress.

See also: Functionalism, History of; Intellectuals, Sociology of; Knowledge, Sociology of; Merton, Robert K (1910–2003); Parsons, Talcott (1902–79); Political Sociology; Simmel, Georg (1858–1918); Socialism; Sociological Theory; Sociology: Overview.

Bibliography

Coser, L.A., 1950. Review of essays in sociological theory, by Talcott Parsons. American Journal of Sociology 55 (5), 502–504.
Coser, L.A., 1956. The Functions of Social Conflict. The Free Press, Glencoe.
Coser, L.A., 1965. Men of Ideas: A Sociologist's View. The Free Press, New York.
Coser, L.A., 1970. The Masters of Sociological Thought: Ideas in Historical and Social Context. Harcourt Bruce Jovanovich, New York.
Coser, L.A., 1974. Greedy Institutions: Patterns of Undivided Commitment. The Free Press, New York.
Coser, L.A., 1984. Refugee Scholars in America: The Impact and Their Experiences. Yale University Press, New Haven.
Coser, L.A., 1988. A Handful of Thistles: Collected Papers in Moral Conviction. Transaction Books, New Brunswick, NJ.
Coser, L.A., 1993. A sociologist's atypical life. Annual Review of Sociology 19, 1–15.
Coser, L.A., Howe, I., 1957. The American Communist Party: A Critical History, 1919–1957. Beacon Press, Boston.
Coser, L.A., Howe, I. (Eds.), 1973. The New Conservatives: A Critique from the Left.
Coser, L.A., Kadushin, C., Powell, W.W., 1982. Books: The Culture and Commerce of Publishing. Basic Books, New York.
Coser, L.A., Larsen, O. (Eds.), 1976. The Uses of Controversy in Sociology. The Free Press, New York.
Coser, L.A., Rosenberg, R., 1957. Sociological Theory. Macmillan, New York.
Cuber, J.F., 1957. Review of functions of social conflict, by Lewis A. Coser. Social Problems 4 (4), 350–351.
Dahrendorf, R., 1959. Class and Class Conflict in Industrial Society. Stanford University Press, Stanford.
Gluckman, M., 1955. Custom and Conflict in Africa. Oxford University Press, Oxford.
Habermas, J., 1991[1962]. The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society (T. Burger, Trans.). MIT Press, Boston.
Howe, I., 1984. Forming *Dissent*. Conflict and Consensus: A *Festscrift* in Honor of Lewis A. Coser. The Free Press, New York.
Jacoby, H., 1984. The Years in France. Conflict and Consensus: A *Festscrift* in Honor of Lewis A. Coser. The Free Press, New York.
Jaworsky, G.D., 1997. George Simmel and the American Prospect. State University of New York Press, Albany.
Jumonville, N., 1990. Critical Crossings: The New York Intellectuals in Postwar America. University of California Press, Berkeley.
Kalberg, S., Jan. 2004. Lewis A. Coser (1913–2003). Perspectives: Newsletter of the ASA Theory Section 27 (2), 9, 20–3.
Martin, D., 12 July 2003. Lewis Coser, 89, Sociologist Who Focused on Intellectuals. New York Times.
Powell, W.W., Robbins, R., 1984. Lewis Coser: Intellectual and Political Commitments. Conflict and Consensus: A *Festscrift* in Honor of Lewis A. Coser. The Free Press, New York.
Rodden, J., 2004. The worlds of Irving Howe. The Kenyon Review 26 (1), 157–174.
Rosenberg, B., 1984. An Interview with Lewis Coser. Conflict and Consensus: A *Festscrift* in Honor of Lewis A. Coser. The Free Press, New York.
Rule, J., 2003. In memoriam Lewis Coser: 1913–2003. Dissent 50, 4.

Cosmopolitanism

Huon Wardle, University of St Andrews, St Andrews, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G.L. Ribeiro, volume 4, pp. 2842–2845, © 2001, Elsevier Ltd.

Abstract

Cosmopolitanism encompasses four distinct but overlapping perspectives: (1) an identification with the world or with humanity in general that transcends local commitments; (2) a position of openness and or tolerance toward the ideas and values of distinct others; (3) an expectation of historical movement toward global peace; and (4) a normative stance advocating cosmopolitan aims and actions. Cosmopolitanism has a collective aspect, theorized among others by Durkheim, and an individual-centered aspect, elaborated particularly by Kant. Wide-ranging debates exist concerning cosmopolitanism as a stance and as an object of study. The period since 1989 has, however, seen a strong resurgence of interest in cosmopolitan themes.

Cosmopolitanism encompasses at least four overlapping perspectives on individual and collective life. Each can follow from the other, but they may also stand alone, and in some formulations may conflict with one another. (1) The first and broadest view holds that there can be commitments to the world at large that transcend loyalty to nation, neighbors, and locally shared conventions. This view, at varying levels of intensity, can be discovered in otherwise quite distinct religious, political, and artistic settings. (2) In a second usage, cosmopolitanism describes an attitude of openness to, tolerance toward, and a desire to comprehend, the standpoints of others and their ways of being. This can range from lighthearted appreciation of the uniqueness and diversity of human lives to more formalized modes of hospitality toward others and their distinct worldviews. (3) Some cosmopolitan views rest on a vision or expectation of historical progress toward peaceful world community. Human conflicts, it is proposed, have been, and can be, surmounted over time through a progressive institutionalization of world citizenship and equitable global governance. (4) A further perspective is that cosmopolitanism provides normative principles for intellectual, political, or aesthetic emancipation. The most general of these is an ethic of world solidarity or *amor mundi*, love of the world. Historically, diverse models of responsibility and reciprocity have been adduced to support normative standpoints. What unifies normative cosmopolitanism is that it does not rest with what *is* or *might be* the case, but instead it prescribes what *should be* done. Noticeably, each of these four perspectives implies processes of subjective and or group change, because each one either encourages or takes for granted practices which will perforce alter the make up of the individual and collective pattern of life.

For social scientists, multiple puzzles present themselves in exploring the meaning and implications of cosmopolitanism. Of the four interrelated views above, each has many possible combinative shades of interpretation: for example, 'openness' may not imply the same attitudes as 'tolerance' even if the two often overlap. Likewise, research will have distinct foci. Where cosmopolitan phenomena are confirmed as valid objects of empirical inquiry, *methodological* questions arise. These include, e.g., how best to distinguish cosmopolitan motives or mechanisms ethnographically, historically or under experimental conditions, and how to gauge the specificity of cosmopolitan commitments. As cosmopolitan concerns cause social scientists to change or rework their models, *ontological* and *epistemological* debates are also foregrounded. These include, for example, whether particular social science disciplines are more cosmopolitan in orientation than others, or what are the consequences of labeling certain objects 'cosmopolitan.' *Normative* or ethical quandaries likewise emerge around whether the social sciences should work toward cosmopolitan goals and if so what this would imply ethically and otherwise.

The current burst of academic interest in cosmopolitanism is of quite recent origin, gathering pace from the 1990s onward, however, critiques of cosmopolitanism are as old as the ancient Greek concept of the 'citizen of the world' (kosmopolitês) itself. Perhaps the most familiar criticism of all is that the cosmopolitan is someone who is loyal to nowhere and feels no obligation to anyone in particular. In the twentieth century, the derisive notion of the 'rootless cosmopolitan' was conjured by the Soviet state with precisely this intention of labeling certain individuals as disloyal to the collective and singling them (Jewish intellectuals in particular) out for repression. In response to the presumption of disloyalty, some recent commentators have argued that cosmopolitanism can coexist with other scales of attachment including to kith and kin, region, city or nation, and need not be merely be an escape from these. Recent studies have also shown how cosmopolitan perspectives are available in diverse contexts and among many kinds of people. We return to these debates later in the article.

Cosmopolitanism and Collective Change

Though the word cosmopolitan still shows its origins in Greek philosophy and culture, cosmopolitan ideas have presented themselves in multiple civilizations and kinds of social organization. Late nineteenth-century sociologist Emile Durkheim theorized that cosmopolitan awareness arises from the aggregation of smaller and simpler social groups into larger more complex entities. In this view, cosmopolitanism is principally a religious sentiment that expresses and reinforces processes of social incorporation. There is a strong affinity between epochs of geo-political expansion and cosmopolitan discourse. The

extension of Chinese territorial control during the first millennia BC, for example, led to philosophical questions concerning the shared human values of *tianxia* or 'all under heaven' and to notions of an ideal cosmopolitan citizen. Another well-recognized case is the emergence of Christian cosmopolitanism during the Roman empire: the image of all Christians as alike in the face of God coincided closely with a model of shared citizenship in a large and diverse imperial ecumene. This image of humans as 'all one in Christ' would return when Bartolomé de las Casas' defended Amerindians from Spanish enslavement during the sixteenth century, and later also when black slaves in the Americas fought for their emancipation.

Cosmopolitanism as the transcendence of local values, in sense (1) above, is found in all the so-named world religions including Buddhism, Judaism, Christianity, and Islam. In turn, all of these worldviews grew through large-scale processes of geo-political interconnection and human mobility. In contrast, the commitment within world religions to cosmopolitanism as openness and/or tolerance to other modes of life, as in sense (2) has been significantly, and perhaps inevitably, more variable. This, in turn, raises a paradox for a cosmopolitan perspective: a universalist cosmopolitanism whether religious or secular (e.g., certain forms of secular liberalism) may insist on doctrines that eliminate the hospitality toward heterodoxy or difference which many would see as the greatest value of cosmopolitanism.

The most recent phase of engagement with cosmopolitan ideas can be understood at least partly in a Durkheimian or collective sense. The late 1980s and early 1990s were marked, among other events, by the beginnings of a new moment, and means, for exchanging knowledge globally. Through the linking of computers into a worldwide web, and the proliferation of mobile phone technology, millions of geographically dispersed people acquired, notionally at least, a new kind of social proximity. Economic and geo-political changes followed, each further shrinking what were in the past significant social and psychological distances. A period of cultural and social questioning began, debates arose about alternative kinds of polity and new modes of citizenship. In turn, unforeseen types of collective social action and institutional control have become visible, as have new ideologies or, in Durkheim's words, revised forms of 'collective consciousness.' The unpredictable pace of economic, political, and also ecological change at the beginning of the twenty-first century has meant that even wealthy polities are less certain of their capacities for governance at the national scale. This, in turn, has led to new quandaries concerning global human development.

Kant's Cosmopolitan Anthropology

From Comte in the nineteenth century onward, French social thought has argued that social relationships can be studied as having a facticity on par with physical facts. Individuals were much less an intrinsic object of inquiry in this school than social relations. Consistent with this approach, Durkheim, the greatest exemplar of the tradition, viewed cosmopolitanism as essentially a product of transformations in collective social structure. In contrast, German anthropology and sociology, after the work of Immanuel Kant in the 1790s, tended to focus on the individual in order to understand society, including cosmopolitan ideals. Kant initiated this tradition with a sequence of writings on anthropology and cosmopolitanism whose starting point was the potential for free-acting individuals to bring into being a cosmopolitan or world community. The cosmopolitanism of classical Greek and Roman philosophy had also centered on individual moral aims and in some cases likewise presented a historical narrative of widening cosmopolitan exchange between peoples; however, Kant introduced radical innovations into these earlier views.

Kant focused, in particular, on how individuals are architects of their own *Weltkenntnis*, the 'world knowledge' that a person develops and adopts as they come to know the world through time in their interactions with others. The cosmopolitan morality of the Stoic philosophers, despite the wider context of Mediterranean trade and conquest, and the growth of urban centers, was grounded in a set of largely fixed and stable virtues. In contrast, Kant's world, that of the late eighteenth century, was already global in shape and extent and witnessing rapid social transformation including the beginnings of industrialization. Kant drew on information supplied by recent voyages of discovery, mercantilism, and colonization to build his view. Controversy surrounds scattered racist and misogynistic opinions that Kant reproduces at various moments in his writing: it remains an open debate how far these statements undermine his cosmopolitan philosophy taken systematically. Certainly, cosmopolitanism should be thought of, in his words, as a 'regulative idea' rather than an established set of virtues. The ability to adopt a cosmopolitan view is indicative of the potential progress of humanity toward peaceful international coexistence. He argued that history is driven by contradictory tendencies. Humans seek their own advantage egoistically and often violently, but they are also socially dependent on others for their own welfare. A cosmopolitan anthropology must comprehend this 'unsocial sociability' of the human individual, through a pluralistic vision of the human condition.

For Kant, a cosmopolitan view describes an individual's attempt to understand, inclusively and hospitably, the diversity of human standpoints; giving full weight to the differences involved while still judging and acting responsibly. Unlike many contemporary Enlightenment philosophers, Kant accepted that human nature was substantially plastic and changeable. Humans can adapt to, and transform, their own environments and in doing so they change themselves: there are potentially many possible rules for interacting and hence varied ways of evaluating human hopes, intentions, and knowledge. Unlike his contemporary Herder, though, Kant did not accept that the best measure of a civilization was the values of that civilization itself. Human beings are capable of thinking beyond the constraint of local circumstances and worldviews by referencing what they have in common with others, their humanity. Cosmopolitanism is 'regulative' because it encourages the individual to open their thinking to the heterodoxy of human perspectives in the process of extending their comprehension of human co-responsibility. Bakhtin's concept of *heteroglossia* (the coexistence of multiple speakers without a single authoritative voice) and Foucault's appreciation of *heterochrony* ('diverse times') and *heterotopia* ('diverse places') also draw on Kant.

One aspect, then, of Kant's cosmopolitanism can be understood as purely formal, and in essence algorithmic. Any person is in a position to adopt an enlarged view of their own life, imagining it as one instance of a common humanity. There are potentially as many such views of how the world is and might be as there are human beings. However, in taking on this cosmopolitan stance as a responsibility, the individual acknowledges their own moral answerability vis-à-vis a heterodox worldwide community. An imaginative re-orientation of this kind can potentiate actual cosmopolitan obligations, rights, and freedoms, he argues. Kant was fully aware that humans welcome the standpoints and principles of others only to a limited extent. However, they also rely on each other. As the interactivity of people around the world grows, so does the urgency for humans to think in global or cosmopolitan terms, he argued. He considered that the world he lived in had already formed an interdependent community; albeit an incipient, often hostile and violent one.

Cosmopolitanism versus Nationalism 1918–89

In the nineteenth century, neo-Kantian social philosophy flourished. As the century ended, two prominent German social theorists, Simmel and Weber, developed aspects of Kant's work with the aim of creating a new discipline, sociology. Weber aimed at a systematic view of the social which, while taking the rational individual as an abstract point of orientation, shifted the analytical center of gravity toward the study of generalized ideological constructs or 'ideal types.' In the process, he moved away from the normative and cosmopolitan aspects of Kant's work toward a value-free comparative study of civilizational *Weltanschauungen*, that is, characteristic and grounding principles by which an individual or group interacts with their surroundings. Weber's stance reflects the growing organizational power of the state in shaping social values during his era. However, in contrast, the social philosophy of Simmel is much closer to Kant's anthropological style and his aim of exploring the emerging *Weltkenntnis* of the free-acting human being. Simmel was preeminently a theorist of urban life and he saw the city, modernity, and elements of a cosmopolitan outlook as broadly coterminous.

For Simmel, urban social life welcomes individual-centered forms of self-expression alongside notions of personal responsibility that transgress group loyalties. In other words, for him modern experience breaks down traditional virtues while encouraging the kind of cosmopolitan transvaluation that Kant emphasized. Simmel notes, however, that detachment from local concerns and attraction to things further afield often leads, not to widened moral commitment, but to an ethically blasé attitude. In a series of interconnected works, Simmel argues that the characteristic qualities of the city and modernity, an emphasis on what is transitory and fashionable, the salience of personal networks, as well as the ubiquitous use of money as a social solvent arise from a universal subjective search for freedom vis-à-vis objective conditions. In turn, these aspects reflect a fundamental human openness to new modes of experience alongside a desire for self-differentiation. The results, as Kant had indicated, include a cosmopolitan awareness of complex forms of social interdependency that extend well beyond immediate relationships.

What is striking, however, is how cosmopolitan debates receded from view in twentieth century sociology. A dominant stance in the social sciences between 1918 and 1989 was that societies or cultures were causal entities and the primary example of this was the relationship between the nation-state and the individual subject. This view responded to the very great significance of nationalism in shaping the history of the period and to the idea that the state was the primary vehicle for individual and social development. Durkheim's emphasis on the power of collective consciousness was pervasive: much less attention was paid to his ideas of social, including cosmopolitan, change. While systematic features of world history including center–periphery relations, migration, and what came to be called globalization were debated, cosmopolitan philosophies and identities were rarely discussed. Some of the most influential neo-Kantian social philosophers of this period, Habermas and Rawls, for example, began to address cosmopolitan concerns only in the closing years of the twentieth century. Even so, many supranational institutions came into being in the mid-century, including the United Nations and the International Court of Justice at the Hague. Likewise, politics took an international turn in the 1960s with the proliferation of Marxist-inspired internationalist liberation movements. Nonetheless it was taken for granted that national citizenship and national identity were the common destiny for most of humanity.

During the citizenship and national identity were the common century at least one social science, sociocultural anthropology, took openness to cultural difference to be a primary method. In this loose sense, anthropology has always been epistemologically cosmopolitan since its primary task is to comprehend the global diversity of social institutions, practices, and worldviews. As with other social science disciplines, though, there was a marked tendency in sociocultural anthropology in the period between 1918 and 1989 to describe culture as geographically localized, internally systematic, and relatively circumscribed. On the other side of the equation, during the same timeframe, certain anthropologists such as Firth also referenced a general human tendency to outstep cultural boundaries and categories. In perhaps the archetypal work of modern anthropology, Argonauts of the Western Pacific, Bronlislaw Malinowski shows the Trobriand Islanders as a classically functioning holistic society, but he also describes their cosmopolitan quests using outrigger canoes to exchange with other groups and individuals overseas.

Since at least the 1950s there have existed social scientific approaches which explore the sociocultural from subjective and open-ended points of view. Transactionalists and interactionists, for example, have understood culture, not as a closed entity, but as a situation of subjective and intersubjective assessment and action. Life histories have often highlighted the heterodoxy of individual experience when compared to the homogeneity of collective worldviews. However, theorization of these and other phenomena gathered pace during the mid to late 1980s as part of postmodern thought-trends. New words such as 'hybridity' and 'creolization' became widespread expressing a sense of the cultural overlap, mixture, and the plurality of contemporary experience. Multiculturalist and

postcolonialist debates also fed into these discussions. Indexings of cosmopolitan phenomena at this time, by social scientists such as Ulf Hannerz, show an invigorated alertness to global flows of people and things, in opposition to styles of analysis that emphasized closed and sedentary aspects of culture.

In retrospect, elements of a genealogy are recognizable from Kant via Simmel and the interactionist views, through postmodern trends, to the newer subject-centered cosmopolitan approaches that have recently reappeared. In the first decades of the twenty-first century, further interest has again been paid to Simmel due to his early and theoretically innovative attention to processes and constraints of network formation. This is significant since there is a close connection between current interest in cosmopolitanism and awareness of new kinds of social network, as mentioned above.

The Formalization of Recent Cosmopolitan Arguments

One of the theorists who has done most to formalize cosmopolitan themes in current social theory is Ulrich Beck. He describes what he sees as a crisis in 'methodological nationalism' which threatens latter-day social inquiry with incoherence. He argues, alongside others, that twentieth century sociology was characterized by a refiguring of social problematics in national terms. However, many of the stock objects of inquiry in twentieth century sociology including the family, the household, and social class have lost their previously taken-for-granted meanings under current conditions. These terms are still used to indicate or contest values that the nation-state is supposed to ensure, but the collective reality to which they refer has changed beyond recognition. For Beck, the symptom of these changes is that in each person's life there exists a conflict of cultures and rationalities which cannot be resolved by a simple resort to a national frame of reference. In effect, Beck proposes, social knowledge is undergoing a renewed phase of cosmopolitanization in which people become aware of interdependencies linking themselves to others beyond national and other boundaries.

Beck's arguments closely conform to, while elaborating, Kant's. He proposes that personal destinies are now globally interlinked to an unprecedented degree. Belonging to a particular state no longer brings with it the feelings of predictability and security that it might have promised in the past. Awareness of global levels of interconnectedness, beyond national citizenship, intensify the individual and collective sense of risk. At the same time this can also lead to a recognition of new kinds of shared responsibility extending beyond and sometimes against the demands of national community. From this perspective, there is a need to reimagine the social sciences through a shift from 'methodological nationalism' to 'methodological cosmopolitanism.' In turn, this move has epistemological and normative consequences. The claims made for a value-free sociology were underwritten by an implicit nationalist morality and metaphysics, Beck argues. Social science should instead now adopt an ethic of cosmopolitan answerability in its inquiries.

Beck's cosmopolitanism makes claims according to all four of the core meanings of cosmopolitanism mentioned at the beginning of this article. He argues (1) for cosmopolitanism as a viewpoint that transcends other kinds of attachments. (2) He emphasizes openness and tolerance to sociocultural difference, though only so far as this does not override a recognition of shared humanity. (3) He indicates general historical trends toward world community. (4) His position is also morally normative in as much as he encourages social science to take an affirmatively cosmopolitan stand. It should be noted that Beck's approach is more individualistic and Kantian than it is collective and Durkheimian. Likewise, Beck is notable for the consistency with which he applies a universalist Kantian cosmopolitanism to the current situation in the social sciences. Criticisms have been leveled at his view at all four of its levels of exposition. It is worth considering these criticisms, and those of latter-day cosmopolitanism generally, both for what they tell us about the assumptions of cosmopolitan theorists and also those of the critics themselves.

Debates around the New Cosmopolitanism

The 1980s and 1990s saw the Soviet Union disintegrate, the end of the apartheid system in South Africa, the break up of states created in 1918 such as Yugoslavia and Czechoslovakia, the formation of the European Union, an unprecedented availability of cheap air travel, and, perhaps most significantly of all, the global dispersal of radically new technologies for communication. We have seen that these transformations were taken as a cue for revised types of cultural inquiry focused on travel, shifting personal identities, and intensified cultural interconnectedness. An initial counterargument to these intellectual foci involved pointing to the fact that, despite high levels of human mobility, most of the world's population remains relatively sedentary: likewise the majority still do not have even approximately equal access to the vaunted new technologies. Likewise processes of globalization do not necessarily lead to a shared vision of, or inclination toward, world community or even to a sense of social interconnection: in many cases, the tendency is contrarily toward fundamentalist localism.

A further critical view holds that recent discussions of cosmopolitanism are in effect benign descriptions of capitalist exploitation. In this counterperspective, the most salient feature of the 1980s, 1990s, and 2000s has been how neoliberal policies were deployed to disassemble social safeguards originally put in place by nation-states after the Second World War. Where supranational institutions such as the World Bank and the IMF intervened in this period it was not to create new rights of citizenship for all, but rather to urge the privatization of state-controlled institutions, especially programs of social welfare, and to insist on unrestricted corporate access to local communities and their markets. Certain individuals and groups benefited from these processes acquiring, in effect, cosmopolitan privileges within the new international political and economic disposition. In sum, the accusation is that world citizenship, and the vista on world affairs it offers, remains the preserve of a wealthy elite and serves its interests: to project cosmopolitanism beyond that grouping is to hide the real damage to local social relationships resulting from the destructive interventions of corporate capitalists.

It has been estimated that for the first time in history, from 2008, more people worldwide live in urban rather than rural settings. We have seen that, in at least some views, what is significant about cities is how they expose city-dwellers to problems of cultural difference and social distance; it has been argued that urbanity corresponds closely to the individualization and cosmopolitanization of experience. From this perspective, a general cosmopolitan trajectory in history seems irreversible. Simmel noted how, in isolated kin-based communities, strangers represent an intrusion of difference into a relatively homogeneous way of life; they inspire combined investments of fear and attraction. In a more intensified way, within the modern city, all urbanites become strangers at varying levels and degrees: all must learn to assess and incorporate differences, balancing this with their desire for social and personal completeness. If only because of the pressure of population density, urbanites must agree to cooperate with culturally distant others; people who their forebears might not even have considered to be fully human. The results of these processes have been new types of commensality and, arguably, historically new cosmopolitan ideals and forms of personhood.

While they might agree that urbanization may lead to some of the cosmopolitan situations and outcomes described by Simmel and others, critics argue against the inevitability, let alone the desirability, of the tendencies involved. A counter-perspective points to the fact that urbanization creates new types of ghetto and gated community, in other words a plethora of no-go-areas and new ways of excluding strangers. Indeed, people may purposefully choose to live in closed social niches as a means of ontological self-preservation. Cities, like human environments in general, are subject to constant change and encompass radically distinct lifeways, status hierarchies, and economic disparities. New communicative technologies, rather than catalyzing cosmopolitan interchange and the appreciation of heterodoxy, can isolate people further within imagined social boundaries. Whether a specific setting evidences growing cosmopolitanization, increasing ghettoization, both or neither, is to a great degree a matter of how questions are formulated and the point at which the phenomena at issue are transected empirically. Here we enter complex territories of translation, interpretation, and evaluation and further thoughtful investigation is needed using new ethnographic, archival, and experimental methods.

The attempt to explore and comprehend the highly diverse worldviews that currently exist is, arguably, an intrinsically cosmopolitan task. At the same time, a paradox is initiated: the more extended the search, the greater the likelihood of encountering practices and ontologies so divergent that expectations concerning further mutual comprehension must be indefinitely deferred. An example is the contrast between the humanist ideals prized in the West versus the animism still valued in many communities worldwide. Within a humanist view, human individuality has moral priority over relationships to other entities. Contrastingly, in an animist perspective, animals, humans, and the material world are understood as spiritually coimplicated; humanness does not have a distinct priority. One kind of answer from Kantian philosophy is that, if nothing else, individual human awareness forms a point of common reference out of which shareable models of responsibility and translatability can grow. However, some critics argue that this attempt to ground cosmopolitan expectations in universal features of human selfhood reproduces a fundamental misunderstanding.

The French tradition of sociology rooted in the work of Comte, Durkheim, and Tarde has provided a number of arguments against Kant's proposal that cosmopolitanism is grounded in universal capacities of human individuality or subjectivity. Recently, following this tradition, Latour has reasserted that there is no meaningful human individuality outside the historically specific networks of agents and things through which subjectivity takes shape. He also argues, though, that there is no bounded society in the holistic causative sense that Durkheim proposed. Cosmopolitanism depends on being able to take a universal or synoptic view on human affairs. However, Latour proposes, this synoptic vision is achievable neither individually nor collectively: it is an illusion. There are, nonetheless, conflicts between different cosmologies, animist versus humanist, for example; these emerge in distinct situations of network formation. So, even though there can be no meaningful cosmopolitanism in either Kant's or Durkheim's sense, there is necessarily cosmopolitics, that is, conflict and negotiation concerning how the cosmos is made up.

Most of the criticisms and debates indicated up to this point turn on the relationship between how the cosmopolitan is imagined and the kinds of empirical phenomena which support these imaginings. Bruno Latour has argued that 'cosmopolitics' is a kind of war between distinct interest groups about how the cosmos is constituted. Latour's thus shifts the question to metaphysical claims and counterclaims. Certainly, one response to Latour's argument might be that when he adverts to the many warring cosmologies in existence he is claiming some type of synoptic, overarching, or cosmopolitan awareness of this manifold. Likewise, if we take from Latour description of cosmopolitics a responsibility to acknowledge the diversity of worldviews, then, again, this is consonant with the principle of cosmopolitan openness and tolerance. His view also fits with cosmopolitanism understood as a theory of, or quest for, change toward peaceful coexistence. This brief delineation of topical arguments and counterarguments by no means exhausts the questions raised, though.

Conclusion: The Applications of Cosmopolitanism

The core ideas and looser trains of thought that give rise to cosmopolitan debate sometimes conflict with each other and sometimes produce paradoxical results. As Derrida among others has pointed out, the urge to be hospitable, a core facet of cosmopolitanism, involves a wish, however generous, to demonstrate one's power over another. In absolute terms, hospitality to others and toward their viewpoints is necessarily double edged and paradoxical. When openness to the other reaches a certain pitch the roles are reversed: the host is now the other's guest and, in that sense, in their thrall. Hence, engagement with cosmopolitanism foregrounds a concern regarding who is host and who guest in today's multipolar, and in certain respects postnational, world. Cosmopolitan fulfillment, as this is subjectively imagined, likewise seems inevitably to run

against the actual global disposition of political and economic power.

Cosmopolitanism can present itself as a moral predicament and it can become an object of inquiry, or it can be both. As a focus of social scientific investigation the tasks involved are extremely varied. Those who currently study cosmopolitanism themselves belong to an increasingly interconnected global public sphere; a network, and in principle a community, of people largely proclaiming secular liberal goals derived from the thinkers of the enlightenment and its aftermath. These features may present both an advantage and a disadvantage. In principle, the demand of a cosmopolitan view is that it should show the fullest awareness of human differences possible, while demonstrating the widest comprehension of human commonality. Theoretically it should be possible to start from any standpoint or global subject position and explore that stance in cosmopolitan terms. The question now, as in the past, is whether there can be a cosmopolitanism that is both intellectually coherent and genuinely open, accommodating, and tolerant toward others.

As a field of inquiry, cosmopolitan studies increasingly recognize that cosmopolitanism is as much a description of change as a universal set of principles. Kant centered his anthropological cosmopolitanism on questions that the self asks itself: 'What can I know?' 'What should I do?' 'What can I hope for?'; uncertainties that alter the individual's relationship to others and to the world at large through time. One way of thinking about cosmopolitanism as a process is to consider how the *Weltkenntnis*, or world knowledge, of the individual or the group can outgrow the worldview, or *weltanshauung*, that nourishes it. While human lives find a center in practical habits, shared values, and reciprocities, a periphery or horizon also continuously unfolds where new ideas and extended networks of relationship come into view affecting the future of self and community as these have been imagined up to now. Inclusion of what was once at the periphery of experience and knowledge provokes a recentering of activity and a widening of scope on the world. This process need not be sparked by any heroic ethical motivation, it may instead come about as a result of curiosity and desire, dissatisfaction and listlessness, or an intensified sense of fear, risk, and uncertainty. At least in one basic formulation, then, cosmopolitanism describes a dynamic of change in the organization of groups and the lives of individuals. However, since methods and concepts for understanding cosmopolitan motivations, perceptions, and institutions are still only sketchily developed, there remains much to be done.

Bibliography

Beck, U., 2006. The Cosmopolitan Vision. Polity, Cambridge.

Delanty, G. (Ed.), 2012. Routledge Handbook of Cosmopolitan Studies. Routledge, London.

Durkheim, E., 2001. The Elementary Forms of Religious Life. Oxford University Press, Oxford.

Firth, R., 1951. Elements of Social Organization. Watts & Co, London.

Friedman, J., 2002. From roots to routes: tropes for trippers. Anthropological Theory 2 (1), 21–36.

Hannerz, U., 1989. Notes on the global ecumene. Public Culture 1 (2), 66–75.

Hannerz, U., 2006. Cosmopolitanism. In: Nugent, D., Vincent, J. (Eds.), Companion to the Anthropology of Politics. Blackwell, Oxford.

Kant, 1983. Perpetual Peace and Other Essays. Hackett Publishing Co., Indianapolis.

Kant, 2006. Anthropology from a Pragmatic Point of View. Cambridge University Press, Cambridge.

Latour, B., 2004. Whose cosmos, which cosmopolitics? Comments on the peace terms of Ulrich Beck. Common Knowledge 10 (3), 450–462.

Rapport, N., 2012. Anyone: The Cosmopolitan Subject of Anthropology. Berghahn, Oxford.

Simmel, G., 1964. The Sociology of Georg Simmel. Free Press of Glencoe, Illinois.

Sturm, T. 'What did Kant mean by and Why did he adopt a cosmopolitan view in history?' OAC Working Paper #12. http://openanthcoop.net/press/2012/03/08/what-did-kant-mean-by-and-why-did-he-adopt-a-cosmopolitan-point-of-view-in-history/.

Wardle, H., 2010. A cosmopolitan anthropology? Introduction to a Special Edition of Social Anthropology 18 (4), 380–389.

Cost–Benefit Analysis

Anthony E Boardman, University of British Columbia, Vancouver, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.H. Haveman, D.L. Weimer, volume 4, pp. 2845–2851, © 2001, Elsevier Ltd.

Abstract

The purpose of cost–benefit analysis (CBA) is to make better-informed, more consistent public policy decisions. CBA is a method for assessing the economic efficiency of proposed policies or programs through the systematic prediction and valuation (i.e., monetization) of all costs and benefits to all members of society. The concepts of 'willingness to pay' and 'opportunity cost' guide the valuation of social benefits and social costs. Future benefits and costs are discounted at the social discount rate.

The purpose of cost–benefit analysis (CBA) is to make better-informed, more consistent policy decisions (*see* Policy Analysis: Evidence-Based Policy Making). CBA is a method for assessing the economic efficiency of proposed public policies through the systematic prediction and valuation (i.e., monetization) of social costs and social benefits. While individual decisions focus on benefits and costs to the individual, CBA considers the benefits accruing to and the costs incurred by all members of society – hence the terms social benefits and social costs. (For expositional convenience, however, from here on we will simply use the terms benefits and costs, although we mean social benefits and social costs.)

The concept of 'willingness to pay' guides the valuation of policy benefits and the concept of 'opportunity cost' guides the valuation of policy costs. CBA values all policy effects on all members of society and aggregates them to yield the net benefits of the policy. A policy with positive net benefits is economically efficient relative to the status quo. When there are many policy alternatives, the decision rule is: choose the policy, or set of policies, that maximizes the net benefits.

The roots of CBA can be traced back to Sir William Petty who urged the evacuation of the residents of London to reduce deaths due to the great plague, 1665–66. CBA gained prominence in the 1930s as a technique for systematically measuring benefits and costs of flood control projects sponsored by the US federal government (for an overview, see Steiner, 1974). In the 1960s, it was used extensively in the UK's Department of Transportation (Colvin, 1985). As the scope of governments grew, and economists gained influence as policy analysts, the use of CBA expanded well beyond the evaluation of public works and other infrastructure projects (*see* Infrastructure Investment). It has also been used to evaluate changes in regulatory policies, environmental policies, health care, welfare-to-work, education, income transfer, privatization, and in many other policy areas (*see* Regulation, Empirical Analysis of; Welfare Economics). In the US, Executive Order 12866 (4 October 1993) requires federal regulatory agencies to assess both the costs and the benefits of proposed regulations.

CBA can be forward looking (ex ante) or backward looking (ex post); that is, it can be used to evaluate whether a proposed project will increase efficiency or to evaluate whether an implemented project did increase efficiency. Ex ante CBA is appropriate for policy decision-making purposes. Ex post CBAs can help guide and improve the quality of ex ante CBAs as well as provide information about the allocative efficiency of actual government policies.

Foundations of CBA

As with the neoclassical economic theory that guides market analysis, CBA rests on the analytical underpinnings of modern welfare economics. Modern welfare economics takes a utilitarian perspective and presumes that aggregating the well-being (utilities) of the individuals who make up society is a valid measure of social well-being. CBA provides a basis for inferring changes in utilities, which are unobservable, from economic choices and outcomes, which are generally observable. It expresses the value of changes in utilities in terms of dollars or some other monetary metric. Thus, CBA can be thought of as the accounting protocol of welfare economics for measuring the changes in utilities caused by changes in public policy.

The principle of Pareto efficiency lies at the heart of welfare economics (*see* Pareto, Vilfredo (1848–1923)). An allocation of goods is Pareto efficient if there exists no alternative allocation that makes at least someone better off without making anyone else worse off. The intuitive appeal of Pareto efficiency is obvious: one would have to be malevolent to oppose a reallocation that betters the situation of someone without harming anyone else.

Although the Pareto efficiency criterion seems unobjectionable as a basis for judging the desirability of public policies, it would be largely useless as the basis for government choice because every policy inevitably has a negative impact on someone. This limitation of Pareto efficiency prompted efforts to develop a criterion with more general applicability to actual policy decisions. More specifically, it led to the 'potential Pareto improvement' criterion, which is also called the 'Kaldor–Hicks criterion' after Nicholas Kaldor and John Hicks. A potential Pareto improvement exists if a change in the allocation of resources allows those who gain to fully compensate those who lose and still be better off. This criterion expands the set of situations and policies about which statements may be made as to whether economic efficiency has been increased, decreased, or left unchanged. If a potential Pareto improvement takes

place, then there is a gain in economic efficiency, also called 'allocative efficiency.'

In order for a potential Pareto improvement to occur, those who gain must gain more than those who lose. Put simply, the benefits must exceed the costs or, equivalently, the net benefits must be positive. CBA attempts to maximize allocative efficiency by choosing the project with the largest net benefits.

Note that it is not necessary for those who gain to actually compensate those who lose. If those who lose were compensated so that they were no worse off than before and the gainers were still better off, then the reallocation would be an actual Pareto improvement. In addition, note that any Pareto improving reallocation would also be a potentially Pareto improving reallocation.

The amount that someone is willing to pay to obtain a policy's impacts depends on his or her wealth. The potential Pareto improvement criterion takes the current distribution of wealth as given. Although improved environmental quality may convey the same physical effect on all people, more wealthy individuals would probably be willing to pay more for this benefit than those who were poorer. The aggregate willingness of a society to pay for a policy impact depends on the level and distribution of wealth among its members. Thus, the benefits used in CBA are contingent upon the particular distribution of wealth that exists in the society. Furthermore, using the potential Pareto criterion as a decision-making rule implies that the current distribution of wealth is legitimate.

Implicitly, CBA assumes that the marginal utility of income or wealth is equal across all members of society. That is, a poor person derives the same utility (benefit) from an additional dollar as does a poor person. A project that transfers a fixed amount of money from one person to another or from one group of people to another group is, from the perspective of CBA, simply a transfer and has a net benefit of zero (ignoring transaction costs).

In summary, CBA concerns allocative efficiency, not equity. According to CBA, a project should be adopted if it increases allocative efficiency, whether equity increases or decreases. Put succinctly, CBA focuses on maximizing the size of the 'economic pie,' not on how the pie is distributed. Later, we discuss distributional CBA and 'passive use' benefits that incorporate some aspects of equity into CBA.

Alternative Projects and the Counterfactual

In ex post CBA, an implemented project is compared to a counterfactual, which reflects what would have happened if the implemented project had not been implemented. If there would have been no change in policy, then the most appropriate counterfactual would be the status quo. Sometimes the status quo would not have been viable and the implemented project would have displaced some other specific alternative. In this case the most appropriate counterfactual would be the displaced specific alternative. In either situation, however, the analyst compares the net benefits with the project to the net benefits without the project. Alternatively, the analyst can calculate the incremental benefits of the project relative to the counterfactual and can calculate the incremental costs of the project relative to the counterfactual. CBA would conclude that the implemented project increased allocative efficiency if the incremental net benefits were positive.

When conducting ex ante CBA there are often many alternative ways of achieving the same broad objective. One of the first tasks in CBA is to describe the various policy alternatives. If there is only one proposed alternative project, the analyst must determine if the net social benefit would be greater with the proposed project than without the project or if the incremental net benefits of the project would be positive.

The Issue of Standing

Costs and benefits in CBA derive from the preferences of the individuals who make up society. But what defines society? And are all preferences legitimate? Borrowing on legal terminology, these questions have come to be known as issues of 'standing' (Whittington and MacRae, 1986).

Recognizing that the citizens of a country share a common constitution, which sets out values and rules for making collective choices, and they participate in the same national economy, most analysts define society at the national level. Thus, the CBA analyst would usually consider all of the benefits received and all of the costs incurred by all of the residents in the country or all of the citizens of the country. However, subnational level governments, such as States in the US or Provinces in Canada, often fund projects that primarily affect residents in their region but might also affect people living in other regions. A subnational government might require the analysis to be conducted at its subnational level. If so, the benefits received and costs incurred by residents of other regions would not be included in the analysis because they would not have standing. An analysis from a regional government perspective might reasonable consider a grant received from the national government as a benefit. However, if the analysis were conducted at the national level, the grant would be treated as a transfer because it involves an offsetting national cost.

Analyses of some environmental policies, such as the reduction of greenhouse gases, and of some health policies, such as the reduction or eradication of polio, are best conducted at the global level because they are global issues. In this case, the analyst should consider all of the benefits and all of the costs on everyone.

The above discussion implies that only people have standing. Plants, fish, and animals do not. Consequently, the cost of a policy that leads to the extinction of a species might reflect how much people are willing to pay to avoid such a loss, but any suffering experienced by members of the species itself would not be included.

Most would agree that an accounting that accepted the preferences of all people as legitimate would be too broad. For example, most observers would not view the pleasure that psychopaths get from physically abusing people as an appropriate part of the utilitarian calculus. However, drawing the line between legitimate and illegitimate preferences is sometimes difficult. One approach is to rely on the existing legal framework for guidance (Zerbe, 1998), yet doing so assumes that existing laws are moral.

Identification and Categorization of Impacts

Analysts need to identify all of the impacts of a project that accrue to those with standing. Most impacts change the observable behaviors of individuals and are called 'active use' impacts. For example, preserving a wilderness area would facilitate bird watching, an observable behavior. Some government policies also produce impacts that benefit people without changing their observable behaviors. Such behavior is described as 'passive use.'

Active Use Impacts

Consider a government project that will provide irrigation to avocado producers. The major inputs would be the resources for the project – design, excavation, construction, land, and transaction costs. The major outputs would be more avocados produced and a lower price of avocados. There might also be some environmental costs or benefits. Analysts should identify each observable impact and classify it as a cost or benefit. Naturally most inputs are costs. Some outputs are a benefit to some people, but a cost to others. For example, draining wetlands to provide housing might be a benefit to people who want housing but a cost to environmentalists who previously enjoyed the presence of ducks or other wildlife.

Comprehensively taking account of the impacts of a policy means that CBA differs from financial analysis. Considering only the financial revenues and costs to government agencies is generally inadequate for CBA because it does not take account of all impacts on citizens. The decision about where to locate some new service-providing branch of a government agency is often based on financial costs. However, some locations are much more accessible by users than others. A CBA would also consider users' transportation costs and their time spent getting to the facility. More accessible locations would have more users and would therefore provide greater benefits to users.

Analysts should be careful not to 'double count' benefits or costs. A new subway station benefits commuters who live nearby because it reduces their travel time. In addition, it will lead to higher property values near the station. Treating both the reduction in travel time and the increase in property values as benefits would be double counting because property values have increased due to the reduced travel time. One impact is a consequence of the other; it is a 'knock on' effect. Another reason why property price increases should not be included is that it is a distributional impact and does not have an aggregate net benefit. Prices increase near the subway station because people want to move there. But these people are moving from somewhere else and prices in those areas will decrease or will not increase as fast as they would have done without the new subway station. In aggregate, there is no net change in the value of property.

The 'knock on' increase in property prices can be thought of as a 'secondary market' effect of a policy (Haveman and Weisbrod, 1975). The benefits of a project can generally be computed from its direct effects in the 'primary market.' For example, the primary benefit of stocking a lake with game fish is to increase the supply of fishing days. But, the improved fishing might lead to the construction of a nearby hotel or the employment of more workers and so on. It is tempting for some to suppose that these knock on effects create additional benefits. However, when the secondary markets are competitive, these effects should be ignored. The reason is the same as the one given above: it is incorrect to attribute any benefits to the secondary effects because they involve offsetting costs. It is especially tempting for politicians to suppose that knock on effects create additional benefits. Indeed politicians are tempted to use multipliers, that is, to multiply the primary market benefits by some usually arbitrary number. While multipliers are often used in economic impact analysis, they should never be used in CBA.

Secondary market effects should be included in CBA where those markets are not perfectly competitive or where the government policy affects the market price. For example, buying many tons of concrete for a dam in a local market may drive up the price of concrete, as the marginal costs of producing concrete increase. However, most of the additional payments, called 'rents,' are transfers from the project budget (government) to existing suppliers and therefore do not affect the net benefits. Furthermore, the effect may be transitory: if the market is competitive, new entrants will lower the market price.

Passive Use Impacts

Analysts have found it useful to identify three distinct categories of passive use. First, a policy may have 'option value' (Weisbrod, 1964). People may value a good because they anticipate the possibility of using it in the future. Only when they actually use it, if ever, will their behavior be observable. For example, a person may be willing to pay something now to preserve a national park in Alaska with the hope of visiting it someday. The person's circumstances, however, may never allow the visit to be made.

Second, a policy may have 'existence value' (Krutilla, 1967). People may value a good even if they anticipate they will never use it. For example, a person may derive benefit from the preservation of a wildlife habitat that would be completely closed to the public. One motivation may be a belief in the intrinsic value of the natural order. Another motivation may be altruism: a desire to make the good available either to the current generation, or to bequeath it to future generations.

Third, a policy may have a 'donor value' (Hochman and Rodgers, 1969). People may derive benefits from a policy that distributes goods in a way they view as desirable. The motivation may be altruistic, say, in providing more to those who otherwise would be consuming less. It may be motivated by values of equality or fairness. The inclusion of donor values integrates individual preferences over the distribution of wealth into the efficiency framework, broadening it from a normative perspective.

Option value is best thought of as an adjustment to net benefits or expected net benefits, as discussed later, although it is often useful to show it as a separate category. In contrast, existence value and donor value can be thought of as additional categories of benefits.

Prediction and Valuation of Policies in Well-Functioning Markets

To begin, consider a 'well-functioning' market where the demand and supply curves are known or could be estimated. The market may be perfectly competitive or there may be some government intervention in an otherwise perfectly competitive market. Also suppose that the proposed policy interventions will affect the market outcome in a predictable way. For example, it might pertain to the introduction of subsidies in a competitive market or policies that regulate a monopoly in different ways. In these situations, the incremental net benefits of the policy can be calculated relatively easily using the tools of welfare economics.

Measuring Benefits

The fundamental concept for guiding the measurement of benefits is the willingness to pay of those affected. The analyst wishes to know the maximum amounts individuals would be willing to pay for policy impacts that they view as beneficial, and the minimum amounts individuals would be willing to accept as compensation for policy impacts that they view as harmful. In other words, the analyst would like to know how much each individual's wealth would have to be adjusted in conjunction with the policy change so that the individual's utility remained unchanged. The algebraic sum of these wealth adjustments for all individuals measures the social value of the benefits of the policy in terms of dollars. Although some theoretical caveats apply (Blackorby and Donaldson, 1985), policies with larger positive sums have larger benefits.

A person's 'Hicksian' or 'compensated variation demand curve' for a good tells how much of a good the person wishes to purchase as a function of the price of the good, holding all other prices and the person's utility constant. In practice, however, analysts usually work with the Marshallian demand curve that holds income (or wealth) constant, not utility. Fortunately, these demand curves are often sufficiently similar so that analysis of the 'Marshallian demand curve' provides a good approximation of individual willingness-to-pay amounts (Willig, 1976).

An individual's 'inverse demand curve' reflects, at each level of consumption, the person's willingness to pay for an additional unit of consumption – their marginal benefit. Adding up the marginal benefit for each of the units consumed gives the total benefit to the person from participation in the market or experiencing services produced by the public sector. (Graphically, the individuals' total benefit is the area under the individual's demand curve from zero units to the number of units actually consumed by the individual.) The (total) benefit to all members of society can be calculated by summing the individual benefits. In effect, it can be calculated as the area under the market demand curve from zero units to the number of units consumed by everyone in society, where the market demand curve is the horizontal sum of the individual demand curves.

If a policy either satisfies a demand that has not been met, or changes the price of a good or service that people are already consuming, the benefits to society will be changed. The change in society's aggregate willingness to pay for the impact of the policy measure provides a measure of the incremental benefits of the policy.

Empirical evidence shows that people's willingness to pay for a gain is generally much less than their willingness to accept in compensation for a loss, even if the gains or losses are small (*see* Behavioral Economics). Furthermore, people may view positive changes as reductions in losses rather than as gains. For example, they may view policies that control air pollution as an attempt to not make pollution worse rather than as improving pollution. In such circumstances, willingness-to-pay estimates of a policy to control pollution are likely to be smaller than what people would be willing to accept to forgo the pollution control policy. The correct measure depends on people's reference point: whether it is the pollution level before or after the policy were implemented (Knetsch et al., 2012).

Measuring Costs

The implementation of policies usually requires the use of inputs that could be used to produce other goods or services. CBA uses the concept of opportunity cost to value these impacts. The opportunity cost of an input is the value of the goods and services that the input would have produced in its best alternative use.

In a perfectly competitive market, the market supply curve measures the marginal cost of the resources used in the production of the good or service. The area under the market supply curve from zero to the number of units produced equals the total variable cost of producing those units. The cost of a policy, assuming there is no 'externality,' equals the change in total variable costs plus any incremental fixed costs.

Most government projects are funded out of taxes. Even if a project is financed in the short run by an increase in the deficit, it will ultimately be paid for by higher taxes. Standard practice in CBA is to treat a dollar of accounting cost raised through taxation as having a cost of a dollar. In effect, this approach treats taxation as a transfer: taxpayers lose, the government gains and there is no impact on net benefits. However, raising taxes has a cost: it reduces allocative efficiency (i.e., it creates a 'deadweight loss') because it affects people's consumption, labor supply, and investment decisions. The amount by which the cost of a dollar of tax revenue exceeds its nominal value is called its 'marginal excess burden.' Expenditures funded by tax increases, and revenues allowing tax reductions, should be adjusted by the marginal excess burden of the relevant tax. Estimates of marginal excess burdens have been made for many different taxes in many different countries (Dahlby, 2008).

Computing the Net Benefits and Social Surplus

Given the incremental benefits of a policy or project and its incremental costs, the incremental net benefits can be obtained by subtraction. Alternatively, the incremental net benefits can be calculated using the concept of 'social surplus.' Most government policies affect consumers, producers, and government. The impact of a new policy on consumers can be measured by the change in 'consumer surplus,' the impact on producers can be measured by the change in 'producer surplus,' and the impact

on government can be measured by the change in 'government surplus.' Social surplus equals the sum of consumer surplus, producer surplus, and government surplus, where the latter should be adjusted by the marginal excess burden of taxation. The advantage of this approach is that it provides information about the distributional implications of a policy.

An individual's consumer surplus for a unit consumed is the difference between what the individual is willingness to pay for that unit and the amount he or she actually pays. Adding up the consumer surpluses for each of the units consumed gives the individual's (total) consumer surplus that results from participation in the market or experiencing services produced by the public sector. Graphically, the individual's consumer surplus is the area under the individual's demand curve but above the market price from zero units to the number of units actually purchased. Aggregate consumer surplus equals the sum of the individual consumer surpluses or, equivalently, it equals the area under the market demand curve but above the market price.

Analogously, the producer surplus equals the area above the market supply curve but below the price. If producers do not incur any incremental fixed costs as a result of a new policy, then the change in producer surplus equals the change in producers' profits. Government surplus equals net government revenue, which is usually negative because most projects are paid for by government.

Prediction and Valuation of Policy Impacts in Poorly Functioning Markets

The preceding section pertains to situations where the relevant market demand and supply curves are known. In practice, this condition is rarely met. Health and transportation policies, for example, affect morbidity and mortality (*see* Transportation, Supply, and Congestion). However, there is no well-functioning market for either of these 'goods' and consequently it is not easy to estimate the required supply or demand curves. The same problem pertains to 'environmental goods,' such as improvements in air quality. In such situations, analysts usually proceed in two steps: first, they predict the benefit and cost impacts and then they value (i.e., monetize) each impact (Boardman et al., 2011).

Prediction of Impacts

The analyst must either predict the impacts with the project and without the project (the counterfactual) or he or she must predict the incremental impacts of the project. For example, in order to estimate the costs and benefits of a new irrigation project, an analyst could predict the yield of crops and the environmental quality that would arise with the project and, for comparison, the yield of crops and the environmental quality that would arise without it. Alternatively, she may predict the change in the yield of crops and the change in environmental quality. Often, the latter approach is easier.

Whichever approach is adopted, analysts use a variety of methods to make predictions. The past is often the best guide. If an intervention saved so many lives when it was introduced in one country, analysts can use that information to predict how many lives it will save in another country. Of course, the prediction must take account of different circumstances, tastes, cultures, and behaviors. Relying on only a single previous study is risky because it may have been poorly conducted and it provides only one data point. More studies would enhance one's confidence. Predicting the impacts of new, unique projects with no previous direct evidence is especially difficult.

Valuation of Benefits Using Revealed Preference Methods

People may convey information about their willingness to pay for changes in some impacts by their action in other markets. Earlier we argued that changes in property values may reflect changes in commuting times. In other situations, changes in property values may reflect changes in the level of noise pollution or in the level of air pollution. Economists can use this information to derive a 'shadow price' for different types of pollution, that is, what the marginal cost would be in a well-functioning market.

People employed in jobs with high 'fatality risk' (i.e., a high chance of suffering a fatality) are generally paid more than people in jobs with low fatality risk. By examining the trade-off between higher salaries and higher fatality risk, economists can infer how much people are willing to pay for increased safety (i.e., for a reduction in fatality risk) (*see* Safety, Economics of; Value of Life, Economics of). In practice, salaries depend on many factors in addition to fatality risk. Analysts control for these and other factors, including selection bias, using the 'hedonic regression method.' The hedonic wage method is one of a number of revealed preference methods used to estimate the 'value of a statistical life' (Viscusi and Aldy, 2003).

Many impacts may require valuation in a CBA. The analyst is unlikely to have the time or other resources to conduct their own study to independently value each impact. Instead, analysts will seek out 'meta-analyses' of previous research or will use a 'plug in' value based on prior research or from some reputable source, a strategy called 'benefits transfer.' For example, the US Environmental Protection Agency (2013) recently estimated the cost of carbon (i.e., the benefit of reducing carbon dioxide emissions by 1 metric ton) at $46 in 2020, stated in 2011 dollars, and using a discount rate of 3%. This value can be used in numerous CBAs pertaining to health or the environment.

When existing shadow prices are used in a new context they must be adjusted to reflect the differences between that context and the context where the value was estimated. The value of a statistical life has a positive 'income elasticity': people are willing to pay more to save their life as their incomes rise. Consequently, people in rich countries are willing to pay more for safety improvements and for improvements in carbon dioxide reduction. Most estimates of the value of a statistical life are derived from the US, which has one of the richest populations in the world. These estimates should be modified for use in countries with lower income levels (Hammitt and Robinson, 2011). They should also be modified if relevant tastes and cultural values are different.

Valuation of Benefits Using Continent Valuation Methods

It would be ideal if analysts could elicit willingness-to-pay amounts directly from individuals through structured

conversations. If securing this information from all citizens were too costly, surveys of population samples could be used as the basis for estimating aggregate willingness to pay. This approach is called 'contingent valuation.' It can be used to value impacts that involve active use benefits. More importantly, it is often the only way to value passive use benefits. Contingent valuation is often used to value environmental impacts (Bateman and Willis, 2000) (*see* Environmental Economics). A blue ribbon panel convened by the US National Oceanic and Atmospheric Administration concluded that contingent valuation could be a reasonable basis for estimating passive use values in natural resource damage assessment (Arrow et al., 1993).

Nevertheless, the use of contingent valuation remains controversial. One source of concern is the cognitive demands that it places on respondents in terms of understanding the policy being valued. Unless respondents understand what is being valued, and believe their answers to be consequential in influencing the policy choice, they cannot be expected to give meaningful answers. Another source of concern is the fear that respondents will answer strategically in the sense of giving false answers that they believe will lead to a better outcome for themselves than would result from truthful answers. Both theory and evidence suggest that giving respondents referendum-type questions about public goods minimizes the dangers of strategic responses. That is, rather than directly asking a respondent to state her willingness to pay, she is given a random dollar amount and asked if she would vote for the policy if it would cost her that amount in higher taxes or increased costs for other goods. An estimate of mean or median willingness to pay for the sampled population can be inferred statistically from sample responses.

Valuation of Costs

The market price of an input provides a good measure of the opportunity cost of that input if it is purchased in a competitive market and the project does not alter the market price of the input. For example, the purchase cost of books bought for a local reading project would equal their opportunity cost. However, the opportunity cost of inputs purchased in noncompetitive markets may be larger, smaller, or equal to their market price. For example, the opportunity cost of the time of people compelled to serve on juries may be larger or smaller than the legally set payments they receive from the court.

A commonly encountered problem in CBA is the measurement of the opportunity cost of labor that would otherwise be unemployed. This cost depends on the value to society of whatever the workers would give up if they were employed. They might not consumer leisure, or they might not clean the house or look after children. The wage they receive is likely to overestimate their opportunity cost. Alternatively, zero would underestimate their opportunity cost. Rarely do analysts have enough information about the value of the forgone opportunities to estimate accurately the opportunity cost of unemployed labor so they usually use some fraction of the accounting cost as an approximation (Haveman and Krutilla, 1968).

Markets in developing countries are often more distorted than markets in developed countries. The labor markets are often segmented and labor mobility may be restricted. Official exchange rates may not reflect the true value of national currencies, and there may be both 'formal' and 'informal' sectors. In developing countries, analysts often compute shadow prices using the LMST method (Little and Mirrlees, 1974; Squire and van der Tak, 1975; Dinwiddy and Teal, 1996).

Sometimes the accounting cost of an input is zero. For example, a government project may not pay any rent because it occupies space owned by a branch of the government. However, the opportunity cost is not zero and may be very large.

Discounting Future Costs and Benefits

The benefits and costs generated by a public program often extend over long periods. Because people are impatient, they would rather consume now than in the future and, therefore, they value goods and services they consume in the future less than if they were consumed in the present. Also, money has a financial opportunity cost: people can invest now and earn a positive real return, which would allow them to consume more in the future. For both of these reasons, analysts discount future benefits and costs.

Markets for lending and borrowing funds (also known as capital markets) balance the preferences of lenders who require a payment for giving up use of their funds (hence, delaying consumption) and the preferences of borrowers who invest in projects with future payoffs. This balancing results in a market interest rate, denoted as r. In an ideal economy, this market interest rate would equal, for each and every individual, their 'marginal rate of substitution' between consumption this year and consumption next year, denoted as p. Since $r = p$, it would be the natural choice for the 'social discount rate' (SDR), that is, the rate to discount benefits and costs in CBA and other methods of government project evaluation. The 'present value' of some benefit received or cost incurred in year n, denoted X_n, would be given by $X_n / (1 + r)^n$. The present value of the net benefits of a policy equals the sum of these amounts over all impacts in all time periods. Maximizing the present value of the net benefits would maximize allocative efficiency.

In practice, however, capital markets are distorted by taxes on savings, externalities, transaction costs, credit rationing, and inconsistencies in individual behavior. When such distortions exist, the choice of the discount rate method becomes more complicated. In fact, it is one of the most controversial topics in CBA (Burgess and Zerbe, 2013; Moore et al., 2013b). There are two main approaches.

The 'social rate of time preference–shadow price of capital' approach begins with the assumption that society should act as if it is maximizing the utility of a representative individual whose utility depends on his or her consumption in each and every time period. From this assumption, it follows that the SDR should equal the 'pure rate of time preference,' which reflects impatience, plus the growth rate of per capita consumption multiplied by the elasticity of the marginal utility of consumption with respect to consumption. The latter term reflects society's preference for more equality in per capita consumption over time than would otherwise occur. Moore et al. (2013a) suggest that, according to this method, the real

SDR should be about 3.5%. However, if this approach were adopted, some public projects might displace some private investments that would yield a higher return on investment. In order to correct this potential problem, flows in or out of private investment should be weighted by the 'shadow price of capital.'

The alternative approach is called the 'social opportunity cost of capital' method. It is based on the view that government projects should have a return higher than alternative (private sector) projects. Unlike the social rate of time preference method, this method derives the value of the SDR based on observed market rates. Specifically, it computes the SDR as a weighted average of the pretax return on investment (reflecting the opportunity cost of private investment), the marginal cost of borrowing (reflecting the cost of incremental foreign funding), and the posttax return on investment (reflecting the opportunity cost of displaced consumption) (Burgess and Zerbe, 2011). The pretax return on investment has the most weight. Using this approach provides estimates of the real SDR of about 5–7%.

Fortunately for practitioners, the difference in the recommended rates are not nearly as large as they used to be in the 1970s, although the choice of the SDR can make a difference between recommending for or against a project, especially for investments with large up-front costs and benefits that extend many years into the future. The decision about what rate to use depends on how one thinks government projects are funded. The social opportunity cost of capital method assumes that government projects primarily displace private investment. In contrast, the shadow price of capital method assumes that government projects are primarily financed by taxes, which reduce consumption.

Many environmental projects have very long-term impacts (*see* Sustainable Development: An Economic Perspective). The impact of climate change policies, for example, may be reflected in events many years from now. Many policy effects may span multiple generations. In these situations, it is generally agreed that analysts should use 'time-declining' discount rates. Thus, for example, an impact that occurs in 200 years would be discounted at a lower rate than an impact that occurs in the 'near future,' say 20 years. One potential criticism of this practice is that it would result in 'time-inconsistent' (or 'hyperbolic') discount rates. However, uncertainty about the future values of the parameters underlying the SDR implies a time-declining schedule of SDRs with only the lowest possible rates applying in the very long term (Weitzman, 2001).

Accounting for Uncertainty

As CBA involves prediction, it can rarely be done with certainty. The standard approach for dealing with 'uncertainty' is to estimate 'expected net benefits.' Analysts convert the problem from one of uncertainty to one of 'risk' by assuming plausible distributions of benefits and costs. They then find the expected value of net benefits. Consider a project with certain opportunity costs of $15 million and uncertain benefits. If there were a 20% chance of obtaining $5 million in benefits and an 80% chance of obtaining $20 million in benefits, the expected benefits would be $17 million and expected net benefits would be $2 million. Under the assumption that society is risk neutral, the analyst can treat the expected values as if they were certain amounts and would recommend a project if it's expected net benefits were positive.

The calculation of expected net benefits, however, does not fully take account of individual preferences for risk. Individuals may not be willing to trade certain amounts of money for expected values of the same magnitude. In such cases, the appropriate measure of benefits is 'option price,' the individual's willingness to pay for the particular distribution of outcomes associated with a policy (Graham, 1981).

As individuals are generally risk averse, policies that reduce risk generally have option prices larger than expected benefits. The difference, 'option value,' is a premium individuals are willing to pay for the reduced risk. As option price can only be elicited through contingent valuation surveys, it is rarely estimated. Unfortunately, relatively little progress has been made in the theoretical determination of the magnitude or even sign of option value. Option value is likely to be reasonably small and expected net benefits is likely to be a reasonably good approximation of option price when it is possible to insure against the relevant risks, different people realize different outcomes from the project and each person is affected by a diversified portfolio of projects whose social returns have low correlations with each other.

The Maximum Expected Net Benefits Rule

Assuming that all of the costs and benefits of a policy have been correctly identified, measured, discounted, adjusted for uncertainty, and aggregated, positive expected net benefits indicate that the policy is potentially Pareto improving, i.e., will increase allocative efficiency. Where there are many alternative, mutually exclusive projects, the analyst should recommend the project with the largest expected net benefits. That project may or may not have the largest benefit–cost ratio. Because benefit–cost ratios vary with the size or scale of projects, they should generally not be used to select among mutually exclusive projects.

In the special case of completely independent projects, all those with positive net benefits should be selected. In this case, this rule corresponds to the rule of selecting all projects with benefit–cost ratios greater than one. If, in addition, government is financially constrained, it should allocate more money to projects with larger marginal net benefits per dollar. In fact, it should allocate money to projects so that the marginal net benefits of an additional dollar of all projects are equal (and greater than one). Of course, if these marginal net benefits per dollar were greater than one, allocative efficiency would be increased by relaxing the fiscal constraint (i.e., by raising taxes).

Sensitivity Analysis and Distributional CBA

Every CBA contains some predictions and some estimates. Impacts may not be predicted accurately and impacts may be valued incorrectly. There is some disagreement over the appropriate value of the SDR. It is common, therefore, for

analysts to test how sensitive their recommendations are to particular assumptions. This process, called 'sensitivity analysis,' typically involves changing the assumed values of a few key parameters to see how net benefits change. Some analysts perform 'Monte Carlo simulations' in which many parameters may vary simultaneously.

One way to incorporate equity into CBA is to attach different weights to different individuals. Thus, for example, disadvantaged children may be given a higher weight than the rest of society. Or, one might want to weight consumers more than producers if, for example, firms are partially foreign-owned. In practice, however, analysts rarely know how much weight should be given to each group. In such situations, it is useful to present the distributional consequences of a proposed policy in addition to the CBA based purely on efficiency. Such a display enables decision makers to perceive fully the distributive implications of the proposals being discussed.

See also: Environmental Economics; Infrastructure Investment; Pareto, Vilfredo (1848–1923); Policy Analysis: Evidence Based Policy-making; Regulation, Empirical Analysis of; Safety, Economics of; Transportation, Supply, and Congestion; Value of Life, Economics of; Welfare Economics.

Bibliography

Arrow, K., Solow, R., Portney, P., Leamer, E., Radner, R., Schuman, H., 1993. Report of the NOAA panel on contingent valuation. Federal Register 58, 4601–4614.

Bateman, I.J., Willis, K.G. (Eds.), 2000. Valuing Environmental Preferences: Theory and Practice of the Contingent Valuation Method in the US, EC, and Developing Countries. Oxford University Press, Oxford, UK.

Blackorby, C., Donaldson, D., 1985. Consumers' surpluses and consistent cost–benefit tests. Social Choice and Welfare 1, 251–262.

Boardman, A.E., Greenberg, D.H., Vining, A.R., Weimer, D.L., 2011. Cost–Benefit Analysis: Concepts and Practice, fourth ed. Prentice-Hall, Upper Saddle River, NJ.

Burgess, D.F., Zerbe, R.O., 2011. Appropriate discounting for benefit–cost analysis. Journal of Benefit–Cost Analysis 2 (2). Article 2.

Burgess, D.F., Zerbe, R.O., 2013. The most appropriate discount rate. Journal of Benefit–Cost Analysis 4 (3), 391–400.

Colvin, P., 1985. The Economic Ideal in British Government: Calculating Costs and Benefits in the 1970s. Manchester University Press, Manchester.

Dahlby, B., 2008. The Marginal Cost of Public Funds: Theory and Applications. MIT Press, Cambridge, MA.

Dinwiddy, C., Teal, F., 1996. Principles of Cost–Benefit Analysis for Developing Countries. Cambridge University Press, New York.

Graham, D.A., 1981. Cost–benefit analysis under uncertainty. American Economic Review 71, 715–725.

Hammitt, J.K., Robinson, L.A., 2011. The income elasticity of the value per statistical life: transferring estimates between high and low income populations. Journal of Benefit–Cost Analysis 2 (1). Article 1.

Haveman, R.H., Krutilla, J.V., 1968. Unemployment, Idle Capacity, and the Evaluation of Public Expenditures. Johns Hopkins University Press, Baltimore, MD.

Haveman, R.H., Weisbrod, B.A., 1975. Defining benefits of public programs: some guidance for policy analysts. Policy Analysis 1, 169–196.

Hochman, H.M., Rodgers, J.D., 1969. Pareto optimal redistribution. American Economic Review 59, 542–557.

Knetsch, J.L., Riyanto, Y.E., Zong, J., 2012. Gain and loss domains and the choice of welfare measure of positive and negative changes. Journal of Benefit–Cost Analysis 3 (4). Article 1.

Krutilla, J.V., 1967. Conservation reconsidered. American Economic Review 57, 777–786.

Little, I.M.D., Mirrlees, J.A., 1974. Project Appraisal and Planning for Developing Countries. Heinemann Educational, London.

Moore, M.A., Boardman, A.E., Vining, A.R., 2013a. More appropriate discounting: the rate of social time preference and the value of the social discount rate. Journal of Benefit–Cost Analysis 4 (1), 1–16.

Moore, M.A., Boardman, A.E., Vining, A.R., 2013b. The choice of the social discount rate and the opportunity cost of public funds. Journal of Benefit–Cost Analysis 4 (3), 401–409.

Squire, L., van der Tak, H., 1975. Economic Analysis of Projects. Johns Hopkins Press, Baltimore, MD.

Steiner, P.O., 1974. Public expenditure budgeting. In: Blinder, A.S., Solow, R.M., Break, G.F., Steiner, P.O., Netzer, D. (Eds.), The Economics of Public Finance. Brookings Institute, Washington, DC.

United States Environmental Protection Agency, 2013. Technical Update of the Social Cost of Carbon for Regulatory Impact Analysis. (Updated and accessed at: 20.02.14.). http://www.epa.gov/climatechange/EPAactivities/economics/scc.html.

Viscusi, W.K., Aldy, J., 2003. The value of statistical life: a critical review of market estimates throughout the world. Journal of Risk and Uncertainty 27 (1), 5–76.

Weisbrod, B.A., 1964. Collective-consumption services of individual-consumption goods. Quarterly Journal of Economics 78 (3), 471–477.

Weitzman, M.L., 2001. Gamma discounting. American Economic Review 91 (1), 260–271.

Whittington, D., MacRae Jr., D., 1986. The issue of standing in cost–benefit analysis. Journal of Policy Analysis and Management 5, 665–682.

Willig, R.D., 1976. Consumer's surplus without apology. American Economic Review 66, 589–597.

Zerbe, R.O., 1998. Is cost–benefit analysis legal? three rules. Journal of Policy Analysis and Management 17 (3), 419–456.

Cost-Benefit and Cost-Effectiveness Analyses in Evaluation Research

Brian T Yates, Department of Psychology, American University, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cost-benefit analysis (CBA) and cost-effectiveness analysis (CEA) are used increasingly in both clinical trials of health and human services, and in evaluations of programs to determine their future funding. Definitions of CBA and CEA are given and contrasted with each other and with requests for 'cost analyses' and 'efficiency' studies. The emerging importance of cost-utility analysis is noted as well. Emerging directions in cost-inclusive assessment are noted, including evaluating costs of using different services to add Quality-Adjusted Life Years and Disability-Adjusted Life Years – nonmonetary measures that can be almost as useful for comparing outcomes of diverse programs as monetary measures, i.e., program benefits. Advantages of conceptualizing 'costs' more broadly as the value of the specific types and amounts of resources used to provide services are explained. Ethical problems in cost-inclusive evaluation are highlighted, and possible solutions suggested. Viewing CBA and CEA as crucial parts of improvement-oriented operations research is recommended. Links for manuals and Web sites for cost-inclusive evaluation in health and human services are provided.

Definitions

Cost-benefit analysis (CBA) and cost-effectiveness analysis (CEA) are two related methods for social scientists and program evaluators to incorporate data on resource requirements of, and resources generated and saved by, social services. At their most basic, both forms of cost-inclusive evaluation compare the value of resources invested in an activity to the value of the outcomes of that activity (Yates, 2009). The activity being examined for cost-benefit or cost-effectiveness can be anything from introducing free health services in a remote village to global eradication of an illness. Persons and organizations can use findings from CBA and CEA to decide whether to engage in or continue an activity and how much time, money, and effort to devote to it. Using different stakeholder perspectives to define, measure, and compare costs, benefits, and effectiveness can result in different findings, different funding, and different impacts for individuals, communities, and human society.

As resource constraints increase, a health or human service tends to be funded only if total benefits of that service exceed total costs of that service by a substantial amount – and relatively soon. Conceptualizing services as human rights or entitlements can drive decision-makers to select the least costly activity, paying less attention to outcomes. This can be unwise: social scientists have developed reliable and valid means of measuring most program outcomes, and of comparing the monetary value of those outcomes over time to program costs over time.

Program Evaluation

Answering questions about how faithfully a program is being implemented, whether its goals are being achieved, and whether the resources devoted to program operations are being used in the most worthwhile manner, all can be considered forms of program evaluation (cf Posavac, 2011). Many research strategies can be used to answer these questions, ranging from expensive and possibly invasive randomized clinical trials to often less expensive but sometimes less valid quasiexperiments (American Evaluation Association, 2013; Shadish et al., 2001). Program evaluation often involves experts from diverse disciplines who apply quantitative and qualitative means of judging programs as they currently operate (*summative* program evaluation), and of improving programs so they operate better (*formative* program evaluation; cf Yates, 1996).

Goals, Effectiveness, Efficacy, and Efficiency

Evaluating the degree to which an activity's desired outcomes or *goals* are being achieved is a judgment of the *effectiveness* of the activity. Activities shown by research to be promising are bundled into programs designed to prevent or remediate problems. Researchers monitor the *fidelity* with which each activity is implemented by highly trained program provider. Persons receiving program services (*clients*) may be assigned randomly to either the program or various control conditions to allow effects of other events, of measurement itself, and of client expectations to be separated from the unique contributions of the program. Participants also may be screened so they have similar characteristics – sometimes so they have only the problem the program is designed to remediate, such as depression, and no other comorbidities, such as anxiety or diabetes. Under these controlled and, some would say, unrealistic conditions, the maximum effectiveness or *efficacy* of the program is measured. When implemented later by providers with varied training and more demands on their time, for clients with combinations of problems, and in programs that must not consume more resources than they will be reimbursed, the *effectiveness* of the program in the 'real-world' is evaluated. Often *effectiveness* is less than *efficacy*.

The 'efficiency' of a program can be distinguished from both its effectiveness and efficacy. In some contexts, 'efficiency' refers to how well the program uses those resources currently available in its environment. This usage resembles *cost-effectiveness* and *cost-benefit* (e.g., Marseille et al., 2004). Other meanings of 'efficiency' may include simple cost per student or cost per patient. Again, 'efficiency' does not have one widely accepted definition, although it appears in speeches about program costs about as often as, and sometimes interchangeably with, 'cost-benefit' and 'cost-effectiveness.' One even hears 'cost-efficiency.'

Actually, *cost*, *cost-benefit*, and *cost-effectiveness* all are terms with specific definitions.

Cost Analysis

An entreaty for 'cost-efficiency,' 'cost-benefit analysis,' or 'cost-effectiveness analysis' often is a request for greater attention to program costs and not a thorough analysis of relationships between costs and benefits or costs and effectiveness. In program evaluation, *costs* usually refer to the total monetary value of the many resources consumed by program activities. A *cost analysis* compares programs purely on their costs, often assuming that any differences in program benefits or effectiveness are minor or unmeasurable. If programs provide entitlements such as income support, aid for children, or health services, program outcomes may not even be measured. Evaluating programs according to their costs but not their outcomes does, of course, risk funding programs that cost the least but also deliver the least, to the extent that program outcomes are functions of the types and amounts of resources available to the program. It also is possible that more-costly programs deliver poorer outcomes than less-costly programs (e.g., Yates, 1994).

Usually *cost analysis* focuses on only the *monetary* value of activities performed by providers. Comprehensive cost analysis performed from the perspective of society considers donated resources such as time volunteered by interns and underpaid providers, and clients.

Benefit Analysis

In cost-inclusive evaluation, *benefits* do not refer to positive outcomes in general, but only monetary outcomes. Most providers of health and human services are trained to focus on changing key client processes and proximal outcomes rather than more distal outcomes. For example, a psychologist may implement activities within cognitive-behavioral therapy and attend primarily to the immediate results of those activities, such as whether the problematic thought patterns changed and anxiety diminished. Outcomes of greater interest to health insurers, employers, and family members might be reduced use of health services due to decreased need, less self-medication with alcohol and other drugs, increased days worked, improved productivity or profitability, and greater availability to provide child- or elder care. These outcomes are typically what programs are funded to achieve. They are either monetary or readily monetizable using methods detailed in monographs (e.g., Gold et al., 1996; Drummond et al., 2005; Yates, 1980, 1996, 1999). All are termed *benefits* and are summed to describe program outcomes to third parties, i.e., to entities other than the client and provider. These can include *negative benefits* such as increased use of health and other human services following treatment, which can be viewed as undesirable by third parties but as ideal by practitioners, clients, and consumer advocates.

Cost-Benefit Analysis

Once summed, the monetary value of activity outcomes can be compared to the total monetary value of resources contributed by providers, clients, and others because those outcomes and resources were measured in the same units. The resulting difference (benefits minus costs or *net benefit*) and ratio (benefits/costs) seem easy to understood and use when choosing among programs to fund (e.g., Yeh, 2009). Programs may differ in whether they are cost-beneficial or not, i.e., whether their net benefit exceeds 0 or their benefit/cost ratio exceeds 1. Programs also can differ in how *much* their benefits exceed costs, either in absolute or relative values, and how soon this can be expected to occur. Because 'benefits' precede 'costs' when describing formulas for both net benefit and benefit-cost ratios, cost-benefit analysis (CBA) might be more understandably termed *benefit-cost analysis* or BCA. The benefit can be seen as a *return* on investment, social or individual, and the amount and speed of that return relative to the value of resources spent is the *return on investment* or ROI. Returns on investment that occur several decades after resource expenditure, as is common in prevention and education programs, are not as valuable as the same returns occurring sooner. This difference is routinely quantified by adjusting temporally distant returns downward for their *net present value* using a *discount rate* to express the return that would have been achieved if the same resources had been invested in alternative activities. Several discount rates are used, resulting in not one but several 'bottom lines' when time to ROI (TROI) and total amount of ROI are reported.

Neglecting certain costs or ignoring particular benefits can, of course, alter indices of cost-benefit to support decisions that later prove to be incorrect. Also, some outcomes are difficult to convert into monetary units, especially in a manner that does not allow under- or overrepresentation due to subjective values of the evaluator or decision-maker. Moreover, some services are considered rights to which all persons are entitled regardless of whether they are or are not found to have benefits which exceed costs. These entitlements include, in many areas of our world, healthy air and water, adequate and nutritious food, opportunities for physical activity, reliable shelter, child- and elder care, protection from physical harm and property theft, comprehensive health care, access to information resources as provided by libraries and the Internet, universal education, opportunities for fulfilling employment, income maintenance when one is not employed, liberty, and happiness (cf United Nations, 2014).

Cost-Effectiveness Analysis (CEA)

When nonmonetary and nonmonetized outcomes are compared to monetary costs for a single program, simple differences or ratios do not yield the same simple decision aids as net benefit or benefit-cost ratios because outcomes are measured in different units than costs. Calculating the 'cost per' unit of nonmonetary outcome may not seem sensical at first unless two or more programs are being compared, but consider the potential usefulness of knowing the *cost per drug-free month* when comparing substance abuse treatment programs, or even the *cost per pound lost* when comparing weight-loss regimens. This is different than 'cost per patient-filled bed,' unless simply filling a bed is the outcome of interest. To be clear: 'Cost per student' would be a measure of cost only. 'Cost per student who graduated with scores exceeding minima required' would be a measure of cost-effectiveness.

Cost-Utility Analysis

An increasingly common form of cost-effectiveness analysis (CEA), *cost-utility analysis (CUA)*, uses nonmonetary measures of outcome that can be applied to many programs. Essentially, program outcomes are measured as not program-specific measures such as 'apples' or 'oranges,' but as more general 'fruits.' (Programs are, of course, free to also report outcomes on more specialized measures, e.g., 'apples,' 'oranges,' or 'lemons.'). *Well-being* is a commonly advocated outcome for which there are useful measures (cf Drummond et al., 2005). Critics deride well-being as too subjective and as difficult to measure with acceptable reliability and validity.

A cost-inclusive index more agreeable to many than net cost or benefit-cost is *cost per year of life saved*. Diverse programs can be compared in terms of years of life saved. In recognition that the subjective value of a year of life can be less if one is seriously ill and suffering, years of life saved usually are adjusted for *quality*. Accepted algorithms are used to calculate *quality-adjusted life years* (QALYs; cf Gold et al., 1996). Careful research can determine the change (hopefully improvement) in QALY produced by different medical or other interventions (e.g., Freed et al., 2007). Combining measures of gain in QALY with information on the costs of activities that seem to cause those gains results in estimates of the cost of adding a quality-adjusted year of life for patients, or *cost per QALY* (e.g., £/QALY). Similar metrics that include costs can be used to compare the cost-benefit of prevention versus treatment, and health interventions versus efforts to improve quality of life through education, the arts, income support, and more.

A related measure is years of life lost due to disability caused by health and mental health problems and premature death: *DALY*. The DALY combines findings regarding program effects on mortality and morbidity into a single variable: a combination that can facilitate higher-level decisions about funding of public health interventions. Costs of efforts to reduce disability and premature cost often are expressed as *cost per DALY* saved.

Schools of Thought and Leading Advocates

CBA and CEA can be viewed as methods of making difficult decisions in which some individuals gain resources or services at the expense of others. CBA and CEA promise a decision process that is fair because it attempts to weigh social investment against social gain objectively and quantitatively. Critics note that this logical positivist framework has been shown to generate decisions that are fundamentally subjective and that often benefit persons and organizations that have the largest monetary stake in the decision (cf Guba and Lincoln, 1989). The critical, constructivist position makes rejection of cost considerations easier for researchers, service providers, advocates, and consumers who would rather not include costs of services or their monetary consequences decisions about which services to fund and offer, or defund and stop.

Nevertheless, as long as there are limits on the availability, quality, and amount of resources needed to provide services and produce products, the amount of those resources needed for a social program will factor into any realistic decision-making. Evaluations and decisions based on at least the monetary value of those resources will continue to be made, informally at least. CBA and CEA can make the measures and biases of each stakeholder group explicit for all to see. The process by which those resources are valued and monetized can be transparent. In addition, CBA and CEA consider outcomes of human endeavors as well as their costs.

Some critics of human services seize upon program evaluation in general, and CBA and CEA in particular, as a means of challenging the worth of human services and especially entitlement programs. Providing treatment to substance abusers, for example, can be challenged as a waste of public funds. This position has been refuted by findings that substance abuse treatment can cause a subsequent *decrease* in use of expensive health and other human services (Ettner et al., 2005). Concern remains about the cost-benefit and cost-effectiveness of other services, such as free health services and income maintenance. Resolving these issues can be facilitated by cost-inclusive evaluations that are well beyond the scope of this article. Finally, many cost-effectiveness and cost-benefit analysts recognize that their findings should not be the sole criteria on which decisions should be based. Again, we all have fundamental rights which no evaluation should deny any of us.

Changes in Focus or Emphasis over Time

CBA can be viewed as having always been a normal part of most individuals' pragmatic decision-making, in which what is required by an enterprise is compared to what may come of that enterprise. In modern European history, the first formal proposal of CBA was made by Dupuit, a French engineer. While cost-inclusive evaluation in health, substance abuse, and criminal justice has proceeded quickly to approximate the complexity of CBA required to justify dams and other projects by the US Army Corps of Engineers in the 1930s (U.S. Flood Control Act), other human services are only now beginning to measure and compare costs and outcomes carefully and comprehensively. Mental health and substance abuse treatment have a surprisingly long history of advocating, training, and using CEA, and to a lesser degree CBA (e.g., Carter and Newman, 1976; Fishman, 1981; Sorensen and Phipps, 1975; Yates, 1977, 1978). Services for children and families have developed a series of manuals to help programs and researchers assess costs (Dickey et al., 1999; Calculating the Costs of Child Welfare Services Workshop, 2013), but efforts to assess cost-effectiveness and cost-benefits of these and many other human services have only thin histories of cost-inclusive evaluation (cf Levin and McEwan, 2001; Persaud, 2007; e.g., Yates, 2005; Yates et al., 1979).

Some state legislatures in the United States have used CBA to control the amount of funds consumed in health and other human services, especially those that might reduce future costs for criminal justice or health services (cf Pew Charitable Trusts, 2013). Best known are the CBA conducted by Steve Aos and colleagues in response to requests from the Washington State legislature (Washington State Institute for Public Policy, 2011).

As in other areas of program evaluation, 'thumbs up or thumbs down' evaluations have only begun to evolve into 'how can we make it (even) better?' evaluations. Because funding of programs increasingly depends on whether the program meets criteria such as a cost-benefit ratio that is well above 1.0, or a maximum cost per QALY (e.g., Towse, 2009), much expertise has been devoted to standardizing the measurement and calculation procedures. Starting with Gold et al. (1996) and continuing through Yates (1999) and Drummond et al. (2005), consensus is being achieved on more methods. The need for cost-inclusive evaluation has been recognized widely, resulting in an increasing number of submissions to professional journals of CBA and CEA in health and other human services (cf Herman et al., 2009; Yates, 1994).

Current and Emerging Directions in Theory and Research

Utility Assessment

A major advance in cost-inclusive evaluation has been the adoption of outcome measures that are not monetary, but that are standardized across diverse services, such as the QALY and DALY introduced earlier. These measures promise to evolve into a common metric of outcomes that will allow a great diversity of human services to be compared for cost-effectiveness.

Willingness-to-pay to obtain a service and willingness-to-accept withdrawal of a service have been advocated as a measure of the worth of a program since CBA was first formalized (cf Johansson, 1995). Part of the appeal is that this outcome is consumer-driven and could be inexpensive to measure. The willingness-to-pay approach to outcome assessment seems particularly oriented toward business models of social enterprises originating in unabashedly capitalist contexts. While consumers' willingness-to-pay for services and products may be a useful metric when the decision-maker is construed as highly rational and able to pay a range of prices, considerable efforts have been exerted by communications specialists in advertising and politics to distort consumer perceptions.

Willingness-to-pay seems a poor choice for valuing outcomes when consumers seek or are prescribed services precisely because their recent choices do not seem entirely rational, as may be the case for substance abuse, mental health, and criminal justice services. Not having sufficient monetary resources to consider investing in services that may not benefit the consumer for years or decades has led many families to educate only some of their children, a choice that reflects willingness-to-pay but seems a poor choice for society in the long run. Asking economically disempowered consumers to 'imagine' they had a certain amount of resources to invest in services may not generate the rational decisions anticipated, but is likely to uncover deep frustration about economic hardships.

BCA Using Nonmonetary Units

As noted earlier, the primary advantage of CBA over CEA is that costs 'input' and outcomes 'output' are measured in the same units. This allows costs to be subtracted from outcomes to calculate net benefit. It also allows calculation of the *ratio* of outcomes to costs. Costs and benefits of different activities within a program can be assessed in the same nonmonetary units, using similar rating scales for *psychological cost* and *psychological benefit* (cf Yates, 1978). The result provided clinically useful information on client perceptions of which activities were best and worst in a weight-loss program. Similar nonmonetary CBA could be used in group decision-making to allow explication, quantification, and comparison of different program activities consumers and providers perceived to be least difficult and most helpful.

Evidence-Based Practices

Although preferences in training as well as funding have emerged for activities that careful research has shown to produce the outcomes they target, information on the costs of those *evidence-based practices* is rarely reported. The only evidence that seems to matter is nonmonetary effectiveness, and neither benefit nor cost. Many activities, done often and well enough by highly trained professionals working intensively and individually with clients in controlled settings, can be shown to meet criteria for specific outcomes. Calls for including costs and benefits among the evidence for specific practices have gone largely unanswered to date (Yates, 2012a).

For replicability alone, it would seem important to include in descriptions of evidence-based practices careful specification of the types and amounts of each of the key resources needed to implement such a practice. Many human service fields could benefit from reporting the resources needed for a specific evidence-based practice so as to allow researchers and practitioners to estimate the costs of those resources in current monetary values and in their settings (Yates, 1996). Including monetary outcomes, i.e., benefits, when measured, as well as common indices such as QALY and DALY added or saved, could advance the adoption and funding of promising practices.

Requiring inclusion of information on service resources consumed, and their costs, also could encourage researchers to not stop once they have found evidence that a program works. The next step in helping a practice become practical may be to then remove activities that are less responsible for program outcomes, leaving only those that are necessary. This 'efficiency services research,' as it might be called, also could determine the intensity of activities needed to achieve outcomes – for example, whether biweekly rather than weekly meetings were sufficient, and whether trained paraprofessionals rather than psychologists and physician's assistants rather than physicians could implement the essential activities.

Of course, one can also ask for evidence that inclusion of costs in evaluation research is itself effective, and whether it helps to maintain and improve service outcomes. It would be hypocritical not to also ask, is CEA cost-effective? … Is CBA cost-beneficial? These issues of meta-cost-inclusive evaluation research have only begun to be broached empirically (cf Herman et al., 2009; Yates and Taub, 2003).

Methodological Issues and Problems

Costs and Outcomes: Biases in Comprehensive Measurement and Monetization

A persistent and widespread problem in cost-inclusive evaluations is careful measurement of both costs *and* outcomes in the same study and in a standard manner (Siegel et al., 1996). Adding monetary outcomes to reports, and contrasting them with monetary costs, has led to major improvements in use of cost-inclusive evaluation. Most cost-inclusive evaluations and research reports focus on measuring, monetizing, and analyzing either resources invested or outcomes produced but rarely both with equal thoroughness.

Unfortunately, although greater attention has been devoted in the past decade to monetizing outcomes of human services such as substance abuse treatment and prevention of criminal behavior, including careful adjustments of outcomes for quality of research design, costs of such services have been accepted largely at face value. Costs not only need to be adjusted for changes in monetary value over time, such as inflation or deflation, but also can vary between urban and rural regions and can depend on whether the program is offered under the auspices of a larger institution or is a 'stand-alone' operation.

Costs: Monetization of Volunteered and Donated Resources

If costs are undermeasured, programs may seem to be more cost-beneficial, cost-effective, or to have greater cost-utility than is actually the case (Yates, 2012b). Because funding flows more to programs whose outcomes are reported to exceed costs to greater extents, there are strong incentives to measure benefits more comprehensively than costs.

Efforts have been made to standardize measurement of costs using principles of accounting and economics (e.g., Caffray and Chatterji, 2009). French's work is the best known (e.g., French, 2003; McCollister et al., 2009). Paper-and-pencil, computerized, and structured interviews seem to generate more reliable cost data than other methods (cf Zarkin et al., 2004).

Budgets and even accounting records can be poor sources for data on how funds were actually spent, however. Most cost measures used in contemporary CBA, CEA, and CUA neglect entirely what is perhaps the most necessary and widely used resource of all: the time of volunteer providers, and of the clients and family members who transport themselves and others to and from service sites, plus costs associated with transportation vehicles. Accounting records rarely include amounts and estimated values of these 'invisible' but necessary resources.

Some researchers have included measures of volunteered and donated resources (e.g., Yates et al., 2011), and resources devoted to treatment by clients as well (e.g., McCollister et al., 2009). These studies have found substantially different relationships between costs and outcomes when these oft-ignored resources are monetized. For example, cost-inclusive evaluations of consumer-operated services found that the monetary value of volunteered and donated resources exceeded the monetary budget of some consumer-operated human services (Yates et al., 2011). Fears that the cost of these programs would be exaggerated relative to programs that did not include volunteered and donated resources prompted. Yates and colleague Freed to report volunteered and donated resources separately from 'cash costs,' allaying some program's concerns and defusing resistance to collection of resource data.

Costs: Itemization in Reporting to Allow Generalization by Readers

Many articles and manuscripts describe in only a summary manner how cost and benefit data were collected. Reporting of cost evaluation methodology often is so minimal that, if same were done for nonmonetary outcomes, reviewers would reject it immediately. If cost- and benefit-assessment methods were reported in detail sufficient to allow replication by other researchers, cost-inclusive assessments could be more useful to readers and researchers in other countries and at future times. Costs and benefits should be itemized when reported, and quantified with their 'native units,' e.g., hours, square meters, and improvement in days employed, rather than just 'dollars invested' and 'dollars generated and saved.' In addition to reporting the total cost of a treatment program, the specific resources used should also be specified (e.g., X hours per week of direct services by a social worker, for instance) along with the pay rate (salary + benefits) in effect at the time data were collected (cf Sava et al., 2009).

This would allow decision-makers in other monetary systems operating a decade or more after publication of research findings to estimate the costs of each resource used and the value of each benefit generated. Without this, an intervention that has superior cost-benefit or cost-effectiveness to another intervention in one economy may be overgeneralized to be superior in other economies, which have such different pay rates as to reverse which intervention is more cost-beneficial and more cost-effective (e.g., Sava et al., 2009; cf Yates, 2011). Similar itemization of benefits would allow adjustment and potential similar reversals of cost-benefit findings for countries that apply different charges for emergency room, inpatient, and outpatient services.

Outcomes: Monetization of Lifetime Earnings Can Exacerbate Discrimination

Monetization of program outcomes often includes not only savings in future health and criminal justice services that will not be required because of program impact, but also increments in future earnings that result from training, education, or longevity enhancement. This income + benefits calculation is common in evaluations of the worth of a college education as well as estimation of DALY generated by some health programs. If some genders, ethnicities, or age groups use more expensive health or criminal justice services, savings in use of those services would be larger for, and favor focusing efforts on, those particular genders, ethnicities, or age groups.

That may seem fair, and certainly monetizing the expected enhancement of lifetime earnings acknowledges the importance of employment and entrepreneurship. This same approach to monetizing outcomes also could perpetuate or worsen historical discrimination that, for example, pays males more than females, some ethnic groups more than others, and

certain age groups more than others. Increments in lifetime earnings usually favor efforts to enhance income of younger individuals because they have more years of life to accumulate larger totals of improved earnings. Some readers may observe that ageist, sexist, and racist preferences simply reflect the values that society has already placed on the worth of different individuals' lives. Others may react to this issue by rejecting cost-inclusive evaluations entirely. It seems possible, however, that adjustments could be made to wage rates so that the same rate was used for all ages, genders, races. The youth preference inherent in measures of lifetime earnings could be reduced by calculating benefits to include (1) the commonly observed diminution of intervention effects over years of life, (2) savings in services to the aging that may result from their continued employment, and (3) monetary valuation of volunteer activities including child care and mentoring by the aging.

Costs and Outcomes: Statistical Models Capture Uncertainty, but Cost-Inclusive Indices Seldom Do

Even if they include volunteered and donated resources, and benefits assessed from multiple perspectives, cost-inclusive evaluations may report only total costs and total benefits for a program. This highly efficient method does not reflect the dramatic variability of response to program activities between individual clients whether those are persons, families, communities, nations, or regions of the world. Expressing that variability can be problematic, in that it introduces uncertainty and another statistic next to what was a rather neat single number: the benefit-cost or cost-effectiveness ratio. High variability that deviates from the typical bell-shaped curve of the 'normal' distribution, and which does not disappear even with large samples, seems all too typical of health and other human services. This interindividual and intersite variability seldom is reported by researchers or recalled by funders when using cost-inclusive indices to make funding decisions.

Even worse, small differences between mean or median costs, benefits, and net benefit for competing programs are too often interpreted as 'real' because of the dollar or pound signs involved, when the uncertainty in costs and benefits introduces sufficient variability as to render statistical tests of these apparent differences profoundly nonsignificant, or at least the differences in their effect sizes trivial.

Uncertainty of cost-benefit and cost-effectiveness can, of course, be expressed with confidence intervals, and its effects can be examined in analyses that simulate cost and outcome variability. One comparison of special interest that is rarely reported is against a net benefit of 0, and a benefit-cost ratio of 1.00. In the author's experience, some apparently substantial mean and median net benefits and promising benefit-cost ratios prove to be statistically or realistically the same as 0 and 1, respectively!

Benefit-Cost and Cost-Effectiveness Ratios Do Not Describe Relationships Over a Range of Values

A long-recognized but seldom solved problem with ratios that include costs is that they are expressions, which discard information on both costs and outcomes and provide only an expression of the relationship between those, and often just in one specific context. For example, Program A might have cost $100 000 per year and have generated $200 000 in outcomes. It would have a net benefit of $100 000 and a benefit-cost ratio of 2:1. Program B might address the same sort of problem but cost only $50 000 with $150 000 in outcomes. The net benefit is still $100 000 but the benefit-cost ratio is 3:1. Is Program B better than Program A? Maybe. But suppose Program B was now funded for $100 000 rather than $50 000. Could we expect that the same benefit-cost ratio of 3:1 would hold at the higher level of funding? Not necessarily. Typically there are diminishing returns on investments in a given program and a given technology, and even on investments in progressively more impactful technologies (cf Yates, 1978). Ratios of benefit/cost, effectiveness/cost, or cost/QALY are but transient slopes at a particular cost value of a potentially complex and even nonmonotonic cost-outcome function at that one moment of measurement (Yates, 1996).

Maximizing Cost-Benefit, Cost-Effectiveness, and Cost-Utility

Operations research has long been used in military and business settings to maximize outcomes, such as product deliveries, within resource constraints. Operations researchers first construct mathematical models that quantify the contribution to overall outcomes made by each activity within the program. These activity-specific cost-effectiveness or cost-benefit indices are part of this model. The amount of each type of resource used in each activity also is captured in the model, along with the total amount of each resource type that is available in the context in which the program operates. Linear programming then manipulates this resource-activity-process-outcome (RAPO) model to either maximize overall outcome within the resource constraints, or minimize the overall cost of achieving a targeted overall outcome.

Cost-inclusive evaluators have long understood that CEA and CBA can be preliminary stages of operations research (Yates, 1980, 1996). Sufficiently detailed models have only recently been applied to determine the most cost-effective solutions to complex and socially significant health problems such as HIV prevention and treatment (cf Lin et al., 2012). Yates and colleagues (Yates, 1997) have attempted to isolate, measure, and quantify relationships between (1) *resources* consumed by a program, (2) the *activities* that program uses to engage clients so that (3) specific biological, psychological, and social *processes* are instigated, enhanced, diminished, or eliminated such that specific (4) nonmonetary and monetary *outcomes* are achieved, both proximally and distally (cf Yates, 1999). This model promises to facilitate application of linear programming and other operations research solutions to problems of program optimization. It has also facilitated understanding of program failures (Yates, 2002).

Social Return on Investment (SROI): CBA Anew

The SROI approach to improving society is used to fund those health, education, and other efforts that are most likely to yield the best return for society. This form of CBA can be used by private and public agencies at local, national, and international levels to decide how to best allocate their limited resources.

Ideally, SROI analyses identify not only the time to return on investment but also use measures that funders have agreed are valid in a global context. Ideally, SROI allows the same funds to be used repeatedly as monetary resources are converted to societal resources that also have monetary value. The later could include traditional benefit measures such as improved productivity and reduced use of health care, both due to the decreased in physical or mental dysfunction and engendered by enhanced well-being and creativity. Questions can be raised about how soon and how completely these and other social resources can be recognized as monetary resources available for new investments. Arguably, private and public entities should allow and fund all efforts that can be shown to have returns that, after adjustment for any delays, exceed by a confident amount the investment require.

Of course, errors and biases in measurement of benefits and costs noted earlier can substantially distort SROI decisions – a problem that proponents of SROI address by advocating involvement of all stakeholder perspectives, careful monetization, and validation of findings. Eschewing the subjective and measuring only what is 'material' is a principle that risks making decisions that damage health while increasing capital. Fortunately, SROI specialists are both aware of this danger and working on solutions. This is, perhaps, the most novel and promising application of CBA and CEA in contemporary settings.

See also: Alcohol Interventions: Disease Models vs. Harm Reduction; Applied Social Research, History of; Behavioral Economics, History of; Ethical Questions in Social and Behavioral Sciences, History of; Evidence-Based Practice; Implementation Science; Methodology of the History of the Social and Behavioral Sciences; Policy Analysis; Pragmatist Social Thought, History of; Prevention Research; Quantification in the History of the Social Sciences; Science and Politics: Value Neutrality; Social Question: Impact on Social Thought; Utilitarian Social Thought, History of.

Bibliography

American Evaluation Association, 2013. An Evaluation Roadmap for a More Effective Government. http://www.eval.org/d/do/472 (accessed 13.02.14.).

APA Presidential Task Force on Evidence Based Practice, 2006. Evidence-based practice in psychology. American Psychologist 61, 271–285.

Caffray, C.M., Chatterji, P., 2009. Developing an internet-based survey to collect program cost data. Evaluation and Program Planning 32, 62–73.

Calculating the Costs of Child Welfare Services Workshop, 2013. Cost Analysis in Program Evaluation: A Guide for Child Welfare Researchers and Service Providers. Children's Bureau, Administration for Children and Families, U.S. Department of Health and Human Services, Washington, DC.

Carter, D.E., Newman, F.L., 1976. A Client-Oriented System of Mental Health Service Delivery and Program Management: A Workbook and Guide. National Institute of Mental Health, Rockville, MD. DHEW Publication No. (ADM), pp. 76–307.

Dickey, B., Beecham, J.K., Latimer, E., Leff, J.S., 1999. Estimating Per Unit Treatment Costs for Mental Health and Substance Abuse Programs. Human Services Research Institute, Cambridge, MA.

Drummond, M.F., Sculpher, M.J., Torrance, G.W., O'Brien, B.J., Stoddart, G.L., 2005. Methods for Economic Evaluation of Health Care Programmes. Oxford University Press, New York, NY.

Ettner, S.L., Huang, D., Evans, E., Ash, D.R., Hardy, M., Jourabchi, M., Hser, Y., 2005. Benefit-cost in the California Treatment Outcome Project: does substance abuse treatment "pay for itself?" Health Research and Education Trust 192–213. http://dx.doi.org/10.1.111/j.1475-6773.2005.00466.x.

Fishman, D.B., 1981. A Cost-Effectiveness Methodology for Community Mental Health Centers: Development and Pilot Test. National Institute of Mental Health, Rockville, MD. DHHS Publication No. (ADM), pp. 81–767.

Freed, M.C., Rohan, K.J., Yates, B.T., 2007. Estimating health utilities and quality adjusted life years in seasonal affective disorder research. Journal of Affective Disorders 100, 83–89.

French, M.T., 2003. Drug Abuse Treatment Cost Analysis Program (DATCAP): User's Manual, eighth ed. University of Miami, Miami, FL.

Gold, M.R., Siegel, J.E., Russell, L.B., Weinstein, M.C., 1996. Cost-Effectiveness in Health and Medicine. Oxford University Press, New York, NY.

Guba, E.G., Lincoln, Y.S., 1989. Fourth Generation Evaluation. Sage Publications, Newbury Park.

Herman, P.M., Avery, D.J., Schemp, C.S., Walsh, M.E., 2009. Are cost-inclusive evaluations worth the effort? Evaluation and Program Planning 32, 55–61.

Johansson, P.O., 1995. Evaluating Health Risks. Cambridge University Press, Cambridge, UK.

Levin, H.M., McEwan, P.J., 2001. Cost-Effectiveness Analysis, second ed. Sage Publications, Thousand Oaks, CA.

Lin, F., Terrell, C., Sansom, S., 2012. HIV Prevention Resource Allocation Model: A Mathematical Model Designed for ECHPP Interventions and Lessons from Philadelphia. Presentation at PS12–1201 Grantee Meeting March 26–28, 2012.

Marseille, E., Daandona, L., Saba, J., McConnel, C., Rollins, B., Gaist, P., et al., 2004. Assessing the efficiency of HIV prevention around the world: methods of the prevent AIDS network for cost-effectiveness analysis (PANCEA) project. Health Services Research 39, 1993–2012.

McCollister, K.E., French, M.T., Pyne, J.M., Booth, B., Rapp, R., Carr, C., 2009. The cost of treating addiction from the client's perspective: results from a multi-modality application of the client DATCAP. Drug and Alcohol Dependence 104, 241–248.

Persaud, N., 2007. Conceptual and Practical Analysis of Costs and Benefits in Evaluation: Developing a Cost Analysis Tool for Practical Program Evaluation. USA: ProQuest Information and Learning Company.

Pew Charitable Trusts, 2013. States' Use of Cost-benefit Analysis: Improving Results for Taxypayers. July 2013. Downloaded 20.02.14 from: MacArthur Foundation http://pewstates.org/resultsfirst.

Posavac, E.J., 2011. Program Evaluation: Methods and Case Studies, eighth ed. Prentice Hall, Boston, MA.

Rogers, P.J., Stevens, K., Boymal, J., 2009. Qualitative cost-benefit evaluation of complex, emergent programs. Evaluation and Program Planning 32, 83–90.

Sava, F.A., Yates, B.T., Lupu, V., Hatieganu, I., Szentagotai, A., David, D., 2009. Cost-effectiveness and cost-utility of cognitive therapy, rational emotive behavioral therapy, and fluoxetine (Prozac ®) in treating depression: a randomized clinical trial. Journal of Clinical Psychology 65, 36–52.

Shadish, W.R., Cook, T.R., Campbell, D.T., 2001. Experimental and Quasi-Experimental Designs for Generalized Causal Inference, second ed. Houghton Mifflin Harcourt, New York.

Siegel, J.E., Weinstein, M.C., Russell, L.B., Gold, M.R., 1996. Recommendations for reporting cost-effectiveness analyses. Journal of the American Medical Association 276 (16), 1339–1341.

Sorensen, J.E., Phipps, D.W., 1975. Cost-finding and Rate-Setting for Community Mental Health Centers. National Institute of Mental Health, Rockville, MD. DHEW Publication No. (ADM), pp. 76–291.

Towse, A., 2009. Should NICE's threshold range for cost per QALY be raised? Yes. British Medical Journal 339, 181.

United Nations, 2014. Universal Declaration of Human Rights. http://www.un.org/en/documents/udhr/index.shtml (accessed 06.03.14.).

Washington State Institute for Public Policy, 2011. Return on Investment: Evidence-Based Options to Improve Statewide Outcomes. Author, Olympia, WA. Available at: wsipp.wa.gov.

Yates, B.T., 1977. A cost-effectiveness analysis of a residential treatment program for behaviorally disturbed children. In: Mittler, P. (Ed.), Research to Practice in Mental Retardation, vol. 1. University Park, Baltimore, pp. 435–445.

Yates, B.T., 1978. Improving the cost-effectiveness of obesity programs: reducing the cost per pound. International Journal of Obesity 2, 249–266.

Yates, B.T., 1980. Improving Effectiveness and Reducing Costs in Mental Health. Thomas, Springfield, IL.

Yates, B.T., 1994. Toward the incorporation of costs, cost-effectiveness analysis, and cost-benefit analysis into clinical research. Journal of Consulting and Clinical Psychology 62, 729–736.

Yates, B.T., 1996. Analyzing Costs, Procedures, Processes, and Outcomes in Human Services: An Introduction. Sage Publications, Thousand Oaks, CA. Chosen for inclusion in the 2010 edition of Sage Research Methods Online.

Yates, B.T., 1997. Formative evaluation of costs, cost-effectiveness, and cost-benefit: toward cost procedure process outcome analysis. In: Bickman, L., Rog, D. (Eds.), Handbook of Applied Social Research Methods. Sage, Thousand Oaks, CA, pp. 285–314.

Yates, B.T., 1999. Measuring and Improving Cost, Cost-Effectiveness, and Cost-Benefit for Substance Abuse Treatment Programs. National Institute on Drug Abuse, Bethesda, MD. NIH Publication Number 99-4518, 124-page manual, initial press run 10,000 copies. Full copy readable and downloadable from NIDA site: http://archives.drugabuse.gov/IMPCOST/IMPCOSTIndex.html.

Yates, B.T., 2002. Roles for psychological procedures, and psychological processes, in cost-offset research: cost procedure process outcome analysis. In: Cummings, N.A., O'Donohue, W.T., Ferguson, K.E. (Eds.), The Impact of Medical Cost Offset on Practice and Research: Making It Work for You. Context Press, Reno, NV, pp. 91–123.

Yates, B.T., 2005. Cost-effectiveness analysis and cost-benefit analysis. In: DuBois, D., Karcher, M. (Eds.), Handbook for Youth Mentoring. Sage Publications, Thousand Oaks, CA, pp. 525–545.

Yates, B.T., 2009. Cost-inclusive evaluation: a banquet of approaches for including costs, benefits, and cost-effectiveness and cost-benefit analyses in your next evaluation. Evaluation and Program Planning 32, 54–56.

Yates, B.T., 2011. Delivery systems can determine therapy costs, and effectiveness, more than type of therapy. Perspectives on Psychological Science 6, 498–502.

Yates, B.T., 2012a. Quantitative approaches to outcome measurement: a commentary. In: Magnabosco, J., Manderscheid, R. (Eds.), Outcomes Measurement in Human Services, second ed. National Association of Social Workers Press, Washington, DC, pp. 47–57.

Yates, B.T., 2012b. Step arounds for common pitfalls when valuing resources used versus resources produced. In: Julnes, G. (Ed.), Promoting Valuation in the Public Interest: Informing Policies for Judging Value in Evaluation, vol. 133. New Directions in Program Evaluation, pp. 43–52.

Yates, B.T., Haven, W.G., Thoresen, C.E., 1979. Cost-effectiveness analysis at learning house: how much change for how much money? In: Stumphauzer, J.S. (Ed.), Progress in Behavior Therapy with Delinquents. Charles C Thomas, Springfield, IL, pp. 186–222.

Yates, B.T., Mannix, D., Freed, M.C., Campbell, J., Johnsen, M., Jones, K., Blyler, C., 2011. Consumer-operated service programs: monetary and donated costs and cost-effectiveness. Psychiatric Rehabilitation Journal 35 (2), 91–99.

Yates, B.T., Taub, J., 2003. Assessing the costs, benefits, cost-effectiveness, and cost-benefit of psychological assessment: we should, we can, and here's how. Psychological Assessment 15, 478–495.

Yeh, S.S., 2009. Shifting the bell curve: the benefits and costs of raising student achievement. Evaluation and Program Planning 32, 74–82.

Zarkin, G.A., Dulap, L.J., Homsi, G., 2004. The substance abuse services cost analysis program (SASCAP): a new method for estimating drug treatment services cost. Evaluation and Program Planning 27, 35–43.

Relevant Websites

http://www.acf.hhs.gov/programs/cb/resource/cost-workgroup – Cost Assessment, Children's Bureau manual for cost assessment, Calculating the Costs of Child Welfare Services., Component-specific costs.

http://www.usda.gov/oce/risk_assessment/ – Government Offices and Agencies US Department of Agriculture.

http://archives.drugabuse.gov/pdf/Costs.pdf – Manuals, NIDA Manual for Measuring Costs, Cost-Effectiveness, Cost-Benefit.

https://research.tufts-nemc.org/cear4/ – Registries, Tufts Cost-Effectiveness Registry.

http://yosemite.epa.gov/ee/epa/eed.nsf/pages/guidelines.html – US Environmental Protection Agency.

https://www.nlm.nih.gov/nichsr/hta101/ta10106.html – US National Institutes of Health Library.

Counseling and Psychotherapy: Ethnic and Cultural Differences

Vivian Ota Wang, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 4, pp. 2854–2859, © 2011, Elsevier Ltd., with revisions made by the Editor.

Abstract

Scholars and mental health professionals have argued that racial–cultural issues are obstacles for those seeking help. When available, cross-cultural counseling has focused on misunderstanding client health beliefs and cultural customs, practitioners' historical, social, and political unfamiliarity with discrimination, professional cultures, biased theoretical models, and failure to create realistic understandings of members of racial–cultural groups in the USA. Recognizing the importance of meeting the needs for all who seek mental health services, cross-cultural scholars and educators have been providing direction in cross-cultural psychology research, practice, and training. Competency guidelines and ethical and professional mandates have come from these efforts. Unfortunately, emanating from a cultural pluralism perspective, cross-cultural psychology has often resulted in overgeneralized and homogenized group stereotypes without attention to within-group variation and individual differences. A transition from cross-cultural toward socially constructivist multicultural psychology has been gaining momentum that shifts learning from cognitively perceived culture-specific group norms to an integrated cognitive-affective racial–cultural identity perspective that includes social, historical, and political frameworks. Understanding personal racial and cultural identities and developing self-acceptance as the means for respectfully and competently working with people from all racial–cultural groups has become the primary and necessary prerequisites for developing multicultural competence in the twenty-first century.

Scholars and mental health professionals have argued that racial–cultural issues are obstacles for those seeking help. When available, cross-cultural counseling has focused on misunderstanding client health beliefs and cultural customs, practitioners' historical, social, and political unfamiliarity with discrimination, professional cultures, biased theoretical models, and failure to create realistic understandings of members of racial–cultural groups in the USA. Recognizing the importance of meeting the needs for all who seek mental health services, cross-cultural scholars and educators have been providing direction in cross-cultural psychology research, practice, and training. Competency guidelines and ethical and professional mandates have come from these efforts. Unfortunately, emanating from a cultural pluralism perspective, cross-cultural psychology has often resulted in overgeneralized and homogenized group stereotypes without attention to within-group variation and individual differences. A transition from cross-cultural toward socially constructivist multicultural psychology has been gaining momentum that shifts learning from cognitively perceived culture-specific group norms to an integrated cognitive-affective racial–cultural identity perspective that includes social, historical, and political frameworks. Understanding personal racial and cultural identities and developing self-acceptance as the means for respectfully and competently working with people from all racial–cultural groups has become the primary and necessary prerequisites for developing multicultural competence in the twenty-first century.

Introduction

Scholars and mental health professionals have argued that racial–cultural issues have been obstacles for those seeking help (Sue and Sue, 1999). When counseling has been available, the focus has often been placed upon failures to create realistic understanding of members of racial–cultural groups in the United States, practitioners' unfamiliarity with historical, social, and political discrimination and oppression (Guthrie, 1998), narrow professional cultures, and biased theoretical models (Carter, 1995).

In many regards, scholars, educators, and practitioners have been focusing upon changing population demographics as a driving force necessitating multicultural inclusion. For example, the US Bureau of the Census data (1990) has shown significant increases in the number of Blacks, American Indians, Asian/Pacific Islanders, and Latinos. Between 1970 and 1990, the Latino population increased 145%, the Asian/Pacific Islander population by 375% with 22.3 million Latinos and 7.2 million Asian/Pacific Islanders now living in the USA. Currently one in three US Americans has been either born outside of the USA or are members of visible racial–cultural groups. Given the present rate of immigration as well as fertility rates, researchers at the Bureau of Immigration have predicted that by the year 2040 Whites will only comprise 62% of the total US population.

Recognizing the importance of meeting the needs for all seeking mental health services, scholars and educators have developed competency guidelines and ethical mandates for integrating multicultural perspectives into psychology research, practice, and training (American Psychological Association, 1991; Sue et al., 1982, 1992, 1998), educational programs (Carter, 1995; Ridley et al., 1994), and evaluation instruments (Pope-Davis and Dings, 1995).

Culture

What is culture? Culture at its core understands human behavior through causes and effects *and/or* content and

processes. While cultural content is focused on knowledge about learned and acquired beliefs, art, morals, customs, and habits, cultural processes attend to the process and perpetuation of pattern transmission. Thus, psychosocial cultural perspectives are socially transmitted attitudes and behaviors emerging from interpersonal interactions transmitted generationally, not through biological inheritance but by formal and informal methods of teaching and demonstration. Thus, culture can include historical dimensions emphasizing traditions, normative dimensions focusing upon unique and shared experiences, psychological dimensions including adjustment, learning and group behaviors, structural dimensions attending to the composition of and interrelated patterns in organizations, and genetic dimensions focusing on the origins of culture as interactive adaptive processes between groups and their environment. Broadly speaking, culture has bound groups of people by common explicit (e.g., tangible artifacts, clothing, food, festivals, and behaviors readily observed by outsiders) and/or implicit (e.g., intangible conscious and unconscious values), assumptions, attitudes, beliefs, and feelings and has endured over time and space. Additionally, while people may be identified with one or more groups, the intensity and resilience of belongingness can vary with the interaction of self-investment and degree of group acceptance.

How people's racial–cultural assumptions shape the way in which culture is embraced, rejected, and ignored has been described in Carter's (1995) classification system. In the *Universal* paradigm, shared similarities are affirmed as universal human experiences and differences as secondary intragroup variations. In the *Ubiquitous* paradigm, socially derived circumstances constitute cultural groups. From this perspective, people embrace one or more socially derived cultural group affiliations. Unlike the Universal paradigm, the Ubiquitous paradigm acknowledges and accentuates differences by promoting cultural sensitivity to increase people's awareness of variation. Influenced by anthropology, the *Traditional* paradigm equates culture with geography (e.g., country) and adaptive socialization (e.g., birth, upbringing). As cultural membership circumscribes personality development by experiencing 'cultures' through exposure and cultural informants, culture is understood as personal circumstances superseded by the dominant culture. In the *Race-based* paradigm, culture is a function of transcending racialized experiences. Becoming aware of the effects of racism, oppression, and promoting racial identity development, culture is understood by how race and racism influence power and authority in the USA. In the fifth *Pan-National* paradigm, culture is constructed as how racial group affiliation and the distribution of power influence global issues.

Cultural Pluralism and the Tyranny of the 'Shoulds'

Historically, group relations between White Anglo and all other racial–cultural groups in the USA have influenced psychological, social, and cultural adjustment for members of dominant (e.g., mainly White Anglo-Saxon) and subordinate cultural (e.g., visible racial–cultural) groups. For example, from the mid-1700s to the early 1880s, the melting pot concept (Zangwill, 1910) assumed all entering US inhabitants would willingly relinquish indigenous heritages and adopt existing White Anglo-Saxon cultural norms. From the late 1880s to the mid-1950s, this biologically blended melting pot became '*Americanized*' (Fairchild, 1926). Through Americanization, the ideology of *cultural pluralism* or *cultural unity* advocated including visible racial–cultural groups into 'mainstream' American culture. Symbolic appreciation for diversity, especially from other civilizations beside White-European societies through congressional legislation and Supreme Court decisions in the 1960s and 1970s furthered the cultural pluralist movement. However, while White European immigrants were expected to assimilate dominant White Anglo-Saxon cultural norms, visible racial–cultural groups were still excluded and institutionally segregated.

Thus, cultural pluralism ideology was contradictory to the realities of the American 'melting pot.' While preserving cultural and social heritages as a means to creating a new society, economic, political, and educational separatism for visible racial–cultural groups was creating a 'tyranny of the shoulds' – idealized images and standards imposed upon visible racial–cultural people by the dominant society resulting in the suppression of the 'true self' (Horney, 1950). Despite conformist trends throughout the nineteenth century, voices of dissent and the 'tyranny of the shoulds' was heard in the words of W.E.B. Dubois as:

> …a peculiar sensation, this double consciousness, this sense of always looking at one's self through the eyes of others, of measuring one's soul by the tape of the world that looks on in amused contempt and pity. One ever feels his two-ness – an American, a Negro; two souls, two thoughts, two unreconciled strivings; two warring ideals in one dark body, whose dogged strength alone keeps it from being torn asunder.
>
> 1989, p. 3

At some level, cultural pluralism may have enabled visible racial–cultural people to exist in threatening environments. However, the consequences of vigilance against repressive environments also created unnecessary personal and group self-doubt, alienation, and isolation from resources and support, thus limiting a person's ability to achieve their full potential.

Cultural Pluralism and Cross-Cultural Psychology

While the expressions such as 'cross-cultural,' 'diversity,' 'multicultural,' and 'cultural pluralism' have been used interchangeably, these terms have generally referred to people who are members of visible racial–cultural groups (e.g., American Indians, Blacks, Asians, Pacific Islanders, Latinos, etc.). Recognizing the health disparities of these visible racial–cultural groups, the cross-cultural psychology movement has been focused upon improving mental health services for members of these groups. Thus, cross-cultural psychology presumed that "cross-cultural counseling problems were most likely to occur when there was a low degree of client–counselor assumed similarity…[where] the practitioner is…a member of the majority group and the client is a minority group member…[and] include[s]…counselor/therapist and client… minority individuals, but represent[ing] different minority groups ([B]lack-Hispanic, Asian-American-American Indian)" (Sue et al., p. 47). By appreciating group differences,

cross-cultural psychologists believed that understanding racial–cultural group specific knowledge (e.g., values, customs, communication patterns) would eliminate racial–cultural mental health disparities. For example, misunderstandings arising from cultural variations in values and language were suggested as antecedents to early psychotherapy terminations at rates greater than 50% as compared with 30% for White clients (Sue and Sue, 1999).

Recognizing that the majority of mental health educators and practitioners are White, professional organizations developed guidelines endorsing cross-cultural culture-specific heritage education about specific visible racial–cultural groups (e.g., Blacks, American Indians, Asian and Pacific Islanders, Latinos) in hopes that sensitizing White practitioners to issues of 'culturally different' people, practitioners would become more effective counselors (American Psychological Association, 1991; Sue et al., 1982, 1992). For example, a plethora of information exists about group-specific racial–cultural knowledge about self-expression, norms, family structures, and values for 'special populations' (e.g., restraint of emotions and hierarchical relationships for Asian American clients, extended family and religion for Latino and Black clients and harmony and cooperation among American Indians). (It should be noted that these categories have been recently expanded to include Whites, socioeconomic status, ethnicity, religion, gender, age, etc. and effects of institutional racism, classism, and sexism (Sue and Sue, 1999).)

Thus, cross-cultural educational strategies have virtually used the 'othered'-focused (Sue and Sue, 1999) or universal (Fischer et al., 1998, p. 561) perspectives when teaching about cultural issues. If cultural pluralism and cross-cultural psychology have been effective in equity and inclusion of underserved visible racial–cultural groups, why does racial–cultural separatism continue to exist?

Several reasons can account for this ongoing separatist phenomenon. Overall, cultural pluralism has not provided meaningful principles about psychological well-being. Beyond obvious axioms such as all humans have language, social systems, emotions, and values and cultural norms, identification of group understanding perceived specific racial–cultural groups by group comparisons has neglected existing within-group variation. The results have been simplistic caricatures about service delivery, cultural norms, and group homogeneity. Consequently, using a social constructivist approach, multicultural psychologists have begun developing more inclusive, affective, and intrapsychic approaches to applied mental health.

Moving from Cross-Cultural to Multicultural Psychology

Understanding people from a multicultural perspective has meant simultaneously recognizing individual and group identities. From this perspective, culture accounts for internal and external personal and social realities by becoming distinct from biological and geographical heritage. In this regard, the importance of racial–cultural self-exploration of understanding explicit content knowledge and implicit affective processes has been most illustrative in the clinical bias literature where scholars have shown mental health professionals' susceptibility to clinical bias (Oakes et al., 1994). Given issues of professional and personal bias, applied psychology and other mental health training programs have been remiss when addressing how personal assumptions influence professional development. Lacking has been a counselor-focused framework incorporating a multiculturally inclusive perspective. By challenging people's beliefs and assumptions about who they are as cultural people, mental health professionals can more realistically contextualize how social, historical, and political processes have influenced their personal and professional development.

While the enumerations of attributes for cross-cultural competencies (Sue et al., 1982) were expanded from 11 to 31 in 1992 (Sue et al., 1992), the 1998 multicultural competency guidelines (Sue et al., 1998) have taken a fundamental shift toward being more culturally inclusive by focusing upon issues of within-group variation (e.g., counselor *and* client racial–cultural identity development into counseling relationships, interventions, and processes) and proposing greater latitude in defining professional roles and responsibilities. Here, multicultural unlike cross-cultural psychology has suggested mere intellectual understanding of visible racial–cultural groups' norms does not meaningfully translate into adequate or sufficient counseling skills.

Thus, the movement from cross-cultural to multicultural psychology has involved a fundamental shift of cognitive learning about (e.g., between-group differences) visible racial–cultural groups to an integrated cognitive-affective racial–cultural identity (within-group/individual variation) orientation. Thus, clarifying ones' own racial and cultural identities, developing a sense of comfort and self-acceptance within a social, historical, and political framework, and respectfully relating and functioning with all racial–cultural groups have become essential prerequisites for developing multicultural competence (Carter, 1995; Sue et al., 1998).

Multicultural Psychology and Racial–Cultural Identity Development

People throughout time have been challenged to understand their sense of self and life experiences in unifying ways. As critical tools for cultural understanding, racial–cultural identity development models have been considered "one of the most promising approaches to the field of multicultural counseling/therapy" (Sue and Sue, 1999, p. 123). By incorporating sociopolitical contexts, racial identity theories have provided a structure of how to understand life experiences, perceptions, attitudes, and behaviors of one's own and other racial–cultural groups, identity formation and intragroup variation.

Diagnostically, racial–cultural identity models also have clinical utility by framing how individuals' experiences of counseling are influenced by their racial–cultural identity and not simply linked to presumed racial–cultural group affiliations (Carter, 1995).

Racial–cultural identity attitude theories are dynamic identity models representing developmental ego differentiation and have existed for Blacks, other visible racial–cultural people (e.g., Asian, Latino, women), and Whites since the early 1970s (Sue and Sue, 1999). Racial identity attitudes are the psychological consequences of racial socialization.

Whereas people are socialized in racialized environments, racial–cultural ego identity development is the psychological reaction to the benefits and consequences of racial–cultural social systems. Involving simultaneous perspectives of oneself in relation to members of dominant and subordinate groups, people with less developed racial–cultural ego identity statuses hold more simplistic and unexamined notions about race and race relations while greater differentiated racial–cultural ego identity statuses are based on more accurate and examined personal and group information and experiences. Thus, a person's presumed racial–cultural phenotype does not preclude racial–cultural orientation. Rather, a person's overall psychological self that has been validated, denied, or ignored influences the manner in which racial identity has been integrated into one's personality and psychological well-being. By exploring the psychological aspects of how feelings, thoughts, and behaviors people have about their own and other's race, a more realistic understanding of *who* a person rather than *what* a person is becomes possible.

Racial Identity Theory

Based upon internal racial identification levels, racial identity models have provided a framework for understanding intragroup socioracial psychological perspectives. According to Black racial identity theory (and more broadly, People of Color Identity Theory), Black people develop from (1) overvaluing and idealizing White cultural norms and internalizing negative stereotypes of Blacks (Preencounter); (2) feeling confusion as they try to understand their previously held beliefs of White superiority and Black inferiority (Encounter); (3) affirming 'Blackness' and denigrating 'Whiteness' by idealizing African cultural heritage and withdrawing from White society (Immersion); (4) accepting strengths and weaknesses of Black people and incorporating a more affirming and realistic Black identity (Emersion); (5) intellectually assessing and responding to Whites (internalization); and (6) integrating a positive Black identity while maintaining a balanced perspective of 'Whiteness' motivated by personal preferences rather than racial group self-denial (Integrative Awareness) (Carter, 1995).

In White racial identity theory, identity development involves the abandonment of racism and the development of a nonracist White identity by moving from (1) being unaware of their own racial group and ignoring the race of others (Contact); (2) becoming aware of the saliency of Whiteness, feeling guilty over personal internal standards and societal norms about race and overidentifying with Blacks (Disintegration); (3) rejecting Blacks and over idealizing Whiteness (Reintegration); (4) intellectualizing rather than emotionally understanding race (Pseudoindependence); (5) exploring biases and redefining Whiteness by trying to change Whites instead of Blacks (Immersion/Emersion); and (6) emotionally and intellectually internalizing a positive White identity through appreciation for and respect of racial differences and similarities (Autonomy) (Carter, 1995).

Overall, for visible racial–cultural people, as members of less powerful sociocultural groups, the primary socioracial identity issue in developing racial–cultural identity means overcoming internalized negative stereotypes associated with their racial–cultural group affiliation. For Whites, as members of the dominant sociocultural group, racial–cultural identity development means overcoming entitled stereotyping, learning self-value and appreciation as a White person. Therefore, in efforts to serve as racial–cultural change agents, all people must seek and promote racial–cultural affirmation – the experience of having various facets of their racial–cultural identities validated for themselves and by others. Thus, the strength of racial–cultural identity models has relied upon their dynamic nature of understanding within-group racial–cultural variation within a social, historical, and political context as well as their diagnostic value for informing counseling processes and outcomes.

Whether its sociopolitical forces versus poor psychological functioning, racial–cultural issues are silent or powerful environmental and psychological realities in a person's development. By examining how racial–cultural issues influence psychological phenomena, therapists can help clients discover how their racial–cultural ego identity forms a filter whose thoughts, feelings, and behaviors organize perceptions and interpretations of thoughts, emotions, and psychological life. To this end, racial–cultural identity models provide a framework understanding personal racial–cultural intrapsychic phenomena within and between people.

Conclusions

Since the 1970s, scholars, educators, and applied mental health professionals have become increasingly aware that for the needs of all clients to be met, multicultural psychology is a needed and essential change for the profession (American Psychological Association [APA], 1991). However, since the original cross-cultural counseling competencies position paper (Sue et al., 1982) to the latest multicultural competences (Sue et al., 1998), educators' ambivalence toward programmatic multicultural education has continued through decreasing multicultural-related courses in counselor preparation programs (Hollis and Wantz, 1994), lack of comprehensive training models (Ridley et al., 1994), and professionals who continue to perceive themselves as having limited multicultural counseling competence providing mental health services (Allison et al., 1994).

As intellectual knowing does not necessarily translate into sufficient multicultural counseling competence, multicultural psychology in the twenty-first century must mean looking beyond the cross-cultural 'tyranny of shoulds' of cultural groups' differences toward racial–cultural self-reflection. As mental health professionals' worldviews are neither neutral nor culture-free; questions posed, studies designed, and results interpreted will be inherently influenced by their racial–cultural perspectives.

As Gordon (1999) so aptly states:

> I like to remind people who cannot recognize it that I am an African-American male, born in the United States in the 20th century, and that these things influence me and what I do. They influence the way I think about things. When I identify a problem, to investigate, I bring my African-American, 20th century, male person to bear on it. That person is likely to be somewhat different from some of my colleagues who do not share my particular background and

> identity….each of us bring to our work may be different. Most important is that those perspectives may influence our respective findings as well as our interpretations–even though we may like to think of findings as being objective (p. 150).

Thus, cognitively and affectively understanding ones' own racial and cultural identities and developing a sense of comfort and self-acceptance are essential prerequisites for developing multicultural competence. Accordingly, cognitive-affective reflective learning of personal and social identities has been shown to be a powerful and effective method of developing multicultural counseling competence (Carter, 1995).

For example, an ecological person-process training framework has been successful in allowing trainees to gain multicultural counseling competence (Tomlinson-Clarke and Ota Wang, 1999). Allowing examination of *who* a person is, this model has emphasized how intra-/intergroup processes influence personal and social racial–cultural identity development. Here, people have been conceptualized within interlocking dynamic systems of: personal individual interactions (microsystem); institutions directly influencing individuals (mesosystem); institutional policies (exosystem); and the overall zeitgeist (Tomlinson-Clarke and Ota Wang, 1999).

Unfortunately, while everyone has potential to develop multicultural competence, sociocultural environments have often imposed barriers of *conformity*, *prejudice*, and *oppression*, limiting essential cognitive and racial–cultural flexibility. Moreover, to varying degrees, people have been overtly or covertly subjected to these obstructions. For example, in *conformity* – the prevalence of expected societal, institutional, or self-imposed mythical ideals (e.g., White, blonde, blue-eyed people are smarter than those who have darker complexions) have often resulted in feelings of inferiority for those who do not fit these ideals and of superiority for those who do. Beyond limiting, these mythical ideals have also prevented others from recognizing the value of diversity. In *prejudice* – those who hold the power in society and in institutions have denied equity to people who are different from them by conveying messages of 'you cannot be successful unless your phenotype, attitudes, and behaviors are similar to ours.' Because prejudice promotes group and cultural superiority, its destructive forces have artificially separated racial–cultural groups. And with *oppression*, this barrier has been socially and personally destructive due to mythical prescriptions for visible racial–cultural people to be *what they are not* while exploiting them for *being who they are* as means for keeping them from true full participation in society.

So how can social scientists nurture the development of multicultural psychology? A discipline hoping to understand, and value diversity must identify and eliminate overt and covert conformity, prejudice, and oppression in its institutions and members. Additionally, programmatic multicultural programs with goals of recognizing, respecting and learning from individual and cultural similarities and differences will combat the present day mythical social ideals and stereotypes resulting in the exploitation, coercion, and oppression of all people. While difficulties exist in eradicating xenophobia and notions of individual, group, and cultural superiority/inferiority, these obstacles will be outweighed by the ultimate rewards of all people having the opportunity to develop into their fullest personal, social, and racial–cultural potential. Thus, educators will need to continue developing training programs to incorporate cognitive/affective racial–cultural learning so counselors will be prepared to realistically work with inherent human variation. By clarifying ones' own racial and cultural identities and developing a self-acceptance, counselors will be able to respectfully and effectively work with all racial–cultural groups. Thus, the growth of multicultural psychology will depend on those who teach, research, and practice to continually develop themselves as racial–cultural people. Institutional support will provide the fertile ground necessary to do this work of moving beyond ignorance, tolerance, to meaningful understanding. Ultimately, the results will be a world where multicultural psychology is a lifestyle lending itself to an openness of understanding *who people are* rather than *what they are imagined to be*.

Disclaimer

The views expressed are those of the author. No official endorsement of the National Human Genomic Research Institute, the National Institutes of Health, or the US Department of Health & Human Services is intended or inferred.

Editor's Note

Since the original publication of Dr Ota Wang's article, the focus in the counseling and larger psychological professions on cultural competence in interventions and a truly cross-cultural psychology has grown dramatically. A now-famous essay (Henrich et al., 2010) has made the point that modern psychology is to a large extent based on "The Weirdest People in the World," where WEIRD stands for Westernized, Educated, Industrialized, Rich, and Democratic – a rather exotic minority group judged by global standards. A useful starting point for readers wishing to access the more current literature on the topic is Roy Moodley and Stephen Palmer, 2006. The most widely cited text on the topic is David Matsumoto and Linda Juang, 2013. Also helpful: J.G. Ponterotto, 2010, Gerstein, L.H. (2011), Pedersen et al. (2008), and Sue, D.W. (2007).

One intervention strategy that responds to the call for cultural competence is that of matching clients and therapists on relevant ethnic, racial, and cultural criteria. An important review essay evaluating the effectiveness of this approach is Raquel R. Cabral and Timothy B. Smith, 2011.

See also: Cross-Cultural Psychology; Social Constructionism.

Bibliography

Allison, K.W., Crawford, I., Echemendia, R.L., Knepp, D., 1994. Human diversity and professional competence. American Psychologist 49, 792–796.

American Psychological Association, 1991. Guidelines for Providers of Psychological Services to Ethnic, Linguistic, and Culturally Diverse Populations. American Psychological Association, Washington, DC.

Carter, R.T., 1995. The Influence of Race and Racial Identity in Psychotherapy: Toward a Racially Inclusive Model. Wiley, New York.

Cabral, Raquel R., Smith, Timothy B., 2011. Racial/ethnic matching of clients and therapists in mental health services: a metaanalytic review of preferences, perceptions, and outcomes. Journal of Counseling Psychology 58 (4), 537–554.
Dubois, W.E.B., 1989. The Souls of Black Folks. Bantam Classic, New York.
Fairchild, H.P., 1926. The Melting Pot Mistake. Little, Brown and Co., Boston, MA.
Fischer, A.R., Jome, L.M., Atkinson, D.R., 1998. The Counseling Psychologist 26, 525–588.
Gordon, E.W., 1999. Education and Justice: A View from the Back of the Bus. Teachers College Press, New York.
Guthrie, R.V., 1998. Even the Rat Was White: A Historical View of Psychology. Allyn and Bacon, Boston, MA.
Gerstein, L.H., 2011. Essentials of Cross-Cultural Counseling. Sage Publications, Thousand Oaks.
Henrich, J., Heine, S.J., Norenzayan, A., 2010. The Weirdest People in the World?. Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139.
Hollis, J.W, Wantz, R.A., 1994. Counselor Preparation 1993–1995, vol. II. Accelerated Development, Muncie, IN.
Horney, K., 1950. Neurosis and Human Growth: The Struggle toward Self Realization. Norton, New York.
Moodley, R., Palmer, S., 2006. Race, Culture and Psychotherapy: Critical Perspectives on Multicultural Practice. Routledge, New York.
Matsumoto, D., Juang, L., 2013. Culture and Psychology, fifth ed. Wadsworth, Belmont, CA.
Oakes, P.J., Haslam, S.A., Turner, J.C., 1994. Stereotyping and Social Reality. Oxford, UK, Blackwell.
Pope-Davis, D.B., Dings, J.G., 1995. Assessment of multicultural counseling competencies. In: Ponterotto, J.G., Casas, J.M., Suzuki, L.A., Alexander, C.M. (Eds.), Handbook of Multicultural Counseling. Sage Publications, Thousand Oaks, CA, pp. 287–311.
Ponterotto, J.G. (Ed.), 2010. Handbook of Multicultural Counseling, third ed. Sage, Thousand Oaks, CA.
Pedersen, Paul B., Draguns, Juris G., Lonner, Walter J., Trimble, Joseph E. (Eds.), 2008. Counseling across Cultures, sixth ed. Sage Publications, Thousand Oaks.
Ridley, C.R., Mendoza, D.W., Kanitz, B.E., 1994. Multicultural training: reexamination, operationalization, and integration. The Counseling Psychologist 22, 227–289.
Sue, D.W., Sue, D., 1999. Counseling the Culturally Different: Theory and Practice, third ed. J. Wiley and Sons, New York.
Sue, D.W., Arredondo, P., McDavis, R.J., 1992. Multicultural counseling competencies and standards: a call to the profession. Journal of Counseling and Development 70, 477–486.
Sue, D.W., Bernier, J.E., Durran, A., Feinberg, L., Pedersen, P., Smith, E.J., Vasquez-Nuttall, E., 1982. Position paper: cross-cultural counseling competencies. The Counseling Psychology 10, 45–52.
Sue, D.W., Carter, R.T., Casas, J.M., Fouad, N.A., Ivey, A.E., Jensen, M., LaFromboise, T., Manese, J.E., Ponterotto, J.G., Vasquez-Nuttal, E., 1998. Multicultural Counseling Competencies. Sage, Thousand Oaks, CA.
Sue, D.W., 2007. Counseling the Culturally Diverse: Theory and Practice, fifth ed. Wiley, New York.
Tomlinson-Clarke, S., Ota Wang, V., 1999. A paradigm for racial–cultural training in the development of counselor cultural competencies. In: Kiselica, M.S. (Ed.), Confronting Prejudice and Racism during Multicultural Training. American Counseling Association, Alexandria, VA, pp. 155–167.
US Bureau of the Census, 1990. Census of the Population: Supplemental Report. Race of the Population by State. US Government Printing Office, Washington, DC.
Zangwill, I., 1910. The Melting Pot. Macmillan, New York.

Counseling Psychology

Charles J Gelso, University of Maryland, College Park, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 4, pp. 2859–2862, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Counseling psychology is a specialty field within the broader discipline of psychology. The distinctiveness of the specialty is clarified in terms of its five unifying themes: focus on intact personalities; focus on human strengths; emphasis on relatively brief interventions; emphasis on person–environment interactions; and emphasis on education, career development, and environments. In terms of basic features, counseling psychology is a doctoral level field that strongly endorses the scientist-practitioner model and pays close attention to human diversity. Typical roles and functions of counseling psychologists are noted. Regarding scientific inquiry, key research issues and topics are discussed: central methodological developments; major research findings and theories in counseling and psychotherapy; multicultural factors in counseling interventions; career development and intervention; research and theory on training in both practice and science; and the fundamental scientific emphasis on healthy functioning and human effectiveness.

This article is about the science and profession of counseling psychology. Key distinguishing features of counseling psychology are described, as are the main roles and work settings of counseling psychologists. In addition, basic areas of scientific inquiry are discussed.

Counseling psychology is a specialty field within the broader discipline of psychology. This specialty is related to, but also different from, several other psychology specialties that are often characterized by the term 'applied.' The psychology specialties to which counseling psychology is most closely related are clinical, community, school, and industrial/organizational psychology. Although each of these fields seeks to study and understand human behavior empirically, the fields also share the mission of applying scientific understandings to the solution of human problems and the betterment of peoples' lives. It is in this sense that they may be termed applied specialties.

As is the case with any field or specialty, counseling psychology originated from a broad matrix of social, economic, and psychological forces. Its beginnings may be traced to the early part of the twentieth century. However, the designation of counseling psychology as a specialty within the American Psychological Association (APA) first occurred in 1952. For historical analyses, the reader is referred to Gelso and Fretz (2000).

Counseling Psychology: Distinctiveness and Diversity

In this section, the features that make counseling psychology a unique specialty within psychology, as well as the breadth and diversity of this field, are described.

The Unifying Themes of Counseling Psychology

As a way of clarifying the features that make for its distinctiveness, Gelso and Fretz (2000) proposed five unifying themes of counseling psychology. These themes, which have received substantial endorsement in the field, are briefly described below. The themes are best viewed as central tendencies, for there is also great variability around each theme.

The first unifying theme is a focus on intact, as opposed to profoundly disturbed, personalities. In earlier days, counseling psychology was thought of as a field that focused only on 'normal' problems. Today, psychologists trained in this field-work with clients at all levels of disturbance. However, in comparison to other health service fields such as clinical psychology and psychiatry, counseling psychologists continue to work more with clients closer to the normal range, and to devote more time to issues of psychological health and wellness.

The second unifying theme is the focus on assets and strengths, regardless of clients' level of disturbance. Indeed, attention to strengths has been a hallmark of this specialty, and is most evident when counseling psychologists engage in preventive and developmental interventions, as opposed to remedial ones. Preventive and developmental interventions, rather than emphasizing emotional problems that have already emerged, seek to prevent problems and enhance growth. A few of the numerous examples of such interventions are enhancement treatments for couples who do well but want to improve further, career counseling for clients who want to understand fields that fit them best, and consultation with business organizations to help create healthy work environments.

The third unifying theme is emphasis on relatively brief interventions. Counseling psychologists engage in a wide range of interventions, for example, counseling, consultation, guidance. Also, many practitioners in this specialty conduct long-term psychotherapeutic treatments. However, a hallmark of the field has always been an emphasis on brevity. The use of the term 'counseling' is a reflection of this tendency, and currently popular brief therapies fit closely with such a counseling orientation. Generally, therapy lasting less than six months is seen as brief, although most counseling psychology interventions are of much briefer duration.

Emphasis on person–environment interactions is a fourth unifying theme. From its inception, counseling psychology has sought to balance the intrapsychic and the environmental, both in understanding development and in formulating interventions. Theories that have focused largely on intrapsychic explanations, such as psychoanalysis, have not had wide appeal. Current attention to the role of culture and multicultural issues is an example of counseling psychology's commitment to incorporating the role of situation and environment in understanding personality and development, and in providing effective interventions.

The final unifying theme, perhaps more than the others, reflects counseling psychology's distinctiveness. The field's emphasis on educational and career development and environments has often been seen as its most distinctive feature. Throughout its history, counseling psychologists have theorized about and empirically investigated issues around career choice, problems, interests, development, and interventions.

Fundamental Features of the Field

There are certain features of this field that, while not necessarily distinctive, are so fundamental as to require mention. First, counseling psychology is a doctoral level field in the USA in that the large majority of US psychologists who use the title counseling psychologist, and who join the Division of Counseling Psychology of APA, hold a doctoral degree. Second, since the specialty's beginnings, it has strongly endorsed the scientist-practitioner model of training, which dictates that doctoral programs seek to train students as scientific researchers as well as practitioners of counseling psychology.

A third fundamental feature of the field is its attention to issues of culture, multiculturalism, and diversity. Although all fields within psychology have begun to pay attention to diversity in terms of race/ethnicity, sexual orientation, gender, religious beliefs, and physical disability, counseling psychology has been at the forefront in these areas for many years. Attention to issues of diversity pertains to, for example, recruitment of doctoral students and faculty, empirical research and theory, and counseling practice. The fundamental emphasis on culture and diversity in the specialty is evidenced in chapters in the most recent edition of the *Handbook of Counseling Psychology* (Brown and Lent 2000) on the role of race and social class in development; gender, sexuality, and sexual orientation; gender issues in counseling; and models of multicultural counseling.

The Roles and Functions of Counseling Psychologists

Information on where counseling psychologists work and what they do is basic to understanding this specialty. Gelso and Fretz (2000) present data on job settings indicating that over half the counseling psychologists who were members of the Division of Counseling Psychology (DCP) of APA were employed in college and university settings, either in academic departments (35%) or counseling centers (17%). A major change in recent years has been the increased appeal of independent practice, with approximately 20% of DCP members claiming this as their primary job. Furthermore, since many counseling psychologists who are in independent practice do not join the DCP, the 20% figure is likely a significant underestimate.

The remainder of counseling psychologists are distributed across a wide range of job settings, including hospitals (e.g., general, psychiatric, Veterans Administration), clinics, government agencies, business and industry, and rehabilitation facilities. Surveys over several decades indicate that these job patterns have been generally quite consistent, except for the increase in independent practice since about the mid-1980s, and a decline in the numbers of counseling psychologists in Veterans Administration centers.

Regarding job activities, the most basic finding of all surveys is that counseling psychologists are involved in an extensive range of activities. A significant number devote at least some of their work time to counseling and psychotherapy (individual, group, couples, family); consultation; psychological testing, assessment, and evaluation; research and writing; teaching, training, and supervision; and administration (Gelso and Fretz 2000). At the same time, the activities performed by the greatest numbers for the greatest amount of time are counseling and psychotherapy.

Scientific Inquiry

The major areas of theory and research in counseling psychology reflect the above discussion. Below are described key scientific issues and topics in this field.

Methodological Considerations

A fundamental quality of science in counseling psychology since the late 1970s has been its emphasis on diversity of research methods and approaches (Heppner et al. 2000). As pointed to by most reviewers, beginning in the late 1970s, there has been an increasing press to incorporate research methods that go beyond the 'received view,' for example, quantitative-experimental research.

Within the context of this methodological pluralism, two notable patterns may be seen. First, stemming from post-modern thought that is evident in all scholarly disciplines, there has been a growing interest in the use of qualitative research approaches (Morrow and Smith 2000). Such approaches use verbal or even visual rather than quantitative data, gather data by observing and listening to what people do and say, and present results in the form of verbal descriptions and theories rather than statistics.

A second pattern actually moves in the opposite direction from qualitative research. That is, there has been a major increase in the use of highly sophisticated quantitative methods. Multivariate statistical methods now appear to be as common as univariate approaches. Especially prominent are techniques that seek to permit causal inferences from correlational data, for example, structural equation modeling.

Counseling and Therapy: Theory and Research

Regarding psychological interventions, counseling psychologists study essentially the same therapeutic process and outcome questions as do other applied counseling fields, for

example, clinical psychology. After many years of outcome research, it is clear that, on the whole, therapy is effective; researchers now seek to discover what treatments are most effective when offered by what therapists for which clients experiencing what problems (Gelso and Fretz 2000).

Beyond this general statement, there are some therapy theories and topics that have been most influential in counseling psychology. Broadly speaking, cognitive-behavioral and experiential-humanistic theories of counseling have had the greatest impact (Heesacker 2000). As part of the strong cognitive-behavioral influence, theories such as Albert Ellis' rational-emotive behavior therapy, Aaron Beck's cognitive therapy, and Albert Bandura's social cognitive theory have been particularly appealing to researchers and practitioners. Psychoanalytic theories have historically been less attractive to counseling psychologists, partly because these theories have conflicted with some of the major emphases of the field (e.g., strengths and assets, brevity, environmental influences). Fundamental paradigmatic changes in psychoanalysis in recent years, however, have made some of its theories (e.g., psychoanalytic self-psychology) more appealing (Gelso and Fassinger 1992). Social psychological conceptions have also had great appeal in recent decades. Theories such as social influence theory have been creatively extrapolated to counseling and therapy, and their propositions extensively investigated (Heesacker 2000).

Multicultural Counseling and Research

As noted, counseling psychologists have closely attended to diversity for many years. Since the early 1980s, however, there has been a mushrooming of theory and research on counseling and development as related to culture and diversity. The topic that has gathered the most scientific momentum is that of race and ethnicity in psychological development and treatment. This momentum is evidenced in the form of the high rate of articles on race, ethnicity, and culture in leading journals; major handbooks on the topic (e.g., Ponterotto et al. 1995); the creation of new journals; and the publication of texts and significant chapters. Although race and ethnicity are the most frequently studied topics under the umbrella of diversity, there are others. In recent years there has been a significant surge of research on gay, lesbian, and bisexual issues in counseling and development. The study of gender, which may be placed within the diversity category, has been central in counseling psychology for nearly three decades (Heppner et al. 2000).

Career Development and Intervention

Whereas the development of theory and research on diversity topics has increased at a dramatic pace in recent decades, the topic of career development and intervention has been a mainstay of counseling psychology since its inception. Many consider this the strongest and most empirically mature area within the specialty (Heppner et al. 2000). As part of this strength and maturity, robust theories of career development and choice guided theory and practice over many years (e.g., theories by Holland, Super, and Krumboltz). There exist a number of journals, handbooks, and texts aimed at studying and elaborating career development and intervention processes. In keeping with the diversity movement, recent decades have witnessed much attention to career development in the lives of marginalized people. The study of career development and intervention in womens' lives has been particularly prominent over many years.

Training in Practice and Science

The tandem topics of counselor training and researcher training are key areas of investigation. Of the two, counselor training and supervision has had the longer life. In particular, theory and research on the process and outcome of counselor supervision have been prominent since the late 1970s. Especially notable have been developmental models of counseling supervision, wherein therapists are theorized as developing in0 certain stages, and supervision approaches need to be matched to those stages (Stoltenberg et al. 1997). Regarding research training, theory and research have examined both person and research training environment factors that contribute to graduate students' and psychologists' interest in, sense of efficacy for, and production of research and science (Gelso and Lent 2000).

Healthy Functioning and Human Effectiveness

Research and theory in this general area are natural extensions of two key unifying themes described earlier: focus on intact personalities and attention to people's strengths. Although attention to positive qualities is a deep and widely shared part of the specialty, theory and research on the topic within counseling psychology had waned for a number of years, to the point that Gelso and Fassinger (1992) viewed this area as counseling psychology's 'unfulfilled promise.' However, there has been an upsurge of scientific attention to well functioning. The study of positive constructs such as subjective well-being, career self-efficacy, and problem solving all fit this thrust. So, too, does the emerging examination by counseling psychologists of biopsychosocial factors that promote positive physical health (Hoffman and Driscoll 2000).

See also: Clinical Psychology in North America, History of; Counseling and Psychotherapy: Ethnic and Cultural Differences.

Bibliography

Brown, S.D., Lent, R.L. (Eds.), 2000. Handbook of Counseling Psychology, 3rd edn. J. Wiley, New York.

Gelso, C.J., Fassinger, R.E., 1992. Personality, development, and counseling psychology: Depth, ambivalence, and actualization. Journal of Counseling Psychology 39, 275–298.

Gelso, C.J., Fretz, B., 2000. Counseling Psychology, 2nd edn. Harcourt College Publishers, Fort Worth, TX. (1st edn. published 1992).

Gelso, C.J., Lent, R.W., 2000. Scientific training and scholarly productivity: The person, the training environment, and their interaction. In: Brown, S.D., Lent, R.L. (Eds.), Handbook of Counseling Psychology, 3rd edn. Wiley, New York, pp. 109–139.

Heesacker, M., 2000. Counseling Psychology: Theories. In: Kazdin, A.E. (Ed.), Encyclopedia of Psychology. American Psychological Association, Washington, DC.

Heppner, P., Casas, M., Carter, J., Stone, G., 2000. The maturation of counseling psychology: Multifaceted perspectives 1978–1998. In: Brown, S., Lent, R. (Eds.), Handbook of Counseling Psychology, 3rd edn. Wiley, New York, pp. 3–49.

Hoffman, M.A., Driscoll, J., 2000. Health promotion and disease prevention: A concentric biopsychosocial model of health status. In: Brown, S.D., Lent, R.L. (Eds.), Handbook of Counseling Psychology, 3rd edn. Wiley, New York, pp. 532–567.

Lyons, Heather Z., Bike, Denise H., 2013. The Use of Multiculturally-Competent Research Methods to Promote Social Justice in Counseling and Psychology. Journal for Social Action in Counseling and Psychology 5 (2), 1–9.

Morrow, S., Smith, M., 2000. Qualitative research in counseling psychology. In: Brown, S.D., Lent, R.L. (Eds.), Handbook of Counseling Psychology, 3rd edn. Wiley, New York, pp. 199–232.

Norcross, John C. (Ed.), 2011. Psychotherapy Relationships that Work: Evidence-Based Responsiveness, 2nd edn. Oxford Univesity Press, Oxford.

Ponterotto, J., Casas, J., Suzuki, L., Alexander, C. (Eds.), 1995. Handbook of Multicultural Counseling. Sage, Thousand Oaks, CA.

Ponterotto, J.G. (Ed.), 2010. Handbook of Multicultural Counseling. Sage, Thousand Oaks, CA.

Stoltenberg, C., McNeil, B., Delworth, U., 1997. IDM Supervision: An Integrated Development Model for Supervising Counselors and Therapists, 1st edn. Jossey-Bass, San Francisco.

Counterculture: 1960s and Beyond

Ralph W Larkin, The City University of New York, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article on counterculture begins with an exploration of the counterculture of the 1960s. The definition of the term 'counterculture' is examined and refined. The legacy of the counterculture of the 1960s is analyzed in light of subsequent social and cultural changes. The English revolution of the mid-seventeenth century is explored as an historical instance of a counterculture. The article concludes with a discussion of the relationship between social movements and countercultures.

In the fall of 1965, young people from the suburbs of San Francisco forsook their comfortable bedroom communities and migrated to the Haight-Ashbury district at the foot of Golden Gate Park to live lives at the edge of poverty. They panhandled for cash, wore secondhand clothes, and lived in 'crash pads' that had little more than mattresses as furniture. They were collectively known as 'hippies' (Cavan, 1972). Hippie culture was subversive of the dominant culture; if contemporary culture emphasized waste consumption, hippies lived with the bare necessities, plus, maybe an electric guitar with an amp system. When the Vietnam War escalated, hippies emphasized living in peace and cooperation. If the dominant culture emphasized the work ethic, hippies were ostensively unemployed. If sobriety and emotional constraint were the dominant norms, hippies were publicly stoned on drugs such as marijuana and lysergic acid diethylamide (LSD) and danced in the streets. If the American culture was sexually repressive, hippies engaged in a variety of tabooed sexual behaviors, including promiscuity, group sex, public intercourse, homosexual coupling, and public nudity (Yablonsky, 1968). Hippies were the precursors to what later came to be known in the English-speaking West as 'the counterculture.'

A year earlier, in the fall of 1964, across the bay at the University of California, the Berkeley Free Speech Movement (FSM) erupted (Freeman, 2004). Members of several protest groups returned to school from the Freedom Summer after registering black voters in the South in the face of white racist terrorism. When campus police confronted them and told them that they could not distribute political literature on campus property, they claimed that their rights to free speech were being abrogated. Based upon their training in nonviolent protest, they staged sit-ins and occupied the administration building; their protests spread throughout the campus. The FSM marked the beginning of the middle-class youth movement in America. The emergence of a dissident subculture in the Haight-Ashbury section of San Francisco among members of essentially the same social category a year later signaled a change in emphasis from overt politics to cultural conflict (Foss and Larkin, 1976).

What Is a Counterculture?

Before defining 'counterculture,' one must know what is meant by 'culture.' A general agreement exists among anthropologists that human culture includes the ideological realm of human behavior, including norms, values, patterns of behaviors, concepts of 'meaning,' and their embodiment in artifacts (Kroeber and Kluckhohn, 1952). Culture includes language, religion, and thought forms.

Complex societies, such as contemporary capitalist societies, contain dominant cultures and a variety of subcultures (Parsons, 1951). The dominant culture is almost always associated with the political and economic elite. Subcultures are at variance from the dominant culture on ethnic, social class, racial, age, sexual orientation, and other social differentiations. Subcultures exist within a larger multicultural context from the dominant culture, with which they share certain core values, but differ over significant points (Yinger, 1960). Most common subcultures are based upon ethnicity, such as African-American and Latino subcultures within the United States. However, there are other subcultures, such as gay 'queer' culture, and religious subcultures such as evangelicals, Jewish Hasidim, and Mormons. A counterculture, however, poses itself in *total opposition* to the dominant culture. It takes the values of the dominant culture and redefines them negatively.

The term 'counterculture' entered the sociological lexicon when social theorist Talcott Parsons (1951: p. 522) used it to describe deviant subcultures, such as delinquent gangs. Milton Yinger (1960), following Parsons, separated 'contraculture' from subcultures as a response to a frustrating environment, citing youth culture as an example. Youth culture tended to have socializing power over its members that contradicted some norms of the parent generation.

Yinger (1977) returned to the topic of counterculture in 1977. By that time, the middle-class youth movement (1964–70) was long over (Foss and Larkin, 1976). With the counterculture in retrospective view, he redefined counterculture as "a set of norms and values that sharply contradict the dominant norms and values of the society of which that group is part" (p. 833). No longer is the counterculture associated with delinquency and adolescent culture, but exists as a critique of the dominant culture.

Since the 1960s, scholars have found countercultures all over the world, within organizations, and in a variety of historical eras (Halldorf, 2010; Kolind, 2011; Morant, 2011; Viehe, 2011). Because the concept of counterculture emerged from within a social category in movement against dominant institutions, its use has been widespread without much consideration of its definition, which has been taken for

granted. Now that countercultures have been defined as a rejection of the dominant culture and exist within the context of a social movement, it must be defined what is meant by a social movement.

Foss and Larkin (1986) defined a social movement as

> the developing collective action of a significant portion of the members of a major social category, involving at some point the use of physical force or violence against members of other social categories, their possessions, or their institutionalized instrumentalities, and interfering, at least temporarily – whether by design or by unintended consequence – with the political and cultural reproduction of society. (p. 2)

All social movements have a cultural dimension; political action is by nature episodic. Typical forms of political action include elections, strikes, protests, civil disobedience, boycotts, insurrections, and so forth. These actions can occur both within and outside social movements. What differentiates social movements from isolated political actions is that movement collectivities have a shared culture that legitimizes their own behavior, identifies opposing institutions and social categories, redefining them as the enemy, and reinterprets social reality (Foss and Larkin, 1986). The movement culture provides the glue that holds the movement together. As the struggle between social movement participants and the opposition intensifies, the movement interpretation of reality tends to become increasingly radical, based upon the experience of the movement participants:

> When collective dissidence arises, the participants come to reject ever-larger portions of the hegemonic ideology. Social relations previously accepted as reasonable, inflexible, inexorable, and primordial now appear as contrived, unnatural, substantively irrational, parasitic, and senseless. [Movement participants'] own subsumption within the relations of exploitation and domination, formerly accepted as part of the natural order of things, is repudiated as unjust, arbitrarily imposed, and void of moral sanction. (Foss and Larkin, 1986: p. 73)

The reinterpretation of social reality that was inextricably tied to the middle-class youth movement formed the basis of the counterculture.

'The' Counterculture

The counterculture emerged in the mid-1960s as a self appellation among young people within the middle-class youth movement as politics merged with cultural issues. The issues of racism, collusion of higher education institutions with the military and corporate worlds in support of the Vietnam War, and *en loco parentis* regulations in colleges and universities fused with struggles over hair length, communal living, musical tastes, drug use, gender roles, and sexuality (Foss, 1972). At the height of the middle-class youth movement, there emerged a new social type, self-designated as the 'freak radical,' whose opposition to the dominant institutions was as much cultural as it was political, and whose critique of state-supported corporate capitalism was total. Political radical Abbie Hoffman (1968) under the pseudonym 'Free' offered social revolution as the meaningful alternative to the structure and ethos of the dominant society that emphasized hierarchy, bureaucracy, competition, upward mobility, the work ethic, and drugs of consciousness constriction (e.g., alcohol and barbiturates).

Countercultural radicals maintained a 'postscarcity' mentality (Bookchin, 1971) based upon their own experiences in an expanding and increasingly prosperous middle class. They perceived that society had reached a point at which the production of material goods was no longer an issue; inequitable distribution was. There was no need for humans to devote the majority of their waking hours obtaining the necessities of existence. Life was to be lived in loving communities that coalesced in a common cause of bringing down a corrupt social system. From the outside, capitalism in its contemporary form was a system that made no sense. Production had been liberated from the satisfaction of basic human needs and had become an end in itself. Humans had become harnessed to an increasingly frenetic system of production, distribution, and consumption of commodities that they had to be convinced they needed. This system of waste production for waste consumption alienated human beings from meaningful social participation, demanded participation in labor that had little redeeming social value, was ecologically destructive, required sexual repression, and drained life of its ecstatic and joyful dimension (Foss, 1972).

By the end of the summer of 1969, after the apocryphal countercultural event, the Woodstock Music Festival, the middle-class youth movement was showing clear signs of its demise (Foss and Larkin, 1976). At this point, the Women's Liberation Movement (WLM), which began as a response to the sexism of male movement participants (Morgan, 1970), was entering its most radical phase. Patriarchy was being diagnosed as the enemy of the liberation of women. The WLM gave a fresh new edge to the counterculture by emphasizing gender equality and promoting feminism. Women's values were offered as an alternative to the competitive, warlike, hierarchical, and patriarchal values of the hypermasculine dominant society. Radical feminists engaged in numerous activities designed to delegitimate patriarchy. They demanded equal pay for equal work; they disrupted the Miss America beauty pageant in 1968. They supported feminist candidates for political office. Feminists formed 'consciousness-raising groups,' analyzing sexist aspects of domestic relations and the sexual bias of the English language.

The Counterculture in the Postmovement Period

Social movements are episodic events that can last from a few weeks (e.g., Paris, France, May–June 1968) to decades (the English Revolution, 1640–60). Their consequences can be as varied as a return to the status quo *ante* or social revolution. When social movements end, the countercultures associated with them must be incorporated in some form into the dominant culture, even if the counterculture has resulted in a cultural revolution. That which was spontaneous becomes institutionalized. Oppositional elements that formerly defined

behaviors and organizations are blunted or eliminated (Foss and Larkin, 1976).

In the wake of a movement, especially a movement that ushered in a cultural revolution, such as the counterculture of the 1960s and early 1970s, movement participants were caught between lives they had rejected and a vision of an alternative life whose possibilities were being foreclosed. Foss and Larkin (1979) described this as a 'life construction crisis' for former social movement participants. In the early 1970s, there arose a plethora of groups that catered to the needs of social movement participants. These assumed the forms of authoritarian communes coalescing around a charismatic leader who controlled the lives of his followers (the Manson family, 'Spirit in Flesh' commune), mechanistic Marxist sects (Weather Underground, National Caucus of Labor Committees, October League, Revolutionary Union), Orientalist sects (Transcendental Meditation, Nichiren Shoshu, Hare Krishnas, Divine Light Mission), and Christian sects (Jesus Freaks, Children of God). In addition, a number of therapeutic cults arose promising surcease from psychological pain, including the Sullivanians on the Upper West Side of New York, and Werner Erhard's Erhard seminars training (est) therapy cult. Already existing organizations, such as Scientology and various evangelical Christian sects experienced a surge in membership consisting of former social movement participants.

In each case, the organization appropriated a fragment of the counterculture vision, such as psychological transcendence, actualizing one's potentials, or revolution, while simultaneously enforcing strict discipline over the membership's lives (Foss and Larkin, 1976, 1979). Sexual and drug abstinence were justified as sacrifices needed to further the revolution or achieve higher consciousness. As noted by several researchers (Judah, 1974; Petersen and Mauss, 1973; Robbins et al., 1975), these groups functioned as conduits from the dying counterculture back to conventional society while assuaging the pain in doing so.

The Legacy of the Counterculture

The counterculture revolutionized contemporary culture in five major ways: (1) deep suspicion of social institutions, especially institutions of the state; (2) the rise of consumer society; (3) the sexual revolution; (4) the emergence of new thought forms; and (5) the advent of the personal computer.

Since the 1960s, Americans' adherence and cohesion to major social institutions has declined quite dramatically, generating a crisis of legitimation. Lipset and Schneider (1983) traced an increase in alienation of Americans from major political, business, and labor organizations from the mid-1960s to 1981, the last year of available data for their analyses. Gallup poll data (2013) have suggested that for major American institutions, including the federal government, organized religion, medicine, news organizations, and public schools, the level of public confidence has continued to decline. The US Congress is at the bottom of the list, at 10% of respondents indicating confidence. In 1980, Ronald Reagan was elected president claiming, "Government is not the solution to our problem. Government IS the problem" (Reagan, 1981). Since declaring that the government is the enemy of the people, the Republican Party has stated that the federal government is incapable of managing anything. In some cases, such as the incompetent declaration and management of the Iraq War by the Bush Administration, it became a self-fulfilling prophecy.

One of the more salient results of the counterculture was the burial of the Protestant ethic. No longer was work ennobling; culturally, it became a means to consume. Daniel Bell (1976) opined, "When the Protestant ethic was sundered from bourgeois society, only the hedonism remained, and the capitalist system lost its transcendental ethic" (p. 21). Bell saw the demise of the Protestant ethic as a consequence of the emergence of a system of mass production and mass consumption, instant credit, and the creation of new wants and needs. Certainly, these provided the material undergirding of the cultural revolution. Once the movement was over and the countercultural critique of contemporary capitalism as waste production and consumption had been effaced, the engine of consumer capitalism could be turned on full speed. As long as it increased sales, merchandisers could embrace the countercultural ethic, 'If it feels good, do it,' which was literally turned into a sales slogan, complete with the Nike swoosh. The hedonistic impulses released by the counterculture had lost their oppositional content and were indulged in for their own pleasures, generating the age of narcissism (Lasch, 1978); the 1970s was labeled the 'me decade' (Wolfe, 1976). Americans could indulge in mass therapy, sex, drugs, higher states of consciousness, and rampant commodity consumption without guilt, but always for a price.

Perhaps the greatest impact of the counterculture was on sex roles. Beginning with hippies' gender bending, the counterculture problematized gender roles, especially male gender role expectations. The countercultural emphasis on subjectivity and feelings as new sensibilities, the rebellion of women against sexism, and the gay liberation movement undermined the patriarchy and paternalism that permeated the dominant culture (Freeman, 1975).

In the 1970s and 1980s, glass ceilings were shattered as women advanced in academia, medicine, law, science, politics, and business. However, role expectations for males also escalated. No longer was it sufficient for males to 'bring home the bacon,' but they also had to be loving and supportive members of their families, which for many men was beyond their emotional capabilities (Faludi, 1999; Kimmel, 1996). The rising expectations of women for male role performance, especially in the family, such as sharing household chores, created male anxiety, especially sexual anxiety. To exacerbate male anxiety, economic changes were eliminating working-class jobs, especially those that produced enough income so that the family could depend on a single wage earner. Simultaneously, jobs in the office, which could be held by either males or females, were increasing. If masculinity was defined as earning enough money to support one's family, throughout the late 1970s to the present, fewer and fewer men were able to meet the criterion. In order to make ends meet, wives had to take up jobs in the labor force. This gave them an independent source of income and moved them out of the house, destabilizing male hegemony in the home. Many men bolted, abandoning their families (Ehrenreich, 1983); others were thrown out (Faludi, 1999). The counterculture

influenced the demise of the nuclear family and the rise of gender wars.

In response to feminism and the emergence of a vigorous gay community demanding civil rights for itself, the cultural and political right has reacted by sponsoring cultures of violence (Gibson, 1994). The elections of Ronald Reagan and George W. Bush created cowboy chic fads, with copious pictures of the presidents on their respective ranches. Both political regimes attacked women's health care rights, were antigay, and attempted to resurrect an outmoded image of masculinity that emphasized emotional restraint, outdoorsmanship, familiarity with guns, military glory, and control of violence. The elections of Democrats Bill Clinton and Barack Obama, both with countercultural credentials, resulted in massive purchases of guns and threats of assassination.

Beginning in the 1980s and rising throughout the 1990s to the present, one of the major indicators of the crisis of masculinity has been the emergence of bullying as a serious social issue. Although the literature is unclear as to whether there has been an increase in bullying over time, what has become clear is that much bullying in school is done by more popular students against their less popular peers (Dijkstra et al., 2008; Holt and Espelage, 2007; Pellegrini et al., 2010). Much high school bullying is done by members of the football team and has the tacit approval of administration (Larkin, 2007; Lefkowitz, 1997).

Rampage shootings by marginalized males have escalated throughout the 1980s to the present. They began in schools and workplaces, subsequently spreading to a myriad of public venues, including colleges, shopping malls, movie theaters, government buildings, military bases, and churches (Ames, 2005; Kellner, 2008). An overwhelming number of rampage shooters were bullied, harassed, and intimidated in school or on the job. Rampage shootings were often revenge for the humiliation and emasculation experienced by the shooter (Kimmel and Mahler, 2003).

Between 1993 and 1996, United States experienced a minor social movement that this author identified as "The Revolt of the Angry White Male" (Larkin, 2011), which began as a response against the assault ban contained in the Brady Bill, passed by Congress during the early days of the Clinton administration. This particular period saw a rise in right-wing hate groups, neo-Nazis, skinheads, weekend warriors, paintball war games, paramilitary groups, 'unorganized' militias, and right-wing terrorist groups. The constituency for this movement was marginal males; their prime target was the federal government, which was seen as protecting 'feminazis,' gays, Jews, and people of color (referred to within the movement as 'mud people'). The movement began with a shootout between the Bureau of Alcohol, Tobacco, and Firearms, the Federal Bureau of Investigation, and a white racist named Randy Weaver in Ruby Ridge, Idaho. This was followed by the siege of the Branch Davidian compound in Waco, Texas, in 1993. The siege was revenged by Timothy McVeigh and Terry Nichols in the bombing of the federal building in Oklahoma City. The movement ended with capture of the Montana Freemen in June 1996.

Meantime, the crisis in masculinity generated by the counterculture continues apace. The political and religious right attempt to limit women's access to abortions, attack Planned Parenthood women's health facilities, assassinate abortion-providing doctors, and attempt to restrict access to contraception. The war against homosexuality has been much less successful, although several Southern states have attempted to pass and implement antisodomy laws. Numerous states have been passing gay marriage laws since the Supreme Court ruled that married same-sex couples are entitled the same federal benefits as heterosexual couples and let stand a California State Supreme Court decision allowing gay marriages in that state.

As noted throughout this article, one of the most prominent ideological features of the counterculture was the denigration of functionally rational thought forms, and the substitution of subjectivism as the most valuable way of knowing (Foss, 1972). The rise of subjectivism in a significant portion of American youth had several consequences. First, it opened up American society to the inclusion of non-Western thought forms, including Hindu spiritualism and Buddhist notions of reality, both of which denigrated the material world. Second, it ushered in notions of ecological or holistic thinking, as opposed to the linearity of scientific thought. Critiques of functional rationalism could be found in the works of Phillip Slater (1975) and Theodore Rosack (1973). TenHouten and Kaplan (1973) developed a synthesis of 'propositional' (scientific) and 'appositional' or 'synthetic' modes of thought, suggesting their lateralization in the brain (right brain versus left brain modes of thinking). They also suggested that human knowledge of a higher level of rationality includes both modalities.

The counterculture spread beyond the confines of middle-class youth to practically every sector of society. In the social sciences, young radicals attacked the dominant paradigm of structural functionalism as inherently conservative and unable to properly conceptualize social conflict (Foss, 1963; Mills, 1959). Marxism and neo-Marxism, especially that of the Frankfurt school, which emphasized cultural issues, challenged the structural functionalism that dominated mid-century social sciences. The whole idea of grand theories of how societies work was challenged by the rise of the school of ethnomethodology, which appealed to the subjectivism of young radicals. Ethnomethodology was the study of the methods by which human beings attempted to make meaning from their environment (Mullins, 1973). The rise of competing schools to structural functionalism in the social sciences has led to what Kuhn (1962) referred to as a 'paradigm crisis,' in which several paradigms contest for dominance in the field, but no consensus exists.

In psychology, the subjectivism of countercultural young psychologists gave rise to humanistic psychology that, borrowing from the work of Abraham Maslow (1954, 1962), emphasized the positive aspects of psychological development, rather than focusing on pathology. The rise of humanistic psychology within the profession dramatically expanded the field, giving rise to the 'Human Potentials Movement,' encouraging psychotherapists to focus on psychological 'growth,' rather than deficits. In the late 1960s and early 1970s, there was an explosion of the psychological marketplace as therapists offered group therapy, marathon weekends, sex therapy, and retreats, for example. They appealed to an affluent middle-class clientele that wanted to actualize their inner potential as human beings.

Within academic psychology, the dominant paradigm of behaviorism was overthrown by cognitive psychology (Berlyne, 1975; Greenwood, 1999). In line with the subjectivism of the counterculture, newly emerging psychologists were not going to treat the cognitive functions of human beings as black boxes and only look at operant conditioning and reinforcement schedules in the mode of B.F. Skinner (1938, 1953).

Within the profession, many professionals began to question the 'dark' side of psychology, that is, its role as a mechanism of social control. Chief among these critics were R.D. Laing (1999) and Thomas Szasz (1961, 1965). Laing discussed schizophrenia as a response to a schizophrenogenic, or an insane, environment. The implication of his theory was to treat the environment rather than the patient, whose responses were rational, given the context. Szasz, on the other hand, viewed psychiatry, especially as it intersected the legal realm, as a process of legitimation of social control over people who are otherwise seen as ungovernable. The anti-psychiatry movement of which Laing and Szasz were part, led to the rise of the Mental Patients Liberation Front (Zito, 2005), in particular, and the disability rights movement in general. These movements led to the passage of the Education for All Handicapped Children (PL 94-142) in 1975, and the Americans with Disabilities Act of 1990.

Literally, no academic discipline was untouched by the countercultural revolution. Conventional economics was disrupted by the intrusion of Marxist-oriented political economists. In 1968, the Union of Radical Political Economists was founded at Harvard University and the University of Michigan (Union of Radical Political Economists, n.d.), rising in popularity through the 1970s, and tapering off after that.

Literature studies were influenced by postmodernism, a direct descendant of countercultural sensibilities. Ethnic studies have emerged on major college campuses, including black studies, and Latino studies, in which minority students have been studying the history, sociology, psychology, and literature of their own diaspora. Most importantly, every major field has developed a feminist component. In addition to women's studies on college campuses, there are feminist sociologists, economists, psychologists, anthropologists, and political scientists. Women have increased their enrollments in the hard sciences as well.

Although the counterculture is generally characterized as antitechnological, young radicals, Vietnam War resisters, and technohippies were heavily involved in developing the personal computer in and around the campus of Stanford University (Markoff, 2008). One instrumental group was called the Homebrew Computer Club, established in the 1970s by war resistor, Fred Moore, who thought he could design a personal computer that could make him organize antiwar protests more efficiently.

The more academic institutes were directed at creating artificial intelligence. One was organized by John McCarthy, a card-carrying member of the Communist Party. The other was operated by Doug Englebart; both were heavily involved in countercultural activities. Markoff noted that often times when walking through the labs, one could smell marijuana being smoked. Both Steve Jobs and Steve Wozniak, founders of Apple Computer, were heavily involved in the counterculture. All these people were advocating open sourcing, which expressed the countercultural notion of democratizing technology and wresting it away from large bureaucratic state and corporate institutions.

Countercultures in History

Human societies have experienced cultural revolutions from time immemorial. Exodus in the Bible is a mythological documentation of a cultural revolution whereby Israelites shed their slave mentalities and their worship of graven images. They become Jews, a monotheistic nation worthy of self-governance and management of their land in harmony with nature. However, the counterculture of the 1960s awakened historical scholars' interest in countercultures and cultural revolutions. Therefore, the study of countercultures is relatively new.

Perhaps the best documented historical counterculture is that which existed in England during the revolution between 1640 and 1660. The two decades of the revolution began with the meeting of the Long Parliament, followed by the interregnum between the beheading of Charles I in 1649, and ended with the restoration of Charles II (Hill, 1961). According to historian Christopher Hill (1972):

> There were … two revolutions of mid-seventeenth-century England. The one which succeeded, established the sacred rights of property …, gave political power to the propertied …, and removed all impediments to the triumph of the ideology of the men of property – the Protestant ethic. There was, however, another revolution, which never happened, though, from time to time it threatened. This might have established communal property, a far wider democracy in political and legal institutions, might have disestablished the state church and rejected the Protestant ethic. (p. 15)

The successful cultural revolution championed those values associated with the merchant class and what Max Weber (1958) referred to as 'worldly asceticism': work as the glorification of God, frugality, rationality, and emotional restraint. This value system became predominant during the Industrial Revolution and was important in the transformation of an agrarian population into an industrial workforce. It also gave rise to a new social type: the rational calculator, for whom time is money, capital accumulation is evidence of God's favor, and all decisions are based upon a cost-benefit analysis.

The second cultural revolution was from below and championed an agrarian communism (Hill, 1972). It also emerged from Protestantism, but it was a much different Protestantism than that espoused by the merchant class. It expressed the notion that if all men are created equal, then why should some men have dominion over others? It was the ideology of the 'nonconformist sects' such as Quakers, and secular rebels, such as Ranters, Lollards, Diggers, Levelers, and other dissident groups that demanded a radical egalitarianism. The radicals often met in ale houses were blasphemous, smoked tobacco as a mind altering drug, and offered a vision of an egalitarian society that was described as 'the world turned upside down.'

Conclusion

'The' counterculture of the 1960s and early 1970s emerged out of the middle-class youth movement, which alternated between periods of political and cultural intensification (Foss and Larkin, 1976). The counterculture was self-defined in opposition to the dominant culture, and framed itself as a culture of life opposing the dominant culture of death. Following the social movement period, numerous aspects of the counterculture were divested of their oppositional content and institutionalized into the dominant society. This created a cultural transformation. It opened up employment opportunities for women and minorities, incorporated new thought forms and technologies, transformed academic disciplines, and led to the development of personal computers.

Countercultures are generated by social movements at a point in time when the old culture becomes an impediment to change. The emergence of Protestantism and the Protestant ethic during the English revolution and throughout Europe in the seventeenth century provided legitimacy for a newly emerging capitalist class that took control over a society formerly run by an agrarian aristocracy. Similarly, in the mid-twentieth century, in a society based on high levels of consumption and production, the Protestant ethic no longer made sense. Although resisted by dominant elites when it was a countercultural value, once the Protestant ethic was dead, capital danced its grave, belatedly understanding that the new sensibility could be exploited in a consumer-based economy. However, the counterculture has been responsible for the generation of 'culture wars' in its wake. The use of psychedelics, marijuana, and hard drugs were rigorously opposed in the dominant culture, resulting in the war on drugs. Liberalized views of sexuality were incorporated into advertising copy and used to sell products. Yet, the fight for women's access to abortion and female health care has been a flash point for wedge issues since the 1970s. Similarly, the crisis of masculinity created by the counterculture has been resisted by the cultural right that has denigrated rape, demanded access to firearms for all, sponsored and lionized violent male subcultures, and has resisted the inclusion of gays into mainstream society.

See also: Alienation, Sociology of; Alienation: Psychosociological Tradition; Asceticism; Attempted Utopias and Intentional Communities; Capitalism: Global; Civil Rights Movements; Cohabitation: United States; Collective Behavior, Sociology of; Collective Beliefs, Sociological Explanation of; Communitarianism, Sociology of; Community, Social Context of: The US Case; Counterculture: The Classical View; Critical Theory: Contemporary; Culture, Sociology of; Disability: Sociological Aspects; Ethnic, Racial, and Nationalist Movements; Family Theory: Competing Perspectives in Social Demography; Fatherhood; Heterosexuality; Identity Movements; Masculinity, Scripts of; Modernization, Sociological Theories of; Movements, Social; Phenomenology in Sociology; Postmodernism in Sociology; Repartnering and Stepchildren; Sexuality: Cultural Aspects; Social Changes: models; Sociological Theory; Subculture, Sociology of; Youth Culture, Sociology of; Youth Movements.

Bibliography

Ames, M., 2005. Going Postal: Rage, Murder, and Rebellion from Reagan's Workplaces to Clinton's Columbine and Beyond. Soft Skull, New York.
Bell, D., 1976. The Cultural Contradictions of Capitalism. Basic Books, New York.
Berlyne, D.E., 1975. Behaviourism? Cognitive theory? Humanistic psychology? To Hull with them all. Canadian Psychological Review/Psychologie Canadienne 16, 69–80.
Bookchin, M., 1971. Post Scarcity Anarchism. Ramparts Press, Berkeley, CA.
Cavan, S., 1972. Hippies of the Haight. New Critics Press, St. Louis, MO.
Dijkstra, J.K., Lindenberg, S., Veenstra, R., 2008. Beyond the class norm: bullying behavior of popular adolescents and its relation to peer acceptance and rejection. Journal of Abnormal Child Psychology 36, 1289–1299.
Ehrenreich, B., 1983. Hearts of Men: American Dreams and the Flight from Commitment. Doubleday Anchor, New York.
Faludi, S., 1999. Stiffed: The Betrayal of the American Man. William Morrow, New York.
Foss, D.A., 1963. The worldview of Talcott Parsons. In: Stein, M., Vidich, A. (Eds.), Sociology on Trial. Prentice-Hall, Englewood Cliffs, NJ.
Foss, D.A., 1972. Freak Culture: Lifestyle and Politics. E.P. Dutton & Co., Inc., New York.
Foss, D.A., Larkin, R.W., 1976. From 'The Gates of Eden' to 'Day of the Locust': an analysis of the dissident youth movement of the 1960s and its heirs in the 1970s – the post-movement groups. Theory and Society 3, 45–64.
Foss, D.A., Larkin, R.W., 1979. The roar of the lemming: youth, postmovement groups, and the life construction crisis. Sociological Inquiry 49, 264–285.
Foss, D.A., Larkin, R.W., 1986. Beyond Revolution: A New Theory of Social Movements. Bergen & Garvey, South Hadley, MA.
Freeman, J., 1975. The Politics of Women's Liberation. Longman, New York.
Freeman, J., 2004. The Berkeley free speech movement. In: Sharpe, M.E. (Ed.), Encyclopedia of American Social Movements. Emmanuel Press, Armonk, NY.
Gallup Inc., 2013. Confidence in Institutions.
Gibson, J.W., 1994. Warrior Dreams. Hill & Wang, New York.
Greenwood, J.D., 1999. Understanding the "cognitive revolution" in psychology. Journal of the History of the Behavioral Sciences 35, 1–22.
Halldorf, J., 2010. Lewi Pethrus and the creation of a Christian counterculture. Pneuma: The Journal of the Society for Pentecostal Studies 32, 354–368.
Hill, C., 1961. The Century of Revolution: 1603–1714. Norton, New York.
Hill, C., 1972. The World Turned Upside Down: Radical Ideas during the English Revolution. Viking Press, New York.
Hoffman, A., 1968. Revolution for the Hell of It. Dial Press, New York.
Holt, M.K., Espelage, D.L., 2007. Perceived social support among bullies, victims, and bully victims. Journal of Youth and Adolescence 36, 984–994.
Judah, J.S., 1974. Hare Krishna and the Counterculture. Wiley, New York.
Kellner, D., 2008. Guys and Guns Amok. Paradigm Publishers, Boulder, CO.
Kimmel, M.S., 1996. Manhood in America. The Free Press, New York.
Kimmel, M.S., Mahler, M., 2003. Adolescent masculinity, homophobia, and violence: random school shootings, 1982–2001. American Behavioral Scientist 46, 1439–1458.
Kolind, T., 2011. Young people, drinking and social class. Mainstream and counterculture in the everyday practice of Danish adolescents. Journal of Youth Studies 14, 295–314.
Kroeber, A.L., Kluckhohn, C., 1952. Culture: A Critical Review of Concepts and Definitions. Harvard University Peabody Museum of American Archeology and Ethnology, Cambridge, MA.
Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Laing, R.D., 1999. The Divided Self: An Existential Study in Sanity and Madness. Routledge, London; New York.
Larkin, R.W., 2007. Comprehending Columbine. Temple University Press, Philadelphia, PA.
Larkin, R.W., 2011. Masculinity, school shooters, and the control of violence. In: Heitmeyer, W., Haupt, H.-G., Malthaner, S., Kirschner, A. (Eds.), Control of Violence: Historical and International Perspectives on Violence in Modern Societies. Springer, New York.

Lasch, C., 1978. The Culture of Narcissism: American Life in an Age of Diminishing Expectations. Norton, New York.
Lefkowitz, B., 1997. Our Guys. Vintage, New York.
Lipset, S.M., Schneider, W., 1983. The decline of confidence in American political institutions. Political Science Quarterly 98, 379–402.
Markoff, J., 2008. What the Dormouse Said: How the 60s Counterculture Shaped the Personal Computer Industry. Penguin, New York.
Maslow, A., 1954. Motivation and Personality. Harper & Row, New York.
Maslow, A., 1962. Toward a Psychology of Being. Van Nostrand, New York.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, New York.
Morant, K.M., 2011. Language in action: funk music as the critical voice of a post – civil rights movement counterculture. Journal of Black Studies 42, 71–82.
Morgan, R., 1970. Sisterhood Is Powerful. Vintage, New York.
Mullins, N.C., 1973. Theories and Theory Groups in Contemporary American Sociology. Harper & Row, New York.
Parsons, T., 1951. The Social System. The Free Press, New York.
Pellegrini, A.D., Long, J.D., Solberg, D., Roseth, C., Dupuis, D., Bohn, C., Hickey, M., Jimerson, S.R., Swearer, S.M., Espelage, D.L., 2010. Bullying and social status during school transitions. In: Jimerson, S.R., Swearer, S.M., Espelage, D.L. (Eds.), Handbook of Bullying in Schools: An International Perspective. Routledge/Taylor & Francis Group, New York, NY.
Petersen, D., Mauss, A., 1973. The cross and the commune: an interpretation of the Jesus people. In: Glock, C. (Ed.), Religion in Sociological Perspective. Wadsworth, Belmont, CA.
Reagan, R., 1981. Government Is Not the Solution…. Quote DB.
Robbins, T., Anthony, D., Curtis, T., 1975. Youth culture religious movements: evaluating the integrative hypothesis. Sociological Quarterly 16, 48–64.
Roszak, T., 1973. Where the Wasteland Ends: Politics and Transcendence in Postindustrial Society. Doubleday & Company, Inc., Garden City, NY.
Skinner, B.F., 1938. The Behavior of Organisms. Appleton-Century-Crofts, New York.
Skinner, B.F., 1953. Science and Human Behavior. Macmillan, New York.
Slater, P., 1975. Earthwalk. Anchor Books, New York.
Szasz, T.S., 1961. The Myth of Mental Illness: Foundations of a Theory of Personal Conduct. Paul B. Hoeber, New York.
Szasz, T.S., 1965. Psychiatric Justice. Collier Books, New York.
Tenhouten, W.D., Kaplan, C.D., 1973. Science and Its Mirror Image: A Theory of Inquiry. Harper & Row, New York.
Union of Radical Political Economists, n.d. General History of URPE.
Viehe, F.W., 2011. The G'hals of New York and other cities at work and play: a flourishing feminine counterculture in mid-nineteenth century urban America. International Journal of Interdisciplinary Social Sciences 6, 1–20.
Weber, M., 1958. The Protestant Ethnic and the Spirit of Capitalism. Scribners, New York.
Wolfe, T., 1976. The Me Decade. New York Magazine, New York.
Yablonsky, L., 1968. The Hippie Trip. Pegasus, New York.
Yinger, J.M., 1960. Contraculture and subculture. American Sociological Review 25, 625–635.
Yinger, J.M., 1977. Presidential address: countercultures and social change. American Sociological Review 42, 833–853.
Zito, A., 2005. A Chronology of the Disability Rights Movements, 1970–1979.

Relevant Websites

Songs That Represent the Counterculture Ethos of the 1960s

http://www.youtube.com/watch?v=9Ha1eqHDu_w – "Ballad of a Thin Man," by Bob Dylan, 1965. Dylan was considered the poet laureate and visionary of the counterculture. The song satirizes Mr. Jones, a straight person who is trying to understand the counterculture but who cannot make sense of it using conventional means. The message is that the alternate reality can only be subjectively experienced.

http://www.youtube.com/watch?v=P-Q9D4dcYng – "A Day in the Life," by the Beatles, 1967. Reminiscent of Bob Dylan's "Ballad of a Thin Man," it satirizes a man who 'made it' but did not realize 'that the light had changed.' It pokes fun at the competitive struggle and bureaucratic functioning. It takes Dylan a step further by celebrating the psychedelic experience both in musical texture and in lyrics when the Beatles tell the listener that they "would like to turn you on," as the music crescendos to a psychedelic peak.

http://www.youtube.com/watch?v=yRhq-yO1KN8 – "Imagine," by John Lennon, 1971. Written after the Beatles had split up and had declared the end of the movement, John Lennon wrote this tribute to the vision of the counterculture.

http://www.youtube.com/watch?v=26bZZwPs5zM – "Let It Be," by John Lennon and Paul McCartney, 1970. This particular sad song announced both the end of the middle-class youth movement and the breakup of the Beatles. The Beatles, who were seen by many as the troubadours of the movement, told their followers that there was nothing that could be done, and to "let it be."

http://www.youtube.com/watch?v=5-0bqonqooo – "Light My Fire," by the Doors, 1967. The name of the band came from Aldus Huxley's book, *The Doors of Perception* (Harper & Row, New York, 1954), in which he described his experiences on mescaline. "Light My Fire" is the ultimate combination of sex and drug ecstasies, as it crescendos to an orgasmic conclusion, much in the way of Maurice Ravel's "Bolero."

http://vimeo.com/11205275 – "The Times They Are a Changin," by Bob Dylan, 1964. This song literally became the anthem of the youth movement during the Civil Rights phase, combining the issues of racism and social protest.

http://www.youtube.com/watch?v=WANNqr-vcx0 – "White Rabbit," by the San Francisco band, Jefferson Airplane, 1967. Using the metaphor of Alice in Wonderland, ostensibly a children's story, but in reality, a satire of Victorian British politics, the Airplane crafted what came to be an advertisement for psychedelic drugs.

Counterculture: The Classical View

Sheila Whiteley, University of Salford, Greater Manchester, UK; Bader International Study Centre, Queen's University, Kingston, ON, Canada; and Hailsham, Sussex, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by B.M. Berger, volume 4, pp. 2862–2864, © 2001, Elsevier Ltd.

Abstract

Counterculture is a sociopolitical term indicating a point of dissent between dominant or mainstream ideologies and alternative value systems, so creating a collective voice that can be considered a significant minority. This article explores the origins of the 1960s, counterculture in the 1950s Beats, and the emergence of youth cultures. With music recognized as integral to hippie culture, protest, and environmentalism, attention is focused on key artists, global communication, and festivals. Liminal aspects of the counterculture are examined, and the continuing relevance of the term as symbolizing the hegemonic struggles that continue to inform everyday life in societies and cultures worldwide.

While the origins of the term 'counterculture' are unclear, as a key sociopolitical term it is generally accepted as indicating a point of dissent between dominant or mainstream ideologies and alternative value systems that are voiced through, for example, music, art, writing, protest, and lifestyles, so creating a collective voice that can be considered a significant minority. With its origins in the 1950s Beats, the term became an integrative label in the 1960s, where it was interpreted by theorists such as Nuttall (1970), Roszak (1970), and Marcuse (2002) as a 'youth' culture.

Beats

The Beats of the 1940s and 1950s are identified as "the catalysts who precipitated the more widespread social rebellion of the sixties and seventies … kindred spirits determined to practice absolute personal freedom within a society governed by stifling conservative attitudes" (Morgan, 2010: p. 247). Originating in postwar Paris and influenced by the existentionalist values of Jean-Paul Sartre, nihilism, and the futility of action, they were also attracted to Eastern mysticism, poetry, jazz, drugs (marijuana), and literature. In the United States, the influence of Jack Kerouac, dubbed by Turner as "the James Dean of the typewriter" (1996: p. 13) "permeated and shaped an emerging youth culture, not least the themes of its rock culture (from the Beatles, Dylan and the Velvet Underground in the 1960s, to David Bowie, Tom Waits and Patti Smith in the 1970s)." The legacy of Kerouac's "key works – from *On the Road* (1957) to the *Dharma Bums* (1958), *Lonesome Traveller* (1960) to *Desolation Angels* (1965) – are transience, impermanence and the prospect of fresh excitements…" and his theme of "the road is sustained in the machismo legend that surrounds rock as a live medium" (Warner, 2013: p. 30) and the associations of Route 66. Beat poet Allan Ginsberg's legacy lies in his reputation as a campaigner for homosexuality and a libertine whose "signature poem 'Howl!' has been recognized both as 'a premature gay *cri de coeur*' and as a precursor to the 1960s sexual revolution" (Warner, 2013: p. 32). As Warner writes:

> There are various other Beat threads, too, that would feed into the history that follows – from the Buddhist affiliations of Kerouac, Ginsberg and Gary Snyder; interests in the ecology exemplified by Snyder and Michael McClure; the DIY enterprise of publishers of books and journals, characterized most prominently by Lawrence Ferlinghetti and City Lights, but also in the many poetry magazines produced by Leroi Jones, Hettie Jones, Diane di Prima and others, representing entrepreneurial action outside the commercial mainstream; and the black radicalism of Leroi Jones, assuming the name Amiri Barak in 1965, and essentially abandoning his Beat associations for the African-American struggle. (Warner, 2013: p. 33)

In the United Kingdom, the propeace, antiwar refrains of American Beat style resonated with the Campaign for Nuclear Disarmament (CND), which was led by philosopher Bertrand Russell, Canon John Collins, the Labor Party's Michael Foot, novelist J.B. Priestly, and emerging musicians, artists, writers, and poets. Jazz trumpeter and writer Jeff Nuttall provides a perceptive reflection on the CND campaign and the rise of the UK underground (a term used in harness with counterculture) in his book, *Bomb Culture* (1970).

The dialogue between Beat and Counterculture is associated with "Bob Dylan's meeting with Allen Ginsberg at the close of 1963" (Wilentz, 2010: p. 69) and was 'followed by the singer's first meeting with the Beatles in the summer of 1964.' Dylan then brought Ginsberg and the Beatles together in May 1965, "inviting them all to join him in his hotel suite after an Albert Hall gig in London" (Miles, 1990: pp. 369–370). "Thus, within a year and a half, the leading players in an increasingly influential generation of popular musicians would cross existing artistic boundaries to make positive contact with the most dynamic and high profile member of the Beat fraternity" (Warner, 2013: p. 238).

The Emergence of Youth Cultures

Of equal importance to the rise of the counterculture was the identification of the teenager as a precursor to 1960s youth culture. This previously unnamed stage (postchildhood, preadolescent), identified and discussed by sociologists and newspapers alike, was triggered by a new emphasis on attitude – attributable in part to the 1950s advent of rock 'n' roll, coffee bars, the jukebox, and such role models as Elvis Presley and

Buddy Holly, and American movies cultivating images of adolescent life from *Rebel Without A Cause* (1955) to *Rock Around The Clock* (1956). Between 1946 and 1951, a record 22 million babies were born in the United States, giving rise to the phrase baby boomers. Born into an affluent economy, and a growing media technology, they were to be an important part of the emergent youth culture of the 1960s.

The United Kingdom, in the 1950s and 1960s was a society in transition and the emerging teenage culture was, in part, a response to this situation. While there was less spending power for British teenagers than their US counterparts, both experienced an overall rise in the standard of living, which was accompanied by a trend toward personal consumption – fashion, records and record players, cinema and other entertainment, including coffee bar culture. What is significant, in this context, are the specifics of consumption, not least lifestyle and leisure, and by the mid-1960s it seemed that the paralyzing grayness that had characterized British society during the post-war period was replaced, almost overnight, by Carnaby Street color, an underground youth culture, and the Beatles. The identification of underground/counterculture with British youth rests more on cultural politics – arts labs, local underground magazines, legalize pot rallies, free pop concerts, and psychedelic 'head' shops. UFO was headlined as London's answer to Haight Ashbury; the Electric Garden, later renamed Middle Earth, opened in Covent Garden. While there were student protests against the American Embassy in Grosvenor Square (1968) at which Mick Jagger was present, and which led to his personal response in 'Street Fighting Man,' sit-ins at various art colleges, and a continuing protest against war (CND, and such slogans as 'Make Love Not War'), as my colleague, historian Steven Fielding observes, despite the visibility of Tariq Ali and the VSC, the motives of most student protesters were closer to the Beatles sentiments than those of Mick Jagger. Their 1968 song, 'Revolution' spelt it out: 'we all want to change the world. But when you talk about destruction, don't you know that you can count me out'.

1960s Counterculture

Roszak identified "the young (qualified as perhaps only a minority of the university campus population) who find themselves cast as the only effective radical opposition within their societies" (1970: p. 2). He described them as 'technocracy's children' – disaffected middle-class youth who wished to break away from the bourgeois and technocratic world of their parents, and the increasing reliance upon technology and rational-scientific reasoning.

> By way of a dialectic Marx could never have imagined, technocratic America produces a potentially revolutionary element among its own youth. The bourgeoisie, instead of discovering the class enemy in its factories, finds it across the breakfast table in the person of its own pampered children. (1969: p. 34)

Nuttall (1970) also draws attention to the escalating Cold War and the constant threat of a nuclear holocaust. As Bennett notes, 'During the late 1960s, the atrocities of Nazi Germany and the horrific realization of the destructive power of the atomic bomb were disturbing facets of a recent past, while the Cold War and the escalating conflict in Vietnam served as current reminders of the highly pathological aspects of the technocratic society (see Bennett, 2005).

The 1960s counterculture has been identified as a generational youth culture that challenged both the potential disasters inherent in an advancing technology and the traditional concepts of career, education, and morality; and a cultural phenomenon with the potential to create a new cultural sphere beyond and ideologically separated from the parent culture. Evidenced in the United States both in New Left Politics (Students for a Democratic Society) and the Hippies, there is the suggestion of an initial duality – with the former embracing political activism and a concern for health, education, and the environment, and the latter a more bohemian lifestyle and cultural politics. In October 1967, while "flower power wilted, a new movement was born… This was the Youth International Party (Yippies). It was not to hit the headlines until August 1968" (Neville, 1971: p. 30) when their revolutionary strategy moved toward an attack on the establishment through aesthetic politics – mocking militarism, exorcising the Pentagon, throwing money on to the floor of the New York Stock Exchange and nominating a pig for President. "At a deeper level both extremes were united in their attack on the traditional institutions that reproduce dominant cultural–ideological relations – the family, education, media, marriage and the sexual division of labor. There was a shared emphasis on the freedom to question and experiment, a commitment to personal action, and an intensive examination of the self" (Whiteley, 1992: p. 83) whether pathologically invasive or creatively expressive. As Roszak wrote at the time: "Beat-hip bohemianism may be too withdrawn from social action to suit New Left Radicalism; but the withdrawal is a direction the activist can readily understand… We see the underlying unity of the countercultural variety, then, if we see beat-hip bohemianism as an effort to work out the personality and total life style that follow from New Left social criticism" (1970: p. 66).

Hippies

The hippies were a significant aspect of the counterculture and were initially centered on San Francisco's Haight Ashbury district, representing a style of cultural politics and ostensibly rejecting mainstream society. Not least they advocated nonviolence and peace, promoting instead openness and tolerance, seeking spiritual guidance from sources outside the Judeo-Christian tradition, particularly Eastern religions, as well as astrology. Their promotion of recreational drugs, such as marijuana and lysergic acid diethylamide (LSD) was justified as a means of attaining an expanded consciousness. The novelist Ken Kesey was one of the best-known literary spokesmen for the movement, and became famous for the bus tours he made with the so-called Merry Pranksters and his advocacy of LSD. Wolfe (1968) documented the travels of Ken Kesey and his 'Merry Pranksters' during the mid-1960s. Similarly, Thompson's (1993) *Fear and Loathing in Las Vegas* also pays due homage to the countercultural legacy and its vision of a new world.

Folk and rock music were an integral part of hippie culture. In a period of heightened social, political, and psychological

awareness, and rapidly increasing global communications technologies, popular music was recognized as a social force, which expressed issues of cultural and political significance across a wide geographic area. As Jeff Nuttall observed, two of the aims of the Underground were "to release forces into the prevailing culture that would dislocate society, untie its stabilizing knots of morality, punctuality, servility and property; and [to] expand the range of human consciousness outside the continuing and ultimately soul-destroying boundaries of the political utilitarian frame of reference" (1970: p. 249), a philosophy which resonated with the identification of rock as a political weapon. In this respect, John Lennon's 'All You Need is Love' is significant as the world's first live global television linkup. Composed as a 'sing along' for the international television program, 'Our World,' the program was broadcast via satellite on 25 June 1967, and was viewed by an estimated 400 million people in 26 countries throughout the world.

As a collective experience, rock appeared to provide the means whereby young people could explore "the politics of consciousness, love, loneliness, depersonalization the search for the truth of the person, and the attempt to set up an alternative life style" (Roszak, 1970: p. 156). Richard Neville comments, "From Berlin to Berkeley, from Zurich to Notting Hill, Movement members exchange a gut solidarity, sharing common inspirations, strategy, style, mood and vocabulary. Long hair is their declaration of independence, pop music their Esperanto and they puff pot in their peace pipe" (1971: p. 14). It is also relevant to note that the distinction between music for living by, and music for leisure was of fundamental importance to the counterculture, highlighting the ways in which music and sociocultural politics could fuse into a collective experience. As such, the impact of noise (inharmonious sound, distortion, dissonance, and the connotations surrounding discord itself) can be interpreted as underpinning a revolutionary agenda suggestive of a state of creative anarchy (to an extent, this can be traced back to the romantic anarchism of the Beats), which is arguably distinct from the more soft-focus connotations of 'All You Need is Love' (The Beatles, 1968) and the pacifist agenda implicit in such slogans as 'Make Love Not War.' Although it is recognized that the fight against middle-class prurience led increasingly toward an explicit identification of sexual freedom with total freedom which, at its extreme, embraced pornography, including the so-termed velvet underground advertisements for blue moves and classified ads, and *play power*'s "Female Fuckability Test" (Neville, 1971: p. 14). As such, while love was fundamental to the philosophy of the counterculture, there was nevertheless a marked difference between the transcendental spirituality promised to followers of the Maharishi Mahesh Yogi and the revolutionary liberation of the Yippie Party's Jerry Rubin and his symbolic call for patricide. If, however, Roszak is correct in identifying "beat-hip bohemianism as an effort to work out the personality and total life style that follow from New Left social criticism" (1970: p. 66), then a movement toward a communality based on love could appear a logical development. Discord and the darker extremes of 'noise' thus come into focus as the first stage in the countercultural agenda of establishing a relevant and alternative lifestyle.

An extreme example comes from the MC5, Detroit's leading underground band, who performed at revolutionary rallies against a political backdrop of racial inequality and suppression, most notably the 1967 riots when confrontations with the US national guard and US army troops resulted in 43 deaths and over 7000 arrests. Managed by John Sinclair, 'Minister of Information' for the White Panthers, the band's second single, 'Looking at You' was supposedly recorded in downtown Detroit "sometime circa the 1967 riots, feeding the myth that the band were a group of 'rock and roll guerrillas' who both fomented and embodied disorder with their rousing performances" (Waksman, 1998: p. 47) and feedback-laden sound. Further violence against protesters at the 1968 Democratic National Convention was highlighted by MC5s revolutionary rallying cry, 'Kick Out the Jams, Motherfuckers.' As Jim Deregotis writes, "Strange as it seems in retrospect, the music industry thought revolution might be the next big marketing concept circa the late '60s, and MC5 were signed to Elektra (along with their younger 'baby brother' band, the Stooges) by Danny Field (A&R) who later managed The Ramones." Jim Deregotis, The 'MC5 Classic still has plenty of kick to it,' Great Albums: The MC5, Kick Out the Jams' 11 August 2002 http://www.jimdero.com/News2001/GreatAug4MC5.htm http://www.rollingstone.com/music/albumreviews/kick-out-the-jams-19690405 accessed 14.04.11.

The Beatles also offered a vision of chaos in the Coda to 'A Day In the Life' (*Sgt. Pepper's Lonely Hearts Club Band*, 1967) where a spiraling ascent of sound challenges "the warm combination of acoustic guitar and piano" (Hannan, 2008: p. 59) that introduces the song, the complacency inherent in contemporary society ('A crowd of stood and stared…,' 'a crowd of people turned away'). The dramatic effect of the indeterminate textures of the ascending orchestral clusters is contextually anarchic in its intrusion into the passacaglia-like countermelody of the first two verses, albeit that Lennon's lyrical vocal "features an unusually expressive use of nonharmonic tones, notably the leaps to dissonant notes in the latter part of the first and fourth bars ('I read the news to-day/oh-boy-/a-about a luck-y man who/made the grade'), so hinting at an underlying disillusionment" (O'Grady, 1983: p. 137). McCartney's second section is more frantic in its delivery, coming to an abrupt and fuzzy conclusion at the words "Somebody spoke and I went into a dream," the vocal line becoming obscured by the progressively louder orchestral unisons on the root of each chord, before a two-bar unison brass motif plunges the key down a minor third for the return of the first section. This time, the bass leads to the intense orchestral build up before the shifting timbres of the final sustained piano chord and its 30-s decay lead finally into "a noisy 15 kHz tone leads to a 2-s piece of gibberish which was cut into the run-out groove of the L.P" (Hannan, 2008: p. 61), so casting doubt as to whether the song's powerful imagery and disruptive musical aesthetics would inspire listeners to wake up and question (as suggested by the ringing alarm clock that heralds McCartney's middle section and the crowing cock in 'Good Morning, Good Morning' or whether it would be interpreted as yet another example of the Beatles' love of the "slicks and tricks of production" (Goldstein, 1967: p. 173). Russell Reising and Jim LeBlanc note that psychedelic rock has used alarm clocks, chimes, bells, the effects of clocks ticking, and other effects to signal psychedelic awakening. "Within and without: *Sgt. Pepper's Lonely Hearts Club Band* and Psychedelic Insight" in Julien (2008: p. 111). It is also reported that

McCartney's interest in the avant-garde, and his admiration for Stockhausen, led to the inclusion of an "instrumental passage… [with] a spiraling ascent of sound… [starting] the passage with all instruments on their lowest note and climbing to the highest in their own time" (Martin and Pearson, 1994: p. 56). (see Whiteley, 2013: for a full discussion of the Beatles and 'A Day in the Life' and the MC5 and 'Kick Out the Jams.')

It is also suggested that songs associated with the counterculture – "through their musical style, their mode of performance, use and interpretation – point towards particular historical moments, knots of collective experience, cultural traditions and so forth. At this level – the level of style-indicators and cultural codes – music can be taken to represent to us the features of a specific conjuncture (neo-Gramscianism), structure of feeling (Raymond Williams), or chronotope (Bakhtin)" (Middleton, 2000: p. 231) and in previous work (Whiteley, 1992) it was suggested that "drug-related experience works through anaphonic musical signs (i.e., signs that work through structural analogy with non-musical experiences of time, space, movement etc), culturally mediated semiotic coding musical features that work through comparison with other, familiar effects in different styles; intra-cultural interpretation (what countercultural listeners said, cultural resonance (analogy between aspects of the music and of the cultural context); technological innovation" (Middleton, 2000: p. 231). While examples can be drawn from, for example, the Grateful Dead, the Beatles, Pink Floyd, the Doors and the Rolling Stones, its most striking focus can be found in the musical and performing body of Jimi Hendrix, where there are intersecting discourses not only of 'progressiveness' and 'psychedelia,' but also of race, gender, and sexuality. Significant examples would include 'Purple Haze,' 'Love or Confusion,' 'Are You Experienced,' and 'Spanish Castle Magic' (Whiteley, 2000: pp. 235–261).

The hippie's identification with the metaphysical is explored in Lennon's 'Tomorrow Never Knows' (The Beatles, *Revolver*, 1966), which set the agenda for psychedelic music in its evocation of the Vedic teaching 'Be Here Now' – living in the present but conscious of both past and future. Rather than the more usual, extravagant colors of psychedelia, the colliding musical circles are introspective, a sensory experience of Timothy Leary's voluntary ego-death. Leary's *The Psychedelic Experience* (Leary et al., 1964) is a paraphrase of *The Tibetan Book of the Dead* which had influenced Lennon's personal experience of LSD and his composition, 'Tomorrow Never Knows.' This is achieved by looped effects which constantly move to the 'eternal' presence of the tonic (an orchestral chord on Bb major; a Mellotron played on flute setting; another Mellotron oscillating in 6/8 over the 4/4 pulse of the melody line, moving from Bb (the flattened 7th) to C (the tonic on its string setting), so suggesting a musical metaphor for the tug between the conscious/unconscious of psychedelic experience (Whiteley, 2002: p. 219). George Harrison's 'Within You Without You' (*Sgt Pepper's Lonely Heart's Club Band*, 1967) also reflects his search for Eastern mysticism. As George Martin reports, "What appealed to George was the sound – the unique sound the instruments made when used in this very ancient (Indian classical) tradition" (Martin and Pearson, 1994: p. 126). Encapsulated by an endless drone and metric shift between fours, fives, and threes, "the vocal line is doubled by two dilrubas (the dilruba is a bowed lute with a buzzy, nasal, hollow-sounding timbre) and, at certain points, by a (Western) string ensemble (eight violins and three cellos) and, at other points, sitar"… "Harrison's vocal sound is mantra-like, sitting in the back of the mix, dominated by the elaborate tabla playing" (Hannan, 2008: pp. 54–55).

While the search for expanded consciousness through both hallucinogenic drugs and the metaphysical of Eastern religions were important aspects of music associated with the hippie counterculture, "the lyrical content of popular music also took on increasingly radical dimensions as the 1960s progressed. Songwriters and lyricists saw it as their artistic responsibility to respond directly to current social and political issues which, during the mid-1960s, were becoming increasingly more turbulent" (Bennett, 2004: p. xv). These included protests against the war in Vietnam, including The Doors 'The Unknown Soldier,' where the promotional film showed singer Jim Morrison being blindfolded and shot, a scene intended to parody the 'summary executions of untried opponents' featured in US media reports about the Vietnam War (Snowman, 1984: p. 163). The anti-Vietnam War protest was supported by many musicians including John Lennon and his wife Yoko Ono, who staged a 10-day bed-in at the Queen Elizabeth Hotel in Montreal, Canada, in the cause of world peace. 'Give Peace A Chance' was written by Lennon, and recorded live in two sessions in the hotel bedroom.

Festivals

Festivals were an important medium for spreading countercultural ideology. The open-air 'Be-In,' Golden Gate Park, San Francisco, which attracted 20 000 people and featured such pioneering acid-rock groups as Country Joe and the Fish, Moby Grape, the Quicksilver Messenger Service, the Steve Miller Band, Jefferson Airplane and the Grateful dead. "An overawed *Time Magazine* recorded, 'The huge crowd was peaceful … an amazing tribute to Haight-Ashbury.' More accurately it was a tribute to LSD and marijuana, which had been distributed by a dazed Santa Claus" (Neville, 1971: p. 28). The scale and coverage of the Monterey Festival of June 1967 made it an event of supreme importance to the rock world. American acts on the bill included the Byrds, the Paul Butterfield Band, the Grateful Dead, the Blues Project, the Mothers of Invention, Simon and Garfunkel, and the Doors.

Woodstock is recognized as the most important festival of the period, attracting between 400 000 and 500 000 people. It is remembered both for its 'bringing together' of the countercultural generation and its music, which was both politically subversive while acting as a springboard for the more commercial rock and pop events that followed. Country Joe McDonald's "I-Feel-like-I'm-Fixin'-to-Die-Rag" was a direct criticism of the Vietnam War and its sing-along chorus was made world-famous through its inclusion in Michael Wadleigh's film *Woodstock (Three Days of Peace and Love)* (1970). Joni Mitchells 'Woodstock,' dedicated to the Woodstock festival, contains a reference to bombers turning into Butterflies, the bombers in question being those attacking North Vietnamese targets (Bennett, 2004: p. xvi). It was, however, Hendrix's performance of 'The Star Spangled Banner'

that provides the most compelling example of rock as a political weapon. Even though the words are not sung, they were familiar to his audience, and the feedback and sustain provide a commentary on the flag itself as, at the words, 'broad stripes and bright stars,' the sounds plummet and waiver. The attack comes at the evocative 'rocket's red glare' where Hendrix creates the sound of a fighter plane "coming out of a deep dive, the impact of bombs striking the earth and the cries of the Cambodian peasants" (Henderson, 1990: p. 245). The straight melody finally comes through on 'gave proof through the night' and the anthem ends to the sounds of feedback and sustain and a final ear-shattering grind as the guitar strings are treated to a crude bottleneck slide against the mike stand (Whiteley, 2004: p. 25). As Charles Shaar-Murray writes: "the ironies were murderous: a black man with a white guitar; a massive almost white audience wallowing in a paddy field of its own making; the clear, pure, trumpetlike sounds of the familiar melody struggling to pierce through clouds of tear-gas, the explosions of cluster-bombs, the screams of the dying, the crackle of flames, the heavy palls of smoke stinking with human grease, the hovering chatter of helicopters" (1989: p. 24) (see Whiteley, 2004: pp. 18–28). Jimi Hendrix, together with Richie Haven and Sly Stone were the only Afro-American artists to appear at the Woodstock, which has been identified as both the "defining moment of the late 1960s and the swansong of the countercultural movement that characterized the era" (Bennett, 2004: p. xviii).

If Woodstock is largely described as utopian, then the Rolling Stones free concert at Altamont heralded the end of the Age of Aquarius. It also raises the question "of the relationship between performer/audience, the performance, and the musical text itself; that the unmitigated violence at Altamont suggests that for many the songs were interpreted as inciting brutality, that they provided a model for behavioral patterns" (Whiteley, 1997: pp. 84–85). While most agree that the reason for the violence at Altamont – which began well before the Stones took the stage – had more to do with the hiring of Hells Angels as security, few would argue that the menacing subject matter and sound of songs like 'Sympathy for the Devil' and 'Midnight Rambler' would "hardly calm the already explosive atmosphere" (1997: p. 84). The Maysles Brothers film of the concert, *Gimme Shelter* (1971) provides a disturbing insight into those attending Altamont: one murdered, total dead four, two run over, one drowned in a drainage ditch, hundreds injured, and countless more stoned on acid.

Altamont was only one of a series of crisis, which signaled the end of the 1960s counterculture. The association of LSD with Satanism which had resulted in the murders of Sharon Tate and four of her friends by members of the Manson 'family' led to a panic reaction to the adverse effects of drugs. Publicity surrounding the death of famous rock stars also led to highly volatile criticisms of the counterculture and in particular the concept of creativity through drug experience. In September 1970, Hendrix had been found dead from inhalation of vomit following barbiturate intoxication. On 4 October of the same year, Janis Joplin died from a heroin overdose. In 1971, Jim Morrison died of a heart attack caused by heroin or alcohol in Paris. Personal survival seemed paramount and the US broadcast license authorities warned radio stations that they would be subverting the government's campaign against drug abuse if they played lyrics, which mentioned dope. Finally, the shooting of four students at Kent State University tragically proved how deep was the gulf between those who wanted to change America and its policies in Vietnam, and the conservative majority who believed in law and order above everything else.

Countercultures: The Broader Picture

While academic attention has been focused largely on the United States and the United Kingdom, the hippie movement quickly established a presence throughout the Western World, with countercultural movements emerging in, for example, Christiania (Denmark) (see Hall, 2014), Naples (Italy) (see Vacca, 2013: pp. 67–84), East and West Berlin (Germany) (see Stahl, 2014), and in some areas of South America, Asia, and the former Soviet bloc. In Paris, university students and workers joined forces to protest about workers' right and the conditions in factories, while across Europe students occupied universities and demonstrated for better facilities. Although activists received considerable attention, as Joel Fort observed, "they probably number no more than tens of thousands in such groups as the Students for a Democratic Society, antiwar and antidraft groups, student power advocates and some elements of the hippies and yippies. There are the equally radical groups of special minorities, namely the Black Panthers, Brown Berets, and others seeking power, dignity and sometimes justice for racial and cultural minorities" (1969: p. 210).

In the United States, attention focused on Civil Rights. "In October 1966 the Student Nonviolent Coordinating Committee (SNC) was established to enable students to become organized in the battle for civil rights. However, neither the SNCC nor President Kennedy's Civil Rights Bill, passed in 1964, served to alleviate the abject poverty and total political impotence of African Americans. Increasing racial discrimination and harassment across the US led to inner-city riots…[as] Civil rights organizations were superseded by the Black Power movement under the leadership of Stokely Carmichael" (Bennett, 2004: p. xvii). Civil Rights activist, Martin Luther King was best known for his advocacy of nonviolent civil disobedience and his stirring speech 'I have a dream.' He was assassinated on 4 April 1968. Malcolm X, the human rights activist was assassinated on 21 April 1965. His father had earlier been murdered by the Klu Klux Klan. Black artists associated with the Black Power movement included James Brown, 'Say It Loud: I'm Black and I'm Proud,' and Aretha Franklin, 'Respect.'

It is also interesting to note the emergence of the so-called Jesus hippies, commonly dubbed 'Jesus freaks.' Questioning "the authority of the church and reinstating biblical authority, most commonly apocalyptic, they simultaneously retained what was widely considered a countercultural aesthetic, often sporting the hippie image while using popular music for Christian proselytizing." As Young observes, "Barry McGuire's classic 'Eve of Destruction' (1965) gained notoriety within the context of radicals, revolutionaries, and those who were convinced the world was on the brink of either cosmic or physical annihilation. Along with Bob Dylan, McGuire's conversion to Christianity gave credibility to the rising Jesus-freak revolution. As celebrities such as Johnny Cash and Eric Clapton converted, revivalistic

Christianity found entry into the entertainment industry, launching a groundswell of activity … Fleetwood Mac's Jeremy Spencer joined the rigid and controversial commune the Children of God, a notable doomsday group. Secular rock groups also explored religious themes, adding to the already enigmatic culture of Christian hippies. These included The Byrds' version of 'Jesus is Just Alright' (1969), Ocean's 'Put Your Hand in the Hand' (1970), Creedence Clearwater Revival's 'Bad Moon Rising' (1969), and Norman Greenbaum's 'Spirit In The Sky' (1969) (see Young, 2013: pp. 51–67).

Despite the breadth of the countercultural movement, research suggests that women were largely marginalized. As Roszak observed, the struggle for liberation was seen mainly "as the province of men who must prove themselves by 'laying their balls on the line'. Too often this suggests that the female of the species must content herself with keeping the home fires burning for her battle-scarred champion or joining the struggle as a camp follower. In either case, the community is saved *for* her, not *by* her" (1970: footnote to p. 65). Despite the emergence of the so-termed women's liberation movement and the questioning of power relations in personal life (in reproduction, sexual relationships, the household division of labor, popular culture, why certain ethnic minorities and working class women were marginalized in their access to the production and consumption of knowledge and culture, and why women generally were not represented within the hierarchies of science and arts) the counterculture remained male-dominated. The breaking down of old-restraints ('free love') privileged a male sexuality, confirming the traditional definitions of masculinity and femininity under the dubious banner of progressiveness, and in the field of music there was little real opportunity to either take control or enjoy the prestige offered to male artists. Apart from the biting social and political commentaries from such performers as Joan Baez and Buffy St Marie and the significance of such frontline artists as Mama Cass (The Mamas and the Papas), Grace Slick (Jefferson Airplane), and Janis Joplin (albeit at a cost, with Joplin dying in 1970 and Cass Elliott in 1974) the musical ethos of the period undermined the role of women, positioning them as desexualized earth mothers, fantasy figures, and easy lays. (Whiteley, 2000: pp. 1–94)

It is thus suggested that while there were highly political elements within the counterculture, the movement was not uniformly politicized in a way that dictated that the only way forward was to overthrow the capitalist system by whatever means. Feminist academics identified patriarchy as the major obstacle in achieving personal freedom, and attitudes toward gay and lesbian artists remained problematic, despite the United Kingdom's Homosexual Reform Bill (1967), the promotion of camp taste and drag culture at Andy Warhol's New York Factory, and the Velvet Underground's uncompromising insights into urban culture, sexuality and voyeurism in the *Velvet Underground and Nico*. Warhol's Factory also provided a refuge for drag queens who are celebrated in *The Philosophy of Andy Warhol* (1975), thus providing a crucial link with the 27 June 1969 police raid on the Stonewall Inn gay bar on Christopher Street in Greenwich Village, and the subsequent riots that were the flashpoint of the American gay liberation movement of the 1970s.

In retrospect, it is apparent that the 1960s counterculture incorporated diverse groupings, including student protest, peace and antiwar movements, the civil rights movement, environmentalists, the utopian, dystopian, apocalyptic, and freaks these are discussed in Whiteley and Sklowers edited text on Countercultures and Popular Music (2013), and could also "manifest itself differently at specific times and within specific places depending on local socio-economic, cultural and demographic circumstances" (Bennett, 2013: pp. 19–31). More recently, the term counterculture has come to symbolize 'forms of social action and/or alternative lifestyles and belief that appear – or can be made to appear – to link with the aesthetic, political and cultural trends associated with the late 1960s manifestation of counterculture', so providing a sociocultural-political analogy for the 'hegemonic struggles that continue to inform everyday life in myriad societies and cultures across the globe' (Bennett, 2013).

See also: Civil Rights Movements; Counterculture: 1960s and Beyond; Culture, Sociology of; Ecofeminism; Feminist Political Theory and Political Science; Gay and Lesbian Movements; Identity Movements; Minorities; Movements, Social; Political Protest and Civil Disobedience; Religion and Youth; Social Protest; Subculture, Sociology of; Urban Sociology; Youth Culture, Anthropology of; Youth Culture, Sociology of; Youth Movements; Youth, Music, and Peace Building.

Bibliography

Arnold, G., 2014. Contradictory cultural rhetoric in 'Woodstock' and 'Gimme Shelter'. In: Whiteley, S., Sklower, J. (Eds.), Countercultures and Popular Music. Ashgate, Aldershot.

Bartkowiak, M.J., 2009. The MC5 and Social Change: A Study of Rock and Revolution. McFarland, Jefferson, NC.

Bennett, A. (Ed.), 2004. Remembering Woodstock. Ashgate, Aldershot.

Bennett, A., 2005. Culture and Everyday Life. Sage, London.

Bennett, A., 2013. Reappraising countercultures. In: Whiteley, S., Sklower, J. (Eds.), Countercultures and Popular Music. Ashgate, Aldershot.

Deregotis, J., 11 August 2002. The 'MC5 Classic Still Has Plenty of Kick to It, Great Albums: The MC5, Kick Out the Jams'. http://www.rollingstone.com/music/albumreviews/kick-out-the-jams-19690405 (accessed 04.04.11.).

Fort, J., 1969. The Pleasure Seekers: The Drug Crisis, Youth and Society. Grove Press, New York.

Ginsberg, A., 1956. Howl and Other Poems. City Lights, San Francisco, CA.

Goldstein, R., 1967. I blew my cool through the New York times. New Times and the Village Voice, 173.

Hall, S., Jefferson, T. (Eds.), Resistance Through Rituals: Youth Subcultures in Post-War Britain. Hutchinson, London.

Hannan, M., 2008. The sound design of sgt. pepper's lonely heart's club band'. In: Julien, O. (Ed.), Sgt. Pepper and the Beatles: It Was Forty Years Ago Today. Ashgate, Aldershot, pp. 45–62.

Henderson, D., 1990. The Life of Jimi Hendrix: 'Scuse Me While I Kiss the Sky, rev ed. Omnibus Press, London.

Hicks, B., 1970[1996]. Jimi Hendrix: a memorial (northwest passage, 29th September 1970). In: Potash, C. (Ed.), The Jimi Hendrix Companion: Three Decades of Commentary. Schirmer, Schirmer, New York, pp. 207–210.

Julien, O. (Ed.), 2008. Sgt. Pepper and the Beatles: It Was Forty Years Ago Today. Ashgate, Aldershot.

Kerouac, J., 1972a. On the Road. Penguin, London (Original published 1957).

Kerouac, J., 1972b. Desolation Angels. Panther, London (Original published 1965).

Kerouac, J., 2007a. The Dharma Bums. Penguin, London (Original published 1958).

Kerouac, J., 2007b. Lonesome Traveller. Penguin, London (Original published 1965).

Kesey, K., 1973. One Flew over the Cuckoo's Nest. Picador, London (Original published 1962).

Leary, T., Metzner, R., Alpert, R., 1964. The Psychedelic Experience: A Manual Based on 'The Tibetan Book of the Dead'. University Books, New Hyde Park, New York.
Marcus, G., 1991. Lester Bang, Psychotic Reactions and Carburettor Dung. Literature as Rock 'n' Roll, Rock 'n' Roll as Literature. Minerva, London.
Marcuse, H., 2002. One-Dimensional Man. Studies in the Ideology of Advanced Industrial Society. Routledge, London and New York (Original published 1964).
Marx, K., 1969. Theories of Surplus Value. Lawrence & Wishart, London.
Martin, G., Pearson, W., 1994. Summer of Love: The Making of Sgt. Pepper. McMillan, London.
Middleton, R. (Ed.), 2000. Reading Pop. Approaches to Textual Analysis in Popular Music. Oxford University Press, Oxford.
Miles, B., 1990. Ginsberg: A Biography. Viking, London.
Morgan, B., 2010. The Typewriter is Holy: The Complete, Uncensored History of the Beat Generation. New York Free Press, New York.
Neville, R., 1971. Play Power. Paladin, London.
Nuttall, J., 1970. Bomb Culture. Paladin, London (Original published 1968).
Olivier, J. (Ed.), 2010. Sgt. Pepper and the Beatles: It Was Forty Years Ago Today. Ashgate, Aldershot.
O'Grady, T.J., 1983. The Beatles: A Musical Evolution. Twayne Publishers, Boston.
Potash, C. (Ed.), 1996. The Jimi Hendrix Companion: Three Decades of Commentary. Schirmer Books, New York.
Reising, R. (Ed.), 2002. 'Every Sound There Is': The Beatles Revolver and the Transformation of Rock and Roll. Ashgate, Aldershot.
Reising, R., LeBlanc, J., 2010. Within and without: sgt. pepper's lonely hearts club band and psychedelic insight. In: Olivier, J. (Ed.), Sgt. Pepper and the Beatles: It Was Forty Years Ago Today. Ashgate, Aldershot, pp. 103–120.
Roszak, T., 1970. The Making of a Counter Culture: Reflections on the Technocrati Society and Its Youthful Opposition. Faber and Faber, London.
Shaar-Murray, C., 1989. Crosstown Traffic: Jimi Hendrix and Post-war Pop. Faber and Faber, London.
Snowman, D., 1984. America Since 1920. Heinemann, London.
Stahl, H., 2014. Auditory traces of subcultural practices in 1960s Berlin. A border-crossing soundscape of pop. In: Whiteley, S., Sklower, J. (Eds.), Counterculture and Popular Music. Ashgate, Aldershot.
Thompson, H.S., 1971. Fear and Loathing in Los Angeles. Random House, New York.
Thorbjorg, D.H., 2013. Countercultural space does not persist: Christiania and the role of music. In: Whiteley, S., Sklower, J. (Eds.), Counterculture and Popular Music. Ashgate, Aldershot.
Turner, S., 1996. Jack Kerouac: Angelheaded Hipster. Bloomsbury, London.
Vacca, G., 2013. Music and countercultures in Italy: the Neopolitan Scene. In: Whiteley, S., Sklower, J. (Eds.), Counterculture and Popular Music. Ashgate, Aldershot.
Waksman, S., 1998. Kick out the jams! The MC5 and the politics of noise. In: Swiss, T., Sloop, J., Herman, A. (Eds.), Mapping the Beat. Popular Music and Contemporary Theory. Blackwell Publishers Inc., Malden MA, pp. 47–76.
Warner, S., 2013. Text and Drugs and Rock 'n' Roll. The Beats and Rock Culture. Bloomsbury, New York.
Whiteley, S., 1992. The Space between the Notes: Rock and the Countercultue. Routledge, London.
Whiteley, 1997. Little red rooster v. the Honky TonkWoman. Mick Jagger, sexuality, style and image. In: Whiteley, S. (Ed.), Sexing the Groove. Popular Music and Gender. Routledge, London, pp. 67–99.
Whiteley, S., 2000. Progressive rock and psychedelic coding the work of Jimi Hendrix. In: Middleton, R. (Ed.), Reading Pop. Approaches to Textual Analysis in Popular Music. Oxford University Press, Oxford, pp. 235–261.
Whiteley, S., 2002. Love is all and love is everyone: a discussion of four musical portraits'. In: Reising, R. (Ed.), 'Every Sound There Is': The Beatles Revolver and the Transformation of Rock and Roll. Ashgate, Aldershot, pp. 209–221.
Whiteley, S., 2004. 1. 2. 3 What are we fighting 4? Music, meaning and the star spangled banner. In: Bennet, A. (Ed.), Remembering Woodstock. Ashgate, Aldershot, pp. 18–28.
Whiteley, S. (Ed.), 2012. Contre-cultures I: Théorie et scènes. Volume ! La revue des musiques populaires, vol. 9. numéro 1. Éditions Mélanie Seteun. www.seteun.net / editions@seteun.net.
Whiteley, S. (Ed.), 2013. Contre-cultures II: Utopies, dystopies, anarchie. Volume ! La revue des musiques populaires, vol. 9. numéro 2. Éditions Mélanie Seteun. www.seteun.net / editions@seteun.net.
Whiteley, S., 2013. Kick out the jams: creative anarchy and noise in 1960s rock. In: Goddard, M., Halligan, B., Spelman, N. (Eds.), Resonances: Noise and Music (Continuum, 2013).
Whiteley, S., Sklower, J. (Eds.), 2013. Counterculture and Popular Music. Ashgate, Aldershot.
Whiteley, S., Sklower, J. (Eds.), 2014. Counterculture and Popular Music. Ashgate, Aldershot.
Wilentz, S., 2010. Bob Dylan in America. Bodley Head, London.
Wolfe, T., 1989. The Electric Coll Aid Acid Test. Black Swan, London (Original published 1968).
Womack, K. (Ed.), The Cambridge Companion to the Beatles. Cambridge University Press, Cambridge.
Young, S.D., 2013. Apocalyptic music and countercultural Christian influences. In: Whiteley, S., Sklower, J. (Eds.), Contre-cultures: Utopies, Dstopies, Anarchie, Éditions Mélanie Seteun, vol. 9–2, pp. 51–67.

Discography

Beatles, All You Need is Love.
Barry McGuire's classic 'Eve of Destruction' (1965).
Hendrix, Star Spangled Banner.
'Purple Haze', 'Love or Confusion' 'Are You Experienced' and 'Spanish Castle Magic' MC5 'Looking at You', 'Kick Out the Jams'.
Tomorrow Never Knows' (The Beatles, *Revolver,* 1966).
The Byrds' version of 'Jesus is Just Alright' (1969), Ocean's 'Put Your Hand in the Hand' (1970), Creedence Clearwater Revival's 'Bad Moon Rising' (1969), and Norman Greenbaum's 'Spirit In The Sky' (1969).
'Within You Without You' (*Sgt Pepper's Lonely Heart's Club Band* 1967) 'A Day In the Life' (*Sgt. Pepper's Lonely Hearts Club Band*, 1967).

Filmography

Maysles Brothers. Gimme Shelter (1970) Maysles Films (USA).
Rebel Without A Cause (1955) to.
Rock Around The Clock (1956).
Woodstock: Three Days of Music, Peace and Love. A Once-in-a-Lifetime Celebration, 1970. Warner Bros. Inc.: A Warner Communication Company.

Counterfactual Reasoning, Qualitative: Philosophical Aspects

Stathis Psillos, University of Athens, Athens, Greece

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Hájek, volume 4, pp. 2872–2874, © 2001, Elsevier Ltd.

Abstract

This article reviews the two major approaches to counterfactual conditionals: the metalinguistic or 'support' approach and the possible worlds approach. It identifies the major problems they face and explores the thought that the core idea behind counterfactual reasoning is to assert that there are not good inductive reasons to affirm simultaneously a generalization and the physical possibility of an exception to it. It also examines the role of counterfactuals in causal inference, causation, and laws of nature.

Subjunctive conditionals or counterfactual conditionals are probably as old as language itself since they give speakers the means to talk about what would or might happen or have happened if certain things were to happen or had happened. In ordinary language, they have the form:

If *x* were (not) the case, then *y* would (not) be the case,
or
If *x* had (not) been the case, then *y* would (not) have been the case.

Subjunctive conditionals leave open the possibility of the realization of whatever is expressed in the antecedent, for example, if John were to come to the party, Mary would not go. Counterfactual (or 'contrary-to-fact') conditionals are such that the antecedent is false; the state-of-affairs expressed in it has not actually obtained, for example, if John had gone to the party, Mary would not have gone. (Here it is an implicit assumption that the actual course of events is that John did *not* go to the party.) Both kinds of conditional contrast to indicative conditionals of the form: *if x is the case, then y is the case*. Although there are differences between them, we will not be detained by them, and instead concentrate on counterfactual conditionals. (From now on, we will follow customary usage and use $\Box\!\rightarrow$ to express the counterfactual 'if…, then…'.)

Counterfactuals fail a number of principles that indicative conditionals satisfy. Most important, they are nonmonotonic, that is they fail the principle of strengthening of the antecedent:

$$X\Box\!\rightarrow Y \text{ does not entail } X\&Z\Box\!\rightarrow Y.$$

Example: If John had been poisoned, he would have died. This does not entail: if John had been poisoned and taken an antidote, he would have died.

Transitivity:

$$X\Box\!\rightarrow Y \text{ and } Y\Box\!\rightarrow Z \text{ does not entail } X\Box\!\rightarrow Z.$$

Example: If John had gone to the market, he would have taken the bus; if John had taken the bus, then he would have gone to his office. These two do not entail: if John had gone to the market, then he would have gone to his office.

Contraposition:

$$X\Box\!\rightarrow Y \text{ does not entail not} - Y\Box\!\rightarrow \text{not} - X.$$

Example: If John had lived in a Euro-zone country, he would have used Euros. This is not equivalent to: If John had not used Euros, he would not have lived in a Euro-zone country.

If we assume that classical semantics apply to indicative conditionals (the indicative conditional is true if either the antecedent is false or the consequent is true), trying to apply classical semantics to counterfactuals leads to their trivialization: Given the actual falsity of the antecedent of a counterfactual, both the counterfactual with the actual consequent and the counterfactual with the negation of the actual consequent end up being true.

Example: Given that the vase was not struck with a hammer, both of the following two conditionals (treated as material conditionals) are true:

If this vase had been struck with a hammer, it would have broken, and
If this vase had been struck with the hammer, it would not have broken.

The failure of the three principles and this unwanted consequence is a *reduction* of the view that classical semantics applies to counterfactuals. But then, what is the right semantics for counterfactuals? What are the truth-conditions of a counterfactual conditional? Or, at least, what are their assertibility conditions? This problem came under sharp focus in the 1940s, when philosophers started to realize that the concept of counterfactual conditionals is instrumental for the explication and understanding of a number of other philosophical concepts. As Nelson Goodman put it in one of the first papers to deal with this issue, "(…) if we lack the means of interpreting counterfactual conditionals, we can hardly claim to have any adequate philosophy of science" (1947: p. 113).

Note that in assessing a counterfactual assertion $X\Box\!\rightarrow Y$, we should replace, as it were, the actual nonoccurrence of X with the supposition that X has occurred. But given that the laws of nature and the actual course of events led to non-X, in supposing the actual occurrence of X, we need to make counterfactual suppositions concerning either the laws or the actual course of events, such that X actually occurred. In particular, we have to assume that either some laws were broken (so that X did happen after all) or that some actual particular matters of fact did not occur. Hence, in specifying the semantics of counterfactuals, we have to take into account considerations concerning the laws of nature and other particular matters of fact prior to the conditions specified in the antecedent of the counterfactuals.

There are two major views concerning the semantics of counterfactuals, the first being introduced by Goodman himself, whereas the second was developed by Robert Stalnaker

and David Lewis (but introduced by William Todd in 1964). Let us examine them in turn.

The Metalinguistic or 'Support' View

On the first major view, known as 'support view' or 'metalinguistic view,' a counterfactual conditional $X\Box\rightarrow Y$ is an elliptic or telescoped argument (or a linguistic construction *about* an argument) such that the antecedent X (taken in its indicative form) together with suitable auxiliary premises entail the consequent (taken in its indicative form). Hence, $X\Box\rightarrow Y$ should not be taken to be a statement at all; its assertoric content is captured by the following argument type:

X&S&L (materially) imply Y,

where L are statements capturing laws of nature and S are singular statements capturing background or collateral conditions that should be 'cotenable' with the antecedent X and express necessary conditions for the consequent to follow.

Example: If this match had been struck (X), it would have lit (Y). For a struck match to light, it is necessary that the match is well made; that it is dry; that there is oxygen; and so on. But these conditions (collectively designated by S) are not sufficient for the lighting of the match; various laws are required (collectively designated by L). Hence, in asserting the counterfactual 'if this match had been struck, it would have lit,' we are committed to the truth of the various statements that describe the laws and the relevant background conditions.

The first general problem with this view concerns the characterization of the relevance relation when it comes to the background–collateral conditions. It cannot be too permissive. If we allowed all true statements to be relevant to the argument, the falsity of the antecedent X (which is *actually* false) would be relevant, too; however, the counterfactual then would be trivially true. It cannot be too restrictive either. The consequent of the counterfactual is false as well. Hence, not-Y is the case. It is not hard to see that given (X&S&L→Y) and not-Y and L, it follows (by obvious steps) that X→not-S. (Assuming, for simplicity that S is 'the match is dry,' the conclusion would be as follows: If the match is struck, it will not be dry!) The point then is that only those background or collateral conditions that are 'cotenable' with the antecedent should be admitted. But which are they? Those conditions S, which are such that if X had been true, S would have been true, too. This is a counterfactual assertion and Goodman thought that this kind of circularity impairs the metalinguistic analysis of counterfactuals.

The second general problem with this view concerns the characterization of laws, which are indispensable for the *connection* between the antecedent and the consequent of a counterfactual. The key thought here is that some generalizations, although true, are unable to 'support' counterfactuals because they are accidental. *Example*: Compare the following two counterfactuals:

A. If x had been a golden sphere, its diameter would not have been more than one mile long.
B. If x had been a plutonium sphere, its diameter would not have been more than one mile long.

(A) is false, while (B) is true – we rightly suppose. And this is because there is a law of nature backing up (B), while the generalization related to (A) is merely accidental (intuitively: if we have had enough gold, we could build a sphere of it with the required diameter; not so with plutonium). So the general statements L that are part of the premises of the argument whose telescopic form is $X\Box\rightarrow Y$ must express *laws of nature* and not merely accidentally true generalizations. But how exactly are we to distinguish between laws and accidents? If we felt that laws are those generalizations that support counterfactuals, whereas accidents are those that do not (see the previous example), then we would move in a(nother) circle. So there is a need to look for ways to distinguish between laws and accidents that do not rely (in the first instance, at least) on their modal force (at least when expressed in their support of counterfactuals). When Goodman brought this problem to the attention of philosophers, the prevailing view of laws was that they are simply regularities (cf Chisholm, 1946); hence, the distinction between laws and accidents (which are regularities too) was taken to be mostly an 'honorific' distinction that is captured by the different epistemic attitudes we have toward them. For instance, laws are those regularities that are projected to the future or that are confirmable by their instances. Sellars (1956: p. 268), however, pointed out that even if laws are taken to be regularities, they are those regularities that are characterized by 'neck-sticking-out-ness,' where this characteristic is captured in the subjunctive mood: 'If this were an A-situation, it would be accompanied by a B-situation.' The counterfactual content of a law then is seen as a 'contextual implication' of a law-statement.

This idea of 'contextual implication' is captured by the supposition view of counterfactuals, which is akin to (although interestingly different from) the metalinguistic view, as this was developed by John Mackie (1973). According to this view, to assert something like $X\Box\rightarrow Y$ is to assert Y *within the scope of the supposition that X*. In other words, we suppose X and then we envisage various possibilities and consequences. This account brings to light the contextuality of counterfactual conditionals, which is not resolvable without some degree of arbitrariness: X did not happen; supposing that X did happen, what else do we have to assume or suppose? What features of the background (including laws and particular matters of fact) should we retain or change? There is no uniquely determined answer to this question, although contextual matters (including a fuller specification of the antecedent of the conditional) might (and as a rule do) help us. *Example*: Is the counterfactual: 'If I had let go of this stone, it would have fallen to the ground' true or false (or assertible or not assertible)? It depends on the context. There are certain conversational contexts in which it would be false to assert it, for example, if this were a precious stone and the owner was very careful with it, so if the stone were to be let go, she would have caught it in midair. *Another example*: Consider the following pair of counterfactuals: 'If Julius Caesar had been in charge of United Nations Forces during the Korean War, then he would have used nuclear weapons' and 'If Julius Caesar had been in charge of United Nations Forces during the Korean War, then he would have used catapults.' Only contextual assumptions can tell us which one, if any, and in what context, is true (or assertible).

The supposition view takes it that counterfactuals are *not* truths about possible words but are ways to express an attitude

toward a possible state of affairs made within the scope of a supposition. The cotenability problem is solved by a legitimate weakening of the cotenability condition: The cotenable premises are taken to be those that are *thought* to be cotenable in a certain conversational context. But how, within this view, can it be explained that laws support counterfactuals while accidents do not? The difference is not in the content of a statement expressing a causal law as opposed to the content of an accidentally true generalization. Rather, the difference is in the circumstances under which it is legitimate (or acceptable) to combine the supposition that X is the case (i.e., the antecedent of the counterfactual $X\Box\rightarrow Y$) with the law L as premises of the relevant argument whose conclusion is Y. Suppose that the sole ground for believing the law L (e.g., All Fs are G) is an enumeration of *all* actual instances (Fai&Gai) of L. Then adding the supposition X, that is, that a *further* a is F, removes the ground for accepting L. We can no longer draw the conclusion that this further a is G. Hence, we cannot assert the counterfactual $X\Box\rightarrow Y$. More generally, if the reasons for accepting L survive placing L within the scope of *the supposition that there are further instances of the law's subject term*, then we can say that the law supports the relevant counterfactual conditional. According to Mackie, the required reasons are ordinary inductive reasons, that is, good inductive evidence for the law. Good inductive evidence, in other words, is evidence for the 'neck-sticking-out-ness' of the law. As Mackie (1974: p. 203) noted, the evidence plays a *double role*. It first establishes inductively a generalization. But then, "it continues to operate separately in making it reasonable to assert the counterfactual conditionals which look like an extension of the law into merely possible worlds" (Mackie 1974).

An interesting related thought comes from Julius Weinberg (1951) who claimed the following: A counterfactual $X\Box\rightarrow Y$ is not best seen as the indicative statement (the statement of a generalization) $X \rightarrow Y$ plus some further antecedent conditions (including that X did not actually happen), but rather as asserting something about the evidence there is for $X\rightarrow Y$, that is, that there is evidence for, and no evidence against, the gereralization: for all X $(X\rightarrow Y)$. Hence, the additional strength a counterfactual is supposed to have over the corresponding generalization is captured by the evidence there is for the generalization.

The Possible Worlds View

Taking literally the view that counterfactuals are used in contemplating *possibilities*, the second major view of the semantics of counterfactuals appeals to possible worlds. In first suggesting this view, Todd (1964: p. 107) noted that when we allow for the possibility that the antecedent of a counterfactual be true, we are "hypothetically substituting a different world for the actual one." On this view, the core meaning of a counterfactual $X\Box\rightarrow Y$ is (roughly): In the possible (but not actual) world where X, Y too.

A possible world is a way the world might be or might have been. For instance, it is possible that gold is not yellow, or that planets describe circular orbits, or that birds do not fly, or that beer does not need yeast to brew. But are there really possible worlds? There are three views here. The *first* is that talk of possible worlds is a mere *facon de parler*, although useful when it comes to assessing counterfactuals (cf Mackie, 1974: p. 199). (I take it that an extension of this view is that possible worlds are useful *fictions*.) The *second* is 'extreme realism,' according to which the way the world *actually* is, is one among the many ways the world could be; hence, the actual world is one among the many possible worlds, the latter being no less real than the actual. The chief advocate of this view was David Lewis (1973). The *third* view is 'abstract realism,' according to which possible worlds are maximally consistent sets of propositions: total ways things might be. A 'possible world' then is fit to represent a concrete reality, but only one of them actually represents anything, that is, the actual world (cf Bennett, 2003).

Stalnaker (1968) developed the core meaning of counterfactuals as follows:

Consider a possible world W in which X is true but otherwise is similar to the actual world @. $X\Box\rightarrow Y$ is true if Y is true in W.

The similarity relation among worlds (a selection function, as Stalnaker put it) is an ordering of possible worlds with respect to their resemblance to the actual world.

Calling an X-world a possible world in which X hold, counterfactuals might be taken to be strict conditionals of the following form:

$X\Box\rightarrow Y$ is true in a world W if Y is true in all X-worlds such-that _____

where the blank is filled by a general condition that X-words should satisfy. Hence, whatever goes into the blank _____ places a restriction on the admissible (or accessible) possible worlds. This idea would model counterfactuals along the lines of strict conditionals of the form:

it is physically necessary that _____
or
it is logically necessary that _____

where the first restriction is to all worlds with the same laws as the actual, whereas the second 'restriction' would be to all possible worlds *simpliciter*.

But this analysis cannot be correct. There is no set of possible worlds W such that $X\rightarrow Y$ throughout W (this is another way to state the fact that counterfactuals are non-monotonic). So Lewis (1973) suggested that counterfactuals $X\Box\rightarrow Y$ are *variably strict conditionals*: each of them is a strict conditional, that is, every X-world *of a certain sort* is a Y-world, but the relevant set of worlds varies with different conditionals.

Like Stalnaker, Lewis took it that worlds are ordered in terms of similarity, or closeness to the actual world. According to this *primitive* notion of 'comparative overall similarity': "we may say that one world is closer to actuality than another if the first resembles our actual world more than the second does, taking account of all the respects of similarity and difference and balancing them off against one another" (1986: p. 163).

But unlike Stalnaker, Lewis took it that in assessing the counterfactual $X\Box\rightarrow Y$, it does not make good sense to talk about *the* closest-to-actual possible X-world. It is not just that there might be more than one closest-to-the actual possible worlds. It is mainly that there might not be even one rightly

deemed *the* closest (even in a limiting sense). Hence, according to Lewis's view:

$X \Box\!\!\rightarrow Y$ is true at a world W if some (accessible) X-world in which Y holds is closer to W than any X-worlds that Y does not hold.

For instance, take the counterfactual that if this pen had been left unsupported (X), it would have fallen to the floor (Y). Neither X nor Y are true of the actual world. The pen was never removed from the table, and it did not fall to the floor. Take all X-worlds. The counterfactual $X \Box\!\!\rightarrow Y$ is true (in the actual world @) if the X-worlds in which Y is true (i.e., the pen is left unsupported and falls to the floor) are closer to @ than any of the X-worlds in which Y is false (i.e., the pen is left unsupported but does not fall to the ground, e.g., it stays still in midair). As Lewis (1986: p. 164) put it, "[A] counterfactual (...) is true if it takes less of a departure from actuality to make the consequent true along with the antecedent than it does to make the antecedent true without the consequent."

The key idea behind the possible-world semantics is that in specifying the truth-conditions of a counterfactual conditional, we should imagine a state of affairs in which X obtains and which is such that all *else is pretty much as they actually were*. But as noted already, this is not quite possible. In the possible world in which X did happen, many other things (including the laws) were different from the actual world @ in which X did not occur. Can we find comfort in the notion of comparative similarity? Now, although 'comparative overall similarity' is not defined strictly, a lot can be said of it. Notably, it imposes a weak ordering on the set of possible worlds that are accessible from @, that is, the relation of comparative similarity is connected and transitive. (It also imposes a centering assumption: @ is closer to itself than any other world is to it.) More important, however, similarity is clearly not one dimensional, but rather it is the result of many component similarities. Lewis (1986: pp. 47–48) ranked possible worlds according to the following dimensions of similarity (put in order of importance):

- Avoid big, widespread violations of the laws of nature of the actual world (very important).
- Maximize the spatiotemporal perfect match of particular matters of fact.
- Avoid small, localized violations of the laws of nature of the actual world.
- Secure approximate similarity of particular matters of fact (not at all important).

So, a world W_1 that has the same laws of nature as the actual world @ is closer to @ than a world W_2 that has different laws. But insofar as there is exact similarity of particular facts in large spatiotemporal regions between @ and a world W_3, Lewis allows that W_3 is close to @ even if some of the laws that hold in @ are violated in W_3.

All this implies that there is quite a lot of vagueness in the notion of overall comparative similarity, which accounts for the fact that counterfactuals themselves are vague, at least in the sense that it is a contextual matter as to what to keep fixed and what to change when we assert a counterfactual conditional. A more serious worry relates to the issue of the motivation behind the foregoing ranking of dimensions of similarity among worlds. It has been observed by many that Lewis's initial theory yielded the wrong truth-values for a type of counterfactual conditional that can be schematized as follows:

$$X \Box\!\!\rightarrow \text{BIG DIFFERENCE}.$$

For instance:

(C) If the president had pressed the button, a nuclear war would have ensued.

Intuitively, (C) is true. But on Lewis's initial account, it would be false. For a possible world W_1 in which the president did press the button and a nuclear war did erupt is more distant from (because more dissimilar to) the actual world than a world W_2 in which the president did press the button but, somehow, a nuclear war did *not* follow. Addressing this worry, Lewis noted that intuitive judgments of the truth and falsity of counterfactuals are prior to the similarity relation that is required for the semantics of counterfactuals; hence, the similarity relation should be such that it tallies with the right intuitive judgments concerning counterfactuals. The similarity ranking is meant to solve this problem. To see how the foregoing counterfactual is indeed true, Lewis invited us to consider the following: Take a world W_1 in which nothing extraordinary happened between the president's pressing the button and the activation of the nuclear missiles. In W_1 the nuclear war did erupt. Take, now, a world W_2 in which the president did press the button, but the nuclear war did *not* follow. For this to happen, many miracles would need to take place (or, to put it in a different way, a really *big* miracle would have to occur). For all the many and tiny traces of the button pushing would have to be wiped out. Hence, appearances to the contrary, W_2 would be more distant from (because more dissimilar to) actuality @ than W_1. The *big* violation of laws of nature in W_2 is outweighed by the maximization of the perfect spatiotemporal match of particular matters of fact between W_1 and @. So, with the help of the refined criteria of similarity among possible worlds, the president counterfactual becomes true. Still, one may follow Horwich (1987: pp. 171–172) in wondering how psychologically plausible Lewis's theory becomes: The similarity criteria are so tailored that the right counterfactuals become true, but they have little to do with our pretheoretical understanding of judgments of similarity.

As noted in relation to the 'support' view, an adequate theory of counterfactuals has to jump two hurdles. The first relates to cotenability. Lewis solved this problem by taking it that some conditions S are cotenable with X (the antecedent of the counterfactual $X \Box\!\!\rightarrow Y$) if some X-world is closer to the actual world than any not-S world. The second hurdle relates to the distinction between laws and accidents. Here, the possible world approach is on safe ground, although the ground can support any decent theory of counterfactuals. David Lewis (1973) revamped a long tradition that goes back to John Stuart Mill, via Frank Ramsey, according to which the regularities that constitute the laws of nature are those that are expressed by the axioms and theorems of an ideal deductive system of our knowledge of the world, and in particular, of a deductive system that strikes the *best* balance between simplicity and strength. Simplicity is required because it disallows extraneous elements from the system of laws. Strength is required because the deductive system should be as informative as possible about the laws that hold in the world. Whatever regularity is not part of this *best*

system, it is merely accidental. The gist of this approach is that no regularity, taken in isolation, can be deemed a law of nature. The regularities that constitute laws of nature are determined in a kind of holistic fashion by being parts of a structure. An advantage of this approach is that it can sustain, in a noncircular way, the view that laws can support counterfactuals for it identifies laws *independently* of their ability to support counterfactuals.

A key objection to the possible world approach to counterfactuals is that counterfactual conditionals are not purely objective; an irremediably subjective element enters into the judgment of similarity (and arguably, to the distinction between laws and accidents). Not only are the truth-conditions of counterfactuals "a highly volatile matter" as Lewis (1973: p. 92) himself noted, but also which counterfactuals are true turns out to depend on various partly nonobjective judgments concerning similarity weights and conversational contexts. This objection, however, might not be as fatal as it first seems precisely because counterfactual conditionals should not be taken to be pointers to necessary connections, powers, and the like but rather (in either of the two theories we have examined) summaries of attitudes we have toward statements that are supposed to express a *connection* between a hypothetical antecedent and a consequent. To exploit an idea of Sellars's, the core idea behind counterfactual reasoning is to assert that there are not good inductive reasons to affirm simultaneously a generalization and the physical possibility of an exception to it.

Counterfactuals and Evidence

Whichever way counterfactuals are treated, a particularly acute problem seems to arise, especially for empiricists: How are they connected with the available evidence, which can be gathered only in the actual world? To put the point differently, there cannot be an empirical test for a counterfactual by realizing its antecedent. The supporters of the 'metalinguistic view,' especially Mackie, avoid this problem by taking the counterfactual claim to be licensed by good inductive evidence. The advocates of the possible worlds approach pin their hopes on the role laws play in the truth-conditions of the counterfactuals. In particular, the thought is that although counterfactuals are about other possible worlds, these are *very much like* the actual world; hence, the evidence gathered in the actual world can be evidence for the truth of a counterfactual, by supporting the laws that back a similarity claim between the actual world and the world in which the antecedent of a counterfactual obtains.

Rubin's and Holland's Model

A promising model of counterfactuals, which aims to offer empirical test-conditions for them, has been developed by Donald Rubin (1978) and Paul Holland (1986) and has attracted increasing interest among statisticians and social scientists. This model focuses on the discovery of the effects of causes. Suppose, to use a simple example, we want to find out whether taking an aspirin makes a difference to a *specific* subject's relief from headache. We would like to give a certain subject u an aspirin to see what happens to the headache episode – let's call the result Y. Ideally, we also would like, *at the same time*, to withhold giving aspirin to the very same subject u, to see what happens to the headache episode – let's call this result Y'. The difference, if any, between Y and Y' naturally would be considered the causal effect of aspirin-taking on the headache episode of subject u. But this kind of experiment is impossible: The experimenter cannot give and *not* give an aspirin to the *same* subject u at the *same* time. Rubin's and Holland's main idea is that an appeal to counterfactuals allows us to make an inference about the causal effect.

Let's consider a population U of individuals, or units, $u \in U$. In a typical experiment, the experimenter applies one treatment, say i, out of a set of possible treatments T, to each unit u and observes the resulting responses Y. The experimental units are chosen and separated into two groups (the experimental group and the control group) by randomization. To simplify matters, let the treatment set T consist of two possible actions (treatment – t, and control – c). For instance, t may be taking the aspirin and c may be taking a placebo. Let Y also consist of two possible responses, for example, headache relief – Y_t, and headache persistence – Y_c. Although it is crucial that each unit is potentially exposable to any one of the treatments, to each unit u just one treatment is *actually* given, that is, either t or c. Similarly, for each unit u, there is just one response that actually is observed, that is, either $Y_t(u) = Y(t, u)$ or $Y_c(u) = Y(c, u)$. Rubin's model defines the two responses in subjunctive–counterfactual terms. $Y(t, u)$ is the value of the response that would be observed if the unit u were exposed to treatment t and $Y(c, u)$ is the value that would be observed *on the same unit* u if it were exposed to c. A key assumption of Rubin's model is that both values $Y(t, u)$ are $Y(c, u)$ are well defined and determined. In particular, it is assumed that even if subject u actually is given treatment t and has response $Y(t, u)$, there is still a fact of the matter about what the subject's u response would have been, had she been given treatment c. The task is to figure out the *individual causal effect*, that is the difference

$$\tau(u) = Y(t, u) - Y(c, u), \qquad [1]$$

which measures the effect of treatment t on u, relative to treatment c.

In each particular experiment, either $Y(t, u)$ or $Y(c, u)$ (but not both) ceases to be counterfactual. Yet, given that one of $Y(t, u)$ and $Y(c, u)$ becomes testable, the other *has to* be untestable. Holland has called a situation such as this "the *fundamental problem of causal inference*." Does it follow that figuring out eqn [1] is impossible?

Suppose that we give treatment t to u and we observe $Y(t, u)$. The question then is how could we possibly figure out the counterfactual value of $Y(c, u)$? According to the present model, when *certain assumptions are in place*, there are ways to assess counterfactuals such as the above. Here is how:

Given that unit u got treatment t, we may try treatment c to a *different* unit u', which is very much like u, except that it was given treatment c instead. That is, instead of testing the counterfactual conditional $Y(c, u)$, which is impossible, we test the indicative conditional $Y(c, u')$ – the response of unit u' if she is given treatment c – and claim that this tells *indirectly* what the value of $Y(c, u)$ is. For this move to be plausible at all, we need an assumption of *unit homogeneity*: that u and u' are so similar

that the actual response of u' to treatment c is the same as the response that unit u would have to treatment c. Under this assumption, we take it that $Y(t, u) = Y(t, u')$ and $Y(c, u) = Y(c, u')$. Then, the individual causal effect can be calculated, since eqn [1] becomes thus:

$$\tau(u) = Y(t, u) - Y(c, u) = Y(t, u) - Y(c, u'). \qquad [2]$$

Although eqn [1] involves essentially a counterfactual conditional ($Y(c, u)$), eqn [2] does not. Eqn [2] is indeed testable, but the counterfactuals are gone. Instead, eqn [2] has two indicative conditionals, one for unit u who received treatment t and another for unit u' who received treatment c. In a sense, the unit homogeneity assumption renders the counterfactual conditional $Y(c, u)$ not so much a claim about the *specific* unit u but rather a claim about *any* of the homogeneous units. It is because of this fact that the counterfactual is supposed to become testable.

We might proceed in another way to calculate $\tau(u)$. Instead of giving treatment t to unit u and treatment c to (uniform) unit u', we give treatment c to unit u at time t_1 and treatment t to the *very same unit* u at a later time t_2. This move requires another assumption, that is, *temporal stability* or the constancy of response over time. It also requires an assumption of 'causal transience' to avoid situations like this: The subject's taking a placebo at time t_1 changes some properties of her enough to affect her response to taking an aspirin at a later time t_2. Under these assumptions, we take it that $Y(t_{t_1}, u) = Y(t_{t_2}, u)$ and $Y(c_{t_1}, u) = Y(c_{t_2}, u)$. If this is so, then the individual causal effect can be calculated, because eqn [1] becomes:

$$\tau(u) = Y(t, u) - Y(c, u) = Y(t_{t_2}, u) - Y(c_{t_1}, u). \qquad [3]$$

The remarks made about eqn [2] can be repeated about eqn [3], too. Eqn [3] has no counterfactuals and it seems that the content of eqn [1] – which does involve the counterfactual $Y(c, u)$ – *reduces* to the joined content of two indicative conditionals $Y(t_{t_2}, u)$ and $Y(c_{t_1}, u)$ together with the two further assumptions of causal transience and temporal stability.

The key point then is that the alleged testability of counterfactual conditionals is predicated on the plausibility and success of certain general assumptions noted previously. These assumptions might fail. If, however, there are reasons to believe they do not, that is, if there is *evidence* for the general assumptions, then causal inference seems quite safe. I would suggest that these assumptions are characteristics of *stable causal-nomological structures*. Consider *unit homogeneity*. For it to hold, it must be the case that two units u and u' are alike in all causally relevant respects other than treatment status. If this is so, we can substitute u for u' and vice versa. This simply means that there is a causal law connecting the treatment and its characteristic effect, which holds for all homogeneous units and hence it is independent of the actual unit chosen (or could have been chosen) to test it. In effect, this holds for temporal stability too, since the latter is the temporal version of unit homogeneity.

Interventionist Counterfactuals

James Woodward recently has introduced the claim that only counterfactuals that are related to *interventions* can be of help when it comes to assessing their test or assertibility conditions. An intervention gives rise to an 'active counterfactual,' that is, to a counterfactual whose antecedent is made true by (hypothetical) interventions.

Woodward (2003: p. 3) characterized the appropriate counterfactuals in terms of *experiments*: They "are understood as claims about what would happen if a certain sort of experiment were to be performed."

Take Ohm's law (that the voltage E of a current is equal to the product of its intensity I times the resistance R of the wire) and consider the following two subjunctives:

1. If the resistance were set to $R = r$ at time t, and the voltage were set to $E = e$ at t, then the intensity I would be $I = e/r$ at t.
2. If the resistance were set to $R = r$ at time t, and the voltage were set to $E = e$ at time t, then the intensity I would be $i^* \neq e/r$ at t.

According to Woodward, we can perform the experiments at a future time t^* to see whether (1) or (2) are true. If, however, we are interested in finding out what *would* have happened, had we performed the experiment in a past time t (although we never did), Woodward invited us to rely on the 'very good evidence' we have that the behavior of the circuit is stable over time. Given this evidence, we can assume that the *actual* performance of the experiment at a future time t^* is as good for the assessment of (1) and (2) as a *hypothetical* performance of the experiment at the past time t.

An obvious advantage of this approach is that the truth-conditions of (the right sort of) counterfactual conditionals are not specified by means of an abstract metaphysical theory, as in the possible worlds approach. But there is a residual tension in this view. Woodward (rightly) insisted that counterfactual conditionals have determinate meaning and truth-conditions independent of the actual and hypothetical interventions. So there is a distinction between truth-conditions and test-conditions for a counterfactual. But then there must be a way for counterfactuals to get their truth-conditions fixed independent of their test-conditions. It is not quite clear what this way might be. The unclarity is accentuated by the fact that if there is such a way to specify the truth-conditions of a counterfactual conditional independently of its test-conditions (which are related to hypothetical interventions), then this way will offer truth-conditions to counterfactuals that do not (or might not) have test-conditions at all.

What if we were to collapse the truth-conditions of counterfactuals to their test-conditions? One can see the prima facie attraction of this move. Because evidence-conditions are specified in terms of actual and hypothetical experiments, the right sort of counterfactuals (the active counterfactuals) *and only those* end up being meaningful and truth-valuable. But there is an important drawback. Recall the subjunctive assertion in (1). On the option presently considered, what makes (1) true is that its evidence-conditions obtain. Under this option, counterfactual conditionals lose, so to speak, their counterfactuality. (1) becomes a shorthand for a future prediction or the evidence that supports the relevant law. If t is a *future* time, (1) gives way to an indicative conditional (a prediction). If t is a past time, then, given that there is good evidence for Ohm's law, all that (1) asserts under the present option is that there has been good evidence for the law.

In any case, Woodward is keen to keep evidence- and truth-conditions apart. One option would be to tie the truth-conditions of counterfactual conditionals to *laws of nature*. It then is easy to see how the evidence-conditions (i.e., actual and hypothetical experiments) are connected with the truth-conditions of a counterfactual: Actual and hypothetical experiments are *symptoms* for the presence of a law. Be that as it may, it is hard to see how a counterfactual can be assessed without taking into account the evidence there is about laws of nature or other general assumptions.

Causation and Powers

Despite the various difficulties we have presented, counterfactual conditionals are a stable part of the philosophical arsenal. Their being modal in character has invited the thought that they can capture a special connection between the antecedent and the consequent – most typically *a causal connection*. According to a popular counterfactual analysis of causation, to say (roughly) that event *c* causes event *e* is to say that *e* is counterfactually dependent on *c*, that is, that if *c* had not happened, *e* would not have happened either. This idea goes back to David Hume, but has been developed by Lewis (1974) into a full-blown theory. The sufficiency part of the definition is straightforward: If two events *c* and *e* are actual, and *e* is counterfactually dependent on *c*, then *c* is the cause of *e*. *Example*: Let *c* be the actual short circuit and *e* be the actual fire. If it is the case that if *c* had not occurred, then *e* would not have occurred, then the short circuit is the cause of the fire. But causation is transitive, whereas (as we have seen) counterfactual dependence is not. *Example*: Let *e′* be an effect of the fire *e*, for example, that the owner of the burnt house got some insurance money. If *c* causes *e* and *e* causes *e′*, then *c* causes *e′*. Although the owner's insurance compensation (*e′*) is counterfactually dependent on his house getting fire (*e*), which, in turn, is counterfactually dependent on the short circuit (*c*), *e′* is not counterfactually dependent on *c*: The owner would have got the insurance compensation (*e′*) even if the short circuit (*c*) had not occurred, assuming that the fire was caused in some other way. To make counterfactual dependence a necessary condition for causation, Lewis introduced a way to enforce the transitivity of counterfactual dependence: The sequence of events must form a *causal chain*. A sequence of events $<c, e, e', \ldots>$ is a chain of causal (counterfactual) dependence if *e* causally (counterfactually) depends on *c*, *e′* causally (counterfactually) depends on *e*, and so on.

This position, intuitively compelling though it may be, faces a number of important difficulties that has led to various ad hoc additions and modifications. (For a discussion of them see my 2002 paper.) The key point is that this approach explains how the effect *depends* on the cause without entailing anything as to how the effect is *connected* to the cause.

Those who think that there is a special kind of *connection* between cause and effect take counterfactual dependence to be a symptom of the presence of a power in the cause to bring about the effect: Causation amounts to a power's producing its manifestation. (This idea goes back to Leibniz who took it that causes are 'producers'). Actually, ascription of dispositions to objects (e.g., fragility or elasticity and the like) has been analyzed in terms of counterfactual (and subjunctive) conditionals. So, to ascribe a disposition F to an object *x* is to say that if *x* were to be given stimulus S, the characteristic result would be R. *Example*: *x* is fragile if and only if *x* were to be struck, it would break. Rom Harré and Edward Madden (1975) offered a general analysis of powers along the foregoing lines: *x* has the power to F if and only if *x* were subject to stimuli or conditions of an appropriate kinds, then *x* would do F, "in virtue of its intrinsic nature." So power-ascriptions to objects is analyzed in terms of (1) a specific counterfactual conditional and (2) an unspecified categorical claim about the nature of the object.

Despite its initial promise, this view faces important counterexamples, most of which point to the claim that the meaning of dispositional ascriptions cannot be captured by counterfactual conditionals. For it is possible either that the antecedent of the counterfactual is realized and the characteristic response does not obtain (e.g., because something else blocks the manifestation of the disposition) or that a disposition exists even if there are no manifestations of it (and hence no relevant counterfactuals to be entailed by it). Hence, the power-based attempts to understand the causal connection should either dissociate causation from counterfactuals or take "being disposed to" as a special (not further reducible) relation (see Mumford and Anjum, 2011).

Acknowledgment

This research has been cofinanced by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program 'Education and Lifelong Learning' of the National Strategic Reference Framework (NSRF) – Research Funding Program: THALIS – UOA – APRePoSMA.

See also: Causation: Physical, Mental, and Social; Causes and Laws: Philosophical Aspects; Counterfactual Reasoning, Quantitative: Philosophical Aspects.

Bibliography

Bennett, Jonathan, 2003. A Philosophical Guide to Conditionals. Oxford University Press, Oxford.
Chisholm, Roderick M., 1946. The contrary-to-fact conditionals. Mind 55, 289–307.
Goodman, Nelson, 1947. The problem of counterfactual conditionals. Journal of Philosophy 44, 113–128.
Harre, Rom, Madden, E.H., 1975. Causal Powers: A Theory of Natural Necessity. Blackwell, Oxford.
Holland, Paul, 1986. Statistics and causal inference. Journal of the American Statistical Association 81, 945–960.
Horwich, Paul, 1987. Asymmetries in Time. MIT Press, Cambridge, MA.
Lewis, David, 1973. Counterfactuals. Blackwell, Oxford.
Lewis, David, 1986. Philosophical Papers, vol. II. Oxford University Press, Oxford.
Mackie, J.L., 1973. Truth, Probability and Paradox. Clarendon Press, Oxford.
Mackie, J.L., 1974. The Cement of the Universe: A Study of Causation. Clarendon Press, Oxford.
Mumford, Stephen, Anjum, Rani Lill, 2011. Getting Causes from Powers. Oxford University Press, Oxford.
Psillos, Stathis, 2002. Causation and Explanation. Acumen, Chesham.
Rubin, Donald B., 1978. Bayesian inference for causal effects: the role of randomization. The Annals of Statistics 6, 34–58.

Sellars, Wilfrid, 1956. Counterfactuals, dispositions, and the causal modalities. In: Feigl, Herbert, et al. (Eds.), Concepts, Theories and the Mind-Body Problem, Minnesota Studies in the Philosophy of Science, vol. II. University of Minnesota Press, MN, pp. 225–308.

Stalnaker, Robert, 1968. A theory of conditionals. In: Rescher, N. (Ed.), Studies in Logical Theory. Blackwell, Oxford.

Todd, William, 1964. Counterfactual conditionals and the presuppositions of induction. Philosophy of Science 31, 101–110.

Weinberg, Julius, 1951. Contrary-to-fact conditionals. Journal of Philosophy 48, 17–22.

Woodward, James, 2003. Making Things Happen: A Theory of Causal Explanation. Oxford University Press, New York.

Counterfactual Reasoning, Quantitative: Philosophical Aspects

Alan Hájek, Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

After a general introduction, historical overview, and a brief discussion of the philosophical roles of counterfactuals and various accounts of them, we emphasize counterfactual reasoning within the quantitative frameworks of probability theory, decision theory, and game theory.

Counterfactuals: History and Background

Counterfactuals are a species of conditionals, although various authors have struggled to delineate exactly *which* species. They are typically sentences with the subjunctive form 'if it were the case that *A*, then it would be the case that *B*,' or 'if it had been the case that *A*, then it would have been the case that *B*,' or equivalents – for example, 'if Gore had won Florida, he would have won the election.' *A* is called the antecedent, and *B* the consequent. Some authors distinguish counterfactuals syntactically from indicative conditionals, regarding the subjunctive mood as the hallmark of the former and the indicative mood as the hallmark of the latter – for instance, 'if *A is* the case, then *B is* the case.' Others draw the distinction semantically, corresponding to two different ways of supposing something to be the case. Counterfactuals usually involve entertaining *hypothetical* or *nonactual* states of affairs: the antecedent is believed or presumed to be false, or contrary to the fact, but its truth is imagined or supposed. Counterfactual reasoning is thus a form of *modal* reasoning, kindred to reasoning about necessity or possibility, and usually in contrast to reasoning about the way things *actually* are, for which indicative conditionals are thought to be appropriate. (Usually, but not always, consider "if the patient had the measles, he would be showing exactly these symptoms" (Anderson, 1951).)

The philosophical study of conditionals goes back at least as far as the Stoics of ancient Greece, although their systems of logic apparently did not accord the counterfactual any emphasis. The rise in interest in counterfactuals has been comparatively recent in the history of philosophy. In the 1940s and 1950s, philosophers began to argue that counterfactuals are closely associated with various important concepts – for example, confirmation, induction, and laws of nature (Goodman, 1947, 1954). Since then, counterfactuals have been employed by some philosopher or other in analyzing most other modal concepts – for example, dispositions (Ryle, 1963), causation (Lewis, 1973), necessity and possibility (Williamson, 2005), and essences (Brogaard and Salerno, 2007). In the past couple of decades, a large literature has emerged on causal modeling, in which counterfactuals play center stage in the interpretation of causal graphs and structural equations (Pearl, 2000; Woodward, 2003). Counterfactuals also feature in the analyses of other concepts that involve some robust, non-accidental basis – often regarded as one of causation or at least stable covariation – such as perception (Lewis, 1980), knowledge (Nozick, 1981), and explanation (Woodward, 2003).

The significance of counterfactuals has also become increasingly appreciated in the social sciences – by psychologists (e.g., in understanding emotions such as regret), by historians (e.g., in studying the incremental contribution by some commodity to economic growth), in political science (e.g., in imagining that a party's rise to power would not have occurred without a certain pivotal event), and so on. In the law, the standard method of establishing a defendant's culpability for some harm is the 'but for' test, which invokes a counterfactual: but for the defendant's action, would the harm have occurred? Sociologists often appeal to counterfactuals in their causal analyses. It is no surprise that counterfactuals are a central area of study in linguistics.

However, some philosophers are suspicious of counterfactual analyses of other notions, and even of counterfactual themselves. Shope (1978) finds many such analyses to commit a 'conditional fallacy' – notably, neglecting cases in which realizing the antecedent of a counterfactual used to analyze such a notion interferes with realizing the notion itself. One may wonder what the 'truthmakers' for counterfactuals could be: in virtue of what are some true and some false? Some authors doubt that they even have truth values at all, as we will see.

This brings us to the so-called *problem of counterfactuals*, raised by Chisholm and Goodman in the 1940s: that of providing adequate (in particular, noncircular) truth conditions for the counterfactual. A widely agreed upon analysis of counterfactuals has proved elusive, and some philosophers are reticent about using them to do foundational philosophical work partly for this reason.

The Analysis of Counterfactuals

Let '$A \mathrel{\Box\!\!\rightarrow} C$' denote a counterfactual with antecedent *A* and consequent *C*. Fixing the truth value of *A* and *C* clearly does not fix the truth value of $A \mathrel{\Box\!\!\rightarrow} C$. For example, we may suppose that 'if I were to drop this fragile glass, it would break' is true, and 'if I were to drop this fragile glass, it would turn into a hamster' is false, yet the antecedents ('I drop this fragile glass') and consequents ('the glass breaks,' 'the glass turns into a hamster') are all false. This rules out as an adequate analysis the *material conditional* $A \supset C$ (defined to be equivalent to 'not-A or C'), which is automatically true when *A* is false.

Nor it seems can the counterfactual be analyzed simply as a *strict conditional* $\Box(A \supset C)$, a material conditional governed

by a *necessity* operator, familiar from modal logic: $A \supset C$ is true not just as a matter of fact, but necessarily so ('in all accessible worlds,' as this is usually understood). Certain inference rules that are valid for the strict conditional are apparently not valid for the counterfactual – for example, antecedent strengthening:

$A \mathrel{\Box\!\!\rightarrow} C$, therefore $(A \,\&\, B) \mathrel{\Box\!\!\rightarrow} C$ (for any B).

Consider: 'if I were to drop this fragile glass *and it were to land on a soft bed of feathers*, it would break.' This may well be false, while the original counterfactual (before the antecedent was strengthened with the italicized words) was true.

According to *metalinguistic* approaches, $A \mathrel{\Box\!\!\rightarrow} C$ is true if and only if A, together with suitable further premises (on one version, stating appropriate laws of nature and initial conditions), implies C. Mackie (1973) takes counterfactuals to be themselves elliptical presentations of such arguments, and thus lacking truth values altogether. The advent of possible worlds semantics for modal logics in the 1960s gave new direction to the analysis of counterfactuals. Truth conditions for them are proposed in classic works by Stalnaker (1968) and Lewis (1973).

Stalnaker motivates his account by the so-called Ramsey test for counterfactual reasoning: first, add the antecedent (hypothetically) to your stock of beliefs; second, make the minimal adjustments required in the rest of your beliefs to maintain consistency; finally, consider whether or not you then believe the consequent. Stalnaker's truth conditions are intended to apply generally to conditionals of the form 'if A, then C,' which Stalnaker denotes '$A > C$,' with just pragmatics distinguishing between indicative conditionals and counterfactuals according to him. He assumes that for each world w there is an ordering of worlds according to their 'closeness' to w – roughly, how little they differ from w (in appropriate respects). Call a world at which A is true an *A-world*; the proposition A is regarded as the set of A-worlds. His guiding idea is that

$A > C$ is true at a world w
if and only if
C is true at the closest A-world to w.

For example, we may say that at the world at which Gore wins Florida and which otherwise differs minimally from the actual world (e.g., he does not lose California, which he won in the actual world), he wins the election. Hence,

Gore wins Florida > he wins the election

is true according to the Stalnaker semantics.

Lewis rejects Stalnaker's assumption that, for any given w and A, there is exactly one closest A-world to w. Lewis's truth conditions are as follows:

$A \mathrel{\Box\!\!\rightarrow} C$ is true at w
if and only if
some AC-world is closer to w than any $A\overline{C}$-world, if there are any A-worlds.

For both Stalnaker and Lewis, the counterfactual is vacuously true if the antecedent is impossible – that is, there are no A-worlds. This yields results that some find unintuitive – for example, 'if 323 were prime, it would be divisible by 2, 4, and 16' seems to be *false* (Nolan, 2013). Cases like this motivate authors such as Nolan (1997) to extend the Stalnaker/Lewis analysis to 'counterpossible' conditionals, appealing to *impossible* worlds that may be closer or further from a given world. On the other hand, Fine (2012) rejects 'possible worlds' accounts of counterfactuals altogether (even restricted to possible antecedents), arguing that counterfactuals are not *intensional*: replacing an antecedent or a consequent with a logically equivalent sentence can change a counterfactual's truth value.

von Fintel (2001) offers a *dynamic semantics* for counterfactuals, according to which the utterance of a counterfactual $A \mathrel{\Box\!\!\rightarrow} C$ shifts the conversational context to one in which the accessible worlds include at least the closest A-worlds; the counterfactual is the strict conditional $\Box(A \supset C)$ *with respect to this newly altered context*. Leitgeb (2012a, b) gives a *probabilistic semantics* for counterfactuals: $A \mathrel{\Box\!\!\rightarrow} C$ is true at w iff the conditional chance of C, given A, at w is very high. 'Chance' is a probability measure that is objective, nonepistemic, and world-relative. This connects counterfactuals to the quantitative framework of probability theory, and in particular the notion of conditional probability. Other such putative connections have driven another important line of research, to which we now turn.

Probabilities of Counterfactuals

There has been much interest in the connection between *probabilities of conditionals* (of which counterfactuals are an important class) and *conditional probabilities*. Let P be the subjective probability function, so that $P(A \mathrel{\Box\!\!\rightarrow} C)$ is the degree of belief in $A \mathrel{\Box\!\!\rightarrow} C$. There are some reasons (linguistic intuitions, considerations stemming from the Ramsey test) for thinking that counterfactual reasoning is, or should be, subject to the following constraint:

(*) $P(A \mathrel{\Box\!\!\rightarrow} C) = P(C \mid A)$ for all A, C in the domain of P such that $P(A) > 0$.

$$P(C|A) = \frac{P(A \,\&\, C)}{P(A)}, \text{ if } P(A) > 0.$$

Stalnaker proposed a version of this identity for his conditional, which can be shown to vindicate his truth conditions above, and to vitiate Lewis's. In response, Lewis (most famously in 1976) proved a number of 'triviality results': under certain assumptions, the only classes of probability functions that can sustain (*) consist solely of members that take at most four values, and that are thus trivial. Since then, a number of authors have strengthened these results – see Eells and Skyrms (1994) for a collection of relevant articles.

One of the assumptions shared by these results is that conditionals have truth values, and correspondingly that the probability of a conditional is the probability of its *truth*. Edgington (2008) denies this, arguing that instead the probability of the counterfactual $A \mathrel{\Box\!\!\rightarrow} C$ is given by "your view about how likely it *was* that *C would have happened*, given that *A* had", a conditional probability relative to "a hypothetical belief state in a different context, normally that of an earlier time". She thus upholds a relative of (*), with suitable understanding of the probabilities on either side.

Moreover, an identity somewhat in the spirit of (*) *does* hold for the Stalnaker conditional (which *has* truth conditions). Start with probability function P. Among other things, P assigns probabilities to individual worlds. Define a new probability function P_A as follows: for each world w, P_A shifts the probability that P assigns to w to the closest A-world to w. P_A is said to be derived from P by *imaging on A*. Some authors model counterfactual reasoning by imaging (eschewing conditional probabilities as in (*)). Lewis (1976) shows that there is an intimate connection between the Stalnaker conditional and imaging:

$$P(A > C) = P_A(C) \quad (\text{for all } P, A, \text{ and } C)$$

The exploration of tenable variants of (*) is likely to continue to be a lively area of research.

Counterfactual Reasoning in Decision Theory

The distinction between probabilities of conditionals and conditional probabilities is important for *decision theory*, a normative theory of how one's beliefs and desires in tandem determine what one should do. Such a theory typically combines an agent's utility function u and probability function P to give a figure of merit for each possible action, called the *expectation*, or *desirability* of that action. Let $S_1, S_2, \ldots, S_n$ be a partition of possible states of the world. *Evidential decision theory*, as presented by Jeffrey (1966), measures the choiceworthiness of action A by a conditional probability-weighted average of the utilities of its possible outcomes (the conjunctions $A \,\&\, S_i$):

$$V(A) = \sum_i u(A \,\&\, S_i) P(S_i | A)$$

It thus prescribes that an agent act so as to maximize V. Various authors believe that evidential decision theory gives incorrect verdicts in Newcomb problem cases, in which an action is thought to be evidence for the obtaining of a given state, without in any way causing that state. This has led to the development of rival versions of decision theory, sharing the name *causal decision theory*, which replace the 'evidential' conditional probability weights with weights that are supposed to capture the agent's degrees of belief about the causal efficacy of the action in bringing about each possible state. Gibbard and Harper (1981), following Stalnaker, use probabilities of counterfactuals as the weights:

$$U(A) = \sum_i u(A \,\&\, S_i) P(A \Box\!\!\rightarrow S_i)$$

They then advocate the maximization of U. In an alternative formulation (e.g., Joyce, 1999), the probabilities of counterfactuals are replaced by probabilities of the states imaged on A. Proponents of the causal decision theory believe that it agrees with the evidential decision theory in unproblematic cases, and that it delivers intuitively correct answers in Newcomb problem cases.

Counterfactual Reasoning in Game Theory

Counterfactual reasoning can play a crucial role in the rational deliberation of two or more agents strategically playing against each other, their pay-offs depending on what they all do – the province of *game theory*. Paralleling the distinction we found in decision theory, Stalnaker (1996) distinguishes two sorts of counterfactual reasoning, *causal* and *epistemic* (the latter could also be called *evidential)*. In the former, agents reason about what would happen if they or others were to act in ways that they know or believe to be nonactual. In the latter, they consider how their own or others' beliefs would change if they were to learn things that they expect not to learn – for example, contrary to what is standardly taken to be common knowledge (and thus regarded as certain) by all the players, an irrational choice is hypothetically made. Indeed, such (epistemic) counterfactual reasoning leads naturally to some important refinements of Nash equilibrium, such as Selten's subgame perfect and ('trembling hand') perfect equilibrium.

Various models have been proposed of how an agent's beliefs are disposed to change in the face of unexpected actions by others, and the consequences of such dispositions for rational action – see, for example, Bicchieri (1988) (for extensive form games) and Stalnaker (1996) (for normal form games). Closely related is Shin's (1989) theory of equilibrium in normal form games. He gives a counterfactual-based account of rationality – a rational agent never acts in such a way that he would have done better had he acted differently – and proves that his account is equivalent, under certain assumptions, to those that employ game theory's standard solution concepts of Nash equilibrium and correlated equilibrium. Skyrms (1998) integrates the themes that we have discussed in this and the previous two sections. He gives a probabilistic theory of subjunctive conditionals that finds its inspiration in (*), with 'assertability value' replacing the 'P' on the left-hand side, and expectation of conditional chance replacing the right-hand side. He then uses this theory to give a unified treatment of counterfactual reasoning in individual decision theory, and in normal form and extensive form games.

Finally, more and more sophisticated probabilistic models have been developed for the representation of conditional beliefs in games. The economists Battigalli and Siniscalchi (1998), for example, construct a space of infinite hierarchies of conditional probability systems in which one can represent any statement about the players' dispositions to hold (arbitrarily high order) beliefs regarding the other players.

See also: Counterfactual Reasoning, Qualitative: Philosophical Aspects; Game Theory.

Bibliography

Anderson, A., 1951. A note on subjunctive and counterfactual conditionals. Analysis 12, 35–58.

Battigalli, P., Siniscalchi, M., 1998. Hierarchies of conditional beliefs and interactive epistemology in dynamic games. In: EUI Working Paper ECO 98/29. Badia Fiesolana, San Domenico.

Bicchieri, C., 1988. Strategic behavior and counterfactuals. Synthese 76, 135–169.

Brogaard, B., Salerno, J., 2007. A counterfactual account of essence. The Reasoner 1 (4), 4–5.

Eells, E., Skyrms, B. (Eds.), 1994. Probability and Conditionals. Cambridge University Press, Cambridge.

Edgington, D., 2008. Counterfactuals. Proceedings of the Aristotelian Society 108 (1 part 1), 1–21.
Fine, K., 2012. A difficulty for the possible worlds analysis of counterfactuals. Synthese. http://dx.doi.org/10.1007/s11229-012-0094-y.
Gibbard, A., Harper, W.L., 1981. Counterfactuals and two kinds of expected utility. In: Harper, W.L., Stalnaker, Pearce, G. (Eds.), Ifs. Reidel, Dordrecht.
Goodman, N., 1947. The problem of counterfactual conditionals. Journal of Philosophy 44, 113–128.
Goodman, N., 1954. Fact, Fiction and Forecast. Harvard University Press, Cambridge, MA.
Harper, W.L., Stalnaker, R., Pearce, G. (Eds.), 1981. Ifs. Reidel, Dordrecht.
Jeffrey, R.C., 1966. The Logic of Decision. Chicago University Press, Chicago, IL.
Joyce, J.M., 1999. The Foundations of Causal Decision Theory. Cambridge University Press, Cambridge.
Leitgeb, H., 2012a. A probabilistic semantics for counterfactuals. Part A. The Review of Symbolic Logic. http://dx.doi.org/10.1017/S1755020311000153.
Leitgeb, H., 2012b. A probabilistic semantics for counterfactuals. Part B. The Review of Symbolic Logic. http://dx.doi.org/10.1017/S1755020311000165.
Lewis, D.K., 1973. Counterfactuals. Blackwell, Oxford.
Lewis, D., 1976. Probabilities of conditionals and conditional probabilities. Philosophical Review 85, 297–315 (reprinted in Harper et al.).
Lewis, D., 1980. Veridical hallucination and prosthetic vision. Australasian Journal of Philosophy 58, 239–249.
Mackie, J.L., 1973. Truth, Probability, and Paradox. Oxford University Press, Oxford.
Nolan, D., 1997. Impossible worlds: a modest approach. Notre Dame Journal of Formal Logic 38 (4), 535–572.
Nolan, D., 2013. Impossible worlds. Philosophy Compass, 8: 360–372.
Nozick, R., 1981. Philosophical Explanations. Harvard University Press, Cambridge, MA.
Pearl, J., 2000. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge.
Ryle, G., 1963. The Concept of Mind. Penguin, London.
Shin, H.S., 1989. Two notions of ratifiability and equilibrium in games. In: Bacharach, M., Hurley, S. (Eds.), Foundations of Decision Theory. Blackwell, Oxford.
Shope, R.K., 1978. The conditional fallacy in contemporary philosophy. The Journal of Philosophy 75 (8), 397–413.
Skyrms, B., 1998. Subjunctive conditionals and revealed preference. Philosophy of Science 65, 545–574.
Stalnaker, R., 1968. A theory of conditionals. In: Rescher, N. (Ed.), Studies in Logical Theory. Blackwell, Oxford (reprinted in Harper et al.).
Stalnaker, R., 1996. Knowledge, belief and counterfactual reasoning in games. Economics and Philosophy 12, 133–163.
von Fintel, K., 2001. Counterfactuals in a dynamic context. In: Kenstowicz, M. (Ed.), Ken Hale: A Life in Language. MIT Press, Cambridge, MA, pp. 123–152.
Williamson, T., 2005. Armchair philosophy, metaphysical modality and counterfactual thinking. Proceedings of the Aristotelian Society 105, 1–23.
Woodward, J., 2003. Making Things Happen. Oxford University Press, Oxford.

Counter-Insurgencies and the Social Sciences

Juan J Navarro, UNCUYO, Argentina
Juan J Morales, CONICET, Argentina

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article takes a look at the contributions made by social sciences to the American counterinsurgency projects initiated during the Cold War in Asia and Latin America. Numerous research projects – many of them canceled before launch – were created to gather demographic, sociological, anthropological, and psychological data to contribute to 'developing' regions' social control.

After playing a peripheral role in World War I (WWI), primarily a chemists' and physicists' war, social sciences acquired higher scientific status and legitimacy as a result, among other reasons, of their use in World War II (WWII). Production planning and supplies' allocation, psychological recruiting and warfare, the use of media to raise both troops' and civilians' morale with propaganda provide but a few examples of social sciences' contributions during WWII. Military institutions profited greatly from sociology in this period. This social science contributed to such institutions by improving areas such as military labor selection and management issues, organizational effectiveness, motivation and morale, leadership, relationships between commands and chief-of-staff personnel, communication, organizational change, and unit efficacy criteria.

After the end of WWII, the race for world supremacy bore its impact not only on science but also on political and intellectual fields. The Cold War changed academic structures and discipline contents considerably. Disciplines concerned with social life aspects were encouraged to build on behavioral sciences in the search for social control and regulation. While, in the early 1940s, the war had been the focus of attention, throughout the (dawn) Cold War years, scientists focused on the revolution–counterrevolution–counterinsurgency triad. And, though the revolution had not been left aside, as it was a central Marxism phenomenon, starting in the 1950s, it garnered growing interest and concern, and it was even explored with greater scientific rigor (Bottomore, 1983).

Social sciences' contributions to counterinsurgency operations can be traced back to countless ventures initiated by former colonial powers. This dictionary entry takes a look at social sciences' contributions in counterinsurgency projects launched by the United States after the beginning of the Cold War. As a result, it also excludes the counterinsurgency efforts by the other dominant bloc after WWII, the Union of Soviet Socialist Republics, due to lack of access to Soviet and Russian archives and language barriers to use literature and documentary sources.

During the Cold War, the knowledge, experience, and fieldwork of social scientists – especially, anthropologists, sociologists, and psychologists – became remarkably important for military operations as well as for espionage and intelligence actions. Particularly, American social sciences stood at the service of their homeland's military apparatus, with a clear focus on foreign policy making. The overriding goal was to stop Communism around the world and to gain a better understanding of 'the Communist Mind' (Solovey, 2001a: 175).

Thus, social sciences turned to serve counterinsurgency, with close ties to decolonization and modernization processes in the so-called Third World during the second half of the twentieth century. (It is impossible to overlook an earlier instance in which military support and funding encouraged the development and growth of some social sciences. In the early twentieth century, anthropological studies garnered significant attention from military agencies wanting to gain a better understanding and insight into *cultural others*. Starting in the 1920s, it was common for many renowned anthropologists to collaborate with the U.S. Office of Strategic Services (OSS) – perhaps the most salient being British anthropologist Gregory Bateson, who collaborated with the American Army on several research studies, including the one in New Guinea that led to the publication of his book *Naven* in 1936. Bateson and his wife, Margaret Mead, conducted field studies, coproduced ethnographic films, and filmed photographic documentaries on Bali. Mead also collaborated with several U.S. military institutions, producing brochures for the Office of War Information and carrying out a study on the feeding and cultural habits of American citizens of foreign stock for the National Research Council. With Bateson and Geoffrey Gorer, she also helped the OSS establish a psychological training unit to prepare individuals for warfare in Asia and the Far West.) American academia largely viewed developing societies as characterized by persisting class, race, religion, or caste notions as well as irrational values. The fact that the political and military power could lure and coopt social scientists meant, in disciplinary terms, that high militarization and social control levels in these developing nations might lead to cultural, democratic, and political stability, while helping to secure strong economic growth. However, data were lacking, and, so, a number of quantitative, demographic, and geographic research programs were launched to raise enough information – figures, data, scientific, and empirical knowledge – to serve international social control. For the military, the promises brought by social sciences – either in terms of statistical analyses, by means of behavioral psychology studies, or anthropological reports on the cultures of the countries and peoples studied – hinged on the possibility to predict insurgent or revolutionary groups' actions and attacks. Thus, the prevailing practice intended to render social sciences much more empirical and quantitative.

More recently, several studies (Chomsky et al., 1997; Diamond, 1992; Simpson, 1998; Solovey, 2001a,b) take a look inside academia, turning them into study objects. These research efforts focus on the relationship between several sciences, universities, and the military – or what has also been called 'the construction of the military – scientific complex.' Thus, these analyses lies a number of questions on the ties between power and knowledge, politics and universities, state funding and the development of study or research fields, as well as the creation of research centers reporting to official, military, intelligence, or propaganda agencies. The intervention of security agencies changed academics and set the pace for its development, particularly in the early Cold War years, providing support to institutions, books, methodologies, programs, and disciplinary branches – including development studies, area studies, research on communications, and Operation Research (Simpson, 1998). This was the setting for a process of increasing respectability for social sciences in the United States, also supported by the construction of the so-called military–industrial–academic complex or the engagement of social sciences in the Cold War under military–government sponsorship (Solovey, 2001). In this process, the existence of a clear boundary between politics and scientific research did not exist – rather, it was actually and often blurred and undistinguishable.

Counterinsurgency in Asia and Vietnam's Centrality

One of the areas that received many counterinsurgency missions as a result of applied military anthropology was Southeast Asia. After WWII, this region became a very relevant economic, geopolitical, and strategic focus for the United States. Japan's reconstruction and its return to the regional economy turned Southeast Asia into a source for raw materials and market opening. It was deemed crucial for U.S. security to forge an agreement with Japan as an alternative to communist China's power in Asia. Indeed, operations were conducted in countries like the Philippines, Malaysia, New Guinea, and Indonesia. The bulk of America's presence and intervention in this Asian region would clearly start to unfold with the Vietnam War.

In the 1960s, U.S. counterinsurgency policies included programs for a type of economic, political, and social development intended to strengthen studied countries' social stability (Solovey, 2001: 195). The United States was back then – as it is now – the leading disseminating hub for the knowledge produced by social sciences (UNESCO, 2010). The theoretical–methodological framework developed by structural functionalism and abstract empiricism dominated that time with *modernization* standing out as prevailing paradigm.

Walt Whitman Rostow had published his *The Stages of Economic Growth: A Non-Communist Manifesto* in 1952, arguing that economic development unfolded in a number of stages that went from the basic economy of more traditional societies to the economy of scale of more developed, capitalist societies. Rostow was appointed national security advisor by President John F. Kennedy in January 1961. Shortly after his appointment, Rostow was asked to meet with Kennedy to discuss several threats to the U.S. national security and foreign policy associated with the Vietnam War. A few days later, Kennedy approved Vietnam's Counterinsurgency Plan.

American Secretary of Defense Robert McNamara characterized the conflict in Vietnam as 'the social scientists' war.' Indeed, throughout its duration, it was clear for the U.S. that a new approach was necessary to address Vietcong's resistance, and the use of social sciences was then deemed strategic. For instance, the Strategic Hamlet program was launched in 1964, initially to study Malaysia and the Philippines, but it later focused almost entirely on Vietnam. The program goal was to massively develop strategic villages and to strengthen farmers' settlements opposing the communist-led guerrilla forces. The underlying theoretical framework was based on a nation-building approach, as the plan hinged on expanding government services to these settlements, providing them with schools and roads as well as fostering local self-governance institutions, and promoting national values and symbols. To execute this program, Maynard Weston Dow, a U.S. Naval Academy graduate recruited by the Defense Department, collected sociogeographic data by means of household surveys for Vietnam's rural population. A while later, Dow himself submitted his doctoral thesis, under an U.S. Air Force Academy scholarship, entitled "Counterinsurgency and Nation Building: A Comparative Study of Post World War II Anti-Guerrilla Resettlement Programmes in Malaya, the Philippines and South Vietnam," at Syracuse University (Martin, 2013). Dow claimed that the Strategic Hamlet program was poorly managed and relatively ineffective in Vietnam (Dow, 1966), but it did account for an early American counterinsurgency attempt merging social sciences and a nation-building project. This type of programs was even viewed as potentially fruitful for other countries in Africa, Latin America, and Asia, when it came to collecting a large number of data on local populations and their loyalty or opposition to insurgent movements.

In 1966, the United States started its Civil Operations and Rural Development Support Program (CORDS) in Vietnam to bring peace to the country, joining military actions with economic and social programs based on theoretical models drawn from political sciences, sociology, and statistics. Specifically, the most outstanding contribution provided by CORDS consisted in a large amount of statistical data produced by the first information program dedicated to insurgent operations (Belcher, 2012: 261). These statistical data were used by the U.S. Army for its military interventions in Vietnam.

At that time, a program of psychological operations and interventions called *Chieu Hoy* (Open Arms) was also launched to persuade Vietcong fighters to desert. By 1967, some 75 000 desertions had taken place, but the program proved a failure, as the Vietnamese people and army was characterized by cultural and psychological resilience (Beckett, 2001). In addition, as revealed by Irving L. Horowitz (1972), a group of social scientists from Michigan received CIA funds to advice South Vietnam's police. At the same time, in 1967, the Advanced Research Projects Agency and the University of Pennsylvania were reported by students from Michigan and Cornell Universities for the development of Project Agile, a research initiative geared to counterinsurgency action in Thailand. In March 1970, a number of documents stolen from a university professor file surfaced, apparently implicating several social scientists in the U.S. counterinsurgency programs in Thailand.

These documents were delivered to the Student Mobilization Committee to End the War in Vietnam and were later published by *The Student Mobilizer* (Wolf and Jorgensen, 1970). A number of anthropologists and social scientists had allegedly gathered data for the U.S. Defense Department and for Thailand Royal house with a view to fostering a counterinsurgency program that would use development aid to ensure that tribal peoples remained loyal to Thailand's government (Council of American Anthropological Association, 1971).

Counterinsurgency in Latin America and the Paradigmatic Project Camelot

In Latin America's case, as politics grew increasingly more radical in some countries after WWII, the United States became more concerned about political developments inside its influence area. As a result, its attention, resources, and research efforts focused on this region. This scenario gathered momentum with the success of the Revolution in Cuba (1959), and the fear of immediate revolutionary guerrilla expansion – an exaggerated expectation that permeated both opposing sides – propelled the measures that had been brewing earlier. In addition to open and covert military interventions, the United States zeroed in on economic cooperation and counterinsurgency to fight communism.

John J. Johnson's *The Military and Society in Latin America* (1964) is a defense of U.S. political influence in favor of military intervention in Latin America. At the same time, a group of young officers – most graduated from West Point – was sent by the Pentagon to the Hilton Institute, a military base in South Carolina, to acquire strategic and sociological knowledge on Latin America to serve as the Pentagon's counterinsurgency 'experts.' Some of these officers would later stand behind the scenes at the military coups staged in the region (Chilcote, 1997).

This setting served as a backdrop for the creation of Project Camelot, a social science research study intended to measure, anticipate, and control internal conflict in peripheral countries. While it just illustrates the variety of research conducted in the Cold War and in the fight against insurrection, Project Camelot may very well be regarded as the most ambitious, straightforward, and crass example of the ties between social sciences and counterinsurgency as well as of the intervention and autonomy loss in science. This project started at American University's Special Operations Research Office and was commissioned by the U.S. Department of Defense to focus on Chile for a 4-year period. After Chilean–American anthropologist Hugo Nutini made several fruitless attempts to recruit Chilean social scientists for the initiative, Project Camelot was doomed to oblivion. When it seemed a thing of the past, the project made the headline at the leftist newspaper *El Siglo* as part of the United States' espionage. The news grew into such a scandal that Chile's political leaders were forced to create a Special Investigation Commission, reporting to the Chilean House of Representatives, to look into the matter.

Project Camelot's ultimate goal was to create knowledge to gauge the potential for domestic unrest, to estimate its effects, and to establish adequate control and normalcy restoration mechanisms in peripheral countries. The U.S. Army was endowed with the ability to "plan and lead operations involving a broad scope of the social and political issues that are part of counter-insurgency operations. Thus, the Army must develop doctrines based on an adequate knowledge of these areas" (Project Camelot Working Papers, 5 December 1964).

The Camelot scandal brought about a number of consequences. On the one hand, it led to the cancellation of similar projects across Latin America. As a result, Project Simpatico in Colombia and Project Colony in Peru did not take off. On the other hand, Camelot fueled a strong debate on external funding for Latin American social sciences. Suspicions about both outside financial support for disciplines and their utilization as espionage or cultural imperialism instruments turned into a myth of scientific research's direct dependence from the United States and its imperialistic 'uses.' These negative representations of science funding and use spread to any type of external support for scientific projects. This debate encompassed financial, structural, agenda setting, and research focus issues, as well as the role academics should play amid political radicalization processes. Its largely antiimperialistic rhetoric focused on the relationship between university and any American – public or private – agency. Philanthropic foundations, scientific institutions, universities, business companies, and government were viewed as cultural facades for the United States' security agencies.

Project Camelot was planned in Latin American social sciences' consolidation period, which came alongside an internationalization process – that is, as new places were created to house and develop them, these disciplines were partaking increasingly of the world's academic system. Counterinsurgency research projects emerged in a setting that welcomed these disciplines – Sociology, particularly in American science – and introduced them to the academic world, albeit without losing their peripheral standing. As these studies became public, a shadow of doubt was cast over all social sciences and their possible uses. However, at the same time, many social scientists were forced to stand up for their work. Sociologist Clodomiro Almeyda, an important Chilean socialist leader, tried to safeguard techniques when he noted that widespread rejection to Project Camelot "has an innocent casualty, smearing social research techniques – and, particularly, surveys" (Última Hora, 07/17/65: 2).

Some Final Remarks and a Brief Digression to the Present

Alas, this short historical review has shown how social sciences – particularly, anthropology, sociology, and psychology – have contributed to the counterinsurgency fight that the United States has waged in the second half of the twentieth century. A bulk of current literature has discussed these phenomena in a conspiracy-riddled tone. However, this analysis has chosen a more objective approach to depict these projects' interventionist and imperialistic nature. In any case, the engagement of social sciences in these studies was, for the most part, intended to produce statistical data – on demographics, geographical information, and on satisfaction/dissatisfaction with political regimes – that would contribute to anticipating and offsetting insurgency movements.

Counterinsurgency and academic knowledge militarization continue to be relevant topics today. The repercussions of the

events of 11 September 2001 marked the beginning of a new stage in the close relationship between counterinsurgency, social, and behavioral sciences, as shown by the U.S. military intervention in Iraq and Afghanistan as part of the so-called war on terror. Starting in 2003 in Afghanistan and in 2005 in Iraq, Provincial Reconstruction Tams, combining military officers and civilians, have been deployed to promote reconstruction efforts, to further security sector reforms, and to help expand central governments' authority in these countries (Marston and Malkasian, 2008).

In 2006, the U.S. Army launched its Human Terrain System, an US$-200-million program intended to combine anthropologists and other social scientists' work with combat units' daily tasks in Iraq and Afghanistan. The goal of the program was to collect data on these societies' human and cultural behavior, storing and disseminating this information to use it as leverage and combat weapons against insurgent forces. The Human Terrain System program was characterized as a mechanism to improve American military knowledge on local cultures and populations, using a software provided by the British firm BAE Systems. These measures replicate the vulgar terms of modern counterinsurgency, developing, creating, and producing statistical algorithms that try to predict insurgent actions but largely fail to capture key cultural, sociological, political, and psychological traits. As a result, these campaigns tend to be doomed, failing to accomplish their goals, as clearly shown by the U.S. failed attempts to sever ties between local insurgency and transnational terrorist networks.

When the U.S. Army and Marine Corps published their Counterinsurgency Manual in December 2006, Baghdad was in flames. Published with the University of Chicago's academic credentials, the manual was praised by think tanks and scholars for its fresh approach to war. Nonetheless, upon its debut in Afghanistan in 2009 under Barack Obama's administration, the doctrine proved to have little tactical success, wiping out villages and leaving a trail of dead civilians on its wake. These events call for a reflection on and a probe into the ethical and intellectual dilemmas brought about by the mobilization of human and social sciences in the field of counterinsurgency.

See also: Center–Periphery Relationships; Colonialism, Anthropology of; Imperialism, History of; Imperialism; Insurgency; Neocolonialism and Social Work; Third World.

Bibliography

Beckett, Ian, 2001. Modern Insurgencies and Counterinsurgencies. Routledge, London.

Belcher, Oliver, 2012. The best-laid schemes: postcolonialism, military social science, and the making of US counterinsurgency doctrine, 1947–2009. Antipode 44 (1), 258–263.

Bottomore, T.B., 1983. Sociology. A Guide to Problems and Literature. Blackie and Son Publishers, Bombay.

Chilcote, Ronald, 1997. U.S. hegemony and academics in the Americas. Under northern eyes: Latin American studies and U.S. hegemony in the Americas, 1898–1990. Latin American Perspectives 24 (1), Liberalism's Revival and Latin American Studies, 73–77.

Chomsky, Noam, et al., 1997. Cold War and Universities. Toward an Intellectual History in the Postwar Years. New Press, New York.

Council of the American Anthropological Association, May 1971. Statement on Ethics: Principles of Professional Responsibility. Recurso Electrónico: www.aaanet.org.

Diamond, Sigmund, 1992. Compromised Campus. The Collaboration of Universities with the Intelligence Community, 1945–1955. Oxford University Press, New York.

Della Faille, Dimitri, 2010. Dissection d'un discours à propos des conflics sociauz: les cas du Project Camelot. Revue canadienne d'études du développement 30 (3–4). Québec.

Della Faille, Dimitri, 2009. América Latina vista a través de la sociología estadounidense (1945–1970). Mimeo.

Dow, Maynard Weston, 1966. Nation Building in Southeast Asia. Pruett Press, Colorado.

Gouldner, Alvin, 1971. The coming crisis of Western sociology. American Journal of Sociology 77 (2), 312–323.

Garretón, Manuel Antonio, Murmis, Miguel, De Sierra, Jerónimo, Trindade, Hélgio, 2005. Argentina, Brazil, Chile, Mexico and Uruguay. Social science in Latin America: a comparative perspective (1930–2003). Social Science Information 44. London.

Gilman, Claudia, 2003. Entre la pluma y el fusil. Debates y dilemas del escritor revolucionario en América Latina. Siglo XXI, Buenos Aires.

Horowitz, Louis Irving, 1972. Foundations of Political Sociology. Harper and Row, New York.

Horowitz, Louis Irving, 1974. The Rise and Fall of Project Camelot. Studies in the Relationship between Social Science and Practical Politics, revised ed. The Massachusetts Institute of Technology Press, Massachusetts.

Horowitz, Louis Irving, 1964. The New Sociology: Essays in Social Science and Social Theory, in Honor of C. Wright Mills. Oxford University Press, New York.

Hounshell, David, 2001. Epilogue: rethinking the Cold War; rethinking science and technology in the Cold War; rethinking the social study of science and technology. Social Studies of Science 31 (2).

Johnson, John, 1964. The Military and Society in Latin America. Stanford University Press, Stanford.

Kuhn, Michael, Weidemann, Doris (Eds.), 2010. Internationalization of the Social Sciences. Asia – Latin America – Middle East – Africa – Eurasia. Transcrip Verlag, Bielefeld.

Lazarsfeld, Sewell, Wilensky, 1972. La Sociología en las instituciones. Editorial Paidós, Buenos Aires.

Mack, Andrew, 1975. Counterinsurgency in the third word: theory and practice. British Journal of International Studies 1 (3), 226–253.

Marston, Daniel, Malkasian, Carter, 2008. Counterinsurgency in Modern Warfare. Osprey Publishing, Oxford.

Martin, Geoffrey, 2013. Maynard Weston Dow (1929-2011) and geographers on film. Annals of the Association of American Geographers 103 (1), 1–4.

Navarro, Juan José, Quesada, Fernando, 2010. El Proyecto Camelot (1964–1965). La dependencia académica, entre el escándalo y el mito en BEIGEL, Fernanda (Directora) Autonomía y Dependencia Académica. Universidad e Investigación Científica en un Circuito Periférico (1950–1980). ISBN: 978-950-786-847-4. Editorial Biblos, Buenos Aires, pp. 145–167.

Needell, Alan, 1999. Project Troy and the Cold War annexation of social sciences. In: Simpson, Christopher (Ed.), Universities and Empire: Money and Politics in the Social Sciences during the Cold War. The New Press, New York.

Nisbet, Robert, 1966. Project Camelot: an autopsy. The Public Interest 5. New York.

Petreaus, David, 2006. Counterinsurgency Manual. University of Chicago Press, Chicago. When the US Army and Marine Corps published the Counterinsurgency Field Manual 3–24 (FM 3–24).

Project Camelot Working Papers, 5 December 1964.

Rohde, Joy, 2007. Counterinsurgency on Contract: Project Camelot, Social Science, and American National Security in the Cold War. University of Pennsylvania (Excerpt from Chapter Four of 'The Social Scientists' War': Expertise in a Cold War Nation, Ph.D. Thesis, University of Pennsylvania).

Selser, Gregorio, 1966. Espionaje en América Latina. El Pentágono y las técnicas sociológicas. Ediciones Iguazú, Buenos Aires.

Simpson, Christopher, 1998. Universities and Empire: Money and Politics in the Social Sciences during the Cold War. The New Press, New York.

Solovey, Mark, 2001a. Project Camelot and the 1960s epistemological revolution: rethinking the politics–patronage. Social science nexus. Social Studies of Science 31 (2).

Solovey, Mark, 2001b. Science and the state during the Cold War: blurred boundaries and a contested legacy. Social Studies of Science 31 (2).

Spenser, Daniela (comp.), 2004. Espejos de la Guerra Fría. Centro de Investigaciones y Estudios Superiores en Antropología Social, México.

UNESCO, 2010. World Social Science Report. UNESCO, París.

Última Hora, 17 July 1965.

Wolf, Eric, Jorgensen, Joseph, November 1970. Anthropology on the warpath in Thailand. New York Review of Books, 26–35.

Courts and Adjudication

Malcolm M Feeley, School of Law, University of California at Berkeley, Berkeley, CA, USA
Carlo Guarnieri, University of Bologna, Bologna, Italy

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M.M. Feeley, C. Guarnieri, volume 4, pp. 2878–2882, © 2001, Elsevier Ltd.

Abstract

This article explores the social logic of adjudication and locates it in context of other forms of legal dispute resolution. In so doing it relates adjudication within the political regime within which a system of adjudication is situated, pressures to settle rather than have a judge-imposed zero-sum solution, and the elusive concept of judicial independence.

Introduction

Adjudication is a type of dispute resolution that relies on an independent, third-party facilitator (Eckhoff, 1966; Shapiro, 1981). Such triadic relationships range from a mutual acquaintance who may serve as a go-between to help conflicting friends, to a powerful state which imposes a resolution on warring neighbors. Within the legal context there are three distinct types of dispute resolution triads: the mediator, the arbiter, and the judge. While all are engaged as third-party resolvers of disputes, they represent a continuum along which the freedom of the parties to the dispute is progressively limited. A mediator will facilitate communication between conflicting parties and perhaps propose a solution, but the parties themselves must agree to this solution. In contrast, an arbiter is a third-party decision-maker, selected and granted authority by the parties to impose a binding resolution. Adjudication is distinguished from both mediation and arbitration in several ways. Judges are usually externally appointed by the state and parties have no or very little control over who is selected to hear their dispute. Judges impose authoritative outcomes, but unlike arbiters, judges are expected to render their decisions according to preestablished norms. Furthermore, unlike mediation, adjudication does not require the active consent of the parties, though great efforts are made to secure consent or at least the appearance of consent.

Because their proceedings do not require the active consent of all parties, courts face a crisis of legitimacy. As a consequence, they must find substitutes for consent.

These substitutes can be understood as a set of principles (Cappelletti, 1989). The more important of these are the principle of party participation, the principle of the rule of law, the principle of impartiality, and the principle of independence. Indeed these principles have come to be understood as the defining characteristics of courts. The principle of party participation assures that those likely to be affected by a court's decision have an opportunity to make arguments and offer proofs in their behalf, but it also requires that the court respond with reasoned arguments to these concerns (Fuller, 1981). This principle may help account for the propensity of courts almost everywhere to encourage settlements in lieu of full-fledged trials and to value confessions of guilt or acknowledgments of liability. Although such practices have the obvious benefit of reducing the costs and risks of trial, they also provide – even if only symbolically – the active consent of the defendant to the process. The principle of the rule of law prohibits judges from making ad hoc decisions; it requires them to render decisions according to preexisting norms, the rules of the game which both parties have used (or are free to use) to organize their expectations and to arrange their affairs. Both party participation and the rule of law can be understood as approximations for consent. Participation provides parties with a voice and requires that their concerns be addressed; the rule of law presumes a shared set of norms or at least a set of rules under which all the parties operate. As we shall see, the rule of law is not a necessary condition for judicial independence, though it is closely associated with the notion of judicial independence (*see* Rule of Law (and Rechtsstaat); Criminal Procedure: Adversarial and Inquisitorial Legal Systems).

The principle of impartiality requires that judges have no interest in the outcome and no reason to favor one party over the other. This is usually achieved through the nature of the office; full-time judges, whose compensation is not dependent upon the outcome of the controversy, do not have any personal interest in the outcome of the controversy or ties to the parties of the dispute.

The related principle of judicial independence holds that the judge is not only independent of one or the other of the parties, but is also independent of other influences as well; the judge's full fidelity is to the rule of law. This poses an obvious and important challenge and may be the most difficult of the principles to implement. At times those not party to a dispute, such as an industry or trade association may have an interest in pressing for a particular outcome in a case or for a particular interpretation of a rule. More challenging still is that, as officials of the state, judges may be under pressure to adhere to political concerns rather than to the rule of law. Indeed, these two concerns are not always clear and separate. Almost everyone would agree that judicial independence is likely to be jeopardized if political authorities dismiss or force the resignation of judges with whom they disagree or if courts routinely uphold edicts of the regime that have not been duly enacted. A more subtle problem of judicial independence arises when the state enacts new laws, resists interpretative approaches of the judges, removes issues from the jurisdiction of courts, and seeks to influence judicial selection.

Similarly the issue of independence arises when judges make policy. Typically judges make policy when they confront

new situations not covered by existing doctrine or when existing doctrine contradicts deeply held personal or widely supported social norms. In such situations, courts may continue to apply outmoded norms. However, courts might alternatively create new doctrine through any of several institutional devices, such as the use of fictions (Fuller, 1981), reliance on metaphors, and reasoning by analogy (Levi, 1961), or treat vague and open-ended provisions in constitutions or statutes as grants of authority authorizing them to make new law (Feeley and Rubin, 1998). In any of these situations, judges can be criticized for abandoning the rule of law and hence for compromising their independence. Yet judges who insist on following outmoded or restrictive laws also run the risk of being criticized for failing to act independently and for being subservient to an unpopular regime or a new regime. Independence surely means more than freedom to apply a narrowly prescribed process of interpretation.

No one seriously argues that the modern state lacks authority to enact a wide variety of laws (though constitutional safeguards, conceptions of individual rights, and the like may severely constrain what the state may do and how it must enact laws). Still, the rule of law requires that laws be duly enacted, and publicized. At a minimum an independent judiciary is one which honors and upholds these procedural limits on state power. Still, governments, particularly new regimes following major political realignments or revolutions, may exert strong pressures on the judiciary to fall in line with the 'spirit' of the new regime. The new regime may be impatient with the slow pace of change through the rule of law or may react to what it perceives as a conservative judiciary wedded to the old order; it might exert pressure upon the judiciary to look at things in a new way. Such pressures can take any number of forms, including requesting judicial resignations, altering the procedures for judicial selection and advancement, selecting new judges whose views are aligned with those of the new regime, demoting judges deemed to be recalcitrant, or simply ignoring judicial actions that are deemed unacceptable. This is more difficult in democratic regimes since they are all more or less committed to respecting judicial independence.

Thus it would be a mistake to treat the issue of judicial independence as a problem faced only by authoritarian or totalitarian regimes (Ginsburg and Moustafa, 2008). In modern states everywhere, judges are governmental officials. In democracies, both the law and the constitution are understood to be products of popular will. Courts everywhere are likely to be affected by, shaped by, and at times in conflict with larger governmental processes in ways that raise issues about the independence of the judiciary. Indeed, since much of the legitimacy of modern government is closely tied to a belief in majoritarian principles and to the rule of law, conflicts between government and courts are all but guaranteed in all sorts of political regimes. This complex relation between law and government, adjudication and politics, and courts and governmental regimes is addressed in the sections below.

Judicial Independence

The quest for judicial independence in common law and civil law systems has taken quite different forms, though the differences may be due less to the particular legal forms than to the political systems in which they are situated. With the rise of absolutist states in early modern Europe and continuing well into the twentieth century, the judiciary was strongly integrated into the machinery of government, and courts were well-known as arms of the ruling regimes. With the rise of support for constitutional democracy in Europe, reformers sought to institutionalize the principles of the rule of law and judicial independence. One of the main objectives of constitutionalism is "to limit the arbitrary exercise of power and make it legally accountable" (Sartori, 1987). Although the rule of law is embraced in all constitutional systems, it has taken quite different forms. In countries with traditions of strong states (which historically have drawn their judiciaries into the service of the state), judicial independence has been pursued by integrating the judiciary into the structure of the civil service, by developing comprehensive codes of law so that laws are understood as bureaucratic-like rules, and by characterizing judges as technical experts. Judges are free to pursue their duties, but within these severe constraints. This conception of the separation of powers and judicial independence assumes the primacy of the legislature; the legislature represents popular will, and courts play a subordinate although independent role in finding facts and applying rules. Thus in strong states, judicial independence is purchased at the price of a narrow conception of the judicial role.

One can see this in the judiciaries of countries with a civil law tradition, where law is understood as a set of technical rules and judges are considered to be technicians (Merryman, 1985). And one can see it in some authoritarian states, where judges have sought to protect themselves by defining their roles narrowly and by insisting on strict interpretations of the law. But where the law involves politically salient issues – such as labor relations or economic regulations – governments everywhere are likely to remove jurisdiction from ordinary courts and establish specialized tribunals more directly responsive to political controls to deal with disputes. Thus in all countries, a large group of disputes between citizens and the state do not fall under the jurisdiction of independent courts but are assigned to specially created administrative tribunals that are more responsive to governmental direction. Similarly, issues of constitutional interpretation and judicial review, i.e., the power of courts to declare actions of other branches of government unconstitutional, have generally been removed from ordinary courts and assigned to more politically responsive bodies (Stone-Sweet, 2000). As a result, many – perhaps most – disputes between citizens and the state do not fall under the jurisdiction of ordinary courts, but are instead assigned to special administrative tribunals which are constructed precisely in order to be responsive to governmental interests.

Thus in disputes between individuals which are not likely to raise political controversy, judicial independence is protected, but where issues which may be politically charged, a more politically responsive judicial institution has been established. In a variation on this theme, Damaska has argued that, in regimes with a tradition of a strong or activist state, courts are designed and expected to implement government policy (1977), while in common law countries with traditions of weak states, judicial independence has been protected in a quite different way. Ironically, such common law political

systems with weak states tend to foster 'strong' courts. Indeed, the law – and courts – have been considered to be to some degree autonomous from the state (Feeley, 2013). In these countries the independence of law is protected not so much through a careful civil service-style judiciary and minutely crafted legal codes, but through a robust legal profession that has close associations with the judiciary. Although here too judges are officials of the state, they are also regarded as an integral part of civil society and are closely associated with an autonomous legal profession (Halliday et al., 2012). Constitutionalism and the rule of law are understood not only as procedural constraints on the government; they are external limits on the state as well. Ironically, this conception of the limited state produces a strong judiciary and an expansive notion of judicial independence (Damaska, 1986).

It is important to underscore some of the salient differences produced by these two conceptions of courts and of judicial independence. In Europe, independence is secured by subordinating the role of judges to the democratic principle – the will of the people expressed through parliament – and by insisting upon detailed codes so that judges have no need to make law but only to apply it (Guarnieri and Pederzoli, 2002). Typically this is reinforced by a career judiciary whose initial selection and subsequent promotions are based on seniority or merit. In contrast, common law systems entrust substantial law-making powers to the courts, since the law itself is conceived as something independent of the state and politics. Although there is an admitted political feature to judicial selection, judges are recruited from senior members of the bar, continue to maintain close ties to the bar, and are not dependent upon political authorities for support. Countries elsewhere have often adopted a hybrid of American-like and Continental institutions, though in all but former British colonies, the result has been much closer to civil law than common law style structures.

The differences between civil law and common law systems are familiar to anyone with some exposure to comparative law. Yet these differences only identify central tendencies; legal systems are too varied and too complex to be divided into any such neat categories. For instance, in continental systems even the most comprehensive code is not likely to anticipate all problems that may come within its scope; as a result a judge may have to make law to fill in the gaps. Similarly, law-making in modern common law countries is no longer primarily judge-made; modern law everywhere is increasingly statutory and administrative. Moreover, since the late twentieth century, the independence of civil law judges has been strengthened. In many countries, the government no longer controls judicial careers and judges have begun to play a more active role in law-making (Guarnieri and Pederzoli, 2002).

Furthermore, judicial independence – of either variety outlined above – is much more likely to occur in regimes with competitive political parties. Almost everywhere, one-party or predominant-party states have constrained the independence of the judiciary and sought to use it for their own political purposes (Ramseyer, 1994). Conversely, countries with competitive parties which alternate in controlling the government or where political parties are weak (e.g., Italy and Israel in the late twentieth century) are more likely to have independent judiciaries. Although all political victors may be tempted to convert the judiciary for their own use, those knowing that they may hold office only temporarily may resist the temptation in order not to jeopardize their long-term interests. Recognizing the harm that each party could impose on the others, parties in competitive political systems all have an incentive to maintain an independent judiciary (Landes and Posner, 1975).

The Business and Organization of Courts

The popular image of trial courts is that they try cases and resolve conflicts. But a careful examination of the process of adjudication reveals a more complex picture. The vast majority of all cases disposed of by courts – perhaps as many as 98% – are resolved by the parties themselves, without a full-dress trial. Indeed, in many instances, there is no expectation of a trial. In civil cases, court filings are often treated as signals to the opposing party that the other person is serious about pursuing his or her grievance. This is often enough to precipitate negotiations and resolution. Similarly, in criminal cases in both civil law and common law countries, most matters are disposed of without sustained judicial involvement in a trial, through unilateral prosecutorial decisions to drop charges, plea negotiations, or simply pleas of guilty. Indeed, in many countries, vast numbers of lesser criminal cases are disposed of through the mail without even a single court appearance by the accused (Feeley, 1979; Felstiner, 1979). In many of these situations, some combination of the prospects of the cost, time, and embarrassment of a trial is enough to encourage the parties to negotiate a settlement in the shadow of the courthouse (Mnookin and Kornhauser, 1979). Still, in such situations, the court may have been actively involved in the preliminary stages of the process, and may have handed down important pretrial rulings dealing with the scope of discovery, the admissibility of evidence, and the like. Such rulings provide cues to the participants as to the likely outcome of their case and thus precipitate negotiations and settlement rather than a prolonged trial process. In such cases, although the court did not formally resolve the conflict, it did nevertheless play an important role in its resolution.

The reasons for negotiations in the shadow of the law are fairly obvious. Filing a case often provides each party with the first opportunity to determine the strength of the opposition. This experience leads weak plaintiffs to drop their cases or to try to settle, and weak defendants to admit liability quickly or to settle. Cases that continue to trial are likely to be close cases in which both parties believe that they stand a good chance of winning. Economists have formalized this model, hypothesizing that, all things being equal, defendants would win (or lose) just as often as plaintiffs in full-fledged trials (Priest and Klein, 1984). Although uncertainty, changes in legal rules, and the degree to which an initiated case is controlled by the court can affect this dynamic, in the long run a stable equilibrium should prevail.

At the outset of this article, we noted that one of the problems of adjudication, in contrast to mediation and arbitration, is that the parties neither resolve their own dispute nor select the third-party decision-maker and that this can generate an aura of illegitimacy around the judicial process. If true, this may account for the ubiquitous tendency for parties to settle in the shadow of the courthouse and for courts to foster

settlement (Shapiro, 1981). Despite this, the tiny fraction of cases that go to full-dress trial provides the public with its perception of the nature of courts and adjudication.

In most countries trial courts are divided into two levels: lower courts which handle minor issues and upper-level courts which are reserved for a handful of more important matters. Although a substantially higher proportion of cases are resolved without formal or full-fledged trial in the lower courts, most issues reaching upper-level courts are also resolved through negotiations. Indeed, when all negotiated settlements and uncontested proceedings at both levels of courts are considered together, one finds that only a tiny portion of cases in trial courts of any sort are resolved through full-fledged adjudication (Trubek et al., 1983). As noted above, however, one should bear in mind that adjudication is a process that involves several steps. A party to a case may agree to settle only after a vigorously contested pretrial hearing that clarifies issues and may point to likely outcomes at trial. Although such dispositions are formally recorded as dropped cases or negotiated settlements, the outcome was in fact the result of an adjudicated decision. This is not at all uncommon; many plea bargains in criminal cases occur only after a contested motion to suppress evidence. As trial procedures have become more complex over the past two centuries, it is not surprising that much important adjudication has been relocated to the pretrial process.

All modern judicial systems provide for some system of appeals. The scope of reasons for appealing a case varies widely in different legal systems, but it is safe to say that a primary purpose of appellate courts is to listen to those who lost at trial and those who claim that the trial court made a mistake. In some legal systems, an appellate court may only consider allegations of error by the trial court. In others, the appellate court can undertake something of a trial *de novo*, and can reconsider facts introduced at the original trial and, in some places, consider additional facts brought to light since trial. In all this, appellate courts service two functions: they can correct errors, and they provide a regime with the ability to oversee and supervise the activities of a highly decentralized system of lower courts. By securing only a handful of appellate court judges sympathetic to its aims, a ruling regime can gain considerable assurance that justice will be administered within the bounds of its tolerance. This may be why civil law systems – that is active, policy-implementing states – tend to have more appeals (Damaska, 1977).

The common law tradition (Llewellyn, 1960) has always acknowledged the vast powers of courts to make policy, but since World War II, courts worldwide have embraced a more activist policy making stance (Tate and Vallinder, 1995; Stone-Sweet, 2000). A primary cause for this is the expansion of judicial review, both in countries with long-standing histories of constitutionalism, and new countries with written constitutions that provide for judicial review. This has resulted in the emergence of a vast literature on the politicization of the judiciary (Ginsburg, 2003; Hirschl, 2004) and the judicialization of politics (Silverstein, 2009). The former is associated with various political issues, such as the war on crime, and with questionable judicial propensities to interpret narrowly rights of the accused and exercise sentencing discretion in tune to harsher popular sentiment, as well as rule in areas with high political saliency, such as expanding tort liability, increasing awards in accident claims, expanding the concept of implied warranties in contracts, expanding (or contracting) doctrine governing employment discrimination, class actions, and remedies in extended structural reform cases such as police use of force or conditions in prisons (Rosenberg, 2008; Epp, 2009; Farhang, 2010). The latter is associated with the extension of due process and related constitutional requirements to governmental agencies such as schools, police, prisons, and welfare agencies that have hitherto not been subject to such detailed scrutiny (Kagan, 2001; Feeley and Rubin, 1998; Silverstein, 2009). While these developments were generally initiated by courts in the United States, they have since spread to courts around the world, and the United States is no longer an outlier or even leader in these areas (Hirschl, 2004; Mate, 2012).

Conclusion

Adjudication is a triadic structure, involving two disputants and a judge who is charged with imposing a solution on the parties. The process is inherently unstable; the losing party is likely to feel that the judge sided with the other party. Courts have responded to this inherent instability in two ways. They have fostered judicial independence through the formalism of the rule of law ('Do not blame me, blame the law!') and everywhere courts tend to encourage parties to arrive at consensual or negotiated outcomes so that they do not have to impose a solution upon them. A cursory look at judicial statistics around the world reveals that courts are fairly successful at this. Indeed, there is now considerable concern about the 'vanishing trial' (Galanter, 2004). Still, it should be remembered that legal claims are often negotiated only after a court has issued a preliminary ruling that may all but signal the outcome after trial, and there is little evidence that precedent-setting appeals have declined. Ironically even as worldwide courts are reducing their adjudicative role in some areas as alternatives to adjudication and negotiated settlements have increased, they have expanded them in others. With the near universality of the embrace of the rule of law and judicial review, courts with greater independence in a great many countries have taken on increased political significance, in some instances handing down decisions that can either destabilize or reinforce the existing regime, and in others subjecting more features of government and the administrative process to more far reaching judicial controls (Kapiszewski et al., 2013).

See also: Common Law; Constitutional Courts; Disputes, Social Construction and Transformation of; Judicial Review; Litigation; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Plea Bargaining: A Comparative Analysis; Political Trials; Rule of Law (and Rechtsstaat); Supreme Courts.

Bibliography

Alter, K., 2009. The European Court's Political Power: Selected Essays. Oxford University Press, Oxford.

Alter, K., 2013. The New Terrain of International Law: Courts, Politics, Rights. Princeton University Press, Princeton.

Cappelletti, M., 1989. The Judicial Process in Comparative Perspective. Clarendon Press, Oxford.
Damaska, M., 1986. The Faces of Justice and State Authority. Yale University Press, New Haven, CT.
Eckhoff, T., 1966. The mediator, the judge, and the administrator in conflict resolution. Acta Socologica 10, 158–166.
Epp, C., 2009. Making Rights Real: Activists, Bureaucrats, and the Creation of the Legalistic State. University of Chicago Press, Chicago.
Farhang, S., 2010. The Litigation State: Public Litigation and Private Lawsuits in the U.S. Princeton University Press, Princeton.
Felstiner, W.L.F., 1979. Plea contracts in West Germany. Law and Society Review 13, 309–325.
Feeley, M.M., 1979. The Process Is the Punishment: Handling Cases in a Lower Court. Russell Sage Foundation, New York.
Feeley, M.M., Rubin, R.L., 1998. Judicial Policy Making and the Modern State. Cambridge University Press, New York.
Feeley, M.M., 2013. Judge and company: courts, constitutionalism, and the legal complex. In: Halliday, T.C., Karpik, L., Feeley, M.M. (Eds.), Fates of Political Liberalism in the British Post-Colony: The Politics of the Legal Complex. Cambridge University Press, New York.
Fuller, L., 1981. Forms and limits of adjudication. In: Winston, K. (Ed.), The Principles of Order: Selected Essays of Lon Fuller. Duke University Press, Durham, NC, pp. 91–134.
Galanter, M., 2004. The vanishing trial: an examination of trials and related matters in federal and state courts. Journal of Empirical Legal Studies 1, 459–570.
Ginsburg, T., 2003. Judicial Review in New Democracies: Constitutional Courts in Asian Cases. Cambridge University Press, New York.
Ginsburg, T., 2009. The clash of commitments at the international criminal court. Journal of International Law 9, 499–514.
Ginsburg, T., Moustafa, T., 2008. Rule by Law: The Politics of Courts in Authoritarian Regimes. Cambridge University Press, New York.
Guarnieri, C., Pederzoli, P., 2002. The Power of Judges: A Comparative Study of Courts and Democracy. Oxford University Press, Oxford.
Halliday, T., Karpik, L., Feeley, M., 2012. Fates of Political Liberalism in the British Post-Colony. Cambridge University Press, New York.
Hilbink, E., 2007. Judges beyond Politics in Democracy and Dictatorship: Lessons from Chile. Cambridge University Press, New York.
Hirschl, R., 2004. Toward Juristocracy: The Origins and Consequences of the New Constitutionalism. Harvard University Press, Cambridge.
Jacob, H., Blanckenburg, E., Krtizer, H., Provine, D.M., Sanders, J., 1996. Courts, Law, and Politics in Comparative Perspective. Yale University Press, New Haven, CT.
Kagan, R., 2001. Adversarial Legalism. Harvard University Press, Cambridge, MA.
Kapiszewski, D., Silverstein, G., Kagan, R., 2013. Consequential Courts: New Judicial Roles in Global Perspectives. Cambridge University Press, New York.
Keller, H., Stone Sweet, A. (Eds.), 2008. A Europe of Rights: The Impact of ECHR on National Legal Systems. Oxford University Press, New York.
Landes, W., Posner, R., 1975. The independent judiciary in an interest-group perspective. Journal of Law and Economics 18, 875–901.
Llewellyn, K., 1960. The Common Law Tradition: Deciding Appeals. Little Brown, Boston.
Mate, M., 2012. Priests in the temple of justice: the Indian legal complex and the basic structure doctrine. In: Halliday, T., Karpik, L., Feeley, M. (Eds.), Fates of Political Liberalism in the British Post-Colony. Cambridge University Press, Cambridge, pp. 112–148.
Merryman, J.H., 1985. The Civil Law Tradition, second ed. Stanford University Press, Stanford, CA.
Mnookin, R., Kornhauser, L., 1979. Bargaining in the shadow of the law. Yale Law Journal 87, 950–987.
Priest, G., Klein, B., 1984. The selection of disputes for litigation. Journal of Legal Studies 13, 1–23.
Ramseyer, M., 1994. The puzzling (in)dependence of courts: a comparative approach. Journal of Legal Studies 23, 721–759.
Rosenberg, G., 2008. The Hollow Hope: Can Courts Bring about Social Change? University of Chicago Press, Chicago.
Sartori, G., 1987. The Theory of Democracy Revisited. Chatam House, Chatam, NJ.
Shapiro, M., 1981. Courts: A Comparative and Political Analysis. University of Chicago Press, Chicago.
Silverstein, G., 2009. Law's Allure: How Law Shapes, Constrains, Saves, and Kills Politics. Cambridge University Press, New York.
Stone-Sweet, A., 2000. Governing with Judges: Constitutional Politics in Europe. Oxford University Press, New York.
Tate, N., Vallinder, T., 1995. The Global Expansion of Judicial Powers. NYU Press, New York.

Craft Production, Anthropology of

Esther N Goody, Cambridge University, Cambridge, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 5, pp. 2883–2888, © 2001, Elsevier Ltd.

Abstract

Social anthropology, grounded in participant observation, has addressed the topic of craft production during the twentieth century in several ways. The theoretical frameworks/approaches discussed here are concerned with: material culture and folklore; ethnographic accounts; Marxist studies; the cultural meaning of commodities; learning and apprenticeship; and craft production as 'niches' in the global economy. The latter is explored further by an examination of some of these niches in both developing and industrialized societies.

The familiar definition of 'craft' as a skill in making something, usually by hand, is essentially that used in anthropology. Political economists have distinguished between craftsmen who make things for use or sale to the consumer, and manufacture where things are made for merchants/capitalists to sell. Manufacture often involves several stages of production, either in the processing of raw materials (e.g., thread into cloth; weaving; dyeing) or of components (parts of a watch each made by a specialist craftsman; parts of a car). Before the industrial revolution the stages of manufacturing were coordinated and controlled by merchants or by the state, who had the necessary capital. While today manufacture refers to making things by machines, the literal meaning is 'handmade' (from Latin manus, 'hand' and facere, factum, to make). Pre-industrial manufacture involved coordinating the stages of handcraft production, and was widespread in some societies with complex economies (India, China, medieval Europe). Production of goods by hand today is often coordinated by middlemen (see below). Thus it is useful to include in craft production both the making of goods by skilled craftsmen, and the coordinated manufacture of handcrafted goods.

Craft Specialization, the Division of Labor, and Commodization

The division of labor by gender and age has been basic in human societies from the simplest hunter-gatherers through pastoralism, simple and more complex forms ofagriculture, to urban societies up until the industrial revolution, and remains important today. In small-scale societies where all adult men and women practice the same craft skills, these tend to be carried out as part of the domestic economy. Kin roles are also economic roles. In this 'domestic mode of production' (Sahlins, 1972), craft products are made for both use and exchange. Craft skills are learned from older same-sex kin in the course of assisting with craft production. The developmental cycle of the domestic group reproduces the unit of craft production.

As societies become larger and more complex, emerging urban centers can support craftsmen who specialize in a single product (iron knives, silk cloth, pottery for cooking). Repeatedly we find that specialized production is associated with refinements in technology. With a specialist division of labor, craft products become commodities as they are exchanged for other commodities or enter the market. Hart writes that the central feature of increasing economic complexity is the integration of the division of labor in a nexus of commodity production and exchange; and that there is a tendency towards the progressive incorporation of ever more aspects of social life as commodities. 'The commodity is human labor embodied in a good or service offered to society rather than be consumed by its producer. As such, com-modisation may be defined as the progressive abstraction of social labor' (Hart, 1982).

As the nexus of production and exchange becomes more elaborate, new patterns of the division of labor appear. This pattern can be seen in the independent towns of medieval Europe and the Mediterranean. In Europe in the Middle Ages, textile production for an expanding market became organized as manufacture, first in specialist guild workshops, and then in rural cottage industry which took two forms. Cottage-craftsmen and women raised or bought the raw materials which they spun and wove into cloth, and sold locally to merchant buyers. The cottage-craftsmen owned raw materials, tools, and the finished cloth, planned their own work, and dealt directly with the market, although vulnerable to its fluctuations. In the other form of cottage industry, merchant manufacturers supplied the thread, often the loom, and owned the finished cloth. Weavers were paid for their work, in effect selling their labors as cottage-laborers. Rural cottage craft production organized by merchant capital is an example of what has been termed 'proto-industrialization.' With the harnessing of water- and then steam-power the invention of machines to replace handcrafts skills became possible. Such machines brought workers into factories, but the organization of components of production in manufacturing remained in many ways the same (machines now did the work of spinning, warping, weaving, finishing, dyeing) (Goody, 1982).

Clearly, the industrial revolution has radically changed the conditions of craft production. It is no longer central to the production and exchange of commodities in industrial societies. Further, Western manufactured goods and industrial production have become integral to the economies of developing countries. Yet a number of niches of craft production remain. Social anthropology, grounded in participant observation, has addressed the topic of craft production during the twentieth century in several ways.

Theoretical Frameworks of the Anthropology of Craft Production

Material Culture and Folklore

The nineteenth century excitement with evolution was shared by anthropology, and carried over into the twentieth century with the Folklore school in which the study of material culture was central, especially in the United States and the Scandinavian countries. The focus was on detailed description and mapping of material culture, with a view of preserving a record of rapidly disappearing 'early' forms of society. Where craft production was treated it was within this framework.

Ethnographies in Social Anthropology

The ethnographies of this tradition took as problematic the social structure and functional integration of societies studied. Ethnographic accounts treat religion, kinship, economics, and political dynamics as all important and related. However social anthropologists differed in the weighting given to each domain, as did the societies they studied. Significantly, those ethnographers who treat craft production as a significant feature tend to have worked in relatively complex societies. Among them, Nadel's account of the kingdom of Nupe (Nigeria) is preeminent for its detailed accounts of a number of craft industries, and the way these relate to the Nupe court. Each craft is discussed in terms of the access to raw materials, technical processes, the organization of the division of labor within the workshop, authority and transmission of skills among craftsmen, design and fashions as these relate to consumer demand, and patronage by nobles and the court where this is relevant. Some crafts are 'open' in the sense that any man can learn by assisting a skilled worker (weaving, bead making); others are 'closed,' restricted to members of certain kin groups (blacksmiths, brass smiths). Craft skills are learned in informal apprenticeship. Closed crafts restrict apprenticeship to kin; this also serves to keep ritual and secrete knowledge within the craft kindered (Nadel, 1942). Miner's account of craft production in Timbuc- too in the same period is less detailed, but closely parallels that for Nupe (Miner, 1965 [1953]). Slave production of food stuffs and cotton was important, particularly in northern Nigeria, but does not appear to have been organized for craft production. However rural industry ofthe cottage craftsman kind in weaving and dyeing is documented around the city of Kano where certain villages were famous as centers of dyeing. The town of Daboya played a similar role as a center of dyeing in the precolonial textile industry of northern Ghana. (See also chapters by Pokrant and by Goody, in Goody, 1982.)

Marxist Studies of Craft Production

With its focus on modes of production, Marxist theory offers social anthropology a framework for the study of craft production. A mode of production consists of the productive forces (natural, material, and technical) and the social relations of production. Most work in this area is by French anthropologists, exemplified by an English collection of papers by French neo-Marxists (Seddon, 1978). The goal of this volume is to lay the groundwork for 'a truly scientific study of society … which will lead to the discovery of laws that underlie the development of … social and economic formations … (and) identify the relations of production characteristics of pre-capitalist formations' (Seddon, 1978: p. ix). There are chapters about self- sustaining agricultural societies, the peasant mode of production, the Asian mode of production, and an African mode of production. Based on detailed ethnography, most are rich and insightful discussions of the relations between economic and political factors, several foregrounding kinship in new ways; this is especially true of those in the section based on fieldwork in Africa. However, perhaps because their goal is the identification of general laws, craft production is not a focus of description or analysis.

Following the Seddon volume a number of anthropologists have written ethnographic studies of craft production from a Marxist perspective. Three can be noted briefly: Kahn's *Minangkabau Social Formations* (1980) is a study of village blacksmiths in Sumatra. He describes in detail technology, organization, and marketing – treating relations of production within the workshop, the pattern of learning craft skills, relations between blacksmiths, external influences such as imported materials, and tools. The analysis is set in the context of Minangkabau matriliny, a subsistence agriculture based on rice, and economic policies of colonial overlordship. Based on individual or small workshop production for a local market or to traders, this is petty commodity production.

The next example is a study of Scottish Hebridean Island weavers of Harris tweed. Ennew (1982) traces the development of this industry from the first sales of Harris tweed in 1881. From the beginning it was based on cottage-laborer production, with merchants supplying the wool, dyes, and often the loom, and deducting the cost from payment for the finished cloth. Demands from the external market (London, overseas markets) led first to the introduction of a wider, foot-powered loom, then to removal of warping to the spinning mills in order to control quality and patterns. 'From the moment of wool-purchase the production process of Harris Tweed is entirely in the control of mill-owning capital, except for the slight loss of control during the outworking [weaving] phase' (Ennew, 1982: pp. 189 – 190).

Very different relations of production characterize the manufacture of carved hardwood furniture in Hong Kong. In the factory which Cooper joined as an apprentice wood carver, identical sets of tables were carved for tourists and export. He speaks of the proletarianization of the industry. Workers were paid daily at an hourly rate; the factory products were standardized, carvers producing endless legs and tops for tables, with no involvement with design or contact with customers. However he does feel that seven months as a factory carver gave him insight into the Maoist distinction between manual labor and mental labor, into proletarian culture as well as ideology (Cooper, in Coy, 1989).

It seems that craft production is incorporated in several modes of production: petty commodity, cottage laborer, and even as proletarian labor.

The Rebirth of Material Culture Studies: The Cultural Meanings of Commodities

Material culture has again become a central concern for anthropology, but now the focus is on the cultural and personal meanings of things as commodities. Objects serve to

objectify cultural meanings in individual experience; individual identities are constructed through the consumption of commodities. Indeed it has been argued that the study of consumption and commodities represent a major transformation in the discipline of anthropology and may come to replace kinship as the core of anthropology. Appadurai, in his Introduction to *The Social Life of Things: Commodities in Cultural Perspective* (1986) reviews economic theories which focus on the production of commodities and argues, following Douglas and Isherwood, that the interesting questions are rather about why and how people consume goods. The factors that drive consumption (the 'demand' of classical economics) are to be found in socially and historically shaped cultural meanings – in the political economy of the social life of things. The Introduction concludes: 'Ever since Marx and the early political economists, there has not been much mystery about the relationship between politics and production. We are now in a better position to demystify the demand side of economic life' (Appadurai, 1986: p. 58). Given this perspective it is not surprising that among these anthropologists the few who are interested in production deplore the scarcity ofrelevant studies (but see chapters by Spooner and Bayly in Appadurai, 1986).

Learning and Apprenticeship in Craft Production

The learning of craft skills has consistently been central to craft production itself – with the domestic mode of craft production in simple societies, in apprenticeships in craft workshops, and again in the domestic mode of cottage industry production. Craft production is labor intensive, and the work of unskilled children and youths frees craftsmen to concentrate on more complex skilled tasks, assembly of components, and organization of both production and distribution to consumers. A Vygotskian view sees learning as based on working alongside an adept: observing, assisting with very simple task elements, gradually taking over responsibility for these elements while continuing to observe, and then assisting with more complex tasks. This model corresponds closely to the learning of weaving by fostered apprentices in the Daboya textile industry (Goody, in Goody, 1982). It also appears to fit accounts of domestic craft production by blacksmiths (Nadel, 1942; Kahn, 1980; Coy, 1989); it is consistent with formal apprenticeship as in industrial wood carvers (Cooper, in Coy, 1989) and Japanese folk potters (Singleton, in Coy, 1989). The contributors to Coy's edited volume on *Apprenticeship* (1989) are all anthropologists who did fieldwork while themselves apprenticed, and several learned handcrafts. Their work supports the view that in craft production the learner is central to the production process – by providing unskilled labor in labor-intensive production; by systematically enforced subordinate status which legitimizes labor demands and supports the expert's control of synchronizing elements of production; and by embodying the social reproduction of craft skills and knowledge.

Trajectories of Craft Production in a Global Economy

Before industrialization, virtually all artifacts – for daily use, exchange, luxuries – and the tools for making these, were the products of craft skills. With the industrial revolution, craft manufacture ('hand making') of most commodities has been replaced by machine production in factories. Yet today craft production still occurs in the many niches left by industrialization, sometimes in surprising places.

There are many craft production niches in both industrial and developing societies; this is an open set. The conditions under which handcraft production has advantages (of cost, of detail work, of skilled knowledge) are varied, and change in response to both local and global factors. Social anthropologists have studied craft production in some of these niches; other awaits research. Contemporary anthropology is responding to the complexity of globalization and postmodern skepticism of both theory and observation with a proliferation of particular interests. While none of the major current themes focuses on production, there are good individual studies in several of these niches. It is only possible here to sketch these briefly.

Niches of Contemporary Craft Production in 'Developing' Societies

Artisan/craft Production for Local Use in Developing Societies

In partially industrialized third world societies many commodities for local use are made by individual craftsmen and workshops, and in small factories, using handcraft tools and techniques. In Bolivia, tinsmiths make pails and basins, blacksmiths make farm implements and locks, others make kerosene lanterns, dancing masks, clothing, shoes, etc. (Buechler, in Coy, 1989). The largest single example must be the Indian Khadi cottage industry making handspun, hand woven cloth. Beginning in the ideology of self-sufficiency with which Gandhi led the movement for independence, this continues to be centrally organized. As with other attempts to make household craft production economically viable alongside industrial production, partially mechanized processes have been introduced to support handcraft skills (e.g., multiple-spindle spinning jennies are wired to electricity by villagers, Bayly, in Appadurai, 1986).

Artisan/craft Production in Developing Countries for the Global Market

There are two links to this market, tourists buying locally, and international buyers for stores in industrialized countries. The two are related, since tourists want to buy similar craft products on returning home, and consumers of craft producers in industrialized countries look for them as tourists. Social anthropologists have recently become interested in tourism. Among the studies which treat craft production are several in Graburn's *Ethnic and tourist arts* (1976); see especially the chapters by Graburn (on Inuit soapstone carving), Kent (Puebio and Navajo weaving in the American Southwest), and Salvador (on Mola applique blouses made by Cuna Indians).

Handwoven and hand-printed textiles have for centuries been among India's most successful export industries. In the twentieth century some firms concentrating on garment manufacture attempted to secure advantages of efficiency and control through factory production. However, there is a pattern of reverting to commissioning both weaving and making up of

garments from smaller units – cottage weavers and small workshops. Swallow's study of one such company finds that in a volatile export market the costs of factory production (maintaining plant and machinery, labor troubles) remain higher than those of cottage craftsmen and cottage laborers managed by intermediate capitalist merchants, under conditions of population pressure and lack of alternative unemployment (Swallow, in Goody, 1982).

An increasingly common pattern is merchant or factory production of crafts sold locally to Western buyers for retail in industrial countries. Best studied is the production and marketing of luxury goods such as oriental rugs/carpets (see Spooner in Appadurai, 1986) and carved hardwood furniture (Cooper, 1980). Less elaborate, cheaper craft products (simple jewelry, pottery, glass, textiles) are made by cottage craftsmen, cottage laborers, or in workshops for sale to wholesale buyers who market them in bulk to retailers at trade fairs in industrialized countries. These are finally sold to consumers in Western gift shops, department stores, and clothing stores (buyer, personal communication). Although this mode of craft production for final sale in the commodity-rich markets of the West probably far exceeds in volume that of luxury goods, it seems as yet not to have been studied by anthropologists.

Development Aid and Craft Production

International development agencies, nongovernmental organizations (NGOs), and national government agencies have seen craft production as a potential focus for income-generating projects, usually for women. This may mean building on existing craft skills; often crafts judged as marketable are unfamiliar locally, and have to be taught. Project effort is directed towards teaching skills, securing access to materials and tools, and marketing. The development of some form of local cooperative structure is often integral to the project. Although anthropologists often participate, little published material is available on such projects, except in the form of reports and evaluations.

Niches of Craft Production in Industrialized Societies

Outwork and Workshop Handcrafts in Industrial Societies

In industrialized societies craftsmen – here usually craftswomen – continue to work at home at piecework rates on a number of crafts: sewing on industrial machines (owned or rented), knitting (often by machine), and the making of small cheap goods like Christmas crackers. This modern cottage industry is cheaper for manufacturers for the familiar reasons: low rates of pay; a dependent labor force lacking employment alternatives; no need to maintain plant or machines; flexibility of labor force which is only given work when demand is high. Some manufacturers, particularly in textiles, bring craftspeople into workshops where sewing and knitting machines are provided and supervision is possible, making changes in design easier.

Production of Folk Crafts for the National Market

The 'Arts and Crafts movement' associated with William Morris in England in the second half of the nineteenth century had counterparts in the United States, and later Japan. This movement reacted to industrial production by romanticizing craft production, seeing it as preserving old techniques and a classical vision of beauty expressed in individual hand skills. In order to manufacture craft products so as to make them accessible to an enlightened market, Morris taught himself to weave, embroider, print textiles, make tied carpets, make glazed tiles, and block print designs used for wallpaper. Apprenticeship was initially rejected, since they were seeking a pure re-creation of the essence of each craft. Although it was Morris' dream to make craft products cheaply so all could afford them, the cost of production with skilled labor and costly materials meant they were always luxury goods.

This vision survived. It subsequently led to handcraft manufacturers using traditional patterns and materials but of relatively modest cost: of textiles (Harris tweed), pottery (exemplified by the craft of Bernard Leach), and in the United States of Quaker furniture. In Japan it was reflected in the deliberate creation of a word of 'folkcraft' *(mingei)* after discussions among artists and philosophers (including Bernard Leach), and the establishing of a national museum of folkcraft which worked especially with traditional potters to develop aesthetic and practical adaptations of old forms. This folk pottery was made in traditional rural workshops, and became commercially popular. In his study of several rural Japanese pottery-making communities, Singleton found that high demand led in some to an opening of apprenticeship to outsiders, with a decline in the authority of the master (Singleton, in Coy, 1989).

Persistence of Customized Craft Production

Factory production for mass consumption has become the major source of commodities in industrialized societies. Customised handcraft production continues alongside this, especially of men's clothing, women's fashions, shoes, and jewelry. Women's fashions have become international, with a few famous designers in Paris, Rome, New York, and Tokyo holding annual shows which set the latest fashion. At the luxury end of this mode of production the craftsmen designs a prototype made in his workshop; this may be commissioned by a customer, offered for sale through a show, or in exclusive shops. Successful designs may be copied in other workshops for sale in select stores; such as garment with a 'designer label' is still considered exclusive. A few designs are more widely copied and sold through chains of stores to the public. This 'trickle-down' of fashions through dressmakers' workshops has always linked dress to social mobility. Lace, from the sixteenth to the nineteenth centuries, moved from courts to town gentry to village bonnets. Maher describes this for Turin in the early twentieth century, where seamstresses copied the dresses made for customers to wear themselves – thus passing socially as ladies (Maher, 1987). Customized craft production also continues to be important, for instance in fine furniture, jewelry, pottery, and textiles. These are sold in specialist shops, and increasingly in shops alongside replicas of museum holdings. Their makers are coming to be known as designer craftsmen,' and also as 'artist craftsmen.' There is a debate, as with some of the ethnic crafts, as to whether this is 'art' or 'craft.'

Individual Craft Production for Pleasure in Using Craft Skill

The anthropological preoccupation with conceptualization of identity in terms of commodities has displaced earlier anthropological interest in occupational and personal identity based on skilled craftsmanship. Yet there is a resurgence of popular interest in learning craft skills like sewing, knitting, embroidery, lace-making, weaving, pottery, jewelry-making, etc. A number of indices reflect this enthusiasm: where they used to be learned from kin or in apprenticeship, such crafts are now taught in community centers and in summer workshops; practitioners form associations concerned with style and pattern, access to tools and materials, teaching and examining craft skills; there are magazines and books devoted to patterns and techniques; practitioners hold exhibitions of their work. The main motivation seems to be development and exercise of individual handcraft skills. Products are used by the maker or given as gifts. While occasionally a seamstress, weaver, or potter becomes skilled enough to sell skills or products on the market, for most commodization is not a factor. Indeed, individual craft production may be seen as the separation of craft skills from the market.

See also: Apprenticeship, Anthropological Aspects; Economic Anthropology; Economic Sociology; Labor History; Labor, Division of; Tourism; Work, Anthropology of; Work, History of.

Bibliography

Appadurai, A. (Ed.), 1986. The Social Life of Things: Commodities in Cultural Perspective. Cambridge University Press, Cambridge, UK.

Cooper, E., 1980. The Wood-carvers of Hong Kong: Craft Production in the World Capitalist Periphery. Cambridge University Press, Cambridge, UK.

Coy, M., 1989. Apprenticeship: From Theory to Method and Back Again. State University of New York Press, Albany, NY.

Ennew, J., 1982. Harris Tweed: Construction, retention and representation of a cottage industry. In: Goody, E.N. (Ed.), From Craft to Industry. Cambridge University Press, Cambridge, UK.

Goody, E.N. (Ed.), 1982. From Craft to Industry: The Ethnography of Proto-Industrial Cloth Production. Cambridge University Press, Cambridge, UK.

Graburn, N.H.H. (Ed.), 1976. Ethnic and Tourist Arts: Cultural Expressions from the Fourth World. University of California Press, Berkeley, CA.

Hart, K., 1982. On commodisation. In: Goody, E.N. (Ed.), From Craft to Industry. Cambridge University Press, Cambridge, UK, pp. 38–49.

Kahn, J., 1980. Minangkabau Social Formations. Cambridge University Press, Cambridge, UK.

Maher, V., 1987. Sewing the seams of society: Dressmakers and seamstresses in Turin between the wars. In: Collier, J., Yanagisako, S., Block, M. (Eds.), Gender and Kinship: Essays Toward a Unified Analysis. Stanford University Press, Stanford, CA.

Miner, H., 1965 [1953]. The Primitive City of Timbuctoo. Anchor Books, New York.

Nadel, S.F., 1942. A Black Byzantium. Oxford University Press, London.

Sahlins, M., 1972. Stone Age Economics. Akline, Chicago.

Seddon, D. (Ed.), 1978. Relations of Production: Marxist Approaches to Economic Anthropology. Frank Cass, London.

Creativity Training

Mark A Runco and Sarah E Sumners, University of Georgia, Athens, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Certain kinds of training purport to enhance creative performance and maximize creative potentials. Training may be short- or long-term and may include explicit directions or may be defined in terms of particular environments and contexts. Because creativity depends on cognitive processes that are not used unless an individual is motivated to use them, training often includes both affective and cognitive facets. The most effective creativity training efforts utilize most or all facets of the creative process. Although creativity can take many different forms, all creativity training assumes that there are in-born creative potentials and thus the most accurate way to describe training is the fulfillment of potential. This view implies that creativity training will have broad benefits, since potentials are universal. This article discusses training as enhancement and then reviews the cognitive, affective, and environmental training intentions and concerns. It concludes with educational recommendations.

One of the most important unanswered questions in the creativity literature is one that focuses on efficient and effective methods for the fulfillment of potential. This is the central question for educators who respect creativity, but it is also critical for organizations that wish to maintain or increase their innovation and productivity. It is relevant to studies of *positive psychology* and optimal development and related to both psychological and physical health, each of which seems to benefit from creative self-expression. In his book, *No Rootless Flower*, Frank Barron, one of the pioneers in the field of creativity studies, once remarked that creative expression and psychotherapy were complementary, the idea being that each could contribute to the other. Psychotherapy might help open the doors for creative self-expression. The more common method for fulfilling creative potential and performance involves systematic training, which is the focus of this article.

Although creativity can take many different forms, all creativity training has at its core the utilization of in-born creative potentials, and indeed the most accurate way to describe training (or education, or any effective treatment) is the *fulfillment of potential*. This indicates that training does not provide something that did not exist before; it merely releases and reinforces what the individual already possesses. It also acknowledges the fact that there are individual differences. Just as individuals vary in terms of their height or weight – each of these being the end result of an interplay of nature and nurture – so too do they differ in the potential for creative behavior. Not everyone could change the world with creative works, like Einstein or Picasso did, but everyone can fulfill potentials and reach the highest level of their creative talents – and thereby change their own worlds, even if not the entire globe.

Creativity is multifaceted. Scientific theories therefore refer to it as a *syndrome* or a *complex*. Most enhancement efforts focus on one single component rather than the entire complex. More often than not, enhancement also focuses on cognitive capacities and skills. This is one issue when considering the effectiveness of training because, although creativity often depends on particular cognitive operations and processes, these are not used unless the individual is motivated to employ them. Creativity is an affective, as well as cognitive phenomenon. Along the same lines, particular personality traits, such as openness to experience, may also be involved in creative work, and there are social, interpersonal, and contextual influences as well. This article focuses on the issue of enhancement and reviews various targets and a range of training methods, by carefully defining creative potential.

Training as the Enhancement of Potential

Defining training as the fulfillment of potential allows the recognition of all enhancement efforts, be they systematic or informal, or short or long term. Besides, by defining enhancement in this way, any controversy over the nature and nurture of creativity can be avoided. This follows from the idea that nearly all human potentials are both genetic *and* experiential, with experience (in the present context, the experience that leads to enhancement) working within genetic boundaries. Biologists refer to the *range of reaction*, with the range genetically determined, and the reaction within that range determined by the individual's experience. In this light, the enhancement of creativity merely fulfills an individual's genetic potential. It does not change the potential but insures that behavior is moved toward the high end of the range of potential. One might say that inborn talents are maximized. Some are born with greater potential than others; however, everyone has some capacity for original and creative expression. Given the importance of creativity for personal growth and social and technological evolution, we should consider carefully how best to fulfill the relevant potentials. As mentioned briefly above, one type of enhancement focuses on the cognitive basis of creative thinking and problem solving.

Cognitive Components

Various cognitive processes contribute to creativity. Some processes are more basic than others. They are purely cognitive (e.g., attention, memory, and information processing), while others involve higher-order thought (e.g., strategic problem

solving, metaphor, and divergent thinking). Metacognitive processes are frequently involved in cognitive theories of creativity because they refer to tactical and intentional operations, and creative work is usually both of these. Many creative persons make a habit of questioning assumptions, for example, or they may focus on doing what has not been done before. Some have as their goals personal paradigm shifts, their intent being to break new ground. They may stretch and extend ideas, approaches, or work in entirely new directions. They intentionally, and often strategically, target creative results. With the intentionality assumed by all strategic and metacognitive efforts, it is likely that the vast majority of all systematic enhancement training efforts attempt to improve creative thinking by providing appropriate tactics.

Table 1 presents a list of some commonly used tactics both in education and the business world employed to foster the creative process. Many of the tactics in Table 1 are interrelated. Many of them, for instance, involve some sort of shift in perspective. This is true of points (1)–(8), (12), (13), (18), (20), (23), (24), and (26) – and the benefit of several others may follow from a change in perspective. New perspectives disrupt mental sets and routines thereby minimizing the chances of fixity; often resulting in original ideas.

Most tactics (and training efforts) focus on originality, which is essential, although when taken alone not a sufficient indicator, for creativity. Keeping in mind that original insights are not always creative, consider points (9), (10), and (20)–(22). Some original insights are unusual, novel, or unique, precisely because they are worthless or inappropriate. They are not creative if only original. Creativity involves more than originality. Truly creative insights are original and somehow fitting or useful. Runco and Jaeger, in the 2012 *Creativity Research Journal*, referred to this as the *Standard Definition of Creativity*.

Table 1 Example tactics to improve creative thinking

1. Change your perspective.
2. Work backward, from the end of the problem (the solution) to the beginning.
3. Invert the problem. Turn it inside out, upside down.
4. Question your assumptions.
5. Get another opinion. Talk to someone from another field, or with a different perspective.
6. Incubate. Put the problem aside for a while.
7. Think like a kid. (What would your younger brother or sister say?)
8. Break the problem into small parts. Deal with the smaller versions.
9. Try new combinations.
10. Explore new connections.
11. Let your mind wander. Trust your intuition.
12. Be playful, foolish, uninhibited.
13. Use trial and error, especially the latter – and then learn from those mistakes.
14. Gather more data.
15. Think big! Think small! "Honey I shrunk the kids!"
16. Find an analogy. (The Wright brothers did it; they looked to birds!)
17. Find useful information or analogies in nature.
18. Visualize. Try 'the mind's eye.'
19. Change the medium. (Draw a picture. Make up a song about that problem.)
20. Explore the bizarre.
21. Be a 'contrarian.' Walt Disney was one – that's how we got Mickey Mouse.
22. Bend the rules! Dr Seuss made up how many new words?
23. Focus on quantity for a while. A high-quality idea may find a way out, eventually.
24. Borrow an idea and then make it better.
25. Ask questions. Ask: Why? How? Why not? What if?
26. Change the problem. Make it one that allows original solutions.

Many training efforts treat creativity as a kind of problem solving. Note, however, that some of the tactics given in Table 1 focus on a problem, and others focus on solutions. This is an important point because many creative ideas are attempts to solve problems. Others are creative precisely because they redefine a problem. With this in mind, it is not surprising that some tactics lead the individual to change the problem or its representation.

For the purposes of training it can be useful to define creativity as problem solving. This allows the training to be focused but it also means that at least two aspects of creativity are ignored. Not all creativity involves problem solving. Even when creativity is a kind of problem solving, the relationship is not a simple one. Problem solving is sometimes creative, but creativity may be manifested when the individual is not working on problems, or at least not solving them. The individual may have the need or urge for self-expression sans problems. From this perspective, creative thinking and problem solving are sometimes unrelated, but creative skills may be used for more than problem solving.

Tactics such as those listed above (and there are many others) are often included in the various systematic programs that are designed to enhance creativity. Gordon's well-known procedure labeled *Synectics* (from 1961), for example, encourages the use of analogical thinking. It is easy to see the power of analogical thinking in inventions such as the steam engine (a tea kettle), Velcro (clingy burrs), and the airplane (birds, to whom the Wright brothers looked many times). Here again there is an alternative, which in this case is that some creative thinking uses analogical reasoning, and some does not.

Even more systematic programs have developed around models of creative problem solving. There is the *Odyssey of the Mind*, for example, *Future Problem Solving*, and *Creative Problem Solving* (CPS). The last of these is probably the best known for its impact on divergent thinking. CPS relies heavily on *brainstorming*, where the quantity of ideas is targeted by a group while the quality is de-emphasized. Judgment and evaluation are postponed, and 'piggybacking' (extending ideas given by others, stretching a line of reasoning) is encouraged. Empirical research on brainstorming suggests that it is only modestly effective (summarized in a chapter by Tudor Rickards and Christian deCock, in 2012). In fact, if an organization really needs a large number of ideas, they may be better off having individuals work alone. Brainstorming may be good for team building, but it does not seem to work as well as for individual problem solving. Too often, individuals in brainstorming groups tend toward 'social loafing' and assume that others will do enough work to get the job done. Likewise, the creative process involves more than the quantity of ideas that brainstorming targets. Min Basadur's *Simplex* training targets problem definition and solution evaluation, as well as idea

generation. Basadur also is explicit in the training about cycling through the process as needed rather than simply going once through it.

Information Processing

Occasionally, enhancement efforts focus on basic facets of information processing rather than the meta-cognitive contributions to creative thinking (e.g., tactics). Note the label, 'information processing.' It is indeed a matter of processing, and usually this kind of training targets changes in information transfer rather than information storage (memory). This follows from the view that information storage capacity is static and largely, or even entirely, universal. The label focuses on 'information' because enhancement is expected when an individual stops treating information as factual in the static sense. This is another way of saying that the individual questions assumptions and shifts his or her perspective. Information is consciously examined and perhaps even altered, rather than merely transferred without question.

At other times, maximization of creative potential is expected when the individual is simply more knowledgeable, or at least has the most relevant data. Good information is very important for many creative solutions, in many domains, although of course quality is more important than quantity. In fact, individuals can apparently have too much information! When this occurs, there is a brand of expertise that leads the expert to assumption and a reliance on routines. In their 1995 article on the psychoeconomics of creativity Rubenson and Runco suggested that this is one reason why creative thinking becomes more difficult as we age. We have more information and make more assumptions. These preclude mindful and spontaneous thinking. Creative thinking is probably most likely where there is an optimal amount of information – more is not always better – and when the individual knows to treat information as contextual rather than absolute. Enhancement is most likely to lead to creative thinking when the individual is informed but, in addition, avoids the assumptions and routines that may preclude original thinking.

Motivation and Attitude

The most effective efforts to maximize creative potential are those that communicate the value of creative solutions and ideas. In addition to simply showing *how* the individual or organization can be more original, they show the *why* as well. This brings us to the extracognitive basis of creativity, including the motivational and attitudinal components of creativity which are important for the development of creative self-efficacy. Attitudinal changes may be necessary because people will need to expend effort to change their ways. They will need to break routines, to question assumptions, and to take risks.

Some individuals and organizations may assume that creative persons are eccentric, or even difficult to work with (because they do things their own way). In this light, one of the critical attitudes targeted in enhancement efforts is *tolerance*. People in organizations wishing to increase innovation and creativity must tolerate independence, risks, and perhaps a lack of structure. Educators must tolerate divergence in the thinking of the students. This can be especially difficult because so much of the educational curriculum (and assessment tradition) is convergent. Tests, for example, are most objective when there is one correct answer, rather than originality.

It would be difficult to overstate the role of motivation. Without it, individuals will not invest in the mindful information processing that was described above. If an individual is already motivated to act creatively, it may be a matter of supplying tactics, direction, or resources. If the individual is not motivated, or motivated to work in other ways (e.g., *satisficing*, or just getting the job done), the motivation to do creative work may be difficult to arouse. This difficulty in arousing motivation occurs because creativity efforts are typically intrinsically motivated by payment or rewards – contingencies and promises made by others. Rather, intrinsic motivation comes from within and is personal and independent of contingencies and promises. Moreover, extrinsic factors can undermine intrinsic motivation. For this reason, the creativity that results from intrinsic motivation to solve a problem or express oneself can be lost if it is not respected and if an effort is not made to reinforce or redirect it. This can make enhancement very difficult. On the one hand, motivation is required, but on the other hand, it is a kind of motivation that comes from within. This is why values should be targeted. If an individual values creativity, he or she will probably be motivated to express and apply it. Instead of rewarding, say, piano playing, it might be better to show how playing the piano improves the quality of life and allows a healthful self-expression. This is just one example, the point being that it is possible to avoid the *overjustification effect*, whereby intrinsic motivation is lost, if the training focuses on general values and the specific application of a tactic or action is left to the individual. Of some relevance is the finding from the meta-analysis of creativity training, prepared by Scott, Leritz, and Mumford and reported in the 2004 *Creativity Research Journal*. This analyzed 70 empirical investigations studies of creativity training and concluded that successful training should include examples of 'real-world' illustrations followed by a series of domain-specific exercises.

Environmental and Interpersonal Factors

Interpersonal and contextual influences on creativity are indicative of the need to respect intrinsic motivation and the need for tolerance in educational settings and organizations. Indeed, maximization of creative potential may result from the careful structuring of context as much as from direct and explicit training.

Creativity is supported by environments that allow autonomy and tolerance and provide the resources necessary for the individual to maintain independence. Tolerance for autonomous and diverse work styles and interests may be vital in most organizations, and any individual with an opportunity to behave in a creative fashion will probably need time. Time allows incubation and the exploration of alternatives. Admittedly, while deadlines and imposed schedules can constrain thinking, some famously creative persons have claimed that they worked best under pressure (e.g., with some competition).

This fits with the idea of creativity as intrinsically motivated. Each individual may have his or her own ideal context for creativity.

Creative processes often produce unexpected results. They are, after all, original, and original results may be at most only loosely associated with previous efforts. But the unexpected nature of creative insights means that there is a risk. Thus, a tolerance of risk and use of creative discretion, on the part of the individual and those working with the individual, is often necessary. Further, organizations must accept that risk if they wish to support creative and innovative thinking.

Conclusions

Enhancement of creative potential through training may involve explicit instructions or it may result from particular environments and contexts. Training efforts can be founded in the cognitive, metacognitive, motivational, or attitudinal facets of the creative process. The cognitive facets may be the easiest to target (e.g., by suggesting tactics) and the motivational, the most difficult, given the role of intrinsic interests, but the most effective enhancement will take most or all facets of the creative process into account.

Educational practitioners and managers should recognize the need for enhancement efforts. They may investigate the programs that are designed to maximize creative efforts, or they may take it upon themselves to create a stimulating environment and provide the respect, resources, and tolerance that will allow creative thinking. Enhancement is probably most effective over the long run if the training covers a period of time or is revisited. If the interest is in fulfilling the potential of children, parents are in the best position to enhance creative thinking. They can model creativity, provide opportunities for original problem solving and self-expression, and appropriately appreciate autonomy and risk taking – and they can do so for an extended period of time. If the benefits of enhanced creativity are recognized on a larger scale, by society as a whole, enhancement is virtually assured. If this occurs, parents, educators, managers, and supervisors will all tolerate, appreciate, support, and benefit from enhanced creative thinking.

See also: Problem Selection in the Social Sciences: Methodology; Self-Determination Theory.

Bibliography

Adams, J.L., 1986. Conceptual Blockbusting: A Guide to Better Ideas. Norton, New York.

Barron, F., 1995. No Rootless Flower: An Ecology of Creativity. Hampton Press, Cresskill, NJ.

Basadur, M., 1994. Managing the creative process in organizations. In: Runco, M.A. (Ed.), Problem Finding, Problem Solving, and Creativity. Ablex, Norwood, NJ, pp. 237–268.

Basadur, M., 2011. Attitudes and creativity. In: Runco, M.A., Pritzker, S. (Eds.), Encyclopedia of Creativity, second ed., vol. 1. Academic Press, San Diego, CA, pp. 85–95.

Fullan, M., 2001. The New Meaning of Educational Change, third ed. Teachers College Press, New York.

Gordon, W.J.J., 1961. Synectics: The Development of Creative Capacity. Harper and Row, New York.

Meador, K.S., Fishkin, A.S., Hoover, M., 1999. Research-based strategies and programs to facilitate creativity. In: Fishkin, A.S., Cramond, B., Olszewski-Kubilius, P. (Eds.), Investigating Creativity in Youth: Research and Methods. Hampton Press, Cresskill, NJ, pp. 389–416.

Rickards, T., deCock, C., 2012. Understanding organizational creativity: towards a multiparadigmatic approach. In: Runco, M.A. (Ed.), Creativity Research Handbook, vol. 2. Hampton Press, Cresskill, NJ, pp. 1–31.

Rubenson, D.L., Runco, M.A., 1995. The psychoeconomic view of creative work in groups and organizations. Creativity and Innovation Management 4, 232–241.

Runco, M.A., Albert, R.S. (Eds.), 2014. Theories of Creativity, rev. ed. Hampton Press, Cresskill, NJ.

Runco, M.A., Charles, R.E., 1993. Judgments of originality and appropriateness as predictors of creativity. Personality and Individual Differences 15 (5), 537–546.

Runco, M.A., Richards, R. (Eds.), 1997. Eminent Creativity, Every-Day Creativity, and Health. Ablex, Norwood, NJ.

Scott, G., Leritz, L.E., Mumford, M.D., 2004. The effectiveness of creativity training: a quantitative review. Creativity Research Journal 16 (4), 361–388.

Creolization: Sociocultural Aspects

Christine Jourdan, Concordia University, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article considers the usage of the concept of creolization in contemporary social sciences and raises questions about its relevance and usefulness for the study of contemporary sociocultural change. Topics covered include the origin of the concept, its appropriation by social sciences including anthropology, its various meanings and usage, and the various critiques that have been addressed to these.

The concept of creolization has been used in anthropology to refer to processes of social and cultural change that take place in societies characterized by rapid social flux and where cultural influences from various origins are integrated into a new locally produced system of meaning. The term is borrowed from linguistics, and used in analogy with the processes of language creolization that appeared in some types of sociolinguistic contact. Cultural creolization takes place during sustained cultural contact, more typically produced within the hegemonical conditions of colonization, World System influences, and globalization.

Over the years, creolization has been used in different ways: as a label indexing sociocultural changes associated with colonization in some parts of the world; as a metaphor capturing the elusive processes of integration of new cultural forms; as a synonym of hybridization associated with globalization; and as a model for the study of cultural change directly derived from the linguistic model of creolization. In all cases, the term refers to processes of social change and restructuration. An example of creolization, certainly not representative of all types and domains of creolization, would be a Solomon Island businessman sending printed invitations to the Catholic wedding of his daughter in the biggest hotel of the city of Honiara; yet, he would make sure to collect the bridewealth from the family of the groom, and would subsequently redistribute it to members of his lineage and clan according to preexisting obligations. The blending of the two practices gives new meanings and new salience to the formalization of the alliance between families, but also between the newlyweds.

The concept of creolization has been 'appropriated' by cultural studies, literary criticism, cultural geography, Latin-American studies, history in addition to anthropology and has provoked healthy debates within these disciplines, but also outside of academia. Indeed, as Romberg points out (2011), creolization and creoles have become a political terrain within academia and local interest groups. In the same vein, Stewart (2007) points to the multiplicity of meanings of 'creolization' and 'creole' and as applied to domains as diverse as race, diaspora, the body and theory. Discussions on the applicability of creolization to the study of social change usually address the nature of the creolization process itself, the relationship of creolization to globalization and to the concepts of identity, culture, and creativity.

One can ask questions about the current interest in creolization. Is this interest, for instance, indexical of current anthropology's interest and fascination for plural and complex societies (Benoist, 2013)? *Is creolization good to think*, as Aisha Khan asks, or is it merely a seductive *floating signifier* (2007: 665)? Does creolization offer a frame that facilitates the analysis of "*fluidity and unbounded nature and cultural processes*" in light of "*pre-existing cultural stability and continuity*" (Eriksen, 2007: 163)? Do creolized societies represent an ideal (Glissant, 2011) of cultural richness, transformation, and creativity that insufflate new blood into old veins?

The life span of creolization is also a subject of debate and a source of questions. Is creolization a transitional period in the life of societies, at particular moments of their history, or is it a permanent state of affairs? And if linguistics proves to be an appropriate model for social creolization, do the creole continuum and the concept of decreolization prove useful for analyzing the transformations of societies whose creolization has peaked, so to speak (Hannerz, 1987)? We can argue, as Eriksen (2007) does, that standardization constitutes a form of decreolization: as fluid cultural forms become cultural models, they are codified. But this begs the question: What is the life span of these codified cultural models? Will they in turn be creolized? Are all cultural models creolizable to the same extent?

The Origin of the Concept

As Bolland (1998) explains, the concept of creolization has meant many things to many people over many centuries. The term is derived from 'creole' itself derived from Spanish 'criollo' which referred to (1) Europeans born in the islands during colonization (thus Joséphine de Beauharnais, the wife of emperor Napoléon, was called a 'creole' as she was born in Martinique, and was not considered totally French anymore); (2) displaced people brought to the Caribbean and Indian Oceans and who contributed centrally to the local plantation economies; (3) the societies that developed during cultural contact in these places; and (4) an islander of mixed 'race.' The overarching theme of 'creole' and creolization thus, is that of transformation. Linguistics borrowed the concept from this social history, but it is worth noting that the concept of creolization appears as early as 1884 in the work of de Quatrefages, a French ethnologist who borrowed it from a letter received from M. Levy (Benoist, 2013: 20). In Linguistics, creolization refers (1) to a process of transformation of a pidgin language into the main language (and often, into the mother tongue) of a new linguistic community, and (2) to the creation of a creole

language out of different languages (at least two) without an intermediate pidgin state (*see* Pidgin and Creole Language Varieties). It is this second meaning that has been adopted by anthropology. A pidgin or a creole developed as a response to communication pressures, in situations of intense cultural contact and often hegemonical conditions such as those found on colonial plantations of the New World, where people from different cultural origins and who spoke mutually unintelligible languages developed a new common language. Human agency and asymmetrical power relationships are central to the creation of these languages. Pidgins and creoles incorporate features from languages available to the pidgin/creole makers: the new language is distinct from the languages that shaped it, yet it often retains a sufficient number of their features, that its origin can be traced and identified. Theories of pidgin and creole genesis are highly debated in creole studies, yet scholars agree that the characteristics of pidgin and creole languages include the (1) speed with which they form and develop; (2) the sociological conditions that have shaped their development, most typically, European colonization, and in some geographical areas such as the Caribbean, the Indian Ocean and the Western Pacific, plantations economies; and (3) an almost instantaneous and continuous creativity. In many instances, creoles coexist with their lexifiers (typically, the colonial language that served as the source of the bulk of the lexicon) and regularly suffer from negative comparisons with them. Purist linguistic ideologies denigrate the hybrid nature of creoles, their supposed lack of syntax, and most of all, their lack of written literature.

Creolization and Anthropology

Despite the epistemological difficulties inherent to the transfer of theories and analytical models across disciplines, the concept of creolization seems to have captured the imagination of some scholars as a way of talking about and analyzing rapid cultural change associated with contact between societies of different ideological and technological orientations.

One wonders why the concept of creolization was applied to these phenomena when the more technically appropriate concept of 'pidginization' was available (see Jourdan, 1987; McKellin, 1991). Knörr (2010) proposed that pidginization would have been more appropriate as the term speaks to the development of common cultural elements amid cultural diversity, whereas the process of creolization entails indigenization that is not always a result of cultural contact. I suspect that the choice of creolization over pidginization is linked to two factors: first, the word 'creole' already existed in social sciences and public discourse to refer to the people and plantation societies of the Caribbean and the Indian Ocean. In these societies, typically characterized by a high degree of cultural diversity and intense contact, creolization took place with the building of communities and the development of indigenization. In other pidgin/creole speaking societies, that was not the case. Intense cultural contacts between different ethnic groups did not erase ethnic differences, but led to the development of new cultural formations. Second, for a long time pidginization has been thought of as a process of linguistic impoverishment, whereas creolization was always associated with linguistic enrichment. Using the former label could have been seen as a pejorative and inadequate label to describe a process of cultural creativity where 'loss' had no place.

The metaphorical power of the concept of creolization, particularly with the implication it carried that new ways of being in the world and living together appeared, made it preferable to the more technically correct pidginization. The earlier examples of such usage appear in the works of Drummond (1980), and Hannerz (1987). In parallel, French linguists, anthropologists, and literary critics, working in West Africa, in the Caribbean, and in the Indian and Pacific Ocean developed other concepts to discuss similar phenomena: *créolité* (Mannessy, 1987; Chaudenson, 1992; Bernabé et al., 1989); creolicity (Jourdan, 1985) and, of cultural *métissage* (Amselle, 1990).

In contemporary social sciences, creolization is being used in different ways, with more or less attention paid to history and power relations between peoples and cultural groups, and greater and lesser claims on recognizing the heuristic dimension of the concept. A first approach uses creolization as a concept that refers to processes of 'transculturation' leading to nativization. It is argued that the concept of creolization is most useful when restricted to those social transformations that take place under the hegemonical conditions of colonization, to the relationship of power and inequality that characterized the prevalent social conditions of colonization, and to the processes of 'transculturation' leading to indigenization during colonial times in the Caribbean and the Indian Ocean. Shih and Lionnet (2011: 31–32) argue that this '*geohistorical specificity*' warrants the restriction of the concept to these areas. But as Stewart (2007: 5) asks "*Can the Caribbean exercise copyright over the meaning of creolization?*" An extension of this restriction is *créolité*, the French Caribbean intellectual movement analyzed in the work of Bernabé et al. (1989). *Créolité* puts an emphasis on identity and serves as a conduit to more ideological considerations of nation-state building and nationalism (see Tauchnitz, 2014). While some authors claim that the use of creolization is appropriate only in such societies (Chaudenson, 1992), others, such as Stuart Hall (2003) and Françoise Vergès (2007), argue that creolization cannot be dissociated from the power relationships that framed it. Thus creolization is not simply a process of social change; it is a process of social change that entails "*inequality, hierarchization, issues of domination and subalternity, mastery and servitude, control and resistance*" (Hall, 2003: 31, cited by Shih and Lionnet, 2011: 25). Using it lightly as a metaphor, it is argued, would obfuscate these power relations and the forced transculturation that it entails (Shih and Lionnet, 2011). This approach stresses that creolization cannot be used as a 'master metaphor' for similar processes around the globe. Indeed, in his important review, Palmié (2006, 2007) warns against the danger of appropriation of the concept by different fields without consideration for its original historical, social, and geographic context.

A second approach uses creolization as a metaphor for capturing elusive processes of integration of new cultural forms, associated with globalization. Used in that way, the meaning of creolization remains very akin to that of hybridization and creates the same type of conceptual ambiguities: is it the process, or is it the results of structural modification or reorganization? The image that there is something that gets

mixed with something else to produce yet another thing (e.g., the new form of marriage in Honiara mentioned above), is inherent to the metaphor. As such, it is an interesting shortcut into the description of phenomena not easily definable, but which seem to be clearly associated with the complexification of social relationships. Creolization, in that sense, is used to refer to a recombination of structures, practices, and meanings: a case of 'new wine in old bottles' or 'old wine in new bottles.' Used only as a metaphor, creolization (and hybridity) fails to explain how these kinds of phenomena warrant the usage of this new term rather than intersystem (Drummond, 1980), syncretism (*see* Syncretism), bricolage, or métissage, those other concepts central to a study of social change that focuses on the in-between. In as much as they focus on the dynamic intermixing and restructuring of meanings, both concepts are similar. Both capture the image of the in-between, the neither/nor, flux, and vanishing boundaries. But while 'hybridity' and 'métisse' focus on the result (Benoist, 1996; Chanson, 2011; Tauchnitz, 2014 for important reviews of these concepts), creolization focuses on the process.

But what process? If the concept is going to be analytical rather than remain epigrammatic, as Stewart (2007) claims it is, it ought to cover a precise semantic field. As he says, "*the term 'creole' has itself creolized, which is what happens to all productive words with long histories*" (2007: 5). But can we be satisfied with using a term whose meaning gets bleached by usage over a short period and is becoming more polysemic with time? Metaphors are not devoid of reality. They are anchored in it. Though creolization implies cultural creativity, its ebullience is not synonymous with chaos; and though it certainly leads to new meaning making systems and new knowledge systems, one cannot be satisfied with defining it solely as an aesthetic movement, or an exercise in identity creation or affirmation, or a metaphor for quick social change. One must find its inner workings, but one should not limit it to expressive interaction. To put emphasis solely on the folkloric dimension of creolization (see Baron and Cara, 2011 for a discussion) is to miss out on its power to evoke the cultural restructuring that is at the basis of the cultural creation that Mintz (1996) writes about. Are we being well served by claiming that creolization is simply a creative, chaotic, and playful response to cultural contact, akin to bricolage? No. Creolization is not some form of *angélisme*, nor is it a form of chaos.

A third approach uses creolization as an analytical model for the study of social change drawing on an analogy with the linguistic model of creolization–pidginization described above. This is a stronger reading of the concept of creolization: It proposes ways of identifying the cultural logic which provides answers to the following nagging questions: What changes and what does not, and under which conditions? What is the nature and the pace of the change? If culture is practice and not simply ideology, how does change affect this practice? If culture is text, is every aspect of culture affected in the same manner and to the same extent? If culture is grammar, are there parameters and constraints for change? What makes the change acceptable? This version of creolization rests on a set of assumptions about the deep grammatical nature of cultural engagement: (1) that moral-existential aspects of culture can be differentiated from instrumental aspects; (2) that the latter change at a faster pace than the former; and (3) that, with regards to the pace of change, they can be compared with the syntax and lexicon of a language respectively. In situations of nascent creolization, local instrumental aspects of culture are often replaced, along with their attending symbols and meanings, with those from another source. The moral-existential aspects of culture are now sustained by new symbols, and may be given slightly different meanings, but will still be perceived as essential to the group's definition of identity (Jourdan, 1987; Mercer, 1988 cited by Stewart, 2007: 6). The bridewealth payment mentioned above is essential to the father's standing in one of his networks (here his lineage) and justifies his traveling back to the home village to collect it. All participants in the transaction agree that the meaning of such bridewealth is now very different from what it used to be, and that the amount is inflated. Yet bridewealth has to be paid, even against the wishes of the parish priest, because this form of exchange is essential to the integration of lineages, and therefore to the father's identity as a member of the group. Critics of this third approach, Briggs (2002) and Palmié (2006, 2007) argue that linguistic theory should not be used as a foolproof mechanism to understand social processes, even more so because the very theories that buttress this approach keep changing. But if creolization is invoked loosely, and if it not to be heuristic, what is its usefulness? As with any form of social process, creolization rests on people's understandings of how they wish to or have to engage the world. It is associated with a reconfiguration of self and of the practice of identity. Undoubtedly, there is a distinct advantage in identifying the processes that lead to a redefinition of people's practice of identity, even if, initially, this identification may start with an analysis of the features and elements that are being transferred across cultures, with or without attendant localization.

Essentialism

Some critics of creolization theory have argued that it is essentialist because it presupposes the existence of pristine cultures that will subsequently join, mix, overlap, and be reshaped (Friedman, 1994). Proponents of creolization reject this notion, and point to culture as a locus of permanent change, made even more spectacular at particular times in the history of societies. They also point to the shortcomings of a conception of culture as a bounded pristine entity. Creolization stresses the importance of the negotiation of meaning and shows that the 'vivre-ensemble' is engaged with and not simply lived-in. If creolization were an essentialist theory, the concept of cultural loss would be integral to its definition. Yet, it is not, for the simple reason that talking about loss in connection with social change implies a move away from the ideal, but unreal, pristine state of tradition. Loss belongs to a theoretical framework in which cultures are seen as essential wholes, neatly bounded self-reproducing structures, impervious to the passage of time, and deriving legitimacy from a primal pristine state. Cultures are no less authentic because they are creolizing. The new cultural practices, and the new cultural forms, that have emerged out of the creolization process are not any more authentic or spurious

than the ones they are derived from, or more spurious or authentic than societies that have changed in other ways. They simply exist. What is revealed by the creolization concept is the endless negotiation of meaning that takes place when groups and individuals engage in social relations within societies. The pace and scope of these cultural negotiations are more important in situations of rapid social change. The strength of creolization theory is that it points to (or foregrounds) elements that are not easily captured by other models: heterogeneity rather than homogeneity, process rather than stasis, fragmentation rather than integration, multiple rather than singular identities, multiplicity of voices, continuity of movement. Above all, it points to the domain of culture as contested ground rather than public space.

The World System and Globalization

Typically, but not exclusively (see Hannerz, 1996) the creolization concept has been applied to the analysis of social change in the so-called Third and Fourth World societies, in colonial and postcolonial situations. Colonial, postcolonial, and hegemonical contact relations are particular in the type and speed of the sociocultural change they provoke, and are best suited to be studied with the paradigm of creolization that includes the speed of change as a defining element. Critics have argued that the quasi-exclusive application of creolization theory to these societies represents them as powerless and reinforces the *fait accompli* of colonization by strengthening stereotypes. The argument is usually framed in relation to the World System theory and to the concepts of core and periphery. There may be a reason for this affinity: the very sociocultural formations and ideologies that have fostered the development of creole languages have reappeared under a new guise, that of the World System, economic and cultural, which creates even more pressures for change in these societies than anywhere else at the same moment. These types of contact-induced ideological and experiential pressures resemble the superstrate and substrate influences at work in the formation of creole languages, in that they juxtapose and confront practices and meanings that are produced by different social systems and ideologies. This raises the question of the relationship of creolization to globalization. Creolization shares with globalization an increase in available modes of social organization and cultural representation (Pieterse, 1995), both offer more options. Both are associated with more culture and not less, even if, on the surface, things look different. But with globalization residing in a movement of ideas, as much as in people's engagement with new sets of ideologies and practices, one could identity, as Hannerz does in Stockholm (1996), creolizing formations in nontypical creolization sites.

Creolization and Fast Social Change

It is in cities that creolization is the most obvious, if only because of the rate and scope of change. As cultural market places, cities are the loci of very fast sociocultural mutations that lead some observers to talk of acculturation and deculturation, and not only of adaptation to an urban way of life. Creolization is neither. When full-fledged towns and cities develop over a mere 50-year period, as is the case in many parts of the Third World, in places where there is no local tradition of urbanization, for instance, it requires on the part of urbanites a reassessment of the parameters of their identity (as with the father's behavior at the time of his daughter's wedding) and subsequent engagement in new types of social relations of the type and scale that is not found in other forms of sociocultural change. What is typical of cities is the ebullience of cultural creativity. Only a close study of the superstrate and substrate influences could reveal why, despite the onslaught of a generic cultural and economic World System (the superstrate) onto local traditions (the substrate), the cultural formations of urban Nigeria (Hannerz, 1987), Solomon Islands (Jourdan, 1985), and Guyana (Drummond, 1980) are similar in very many ways, and yet very different from one another. This begs the question of the specificity of creolized societies: Are they any different from other types of societies, given the fact that all societies are changing, and that many contemporary non-Third World societies have, at some point in their history, lived in situations of colonization, invasion, or hegemony? Maybe they are, if only because of the speed and the depth of change that characterize them, and give the impression of a permanent flux.

Final questions remain. Will the concept prove more useful when applied to some social contexts of change and not to others, defined geographically and historically? Will anthropology be better served by using creolization as a universal synonym for fast cultural change associated with globalization and characterized by the blending of forms, meanings, knowledge systems, and structures? Will it be better served if creolization proves a heuristic concept that goes beyond the metaphor but helps understand and describe contemporary processes of change and mixing in light of increasing cultural complexity?

See also: Authenticity, Anthropology of; Center–Periphery Relationships; Colonialism, Anthropology of; Cosmopolitanism; Culture Shock; Diffusion: Anthropological Aspects; Globalization and World Culture; Hegemony and Cultural Resistance; Hybridity; Identity in Anthropology; Intercultural Communication: An Anthropological Perspective; Pidgin and Creole Languages; Practice: An Anthropological Approach; Syncretism.

Bibliography

Amselle, J.L., 1990. Logiques Métisses. Payot, Paris.
Baron, R., Cara, Ana C. (Eds.), 2011. Creolization as Cultural Creativity. University Press of Mississipi, Jackson.
Benoist, J., 1996. Métissage, syncrétisme, créolisation: métaphores et dérives. Etudes créoles 19 (1), 47–60.
Benoist, J., 2013. La créolisation: locale ou mondiale? Archipelies 3&4, 19–30.
Bolland, N., 1998. Creolisation and creole societies. A cultural nationalist view of Caribbean social history. Caribbean Quarterly 44 (1/2), 1–32.
Briggs, C., 2002. Linguistic magic bullets in the making of a modernist anthropology. American Anthropologist 104, 481–498.
Bernabé, J., Chamoiseau, P., Confiant, R., 1989. Eloge de la créolité. Editions Gallimard, Paris.

Chanson, P., 2011. Variations métisses: dis metaphors pour penser le métissage. Academia Bruylant, Louvain-la-Neuve.
Chaudenson, R., 1992. Des îles, des hommes, des langues. L'Harmattan, Paris.
Drummond, L., 1980. The cultural continuum: a theory of intersystems. Man 15, 352–374.
Eriksen, T.H., 2003. Creolization and creativity. Global networks 3 (3), 223–237.
Eriksen, T.H., 2007. Creolization in anthropological theory and in Mauritius. In: Stewart, C. (Ed.), Creolization: History, Ethnography and Theory. Left Coast Press, Walnut Creek, CA, pp. 153–177.
Friedman, J., 1994. Cultural Identity and Global Process. Sage, London.
Glissant, E., 4 février 2011. Entrevue Avec Edouard Glissant. Le Monde.
Hall, S., 2003. Creolization, diaspora an hybridity in the context of globalization. In: Enwezor, Okwui, et al. (Eds.), Créolité and Creolization. Hatje Cabtze, Ostfilden Huit, pp. 185–198.
Hannerz, U., 1987. The world in creolization. Africa 57, 546–559.
Hannerz, U., 1996. Transnational Connections. Routledge, London.
Jourdan, C., 1985. Creolization and Urbanization in the Solomon Islands (Ph.D. thesis). The Australian National University, Canberra, ACT.
Jourdan, C., 1987. Des plantations à la ville. Journal de la Société des Océanistes 85, 243–253.
Khan, A., 2007. Good to think? Creolization, optimism and agency. Current Anthropology 48 (5), 653–673.
Knörr, J., 2010. Contemporary creoleness. Or the world in pidginization? Current Anthropology 51 (6), 731–759.
Mannessy, G., 1987. Créolisation et créolité. Études Créoles 10, 25–38.
McKellin, W., 1991. Hegemony and the language of change: the pidginization of land tenure among the Managalase of Papua New Guinea. Ethnology 30, 313–324.
Mercer, K., 1988. Diaspora culture and the dialogic imagination. In: Cam, M., Andrade-Watkins, C. (Eds.), Blackframes: Critical Perspectives on Black Independent Cinema. The MIT Press, Cambridge, MA, pp. 50–61.
Mintz, S., 1996. Enduring substances, trying theories. Journal of the Royal Anthropological Institute 2, 289–311.
Palmié, S., 2006. Creolization and its discontent. Annual Reviews in Anthropology 35, 433–456.
Palmié, S., 2007. The 'C-Word' again: from colonial to post-colonial semantics. In: Stewart, C. (Ed.), Creolization: History, Ethnography and Theory. Left Coast Press, Walnut Creek, CA, pp. 66–83.
Pieterse, J.N., 1995. Globalization as hybridization. In: Featherstone, M., Lash, S., Robertson, R. (Eds.), Global Modernities. Sage, London, pp. 45–68.
Romberg, R., 2011. Ritual piracy: or creolization with an attitude. In: Baron, R., Cara, A. (Eds.), Creolization as Cultural Creativity. University of Mississipi Press, Jackson, pp. 109–136.
Shih, S.-M., Lionnet, F. (Eds.), 2011. Introduction. In: Lionnet, F., Shih, S.M. (Eds.), 2011, The Creolization of Theory. Duke University Press, Durham and London, pp. 1–33.
Stewart, C., 2007. Creolization: history, ethnography and theory. In: Stewart, C. (Ed.), Creolization: History, Ethnography and Theory. Left Coast Press, Walnut Creek, CA, pp. 1–25.
Tauchnitz, J., 2014. La créolité dans le context international et postcolonial du métissage et de l'identité. L'Harmattan, Paris.
Vergès, F., 2007. Indian-oceanic creolizations: processes and practices of creolization on Réunion Island. In: Stewart, C. (Ed.), Creolization: History, Ethnography and Theory. Left Coast Press, Walnut Creek, CA, pp. 133–152.

Crime and Class

Roland Chilton, University of Massachusetts Amherst, Amherst, MA, USA
Ruth Triplett, Old Dominion University, Norfolk, VA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C. Uggen, volume 5, pp. 2906–2910, © 2001, Elsevier Ltd.

Abstract

To assess the importance of social class in the production of some kinds of crime, we define social class and discuss the many kinds of conduct called crime. After a review of theoretical approaches to class and crime, we examine major research approaches to the issue, starting with studies of business cycles, unemployment rates, prisoner characteristics, and their relationship to crime rates. We then review studies of the distribution of crime rates in geographic areas, studies based on victim and self-report surveys, and close with a promising new approach, the creation and analysis of official registries.

A Basic Question

In 2001 Uggen asked, "Does social class affect criminal behavior?," and pointed out that criminologists have wondered about the importance of social class in the production of some kinds of crime for over a century. The lack of consensus on the issue is due, in part, to differences in the definition and measurement of crime and in part to disagreement about the meaning and measurement of social class. Social class usually refers to the social and economic status of individuals or families, but in criminological discussions it more often suggests a set of categories used to classify victims or offenders based on their wealth and income. There are discussions of values and culture, such as 'middle-class values' and 'lower class culture,' but most treatments of crime and class use occupation, employment, or education as indicators of income and wealth. In general, assessments of the social class of individuals or groups of individuals in criminology have been based on personal or family income, unemployment, education, occupation, or parent's occupation or education. At times combinations of these measures have been used. In general, the social class categories used in discussions of crime and class have been broad and sometimes vague, and there have been good reasons for this. We can say in a general way that those who own a great deal of property and have high incomes are rich or upper class and those who own less or nothing and have low incomes are poor or lower class, but beyond this general notion the issue is quickly complicated. No commonly accepted set of classes exists.

Changing definitions of crime have complicated the issue as well. Conduct officially defined as crime is usually the result of a legislative process, but in practice crime is defined by administrative policies and enforcement procedures. This can be seen most easily when considering crimes such as drug possession, gambling, and prostitution. In such cases, there is generally no identifiable victim and thus no complaining witness. The number of arrests and prosecutions will depend on enforcement policies and practices rather than calls to the police. Since those arrested and incarcerated for such offenses are disproportionately people with low incomes, limited education, and histories of unemployment, a link between class and crime is a plausible conclusion as is bias in the operation of the system of justice. The relationship is less obvious when assaults and thefts, crimes clearly involving direct victimization, are considered.

However, definitions of crime also were complicated by criminologists in the 1940s and 1950s when they started to ask young people to report their own illegal or improper behavior. In some of the early self-report studies so much youthful conduct was defined as delinquent that almost any child could be said to have committed a delinquent act, and this weakened the link between class and crime. Moreover, the early self-report studies did not develop a uniform or standard self-report format, and this, combined with the use of different social class measures, produced conflicting conclusions about the importance of social class as a predictor of crime.

Prior to the widespread use of self-report studies, the link between class and crime was studied by looking at the possible impact of economic conditions on crime rates, asking if economic downturns are followed by higher crime rates, and focusing on the social class of prisoners. By the 1930s, comparing crime rates for specific geographic areas with the social and economic characteristics of the areas was added to this set. Such studies asked if areas with indicators of high-poverty rates and low-social class were also areas with high-crime rates. The answers produced by these older approaches probably influenced the development of theories attempting to explain the reasons for the crime and class relationships. A brief history of important theoretical perspectives on social class and crime is presented below, followed by a review of the research tradition on this topic. Both discussions should highlight important definitional and measurement issues.

Class and Crime: The Theoretical Tradition

Prominent twentieth-century criminologists generally accepted the suggestion that class or some approximation of it was important in the generation of many forms of crime. Shaw and McKay found that crime and delinquency were linked, at least geographically, to low income, unemployment, dilapidated housing, and a lack of neighborhood organization. In addition, Shaw's use of life histories, presented by some young delinquents, provided detailed accounts of ways in which

impoverished areas provided young people with criminal opportunities. Their theory tried to explain why some area characteristics produced higher delinquency referral rates.

Robert Merton set out to explain why criminal conduct was so often an activity of poor people. His explanation is so plausible that some have wondered why it was not presented sooner. He suggests that almost all Americans come to believe that the primary goal in life is the achievement of monetary success and, at the same time, learn that there are culturally prescribed rules for doing this. More importantly, he argued that the economic position of some people effectively eliminates the possibility that they will achieve monetary success if they follow the rules and that this causes some individuals to increase their income by ignoring the rules. His discussion is more complex than this, but it provides an explicitly structural explanation for much crime and suggests that affluent individuals will commit less crime than poor people. Edwin Sutherland developed a more complex social psychological theory suggesting that individuals learn to accept or reject beliefs conducive to or hostile to criminal conduct based on the views of those with whom they associate and depending on the strength and priority of those contacts. However, he also introduced an activity he called 'white collar crime' in which crime is not committed by the poor but is engaged in by high-status individuals in connection with their occupations. From this perspective, asking about the impact of class position on specific types of crime may be more useful than asking whether social class position affects criminal conduct in general.

The gap between Merton's and Sutherland's approaches is not as wide as it seems, if we focus on Merton's suggestion that when people find themselves unable to achieve monetary success, if they follow the rules, one option is to innovate and break the rules. However, even people with large incomes and great wealth may want to increase their wealth and the status that comes with it. If they find their ability to do this limited by rules regarding financial transactions, they too may innovate and break the rules. Therefore, by viewing the differences between the illegal conduct of lower class people and the illegal conduct of upper class people as a matter of scale, it is still possible to attribute some crime to the impact of institutional constraints on conduct in the face of culturally prescribed goals. That is, it is still possible to see illegal conduct in structural terms. To be sure, the link between social class and crime is different in those rare cases when high-status people commit high-level financial crimes. We describe these crimes as rare because only a small percentage of the population find themselves in positions that allow them to steal very large sums of money. This suggests that high status can lead to crime but that, for people at the bottom, class and crime are linked because their position in the economic system limits their ability to succeed financially if they follow the rules.

Cloward and Ohlin suggested a different way to bridge the theories of Merton and Sutherland by focusing on the importance of opportunity in a variety of forms. However, they too assumed that the lack of economic opportunity, legal or illegal, was an important contributor to much criminal conduct. Like Cloward and Ohlin, Albert Cohen used ideas form Merton and Sutherland to explain the conduct of lower class boys in urban areas. He relied heavily on material from Shaw's life histories, for examples of delinquent conduct, and his book is notable for its careful description of the behavior to be explained. Cohen suggested that many lower class boys engage in delinquent conduct, in part, because of their desire to be accepted by teachers and other middle-class adults and by their lack of qualities that middle-class people admire. His argument is more nuanced but he assumes that much if not most urban, male, gang delinquency is lower class in origin. In a critique of Cohen's approach, Walter Miller suggested that it was lower class culture itself that created higher rates of misconduct among lower class boys. His conclusions were based on the comments of gang workers about the basic concerns of lower class children and their parents. Gang workers were people trained to locate, work with, and try to redirect the activities of urban gangs. Neither Cohen's nor Miller's theory generated a research tradition, but both may have inspired other theories of crime.

Walter Reckless was intrigued by the fact that some boys in high-delinquency areas did not engage in delinquency and developed what he called a containment theory to explain it. Basically, he suggested crime is the result of the failure of internal and external controls of conduct. At about the same time Short and Nye modified a technique for studying delinquency by asking young people to answer questions about their own misconduct. Because such self-reported delinquency studies frequently used questions about parental occupation and education many of the early uses of this procedure suggested a much weaker link between class and crime than earlier approaches. The self-report approach encouraged the use of questions about the child's relationship to parents and teachers and allowed the children involved to give some indication of their feeling about rules of conduct in general.

Perhaps the most salient theory to come out of this approach is the version of social control theory, which Travis Hirschi presented in 1969 in his book, *Causes of Delinquency*. His version of the theory proposes that it is attachment to teachers and parents, involvement in conventional activities, belief in the validity of rules, and fear of punishment that prevents boys from engaging in some forms of delinquent conduct. Later versions of the theory suggested that these factors or closely related factors had more to do with involvement in criminal conduct than class position or class origin.

Class and Crime: The Research Tradition

The Business Cycle, Unemployment, and Prisoner Characteristics

Although Bonger focused on the impact of economic conditions on crime early in the twentieth century, studies of the linkage between economic cycles and crime have been relatively rare. However, a 2006 study of business cycles and street crime compared trends in the 'Gross State Product' with changes in state-level robbery and property crime rates from 1980 to 2000. The authors concluded "that an improving economy has a negative and statistically significant effect on all four … property crimes and robbery" (Arvanites and Defina, 2006). In addition, there have been numerous studies of unemployment and crime. Ted Chiricos in a review of 63 such studies reported a consistent and positive relationship between unemployment and crime (1987), and a review of articles

published later found a relationship between unemployment and property but not violent crime.

Other early studies looked at the social class of prisoners and found that convicts were generally poor. In this tradition, a 1991 survey of state prison inmates in the United States found that, prior to their most recent arrest, only one-third of the prisoners had completed high school, one-third of them were unemployed for the month before they were arrested, and only about half of them reported annual incomes of $10 000 or more (U.S. Department of Justice, 1993). In addition, the Bureau of Justice Statistics report on the characteristics of the inmates interviewed in 1991 noted that they were mostly male (95%), minority (65% black or Hispanic or other race), and young (68% were under 35). Interviews of prisoners in state prisons that were conducted in 2004, indicated that only 40% had completed high school, 30% were unemployed for the month before they were arrested, and only 52% of them reported annual incomes $14 000 or more (National Archive of Criminal Justice Data).

Social Area Analysis

When Shaw and McKay developed Social Disorganization theory they were trying to explain why most children referred to the Cook County Juvenile Court during the first quarter of the twentieth century lived in economically disadvantaged areas of Chicago. Their Social Disorganization theory grew out of a systematic comparison of referral rates with the social and economic characteristics of geographic areas in Chicago from 1900 through the 1930s (Shaw and McKay, 1969). Although Shaw supplemented the statistical analyses with a series of life histories of children referred to the court, his work with McKay remains one of the best examples of the use of factual information in the study of crime and delinquency. Their work, and the work of others who repeated the process for different cities in the 1950s, 1960, and 1970s, clearly suggests that delinquency referral is related to social class.

However, in 1962 Karl Schuessler altered the approach by using the characteristics of 105 cities rather than the characteristics of neighborhoods to examine the correlates of specific crime rates for 1950. One of his basic findings was that "specific offenses may be combined into broader categories, but that a single, composite index of crime would be of doubtful value." Another was that the percent of the population engaged in manufacturing contributed most heavily to variations in crime rates. Where manufacturing was high, all seven offense rates were low. Then using a correlation technique designed to indicate whether or not a small number of factors might explain the variation in crime rates for these cities, he found that one important general factor might be called 'income and race,' and a second factor could be labeled 'education and manufacturing.' His general conclusion was that the incidence of crime is a function of a city's social and economic characteristics.

Between 1960 and 1990, a number of studies of the correlates of crime were conducted. Some investigators used the delinquency rates for census tracts and others used city-level, county-level, or state-level rates. The variable to be explained in these studies was sometimes an overall crime rate reflecting UCR counts of violent and property crimes. In other cases, researchers focused on homicide or robbery. Homicide was a frequent choice because it is the most serious crime in the UCR set and because it is believed to be accurately reported. The creation of the Supplementary Homicide Report program in the late 1970s may have drawn additional attention to homicide. This made it possible for Land and his colleagues to review the published results of 21 studies that looked at the social and economic correlates of homicide. Some of the studies looked at the homicide rates for U.S. cities, some used metropolitan area rates, and some used state rates. Some of the studies used 1960 data, some 1970, and some 1980 data. Land et al. found great inconsistency in the results produced by the earlier studies and set out to create a standard set of variables that could be used in the same way for each period for a set of cities, a set of metropolitan areas, and all 50 U.S. states.

However, the major change they introduced in their own analysis was to use a new variable created by combining a set of highly inter-correlated variables and calling it a 'resource deprivation/affluence' component. The variables used were 'percent black,' 'percent of children not living with both parents' (essentially 'mother–child families'), and three income measures. The variables most highly correlated with the overall homicide rate for all four periods were percent black, mother–child families, and percent in poverty. The new composite variable appeared to be a way to deal with the interconnections of these three variables, but one impact was to draw attention away from attempts to understand the role of family structure and racial composition in the production of high homicide rates.

To avoid this blind spot, some criminologists started analyzing black and white crime rates separately. Peterson and Krivo found that residential racial segregation was a strong and significant predictor of high-black homicide rates and that the percent of black residents in professional and managerial occupations predicted a lower black crime rate (1993). Parker and McCall found that economic deprivation affects the within-race homicide rates for white and black city residents (1999). Earlier, Sampson found that black male unemployment contributed to an increase in black mother–child families and that mother–child families contributed to high homicide and robbery rates (1987).

In 2009 and 2010, a group of criminologists returned to the questions raised in the 1990 Land et al. study to see what had changed since 1990. In a version presented at a 2009 Annual American Society of Criminology (ASC) meeting, the authors introduced additional variables and removed percent black from the deprivation/affluence measure, using it with percent foreign born to create a 'heterogeneity' variable (McCall et al., 2009). They reported that their analysis suggested that the 'resource deprivation/affluence' component, even without percent black, was a statistically significant and positive predictor of the overall homicide victim rate. However, in the published version of the paper, when percent black was put back in the deprivation component and the segregation measure was not used, the deprivation component was positive and statistically significant for all four time periods (McCall et al., 2010). The standardized coefficients for this version indicated that deprivation was the strongest predictor of the homicide rate for all four analyses. Taken together, these studies of the correlates of urban crime rates provide support

for the suggestion that crime rates, especially violent crime rates, are higher for urban areas with large lower class populations than they are for areas with more affluent populations.

Victim and Self-Report Surveys

Two kinds of surveys provide estimates of the social class of victims and offenders. In victim surveys, respondents provide information about themselves and their families and answer questions about their experience as crime victims during the preceding 6 months. Reports of a 2005 victim survey in the United States suggested a violent crime victimization rate of 38 per 1000 households with incomes below $7500. This was twice as high as the crime rate for households with incomes of $75 000 or more. This 2:1 ratio was constant from 1993 to 2005, the last year that income information was published for violent victimizations. The pattern was different for property crime rates. For 1993, the property crime rate was 400 per 1000 households with incomes of $75 000 or more and 306 for households with incomes of less than $7500. However, the rates for these two income categories gradually changed so that by 2010, the property crime rate for households with low incomes was almost twice as high as the rate for high-income households. This suggests that by 2010, low-income respondents were more likely than high-income respondents to be victims of crime.

Surveys asking individuals to indicate how often they have engaged in criminal or delinquent conduct usually include questions about income and education. Those most often completing such surveys have been juveniles, so they were usually asked about the income and education of their parents. Self-reported crime surveys have not created a standard format, so the kinds of misconduct studied and the measures of social class used have been similar but varied. Nevertheless, a large number of self-reported delinquency studies have suggested a very weak relationship between social class and delinquency. A review of such studies led Tittle et al. to conclude that the class-crime link was a myth (1978). However, a comprehensive international review by Braithwaite, that included self-report studies, studies of social areas, and other research approaches, found that social class is a strong predictor of crime in the United States and around the world (1981).

A self-report survey conducted in Greece, Russia, and Ukraine in 2006 asked adults about their criminal and deviant conduct. The use of adults was an important break with the traditional use of juveniles. Believing that questions about education or occupation in these countries would not be accurate measures of social class, the investigators used questions asking respondents if they could afford food, clothing, housing, medicine, and travel to classify them by social class. The number of 'yes' answers to these questions indicated their class. Greek respondents with high-social class scores reported fewer crimes than those with low-social class scored, but Russian and Ukraine respondents with high-social class scores reported committing more crime than those with low scores. It is hard to say whether it was the use of adults, the use of 'telling a lie' as part of the crime and deviance measure, or the use of a new measure of social class that produced these interesting results suggesting that in some countries, crime is predicted by high income.

A Promising Approach

The advantage of self-report studies is that they do not depend on the police to provide the crime information they collect. The disadvantage of the approach is that key information provided by the respondents cannot be verified or corrected. However, another survey approach being used in several European countries begins with a large representative set of individuals that is supplemented by arrest information. A 2012 study of male violence in Finland compiled income, education, and unemployment information from 'several different administrative registers' for over 48 000 men born in Finland. This file was then expanded to include police information about arrests for violence of men in the file. Analysis of the expanded registry indicated that reports of violence were strongly associated with low socioeconomic status and a prior criminal record. Education, across all ages was the strongest predictor of all types of violence, while income and unemployment presented a weaker relationship for young offenders (Aaltonen et al., 2012). This approach may replace or drastically alter victim and offender surveys because it reduces or avoids problems of missing data caused by the unwillingness of some respondents to answer questions about income, education, and employment.

In the United States, it would be possible for four federal agencies, working together, to create a registry that would contain income, employment, occupation, and arrest information for a representative set of residents. The Census Bureau's American Community Survey already collects much of the nonarrest information needed, but with the cooperation of the Social Security Administration and the Internal Revenue Service the income and employment information could be augmented. After arrests in the FBI's Records of Arrest and Prosecution file are added, the registry would provide a very good indication of the linkages of social class with 43 different offenses. While such an approach is technically possible, it is not currently feasible because of laws prohibiting data exchanges among these federal agencies. Nevertheless, this is a promising approach that may eventually be adopted as more European countries develop such registries.

Overview

The role of social class in the production of some kinds of crime is important because of its implication for crime reduction. If rates of serious crime can be reduced by moving more people out of poverty, programs that work in this direction make practical sense. The addition of empirical support to ethical and political arguments for creating a more equal society might produce beneficial changes in many communities. This review of the evidence for a class crime linkage and its theoretical synthesis suggests that social class is important in the production of serious crime. The unemployment and business cycle studies, the Bureau of Justice Statistics victim and prisoner surveys, and the social area analyses almost all suggest this conclusion.

The evidence supporting the opposite conclusion is limited but strong enough to explain why there is a lack of consensus among criminologists. One issue that must be resolved to arrive at a more definitive conclusion is the impact of age on

self-report studies. When self-report approaches are used, it will be important for the surveys to include people other than juveniles and young adults. A second impediment to consensus is the wide variety of types of crime that can be studied. Focusing on crimes that can only be carried out by high-status people will indicate that one need not be poor to break the law, but it will not tell us much more about the crime and class relationship. A third impediment to consensus is the fact that criminology is data starved. Data about the income and assets of individuals that is linked to the criminal activity of the same individuals is quite limited. Moreover, some measures of crime, such as offenses known to the police, are simply measures of victimization, and none of the large-scale reporting programs, except the National Prisoner Survey, have information on the social class of offenders.

The issue may be complicated by the fact that crime and the response to it may have impact on an individual's life chances. That is, crime may impact social class in addition to social class impacting crime. Sampson and Laub's 1993 study of boys sent to prison for juvenile offenses suggests that being incarcerated at an age when other young men are finding work and starting families may have had an impact on their economic status in the years following incarceration. In addition, as we have seen, the class-crime link is certainly different for different kinds of crime. However, the crime and class issue is important in spite of the complications and difficulties in assessing the impact of class. As Hagan noted in his 1991 ASC Presidential address, "the simple omission of class from the study of crime would impoverish criminology."

See also: Communities and Crime; Crime, Geography of; Crime, Race and Ethnicity; Criminology: Development and Current Status of the Discipline; Data Bases and Statistical Systems: Crime Measurement; Homicide.

Bibliography

Aaltonen, M., Kivivuori, J., Martikaien, P., Salmi, V., 2012. Socio-economic status and criminality as predictors of male violence. British Journal of Criminology 52, 1192–1211.

Antonaccio, O., Title, C.R., Botchkovar, E., Kranidiotis, M., 2010. The correlates of crime and deviance: additional evidence. Journal of Research in Crime and Delinquency 47, 297–328.

Arvanites, T.M., Defina, R.H., 2006. Business cycles and street crime. Criminology 44, 139–164.

Bonger, W., 1969. Criminality and Economic Conditions: Abridged and with an Introduction by A.T. Turk. Indiana University Press, Bloomington, London.

Braithwaite, J., 1981. The myth of social class and criminality reconsidered. American Sociological Review 46, 36–57.

Chiricos, T., 1987. Rates of crime and unemployment: an analysis of aggregate research evidence. Social Problems 34, 187–212.

Cloward, R.A., Ohlin, L.E., 1960. Delinquency and Opportunity: A Theory of Delinquent Gangs. The Free Press, New York.

Cohen, A.K., 1955. Delinquent Boys. The Free Press, New York.

Elliott, D.S., Ageton, S.S., 1980. Reconciling race and class differences in self-reported and official estimates of delinquency. American Sociological Review 45, 95–110.

Hagan, J., 1992. The poverty of a classless criminology: American society of criminology 1991 presidential address. Criminology 30, 1–19.

Hirschi, T., 1969. Causes of Delinquency. University of California Press, Berkeley, Los Angeles.

Land, K.C., McCall, P.L., Cohen, L.E., 1990. Structural covariates of homicide rates: are there any invariances across time and social space? American Journal of Sociology 95, 922–963.

McCall, P.L., Parker, K.F., Land, K.C., 2009. An empirical assessment of what we know about structural covariates of homicide rates: a return to a classic twenty years later. Paper presented at the annual meetings of the American Society of Criminology, Philadelphia, PA.

McCall, P.L., Land, K.C., Parker, K.F., 2010. What do we know about the structural covariates of homocide rates?: a return to a classic twenty years later. Homicide Studies 14, 219–243.

Merton, R.K., 1938. Social structure and anomie. American Sociological Review 3, 672–682.

Miller, W., 1958. Lower class culture as a generating milieu of gang delinquency. Journal of Social Issues 14, 5–19.

Nye, F.I., Short, J.F., 1957. Scaling delinquent behavior. American Sociological Review 3, 326–331.

National Archive of Criminal Justice Data, 2004. Survey of Inmates in State and Federal Correctional Facilities. Inter-University Consortium for Political and Social Research, Study Number ICPSR 4572.

Ousey, G.C., 1999. Homicide, structural factors, and the racial invariance assumption. Crimiology 37, 405–426.

Parker, K.F., McCall, P.L., 1999. Structural covariates and racial homicide patterns: a look at the multiple disadvantages in urban areas. Criminology 37, 447–477.

Peterson, R.D., Krivo, L.J., 1993. Racial segregation and black urban homicide. Social Forces 71, 1001–1026.

Sampson, R.J., 1987. Urban black violence: the effect of male joblessness and family disruption. American Journal of Sociology 93, 348–382.

Sampson, R.J., Laub, J.H., 1993. Crime in the Making: Pathways and Turning Points through Life. Harvard University Press, Cambridge, MA.

Sampson, R.J., Wilson, W.J., 1990. A theory of race, crime, and urban inequality. In: Hagan, J., Peterson, R. (Eds.), Crime and Inequality. Stanford University Press, Stanford.

Schuessler, K., 1962. Components of variation in city crime rates. Social Problems 9, 314–323.

Shaw, C., McKay, H.D., 1969. Juvenile Delinquency and Urban Areas. University of Chicago Press, Chicago.

Tittle, C.R., Villemez, W.J., Smith, D.A., 1978. The myth of social class and criminality: empirical assessment of empirical evidence. American Sociological Review 43, 643–656.

Uggen, C., 2001. Crime and class. In: International Encyclopedia of the Social and Behavioral Sciences. Elsevier, London, pp. 2906–2910.

U.S. Department of Justice, 1993. Survey of State Prison Inmates, 1991. Bureau of Justice Statistics, Washington, DC.

U.S. Department of Justice, 1999. Criminal Victimization 1998. Bureau of Justice Statistics, Washington, DC.

Crime and Criminal Justice: Applied Research from Routine Monitoring to Evidence-Based Practices

Roberto Hugh Potter and Gail S Humiston, University of Central Florida, Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Applied research has been at, and remains at, the heart of the fields of criminology and criminal justice since the days of the 'moral statisticians.' This article examines some of the history of applied research in these fields, and the roles of government, university-based, and nonuniversity-based researchers. The impact of different sources of funding and the challenges that funding produce to the integrity of applied research are canvassed. Particular attention is paid to applied research in the areas of prevention, intervention, and evaluation studies. The development of evidence-based practices, implementation research, and translational research is also covered as part of the evolution of applied research in criminal justice and criminology. Exemplars in the areas of corrections and policing are offered to highlight the continuing development of this important aspect of the application of social science research to applied problems.

Applied research in criminology and criminal/juvenile justice can be traced back to the foundations of modern social sciences. The work of the 'moral statisticians,' such as Quetelet, often dealt with aspects of crime (Akers et al., 2013). The 1925 work on delinquency prevention and treatment by Burt (Lodge, 1974) in the United Kingdom and the Cambridge–Somerville (Glueck and Glueck, 1959) research in the United States provide examples of research undertaken by academics to address the problem of juvenile delinquency.

Our definition of applied research follows the tradition set forward by Kurt Lewin (1946), and reflected in Nettler's (1984) formulation of description, risk assessment, program evaluation, and explanation in criminal justice research (see also Potter and Krider, 2000). Likewise, the development of program evaluation methods and research projects plays a strong role in our approach here (Rossi et al., 1999).

There are three themes in the applied research worlds of criminology and criminal justice since World War II: corrections, prevention, and evaluation. As we will explore below, much of the early program evaluation was applied to the results of correctional interventions. Prevention programs and the three levels of prevention (primary, secondary, tertiary) were introduced to the justice fields in the mid-1970s (Brantingham and Faust, 1976), and the evaluation of those programs provided one more impetus to applied research. These themes have now led to the development of 'implementation research' (Fixsen et al., 2005; Taxman and Belenko, 2012), and the broader 'translational research' (e.g., Neyroud and Weisburd, 2014).

The development of applied research in these fields has been uneven, however. In this section, we will examine some of the influences of government monitoring, policies, and funding on criminal justice and criminological applied research. We will also examine the changing roles of the academic, government, and private sectors in the provision of applied research. Then, we will consider one researcher's work which forged an intellectual fork in the road, altering the course of criminal justice policies and research, carving new pathways. Finally, we will examine current applied research in corrections and policing, and the potential for future directions of untapped information on issues of criminal justice personnel. In the end, we hope to provide the reader with an understanding of how far we have come and how far we might be able to push applied research in this area of social control.

Is Crime Monitoring Applied Research?

One of the distinctions that might be made when discussing applied research in criminal justice and criminology is that between crime monitoring (or surveillance), especially 'passive surveillance,' and more theory-driven research. By 'monitoring,' we mean the systematic collection of data about crime events that can be utilized to describe the incidence of crime and allow the data to be utilized for attempts to predict and control crime (Potter and Krider, 2000). Unlike crime monitoring research, theory-driven research purposefully manipulates variables to observe an impact on a crime-related outcome.

Monitoring crime has been a long-term endeavor of many national governments. The French began systematic collection of criminal court data in the 1820s, and continue to the present, while the Uniform Crime Report (UCR) provides an exemplar of one of the oldest surveillance systems in the United States. For our purposes here, it is enough to point out that the UCR was developed in the late 1920s and published annually by the Federal Bureau of Investigation (FBI) since 1930. The UK Home Office began publishing official crime data in the 1950s, and Australia in the 1960s (Mukherjee, 1986). Although some may not consider these crime monitoring systems and their reporting to be research, without these efforts national- and international-level crime information would be even more fragmented than it is to date.

In the United States, the analysts at the FBI generally report only the basic observations and trends across time. They leave it for others to utilize the data for exploratory and theory-driven research purposes (see Akers et al., 2013 for a discussion of surveillance and monitoring systems). We would argue that while some may not view crime monitoring as applied research, it does form the foundation for others to conduct

applied research on crime trends and their relationships to other social facts. This holds for monitoring systems on crimes, criminals, and victims, as well as statistics related to the formal organizations involved in the criminal justice process, such as police, corrections, and courts.

The Government Wants Answers!

The role of government funding in applied criminological/criminal justice research cannot be understated. Perhaps the first nation to invest formally in applied criminological research was the United Kingdom. The Criminal Justice Act of 1948 established the Home Office Research Unit, which authorized the expenditure of funds for criminological research (Lodge, 1974).

In the United States, the President's Commission on Law Enforcement and Administration of Justice (1965), and its landmark 1967 report *The Challenge of Crime in a Free Society* (USGPO, 1967), led to the establishment of the Law Enforcement Assistance Administration (LEAA). The LEAA program provided substantial funds for criminal justice-related research to universities and consulting agencies, as well as funds to state and local agencies for personnel and hardware (sometimes referred to disparagingly as 'toys for the boys').

In Australia, the Federal Parliament passed the *Criminology Research Act 1971* to establish the Australian Institute of Criminology and the Criminology Research Council (CRC). These federal-level agencies conducted in-house research and communicated that research to states, territories, and the public. The CRC also provided grants to universities, state agencies, and local organizations to provide research and program services, though the impact of that funding was being questioned into the early 1990s (Chappell and Wilson, 1993).

This government involvement in setting criminological/criminal justice research agendas through funding has not been without critics, of course. Potter remembers lively discussions in the United States of the perils of allowing government funding to set the criminological research agenda during the late 1970s and early 1980s. Similar concerns were noted in conference panels in Australia during the late 1980s. At the time, however, many of these doubts were general in nature and few seem to have been published.

More recently, the voices of concern have become more explicit. In the United Kingdom, several authors noted the influence of government policy on the types of applied research being conducted (e.g., Jefferson and Shapland, 1994; Jupp et al., 2000). In the United States, Chilton (2001) used part of his American Society of Criminology Presidential Address to argue that such peak bodies should be advising legislators on what areas of criminal justice deserved research funding. Likewise, Israel (2000) lamented the increasing pressure on Australian universities to achieve greater external funding for research purposes and the contradictory pressures such as demands placed on the independence and quality of criminological and criminal justice research. Geis et al. (1999) were among the few who expressed concern about the influence of private-sector funding on criminological research.

Contributions from private foundations continue to be overlooked in the funding of applied (and theoretical) research in criminal justice. Over the past few decades several philanthropic foundations have provided funding to university-, private-, and nongovernmental organizations (NGO)-based researchers to conduct applied research across a broad range of justice issues. Some of that funding has spun off research organizations listed below, while other funding has even begun to comingle with federal funds to push research into particular arenas. The MacArthur Foundation 'Models for Change' (MacArthur, n.d.) initiative, for example, seeks to move evidence-based juvenile justice reforms across the United States.

This previous discussion focused primarily on the concerns of academic-based researchers. Whether or not all research funding carries with its implicit cues from the funding agency, or is truly 'objective,' remains an empirical question. An additional question becomes, Is it only academics who do applied research in these areas, and do they have the same concerns?

Who Does Applied Criminal Justice/Criminological Research?

Probably all Western governments have some crime monitoring and applied research capacity at the federal/national government levels (e.g., National Crime Prevention Center, n.d.; National Institute of Justice, n.d.). In Australia, Canada, and the United States, most states or provinces also have units in various justice-related agencies that conduct applied research and monitoring of criminal justice trends and efforts (several are listed in the reference section). For example, most state-level correctional systems in the United States have some form of statistics and research unit that generates annual data and conducts evaluations of certain interventions. Some provide funds for university and other researchers to conduct contract applied research, though such funds are generally much smaller than those provided by federal governments.

A great deal of applied criminal justice research is conducted by university-based faculty around the world. Such projects range from small-scale evaluation and organizational studies, often conducted pro bono or for small sums of funding or publishable data alone, to large-scale multisite and multi-university collaborations attracting large sums of funding. Because there is no centralized repository of total funds paid for university-based criminal justice research, much less separated by applied versus theoretical, it is impossible to state the total value of such to the criminal justice world.

In the United States at least, the private research sector, both not-for-profit and for-profit, have been involved in applied criminal justice research from at least the LEAA days. The end of LEAA brought a temporary movement away from the corporate research sector as large accounting and consulting firms shed their social science researchers in the absence of government contracts. In the current environment, these private organizations may consist of small collaborations between university-based researchers. For example, the Jill Dando Institute of Crime Science at University College London (JDI, n.d.) is among the hybrid academic–private sector entrants into the area of applied research. Larger research organizations also exist, such as the privately owned Abt Associates (Abt, n.d.) and employee-owned Westat (Westat, n.d.).

A range of not-for-profits (NGOs) have also played a significant role in applied criminal justice research over the past century. In the United States, the National Council on Crime and Delinquency began conducting applied research shortly after its formation in 1907 (Krisberg et al., 2007; NCCD, n.d.) and continues to provide such services. The Vera Institute is now more than 50 years old and continues to play a major role in criminal justice applied research and policy development (Vera, n.d.). These nonprofits have been joined by a myriad of others over the past few decades, with newer entrants such as the Urban Institute (Urban Institute, n.d.), Research Triangle Institute (RTI, n.d.), and the Pew Center on the States (Pew Charitable Trusts, n.d.) playing increasingly important roles in applied research in criminal justice settings.

A Fork in the Road

National governments spent a great deal of public dollars to research solutions to crime and to increase efficiency in their criminal and juvenile justice systems from the late 1960s to the early 1980s. Then came a rude shock in the form of the 'nothing works' revelation attributed to Martinson (1974). Martinson's brief was fairly broad, to study the effectiveness of offender treatment programs. His 'bald summary' with regard to the impacts on recidivism was: "With few and isolated exceptions, the rehabilitative efforts that have been reported so far have had no appreciable effect on recidivism" (Martinson, 1974, p. 25). In the paragraph prior to that statement, however, he conditioned what was to follow by stating: "The use of even this one measure (recidivism) brings with it enough methodological complications to make clear reporting of the findings most difficult… so that it is hard to tell whether 'what works' for one kind of offender also works for others." Whether or not Martinson's statements amounted to 'nothing works,' they had the dual function of spurring a legislative landslide to execution and incarceration regimens, while at the same time sharpening the resolve of criminal justice researchers to determine what might work. On an aside, 40 years later, Durose et al. (2014) observed that post-release offending patterns in the majority of US states have not changed appreciably from observations made during Martinson's review. This is regardless of employing either a retributive or rehabilitative correctional intervention approach.

Moving to 'What Works' and Evidence-Based Practices

Throughout the 1980s it appeared that the 'nothing works' and retributive philosophies guided legislative funding of criminal justice and criminological research, especially in the United States. At the same time, criminal justice/criminological researchers around the world were engaged in applied research to determine effective practices. The search for 'evidence-based' programs which decisively achieve their intended outcomes, such as crime reduction, continues today. The remainder of this article will deal with exemplar strands of research on evidence-based programs and practices in the areas of policing and corrections.

Corrections

As a direct response to Martinson's (1974) critique, a broad range of researchers in the corrections area, especially in Canada, began to explore the characteristics of individuals, groups, and communities which may affect the likelihood that persons convicted of criminal acts would continue to offend. Empirical research has shown that certain characteristics, or 'criminogenic risk' factors, are associated with the increased likelihood of offending. As a result, 'front-end' assessments and 'back-end' interventions were developed at a new pace. These efforts continue to the present and impact the practices of corrections at every level (local, state/province, and national).

Research has culminated in meta-analyses of correctional interventions in adult and juvenile justice settings, which have provided a set of principles of risk for offending and the development of a range of interventions to address the dynamic risks (Andrews and Bonta, 2010; Farrington and Welsh, 2007; Lipsey, 1992; Lowenkamp et al., 2006; Smith et al., 2009). These have led to the development of the 'Risk-Need-Responsivity' (RNR) actuarial approach to assessing the dynamic criminogenic needs of offenders and to placing them into interventions appropriate to the need and individual offender's readiness for change (Andrews and Bonta, 2010).

The development of risk assessments and evidence-based interventions, as well as the subsequent RNR process, have led to a concern with what Gottfredson (1984) termed 'fidelity' and implementation. In Gottfredson's approach, fidelity had to do with the integrity of the intervention being implemented and moving into what we would now recognize as a form of continuous quality improvement. These ideas have been wrapped together into 'implementation science' (Fixsen et al., 2005; Taxman and Belenko, 2012). Implementation science suggests that we should treat the outcomes of interventions and implementation as being equally problematic and deserving of study. After all, if the notions of 'intervention' rely on fidelity of 'implementation,' then the two are inextricably linked in theory and practice.

As noted earlier, the lack of significant change in the proportion of prisoners who recidivate within 3 years of release from prison in the United States suggests we have a way to go before documenting the impact of our current assessment and intervention programs. Likewise, recidivism in Australia (Payne, 2007) and Canada (Nouwens et al., 1993; Bonta et al., 2003) remains stubborn. This is an area of applied research across risk, responsive interventions, and implementation, which is sure to continue as we seek ways to respond to those for whom prevention fails.

Law Enforcement/Policing

For most citizens, police are the most likely point of contact they will have with the criminal justice system. In terms of sheer numbers (though not necessarily budgets), police are the single largest segment of the criminal justice system. Because of the discretion afforded police officers, they are also the gatekeepers to the process. It is not difficult to see why so much applied research focuses on policing. This ranges from assessments of

technological innovation on police activity and effectiveness (e.g., Chan, 2001; Grossi, 2008; Terrell et al., 2008; Wolf and Gordon, 2007), to studies of organizational factors associated with effective policing (e.g., Lambert and Paoline, 2010), to citizens' perceptions of police legitimacy and impact on crime (e.g., Gau, 2013).

One exemplar of applied policing research that has spanned more than 20 years is the police response to domestic violence incidents (Sherman, 1992; Sherman and Flynn, 2014). The 'Minneapolis experiment' combined the gold standard experimental assignment of cases to arrest or no arrest, and recorded subsequent behaviors by the alleged offender to the nonarrested partner (and in other sites other victims). The results reported were varied in terms of reducing subsequent violent behaviors by those arrested. Yet, for many years, these experiments led to statutory and departmental policies that required the arrest of accused domestic offenders.

Recently, however, an analysis of follow-up data from one of the original sites (Milwaukee, WI), had shown that death rates were twice as high among African-American women whose partners were arrested during the experiment as compared to those African-American women whose partners were not arrested (Sherman and Flynn, 2014). While these deaths are not directly attributable to domestic violence or the arrested partner, they do raise questions about unintended outcomes from applied research – or, adverse events. The study of adverse events of our intended interventions is among those areas in which applied criminal justice researchers are lagging.

The entire area of 'intelligence-led' policing (Ratcliffe, 2008) is another evidence-based subject where decades of applied research have come together to inform crime prevention and intervention efforts at the local level. Employing an experimental design, Ratcliffe et al. (2011), demonstrated the impact of foot patrols aided by intelligence systems and analytics over automobile-based patrol activities in the city of Philadelphia. Along with the observed intended crime reduction outcomes, the researchers also noted public health interventions and positive outcomes achieved by police being closer to the public they served.

Personnel Matters!

One of the often overlooked areas of applied research in criminal justice involves personnel studies. Issues of personnel recruitment (Stinchcomb, 2008; Stinchcomb et al., 2006), selection, early warning systems (Walker et al., 2000), and retention are of great importance to criminal justice and allied organizations. While some of these are destined to be published in academic peer-reviewed outlets, many more are produced for individual agencies or peak organizations. The US National Institute of Corrections (www.nicic.gov) is one of the best repositories of such reports available. Just about every criminal justice sector will also have a professional magazine (e.g., *Police Chief*, *Corrections Today*, etc.) or professional association Web site (e.g., American Correctional Association, American Jail Association, etc.) where articles regarding applied studies of personnel issues are published.

Conclusion

This article provides a very brief overview of some areas of applied research in the criminal justice/criminology area. As noted earlier, such research is carried out regularly around the world by governments, university-based faculty researchers, and private-sector researchers. Much applied research is intended to address specific issues and problems faced by agencies, not necessarily for peer-reviewed articles and texts. The research skills and knowledge necessary to conduct applied research on behalf of agencies are the same as those required to produce 'scholarly' research. To paraphrase the old *Dragnet* opening, 'only the audience is changed' in most applied research. Rather than informing other academics, the purpose is first to inform decision makers at organizational levels.

In the end, applied research of all types still needs to be rigorous and ethical. How it is valued by the academic world and the applied world is often a contentious matter, especially when it does not undergo peer review. For the end consumers of the information, the rigor and the utility of the information for their interests will ultimately dictate the value of the work.

> The problems of crime bring us together. Even as we join in common action, we know there can be no instant victory. Ancient evils do not yield to easy conquest. We cannot limit our efforts to enemies we can see. We must, with equal resolve, seek out new knowledge, new techniques, and new understanding.
>
> Message from President Johnson to the Congress, 9 March 1966

Acknowledgments

The authors thank Dr Timothy Coupe, Professor Sir Anthony Bottoms, and Dr John Walker for their assistance in preparing this article.

See also: Applied Sociology; Evidence-Based Practice; Implementation Science; Organizational Climate in the Work Setting; Personnel Selection, Psychology of; Prevention Research; Program Evaluation.

Bibliography

Akers, T.A., Potter, R.H., Hill, C.V., 2013. Epidemiological Criminology: A Public Health Approach to Crime and Violence. Jossey-Bass, San Francisco.

Bonta, J., Rugge, T., Dauvergne, M., 2003. The Reconviction Rate of Federal Offenders (User Report 2003-02). Solicitor General Canada, Ottawa. Available at: http://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/rcvd-fdffndr/index-eng.aspx.

Andrews, D.A., Bonta, J., 2010. The Psychology of Criminal Conduct, fifth edition. Lexis-Nexis Group, Providence, NJ.

Brantingham, P., Faust, F., 1976. A conceptual model of crime prevention. Crime & Delinquency 22 (3), 284–296.

Chan, J.B.L., 2001. The technological game: how information technology is transforming police practice. Criminology and Criminal Justice 1 (2), 139–159.

Chappell, D., Wilson, P., 1993. The Australian Criminal Justice System: The Mid 1990s. Butterworth, Sydney.

Chilton, R., 2001. Viable policy: the impact of federal funding and the need for independent research agendas—the American Society of Criminology 2000 presidential address. Criminology 39 (1), 1–8.

Durose, M.R., Cooper, A.D., Snyder, H.N., 2014. Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010. Bureau of Justice Statistics, Washington, DC. At: http://www.bjs.gov/content/pub/pdf/rprts05p0510.pdf (accessed 29.04.14.).

Farrington, D.P., Welsh, B.C., 2007. Saving Children from a Life of Crime: Early Risk Factors and Effective Interventions. Oxford University Press, New York.

Fixsen, D.L., Naoom, S.F., Blase, K.A., Friedman, R.M., 2005. Implementation Research: A Synthesis of the Literature. University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network, Tampa, FL.
Gau, J.M., 2013. Procedural justice and police legitimacy: a test of measurement and structure. American Journal of Criminal Justice 1–19. http://dx.doi.org/10.1007/s12103-013-9220-8.
Geis, G., Mobley, A., Shichor, D., 1999. Private prisons, criminological research, and conflict of interest: a case study. Crime & Delinquency 45 (3), 372–388.
Glueck, S., Glueck, E., 1959. Predicting Delinquency and Crime. Harvard University Press, Cambridge, MA.
Gottfredson, G.D., 1984. A theory-ridden approach to program evaluation: a method for stimulating researcher-implementer collaboration. American Psychologist 39 (10), 1101–1112.
Grossi, D., 2008. Less lethal use of force: the continuum of force and less-lethal options. Law Enforcement Executive Forum 8 (4), 1–6.
Israel, M., 2000. The commercialisation of university-based criminological research in Australia. Australian & New Zealand Journal of Criminology 33 (1), 1–20.
Jefferson, T., Shapland, J., 1994. Criminal justice and the production of order and control: criminological research in the UK in the 1980s. British Journal of Criminology 34 (3), 265–290.
Jupp, V., Davies, P., Francis, P., 2000. Doing Criminological Research. Sage Publications, Inc., London.
Krisberg, B., Marchionna, S., Baird, C. (Eds.), 2007. Continuing the Struggle for Justice: 100 Years of the National Council on Crime and Delinquency. Sage Publications, Inc., Thousand Oaks, CA.
Lambert, E., Paoline III, E.A., 2010. Take this job and shove it: an exploratory study of turnover intent among jail staff. Journal of Criminal Justice 38 (2), 139–148.
Lewin, K., 1946. Action research and minority problems. Journal of Social Issues 2, 34–46. http://dx.doi.org/10.1111/j.1540-4560.1946.tb02295.x.
Lipsey, M.W., 1992. The effect of treatment on juvenile delinquents: results from a meta-analysis. In: Lösel, F., Bender, D., Bliesener, T. (Eds.), Psychology and Law: International Perspectives. Walter de Gruyter, New York, pp. 131–143.
Lodge, T.S., 1974. The founding of the home office research unit. In: Hood, R. (Ed.), Crime, Criminology and Public Policy: Essays in Honour of Sir Leon Radzinowicz. The Free Press, New York, pp. 11–24.
Lowenkamp, C.T., Latessa, E.J., Holsinger, A.M., 2006. The risk principle in action: what have we learned from 13,676 offenders and 97 correctional programs? Crime & Delinquency 52 (1), 77–93.
Martinson, R., 1974. What works?: questions and answers about prison reform. The Public Interest 35, 22–54.
Mukherjee, S.K., 1986. The nature of crime trends in Australia. In: Chappell, D., Wilson, P.R. (Eds.), The Australian Criminal Justice System – The Mid-1980s. Butterworths, Sydney, pp. 18–38.
Nettler, G., 1984. Explaining Crime, third ed. McGraw-Hill, New York.
Neyroud, P., Weisburd, D., Spring 2014. Transforming the police through science: some new thoughts on the controversy and challenge of translation. Translational Criminology 16–18. Available at: http://cebcp.org/wp-content/TCmagazine/TC6-Spring2014 (accessed 30.04.14.).
Nouwens, T., Motiuk, L., Boe, R., 1993. So you want to know the recidivism rate. FORUM on Corrections Research 5, 3.
Payne, J., 2007. Recidivism in Australia: Findings and Future Research. Australian Institute of Criminology, Canberra. Available at: http://www.aic.gov.au/documents/0/6/B/%7B06BA8B79-E747-413E-A263-72FA37E42F6F%7Drpp80.pdf.
Potter, R.H., Krider, J., 2000. Teaching about violence prevention: a bridge between public health and criminal justice educators. Journal of Criminal Justice Education 11 (2), 339–351.
Ratcliffe, J., 2008. Intelligence-Led Policing. Willlan Publishing, Portland, OR.
Ratcliffe, J., Taniguchi, T., Groff, E.R., Wood, J.D., 2011. The Philadelphia foot patrol experiment: a randomized controlled trial of police patrol effectiveness in violent crime hotspots. Criminology 49 (3), 795–831.
Rossi, P.H., Freeman, H.E., Lipsey, M.W., 1999. Evaluation: A Systematic Approach, sixth ed. Sage Publications, Thousand Oaks, CA.
Sherman, L.W., 1992. Policing Domestic Violence: Experiments and Dilemmas. Free Press, New York.
Sherman, L.W., Flynn, E., Spring 2014. Arrests for misdemeanor domestic abuse: a crucible of evidence-based policing. Translational Criminology 20–21. Available at: http://cebcp.org/wp-content/TCmagazine/TC6-Spring2014 (accessed 30.04.14.).
Smith, P., Genreau, P., Swartz, K., 2009. Validating the principles of effective intervention: a systematic review of the contributions of meta-analysis in the field of corrections. Victims and Offenders 4 (2), 148–169.
Stinchcomb, J.B., 2008. The Future Is Now: Recruiting, Retaining, and Developing the 21st Century Jail Workforce. Florida Atlantic University, Ft. Lauderdale. Available at: http://www.cipp.org/pdf/JailsWorkforceLiteraturereview71308.pdf (accessed 30.04.14.).
Stinchcomb, J.B., McCampbell, S.W., Layman, E.P., 2006. Future Force: A Guide to Building the 21st Century Community Corrections Workforce. Center for Innovative Public Policies, Inc., Naples, FL. Available at: https://www.ncjrs.gov/App/abstractdb/AbstractDBDetails.aspx?id=237476 (accessed 30.04.14.).
Taxman, F.S., Belenko, S., 2012. Implementing Evidence-Based Practices in Community Corrections and Addiction Treatment. Springer, New York.
Terrell, W., Paoline, E.A., Ingram, J.R., 2008. Nonlethal use-of-force policy: perceptions from the front line. Law Enforcement Executive Forum 8 (4), 7–20.
United States Government Printing Office, 1967. The Challenge of Crime in a Free Society. USGPO, Washington, DC. At: https://www.ncjrs.gov/pdffiles1/nij/42.pdf (accessed 10.04.14.).
Walker, S., Alpert, G.P., Kenney, D.J., 2000. Early warning systems for police: concept, history and issues. Police Quarterly 3 (2), 132–152.
Wolf, R., Gordon, A., 2007. License plate recognition (LPR) technology: practical and legal issues for police implementation. Law Enforcement Executive Forum 7 (7), 127–142.

Relevant Websites

http://www.abtassociates.com/ – Abt Associates.
http://www.aic.gov.au – Australian Institute of Criminology.
http://www.comlaw.gov.au/Details/C2012C00049/Html/Text#_Toc310437623 – Commonwealth of Australia.
Jill Dando Institute: http://www.ucl.ac.uk/jdi (accessed 30.04.14.).
MacArthur Foundation: http://www.macfound.org/programs/juvenile_justice/strategy/ (accessed 30.04.14.).
National Council on Crime & Delinquency: http://www.nccdglobal.org/ (accessed 30.04.14.).
National Crime Prevention Centre: http://www.publicsafety.gc.ca/cnt/cntrng-crm/crm-prvntn/ntnl-crm-prvntn-cntr-eng.aspx (accessed 14.05.14.).
National Institute of Corrections: http://www.nicic.gov.
National Institute of Justice: http://www.nij.gov/Pages/welcome.aspx (accessed 14.05.14.).
Pew Center on the States: http://www.pewstates.org/issues/public-safety-328137 (accessed 30.04.14.).
Research Triangle Institute: http://www.rti.org/ (accessed 30.04.14.).
Urban Institute: http://urban.org/justice/index.cfm (accessed 30.04.14.).
Vera Institute: http://www.vera.org/about-us (accessed 30.04.14.).
Westat: https://www.westat.com/ (accessed 30.04.14.).

Crime and Gender

Samantha R Cumley, University of Michigan-Flint, Flint, MI, USA
Karen Heimer, University of Iowa, Iowa City, IA, USA
Stacy De Coster, North Carolina State University, Raleigh, NC, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by K. Heimer, S. De Coster, volume 5, pp. 2918–2921, © 2001, Elsevier Ltd.

Abstract

Research consistently finds that females are much less likely than males to engage in illegal behavior. This article summarizes scholarship examining the relationship between gender and crime, which often follows one of three approaches: (1) Macro-level research examines national trends in women's offending compared to men's, emphasizing explanations for changes in the gender–crime gap over time. (2) Qualitative research on women's offending focuses on how social environments influence women's decisions to commit crime. And, (3) Individual-level analyses of survey data center on sociopsychological explanations of why males are more likely than females to break the law.

One of the strongest findings in the criminological literature is that there is a gender gap in crime and delinquency. Research consistently finds that females are much less likely than males to engage in illegal behavior. Scholarship examining gender and crime tends to follow one of three approaches: (1) Research at the macro-level examines national trends in women's offending compared to men's, assessing explanations of the change over time in the gender–crime gap. (2) Qualitative research illuminates how social environments shape women's decisions to commit crime. (3) Individual-level analyses of survey data center on sociopsychological explanations of why males are more likely than females to break the law. This article summarizes these approaches to studying gender and crime. It also discusses avenues for bridging macro- and individual-level explanations. The conclusions suggest areas in need of further research.

The 'gender gap' – or the significantly greater involvement of males than females in crime and delinquency – is one of the most robust findings in criminology. This gap has been identified across all types of data sources, including official data on arrests and criminal convictions, self-report data on delinquency and crime, and victimization data in which victims report the sex of perpetrators. Moreover, the gender gap appears to transcend time and space, as it has been documented cross nationally and over time (e.g., Smith and Visher, 1980; Lauritsen et al., 2009).

Although the persistence of a crime–gender gap has long been widely acknowledged, gender differences in crime received little attention in the empirical literature until the 1980s. Following the social movements of the 1960s, feminist criminologists in the 1970s critiqued the absence of research on female offending (e.g., Adler, 1975; Simon, 1975; Smart, 1976). Since that time, there has been a movement toward more serious considerations of female offending and the gender gap in crime and delinquency.

This article summarizes the research on crime and gender that has been conducted following the feminist critiques of the 1970s. It is organized around three typical approaches to the study of gender and crime and discusses key findings and arguments of each. Given space limitations, the article focuses on research on the causes of gender differences in offending and does not address gender and victimization or gender and criminal justice processing. For a review of those topics, see Kruttschnitt (1996).

Approaches to Studying Gender, Crime, and Delinquency

As with much social scientific research, empirical studies of gender and offending can be categorized according to levels of analysis, which guide the specific research questions addressed. One line of research focuses on the macro-level of analysis, examining rates of crime over time and across geographical areas. Research at this level seeks to explain the crime–gender gap as it is rooted in broad social patterns. Much of this macro-level research focuses on changes in the gender gap in crime rates over time and asks whether there has been convergence between female and male crime rates. This line of research suggests that there has been some convergence since the 1950s and 1960s in various nations, including England and Wales, Canada, and the United States (Box and Hale, 1983; Fox and Hartnagel, 1979; Steffensmeier, 1993). In some years during this time period, female crime rates increased faster than male rates, and at other times, male crime rates experienced sharper declines than female rates (Heimer, 2000). Moreover, the convergence in female and male rates is more pronounced during some periods for certain property crimes, such as larceny and fraud (Steffensmeier, 1993). Macro-level studies most often examine annual arrest or conviction rates, which are aggregated to the national level. The explanations offered tend to focus on women's liberation and worsening economic circumstances due to increases in female headship and other factors.

A second line of research utilizes individual-level data and sociopsychological explanations to account for gender differences in offending. Studies of this type often use quantitative data from self-report surveys of large national or local samples; most focus on juvenile delinquency. These surveys demonstrate that males are more likely than females to commit both serious

and minor violations of the law, but that the gender gap is most pronounced for the most serious and violent offenses. This line of work focuses on how sociopsychological factors – including family dynamics, beliefs, and attitudes – may contribute to the gender gap in crime and delinquency.

Beyond the distinction based on levels of analysis, research on gender and crime can be categorized by the type of methodology used. While the two lines of research discussed previously rely almost exclusively on quantitative data, a third line of research develops rich explanations for female offending based on qualitative analyses, reporting findings from open-ended interviews and ethnographic observations. Rather than examining female offending vis-à-vis male offending, this line of research most often focuses on female offenders, thus addressing differences in crime between groups of women, rather than comparing across gender.

The remainder of this article briefly reviews the three lines of research identified herein. It then examines attempts to bridge the macro- and individual-levels of analysis. It concludes by suggesting that future research continue to attempt to bridge levels of analysis, and to synthesize quantitative and qualitative research findings. In particular, the article argues that qualitative studies offer great potential to uncover the context and content of race, class, and gender inequalities that affect decisions to engage in crime and delinquency.

Research on the Narrowing of the Gender Gap over Time

Research on the narrowing of the gender gap in crime rates over time offers four explanations – women's liberation, economic marginalization, decline in chivalry, and domestic violence law enforcement. The liberation hypothesis proposes that women's crime rates increased during the 1960s and 1970s because women gained increased economic and social independence from men. One variant proposes that a convergence in gender roles led to a convergence in the gender gap in all types of crime (Adler, 1975; Hunnicutt and Broidy, 2004). Another variant proposes that women's increased labor force participation created new opportunities for crime in the workplace (Simon, 1975). Both variants have been criticized for assuming that labor force participation and improved economic conditions increase crime by females, whereas these conditions reduce crime by males (Chesney-Lind, 1997). Moreover, the narrowing of the gender gap is most pronounced for crimes that are consistent with women's preliberation roles – including larceny (theft), fraud, and forgery (see Steffensmeier, 1993). In addition, empirical research provides little evidence that women's labor force participation increases their crime (e.g., Box and Hale, 1983; for review, see Steffensmeier and Allan, 1996).

Consequently, contemporary studies often focus on the economic marginalization hypothesis, which proposes that the increased financial hardship of women relative to men has contributed to the narrowing gender gap in crime (Box, 1987; Steffensmeier, 1993). In general, these studies reveal that female arrest rates are often higher in cities where there are higher levels of poverty, joblessness, female-headed households, and other indicators of socioeconomic disadvantage, just as is the case for male arrest rates (Steffensmeier and Haynie, 2000; Heimer et al., 2006). While these findings suggest similarity in the structural factors affecting men's and women's arrest rates, little research has examined the effects of sex-specific measures of economic marginalization on women's arrest rates vis-à-vis men's. Future research should address the interplay between the specific elements of women's poverty that affect their crime rates compared to men's, including increasing rates of single motherhood, persistent gender inequality in wages, and changes in social welfare policies (Heimer, 2000; Heimer et al., 2006).

A third explanation of the narrowing of the gender–crime gap is the decline of chivalry hypothesis, which proposes that the convergence of female and male arrest rates is due in part to changing attitudes among law enforcement personnel. Specifically, the gender gap in female-to-male arrest rates reflects the chivalry of law enforcement personnel toward women (e.g., hesitation to arrest women), and the narrowing of the gap reflects a decline in this attitude over the years (Steffensmeier et al., 2005, 2006). Empirical evidence suggests that the decline of chivalry likely has had some effect on changes in the gender gap in crime, although economic factors are at least as important (Box and Hale, 1983; Steffensmeier and Allan, 1996; Steffensmeier and Haynie, 2000). However, the veracity of this hypothesis remains the subject of scholarly debate, with some researchers arguing that females are treated more harshly than males for some types of offenses (Chesney-Lind, 1997; Steffensmeier and Schwartz, 2004).

Relatedly, some scholars have argued that mandatory arrest policies in domestic violence cases may have led to more arrests of women over time as the arrest of both parties in these cases increased (Chesney-Lind, 2002; Steffensmeier et al., 2006). As a result of this shift, female arrest rates would presumably increase more than male rates. However, recent research by Lauritsen and colleagues (2009) compares Uniform Crime Report (UCR) violent crime arrest rate data over time with estimates of violent offending rates based on victim reports of offenders' gender using the National Crime Victimization Survey (NCVS) data over time. This study demonstrates that UCR arrest data and NCVS victim reports of offenders' sex produce very similar trends over time in the gender gap (Lauritsen et al., 2009). Given that victim report data are not subject to changes in police practice (Steffensmeier et al., 2006), the findings of Lauritsen and colleagues (2009) suggest that gender gap convergence in arrest rates is not entirely attributable to changes in police behavior.

Qualitative Research on Women's Offending

Qualitative studies of gender and crime typically do not focus simultaneously on male and female offenders. Instead, scholars interested in gender recently have begun to focus particularly on women offenders without directly confronting the gender gap in offending. This research offers in-depth analyses of the structural and situational factors associated with chronic offending among women. As a group, these studies suggest that poverty, unemployment, and single motherhood combine to create conditions of extreme economic hardship, which in turn are associated with events

and conditions of women's lives that increase the chances of a life of crime (Miller, 1986; Carlen, 1988; Daly, 1994; Richie, 1996; Maher, 1997; Baskin and Sommers, 1998; Miller, 2001). Consistent with the economic marginalization hypothesis, many qualitative studies find that female offenders are often socioeconomically disadvantaged. In general, qualitative studies report that female offenders are most often poor, unmarried, unemployed, and less educated (e.g., Wolfe et al., 1984). Of the women in this group who are employed, many work in low-paid service sector jobs without benefits. Given labor market inequalities along racial and class lines, women of color are also overrepresented (Campbell, 1984; Miller, 1986; Carlen, 1988; Daly, 1994; Maher, 1997; Baskin and Sommers, 1998). Similarly, other studies examine how living in a marginalized community facilitates entry into criminal networks and peer groups, which play an important role in initiating women into lives of crime and sustaining these lifestyles in the future (Miller, 1986; Baskin and Sommers, 1998; Miller, 2001). For example, several studies suggest that turning to crime is a rational attempt to support children and other family members when confronted with severe economic hardship (Miller, 1986; Carlen, 1988; Mullins and Wright, 2003).

In addition, many studies emphasize the connections between abuse during childhood, economic hardships, and criminal careers among women (Chesney-Lind and Rodriguez, 1983; Miller, 1986; Daly, 1994; Richie, 1996; Maher, 1997). Specifically, some researchers report that between one-third and nearly one-half of the women and girls participating in their studies experienced sexual and/or physical abuse before becoming adults (Daly, 1994; Miller, 2001). However, some scholars have proposed that focusing on women's victimization experiences as a primary factor driving their criminal offending is overly deterministic and obscures the choices women and girls make when becoming involved in crime and delinquency (Maher, 1997; Miller, 2001). Indeed, many studies find that women and men perpetrate crime and violence for similar underlying reasons (e.g., financial gain, perceived disrespect) (Miller, 2001, 2008; Mullins and Wright, 2003; Miller and Mullins, 2005; Kruttschnitt and Carbone-Lopez, 2006; Jones, 2010). These studies offer rich insight into the social world of female chronic offenders and, as we note in Section Bridging Levels of Explanation, open the door for the development of theories that can address macro-level as well as sociopsychological processes.

Self-Report Research on Gender and Juvenile Delinquency

Most research based on self-report surveys focuses on juvenile delinquency, especially common and less serious forms of delinquency. Much of the research in this tradition has focused on the application of and extension of existing criminological theory to explain sex differences in juvenile delinquency (for a review, see De Coster et al., 2012). The emphasis in this line of work typically is on understanding the sociopsychological mechanisms that contribute to the gender gap in delinquency. Recent research on these mechanisms suggests that the gender gap reflects both gender differences in exposure to factors that make law violation more likely as well as gender differences in the impact of these factors.

The sociopsychological mechanisms that have received the most attention in research on gender and delinquency occur within the family. Specifically, some research shows that part of the gender gap in delinquency is explained by the greater exposure of girls than boys both to direct parental controls, such as supervision and monitoring, and to indirect controls, such as emotional bonding (e.g., Jensen and Eve, 1979; Hagan et al., 1985, 1987; Chapple et al., 2010). Other research shows that the gender gap is not the simple outcome of girls being subject to higher levels of familial controls but also reflects the differential impact of these controls on girls versus boys. For instance, even though girls are more closely supervised, boys' delinquency appears to be influenced more strongly by supervision and monitoring; in contrast, girls' delinquency is influenced more strongly by emotional bonds to their families (Heimer and De Coster, 1999; Hagan, 1989; Chapple et al., 2005; alternatively see Kruttschnitt, 1996). In short, it may be that parents control girls' misbehavior through subtle, covert control mechanisms (e.g., emotional bonding), whereas controlling boys' misbehavior requires more direct, overt strategies (e.g., supervision).

Another individual-level factor that has received attention is gender roles. The argument, related to the liberation thesis discussed before, is that feminine roles restrain delinquency more than masculine roles do (e.g., Shover and Norland, 1979). Many studies of gender roles and delinquency appeared in the late 1970s. However, the empirical research was inconsistent, perhaps because gender roles were operationalized in diverse ways across studies, often including traits, attitudes, and familial controls as well as behavioral expectations or roles.

The more recent trend has been to focus on cultural definitions of gender or hegemonic gender definitions, which are widely accepted beliefs and attitudes that support the subordination of females to males under patriarchy. Hegemonic gender definitions depict a feminine ideal that includes a high capacity for nurturance, passivity, connectedness to others, and physical weakness; in contrast, the masculine ideal emphasizes competitiveness, aggressiveness, independence, and strength. Definitions of femininity thus are more inconsistent than definitions of masculinity with victimizing others physically or taking their property. Some empirical research shows that acceptance of these gender definitions is an important contributor to the gender gap in law violation (Simpson and Elis, 1995; Heimer and De Coster, 1999; McCarthy et al., 1999). Other research suggests that boys who have internalized hegemonic definitions of masculinity are more likely to engage in crime and violence (Messerschmidt, 1993, 2000; Simpson and Elis, 1995; Heimer, 1996; Miller, 2001; Mullins et al., 2004; alternatively, see Morash and Chesney-Lind, 1991).

A third individual-level factor that has received attention in the literature on gender and delinquency is attitudes toward risk taking. The hypothesis here is that youths who develop a taste for risk will be more likely than others to engage in thrill-seeking behavior, including delinquency. The power-control theory of gender and delinquency, discussed

following, proposes that one reason males are more likely than females to be delinquent is that males are socialized to prefer risk taking (Hagan 1989; McCarthy et al., 1999; Blackwell et al., 2002; Hagan et al., 2004).

Bridging Levels of Explanation

To date, most research on gender and criminal offending has focused on either the macro-level or sociopsychological explanations discussed previously and rarely has pursued links between the two. One notable exception is Hagan's power-control theory, which argues that parents' positions in the workplace shape their parenting practices, including their supervision of children, emotional bonding with children, and the gendered socialization of children (Hagan et al., 1985, 1987). More specifically, when fathers have more power than mothers in their work or professional lives, mothers experience paternalistic control within the family. Mothers reproduce this arrangement with their children by subjecting daughters to higher levels of control than sons, encouraging risk preferences more among sons than daughters, and inculcating sons and daughters with traditional gendered schemas or ideologies (Hagan et al., 1987; McCarthy et al., 1999; Hagan et al., 2004). When parents have more equal power in their work lives, they engage in a more egalitarian partnership that they pass on to their sons and daughters. That is, sons and daughters in these more egalitarian families are treated rather similarly, resulting in a narrowing of the gender gap in law violation. Some research supports these arguments (Blackwell and Reed, 2003); however, others fail to replicate key findings from the theory (Singer and Levine, 1988; Jensen and Thompson, 1990; Blackwell, 2000).

Another emerging perspective on gender and crime with the potential to bridge levels of analysis is the work of Messerschmidt (e.g., 1993, 2000). He argues that males may attempt to display masculinity through crime and violence when legitimate channels for claiming gender are blocked by structural barriers, such as unemployment and poverty (Messerschmidt, 1993). Messerschmidt (2000) further explores links between structural barriers and sociopsychological mechanisms in his analysis of boys' violence. He argues that much of this violence is a result of the interplay between patriarchal structures, social class, and sociopsychological processes, such as alienation at school, victimization, body image, and hegemonic masculine ideals. While this work moves toward bridging levels in analyzing masculinity and crime, it does not address femininity and crime. Jody Miller (2002) argues, for instance, that gender is overemphasized in Messerschmidt's structured action theory, suggesting that violence among women and girls cannot be explained effectively with an emphasis on gendered displays. For example, several studies find that girls who engage in violent behavior are not necessarily attempting to claim gender, but rather to maintain their hard-earned respected positions as females in male-dominated groups (Messerschmidt, 1997, 2012; Miller, 2001; Jones, 2010).

Some of the qualitative research discussed in Section Qualitative Research on Women's Offending also could serve as a springboard for the development of theories that bridge the macro- and sociopsychological levels of analysis. Many of these studies examine how economic hardship and cultural contexts influence individuals' perceptions and decisions about crime (e.g., Miller, 1986; Maher, 1997; Baskin and Sommers, 1998; Miller, 2001; Kruttschnitt and Carbone-Lopez, 2006). As De Coster and Heimer (2006: 147) observe, intersecting inequalities influence behavior in ways that scholars are only beginning to understand. Future research must consider definitions of situations as well as role-taking processes to understand the ways that social structures, such as race, class, and gender inequality influence individual actions. The next step would be to account for these findings by specifying a theory of the links between the macro- and individual-level mechanisms leading to women's crime, and then broaden this explanation to address gender differences in crime.

Directions for Future Research

Up to the end of the twentieth century, the strands of research discussed herein have remained relatively isolated from one another. In addition, research was focused more on uncovering empirical relationships than on building comprehensive theories of gender, crime, and delinquency. Future research will need to develop rigorous theoretical arguments that link levels of analysis, drawing links between the mechanisms identified by macro-level research on changes in the gender gap and individual-level survey research. Moreover, research must bridge the divide between quantitative and qualitative research, combining findings from both to generate more complete explanations of gender, crime, and delinquency. In particular, quantitative methods may be more useful for uncovering the factors underlying structural-level patterns of crime as they emerge from the intersections of race, class, and gender inequalities, and qualitative research is critical for adding depth and greater understanding of how these factors combine to explain offending in the daily life experiences of women.

See also: Crime and Class; Crime, Race and Ethnicity; Crime, Sociology of; Delinquency, Sociology of; Domestic Violence: Sociological Perspectives; Feminist Criminology and Gender Studies; Intersectionality and Crime.

Bibliography

Adler, Freda, 1975. Sisters in Crime. McGraw-Hill, New York.

Baskin, Deborah R., Sommers, Ira, 1998. Casualties of Community Disorder. Westview Press, Boulder, CO.

Blackwell, Brenda Sims, 2000. Perceived sanction threats, gender, and crime: a test and elaboration of power-control theory. Criminology 38, 439–488.

Blackwell, Brenda Sims, Reed, Mark D., 2003. Power-control as a between- and within-family model: reconsidering the unit of analysis. Journal of Youth and Adolescence 32, 385–400.

Blackwell, Brenda Sims, Sellers, Christine S., Schlaupitz, Sheila M., 2002. A power control theory of vulnerability to crime and adolescent role exits—revisited. Canadian Review of Sociology and Anthropology 39, 199–218.

Box, Steven, 1987. Recession, Crime and Punishment. MacMillan Education, London.

Box, Steven, Hale, Chris, 1983. Liberation and female criminality in England and Wales. British Journal of Criminology 23, 35–49.

Campbell, Anne, 1984. The girls in the gang: a report from New York City. Basil Blackwell, Oxford, UK.

Carlen, Pat, 1988. Women, Crime and Poverty. Open University Press, Milton Keynes, UK.

Chapple, Constance L., McQuillan, Julia A., Berdahl, Terceira A., 2005. Gender, social bonds, delinquency: a comparison of girls' and boys' models. Social Science Research 34, 357–383.

Chapple, Constance L., Vaske, Jamie, Hope, Trina L., 2010. Sex differences in the causes of self-control: an examination of mediation, moderation, and gendered etiologies. Journal of Criminal Justice 38, 1122–1131.

Chesney-Lind, Meda, 1997. The Female Offender: Girls, Women and Crime. Sage, Thousand Oaks, CA.

Chesney-Lind, Meda, 2002. Criminalizing victimization: the unintended consequences of pro-arrest policies for girls and women. Criminology & Public Policy 2, 81–90.

Chesney-Lind, Meda, Rodriguez, Noelie, 1983. Women under lock and key: a view from the inside. The Prison Journal 63, 47–65.

Daly, Kathleen, 1994. Gender, Crime, and Punishment. Yale University Press, New Haven.

De Coster, Stacy, Heimer, Karen, 2006. Crime at the intersection: gender, race, and violent off ending. In: Peterson, Ruth, Krivo, Lauren, Hagan, John (Eds.), The Many Colors of Crime: Inequalities of Race, Ethnicity and Crime in America. New York University Press, New York.

De Coster, Stacy, Heimer, Karen, Cumley, Samantha, 2012. Girls, gender, and delinquency. In: Wilcox, Pamela, Cullen, Francis T. (Eds.), The Oxford Handbook of Criminological Theory. Oxford University Press, New York, pp. 313–330.

Fox, John, Hartnagel, Timothy, 1979. Changing social roles and female crime in Canada: a time series analysis. Canadian Review of Sociology and Anthropology 16, 96–104.

Hagan, John, 1989. Structural Criminology. Rutgers University Press, New Brunswick, NJ.

Hagan, John, Boehnke, Klaus, Merkens, Hans, 2004. Gender differences in capitalization processes and the delinquency of siblings in Toronto and Berlin. British Journal of Criminology 44, 659–676.

Hagan, John, Gillis, A.R., Simpson, John, 1985. The class structure of gender and delinquency: toward a power control theory of common delinquent behavior. American Journal of Sociology 90, 1151–1178.

Hagan, John, Simpson, John, Gillis, A.R., 1987. Class in the household: a power control theory of gender and delinquency. American Journal of Sociology 92, 788–816.

Heimer, Karen, 1996. Gender, interaction, and delinquency: testing a theory of differential social control. Social Psychological Quarterly 59, 39–61.

Heimer, Karen, 2000. Changes in the gender gap in crime and women's economic marginalization. In: Criminal Justice 2000, vol. 1. The Nature of Crime: Continuity and Change, Washington, D.C.

Heimer, Karen, De Coster, Stacy, 1999. The gendering of violent delinquency. Criminology 37, 277–317.

Heimer, Karen, Wittrock, Stacy, Unal, Halime, 2006. Crimes of poverty: economic marginalization and the gender gap in crime. In: Heimer, Karen, Kruttschnitt, Candace (Eds.), Gender and Crime: Patterns of Victimization and Offending. New York University Press, New York.

Hunnicutt, Gwen, Broidy, Lisa M., 2004. Liberation and economic marginalization: a reformulation and test of (formerly?) competing models. Journal of Research in Crime and Delinquency 41, 130–155.

Jensen, Gary, Eve, Raymond, 1979. Sex differences in delinquency: an examination of popular sociological explanations. Criminology 13, 427–448.

Jensen, Gary, Thompson, Kevin, 1990. What's class got to do with it? A further examination of power-control theory. American Journal of Sociology 95, 1009–1023.

Jones, Nikki, 2010. Between Good and Ghetto: African American Girls and Inner-City Violence. Rutgers University Press, New Jersey.

Kruttschnitt, Candace, 1996. Contributions of quantitative methods to the study of gender and crime, or bootstrapping our way into the theoretical thicket. Journal of Quantitative Criminology 12, 135–161.

Kruttschnitt, Candace, Carbone-Lopez, Kristin, 2006. Moving beyond the stereotypes: women's subjective accounts of their violent crime. Criminology 44, 321–348.

Lauritsen, Janet L., Heimer, Karen, Lynch, James P., 2009. Trends in the gender gap in violent offending: new evidence from the national crime victimization survey. Criminology 47, 361–399.

Maher, Lisa, 1997. Sexed Work: Gender, Race, and Resistance in a Brooklyn Drug Market. Clarendon Press, New York.

McCarthy, Bill, Hagan, John, Woodward, Todd S., 1999. In the company of women: structure and agency in a revised power-control theory of gender and delinquency. Criminology 37, 761–788.

Messerschmidt, James, 1993. Masculinities and Crime: Critique and Reconceptualization of Theory. Rowman & Littlefield, Lanham, MD.

Messerschmidt, James, 1997. Crime and Structured Action: Gender, Race, Class, and Crime in the Making. Sage, Thousand Oaks, CA.

Messerschmidt, James, 2000. Nine Lives: Adolescent Masculinities, the Body and Violence. Westview Press, Boulder, CO.

Messerschmidt, James, 2012. Gender, Heterosexuality, and Youth Violence: The Struggle for Recognition. Rowman and Littlefield, Lanham, MD.

Miller, Eleanor, 1986. Street Woman. Temple University Press, Philadelphia.

Miller, Jody, 2001. One of the Guys: Girls, Gangs and Gender. Oxford University Press, New York.

Miller, Jody, 2002. The strengths and limits of 'doing gender' for understanding street crime. Theoretical Criminology 6, 433–460.

Miller, Jody, 2008. Getting Played: African American Girls, Urban Inequality, and Gendered Violence. New York University Press, New York.

Miller, Jody, Mullins, Christopher W., 2005. Stuck up, telling lies and talking too much: the gendered context of young women's violence. In: Heimer, Karen, Kruttschnitt, Candace (Eds.), Gender and Crime: Patterns of Victimization and Offending. New York University, New York, pp. 41–66.

Morash, Merry, Chesney-Lind, Meda, 1991. A reformulation and partial test of the power-control theory of delinquency. Justice Quarterly 8, 347–377.

Mullins, Christopher, Wright, Richard, 2003. Gender, social networks, and residential burglary. Criminology 41, 813–839.

Mullins, Christopher, Wright, Richard, Jacobs, Bruce, 2004. Gender, streetlife and criminal retaliation. Criminology 42, 911–940.

Richie, Beth, 1996. Compelled to Crime: The Gender Entrapment of Battered Black Women. Routledge, New York.

Shover, Neal, Norland, Stephen, 1979. Sex roles and criminality: science or conventional wisdom? Sex Roles 4, 111–125.

Simon, Rita, 1975. Women and Crime. Lexington Books, Lexington, MA.

Simpson, Sally, Elis, Lori, 1995. Doing gender: sorting out the caste and crime conundrum. Criminology 33, 47–81.

Singer, Simon, Levine, Murray, 1988. Power-control theory, gender, and delinquency: a partial replication with additional evidence on the effects of peers. Criminology 26, 627–648.

Smart, Carol, 1976. Women, Crime and Criminology: A Feminist Critique. Routeledge and Kegan Paul, London.

Smith, Douglass, Visher, Christy, 1980. Sex and involvement in deviance/crime: a review of the empirical literature. American Sociological Review 45, 691–701.

Steffensmeier, Darrell, 1993. National trends in female arrests: 1960–1990: assessments and recommendations for research. Journal of Quantitative Criminology 9, 411–441.

Steffensmeier, Darrell, Allan, Emilie, 1996. Gender and crime: toward a gendered theory of female offending. Annual Review of Sociology 22, 459–487.

Steffensmeier, Darrell, Haynie, Dana, 2000. Gender, structural disadvantage, and urban crime: do macrosocial variables also explain female offending rates? Criminology 38, 403–438.

Steffensmeier, Darrell, Schwartz, Jennifer, 2004. Trends in female Crime: is crime still a man's world? In: Price, Barbara, Sokoloff, Natalie (Eds.), The Criminal Justice System and Women: Offenders, Victims, and Workers. McGraw-Hill, New York, pp. 96–127.

Steffensmeier, Darrell, Schwartz, Jennifer, Zhong, Hua, Ackerman, Jeff, 2005. An assessment of recent trends in girls' violence using diverse longitudinal sources. Is the gender gap closing? Criminology 43, 355–405.

Steffensmeier, Darrell, Zhong, Hua, Ackerman, Jeff, Schwartz, Jennifer, Agha, Suzanne, 2006. Gender gap trends for violent crimes, 1980–2003. Feminist Criminology 1, 72–98.

Wolfe, Nancy T., Cullen, Francis T., Cullen, John B., 1984. Describing the female offender: a note on the demographics of arrests. Journal of Criminal Justice 12, 483–492.

Crime, Desistance, and Persistence

Michael E Ezell, Northern Illinois University, DeKalb, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article on persistence and desistance in criminal offending examines key areas related to the question of why some criminal offenders stop their offending (desist) while other criminal offenders continue to offend for longer periods of time (persist). Included in the article are discussions of this research literature's foundations in the age–crime curve, common theoretical explanations, and key methodological issues including both definitional and modeling challenges. The discussion concludes by describing key empirical findings related to the association of adult life events with changes in offending patterns and finally identifies areas for future inquiry.

Introduction

Broadly understood, the issues of persistence and desistance boil down to a simple question about criminal offenders: why do some continue to offend (persist) while others stop offending (desist)? Answers to this seemingly simple question have exposed a number of issues that have now generated over two decades of debate among criminologists. The debates have included theoretical explanations, how to define the two terms, the appropriate methodology to study them, and the relevance of adult life events.

The Rise of Interest

Interest in persistence and desistance grew rapidly beginning in the mid-1980s, largely motivated by a desire to understand one of the oldest 'facts' of criminology – the age–crime curve (Piquero et al., 2003). The age–crime curve depicts the following empirical relationship between age and crime: in the aggregate, criminal offending rates rapidly increase during early adolescence, peak during late adolescence or early adulthood (violent crimes typically peak later than property crimes), and then subsequently decline over the remaining age distribution, with the decline generally steeper during the 20s and 30s. Figure 1 graphically illustrates the age–crime curve using the aggregate age-specific arrest rates for burglary in the United States in 2010.

Indeed, it is the shape of the age–crime curve that has yielded interest in the nature of criminal offending across age. Although historically criminologists have focused on adolescent offenders and the differences between those with higher and lower rates of offending, over the last 25 years, more interest has shifted to understanding the 'right tail' of this distribution. Looking beyond the key finding that offending is nearly ubiquitous during adolescence, the declining arrest rate in the right tail indicates that individuals either stop or reduce their offending as they age. Questions are thus posed about why some offenders change (i.e., desist) while others continue offending into adulthood (i.e., persist). Explaining these patterns largely focuses on the role of adult life events (such as marriage and employment) for understanding the processes of both persistence (continuity) and desistance (change).

The relationship between age and crime has also generated considerable debate and controversy, especially in the late 1980s and early 1990s (see Piquero et al., 2003). Much of the controversy surrounding the age–crime curve focuses on a key question: does the aggregate pattern seen in Figure 1 uniformly apply at the individual level? This question posits the possibility that the relationship between age and crime could vary either between individual offenders or between different groups of offenders. This possibility is important because if the relationship depicted at the aggregate level does indeed apply to everyone at the individual level, then the relationship between age and crime would be 'inexplicable' because it does not vary, which would render desistance and persistence as

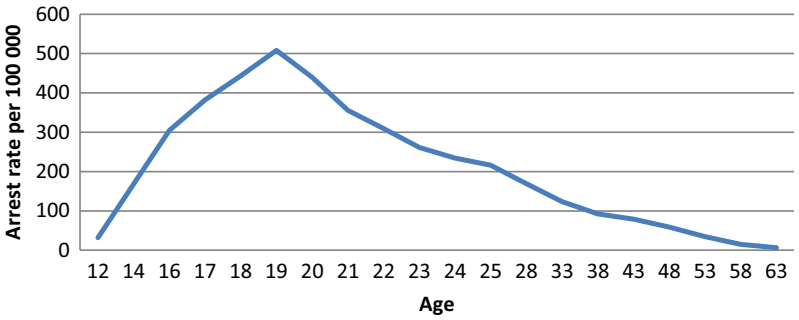

Figure 1 Age-specific arrest rates for burglary, 2010. Source: Snyder, H.N., 2012. Arrest in the United States, 1990–2010. http://www.bjs.gov/index.cfm?ty¼pbdetail&iid¼4515 (accessed 28.10.13.).

uniform processes; however, if it varies at the individual level and/or varies between different groups of offenders, then there would be differences that require both empirical investigation and theoretical explanation.

Three Key Theories

Research investigating issues related to persistence or desistance have often been examined through three broad perspectives: population heterogeneity, state dependence, and developmental taxonomic. These three perspectives are often represented using three specific theoretical explanations, respectively: Gottfredson and Hirschi's (1990) self-control theory, Sampson and Laub's (1993; Laub and Sampson, 2003) informal social control theory, and Moffitt's (1993) dual taxonomy theory. Each of these theories differs in their explanations of the age–crime relationship, the possibility of and/or the explanations for change, and whether life events matter for explaining persistence or desistance.

Gottfredson and Hirschi's self-control theory (1990) articulates the three defining aspects of population heterogeneity theories (Thornberry et al., 2012). First, the causes of crime develop early in the life course; second, these causes are stable (in a relative sense) across age and differences between people persist across time; and third, observed changes with age reflect the direct effect of aging/maturational processes that occur uniformly for all offenders. They argue that a person's level of self-control develops by approximately age 8 years as a consequence of the degree to which they are exposed to effective parenting. After childhood, however, there is relative stability between people in their levels of self-control. Therefore, a person's criminal propensity, once formed, is viewed as relatively immutable to meaningful change. More importantly, all dimensions of criminal offending, including onset, frequency, persistence, and desistance, are argued to be explained by a person's self-control.

Individuals with low self-control are argued to have a higher criminal propensity and to take advantage of opportunities to engage in criminal behavior earlier in the life course, at a higher rate, and for longer periods of time compared to individuals with higher self-control levels. Thus, persisters are individuals with lower levels of self-control, while desisters have higher levels of self-control. Gottfredson and Hirschi argue that all offenders reduce their offending as they age, thereby representing a uniform, natural aging process affecting all offenders at similar ages and rates. Therefore, the relationship between age and crime is said to be 'invariant' between people and 'inexplicable' by sociological and psychological processes.

Further, the authors of this theory argue that people self-select into both life experiences (e.g., whether they are employed in a stable job or in a stable marriage) and social environments on the basis of their level of self-control, and thus the association of life events with crime is spuriously due to self-control differences. Stated differently, the processes of both persistence and desistance are the outcomes of a person's level of self-control.

Sampson and Laub's theory, on the other hand, takes an opposite position and maintains that adult life events, or the lack thereof, are what matter most for explaining both desistance (change) and persistence (continuity). They draw on the core underlying assumptions of both the life course perspective and state dependence theories to argue that a person's level of criminal propensity can change as a consequence of changing social environments and lived social experiences. Furthermore, because the presence, absence, and timing of these changing social environments and life experiences will vary substantially across people, there will be expected variability between people in the relationship between age and crime (Thornberry et al., 2012). Thus, rather than envisioning uniformity in the desistance process, this theoretical framework expects variability between people in their patterns of desistance in terms of both the age at which it begins and the rate at which it progresses.

Sampson and Laub's theory is an informal social control theory that proposes a person's willingness to engage in crime is a function of the current strength of their bonds to the age-graded institutions of informal social control. They primarily focus on explaining patterns of persistence and desistance in adulthood by focusing on two key adult social institutions: marriage and employment. Experiencing these key adult life events, whether by chance or choice, results in changing levels of informal social control in offenders' lives. Importantly, they argue that the quality of the relationships and attachments to these social institutions matters more than their mere occurrence. Desistance is facilitated to the degree that getting married or finding a stable job results in the offender becoming invested in those relationships. Crucial to this perspective is recognizing that changes in offending patterns are characterized as a dynamically evolving process rather than an isolated discrete event.

Persistence in crime, on the other hand, results from the lack of informal social control due to the negative consequences of prior criminal behavior, including prison terms and concomitant labeling effects, that serve to further attenuate a person's exposure to key institutions of informal social control. Engaging in crime is thus seen as exerting a causal effect on the likelihood of future crime as it both weakens the offender's existing social bond and reduces opportunities to experience other life events. In summary, Sampson and Laub envision that within-individual changes in the strength of social bonds across age are the primary mechanisms for both persistence and desistance, and that change is ever present.

Despite the obvious contrasts, both of the prior theories use the same general theoretical framework to explain both persistence and desistance processes. Moffitt's dual taxonomy theory, on the other hand, explains each process separately, with each explanation attached to a single 'type' of offender.

The offender type labeled 'life-course-persistent' is argued to be responsible for the pattern of persistence in offending, as the name aptly implies. Argued to represent less than 10% of the population, this group's persistence in criminal and antisocial behavior is explained as the result of the negative, reciprocal interaction between a child with neuropsychological deficits and parents with ineffective parenting skills and limited resources; this reciprocal interaction ultimately causes the socialization process to fail and yields a person with a stable antisocial behavioral repertoire impervious to change and likely expressed in all life domains. This group is expected to fail in a variety of familial, educational, and interpersonal

outcomes, and indeed has been argued to likely select life experiences that are either consistent with or serve to reinforce their antisocial tendencies rather than provide any opportunity for change; thus, persistence in crime during adulthood defines this group.

Patterns of desistance are explained by the more common offender type, 'adolescence-limited' offenders. Compared to life-course persistent offenders, this group (about 90% of the population) both begins and ends their offending during adolescence and commits less serious offenses. Their rapidly changing likelihood of offending is due to the strain-inducing 'maturity gap' between biological and social maturity found in modern adolescence. Although they have become biologically mature like adults (and capable of reproducing), they remain socially restricted from acting like autonomous adults who get to decide when, where, and what activities they do. This strain is reduced, however, as they legitimately gain access to the adult maturity they have so desired. Thus, with increased independence, the benefits of delinquent/criminal behavior become less rewarding and also threaten their newfound autonomy, and they desist as a result. Thus, desistance defines this group, and given the size of this group, it is their rapidly changing offending patterns that are argued to largely generate the shape of the age–crime curve.

Methodological Issues

The Definition Problem

Among criminologists, there is at least one area of agreement as it relates to both desistance and persistence – scholars agree that there is a lack of consensus on how to define the terms conceptually and operationally (Brame et al., 2003; Bushway et al., 2001, 2003, 2004; Kazemian, 2007; Laub and Sampson, 2001, 2003; Piquero, 2004, 2009; Piquero et al., 2003, 2012). The most central issue in this definition problem concerns the distinction between 'static' and 'dynamic/process' definitions of desistance.

The static definition, which has been the most frequently applied definition in the research, "… counts everyone who offends at least once before a specified cutoff in a follow-up period but not afterward as a desister. Individuals who offend in both periods are termed persisters" (Bushway et al., 2003: 130). In the static definition, desisters are akin to smokers who go 'cold turkey,' and they move abruptly from 'offender' to 'nonoffender,' thereby emphasizing the qualitative state of having 'terminated' their offending. The focus of studies using this definition is a 'between-groups' analysis to examine the measured characteristics associated with membership in the two mutually exclusive groups (desisters and persisters). One significant limitation of this definition is that the short-term follow-up periods in most studies do not allow researchers to distinguish between 'true desistance' (or the true, complete 'termination' of offending) and 'false desistance' (a nonpermanent lull in offending that is eventually followed by subsequent criminal offending) (Bushway et al., 2003).

Broadly stated, the dynamic or process definition views desistance as a discernible reduction in criminal offending. Thus, rather than emphasizing the complete absence of subsequent offending during a particular time period, as with the static definition, the dynamic/process definition emphasizes the "quantitative change in frequency [of offending] from higher to lower levels" (Bushway et al., 2001: 495). Here, desisters are akin to smokers who eventually become nonsmokers by gradually reducing the number of cigarettes smoked over time. Desistance, therefore, is a process unfolding gradually over time and indeed may unfold faster or slower, earlier or later, for different offenders. The focus here is largely on the nature of the changing offending patterns 'within person' over time, and studies using this definition are more interested in examining the observed events and changing characteristics within a person over time that are associated with either initiating and/or enhancing the reduction in offending.

Although the field has expressed a preference for the dynamic definition, disagreement still remains over the precise definition of 'desistance as a process' (Mulvey et al., 2004). At issue is whether desistance as a process means (1) a reduction or a deceleration in frequency or severity of offending (Fagan, 1989), (2) offending must be reduced to a level near zero but not necessarily zero (Brame et al., 2003), (3) a reduction in a latent criminal propensity to a level not significantly different from zero (Bushway et al., 2001), (4) a discrete period of time where offenders maintain the state of being crime-free (Maruna, 2001), or (5) a specific theoretical causal process that underlies the termination of offending (Laub and Sampson, 2001, 2003).

Similar challenges emerge when defining 'persistence' in offending, as it is 'seemingly the conceptual reverse' of desistance (Laub and Sampson, 2003: 17; Piquero, 2009). Laub and Sampson (2003: 150) associate the following words with persistent offending: 'enduring, repetitious, and tenacious.' But exactly how 'enduring,' how 'repetitious,' and how 'tenacious' are not agreed upon, and thus there is no commonly or broadly accepted definition of persistence (Laub and Sampson, 2001; Piquero, 2009). Piquero (2009: 272) asks, "Is persistence made up of offense frequency, chronicity, or a time-span or length of criminal involvement, or some combination thereof?" He continues by positing at least two key conditions necessary to generate a valid definition of persistence. First, persistent offenders, must meet some minimum threshold of offending, and second they must be involved in crime for some minimum number of phases of the life course. However, the specific minimums for those have not been specified or agreed upon.

The concept of 'intermittent offending' adds further ambiguity to these definitional conundrums. Intermittent offenders are those who drift in and out of criminal and noncriminal behaviors in an often termed 'zigzag' style or pattern of offending (Laub and Sampson, 2003). Intermittent offending thus specifically acknowledges the 'lulls' in offending over time. The goal of investigating such intermittent offending patterns is to uncover the factors associated with the back and forth movement between offending and nonoffending. But much like persistence and desistance, there is no accepted definition of intermittent offending (Piquero, 2004). Further, the manner in which it has often been empirically operationalized, such as "people who are exhibiting longer waiting times between offenses" (Baker et al., 2013: 19), can clearly blur the

boundaries between intermittence and both persistence and desistance. Indeed, as Mulvey et al. (2004) discuss, one appealing aspect of defining desistance as a decrease in offending over time is that it specifically allows for the inclusion of intermittent offending.

One of the main reasons why this definition problem is so critical is the possibility for 'overlap' in the definitions. Indeed, using the definitions of desistance as a declining rate of offending, persistence as criminal behavior that occurs over multiple phases of the life course, and intermittence as increased waiting times between offenses, it is entirely possible that one individual offender, using the exact same measured data, could be simultaneously defined as an offender in the process of desistance, as a persistent offender, and as an intermittent offender. One researcher's persister can be another researcher's desister and yet another researcher's intermittent offender. Importantly, however, this overlap of definitions emphasizes a critical point raised by both Laub and Sampson (2003: 287) and Maruna (2001: 75): while the dichotomy of 'persister/desister' is often used to imply that these labels apply to distinct kinds of people, these terms should only be used at most as 'heuristic devices' because they may in fact be the same 'individuals in different stages of the process of change.'

Operationalization Complications

Besides the problem of conceptually defining desistance and persistence, there are also a number of important issues related to the operationalization and empirical measurement of them. First is the issue of data, whether to measure criminal behavior using official (arrest) data or self-report data. Compared to official data, self-report data may find more persistence or 'delayed desistance' because some crimes are less likely reported to the authorities and therefore less likely to be captured in official data, but still likely to be captured in self-report data (Farrington, 2007). Official data, on the other hand, carries the bias of decision makers in the criminal justice system. Ideally, studies would use both measures of criminal behavior, but that is rare.

Other questions emerge about the nature of the criminal offenses committed over time and how they should be included in measuring desistance. In addition to offending frequency and seriousness, is versatility/specialization in offending relevant? Should measures of desistance allow for 'offense displacement' (Massoglia, 2006), which describes how offenders may vary or shift the types of offenses they commit across age? If offenders shift from violent crimes to drug crimes, is this evidence of desistance? And perhaps most importantly, can one simultaneously combine many of these issues (frequency, seriousness, versatility, displacement) into a more holistic view of desistance?

A third measurement issue concerns exactly who should be included in studies of desistance. Laub and Sampson (2001) argue for a minimum threshold of prior offending to be met in order to be included in a study of desistance and specifically argue that occasional, low-rate offenders should be excluded from studies of desistance. Not all criminologists agree. Finally, given that individuals highly involved in crime are more likely to experience both death and incarceration, it has been argued that adjustments for these outcomes should be made in order to not explain a lack of new offenses as desistance when in fact the offender was either incarcerated or deceased.

These many issues concerning definitions and measurement may seem overly complicating; however, definitions and operationalizations matter. For example, both Bushway et al. (2003) and Brame et al. (2003) find that different definitions applied to the same data can yield different conclusions about how many offenders 'desist' and exactly who is identified as a 'desister.' Different concepts, different data, and different empirical measures all express particular ideas and ultimately may mean different answers. Examining the definitions, samples, and measures used in any study is critical to contextualizing the results because both findings and generalizability can be affected by them (Piquero et al., 2012). Further research is needed to explore conceptual and operational consequences on individual samples. Yet, until a greater understanding and consensus emerges, readers of the literature would be well served to scrutinize the definitions, samples, and measures in any given study and to not overgeneralize results.

Methods of Analysis

Since most studies of desistance and persistence employ the static definition, they generally analyze the data using standard methods of analysis (such as logistic regression) that allow for a between-group analysis of the characteristics significantly associated with distinguishing desisters from persisters. In contrast, there is more uncertainty about the appropriate method of analysis when one adopts the process definition. Studies using the process definition typically choose between one of two statistical approaches to modeling changing offending patterns over time: the group-based trajectory model (GBTM) and the growth curve model (GCM). Both methods examine heterogeneity in offending trajectories over time while also allowing for examining the effects of time-varying covariates on offending patterns, but they do so using different assumptions. Of particular note is the manner in which each accounts for heterogeneity in the distribution of offending trajectories. This brief description of these two approaches draws on several excellent, relatively nontechnical overviews found in Brame et al. (2012), Nagin (2010), Osgood (2010), and Sullivan and Piquero (2011).

The GBTM accounts for the heterogeneity by 'grouping' individuals together into a finite number of groups, with each group possessing a distinct offending trajectory in terms of its level of offending and/or its developmental pattern over time. This allows each of the estimated groups to have a distinct trajectory shape over time.

The number of groups in any data set is empirically determined based on a number of considerations, including the fit of the model. After the necessary number of groups has been decided and the model estimated, each individual is assigned to the group that most closely resembles their observed offending pattern. The model estimates also allow for a graphical depiction of each group's trajectory, as well as the expected percentage of the population belonging to each group. Analyses can also examine whether specific predictors are associated with membership in the different groups.

Further, since individuals with similar offending patterns over time are assigned to the same group, a variety of methods are then available to control for possible selection effects. For example, analyses have been completed using just the members of particular groups because all members within each group have a similar criminal propensity. Similarly, propensity score methods (which match people on the basis of observed variables) have also been used to match members within a trajectory group (Nagin, 2010). Finally, time-varying variables can also be included in the estimation of the GBTM itself to examine whether the time-varying variable impacts the offending rate itself, and separate effects can also be estimated for each group, which also allows one to control for criminal propensity differences.

The main controversy of this method, however, concerns the estimated 'groups.' The notable concern has been over the possible reification of the groups as real, a priori distinct groups, which has significant implications for both theory and policy. If the estimated groups arise out of the 'approximation' process of the GBTM whereby the continuous distribution of trajectories is approximated through the use of discrete groups (see Nagin, 2010), then the groups are not real, distinct groups, but they may still be treated as such. This concern has been raised in light of the 'chronic offender' groups that are often found, especially since the chronic offender group has long been a seductive target for selective incapacitation policies.

The GCM approach, on the other hand, accounts for the heterogeneity in offending by imposing mathematical assumptions about distributions of the individual offending trajectories, or more specifically, the distribution of the intercept and growth (slope) parameters. As part of this assumption of the distribution of the trajectories, the method assumes there is a common, average trajectory descriptive of the population as a whole, as well as usually specifying that the individual trajectories are distributed about the overall average trajectory according to the multivariate normal distribution. Thus, in comparison to the GBTM, the GCM approach assumes that all individual trajectories are distributed about a common, average trajectory. The power of these models principally lies in their ability to examine what individual characteristics can account for the variability of the trajectories about the average trajectory, both between persons (using individual characteristics that do not vary across time) and within individuals (using individual characteristics that vary across time). It is the within-individual part of these models that has been creatively exploited to examine the effect of life events on changing offending patterns as it allows for examining the effects of events while holding constant all of the time-stable characteristics of the person, both measured (such as gender) and unmeasured (such as a time stable criminal propensity) (Osgood, 2010); in effect, because each person is essentially acting as their own control in this part of the model, this allows one to account for the selection effect argument that argues time-stable traits explain the occurrence of both life events and criminal offending patterns. However, it cannot account for time-varying selection effect processes or unmeasured time-varying characteristics (Osgood, 2010; Siennick and Osgood, 2008).

It is perhaps best to think of these two approaches as mutually beneficial (Nagin, 2010). The GBTM's key strength lies in its ability to uncover distinct trajectory shapes that are allowed to vary in terms of their developmental shape and to examine how characteristics are associated with and predict group membership. The key strength of the GCM is its ability to examine how the time-varying characteristics of the individuals can be used to explain within-person changes in offending (Nagin, 2010). Since there will be continuing interest in both discovering and examining trajectories displaying patterns of persistence and/or desistance (the strength of GBTM), as well as examining the characteristics associated with within-individual changes in offending patterns (the strength of GCM), both methods will continue to add insight to our understanding of the complex processes of desistance and persistence in unique, yet complementary, ways.

Key Empirical Findings: Trajectories of Criminal Offending, Salient Life Events, and the Prediction Problem

Key conclusions and empirical findings of three research areas related to the empirical study of persistence and desistance are briefly and broadly discussed here. Key studies in each area are cited for further background information and details on these topics.

Criminal Offending Trajectories

Rooted in the age–crime curve debate and largely utilizing the group-based trajectory methods described above, there were already over 80 studies on the nature of criminal offending trajectories by 2008 (Piquero, 2008), and more have appeared since then. Key reviews can be found in Piquero (2008), Van Dulmen et al. (2009), and Jennings and Reingle (2012).

Several notable, generally consistent conclusions emerge from the studies examining criminal trajectories. First, there is a considerable amount of heterogeneity with respect to the nature of criminal trajectories across age; this relevant heterogeneity in the offending patterns can be specified in terms of when they begin, the rate at which they increase, the age at which they peak, the age at which they begin to decrease, the rate at which they decrease, and when/if they reach a point where the offending rate is approximately zero. This heterogeneity sometimes results in trajectories that, although very similar at one point in time, ultimately become quite dissimilar at a later point in time, including trajectories crossing over each other (e.g., see Bushway et al., 2003: Figure 1). Ultimately this broad conclusion sheds doubt on the arguments that the relationship between age and crime is invariant and that the aggregate age–crime curve applies at the individual level; further, it supports arguments of Sampson and Laub that differences in adult offending patterns cannot be explained purely on the basis of earlier preexisting differences in offending.

Second, the vast majority of offending trajectories rise to their peak rates and then decline with age, although the specific peak age and rate of this decline varies across trajectories. Regardless, by early to midadulthood, most estimated

trajectories are either in the 'process of desistance' or appear to have terminated their offending, but the process itself is clearly variable across the groups. Third, at least one or two of the estimated groups typically engage in crime rather persistently across age, and are often labeled with the modifiers 'persister,' 'chronic,' or 'high-rate' (see Table 1 of Jennings and Reingle, 2012 and Table 14.1 of Van Dulmen et al., 2009). Thus, although it has been argued that it is inherently problematic to find the 'life-course-persistent' offender type because of censoring issues such as short follow-ups, death, and incarceration and because the GBTM/GCM methods have been argued to be biased against finding such offenders (Bushway, 2013), it does appear to be that there is a small group of offenders who, at least more or less, engage in crime rather persistently across age. The second and third conclusions noted here, however, are not inconsistent with each other, especially when one contextualizes the problem of defining terms like desistance, persistence, and chronic. For example, a persistent offender group could be criminally active in a number of different decades in life, but they could also be much more involved in crime during their earlier periods of life.

Marriage, Employment, and Other Adult Life Events

Positive life events experienced during adulthood are the primary focus for explaining the heterogeneity in offending trajectories across age because the declines in crime occur at times when adult life event transitions are common. Broad, summary conclusions from this research, especially those regarding marriage and employment, are included here. Excellent detailed reviews on these topics and their methodological challenges can be found in Laub and Sampson (2001), Siennick and Osgood (2008), Uggen and Wakefield (2008), and Horney et al. (2012). To be clear, determining the 'effect' of life events is methodologically challenging because one must effectively account for the 'selection' arguments that the same individuals who experience various life events could be the same people likely to desist for other reasons (e.g., because they have higher self-control).

Marriage is both the most studied life event and the one demonstrating the largest and most consistent association with reductions in criminal offending. This 'good marriage' effect has proven robust across studies that differ in research design (cross-sectional and longitudinal designs), time period, geographic area, operationalization of crime (self-report and official data), and sample type (general population and offender-based samples) (Horney et al., 2012; Siennick and Osgood, 2008). This robust marriage effect is further strengthened by the many studies employing the best and most powerful methods of accounting for possible selection effects, including propensity score methods, estimating the effect within distinct trajectory groups, and through analyses of within-individual change that control for all time-stable differences between people.

However, the marriage effect may in fact be qualified by gender because it appears to be more beneficial for men, but further research is needed. Future additional research also needs to expand to include romantic partners and cohabitation more generally, as well as whether the effect of marriage depends on the level of prior criminal activity (whether it is stronger for less/more serious offenders), the possible reciprocal relationship between crime and marriage, and how the level of criminality/deviance of the marital/romantic partner may alter the effect size or direction. And despite considerable efforts into determining the specific causal mechanisms underlying why marriage matters, researchers admit a variety of sources could potentially explain it (Laub and Sampson, 2003): the investment and strength of the marriage, changing patterns of associations with (deviant) peers, changing routine activities, changing places of residence and neighborhoods, increased levels of supervision and monitoring, and/or as a result of cognitive identity transformations. Future research is needed to help determine the specific intervening, mediating mechanism(s) through which the marriage effect is transmitted.

Although studied less than marriage, employment is another life event subjected to empirical examination for its association with desistance/persistence, but its effects are both less consistent and weaker than those associated with marriage; further, estimated effects are often found to 'depend on something else' (Siennick and Osgood, 2008: 180). For example, employment appears to have differing effects across age. After employing effective statistical controls for possible selection effects, there appears to be no association of employment with either increased or decreased levels of criminal activity for adolescents. For older adults (especially those beyond age 26 years), however, employment appears to significantly help facilitate desistance. Conclusions on the importance of job stability and job quality remain unclear, including how job quality could influence the overall relationship between employment and crime. Additionally, the relationship between employment and crime may vary by offense type, with employment possibly increasing property offending and decreasing violent offending. Because studies of employment tend to not be as methodologically rigorous as those examining the effect of marriage, conclusions regarding the effects of employment also tend to be more tentative, lending greater uncertainty to the general conclusions and leaving wide open the areas for much future research.

Beyond marriage and employment, however, many other theoretically interesting adult life events have yet to be studied as frequently or extensively. For example, whether/how parenthood, military service, religion and spirituality, and penal sanctioning and imprisonment are associated with patterns of desistance/persistence are all issues awaiting additional research. For example, initial findings examining the link of both military service and parenthood to changes in offending have produced variable, mixed findings. Also, it has been noted that because some individuals begin their pattern of desistance during adolescence, prior in time to the occurrence of any key adult life events, future research needs to delineate and study the role of important adolescent life events and processes as well (Laub and Boonstoppel, 2012). Therefore, as the field continues to examine these issues, the many and nuanced ways in which many life events may or may not be associated with persistence and desistance will certainly become clearer.

The Prospective Prediction Problem

The discovery of the 'chronic offender' was critical for generating interest in whether it was possible to identify these persistent 'career criminals' early in their life courses in order to incapacitate them. Subsequent research indicated that the early prediction of chronic offenders was fraught with prediction errors, especially false-positive errors (falsely predicting someone to be a chronic offender).

The issue of whether it is possible to prospectively predict long-term crime outcomes is still present today. The ascent of developmental theories that predict distinct groups of offenders along with the rise of the GBTM methodology, which generates 'groups' (whether real or approximations), has generated interest in whether membership in the groups could be prospectively predicted on the basis of early risk factors. Although groups with distinctive offending trajectories can clearly be retrospectively identified, including groups labeled as 'persisters' and 'desisters,' subsequent attempts to accurately predict group membership on the basis of early (childhood) risk factors have resulted in the conclusion that such prospective prediction is not possible with the currently and commonly available predictors (Bersani et al., 2009; Sampson and Laub, 2003). Although certain variables have been found to have significant associations with membership in certain groups, the predictive results from those associations are ineffective at accurately discriminating group membership. Thus, although we know there are people who desist earlier in the life course and people who will persist over longer periods of the life course, studies have not been able to effectively predict which offenders will be the persisters or desisters. Consider three important, interrelated points with respect to this finding. First, it is consistent with the previous discussion that there is heterogeneity in long-term offending patterns and that the relationship between age and crime is variable. Second, it indicates that even for high-rate offenders, change over the life course is possible. Finally, it highlights the importance of life experiences in late adolescence and adulthood; unless one knows prospectively what life events a person will experience, how those events are specifically related to crime, and how that person will respond to those events, it will be impossible to predict their specific offending rate in the future (see Laub et al., 2006). Further, it reinforces the previous conclusion of the need for further research to empirically uncover the specific life events, beyond marriage, that are significantly associated with changes in offending.

Key Emerging Issues

Despite the historically heavily quantitative nature of studies on desistance and persistence, a budding interest in the subjective accounts of the desistance and persistence processes has increased interest in qualitative methods, especially life history narratives. Qualitative methods allow for important in-depth analysis from the offenders' subjective perspective, allowing offenders to interpret their present in light of both their past and their anticipated future to explain how they 'make sense of their lives' (Maruna, 2001: 7). These narratives allow for examination of possible changes in cognitive processes and identity and specific causal mechanisms though which salient life events are linked to crime, including the interplay between human agency (purposeful choice) and constrained structural contexts (Laub and Sampson, 2003).

Indeed, the issues of the future self, cognitive processes, identity change, and human agency all strike at the core of the disputes between differing perspectives on the relevant subjective factors and processes that may play crucial roles in the desistance/persistence process. For example, Maruna (2001) argues that offenders who desist have been able to cognitively reconstruct themselves as 'making good' through a 'redemption script,' but they do not literally cast off their old identity. Paternoster and Bushway (2009), on the other hand, argue that a necessary, gradual, and complete identity change from offender to nonoffender results from the 'fear of the future self,' as well as that beginning the process of such identity change is required prior in time to both accessing and benefitting from adult life events. Laub and Sampson (2003), in contrast, argue that it is not necessary for offenders to undergo an identity change to benefit from life events or desist, and that, in fact, life events themselves can generate 'desistance by default.' So, although there is growing interest in the importance of subjective processes, answers as to which perspectives have more empirical validity await future qualitative (and quantitative) research.

Moving forward, three empirical issues will likely set the stage upon which many of the interesting questions discussed above concerning persistence and desistance will play out. First, studies on these topics nearly always note that the generalizability of findings remains a critical empirical limitation because most analyses are based on predominantly white, male samples from the United States or other Western industrialized nations and are specific to time and place. The degree to which any reported findings generalize across other important demographic subgroups, including Hispanics, African-Americans, and females, as well as other places and times, may either unlock new findings or fortify current findings about how and why criminal offenders persist or desist. For example, what are the consequences of recent changes in patterns of marriage, childbirth, education, and employment for the desistance processes of young adults (Horney et al., 2012)?

Second, there is a need to understand the nuanced ways in which life events and criminal behavior are related. For example, data and theory must delve into the nuances of the timing of life events and crime to clarify both causal ordering issues and possible reciprocal effects of crime on the events themselves.

Finally, one important topic that is especially likely to be the source of future research is whether and how brain development during late adolescence and early adulthood plays a role in the timing and patterning of desistance. For example, the relevance of brain development can be seen in light of the role it played in the *Roper v. Simmons* (2005) US Supreme Court case that abolished the death penalty for juveniles. Given recent advances in brain imaging technologies and recent findings that the brain undergoes significant structural and functional changes during late adolescence and early adulthood,

especially in the prefrontal cortex area related to reasoning and impulse control, this topic is ripe for empirical investigation. This likely research on brain development will join the traditional social science literature along with the more recent interest in both biosocial and genetic explanations for explaining patterns of persistence and desistance among criminal offenders.

See also: Age and Crime; Biosocial Approaches: Crime; Causal Counterfactuals in Social Science Research; Control Theories and Crime; Crime: Knowledge about and Prevalence; Delinquency, Sociology of; Growth Curve Analysis; Hierarchical Models: Random and Fixed Effects; Imprisonment: Sociological Aspects; Life Course Criminology; Mass Imprisonment and Its Consequences; Mixture Models in Statistics; Quantitative Methods and Crime; Re-entry and Reintegration After Incarceration; Recidivism; Rehabilitation of the Offender; Self-Control Theory and Crime; Statistical Analysis: Multilevel Methods; Transition to Adulthood.

Bibliography

Baker, T., Falco Metcalfe, C., Piquero, A.R., 2013. Measuring the intermittency of criminal careers. Crime and Delinquency, doi:10.1177/0011128712466382. http://cad.sagepub.com/content/early/2013/01/28/0011128712466382.abstract.

Bersani, B.E., Nieuwbeerta, P., Laub, J.H., 2009. Predicting trajectories of offending over the life course: findings from a Dutch conviction cohort. Journal of Research in Crime & Delinquency 46, 468–494.

Brame, R., Bushway, S.D., Paternoster, R., 2003. Examining the prevalence of criminal desistance. Criminology 41, 423–448.

Brame, R.B., Paternoster, R., Piquero, A.R., 2012. Thoughts on the analysis of group-based developmental trajectories in criminology. Justice Quarterly 29, 469–490.

Bushway, S.D., 2013. Life-course-persistent offenders. In: Cullen, F.T., Wilcox, P. (Eds.), The Oxford Handbook of Criminological Theory. Oxford University Press, Oxford, pp. 189–204.

Bushway, S.D., Piquero, A.R., Broidy, L.M., Cauffman, E., Mazerolle, P., 2001. An empirical framework for studying desistance as a process. Criminology 39, 491–516.

Bushway, S.D., Thornberry, T.P., Krohn, M.D., 2003. Desistance as a developmental process: a comparison of static and dynamic approaches. Journal of Quantitative Criminology 19, 129–153.

Bushway, S.D., Brame, R., Paternoster, R., 2004. Connecting desistance and recidivism: measuring changes in criminality over the lifespan. In: Maruna, S., Immarigeon, R. (Eds.), After Crime and Punishment: Pathways to Offender Reintegration. Willan Publishing, Portland, pp. 85–101.

Fagan, J., 1989. Cessation of family violence: deterrence and dissuasion. In: Ohlin, L., Tonry, M. (Eds.), Crime and Justice: An Annual Review of Research, vol. 11. University of Chicago Press, Chicago, pp. 377–425.

Farrington, D., 2007. Advancing knowledge about desistance. Journal of Contemporary Criminal Justice 23, 125–134.

Gottfredson, M.R., Hirschi, T., 1990. A General Theory of Crime. Stanford University Press, Stanford, CA.

Horney, J., Tolan, P., Weisburd, D., 2012. Contextual influences. In: Loeber, R., Farrington, D.P. (Eds.), From Juvenile Delinquency to Adult Crime: Criminal Careers, Justice Policy, and Prevention. Oxford University Press, London, pp. 86–117.

Jennings, W.G., Reingle, J.M., 2012. On the number and shape of developmental/life-course violence, aggression, and delinquency trajectories: a state of the art review. Journal of Criminal Justice 40, 472–489.

Kazemian, L., 2007. Desistance from crime: theoretical, empirical, methodological, and policy considerations. Journal of Contemporary Criminal Justice 23, 5–27.

Laub, J.H., Boonstoppel, S.L., 2012. Understanding desistance from juvenile offending. In: Feld, B.C., Bishop, D.M. (Eds.), The Oxford Handbook of Juvenile Crime and Juvenile Justice. Oxford University Press, Oxford, pp. 373–394.

Laub, J.H., Sampson, R.J., 2001. Understanding desistance from crime. In: Tonry, M. (Ed.), Crime and Justice: An Annual Review of Research, vol. 28. University of Chicago Press, Chicago, pp. 1–69.

Laub, J.H., Sampson, R.J., 2003. Shared Beginnings, Divergent Lives: Delinquent Boys to Age 70. Harvard University Press, Cambridge.

Laub, J.H., Sampson, R.J., Sweeten, G.A., 2006. Assessing Sampson and Laub's life-course theory of crime. In: Cullen, F.T., Wright, J.P., Blevins, K.R. (Eds.), Taking Stock: The Status of Criminological Theory, vol. 15. Transaction Publishers, New Brunswick, pp. 313–333.

Maruna, S., 2001. Making Good: How Ex-convicts Reform and Rebuild Their Lives. American Psychological Association, Washington, DC.

Massoglia, M., 2006. Desistance or displacement? The changing patterns of offending from adolescence to young adulthood. Journal of Quantitative Criminology 22, 215–239.

Moffitt, T.E., 1993. Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy. Psychological Review 100, 674–701.

Mulvey, E.P., Steinberg, L., Fagan, J., Cauffman, E., Piquero, A.R., Chassin, L., Knight, G.P., Brame, R., Schiubert, C.A., Hecker, T., Losoya, S.L., 2004. Theory and research on desistance from antisocial activity among serious adolescent offenders. Youth Violence and Juvenile Justice 2, 213–236.

Nagin, D.S., 2010. Group-based trajectory modeling: an overview. In: Piquero, A.R., Weisburd, D. (Eds.), Handbook of Quantitative Criminology. Springer, New York, pp. 53–67.

Osgood, D.W., 2010. Statistical models of life events and criminal behavior. In: Piquero, A.R., Weisburd, D. (Eds.), Handbook of Quantitative Criminology. Springer, New York, pp. 375–396.

Paternoster, R., Bushway, S., 2009. Desistance and the "feared self": toward an identity theory of criminal desistance. The Journal of Criminal Law & Criminology 99, 1103–1156.

Piquero, A.R., 2004. Somewhere between persistence and desistance: the intermittency of criminal careers. In: Maruna, S., Immarigeon, R. (Eds.), After Crime and Punishment: Pathways to Offender Reintegration. Willan Publishing, Portland, pp. 102–125.

Piquero, A.R., 2008. Taking stock of developmental trajectories of criminal activity over the life course. In: Liberman, A.M. (Ed.), The Long View of Crime: A Synthesis of Longitudinal Research. Springer, New York, pp. 23–78.

Piquero, A.R., 2009. Methodological issues in the study of persistence in offending. In: Savage, J. (Ed.), The Development of Persistent Criminality. Oxford University Press, New York, pp. 271–287.

Piquero, A.R., Farrington, D.P., Blumstein, A., 2003. The criminal career paradigm. In: Tonry, M. (Ed.), Crime and Justice: An Annual Review of Research, vol. 30. University of Chicago Press, Chicago, pp. 359–506.

Piquero, A.R., Hawkins, J.D., Kazemian, L., 2012. Criminal career patterns. In: Loeber, R., Farrington, D.P. (Eds.), From Juvenile Delinquency to Adult Crime: Criminal Careers, Justice Policy, and Prevention. Oxford University Press, New York, pp. 14–46.

Sampson, R.J., Laub, J.H., 1993. Crime in the Making: Pathways and Turning Points through Life. Harvard University Press, Cambridge.

Sampson, R.J., Laub, J.H., 2003. Life course desisters? Trajectories of crime among delinquent boys followed to age 70. Criminology 41, 555–592.

Siennick, S.E., Osgood, D.W., 2008. A review of research on the impact on crime of transitions to adult roles. In: Liberman, A.M. (Ed.), The Long View of Crime: A Synthesis of Longitudinal Research. Springer, New York, pp. 161–185.

Snyder, H.N., 2012. Arrest in the United States, 1990–2010. http://www.bjs.gov/index.cfm?ty=pbdetail&iid=4515 (accessed 28.10.13.).

Sullivan, C.J., Piquero, A.R., 2011. Criminal career research: a statistical and substantive comparison of growth modeling approaches. In: MacDonald, J. (Ed.), Measuring Crime & Criminality: Advances in Criminological Theory. Transaction, New Brunswick, pp. 267–297.

Thornberry, Terence P., Giordano, P.C., Uggen, C., Matsuda, M., Masten, A.S., Bulten, E., Donker, A.G., 2012. Explanations for offending. In: Loeber, R., Farrington, D.P. (Eds.), From Juvenile Delinquency to Adult Crime: Criminal Careers, Justice Policy, and Prevention. Oxford University Press, New York, pp. 47–85.

Uggen, C., Wakefield, S., 2008. What have we learned from longitudinal studies of work and crime? In: Liberman, A.M. (Ed.), The Long View of Crime: A Synthesis of Longitudinal Research. Springer, New York, pp. 191–209.

Van Dulmen, M.H.M., Goncy, E.A., Vest, A., Flannery, D.J., 2009. Group-based trajectory modeling of externalizing behavior problems from childhood through adulthood: exploring discrepancies in the empirical findings. In: Savage, J. (Ed.), The Development of Persistent Criminality. Oxford University Press, New York, pp. 288–314.

Crime, Geography of

David Herbert, Swansea University, Swansea, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The geography of crime has been recognized for a considerable time but has only in relatively recent years moved into a more central position as a significant part of criminology. There is a strong academic strand that has examined concepts such as crime areas and vulnerable areas and has developed theories to account for their existence in cities. These studies have drawn on geographical concepts such as the importance of place and the dynamics of spatial analysis but have also related to established social science benchmarks such as social disorganization and deviant subcultures. Additionally, a strong applied approach has developed that centers around the concept of geographical profiling and the application of techniques and algorithms of geographical information systems (GIS) to crime data.

Crime has its geographies and is a form of behavior that lends itself to the principles and methods of geographical analysis. Interestingly, the initial awareness of geographical crime patterns had a practical element; it was developed as an analysis with the potential for applications to problem solving. This applied aspect has become more and more predominant in recent times with the advent of procedures such as geographical crime profiling (the application of geographical information systems (GIS) techniques to crime data to allow the probable location of offenders to be identified. It can also be used to predict where offenses are most likely to occur) and the continued evolution of the family of techniques known as GIS (method for the collection and analysis of georeferenced data. It is now widely used by police forces in their use of crime statistics). Many government agencies and police forces now use crime mapping in its various forms as a standard tool in the task of criminal investigations. There are also what might be termed more 'academic' studies with no necessary practical outcomes but even they have the potential to contribute to the policy debate of understanding crime and criminals.

The recognition of crime patterns at a variety of scales can be traced to studies and reports completed in the early part of the nineteenth century. The geography of crime as a modern subdiscipline has moved from these identifications of patterns to their detailed representation, interpretation, and understanding. There are studies of the places where offenders live and the ways in which places that may be described as crime areas (those parts of the city that persistently suffer the most crime and tend to contain disproportionate numbers of offenders) emerge and persist. There are studies of the places that experience most crime and of the conditions that underpin their vulnerability. Different types of crime have different geographies and even the most simple forms of mapping will reveal these differences. Environmental criminology focuses on crime and place and has extended beyond an interest in the designed or built environment into the role of a more holistic concept of place. People experience crime and they also experience fear of crime and possible victimization, both of these affect their images of the environments in which they live. The neighborhood or residential area is one highly relevant environment and others include the places where they work or enjoy leisure and the 'pathways' they follow to reach such places. If some of the conditions that contribute to the incidence of crime in space, both real and perceived, can be identified, it is a short step to ameliorative policy formation. Criminology is a multidisciplinary field but it does now encompass a geographical perspective.

Crime is a form of deviance that involves the infringement of rules or laws that a specific society has created; both the act of a crime and the behavior of the criminal are therefore socially defined and those definitions will vary from one society to another and also over time within one society. A change in the law relating to drug abuse or traffic offenses, for example, will have significant impact on the geography of crime. The geographic study of crime draws a distinction between the geography of offenses, where crimes are committed, and the geography of offenders, where criminals live. As suggested, there are many different geographies of crime. The places where shoplifting, robbery, and residential burglary occur are all different: the teenage mugger is likely to come from a different kind of neighborhood than the corporate lawyer engaged in tax fraud. This diversity has stimulated study of the geography of crime at a variety of scales. Environmental criminology, with its focus on the places at which crimes occur, offers one organizing framework; the problematic concept of the 'crime area' offers another.

Origins

The origins of the geography of crime are found in early nineteenth century interest in cartographic criminology. A French statistician, Guerry (1833) who was employed in public service, produced maps showing the regional concentrations of crime. His observation that crimes of violence were disproportionately represented in the south of France led him to link criminal behavior and climate. In England, the social reformers such as Mayhew (1861–62) noted the strong clusters of crimes and criminals in specific areas of cities. Mayhew described the 'rookeries' of London where children were "born and bred to the business of crime." The Chicago studies of Shaw and McKay (1942) established a social science-based perspective on crime or delinquency areas, urban spatial patterns, and gradients and a set of 'theories,' principally Social Disorganization, that underpinned the ecology of crime in the city. Their studies gave

a strong impetus to an academic and theoretical analysis that had practical implications but also conceptual links to the evolving theories of the contemporary Chicago School of Social Ecology. The basic premises were that crime had a spatial order; it was not distributed randomly but showed strong and persistent trends to cluster in poorer, disadvantaged environments. Clifford Shaw moved from this analysis to found the Chicago Area Project aimed at helping youths in disadvantaged neighborhoods and this work has continued to the present day. Subsequent studies have shown a remarkable persistence of these basic qualities of a geography of crime and, further, have also demonstrated that spatial order, clustering, and a link with 'place' typify related aspects of the criminal justice system such as sentencing patterns. Policing policies and practices are other areas where these characteristics are evident and worthy of detailed study.

Information Bases

The geography of crime shares with criminology in general the problematic nature of the available data. Official crime statistics are suspect because the police record a relatively small minority of offenses that actually occur and apprehend an even smaller number of criminals. Approximately 80% of offenses never enter police records. This 'dark area' varies by type of offense. It is very high for offenses such as shoplifting or vandalism and low for car theft or homicide and arises from the unwillingness of the public, as victims or observers, to report crime to the police. Reasons for underreporting range from embarrassment to fear of reprisal to apathy. Police themselves actually observe few criminal events and therefore rely on the public, broadly defined, to notify them when offenses occur. Even the remaining 20% is further diminished by offenses not proceeded with. This underrecording of crime may mean that official data contains biases and deficiencies. Does it overrecord offenses by disadvantaged youths? Does it underrecord white-collar crime? Official crime statistics may reveal more about those who define criminality and administer the criminal justice system than about the actual incidence and seriousness of crime.

The official 'antidote' to these concerns about criminal statistics has been the introduction of national crime surveys. Dating in the main from the 1980s, these build upon an older ad hoc tradition of self-report studies and take the form of sample surveys of victimization experience and related attitudes toward crime, justice, law, and order. National crime surveys have confirmed underrecording and have qualified statements on rising crime rates, which may reflect differential recording practice rather than actual changes in crime. They have also augmented our knowledge of topics such as fear of crime. Whereas national crime surveys have not remedied the deficiencies of official criminal statistics, they have allowed for their interpretation in more refined and constructive ways. For geographers, official statistics and national crime surveys are useful for studies of national, regional, and local patterns of criminal activity and the quality of these databases are now well understood. There is continuing interest in identifying new historical data sources. The Bodleian Library at Oxford has a project to digitize records on crimes, murders, executions, and transportations in nineteenth-century England from the John Johnson Collection and provides links both to the Old Bailey proceedings web site and to sources at Harvard and Austin, Texas, in the United States.

Theoretical Bases

The geography of crime draws upon the generic theories of criminology and social science and its own disciplinary concepts. Criminology has nurtured many theories and a significant lack of consensus. Most embrace a 'multifactor' explanation, which, despite its diffuseness, is probably correct. There is no single causal explanation for criminality but a 'package' of interrelated conditions of varying applicability. Structural theories see criminal behavior as the inevitable outcome of societies that create inequalities. Yet these macro-theories fail to account for the fact that some individuals who suffer from inequality will react criminally, whereas others will not. Geographers have favored those criminological theories that lend themselves to analyses that allow a focus on place. Social disorganization is one such theory. Originating in the Chicago studies of the 1920s and 1930s, this theory linked crime with areas that lacked social order, established values, leaders, and social networks. Comparisons were made with residential areas that did possess these qualities, were 'socially organized,' and experienced low rates of crime. Subcultural theory, proposing that crime may be prevalent because it is part of the normal behavior pattern within specific groups, can be found in embryo form in Mayhew's study of London's rookeries. Although now argued most consistently in relation to social groups, subcultural theory can be articulated in terms of areas or neighborhoods. Social control theory shares common strands with social disorganization and subculture, emphasizing the positive impacts of social control mechanisms in reducing the incidence of criminal behavior. The mechanism may be good parental supervision, an effective school or teacher, or the effects of voluntary organizations or community groups. It is again a theory that can be adapted to a focus on place and has a central concern with the offenders and where they live.

Environmental criminology offers the closest approximation to a geographical 'theory' of crime. It emphasizes the criminal act and the location at which it occurs, the 'where' rather than the 'why' of crime (Bottoms, 1994). Place is central to an approach that acknowledges that each offense leaves a new point on the map and is a clue to the activity space of the offender. The approach draws in considerations of the routine spatial activities of offenders and victims, the concept of journey to crime, and related ideas of territoriality and surveillance. At the present time environmental criminology is more a methodology than a theory. As such it draws heavily on basic geographical concepts such as space and place, and lends itself to detailed spatial analysis through GIS. Conceptual issues include the problem of attaching any independent significance to space and understanding the full range of meanings attached to place, though these are issues that are generic to the study of human geography and are by no means specific to the geography of crime.

Approaches to the Geography of Crime

Studies that merit the description of a geography of crime have changed along with broader paradigm shifts within the social sciences. Cartographic criminology, basically concerned with the mapping of criminal statistics and the representation of spatial patterns, has evolved to reflect wider use of statistical methods including the sophisticated technology of GIS and will be discussed more fully in the next section.

The study of crime areas was an early focus for a geography of crime, arising from the persistent evidence that the mapping of both offenders and offenses revealed concentrations or clusters in space and indeed in disadvantaged neighborhoods. This evidence was apparent at a variety of geographical scales from regions to localities within cities but the most significant work arose from Shaw and McKay's studies of delinquency and crime areas in Chicago. There is persistent evidence for 'crime areas' and a simple exercise of mapping crime statistics makes this clear. Although crime areas can be identified in a reasonably clear way, the task of explaining their existence is less straightforward. Statistical analyses reveal strong correlations between the incidence of criminal residence and indicators of a poor residential environment. Measures such as high unemployment rate, low income, substandard housing, and low educational attainment recur and, less frequently, indicators such as lone-parent families and broken homes. Correlates of this kind have been interpreted as indicative of some more general condition such as social disorganization or weak social control, and particular significance is often attached to single measures such as high unemployment. For the most part, however, the results of this type of analysis remain descriptive; they have value in identifying groups and areas from which offenders are likely to come but are in effect markers of some more general social condition as is the crime rate itself.

Statistical analyses have also been used as a first stage in the study of crime or delinquency areas. After identifying clusters of offenders and the conditions under which they live, the research task then is to probe more deeply with different research methodologies into the nature of the 'crime area.' One approach has identified particular neighborhood characteristics that might predispose residents to criminology. Herbert (1976) compared 'crime' and 'non-crime' (or at least much less crime) areas in a city, controlling for sociodemographic profiles, and found a consistent difference in terms of sets of values and forms of behavior between the two types of area. In the 'crime' areas, parents adopted different approaches to control, had a clear tendency to accept some forms of misbehavior, and the children were lower achievers educationally. Other studies have examined this 'neighborhood effect' and the extent to which an individual's residential area has influence on criminal activity. The Edinburgh Study of Youth Transitions and Crime was a longitudinal study of the pathways into and out of offending for a cohort of 4328 children who started secondary school in the City of Edinburgh in 1998. One finding was that the characteristics of neighborhoods did play a role, albeit weaker than individual traits such as gender and personality, in influencing their delinquent and drug-taking behavior (McVie and Norris, 2006a). Delinquency and hard drug use were influenced by negative conditions such as high levels of deprivation. Another finding was that neighborhood effects were strongest in the earlier years of the cohort at around the age of 12 and became less significant as they became older (McVie and Norris, 2006b). Factors such as low social control and less social organization seemed relevant. Other neighborhood studies have focused on specific issues such as lack of parental supervision. More recently, what might be termed 'crime area' studies have taken a known problem area and sought to unravel both its origins and present form. In the United Kingdom, the so-called 'problem housing estates' lend themselves to this kind of analysis. Such estates often originated as local housing managers 'dumped' known problem families or relocated a slum clearance population to a specific estate. Early episodes of misbehavior reinforced labeling and reputations that have persisted. The Meadowell Study (Barke and Turnbull, 1992), based on an estate of around 2000 homes, demonstrated a lifestyle that was permeated by welfare dependency, drug-trading, petty crime, and vandalism. Often self-selection of tenants who opt for this kind of lifestyle builds the area profile. In late 1991, there were major riots and these events acted as a catalyst for change. Issues such as quality of housing, jobs, and local facilities were addressed with major changes to the physical environment. By mid-2000, significant improvements were being reported and Meadowell was being described as a well-cared for and thriving community (BBC, 2012). Crime areas change over time and as this example shows, they can be recovered by determined social intervention. There is also evidence that they are cyclical in the sense they are worse at specific points of local demographic change. The concept of a 'community career' suggests that some localities move through a form of life cycle with highs and lows of criminal activity (Bottoms et al., 1992). The impact of deprivation is far from removed from the concept of crime areas. Sections of urban populations are still being disenfranchised from the rest of society and the notion of a section of society that can be described as the 'truly disadvantaged' and the wider idea of an urban underclass exemplify this.

Understanding why criminal events occur where they do require different approaches. Criminals respond to opportunities or targets in the environment and the task is to identify these 'cues.' A large majority of such crimes are opportunist and involve minimal planning; a minority is 'professional,' involving experienced criminals who may spend significant amounts of time preparing for a single criminal event. It is the former that lends itself to analyses based on opportunism and environmental cues. What makes some areas or residential units more vulnerable to crime? First, areas that offer more obvious or lucrative targets for a crime such as burglary are often not likely to experience most offenses. In part this is a function of security, wealthier neighborhoods are able to defend themselves with security devices and forms of control such as block or neighborhood watch and gated communities. Research also shows that burglars tend not to travel far to commit an offense (Herbert and Hyde, 1985). The offender–residence hypothesis suggests that vulnerable areas are either the areas or are close to areas where offenders live.

Neighborhoods have 'cores' and 'borders' and the border-zone hypothesis suggests that the borders are more vulnerable. The local social control hypothesis suggests that ordered, well-organized neighborhoods with a strong sense of community identity experience less crime; the area variability hypothesis suggests that 'mixed' areas experience more. Hypotheses of these kinds, focusing on the places at which criminal events occur, blend into the general approach of environmental criminology. So too do studies of incivilities or signs of disorder in the urban environment (Herbert, 1993) and the idea that these may indicate a deteriorating situation in neighborhoods. The well-known study of 'broken windows' (Wilson and Kelling, 1982) instigated this line of thinking. Allied to the idea of vulnerable areas that suffer disproportionately from crime are studies of the movements of offenders, sometimes referred to as journeys to crime. There is ample evidence that the majority of offenders do not travel great distances to commit crimes and this is especially true of youths and opportunistic offending. Known areas are the favorite targets and these often correspond with the places and routes offenders travel in their normal everyday lives. For the small minority of professional criminals, the ranges can be much greater and involve planning and observation.

As criminals react to cues in the environment so do potential victims. Fear of crime arises from the perception that parts of the city, types of individuals or groups, and times of the day are dangerous and there is evidence for a steady increase in public anxiety about crime and safety (Smith, 2010). People will modify their behavior to cope with fear and to achieve feelings of community safety. Early studies identified residents' perceived stress areas and showed that these were avoided in everyday routines. Fear of crime is known to be high among old people and women with young children. Victimization experience and fear are linked to deprived areas but also to the routine forms of some urban lifestyles and specific demographic groups. Groups such as the elderly in fact experience less crime than youths but tend to adopt more comprehensive avoidance strategies to reduce risk. More broadly, studies of perceived 'dangerous places' suggest that feelings of fear are dynamic and contextual and are constantly being reassessed depending on factors such as time of day or night, past experiences, local knowledge, and presence of others (Lupton, 1999). Fear may often be stimulated by perceived disorders or negative cues (broken windows idea) in a neighborhood though Brunton-Smith (2011) argued that it was difficult to say which came first. Was it the perceived disorder that prompted fear or did fear lead to the perception of disorders?

Police have responded to evidence on fear and victimization studies with initiatives such as 'do not speak to strangers' to protect children, but part of the problem may lie in the design of city environments and conventional attitudes toward time and place. The step from environmental criminology to 'designing out crime' and crime prevention has proved a short one. In his thesis on 'defensible space,' Newman (1972) argued that places became more vulnerable if they had bad design features, lacked a sense of territoriality, had poor surveillance, and were stigmatized by their external images. If design could incorporate more positive qualities, it could create defensible space and become less vulnerable to crime. The most comprehensive test of defensible space principles was provided by Coleman's (1985) study of housing projects in inner London. She developed a 'disadvantagement' score that was made up of measures on a range of indices from presence or absence of overhead walkways to alignments of doors and windows. Coleman demonstrated close correlations of disadvantagement with crime and argued that less disadvantagement meant less crime. Although criticized for neglect of social dimensions, this work had a high impact on policy and on perceived community safety and quality of life. More restricted research designs, such as those to assess the impact of improved street lighting or CCTV camera surveillance, have produced similar positive results. The caveat with such measures is that instead of reducing crime they can simply displace it to other less well-protected areas. The ultimate in designing out crime is evident in rich suburban enclaves where high technology surveillance devices, walls, gates, and guards are used to protect residents. Davis (1990) wrote of fortress Los Angeles, divided between the protected enclaves of the very rich and the territories of the criminalized poor. Geographical research on social conditions also translates easily into crime prevention policies. Block or neighborhood watch is a direct transfer of ideas on community, surveillance, and social control. By mid-2000, 27% of British households (about 6 million) and 40% of the American population were participating in neighborhood watch schemes (Bennett et al., 2008). This analysis showed that there was sufficient evidence to state that neighborhood watch schemes are effective in reducing crime but that the processes that allow that reduction are not well understood.

The geography of crime has extended to consideration of policing, sentencing, and the criminal justice system. Early studies showed that sentencing was harsher in certain states and localities: the place where one committed a crime affected the sentence received. At international scale there are clear differences; Islamic law, for example, punishes an offense differently from European law. Within the United States, Florida tends to have more prisoners on death row and more executions (and Jury composition may affect verdicts) and in the United Kingdom, magistrates drawn from one class of society judge offenders from another. Policing as an activity has strong spatial connotations. Police carry images of the city, derived from their experience and passed-on wisdom; these influence the ways in which they police the city. As police patrol areas, they need a detailed understanding of those places; the distributions in space of offenders and offenses are key considerations for them. The police often focus their resources on specific areas and/or crimes for limited periods both in terms of community policing initiatives and crime prevention or target hardening. The policy known as 'zero tolerance' has been used to apparently good effect. Geographers have studied policing initiatives such as police consultative committees and the ways in which they are constructed. Studies of the so-called 'victimless crimes' of prostitution (Shumsky and Springer, 1981) suggest that policing practice and the variable workings of the criminal justice system shape their geographies. Vice areas emerged and changed within the city in response to varying policing strategies. Beyond policing

and the courts is the penal system, and as prison populations rise, their distribution and access are significant.

Applied Aspects of the Geography of Crime

Crime mapping has moved to higher levels of diagnostic investigation with the development of GIS and its application to crime data. GIS is a technology that is now available in many software packages and these can be customized to serve specific needs. There are generic packages such as Arc Info and Map Info that are widely used in teaching and research. GIS is designed to capture, store, and manipulate all kinds of geographical data and digitally creates integrated data systems. It allows us to view, understand, interpret, and visualize any georeferenced data, and this ability to display and identify relationships is a key quality. Police forces now recognize the value of GIS as a part of crime data analysis and the geocoding of criminal events for computer storage has become standard practice. There are clear advantages in the rapid portrayal of crime patterns and the related ability both to measure the features of the spatial pattern and to explore correlation with other environmental indicators. One police analysis of 79 crimes involving a single serial rapist was reported as involving 790 000 calculations using probabilities of journey to crime to link all points on a grid to the known crime sites (Grescoe, 1996).

Perhaps the clearest expression of the applied qualities of geography to the study of crime has been the emergence of geographic crime profiling. This approach uses quantitative models and algorithms of varying levels of sophistication to aid the process of criminal investigation. Data that show where a series of offenses have occurred form the input to a procedure that allows investigators to identify the likely locations or 'hot-spots' where the perpetrators of the offenses may be found. Awareness of geographic crime profiling has entered popular crime fiction. Val McDermid's novels revolve around a psychological profiler named Dr Tony Hill. In one of his investigations (McDermid, 2009) he calls in a Geographic Profiler, who uses an algorithm that he cannot understand, to help pin down the locale of the offender. Crime mapping not only identifies where the actual crime took place but also looks at where the perpetrators and the victims spend their normal lives. Therefore, where they live is of particular relevance and to a lesser, though still significant extent, the places where they work and spend leisure time. Most criminals operate within their 'comfort zone,' probably not far from where they live and in environments with which they are familiar. Simple geographic approaches such as centrographic techniques that allow comfort zones to be demarcated were early tools in this type of study. Neil Trainor, a police profiler at Bramshill College, spoke of the relevance of the principle of least effort to criminals; they work close to home, in places with the easiest access and least risk. Rossmo (1999) in his book titled *Geographic Profiling*, defines it as "an investigative methodology that uses the locations of a connected series of crimes to determine the most probable area of offender residence." As a police officer, he is clear that geographic profiling has an important supportive role when placed alongside the range of policing strategies. Among many other approaches, nongeographic crime profiling is a separate approach that continues to prove its value to many forms of criminal investigation. Yet, Porter (2010) notes that in 1996 the United States National Institute of Justice allocated $15 million to establish a crime mapping research center to promote crime mapping as an analytical tool and this center was renamed Mapping and Analysis for Public safety in 2002.

Underpinning the growth of geographic crime profiling has been the broader development and application of the methodology of GIS and its relevance to analysis of crime data. Haining and Ceccato (2010) have conducted many studies based on large georeferenced crime data sets from different countries. The methodology of all this work is quantitative using GIS and spatial modeling and includes studies of car theft and burglary in Stockholm and vandalism in Malmo. Some of this research is not driven by applied dimensions; studies of small area offense rates in the 'transition states' of Estonia, Latvia, and Lithuania, for example, were designed to test Durkheim's hypothesis that these political and social changes led to alienation and greater chance of criminality.

Future Directions

The geography of crime started from a few simple observations of spatial distributions and extended to a variety of applications of basic disciplinary concepts of space, place, and environment. Current thrusts involve developing pattern analysis through sophisticated information technologies, focusing on the meanings and relevance of place, and examining the interfaces between crime, criminals, and the criminal justice system. Clearly, the basic geographical approach of mapping crime in its various forms and expressions has become a prime focus of attention. Whereas there remain academic questions, such as the significance of place and neighborhood, the epidemiologies of criminal activity, fear and the experience of victims, and the applications of the criminal justice system including policing, the value of crime mapping as an investigative tool is a central concern.

See also: Crime Trends and Debates; Fear of Crime; Policing.

Bibliography

Barke, M., Turnbull, G., 1992. Meadowell: The Biography of an Estate with Problems. Avebury, Aldershot, UK.

Bennett, T.H., Holloway, K.R., Farrington, D.P., 2008. The effectiveness of neighbourhood watch in reducing crime. Campbell Systematic Reviews.

Bottoms, A.E., 1994. Environmental criminology. In: Maguire, M., Morgan, R., Reiner, R. (Eds.), The Oxford Handbook on Criminology. Clarendon Press, Oxford, UK, pp. 585–656.

Bottoms, A.E., Claytor, A., Wiles, P., 1992. Housing markets and residential community crime careers: a case study from Sheffield. In: Evans, D.J., Fyfe, N.R., Herbert, D.T. (Eds.), Crime, Policing and Place: Essays in Environmental Criminology. Routledge, London, pp. 118–144.

British Broadcasting Corporation (BBC), 2012. Mapping Crime in the Meadowell Estate. CLIP 1129.
Brunton-Smith, I., 2011. Untangling the relationship between fear of crime and perceptions of disorder. British Journal of Criminology 51, 885–899.
Coleman, A., 1985. Utopia on Trial. Shipman, London.
Davis, M., 1990. City of Quartz: Excavating the Future of Los Angeles. Verso, London.
Grescoe, T., September/October 1996. Murder he mapped. Canadian Geographic, 49–52.
Guerry, A.M., 1833. Essai sur la Statistique Morale de la France avec Cartes. Crochard, Paris.
Haining, R.P., Ceccato, V., 2010. Applying geo-statistical analysis to crime data: car-related thefts in the Baltic States. Geographical Analysis 42, 53–77.
Herbert, D.T., 1976. The study of delinquency areas: a social geographical approach. Transactions of the Institute of British Geographers, NS 1, 472–492.
Herbert, D.T., 1993. Neighborhood incivilities and the study of crime in place. Area 25, 45–54.
Herbert, D.T., Hyde, S.W., 1985. Environmental criminology: testing some area hypotheses. Transactions of the Institute of British Geographers 10, 259–274.
Lupton, D., 1999. Dangerous places and the unpredictable stranger: constructions of fear of crime. Australian and New Zealand Journal of Criminology 32, 1–15.
Mayhew, H., 1861–62. London Labour and the London Poor. Griffin, London, Bohn.
McDermid, V., 2009. Fever of the Bone. Sphere, Little Brown Book Group, London.
McVie, S., Norris, P., 2006a. Neighborhood Effects on Youth Delinquency and Drug Use. Edinburgh Study of Youth Transitions and Crime, Edinburgh. Publication 10.
McVie, S., Norris, P., 2006b. The Effects of Neighborhood on Adolescent Property Offending. Edinburgh Study of Youth Transitions and Crime, Edinburgh. Publication 11.
Newman, O., 1972. Defensible Space. Macmillan, New York.
Porter, J.R., 2010. Tackling the Mobility of Crime: New Methodologies and Geographies in Modeling the Diffusion of Offending. Cambridge Scholars' Publishing, Newcastle-on-Tyne.
Rossmo, D.K., 1999. Geographic Profiling. CRC Press, New York.
Shaw, C.R., McKay, H.D., 1942. Juvenile Delinquency and Urban Areas. University of Chicago Press, Chicago.
Shumsky, N.L., Springer, L.M., 1981. San Francisco's zone of prostitution. Journal of Historical Geography 7, 71–89.
Smith, S., 2010. Crime, Space and Society. Cambridge University Press, Cambridge.
Wilson, J.Q., Kelling, G.I., 1982. Broken windows: the police and neighborhood safety. Atlantic Monthly 255, 29–38.

Crime, Immigration, and Ethnicity

Sandra M Bucerius and Katherine Hancock, University of Alberta, Edmonton, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.P. Brodeur, volume 7, pp. 4828–4831, © 2001, Elsevier Ltd.

Abstract

Despite much evidence to the contrary, the general public and policy makers in the US often assume that immigrants have higher crime rates than the native-born population and that the process of immigration itself increases the country's crime rates. In recent years, criminological researchers have produced many studies showing that the opposite might be true: immigrants are – on average – less involved in crime than the native-born population and the process of immigration has no or a negative effect on the national crime rate. In this article, we will highlight and summarize some of the existing research findings – both historical and contemporary – while at the same time pointing to areas in the field of immigration and crime that are much less understood to date. In addition, we provide a condensed overview of the existing knowledge on race and crime, and offer possible explanations for the overrepresentation of some ethnic and racial groups in the criminal justice system.

Members of the public and policy makers often assume a link between ethnic minorities, immigration, and crime, expecting rising immigration to produce greater crime and violence. In spite of these commonly held beliefs, there is a lot about the nature of the immigration and crime nexus that we do not yet fully understand. What we do know for certain is that among Western countries, *some* disadvantaged, minority groups tend to be overrepresented in arrests, convictions, and imprisonment (Tonry, 1997: p. 1). The affected groups differ in origin and characteristic across nations. In Canada and Australia, Aboriginal populations are disproportionately involved in the criminal justice system. In other places, this applies to recent immigrants (e.g., Moroccans in the Netherlands, Turks in Germany) and to visible ethnic/racial minorities (African-Americans in the United States, Afro-Caribbeans in England). Notably, the groups that face higher involvement in the justice system tend to also experience various forms of marginalization and disadvantage in their respective societies (e.g., in the educational system, in the formal economy, in access to health care, etc.). In spite of the disadvantage that visible minority and immigrant groups often face, *most* immigrant groups have lower crime rates than the native-born population.

One of the most important and least contested findings in contemporary research on immigration and crime is that broadly, immigration does *not* increase crime on the macro-level, especially violent crime, and may actually be associated with lower crime rates in neighborhoods (Butcher and Piehl, 1998; Kubrin and Trager, 2013; Reid et al., 2005; Rumbaut and Ewing, 2007). Even when "controlling for demographic and economic characteristics associated with higher crime rates, immigration either does not affect crime, or exerts a negative effect" (Reid et al., 2005: p. 775). This finding also holds for studies that put neighborhoods and cities at the center of their analysis (for a review see Ousey and Kubrin, 2009). Similarly important, research has consistently shown that first-generation immigrants are – on average – *less* likely to be involved in crime than their native-born counterparts. The observation that immigration is not associated with more crime but with less has now become the new intellectual canon (Lee and Martinez, 2009). Despite this, immigration policies in many US states have become more punitive in recent years (e.g., Arizona's SB1070, and similar laws in Winders, 2013), portraying the immigrant still as potential threat that needs to be controlled.

Overall, we need to acknowledge that the relationship between ethnicity, immigration, and crime is neither simple nor consistent across countries, and that we still have much to learn. For example, on the individual level, we need to better understand that the category of 'immigrant' is not a homogeneous one and factors like immigrants' ethnicity, age, gender, immigration history, family structure, and the quality of their reception in a host country make the study of immigration and crime complex and nuanced. On the neighborhood and city level, we still lack a good understanding as to why some immigrant neighborhoods are safer than others (Kubrin and Trager, 2013; Kubrin and Ishizawa, 2012). Lastly, our knowledge about immigrant victimization remains scarce (Stowell and DiPietro, 2013).

In this article, we will provide an overview of key issues that arise in the study of ethnicity, race, immigration, and crime, including definitional inconsistencies and challenges with cross-national comparisons. We will first offer some historical background on the immigration–crime nexus and present historical explanations that researchers and the public have used to explain linkages between immigration and crime. Subsequently, we will present empirical, contemporary findings on the immigration–crime nexus, including generational differences in offending. Further, we will turn to research on race, ethnicity and crime, present findings on ethnic or racialized discrepancy in victimization and crime rates in the United States and Canada, and will discuss potential areas of discriminatory treatment by the criminal justice system. Finally, we will identify some areas of research that we believe have been inadequately investigated to date.

Key Issues with the Study of Ethnicity, Race, Immigration, and Crime

Examining the links between immigration, race, ethnicity, and crime comes with certain challenges. For example, the concepts

of race and ethnicity are often conflated by researchers but are theoretically distinct. A common theoretical understanding of 'race' focuses on physical attributes, biology, genetics, and biologically induced behavior (Kalunta-Crumpton, 2012). 'Ethnicity' encompasses cultural traits, social interaction, nation-state, or place of origin and is related to the identity among those who share a set of cultural traits, histories, group dynamics, and experiences (Kalunta-Crumpton, 2012). Immigration usually refers to a legal status. These definitions are useful but do not rectify the difficulties in studying ethnicity, race, immigration, and crime. In practice, the distinction between race and ethnicity is often blurred or entirely ignored.

Another challenge is the difficulty of cross-national or cross-regional comparisons of empirical findings on ethnicity, race, immigration, and crime; it is imperative that contextual differences in each country or region are considered. Policies and standards for who are categorized as 'immigrant' or native born vary broadly across countries, making such comparisons ever more difficult. More importantly, there are not necessarily hard rules of definition in place. For example, in immigrant nations like the United States and Canada who are considered 'immigrant' and 'native born' are not automatically obvious: members of the second generation could be classified as 'native born' or 'second-generation immigrant' (Bucerius, 2011).

Elsewhere, statistics on immigration may be even more variable. Germany's official crime statistics only distinguish between German citizens and foreigners. Very restrictive immigration policies (like that of Germany) will likely result in many immigrants never gaining citizenship, whereas citizenship may be relatively quick to attain in countries like Canada. This makes cross-national comparisons difficult as the people categorized as 'immigrants' and 'native born' in each country are very different – making researchers literally compare 'apples' with 'oranges.' Additionally, crime statistics, definitions, and reporting practices are not consistent across borders.

Yet another difficulty for systematic cross-national comparison is that in most Western nations, justice statistics generally do not specify ethnicity and race of the offender because of the belief that collecting data on race and ethnicity would violate civil rights or incriminate certain ethnic or racial groups. More recently, many minority groups who were previously opposed to the collection of race-based statistics now advocate their collection as a means to "redress racial discrimination in the administration of justice" (Owusu-Bempah and Millar, 2010: p. 97). The United States and the United Kingdom stand alone among Western countries in their collection of race and ethnicity in crime data. Still, census and crime data collected in the United States include only broad racial categories (black, white, Asian, American/Alaskan Indian, Pacific Islander) that do not acknowledge ethnic differences, conflating the experiences of immigrants from very different cultural backgrounds and migration histories on the basis of skin color. Ultimately, we would argue that using the category of 'ethnicity' would provide a much more nuanced and detailed analysis than 'race.'

Another key issue in relation to studying immigration, ethnicity, race, and crime relates to age and gender, as they are frequently identified as the chief predictors of criminal involvement in the criminological literature. Young males have a higher propensity for criminal behavior. In his look at the relationship of immigration and crime in the United States, Waters (1999) found that during large immigration waves, immigrant groups were disproportionately young males (as cited in Bucerius, 2011: p. 389). Since recent immigrants are still disproportionately young and male, it is critical that researchers control for age and sex when attempting to discern a connection between immigration and crime (Bucerius, 2011). Similarly important is to keep in mind that most studies solely focus on the relationship of violent crime and immigration (Martinez and Lee, 2000; Lee and Martinez, 2002; Martinez, 2002; Feldmeyer, 2009; Stowell et al., 2009), ignoring other types of criminal activity.

Aside from age and sex differences, immigrants' ethnic and socioeconomic backgrounds and reasons for immigrating must be considered. Recent immigrants are a diverse group, and these factors may be salient in the study of an immigration–crime nexus. Likewise, there are significant generational differences in criminal involvement among immigrants. In Western countries, second-generation immigrants typically have higher crime rates relative to their first-generation counterparts. The second-generation crime rates relative to the native-born population differ across countries. For example, on average, second-generation immigrant groups in Canada and the US tend to have lower crime rates than the native-born population. In Europe, some second-generation immigrant group crime rates exceed those of the native born.

Lastly, studies about the relationship of immigration and crime are often insensitive to differences within the 'immigrant' category. Most do not (or cannot) distinguish between legal and illegal immigrants, partially because reliable estimates for the number of undocumented immigrants are not available. Estimates for the number of undocumented immigrants in the United States fluctuate by millions (Bucerius, 2011), and equivocal claims for their tendency to engage in crime are possible. In line with economic (Becker, 1968; Ehrlich, 1973) and strain theories (Merton, 1938; Cloward and Ohlin, 1960), immigrants with limited, legitimate economic opportunities may be expected to resort to criminal activity. Early work has emphasized the 'difficulties of adjustment' (Taft, 1933) to explain immigrant crime. Conversely, it can be argued that their vulnerable (illegal) status makes undocumented immigrants eager to avoid contact with law enforcement.

Overall there are many peculiarities and limitations that the reader needs to keep in mind when looking at research on immigration, ethnicity, race, and crime, and many research questions that remain relatively understudied. Below, we will first turn to historical and contemporary research on immigration and crime and then turn to research on race, ethnicity, and crime.

Immigration and Crime

Historical Background on Immigration and Crime

The notion of a 'criminal immigrant' appearing in the media, political discourse, and public opinion is not new. During the mass immigration of Eastern Europeans to America at the turn of the twentieth century, concern about immigrants contributing to social problems, like crime, was great (Reid et al., 2005). In response, legislation to hamper or stop certain

immigrants from entering the United States was introduced (Hagan and Palloni, 1999). Concern about immigration and crime declined after the Second World War when the US economy was strong. Recently, public attitudes have shifted once more. Indeed, from the recent General Social Survey, nearly three quarters of Americans believe immigration and higher crime rates are positively correlated (Rumbaut and Ewing, 2007; Feldmeyer, 2009: p. 718). These concerns are triggered by and exacerbated in response to immigration influxes and crises (e.g., the 9/11 terrorist attacks in the United States).

Historically, a series of federal commissions in the United States (the US Industrial Commission in 1901; the 1911 Immigration Commission; and the 1931 National Commission on Law Observance and Enforcement) tried to shed light on the immigration and crime relationship. The 1901 Industrial Commission used prison statistics as a measure of criminality and found foreign-born whites were less criminal than native-born whites (Industrial Commission, 1901). Because the native-born prison population was composed of a large number of second-generation immigrants, these findings were interpreted as supporting the 'injurious effect of immigration' (Industrial Commission 1901 in Tonry, 1997: p. 20). The 1911 Immigration Commission had similar aims but divergent interpretations from the earlier Industrial Commission. The commission concluded that immigration did not increase crime significantly or at all. The 1931 National Commission on Law Observance and Enforcement used crime and arrest statistics and found first-generation immigrants proportionately committed fewer major offenses than native-born persons of the same age and sex. In an effort to measure the criminal involvement of the second-generation, the commission interviewed criminal justice officials and concluded that the 'immigrant crime problem' was not due to immigrants 'but to their sons' (Tonry, 1997: p. 20).

Early academic work sometimes relied on biological explanations to account for immigrant crime, purporting that immigrants were biologically inferior to nonimmigrants (Martinez and Lee, 2000: p. 488). These explanations have been discredited as racist and lacking empirical and theoretical value (Gabor and Roberts, 1990; Wortley, 2003). Researchers have also attributed cultural differences to explain differences in crime rates between ethnic and racial groups (Taft, 1933; Sellin, 1938; Hawkins, 1994; Erlanger, 1995). Similar to biological explanations, cultural approaches have been criticized by academics for failing to explain nonimmigrant offending (Bucerius, 2011).

A historical examination that still holds much value in contemporary research goes back to Edwin Sutherland's (1924) criminological examination of immigration and crime, which found that immigrants in the United States engage in crime less than the native born. Sutherland found first-generation immigrants had, on average, lower crime rates than subsequent generations. As well, those who immigrated to the United States as children were more likely to be later incarcerated than those who immigrated as adults. According to Sutherland, immigrant criminal involvement increased with time spent in the United States and with later generations – a finding that still holds true in most empirical studies today. He identified that acculturation, and not immigration itself, were to be blamed for increased incarceration and arrest. In other words, the more Americanized immigrants became, the more their crime rates resembled the (on average, higher) crime rates of native-born Americans.

Contemporary Empirical Findings on Crime and Immigration

Discussions of an immigration–crime nexus frequently center on immigrants' involvement in crime or on the influence of immigration on crime. If something intrinsic to immigration produced a greater proclivity for criminal offending, countries and cities with higher immigration rates should have elevated crime rates. That is to say, universally, immigrants should offend more often than native born. Most scholarly work on immigration and crime has debunked this supposition (Reid et al., 2005; Ousey and Kubrin, 2009; Stowell et al., 2009).

Contemporary studies at the individual level repeatedly indicate that immigrants have lower offense, arrest, or incarceration rates than the native-born population (Bucerius, 2011). There is a dearth of statistics in this area looking at ethnic group differences and immigration status, save for incarceration data (McDonald, 1997). Using incarceration rates as a proxy for offending is problematic since immigrants are at greater risk of incarceration than native born (Hagan and Palloni, 1999). In spite of this disproportionate risk, in actuality, immigrants are less likely to be incarcerated than the native-born population. Butcher and Piehl (2007) analyzed census data over 1980–2000 and found the native-born population was more likely to be institutionalized (incarcerated) than immigrants. They concluded the process of immigration selects individuals who are more amenable to deterrence. Similarly, Rumbaut and Ewing (2007) found the foreign-born immigrants in the United States were less likely to be incarcerated than the native-born population. This pattern was found for all ethnic groups.

The existing studies on the effect of immigration on aggregate criminal offending provide support that immigration does not increase, and may even reduce, violent crime rates (Butcher and Piehl, 1998; Hagan and Palloni, 1999; Lee et al., 2001; Martinez, 2002; Reid et al., 2005; Ousey and Kubrin, 2009; Stowell et al., 2009). Most notably, research by Sampson across 180 communities in Chicago has shown that concentrated immigration predicted lower rates of violence, with the relationship strongest in poor neighborhoods. Sampson argues in a now-famous New York Times Op-Ed that immigration not only appears to be 'protective' against violence in poverty areas, but that immigration as a whole might have been responsible for the overall crime drop in the United States in the 1990s (Sampson, 2006).

To date, we know much less about the victimization of immigrants. Stowell and DiPietro's (2013) descriptive examination of immigrant victimization reveals that foreign-born individuals (i.e., first-generation immigrants) have a slightly higher homicide risk than their native-born counterparts; with Latino immigrants having the highest risk of homicide victimization among the ethnic groups in their analysis. However, homicide victimization rates for immigrants falls more quickly than for native-born individuals.

While research in the US context has debunked the idea that immigrants are more criminal than the native born and that

immigration leads to elevated crime rates, the picture is not as clear in the European context, with some disadvantaged minority groups being disproportionally involved in the criminal justice system. Is it important to avoid reductionist explanations of the role immigrant status and ethnicity play in criminal involvement and justice system experiences. Tonry (1997) advises that not all economically and socially disadvantaged groups are disproportionately involved in the criminal justice system. For example, despite arriving in England and Wales during the same time period and being subject to comparable forms and levels of discrimination, Afro-Caribbean and Indian migrants' involvement in the justice system are extremely dissimilar (Tonry, 1997: p. 2). Similarly, the experiences of Moroccans and Turks in the Netherlands are also divergent, in spite of similar economic and social disadvantage (Moroccans, but not Turks, have much higher crime and incarceration rates than the Dutch population) (Tonry, 1997; Junger-Tas, 1997; van Gemert, 1998). Meanwhile, although Dutch Turks and German Turks share similar migration histories and socioeconomic backgrounds, the incarceration rate for Turks in Germany is much higher than the native population (Koopmans, 2003). We believe that immigrant incorporation strategies and discrimination against certain immigrant and/or ethnic groups in different countries significantly contribute to the overrepresentation of ethnic groups in some countries but not others (see also Martinez and Stowell, 2012).

Generational Differences

The acculturation thesis described by Sutherland remains relevant and supported by many studies on generational differences (for an overview on generational differences, see Berardi and Bucerius, 2013). In spite of improved conditions relative to the first-generation, second-generation immigrants tend to fare more poorly with regard to crime rates (see Berardi and Bucerius, 2013; Picot and Hou, 2011; Portes and Fernandez-Kelly, 2008; Rumbaut and Ewing, 2007). This downward trend may be partly explained by the process of acculturation. Rumbaut and Ewing (2007) describe a 'paradox of assimilation.' The process of assimilation is often touted as a positive transformation for immigrants (e.g., gaining English language proficiency, education, job skills) that facilitates their settlement and eventual success in the economy and society of the host country. Ironically, assimilation appears to have a sinister aspect as well. Immigrants' criminal behavior and health status worsen and mimic those of the native-born population. Mirroring Sutherland's findings, second- and third-generation immigrants have greater risk of incarceration relative to the first generation. First-generation immigrants' risk of incarceration also increases along with their stay in the United States.

In a longitudinal study of neighborhood in Chicago from 1995 to 2002, first-generation immigrants of several racial and ethnic groups were nearly 50% less likely to violently offend than third-generation comparators (Sampson et al., 2005). As would be expected from the acculturation thesis, second-generation immigrants were more likely to commit violent crime than the first generation, but less likely than the third generation. Racial group differences observed between black and whites could also be explained by immigrant status, which was identified as a protective factor against violence. For example, whites in the study were more likely to be recent immigrants than blacks, and blacks demonstrated increased violent crimes. Controlling for immigrant generational status, the disparity in violence shrunk. The limited research on immigrant involvement in delinquency and substance abuse in the United States is fairly consistent with research findings on violence – first-generation immigrants tend to be less likely delinquent or use illicit drugs than the second generation (see examples, Vega et al., 1993; Bui, 2009; Gfroerer and Tan, 2003).

Exceptionally, second-generation Asian-Americans are less likely than the first generation to commit homicide, suggesting that they might also score lower on other types of crime in comparison to their parents' generation. Factors like cultural norms, parental oversight, and structural forces may account for the success achieved by Asian immigrants in academics and crime avoidance; the so-called 'Asian advantage' (Zhou and Kim, 2006; Min and Kim, 2009) has been used to explain why second-generation immigrants of Asian heritage tend to outperform other immigrant and native-born populations academically.

The trend for first-generation immigrants to be less violent and involved in crime than native born and subsequent generations is generally true for most ethnic groups in the United States. First-generation immigrants who consciously choose to immigrate to a new country may be more dedicated to succeed in the new country and prepared to defer short-term gratification (which crime might offer) with this objective in mind (Berardi and Bucerius, 2013). First-generation immigrants attribute their experiences of exclusion as exceptional and the fault of individuals (Viruell-Fuentes, 2007). Subsequent generations of immigrants seem to be more affected by experiences of exclusion than the first generation and more readily identify these incidents as evidence of systemic discrimination and social exclusion (Bucerius, 2011).

Another explanation could be that first-generation immigrants compare their living situation in the host country to their country of origin. The second generation may refer to a different comparator – the native born. Having no (or limited) exposure to their parents' country of origin, the second-generation immigrants may suffer more by the unequal distribution of opportunities. Feelings of deprivation may propel the second generation to innovate ways to achieve successes that are precluded for them (Merton, 1938). While these explanations offer some insight into the generational increase in crime rates, they cannot fully account for the relationship of generation and offending; some second-generation immigrants do not show elevated crime rates. Additional factors such as opportunity structures (Engbersen et al., 2007), family structure (Portes and Zhou, 1993; Portes and Rumbaut, 2001), and neighborhood (Shaw and McKay, 1969) may also be impactful.

Countries with immigrant incorporation policies that promote integration can reduce the likelihood of immigrants becoming involved in crime. Research from Sweden shows that the country's inclusive immigration policies help reduce second-generation offending (Bucerius, 2011). Specifically, studies have shown that second-generation immigrants

experience disadvantage and exclusion in education (Portes and Rumbaut, 2001; Farley and Alba, 2002; Bucerius, 2009), suggesting that the school system is one important area in which carefully considered incorporation policies can make a significant difference (Bucerius, 2009). Telles and Ortiz (2008), for example, demonstrated that Mexican immigrants have faced systemic exclusion for multiple generations in the United States and suggest that this group could benefit from attention. Dinovitzer et al. (2009) also attributed education as being protective against criminal involvement for certain immigrant groups. Subsequently, some researchers have recommended that policy makers aiming to reduce immigrant crime redirect their attention to bridging the educational gap between second-generation immigrants and nonimmigrants.

Ethnicity, Race, and Crime: Discrimination and Disparity

The US Context

The relationship of race and crime has long interested American scholars and policy makers. Arrest and incarceration data from the United States has consistently shown that blacks tend to be overrepresented as offenders and inmates, while Asians are particularly underrepresented. Uniform Crime Reports ('UCR') from the United States provide some insight to the race of offenders and victims but only capture crimes that are reported or where an offender was identified and arrested. In 2011, the UCR indicated that 69.2% of arrests in the United States were whites, whereas blacks represented 28.4% of the arrests (FBI, 2011). Meanwhile, census data from the same period estimates that around 13% of the general population is black (US Census Bureau, 2012), pointing to an overrepresentation of blacks in the criminal justice system. Importantly, in addition to racial categories, the UCR distinguishes between two ethnic categories: Hispanic and Non-Hispanic. Hispanics, as the largest minority group in the United States, can be captured in different racial categories (i.e., Hispanic offenders can be classified as white or black). As of 2013, the Hispanic category will be changed to 'Hispanic and Latino origin.'

Similarly to the UCR, the National Crime Victimization Survey, looking at violent crime, property crime, and personal theft, indicates not only an overrepresentation of blacks in arrests but also with respect to victimization. From the 2009 survey, for example, blacks were more likely than whites to be victims of violent crime. While there was no significant difference in the likelihood of Hispanics and non-Hispanics to experience violent crime (including rape, sexual assault, and assault), Hispanics were more likely to be victims of robbery. Native-American women had higher rates of victimization for rape and sexual assault than white or black women (Kalunta-Crumpton and Eijogu, 2012).

Canada

Canada's international reputation as a diverse and tolerant nation has not protected it from moral panics prompted by minority criminality nor is its criminal justice system free of discrimination toward certain racial minorities (Wortley and Owusu-Bempah, 2012). Since official race–crime statistics are not released in Canada, it is difficult to produce an accurate account of the victimization and offending experiences of racial and ethnic minorities. In 2004, Canada's General Social Survey (a national, random sample survey) found no significant difference among visible minorities and whites with regard to property and violent victimization. This may be because the aggregate 'visible minority' category failed to distinguish between racial and ethnic groups (Wortley and Owusu-Bempah, 2012: p. 15).

Like in the United States, some studies in the Canadian context have shown that certain ethnic groups are overrepresented as victims of crime. While much of the discussion on overrepresentation in the US centers on blacks, much of the focus in the Canadian context lies on the overrepresentation of Aboriginals (both with respect to offending and victimization). Nationally, the homicide victimization rate of Aboriginal people is up to eight times that of non-Aboriginals (Brzozowski et al., 2006). Serious violence in large Canadian cities is also disproportionately concentrated in the black community (though to a lesser extent than in the Aboriginal community), and the homicide victimization rate for blacks in Toronto was four times the city average (Gartner and Thompson, 2004). Black males are overrepresented as homicide victims (4% of the city population, but 40% of homicide victims) (Wortley, 2008).

While census and federal correctional data suggest that Aboriginal and blacks are grossly overrepresented in the Canadian prison system, the representation varies across regions (Wortley and Owusu-Bempah, 2012: p. 17) and attempts to explain this discrepancy are made more difficult by the dearth of relevant statistics. Some research indicates that both overrepresented groups (black and Aboriginal people) *may* commit certain types of crimes more often than other racial groups, but this is inconclusive and their different rates of offending (especially for street-level crime and violence) can be explained by socioeconomic disadvantage (Wortley and Owusu-Bempah, 2012).

Possible Explanations for Overrepresentations

A potential explanation for racialized differences in victimization could be the intersection between race and class. Much criminological research has shown that most crimes are intragroup offenses; i.e., the offender and victim often come from the same ethnic background. Moreover, persons with lower income are more at risk of property and violent crime than those with high incomes, particularly for persons who are not of Asian origin. For example, blacks in low-income households face greater risk of violence than those in higher income households (Kalunta-Crumpton and Eijogu, 2012: p. 50). Racialized minorities still tend to occupy the more marginalized positions in society. In other words: they are more likely to be classified as low income, and thus, face a higher risk of both criminalization and victimization.

Aside from observed differences in the offending and victimization of certain ethnic minorities, these groups are also subject to differential treatment by the criminal justice system as a consequence of systemic disadvantage or more overt forms of racism. Decisions made in the criminal process that seem innocuous and neutral can contribute to the systemic

disadvantage of members of certain minority groups. For example, the decision to detain someone before trial, even when ostensibly based on an accused person's stability or likelihood to appear, can disadvantage minority groups who are more likely to be low income, unemployed, or lack a permanent residence (Tonry, 2011: p. 16). These disparities early in criminal proceedings have a cumulative effect; persons detained before trial are more likely to be incarcerated after trial. Studies on pretrial detention have been performed in the United States, Canada, and Europe (see Tonry, 2011 for more extensive coverage).

A popular example of racism in official crime policy is the differences in sentencing for crack cocaine vs powder cocaine charges in the United States. Crack cocaine, associated with low income, racialized users, is associated with much harsher sentences than the more expensive powder cocaine. Another area of contention are 'three strikes' laws that impose significantly lengthier sentences for third offenses. Some have argued that these laws unfairly impact blacks in the United States because of their overrepresentation in violent crime (Tonry, 2011). The continued use of the death penalty in certain American states has also been criticized as biased toward black offenders (Zeisel, 1981). In the Canadian context, racial profiling by the police and biased treatment in the criminal justice system (e.g., in sentencing, pretrial detention decisions, police use of force) also unevenly impacts certain groups. A study from Kingston, Ontario analyzed police stops and found black residents were three times more likely to be stopped than whites, and the difference in stop rates could not be explained by age, gender, location, or reason of the stop (Wortley and Marshall, 2005).

While racial profiling and disparities at the pretrial and sentencing stage may partly explain the overrepresentation of racialized minorities in certain crimes, there is still much to learn with respect to the degree in which these factors play a role and to what degree with need to turn to other structural forces (such as income disparities, segregation, historical and contemporary racism, etc.). One consequence of differential treatment in the criminal justice system and racial profiling is that it may further alienate communities who are already distrustful of the state and the police by confirming suspicions of discriminatory practices.

See also: Discrimination: Racial; Ethnicity, Sociology of; Income Inequality; Inequality, Social; Integration, Social; Migration and Social Capital; Migration: Motivations; Minorities; Prejudice in Society: Psychological Perspectives; Prejudice in Society: Sociological Perspectives; Racial Relations; Racism, Sociology of; Residential Segregation in the United States; Social Justice: Historical and Theoretical Considerations.

Bibliography

Becker, G.S., 1968. Crime and punishment: an economic approach. Journal of Political Economy 76 (2), 169–217.

Berardi, L., Bucerius, S.M., 2013. Immigrants and their children: evidence on generational differences in crime. In: Bucerius, S.M., Tonry, M. (Eds.), The Oxford Handbook of Ethnicity, Crime, and Immigration. Oxford University Press, New York.

Brzozowski, J., Taylor-Butts, A., Johnson, S., 2006. Victimization and Offending among the Aboriginal Population in Canada [pdf]. Canadian Centre for Justice Statistics. Available at: http://publications.gc.ca/Collection-R/Statcan/85-002-XIE/85-002-XIE2006003.pdf (accessed 15.10.13.).

Bucerius, S.M., 2011. Immigrants and crime. In: Tonry, M. (Ed.), The Oxford Handbook of Crime and Criminal Justice. Oxford University Press, New York, pp. 385–419.

Bucerius, S.M., 2009. Fostering academic opportunities to counteract social exclusion. In: Frost, N.A., Freilich, J.D., Clear, T.R. (Eds.), Contemporary Issues in Criminal Justice Policy: Policy Proposals from the American Society of Criminology Conference. Cengage/Wadsworth, Belmont, CA, pp. 235–245.

Bui, H.N., 2009. Parent-child conflicts, school troubles, and differences in delinquency across immigration generations. Crime & Delinquency 55 (3), 412–441.

Butcher, K.F., Piehl, A.M., 2007. Why are Immigrants' Incarceration Rates so Low? Evidence on Selective Immigration, Deterrence, and Deportation. Rutgers University Economics Department Working Paper. Rutgers University, Newark, NJ.

Butcher, K.F., Piehl, A.M., 1998. Cross-city evidence on the relationship between immigration and crime. Journal of Policy Analysis and Management 17 (3), 457–493.

Chui, T., Tran, K., Maheux, H., 2008. Canada's Ethnocultural Mosaic, 2006 Census. Statistics Canada. Catalogue, 97-562-X.

Cloward, R.A., Ohlin, L.E., 1960. Delinquency and Opportunity. Free Press, New York, NY.

Dinovitzer, R., Hagan, J., Levi, R., 2009. Immigration and youthful illegalities in a global edge city. Social Forces 88 (1), 337–372.

Ehrlich, I., 1973. Participation in illegitimate activities: a theoretical and empirical investigation. Journal of Political Economy 81 (3), 521–565.

Engbersen, G., Van der Leun, J., De Boom, J., 2007. The fragmentation of migration and crime in the Netherlands. Crime and Justice 35 (1), 389–452.

Erlanger, H., 1995. Estrangement, machismo and gang violence. In: Lopez, A. (Ed.), Latinos in the United States, Criminal Justice and Latino Communities. Garland, New York, NY, pp. 207–220.

Farley, R., Alba, R., 2002. The new second generation in the United States. International Migration Review 36 (3), 669–701.

FBI, 2011. Crime in the United States 2011. Available at: http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2011/crime-in-the-u.s.-2011/tables/table-43 (accessed 15.10.13.).

Feldmeyer, B., 2009. Immigration and violence: the offsetting effects of immigrant concentration on Latino violence. Social Science Research 38 (3), 717–731.

Gabor, T., Roberts, J.V., 1990. Rushton on race and crime: the evidence remains unconvincing. Canadian Journal of Criminology 32, 335.

Gartner, R., Thompson, S., 2004. Trends in Homicide in Toronto. From Enforcement and Prevention to Civic Engagement: Research on Community Safety. Centre of Criminology, University of Toronto, Toronto, ON.

Gfroerer, J.C., Tan, L.L., 2003. Substance use among foreign-born youths in the United States: does the length of residence matter? American Journal of Public Health 93 (11), 1892–1895.

Hagan, J., Palloni, A., 1999. Sociological criminology and the mythology of Hispanic immigration and crime. Social Problems 46 (4), 617–632.

Hawkins, D.F., 1994. Ethnicity: the forgotten dimension of American social control. In: Bridges, G., Myers, M. (Eds.), Inequality, Crime and Social Control. Westview, Boulder, CO, pp. 99–116.

Industrial Commission, 1991. Special Report on General Statistics of Immigration and the Foreign Born Population. US Government Printing Office, Washington, DC.

Junger-Tas, J., 1997. Ethnic minorities and criminal justice in the Netherlands. Crime and Justice 21, 257–310.

Kalunta-Crumpton, A., 2012. Race, Ethnicity, Crime and Criminal Justice in the Americas. Palgrave Macmillan, New York, NY.

Kalunta-Crumpton, A., Ejiogu, K., 2012. Race, ethnicity, crime, and criminal justice in the United States. In: Kalunta-Crumpton, A. (Ed.), Race, Ethnicity, Crime, and Criminal Justice in the Americas. Palgrave Macmillan, New York, NY, p. 41.

Kleck, G., 1981. Racial discrimination in criminal sentencing: a critical evaluation of the evidence with additional evidence on the death penalty. American Sociological Review 46 (6), 783–805.

Koopmans, R., 2003. Good intentions sometimes make bad policy: a comparison of Dutch and German integration policies. In: Cuperus, R., Duffek, K., Kandel, J. (Eds.), The Challenge of Diversity: European Social Democracy Facing Migration, Integration, and Multiculturalism. Studienverlag, Innsbruck, Austria, pp. 163–168.

Kubrin, C., Trager, G., 2013. Immigration and crime in US communities: charting some promising new directions in research. In: Bucerius, S.M., Tonry, M. (Eds.), The Oxford Handbook of Ethnicity, Crime, and Immigration. Oxford University Press, New York.

Kubrin, C., Ishizawa, H., 2012. Why some immigrant neighborhoods are safer than others: divergent findings from Los Angeles and Chicago. The Annals of the American Academy of Political and Social Science 641, 148–173.
LaLonde, R.J., Topel, R.H., 1991. Immigrants in the American labor market: quality, assimilation, and distributional effects. The American Economic Review 81 (2), 297–302.
Lee, M.T., Martinez Jr., R., 2002. Social disorganization revisited: mapping the recent immigration and black homicide relationship in Northern Miami. Sociological Focus 35 (4), 363–380.
Lee, M.T., Martinez, R., 2009. Immigration reduces crime: a review of the emerging scholarly consensus. In: Mcdonald, W.F. (Ed.), Immigration, Crime, and Justice (Sociology of Crime, Law, and Deviance, vol. 13). Emerald/JAI Press, Bingley, UK, pp. 3–16.
Lee, M.T., Martinez, R., Rosenfeld, R., 2001. Does immigration increase homicide? The Sociological Quarterly 42 (4), 559–580.
Martinez Jr., R., 2002. Latino Homicide: Immigration, Violence, and Community. Routledge Press, New York, NY.
Martinez Jr., R., Lee, M.T., 2000. On immigration and crime. Criminal Justice 1 (1), 486–524.
Martinez Jr., R., Stowell, J.I., 2012. Extending immigration and crime studies: national implications and local Settings. Annals of the American Academy of Political and Social Science 641 (1), 174–191.
McDonald, W.F., 1997. Crime and illegal immigration: emerging local, state, and federal partnerships. National Institute of Justice Journal 232, 2–10.
Merton, R.K., 1938. Social structure and anomie. American Sociological Review 3 (5), 672–682.
Min, P., Kim, Y., 2009. Ethnic and sub-ethnic attachments among Chinese, Korean, and Indian immigrants in New York City. Ethnics and Racial Studies 32 (5), 758–780.
Newburn, T., Shiner, M., Hayman, S., 2004. Race, crime and injustice? Strip search and the treatment of suspects in custody. British Journal of Criminology 44 (5), 677–694.
Ousey, G.C., Kubrin, C.E., 2009. Exploring the connection between immigration and violent crime rates in US cities, 1980–2000. Social Problems 56 (3), 447–473.
Owusu-Bempah, A., Millar, P., 2010. Research note: revisiting the collection of justice statistics by race in Canada. Canadian Journal of Law and Society 25 (1), 97–104.
Passel, J.S., 2006. The Size and Characteristics of the Unauthorized Migrant Population in the US. Pew Hispanic Center, 7.
Picot, G., Hou, F., 2011. Preparing for Success in Canada and the United States: The Determinants of Educational Attainment Among the Children of Immigrants. Statistics Canada Analytical Branch Studies Working Paper, no. 332.
Portes, A., Fernandez-Kelly, P., 2008. No margin for error: educational and occupational achievement among disadvantaged children of immigrants. The Annals of the American Academy of Political and Social Science 620 (1), 12–36.
Portes, A., Rumbaut, R.G., 2001. Legacies: The Story of the Immigrant Second Generation. University of California Press, Berkeley, CA.
Portes, A., Zhou, M., 1993. The new second generation: segmented assimilation and its variants. The Annals of the American Academy of Political and Social Science 530 (1), 74–96.
Reid, L.W., Weiss, H.E., Adelman, R.M., Jaret, C., 2005. The immigration–crime relationship: evidence across US metropolitan areas. Social Science Research 34 (4), 757–780.
Rowe, M., 2004. Policing, Race Racism. Willan Publishing, Portland, OR.
Rumbaut, R.G., Ewing, W.A., 2007. The Myth of Immigrant Criminality and the Paradox of Assimilation: Incarceration Rates Among Native and Foreign-born Men. Immigration Policy Center Special Report. American Immigration Law Foundation, Washington, DC.
Sampson, R.J., March 11, 2006. Open Doors Don't Invite Criminals. New York Times.
Sampson, R.J., Morenoff, J.D., Raudenbush, S., 2005. Social anatomy of racial and ethnic disparities in violence. American Journal of Public Health 95 (2), 224–232.
Sellin, T., 1938. Culture conflict and crime. The American Journal of Sociology 44 (1), 97–103.
Shaw, C.R., McKay, H.D., 1969. Juvenile Delinquency and Urban Areas. University of Chicago Press, Chicago, IL.
Stowell, J.I., Messner, S.F., McGeever, K.F., Raffalovich, L.E., 2009. Immigration and the recent violent crime drop in the united states: a pooled, cross-sectional time-series analysis of metropolitan areas. Criminology 47 (3), 889–928.
Stowell, J.I., DiPietro, S., 2013. Ethnicity, crime, and immigration in the United States: crimes by and against immigrants. In: Bucerius, S.M., Tonry, M. (Eds.), The Oxford Handbook of Ethnicity, Crime, and Immigration. Oxford University Press, New York.
Sutherland, E., 1924. Criminology. Lippincott, Philadelphia, PA.
Taft, D.R., 1933. Does immigration increase crime? Social Forces 12, 69–77.
Telles, E.E., Ortiz, V., 2008. Generations of Exclusion: Mexican Americans, Assimilation, and Race. Russell Sage Foundation, New York, NY.
Tonry, M., 1997. Ethnicity, crime, and immigration. Crime and Justice 21, 1–29.
Tonry, M., 2011. The Oxford Handbook of Crime and Criminal Justice. Oxford University Press, New York, NY.
Travis, A., 24 August 2002. After 44 Years Secret Papers Reveal Truth about Five Nights of Violence in Notting Hill. The Guardian. http://webcache.googleusercontent.com/search?q¼cache: http://www.theguardian.com/uk/2002/aug/24/artsandhumanities.nottinghillcarnival2002.
Troper, H., 1993. Canada's immigration policy since 1945. International Journal 48 (2), 255–281.
US Census Bureau, 2012. Population Estimates. http://www.census.gov/popest/data/index.html (accessed 15.10.13).
van Gemert, F., 1998. Ieder voor zich: kansen, cultuur en criminaliteit van Marokkaanse jongen. Het Spinhuis, Amsterdam.
Vega, W.A., Zimmerman, R.S., Warheit, G.J., Apospori, E., Gil, A.G., 1993. Risk factors for early adolescent drug use in four ethnic and racial groups. American Journal of Public Health 83 (2), 185–189.
Viruell-Fuentes, E.A., 2007. Beyond acculturation: immigration, discrimination, and health research among Mexicans in the United States. Social Science & Medicine 65 (7), 1524–1535.
Waters, T., 1999. Crime & Immigrant Youth. Sage Publications, Thousand Oaks, CA.
Winders, J., 2013. Criminalizing settlement: the politics of immigration in the American South. In: Bucerius, S.M., Tonry, M. (Eds.), The Oxford Handbook of Ethnicity, Crime, and Immigration. Oxford University Press, New York.
Wortley, S., Marshall, L., 2005. Race and Police Stops in Kingston, Ontario: Results of a Pilot Project. Kingston Police Services Board, Kingston, ON.
Wortley, S., 2008. A province at the crossroads: statistics on youth violence in Ontario. In: McMurtry, R., Curling, A. (Eds.), Review of the Roots of Violence, vol. 5. Queen's Printer for Ontario, Toronto, ON, pp. 1–64.
Wortley, S., 2003. Hidden intersections: research on race, crime, and criminal justice in Canada. Canadian Ethnic Studies 35 (3), 99–117.
Wortley, S., 1999. Northern taboo: research on race, crime, and criminal justice in Canada. Canadian Journal of Criminology 41, 261.
Wortley, S., Owusu-Bempah, A., 2012. Race, ethnicity, crime, and criminal justice in Canada. In: Kalunta-Crumpton, A. (Ed.), Race, Ethnicity, Crime, and Criminal Justice in the Americas. Palgrave Macmillan, New York, NY, p. 11.
Zeisel, H., 1981. Race bias in the administration of the death penalty: the Florida experience. Harvard Law Review 95, 456.
Zhou, M., Kim, S.S., 2006. Community forces, social capital, and educational achievement: the case of supplementary education in the Chinese and Korean immigrant communities. Harvard Educational Review 76 (1), 1–29.

Relevant Websites

FBI, 2011. http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2011/crime-in-the-u.s.-2011/tables/table-43.
Travis, A., August 24, 2002. After 44 Years Secret Papers Reveal Truth about Five Nights of Violence in Notting Hill. The Guardian. Retrieved from: http://webcache.googleusercontent.com/search?q=cache:http://www.theguardian.com/uk/2002/aug/24/artsandhumanities.nottinghillcarnival2002.
US Census Bureau, 2013. http://www.census.gov/popest/data/index.html.

Crime in the Rural Context

Joseph F Donnermeyer, School of Environment and Natural Resources, The Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of crime in a rural context, or rural criminology, is advancing rapidly, even though it was slow to develop during much of the twentieth century. This article addresses five key areas within rural criminology, including agricultural crime, environmental crime, drugs, violence, and rural communities and crime. The literature in each area is reviewed, and implications of rural crime scholarship for criminological theory are noted, especially for how this research reveals fallacies and suggests revisions to place-based theories and concepts like social disorganization theory and collective efficacy. This article on crime in a rural context is not comprehensive, however, but observes that there are substantial bodies of literature on issues associated with hate crime and hate groups, crime at rural-located parks, recreation areas, tourist sites, and other public places, gangs, other forms of organized crime, and issues associated with policing, courts, jails, and other infrastructural aspects of the criminal justice system.

Introduction

Until very recently, the world's population was a majority rural population. Even though it was estimated that this would change sometime during 2008 (for the first time in human history), the percentage of people who reside somewhere other than cities and suburbs will remain a significant share of all people who live on this planet for quite some time (Population Reference Bureau, 2007). In 2015, about 3.36 billion (48.1%) individuals were estimated to be rural, and rural people will exceed 3 billion throughout the first half of the twenty-first century, even though the non-urban share of the world population will decline to about 33% (Department of Economic and Social Affairs, United Nations, 2012).

Even in highly urbanized countries like Australia, Canada, Great Britain, and the USA, where so much of criminological scholarship is focused, rural people are a significant share of the citizenry. For example, in the USA, about one in six residents live in areas defined as rural, according to the US Bureau of the Census (2012). This represents an astonishing 52 plus million individuals, which exceeds the total populations of countries like Kenya, South Korea, Spain, and Ukraine, and nearly equal to the population of South Africa. If the US rural population was a 'stand alone' country, its size would rank it as about the 26th most populous. Likewise, rural people compose about 10% of all Australians, nearly 19% of Canada's inhabitants, and 20% of people who live in the UK (Department of Economic and Social Affairs, United Nations, 2012).

Even though a great many of these rural people live within easy driving distance of large, urban centers, it remains true that despite a continuing transformation of the world from rural and agricultural to urban and industrial/postindustrial, crime in the rural context is important for understanding crime in general. Yet, the criminological communities in these four countries, and most other regions of the world, have virtually ignored rural crime; that is, until recently.

It is indeed ironic that the strongly held urban bias of criminology arose during the very time when these countries were still majority rural societies, and when accounts of highway robbery and frontier outlaws were closer in time than today. The second irony is that during the past three decades, when the urban transformation of the world has continued at an unerring pace, rural criminology as a specialization within criminology has begun to emerge (Dingwall and Moody, 1999; Hogg and Carrington, 2006; Donnermeyer and DeKeseredy, 2014).

Rural Crime

Perhaps one reason that rural crime was ignored in the past by criminologists is the difficulty in defining and specifying what constitutes rural. Certainly, where census or population data is available, criteria for defining rural can be clear, although the definitions vary from country to country. As well, within a single country like the USA, there are multiple definitions of rural, even though the two most popular are: (1) a traditional census definition of rural as any place with less than 2500 people that is not contiguous to a larger, urban locality; and (2) 'rural' counties that are defined as either micropolitan (with a place larger than 10 000 persons but no bigger than 50 000) and non-metropolitan (without a place larger than 10 000 inhabitants) (US Bureau of the Census, 2010). Hence, cross-county comparative statistics on official rates of crime and arrests between rural and urban places becomes problematic, both empirically and conceptually.

Another layer of complexity lies in interpretation of the statistics, since it can be argued that there is an inherent bias in police records that belie objective comparisons because of the generally larger per capita presence of police in urban communities, hence, rendering the higher rates so often found for urban crime as an artifact of the reporting system itself (Maltz and Targonski, 2002). In short, where there are more police, there will be more reported crime.

A second reason is the nearly impossible task of differentiating clearly the generalizing features of rural places from urban places. In particular, the growth and spread of electronic media and the rise of mass society complicates further the sociological differences of rural from urban. Since the establishment of the Chicago School of Sociology in the first third of the twentieth

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45049-X

century and the publication of the still frequently cited work of Louis Wirth (1938) on 'Urbanism as a Way of Life,' coupled with the proportionate size of the urban population in countries where criminology is the most highly developed and where the official statistics suggest a higher rate of crime for urban places, rural was mostly ignored.

The problem is that many criminologists confuse what rural means, and by extension, what urban means. Many fall back on simplistic dichotomies which conflate the geography of 'space' with the economic, cultural and sociological characteristics of 'place' (Liepins, 2000). The most famous of these dichotomies is the well-known distinction by Tonnies – gemeinschaft vs gesellshaft – with the former representing in the minds of many mainstream criminologists the concepts of rurality, cohesive places, high levels of collective efficacy, and relatively little crime; and the latter linked to urbanity, social disorganization, and relatively more crime. Today, a concept like collective efficacy, which means "social cohesion combined with shard expectation" is given a one-dimensional treatment; that is, it must only be associated with less crime (Sampson, 2012, p. 27).

Another conceptual fallacy is made by criminologists who attempt to define rural and urban distinctions based on theories and concepts of social change (Donnermeyer, 2012). Hence, rural comes to mean preindustrial, agrarian, small, 'folk'-oriented groups with mechanical solidarity and urban depicts industrial/postindustrial, and large-scale forms of social organization with organic solidarity.

The rural (and urban) reality is that in order to advance criminological thought, dichotomies should be replaced by more complex conceptualizations that start with a geographically-based definition of rural, and by extension, urban. Specifically, rural places have smaller populations and/or population densities, and this is perhaps the only difference between rural and urban spaces common across the spectrum of developed and developing societies. The real point is that the specific social organizational, cultural, and economic profile of rural areas varies as greatly as do the specific characteristics of urban neighborhoods. Hence, crime varies as well. For example, rates of robbery, according to the FBI's (2012) *Uniform Crime Reports*, is many times lower at rural places when compared to urban localities, but the percentage of burglaries experienced by farmers may be higher than all but a few urban neighborhoods (Barclay and Donnermeyer, 2007). Hence, from this point of view, what the field of criminology has actually ignored is the evidence of expressions of crime at smaller places, and by doing so, fails to establish a more generalized validity to criminological theories like social disorganization, and frequently cited concepts, like collective efficacy. Even though the rural world is shrinking proportionately, it still represents a significant share of the population of both developed and developing countries, and criminology cannot pretend to have valid theories if it cannot make these theories work outside of an urban context.

There are five areas where the rural criminology literature has focused most keenly in recent times. Each describes a rural reality different from anything assumed by dichotomous thinking and simplistic theorizing. They include: agricultural crime, environmental crime, drugs, violence, and rural communities and crime.

Agricultural Crime

The industrialization of farming in advanced capitalist countries has had two concurrent consequences. First, the proportion of the population which lives on a farm has declined to such an extent, that in a country like the USA, less than 2% of its residents are directly involved in commercially grown/raised food (even though there are many more people who have food gardens in their backyards or raise some livestock part-time, perhaps as a hobby/leisure pursuit). Second, to maintain efficiency in food production under globalized market conditions requires a huge economic investment in machinery, equipment, and supplies.

Even though 'rustling' livestock forms a romanticized part of the frontier lore in Australia, Canada, the USA, and other places, few systematic studies of agricultural crime have ever been conducted. However, those that have been completed paint a picture of extensive victimization experienced by agriculturalists. Research from Australia, the UK, the USA (Barclay and Donnermeyer, 2011), and the east African country of Kenya (Bunei et al., 2013), show annual rates of victimization as follows: for machinery and equipment theft – 10–25%; for burglary – 6–12%; for theft of farm animals – 1–13%; and for vandalism – 13–24%. The wide range (ratio of low to high estimate) of animal theft is partly explained by the type of agricultural operations that occupy an area, which in turn, depends partly on climate and topography. Some areas are mostly devoted to crops and vegetables, while other localities are predominately dedicated to livestock. In these latter areas, the theft of animals can be much more extensive.

Some farm crime may be the case of a food producer whose livestock is stolen by a neighbor who is also a farmer, as Barclay (2003) found in her study of agricultural victimization in New South Wales. Also remarkable in Barclay's (2003) research was the pressure on the farmer-victims to not inform the police ('dob in') about their neighbors-as-suspects, which is contradictory to the notion that cohesive rural communities where everyone knows everyone else form the conditions under which crime is restrained or controlled. Simply put, in smaller groups, in this case, smaller communities, one faces ostracization for snitching. Second, the research indicates that forms of violent crime victimization (with one exception) are low to nearly nonexistent for food producers (Barclay and Donnermeyer, 2011). The exception (discussed below) is domestic violence: that is, how various economic and social forces related to farming facilitate abuse of farm women by their partners (Carrington et al., 2013).

There are no reliably accurate estimates of the economic costs of agricultural crime. However, Swanson et al. (2002) suggested losses to agriculturalists in the USA alone from crime at about $5 billion annually. More recently, Mears et al. (2007) estimated both high rates of victimization and over $10 million in stolen property annually for a single nine-county area of central California. The list of items stolen is remarkably diverse, but all of the property had economic value and could be converted to cash, including scrap metal, barbwire, saddles and tack, irrigation pipe, water pumps, diesel fuel, beef cattle and other farm animals, nearly ripened fruit from groves of trees and bushes, chemicals, and welders. Echoing these claims is a survey of farmers in Hawaii by the National Agricultural Statistics Service of the US Department of

Agriculture, which found an overall victimization rate to agricultural operations of 17% and a total dollar cost equivalent to 3% of net farm income, plus another 6% for the cost of security (NASS, 2005).

The extant criminological literature on agricultural victimization, especially the work by Mears et al. (2007), demonstrates the relationship between social change and crime. There are two primary factors that explain why farm crime should be taken seriously, both of which are commensurate with routine activities theory (Barclay and Donnermeyer, 2011). First, the industrialization of agriculture in many parts of the world has made farm property increasingly attractive for theft. Simply put, if the property is valuable and can be liquidated for cash, it stands a greater chance of being stolen. Second, as the number of farms in many parts of the world has declined in response to the industrialization of agriculture, the size of farms has increased. Hence, the physical or geographic distribution of farm property is spread out in ways that make it difficult for the agriculturalist, other family members, and even farm employees/laborers to maintain a natural form of surveillance, that is, as they go about their daily routines. The consequence is an increase in opportunities for crime as guardianship declines with farm size.

Environmental Crime

One of the fastest growing specializations in criminology is the study of environmental crime, which has come to be known as 'Green Criminology.' The adoption of the moniker 'Green Criminology' also expands the scope of meaning of criminological scholarship to include not only actions by individuals, corporations, and states which violate the law, but actions which harm peoples and communities through environmental degradation and threatened livelihoods (White, 2009; South and Brisman, 2013).

Rural criminology is at its core place based, hence, a great deal of environmental crime can also be regarded as rural crime; or, in cases of transnational networks for the illegal smuggling of flora and fauna, as equally rural as urban in context. For example, illegal logging is a rural located crime, but with obvious urban linkages and connections to other forms of crime as well. A recent news story describing the poaching of redwoods in northern California illustrates this connection. The trees are stripped of their burls and other knotty wood by thieves who sell the pieces for money. Burls can fetch over $2 a pound, and a redwood table on E-Bay can be worth $1000 plus. One law enforcement officer who has interviewed those apprehended for the offense indicated that most said they use the money to support their drug habits (Graham, 2014). McSweeney et al. (2014) note that clandestine drug operations in central America have tipped the balance of nature to such an extent that whole swaths of rain forest are destroyed or denuded, which has rippled through the natural ecology of the region, affecting other plants and animals as well, plus the people who live there. Also common are logging companies who violate state and national laws regulating tree harvesting, and in many parts of the world, harvest rare and endangered tree varieties, and often trespass in preserves, national parks, and protected areas to find valuable wood for sale on world markets (Green et al., 2009).

Poaching and various violations of hunting and fishing regulations and violations of harvesting quotas for animals are two other areas where most of the offenses are rural located. In these particular cases, the offenders can be thought of as mostly individual violators – such as hunters and individuals who make a living by harvesting fresh and seawater fishes – but both corporations and governments can also play roles in poaching.

As well, violations of environmental regulations for the use of chemicals by farmers, illegal dumping by individuals and corporations, and land-grabbing/theft, especially from indigenous peoples in countries like Brazil (White, 2012), adds to the list of serious environmental crimes which are also rural located. Finally, environmental justice issues have a significant rural dimension. The location of landfills in rural areas (or near working class/underclass neighborhoods in cities), the externality costs of fracking, oil drilling, and other forms of natural resource development, and impacts on crime and fear of crime from the boom and bust cycles created by energy development near rural communities, and many more – all point to significant problems associated with rural places whose long-term residents often lack the political capital to withstand environmental victimization and harm.

It could be argued that since a great deal of environmental crime is committed by individuals who are not permanent residents and that much of it represents complex crime networks that are regional and worldwide in scope, it is not actually rural crime. The counterargument is to apply the same logic in the other direction by pointing out that a crime committed in downtown Chicago by a tourist or visitor who lives at a rural place is in fact an example of rural, not urban, crime. A positive answer to the latter is nonsensical, as is the assertion made in the former. Hence, the ultimate solution is simple: first, ask where the crime or offense occurred; and secondly, then consider both the micro- and marcocultural, economic, and sociological factors associated with the crime at its place of commission.

Drugs

Drug use, production and trafficking in the rural context are perhaps the most thoroughly researched subjects of any in rural criminology. However, much of the research is not conducted by criminologists, but by scholars in the fields of public health, psychology, social psychology, and sociology. Nonetheless, along with the criminologists who pay attention to issues of drugs in the rural context, there is substantial evidence for the argument that it is at least as great a problem affecting rural peoples and communities as the urban populations of many countries. Consider, for example, evidence from the *Monitoring the Future* (MTF) research (Johnston et al., 2014). This annual study among a nationally representative sample of adolescents and young adults in the USA shows that today, there is little difference in rates of usage between respondents living in urban counties when compared to youth from rural counties for most illicit substances. Since the MTF has been continuously collecting this data, beginning in the late 1970s, it is also possible to trace when this convergence occurred, which indicates sometime during the 1990s. The results suggest that whatever immunities rural communities had at one time, which may be

based more on popular and criminological mythologies about the rural, those times are now lost (Donnermeyer, 2012).

A report by the Carsey Institute (Van Gundy, 2006) suggests how time and change affect the criminology of places, large and small. The report describes how mining companies in places like Harlan County, Kentucky dispensed painkillers as a way to squeeze more profit out of the labor of miners who had to crouch in the shafts of mines for long hours to extract the coal. Soon many of the locals were learning how to remove the time release coating on the pills to achieve a more immediate effect. With the emergence of a culture of drug use, the spread of newer and more powerfully addictive drugs like oxycontin and heroin was made easier. Similarly, Stallwitz and Shewan (2004) document the relationship of energy resource development as a source of social change and drug use. Specifically, in the Shetland Islands of the UK, the arrival of shipyard workers associated with port facilities constructed for the transport of oil from the North Sea brought with them the drug of heroin. However, it did not cause a moral outcry or panic, as the local culture, which was already attuned to pubs and heavy alcohol use, was tolerant of the new addiction with the rationalizations that the users were gainfully employed and did not engage in other crimes, such as burglary and unacceptable forms of public disorder. Hence, like the research by Barclay (2003), there are clear links between local forms of collective efficacy and crime.

Weisheit's work on both marijuana (1992) and methamphetamine (2008) production portrays the extent to which the supply of illicit drugs has a decided rural dimension, as does the previously cited work of McSweeney et al. (2014). Many rural places in the USA can be described as 'persistent poverty' counties, meaning that they have experienced above average levels of poverty for not only many continuous years, but for several decades (Economic Research Service, 2011). It is at these rural localities where drug use and drug production often can be found together. Garriott (2011) presents an intimate portrayal of methamphetamine's impact on the cultural and social life in a rural region of West Virginia. However, he does not demonstrate a relationship of social disorganization and crime, but instead how the social organization of a rural place was changed and reorganized by its introduction, including the ways local police operate, the emergence of drug clinics, and other coping mechanisms among its citizens.

Violence

There are two major strands of rural-located criminological science related to the subject of violence. The first is the work of criminological scholars who focus on violence against women, and in particular, violence by intimate partners. For example, DeKeseredy and Schwartz (2009) studied a rural region within the hilly or so-called 'Appalachian' area of southeastern Ohio. Although it is impossible to estimate the true rate of intimate partner violence, their research and previous work by other scholars (Donnermeyer and DeKeseredy, 2014) suggests that domestic violence may be much higher in poor, rural communities of the Appalachian region than in the city. More importantly, what they discovered is the extent to which forms of patriarchy pervade the social fabric of these communities. Based on their interviews, they found a 'male peer support system' important for enabling this form of violence. The support system consisted not only of learned rationalizations by men for why it is okay to physically and verbally abuse their girlfriends and wives, but of cohesive networks of peers who shared this information on a face-to-face basis. Abusers would meet at a local restaurant or other spots, sharing coffee, beer, rationalizations, and techniques to abuse without being detected by others. As well, violence against women was tolerated, that is, part of the normative culture of many of these small communities, and local law enforcement was sometimes reluctant to recognize domestic violence as an offense worth responding to.

What is perhaps most significant about their research is the spotlight they shine on the diversity of smaller places. This diversity is not in the traditional way criminologists think about it – which is racial, ethnic, or some other form of demographic variety in a population – but instead, is a form of sociological diversity. The research by DeKeseredy and Schwartz (1996) of violence against women at places with smaller populations helps to illustrate the geographic simultaneity of social organization which both enables and constrains crime. In essence, they and others who have focused on this subject in the rural context have found forms of collective efficacy associated with violence existing alongside forms of collective efficacy that constrain violence and other types of crime. Hence, violence against women is frequent, but robbery is infrequent. Similarly, research on violence against women in a number of other localities, from Australia to India to various African countries, demonstrates the relationship of social organization and culture to crime at specific places with smaller populations (Donnermeyer and DeKeseredy, 2014).

Another strand of research on violence in the rural context comes from those who look more generally at interpersonal violence, or more specifically at homicide (Donnermeyer and DeKeseredy, 2014). Some work on the culture of violence has examined possible statistical links of regional expressions of cultural themes, such as a 'culture of honor' in the southern USA (Nisbett and Cohen, 1996) to rates of assault and homicide. Others have taken a more qualitative approach and looked at how 'pub'lic (i.e., pubs as public places) expressions of masculinity are associated with interpersonal violence in many rural communities (Campbell, 2000). Still others have attempted large-scale statistical examinations of rural places and rates of homicide using the logic of social disorganization theory or related theories, such as civic community theory (a summary and test of both theories can be found in Wells and Weisheit, 2012), the results of which are discussed in the next section.

Rural Community Characteristics and Crime

Nowhere is the need for criminology to overhaul its theoretical focus more apparent than in studies of rural communities and crime. Further, nowhere has the burgeoning field of rural criminology presented a critique of mainstream criminological theory than the accumulated work on rural communities and crime. This body of research uses a very standard approach insofar as the research is: (1) regional or national in scope with a large number of cases; (2) the unit of analysis or case is a county or something else reflecting governmental jurisdictions

and boundaries; (3) the use of official police statistics for calculation of crime or arrest rates, combined with census or other secondary source data for measuring community characteristics; and (4) the employment of the logic of social disorganization theory and concepts like collective efficacy, social capital, and civic community (Donnermeyer and DeKeseredy, 2014).

This rural-based literature, mostly from studies conducted in the USA, consistently finds that economic indicators are not as strong in predicting variations in crime within rural communities when compared to urban-based studies. Specifically, the research has discovered that high poverty rural communities had lower official rates of crime. Further, these communities had relatively lower rates of population turnover or instability. Both findings are contradictory with the urban-based research, which claims that poverty and population turnover are forms of disorganization and conditions that weaken collective efficacy, hence, crime rates should be higher.

Since the research on rural violence, as noted above, seems to suggest the opposite finding for rural communities, the issue boils down to one of two explanations. One is methodological – the quantitative studies versus the more qualitative approaches of rural work on violence against women produces differing results because of the ways the data was collected and analyzed (police data versus personal interviews with victims). The other is theoretical, that is, theories like social disorganization are not generalizable to places with smaller and less dense populations without significant modification. Wells and Weisheit (2012) helped solve the dilemma by testing both social disorganization and a related theory known as civic community (Lee and Thomas, 2010) on arrest rate variations. As the population size of the county declined, so too did the predictive ability of both theories. As well, Kaylen and Pridemore (2012) addressed a series of inconsistencies in the rural findings that adopt a social disorganization perspective, and concluded that even though the mediating factors (i.e., cultural and social dimensions – the collective efficacy which is internal to a place) may be the same across rural and urban localities, the relationship of antecedents like poverty and population turnover (in the sense of antecedents in a regression analysis) to crime may be different between rural and urban communities. Hence, the problem has more to do with the theory of social disorganization and its reliance on the establishment of dependent variables derived from official police statistics.

In the past, some researchers (Donnermeyer, 2012; Weisheit et al., 2006) have noticed that there are indeed rural communities who manifest higher rates of official crime and arrests, such as recorded in the FBI's *Uniform Crime Reports*, than metropolitan communities. Hence, there is overlap in rates, which belie dichotomous thinking about rural and urban crime. Together, both the quantitative research on crime and rural communities (Wells and Weisheit, 2012; Kaylen and Pridemore, 2012) and scholarship associated with farm victimization by Barclay (2003), drug use by Stallwitz (2014) and Garriott (2011) and violence against women by DeKeseredy and Schwartz (1996) challenge directly the validity of social disorganization theory. This is important for understanding crime in the rural context because the deconstruction of urban-biased approaches to explaining crime and place are necessary before a full appreciation of the importance of crime in the rural context can occur. If sometime in the past, dichotomous thinking about rural = gemeinschaft = cohesion = less crime versus urban = gesellschaft = disorganization = more crime, had not yet been firmly set in place, and the comparative study of crime in places of different population sizes and densities had been more frequent throughout criminology's development, this challenge would not be necessary. Unfortunately, the growth of rural criminology was stunted by a past urban-centric approach to crime, and now that it has emerged as a vital field of inquiry, it is finding that place-based criminological theory needs revising, if not complete rewriting.

In Conclusion: Rural Mythologies

Rural crime should be taken seriously because it affects a significant share of the world's population. Even in countries with large criminological communities – Australia, Canada, the UK, and the USA – a significant share of citizens live and work at smaller, more rural places. This encyclopedia article has reviewed five primary areas where crime has a significant rural dimension, namely, agricultural crime, environmental crime, drugs, violence, and variances in crime and arrest rates across a diversity of rural community characteristics. There are other areas not reviewed here where the rural dimension needs more consideration, such as hate crime and significant presence of hate groups in rural places; crime at rural-located parks, recreation areas, tourist sites, and other places used by the general public; gangs and other forms of organized crime; arson; and issues associated with policing, courts, jails, and other infrastructural aspects of the criminal justice system.

Without a doubt, however, the emergence of rural criminology has already helped dispel two fundamental mythologies. The first concerns the weakness of a rural–urban dichotomy as an analytic shortcut to theorizing about crime. Rural places are incredibly diverse in their social, normative, and economic structures. It follows that expressions of crime vary across these places as well. In the earlier days of criminology's development, especially as an empirically-based science with testable theories, criminologists tended to relegate rural to a type of homogeneous residual or counterpoint to investigations of urban crime (Weisheit et al., 2006). Hence, theories such as social disorganization and concepts such as collective efficacy have one-sided interpretations. We now know that some forms of collective efficacy enable crime, while other forms constrain crime, and we know this from the work on rural crime. However, this truism applies to both rural and urban places, but in specific reference to rural crime, it demonstrates how a diversity of social structures exist, simultaneously, even in places that lack the outward complexity of cities. Whether the reference is to rural or urban places, there is really no such thing as social disorganization, only forms of organization which covary with expressions of specific crimes, which themselves may not always be positively correlated or covary in the same direction.

Second, places are rural (or urban) by virtue of their population size and population density. Nothing more! Within these smaller places is a greater heterogeneity of social organization, influence by normative orders and economic conditions,

and continuously changing, as do larger places. Hence, how they change overtime affects all aspects of crime, from crimes known to the police and those less likely reported by citizens to law enforcement, and to citizens' perceptions of what is a crime, what is safety, and what is the appropriate response. A rise in crime at a rural place cannot be merely interpreted as a feature of its urbanization, as this presents not a place-based or static dichotomy which has little value, but an equally useless time-based dichotomy with presumptions that the yesteryear of any place or of any society was relatively crime-free, and that it is only with industrialization and other aspects of modernization that crime arose. We know from history that rural places were never crime-free. Instead, the changing context of rural crime is a matter of examining the changing characteristics of smaller places (Liepins, 2000), just as the changing characteristics of urban places provides the context for understanding crime's ever-changing expression in the city.

See also: Collective Efficacy and Crime; Communities and Crime; Crime Trends and Debates; Crime, Geography of; Crime, Sociology of; Crime: Knowledge about and Prevalence; Criminology: Development and Current Status of the Discipline; Critical Criminology; Data Bases and Statistical Systems: Crime Measurement; Domestic Violence: Sociological Perspectives; Drugs: Illicit Use and Prevention; Routine Activities, Criminal Opportunities, Crime and Crime Prevention.

Bibliography

Barclay, E., 2003. Crime with Rural Communities: The Dark Side of Gemeinschaft (Ph.D. dissertation). The University of New England, Armidale, New South Wales.

Barclay, E., Donnermeyer, J.F., 2007. Farm victimisation: the quintessential rural crime. In: Barclay, E., Donnermeyer, J.F., Scott, J., Hogg, R. (Eds.), Crime in Rural Australia. Federation Press, Annandale, NSW, pp. 57–68.

Barclay, E., Donnermeyer, J.F., 2011. Crime and security on agricultural operations. Security Journal 24, 1–18.

Bunei, E.K., Rono, J.K., Chessa, S.R., 2013. Factors influencing farm crime in Kenya: opinions and experiences of farmers. International Journal of Rural Criminology 2, 75–100.

Campbell, H., 2000. The glass phallus: pub(lic) masculinity and drinking in rural New Zealand. Rural Sociology 65, 532–536.

Carrington, K., McIntosh, A., Hogg, R., Scott, J., 2013. Rural masculinities and the internationalisation of violence in agricultural communities. International Journal of Rural Criminology 2, 3–24.

DeKeseredy, W.S., Schwartz, M.D., 2009. Dangerous Exists: Escaping Abusive Relationship in Rural America. Rutgers University Press, New Brunswick, NJ.

Department of Economic and Social Affairs, United Nations, 2012. World Population Prospects: The 2012 Revision. United Nations, New York.

Dingwall, G., Moody, S.R. (Eds.), 1999. Crime and Conflict in the Countryside. University of Wales Press, Cardiff, Wales.

Donnermeyer, J.F., DeKeseredy, W.S., 2014. Rural Criminology. Routledge, London.

Donnermeyer, J.F., 2012. Rural crime and critical criminology. In: DeKeseredy, W.S., Dragiewicz, M. (Eds.), Routledge Handbook of Critical Criminology. Routledge, London, pp. 290–302.

Economic Research Service, U.S. Department of Agriculture, 2011. Rural America at a Glance 2011 Edition. Economic Research Service, U.S. Department of Agriculture, Washington, DC.

Federal Bureau of Investigation, U.S. Department of Justice, 2013. Crime in the United States 2012. Federal Bureau of Investigation, U.S. Department of Justice, Washington, DC. From: http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2012/crime-in-the-u.s.-2012 (retrieved 12.03.14.).

Garriott, W., 2011. Policing Methamphetamine: Narcopolitics in Rural America. New York University Press, New York.

Graham, K., 2014. California redwoods being poached as a source of drug money. Digital Journal. From: http://digitaljournal.com/news/environment/california-redwoods-being-poached-as-a-source-for-drug-money/article/374897 (retrieved 25.03.14.).

Green, P., Ward, T., McConnachie, K., 2009. Logging and legality: environmental crime, civil society, and the state. In: White, R. (Ed.), Environmental Crime: A Reader. Willan Publishing, Cullompton, UK, pp. 116–130.

Hogg, R., Carrington, R., 2006. Policing the Rural Crisis. Federation Press, Sydney, Aus.

Johnston, L.D., O'Malley, P.M., Miech, R.A., Bachman, J.G., Schulenberg, J.E., 2014. Monitoring the Future National Results on Drug Use: 1975–2013: Overview, Key Findings on Adolescent Drug Use. Institute for Social Research, The University of Michigan, Ann Arbor, MI.

Kaylen, M.T., Pridemore, W.A., 2012. Systematically addressing inconsistences in the rural social disorganization and crime literature. International Journal of Rural Criminology 1, 134–152.

Lee, M.R., Thomas, S.A., 2010. Civic community, population change, and violent crime, in rural communities. Journal of Research in Crime and Delinquency 47, 118–147.

Liepins, R., 2000. New energies for an old idea: reworking approaches to "community" in contemporary rural studies. Journal of Rural Studies 16, 23–35.

Maltz, M.D., Targonski, J., 2002. A note on the use of county-lvel UCR data. Journal of Quantitative Criminology 18, 297–318.

Mears, D.P., Scott, M.L., Bhati, A., 2007. A process and outcome evaluation of an agricultural crime prevention initiative. Criminal Justice Review 18, 51–80.

McSweeney, K., Nielsen, E.A., Taylor, M.J., Wrathall, D.J., Pearson, Z., Wang, O., Plumb, S.T., 2014. Drug policy as conservation policy: narco-deforestation. Science 343, 489–490.

Nisbett, R.E., Cohen, D., 1996. Culture of Honor: The Psychology of Violence in the South. Westview Press, Boulder, Colorado.

National Agricultural Statistics Service, U.S. Department of Agriculture, 2005. Hawaiian Agricultural Theft and Vandalism. NASS Field Office, Honolulu.

Population Reference Bureau, 2007. Urban Population to Become the New Majority Worldwide. Population Reference Bureau, Washington, DC. From: http://www.prb.org/Publications/Articles/2007/UrbanPopToBecomeMajority.aspx (retrieved 22.03.14.).

Sampson, R.J., 2012. Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago Press, Chicago.

South, N., Brisman, A. (Eds.), 2013. Routledge International Handbook of Green Criminology. Routledge, London.

Stallwitz, A., 2014. Community-mindedness: protection against crime in the context of illicit drug cultures? International Journal of Rural Criminology 2, 166–207.

Stallwitz, A., Shewan, D., 2004. A qualitative exploration of the impact of cultural and social factors on heroin use in Shetland (Scotland). Journal of Psychoactive Drugs 36, 367–378.

Swanson, C.R., Chamelin, N.C., Territo, L., 2002. Criminal Investigation. McGraw-Hill, New York.

U.S. Bureau of the Census, 2010. 2010 Urban and Rural Classification Main Page. From: http://www.census.gov/geo/www/ua/2010urbanruralclass.html (retrieved 11.10.12.).

U.S. Bureau of the Census, 2012. The 2012 Statistical Abstract of the United States. From: http://www.census.gov/compendia/statab/2012edition.html (retrieved 05.03.14.).

Van Gundy, K., 2006. Substance Abuse in Rural and Small Town America. Reports on Rural America. The Carsey Institute, Durham, New Hampshire.

Weisheit, R.A., 1992. Domestic Marijuana: A Neglected Industry. Greenwood, Westport, CT.

Weisheit, R.A., 2008. Making methamphetamine. Southern Rural Sociology 23, 78–107.

Weisheit, R.A., Falcone, D.N., Wells, L.E., 2006. Crime and policing in rural and small-town America, third ed. Waveland Press, Long Grove, Illinois.

Wells, L.E., Weisheit, R.A., 2012. Explaining crime in metropolitan and non-metropolitan communities. International Journal of Rural Criminology 1, 154–183.

White, R. (Ed.), 2009. Environmental Crime. Willan Publishing, Cullompton, UK.

White, R., 2012. Land theft as rural eco-crime. International Journal of Rural Criminology 1, 203–217.

Wirth, L., 1938. Urbanism as a way of life. American Journal of Sociology 40, 1–24.

Crime: Knowledge about and Prevalence

Rosemary Gartner, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Crime is a core concept and concern in many societies, although what constitutes crime can vary substantially across time and place. This is because definitions of crime are the outcome of political processes and power relations that differ among societies as they struggle to achieve social order. Most of our systematic knowledge about crime and its prevalence relies on legal definitions that, like common-sense definitions, are socially constructed. After describing sources of information on crime, this article summarizes what is known about its prevalence and distribution over time, across and within societies. It concludes with an overview of explanations for historical and contemporary trends and patterns in crime.

Defining Crime

People in all societies, in the course of their everyday lives, interact and form relationships – with each other, with larger groups and organizations, and with social institutions, including the state – that can be marked by conflict, hostility, and exploitation. At times, the behaviors or situations that arise from these are termed 'crimes' or 'criminal.' In popular discourse, crimes are acts or events that violate fundamental moral precepts and deserve to be publicly denounced and punished. This common-sense definition implies that certain behaviors are inherently criminal and necessarily invite sanctions. However, because moral precepts – as well as responses to their violation – vary substantially across time, place, and people, common-sense definitions of crime cannot denote a fixed and natural category of behaviors.

In legal discourse, crime is behavior that leads to particular legal procedures and for which the law permits the imposition of criminal penalties. Here crime is constituted by legal definitions and reactions, not by characteristics inherent in behaviors. Legal definitions of crime appear both more precise and more objective than common-sense definitions: crime is what the law defines and treats as such. However, because legal systems are products of particular historical epochs, and legal definitions are created and applied through social and political processes, crime in this definition is both historically and socially constructed and contingent. What was once criminal (e.g., possession of small amounts of marijuana or certain sexual activities between consenting adults) is now legal, and what once was legal (e.g., dumping toxic chemicals into rivers or marital rape) is now criminalized – at least in some jurisdictions. As a consequence of the slippage between common-sense and legal definitions of crime, in every society many acts or conditions that are seen as harmful or morally repugnant (at least by some) are not legally defined as crimes; and some acts that are legally defined as crimes are not uniformly seen as harmful or immoral.

Since the late twentieth century, as social interaction and networks became more diverse and extensive and as new social movements gained force, new crimes and new categories of existing crimes have been created (e.g., cyber crimes, hate crimes, and transnational crimes). In addition, institutions that legislate and sanction these crimes have also emerged and grown – some of them, such as the International Criminal Court, in recognition of the increasingly global character of crime. These trends will continue to affect legal and common-sense definitions of crime, as well as debates over them, for decades to come.

Many social and behavioral scientists who study crime rely on legal definitions and restrict their attention to acts defined as violations of the criminal law. They differ, however, on the questions they ask about crime. Some are primarily interested in the behaviors defined as crime and the people who engage in them, whereas others are interested in the processes through which particular behaviors and people are defined as criminal.

Determining how much crime occurs and explaining why people commit it have traditionally been one of the principal concerns of those who study crime. Because there is no common behavioral attribute that ties all crimes together, the task of explanation is quite complex. How, for example, can one theory account for behaviors as diverse as possession of heroin, stock fraud, vandalism, bribery of public officials, and spouse assault – each of which violates the criminal law of many countries?

The difficulty of explaining crime as behavior suggests why other researchers direct their attention to the social and political processes whereby acts and people are criminalized. Here, crime is both a product of, and a discourse used in, struggles over how social order is defined and achieved; its lack of any essential attribute highlights its ideological character. This approach downplays the near-universal criminalization of certain behaviors, such as murder, assault, theft, and arson, and would point out that the acts included in even the most serious crimes, such as murder, can vary across societies.

A third group of legal scholars, criminologists, and philosophers engage in debates over the fundamental meaning of crime and the consequences of either restricting its definition to legal categories or expanding it, for example, to include potential or future behaviors. Those who object to more restrictive definitions argue that a range of harms to humans or violations of human rights are excluded from many criminal codes and therefore escape the attention of both legal officials and criminologists (Hillyard et al., 2004). Those concerned about more expansive definitions point to, among other things, the increased use of and confidence in actuarial tools and risk profiling technology to identify who and what may become criminal (Zedner, 2007).

Notwithstanding the conceptual challenges of defining crime, the combination of two elements is ordinarily recognized as critical: human behavior and a state-sanctioned response to it. Crimes of greatest concern in most societies also involve a third element: the intention to harm or wrong another. The more serious of these crimes – what are called *mala in se* (acts wrong in themselves) – are the target of substantial measurement efforts designed to answer questions about their prevalence, distribution, and causes.

Measuring Crime

Efforts to measure crime usually are guided by legal definitions of the jurisdiction of interest and rely on statistics and other information generated by criminal justice and legal officials. Most nations compile annual reports on the number of crimes known to police and the number of persons arrested or otherwise proceeded against by the criminal justice system. For example, in the United States, the Federal Bureau of Investigation (FBI) has compiled and published the Uniform Crime Reports (UCR) annually since the 1930s, based on data from several thousand police agencies; more recently, it has provided detailed data on lethal criminal violence in its Supplemental Homicide Reports. Vital statistics agencies, such as the National Vital Statistics System in the United States, are another source of official data on homicide in many countries. The World Health Organization's data on homicide mortality frequently are used to make comparisons among nations on levels of lethal criminal violence. Other important sources of official crime data from different countries include the International Police Organization (Interpol), the United Nations Survey of Crime Trends and Operations of Criminal Justice Systems, and the European Sourcebook of Crime and Criminal Justice Statistics.

While official statistics on crime from these and other sources are a mainstay of criminological research, they present three principal difficulties in measuring crime and documenting its prevalence. First, while legal definitions and official statistics provide a reasonably reliable guide to what is designated as criminal behavior within a given society at one point in time, they may not accurately measure differences across societies in similar types of behavior. This is because (as noted before) legal definitions of even relatively serious crimes vary across societies.

The second difficulty is that, for a number of reasons, official statistics do not accurately measure levels of crime even within a single society. Characteristics of official agencies that collect information on crime, such as the police, inevitably influence how these acts are interpreted and recorded. Some acts that fit legal definitions of crime may only infrequently be subject to criminal enforcement, such as many white-collar, corporate, or political crimes. Because the public is the primary means by which officials learn about crime, popular views on crime and attitudes toward the agencies that deal with it also influence which acts are reported to legal officials. So, for example, well under half of all physical and sexual assaults are reported to police in most nations (van Dijk et al., 2007). Police and other legal officials also have considerable discretion in deciding which acts or people to proceed against and how to categorize these in official records. As a result, official statistics on crime reflect less than its true prevalence as well as decision makers' perspectives on how to classify what are often complex behaviors and situations.

Relatedly, legal definitions, public and official perceptions, and public willingness to report crime also change over time within a society. Consequently, documenting even short-term trends in the prevalence of crime is subject to error in ways that are difficult to correct. These problems are compounded when official record keeping systems alter their methods of data collection or their coverage of law enforcement agencies, both of which have affected the FBI's UCR.

Unofficial sources of crime data, such as surveys of the general population, avoid some of these problems and shed light on the 'dark figure of crime' – that is, crimes that do not come to the attention of legal officials. Self-report victimization and offending surveys, which ask people about their criminal victimization or offending over a specified period of time, yield higher estimates of the prevalence of crime than do official statistics. Victimization surveys also show that the likelihood that a crime will become known to criminal justice officials depends in part upon characteristics of the victim, the offender, and their relationship – and that the more serious the crime, the more likely it is to be reported to and recorded by officials.

Since crime victimization surveys were first fielded in the late 1960s, they have been adopted as one of the major methods of measuring crime within many countries and at the international level. The National Crime Victimization Survey (previously the National Crime Survey), which has been conducted annually in the United States since the early 1970s, is the largest and longest running national victimization survey. Other national surveys are carried out regularly in Great Britain (the British Crime Survey) and in Canada (the General Social Survey). Because national surveys conducted in different countries have used different methods and measures, the International Crime Victims Survey (ICVS) was launched in 1989 to gather data that could allow comparisons among different nations. It has since been conducted about every five years and has collected data in over 70 countries and 35 capital or major cities. Modeled after the ICVS, the European Survey on Crime and Safety was launched in 2005 to document the extent and nature of crime in 18 member countries of the European Union.

Self-report offending surveys are much less extensive, in part because of the variety of problems involved in asking people to reveal their own criminal behaviors. The most common populations sampled by self-report surveys are young people, particularly high school or college students. One of the largest and longest running such surveys in the United States is the Monitoring the Future Study, which has been conducted annually since 1975 among high school seniors and since 1991 among students in the 10th grade and higher. The International Self-Report Delinquency Study has been carried out twice (in 1988 and between 2005 and 2007) among 12–15 years olds, most recently in 30 European countries and North America.

Because they rely on the honesty, memory, and perceptions of respondents, self-report victimization and offending surveys have their own limitations and biases. For example, asking someone whether they have been the victim of spousal assault in the last year may elicit a very different answer than asking the

same person if her husband has slapped her in the last year. Furthermore, victimization and self-report surveys can tap only certain types of crimes, that is, crimes with identifiable victims who realize they have been victimized (in the case of victimization surveys) or crimes with identifiable offenders who recognize they have committed a crime and are willing to admit to it (in the case of self-report surveys). As a consequence, some types of crime cannot be measured with survey data.

For these reasons, researchers sometimes rely on indirect and unobtrusive measures to document the prevalence of crime. These have included hospital emergency room records (to measure violent crimes), insurance company records (to measure property crimes), and air or water pollution scans (to measure environmental crimes). Participant or covert observation has also yielded valuable evidence about the extent of illegal behavior by police, illicit substance use and sales, and illegal sex work, among other criminal acts.

All efforts to measure crime inevitably confront the problem of defining a crime. Someone – a victim, a police officer, an offender, a witness, or a researcher – must apply a legal definition to a complex set of circumstances and human relationships. The difficulties in measuring crime, then, are not simply technical or statistical, but conceptual – and unavoidable. Consequently, it is impossible to determine the total amount of crime in any jurisdiction at any time. One can, however, count and describe with some accuracy events that are crimes according to the law and that come to the attention of legal officials, as well as criminal events which people report to survey researchers. These estimates form the basis of knowledge about the prevalence and distribution of crime over time, place, and persons.

The Distribution of Crime

Is there more crime today than in the past? For reasons already outlined, this question has no simple answer. To the extent that contemporary criminal law defines a wider range of behaviors as criminal, there is likely to be more crime. However, if we focus on the types of behaviors that most societies have defined as criminal or deserving of punishment and control, we can draw some conclusions about the distribution of crime over time.

The most reliable historical evidence pertains to serious violent crimes – such as homicide, robbery, and serious assault – in Western Europe and North America. By drawing on coroners' and court records, historians have documented levels of serious violence in a range of European jurisdictions at various time points over the last few hundred years. Piecing these records together shows that homicide rates were exceptionally high in the thirteenth and fourteenth centuries and then declined through the end of the seventeenth century (Eisner, 2003). After this point the downward trend became more gradual, although it was punctuated by short-term waves after wars and other major social dislocations. By the mid-twentieth century, most Western European nations were experiencing some of the lowest levels of serious criminal violence in their histories.

In the United States, after reaching relatively high rates in the first half of the seventeenth century, rates of homicide declined and – unlike today – were comparable to rates in Europe up until the early nineteenth century. At that point, lethal criminal violence in the United States took an upturn due to social and economic problems associated with immigration and urbanization and to the growing availability of hand guns. However, by the late nineteenth century in the United States, rates of homicide and other serious violence stabilized and in the late 1920s began a decline that lasted until the mid-1960s (Roth, 2009).

The centuries-long decline in criminal violence in Europe and North America reversed somewhat in the last third of the twentieth century. By this point, relatively reliable and comparable data on a range of violent crimes became available for many countries, allowing conclusions about trends in nonlethal as well as lethal violence. Beginning in the late 1960s, violent crime rates rose and then plateaued in most Western nations, whereas in many Eastern European nations, these rates increased dramatically beginning in the late 1980s and continued their upward spiral with the breakup of the Soviet Union. Trends in violent crime in the United States were similar to those in some Western European nations, with peaks in the early 1980s and again in the early 1990s.

Since the mid-1990s, however, both official statistics and data from victimization surveys show a steady, and in some cases dramatic, decline in violent crime in the United States and other Western developed nations. In some cities in the United States, the reduction in serious violence was remarkable. For example, in New York City there were over 2200 homicides known to police in 1990; in 2010 the number of homicides had dropped to just over 530. Similar declines occurred in aggravated assault and robbery and, to somewhat lesser extent, rape in New York City and many other places in the United States.

This 'crime drop,' as it has been termed by criminologists, has characterized a number of other countries as well. For example, official statistics and victimization survey data in Canada, England and Wales, and all of the Scandinavian countries show similar reductions since the early or mid-1990s. Similar to the pattern in the United States, in 2010 Canada's homicide rate reached its lowest level in 44 years, and the number of victims (554) was 25% smaller than the number of victims in the peak year of 1991.

It seems unlikely that the centuries-long decline in violent crime has been accompanied by a similar drop in property crime, simply because the growth in consumer goods and individual wealth means opportunities for property crime have increased greatly over the last few hundred years. However, reliable historical data on property crime do not exist. Before the advent of victimization surveys, even tracking more recent trends in property crime with any reliability was difficult because of the limitations of official statistics discussed previously. Since the late 1970s, however, both victimization survey data and official statistics show a drop in property crime (i.e., burglary, larceny, and motor vehicle theft) in the United States, paralleling the drop in violent crime. With the development of the ICVS, it has also been possible to estimate trends in property crimes in the developed countries where the survey has been carried out multiple times. These data show a substantial downturn since the mid-1990s in a range of property crimes, particularly burglary and motor

vehicle thefts, a finding paralleled by statistics from other sources (Harrendorf et al., 2010).

Are some societies more crime prone than others? Cautious comparisons can be made of direct contact, predatory crimes of individuals using victimization surveys and the data on homicides discussed earlier. According to these, rates of serious violent crime vary greatly, even across relatively similar societies, although these differences have diminished somewhat with the recent crime drop. For example, in 2009 Finland's homicide rate (2.3 per 100 000 population) was almost four times greater than Norway's (0.6), Belgium's rate (1.7) was more than three times greater than Austria's (0.5), the Russian Federation's rate (11.8) was more than twice as large as the Ukraine's (4.8), and the rate in the United States (5.0) was almost three times greater than Canada's (1.8) (UNODC, 2011). Among advanced capitalist nations, homicide rates are higher in the United States, Canada, and Australia. These rates are exceeded by those in Latin and South America and in Africa, where homicide rates are about twice as high as the global average. In contrast, homicide rates in Asia and Europe are about half the global average. The distribution of other serious violent crimes tends to follow a similar pattern.

Homicides and other serious violent crimes make up only a small part of the total crime rate, however. In all nations, most crimes are crimes against property. Traditionally, burglary and larceny (or theft) have been the major property crimes; of the 30 countries participating in the most recent wave of the ICVS (which does not include the United States), England and Wales had one of the highest rates of property crime victimization and Japan had among the lowest. Since the turn of the twenty-first century, consumer frauds of various sorts have become more widespread in many countries and may overtake more traditional property crimes, in large part because of the growth of the Internet, increased use of credit and debit cards, and other changes in banking and purchasing practices.

Although there is only limited research and almost no systematic data on national differences in white-collar, corporate, and political crimes – for example, theft by corporations, environmental destruction, or illicit violence by state agencies – these appear to be more common in developing countries. In part, this results from limited resources for enforcement and the ability of offenders to avoid prosecution. For similar reasons, activities of criminal organizations, such as trafficking in drugs, people and weapons, money laundering, and cross-border terrorism, may also be more prevalent in countries with emerging or transitional economies.

As this suggests, the economic and political organization of a country produces particular patterns of crime. For example, although the pre-1990s socialist countries of Eastern Europe had low rates of violent and predatory street crimes, other behaviors – worker sabotage, theft of nationalized property, bribery, and black market activity – were relatively widespread; and violent crime rates are higher in nations with greater inequalities in income and wealth and with fewer restrictions on the ownership of firearms.

Many of the social processes, conditions, or policies commonly thought to affect crime rates often do not do so, or affect only some types of crime. Urbanization and unemployment, for example, are not consistently accompanied by higher rates of violent crime, although they are associated with higher rates of property crime. Conversely, some social policies intended to decrease crime do not. For example, there is little evidence that the death penalty deters homicides or that high imprisonment rates are a cost-effective way to reduce crime. During the twenty-first century, the social process likely to have the greatest impact on the prevalence and distribution of crime – from street crimes to crimes of transnational corporations – is globalization (Natarajan, 2011). While there is nothing new about crime crossing national borders, unique opportunities for truly global crime have emerged with the rise of global capitalism, the expansion of electronic information networks, and widespread immigration and travel between countries and continents.

Just as certain national or global conditions or processes create opportunities and motivations for crime, so do more localized environments within countries. At the level of cities and neighborhoods, the prevalence and distribution of crime are not random. However, national context and culture affect the relationships between community characteristics and crime in complex ways, and so some of the relationships observed in the United States hold in other countries whereas others do not. Within the United States, for example, larger urban areas tend to have higher rates of violent and property crimes, but this association does not exist in all countries, even Western democracies; some of the least urbanized areas in Canada have among the highest homicide rates in the country.

Nevertheless, decades of criminological research have identified a set of characteristics consistently associated with high rates of serious, direct-contact predatory crime within the United States (i.e., at the state, urban, and neighborhood levels) and many other Western countries, including economic deprivation (e.g., absolute and relative deprivation), family disruption (e.g., divorce rates, single-parent families), and the geographic isolation of marginalized racial minority groups (Sampson, 2011). Some features of urban areas often assumed to be 'criminogenic' – such as population density, unemployment, or residential mobility – are related to some types of crimes but not others. Other characteristics are linked to crime, but in the direction opposite of what many people believe. For example, immigration is associated with lower levels of crime in many urban areas of the United States.

Neighborhoods and other local communities do not exist in vacuums. Consequently, the political, economic, and social processes and policies that allow local criminogenic conditions to develop and persist are generated by governments and major economic actors, which are in turn affected by national and global factors. Understanding the prevalence and distribution of crime at the local level therefore requires attention to multiple levels of analysis and processes that are at times quite distant from specific communities, even if their consequences are not.

Just as some communities are characterized by what are termed criminogenic structures and cultures, so are some organizations, both public and private (Friedrichs, 2010). Corporate crimes, state crimes, and other organizational crimes are also patterned by internal organizational characteristics (such as normative culture, lines of communication and authority, salary and other incentive structures) and external environmental characteristics (such as characteristics of competitors and regulatory agencies). The complexity of many

organizations and economic sectors – such as the financial sector – can also increase opportunities for criminal activity as well as decrease the ability of regulatory organizations to detect it. As a result, corporate and other organizational crimes proliferate and are more costly than all other types of crime combined. For example, the global financial crisis that began in 2008, which was prompted in part by fraudulent accounting and lending practices by major United States banks and financial firms, is estimated to have cost US taxpayers almost one-hundred billion dollars in 2009 alone.

State crime – that is, acts by state agents in the course of their duties that violate national or international criminal laws – can also be extremely costly, both financially and in human lives. Weak states run by autocrats can permit certain groups to expropriate public and private citizen's wealth and other resources for personal gain. At their worst, state crimes – from the torture of political prisoners to genocide – violate fundamental human rights in vastly more serious ways than any other type of crime. In recognition of this, in 2002 the International Criminal Court was established to prosecute individuals for genocide, crimes against humanity, war crimes, and the crime of aggression. By mid-2012 more than 120 countries were parties to the Statute of the Court.

Research on the prevalence of crime often focuses on the individual level and the characteristics of those who engage in criminal behavior and asks if some people are more likely to engage in crime than others. Once again, the heterogeneity of the acts comprising crime demands a cautious answer. Serious, direct-contact predatory crimes are the domain of young, economically disadvantaged males in virtually all nations. White-collar, corporate, and political crimes are also male dominated, though these are, for obvious reasons, rarely the preserve of the young and the poor. It is largely for victimless crimes – such as prostitution or drug use – and some low-level property crimes – such as minor thefts and frauds – that women's involvement at times rises to the level of men's.

Although some criminologists argue that there are few individual characteristics that consistently distinguish between people who engage in crime and people who do not, others disagree. The latter have highlighted psychological traits or characteristics – such as low self-control, impulsivity, psychopathy, or certain neurocognitive deficits – that are associated with many types of violent and other criminal behaviors. An important limitation of this research is that it focuses on a range of behaviors and actors that are unlikely to be representative of the entire population of crimes and offenders.

Explaining the Distribution of Crime

The decline in serious violence in Western Europe from the thirteenth century (and somewhat later in North America) up until the mid-twentieth century has been attributed to a number of different factors (Eisner, 2003). Certainly, the reduction in deaths due to homicide is at least partly the result of advances in medical technology, such that assaults are less lethal now than in the past. However, this cannot explain an apparently substantial decrease in nonlethal violence, which typically is ascribed to a set of broad cultural and social processes. Increases in self-control – which are captured by the concept of the 'civilizing process' (Elias, 1978) – as well as in informal and formal social controls, are central to explanations that focus on fundamental cultural changes.

While explanations that focus on large-scale cultural change can account for the centuries-long decline in violence, explanations of shorter-term deviations from that decline tend to focus on more structural factors specific to certain times and places. The widespread increases in both violent and nonviolent crime that began in the 1960s have yet to be fully explained, but are due at least in part to demographic changes (i.e., a growth in the population aged 15–29, the peak years for crime), shifts in routine daily activities away from the home, high rates of marital disruption, and growing economic inequality. In the United States, the emergence and expansion of crack cocaine sales and use also helped drive crime rates up in the late 1980s and early 1990s.

What is less clear is why both violent and nonviolent crime rates dropped not just in the United States but in a number of other Western developed countries, beginning in the mid-1990s. Many explanations have focused on factors specific to the United States (such as rising imprisonment rates, changes in policing practices, declines in crack cocaine use, etc.), but these cannot account for the crime drop in other countries that did not experience these changes. Perhaps there was a reemergence of the broader cultural shifts responsible for the long-term decline in crime, combined with a set of other, more country-specific factors (such as a reduction in childhood exposure to lead). Criminologists continue to explore a number of possibilities but as yet have reached no consensus about the driving forces behind the downward trend that continued throughout the first decade of the twenty-first century (Baumer, 2011).

Most explanations for variations in crime rates among and within countries or between individuals feature one of three mechanisms: motivations, controls, or opportunities. Motivational explanations focus on social conditions or processes that generate criminal tendencies in societies, groups, or individuals. Criminal motivations have their roots both in social structure – institutional patterns and processes that organize interactions and relationships – and in culture – accumulated traditions and shared experiences that shape the values of a society or group. Economic inequality is a structural characteristic strongly associated with high crime rates and is thought to increase motivations for crime by fostering alienation and hopelessness. Norms that advocate criminal acts under certain conditions are a cultural factor that may motivate criminal behavior.

Control explanations focus on social conditions and processes that weaken internal and external, informal and formal controls on crime. Both structural and cultural factors can affect social control. Crime tends to be lower in societies with stable families and integrated community networks because these strengthen controls on criminal behavior. Crime tends to be higher where group norms excuse or tolerate crimes against certain victims (e.g., against spouses, organizations, sexual minorities, etc.) because these cultural features weaken controls.

While motivational and control perspectives are primarily concerned with potential offenders, opportunity explanations

focus on the circumstances in which crimes occur. Situations in which a motivated offender encounters a suitable target in the absence of capable guardians create opportunities for crime (Cohen and Felson, 1979). Patterns of routine daily activity that take people away from home foster opportunities for burglary, mugging, and auto theft. Criminal opportunities are also greater where small consumer items are mass produced.

Some general theories of the prevalence or distribution of crime have been widely tested and command substantial support. At the same time, there is considerable debate over whether a general explanation of crime is possible or even desirable, given the heterogeneity of crime. Nevertheless, some generalizations about the sociocultural contexts that appear to discourage direct contact, predatory crimes can be drawn. These are characterized by (1) strong systems of informal social control reinforced by highly consensual normative systems; (2) strong networks of communal obligation and mutual interdependence; and (3) cultural orientations that discourage cynicism about mainstream social institutions. Societies with these characteristics tend to be organized around collectivist rather than individualistic principles. These principles in turn are reflected in political, economic, and legal ideologies and practices and generate similar characteristics at the local or community level. These generalizations are grounded in research on a limited range of criminal behaviors, however. Too little is known to develop similar generalizations about many types of crime, such as those by transnational corporations and national governments, although this is changing. Until these types of crime can be more systematically documented, understanding of the prevalence and distribution of crime will be incomplete.

See also: Crime and Class; Crime and Gender; Crime, Geography of; Crime, Race and Ethnicity; Crime, Sociology of; Crime: Organized; Crime: Whitecollar; Data Bases and Statistical Systems: Crime; Delinquency, Sociology of; Prevention of Crime and Delinquency; Rape and Sexual Coercion; Youth Gangs.

Bibliography

Baumer, Eric, 2011. Crime trends. In: Tonry, Michael (Ed.), The Oxford Handbook of Crime and Criminal Justice. Oxford University Press, New York, pp. 26–59.

Cohen, Lawrence, Felson, Marcus, 1979. Social change and crime rate trends: a routine activity approach. American Sociological Review 44, 588–608.

Elias, Norbert, 1978. The Civilizing Process, vols. 1–2. Oxford University Press, Oxford.

Eisner, Manuel, 2003. Long-term historical trends in violent crime. In: Tonry, Michael (Ed.), 2003. Crime and Justice: A Review of Research, vol. 30. University of Chicago Press, Chicago IL, pp. 83–142.

Friedrichs, David O., 2010. Trusted Criminals: White Collar Crime in Contemporary Society. Wadsworth, Belmont, CA.

Harrendorf, Stefan, Heiskanen, Markku, Malby, Steven (Eds.), 2010. International Statistics on Crime and Justice. European Institute for Crime Prevention and Control, Helsinki.

Hillyard, Paddy, Pantazis, Christine, Gordon, Dave, Tombs, Steve (Eds.), 2004. Beyond Criminology: Taking Harm Seriously. Pluto Press, London.

Natarajan, Mangai (Ed.), 2011. International Crime and Justice. Cambridge University Press, Cambridge.

Roth, Randolph, 2009. American Homicide. The Belknap Press of Harvard University Press, Cambridge, MA.

Sampson, Robert J., 2011. Community. In: Wilson, James Q., Petersilia, Joan (Eds.), Crime and Public Policy. Oxford University Press, New York, pp. 210–235.

United Nations Office on Drugs and Crime, 2011. Global Study on Homicide 2011: Trends, Contexts, Data. UNODC, Vienna.

van Dijk, Jan, van Kesteren, John, Smit, Paul, 2007. Criminal Victimization in International Perspective. Eleven International Publishing, The Haag, The Netherlands.

Zedner, Lucia, 2007. Pre-crime and post-criminology? Theoretical Criminology 11, 261–281.

Crime: Organized

Andrew V Papachristos and Sandy Y Zhao, Yale University, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Steffensmeier, F. Martens, volume 5, pp. 2928–2934, © 2001, Elsevier Ltd.

Abstract

The problem of organized crime has captured the attention of scholarly research, federal legislation, popular culture, and the imagination of the public for decades. The definition, conceptualization, activities, and overall analysis of organized crime have also undergone important developments in sociology, political science, economics, law, and other interdisciplinary fields. This article offers an overview of this body of research with descriptions of the current understandings of (1) what constitutes organized crime/organized criminal groups and their activities and (2) different avenues of analysis and approaches to research on organized crime.

Introduction

Definition

The idea of organized crime generally refers to crime that is organized (i.e., committed by groups of more than two persons) as well as to various other typologies and forms of criminal groups such as street gangs or youth gangs (Varese, 2001). That is, organized crime refers to both types of activities as well as types of groups. Organized crime groups are typically characterized by

1. a formalized structure with some sort of governance system that is maintained by some combination of force, the use of violence, and appeals to personal and kinship ties;
2. monopolization of the production and distribution of some commodity, usually illegal goods and/or services;
3. resilience in time, with development, expansion, or reorganization, sometimes into legitimate (noncriminal) sectors of politics and/or the economy; and
4. a sense of collective identification as a group, society, or organizational entity (Gambetta, 1993; Schelling, 1984; Lupsha, 1995; Ayling, 2011).

The most commonly portrayed (and studied) organized crime groups are often generally termed 'mafia,' a term originally applied to describe Italian criminal groups, but has since been challenged as a misnomer and applied cautiously (Varese, 2001). However controversial or inadequate, the mafia is distinguished from organized crime as a whole in its identity as a separate relatively autonomous entity, such as a particular crime group like the Russian *vory v zakone*, 'relatively autonomous' in its stance with respect to legitimate and illegitimate criminal organizations. Most of empirical research (not to mention, popular culture) considers the mafia to be the 'archetype' of the organized crime group. But in fact, we know very little about other types of organized crime groups. In this sense, while mafias are integral to the study of organized crime, further investigation is needed to better understand organized crime as a whole.

Types of Activities

Mafias and criminal groups defined as organized crime have historically been involved in criminal activities such as racketeering, money laundering, bootlegging, and drug trafficking. In recent decades, there has been increased concern about the involvement of organized crime groups in human smuggling and sex trafficking (Ayling, 2011) in addition to ongoing issues with drugs and arms trafficking. In general, organized crime groups engage in providing illicit goods and services, and are often able to infiltrate legitimate businesses and institutions.

Criminal activities such as drug trafficking are characterized as 'organized' primarily due to more sophisticated levels of logistics and delegation of tasks as opposed to, say, street crimes such as robberies that often involve little planning and organizational wherewithal. Drug trafficking at the level of organized crime is therefore distinguished in complexity from street corner drug sales by local gangs because it involves moving large quantities of narcotics often across state and international borders – a task that requires both organizational capacity as well as connections to noncriminal entities. The scale of such drug trafficking by organized crime groups requires more planning and organizational structure, similar to that of a legitimate corporation.

Roots of Organized Crime

Social Disorganization

The most common theories on the origins of organized crime rely on three main themes: social disorganization, existence of weak government, and functional theories. First, social disorganization refers generally the inability of a community to realize the common values and goals of its residents to maintain effective social controls (see Sampson, 2012: p. 37, for a review). Typically, social disorganization arises in communities with high population turnover, a lack of resources, and a greater level of population heterogeneity, and more broadly from consequences of industrialization, urbanization, modernization, and the growth of free-market capitalism in democracy, in parts of the world. The original Chicago School interpretation of social disorganization held that such shifts in communities diminished the functioning of social institutions such as families and community organizations that would otherwise monitor resident behavior (e.g., Shaw and McKay, 1942).

The consequences of social disorganization include increased poverty, social injustice, poor health conditions, and violence. In particular, conditions of social disorganization often stir the development of criminal subcultures and groups of all types as a means not simply to deal with harsh living conditions but also to fulfill needs otherwise left unmet (Steffensmeier and Martens, 2001). Much organized crime, particularly in the United States, arose from the impetus to seek financial prominence and elevated status in the wider community (not necessarily criminal). Indeed, sociologist Robert K. Merton uses none other than the notorious gangster Al Capone as a prime example of how organized crime can be an adaptation to structural conditions and an illegal means to the American Dream (see also Bell, 1953).

Weak Government

The second main theory of the origins of organized crime is the concept of weak government. Criminal organizations are more likely to emerge and develop where government is seen as corrupt, unstable, or insufficient/inadequate over territory under its control. Organized crime groups arise to ostensibly fill in this governance vacuum, sometimes addressing real needs for a certain type of stability, but more often than not exploiting this power gap to advance organizational goals or status. Organized crime in many countries, as in Sicily and during the Stalinist period in the former Soviet Union, has stemmed from rule by authoritarian or oppressive regimes. Such governments generate conditions that enable groups to form for self-protection and provision of quasi-governmental services (Steffensmeier and Martens, 2001).

Conditions such as these, in turn, give rise to a system of beliefs and attitudes toward society and government (distrust, clan mentality, inclination to use violence and vigilante tactics, corruption of law and influential figures in politics, exploitation and use of local – and in the case of ethnic criminal groups, members of the ethnic group – residents to manipulate influence), all of which is conducive to the formulation and growth of organized crime (Cressey, 1969; Pennsylvania Crime Commission, 1991; Steffensmeier, 1995; Steffensmeier and Martens, 2001). For instance, criminal syndicates in Russia after the Bolshevik Revolution (Varese, 2001; Solomon, 1996) and former Soviet Republics such as Ukraine in the 1990s (Serio, 2008; Humphrey, 2002) after the collapse of the Soviet Union arguably arose from instability, rising corruption, and a demand for access to basic goods (bread, alcohol, meat) which otherwise were in short supply or available only to the political elites. In cases such as these, and many others, organized crime groups exploit the opportunity to insert themselves within the social organization of a community.

Without a strong or functioning government, organized crime groups can posture themselves as *de facto* leaders/rulers, i.e., preside over kangaroo courts, mete out justice, and provide basic services, but at a cost.

Functional Theories

Organized crime is also thought to stem from a number of 'functional' reasons in the sense that its emergence serves a basic function or multiple functions – positive or negative – for some segment in society (Cressey, 1969). Such functions can include providing illicit incomes in communities lacking legitimate employment opportunities, entry into political and economic spheres for disenfranchised populations, and social and financial aid as a form of off-the-books welfare systems (Steffensmeier and Martens, 2001).

Several scholars contend that much of the illegal activity conducted by member of organized criminal groups is connected to, and often dependent on in a practically symbiotic sense, legitimate (legal) organizations, institutions, and actors (Steffensmeier, 1986; Steffensmeier and Martens, 2001; National Research Council, 1999). That is, organized criminal activities often exist precisely because they require connections with the legitimate world not only to market illicit services and goods but also to maintain connections to public officials who can help organized criminals avoid detection and apprehension. One of the defining characteristics of organized crime is its ability to forge such connections and interdependencies, interacting with and infiltrating mainstream social institutions through networks (see more in Section Organized Crime as Networks, also Papachristos and Smith, 2013).

In this sense, organized crime is not so much a separate or entirely autonomous group apart from mainstream society. Rather, organized crime is integrated into many social and political components that influence the effective control and functioning of greater society. These deep-seated issues are also evidenced by organized crime's role in certain economic sectors such as the prominent issue of labor racketeering into organized labor unions in mid-twentieth-century America saw crime groups like the Cosa Nostra exerting influence on a movement with major impact on the U.S. economy (Jacobs, 2006).

The 'Gray Economy'

One of the more studied activities of organized crime is its role and functioning in illegal markets and economics which often highlight the role of organized crime in a community as well as its relationship with legitimate social institutions. In terms of organized crime and quasi-legal activities in general, this is referred to as the 'gray economy' or the 'shadow market.' The gray or 'shadow' economy is typically defined as market economic activity that goes unreported to the government (e.g., Feige, 1989). Such economic activities are those that may not necessarily be illegal, but are 'under the radar' of typical economic oversight and measurement (i.e., GDP). The provision and exchange of goods and services by criminal groups can be considered to fall into this category of economic analysis, which reiterates the interconnectivity of organized crime and legitimate institutions and actors in society. For example, in the post-Soviet period in the 1990s, criminal groups in former Soviet states such as Ukraine and regions of eastern Russia regularly participated in an underground economy, providing everything from food products to illegal substances, effectively evading enforcement from governments that were themselves economically unstable and corrupt (Humphrey, 2002). Common organized crime gray market activities include tax evasion (Ayres, 1996), money laundering, arms smuggling (Nordstrom, 2000), and many other illegal to borderline illegal activities that have escaped government monitoring or legitimate authorization (see Bucklin, 1993).

Scope and Structure of Organized Crime

The scope and structure of organized crime is generally understood through four different conceptual paradigms: as ethnic subcultures, as hierarchies and bureaucracies, as patron–client structures, and as networks. These perspectives in understanding organized crime are a condensed summary of the relatively wide variation in the composition, structure, and development of criminal groups whose relations to mainstream society and broader organized crime syndicates are becoming increasingly complex due to resilience and varying degrees of institutional evolution (Ayling, 2011).

Organized Crime as Ethnic Subcultures

Organized crime can be understood in terms of different groups acting as ethnic subcultures or, at least, as being dominated by distinct ethnic subcultures within the larger society in which a particular organized crime group operates. For instance, the Mafia in the United States is often depicted and studied as a set of distinct ethnic groups, but especially Italian and Sicilian (Gambetta, 1993, and others). As mentioned above, activities of organized criminal groups have at times been connected with an ethnic group's need to survive and thrive in a foreign and changing environment. Particular ethnicities and cultural practices provide a repertoire of already existing traditions, values, rituals, and common understandings that can be adapted or manipulated in establishing and promoting activities of organized crime. Within the subgroup, especially a disenfranchised and marginalized subgroup, one would presumably encounter resistance or distrust of official authority and government, perceptions of inadequacy toward the law (i.e., 'legal cynicism'), in-group trust, and insularity, which are all conducive to the reinforcement and development of the overarching 'culture' of organized crime (Steffensmeier, 1995; Steffensmeier and Martens, 2001).

Organized Crime as (Ethnic) Hierarchies/Bureaucracies

The structure of organized crime groups is often conceptualized as hierarchical or bureaucratic systems, where decisions are made top–down, to the benefit of the organization and its survival. In scholarship/research on the Italian Mafia and the Russian Mafia, this nearly institutionalized structure can be seen, particularly when focusing on the criminal activities within the 'gray market' of the economy (see Gambetta, 1993; Varese, 2001). The main idea is that actors involved in organized crime and illicit goods and services follow a similar script of rational behavior indicated by mainstream/legal economic actors and institutions. To illustrate this point further, organized crime has been analyzed as an 'enterprise,' (see Steffensmeier and Martens, 2001) especially in activities such as racketeering.

In racketeering and other entrepreneurial activities, organized crime can be seen as a 'business' that deviates in its legality from commonly understood business activities but shares much of the ideology and ethos of legal businesses (Smith, 1980; Steffensmeier and Martens, 2001). In regarding organized crime as a business, however illegitimate, focus is directed again to the functions and benefits to the organization itself, and the necessary arrangements and relationships it has with legitimate businesses. Chinatown gangs are a prime example: they often rely on extortion of Chinese businesses in Chinatown for revenue and establishment of authority (Chin, 1996). Chinatown organized crime groups therefore infiltrate legitimate businesses, providing policing and loans, but at a high cost of exploitation. Bootlegging during Prohibition was so large and extensive that many operations ran with businesslike precision, orchestrating international production and distribution systems, operating local speakeasies, and literally paying hundreds of employees, including brewers, truck drivers, hired goons, crooked cops, and even politicians (Haller, 1990; Papachristos and Smith, 2013). In short, many organized criminal activities require the sort of organizational structure and business acumen often valued in legitimate business ventures (Merton, 1938).

Often organized criminal activities are so similar to those of legitimate businesses that distinguishing illegal groups becomes difficult at the expense of enforcement and legal prosecution. For instance, in China, such ambiguities in definition make it difficult for legal authorities to properly understand what they are up against, with most assuming that such groups are hierarchical, when in fact they are widely varied from market-based enterprises to personal network systems (Xia, 2006; Chin and Zhang, 2008).

Organized Crime as Patron–Client Structures

Similarly, organized crime can be structurally understood in terms of a patron–client relationship. Patron–client relationships operate in such a way as power and access to particular resources rests in the hands of a central figure (the patron) who manages access to resources and brokers relationships. Those seeking access to said resources (the clients) must negotiate with the patron, often indebting themselves or otherwise entering into a subordinate position in the network, offering loyalty and allegiance in exchange for the patron's attention and resources (see Steffensmeier and Martens, 2001). In such relationships, connections are invaluable for access to illegal activities (perhaps through legal or quasi-legal methods) (Albini, 1971; Ianni and Reuss-Ianni, 1972; see Steffensmeier and Martens, 2001). Key individuals in positions of power and prominence are approached for favors and other services such as protection, important connections, and resources (Albini, 1971; Steffensmeier and Martens, 2001). More recently, research is redirecting the focus from patron–client relationships, where ties are already complex to a focus on the interconnectivities of networks dictated not by a strict top–down hierarchy but rather by a symbiosis of social circles and exchange of influence (Papachristos and Smith, 2013).

Organized Crime as Networks

An emergent field in understanding organized crime is applying social network analysis to the study of organized crime (see Morselli, 2003). Understanding organized crime syndicates as networks allows research to expand and transcend traditional perspectives on the structure and activities of organized crime groups. In fact, the application of social networks

to organized criminal groups is arguably intuitive: such groups, varied as they are along social, ethnic, and historical lines, are characterized by network relationships within and without the group. The Cosa Nostra relies on connections, affiliations, and exchanges as much as Chinese criminal syndicates rely on 'guanxi,' the Eastern conceptual equivalent (McIllwain, 2000). Similarly, the transcendental nature of using social networks in analyzing organized crime groups highlights and bridges entrepreneurial elements and activities within the criminal group with noncriminal institutions and actors such as legitimate businesses and politicians. The New York Gambino crime family (among others) had intragroup and intergroup connections that enabled upward mobility from within the family ranks as well as development of the criminal enterprise to infiltrate legitimate businesses (Morselli, 2003). Another example is the reliance on personal networks of family, friends, and acquaintances that allow criminal actors to place themselves in strategic positions of influence and control over resources, licit and illicit, in order to both expand and strengthen a criminal group's hold over a certain region or community such as with Hong Kong Triads in U.S. Chinatowns and the Russian 'Mafia' in the U.S. (Chin, 1996; McIllwain, 2000; Morselli, 2003; Finckenauer and Waring, 1998).

Recent research has also focused on the dual concepts of 'embeddedness' and 'multiplexity' (Granovetter, 1985; Papachristos and Smith, 2013) within social network analysis and organized crime research. The placement of different social relationships in multiple types of networks is analytically combined to better illustrate and understand that the social circles in which individuals operate and are situated often not just involve multiple realms of politics and society but also interact and influence one another in a variety of possible ways (Papachristos and Smith, 2013). In short, emphasis on interdependencies of actors in criminal and noncriminal networks provides a helpful method to thinking about and researching organized crime that reaches beyond classical conceptualizations of structure and criminal activity in such groups.

Organized Crime in the United States

Much of organized crime research and theory has focused on organized crime in the United States, particularly mafias of various ethnic subgroups located in the country. In recent years, as organized crime has grown along with new advancements in technology and communication, more and more research is broadening to include previously overlooked groups in their apparently nascent stages and the issue of transnational organized crime. As previously stated, not only are such organizations expanding but their activities and practices are increasingly merging internally and externally with mainstream social, political, and economic activities (Steffensmeier, 1995; see Steffensmeier and Martens, 2001). Organized crime as treated in this article refers only to large-scale enterprises with highly organized structures and sophisticated distribution and application of resources for illegal activities. While biker gangs and drug cartels are part of the larger umbrella of organized crime, discussion of these criminal groups in the U.S. is outside the scope of our analysis.

Mafia in the United States

Organized crime in the U.S. has historically centered on the early immigrants to the country: Irish, Eastern Europeans, and Italians. The Italian-American Cosa Nostra has dominated the conception and study of organized crime in scholarship and popular media, but increasingly more research is focusing on Chinese and Hispanic involvement in organized crime, particularly as immigration (and controversial immigration policies) has increased as well.

In addition to the U.S. Mafia associated with the Cosa Nostra, there have also been studies on Chinatown gangs (from Hong Kong Triads and 'tongs'; see Chin, 1996) and Asian-American organized crime as a whole, mostly located in Los Angeles, San Francisco, and New York, where Asian populations (East Asian, Southeast Asian) are typically high. These criminal groups, along with broader examination of organized crime in other subgroups (bikers, Blacks, Colombians, Dominicans, Jamaicans, Mexicans, Syrians, Lebanese, and Russians) previously underresearched and seen as 'emergent' subgroups, are gaining prominence in the literature (see Finckenauer and Waring, 1998). These groups are typically involved in trafficking illicit goods, especially drugs, but in recent years, human smuggling and sex trafficking have also come to greater attention, particularly as these criminal activities have reached the global level.

Transnational Organized Crime and Organized Crime in Other Countries

China (Mainland, Hong Kong, and Taiwan)

While research in Asian organized crime as a whole has increased, most of it has focused exclusively on human smuggling and drug trafficking (Chin and Zhang, 2008). Aside from a few exceptions (Chin, 1996, 1999 and others), most studies of Chinese organized crime have also largely depended upon secondary sources, particularly analysis of media coverage and government reaction/response. Empirical research that has relied on more direct contact with victims and perpetrators of criminal activity has dispelled traditional assumptions that single sophisticated organized criminal groups – such as the Hong Kong Triads – were responsible for the majority of human smuggling (Chin and Kelly, 1997). Most smuggling operations are conducted by individuals known as 'snakeheads,' who are locally or often familiarly connected to the illegal immigrants (Chin and Zhang, 2008).

Other important headway made in Chinese organized crime research is focus on the often symbiotic relationship between the Chinese government and criminal organizations. The Chinese government itself has not officially recognized the problem/existence of underworld organizations originating and stretching from Mainland China to Taiwan to Hong Kong and Macao (Chin and Godson, 2006). The current political system and situation in China are impediments to deeper research and acknowledgment of real and widespread issues pertaining to such criminal groups. In the meantime, groups from the Mainland are finding it increasingly easy to gain legal protection and legitimacy for their illegal conduct from the government (Chin and Zhang, 2008).

Russia

Deriving its historical roots from criminal organizations in Stalinist prisons, the Russian Mafia grew from multiethnic roots to a transnational criminal network. The structure of the organization has arguably remained hierarchical, but is now more flexible (Varese, 2001). The Russian Mafia is not a single organization, but a combination of criminal leaders with historical origins in Soviet criminal groups (Varese, 2001). Because of the very historical and (anti) political origins of the Russian Mafia, there are legacies of the Mafia's foothold in the government and the economy. Despite the lack of unified identity in the proliferation of the Russian Mafia (with a composite of multiple subgroups of different ethnicities and spreading to countries now independent from the former Soviet Union), the Mafia itself has lasted for many years and has the advantage of an international network.

Italy

Most research on organized crime in Italy has centered on Sicily. In extant research, Sicilian organized crime is characterized as a formalized, structured group, relying on codes of communication that enable effective signaling within and outside those groups (Gambetta, 1993, 2009). Communication among criminals is necessary not just in order for a heist or crime to go smoothly, or even so much to promote mutual trust, but rather to be able to signal their own 'criminal credentials' (Gambetta, 2009). Intuitively, the more structured a group is, the more structured its signaling system. Similar to research on organized crime in China and Russia (and the former Soviet Republics), research on organized crime in Italy emphasizes complex in- and out-group interactions and ties with legitimate institutions, which appear to be the driving force behind the continuity of organized crime.

Other Countries

The extent of organized crime as a global, transnational problem is evidenced in its appearance in various forms in many different countries, spanning continents. In addition to the Chinese organized crime groups mentioned above, the Yakuza are the main organized criminal entity in Japan. The Yakuza have a long history and comparatively formal structure, with recognizable gang regalia (tattoos, hairstyles, and clothing) and participation in gambling, prostitution, and trafficking (Hill, 2003). Other criminal groups can be found in South America, namely Brazil; Central America and the Caribbean (Mexican drug cartels, gangs in Jamaica); and the Middle East (Israel) (Manwaring, 2007; Siegel, 2003).

See also: Communities and Crime; Control Theories and Crime; Crime, Geography of; Crime, Race and Ethnicity; Crime, Sociology of; Crime: Knowledge about and Prevalence; Data Bases and Statistical Systems: Crime and Delinquency; Drug Control: International Aspects; Economics and Crime; Gangs, Sociology of; Mafias, Criminology of; Prevention of Crime and Delinquency; Transnational Organized Crime.

Bibliography

Albini, J.L., 1971. American Mafia: Genesis of a Legend. Appleton-Century-Crofts, Inc., New York.
Arlacchi, P., 1986. Mafia Business: The Mafia Ethic and Spirit of Capitalism. Biddles, Guildford, UK.
Ayling, J., 2011. Gang change and evolutionary theory. Crime, Law and Social Change 56, 1–26.
Ayres, E., 1996. The expanding shadow economy. World Watch 9 (4) (cover article).
Bell, D., 1953. Crime as an American way of life. The Antioch Review 13 (2), 131–154.
Bucklin, L.P., 1993. Modeling the international gray market for public policy decisions. International Journal of Research in Marketing 10 (4), 387–405.
Chin, K., 1996. Chinatown Gangs: Extortion, Enterprise and Ethnicity. Oxford University Press, New York.
Chin, K., 1999. Smuggled Chinese: Clandestine Immigration to the United States. Temple University Press, Philadelphia.
Chin, K., 2003. Heijin: Organized Crime, Business, and Politics in Taiwan. M.E. Sharpe, Armonk, NY.
Chin, K., Godson, R., 2006. Organized crime and the political-criminal nexus in China. Trends in Organized Crime 9, 5–44.
Chin, K., Kelly, R., 1997. Human Snakes: Illegal Chinese Immigrants in the United States, Law and Social Science Program. National Science Foundation, Washington, DC.
Chin, K., Zhang, S.X., 2008. Snakeheads, mules, and protective umbrellas: a review of current research on Chinese organized crime. Crime, Law and Social Change 50, 177–195.
Cressey, D., 1969. Theft of the Nation. Harper and Row, New York.
Feige, E.L., 1989. The Underground Economies: Tax Evasion and Information Distortion. Cambridge University Press, Cambridge.
Feige, E.L., Ott, K. (Eds.), 1999. Underground Economies in Transition: Unrecorded Activity, Tax Evasion, Corruption and Organized Crime. Ashgate Publishing, Hants.
Finckenauer, J.O., Waring, E.J., 1998. Russian Mafia in America: Immigration, Culture, and Crime. Northeastern University Press, Boston.
Fiorentini, G., Peltzman, S. (Eds.), 1995. The Economics of Organised Crime. Cambridge University Press, Cambridge.
Friedman, R.I., 2000. Red Mafiya: How the Russian Mob Has Invaded America. Little, Brown & Co., Boston.
Gambetta, D., 1993. The Sicilian Mafia: The Business of Private Protection. Harvard University Press, Cambridge.
Gambetta, D., 2009. Codes of the Underworld: How Criminals Communicate. Princeton University Press, Princeton.
Granovetter, M., 1985. Economic action and social structure: the problem of embeddedness. American Journal of Sociology 91 (3), 481–510.
Hagedorn, J., 2002. Gangs and the informal economy. In: Huff, C.R. (Ed.), Gangs in America, third ed. Sage Publications, Thousand Oaks.
Haller, M.H., 1990. Illegal enterprise: a theoretical and historical interpretation. Criminology 28 (2), 207–235.
Hill, P., 2003. The Japanese Mafia: Yakuza, Law, and the State. Oxford University Press, Oxford.
Humphrey, C., 2002. The Unmaking of Soviet Life: Everyday Economies after Socialism. Cornell University Press, Ithaca.
Ianni, F.A.J., Reuss-Ianni, E., 1972. A Family Business: Kinship and Social Control in Organized Crime. Russell Sage Foundation, New York.
Jacobs, J.B., 2006. Mobsters, Unions, and Feds: The Mafia and the American Labor Movement. New York University Press, New York.
Lupsha, P., 1995. Organized crime. In: Bailey, W. (Ed.), The Encyclopedia of Police Science. Garland, New York, pp. 493–497.
McIllwain, J.S., 2000. Organized crime: a social network approach. Crime, Law and Social Change 32, 301–323.
Manwaring, M.G., 2007. A Contemporary Challenge to State Sovereignty: Gangs and Other Illicit Transnational Criminal Organizations in Central America, El Salvador, Mexico, Jamaica, and Brazil. U.S. Army War College, Carlisle, PA.
Maxson, C.L., Klein, M.W., 1995. Investigating gang structures. Journal of Gang Research 3, 33–40.
Merton, R., 1938. Social structure and anomie. American Sociological Review 3 (5), 672–682.
Morselli, C., 2003. Career opportunities and network-based privileges in the Cosa Nostra. Crime, Law and Social Change 39, 383–418.
National Research Council, 1999. Transnational Organized Crime, Commission on Behavioral and Social Sciences and Education. National Academy Press, Washington, DC.

Nordstrom, C., 2000. Shadows and sovereigns. Theory, Culture, and Society 17 (4), 35–54.
Paoli, L., 2002. The paradoxes of organized crime. Crime, Law and Social Change 37, 51–97.
Papachristos, A.V., Kirk, D.S., 2011. Cultural mechanisms and the persistence of neighborhood violence. American Journal of Sociology 116, 1190–1233.
Papachristos, A.V., Smith, C.M., 2013. The Embedded and Multiplex Nature of Al Capone.
Park, R.E., Burgess, E.W., McKenzie, R.D., 1925. The City. University of Chicago Press, Chicago.
Pennsylvania Crime Commission, 1991. The 1990 Report – Organized Crime in Pennsylvania: A Decade of Change, Pennsylvania Crime Commission. Commonwealth of Pennsylvania, Conshohocken, PA.
Sampson, R.J., 2012. Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago Press, Chicago, IL.
Schelling, T., 1984. Choice and Consequence. Harvard University Press, Cambridge.
Schulte-Bockholt, A., 2006. The Politics of Organized Crime and the Organized Crime of Politics: A Study in Criminal Power. Lexington Books, Lanham.
Serio, J., 2008. Investigating the Russian Mafia. Carolina Academic Press, Durham.
Shaw, C.R., McKay, H.D., 1942. Juvenile Delinquency in Urban Areas. University of Chicago Press, Chicago.
Siegel, D., 2003. Global Organized Crime: Trends and Developments. Springer, Netherlands.
Smith, D., 1980. Paragons, pariahs, and pirates: a spectrum-based theory of enterprise. Crime and Delinquency 26 (4), 358–386.
Solomon Jr, P.H., 1996. Soviet Criminal Justice Under Stalin. Cambridge University Press, Cambridge.
Steffensmeier, D., 1986. The Fence: In the Shadow of Two Worlds. Rowman and Littlefield, New Jersey.
Steffensmeier, D., 1995. Public policy agenda for combating organized crime. In: Barlow, H. (Ed.), Crime and Public Policy. Westview Press, Boulder, CO, pp. 269–295.
Steffensmeier, D., Martens, F., 2001. International Encyclopedia of the Social & Behavioral Sciences 5, 2928–2934.
Sullivan, J.P., 2006. Maras morphing: revisiting third generation gangs. Global Crime 7 (3/4), 487–504.
Van Calster, P., 2006. Re-visiting *Mr. Nice*: on organized crime as conversational interaction. Crime, Law and Social Change 45, 337–359.
Varese, F., 2001. The Russian Mafia: Private Protection in a New Market Economy. Oxford University Press, Oxford.
Xia, M., 2006. Assessing and explaining the resurgence of China's criminal underworld. Global Crime 7, 151–175.

Crime Policy in the United States

Candace McCoy, John Jay College and the Graduate Center, City University of New York, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The goal of anti-crime public policy is usually understood to be the interception of and/or prevention of criminal acts. At first impression, this statement is simple and straightforward. However, even a cursory overview of crime policy in operation presents a much more complicated picture.

Briefly, four topical areas illustrate the difficulty of determining how government can best respond to crime (and there are surely others not specifically covered in this article). These four categories of issues in crime policy can be labeled 'sociological, legal, governmental, and practical.' This article will consider each topical area in turn, offering examples to illustrate major issues and problems.

The goal of anti-crime public policy is usually understood to be the interception of and/or prevention of criminal acts. At first impression, this statement is simple and straightforward. However, even a cursory overview of crime policy in operation presents a much more complicated picture.

Briefly, four topical areas illustrate the difficulty of determining how government can best respond to crime (and there are surely others not specifically covered in this article). These four categories of issues in crime policy can be labeled 'sociological, legal, governmental, and practical.' The sociological approach regards crime as a matter of social deviance from the norm, and it raises two fundamental issues: (1) what is the definition of crime? and (2) what causes people to commit crimes? The state defines the crime and also develops policies and programs based on what the legislative and executive branches' beliefs about 'why people commit crimes' may be. The legal approach proceeds from a positivist assumption that crime is what the legislature says it is, so laws and their interpretation are its focus. Its fundamental threshold issue is how can the law (most importantly, the nation's Constitution) be written and applied so as to reduce crime while simultaneously providing the fullest measure of personal and political liberty? The governmental approach has as its goal the enforcement of this law, and its primary concern is which agencies of government – and perhaps private parties acting in partnership with them – can be mobilized most efficiently and effectively to achieve that goal? Practitioners in the field are generally most concerned with the fourth topical area, the 'practical,' which concerns the structure and operations of governmental agencies of criminal justice. The fundamental underlying question for them is: what works to reduce crime, and how can the agencies of criminal justice be improved?

This article will consider each topical area in turn, offering examples to illustrate major issues and problems. These examples are not meant to comprise an exhaustive list. In fact, they constitute only a very small number of criminal justice policies that could be cited and explored. They are offered only as useful illustrations that are chosen because they have been recently discussed widely in popular media and scholarly research. Most of the examples are taken from US crime policies, but similar questions and programs are observable internationally.

Taking the four topical areas in order, hundreds of examples of crime policies under each could be cited, but this article covers only a few which can serve as illustrations. Examples of the sociological concern about 'what is crime?' are the varied policies about use of drugs, immigration, or white-collar deviance. (The second sociological concern, 'what causes people to commit crime?' is important because any particular crime suppression policy must be rooted in the expectation that it will affect a cause of crime; however, this concern is primarily criminological and thus is beyond the scope of consideration in this article.) An example illustrating the tension between security and liberty is legal policy about surveillance of electronic communications, Internet use, and merging of data about consumer choices into metadata made available to law enforcement for crime suppression. Also involved is the classic tension between due process and crime control. Examples of governmental structural concerns involve the proper role of the various agencies of the criminal justice system in crime reduction: police, prosecutors, courts, and agencies designed to punish criminals often overlap in their functions or take on characteristics of each other which may violate separation of powers necessary for checks and balances. Examples of practical policies are laws or programs which are expected to reduce crime – for instance: capital punishment, policing through aggressive surveillance and intervention in street encounters, gun control, diversion from prosecution of persons charged with drug or property crimes, granting immunity from prosecution in return for a defendant's cooperation in convicting others, operation of treatment or 'problem-solving' courts, reduction of bloated prison populations, and providing social support for ex-offenders upon reentry into public life after having served their sentences in prison. Any of these programs or policies – and there are many, many more – can be evaluated in cost–benefit terms, in terms of fidelity to the law, in terms of management efficiency, and/or by other methods.

Finally, note that all these public policies and governmental operations function at several different jurisdictional levels. Crime is generally regarded as primarily a matter for local government, considering that individual criminals live in local communities and that particular conditions and neighborhood knowledge are critical in responding to criminal acts. However, local law enforcement and jails must operate within the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86155-3

framework required by state or provincial law, and authority to operate prison systems usually is lodged in the law of the larger jurisdiction. Similarly, a province or state usually has its own police agency with statewide jurisdiction over crimes affecting the public in the many local jurisdictions (for instance, traffic enforcement on state highways, or consumer frauds committed by the same criminal enterprise in many different cities). Similarly, in the United States, a federal criminal justice system takes jurisdiction over crimes that occur in multiple states (for instance, drug smuggling, or frauds committed by corporations doing business in many states) or crimes against federal entities (e.g., corruption among elected federal officials) or crimes committed on federal lands (e.g., Indian reservations or the District of Columbia).

Similarly, international crime policy addresses matters of genocide, human smuggling, international terrorism, and other crimes. The international law principle of universal jurisdiction holds that a former head of state can be held to account by public trial for human rights violations – war crimes, crimes against humanity, genocide – which occurred during the official's term of office. A permanent International Criminal Court (ICC) has recently begun operations. Not only state officials but also persons acting in their individual capacities as war lords or torturers are also subject to the ICC jurisdiction (Schabas, 2011). The ICC handed down its first conviction against Thomas Lubanga of the Congo, for war crimes, in 2012. The United Nations can call upon the Court to begin proceedings – as, for example, against North Korea for committing crimes against humanity including rape, forced abortions, enslavement, and religious persecution 'meant to sustain the current political system' (United Nations, Human Rights Council, 2014). The power of this new Court is likely diminished by the refusal of several powerful counties, specifically the United States and China, to support it. One reason the United States might not be willing to support the International Criminal Court is that the Court may prosecute heads of state who are alleged to have engaged in torture for political ends, so ex-President George W. Bush might arguably have been subjected to prosecution for interrogation practices used against alleged terrorists incarcerated at Guantanamo Bay. Because this article covers crime policy in the United States, these international topics are not covered further here.

Policy Defines Crime

Universally, *mala in se* crimes (translated: 'evil in itself,' or 'intrinsically bad') are regarded as violations of the law in all nations. These roughly correspond to violent felony crimes: homicide, rape, robbery, and assault. In addition, theft is universally considered to be a crime, although the conditions under which the property is taken may or may not cause the action to fall under the category of crime.

Nevertheless, the assumption of universality is challenged when legislation and law enforcement policies are seen to vary across jurisdictions and time. The laws about whether a homicide is justifiable, for instance, vary among states within the United States. Recently, the 'Stand Your Ground' law was hotly debated in a Florida case that drew considerable public attention. 'Stand Your Ground' laws have been passed in approximately half of states in the United States. These laws state that a person who reasonably perceives him- or herself to be at risk of bodily harm, in any place where he or she is lawfully present, may defend using any force necessary. In these jurisdictions, a person is under no obligation to yield to threats even if the assailant can be avoided. If the local prosecutor would determine that a person perceived that an alleged attacker presented a serious threat of bodily harm, no murder has occurred if the former kills the latter – even if it would have been easy to avoid it.

In the Florida case, Trayvon Martin, an unarmed black teenager, was shot and killed by a white Hispanic neighbor, George Zimmerman, who said he believed that Martin was suspicious and that the teenager had threatened him when Zimmerman attempted to question him on the street. The prosecutor charged Zimmerman with murder, but only after public outcry at her initial refusal to do so. The prosecutor had to prove that Martin had not threatened Zimmerman; if he had, the 'Stand Your Ground' law would have applied and the threats would have constituted a defense to the killing. The jury found that Zimmerman reasonably believed he was being threatened and acquitted him. (This verified the prosecutor's original opinion, but public outrage had to be taken into account.) The case highlighted racial animosity in Florida and sparked protests about laws that seem to be aimed at racial minorities.

Evaluations of Stand Your Ground laws in practice are contentious. However, states do report to the Federal Bureau of Investigation every homicide that occurs and whether it was ruled justifiable. It is impossible to prove whether the laws caused the trends, but the data do indicate that homicides rose in many states that passed them, even as they declined in states without them, and that prosecutorial findings of justifiable homicide are significantly higher in those states. (Interestingly, Florida is the only state that does not report these data to the F.B.I.) Furthermore, studies have found "substantial evidence of racial disparities in justifiable homicide determinations. Regardless of how the data are analyzed, substantial racial disparities exist in the outcomes of cross-race homicides … [and] the recent expansion of Stand Your Ground laws in two dozen states appears to worsen the disparity" (Roman, 2013).

This example of variation in state laws illustrates the sociologists' point: crime is defined differently depending on time and place and culture. Stand Your Ground laws, which essentially define a killing as not murder if done under any circumstances presenting a threat, are generally on the books in states where voters support gun ownership. What act the law will regard as a crime is not universal even in the most serious of felonious actions: murder. Crime policy as set by state law varies accordingly.

Perhaps the most contentious contemporary example of the difficulties of definition is whether drug use is a crime. Drug use has generally been considered a 'victimless crime,' – along with prostitution, gambling, production, and sale of pornography, and so on – because the person engaging in the acts is said to harm only himself or herself, but laws will often define the acts as criminal if legislators believe they cause great social and moral harm to the general public. (Indeed there may be a victim, though, depending on one's view as to how harmful these private acts are to the larger society. If the public

in general is the victim of these acts, they would not be regarded as 'victimless.' Thus, alternative terms could be 'consensual crime,' or 'vice crime' in the examples given here.) Indeed there may be a victim, although, depending on one's view as to how socially or morally harmful, these private acts are to the larger society. If the society in general is considered the victim of these acts, alternative terms such as 'consensual crime' or 'vice crime' or 'morals crimes' may be preferred. That these 'morals' laws vary over time is well understood. Modern Americans may scoff at the moral crusading of the Temperance movement which advocated prohibition of the manufacture, transportation, and sale (though not the use) of beer, wine, and liquor and was successful in bringing about the passage of the 20th Amendment to the US Constitution in 1920. Although Prohibition was repealed 13 years later, the policy of demanding abstinence indeed had the effect of significantly lowering the rate of illness, deaths, and violence associated with alcohol consumption; perversely, it also had the effect of empowering gangster syndicates that built an underground booze-producing economy and fostered corruption in the law enforcement agencies that had to enforce the law (Aaron and Musto, 1981). The production and sale of alcohol was a crime in 1920 and was not a crime in 1933. Definitions of crime can shift.

The contemporary equivalent of Prohibition is 'The War on Drugs.' Although production, sale, and use of drugs ranging in intensity of effects from marijuana to heroin to psychedelics have variously been defined as against the law throughout the nineteenth and twentieth centuries, any could be 'decriminalized' by simply not enforcing these drug prohibition laws. Conversely, drug use, production, and/or distribution could be defined and prosecuted and punished as serious felonies akin to homicide or sexual assault. That is indeed what happened as the 'War on Drugs' proceeded through the late 1980s until late 2000s. Today, sentencing policy about production, sale, and use of various drugs has become generally less punitive but also widely divergent among the states and federal jurisdictions. "[A]t least 29 states have taken steps to roll back mandatory sentences since 2000 … Much of this legislative activity has taken place in the last 5 years and most changes affect nonviolent offenses, the vast majority of which are drug-related" (Subramanian and Delaney, 2014). Indeed, in 2012, the states of Colorado and Washington repealed their laws against the use of marijuana. At the same time, it became apparent that the nation's most dangerous drug abuse problem has become the use of prescription opiates, which are, of course, usually prescribed and used legally but can be diverted to recreational uses.

Definitions of business-related crime also shift over time and are subject to nonenforcement, even if they are 'on the books.' A common enraged outcry following the financial crisis of 2008 was that 'none of those guys went to prison.' However, making mortgage loans while ignoring the fact that the mortgagees were unlikely to be able to repay them – with the intent of selling the loans as worthless mortgage-backed securities to unsuspecting stock investors – was not against the law. When the economy went into 'freefall' as a result of collapsing credit markets and shrinking consumer demand, the public discovered that these mortgage practices were not subject to criminal laws; indeed, they were not even regulated closely by administrative agencies. Another example is the story of the insurance company AIG. Reckless gambling on what banks would do in credit markets, using 'credit default swaps,' caused AIG to lose $200 billion. The federal government took on the debt in a 'bailout' to prevent the nation's economy from crumbling, but none of AIG's acts had previously been defined as crimes in federal law.

Although the laws on the books have not changed much since 2008, prosecutorial discretion as to how vigorously to apply those laws operates such that actions previously overlooked will be prosecuted if they have caused great harm and are more widely publicly condemned. As Khanna points out, "One of the first things to note about corporate crime legislation is that enforcement has traditionally been quite thin – indeed it may appear largely symbolic…" (Khanna, 2004), so in this example the issue is not that the laws are not on the books, but that they are not enforced. Furthermore, from what we saw in the mortgage meltdown, most of that activity was not against the criminal law – regulatory law, perhaps, but not criminal. Thus, it is apparent that policy choices about what acts are crimes vary depending on time, place, and circumstance. Sometimes policy is quite punitive and at other times it is almost nonexistent. Setting the degree of social control a policy is expected to exert over particular acts – labeling them criminal or not, or choosing to enforce the law or not – is a political process influenced deeply by the cultural and historical assumptions of the age.

Suppressing Crime and Protecting Liberty

In liberal democracies, the omnipresent tension between preserving liberty and providing security lies at the heart of crime policy debates. The word 'liberal' derives from Enlightenment philosophy regarding 'liberty,' which social contractarians defined as the capacity to live and work freely in civil society subject to the constraints of law. The law provided the security of protection from harm by others, but it also controlled government so that it could not intrude on citizens' personal lives (Locke, 1690). Those 'personal' lives are roughly set out in the Bill of Rights, the first 10 amendments to the 'social contract' of the United States Constitution: freedom of religion, freedom of speech, freedom not to be arrested on suspicion of criminal activity unless the government has probable cause to conclude that the act was criminal and the suspect was the one who committed it, the right not to be coerced to confess to crime, freedom from being punished cruelly, and so on. These 'liberal' or 'liberty' protections must operate at the same time as government sets out to provide personal security against crime for its citizens. That liberty and security often clash is well understood, but the best way to formulate policies that will provide the greatest amount of both liberty and security simultaneously is not as easily grasped.

For example, recent controversies about the proper scope of the collection and use of data by the National Security Agency raise serious concerns about whether the federal anti-terrorism program has resulted in violation of 4th Amendment rights of US citizens. (The 4th Amendment to the United States Constitution provides "The right of the people to be secure in

their persons, houses, papers and effects, against unreasonable searches and seizures, shall not be violated ….")

Must law enforcement officers suspect that a particular person has committed a crime, such as aiding terrorists, before investigators will be permitted to gather information which might eventually ripen into sufficient evidence to constitute probable cause to arrest? This would seem to be the legal standard, but the technology of information gathering has progressed so quickly and deeply that it is now possible to track the phone calls, physical locations, and online activity of virtually all Americans. Patterns of online behaviors that might lead to knowledge about crimes such as terrorism – or, arguably, political activity or reading habits – are observable through algorithmic data-mining, even though the individuals whose activities are surveilled had not been linked to or suspected of any crime in particular at the beginning of the data analysis. Connections between phone numbers of people known to have links to terrorism and everyone those suspects have called constitute one 'hop' when all data on the contacts are gathered and examined for patterns. Next, all phone numbers connected to those contacts are also examined: the second 'hop.' The National Security Agency is said typically to make at least four 'hops' examining the phone call patterns in all phones branching from the original suspect, subsequently 'fusing' the information from a great many different data sources together around a common personal identifier such as a social security number or an email address if suspicious phone call patterns of a particular number are observed. The policy question is how to find the point at which personal freedom is preserved at the same time that security from acts of terrorism is enhanced. In this context, the question would be: "how many hops are too many, and how far can government go in fusing personal information to a particular phone number/identifier?" This problem has recently been litigated and conflicting opinions from federal district courts indicate that the issue will probably find its way to the US Supreme Court in the future. (See, e.g., the memorandum opinion in *Klayman v. Obama*, 2013.)

This issue is scarcely new. It has always been the case that law enforcement pushes to investigate and arrest, while Constitutional law demands that these activities be conducted only when controlled by liberty-preserving standards and procedures. A common way of describing this tension is to say that crime control must be accomplished under the constraints of due process (Packer, 1968). Most commentators today state that due process has been swept away in favor of a crime control apparatus that observes the forms of due process (i.e., requires probable cause to arrest, prohibits brutal interrogations, does not coerce confessions, provides the right to a jury trial) while substantively supporting crime control ideology. (Stuntz, 2011; see also Greenspan, 2014, arguing that the standard of proof 'beyond a reasonable doubt' historically required to prove criminal culpability has eroded down to the lower standard of 'preponderance of evidence' typical of administrative proceedings.) Several scholars state that crime control strategies have trumped liberty interests because actuarial risk management techniques have become the most common tools for reducing harm in all areas of contemporary life; we manage and spread risk, even the risk of crime, in much the same way as insurance companies manage and control the risk of automobile accidents. (O'Malley, 2010; see also Feeley and Simon, 1992, arguing that criminal justice agencies now operate under a 'new penology' designed to 'manage the underclass' using actuarial techniques of prediction and control.)

Some legal scholars, by contrast, tend to criticize the due process model for its very success, which led to its elimination at the hands of crime prevention. That is, due process became so cumbersome and technical that the system evolved to avoid it (Stuntz, 2011). One important example: in 1975, approximately 20% of felony cases went to trials before juries. In 2010, only 1% of them did (Galanter, 2005). A guilty plea is technically a confession to the crime, which means that the state need not meet its burden of producing evidence in public trial which proves beyond a reasonable doubt that the defendant is indeed guilty. It is generally well known that defendants often plead guilty because conviction after a jury trial will produce significantly harsher penalties than conviction from a guilty plea will (McCoy, 2005). This is a strong example of criminal justice policy evolving toward crime prevention and control at the expense of personal freedom (in this case, the right to trial).

In practice, the government has little incentive to produce laws that would evince a policy of providing both security and liberty simultaneously. When it moves too far in devising and implementing policies meant to provide security, Constitutional standards as interpreted and applied in the courts, via the doctrine of judicial review, can insert elements of personal liberties in the process. Furthermore, legislation covering security matters is drafted so as to take liberty concerns into account. In the example of surveillance of communication technologies, see 'Liberty and Security in a Changing World,' legislative standards recommended by a Presidential review panel (President's Review Group on Intelligence and Communication Technologies, 2013).

Perhaps the most well-known metaphor for describing the tension between crime control and liberty protection in the field of criminal justice policymaking is 'the net.' The state seeks to catch and control deviant people in its 'net,' and the liberty issue concerns how wide the net will be, and how small its webbed openings will be. In other words, how many people will be caught and kept in the net of the criminal control system? Policies may be seen as 'widening the net' or narrowing it and/or as 'broadening the sieve's holes' or closing them.

For example, policies about appropriate punishment following conviction – criminal sentencing in general, but also administrative procedures of particular programs such as parole from imprisonment – become more or less liberty preserving depending on how intrusive the state is permitted to be in demanding that convicted offenders submit to programs or incarceration. Julian Roberts gives an example: the net is widened when a 'community sanction' such as electronic monitoring combined with house arrest is applied "to offenders who were not at risk for a term of custody in the first place" (Roberts, 2004: 117). Supposedly for the goal of reducing reliance on imprisonment as a criminal sanction, a variety of nonincarcerative sentencing options are developed (electronic monitoring, work programs, drug treatment combined with halfway house nighttime custody, fines and restitution via restorative orders, home custody, etc.). If these options are not used instead of imprisonment, or are added on

to parole orders after release from prison, the net has widened. If judges send the same offenders who would have gone to prison to these in-community punishments instead, the net has narrowed. If an offender commits a new crime or violates a condition of the program and is sent to prison, the 'hole of the sieve' is tight; if the offender is kept in the community and program restrictions tightened, it is wider. Whether the net widens or loosens is a matter of policy choice by legislators and executive branch corrections managers.

Branches of Government that Design and Implement Crime Policy

These observations lead to the third topical issue: the exercise of state power to enforce the law. Law enforcement is the province of the executive branch of government. Law enforcement is explicitly concerned with crime control, and its agencies are the police (conducting investigations and also response to crimes that have occurred) and prosecutors. Although prosecutors are often regarded as members of the judicial branch because they are lawyers, in fact they perform law enforcement functions of criminal investigation and response to crimes, as well, when they hold criminal suspects to account through court proceedings. The prosecutorial power is widely regarded as the most powerful function of all criminal justice officials, insofar as it touches the greatest number of people and can exert considerable control over any of them. (The judiciary is also quite powerful, but judges consider only those cases which prosecutors bring to them, and prosecutors can investigate and also make agreements with suspects that are not reviewed in court.) Crime control is conceptualized as crime prevention and also response to crimes that have already occurred, and it can include social aid under a rehabilitative or restorative rubric, under the assumption that helping offenders improve their lives will prevent them from committing more crimes.

Law enforcement policy shifts over time, as do policies in all branches of government. An example is policies in local jurisdictions related to street crime. Local police traditionally have operated reactively; they respond to victims' calls for help. But police duties have always also included patrol, either on foot or in vehicles. Twenty-first century policing has evolved to emphasize crime prevention as well as response. This is often called 'proactive policing' as opposed to 'reactive.' Environmental crime prevention through 'target hardening' (i.e., improving, for example, locks and alarms) as well as installing closed-circuit television surveillance on public streets are examples of earlier crime prevention programs that public police have undertaken, but a more fundamental repurposing of patrol itself has now become more common. Stop-question-and-frisk practices, by which patrol officers confront persons on the street and ask for their identification and information about their activities, which might lead to suspicion of wrongdoing and thus to a body frisk for contraband such as weapons or drugs, have become much more common. Although the Supreme Court case of *Terry v. Ohio*, 392 U.S. 1 (1968) held decades ago that police officers may frisk suspects only if the officers believe that their lives or the lives of others are in danger, courts seldom question the legality of this practice when arrested suspects are brought before them. Nevertheless, in some jurisdictions the practice has become so controversial that residents of neighborhoods in which stop-question-frisk is a very common tactic of crime prevention sometimes protest bitterly that the stops are racially pretextual rather than rationally crime related. Federal courts may require that police policy on the matter be much more carefully controlled. (See: *Floyd v. City of New York*, Southern District of New York, decided 12 August 2013 cite not yet available.) These controversies are very similar to earlier policy debates about whether police should arrest and detain people for disorderly activities such as panhandling or drinking in public, rather than referring the problems to city agencies dealing with the homeless and addicted, on the assumption that vigorous intervention in 'quality of life crimes' will prevent more serious crime in the future. (Compare Kelling and Coles, 1996 with Harcourt, 2001.)

Of course, the executive is not the only branch of government which works on crime policy. (Criminal corrections agencies responsible for the delivery of punishment through in-community programs or incarceration are executive agencies and are discussed below.) The legislature writes the criminal code, which might be seen as a 'meta-policy.' The judiciary is more concerned with crime on a daily basis, since judges subject each case to scrutiny in the usual criminal procedures. Of course, crime policy like any other public policy is shaped by its budget. It is generally agreed that the quality and availability of criminal defense attorneys are highly correlated with the amount of public funding available. If defense attorneys were to demand that the state demonstrate the strength of its evidence more often than the current plea bargaining system allows, courts' policies about proof and speedy trial would probably change.

What Works?

Finally, which crime policies work best? Legislatures will pass laws to support those programs that are seen to be effective, and the programs will be required to design their own daily standard operating procedures (which might be regarded as policies in themselves).

The goals of crime policy are several and distinct depending on the problem to be solved, as the previous discussion of liberty and security shows. The goal may be to provide fairness in criminal adjudication, to punish effectively and humanely, to reduce crime and recidivism, or to achieve other goals. Obviously, inquiry into whether a policy 'works' or not begins with a statement of what goal it is intended to achieve. Sometimes, this is difficult.

For example, what is the goal of capital punishment? It may be to deter future murderers or it may be to provide retribution, or it may be simply to incapacitate the criminal in the most drastic way to prevent recidivism. If a state legislature determines that none of these purposes are sufficiently met to offset the costs of the death penalty (both in moral terms criticizing the state for taking any life no matter how condemnable, and in terms of the high costs of adjudicating these cases in court), the state law will not permit capital punishment. Currently, 32 state criminal codes permit the death penalty, but in only 14 have executions actually been carried out in the past 5 years.

(Death Penalty Information Center, 2014; for a history of policymaking about the death penalty in America, see Mandery, 2013.) Legislators' policy decisions are evident in capital statutes, while prosecutors' policy decisions about who to charge with a capital offense, whether to accept a guilty plea that will produce a sentence of life imprisonment without parole, and whether to give leniency to codefendants who testify against named defendants are all policy choices highlighting the nature of official discretion. If the state law does provide for the death sentence and prosecutors charge it and judges impose it, another discretionary policy choice falls to corrections officials in deciding how to carry out the sentence. (Electric chair? Firing squad? Lethal injection and if so, with what drugs?)

This is perhaps the most extreme example of crime policy, but the dynamics are the same throughout all agencies of criminal justice and can be described in the thousands of instances of policymaking within the many agencies of government. Simply put: crime policy is the outcome of official discretion to decide which course of action the state will take. That discretion itself may be more or less constrained, but it is always at the base of any decisions about what policy to set and how to implement it.

See also: Criminal Liability and Criminal Defenses; Death Penalty; Drugs, Decriminalization of: Cultural Concerns; Due Process in Law; Economics and Crime; Policing; Privacy: Theoretical and Legal Issues; Prosecution; Punishment: Social and Legal Aspects; Victimless Crime; White Collar Crime: Recent Trends and Debates.

Bibliography

Aaron, Paul, Musto, David, 1981. Temperance and prohibition in America: a historical overview. In: National Research Council. Alcohol and Public Policy: Beyond the Shadow of Prohibition. The National Academies Press, Washington, DC.

Death Penalty Information Center. http://www.deathpenaltyinfo.org/executions-year (accessed March 2014).

Feeley, Malcolm M., Simon, Jonathan, 1992. The new penology: notes on the emerging strategy of corrections and its implications. Criminology 30 (1).

Galanter, Marc, April 2005. The hundred-year decline of trials and the thirty years war. Stanford Law Review 57 (4).

Greenspan, Rosann, 2014. The Transformation of Criminal Due Process in the Administrative State. Quid Pro Books, New Orleans, LA.

Harcourt, Bernard E., 2001. Illusion of Order: The False Promise of Broken Windows Policing. Harvard University Press.

Kelling, George L., Coles, Catherine M., 1996. Fixing Broken Windows: Restoring Order and Reducing Crime in Our Communities. Touchstone Books.

Khanna, Vikramaditya S., 2004. Corporate crime legislation: a political economy analysis. Washington University Law Review 82 (1).

Klayman v. Obama, memorandum opinion filed, 16 December 2013, United States District Court for the District of Columbia, Richard J. Leon, District Judge. Accessed at: http://sensenbrenner.house.gov/uploadedfiles/klayman_v._obama.pdf (accessed 20.03.14.).

Locke, John. The Second Treatise on Government (originally published 1690, relevant excerpts republished in C.B. McPherson, Hackett Publishing Company, Indianapolis and Cambridge, 1980).

Mandery, Evan J., 2013. A Wild Justice: The Death and Resurrection of Capital Punishment in America. W.W. Norton and Co, New York.

McCoy, Candace, 2005. Plea bargaining as coercion: the trial penalty and plea bargaining reform. Criminal Law Quarterly 50 (1–2), 67–107.

O'Malley, Pat, 2010. Crime and Risk. SAGE Publications.

Packer, Herbert L., 1968. The Limits of the Criminal Sanction. Stanford University Press, Palo Alto.

President's Review Group on Intelligence and Communication Technologies, 13 December 2013. Liberty and Security in a Changing World. Accessed at: http://www.whitehouse.gov/sites/default/files/docs/2013-12-12_rg_final_report.pdf (accessed March 2014).

Roberts, Julian V., 2004. The Virtual Prison: Community Custody and the Evolution of Imprisonment. Cambridge University Press.

Roman, John K., 2013. Race, Justifiable Homicide and Stand Your Ground Laws: Analysis of the FBI Supplementary Homicide Report Data. Urban Institute Justice Policy Center, Washington, DC.

Schabas, William A., 2011. An Introduction to the International Criminal Court, fourth ed. Cambridge University Press.

Stuntz, William J., 2011. The Collapse of American Criminal Justice. Harvard University Press.

Subramanian, R., Delaney, R., 2014. Playbook for Change? States Reconsider Mandatory Sentences. VERA Institute, New York. Available at: http://www.vera.org/pubs/mandatory-sentences-playbook-for-change (accessed 24.02.14.).

United Nations Council on Human Rights, Report of the Commission of Inquiry on Human Rights in the Democratic People's Republic of Korea. Available at: www.ohcr.org (accessed 20.03.14.).

Crime, Race and Ethnicity

Malcolm D Holmes, University of Wyoming, Laramie, WY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The relationship of ethnicity/race to crime is a longstanding concern of criminology. Contemporary research identifies social characteristics of neighborhoods and differential treatment in the criminal justice system as the primary causes of higher crime rates among certain ethnic/racial minority groups. This article explores that research.

Introduction

A central concern of criminology is the differential patterns of crime and justice across various ethnic and racial groups in the United States. These groups are defined by the cultural (ethnic) and physical (racial) characteristics that people use to categorize and respond to themselves and to others. These categories are not always clear-cut. For example, Hispanics (or Latinos) may see themselves as members of a traditional racial category (e.g., White) or a distinct racial category (e.g., many respond Other Race in the US Census), and they will also possess a distinct ethnic identity (e.g., Puerto Rican or Mexican). Data pertaining to crime generally rely on broad categories that mask much of the variation in ethnic/racial identity, limiting analysis to comparisons among major distinctions, such as Whites and Blacks (or African-Americans). While various scholars and data sources employ different designations for ethnic/racial groups, here terms (e.g., Black and Hispanic) are used consistently for clarity.

Research on ethnicity/race focuses primarily on 'street' crimes (e.g., murder, rape, robbery, burglary) that are the mainstay of criminological research and, accordingly, are the focus of this discussion. The relationship of ethnicity and race to such crimes has constituted a major area of investigation for criminologists ever since the advent of scientific criminology during the late nineteenth century. Quite different viewpoints have characterized this work over time. For example, early research by the Italian physician/criminologist Cesare Lombroso suggested that non-European 'colored races' failed to fully evolve and, therefore, remain innately inclined toward 'savagery' (Horn, 2003). He thought that Blacks in the United States posed a social danger and were responsible for an increasing crime problem because of their inherent savagery. His ideas influenced the development of criminal anthropology in the United States, which achieved considerable popularity and helped fuel the eugenics movement in the early twentieth century. In this view, crime is heavily influenced by biologically based racial differences. These ideas were challenged by research conducted by sociologists such as Shaw and McKay (1969) at the University of Chicago. Relying on the approach of human ecology, they established that the incidence of crime is associated with certain characteristics of neighborhoods – poverty, heterogeneity, and transience – that appear unrelated to specific ethnic/racial groups. Over the past century, the viewpoint that such social factors explain variation in offending by ethnic and racial group has come to dominate the discipline, although some researchers continue to postulate that biological race contributes to criminality (e.g., Rushton, 1996).

Contemporary criminologists largely disavow a significant role for racial characteristics as causal factors in crime because commonly used racial distinctions defined by visible physical characteristics are social constructions rather than meaningful biological categories. Group categorizations based on gene distributions in populations do not correspond well with the categories of race in popular use; people with similar outward physical features used to identify race can be quite different genetically (e.g., Bamshad and Olson, 2003). Moreover, numerous studies in the tradition of the Chicago School have linked crime to the social structural and cultural characteristics of disadvantaged urban neighborhoods. Factors such as racial segregation, persistent poverty, lack of educational and employment opportunities, cultural social isolation, and discrimination appear to be the major causes of crime rate differentials among ethnic and racial populations (e.g., Peterson and Krivo, 2010; Sampson and Wilson, 1995).

Another possibility is overlooked in the focus on the causes of crime – the differential treatment of ethnic/racial minorities in the criminal justice system (CJS) may inflate crime rates for certain offenses among some groups (Tonry, 2011). Rather than representing underlying patterns of criminality, data on ethnicity/race and crime may reflect relatively greater social control of minority populations. A large body of evidence lends partial support to this proposition, but findings are not always consistent and appear to be contingent on factors such as type of offense. It appears that crime rates are explained both by variations in criminogenic social characteristics of urban neighborhoods and by differential treatment of ethnic/racial groups in the CJS.

Patterns of Crime by Ethnicity and Race

Research on patterns of crime among minority populations focuses primarily on the United States. The 2010 US Census reports that Blacks (12.6%) and Hispanics (16.3%) comprise the largest minority populations in the nation. Blacks are a diverse group, although most trace their ancestry to Africans who came to America via the slave trade. Hispanic is an ethnic identifier that encompasses numerous nationalities (those of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45002-6

Mexican origin make up 63% of Hispanics) and different racial identities (most Hispanics self-identify as White in the Census). Much of the contemporary research on race, ethnicity, and crime focuses on these two groups because of their substantial representation in the population, the various social and economic disadvantages they experience, their disproportionately high crime rates, and the discrimination they experience in the CJS.

Efforts to study the relationships of ethnicity, race, and crime are confounded by the limitations of available data. The most systematic national data consist of arrest rates and victimization surveys. Beginning in 1930, the Federal Bureau of Investigation began compiling annual data on crimes known to police, including seven serious index crimes – murder/nonnegligent manslaughter, forcible rape, robbery, aggravated assault, burglary, larceny/theft, and motor vehicle theft (arson was added in 1979) – as well as various lesser offenses (e.g., drug abuse violations). Nearly all US law enforcement agencies report known offenses occurring in their jurisdictions. These data are reported in the Uniform Crime Report (UCR), but they do not provide information on ethnic/racial identity of offenders. Arrest data are also reported in the UCR. These are broken down by four major racial categories but not by Hispanic and non-Hispanic (data on Hispanics were collected for brief periods during the 1930s and 1980s, but these offer little insight into contemporary arrest patterns). These data are limited by the exclusion of unreported crimes and reported crimes that are not cleared by arrest, which involve most crimes in many offense categories.

The other major data source on crime is the National Crime Victimization Surveys (NCVSs) conducted by the Bureau of Census for the Bureau of Justice Statistics. The surveys have been conducted since 1973. They are designed to be nationally representative and are used to make annual estimates of the number and rate of personal (rape, robbery, assault, and personal larceny) and household (household burglary, household larceny, and motor vehicle theft) victimizations. These victimization data have the advantages of capturing unreported crimes and of providing information on Hispanics, although the survey methodology also has inherent limitations (e.g., memory loss, misunderstanding about crime categories by respondents).

Black Crime

Table 1 presents race-specific arrest rates for 1990, 2000, and 2010, which include UCR index offenses (except arson) and drug violations. The rates are broken down into the four ethnic/racial categories used in the UCR. These data reveal that Blacks are arrested for major personal and property crimes at higher rates than those in other ethnic/racial categories. Caution must be used in interpreting these data given that many categories (e.g., larceny-theft) have low rates of reporting and arrest. In general, however, it appears that Blacks have disproportionately high rates of index offenses. Particularly notable is that murder rates among Blacks are several times greater than for Whites in each time period, and this offense category has high reporting and clearance rates.

It may be observed in Table 1 that murder rates, as well as those for most other personal and property crimes, declined precipitously from 1990 to 2010 (for a detailed analysis of arrest-rate trends, see Snyder, 2011). As arrest rates have declined, there has been some convergence between Black and White rates. For example, in 1990 the Black murder arrest rate was more than eight times that of Whites, but by 2010 the difference was a factor of less than six. Similarly, in 1990 the robbery arrest rate of Blacks was more than 10 times that of Whites, but by 2010 the difference was a factor of about 7. Most index crimes saw a noticeable relative reduction in Black–White arrest rates, although it is less pronounced for aggravated assault and burglary.

Another marked contrast among the ethnic/racial categories concerns drug abuse violations. Again Blacks have an arrest rate several times higher than those in other ethnic/racial categories. Although there is some convergence of Black and White drug arrest rates by 2010, this is largely attributable to a 1.5 times increase in the rate for Whites rather than the small reduction in the rate for Blacks. These observations, along with those for murder rates, underscore an important issue. Given high reporting and clearance rates for murder, there is little doubt

Table 1 Race-specific arrest rates (per 100 000 population) for selected UCR offenses[a]

	1990				*2000*				*2010*			
Offenses	*W*[b]	*B*	*AI*	*API*	*W*	*B*	*AI*	*API*	*W*	*B*	*AI*	*API*
Murder[c]	4.0	33.2	6.7	2.2	2.0	12.2	3.5	1.2	1.9	10.8	3.1	0.5
Forcible rape	8.5	44.4	12.7	3.7	5.4	17.6	8.0	1.8	4.6	12.6	7.3	1.2
Robbery	25.6	277.3	24.6	14.1	15.1	112.2	18.0	8.3	17.0	123.7	21.0	6.0
Aggravated assault	112.2	478.7	181.1	38.9	94.9	310.2	143.0	38.1	90.5	273.3	165.5	25.8
Burglary	115.3	339.7	153.2	40.3	62.0	154.6	72.2	21.8	68.1	178.6	66.9	13.6
Larceny-theft	414.6	1250.5	681.8	207.2	245.8	683.2	400.5	120.1	307.6	725.0	488.5	94.0
Motor vehicle theft	50.0	215.2	64.5	31.6	25.8	118.0	44.4	17.3	15.7	48.3	23.7	5.1
Drug abuse violations	252.0	1167.1	149.8	52.3	315.6	1034.6	224.0	70.3	378.8	1039.3	299.0	67.9

UCR, Uniform Crime Report.
[a]Offense categories include UCR index crimes (except arson) and drug violations (nonindex crime); calculations are based on data from UCR and US Census reports for the respective years.
[b]Race categories are: W – White; B – Black; AI – American Indian; API – Asian or Pacific Islander.
[c]Includes nonnegligent manslaughter.

that a disproportionately high incidence of homicide among Blacks exists. Drug violations, however, are rarely reported by citizens and enforcement largely involves proactive policing strategies, which raises the issue of whether the apparent high incidence of drug use among Blacks is a product of selective enforcement. These two offense categories highlight the complexities of explaining crime differentials and are discussed further below.

Additional evidence regarding crime among Blacks comes from victimization data. In 2005, Blacks accounted for 49% of all homicide victims (Harrell, 2007), which is comparable to UCR statistics for murders by Blacks (48.8% of total murders in 2000; 48.7% in 2010). These data reflect the intraracial nature of homicide. NCVS data show that in 2007, the total Black victimization rate (26.1) for personal crimes was higher than the rate of White victimization (20.5) (BJS, 2010). The higher rates of robbery (4.9) and aggravated assault (4.4) victimization among Blacks compared to Whites (1.9 and 3.2, respectively) are particularly noticeable. With respect to property crime, the victimization rate for Blacks (170.0) was somewhat higher than for Whites (141.9), which reflects higher rates of burglary (37.2) and motor vehicle theft (14.2) victimization among Blacks compared to Whites (25.1 and 7.1, respectively). Taken together, the rates of arrest and victimization indicate that Black communities are disproportionately affected by crime.

Explaining Index Crimes of Blacks

The substantial differences in arrest rates of Blacks and Whites for serious crimes provide fuel for the debate about the causes of Black criminality. As noted earlier, the dominant perspective in contemporary criminology is that social and economic disadvantages of segregated Black neighborhoods produce the disproportionately high rates of crime among Blacks. While data collected within individual cities lend support to that argument, the lack of data from a broader array of cities has not permitted a more definitive conclusion about the effects of neighborhood characteristics on crime until recently.

To address the issue, Ruth Peterson and Lauren Krivo (2010) conducted the National Neighborhood Crime Study (NNCS) to obtain census tract-level data on reported violent (homicide and robbery) and property (burglary, larceny, and motor vehicle theft) crimes and on social and demographic characteristics for 91 cities. The sample is representative of the US cities with 100 000 or more population. Analysis of the NNCS data reveals that whereas White neighborhoods have very low levels of economic and social disadvantage, most African-American neighborhoods have very high levels. When a composite disadvantage index is included in statistical models to explain the overall ratio of violent crime rates between Black and White neighborhoods, the vast majority of the difference between them is accounted for (although the rate in Black neighborhoods remains 65% higher than in White neighborhoods). The property crime rate for Black neighborhoods is just 10% higher than for White neighborhoods after the disadvantage index is added to the model. When the spatial characteristics (e.g., disadvantage, violent crime rate) of surrounding neighborhoods are included in the analysis, the Black violent crime rate is just 13% higher than in White neighborhoods, and the property rate is slightly lower.

These analyses provide strong evidence that structural characteristics of US cities, namely the disadvantages disproportionately borne by segregated Black neighborhoods, largely account for the exceptionally high crime rates observed for Blacks. It appears, therefore, that much of the serious crime problem that exists among African-Americans can be explained by the historical segregation of a substantial part of that population in disadvantaged urban neighborhoods (see also Massey and Denton, 1993).

Blacks and the War on Drugs

The War on Drugs, which began during President Reagan's administration, profoundly affected the Black population. The greatly disproportionate Black arrest and imprisonment rates for drug abuse violations during the past three decades would seem to suggest that drug use is a substantial problem in that population. However, that is not the case. While arrests for drug violations are much higher among Blacks, their overall use of most drugs is lower than among Whites (Tonry, 2011). For example, cocaine use (powder and crack combined) is appreciably higher among Whites than Blacks. Moreover, Blacks are not more involved than Whites in the drug trade. One reason for the higher arrest rate involves the social organization of drug distribution. The anonymous transactions in the public drug markets of urban Black neighborhoods are more susceptible to surveillance than are the personal transactions among acquaintances within existing social networks of White neighborhoods. Moreover, police stop and search Blacks more often than Whites in public places, although these contacts result in lower rates of contraband seizure (i.e., guns and drugs) for Blacks than Whites, suggesting that race per se is often the reason for stopping Blacks.

Hispanic Crime

Research on crime among Hispanics is complicated by the UCR reporting system. Most of the arrest data for Hispanics are folded into the White category, disallowing any direct comparisons with other ethnic/racial groups. However, Peterson and Krivo's (2010) NNCS included data on crime in Hispanic neighborhoods. Hispanics tend to reside in relatively disadvantaged neighborhoods compared to Whites, but the degree of Hispanic neighborhood disadvantage does not reach to that experienced by Blacks. Crime rates also occupy an intermediate position. When neighborhood disadvantage is included in statistical models to predict neighborhood crime rates, the violent crime rate remains 29% higher than in White neighborhoods, and the property crime rate is 8% higher. With the addition of spatial characteristics of surrounding neighborhoods to the analysis, the differences between Hispanic and White neighborhoods for both violent and property crimes become negligible.

Victimization data indicate that Hispanics have a slightly lower personal crime victimization rate (19.6) than Whites (see above). The Hispanic rate of property crime victimization (190.5), however, is higher than for Blacks or Whites. The higher rate of property victimization is particularly evident for motor vehicle theft (over twice as high as Whites) and theft (one and a third times that of Whites).

Hispanic Immigration and Crime

In recent years a significant debate has centered on whether illegal Hispanic immigrants (especially from Mexico) to the United States are disproportionately involved in crime. The concern about immigrant crime provided the impetus for controversial legislation aimed at controlling the alleged problem in Arizona. The 2010-law included the requirement that law enforcement officials attempt to ascertain individuals' immigration status during a lawful stop or detention and made it a misdemeanor offense to be an alien in Arizona without carrying required documents (e.g., an Arizona driver's license). The Supreme Court (*Arizona v. United States*, 567 U.S.___ 2012) ruled that key provisions of the Arizona statute are preempted by federal law, but illegal immigration and crime remain hot-button issues. While some are convinced that immigrant crime poses a serious problem, others are concerned that Hispanics are being indiscriminately subjected to racial profiling.

Criminologists question whether a crime problem actually exists. Research generally finds little evidence that immigrants are disproportionately involved in crime (for a review, see Ousey and Kubrin, 2009). Contemporary public perception of the problem is fueled in part by the large number of incarcerated Mexican immigrants. Immigrants are largely young males, mirroring the demographic categories generally associated with crime. Once the age and sex distribution of the immigrant population is taken into account, Mexican immigrants are incarcerated in state prisons at rates not much different than U.S. citizens (Hagan and Palloni, 1999). That study also found that immigrants in two US–Mexico border cities (El Paso and San Diego) are more likely to be detained before adjudication, which increases their chances of conviction and imprisonment. A longitudinal examination of the connection between immigration and violent crime rates in cities of 100 000 or more population shows that within-city change in immigration is negatively related to within-city change in violent crime rates, after important sociodemographic characteristics of the cities are controlled (Ousey and Kubrin, 2009). In other words, immigration is related to lower violent crime rates.

Crime among Other Ethnic/Racial Groups

As shown in **Table 1**, in addition to arrest data on Whites and Blacks, the UCR includes American Indian and Asian or Pacific Islander categories. These groups comprised .9% and 5.0% (4.8% Asian, 0.2% Pacific Islander) of the US population in 2010. It may be seen that the 2010 index offense arrest rates (except burglary) for American Indians are somewhat higher than for Whites, but appreciably lower than for Blacks. Drug abuse violation arrests are lower than for Whites and Blacks. Research on homicide among American Indians shows that social structural characteristics contribute to their high homicide rate (e.g., Lanier and Huff-Corzine, 2006), consistent with research on violent crime among other disadvantaged ethnic/racial groups.

All crime rates for those of Asian and Pacific Island ancestry are appreciably lower than for any other racial category. Although there is little criminological research on this population, these rates likely reflect the relatively economically advantaged position of many Asians compared to other ethnic/racial minority populations in the United States.

Differential Treatment in the CJS

The data presented above indicate the existence of crime differentials among various ethnic/racial groups. However, a critical question for researchers is the degree to which differential arrest rates reflect disparities of treatment in the CJS. While arrests and subsequent CJS decisions may be attributable to differences in legally relevant characteristics across groups (i.e., their involvement in crime), it also appears that discrimination based on offenders' extralegal social characteristics is at work. Historically, there has been considerable debate about the degree to which the more severe outcomes of Blacks and Hispanics in the CJS reflect disparate involvement in serious crime vs discriminatory treatment (Spohn, 2000).

Case Processing

While theories of crime causation have dominated the field of criminology since its inception, theoretical perspectives emphasizing the role of extralegal considerations in differential outcomes in the CJS gained greater prominence in criminology beginning in the 1960s. One such perspective maintains that crime is socially constructed by agents of the CJS. Stereotypes of crimes and offenders are said to influence decisions at various stages of the processing of defendants (e.g., Steen et al., 2005). Another approach argues that economically and socially disadvantaged defendants have relatively limited access to legal resources (e.g., pretrial release), which indirectly influences final dispositions (e.g., Holmes et al., 1996). While research consistently demonstrates that legal considerations – notably offense severity and prior criminal record – influence defendants' outcomes, these perspectives explain how social characteristics such as ethnicity/race also may influence outcomes at various stages of case processing. Insofar as differential patterns of incarceration may involve cumulative disadvantage accruing from decisions at various stages of adjudication, a process-oriented approach that considers patterns of arrest, pretrial decisions, and sentencing is necessary.

Arrest

A large body of literature indicates that Blacks and Hispanics are disproportionately stopped, searched, and arrested by police (for a review, see Smith and Alpert, 2007). These disparities exist across various regions and different types of law enforcement agencies. An obvious explanation for this pattern is the differential involvement of minorities in crime, but it appears that disparate treatment of minorities is not simply a by-product of differential offending. Historically, the racial animus of White police officers undoubtedly affected the differential treatment of minorities, but now greater representation of minorities in policing and less tolerance for overt discrimination by police departments may have reduced, but hardly eliminated, open discrimination against minority citizens. Researchers continue to document very aggressive, often legally impermissible, strategies of policing in disadvantaged minority neighborhoods (e.g., Brunson and Miller, 2006). Policymakers may deploy more police resources or implement aggressive stop and frisk strategies in minority neighborhoods, and individual officers may unconsciously engage in racial profiling irrespective of their

own racial identity (Smith and Alpert, 2007). These practices may increase discretionary stops of minority citizens, which, in turn, may increase arrest rates for crimes such as drug possession.

Pretrial decisions

Once an arrest takes place, several important decisions occur before formal adjudication occurs. The first is the setting of bail. The constitutionally sanctioned purpose of bail is to ensure a defendant's appearance at subsequent proceedings, except in cases where the defendant is a flight risk or poses a significant danger to the community. Yet, bail setting in urban jurisdictions, where most criminal cases are adjudicated, often involves perfunctory consideration of a defendant's background and ability to make bail. An analysis of a sample of formally charged felony defendants in the 75 most populous counties in the United States shows that Black and Hispanic defendants obtain less favorable pretrial release decisions and outcomes (released or not released) compared to Whites after controlling legally relevant factors such as criminal history (Demuth, 2003). Minority defendants may be disadvantaged in bail decisions because of stereotypes about criminality or a relative lack of economic resources to make cash bail. When defendants are not released prior to adjudication, final case dispositions may be adversely affected, although research on this issue remains limited and provides mixed findings (see, e.g., Holmes et al., 1996). Pretrial detention may, for example, reduce family stability or result in loss of employment, which constitutes ties to the community that are considered in subsequent decisions about probation vs imprisonment.

Charging decisions and plea negotiations have also been identified as possible sources of discrimination against minority defendants, but the influence of ethnicity/race on these decisions has received limited attention from researchers. Available studies provide mixed evidence that ethnicity/race affects these outcomes, but the paucity of research precludes any definitive conclusions (for a review, see Shermer and Johnson, 2010).

Sentencing

One of the most extensively researched and debated issues regarding ethnicity/race and crime is disparities in sentencing. Studies of sentencing focus on two outcomes: probation vs parole or length of sentence. The former in/out decision is particularly important because incarceration conveys significant disadvantages compared to release on probation, which allows continued employment, facilitates stable family relationships, and averts the difficult conditions of prison life. Both of these decisions are heavily influenced by legal considerations, notably defendants' prior criminal histories and present offenses, as expected. However, after controlling legally relevant variables, does ethnicity/race affect sentencing decisions?

Cassia Spohn's (2000) systematic analysis of the literature answers that question in the affirmative. For state courts, she reports that 55.5% of reported statistical effects show that Blacks, and 41.7% that Hispanics, are more likely than Whites to be sentenced to prison than to probation. Relatively few demonstrate that Black or Hispanic defendants receive longer sentences in state courts. At the federal level, some research finds evidence that Blacks (42.8% of reported effects) and Hispanics (also 42.8%) are more likely to be sentenced to prison, and most show that Blacks (66.7%) but not Hispanics receive longer sentences on average. In addition to the main effect of race, many of the studies Spohn reviews consider the contexts in which ethnicity/race influences sentencing. She identifies four patterns of contextual effects reported in the literature: (1) certain combinations of ethnicity/race and other extralegal offender characteristics (e.g., employment status) produce greater disparity than ethnicity/race alone does; (2) process-related factors (e.g., pretrial detention) disadvantage minority defendants at sentencing; (3) some combinations of offender–victim race (i.e., Black offender–White victim) may produce harsher sentences; and (4) sentencing differentials may depend on the type of crime (e.g., minorities may be sentenced more harshly for drug offenses). While noting that findings are not entirely consistent and significant effects of ethnicity/race are sometimes modest, she concludes that the disproportionate incarceration of minorities reflects more than legally relevant considerations. A subsequent meta-analysis of sentencing research similarly concludes that most statistical effects indicate that Blacks are punished more severely than Whites (Mitchell, 2005).

Incarceration

Data on incarceration in the United States reflect the patterns of arrest and adjudication described above. The higher arrest rates and probabilities of prison sentences for Blacks and Hispanics are reflected in much higher incarceration rates per 100 000 males (3119 and 1193, respectively) compared to White males (487) in state and federal prisons in 2009 (West et al., 2010). (Female incarceration rates are much lower than males, but somewhat smaller ethnic and racial disparities still exist.)

Focusing on the Black–White disparity in incarceration rates, Tonry (2011) demonstrates that since 1950 Black incarceration rates per 100 000 increased dramatically compared to White rates, and stood at roughly eight times greater by 2008. Part of the higher incarceration rates for Blacks is a reflection of higher arrest rates, particularly for violent crimes that often result in imprisonment; however, a sizable part of the higher rates of incarceration remains unexplained by arrest rates. Moreover, Blacks are arrested at higher rates for the personal offenses of robbery and aggravated assault than their involvement in those offenses would predict. Of particular significance is the greatly disproportionate rate of incarceration of Blacks for drug offenses, which is at odds with their patterns of use and distribution. Rather than reflecting differential involvement in crime, the rate is a product of the harsher sentence structure for drug offenses more common among Blacks (e.g., longer sentences for crack compared to powder cocaine under federal sentencing guidelines) and the practices of police (e.g., racial profiling) that result in much higher arrest rates for Blacks.

It appears that Hispanics are also disproportionately incarcerated relative to their involvement in crime. Research shows that, after controlling for legally relevant factors, young Hispanic males are more likely to receive a prison sentence (Spohn and Holleran, 2000). Moreover, relatively large numbers of Hispanics are incarcerated in state prisons for drug offenses (see West et al., 2010), although Hispanic rates of drug use are generally lower than that of Blacks or Whites (Warner et al., 2006).

The Death Penalty

Certainly the death penalty is the most severe and irrevocable criminal sanction, and its use highlights concern about ethnic/racial injustice in the CJS. At the end of 2009, there were 3173 prisoners under sentence of death in the United States (Snell, 2010). Of those, 56.1% were White and 41.5% Black (Hispanics are included in these calculations). Thus, the proportion of Blacks on death row is 3.3 times their representation in the US population, whereas the representation of Whites on death row is substantially lower than in the overall population (75.1%). Hispanics make up 13.5% of death row inmates, which is somewhat below their representation in the population in 2010. Since 1977, 1188 prisoners have been executed in the United States. Of these, 56.6% were non-Hispanic Whites, 34.6% were non-Hispanic Blacks, and 7.7% were Hispanic (all others were 1.2%).

Supreme Court decisions of the past four decades call attention to the racial disparity involving Blacks on death row. In *Furman v. Georgia* (408 U.S. 238, 1972), the Court narrowly ruled that the arbitrary, capricious, and discriminatory nature of the imposition of the death penalty violated the prohibition against cruel and unusual punishment of the Eighth Amendment, but the question of racial discrimination in its application was not the central issue in the ruling. Subsequent decisions allowed reinstatement of the death penalty so long as juries consider statutorily defined aggravating and mitigating factors in reaching their decisions (*Gregg v. Georgia*, 428 U.S. 153, 1976), but disallowed the use of the death penalty in rape cases because rape victimization does not involve 'serious' injury, and the death penalty is grossly disproportionate to the severity of the crime (*Coker v. Georgia*, 433 U.S. 584, 1977). Historically, the death penalty for rape had been applied almost exclusively in cases involving Black defendants and White victims, but this was not the central issue in the Court's decision (both the defendant and victim were White in the *Coker* case).

The Supreme Court directly confronted the issue of racial discrimination in the death penalty in *McClesky v. Kemp* (481 U.S. 279, 1987). Based on a sophisticated statistical study conducted by David Baldus and associates, the case of Warren McClesky (a Black defendant convicted of killing a White police officer) was appealed on the grounds that it violated provisions of the Fourteenth (equal protection) and Eighth (cruel and unusual punishment) Amendments. The study used to support these claims demonstrated that, after controlling numerous legally relevant factors, cases involving White victims were several times more likely than those with Black victims to result in a death sentence in Georgia (defendants' race was not a factor). Little disparity existed at the extremes of the aggravation continuum, with pronounced differences observed in cases falling in the midrange of aggravation, where cases might go either way. Although acknowledging the validity of these data, the Court affirmed McClesky's death sentence in a controversial 5–4 decision. It ruled that the statistical data did not provide clear proof that racial animus influenced the sentence of death in McClesky's case, which established the standard to prevail on appeal. Given that seemingly race-neutral court proceedings can mask race-based decision making, the ruling set a very high bar for successfully challenging death penalty sentences by requiring case-specific evidence of discrimination.

While the standard set in the *McClesky* case makes it more difficult for death penalty appellants to prevail, the Supreme Court has subsequently prohibited the use of racial criteria when credible evidence specific to the case at hand is presented. In the case of *Miller-El v. Cockrell* (537 U.S. 322, 2003), the Court held that the use of peremptory challenges to systematically eliminate Blacks from a capital murder jury violates the equal protection clause of the Fourteenth Amendment. In this instance, the statistical evidence of discrimination during *voir dire* (91% of Blacks eligible to serve on the jury were stricken by peremptory challenge) was specific to the case of Miller-El, and in the view of the Court the process of questioning prospective jurors also demonstrated purposeful discrimination by the prosecution in exercising peremptory challenges.

As it stands today, minority defendants will rarely succeed in appealing a death sentence on the grounds of racial discrimination because of the challenge of providing credible evidence of purposeful discrimination during court proceedings. Legislative and executive actions are a more likely avenue of relief from discriminatory practices, particularly given that the Supreme Court maintains legislatures are the appropriate venue for addressing the issue (see *McClesky v. Kemp*). Former Illinois Governor Ryan reluctantly commuted all death sentences in the state before leaving office in 2003, because of the evidence of racial discrimination, procedural irregularities, and wrongful convictions. Such intervention remains the rare exception, however, and it appears likely that the death penalty will continue to disproportionately impact minority defendants for the foreseeable future. Death penalty opponents remain particularly concerned about wrongful convictions. The Death Penalty Information Center reports that between 1973 and 2012, 142 death row inmates were exonerated, of whom 50% were Black.

Conclusion

The relationship of ethnicity/race to crime and justice has been a preoccupation of criminologists since the founding of the field. This article demonstrates that the problem is complex. Arrest and victimization data suggest that America's largest minority groups – Hispanics and Blacks – have high rates of serious crime relative to Whites. A substantial body of research supports the proposition that the disadvantaged conditions of Black and Hispanic neighborhoods explain many of these differences. Another body of research provides evidence that Blacks and Hispanics experience discriminatory treatment in the CJS, which inflates their apparent involvement in some crimes, notably drug violations. Taken together, criminogenic conditions of urban neighborhoods and discriminatory treatment in the CJS produce higher rates of imprisonment for Blacks and Hispanics.

This article outlines the major issues involved in the study of crime, justice, and ethnicity/race, but there are important intricacies that are beyond its scope. For example, there are extralegal activities of the CJS that do not influence the production of crime rates but adversely impact minority communities, create tensions with CJS agents, and undermine

the legitimacy of the system. The disproportionate use of excessive force by police in minority communities epitomizes the problem (e.g., Holmes and Smith, 2008). In addition, the focus of this discussion is limited to the United States. Research on other industrialized nations suggests that the patterns of crime and justice discussed here occur in many other nations as well (Tonry, 1997; Gabbidon, 2010). Studies of other nations remain sparse compared to the United States, however, partly because of limited availability of reliable data on ethnicity/race, crime, and justice. Another issue involves the social policy approaches that have been taken to address the problem. The piecemeal solutions commonly advocated by criminologists and policymakers will likely have little impact because they do not address the underlying problem of a racially and ethnically stratified society (Holmes and Smith, 2008; Peterson and Krivo, 2010). Clearly, the connection of crime, justice, and ethnicity/race will remain a critical area of criminological investigation for the foreseeable future.

See also: Crime, Sociology of; Crime: Knowledge about and Prevalence; Criminal Justice, Sociology of; Homicide; Immigration and Crime in the United States; Imprisonment: Sociological Aspects; Mass Imprisonment and Its Consequences; Policing; Prisons and Imprisonment; Race: Prespectives and Crime; Re-entry and Reintegration After Incarceration; Sentencing, Trends, and Disparities: Crime; Wrongful Convictions.

Bibliography

Bamshad, M.J., Olson, S.E., 2003. Does race exist? Scientific American 289, 78–85.

Brunson, R.K., Miller, J., 2006. Young black men and urban policing in the United States. British Journal of Criminology 46, 613–640.

Bureau of Justice Statistics, 2010. Criminal Victimization in the United States, 2007 Statistical Tables. U.S. Department of Justice, Office of Justice Programs, Washington, DC.

Demuth, S., 2003. Ethnic and racial differences in pretrial release decisions and outcomes: a comparison of Hispanic, black, and white felony arrestees. Criminology 41, 873–908.

Gabbidon, S.L., 2010. Race, Ethnicity, Crime, and Justice: An International Dilemma. Sage, Los Angeles.

Hagan, J., Palloni, A., 1999. Sociological criminology and the mythology of hispanic immigration and crime. Social Problems 46, 617–632.

Harrell, C., 2007. Black Victims of Violent Crime. U.S. Department of Justice, Office of Justice Programs, Washington, DC.

Holmes, M.D., Smith, B.W., 2008. Race and Police Brutality: Roots of an Urban Dilemma. State University of New York Press, Albany.

Holmes, M.D., Hosch, H.M., Daudistel, H.C., Perez, D.A., Graves, J.B., 1996. Ethnicity, legal resources, and felony dispositions in two southwestern jurisdictions. Justice Quarterly 13, 11–30.

Horn, D.G., 2003. The Criminal Body: Lombroso and the Anatomy of Deviance. Routledge, New York.

Lanier, C., Huff-Corzine, L., 2006. American Indian homicide: a county-level analysis utilizing social disorganization theory. Homicide Studies 10, 181–194.

Massey, D.S., Denton, N.A., 1993. American Apartheid: Segregation and the Making of the Underclass. Harvard University Press, Cambridge, MA.

Mitchell, O., 2005. A meta-analysis of race and sentencing research: explaining the inconsistencies. Journal of Quantitative Criminology 21, 439–466.

Ousey, G.C., Kubrin, C.E., 2009. Exploring the connection immigration and violent crime rate in U.S. cities, 1980-2000. Social Problems 56, 447–473.

Peterson, R.D., Krivo, L.J., 2010. Divergent Social Worlds: Neighborhood Crime and the Racial-Spatial Divide. Russell Sage Foundation, New York.

Rushton, J.P., 1996. Race, Evolution, and Behavior: A Life History Perspective. Transaction, New Brunswick, NJ.

Sampson, R.J., Wilson, W.J., 1995. Toward a theory of race, crime, and urban inequality. In: Hagan, J., Peterson, R.D. (Eds.), Crime and Inequality. Stanford University Press, Palo Alto, CA, pp. 37–54.

Shaw, C.R., McKay, H.D., 1969. Juvenile Delinquency and Urban Areas: A Study of Rates of Delinquency in Relation to Differential Characteristics of Local Communities in American Cities. University of Chicago Press, Chicago, IL.

Shermer, L.O., Johnson, B.D., 2010. Criminal prosecutions: examining prosecutorial discretion and charge reductions in U.S. Federal District Courts. Justice Quarterly 27, 394–430.

Smith, R.P., Alpert, G.P., 2007. Explaining police bias: a theory of social conditioning and illusory correlation. Criminal Justice and Behavior 34, 1262–1283.

Snell, T.L., 2010. Capital Punishment, 2009—Statistical Tables. U.S. Department of Justice, Office of Justice Programs, Washington, DC.

Snyder, H.N., 2011. Arrest in the United States, 1980-2009. U.S. Department of Justice, Office of Justice Programs, Washington, DC.

Spohn, C., 2000. Thirty years of sentencing reform: the quest for a racially neutral sentencing process. In: Horney, J. (Ed.), Policies, Processes, and Decisions of the Criminal Justice System, Criminal Justice 2000, vol. 3. National Institute of Justice, Washington, DC, pp. 427–501.

Spohn, C., Holleran, D., 2000. The imprisonment penalty paid by young, unemployed Black and Hispanic male offenders. Criminology 38, 281–306.

Steen, S., Engen, R.L., Gainey, R.R., 2005. Images of danger and culpability: racial stereotyping, case processing, and criminal sentencing. Criminology 43, 435–468.

Tonry, M. (Ed.), 1997. Ethnicity, Crime, and Immigration: Comparative and Cross-National Perspectives. Crime and Justice, vol. 21. University of Chicago Press, Chicago, IL.

Tonry, M., 2011. Punishing Race. Oxford University Press, Oxford.

Warner, L.A., Valdez, A., Vega, W.A., de la Rosa, M., Turner, R.J., Canino, G., 2006. Hispanic drug abuse in an evolving cultural context: an agenda for research. Drug and Alcohol Dependence 84 (Suppl.), S8–S16.

West, H.C., Sabol, W.J., Greenman, S.J., 2010. Prisoners in 2009. U.S. Department of Justice, Office of Justice Programs, Washington, DC.

Crime, Sociology of

Lorine A Hughes, University of Nebraska at Omaha, Lincoln, NE, USA
James F Short, Jr., Washington State University, Pullman, WA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.-P. Brodeur, volume 5, pp. 2937–2941, © 2001, Elsevier Ltd.

Abstract

This article summarizes theoretical and empirical developments and trends in the sociology of crime. The history of the field is reviewed first, followed by discussion of its primary methods, data sources, and levels of explanation. Attention then turns to the social distribution of crime, dominant theoretical traditions, and issues of crime control. We close with brief comments concerning continuity and change in the contexts and processes associated with crime.

Introduction

Early studies of the sociology of crime were often combined with study of other phenomena such as suicide, juvenile delinquency and dependency, poverty, and child abuse (Short, with Hughes, 2007). Although efforts to understand and control these disparate phenomena have elements in common, separate disciplinary approaches to their study soon developed. Boundaries between and within the social and behavioral sciences have become less distinct, however, as knowledge has accumulated and research programs have become more interdisciplinary. Crime control programs have also changed, often in response to discoveries and insights from the social and behavioral sciences. Relationships between sociological scholarship and crime control programs have been controversial but important to both scholarly and control efforts.

History of Sociological Issues in the Study of Crime and Criminals

The sociology of crime is marked by controversy, beginning with definitional issues. The focus, on law – law making and breaking, lawbreakers and their victims, and the control of law breaking and lawbreakers – serves as a unifying theme for an otherwise extremely diverse discipline. This theme is increasingly challenged, however, by the continually evolving nature of criminal and regulatory law, the diversity of substantive interests associated with expansion of the scope of behaviors defined as criminal, and differences within and between nations in law enforcement and justice practices. Such centrifugal tendencies in the sociology of crime have occurred even as study of crime has become more interdisciplinary, less identified as a special field of sociology, and more closely related to legal and extralegal crime control efforts. Diversity and change, in turn, have made the full range of the sociological *oeuvre* more relevant to the explanation of crime.

The sociology of crime is often traced to the emergence, in the nineteenth century, of the positivist view that crime and criminals could be objectively identified and measured. Attracted to the positivist thinking of Darwin, the Italian military physician Cesare Lombroso studied physical stigmata among Italian prisoners and soldiers. His attribution of crime to atavism, i.e., reversion to an earlier biological stage of development, was later modified to recognize the influence of social and economic circumstances and of mental illness.

Emphasis on social factors was inspired by journalistic accounts of social conditions associated with crime and child dependency and by systematic research conducted by André-Michel Guerry (1833), Adolphe Quetelet (1831, 1835), Émile Durkheim (1951/1897), and Gabriel Tarde (1886) documenting regularities in the distribution of social phenomena. Durkheim's insistence that social facts such as suicide and homicide rates were *sui generis* drew a boundary between sociology and psychology that profoundly affected the development of both disciplines.

As one of the earliest disciplines to include crime and criminals among its research and theoretical interests, sociology continues to be the primary discipline of many criminologists. Increasingly, however, sociologists of crime are associated with interdisciplinary research programs and criminal justice departments. Substantive interests also continue to evolve, largely in response to changing social conditions and legal criteria. Many white-collar crimes involve violations of regulatory law, for example, requiring relaxation of the stipulation that crime is defined as violation of criminal law. Post–World War II developments in international law defining war crimes, genocide, and other human rights violations further broaden the scope and interdependency of law and the social sciences. The complexity of such phenomena has spurred a variety of interdisciplinary sociolegal efforts and innovative methodologies and theoretical approaches. Sociologists remain at the forefront of such efforts (Hagan and Rymond-Richmond, 2009; Savelsberg, 2010).

Issues of Method, Measurement, and Levels of Explanation

Because crime consists of a very large number of heterogeneous criminal acts, no single measure or methodology can fully capture its nature. Historically, the full armamentarium of methodologies of the social and behavioral sciences has been employed in pursuit of measuring and understanding crime and criminals. We begin with attempts to measure crime based on actions of law enforcement and other criminal justice agencies.

Many countries now collect data on crimes known to the police, persons arrested, persons appearing before juvenile and adult courts, and persons convicted and sentenced to various forms of correctional sanctions. However, because data collection and reporting systems reflect the laws and practices of the jurisdiction with which they are associated, they vary considerably in scope and quality. Another challenge is that official systems of common crime reporting are designed primarily to serve institutional purposes, such as documenting police or court activity, justifying budget requests, and evaluating institutional performance, rather than to satisfy scholarly interests.

A great deal of scholarly as well as official effort has gone into standardizing definitions and reporting systems in order to enhance their comparability and research usefulness. Principles guiding efforts to establish these systems focus primarily on distinctions between *crimes against persons* and *property crimes* and on assumptions concerning data reliability. The further removed from the behavior being reported, the less reliable it is assumed to be as a measure of that behavior; and because officials such as the police must rely on victims and others to report most crimes, crimes that are likely to be reported are the most reliable as measures of those crimes. Thus, measures based on serious *crimes reported to the police* (or in surveys of victims of crime) are favored when the object of study is the distribution of crimes. For purposes of studying offenders, data on *persons arrested* for serious crimes are regarded as superior to data on persons incarcerated for those crimes.

Dissatisfaction with official data sources has motivated the search for other measures of crime. *Self-report* surveys of individual involvement in criminal/delinquent behaviors and of victimization were developed in the US in the mid-twentieth century and continuously are being refined and solicited from better samples, over time, and in many countries. Data collection using this technique requires that sampled respondents estimate, retrospectively or prospectively, their offending and victimization experiences within a specified time frame. Convergence with official data sources usually is interpreted as evidence of differential behavior rather than discriminatory police treatment of racial/ethnic minority groups and lower-class populations. Self-reports have been criticized, however, for focusing disproportionately on trivial offenses, rather than on more serious forms of crime included in official data (Elliott and Ageton, 1980).

Because of their power and relative ease of administration and analysis, self-report survey research methods are the primary basis of most measures of official, self-, and victim-reported common crime. Self-reports are also employed in many tests of theories advanced to explain crime and in evaluations of the effectiveness of crime-control programs. Extending decades of findings based on cross-sectional research, *longitudinal* designs that follow participants over time have significantly enhanced understanding of the causes and consequences of crime throughout the life course.

A primary weakness of self-reports is that they are subject to multiple forms of distortion and error, including problems of recall and respondent exaggeration. Because they focus primarily on individuals, self-reports also omit crucial information about the micro- and macro-social contexts of crime and offending. Although typologies of criminal behavior systems have been proposed based on ethnographic studies of criminal techniques, careers, and lifestyles, none has succeeded in addressing the full range of criminal behavior. Importantly, self-reports and other research methods have demonstrated that most common crimes are committed not by specialists but by persons who vary the crimes they commit, contingent on their needs, abilities, opportunities, and internal and external constraints. Such versatility in offending poses special challenges to efforts to control crime through selective incapacitation of specific *types* of offenders.

Historically, the primary data employed by sociologists have been official *crime rates*. This *macrosocial* level of explanation seeks to determine what it is about groups, organizations, institutions, neighborhoods, and communities – their structural conditions, cultures, and subcultures – that leads to aggregate differences in crime and other social problems. Sociologists who focus on the *individual* level of explanation, by contrast, study the effect of social structural and cultural factors, social interaction, and learning experiences on individual behavior. They differ among themselves, however, in the extent to which they assume crime to be innate or the product of learning in the course of life experiences, i.e., 'persistent heterogeneity' in the population versus 'state-dependent' offending (cf. Akers, 1998; Gottfredson and Hirschi, 1990). A third level of explanation, the *microsocial*, focuses on criminal behavior as a function of the characteristics of situations and ongoing interaction among social actors. Arguably, this level of explanation may serve as one means to bridge the macro- and individual-levels of explanation. Research and theory at this level are based largely on ethnographic and other means of directly observed behavior, including laboratory experiments (Horne and Lovalgia, 2008).

The Social Distribution and Explanation of Crime

The basic data for sociological explanations of crime rest heavily on how crimes are distributed among such social categories as age, gender, social class, racial, and ethnic status, over time and in differing social and cultural locations. Most common crime in all modern societies is committed primarily by young males between the ages of 15 and 25, for example, and is concentrated among those who are disadvantaged by economic and/or racial and ethnic status (Peterson, 2012). Two types of crimes are major exceptions: (1) the aforementioned 'white-collar' crimes, first identified as crimes committed by corporations or persons of high status but later focusing primarily on violations of fiduciary responsibility or public trust; and (2) the broad spectrum of crimes included under the rubric of genocide, human rights violations, terrorism, and war crimes. The distinguishing character of these exceptions is that neither is encompassed by criminal law, the former because its legal status is established in regulatory as well as criminal law, the latter because it is a product of international law as well as the criminal statutes of nation states. The *consequences* of these crimes have evolved in direct relationship with legal definitions and social control actions within and between nations.

These and other expansions of legally proscribed behaviors have made the study of crime more truly interdisciplinary and

even more closely allied with law and its enforcement. Although sociologists are among the most persistent and innovative in this respect (see, e.g., Sampson, 2012; Tittle, 1995; Wikström, 2012), extension of criminal behaviors beyond common crimes, interdisciplinary collaborations, and methodological innovations make for an extraordinarily complex mix of explanatory approaches.

Several early strands of sociological research nevertheless form the basis of much current theorizing about crime. The first, building on the earlier work of the moral statisticians, sought to establish and interpret the social distribution of crime and delinquency. Following the 'Chicago School' framework of sociological and ecological theory developed by Thomas and Znaniecki (1920) and Park and Burgess (1924), Clifford Shaw and his associates (Shaw et al., 1929; Shaw and McKay, 1942) related crime and delinquency rates to social conditions in community areas in the city of Chicago and elsewhere. Their chief finding – that rates of delinquency were highest in local communities characterized by poverty and high rates of population heterogeneity and turnover – and the theory advanced to account for it – social disorganization and its successors – remain of vital importance (Sampson, 2012).

Fundamentally a control theory, social disorganization theory hypothesized that socially disorganized areas lacked the means of formal and informal social control. At both macrosocial and individual levels, the basis for control lay in processes considered to be vital to the supervisory capabilities of communities, viz., population turnover and heterogeneity, which weaken social institutions, networks of primary relationships, and channels of communication.

Although study of the social distribution of crime took on a life of its own, social disorganization theory was soon challenged by research documenting highly organized groups and institutions in communities with high rates of crime and delinquency (Suttles, 1968; Whyte, 1943; see also Patillo, 2007). Revisions of the theory thus focused on the nature of relational networks in these organizations and institutions and on their ability to mobilize internal support for goals such as crime control (Bursik and Grasmick, 1993; Sampson, 2012). Complementing and extending William Wilson's (1987) highly influential account of the 'truly disadvantaged' urban poor, this *systemic* perspective once again brought together scholarly work on social problems such as public health, cognitive development, drug abuse, and initiation of sexual behaviors as well as a variety of types of crime.

The systemic perspective advanced by Sampson and colleagues extended observations by James Coleman (1988) and others regarding social capital to demonstrating that 'collective efficacy' – intergenerational closure and reciprocal exchange in support of shared control of children – was effective in controlling youth violence. Importantly, close interpersonal relationships were not the primary basis for this finding; most important, rather, were mutual trust and willingness to intervene in common pursuit of neighborhood social control.

The theory of collective efficacy has been supported for problems other than crime and in places other than Chicago, although Sampson (2012: p. 167) notes that "Latin American countries may prove to be the exception." Moreover, contrary to Wilson and Kelling's (1982) 'broken windows' thesis, which attributes crime to objective signs of disorder and the message sent to potential offenders that neighborhood social control is lacking, collective efficacy and historically shared (and often racially tinged) neighborhood perceptions and meanings have been found to predict crime and poverty rates better than either prior crime or observed physical and social disorders (Sampson, 2012). The broken windows thesis nonetheless has profoundly influenced urban policy, as evidenced by recent police crackdowns on 'public incivilities,' such as loitering, vandalism and other petty crimes, and hooliganism.

A second major line of theory developed from symbolic interaction foundations and observations of the group nature of delinquent behavior and the effects on delinquent boys of legal definitions and enforcement actions. Case studies suggested ways in which children learn delinquent behaviors from one another and the manner in which, through encounters with the law and other authority figures, they learn they are *defined* as delinquent. Evidence of this sort led Edwin Sutherland (1939) to focus on the process by which individuals become criminals, a process he called 'differential association.' Recognition that definitions and labeling as criminal (in law, in interaction with others, and by authority figures) had behavioral consequences supported social learning, labeling, and microsocial perspectives stressing the importance of interactions with others in the course of socialization and in the production of particular behaviors. Sutherland, reacting against multifactor theories that mechanistically related factors associated with offending as 'causes' of crime, applied differential association to crimes that were significantly different from common crime, notably white-collar and professional crimes. Extension of the theory to differential social organization addressed macrosocial issues and enhanced application to sociological interests in groups, organizations, and institutions.

Yet another major sociological perspective on crime causation appeared with the publication of Robert Merton's (1938) seminal theory of 'Social structure and anomie,' which focused on the consequences of disparities between culturally emphasized goals and structural limitations on the means for their achievement. Elaborating Durkheim's concept of anomie, Merton noted that the cultural emphasis on achievement of economic success posed special problems for those who were disadvantaged as a result of their social class, racial, or ethnic status. He predicted that, for these populations, the most likely 'modes of adaptation' were 'innovation' using illegitimate means or complete rejection of cultural goals (e.g., 'retreatism' through participation in a drug use subculture), as opposed to 'conformity' with conventional rules of behavior.

Theoretical elaboration of anomie theory, as Merton's formulation has come to be known, has taken a variety of forms. Cloward and Ohlin's (1960) 'opportunity theory' joins anomie theory with both social disorganization and differential association to explain variations in delinquent subcultures. Messner and Rosenfeld (1994) theorize that the domination of institutional life by economic institutions weakens the ability of the family, school, and other institutions to control crime. Additionally, the individual-level counterpart to their systems level theory, institutional-anomie theory, implies that "the risks of committing violent crimes will be high for actors: (1) who prioritize economic roles over non-economic

roles (perceptually and behaviorally); (2) who are insensitive to the moral status of the means of action; and (3) whose enactment of economic roles is high on the 'marketness' continuum" (Messner et al., 2008: p. 174). Still others shift the focus of Merton's formulation to individual strain, viewing crime as a means for reducing psychological strain caused by failure to achieve a variety of goals, such as autonomy, status, respect, and reciprocity in relationships with others (Agnew, 1999).

While much sociological theory about crime continues to be based on these theoretical traditions, methods and perspectives that are not explicitly tied to them may serve as important conceptual bridges. 'Routine activities,' a macrolevel perspective, explains rates of crime in terms of the convergence of motivated offenders, the absence of capable guardians, and suitable targets for criminal victimization (Cohen and Felson, 1979). Similarly, at the microsocial level, the routine activity theory of general offending (Osgood et al., 1996) and the 'group process' perspective (Hughes and Short, 2005; Short and Strodtbeck, 1965) seek to identify mechanisms – such as status management and time spent in unstructured and unsupervised socializing with peers – that produce different types of crime under varying conditions.

As support for large-scale data sets, multimethod studies, and experimental programs designed to control crime increases, and compensating strengths and limitations of research methods become apparent, theories at each level of explanation are enriched and the efficacy of policy evaluations strengthened. Sampson demonstrates that findings from field experiments oriented toward individual choice and behavior, for example, neglect important neighborhood- and community-level effects (cf, Ludwig et al., 2008; Sampson, 2012). Using a variety of methodologies, he develops the argument that "neighborhoods, ecological, and other collective phenomena demand their own measurement" – ecometrics, in contrast to psychometrics, which focus on individual differences – "since they are not stand-ins for individual-level traits" (2012: p. 60). By this logic, even findings from laboratory experiments, long considered to be the 'gold standard' of research, require contextual specification.

Issues Related to Crime Control

Crime control brings the sociological study of crime full circle, back to the law. The effects of law, law enforcement, and the imposition of legal sanctions on offenders and others are major sociological concerns. Although the central assumption of deterrence theory – i.e., that human beings are rational actors whose criminal behavior can be influenced by punishment – remains at the heart of many modern systems of law and justice, few firm conclusions have resulted from extensive research on the effects of legal sanctions on criminals (general deterrence) or on those who might offend (specific deterrence). Public concern with rising crime rates during the second half of the twentieth century resulted in higher rates of incarceration of convicted offenders which, despite falling crime rates, accelerated during the last quarter of the century. A National Research Council report that concluded that increments of crime control from incapacitation of offenders were modest (the size of the inmate population in the United States doubled over a decade, resulting in an estimated reduction in crime of 10%), extremely expensive, and socially destructive remains valid (Blumstein et al., 1986). Evaluations of other 'get tough' measures of crime control reach similar conclusions.

As crime increasingly has become an issue of political contention, the relationship between knowledge about crime, criminals, and crime control has become controversial among scholars, criminal justice professionals, and the public (Garland and Sparks, 2002). Efforts to bridge these sometimes-conflicting interests nevertheless continue, theoretically and in the form of experimental programs supported by governmental agencies in many countries. A prime example is Jon Braithewaite's (1989) Reintegrative Shaming Theory and related 'restorative justice' practices involving offenders, victims, and criminal justice and community members.

Challenges of Change and Continuity

Expansion of behaviors defined as criminal requires close attention to legal definitions and mechanisms of discovery, as well as enforcement options and practices. Major developments have occurred in international law and tribunals designed to control offenses, such as trafficking in prostitution, illegal migration, offenses committed by global corporations, and state involvement in genocide, war crimes, and human rights violations (Hagan and Greer 2002; Savelsberg, 2010). Traditional distinctions continue to change between individual and corporate behavior defined as crime and in the behavior of nations that, although harmful to citizens of those nations, previously was considered beyond the purview of other nations. In addition, new laws and new rules evolve as global markets alter commercial activity.

Rapid social change, driven by advances in science and technology, creates new opportunities for crime even as it enhances crime control efforts. Communication and data transfer technologies, for example, increase opportunities for inter- and intra-agency collaboration while exposing these agencies to cybercrimes and other security breaches. New social media facilitate formation of interest groups and other collective behaviors while ironically promoting individual interests. Other technologies, such as nuclear and biotechnologies, lend themselves to widespread victimization by blackmail or terrorist attack.

These and other changes require conceptual as well as methodological and measurement innovation, as has occurred with social network analyses as well as multilevel modeling, both of which have raised new questions and informed old debates at the macrosocial, individual, and microsocial levels of explanations. There is, however, continuity as well as change in the contexts and processes that produce and reproduce crime. The challenge for sociologists of crime, lawmakers, and criminal justice scholars and professionals is to recognize and account for these diverging influences.

See also: Control, Social; Crime and Class; Delinquency, Sociology of; Law, Sociology of.

Bibliography

Agnew, Robert, May 1999. A general strain theory of community differences in crime rates. Journal of Research in Crime and Delinquency 36, 123–155.

Akers, Ronald. L., 1998. Social Learning and Social Structure: A General Theory of Crime and Deviance. Northeastern University Press, Boston, MA.

Blumstein, Alfred, Cohen, Jacqueline, Roth, Jeffrey A., Visher, Christy A., 1986. Criminal Careers and Career Criminals. National Academy Press, Washington, D.C.

Braithewaite, Jon, 1989. Crime, Shame, and Reintegration. Cambridge University Press, Cambridge, UK.

Bursik, Robert J., Grasmick, Harold G., 1993. Neighborhoods and Crime: The Dimensions of Effective Community Crime. Lexington Books, New York.

Cloward, Richard A., Ohlin, Lloyd E., 1960. Delinquency and Opportunity: A Theory of Delinquent Gangs. Free Press, Glencoe, IL.

Cohen, Lawrence E., Felson, Marcus, 1979. Social change and crime rate trends: a routine activity approach. American Sociological Review 44, 588–608.

Coleman, James S., 1988. Social capital in the creation of human capital. American Journal of Sociology 94, S95–S120.

Durkheim, Émile, 1951/1897. Suicide: A Study in Sociology (J.A. Spaulding, G. Simpson, Trans.). The Free Press, Glencoe, IL.

Elliott, Delbert S., Ageton, Suzanne S., 1980. Reconciling race and class differences in self-reported and official estimates of delinquency. American Sociological Review 45 (1), 95–110.

Garland, David, Sparks, Richard, 2000. Criminology and Social Theory. Oxford University Press, Oxford.

Gottfredson, Michael R., Hirschi, Travis, 1990. A General Theory of Crime. Stanford University Press, Stanford, CA.

Guerry, André-Michel, 1833. Essai sur la statistique morale de la France. Crochard, Paris.

Hagan, John, Greer, Scott, 2002. Making war criminal? Criminology 40 (2), 231–264.

Hagan, John, Rymond-Richmond, Winona, 2009. Darfur and the Crime of Genocide. Cambridge University Press, Cambridge.

Horne, Christine, Lovalgia, Michael J., 2008. Experiments in Criminology and Law: A Research Revolution. Rowman & Littlefield, Lanham, MD.

Hughes, Lorine A., Short Jr., James F., 2005. Disputes involving youth street gang members: micro-social contexts. Criminology 43 (1), 43–76.

Ludwig, Jens, Liebman, Jeffrey B., Kling, Jeffrey R., Duncan, Greg J., Katz, Lawrence F., Kessler, Ronald C., Sanbonmatsu, Lisa, 2008. What can we learn about neighborhood effects from the moving to opportunity experiment? American Journal of Sociology 114, 144–188.

Merton, Robert K., 1938. Social structure and anomie. American Sociological Review 3, 672–682.

Messner, Steven F., Rosenfeld, Richard, 1994. Crime and the American Dream. Wadsworth, Belmont, CA.

Messner, Steven F., Thorne, Helmut, Rosenfeld, Richard, 2008. Institutions, anomie, and violent crime: clarifying and elaborating institutional-anomie theory. International Journal of Conflict and Violence 2 (2), 163–181.

Osgood, D. Wayne, Wilson, Janet K., O'Malley, Patrick M., Bachman, Jerald G., Johnston, Lloyd D., 1996. Routine activities and individual deviant behavior. American Sociological Review 61, 635–655.

Park, Robert E., Burgess, Ernest W., 1924. Introduction to the Science of Sociology. University of Chicago Press, Chicago, IL.

Patillo, Mary, 2007. Black on the Block: The Politics of Race and Class in the City. University of Chicago Press, Chicago, IL.

Peterson, Ruth D., 2012. The central place of race in crime and justice. Criminology 50, 303–327.

Quetelet, Adolphe, 1831. Recherches sur le penchant au crime aux différents âges. Hayez, Brussels.

Quetelet, Adoplpe, 1835. Sur l'homme et le développement de ses facultés, ou Essai d'une physique sociale. Bachelier, Paris.

Sampson, Robert J., 2012. Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago Press, Chicago, IL.

Savelsberg, Joachim J., 2010. Crime and Human Rights: Criminology of Genocide and Atrocities. Sage, Los Angeles, CA.

Shaw, Clifford R., McKay, Henry D., 1942. Juvenile Delinquency and Urban Areas. University of Chicago Press, Chicago, IL.

Shaw, Clifford R., Zorbaugh, Harvey, McKay, Henry D., Cottrell, Leonard S., 1929. Delinquency Areas. University of Chicago Press, Chicago, IL.

Short Jr., James F., Hughes, Lorine A., 2007. Criminology, criminologists, and the sociological enterprise. In: Calhoun, C. (Ed.), Sociology in America: A History. University of Chicago Press, Chicago, pp. 605–638.

Short, Jr., James, F., Strodtbeck, Fred L., 1965. Group Process and Gang Delinquency. University of Chicago Press, Chicago, IL.

Sutherland, Edwin H., 1939. Principles of Criminology. Lippincott, Philadelphia, PA.

Suttles, Gerald, 1968. The Social Order of the Slum: Ethnicity and Territory in the Inner City. University of Chicago Press, Chicago, IL.

Tarde, Gabriel, 1886. Le Criminalité Comparée. Félix Alcan, Paris.

Thomas, William I., Znaniecki, Florian W., 1920. The Polish Peasant in Europe and America. Richard Badger, Boston, MA.

Tittle, Charles R., 1995. Control Balance: Toward a General Theory of Deviance. Westview Press, Boulder, CO.

Whyte, William F., 1943. Street Corner Society: The Social Structure of an Italian Slum. University of Chicago Press, Chicago, IL.

Wikström, Per-Olof H., Treiber, Kyle, Oberwittler, Dietrich, Hardie, Beth, 2012. Breaking Rules: The Social and Situational Dynamics of Young People's Urban Crime. Oxford University Press, Oxford.

Wilson, William J., 1987. The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy. University of Chicago Press, Chicago, IL.

Wilson, James Q., Kelling, George, 1982. Broken windows: the police and neighborhood safety. Atlantic 127, 29–38.

Crime Trends and Debates

Patricia L McCall and Joshua A Hendrix, North Carolina State University, Raleigh, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The reasons offered by criminologists for the greatest crime decline in the twentieth century, occurring between the early 1990s and 2000, have been varied and have led to a number of debates in the field. The crime decline during this period, explanations offered to account for this trend and related debates are the focus of this article. We discuss the demographic influence of age structure and ethnic immigration, incarceration, police forces and strategies, the waning crack-cocaine markets, the economy and abortion laws. No one of these by itself can account for the crime decline; but they all likely contributed to some extent.

Crime trends capture the attention of the media, public and politicians who demand to know what drives criminal behavior. It is the challenge for criminologists to account for changes in crime rates over time. When crime rates across the U.S. began climbing in the 1960s and into the 1970s, a number of criminologists accounted for this as the result of the relatively large baby boom population 'coming of age' – that is, the crime-prone ages of juveniles and young adults. This explanation seemed to account for a great deal of the variation in crime trends during that period. The crime trends of the mid-1980s were more baffling. During a period when the baby boom generation was 'aging out' of crime-prone years, the crime rate began rising once again. These rising crime trends also were occurring during a period when 'get tough on crime' policies had been implemented and a growing proportion of the adult population was being incarcerated. In retrospect, many criminologists agreed that, in large part, the increasing crime trends in the 1980s and early 1990s were due to the crack-cocaine trade and to gun-carrying youth who had been indoctrinated into the drug trade especially after many adult drug traffickers were arrested and imprisoned. The reasons offered by criminologists for the greatest crime decline in the twentieth century that occurred between the early 1990s and 2000 have been varied and have led to a number of debates in the field. The crime decline during this period and related debates are the focus of this article. Homicide is the primary crime type used for this discussion as it represents the most serious and best measured of street crimes, and the FBI's Uniform Crime Report is the source for these crime data.

Crime Decline: Trends and Patterns

Regardless of the data source – either the Federal Bureau of Investigation (FBI)'s Uniform Crime Reports or the National Crime Victimization Survey, index crime rates declined more than 40% between the early 1990s and 2000 to lows that had not been realized since the mid-1960s. This plummeting trend represented the most prolonged and the greatest magnitude of decline in crime since 1933 when the FBI began keeping record of national crime rates. **Figure 1** displays homicide and robbery trends for the period between 1973 and 2010 (U.S. Department of Justice). These predatory crime rates share remarkably similar trends over this period. Though not so closely shared among other serious crime series, all of the major index crimes declined during the last decade of the twentieth century.

Figure 1 shows that the two decades leading up to the crime decline were characterized primarily with an oscillating pattern. Homicide peaked in 1980 at 10.2 per 100 000 population – its highest rate in the twentieth century, then fell 23% in the next 5 years to 7.9. Between 1985 and 1991, the homicide rate rose almost 25% to 9.8. Unlike homicide trends, robbery peaked in 1990, otherwise mirroring the homicide pattern since 1980. Since 2000, the end of the great American crime decline, these rates stabilized through 2007 after which they declined until 2011, our latest available data point – dropping about 20%.

Remarkably, these general trends were not solely driven by a few of the largest U.S. cities but were found in cities across the country although the magnitude of the crime rates varied across cities (McCall et al., 2011). City-level studies of this period found that larger cities were more likely to exhibit this pattern and that the timing of the rise and fall of crime rates was earlier for more highly and densely populated cities with higher levels of deprivation (Messner et al., 2005). Messner and his colleagues (2005) also found a geographical pattern among those cities following the national trend – that is, the upturn in homicide rates began in the early 1980s on the west coast and the downturn ensued among east coast cities with patterns reflecting a dispersion of effects.

The close parallel between the homicide and robbery series shown in **Figure 1** suggests that common forces drive these trends, but criminologists have yet to arrive at any consensus on what these causal influences may be. The structural and situational factors that have been proposed as plausible drivers of these trends and related scholarly debates are the subject of this article. Before moving to those forces, we provide a brief discussion of the age- and race-specific homicide trends between 1980 and 2008.

Age-Specific Homicide Offending Trends

By disaggregating the total homicide rates shown in **Figure 1**, we discover that certain age groups, in particular, were driving these crime trends. In **Figure 2**, age-specific homicide offending rates for select age groups are displayed (U.S. Department of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45050-6

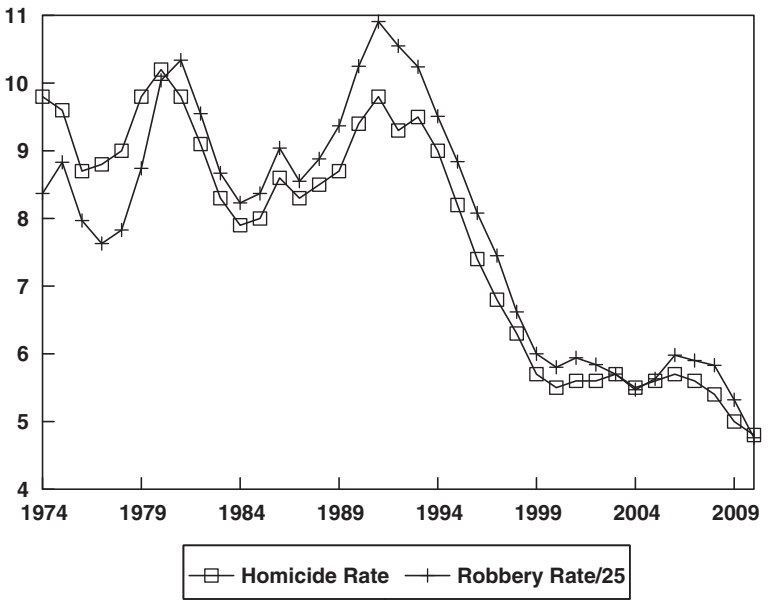

Figure 1 U.S. Homicide and Robbery* Rates, 1974–2010. *Note: Robbery rate divided by 25. Source: Department of Justice, FBI's Uniform Crime Reports.

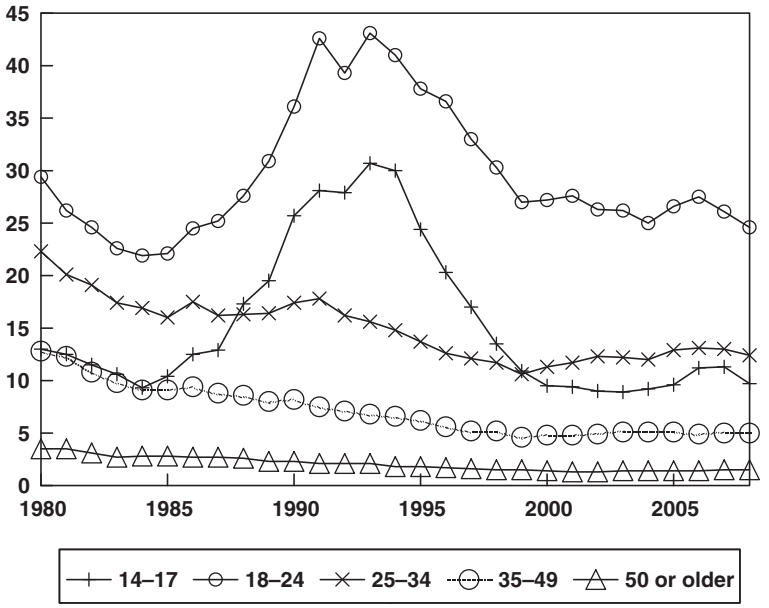

Figure 2 Age-specific homicide-offending rates, 1980–2008. Source: Department of Justice, FBI's Uniform Crime Reports.

Justice, 2011). The distinct patterns among these age groups are quite striking as the younger populations exhibited remarkable changes in their homicide offending patterns. In particular, the 14–17 year olds began the period with rates comparable to the relatively lower rates of the 35–49 year olds, but then by 1988, surpassed the rates of the typically higher homicide offending rate group, the 25–34 year olds, and maintained that higher rate until 1999 – toward the end of the crime decline.

The highest offending group is the 18–24 year olds with the steepest growth and decline in their homicide trends during this 30-year period – almost doubling between 1984 and 1993 (from 21.9 to 43.1) and then falling by 42% to their low of 25 per 100 000 in 2004. The youngest age group saw similarly dramatic swings, beginning in 1984 with homicide rates at 9.3 and tripling by 1993 when homicide rates hit a high of 31 per 100 000 14–17 year olds. These rates grew as many young people were affected or caught up in the crack-cocaine markets and began arming themselves with handguns. Their homicide rates then dropped below the 1984 rates to their low of 8.9 in 2003. Interestingly, the two older age groups, 35–49 and 50 years and older, realized stable or decreasing homicide trends over 72% of this 30-year period. There are no clear-cut answers

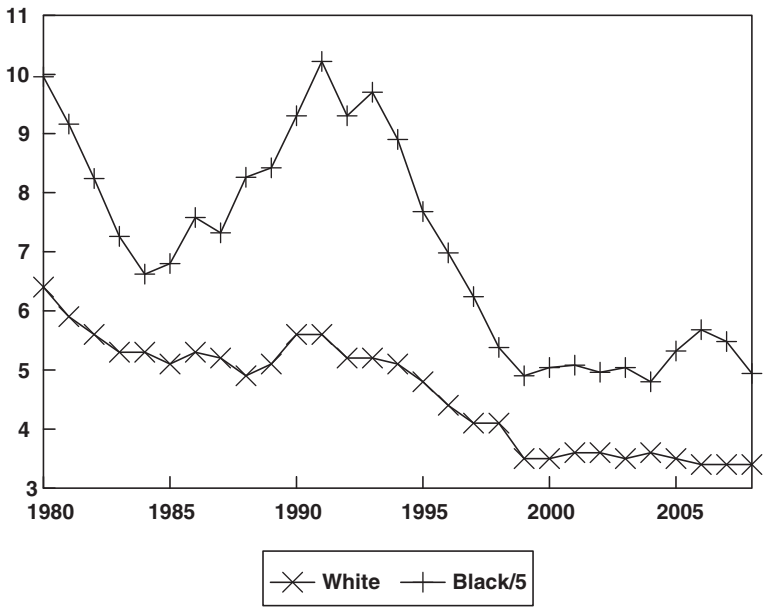

Figure 3 Race-specific (White and Black*) homicide-offending rates, 1980–2008. *Note: Black homicide offending rate divided by 5. Source: Department of Justice, FBI's Uniform Crime Reports.

to explain these shifting violent crime rates, and more puzzling to criminologists is what accounted for the unanticipated declines of the 1990s.

Race-Specific Homicide Offending Trends

Disaggregating total homicide rates by race indicate that the African American homicide rate has been considerably higher than that of whites since 1980. (Note that the African American rate has been divided by five so the variation in the white homicide rate is not dwarfed by the substantially higher minority rate.) Moreover, the white homicide rate has remained relatively more stable, fluctuating between 6.4 (highest rate, 1980) to 3.4 (lowest rate, 2008), compared with African American homicide, which fluctuated from 51 (10.2 in figure after being divided by 5) (highest rate, 1991) to 24.5 (lowest rate, 1999). The significantly steeper growth and decline in African American homicide rates suggests that the crime decline in the 1990s was disproportionately influenced by falling homicide rates among African Americans. Again, although a parsimonious example for why African American homicide rates fell dramatically during the 1990s does not exist, the following section sheds light on possible explanations, all of which contributed in part to this decline. (Figure 3).

Crime Decline: Explanations

The major hypotheses that have been proposed to explain the crime decline of the 1990s are delineated below. First, we discuss the demographic influence of age structure and ethnic immigration. Second, we present the role of criminal justice factors – such as incarceration, police forces and strategies, as well as gun control interventions – which may have contributed to the declining violent crime rates. Finally, we shift to social and economic explanations that have been offered, such as the waning crack-cocaine markets, the economic boom of the 1990s and the impact of legalized abortion that led to the drop in numbers of unwanted births who would have been in their crime-prone ages during the 1990s – about 20 years after the passage of those abortion laws. We conclude with summary discussions of each of these proposed explanations.

Age Structure

Criminologists widely acknowledge the existence of an 'age–crime curve,' or the tendency for individual criminality to peak in adolescence and early adulthood and decline thereafter (Hirschi and Gottfredson, 1983). The pervasiveness of the age–crime curve suggests that a larger proportion of young people in the total population should equate to a larger pool of criminal victims and offenders and thus a higher crime rate more generally. Early articles reported considerable support for this idea (see Sagi and Wellford, 1968), and the youth age structure trend was found to closely mirror the precipitous increase in crime rates of the 1960s and 1970s (besides a dip in 1975) and subsequent decline beginning around 1980 (Cohen and Land, 1987). Therefore, many criminologists explored this demographic force to explain the crime decline of the 1990s.

More recently, Marvell and Moody's (1991) extensive review identifies statistically significant and positive relationships between the relative size of the youth population and crime rates (i.e., homicide and burglary) in almost all studies under consideration. Some scholars suggest that future crime rates can be forecasted on the basis of the projected size and age composition of juvenile and adult populations. For instance, Blumstein et al. (1980) accurately

predicted that crime rates would peak in 1980 as the postwar baby boom birth cohort moved out of high crime ages.

Although the aforementioned studies would suggest that the crime drop of the 1990s had much to do with changes in age structure, some scholars express doubts. Zimring (2007) argues that New York City experienced only a slight reduction in the total number of males in their late teens to late twenties in the mid-1990s, yet also experienced a major crime decline. Others argue that when crime rates were declining in the late 1990s, the number of 15- to 24-year olds in the U.S. population was increasing and continued to do so into the twenty-first century while crime rates continued to decline (Blumstein and Rosenfeld, 1999, 2008; Levitt, 2004). Additionally, the effects of age composition on crime rates can be offset when age compositional changes are slow or when such changes occur along with other demographic changes, such as changes in the relative size of racial minority groups (Blumstein and Rosenfeld, 2008; Levitt, 2004).

Also problematic is that a number of forecasts prior to the crime drop based on age composition projections were largely unsuccessful. Steffensmeier and Harer (1987) predicted significant declines in violent and property crime from 1980 to 2000 and thus unanticipated the crime rise beginning in the mid-1980s. Others predicted that anticipated massive increases in the proportion of youths in the late 1990s would produce a '"crime bomb' and the emergence of young, ultraviolent 'superpredators' after the turn of the century (Dilulio, 1996). However, proponents of the crime storm based their predictions solely on the size of the youth population and did not take into account relative sizes of middle-aged and elderly populations (Steffensmeier and Harer, 1999). Additionally, forecasts are strongly influenced by recent crime trends, which produce projections that are based on expected, not actual age-specific crime rates. Thus, although age composition is certainly an important covariate of crime rates, a large share of youth in the population does not inevitably translate into higher crime rates. In fact, the recent increase in the 15- to 29-year old population from 2000 to 2010 has been accompanied by a leveling and then declining crime trend. Hence, evidence points to other forces that define the crime decline of the 1990s.

Immigration

Conventional wisdom has supported the view that immigration promotes crime (Gurr, 1989). Social disorganization theory suggests that immigration influxes introduce cultural and language differences among community residents, and thus attenuates community bonds and informal controls and in turn enhances social disorder and crime levels (Shaw and McKay, 1969 [1942]). However, violent crime reductions in recent decades have coincided with the influx of foreign-born individuals (Simanski, 2005). Additionally, many U.S. cities with high concentrations of immigrants have relatively low crime rates and experienced some of the largest crime reductions in the 1990s (Martinez and Lee, 2000). A number of studies based on longitudinal and time-series data find that increases in immigration contributed to the crime decline in the 1990s (Ousey and Kubrin, 2009; Stowell et al., 2009; Wadsworth, 2010). Accordingly, it may no longer be justifiable to assume that immigration and diversity are automatic determinants of social disorganization and crime.

A number of explanations have emerged for why higher immigration levels may lead to lower levels of crime. Many scholars point out that immigrants are less likely to engage in criminal behavior than other racial groups; thus, a larger proportion of law-abiding citizens logically equates to a lower overall crime rate (Butcher and Piehl, 1998; Hagan and Palloni, 1998; Sampson and Bean, 2006). Others point out culturally-based protective factors such as marital/family stability, extended family, religiosity, employment networking, and collective efficacy among immigrant enclaves (Ousey and Kubrin, 2009; Sampson et al., 2005; Wadsworth and Kubrin, 2007). Some posit that immigration promotes positive changes in the general character of communities by stimulating the creation of new institutions such as churches and self-help groups (Martinez and Valenzuela, 2006). Logan et al. (1994) argue that low-skilled service markets heavily depend on immigrant labor, and thus reduce unemployment rates in immigrant communities. Finally, some scholars have pointed to a selection effect, which suggests that foreign-born individuals who come to the U.S. are relatively more motivated, talented, and law-abiding than those who do not (Martinez, 2002).

Despite empirical support, some scholars have been skeptical of the crime-reducing potential of immigrant populations. For instance, Skogan (2009) argues that two-thirds of the crime decline in Chicago since 1991 transpired in African American neighborhoods, whereas gang activity and crime increased in Hispanic neighborhoods. Accordingly, Skogan refutes the notion that immigrants exhibit lower levels of criminal behavior, but rather, that undocumented immigrants go to considerable lengths to avoid attention from law enforcement officials. Also, a number of cross-sectional (Reid et al., 2005) and longitudinal studies (Baumer, 2008; Butcher and Piehl, 1998) find that increases in immigration have had a negligible effect on recent crime trends. Finally, some studies find that although first-generation immigrants display lower levels of criminal involvement, criminal behavior is considerably more common among second- and third-generation immigrants (Hagan et al., 2008; Hagan and Palloni, 1999). Despite contradictory findings, there is enough evidence to suggest that growing immigration patterns in some areas contributed to declining crime rates during the 1990s. The immigrant population has continued to grow since the 1990s during a time when the crime trend plateaued until 2006 and subsequently declined.

Moving from demographic factors, many politicians and police chiefs have claimed that criminal justice practices have resulted in curbing crime rates during the crime decline period.

Incarceration

Beginning in the early 1970s, imprisonment rates in the U.S. soared. By the end of the twentieth Century, the incarcerated population had quadrupled with the number imprisoned climbing from 500 000 in 1972 to two million (Spelman, 2000). Incarceration could affect crime rates via incapacitation,

deterrence, or rehabilitation. Distinguishing among these forces has proven to be a difficult if not insurmountable statistical challenge to date (Blumstein et al., 1978). Scholars have long debated the effects of incarceration on crime rates – some arguing that criminal justice sanctions deter criminal offending (Ehrlich, 1973) and others arguing that prisons have a criminogenic effect (Vieraitis et al., 2007). Scholars have established a link between incarceration and crime rates (Levitt, 2004; Stemen, 2007; Zimring and Hawkins, 1973), and incarceration is listed among the major factors credited for declining crime trends of the 1990s (Levitt, 2004). By and large, the weight of evidence supports the dampening effect of incarceration rates on subsequent crime rates (as well as the positive effect of growing crime rates on imprisonment rates). Recent research by Levitt (2004) and Baumer (2008) suggest that imprisonment (the effect of stock and flow, respectively) accounts for a relatively large percentage of the drop in crime rates during the 1990s (estimates ranging from about 10 to 35% of the decline) making incarceration one of the substantial contributors to the crime decline. Studies of the elasticity of incarceration suggest that the increased effectiveness of incarceration in reducing crime rates may be due to the growth in the scale of imprisonment, the proportion of adults represented among those incarcerated and the focused efforts of law enforcement to get the more serious offenders off the street (Spelman, 2000). On the other hand, some scholars provide evidence of the diminishing returns of these rising imprisonment trends that suggests, after a point, growing imprisonment may actually increase crime rates as optimum deterrent and incapacitation effectiveness is surpassed (Liedka et al., 2006). There are growing concerns among scholars that the release of these prisoners (about 500 000 released in 2000) back into their home communities may have deleterious effects on crime trends (Rose and Clear, 1998).

Even though there appears to be a general consensus in the literature about the effect incarceration has had on reducing crime rates during the 1990s (Blumstein and Rosenfeld, 2008), researchers note that the incarceration rate was climbing long before the precipitous decline of the 1990s crime rates (Blumstein and Rosenfeld, 1998). Furthermore, since 2000, the incarcerated population continued to climb until 2009 when the numbers began to drop – the period when the crime trend leveled off, and renewed its decline. Scholars continue their work to hone the precision of statistical techniques used to assess this relationship; and the growing knowledge of the limitations of previously employed model specifications call into question the confidence some scholars have in inferences drawn from the current body of evidence (Spelman, 2008).

Police Force and Policing Strategies

A great deal of credit has been claimed by police commissioners and politicians for the role that the police had on the declining crime trends of the 1990s. Attention not only has been given to changes in the size of the police force but also to changes in the focus of policing efforts and policing strategies. Many police departments shifted from strictly reactive policing as they adopted new innovations in police tactics in the 1980s and 1990s. A myriad of strategies were implemented, including community-oriented policing, problem-solving policing, zero-tolerance policing, and targeting public disorder and weapons carrying. Therefore, the impact of law enforcement is among factors identified in the crime decline literature. Amidst claims of law enforcement's successful efforts in reducing crime, relatively few carefully designed studies have been conducted to test this relationship during the years of the crime decline. Scholars have called into question the influence law enforcement efforts have on curbing crime because the widespread crime decline occurred in many U.S. cities where such tactics had not been introduced (Blumstein and Rosenfeld, 1998). Furthermore, a direct link between policing and crime rates is difficult to determine as many cities' crime rates were already declining by the time policing initiatives were introduced (Eck and Maguire, 2000).

Eck and Maguire (2000) had conducted a thorough review of research on the impact of the number of police personnel on crime and found mixed evidence even after assessing only the more rigorous studies that addressed methodological issues such as simultaneity. Nevertheless, more recent research focusing on the period of the crime decline has produced evidence that the size of the police force (officers on the street) had a dampening, although limited, effect on violent crime rates (Baumer, 2008; Levitt, 2004).

Research examining more targeted police initiatives has provided some support for their contribution to falling crime rates, such as Boston's Operation Ceasefire (Braga et al., 1999), New York City's Compstat (Messner et al., 2007), Richmond's Project Exile (Raphael and Ludwig, 2003), Kansas City's Gun Experiment (Sherman et al., 1995) and St. Louis' Firearm Suppression Program (Rosenfeld and Decker, 1996). Rosenfeld et al., (2005) compared police strategies and their effectiveness in reducing violent crime trends among the programs in Boston, New York City, and Richmond along with a sample of cities that did not implement these police tactics. They found modest support (only for Richmond) that new policing interventions differentiated falling violent crime rates across U.S. cities during the 1990s. The effect of policing on the crime decline warrants further scholarly attention in order to provide definitive evidence about the nature and magnitude of this relationship (National Research Council, 2004). Eck and Maguire (2000) concluded that law enforcement's focused attention on high-crime neighborhoods likely contributed to the reduction in crime rates; and also that overall effectiveness of policing is realized typically as part of a coordinated network of institutions, including formal (e.g., courts and schools) and informal (families and churches) institutions. Other evidence suggests that successful police tactics are those that target specific hot spots and violent offenders (National Research Council, 2005). Another review suggests that the diverse police strategies introduced in the 1990s likely contributed very little to the crime drop (Eck and Maguire, 2000).

The majority of the decline in homicides during the 1990s was accounted for by gun-related homicides (Fagan et al., 1998); therefore, some have given credit to gun control legislation and police strategies that targeted illegal possession of guns and gun-related crimes during this period (Sherman and Rogan, 1995).

Like most studies of police interventions, initiatives toward gun control and the reduction of gun-related violence were seldom evaluated using systematic and multivariate methodological techniques. Most reports are descriptive in nature and only a small percentage of these studies would pass the basic requirements of sound statistical analysis for attributing success to these programs (Wellford, 2005: 674). As stated among the concluding remarks from a critical review of firearms and violence studies published by the National Research Council with regard to place-based policing of firearms and targeted violent gun offenders:

> … in the committee's view, the existing data and methods make it difficult to assess how … policing programs affect crime. Researchers cannot hope to credibly control for the many confounders that influence violence and crime using simple time-series comparisons (National Research Council, 2005: 241)

Although a number of these initiatives were introduced during the 1990s, it is difficult if not impossible to separate the unique effects of these policing and firearm strategies from other concurrent effects on the declining homicide rates without different methodological designs and statistical tests (National Research Council, 2005). Without clear assessments of these programs' effectiveness in reducing crime, it is certainly difficult to determine the extent to which these programs had an impact on reducing crime rates in the 1990s net of the effects of other influences. Some scholars attribute limited significance to interventions such as policing strategies, gun control, and concealed weapons laws on influencing the crime decline (Levitt, 2004).

Aside from criminal justice policies and practices, criminologists have examined the influence of social, economic, and situational forces to explain the crime decline.

The Rise and Fall of the Crack-Cocaine Market

The link between U.S. crack-cocaine markets and violent crime rates is a salient theme in the crime decline debate (Blumstein and Rosenfeld, 1998; Levitt, 2004). Blumstein's (1995) 'crack-cocaine thesis' argues that the intensified demand for crack-cocaine accounts for the substantial increase in homicides and robberies in urban America during the 1980s and early 1990s and that violent crime rates declined with the increase in drug-related arrests and the subsiding crack-cocaine demand in the 1990s. The following section discusses characteristics of the crack-cocaine market that influenced an increase, followed by a subsequent decrease in violent crime.

In an illegal drug market, disputes involving drug transactions, territories, and other matters are unlikely to be mediated by the police or other formal modes of control. Consequently, the growth in the crack-cocaine market is purported to have increased violent crime because it heightened the use of informal and violent forms of conflict resolution. Beyond this, crack-cocaine transactions are likely to transpire between strangers, a characteristic that may heighten feelings of suspicion and lead to a "shoot first and ask questions later" mentality (Ousey and Lee, 2007). More importantly, intensified law enforcement efforts that incarcerated early forerunners of crack markets created a 'replacement effect,' whereby younger and more inexperienced individuals took over control of crack-cocaine markets. Compared with their predecessors, these replacements were more reckless, more likely to carry handguns, and quicker to resort to lethal forms of conflict resolution (Blumstein, 1995). As the crack-cocaine trade became more lucrative, it also became more violent - characteristics that regularized gun-toting not only among those directly involved in trafficking, but also as a means of self-defense among uninvolved individuals who lived in geographical areas deeply implicated in the crack-cocaine trade (Blumstein, 1995).

Some scholars attribute the decline in violent crime in the 1990s to decreasing levels of crack-cocaine market activity following intensified law enforcement efforts in addition to the declining numbers of new users as information regarding the negative consequences of crack-cocaine involvement became widespread (Blumstein and Rosenfeld, 1999). Ousey and Lee (2007) argue that the violent crime decline can be attributed not only to the subsiding of drug markets, but also to the attenuation of the drug market-lethal violence relationship, which had created a "kinder and gentler" drug market than that of the 1980s and early 1990s. Specifically, as marijuana became the drug of choice among young drug users in the 1990s, the crack-using population became increasingly characterized by older, less impulsive, and less violent individuals compared with their younger counterparts during the peak of the "crack era" in the late 1980s (see Johnson et al., 2000). At the same, drug transactions were becoming significantly less likely to transpire in visible street markets, instead moving to private, indoor settings among trusted friends or regular customers (Curtis, 1998).

Several studies indicate that U.S. cities that experienced high levels of crack activity or significant increases in crack use in the 1980s also experienced momentous upsurges in violent crime (Baumer et al., 1998; Messner et al., 2005; Messner et al., 2007). However, other studies find less definitive support. Kelling and Sousa (2001) find that the declining use of crack cocaine, as represented by the number of hospital discharges for cocaine-related incidents, is not associated with the decline in violence in New York City. Corman and Mocan (2000) employ a time-series analysis of crime and drug use in New York City, and report that murder and assault rates are not related to changes in the rate of drug use. The authors speculate that the significant increase in the supply of crack cocaine during the 1980s created cocaine surpluses among drug dealers, and thus reduced violent interactions that were once again prevalent among drug dealers due to a more limited supply of available cocaine.

Economic Growth

Improving economic conditions during the 1990s have been proposed to account for the declining crime rates during that decade. Criminologists have theorized the link between economic conditions and crime rates arguing that offenders are motivated by economic strain (Merton, 1938) and/or that routine activities facilitate opportunities for crime during economic upswings (Cohen et al., 1980). Typically using

unemployment rates as their economic indicator, criminologists have generated mixed support for the relationship between economic conditions and crime rates (Chiricos, 1987; see Cantor and Land, 1985 for countervailing forces explaining these mixed findings). Some individuals who lack lucrative, legitimate employment opportunities (especially youth and minorities) may turn to illegitimate opportunities for economic gain when those illicit opportunities present themselves as was the case in many inner cities during the crack-cocaine epidemic of the mid-1980s extending to early 1990s (Freeman, 1996).

Between 1991 and 2001, gross domestic product (GDP) per capita increased by 30% in the U.S.; and not only did the unemployment rate drop from 6.8 to 4.8%, but real wages increased and legitimate labor market opportunities improved. Even though many of these job opportunities were low-wage, unskilled positions, they provided many youth with a safer work environment than what the illegal drug market had afforded them during the heyday of the crack-cocaine market (Blumstein and Rosenfeld, 1998). Yet, empirical support found for the relationship between economic expansion of the 1990s and the declining crime rate has been modest and applies primarily to property crimes and robbery. Researchers examining this relationship have operationalized economic forces using measures such as the unemployment rate (Gould et al., 2002), business cycles (Arvanites and Defina, 2006), and real wages (Gould et al., 2002). There is essentially no systematic support for the effect of improved economic conditions and the declining homicide trends of the 1990s. Levitt (2004) argues that the economy does not have a direct impact on crime rates but may work largely through increased resources available to criminal justice agencies (police and prisons) during these times of economic upswings. This may account for the fact that the economic crises of the twenty-first century have not been accompanied by rising crime rates, but instead with a flattened trend through 2006 followed by another crime decline.

Another indicator of economic health, consumer sentiment, has been proposed as a better predictor of crime trends because it captures the subjective aspects of economic conditions. Rosenfeld and Fornango (2007) find robust evidence for the effect of their consumer sentiment summary index on property crime and robbery trends. Although their study does not test the association between consumer sentiment and homicide trends, Blumstein and Rosenfeld (2008) argue that the close parallel between homicide and detrended consumer sentiment index trends and, likewise, the similarity between the robbery and homicide trends during the last few decades of the twentieth century suggest there may exist a link between economic conditions and homicide trends using this index of consumer sentiment. Although there appears to be consistent empirical evidence for a modest influence of economic forces on property crime and robbery trends, as yet there is no support for its impact on other violent crime trends.

Legalized Abortion

One of the most contentious explanations for the crime drop is Donohue and Levitt's (2001) "abortion dividend" thesis, which attributes up to half of the crime decline in the 1990s to the legalization of abortion in the early 1970s. The authors argue that changes in abortion laws produced smaller cohorts of teenagers and young adults in the 1990s in addition to fewer numbers of unwanted children who would have been born to high-risk mothers (i.e., young, unmarried, low educational attainment and economic status) and in high-risk birth conditions (i.e., low birth weight, prenatal exposure to drugs and alcohol, neurological deficits) (Levitt, 2004). To support their claim, Donohue and Levitt argue that states that legalized abortion before the 1973 *Roe vs Wade* decision experienced crime declines earlier than other states, and states with higher abortion rates in the 1970s saw the most significant crime declines in the 1990s. Other studies also find that abortion laws contributed to crime declines in recent decades, although most find its effect to be much lower in magnitude than was estimated by Donohue and Levitt (Baumer, 2008; Berk et al., 2003).

The abortion-dividend thesis has also attracted a great deal of skepticism. Many scholars note that the timing of abortion laws and the crime decline in the 1990s were temporally out-of-synch (Rosenfeld, 2004). Specifically, they argue that crime declines occurred several years later than they should have based on the timing of abortion law implementation and also that increases in violent crime in the early 1990s were greatest among African Americans adolescents, although African American women were more likely to have abortions than other racial groups (Blumstein and Wallman, 2006). Also, Zimring (2007) points out that Canada did not remove restrictions on abortion until 1989, yet experienced significant crime reductions at the same time that the U.S. did. Others claim that the abortion-crime link is not causal, but rather is confounded with the rise and fall of crack markets (Blumstein and Wallman, 2006). For instance, Cork (1999) argues that even though states that legalized abortion earlier than 1973 (e.g., New York and California) did experience earlier starts to the crime drop; this was because these states underwent the rise and fall of the crack epidemic earlier than other states.

Conclusion

The major hypotheses that have been proposed to explain the crime decline of the 1990s are varied and some have generated considerable debate among scholars. The demographic factors including age structure likely have little direct bearing on the decline; and immigration trends may have contributed to some of the decline, but research suggests countervailing forces related to immigration may limit the extent to which these patterns accounted for the decline. Criminal justice policies – that led to unprecedented proportions of the population being incarcerated and to innovative policing strategies and gun control interventions – have been shown to contribute to the declining violent crime rates – in particular, incarceration trends. Unfortunately, limitations in research design of policing and gun control tactics leaves out the extent to which these factors curbed crime rates in question. The diminishing crack-cocaine markets may be linked to the crime decline although the widespread handgun availability that seemed interrelated with drug market activity showed no

sign of fading. The economic growth of the 1990s provided legitimate labor market opportunities to many youth who, in prior years, were attracted to the lucrative, though dangerous, illicit drug trade; yet research showed that economic forces impacted declining property and robbery crime trends rather than other personal crime trends such as homicide. The impact of legalized abortion was one of the explanations for the crime decline that generated lively scholarly debates in the literature, and some support emerged that it may have accounted for some of the crime decline – indirectly, if not directly. No one of these explanations by itself can account for the crime decline; but likely in combination – and some of which interacting – they all contributed to some extent. A body of literature accounting for crime trends since the turn of the twenty-first century has yet to emerge, but we can expect to revisit many of these social and economic forces when scholars attribute their influence to account for recent fluctuations.

See also: Age and Crime; Communities and Crime; Drugs: Illicit Use and Prevention; Economics and Crime; Guns and Criminals; Homicide; Immigration and Crime in the United States; Mass Imprisonment and Its Consequences; Property Crime: Victimization and Trends.

Bibliography

Arvanites, T.M., Defina, R.H., 2006. Business cycles and street crimes. Criminology 44, 139–164.

Baumer, E., Lauritsen, J.L., Rosenfeld, R., Wright, R., 1998. The influence of crack cocaine on robbery, burglary, and homicide rates: a cross-city, longitudinal analysis. Journal of Research in Crime and Delinquency 35, 316–340.

Baumer, E., 2008. An empirical assessment of the contemporary crime trends puzzle: a modest step toward a more comprehensive research agenda. In: Goldberger, A.S., Rosenfeld, R. (Eds.), National Research Council's Understanding Crime Trends: Workshop Report. Committee on Law and Justice, Division of Behavioral and Social Sciences and Education. The National Academies Press, Washington, DC, pp. 127–176.

Berk, R., Sorenson, S., Wiebe, D., Upchurch, D., 2003. The legalization of abortion and subsequent youth homicide: a time series analysis. Analyses of Social Issues and Public Policy 3, 45–64.

Blumstein, A., 1995. Youth violence, guns, and the illicit-drug industry. Journal of Criminal Law and Criminology 86, 10–36.

Blumstein, A., Cohen, J., Miller, H., 1980. Demographically disaggregated projections of prison populations. Journal of Criminal Justice 8, 1–25.

Blumstein, A., Cohen, J., Nagin, D., 1978. Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates. National Academy of Sciences, Washington, DC.

Blumstein, A., Rosenfeld, R., 1998. Explaining recent trends in U.S. homicide rates. Journal of Criminal Law and Criminology 88, 1175–1216.

Blumstein, A., Rosenfeld, R., 1999. Trends in rates of violence in the USA. Studies on Crime and Crime Prevention 8, 139–167.

Blumstein, A., Rosenfeld, R., 2008. Factors contributing to U.S. crime trends. In: Goldberger, A.S., Rosenfeld, R. (Eds.), National Research Council's Understanding Crime Trends: Workshop Report. Committee on Law and Justice, Division of Behavioral and Social Sciences and Education. The National Academies Press, Washington, DC, pp. 13–44.

Blumstein, A., Wallman, J., 2006. The crime drop and beyond. Annual Review of Law and Social Science 2, 125–146.

Braga, A.A., Weisburd, D.L., Waring, E.J., Mazerolle, L.J., Spelman, W., Gajewski, F., 1999. Problem- oriented policing in violent crime places: a randomized controlled experiment. Criminology 37, 541–580.

Butcher, K.F., Piehl, A.M., 1998. Cross-city evidence on the relationship between immigration and crime. Journal of Policy Analysis and Management 17, 457–493.

Cantor, D., Land, K.C., 1985. Unemployment and crime rates in the post-World War II United States: a theoretical and empirical analysis. American Sociological Review 50, 317–332.

Chiricos, T.G., 1987. Rates of crime and unemployment: an analysis of aggregate research evidence. Social Problems 34, 187–212.

Cohen, L.E., Felson, M., Land, K.C., 1980. Property crime rates in the United States: a macrodynamic analysis, 1947–1977, with ex–ante forecasts for the mid–1980's. American Journal of Sociology 86, 90–118.

Cohen, L.E., Land, K.C., 1987. Age structure and crime: symmetry versus asymmetry and the projection of crime rates through the 1990s. American Sociological Review 52, 170–183.

Cork, D., 1999. Examining space-time interaction in city-level homicide data: crack markets and the diffusion of guns among youth. Journal of Quantitative Criminology 15, 379–405.

Corman, H., Mocan, N., 2000. A time-series analysis of crime, deterrence, and drug abuse in New York City. American Economic Review 90, 584–604.

Curtis, R., 1998. The improbable transformation of inner-city neighborhoods: crime, violence, drugs and youth in the 1990s. Journal of Criminal Law and Criminology 88, 1233–1276.

Dilulio, J.J., 1996. They're Coming: Florida's Youth Crime Bomb. Impact spring pp. 25–27.

Donohue, J.J., Levitt, S., 2001. The impact of legalized abortion on crime. Quarterly Journal of Economics 116, 379–420.

Eck, J.E., Maguire, E.R., 2000. Have changes in policing reduced violent crime? an assessment of the evidence. In: Blumstein, A., Wallman, J. (Eds.), The Crime Drop in America. Cambridge University Press, New York, NY, pp. 207–265.

Ehrlich, I., 1973. Participation in illegitimate activities: a theoretical and empirical investigation. Journal of Political Economy 81, 521–565.

Fagan, J., Zimring, F.E., Kim, J., 1998. Declining homicide in New York: a tale of two trends. Journal of Criminal Law and Criminology 88, 1277–1324.

Freeman, R.B., 1996. Why do so many young American men commit crimes and what might we do about it? Journal of Economic Perspectives 10, 25–42.

Gould, E.D., Weinberg, B.A., Mustard, D.B., 2002. Crime rates and labor market opportunities in the United States: 1979–1997. The Review of Economics and Statistics 84, 45–61.

Gurr, T.R., 1989. The history of violent crime in America. In: Gurr, T.R. (Ed.), Violence in America. Sage, Newbury Park, CA.

Hagan, J., Levi, R., Dinovitzer, R., 2008. The symbolic violence of the crime-immigration nexus: migrant mythologies in the Americas. Criminology & Public Policy 7, 95–112.

Hagan, J., Palloni, A., 1998. Immigration and crime in the United States. In: Smith, J.P., Edmonston, B. (Eds.), The Immigration Debate. National Academy Press, Washington, DC.

Hagan, J., Palloni, A., 1999. Sociological criminology and the myth of Hispanic immigration and crime. Social Problems 46, 617–632.

Hirschi, T., Gottfredson, M., 1983. Age and the explanation of crime. American Journal of Sociology 89, 552–584.

Johnson, B., Golub, A., Dunlap, E., 2000. The rise and decline of hard drugs, drug markets, and violence in inner-city New York. In: Blumstein, A., Wallman, J. (Eds.), The Crime Drop in America. Cambridge University Press, Cambridge, UK.

Kelling, G., Sousa, W.H., 2001. Do police matter? An analysis of the impact of New York City's police reforms. Manhattan Institute Civic Report, 22 (December). Available: http://www.manhattan-institute.org/pdf/cr_22.pdf (accessed August 2008).

Levitt, S.D., 2004. Understanding why crime fell in the 1990s: four factors that explain the decline and six that do not. Journal of Economic Perspectives 18, 163–190.

Liedka, R.V., Piehl, A.M., Useem, B., 2006. The crime control effect on incarceration: does scale matter? Criminology and Public Policy 5, 245–276.

Logan, J.R., Alba, R.D., McNulty, T.L., 1994. Ethnic economies in metropolitan regions: Miami and beyond. Social Forces 72, 691–724.

Martinez, R., 2002. Latino Homicide: Immigration, Violence, and Community. Routledge, New York, NY.

Martinez, R., Lee, M.T., 2000. On immigration and crime. In: Lafree, G., Bursik Jr., R.J. (Eds.), Criminal Justice 2000: the Changing Nature of Crime, vol. I. National Institute of Justice, Washington, DC.

Martinez, R., Valenzuela, A., 2006. Immigration and Crime: Race, Ethnicity, and Violence. New York University Press, New York, NY.

Marvell, T.B., Moody, C.E., 1991. Age structure and crime rates: conflicting evidence. Journal of Quantitative Criminology 7, 237–273.

McCall, P.L., Land, K.C., Parker, K.F., 2011. Heterogeneity in the rise and decline of city-level homicide rates, 1976–2005: a latent trajectory analysis. Social Science Research 40, 363–378.

Merton, R.K., 1938. Social structure and anomie. American Sociological Review 3, 672–682.

Messner, S.F., Deane, G.D., Anselin, L., Pearson-Nelson, B., 2005. Locating the vanguard in rising and falling homicide rates across U.S. Cities. Criminology 43, 661–696.

Messner, S.F., Galea, S., Tardiff, K.J., Tracy, M., Bucciarelli, A., Piper, T.M., Frye, V., Vlahov, D., 2007. Policing, drugs, and the homicide decline in the 1990s. Criminology 45, 385–414.

National Research Council, 2004. Fairness and effectiveness in policing: the evidence. In: Skogan, W., Frydl, K. (Eds.), Committee on Law and Justice, Division of Behavioral and Social Sciences and Education. National Academies Press, Washington, DC.

National Research Council, 2005. Firearms and violence: a critical review. Committee to Improve Research Information and Data on Firearms. In: Wellford, C.F., Pepper, J.V., Petrie, C.V. (Eds.), Committee on Law and Justice, Division of Behavioral and Social Sciences and Education. The National Academies Press, Washington, DC.

National Research Council, 2008. Understanding crime trends: workshop report. Committee on Understanding Crime Trends. In: Goldberger, A.S., Rosenfeld, R. (Eds.), Committee on Law and Justice, Division of Behavioral and Social Sciences and Education. The National Academies Press, Washington, DC.

Ousey, G.C., Kubrin, C.E., 2009. Exploring the connection between immigration and violent crime rates in U.S. cities, 1980–2000. Social Problems 56, 447–473.

Ousey, G.C., Lee, M.R., 2007. Homicide trends and illicit drug markets: exploring differences across time. Justice Quarterly 24, 48–79.

Raphael, S., Ludwig, J., 2003. Prison sentence enhancements: the case of project exile. In: Ludwig, J., Cook, P.J. (Eds.), Evaluating Gun Policy: Effects on Crime and Violence. Brookings Institution Press, Washington, DC.

Reid, L.W., Weiss, H.E., Adelman, R.M., Jaret, C., 2005. The immigration-crime relationship: evidence across U.S. metropolitan areas. Social Science Research 34, 757–780.

Rose, D.R., Clear, T.R., 1998. Incarceration, social capital and crime: examining the unintended consequences of incarceration. Criminology 36, 441–479.

Rosenfeld, R., 2004. The case of the unsolved crime decline. Scientific American February, 82–89.

Rosenfeld, R., Decker, S.H., 1996. Consent to search and seize: evaluating an innovative youth firearms suppression program. Law and Contemporary Problems 59, 197–220.

Rosenfeld, R., Fornango, R., 2007. The impact of economic conditions on robbery and property crime: the role of consumer sentiment. Criminology 45, 735–769.

Rosenfeld, R., Fornango, R., Baumer, E., 2005. Did Ceasefire, Compstat and exile reduce homicide? Criminology and Public Policy 4, 419–450.

Sagi, P.C., Wellford, C.F., 1968. Age composition and patterns of change in criminal Statistics. Journal of Criminal Law, Criminology and Police Science 59, 29–36.

Sampson, R.J., Bean, L., 2006. Cultural mechanisms and killing fields: a revised theory of community-level racial inequality. In: Peterson, R., Krivo, L., Hagan, J. (Eds.), The Many Colors of Crime: Inequalities of Race, Ethnicity and Crime in America. New York University Press, New York, NY.

Sampson, R.J., Morenoff, J.D., Raudenbush, S.W., 2005. Social anatomy of racial and ethnic disparities in violence. American Journal of Public Health 95, 224–232.

Shaw, C.R., McKay, H.D., 1969 [1942]. Juvenile Delinquency and Urban Areas. University of Chicago Press, Chicago, IL.

Sherman, L.W., Rogan, D.P., 1995. Effects of gun seizure on gun violence: 'hot spots' control in Kansas City. Justice Quarterly 12, 673–693.

Sherman, L.W., Shaw, J.W., Rogan, D.P., 1995. The Kansas City Gun Experiment. NCJ-150855. National Institute of Justice, Washington, DC.

Simanski, J., September 2005. Mapping Trends in U.S. Legal Immigration: 1980 to 2003. Working Paper, Department of Homeland Security, Office of Immigration Statistics, Management Directorate.

Skogan, W.G., 2009. Police and Community in Chicago a Tale of Three Cities. Oxford University Press, New York, NY.

Spelman, W., 2000. The limited importance of prison expansion. In: Blumstein, A., Wallman, J. (Eds.), The Crime Drop in America. Cambridge University Press, Cambridge, UK, pp. 97–129.

Spelman, W., 2008. Specifying the relationship between crime and prisons. Journal of Quantitative Criminology 24, 149–178.

Steffensmeier, D., Harer, M.D., 1987. Is the crime rate really falling? Journal of Research in Crime and Delinquency 24, 23–48.

Steffensmeier, D., Harer, M.D., 1999. Making sense of recent U.S. crime trends, 1980 to 1996/98: age-composition effects and other explanations. Journal of Research in Crime and Delinquency 36, 235–274.

Stemen, D., 2007. Reconsidering Incarceration: New Directions for Reducing Crime. Vera Institute, New York, NY.

Stowell, J.I., Messner, S., McGeever, K., Raffalovich, L., 2009. Immigration and the recent violent crime drop in the United States: a pooled, cross-sectional time-series analysis of metropolitan areas. Criminology 47, 889–928.

U.S. Department of Justice, various years. Federal Bureau of Investigation. Crime in the United States. Retrieved from FBI website: http://www.fbi.gov/stats-services/crimestats.

U.S. Department of Justice, Bureau of Justice Statistics, 2011. Homicide Trends in the United States, 1980–2008. NCJ 236018. Retrieved from BJS website: http://bjs.gov/content/pub/pdf/htus8008.pdf.

Vieraitis, L.M., Kovandzic, T.V., Marvell, T.B., 2007. The criminogenic effects of imprisonment: evidence from state panel data, 1974–2002. Journal of Quantitative Criminology 6, 589–622.

Wadsworth, T., 2010. Is immigration responsible for the crime drop? an assessment of the influence of immigration on changes in violent crime between 1990 and 2000. Social Science Quarterly 91, 532–553.

Wadsworth, T., Kubrin, C.E., 2007. Hispanic suicide in U.S. metropolitan areas: examining the effects of economic disadvantage, immigration, and cultural assimilation. American Journal of Sociology 112, 1848–1885.

Wellford, C., 2005. Editorial introduction: special issue on gun policy. Criminology and Public Policy 4, 673–676.

Zimring, R.E., Hawkins, G., 1973. Deterrence: The Legal Threat in Crime Control. University of Chicago Press, Chicago, IL.

Zimring, F.E., 2007. The Great American Crime Decline. Oxford University Press, New York, NY.

Crime: Whitecollar

Susan P Shapiro, American Bar Foundation, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term 'white-collar crime' was coined by criminologist Edwin H. Sutherland, who argued that, because the offenses of corporations and white-collar workers are not represented in official crime statistics, scholars misunderstand the nature, causes, and consequences of crime and its control. This article describes how some of the standard topics in criminology – the social organization of offending, victimization, enforcement, prosecution, punishment, and criminological theory – are transformed when one looks in the unconventional places captured by the varied conceptions of white-collar crime. The article also considers the methodological and normative impediments to the study of crime and justice in the suites.

The term 'white-collar crime' was coined by Edwin H. Sutherland, in a 1939 presidential address to the American Sociological Society, as "crime committed by a person of respectability and high social status in the course of his occupation" (Sutherland, 1940, 1949, p. 9). In the better part of the century that followed, scholars continued to disagree about this definition of the concept – from its focus, not on illicit acts, but on specific characteristics of individual lawbreakers, to its characterization of their misdeeds as 'crimes.' Although more eschewed than embraced his early definition, Sutherland's compelling label persists. But a vast collection of synonyms giving linguistic expression to conceptual, theoretical, and ideological differences proliferated – upperworld, elite, suite, of-the-powerful, corporate, organizational, occupational, employee, avocational, trust, business, economic, official, governmental, or political crimes, illegalities, offenses, violations, deviance, misconduct, harms, or violence. After decades of fruitless conceptual debate, most scholars have agreed to disagree.

This babel of competing verbiage and the disparate phenomena it embraces would be a mere annoyance were it not for the fact that the term was born, as Robert Merton once observed (Merton, 1957, pp. 90–92), of an effort to 'liberate' scholarship from 'cognitive misbehavior' reflected in the well-accepted but spurious correlation between poverty and crime. Concepts tell us where to look, and where we look, affects what we find. Sutherland asserted that criminologists, obstructed by conceptual blinders, had been looking in the wrong places:

> …[white-collar crimes] are not included in the criminal statistics nor are individual cases brought to the attention of the scholars who write the theories of criminal behavior. The sample of criminal behavior on which the theories are founded is biased as to socio-economic status, since it excludes these business and professional men. This bias is quite as certain as it would be if the scholars selected only red-haired criminals for study and reached the conclusion that redness of hair was the cause of crime. (Sutherland, 1949, pp. 8–9)

In other words, if scholars look only at crimes committed by the poor, it is not surprising that they conclude that poverty and its associated pathologies cause crime. For Sutherland, this selection bias is the *raison d'être* of inquiry into white-collar crime. These offenses evoke that nagging sense of all that is ignored by traditional criminology and that cause us to misunderstand the nature, causes, and consequences of crime and its control.

In the ensuing years, white-collar crime has become a subject of inquiry in its own right, of course. But selection bias continues to symbolize the essence of white-collar crime scholarship, simultaneously signaling its promise and foreshadowing its vulnerability, accounting for why the work is provocative and groundbreaking, and yet so often misguided. This article explores this dual tension, thereby featuring a handful of the major contributions of this scholarship and the challenges that social scientists face in doing it well.

Removing the Blinders

We begin by considering how some of the standard topics in criminology are transformed when one looks in the unconventional places captured by the varied conceptions of white-collar crime.

The Social Organization of Offending

Although Sutherland's early definition explicitly referred to individuals, his empirical work (Sutherland, 1949) on the offenses of the 70 largest manufacturing, mining, and mercantile corporations in the United States took organizations as the unit of analysis, a legacy reflected in the generations of scholars he spawned. Although organizations surely commit misdeeds, they cannot do so without the complicity of individuals, often countless individuals plucked from disparate branches of the corporate tree and from tiers spanning the organizational hierarchy, indeed individuals often situated outside of the organization altogether – lawyers, accountants, investment bankers, brokers, agents, consultants, advertisers, competitors, even regulators. But coordinating the activities of such a diverse collection of collaborators situated across the globe, many of them ignorant of the illicit activities to which they contribute and often engaged in concerted action spanning years or decades, is not easy. Organizations need to maintain secrecy, lest the activities be exposed. They must ensure that some conspirators remain unwitting of the consequences of their

actions, lest they refuse to participate or blow the whistle to regulators inside or outside the organization. They must maintain loyalty and commitment and cool out those with little taste for offending, lest disgruntled or disaffected conspirators sell out their colleagues. They must buffer those in positions of power from those carrying out their directives to shield the former from accusations of culpability or to build a paper trail that deflects responsibility away from those most complicit. They must deceive and cover up their activities from victims to ensure their continued victimization and to postpone discovery. They must destroy or doctor evidence to subvert regulatory discovery and to undermine the prosecution of their misdeeds, if discovered. These are massive organizational tasks perhaps better suited to the theoretical paradigms emanating from organizational sociology than from traditional criminology and that provoke unique challenges for systems of social control (Stone, 1975).

Victimization

White-collar crimes exploit impersonal trust, asymmetric relationships in which principals entrust strangers to provide specialization, intermediation, and collectivization (Shapiro, 1990). Victims of white-collar crime, therefore, typically lack expertise and access to the information necessary to evaluate the conduct of trustees. Because many participate in 'futures transactions' such as investment and insurance where payoff is delayed or protracted, victims have no tangible data in the interim to assess the likelihood that commitments will eventually be honored. Victims of white-collar crimes, therefore, tend to be unwitting. Because victims of common crimes are readily aware that they have been robbed, burglarized, assaulted, or victims of larceny, and members of their social networks know when associates have been killed, systems of social control (and researchers as well) tend to be reactive. They wait for victims to tell them that a crime has occurred. Reactive systems of social control yield little, stale, and biased data (about more blatant misdeeds) when most victims are unwitting. The policing (and study) of white-collar crime therefore requires different methods than those characteristic of common crimes.

White-collar offenses also differ from common crimes in the recruitment, pattern, scale, and impact of victimization, and in the relationships between victims and offenders. Because some relationships of trust collectivize the interests of multiple principals – in financial institutions, mutual and pension funds, insurance pools, and the like – victimization from the abuse of trust is often massive. Moreover, because white-collar crimes tend to be delivered by organizations whose activities touch countless consumers, clients, employees, investors, and neighbors (who drink the water and breathe the air poisoned by the corporation), victimization is often extensive and its cost enormous. Research finds that the organizational complexity of an offense is "the single most powerful predictor of victimization" (Weisburd et al., 1991, p. 93). Hence, the organization itself becomes the 'weapon' in white-collar crime. It is "for white-collar criminals what the gun or knife is for the common criminal – a tool to obtain money from victims" (Wheeler and Rothman, 1982, p. 1406).

Enforcement

Systems of social control can be differentiated into those which seek 'compliance' – preventing the occurrence of law violations – or 'deterrence' – reacting to violations that have already occurred (Reiss, 1984). Because white-collar offenses portend significant victimization with respect to cost, harm, consequences, irreversibility, and numbers of victims, enforcement systems frequently opt for compliance over deterrence. Regulators would rather prevent toxic waste or abort a securities swindle than punish those responsible, especially when the unwittingness of victims often means that significant harm has been done by the time reactive deterrence systems learn of the illegality. And many find that cooperative enforcement techniques are better able to secure compliance than those structured to deliver punishments (e.g., Grabosky and Braithwaite, 1986).

Compliance systems are more common in the enforcement of white-collar than common crimes as well because the former are more predictable. Because would-be organizational offenders are relatively stable and visible, as they seek to market their wares, they can be licensed, given preclearance, inspected, monitored, and subject to regulations and mandated procedures.

The policing of white-collar crime is also more diversified than that of common crime, which is undertaken largely by the state and private police (e.g., security guards). The costs and consequences of white-collar crimes, the penalties imposed, and the loss of reputation and the diminution of institutional trust (which are the lubricants of continued business) are so great that those likely to bear these costs often undertake their own social control. Insurers, stock and commodity exchanges, professional or trade associations, and organizations themselves undertake or mandate self-regulation or buy the regulatory services of outside auditors, inspectors, rating firms, and the like.

Prosecution and Punishment

Scholars object, with good reason, to Sutherland's classification of white-collar offenses as 'crimes.' Many – perhaps most – so-called white-collar crimes are, in fact, not prosecuted as crimes at all or subject to criminal penalties. Compliance systems of social control have little use for the criminal justice system (aside from the incentive that potential criminal penalties give for undertaking self-regulation or abstinence). Criminal sanctions come too late, after the damage is done; they require standards of proof that are difficult to meet, especially with extensive obfuscation and cover-up that shield those most complicit; they do little to remedy the harm, lessen its impact, or impede future wrongdoing. Moreover, the choice between imprisonment and relatively small criminal fines offers limited options when the criminals are wealthy organizations. White-collar offenses are often dealt with by civil injunctions or administrative remedies that are better able to abort incipient or ongoing violations, contain the damage quickly, remedy the harm, provide restitution to victims, and fashion remedial solutions that will reduce opportunities for future wrongdoing without exacerbating the harm to innocent parties. Moreover, because wayward organizations or their insurers often have

deep pockets, victims and the private plaintiff's bar have strong incentives to undertake civil litigation against white-collar malefactors. Private litigation often plays a significant role in rectifying the misdeeds, securing compensation, meting out punishment, and deterring future misconduct.

Criminological Theory

As even these few selected features of white-collar crime scholarship make clear, Sutherland's *raison d'être* for drawing attention to these offenses was prescient. By broadening the expanse of crime and its control from those patterns and institutions evoked by the stuff of traditional crime statistics to those insinuated in organizational offending and abuse of trust, the limitations of criminological theory are readily exposed. The challenge of incorporating into criminological theory the phenomena observed once the conceptual blinders were removed has been less successful.

Over the years, there have been sporadic attempts to develop a general theory of crime that accounts for both white-collar and common crimes. Sutherland's theory of differential association – that the techniques and motivations for criminal behavior are learned through contact with others – represents the earliest attempt (Sutherland, 1949, p. 234). Some scholars have extended Merton's anomie theory of deviance or variants of strain theory, proposing that, like ordinary deviance, white-collar crimes arise when legitimate means to achieve material success or to attain organizational goals are blocked or unavailable. And other theories of white-collar lawbreaking have drawn on and refined rational choice, self-control, neutralization, conflict, and especially variations on opportunity theories (Benson and Simpson, 2009; Shover and Hochstetler, 2006), all originally used to account for traditional crimes.

These varied theoretical paradigms have been subject to extensive and justifiable criticism for imprecision, tautological reasoning, inability to account for compliance and law-abiding behavior, confusing correlation and cause, or lack of empirical support. While some scholars continue to refine grand theories of white-collar and common crimes, others eschew such theories, developing explanations for specific misdeeds or those that account for variation within particular datasets.

Methodological Pitfalls

The empirical challenges to the study of white-collar crime far outweigh the theoretical ones. First, because of victim unwittingness, relationships of impersonal trust, the protection of crime in the suites by institutions of privacy, complex cover-up schemes, the commission of crime with organizational weapons rather than physical ones, and the ability to buffer misdeeds in long-latent futures transactions, the dark figure of undetected white-collar crime probably exceeds that of all other crimes combined. And even where law enforcement officials have discovered white-collar offenses, they are often unable to ascertain or pin the blame on those most responsible. The most common sources of data on traditional crimes – arrest records and crime statistics, victimization surveys, and observation studies – therefore yield little and very biased data on white-collar crime.

With limited access to data, researchers often rely on materials in the public domain – typically, sentencing data, reports of administrative agencies, journalistic accounts, hearings, and official inquiries. But selection bias is the price of easy access. Journalists produce what sells – the egregious, scandalous, shocking, clever, costly, and devastating offenses committed by notorious or colorful characters, where those in high stations fell from grace, where safety nets failed, where victims should have known better, where offenders received excessively lenient treatment – many of the same features of offenses that incite public hearings and legislative inquiry. Hardly the mundane stuff of white-collar crime day in and day out. In selecting research subjects from the news media and political tribunals, scholars replicate Sutherland's red-haired crime fallacy. We come to believe that these crimes are committed (with impunity) by the rich and powerful, when a less biased data source would reveal that they are typically carried out by members of the middle class, and that, contrary to expectation, the severity of punishment of white-collar criminals appears to increase with social standing (Weisburd et al., 1991).

Turning to public records of convicted criminals incurs other selection biases. Convictions represent a very small and unrepresentative tip of the iceberg. Research on the US Securities and Exchange Commission, for example, indicates that not even 5 of every 100 parties found after investigation to have violated the securities laws are convicted of a crime (Shapiro, 1984; pp. 155–161). Criminal courts are less likely to process the misdeeds of organizations (than of individuals), that were uncovered by compliance (rather than deterrence) systems of social control, and that were committed by those atop corporate hierarchies who were able to conceal their responsibility or inculpate others in exchange for immunity from criminal prosecution. When offenses are ongoing, victimization is extensive, and the impact is significant, law enforcers opt for compliance – halting the misdeeds, forestalling, and remedying the harm, minimizing the likelihood of recurrence, and doing so with more immediacy, less litigation, and lower standards of proof – over criminal proceedings. Therefore, although serious significant offenses can certainly be found at the tip of the iceberg, so can some of the most trivial for which no remedy other than criminal prosecution is available. Researchers concentrate on the tip of the iceberg at their peril.

Because of these biases embedded in the social construction of white-collar offenses as 'crimes,' many scholars choose not to concentrate solely on criminal cases. Some study offenses that carry criminal penalties, even if they have been resolved instead by private settlements, warnings, injunctions, civil or administrative proceedings, lawsuits, or other remedies. But a slippery slope looms below the tip of the iceberg. As the labels shift from 'crimes' to 'violations,' 'offenses,' 'misconduct,' 'harms,' 'wrongs,' 'deviance,' 'violence,' and so on, the literature becomes cluttered with studies of torts, liability, malpractice, accidents, mistakes, or inadvertence committed by corporations or those in high-status occupations. We dilute white-collar crimes with all manner of misfortune. To include these harms among crimes in the suites is like including fender benders in an ice storm among crimes in the streets.

Sadly, the ideological passions that animated Sutherland's work and that of many of his intellectual progeny, which

cause us to be suspicious of corporations and individuals of wealth, status, and power, also cause us to get sloppy. They permit us to confuse mistake or accident with intentional noncompliance. They lead us to believe what comports with our assumptions, relying on the thin anecdotal data that are readily available rather than probing deeper, using published reports and secondary data rather than collecting primary materials and triangulating data sources. They blind us from questioning and making problematic what seems intuitively true. As a result, we are destined to learn that some of the classic and repeatedly cited 'landmark narratives' of corporate crime – for example, the Ford Pinto affair and the Challenger disaster – actually got the story wrong, confusing misconduct with mistake, and venality and amoral calculation with banal normalized organizational routines (Lee and Ermann, 1999; Vaughan, 1996).

We are also learning what it takes to get the story of long-running technologically, financially, and organizationally complex potential white-collar crimes right. In the case of the Challenger disaster, it took 10 years of persistent, painstaking, methodical research that probed the entire decade that preceded the shuttle explosion that cold January morning; systematic analysis of more than 122 000 pages of documents, 2500 pages of hearing transcripts, and 9000 pages of depositions; dozens of new interviews and site visits; the commitment to learn technical science and engineering; and the willingness of a publisher to print an affordable 600-page scholarly book with almost 100 pages of technical charts, diagrams, memos, and thousands of footnotes to transform a seductive but erroneous account of organizational misconduct into a creditable one of the normalization of deviance and the sociology of mistake (Vaughan, 1996). Vaughan's chronicle, with the conventional (erroneous) account in boldface type juxtaposed against the fuller complex documentary and ethnographic narrative, starkly demonstrates what is at stake in the social construction of white-collar crime scholarship and what is required of future scholars.

The Future of White-Collar Crime

White-collar crimes are dynamic, continually responding to changing features of trust relationships, structures of opportunity for abuse, and strategies of social control. But unprecedented transformations in each of these factors portend major changes in the future course of white-collar crime and its control.

First, relationships of trust are increasingly asymmetric, enlarging opportunities for abuse at the same time that they better conceal misdeeds from victims and social control organizations. On the one hand, social and especially economic interactions are increasingly enacted in virtual space among anonymous parties in often one-shot exchanges. On the other, the objects of exchange have become more complex, abstract, and opaque. For example, mortgages backed by real property are securitized and combined with other pools of securitized assets and then offloaded to countless investors. These securities are then repackaged, sliced, and diced into still other instruments. Others are transformed into derivatives, bets, or other synthetic securities. The most risky slices (or 'tranches') of these instruments are diversified and combined into yet other instruments, now magically stripped of risk. The owners of shares of these varied pools of abstract wealth do not know what they own, how risky it is, or even that they may own the same thing several times over in different incarnations or are even betting against themselves.

Second, the economic landscape is increasingly globalized. Complex, abstract financial instruments are marketed, bought, and sold throughout the world, often over virtual marketplaces. Capital is highly mobile; individual and organizational victims worldwide gravitate to more secure currencies and higher rates of return in other jurisdictions. Corporations and the lawyers, accountants, and other professionals who facilitate their misdeeds are increasingly multinational. Global financial institutions are interdependent. Globalization increases the opportunities to conceal or cover up misdeeds from unwitting facilitators of white-collar crime, victims, and oversight organizations alike. It also creates opportunities for wrongdoers to situate their activities in jurisdictions with more favorable laws or less regulation. Because multinational criminal justice or regulatory organizations (both public and private) are less developed, resourced, and coordinated than their domestic counterparts, globalized misdeeds more easily escape detection or meaningful sanctions. And, of course, white-collar crimes that cross the globe often leave even more victimization and loss in their wake. Great challenges in the social control of white-collar crime therefore lie ahead for scholars to investigate.

White-Collar Crime and the New Social Science

> … societies have been undergoing an organizational revolution. Just as the forests and fields of the physical environment are being replaced by streets and skyscrapers, the primordial institutions around which societies have developed are being replaced by purposively constructed social organization. (Coleman, 1990, p. xv)

Perhaps nowhere has James S. Coleman's influential call for a new social science attuned to the transformation of social life occasioned by the displacement of natural persons by 'purposive corporate actors' been better realized than in the study of white-collar crime. White-collar crime scholarship does not hold a monopoly on the study of organizational offending, of course. Even traditional criminology (which concentrates on the natural persons who populate the primordial social organization which Coleman argues is 'withering' away) has produced the occasional study of gangs, organized crime, terrorism, the distribution of illicit goods (e.g., drugs, gambling) and services (e.g., prostitution, fencing), and the like. But white-collar crime scholarship more than any other area of criminological inquiry has made problematic the difficulties of regulating organizational life and insuring norm compliance within organizational contexts. White-collar crime offers a bridge to the new social science and a compelling window onto the changing structure of social life. As we gaze through the window, we realize the myopia our blinders have imposed.

See also: Anomie; Coleman, James Samuel (1926–95); Control, Social; Corporate Governance; Crime and Class; Deterrence Theory: Crime; Economics and Crime; International and Comparative Criminology; Organizations, Sociology of; Property Crime: Victimization and Trends; Punishment: Social and Legal Aspects; Regulation, Empirical Analysis of; Self-Control Theory and Crime; Strain Theories and Crime; Trust, Sociology of; White Collar Crime: Recent Trends and Debates.

Bibliography

Benson, Michael L., Simpson, Sally S., 2009. White-Collar Crime: An Opportunity Perspective. Routledge, New York, NY.

Coleman, James S., 1990. Foundations of Social Theory. Belknap Press of Harvard University Press, Cambridge, MA.

Grabosky, Peter, Braithwaite, John, 1986. Of Manners Gentle: Enforcement Strategies of Australian Business Regulatory Agencies. Oxford University Press, Melbourne, Australia.

Lee, Matthew T., Ermann, M., David, 1999. Pinto 'madness' as a flawed landmark narrative: an organizational and network analysis. Social Problems 46, 30–47.

Merton, Robert K., 1957. Social Theory and Social Structure. Free Press, Glencoe, IL.

Reiss Jr., Albert. J., 1984. Selecting strategies of social control over organizational life. In: Hawkins, Keith, Thomas, John M. (Eds.), Enforcing Regulation. Kluwer-Nijhoff, Boston, pp. 23–35.

Shapiro, Susan P., 1984. Wayward Capitalists: Target of the Securities and Exchange Commission. Yale University Press, New Haven, CT.

Shapiro, Susan P., 1990. Collaring the crime, not the criminal: reconsidering the concept of white-collar crime. American Sociological Review 55, 346–365.

Shover, Neal, Hochstetler, Andy, 2006. Choosing White-Collar Crime. Cambridge University Press, New York, NY.

Stone, Christopher D., 1975. Where the Law Ends: The Social Control of Corporate Behavior, first ed. Harper and Row, New York.

Sutherland, Edwin H., 1940. White-collar criminality. American Sociological Review 5, 1–12.

Sutherland, Edwin H., 1949. White Collar Crime. Holt, Rinehart and Winston, New York.

Vaughan, Diane, 1996. The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA. University of Chicago Press, Chicago.

Weisburd, David, Wheeler, Stanton, Waring, Elin, Bode, Nancy, 1991. Crimes of the Middle Class: White-Collar Offenders in the Federal Courts. Yale University Press, New Haven, CT.

Wheeler, Stanton, Rothman, Mitchell, L., 1982. The organization as weapon in white-collar crime. Michigan Law Review 80, 1403–1426.

Criminal Justice and Social Work Practice

Roger Smith, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article offers an overview of the complex and conflictual relationship between the ideas and values underpinning social work and those informing the operation of the criminal justice system. It describes the ways in which these tensions have been played out historically, before offering a typology, which seeks to account for the different approaches taken to social work practice with people in conflict with the law. The article considers the role of social work in challenging discrimination endemic to the justice system before briefly considering how to understand 'what works.' In concluding, the article asserts the legitimacy of the social work project in the context of the criminal justice system and restates the importance of retaining a commitment to core professional values, rather than becoming subsumed by the alternative logics of control and punishment. The material here is largely based on experience in the United Kingdom, supplemented with international examples where possible.

Conceptual Challenges and Practice Dilemmas

The juxtaposition of social work and the criminal justice system undoubtedly poses a range of fundamental conceptual and practical challenges. In essence, these relate to the tensions arising from different ways of thinking about and classifying human characteristics and human behavior. It is important, therefore, to start with a brief outline of these inherent tensions, before going on to develop a more detailed understanding of the challenges and possibilities manifesting themselves in the context of social work practice in this field.

At the risk of oversimplification, the criminal process is fundamentally concerned with problematic behavior, holding its perpetrators to account, and imposing measures of control and sanctions proportionate to the identified transgression. Social work, on the other hand, is interested in addressing the problems experienced by those affected by inequality and oppression, and resolving these based on their needs and aspirations. It is likely, therefore, to have a central role in enabling, empowering, and advocating for those who are 'victimized,' particularly those who experience abuse or exploitation, as much as it also has a role in addressing the welfare needs and rights of those who are identified and treated as offenders by the justice system. The place of social work within criminal justice is therefore, for this and other reasons, fundamentally ambiguous.

In addition, social work is not solely concerned with the attributes and actions of individuals, but sees the person in his or her context, and equally acknowledges that the focal point for change is as much social as it is personal. The individual is not the sole focus for intervention and change, and at this basic level there appears to be a significant mismatch between the ambit and activities of the justice system and the objectives and practices of social work.

In the following account, the ways in which these tensions have played themselves out will be explored. This will lead to a consideration of differing modes of accommodation between the two, which enable social work practice to gain some purchase in the criminal justice arena, while remaining true to its underlying professional principles and core values.

The Origins and Development of Social Work with Offenders

It was during the nineteenth century that professional social work (as it is now known) began to gain a recognized place in the criminal justice system. This coincided with the wider emergence of movements directed at meeting the welfare needs of people in difficult and oppressive settings, particularly experiencing the forms of poverty associated with rapid economic change and the formation of modern capitalist societies. In the sphere of criminal justice, the issue of welfare need was closely bound up with the wider emerging distinction between the 'deserving' and 'undeserving' poor. The category to which those in need were assigned had a significant impact on the kind of services to be provided for them, and irrespective of whether or not they had committed a criminal offense, whether they would be punished or helped. Criminalization, however, tipped the balance sharply toward a punitive form of intervention. Gradually, however, as in other areas of welfare practice, such as the treatment of mental ill health, distinctions began to emerge. Certain groups came to be viewed effectively as less than fully responsible for their behavior, simply by reason of their status, and would thus merit by way of intervention something other than simple punishment. Their delinquent behavior might be attributable to their circumstances, or certain characteristics that resulted in diminished responsibility, and so other forms of intervention could legitimately be brought into play to take these factors into account. If they were not, the effects of punishment alone might be to confirm, rather than prevent, a life of crime.

> The want of employment – the prevalence of improvident marriages – the degrading tendencies of the poor laws – and the increased facilities for the consumption of spirituous liquors have doubtless contributed much to weaken the natural affections of the lower classes of society… Dreadful is the situation of the young offender: he becomes the victim of circumstances over which he has no control. The laws of the country operate not to constrain but to punish him… If when apprehended, he has not attained the full measure of guilt, the nature of his confinement is almost sure to complete it; and discharged, as he frequently is, penniless, without friends, character or employment, he is driven for a subsistence to the renewal of depredations. (Source: Report of the Committee into Juvenile Delinquency, London, 1816 quoted in Muncie, 2009: p. 53.)

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28020-3

For children, the mitigating factors identified might be associated with the circumstances of their upbringing, that is, the effects of parental irresponsibility and wider influences such as mass unemployment, none of which were under their control. Increasingly, too, age alone was seen as a factor that lessened children's level of independent responsibility. For other groups to be seen similarly as acting under diminished responsibility, specific conditions would need to apply, such as mental ill health, or what would now be known as learning disabilities. In some cases, too, those with alcohol problems, or extreme personal circumstances, might also be held as less than fully responsible for any criminal actions they might commit. It was partly because these circumstances or characteristics might not always be clear cut that some of the subsequent machinery for assessing and classifying offenders (welfare reports) began to be developed. In the United Kingdom, for example, the emergence of principles of 'rehabilitation' and 'treatment' of offenders affected by neglect and destitution (especially young offenders) is associated with the introduction of the Probation Act 1907 and the Children Act 1908, which "was notable for its attempt to reconcile welfare and justice imperatives through the establishment of the 'juvenile court' with both civil jurisdiction over the 'needy' child and criminal jurisdiction over the offending child" (Hendrick, 2006: p. 8).

It was in this way that space began to open up for welfare as both a legitimate form of discourse and as a vehicle for the provision of help and support to identified offenders. While from a social work perspective welfare interventions to resolve the problems of those in need are inherently justifiable, from a criminal justice perspective their justification is that they will contribute to reducing or eliminating the prospect of further offending behavior by the individual concerned. It is in this way that a kind of accommodation between the competing discourses of welfare and justice began to emerge. Nonetheless, this rapprochement of ideas was neither clear cut nor stable, and has been the subject of continual renegotiation and contestation subsequently.

Early forms of social work with offenders emerged in the form of targeted services, typically offered by philanthropic organizations, which in turn had strong roots in religious movements. Police court missionaries, for example, took the language of salvation into their practice, as they began to offer advice and support to those identified as offenders in and around the magistrates' courts, seeking to identify those with particular needs and then seeking authority to address these needs in place of punitive sanctions.

1876: Hertfordshire printer Frederic Rainer, a volunteer with the Church of England Temperance Society (CETS), writes to the society of his concern about the lack of help for those who come before the courts. He sends a donation of five shillings (25p) towards a fund for practical rescue work in the police courts. The CETS responds by appointing two "missionaries" to Southwark court with the initial aim of "reclaiming drunkards". This forms the basis of the London Police Courts Mission (LPCM), whose missionaries worked with magistrates to develop a system of releasing offenders on the condition that they kept in touch with the missionary and accepted guidance. (*The Guardian*, 2 May 2007)

Institutionalizing Social Work in Criminal Justice

From the early philanthropic initiatives, there subsequently developed a much more formalized approach to social work provision in the justice system, with the establishment and subsequent expansion of statutory agencies to take on this role. Probation services have progressively been established around the world, with the remit to 'advise, assist, and befriend' offenders where they may have welfare needs. A special issue of the *Probation Journal* in 2011, for example, highlighted developments in countries as diverse as South Korea, France, Russia, Australia, and Japan, indicating that the 'space' for a welfare approach to offending behavior can be achieved across a wide range of cultural variations. On the other hand, it is also acknowledged at the same time that more recent developments, oriented toward risk and behavior management, surveillance, and corrections, have influenced practice across national borders, and there appears to be a widespread trend toward the diminution of the social work role in probation, as in France, for example (Herzog-Evans, 2011).

Over time, too, there has emerged a clear distinction between social work with adult offenders and with children and young people in trouble. This mirrors the distinction between adult and children's jurisdictions and the effective establishment of children as a special interest group by way of international conventions, including the United Nations (UN) Convention on the Rights of the Child, on their rights and just treatment. In effect, with the institutionalization of such measures, children in the justice system do not, as adults do, have to justify their entitlement to special consideration and the prioritization of their welfare interests, since they are *de facto* assumed to be entitled to support and assistance, in principle, at least. In practical terms, this has led to the establishment of entirely separate systems and processes for dealing with young people who get into trouble with the law, including separate jurisdictional procedures and courts, and distinctive institutional arrangements for addressing their welfare needs, such as those afforded by child welfare services.

Consistent with the international principles and objectives articulated under the aegis of the UN, that organization has also taken an active role in promoting and sharing evidence of best practice in 'juvenile justice' with the aim of ensuring that there is a general commitment to the embedding of welfare interventions at the heart of the justice process.

In the area of **juvenile justice**, UNICEF aims to reduce incarceration while protecting children from violence, abuse and exploitation. It promotes rehabilitation that involves families and communities as a safer, more appropriate and effective approach than punitive measures. Justice systems designed for adults often lack the capacity to adequately address these issues and are more likely to harm than improve a child's chances for reintegration into society. For all these reasons, UNICEF strongly advocates **diversion** (directing children away from judicial proceedings and towards community solutions), **restorative justice** (promoting reconciliation, restitution and responsibility through the involvement of the child, family members, victims and communities), and **alternatives to custodial sentencing** (counselling, probation and community service). (UNICEF Child Protection Information Sheet, 2006)

A Typology of Social Work in Criminal Justice

Despite the aspirations of the UN and others to embed effective welfare practice in the justice system, social work continues to sit in an ambiguous position under the shadow of the logic and practices of criminal justice processes, as the precarious position of welfare in probation services appears to underline. There is no doubt that social work practice in criminal justice has been infused with and modified by the logic and rationale of criminalization and the principles of individual accountability. Social work, therefore, often appears to encompass a dual mandate, with welfare intervention predicated on its capacity not just to improve the lives of children and young people, but also to prevent them reoffending, in their own interests. Once this logical leap is made, it becomes practicable and acceptable to incorporate measures of control and correction into 'social work' practice. Despite this elision, and the occasional explicit argument that welfare and justice objectives are coterminous, it remains difficult to sustain the argument that they can be fully aligned in practice, or that this does not simply lead to the co-option of social work in the service of justice-based objectives, in terms of containment and control.

Instead, it is suggested, practitioners should focus first and fundamentally on the values and purposes of social work to develop a typology of appropriate and genuinely enabling forms of practice within the justice system. In certain circumstances, it is likely that these forms of intervention will also support wider aims, such as the prevention of reoffending, but in social work terms these are subsidiary and not principal objectives. We may, for instance, be able to articulate a typology of criminal justice-based social work as follows: resettlement, rehabilitation, socialization, and, emerging more recently, restoration and reconciliation. It has also suggested that social work could and should adopt a 'welfare rights' approach to working in the justice system, although this is more evident as a statement of principle than an established form of practice (see Scraton and Haydon, 2002).

Resettlement

Since the early days of its involvement in criminal justice, social work has been centrally involved in the process of supporting exoffenders to resettle, particularly on release from custodial or other secure institutions. Given that reoffending rates of those released from custody are often very high, and that is linked with a lack of security, it seems that the role of social work in this context is particularly important. Not only does it offer the possibility of supporting exoffenders through difficult transitions, but it may also contribute to reducing the likelihood of further involvement with the justice system.

Resettlement has been established as a central aim of social work intervention with exoffenders for a considerable length of time. As Maguire and Raynor state "the main focus of attention has understandably been on the pressing practical problems faced by many ex-prisoners, and interventions have been driven as much by the charitable aim of helping people in need, as by criminal justice goals" (2006: p. 20). Organizations such as Discharged Prisoners' Aid Societies, originating in the nineteenth and early twentieth centuries in Western societies provided an early stimulus for this kind of social work intervention, which was subsequently built upon and developed by both statutory probation services and organizations such as the UK-based National Association for the Care and Resettlement of Offenders.

This emphasis on aftercare and the provision of effective practical help for those passing through the justice system has continued to form a key element of service provision internationally. In Kenya, for example, the Probation and Aftercare Service retains responsibility for the resettlement of offenders, operating a range of services including probation hostels that provide therapeutic services and vocational training (Obondi, 2010).

Rehabilitation

Offering a broader range of interventions, rehabilitation practice encompasses the objective of enabling former offenders to reestablish themselves with their families and communities, and to become productive contributors to society. Interventions under this heading may incorporate assistance in seeking training or employment opportunities, support to address problems such as substance misuse, counseling and family support, resolving problems linked to offending, and personal advice and support.

The origins of the rehabilitative ideal can also be located in nineteenth-century industrial societies, with the establishment, for example, of the police court missionary service in England and Wales in 1876. Rehabilitation was not seen simply as a matter of solving the practical problems faced by exoffenders, but of redemption; it was 'for the good of their souls, as well as to reduce the harm they would otherwise continue to do themselves and others, such as their families' (Raynor and Robinson, 2009: p. 5). Importantly, the rehabilitative perspective focuses on providing 'treatment' to address the underlying causes of crime, presumed to be located squarely with offenders themselves, whether these be associated with educational deficits, mental health problems, or possibly substance or alcohol misuse (Patterson, 2012: p. 11). The concept of treatment sat well with the emerging emphasis on psychotherapeutic interventions as social work developed and became increasingly professionalized during the early to mid-twentieth century. This strand of thought remains influential in shaping models of intervention directed at changing the lives, capabilities, and circumstances of those whose criminal behavior gives cause for concern.

The 'Good Lives' model of offender rehabilitation, for example, is organized around developing the strengths and abilities of those undergoing the program:

> Intervention should be viewed as an activity that should add to an individual's repertoire of personal functioning, rather than an activity that simply removes a problem, or is devoted to managing problems, as if a lifetime of restricting one's activity is the only way to avoid offending. (www.goodlivesmodel.com, accessed 18.06.13.)

Socialization

A strand of practice has emerged that takes a more behavioral approach to resolving the difficulties faced by offenders, and relies on the assumption that there is a direct connection between behavior change and improving well-being. Under

this heading, for example, interventions may include social or life skills training, anger management programs, offending behavior programs, and other forms of behavioral intervention designed to reduce or eradicate problem behaviors, using either positive or negative reinforcement (Madoc-Jones, 2008: p. 132). These interventions are geared not primarily toward promoting the individual's welfare, but to ensuring that they are reintegrated into society as 'good citizens.' The assumptions behind this kind of approach incorporate the notion that behavior can be learned and reinforced either positively or negatively, and that 'problem behavior' is simultaneously socially undesirable and contrary to the best interests of its perpetrator. As behavioral interventions themselves have become increasingly central to the social work repertoire, so they have equally come to be seen as particularly suited to the challenge of seeking to inhibit antisocial tendencies. As early as 1964, studies were being undertaken into the effectiveness of positive reinforcement, and reporting beneficial outcomes (Trotter, 2009: p. 143).

Increasingly, this form of intervention has come to be associated with the notion of 'prosocial modeling', based on the assumption that positive reinforcement is known to be more effective than behavioral programs based on negative sanctions. This term is

> interpreted… broadly to include a group of skills which include supervisors modelling pro-social values, reinforcing clients' pro-social expressions and actions and negatively reinforcing or confronting pro-criminal actions and expressions of those clients'. p. 142)

Restoration

Another relatively recent development has been the emergence of restorative justice, which arguably represents a contemporary attempt to integrate social work methods with the expectations of the justice system. Unlike other approaches, which are concerned with changing aspects of offenders' behavior or circumstances, the focus here is on the offense and the 'harms' arising from it. Interventions are geared toward problem solving, and enabling all those affected by the offense to come together to achieve a negotiated settlement. The offender has an opportunity to 'put right' the damage done, and also, in doing so, to be 'reintegrated' and effectively forgiven by the community. Restorative justice derives substantially from the experiences of societies seeking to achieve reconciliation following periods of conflict, and it is informed by a commitment to utilizing the justice process to rebuild and support healthy communities.

The modern widespread adoption of restorative justice practices is commonly attributed to the (re)discovery of a historic Maori tradition, grounded in principles of community justice, consultation, and settlement of any damage done: "consultation between the parties would take place as to a suitable way of dealing with the matter, so as to heal any hurts and restore things to 'normal' again" (Considine, 2003: p. 153). The 'healing' process offers a means to secure reconciliation and reintegration for the offender, at the same time as the offense victim's needs are directly addressed. The interests and voices of the offender and victim are brought together and considered in tandem, so that both sides are enabled to gain from the resolution process, and the well-being of the community in general is enhanced.

> Through restorative justice, the dignity of both the offender and victim are maintained through a process that is the opposite of customary criminal justice proceedings… The focus of restorative justice is on the offender's whole personality, not only on the acts that have caused the harm. Ideally, this informal but emotionally intense process will have a humanizing effect on all participants. (van Wormer, 2002)

Welfare + Rights

An alternative approach to social work in criminal justice, albeit relatively underdeveloped, is based on principles of 'welfare + rights,' underpinned by the recognition that many offenders are significantly disadvantaged, and that it is only through a process of empowerment and capacity building that they will be able to overcome the inequalities that are associated with their criminalization. In other words, social work practice with people in the justice system should not collude with the labels associated with their status, but should work with them to ensure that they are able to challenge the discriminatory processes they experience.

Scraton and Haydon (2002), for example, have argued for a 'positive rights-based welfare approach' to intervention with children in the criminal justice system; and the Howard League has developed a form of intervention with young people, which is a form of enhanced legal service, combining 'welfare + rights' (Smith and Fleming, 2011). While the principles of a rights-based approach are quite widely accepted, as in South Africa, for instance, implementation is much more tentative and precarious (Stout, 2008). Wales, too, is reported to have made progress toward a more inclusive and participatory approach to the delivery of youth justice (Nacro, 2009).

Social Work, Criminal Justice, and Antidiscriminatory Practice

As the 'welfare + rights' perspective recognizes, the argument for a social justice approach is further supported by the principles of antioppressive and nondiscriminatory practice, which are so central to social work (Dominelli, 1988). Not only is criminalization itself associated with inequality and discriminatory treatment, but the justice system, too, is characterized by practices that lead to additional harms.

Figures from the United Kingdom, for example, demonstrate that black young people are significantly overrepresented in the justice system (Ministry of Justice, 2010). This is also confirmed as a widespread problem internationally by the Global Commission on Drugs Policy (Stevenson, 2011). Discrimination is evident, too, in other respects, as the profile of those who make up the population of officially recorded offenders is shown to be very different in many respects from the population in general, for example, in respect of previous care histories, learning and communication difficulties, mental health status, income, and class background. Against this backdrop, it seems clear that social work with offenders must incorporate a commitment to advocacy and action on

behalf of those who experience the discriminatory effects of being criminalized, as well as the harms that are sometimes associated with the activities of the justice system itself (see Carlile, 2006, for example).

'What Works' in Social Work in Criminal Justice Settings?

There has been a considerable emphasis in recent years on the importance of providing evidence of 'what works' in criminal justice interventions, as in other areas of welfare service delivery. This is often narrowly defined in terms of reoffending rates, to the exclusion of those aspects of change in people's lives that might be thought of primary interest to social workers, such as improved quality of life, well-being, and social inclusion.

It is argued, for instance, that it may be unrealistic to expect probation supervision to make a substantial contribution to a reduction in reoffending, but it can help to build 'human capital' through provision of practical help, to increase employment prospects, or to improve family circumstances (Farrall, 2002). McNeill (2006) also argues that a narrow focus on correctional aims and methods is counterproductive, and unlikely to succeed even in its own terms. He argues instead that "some things 'work' with some people some of the time" (2006: p. 130), and this is perhaps unsurprising given the diversity of circumstances, characteristics, and needs represented by those who come into contact with the justice system. Instead,

> It is a recurring finding that no method of intervention is, in and of itself, any more 'effective' than any other; rather, there are common features of each intervention that are most likely to bring about positive change… These 'core conditions' of effectiveness include: empathy and genuineness; the establishment of a working alliance; and the adoption of person-centred, collaborative and 'client-driven' approaches. (McNeill, 2006: p. 131)

In other words, whether the objectives are to reform offenders, or to improve their lives and meet welfare needs, it is the application of core social work values and methods that will provide the basis for effective interventions. The precise nature of the intervention should be guided by the context and specific features of the offense in question, but the underlying principles and strategies should be derived from recognition of the need for inclusive practice based on a spirit of partnership and empowerment.

Social Work and Criminal Justice: Where Are We Now?

While there is some evidence from the typology and practices outlined above that social workers have taken a proactive and creative approach to negotiating a place for welfare practices in the criminal justice system, it also seems that this has been (and will be) a continuing struggle between those perspectives that locate 'the problem' with those identified as offenders, on the one hand, and on the other, those perspectives that recognize the role of social systems and structures in the process of 'constructing' criminals. The attempt to assert and establish the legitimacy of a distinctive language of welfare, need, and indeed, rights, against a system preoccupied with risk, responsibilization, punishment, and control will remain problematic. Where it has been possible to achieve an accommodation, it is usually on the grounds that a distinctive group (children, people with learning difficulties, or those with mental health issues, for example) are entitled to special treatment on the grounds of 'diminished responsibility'. For those who are held fully accountable for their actions, the case for welfare intervention is much harder to make and is reflected in practice by the continuing erosion of the social work element of probation service practice, internationally (Burke, 2011).

In accepting a place in the justice system on the basis of identifiable 'vulnerability' or 'special needs,' however, social work puts itself at risk of colluding with wider discriminatory processes and systemic disadvantage, and undermining its own core values. Whether or not some individuals and groups have particular needs, social work practice should not lose sight of the fact that the criminal justice system itself generates needs and creates problems for those who come into its ambit, and that therefore there is a legitimate place for welfare intervention and advocacy across the entirety of that system. When social workers identify aspects of the justice system as oppressive and in breach of human rights, then as a profession social work has a responsibility, in line with its core values, not to collude with, but to challenge unfair and discriminatory treatment of those classified as 'offenders.'

See also: Antisocial Behavior; Critical Social Work Practice; Probation, offenders and Risk; Social Work Theory.

Bibliography

Burke, L., 2011. A comparative imagination. Probation Journal 58, 311–316.
Carlile, Lord, 2006. An Independent Inquiry into the Use of Physical Restraint, Solitary Confinement and Forcible Strip Searching of Children in Prisons, Secure Training Centres and Local Authority Secure Children's Homes. Howard League, London.
Considine, J., 2003. The Maori restorative tradition. In: Johnstone, G. (Ed.), A Restorative Justice Reader. Willan, Cullompton, pp. 152–157.
Dominelli, L., 2002. Anti-Oppressive Social Work Theory and Practice. Palgrave Macmillan, Basingstoke.
Farrall, S., 2002. Rethinking What Works with Offenders: Probation, Social Context and Desistance from Crime. Willan, Cullompton.
The Guardian, 2007. Timeline: a history of probation, May 2.
Hendrick, H., 2006. Histories of youth crime and justice. In: Goldson, B., Muncie, J. (Eds.), Youth Crime and Justice. Sage, London, pp. 3–16.
Herzog-Evans, M., 2011. Judicial rehabilitation in France: helping with the desisting process and acknowledging achieved desistance. European Journal of Probation 3 (1), 4–19.
Madoc-Jones, I., 2008. Models of intervention. In: Green, S., Lancaster, E., Feasey, S. (Eds.), Addressing Offending Behaviour: Contexts, Practice and Values. Willan, Cullompton, pp. 128–153.
Maguire, M., Raynor, P., 2006. How the resettlement of prisoners promotes desistance from crime: or does it? Criminology & Criminal Justice 6 (1), 19–38.
McNeill, F., 2006. Community supervision: context and relationships matter. In: Goldson, B., Muncie, J. (Eds.), Youth Crime and Justice. Sage, London, pp. 125–138.
Muncie, J., 2009. Youth and Crime, third ed. Sage, London.
Ministry of Justice, 2010. Breaking the Cycle. The Stationery Office, London.
Nacro Cymru Youth Offending Unit, 2009. Youth Justice and Participation in Wales. Nacro Cymru Youth Offending Unit, Port Talbot.
Obondi, C., 2010. Effective resettlement of offenders by strengthening 'community reintegration factors': Kenya's experience. In: United Nations Asia and Far East Institute (Ed.), Resource Material Series No. 82. United Nations Asia and Far East Institute, Tokyo, pp. 58–65.

Patterson, G., 2012. Social Work Practice in the Criminal Justice System. Routledge, Abingdon.

Raynor, P., Robinson, G., 2009. Rehabilitation, Crime and Justice. Palgrave Macmillan, Basingstoke.

Scraton, P., Haydon, D., 2002. Challenging the criminalisation of children and young people: securing a rights-based agenda. In: Muncie, J., Hughes, G., McLaughlin, E. (Eds.), Youth Justice Critical Readings. Sage, London, pp. 311–328.

Smith, R., Fleming, J., 2011. Welfare + Rights: An Interim Evaluation Report into the Howard League's Legal Service for Children and Young People. Howard League, London.

Stevenson, B., 2011. Drug Policy, Criminal Justice and Mass Imprisonment. Global Commission on Drugs Policy, Geneva.

Stout, B., 2008. Children's rights, restorative justice and South African youth justice law reform. BSU/IUC Journal of Social Work Theory and Practice 15. http://www.bemidjistate.edu/academics/publications/social_work_journal/issue15/articles/stout.htm (accessed 18.06.13.).

Trotter, C., 2009. Pro-social modelling. European Journal of Probation 1 (2), 142–152.

van Wormer, K., 2002. Social work and restorative justice. Social Work Today 2 (1). http://www.restorativejustice.org/10fulltext/vanwormer (accessed 18.06.13.).

Criminal Justice, Sociology of

Myron Orleans, California State University, Fullerton, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Szabo, volume 5, pp. 2954–2958, © 2001, Elsevier Ltd.

Abstract

The relevance of sociology to the study of criminal justice is assessed and analyzed. The Criminal Justice System (CJS) is conceptualized and linked to broader sociopolitical factors and influences. Components of the CJS are examined in light of sociological perspectives. Trends and directions of research, theory, and policy formation are explored.

In the last 40 years and especially since the turn of the twenty-first century, the study of criminal justice (CJ) has burgeoned with increasing numbers of programs at all academic levels devoted to this specialty. This trend has resulted in a progressive growth of CJ doctorates and concomitant reductions in the influence of other disciplinary professionals on the course and trajectory of theory, research, and policy formation in the field. While sociology departments have retained perhaps the most substantial outsider effect on CJ study of all disciplinary approaches, some scholars suggest that this is a time-limited effect that will diminish as CJ establishes its full autonomy (Wrede and Featherstone, 2012). Alongside the remarkable development of CJ programs, sociology departments have retained much of their criminology curricula, resulting in extensive duplication of content (Maddan and Hartley, 2011). Students and scholars with a greater policy orientation tend to gravitate to CJ specialization while sociological criminologists, especially those theoretically inclined, tend to study the contextual, ideological, and cultural dimensions of both crime and its study by CJ practitioners (Hauhart, 2012).

This article will utilize sociological perspectives to examine: (1) the roots of the idea of a criminal justice system (CJS); (2) the implications of this notion in practice; (3) its elements and functions; (4) its relation to the courts; (5) the distinction of crime and deviance; (6) implications for corrections; (7) the interplay of legal and medical models; (8) research priorities in the field; and (9) future directions in the CJS. The emphasis here will be on how a sociological approach affords certain analytical benefits not achievable from a purely criminal justice perspective. The intent here is to demonstrate the independent utility of a sociological study of criminal justice while not challenging the legitimacy of the concrete study of this area. A distinction should be drawn at the start between criminal justice as a study and the CJS. Criminal justice studies the CJS for the purpose of preparing and informing practitioners, enhancing the effectiveness of programmatic efforts to address the social problems of crime by providing data, explanation, prediction, and control. Its task is to support the criminal justice providing the broader and more fundamental surround within which criminal justice operates as both study and system. To posit that there is in fact a CJS is to construct it ontologically as an interconnected, interactive, functional feature of a broader social system. Whether criminal justice does indeed function systemically is itself a very general but still empirical question, which has yet to be fully or persuasively established by the field of criminal justice study. Indeed, the legitimacy of criminal justice study does not depend on the actuality of the CJS, but it does generate a paradigmatic approach that an emerging field of theory and research might use effectively to organize its agenda.

The Origins of the Concept

CJS is a fairly recent acquisition of the scientific and professional vocabulary. Since the publication of the United States, President's Commission on Law Enforcement and Administration of Justice report 'The Challenge of Crime in a Free Society' (1967) (https://www.ncjrs.gov/pdffiles1/nij/42.pdf), the term has been used with a somewhat precise meaning.

However, in the process, the Commission's investigation led to the claim that the goal of the 'war on crime' – the fight against insecurity and the foundation of public safety of the persons, goods, and moral integrity of the citizenry – necessitated a comprehensive concept inclusive of all the institutions concerned in this objective. The breadth and complexities of this conceptualization led to the identification of incongruities between some of its parts, allowing critics to call it a 'nonsystem,' referring to the apparent incoherences and contradictions in the functioning of the overall system.

It became evident for the Commission that there existed a real and a virtual interrelationship, reflecting an interdependency between legislations, institutions, organizations, and scattered offices throughout the machinery of criminal justice, whose goal is the maintenance of public order, the safeguard of liberty, physical integrity, and concern for the victims of the failings of the protective shield, extended by the law and protected by the Constitution. Recognizing these historically and nonlogically ordered institutions, laws, services, and so forth, the Commission wanted to give a degree of coherence to all of them in order to secure a better, more efficient coordination, built around a common goal: assuring, under the supremacy of the law and the Constitution, equal justice and protection to everybody. Despite the criticism, the concept of the CJS was born. Thanks to this organizational principle (a 'system' necessitates some logic), overlapping jurisdiction, duplication of services, and irrelevant goals or functions were identified. The Commission's logo – with its input and output/entrance and exit design – illustrated graphically how, from arrest to release, from sentence to punishment, the whole process of

criminal justice constituted an administrative and logical 'whole.' This holistic approach provided the reformer with a blueprint and a compelling agenda for reform.

However, these origins of the CJS should not hide the basic fact that the inspiring and commanding center of their systematizing approach can be understood as a development of the *code d'instruction criminelle of 1808*. This code was characterized by: (1) rigorous penal procedural rules; (2) the availability of material and technical conditions to implement those rules; (3) a reasonable degree of coherence among the constitutive elements of the system, principally the offices of the prosecutors, magistrates, and other officers of the judiciary. Similar transformations took place in other Western countries that incorporated a systematic approach. The principle of the separation of powers, the presumption of innocence, and the independence of the judiciary were nowhere threatened by the emergence of the CJS, with the notable exception of course in the totalitarian countries.

Practical Impact of the Concept of the CJS

A radical transformation in public administration began as early as the middle of the nineteenth century, accelerating after World War I and continuing beyond World War II. One result was the emergence of an activist and interventionist social policy. A similar transformation in criminal justice administration did not occur until later. In the process of its consolidation, it confronted the cultural revolutions of the 1960s and 1970s, and even some basic principles of the penal procedural law underwent some changes.

Among the consequences of the development of legal structures and philosophies were a loosening up of the system, allowing an increased role for a defense attorney (human rights legislation had a major impact on the proactive role of the defense councils), and resulting in an increase in the burden carried by law enforcement agencies to accommodate the increasing demands of defense attorneys as well as rules edited by commissions of human rights (the need for 'foolproof' proofs required by tribunal became more cumbersome and proceedings slowed down notably). The increase of the workload in the prosecutor's offices, as well as at the courts, was not in step with the necessary budgetary requirements allowing a decent discharge of the increased workload devoluted to the CJS in charge of the 'postarrest' phases. The well-known, yet little understood, consequences of overloaded courts; plea bargaining; vagaries of correctional strategies, inequalities, and extravagant; burdensome cost increments escalated from the 1980s onward.

Additionally, there was a growing incoherence between segments inside the diverse components of the CJS. Undue competition has arisen between services, which should have been submitted to rules prescribed by their common goals. Moreover, a growing politicization has taken place. The power of unions is increasing everywhere: the judges, the prosecutors, and the police are organized in their own unions, and correctional officer unions in some states, particularly in California, constitute major veto groups. Some unions of magistrates, particularly in France and Italy, are playing an active and sometimes radical role inside of the CJS. Thus the coherence of the CJS has suffered from the centrifugal forces in its own ranks. This worldwide phenomenon is affecting open societies, and while the European continental nations, in-line with their centralizing traditions, are seeking new measures in tune with their traditions, all of them are also facing similar problems. Problems of CJS incoherence have become exacerbated by severe budgetary constraints since 2008, particularly in the US with its comparatively complex legal/juridical system, diverse demographics, diverging yet stringent crime philosophies, and harsher sentencing practices.

Components and Goals of the CJS

The 1967 Presidential Commission, which is *de facto* the originator of modern research on crime, delinquency, victimology, and criminal justice agencies, introduced and somehow accredited a rather artificial distinction between research carried out on the CJS and criminology. This discipline was related, if not assimilated, to etiological research on crime causation. Its continental origins were stressed. Criminal justice research, born of the common law tradition, postulated the existence and interdependence of all components of the CJS, including law enforcement, criminology, criminal justice administration, prosecutor's office management, sentencing guidelines for judges, and so forth (See Section Criminal Justice Research).

Many of the major changes that the CJS underwent are related to the sudden surge of the world economy and commerce in the direction of globalization. Accordingly, the problems facing the CJS became global, embracing the whole world. One has to face the fact that crime and justice have to be addressed in an open world, at the level of the global economic and financial system working with the tools of the newest information and technology. The sociopolitical system and the political power structures are closely related to the new financial, commercial, and economic trends and powers. The magnitude of the challenge facing all the components of the CJS is tremendous. Consider, after 1980, that the volume of the sales of foreign stocks and bonds increased drastically on the market (Adler et al., 1994). Moreover, capital movements, legal and nonlegal labor migrations, corporate alliances and partnerships, proliferation of tax havens, money-laundering practices and controls, cyber attacks, hacking, pirating, and so forth, have rendered many crimes global in nature.

However, many countries are largely decentralized and have residues of common law traditions. There, the integration and internal coherence of the CJS is limited; rather the tendency is toward internal incoherence and diversity. Thus it is not surprising that these countries, illustrated by the USA, are much slower to ratify or contribute to international criminal and penal jurisdiction. An example is the US government's refusal to adhere to the Rome convention sponsored by the UN to create an International Criminal Court (ICC) although a working relationship of engagement has been established since 2010 (http://www.state.gov/j/gcj/us_releases/remarks/2010/143178.htm). In the second decade of the century, the US has extended greater efforts to collaborate with international bodies such as ICC but a delicate line has to be threaded

due to both the mode of operation of the utter complexity US crime control policy and political imperatives and ideologies.

CJS and the Courts

In this context of global integration, the thrust toward internationalization increasingly impacts local crime scenes (arms dealing, Internet crime, terrorism, migration, etc.) so that, directly or indirectly, they become integrated into international activities through the operation of offshore banking; industrial espionage; unofficially sponsored hacking into corporate, financial, and national security institutions among other devices; intermingling, sometimes, with secret service operations (e.g., 'evidence' of Saddam Hussein's WMDs) (Weiner, 2008); and with illegal merchandising or human trafficking, tax frauds, illegal financial contributions to political campaigns or parties, and so forth. Money laundering is the lifeline of all these criminal and illegal activities. Out of necessity, a new international CJS is emerging with the support of the UN and the European Union (EU) (http://www2.icc-cpi.int/Menus/ICC/About+the+Court/). Other tribunals in Africa and Asia are in process of being established. Especially in the face of international terrorism, support for such tribunals is growing in the US along with opposition to internationalization of the CJS (http://voices.yahoo.com/moving-towards-international-criminal-justice-system-649697.html).

One can see that open democratic societies are building new lines of defenses using legislations and sharpening legal tools available to the police and prosecuting attorneys. A new Europe-wide legal system is being created and is slowly and systematically building structures and defenses, which are both reactive and proactive. The building of such a coherent system was, in part, a response by the CJS to challenges initiated in continental countries under the auspices of the EU (http://www.haguejusticeportal.net/?id=2). Threats to the Euro currency by failures of and fraud by financial institutions, national banking crises, and fear of extreme exchange rate fluctuations also are precipitating higher levels of cross-national collaboration than in previous decades.

The Debate on the Definition of Deviance and Criminality

In the 1970s, sociology questioned what distinguishes a 'criminal' act from a 'deviant' one, thereby adding a substantial burden to the functioning of the CJS. Such questioning created difficult dilemmas for the CJS, whose task is to enforce the existing laws. Problems arose as to where to draw the limits of the interventions of the jurisdictions of the CJS agencies in controlling or repressing the 'criminal' and unlawful activities of large, and certainly vocal, minorities. The frontiers between crime and deviant behavior became blurred. The law enforcement agencies, occupying the first line of defense in securing public safety, became hesitant in the application of the law. This environment of questioning and doubt encouraged the increasing activism of human rights defense groups, critical of the law. And when the police tried to enforce questionable laws, they were frequently confronted by activist groups, sometimes with violence. Many court actions followed violent confrontations in the streets. Also, accusations against the police for being prejudicial created a major impediment to efficient policing.

Variations in local CJS and community relations, contrasting values between law enforcement personnel and civilians, and inconsistencies of enforcement, prosecutorial, and judicial and correctional policies have all contributed to confusions, conflicts, and unpredictability of the entire range of crime response approaches. External review boards, internal examiners, accountability requirements, unending assessments, and so forth all encumber the task of securing public safety within a civil rights framework.

The manifold increase in the charges burdening the prosecutor's offices, mainly in major cities, reflected itself in the overflow of cases brought in front of the judges. Plea bargaining before the trial (essentially a kind of out of court settlement between the prosecutor and the defense council) became an accepted way of proceeding. For several decades it has been the only expedient to avoid the complete paralysis of the whole judicial process. These plea bargains have resulted in some gross miscarriages of justice that have become broadly recognized (Bibas, 2004).

Morally as well as legally, and sometimes clearly illegal, plea bargaining is still currently and fairly widely practiced in the US courts system. One notes that sociologists frequently criticize this practice, while many CJ professionals accept the expedient as at least a necessary one. Certainly in the US there is more support for plea bargaining than elsewhere (http://www.debate.org/opinions/does-plea-bargaining-undermine-the-criminal-justice-system). In contrast to the US, these practices are strongly rejected by the European continental legal system, but the future of its integrity remains an open question. The caseload of the courts is increasing constantly and political power negatively confronts the judiciary in several important Western European countries. The main issues are political corruption, bribery, financial deception, and abuse of power. We have yet to see the outcome of this type of conflict inside of the continental system where political power negatively confronts the CJS.

Changes in the Philosophy of Punishment

Imprisonment was a notable advance over the all-too-common previous method of punishment: the death penalty. Virtually from its beginnings imprisonment developed in response to religious concerns (largely from nonconformist denominations – Methodist, Congregationalist, Presbyterians, Quakers, Baptists, etc.), and was informed by ideas of the continental Enlightenment, as well as English social contract and utilitarian theories, and the views of Scottish moral philosophers.

In the north-east sector of the US these religious and intellectual forces combined to encourage a massive social experiment in punishment, which was grounded in the emerging belief that humans are social beings possessing 'natural' or inalienable rights, capable of reason, and inherently good or virtuous: the infant cannot be evil. From this perspective, the logic of imprisonment would not only punish the wrongdoer,

but seek also to rehabilitate that wrongdoer, as well as to deter others. Elmira, Auburn, and Cherry Hill were famous attempts to express this evolving penal philosophy, and de Tocqueville's tour of these institutions was an example of European continental interest.

This complex and integrated humanistic view, issuing from the Enlightenment, was challenged by developments in sciences such as medicine and psychiatry. These disciplines understood criminal acts to arise from personality disorders rooted essentially in social, biological, or psychological conditions. In a sense, the 'medical model' retained but transformed elements of the 'humanistic model.' The goal of rehabilitation was retained, but its justification was to be found and legitimized in scientific methodology. One consequence of this scientific approach, as it gained influence, was to impact the CJS through legislative reforms and judicial practices, for example, the rapid spread of indeterminate sentences, and the transferal of the principle of rehabilitation from being a concern of the larger society (religious activists, legislators, educators, etc.) to become the concern of scientific practitioners encompassing all sorts of therapies, treatments, and rehabilitative initiatives.

Emergence and Hegemony of the Medical Model: Conflicts of Goals with the Legal Tradition

The nature of the scientific method is the permanent questioning of existing proofs and practices. For example, in medicine, it results in a continuing revision of therapeutic practices and rehabilitative techniques. Naturally, a continuing tension and conflict develops between the ever-changing medical therapeutic approach and the desire of the CJS to retain humanistic principles of equity, rights, and equal treatment before the law. This situation threatens its coherence and creates an increasing vulnerability for the CJS, especially to its claims of legitimacy.

The challenge to legitimacy was not limited to the confines of the CJS, but permeated the entire open, pluralistic, humanistic society. Everywhere, a profound crisis of values occurred: in the family, and in institutions of authority, including the medical and academic communities. The identification of who or what was responsible for this turmoil emerged from a variety of diverse sources, making it impossible for a generally agreed upon understanding to be reached, and so disarray prevailed.

The impact on the CJS, which became closely associated with the medical model, was disastrous in that it allowed for both an immunization of the 'underclass' against claims of criminal behavior, and the emergence of new prophets with new social and cultural constructs. An attempt to correct this situation, and to devise a means of bridging the chasm separating the reality of crime and the CJS, was Von Hirsch's (1976) doctrine of 'just deserts.' This doctrine compromised law and order while accommodating the new forms of behavior. The end result was a temporary compromise in which the CJS maintained and protected the core values concerning material and personal security, while adapting itself to the demands of the 'post-Woodstock' society.

Reactions counter to the delegitimation process rose in the 1990s and dominated thinking for about 20 years. This reaction emphasized a return to a simpler certainty about the proper response to crime with determinate sentencing, reduction of funding for rehabilitative and educative focus, de-emphasis on alternative correctional approaches, treating juvenile offenders as adults, etc. More recently, efforts have been directed toward addressing deep and long-term concerns regarding overcrowding, high rates of recidivism after long sentences, juvenile transitioning to adult crime, funding reductions for corrections, gun violence, victim support, and punitive approaches to drug-related offenses (http://www.thecrimereport.org/news/inside-criminal-justice/2012-04-crime-and-justice-in-america-1975-2025).

Criminal Justice Research

While criminological research of a sociological stripe has focused on etiological areas, CJS research has been primarily institutional. As in all such distinctions, the differences in fact are less determinative than in definition. 'Corrections,' a new designation for penology, was a major aspect of research on the CJS. These distinctions also reflect a growing attention to the evolution of ideologies, to new tendencies of penal philosophies, justifying research on alternatives to the actual CJS, which answers to the needs of victims and new sentencing philosophies. The analysis of the performance of the CJS, evaluation of alternatives for this system, and all kinds of innovations, were also an integral part of research on the CJS. Also, the treatment of offenders, their resocialization, programs of crime prevention, and experiences with 'restorative justice' became the spearheads of research in criminal justice.

Some Future Perspectives

The real and symbolic ruins that were the reality of the CJS under the conditions of this temporary compromise were, fortunately, addressed by H.L. Packer (1968). Packer identifies two models. The first is the 'due process model,' characterized by the jurisprudence of the Warren Court, in which most of the odds were placed against the defender of public safety and order. This model amplified the 'rights' of the individual, and encouraged a critical, reactive attitude on the part of the judiciary. This model has a rule-oriented rather than a substantive-oriented approach.

The second is the 'crime control model,' which emphasizes the existence and exercise of official power. Its validating authority is ultimately legislative (although approximately administrative). It can be viewed as consistent with, and a continuation of, humanistic-Enlightenment principles. In light of the lessons of the eighteenth century, and because Packer's theories are deeply rooted in humanistic-Enlightenment principles and now associated with conservative political orientations, his crime control model prevailed in the recent past but has come up against popular sentiment that counters mass incarceration, particularly for nonviolent offenders. A synthesis of models seems to be taking shape as crime control remains a significant concern, especially in relation to transnational crime, while alternative means of criminal justice are coming to the fore, including drug courts, restorative justice, community-based corrections, electronic supervision, community and military service, etc.

Conclusion

In conclusion, the contributions of the social sciences to the CJS are varied. In continental Europe, the social sciences, as an academic discipline, had limited influence, notwithstanding contributions of outstanding scholars. In North America the social sciences have established a strong academic position and exercised a direct influence on the CJS. This influence also included a contribution to the intellectual and moral crises, which characterized the 1960s and 1970s. This was largely due to an epistemological crisis, which shook the very foundations of social and behavioral sciences. Whether or not they will recover their empirical scientific tradition is still an open question.

That crisis has largely been resolved in the intervening decades as varieties of criminological and criminal justice studies have applied postmodern, poststructural, cultural, peacemaking, radical perspectives, etc., as well as conventional crime control approaches in numerous and proliferating theoretically informed empirical studies. As the study of the CJS has grown and absorbed diverse tools, sociological influences have both expanded and occupied a distinctive, even privileged location in this emergent, no longer nascent field (Eagle and Barnes, 2008). Alongside other disciplines including psychology, anthropology, biology, sociology have contributed substantially to the advancement of criminal justice study and to the CJS itself, and will continue to do so deeper into the twenty-first century.

See also: Control, Social; Courts and Adjudication; Crime and Class; Crime, Geography of; Crime, Race and Ethnicity; Crime, Sociology of; Crime: Knowledge about and Prevalence; Delinquency, Sociology of; Imprisonment: Sociological Aspects; Justice and Law; Juvenile Justice: International Law and Children's Rights; Law, Sociology of; Law: Economics of Its Public Enforcement; Police: A Sociology of Knowledge Approach; Policing; Punishment: Comparative Politics of; Punishment: Social and Legal Aspects; Rehabilitation of the offender.

Bibliography

Adler, F., Mueller, G.O.W., Laufer, W.S., 1994. Criminal Justice. McGraw-Hill, New York.

Bibas, S., 2004. Plea bargaining outside the shadow of trial. Harvard Law Review 117 (8), 2464–2547.

Bridges, G.S., Weis, J.G., Crutchfield, R.D. (Eds.), 1996. Criminal Justice Readings. Pine Forge Press, Thousand Oaks, CA.

Cohen, J., 1983. Incapacitation as a strategy for crime control: possibilities and pitfalls. In: Tonry, M., Morris, N. (Eds.), Crime and Justice: An Annual Review of Research. Chicago University Press, Chicago, IL.

Cullen, F.T., Gilbert, K.E., 1982. An American History of Correction. Anderson Publications, Cincinnati, OH.

Eagle, S.J., Barnes, J., 2008. Reexamining criminal justice 'star power' in a larger sky: a belated response to Rice et al. on sociological influence in criminology and criminal justice. Journal of Criminal Justice Education. 19 (2), 213–226.

Hauhart, R.C., 2012. Toward a sociology of criminological theory. American Sociologist 43, 153–171.

Jacoby, J., 1979. The changing policies and prosecutors. In: McDonald, W.F. (Ed.), The Prosecutor. Sage, Beverly Hills, CA.

Maddan, S., Hartley, R., 2011. Lawyers practicing medicine: criminal justice, criminology, sociology and differential curricula. Journal of Criminal Justice Education 22 (3), 440–465.

Maguire, M., Morgan, R., Reiner, R. (Eds.), 1994. The Oxford Handbook of Criminology. Clarendon Press, Oxford.

Menninger, K., 1968. The Crime of Punishment. Viking Press, New York.

Mouledoux, J., 1972. E. Livingston (1764–1836). In: Mannheim, H. (Ed.), Pioneers in Criminology, second ed. enl. Patterson Smith, Montclair, NJ.

Packer, H.L., 1968. The Limits of the Criminal Sanction. Stanford University Press, Stanford, CA.

Schrag, C., 1961. Foundations for a theory of corrections. In: Cressey, D.R. (Ed.), The Prison: Studies in Institutional Organization and Change. Holt, Rinehart & Winston, New York.

Sykes, G.M., 1958. The Society of Captives. Princeton University Press, Princeton, NJ.

United States. President's Commission on Law Enforcement and Administration of Justice, 1967. The Challenge of Crime in a Free Society. U.S. Government Printing Office, Washington, DC.

Von Hirsch, A., 1976. Doing Justice: The Choice of Punishment, first ed. Hill and Wang, New York.

Weiner, T., 2008. Legacy of Ashes: A History of the CIA. Anchor, New York.

Wrede, C., Featherstone, R., 2012. Striking out on its own: the divergence of criminology and criminal justice from sociology. Journal of Criminal Justice Education 23 (1), 103–125.

Relevant Websites

http://www.everydaysociologyblog.com/2011/07/the-criminal-justice-system-in-context.html – Blog on the CJS in Context.

http://www.rethinking.org.nz/assets/Newsletter_PDF/Issue_101/NAO_Briefing_Comparing_International_Criminal_Justice.pdf – Briefing to House of Commons England and Wales 2012: Comparing International Criminal Justice Systems.

http://www.civilrights.org/criminal-justice/ – Civil and Human Rights Coalition of the Leadership Conference.

http://www.justice.gov/cjs/ – The CJS as a Counter-Terrorist Tool.

http://manhattanda.org/criminal-justice-system-how-it-works – Criminal Justice System: How It Works in New York County.

http://www.nami.org/Content/ContentGroups/Helpline1/Dealing_with_the_Criminal_Justice_System.htm – National Alliance on Mental Illness on Dealing with the Criminal Justice System.

http://thesocietypages.org/sociologylens/2012/07/03/deinstitutionalization-and-the-criminal-justice-system-an-unfortunate-correlation/ – Sociology as a Stem Discipline for CJS.

http://globalsolutions.org/law-justice/international-criminal-court – US Support for an International Criminal Justice System.

Criminal Liability and Criminal Defenses

Gerry Ferguson, University of Victoria, Victoria, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

All countries have criminal laws. Those laws define certain behavior that is forbidden – not just any behavior, but behavior that violates a country's core values and interests. Criminal laws also set out the punishments for those who violate the criminal laws without lawful authority, justification, or excuse. This article examines the principles that govern criminal liability and the defenses that allow a person to avoid conviction and punishment. A fundamental principle of criminal liability is that no person shall be convicted of a crime unless that person has the ability to reason right from wrong and the freedom to choose whether to do right or wrong. Many specific defenses find their rationale in these dual abilities. Put another way, there must be both unlawful conduct and an unlawful state of mind when engaging in that conduct. With that in mind, this article will look at defenses of automatism, sleepwalking, and other forms of involuntary conduct, mistakes as to an essential element of an offense, intoxication, insanity, provocation, duress, and justifications such as use of lawful force, self-defense, and necessity.

Criminal liability requires proof of three interrelated components: (1) commission of the prohibited conduct specified in the offense (also called *actus reus*), (2) committed with a particular mental state (also called *mens rea*), (3) without a recognized legal defense. Broadly conceived, criminal defenses include all legally recognized factors or rules that prevent a person from being convicted of an offense. Defenses may negate the elements of *actus reus* or *mens rea*, or may operate notwithstanding the fact that a person has committed both the *actus reus* and the *mens rea* of an offense. This article begins with the observation that while there are a number of different legal systems in the world that differ very significantly in many respects, when it comes to basic principles of criminal liability and defenses, they are surprisingly more similar than different. After providing a few other preliminary comments on criminal liability, this article then provides an overview of defenses in general, including a bit more detailed analysis of a few defenses such as provocation, self-defense, and insanity, which have been particularly controversial in many countries around the world.

Preliminary Comments

Major Legal Systems and Criminal Responsibility

Today there are four major legal systems operating in different parts of the world, not counting Hindu law in parts of India's legal system and customary law operating in whole or in part in various parts of the world (Menski, 2006; Kritzer, 2002). Those four systems are common law, civil law, Islamic law, and socialist law systems.

With extensive colonization in the eighteenth and nineteenth centuries of North, South, and Central America, Africa, South Pacific, and Asia by England and other European countries, especially France, Spain, and Holland, the common law and civil law systems, and in particular their criminal laws, were adopted or imposed in most countries throughout these regions (Ferguson, 1998; Petchsiri, 1987). In many instances, these laws and legal systems remained in place after the colonizers left and the countries had declared independence. The legal systems noted above are different from one another in a multitude of ways, including substantively, procedurally, and institutionally. While each legal system has criminal (penal) laws, the procedures for investigating, prosecuting, trying, and punishing crimes are significantly different in many respects in each system. What is similar is that each country creates criminal laws that are designed to protect fundamental values in that country by forbidding certain behavior that threatens or violates those values. In this regard there is a lot of commonality between these legal systems as to basic crimes against persons and property (murder, rape, robbery, arson, fraud, etc.). On the other hand, there are significant differences in crimes against public morality, state order, and state security when one compares socialist, Islamic, and civil law/common law criminal laws (Butler, 2003; Zhu and Han, 2013; Lindsey, 2008).

In regard to the subject of basic principles of criminal responsibility, there is more similarity than difference between nations. All systems recognize the same three components for criminal liability: prohibited conduct (acts or omissions of legal duties), blameworthy mind, and absence of any legal justification or excuse (LaFave, 2003, vol. 1; Ashworth, 2009; Bohlander, 2009; Butler, 2003, Chapter 17; Badar and Marchuk, 2013). In short, each system operates on the principle that a person deserves to be convicted and punished when he/she knowingly and willingly violates criminal prohibitions without any lawful excuse or justification. Badar and Marchuk (2013) make this same point in their comparison of principles of criminal responsibility in common law, civil law, Islamic law, and socialist legal systems such as Russia and China.

Questionable Assumptions about Autonomy and Criminal Liability

The criminal justice systems around the world continue to base decisions on criminal liability and punishment on rather simplistic views of humans as rational and autonomous persons. For Kant and many other moral and religious

philosophers, the belief that humans generally have the capacity to reason right from wrong and the capacity to choose right or wrong justifies imposing criminal responsibility and punishment on persons who chose to do wrong.

Sadly, the law's current approach to autonomy or capacity to choose has remained largely unchanged for the past two or three centuries. In this regard, the law acts as if the past 150 years of social science inquiry and knowledge about human conduct had never occurred. The alternative notion that human behavior could be more a function of biology and environmental forces than free choice is a direct threat to a society's deep-seated need to blame someone other than themselves for criminal harms that occur (Bargh, 2008; Montague, 2008; Jones, 2003; Friedland, 1998; Johnson, 1998; Grant, 2006; Morse, 2004).

That is a form and degree of blame which most societies do not seem ready to accept. It is easier and simpler for a society and its law to assume that people who can reason – i.e., know what they are doing – can also choose to do or not to do the criminal act in question. In that sense, the individual who is found criminally liable is a convenient scapegoat for societal inaction or neglect. As Herbert Packer (1968: 74–75) famously said, "The law treats man's conduct as autonomous and willed, not because it is, but because it is desirable to proceed as if it were."

While the law may not provide a substantive defense – i.e., a full or partial acquittal – to a person whose criminal conduct is largely influenced by biological or environmental factors, our sentencing laws are somewhat more forgiving. A sentence is, among other matters, the law's denunciatory and punitive response to a person who has been found guilty of a crime and therefore deserving of blame for causing that criminal harm. External factors are often viewed as mitigating circumstances when a judge engages in the difficult task of determining a fit and proportionate sentence. Notions of individual blameworthiness allow for the consideration of at least a crude account of other influencing factors in determining 'how much' the offender is to blame and therefore 'how much' punishment he/she deserves (von Hirsch and Ashworth, 2005).

Criminal Liability and Defenses

Defenses negate criminal liability, either fully or, in some cases, partially. Thus defenses are central to a determination of criminal liability. There are a number of different ways to distinguish and classify defenses. The first way is to organize defenses with regard to how they operate on the three elements of criminal liability: they can negate the existence of the prohibited conduct, they can negate the existence of the requisite blameworthy mind, or alternatively they can apply notwithstanding the existence of the prohibited conduct and the required state of mind. Since the presumption of innocence normally requires the state to prove both the prohibited conduct and the blameworthy mind before imposing a finding of guilt, defenses that negate the prohibited conduct or blameworthy mind are also referred to as 'failure of proof' defenses in some jurisdictions (Robinson, 1984).

Secondly, defenses can also be classified as exculpatory and nonexculpatory. In other words, some defenses exist to reflect society's view that the actor does not deserve to be found criminally liable for his/her conduct because the actor's conduct or state of mind was exculpatory, i.e., not blameworthy in the circumstances in which the conduct occurred. For example, applying force to another person in lawful self-defense or while insane renders the actor's conduct exculpatory or nonculpable, i.e., not deserving of criminal liability and blame, albeit for very different reasons. On the other hand, some defenses exist notwithstanding the actor's blameworthiness. These defenses are justified on the basis of other policy objectives. For example, diplomatic immunity and statutory time limitations on prosecuting crime are two examples of nonexculpatory defenses. Those defenses exist to pursue policy objectives that are deemed to be of greater general importance than convicting an otherwise blameworthy actor. These defenses will not be further discussed in this article due to space limitations.

Thirdly, exculpatory defenses are sometimes further classified as either justifications or excuses. Justifications render *conduct* lawful and therefore not a crime; excuses render the *actor*'s otherwise unlawful conduct not deserving of punishment. Justifications focus on the *legality of the act*; excuses focus on *the blameworthiness of the actor* (Fletcher, 1978; Bohlander, 2009). The above three classification schemes for defenses can be a useful way to compare and contrast different defenses. However, these categories should not be seen as airtight compartments, and some jurisdictions, like Russia (Butler, 2003), do not subdivide their exculpatory defenses into justifications and excuses. In addition, the same defense may be classified into one category or another in different countries because the requirements for that defense can vary from one country to another (Robinson, 1984, vol. 1).

Overview of Defenses

The following overview of defenses is based primarily on the first distinction, namely the manner in which defenses operate in negating *actus reus* or *mens rea*, or operate notwithstanding the existence of *actus reus* and *mens rea*.

Defenses Negating *Actus Reus*

The *actus reus* or prohibited conduct of each offense is described in the statute creating the crime. Most jurisdictions have abolished unwritten common law crimes, though England is an example of a country that still has a number of them (Ormerod, 2011; Baker and Williams, 2012). Violation of prohibited conduct in a penal statute or code will constitute a crime unless there is a structural defect in the offense itself. Thus before being convicted of a crime in a code or statute, the offense in question must be lawfully enacted and proclaimed in force, not be applied retroactively, nor be unduly vague or a violation of constitutional rights such as free speech, privacy, and security of the person. If the offense in question passes those structural defenses, then the analysis of criminal liability will shift to the existence of the *actus reus*. The *actus reus*, depending on the crime in question, can be broken into the following elements:

1. an act and/or an omission of a legal duty to act,
2. which is voluntarily committed,

3. which occurs in the prohibited circumstances described in the offense, and
4. causes or results in any harm described in the offense.

In all legal systems, the prosecutor is normally required to prove each of these *actus reus* elements although there may be limited exceptions to that rule for one or more *actus reus* elements for some offenses. Thus a*ctus reus* defenses negate an essential element of the prohibited conduct and harm and they include the following:

1. *Not me*. The person charged does not deny that an offense occurred, but claims instead "it was not me who committed it." The accused may present evidence of an *alibi* (it could not be me, because I was in a different place at the time of the offense), or may argue that the eyewitnesses are mistaken, or that other witnesses are lying or that the forensic evidence obtained from the crime scene or from the accused is not conclusive in proving he was the person who committed the offense.
2. *No legal duty to act*. In general, criminal liability arises out of acts, not omissions, and this is particularly the case in common law countries. But all countries impose liability for some omissions. The question then is: in what circumstances does the law impose a duty on a person to act? For example, it is common in all countries to impose a duty on parents to provide care and protection to their children, or for a jailor, nurse, or guardian to provide care and protection to those in their charge. Failure to fulfill these legal duties will result in criminal liability if that omission facilitates or results in criminal harm. Common law countries (and some socialist system countries (Butler, 2003)) do not usually impose duties of care on persons who are not in a special relationship with one another. Civilian continental jurisdictions such as Germany (Bohlander, 2009) and France (Cairns and McKeon, 1995) do have broader legal duties to take action to save a stranger from serious harm or death if that action can be safely undertaken with little or no cost or inconvenience. But notwithstanding the differences in the nature and extent of legal duties in different countries, it is always a defense in every country if the prosecutor fails to prove that the accused omitted a legal duty imposed on him or her.
3. *The prescribed harm or circumstance has not occurred*. The definition of each offense sets out the harm or circumstances that must exist for that offense to occur. For example, the accused may have possessed sexually explicit material but that does not constitute the offense of possession of obscene material unless the sexually explicit material also meets the legal definition for obscenity in that particular country. Likewise, the offense of assault causing serious bodily harm is not proven if the complainant only suffered minor harm as a result of the accused's assault on him. Thus denial of the prescribed harm or circumstances is a very offense-specific defense.
4. *My act or omission did not cause the prescribed harm*. Some offenses only exist in circumstances whether the accused's act or omission has caused the prescribed harm. For example, the various forms of unlawful homicide – whether murder, manslaughter, or death by criminal negligence – all require the accused to commit an act, or omit a legal duty, that causes the death of another person. Thus if a wife secretly administers a deadly poison to her husband with the intention of collecting money from his life insurance policy, but the husband dies from an unrelated heart attack before the poison takes effect, then the wife's act of poisoning did not cause or contribute to the husband's death. Much more difficult causal cases arise in law when the accused's conduct has played some role, but not the primary role, in causing the prescribed harm. Different countries have different standards for 'how much causal connection to the prescribed harm' is required in order to hold a person criminally liable for that harm. Clearly the accused's act does not have to be the only cause, or the main cause. In many countries it is sufficient if it is a substantial and operating cause, or a significant contributing cause, or a real cause, or a proximate cause, or a cause that is more than trivial. If the accused's conduct does not meet a country's requisite causal standard, the accused is not guilty of that particular offense.
5. *My act or omission was not voluntary*. It is a widely accepted fundamental principle that acts or omissions are not criminal unless they are voluntary. But the word voluntary has a narrow legal meaning in this context. It refers to physical voluntariness in which the accused's conduct is a willed, physical act or omission in the sense that the person has consciously chosen to act or omit to act. By that definition, the accused's conduct is not voluntary if it is performed while the accused is unconscious or sleepwalking. Involuntary unconscious conduct is also referred to as automatism. Conduct can also be involuntary in some cases even though the accused is conscious, such as where the action is caused by a force beyond the actor's control.

Defenses Negating *Mens Rea*

For several centuries a Latin maxim has been used in many legal systems to capture two of the essential requirements for criminal liability – *actus non facit reum nisi mens sit rea*. Literally translated, this means "the act does not make it a crime unless the mind is also criminal." This maxim encapsulates the fundamental principle of fault – acts or omissions that cause harm are not subject to criminal liability unless they are done with a guilty mind. While the precise meaning of a guilty mind has changed over the centuries, there are nonetheless two broad categories of mental fault: (1) subjective and (2) objective. Subjective fault normally includes intention and recklessness. In cases of intent, the accused knows what he/she is doing and knows that it may cause or bring about the prohibited consequences or circumstances. In doing the act, the accused means to or intends to cause the prohibited harm. In cases of recklessness, the accused foresees the harm as a possible consequence of his conduct but does not necessarily want the harm to occur. Nonetheless the accused does the act for another purpose, recognizing that it is either possible or probable that the prohibited harm will also occur as a result of his actions. If the harm does occur, the accused has recklessly caused it.

On the other hand, objective fault does not focus on the accused's knowledge or state of mind, but rather the state of mind that a reasonable person would likely have if he/she

engaged in similar conduct in similar circumstance as the accused did. Objective fault is most frequently referred to as negligence. Serious crimes (or felonies) generally require subjective fault. If a legislature creates a serious crime punishable by objective fault, it is normally assumed that the requisite degree of negligence must be quite high – i.e., gross negligence or criminal negligence. Less serious crimes may require either subjective or objective fault depending on the nature of the offense and the words in the statute, creating the offense. In addition, where less serious offenses require objective fault, that objective fault may be of a lower level than criminal negligence. Ordinary negligence – i.e., lack of reasonable care or due diligence – may suffice. Thus in looking at defenses that negate *mens rea*, it is essential to determine whether the offense in question requires subjective or objective mental fault, and then to determine the exact form and nature of the requisite subjective or objective state of mind. The following defenses negate *mens rea*.

Mistake of Fact as a Defense

For subjective *mens rea* offenses, if an accused has made an honest mistake of fact in respect to one of the essential elements of the offense, the accused does not have the required *mens rea* and therefore he is not guilty of that offense. For example, the essential elements of theft are (1) taking (2) property (3) that belongs to someone else (4) without that person's consent. If an accused takes an umbrella from an umbrella rack at work thinking it is his umbrella, he *intends* to take the umbrella, but he does not *intend* to steal it. His mistake of fact in respect to element (3) of the offense (he thinks the umbrella is his) negates the intent to steal. But not all mistakes of fact provide a defense. Suppose an accused has always wanted a black umbrella. When an opportunity arises, he takes an umbrella thinking it is black and knowing it is not his and that he does not have the owner's consent. Suppose he is color blind and by mistake takes a blue umbrella thinking it is black. The accused's mistake of fact as to the color of the umbrella is no defense. It is true that he did not subjectively intend to steal a blue umbrella. However, proof of the color of the stolen property is not an essential element of the offense of theft; therefore a mistake as to its color does not negate the *mens rea* for theft.

While a mistake of fact as to an essential element of an offense is a defense, it can sometimes create a *mens rea* dilemma. Suppose the accused took the umbrella honestly thinking it was his, but his mistake was very unreasonable (i.e., a reasonable person in his position would have immediately recognized it was not his umbrella). Theft requires subjective *mens rea* – the accused must know (or suspect) that it is not his umbrella. In this example, he did not know it was someone else's umbrella, even though his mistake was negligent (highly unreasonable by an objective reasonable person standard). The dilemma then is whether the objective fault level of negligence can be used to satisfy the subjective *mens rea* required for the theft. Generally courts have said no and acquitted the accused of theft in those circumstances, but there have been odd and unprincipled exceptions to that rule. Once again, the better solution is to acquit the accused of the subjective offense of theft and allow the legislature to create a new offense of negligent taking of property if they want to criminalize such behavior. This type of *mens rea* dilemma does not arise in the case of objective fault crimes since a mistake of fact as to an essential element of the offense must be both honest and reasonable before it can negate the objective fault for the offense in question.

Mistake of Law and Color of Right

Since at least Roman times, it has been declared in most legal systems that mistake or ignorance of the law is no defense to committing a crime. This rule made some sense when it was only applied to serious and well-known crimes (as was the case in Roman law). But the rule can be harsh and unjust when applied to thousands and thousands of regulatory or quasi-criminal offenses, which are now part of the modern industrial state and are not generally well-known, nor easy to find. The harshness of the rule that ignorance of the law does not excuse has been avoided in some instances by creating an exception to that rule. For example, the definitions of some property offenses have a defense of mistake of law built into them. The law of theft in many jurisdictions is defined as intentionally taking another's property without consent and 'without color of right.' Color of right is an honest but mistaken belief on the part of the accused that he/she has a legal right to take or possess the property. In effect, the common law has recognized that property rights are complex and that ordinary citizens might understandably be mistaken about them. So an honest mistake of law about property rights is a defense.

Intoxication

Much like insanity, intoxication as a defense has been the subject of intense analysis and debate over the past 200 years. Prior to that time, intoxication was not recognized as a defense and in some instances was seen as an aggravating factor. However as the importance of the mental element in criminal liability increased, so also did the analysis of what should be done if the accused lacks the mental element for that crime due to intoxication.

Intoxication can be involuntary (for example, when someone surreptitiously puts a drug in another's food or drink). But there are very few instances of an involuntarily intoxicated person committing a crime. If that does occur and the drug puts the accused in a state of automatism or unconsciousness, then the accused is not criminally liable because the accused's conduct in committing the *actus reus* of the crime was not voluntary. Alternatively, if the drug does not render the accused unconscious, but it results in the accused not having the *mens rea* for the offense committed, then the accused is entitled to an acquittal. The only difficult scenario is if the involuntary intoxication does not negate voluntariness or the requisite intent, but only loosens the accused's inhibitions so that he/she commits the offense in circumstances where he/she would not have committed it if he/she had not been drugged. Most legal systems provide no defense in that scenario. However, since the accused's ability to control his/her behavior was reduced through no fault of his/her own, justice may suggest that a partial defense, or at least a lower sentence, is warranted (Ferguson, 2012).

While involuntary intoxication is rare, voluntary intoxication is a very frequent condition in the commission of offenses. It does not usually reach the very high level of intoxication that

is necessary to negate voluntary conduct or the requisite mental element. Lesser intoxication is not a defense. However, in the relatively small number of cases when the level of intoxication is sufficiently extreme to render the accused's criminal conduct involuntary or unintentional, different jurisdictions take significantly different approaches to those circumstances.

Exculpatory Defenses That Do Not Negate *Actus Reus* or *Mens Rea*

There are a range of other exculpatory defenses that operate notwithstanding the fact that the *actus reus* and *mens rea*, for a specific offense, have been committed. Some of these defenses are justifications; others are excuses.

Justifications

Lawful Use of Force

There are many occasions where the law authorizes the person to use force where necessary to achieve a socially legitimate purpose. For example, a person is entitled to use force to defend himself or herself or others, or to protect their property from unlawful interference from others, or to enforce the law, such as making an arrest, or in limited circumstances to control or discipline children under one's care. The lawful force must be necessary and reasonable in the circumstances. In addition, reasonableness implies some degree of proportionality.

In the above instances, the intentional use of force meets the *actus reus* and *mens rea* requirements for assault, but if the force is lawfully employed for one of the listed purposes, that use of force is neither wrong nor unlawful in the eyes of the law; indeed it is seen as lawful and proper – justified – conduct.

What is considered reasonable and therefore justified varies from time to time and country to country. If a person uses more force than necessary in any of these justified circumstances, that excessive force is not justified and the person may be found criminally liable for its use. In some countries, excessive force in these circumstances is recognized as a partial defense; in other countries it is not. In countries where excessive force in justified circumstances is a partial defense, should that form of self-defense be classified as an excuse, rather than a justification?

Self-Defense

Self-defense is a common example of the general defense of lawful use of force. Its legitimacy as a defense comes with its own set of challenges and controversies. In particular, is a person who honestly but mistakenly believes he is about to be assaulted entitled to rely upon self-defense (Conde, 2008)? To what extent do notions of imminence of the unlawful assault allow or prevent preemptive use of force in self-defense? To what extent should self-defense be justified if it is not necessary in the sense that the person could safely retreat and thereby avoid using force against the potential assaulter? And finally to what extent should the force being used in self-defense be proportionate to the unlawful force that is threatened or being used? In other words, would a person in a wheelchair, or a weak and feeble person, be justified in shooting another person who was about to unlawfully slap the feeble person in the face in circumstances where the feeble person had no other means to prevent the unlawful slap? The law's response to questions of this sort reveals a fair bit about a country's cultural beliefs on community peace and protection of life (including the life of the unlawful attacker), as opposed to cultural beliefs that citizens have the right to stand their ground at any cost against unlawful aggression.

1. The issue of retreat vs 'stand your ground' has become a key issue in regard to self-defense laws in the United States in recent years. For many years, and in many countries, the law of self-defense did not authorize a person to defend himself by use of deadly force unless he reasonably believed that his life was at risk and that there was no other reasonable and safe option other than the use of deadly force. Thus, if retreat was a viable and safe option, a citizen was expected to retreat. The only exception was that a citizen did not have to retreat from her home, or in the language of the common law, her castle. These self-defense principles are incorporated in the U.S. *Model Penal Code*. But only a minority of states still adhere to those principles (LaFave, 2003). Led by Florida in 2005, and fueled by support from the National Rifle Association, the majority of states have adopted 'stand your ground' self-defense laws. Invoking the right to defend oneself and others, and the right to bear arms to do so, states have abolished the requirement to retreat if possible before resorting to the use of deadly force in self-defense. Critics of these new laws claim that they result in a net decrease in public safety, that they encourage aggressive violence under the guise of self-defense and that such laws are unnecessary in the sense that there is no data indicating that persons were being prosecuted unjustly for defending themselves.

Necessity

The defense of necessity contemplates that in particular circumstances criminal conduct may be necessary to avoid the occurrence of a greater imminent harm, if there is no other reasonable and safe way to avoid this greater harm. The accused is between a rock and a hard place and the law should expect him/her to choose the lesser of the two harms in working his/her way out of that situation. In many countries, necessity has been successfully used as a defense to breaking traffic laws (e.g., speeding) to escape from a murderous attacker or to rush a sick child to the hospital where there are no other reasonable options (e.g., no phone to call emergency services or no emergency services readily available). Likewise, most people would argue that a hopelessly lost hiker, who is weak, cold, and starving to death, should be entitled, if he stumbles upon a cabin, to break into it to obtain food and shelter. While the defense of necessity is considered an excuse in some jurisdictions such as Canada and England, it seems more appropriate to treat it as a justification when the person has no alternative but to choose between the occurrence of two harms and in that situation chooses the lesser harm. If the accused chooses the lesser harm, it seems to be the right thing to do and in that sense his choice of conduct should be seen as justified, as it is in the majority of countries (LaFave, 2003, vol. 2; Ghanayim, 2007).

2. Whether necessity should be an available defense when the harm caused by breaking the law is roughly equivalent to the harm avoided is a much more controversial issue. When the two harms at stake involve human life, great moral debates are unavoidable. For example, is a doctor justified

or excused if he/she separates conjoined twins when one twin will surely die, but without the separation both twins will die within a relatively short time because the other twin's heart is not strong enough to support both of them? Is a firefighter who is trying to put out a fire in a nursing home justified in closing a fire door between two sections of the building, if the fire door is the only escape route for a patient who is still in that section of the building and will be unable to open the door on his/her own? There is no limit to the number of compelling examples one can imagine of situations where a choice may have to be made between the death of one person in order to save one or more other persons. Thankfully these examples rarely arise in actual practice.

Excuses

Apart from justifications, there are other exculpatory defenses that do not make the accused's acts right or lawful, but do nonetheless provide a socially acceptable reason to excuse the actor from conviction and punishment for his/her otherwise criminal conduct. These excuses include (1) criminal exemptions for the insane and children under a specified age, (2) provocation (and in some jurisdictions a broader defense of extreme emotional disturbance), (3) duress, and (4) mistakes of law based on officially induced error.

Insanity

The insanity (or mental disorder) defense has been the subject of voluminous discussion and concern for the past 200 years. In some legal systems, insanity is treated as a negation of the requisite *mens rea*. The accused's intent to kill a person whom he falsely believes is a devil or an assassin is not treated in law as a sufficient *mens rea* because it is not formulated in a rational, reality-based mind. On the other hand, many legal systems treat insanity as an excuse, separate and apart from whether the insane accused had *mens rea*. In these countries, the accused's insanely formulated intent is treated as a legally adequate *mens rea* but the accused is nonetheless excused from criminal liability. That excuse is based on the finding that due to insanity he/she was unable to understand the nature of his/her conduct, or that it was wrong, or in some jurisdictions, that he/she was unable to control his/her conduct. (Whether insanity is treated as a negation of *mens rea* or as an excuse, that fact does not significantly alter its elements.)

The various standards used in different countries from time to time for meeting the insanity test have always been quite high. This means that the insanity defense seldom succeeds. Many accused who are mentally ill are convicted and sentenced to imprisonment because their significant mental impairment has not risen to the required heights to satisfy the insanity defense.

If insanity is proven (and many legal systems require the accused to prove insanity on a balance of probabilities), the judge or jury must return a verdict of 'not guilty due to insanity' or 'not criminally responsible due to mental disorder' or some similarly worded verdict. But the verdict of insanity is not an outright acquittal as other verdicts are. The insanity acquittee is often subject to further state scrutiny, usually by way of a subsequent judicial hearing to determine whether the person is a continuing danger and therefore needs to be committed to a psychiatric institution or needs to be subject to other forms of control in the community.

Duress

A classic example of duress is a case where a robber holds a gun to the head of a driver and says "drive me to that bank which I am going to rob, or I will shoot you." If the driver does so, the driver has committed the *actus reus* and *mens rea* of robbery by knowingly assisting its commission. But he should be excused for his behavior because it was not the product of a meaningful choice. In most countries, the defense is invoked by an accused who claims he was threatened (or a member of his family was threatened) with immediate death or great bodily injury if he refused to participate in a crime; he had no lawful alternatives available, he reasonably believed the threat would be carried out, and he was not responsible for placing himself in a position where such a threat was likely.

Duress can be classified as a justification or an excuse depending upon the manner in which it is defined in each country. In France, it is properly considered a justification because proportionality is an additional requisite element. In other words duress is only available as a defense in France if the harm that is threatened if the accused does not comply is greater than the harm that the accused causes by doing what he has been ordered to do. On the other hand, in the United States duress is properly classified as an excuse because there is no strict requirement of proportionality. However lack of proportionality is nonetheless relevant as a factor that can be considered in deciding whether another legal requirement of duress in the United States has been met, namely, was the threat grave enough that a person of reasonable firmness would not have resisted it? In South Africa, there is no requirement of proportionality, only a requirement of reasonableness, but South Africa does not divide its exculpatory defenses into justifications and excuses (van der Merwe and du Plessis, 2004, Chapter 13). Conversely, many common law countries impose a proportionality requirement on the defense of duress yet inconsistently insist on referring to it as an excuse (Stuart, 2011). In these countries the defense of duress is very similar in character to the defense of necessity.

Provocation (Extreme Emotional Disturbance)

Provocation is a qualified or limited defense in two different respects. First, it is only a defense to one offense – murder. Provocation in respect to other offenses is no defense but it is a mitigating factor in sentencing (Stuart, 2011; Baker and Williams, 2012). Second, it does not result in a complete acquittal. Instead it reduces murder to manslaughter (or culpable homicide not amounting to murder) in common law countries and it reduces murder to 'mitigated murder' in continental countries. In both systems, the reduced offense provides for a lesser punishment than the punishment for murder. The rationale for provocation as a partial mitigating excuse is the belief that there are certain provocative incidents that are so strong that it is understandable (though not entirely excusable) that an ordinary person in those circumstances could temporarily lose his/her normal powers of self-control and lash out at the provoker with deadly force. Obviously the types of incidents that the law recognizes as sufficient to temporarily impair a person's power of self-control are premised on a society's dominant cultural beliefs at any given time. In England, provocation as

a defense was originally recognized if the killing occurred as a result of a sudden quarrel or if a husband caught his wife in an act of adultery. Slowly the circumstances where provocation was recognized expanded to any acts or words that might be sufficient to cause an ordinary person in circumstance similar to those of the accused to temporarily lose the normal degree of self-control expected of all of us.

The defense of provocation has received sustained criticism and has been the subject of several law reform efforts in the past 20 years (Ormerod, 2011; Bronitt and McSherry, 2010; Simester and Brookbanks, 2007). The major criticism of provocation is its disproportionate use in cases of males killing their wives, partners, or ex-partners in the context of jealousy, domestic discord, and abuse, or males killing other males based on a homophobic claim that the deceased provoked the attack by making an unwanted sexual advance on the accused (Warrick, 2011; Hemming, 2010). For example, in one Canadian study of 114 provocation cases in the 1990s, the accused were males in 107 of those cases and females in only seven cases; in just over one-half of those cases, the male accusers killed their intimate female partners or ex-partners. The claim of provocation succeeded in about 25% of these types of cases. The fact that provocation is raised in these types of cases leads to the criticism that the provocation defense is condoning male violence against women, at a time when it should be denouncing it in an uncompromising way.

The Australian State of Victoria significantly altered its provocation defense in 2005 to respond to criticisms about its use by males in family violence homicides and the relative ineffectiveness of the defenses of self-defense and provocation for women who kill their male partners as a defensive response after years of spousal abuse. The new legislation abolished the defense of provocation and, among other matters, enacted a new partial defense of 'excessive self-defense' that, if successful, results in a conviction of the lesser offense of defensive homicide. The new defense was designed to assist women who kill as a form of defensive response to (male) domestic violence in circumstances where the strictures of self-defense laws do not provide a defense. There is some controversy as to whether these reforms have been successful (Hemming, 2010).

New Zealand abolished the provocation defense in 2008. Provocation in New Zealand is now simply one of the mitigating factors a judge can take into account in sentencing the accused for murder. This reform option is viable in New Zealand because, unlike most other common law countries, New Zealand does not have a mandatory minimum sentence of life imprisonment for murder. In countries where murder is only punishable by life imprisonment, it is not possible to reduce that sentence regardless of the number or type of mitigating factors that exist (Simester and Brookbanks, 2007). In 2009, England abolished the defense of provocation and replaced it with a new defense called loss of control. Besides a change in name, the new law is an attempt to deal with some of the criticisms of England's former provocation defense (Ormerod, 2011). There is considerable debate as to whether the new law will result in much practical change. No doubt, in the years to come, other countries will continue to explore abolition or amendment of their provocation defense.

Conclusion

The above discussion of criminal liability and defenses demonstrates that the laws' definition and requisite elements are intended to reflect the circumstances in which a given society considers it fair and proper to assign blame and impose punishment for breach of its criminal laws. As society's views and values change, the legal requirements for a crime and a defense can and should change to match those new views and values. We have seen this process at work in the past 20–30 years in a number of areas including the defense of provocation, self-defense, necessity, and duress. In addition to changes in these defenses, there are other challenges to our articulation of defenses that are still developing. For example, to what extent should the law recognize an accused's particular cultural beliefs and values as relevant to self-defense, provocation, and necessity when the accused's beliefs are far different, or directly in conflict, with the majority's beliefs? And to what extent if any does the law's notion of voluntary conduct have to change as science continues to document how the ability to understand and control one's behavior are affected when an accused suffers from conditions such as fetal alcohol syndrome, attention-deficit disorder, and post-traumatic stress disorder? And as neuroscience continues its rapid progress in mapping the brain and measuring its activity, the law may need to abandon or alter some of its current assumptions about the nature of voluntary conduct, which underlies various defenses. These advances, and others not yet commonly understood, will continue to challenge society and its lawmakers in the creation of a fair and proper system of defenses for the imposition of criminal liability and punishment.

See also: Applied Criminology and Criminal Justice; Law And society; Law, Autopoiesis In; Law: The Socio-Legal Perspective; Punishment: Comparative Politics of; Punishment: Social and Legal Aspects; Punishment.

Bibliography

Ashworth, A., 2009. Principles of Criminal Law, sixth ed. Oxford University Press, New York.

Badar, M., Marchuk, I., 2013. A comparative study of the principles governing criminal responsibility in the major legal systems of the world (England, United States, Germany, France, Denmark, Russia, China, and Islamic legal traditions). Criminal Law Forum 24, 1–48.

Bahnassi, A.F., 1982. Criminal responsibility in Islamic law. In: Bassiouni, M.C. (Ed.), The Islamic Criminal Justice System. Oceana Publications, New York, NY, pp. 171–193.

Baker, D., Williams, G., 2012. Textbook of Criminal Law, third ed. Sweet & Maxwell, London.

Bargh, J., 2008. Free will is un-natural. In: Baer, J., Kaufman, J., Baumeister, R. (Eds.), Are We Free? Psychology and Free Will. Oxford University Press, New York, pp. 128–154.

Bohlander, M., 2012. Principles of German Criminal Procedure. Hart Publishing, Portland, OR.

Bohlander, M. (Trans.), 2008. The German Criminal Code: A Modern English Translation. Hart Publishing, Portland, OR.

Bohlander, M., 2009. Principles of German Law. Hart Publishing, Portland OR.

Bronitt, S., McSherry, B., 2010. Principles of Criminal Law, third ed. Thomson Reuters, Sidney, Australia.

Butler, W., 2003. Russian Law, second ed. Oxford University Press, New York.

Cairns, W., McKeon, R., 1995. Introduction to French Law. Cavendish Publishing, London.
Conde, F.M., 2008. Putative self-defense: a borderline case between justification and excuse. New Criminal Law Review 11, 590–614.
Elliott, C., 2000a. A comparative analysis of defences in English and French criminal law. European Journal of Crime, Criminal Law and Criminal Justice 8, 319–326.
Elliott, C., 2000b. The French law of intent and its influence on the development of international criminal law. Criminal Law Forum 11, 35–46.
Ferguson, G., 1998. Legal regulation of mental disorder: looking East and West. In: Johnston, D.M., Ferguson, G. (Eds.), Asia-Pacific Legal Development. UBC Press, Vancouver, BC, pp. 269–352.
Ferguson, G., 2011. Insanity. In: Chan, W.-C., Wright, B., Yeo, S. (Eds.), Codification, Macaulay and the Indian Penal Code. Ashgate, Farnham, UK, pp. 231–255.
Ferguson, G., 2012. The intoxication defence: constitutionally impaired and in need of rehabilitation. Supreme Court Law Review 52, 111–148.
Fletcher, G., 1978. Rethinking Criminal Law. Little, Brown & Company, Boston, MA.
Friedland, S., 1998. The criminal law implications of the human genome project: reimagining a genetically oriented criminal justice system. Kentucky Law Journal 86, 303–366.
Ghanayim, K., 2007. Excused necessity: a defence in the criminal code – a comparative doctrinal study. Canadian Criminal Law Review 11, 53–96.
Grant, J., 2006. Determinism, neuroscience and responsibility. International Journal of Law in Context 2, 221–231.
Hemming, A., 2010. Provocation: a totally flawed defence that has no place in Australian criminal law irrespective of sentencing regime. University of Western Sydney Law Review 14, 1–44.
Johnson, M., 1998. Genetic technology and its impact on culpability for criminal actions. Cleveland State Law Review 46, 443–470.
Jones, M., 2003. Overcoming the myth of free will in criminal law: the true impact of the genetic revolution. Duke Law Journal 52, 1031–1053.
Kairns, W., McKeon, R., 1995. Introduction to French Law. Cavendish Publishing Ltd, London, UK.
Kritzer, H.M. (Ed.), 2002. Legal Systems of the World: A Political, Social and Cultural Encyclopedia, 1–4 vols. ABC-CLIO, Santa Barbara, CA.
LaFave, W., 2003 (with pocket supplement to 2013–2014). Substantive Criminal Law, second ed., 3 vols. Thomson West, Eagan, MN.
Lindsey, T., 2008. Indonesia: Law and Society, second ed. Federation Press, Leichhardt, NSW, Australia.
Menski, W., 2006. Comparative Law in Global Context: The Legal Systems of Asia and Africa. Cambridge University Press, Cambridge, UK.
Merryman, J., Pérez-Perdomo, R., 2007. The Civil Law Tradition: An Introduction to the Legal Systems of Europe and Latin America, third ed. Stanford University Press, Stanford, USA.
Montague, R., 2008. Free will. Current Biology 18, R584–R585.
Morse, S., 2004. New neuroscience, old problems. In: Garland, B. (Ed.), Neuroscience and the Law: Brain, Mind and the Scales of Justice. Dana Press, New York, pp. 157–198.
Ormerod, D., 2011. Smith and Hogan's Criminal Law, thirteenth ed. Oxford University Press, New York.
Packer, H., 1968. The Limits of Criminal Sanction. Stanford University Press, Stanford, CA.
Petchsiri, A., 1987. Eastern Importation of Western Criminal Law: Thailand as a Case Study. Rothman, Littleton, CO.
Rayar L., Wadsworth, S. et al. (Trans.), 1997. The Dutch Penal Code. Fred B Rothman, Littleton, CO.
Robinson, H., 1984. Criminal Law Defences, 2 vols. West Publishing, St. Paul, MN.
Röhl, W. (Ed.), 2005. History of Law in Japan since 1868. Brill, Leiden, Netherlands.
Simester, A., Brookbanks, W., 2007. Principles of Criminal Law, third ed. Thomson Brookers, Wellington, NZ.
Stuart, D., 2011. Canadian Criminal Law, sixth ed. Thomson Reuters Canada Ltd, Toronto, ON.
Tomlinson, E. (Trans.), 1999. The French Penal Code. Fred B Rothman, Littleton, CO.
van der Merwe, C., du Plessis, J. (Eds.), 2004. Introduction to the Law of South Africa. Kluwer Law International, The Hague.
von Hirsch, A., Ashworth, A., 2005. Proportionate Sentencing: Exploring the Principles. Oxford University Press, Oxford, UK.
Warrick, C., 2011. Not in our right minds: the implications of reason and passion in the law. Politics & Gender 7 (2), 166–192.
Yeo, S., 2008. The insanity defence in the criminal laws of the commonwealth of Nations. Singapore Journal of Legal Studies, 241–263.
Zhu, J., Han, D. (Eds.), 2013. (X. Zhang, Trans.). Research Report on the Socialist Legal System with Chinese Characteristics, vol. 2. Enrich Professional Publishing, Singapore.

Criminal Procedure: Adversarial and Inquisitorial Legal Systems

Stephen C Thaman, Saint Louis University School of Law, St. Louis, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explains the historical development of the two main types of court procedures used in criminal, and to a lesser extent in civil, cases around the world. The adversarial system developed from the English jury trial and envisions a passive judge and an active role for the parties. The inquisitorial system sees the judge as active investigator of the truth and mistrusts lay participation. A third tradition, that of consensual resolution of cases and plea bargaining, grew not only out of the adversarial tradition but also, through its inducement of pleas of guilty, smacks of the inquisitorial tradition.

Introduction

There has always been a multiplicity of penal court procedures. The differences in how criminal wrongdoings were unraveled and settled often depended on the mode of their commission, the relationship between victim and perpetrator, and the seriousness of the offense. Three models for the resolution of conflicts arising from criminal wrongdoings have their roots in ancient procedures that have perhaps been common to all cultures. All were attempts to avoid the most primordial response to criminal wrongdoing: self-help and, in the case of homicide, blood revenge. All are relevant in understanding modern criminal procedure. These are (1) consensual semi-private resolution of the conflict through negotiations between victim or prosecutor and the accused, whereby 'consensual' refers to court procedures where the parties agree to a resolution without requiring a full trial; (2) adversarial resolution of the dispute in a public oral trial often held before a panel of lay judges (jury), whereby 'adversarial' means that evidence is produced by the parties (prosecution, defense) and the judge plays a neutral, more passive role; (3) inquisitorial investigation and decision of criminal cases, conducted in its heyday by state officials obligated to ascertain the truth (i.e., judges play a more active role in investigating the issues and ascertaining the truth and the parties play a more passive role).

All these procedural modes are found in varying mixtures in most modern penal procedures, albeit in different combinations, usually with one procedural mode dominating for cultural and historical reasons. They, and the principles derivative therefrom, inform contemporary discussions about criminal procedure reform in Asia, Latin America, the former Soviet Union, and other parts of the globe.

Criminal procedure consists in a sequence of acts or procedures conducted by officials aimed at determining whether crimes were committed, who committed them, and what their punishment, if any, should be. Criminal procedure reformers must assess the applicability of the three procedural modes to (1) the decision to arrest and commence a preliminary investigation (i.e., the procedure of investigating a crime, determining the identity of the perpetrator/s, and the legal qualification of the criminal act, before the case goes to a trial court); (2) the decision to detain the suspect pretrial; (3) the decision to charge a suspect with the commission of a crime; (4) the pretrial assessment of the sufficiency of evidence to allow the case to proceed to trial (preliminary hearing); (5) the decision as to guilt (the trial); and (6) decisions post-trial as to whether there was sufficient evidence to convict, whether errors were made that require a new trial, or whether a person acquitted may be retried.

Some Historical Watersheds in Criminal Procedure Reform

Early modes of consensual and adversarial resolution of criminal disputes predominated in Europe up until the late middle ages and the Renaissance, and exist in some communities in Africa, Asia, Latin America, and other parts of the world that still use customary law ('customary law' refers to procedures that were used before inquisitorial and adversarial procedures developed, such as ordeals, duels, and use of oaths). In both systems, the victim (or his or her tribe) made an accusation and the suspect-defendant responded. These early procedural systems were usually accusatorial in nature, that is, initiated by the victim as accuser. A variety of influences, both political and cultural, led to the development of a new system of criminal procedure on the European Continent, what is now called the inquisitorial procedure. Politically, the rise of the nation-state and absolute monarchy led to a politicization of criminal procedure, with the state replacing the victim as initiator of criminal proceedings. The vertical hierarchy of such systems was reflected in a vertical hierarchy of courts and a huge royal judiciary to administer it.

Culturally, the law was heavily influenced by the Roman Catholic canon law's inquisitorial procedures and formal rules of evidence and by the rediscovery of Roman law in the Italian universities. Judges began to see themselves (or to be seen) as truth seekers who, armed with the newly articulated Roman legal principles emerging from the Italian universities and with the rules of formal proof emanating from the Canon law, could achieve a quality of justice superior to that achieved by the irrational forms of procedure used by the Germanic tribes: ordeals, oath helpers, duels, and lay decision makers.

The judges were royal officials beholden to the monarch, with no fealty to the communities in which their judgments would resonate. Punishments were considerably more Draconian than in the early middle ages. Procedurally, the most radical new innovation was that of a formal preliminary investigation initiated *ex officio* and conducted by a judicial official who would examine the suspect and the witnesses,

conduct other investigative acts, and reduce all of the findings to writing. Because no accusation by the victim was required, the new procedure was called inquisitorial. The written results of this preliminary investigation were compiled in an investigative dossier, a file compiled by the investigative official during the preliminary investigation that contains all the evidence to be presented at trial. The investigative dossier became the exclusive record upon which the decision of guilt or innocence would be based. If, under the formal rules of evidence, enough indicia existed to constitute what one might today call probable cause or reasonable suspicion to believe that the suspect was guilty, then the investigating magistrate (the judge who carried out the preliminary investigation in early inquisitorial systems, and still does so in some countries such as France or Spain) could order the suspect to be tortured if he or she refused to be examined. There was no oral, public trial before a jury. Instead, the proceedings consisted of a secret written review of the findings of the investigating magistrate contained in the written investigative dossier. There was no right to counsel. There was no confrontation of defense and public prosecutor (or victim) as one knows it today, where the 'right to confrontation' denotes the right of the defendant to be present when the witnesses of the prosecution testify, and to question them.

Judges had no discretion to mitigate the harshness of the findings. Unlike the jury, which continued to function in England, inquisitorial judges could not reject the result mandated by the rules of evidence or acquit against the strictures of the law. The lay judges and litigious procedures that dominated in the earlier procedures were gradually eliminated in favor of a purely 'professional' and more administrative form of justice.

Yet the oral tradition persevered in England and to an extent in Scandinavia. Oral, public trials by a jury of 12 gradually replaced ordeals, duels, and the use of compurgators or oath helpers, friends of the accused or victim who would swear to their honesty and uprightness, by the thirteenth century. Verdicts had to be unanimous and the ensuing judgments were final, not allowing for any posttrial review. Other than a brief examining of witnesses and suspects for purposes of determining release on bail, there was no preliminary investigation. Victims or their lawyers brought the case and the accused defended himself, usually without a lawyer, until the eighteenth century.

With the advent of the French Revolution, these two systems confronted each other and led to a new mixed form of procedure on the European Continent. The French revolutionaries were inspired by Enlightenment criticism of the brutality of inquisitorial criminal procedure (invoking great thinkers like Beccaria, Montesequieu, and Voltaire) and were influenced by the antiauthoritarian credentials of the English jury trial (provided by the notorious acquittals of William Penn and others dissidents against the instructions of royal judges). As a result, they introduced to France the public, oral English jury system, eventually grafting it on to the secret, written, inquisitorial preliminary investigation.

The French Declaration of the Rights of Man in 1789 and the American Bill of Rights of 1791 were products of the first great human rights revolution that impacted criminal procedure, leading to its gradual humanization by protecting the criminal suspect from the arbitrary and pitiless power of the state. The American Bill of Rights proclaimed the protection against cruel and unusual punishments and unreasonable searches and seizures, the privilege against self-incrimination, and a host of trial rights, including the right to counsel, to subpoena witnesses for the defense, to cross-examine or confront the witnesses of the prosecution, and the right to a speedy, public trial by a jury of the defendant's peers. All these rights were meant to protect against overreaching by the state, either by securing important rights of suspects during the gathering of evidence, or by assuring that the defendant would not be tried on the basis of written evidence prepared by the state in secret, without input or checking by the defendant through cross-examination. The presumption of innocence was proclaimed by the French Declaration and was held to be implicit in the due process rights granted by the American Bill of Rights.

Although jury verdicts in France were final, like in England, French juries were asked to answer a list of specific questions related to the elements of the charged crime(s) and the guilt of the defendant(s) contained in a special verdict, which sometimes contained dozens – if not hundreds – of questions. The inquisitorial, written tradition on the European Continent preferred reasoned judgments; a particularized special verdict was intended to make the jury's reasoning process clear and better facilitate appellate review of their verdicts. Often, the professional judges would evaluate the substance of the jury's factual answers and pronounce judgment, thus limiting the autonomy of the jury's finding of guilt. The French jury model, in which a jury of twelve was presided over by three professional judges, and in which only a majority verdict was required, was adopted by nearly all European countries in the nineteenth century. In Germany, a mixed court, composed of one professional judge and two assessors, in which lay and professional judges collegially decide questions of fact, law, guilt, and sentence, developed in the mid-nineteenth century for the trial of lesser offenses (the term 'mixed court' in general refers to courts made up of one or more professional judges and a group of lay judges who deliberate together and together decide all issues of fact, law, guilt, and sentence).

The continental European mixed system and the Anglo-American system of adversarial jury trial were spread through colonization, imperialism, and systematic borrowings into Asia, Africa, and Latin America. Ironically, many of the Latin American countries that liberated themselves from Spanish and (in the case of Brazil) Portuguese rule in the early nineteenth century kept the purely written inquisitorial procedures conducted exclusively by professional judges and rejected the reforms that swept the European continent. Only Brazil, Panama, Nicaragua, and El Salvador introduced trial by jury, but these nations largely limited the jury's role to reading the contents of the inquisitorially prepared investigative dossier.

Jury courts were always controversial on the European continent and many jurists sought to abolish them or to convert them into mixed courts. This was finally accomplished by the totalitarian regimes of the first half of the twentieth century. Juries vanished in many countries, replaced by mixed courts or purely professional panels, in 1917 in Russia, in 1931 in Italy, in 1939 in Spain, and in 1941 in Vichy France. In 1943 in Japan, jury trials, which had been introduced in 1928, were suspended.

Jury trial was eliminated in Germany in favor of the mixed court before the rise of Nazism to power, by decree in 1924.

After the defeat of Nazism and Fascism in World War II, Italy, France, and Germany did not return to the classic jury, preferring the mixed court. The mixed court cemented the dominance of the professional bench over its lay component by enabling the judge to personally influence the lay judges in how they decided the question of guilt. It also was more consistent with the necessity of giving reasoned judgments in all criminal cases.

The horrors of World War II and the Holocaust and the crimes of the totalitarian regimes led to a new era of human rights in the latter half of the twentieth century, with the adoption of the Universal Declaration of Human Rights, the International Covenant for Civil and Political Rights, and the European, American, and African Conventions on Human Rights. The European Court of Human Rights has greatly influenced the reform of European criminal procedure through its interpretations of the European Convention. In the wake of this revolution, and with the postwar democratization of Germany, Italy, France, and Japan and of Spain after Franco's death, many European countries began to systematically eliminate many of the negative vestiges of the old inquisitorial systems that had survived since the French Revolution.

A distinct shift toward adversary and consensual procedures has been noticeable since the late 1980s. The 1988 Italian Code of Criminal Procedure introduced an adversarial trial system and several forms of consensual resolution of the guilt issues; it did not, however, adopt trial by jury, because the Italian Constitution requires that judges give reasons for criminal judgments, and the authors of the Constitution felt that juries were unable to give such reasons. Russia, on the other hand, reintroduced adversary procedure and the jury in 1993 and a form of consensual resolution of cases in 2001. The majority of the former Soviet Republics have followed the Russian model, although only Georgia has introduced trial by jury. Azerbaijan and Kyrgyzstan have, however, passed laws aimed at its eventual introduction.

Spain strengthened the adversarial nature of its trial system by introducing trial by jury in 1995. It then took a daring step by requiring its juries, consisting of nine citizens needing seven votes to convict, to give concise reasons for their answers to the questions contained in the special verdicts. This innovation proved to be critical for the whole continent; the European Court of Human Rights, in the case of *Taxquet v. Belgium* (2009), overturned the conviction of a man for murder of a Belgian politician because the jury's verdict did not give any indication of why they chose to convict. The Court made clear that the decision was not calling into question the Anglo-Saxon jury system as such, as long as clear judicial instructions to the jury made verdicts understandable to the public and to the defendant. However, the case did induce Belgium to require its juries in the future to provide reasons for their verdicts.

The Adversarialization of the Inquisitorial Preliminary Investigation

The centerpiece of inquisitorial procedure was always the preliminary investigation, traditionally carried on by an investigating magistrate who was usually a member of the judiciary. Even after the inquisitorial systems reformed to include an oral trial by jury, most systems allowed the trial court to read the reports, including those of forensic experts, contained in the investigative dossier. This constituted a prepackaging of evidence by law enforcement agencies with no opportunity for the defense to see or confront the witnesses. Some countries allowed a minimal amount of participation by counsel. For instance, since 1897, France has allowed counsel to participate during judicial interrogations of the suspect, but this was of little help to the suspect because his or her interrogation by the police was always conducted in secret, with no right to counsel.

Substantial changes in the inquisitorial preliminary investigation only occurred after the second human rights revolution following World War II. The role of a judicial official as investigator began to be questioned, for how can a judge, who is following his or her particular theory of guilt in a case, be independent, neutral, and judicial when issuing arrest, search, or wiretapping warrants? Since it was the police who, in reality, did the lion's share of actual criminal investigation, and since it was the prosecutor who was responsible in most countries for preferring the charges, it was decided to put the prosecutor in charge of the preliminary investigation and to reduce the judicial role to that of a liberty or control judge, a neutral arbiter of invasions of human rights (authorizations of arrests, seizures, searches, and wiretaps) who could also preside over interrogations. This step was taken in Germany in 1974, in Italy in 1988, in Venezuela and other Latin American countries in the 1990s. The prosecutor still maintained the inquisitorial control of the investigation and determined which witnesses and experts would testify in the case. In Spain, new legislation, such as the 1995 jury law, has transformed the investigating magistrate into a more neutral pretrial judge by allowing him or her to investigate only in response to a request by either the public or private prosecutor (victim) or the defense.

In the Soviet Union and formerly socialist Eastern Europe, it was the public prosecutor who not only generally directed the preliminary investigation (which was in the hands of an investigator provided by the Ministry of the Interior) but also authorized all invasions of protected human rights of suspects. Former socialist countries that have since ratified the European Convention of Human Rights have gradually introduced new legislation requiring judges to perform these important functions, and the ex-Soviet republics of Central Asia have by and large followed suit.

Protection of suspects from secret, custodial, counselless interrogations was greatly influenced worldwide by the famous decision of the US Supreme Court in *Miranda v. Arizona*, 384 U.S. 436 (1966). Today, most of the formerly inquisitorial systems of continental Europe, the former Soviet Union, Latin America, and South Korea require law enforcement officials to advise suspects of their right to remain silent and their right to counsel before being questioned, and many new codes of criminal procedure, such as those of Italy (1988), Venezuela (1998), and even the Kyrgyz Republic (1999) require counsel to be present during all interrogations. (This is a requirement that extends beyond what US law requires.) This trend has also been substantially strengthened by a decision of the European Court of Human Rights, *Salduz v. Turkey* (2009), which has

recognized a firm right to counsel during the preliminary investigation and especially during police interrogations.

Increased recognition of the right to confront witnesses, guaranteed by Art. 6(3)d of the European Convention of Human Rights, has led formerly inquisitorial countries to require, where possible, the presence of the defendant and/or defense counsel during the examination of prosecution witnesses during the preliminary investigation. The European Court of Human Rights has repeatedly condemned the use of written statements where they are the sole or main evidence of guilt in criminal cases, if the defendant had no chance pretrial to confront and examine the witness. The 1988 Italian Code requires the prosecutor to initiate a hearing to preserve witness testimony in cases where there is fear the witness may not be available for trial. At this hearing, the defendant and victim have the right to be present and to examine the witness; failure to honor this right means the statement cannot be admitted at trial. Similar provisions exist in the 1995 Spanish Jury Law and in new Latin American codes of criminal procedure. Thus, procedure in formerly inquisitorial countries is approximating that in the United States, where the inadmissibility of pretrial, out of court, unconfronted statements was strengthened following the decision of the US Supreme Court in *Crawford v. Washington*, 541 U.S. 36 (2004).

In Italy and in cases subject to Spain's new jury law, the preliminary investigation is not supposed to be the stage for preparing evidence to be used at the trial, but only for determining whether sufficient probable cause exists to charge the suspect. Any evidence that could be presented at trial and is not required for a showing of probable cause need not be taken during the preliminary investigation. In this sense, the preliminary investigation tends to resemble the American grand jury or preliminary hearing. In fact, the investigative dossier is, in principle, not to be used during the trial as a source of evidence. A special trial dossier is compiled, including the accusatory pleading and any evidence that has been properly preserved for trial, guaranteeing defense rights of confrontation as noted above.

Some formerly inquisitorial countries have gone a step further, however, by allowing parallel defense investigations as exist in the United States. The 1988 Italian Code provides for defense gathering of evidence in preparation for the trial phase, while amendments in 1999 provide for detailed procedures to regulate the gathering of this evidence and its eventual unification with the prosecution evidence in a common investigative file. The Russian code of 2001 also allows defense investigations.

The Decline of the Inquisitorial Trial Judge

The trial judge in the nineteenth-century mixed systems on the European continent acted like the quintessential investigator. After reviewing the investigative dossier, the judge decided which witnesses would be called, and examined the witnesses at trial. Since the judge had reviewed the investigative dossier and had it at hand, he or she was aware of the inculpatory premises put forward by the investigating magistrate and usually adopted them, for he or she had to decide pretrial whether there was sufficient evidence to set the case for trial. It stretches the imagination to believe that such a trial judge actually entertained a presumption of the innocence of the defendant in such a procedure. Predictably, the written judgment would often closely follow the language of the written accusatory pleading.

On top of that, the defendant was called to answer the charges at the beginning of the trial in front of the jury and/or judge and asked to give a statement before any other witnesses were called or evidence presented, seemingly belying the fact that the burden was on the state to prove the charges. The judge then examined the defendant, using the materials in the dossier to guide him or her in the 'search for the truth'. The prosecutor and defense counsel would be able to submit questions (often only in writing to the presiding judge) to supplement the examination of the judge. If witnesses did not appear, the presiding judge would merely read the statements they had made to the investigating magistrate during the preliminary investigation.

As long as European systems had juries, the trial judge was not a judge of the facts, so the judge's lack of neutrality did not necessarily directly affect the outcome of the case. But when most European systems and Japan eliminated the jury in favor of professional or mixed panels, the same judge who 'investigated' the case at trial using his or her inquisitorial skills (and who may have adopted the findings of the investigating magistrate) also decided the case. This practice appears to violate the presumption of innocence.

This system, which still largely exists in France, Germany, the Netherlands, Belgium, and other countries, is gradually being replaced by the adversarial model, where prosecution and defense are responsible for preparing and presenting the evidence and witnesses, including expert witnesses. In this model, the judge assumes a more passive role, deciding questions of admissibility of evidence and making sure the parties have 'equality of arms' in presenting their cases.

In 1998, Italy was the first to eliminate the preliminary investigation dossier from the courtroom. Spain followed suit in its 1995 jury law, as did Venezuela in 1998. Italy also eliminated the presiding judge's duty to determine the truth, which was important because, unlike in Spain and Russia, Italy did not return to the jury trial, which means the judge remains a trier of fact. Already in 1992, the Russian Constitution was amended to provide for adversarial procedure and trial by jury. Jury trials began in Russia in 1993; since the 2001 code of criminal procedure was enacted, they now exist throughout the country.

Since 1990, most Spanish-speaking Latin American countries have moved away from the old pre-Napoleonic inquisitorial systems to an accusatorial-adversarial system. Latin American countries, however, have by and large kept their professional judges, although the Argentinean province of Cordoba, Venezuela, and Bolivia have introduced mixed courts and the Province of Buenos Aires, in 2013, passed legislation to introduce jury trial. El Salvador and Nicaragua have modernized their old inquisitorial jury systems within new adversarial codes of criminal procedure. Japan and South Korea also introduced new mixed court systems in 2009. The Soviet mixed court, which replaced the Russian jury court in 1917, was adopted in nearly all countries in the socialist bloc following World War II; many of these countries, whether or not they

have become democracies, continue to use a court composed of one professional judge and two lay assessors in criminal cases (Poland, Czech Republic, Hungary, China, Vietnam, among others).

The 'search for truth' at all costs during the criminal trial has also come under attack, following growing recognition that evidence gathered in violation of the human rights of criminal suspects (usually violations of the right of privacy, the right to human dignity, or the privilege against self-incrimination) should not be used in the criminal trial even if it is otherwise relevant and credible evidence of guilt. The United States took this step in 1961 with the landmark decision of *Mapp v. Ohio*, 367 U.S. 643 (1961), which recognized an exclusionary rule in relation to illegally gathered evidence (an 'exclusionary rule' prevents the use of evidence, even if relevant and reliable, which was seized in violation of important constitutional rights of the defendant). In most Western countries, however, the search for truth still prevails because the courts take a very cautious approach to exclusion of evidence, especially when it is very credible physical evidence. They engage in an elaborate balancing process that in the end usually leads to exclusion of evidence only when the violations have been egregious, such as the case of torture (Germany, England and Wales, Canada, Australia, and New Zealand). In Italy and Spain, legislatures have passed laws that require exclusion of such evidence, yet only the Spanish courts have enforced this law with any vigor. The Russian Constitution of 1993 mandates exclusion of illegally gathered evidence, as do most of the new post-Soviet constitutions and many new Latin American constitutions.

The Eclipse of Adversarial and Inquisitorial Procedure?

Although adversarial principles appear to have triumphed over their inquisitorial counterparts in modern criminal procedure reforms, criminal procedure today is increasingly marked by avoidance of the full trial, through plea bargaining or other consensual procedures. Plea bargaining is a type of consensual procedure in criminal cases where the defendant agrees to plead guilty to a charge and punishment that are arrived at through negotiations between prosecution and defense, and where not trial takes place. In the United States, over 95% of all cases are decided by guilty pleas, which are often prompted by the threat of Draconian sentences if the defendant goes to trial and loses. Since Italy enacted its code of criminal procedure of 1988, introducing several types of consensual avoidance of a full trial that were similar to plea bargaining, nearly all European countries have followed suit, and today guilty pleas or stipulations are rapidly replacing trial as the dominant mode of resolving criminal cases. Although the agreement between the parties smacks of adversarial procedure, the pressure placed on the defendant to admit guilt reminds one of old inquisitorial practices of inducing confessions.

Although in the United States there are few restrictions on the sorts of cases that can be bargained (one can even strike bargains in death penalty cases), most countries use procedures similar to guilty pleas only for lesser charges. Where the punishment consists only of a fine or a suspended jail sentence, many countries allow the prosecutor to issue a 'penal order,' which contains a stipulated charge and sentence, and the defendant has a period of time, say 2 weeks, to accept the order or to request a trial. Italy, France, Russia, and several former Soviet republics have introduced a system whereby a defendant who is facing less than 10 years imprisonment (Russia), or 5 years (France and Italy), agrees not to contest the charges, and in exchange receives a one-third discount on what would have been the sentence. An increasing number of countries are also allowing 'cooperation agreements' in very serious cases, which can lead to large discounts in sentencing, or even dismissal, if a suspect or defendant cooperates with the prosecution to help convict a more serious criminal. Russia took this step in 2009.

Notably, these abbreviated and consensual procedures induce defendants to give up the broad protective rights they have only earned after centuries of struggle, and to accept guilt without a trial in order to avoid the threat of a higher punishment. Because of plea bargaining, trial by jury, perhaps the most protective procedure, is virtually disappearing in the United States, while jury trials in countries that have only recently returned to them, such as Russia or Georgia, are becoming few and far between.

See also: Civil Liberties and Human Rights; Courts and Adjudication; Due Process in Law; Human Rights: Legal Aspects; Judges; Juries; Plea Bargaining: A Comparative Analysis; Prosecution; Torture: Legal Perspectives; Trials.

Bibliography

Bradley, Craig M. (Ed.), 2007. Criminal Procedure: A Worldwide Study, second ed. Carolina Academic Press, Durham, NC.
Damaška, Mirjan R., 1986. The Faces of Justice and State Authority. Yale University Press, New Haven, CT.
Damaška, Mirjan R., 1997. Evidence Law Adrift. Yale University Press, New Haven, CT.
Damaška, Mirjan R., 2001. Models of criminal procedure. Zbornik Pravnog fakulteta u Zagrebu 51, 477–516.
Dawson, John Philip, 1960. A History of Lay Judges. Harvard University Press, Cambridge, MA.
Esmein, Adhemar, 1913. A History of Continental Criminal Procedure with Special Reference to France. Little, Brown, Boston, MA.
Hatchard, John, Huber, Barbara, Vogler, Richard (Eds.), 1996. Comparative Criminal Procedure. British Institute of Comparative Law, London.
Jackson, John, Langer, Máximo, Tillers, Peter, 2008. Crime, Procedure and Evidence in Comparative and International Context. Essays in Honor of Professor Mirjan Damaska. Hart Publishing, Oxford.
Langbein, John H., 1977. Torture and the Law of Proof. University of Chicago Press, Chicago, IL.
Langbein, John H., 2003. The Origins of Adversary Criminal Trial. Oxford University Press, Oxford.
Thaman, Stephen C., 2008. In: Comparative Criminal Procedure: A Casebook Approach, second ed. Carolina Academic Press, Durham, NC.
Thaman, Stephen C. (Ed.), 2010. World Plea Bargaining: Consensual Procedures and the Avoidance of the Full Criminal Trial. Carolina Academic Press, Durham, NC.
Thaman, Stephen C. (Ed.), 2013. Exclusionary Rules in Comparative Law. Springer, Dordrecht.
Trechsel, Stefan, 2005. Human Rights and Criminal Procedure. Oxford University Press, Oxford.
Vidmar, Neil (Ed.), 2000. World Jury Systems. Oxford University Press, Oxford.
Vogler, Richard, 2005. A World View of Criminal Justice. Ashgate Press, Burlington, VT.

Criminology: Development and Current Status of the Discipline

Charles F Wellford, University of Maryland, College Park, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The development of criminology as a separate academic and research discipline is traced from the late 1700s to the present. The major conclusions for criminology are (1) causal theories have not proven to be very robust and powerful. Future theory development will require the development of interdisciplinary theory. (2) Criminology has not experienced significant growth in the last 20 years. The needed expansion of the field will require a new federal effort to build the infrastructure for crime and justice research. (3) The changing nature of crime and scientific methods in criminology, the internationalization of crime, and the demand for more evidence-based policies and practices present new challenges for the discipline of criminology.

Introduction

The goal of this article is to provide an analysis of the development and current status of the discipline of criminology. Criminology is a behavioral and social science. It is not the investigation of specific crimes. Too often the public, under the influence of the media, confuse criminology as a science of behavior, and criminalistics and forensic science as they are applied in the investigation and processing of criminal cases. Here we consider criminology in the sense defined by Sutherland, one of the pioneers in establishing the science of criminology. Sutherland (1934) defined criminology as the scientific study of the making of law, the breaking of law, and the reaction of society to law breaking. Criminology is one of the fastest growing social sciences and at many universities is one of the largest programs in terms of number of majors (between 1970 and today, it is estimated that enrollment in criminology programs at universities in the United States have increased by a factor of 1000, partly because of the rapid growth in the number of such programs).

How we understand the past is often influenced by how we see the present. Today criminology finds itself defined by three major themes: the steady movement toward a more rigorous science; a commitment to the development of rigorously tested theories of crime and criminal behavior; and the establishment of a demand for evidence-based crime control and justice policies and practices. However, the history of the discipline is not as unified. Two, in part, conflicting approaches define the foundation of criminology. The next section describes those approaches and their influence today.

Two Points of Departure for the Beginning of Criminology

The Classical School of Criminology

This pillar of criminology begins with the work on an Italian lawyer, Cesare Beccaria. Working with a small group of his friends and fellow lawyers, Beccaria published what was to become one of the most important works for criminology, and more importantly for criminal justice reform. Published in 1764 this book would become one of the most influential legal treatises of the eighteenth century (*Essays on Crime and Punishment* (hereinafter referred to as Essays)). (He was the primary author but contemporary scholarship suggests that others were intimately involved in the development and writing of the book. While that is not critical for this article, you can find a discussion of this in the forward to the Paolucci edition (1963).) Influenced by the Enlightenment philosophers, Beccaria sought to reform the criminal justice system to make it more humane and fair. He argued for punishments other than corporal punishment and death by embedding punishment in an enlightened legal system. Within 10 years of its publication the book was translated into all European languages; Beccaria was celebrated as a profound new legal thinker; and, the work influenced the governments of numerous countries including England and the United States. As early as 1775 John Adams referenced this work in his justification for accepting the unpopular and politically dangerous task of defending the British soldiers who fired on the citizens of Boston who charged the arms depot atop Bunker Hill. Adams (quoted in McCullough, 2001) in explaining this decision quoted the following from Beccaria:

> If, by supporting the rights of mankind, and the invincible truth, I shall contribute to save from the agonies of death one unfortunate victim of tyranny, or of ignorance, equally fatal, his blessing and years of transport will be sufficient consolation to me for the contempt of all mankind.

Beccaria offered four central tenets of a new approach to criminal justice. First, he rejected the then dominant idea that crime represented a moral failure. With religion as the source of power in society it was understood that personal and social failures resulted from a falling from grace. Beccaria accepted the notion from Bentham (1789) that behavior was always a choice and that choices were determined by a 'moral calculus.' Crime resulted when the rewards of crime outweighed the costs (pains) resulting from the commission of crime. Second, Beccaria argued that because crime was a choice influenced by rewards and punishments, punishments need be just severe enough to offset the rewards of the crime. At a time when the death penalty and corporal punishment were common, he argued for a more

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45052-X

rational system of punishments, one in which there would be no death penalty. Third, Beccaria called for the application of science to determining the exact level of punishment for each crime. These punishments were to be determined with 'geometric precision.' Finally, Beccaria understood that any science of punishment could be seen as challenging the established powers and their interests. So he argued that his new law system was consistent with the teachings of the church. These central ideas formed the basis for criminal justice systems in the emerging western democracies and an agenda for criminology that focused on rationalizing the criminal justice system.

In the second foundation for criminology we see a very different set of assumptions and foci. This approach pays little attention to the criminal justice system and the focus is almost exclusively on the causes of crime.

The Causes of Crime

Cesare Lombroso was an Italian physician, who worked in local prisons. Apparently heavily influenced by the recently published work on Darwin on evolution, Lombroso considered how criminality could be explained within an evolutionary perspective. Like Beccaria, his contribution was to take an emerging world view and apply it to the problem of crime. As he worked with criminals he began to notice physical characteristics that suggested to him that they had not fully developed as humans – their criminality was a result of this lack of development. This idea was summarized with numerous examples in his book *Criminal Man* published in 1876. This brief book (approximately 250 pages) argued that crime was caused and that it was caused by a biological developmental anomaly. Criminals did not choose to commit crime; their behavior was caused by factors beyond their control. Until the beginning of the twentieth century the idea that biology was destiny for criminals became increasingly popular. Not until Charles Goring tested this approach by conducting a statistical analysis of the physical characteristics of criminals and noncriminals did students of crime begin to explore other approaches to explaining crime (Goring, 1913). Still, Lombroso was successful in directing the study of crime to a new set of questions: what causes crime and what does our knowledge of crime indicate for a system of crime prevention and control – the second approach influencing criminology as it emerged as a separate discipline. (Lombroso gradually began to understand that factors other than biology influenced crime. By the time he wrote the fifth edition of his book it had grown to 1903 pages and identified hundreds of causes of crime. Still he is most remembered for his emphasis on the biology of crime. For a detailed discussion of this period in criminology see Chapters 3 and 4 in Vold (1979).)

In fact, by the time this work was published Lombroso had published the fifth edition of his book, each one getting longer and noting other possible explanations of crime (the fifth edition had grown to 1903 pages and listed hundreds of causes of crime). (This in part explains Lombroso's new appreciation in contemporary biological criminology (DeLisi, 2013).) In 1885, Rafaele Garofalo (a student of Lombroso) introduced the word criminology in his influential book by the same name (*Criminology*). It would be over 50 years from then before a professional society devoted to crime and justice would emerge, and separate academic departments of criminology would be created in American universities. As it emerged some 75 years ago as a new social and behavioral science, it stood on two foundations: the science of criminal justice and the science of criminal behavior. In very important ways they continue to be the two, sometimes conflicting, foci on contemporary criminology.

Explaining Crime and Criminal Behavior

Even though Beccaria wrote first, the emergence of a science of criminology started with the focus on causes of crime introduced by Lombroso. It was not until the mid-twentieth century when criminology began to pay scientific attention to testing and improving the criminal justice system. While criminological theory has been less cumulative than you would expect in an emerging science, it has accumulated. The path of theories of crime has moved through four distinct forms of theory development: single factor reductionism; systemic reductionism; multidisciplinary theories; and interdisciplinary theory. (These are themes that seem more prominent at different points in time. This is not to imply they have occurred in a liner fashion. In fact, all approaches exist today.) Today, interdisciplinary theory stands as the dominant model for explaining crime (and other human behaviors). It is also the case that while Lombroso focused on why individuals commit crimes, crime theories are also used to explain why crimes occur at different levels in different sections of the population, geographical areas, and/or times. Being clear about what the theory is intended to explain is vital to understanding its strengths and weaknesses (for a more detailed discussion *see* Sociology of Crime). In addition, theories tend to not distinguish the kind of crime being explained, as if one theory could explain homicide and use of controlled substances. (See articles in this volume on specific crimes for more details on their distribution and causes.)

Single Factor Reductionism

In an effort to explain crime in the simplest way, some have offered a single factor or idea to explain all crime. Following Lombroso numerous authors offered explanations of criminal behavior: 'feeblemindedness' and family transmission, poverty, single parent families, abortion, lead paint – the list could go on and on. The most positive observation on these approaches is that they identify characteristics of crime that a theory should address. More to our point, these approaches tend to distract criminology into research to test ideas that even Lombroso understood were limited as he expanded his notion of what could be involved in causing crime. Still these approaches have the advantage of simplicity, and they have clear policy implications. By the end of the first quarter of the twentieth century it was understood that no single factor could explain criminal behavior and crime rates. (This approach continues today in some public and political discourse. Too frequently we hear that crime is caused by 'welfare moms,' lead paint, unemployment, and a variety of other factors. Criminologists can usually respond yes, but that is just one of a myriad of characteristics that may cause crime and the problem is we do not know which is most important and how these many factors interact. This type of response reinforces the idea held by some

that criminologists do not know much about the causes of crime.) However, just as difficult as saying a single dimension explains crime is to say everything causes crime. In the search for some way to limit and organize explanations the period of systemic reductionism emerged.

By systemic reductionism I mean the approach that still dominates the field, the assertion that one discipline can explain crime. Hence, in many recent introductory texts there are separate chapters or sections on biological theories, psychological theories, sociological theories, and cultural theories. Since criminology emerged in the United States in departments of Sociology, that approach dominated theories of crimes. Along with this focus, came a focus on crime rate variation as opposed to Lombroso's focus on individuals who committed crime. Most of contemporary criminological theory contains a strong if not exclusive sociological core. The works of Sutherland (1934), Shaw and McKay (1942), Merton, Schur (1971), and those who tested and elaborated their works form the core of much of contemporary theory in criminology.

Multidisciplinary Approaches

It took a husband and wife team trained in education, social work, law, and social ethics (Sheldon and Eleanor Glueck), and heavily influenced by a psychiatrist brother (Bernard Glueck) to challenge the notion of systemic reductionism. For the Gluecks it was obvious that the work of psychiatry, biology, and psychology was as important as sociology in understanding all human behavior including crime. They set out to understand all of the factors that distinguished criminals and delinquents for those who did not engage in those behaviors. In one of the clearest statements of the problem of systemic reductionism they stated:

> It is common knowledge that the great majority of children growing up in urban slums do not become delinquents despite the depravations of a vicious environment, despite the fact that they and their parents have not had access to fruitful economic opportunities, despite the fact that they are swimming around in the same antisocial subculture in which delinquents and gang members are said to thrive. Why?
>
> Glueck and Glueck, 1968

Sociological approaches might be appropriate to explain the variation in rates across class levels, but it could not explain why similarly situated individuals did or did not commit crime. To address that question, the Gluecks turned to all of the social and behavioral sciences as they were practiced in the 1930s to the 1970s. In a series of retrospective longitudinal studies they sought to determine how to distinguish those who did and did not violate the law by measuring every dimension of large samples regardless of the discipline the distinguishing factor might be drawn from. Unfortunately, this approach offers little basis for theory development or explanation. It ends up being an approach to prediction not explanation. So in their best-known work, *Unraveling Juvenile Delinquency* (1950) and its elaboration *Delinquents and Nondelinquents in Perspective* (1968), the Gluecks did offer a theory of delinquency but rather a 'social prediction' scale that would achieve their goal of distinguishing law violators from others.

Of course, we now understand that multidisciplinary approaches are not theories. They are the search for all of the factors that need to be addressed by theory. There is always another factor to consider. There is no logical end to the effort. The only end is the explanation of all variation. In that sense, it is more like what today we call data mining than it is explanation. However, the demonstration that there is more to explanation than sociology ushered in the current focus for explanation in criminology – interdisciplinary theory.

Interdisciplinary Explantations

Although the growth of criminology was stimulated by federal programs aimed at improving justice, as it emerged as a separate field of study it defined itself as interdisciplinary – not multidisciplinary. Without this understanding of its central characteristic there would be no need for a separate discipline. The study of crime and justice could continue in the various traditional, social, and behavioral sciences. While many academic criminologists continue to be located in the more traditional departments, the field of criminology understands itself to be interdisciplinary. So when the American Society of Criminology emerged as a scientific, scholarly organization, it placed an interdisciplinary perspective as being central to criminology. This is best exemplified in *Criminology*, the official journal of the American Society of Criminology, which has its subtitle *An Interdisciplinary Journal*.

As noted earlier this does mean that all theory in criminology is now interdisciplinary. In fact, earlier theories continue to dominate, but increasingly, we and others understand their limits. A recent article, by Weisburd and Piquero (2008), is an indicator of the problem. They perform an analysis of all recent tests of contemporary criminological theory. They empirically demonstrate that:

> The overall level of variance explained is often very low with 80–90% unexplained. There has been no improvement over time.

Interdisciplinary theory is difficult to define and more difficult to achieve. We know it must contain more than one discipline, and that the contributions from each discipline must be independent but related (for a complete discussion, see Wellford, 1989). An example of contemporary interdisciplinary theories is the work of Terrie Moffitt (2001). This theory includes all disciplinary levels, recognizes the changing nature of criminal behavior and its causes over time, and seeks to identify different patterns of crime to explain. For Moffitt (2001), there are two patterns of antisocial behavior: life course persistent where offenders start early and maintain their involvement in antisocial behavior throughout the life course; and, adolescent-limited whose antisocial behavior emerges and ends during that period of the life course. Moffitt observes that the persistent ones result from "childhood neuropsychological problems interacting cumulatively with their criminogenic environments producing a pathological personality." With the adolescent-limited pattern "a contemporary maturity gap encourages teens to mimic antisocial behavior in ways that are normative and adjustive." For Moffitt, explanation of these patterns included biology, psychology, and social levels of explanation. The history and

research in criminology on causes of crime point to the importance of interdisciplinary models; the contention that criminology is a separate and distinct discipline rests on a recognition that interdisciplinary theory is at the core of criminology.

The Beccarian Revival

Criminology returned to a focus on the justice system in the 1950s and 1960s. Studies on arrest, prosecution, and imprisonment began to emerge in this period that pointed to clear problems in almost every aspect of the criminal and juvenile justice systems. Not only were they failing to prevent and control crime, they were found to be disparate in their application. They were neither effective nor fair. When the urban riots of the 1960s and law enforcements reaction to the aroused national concern, a Presidential Commission was established to consider how to improve the system. The report of that commission (1967), titled the *Challenge of Crime in a Free Society*, while acknowledging the need for better understanding of the causes of crime, placed greater emphasis on the important of reforming criminal and juvenile justice. At the same time the report acknowledged how little we knew about these systems and the importance of research to gain the understanding that could support reform. The result was a substantial new effort by the federal government to support research and grow the research infrastructure to conduct research on crime and justice. The development of criminology was directly stimulated by this call for research to reform justice in America. Since the actions taken to implement the recommendations of the President's Commission, criminology has emerged as a new field at most universities in the United States, a source of major funding for the social and behavioral sciences, and a source of findings and ideas that have profoundly influenced criminal and juvenile justice here and around the world.

Few major initiatives in criminal law and criminal justice during this period have been developed without significant criminological involvement. Hot spots policing (Sherman et al., 1989), problem oriented policing (Goldstein, 1990), crime mapping and analysis (Harries, 1974), sentencing guidelines, specialized courts, sex offender programing, effective models of rehabilitation, reentry programs, problems of eyewitness identifications, procedures for police lineups and interrogations, the role of DNA in courts and policing – I could go on and on (*see* Criminal Justice, Sociology of for more a fuller discussion). The point is that criminology is making a difference. As we note in the following section, this has culminated in a new emphasis throughout the justice system – evidence-based policy and practice.

Contemporary Issues in Criminology

Criminology today confronts many challenges and issues. In this section, we discuss some of the most important issues facing the field today. Of course, one is the further development of interdisciplinary theory of crime discussed above. We do not repeat that discussion but note its centrality to the further development of criminology as a discipline. Advances in interdisciplinary theory and the resolution of the following issues will define how well criminology does in the future.

Growth of the Discipline

While undergraduate enrollments in criminology programs have surged in recent years, the number of universities with criminology departments offering advanced degrees has hardly grown in the last 15 years (during the last rankings of graduate education in the United States by the National Research Council there were still not enough doctoral programs to warrant the inclusion of criminology in the rankings). Few elite universities offer degrees in criminology. Membership in criminological professional societies is stagnating. Federal funding for crime and justice research has actually decreased as a percentage of federal spending. Today, some of the most important advances in criminology come from researchers in other disciplines and in the nonuniversity research community. This is occurring as the demand for criminological research is increasing. It will fall to the leaders of the leading criminological professional societies to generate support for a new period of federal funding for expanding the infrastructure for criminological research. Just as they have done in many other research and applied fields, the federal government must lead the way in the expansion and improvement of criminological higher education.

The Measurement of Crime

In the late 1920s a group of police officials and academics created, with no empirical research, an index of crime that exists to essentially unchanged. Managed by the Federal Bureau of Investigation (FBI), this voluntary but hopefully uniform system assumes that a count of homicides, serious sexual and other assaults, robbery, burglary, larcenies where the loss is above $500, and auto theft can provide an index to changes in all crime. Only in a country with as weak a statistical system to measure crime as the United States would take seriously this index of crime as a serious measure. In recognition of this fact, in 1980 a new system was proposed (the National Incident Based Reporting System) which has not been adopted by even a majority of jurisdictions in the United States. In addition, in the over 80 years since the creation of the FBI's uniform crime reporting system crime itself has changed. Cybercrime, Internet fraud, business-related crimes, etc., were unheard of in 1929. These crimes can be construed as crimes in the FBI index of crime (i.e., Internet fraud is a form of theft), however, most are not reported to police (in fact are not even known to the victim). In recognition of these changes, the Bureau of Justice Statistics has recently launched an effort to reconsider the kinds of crimes counted in an index and the mechanisms that would be used to count those crimes (NCS-X, 2012). Developing new and more robust measures of crime and their characteristics is essential to the future of criminology as a science of crime and justice. These ideas are more fully developed in the article on 'Crime Trends and Debates' in this Encyclopedia.

Changes in the Science of Criminology

In two recent reports from the National Academies of Science (NAS) challenged the notion that important criminological questions can be answered with what has been a central component of the criminological toolbox, statistical modeling of existing data. In consideration of the research on the impact

of right to carry laws on crime, the NAS committee on firearms (National Research Council, 2005) "concluded that with the current evidence it is not possible to determine that there is a causal link between the passage of right-to-carry laws and crime rates" (p. 150). In an Appendix to that report, Joel Horowitz explained that "In summary, the problems posed by high-dimensional estimation, misspecified models, and lack of knowledge of the correct set of explanatory variables seem insurmountable with observational data" (p. 308). Similarly, in an NAS report on the death penalty (NRC, 2012) the committee reached the conclusion that the research on the death penalty was so flawed (by characteristics similar to those discussed by Horowitz) that it could not be determined if the death penalty had any effect on crime rates. Research using observational data, with weak or no theory, and models and specifications that are highly dimensional are not likely to produce useful knowledge. Unfortunately this characterizes much criminological research (*see* Quantitative Methods and Crime).

It may be that two recent developments are in reaction to the problems noted above. First is the increased emphasis on experimentation as the preferred method for causal research. Of course, randomized controlled trials have long been understood to be the gold standard for research. However, in criminology the use of experiments has been rare until the last 20 years. Now there is a *Journal of Experimental Criminology* and a section of the American Society of Criminology devoted to experimental research. While focused on the testing of intervention effects, the introduction of experimentation as a focus of criminology has contributed to the challenge to more traditional research that relies on existing data and the use of modeling to account for problems in the data. The second development is the application to data mining with a corresponding shift from causal to predictive understanding (e.g., see Berk, 2009; Brantingham, 2011). This approach seeks only to predict outcomes and thus overcome the limitations of modeling of existing data that comes with a search for causation with nonexperimental tools.

The Internationalization of Crime and Criminology

While criminology began in Italy and other parts of Europe, it developed and flourished in the United States. While criminology was developing as a social and behavioral science in the United States, in the rest of the world it continued, if it existed at all, as a part of the study of law. In recent years, two changes have occurred, a change more fully developed in the article on 'International/Comparative Criminology' in this Encyclopedia. Criminology has emerged as a social science with an empirical research focus in countries around the world (in England and some Scandinavian countries there is a longer history of criminology, but even in those countries research criminology has expanded in the last 50 years). Crime has become more international. These two developments have resulted in a growing recognition of the importance of an international perspective in criminology. Even in countries without an established field of criminology, there is strong interest in understanding the discipline and how it might be used in their countries (for example, both China and Vietnam have developed in conjunction with the University of Maryland Criminology master's degrees in their countries as a way to stimulate the development of the criminological study of crime and justice (www.oiep.umd.edu)). The study and control of organized crime, drug markets, human trafficking, money laundering, terrorism, and cybercrime inevitably involve an international focus. In short, criminology can no longer be the study of crime only in the United States.

Evidence-Based Policy

In 1984 the United States Congress adopted the Comprehensive Crime Control, which was signed into law by President Regan. One aspect of this law was the requirement that funds distributed by the federal government to the states to address crime had to be used to support programs of proven effectiveness. Only those programs demonstrated to be effective at reducing crime or improving justice could be funded (as with most legislation there were some caveats to this requirement but this intent was very clear (see Title V on the Comprehensive Crime Control Act of 1984)). This requirement was never met primarily because the science required to implement it did not exist. Today, at every level of government the demand is for evidence that existing and new programs 'work.' In the field of criminal and juvenile justice that means the need for more and more research on interventions to control crime and improve justice. The efforts of the Office of Justice Programs to promote this approach are most evident in a Web site that seeks to identify all successful and unsuccessful programs (www.CrimeSolutions.org). The pressure on criminology to conduct rigorous evaluations requires more criminological research and better ways to assemble and assess that research.

The Importance of Partnerships

Criminology is no longer a science that can be done with access to existing data in archives or on shelves. It requires direct access to the objects it studies. This means developing partnerships with law enforcement agencies to gain access to personnel and records, and directly studying and observing offenders. Fortunately, law enforcement agencies have understood the value of criminological research and have encouraged criminologists to partner with them. This is especially true in policing, where, for example, the International Association of Chiefs of Police have produced a guide for chiefs to develop effective research partnerships and have started an annual research award that has as one of its central criteria partnerships with researchers. While prosecutors and correctional officials have not been as welcoming, they too have recognized the necessity for evidence-based policies and practices and thus the need for research enhancements. As for the study of causation, criminology may be approaching the situation in suicide research where the leader of that research community is quoted in the *New York Times* magazine as follows:

> We've never gone out and observed, as an ecologist would or a biologist would go out and observe the thing you're interested in for hours and hours and hours and then understand its basic properties and then work from that, Matthew K. Nock, the director of Harvard University's Laboratory for Clinical and Developmental Research, told me. We've never done it. (June 2013)

The divide between in-depth case and qualitative studies and quantitative research in crime causation is challenged not only by the failure of theory, but the recognition in economics and other social and behavioral sciences that understanding behavior requires deep observation of that behavior and those who engage in it. This may the biggest challenge for criminology in the future.

Bibliography

Beccaria, C., 1963. Essays on Crime and Punishment (H. Paolucci, Trans.). Bobbs-Merrill, Indianapolis.

Berk, R., 2009. Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning (with Lawrence Sherman, Geoffrey Barnes, Ellen Kurtz, and Lindsay, Ahlman). Journal of the Royal Statistical Society (Series A) 172 (Part 1), 191–211.

Brantingham, J.P., 2011. Crime emerges. In: McGloin, J., Sullivan, C., Kennedy, L. (Eds.), When Crime Appears. Routledge, New York.

Bureau of Justice Statistics, 2012. NCS-X Project. Washington, DC.

DeLisi, M., 2013. Revisiting Lombroso. In: Cullen, F., Wilcox, P. (Eds.), The Oxford Handbook of Criminological Theory. Oxford University Press, New York.

Garofalo, R., 1885. Criminology. Reprinted in 1968. Patterson Smith, Montclair.

Goldstein, H., 1990. Problem Oriented Policing. McGraw-Hill, New York.

Goring, C., 1913. The English Convict. Reprinted 1972. Patterson Smith, Montclair.

Glueck, S., Glueck, E., 1950. Unraveling Juvenile Delinquency. Commonwealth Fund, New York.

Glueck, S., Glueck, E., 1968. Delinquents and Nondelinquents in Perspective. Harvard University Press, Cambridge.

Harries, K., 1974. The Geography of Crime and Justice. McGraw-Hill, New York.

Lemert, E., 1972. Human Deviance, Social Problems and Social Control, second ed. Prentice-Hall, Englewood-Cliffs.

Lombroso, C., 1876. Criminal Man. Hoepli, Milan.

McCullough, D., 2001. John Adams. Simon and Schuster, New York.

Merton, R., 1938. Social structure and anomie. American Sociological Review 3, 672–682.

Moffitt, T., 2001. Adolescence-limited and life course persistent antisocial behavior. In: Piquero, A., Mazerolle, P. (Eds.), Life-Course Criminology. Wadsworth, Belmont, pp. 91–145.

National Research Council, 2005. In: Wellford, C., Pepper, J., Petrie, C. (Eds.), Firearms and Violence. National Academies Press, Washington, DC.

National Research Council, 2012. In: Nagin, D., Pepper, J. (Eds.), Deterrence and the Death Penalty. National Academies Press, Washington, DC.

President's Commission on Law Enforcement and the Administration of Justice, 1967. The Challenge of Crime in a Free Society. GPO, Washington, DC.

Sampson, R., 2001. Foreword. In: Piquero, A., Mazerolle, P. (Eds.), Life-course Criminology. Wadsworth, Belmont, pp. 91–145.

Schur, E., 1971. Labeling Deviant Behavior. Prentice-Hall, Englewood-Cliffs.

Shaw, C., McKay, H., 1942. Juvenile Delinquency and Urban Areas. Chicago University Press, Chicago.

Sherman, L., Gartin, P., Buerger, M., 1989. Hot spots policing. Criminology 27, 27–55.

Sutherland, E., 1934. Principles of Criminology. Lippincott, Philadelphia.

Vold, G., 1979. Theoretical Criminology. second ed., prepared by T. Bernard. Oxford University Press, Oxford.

Weisburd, D., Piquero, A., 2008. How well do criminologists explain crime?. In: Crime and Justice: A Review of Research, vol. 37, pp. 453–477.

Wellford, C., 1989. Towards an integrated theory of criminal behavior. In: Liska, A., Krohn, M., Meissner, S. (Eds.), Theoretical Integration in the Study of Deviance and Crime. SUNY Press, Albany, New York.

Relevant Websites

Professional Societies

www.acjs.org – Academy of Criminal Justice Sciences.
www.asc41.com – American Society of Criminology.

National Organizations

www.bjs.gov – Bureau of Justice Statistics.
http://sites.nationalacademies.org/DBASSE/CLAJ – Committee on Law and Justice, National Research Council.
www.ncjrs.gov – National Criminal Justice Reference Service.

International Programs

www.oiep.umd – Office of International Programs.

Criminology, History of

Joachim J Savelsberg, University of Minnesota, Minneapolis, MN, USA
Lorine A Hughes, University of Nebraska, Lincoln, NE, USA
Janne Kivivuori, National Research Institute of Legal Policy, Helsinki, Finland
James F Short Jr., Washington State University, Pullman, WA, USA
Maximo Sozzo, Universidad Nacional del Litoral, Santa Fe Province, Argentina
Richard Sparks, University of Edinburgh, Scotland, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The history of criminology is examined comparatively for four countries or regions: the United States, Latin America, Scandinavia, and the United Kingdom. Each case history considers a common set of analytic dimensions: origin and takeoff (time, discipline, political context); changing shapes (themes, theoretical orientations, data); changing organization (associations, journals, position in universities, government and nonprofit institutes); and the political-economic environment of criminology. Commonalities across national and regional developments and particularities of each region speak to the conditions of criminology. Globalization is at work, partly under US guidance, but academic exports are challenged and adapted to nation-level political, ideological, and institutional contexts.

Criminology, the study of the causes and consequences of crime and societal and governmental responses to crime, is today represented around the globe. Yet, its history and institutionalization vary by country and region. The relative weight of the social science component in this multidisciplinary field has grown over the past century, but its prominence varies as the field cohabits and competes with legally oriented or biologically oriented criminologies. The history and current shape of criminology (from here on referring to social science criminology if not otherwise specified) hinge on economic and political conditions common to all sciences and on the organization of academia within universities, government-sponsored and private institutes, and disciplines already established when criminology emerged.

Four case histories are briefly described below, specifically those of the United States (US), Latin America, Scandinavia, and the United Kingdom (UK). They display commonalities and particularities of criminologies across the globe. Commonalities of academic development across countries and world regions then point at global forces that affect criminology irrespective of its geographic location. Particularities speak to national and regional factors that contribute to local distinctions. Examining commonalities and particularities thus helps us understand the social conditions of criminology and its history. As we examine our four countries/regions, each case history considers a common set of analytic dimensions:

1. Origin and takeoff: time, discipline, political context;
2. Changing shapes: themes, theoretical orientations, data;
3. Changing organization: associations, journals, position in universities, and government and nonprofit institutes. Of interest in its own right, the organization of criminology also affects the substance of criminological work. Relatively autonomous academic organizations free scholars from immediate political and economic concerns and allow for the formation of networks of competition and collaboration (Collins, 1998). Within universities, organizational competition shapes the nature of newly emerging disciplinary units as demonstrated for early twentieth century sociology (Camic, 1995). The organization of departments and journals subsequently affects the content of scholarly work (for criminology see Laub and Sampson, 1991; Savelsberg et al., 2004; Short and Hughes, 2007).
4. Political-economic environment of criminology: While economic surplus is a prerequisite for the differentiation and expansion of science, political context can be restricting and enabling. Historical research on criminology under totalitarianism depicts limitations, while simultaneously showing niches of free exploration within the specialized academic field (Wetzell, 2000 on Nazi Germany). Under liberal democratic regimes, governments affect the substance of criminology through funding of institutions and research or through the stimulation of demand for professional degree programs (Savelsberg and Flood, 2011).

Criminology in the United States

Criminology in the US emerged in the late nineteenth and early twentieth centuries, accompanying social changes that highlighted the need for scientific studies of crime and criminals. Police departments in rapidly growing cities broadened the scope of public responsibility for crime and made that responsibility problematic. Theories concerned with the effects of prison conditions were hotly debated, and innovations proposed; participants in these debates were part of a social movement to construct 'asylums' for criminals and other unfortunates (Rothman, 1971). Concerns for children prompted legal changes, special courts and institutions.

Journalists highlighted the association of crime and social conditions, stimulating government and citizen, as well as scholarly activities. Stirrings of social science concerns with crime in the US are reflected in the 1865 constitution of the American Association for the Promotion of Social Science that referred to goals such as 'Prevention and Repression of Crime,'

and 'Reformation of Criminals' (Lengermann and Neibrugge, 2007: p. 74).

Programmatic scientifically inspired empirical research on the social distribution of juvenile delinquency (Shaw et al., 1929) led to sociological hegemony in criminology, albeit not without challenges from biology (Rafter, 1997), while academic texts marked the autonomy of the field. Criminology was integral to the dramatic growth of social sciences in the US through the 1920s and 1930s (Rice, 1931).

The inherently interdisciplinary nature of criminology created a great deal of theoretical ferment. Robert Merton's (1938) theory of deviant adaptations to structural limitations on goal attainment marked a shift toward an emphasis on social structure. Advances in survey methodology made study of individual delinquents and criminals more efficient. Individual-level research and theory flourished with advances in psychology and social psychology, and economics brought the rational action paradigm to the field.

Criminological theory over the past half century elaborates and adds to insights from these disciplines and from advances in the biological sciences, an early precursor of criminology. Social disorganization theory and Sutherland's (see Cohen et al., 1956) ideas of differential association and differential social organization began a process of formalizing theory in criminology that continues even today. Social learning theory, borrowed from psychology, became a major perspective, as did control and labeling theories, which owed much to their roots in symbolic interaction theory. In recognition of failures to explain variations of regulatory capacities of social relationships and neighborhoods, 'social capital,' an addition to human capital, was in turn added to as 'collective efficacy,' and social disorganization became systemic disorganization.

While macro- and individual-levels of explanation were strengthened by these developments, situational contexts and group processes received less emphasis. However, methodological and theoretical innovations in network analysis and analysis of spatial relationships, together with new technologies for qualitative data analysis, offer promise of more holistic scholarship in criminology (Sampson, 2012).

Formal organization of criminology came about in conjunction with organizations devoted to law enforcement and corrections. The American Society of Criminology evolved from an organization formed initially for the purpose of professionalization of police. The constitution of one of its progenitors defined 'criminology' as 'the study of the causes and prevention of crime, including, but not restricted to,' several law enforcement and crime control goals (Morris, 1975: p. 128).

Implications of research for crime prevention and control have been problems for criminology from its beginnings. Debate also centers on whether criminology is a fundamental discipline or a special interdisciplinary field; and/or an applied vs a theoretical- and research-oriented enterprise. Partisans of the disciplinary view argue that criminology's association with law and its multimethod, theoretical and empirical maturity warrants such recognition. The dependence of criminology on the social science and behavioral science disciplines as well as the influence of extrascholarly influence on the field's subject matter challenge this view. In either case, the lines between criminology and its supporting disciplines increasingly are blurred, as criminologists continue to benefit from and contribute to developments in theories and methods in the social sciences, law, and other special fields, including comparative criminology and Third World human rights and conflict (Hagan and Rymond-Richmond, 2009; Savelsberg and King, 2011).

Criminology in Latin America

Criminology in Latin America was born when ideas of the Italian 'Positivist School' were imported in the 1880s, both in the fields of law and medicine. This importation was stronger in some national contexts such as Argentina, where the Society for Juridical Anthropology was founded in 1888. In the same year, the first Latin American book on criminology was published. Entitled *Los Hombres de Presa* and authored by Luis M. Drago, it was translated into Italian and published with a highly favorable prologue of Cesare Lombroso. The year 1898 witnessed publication of the first scientific journal on criminology in Latin America, entitled *Criminalogia Moderna*. In national contexts like Brazil, Mexico, Cuba, or Argentina, between the end of the nineteenth century and the 1940s, Lombrosian ideas about the 'born criminal' were widely diffused and associated with strong racist connotations.

In reaction to the biological school, challengers emerged in the early twentieth century. They were inspired by the 'French School' that stressed the psychological and sociological dimensions of crime causation (e.g., Debuyst, 1998; Mucchielli, 1994a,b). Most prominent among these scholars was Jose Ingenieros who developed a new classification of criminals. He did not discard the idea of criminals as abnormal, but he explained abnormality entirely in a psychopathological way. Guided by this framework, Latin American universities created first institutional spaces for criminal anthropology and criminology, particularly in Schools of Law and Medicine. Further schools emerged in affiliation with police and prison institutions. For example, in 1907 the Institute of Criminology was created within the National Penitentiary of Buenos Aires, and Ingenieros was appointed its first director. Throughout the first half of the twentieth century, positivist criminology was consistently promoted by dictatorial and authoritarian governments – albeit with important differences across national contexts. This movement's impact on penal laws, culture, and practices was substantial and continues to be recognizable.

This 'golden age' of positivist criminology reached its peak during the 1940s. Following trends in post–World War II Europe, psychiatric and psychological approaches were emphasized. The focus was always on the individual. Yet, this era also witnessed a general decay in the volume and quality of intellectual production. In some countries, this moment was even labeled the 'era of the criminological silence.' Nevertheless, this psychopathologically reoriented positivist criminology generated new degree programs in criminology. Examples are the University of Buenos Aires (Argentina) in 1966 and the University of Nuevo Leon (Mexico) in 1974. These programs, following a medical model, sought to generate professionals for the production of diagnostic and prognostic reports on 'criminals' in judicial and

prison institutions. Yet, these experts barely influenced the construction of crime control policies, a role that continued to be monopolized by criminal law professors.

Also in the 1960s, some Latin American countries began to import sociological criminology, especially from the US. In a further step, the 1970s witnessed a critical criminology taking hold when new ideas from English speaking countries and Italy were translated into Spanish. Law schools were at the epicenter of this movement, particularly in Venezuela, Mexico, and Colombia. The strong connection of this critical criminology with neo-Marxist political and theoretical languages, however, blocked these developments in countries governed by military dictatorships like Brazil and Argentina.

Since the 1980s criminology's opening toward social science and critical theory, in a field previously dominated by positivist criminology, was promoted by the processes of democratization. Critical vocabulary was diffused and institutional spaces multiplied in law schools across Latin America. Nevertheless this intellectual movement did not translate into the creation of new criminology degree programs, while earlier degree programs in the tradition of the positivist school declined and, in some cases, were closed.

At the beginning of the 1990s, building on initiatives of the preceding two decades, numerous researchers from different social sciences (sociology, anthropology, and political science) began to conduct research on crime and crime control, in many cases beyond the label 'criminology' as an institutionalized academic field. These scholars contributed new theoretical perspectives and empirical research. New institutional spaces were opened in schools of social sciences. In general, the relevance of these types of studies increased significantly, beginning in the second half of the 1990s, as issues of crime and crime control moved center stage in political and public debates.

As a consequence, this field of studies assumed, during these last 20 years, a polyvalent and complex structure. Many researchers today produce knowledge in the name of 'criminology,' others work on similar themes but under different disciplinary labels. Again, as empirical research and theoretical discourse expanded, many Latin American universities created new master's degrees in criminology. Law Schools continued to be their prime institutional location. Simultaneously, scientific journals emerged such as *Capítulo Criminológico* (1973) and *Delito y Sociedad* (1992). In short, the newly emerging field of criminology in Latin America is a growing, complex, and polyvalent field. It entails strong contradictions and debates within and faces substantial challenges regarding both knowledge production and policy relevance.

Criminology in Scandinavia

Criminology in Scandinavia, in its intellectual development, conforms to the overall Continental European patterns. Before the establishment of an independent social science criminology, studies of criminal behavior, policy, and law were conducted by scholars with background in psychiatry or law. Yet, after World War II, criminology took off as an independent social science discipline, as distinct from law and psychiatry.

In terms of core ideas and orientations, the Scandinavian development similarly followed the general shifts of twentieth century criminology from the abnormality to the normality paradigm: While earlier scholars considered crime as a manifestation of individual physical or psychological abnormality, criminologists in the postwar years increasingly redefined it as a normal social phenomenon. The aim of understanding crime as a societal phenomenon was adopted early on by progressive lawyers who established so-called associations of criminalists in Denmark (1889), Norway (1892), Sweden (1911), Finland (1934), and Iceland (1949). In the early phase, the gradual 'sociologization' of criminology was spearheaded by socially minded lawyers. In the postwar era, sociologization gathered momentum as sociology proper emerged as the intellectual backbone discipline of criminology. During the 1960s and 1970s, the profession of psychiatry was sidelined from the core of criminology, while criminal law experts have remained central.

Two factors help explain the postwar modernization and empirical sociologization of Scandinavian criminology. First, World War II implied a drastic break and cultural reorientation from the Continental European tradition to the Anglophone world. Most scientific influences came from the US, and Nordic social science rapidly became 'Americanized' (Haney, 2009). Second, since the early 1960s the Nordic governments created a new international organ of cooperation, the Scandinavian Research Council for Criminology (SRCC). Established in 1961, the SRCC has for more than 50 years channeled government funding for social science criminology. In contrast, the weight of university-based criminology varies between the Nordic countries, ranging from Sweden with a very strong university-based criminology to Finland where criminology is nearly nonexistent in universities.

After the late 1960s, the uniformly 'Americanized,' quantitative and positive approach of Nordic criminology ended as the critique of positivism inspired some criminologists to abandon quantitative criminology for the benefit of qualitative and critical approaches. To some extent, national criminology cultures diverged as a result of this antipositivist turn. Over the recent years, the internal cleavages of Scandinavian criminology have widened, a process that culminated in Sweden's secession from the SRCC in 2012.

Irrespective of such rifts, government support remains the most important source of research funding in Scandinavia. It is channeled through the SRCC, government research institutes, government grants toward evaluation studies, and publicly funded universities. An important outlet for publications is now the *Journal of Scandinavian Studies in Criminology and Crime Prevention* (founded in 2000).

In terms of research substance, Nordic criminology is characterized by the paradigm shift from abnormality constructions to the normality perspective. This shift was heavily embedded in the corresponding shift of the preferred data basis of criminology: away from official statistics to survey data on crime. The very first project of the SRCC was the Nordic Draftee Research (NDR) project (1961–64) that also marks the world's first internationally comparative crime survey. The NDR was launched to show empirically that crime was very prevalent and, therefore, statistically normal (for an in-depth history of the normality paradigm shift in US-Scandinavian criminology,

see Kivivuori, 2011). Since then, crime measurement and analysis of real trends, apart from recorded changes, have remained a central concern of Nordic criminology. Victim surveys and self-report surveys have therefore remained favored data sources (Kivivuori and Bernbrug, 2011). The Nordic homicide research tradition has also been strong. It includes groundbreaking internationally comparative work of Finnish criminologist Veli Verkko (1893–1955) who created the world's longest uninterrupted national homicide time series for Sweden and Finland (since 1754). Throughout, Nordic criminology has been concerned with the practical goal of supporting an evidence-informed and nonpunitive penal regime (Kyvsgaard, 2012). Much of the internal cohesion and emotional energy of the Scandinavian criminological research community derives from this motive.

Criminology in the United Kingdom

Criminology in the United Kingdom is an activity with extended antecedents but a relatively short history. As Garland (1988) has shown, the scientific, clinical, and reformist interests of a wide range of practitioners, but initially perhaps especially psychiatrists and social workers, long predated the development of research institutes, or university degree programs, or conferences devoted to the promotion of criminology as a 'subject.'

The generative context for the formation of criminological discourse in Britain thus includes the activities of a range of (mainly quite gradualist) movements for social reform and 'improvement,' often of Protestant religious affiliation; the widespread and intensifying attachment of respectable opinion throughout the later nineteenth and early twentieth centuries to science; and the specific development of a range of institutions and practices aimed at correcting the habits, morals, and lifestyles of the disreputable poor. Emblematic here are such innovations as the Probation of Offenders Act (1907) with its resonant injunction to 'advise, assist, and befriend' the wayward, the 'delinquent,' the inebriated.

In its articulations with the emergent 'penal welfare complex' (Garland, 1985) criminology in Britain thus begins both within and outside 'the state.' These pluralist and interventionist characteristics help to explain why there has never been a predominance either of Beccarian 'Classicism' or of programmatic Positivism in British criminology despite its attachments to science and its proximity to clinical practices. Neither do the textbook oppositions between such 'paradigms' very well describe its eclectic, pragmatic, empiricist, and reformist preferences.

The institutionalization of criminology as a fit subject for study in British universities was profoundly influenced and advanced in the mid-twentieth century by a number of key European émigré intellectuals. Hermann Mannheim established the presence of criminology at the London School of Economics, the single most vital site for social science research in the UK for much of the century, and one with a close set of associations with influential policy communities especially in the Labor and Liberal parties. Mannheim's scholarly work, including the brilliant but now little-consulted *Dilemma of Penal Reform* (1939) and the prescient *Comparative Criminology* (1965) were among the most substantial, and certainly most sociologically informed, of the period. Meanwhile in Cambridge, Leon Radzinowicz set about developing a permanent institutional home for criminology in the unlikely setting of the Law Faculty of Cambridge University. In an autobiography published as late as 1999, Radzinowicz recalls his attraction to English political culture as he encountered it in and after the 1930s. It was, in his eyes, "down to earth, practical, observant, critical" (Radzinowicz, 1999: p. 376). 'England,' he asserts, "made me a realist and a pragmatist" (see further Sparks, 1999: p. 454).

The influence of Radzinowicz on the subsequent development of the subject was distinctive and substantial but nowadays culturally remote. Radzinowicz set out to create an academic field in the image of the virtues that he attributed to British institutions – thoroughness, decency, probity, and a lack of ideological enthusiasm. The formation in 1960 of the Institute of Criminology, with significant and sustained backing from the Home Office, was a watershed in the institutionalization of British criminology. It marks a transition from amateur enthusiasm and 'cottage industry' (Downes, 1988) to academic enterprise. What might seem more unfamiliar to observers beyond the UK (and frequently lamented nowadays by many within it) is the extent to which during the same period the internal research branch of government, the Home Office Research and Planning Unit, also became a major producer of criminological research (Walker, 1988; Newburn, 2011).

Yet even if the 'pragmatic frame of reference' seen as characterizing these institutions (Cohen, 1988: p. 39) was initially dominant it did not take long for it to receive certain significant shocks and challenges. Indeed, criminology rapidly became a focal site for the political and intellectual debates and conflicts that gripped British society in the 1960s. A new generation of social scientists, embarking on academic careers in newly founded universities, engaged with the political and cultural tumult of that decade, and freed from the direct dependency of bodies like the Cambridge Institute on short-term government research funding, brought new objects of inquiry to the fore: deviance, social control, political conflict and state violence, police power, ideology, and the manufacture of news, not to mention the reflexive and intentionally subversive reexamination of the claims and effects of criminology itself.

New theoretical perspectives and concepts took center stage: appreciation, labeling, moral panics, symbolic interactionism, Marxism, and feminism. New institutions were assembled to provide a platform for these perspectives and to build alliances with radical social workers and lawyers, students, prisoners' organizations and campaigns against state violence. Foremost among these were the National Deviancy Conference (1968–79) and the European Group for the Study of Deviance and Social Control (founded in 1973 and still in existence). The radical movement sought to extend criminology's intellectual range and reach, its theoretical ambitions and disciplinary frame of reference as well as adding hugely to its scholarly productivity. For example, in an emblematic contribution *The New Criminology* (Taylor et al., 1973) set out a bold and influential prospectus for a 'fully social' theory of

deviance. The new generation also promoted a more cosmopolitan approach, forming connections with European abolitionists, antipsychiatrists, and a range of other activists and social movements. The small, somewhat insular, rather atheoretical world of 'mainstream' British criminology had been subverted almost as soon as it had been established (see further Loader and Sparks, 2012).

The resulting ironies and fissures are never far to seek in the subsequent development of criminology as a 'discipline.' As the most acute of its radical observer-practitioners notes: "The more successful our attack on the old regime, the more we received PhDs, tenure, publishers' contracts and research funds, appeared on booklists and examination questions, and even became directors of institutes of criminology" (Cohen, 1988: p. 8). Thus, in one of the places in the world where criminology has expanded most rapidly, allegations of its failure, or betrayal or irrelevance are most persistently debated (Loader and Sparks, 2010: Chapter 1). There is a great deal more that might be said – not least that Britain is not now, and has really never been, a single place; and that the criminological cultures of Scotland, Wales, and Northern Ireland merit attention that space here forbids (see, e.g., McAra, 2008).

Criminology in Britain appears to have become one of those academic formations whose 'health' thrives upon what are often negative developments in the institutions it studies – their politicization, the repeated shocks to their legitimacy. It exhibits a permanent duality, as both a critical and a governmental discourse (Garland and Sparks, 2000; Garland, 2011), and to invite takeover bids and denunciations from partisans on either side. Its relationship to government and public policy remains integral, yet unlikely ever to recapture the intimacy, or easy alignment of worldviews, that characterized its birth.

Comparative Notes and Conclusions

Examining the history of criminology in four countries or regions shows common patterns and differences. Common trends during the twentieth century include shifts away from biologically and psychiatrically toward sociologically inspired ways of thinking and conducting research on issues of crime and crime control. They also include an increasing institutionalization of the field, with its own departments, degree programs, journals, and funding sources.

Despite such secular commonalities, differences in the course of these developments and in the current state of criminology are substantial. A social scientifically oriented criminology emerged first and most prominently in the US, reflective of a weak central state in times of rapid societal transformations and massive social problems (Hamilton and Sutton, 1989). Its emergence in Europe coincides only with the end of World War II, and its establishment in Latin America lags even more. The country/region accounts speak to potential explanations. The political context of an academic field with applied leanings obviously matters. A sociologically oriented criminology had chances in Latin America only with the demise of military dictatorships. In Scandinavia, criminology with its centralized government funding takes on a particular character, with specific methodological and thematic foci to reflect concerns of the Scandinavian welfare states. Also in the UK government funding contributed to the establishment of major research initiatives and institutions – even if an academic counterculture soon reacted to this alliance between government and the academy.

We see further that academic fields do not exist in national isolation. The stories about Scandinavia and Latin America are stories of a movement toward globalized criminology, to some degree an example of 'Americanization.' In the UK, given its close historical and linguistic associations with the US, there have also been profound influences, but these have been mediated by markedly different origins, and sharply distinct institutional and political conditions. In short, globalization leaves different traces in different national contexts. Latin American criminology appears to supplement the 'American' type of sociological criminology with a critical criminology that was inspired by translations of Italian and British writings with neo-Marxist leanings. Counter-hegemonic movements in Latin America and the massive tensions and inequalities in its countries appear to inspire this trend within academia. In Scandinavia, American criminology was incorporated in the welfare state concerns of the Nordic countries.

In short, a comparative study of the history of criminologies across countries and regions shows common trends *and* remarkable distinctions. Economic-political contexts, the nature of academic institutions and of the linkage between political-administrative systems and academia contribute to shaping the specific character criminology takes in each country and region. Globalization is at work, partly under US guidance, but academic exports are challenged and adapted to nation-level political, ideological, and institutional contexts.

See also: Anomie; Class: Social; Community Sociology; Control, Social; Counterculture: The Classical View; Merton, Robert K (1910–2003); Norms; Police: A Sociology of Knowledge Approach; Prison Life, Sociology of; Social Justice: Historical and Theoretical Considerations; State, Sociology of The; Subculture, Sociology of; Underclass; Urban Violence.

Bibliography

Anitua, Gabriel I., 2005. Historias de los Pensamientos Criminológicos. Editores del Puerto, Buenos Aires.

Camic, Charles, 1995. "Three departments in search of a discipline." Localism and interdisciplinary interaction in American sociology, 1890–1940. Social Research 62, 1003–1033.

Cohen, Albert K., Lindesmith, Alfred, Schuessler, Karl (Eds.), 1956. The Sutherland Papers. Indiana University Press, Bloomington.

Cohen, Stanley, 1988. Against Criminology. Transaction, New Brunswick, NJ.

Collins, Randall, 1998. The Sociology of Philosophies: A Global Theory of Intellectual Change. Belknap/Harvard, Cambridge, MA.

Debuyst, Christian, 1998. L'ecole francaise dite du 'millieu sociale'. In: Debuyst, Christian, Digneffe, Francois, Pires, Alvaro (Eds.), Histoire des savoirs sur le crime et la peine. DeBoeck, Paris/Bruxelles, pp. 301–356.

Del Olmo, Rosa, 1981. América Latina y su criminología. Siglo XXI, México.

Del Olmo, Rosa, 1999. The development of criminology in Latin America. Social Justice 26 (2), 19–45.

Downes, David, 1988. The sociology of crime and social control in Britain, 1960–87. British Journal of Criminology 28 (2), 45–57.

Garland, David, 1985. Punishment and Welfare: A History of Penal Strategies. Gower, Aldershot.
Garland, David, 1988. British criminology before 1935. British Journal of Criminology 28 (2), 1–17.
Garland, David, 2011. Criminology's Place in the Academic Field. In: Bosworth, M., Hoyle, C. (Eds.), What Is Criminology? Oxford University Press, Oxford.
Garland, David, Sparks, Richard, 2000. Criminology, social theory and the challenge of our times. In: Garland, David, Sparks, Richard (Eds.), Criminology and Social Theory. Oxford University Press, Oxford.
Hagan, John, Rymond-Richmond, Wenona, 2009. Darfur and the Crime of Genocide. Cambridge University Press, Cambridge.
Hamilton, Gary G., Sutton, John, 1989. The problem of control in the weak states: domination in the United States, 1880–1920. Theory and Society 18, 1–46.
Haney, D.P., 2009. The Americanization of Social Science. Intellectuals and Public Responsibility in the Postwar United States. Temple University Press, Philadelphia.
Kivivuori, Janne, 2011. Discovery of Hidden Crime: Self-Report Delinquency Surveys in Criminal Policy Context. Oxford University Press, Oxford.
Kivivuori, Janne, Bernburg, J.G., 2011. Delinquency research in the Nordic countries. In: Tonry, M., Lappi-Seppälä, T. (Eds.), Crime and Justice in Scandinavia. Crime and Justice – A Review of Research. University of Chicago Press, Chicago, pp. 405–478.
Kyvsgaard, B., 2012. Criminology past and present: worries and achievements. Journal of Scandinavian Studies in Criminology and Crime Prevention 13 (Suppl. 1), 4–14.
Laub, John H., Sampson, Robert J., 1991. The Sutherland-Glueck debate: on the sociology of criminological knowledge. American Journal of Sociology 96, 1402–1440.
Lengermann, Patricia, Neibrugge, Gillian, 2007. Thrice told: narratives of sociology's relation to social work. In: Calhoun, Craig C. (Ed.), Sociology in America: A History. University of Chicago Press, Chicago, pp. 63–114.
Loader, Ian, Sparks, Richard, 2010. Public Criminology? Routledge, London.
Loader, Ian, Sparks, Richard, 2012. Situating criminology. In: Maguire, M., Morgan, R., Reiner, R. (Eds.), Oxford Handbook of Criminology, fifth ed. Oxford University Press, Oxford.
Mannheim, Herman, 1939. The Dilemma of Penal Reform. Allen and Unwin, London.
Mannheim, Herman, 1965. Comparative Criminology. Routledge and Kegan Paul, London.
McAra, Lesley, 2008. Crime, criminology and criminal justice in Scotland. European Journal of Criminology 5 (4), 481–504.
Merton, Robert K., 1938. Social structure and anomie. American Sociological Review 3, 672–682.
Morris, Albert, 1975. The American society of criminology: a history, 1941–1974. Criminology 13 (2), 123–167.
Mucchielli, Laurent, 1994a. Hérédité et millieu: le faux antagonisme franco-italien. In: Mucchielli, Laurent (Ed.), Histoire de la criminologie francaise. L'Harmattan, Paris, pp. 189–214.
Mucchielli, Laurent, 1994b. Naissance et decline de la sociologie criminelle (1890–1940). In: Mucchielli, Laurent (Ed.), Histoire de la criminologie francaise. L'Harmattan, Paris, pp. 287–312.
Newburn, Tim, 2011. Criminology and government: some reflections on recent developments in England. In: Bosworth, M., Hoyle, C. (Eds.), What Is Criminology? Oxford University Press, Oxford.
Radzinowicz, Leon, 1999. Adventures in Criminology. Routledge, London.
Rafter, Nicole, 1997. Creating Born Criminals. University of Illinois Press, Champaign, IL.
Rice, Stuart A. (Ed.), 1931. Methods in Social Science. University of Chicago Press, Chicago.
Rothman, David J., 1971. The Discovery of the Asylum: Social Order and Disorder in the New Republic. Little, Brown and Company, Boston.
Sampson, Robert J., 2012. Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago Press, Chicago.
Savelsberg, Joachim J., Flood, Sarah M., 2011. American criminology meets Collins: global theory of intellectual change and a policy-oriented field. Sociological Forum 26, 21–44.
Savelsberg, Joachim J., King, Ryan D., 2011. American Memories: Atrocities and the Law. Russell Sage Foundation, New York.
Savelsberg, Joachim J., Cleveland, Lara L., King, Ryan D., 2004. Institutional environments and scholarly work: American criminology, 1951–1993. Social Forces 82 (4), 1275–1302.
Shaw, Clifford R., Zorbaugh, Frederick M., McKay, Henry D., Cottrell, Leonard S., 1929. Delinquency Areas. University of Chicago Press, Chicago.
Short Jr., James F., Hughes, Lorine A., 2007. Criminology, criminologists, and the sociological enterprise. In: Calhoun, Craig (Ed.), Sociology in America: A History. The University of Chicago Press, Chicago, pp. 605–638.
Sozzo, Máximo, 2006. "Traduttore Traditore". Traducción, importación cultural e historia del presente de la criminología en América Latina. In: Sozzo, Máximo (Coordinador), Reconstruyendo las Criminologías Críticas. Ad-Hoc, Buenos Aires.
Sozzo, Maximo, Salvatore, Ricardo, 2009. Criminología moderna en Estados Unidos y América Latina (1880–1940). In: Dammert, Lucia (Ed.), Crimen e Inseguridad. Políticas, temas y problemas en las Americas. Catalonia-FLACSO Chile, Santiago de Chile.
Sparks, Richard, 1999. Review of L. Radzinowicz, adventures in criminology. British Journal of Criminology 39 (3), 452–455.
Taylor, Ian, Walton, Paul, Young, Jock, 1973. The New Criminology. Routledge, London.
Walker, Robert, 1988. Perhaps minister: the messy world of in-house social research. In: Bulmer, M. (Ed.), Social Science Research and Government: Comparative Essays on Britain and the United States. Cambridge University Press, Cambridge.
Wetzell, Richard F., 2000. Inventing the Criminal: A History of German Criminology, 1880–1945. University of North Carolina Press, Chapel Hill.

Critical Criminology

Bruce A Arrigo and Heather Y Bersot, University of North Carolina at Charlotte, Charlotte, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Central to critical criminology's reformist agenda is the status of the self in society and the structural forces that undermine or thwart human capital. This article historicizes and culturalizes these concerns, mindful of the global political economy. Four specific aspects of the critique include (1) the problem of reification; (2) the risk, captivity, and harm thesis; (3) the criminology of the shadow; and (4) the criminology of the stranger. These issues are discussed in relation to critical criminology's evolving commitment to humanism, citizenship, and social justice.

Introduction: New Developments and Emerging Trends

While traditional criminology seeks to empirically account for the phenomena of crime and deviance, its critical counterpart endeavors to systematically explain the structural problems that plague veritable existence. Thus, at its philosophical core, critical criminology intends to identify and expose extant sociocultural conditions and arrangements that undermine or thwart the potential forces of human capital. This potential extends from the aesthetical (i.e., the realm of consumerism) to the epistemological (i.e., the realm of politics), from the ethical (i.e., the realm of technology) to the ontological (i.e., the realm of culture). This article describes the interdependent and interconnected flows, intensities, and assemblages that situate or define human capital (i.e., possibilities in being/becoming), and from which this unrealized in-and-of-society self makes choice and undertakes action. In addition, the article discusses the present-day challenge of critical criminology to critique and assess the human agency – social structure dynamics in a globalized and informationalized political economy. This is the challenge of 'ultramodern' society.

The Problem of Reification

At the core of critical commentary is the notion of reification. According to Hungarian philosopher Georg Lukacs, reification occurs when "the subject and social relations become object-like, thing-like, abstract, and commodity-like" (Baldwin 2009: p. 381; see also Lukacs, 1971). This "thingification" (Baldwin, 2009: p. 381) process unfolds when humans assign objectlike qualities to their constructions and ideologies, which are then perceived as "facts of nature, [the] result of cosmic laws, or manifestations of divine will" (Berger and Luckmann, 1966: p. 33). In other words, reification fosters a state in which the distinction between the social person and objects outside the self are, in essence, no longer discernible. This section delineates the three historical and cultural shifts that have abstracted the social person's humanness. These abstractions constitute inversions in being and becoming; harms to human capital.

Marx and Reification

German philosopher Karl Marx's economic theory concerning class consciousness and the "abstraction of the worker" (Arrigo, 2006: p. 45) principally defined the first shift. In the political economy of capitalism, he argued that an economic formula of this kind required the perpetual circulation of money (i.e., capital). However, as Marx reasoned, the accumulation of wealth generated by labor mostly served the interests of bourgeoisie or capitalist interests (i.e., acquiring through production) rather than proletariat or laborer needs (i.e., being self and other sufficient through production). The dialectical flows and intensities that followed from such conflicting worldviews fueled collectivist struggle to overcome the harm-sustaining acquisitive conditions of production (see e.g., Lynch and Michalowski, 2006). Indeed, as Marx (1990) asserted, in the capitalist mode of production, the intrinsic use-value of one's labor is assigned a fixed, although artificial, exchange (i.e., monetary) value capable of being purchased in the marketplace. In this way, the worker and the product of one's labor are, in essence, rendered a commodity. Marx (1990) described this concept as follows:

> Not an atom of matter enters into the objectivity of commodities as values; in this it is the direct opposite of the coarsely sensuous objectivity of commodities as physical objects. We may twist and turn a single commodity as we wish; it remains impossible to grasp it as a thing possessing value. However, let us remember that commodities possess an objective character as values only in so far as they are all expressions of an identical social substance, human labour, that their objective character as values is therefore purely social. From this it follows self-evidently that it can only appear in the social relation between commodity and commodity. (pp. 138–139)

Made abstract, Marx (1995) posited that the worker "sinks to the level of a commodity and becomes indeed the most wretched of commodities" (p. 70). Accordingly, qualities inherent to the worker/social person's humanness, such as love, virtue, and conscience, "enter into commerce" (Marx, 1995: p. 36), altering the nature of the human relationship

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45053-1

from one of authentic interaction to one of mere transaction. As Fromm (1994) noted,

> It becomes [the individual's] fate to contribute to the growth of the economic system, to amass capital, not for the purpose of [one's] own happiness or salvation, but as an end in itself. [The individual thus becomes] a cog in the vast economic machine – an important one if he [or she] has capital, an insignificant one if he [or she] has none – but always a cog to serve a purpose outside of the [self]. (p. 110)

In the political economy of capitalism, then, the notion of *having* rather than *being/becoming* (or acquiring rather than creating) dangerously and perniciously prevails, thereby disrupting intrinsic social bonds and alienating possibility/potential for collectivist good. Critical commentators have contended that this state of captivity through consumption (see e.g., Chambliss and Seidman, 1971; Quinney, 1974) – which Marx (1990) described as commodity fetishism – extends from "the kept (workers) to their keepers (factory owners), from their managers (institutional decision-brokers) to their watchers (state government bureaucrats)" (Arrigo and Milovanovic, 2009; see also e.g., Rusche and Kirchheimer, 1968; Pashukanis, 2002).

The Situationists and Reification

Drawing upon Marx's commodity concept, a group of predominately European activists known as the Situationists responded to what they viewed as the second major shift or transition concerning the problem of reification. As information technology advanced under late capitalism, these social and cultural theorists sought to challenge the ways in which the mass media manufactured imitative realities (e.g., by way of government propaganda) and stylized consumerism (e.g., by way of crass commercialism). The conspicuous presence of these political economic conditions sustained the decline of being/becoming into having, and transformed the ontological status of having into vacuous appearing (Debord, 1983). In response to this state of affairs, the Situationists mobilized, organized, and promoted the notion of constructing situations, or environments, to counter what they took to be the increasing degradation or abstraction of the social person (Debord, 1983; Lukacs, 1971).

Guy Debord (1983), a leader in the Situationist movement, developed the concept of the spectacle to explain the way in which late capitalism engendered an insidious form of hegemonic repression and alienation. In his compilation of theses, entitled *The Society of the Spectacle*, Debord (1983) contended that "[i]n societies dominated by modern conditions of production, life is presented as an immense accumulation of spectacles. Everything that was directly lived has receded into a representation" (thesis 1). Thus, following Debord, the social person's humanness is further reified by way of images that depict only fabricated and fictionalized constructions (i.e., spectacles) of meaningful life. When these contrived representations are culturalized (i.e., circulated for mass consumer consumption), they become the image-crafted commodities of a consumer society driven insatiably by illusions. This, then, is the "moment when the commodity [as appearance] attains the total consumption of social life" (Debord, 1983: p. 42). Under these conditions, matter and materiality are supplanted by replicas and façades. Indeed, the self in society as human capital is further abstracted through the iterative reproduction of liquid identities and simulated realities that caricaturize, sensationalize, and derealize visceral (i.e., embodied) existence (Bauman, 2000).

The Hyper-Realists and Reification

The third historical and cultural shift was marked by the hyperrealists. Following the Situationists' notion of the spectacle (appearing over being/becoming), these thinkers chiefly concerned themselves with discerning whether contrived representations (i.e., images of reality) were more real than the reality on which they (these representations) were based. In response, the hyperrealists proposed a novel social theory that "elevated the commodification of illusion to a novel and disturbing (alienating) level of abstraction" (Arrigo and Bersot, 2013).

One of the preeminent hyperrealists, French philosopher Jean Baudrillard, asserted that postmodern societies are not organized around production and consumption (i.e., having) as Marx maintained. Rather, Baudrillard (1968, 1972) argued that social life in the new reification era was mediated by simulation. This was an era that no longer made sustainable the replication of appearance over being/becoming as a cultural artifact. In its place, simulation (i.e., the replication and proliferation of image) made coded meanings (systems of communication) the essential artifacts of culture. Indeed, in a digitized information age, reality (or the appearance of it) becomes irrelevant (Baudrillard, 1983a, 1983b). The ascendancy of the code's status (and with it the ungrounding of existence and its manifold illusions), led Baudrillard (1993) to conclude that social life as we knew it had ended. As he proclaimed,

> The end of labor. The end of production. The end of political economy. The end of the signifier/signified dialectic which facilitates the accumulation of knowledge and of meaning, the linear syntagma of cumulative discourse. And at the same time, the end simultaneously of the exchange value/use value dialectic which is the only thing that makes accumulation and social production possible. The end of linear dimensions of discourse. The end of the linear dimension of the commodity. The end of the classical era of the sign. The end of the era of production. (p. 8)

As such, Baudrillard (1983a) suggested that within postmodern culture, significance, or value, could only be expressed through sign-exchange-value. In this new epoch, commodity forms "circulate in the marketplace of signs, anchored, although temporarily, in the dominant sign meanings assigned to them in a particular political economy" (Arrigo, 2006: p. 48). Consequently, as several critical criminologists and cultural critics have warned, this consumption, hypervigilantly reproduced and panoptically (Foucault, 1977) and synoptically (Mathiesen, 1997) disseminated, signifies nothing real other than the fictionalized fear that it feeds for a simulated society of commodified captives (Arrigo and Bersot, 2013; Delanda,

2006; Dyer-Witheford, 1999). Thus, in the hyperreal society, "there is no longer a real to be recovered behind the illusion [and, thus,] there is no illusion either" (Best, 1989: p. 37). Given these conditions of reification, human capital's potential is obliterated.

The Risk–Captivity–Harm Thesis

Recent developments in critical criminology point to reification and how this problem is related to a larger cultural crisis (Arrigo, 2010, 2011, 2012). Specifically, commentators assert that what is reified is (1) risk or excessive investments in fearmongering that manages difference or human capital as sameness through reductive/repress mechanisms and techniques; (2) captivity or the confinement of individual and collectivist consciousness that follows from nurturing or sustaining such mundane risk management practices; and (3) harm or the recursive restrictions that limit possibilities in being and/or deny potentials in becoming for and about one and all when difference is reduced to and repressed by serialized codes. These reifications constitute the cultural crisis of the present ultramodern era. In this section, risk, captivity, and harm and the unreflective maintenance of the thesis are discussed.

Contemporary critical thinkers contend that what is cultivated within a culture of consumption is risk-minded systemic pathology (Fromm, 1994) that silences dissent and suppresses difference. Indeed, current political economic conditions worldwide are indicative of what Beck (1992; see also 2009) termed the 'risk society.' The resulting social dis-ease fosters negative freedom (freedom from rather than freedom to) (Fromm, 1994), and is perhaps made most disturbingly apparent in the industry of crime. The rapid development of penal institutions and their exploitation of inmate labor, the trade in diversion courts, and the vast expansion of surveillance policing are deemed "healthy, natural and inevitable threat-avoidance progress" (Arrigo and Bersot, 2013). Although such measures promote economic growth consistent with capitalism in the ultramodern era, some critical scholars posit that the industry of crime normalizes violence by way of engendering a palpable sense of abject fear among the citizenry (Simon, 2009). This is fear with transnational harm-intensifying human consequences (Bauman, 1998).

To illustrate, consider the dramatic surge in surveillance technology. These monitoring mechanisms and techniques – defined by some as emblematic of the "prison-release-reincarnation abstract machine" (Arrigo, 2012: p. 441; see also Arrigo and Milovanovic, 2009; Deleuze and Guattari, 1987) – include radio frequency identification (e.g., to supervise registered sex offenders within the community) and global positioning systems (e.g., to locate diverted juveniles) (Bowling et al., 2009; Grabosky, 1998). While the expressed aim of these measures may be to protect the public, "risk" is "connected to future offenders, not victims" (Schinkel, 2011: p. 370). Thus, they (these mechanisms and techniques of surveillance) are deployed to manage potential crime rather than respond to actual crime.

Derivative in its consumerism and digital in its politics, this machine serializes (i.e., makes abstract, deontologizes) the experiences of crime and justice, law and order, violence and victimization by semiotically reducing and repressing objects, events, and individuals to nothing more than informational (i.e., coded) utility. Biopolitical tattooing, social sorting, and prepression are all measures that have been created and/or reintroduced as means to manage risk – eliminating difference and imposing sameness within society. Biopolitical tattooing involves inspecting and monitoring the very body of an individual by way of subcutaneous tattoos and finger and retina prints (Agamben, 2004). Social sorting refers to the way in which (inter)national surveillance measures, particularly in the post-9/11 world, are "coded to categorize personal data such that people [digitally identified as potential threats]…may be treated differently" (Lyon, 2007: p. 162; see also 2003 and 2006). Prepression is a "combination of prevention and repression [that] entails the archiving of risky individuals (e.g., children who display problematic behavior) and their selection for 'early intervention'" (Schinkel, 2011: p. 365).

What emerges from these hyperreal meaning-making/generating risk practices is captivity. The harm of this captivity extends to and from the kept (i.e., those incarcerated, civilly committed, and otherwise subjected to inspection and monitoring within the community), their keepers (e.g., correctional staff and police officers), their regulators (e.g., policy makers, judges, and correctional administrators), and their watchers (i.e., the public) (Arrigo et al., 2011). Indeed, the quality of this captivity is one that confines individual and collectivist consciousness. Drawing upon Foucault's (1977) work on the disciplinary society, French philosopher Gilles Deleuze (1992) proposed the notion of the control society. He described this intellectual transition as follows:

> In the societies of control, on the other hand, what is important is no longer either a signature or a number, but a code: the code is a password, while on the other hand disciplinary societies are regulated by watchwords (as much as from the point of view of integration as from that of resistance). The numerical language of control is made of codes that mark access to information, or reject it. We no longer find ourselves dealing with the mass/individual pair. Individuals have become "dividuals," and masses, samples, data, markets, or "banks." […] The disciplinary [subject] was a discontinuous producer of energy, but the [subject or person] of control is undulatory, in orbit, in a continuous network. (Deleuze, 1992: p. 6)

Thus, the society of control reduces and represses human subjectivity by way of propagating fear about would-be offending/offenders through the numeric serialization of human capital. This fear induces a sense of complacency toward the ubiquity of surveillance technology and its disturbing risk-management flows, intensities, and assemblages (i.e., tagging, tracking, neutralizing, and eliminating would-be offenders). Although such technologies seemingly ameliorate collective anxieties about the deviant (e.g., juveniles whose violent tendencies belie their youthful innocence), the diseased (e.g., sex offenders who are incapable of being cured from their perverse proclivities), and the *potential* dangerousness of both constituencies, the 'informationalizing' that follows from the control society is normalized harm. The violence of this

informationalizing implicates – and captivates – one and all. By vanquishing difference (i.e., derealizing it through its serialized coding), the resulting harm territorializes freedom, citizenship, and social justice and their embodied human capital significations. Thus, greater possibilities in being and increased prospects for becoming are constrained or foreclosed.

The Criminology of the Shadow

Amidst the ultramodern crisis are troubling conditions of control that interdependently sustain the criminology of the shadow (Arrigo, 2008, 2012; Arrigo and Williams, 2009). Finding its origins in Plato's (2008) appearance/reality thesis, the shadow represents a critique of the in-and-of-society self (Arrigo, 2011, 2012). The criminological status to which this crisis refers is the social person's reification captured within and subjected to four forces of destabilizing influence. These interdependent, mutually supporting, and coproductive forces or control conditions consist of symbolic (aesthetic), linguistic (epistemological), material (ethical), and cultural (ontological) flows, intensities, and assemblages. The harms sustained by these influencing conditions extend to and from the kept (i.e., those confined), their keepers (i.e., those who confine), their managers (i.e., those who administrate confinement), and their watchers (i.e., those who observe confinement) (Arrigo et al., 2011; Arrigo and Milovanovic, 2009). The section below delineates the conditions of control and the violence that is normalized (i.e., totalizing madness that is made healthy and normative) by way of these captivating forces.

The Symbolic Realm and the Ultramodern Shadow

The symbolic realm consists of sign-exchange-value. Sign-exchange-value refers to the images, or symbolic constructions, that are mobilized and activated in the unconscious. Although always fragmented or incomplete, these images (i.e., pictures in our minds) mediate prevailing articulations (i.e., master discourses) (Lacan, 1977, 1981) for and about the same. This includes mental constructions for and about offenders and victims; institutional and community-based treatment for these citizen groups or from these governmental entities; and agency, responsibility, and choice for all stakeholders. Thus, for example, what images are conjured in one's mind (dominate individual and collectivist consciousness) when it comes to juvenile delinquents, mentally ill individuals, and sexually violent predators? If the imagined construction tells the story of delinquency, disease, and dangerousness (i.e., the social person is defined only as the 'lack' in discourse), then much is missing from such psychic configurations. This is the self in society as mere shadow in mental representation. Conversely, image-crafting the social person as delinquent, diseased, and dangerous human potential involves the production and circulation of speech that, although incomplete, affirms this yet-to-be-fully envisioned subjectivity (Lacan, 1977, 1981). What is hidden through the ascendancy of the master discourse is a "struggle whose aesthetic intensity (about human agency) and force (concerning structural meaning) dangerously (i.e., panoptically and synoptically) linger as desperation and fear respectively" (Arrigo and Bersot, 2013). Thus, the struggle in being/becoming entails recognizing and transforming circumscribed, although dominant, symbolic constructions that are consumed by way of sign-exchange-value. These images incompletely symbolize human capital's potential.

The Linguistic Realm and the Ultramodern Shadow

The linguistic realm consists of sign-exchange-value emanating from the unconscious that is translated into speech or the written word. Because the symbolic realm consists of inadequate and partial representations, the text that emerges from this imaging is likewise unfinished and insufficient. Indeed, the unconscious narrative that flows from these symbolizations "communicates less than (produces less about) who we are or could become" (Arrigo and Bersot, 2013). What is produced through this semiotic process is the mass-marketing of fictions about human capital manifested by way of the 'presencing' (privileging) of selected knowledge. "As narrative, [this preferred text] enters the marketplace of sign-exchange-value and recounts the struggle of the [social person] in its own governing voice. This voice tells of desperation and fear through the epistemology of logocentrism and the metaphysics of presence" (Arrigo and Bersot, 2013; see also Derrida, 1973, 1977, 1978). The concept of logocentrism suggests that what is most readily accessible to consciousness is that which is fundamentally defined as good and virtuous. Thus, the values attached to these discernible and manifest notions are rendered true and absolute. However, under these textual conditions, the narrative that is produced and reproduced is one that tends toward finalization (Bakhtin, 1982). This is the self in society as mere shadow in words. When these fragmented renditions of truth, progress, science, and human potential are written and rewritten, the possibilities for articulating alternative narratives are forestalled or, worst, excluded (Deleuze, 1983). As such, the texts for and about the social person's emergent possibilities in being and becoming await emancipation and inscription.

The Material Realm and the Ultramodern Shadow

The ethos of the material realm entails inhabiting the favored texts of human potential – and all their incompletenesses and partialities – as lived history, as embodied truth. These truths materialize as systems of thought and bodies of knowledge (e.g., in psychiatry, law, penology, social work, education, and the like) that domesticate and discipline the social person's humanness. This domestication and disciplining is power materialized corporeally. Biopower relies on techniques and mechanisms that depathologize (Foucault, 1965, 1977) and territorialize (Deleuze, 1983) being/becoming. Under these material conditions, "difference is vanquished…, identity is homogenized, and community is sanitized" (Arrigo et al., 2011: p. 165; Arrigo, 2010). This is the self in society as mere shadow in its embodiments.

Indeed, as Fromm (1994) cautioned when diagnosing the status of the captured social person,

> The individual ceases to be [and] adopts entirely the kind of personality offered to him [or her] by the cultural patterns; and [the individual] therefore becomes exactly as all others are and as they expect [the individual] to be. The discrepancy between "I" and the world disappears and with it the conscious fear of aloneness and powerlessness…[t]he person who gives up his [or her] individual self becomes an automaton, identical with millions of other automatons…[and] need not feel alone or anxious anymore. But the price [the individual] pays, however, is high; it is the loss of the self. (p. 184)

Under these material conditions, then, subjects are rendered docile bodies amenable to the will of the state (Foucault, 1977: p. 210; see also Arrigo et al., 2005). What is normalized by way of this domestication and disciplining is violence corporealized (i.e., the self and other habitations of captivity's shadow).

The Cultural Realm and the Ultramodern Shadow

The cultural realm encompasses the sphere of replication and dissemination – in image, text, and bodies of knowledge. Given the partial portraits, incomplete texts, and disembodied systems of knowledge that govern human capital, caricatured representations of each prevail in the ultramodern age. This is the era of digital interconnectivity and hyperreal information-only interdependency. This is the condition of carnival consumption (Carrabine, 2008), fast capitalism (Agger, 1989), and liquid identities (Bauman, 2000) in which the virtual rather than the visceral, the sensational over the sensual is ontologized (Arrigo et al., 2011). What is absent in this culturalizing of the social person is human potential as being/becoming – as yet-to-be imagined, spoken, scripted, and inhabited productivity. What remains, then, is the self in and of society as mere shadow in its cultural artifacts and political economic abstractions. This is human potential restricted in images, texts, and systems of thought that are simulated derivatively serially, and ubiquitously.

When these collective conditions of control are maintained and nurtured as healthy and normative pathways to progress, the all-encompassing consumption of the criminological shadow is made more immanent. Currently, the intersection of image, language, embodied subjectivity, and their cultural reproductions reify fragments and fictions of human capital. Sustaining these conditions is madness. Succumbing uncritically to these limits on and denials of being/becoming is totalizing madness (Arrigo et al., 2011). Recognizing and overcoming the human capital harm of this captivity is a journey in the making; it is a revolution for people yet to come (Deleuze and Guattari, 1987).

The Criminology of the Stranger

As described previously, the coproductive, mutually supporting, and interdependent flows, intensities, and assemblages of ultramodernity exert symbolic, linguistic, and material force on the self in society. The current cultural artifacts and political economic abstractions that emerge from these exertions sustain harm-intensifying/generating conditions of control. These influencing conditions reduce possibilities in being and repress prospects for becoming. Consequently, the quality of citizenship and the character of social justice are less than what they could be for, by and about one and all. The ultramodern journey that awaits development is the exploration, cultivation, and occupation of novel and derelict flows, intensities, and assemblages. This is habitus – the horizon of perception, thought, and action (Bourdieu, 1977) – as experimental, creative, and dynamic. When habitus is envisioned, spoken of, and practiced as such, it provides the positional, relational, and provisional grounding from within which human capital's potential as productive change (as freedom) can emerge. This undertaking, then, is a mutating and unfamiliar revolution for people yet to be (Arrigo and Milovanovic, 2009).

The counterpoint to the criminological shadow is the criminological stranger (Arrigo and Milovanovic, 2009). The stranger seeks to liberate the desiring subject (i.e., the social person in captivity) from harm-intensifying/generating abstractions. The stranger's evolving flows, intensities, and assemblages offer a way out of the ultramodern crisis of reified madness that the criminological shadow fetishizes and normalizes. This 'way out' includes departures in theory (the consumerism of critical mindfulness), method (the politics of difference), and praxis (the corporealizations of excellence). These exits are briefly discussed, especially in relation to their capacity to reconceive critical criminology in ways that hold the nearest promise to unleash and harness dormant human capital.

As a departure in theory, the journey entails embracing a cosmopolitan imagination (Delanty, 2009). This is image, language, and embodied subjectivity reconceived within the jurisprudence, the landscape, of the mind. This cosmopolitanism symbolizes (image crafts) 'offenders,' 'victims,' 'crime,' 'punishment,' 'recovery,' 'human agency,' 'social structure,' etc. differently (e.g., doing drug treatment and recovery another way; doing victim offender mediation another way; doing community policing another way). These are mental constructions that are resistant to closure, to the finalization of being, and to the criminological shadow's reification. To access this "reservoir of desire" (Arrigo and Bersot, 2013), the master discourses that govern consciousness must be excavated and deterritorialized (Deleuze and Guattari, 1987) for their harm-producing tendencies. These tendencies include the normalization of violence that renders the social person's difference (one's human potential) as lacking in consciousness. Deterritorialization is genealogical in its operation. It involves mining and uprooting the unconscious, its topography, and the Oedipal and capital logics that the psychic apparatus both privileges and endorses (Arrigo and Milovanovic, 2009). Critical criminology's challenge by way of cosmopolitanism, then, is to disassemble the hyperreal shadows of difference that criminology reproduces, and to reterritorialize (i.e., reassemble, reconstruct) criminology's serialized terrain in ways that harness will to power (Nietzsche, 1968). Power refers to desire (latent being/becoming) mobilized in the service of unleashing human capital's potential. The consumerism of critical mindfulness awaits additional theoretical development and amplification.

As a methodological exit, the criminology of the stranger entails movement outside of consciousness to the texts that narrate or tell the story of law and order, freedom and punishment, offenders and victims, treatment and recovery, self and society, etc. This movement – situated within the flows of deterritorialization and reterritorialization – entails journeying as a being toward difference, tending ever more so toward dynamic and transformative (rather than unchanged and uniformed) descriptions of human capital. These are the narratives of how the social person (e.g., as a recovering drug abuser, exincarcerate, and mental health consumer) makes overcoming possible because of, rather than despite, one's status as addict, convict, or patient. The challenge of critical criminology is to speak of, write about, and circulate these creative, provocative, and experimental stories of overcoming through difference. These narratives constitute a 'strange' humanness whose inscriptions tell of human capital's uncharted possibilities in being and unexplored prospects for becoming. This difference implicates one and all in an era that reifies hyperreality; derealizes risk, captivity, and harm; and serially codes the criminological shadow. The 'politics' of difference as method awaits further analytical refinement.

As a venture in praxis, the criminological stranger seeks to produce different lived histories about crime making and peace building, citizenship and community, and agency and structure. This is embodied subjectivity steeped in replacement systems of thought and alternative bodies of knowledge that manufacture mutating habits of excellence (Aristotle, 1976). This excellence or flourishing is human capital corporealized differently, as being/becoming virtuous. What is corporealized is human capital's dynamic potential lived ever more courageously, compassionately, responsibly, generously, etc. when studying crime; making peace; doing citizenship; residing within neighborhoods; and inhabiting class, race, and gender categories or their intersectional configurations. These embodiments transform the social person into a subject in-process (i.e., embodying one's difference another way). Critical criminology's challenge, then, is to occupy (territorialize) absent bodies of knowledge and heterodox systems of thought through mutating habits of excellence. This is the evolving journey of becoming ethically revolutionary. The proposed revolution dignifies, honors, and affirms all those within the society of captives and extols the potential forces that emancipate human capital. The corporealizations of excellence as praxis await supplementary and systematic delineation.

See also: Conspicuous Consumption and Distinction, History of; Marx, Karl (1818–83); Postmodernism: Philosophical Aspects; Self: Philosophical Aspects; Virtue Ethics.

Bibliography

Agamben, G., 10 January 2004. No to bio-political tattooing. Le Monde.

Agger, B., 1989. Fast Capitalism: a Critical Theory of Significance. University of Illinois Press, Urbana-Champagne, IL.

Aristotle, 1976. Ethics (J.A.K. Thomson, Trans.). Penguin, New York.

Arrigo, B.A., 2006. The ontology of crime: on the construction of the real, the image, and the hyper-real. In: Arrigo, B.A., Williams, C.R. (Eds.), Philosophy, Crime, and Criminology. University of Illinois Press, Urbana and Chicago, pp. 41–73.

Arrigo, B.A., 2008. Crime, justice, and the under-laborer: on the criminology of the shadow and the search for disciplinary legitimacy and identity. Justice Quarterly 25 (3), 439–468.

Arrigo, B.A., 2010. De/reconstructing critical psychological jurisprudence: strategies of resistance and struggles for justice. International Journal of Law in Context 6 (4), 363–396.

Arrigo, B.A., 2011. Madness, citizenship, and social justice: on the ethics of the shadow and the ultramodern. Law and Literature 23 (3), 404–441.

Arrigo, B.A., 2012. The ultramodern condition: on the phenomenology of the shadow as transgression. Human Studies: A Journal for Philosophy and the Social Sciences 35 (3), 429–445.

Arrigo, B.A., Bersot, H.Y., 2013. intellectual history and ultramodern developments. In: Arrigo, B.A., Bersot, H.Y. (Eds.), The radical philosophy of criminology culturalized, The Routledge Handbook of International Crime and Justice Studies. Taylor and Francis, UK.

Arrigo, B.A., Bersot, H.Y., Sellers, B.G., 2011. The Ethics of Total Confinement: A Critique of Madness, Citizenship, and Social Justice. Oxford University Press, New York.

Arrigo, B.A., Milovanovic, D., 2009. Revolution in Penology: Rethinking the Society of Captives. Rowman and Littlefield, Lanham, MD.

Arrigo, B.A., Milovanovic, D., Schehr, R.C., 2005. The French Connection in Criminology: Rediscovering Crime, Law, and Social Change. State University of New York Press, New York.

Arrigo, B.A., Williams, C.R., 2009. Existentialism and the criminology of the shadow. In: Lippens, R., Crewe, D. (Eds.), Existentialist Criminology. Routledge-Cavendish, UK, pp. 222–248.

Bakhtin, M., 1982. The Dialogic Imagination: Four Essays. University of Texas Press, Austin.

Baldwin, J., 2009. Exchange and subjectivity, commodity and gift. Semiotica 173 (1/4), 377–396.

Baudrillard, J., 1968. Le System des Objects. Denoel-Gonthier, Paris.

Baudrillard, J., 1972. For a Critique of the Political Economy of the Sign. Telos Press, St. Louis, MO.

Baudrillard, J., 1983a. In the Shadow of the Silent Majorities. Semiotiext(e), New York.

Baudrillard, J., 1983b. Simulations. Semiotext(e), New York.

Baudrillard, J., 1993. Symbolic Exchange and Death. Sage, London.

Bauman, Z., 1998. Globalization: The Human Consequences. Blackwell, Oxford.

Bauman, Z., 2000. Liquid Modernity. Polity Press, Cambridge.

Beck, U., 1992. Risk Society: Towards a New Modernity. Sage, London.

Beck, U., 2009. World at Risk. Polity Press, Cambridge, UK.

Berger, P., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Anchor/Doubleday, New York.

Best, S., 1989. The commodification of reality and the reality of commodification: Jean Baudrillard and post-modernism. Current Perspectives in Social Theory 9, 23–50.

Bowling, B., Marks, A., Murphy, C.C., 2009. Crime control technologies: towards an analytical framework and research agenda. In: Brownsword, R., Yeung, R. (Eds.), Regulating Technologies. Hart, Oxford, pp. 51–78.

Bourdieu, P., 1977. Outline of a Theory of Practice. Cambridge University Press, UK.

Carrabine, E., 2008. Crime, Culture, and the Media. Polity Press, UK.

Chambliss, W., Seidman, R., 1971. Law, Order and Power. Addison-Wesley, Boston.

Debord, G., 1983. Society of the Spectacle. Black and Red, Detroit.

Delanda, M., 2006. A New Philosophy of Society: Assemblage Theory and Social Complexity. Coninuum Books, New York.

Delanty, G., 2009. The Cosmopolitan Imagination: The Renewal of Critical Social Theory. Cambridge University Press, New York.

Deleuze, G., 1983. Nietzsche and Philosophy. Columbia University Press, New York.

Deleuze, G., 1992. Postscript on the society of control. October 59 (Winter), 3–7.

Deleuze, G., Guattari, F., 1987. A Thousand Plateaus: Capitalism and Schizophrenia. University of Minnesota Press, Minneapolis.

Derrida, J., 1973. Speech and Other Phenomena. Northwestern University Press, Evanston.

Derrida, J., 1977. Of Grammatology. Johns Hopkins University Press, Baltimore.

Derrida, J., 1978. Writing and Difference. University of Chicago Press, Chicago.

Dyer-Witheford, N., 1999. Cyber-Marx: Cycles and Circuits of Struggle in High Technology Capitalism. University of Illinois Press, Urbana-Chicago, IL.

Foucault, M., 1965. Madness and Civilization: A History of Insanity in the Age of Reason. Vintage, New York.

Foucault, M., 1977. Discipline and Punish: The Birth of a Prison. Pantheon, New York.

Fromm, E., 1994. Escape from Freedom. Henry Holt and Company, New York.
Grabosky, P., 1998. Technology & crime control. Trends & Issues in Crime and Criminal Justice 78, 1–6.
Lacan, J., 1977. Ecrits: A Selection. W. W. Norton, New York.
Lacan, J., 1981. The Four Fundamental Concepts of Psycho-Analysis. W.W. Norton, New York.
Lukacs, G., 1971. History of Class Consciousness. MIT Press, Cambridge.
Lynch, M.J., Michalowski, R.J., 2006. Primer in Radical Criminology: Critical Perspectives on Crime, Power, and Identity, fourth ed. Criminal Justice Press, New York.
Lyon, D., 2003. Surveillance as social sorting: computer codes and mobile bodies. In: Lyon, D. (Ed.), Surveillance as Social Sorting. Routledge, London, pp. 13–30.
Lyon, D., 2006. 911, Synopticon, scopophilia. In: Haggerty, K., Ericson, R. (Eds.), The New Politics of Surveillance and Visibility. University of Toronto Press, Toronto, pp. 199–224.
Lyon, D., 2007. Surveillance Studies: An Overview. Polity Press, Cambridge.
Marx, K., 1990. Capital, vol. I. Penguin Books, London.
Marx, K., 1995. The Poverty of Philosophy. Prometheus Books, New York.
Mathiesen, T., 1997. The viewer society: Michel Foucault's 'panopticon revisited'. Theoretical Criminology 1 (2), 215–234.
Nietzsche, F., 1968. In: Kaufmann, William (Ed.), The Will to Power. Vintage Books, New York.
Pashukanis, E., 2002. The General Theory of Law and Marxism. Transaction Books, New Brunswick, NJ.
Plato, 2008. Republic (R. Waterfield, Trans.). Oxford University Press, New York, NY.
Quinney, R., 1974. Critique of the Legal Order. Little Brown and Co, Boston.
Rusche, G., Kirchheimer, O., 1968. Punishment and Social Structure. Transaction Publishers, Totowa, NJ.
Schinkel, W., 2011. Prepression: the actuarial and new technologies of security. Theoretical Criminology 15 (4), 365–380.
Simon, J., 2009. Governing through Crime: How the War on Crime Transformed American Democracy and Created a Culture of Fear. Oxford University Press, New York.

Relevant Websites

http://www.critcrim.org – Critical Criminology Information and Resources.
http://theory.eserver.org – Cultural Studies and Critical Theory.
http://blogs.kent.ac.uk/culturalcriminology – Cultural Criminology.
http://www.uta.edu/huma/illuminations – Illuminations: The Critical Theory Website.
http://plato.stanford.edu/entries/marx/ – Karl Marx.

Critical Legal Studies

Robert W Gordon, Yale University Law School, New Haven, CT, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 5, pp. 2967–2972, © 2001, Elsevier Ltd.

Abstract

Critical legal studies is a label for movements of leftist legal academics in the USA, Britain, and France which reached their height of influence in the 1980s and 1990s. Critics target their critiques to the dominant modes of legal reasoning, scholarship, and teaching in their legal cultures. These modes usually pretend to be neutral and apolitical, but critics see them as expressions of conflicting political ideologies. Critics do not assume law is merely a superstructural reflection of a material base of determining economic relations, or as simply an instrument of ruling-class power, but treat the content and forms of law and legal reasoning as relatively autonomous active force fields in the structuration and interpretation of the economy and social world. Ideologies hidden in law often serve to produce, reproduce, and justify inequalities and illegitimate hierarchies; but they also express utopian aspirations and point to plausible alternative social arrangements. The critics' aims have been to bring into view the latent ideologies and contradictory structures of the legal order, and to capture and reorient enclaves of legal scholarship, teaching, and practice, and to use them to reimagine and incrementally put into practice alternative legal ideas and policies.

'Critical legal studies' (CLS) is an umbrella term for a variety of movements of legal intellectuals, most of them in law faculties in the USA, Britain, and France. The movements originated in the late 1970s and achieved their peak of influence in the 1980s and 1990s. Their adherents are generally on the political left, critical of existing distributions of wealth, power and status rewards and of patterns of domination and subordination by class, race, gender, and other 'illegitimate hierarchies.' Unlike classical Marxists, they do not treat law as a merely 'superstructural' reflection of a material' base' of determining economic relations, or as simply an instrument of ruling-class power; they take the content and forms of law and legal reasoning seriously, as relatively autonomous active force fields in the construction and interpretation of the social world. Critics (or 'crits') target their critiques to the dominant modes of legal reasoning, scholarship, and teaching in their legal cultures, which usually pretend to be neutral and apolitical, but which the critics see as expressions of conflicting political ideologies. The ideologies hidden in law often serve to produce, reproduce, and justify inequalities and illegitimate hierarchies, but they also express utopian aspirations and point to plausible alternative social arrangements. The critics' general aims have been to bring into view the hidden ideological biases and effects of the existing legal order, to capture and re-orient enclaves of legal scholarship, teaching, and practice, and to use them to re-imagine and to implement alternative legal ideas and policies.

Within these broad commonalities of outlook and concern, the CLS movements have tended to fragment into many different schools, informed by diverse intellectual antecedents and committed to divergent views of the role of law in social life and of the political ambitions of intellectual critique and reform.

The US movements receive most attention here, since they attracted the most members, produced the largest literature, and penetrated the furthest into centers of legal education; also, their types of scholarship and internal quarrels parallel those of CLS movements elsewhere. CLS-USA came into being in 1977 at a conference of law teachers rebelling against their own legal education and its seeming irrelevance to the contemporary ferment of social movements (anti-Vietnam War, civil rights, feminist, environmental, etc.). By the 1980s critical teachers made up a significant fraction (four or more faculty) at Buffalo, Harvard, Miami, Northeastern, Stanford, UCLA, and Wisconsin law schools and a token presence at many more, eventually producing students who went into law teaching themselves, a substantial bibliography (Kennedy and Klare, 1984; Bauman, 1996), and a series of national conferences.

Antecedents of CLS

To critique the orthodoxies of their legal culture, American critics in the first phase of their movement found resources in various European theories, especially Western neo-Marxist theories of law as 'ideology' and an instrument of 'hegemony' (Gramsci, 1971), humanist social histories of subordinated groups (Hay, 1975; Thompson, 1975), structuralism, and phenomenology. But their main resource turned out to be their domestic forerunners, the American legal realists of the 1920s and 1930s who critiqued the classical-formalist legal orthodoxy of their own day (*see* Legal Realism; Legal Formalism). That orthodoxy had claimed that the operating rules of the legal system could be derived from a few basic principles induced from a scientific arrangement of precedents, such as the protection of property, the basing of liability to others on 'fault,' and the enforcement of the 'free will' of parties to contracts. This system added up to a natural and neutral mode of ordering private life, and legislative interferences with this system of rules would be struck down as unconstitutional unless they met strict tests of validity.

The realists' critique, in brief, was that (1) the principles were too vague, and the precedents too flexible, to determine outcomes, and in actual operation produced results varying over time and with contexts; (2) the actual determinants of results were latent and unarticulated policies, ideologies, and social vision; and (3) the 'private' law rules of property-tort-contract that undergirded the natural-seeming private sphere of

free choice in free markets, protected against interference by constitutional law, were no less artificial, no less exercises of state policy, and no less coercive, than the 'public' rules of legislation and administration. A property right, for example, delegated to its owner the ultimate power to use state force (self-help, private guards, the police or the army) to exclude others (such as union members or black people) he did not want to hire or deal with, and to condition access to his property (tools, land, wages, the assets of a marriage) on those others' submitting to his orders. The coercive state was always involved in the market and business firms and family life; the issues for legal policy were thus not *whether* the state should be involved, but *how*, and *to what ends* – issues that courts, as well as legislatures and administrative agencies, could not avoid and had to resolve.

By the 1950s and 1960s, however, the leading currents of American legal thought had settled into a new orthodoxy, one that absorbed elements of the realist critique but blunted its more radical implications (*see* Legal Scholarship). Some postwar schools held that law (including law made by judges), although not autonomous from clashing social interests, had neutral craft techniques for balancing those interests; others that law properly interpreted served social purposes on which there was presumed to be a general consensus; and others that courts differed from legislatures in that judge-made law promoted general 'principles' rather than specific 'policies.' In the late 1960s, emerging schools of law-and-economics sought to rationalize legal doctrine, especially the private law of property, tort, and contract, as serving the aims of efficiency or wealth maximization. (For an overview see Duxbury, 1995; *see* Law and Economics.)

Early CLS

CLS began as a set of critiques of these postwar modes of legal reasoning. Unlike other movements that carried on the legal-realist critiques, such as the law-and-society movement (which investigated empirically the social influences that helped to form, and the actual social effects of, legal rules), the critics deployed an internal critique: attacking the prevalent modes of doctrinal and policy argument practiced by law teachers, judges, and other officials on their own terms, aiming to develop counterarguments to teach to law students and practicing lawyers.

Following neo-Marxist writers and the critical social theorists of the Frankfurt School, the critics pictured legal doctrines and arguments ('legal discourses') as ideology, a partial vision of the world that presents itself as a universal vision. These discourses function so as to portray existing constellations of rights, powers, privileges, and immunities recognized by the legal system as if they were natural and inevitable ('false necessity'), and as close to being good as they could be ('false legitimation'). By such means the law 'reifies' and thus helps to freeze in place scripted social roles such as manager and worker, landlord and tenant, husband and wife, consumer debtor and bank creditor, property owner and picketing laborer, that blind the ordinary people who play those roles to alternative possibilities for action and organization (see, e.g., Gabel, 1980).

Like the legal realists, the critics' method was to unearth and make transparent the latent subtexts of legal doctrines and policies, the hidden framing conceptions, empirical assumptions, and social visions on which conventional legal reasoning, and the scripted social roles it gave people to play, relied. But the aim was not simply to unmask; it was to develop alternative doctrinal and policy arguments in particular fields of law, and by so doing to open a path for a revisionary politics based on different views of social reality and human potential.

In its first decade, CLS developed a distinctive model of such critiques, which, taught to and adapted by adherents and disciples, was replicated in about 100 articles covering a wide variety of legal fields. The model critique identified some deep-level tensions, polarities, or contradictions underlying a field of legal policy or argument. In a pioneering article, Kennedy argued that contract law doctrines tended to cluster around opposing conceptions of human nature, social obligation, and the functions of law: one favoring formal rules and 'individualist' views of economic transactions and social duty, the other favoring informal standards and 'altruist' ethics of sharing and cooperation (Kennedy, 1976). The law in force at any given time oscillates between the opposing poles as it moves from one subject to another, but tends to privilege one pole and suppress the other, reserving it for exceptional or marginal situations.

Doctrines from the suppressed pole, however, are always available as a 'dangerous supplement' which at any moment may, and sometimes do take over the privileged position (Kelman, 1987). One pioneer (Unger, 1975) generalized the critique across legal fields to 'liberalism' as a whole; but most practiced it as a form of local critique, whose content varied with the field of law that was being criticized. The model also spawned its own genre of structuralist legal history, showing how legal fields come to be structured around contradictory organizing principles and social visions. For example, doctrines reflect people's need for other people and fear of them; the desire for a state powerful enough to protect rights that will not also trample them, and for contract and property law that will carry out parties' intentions and underwrite their freedom of action but also protect their (and others') security. The legal system tries to 'mediate' the contradictory doctrines by confining each to a limited field of operation, yet such systems become unstable as the mediating devices become incoherent and exceptions spill over limits to swamp the rules (see, e.g., Alexander, 1987; Kennedy, 1979; Mensch, 1982).

This strand of CLS was often misunderstood. Critics argued that the structuring of legal fields around such 'contradictions' meant that law was 'indeterminate' because it could (and historically often did) enable judges, advocates, or reformers to make valid arguments (according to internal criteria of the legal system) for opposed positions. Doctrinal or policy arguments could always be 'flipped' by appealing to principles at the opposite pole. This did not mean that law was a facade, sham, or arbitrary meaningless jumble, or that doctrine was just a mask for raw political choice; rather, its indeterminacy was patterned and systemic, an effect of law's (or liberalism's) contradictory structures. Nor did it mean that law was unpredictable, because arguments drawn from one pole or another tended to be stable in the medium term. Nor, finally, was it an indictment of a legal system to say that it was 'indeterminate.' It

was merely a fact, one papered over by mediating devices leading to the appearance of natural, necessary, and almost-as-good-as-it-could-be law.

Later CLS

CLS also extended the method beyond private law doctrine to the powerful emerging new forms of policy analysis, legal economics, attacking certain core premises of the new science including the Coase theorem, cost-benefit analysis of legal entitlements, public choice theory, and submission to the will of others as rational consent (see, e.g., Kennedy and Michelman, 1980). Others extended it to institutions: Simon identified opposed models of welfare administration, mechanical rule-application by low-skilled workers and the professional informal discretion of caseworkers. He showed how the welfare system had evolved from discretion to rules with damaging results for clients and argued for restoration of discretion (Simon, 1983).

Critics also revived and extended the legal realists' critique of the distinction between 'public' (coercive regulatory) and 'private' (freely willed contractual consensual market) relations. Following the realist R. L. Hale (see Fried, 1998 for an intellectual biography), they argued that all 'free choice' is made by actors with property and power endowments and under baseline conditions structured by the legal system, even choice in the most intimate spheres of life such as marriage relations. Their aim was to supplement Foucault's (1977) account of the pervasive regulation of social interactions through 'disciplines' with an account of the pervasive structuration of transactions and interactions in business and personal life through law.

Much critical effort went into history (see Gordon, 1984). Some critical historians worked in the tradition of the social history of subordinated groups, using historical examples to show how ruling groups had used neutral- and equal-sounding legal doctrines to extend and justify their rule. Horwitz (1977) showed how early-nineteenth-century law had reworked private law rules of property, tort, and contract to favor entrepreneurs in the early nineteenth century, then later froze the new rules against redistributive revision by imposing constitutional restraints on legislation. Klare (1978) showed how conservative judges had drained the National Labor Relations (Wagner) Act of the New Deal of its potential as a charter of workers' rights to participate in management. Critical historians often read the past as a repository of lost social experiments, long since crushed or forgotten, which might serve as models for reconstructing aspects of social life – as did Frug with the nineteenth-century model of powerful autonomous city governments (Frug, 1980), and Simon did with 'social-republican' property models of landholding and producers' cooperatives (Simon, 1991). Other histories explored the legal structuration of social relations, especially those of master and slave, employer and employee, husband and wife (see, e.g., Steinfeld, 1991; Olsen, 1983).

By the mid-1980s CLS had entered a new phase. One group of critics, perhaps responding to accusations that CLS was 'nihilist' in its attachment to destructive criticism ('trashing'), turned to concrete programmatic work. Unger's (1998) was once again the most ambitious, proposing large-scale programs for the democratization of the economies of both developed and developing societies; most others were small-scale proposals for reforming specific legal policy fields, such as worker participation in ownership and management (Stone and van Wezel, 1988), treatment of workers in plant closings (Singer, 1988), regulation of injurious speech (Abel, 1994), intellectual property (Boyle, 1996), disability law (Kelman and Lester, 1997), and property rights in ex-Communist societies (Alexander and Skapska, 1994).

Another group focused on developing critical theories of law to a new level of sophistication, integrating them with postmodernist and poststructuralist theories, especially those of Foucault and Derrida (see, e.g., Balkin, 1987; Heller, 1984; Peller, 1985).

Challenges and Erosion

Around the same time, CLS experienced intensifying challenges from neighboring left legal-intellectual movements, especially feminism and critical race theory. The new movements asserted differences from CLS in focus and method – differences that provoked vigorous arguments throughout the 1990s. CLS had criticized liberal legalism for its claims to have solved the problem of domination by constructing 'rights.' Some critics said rights increased the alienation of individuals by constituting them as monads isolated from one another (Gabel, 1984). Others argued that rights mystified the realities of power relations, while still others argued that social movements' crusades for rights (such as the black civil rights movement) often ended up giving them only symbolic victories and unenforceable claims still others contended that rights, like other legal doctrines and policies, were always in conflict – the right to equal treatment in accommodations, housing, and employment, for example, with the right to associate – and that the conflicts could not be resolved by appeals to rights (Tushnet, 1984). Race theorists (e.g., Crenshaw, 1988) counterargued that whatever might be the weaknesses of rights-rhetoric, claims using the language of rights had been essential to the gains of subordinated groups. CLS scholars had also stressed the indeterminacy, manipulability, and contingency of social arrangements and legal discourses, the fluidity of meanings and identities. Some feminist and critical-race theorists, by contrast, wished to insist at least provisionally on the grounded objectivity of experiences of subordination, the common shared elements that made up sexual or racial identity and the authenticity of personal narratives of the lives of the subordinated (*see* Critical Race Theory; Feminist Legal Theory).

CLS had helped to open up and to institutionalize space – in the form of places for faculty positions, room for articles in law reviews, slots on the programs of conferences – for a legal-academic left, a novel phenomenon in the USA (and wholly missing in other professional faculties such as business schools). By the end of the 1990s, however, most of that space, like most of the energy going into novel work, was occupied by gender and race studies. CLS ideas continued diffusely to pervade both mainstream and critical fields of scholarship and policy analysis – notably its critiques of the public-private distinction and of the legal foundations of neo-liberal political economy. But as a distinct school CLS had faded away.

European CLS

European critical schools of law followed similar paths, with important differences due to national variations in the strength and programs of left-wing political parties and intelligentsia, and the objects of critique – traditional modes of legal training and reasoning – on the other. (A European Conference on Critical Legal Studies was organized in 1981.) In Britain, CLS were first concentrated in social science faculties and grew out of the radical sociology of law, oriented in its early years to the critique of the 'State' and its legal controls of poor and working people through labor, policing, criminal law, and professionalized social services, later evolving into broader critiques of law as ideology (see, e.g., Cain and Hunt, 1979). In the 1980s CLS was taken up by law teachers, especially in the polytechnics. They organized the Critical Legal Conference in 1984, proclaiming "a shared rejection of the dominant tradition of Anglo-American legal scholarship, the expository orthodoxy or … 'black-letter law' tradition" (Fitzpatrick and Hunt, 1987, p. 1). Believing that their American counterparts focused too narrowly on the internal critique of legal doctrine and reasoning, British critics maintained their connections to socio-legal studies and to radical law practices; they also developed new alliances with feminist, deconstructionist, psychoanalytic, and literary approaches. CLS in the UK, as in the USA, are a very fragmented enterprise, but unlike its American counterpart, it is flourishing (for exemplary collections and overviews see Fitzpatrick and Hunt, 1987; Douzinas et al., 1994).

In France the combined influence of (1) Marxist traditions treating law as 'merely superstructural' and (2) separation of faculties of legal philosophy and theory from faculties of law, held back the development of left-wing legal studies until the late 1960s, when radical social movements and the surge of interest in the social theories of Althusser, Gramsci, and Foucault led to critiques of law as forms of ideology that structure and legitimate the functioning of capitalist order. After the 1980s the movement shifted emphasis from general theory to concrete case studies of the norms, reasoning modes, and institutions that went into the construction of everyday legal relations. French CLS has not deeply penetrated the centers of university training of the legal elite, but has gained a foothold in some faculties and in research centers at St. Etienne, Lyon, and Montpellier (see Arnaud, 1993 for a concise summary).

See also: Critical Legal Studies; Critical Race Theory; Critical Theory: Contemporary; Feminist Legal Theory; Law: Economics of Its Public Enforcement; Legal Formalism; Legal Realism.

Bibliography

Abel, R., 1989. American Lawyers. Oxford University Press, New York.
Abel, R., 1994. Speech and Respect. Sweet & Maxwell, London.
Alexander, G., 1987. The transformation of trusts as a legal category. Law & History Review 5, 303.
Alexander, G., Skapska, G., 1994. A Fourth Way? Privatization, Property and the Emergence of the New Market Economies. Routledge, New York.
Arnaud, A-J., 1993. Critique du Droit. In: Arnaud, A.-J. (Ed.), Dictionnaire Encyclopedique de Théorie et de Sociologie de Droit, second ed. Libre générale de droit et de jurisprudence, Paris.
Balkin, J., 1987. Deconstructive practice and legal theory. Yale Law Journal 96, 743.
Bauman, R.W., 1996. Critical Legal Studies: A Guide to the Literature. Westview Press, Boulder, CO.
Boyle, J., 1996. Shamans, Software and Spleens. Harvard University Press, Cambridge, MA.
Cain, M., Hunt, A. (Eds.), 1979. Marx and Engels on Law. Academic Press, London.
Crenshaw, K., 1988. Race, reform and retrenchment. Harvard Law Review 101, 1331.
Douzinas, C., Goodrich, P., Hatchamovitch, Y. (Eds.), 1994. Politics, Postmodernity, and Critical Legal Studies. Routledge, London.
Duxbury, N., 1995. Patterns of American Jurisprudence. Oxford University Press, New York.
Fitzpatrick, P., Hunt, A. (Eds.), 1987. Critical Legal Studies. Blackwell, Oxford, UK.
Foucault, M., 1977. Discipline and Punish: The Birth of the Prison. Pantheon, New York.
Fried, B.H., 1998. The Progressive Assault on Laissez-Faire: Robert Hale and the First Law and Economics Movement. Harvard University Press, Cambridge, MA.
Frug, G., 1980. The city as a legal concept. Harvard Law Review 93, 1057.
Gabel, P., 1980. Reification in legal reasoning. Research in Law and Sociology 3, 25.
Gabel, P., 1984. The phenomenology of rights-consciousness and the pact of the withdrawn selves. Texas Law Review 62, 1563.
Gordon, R.W., 1984. Critical legal histories. Stanford Law Review 36, 57.
Gramsci, A., 1971. Selections from the Prison Notebooks. In: Hoare, Q., Nowell-Smith, G. (Eds.). International Publishers, New York.
Hay, D., 1975. Property, authority and the criminal law. In: Hay, D., Linebaugh, P., Rule, J.G., Thompson, E.P., Winslow, C. (Eds.), Albion's Fatal Tree. Pantheon, New York.
Heller, T.C., 1984. Structuralism and critique. Stanford Law Review 36, 127.
Horwitz, M.J., 1977. The Transformation of American Law, 1780–1860. Harvard University Press, Cambridge, MA.
Kelman, M., 1987. A Guide to Critical Legal Studies. Harvard University Press, Cambridge, MA.
Kelman, M., Lester, G., 1997. Jumping the Queue. Harvard University Press, Cambridge, MA.
Kennedy, D., 1976. Form and substance in private law adjudication. Harvard Law Review 89, 1685.
Kennedy, D., 1979. The structure of Blackstone's commentaries. Buffalo Law Review 28, 205.
Kennedy, D., Klare, K., 1984. A bibiliography of critical legal studies. Yale Law Journal 94, 461.
Kennedy, D., Michelman, F., 1980. Are property and contract efficient? Hofstra Law Review 8, 711.
Klare, K., 1978. The judicial deradicalization of the Wagner Act and the origins of modern legal consciousness. Minnesota Law Review 62, 265.
Mensch, E., 1982. The colonial origins of liberal property rights. Buffalo Law Review 31, 635.
Olsen, F., 1983. The family and the market: a study of ideology and legal reform. Harvard Law Review 96, 1497.
Peller, G., 1985. The metaphysics of American law. California Law Review 73, 1151.
Simon, W.H., 1983. Legality, bureaucracy and class in welfare administration. Yale Law Journal 93, 1198.
Simon, W.H., 1991. Social republican property. UCLA Law Review 38, 1335.
Simon, W.H., 1992. The prospects of pension-fund socialism. Berkeley Journal and Employment & Labor Law 14, 251.
Singer, J.W., 1988. The reliance interest in property. Stanford Law Review 40, 611.
Steinfeld, R., 1991. The Invention of Free Labor. University of North Carolina Press, Chapel Hill, NC.
Stone, K., van Wezel, 1988. Labor and the corporate structure. University of Chicago Law Review 55, 73–173.
Thompson, E.P., 1971. Whigs and Hunters. Pantheon, New York.
Tushnet, M., 1984. An essay on rights. Texas Law Review 62, 1363.
Unger, R.M., 1975. Knowledge and Politics. Free Press, New York.
Unger, R.M., 1998. Democracy Realized. Verso, New York.

Critical Planning Theory

Michael Gunder, University of Auckland, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article appraises the evolving nature of critical planning theory. It begins with an interrogation of critical planning theory derived from political economy. It then considers liberal pragmatic theory, followed by poststructuralist critical theory. This is followed by an examination of the contemporary range of critical critiques of the role of planning under neoliberalism. The article will conclude with a brief consideration as to why critical planning theory has continued for over four decades to provide an important ethical compass for the planning discipline.

Introduction

Planning theory may be conceptualized broadly under a contemporary triplet of labels: rational, communicative, and critical, which corresponds loosely to a topology of planning theory addressing three normative issues: 'What is planning's purpose?', 'What constitutes good planning process?' and 'What constitutes good planning?' (Yiftachel, 1989: p. 26). This article is principally concerned about the latter question. Critical planning theory (CPT) engages with questions of power, equity, language, knowledge construction, and related issues so as to test "professional concepts, models and 'gospels' against their real world material and discursive consequences" (Yiftachel and Huxley, 2000: p. 910).

Spatial and related types of planning, by their very nature, are largely apparatus of state government and governance (*see* Planning Theory: The Good City and Its Governance); hence, CPT inherently tends to interrogate the role of the state and its overarching constituting ideology to evaluate who benefits and loses through the application of planning's diverse activities. CPT might be further deconstructed under a range of categories attributable to the diverse fields of thought that have been applied to its theorization over the last several decades. Vastly simplified and aggregated, these include structuralist political economy, liberal pragmatism, poststructuralism, as well as a range of interrelated contemporary critiques of neoliberalism, with this last category largely emerging in the new millennium and yet to effectively formulate a singular overarching definitive label. This article will briefly consider separately each of these categories of thought, which have sequentially evolved over the last four decades in CPT, even though, in actuality, they often tend to overlap in their application and blur in their temporality. The article will then conclude with a consideration as to why CPT has continued to evolve over this period and suggest that this is because it provides an important and necessary ethical consciousness for planning that helps to provide a reflective "humanist philosophy to guide planners in their work" (Friedmann, 2008: p. 255).

Structuralist Political Economy Interrogation

Arising out of the popular dissatisfaction with Western governments of the late 1960s, and often drawing on the Marxist and at times contradictory positions of Manuel Castells (1977) and Henri Lefevbre (1991[[1974]], 1996, 2003[[1970]]), CPT, particularly in North America, drew on structuralist political economy to interrogate planning's role of facilitating the state and called in its day for a new revolutionary urbanism predicated on social justice (Harvey, 2009[1973]). Scott and Roweis (1977: p. 1099) best summarized this critical position for many theorists, by deploying Marxian critique to explain "urban planning as an observable social and historical phenomenon." In doing so, the two authors critiqued "noncritical" planning theory's underlying purpose, which they argued served and largely justified "the ideological and apologetic functions" of planning practice necessary to justify state interventions in capitalist "social and property relations" (p. 1114), while constituting a naive idealist-utopian urban problematic, which then calls out for a humanistic prescribe of what "*ought* to be" (pp. 1116–1117).

David Harvey (1985: p. 177) further engaged with critical political economy to argue that planners indeed shape and operate under a "progressive ideology" that purports to maintain a balance in society, predicated on rationality, in regard to the built environment so as to "reduce civil strife and to maintain the requisite conditions for the steady accumulation of capital" and this is achieved by placing the planner in the role of "defender of the public interest." Of course, what goes unstated in this progressive ideology is that the ultimate public interest is the creation of wealth, but not its equitable distribution. In this regard, Peter Marcuse (1976: p. 272) had a decade before being called for a structural approach to planning ethics, where planning "ethics could logically deal with issues of power and act directly to further values such as equity and democracy." Norman and Susan Fainstein (1979: p. 399) argued from this ethical perspective for the value of a working-class orientated planning practice that could "show the consequences for the lower classes of solutions proposed by monopoly capital" and which could create sensitive planning "solutions which address systemic problems … with minimal harm to the lower classes" so as to "advance working class interests."

Liberal Pragmatic Interrogation

Not all critical planning theorists of that period based their approach purely on structuralist political economy; many took

a more pragmatic liberal approach, all-be-it often influenced by the critical communicative theory of the German sociologist Jürgen Habermas (Forester, 1980). John Forester (1982: p. 76), following both Habermas and Lukes (2005[1974]), considered three types of mechanisms for exercising power in planning-related matters: through "[1] the ability to prevail in decision making…[2] the setting of agendas…[3] the ability of major institutions and actors to shape the felt needs and self-conceptions of citizens: for example, that they must acquiesce in the face of big government and big business…that individual market consumption will provide for the satisfaction of all needs, and that collective action is not a public responsibility but a nuisance." Taking an ethnographic approach to researching American planning practice, Forester (1989) found that practitioners often act in opposition to dominant power interests and he documented the planner's often attention-shaping role in seeking to resist domination so as to achieve more equitable planning outcomes.

Susan Fainstein (2000: p. 468, 2010), subsequent to her earlier pro-Marxist engagement, developed a liberal argument during the 1990s and 2000s for 'the just city,' one with an entrepreneurial state projecting 'a future embodying a middle-class society.' Fainstein (2010) established three interlinked criteria, or primary qualities constituting urban justice, for adjudicating the just city. These were 'democracy' and 'equity,' drawn from the American liberal perspective, and, attributed to poststructuralism, rather than liberal identity politics, that of 'diversity.' The author developed a list of principles for the furtherance of these three criteria, which were largely focused on planning for a more equitable American city, but one that used the European city of Amsterdam as its template. These ranged from prescriptive and detailed principles for the provision of affordable housing and public transit, public space provision, employee rights, through to calls for increased meaningful public consultation, consistent with the liberal-democratic American political tradition. The work called for more regulation and public ownership of resources, particularly by local government, while still operating within a capitalist market. Throughout her work, Fainstein has always argued for planners to help shift the urban problematic toward a greater concern for social equity.

With less optimism, Oren Yiftachel (1995a,b) explored planning's use as a state mechanism of society control, not for positive societal reform, but for the domination of a minority population. This was a use of planning that was not enlightened, but rather oppressive – hence planning's 'dark side'– and he empirically documented how state planning processes were deployed to spatially, procedurally, and socioeconomically oppress the Arab minority in the deeply divided society of Israel. Yiftachels (1998: p. 399) then universalized this critique and considered further "the dialectical tension between reform and control" by including a fourth dimension of potential oppression, implicitly, but not explicitly, categorized in the prior works: planning's potential cultural impact on minority groups in any society.

Bent Flyvbjerg (1996) also explored the 'dark side' of planning through his critique of normative idealism in modern social science and his development of an in-depth case study methodology, which he labeled phronetic social science. This was a social science approach that drew on the Aristotelian intellectual virtue of *phronesis*, so as to provide a method of social enquiry "to carry out analyses and interpretations of the status of values and interests in society aimed at social commentary and social action" that evaluates if particular public policies, such as a plan, are desirable and also determines "who gains and who loses" as a consequence and "by which mechanisms of power" (Flyvbjerg, 2001: p. 60). This is a CPT methodology that he deployed in his book-length critique of planning practice in Aalsbourg, Denmark, which clearly illustrated how dominant power was deployed in a Foucauldian manner to shape urban rationality (1998a).

Poststructural Interrogation

As well as Flyvbjerg, a number of other CPT authors placed Foucauldian analysis at the center of their CPT research. They ask the key critical question 'what is actually done' so as to provide necessary understanding and guidance to develop the 'strategies and tactics needed' to counter existing oppression and/or power imbalances, rather than the normative idealist question of much non-CPT of simply 'what should be done' (Flyvbjerg and Richardson, 2002: p. 61). Foucault, with his focus on power, is the predominant poststructuralist deployed for this critique, although others such as Derrida, Lacan, and Deleuze have also been called upon (Gunder and Hillier, 2009). Tim Richardson (1996: p. 280) introduced Foucauldian critique to planning theory as a set of tools to analyze the agonistic "dynamics of power and knowledge in policy making," so as to unmask power by grasping how "power appropriates knowledge, and weaves it into discourses" in the achievement of particular objectives. Richardson's discourse analysis sought to "ask how, why and by who, truth is attributed to particular arguments and not to others" (p. 283). He suggested that this Foucauldian approach allowed a new critical perspective from which to interrogate planning policy formulation by asking "how arguments are transformed into 'knowledge'" and that this perspective could form "an agenda of unmasking and challenging power, and the consequent empowering of excluded or weak minorities" (p. 290).

Unsurprisingly, several authors deployed poststructural critique to demonstrate the inherent failure of Habermasian-derived communicative planning, which had emerged in the 1990s as a prominent form of planning process concerned with participatory consensual decision making (Healey, 1992; Innes, 1995), to adequately deal with and engage with the concept of power. In this regard, Fischler (2000: p. 362) raised concerns about what Foucault called governmentality, that is, "the modern rationality of government—by revealing the origins of the ways in which people govern themselves, govern other individuals, and govern society at large" and cautioned that "we ought to think critically about communicative theory as a new theory of government and about collaborative action as a new form of governmental action." In contrast to a Habermasian seeking of utopian consensus, Flyvbjerg (1998b) drew on Foucault to argue for the importance of conflict to ensure democratic process, for from a Foucauldian perspective the suppression of conflict is also the suppression of personal choice for difference and the freedom to argue for

this choice. Perhaps the most critical nail in the coffin of communicative planning, at least from a poststructuralist perspective, was not derived from Foucault, but from Lacan. Hillier (2003: p. 46), drawing on Lacan, demonstrated how the very incompleteness and ambiguity of language makes the idea of consensus an impossibility, as "[t]here is always an excess of meaning that eludes representation."

Foucault's concept of governmentality has been deployed by several CPT authors in addition to Fischler. Gunder and Mouat (2002) demonstrated how governmentality was deployed in a manner "generating domination and oppression for the majority of the population within planning processes in New Zealand by subtle accumulations of minor manipulations, or pragmatic administrative practices of government, resulting in an aggregate of symbolic violence and the frustration of institutional victimization for the majority." Huxley (2002) explored the "history of the present" that interrogated planning as a form of governmentality; that is, a 'conduct of conduct' evolved genealogically through the codification of the social sciences that structures our taken for granted, or normalized, use of space. This is a normalized use of space that simply reproduces 'gendered regimes of domination' that produce the normalized female subject of 'compulsory heterosexuality, motherhood and reproduction' (p. 151). Consistent with Flyvbjerg (1998b), Pløger (2004: p. 73) also argued for resistance to mainstream planning "'governmentalities' framing urban governance initiatives and experiments" so as to engage with and respect the important role of strife and agonism in real-world planning practice. This was to positively empower local communities so as to ensure than planning actually created communities that are wanted by their inhabitants.

Contemporary Neoliberal Interrogation

Over the last decades, CPT's interrogation of planning has changed significantly as the postwar welfare state has evolved into the contemporary entrepreneurial neoliberal state. As outlined in the prior sections, "[s]ince the 1970s, the claims of planning to moral rightness, influence and expertise have come under scrutiny, with the result that the" traditional "legitimacy of the profession and public trust in its institutions have both been challenged" (Laurian, 2009: p. 369). Consistent with the dominance of the public choice school in economics and the rise of acceptance of neoliberal ideology, trust in bureaucratic planning prescription was eroded as an alternative to market solutions, with this arguably constricting 'the planner's role' to that of a 'designer of efficient and democratically acceptable market institutions' (Sager, 2009: p. 66).

Planning has come under further challenge as a consequence of a range of other, often related, issues. These include the challenge to "notions of exclusivity of knowledge"; the "displacement of the power of the bureau-professions by managerial [ism]"; the "de-professionalising tendencies of government through the imposition of performance management techniques"; and, not to mention, the "increasing specialisations within the domains of practice" (Vigar, 2012: p. 363). The latter includes the emergence of urban design, with its promarket/client orientation, as a new and separate profession to that of urban planning, which still tends to retain its orientation to achieve public rather than private goods (Gunder, 2011). Laurian (2009: p. 369) further asserts that planning has also come under challenge from an academic perspective derived from "the 'radical' critique of instrumental rationality and [planning's sometimes] monolithic, exclusive views of the public interest," as outlined in the prior sections.

Swain and Tait (2007: p. 236) contextualize planning in the context of modern risk society, where 'thick' trust of traditional expertise and authority is at best displaced into a 'thin' faith in expert systems. This was a consequence of an aggregate societal loss of individually perceived ontological security in the latter part of the twentieth century. This loss occurred as a consequence of heightened reflexive individual awareness that many issues of complex societal and planning concern, such as that of race, gender, or environmental risk, where not being adequately addressed by government, or other societal institutions (Beck, 1992). These were often concerns of great complexity, in contrast to the more 'manageable' concerns of traditional instrumentalist planning, where readily calculable best means could generally be determined to achieve measurably successful end states, such as the best means to calculate and then allocate vacant land necessary to meet forecast housing needs.

These new concerns were often 'wicked problems' (Ritter and Webber, 1973: p. 160) without readily calculable best means of solution or, in some cases, even fully comprehensible issues capable of having definable end states of resolution; hence, these wicked problems were not solvable, 'at best they are only re-solved – over and over again.' Accordingly, traditional prescriptive planning, predicated on instrumental rationality, was perceived to fail in this resolution of complex problems. In some instances, such as that of the United Kingdom, this gave ground to new planning approaches, predicated on communicative rationality (Gunn and Vigar, 2012). Yet, more often and more progressively over time, direct planning 'rational' prescription was simply displaced from the agenda, leaving ground for only market-driven issues and their market-driven 'solutions,' with difficult nonmarket-orientated problems largely displaced from the governance agenda, or, at best, they were dealt with in a nonspecific and generally meaningless 'feel good' manner (Allmendinger and Haughton, 2010: p. 804).

In this light, Sager (2009: p. 65) contends that under the global dominance of neoliberal ideology in recent years, New Public Management has become the "most important economic-political influence" on public sector institutional behavior and that this has transformed "the planning bureaucracies of many countries in line with principles of competitiveness, market-orientation, and economic accountability." In this context, the planner's role is perceived "as providing expertise in substantive and legal-procedural matters," with planned solutions being "in harmony with the market… thus leaving less room for politics" (Sager, 2009: p. 66). Further, this new planning practice also did not necessarily attempt to resolve 'wicked problems,' but often tends to obscure them, or, at best, give the appearance that something is being done – or managed – about them. This most often occurs by labeling these issues or solutions in desirous, but ultimately meaningless terms, such as sustainability, livability, resilience, or 'smart growth' (Gunder and Hillier, 2009).

Recent trends in the neoliberal turn of governance have resulted in a "managerial restructuring of government [which] has aimed at depoliticizing decisions by making them a matter of operational management…[where] [s]ocial and political issues are reduced to technical and procedural matters; they are translated into problems to be managed" (Sager, 2009: p. 73). Indeed, planning "has long stood accused of acting as a mechanism for depoliticisation, masking value-based decisions in rational—technical or professional justifications as a means of defusing conflict and imposing development without fully examining its social or environmental consequences" (Inch, 2012a: p. 523). Moreover, the communicative turn in planning's seeking of consensus further depoliticized contentious issues, as the "cooperative search for shared understanding and agreement cannot foster the kind of counter-hegemonic politics we require to challenge neoliberalization" (Purcell, 2009: p. 152).

This depoliticizing role of planning is often attributed to what has emerged in the last decade under the label spatial planning. Most commentators agree that 'spatial planning,' while comprising a wide range of diverse practices, is predominantly "predicated upon a role for planning of shaping economic, social, cultural, and ecological dimensions of society through 'place making'," with it generally having the characteristics of "encouraging long term strategic visions; providing the spatial dimension to improving integration across a range of sectoral plans and activity; supporting 'balanced' approaches to sustainable development; and improving engagement with stakeholders and the public" (Allmendinger and Haughton, 2010: p. 803).

Yet, Allmendinger and Haughton (2010: p. 804) go further and are critical of spatial planning, suggesting that spatial planning in its postideological, or postpolitical condition, is largely "organised around loosely defined and hard to refute 'feel good' issues and labels such as 'sustainable development'" that tend to promote "the use of consensus-based approaches to reinforce mainstream growth-led ideologies," which "in effect deny legitimacy and influence to more radical alternatives" and marginalize effective opposition. They continue (p. 809): "[f]ar from being strategic and steering, spatial planning could just as easily be portrayed as supine and subservient to powerful vested interests in the development process." Indeed, consistent with Fischler's (2000) concerns about communicative planning creating the grounds for a new form of governmentality, these authors go so far as to assert that "spatial planning…needs to be analysed as a form of neoliberal spatial governance" (Allmendinger and Haughton, 2012: p. 89).

In this regard, "[t]he problematisation of the planning system itself can therefore be seen as a means of covering over the underlying causes of conflict with fantasies of consensus that a fit-for-purpose planning process will create" (Inch, 2012a: p. 532). Indeed, a number of authors have documented, at least in the United Kingdom, that planning itself, or at least the prevailing planning system, has often been portrait in the role of scapegoat for both the state and the media when difficult to resolve contradictions occur between spatial activities, such as that of countryside housing provision, especially in the south of England (Inch, 2012a).

In professional planning practice, the "authenticity of knowledge circulating…is insured by the presence of professional standards" (Amin and Roberts, 2008: p. 359). However, within such communities of professional practice innovation "tends to be incremental rather than radical," where "the protectionist role of professional associations can act as a barrier to radical change" with at best professional associations promoting the dissemination "of new incremental knowledge across the sector" (Amin and Roberts, 2008: p. 359). Further, as the following authors contend, this institutional professional barrier to radical change actually facilitates the adoption of planning practice that tends to conform to the values of the dominant actors shaping the *doxa* of contemporary society. Further, this conformist practice generally fails to engage with unresolvable, or wicked problems, rather its consensus-based approach tends to, at best, paper over these difficult to resolve concerns and issues and concentrate on what really matters: market success and resultant economic growth.

While Amin and Roberts (2008) argue that the protectionist role of professional planning associations tend to act as a barrier to significant planning change, Andy Inch (2012b) documents how British planners, led by their professional institute – the Royal Town Planning Institute – sought to redefine, or reterritorialize, their professional practice during the first decade of the new millennium, as 'spatial planning,' rather than the more traditional 'land-use planning' or 'town planning.' This *cultural work*, as Inch (2012b: p. 528) calls it, required the profession "to reshape the knowledge, social relations and identities underpinning their practices, organising the attention of other actors, and altering existing conceptions of planning's role" resulting "in reworking the boundaries of the field of professional practice, at various levels from the national to the local."

This process was also co-opted in the interests of the state and used as a mechanism to further facilitate planning as a largely postpolitical process (Allmendinger and Haughton, 2012). Indeed, rather than increasing and empowering the discretionary professional space for effective planning practice, some contend that the adoption of a spatial planning paradigm has actually reduced it (Gunn and Vigar, 2012).

Yet, planning is still often perceived as making a positive contribution to society. But this is because today's successful planners are often empowered by market success, that is, they are seen as contributing to the neoliberal success of their polity. In this light, "[c]ontributing to the overall 'gain' for the city, to the 'good' for the city, as defined by neoliberal planning principles, acts as an aphrodisiac for planners and other policy makers, since those principles form the parameters of urban success" (Baeten, 2011: p. 209). That these principles and the plans and projects that they often create have a tendency to negate anything that does not contribute to economic gain is just not part of the consideration (Gunder, 2006).

Indeed, within the neoliberal world of the new millennium, the traditional authority and disciplinary rules of government and that of its planning experts providing public welfare have been displaced. Rather than traditional government facilitating a role of pastoral care for its Foucauldian docile and obedient population in creating a 'better' world for all, a new ideal has emerged of the state facilitating socially mobile competitive individualism. As a consequence, the "*symbolic* prohibitive

norms" of the welfare state have been "increasing replaced with *imaginary* ideals (of social success," of smart growth, of livability, of sustainable development – which Swyngedouw (2010) refers to as the new opiate of the masses – not to mention, being a globally competitive city to ensure wealth creation), where the neoliberal subject is actually driven by a societally imposed injunction to no longer simply 'obey' but rather to ultimately 'enjoy' in his/her 'sustainable' consumption (Žižek, 1998: p. 480 – emphasis in original). This contemporary world is no longer simply a world of government, but rather one of governance, and the ideological role, or appearance, of planning is no longer one of control so as to channel balanced development for a collective public interest, but rather to facilitate the success of a polity by ensuring the appropriate environment for the competitively achieved market-derived enjoyment for its residents (Gunder, 2010). Planning's contemporary mainstream success is largely adjudicated accordingly. Consequently, CPT's contemporary task now is to open this adjudication to challenge where and as necessary.

The Importance of CPT

A small group of planning theorists and related scholars have asked critical questions of planning and planning's role in wider society over, at least, the last four decades. Their arguments and debates have been derived from a range of schools of thought, but generally they have been concerned about equity, fairness and social justice. They have all interrogated the role of planning as an agent of the state facilitating public policy as to how the future city should be constituted. CPT has endured because it provides an important and necessary ethical consciousness that questions planning practice so as to help provide a reflective "humanist philosophy to guide planners in their work" (Friedmann, 2008: p. 255). Indeed, CPT repeatedly reminds planners for the need to continuously reflect on how power is deployed both by others and by their own practice on and in their communities and throughout wider society.

CPT has sought to remove the ideological veils that shield and shape the mechanisms that drive the capitalist global culture forward and that often induce distortions and disparities within and across societies. Indeed, planning is largely about trying to change the way things are for something better, which is inherently a normative task. These good ideas of how to achieve a better future community are highly subject to capture by the powerful in the seeking of their own interests. CPT is also an important and necessary check and balance on planning practice itself. It provides a means to understand what is actually being done, who actually benefits, and who may actually loose. With this questioning and resultant understanding comes the potential to develop tactics and strategies that may make planning a more equitable process that indeed can lead to more effective planning practice, and potentially better communities for all inhabitants; however, each polity may wish to collectively define its normative concept of 'better.'

See also: Just City; Planning Theory: The Good City and Its Governance; The Right to the City.

Bibliography

Allmendinger, P., Haughton, G., 2010. Spatial planning, devolution, and new planning spaces. Environment and Planning C: Government and Policy 28 (5), 803–818.

Allmendinger, P., Haughton, G., 2012. Post-political spatial planning in England: a crisis of consensus? Transactions of the Institute of British Geographers 37 (1), 89–103.

Amin, A., Roberts, J., 2008. Knowing in action: beyond communities of practice. Research Policy 37 (2), 353–369.

Baeten, G., 2011. Neoliberal planning: does it really exist? In: Taşan-Kok, T., Baeten, G. (Eds.), Contradictions of Neoliberal Planning. Springer, Dordrecht, pp. 205–211.

Beck, U., 1992. Risk Society: Towards a New Modernity. Sage, London.

Castells, M., 1977. The Urban Question: A Marxist Approach. MIT Press, Cambridge, MA.

Fainstein, S., 2000. New directions in planning theory. Urban Affairs Review 35 (4), 451–478.

Fainstein, S., 2010. The Just City. Cornell University Press, London.

Fainstein, N., Fainstein, S., 1979. New debates in urban planning: the impact of Marxist theory within the United States. International Journal of Urban and Regional Research 3 (1–4), 381–403.

Fischler, R., 2000. Communicative planning theory: a Foucauldian assessment. Journal of Planning Education and Research 19 (4), 358–368.

Flyvbjerg, B., 1996. The dark side of planning: rationality and realrationalität. In: Mandelbaum, S., Mazza, L., Burchell, R. (Eds.), Explorations in Planning Theory. Center for Urban Policy Research Press, New Brunswick, NJ, pp. 383–394.

Flyvbjerg, B., 1998a. Rationality and Power. University of Chicago Press, Chicago.

Flyvbjerg, J., 1998b. Habermas and Foucault: thinkers for civil society? British Journal of Sociology 49 (2), 210–233.

Flyvbjerg, B., 2001. Making Social Science Matter. Cambridge University Press, Cambridge.

Flyvbjerg, B., Richardson, T., 2002. Planning and Foucault: in search of the dark side of planning theory. In: Allmendinger, P., Tewdwr-Jones, M. (Eds.), Planning Futures: New Directions for Planning Theory. Routledge, London, pp. 44–62.

Forester, J., 1980. Critical theory and planning practice. Journal of the American Planning Association 46 (3), 275–286.

Forester, J., 1982. Planning in the face of power. Journal of the American Planning Association 48 (1), 67–80.

Forester, J., 1989. Planning in the Face of Power. University of California Press, Berkeley.

Friedmann, J., 2008. The uses of planning theory: a bibliographic essay. Journal of Planning Education and Research 28 (2), 247–257.

Gunder, M., 2006. Sustainability: planning's saving grace or road to perdition? Journal of Planning Education and Research 26 (2), 208–221.

Gunder, M., 2010. Planning as the ideology of (neo-liberal) space. Planning Theory 9 (4), 298–314.

Gunder, M., 2011. Is urban design still urban planning? An exploration and response. Journal of Planning Education and Research 30 (4), 184–195.

Gunder, M., Hillier, J., 2009. Planning in Ten Words or Less: A Lacanian Entanglement with Spatial Planning. Ashgate, Farnham.

Gunder, M., Mouat, C., 2002. Symbolic violence and victimisation in planning processes: a reconnoitre of the New Zealand Resource Management Act. Planning Theory 1 (2), 125–146.

Gunn, S., Vigar, G., 2012. Reform processes and discretionary acting space in English planning practice, 1997–2010. Town Planning Review 83 (5), 533–551.

Harvey, D., 1985. On planning the ideology of planning. In: Harvey, D. (Ed.), The Urbanization of Capitalism. Basil Blackwell, Oxford, pp. 165–184.

Harvey, D., 2009[1973]. Social Justice and the City. The University of Georgia Press, London.

Healey, P., 1992. Planning through debate: the communitive turn in planning theory. Town Planning Review 63 (2), 143–162.

Hillier, J., 2003. Agonising over consensus: why Habermasian ideals cannot be 'real'. Planning Review 2 (1), 37–59.

Huxley, M., 2002. Governmentality, gender, planning: a Foucaultian perspective. In: Allmendinger, P., Tewdwr-Jones, M. (Eds.), Planning Futures: New Directions for Planning Theory. Routledge, London, pp. 136–153.

Inch, A., 2012a. Creating 'a generation of NIMBYs'? Interpreting the role of the state in managing the politics of urban development. Environment and Planning C: Government and Policy 30 (3), 520–535.

Inch, A., 2012b. 'Cultural work', spatial planning and the politics of renewing public sector planning professionalism in England. Town Planning Review 83 (5), 513–532.

Innes, J., 1995. Planning theory's emerging paradigm: communicative action and interactive practice. Journal of Planning Education and Research 14 (3), 183–189.

Laurian, L., 2009. Trust in planning: theoretical and practical considerations for participatory and deliberative planning. Planning Theory & Practice 10 (3), 369–391.

Lefevbre, H., 1991[1974]. The Production of Space. Blackwell, Oxford.

Lefevbre, H., 1996. Writings on Cities. Blackwell, Oxford.
Lefevbre, H., 2003[1970]. The Urban Revolution. University of Minnesota Press, Minneapolis, MN.
Lukes, S., 2005[1974]. Power: A Radical View. Palgrave Macmillan, London.
Marcuse, P., 1976. Professional ethics and beyond: values in planning. Journal of the American Institute of Planners 42 (3), 264–274.
Pløger, J., 2004. Strife: urban planning and agonism. Planning Theory 3 (1), 71–92.
Purcell, M., 2009. Resisting neoliberalization: communicative planning or counter-hegemonic movements. Planning Theory 8 (2), 140–165.
Richardson, T., 1996. Foucaultian discourse: power and truth in urban and regional policy making. European Planning Studies 4 (3), 279–292.
Ritter, H., Webber, M., 1973. Dilemmas in a general theory of planning. Policy Sciences 4 (1), 155–169.
Sager, T., 2009. Planner's role: torn between dialogical ideals and neo-liberal realities. European Planning Studies 17 (1), 65–84.
Scott, A., Roweis, S., 1977. Urban planning in theory and practice: a reappraisal. Environment and Planning: A 9 (10), 1097–1142.
Swain, C., Tait, M., 2007. The crisis of trust and planning. Planning Theory & Practice 8 (2), 229–247.
Swyngedouw, E., 2010. Trouble with nature: 'ecology as the new opium for the masses'. In: Hillier, J., Healey, P. (Eds.), The Ashgate Research Companion to Planning Theory. Ashgate, Farnham, pp. 299–318.
Vigar, G., 2012. Planning and professionalism: knowledge, judgement and expertise in English planning. Planning Theory 11 (4), 361–378.
Yiftachel, O., 1989. Towards a new typology of urban planning theories. Environment and Planning B: Planning and Design 16 (1), 23–39.
Yiftachel, O., 1995a. The dark side of modernism: planning as control of an ethnic minority. In: Watson, S., Gibson, K. (Eds.), Postmodern Cities and Spaces. Blackwell, Oxford, pp. 216–242.
Yiftachel, O., 1995b. Planning as control: policy and resistance in a deeply divided society. Progress in Planning 44 (2), 115–184.
Yiftachel, O., 1998. Planning and social control: exploring the dark side. Journal of Planning Literature 12 (4), 395–406.
Yiftachel, O., Huxley, M., 2000. Debating dominance and relevance: notes on the 'communicative turn' in planning theory. International Journal of Urban and Regional Research 24 (4), 907–913.
Žižek, S., 1998. Cyberspace, or, how to traverse the fantasy in the age of the retreat of the big other. Public Culture 10 (3), 483–513.

Critical Psychology

Horst Gundlach, Adolf-Würth-Zentrum für Geschichte der Psychologie, Würzburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

During the student revolts in the late 1960s arose the requests for a leftist critique of so-called bourgeois psychology and for the construction of a new psychology in the service of the repressed proletariat, repressed women, and repressed ex-colonial peoples. High hopes were set on Marxism as the basis for both objectives, especially of the kind developed by the Frankfurt School. Various alternatives of psychology that followed a Marxist model came to be known under the term 'critical psychology.' The most prominent school of critical psychology originated in West Berlin under the leadership of Klaus Holzkamp. They adopted the Soviet modification of Marxism, Marxism–Leninism including historical and dialectical materialism, as their basis for what became known as Critical Psychology, capitalized. There are still other critical currents in psychology calling themselves 'critical psychology,' generally, however, without such a definitive philosophical underpinning and such a comprehensive scope.

The term 'critical psychology' has been applied to various approaches to psychology which, in the last decades of the twentieth century, typically share hardly anything more than a resistance to so-called 'mainstream or traditional psychology,' an affinity to the moral ingredient of the Marxist critique of capitalism, and an aspiration to fight for the oppressed and the excluded. The connection to Marxist thought was further mitigated in the twenty-first century.

Critical psychologies of whatever nuance are not connected to I. Kant's critical philosophy or to neo-Kantian approaches to psychology despite the fact that one of the authorities of critical psychology, S.L. Rubinshtein, started out as a student of neo-Kantianism in Marburg.

One division in the heterogenous stream of critical psychologies is named 'Critical Psychology.' This is a school of psychology that chose to spell its name with capital letters; it originated in the 1970s in West Berlin, and is centered in the work of K. Holzkamp. It describes itself as being Marxist–Leninist, scientific (in a Marxist sense), and critical of what it calls bourgeois psychology. Critical Psychology will be the primary object of this article as it is so far the branch of critical psychology with the most noticeable degree of social coherence and of theoretical consistency.

Origins and Early History of Critical Psychology

The second half of the 1960s witnessed a rapid increase in the university student population in the Western world. The resulting tension and uneasiness combined with the televised atrocities of the Vietnam War to produce an international agitation for global revolutionary changes in a world engaged in Cold War.

An important factor in the student revolt, especially in Germany, and also in the formation of critical psychology and Critical Psychology, was the Critical Theory of the Frankfurt School, a strain of Marxism embodied by T.W. Adorno, M. Horkheimer, H. Marcuse, and J. Habermas. Especially pertinent at universities was their critique of positivism in sociology and in other social sciences, and generally of positivism in the sciences as prominently advocated by K. Popper. The positivist emphasis on the priority of observable events and empirical data over hypothetical constructs and theoretical terms was criticized as a deceptive disguise for sinister theory, especially for reactionary ideology in the service of capitalism. Therefore, the key reproach against positivism was its presumed lack of scientific relevance and of relevance for social change.

The Frankfurt School itself was not interested in the academic discipline of psychology. But psychology students watched the positivism controversy and applied it to their own field. A student Action Committee Critical Psychology disturbed the customary serenity of the congress of the German Society for Psychology at Tübingen in 1968, and, in order to coordinate various local initiatives, students met for a singular and tumultuous Congress of Critical and Oppositional Psychology in Hanover in 1969.

The term 'critical psychology' appeared for the first time as a book title *Kritische psychologie* in 1969. This is an anthology compiled by Bochum University psychology students comprising pirated texts by T.W. Adorno, W. Reich, K. Holzkamp and others, and various resolutions exchanged at the Hanover Congress and at earlier occasions. Sold in many psychological institutes, it established the term as a trademark for revolt, and stepped up rebellious behavior which few university teachers could stomach. The concept of critical psychology covered a multitude of different convictions, from an engagement for the weak and the betterment of society, to the use of psychology in a Marxist class struggle, to the conviction that psychology is unredeemably reactionary and should be dumped on the dungheap of history. Views of the future of psychology varied accordingly.

Among the few university professors who engaged in extended dialog with critical psychology students was Klaus Holzkamp (1927–95), holder of a chair of psychology at West Berlin's Free University. He had, in 1964 and in 1968, published a critique of psychological experimentation wherein he denounced the lack of criteria for the evaluation of the relevance of psychological experiments for psychological theory. He tried to improve the diagnosed situation by following the constructivist philosophy of science of

H. Dingler, the most prominent European proponent of operationalism.

In an invited speech at the Berlin section of the German association of professional psychologists in I968, Holzkamp applied the concepts of technical and emancipatory relevance, which he had taken from J. Habermas, to psychology and declared emancipatory relevance superior to all other forms of relevance. Through his discussions with critical students, Holzkamp was introduced to an analysis of society and its dominance relations according to Marxist lines. In his collection of essays, *Critical Psychology* (Holzkamp, 1972), he called his change in outlook the *Critical-Emancipatory Turn of Psychology*. Activity at the West Berlin psychological institute was characterized by studies in the history of psychology, which strove to uncover its quality as a tool of oppression in the ongoing class struggle (Geuter and Mattes, 1984), and by practical engagement for the underprivileged. This approach to the practice of psychologists remained what Holzkamp later called 'voluntaristic,' that is, appealing to the good will of psychologists and a willful decision for partisanship for the weak and needy.

As the antiauthoritarian and nonconformist movement slowly became organized by political parties and splinter groups in the early 1970s, so were the critical psychologists at the psychological institute in West Berlin Free University. One faction drifted into the sphere of influence of the Marxist–Leninist party (SED) ruling East Germany (DDR) and its West Berlin vassal party (SEW); another faction gathered around Beijing-oriented Maoist lines; others become Trotskyites, while a good many remained nonaligned. The psychologists around Holzkamp enlisted in the Marxist–Leninist camp. They withdrew from dialog with other factions, fashioned their own publication outlets and, to emphasize their distinctness, from the middle of the 1970s spelled their field as Critical Psychology. This group has also been known as the Holzkamp School, or the Berlin critical psychologists, despite other critical psychologists protesting that it was merely one offshoot of a much broader critical psychology unjustifiably trying to usurp the appellation and to eclipse the other approaches. The Marxist–Leninist turn has led Critical Psychology indubitably away from Adorno, Critical Theory, and the New Left.

Creation of Capitalized Critical Psychology

Capitalized Critical Psychology defines itself as a Marxist–Leninist psychology (Holzkamp, 1983: p. 27). It looks toward the writings of K. Marx, F. Engels, and V.I. Lenin for orientation, and adopts as its scientific and epistemological foundation or superstructure dialectical materialism and historical materialism. It embraces as pioneering works in this kind of psychology the publications of the historicocultural approach devised during Stalin's reign by L.S. Vygotsky and his disciples, in particular, S.L. Rubinshtein and A.N. Leont'ev. These authors were *de rigueur* for prefaces to be considered ideologically impeccable in the psychology of the DDR and the USSR, even though few serious scientists paid more than lip service to the ideological superstructure.

Holzkamp's turn away from critical emancipatory psychology toward Critical Psychology became visible with his book on *Sensory Cognition* (1973), where K. Marx is the most often cited author, and *Das Kapital* provides the model of human consciousness as dependent on and determined by the material and historical foundation of economic and social practice. Traditional psychology, now named bourgeois psychology, however, is perceived as erroneously proclaiming the independence of consciousness from this material basis.

Critical Psychology had not only an ideological superstructure but also a rarely considered material basis as well, namely the chair of psychology at the Free University in West Berlin held by Holzkamp and the power and privileges that went with it. A German university professor is a civil servant, but at the same time enjoys a very high degree of academic freedom in research and in teaching, as guaranteed by the constitution. This freedom, however, may have one limitation. A professor of psychology is obliged to train and examine future psychologists according to the canon of the examination Regulations.

Critical psychology's critique of traditional or bourgeois psychology had purportedly unveiled the latter's theoretical hollowness and its practical use as a weapon of the ruling classes against the proletariat. This brought Holzkamp into a 'dilemma' (1973: p. 13). Either he could go on teaching and training in bourgeois psychology and thereby lose his credibility as a critical psychologist, or he could stop training psychologists altogether, limit himself to criticizing psychology, and face the loss of his professorship and its benefits. He resolved this 'fatal alternative' (1973: p. 15) in a way that he thought allowed him to preserve critical credibility and to keep the material basis as well. He devised Critical Psychology, a new brand of psychology professedly consistent with the results of the critique of traditional psychology and also suitable as a foundation for the training of future practical psychologists.

Holzkamp and his coworkers had to prove that Critical Psychology could deliver. They formed a secluded group and started out to present to the public the various areas of psychology mentioned in the Examination Regulations in a new fashion according to Critical Psychology.

Organization of the School of Critical Psychology

Critical Psychology organized itself as a school of psychology with defined teacher–pupil relations and clear boundaries between insiders and outsiders. Its center was the chair of psychology occupied by Holzkamp, who was surrounded by assistants, doctoral students, and students. The notable names are gathered in the reader *Critical Psychology* (Tolman and Maiers, 1991). The West Berlin case remained unique. Critical Psychology could never gain a similar foothold in another university institute.

A series of textbooks, *Texte zur Kritischen Psychologie*, started to appear in 1973. It was supplanted in 1977 by another series, *Studien zur Kritischen Psychologie*, which appeared in the publishing house of Marxist–Leninist orthodoxy. The journal *Forum Kritische Psychologie* began to appear irregularly in 1975, and was turned into a biannual journal in 1977. A congress of Critical Psychology was held in the West German university town of Marburg in 1977; further congresses, now named

'International,' were held there in 1979 and 1984, and a fourth congress, not 'International,' in Berlin in 1997.

The 1980 and 1990s seem to have been a time of stagnation for Critical Psychology. But steps were undertaken to introduce an international public to Critical Psychology (Tolman and Maiers, 1991; Tolman, 1994).

In the 2010s, Critical Psychology is still taught at some places. The Psychological Institute at the Berlin Free University harbored until recently a subdivision for *Research on the Subject and Critical Psychology*, and still has a *Klaus Holzkamp Institute for the Science of the Subject.* Holzkamp is respected as the founding father of Critical Psychology, and an at-present unfinished edition of his *Collected Works* (*Schriften*) in seven volumes is in preparation at the *Institute for Critical Theory* (Holzkamp, 1997, etc.). His *Grundlegung der Psychologie* (*Foundation of Psychology*, Holzkamp,1983) is considered authoritative in the field, and in order to enhance the reader's understanding of his cloudy prose, a separate introduction into the matter has been published (Meretz, 2012). The latest developments may be followed at www.critical-psychology.de or www.kritische-psychologie.de.

Themes and Topics of Critical Psychology

Critique not only has to scrutinize the 'bourgois' or mainstream psychology but also must reflect the subject matters of psychological research, because these are not unconditional givens, but are the products of the economical and societal conditions under which individuals come into being and are formed.

Work done in Critical Psychology is to an uncommon degree conceptual work. Basic categories of psychology are put under scrutiny in the so-called functional-historical analysis of categories with the aim of reconstructing them under more satisfactory conditions. Psychological categories (or rather what they contain), like perception, motivation, learning, emotion, and cognition, are considered in their phylogenic, historical, and ontogenic evolution. At the end of this process, newly defined or rehabilitated categories are offered for the construction of an encompassing system of psychology.

Functional-historical analysis blends two things. One is C. Darwin's theory of evolution. It is utilized mainly for armchair paleozoology and paleoanthropology, and amplified and revised by the other constituent, Hegelian dialectics in its specific modification formulated by K. Marx, F. Engels, and later luminaries of Dialectical Materialism like G.V. Plekhanov. Functional-historical analysis, therefore, includes controversial Marxist tools of thought like negation in nonconceptual matters (antithesis), contradiction as causal agent (thesis–antithesis clash), or qualitative leaps or transitions (resulting in syntheses as unities of opposites, and forming the new theses for more dialectics). Such qualitative leaps are, among others, the transition from prepsychical to psychical organisms, the differentiation of special psychical functions, the evolution of learning, and the emergence of the societal nature of human beings.

These tools of thought are usually disregarded in science. It is therefore difficult for the reader not trained in their applications to follow the train of cogitations in which they are used. It is also not feasible to paraphrase them adequately in non-dialectical prose, and therefore, it may be fitting to let Critical Psychology speak for itself. This is a summary of the five steps of the functional-historical method according to Holzkamp (1983: p. 78):

1. Identify the real historical dimensions within the preceding developmental stage of the organism at which the quantitative transformation in question took place; that is, determine the 'position' that is dialectically 'negated' by the qualitative transformation and thereby bring the specificity of the new developmental stage into relief.
2. Identify in the external conditions the objective alterations that constitute the 'environmental pole' of the inner developmental contradiction that causes the new quality to emerge. Such alternations involve a 'moderate discrepancy' that, on one hand, demands 'compensations' from the organismic system, yet is supportable within its capacity. If it is not, the overall organism–environment system collapses, the inner contradiction is turned into an external opposition, and its antipodes will no longer be mediated as 'poles' in further development.
3. Identify the change in function of the relevant dimension demonstrated in Step 1 as the 'organism pole' of the developmental contradiction and with it the origin of the first qualitative leap, that is, the development of the specificity of the new function under the altered conditions. This dialectical negation of a previously prevalent function does not determine the overall process but, in a sense, still serves to maintain the system at the earlier stage.
4. Identify the change of dominance between the earlier function characteristic of the system's maintenance and the new function, which then determines the specifics of the system. (This change presents itself as a discontinuous reversal of the relationship between two continuously changing dimensions.)
5. Identify the ways in which the overall development of the system is restructured and assumes new direction (that is, analyze the 'specific secondary alterations,' and so on) after the qualitatively more specific function has become determining for the system's maintenance. With this identification of the qualitatively new dimensional structure, in which further qualitative transformations will occur, follows a return to the first step of the analysis but at a higher level (Tolman and Maiers, 1991: p. 39f.).

Categorical analysis leads to three characteristics of the human societal mode of existence: action possibility, action potency, and subjective situation. The latter is a personal assessment of one's own action possibilities and of where one stands in a particular objective situation. This derivation leads to a novel understanding of subjectivity and of the famous quote from Marx, "The essence of man is conscious activity."

In personality psychology and psychotherapy, Critical Psychology tries to utilize Freudian psychoanalysis. After a deft functional-historical analysis of its categories and of the fumbling efforts of Freudo-Marxists, it presents itself as the Marxist alternative to psychoanalysis.

The enthusiasm of early critical psychology to investigate the history and function of psychology is not followed by Critical Psychology. There is no critical historical analysis of the

fate of psychology and psychologists in the USSR, especially during the times of Stalinism and anticosmopolitanism when Vygotsky initiated the historicocultural approach, and Rubinshtein and Leont'ev became its proponents. Nor is the practical application of psychology in the USSR or DDR, particularly in their armed forces and secret services, examined.

Critique of Critical Psychology

There have been numerous controversies around Critical Psychology, among them the following.

Holzkamp and the not-yet capitalized critical psychology came under attack by H. Albert and others (Albert and Keuth, 1973), who championed a Popperian position for psychology and the social sciences. The ensuing debate in psychology reiterated the main positions of the positivism controversy between Popper and the Frankfurt School.

Holzkamp's turn to Critical Psychology and the associated Marxist–Leninist position provoked the criticism of other tendencies in the spectrum of critical psychology. This resulted in a number of more or less ephemeral publications, the most accessible of which is edited by Busch et al. (1979). Points of criticism were the usurpation of the name 'critical psychology,' the political orientation on Marxist–Leninist lines, and a perceived retrocession to bourgeois conceptions of psychology.

Disapproval of Critical Psychology by psychologists working in the DDR could not be uttered before the collapse of Real Socialism and the Soviet empire. But then they could report that they felt a virulent distaste for it as DDR party officials reprimanded their local psychologists for staying aloof from party doctrine and advised them unpleasantly to follow the ideologically blameless example of Critical Psychologists.

Other Critical Psychologies

It is not appropriate to call Critical Psychology 'German critical psychology,' as T. Teo (1998) does. There are and always have been other critical psychologists in Germany who have no affiliations with Critical Psychology. They gathered around the journal *Psychologie und Gesellschaft*, which was established in 1977 and renamed *Psychologie und Gesellschaftskritik* in 1978. Rexilius (1988) assembled some of the nonaligned voices.

An early voice in critical psychology was that of Peter Brückner (1922–82). From 1967, he occupied the chair of psychology at the Technical University in Hanover. There was no professional training of psychologists. Brückner did not publish systematic treatises on his view of psychology or on standard examination topics, but he authored a *Social Psychology of Capitalism* (1972), and papers on political psychology and on psychoanalysis (1983). Although he had some disciples, they did not organize a school.

A comparative portrayal of three early critical psychologies, Vygotsky's, Holzkamp's, and K.F. Riegel's dialectical psychology, was published by Ijzendoorn et al. (1984).

Another kind of critical psychology arose at Salzburg, Austria, in the footsteps of Igor Caruso. It attempts to eventually combine Marxism and Freudian psychoanalysis, an endeavor carried on by leaps and bounds since the 1920s, most prominently by Wilhelm Reich. The contemporary labor in this vast field is presented in a *Kritische Psychologie* (Abl, 2007), which is critical of Holzkamp's approach.

There is a tendency in English-speaking countries for various New Left tendencies to rally under the name of 'critical psychology.' Information on present trends may be found in Fox et al. (2009), Sloan (2000), the *Annual Review of Critical Psychology* (founded in 1999), the *International Journal of Critical Psychology* (2000 till 2008), continued in *Subjectivity, International Journal of Critical Psychology* (founded in 1986), and the *Journal of Critical Psychology, Counseling and Psychotherapy* (since 2001).

Concluding Remarks

At present, three hardly compatible currents that make use of the expression 'critical psychology' may be observed.

There is the capitalized Critical Psychology created by Holzkamp. Its trademark is the application of the Marxist–Leninist conceptual toolbox of dialectical and historical materialism described above. Its established introductory textbook is the *Introduction into Critical psychology* by M. Markard (2012), now in its fifth edition.

There is a less circumscribed critical psychology that is carried by a consensus that a Marxist inspired moral critique of capitalist societies, sporadically combined with the moral portion of Lenin's critique of imperialism, could bring about a psychology for the betterment of the human condition. Here, the theoretical pillars of Marxism like the labor theory of value, the theory of class struggle as the motor of social change, the theory of the approaching collapse of capitalism, and dialectical and historical materialism seem to be have lost much of their plausibility and currency. Sometimes this current is called 'radical psychology,' voluntaristic impulses seem to rule, and constructivism and feminism are typical influences.

There is a nonspecific employment of the expression 'critical psychology' as an abbreviation for multifaceted criticism regarding the theories, methods, and the usefulness or social relevance of the science and the discipline of psychology. This type of critical psychology is to be found all over the globe in such diverse countries like Italy (Colucci and Montali, 2013), Japan (Igarashi et al., 2013), and Sri Lanka (Vithanapathirana, 2013) or Brazil (Lacerda, 2013). It is difficult not to generate the notion that any sensible psychologist in the history of psychology has already contributed her/his share of what this current calls critical psychology.

From what can be known at present about the not-yet published international fourth volume of *Encyclopedia of Critical Psychology* under the editorship of Teo (2014), it seems fair to regard it as a commixture of the second and the third currents of critical psychology.

The various versions and mutations of critical psychologies present worthwhile case-study material for research in the predicament of the discipline of psychology and the dialectics between psychology as a science and as a profession. When in the twentieth century psychology was hastily metamorphosed into a practical profession, the university institutes had to be converted from pure research institutions into training

establishments – and they had to somehow cope with the egregious gap between the high expectations of application-minded students and the comparatively limited amount of effective tools psychology had to offer. For some of the discontented, Marxism–Leninism or pure Marxism promised to be the answer for the twentieth century. In the twenty-first century, while still receptive to calling one's own stand 'critical,' these theoretical edifices seems to have lost much of their past appeal.

See also: Critical Theory: Contemporary; Critical Theory; Field Theory; Frankfurt School: Institute for Social Research; Hermeneutics; Marx, Karl (1818–83); Marxism–Leninism: The Ideology of Twentieth-Century Communism; Representations, Social Psychology of; Social Constructionism; Social Psychological Theory, History of; Social Psychology; Vygotsky, Lev Semenovic (1896–1934).

Bibliography

Abl, G., 2007. Kritische Psychologie. Eine Einführung. Schmetterling Verlag, Stuttgart, Germany.

Albert, H., Keuth, H. (Eds.), 1973. Kritik der kritischen Psychologie. Hoffmann und Campe, Hamburg, Germany.

Brückner, P., 1972. Zur Sozialpsychologie des Kapitalismus: Sozialpsychologie der antiautoritären Bewegung. Europäische Verlagsanstalt, Frankfurt am Main, Germany.

Brückner, P., 1983. Zerstörung des Gehorsams: Aufsätze zur politischen Psychologie. Wagenbach, Berlin.

Busch, T., Engelhardt, W., Geuter, U., Mattes, P., Schulte, D., 1979. Zur Kritik der Kritischen Psychologie. Psychologie, Erkenntnistheorie und Marxismus. Oberbaumverlag, Berlin.

Colucci, F.P., Montali, L., 2013. The origins, characteristics and development of critical psychology in Italy. Annual Review of Critical Psychology 10, 596–621.

Fox, D., Prilleltensky, I., Austin, S., 2009. Critical Psychology: An Introduction, second ed. Sage, London.

Geuter, U., Mattes, P., 1984. Historiography of psychology in West Germany: approaches from social history. Storia e Critica della Psicologia 5, 111–126.

Holzkamp, K., 1972. Kritische Psychologie: Vorbereitende Arbeiten. Fischer-Taschenbuch-Verlag, Frankfurt am Main, Germany.

Holzkamp, K., 1973. Sinnliche Erkenntnis – Historischer Ursprung und gesellschaftliche Funktion der Wahrnehmung. Athenäum Fischer, Frankfurt am Main, Germany.

Holzkamp, K., 1978. Gesellschaftlichkeit des Individuums: Aufsätze 1974–1977. Pahl-Rugenstein, Cologne, Germany.

Holzkamp, K., 1983. Grundlegung der Psychologie. Campus Verlag, Frankfurt am Main, Germany.

Holzkamp, K., 1997. Schriften (Writings, 7 vols. planned, 5 published). Argument-Verlag, Hamburg, Germany.

Igarashi, Y., et al., 2013. Critical psychology in Japan: voices for change. Annual Review of Critical Psychology 10, 622–641.

Ijzendoorn, M., van, H., Veer van der, R., Goossens, F.A., 1984. Main Currents of Critical Psychology: Vygotskij, Holzkamp, Riegel. Irvington, New York.

Kritische Psychologie, 1969. Fachschaft Psychologie. Bochum University, Germany.

Lacerda, F., 2013. Critical psychology in Brazil: a sketch of its history between the end of the 20th century and the early 21st century. Annual Review of Critical Psychology 10, 110–149.

Markard, M., 2012. Einführung in die Kritische Psychologie: Grundlagen, Methoden und Problemfelder marxistischer Subjektwissenschaft, fifth ed. Argument Verlag, Hamburg (first edition 2009).

Meretz, S., 2012. Die »Grundlegung der Psychologie« lesen: Einführung in das Standardwerk von Klaus Holzkamp. Books on Demand, Norderstedt, Germany.

Rexilius, G. (Ed.), 1988. Psychologie als Gesellschaftswissenschaft: Geschichte, Theorie und Praxis kritischer Psychologie. Westdeutscher Verlag, Opladen, Germany.

Sloan, T. (Ed.), 2000. Critical Psychology: Voices for Change. St. Martin's Press, New York.

Teo, T., 1998. Klaus Holzkamp and the rise and fall of German critical psychology. History of Psychology 3, 235–253.

Teo, T. (Ed.), 2014. Encyclopedia of Critical Psychology, 4 vols. Springer-Verlag, Heidelberg, Germany (To be published in February 2014).

Tolman, C.W., 1994. Psychology, Society and Subjectivity: An Introduction to German Critical Psychology. Routledge, London.

Tolman, C.W., Maiers, W. (Eds.), 1991. Critical Psychology: Contributions to an Historical Science of the Subject. University Press, Cambridge, UK.

Vithanapathirana, M., 2013. Critical psychology in Sri Lanka: the Buddhist perspective. Annual Review of Critical Psychology 10, 870–885.

Critical Race Theory

Angela P Harris, University of California, Davis, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Critical race theory is a body of work that first emerged in American legal scholarship in the late 1980s and has since spread to other disciplines. It investigates a paradox: How does racism persist despite its nearly universal condemnation by state policy and by the norms of polite society? Rejecting the conventional liberal position that racism survives only as a relic from a less-enlightened time or as a characteristic of poorly educated or troubled individuals, critical race theorists take the position that racism is ordinary and normal in contemporary society, indeed perhaps integral to social practices and institutions.

Critical race theory investigates a paradox: How does racism persist despite its nearly universal condemnation by state policy and by the norms of polite society? Rejecting the conventional liberal position that racism survives only as a relic from a less-enlightened time or as a characteristic of poorly educated or troubled individuals, critical race theorists take the position that racism is ordinary and normal in contemporary society, indeed perhaps integral to social practices and institutions. Critical race theory can thus be understood as a study of hegemony: how domination can persist without coercion. It can also be understood as a study of collective denial.

Critical race theory originated in American legal studies during the post-Civil Rights Movement period, out of a sense of frustration that the era's 'Second Reconstruction' had failed to extirpate racism (see Crenshaw, 1988; Ansley, 1989). Its central target of critique has been state law, and its primary methodological innovations in legal scholarship have been the use of storytelling, fictional or anecdotal, to criticize legal reasoning and legal doctrine; and a method of reading against the grain that refuses to take legal doctrines at face value. Substantive themes of critical race theory in law include a sustained attack on the US Supreme Court's contemporary jurisprudence of race, particularly the position that the state can and should be colorblind; attempts to theorize unconscious racism or implicit bias as well as structural racism; the examination of race relations beyond the black/white paradigm; and efforts to understand how race interlocks with other forms of oppression. Future directions for critical race theory include its migration into and collaboration with other disciplines and scholarly fields, and an interest in placing US race relations into a global political-economic context.

In contemporary US social life, it is popularly accepted that racism is a form of injustice and that state law has the moral obligation not only to refrain from perpetrating racist acts against citizens but to provide redress for certain victims of nonstate racism as a matter of public policy. Most debates over antidiscrimination law and policy focus on the scope of these obligations. With respect to the state's obligation not to perpetrate racism itself, a question is whether antidiscrimination norms are satisfied when the state eschews all race-conscious state and nonstate public action, or whether the state should undertake race-conscious actions in certain circumstances in order to remedy its own past and continuing racial discrimination. With respect to the state's obligation to remedy nonstate discrimination, questions include to what extent the state should remedy private discrimination at all, and where the lines between public and private behavior should be drawn. These debates also require a definition of what 'discrimination' means.

Writers who associate themselves with critical race theory take the position that racism is ordinary and normal in American society (Delgado, 1994), indeed perhaps integral to it. Accordingly, critical race theorists attempt to show how contemporary law – including contemporary antidiscrimination law – paradoxically accommodates and even facilitates racism. The critical race theory project inevitably calls into question fundamental assumptions about legal institutions and legal reasoning, about the trajectory of American history, and about the nature of injustice.

One tendency within critical race theory, deriving from the tradition of civil rights activism and jurisprudence, is strongly reformist in bent and seeks to galvanize moral outrage and direct it toward constructive legal change. A second influence on critical race theory is the poststructuralist tradition, within which the very possibility of overcoming racism through discursive structures such as law is considered a false hope. Critical race theory thus reflects a distinctive mix of reformist zeal and critical pessimism (A. Harris, 1994). For example, Derrick Bell has urged civil rights activists to accept that, although there is a moral duty to challenge racism, racism is a permanent feature of American society (Bell, 1992). Other scholars influenced by critical race theory pursue the hope that innovations such as the theory of 'implicit bias' can successfully reshape antidiscrimination law into a more effective tool against racism (Kang et al., 2012).

Origins of Critical Race Theory

Critical race theory emerged in the United States in the 1980s as a product of political, intellectual, and sociological developments in American legal academia. Politically, the 1980s were a time when many American civil rights activists and left-wing legal scholars felt themselves caught up in a conservative backlash against the gains of the 1960s. The scholars who would later call themselves critical race theorists sought explanations for why formal legal equality had produced only

modest success in improving the lived experience of most African Americans and other people of color, and why hopes for social integration with whites seemed largely to have faded (Crenshaw, 1988; Ansley, 1989).

Intellectually, the early 1980s saw the heyday of critical legal studies (CLS) in the American legal academy. CLS introduced into legal scholarship poststructuralist ideas that had already permeated other disciplines focused on the interpretation of texts, including philosophy, literary criticism, and anthropology. Within legal studies, CLS scholars challenged the divide between law and politics, arguing that legal rules were radically indeterminate. CLS called for the deconstruction or trashing of legal doctrine in order to expose the reification of legal concepts and the hegemonic function of traditional legal reasoning (Kelman, 1987). CLS also provided an intellectual opening for broad-based critiques of law based on hidden norms of race and gender exclusion. For example, the first collection of essays in critical race theory appeared under the title *Minority Critiques of the Critical Legal Studies Movement*; its authors were CLS sympathizers who nonetheless were disappointed with CLS's failure to extensively engage with racism (Delgado, 1987; Williams, 1987).

Sociologically, the 1980s saw the gradual entry into legal academia of African Americans and other people of color, albeit in small numbers. Many of these new scholars were drawn to study race and the law, were frustrated by the disappointing outcomes of traditional avenues of legal reform in an era of backlash, and perceived traditional civil rights scholarship as a literature that ignored the voices of people of color (Delgado, 1984). Several key figures in what would become critical race theory were radicalized as students by the departure of Derrick Bell, one of the pioneering African American scholars critical of existing civil rights jurisprudence, from Harvard Law School because of its failure to hire an African American woman to the tenure-track faculty. These Harvard students created their own course on race, racism, and the law, and staffed it with visiting scholars from around the country. The movement known as *critical race theory*, and the term itself, emerged from this working group of students and teachers (Crenshaw et al., 1995).

Themes of Critical Race Theory

One of the earliest themes to appear within critical race theory was epistemological in nature and was reflected in a distinctive method. Why were people of color much more likely than whites to see American society as racist and to be pessimistic about the possibility of eliminating racism? Did they know something whites did not? Many critical race theorists argued that existing antidiscrimination law omitted the perspectives of the racially subordinated, and that subordinated groups in any social system have knowledge that the privileged lack. Thus, in a famous essay, Mari Matsuda argued that civil rights scholarship and practice should 'look to the bottom' of society for leadership and insight (Matsuda, 1987). Some scholars argued for the existence of a distinctive 'voice of color' that needed to be heard in public policy and legal scholarship (e.g., Johnson, 1991).

Critical race theorists also frequently abandoned the third-person objective voice in their legal writing in order to openly tell stories. For example, Derrick Bell's *Civil Rights Chronicles* featured an imaginary character who could travel through time and who used her magical powers to criticize American civil rights doctrine from the drafting of the Constitution to the present (Bell, 1985, 1992). The telling of personal anecdotes or fables was intended not only to convey ideas that could not be expressed in traditional scholarly language but also to criticize the assumption that norms of legal scholarship, legal reasoning, and legal institutions operated objectively and produced neutral assessments of public policy. Critical race theorists hoped to show that the legal scholar's 'perspectivelessness' (Crenshaw, 1989b) was in fact a perspective aligned with white privilege. Thus, for example, Patricia Williams, trained in poststructuralist theory and CLS, used personal stories, a dense web of allusions, and literary techniques to critique the social and legal construction of racial realities (P. Williams, 1991). The assertion of a distinctive voice of color proved controversial; critical race theory, like CLS before it, was bitterly attacked as racial 'special pleading' and as degrading the rigorous standards of argument and research established by traditional legal scholarship (see, for example, Kennedy, 1989). Legal storytelling was similarly attacked as sloppy scholarship and as an attempt to destroy the legacy of the Enlightenment. In this attack, critical race theory was taken to be a postmodernist assault on truth itself (see, for example, Farber and Sherry, 1997).

Drawing on the CLS tradition, critical race theorists have developed a distinctive style of reading against the grain, or, put another way, have adopted a hermeneutics of skepticism with respect to legal doctrine. This is often done by placing jurisprudential developments in a larger political and historical context in order to identify the continuity of racial oppression across time despite changing legal and political regimes. Thus, for example, Michelle Alexander argues that mass incarceration in the contemporary United States can properly be called 'the new Jim Crow' because it restricts the life chances of African Americans in the same way as *de jure* racial segregation in the pre-World War II South (Alexander, 2010). Ruth Gordon argues that the multilateral regime of the World Trade Organization has set sub-Saharan Africa up to remain poor and disorganized (Gordon, 2006). Derrick Bell argues that *Brown v. Board of Education*, the landmark constitutional case that struck down *de jure* segregation in elementary education, has become a historical artifact without contemporary significance given the resegregation of American schools (Bell, 2008). Critical race theorists thus set jurisprudence within a narrative of preservation through transformation – the idea that racism constantly changes in form but not effect (Siegel, 1997; see also Calmore, 1997). In this way, critical race theory represents a critique of the conventional liberal faith in social progress.

Most of the writing within critical race theory, however, has not taken on philosophical questions or even strayed very far from conventional methods of legal scholarship but rather has consisted of a sustained attack on the US Supreme Court's race jurisprudence, particularly its constitutional jurisprudence. For example, as one scholar has it, throughout the 1980s and 1990s the Court asserted its fidelity to four constitutional traditions

(Hayman Jr., 1995). First, the Court increasingly favored a position of colorblindness, under which all government racial classifications are presumptively unconstitutional, except perhaps when necessary to compensate a group of plaintiffs who can demonstrate the existence of conscious, deliberate racial bigotry that caused material disadvantage. Second, the Court frequently deferred to legislatures, to state and local governments, and to the discretion historically exercised by public actors such as criminal prosecutors and juries when these actors were charged with racial discrimination. Third, the Court deferred to private choices and free markets to produce the optimal allocation of resources. Fourth, the Court supported traditional indicators of merit when these were challenged as racially biased.

Critical race theorists have painstakingly attempted to show how these jurisprudential developments, separately and together, block the possibility of political, economic, and social redistributions that would disrupt systemic racism. Some critiques have focused on unpacking the history and rhetorical structure of the Court's contemporary race jurisprudence to show how it redirects but does not challenge racism (Haney Lopez, 2007). Other scholars have mounted sustained attacks on particular doctrinal developments. For example, Devon Carbado and Cheryl Harris argue that by accepting the principle that the government can legitimately employ race when it is enforcing immigration laws, the Supreme Court has sanctioned racial profiling against Latinos, facilitated the merger of criminal law and immigration law with an expansion of police and prosecutorial power, and weakened constitutional protections against search and seizure in ordinary policing (Carbado and Harris, 2011).

A second theme of critical race theory is the attempt to articulate more expansive theories of race and racism than the Court's focus on conscious, intentional prejudice. Some critical race theorists have turned to economics to show how structural incentives in sites such as the workplace perpetuate racial stereotyping and racial cartels (Carbado and Gulati, 2000, 2003, 2004; Roithmayr, 2010). Others draw on sociology, history, and/or geography to theorize institutional or structural racism (Haney Lopez, 2000; powell, 2008). A number of critical race theorists have drawn on cognitive psychology and the theory of implicit bias to develop a theory of unconscious racism (Krieger, 1995; Robinson, 2008; Kang, 2010). This work builds on an early influential article by Charles Lawrence III, who attacked the Supreme Court's use of conscious intent to discriminate as the touchstone for Fourteenth Amendment equal protection clause jurisprudence (Lawrence, 1987).

A third theme within critical race theory is a move beyond the traditional black/white paradigm in the study of race relations. Much traditional US civil rights jurisprudence, and much early critical race theory, focused on caste racism against African Americans. A number of other critical race theorists, particularly those associated with the Latina/o Critical Theory (LatCrit) movement, have criticized this focus (R. Williams, 1991; Perea, 1997), and explored the racialization of other nonwhite groups in the United States, including Latinos, Asian Americans, American Indians, and mixed-race people (e.g., Chin et al., 1996; Berger, 2009; Hernandez, 1998, 2007). These scholars identify nativist racism as a phenomenon distinct from, although related to, caste racism and explore the political and social triangulation of racial groups defined as neither black nor white (Chang and Aoki, 1998; Chang and Gotanda, 2006–07). They have sought to recover and examine aspects of American race history such as the Chicano struggle for civil rights, Asian exclusion, and the colonization of Puerto Rico that have been neglected because they do not fit the traditional emphasis on African American civil rights (Gomez, 2007; Haney Lopez, 2003; Chin, 1998; Roman, 1997). Finally, this strand of critical race theory explores the convergences and divergences between race law, immigration and naturalization law, and federal Indian law (Villazor, 2008, 2010; Johnson, 2008, 2010).

A fourth theme within critical race theory has been an effort to understand how racism interlocks with other forms of oppression, including gender, class, and sexuality. From the beginning, many critical race theorists regarded themselves as feminists, and several addressed the failure of mainstream feminism to respond to the interests and perspectives of women of color (Crenshaw, 1989a; A. Harris, 1990). Later writers in this tradition sometimes refer to themselves as 'critical race feminists' (Wing, 2003), and have produced a body of work that criticizes legal institutions and legal doctrine from the perspective of women of color. The work in this tradition has come to treat race as only one axis of a multidimensional system of status oppression (Hutchinson, 1997), and, drawing on an early article by Crenshaw, much of it goes under the label of 'intersectionality' (see Crenshaw, 1989a). The LatCrit offshoot of critical race theory has devoted itself explicitly to intersectionality, exploring the interplays of race, ethnicity, religion, national origin, class, and sexuality (Valdes, 1998). More recently, critical race scholarship has explored the relationship between race and class (Jones, 2009).

Each of these thematic strands within critical race theory has engaged with the effort to work out the implications for legal theory of understanding race as a social construction rather than a natural fact. In this broader project, some critical race scholars have imported the work of sociologists Michael Omi and Howard Winant on 'racial formation' into the legal academy, studying the influence of state law on the racialization of whites and Mexicans (Gomez, 2007; Haney Lopez, 1996, 1998). Others have traced the differing meanings of *race* within judicial discourse and within social and political discourse to show how policy and legal conflicts often involve an unacknowledged manipulation of the term (Gotanda, 1991; Hernandez, 1998).

Future Directions for Critical Race Theory

One recent development within critical race theory has been an interest in the flip side of minority oppression: white privilege. One aspect of this development is an interest in exploring white identity, facilitated by a growing literature on critical white studies (Delgado and Stefancic, 1997; Flagg, 1998) and an interest in how racial discrimination affects whites (Rich, 2010). Additionally, there is growing interest in taking racial privilege rather than discrimination as the

object of study and critique (Wildman, 2005, 1996). Here, the argument is that the traditional focus on racism's victims leaves unexamined the everyday ways in which those who are not victimizers nevertheless benefit from racial hierarchies. This development is in keeping with the turn away from intentional discrimination as the touchstone for racism.

A second recent development within critical race theory is an expansion into private law. Some participants in an emerging critical tax studies movement have imported the themes and concerns of critical race theory into their work (Moran and Whitford, 1996; Brown, 1997). A few scholars of corporate law have examined corporate law doctrine and racial diversity issues within corporations from a critical race theory perspective (Wade, 2002; Carbado and Gulati, 2004). Others have incorporated critical race theory's concerns into examinations of contract, tort, and property law (Houh, 2005; Chamallas and Wriggins, 2010; Brophy et al., 2011).

A third development within critical race theory is a movement beyond the borders of the United States. One example of this move is a growing literature in critical comparative studies of race and law (Oppenheimer et al., 2012; Hernandez, 2003; Wing, 2001). Another example is the incorporation of themes from critical race theory into the international law context (Richardson, 2008; Gordon, 1997) and into discussions of international human rights, foreign policy, and international trade relations (Hernandez-Truyol, 2008; Gordon, 2009; Gordon and Sylvester, 2004).

A final recent development is the migration of critical race theory into other fields of legal scholarship and other disciplines. Critical race theory explores a subject matter of interest to sociologists, political scientists, historians, and anthropologists. Laura Gomez has noted the overlap in particular between critical race theorists and law and society scholars, advocating more collaboration (Gomez, 2004). Other scholars have argued for greater collaboration between critical race theorists and empirical researchers generally, citing the benefits to both types of scholarship (Parks, 2008; see Twine and Warren, 2000). As previously noted, a fruitful emerging collaboration between critical race theory and cognitive psychology concerns the concept of implicit bias (Kang, 2010), although others lament that "a critical race psychology is not yet born" (Adams and Salter, 2011). Critical race theorists and economists have worked together (Carbado and Gulati, 2000, 2003, 2004; see also Ayres, 2007). Finally, ethnic studies, cultural studies, postcolonial studies, and American studies have concerns, subject matter, and methods closely allied with those of critical race theory, and scholars in these fields freely borrow from one another's work.

Critical race theory has also moved further afield, where it is sometimes referred to as critical race studies. In particular, it has taken off in the discipline of education (Ladson-Billings and Tate, 1995; Ladson-Billings, 2011). A handful of philosophers exploring questions of racial equality have also drawn on critical race theory in their work (e.g., Pateman and Mills, 2007).

Finally, at least one law school has established a critical race studies certificate program for its students with a program that seeks, among other things, to promote a more comprehensive and comparative approach to the study of race and law (C. Harris, 2002). This development may be a harbinger of a shift of critical race theory from the margins to the mainstream of research and teaching.

See also: Affirmative Action: Comparative Policies and Controversies; Civil Liberties and Human Rights; Civil Rights; Critical Legal Studies; Discrimination and the Law; Equality and Inequality: Legal Aspects; Feminist Legal Theory; Gender and the Law; Injustice: Legal Aspects; Justice and Law; Law and Society: Development of the Field; Law as Constitutive; Law as an Instrument of Social Change; Legal Scholarship; Postcolonial Law; Race and the Law.

Bibliography

Adams, G., Salter, P.S., 2011. A critical race psychology is not yet born. Connecticut Law Review 43, 1355–1377.

Alexander, M., 2010. The New Jim Crow. The New Press, New York.

Ansley, F.L., 1989. Stirring the ashes: race, class, and the future of civil rights scholarship. Cornell Law Review 74, 993–1077.

Ayres, I., 2007. Market power and inequality: a competitive conduct standard for assessing when disparate impacts are unjustified. California Law Review 95, 669–719.

Bell Jr, D.A., 1985. Foreword: the civil rights chronicles. Harvard Law Review 99, 4–83.

Bell Jr, D.A., 1992. Faces at the Bottom of the Well. Basic Books, New York.

Bell Jr, D.A., 2008. Racial equality: Progressives' passion for the unobtainable. Virginia Law Review 94, 495–520.

Berger, B.R., 2009. Red: racism and the American Indian. UCLA Law Review 56, 591–656.

Brophy, A., Lopez, A., Murray, K., 2011. Integrating Spaces: Property Law and Race. Aspen Publishers, New York.

Brown, D.A., 1997. Race, class, and gender essentialism in tax literature: the joint return. Washington and Lee Law Review 54, 1469–1512.

Calmore, J.O., 1997. Exploring Michael Omi's "messy" real world of race: an essay for "naked people longing to swim free." Law and Inequality: A Journal of Theory and Practice 15, 25–82.

Carbado, D., Gulati, M., 2000. Working identity. Cornell Law Review 85, 1259–1308.

Carbado, D., Gulati, M., 2003. The law and economics of critical race theory. Yale Law Journal 112, 1757–1828.

Carbado, D., Gulati, M., 2004. Race to the top of the corporate ladder: what minorities do when they get there. Washington and Lee Review 61, 1645–1693.

Carbado, D., Harris, C., 2011. Undocumented criminal procedure. UCLA Law Review 58, 1543–1616.

Chamallas, M., Wriggins, J., 2010. The Measure of Injury: Race, Gender, and Tort Law. New York University Press, New York.

Chang, R.S., Aoki, K., 1998. Centering the immigrant in the inter/national imagination. La Raza Law Journal 10, 1395–1447.

Chang, R.S., Gotanda, N., 2006–2007. The race question in LatCrit theory and Asian American jurisprudence. Nevada Law Journal 7, 1012–1029.

Chin, G.J., 1998. Segregation's last stronghold: race discrimination and the constitutional law of immigration. UCLA Law Review 46, 1–74.

Chin, G.J., Cho, S., Kang, J., Wu, F., 1996. Beyond self-interest: Asian Pacific Americans toward a community of justice, a policy analysis of affirmative action. Asian Pacific American Law Journal 4, 129–162.

Crenshaw, K.W., 1988. Race, reform, and retrenchment: transformation and legitimation in antidiscrimination law. Harvard Law Review 101, 1331–1387.

Crenshaw, K.W., 1989a. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine. University of Chicago Legal Forum, 139–167.

Crenshaw, K.W., 1989b. Foreword: toward a race-conscious pedagogy in legal education. National Black Law Journal 11, 1–14.

Crenshaw, K., Gotanda, N., Peller, G., Thomas, K., 1995. Critical Race Theory: The Key Writings that Formed the Movement. The New Press, New York.

Delgado, R., 1984. The imperial scholar: reflections on a review of civil rights literature. University of Pennsylvania Law Review 132, 561–578.

Delgado, R., 1987. The ethereal scholar: does critical legal studies have what minorities want? Harvard Civil Rights—Civil Liberties Law Review 22, 301–322.

Delgado, R., 1994. Bibliographic essay: critical race theory. SAGE Race Relations Abstracts 19, 3–28.

Delgado, R., Stefancic, J., 1997. Critical White Studies: Looking behind the Mirror. Temple University Press, Philadelphia, PA.

Farber, D., Sherry, S., 1997. Beyond All Reason: The Radical Assault on Truth in American Law. Oxford Press, New York.

Flagg, B., 1998. Was Blind but Now I See: White Race Consciousness and the Law. New York University Press, New York.

Gomez, L., 2004. A tale of two genres: on the real and ideal links between law and society and critical race theory. In: Sarat, Austin (Ed.), The Blackwell Companion to Law and Society. Blackwell Publishing, Oxford.

Gomez, L., 2007. Manifest Destinies: The Making of the Mexican American Race. New York University Press, New York.

Gordon, R., 1997. Saving failed states: sometimes a neocolonialist notion. American University Journal of International Law and Policy 12, 903–974.

Gordon, R., 2006. Sub-Saharan Africa and the brave new world of the WTO multilateral trade regime. Berkeley Journal of African American Law and Policy 8, 79–121.

Gordon, R., 2009. The dawn of a new, new, international economic order? Law and Contemporary Problems 72, 131–162.

Gordon, R.E., Sylvester, J., 2004. Deconstructing development. Wisconsin International Law Journal 22, 1–98.

Gotanda, N., 1991. A critique of our constitution is colorblind. Stanford Law Review 44, 1–68.

Haney Lopez, I.F., 1996. White by Law: The Legal Construction of Race. New York University Press, New York.

Haney Lopez, I.F., 1998. Race, ethnicity, erasure: the salience of 'race' to LatCrit theory. La Raza Law Journal 10, 57–125. California Law Review 85, 1143–1211.

Haney Lopez, I.F., 2000. Institutional racism: judicial conduct and a new theory of racial discrimination. Yale Law Journal 109, 1717–1884.

Haney Lopez, I.F., 2003. Racism on Trial: The Chicano Fight for Justice. Harvard University Press, Cambridge, MA.

Haney Lopez, I.F., 2007. 'A nation of minorities': race, ethnicity, and reactionary colorblindness. Stanford Law Review 57, 985–1063.

Harris, A., 1990. Race and essentialism in feminist legal theory. Stanford Law Review 40, 581–616.

Harris, A., 1994. Foreword: the jurisprudence of reconstruction. California Law Review 82, 741–785.

Harris, C., 2002. Critical race studies: an introduction. UCLA Law Review 49, 1215–1236.

Hayman Jr, R.L., 1995. The color of tradition: critical race theory and postmodern constitutional traditionalism. Harvard Civil Rights—Civil Liberties Law Review 30, 57–108.

Hernandez, T.K., 1998. Multiracial' discourse: racial classifications in an era of colorblind jurisprudence. Maryland Law Review 57, 97–173.

Hernandez, T.K., 2003. Multiracial matrix: the role of race ideology in the enforcement of antidiscrimination laws, a United States-Latin American comparison. Cornell Law Review 87, 1093–1176.

Hernandez, T.K., 2007. Latino inter-ethnic employment discrimination and the 'diversity' defense. Harvard Civil Rights—Civil Liberties Law Review 42, 259–316.

Hernandez-Truyol, B.E., 2008. The gender bend: culture, sex, and sexuality – a Latcritical human rights map of Latina/o border crossings. Indiana Law Journal 83, 1283–1331.

Houh, E.M.S., 2005. Critical race realism: re-claiming the antidiscrimination principle through the doctrine of good faith in contract law. University of Pittsburgh Law Review 66, 455–520.

Hutchinson, D.L., 1997. Out yet unseen: a racial critique of gay and lesbian legal theory and political discourse. Connecticut Law Review 29, 561–645.

Johnson Jr, A.M., 1991. The new voice of color. Yale Law Journal 100, 2007–2063.

Johnson, K., 2008. A handicapped, not 'sleeping,' giant: the devastating impact of the initiative process on Latina/o and immigrant communities. California Law Review 96, 1259–1297.

Johnson, K., 2010. How racial profiling became the law of the land: United States v. Brignoni-Ponce and Whren v. United States and the need for truly rebellious lawyering. Georgetown Law Journal 98, 1005–1077.

Jones, T., 2009. Foreword: race and socioeconomic class: examining an increasingly complex tapestry. Law and Contemporary Problems 72 (Fall), i–xii.

Kang, J., 2010. Seeing through colorblindness: implicit bias and the law. UCLA Law Review 58, 465–520.

Kang, J., Judge Bennett, M., Carbado, D., Casey, P., Dasgupta, N., Faigman, D., Godsil, R., Greenwald, A.G., Levinson, J., Mnookin, J., 2012. Implicit bias in the courtroom. UCLA Law Review 59, 1124–1186.

Kelman, M., 1987. A Guide to Critical Legal Studies. Harvard University Press, Cambridge, MA.

Kennedy, R.L., 1989. Racial critiques of legal academia. Harvard Law Review 102, 1745–1819.

Krieger, L.H., 1995. The content of our categories: a cognitive bias approach to discrimination and equal employment opportunity. Stanford Law Review 47, 1161–1248.

Ladson-Billings, G., Tate, W.F., 1995. Towards a critical race theory in education. Teachers College Record 97, 47–68.

Ladson-Billings, G., 2011. Race to the top, again… comments on the genealogy of critical race theory. Connecticut Law Review 43, 1439–1457.

Lawrence III, C.R., 1987. The Id, the Ego, and equal protection: reckoning with unconscious racism. Stanford Law Review 39, 317–388.

Matsuda, M.J., 1987. Looking to the bottom: critical legal studies and reparations. Harvard Civil Rights—Civil Liberties Law Review 22, 323–399.

Moran, B., Whitford, W., 1996. A black critique of the Internal Revenue Code. Wisconsin Law Review 1996, 751–820.

Oppenheimer, D., Foster, S., Han, S., 2012. Comparative Equality and Anti-discrimination Law: Cases, Codes, Constitutions and Commentary. Foundation Press.

Parks, G.S., 2008. Toward a critical race realism. Cornell Journal of Law and Public Policy 17, 683–745.

Pateman, C., Mills, C., 2007. The Contract and Domination. Polity Press, Cambridge.

Perea, J., 1997. Immigrants Out! The New Nativism and the Anti-immigrant Impulse in the United States. New York University Press, New York.

powell, J.A., 2008. Structural racism: Building upon the insights of John O. Calmore. North Carolina Law Review 86, 791–816.

Rich, C.G., 2010. Marginal whiteness. California Law Review 98, 1497–1593.

Richardson III, H.J., 2008. The Origins of African-American Interests in International Law. Carolina Press, Durham, NC.

Robinson, R., 2008. Perceptual segregation. Columbia Law Review 108, 1093–1180.

Roithmayr, D., 2010. Racial cartels. Michigan Journal of Race and Law 16, 45–79.

Roman, E., 1997. Empire forgotten: the United states' colonization of Puerto Rico. Villanova Law Review 42, 1119–1211.

Siegel, R., 1997. Why equal protection no longer protects: the evolving forms of status-enforcing state action. Stanford Law Review 49, 1111–1148.

Twine, F.W., Warren, J.W., 2000. Racing Research, Researching Race: Methodological Dilemmas in Critical Race Studies. New York University Press, New York.

Valdes, F., 1998. Under construction: LatCrit consciousness, community, and theory. La Raza Law Journal 10, 1–56.

Villazor, R.C., 2008. Blood quantum land laws and the race versus political identity dilemma. California Law Review 96, 801–837.

Villazor, R.C., 2010. Rediscovering Oyama v. California: at the intersection of property, race, and citizenship. Washington University Law Review 87, 979–1042.

Wade, C.L., 2002. Racial discrimination and the relationship between the directorial duty of care and corporate disclosure. University of Pittsburgh Law Review 63, 389–440.

Wildman, S.M., 1996. Privilege Revealed: How Invisible Preference Undermines America. New York University Press, New York.

Wildman, S.M., 2005. The persistence of white privilege. Washington University Journal of Law and Policy 18, 245–265.

Williams, P.J., 1987. Alchemical notes: reconstructing ideals from deconstructed rights. Harvard Civil Rights—Civil Liberties Law Review 22, 401–433.

Williams, P.J., 1991. The Alchemy of Race and Rights. Harvard University Press, Cambridge, MA.

Williams Jr, R.A., 1991. Columbus's legacy: law as an instrument of racial discrimination against indigenous people's rights of self-determination. Arizona Journal of International and Comparative Law 8, 51–75.

Wing, A.K., 2001. Polygamy from southern Africa to black Britannia to black America: global critical race feminism as legal reform for the twenty-first century. Journal of Contemporary Legal Issues 11, 811–863.

Wing, A.K., 2003. Critical Race Feminism: A Reader, second ed. New York University Press, New York.

Critical Rationalism

Volker Gadenne, Johannes Kepler University Linz, Linz-Auhof, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Critical rationalism is a philosophy of rational thought and action founded by Karl Popper. It arose out of Popper's idea that falsifiability is the characteristic of empirical science. According to critical rationalism, there is no certain knowledge. All human thoughts and actions are fallible. Yet rationality and knowledge are possible. Rationality is openness to criticism, that is, the readiness to listen to critical argument and evidence. Critical rationalism has been criticized as being a form of skepticism and incompatible with the history of science.

Critical rationalism (CR) is a philosophy of rational thought and action founded by Karl Popper. By rationalism, Popper meant the attitude to solve problems by an appeal to reason, that is, to clear thought, experience, and argument. Rationalism in this wide sense covers intellectualism (rationalism in the narrow sense) as well as empiricism. In the history of philosophy, the idea of rationalism usually implied that we can achieve certain knowledge and that we should accept only what has been proven. CR, however, denies that certain knowledge is possible. All hypotheses and theories – moreover, all human attempts of solving problems – are uncertain and fallible. Therefore, we should always be prepared to correct our views. According to CR, rationality is the readiness to listen to critical argument and to learn from experience. CR arose out of Popper's (1934) methodology of science, and has been developed further by some of Popper's successors.

The Methodology of Falsificationism

In 1934, Popper published his classical text *Logik der Forschung*, which was translated into English in 1959 as *The Logic of Scientific Discovery* (hereafter, LScD). In this book, Popper posed the *problem of demarcation*. What is the characteristic of empirical science? How can science be demarcated from metaphysics and pseudoscience, on the one hand, and logic and mathematics, on the other hand? Popper did not propose as a criterion of demarcation the verification or proof of empirical theories. Theories are systems of universal statements, and such statements cannot be verified by observation. For example, no matter how many white swans we have observed, this does not prove that all swans are white. With reference to Hume, Popper rejected as invalid all kinds of inductive inference. Instead he proposed as a suitable criterion of demarcation the *falsifiability* of empirical theories. The hypotheses and theories of science must be refutable by experience. For instance, 'All swans are white' is falsifiable, since it can be refuted by the observation of a black swan. Similarly, the universal statements of physics can be refuted by contradicting experimental results. The observational statements that serve to falsify a theory must themselves be intersubjectively testable (e.g., by repeating the respective experiment).

For scientific research, Popper advocated the *hypothetico-deductive method*. Theories are posed without prior justification, and then critically tested. From a theory T, with the help of some background knowledge A (assumptions regarded as unproblematic in the given situation, e.g., auxiliary hypotheses and singular statements), a test prediction P is deduced. P is compared with the results of observation. If P is accepted as true, T is regarded as *corroborated*, provided the test was a serious attempt to falsify T. If P is rejected, T is considered as falsified.

Theories should be *severely* tested. They should, in every conceivable way, be exposed to falsification. However, theories can always be saved from refutation by introducing ad hoc hypotheses. Any hypothesis or theory can be immunized against falsification. In order to prevent this, Popper formulated a set of methodological rules. These rules propose how to proceed in specific research situations, and together define the critical method of science.

It is a major point of Popper's methodology that corroboration is not mere correspondence between theory and observation. It is the result of a failed attempt to falsify a theory. The degree of corroboration increases with the number and kind of tests the theory has passed. Once a theory has passed some tests, additional tests of the same kind raise the degree of corroboration very little. The theory may, however, gain additional corroboration by application in new fields.

It is not possible to prove with certainty that a statement is true or false. This holds for statements of any kind, including observation statements. Human cognition is fallible. Certain knowledge is impossible. Popper noted that fallibilism was stated more than 2000 years ago by Xenophanes. However, the characteristic of Popper's view is not fallibilism by itself but the combination of fallibilism with rationalism. Unlike skeptics and relativists, who also embrace fallibilism, Popper insisted that rationality and knowledge are possible. In his LScD, he wanted to point out what it is that constitutes rationality in empirical science. The two central elements of falsificationism are fallibilism plus the assumption that proceeding in accordance with the rules of the critical method is conducive to scientific progress.

LScD contributed to the debates of the Vienna Circle, a group of philosophers and scientists around Moritz Schlick. Members of the Vienna Circle had proposed to consider as meaningful only statements that could be verified by observation. On this view, metaphysical statements turn out to be meaningless. Popper rejected this view, which would also render the laws of nature as meaningless since they cannot be verified either. Popper loosened the standard of scientific rationality by replacing verifiability with falsifiability.

Furthermore, he insisted that what is unfalsifiable is not necessarily meaningless. He even granted that metaphysical ideas may be fruitful. Yet he thought it necessary to demarcate science from metaphysics. He believed that rational discussion is possible in science but not in metaphysics.

From Falsificationism to Critical Rationalism

Roughly 10 years later, Popper was concerned with problems of social and political philosophy. In his *The Open Society and Its Enemies*, he criticized the political theories, and their philosophical roots, of Plato, Hegel, and Marx. He also defended what he called the *open society*, in which individuals are personally responsible for their decisions, and in which it is allowed to put forward criticism, including criticism of the government. Popper considered as the basic principle of democracy that institutions are organized in such a way that people can replace governments without resorting to violence. In politics, as in science, Popper suggested replacing questions of justification with questions of criticism. Instead of asking the traditional question 'Who should rule?' we should ask how to organize institutions so that we can get rid of bad rulers.

Popper also addressed problems of social engineering. He considered as most dangerous, and vehemently rejected, what he called utopian engineering, which means the reconstruction of a society as a whole in order to realize an ideal state. We do not possess the knowledge necessary to achieve such an ambitious aim. Social action has unintended consequences. Utopian engineering will probably destroy the existing institutions and may lead to a dictatorship. Popper regarded as rationally justified only *piecemeal engineering*, which is small scale and continuously corrected in the light of experience.

In this context, Popper was confronted with an epistemological difficulty (Wettersten, 1992, Chapter 9). He had stated in LScD what rationality is in empirical science. But how could he argue for or against assumptions that are not empirical theories (e.g., assumptions of social philosophy or ethics)? And how can methodology itself claim rationality if its statements are not falsifiable, as are the statements of science?

Popper had loosened the standard of scientific rationality from proof to falsifiability, but he lacked a criterion of rationality applicable in other areas than science. Falsifiability could not serve that purpose. It would be too high a standard of rationality in metaphysics, ethics, or social philosophy. In his *The Open Society and Its Enemies*, Popper defines rationality as an attitude of *being open to criticism*. Rationality is the readiness to listen to critical argument and to learn from experience. It is "an attitude of admitting that I may be wrong and you may be right, and by an effort, we may get nearer to the truth" (Popper, 1945, vol. 2, p. 213). The idea of being open to criticism is, of course, somewhat vague. It has to be specified for the respective domain in which it is applied. It is a central concern of the CR program to specify what it means (e.g., in science, politics, or law) to proceed in accordance with the critical method.

Popper conceded that the rational attitude cannot itself be proven by argument, since only those who have adopted it will be impressed by argument. He concluded that rationality must be based on a decision, and on a faith in reason that must itself be described as irrational.

In later works, Popper (1963, Chapter 8) explained how philosophical or metaphysical theories can be rationally discussed even though they are not falsifiable. One can raise the question whether, and to what degree, a theory is able to solve the problems it is said to solve. Such an analysis may lead to the conclusion that a metaphysical theory does not solve any problem at all, or that one metaphysical theory is able to solve a problem better or solve more problems than another, and should therefore be preferred. Popper demonstrates, for example, how determinism and idealism can be assessed, and he rejected these views. Similarly, he argues against instrumentalism and for realism (Popper, 1963, Chapter 3).

So far, two components of CR have been pointed out. The first is *fallibilism*. The second is the *plea for rationality*, where rationality is conceived of as the critical attitude. Alternatively, one could state the second component by saying that the critical method is the best way of addressing problems of knowledge or action. According to Hans Albert (1999, Chapter 1), there is a third component of CR, namely, *realism*. Realism is conceived of as the views that there is an objective world existing independently of mind and language, and that we can achieve knowledge about this world. It is important to note here that, for CR, knowledge does not mean certain belief but corroborated yet fallible conjecture. For example, we know that the earth is round and not flat in the sense that the first hypothesis is highly corroborated while the second has been falsified.

In connection with realism, most critical rationalists adopt the *correspondence theory of truth*, the idea of truth as correspondence to the facts. That is, a statement is defined as true if and only if what it states is really the case. Otherwise it is false. This definition says only how truth and falsity are conceived of, and does not answer the methodological question of how truths and falsehoods are to be detected.

Fallibilism represents the (moderately) skeptical aspect of CR, and realism represents its optimistic aspect. Popper emphasized that there is no certainty, yet he declared truth (or truthlikeness) as the aim of science. CR vehemently rejects any kind of dogmatism or fundamentalism, and also rejects radical skepticism, relativism, and antirealism (including the various forms of constructivism).

CR is Popper's main contribution. But others have also contributed to CR as it exists today. Among them are Joseph Agassi, Hans Albert, Gunnar Andersson, William W. Bartley, David Miller, Alan Musgrave, and John Watkins. Imre Lakatos' (1970) methodology of scientific research programs is not regarded as a version of CR but as an alternative approach since it assumes that the hard core of a research program is immune to criticism, an assumption that is not in accordance with CR.

It should be noted that CR is not the same as Popper's philosophy in general. Popper wrote on many different subjects and developed many views that are independent of the basic assumptions of CR. Some of these views (e.g., his 'logic of situation' or his theory of the three worlds) are not shared by many critical rationalists.

Criticizing Critical Rationalism

Popper's *Open Society* became widely known and made its author famous. Some critics found his interpretation and

devastating critique of Plato and Hegel unconvincing. But his plea for the open society and for the critical attitude was widely accepted.

When Popper had developed CR, he had not yet reconsidered and improved his earlier methodology of falsification. This methodology was embraced by many scientists, but rather criticized by philosophers. The critics argued that scientific theories were not conclusively falsifiable since one could always decide to give up auxiliary hypotheses instead of the central laws of a theory. In the history of science, major theories were protected from falsification. Theories were retained, the empirical results that contradicted them were put into question, and often it turned out later that these results had been based on false assumptions (Chalmers, 1999, pp. 87–91). And, most importantly, such developments happened in periods in which great progress was achieved. CR was therefore said to be incompatible with the way famous scientists like Galilei and Newton proceeded. Popper's main critics who argued from the history of science were Thomas Kuhn (1962) and Paul Feyerabend (1975).

Some critical rationalists, such as Andersson (1994) and Musgrave (1999, Chapter 11), also studied the respective historical facts but came to another conclusion: Popper's methodology is quite in agreement with the history of science, and the criticism by Kuhn and Feyerabend is unfounded. Admittedly, scientists sometimes retain theories and reject results of observation, but this is quite in agreement with CR, for CR declares observations and auxiliary hypotheses as fallible.

Critics responded by asking what CR actually contributes if it allows rejecting either theories or observational results. Is such a methodology really different from Feyerabend's 'Anything goes'? Now CR commits to the hypothetico-deductive method. From a theory T and some background knowledge A (hereafter, T & A), a test statement P is deduced. P is compared with the results of observation. If P is rejected, some statement that belongs to T & A has to be given up, for, if non-P is true, T & A are necessarily false. The conjunction of T & A with non-P is inconsistent. Something has to be changed in order to regain consistency.

This is clearly different from 'Anything goes.' But it is not different from the view most philosophers of science hold. Does CR offer anything more specific than the hypothetico-deductive method?

Improving the Methodology of Critical Rationalism

The answer to the last question lies in the *methodological rules* of CR. But there is no single answer. One has to distinguish between Popper's view in LScD, his later view, and the views of other critical rationalists.

The hypothetico-deductive method leaves many decisions to the researcher. For example, any theory has infinitely many logical consequences. We may therefore ask, first, whether CR can give some advice regarding which consequence should be selected for testing. Second, if the result of the test is non-P, which part of the system T & A should be declared as false? This is known as Duhem's problem. Third, once it has been decided that T is to be regarded as falsified, is this judgment final or may it be revised?

As to the first question, Popper specifies that P not only should be derivable from T & A but also must not be derivable from other theories than T (Popper, 1959, p. 33). Even better, P should contradict other theories, especially those competing with T. Thus, P should be an implication from T that is extremely unlikely in the light of other views we have held true. We may call this the *rule of critical testing*, because the test is planned in such a way that the empirical result either falsifies T or teaches us that some well-corroborated theory T′ seems to contain a mistake. In each case, the result fosters the growth of knowledge.

Concerning the second question, Popper allows the introduction of new auxiliary hypotheses in order to retain the theory that is contradicted by observations. However, he sets some restrictions. No ad hoc hypotheses must be introduced. Popper also decides that new auxiliary hypotheses must not decrease the empirical content (or degree of falsifiability) of T & A, but rather should increase it (Popper, 1959, pp. 82–83). And he rejects the strategy of explaining away threatening observations by appeal to some unknown disturbances, or to the incompetence of researchers.

The last point seems uncontroversial. Saving theories from refutation with the help of untestable speculations does not further the growth of knowledge. However, the rule never to decrease the content of T & A appears to be too strong, and often not applicable. If non-P contradicts T & A, and T is retained, some assumption within A must be declared as false. The new system T & A′ must be consistent with non-P, and is therefore inconsistent with T & A. This, however, renders it very difficult to compare the two systems with respect to empirical content. (Comparing the content of systems is easy if they are consistent with each other and one of them implies the other one logically.) Furthermore, it is very restrictive to require that the empirical content must never decrease. Empirical content is an important aim in theory building, but there are also other aims (e.g., simplicity) and, primarily, correspondence to the empirical facts. In the course of scientific research, it is not always possible to achieve an increase in all dimensions of progress.

With respect to the third question, Popper wrote (1959, p. 268), "In general we regard an inter-subjectively testable falsification as final (provided it is well tested)." It is not easy to interpret what Popper wanted to say here. It seems that this statement is not in accordance with other parts of LScD or with the spirit of fallibilism. How can a falsification, that is, a decision to declare a theory as false, be final or conclusive if auxiliary hypotheses as well as observation statements are fallible? In his postscript to LScD, Popper made clear "that it is never possible to prove conclusively that an empirical theory is false" (Popper, 1983, p. xxii). This is the view to which all critical rationalists have agreed. Since observations and auxiliary hypotheses may be erroneous, the judgment that a theory is false is itself fallible and may be revised.

As to Duhem's problem, the answer is systematic trial and error. If non-P obtains, it is up to the researchers to speculate where the false assumption(s) is located. They have to decide to modify either T or A and to replace the respective assumption by a new one. The result, T′ & A, or T & A′, is then again tested in order to find out whether T′ & A, or perhaps T & A′, is empirically more successful than T & A. This procedure is continued

until the system is corroborated, or the whole theory T is given up in favor of a new approach. However, CR does not specify when exactly a theory should be given up. There is no criterion of final falsification. But CR assumes that science is quite successful in deciding when it is reasonable to give up a certain theory and replace it with a better one.

So far, some methodological rules of CR have been discussed. There are many further rules in Popper's works (Jarvie, 2001). Most of them can be interpreted as *heuristic principles*. They recommend thinking in a certain direction in order to solve the problem at issue, but they do not exactly prescribe what to do. For example, according to the rule of critical testing, one should try to derive from T a test statement that does not follow from other theories. This excludes many logical consequences from T but still leaves the selection of the best test statement to the imagination of the researchers. Hans Albert (1999, Chapter 2) therefore conceived of methodology as a kind of heuristics.

As to the logical status of methodological rules, Popper realized that they are not analytical statements and not empirical statements (descriptions of what scientists do). Popper (1959, p. 53) therefore declared them as conventions that define the game of science. Albert (1999, Chapter 2) instead proposed to regard methodology as a special kind of technology. The statements of methodology say that certain procedures are conducive to the aim(s) of science. Suppose the aim of science is the development of true theories of high explanatory power. The idea of critical testing may then be formulated as follows: 'Critical testing contributes to achieving true, powerful theories' (Gadenne, 2006).

The methodological rules cited here are considered as relevant for natural science as well as for the social and cultural sciences. Of course, the various disciplines use different research techniques. They are concerned with different objects, and they therefore need different methods of observation, measurement, and statistical analysis. However, the fundamental methodology is the same in all empirical sciences. The hypothetico-deductive method, and the specific rules associated with that method, can be used effectively in physics and biology as well as in the apprehension of the meaning of actions and the meaning of texts (Mantzavinos, 2005).

Thus, CR has more to offer than mere deductive thinking. Critical rationalists differ, however, in regard to the significance of methodological rules. Some think that the hypothetico-deductive method is the main answer to the questions of methodology, and the additional rules are of secondary importance. Others believe that Popper made an important step when he pointed out that deductive logic is not sufficient for methodology. There are many pragmatic aspects of research to which deductive logic has nothing to say but that are central for the rationality of science. They do not concern the logical properties of statements but the way scientists proceed when they test, discuss, and evaluate these statements.

Popper also introduced the *social* dimension into the methodology of science. He considered science as a social activity and was convinced that rationality requires cooperation. Robinson Crusoe could not do science, Popper said. One individual is not able to both develop and test a theory effectively. Jarvie (2001) pointed out that these thoughts laid the foundation for a theory of the institutions of science. It depends on social rules and on the institutions of science to which degree scientists can follow the methodological rules that CR declares to be conducive to progress. These institutions can foster or hinder rationality. For example, in some domains there is a tendency to appreciate more, and prefer for publication, data that confirm a hypothesis to data that refute a hypothesis. However, if 'mere criticism' (instead of 'positive results') is considered by a community as not worth being published, this hinders learning from experience. In this context, CR encourages the study of the institutions of science (Wettersten, 2006; Albert, 2010).

When Popper developed falsificationism, his main concern was the demarcation of science. Is this still a major point of CR? Popper had granted that metaphysical (or philosophical) ideas may be fruitful for science. Some falsifiable scientific theories arose out of metaphysical ideas. Atomism is an example. Later, Popper came to the conclusion that metaphysical theories could even be rationally assessed. Agassi (1975) went one step further. On his view, most scientific theories were attempts to give answers to metaphysical questions (e.g., 'From what is everything made?'). Metaphysics is also a source of ideas for science. It can guide scientific research by giving standards for desirable theories and explanations. Metaphysical theories can then become research programs. From the critical perspective, it is advantageous if two or more research programs compete with one another. Agassi expounded how Faraday used his metaphysical field theory as a framework within which he developed physical field theories. (Agassi's view influenced Lakatos's methodology of research programs.)

In psychology, the metaphysical idea that cognition is information processing has become important as a framework for the construction of many theories. A rival program is based on the (metaphysical) view that all mental processes are nothing over and beyond brain processes, and should correspondingly be studied by neuropsychology. The metaphysical assumption that humans always try to maximize utility has been a framework for economic theories. From the viewpoint of CR, it is unfortunate that in the latter case, there is no competing program of equal influence. If metaphysical views become too dominant in science, this may impede the growth of knowledge.

This gives rise to reconsidering the problem of demarcation. When Popper developed CR, he continued to regard the demarcation of science as of the utmost importance. But is this still justified? In CR, the boundary of science is no longer the boundary of rationality. So why still demarcate science from metaphysics? According to Wettersten (1992), this demarcation is a residue of positivistic thought. Albert (1985, Chapter 5) argued against drawing boundaries between disciplines. Such bounds often foster the immunization of theories against criticism from other perspectives.

One demarcation that remains significant is the idea of rationality as openness to criticism. CR rejects all kinds of dogmatism and fundamentalism, which refuse to take note of any unwanted evidence or argument, and consequently rejects such metaphysical approaches that are essentially dogmatic. But here the boundary is not between science and metaphysics but between the critical attitude and dogmatism. Dogmatism can sometimes be found even in science, and metaphysical traditions may place great importance on assessing their views critically.

Today many philosophers of science hold that philosophy should only try to understand, describe, or reconstruct what science does, but never evaluate or criticize it. Some even consider it as presumptuous on the philosophers' part to tell scientists what to do or not to do. By contrast, CR deems its task a critical one. Philosophy should critically analyze the developments in science, point out problems from the metascientific perspective, and also be open to criticism from science. In accordance with this view, Popper (1982) criticized the Copenhagen interpretation of quantum mechanics by Niels Bohr. Albert criticized strategies of immunization in economics, of making theories and models irrefutable, for instance, by using unspecified ceteris paribus clauses (Albert et al., 2012).

Reason and Nonjustificationism

Most philosophers closely associate rationalism with the idea of justification, of giving reasons for what one believes. CR, however, rejects justificationism because the search for proof leads to unsolvable problems. Albert (1985) demonstrated this by considering the principle that one should accept a statement if and only if one can give a sufficient reason for it. A reason for a statement A is regarded as sufficient if it makes A absolutely certain, or guarantees that A is true. This, however, leads into a fatal situation that Albert called the 'Münchhausen trilemma.' One has three possibilities, none of which solves the problem. In order to prove A, one has to accept some other assumption B as a reason for A, which has in turn to be proven, so that one gets into an infinite regress. Or one has to argue in a circular way. Or one has to stop the process of justification by adopting some assumption dogmatically. In each case, one has not succeeded in sufficiently justifying A. CR offers as a solution to give up the principle of sufficient reason and replace it with the principle of criticism.

But is criticism an end in itself? Must we abandon any justification by giving reasons once we have accepted fallibilism? Or can we conceive of an idea of justification that does not involve certainty? Isn't it, for instance, justified to believe that the earth is round and not flat, though this belief is fallible? Does CR hold that there are no sufficient or certain reasons, or that there are no reasons, sufficient or not, for anything at all? With respect to this problem, there are two groups of critical rationalists who maintain different views.

Musgrave (1999) interprets CR as saying that it is not possible to justify a hypothesis (or theory), but it is possible to justify belief in, or acceptance of, a hypothesis. A universal statement cannot be inferred, by deduction or induction, from observation statements. In this sense, it cannot be justified. Yet it is reasonable to believe in a hypothesis or theory, or to adopt it as true, if it has withstood serious criticism. If there are several competing hypotheses, it is reasonable to adopt the best-corroborated one. More exactly, *it is reasonable to (believe) adopt as true A (at time t) if and only if A is the hypothesis that has (at time t) best withstood serious criticism*. According to Musgrave (1999, p. 324), this is the central principle of CR.

Popper's view is slightly different. He regarded as incautious the claim that a theory is true. According to Popper, truth is a regulative idea, a standard by which we usually fall short. Even the best scientific theories are oversimplifications, and therefore we never have reason to believe in their complete truth. But we may have reason to believe that one theory is a better *approximation of the truth* than another. More exactly, Popper (1972, p. 82) said we could have reasons for the following claim: "This theory seems at present, in the light of a thorough critical discussion, and of severe and ingenious testing, by far the *best* (the strongest, the best tested); and so it seems the one nearest to truth among the competing theories." Such a claim is fallible but may be justified by the current state of the discussion.

Popper also gave a formal definition of *truthlikeness* or *verisimilitude* (1963, Chapter 10) that, however, turned out to be untenable. Others have tried to improve Popper's attempt, but all proposals raise difficult new problems. Popper (1983, p. xxxvi) finally came to the conclusion that a formal definition is not necessary, and that the idea of truthlikeness could be made sufficiently clear with the help of examples. For instance, Kepler's theory that the planets move in (not very elongated) ellipses with the sun in their common focus is a better approximation to truth than Copernicus' view according to which the planets move in circular orbits. It is controversially discussed whether methodology needs the idea of truthlikeness, and whether a formal definition is necessary.

Thus, with respect to justification, Popper's view is similar to Musgrave's, but Popper preferred to talk of approximation to the truth (or verisimilitude) instead of truth. He (1972, p. 103) also said that the degree of corroboration of a theory could be taken as an indication of its verisimilitude.

Does this mean that CR needs and (implicitly) accepts a form of *induction*? This depends on how induction is conceived of. There are at least three kinds of induction that CR can dispense with. First, as Hume and Popper pointed out, the truth of observation statements never guarantees the truth of a universal theory. Second, attempts to define the confirmation of a theory as an inductive probability have not yet been successful. Third, the major theories of science have not been discovered by applying formal inductive rules. (Otherwise, anybody could have developed these theories.) However, if any principle for the acceptance or preference of evidence-transcending statements is called inductive, then Musgrave's (and, as it seems, Popper's) CR includes an inductive principle.

Now there is another group of critical rationalists who abandon not only the aim of certainty but any kind of justification. Bartley (1984, p. 104) criticizes Popper's view for being not entirely consistent. Popper proposes that rationality is openness to criticism, yet his view is not completely nonjustificational since he thinks that the critical attitude had to be based on a faith in reason. Bartley calls such a view *fideistic*. As an alternative, he offers his *pancritical rationalism*, which does not even seek to justify itself. Pancritical rationalism keeps everything open to criticism, including itself. The traditional demand 'How do you know?' is answered by 'I do not know.' The request for justification no longer arises. Instead, it is asked, "How can our beliefs, conjectures, policies, traditions, and the like be exposed to maximum criticism?"

Miller (1994) holds a similar view. There is reason and rationality, but there are no good reasons, no sufficient or

partly sufficient reasons for accepting or rejecting any assumption. The aim of science is to retain as many true statements and to single out as many false statements as possible. The appropriate strategy is Popper's method of conjectures and refutations. Hypotheses are admitted to science if they are falsifiable by experience. If a hypothesis fails the tests to which it is submitted, it is removed from science. If it passes the tests, it is retained. But, most importantly, the passing of tests makes no difference to the status of a hypothesis. That is, the hypothesis does not become more plausible or more probable. Neither does it gain any kind of support or credibility. It remains forever as uncertain as it was when it was invented and yet untested. Neither theories nor observation statements can be justified. If a theory is classified as false, this is done without the claim that one could justify the respective observation statements. The latter are themselves fallible and open to dispute.

Miller (2006, Chapter 7) conceded that this is a form of skepticism. He emphasizes, however, that it is not relativism. Relativism claims that there is no absolute truth, though there are many relative truths. Miller holds that there is absolute truth to be had, but insists that we can achieve it only by making lucky guesses, and can never justify these guesses.

The view that theories can never gain any empirical support is, of course, different from what most scientists and philosophers think. Some critics deem CR a modern kind of irrationalism. The group of critical rationalists represented by Bartley and Miller, however, think that CR presented an ingenious solution of the problem of induction, which has not yet been understood by the majority of scientists and philosophers, because they have not yet escaped justificationist thought. According to Miller (2006, p. 112), this is a 'scandal of philosophy.'

See also: Critical Theory; Logical Positivism and Logical Empiricism; Popper, Karl Raimund (1902–94); Scientific Explanation.

Bibliography

Agassi, J., 1975. The nature of scientific problems and their roots in metaphysics. In: Agassi, J. (Ed.), Science in Flux. Reidel, Dordrecht, pp. 208–233.
Albert, H., 1985. Treatise on Critical Reason. Princeton University Press, Princeton.
Albert, H., 1999. Between Social Science, Religion and Politics. Rodopi, Amsterdam.
Albert, H., Arnold, D., Maier-Rigaud, F., 2012. Model Platonism: neoclassical economic thought in critical light. Journal of Institutional Economics 8, 295–323.
Albert, M., 2010. Critical rationalism and scientific competition. Analyse & Kritik 24, 247–266.
Andersson, G., 1994. Criticism and the History of Science. Brill, Leiden.
Bartley III, W.W., 1984. The Retreat to Commitment. Open Court, La Salle.
Chalmers, A.F., 1999. What Is This Thing Called Science? third ed. Hackett Publishing Company, Indianapolis.
Feyerabend, P., 1975. Against Method. New Left Books, London.
Gadenne, V., 2006. Methodological rules, rationality, and truth. In: Cheyne, C., Worrall, J. (Eds.), Rationality and Reality: Conversations with Alan Musgrave. Springer, the Netherlands, pp. 97–107.
Jarvie, I.C., 2001. The Republic of Science. Rodopi, Amsterdam.
Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Lakatos, I., 1970. Falsification and the methodology of scientific research programmes. In: Lakatos, I., Musgrave, A. (Eds.), Criticism and the Growth of Knowledge. Cambridge University Press, Cambridge.
Mantzavinos, C., 2005. Naturalistic Hermeneutics. Cambridge University Press, Cambridge.
Miller, D., 1994. Critical Rationalism: A Restatement and Defence. Open Court, Chicago.
Miller, D., 2006. Out of Error: Further Essays on Critical Rationalism. Ashgate, Aldershot.
Musgrave, A., 1999. Essays on Realism and Rationalism. Rodopi, Amsterdam.
Popper, K., 1934. Logik der Forschung. Springer, Wien (tenth ed. Tübingen: Mohr Siebeck, 1994).
Popper, K., 1945. The Open Society and Its Enemies. Vol. 1: The Spell of Plato. Vol. 2: The High Tide of Prophecy. Routledge, London.
Popper, K., 1959. The Logic of Scientific Discovery. Hutchinson, London (Originally published as Logik der Forschung, 1934).
Popper, K., 1963. Conjectures and Refutations. Routledge, London.
Popper, K., 1972. Objective Knowledge. Clarendon Press, Oxford.
Popper, K., 1982. The Postscript to the Logic of Scientific Discovery. vol. 3. Quantum Theory and the Schism in Physics. Hutchinson, London.
Popper, K., 1983. The Postscript to the Logic of Scientific Discovery. vol. 1. Realism and the Aim of Science. Hutchinson, London.
Wettersten, J.R., 1992. The Roots of Critical Rationalism. Rodopi, Amsterdam.
Wettersten, J.R., 2006. How Do Institutions Steer Events? An Inquiry into the Limits and Possibilities of Rational Thought and Action. Ashgate, Aldershot.

Critical Realism in Geography

Andrew Sayer, Lancaster University, Lancaster, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The nature of critical realist philosophy and its relevance to geography are outlined, including its fallibilism and distinctive way of understanding causation, which is especially appropriate for the study of the irregular behaviors that typically confront geographers. It also acknowledges that the intrinsically meaningful character of social phenomena requires interpretive understanding of meaning in society. Further, it defends a critical stance toward the ideas and practices studied by social researchers. Some implications for methodology are sketched, and illustrations are given of research informed by critical realism in geography. Finally, critical responses to critical realism are summarized.

Introduction

Critical realism is a philosophy of and for the social sciences that has had an impact on methodology in geography since the 1980s, although there have always been realist elements in geographical thought. With the rise of critical realism, these and further elements were developed and promoted self-consciously, while non or antirealist elements were attacked. Since its high point in the 1980s and 1990s various nonrealist developments and other kinds of realism have come to the fore.

A Definition

Critical realism argues that the natural and social worlds we study exist largely independently of the researcher, and that there is necessity in the world which enables and constrains what can happen. Causation is to be understood in terms of the powers that objects possess rather than in terms of empirical regularities among events; if and when these 'causal powers' are activated, the results depend on contexts, and are unlikely to be regular. Since social phenomena such as actions and ideas are intrinsically meaningful, unlike the subject matter of the physical sciences, their meaning also has to be interpreted. The world can only be understood in terms of available discourses. While there are no privileged routes for discovering the truth about the world, and empirical tests are themselves fallible, it is often possible to distinguish better from worse accounts in terms of their adequacy for informing material practice.

Intellectual Context

Critical realist philosophy was developed by Roy Bhaskar out of his earlier work on the philosophy of natural science ('transcendental realism') (Bhaskar, 1978, 1989). The latter sought an alternative to both those positions which treat knowledge as straightforwardly based on observation and experience (empiricism) and for which truth is a simple matter of correspondence between statements and the objects to which they refer, and various relativist positions which treat the truth of theories as purely relative to discourses. Critical realism also steers a course between positions which regard social science as the same as natural science and those which treat it as entirely different.

Realism and Fallibilism

The most fundamental realist assumption is that much of the world can exist independently of an observer's knowledge of it. The plausibility of this assumption rests upon the fallibility of knowledge; the fact that knowledge claims sometimes prove to be mistaken and generate false expectations implies that the world is not a product of our thought but is whatever it is regardless of what we think about it. Realists reject the empiricist view that reality is whatever we can experience since this implies an improbable coincidence between the world and the limits of our sensory powers, and implies that either knowledge is already complete, or that new knowledge creates new objects. Critical realists accept that our observations and knowledge are always formed through available discourses or are 'theory laden,' so that we can never escape from our conceptual frameworks to see how they compare with the world they seek to represent. However, that observation is theory laden does not necessarily mean that it is entirely determined by theories and so it can still be possible to compare the adequacy of different theories from within discourses and find observations which contradict expectations. This offers a way of avoiding the relativist view that truth is merely relative to theoretical systems. It also implies that science can make progress in explaining the world even though absolute truth is not an intelligible goal.

Critical realism therefore does not involve claims to have a privileged access to the truth about the world, a uniquely 'realistic' view, for the very independence of the world from our knowledge of it renders such claims problematic. It must necessarily view all knowledge as in principle fallible, although this does not mean, of course, that it is all false or all equally true or untrue.

Key Arguments and Their Influence on Geography

Critical realism became influential in geography partly for the above reasons, for geographers, like other scientists, were

seeking a way of understanding science that avoided the unappealing alternatives of empiricism and relativism. However, a further important factor was probably its view of causation and systems. Some of the first discussions of critical realism in geography emphasized this issue and its implications for the nature of theorizing in geography (Sayer, 1982, 1985).

The scientific revolution in geography, exemplified by *location theory* and *spatial analysis*, had largely assumed that causation was a matter of empirical regularities or 'constant conjunctions' among causes and effects. It was expected that understanding of the regularities in spatial organization of phenomena would grow to the point where it would be possible to formulate laws about them. This regularity-based view of causation had dominated the philosophy of science despite the fact that, particularly in social sciences like human geography, there appear to be few enduring precise empirical regularities, so that the search for laws of social and spatial behavior has been singularly unsuccessful. For realists this is unsurprising. For them, causation is a matter of what produces change, not a matter of empirical regularities in sequences of events. Objects have causal powers, that is, the power to act in certain ways. A seed has the power to germinate, a person the power to work. These objects have these powers in virtue of their structures, such as the biochemical constitution of the seed, or the anatomical and mental structures of a person. Whether these powers are ever activated depends on conditions whose presence is contingent, that is, neither necessary nor impossible, such as moisture and solar energy in the case of the seed or a need for food in the case of the person. Furthermore, the particular effects which follow when these causal powers are activated depends on conditions which again are contingent. The germination of the seed may be encouraged by fertilizer or halted by herbicides, the attempt of the person to work may be hampered by lack of tools or encouraged by their availability.

This alternative conception of causation has far-reaching implications. First, since it dissociates causation from empirical regularities, it enables understanding of how unique events, as much as repeated ones, can be caused. Second, it focuses on necessity and contingency rather than regularity or the search for empirical order. Third, in view of this, explanation does not require repeated observations: what makes something happen has nothing to do with the number of times it has or has not been observed to happen. It follows from this that the old debate in geography between idiographic and nomothetic approaches – the former a supposedly nonscientific study of the unique, the latter a supposedly scientific search for laws – was falsely based. Regardless of whether phenomena were unique or widely replicated, they could be explained in the same way.

The uniqueness of many geographical phenomena no longer posed a threat to geography's scientific status. Fourth, it became possible to see that the search for generalizations is distinct from explaining how causes work; generalizations, as the term suggests, attempt to tell us how common or extensive certain phenomena are, and this is different from explaining how they are produced. The greater part of standard literature on the philosophy and methodology of science had confused these two things, and many geographers had done so too. Fifth, the realist account offers a way of understanding how it is possible for the same cause to produce different effects and for different causes to produce the same effects, since the relationship between causal powers and empirical outcomes is not one of automatic regularity but is dependent on the conditions within which they are located.

An example of this can be given from a kind of economic geography which was popular at the same time critical realism became influential. The geography of industrial restructuring looked at changes in economic geography in terms of how competitive pressures upon firms to restructure produced responses (e.g., Massey and Meegan, 1982). In realist terms, competitive pressures on profit margins were considered as the main causal power of interest here, with their effects being mediated by a wide range of contextual circumstances, such as the particular cost structures of the firms under pressure, the scope for technological change, and the availability of various kinds of labor. Thus, in an industry such as clothing, which had little scope for further automation or speeding up work, and in which wages were a major element of costs, a common response was to seek out cheaper sources of labor, sometimes by relocating.

By contrast, in industries which were more technology intensive and had scope for further automation, a more common response was technological innovation. By such means, it was possible to explain how the same or similar causes could produce radically different results according to context, whereas approaches which rested upon the discovery of empirical regularities foundered. Similar research also illustrated how a single kind of outcome – job losses – could have completely different causes, for example, a drop in demand for the firm's products, or the introduction of automation during steady or rising demand, or reorganization of work routines to eliminate idle time and intensify work (see also Henry, 1992). For further discussion of these topics *see* Economic Restructuring: Geographic Aspects.

One of the most important and lasting influences of critical realism on geography has been on its empirical research designs (Sayer, 1992, 2010). Hitherto, the dominance of conceptions of explanation as tied to generalizations about empirical regularities had been used to privilege 'extensive' research designs which sought to uncover such regularities or patterns of covariation. Although these are, as the term indicates, useful for establishing how extensive certain phenomena are, they do not necessarily identify causes, as opposed to spurious correlations. The latter are likely to be thrown up by extensive research because it preselects groups of objects taxonomically, that is, on the basis of similarities (e.g., industries of a certain kind), rather than functional connection. By contrast, intensive research designs, including ethnographic case studies, focus on causal rather than taxonomic groups and follow up causal connections wherever they lead – often between heterogeneous objects, such as firms of different kinds. Intensive research is typically more exploratory, since the purpose is to discover the relevant causal groups. It is often time consuming to do this for large numbers of cases, and therefore it is not always a good basis for generalization. However, as we have seen, this lack of generalizability does not undermine the status of causal explanations of the

particular cases which are studied intensively. Consequently, having formerly been regarded as unscientific, intensive studies have now become more accepted in geographical research.

Critical realism acknowledges that the methodology of social science must diverge from that of the natural sciences because human, social phenomena are intrinsically meaningful. While the objects that, say, geologists study, are what they are regardless of any meaning people may give to them, objects such as socially segregated urban areas depend on the meanings actors give to them. To understand what these are, one has to know what their meaning is in society, whereas the nature of a rock does not depend on how its constituents understand one another, as they do not. Consequently, an important part of the methodology of human geography, no less than other social sciences, is interpretive understanding. For example, to understand the geography of the Middle East, we would have to understand what significance its land had for Jews and Arabs in terms of their identities.

Although this interpretive understanding has also been the focus of *humanistic geography*, critical realists make certain qualifications which the latter might not. The three most important are as follows. First, critical realists argue that although we have to interpret the understandings that actors have of their situations in order to make sense of how they act, their understanding is not necessarily a good one, indeed it may be systematically flawed. Moreover, these misunderstandings may have important effects which actors cannot explain. Thus a misogynist may believe that the proper spatial location for women is in the home, on the false grounds that they are incapable of coping with life in the public sphere; the falsity of this understanding makes a difference to what happens and therefore has to be acknowledged in scientific accounts. Further, different actors may give contradictory accounts of the same situation: where this is the case the researcher cannot sit on the fence and say both are correct. Actors' accounts of situations therefore face social scientists' accounts both as objects of study and rivals. It is therefore likely that researchers' accounts will be at least implicitly critical of actors' accounts – a conclusion which humanistic geographers are generally unwilling to accept. It is primarily for this reason that Bhaskar's philosophy of social science is called *critical* realism.

Second, among the intrinsically meaningful aspects of social phenomena are actors' reasons. Critical realists argue that reasons can be causes, since they can be responsible for producing change, by prompting actions. This, of course, implies that realists do not limit causation to physical processes. To deny that reasons are causes is to imply that the giving of reasons by actors is ineffectual and redundant, which of course makes it unclear why they are given. Third, an insistence on the necessity of interpretive understanding should not allow us to forget that social life is also material, embodied, and spatially situated, and that things happen *to* people, so that, for example, they can be dominated, sometimes regardless of their understanding of the situation.

The realist facility for explaining irregular processes was particularly attractive to a discipline such as geography, concerned as it is with spatial differentiation and uneven development, and hence unable to ignore – as, say, economics tries to do – the problems of applying equilibrium models or making generalizations which are supposed to hold across time and space. The systems which geography and other social sciences study are *open*, that is, their elements can themselves be undergoing internal qualitative change – actors aging and learning, for example – and the relations among the elements are not constant, with the result that any empirically regular behaviors are unlikely to be durable or widespread. Evolutionary change and disequilibrium rather than equilibrium are the norm.

Where previous attempts at scientific explanation in geography had tried to evade the prevalence of differentiation and irregularity, on the mistaken grounds that explanation required the discovery of regularity, human geography began to reembrace differentiation, nowhere more strikingly than in Doreen Massey's influential *Spatial Divisions of Labour* (Massey, 1984). This was closely followed by a burgeoning literature on locality studies, which typically focused on how economic restructuring worked out in different localities, and how those differences impacted back on the restructuring process (e.g., Cooke, 1986).

Critical Reception

Some geographers feared that such concerns heralded an abandonment of theory and generalization and a slide toward the empiricist documentation of facts (e.g., Harvey and Scott, 1989), indicating a resistance to the critical realist de-emphasis of generalization. These fears were not merely methodological but partly about a weakening attachment in economic geography to a particular kind of theory – Marxism – and also a tendency to treat critical realism as a source of substantive theory rather than as a philosophy, when it had no such pretensions. At the same time, some of the attempts to use critical realism in localities studies invited justified criticism by arbitrarily translating it into the study of general, supralocal causal processes interacting with local contingent conditions. However, critical realism did not license such an extraordinary alignment of scale and necessity: the supralocal is no more or less the realm of necessity than the local, nor is the local any more the realm of contingency than the supralocal.

Toward the end of the 1980s a newer generation of geographers became influenced by postmodernism (*see* Postmodernism in Geography). In some ways this continued the emphasis on differentiation – only phrased as 'difference' which had connotations of social differences of identity. One of the positive features of this literature was that it gave greater concern to the textual nature of accounts of the world (Sayer, 1989; Stones, 1996). Particularly in actor network theory, the world was seen as inherently messy, and skepticism toward grand universalizing accounts of the world grew. At the same time, some postmodernist writers rekindled relativist, antirealist tendencies (e.g., Barnes, 1996; Graham, 1990), although sometimes by mischaracterizing realism as claiming privileged access to the truth rather than as having a fallibilist view of knowledge.

In recent years interest has grown in 'assemblage theory' (Anderson et al., 2012; Saldanha, 2006). This derives from the

work of Gilles Deleuze, and, as Manuel DeLanda has argued, represents a different kind of realism (DeLanda, 2006). Where critical realism has tended to assume that the world is made up primarily of relatively durable structures of interlocking parts such as capitalist firms or other organizations, assemblage theory treats the world as a consisting in combinations of externally related elements that interact in various ways that depend on the manner of their combination, but which can also break away and form new assemblages. Thus the emphasis is on processes of formation and transformation or dissolution. This has obvious appeal for understanding 'messy' objects such as cities and regions. Like critical realists, assemblage theory acknowledges the possibility of 'emergent powers,' that is, powers that emerge from particular relationships between elements and which are irreducible to the powers of their constituents. However, for assemblage thinkers, applying theoretical concepts of familiar structures in empirical research to actual cases is suspect: it is liable to underestimate the open-endedness of processes of formation of assemblages from preexisting elements. Instead, the assemblage approach suggests a more cautious and bottom-up style of investigation into the formation of assemblages, rather than one which 'draws down' and applies familiar concepts such as 'neoliberalism' or 'globalization' to cases.

This is a new incarnation of an old concern in geography about theory and empirical research whose previous incarnations include the debate between idiographic and nomothetic approaches in the 1960s and 1970s, and between narrative and analysis in the 1980s and 1990s (Sayer, 2000). Each incarnation adds something different. Where critical realism emphasizes the centrality of conceptualization to theorizing, postmodern, and post-structuralist sensibilities push further the questioning of familiar categories, for example, challenging familiar dualisms such as nature/culture and problematizing the nature of boundaries between things so that they are seen as contingent outcomes rather than as givens. Perhaps too much emphasis on the messiness of the world may lead us to neglect the power of our best theories to help us see the wood for the trees, so that we render what may be recurrent structures as ad hoc assemblages. Whatever the case, realism's emphasis on ontology, on theorizing the basic composition of the world, evident in its focus on necessity, causal powers, and structures, makes it particularly well equipped for continuing these discussions.

Even though critical realist philosophy is now acknowledged infrequently in geography, most geographical research is at least implicitly realist. Antirealist and relativist approaches tend to be more preached than practiced – not surprisingly, for when it comes to doing actual research, we have to make claims about what is the case and assume that the world is not merely the product of our ideas, and that our theories are susceptible to empirical assessment. Intensive research designs, defended by critical realism, have become common and indeed orthodox in human geography, and at least implicitly, explanations are usually in terms of the causal powers of things. Recently, the emergence of the assemblage approach as a different form of realism invites dialogue with critical realism.

See also: Phenomenology: Philosophical Aspects.

Bibliography

Anderson, B., Kearnes, M., McFarlane, C., 2012. On assemblages in geography. Dialogues in Geography 2 (2), 171–189.
Barnes, T.J., 1996. Logics of Dislocation. Guilford Press, New York.
Bhaskar, R., 1978. A Realist Theory of Science. Harvester Press, Hassocks, Sussex.
Bhaskar, R., 1989. The Possibility of Naturalism, second ed. Harvester Wheatsheaf, Hemel Hempstead, UK.
Cooke, P., 1986. Localities: The Changing Face of Urban Britain. Unwin Hyman, London.
DeLanda, M., 2006. A New Philosophy of Science. Continuum.
Graham, J., 1990. Theory and essentialism in Marxist geography. Antipode 22, 53–66.
Harvey, D., Scott, A.J., 1989. The practice of human geography: theory and empirical specificity in the transition from Fordism to flexible accumulation. In: Macmillan, B. (Ed.), Remodelling Geography. Blackwell, Oxford, UK, pp. 217–229.
Henry, N., 1992. The new industrial spaces: locational logic of a new production era. International Journal of Urban and Regional Research 16, 375–396.
Massey, D., 1984. Spatial Divisions of Labour. Macmillan, London.
Massey, D., Meegan, R., 1982. Anatomy of Job Loss. Methuen, London.
Pratt, A.C., 1994. Uneven Re-production: Industry, Space and Society, first ed. Pergamon, Oxford, UK.
Saldanha, A., 2006. Reontologising race: the machinic ontology of phenotype. Environment and Planning D: Society and Space 24, 9–24.
Sayer, A., 1982. Explanation in economic geography. Progress in Human Geography 6, 68–88.
Sayer, A., 1985. Realism and geography. In: Johnston, R.J. (Ed.), The Future of Geography. Methuen, London, pp. 159–173.
Sayer, A., 1989. The 'new' regional geography and problems of narrative. Environment and Planning D: Society and Space 7, 253–276.
Sayer, A., 1992/2010. Method in Social Science: A Realist Approach, second ed. Routledge, London.
Sayer, A., 2000. Realism and Social Science. Sage, London.
Stones, R., 1996. Sociological Reasoning: Towards a Post-modern Sociology. St Martin's Press, New York.

Critical Reflexivity and Social Work Practice

Christine Morley, University of the Sunshine Coast, Maroochydore DC, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Critical reflexivity is an epistemological position that brings together the social constructionist stance of reflexivity with the emancipatory goals of critical social science. Critical reflexivity is a vial precondition for critical reflection, yet not sufficient in its own right to activate the transformative deconstruction and reconstruction processes that critical reflection enables. The related concepts of reflection, reflexivity, and critical reflection will also be discussed. A critically reflexive stance is essential to further a critical social work agenda that is committed to social justice and human rights, despite the challenges presented by the contemporary, neoliberal context.

Introduction

Critical reflexivity is an epistemological position that brings together the social constructionist stance of reflexivity with the emancipatory goals of critical social science. It is a contested concept and often conflated with critical reflection. Critical reflexivity is a vial precondition for critical reflection, yet not sufficient on its own to activate the transformative deconstruction and reconstruction processes that critical reflection enables. In order to clarify the concept of critical reflexivity, the related concepts of reflection, reflexivity, and critical reflection will also be discussed. A critically reflexive stance is essential to further a critical social work agenda that is committed to social justice and human rights, because it enables critical reflection. These concepts are intimately connected as critical reflection can support social work practitioners to maintain a critically reflexive stance, despite the challenges presented by contemporary, neoliberal contexts that privilege objectivist paradigms including evidence-based practice and research.

Several writers have noted the distinct lack of consensus over definitions of reflection, reflexivity, and critical reflection (Fook et al., 2006), and critical reflexivity (Kessl, 2009: p. 305). As Askeland and Fook (2009: p. 289) state, "The terms reflection, reflective practice, reflexivity and critical reflection/reflexivity are often conflated or used indiscriminatingly, without clear definition." D'Cruz et al. (2007) similarly point out that "'Reflexivity' is often used interchangeably with other terms such as 'reflectivity' 'reflection' and 'critical reflection'" (see, e.g., Jessup and Rogerson, 1999: p. 176; Leonard, 1997: p. vii). Despite the complementarity of these terms, articulating clearer delineation between them is important for social work as they explicitly relate to improving professional practice and facilitating social change (D'Cruz et al., 2007).

Reflection

Reflection, also referred to as reflectivity (Fook, 1999), the reflective approach (Argyris and Schon, 1976; Schon, 1987), the reflective model (Pray, 1991), and reflective practice (Fook, 1996; Gould and Taylor, 1996), is "a process of thinking, of comparing and verifying for the purpose of learning about and improving practice, developing practice-based theory, connecting theory to practice, and improving and changing practice" (Fook and Askeland, 2006: p. 12).

The ideas about a reflective practice are usually traced to educationalists Argyris and Schon (1976) (later developed by Schon, 1987), who proposed a reflective process for uncovering hidden assumptions, particularly regarding professional learning. These authors were critical of professionals, identifying wide-ranging differences between their espoused theories and their actual practice (Argyris and Schon, 1976; Schon, 1987).

Schon (1987) questioned the usefulness of traditional approaches to building knowledge, and was dubious of the capacity of technical rationality to respond to the relationship between theory and practice. Some authors credit the earlier work of Dewey (1916, 1933) in developing current understandings of reflective practice (Mezirow et al., 1990; Fook et al., 2006: p. 8).

Whilst closing the gap between theory and practice is the ultimate goal of reflection (Schon, 1987), this may simply reinforce conservative practice. If a mental health practitioner espouses an acceptance of psychiatry, then deferring to medical experts that accept the status quo, and individualize social problems, is not identified as problematic practice by a reflective approach (Taylor and White, 2000). As Taylor (2013) explains reflective practice potentially occupies an apolitical space, over emphasizing the role of individual relationships and presenting an inadequate view of the impact social structures. This point clearly demarcates reflection from critical reflection, and some authors, suggest that such limitations can be enhanced by reflexivity (Lipp, 2007; Finlay, 2008).

Reflexivity

Reflexivity refers to one's capacity to look or act back upon oneself. The concept of reflexivity has figured prominently in epistemological debates concerning the relationships between knowledge, self, and society; and has emerged as a key concept in social constructionist approaches. The basic constructionist insight posits that knowledge of the world does not exist apart from people's involvement in it, which in turn shapes knowledge. Socially conditioned subjectivity decisively

affects knowledge production and truth claims (Seidman, 2013). Reflexivity is the self-conscious application of this insight, that social scientists apply to the truth claims of themselves and others.

One of the first writers to coin the term 'reflexivity' was Kelly (1955), who referred to the reflexive qualities of his work based on his acknowledgment of personal constructs (cited in Holland, 1999: p. 465). Later, in developing ethnomethodology, Garfinkel (1967) employed self-reflexive insight to show how the meaningfulness of statements and practices within academic disciplines are contingent upon the context of their production. This anticipated the contemporary movement from a social science based on establishing universal and objective knowledge to recognition of emergent, situated, and partial truths.

Gouldner (1970), in formulating a critical sociology as an alternative to systems theory and functionalism, contrasted its self-critical 'reflexivity' (explicating the social scientists' own assumptions and positionality in the construction of knowledge) with the contrived objectivism of mainstream sociology. Throughout the social work literature, reflexivity has been defined in myriad ways, yet most writers accentuate the impact and influence of the researcher on the research, especially motivations for undertaking the research, the design and methodology of the research, and how 'findings' are interpreted and presented, e.g., the privileging or exclusion of particular data in the final presentation of the research (Fook, 1999). This applies to both formal research in social work, and informal research that practitioners undertake on a daily basis, i.e., the knowledge that social workers produce about their practice and clients when interpreting situations to inform their responses to clients. Out of the multitude of ways that the worker could understand the client's narrative, how has the practitioner's values, assumptions, and interpretations shaped the information they have chosen to privilege, marginalize, or exclude in their presentation of the client's story in their case notes, research reports, or assessments? (This is discussed further below.)

Reflexivity exposes the researcher as the main instrument of data collection and a primary factor in the production of the knowledge (Fook, 2013). It asks researchers and practitioners to cease impersonating "'shamans' of objectivity" (Ruby, 1980: p. 154 cited in Smyth and Shacklock, 1998: p. 7) and transparently acknowledge that "the person producing the theory is included within the subject matter [s/]he attempts to understand" (Holland, 1977: p. 82 cited in Holland, 1999: p. 465). Instead of indulging the pretense of objectivity, and "engaging in futile attempts to eliminate the effects of the researcher … reflexivity enables researchers to…set about understanding them" (May, 1993: p. 75 cited in Smyth and Shacklock, 1998: p. 3) and use these to develop deeper insights into the social phenomena being explored (Morley, 2013).

Reflexivity is increasingly a compulsory requirement in evaluating qualitative social work research (Daley, 2010; Lipp, 2007). It appears in the research literature as an alternative to objectivist notions of reliability and validity, offering another approach to conceptualize ethical and epistemologically and methodologically sound research (D'Cruz and Jones, 2004).

Lipp (2007) argues that complexities associated with developing a clear, universal definition of reflexivity lie in acknowledging its use in different situations for various purposes. Whilst originating from research to "allow the researcher to examine how their research acts on the world and how the world acts on their research" (Lipp, 2007: n.p.), its applicability to enhance professional learning and social work practice is equally profound.

Reflexivity and Social Work Practice

More recently, reflexivity has emerged within the social work literature (Taylor and White, 2000; Powell, 2002; D'Cruz et al., 2007). For social work, reflexivity carries significant implications for informing research and critical practice. In the context of social work practice, being reflexive requires "an ability to recognize our own influence – and the influence of our social and cultural contexts …. In this sense, then, it is … about factoring ourselves as players into the situations we practice in" (Fook and Askeland, 2006: p. 12). Other writers further articulate: "Reflexivity is manifested in asking fundamental questions about assumptions generated by formal and practice theories; it addresses the multiple interrelations between power and knowledge, and acknowledges the inclusion of self in the process of knowledge creation in social work practice" (Lam et al., 2007: p. 91).

Paralleling this, Otto et al. (2009: p. 476) have drawn on work by Taylor and White (2000) to define reflexivity as a "primarily practical and ethical skill in social work … one that is fundamental for judging what is desirable and appropriate in specific circumstances by applying particular informal heuristics, rules and values, and the one that is particularly important in ambiguous and undetermined settings, circumstances, and contexts." Other important work on reflexivity in social work argues that reflexivity demands recognition of practitioners' role in engaging with, interpreting, and therefore recreating, and reproducing understandings of reality (Taylor, 2006).

Given that reflexivity highlights "he relevance of understanding how one's self can influence what and how knowledge is made, and can therefore shed light on how specific assumptions we make can arise from our own background and experience" (Fook, 2013: p. 6), it holds particular implications for practice, especially in relation to how social workers conduct assessments. In taking a reflexive position, social work practitioners choose from competing sets of discourses to assess the nature of issues presented by the person, group, or community they are working with. Each story the social worker collects has been assembled from a series of social constructions that are elevated to the status of facts and privileged in a coconstruction of the service user's particular account of their experience and the worker's engagement with it. Therefore, as the person to whom the stories are being told, the practitioner is implicated in constructing the service user's narratives (Frost, 2006). Developing a reflexive awareness about intrinsically political process of meaning making reveals that the practitioner's own subjectivity inevitably influences how they interpret and assess situations or behaviors.

The implications of this are troubling for the dominant objectivist paradigms that contend it is possible to isolate the object from the subject, the researcher from the researched, and the social worker from their assessment (Meinert et al., 2000).

That which can be known, from an objectivist epistemological position, is assumed to exist independently of the knower (Arnd-Caddigan and Pozzuto, 2006).

Conversely, reflexive practitioners recognize the impact of self on the knowledge that is created and factor this understanding into how professional decisions are made (Fook, 2013). Thus, the practitioner becomes the main tool of data generation during an assessment, which arguably contributes to a much deeper understanding of the service user's situation (Powell, 2002; White, 2001). Hence, the objectivist notion that assessments are fact-finding missions that can be undertaken as neutral and detached, technical activities is flawed (Rodwell, 1987; Fook, 2012). This type of reflexivity challenges the instrumental accountability inherent in evidence-based practice protocols and guidelines. For example, a child protection worker who undertakes an assessment using an objectivist/realist paradigm would attempt to gain information through the use of technical tools such as risk assessments, believing that it is possible to gain objective knowledge about reality that is untouched by their interpretation (Meinert et al., 2000; Loftus et al., 2011). In contrast, reflexive practitioners recognize that they are the key tool in any assessment process and so they scrutinize the ways their interpretive lens has informed the knowledge they produce.

When Kelly (1955 cited in Holland, 1999) first referred to the reflexive qualities of his work, he did so to differentiate himself from the behaviorists and psychoanalysts that he detested. This need for differentiation closely relates to the more contemporary debates that position reflexivity as a counter discourse to objectivist paradigms (D'Cruz and Jones, 2004); or reflective practice as a "new epistemology of practice" (Napier and Fook, 2000) and a refusal to reproduce the assumptions embedded within the natural sciences within social work practice (Taylor, 2013); and critical social work as an agent of critique and resistance to dominant discourses such as objectivism, evidence-based practice, and neoliberalism (Dominelli, 2004; Gray and Webb, 2013). Indeed, "In the context of the welfare state the definition of critical or radical social work – and therefore critical reflexivity – seem[s] to be relatively evident as 'oppositional' to the dominant paradigms" (Kessl, 2009: p. 312). Gray and Webb (2013) concur with this view, arguing that the new politics of critical social work take an innately political stance that that challenges capitalism and neoliberalism whilst facilitating emancipatory practices that address structural inequalities and injustices.

Critical Reflexivity

It is this political orientation that enables reflexivity to become critical reflexivity: by challenging and contesting the truth claims of dominant paradigms and the associated power relations; by elucidating the socially constructed nature of reality; and creating the conceptual space for alternative emancipatory discourses to emerge (Dominelli, 2004). For some writers, an analysis based on both modernist and postmodernist notions of critical social theory, implicitly, yet distinctively underpin their use of reflexivity in social work. These writers do not use the term critical reflexivity, but define reflexivity in critical terms. For example, Giles and Pockett (2013: p. 209) describe reflexivity as "an active process of critically analysing the knowledge and values that influence practice." Related to this, Ferguson (2003: p. 199) argues for a definition of reflexivity that draws on Giddens (1990) and Beck (1992) suggesting that reflexivity is "the ability to act in the world and to critically reflect on our action and in ways that may reconstitute how we act and even reshape the very nature of identity itself." Taylor and White (2000) refer to 'epistemic reflexivity' that promotes self-awareness of the impact of dominant discourses on tacit assumptions (Polanyi, 1966; Giles and Pockett, 2013: p. 209).

Gouldner (1970: pp. xiii, 489), writing prior to these debates, conceptualized "the historical mission of a 'reflexive sociology' as the 'transcend[ence] [of] sociology as it now exists." This representation of reflexivity invokes notions of fundamental transformations of the discipline, and is aspirational about social change. Lipp (2007) similarly devises a critical model of reflexivity that utilizes a form of reflective knowledge that takes account of technical practice and emancipatory dimensions of knowledge through critical theory. In addition, Freshwater and Rolfe's model of reflexivity implicitly incorporates a critical perspective including three types of reflexivity: a deep level of reflection; an account of the socio-political context; and reflection-in-action (cited in Lipp, 2007). Finlay (2008: p. 5) also depicts reflexivity as involving both 'personal reflection and social critique.' In her construction of reflexivity she asserts that "Reflexive practitioners engage in critical self-reflection: reflecting critically on the impact of their own background, assumptions, positioning, feelings, behaviour while also attending to the impact of the wider organisational, discursive, ideological and political context" (Finlay, 2008: p. 6). Such interpretations of reflexivity are clearly informed by the emancipatory goals of critical social work, despite the language of critical reflexivity not being explicitly used.

However, some discussions of both reflection and reflexivity have the potential to contribute to a critical emancipatory agenda, but are not inherently critical of their own volition (Lay and McGuire, 2009). Without explicit links to a critical theoretical analysis, these two concepts can become technicized and conservative (Taylor, 2013). For example, for Thomson (2013: p. 54) "the essence" of reflexivity in social work concerns "understanding how I am perceived by others. How others respond to me. And how I perceive myself working with others. It also concerns how I think and feel about my work, how I perceive myself in the job, and what the job means to me." This representation of reflexivity is neither politically conservative nor progressive and therefore, is inadequate for a critical social work project that "is primarily concerned with practising in ways which further a society without domination, exploitation or oppression" (Fook, 2012: p. 18). Whilst reflexivity involves an awareness of how a biographical lens shapes interpretations of individuals' constructions of their 'reality' over and above a range of competing perceptions, ultimately they need to be able to make decisions about which interpretations they privilege and why (Taylor and White, 2000). Critical theories provide an important framework on

which to base such judgments if individuals are committed to critical social work practice (Fook, 2012).

Critical Reflexivity and Social Work

This emphasis on the 'critical' is at the core of the difference between reflexivity and critical reflexivity. Although many authors have historically used an implicit critical lens when discussing reflexivity, the term critical reflexivity has more recently started to appear in the social work literature. Trevelyan et al. (2012: n.p.) position critical reflexivity as a principle aim of social work education, and argue that "critical reflexivity can be distinguished from non-critical forms of reflexivity by its emphasis on the socially constructed, power-laden nature of knowledge and subjectivity, and its embracement of anxiety and other difficult emotions in social work practice."

Kessl (2009: p. 307) positions critical reflexivity as a way for social work to locate itself within a 'transformatory agenda.' Kessl (2009) links critical approaches and reflexivity by drawing on Fook et al. (2006: p. 118) who connect the goals of social change to social workers that "are able to create critical knowledge which potentially challenges and resists current forms of domination." Kessl (2009) argues that a systematic approach to critical reflexivity in social work needs to be "augmented by and reconnected to a political, theory-based analysis of the actual contexts in which social work operates" and "need to be seen in connection with a fundamental transformation of the social order" (Kessl, 2009: p. 307).

Kondrat (1999) critiques the notion of reflexive self-awareness to develop critical reflexivity, which can be "taught as a skill … as a means for emancipation and self-actualization" (cited in D'Cruz et al., 2007: p. 77). Paralleling this, Brown (2012) posits critical reflexivity as a conceptual tool to identity the unintended, consequences of modernist discourses of power, which limits antioppressive practice. This is instructive for well-intentioned practitioners who seek to empower clients by giving them power. Brown (2012) highlights how assumptions embedded in modernist constructions of power as a commodity can inadvertently reproduce dominant power relations. In the example above, a lack of critical reflexivity could further disempower the client by reinforcing the idea that the client can only become empowered through practitioners' intervention, rather than their own actions. A critically reflexive practitioner would be more attentive to how the processes of their practice can be empowering, rather than assuming their intention will automatically produce an empowering outcome. Critical reflexive practitioners are more sensitive to unintended harm that social workers can cause through unreflexive practice.

Critical Reflexivity and Critical Reflection

This emphasis on the 'critical' is a key difference between reflection and critical reflection. Whilst some authors conflate the two concepts and use them interchangeably (e.g., Kondrat, 1999; Ruch, 2002), others debate whether they are the same (Parton and O'Byrne, 2000; Taylor and White, 2000), and some writers explicitly distinguish between them (e.g., Fook, 1999, 2012; Powell, 2002; D'Cruz et al., 2007; Finlay, 2008).

Both reflection and critical reflection are primarily identified as educational tools to improve congruence between professionals' use of theory and practice. However, more explicitly aligned with a radical politics, "[c]ritical reflection … is about understanding the technologies of power, language and practice that produce and legitimate forms of moral and political regulation" (Fook et al., 2006: p. 10). As Fook (2012: p. 113) claims: "The prime purpose of critical reflection is to unearth how we ourselves participate in discourses which shape existing power relations."

The critical component of critical reflection essentially challenges social inequalities that create injustice. This includes a key emphasis on power relations and domination (Gray and Webb, 2013). Critical reflection exposes the ways in which social work practitioners may be implicated in, or complicit with, dominant social power relations and structures that are unjust, inequitable, and/or undemocratic (Fook, 2012).

To critically reflect on implicit values and assumptions and scrutinize how these affect professional practice, practitioners can adopt a critically reflexive epistemological position. A critically reflective process cannot be undertaken from a position of objectivity. Critical reflexivity is a central component and necessary precondition of critical reflection. Both critical refection and critical reflexivity are underpinned by social constructivism, which fundamentally informs post-modernism. However, the key difference remains that critical reflexivity enables a practitioner to undertake critically reflective practice. Critical reflection incorporates a critically reflexive stance that activates broader transformative processes of deconstruction and reconstruction and takes account of critical postmodern theoretical frameworks (e.g., Allan et al., 2009; Fook, 2012); constructivist and experiential approaches to adult learning (Mezirow et al., 1990); and the reflective approach (Argyris and Schon, 1976; Schon, 1987) to generate theory inductively through critical reflection on practice.

Differentiation between critical reflexivity and critical reflection is contested, and some authors use the term critical reflexivity, as documented by Fook (2012) as critical reflection. Allan (2009: p. 1), in contrast, describes her auto-ethnographic study as one that used a "critically reflexive methodology to investigate [her] underlying assumptions about spirituality in social work and their impact on [her] practice with women survivors of sexual trauma." As in critical reflection, Allan (2009: p. 1) refers to analyzing 'critical practice moments' to identify 'existing and newly emerging assumptions' and refers to identifying the 'intended and unintended consequences of these assumptions' for her practice and engages in a 'deconstruction/reconstruction process' that develops new knowledge to inform her future practice. Hence critical reflexivity as defined here, is how others describe critical reflection.

Alternatively, other writers have presented reflection, critical reflection, and reflexivity on a continuum (Finlay and Gough, 2003). They contend that reflection stands at the

less sophisticated end of the continuum. "At the other end stands reflexivity: a more immediate and dynamic process which involves continuing self-awareness. Critical reflection lies somewhere in between" (Finlay, 2008: n.p.). The concepts of reflection, critical reflection, reflexivity, and critical reflexivity are presented, interpreted, and used differently, depending on the context and position of the author, and thus remain relatively elusive concepts, resisting universal definition.

Challenges to Critical Reflexivity in the Current Context

This article has discussed the development of critical reflexivity in social work, considering its implications for the practices of research and assessment. The related concepts of reflection, reflexivity, and critical reflection have been explored, demonstrating that the relationship between these concepts and critical reflexivity is contested and contentious. Regardless of the complexities and nuances between interpretations, most authors are in broad agreement that these concepts can support the emancipatory goals of critical social work.

The impact of neoliberalism has been well documented (Holscher and Sewpaul, 2006; Madhu, 2011) with the consensus being that neoliberal priorities can undermine critical social work objectives (see, e.g., Gray and Webb, 2013). These neoliberal priorities are augmented by imposing the practices of cost containment, standardization, compliance, an increased emphasis on risk assessment, and a quest of objectivity to eliminate uncertainty (Dominelli, 2004). This supports a view of social work that is less concerned with social justice and human rights, and more focused on individualizing social problems through developing technical proficiency in procedural knowledge that treats and manages apparently dysfunctional others (Holscher and Sewpaul, 2006; Madhu, 2011).

Within this context, many challenges exist for social work practitioners, researchers, and educators alike, who are committed to a critically reflexive stance. The dominant discourses are hostile toward this critical position, which is essentially constructed as at odds with the prevailing paradigms (Kessl, 2009; Gray and Webb, 2013). The debates about critical reflexivity, and attempts to develop further the concept, occur alongside policy mandates to make practice more scientific, objective, and evidence based (Taylor, 2013). Kessl (2009: p. 309) describes this as an "evidence-based practice or the 'what works agenda', that is designed to elicit a 'direct knowledge transfer' from positivist, experimental sources into the particular social work intervention" (Kessl, 2009: p. 309). Within this context, Kessl (2009: p. 309) argues that the concepts and practices of reflexivity (and critical reflexivity and critical reflection) are "under siege … and in danger of becoming more and more substituted by a technical-rational intervention agency which is still called 'social work', but represents a mode of intervention which is reduced to finding ways to achieve predetermined goals."

Arguably critical reflexivity and critical reflection are precisely what is needed to expose and oppose this situation: critical reflexivity to assist social workers to avoid becoming complicit in dominant power relations and being seduced by the 'tactics of governmentality' (Foucault cited in Madhu, 2011: p. 9); and critical reflection to reconstruct opportunities to create discretionary practice spaces for resistance and change. A lack of reflexivity, and particularly, critical reflexivity allows social work to function in a depoliticized mode (Taylor, 2013).

Both critical reflexivity and critical reflection are vital if social work is to strengthen a critical agenda. As Gray and Webb (2013) argue social work is charged with the responsibility to militantly challenge social injustices and the systems of domination that create them. The very acts of taking a critically reflexive epistemological position and developing the courage and humility to engage in critical reflection can in themselves, constitute undertakings of resistance to the dominant paradigms, and facilitate capacity to work toward emancipatory social change.

Acknowledgments

Thanks to Dr Phillip Ablett, University of the Sunshine Coast, for helping me to formulate and refine my ideas for this article. Acknowledgments to Judi Moir, University of the Sunshine Coast, for her assistance with the library research undertaken for this article.

See also: Critical Social Work Practice; Reflexivity; Social Work Theory.

Bibliography

Allan, J., 2009. Constructing the Spiritual in Social Work: A Critically Reflexive, Autoethnographic Process. Dalhousie University, Halifax, NS.

Allan, J., Briskman, L., Pease, B. (Eds.), 2009. Critical Social Work, second ed. Allen and Unwin, Crows Nest, NSW.

Argyris, C., Schon, D., 1976. Theory in Practice: Increasing Professional Effectiveness. Jossey-Bass, San Francisco, CA.

Arnd-Caddigan, M., Pozzuto, R., 2006. Truth in our time. Qualitative Social Work 5 (4), 423–440.

Askeland, G.A., Fook, J., 2009. Critical reflection in social work. European Journal of Social Work 12 (3), 287–292.

Beck, U., 1992. Risk Society: Towards a New Modernity. Sage, London.

Brown, C.G., 2012. Anti-oppression through a postmodern lens: dismantling the Master's conceptual tools in discursive social work practice. Critical Social Work 13 (1), n.p.

Daley, A., 2010. Refections on reflexivity and critical reflection as critical research practices. Journal of Women and Social Work 25 (1), 68–82.

D'Cruz, H., Jones, M., 2004. Social Work Research: Ethical and Political Contexts. Sage, London.

D'Cruz, H., Gillingham, P., Melendez, S., 2007. Reflexivity, its meanings and relevance for social work: a critical review of the literature. British Journal of Social Work 37 (1), 73–90.

Dewey, J., 1916. Democracy and Education: An Introduction to the Philosophy of Education. The Free Press, New York.

Dewey, J., 1933. How We Think: A Restatement of the Relation of Reflective Thinking the Education Process. Heath, Boston.

Dominelli, L., 2004. Social Work: Theory and Practice for a Changing Profession. Polity Press, Cambridge.

Ferguson, H., 2003. Welfare, social exclusion and reflexivity: the case of child and woman protection. Journal of Social Policy 32 (2), 199–216.

Finlay, L., January 2008. Reflecting on 'Reflective Practice' PBPL Paper No. 52. A discussion paper prepared for PBPL CETL. www.open.ac.uk/pbpl. http://www.open.ac.uk/cetl-workspace/cetlcontent/documents/4bf2b48887459.pdf (accessed 01.08.12.).

Finlay, L., Gough, B., 2003. Reflexivity: A Practical Guide for Researcher's in Health and Social Sciences. Blackwell Publishing, Oxford.

Fook, J., 1996. The reflective researcher: developing a reflective approach to practice. In: Fook, J. (Ed.), The Reflective Researcher: Social Workers' Theories of Practice Research. Allen and Unwin, Sydney, pp. 1–8.
Fook, J., 1999. Reflexivity as Method. La Trobe University, Bundoora, Melbourne.
Fook, J., 2012. Social Work: A Critical Approach to Practice, second ed. Sage, London.
Fook, J., 2013. Critical reflection in context: contemporary perspectives and issues. In: Fook, J., Gardner, F. (Eds.), Critical Reflection in Context: Applications in Health and Social Care. Routledge, Oxon, UK.
Fook, J., Askeland, G., 2006. The 'critical' in critical reflection. In: White, S., Fook, J., Gardner, F. (Eds.), Critical Reflection in Health and Social Care. Open University Press, Maidenhead, pp. 40–53.
Fook, J., White, S., Gardner, F., 2006. Critical reflection: a review of contemporary literature and understandings. In: White, S., Fook, J., Gardner, F. (Eds.), Critical Reflection in Health and Social Care. Open University Press, Maidenhead, England, pp. 3–20.
Frost, S., 2006. Recasting individual practice through reflection on narratives. In: White, S., Fook, J., Gardner, F. (Eds.), Critical Reflection in Health and Social Care. Open University Press, Maidenhead, pp. 107–117.
Garfinkel, H., 1967. Studies in Ethnomethodology. Prentice-Hall, Englewood Cliffs, NJ.
Giddens, A., 1990. The Consequences of Modernity. Polity Press/Blackwell, Oxford.
Giles, R., Pockett, R., 2013. Critical reflection in social work education. In: Fook, J., Gardner, F. (Eds.), Critical Reflection in Context: Applications in Health and Social Care. Routledge, Abingdon Oxon, UK.
Gould, T., Taylor, I. (Eds.), 1996. Reflective Learning for Social Work. Arena, Aldershot.
Gouldner, A., 1970. The Coming Crisis of Western Sociology. Basic Books, New York.
Gray, M., Webb, S., 2013. Towards a new politics of social work. In: Gray, M., Webb, S. (Eds.), The New Politics of Critical Social Work. Palgrave Macmillan, Houndmills.
Holland, R., 1999. Reflexivity. Human Relations 52 (4), 463–484.
Holscher, D., Sewpaul, V., 2006. Ethics as a site of resistance: the tension between social control and critical reflection. Research Reports 48 (1), 251–272.
Jessup, H., Rogerson, S., 1999. Postmodernism and the teaching and practice of interpersonal skills. In: Pease, B., Fook, J. (Eds.), Transforming Social Work Practice: Postmodern Critical Perspectives. Allen and Unwin, St Leonards, NSW.
Kessl, F., 2009. Critical reflexivity, social work, and the emerging European post-welfare states. European Journal of Social Work 12 (3), 305–317.
Kondrat, M.E., 1999. Who is the "self" in self aware: professional self awareness from a critical theory perspective. Social Service Review 74 (4), 451–477.
Lam, C.M., Wong, H., Leung, T.T.F., 2007. An unfinished reflexive journey: social work students' reflection on their placement experiences. British Journal of Social Work 37, 91–105.
Lay, K., McGuire, L., 2009. Building a lens for critical reflection and reflexivity in social work education. Social Work Education 29 (5), 539–550.
Leonard, P., 1997. Postmodern Welfare. Sage, London.
Lipp, A., 2007. Developing the reflexive dimension of reflection: a framework for debate. International Journal of Multiple Research Approaches 1 (1), 18–26.
Loftus, S., Higgs, J., Trede, F., 2011. Researching living spaces: trends in creative qualitative research. In: Higgs, A., Titches, D., Horsfall, Bridges, D. (Eds.), Critical Spaces for Qualitative Researching: Living Research. Sense Publishers, Rotterdam, The Netherlands, pp. 3–12.
Madhu, P., 2011. Praxis Intervention: Towards a New Critical Social Work Practice, 20th February. Social Science Research Network. http://ssrn.com/abstract=1765143 (accessed 14.02.12.).
Meinert, R., Pardeck, J., Kreuger, L., 2000. Social Work: Seeking Relevancy in the Twenty-First Century. Hapworth Press, Binghampton, NY.
Mezirow, J., et al., 1990. Fostering Critical Reflection in Adulthood. Jossey-Bass, San Francisco, CA.
Morley, C., 2013. Some methodological and ethical tensions in using critical reflection as a research methodology. In: Fook, J., Gardner, F. (Eds.), Critical Reflection in Context: Applications in Health and Social Care. Routledge, Abingdon Oxon, UK.
Napier, L., Fook, J., 2000. Reflective practice in social work. In: Napier, L., Fook, J. (Eds.), Breakthroughs in Practice: Theorising Critical Moments in Social Work. Whiting and Birch, London, pp. 1–11.
Otto, H., Polutta, A., Ziegler, H., 2009. Reflexive professionalism as a second generation of evidence-based practice: some considerations on the special issue "what works? Modernizing the knowledge-base of social work". Research on Social Work Practice 19 (4), 472–478.
Parton, N., O'Byrne, 2000. Constructive Social Work; Towards a New Practice. Allen and Unwin, St Leonards, NSW.
Polanyi, M., 1966. The Tacit Dimension. Anchor Books, New York.
Powell, J., 2002. The changing conditions of social work research. British Journal of Social Work 32 (1), 17–33.
Pray, J.E., 1991. Respecting the uniqueness of the individual: social work practice within a reflective model. Social Work 36 (1), 80–85.
Rodwell, K., June 1987. Naturalistic inquiry: an alternative model for social work assessments. Social Service Review 61 (2), 233–246.
Ruch, G., 2002. From triangle to spiral: reflective practice in social work education, practice and research. Social Work Education 21 (2), 199–216.
Schon, D., 1987. Educating the Critically Reflective Practitioner: Toward a New Design for Teaching and Learning in the Professions. Jossey-Bass, San Francisco, CA.
Scott, D., 1989. Meaning construction and social work practice. Social Service Review 63 (1), 39–51.
Seidman, S., 2013. Contested Knowledge: Social Theory Today, fifth ed. Cambridge University Press, New York.
Smyth, J., Shacklock, G., 1998. Behind the cleansing of socially critical research accounts. In: Shacklock, G., Smyth, J. (Eds.), Being Reflexive in Critical Educational and Social Research. Falmer Press, London, pp. 1–13.
Taylor, C., 2006. Practicing reflexivity: narrative, reflection and the moral order. In: White, S., Fook, J., Gardner, F. (Eds.), Critical Reflection in Health and Social Care. Open University Press, Maidenhead, UK, pp. 73–88.
Taylor, C., 2013. Critically reflective practice. In: Gray, M., Webb, S. (Eds.), The New Politics of Critical Social Work. Palgrave Macmillan, Baskingstoke.
Taylor, C., White, S., 2000. Practising Reflexivity in Health and Welfare: Making Knowledge. Open University Press, Maidenhead, UK.
Thomson, G., 2013. Not beating around the bush: critical reflection in a rural community health service. In: Fook, J., Gardner, F. (Eds.), Critical Reflection in Context: Applications in Health and Social Care. Routledge, Abingdon Oxon, UK.
Trevelyan, C., Crath, R., Chambon, A., 2012. Promoting critical reflexivity through arts-based media: a case study. British Journal of Social Work. http://dx.doi.org/10.1093/bjsw/bcs090. First published online: June 24.
White, S., 2001. Auto-ethnography as reflexive inquiry: the research act as self-surveillance. In: Shaw, I., Gould, N. (Eds.), Qualitative Research in Social Work. Sage, London, pp. 100–115.

Critical Social Work Practice

Jan Fook, Kingston University, Surrey, UK; and University of London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the history and development of critical social work as an approach and a practice. It reviews the major historical perspectives on a critical social work approach and then outlines an approach to critical social work based on the major tenets of critical social theory, which includes postmodern thinking about how beliefs about how knowledge is made and is inextricably linked to power. Specific concepts (knowledge, power, language, and identity) and some specific ways that might translate into the practice of critical social work are also considered. These include critical reflection and deconstruction/reconstruction, problematization and research, narrativity, and contextuality.

Introduction

Critical social work practice is still developing as an approach, and there are varied perspectives on exactly what theoretical frameworks underpin it. In this article, I outline a framework using a conception of critical social theory which incorporates common aspects from several related contemporary theoretical approaches such as postmodernism and feminism.

History of Critical Social Work

The idea of a critical social work is relatively recent idea in the history of social work. It is an idea and an approach, which is still developing, and this means that there are many different meanings of the term.

Most writers would, however, agree that critical approaches to social work were born around the 1970s. They were based on the common critique during that era. Social work was characterized, like many other helping professions, as an agent of social control and the state. Contrary, therefore, to its stated helping ideals, the social work profession was seen to support the status quo, and social structures, which worked against the best interests of the very communities and people it purported to help. The social fabric, which justified and supported social inequalities, was viewed as the cause of social problems themselves, and so social workers could not really help people unless they assisted in dismantling the very structures which brought about social ills in the first place. Social change was, therefore, seen as an imperative of social work. This included challenging social oppressions, and empowering disadvantaged groups. Most criticism was directed toward casework, or individualized work, which was seen as founded on personal pathology. Individualized practice was seen as 'blaming the victim,' or as pathologizing and changing the people who were actually the blameless victims of a faulty system. This individualized type of approach was regarded as the major influence in the 1950s when the 'psychiatric deluge' swamped social work's initial conceptions of the 'person-in-context' (see Fook, 1993 for a full discussion of these terms and trends). These early criticisms were labelled as radical, and spawned a whole body of literature on radical social work, largely underpinned by Marxist analysis (e.g., Bailey and Brake, 1975; Corrigan and Leonard, 1978). These Marxist critiques tended to focus on macro level criticisms directed at material structural inequalities, embodied particularly in class differences. Much of this literature was subsequently criticized for its lack of attention to social work with individuals (Fook, 1993) and indeed for its lack of attention to practice per se.

There was room, therefore, for the radical critique of social work to become more sophisticated. There was a need to include the importance of gender, and gendered inequalities, as an additional aspect of the social structure. Feminist social work (e.g., Dominelli and MacLeod (1982) therefore developed in the 1980s as an important approach in addressing structural inequalities. Critiques of racist and ableist social work were also relevant to this (Dominelli, 1988; Oliver, 1990; Morris, 1991). This development continued, as further types of analyses were introduced which better theorized the nature and causes of social inequalities, and which allowed more sophisticated understandings of power, and the relationship of individuals to their social contexts. How these issues are theorized accounts to a large extent for different perspectives on critical social work which are in contemporary development.

Contemporary Understandings

There are now a number of perspectives on critical social work, which include: postmodern, antioppressive, structural, and new radical perspectives based on neo-Marxist analysis. What each of these shares is a commitment to understanding how power operates to sustain social inequalities. However the specific analysis used, and therefore the specific actions recommended, and the types of social changes called for, vary depending on the perspective taken.

Structural perspectives emphasize the importance of a structural analysis in social work practice, and are more materialist in their orientation (Allan, 2009). A key contemporary Canadian author in this genre is Mullaly (2007). Antioppressive approaches emphasize intersectionality and the imperative against oppression of all types (e.g., Dominelli (2002). Postmodern, and postmodern critical perspectives tend to include more complex analyses of power and difference, and therefore are useful for understanding the operation of exclusion at the everyday level. In social work, the major exponents of this

approach include Australians Healy (2000), Pease and Fook (1999) and Fook (2002, 2012). Newer perspectives including human rights, reflexive materialism, and critical race theory are covered in a volume, which includes authors from several different country contexts (Hick et al., 2005).

To a large extent, these varied perspectives are also associated with different country contexts, although there is some overlap. For instance, antioppressive social work is mostly written about from Britain, and is used to some extent in Canada. The term 'structural social work' originated from Canadian writers as noted above. Postmodern and postmodern critical perspective have largely been developed by Australian writers (e.g., Fook, 2002; Healy, 2000; Pease and Fook, 1999). Neo-Marxist approaches are having a comeback in the United Kingdom (e.g., Ferguson, 2008). In the United States, it is possible to argue that no specific approach exists, which even roughly equates to critical perspectives (Reisch and Andrews, 2001), although there is significant writing on empowerment (e.g., Simon, 1994). In the United States, there is some argument that critical perspectives have adhered most strongly with postmodern and social constructionist perspectives and it is noteworthy that the term 'critical' is rarely used.

I take a critical perspective, which incorporates postmodern thinking (Fook, 2002, 2012). My approach to critical social work practice is based on an understanding of critical social theory as espoused by Agger (2006). I cover this in the following section.

Key Concepts

Common Ideas Underpinning Critical Social Theory

Agger (2006) argues that there are common ideas shared by many social theorists (e.g., Foucault, Derrida, the Frankfurt School, feminism), which can be loosely grouped together as critical social theory. These include theories termed critical, postmodern, feminist, multicultural, and selected theories from cultural studies. He terms this a critical social theory "theory cluster" (2006: p. 4), which can be identified by seven common features (Agger, 2006: pp. 4–5):

1. Opposition to positivist understandings of knowledge making. Knowledge is not wholly 'objective' (i.e., simply constructed as an empirical 'truth' distinct from and outside of ourselves as knowing human beings). It is also actively made by people (e.g., researchers, theorists, practitioners). We, as members of society are all active participants in creating knowledge from our own social positions and perspectives, and cannot help but bring inherent value stances and biases. (The pronoun 'we' is used to refer to us all as people who are 'social subjects' i.e., as individuals who live in a society in which we participate to cocreate a shared sense of reality.) Knowledge cannot be value free.
2. The possibility of progress. Critical social theories build on the past and present to envisage and create a better future, to be achieved through social and political action. Critical social theory both raises awareness of oppression, and demonstrates the possibility of a different nonoppressive society. It not only seeks to facilitate social change, but also offers an analytical base for change.
3. Domination is both personal and structural, that is, people's everyday lives are affected by structural institutions and constructions such as gender and race.
4. Structural domination is also sustained at the individual level through false consciousness or ideology, (or discourse). Ideological or discursive thinking involves believing that there are ideas which are unchangeable, and therefore, taken-for-granted. Continuing to believe that these ideas cannot be changed therefore serves to maintain existing power relations and inequities. The role of critical social theory is to expose these beliefs (and false consciousness or discourse), thereby creating the basis for social change and transformation. By undermining false consciousness (and revealing the discursive functions of ideas, personal, and collective agency are enabled, bringing about personal and social change.
5. The belief that social change can be brought about by exposing false consciousness at individual levels implies voluntaristic (rather than deterministic) assumptions about people and the possibility of change. It opens the way for social changes to begin at the site of everyday life.
6. A dialectical relationship exists between structure and agency. Although social structures do influence everyday experience and lives, knowledge of how this happens also enables people to change the ways this happens.
7. The focus on the connection between everyday life and social structures also places responsibility on personal liberation, rather than on oppressing others in the name of distant future liberation.

Critical social work, based on the foregoing understanding of critical social theory, is therefore an optimistic approach, based on the possibility of personal and social change for the better, achieved through working together in dialogue. It can be summarized as follows. Critical social work is as primarily concerned with practising in ways, which work toward a society without domination, exploitation, and oppression. It involves understanding how structures dominate and also how people construct and are constructed by changing social structures and relations, recognizing that there may be multiple and diverse constructions of ostensibly similar situations. This understanding of social relations and structures disrupts dominant arrangements, and changes these so that they are more inclusive of different interest groups and perspectives.

Knowledge to inform this understanding is derived in different ways. Empirical knowledge is needed in order to understand how material structures shape lives. But processes of critical self-reflection are also crucial to ensure that dominant structures and relations are challenged in the way they are implicitly enacted in individual people's everyday lives. Communication and dialogue are necessary in order to ensure that diverse perspectives are included in forging new and inclusive ways of working. Thus, in a postmodern and critical social work both the kinds of social changes that are sought, and the ways in which they are enacted, are important. Outcome and process are integral to each other (Fook, 2012: p. 18).

Common Ideas Underpinning Critical Social Work Practice

This approach and these common theoretical underpinnings can be distilled to four key concepts, which are crucial to the

practice of critical social work, namely, knowledge, power, language, and identity creation.

Knowledge

The nature of knowledge and how it is generated are both technical and political issues in critical social work. What constitutes legitimate knowledge (and truth?); who accepts it as such; and what are the legitimate ways to generate and create it involves questions of hierarchies of knowledge and people (and therefore power).

It is important to note that critical perspectives recognize that knowledge can be generated by more scientific; means – commonly accepted methods which purport to discover the nature of 'facts' about our empirical world. However, these 'facts' may also be disputed because not everyone will agree with the methods used to develop them, or indeed, some 'facts' may not be relevant in different historical times, social contexts, or for different players. In this way, the more subjective nature of knowledge, and its generation, must also be taken into account. Knowledge in this sense can never be value free as it is necessarily influenced by who creates it, using what type of methodology, and in what context.

Reflexivity, or our ability to recognize how aspects of ourselves and our contexts influence the kinds of knowledge we produce and value, is an important concept in understanding how we as individuals can actively participate in creating and maintaining knowledge (which may or may not be relevant for other people in other contexts).

Additionally, to promote particular social actions, the knowledge that supports them is promulgated as factual, and therefore, universally agreed upon as true. To do this successfully, alternative perspectives are often suppressed or denied. Thus, knowledge and power are inextricably linked. In this way, views which are touted as 'true' may simply reflect the views of whatever group is dominant. This concept is highly relevant in understanding the operation of knowledge and power in professional life. For example, service users must often conform to a predetermined set of criteria about their situation and themselves as people, in order to qualify for legitimate services under particular policy provisions.

Furthermore, a similar process can occur in a more subtle way, with individual people in their personal and professional lives. For instance, most of us, base our actions, at the micro-level, on barely recognized assumptions, which mirror and support social arrangements. These may not be of our explicit choosing, and can work against our interests, or the best interests of the people we, as social workers, are working to assist. An everyday example of this is the decisions about how we behave in meetings. What culture(s) are we choosing to go along with when we decide where to sit, or who and what to support?

Critical reflection is an important ingredient in becoming aware of how we as individuals may create and support knowledge which excludes different (perhaps marginal) perspectives and which may therefore work to maintain social inequalities. I will discuss this in more detail in the last section on practice applications.

In social work practice, assumptions about knowledge are linked with assumptions about theory, what constitutes legitimate theory, and how it is made distinctive from practice. How we construct theory and practice, and the relationship between the two, is therefore a critical issue (Fook, forthcoming). How ways of understanding the relationship between theory and practice can be empowering or disempowering is an important site for social change.

Power

The concept of power, from a postmodern critical perspective, is seen as a dynamic force, which can be shaped and created differently in different contexts. The Foucauldian concept of power, as something, which is exercised, rather than something which is possessed, captures the essence of this idea. This conception of power as something, which can be made in social interactions is vastly different from the more commodified view, which seems to dominate accepted thinking. In dominant thinking, power is seen to reside with social position, and therefore, empowerment involves transferring power from one (more powerful) person to another (less powerful) person or group. This finite view of power is problematic because it infers that empowering one group necessarily involves disempowering another. This means that all situations, which involve empowering people are necessarily conflictual. However, in a postmodern view, this need not be the case. Different people and groups, starting from different social and power positions, may actively work together to create more power and in the process empower each other.

We discussed the relationship between knowledge and power above and it is important to remember that they should not be separated. A site for creating power and empowering others or ourselves, therefore, lies in how we conceptualize power, other people's power, and our own abilities to influence or act in a situation. This understanding of personal agency is crucial to a critical perspective on power.

Language

How we speak about our world and social phenomena is integral to our understanding of how we actively participate in shaping social structures and relations. Firstly, language involves labelling and categorizing which orders the phenomena we experience. Secondly, language functions to socially recognize phenomena – if something is not named, it is harder to speak or communicate about it, and therefore harder to legitimize or value it.

Furthermore, language is about more than words or phrases. It also includes implications about the phenomena we speak about. This is where the concept of discourse becomes important. Discourse refers to the ways of speaking about, and the messages, which are conveyed about phenomena. In this sense, it is about more than the specific words used in the language itself, but may involve nonverbal messages, implied messages, or messages communicated by the way a situation is structured physically. It may also be about more hidden meanings that are assumed in any given cultures, which are operating in a particular situation. For instance, when binary ('either – or' type) categories are used, it is implied that there is no other category or way of being. Often the two elements of the binary are pitted against each other, which suggests that they are mutually exclusive and sometimes opposed to each other. Hence, when we use binary categories, such as 'theory-practice,' we get the message that the two are very different and that much education and intellectual gymnastics go into

'integrating' the two. Such a split can support other binary categorizations (e.g., the split between academia and practice, or the split between research and practice). So where does 'practice theory' fit in such a split? In this way, it is easy to see how any more complex ways of understanding theory and practice may be easily left out. It is this leaving out of other perspectives, categories, or experiences which can also act as a form of power, to silence or devalue alternative views.

When we recognize the power of discourse, and also the power of dominant discourses to discount different perspectives, the concept of narrative becomes important. Narrative in a simple sense refers to the idea that whatever perspective is being touted represents that person's own story, told from their own social position, experience, and background. Thus, narratives may only represent one perspective, but may also represent a perspective, which has not formerly been recognized. This becomes particularly significant when people have not been allowed to tell their own narrative about themselves, their own biography, and identity. The concept of narrative is an important concept in applying the notion of discourse to practice. This will be discussed further in the last section.

Identity Creation

People's sense of themselves and who they are socially and personally can change over time, and can also include contradictions. Identity making involves both personal and social elements, and these can be reworked in the light of new experiences. This means that people do have the opportunity to remake a sense of themselves in relation to changing social contexts. It is possible that identities can:

- Change
- Be contradictory
- Be multiple

This perspective represents a major departure from the accepted way of understanding identity, which assumes that people hold an idea of themselves, which remains relatively coherent and constant over time. Indeed, it may sometimes be assumed that a stable identity is crucial to mental health. However, from a postmodern and critical perspective, we acknowledge that because identity is made in context, then an important aspect of coping in life may involve adjusting a sense of self in relation to changing experiences and contexts.

Since identities are changeable, it is important to understand how identities are made, and how these processes can be crucial in constructing social difference (and also in creating discrimination, inclusion, and exclusion). The process of 'othering' is a process whereby individual people or groups reinforce their existing identities by creating binary oppositional categories for other people or phenomena, which do not fit their preconceived notions. The 'other' group is by definition different. This type of thinking preserves and strengthens boundaries, and presumably, social comfort and certainty. This then forms a basis for discrimination.

Identity politics, as mentioned in relation to the concept of narrative, is therefore important in critical social work. How, and to whom, we accord power to make their own identities, and to coin their own language, becomes an important aspect of making social changes in our everyday lives. And how and whether we recognize and value these new categories are also an important part of critical practice. (see Fook, 2012: Chapter 6 for a fuller discussion of terms and concepts used in this section.)

Major Practice Applications

There are four major practice strategies which derive from the key concepts as follows: critical reflection and deconstruction/reconstruction, problematization and research, narrativity, and contextuality.

Critical Reflection and Deconstruction/Reconstruction

As discussed earlier, critical social work includes the notion that knowledge cannot be value free, and that the way knowledge is constructed, and the sorts of knowledge (theory), which is believed to be true, often reflects the exercise of power. In the practice of critical social work, it is therefore important to be able to uncover how this power operates through the knowledge that we choose to believe. Deconstructing language and ways of thinking to discover the operation of power is a method integral to critical social work practice.

In essence, deconstruction is a form of analysing discourses, or "uncovering the ways we talk about and choose to label experience, and how these shape experience" (Fook, 2012: p. 105). In other words, deconstruction involves uncovering the ways that power operates through the ways we label and interpret our experience. What preconceived labels or categories do we accept and choose to maintain? What are we leaving out? What is being distorted? What social rules are being affirmed or created? Whose perspectives are they and what do they have to do with power differences and inequalities? These are the sorts of questions, the answers to which help pull apart our own language and thinking to see how power operates in our own experience.

In the critical reflection model I developed (e.g., Fook and Gardner, 2007), the discourse being analyzed is a person's narrative about their own experience. Although there are many different perspectives on critical reflection, in critical social work practice, it is relevant to theorize critical reflection as a type of deconstruction of the narratives of personal experience, in order to uncover the hidden assumptions, which inform our practice, and therefore the hidden ways we participate in creating power. These can exist at a very deep and personal level, and uncovering these may help reveal how we can be complicit in supporting views or actions, which can also work against us or the people we as social workers are trying to help.

My model includes two stages (Fook, 2012): the first is deconstruction, where hidden assumptions are uncovered and examined; the second is reconstruction when assumptions (thinking) may be changed, and new ways of practice developed accordingly. This second stage of reconstruction must be an integral part of any critical social work practice. Deconstructing practice experience and thinking must be accompanied by concrete ideas of how practice can be rebuilt on the basis of what is discovered. Critical social theory in this sense provides an analysis, a way of enacting this analysis, and a blueprint for making social changes toward a more equitable and inclusive

society. These are contained in a structured and purposeful way, through critical reflection.

Problematization and Research

How we conceptualize and understand the daily 'problems' that we deal with as social workers is an integral part of the way we work as practitioners and researchers. Much has been made in the past of the way professionals 'pathologize' people and problems, often through a process of 'othering' so that any behavior or culture, which appears deviant from the norm is categorized as problematic. How then do we choose to understand and categorize social and personal issues in a way which respects them from an insider's perspective, but which also values the uniqueness of the perspective enough so that it can be acted upon effectively in a broader social context?

A first step is to see the task of conceptualizing and understanding an issue first and foremost as a research task. The aim of the research is to form a complex and deep understanding of the issue from the point of view of the person/people experiencing it. The task of the researcher is to stay open to discovering what this perspective is, and to use methods, which enable this to occur. This means starting with as little preconceived idea as possible, to listen to and accept the language of the person/people involved. It may also mean using communicative and investigative methods, which are most culturally comfortable or appropriate. These will vary greatly depending on your (the researcher's) gender, social context, age, ethnicity, and so forth, and those of the person/people you are interacting with. It may also mean accepting apparent contradictions (or even new viewpoints which differ vastly from your own), multiple perspectives, and changing interpretations (in line with the postmodern acceptance of fluid identities). Rather than glossing over or leaving out contradictions or changing viewpoints, the task of the researcher is to provide a communicable understanding of the issue or experience, which includes these differences. A complex picture (which derives from the person or situation itself) is aimed at, rather than a consistent or even balanced view.

The second task of the practitioner/researcher is then to mold this understanding into a narrative, which can be communicated to those who need to hear it. Who this audience is will depend on context. For instance, a social worker developing an assessment of a person to receive some form of social assistance may wish to use the language of the relevant policy, and choose to leave out aspects, which will diminish the person's chances of receiving the assistance. Or aspects that appear not to fit may be reworked in appropriate language. An audience of policy makers may need to hear about aspects of individual experiences, which sit outside the norm, so that their own critical thinking can be stimulated. A target group of social work colleagues might need to know about supporting statistics which will broaden their understanding of the social extent of the issue at hand, and give a context to their practice experience.

This process of coming to a comprehensive appreciation of social problems and issues from insiders' perspectives may be used equally well for assessment of individual or community situations. The key is to view the assessment not as an attempt to pathologize or fit the issue at hand into a preconceived category, but to be open to creating new categories or labels, and ways of communicating about the issue or situation, which suit the context in which the assessment is being made. This recognizes that conducting an 'assessment' is really more of a construction of a professional narrative, constructed to be effective, and relevant in a particular context.

Narrativity

The concept of narrativity refers to an understanding that a perspective on particular phenomena may represent only one perspective from a particular vantage point – social, historical, or political. Narratives, constructed as they are by particular people in particular contexts, function as a form of meaning making. Since the meaning is made in a social context, narratives also perform social functions. Narratives, like discourses, embody power. The principle of narrativity builds on this awareness to suggest strategies for destabilizing this power, and this provides the opportunity to remake it.

The principle of narrativity is related to concepts of discourse analysis, deconstruction, and critical reflection. However, it has special mention here as narrative therapy (see White and Epston, 1990) has become particularly well developed as a movement in its own right. The approach is based on the idea of examining and changing harmful narratives. In summary, it involves: identifying narratives, dominant, empowering, and disempowering; validating the positive ones and externalizing the harmful ones; and rebuilding alternative narratives and validating these further by creating an audience. Such an approach has been used successfully in many different fields of therapeutic work (Duvall and Beres, 2011). Narrative methods, however, can also be used more broadly, in work with communities and also in organizations, and are, therefore, an important component of the critical social worker's repertoire.

Contextuality

The principle of contextuality arises through recognizing that context influences much of analysis and practice in social work. Contextual practice (practice both within and with context) emphasizes several new ways of approaching practice (Fook, 2012: p. 162):

- Understanding the nature of the different contexts in which we work at different times and to developing relevant practice strategies. This may involve recognizing the role of a particular context in shaping certain attitudes or practices; or it may require adapting our ways of working to incorporate methods from past times or other contexts.
- Positionality – assuming a reflexively aware stance, simultaneously being able to see outside but also within contexts. This involves being able to understand and identify how we are inextricably part of a situation can also allow us to see outside it, and to use this dual awareness to be sympathetic to any resistance to change, but courageous enough to seek change.
- Working with whole contexts (rather than a number of disparate players or aspects). This involves the ability to see beyond the separate elements of a context to appreciate how they work together to create a dynamic whole. This enables working with how people work together, rather than working separately with the people themselves.

- Developing practice knowledge/theory that is transferable (not generalizable) between different contexts. This involves understanding that preexisting knowledge is not imposed top-down or from outside a situation, but rather preexisting knowledge is remade as relevant in new situations. New contexts are not made to fit preexisting knowledge, but rather preexisting knowledge is constantly remade from experience with new contexts.
- Reframing skills in contextual terms – developing 'contextual competence' is the ability to read the cultural climate of different contexts and to practice effectively with and within these contexts (Fook, 2012: p. 187).

Conclusion

Critical social work practice is still developing. Outlined here are some ideas about what it might look like using a conception of critical social theory, which incorporates common aspects from a number of related contemporary theoretical approaches (including postmodernism, feminism). An important concluding thought is that the practice of critical social work requires an openness to new experiences, people, and situations which allows engagement with each other in contexts of uncertainty which are often outside our control.

See also: Intersectionality; Radical Social Work; Social Work Theory.

Bibliography

Agger, B., 2006. Critical Social Theories, second ed. Paradigm, Boulder.
Allen, J., Briskman, L., Pease, B. (Eds.), 2009. Critical Social Work, second ed. Allen & Unwin, Syndey.
Bailey, R., Brake, M. (Eds.), 1975. Radical Social Work. Edward Arnold, London.
Corrigan, P., Leonard, P., 1978. Social Work Practice under Capitalism. Macmillan, London.
Dominelli, L., 1988. Anti-Racist Social Work. Macmillan, London.
Dominelli, L., 2002. Anti-Oppressive Social Work. Palgrave Macmillan, Basingstoke.
Dominelli, L., MacLeod, E., 1982. Feminist Social Work. Macmillan, London.
Duvall, J., Beres, L., 2011. Innovations I Narrative Therapy. Norton, New York.
Ferguson, I., 2008. Reclaiming Social Work: Challenging Neo-Liberalism and Promoting Social Justice. Sage, London.
Fook, J., 1993. Radical Casework: A Theory of Practice. Allen & Unwin, Sydney.
Fook, J., 2002. Social Work: Critical Theory and Practice. Sage, London.
Fook, J., 2012. Social Work: A Critical Approach to Practice, second ed. Sage, London.
Fook, J., forthcoming. The relationship between theory and practice. In: Pinto, C., Carvalho, M. (Eds.), Social Work Themes: Research and Practices, PACTOR.
Fook, J., Gardner, F., 2007. Practising Critical Reflection. Open University Press, Maidenhead.
Healy, K., 2000. Social Work Practices. Sage, London.
Hick, S., Fook, J., Pozzuto, R. (Eds.), 2005. Social Work: A Critical Turn. Thompson Educational Publishers, Toronto.
Morris, J., 1991. Pride against Prejudice: Transforming Attitudes to Disability. The Women's Press, London.
Mullaly, B., 2007. The New Structural Social Work, third ed. Oxford University Press, Toronto.
Oliver, M., 1990. The Politics of Disability. Macmillan, London.
Pease, B., Fook, J. (Eds.), 1999. Transforming Social Work Practice. Allen & Unwin, Sydney.
Reisch, M., Andrews, J., 2001. The Road Not Taken: A History of Radical Social Work in the USA. Routledge-Brunner, Philadelphia.
Simon, B., 1994. The Empowerment Tradition in American Social Work. Columbia University Press, New York.
White, M., Epston, D., 1990. Narrative Means to Therapeutic Ends. Norton, New York.

Critical Theory

Darrell P Arnold, St. Thomas University, Miami Gardens, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Definitions

Colonization of the lifeworld The system/lifeworld distinction is fundamental to Habermas's understanding of social coordination in modern societies. In such societies this coordination occurs, on the one hand through social systems, which embody instrumental knowledge, and on the other through the conscious steering of individual agents who are guided by norms that have been learned in the course of cultural evolution in reference to 'communicative action' and that have become embodied in the lifeworld. In Habermas's view the 'colonization of the lifeworld' occurs in modern capitalist societies as the instrumental and commodity logic of the capitalist system comes to dominate lifeworld relations, creating social pathologies. Within critical theory it is widely thought that such a 'colonization of the lifeworld' results in what Horkheimer describes as the 'fully administered life' and Marcuse identifies with 'one dimensionality.'

Discourse ethics This neo-Kantian form of ethical thinking is 'deontological, cognitivist, formalist, and universalist.' Rather than dealing with all questions of a good life, its focus is on problems of right or just action. Discourse ethics treats normative claims of rightness analogously to truth claims. In the place of Kant's categorical imperative, its principle is "only those norms may claim to be valid that could meet with the consent of all affected in their role as participants in a practical discourse" (Habermas, 1990, 196–7).

Ideology Ideology in the Left Hegelian tradition is not identified with 'worldviews' as it is in everyday language. Rather the term is used pejoratively to describe distorted worldviews wherein ideas are represented as universally valid or true when they in fact reflect particularistic interests of dominant social groups.

Ideology critique Rather than examining ideas as embodiments of value-neutral truths or looking for how they reflect universal characteristics of the transcendental ego, those engaged in ideology critique examine ideas and prevailing interpretive schema to see the ways in which they reflect the interests of socially dominant groups. In the tradition of critical theory, it has typically been maintained that unveiling the particularistic interests behind such allegedly universal ideas serves to empower social agents to enact social change.

Immanent transcendence In a formulation by Hiliary Putnam that Habermas approvingly cites, "Reason is…both immanent (not to be found outside of concrete language games) and transcendent (a regulative idea that we use to criticize the conduct of all activities and institutions)" (cp. Habermas, 1992, 139). The concept of immanent transcendence, also often referred to as intermundane transcendence, implies in Habermas's terms that "validity claims raised hic et nunc and aimed at intersubjective recognition or acceptance can, at the same time, overshoot local standards…." (Habermas, 1988, 15).

Recognition Recognition is the English translation of the German term 'Anerkennung,' which serves in Hegel's *Philosophy of Spirit* as a goal toward which mind is moving in the process of cultural and social evolution. In Hegel's system, upon achieving recognition, mind comes to see itself in all of its creations. Recently, Axel Honneth has made a reformulation of this Hegelian idea the centerpiece of his critical theory, maintaining that it is anchored in human anthropology, establishing a pretheoretical desire for (1) care, (2) self-respect, and (3) self-esteem. The concrete understanding of what recognition in these specific areas entails is worked out in human interactions under specific historical and social conditions.

Rational reconstruction Rational reconstruction is a method of theory construction that aims to formulate implicit, pretheoretical knowledge, practical know-how, or rules needed for engaging in social practices or communication. Habermas views rational reconstruction as allowing 'weak transcendental arguments,' which, however, are fallibilist. Often such reconstructions aim to disclose universal or species competencies or rules. Among other things, pretheoretical knowledge in the existing forms of intramundane transcendence is made explicit in reconstructions (see Habermas, 1990, 32f.).

Abstract

Critical theory is a philosophical social scientific research program that engages in explanation against a background of normative interests. While Max Horkheimer conceived of the original research program as an attempt to revive a form of nonorthodox Marxism, Jürgen Habermas, the main representative of the second generation of critical theory, has developed a research program on the basis of a rational reconstruction of competencies of communicative action while increasingly assimilating the work of American pragmatists. Current critical theorists – besides Habermas – see a need to reconstruct the original research program, while retaining its orientation as a normatively driven theory.

Introduction

'Critical theory' refers first and foremost to the social scientific research program guided at its inception especially by the vision of Max Horkheimer, who was an early director of the Frankfurt Institute for Social Research, later known as the 'Frankfurt School.' Horkheimer envisioned an interdisciplinary research program, at once philosophically normative and empirically based. The research program was conceived as an attempt to revitalize Marxist analysis, while rejecting orthodox Marxism and critically appropriating especially the work of Freud, whose approach was viewed as complimentary to Marx's in that it too was both socially scientific and oriented toward human emancipation. Rather than focusing exclusively on issues of economic redistribution, critical theorists further emphasize 'psychological, cultural, and anthropological sources of unhappiness' (Bronner 2011, 29). The further development of the work of critical theory is represented especially in the social research program of Jürgen Habermas and the moral philosophical work of Axel Honneth.

More recently, the terms 'critical theory' and 'critical social inquiry' have been used to refer to various approaches in the social sciences and humanities that have family resemblances to the original work of the Frankfurt School. Such critical theories include critical race theory; some feminist work in the philosophy of science; the socially critical work of Michael Foucault, Jean Baudrillard, and some other postmodern social theorists; and the work of various contemporary nonorthodox Marxist social critics. Given the lack of a specific unifying research program, it is most appropriate now 'to speak of critical theory in the plural' (Müller-Doohm, 2004, 294).

Since its onset critical theory has normally shared certain characteristics with Marx's criticism: (1) Critical theorists have viewed themselves as continuing the modernist project. (2) They have linked theory and practice, both noting the effects of social practices on theory formation and formulating theory with a view to emancipating the marginalized. (3) Most often they have proposed dialectic theories. (4) They have operated with a view that there are forms of immanent transcendence available that facilitate critique. Consequently, an effective critique of society is not reliant on external criticism, but is possible using norms already accepted within a given society. (5) They have engaged in ideology critique, attempting to unveil the ways in which prevalent ideas in many social domains illegitimately facilitate the control of some groups by others. Unveiling such ideology has been viewed as having the potential to empower those who see the ideology to engage in social transformation. (6) Most critical theorists have assumed that there are pathological features specific to capitalism that prevent the self-actualization of many members of society and that changing the economic system will eradicate many of the pathological features of modern life.

Axel Honneth, the current director of the Frankfurt Institute for Social Research, expresses a widely held view among critical theorists that the original program of critical theory is no longer plausible and that in order for the tradition to continue, it must be reconstructed. In Honneth's view, any useful reconstruction will need to take up themes of Left Hegelianism, which differentiate critical social inquiry from other forms of social criticism: A viable critical theory will show how reason is embedded in history (the idea of immanent transcendence), indicate how the lack of the full realization of reason in society leads to social pathologies, and demonstrate that humans have some form of 'emancipatory interest' in overcoming such social pathologies. Such a reconstruction, integrating several of the elements noted above, is what he envisions as the task for the heirs of critical theory in the twenty-first century (cp. Honneth, 2009).

Intellectual Backgrounds

The Frankfurt School

The Frankfurt Institute for Social Research was founded in 1923 as a Marxist think tank under the direction of the economist Kurt Albert Gerlack. Important members and participants in the early work group included two of the most important 'Western Marxists' from the early twentieth century, Karl Korch and György Lukács, along with Friedrich Pollock, who continued to contribute to research in the school after Horkheimer assumed directorship of the institute in 1931. Only Lukács's work is read today. While the institute's early work was 'orthodox Marxist' in orientation, once Horkheimer's assumed the directorship, the institute moved to a Freudian-Marxist influenced position. Under Horkheimer's direction, the institute published the *Zeitschrift für Sozialforschung*, the journal that served as a conduit for seminal papers reflecting the research program. During the Nazi regime, the institute was relocated to Columbia University. In 1949, Max Horkheimer and Theodor Adorno returned to Frankfurt to reopen the institute there. Various thinkers who had become identified with the Frankfurt School (as the school came to be known under Horkheimer's direction) went their separate ways throughout this period of exile. Throughout the tumultuous history, the originally shared research program largely broke down, and even Horkheimer increasingly distanced himself from it.

The interdisciplinary early research in the Frankfurt School included Theodor Adorno's work on aesthetics and on the authoritarian personality; Herbert Marcuse's writings linking material structures to the phenomenological structures of human thought in *Ontology and the Theory of Historocity* (1932); Friedrich Pollock's work on 'state capitalism,' a term that referred to the administration of the economy even through the apparatus of the state, which is lead by instrumental logic of capital and results in conditions of alienation and exploitation; Leo Lowenthal's sociology of literature and mass culture; and Otto Kirchheimer's and Franz Neumann's writings on philosophy of law. Eric Fromm was formally a part of the institute for a short time until 1940, and his treatment of authoritarianism in *Escape from Freedom* (1941) and later work after a break from the institute largely reflect many of the concerns of the research program. Walter Benjamin never formally belonged to the Institute for Social Research, but he was intensely involved with several of the institute's members, especially Adorno.

Horkheimer's Early Research Program

The main outlines of Horkheimer's initial research program were clearly articulated in essays throughout the 1930s, including 'Materialism and Metaphysics' (1933) and 'Traditional and

Critical Theory' (1933). In general, Horkheimer is critical both of what he views as irrationalist forms of hermeneutical thought as well as the scientistic tendency of positivism, which reduces reason to its instrumental function and brackets rational reflection on social questions. In contrast to 'traditional theorists,' Horkheimer thinks that the rational development of ideas does not occur in a normative-free space but against a background of real material interests. Rejecting the view that knowledge develops unhinged from interests and values, critical theory aims not just at an abstract truth but at a theory that can 'liberate human beings from the circumstances that enslave them' (246). The themes of Horkheimer's early program are Left Hegelian. In fact, he emphasizes the need for a dialectical philosophy that recognizes 'that the free development of individuals depends on the rational constitution of society' (Horkheimer and Adorno, 1972, 246). Horkheimer's view unites theory and practice, for one by aiming theory at practical (normative) results, but for another by unmasking traces of human practices (and norms) in the world of theory.

Horkheimer's critical philosophy and social theory does accept in broad outlines the Kantian and Hegelian view that a full scientific understanding of the world requires an understanding of the subjects that produce knowledge. While in Kant, this famously entailed an analysis of the universal categorical structure required for generating human knowledge, in critical theory the very structure of human thought determinations itself is viewed as socially conditioned and as especially related to the economic base of society. Therefore, the mental schema necessary for an interpretative experience of reality are not viewed as ahistorical or transhistorical elements of a transcendental ego but as part of a socially conditioned and historically malleable mental architecture that, among other things, reflects power schema in the social world. As a result of this, any sufficiently critical philosophy will entail not just an understanding of the prevalent views of science, and not just an understanding of thought schema used by human subjects (the focus of the 'critical' studies of rationality in German idealism), but more fully an understanding of the social conditions that come to be reflected in the thought schema necessary for generating theory and in the theories generated with such mental schema.

The approach does retain something of the attempt to German ideology to think the entire system of reality. However, Kant's category teaching gets replaced by a sociology of knowledge – and in particular one at that entails ideology critique or a 'hermeneutics of suspicion' while aiming at social emancipation. Rational reflection on the social basis for the mental schema is thought to be a necessary supplement to reflection on the results of science. Therefore, the products of law, art, media, and science are not to be evaluated only on discipline-internal criteria. While there is social learning according to discipline-internal standards, it occurs always against a backdrop of human interests and values.

Moreover, critical theory tends to entail that more rational forms of social interaction are a prerequisite for more adequate forms of theory. In short, the counterpart to the idea that bad social practices become mirrored in theoretical ideas, and in the very theoretical schema used in the appropriation of knowledge, is that good ones do as well. Therefore, in the early work an impulse is found that adequate theory requires more humane social practices. Truth is linked to justice. Indeed this is a view already implied in the work of Hegel and Marx.

The Negative Turn in Horkheimer and Adorno

Although Horkheimer set the agenda for the first-generation critical theoreticians while in Frankfurt, he progressively was influenced by Adorno, and by the time of the writing of the Horkheimer–Adorno coauthored *Dialiectic of Enlightenment* (published in mimeograph form in 1944 and as a first-edition book in 1947) and Horkheimer's *Eclipse of Reason* (1947) much of Horkheimer's original agenda had been set aside, as he had become increasingly negative about the possibilities for escaping the 'iron cage' of instrumental rationality in late capitalism.

The *Dialectic of Enlightenment* aimed to provide 'the discovery of why mankind, instead of entering into a truly human condition, is sinking into a new kind of barbarism' (Horkheimer and Adorno, 1972, xiv). While Horkheimer and Adorno by no means come to reject rationality and indeed go so far in their support of the enlightenment as to maintain that 'freedom in society is inseparable from enlightenment thinking' (xvi), they view the *de facto* course of the development of reason as having resulted in the domination of nature and of man. They thus engage in their work in the *Dialectic of Enlightenment* for the somber reason that it is the demand of critical reason 'to take up the cause of the remnants of freedom' even if it seems 'powerless in the face of the great historical trend' (Horkheimer and Adorno, 1972, xi).

While indicating that the instrumentalization of reason and the administration of all of life do indeed come to a head in capitalism, *Dialectic of Enlightenment* differs from the earlier work of the institute because it offers not a local analysis of forms of thought and problems in late capitalism, but a world-historical view. The roots of the process of the instrumentalization of reason and of the present crisis have a much older origin. In fact, early myth is even implicated in the process as a first step toward the emancipation from nature.

A description of the rationalization process of Western civilization – building on Weber's work – is central to the book, accompanied by an explication of the domination of nature and of man that results from this. While the enlightenment has aimed at autonomy, in its *de facto* development it 'radiates disaster triumphant' (Horkheimer and Adorno, 1972, 1) and has even become 'totalitarian' (Horkheimer and Adorno, 1972, 4). The instrumentalization of reason leads to a fully administrated world and 'makes humans into beasts' (Horkheimer and Adorno, 1972, 185). The end result, as they discuss it, could have been written by Heilbroner or other later environmental survivalists: "Either the human species will tear itself to pieces or it will take all the earth's fauna and flora down with it…" (Horkheimer and Adorno, 1972, 186).

Despite this prevailing pessimism, the book and this period of the writing of Horkheimer and Adorno more generally do contain a dichotomy between the radically negative characterization of the results of the domination of instrumental rationality and a quite different positive and expanded characterization of a potentially emancipating reason, which always exists as an alternative to the instrumentalizing concept, as a parallel running enlightenment that could control the

domination of man and nature, but that never manages to do so (cp. Wiggershaus, 2007, 332ff.).

While Horkheimer turned increasingly to theology in his late work, Adorno, in *Minima Moralia: Reflections from Damaged Life* (1951) and *Negative Dialectics* (1966), if anything, became even more pessimistic. His view progressively became that contemporary life is so damaged that the most that can be hoped for is that individuals engaged in nonpurposive activity – in art, in acts of love, in unselfish giving – achieve some 'human cells within universal inhumanity' (Adorno, 2005, 31).

Current Knowledge

Habermas

Habermas, the most well known of the second-generation critical theorists, viewed this negative turn and abandonment of the earlier social scientific research program as unfortunate. Noting that there was no existing social research program when he arrived at the Institute for Social Research (Wiggershaus, 2007, 544), he did begin to revive critical theory in alignment with the scientific aspirations of the earlier work, even while developing this in a quite different direction than the first-generation theorists. His work is encyclopedic, like that envisioned as programmatic for the early Frankfurt School, including work on numerous aspects of the history of philosophy, ethics, political theory, philosophy of law, philosophy of mind, psychology, and religion. His most influential writings, however, focus on communicative reason. His own research program, which he views as a fallibilistic continuation of the enlightenment project, has developed in part through the critical appropriation of work in German social theory, American pragmatism, and Kantianism.

While Habermas does take up the view implicit in Horkheimer and Adorno's earlier work, he, in contrast to his former teachers, engages in a justification of the enlightenment project. Like the earlier critical theorists, Habermas is critical of positivism for the ways that it views scientific knowledge as interest-free and of hermeneutics and its heirs for its relativism. He thinks that in the end both traditions undermine rational thought about ethics – the former tradition because 'ethical thought' is not scientific, the latter because it views ethical commitment as based on decisionism. In his early epistemological work *Theory and Practice* (1963) and later in *Knowledge and Human Interest* (1968), he draws up a position counter to both positivist and hermeneutic theories while working through various issues that arose in *The Positivist Dispute in German Sociology* (1969/1976). In particular, under the influence of American pragmatism, Habermas argues for the interest-laden character of all knowledge, maintaining that knowledge in the various human domains is rooted in three 'anthropologically deep-seated interests': the interest in instrumental control, in understanding, and in emancipation. Each of these anthropologically seated interests tracks to a particular intellectual domain. The interests in instrumental control is at the basis of the development of the natural sciences. The interest in understanding serves as the basis for the human sciences. The interest in emancipation serves as the basis for moral-practical knowledge. Herewith he slices up intellectual activity in alignment with Dilthey's view that there are different methods for the natural and the human sciences, while adding an 'emancipatory interest' as the basis of a third form of knowledge. This latter interest is especially evident in ideology critique and psychoanalysis, both of which aim at the emancipation of the individual from social and individual psychological burdens.

Although Habermas later abandons this anthropological oriented theory of this early work for a more Kantian quasi-transcendental grounding of knowledge in reference to forms of social action based in linguistic competencies, he retains the intellectual division of labor of the earlier schema. Habermas employs a method of rational reconstruction in which he elucidates the presuppositions for various types of human action. These allow a form of 'weak transcendental' argument. However, unlike the Kantian program, the rational reconstruction results in falsifiable claims, not elements of a 'first philosophy' in the tradition of the philosophy of the subject, which Habermas thinks brackets social inquiry. The forms of action in Habermas's mature theory are viewed as preconditions for the type of social action that we already *de facto* engage in. In our social life, we do not merely use reason in a rational purposive manner (in action oriented toward success). We also give and take reasons in an attempt to reach agreement about what is right (in action oriented toward consent) (cp. Habermas, 1984, 285). This act of giving and accepting reasons entails various presuppositions. We presuppose our own free agency. We presuppose that there is one world that we are referring to. We presuppose that there are better and worse answers not only to scientific but also to the ethical/social questions we pose.

More specifically, communicative action involves three types of validity claims, which for their part are tracked to three types of speech acts: constative speech acts entail a claim to truth, regulative speech acts contain a claim to normative rightness, and expressive speech acts entail a claim to truthfulness (Habermas, 1984, 319–20). In the claim to normative rightness, later called the 'discourse principle' (Habermas, 1998, 41), a claim is justified only if all could agree to it in an unconstrained dialogue, or an ideal speech situation. Those who do provide or accept reasons for actions are thought to necessarily presuppose these to make sense of their own reason giving. This latter form of action, similarly to the earlier 'emancipatory interest,' tracks to the 'subterranean enlightenment' of Horkheimer and Adorno's *Dialectic of the Enlightenment*. In Habermas's view, it is not merely subterranean and forever unable to gain control over rational purposive action. Rather, it too has permeated civil society, and it influences law and politics and provides real possibilities for emancipation.

In the fuller rational reconstruction of our competencies, Habermas argues that in the act of giving and accepting reasons for action, we *de facto* assume that we are agents, responsive to reasons. We assume active stances toward our desires and are not merely passive victims of them. We assume that through a process of rational reflection, we are able to come to 'reflectively endorse' certain actions, that is, that we can commit ourselves to them as being rationally worthy of support. Finally, we understand this process of reflectively endorsing ideas not as merely playing out in a human subject, cut off from the social world, but rather as subjects engage in reflection with other reflective agents in a social world where certain norms

and reasons are considered acceptable (cp. Baynes, 2004, 202). In contrast to the view that these presuppositions for action adhere merely because we are heirs to the practices of a given tradition, the quasi-transcendental view sees such presuppositions as universal. In this way, Habermas can avoid falling back into a mere hermeneutics or a relativist pragmatism a la Rorty.

Throughout his career, Habermas attempted to avoid relativist conclusions like those he sees as inherent in hermeneutics, yet he does not for that dismiss hermeneutics, but rather critically appropriates it. In Habermas, the giving and taking of reasons plays out in the lifeworld, where the interpretive schema of our traditions play a fundamental role. They serve as backdrops for viewing the world and imagining possible social action. While we cannot escape this backdrop, neither are we completely bound by it. Rather, it functions by establishing constraints and enabling conditions for our action and our interpretations and explanations of the world. Habermas's view is that individuals not only actively appropriate ideas from their traditions but also critically reflect upon them. In the give and take of rational argumentative processes, we are always open to the fact that justifications might challenge the prevailing interpretations of the world, as embodied in our traditions, and lead to a revision of the interpretive framework against which we view the world and against which our wills are formed.

Habermas and German Social Theory

Habermas's theory does develop through a critical appropriation of various stands of social theory and philosophy of the late twentieth century besides hermeneutics. In contrast to the proponents of hermeneutics, he argues for a thicker view of reason. He also argues that their focus on linguistic 'interpretation' renders hermeneutics a form of idealism that fails to duly account for the ways in which social institutions come to embody certain rational norms and values and to influence our lives.

Besides appropriating elements of hermeneutics, Habermas also integrates work from sociological systems theory, although, in the end, adopting a version closer to that of Parsons than of Parsons's German student, Niklas Luhmann, the most important figure in European social systems theory in the late twentieth century. While Habermas recognizes that system dynamics establish something like hard-wired tendencies for certain domains, like Parsons, he still maintains a fundamental role for human agency.

Luhmann's version of systems theory views social systems as functioning on the model of cybernetic theory. Social systems are organized wholes that assimilate information from the world according to simple binary codes; in this way, they simplify the world in such a manner as to facilitate social coordination. Each of these systems works according to its own code. The economy operates on the codes of profit and loss, while the knowledge system operates on the code of true and false, and democratic political systems operate on the code of government and opposition. In Luhmann's view, we must simply recognize these as independent domains, and in fact we have no reason to choose one system over another.

Because the position ultimately offers an extremely truncated space for human agency, and indeed flies in the face of how we understand ourselves as human agents, Habermas contrasts system developments with those in the lifeworld.

While there is a decoupling of system and lifeworld in modern societies, which contributes to social efficiency, learning occurs not only with a view to instrumental action that comes to be embodied in systems but also in reference to communicative action in the lifeworld, allowing democratic, as opposed to merely technocratic social action coordination. It is in the lifeworld where symbolic meaning systems are negotiated that allow collective control over systems.

While Habermas sees a place in the world for social systems, he sees a threat in their overstepping their rightful bounds. A fundamental problem specific to the late capitalist social world is that the system of the economy becomes increasingly pervasive in all spheres of life. The logic of that system comes to dominate lifeworld interactions, and the 'iron cage' of instrumental reason – emphasized by Weber and taken up by Horkheimer and Adorno – thus becomes a dominant filter through which we interpret all social life as we experience the 'colonization of the lifeworld' by the logic of the capitalist economy. This is a tendency that must be overcome if we wish to have autonomy.

Habermas and Postmodernism

Because Habermas views social learning also as occurring in the lifeworld, Habermas is more optimistic than Horkheimer and Adorno, and also the postmodern critics of rationality. In Habermas's view, the postmodern critics have a truncated understanding of what reason is. In the end, they exhibit a sensitivity to difference and to the dominance of instrumental reason, but they undermine any criteria that might be available for supporting their own views that the dominance of instrumental reason is a problem to be taken seriously. Habermas thus views these critics of reason as ultimately involved in a 'performative contradiction.' They argue for the weaknesses of reason, generally viewing the enlightenment tradition as merely the most recent of numerous forms of mythology. If this is the case, there is no need to pay heed to their rational arguments at all. If the enlightenment system of reason giving is a mythological system like any other, then why prefer it to other myths? Habermas's own view is that in giving and accepting reasons, we already do presuppose the transcendental conditions that were earlier outlined. If we fail to do so, we are left with decisionism. Among other things, we thereby also undermine the self-understanding we have as agents.

Procedural View of Reason and Justice

Against postmodern critics, Luhmannian systems theoreticians, and hermeneutists, Habermas argues for continuing the enlightenment project. Rather than arguing for a substantive view of rationality, for example, based in human anthropology, he ultimately argues for a procedural view of reason. Rational procedures are those that promote 'unconstrained dialogue' aimed at understanding. This requires that no individuals be prevented from taking part in the processes of social decision making. A rational, and consequently just, society thus has social institutions that facilitate the collective control of life. In Habermas's view, this provides the basis for the acceptance of

certain procedures of justice in society but does not provide for the mutual acceptance of metaphysically laden thick views of a good life (cp. Habermas, 2005).

This emphasis on the procedural view of reason and justice is characteristic of Habermas's political writings and ethics. His view is that in a pluralistic postmetaphysical political reality, we cannot base a political program on shared thick views of the good. Rather, communicative action can at best secure procedures for social decision making that make a place for all individuals affected by social decisions. This leads him to support a form of deliberative democracy that is inclusive, embracing contemporary values of religious liberty and respect for diversity. He does not presuppose that we come to the social dialogue with a ready-made understanding of what is in the general interest, but that through fair, inclusive procedures we can reach conclusions about what legitimate social and political actions should be undertaken.

While Habermas has increasingly emphasized the importance of specific religious and communal identities to action in civil society and that these are not to be bracketed from in democratic societies, he, like Rawls, continually maintains that the legislative process itself should only reflect those considerations that can be put in language of the public sphere – the language of reason – amenable to all regardless of their thicker views of the good.

Current and Emerging Directions

Contemporary Critical Theory after Habermas

Axel Honneth

The main strands of critical theory today view the first-generation research program as requiring a fundamental reconstruction. Axel Honneth, probably the best known living representative of critical theory after Habermas, has long made strong charges against the Frankfurt School, going so far as to say that "critical theory appears to have become an intellectual artifact" (Honneth, 2009, 336) and noting that "as an interdisciplinary endeavor to critically diagnose social reality, this tradition ceased to exist some time ago" (Honneth, 2007, 63). Still, he thinks that the tradition can be usefully reconstructed for our times.

For one, Honneth argues that first-generation critical theorists largely simply presupposed forms of 'intramundane transcendence' for which the Hegelian tradition had attempted to provide a metaphysical grounding. A major task for critical theory is now to justify the standards of critique in society. Habermas has taken this up, but so have Honneth and others among the new generation of critical theorists.

As Honneth sees it, two major lines have emerged in critical theory in reference to the question of immanent transcendence, each characterizing one of the main tendencies in critical theory today (Honneth, 2007, 66). On the one hand, there is the negativist turn, following Adorno. The most well-known theoretician of this ilk is Foucault, especially in his middle-period writings. There he indicates that we are always subject to exertions of power in oppressive ways but indicates no way to escape this situation. Another of the postmodern critical theorists is Jean Baudrillard, who emphasizes the impossibility of escaping mere simulation in the contemporary world in which symbolic forms are proliferated and reproduced, in the end freed from their referents (Honneth, 2007, 67). Such postmodern forms of thought wield an important critique of power aligned with fundamental tendencies of Frankfurt School critical theory, in particular integrating the genealogical critique traceable back to Nietzsche (cp. Honneth, 2009, 243–253), yet they fail to secure the basis on which critique is possible. Honneth argues, however, that the type of genealogical element found in this work – which examines ways that norms have lost their original meanings and serve as instruments of oppression – is integral to any critical theory today (Honneth, 2009, 51ff.).

On the other hand, there are attempts, like Habermas's, that provide a justification for rational critique. Honneth's work builds on Habermas's as he attempts to show the types of virtue that are necessary for individuals to participate in the ideal forms of dialogue that Habermas envisions and as he argues that "the normative presupposition of all communicative action is to be seen in the acquisition of social recognition" (Honneth, 2007, 71). While Honneth builds on Habermas's theory in fundamental ways, Honneth is critical of Habermas for failing to give an account of individuals' experiences of injustice. This severely limits Habermas's ability to shed light on how social conflict plays a role in the shaping of norms in the lifeworld.

This is fundamental, Honneth argues, because progress in our moral understanding is pushed along by social conflict, and the social protest that in part drives this social learning process occurs not because individuals have a sense that communication is distorted, but because their senses of justice have been violated. Honneth argues that the feelings of injustice emerge because of a lack of recognition – manifest as affronts to dignity, honor, or integrity. It is such feelings of 'social disrespect' that drive protest.

Honneth's system thus elevates the phenomenological moral experience of those involved in social interaction, hereby providing an alternative to Habermas's linguistic-based justification of moral theory. Where Habermas takes a Kantian turn, Honneth takes a Hegelian one, making the concept of recognition the basis of his justification of the moral-critical point of view. He also notes, however, that he is not yet able to justify that the desire for recognition is basic to communicative action. Consequently, his theory is based on an assumption that a desire for recognition is pretheoretically anchored, as a part of basic human anthropology. In the course of the sociohistorical development, specific understandings of what recognition entails are worked out.

The point he would like to emphasize is that when entering into communicative relations, we bring expectations of social recognition (Honneth, 2007, 72). Fundamental to human life is not merely our interest in our individual lives but also an interest in our collective life. We form our identities in intersubjective processes, and our self-actualization occurs not in isolation, but in society. Honneth specifies three dimensions of this concern with recognition in a way that facilitates insights into moral theory from an ethics of care, Kantian or other rights approaches, and the neo-Aristotelian capabilities approach. These dimensions entail (1) 'emotional concern,' (2) 'rights-based recognition,' and (3) 'the social esteem of individual achievements and abilities,' respectively (Honneth, 2007, 74). Recognition in each dimension is necessary for successful self-development.

Nancy Fraser

Other important critical theorists are involved in endeavors similar to Honneth's, focusing on moral argumentation. While Honneth claims that his recognition concept includes issues of redistribution, Nancy Fraser argues that it systematically underplays the significance of economic issues. In short, Honneth extends the concept of recognition 'beyond all recognition' when claiming that it encompasses economic concerns in the same way as the 'Old Left' (see Fraser and Honneth, 2003). Fraser, for her part, develops a model based on a tripartide concern for matters of (1) recognition, (2) redistribution, and (3) representation. Honneth's theory, she argues, reflects the turn toward a politics or recognition that occurred with the 'New Left.' This turn from traditional concern with economic issues did indeed help in the understanding of social justice, but one should not for that downplay economic justice. Beyond these two concerns, Fraser argues, in part assimilating Habermas's work, it is also fundamental that there is fair access to political and cultural processes that shape the framing of political and social issues. In these processes, among other things, we delimit the concerns for justice (cp. Fraser, 2008, 2, 5).

New Issue Areas

Gender issues, along with issues of environmentalism and global governance, are among the areas of special importance in developments in critical theory in recent years.

Gender Issues Reflected in Feminist Philosophy of Science

Seyla Benabib and Nancy Fraser have taken up gender issues in developing critique that clearly identifies with the tradition of critical theory. However, specifically in the area of gender studies other interesting work clearly aligned with critical theory has been done in philosophy of science by theoreticians who do not specifically claim identification with the tradition of the Frankfurt School, but whose work provides a clear example of critical social inquiry. In particular, the works of Sandra Harding and Helen Longino are exemplary of this. Like first-generation critical theorists, both Sandra Harding and Helen Longino counter the positivist view of the value-neutral character of observation, emphasizing that observation occurs against historically contingent background conditions. As Longino notes, "Scientific knowledge…is a body of diverse theories and their articulations onto the world that changes over time in response to the changing cognitive needs of those who develop and use the theories" (Longino, 1996, 276).

It is a mistake, Harding similarly argues, to view the 'facts' produced in the sciences apart from the conditions under which they are produced, for there are social conditions at play at every step in the knowledge formation process. "Culturewide (or nearly culturewide) beliefs function as evidence at every stage in scientific inquiry; in the selection of problems, the formation of hypotheses, the design of research (including the organization of research communities), the collection of data, the interpretation and sorting of data, the decision about when to stop research, the way results of research are reported, and so on" (Harding, 1993, 69). Longino highlights the historical malleability in the standards for research: "The kind of knowledge a community seeks, the purposes for which it seeks it…guide the development of the community's standards" (Longino, 2001, 188). Both thinkers reject the view of the 'unconditioned subject,' who is 'guided by context-independent and value-neutral methodological rules,' and they focus on epistemic communities: since there are no raw facts, but only facts produced against the background of social beliefs, needs, and interests of an epistemic and social community, we must evaluate the background conditions in such a community. This constitutes the 'logic of discovery' that Harding discusses.

Specifically, given the subjective elements in science, Harding and Longino emphasize that scientific institutions would do well to integrate more women at our current juncture. Inclusive scientific communities allow the background conditions to be more adequately examined and allow shared values and interests to emerge. Far from distorting objectivity, introducing a 'women's perspective' thus can improve the quality of science. Concrete examples of this are not lacking, for instance, in anthropology, where opening the field to women researchers unveiled the presence of male biases in research that systematically ignored potential roles of women in the technological development of society by focusing on the development of weapons rather than household utensils, and in sociobiological research, where women working in the field unveiled similar distortions. These arguments in general are more forceful in the human than in the natural sciences.

Much of the work of these feminist philosophies of science corresponds with the basic impulses of the Frankfurt School. They view science and society in some ways as 'co-constituting' of each other, and carry out a hermeneutics of suspicion with the intention of empowering those who gain this knowledge to bring about changes in the institutions of science so that it can better overcome systematic distortions to knowledge that can occur in light of male power structures dominant in society.

Environmental Concern

Critical theorists are also increasingly working in environmental philosophy and environmental political theory. Here in particular two strands of work in environmental philosophy build on the traditions of critical theory. One strand (see Biro, 2011) views Habermas's theory as largely lacking the resources needed for adequate reflection on the environmental problematic. These theoreticians criticize Habermas for viewing our approach to nature as by necessity instrumental, anchored as this is in the view that rational-purpose action is basic to constative speech acts themselves. They also emphasize his lack of general concern with the environmental problematic, which he only addresses in passing references. Consequently, they would like to go back before his work to critically appropriate many of the ideas of the first-generation Frankfurt School. While there are many dimensions to this work, one of the most important is that they view critical theory as providing a sophisticated view of the social construction of reality that steers between the naïve realism found in many varieties of environmental thought and an overly linguistically oriented social constructivism found in others. While, like Kant, critical theorists view 'nature' as constructed, in contrast to Kant, they

emphasize that the human subject that constructs that 'nature' is also constructed: "Neither subject (society, the individual) nor object (nature) is something given; it is always connected with its opposite" (Görg, 2011, 52). As such theoreticians see it, it is thus important to shape societal relationships that are shaping the subject and the subject's understanding of and interaction with the world. Many of these critical theorists attempt to identify the effects of the 'capitalization of nature' on how nature is being constructed in environmental politics – for example, on how capitalization results in the 'selective treatment' of environmental issues in late capitalism (Görg, 2011, 60). In general agreement with Walter Benjamin, Görg argues a point that many of these environmental philosophers accept: "it is not nature we need to master, but rather our societal relationships with nature – that is, we need to control the impact those relationships have on nature" (Görg, 2011, 47).

Another stand of environmental philosophy, represented by John Dryzek and Robyn Eckersley, is more optimistic about the possibility of extending Habermas's model so as to make a place in our political and civic life for those who represent all the 'communities of fate,' including nonhuman members of it, along with members of future generations. Extending critical theory along these lines, John Dryzek argues for a form of strong ecological modernism, which restructures capitalism so as to become environmentally sound. This entails (1) strong changes to the infrastructure and economic systems to make them more responsive to ecological concerns; (2) more open democratic decision-making procedures, also with increased communication about ecological issues; and (3) concern with international aspects of environmental and development problems (Dryzek, 2005a, 173ff.). Given that ecological concerns are not nationally contained, state policies and state publics must take up concerns that transcend their own borders.

Globalization

Much of the work on green politics that is coming from critical theorists draws on further work on deliberative democracy and on the global public sphere that has been done by Habermas and other critical theorists such as James Bohman (2010) and Seyla Benhabib (2006). A lot of this work focuses on how transnational civil society can be used to shape public opinion, determine national policies on transnational matters, and affect the policy of the United Nations (UN) and other transnational governmental bodies.

Most of these critical theorists, like Dryzek, follow Habermas in rejecting David Held's call for a world parliament as impracticable and propose instead forms of transnational cooperation among states, regional governments, and civil society organizations, along with attempts to influence the existing transnational governmental bodies such as the UN and the European Union. Dryzek is skeptical that enough can be done within states, noting 'all liberal democracies currently operate in the context of a capitalist market system' and are consequently 'greatly constrained' in determining policies (Dryzek, 2005b, 634). Consequently, rather than focusing on the state or on extending the mechanism of the state globally in accord with Held's model, Dryzek proposes a grassroots strategy in global civil society, emphasizing the importance of nongovernmental organizations and informal international networks (cp. Bohman, 2010, 42–44). Here, Dryzek emphasizes the value of radicals in the public sphere. Since the radicals are not working within the system of government, they are not as easily forced to compromises, and can more easily speak the truth to power (Dryzek, 2005a, 218–227). It is radicals in the public sphere, he maintains, "who have made the running in instigating change in ideas and attitudes, which extends to those who do not vote for green parties, still less join more radical green actions" (Dryzek, 2005a, 222). Of course, if these attitudes cannot change institutions, then there is little hope that the response will be sufficient.

Taking up global governance in reference to the environmental problematic and other issues is certainly important in globalizing critical theory. However, even the perspective on these issues that is dominant in Frankfurt School-oriented critical theory is largely Eurocentric. Although many in Asia, the Middle East, Africa, and Latin America are strongly influenced by critical theory and are working largely as critical social theoreticians, European and American scholars rarely engage in exchanges with these thinkers. A couple of the theoreticians who have begun looking beyond the Euro-Anglo world and are engaged in critical social inquiry in ways that may facilitate a different kind of global critical theory are Susan Buck-Morss (*Thinking Past Terror: Islamism and Critical Theory on the Left*, 2003) and Eduardo Mendieta (*Global Fragments, Latinamericanisms, and Critical Theory*, 2008).

Changing in Focus or Emphasis over Time

Beyond the shifts in emphasis in reference to gender issues, environmental concerns, and issues of globalization, three of the other biggest alterations in critical theory have occurred in reference to Marxism, pragmatism, and methodological pluralism.

The End of Marxism?

While critical theorists are critical of the view that often is thought to be implicit in Marx – namely, that a shift in the means of production would bring an end to nearly all forms of psychic discomfort – they still have always traced many forms of social pathology to a basis in the existing economic form of existence. Marcuse spoke of the conditions of capitalism as causing one dimensionality, and restricting the type of pleasure possible in a postscarcity society. Horkheimer and Adorno spoke of how capitalism leads to a 'fully administered life.' Habermas has long spoken of the 'colonization of the lifeworld' leading to distorted communication.

While there is still a tendency to perceive the dominant forms of life in capitalism as impinging on as full of an expression of reason as is possible, few of the Frankfurt School-influenced critical theorists today speak of a full-throated socialist alternative to capitalism. More typically they write of the need for 'reflexive modernization' (Dryzek, 2000, 163–66), 'economic restructuring' (Fraser and Honneth, 2003, 75), or the 'taming' of capitalism (Habermas). This shift has accompanied the shift toward a more optimistic view – evident most

clearly in Habermas – that there have been advances in moral-critical reasoning accompanying the development in the forms of instrumental technical mastery. Consequently, the moral and political ideas in modern political and legal institutions are not viewed as merely entailing potential for domination but are seen as also embodying potentials for emancipation.

This general turn toward liberalism in Habermas-influenced critical theory can be contrasted with the recent work of various leftist theoreticians proposing a renewal of the 'communist idea' (see the selection of papers in *The Idea of Communism*). Such work runs the gamut from Zizek's farcical embrace of Stalinism through Susan Buck-Morss's look at 'historical pragmatics' to Vattimo's plea for a kinder, gentler, 'weak communism.' While some of this work goes much further in the support of violence than the Frankfurt School ever did, some of it clearly aligns with the basic reformist tendency that has characterized most thoughts of critical theory.

Pragmatist Critical Theory

While first-generation critical theory was critical of pragmatism – with Horkheimer viewing it as having no theory of society as a whole – there has been a turn to pragmatism in the course of the development of critical theory. Habermas began referring to his own views as 'Kantian pragmatist' in the 1990s, but he had long been influenced by pragmatism. *Knowledge and Human Interests* had already integrated the essentially pragmatic idea of various interests at the basis of knowledge generation. Further, the idea fundamental to Habermas's consensus theory of truth – that truth is what we are bound to agree to over the long run – is adopted from Pierce. In addition, Habermas integrated work from Mead on the social construction of individual identity in the give and take of dialogue in a social world (cp. Strydom, 2011, Chapter 3). Indeed, Habermas has come to see pragmatism as the one school of thought among the Hegelian legacy that shares his own spirit of liberalism. He notes, "Pragmatism constitutes, besides Marx and Kierkegaard, the third Young Heglian tradition, and the only one that convincingly develops the liberal spirit of radical democracy" (Habermas 2002, 228; see Strydom, 2011, 64).

Steven White and James Bohman emphasize a 'pragmatic ethos' at the root of contemporary critical theory that underlies the acceptance of plural perspectives, the social character of the self and of scientific inquiry, along with the connection between scientific inquiry and the community (cp. White, 2004; Bohman, 1999). Both thinkers suggest that a critical social theory informed by such a pragmatic ethos should be more 'problem-driven' than the current 'theory-driven' or 'method-driven' inquiry characteristic of critical theory, including Habermas's version. This move toward pragmatism does give rise to various methodological questions.

Methodological Issues

Methodological Pluralism

While Habermas is widely known for emphasizing that the possibility of an ultimate consensus is viewed as a presupposition to communicative action, he has always also recognized that the path to this consensus is paved with debate and disagreement. Still, his position has been that since we rationally are trying to explain and interpret the same social world, we should expect ultimate convergence in the best possible explanation. This is a point of contention both within critical theory and beyond. Many critical theorists emphasize that critical social inquiry is not in a position to replace existing research programs in the social sciences so much as to work in connection with them (cp. Honneth, 2007, 63; Bohman, 1999). Given the pluralism in the social sciences, however, this entails that we might well expect multiple viewpoints to prevail within the social sciences indeterminately. Longino formulates one position on this debate, noting her approach "does not preclude the possibility of some single all-encompassing theory being the final result of inquiry nor does it preclude the possibility of multiple accounts" (Longino, 2001, 140). Indeed, if philosophy is no longer able to stand as judge over the sciences, but must work as stand-in, and do that within a fallibilist framework, this seems a likely conclusion. While the work that a philosopher does as a stand-in should not render philosophy so impotent that it cannot, for example, criticize social scientists for working with methodologies that are themselves not falsifiable, it does not appear to leave philosophy so empowered that it can justify an ultimate judgment about whether a consensus will emerge in the social sciences, above all given the rather nonideal conditions in which theorizing does occur.

Explanation and Interpretation/Theoretical and Practical Reasons

The issue of methodological pluralism is related to questions of the separation between forms of understanding and forms of explanation in Habermas's formal description of his theory and in the work of other critical theorists. Habermas has come under considerable criticism for implying that there is such a neat division of labor between the various fields of inquiry, with certain interests tracking to specific domains (Detel, 2007). A myriad of questions emerge. For one, is not explanation used in the 'interpretative sciences' the domain evidently of understanding? Do not social scientists formulate hypothetical deductive theories as a matter of course so that the claim that they operate fully under the influence of a different method is clearly wrong from the outset? (cp. Mantzavinos, 2006). Similarly, are not interpretations a part of the background for theoretical explanations in the natural sciences? In short, it appears that in his theoretical writing Habermas illegitimately divvies up intellectual tasks in a too neat and unrealistic manner.

While a generous reading of Habermas is that he sees these spheres, not as completely isolated from one another, and indicating that rightness claims can have nothing to do with truth claims, or that strategic or instrumental reason can have nothing to do with communicative reason, but as indicating an 'idealizing supposition' generally applicable for certain kinds of action (Baynes, 2004, 198), this will not dissolve all the issues. Connected to this division of the intellectual tasks are questions about whether rightness and truth claims are best treated as creating a fundamental division between practical and theoretical reasons at all, or whether it is more suitable to

follow Dewey in embracing a more unified view of reason, which sees the tasks in these domains not as separate but as existing on continuum (cp. Berstein, 2010, 188ff.). Therefore, along with questions about the justification of the separation between explanation and understanding are questions about the separation of theoretical and practical reasons.

Further Questions of Habermas's Moral Theory

A more general criticism of Habermas's moral theory has concerned its formalism since the Kantian turn. In short, many have questioned how useful the idea of unconstrained dialogue is, or more specifically whether it is utopian. Since we will never have a condition of absolutely unconstrained dialogue, there are obvious difficulties knowing what normative claims might be reached under such an ideal speech situation. Does it get us further than Rawls' thought experiment of the veil of ignorance? Many of the criticisms along these lines have the character of Hegel's criticisms of Kant. While these criticisms come from some opponents to Habermas's project, they are also waged by those critical theorists who think that Habermas needs a still stronger pragmatist turn.

A further influential criticism of Habermas's ethics has been leveled by Ernst Tugendhat, a German ethicist who, like Habermas, develops a Kantian-influenced moral theory. Tugendhat (1994) famously questioned whether Habermas's ethics involves him in a form of circular reasoning. According to Habermas's argumentation, rational dialogue requires unconstrained agreement, a condition that for its part can only be guaranteed if the participants in moral dialogue have equal possibilities in social dialogue. This requires rather robust conditions of social equality. Consequently, the egalitarian form of society that Habermas desires to achieve is simply elevated to a presupposition for rational dialogue. As Tugendhat expresses this, the form of the argument is essentially if p and q, then q. If rational dialogue is only achievable with equality, then equality – a goal of Habermas's theory – is *de facto* not justified but presupposed. It is written into his definition.

While the force of the argument is not to be denied, it is clear that Habermas does not simply aim at generating equality from his moral argument. Rather, unconstrained dialogue allows decisions to be made about an entire plethora of moral questions in a fair and open manner. Further, the radicalism of the equality demanded by Habermas's theory is not entirely clear. Indeed, it is not clear that the theory entails a more fundamental form of equality than Rawls's theory, or even Tugendhat's.

Views of Reason and the Linguistic Turn

Further criticisms concern Habermas's linguistic turn. One issue is that it has neglected the experience of human subjects. Honneth is among those calling for a more phenomenological turn within critical theory, indicating the need to refer to first-person experience to make sense of movements of social justice that he thinks stand behind the growth in our moral-practical knowledge. Various proponents of the philosophy of the consciousness similarly criticize Habermas for the lack of a place for elements of the philosophy of consciousness, noting that even the work in rational reconstruction that Habermas engages in presupposes the privileged access that cognitive subjects have to their own experience, despite Habermas's rejection of such elements (cp. Freundlieb et al., 2004). Quite recently, Matthias Vogel (2012) has presented challenging arguments to all of the views of reason that focus solely on linguistic competencies for failing to provide the means needed to explain nonlinguistic human communication, such as that which occurs in music, art, and communication with babies. Related to such criticism is the question of whether Habermas's critical theory has not more generally ignored our embodiment. The characteristic ignorance of critical theorists of work in philosophy of biology and neurophilosophy is not unrelated.

Conclusion

The idea that critical theory does not provide a theoretical basis that can be definitively formulated, but that it should be reconceived to deal with emergent sociohistorical realities and developments in the sciences, was embraced even by the earliest generation of critical theorists. When publishing the English language volume *Critical Theory* in 1968 Horkheimer himself noted that many of the ideas "no longer have any direct application" and that "to relate them properly to the present situation requires careful reflection" (Horkheimer, 1982, v). Indeed, the original program itself was sensitive to the historical nature of thought, viewing the frameworks for interpreting reality and against which explanations occur as reflective of shifts in material conditions of reality. Those involved in critical theory today would thus be well served by accepting Horkheimer's warning of the dangers of a "thoughtless and dogmatic application of the critical theory to practice in changed historical circumstances" (Horkheimer, 1982, v).

Most critical theorists today are of the view that critical theory needs to be reconstructed. Greater attentiveness to our embodiment, to subjective experience, to particular emerging social issues, and to the developments in the natural sciences will mean that critical theory will develop not only in directions differing from first-generation critical theorists but also in directions differing from Habermas's influential form of critical social inquiry.

See also: Hermeneutics; Marx, Karl (1818–83); Phenomenology: Philosophical Aspects Postmodernism: Philosophical Aspects; Pragmatism: Philosophical Aspects; Scientific Explanation.

Bibliography

Adorno, Theodor, 2005. Minima Moralia: Reflections from Damaged Life. Jephcott, E. F. N. (trans.). Verso, New York.

Baynes, Kenneth, 2004. The transcendental turn: Habermas's Kantian pragmatism. In: Rush, Fred (Ed.), The Cambridge Companion to Critical Theory. Cambridge U P, Cambridge.

Benhabib, Seyla, 2006. Another Cosmopolitanism. Oxford U P, New York.

Berstein, Richard J, 2010. The Pragmatic Turn. Polity, Malden, MA.

Biro, Andrew (Ed.), 2011. Critical Ecologies. The Frankfurt School and Contemporary Environmental Crises. University of Toronto Press, Toronto.
Bohman, James, 1999. Theories, practices, and pluralism: a pragmatic interpretation of critical social science. Philosophy of the Social Sciences 29, 459–480.
Bohman, James, 2010. Democracy Across Borders: From Dêmos to Dêmoi. MIT Press, Cambridge, MA.
Bronner, Stephen Eric, 2011. A Very Short Introduction to Critical Theory. Oxford U P, New York.
Detel, Wolfgang, 2007. Habermas und die methodologie kiritischer theorien. In: Winter, Rainer, Zima, Peter V (Eds.), Kritische Theorie Heute. Transcript, Bielefeld.
Dryzek, John, 2000. Deliberative Democracy and Beyond. Liberals, Critics, Contestations. Oxford U P, New York.
Dryzek, John, 2005a. The Politics of the Earth. Environmental Discourses. Oxford U P, New York.
Dryzek, John, 2005b. Political and ecological communication. In: Dryzek, John (Ed.), Debating the Earth. Oxford U P, New York.
Fraser, Nancy, 2008. Scales of Justice. Columbia U P, New York.
Fraser, Nancy, Honneth, Axel, 2003. Redistribution or Recognition? A Political-Philosophical Exchange. Verso, New York.
Freundlieb, Dieter, Hudson, Wayne, Rundell, John, 2004. Reasoning, language and intersubjectivity. In: Freundlieb, Dieter, et al. (Eds.), Critical Theory After Habermas. Brill, Boston.
Görg, Christoph, 2011. Societal relationships with nature: a dialectical approach to environmental politics. In: Biro, Andrew (Ed.), Critical Ecologies. The Frankfurt School and Contemporary Environmental Crises. University of Toronto Press, Toronto.
Habermas, Jürgen, 1984. McCarthy, Thomas (trans.). Theory of Communicative Action, vol. 1. Beacon Press, Boston.
Habermas, Jürgen, 1988. Between Facts and Norms. Rehg, William (trans.). MIT Press, Cambridge, MA.
Habermas, Jürgen, 1990. Moral Consciousness and Communicative Action. Lenhardt, Christian, Weber Nicholsen, Shierry (trans.). MIT Press, Cambridge, MA.
Habermas, Jürgen, 1992. Postmetaphysical Thinking. Hohengarten, William Mark (trans.). Polity, Cambridge.
Habermas, Jürgen, 1998. The inclusion of the other. In: Ciaran, Cronin, Pablo, De Greiff (Eds.), Studies in Political Theory. MIT Press, Cambridge, MA.
Harding, Sandra, 1993. Rethinking standpoint epistemology. In: Alcoff, Linda, Potter, Elizabeth (Eds.), Feminist Epistemologies. Routledge, New York, NY.
Honneth, Axel, 2007. Disrespect. Polity, Cambridge.
Honneth, Axel, 2009. Pathologies of Reason. On the Legacy of Critical Theory. Ingram, James (trans.). Columbia U P, New York.
Horkheimer, Max, 1982. Critical Theory. O'Connell, Matthew J and others (trans.). Herder and Herder, New York.
Horkheimer, Max, Adorno, T.W., 1972. Dialectic of Enlightenment. Seabury, New York.
Longino, Helen, 1996. Subjects, power, and knowledge: description and prescription in feminist philosophies of science. In: Keller, Evelyn Fox, Longino, Helen E (Eds.), Feminism and Science. Oxford U P, Oxford.
Longino, Helen, 2001. The Fate of Knowledge. Princeton University Press, Princeton.
Mantzavinos, C., 2006. Naturalistic Hermeneutics, Arnold, Darrell P in collaboration with the author (trans.). Cambridge U P, Cambridge.
Müller-Doohm, Stefan, 2004. The critical theory of society as reflexive sociology. In: Huhn, Tom (Ed.), The Cambridge Companion to Adorno. Cambridge U P, Cambridge.
Strydom, Piet, 2011. Contemporary Critical Theory and Methodology. Routledge, New York.
Tugendhat, Ernst, 1994. Vorlesungen über Ethik. Suhrkamp, Frankfurt.
Vogel, Matthias, 2012. Media of Reason, Arnold, Darrell P. (trans.). Columbia University Press, New York.
White, Stephen K, 2004. The very idea of a critical social science. A pragmatist turn. In: Rush, Fred (Ed.), Cambridge Companion to Critical Theory. Cambridge University Press, Cambridge.
Wiggershaus, Rolf, 2007. The Frankfurt School. Robertson, Michael (trans.). MIT Press, Cambridge, MA.

Relevant Websites

http://pages.gseis.ucla.edu/faculty/kellner/Illumina%20Folder/ – Illuminations. The Critical Theory Project.
http://www.marxists.org – Search for Original Source Texts at Marxist.org.
http://www.ifs.uni-frankfurt.de/archiv/index.htm – Theodor Adorno Archive.
http://www.habermasforum.dk – The Habermas Forum.
http://www.walterbenjamin.org – The International Walter Benjamin Society.
http://carbon.ucdenver.edu/~mryder/itc/postmodern.html – Website on Contemporary Philosophy, Critical Theory, Postmodern Thought, Maintained by Mitch Ryder.
http://plato.stanford.edu/entries/critical-theory/ – The Entry on Critical Theory at the Stanford Encyclopedia of Philosophy.
http://www.marcusesociety.org – International Marcuse society.

Critical Theory: Contemporary

Larry Ray, University of Kent, Canterbury, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article is a discussion of the formation of contemporary critical theory with particular reference to the work of Habermas and Honneth and the debates that have arisen from these. They have both extensively recast the earlier critical theory of Adorno and Horkheimer through Habermas's communicative turn and then Honneth's theory of recognition. However, core themes of critique of domination and the conditions for an emancipated form of life continue to inform contemporary debates.

Critical theory is a term adopted to situate social philosophy and theory in the heterodox Marxist intellectual tradition of the Frankfurt School (especially Adorno and Horkheimer). Contemporary theories might develop different analyses from earlier versions, although a critical theory will aim to explore the conditions of both domination and critique of obstacles to emancipation and a 'good society.' Critical theory was originally associated with the Frankfurt Institute in the 1930s, the key figures of which included Max Horkheimer, Theodor Adorno, Walter Benjamin, and Herbert Marcuse. Critical theory embodied the Hegelian and psychoanalytic reorientation of Western Marxism during the 1930s in which theory was transformative, rather than contemplative, engaging with historical processes to enhance human emancipation (Horkheimer, 1972). Unlike more orthodox Marxism, it accorded central significance to culture (including mass media), authoritarianism, science and technology, and gender. Central to their thinking was the apparent eclipse of the revolutionary proletariat in the wake of Stalinism, Nazism, and mass consumption capitalism. Many of these themes, expressed for example in Horkheimer and Adorno (1973), were to become central to social theory some 25 years later. Many of its original themes (such as the demise of the historical subject, critique of the culture industry, attention to psychosocial processes, critique of technology, science, and domination) have entered mainstream social theory and it might be less clear now what its particular contribution is. Nonetheless, these themes remain important foci of critical theoretical debate about the possibilities for a better society.

Of central concern for earlier critical theorists was the materiality of capitalism and its distorting effects on critical consciousness, rational reflection, and the capacity for organized political opposition. When theorists turned to areas previously neglected by Marxism, such as Adorno's extensive critiques of the culture industry, mass media, and positivist epistemology, the critique of commodification retained central importance. This was moreover a critique informed by an aesthetic and affective sensibility, which Adorno hoped would "lend suffering a voice" (Adorno, 1974, p. 247). Further, while the aesthetic experience of critical art offered the 'tremor of the possibility' of a better society (Adorno, 2003, p. 14), popular culture was commodified and produced according to the logic of mechanization. By contrast, later the critical theory of Jürgen Habermas refocused attention on issues of the public sphere and emancipation, linguistics, contemporary sociological theory, and the limits of radical critique. Habermas's work has been controversial. Some of the most interesting critiques of Habermas have come from feminist critical theorists, who share his commitment to the explication of power, knowledge, and morality, but differ with him on other issues. This has, in turn, raised questions about the limitations of Habermasian universalism as opposed to more embodied and concrete concepts of the self and discourse. Subsequently, Axel Honneth (b. 1949) who is currently Professor of Philosophy at the University of Frankfurt and Director of the Institute for Social Research has shifted the metatheoretical foundations of critical theory from communication to recognition and moral experience via a return to the Hegelian concept. He treats recognition as a need that is anthropologically grounded in the constitution of society. Here, 'recognition' is used in the Hegelian sense as a concept to describe the birth of consciousness through mutual recognition between two subjects. More recently, Honneth has provided an extensive discussion of recognition aiming to show the motivational and normative role recognition can play in stimulating social movements. Following Hegel and G. H. Mead, Honneth identifies three 'patterns of recognition' necessary for a positive relation to self. These are love, rights, and solidarity, which are discussed later.

Contemporary Critical Theory

Jürgen Habermas (b. 1929) claims that it was not his intention to continue the work of the Frankfurt School, which he says was a response to historically specific experiences of fascism, Stalinism, and the Holocaust (Anderson, 2011, p. 31). Nonetheless, he has been the dominant figure in contemporary critical theory and is widely regarded, along with some contemporaries, such as Albrech Wellmer, Karl-Otto Apel, and Oskar Negt as representing a 'second generation' of their thinking. Central to his work has been the aim to specify the conditions under which human life would be free from domination. He has, however, criticized many of its earlier claims. He regards Horkheimer and Adorno's apocalyptic despair as having ignored the

'unfulfilled potential' of Western modernity to expand rationality and democracy (Habermas, 1987, p. 113). They too readily accepted a Weberian nightmare of an iron cage of total domination, in which philosophy could only keep alive the possibility of a better society while lamenting the lost opportunity for emancipation (Benhabib, 1981). Further, Habermas has effected a comprehensive project critically synthesizing contemporary sociology and philosophy that not only abandons Adorno's aversion to theoretical wholes but also positively incorporates theorists whose nondialectical and 'affirmative' systems thinking Adorno rejected. This notably includes Talcott Parsons who is central to Habermas's development of a theory of communicative action in the 1980s (Habermas, 1987). Nonetheless, for Habermas, the very possibility of critique must presuppose the existence of rational criteria of a better society against which the present is being judged (Habermas, 1987, p. 113). He aims to recover and explicate this potential, to which end he has undertaken an elaborate reconstruction of the grounds of everyday communication.

Communicative Turn

Habermas attempts to invigorate critical theory by embracing the real achievements of bourgeois democracy while salvaging the project of modernity specifically through a mobilized public sphere, revitalized public discourse, and personal involvement in politics. Habermas's early work focused on the constriction of the public sphere (*Öffentlichkeit*) and democratic deliberation by technocratic politics and commercial mass media. (The public sphere is seen as a discursive arena that facilitates citizen debates, deliberation, agreement, and action where individuals are able to freely share their views with one another and see the effects of their deliberations in a process that approximates a participatory democracy.) Although less apparent in his later work, the encroachment of democratic will formation was encapsulated in his highly influential distinction between two social orders, of lifeworld and system. The former is the sphere of communication and culture, mediated by norms and language. The latter refers, using Parson's concept, to forms of social 'steering' by money and power, which are not available to normative discourse. While Habermas insists that the market and state are essential in modern societies, the notion of steering media appear no longer as benign processes reducing social complexity but if unleashed threaten to undermine communicatively secured solidarities. There is then a continual tension along the boundaries of lifeworld and system where new social movements arise, potentially pushing back systemic intrusion and opening spaces for democratic deliberation. An example would be environmental protests over nuclear reprocessing that open debates about the rationality and morality of projects previously driven by technological and financial criteria. This approach was closely linked in the 1980s to 'new social movement' theories, the underlying assumption of which was that new conflicts no longer arise in areas of material reproduction, are not now channeled through parties and organizations, and can no longer be alleviated by compensations that conform to the system. Rather, "the new conflicts arise in areas of cultural reproduction, social integration, and socialization" (Habermas, 1981). However, while this approach was highly influential and cognate to the emerging identity politics in the same period, the idea of postmaterial conflicts looks less relevant in a period of economic crisis and rapidly growing national and global inequalities, a view which is reflected in more recent debates. Thus, the critique of the disembedding and unleashing of markets and political power in the early twenty-first century has been made more with reference to global neoliberalism and, for example, how the weak publics of transnational or global civil society have failed to generate levels of postnational solidarity that would even begin to approximate those of nation-states. This disembedding of economic and political power reflects how social solidarities have not kept pace with the globalization of power, law, and money (Brunkhorst, 2007).

In a further departure from the earlier critical theory, Habermas identified potential emancipatory practice not by appeal to putative agents of change (such as the proletariat) but through 'universal pragmatics' – a move that was signaled by his critical engagement with Peirce in Habermas (1973). Later, Habermas sees in the pragmatism of Pierce and Dewey a postmetaphysical philosophy that can address the very foundations of the critique of domination (Ray, 2004). This is because for Habermas the potential for rational agreement is present in sociolinguistic rules of communication. He uses Chomsky's linguistics and Austin and Searle's theory of speech acts to reconstruct communicative rules, arguing that the elementary units of communication (speech acts) involve validity claims that are 'naively accepted' in conversations. For conversations to occur, one assumes agreement about grammar and illocution, that is, one recognizes something as a promise, an assertion, an order, etc. Similarly, judgments about the comprehensibility of actions involve evaluations as to whether good reasons have been provided. Although moral choices are often regarded as matters of personal conviction, Habermas agues that the requirement to provide good reasons, subject to public scrutiny, renders these choices 'capable of truth.' The evaluation of moral choices is comparable to the evaluation of empirical claims in that both the proffered reasons are subject to critical testing (Habermas, 1992, p. 76). Thus, Habermas seeks a practical and cognitive basis for critical theorizing.

These are not timeless linguistic structures, but unfold historically through the formation of posttraditional (modern) worldviews. Unlike Marxist theories, which view social development as the succession of modes of production, Habermas's theory refers to stages of cognitive and moral learning, which parallel childhood development. Premodern forms of moral authority involve conventional attachments to externally given value systems, underwritten by sacred beliefs. Echoing Kant, 'maturity' involves postconventional distance from rules and the capacity to evaluate and argue whether adherence to a moral injunction is appropriate. Successful posttraditional sociality institutionalizes new levels of cognitive and moral learning, thus normative structures are the 'pacemaker of social evolution' (Habermas, 1979, p. 120). Modern worldviews dissolve sacred authority into rational justifications of morality.

This process of the 'linguistification of the sacred' is potentially liberating because it releases the 'argumentative force of language' grounded in the capacity of assent or dissent

from any proposition, whether empirical, moral, or aesthetic. ('Linguistification of the sacred' refers to a process in post-traditional societies whereby actions and norms cease to be guided by fear of the sacred but rather are open to rational validity claims. Actions are less governed by unquestionable authorities and are instead open to rational, communicative action, which involves a progressive expansion in critical reflection on norms and practices in everyday life.) In each of these spheres, any attempt to reach an understanding of the truth of a proposition, the sincerity of the speaker, or the appropriateness of the utterance can be disputed. Habermas claims that entering into dialogue about the validity of utterances involves a specific type of discussion (*Diskurs*) oriented toward examining these conditions (Habermas, 1984, p. 28). But here is the nub of the argument. Such practices will be successful only if one assumes equality of access to speech acts and freedom to move from level to level in *Diskurs*, since it is only under such conditions that consensus could be motivated by the force of better argument. This enables Habermas to introduce the 'general symmetry requirement' (also called the 'ideal speech situation'), which states that if a consensus is to be reached guided only by the force of better argument, then whether one recognizes it or not, one routinely assumes certain conditions. That is,

- everyone has an equal chance to deploy, initiate, and perpetuate speech acts,
- utterances are comprehensible,
- their propositional content is true,
- what is said is legitimate and appropriate, and
- it is sincerely spoken.

Habermas's concept of 'ideal speech situation' (also referred to in his work as 'discourse ethics' and 'communicative competence') is used to depict a counterfactual situation in which all participants to communication must have equal opportunity to contribute. They must have the right to assert, defend, or question any factual or normative claim and be motivated solely by the desire to reach a consensus about the truth of statements and the validity of norms. For Habermas, these norms are 'always already' implied in linguistic communication and provide a basis for critique of domination that distorts actual communications in everyday life. Habermas realized, of course, that these conditions of communicative competence are counterfactual and rarely present in actual speech, but it is precisely this that justifies two further claims. The first is that the goal of normal communication is disturbed by power relations, which intrude to prevent questions being raised or validity claims being tested. Second, these constraints on communication are self-defeating since they contradict the underlying assumptions that make dialogue possible at all. From this, it follows that rationality can be measured by the degree of openness or closure in communication. The goals of truth, freedom, and justice are not mere utopian dreams, but are anticipated in ordinary communication, and therefore the goal of emancipation is presupposed in the constitution of the species as linguistic beings (Habermas, 1991, p. 244). If the unspoken authority of all communications can potentially be challenged, then perhaps democratic social relations are already implicit in everyday communications. For this insight to have practical significance, however, it would need to be linked to the formation of emancipatory social movements: feminism, ecology, civil rights, and democratization move ments (Ray, 1993; pp. 57–78).

This analysis and the emancipatory promise it contained, invested in extended democratization of the grammar of everyday, became distinctive of a new turn in critical theory in the later twentieth century. However, the concept of the public sphere in Habermas was grounded in face-to-face interactions and a possibly rather idealized view of the eighteenth century bourgeois 'public.' Prior to the digital revolution and social media, which introduced further dramatic shifts in modes of communication, critics already suggested that Habermas had not given sufficient attention to mass media's obstruction of democratic debate (e.g., Kellner, 1990). Moreover, as Kellner (n.d.) also argues, "the development of new global public spheres with the Internet and new multimedia technology require further development of the concept of the public sphere today and reflection on the emerging importance of new technologies within democracy." But, while Kellner (n.d.) suggests that "The rise of the Internet expands the realm for democratic participation and debate and creates new public spaces for political intervention," others are more pessimistic. For example, Jodi Dean's (2009) concept of 'communicative capitalism' points to a process that subdues deliberative democracy with mass participation in infotainment (Dean, 2009, p. 2). The paradox here is that, far from being de-democratized, the contemporary ideological formation of communicative capitalism "fetishizes speech, opinion, arid participation." It "embeds us in a mindset wherein the number of friends one has on Facebook or MySpace, the number of page-hits one gets on one's blog, and the number of videos featured on one's You Tube channel are the key markers of success." But at the same time meaningful debate, "details such as duration, depth of commitment, corporate and financial influence, access to structures of decision-making, and the narrowing of political struggle to the standards of do-it-yourself entertainment culture become [seen as] the boring preoccupations of baby-boomers stuck in the past" (Dean, 2009, p. 17). In an echo of the pessimism of earlier critical theory she calls this "talk without response" (2009, p. 24).

How Critical Is Critical Theory?

Contemporary critical theory has developed a wide range of debates and projects in the course of which Habermas's has been criticized. For some (Hullot-Kentor, 1989), he has abandoned the earlier critical thrust of Horkheimer and Adorno. For others, he privileges communicative functions of language over others, such as irony or aesthetic expression (Thompson, 1982) and has misunderstood or misused speech act theory (Rasmussen, 1990, p. 37). One of the most frequent criticisms is that if Habermas envisages domination-free dialogue as an attainable state, this is an implausible utopian dream (Jay, 1988, p. 31). If it is not at least potentially an attainable state, then the force of his critique is lost. Similarly, Habermas's emphasis on communication and language is criticized as being idealist and ignoring the importance of material structures (Dux, 1991). Brian Fay suggests that emancipation through rational reflection

encounters limitations imposed by 'embodiment,' since authority is inscribed into conditioned patterns of activity, which cannot readily be subject to rational refection (Fay, 1987, p. 148), thereby signaling the corporal turn in later twentieth and early twenty-first century sociology, with which Habermas has not engaged. Two areas of controversy in particular illustrate the relevance of critical theory to contemporary debates.

Critical Theory and Feminism

Some of the most interesting critiques of Habermas have come from feminist critical theorists, who share his commitment to the explication of power, knowledge, and morality, but differ with him on other issues. A central issue for some is that Habermasian theory stands in the Enlightenment tradition, which many feminists have rejected (Meehan, 1995), while for others he has turned away from the emphasis placed on gender oppression in classical works of critical theory such as the *Dialectic of Enlightenment* (Fischer and Tepe, 2011). Universal pragmatics deals with highly abstract and depersonalized speech rules, which is open to the objection that the materiality of class, gender, and other social divisions remain at the core of society and therefore should be at the center of a critical theory. By contrast, feminist critics have identified discursive forms that represent the experience of embodied female subjects. For example, Nancy Fraser (1985) argues that Habermas fails to examine the gendering of his core concepts such as citizenship, universalism, public sphere, and civil society. The role of citizen is premised on male occupation of the public sphere in which a civil contract among brothers is contrasted with the feminization of the private sphere.

Seyla Benhabib criticizes what she sees as the disembodied transcendental rational ego that is the subject of ideal speech. Drawing on communitarian critics of the 'unencumbered self' (Taylor, MacIntyre), she sees in Habermas's attempt to reconstruct universal morality an autonomous male ego, which universalism privileges as a point of reference. Since subjects are finite and social, contexts of gender and community are central and people are not only generalized others (moral persons with reason) but also concrete others. The negotiation between the 'generalized other' and collective concrete others takes place in a male public sphere from which women disappear (Benhabib, 1992, p. 13). Nonetheless, Benhabib applauds Habermas's shift to communication, which departs from reason as a psychological attribute of consciousness to an intersubjective view of ethics and dialogue.

A similar point is made by Kate Soper, who notes the "all male monastical atmospherics of the Frankfurt School," but sees nonadversarial dialogue as similar to consciousness raising, and his distinction between normatively secured and communicatively achieved dialogue as a means of opening up unquestioned conventions (e.g., family) to explicit discussion. Indeed, there is some anxiety among these critics that the theory risks losing sight of universalism in favor of Romantic particularism. Benhabib asks, "would a moral theory restricted to the standpoint of the concrete other be sexist, racist, culturally relative, and discriminatory?" "No," she replies, "rights must be an essential component of any such theory." The problem here (cf. Soper, 1989) is that 'rights' must involve abstraction from particularity otherwise the grounding for the demand for equal treatment is removed.

In similar vein, Sorial (2011) argues that Habermas's proceduralist model of law, developed in Habermas (1998) can be put to feminist ends in at least two significant ways. First, by presenting an alternative to the liberal and welfare models of laws, the proceduralist model offers feminism a way out of the equality/difference dilemma. This dilemma, which Habermas addresses, arises from the tension between a juridical-liberal model of equality based on equal rights on the one hand and the welfare state model on the other "that developed protective norms in relation to pregnancy and maternity, custody rights, divorce, domestic violence and sexual assault." Both these, however, "merely reinforced existing stereotypes of gender identity" and require women to adapt to patriarchal institutions that they do not in themselves challenge. The "proceduralist model renders this dilemma irrelevant [because] … it proceeds from the fact of sexual difference" and requires that all parties themselves engage in public discussions in which they "articulate their needs, interpret their rights, and take responsibility for the law making process about issues that affect them." Second, Sorial defends Habermas against feminist critics such as Fraser and Dean, claiming that his deliberative model is not inherently masculine or exclusionary. On the contrary, the "requirement that all those affected by a particular norm participate in the debate about that norm not only offers a way of overcoming the equality/difference dilemma … but also necessitates an important institutional shift" in legal institutions themselves by opening up multiple avenues of debate and reconnecting notions of private and public autonomy. Even so, it is not clear to what extent Habermas or Sorial are describing *actual* legal systems or an ideal(ized?) model of a system which, like Hegel's *Rechtsstaat*, did not actually exist.

The Debate over Modernity

In pursuit of his project of salvaging the project of modernity, Habermas confronted what he regarded as another reversion to irrationalism and pessimism in the postmodern theories of the 1980s and 1990s. He appeared to be one of the few social theorists at the time who was willing to defend the tradition of modernity, and was frequently called to do so in debates with theorists like Lyotard, Gadamer, and Foucault. In some ways, the two intellectual movements, critical theory and postmodernism, did share an attention to knowledge and communication media, the fragmentation of culture and identity, and the end of historical narratives. (To be clear, postmodernism was a diffuse intellectual movement in the social sciences, arts, architecture, and literature that rejected universalist Enlightenment 'narratives,' such as reason and progress, in favor of celebration of diversity, fragmentation, and aesthetic irony.) Nonetheless, for critical theory, postmodernism denies the grounds of its own critique and permits celebration of spontaneous power of imagination. For Habermas, that modernity is not exhausted but unfulfilled was

demonstrated by the anticommunist 'rectifying revolutions' (*die nachholende Revolution*) of 1989. These posed anew the core issues of modernity – rights, civil society, market governance, cultural autonomy, and democratic association (Habermas, 1990, 1994, p. 62). Even so, Habermas's (2001) hope that the 1989 revolutions would enable the emergence of a "European civil society; the construction of a European-wide public sphere; and the shaping of a political culture that can be shared by all European citizens resurgence of civil society" in both the east and the west proved too optimistic. In retrospect, many of the utopian hopes that surfaced briefly were disappointed (Ray, 2009; Vogt, 2005), which leaves open the question of what are the conditions for emancipatory critique and practice?

The details of debate over postmodernism subsequently became a little passé since postmodernism as Matthewman and Hoey (2006) suggest "never made it into the 21st century." Nonetheless, the sociological issues that stimulated this debate remain relevant and the question of the possible potentials for emancipatory practice central to the project of critical theory. Moreover, Habermas's robust defense of modernity and enlightenment is as much a challenge to earlier critical theory as to postmodernism. An evaluation of contemporary critical theory should assess its success in identifying emancipatory impulses neglected by earlier theorists. Habermas gestures toward social movements as representing new communicative structures, although he says very little about these in detail. The key issue is whether the theory of communicative action is a plausible reconstruction of the emancipatory potentials inherent in everyday speech. Even if one finds these arguments persuasive, the question remains whether these are cultural conventions or deeper structures constitutive of discourse.

Communication to Recognition?

Habermas has been the dominant theorist of contemporary critical theory for several decades. However, Anderson (2011, p. 46) points out that there is a 'third generation' of critical theorists, among whom Axel Honneth "is the undisputed gravitational center," whose theoretical concerns were shaped by formative experiences of the counterculture of the late 1960s, new social movements, fall of Berlin Wall, and the resurgence of nationalism and globalization. These brought to the fore postmaterial issues of identity, lifestyle, and belonging. Like earlier critical theorists, the new generation draws upon modern Western philosophy, seeks engagement with social science research, and aims to develop a normative critical theory. Honneth's social philosophy places mutual recognition at its core, while retaining a metatheoretical approach to specifying the conditions for social critique. Honneth claims that Habermas is concerned with the social and cognitive restrictions that place limits on unimpeded application of linguistic rules but these do not appear in the moral experiences of subjects (2007, p. 70). He attempts to rectify this by shifting the focus of critical theory from communication to experiences of recognition. This is the key process of social integration while conflict is motivated by *moral* feelings of lack of recognition. Further, Honneth returns to the conceptual lexicon and mood of earlier critical theory, for example, in his reformulation of the concept of reification (Honneth, 2008). In this way, he aims to "close the gap left open by Habermas in his further development of Horkheimer's programme" (2007, p. 72). So, "In place of Habermas's focus on undistorted relations of communication as revealing a standard of justification, Honneth focuses on the progressive overcoming of barriers to full interpersonal recognition, barriers such as legal exclusion and cultural denigration" (Anderson, 2011, p. 54).

Honneth follows Charles Taylor in arguing that recognition is essential to self-realization, although he returns more directly to the Hegelian concept while attempting to explain the normative role of recognition in the formation of social movements. Critical for motivating resistance is the experience of disrespect and denial of recognition. Like Hegel and George Herbert Mead, Honneth identifies three levels of mutuality, namely love, rights, and solidarity (e.g., Honneth, 2007, pp. 129–142).

- Love refers to mutual affirmation and recognition of physical needs, which, especially in the parent–child relationship, is essential to one's ability to achieve both dependence and autonomy. It is located in processes of socialization, family forms, and friendship and it is through the child's emotional attachment to another person that a world of meaningful qualities is disclosed. This is destroyed by physical violence and abuse that will cripple one's trust and ability to form relationships in the world.
- Rights refers to freedom and the absence of arbitrary domination combined with access to the political process and moreover the development of moral responsibility in one's relationships with others. Its context is the institution and application of laws. The denial of rights through legal and social exclusion threatens a person's moral standing and autonomy.
- Solidarity refers to the affirmation and self-esteem that arises from finding one's particularity affirmed within communities. This is essential for developing self-esteem and for how one becomes 'individualized.' For an example of the consequence of refusal of recognition, Honneth cites Orlando Patterson's (1982) concept of 'social death' among Afro-American slaves. The slave was so dependent on the master's authority that he/she became a nonperson in the face of systematic denigration and insult.

For Honneth, negative emotional states of shame, anger, and frustration might make us conscious of an injustice, although these do not automatically disclose relations of disrespect (2007, p. 72). However, they are a "pre-theoretical fact" on the "basis of which a critique of relations of recognition can identify its own theoretical perspective in social reality" (2007, p. 72). So he seems to be claiming that first, these emotional states indicate the potential for naming an injustice based on disrespect and from this the formation of social struggles, and second, that a critical theory is founded on these experiences as well. These expectations of mutual recognition are 'quasi-transcendental interests' of the human race – a claim that locates the critique of their violation in a concept of species being. Like Habermas, Honneth argues that these struggles are not exclusively conflicts over interests, since as

Anderson (1996) points out, the 'grammar' of such struggles is 'moral' in the sense that the feelings of outrage and indignation generated by the rejection of claims to recognition imply normative judgments about the legitimacy of social arrangements. In this way, the normative ideal of a just society is empirically confirmed by historical struggles for recognition.

These views, like Habermas's universal pragmatics, have generated extensive debate and stimulated further theorization of critique and resistance. There has been debate about Honneth's neglect of distributive injustices (e.g., Fraser and Honneth, 2003; Zurn, 2003, 2005). Rogers (2009) argues that focusing on which category – distribution or recognition – is more fundamental misses the complexities of historical political action that presumes they are linked, although this does seem to be the gist of the exchange in Fraser and Honneth (2003). Further, Honneth does regard the subjective experience of injustice (lack of respect) to be a consequence of material inequalities (2007, p. 81ff) but regards the experience of disrespect to be the condition for resistance. However, it has further been argued that shame and humiliation are not necessarily routes to movements for social justice and indeed Thomas Scheff's extensive research on shame and violence puts this claim in considerable doubt (e.g., Scheff, 2000). Shame and humiliation as Sayer (2005) suggests are in some ways the opposite of self-respect and pride and as likely to generate self-contempt as active resistance.

These observations are part of a wider critique that identifies difficulties in his attempt to "fix recognition as a kind of anthropological constant underneath social interaction" (Foster, 2011). O'Mahony (2011) argues that while Honneth's theory provides a rich starting point for sociological analysis it does also have certain weaknesses. First, his action-theoretical orientation does not address institutional and systemic levels of sociological analysis (see also Strydom, 2011). Second, his spheres of recognition are an "uneasy mix of anthropological constants and socio-historically varying standards that inhibit their value for concrete analysis." If they are anthropological, he says, they are 'always already' present in social life but it is then not clear how they vary historically. Third, they are presented as typologically distinct rather than institutionally interdependent in shaping practices across different social spheres. Foster (2011) suggests that Honneth underplays the transformative power of struggle for recognition on the self in the process of working through misrecognition. He further suggests that the Honneth's sociological account of reification is a narrowing of critical theory that could be broadened by an understanding of overcoming misrecognition as a struggle for 'self-awareness of damaged selves.' Recognition is a struggle in which reification and subject formation are entwined and in which recovery of empathetic attunement (the ability to see the other) must be part of a process of self-transformation in which one acknowledges one's own disfigurement. This implies a further return to older psychoanalytical Frankfurt themes entwined with a critique of aesthetic and meaningful senses in capitalism. Finally, here, Rogers (2009) sees in Honneth a paradox of recognition in that psychological integrity and autonomy depend on recognition by the larger political community. This means that subjectivity has to be constructed through the vocabulary of those in power who deny the needs and integrity of the other. He challenges Honneth's account of 'social death' of African-American slaves that Rogers suggests would leave slaves unaware of what to struggle for. This obscures the active process through which oppressed groups reimagine their self-understanding and exhibit resistance, which invokes parallel discursive arenas for developing counterdiscourses. This wide-ranging debate, however, indicates how Honneth's work will become the focus of debates in critical theory in the coming years.

Conclusion

The debates around both Habermas and Honneth pose the question of how to extend the critical theory beyond political philosophy and connect with critique of social conditions and practical engagement. Early critical theory addressed a wide range of social, political, and intellectual issues but central within these was the Marxist critique of the dominance of the commodity form and the problem of consciousness and historical agency. In later critical theory, these problems have been fundamentally recast in that the materiality of the commodity and critique of consciousness has given way to the critique of distorted communication, the program of democratization, and the return to recognition as the basis of sociality. While the critical theory of Habermas, Honneth, and other contemporary representatives has increased in analytical clarity and comprehensive engagement with philosophy and social theory, much of the passion and outrage with which works such as *Dialectic of Enlightenment* were infused has ebbed. Critical theory is rigorous and critical within the confines of scholarly academic work, while attempting to set out the metatheoretical conditions for critique of injustice. However, one will see over the coming years whether the urgency of engaging with early twenty-first century crises of economy, ecology, social integration, democracy, public spaces, and institutions provokes new modalities of theory and praxis. When, as Dean says, one has "talk without response" there is an urgent need to locate the agents of critical and radical social change.

See also: Cultural Critique, Anthropological; Field Theory; Modernity; Postmodernism: Philosophical Aspects; Rationality in Society; Sociological Theory.

Bibliography

Adorno, T.W., 1974. Minima Moralia. Verso, London.

Adorno, T.W., 2003. In: Tiedemann, Rolf (Ed.), Can One Live after Auschwitz? A Philosophical Reader. Stanford University Press, Stanford.

Anderson, J., 1996. 'Translator's Introduction' to Honneth, The Struggle for Recognition, pp. x–xxi.

Anderson, J., 2011. Situating Axel Honneth in the Frankfurt School tradition. In: Petherbridge, D. (Ed.), Axel Honneth: Critical Essays. Brill, Boston, pp. 31–58.

Benhabib, S., 1981. Modernity and the aporias of critical theory. Telos 49, 38–61.

Benhabib, S., 1992. Autonomy, Feminism and Postmodernism. Blackwell, Oxford, UK.

Brunkhorst, H., 2007. Globalizing solidarity: the destiny of democratic solidarity in the times of global capitalism, global religion, and the global public. Journal of Social Philosophy 38 (1), 93–111.

Dean, J., 2009. Democracy and Other Neoliberal Fantasies – Communicative Capitalism and Left Politics. Duke University Press.
Dux, G., 1991. Communicative reason and interest. In: Honneth, A., Joas, H. (Eds.), Communicative Action. Polity, Oxford, pp. 74–96.
Fay, B., 1987. Critical Social Science—Liberation and Its Limits. Cornell University Press, Ithaca, NY.
Fischer, A., Tepe, D., 2011. What's "critical" about critical theory: capturing the social totality (*das Gesellschaftliche Ganze*). Journal of International Relations and Development 14, 366–375.
Foster, R., 2011. An Adornian theory of recognition? A critical response to Axel Honneth's reification: a new look at an old idea. International Journal of Philosophical Studies 19 (2), 255–265.
Fraser, N., 1985. What's critical about critical theory? Habermas and gender. New German Critique 35, 97–131.
Fraser, N., Honneth, A., 2003. Redistribution or Recognition? A Political-Philosophical Exchange. Verso, London.
Habermas, J., 1973. Knowledge and Human Interest. Heinemann, London.
Habermas, J., 1979. Communication and Evolution of Society. Heinemann, London.
Habermas, J., 1981. New social movements. Telos 49, 33–37.
Habermas, J., 1984. The Theory of Communicative Action: Reason and the Rationalization of Society, vol. 1. Heinemann, London.
Habermas, J., 1987. Philosophical Discourse of Modernity. MIT Press, Cambridge, MA.
Habermas, J., 1990. What does socialism mean today? the rectifying revolution and the need for new thinking on the left. New Left Review 183, 3–21.
Habermas, J., 1991. A reply. In: Honneth, A., Joas, H. (Eds.), Communicative Action. Polity, Oxford, pp. 214–264.
Habermas, J., 1992. Moral Consciousness and Communicative Action. Polity, Cambridge, UK.
Habermas, J., 1994. The Past as Future. Polity, Cambridge, UK.
Habermas, J., 1998. Between Facts and Norms. MIT Press, Cambridge, MA.
Habermas, J., 2001. Why Europe needs a constitution. New Left Review 11, 5–26.
Honneth, A., 2007. Disrespect: The Normative Foundations of Critical Theory. Polity, Cambridge, UK.
Honneth, A., 2008. Reification. Oxford University Press, Oxford.
Horkheimer, M., 1972. Traditional and critical theory. In: Horkheimer, M. (Ed.), Critical Theory. Herder and Herder, New York.
Horkheimer, M., Adorno, T.W., 1973. Dialectic of Enlightenment. Allen Love, London.
Hullot-Kentor, R., 1989. Back to Adorno. Telos 81, 5–29.
Jay, M., 1988. Fin de Siècle Socialism. Routledge, London.
Kellner, D., n.d. Habermas, the Public Sphere, and Democracy: A Critical Intervention. http://www.gseis.ucla.edu/faculty/kellner/kellner.html.
Kellner, D., 1990. Television and the Crisis of Democracy. Westview Press, Boulder.
Matthewman, S., Hoey, D., 2006. What happened to postmodernism? Sociology 40 (3), 529–547.
Meehan, J., 1995. Introduction. In: Meehan, J. (Ed.), Feminists Read Habermas: Gendering the Subject of Discourse. Routledge, New York.
O'Mahony, P., 2011. Social theory and the critique of capitalism in a communication society. Irish Journal of Sociology 19 (1), 150–175.
Patterson, O., 1982. Slavery and Social Death: A Comparative Study. Harvard University Press, Cambridge, MA.
Rasmussen, D.M., 1990. Rereading Habermas. Blackwell, Oxford, UK.
Ray, L.J., 1993. Rethinking Critical Theory. Sage, London.
Ray, L.J., 2004. Pragmatism and critical theory. European Journal of Social Theory 7 (3), 307–321.
Ray, L.J., 2009. At the end of the post-Communist transformation? Normalization or imagining utopia? European Journal of Social Theory 2, 321–336.
Rogers, M.L., 2009. Rereading Honneth: Exodus politics and the paradox of recognition. European Journal of Political Theory 8 (2), 183–206.
Sayer, A., 2005. Class, moral worth and recognition. Sociology 39 (5), 947–963.
Scheff, T., 2000. Bloody Revenge – Emotions, Nationalism and War. Authors Guild, Lincoln NE.
Soper, K., 1989. Feminism as critique. New Left Review 176, 91–112.
Sorial, S., 2011. Habermas, feminism and law: beyond equality and difference? Ratio Juris 24 (1), 25–48.
Strydom, P., 2011. Towards a cognitive sociology for our time: Habermas and Honneth or language and recognition…and beyond. Irish Journal of Sociology 19 (1), 176–198.
Thompson, J., 1982. Universal pragmatics. In: Thompson, J.B., Held, D. (Eds.), Habermas Critical Debates. Macmillan, London, pp. 116–133.
Vogt, H., 2005. Between Utopia and Disillusionment: A Narrative of the Political Transformation in Eastern Europe. Berghahn, Oxford.
Zurn, C., 2003. Identity or status? In: Fraser, N., Honneth, A., Taylor, C. (Eds.), Constellations, vol. 10, pp. 519–537.
Zurn, C., 2005. Recognition, redistribution and democracy: dilemmas of Honneth's critical social theory. European Journal of Philosophy 13, 89–126.

Cross-Cultural Psychology

Ype H Poortinga, Tilburg University, Tilburg, Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H.C. Triandis, volume 5, pp. 2994–2999, © 2001, Elsevier Ltd.

Abstract

In this article, an inclusive approach is taken, covering the study of relationships between behavior and cultural context from three perspectives: the more classical culture-comparative school, cultural psychology, and indigenous psychologies. After a brief review of the history of cross-cultural psychology over the last half century, conceptualizations following the three perspectives are outlined. There are further sections on methodology, empirical research addressing controversial issues, and applied cross-cultural psychology in six areas (acculturation, intercultural relations, intercultural communication and training, work and organizations, health, and national development). A final section gives a brief outlook on the future.

Cross-cultural psychology (CCP) is one of several areas of scholarship that focuses on the relationship between human behavior and cultural context. One definition, formulated by Berry et al. (2011: p. 5), reads:

> Cross-cultural psychology is the study: of similarities and differences in individual psychological functioning in various cultural and ethnocultural groups; of ongoing changes in variables reflecting such functioning; and of the relationships of psychological variables with sociocultural, ecological and biological variables.

With an emphasis on similarities (i.e., cultural invariance) as well as on differences, and the explicit inclusion of ecological and biological variables this is a broad definition. CCP as defined here attempts to encompass three major orientations in the psychological study of behavior–culture relationships: culture-comparative research (the defining and initially dominant approach), cultural psychology, and indigenous psychologies. There is also overlap with entries for various fields of psychology, such as *Developmental Psychology*.

History

Themes found in contemporary thinking have a long history (Jahoda, 1992), but as a subdiscipline of psychology fitting the definition mentioned above CCP started to develop around the mid-twentieth century. In the early 1970s, the field became organized with the *Journal of Cross-Cultural Psychology* (first volume in 1970) and the *International Association for Cross-Cultural Psychology* (first congress in 1972). At that time, perception and cognition were major areas of study. Studies examining the susceptibility for visual illusions of rectangular buildings and streets (carpenteredness of the environment), and of open vs close vistas found variations related to characteristics of the natural environment (Segall et al., 1966). Lack of exposure to pictures (i.e., two-dimensional representations of three-dimensional objects and scenes) was shown to affect the perception of depth (see Deregowski, 1980). Such research demonstrated the relevance of external context and experience on how humans perceive the world.

Research on cognition had moved away from the racist agenda of comparing groups differing in skin color on IQ tests, a prominent, even though hotly debated, activity of (Western) psychology in the first half of the twentieth century. The new research in cognition was based on cognitively simple tasks and, at that time, recent models of information transmission. One example is the distinction between two cognitive styles, called field-independence and field-dependence; their relative prevalence was related to the mode of economic subsistence. Hunter-gatherers tended to be more field-independent (i.e., they identified more easily the target figure in embedded figures tests), than traditional agriculturists (Witkin and Berry, 1975). A second example is research conducted in Africa on recall and categorization in which the importance of local activities and modes of classification was demonstrated (Cole et al., 1971).

An important feature of projects as mentioned was that unschooled traditional farmers and even hunter-gatherers were included in the research design. More than in many recent studies, which tend to be limited to literate samples, it was recognized that a major strength of the cross-cultural method lies in extending the range of cultural variation to the maximum. Much of the early work was summarized in the first edition of the *Handbook of Cross-Cultural Psychology*, edited by Triandis et al. (1980).

After 1980, there have been two notable developments. First, the research agenda became predominantly social–psychological; the self-report questionnaire became the main method of data collection and differences in mean scores between cultures were interpreted in terms of broad cultural dimensions. Four major value dimensions were postulated by Hofstede (1980). One of these was further developed into the individualism–collectivism dimension of which its doyen, Triandis (1995), could rightly claim that it has become the most important syndrome of culture in the literature on CCP.

The second development had to do with a shift in cultural anthropology. In traditional ethnography, a good deal of attention was paid to culture as external context, including kinship relations and economic subsistence (i.e., the 'human-made part of the environment'). The shift was toward a greater focus on culture as internal context ('culture in the minds

of people'). Each separate culture became to be seen as constituting a set of meanings, or even a system of meanings, that should be understood in its own terms. The method of access was through 'thick description' (Geertz, 1973). This reorientation was at the basis of the development of cultural psychology, with the adage that behavior and culture makes each other up (Shweder, 1991). Such ideas were in tune with the concerns of psychologists in non-Western countries who observed that Western instruments, concepts, and theories did not fit the cultural realities of their societies. The rejection of psychological concepts originating in the West has been a major factor in the development of local indigenous psychologies, for example in India (Sinha, 1997).

In the 1990s, most students of behavior and culture belonged to one of two camps: universalism (a theoretical approach that considers basic psychological processes to be shared by all people, and cultures as influencing their manifestation and salience) and relativism (a theoretical approach emphasizing that each culture should be understood in its own terms, rather than being assessed with standards and instruments from elsewhere). In universalism, there is an emphasis on the 'psychic unity' of the human species; cross-cultural differences in manifest behavior should be explained in terms of common underlying psychological mechanisms and processes that are common to all humans. One might say that universalism builds on the assumed primacy of psychological principles. In relativism, such principles are seen as inherently cultural; the primacy of culture is assumed. In the volume on theory and method of the second edition of the *Handbook of Cross-Cultural Psychology* (Berry et al., 1997) the universalism–relativism distinction was the main theme of discussion. Since then, the underlying issues have not been resolved, but especially in applied research they have become less focal. There is now growing recognition that culture-common and culture-specific are the two sides of a single coin.

Theory

This growing respect for theoretical diversity hardly has led to more homogeneity in outlook. Established traditions continue to influence current ideas, and new formulations are adding to the existing variation. The following, often somewhat overlapping, approaches to behavior–culture relationships may be mentioned.

Cultural variables can be defined as *antecedents* of behavior outcomes. Such antecedents need not be deterministic causes; they can also refer to probabilistic and enabling conditions. Important antecedents are affluence (a compound variable covering income, formal education, access to media, etc.) and climate. Such external conditions are supposed to have effects that are internalized and may persist over longer historical periods. However, the resulting cross-cultural differences are interpreted largely in quantitative terms; i.e., differences in manifest behavior are explained in terms of common psychological mechanisms and processes that vary in the salience of their expression across cultures. For example, individualism tends to be conceived of as a broad value dimension with deep historical roots for which higher scores are found in Western societies than elsewhere.

Culture can also be defined as *mediating* between context and individual outcomes. This idea was expressed strongly by Vygotsky (1978) who saw the development of higher mental functions (especially abstract thinking) as a historical process at the level of societies that led to essential cross-cultural inequalities in these functions, especially between literate and illiterate societies. The influence of Vygotsky is explicit in the cultural-historical school of Cole (1996), but with one important modification: cultural context is not postulated to be operating at the level of broadly defined functions, but at the level of specific competences and skills, organized in fields of activity. Various other contemporary approaches, labeled 'sociocultural psychology' in Valsiner and Rosa (2007), such as social representation and action theories, share the basic premise that the culture of a group lies in the understanding and construction of their reality. For example, in action theories the unit of analysis in psychology is taken to be an action, which invariably is embedded in cultural and historical context, and encompasses the intentions of the actor.

The theoretical position of cultural psychology is that *psychological process and behavior content are inseparable*. At an elementary level, there is no disagreement on this; even the unequal pronunciation of the word 'table' in English and in French points to some (minute) differences in the innervations of the speech apparatus between speakers of these two languages. Controversy between schools of thought, or at least a difference in emphasis, emerges at more inclusive levels of conceptualization and explanation. The most important distinction in cultural psychology (overlapping with individualism vs collectivism) is between two modes of construal of the self: the independent and the interdependent mode (Markus and Kitayama, 1991). Methods of neuroscience have been used to find differences in brain functioning that could be characteristic of interdependent self-construal (in East Asian populations), as opposed to independent self-construal (Americans of European descent in the US).

Recent research in cultural psychology tends to postulate less overarching distinctions between cultural populations. For example, culture has been defined as *shared knowledge in a group* (Hong et al., 2000). Such knowledge, which is the outcome of cultural constructions, tends to be domain-specific and even situation-specific. This is shown by the ability of bicultural persons to switch between their two cultural frames when presented with 'primes' (e.g., icons or symbols) representing either of the two cultures. Still, in the *Handbook of Cultural Psychology* (Kitayama and Cohen, 2007) the theoretical position of inseparableness of process and behavior manifestation is largely maintained. At the same time, most empirical research is seeking to follow the experimental method of psychological science with administration of the same tasks across cultures followed by direct comparison of scores. This entails an important shift away from the initial position of cultural psychology, insofar as the use of common methods and concepts only makes sense under assumptions of universalism (see below).

A relativist position, rejecting the use of the same methods and concepts across cultures as (Western) cultural impositions and advocating *culture-specific and culture-appropriate* methods and concepts, continues to be maintained in schools

of indigenous research. Various levels can be distinguished in the evidence for culture-specific functioning. Most frequently reference is made to specific categorizations (e.g., emotion terms, color terms). Broader concepts include pathological syndromes, stages of ontogenetic development, and personality traits that are said to be unique for some culture. At the most general level, there are broad concepts of functioning that tend to be seen as incommensurable across cultures, such as *Ubuntu* in Africa, a form of social orientation that is reminiscent of collectivism (Mkhize, 2004). Evidence for cultural specificity often hinges on the absence of terms corresponding in meaning across languages (see below).

Methodology

Research in CCP has to deal with difficulties in design and measurement that are less central in other fields of psychology. First, experimental designs for causal inference are generally not available. In an ideal experiment, the experimenter has control over both the treatment conditions and the assignment of subjects to these conditions. In CCP, the treatment (i.e., some aspect of living in a certain cultural context) is often vague and imprecise, and participants' assignment to a culture is fixed by virtue of their membership of that culture. This implies that alternative explanations of a finding are difficult to rule out. The onus is on researchers to strengthen their interpretations through replication and checks on alternative explanations, but designs seeking discriminant validation of a preferred explanation continue to be rare in CCP.

Second, any grouping of humans can be referred to as a culture. Ideally, for a comparative study groups are selected that show clear between-group variance and within-group homogeneity for the domain or variable of interest (as for example, with language groups in research on psycholinguistics). The most common grouping in CCP is countries for which such homogeneity obviously is questionable; moreover, the selection of countries is often on and ad hoc basis. In studies, seeking to explore the universality of some theory or relationship between variables a random sample of all cultures would be required. Even larger studies with numerous samples tend to show a bias; as a rule illiterate groups and groups with traditional means of economic subsistence are underrepresented or even totally absent.

The third and most studied methodological issue concerns the question whether or not it makes sense to use a psychological instrument (psychometric test or questionnaire) constructed in one cultural context in some other cultural context. When cultural specificity of behavior is emphasized, virtually any cross-cultural use of instruments is ruled out. On the other hand, researchers with strong universalist views tend to assume that only minor (linguistic) adaptations are needed. The key question is whether psychological data obtained from subjects in different cultures can be considered equivalent (or comparable). For most researchers in CCP, equivalence has become an empirical issue. Over several decades, an increasingly sophisticated set of psychometric approaches has been developed to examine cultural bias (i.e., differences in scores on tests and questionnaires that do not reflect corresponding differences in the trait or concept presumably measured and distort the interpretation of the scores) (Van de Vijver and Leung, 1997). The old dichotomy of a scale being equivalent or inequivalent has been replaced by distinctions between various levels of equivalence that can be examined in a hierarchical sequence of analyses. It may be found, for example, that a cross-cultural data set does not meet psychometric conditions for full-scale equivalence, but only conditions for structural equivalence. Full-scale equivalence implies that a score of a certain value corresponds to the same standing on the trait being assessed, independent of the culture of the respondent. Structural equivalence implies that the same trait is being assessed, as demonstrated, for example, by high correspondence of factor structures. But structural equivalence does not imply that across cultures test takers with the same scores on the instrument will have the same standing on the underlying trait. Evidence of inequivalence, or cultural bias, imposes constraints on cross-cultural comparison of data. It may be noted that bias refers to systematic variance (rather than error variance) and for this reason further analysis is indicated. In such attempts, a combination of qualitative and quantitative approaches, called 'mixed methods' are used to trace the cultural factors in terms of which observed bias can be interpreted (Van de Vijver, 2011).

In a carefully designed culture-comparative study researchers from the various participating cultures should be involved in the design and formulation of items, optimizing translation equivalence. Post hoc analyses after collection of the data will determine whether or not conditions for psychometric equivalence have been met. Large-scale projects for international assessment of quality of education, such as PISA (Programme for International Student Assessment) and TIMMS (Trends in International Mathematics and Science Study), follow this methodology. Such careful design is still exceptional and most studies in CCP are based on existing instruments. Even then, the local experts can help in the translation and adaptation of an instrument to a new target culture. The International Test Commission (2011) has published a set of Guidelines on Adapting Tests that is available at www.intestcom.org/guidelines.

Topics in Empirical Research

The relevance of cultural context for behavior is widely recognized and has led to a large research literature that is impossible to summarize in a brief review. Here three examples will be given of how empirical findings are advancing insights in behavior–culture relationships. The first example reflects on differences between cultures relative to individual differences within cultures. Recent estimates show that the ratio of between-culture variance and within-culture variance (i.e., individual variance) is in the order of 1:10 for self-reports of value types (e.g., Fischer et al., 2010) and 1:8 for dimensions of personality (McCrae et al., 2005). Such estimates of cross-cultural differences are much smaller than would be expected by both experts and the public at large. What are the implications? Of course, there are variables for which there is large heterogeneity between cultures and homogeneity within cultures. This is the case with many conventions (i.e., agreed upon ways how to act in certain situations, what to believe,

'default strategies' in responding, etc.). Also, substantial differences are found in mean ratings of happiness and well-being that are closely associated with the affluence and security, which societies provide for their members (Diener et al., 2010). Together such findings suggest substantial differences in manifest behavior as a function of features in the external context, while psychological mechanisms and processes underlying manifest behavior may be more invariant across cultures than thought in the past.

The second example is the extensive debate on emotions. Words for emotions are found in all cultures, but for some such words in a language there may not be translation equivalent words in some other language (usually English). In cultural anthropology and cultural psychology, it has been argued that such differences point to emotions being cultural constructions that form variable categories. In culture-comparative research emotions tend to be conceived of as biologically rooted discrete categories; here cross-cultural differences are associated with differences in the salience of a certain emotion and with display rules, which proscribe the expression of emotions in various situations (Matsumoto et al., 2008). Older evidence of culture-specificity of emotions hinging on the specificity of emotion terms has been largely refuted. At the same time, a more complex view of emotions has come from componential theory, in which cultural variation can occur in each of several distinguishable features, such as the antecedents events leading to an emotion, bodily reactions, and action tendencies. However, ratings of emotion terms on numerous emotion features across numerous countries and languages have shown limited cultural variation (Fontaine et al., 2013). In turn, a new challenge has emerged from a failure in neuroscience to identify features in the brain characteristic for various emotion categories, which is taken as evidence for constructionist (relativist) views of emotions (Lindquist et al., 2012). All in all, viewpoints appear to be less extreme than a few decades ago, but a unified perspective on culture and emotion is still far off.

Research on relationships between language and thinking, the third example, also can be characterized as gradually having moved from more extreme to more moderate positions both in universalist and in relativist approaches. Much research has been inspired by Whorf's hypothesis that the vocabulary and grammar of a language are determinants of the ideas and thoughts found with the speakers of that language. Such extensive claims by Whorf soon were shown not to stand upto empirical scrutiny and were followed by strong universalist positions, exemplified in the notion of a universal grammar and other invariant features. In turn, for various allegedly invariant properties of languages, exceptions have been found in comparative linguistics, challenging universality (Evans and Levinson, 2009). Research on spatial orientation provides an illustration how careful analysis can point the way to convergence of viewpoints. In Indo-European languages, the location of objects mostly is expressed using an ego-referenced orientation. The statement: 'the chair is to the left side of the table' will change to 'the chair is to the right side of the table' when the speaker moves to the opposite side of the table. In some other language families, there is a preference for an absolute orientation that is independent of the position of the viewer (e.g., 'the chair is to the North of the table,' or 'to the side of the rising sun'). The two frames of orientation have been associated with performance on nonlinguistic tasks and have been taken to support the argument that language has profound effects on cognition. Dasen and Mishra (2010) carried out a series of careful studies in which language was seen as one aspect of a more general web of ecological and sociocultural factors. They found that the use of a particular frame by a respondent depended not only on the dominant orientation mode of the language, but also on features of the task, such as the ease of encoding either with an ego-referenced or an absolute orientation. Their conclusion was that across languages, individuals possess the basic processes needed for either frame.

Each of the three research topics mentioned is an illustration of research in action. Theoretical controversies have not been resolved fully in either of the three, but empirical evidence is slowly shaping insights. Research designs are becoming more sophisticated and there appears to be a shift from claims about major and extensive cross-cultural differences toward more modest interpretations.

Applied Cross-Cultural Research

Most applied research in CCP has a pragmatic agenda; the main question is not how or how much peoples across cultures differ, but how differences can be best addressed and dealt with. There are various fields of applied CCP; in their textbook Berry et al. (2011) present six such areas: acculturation, intercultural relations, intercultural communication and training, work and organizations, health, and national development.

Research on acculturation is rapidly expanding in response to the increase in worldwide migration. The earlier approach to acculturation, which assumed that there would be more or less successful assimilation by migrants to the new culture, has been largely abandoned. It is now recognized that migrants can follow various acculturation strategies. They can focus on the maintenance of what they consider positive in their culture of origin (such as their religion) or on gaining access to the new culture, or on both. There need not be one uniform strategy; there can be variations across domains, such as the public domain and the private domain (with family and friends). In addition, it is recognized that acculturation does not only depend on the migrants themselves, but also on attitudes toward migrants in the larger society and the kinds of settlement policies of a country. Further diversification in research has resulted from differences in theoretical approaches that emphasize stress and coping, cultural learning or social identification (Ward, 2001).

The area of intergroup relations has an international dimension, but most psychological research deals with the relations between groups within culturally plural societies. It is widely held that the best arrangement is multiculturalism, which presupposes an overarching framework of shared values and goals in a society, and mutual tolerance and respect of ethnic and religious groups for each others' cultures and identities. Positive intergroup relations are threatened by psychological mechanisms expressed in concepts such as

stereotypes (cognitive), prejudices (evaluative), and discrimination (overt behavior). The most well-known body of work in this area, aimed at overcoming negative attitudes and discrimination, is based on Allport's contact hypothesis. This hypothesis holds that, under favorable conditions, contact and sharing among members of various groups will promote mutual acceptance. There is extensive evidence supporting Allport's conjecture (Pettigrew and Tropp, 2011) and this has opened up promising perspectives for successful social intervention programs.

Intercultural communication is addressed in a range of disciplines, including linguistics and sociolinguistics, cultural anthropology and communications research. The relevance of this area derives from growing globalization and internationalization. Presence or absence of a shared language is a key factor in communication, but misunderstandings also may follow from subtle factors, such as paralinguistic differences (e.g., pragmatics, such as turn-taking) and mismatches in conventions. In contrast, theories of intercultural communication tend to provide explanations for communication difficulties in terms of dimensions at a high level of abstraction, with individualism–collectivism as the main distinction (Gudykunst, 2005). The main field for intervention is the training of sojourners or expatriates who are preparing to live and work in another culture. There is a host of training programs; one dimension on which they vary is from culture-general (promoting a deeper understanding of both self and cultural others) to culture-specific (information about rules and customs in a society). Unfortunately, for most interventions it is difficult to estimate their effectiveness. Positive reactions of participants form the main evidence for success; few programs have been validated with less impressionistic data.

Cross-cultural research on work and organizations is a large and active area with research ranging from structural characteristics of organizations as they are found in various societies to individual work-related psychological variables. One persistent question is whether across countries the structure of organizations is converging in line with technological demands or whether cultural characteristics prevent such convergence. The largest volume of research is on work-related values across countries, following Hofstede's (1980) formulation of four value dimensions. Despite increasing criticisms this set of dimensions still provides an explanatory framework for a host of differences between countries not only in psychology, but also in other disciplines, such as economics and management research. Another major focus is leadership behavior with topics such as decision making and leadership styles. In the latter domain, the Global Leadership and Organizational Behavioral Effectiveness (GLOBE) project (House et al., 2004) has been a major effort. Studies of work-related individual psychological variables include motivation, job attitudes, organizational commitment, etc. (e.g., Smith et al., 2008).

Applied cross-cultural research on health is the most diversified of the six areas mentioned here. Perhaps more than in other areas the whole range of conceptual and methodological issues is present, including conceptual discussions on the nature of culture, cultural specificity of illnesses and interventions, and (in)equivalence of diagnostic procedures and assessment scales. The area as a whole has been moving from illness as a pathological state to health as a state of well-being with a simultaneous extension from treatment to prevention. Classical topics are the universality of the major mental disorders, schizophrenia and depression, and the place of (non-Western) culture-bound syndromes in Western psychiatric classifications. Another topic, challenging dominant conceptions of psychotherapy, is the study of indigenous approaches to mental healing, with religious principles being emphasized in Muslim countries, the spiritual principle of ancestors in Africa, and retribution and punishment in some Japanese schools of psychotherapy. Prevention research and interventions address severe real-life problems, including malnutrition and transmission of HIV and other sexual diseases. In the more recently developed approach of positive mental health, there are applications on quality-of-life, stress and posttraumatic stress, and on a host of intervention programs seeking to enhance empowerment and well-being of poor and marginalized groups, especially children and women.

CCP and national development is the smallest of the six areas of applied research mentioned here. For a long time, national development was mostly owned by economists and only since it has become evident that financial and technical support are not sufficient to bring about lasting change, has there been more interest in the cultural and psychological side. Still, most of the contributions to the understanding of development come from developmental economists, such as Sen (1999) who views expansion of 'freedoms' (psychologists might say the 'capability of making choices') as the principal route of development. This suggests that intervention programs seeking to enhance empowerment and autonomy might include economic as well as health-related behaviors (Pick and Sirkin, 2010).

Future Outlook

An array of achievements can be listed showing how over the last half century CCP has contributed to insights in the relationships between behavior and culture, but is this reason for satisfaction? *Eminence grise* Gustav Jahoda (2011) recently argued that CCP has become unduly narrow and inward looking. At the same time, he is positive about the potential of CCP, outlining a promising future with theory-driven experimental research, the inclusion of samples from nonliterate societies, and the use of methods that require more direct contacts with participants than the frequently used questionnaire.

A more outward looking CCP will have to incorporate biological and evolutionary thinking. Traditionally, the concept of culture has been closely associated with differences in behavior between groups. Despite theoretical disagreements, outlined earlier, there continues to be broad consensus between researchers of various schools that the interest of research on culture and behavior mainly lies in on how, or how much, cultural populations differ from each other. There is evidence that researchers are biased toward finding and interpreting differences while ignoring evidence of cross-cultural invariance. A further bias is toward interpretations in terms of broad and inclusive concepts. In the history of CCP, it has happened time and again that initially far-reaching

generalizations of cross-cultural differences came to be seen as overgeneralizations in the light of subsequent more precise data (Poortinga, 2003).

The future relevance of CCP may well lie in conceptualizing culture as a common human propensity and understanding cross-cultural variance on psychological variables as variations on common themes. If so, CCP will have to complement its legacy from cultural anthropology with approaches rooted in biology. In that way, the field is well placed to contribute to and counterbalance the biological revolution in psychology. Cross-cultural research brings a much needed range of variation, which can help to put into question too easy invocations of genetic determinants and phylogenetic processes as long as the complexities of environmental influences on gene expression and epigenesis are poorly understood.

This does not mean that CCP should adopt the methods of current biology. Various alternatives are imaginable, one of which would be to follow an integrative approach as envisaged for ethology by Tinbergen (1963) at a time when in that discipline direct observation was the major method. Tinbergen argued that four questions have to be answered in the analysis of any behavior pattern: (1) the mechanisms or causes of the behavior, (2) its evolutionary history, (3) its ontogenetic development, and (4) the function it supposedly serves. There has been debate on the specific questions raised by Tinbergen and in CCP a question about the historical origin of a behavior pattern might have to be added. The point is that in an integrative approach, more questions have to be answered simultaneously and that this urges researchers to consider multiple perspectives.

Whatever vision one may hold on the future of CCP there remains a need to clarify basic issues. In a recent chapter Van de Vijver et al. (2011) formulated four such questions: (1) how 'deep' or profound are cross-cultural differences in psychological functioning; (2) what are the main methodological challenges to the field; (3) how does culture become engrained in human development; and (4) what is the relationship between individual and culture. It is unlikely that agreement will be reached on these questions anytime soon, but that does not rule out further convergence to at least partly shared views.

See also: Acculturation; Collectivism and Individualism: Cultural and Psychological Concerns; Cross-Cultural Research Methods in Sociology; Cross-Cultural Study of Education; Cultural Influences on Interpersonal Relationships; Cultural Psychology; Culture and Emotion; Culture and the Self: Implications for Psychological Theory; Immigration: Social Psychological Aspects; Indigenous Psychology; Intergroup Relations; Quantitative Cross-national Research Methods; Representations, Social Psychology of; Research Ethics, Cross-Cultural Dimensions of; Translating Sociology into Japanese, Chinese and Korean Languages.

Bibliography

Berry, J.W., Poortinga, Y.H., Breugelmans, S.M., Chasiotis, A., Sam, D.L., 2011. Cross-Cultural Psychology: Research and Applications, third ed. Cambridge University Press, Cambridge.

Berry, J.W., Poortinga, Y.H., Pandey, J., Dasen, P.R., Saraswathi, T.S., Segall, M.H., Kagitcibasi, C. (Eds.), 1997. Handbook of Cross-Cultural Psychology, second ed., vols. I–III. Allyn & Bacon, Boston.

Cole, M., 1996. Cultural Psychology: A Once and Future Discipline. Belknap, Cambridge, Mass.

Cole, M., Gay, J., Glick, J., Sharp, D., 1971. The Cultural Context of Learning and Thinking. Basic Books, New York.

Dasen, P.R., Mishra, R.C., 2010. Development of Geocentric Spatial Language and Cognition. Cambridge University Press, Cambridge.

Deregowski, J.B., 1980. Illusions, Patterns and Pictures: A Cross-Cultural Perspective. Academic Press, London.

Diener, E., Ng, W., Harter, J., Arora, R., 2010. Wealth and happiness across the world: material prosperity predicts life evaluation, whereas psychosocial prosperity predicts positive feeling. Journal of Personality and Social Psychology 99, 52–61.

Evans, N., Levinson, S.C., 2009. The myth of language universals: language diversity and its importance for cognitive science. Behavioral and Brain Sciences 32, 429–448.

Fischer, R., Vauclair, C.M., Fontaine, J.R.J., Schwartz, S.H., 2010. Are individual-level and country-level value structures different? Testing Hofstede's legacy with the Schwartz Value Survey. Journal of Cross-Cultural Psychology 41, 135–151.

Fontaine, J.R.J., Scherer, K.R., Soriano, C. (Eds.), 2013. Components of Emotional Meaning: A Sourcebook. Oxford University Press, Oxford.

Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.

Gudykunst, W.B. (Ed.), 2005. Theorizing about Intercultural Communication. Sage, Thousand Oaks, Calif.

Hofstede, G., 1980. Culture's Consequences: International Differences in Work-Related Values. Sage, Beverly Hills.

Hong, Y., Morris, M., Chiu, C., Benet-Martínez, V., 2000. Multicultural minds: a dynamic constructivist approach to culture and cognition. American Psychologist 55, 709–720.

House, R.J., Hanges, P.J., Javidan, M., Dorfman, P.W., Gupta, V., 2004. Culture, Leadership and Organizations: The GLOBE Study of 62 Societies. Sage, Thousand Oaks, Calif.

International Test Commission, 2011. International Test Commission Guidelines for Translating and Adapting Tests. http://www.intestcom.org.

Jahoda, G., 1992. Crossroads between Culture and Mind. Harvester Wheatsheaf, New York.

Jahoda, G., 2011. Past and present of cross-cultural psychology. In: Van de Vijver, F.J.R., Chasiotis, A., Breugelmans, S.M. (Eds.), Fundamental Questions in Cross-Cultural Psychology. Cambridge University Press, Cambridge, Mass, pp. 37–63.

Kitayama, S., Cohen, D. (Eds.), 2007. Handbook of Cultural Psychology. Guildford Press, New York.

Lindquist, K.A., Wager, T.D., Kober, H., Bliss-Moreau, E., Feldman Barrett, L., 2012. The brain basis of emotion: a meta-analytic review. Behavioral and Brain Sciences 35, 121–202.

Markus, H.R., Kitayama, S., 1991. Culture and the self: implications for cognition, emotion and motivation. Psychological Review 98, 244–253.

Matsumoto, D., Yoo, S.H., Fontaine, J., et al., 2008. Mapping expressive differences around the world: the relationship between emotional display rules and individualism v. collectivism. Journal of Cross-Cultural Psychology 39, 55–74.

McCrae, R.R., Terracciano, A., 79 Members of the Personality Profiles of Cultures Project, 2005. Personality profiles of cultures: aggregate personality traits. Journal of Personality and Social Psychology 89, 407–425.

Mkhize, N., 2004. Psychology: an African perspective. In: Ratele, K., Duncan, N., Hook, D., Mkhize, N., Kiguwa, P., Collins, A. (Eds.), Self, Community and Psychology. UCT Press, Cape Town, pp. 1–29.

Pettigrew, T., Tropp, L., 2011. When Groups Meet. Psychology Press, London.

Pick, S., Sirkin, J., 2010. Breaking the Cycle of Poverty: The Human Basis for Sustainable Development. Oxford University Press, New York.

Poortinga, Y.H., 2003. Coherence of culture and generalizability of data: two questionable assumptions in cross-cultural psychology. In: Berman, J., Berman, J. (Eds.), Cross-Cultural Differences in Perspectives on the Self, Nebraska Symposium on Motivation, vol. 49. University of Nebraska Press, Lincoln, Nebr, pp. 257–305.

Segall, M.H., Campbell, D.T., Herskovits, K.J., 1966. The Influence of Culture on Visual Perception. Bobbs-Merrill, Indianapolis.

Sen, A., 1999. Development as Freedom. Oxford University Press, Oxford.

Shweder, R.A., 1991. Thinking through Cultures: Expeditions in Cultural Psychology. Harvard University Press, Cambridge, Mass.

Sinha, D., 1997. Indigenizing psychology. In: Berry, J.W., Poortinga, Y.H., Pandey, J. (Eds.), Handbook of Cross-Cultural Psychology, Theory and Method, second ed., vol. I. Allyn & Bacon, Boston, pp. 129–169.

Smith, P.B., Peterson, M.F., Thomas, D.C., 2008. The Handbook of Cross-Cultural Management Research. Sage, Los Angeles.

Tinbergen, N., 1963. On aims and methods of ethology. Zeitschrift fuer Tierpsychologie 20, 410–433.
Triandis, H.C., 1995. Individualism and Collectivism. Westview, Boulder, Colo.
Triandis, H.C., et al. (Eds.), 1980. Handbook of Cross-Cultural Psychology, vols. 1–6. Allyn & Bacon, Boston.
Valsiner, J., Rosa, A. (Eds.), 2007. The Cambridge Handbook of Sociocultural Psychology. Cambridge University Press, Cambridge.
Van de Vijver, F.J.R., 2011. Bias and real differences in cross-cultural differences: neither friends nor foes. In: Van de Vijver, F.J.R., Chasiotis, A., Breugelmans, S.M. (Eds.), Fundamental Questions in Cross-Cultural Psychology. Cambridge University Press, Cambridge, pp. 235–257.
Van de Vijver, F.J.R., Chasiotis, A., Breugelmans, S.M. (Eds.), 2011. Fundamental Questions of Cross-Cultural Psychology. Cambridge University Press, Cambridge, pp. 9–34.
Van de Vijver, F.J.R., Leung, K., 1997. Methods and Data Analysis for Cross-Cultural Research. Sage, Newbury Park, CA.
Vygotsky, L.S., 1978. Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA.
Ward, C., 2001. The A, B, Cs of acculturation. In: Matsumoto, D. (Ed.), The Handbook of Culture and Psychology. Oxford University Press, Oxford, pp. 411–445.
Witkin, H., Berry, J.W., 1975. Psychological differentiation in cross-cultural perspective. Journal of Cross-Cultural Psychology 6, 4–87.

Cross-Cultural Research Methods in Psychology

Fons van de Vijver, Tilburg University, Tilburg, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp. 2999–3003, © 2011, Elsevier Ltd., with revisions made by the Editor.

Abstract

Cross-cultural studies involve persons from different countries and/or ethnic groups. One of the central methodological problems of these studies is bias, the generic term for multiple explanations of cross-cultural differences. Three different types of bias are distinguished, depending on whether the source of interpretation problems derives from the construct, method of the study, or specific items (called construct, method, and item bias or differential item functioning, respectively). Equivalence refers to the implications of bias on score comparability. Linguistic, structural, measurement unit, and full score equivalence are described. Issues in test translation (translation – back-translation, committee designs, decentering) are discussed. Common subject- and culture-sampling schemes in cross-cultural research are mentioned. The article ends with a discussion of issues in combining individual- and country-level characteristics.

Cross-cultural studies involve persons from different countries and/or ethnic groups; a defining characteristic is their comparative nature. Most studies employ quantitative methods of data collection and analysis. Studies of cultural topics that are noncomparative and apply a qualitative methodology can be found in sociology ('cultural studies,' e.g., Barker, 2003), cultural psychology (Greenfield et al., 2003), and cultural anthropology (Angrosino, 2006).

The range of instruments used in comparative studies is very broad, ranging from highly standardized psychological tests, to observation schedules, and free interviews. In many studies existing Western instruments (mental tests, survey questionnaires, personality inventories) are administered either in a new cultural context, or not adapted to enhance their cultural appropriateness (Hambleton et al., 2004).

If two persons from different cultural groups show different scores on a reliable and valid measure of subjective well-being, these score differences may refer to individual differences in subjective well-being. However, the score differences may also arise from differential social desirability or some other response style, inappropriate translation, or inadequacy of the item to measure well-being in both groups. The example illustrates a central problem in cross-cultural research: observed score differences are often susceptible to multiple explanations (Creswell, 2013). When the same instrument has been administered to persons from different ethnic groups, it cannot be taken for granted that the same scores obtained in different cultural groups have the same psychological meaning.

The ambiguity of interpretation is a consequence of the methodological nature of culture as an independent variable. In laboratory studies researchers randomly assign subjects to experimental treatments. The random assignment leads to a firm control of ambient variables; ideally an experimental and control group are matched on all outcome-relevant characteristics (e.g., personality characteristics and socioeconomic status), except for the treatment variable studied (*see* Internal Validity). However, like gender and other intrinsic subject characteristics, culture is not an experimental treatment that can be manipulated. Groups with a different cultural background tend to differ on a variety of outcome-relevant characteristics. These differences may constitute rival explanations of observed cross-cultural differences. Without precautions to rule out these rival explanations, observed cross-cultural differences are open to multiple interpretations. Findings in cross-cultural research are more convincing when rival explanations have been more adequately dealt with.

Bias is the generic name of an important family of rival explanations (*see* Psychometrics). It refers to the common problem in the assessment of nonequivalent groups that scores obtained in different cultural groups are not an adequate reflection of the groups' standing on the construct underlying the instrument. If scores are biased, their psychological meaning is group dependent and group differences in assessment outcome are to be accounted for, at least to some extent, by auxiliary psychological constructs or measurement artifacts. A closely related concept is equivalence which refers to the absence of bias and hence, to similarity of meaning across groups. The two concepts have somewhat different historical roots and areas of application. Whereas bias usually refers to nuisance factors, equivalence has become the generic term for metrical implications of bias.

Bias and equivalence are not inherent properties of an instrument but arise in a group comparison with a particular instrument. Score comparisons of groups that differ in more test-relevant aspects will show a higher susceptibility to bias.

Sources of Bias

There are three bias sources in cross-cultural research. The first is called construct bias; it occurs when the construct measured is not identical across groups. Work on filial piety (psychological characteristics associated with being a good son or daughter) provides a good example (e.g., Yeh and Bredford, 2003). The Western conceptualization is narrower than the Chinese, according to which children are supposed to assume the role of caretaker of their parents when these grow old and become needy. Construct bias precludes the cross-cultural measurement of a construct with the same measure. An

inventory of filial piety based on the Chinese conceptualization will cover aspects unrelated to the concept among Western subjects, while a Western-based inventory will leave an important Chinese aspect uncovered.

An important source of bias, called method bias, can result from sample incomparability, instrument characteristics, tester and interviewer effects, and the method (mode) of administration. Examples are differential stimulus familiarity in mental testing and differential social desirability in personality and survey research. Some sources of method bias can be dealt with by careful preparation of the assessment instrument and its instruction manual (e.g., proper test instruction with a clear specification of what is asked from participants, standardization of administration, and adequate training of testers and interviewers). Yet, it may be impossible to eliminate all outcome-relevant sample characteristics, in particular when the cultural distance of the countries involved is large. There are indications that a country's Gross National Product (per capita) is positively related to its mean score on mental tests and negatively to its mean score on social desirability. Particularly in comparisons of culturally highly dissimilar groups it may be hard or even impossible to eliminate the impact of sources of method bias such as sources familiarity and social desirability.

Finally, bias can be due to anomalies at item level (e.g., poor translations); this is called item bias or differential item functioning. According to a definition that is used widely in psychology, an item is biased if persons with the same standing on the underlying construct (e.g., they are equally intelligent) but coming from different cultural groups, do not have the same average score on the item. The score on the construct usually is derived from the total test score. If a geography test administered to pupils in Poland and Japan, contains the item 'What is the capital of Poland?' Polish pupils can be expected to show higher scores on the item than Japanese students, even when pupils with the same total test score would be compared. The item is biased as it favors one cultural group across all test score levels. Of all bias types, item bias has been the most extensively studied; various psychometric techniques are available to identify item bias (e.g., Obinne and Amali, 2014).

An overview of common ways of addressing bias is given in Table 1.

Types of Equivalence

Elaborating on categorizations in the literature, four different types of equivalence are proposed here. The first type is labeled construct inequivalence. It amounts to comparing apples and oranges (e.g., the comparison of Chinese and Western filial piety, discussed above). Comparisons lack an attribute for comparison, also called tertium comparationis (the third term in the comparison). The second is called structural (or functional) equivalence. An instrument, administered in different cultural groups, shows structural equivalence if it measures the same construct in both groups (e.g., Raven's Standard Progressive Matrices test has been found to measure intelligence in various cultural groups). Exploratory factor analyses followed by target rotations or confirmatory factor analysis of correlation matrices (structural equation modeling) may be applied to examine structural equivalence. Structural equivalence does not presuppose identity of measures across groups. The measures may use different stimuli across groups. If different operationalizations have been chosen, structural equivalence can be examined by comparing nomological networks across groups.

The third type of equivalence is called measurement unit equivalence. Instruments show this type of equivalence if their measurement scales have the same units of measurement and a different origin (such as the Celsius and Kelvin scales in temperature measurement). This type of equivalence assumes interval- or ratio-level scores (with the same measurement units in each culture). It applies when a bias factor with a fairly

Table 1 Strategies for identifying and dealing with bias in cross-cultural research

Type of bias	*Strategies*
Construct bias	Decentering (i.e., simultaneously developing the same instrument in several cultures)
	Convergence approach (i.e., independent within-culture development of instruments and subsequent cross-cultural administration of all instruments)
Construct and/or method bias	Use of informants with expertise in local culture and language
	Use samples of bilingual subjects
	Use of local surveys (e.g., content analyses of free-response questions)
	Nonstandard instrument administration (e.g., 'thinking aloud')
	Cross-cultural comparison of nomological networks (e.g., convergent/discriminant validity studies, monotrait-multimethod studies, connotation of key phrases)
Method bias	Extensive training of administrators (e.g., increasing cultural sensitivity)
	Detailed manual/protocol for administration, scoring, and interpretation
	Detailed instructions (e.g., with sufficient number of examples and/or exercises)
	Use of subject and context variables (e.g., educational background)
	Use of collateral information (e.g., test-taking behavior or test attitudes)
	Assessment of response styles (e.g., social desirability, acquiescence)
	Use of test–retest, training and/or intervention studies
Item bias	Judgmental methods of item bias detection (e.g., linguistic and psychological analysis)
	Psychometric methods of item bias detection (e.g., differential item functioning analysis)
	Error or distracter analysis

Source: Van de Vijver and Tanzer, 1997. European Review of Psychological Assessment (reproduced with the permission of Swets and Zeitlinger).

uniform influence on the items of an instrument affects test scores of different cultural groups in a differential way. Social desirability and stimulus familiarity may exert this influence. Observed group differences in scores are then a mixture of valid cross-cultural differences and measurement artifacts. When the relative contribution of both sources cannot be estimated, the interpretation of group comparisons of mean scores remains ambiguous. Multigroup comparisons of confirmatory factor analytic models have been utilized to examine measurement unit equivalence. This technique, based on comparisons of covariance matrices, can identify bias sources that affect the covariance of items or tests but it cannot differentiate between valid differences in mean scores and bias sources with a uniform influence on all parts of an instrument.

Only in the case of scalar (or full-score) equivalence can direct comparisons be made; it is the only type of equivalence that allows for the conclusion that average scores obtained in two cultures are different. This type of equivalence assumes the same interval or ratio scales across groups. Conclusions about which of the latter two types of equivalence applies are often difficult to draw and controversial. For example, racial differences in intelligence test scores have been interpreted as due to valid differences (scalar equivalence) and as reflecting measurement artifacts (measurement unit equivalence). Scalar equivalence assumes that the role of bias can be safely neglected. The demonstration of scalar equivalence draws on inductive argumentation. Therefore, it is easier to disprove than to prove scalar equivalence. This can be made plausible by measuring presumably relevant sources of bias (such as stimulus familiarity or social desirability) and showing that they cannot statistically explain observed cross-cultural differences in a multiple regression or covariance analysis.

The distinction between the latter two types of equivalence is immaterial when comparing experimental conditions or changes across cultures (e.g., developmental trajectories or training effects). In respect of scales that show measurement unit equivalence, measure changes at the level of full score equivalence.

Structural, measurement unit, and scalar equivalence are hierarchically ordered. The third presupposes the second, which presupposes the first. Moreover, higher levels of equivalence are more difficult to establish. It is easier to demonstrate that an instrument measures the same construct in different cultural groups (structural equivalence) than to demonstrate numerical comparability across cultures (scalar equivalence). On the other hand, higher levels of equivalence allow for more detailed comparisons of scores across cultures. Whereas in the case of structural equivalence, the only factor on which structures and nomological networks can be compared, scalar equivalence allows for more fine-grained analyses of cross-cultural similarities and differences, such as comparisons of mean scores across cultures in *t* tests and analyses of (co)variance.

Linguistic Equivalence

The concept of linguistic equivalence is developed in multilingual studies. Versions of an instrument in different languages show linguistic equivalence if these have the same characteristics that are relevant for the measurement outcome such as meaning, connotations of words and sentences, comprehensibility, and readability. Linguistic equivalence can be jeopardized by various sources, such as incorrect translations, words that are hard or impossible to translate (e.g., the English expression 'distress' does not have an equivalent in many languages), idiomatic expressions and metaphors (e.g., 'feeling blue'), and imprecise quantifiers ('rather often').

Because linguistic equivalence is not always guaranteed by a literal translation, it has become increasingly popular to utilize adaptations. In an adaptation, parts are changed (instead of literally translated) with the aim to improve an instrument's suitability for a target group.

Many studies employ a translation–back-translation procedure (Sperber, 2004). This amounts to a forward translation, followed by an independent back-translation and a comparison of the original and back-translated version, possibly followed by some alterations of the translation. Such a procedure provides a powerful tool to enhance the correspondence of original and translated versions that is independent of the researcher's knowledge of the target language. Yet, it also has some disadvantages. It puts a premium on literal reproduction; this may give rise to a stilted language use in the target version that lacks the readability and natural flow of the original. The problem may be compounded by translators' awareness of their involvement in a translation–back-translation procedure.

A second problem involves translatability. The use of idiom (e.g., the English 'feeling blue') or references to cultural specifics (e.g., references to country-specific public holidays) or other features that cannot be represented adequately in the target language challenges translation–back-translations designs (and indeed all procedures in which existing instruments are translated). During the 1990s there was a growing awareness that translations and adaptations require the combined expertise of social and behavioral scientists (with a competence in the construct studied) and experts in the target language(s) and culture(s). In this so-called committee approach in which the expertise of all relevant disciplines is combined, there is usually no formal accuracy check of the translation. Usage of the committee approach is popular in large international bodies in which texts are translated in many languages like the United Nations and the European Union.

When versions in all languages are developed simultaneously, a procedure called 'decentering' can be used. No single instrument or cultural group is then taken as starting point; individuals from different cultures jointly develop an instrument, thereby reducing the risk of introducing unwanted references to a specific culture.

The judgmental evidence of the designs to enhance linguistic equivalence can be easily combined with statistical approaches to establish equivalence. Reports of multilingual studies often provide a combination of judgmental–linguistic and empirical–statistical evidence to demonstrate the adequacy of an instrument and its translation. There is a tradeoff between the type of translation and the level of equivalence that can be obtained. When most or all questions of an instrument are adapted, structural equivalence is the highest level possible. When most or all items are literally translated, measurement unit and structural equivalence can be obtained. Recent statistical advancements in item response theory and structural equation modeling have made it possible to retain scalar equivalence,

even when not all stimuli are literally translated (provided that all items measure the same underlying construct in each group).

Sampling Cultures and Subjects

Cross-cultural studies can apply three types of schemes to sample cultures. Three types of sampling can be envisaged. The first is probability (or random) sampling. Because of the large cost of a probability sample from all existing cultures, it often amounts to stratified (random) sampling of specific cultures (e.g., Western cultures). The second and most frequently observed type of culture sampling is convenience sampling. The choice of cultures is governed here by availability and cost efficiency: researchers decide to form a research network and all participants collect data in their own country. In the third type, called systematic sampling, the choice of cultures is more based on substantive considerations. A culture is deliberately chosen because of some characteristic, such as in Nisbett and Miyamoto's (2005) study.

In survey research there is a well-developed theory of subject sampling (Scheaffer et al., 2011). In the area of cross-cultural research three types of sampling procedures of individuals are relevant as they represent different ways of dealing with confounding characteristics. The first is probability sampling. It consists of a random drawing from a list of eligible units such as persons or households. Confounding variables are not controlled for. The second type is stratified sampling. A population is stratified (e.g., in levels of schooling or socioeconomic status) and within each stratum a random sample is drawn. The purpose of stratification is the control of confounding variables (e.g., matching on number of years of schooling). The procedure cannot be taken to adequately correct for confounding variables when there is little or no overlap of the cultures (e.g., comparisons of literates and illiterates). The third procedure combines random or stratified sampling with the measurement of control variables. The procedure enables a statistical control of ambient variables (e.g., using an analysis of covariance).

Individual- and Country-Level Studies

The 1990s saw a growing interest in studies combining individual- and country-level data. Two kinds of studies have been reported. In multilevel hierarchical models (Raudenbush and Bryk, 2002), a regression model is used to explain individual variation (e.g., explaining pupils achievement scores on the basis of their intelligence) and class, district, or country variation (e.g., the explanation of the relationship between intelligence and achievement by means of school quality indicators). The second line of research involves the study of the same phenomena at different levels of aggregation (multilevel covariance structure analysis or multilevel factor analysis, Hox, 2002). From a conceptual point of view this line is more complicated, because it is well documented that (dis) aggregation can lead to methodological artifacts such as the ecological fallacy (*see* Ecological Fallacy, Statistics of), the incorrect application of culture-level characteristics to individuals. In each country a proportion of women are pregnant, but obviously, this proportion does not apply to any individual woman. In addition, equivalence issues have to be dealt with (Van de Vijver and Poortinga, 2002). Hofstede's (2001) famous study of values is based on country characteristics; the four dimensions he reported (individualism, power distance, uncertainty avoidance and masculinity–femininity) are characteristics of countries and their applicability to individual behavior cannot be taken for granted and has to be established. Multilevel covariance structure and factor analysis are powerful tools to examine the equivalence of phenomena at different levels of aggregation, answering questions such as the structural (in)equivalence of concepts like individualism–collectivism at individual and country level. For these and various other concepts the structural equivalence at different levels of aggregation is still unresolved.

See also: Psychometrics.

Bibliography

Angrosino, Michael V., 2006. Doing Cultural Anthropology: Projects for Ethnographic Data Collection. Waveland Press.

Barker, Chris, 2003. Cultural Studies: Theory and Practice. Sage.

Creswell, John W., 2013. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage.

Greenfield, Patricia M., et al., 2003. Cultural pathways through universal development. Annual review of psychology 54 (1), 461–490.

Hambleton, Ronald K., Merenda, Peter F., Spielberger, Charles D. (Eds.), 2004. Adapting Educational and Psychological Tests for Cross-cultural Assessment. Psychology Press.

Hofstede, Geert H., 2001. Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations across Nations. Sage.

Hox, Joop J., 2002. Multilevel Analysis: Techniques and Applications. Psychology Press.

Nisbett, Richard E., Miyamoto, Yuri, 2005. The influence of culture: holistic versus analytic perception. Trends in cognitive sciences 9 (10), 467–473.

Obinne, A.D.E., Amali, A.O., 2014. Differential item functioning: the implication for educational testing in Nigeria. International Review of Social Sciences & Humanities 7 (1).

Raudenbush, Stephen W., Bryk, Anthony S., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods, vol. 1. Sage.

Scheaffer, Richard, et al., 2011. Elementary Survey Sampling. Cengage Learning.

Sperber, Ami D., 2004. Translation and validation of study instruments for cross-cultural research. Gastroenterology 126, S124–S128.

Van de Vijver, Fons J.R., Poortinga, Ype H., 2002. Structural equivalence in multilevel research. Journal of Cross-cultural Psychology 33 (2), 141–156.

Yeh, Kuang-Hui, Bedford, Olwen, 2003. A test of the dual filial piety model. Asian Journal of Social Psychology 6 (3), 215–228.

Cross-Cultural Research Methods in Sociology

Fons van de Vijver, Tilburg University, Tilburg, The Netherlands; North-West University, Potchefstroom, South Africa; and University of Queensland, Brisbane, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

An overview is given of the major methodological issues in cross-cultural research. Different types of bias (at the level of the construct, method, and item) are described; a taxonomy of hierarchically ordered levels of equivalence (comparability) is presented that distinguishes between structural, measurement unit (metric), and score (scalar) equivalence. Sampling of cultures and subjects is discussed. Finally, issues in multilevel modeling are mentioned.

Introduction

Cross-cultural studies involve persons from different countries and/or ethnic groups. One of the central methodological problems of these studies is bias, the generic term for rival explanations of cross-cultural differences. For example, cross-cultural differences found for an inaccurately translated item do not reflect differences in the construct the item is supposed to measure; the cross-cultural comparison is then biased and scores cannot be compared across cultures. Three different types of bias are distinguished, depending on whether the source of interpretation problems derives from the construct, method of the study, or specific items (called construct, method, and item bias or differential item functioning, respectively). Equivalence, the second pivotal concept in cross-cultural research methods, refers to the implications of bias on the comparability of constructs and/or scores. Linguistic, structural, measurement unit, and full-score equivalence are described. Issues in test translation (such as translation, back-translation, and committee designs) are discussed. Common subject- and culture-sampling frames in cross-cultural research are mentioned. The article ends with a discussion of issues in combining individual- and country-level characteristics.

Cross-cultural studies involve persons from different countries and/or ethnic groups; a defining characteristic is their comparative nature. Most studies employ quantitative methods of data collection and analysis, which are also the topic of this article. Studies of cultural topics that are comparative and apply a qualitative methodology have been described by Ragin (1987). Combinations of quantitative and qualitative approaches are slowly emerging in mixed-methods research (Teddlie and Tashakkori, 2009). The range of instruments used in comparative studies is broad – encompassing, among other things, standardized psychological tests, observation schedules, and free interviews. In many studies, existing Western instruments (mental tests, survey questionnaires, and personality inventories) are administered in a new cultural context, either adapted or not to enhance their cultural appropriateness. The methodological framework outlined in this article applies to all these types of instruments. The aim of the framework is to achieve valid inferences about cross-cultural similarities and differences.

The Interpretation of Cross-Cultural Differences

If two persons from different cultural groups show different scores on a reliable and valid measure of subjective well-being, these score differences may refer to individual differences in subjective well-being. However, the score differences may also arise from differential social desirability, acquiescence, extremity scoring, inappropriate translation, or inadequacy of the item to measure well-being in both groups. The example illustrates a central problem in cross-cultural research: Observed score differences are often susceptible to multiple interpretations (Shadish et al., 2002). When the same instrument has been administered to persons from different cultural groups, it cannot be taken for granted that the same scores obtained in different cultural groups have the same psychological meaning.

The ambiguity of interpretation is a consequence of the methodological nature of culture as an independent variable. In laboratory studies, researchers randomly assign subjects to experimental treatments. The random assignment leads to a firm control of ambient variables; ideally, the experimental and control groups are matched on all outcome-relevant characteristics (e.g., personality characteristics and socioeconomic status), except for the treatment variable studied (*see* Internal Validity). However, like gender and other intrinsic subject characteristics, culture is not an experimental treatment that can be manipulated. Groups with a different cultural background tend to differ on a variety of outcome-relevant characteristics. These differences may constitute rival explanations of observed cross-cultural differences. Without precautions to rule out these rival explanations, observed cross-cultural differences are open to multiple interpretations. Findings in cross-cultural research are more convincing when rival explanations have been more adequately dealt with.

Concepts of Bias and Equivalence

Bias is the generic name of an important family of rival explanations. It refers to the common problem in the assessment of nonequivalent groups that scores obtained in different cultural groups are not an adequate reflection of the groups' standing on the construct underlying the instrument.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.44014-6

If scores are biased, their psychological meaning is group dependent and group differences in assessment outcome are to be accounted for, at least to some extent, by auxiliary psychological constructs or measurement artifacts. A closely related concept is equivalence, which refers to the absence of bias and, hence, to similarity of meaning across groups. The two concepts have somewhat different historical roots and areas of application. Whereas bias usually refers to nuisance factors, equivalence has become the generic term for metrical implications of bias.

Bias and equivalence are not intrinsic properties of an instrument but arise in a group comparison with a particular instrument. An instrument that works well in the comparison of two groups may not work well when a third country is involved. The literature on personality is dominated nowadays by the Big Five Model, according to which personality structure can be described by five traits: neuroticism (or its opposite, emotional stability), extraversion, openness, agreeableness, and conscientiousness. There is evidence from many cross-cultural studies that Big Five measures of personality show little bias (McCrae et al., 2005); however, this remarkable absence of bias in applications of studies with Big Five measures does not imply that applications of other measures of personality in the same countries would also be free of bias. Bias is more likely to affect comparisons of groups that differ in more test-relevant aspects.

Sources of Bias

There are three bias sources in cross-cultural research. The first is called construct bias; it occurs when the construct measured is not identical across groups. Ho's (1996) work on filial piety (psychological characteristics associated with being a good son or daughter) provides a good example. The Western conceptualization is narrower than the Chinese, according to which children are supposed to assume the role of caretaker of their parents when the latter grow old and become needy. Construct bias precludes the cross-cultural measurement of a construct with the same measure. An inventory of filial piety based on the Chinese conceptualization will cover aspects unrelated to the concept among Western subjects, while a Western-based inventory will leave important Chinese aspects uncovered.

An important source of bias, called method bias, can result from sample incomparability, instrument characteristics, tester and interviewer effects, and the method (mode) of administration. Examples are differential stimulus familiarity in mental testing and differential social desirability in personality and survey research. Some sources of method bias can be dealt with by careful preparation of the assessment instrument and its instruction manual (e.g., proper test instruction with a clear specification of what is asked from participants, standardization of administration, and adequate training of testers and interviewers). Yet, in practice it is often impossible to eliminate all outcome-relevant sample characteristics, particularly when the cultural distance of the countries involved is large. There are indications that a country's Gross National Product (per capita) is positively related to its mean score on mental tests (Brouwers et al., 2009) and negatively related to its mean score on acquiescence and social desirability (Van Hemert et al., 2002). Particularly in comparisons of culturally highly dissimilar groups, it may be hard or even impossible to eliminate the impact of sources of method bias such as source familiarity and social desirability.

Finally, bias can be due to anomalies at the item level (e.g., poor translations); this is called item bias or differential item functioning. According to a widely used definition, an item is biased if persons who have the same standing on the underlying construct (e.g., they are equally intelligent) but come from different cultural groups do not have the same expected score on the item. The expected item score is usually derived from the total test score. If a geography test administered to pupils in Poland and Japan contains the item 'What is the capital of Poland?' the test would be said to be biased in favor of Polish students, who can be expected to show higher scores on the item than Japanese students, even when pupils with the same total test score are compared. The item is biased as it favors one cultural group across all test score levels. Of all bias types, item bias has been the most extensively studied; various psychometric techniques are available to identify item bias (e.g., Osterlind and Everson, 2009). An overview of common ways of addressing bias is given in Table 1.

Types of Equivalence

Elaborating on categorizations in the literature, four different types of equivalence are proposed here (Matsumoto and Van de Vijver, 2011; Van de Vijver and Leung, 1997). The first type is labeled construct inequivalence. It amounts to comparing apples and oranges (e.g., the comparison of Chinese and Western filial piety, discussed in this article). Comparisons lack an attribute for comparison, also called *tertium comparationis* (the third term in the comparison). Such views can be found among authors who argue that human experiences are made up by and therefore bound to their cultural context; depression is incomparable across cultures as it entails experiences that are tied to a specific cultural and historical context. Proponents of such a view, according to whom psychological phenomena are fundamentally incomparable across cultures, would argue in favor of construct inequivalence. The second is called structural (or functional) equivalence. An instrument administered in different cultural groups shows structural equivalence if it measures the same construct in each group (e.g., Raven's Standard Progressive Matrices test has been found to measure intelligence in various cultural groups). Exploratory factor analyses followed by target rotations or confirmatory factor analysis of covariance matrices (structural equation modeling) may be applied to examine structural equivalence. Structural equivalence does not presuppose identity of measures across groups. The measures may use different stimuli across groups. If different operationalizations have been chosen, structural equivalence can be examined by comparing nomological networks across groups.

The third type of equivalence is called measurement unit or metric equivalence. Instruments show this type of equivalence if their measurement scales have the same units of measurement but a different origin (such as the Celsius and Kelvin scales in temperature measurement). This type of equivalence assumes interval- or ratio-level scores (with the same

Table 1 Strategies for identifying and dealing with bias in cross-cultural research

Type of bias	*Strategies*
Construct bias	Simultaneous development of an instrument in several cultures
Construct and/or method bias	Use of informants with expertise in local culture and language
	Use of samples of bilingual subjects
	Use of local surveys (e.g., content analyses of free-response questions)
	Cognitive interviewing
	Cross-cultural comparison of nomological networks
Method bias	Extensive training of administrators (e.g., increasing cultural sensitivity)
	Detailed manual or protocol for administration, scoring, and interpretation
	Detailed instructions (e.g., with a sufficient number of examples and/or exercises)
	Use of subject and context variables (e.g., educational background) that can be used for statistical control
	Use of collateral information (e.g., test-taking behavior or test attitudes)
	Assessment of response styles (e.g., social desirability or acquiescence)
	Use of test–retest, training, and/or intervention studies
Item bias (differential item functioning)	Judgmental methods
	Psychometric methods

Adapted from Van de Vijver, F.J.R., Tanzer, N.K., 2004. Bias and equivalence in cross-cultural assessment: an overview. European Review of Applied Psychology 54, 119–135.

measurement units in each culture). It applies when a bias factor with a fairly uniform influence on the items of an instrument affects test scores of different cultural groups in a differential way. Social desirability and stimulus familiarity may exert this influence. Observed group differences in scores are then a mixture of valid cross-cultural differences and measurement artifacts. When the relative contribution of both sources cannot be estimated, the interpretation of group comparisons of mean scores remains ambiguous. Multigroup comparisons of confirmatory factor analytic models have been utilized to examine measurement unit equivalence (Byrne, 2009). This technique, based on comparisons of covariance matrices, can identify bias sources that affect the covariance of items or tests, but it cannot differentiate between valid differences in mean scores and bias sources with a fairly uniform influence on all parts of an instrument (examples of such influences would be response styles like acquiescence or extremity scoring). Only in the case of scalar (or full-score) equivalence can direct comparisons be made; it is the only type of equivalence that allows for the conclusion that average scores obtained in two cultures are different. This type of equivalence assumes the same interval or ratio scales across groups. Conclusions about which of the latter two types of equivalence applies are often difficult to draw and subject to interpretation differences and controversy. For example, racial differences in intelligence test scores have been interpreted as due to valid differences (scalar equivalence) and as reflecting measurement artifacts (measurement unit equivalence). Scalar equivalence assumes that the role of bias can be safely neglected. The demonstration of scalar equivalence draws on inductive argumentation. Therefore, it is easier to disprove than to prove scalar equivalence. This can be made plausible by measuring presumably relevant sources of bias (such as stimulus familiarity or social desirability) and showing that they cannot statistically explain observed cross-cultural differences in a multiple regression or covariance analysis. The distinction between the latter two types of equivalence is immaterial when comparing experimental conditions or changes across cultures (e.g., developmental trajectories or training effects).

Structural, measurement unit, and scalar equivalence are hierarchically ordered. The third presupposes the second, which presupposes the first. Moreover, higher levels of equivalence are more difficult to establish. It is easier to demonstrate that an instrument measures the same construct in different cultural groups (structural equivalence) than to demonstrate numerical comparability across cultures (scalar equivalence). However, higher levels of equivalence allow for more detailed comparisons of scores (and, hence, constructs) across cultures. Whereas in the case of structural equivalence, only factor structures and nomological networks can be compared, scalar equivalence allows for more fine-grained analyses of cross-cultural similarities and differences, such as comparisons of mean scores across cultures in *t* tests and analyses of (co) variance.

In the last 10 years, it has become increasingly popular to apply structural equation modeling to investigate equivalence in so-called multigroup comparisons (Byrne, 2009). This advanced type of statistical modeling can be used to test different levels of equivalence, ranging from structural equivalence to scalar equivalence. The former is taken to be supported if factor loadings are identical across cultures. The latter is usually operationalized as identical intercepts in regression lines that link items to their construct. In all major software packages of structural equation modeling, multigroup options are available. On the one hand, these techniques are rigorous, have a rather firm statistical basis, and allow for a detailed test between different levels of equivalence; on the other hand, in these procedures it is easy to overlook sources of bias that were not measured. For example, if an instrument to assess extraversion is administered in different cultures, a multigroup comparison may provide a detailed view on sources of incomparability in items, but such an analysis will not reveal that important indicators of extraversion (such as specific manifestations of extraversion in that culture) may have been left out. An item about being the life of the party may be found to perform poorly in a particular culture, but such an analysis will not suggest alternatives, so multigroup modeling does not replace the value of a pilot test or focus groups to identify culture-relevant markers of constructs.

Linguistic Equivalence

The concept of linguistic equivalence is relevant in multilingual studies. Versions of an instrument in different languages show linguistic equivalence if these have the same characteristics that are relevant for the measurement outcome such as meaning, connotations of words and sentences, comprehensibility, and readability. Linguistic equivalence can be jeopardized by various sources, such as incorrect translations, words that are hard or impossible to translate (e.g., the English 'distress' does not have an equivalent in many languages), idiomatic expressions and metaphors (e.g., 'feeling blue'), and imprecise quantifiers ('rather often').

Because linguistic equivalence is not always guaranteed by a close translation, it has become increasingly popular to utilize adaptations. In an adaptation, parts are changed (instead of literally translated) with the aim to improve an instrument's suitability for a target group. Many studies employ a translation–back-translation procedure (Brislin, 1986). This amounts to a translation, followed by an independent back-translation and a comparison of the original and back-translated version, possibly followed by some alterations of the translation. Such a procedure provides a powerful tool to enhance the correspondence of original and translated versions that is independent of the researcher's knowledge of the target language. Yet, it also has some disadvantages. It puts a premium on literal reproduction; this may give rise to a stilted language use in the target version that lacks the readability and natural flow of the original. The problem may be compounded by translators' awareness of their involvement in a translation–back-translation procedure. Translators who know that the criterion for judging the quality of translation will be the equality of original and back-translated versions may focus more on compliance with this criterion and less on the quality of the text in the target language.

A second problem involves translatability. The use of idioms (e.g., the English 'feeling blue'), references to cultural specifics (e.g., references to country-specific public holidays), or other features that cannot be represented adequately in the target language challenges translation and back-translation designs (and, indeed, all procedures in which existing instruments are translated).

During the 1990s, there was a growing awareness that translations and adaptations require the combined expertise of social and behavioral scientists (with a competence in the construct studied) and experts in the target language(s) and culture(s). This awareness led to the use of new translation designs. An example is the so-called committee approach in which the expertise of all relevant disciplines is combined and in which there is usually no formal accuracy check of the translation. Combining different types of expertise that may be needed for a translation, such as linguistic and technical knowledge in the translation of technical documents, is relatively easy in a committee approach. Another approach is to commission multiple, independent translations of an instrument to a target language so that multiple versions can be compared and, if differences are found, the best can be chosen. An important characteristic of current large-scale cross-cultural studies is their awareness of the need for thorough quality checks of the translations, as poor translations can challenge the validity of cross-cultural studies. All kinds of procedures have been proposed to check the accuracy of translations. An example is Harkness's (2003) five-step TRAPD procedure, an acronym for translation, review, adjudication, pretesting, and documentation. These quality checks usually go much beyond verifying the accuracy of the translation by linguists and entail the administration of the new version in the target group and an evaluation of the psychometric properties of the instrument to ensure that the construct as measured is the construct as intended.

The judgmental evidence of the designs to enhance linguistic equivalence can be easily combined with statistical approaches to establish equivalence. Reports of multilingual studies often provide a combination of judgmental–linguistic and empirical–statistical evidence to demonstrate the adequacy of an instrument and its translation. There is a trade-off between the type of translation and the level of equivalence that can be obtained. When most or all questions of an instrument are adapted, structural equivalence is the highest level possible. When most or all items are literally translated, measurement unit and structural equivalence can be obtained. Recent statistical advancements in item response theory and structural equation modeling have made it possible to retain scalar equivalence, even when not all stimuli are literally translated (provided that all items measure the same underlying construct in each group). It is one of the achievements of recent decades in cross-cultural research methods that previously unlinked concepts (equivalence in a quantitative framework and in a linguistic framework) are now more linked in both the design and analysis stages of cross-cultural studies.

Sampling Cultures and Subjects

Cross-cultural studies can apply three types of schemes to sample cultures. Three types of sampling can be envisaged. The first is probability (or random) sampling. Because of the large cost of a probability sample from all existing cultures, it often amounts to stratified (random) sampling of specific cultures (e.g., Western cultures). The second and most frequently observed type of culture sampling is convenience sampling. The choice of cultures is governed here by availability and cost efficiency: researchers decide to form a research network, and all participants collect data in their own country. In the third type, called systematic sampling, the choice of cultures is more based on substantive considerations. A culture is deliberately chosen because of some characteristic.

In survey research, there is a well-developed theory of subject sampling (Kish, 1965). In the area of cross-cultural research, three types of sampling procedures of individuals are relevant as they represent different ways of dealing with confounding characteristics. The first is probability sampling. It consists of a random drawing from a list of eligible units such as persons or households. Confounding variables are not controlled for. The second type is stratified sampling. A population is stratified (e.g., in levels of schooling or socioeconomic status), and within each stratum a random sample is drawn. The purpose of stratification is the control of confounding variables (e.g., matching on number of years of schooling). The procedure cannot be taken to adequately

correct for confounding variables when there is little or no overlap of the cultures (e.g., comparisons of literates and illiterates). The third procedure combines random or stratified sampling with the measurement of control variables. The procedure enables a statistical control of ambient variables (e.g., using an analysis of covariance).

Individual- and Country-Level Studies

The 1990s saw a growing interest in studies combining individual- and country-level data. Two kinds of studies have been reported. In multilevel hierarchical models (Snijders and Bosker, 2011), a regression model is used to explain individual variation (e.g., explaining pupils' achievement scores on the basis of their intelligence) and class, district, or country variation (e.g., the explanation of the relationship between intelligence and achievement by means of school quality indicators). The second line of research involves the study of the same phenomena at different levels of aggregation (multilevel covariance structure analysis or multilevel factor analysis; Hox, 2010). From a conceptual point of view, this line is more complicated because it is well documented that (dis)aggregation can lead to methodological artifacts such as the ecological fallacy (*see* Ecological Fallacy, Statistics of), which is the incorrect application of culture-level characteristics to individuals. As an example, in each country a proportion of women is pregnant, but obviously, this proportion does not apply to any individual woman. In addition, equivalence issues have to be dealt with. Hofstede's (2001) famous study of values is based on country characteristics; the four dimensions he reported (individualism, power distance, uncertainty reduction, and masculinity–femininity) are characteristics of countries, and their applicability to individual behavior cannot be taken for granted and has to be established. Multilevel covariance structure and factor analysis are powerful tools to examine the equivalence of phenomena at different levels of aggregation, answering questions like the structural (in)equivalence of concepts like individualism–collectivism at individual and country levels. For these and various other concepts, the structural equivalence at different levels of aggregation is still unresolved.

See also: Psychometrics.

Bibliography

Brislin, R., 1986. The wording and translation of research instruments. In: Lonner, W., Berry, J. (Eds.), Field Methods in Cross-Cultural Research. Sage, Beverly Hills, pp. 137–164.

Brouwers, S., Van de Vijver, F., Van Hemert, D., 2009. Variation in Raven's progressive matrices scores across time and place. Learning and Individual Differences 19, 330–338.

Byrne, B., 2009. Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming. Erlbaum, Mahwah.

Harkness, J., 2003. Questionnaire translation. In: Harkness, J., Van de Vijver, F., Mohler, P. (Eds.), Cross-Cultural Survey Methods. Wiley, New York, pp. 35–56.

Ho, D., 1996. Filial piety and its psychological consequences. In: Bond, M. (Ed.), Handbook of Chinese Psychology. Oxford University Press, Hong Kong, pp. 155–165.

Hofstede, G., 2001. Culture's Consequences: International Differences in Work-Related Values. Sage, Beverly Hills.

Hox, J., 2010. Multilevel Analysis: Techniques and Applications, second ed. Routledge, New York.

Kish, L., 1965. Survey Sampling. Wiley, New York.

Matsumoto, D., Van de Vijver, F. (Eds.), 2011. Cross-Cultural Research Methods in Psychology. Cambridge University Press, New York.

McCrae, R.R., Terracciano, A., with 78 members of the Personality Profiles of Cultures Project, 2005. Universal features of personality traits from the observer's perspective: data from 50 cultures. Journal of Personality and Social Psychology 88, 547–561.

Osterlind, S., Everson, H., 2009. Differential Item Functioning, second ed. Sage, Newbury Park.

Ragin, C., 1987. The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies. University of California Press, Berkeley.

Shadish, W., Cook, T., Campbell, D., 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton-Mifflin, Boston.

Snijders, T., Bosker, R., 2011. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, second ed. Sage, London.

Teddlie, C., Tashakkori, A., 2009. Foundations of Mixed Methods Research: Integrating Quantitative and Qualitative Approaches in the Social and Behavioral Sciences. Sage, Thousand Oaks, CA.

Van de Vijver, F., Leung, K., 1997. Methods and Data Analysis for Cross-Cultural Research. Sage, Thousand Oaks, CA.

Van de Vijver, F.J.R., Tanzer, N.K., 2004. Bias and equivalence in cross-cultural assessment: An overview. European Review of Applied Psychology 54, 119–135.

Van Hemert, D., Van de Vijver, F., Poortinga, Y., Georgas, J., 2002. Structural and functional equivalence of the eysenck personality questionnaire within and between countries. Personality and Individual Differences 33, 1229–1249.

Cross-Cultural Study of Education

Akane Zusho and Karen E Clayton, Fordham University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Trommsdorff, P. Dasen, volume 5, pp. 3003–3007, © 2001, Elsevier Ltd.

Abstract

This article summarizes the contemporary research on cross-cultural study of education. Specifically, a summary of large-scale international assessments of academic achievement and teaching is presented, along with a cultural analysis of how home, school, and individual factors relate to academic performance.

An analysis of cross-cultural studies of education first necessitates an operational definition of culture. Culture is generally defined as a universal way of life, one that consists of people collectively using all resources in their environment to achieve; is learned, shared, and regulated by political, legal, and social systems; is socially transmitted; is represented by both external (observable behaviors) and internal (inferred traits) aspects of an individual; and is an abstraction of people's knowledge and beliefs about themselves, other people, and the world (Zusho and Clayton, 2011). In cross-cultural research, culture is often operationalized in terms of global categories such as nation of origin, social class, and race/ethnicity. Increasingly, however, researchers are recognizing that defining culture entirely in terms of these global categories promotes a limited, static approach to culture (Zusho and Clayton, 2011). Accordingly, more and more emphasis is being placed on capturing the heterogeneity within groups in cross-cultural research.

When broadly conceived, education can be equated with the interwoven processes of enculturation and socialization – as the cultural transmission of knowledge, skills, customs, and values from one generation to another. In the contemporary world, education is most often associated with schooling, both formal and informal. Cross-cultural studies of education therefore concerns the comparative study of the similarities and differences in schooling, including achievement, instruction, and beliefs about learning and instruction. The current focus within this area of research is an examination of how schooling is impacted and changed by sociocultural, ecological, and to a lesser extent, biological forces.

In general, cross-cultural studies of education are guided by three main goals (Berry et al., 2002). The first goal, the 'transport and test' function, concerns the testing of the generalizability of mainstream (typically Western, i.e., American, European, and Australian) theories to other cultures, for example, to test theories of motivation and learning across cultures. The second goal – the 'exploration and discovery' goal – in turn, entails exploring other cultures to discover variations absent in mainstream theories. An example of the second goal would be to explore other cultures to see if there are culture-specific motivations for learning. The second goal usually results from failure to find similarities across cultures as reflected in the transport and test function. Thus, the exploration and discovery goal largely aims to understand why theories and hypotheses fail to generalize in a particular culture by examining its ecological, historical, and cultural contexts. Finally, the third goal concerns the development of universal theories that is acceptable across cultural contexts via integration of the first two goals, for instance the development of universal theory of motivation.

International Assessments of Achievement

Although comparative studies of education can be traced back to the nineteenth century (Rust et al., 2009), cross-cultural studies of education did not begin to take center stage until the 1960s when the International Association for the Evaluation of Educational Achievement (IEA) conducted its first large-scale international assessment of educational achievement. Building on a previous pilot study, the IEA launched the First International Mathematics Study in 1964, which was subsequently followed up by the First International Science Study and the Study of Reading Comprehension in 1968. Since then, the IEA has commissioned numerous international assessments, including the Second and Third International Mathematics and Science Studies (SIMS, SISS, TIMSS). Current IEA studies of student achievement include assessments of computer and information literacy (International Computer and Information Literacy Study, 2013), mathematics and science (TIMSS, 2011, now known as Trends in International Mathematics and Science Study), reading literacy (Progress in International Reading Literacy Study (PIRLS), 2011), and civic and citizenship education (International Civic and Citizenship Education Study, 2009).

The Organisation for Economic Co-operation and Development (OECD) is also responsible for a number of international comparisons of education, including the Programme of International Student Achievement (PISA) and the Programme for the International Assessment of Adult Competencies (PIAAC). Since 2000, the OECD has administered PISA every 3 years with the intention of measuring how well member nations educate their youth in the areas of mathematics, reading, and science. Specifically, PISA assesses to what extent students can apply what they learned in school to real-life situations. PIAAC, in turn, is designed to assess the life skills of adults (aged 16–65 years), particularly those related to success in a twenty-first-century global economy.

Overall, these large-scale international assessments appear to be motivated by several interrelated goals. First is the goal to broaden knowledge of educational systems around the world

and to discern general educational trends across participating nations. A second, more pragmatic, goal is to identify characteristics of 'successful' educational systems (as indicated by consistent, superior performance on standardized educational assessments irrespective of socioeconomic background), so that it can be used as a model for others. It is worth mentioning that these studies generally assume that it is possible to reliably measure academic achievement across cultural contexts since it is possible to establish, at least partially, that students in different countries respond to achievement items in similar ways (i.e., metric invariance, see Steenkamp and Baumgartner, 1998). Indeed, items on TIMSS assessments are routinely scrutinized for whether or not they are biased toward or against individual countries (Martin et al., 2004). Another commonly held assumption is that rankings of nations on assessments will change over time, and can be used to encourage participating countries to improve their own teaching and student performance (OECD, 2010c).

Across the international assessments, several conclusions can be drawn. First, students from East Asian countries (e.g., Korea, Shanghai-China, Hong Kong, and Singapore) typically achieve the highest scores on tests of mathematics and science (Martin et al., 2008; Mullis et al., 2008; OECD, 2010c). Second, girls, on average, score higher than boys in reading (Mullis et al., 2007; OECD, 2010c), whereas males outperform females in mathematics by the time they are 15 years of age (OECD, 2010c). Third, although achievement rankings may not be tied to a nation's overall wealth (OECD, 2010c), there appears to be a positive relationship between students' socioeconomic background and achievement scores. For example, students who attend schools with a higher proportion of economically disadvantaged students generally do more poorly than students who attend schools with a lower proportion of students who come from economically disadvantaged backgrounds (Martin et al., 2008; Mullis et al., 2008; OECD, 2010c). Finally, there is a general trend toward improvement in achievement scores among participating nations over time (Martin et al., 2008; Mullis et al., 2008; OECD, 2010c). Specifically, there appears to be a narrowing of the gap in achievement scores among students who score at the top and bottom (OECD, 2010c).

Support for Learning: Schools and Teachers

One major focus of cross-cultural studies of education is schools and teachers. Specifically, these studies generally explore how and to what extent do schools and teachers provide all students the opportunity to learn. To this end, studies have been conducted on various aspects of schooling and instruction, including curricular content, quality of instruction, as well as other teacher-related variables (e.g., teacher beliefs and preparation of teachers).

In general, studies of instructional time and curricular content reveal more similarities than differences across countries. On average, studies report that teachers spend approximately 23–24 h a week in classroom instruction (Mullis et al., 2008; OECD, 2004). Most of the instructional time at the younger grades is, not surprisingly, focused on reading and language instruction (Mullis et al., 2007), followed by mathematics and then science (Martin et al., 2008; Mullis et al., 2008). In addition, most countries report teaching the topics presented on international assessments, indicating that there are commonalities in curriculum content across participating nations (Martin et al., 2008; Mullis et al., 2008). For science, however, there appears to be one major difference; countries vary in whether they teach science as a single integrated subject, or as separate subjects. As for specific instructional practices, teachers in most countries report spending the most instructional time on structuring practices, such as stating learning goals, summarizing and reviewing previous lessons and homework, and checking for student understanding (OECD, 2009).

Although the content of curriculum does not appear to vary considerably across countries, there is some emerging evidence to suggest that cultural differences may exist in curricular *coherence* – that is the extent to which the sequence of topics presented in the curriculum is logical, progressive, and corresponds to the natural structure of the discipline from which it is drawn (Schmidt et al., 2005). Schmidt et al. (2005) argue that countries that perform at the top of the TIMSS comparisons generally have mathematics and science curricula that are coherent and promote conceptual understanding. They further argue that the US math and science curricula are generally organized in an ad hoc fashion, and attempt to cover almost every topic at every grade, resulting in a 'mile-wide inch-deep curriculum.'

There is, too, limited evidence that the quality of instruction may vary culturally (Tatto et al., 2012). For example, the results of the TIMSS video case studies (Stigler et al., 1999; Hiebert et al., 2003) suggest that the mathematics instruction in certain countries may be of higher quality than others. The 1999 study, which examined eighth-grade mathematics lessons from six countries (Australia, Czech Republic, Hong Kong Special Administrative Region, the Netherlands, Switzerland, and the United States), rated the quality of instruction to be highest in Hong Kong and lowest in the United States. The OECD studies further suggest that disciplinary climate, as measured by time on task, frequency of classroom distractions, teacher control, and other indices of classroom management, is an important predictor of achievement test scores (OECD, 2010a).

Recent studies such as the OECD's Teaching and Learning International Survey (TALIS) (OECD, 2009) and the IEA's Teacher Education and Development Study in Mathematics (TEDS-M) focus on the working conditions and preparation of teachers, respectively. TALIS had four major foci: teachers' professional development (PD); teachers' teaching practices, beliefs, and attitudes; teacher appraisal and feedback; and school leadership. Overall, TALIS findings indicate that most teachers are satisfied with their jobs and feel that they make a considerable impact on their students. However, despite varying levels in the intensity and level of participation in PD, most teachers, on average, report that their PD needs are generally not being met; for example, more than half of the teachers surveyed reported wanting more PD that they had received, specifically related to teaching students with special needs (OECD, 2009). TALIS results also indicate that with the exception of Italy, teachers, particularly those in Austria, Australia, Belgium, Denmark, Estonia, and Iceland, generally perceive that their role is more to support students'

construction of knowledge rather than to simply present facts and giving students opportunities to practice.

The results of TEDS-M, in turn, indicate that there is variability in how countries prepare their teachers (Tatto et al., 2012). First, teacher education systems vary in terms of teacher selectivity and status; for example, more stringent admissions criteria are observed in countries with an oversupply of potential candidates (e.g., Chinese Taipei). The general background of teacher candidates, therefore, varies across nations – for example, the proportion of teacher candidates who possess advanced degrees in mathematics and consider themselves to be mathematics specialists are much higher in countries where teaching is a more popular profession. Teacher education systems also differ in terms of the length of program, the amount of class time allocated to mathematics and mathematics pedagogy, as well as field experiences. The results of TEDS-M also provide initial evidence that variability within and across teacher education programs are partly attributable to differences in mathematics content knowledge, teacher beliefs, and general opportunities to learn (Tatto et al., 2012).

In short, the collective studies on teaching and schooling suggest that a combination of factors related to teaching and schooling are likely to be related to higher student performance. The OECD (2010b) characterizes 'successful' school systems as those who provide *all* students equal opportunities to learn; in other words, school systems with limited tracking where economically advantaged and disadvantaged students attend the same schools, and schools with low transfer rates related to poor academic performance or behaviors. Successful school systems also generally provide greater autonomy and resources to individual schools.

Support for Learning: Home and Family

Research indicates that it is not only teachers and schools that have an effect on students' performance. Indeed, mounting evidence suggests that students' home background may make an equal, if not more, significant impact on students' learning outcomes (OECD, 2003). Thus, increasing attention is being paid to investigating how various home factors relate to learning. For example, the results of the TIMSS studies largely indicate that students who achieve at the top of international comparisons speak the language of the test at home, have parents with relatively higher levels of education, and also have access to computers and books at home (Martin et al., 2008; Mullis et al., 2008).

Perhaps the greatest evidence of the impact of home factors on achievement comes from studies of reading. For example, PISA data consistently demonstrate a positive relationship between socioeconomic status and literacy performance (OECD, 2003). It appears that the impact of advantaged family background is present in every country, but particularly pronounced in Latin American countries such as Peru and Argentina. Results of PIRLS 2006 continue to demonstrate a positive relationship between students' fourth-grade reading achievement and the extent to which parents engage their children in early literacy activities (e.g., reading books, telling stores, singing songs, and playing with alphabet toys and other word games) before starting school (Mullis et al., 2007). Students who come from homes with more than 100 books were also found to achieve, on average, close to 1 standard deviation higher on the PIRLS reading assessment than children who come from homes with 10 or fewer books. Correspondingly, the average reading achievement was highest among students who had parents who were frequent readers and who had favorable attitudes toward reading (Mullis et al., 2007).

Taken together, these findings demonstrate that the home is an important context for socialization and that family characteristics such as wealth, parental education, and access to and support of early literacy activities are a major source of disparity in students' reading achievement.

Support for Learning: Individual Attitudes and Beliefs about Learning

Apart from school- and home-related contextual factors, there is ample evidence that students' personal attitudes and beliefs about learning also play an important role in explaining learning outcomes. To date, most of this research has focused on examining cross-cultural differences in students' motivational beliefs, as well as learning strategies (Zusho and Clayton, 2011; McInerney, 2011). Briefly, motivational beliefs are usually composed of students' interest and enjoyment in school-related matters, learning goals, beliefs about identity (self-concept) and ability (self-efficacy), and attributions for success and failure, specifically the role of effort and ability. Similarly students' learning strategies are usually characterized as being deep or surface level (Biggs, 1987). Deeper processing learning strategies promote conceptual understanding and typically involve the use of elaborative cognitive and metacognitive strategies such as making connections between new material and prior knowledge, or regulating and monitoring one's understanding of material during study (Pintrich and DeGroot, 1990; Zimmerman and Martinez-Pons, 1986; Zimmerman, 1990). In contrast, surface-level learning strategies generally focus more on task completion, and rote memorization.

Based on this emerging area of study, certain generalizations can be made about the relationship of motivational beliefs, learning strategies, and academic performance across cultures. First, research on both Eastern and Western students suggests that students who display more adaptive motivational profiles (i.e., higher levels of intrinsic motivation, focus on goals focused on learning and understanding) generally employ more deeper processing learning strategies than those students who display more maladaptive profiles (Lepper et al., 2005; Schmuck et al., 2000). Second, research confirms that deep learning strategies are positively related to academic achievement across cultures (Watkins, 2000, 2001). Moreover, results from OECD (2010d) found that students who enjoy reading, are highly engaged in the learning process, monitor and evaluate their learning, and are aware of what strategies are effective for learning tend to perform better in the PISA reading assessment.

Such generalizations notwithstanding, studies on Eastern and Western countries have also identified cross-cultural

differences in the relations of certain motivational constructs and learning outcomes. For example, in Western countries, empirical findings typically find student self-perceptions (e.g., self-concept, self-efficacy) to be a strong predictor of individuals' effort, cognitive engagement, academic achievement, and subjective well-being (Bandura, 1997; Linnenbrink and Pintrich, 2003). Paradoxically, however, it is typically found that in Eastern countries, lower levels of self-concept or self-efficacy are oftentimes a better predictor of academic performance (Klassen, 2004a,b; Leung, 2001; Mau, 2000).

Similarly, differences have been noted in the attributional styles of Eastern and Western students. For example, a number of studies report Eastern students to attribute academic success more to effort than to ability, in comparison to Western students who make more ability-based attributions for success (e.g., Ho et al., 1999; Holloway, 1988). It is important to note, however, that it may be somewhat premature to make cross-cultural generalizations about attributional style. For example, Bempechat and colleagues caution that Asian students do not always emphasize effort over ability, and that ability attributions may play a more significant role in explaining Asian and Asian-American students' academic performance (Bempechat and Drago-Severson, 1999; Choi et al., 1994). Western students, too, appear to emphasize effort over ability at times; British children in particular have been noted to attribute success in school to effort rather than ability (Gipps and Tunstall, 1998; Lightbody et al., 1996). Such findings suggest that there is within-group variability in attributional beliefs.

Similarly, the predictive ability of effort attributions to academic achievement has been noted to vary cross-culturally and to be inconsistent with theoretical expectations that effort attributions can sustain academic performance. For example, Eaton and Dembo (1997) found effort attributions to predict higher levels of achievement, but somewhat unexpectedly, this pattern was only found among their non-Asian participants. Elliott et al. (1999) also found English and American students to report effort to be a major determinant of their academic success; it was noted, however, that these students' commitments to educational endeavors were generally less strong than those of their Russian counterparts. Furthermore, Bempechat and her colleagues generally suggest ability attributions to be a better predictor of academic achievement than effort attributions (Bempechat and Drago-Severson, 1999; Bempechat et al., 1996), pointing to empirical findings that demonstrate that effort attributions are typically either negatively related or unrelated to academic achievement (Holloway, 1988; Marsh, 1984).

As for cross-cultural differences in learning strategies, some studies suggest Asian students to focus more on rote memorization or rehearsal strategies during study (e.g., Purdie and Hattie, 1996), and that use of these strategies does not seem to negatively impact their achievement. However, other studies have found that Asian students do not use memorization strategies more frequently than non-Asian students (Chiu et al., 2007; Neber et al., 2008; OECD, 2010d). For example, Chiu et al. (2007) found that students from Asian countries who reported greater use of learning strategies focused on memorization performed worse academically. Thus, the empirical findings on cultural differences in learning strategies are somewhat mixed. Nevertheless, there continues to be speculation about whether Asian students' conceptions of memorization differ from those of Western students. In particular, it has been suggested that Asian students conceptualize memorization and understanding as interlocking processes that are both important for effective learning outcomes (Kember, 1996; Watkins, 2000), whereas Western students are more likely to view understanding in terms of sudden insight (Dahlin and Watkins, 2000).

In short, this body of work suggests that there may be both universal and culture-specific aspects to students' attitudes and beliefs about learning. However, it is important to note that given the subjective nature of individual attitudes and beliefs about learning, precise measurement is much more difficult to achieve cross-culturally, which often compromises comparisons and interpretation of results. Thus, interpretations of cultural differences in beliefs and attitudes should be made with caution.

For example, it has been noted that the meaning of motivation may vary culturally (Markus and Kitayama, 1991; Zusho, 2008; Zusho and Njoku, 2007). Explanations of differences in meaning typically relate to research on individualism and collectivism, defined, respectively, in terms of autonomous and interdependent self-views, as well as the research on the value placed on personal goals versus group goals, and the emphasis on conforming to communal norms and duties (Oyserman et al., 2002; Triandis, 2001). There is, for instance, evidence to suggest that students from collectivist societies define academic success more in terms of affiliative goals, such as maintenance of good family and social relationships (Chan et al., 2005; Salili, 2009). As a result, it is assumed that individuals from collectivistic societies may experience difficulties separating individual goals from socially based goals (e.g., Bernardo, 2008; Chang and Wong, 2008), and may consequently interpret survey items in a different manner than individuals from individualistic societies.

There is also emerging evidence to suggest that there may be cross-national differences in how individuals respond to Likert scale survey items more generally (Harzing et al., 2012). Specifically, it has been found that East Asian individuals tend to select responses toward the middle of the scale, in comparison to US and Canadian respondents who are more likely to select responses at the extreme ends of the scale (Harzing et al., 2012). Furthermore, unlike the aforementioned large-scale studies of achievement, the majority of cross-cultural studies on attitudes and beliefs employ nonstandardized assessments with compromised reliability and construct validity (McInerney, 2011). Taken together, such findings call into question the construct equivalence of measures used in cross-cultural studies of attitudes and beliefs, further complicating the interpretation of results.

General Trends and Issues

The collective research to date on cross-cultural studies of education has uncovered some important findings about

educational processes and the socialization practices that support them. In general, the findings suggest that academic performance can be explained, in part, by coherent and high-quality instructional practices that promote conceptual understanding. Similarly, home factors such as support of early literacy have been demonstrated to play an important role in students' academic outcomes. Finally, a focus on adaptive motivational beliefs and deeper processing strategies has been found to lead to higher levels of academic performance. Such findings also fall in line with Hattie's (2009) meta-analytic studies, which find that the largest effect sizes associated with achievement are related to quality of teaching, particularly an emphasis on instructional processes that support metacognition and self-regulated learning. Cross-cultural studies of education have also been useful in demonstrating how variation in cultural practices and beliefs impact educational outcomes. Thus, the cross-cultural approach to the study of education has been proven to be advantageous in highlighting both universal (etic) and culture-specific (emic) aspects of education across cultures.

Such findings, notwithstanding, the cross-cultural study of education are a fairly new area of inquiry. As such, there is much that can be improved, both theoretically and methodologically.

Theoretically, Berry et al. (2002) propose universalism to be the most appropriate metatheoretical tradition for the cross-cultural study of differences. Universalism assumes that there are some basic processes that are species-wide and therefore, universal. However, it recognizes the importance of examining larger contextual and societal forces in explaining variation in human behavior, thoughts, and attitudes. Universalism also underscores the importance of reducing ethnocentric methodologies and cultural biases whenever possible, a task that is acknowledged to be difficult but not necessarily impossible (Zusho and Clayton, 2011).

Methodologically, then, one of the greatest challenges facing a universalist approach to the study of education is the operationalization of culture. Many of the cross-cultural studies, in particular those that focus on individual beliefs and attitudes, operationalize culture in terms of race/ethnicity/nation of origin, which can inadvertently promote stereotypes about cultural groups by prioritizing between-group differences over within-group differences. Thus it is important for cross-cultural studies of education to place equal emphasis on investigating both between and within-group differences. For example, the findings of TALIS survey indicate that within-country variation in the intensity of PD can be quite high, and it is greatest in Italy, Korea, Mexico, Poland, and Spain (OECD, 2009). Muthén (1989), too, cautions against the homogeneity assumption – that is assuming that data are obtained from a single population – particularly in studies of education where students often are exposed to varying curricula and instructional approaches, sometimes even within a single school.

It is also clear that the preponderance of research on culture and education to date has selectively focused on restricted populations. For example, Heinrich et al. (2010) highlight that most of the research in cross-cultural psychology focus on individuals living in Western, educated, industrialized, rich, and democratic (W.E.I.R.D) societies, and thus claims about universality of psychological processes may be at best premature, and at worst, simply wrong. Similar claims can be made about the cross-research in education, particularly the studies on individual attitudes and beliefs. Comparisons between Western (typically US or European) and Eastern cultures continue to dominate the field, and smaller scale studies also often rely on convenience sampling (Zusho and Clayton, 2011).

Finally, cross-cultural comparisons are meaningful only to the extent that constructs can be measured reliability and validly across cultures. The various large-scale international comparisons of achievement suggest that it is possible to do this, for the most part, for academic achievement. Indeed, systematic procedures have been put into place to validate assessments of academic achievement. Typically, test questions are developed collaboratively with representatives from participating nations, and field tested prior to use. In addition, the items undergo an extensive translation verification process to ensure that the items have the same meaning across cultures. Procedures for administration of the assessments are also standardized across the various countries. Once the data are collected, care is taken to establish both structural and functional equivalence of the instrument across cultures through factor analytic techniques and by demonstrating that the instrument displays comparable relations with related measures in the hypothesized direction across cultures (Olson et al., 2008).

Such procedures are clearly labor intensive, timely, and costly. Thus, it is hardly surprising that such level of care is not usually seen in smaller scale studies, particularly those that investigate cultural differences in beliefs and attitudes. It is therefore important for future research to continue to investigate and improve the cultural equivalence of its measures before use in comparative studies. In particular, there is great need to standardize the measurement of attitudinal measures such as those related to motivation and learning strategies. It is not uncommon for cultural studies of motivation and learning strategies to employ a variety of instruments, which vary not only in its theoretical orientation but also in the degree of cultural bias inherent in the items. Thus, much of the cross-cultural research is related to differences in students' motivational beliefs and use of learning strategies remain largely uninterpretable.

See also: Academic Achievement Motivation, Development of; Academic Self-Concept and Achievement; Comparative Research in Education: Iea Studies; Content Validity; Cross-National Comparisons in Education: Findings from Pisa; Illiteracy, Sociology of; Literacy Education; Motivation and Actions, Psychology of; Motivation, Learning, and Instruction; Motivational Development, Systems Theory of; Race and Academic Motivation; Reliability and Validity Assessment: New Approaches; School Achievement: Motivational Determinants and Processes; Schooling: Impact on Cognitive and Motivational Development.

Bibliography

Bandura, A., 1997. Self-efficacy: The Exercise of Control. Freeman, New York, NY.

Bempechat, J., Drago-Severson, E., 1999. Cross-national differences in academic achievement: beyond etic conceptions of children's understanding. American Educational Research Association 69, 287–314. http://dx.doi.org/10.2307/1170541.

Bempechat, J., Nakkula, M., Wu, J., Ginsburg, H., 1996. Attributions as predictors of math achievement: a comparative study. Journal of Research and Development in Education 29, 53–59.

Bernardo, A.B., 2008. Individual and social dimensions of Filipino students' achievement goals. International Journal of Psychology 43, 886–891. http://dx.doi.org/10.1080/00207590701837834.

Berry, J.W., Poortinga, Y.H., Segall, M.H., Dasen, P.R., 2002. Cross-cultural Psychology: Research and Applications, second ed. Cambridge University Press, Cambridge, MA.

Biggs, J.B., 1987. Student Approaches to Learning and Studying. Australian Council for Educational Research, Melbourne, Australia.

Chan, K., Lai, P., Leung, M., Moore, P.J., 2005. Students' goal orientations, study strategies and achievement: a closer look in Hong Kong Chinese cultural context. The Asian-Pacific Education Researcher 14, 1–26.

Chang, W.C., Wong, K., 2008. Socially oriented achievement goals of Chinese university students in Singapore: structure and relationships with achievement motives, goals, and affective outcomes. International Journal of Psychology 43, 880–885. http://dx.doi.org/10.1080/00207590701836398.

Chiu, M.M., Chow, B.W.-Y., McBride-Chang, C., 2007. Universals and specifics in learning strategies: explaining adolescent mathematics, science, and reading achievement across 34 countries. Learning and Individual Differences 17, 63–76.

Choi, E., Bempechat, J., Ginsburg, H., 1994. Educational socialization in Korean-American children: a longitudinal study. Journal of Applied Developmental Psychology 15, 313–318. http://dx.doi.org/10.1016/0193-3973(94)90033-7.

Creed, P.A., Patton, W., Watson, M.B., 2002. Cross-cultural equivalence of the career decision making self-efficacy scale – short form: an Australia and South African comparison. Journal of Career Assessment 10, 327–342.

Dahlin, B., Watkins, D.A., 2000. The role of repetition in the process of memorizing and understanding: a comparison of German and Chinese secondary school students in Hong Kong. British Journal of Educational Psychology 70, 65–84. http://dx.doi.org/10.1348/000709900157976.

Eaton, M.J., Dembo, M.H., 1997. Differences in the motivational beliefs of Asian American and non-Asian students. Journal of Educational Psychology 89, 433–440. http://dx.doi.org/10.1037/0022-0663.89.3.433.

Elliott, J., Hufton, N., Hildreth, A., 1999. Factors influencing educational motivation: a study of attitudes, expectations and behavior in Sutherland, Kentucky and St. Petersburg. British Educational Research Journal 25, 75–94.

Gipps, C., Tunstall, P., 1998. Effort, ability and the teacher: young children's explanations for success and failures. Oxford Review of Education 24, 149–165.

Hattie, J.A.C., 2009. Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement. Routledge, Abingdon, Oxon.

Harzing, A., Brown, M., Koster, K., Zhao, S., 2012. Response style differences in cross-national research. Management International Review 52, 341–363. http://dx.doi.org/10.1007/s11575-011-0111-2.

Heinrich, J., Heine, S.J., Norenzayan, A., 2010. Most people are not WEIRD. Nature 466, 29.

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K.B., Hollingsworth, H., Jacobs, J., Chui, A.M., Wearne, D., Smith, M., Kersting, N., Manaster, A., Tseng, E., Etterbeek, W., Manaster, C., Gonzales, P., Stigler, J., 2003. Teaching Mathematics in Seven Countries: Results from the TIMSS 1999 Video Study. NCES (2003-013), U.S. Department of Education. National Center for Education Statistics, Washington, DC. Available online at: http://timssvideo.com/sites/default/files/TIMSS%201999%20Math%20Report.pdf.

Ho, I.T., Salili, F., Biggs, J.B., Hau, K., 1999. The relations among causal attributions, learning strategies and level of achievement: a Hong Kong Chinese study. Asia Pacific Journal of Education 19, 44–58.

Holloway, S.D., 1988. Concepts of ability and effort in Japan and the United States. Review of Educational Research 58, 327–345.

Kember, D., 1996. The intention to both memorize and understand: another approach to learning? Higher Education 31, 341–354.

Klassen, R.M., 2004a. A cross-cultural investigation of the efficacy beliefs of South Asian immigrant and Anglo Canadian nonimmigrant early adolescents. Journal of Educational Psychology 96, 731–742. http://dx.doi.org/10.1037/0022-0663.96.4.731.

Klassen, R.M., 2004b. Optimism and realism: a review of self-efficacy from a cross-cultural perspective. International Journal of Psychology 39, 205–230.

Law, Y.-K., Chan, C.K.K., Sachs, J., 2008. Beliefs about learning, self-regulated strategies and text comprehension among Chinese children. British Journal of Educational Psychology 78, 51–73.

Lepper, M.R., Corpus, J.H., Iyengar, S.S., 2005. Intrinsic and extrinsic motivational orientations in the classroom: age differences and academic correlates. Journal of Educational Psychology 97, 184–196. http://dx.doi.org/10.1037/0022-0663.97.2.184.

Leung, C., 2001. The psychological adaptation of overseas and migrant students in Australia. International Journal of Psychology 36, 251–259. http://dx.doi.org/10.1080/00207590143000018.

Lightbody, P., Siann, G., Stocks, R., Walsh, D., 1996. Motivation and attribution at secondary school: the role of gender. Educational Studies 22, 13–25. http://dx.doi.org/10.1080/0305569960220102.

Linnenbrink, E.A., Pintrich, P.R., 2003. The role of self-efficacy beliefs in student engagement and learning in the classroom. Reading & Writing Quarterly 19, 119–137. http://dx.doi.org/10.1080/10573560308223.

Luszczynska, A., Gutierrez-Dona, B., Schwarzer, R., 2005. General self-efficacy in various domains of human functioning: evidence from five countries. International Journal of Psychology 40, 80–89. http://dx.doi.org/10.1080/00207590444000041.

Markus, H.R., Kitayama, S.K., 1991. Culture and the self: implications for cognition, emotion, and motivation. Psychological Review 98, 224–253. http://dx.doi.org/10.1037/0033-295X.98.2.224.

Marsh, H., 1984. Relations among dimensions of self-attribution, dimensions of self-concept, and academic achievement. Journal of Educational Psychology 76, 1291–1308. http://dx.doi.org/10.1037/0022-0663.76.6.1291.

Martin, M.O., Mullis, I.V.S., Chrostowski, S.J. (Eds.), 2004. TIMSS 2003 Technical Report. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA. Available online at: http://timssandpirls.bc.edu/timss2003i/technicalD.html.

Martin, M.O., Mullis, I.V.S., Foy, P. (with Olson, J.F., Erberber, E., Preuschoff, C., Galia, J.), 2008. TIMSS 2007 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA. Available online at: http://timssandpirls.bc.edu/TIMSS2007/PDF/TIMSS2007_InternationalScienceReport.pdf.

Mau, W., 2000. Cultural differences in career decision-making styles and self-efficacy. Journal of Vocational Behavior 57, 365–378. http://dx.doi.org/10.1006/jvbe.1999.1745.

McInerney, D., 2011. Cultural influences on self-regulation. In: Zimmerman, B., Schunk, D. (Eds.), Handbook of Self-Regulation of Learning and Performance. Routledge, New York, pp. 442–464.

Mullis, I.V.S., Martin, M.O., Foy, P. (with Olson, J.F., Preuschoff, C., Erberber, E., Arora, A., Galia, J.), 2008. TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA. Available online at: http://timssandpirls.bc.edu/TIMSS2007/PDF/TIMSS2007_InternationalMathematicsReport.pdf.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., Foy, P., 2007. PIRLS 2006 International Report: IEA's Progress in International Reading Literacy Study in Primary Schools in 40 Countries. Boston College, Chestnut Hill, MA. Available online at: http://timssandpirls.bc.edu/PDF/PIRLS2006_international_report.pdf.

Muthén, B.O., 1989. Latent variable modeling in heterogeneous populations. Psychometrika 54, 556–585.

Neber, H., He, J., Liu, B.-X., Schofield, N., 2008. Chinese high-school students in physics classroom as active, self-regulated learners: cognitive, motivational and environmental aspects. International Journal of Science and Mathematics Education 6, 769–788.

OECD, 2003. Literacy Skills for the World of Tomorrow: Further Results from PISA 2000. Available online at: http://www.oecd.org/edu/preschoolandschool/programmeforinternationalstudentassessmentpisa/33690591.pdf.

OECD, 2004. Learning for Tomorrow's World – First Results from PISA 2003. Available online at: http://www.oecd.org/edu/preschoolandschool/programmeforinternationalstudentassessmentpisa/34002216.pdf.

OECD, 2009. Creating Effective Teaching and Learning Environments: First Results from TALIS. Available online at: http://www.oecd.org/edu/preschoolandschool/43023606.pdf.

OECD, 2010a. Mathematics Teaching and Learning Strategies in PISA. Available online at: http://www.oecd.org/pisa/pisaproducts/pisa2003/mathematicsteachingandlearningstrategiesinpisa.htm.

OECD, 2010b. PISA 2009 Results: What Makes a School Successful? Resources, Policies and Practices, vol. IV. Available online at: http://www.oecd.org/pisa/pisaproducts/48852721.pdf.

OECD, 2010c. PISA 2009 Results: What Students Know and Can Do - Student Performance in Reading, Mathematics, and Science, vol. 1. Available online at: http://www.oecd.org/pisa/pisaproducts/48852548.pdf.

OECD, 2010d. PISA 2009 Results: Learning to Learn: Student Engagement, Strategies and Practices, vol. 111. Available online at: http://www.oecd.org/pisa/pisaproducts/48852630.pdf.

Olson, J.F., Martin, M.O., Mullis, I. (Eds.), 2008. TIMSS 2007 Technical Report. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA. Available online at: http://timss.bc.edu/timss2007/techreport.html.

Oyserman, D., Coon, H.M., Kemmelmeier, M., 2002. Rethinking individualism and collectivism: evaluation of theoretical assumptions and meta-analyses. Psychological Bulletin 128, 3–72. http://dx.doi.org/10.1037/0033-2909.128.1.3.

Pintrich, P.R., DeGroot, E.V., 1990. Motivational and self-regulated learning components of classroom academic performance. Journal of Educational Psychology 82, 33–40. http://dx.doi.org/10.1037/0022-0663.82.1.33.

Purdie, N., Hattie, J., 1996. Cultural differences in the use of strategies for self-regulated learning. American Educational Research Journal 33 (4), 845–871.

Rao, N., Moely, B.E., Sachs, J., 2000. Motivational beliefs, study strategies, and mathematics attainment in high-and low-achieving Chinese secondary school students. Contemporary Educational Psychology 25, 287–316.

Rust, V., Johnstone, B., Allaf, C., 2009. Reflection on the development of comparative education. In: Kazamias, A., Cowan, R. (Eds.), International Handbook of Comparative Education. Springer, London, pp. 121–138.

Salili, F., 2009. A model of culture and achievement behavior. In: Kaplan, A., Karabenick, S., De Groot, E. (Eds.), Culture, Self, and Motivation: Essays in Honor of Martin L. Maehr. Information Age Publishing, Charlotte, NC, pp. 183–212.

Schmidt, W., Wang, H.C., McKnight, C.C., 2005. Curriculum coherence: an examination of U.S. mathematics and science content standards from an international perspective. Journal of Curriculum Studies 37, 525–559.

Schmuck, P., Kasser, T., Ryan, R.M., 2000. Intrinsic and extrinsic goals: their structure and relationship to well-being in German and U.S. college students. Social Indicators Research 50, 225–241. http://dx.doi.org/10.1023/A:1007084005278.

Steenkamp, J.E.M., Baumgartner, H., 1998. Assessing measurement invariance in cross-national consumer research. Journal of Consumer Research 25, 78–107.

Stigler, J.W., Gonzales, P., Kawanaka, T., Knol, S., Serrano, A., 1999. The TIMSS Videotape Classroom Study: Methods and Findings from an Exploratory Research Project on Eighth-Grade Mathematics Instruction in Germany, Japan, and the United States, NCES 99-074. US Department of Education. National Center for Education Statistics, Washington, DC.

Tatto, M.T., Schwille, J., Senk, S.L., Ingvarson, L., Rowley, G., Peck, R., Bankov, K., Rodriguez, Reckase, M., (with Dumais, J., Carstens, R., Brese, F., Meinck, S., Berzina-Pitcher, I., Lu, Y., Holdgreve-Resendez, R.), 2012. Policy, Practice, and Readiness to Teach Primary and Secondary Mathematics in 17 Countries: Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-m). IEA, Amsterdam, NL. Available online at: http://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/TEDS-M_International_Report.pdf.

Triandis, H.C., 2001. Individualism and collectivism. Past, present, and future. In: Matsumoto, D. (Ed.), The Handbook of Culture and Psychology. Oxford University Press, New York, NY, pp. 35–50.

Watkins, D.A., 2000. Learning and teaching: a cross-cultural perspective. School Leadership & Management 20, 161–173. http://dx.doi.org/10.1080/13632430050011407.

Watkins, D.A., 2001. Correlates of approaches to learning: a cross-cultural meta analysis. In: Sternberg, R.J., Zhang, L.F. (Eds.), Perspectives on Thinking, Learning, and Cognitive Styles. Lawrence Erlbaum, Mahwah, NJ, pp. 165–195.

Zimmerman, B.J., 1990. Self-regulated learning and academic achievement: an overview. Educational Psychologist 24, 3–17. http://dx.doi.org/10.1207/s15326985ep2501_2.

Zimmerman, B.J., Martinez-Pons, M., 1986. Development of a structured interview for assessing student use of self-regulated learning strategies. American Educational Research Journal 23, 614–628. http://dx.doi.org/10.2307/1163093.

Zusho, A., 2008. Cultural variation in the motivational standards of self-enhancement and self-criticism among bicultural Asian American and Anglo American students. International Journal of Psychology 43, 904–911. http://dx.doi.org/10.1080/00207590701838121.

Zusho, A., Clayton, K., 2011. Culturalizing achievement goal theory and research. Educational Psychologist 46, 239–260.

Zusho, A., Njoku, H., 2007. Culture and motivation to learn: exploring the generalizability of achievement goal theory. In: Salili, F., Hoosain, R. (Eds.), Culture, Motivation, and Learning: A Multicultural Perspective. Information Age, Greenwich, CT, pp. 91–113.

Cross-Cultural Training in Industrial and Organizational Psychology

Ariel Lelchook, Gettysburg College, Gettysburg, PA, USA
Mary Sully de Luque, Thunderbird School of Global Management, Glendale, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The focus of this article is on the current state of the literature in cross-cultural training in Industrial–Organizational Psychology. There has been over four decades of research on cross-cultural training; however, the increasing globalization of organizations has presented new challenges in the past decade that need further review. Future areas of cross-cultural training are discussed in relation to issues such as cultural intelligence, virtual teams, and a framework for global leadership.

For over four decades, cross-cultural training (CCT) research has focused on a variety of topics from differences across national boundaries (Berry, 1974; Brislin and Pedersen, 1976) to expatriate deployment (Littrell et al., 2006) to multicultural virtual teams (Brandl and Neyer, 2009). Through the 1990s, there was an intense interest in CCT, which was increased through expatriate assignments. This past decade, CCT has seen the effects of globalization as CCT has been connected to cultural intelligence (CQ) (Earley and Ang, 2003; MacNab, 2012), societal differentiation (Berry, 2008), and leadership appreciation (Holt and Seki, 2012), among other topics. As organizations continue to become increasingly global with people from various cultures working together, it is important to consider factors that influence successful communication and collaboration between diverse groups of people. Although the effectiveness of CCT has been the source of much debate, interest in this topic endures. In this article, we review the major literature on the topic and discuss future research directions.

One challenge that has resulted in studying globalization stems from the semantics used to conduct multicultural research. Early training programs were often referred to as "cross-cultural or intercultural orientation programs" (Bhawuk and Brislin, 2000: p. 162). Terms such as multicultural, cross-cultural, intercultural, and transnational have at times been used to mean the same thing, and at other times referred to different concepts (Holt and Seki, 2012). In a review of the literature through the early 2000s, Littrell et al. (2006) defined CCT as the "educative process used to improve intercultural learning via the development of the cognitive, affective, and behavioral competencies needed for successful interactions in diverse cultures" (p. 356). Other reviews and discussions of cross-cultural research have also defined CCT in a similar manner (e.g., Landis, 2008; Landis and Brislin, 1996; Morris and Robie, 2001). We are adopting the term *cross-cultural* when referring to the study of cultures.

Expatriates have been a focus of many different CCT theoretical and empirical studies (see Bhawuk and Brislin, 2000, for a review). The primary goals of CCT largely include components about how to learn, being adaptable, and processing disconfirmed expectations; with the ultimate aim of making isomorphic attributions rather than rushed assumptions about host culture behaviors (Bennett, 1986; Brislin and Bhawuk, 1999). While many methods can be used to deliver CCT, there is still a lack of information about whether the training organizations conduct is successful (Littrell and Salas, 2005). Studies have found positive relationships between effectiveness of CCT and global assignment adjustment (Brewster, 1995), performance (Caligiuri et al., 2001), and cultural awareness and overall well-being (Black and Mendenhall, 1990). In a meta-analysis examining expatriates, CCT was found to positively relate to performance and adjustment (Morris and Robie, 2001). However, these results were lower than in a previous meta-analysis (see Deshpande and Viswesvaran, 1992), and there was insufficient data to examine differences in training method or amount of time spent in training. Recent research has continued to examine the effectiveness of CCT (e.g., Sanchez-Burks et al., 2007; Okpara and Kabongo, 2011; Puck et al., 2008); however, more empirical work is needed.

The term expatriate has been used throughout the CCT literature to describe people living outside their country of origin. Changes in global organizations show that being a global worker does not necessitate becoming an expatriate. Global workers can include workers that travel for a short period of time (e.g., a one-day business trip), or even workers that remain in their country of origin, but frequently interact with others from another culture (e.g., through phone calls, e-mails, and other virtual communication). Due to these continuing developments in globalization, there is a need to examine workers that do not qualify as expatriates, but nonetheless interact with people from other cultures (e.g., Brandl and Neyer, 2009). Cosmopolitan is a term that refers to someone who has recurrent experience with many different cultures. A global worker today may be a cosmopolitan employee, without necessarily being an expatriate.

Integrating Theoretical Contributions

Many theories have been examined in relation to CCT. While there is no unifying theoretical framework, two important concepts are *culture shock* (stress experienced by people living in a different culture) and *culture distance* (the amount of distance and differences between two cultures; Bhawuk and Brislin, 2000). The two concepts are significant because they appear in many of the frameworks that have been used to examine CCT, and practically they provide ready reasons for conducting CCT (Bhawuk and Brislin, 2000). These two concepts can also be directly linked since more culture shock will be experienced when there is a greater distance between someone's culture of origin and host culture.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.22022-9

U-Curve Theory of Adjustment

The concepts of culture shock and culture distance can be seen in early theories of expatriate adjustment. One such theory is the U-curve theory of adjustment, which is a framework that focuses on how people adjust to being in a new culture (Lysgaard, 1955). The U-curve theory of adjustment consists of four stages of adjustment (honeymoon, culture shock, adjustment, and mastery; see **Table 1**). The theory was named based on the U-shaped pattern that emerged as expatriates adjusted to another culture (Lysgaard, 1955). Adjustment is considered relatively easy during the initial contact with a new culture, as many things are new and exciting. Adjustment becomes more difficult as people become frustrated with the new culture. Following this period of frustration, people become accustomed to the host culture, and adjustment again becomes positive. Recent research has continued to expand on the U-curve theory and examines how time can interact with other factors. Assessing expatriate adaptation during the early phases of international assignments, Firth et al. (2014) found that expatriates high in cross-cultural motivation and psychological empowerment were more likely to demonstrate a higher level of work adjustment in early stages. The results also showed that these same employees need to continually be challenged in their expatriate assignments over time, to maintain long-term adjustment. These findings provide support for screening of potential employees who are higher on these characteristics, to help ease the adjustment process.

The concept of culture distance is reflected in the U-curve theory, as adjustment is supposedly easier if the unfamiliar culture is closer to one's own, as there would be less uncertainty and ultimately less stress (Liu and Lee, 2008). As a more general explanation for CC adjustment that focuses on longer periods of time in a host country, this theory has had much utility. It may help explain the experience of culture and suggest methods of CCT for expatriate adjustment. However, other theories may offer more specific suggestions for CCT methods when preparing for short term or virtual assignments (e.g., working with global virtual teams, working briefly in a company's foreign location, adjusting to an expatriate in their own country).

Table 1 U-curve theory of adjustment

Stage	*Definition*
Honeymoon	Expatriates are interested in all the new and different aspects of the culture. (Occurs during the first week.)
Culture shock	Expatriates become frustrated when trying to adapt to the cultural norms of a new place. They may express anger and hostility to the culture and people. (May last from 1 to 6 months.)
Adjustment	Expatriates are adjusting to the differences in the host culture. (May occur from about 6 to 36 months.)
Mastery	Expatriates are able to behave appropriately in a different culture and have an understanding of the host culture's norms and values. (May occur starting around 25 months.)

As summarized in Liu, C.-H., Lee, H.-W., 2008. A proposed model of expatriates in multinational corporations. Cross Cultural Management 15, 176–193.

Social Learning Theory

Social learning theory (SLT) is a cognitive and behavioral approach that has provided an informative theoretical framework for some CCT methods (Black and Mendenhall, 1990; Morris and Robie, 2001). Proposing that people learn by both observation and experience, Bandura (1977) introduced SLT comprised of four main components: attention (behavior must first be noticed before it can be copied), retention (behavior must be remembered), reproduction (behaviors in memory must be put into action), and incentives (behaviors that are perceived to be positively rewarded are enacted). SLT can inform cross-cultural adjustment, as different behavior patterns may occur across cultures. SLT helps explain how people adjust to different cultures, as well as challenges they face with adjustment. The U-curve theory of adjustment provides a general timeline of adjustment challenges; SLT can help explain what is needed for successful adjustment.

Black and Mendenhall (1990) proposed that SLT can influence cross-cultural adjustment. As SLT suggests that people learn from both observation and experience, these scholars posited that CCT would be most beneficial if it includes models (either hypothetical or simulations) of appropriate behavior. Building on their work, Liu and Lee (2008), described four dimensions of successful cross-cultural acculturation. These four dimensions include self-orientation, other-orientation, perceptual skills, and cultural toughness. Self-orientation increases self-esteem and confidence, which helps people deal effectively with stress and adjust to a different culture. Other-orientation is the ability to develop relationships in other cultures (especially mentoring). Perceptual skills are being nonjudgmental and more willing to align beliefs with a host country. Cultural toughness means adapting to the living standard in another country. Together, this is related to the concept of cultural distance, because adjustment proves more difficult when discrepancies in living standards are larger.

Applying the four-dimension learning process of acculturation with SLT can enhance CCT (Liu and Lee, 2008) by guiding training to incorporate methods, which allow for both observation and behavior modeling (e.g., experiential training, simulations, etc.). Training that incorporates observation and experience can increase successful acculturation by helping participants associate positive outcomes with culturally appropriate behaviors, thereby increasing self-orientation (which has been recently advocated by others (e.g., Chao et al., 2011)). This experience can provide individuals with an initial positive interaction that can help develop future successful relationships, enhancing their other-orientation. Accordingly, when training participants encounter a cross-cultural social interaction that they observed or experienced in training, they are more likely to remember and be able to reproduce this behavior. Perceptual skills can be helped by CCT that includes information about how others view the world (e.g., in culture assimilator training) and draw attention to appropriate models of behavior. Making the assumption that trainees receive accurate information, they will develop accurate cognitive maps of appropriate behavior and be able to envision the appropriate behavior in a cross-cultural interaction (Black and Mendenhall, 1990). The role of toughness in relation to SLT needs further study, but holds much promise.

While cross-cultural scholars have started to integrate theories, there is still more work to be done (Landis, 2008).

Cultural Intelligence

As organizations become more global, theories focusing on how to interact within specific cultures (i.e., SLT and creating cognitive maps) may not fully explain successful cross-cultural interaction. While it is possible to develop cognitive maps for a few specific cultures, it becomes increasingly difficult to retain accurate cognitive maps for many different cultures. Further advancing research on cross-cultural interaction, scholars have introduced the concept of CQ (Ang et al., 2007; Brislin et al., 2006; Earley and Ang, 2003; Ng and Earley, 2006). CQ has commonly been defined as successful adjustment to another culture (Brislin et al., 2006). Alternatively, it has been defined as behaviors that are intelligent or appropriate in different cultures (Brislin et al., 2006), or "adjustments a person can make to be effective across cultures" (MacNab et al., 2012: p. 68). While the exact definitions of CQ vary, the most generally applicable description refers to developing knowledge and enacting behavior appropriate in different cultural contexts (without requiring someone to live in another culture). In the current global environment, people may need to engage in cross-cultural behaviors while working as an expatriate, interacting with an expatriate in their own culture, or are temporarily traveling or engaging in virtual work with someone from another culture. Being able to identify the norms, values, and behaviors of a new culture may be a useful skill and one that can be taught by CCT (MacNab, 2012). Initial research using CQ as a framework for experiential training has provided basic support for CQ (MacNab et al., 2012; Ng et al., 2009).

Among this research, Earley and Peterson (2004) developed a three-part framework for CQ consisting of cognitive/metacognitive, motivation, and behavior components. Cognitive refers to thinking, learning, strategizing, and interpreting information (e.g., the meaning of gestures and have them make sense). Motivation refers to appropriate persistence or realistic goal setting in the context of intercultural interactions, especially when they become confusing or the meaning is not clear, as happens in new contexts. The behavior dimension refers to consciously adapting behaviors appropriate for the cultural context. Using CQ as a framework, CCT methods may be specifically tailored to these three dimensions. These concepts build on the cognitive and behavior components of SLT. Including motivation is an important factor when considering global workers' successful adaptation to cross-cultural situations, even if they have limited existing cognitive maps.

Types of Training

There are two main categories of CCT that are frequently discussed: the *mode* of the training, and the *focus* of the training. The mode of training refers to how information is delivered (e.g., conventional vs experiential training). The focus of training refers to the content or information contained within a training course (e.g., increase cultural awareness, teach culturally appropriate behaviors, etc.). **Table 2** provides an

Table 2 Mode and focus of training

Training mode/focus	*Definition*
Lecture/conventional training	This mode of training involves passive learning. Through this mode, the instructor gives a lecture or talk to participants. It often involves problem-solving activities provided by the instructor, but the problems are not in a realistic setting. This mode is one of the simpler and inexpensive methods of CCT, often delivered as moral training (Bennett, 1986). Compared with no training, this mode may provide benefits to employees in cross-cultural interactions (Bhawuk and Brislin, 2000).
Experiential training	This mode of training was developed in response to the traditional lecture mode of training. These methods include substantial involvement from participants and may include simulation exercises. Simulation games often divide participants into two groups, with each group receiving guidelines about an imaginary culture of which they are members. Helpful for highlighting differences, simulations lack the authenticity of being in a different culture. Other modes of experiential training may include simulating target cultures, or helping participants become more self-aware (Bhawuk and Brislin, 2000; Ng et al., 2009).
Culture assimilator	Cultural assimilator is a training mode where real-life misunderstandings in cross-cultural interactions are described (Fiedler et al., 1971). After reading a critical incident, participants are asked to think about the misunderstandings. Four or five alternative choices of action follow the critical incident. One choice reflects one culture involved in the critical incident and another choice reflects the second culture involved in the critical incident. The other two to three choices are less optimal, but tend to reflect a variety of individual differences (Bhawuk and Brislin, 2000).
Cultural awareness training	Cultural awareness training focuses on teaching participants to be aware of aspects of their own culture that are not necessarily generalizable to other cultures (Morris and Robie, 2001). This can help participants identify how cultures differ from their own.
Attribution training	This mode of training helps participants think in ways more similar to someone of another culture (Morris and Robie, 2001). This can include behavioral modeling, if the participant is expected to interpret the situation from the perspective of a member of another culture (Morris and Robie, 2001). Based on SLT, observation can help people learn about appropriate behaviors (Bhawuk and Brislin, 2000).
Language training	Language training is an important focus because of its significance for cross-cultural adjustment. Although participants may not achieve fluency of the language, gaining the proficiency to communicate at a basic level of the host language, such as exchanging common courtesies, assists in learning another culture (Caligiuri et al., 2001; Puck et al., 2008).

overview of different types of CC training that have been discussed in the literature. These training modes do not necessarily have to compete with each other, but may actually complement each other.

Specific vs General Training

Discussion of training methods emphasizes the classification of training into two main categories: culture-specific or culture-general training. Culture assimilator training is often culture-specific. It provides information to participants about what they might expect in realistic social situations and imparts knowledge about behavioral expectations. While this type of training may be helpful for someone moving abroad and likely interacting with people from one host culture, it may not be as effective for someone regularly interacting with people from a variety of different cultures. Cultural assimilator training is one of the earliest, most widely tested, and most commonly used methods of CCT (Bhawuk and Brislin, 2000; Morris and Robie, 2001). However, limitations of this method's effectiveness have been noted, and this is especially so with the increase in cosmopolitan employees. Thus, more recent research has focused on alternative types of training.

The more general culture awareness training may have greater applicability for someone interacting with many cultures. It is often time consuming and unrealistic for employees to remember how to appropriately interact with many different cultures. Other more general methods of CCT may be of higher utility in improving cross-cultural communication. Cultural awareness training helps employees be more aware of when their cultural assumptions might lead to miscommunication. Thus, it might help employees be aware of important aspects of cross-cultural communication, even if they are unfamiliar with a specific culture. This is illustrated in the work by House et al. (2004) who demonstrate that culture and leadership differences influence management and communication across societies, with much of this research currently assisting in cultural awareness training.

Evaluating the Evidence: Recent Empirical Research in CCT

While many different types of training have been proposed, it is helpful to examine several studies that have compared training processes. In this section, we review more recent work in CCT research. For a review of earlier studies, please see exemplary previous reviews of the theoretical and empirical literature (e.g., Bhawuk and Brislin, 2000; Landis, 2008; Landis and Brislin, 1996; Morris and Robie, 2001; Littrell and Salas, 2005). While empirical studies address many different topics, most studies can be classified as assessing either pre- or postdeparture training, evaluating culture-specific or culture-general training, and addressing conventional or experiential training (see **Table 3** for a classification of select recent articles). In addition to these three components, we have reviewed some recent empirical research focusing on supplementary factors relevant to CCT in more detail.

Non-Expatriate Research

Many research studies (including many of the empirical studies discussed below) examine CCT in relation to expatriates. However, Sanchez-Burks et al. (2007) conducted a study examining American cosmopolitan employees preparing for a short-term six-week international project (where 2 weeks were spent on a short trip working in China or Chile). They examined how cultural assimilator training compared with relational ideology training, focusing on non-expatriates. Overall, their results suggested that relational ideology training (which is a form of experiential training that includes a cultural awareness component) has benefits over cultural assimilator training. While the results supported the benefits of experiential training over cross-cultural assimilator training, it is important to take into account this training was specifically geared toward the three cultures of interest (US as the culture of origin and East Asia and Latin America as the host cultures).

Focus on Culture-Specific and Culture-General Training

Among studies assessing a non-US sample, Puck et al. (2008) examined types of training methods comparing culture-specific with culture-general training. Examining predeparture CCT on adjustment (general, work, and interactional adjustment) of German expatriates working in different countries, a comprehensive measure of training was examined in addition to examining general CCT participation and other factors relevant to adjustment. To create the comprehensive measure, CCT was categorized into four groups (general conventional, specific conventional, general experiential, and specific experiential training), and each participant received a score (1–4) based on the number of types of training completed. The results of the

Table 3 Training components addressed in article[a]

Study	*Pre/postdeparture*	*Culture specific/general*	*Conventional/experiential*
Sanchez-Burks et al. (2007)	Predeparture	Specific	Experiential
Puck et al. (2008)	Predeparture	Specific and general	Conventional and experiential
Lievens et al. (2003)	Predeparture	Specific	Experiential
Okpara and Kabongo (2011)	Predeparture	Specific and general	Conventional and experiential
MacNab et al. (2012)	Postdeparture	Specific	Conventional and experiential (mostly experiential)
Osman-Gani and Rockstuhl (2009)	Pre- and postdeparture	Not specified	Not specified
Liu and Lee (2008)	Predeparture	Specific	Not specified
Tarique and Caligiuri (2009)	Postdeparture	Specific and general	Conventional and experiential

[a]These studies are discussed below or elsewhere in this article.

study suggested that language skills were important in predicting adjustment, but adjustment was not predicted by predeparture training, training length, nor training comprehensiveness. Future research may consider how else data can be collected and analyzed to further aid researchers' and practitioners' understanding of the benefits and limitations of different types of training.

Okpara and Kabongo (2011) further extend the literature on CCT, as they also examined the effects of general conventional, specific conventional, general experimental (experiential), and specific experimental (experiential) training on cross-cultural adjustment. Rather than creating a comprehensive measure of training, they examined the four types of training separately, by assessing expatriate managers working in Nigeria. Overall, they found that all four types of CCT somewhat positively relate to expatriate adjustment (work adjustment, general adjustment, interaction adjustment, and psychological adjustment). Specific experimental training explained slightly larger amounts of variance in adjustment than the other types of training. The relationships between general experimental CCT and adjustment showed slightly weaker associations. This highlights the need for future studies to explore the effects of culture-specific compared to culture-general training.

Inclusion of Contextual and Personality Variables

Rather than just looking at the success of a specific type of training program, some research has also examined personality or contextual factors that impact training effectiveness. Lievens et al. (2003) conducted a study assessing predictors of CCT. They looked at the predeparture experiential training program and language executive training program for European managers being assigned to Japan. Openness predicted training performance, as well as adaptability, teamwork, and communication. Language acquisition (which has previously been related to adjustment) was predicted by cognitive ability. In another study, Osman-Gani and Rockstuhl (2009) examined the relationship between perceived pre- and post-CCT effectiveness, self-efficacy, and adjustment of expatriates working in Asia. Expatriate managers, originating from four different nations, showed that CCT facilitated adjustment through an increase in their self-efficacy. This study helped identify a construct that mediates the effectiveness of CCT on adjustment. Further research should examine other factors that impact successful adjustment in addition to and conjunction with CCT.

Postdeparture Training

In contrast to research that has focused on predeparture training and mainly used survey methodology that relies on participant self-recall to study training, research has also examined postdeparture CCT for expatriates. MacNab et al. (2012) created an experiential training program to help develop the CQ of expatriate university students (students not originally from Australia). They developed a seven-step experiential training program that occurred across 8 weeks and was embedded within an international management course and cross-cultural management course (see MacNab et al., 2012, for the detailed seven-step program). Overall, they found support for this seven-step experiential training program as affecting CQ. They also included individual (e.g., general self-efficacy) and contextual (relationship of someone with the host culture based on contact theory) factors in their study. They found that higher general self-efficacy and more positive contacts with people from the host country increased internalization of CQ. While the results for using a CQ framework and experiential training were positive, future empirical research should further examine CQ as a useful framework for CCT, especially beyond a student population.

Tarique and Caligiuri (2009) argued that in-country CCT would increase cultural knowledge if the training units were divided and disseminated over time. This empirical experiment demonstrated that, similar to predeparture CCT, in-country CCT is another worthwhile intervention for knowledge acquisition for expatriates. Future research would benefit from extrapolating this study beyond the student population utilized in this study. Additional research should also consider if there are forms of postdeparture training that would assist cosmopolitan employees embarking on shorter visits.

Leadership

Interest in cross-cultural leadership has grown exponentially with the advent of globalization. Leaders need to be aware of local pursuits, yet enact global thinking; this is augmented by intensified competition, as companies are able to distribute products and services with increasing ease. The resulting challenges of this increased competition have moved organizations to focus on and invest in leader effectiveness (Caligiuri, 2006). This has resulted in lively discussions of what constitute effective leaders as well as how to develop global leaders (Butler et al., 2012; Holt and Seki, 2012; Ng et al., 2009). Yet attaining a unified definition of a global leader and how to develop global leaders remains somewhat elusive (Bird et al., 2010; Gundling et al., 2011). The number and scope of the tools developed to assess assorted aspects of global leadership is impressive. The list is long with tools such as the Global Mindset Inventory (Javidan and Teagarden, 2011), the Global Competencies Inventory (Bird et al., 2010), CQ (Van Dyne et al., 2008), the Leadership Versatility Inventory (Kaplan and Kaiser, 2003), Transition Leadership Wheel (Bunker and Wakefield, 2006), and the Opposite Strengths Inventory (Thomas and Thomas, 2006) being just a few examples of tools that exist to measure various models of global leadership. Often, leadership training is determined by the outcomes of these assessments.

Most models of global leadership assignments have concentrated on expatriate performance and adjustment, not leadership development (Ng et al., 2009). International assignments are seen as ways to develop global leaders; however, few studies have assessed how leaders turn their expatriate learning into experiences that enhance the effectiveness of their actual leadership. From this perspective, the expatriate assignment functions as a form of CCT. Through the integration of research on CQ and experiential learning theories, Ng et al. (2009) address this question of how and why global leaders learn from their international work assignments. In their model, they propose that CQ will act as a moderator,

increasing the probability that individuals (expatriates or cosmopolitans) on assignments internationally will be part of four stages of experiential learning theory. Further, they contend that participation in these four stages (experience, reflect, conceptualize, and experiment) will allow leaders to developmentally gain from their experiences during their international assignments. Adopting this developmental perspective of expatriates and cosmopolitans (short-term travelers) is worthy of future study.

Some scholars have made compelling arguments that effective global leadership involves a developmental shift beyond the current competency models. Notably, Holt and Seki (2012) argue that such a shift requires us to question the use of some outdated models of leadership; questioning if we have a Western-centric approach to leadership that is applicable in a global context. They acknowledge that a separate stream of literature about intercultural communication has emerged over the past 30 years; however, even when leadership has been acknowledged, the complexities of leadership have not been addressed in these intercultural communication models. Further, Holt and Seki (2012) suggest that today most leaders operate in a global context and that being successful as a leader in a single context does not equate to being an effective global leader. Instead of simply adding a competency to other leadership abilities, being an effective global leader requires a developmental shift. They propose four developmental shifts including: "(a) developing multicultural effectiveness (MCE), (b) becoming adept at managing paradoxes associated with global work, (c) cultivating the 'being' dimension of human experience, and (d) appreciating individual uniqueness in the context of cultural differences" (Holt and Seki, 2012: p. 197). These dimensions cannot be addressed by simply adding competencies to existing leadership models.

There are, however, some existing leadership models that account for cross-cultural differences within their framework (Inceoglu and Bartram, 2012). One of the most noteworthy and comprehensive studies assessing cross-cultural leadership is Project GLOBE (Chhokar et al., 2007; House et al., 2004; House et al., 2014). Carried out in three phases, the first phase was focused on the development of research instruments to assess both societal culture and leadership, which was conducted through the participation of researchers from around the world. From this initial phase, nine cultural dimensions and six global leadership dimensions (comprised of 21 first-order dimensions) emerged. Using the scales developed in the first phase, researchers in the second phase sought to further differentiate attributes of culture and leadership at both the organizational and societal levels focusing on mid-level managers in 62 societies. The second phase was devoted to implementing the surveys to assess the nine core attributes of cultures and six broad dimensions of leadership, with a keen interest in noting whether universal and/or culturally contingent leader characteristics emerged across cultures. Universal leadership characteristics are those seen as important across cultures, while culturally contingent leadership characteristics are seen as concepts influenced by values, beliefs, cognitive schemas, and behaviors that are unique to a culture (Bass, 1997; Den Hartog et al., 1999; Dickson et al., 2012). Culturally contingent leadership characteristics are considered desirable in some cultures, but not others (Den Hartog et al., 1999; Dorfman et al., 2004). The analyses conducted in this stage of the project revealed the existence of both culturally universal and culturally contingent leadership attributes, as well as their impact on leadership. Most recently, the third phase was dedicated to examining the impact and effectiveness of specific CEO leader behaviors on top management team attitudes and firm performance (House et al., 2014). Assessing actual CEO behavior across 24 countries, the results from this phase supported the idea of universal and culturally contingent leader attributes. Together these studies are important for CCT in global leadership, as they greatly enhance our understanding of culture and leadership expectations across many societies. Unfortunately there are a number of practitioners who do not know about this model, and ironically if they do it is the scope of this work that makes it intimidating for practitioners (Holt and Seki, 2012). In order to effectively use the leadership framework promoted in project GLOBE, future CCT methods should be constructed with this model in mind, and be communicated in a way that encourages practitioners to use this research to inform leadership frameworks (e.g., Javidan et al., 2006).

Teams

As multicultural teams become more common, the need for CCT has increased. Cross-cultural team training can help with problems experienced by teams when members interact with other cultures (Littrell et al., 2006). Multicultural teams commonly form through the presence of expatriates in organizations or through virtual teams, with the resulting culture differences affecting the interaction across the teams (Martins et al., 2004). Global virtual teams usually consist of individuals that are from different geographic areas who rarely (if ever) meet in face-to-face communication. With continued technological changes and ever-increasing ways to communicate with people, the need for CCT in virtual teams continues to intensify. Leaders have considerable influence on teams and training for multicultural teams, which impacts the overall team effectiveness (Burke et al., 2005). However, as discussed above, the competencies for intercultural leadership may differ from intracultural leadership.

Since virtual teams rely on technology to communicate, the role of a virtual team leader may be different from a traditional team leader. A leader of a virtual team may need to communicate differently in order to lead, make decisions, and build trust (Meyer, 2010). With online teams, the leader needs to help provide clear direction and be the point person for coordination. In addition to dealing with all interaction occurring through technology, the team leader is often dealing with cultural differences if team members are from different countries and backgrounds. Team members from dissimilar cultures may have different expectations about the norms surrounding appropriate organizational behavior. Factors such as how to arrive at decisions, behaviors that represent trust, and the types of communication that demonstrate politeness and professionalism may differ between cultures (Meyer, 2010). Trust may be especially challenging for global virtual teams engaged in technology-mediated communication (Jarvenpaa and Leidner, 1999). Face-to-face and informal communication have been identified as important for trust, but these forms of

communication are not often part of virtual team interactions (Oertig and Buergi, 2006). As virtual team members are typically cosmopolitan employees (non-expatriates), the intensity and type of training may need to differ from expatriates.

When working in teams with people from different cultures there is a need to adjust to unknown situations and manage uncertainty (Gudykunst, 1998). Brandl and Neyer (2009) propose the importance of cognitive adjustment in CCT, especially as it pertains to newcomers. The concept of cognitive adjustment dates back decades and in this context it reflects team members' ability to understand someone else's perspective (Schutz, 1944). This should be considered when selecting a training program for virtual team members. The usefulness of culture-specific orientation is of limited utility when preparing for the complexities of engaging in multinational team interaction (Earley and Peterson, 2004). General cultural awareness training may be better suited for employees working on cross-cultural teams by enhancing ones' ability to adjust to unknown and uncertain situations (Brandl and Neyer, 2009). This training may help team members be more aware that uncertainty is a normal part of virtual teams. In turn, this makes them feel like they have more support and thus are more willing to explore unknown situations (Glanz et al., 2001), increasing their ability to make mental connections (Osland and Bird, 2000). As virtual teams are progressively either complimenting or even replacing traditional types of global assignments, the necessity for CCT to expand beyond traditional contexts is crucial.

Future Research

Building on recommendations for future research discussed throughout this article, several additional points are worth noting. First and foremost, future research should continue to focus on the evaluation of CCT methods. As more empirical studies are conducted, it would be helpful to examine if there are disparities in the usefulness of different types of training. Other factors to further examine include determining optimal types of training best matched to certain situations (e.g., cross-cultural manager training, training for cross-cultural virtual teams, training for cosmopolitan employees vs expatriates, etc.), pre- vs postdeparture training, individual and contextual factors, as well as training mode and focus.

Future research may also benefit from examining and incorporating, when relevant, components from other training literature such as diversity training. The work on cross-cultural issues and diversity in organizations are interconnected in many ways, yet the theory, research, and practice in the two fields have different frameworks, approaches, and goals. Ferdman and Sagiv (2012) discuss some similarities and differences between CCT and diversity training, and propose how both fields could benefit from a collaboration. CCT focuses on preparing individuals to work in cross-cultural contexts and with people from other countries. Diversity training looks at diversity in organizations in terms of race, ethnicity, gender, or profession. This training addresses issues such as identity, power, privilege, stereotyping, leadership, conflict management, and inclusive behavior. Both fields provide information for learning and interaction, dealing with conflict, and collaboration. CCT can benefit from including other factors (beyond cultural dimensions) and focusing on how those factors can influence interpersonal relationships between people from different cultures. This is an area that may be especially relevant for CCT and research to address.

Future research may further examine how different theories influencing CCT may be integrated to ultimately create more effective CCT methods. Several challenges have been noted when assessing MCE and global leadership development. These challenges include (1) an increasing number of competing frameworks and models; (2) an insufficient level of validation of assessment tools; (3) an inadequate focus on performance of various intercultural models; and (4) along with an insufficient multicultural focus in some leadership models (Holt and Seki, 2012). Future research should focus on integrating performance and multicultural dimensions into the constructs of global leadership competency models.

Increasingly, researchers argue that CCT alone might not always enhance cross-cultural adjustment (Puck et al., 2008; Tarique and Caligiuri, 2009), so companies may want to implement posttraining interventions to assist in learning transfer and adjustment. Future research could assess how organizations might use posttraining coaching, language training, and mentoring as possible interventions to facilitate cross-cultural adjustment. This relates to a final recommendation, which is the need for more longitudinal studies in CCT that encompass many of the aforementioned factors. The more we understand the complexities of CCT, the greater the benefits to the employees and our increasingly global organizations. With the acceleration of globalization, the field is more informed and the challenges ahead more multilayered. Reviewing the field of CCT over the past four decades provides a compelling reflection of what we have learned and where we are progressing.

See also: Diversity in Organizations; Industrial–Organizational Psychology: Science and Practice; Managerial and Organizational Cognition; Negotiation and Conflict, The Psychology of; Personality and Values at Work; Teamwork and Team Performance Measurement.

Bibliography

Ang, S., Van Dyne, L., Koh, C., Ng, K.Y., Templer, K.J., Tay, C., Chandrasekar, N.A., 2007. Cultural intelligence: its measurement and effects on cultural judgment and decision making, cultural adaptation, and task performance. Management and Organization Review 3, 335–371.

Bandura, A., 1977. Social Learning Theory. Prentice-Hall, Englewood Cliffs, NJ.

Bass, B.M., 1997. Does the transactional-transformational leadership paradigm transcend organizational and national boundaries? American Psychologist 52, 130–139.

Bennett, J.M., 1986. Modes of cross-cultural training: conceptualizing cross-cultural training as education. International Journal of Intercultural Relations 10, 117–134.

Berry, J.W., 1974. Psychological aspects of cultural pluralism: unity and identity reconsidered. Topics in Culture Learning 2, 17–22.

Berry, J.W., 2008. Globalisation and acculturation. International Journal of Intercultural Relations 32, 328–336.

Bhawuk, D.P.S., Brislin, R.W., 2000. Cross-cultural training: a review. Applied Psychology: An International Review 49, 162–191.

Bird, A., Mendenhall, M., Stevens, M.J., Oddou, G., 2010. Defining the content domain of intercultural competence for global leaders. Journal of Managerial Psychology 25, 810–828.

Black, J.S., Mendenhall, M.E., 1990. Cross-cultural training effectiveness: a review and a theoretical framework for future research. Academy of Management Review 15, 113–136.

Brandl, J., Neyer, A.K., 2009. Applying cognitive adjustment theory to cross-cultural training for global virtual teams. Human Resource Management 48, 341–353.
Brislin, R.W., Bhawuk, D.P.S., 1999. Cross-cultural training: research and innovations. In: Adamopoulos, J., Kashima, Y. (Eds.), Social Psychology and Cultural Context. Sage, Thousand Oaks, CA, pp. 205–216.
Brislin, R.W., Pedersen, P., 1976. Cross-cultural Orientation Programs. Gardner, New York.
Brislin, R., Worthley, R., MacNab, B., 2006. Cultural intelligence: understanding behaviors that serve people's goals. Group and Organization Management 31, 40–55.
Brewster, C., 1995. Effective expatriate training. In: Selmer, J. (Ed.), Expatriate Management: New Ideas for International Business. Quorem, Westport, CT, pp. 57–71.
Bunker, K.A., Wakefield, M., 2006. Leading in times of change. Harvard Management Update 11, 1–4.
Burke, C.S., Hess, K.P., Priest, H.A., Rosen, M., Salas, E., Paley, M., Riedel, S., 2005. Facilitating leadership in a global community: a training tool for multicultural team leaders. Proceedings of the 2005 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL, pp. 1–12.
Butler, C., Zander, L., Mockaitis, A., Sutton, C., 2012. The global leader as boundary spanner, bridge maker and blender. Industrial and Organizational Psychology 5, 240–243.
Caligiuri, P., 2006. Developing global leaders. Human Resource Management Review 16, 219–228.
Caligiuri, P., Phillips, J., Lazarova, M., Tarique, I., Burgi, P., 2001. The theory of met expectations applied to expatriate adjustment: the role of cross-cultural training. International Journal of Human Resource Management 12, 357–372.
Chao, M.M., Okazaki, S., Hong, Y., 2011. The quest for multicultural competence: challenges and lessons learned from clinical and organizational research. Social and Personality Psychology Compass 5, 263–274.
Chhokar, J.S., Brodbeck, F.C., House, R.J., 2007. Culture and Leadership across the World: The Globe Book of In-depth Studies of 25 Societies. Erlbaum, New York.
Den Hartog, D.N., House, R.J., Hanges, P.J., Ruiz-Quintanilla, S.A., Dorfman, P.W., 1999. Culture-specific and cross-culturally generalizable implicit leadership theories: are attributes of charismatic/transformational leadership universally endorsed? The Leadership Quarterly 10, 219–257.
Deshpande, S.P., Viswesvaran, C., 1992. Is cross-cultural training of expatriate managers effective? A meta-analysis. International Journal of Intercultural Relations 16, 295–310.
Dickson, M.W., Castano, N., Magomaeva, A., Den Hartog, D.N., 2012. Conceptualizing leadership across cultures. Journal of World Business 47, 483–492.
Dorfman, P.W., Hanges, P.J., Brodbeck, F.C., 2004. Leadership and cultural variation: the identification of culturally endorsed leadership profiles. In: House, R., Hanges, P., Javidan, M., Dorfman, P., Gupta, V. (Eds.), Leadership, Culture, and Organizations: The GLOBE Study of 62 Societies. Sage, Thousand Oaks, CA, pp. 668–720.
Earley, C.P., Ang, S., 2003. Cultural Intelligence: Individual Interactions Across Cultures. Stanford Business Books, Palo Alto, CA.
Earley, C., Peterson, R., 2004. The elusive cultural chameleon: cultural intelligence as a new approach to intercultural training for the global manager. Academy of Management Learning and Education 3, 100–115.
Ferdman, B.M., Sagiv, L., 2012. Diversity in organizations and cross-cultural work psychology: what if they were more connected? Industrial and Organizational Psychology 5, 323–345.
Fiedler, F.E., Mitchell, T.R., Triandis, H.C., 1971. The culture assimilator: an approach to cross-cultural training. Journal of Applied Psychology 55, 95–102.
Firth, B.M., Chen, G., Kirkman, B.L., Kim, K., 2014. Newcomer abroad: expatriate adaptation during early phases of international assignments. Academy of Management Journal 57, 280–300.
Glanz, L., Williams, R., Hoeksema, L., 2001. Sensemaking in expatriation: a theoretical basis. Thunderbird International Business Review 43, 101–119.
Gudykunst, W.B., 1998. Applying anxiety/uncertainty management (AUM) theory to intercultural adjustment training. International Journal of Intercultural Relations 22, 227–250.
Gundling, E., Hogan, T., Cvitkovich, K., 2011. What Is Global Leadership: 10 Key Behaviors that Define Great Global Leaders. Nicholas Brealey, Boston, MA.
Holt, K., Seki, K., 2012. Global leadership: a developmental shift for everyone. Industrial and Organizational Psychology 5, 196–215.
House, R., Dorfman, P., Javidan, M., Hanges, P., Sully de Luque, M., 2014. Strategic Leadership across Cultures: The GLOBE Study of CEO Leadership Behavior and Effectiveness in 24 Countries. Sage Publications, Thousand Oaks, CA.
House, R.J., Hanges, P.M., Javidan, M., Dorfman, P., Gupta, V., 2004. Culture, Leadership and Organizations: The GLOBE Study of 62 Societies. Sage Publications, Inc, Thousand Oaks, CA.
Inceoglu, I., Bartram, D., 2012. Global leadership: the myth of multicultural competency. Industrial and Organizational Psychology 5, 216–218.
Javidan, M., Dorfman, P., Sully de Luque, M., House, R.J., 2006. In the eye of the beholder: cross cultural lessons in leadership from Project GLOBE. Academy of Management Perspective 20 (1), 67–90.
Javidan, M., Teagarden, M.B., 2011. Conceptualizing and measuring global mindset. In: Mobley, W.H., Li, M., Wang, Y. (Eds.), Advances in Global Leadership, vol. 6. Emerald Group, Bingley, UK, pp. 13–39.
Jarvenpaa, S.L., Leidner, D.E., 1999. Communication and trust in global virtual teams. Organization Science 10, 791–815.
Kaplan, R.E., Kaiser, R.B., 2003. Developing versatile leadership. MIT Sloan Management Review 44, 19–26.
Landis, D., 2008. Globalization, migration into urban centers, and cross-cultural training. International Journal of Intercultural Relations 32, 337–348.
Landis, D., Brislin, R.W. (Eds.), 1996. Handbook of Intercultural Training, second ed. Sage, Thousand Oaks, CA.
Lievens, F., Harris, M., Van Keer, E., Bisqueret, C., 2003. Predicting cross-cultural training performance: the validity of personality, cognitive ability, and dimensions measured by an assessment center and a behavior description interview. Journal of Applied Psychology 88, 476–489.
Littrell, L.N., Salas, E., 2005. A review of cross-cultural training: best practices, guidelines, and research needs. Human Resource Development Review 4 (3), 305–334.
Littrell, L.N., Salas, E., Hess, K.P., Paley, M., Riedel, S., 2006. Expatriate preparation: a critical analysis of 25 years of cross-cultural training research. Human Resource Development Review 5, 355–388.
Liu, C.-H., Lee, H.-W., 2008. A proposed model of expatriates in multinational corporations. Cross Cultural Management 15, 176–193.
Lysgaard, S., 1955. Adjustment in a foreign society: Norwegian Fulbright grantees visiting the United States. International Social Science Bulletin 7, 45–51.
MacNab, B.R., 2012. An experiential approach to cultural intelligence education. Journal of Management Education 36, 66–94.
MacNab, B., Brislin, R., Worthley, R., 2012. Experiential cultural intelligence development: context and individual attributes. International Journal of Human Resource Management 23, 1320–1341.
Martins, L.L., Gilson, L.L., Maynard, M.T., 2004. Virtual teams: what do we know and where do we go from here? Journal of Management 30 (6), 805–835.
Meyer, E., 2010. The Four Keys to Success with Virtual Teams. http://www.forbes.com/2010/08/19/virtual-teams-meetings-leadership-managing-cooperation.html (accessed 12.02.14.).
Morris, M.A., Robie, C., 2001. A meta-analysis of the effects of cross-cultural training on expatriate performance and adjustment. International Journal of Training and Development 5 (1), 112–125.
Ng, K., Earley, C., 2006. Culture + intelligence: old constructs, new frontiers. Group and Organizational Management 31, 4–19.
Ng, K.-Y., Van Dyne, L., Ang, S., 2009. From experience to experiential learning: cultural intelligence as a learning capability for global leader development. Academy of Management Learning and Education 8, 511–526.
Oertig, M., Buergi, T., 2006. The challenges of managing cross-cultural virtual project teams. Team Performance Management 12, 23–30.
Okpara, J.O., Kabongo, J.D., 2011. Cross-cultural training and expatriate adjustment: a study of western expatriates in Nigeria. Journal of World Business 46, 22–30.
Osland, J.S., Bird, A., 2000. Beyond sophisticated stereotyping: cultural sensemaking in context. Academy of Management Executive 14, 65–87.
Osman-Gani, A.M., Rockstuhl, T., 2009. Cross-cultural training, expatriate self-efficacy, and adjustments to overseas assignments: an empirical investigation of managers in Asia. International Journal of Intercultural Relations 33, 277–290.
Puck, J.F., Kittler, M.G., Wright, C., 2008. Does it really work? Re-assessing the impact of pre-departure cross-cultural training on expatriate adjustment. International Journal of Human Resource Management 19, 2182–2197.
Sanchez-Burks, J., Lee, F., Nisbett, R., Ybarra, O., 2007. Cultural training based on a theory of relational ideology. Basic and Applied Social Psychology 29, 257–268.
Schutz, A., 1944. The stranger: an essay in social psychology. American Journal of Sociology 49, 499–507.
Tarique, I., Caligiuri, P., 2009. The role of cross-cultural absorptive capacity in the effectiveness of in-country cross-cultural training. International Journal of Training and Development 13, 148–164.
Thomas, J.W., Thomas, T.J., 2006. The Power of Opposite Strengths: Making Relationships Work. Thomas Concept, Austin, TX.
Van Dyne, L., Ang, S., Koh, C., 2008. Development and validation of the CQS: the cultural intelligence scale. In: Ang, S. (Ed.), Handbook of Cultural Intelligence: Theory, Measurement and Application. M.E. Sharpe, Armonk, NY, pp. 16–38.

Cross-National Comparisons in Education: Findings from PISA

Ming M Chiu, Purdue University, West Lafayette, IN, USA
Sung W Joh, Seoul National University, Seoul, Korea

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Multilevel analyses of large-scale international studies' data have shown how characteristics at multiple levels (country, family, school, student) influence student achievement. Students in countries that are richer academically outperform other students, in part because they have more resources at home (family socioeconomic status, books, etc.) and school (teachers, schoolmates, educational materials, etc.). Students in countries with greater household inequality, inequality across schools, or economic segregation of richer students from poorer students show lower academic achievement than other students. Cultural values and family characteristics showed interaction effects. Together, these results show the importance of an ecological approach to understanding student achievement.

Introduction

Student achievement on international tests differs sharply across countries. Students in Finland, Korea, Japan, and Canada consistently score over 2 standard deviations higher than students in Indonesia, Panama, Peru, and Kyrgyzstan in mathematics, reading, and science across several international comparisons. These differences in academic achievement might stem from the countries' economic and cultural differences. First, the countries with higher average test scores are generally much wealthier (e.g., as measured by gross domestic product (GDP) per capita) and more equal economically (e.g., income Gini) than the countries with lower average test scores. These differences suggest that both greater educational resources and more equal distribution contribute to greater student learning and academic achievement. Second, the countries within each set are culturally diverse, suggesting that cultural values do not account for differences in test scores.

We examine the antecedents of these differences more systematically by adopting Bronfenbrenner's (2005) ecological framework and discussing the results of previous international studies. Bronfenbrenner (2005) modeled children's interactions with surrounding environments including persons and entities within immediate and direct contexts (such as family and school), the relationships across these contexts, and the broader country context (such as economic and cultural differences). Economic and cultural differences across and within countries can fuel differences in resources across families, schoolmates, schools, and students that influence students' learning opportunities and subsequent achievement. Lastly, we discuss the methodologies involved in collecting and analyzing large-scale international data.

Ecology of Student Achievement

Student interactions with their environment can influence their academic achievement (see **Figure 1**; Bronfenbrenner, 2005). Students have frequent and direct contacts with family (or any immediate context to create a *microsystem*), which offers learning opportunities on which they can capitalize to learn and achieve in school. For instance, students living in families with higher socioeconomic status (SES) tend to outperform other students on achievement tests (Baker et al., 2002). Connections between family and school contexts (or any two microsystems that form a *mesosystem*) can also influence academic achievement. For example, when parents are more involved in children's school activities, their children show higher academic performance (Chiu, 2010). Thus, both the immediate contexts and the interactions among them influence

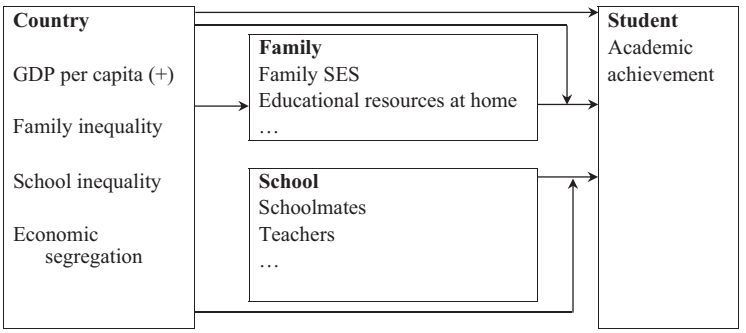

Figure 1 Ecological model of relationships among country, family, school, and student academic achievement. GDP, gross domestic product; SES, socioeconomic status.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92144-5

students' academic achievement. A more distal context, such as a parent's stress at work, might result in less school involvement and their children's poorer academic performance (links between distal and immediate contexts are called *exosystems*, Bronfenbrenner, 2005). These microsystems, mesosystems, and exosystems are all under the influences of the larger cultural contexts, such as the economy and cultural values of the society (*macrosystem*). For example, students in richer countries showed higher academic achievement, and had stronger links between family cultural communication and academic performance (Chiu, 2007). Hence, ecologies of microsystems, mesosystems, exosystems, and macrosystems can help us understand the factors that influence student achievement within and across countries.

The results discussed in this article are largely drawn from the Organization for Economic Cooperation and Development's Program for International Student Assessment (OECD-PISA; OECD, 2010). OECD-PISA assessed 193 076 15-year-old students from 41 countries in 2000–02; 276 165 15-year-old students from 41 countries in 2003; and 475 760 15-year-old students from 65 countries in 2009 (see **Table 1**). They also asked students and principals to fill out questionnaires. International experts from participating OECD and non-OECD countries defined mathematics, reading, and science achievement; built assessment frameworks; created test items; forward translated them; backward translated them; and pilot tested them to check their validity and reliability (for details and sample items, see OECD 2002, and www.pisa.oecd.org). Participating students completed a 2-h assessment booklet and then, a 30- to 40-min questionnaire.

Studies using OECD-PISA data have shown that differences in student achievement at the country macrosystem, school microsystem, and student levels are all substantial (Chiu and Khoo, 2005). Country macrosystem differences account for 31% of the differences in students' mathematics achievement (24% for reading achievement and 22% for science achievement). Meanwhile, school differences account for 25, 30, and 26% of the differences in mathematics, reading, and science achievements, respectively. Lastly, student-level differences account for 44, 46, and 52% of the differences in mathematics, reading, and science achievements, respectively. Together, these results show the importance of examining factors that influence student academic achievement within a country's ecology.

Economic Differences

Students in countries with greater economic productivity (real GDP per capita) show higher academic achievement (Chiu and McBride-Chang, 2010). Richer countries can raise student achievement through extra resources at the country, family, and school levels (Chiu, 2009). These resources provide students with more learning opportunities. In richer countries, governments tend to invest more in education by building more educational infrastructure (schools, libraries, and museums) and providing better educational resources and teacher training. The availability of these educational resources and opportunities can directly enhance students' learning. Students in richer countries can also benefit indirectly through higher nutritional standards or better health care. These help ensure

Table 1 The years in which the following countries participated in OECD-PISA

Country	*2000–02*	*2003*	*2009*
Argentina	X		
Albania	X		X
Argentina			X
Australia	X	X	X
Austria	X	X	X
Azerbaijan			X
Belgium	X	X	X
Brazil	X	X	X
Bulgaria	X		X
Canada		X	X
Chile	X		X
Colombia			X
Croatia			X
Czech Republic	X	X	X
Denmark	X	X	X
Dubai (UAE)			X
Estonia			X
Finland	X	X	X
France	X	X	X
Germany	X	X	X
Greece	X	X	X
Hong Kong–China	X	X	X
Hungary	X	X	X
Iceland	X	X	X
Indonesia	X	X	X
Ireland	X	X	X
Israel	X		X
Italy	X	X	X
Japan	X	X	X
Jordan			X
Kazakhstan			X
Korea	X	X	X
Kyrgyz Republic			X
Latvia	X	X	X
Liechtenstein	X	X	X
Lithuania			X
Luxembourg	X	X	X
Macao–China		X	X
FYR Macedonia	X		
Mexico	X	X	X
Republic of Montenegro			X
The Netherlands	X	X	X
New Zealand	X	X	X
Norway	X	X	X
Panama			X
Peru	X		X
Poland	X	X	X
Portugal	X	X	X
Qatar			X
Romania	X		X
Russian Federation	X	X	X
Republic of Serbia			X
Serbia and Montenegro		X	
Shanghai (China)			X
Singapore			X
Slovak Republic		X	X
Slovenia			X
Spain	X	X	X
Sweden	X	X	X

(Continued)

Table 1 The years in which the following countries participated in OECD-PISA—cont'd

Country	*2000–02*	*2003*	*2009*
Switzerland	X	X	X
Chinese Taipei			X
Thailand	X	X	X
Trinidad and Tobago			X
Tunisia		X	X
Turkey		X	X
United Kingdom	X	X	X
United States	X	X	X
Uruguay		X	X

students are healthy and have adequate energy when they study, which in turn contributes to better academic performance in richer countries.

Family SES

Families in richer countries tend to have more material, human, and cultural capitals that help their children learn more (Chiu and Khoo, 2005). In richer countries, families often have more educational resources at home (such as books, computers, and tutors), which provide students with more learning opportunities. By capitalizing on these learning opportunities at home, students in richer countries can academically outperform students from poorer countries.

Apart from material support, children also benefit from family members' human capital (Chiu and McBride-Chang, 2010). As the overall education level is higher in richer countries, parents in these countries often completed more years of schooling and have higher aspirations for their children's educational attainment. As a result, these parents encourage their children to study, excel academically, and complete higher levels of education. Furthermore, families with both parents are more likely to have more and better educational resources at home and are more involved in their children's schooling.

Parents in richer countries also often give their children more opportunities to learn cultural knowledge, skills, and values, which contribute to better academic performance (Chiu and Chow, 2010). Cultural possessions at home, such as paintings and poetry books, can highlight the importance of one's culture and facilitate communication about cultural values and norms. This family cultural capital can influence students' academic performance in two ways. First, families with more cultural capital often understand their children's schools better and are more connected with these schools, thereby are more likely to be involved in activities and works in these schools. Second, this cultural capital also helps children better understand school norms and their teachers and classmates' expectations in school. So, these children often behave appropriately in school and have better relationships with teachers and classmates. With greater support from their parents, teachers, and fellow students, these children feel a stronger sense of belonging at school, are more motivated to learn, and have greater academic success than other students (Chiu et al., 2012).

School Resources

At the school level, students in richer countries often have more school resources, teacher resources, and schoolmates who share educational resources with one another, compared to students in poorer countries (Chiu, 2010). Schools in richer countries tend to offer more hours of classroom instruction and have better facilities and teaching materials. By utilizing these school facilities and materials, students in richer countries can learn more than those in poorer countries.

Schools in richer countries often have teachers with higher qualifications and better training (e.g., teacher certification), so their students typically show higher academic achievement (Chiu and Khoo, 2005). These teachers are more capable of using effective teaching methods and raising students' learning interest, all of which enhance student learning. Also, these teachers are more likely to create a supportive classroom atmosphere of mutual respect and caring, and foster students' feelings of being valued for academic efforts, reinforcing their students' school engagement. As discussed earlier, students who have better relationships with their teachers feel a greater sense of belonging at their school and higher academic achievement (Chiu et al., 2012). Furthermore, teachers in wealthier nations are more likely to choose a broader degree of curriculum coverage, a more advanced level of learning activities, and more advanced materials, compared to those in poorer countries. These teachers often have higher expectations of their students and help them realize those expectations through higher academic achievement. Students enrolled in privileged schools are perceived as smarter. Their teachers are likely to have higher expectations of them and are willing to invest more in their learning compared to teachers of other students. Also, teachers can directly or indirectly communicate these expectations to their students, such as adopting a more advanced level of learning activities, thereby affecting their expectations and behaviors. These advantages help students in privileged schools learn more than their counterparts in poorer schools.

Students in schools with good disciplinary climates also show higher academic achievement (Chiu and Chow, 2011). A good disciplinary climate at school provides students with a physically and psychologically safe school environment. These feelings of safety throughout a school contribute to stronger school engagement, and students can concentrate on their learning in this environment. As schools in richer countries often have better disciplinary climates, students in these schools tend to have greater school engagement and higher academic achievement.

Schoolmates

Schoolmates can affect one another's learning through their educational resources and group culture (Chiu and Khoo, 2005). Compared to their counterparts in poorer countries, families in wealthier nations are more educated, are more capable in creating an enriched learning environment, and have more positive attitudes toward formal education. When students share their resources and experiences with their

schoolmates, they learn more and show higher academic achievement. Students with proportionately more high-SES and higher-achieving schoolmates showed better academic performance than those with more low-SES and lower-achieving schoolmates. Higher-achieving students are more likely than other students to value learning and academic success, and thereby support such a school culture at school, which supports other students' learning.

Microsystem Effects in Richer vs Poorer Countries

Country wealth also moderates family or school characteristics' links with student learning (Heyneman and Loxley, 1983). Distinguishing between different types of family resources can explain the apparently contradictory results regarding whether the link between family resources and student achievement is stronger or weaker in richer countries (Chiu, 2007). In richer countries, physical resources at home, such as books and computers, have weaker links to student achievement, but intangible family resources, such as family involvement and communication, have stronger links to student achievement.

As richer countries often have more public physical resources (e.g., library books, museum exhibits) that can substitute for a family's physical resources, these public resources dilute the importance of family learning opportunities, thereby weakening their links to student achievement (*modernization*). Family's physical resources (e.g., books) are strongly linked to student achievement in poorer countries, but much less so in richer countries whose GDP per capita exceeds $16 000 per year (in 1990 US dollars; Gamoran and Long, 2006). While basic physical educational resources (textbooks, maps) are widely available in public arenas (schools, libraries, etc.) in richer countries, these physical resources are not as publicly available in poorer countries (especially schools in Latin American countries), allowing family physical resources to have larger effects on student achievement.

On the other hand, the widespread availability of public physical resources (such as public libraries and museums) in richer countries also raises the value of intangible processes such as family communication (*complementary intangible processes*, Chiu, 2007). Unlike physical resources (e.g., textbooks), processes are intangible and dynamic (e.g., family communication, teacher–student). For example, a student benefits from reading an extra book, but that benefit can be magnified substantially by discussing the book with a parent or a teacher (like a left shoe is more useful with a right shoe). Indeed, children in richer countries often spend more time with their parents due to fewer competing siblings, less parent time on housework, and multitasking parents. Thus, family intangible processes' links to student achievement in richer countries are stronger, while those of family physical resources are weaker.

Inequalities and Their Mechanisms

In addition to a country's total educational resources, the distribution of these resources influences student achievement (Chiu and Khoo, 2005). Specifically, inequality of family income (*family inequality*), inequality of school resources (*school inequalities*), and degree of separation of rich students from poor students into different sets of schools (*economic segregation*) show effects on student achievement.

Family Inequality

National and school policies can influence students' family inequality, school inequalities, and economic segregation (Chiu, 2009). Regressive taxes, corruption, and a weak legal system can increase inequality of family incomes (*family inequality*), while progressive taxes, social welfare, transparency, and health care and safety standards (e.g., clean water) can reduce it. Influenced by national policies, family inequality reduces overall student achievement through fewer overall educational resources, diminishing marginal returns, and privileged student bias.

Greater family inequality results in less overall investment in educational resources, which reduces academic achievement among both poor and rich students (*fewer educational resources*, Chiu, 2009). The elite of very unequal countries can send their children to private schools, so they might advocate for fewer public school resources and greater family payment for each child's education. Even if the elite send their children to public schools, they have many more educational resources (e.g., books or tutors) at home than others, so they may not advocate for more public spending. When students' families have to pay for much of their education, poorer parents cannot afford the optimal amount of education for their children, resulting in suboptimal educational resources in the society. As a result, poorer students have fewer educational resources and learn less, thereby reducing overall achievement.

Family inequality lowers overall student achievement due to diminishing marginal returns (Chiu and Khoo, 2005). Consider a thirsty woman and two glasses of water. She greatly values the first glass of water and drinks it all. Her thirst quenched, she does not finish the extra glass of water, revealing its lower value (*diminishing marginal returns*). Likewise, rich students have more educational resources than poor students do, and a richer student typically benefits less from an extra book than a poorer student does. With a less equal distribution of a fixed set of resources, richer students have extra resources and obtain less utility from them compared to poorer students who have fewer resources, resulting in less efficient utility of resources and lower education outcomes overall.

In school systems with larger distribution inequality, privileged parents have more incentive to use their capital to get more educational resources for their children (Chiu and Khoo, 2005). Schools that allocate more resources to richer students exacerbate the effect of disparities in income. Because of diminishing marginal returns, greater *privileged student bias* results in lower overall student achievement.

Some methods of resource allocation (open markets, cronyism, and social–cultural affinity) allow privileged parents to move more resources to their children compared to others (uniform or truly random allocations; Chiu, 2009). In open markets, privileged parents can use their financial capital by moving to a rich neighborhood to give their children richer schoolmates for example. They can use their social capital through cronyism, for example, by asking the vice principal to

assign their child to a course with a better teacher. Through their human capital, privileged parents can teach their children social norms and cultural capital to create greater affinity with teachers. Teachers have higher expectations of these students and often give them more attention and assistance (*cultural gatekeeping*).

School Inequality

Countries with greater family inequality often have greater school inequalities (Chiu, 2010). A system of strictly free market, private schools without national standards (e.g., no teacher certification) can magnify the impact of family inequalities as richer families pay higher tuition for their children to attend richer schools with more educational resources (teacher quality, computers, richer schoolmates, and so on). In contrast, a universal public school system with required teacher certification standards can reduce inequalities among students.

Schools with greater inequalities showed lower academic achievement among mostly poor students (Chiu, 2009). As with family inequality, school inequalities are accompanied by diminishing marginal returns, which result in lower overall achievement among poorer students. Furthermore, school resources (books, teachers, classmates) and family resources (books, siblings, parents) are mostly equivalent substitutes for one another; richer students often have a book at home that is largely equivalent to a book at school, so the latter's absence does not affect the richer student's academic achievement (*substitution effect*). Hence, fewer school resources did not reduce richer students' academic achievement.

Economic Segregation

Economic segregation is also related to academic achievement (Chiu and Khoo, 2005). The degree of economic segregation differs across school systems. In the US, schools are often funded through local taxes, so richer students often attend one set of schools in richer neighborhoods, while poorer students attend a separate set of schools in poorer neighborhoods. In contrast, rich and poor students often mix together in the same schools in other countries such as Japan and Finland. Economic segregation reduces interactions between richer and poorer students, resulting in less diverse interactions for all students. Less exposure to experiences of students in other economic strata results in less shared knowledge, less learning, and lower academic achievement among all students (*diversity effect*; Chiu, 2009).

Cultural Differences

Countries have different cultural values and the ways they address basic societal issues (Hofstede, 2003). People learn these values through both formal socialization, including direct teachings of parents and teachers, and informal socialization, such as daily exposure to social norms. Among different cultural values, two basic societal issues are particularly important to students' academic achievement, and they are the degree of hierarchy and obedience to authority vs equality (hierarchical vs egalitarian), and favoring group interests vs individual interests (collectivism vs individualism). A hierarchical nation emphasizes authority, obedience and hierarchical status (e.g., Russia), while an egalitarian country emphasizes equal status and treatment of one another (e.g., Sweden). Also, a collectivist society favors group interests (e.g., Hong Kong), while an individualistic nation favors individual interests (e.g., New Zealand).

Cultural values did not directly affect students' test scores, but they interact with the effects of family characteristics on academic achievement (Chiu, 2007). As egalitarian societies have less rigid societal status, individuals have greater economic and social mobility through their individual skills and talents. Thus, students in these countries often view academic achievement as linked to future success, so they are more likely to study harder and achieve greater academic success. Furthermore, students are more motivated to make better use of their resources in egalitarian societies, yielding stronger links between family resources and student achievement (*egalitarian motivation*). Specifically, family SES and numbers of books at home had stronger links to academic achievement in more egalitarian cultures.

Meanwhile, students in collectivist societies are more inclined to cooperate, support, and learn from their schoolmates and family members (Inglehart et al., 2004). In more collective societies, people tend to rely more on their extended family members who often live nearby, and so students are more likely to benefit from them. As extended family resources dilute the effects of immediate family resources in collective societies, the link between immediate family resources and academic achievement is weaker than in individualistic societies (*collectivist dilution*; Chiu, 2007). Specifically, single parents, family SES, resident grandparents, and birth order were all less negatively linked to test scores in more collectivist cultures.

Methodological Issues

International studies require appropriate data samples, precise test and questionnaire designs, and suitable analytic tools (see **Table 2**; Chiu, 2010). Ideally, data samples are representative of the populations, with few sampling errors and no missing data. Meanwhile, tests and questionnaires aim to maximize data validity by minimizing measurement errors and maximizing equivalence across countries. Tests seek to optimize coverage of target concepts and skills. Analyses can model multiple levels of data, cross-level moderation and mediation, and multiple outcomes.

Data Samples

To create a representative sample of a country, researchers typically categorize schools (e.g., by neighborhood SES), randomly select schools from each category, and then randomly select students from each selected school (*stratified samples*) with corresponding *weights* for each student. To reduce sampling error, researchers select several subsamples to create *plausible values*, which must be modeled in the subsequent analyses.

In such a huge data collection enterprise, missing data are typically unavoidable. Missing data can reduce estimation

Table 2 Strategies to address each analytical difficulty

Analytical difficulty	Strategy
● Representative sample	● Stratified sample
● Sampling error	● Plausible values
● Missing data	● Markov Chain Monte Carlo multiple imputation
● Measurement equivalence of test and questionnaire items	● Forward and backward translation of each item; ● Multigroup Rasch model
● Wide coverage of skills while minimizing student fatigue and learning effects	● Balanced incomplete block test
● Competence estimates from test answers	● Factor Analysis ● Item Response models
● Measurement errors on questionnaires	● Factor Analysis ● Item Response models
● Nested data of students within schools within countries	● Multilevel analysis (aka Hierarchical linear modeling)
● Cross-level moderation	● Random effects model
● Indirect, mediation effects	● Multilevel mediation tests ● Structural equation models
● False positives	● Two-stage linear step-up procedure
● Simultaneous outcomes	● Multivariate outcome models

efficiency, complicate data analyses, and bias results. Markov chain Monte Carlo multiple imputation estimates the values of the missing data, which addresses these missing data issues more effectively than deletion, mean substitution, or simple imputation.

Tests and Questionnaires

Tests and questionnaires have special designs. Test items are divided into overlapping subsets (*subtests*) in a *balanced incomplete block* test design to (1) maximize coverage of target concepts and skills, (2) reduce student fatigue, and (3) reduce test-learning effects. To reduce measurement error on questionnaires, multiple items were used for each theoretical construct (e.g., SES) to create an index that is more precise than any single variable.

To test for data validity, translations and a multigroup Rasch model are used. Each item on a test or a questionnaire is forward translated into the new language of the test by one person and then backward translated by another to check the quality of the translation. Multigroup Rasch models for each item in each country test whether the measures are equivalent across countries. Unlike factor analysis, a multigroup Rasch model has the advantages of requiring only one invariant anchor item across countries and modeling heterogeneous use of the ordinal rating scale.

To capitalize on the special designs of the tests and questionnaires to reduce measurement error, researchers use factor analyses and item response models. A factor analysis tests whether a set of test items reflects one or more underlying achievement dimensions (e.g., mathematics achievement vs reading achievement). For each construct, its test item's characteristics (difficulty, discrimination, chance) are estimated with an item response model. Some test items capture higher levels of a construct more precisely, while others capture lower levels of a construct more precisely (*item difficulty*). Likewise, each test item's precision for distinguishing each level of a construct can differ (*discrimination*). The chance parameter models the likelihood of successfully guessing a correct answer for a test item.

For questionnaire items, factor analysis and item response models are also used. A factor analysis tests whether a set of questions that appears to test similar constructs (e.g., mother's and father's years of schooling) reflects one or more underlying constructs (e.g., SES). For each construct, its questionnaire item's characteristics (difficulty, discrimination) are estimated with an item response model. Some questionnaire items capture higher levels of a construct more precisely, while others capture lower levels of a construct more precisely (*item difficulty*). Likewise, each question's precision for distinguishing each level of a construct can differ (*discrimination*).

Analysis

As the data are nested (students within schools within countries), a multilevel analysis of weighted, plausible values is needed (also known as hierarchical linear modeling). Students in the same school (or country) likely resemble one another more than students in different schools (or countries, *group heterogeneity*). Applying an ordinary least squares regression to nested data yields biased standard errors. In contrast, a multilevel analysis yields appropriate standard errors. For multilevel analyses, at least partial metric and partial scalar invariance of the measurement instruments are needed to compare latent means and regression coefficients over countries (Davidov et al., 2011).

Testing many hypotheses increases the likelihood of a false positive. To control for the false discovery rate, the two-stage linear step-up procedure is used, which outperformed 13 other methods in computer simulations.

As the data are nested, cross-level moderation and mediation tests are needed. A random effects model precisely tests whether the relationship between an explanatory variable and an outcome variable differs across contexts or interacts with another explanatory variable via random effects models. The empirical M-test corrects for potential nonnormal distributions in nested data by using multiple data simulations.

Multivariate outcome models can be used to analyze multiple outcome variables simultaneously (e.g., mathematics, reading, and science tests). Ignoring the relationships among outcome variables can result in correlated residuals that underestimate standard errors. Multilevel, multivariate outcome models address this issue to yield proper standard errors.

In short, international studies require high-quality data, precise tests and questionnaires, and appropriate analytic tools. Stratified samples, representative weights, plausible values, and Markov chain Monte Carlo multiple imputation can improve their data quality. A balanced incomplete block test design, construct indices based on sets of questionnaire items, forward and backward translations, multigroup Rasch models, factor

analyses, and item response models can help increase data validity. Lastly, multilevel, multivariate outcome analysis of weighted, plausible values; two-stage linear step-up procedure; random effects model; and empirical M-tests can be used to analyze these data.

Future Research

The availability of international data sets, their data beyond student achievement, and the potential expansion of databases portend tremendous growth in international research in the coming decade. Research on international comparisons of education is accelerating through large, international teams of scholars collecting enormous, high-quality data sets and making them freely available and easily accessible by Website downloads (e.g., www.pisa.oecd.org). Thus, scholars can use them readily both for research and teaching. The data can benefit young scholars, especially students, by providing data with which they can immediately test their hypotheses.

While most research has focused on the student test data, the extensive student questionnaire data present invaluable opportunities for analyses of other aspects of schooling aside from achievement, such as social relationships, student discipline, emotional sense of belonging at school, and so on. Furthermore, the questionnaire responses of parents, teachers, and principals can be examined as outcome variables in addition to their relationships to student outcomes.

Future research will link large-scale, cross-country data sets and create longitudinal data sets. Linked data sets combine a student achievement database with school databases or country databases to model the effects on student outcomes of school or country variables that were not available in the original student database. Longitudinal data sets follow individual students, which both control for individual student differences for more precise results and allow testing of hypotheses involving time. Altogether, these are promising developments for large-scale international research.

See also: Comparative Research in Education: Iea Studies; Cross-Cultural Study of Education; Education, Economics of; Family and Schooling; Teacher Behaviours and Student Outcomes.

Bibliography

Baker, D.P., Goesling, B., Letendre, G.K., 2002. Socioeconomic status, school quality, and national economic development: a cross-national analysis of the "Heyneman-Loxley effect" on mathematics and science achievement. Comparative Education Review 46, 291–312.

Bronfenbrenner, U., 2005. Making Human Beings Human: Bioecological Perspectives on Human Development. Sage, Thousand Oaks, CA.

Chiu, M.M., 2007. Families, economies, cultures and science achievement in 41 countries: country, school, and student level analyses. Journal of Family Psychology 21, 510–519.

Chiu, M.M., 2009. Inequalities' harmful effects on both disadvantaged and privileged students: sources, mechanisms, and strategies. Journal of Education Research 3, 109–128.

Chiu, M.M., 2010. Inequality, family, school, and mathematics achievement. Social Forces 88 (4), 1645–1676.

Chiu, M.M., Chow, B.W.-Y., 2011. Classroom discipline across 41 countries: school, economic, and cultural differences. Journal of Cross-Cultural Psychology 42, 516–533.

Chiu, M.M., Chow, B.W.-Y., 2010. Culture, motivation, and reading achievement: high school students in 41 countries. Learning and Individual Differences 20, 579–592.

Chiu, M.M., Khoo, L., 2005. Effects of resources, inequality, and privilege bias on achievement: country, school, and student level analyses. American Educational Research Journal 42, 575–603.

Chiu, M.M., McBride-Chang, C., 2010. Family and reading in 41 countries: differences across cultures and students. Scientific Studies of Reading 14, 514–543.

Chiu, M.M., Pong, S.L., Mori, I., Chow, B.W.Y., 2012. Immigrant students' cognitive and emotional engagement at school: a multilevel analysis of students in 41 countries. Journal of Youth and Adolescence 41, 1409–1425.

Davidov, E., Datler, G., Schmidt, P., Schwartz, S.H., 2011. Testing the invariance of values in the Benelux countries with the European Social Survey: accounting for ordinality. In: Davidov, E., Schmidt, P., Billie's, J. (Eds.), Cross-Cultural Analysis: Methods and Applications. Routledge, London, pp. 149–172.

Gamoran, A., Long, D.A., 2006. Equality of educational opportunity: a 40 year perspective. In: Teese, R., Lamb, S., Duru-Bellat, M. (Eds.), Education and Equity: International Perspectives on Theory and Policy. Springer, New York.

Heyneman, S.P., Loxley, W.A., 1983. The effect of primary-school quality on academic performance across twenty-nine high- and low-income countries. American Journal of Sociology 88, 1162–1194.

Hofstede, G., 2003. Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations across Nations. Sage, Thousand Oaks, CA.

Inglehart, R., Basanez, M., Diez-Medrano, J., Halman, L., Luijkx, R., 2004. Human Beliefs and Values: A Cross-Cultural Sourcebook. Siglo XXI, Mexico City.

Organization for Economic Cooperation and Development (OECD), 2010. PISA 2009 Technical Report. OECD, Paris.

Relevant Websites

International Studies of Students

www.pisa.oecd.org – OECD Program for International Student Assessment (PISA).

TIMSS & PIRLS studies of students. timss.bc.edu/. Trends in international math and science study & Progress in international reading literacy study.

Country Data

geert-hofstede.com/dimensions.html – Hofstede Cultural values.

http://www.oecd.org/pisa/pisaproducts/ – OECD statistics.

data.worldbank.org/ – World Bank Data.

Crossover Designs

Michael G Kenward, London School of Hygiene and Tropical Medicine, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In a crossover design, each unit is randomized to a sequence of experimental conditions or treatments, with the aim of comparing the effect of individual treatment assignment. When within-unit dependence of the outcome is relatively large, such designs can be much more efficient than parallel group studies with the same number of units. One potential issue with such designs is the possibility that earlier treatment responses can affect later outcomes, the so-called carryover effect. Many possible designs exist, ranging from the simple two-period two-treatment designs to those with many treatments and periods. Efficient designs can be constructed that allow adjustment for carryover effects, but only when expressed in particular, limited, mathematical forms. Data from crossover designs are examples of conventional repeated measures, and analyses can be taken from the range of tools available for these, both with continuous and discrete outcomes. Methods that perform well in small samples are essential because of the common use of such small sample sizes in the crossover context.

Introduction

In the familiar randomized parallel group design, each experimental unit, often a subject, is randomly allocated to one of the experimental interventions, or treatments. By contrast, in a crossover design, each unit is randomized to a *sequence* of treatments; a special case of a *repeated measures* design. In spite of this difference, however, the goal of a crossover study, in nearly all settings, remains the same as that of a parallel group study: to compare the effects of the *single* treatments, not the effects of the *sequences* to which the units have been randomized. Hence, in a crossover trial, units are not randomized to the interventions under comparison, and this has important implications for the design, analysis, and interpretation.

The main advantage of a crossover design, over the parallel group, is the opportunity it provides to compare the effects of treatments within-units. That is, any component of a unit's response that is consistent across time is removed from the treatment comparison. The degree of such consistency is measured by the within-unit, or intraclass, correlation. As a consequence, the precision of the treatment effect from a crossover study will typically be larger than that from a parallel group study with the same numbers of units, and will have the greatest advantage when the within-unit correlation is high. In this respect, the crossover study has much in common with a blocked or matched study in which the unit is the blocking/matching variable. A second advantage in terms of precision per unit, which is less often remarked in this context, is that each unit contributes observations under more than one treatment. So, although a simple baseline measurement can be used to remove between-unit variability in a parallel group trial, the precision will still be substantially less than the corresponding crossover trial. The main reason therefore for considering the use of a crossover design is the ability to produce sufficiently high precision with comparatively small numbers of units. Indeed, there are examples where the precision from a crossover trial is in orders of magnitude greater than a parallel group study of equivalent size.

This potential gain in precision comes with a price, however. Clearly, such a design cannot be used with treatments that irreversibly change the condition of the unit.

Once treatment has ceased, the unit must return to the original condition, at least approximately. For the same reason, such designs are only appropriate for measurements that are reasonably stable over time. It is always possible with such a design that some consequence of earlier treatment may still be influential later in the trial. In the crossover context, this is called a carryover effect. In principle, such effects may be due literally to a physical carryover; a drug is a good example of this possibility. Or, it may be due to a less direct cause such as a psychological effect of prior treatment. The key point is that units in the second and later periods of a crossover trial have systematic differences in their past experience in the trial. This potential source of bias is akin to confounding in an epidemiological study and implies that to some extent the analysis of data from a crossover trial will inevitably rely more on modeling assumptions, and less directly on the randomization, than a conventional parallel group study. This issue is particularly acute in the so-called two-period two-treatment or 2×2 or AB/BA design. This is the simplest crossover design and is particularly common in a clinical setting. In this design, each unit receives two different treatments, which are conventionally labeled as A and B. Half the units are randomized to receive A first and then, after a suitably chosen period of time, *cross over* to B. The remaining units receive B first and then cross over to A. Because this particular design is commonly used, and because it raises very special issues in its own right, the next section is devoted specifically to it. There are many other so-called *higher-order* designs, with more than two periods, and/or treatments, and/or sequences. The choice of such designs and associated analyses are considered in the remaining sections. Much of what follows is elaborated in two standard texts on crossover trials, Senn (2002) and Jones and Kenward (2014), and we refer the reader to these for further details.

The Two-Period Two-Treatment Design

In the two-period two-treatment, or 2×2, design each unit is randomized to one of two treatment sequences A then B, or B then A. The term period is used for the time interval in which the unit receives a particular treatment, hence this is a two-

period design. In a clinical setting, there is often an additional period between the two treatment periods in which units do not take either treatment, known as the washout period. Its purpose is largely to reduce the chance of bias due to carryover. We return below to such concerns, as well as to the incorporation into the analysis of measurements made in the washout period. To begin with, it is assumed that there is no carryover effect present, and no treatment-by-period interaction, and the focus is on the use of the measurements made only in the treatment periods. To represent these, we introduce some notation. Let Y_{ijk} denote the outcome from unit k, $(k = 1, \ldots, n_i)$ on sequence i, $(i = 1, 2)$ in period j, $(j = 1, 2)$. It is assumed that n_i units are randomized to the ith sequence. This outcome might, for example, be a summary measure from all or part of the relevant treatment period or a measure from a particular time, often the end, of the period. When the treatment effect is estimated from these, it is necessary to acknowledge the repeated measures structure of the design. This implies that it is not possible to rule out a priori an overall change in response between the two periods: the so-called period effect. Hence on design considerations alone, it is necessary to adjust for a period effect in the analysis. It is the requirement for this adjustment that means that, in general, any crossover design must have at least as many sequences as treatments.

Fortunately, the adjustment is very simple to do in practice, particularly so in the 2×2 design. Denote the averages from the four sequence-by-period means by $\overline{Y}_{ij}$. The B–A mean treatment difference, adjusted for period effect (τ say) is then estimated as

$$\widehat{\tau} = \frac{1}{2}\left(\overline{Y}_{12} + \overline{Y}_{21} - \overline{Y}_{11} - \overline{Y}_{22}\right) \qquad [1]$$

To make inferences about τ we note that, if we define the within-unit differences $D_{ik} = Y_{i2k} - Y_{i1k}$, then $\widehat{\tau}$ is simply the mean difference of these between the two sequence groups:

$$\widehat{\tau} = \frac{1}{2}\left(\overline{D}_2 - \overline{D}_1\right)$$

For inferences therefore standard two-sample procedures can be applied to the differences D_{ik}. Most commonly, the standard two-sample t-test framework is used to provide test statistics, confidence intervals, and power calculations. For the latter, it must be remembered that the relevant variance is that of the within-unit differences D_{ik}, not the raw observations Y_{ijk}. Alternatively, nonparametric procedures can be applied to the same differences.

It can therefore be seen that, with complete data, under the assumption of no carryover or treatment-by-period interaction, the analysis for the 2×2 design is a very straightforward application of a two-sample procedure applied to the within-unit differences. However, there has long been concern in this setting about the carryover assumption. If there is such a carryover effect, the simple estimator in [1] will be biased for τ. An early and influential paper, Grizzle (1965), made the suggestion that the analysis should therefore begin with a test for the presence of such a carryover and, only if nonsignificant, should the estimator in [1] be used. If found significant, on the other hand, then the treatment effect should be estimated from the first period data only (i.e., $\overline{Y}_{21} - \overline{Y}_{11}$). This two-stage procedure was widely adopted. However, it suffers from two very serious problems, which mean, in practice, it should be avoided.

The first problem lies with the power of the initial test for carryover, which, in the 2×2 trial, turns out to be equivalent to a test for treatment-by-period interaction. This test parallels that for the direct treatment effect above, but is calculated using the unit sums $S_{ik} = Y_{i1k} + Y_{i2k}$ in place of the differences D_{ik}, and there is no factor of 1/2. Suppose that, an observation Y_{ijk} has variance σ^2, and the within-unit correlation is equal to ρ, then $\mathrm{V}(D_{ik}) = (1-\rho)\sigma^2$ and $\mathrm{V}(S_{ik}) = (1+\rho)\sigma^2$. It follows that the ratio of the variances of the carryover estimator and the direct treatment estimator is

$$R = 2 + \frac{\rho}{1-\rho} \qquad [2]$$

As the within-unit correlation increases, this ratio grows without limit. In a typical crossover setting, the within-unit correlation is large, and the trial is powered for the direct treatment effect, leading to comparatively small numbers of units. The ratio [2] grows without bound as ρ increases, and this means that the variance of the carryover estimator can become very large in a typical trial. Consequently, the power for the carryover effect is often very low indeed. As an example, consider the 2×2 crossover trial on treatment for asthma described in Jones and Kenward (2014, Section 2.10). This had 17 units with the outcome, a lung-function measure: forced expired volume in 1 s. In this trial, ρ was estimated as 0.96, with corresponding R equal to 28.6. As a consequence, 28.6 times as many units (476) would have been required to estimate the carryover effect with the same precision as the direct treatment effect and, in terms of power in the actual sample, the test for carryover is effectively useless. Because of the typical low power of the test for carryover, it was suggested in the original two-stage proposal (Grizzle, 1965) that this might be made at a higher nominal size, such as 10%. However, any such modification is insufficient in most cases to rescue this test and negative results from such low-powered tests convey no useful information.

The second major problem with the two-stage procedure was pointed out by Freeman (1989). Because of the statistical dependence between the preliminary test for carryover and the first-period treatment comparison, the two-stage procedure leads to both bias in the resultant treatment estimator and increases the probability of making a Type I error. In other words, the actual significance level of the direct treatment test is higher than the nominal one chosen for the test. The first-period comparison would only be unbiased were it always used, or chosen for a reason independent of it. Unfortunately, this is not the case in the two-stage procedure, in which it is only used following a significant (hence large) carryover. Although attempts have been made to circumvent this problem, there is no solution that does not involve introducing information about the carryover effect that is not contained in the data. Given that a crossover trial will be powered for within-unit estimation of the treatment effect and so will not have the sensitivity to provide useful information on the carryover, it has been argued that the best approach to this problem is to take all practical steps to avoid possible carryover, such as using washout periods of adequate length between the treatment periods, and then assume that this effect is negligible and use the simple estimator in [1] and the associated analysis. This in turn requires a good working knowledge of the treatment

effects, which will most likely be based on prior knowledge of, and/or the mechanisms behind, these.

Many 2×2 trials have additional measurements taken before the first treatment period, at the end of the so-called run-in period and at the end of the washout period. These are called baseline measurements, noting that, in contrast to a parallel group design, such baseline measurements (the ones from the washout period) can be collected after randomization. Although it is tempting to introduce these baselines into the analysis in a fashion analogous to a parallel group design, for example using change from baseline as response and/or with baselines as covariates, some care is needed in their introduction. Because of the within-unit nature of the primary analysis, these baselines do not have the usual role familiar from parallel group studies. Between-unit variability as existing in the treatment periods is already eliminated from the treatment comparison and for the baselines to make a contribution over and above this, they must explain additional variability. This will happen only when the correlation between an outcome Y_{ijk} and its associated baseline is considerably greater than between the two outcomes Y_{i1k} and Y_{i2k}. This might happen, for example, if the time interval between the treatment measurements is much greater than between baseline and associated treatment measurement. On the other hand, if the correlations among the four measurements from a unit are all similar, then not only will no improvement in precision be obtained from using baselines in the analysis, but also their introduction is likely to actually *decrease* precision. For example, if all within-unit correlations are exactly equal, then the variance of the treatment estimator obtained using the change from baseline as response is exactly double that of the estimator [1] that uses only the treatment outcomes.

The baselines have much greater potential value when estimating the carryover effect, because they can be used to eliminate between-unit variability from this. For further details, we refer to Jones and Kenward (2014: Section 2.11). A further and important issue arises with the use of baselines in crossover designs that applies both to the 2×2 and higher-order designs. This is discussed below when model-based analyses are considered.

The very simple analyses so far described rest on the assumption that the data are continuous. The analogue of the two-sample t-test of the differences D_{ik} for binary data is given by the Mainland–Gart test, which is a simple extension of McNemar's test that incorporates adjustment for a possible period effect. The Mainland–Gart test is just that for association in a 2×2 contingency table that is constructed as follows. When the outcome is binary, each unit presents one of four possible responses for the two periods (Y_{i1k}, Y_{i2k}): (0,0), (0,1), (1,0), and (1,1). The so-called nonpreference responses (0,0) and (1,1) are discarded and the columns of the table are correspond to the *preference* responses (0,1) and (1,0). The rows are then given by the two sequence groups:

	Outcome	
Sequence	*(0,1)*	*(1,0)*
AB	n_{11}	n_{12}
BA	n_{21}	n_{22}

In this table, n_{11} is, for example, the number of units who responded with a 0 in period 1 and a 1 in period 2 in sequence group AB. The test for treatment effect, ignoring carryover, is then the test for association in table; for example Pearson's chi-squared statistic. If numbers in this table are small, as they often are in a crossover trial, especially after discarding the (0,0) and (1,1) responses, then a small-sample exact procedure such as Fisher's can be used. Other outcomes, such as ordinal and count are best handled using appropriate model-based methods, to which we return briefly below.

Finally, it should be noted that the simplicity of the analyses met so far rests on the completeness of the data. If units drop out of the trial, then there will be left some singleton measurements. The decision may be made leaving these out of the analysis, and use the same simple procedures with the completers. In terms of precision, if the within-unit correlation is high, then these singleton measurements contribute very little to the direct treatment effect and little is lost through their omission. On the other hand, to use the singletons in the analysis, a model-based analysis is needed that can combine within- and between-unit information. Suitable tools for this are met below.

Higher-Order Designs

As argued above, that, when using the 2×2 crossover design, it is best to assume that any carryover effect is negligible. This is because, in nearly all examples of its use, the design provides insufficient information on such a carryover to allow either adjustment for it or its elimination. Several higher-order designs for two treatments have been proposed that attempt to circumvent this problem. However, to be effective, these require strong assumptions about the nature of the carryover effect. These designs have extra sequences and/or extra periods. For example, extending the design to three periods means that the following three alternative designs with two sequences might be considered: ABB/BAA, ABA/BAB, or AAB/BAA. In the absence of a carryover effect, these designs are effectively equivalent for estimating the direct treatment effect. However, more importantly, they also allow the treatment effect to be estimated using within-unit information after adjustment for a particular form of simple carryover effect, something that is impossible with the 2×2 design. In these circumstances, the three designs behave rather differently in terms of precision, and it turns out that the first design, ABB/BAA, is most efficient among the three. At first sight, this appears to provide a solution to the main problem associated with the 2×2 design. However, this advantage depends on strong and potentially unrealistic assumptions about the form of the carryover: namely that the carryover effect is the same whether carryover is occurring into A or into B. This is an example of a general problem with the design of higher-order crossover design: many arrangements are possible and some can be shown to have particularly good mathematical properties under specific modeling assumptions (especially concerning carryover). But these assumptions themselves are not usually open to sufficiently precise assessment from the data generated by the design, so the same general issue that surrounds the 2×2 design is still present, albeit in a more complicated and less

accessible form. For this reason, great care is needed in interpreting the term 'optimal design' in this setting.

The higher-order two-treatment designs can be extended in a number of ways with more sequences and periods, allowing further effects, such as treatment-by-period interaction, second-order carryover, and carryover-by-treatment interaction to be estimated. They are not widely used in practice, however, beyond very specific application areas.

For a design with more than two treatments (t say) it is usual, where practical, for each unit to receive all treatments, implying that the number of periods is equal to t. In addition, because adjustment for period effects is the norm, it is important for reasons of efficiency to balance treatments within periods. A design in which this balance is exact is called a Latin square design.

In such a design, every unit receives each treatment and each treatment occurs equally often in each period. This is an example with four sequences, four treatments (A, B, C, and, D), and four periods:

	Period			
Sequence	*1*	*2*	*3*	*4*
1	A	B	C	D
2	D	A	B	C
3	C	D	A	B
4	B	C	D	A

A design based on this scheme would have a multiple of four units in total with one quarter randomly assigned to each sequence. Provided carryover is not a concern, any such square would be equally efficient for estimation of the direct treatment effects, and similar arrangements can be constructed with any number of treatments.

If, on the other hand, adjustment is to be made for carryover, then this design is very poor. It can be seen that each treatment is preceded by the *same* treatment in the three sequences in which it does not occur first. This means that treatment and carryover effects are highly nonorthogonal and adjustment causes a severe loss in precision. This can be largely resolved by choosing a special Latin square arrangement in which every treatment is preceded equally often by each other treatment. For example, in the four treatment case we might consider the following design:

	Period			
Sequence	*1*	*2*	*3*	*4*
1	A	B	C	D
2	B	D	A	C
3	C	A	D	B
4	D	C	B	A

Latin square designs with such balance are called Williams square designs (Williams, 1949). In fact, they are only square when t is even; for t odd, a Williams square design has $2t$ sequences. Examples of Williams squares designs for a range of values of t are given in Jones and Kenward (2014: Chapter 4). Even if adjustment for carryover is not to be used in the analysis, these designs have the advantage of having a degree of robustness to carryover: an unadjusted analysis will produce less-biased estimates of the direct treatment effects even when carryover is present in the data compared with the earlier Latin square design. However, as above, these remarks refer strictly to a specific form of carryover: simple first order. Many alternative forms are conceivable, and a design that is chosen with one type in mind may be poor for another.

Not all higher-order crossover designs are appropriate for undifferentiated sets of treatments, as are the Williams squares. There are particular designs for comparisons with a control, and for factorial designs. Details of these and other special designs can be found in Jones and Kenward (2014: Chapter 4). It is a good choice where possible to have each unit receive every treatment in a crossover trial, as in a Williams square. Designs for which this is not true are sometimes called incomplete designs, and these are generally less efficient.

Model-Based Analyses

The simple two-sample analyses that were used above with complete data from the 2×2 design need generalization for other designs and for incomplete data. Because the data from crossover trials are examples of repeated measures data, we can draw on the body of analysis tools available for this setting (e.g., Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005). Accordingly, with continuous data, linear regression models for dependent data can be used and, for discrete data, the appropriate analogues of these. In many settings with continuous data, we can in fact use very simple, ordinary, regression models by analogy with randomized block analyses. With crossover trials, one normally expects fairly stable responses over time, and it is often sufficient therefore to capture the within-unit dependence through the introduction into a linear model of a simple unit effect. This effect can be treated as fixed or random in the analysis. To understand the implications of choosing one or other of these, we need to consider the source of the information on treatment effects in the given crossover design. It is natural and convenient to divide this information into two independent components: the between- and within-unit information. The between-unit information is that provided by the sums (or means) of all the observations from each unit. By contrast, the within-unit information is that contained among all differences of observations from each unit. The information available from these two sources depends on both the design and the model fitted. This division of information is also dependent on the size of the within-unit correlation: as this correlation increases, the comparative variability of the between-unit sums grows. In terms of models, in many settings, a very simple, or minimal, form will suffice. This will include, in addition to the unit effects, categorical effects for period and treatment. On occasion, this will be augmented by carryover effects and, more rarely, other terms. In a fully balanced Latin square design, with the minimal model, all information on treatments is contained within-units. When some form of imbalance is introduced, due to, for example, dropout, the introduction of additional model

terms, or the use of an incomplete design, then some information on treatment effects will 'move' to the between-units stratum.

We are now in a position to consider the choice of fixed or random unit effects. The use of fixed unit effects, that is, treating the unit effects as any other linear-model parameter, means that all between-unit information is discarded. However, the resulting model is an example of a standard linear regression model and all the familiar regression procedures are available, including exact small sample *F*- and *t*-based inferences. Moreover, in such analyses, no assumption needs to be made about the distribution of the unit effects implying an additional degree of robustness in these analyses. They can be seen as crossover versions of conventional randomized block analyses in which units are blocks. There are many advantages to keeping the crossover analysis within this framework, not least the simplicity. So this should be considered the approach of first choice, provided it is appropriate, that is, when there is negligible between-unit information. In contrast, the use of random unit effects allows the recovery of the between-unit information, but this moves the analysis framework to that of the linear mixed model. This is less robust that the fixed effects model, because of the need to assume some distribution for the unit effects and, in addition, because small-sample inferences must rely on approximate methods (Kenward and Roger, 1997; Skene and Kenward, 2010). Hence this approach should be used only when strictly necessary, that is, when there is nonnegligible between-unit information is available for recovery, with sufficiently large-sample sizes. For the reasons given above, this will not be the case in the typical efficient, complete, crossover design with high within-unit correlation. Hence, careful consideration needs to be given to the data, design, and analysis model, before choosing the random unit effects approach. In specific settings, this approach has benefits, but these tend, in practice, to be the minority.

Another important distinction arises between the use of fixed and random effects when baselines are incorporated into the analysis. It is shown in Kenward and Roger (2010) that if random unit effects are used, then to avoid the so-called cross-level bias, an additional covariate must be included that consists of the sum of the baselines over all periods. With fixed unit effects, this is unnecessary.

Similar issues apply for discrete data, with some important distinctions. First, asymptotic arguments become the norm with the discrete data analogues of these analyses. This implies, among other things, that the fixed unit model cannot be used in the same way as above, by simply estimating an effect for each unit. Instead, the so-called conditional likelihood analyses are the proper analogues of these and are available only for certain models, namely logistic regression for a binary outcome, multivariate logistic regression for a nominal outcome, and log-linear regression for a Poisson (count) outcome (Jones and Kenward, 2014: Sections 6.2, 6.4, and 6.5.1, respectively). There is no conditional likelihood analogy for a proportional odds model for an ordinal outcome. Random effects models can be used however with all such classes.

A second important distinction concerns the variety of classes of model available for nonnormal outcomes, in contrast to the almost automatic use of the linear regression basis of models for continuous data. This is largely a consequence of the nonlinearity in the link between the model effects such as treatment and period and the mean outcome, which is necessary in such nonnormal settings. The distinction between two important classes of model, the so-called marginal and unit-specific, is especially relevant in the repeated measures setting and needs to be fully appreciated when analyzing discrete crossover data. These classes are discussed in the crossover setting in Jones and Kenward (2014: Chapter 6) and in a more general repeated measures setting in Molenberghs and Verbeke (2005). A brief consideration other types of outcome, such as event time, is given in Jones and Kenward (2014: Section 6.5).

See also: Analysis of Variance and Generalized Linear Models; Experimental Design: Overview; Generalized Linear Mixed Models; Hierarchical Models: Random and Fixed Effects; Linear Mixed Models; Optimal Experimental Design.

Bibliography

Freeman, P., 1989. The performance of the two-stage analysis of two treatment, two period crossover trials. Statistics in Medicine 8, 1421–1432.

Grizzle, J.E., 1965. The two-period change-over design and its use in clinical trials. Biometrics 21, 467–480.

Jones, B., Kenward, M.G., 2014. Design and Analysis of Crossover Trials, third ed. Chapman & Hall/CRC, London.

Kenward, M.G., Roger, J.H., 1997. Small sample inference for fixed effects estimators from restricted maximum likelihood. Biometrics 53, 983–997.

Kenward, M.G., Roger, J.H., 2010. The use of baseline covariates in cross-over studies. Biostatistics 11, 1–17.

Molenberghs, G., Verbeke, G., 2005. Models for Discrete Longitudinal Data. Springer, New York.

Senn, S.J., 2002. Crossover Trials in Clinical Research, second ed. Wiley, Chichester.

Skene, S., Kenward, M.G., 2010. The analysis of very small samples of repeated measurements. II: a modified box correction. Statistics in Medicine 29, 2838–2856.

Verbeke, G., Molenberghs, G., 2000. Linear Mixed Models for Longitudinal Data. Springer, New York.

Williams, E.J., 1949. Experimental designs balanced for the estimation of residual effects of treatments. Australian Journal of Scientific Research 2, 149–168.

Cross-Sectional Methodologies in Developmental Psychology

Victoria CP Knowland, Division of Language and Communication Science, City University London, London, UK
Harry Purser, Kingston University, London, UK
Michael SC Thomas, Birkbeck College, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cross-sectional methodologies are typically adopted in the field of developmental psychology when assessing the impact of developmental age or stage on some cognitive measure. Here, we describe the statistical techniques that form the foundation of cross-sectional studies and illustrate their use through recent examples from the literature. We present both behavioral and brain-imaging work and emphasize the value of cross-sectional designs, with particular reference to studies that explore atypical development and consider change across the life span.

Introduction

A *cross-sectional* experimental design is one that involves taking a snapshot of information at a given point in time. In the context of developmental psychology, this broad approach is typically utilized to assess a cross-section of development, with a variable of interest being studied in children at different ages. For example, a study might measure how children's ability to perform on a memory task varies according to age, while keeping other factors such as socioeconomic status (SES) as consistent as possible. Although usually a design might examine performance across age, it could be across any continuous variable (height, intelligence, or time of day). For example, the measurement of interest could be performance on the memory task as it varies by SES, while age is kept consistent. However, generally speaking, cross-sectional designs in this branch of psychology refer to studies that take age or stage of development to be the continuously varying, predictor measure. In this chapter, we outline the statistical basis and value of cross-sectional designs for developmental psychology, and draw out the limitations and challenges inherent in them. We take specific examples from recent research in the field to illustrate the methodology, each of which takes data collected at a single point in time to understand the processes of change in cognitive systems.

The first systematic analysis of cognitive development arguably came not long after the advent of experimental psychology, when Alfred Binet attempted to measure average cognitive functioning in the domains of sensorimotor processing, language, memory, and logic in children between 6 and 15 years of age. This work resulted in the publication of the first intelligence test in 1905 (Binet and Simon, 1905), and is an excellent example of the early adoption of a cross-sectional methodology. Developmental psychology as a field in its own right did not really gather pace, however, until Jean Piaget's work from the 1930s onward. Piaget made intricate observations of his own children and used them as a basis for his hypothesis that cognitive development is staged and hierarchical (e.g., see Piaget, 1936). His work sparked an explosion of studies addressing cognitive development and, in particular, the underlying mechanisms of change. Developmental psychology is increasingly now thought of as the study of change in cognitive systems, regardless of age; development is a lifelong process.

Variation over developmental time can be recorded in one of two ways: either by studying individuals at different stages of development at one point in time, as discussed here, or by following the same set of individuals over multiple points in time. This latter, longitudinal approach is discussed elsewhere in the current volume. The establishment of statistical measures such as correlation and linear regression, based on the influential work of Karl Pearson at the turn of the twentieth century (see Pearson, 1896), allowed for the formalization of theoretical notions of development. Indeed, theory has driven, and has been driven by, the advance of statistical methods in every area of psychology. Taking once more the example of intelligence research, the establishment of modern notions of the structure of cognition went hand in hand with the development of the statistical technique of factor analysis (see Spearman, 1904). In the remainder of this article, we discuss the statistical measures that have been developed in parallel with the theory and practice of cross-sectional designs in developmental psychology. The addition of each statistical technique will allow us to elaborate from the basic concept of cross-sectional studies to a more complex and powerful set of methodologies. We begin with the roles of correlation and regression.

Correlation

Correlation describes the strength and direction of linear dependence between two variables. The statistic used to describe the relationship is most typically Pearson's r, which is a measure of the covariance of the two variables divided by the product of their standard deviation. The obtained value ranges from -1.0 to 1.0, from a perfect negative correlation, through no dependence between the variables, to a perfect positive correlation. As a measure of the strength of a relationship, r is used as an effect size. With respect to developmental psychology, correlations are frequently used to analyze the relationship either between age and performance on a cognitive measure, or between two cognitive measures at different ages. Cross-sectional designs lend

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23235-2

themselves well to correlational analysis as the predictor variable tends to be continuous and have a wide range. Here, we will explore some of the applications of correlational analyses in developmental psychology in the context of the relationship between month of birth and academic performance.

Throughout primary and secondary school, there is a significant correlation between children's performance on formal academic tests and their month of birth. This relationship has been established by running cross-sectional studies looking at the outcomes of national curriculum tests sat at ages 7, 11, 14, and 16 in the United Kingdom (e.g., Crawford et al., 2010). Children who are born at the end of the academic year tend to have lower educational attainment than children born at the start of the academic year. Equivalent relationships have been repeatedly found around the world, including in the United States (Elder and Lubotsky, 2008). Month of birth therefore has long-term implications for children's academic achievement and life outcomes as, among other things, it impacts who is likely to finish school and thereby find employment. Readers interested in the details of this relationship and what drives it are directed to Crawford et al. (2013); here, we use it to demonstrate some of the key principles of correlation.

A correlational analysis allows us to probe the nature of the relationship between month of birth and academic achievement. One important question is which aspect of month of birth drives, or mediates, the relationship. The two prime candidate factors are age at starting school and age at which the tests are sat. As these two factors are themselves not perfectly correlated in the United Kingdom, Crawford and colleagues (Crawford et al., 2010) were able to separate out the impact of each. By controlling for each in turn, the authors found that the relationship between month of birth and academic test score is largely driven by, or mediated by, the age at which the test is sat. For the conditions for mediation, see Baron and Kenny (1986) and Holmbeck (1997). Another question that can be answered through a correlational analysis is whether variables exist that impact, or moderate, the strength of the correlation under investigation. In the case presented here, multiple factors could theoretically be moderators. For example, the relationship might ameliorate as children get older, such that age acts as a moderator. In actual fact, the effect of month of birth on academic achievement does lessen over time, but it remains statistically significant to the point of college entry. One paper has found that gender is another moderating variable, with exam results at age 16 from children in the United Kingdom showing that boys born in the summer had the greatest disadvantage and girls in the autumn the greatest advantage (Sharp, 1995).

Correlation, then, can be a powerful tool to establish a relationship between two variables. This method of analysis does have considerable limitations, however, including the assumption of linearity (for nonlinear relationships, growth-curve modeling is more suitable). What correlation cannot tell us is whether the relationship between two variables is causal. This is a difficult problem to overcome without either a longitudinal data set to run time-lagged correlations, longitudinal regression, or the experimental manipulation of variables.

Regression

Simple regression determines the extent to which a value of the dependent (outcome) variable can be predicted based on the predictor variable. This technique differs from correlation in that it tacitly assumes a directional causal relationship between the predictor and outcome variable. The distinction is perhaps clearest when age and cognitive task performance are considered: increasing age indirectly leads to improvements on cognitive tasks, but improvements on cognitive tasks cannot lead to augmentation of age. It is worth noting that changes in chronological age do not directly cause improvements in task performance; age is associated with maturation and experience-dependent learning, which, as aspects of cognitive development per se, may be considered more legitimate direct causes of task performance improvement.

It is common for researchers to use performance on one task to predict performance on another. Note that, according to the logic discussed here, such researchers are tacitly stating that the predictor variable causally determines the outcome variable to an extent. For example, Purser et al. (2012) investigated whether measures of components of Baddeley's (1986) model of working memory predicted the route-learning ability (acquiring knowledge about routes through space) of typically developing children aged 5–11 years. Baddeley's model features both verbal and visuospatial short-term storage components, and a 'central executive' that is concerned with controlling attention (among other things). Verbal short-term memory was assessed with digit span, a task in which participants must repeat back a list of spoken numbers in serial order; visuospatial short-term memory was indexed by Corsi span (Corsi, 1972), in which the participant attempts to reproduce a sequence of spatial locations. The 'executive' component was measured with the Go/No Go task, in which a pseudo-random series of differently colored circles is presented on a computer; participants must press a key as quickly as possible on seeing each circle, unless it is red, in which case they should refrain from pressing the key. Route learning was assessed by the number of errors made in the course of learning a route through a virtual environment maze.

A series of linear regressions indicated that the measures of all three model components – verbal and visuospatial short-term memory and the central executive – were statistically significant predictors of children's route-learning ability. However, one would expect every cognitive function tested to improve with age in this cross-sectional sample and hence be intercorrelated, which was indeed the case. Stepwise multiple regression was therefore used to investigate the independent contributions of each cognitive component to children's route learning.

In forward stepwise multiple regression, predictors are added one by one to the regression model. Due to the fact that multiple regression tests for the unique variance in an outcome variable explained by each predictor, it is important to enter predictor variables in a theory-sensitive manner for this kind of analysis. Both memory tasks must have involved some degree of attentional control, because the stimuli could not be recalled if they were not attended to. The executive task, however, had no clear short-term storage demands. Therefore, the executive task was entered as the first predictor, accounting for

a significant proportion of variance in route learning (40%). Neither of the storage tasks contributed significant additional predictive power, suggesting that their predictive relationships with route learning, as discussed in this section, were mediated by the executive control demands of the tasks.

Despite the tacit assumptions made when using regression, it is actually very hard to establish causality. The oft-used phrase 'correlation does not equal causation' should be extended to 'neither correlation nor regression equals causation.' Using a regression model presupposes a causal relationship between the predictor and the outcome variable, but it cannot establish it. Longitudinal methods are better suited to test such hypotheses.

Matching

Matching is the equating of groups on some variable – usually chronological or mental age – to afford a meaningful comparison. It is frequently utilized in developmental disorder research, whether cross-sectional or longitudinal, with the aim of discovering whether a group with a disorder is above or below the level of task performance expected for their age or for their ability in some domain(s). The control group, then, acts a reference point for the disorder group, in order to rule out candidate explanations for any resulting group differences. Matching has become controversial, due to the fact that it ignores variability in the matching variable and is not developmental in its emphasis (see Thomas et al., 2009).

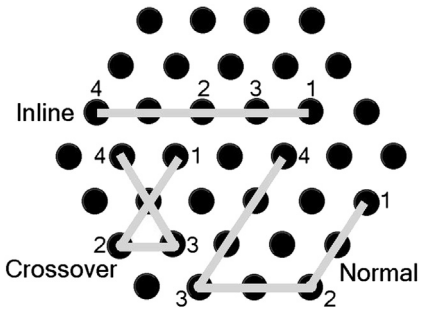

Figure 1 The task display showing examples of the paths in each condition of the experiment (the order of presentation is illustrated in the figure, but no digits were actually shown on the display; although path lengths were matched across conditions, a shorter Crossover path is shown for the sake of overall clarity).

The implicit logic behind group matching, noted by Jarrold and Brock (2004), is that matching will equate for 'noncentral' task demands: understanding instructions, controlling response behavior, selecting and using appropriate strategies, and so on. However, the mental age measures most commonly used for matching are very indirect means of achieving this: the Peabody Picture Vocabulary Test (PPVT; Dunn and Dunn, 1997) and Raven's Colored Progressive Matrices (RCPM; Raven et al., 1998) are measures of receptive vocabulary and nonverbal reasoning, respectively. Jarrold and Brock, instead, advocate control conditions that are essentially the same as the experimental task, differing only in that the target cognitive ability is not required for successful performance.

Purser (2006) investigated whether individuals with Down syndrome (DS) rely on a visual strategy to support their visuospatial short-term recall, relative to a typically developing control group. There were three conditions, presented on a honeycomb-like grid on a computer touchscreen: 'Normal' trials on which the path traced by the visuospatial sequence could be represented as a regular four-sided shape, 'Crossover' trials on which the path crossed over itself once, and 'Inline' trials on which the path fell on a single line, so that no two-dimensional shape could be represented (see **Figure 1**). The sequences were presented by circles momentarily changing color, after which the participant attempted to touch the circles in correct serial order.

Participants from DS and typically developing groups were matched on the Normal version of the task, ensuring that the two groups were matched for general factors related to successful task performance in the other two conditions. **Figure 2**a shows each group's average recall over the three conditions. The DS group was significantly poorer on the Inline version of the task than the TD group. An error analysis (**Figure 2**b) indicated that the DS group made more order errors than the TD group in both the Crossover and Inline conditions. These results indicated that the DS group found path crossing more detrimental to recall than the TD group,

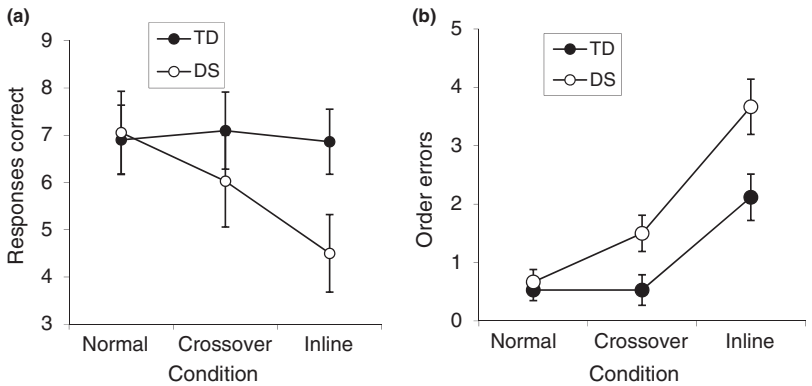

Figure 2 (a) Condition effects by group in the visuospatial recall task. Vertical lines depict standard errors of the means. Maximum score = 16. TD = typically developing; DS = Down syndrome. (b) Condition effects by group for order errors in the recall task. Vertical lines depict standard errors of the means.

consistent with relying on a visual strategy. Importantly, due to the 'task-matching' method, these differences cannot be attributed to general factors differentially affecting the groups.

Trajectory Analyses

Trajectory analyses (Thomas et al., 2009) are essentially modified forms of analysis of variance (ANOVA). Instead of comparing group means, however, the analyses involve the comparison of regression lines, or 'developmental trajectories,' between groups, across conditions, or both. Trajectories are linear functions that vary in terms of both their gradients (rates of change) and intercepts (initial levels of performance).

Trajectories are generally used to relate task performance to either chronological or mental age; they are especially useful for investigating the developmental relations that exist within developmental disorders that show uneven cognitive profiles, or *developmental dissociations*. Although longitudinal methods would ideally be used for such investigations, the cross-sectional approach can give an approximation of developmental trajectories, which can subsequently be validated by longitudinal research.

Trajectories help to answer the question 'Do individuals with a disorder perform at an age-appropriate level?' In a simple example of the trajectories approach, Purser and colleagues (Purser et al., 2011) compared word knowledge and vocabulary age in the rare genetic disorder Williams syndrome (WS), and a typically developing (TD) control group. In individuals with Williams syndrome, language development can appear a relative strength, with language level exceeding overall mental age, but the exact nature of these language abilities has been the subject of much research. In Purser et al.'s study, word knowledge was assessed with a definitions task (in which participants are asked to define words; e.g., 'What is an elephant?'). Vocabulary age was assessed via the British Picture Vocabulary Scale (BPVS; Dunn et al., 1997). **Figure 3**a shows the two groups' performance on the definitions task: the WS group's performance began at a level appropriate for vocabulary age, but then the TD group improved at a faster rate than the WS group. The gradients of the trajectories differed but not the intercepts.

Participants also completed a categorization task, which was also a measure of word knowledge but avoided some of the metacognitive demands of the definitions task, such as understanding what a definition is and how to respond appropriately. In the categorization task (**Figure 3**b), the performances of both groups developed at similar rates, but the WS group was markedly poorer than the TD group, on average, than predicted by vocabulary age. Here, the gradients did not differ, but the intercepts did.

Taken together, these results indicated that individuals with WS have poorer word knowledge than predicted by their vocabulary age, but this word knowledge improves at a similar rate with increasing vocabulary age in both WS and typical development. The 'falling behind' of the WS group on the definitions task, relative to the TD group with advancing vocabulary age, was likely due to older TD children understanding the metacognitive aspects of the definitions task better than participants with WS of a similar vocabulary age.

Brain Imaging

Although all the data we have considered thus far have related to behavioral measures, researchers are increasingly turning to brain-imaging techniques such as functional and structural magnetic resonance imaging (MRI) (see Holland et al., 2007; Knickmeyer et al., 2008, for examples of functional and structural studies, respectively) and electroencephalography (EEG) (e.g., Thatcher et al., 2008) to inform developmental theory. Such techniques allow an examination of not just how behavior and cognition change over development, but also how the brain changes, and how the relationship between the brain and behavior may alter, too. Questions being addressed include the following: what are the typical functional relationships between different brain regions over development? How do functional brain activity and the structure of the brain relate to the development of specific skills over childhood, and indeed the lifespan? What can differences in the brain tell us about why behavior is atypical

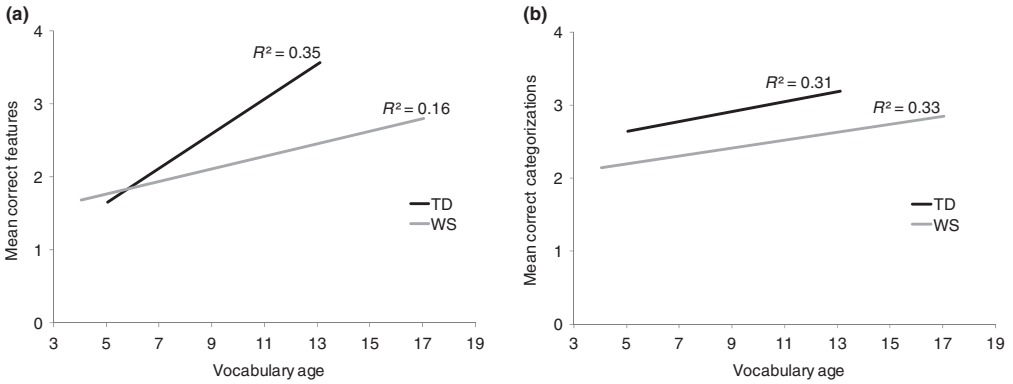

Figure 3 (a) Mean number of features given by participants in the definitions task plotted against vocabulary age in years. (b) Mean number of correct categorizations plotted against verbal mental age in years. WS = Williams syndrome; TD = typically developing. Data from Purser, H.R.M., Thomas, M.S.C., Snoxall, S., Mareschal, D., Karmiloff-Smith, A., 2011. Definitions versus categorization: assessing the development of lexicosemantic knowledge in Williams syndrome. International Journal of Language & Communication Disorders 46 (3), 361–373.

in children with developmental disorders such as dyslexia? The questions addressed using brain imaging, and the data acquired, are complex and require careful thought during the design of studies and the interpretation of results. This is especially true in the light of the many-to-many relationships between brain and behavioral measures. Individual behaviors are generated by networks of brain regions working together, and an individual brain region may be involved in generating more than one behavior. Moreover, a developmental perspective must take into account that the experienced environment may continually change over the lifespan.

While adopting brain-imaging techniques permits some key insights into the development of cognition, using these techniques with children also raises practical and theoretical issues (see Davidson et al., 2003, for a discussion of the use of functional MRI (fMRI) techniques with children). The majority of cross-sectional studies in developmental psychology require a wide age range in order to capture change over development. This presents a challenge in that the paradigm must be suitable for a range of ages, and it must be sensitive enough to measure performance at each age. Being sensitive across the age range requires that no set of participants performs at either floor or ceiling on any task. With brain imaging, this challenge is compounded by changes in physical properties such as skull thickness and brain size, as well as the propensity for young children to move about during testing. All of these factors, and many more besides, influence the quality of the imaging signal and show both individual differences and age effects. Although these issues are always relevant when working with a developmental sample, they are easier to take into account if testing the same children repeatedly, that is, if using a longitudinal design. Nevertheless, cross-sectional studies of development involving brain imaging are relatively sparse from ages 2 to 6 years, where the practical challenges of obtaining measurement are most severe.

Cross-sectional techniques can be applied to any brain-imaging study, just as they can to any behavioral study. Here we take an example of a lifespan study, emphasizing that the study of developmental processes considers trajectories beyond the end of childhood. Richardson et al. (2010) gathered receptive vocabulary and nonverbal IQ scores for 47 individuals aged between 7 and 73. These participants were then scanned using MRI to look in detail at the structure of their brains while they rested. When the relationship between brain structure and performance on the behavioral measures was analyzed, a significant correlation between vocabulary score and gray matter density was revealed in both left posterior superior temporal sulcus and the left posterior temporal–parietal junction. In the teenage participants alone, an additional correlation was discovered between vocabulary score and gray matter density in the left posterior supramarginal gyrus. When Richardson et al. examined which of these regions were activated more during auditory and visual language comprehension than control tasks, only the first two regions showed significant effects. What, then, is the function of the left posterior supramarginal gyrus, where more gray matter was observed in teenagers with higher vocabulary abilities? The authors suggested that the developmental shift in the relationship between brain structure and behavior was driven by the way in which vocabulary is learned during the teenage years; specifically, that the relationship seen in teenagers is driven by the learning style of explicit instruction through lexical or conceptual equivalents common in secondary school, as distinct from incidental vocabulary learning through interaction that is more typical in younger children and adults.

In many ways, brain imaging lends itself as well to cross-sectional designs as do behavioral measures, but it provides an extra level of understanding and complexity that offers insights different to, and arguably beyond, behavioral measures. However, the difficulties inherent to all developmental studies can be exaggerated by the demands of adopting complex imaging techniques.

Benefits and Limitations

The major benefits of cross-sectional techniques in developmental psychology are practical in nature. The first benefit is that cross-sectional work is relatively inexpensive. The usual aim of cross-sectional studies is to get a measure of change with development, which may also be achieved by repeatedly testing the same children over time rather than testing children of different ages at one point. Such longitudinal studies require repeated testing over many years, resulting in high staff and laboratory costs, a significant commitment from participants, and the risk of data loss where that commitment cannot be delivered. The second benefit relates to the difficulty of selection bias, which is common to almost all psychology experiments. Researchers often invite participants who are close by, accessible, and likely to comply with studies. In addition, people who are proactive and interested in psychology are much more likely to take up invitations or get in touch with labs. This has some serious implications for the whole field (see Jones, 2010, for a discussion of the impact of overusing certain demographics in psychology studies). A specific problem for longitudinal studies is that people drop out over time, so for some participants a researcher might have just one data point, while for others they have several. Unfortunately, who stays in the study and who drops out are often not random. Participants might drop out because they find the study challenging, or perhaps because of difficulties with travel or unemployment. This means not only that data sets are often incomplete, but also that the data that are available are particularly prone to selection bias. Cross-sectional studies avoid at least some of this difficulty.

Despite the clear benefits of adopting a cross-sectional approach when running a developmental study, there are equally a number of important limitations. Perhaps the primary limitation is that when looking for the effect of age on some cognitive measure with a cross-section of children, there is an inevitable confound of individual differences. Do these children differ on the task because one is older or as a result of some other variable, such as nonverbal IQ or attentional control, which has not been measured? This will manifest as 'error' in the statistical models used, reducing the ability to find effects of interest in the variables that have been measured. In contrast, longitudinal methods effectively control for many of these unmeasured variables, to the extent that they are stable across the time points of measurement involved in the study.

The second major challenge of cross-sectional methods is the development of an appropriate paradigm (as briefly discussed in this article). Cross-sectional methods assume that the cognitive function in question is captured in the same way across the whole range of chronological or mental ages considered. However, it is not always clear that this is, in fact, the case. There are two main situations in which the assumption may not be met. The first is when different tasks are used for participants of different ages or ability levels. Consider how one might test the language ability of an 18-month-old, a 5-year-old, and a 15-year-old. The first might rely on a parental questionnaire of the vocabulary that the child produces; the second can focus on more directly on the child's language skills in the oral domain, including both vocabulary and syntax; and the third may assess more complex aspects of, say, syntax or pragmatics in the written domain. However, the problem of using different tasks at different ages is obvious: if the task is different, then the demands must be different. This creates problems of interpretation because these differences in task demands may lead to differences in scores across the range of ages measured, rather than (or in addition to) any underlying differences in the target cognitive function.

A more basic problem is that the different tasks may be scored in different ways. For example, the RCPM (Raven et al., 1998) is a test of nonverbal reasoning for children and is scored out of 36 items. Raven's Standard Progressive Matrices (RSPM; Raven et al., 2003) is a broad equivalent for adults and is scored out of 60. How should one compare scores from each test? In this case, there is a solution, because norms are available that indicate what score the norming samples achieved on average across the whole age range, with standard deviations for each age. For a given age, participants' score can be converted into how many standard deviations they scored above or below the norm average. When such norms are not available, researchers can standardize scores against their own sample, although large numbers are generally required.

A subtler situation arises when the task used does not change but the determinants of task performance do. A clear example is the decline in working memory capacity observed in older adults. According to Baddeley's (1986) model of working memory, there is an attentional control component, the 'central executive,' and two 'slave' passive storage systems, the phonological loop and visuospatial sketchpad, which temporarily hold verbal and visuospatial representations, respectively. Older adults perform similarly to younger adults on tasks that require few executive demands (i.e., they require little attentional control), but are poorer than younger adults on tasks that do make such demands (e.g., Craik and Byrd, 1982). Thus, the poorer working memory scores of older adults reflect limitations of attentional control, but not of working memory per se.

In order to conduct valid research comparing task performance across the lifespan, researchers must analyze their main task in terms of the cognitive demands that it makes. If there is reason to believe that any of these demands, other than the target cognitive function, might change across the chronological or mental age range involved in the research, then these component demands should be independently measured so that any observed differences in the main task can be confidently attributed to the target ability rather than to the secondary demands.

A final limitation is the widespread use of linear modeling in cross-sectional studies of development. Although some cognitive changes may approximate straight lines, many may not. However, although some challenges in interpretation may arise, most linear methods, including trajectory analyses, can be adapted to nonlinear models simply by altering the model options in the statistical software used (see Thomas et al., 2009). As noted by Thomas and colleagues, the principle of parsimony may be invoked when deciding whether a linear or nonlinear model is preferable: nonlinear models involve a greater number of parameters than linear models.

Getting the most out of cross-sectional methodologies will mean taking advantage of the benefits and minimizing the impact of the limitations. Ways of minimizing the impact of the limitations might include taking care to abate potentially confounding factors where possible, and keeping response demands simple for all paradigms, while ensuring as wide a performance range as possible; using adaptive tasks is one way to achieve this.

Conclusion

The aim of this article was to describe the use of cross-sectional methodologies in developmental psychology, and to explain their origin, their strengths, and their weaknesses. We have discussed the theory behind the statistical techniques adopted in cross-sectional studies and used examples from the literature to illustrate their use.

Cross-sectional methodologies provide substantial scope for the development of highly informative studies without the cost associated with longitudinal work. The development of trajectory modeling allows a developmental perspective to be taken, at the heart of which is the aim of determining the mechanisms of change in cognitive systems. Throughout the history of developmental psychology, the establishment of statistical methods such as regression has been intimately tied up with theoretical advancements and new conceptual understanding. This is certainly apparent with respect to cross-sectional studies, although equally the statistical limitations inherent in data collection at a single point in time must be borne in mind. Development is, in its most fundamental sense, a dynamic process of change. Indeed, the word itself comes from the French *développer*, to unfold. Research into the process of cognitive development must, therefore, balance practicality with the theoretical rigor of longitudinal studies.

The future of cross-sectional methodologies lies in researchers pushing the boundaries of what questions can be asked: using trajectory analyses to trace the patterns of cognitive change, using brain-imaging techniques to answer questions about the neuroanatomical and neurophysiological underpinnings of cognition, and bearing in mind that developmental psychology is not just about cognitive development in children but also about change throughout the lifespan.

See also: Cross-Cultural Research Methods in Psychology; Equivalence and Transfer Problems in Cross-Cultural Research; Longitudinal Analyses of Sexual Development through Early Adulthood; Qualitative Methodology in Developmental Psychology.

Bibliography

Baddeley, A.D., 1986. Working Memory. Oxford University Press, New York.

Baron, R.M., Kenny, D.A., 1986. The moderator mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology 51, 1173–1182.

Binet, A., Simon, T., 1905. The Development of Intelligence in Children: (The Binet-Simon Scale). Vineland, NJ: Publications of the Training School at Vineland. Department of Research, NJ.

Corsi, P.H., 1972. Human memory and the medial temporal region of the brain. Unpublished doctoral dissertation, McGill University, Canada.

Craik, F.I.M., Byrd, M., 1982. Aging and cognitive deficits: the role of attentional resources. In: Craik, F.I.M., Trehub, S.E. (Eds.), Aging and Cognitive Processes. Plenum, New York.

Crawford, C., Dearden, L., Greaves, E., 2013. Identifying the Drivers of Month of Birth Differences in Educational Attainment. Institute for Fiscal Studies, London.

Crawford, C., Dearden, L., Maghir, C., 2010. When You Are Born Matters: The Impact of Date of Birth on Educational Outcomes in England. Institute for Fiscal Studies (IFS), London.

Davidson, M.C., Thomas, K.M., Casey, B.J., 2003. Imaging the developing brain with fMRI. Mental Retardation and Developmental Disabilities: Research Reviews 9, 161–167.

Dunn, L.M., Dunn, L.M., Whetton, C., Burley, J., 1997. The British Picture Vocabulary Scale, second ed. NFER Nelson, Swindon, UK.

Dunn, M., Dunn, L.M., 1997. Peabody Picture Vocabulary Test – 3. AGS, Circle Pines, MN.

Elder, T., Lubotsky, D., 2008. Kindergarten Entrance Age and Children's Achievement: Impacts of State Policies, Family Background and Peers. Mimeo.

Holland, S.K., Vannest, J., Mecoli, M., Jacola, L.M., Tillema, J.-M., Karunanayaka, P.R., Schmithorst, V.J., Yuan, W., Plante, E., Byars, A.W., 2007. Functional MRI of language lateralization during development in children. International Journal of Audiology 46 (9), 533–551.

Holmbeck, G.N., 1997. Toward terminological, conceptual, and statistical clarity in the study of mediators and moderators: examples from the child-clinical and pediatric psychology literatures. Journal of Consulting and Clinical Psychology 65, 599–610.

Jarrold, C., Brock, J., 2004. To match or not to match? Methodological issues in autism-related research. Journal of Autism and Developmental Disorders 34, 81–86.

Jones, D., 2010. A WEIRD view of human nature skews psychologists' studies. Science 328, 1627.

Knickmeyer, R.C., Gouttard, S., Kang, C., Evans, D., Wilber, K., Smith, J.K., Hamer, R.M., Lin, W., Gerig, G., Gilmore, J.H., 2008. A structural MRI study of human brain development from birth to 2 years. The Journal of Neuroscience 28 (47), 12176–12182.

Pearson, K., 1896. Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. Philosophical Transactions of the Royal Society A 187, 253–318.

Piaget, J., 1936. La naissance de l'intelligence chez l'enfant (The Origin of Intelligence in the Child). Routledge and Kegan Paul, London.

Purser, H.R.M., 2006. Short-term memory in Down syndrome. Unpublished PhD thesis. University of Bristol.

Purser, H.R.M., Farran, E.K., Courbois, Y., Lemahieu, A., Sockeel, P., Blades, M., 2012. Short-term memory, executive control and children's route learning. Journal of Experimental Child Psychology 113, 273–285.

Purser, H.R.M., Thomas, M.S.C., Snoxall, S., Mareschal, D., Karmiloff-Smith, A., 2011. Definitions versus categorization: assessing the development of lexico-semantic knowledge in Williams syndrome. International Journal of Language & Communication Disorders 46 (3), 361–373.

Raven, J., Raven, J.C., Court, J.H., 1998. Coloured Progressive Matrices. Oxford University Press, Oxford, UK.

Raven, J., Raven, J.C., Court, J.H., 2003. Manual for Raven's Progressive Matrices and Vocabulary Scales. Harcourt Assessment, San Antonio, TX.

Richardson, F.M., Thomas, M.S., Filippi, R., Harth, H., Price, C.J., 2010. Contrasting effects of vocabulary knowledge on temporal and parietal brain structure across lifespan. Journal of Cognitive Neuroscience 22 (5), 943–954.

Sharp, C., 1995. What's age got to do with it? A study of patterns of school entry and the impact of season of birth on school attainments. Educational Research 37, 251–265.

Spearman, C., 1904. "General intelligence," objectively determined and measured. The American Journal of Psychology 15 (2), 201–259.

Thatcher, R.W., North, D.M., Biver, C.J., 2008. Development of cortical connections as measured by EEG coherence and phase delays. Human Brain Mapping 29 (12), 1400–1415.

Thomas, M.S.C., Annaz, D., Ansari, D., Scerif, G., Jarrold, C., Karmiloff-Smith, A., 2009. Using developmental trajectories to understand genetic disorders. Journal of Speech, Language and Hearing Research 52, 336–358.

Crowds in History

M Gailus, Technical University of Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 5, pp. 3022–3027, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Crowds continue to play a major role in politics and society, just as they have done since the first forms of urban living were established in ancient times (although they can also be found as forms of protest in agricultural societies). As such the issue of crowds and how they behave is a significant area of research. This article outlines the 'Roman school' of mass psychology around 1900 (LeBon), the social–historical school of French Revolution research (Lefebvre), the modern classic 'history from below' approach (George Rudé et al.), and the quantitative American school of historical sociology (Charles Tilly). The claim has often been made that crowd rebellions can normally be reduced to two main essentials: 'subsistence' and 'liberties.' More recent research, joining the modern classic school, has increasingly included other aspects of crowds that have been too neglected, including violent actions against religious or ethnic minorities.

After stating the historical relevance of crowds, they are defined as relatively sudden gatherings of large groups of people whose actions are goal oriented and conflictual. They interrupt the regular course of everyday life and disturb, or threaten to disturb, public order. Crowd history has primarily been indebted to the social–history paradigm, ranging methodologically from 'thick description' to extended protest statistics. Central questions concern times and places of crowds, their physical composition, forms, and motives of action, their targets, the role of peacekeepers and governments, and their long-term effects on high politics and society. Describing different schools, four crucial phases are identified: the 'Roman school' of mass psychology around 1900 (LeBon), the social–historical school of French Revolution research (Lefebvre), the modern classic 'history from below' (Rudé et al.), and the quantitative American school of historical sociology (Tilly). Focusing on early-modern Europe, the modern classic school reconstructed a peculiar repertoire of crowd actions. At the center were food riots, a long lasting European conflict on values and goals that is best explained by Thompson's path-breaking concept of the moral economy. Taking the whole spectrum of crowd actions together, their unwritten program can be reduced to two main essentials: 'subsistence' and 'liberties.' These issues were defended in a specific mode of popular 'rebellious traditionalism' against the two great dynamic powers of expanding markets and national states. More recent research, joining the modern classic school, has increasingly included other aspects of crowds that have been too neglected: violent actions against religious or ethnic minorities (especially anti-Jewish riots), monarchistic, and conservative crowds within dynastic ceremonies or national movements, or twentieth-century crowds in 'street politics' and party political action. There has also been a broadening of perspectives on non-European crowds, such as peasant resistance in Asia or historical food rioting in China or India, and very recent conflicts in Africa and Latin America. The long lasting interpretive struggle (*Deutungskampf*) on crowds is marked by a double disappointment: neither have crowds turned out to be the great culture-destroying evil, nor have they fulfilled any 'higher' historical missions. However, a certain ambivalence remains: although emancipating aspects may dominate, in their repertoires there were (and are) also dangerous exclusionary efforts and violence. But following their unwritten program of 'subsistence' and 'liberties' in the main, crowds – as paradoxical as it may seem – were indirect reformers, cowriting the present day agenda of social and political human rights. There has been some transformation of crowds in Western societies: they have become normal parts of political culture. Social- and political-egalitarian demands have been shrinking for some time now, ranking beside (or behind) new values and demands within 'new social movements.' Becoming less class specific, crowds today no longer consist exclusively of the lower orders, but are more likely to be mixed. As an elementary form of face-to-face public, crowds continue to persist in the age of mass media.

Crowds do play an important role in history. Particularly during times of crisis and revolution, they can function as prime movers in historical–political conflicts. As a constituent feature of everyday public life, the omnipresent crowd is a highly complex phenomenon, especially in modern urbanized societies. In view of the innumerable forms that crowds take, in the most varied of contexts, for pragmatic reasons it is sensible to distinguish between: (1) more trivial everyday crowds as well as crowds of a recurring, institutionalized, and routine nature; and (2) exceptional, unforeseeable crowd events.

For research purposes, crowds can be defined as relatively short-term gatherings of large groups of people (on streets or squares or in the countryside – in any event, in the open air), whose actions are goal oriented and, as a rule, conflictual. Crowds are assemblages that interrupt the regular course of everyday life and disturb the public order, or at least threaten to disturb it. Under certain circumstances, of course, an everyday or routine crowd can also transform itself into one that will be conflictual.

Crowd history was (and will likely remain) primarily indebted to the social–history paradigm: theory oriented, but seldom fixated on theory – in fact, with a fair bit of skepticism regarding theory; and utilizing a variety of empirical methods, singularly or in combination, from 'thick description' to

extensive statistics on protest. Central concerns are the precise reconstruction of crowd events, the involved groups and their identities, and a microhistorical embedding in local contexts. Dominating questions concern exact times, size, duration, frequency, and topography or geography of crowds. Research topics are their physical composition (social status, ethnic identity, religion, age, sex) and the – considerably heterogeneous – behavior of individuals or smaller groupings in these events.

Other focuses are the forms and motives of action (violent tendencies, emotionality, resentment, mentalities, visions), in addition to occasions and possible structural reasons for common action. Furthermore, crowd history investigates the targets of crowd action, as well as the presence and function of peacekeepers in such confrontations. And last but not least, crowd history discusses the results and long-term consequences of crowd events in consideration of high politics and society.

Phases and Schools: Changing Images of Crowds

Academic crowd research was first begun toward the end of the nineteenth century, and its beginnings are conspicuously multidisciplinary: philosophy, psychology, criminology, sociology, and anthropology occupied themselves with the phenomenon; only gradually did historical studies turn their attention to the subject. Against the background of opposing political–ideological positions, several academic schools disputed the 'true essence' or the 'actual character' of crowds. Depending on one's position, interpretations ranged from an extreme damning of crowds as purely negative and regressive destroyers of civilization, to their enjoying an idealized historical–philosophical role as genuine creative agents, as essential driving forces of historical progress. From the historian's perspective, there are at least four important phases.

The starting point was the so-called 'Latin School' of mass psychology around 1900, represented by such authors as the Italian criminologist Scipio Sighele (*La Folla Delinquente*, 1891), the French popular scientist Gustave LeBon (*La Psychologie des Foules*, 1895), and the French criminologist and sociologist Gabriel Tarde (*L'Opinion et la Foule*, 1901). Relying for their historical evidence primarily on one-sided studies of the French Revolution by the conservative historian and philosopher Hippolyte Taine, their speculative theories concerning crowds were a projection of bourgeois-conservative fears of social subversion at the *fin de siècle*: "Their crowds loomed as violent, bestial, insane, capricious beings whose comportment resembled that of mentally ill, women, alcoholics, or savages" (Barrows, 1981: p. 5).

The emerging social–historical school of French Revolution research opposed the conservative LeBon mass psychology by de-demonizing the crowd debate through empirical case studies addressing the character and role of crowds during the French Revolution. One of this school's path-breaking authors was Georges Lefebvre. In *La Grande Peur* (1932) and especially in *Foules Révolutionnaires* (1934), he showed how rebellious groupings were stimulated through various types of oral communication (e.g., spreading of rumors, suspicions, and mental images), and how the *foules pures* of everyday life transmogrified into the *foules révolutionnaires* when seized by a 'collective mentality.'

After the World War II, a modern classic school of crowd research was constituted that traced its ancestry partly to the social–historical school of French Revolution research. The classic school's first important propagator was an occasional student of Lefebvre's, George Rudé, whose 1950 doctoral dissertation treated the Paris workers and their uprisings of 1789–91. In his studies on crowds in the French Revolution and on crowds generally in France and England from 1730 to 1848 (Rudé, 1964), he identified for the first time the 'faces in the crowd.' Together with Eric J. Hobsbawm's early studies of the pre- and early-industrial workers groups whose rebellious behavior was plausibly labeled 'collective bargaining by riot' (Hobsbawm, 1952), as well as the related work of Edward P. Thompson, especially his paradigmatic study of *The Moral Economy of the English Crowd in the Eighteenth Century* (Thompson, 1971), the classical canon of a new history from below emerged. These historians understood themselves to be an open and undogmatic variant of a Marxist historiography of the West. At first an oppositional subdiscipline, from the 1960s to the 1990s it quickly gained a wide reputation in academic circles, inspiring numerous other studies.

There was a parallel development in the American social sciences, using modernization theory and quantitative measurements in historical–sociological research on crowds. Crucial was a more systematic investigation of possible causal connections between large processes, such as industrialization, urbanization, or social crises, and the frequency and geography of crowd occurrences. In a critical test of theory, then-current concepts of collective behavior and of social movement research (relative deprivation, among others) were checked against historical evidence. This led to more refined methods of crowd event analysis, accumulating more precise protest statistics over longer periods and/or testing its geographic diffusion for strategically chosen regions. Exemplary is the work of Tilly et al., in particular the comparative study of France, Italy, and Germany in *The Rebellious Century* (1975), or the strictly quantitative and simultaneously Thompsonian case study of English riots 1790–1810 (Bohstedt, 1983).

The Unwritten Program: 'Subsistence' and 'Liberties'

For the longest time, European societies of the early-modern period, including the nineteenth century, figured as the classic time frame for the bulk of crowd research. The essential scope of crowd activities was above all defined by the secular processes of state centralization and national state-building, the expansion of market economies, social polarization, and proletarianization. Viewed from the lower-class angle, one is able to speak of a European age of persistent material shortages. Under conditions of subjection and/or political immaturity, the lower classes found themselves in permanent conflict with authorities and social elites for subsistence guarantees, communal rights and freedoms, and cultural autonomy.

Rudé's and Hobsbawm's 'new picture of the crowd' was and remains largely a counterimage to the previously dominant

LeBon tradition. Crowds, it was now shown, were not – as a rule – criminal rabble but people of established residence, average working men and women, poor but orderly folk. Their motivations were a frequent mixture of Jacobin egalitarianism and conservative monarchism. Their collective behavior was one of direct action, and they pursued their thoroughly rational goals with fitting means, this sometimes translating into violence. However, it is said that their violence was predominantly symbolic and a question of efficacy. Moreover, their successes were of limited nature, particularly as their actions were so short lived and lacking of far-reaching visions. Thus, conflict crowds were to be considered forerunners of later social and political movements with lasting organizational structures and long-term social–political programs.

These and related studies established a solid canon of conflicting crowds, a European repertoire of riots. These included: peasant uprisings against noble estates, antitax protests against mounting state prerogatives, religious riots between competing confessions, subsistence-based actions of the rural and urban lower classes, moral rebukes in the tradition of *charivari* (rough music), strikes and boycotts of artisans and workers, machine-wrecking, but also aggressive acts of popular monarchic loyalty against alleged enemies of king and church, and xenophobic outbursts against 'strangers.' As the latest research for the UK, France, and Germany has shown, in no other early-modern conflict of any duration were crowds so involved as in food riots.

More than 3000 food riots have been documented for France between 1560 and 1860. At their center were the blockades of interregional grain trade to prevent scarce resources from leaving their place of origin. Participating in these blockades were mainly the rural and urban lower classes from the highly commercialized farming regions of the far-reaching agrarian hinterlands of Paris. For four decades, ever since 1760, the power struggle over access to limited food resources became permanent in France. The attempt of Controller General Turgot in the fall of 1774 to deregulate the national grain trade according to physiocratic liberal doctrines, ended in the fiasco of the Guerre des Farines (Flour War) of 1775, when over 300 food rebellions forced the government to back down its policy (Bouton, 1993).

There have been about 1000 food riots documented for England in the period 1530–1820, the majority of them taking place in the last 80 years of that time span. With regard to the years around 1800, Bohstedt speaks of a communally anchored 'protocol of riot,' of 'orderly disorders,' with widely accepted public rules for all of the participants. With crowd action as an essential element of community politics, poorer consumer groups, especially in small and medium-sized market towns, participated successfully in the creation of reasonably tolerable living conditions (Bohstedt, 1983).

For German regions (1700–1870) over 400 food conflicts have been substantiated. They culminated in the 1840s (shortly before and during the Revolution of 1848), a decade that has been characterized as the 'angry forties.' Particularly in Prussia's East-Elbian provinces, the eighteenth-century myth of a paternalistic 'social monarchy' quickly faded. Market-driven resource scarcities, not least the result of agricultural exports to Western Europe, forced a sharp internal 'dissent about markets.' Even more so than the urban lower classes, the rural poor frequently saw themselves robbed of whole food entitlements (Gailus, 1990).

All in all, these recurring pre- and early-industrial mass conflicts attested to not only a simple struggle for calories among the abject poor and hungry, but symbolized a central conflict of values and goals. On the one side were liberal principles and Utopias envisioning abundant markets and property for all, while opposed to it were nontransferable subsistence rights and the blows to the human dignity of rapidly expanding market-dependent classes in an increasingly urbanizing market society.

Probably no other model has so persuasively explained the social and cultural meaning of pre- and early-industrial crowd activities as Edward P. Thompson's classic notion of moral economy, which he developed through an analysis of the rebellious behavior of the working poor in eighteenth-century England. According to his study, the customs and rituals practiced by the crowds in their numerous food riots, their angry outbursts and anxious visions, their claims and demands, were embedded in "a popular consensus as to what were legitimate and what were illegitimate practices in marketing, milling, baking, etc. This [consensus] in its turn was grounded upon a consistent traditional view of social norms and obligations, of the proper economic functions of several parties within the community" (Thompson, 1971: p. 79). For Thompson the word 'riot' is too narrow a concept to adequately describe all the issues at stake in these confrontations. In order to assure their 'way of survival,' crowds (or people) were in permanent collective negotiations with aristocracy and gentry, with officials and the middle classes. The crowds would typically remind the authorities of their traditional paternalistic duties vis-à-vis the weaker members of the community. In the ritualized forms of conflict with ruling elites, in expressive and symbolic confrontations of political 'theater' and 'counter-theater,' they actively participated in the social organization of their own living conditions and thus functioned as an influential collective subject. Crowds were exercising direct influence on the epoch's balance of power, which was shaped by the political hegemony of the gentry (Thompson, 1991). One can with good reason speak of 'participation by riot' or of popular 'participation before participation.' Despite some critical objections, the moral economy model has asserted itself as the most successful interpretive concept for historical studies of conflict crowds – and not solely for food riots and for European conflict patterns. More recent work attests to its yet fertile analytic potential (Randall and Charlesworth, 1996).

Focusing the implications of these old European crowds on a few essential aspects would above all include 'subsistence' (entitlements to food, rights of social share) and 'liberties' (in the sense of additional rights and popular customs). This collective establishing of values was both asserted and defended against outside 'disturbances' and all actual and imagined threats to existence through dynamic penetration of the two great powers, the national state and the expanding modern market. This culture of resistance of conflicting crowds has contributed in manifold ways to the formation of basic political and social rights of modern constitutional states.

Conceptual Constrictions – Widening of Perspectives

The advance of the modern social–historical school (1960–90) had a lasting effect on the historical image of crowds: the 'bad crowds' of the older conservative orthodoxy turned into the 'good crowds,' possessing a predominantly positive, emancipating function. However, completely or partially obscured were other aspects of crowd history: aggressive, conservative, and nationalistic or chauvinistic mass events, as well as excessively violent and ethnically motivated outbursts. Also criticized has been crowd history's exclusive concentration on rebellious crowds. Too little attention has been given to other forms of nonconflicting crowds. In addition, the 'great watershed' of 1850, allegedly occasioning a transformation of crowds, has hardly been proven. Despite a sneaking institutionalization of crowds over the following decades, they have remained critically important in the industrial societies of the late nineteenth and twentieth centuries (Holton, 1978; Harrison, 1988).

Crowd history has to pay more attention to violent actions against minorities and outsiders. Excessive crowd violence was mostly reserved for pogrom-like persecutions of the Jewish minorities. These assumed an almost nationwide character at least twice in nineteenth-century Germany: in the form of 'Hep-hep riots' (1819) against the political emancipation of the Jews, and the forcible expulsion of south German village Jews within the wider context of the rural mass movements of the 1848 revolutions (Rohrbacher, 1993). As of the second half of the nineteenth century, pogroms established themselves in eastern Europe with much greater militancy – in Russia, especially the Ukraine, and in parts of Poland – its chief phases encompassing the periods 1881–84, 1903–06, and 1919–21. Until the end of the nineteenth century these actions were perpetrated by crowds of similar compositions to those in other riots (Klier and Lambroza, 1992).

There has been an increasing shift of focus to nonconflicting crowds, a development promising to greatly expand the field of research. Dynastic governmental ceremonies, middle-class-inspired national festivals, traditional celebrations, religious feast days, and new style leisure activities could easily assume the form of mass events, though in different guises and in radically transformed public functions: staged by the ruling class and social elites and hierarchically structured; composed of a mixture of social groups, neither exclusively nor primarily from the lower classes; symbolically strengthening the existing rule and its established social order, as opposed to rebellious behavior protesting destitution and political impotence; in their effects, socially integrative rather than polarizing. The lower classes, usually self-determining actors at the very nucleus of crowd activities, figured here primarily as public, as mere backdrop.

A study of Bristol 1790–1835 shows that next to peaceful 'election crowds,' 'royal and military crowds,' and 'recreational crowds,' unruly 'crowd riots' were among those events occurring with the highest frequency in the city. However, at the same time, 'crowd riots' formed only a minority (22%) of a sample that covered all types of crowd occurrences taking place within the city. The local elites of Bristol had ample skill and experience in successfully mobilizing large crowds onto the main streets and squares for purposes of their own (Harrison, 1988).

The general European 'nationalization of the masses' in the nineteenth century included a wealth of identity-forming national celebrations and mass manifestations.

A comparative study of middle-class movements for national memorials in France and Germany has demonstrated the precarious and paradoxical strategy of including the lower classes in the mythical 'nation,' while simultaneously drawing sharp social distinctions over and against these very same lower classes. It was above all through the militarization of the bourgeois festival culture that the elites were able to overcome their deep-rooted fear of the masses (Tacke, 1995).

Expanded political participation, as well as transformations within the public sphere in industrial and postindustrial societies of the twentieth century in no way contributed – as often asserted – to the disappearance of protesting crowds. Rather, one can speak of modifications, and of coexistence of direct action with newly learned and more organized and representative political forms. After 1850, aside from periodic recurrences of crowd activity in line with the older model (for example, food riots in France, and especially in Italy and Spain), one finds numerous newly invented forms of crowd action, above all strikes, boycotts, mass manifestations and the like. However, soon after 1900 the power potential of crowds began to appear less threatening to the established order (in the sense of general overthrow). Crowds were governable, crowds could be manipulated – this was the new slogan of ruling elites.

An empirical study of Berlin 1900–14 shows the high presence of crowd events in early twentieth-century streets. The events fell into three major categories – 405 street disorders, 74 violent strikes, and 61 political mass manifestations. The ongoing struggle for the city streets is interpreted as conflict between a 'street politics from below' and a restrictive 'street politics from above' carried out by the state power and its police. It was on the political terrain of the street that exaggerated fear of subversion, strict antirevolutionary measures, and socially discriminatory police practice clashed with collective demands for greater social justice and expanded political participation (freedom of assembly, the right to demonstrate, universal suffrage). In a much wider sense, these were demands for respect and self-determination fought out by the proletarian majority in the public forum of the street, the traditional political arena of the common people (Lindenberger, 1995).

For the extremely crisis-ridden interwar period, one can even speak of a culmination of 'street politics,' increasingly transformed (militarized) through the advent of radical political–ideological movements from both the left and the right. In this civil war atmosphere of rabid party political action, with the air filled with bellicose slogans, such as 'self-help' and 'self-defense,' authentic collective crowd action threatened to be instrumentalized and finally suffocated in the battle of totalitarian movements bent on the conquest of state power (for a comparative study of Berlin and Paris 1919–33/39, see Wirsching, 1999).

For some time now the broadening of perspectives has meant the inclusion of non-European regions of Asia, Africa, and Latin America, which have a history of crowds of their own. Important basic forms resemble the European riot patterns: in

the center stood collective claims of peasant producers who rebelled against landlords or large land-holding farmers, urban moneylenders or state organs for reasons of land scarcity, imposed duties, debts, or tax pressure. In the East Asian context, in a series of studies, James C. Scott has shown how peasant behavior was culturally embedded in an all-encompassing subsistence ethic with precise norms regarding access to land, acreage, and a fair apportionment of their own produce. The threats posed to these cultural norms in the context of firmly established clientele relationships might have led to open collective resistance, even if – in consideration of the manifold covert and individual 'weapons of the weak' – this remained more the exception than the rule (Scott, 1976).

In some regions (particularly of China since the eighteenth century, and to some degree of Japan and of India in the nineteenth and early twentieth centuries) the food riot tradition was of considerable historical importance, thus offering comparisons with early-modern Europe (Wong, 1983). At the same time, subsistence conflicts of many Third World regions continue to this day, as confirmed by many contemporary eruptions since the 1970s in North and Central Africa (Egypt, Morocco, Sudan, and Tunisia) and Latin America (Dominican Republic, Venezuela, Brazil, and Argentina). In several cases it was food-riot crowds – whose agitation came in the wake of reduced state food subventions and resulting higher prices – that shook and even toppled ruling governments.

Outlook

The long and hard-fought interpretive battle concerning crowds has been marked by a double disappointment: neither have crowds been shown to be the great negative power, an absolute culture-destroying evil; nor have they been – teleologically speaking – a collective subject to fulfill any historical missions. Crowds have always shown themselves to be multilayered, ambiguous, and variously interpretable phenomena. And naturally, a certain ambivalence remains. For although emancipatory aspects in the sense of subsistence gains, political participation, and self-determination may dominate the phenomenon of crowds, dangerous exclusionary efforts at expulsion or even destruction of religious, ethnic, and other minorities can also be a part of the crowd repertoire. The demythification of the crowd, the gradual opening of political structures to greater popular participation, the modification of collective behavior through historical experience (learning processes that include a growing self-control and adoption of the principle of nonviolence), and other factors may have contributed to crowds having increasingly forfeited their attractiveness as a catch-all for fearful visions or as an instrument for political and ideological agendas imposed from without.

As historical actors in their own right, crowds lay claim to a considerable degree of legitimacy. At the center of this legitimacy stood the collective demands for 'subsistence' (later on: social sharing) and 'liberties' (communal rights, cultural autonomy and, later, political participation) in the context of given governmental and living conditions. Insofar, conflicting crowds were always – more unconsciously than intentionally – indirect collective reformers. Obviously, this seemingly paradoxical role is part of the peculiarities of rebellion and revolution. The political culture of the West – with its human rights norms embodied in the welfare state and democratic institutions since the end of the nineteenth century – would have never attained to the stature it occupies today had it not been for persistent collective intervention from below. Crowds have helped to draw up the agenda for social and political human rights, and they continue to do so.

Where comprehensive systems of social sharing and political participation have been historically established and long in place, crowds tend toward a shift in their forms and motives. From a purely quantitative standpoint, protesting crowds in recent times have been anything but on the decline. To the contrary, they have been on the upswing, becoming a conventional part of political culture. Although violence has receded, it has not fully disappeared. Rating high are questions other than elementary ones of subsistence and freedom. Demands for social and political egalitarianism have for some time taken a backseat to new values and fresh demands based on 'new social movements.' Those groups affiliated with crowd events are no longer predominantly or exclusively from the lower classes, but are a great mix of people with considerable representation from the public sector, education, science, and culture. Fitting in well with these trends is also the upsurge of class nonspecific leisure crowds in the context of new event cultures.

As an elementary form of the public sphere (*Versammlungsöffentlichkeit*), crowds will endure. Certainly modern mass media have continued to occupy strongly the public arena, becoming themselves more and more an embodiment of the entire public. But talk shows have hardly rendered crowds superfluous. Albeit today's media presence largely determines the public perception of crowd events, thereby heavily influencing their effectiveness. In their endless variety, crowds represent the unmediated, face-to-face public. Such gatherings of people for purposes of social and communicative self-awareness are apparently a deep-seated human need that – despite the relative dwindling of traditional conflicting crowds – will not soon disappear.

See also: Collective Behavior, Sociology of; Groups, Sociology of; Public Sphere: Nineteenth and Twentieth-Century History; Revolution; Revolutions, History of; Revolutions, Sociology of; Social Movements, Geography of; Social Movements, History of: General; Social Protest.

Bibliography

Barrows, S., 1981. Distorting Mirrors: Visions of the Crowd in Late Nineteenth-Century France. Yale University Press, New Haven, CT and London.

Bohstedt, J., 1983. Riots and Community Politics in England and Wales 1790–1810. Harvard University Press, Cambridge, MA.

Bohstedt, J., 2010. The Politics of Provisions: Food Riots, Moral Economy, and Market Transition in England, c. 1550–1850. Ashgate, Burlington, VT.

Borch, C., 2012. The Politics of Crowds: An Alternative History of Sociology. Cambridge University Press, Cambridge.

Bouton, C.A., 1993. The Flour War: Gender, Class, and Community in Late Ancien Regime French Society. Pennsylvania State University Press, PA.

Buoye, T.M., 2000. Manslaughter, Markets, and Moral Economy: Violent Disputes over Property Rights in Eighteenth Century China. Cambridge University Press, Cambridge.
Canetti, E., 2000. Crowds and Power. Phoenix, London.
Gailus, M., 1990. Strasse und Brot: Sozialer Protest in den deutschen Staaten unter besonderer Berücksichtigung Preussens 1847–1849. [English Translation: Street and Bread: Social Protest in the German States with Special Reference to Prussia 1847–1849]. Vandenhoeck & Ruprecht, Göttingen, Germany.
Harrison, M., 1988. Crowds and History: Mass Phenomena in English Towns 1790–1835. Cambridge University Press, Cambridge, UK.
Hobsbawm, E.J., 1952. The machine breakers. Past and Present 1, 57–70.
Holton, R.J., 1978. The crowd in history: some problems of theory and method. Social History 3, 219–233.
Jonsson, S., 2013. Crowds and Democracy: The Idea and Image of the Masses from Revolution to Fascism, Columbia University Press.
Jupp, P., Magennis, E., 2000. Crowds in Ireland, c. 1720–1920. Macmillan Press, Houndmills, Basingstoke, Hampshire.
Klier, J.D., Lambroza, S. (Eds.), 1992. Pogroms: Anti-Jewish Violence in Modern Russian History. Cambridge University Press, Cambridge, UK.
Lefebvre, G., 1934. Foules révolutionnaires [English translation: Revolutionary crowds]. Annales historiques de la Révolution franaise 11, 1–26.
Lindenberger, T., 1995. Strassenpolitik: Zur Sozialgeschichte der öffentlichen Ordnung in Berlin 1900 bis 1914. [English Translation: Street Politics: A Social History of Public Order in Berlin 1900 to 1914]. Dietz Nachfolger, Bonn, Germany.
Randall, A., Charlesworth, A. (Eds.), 1996. Markets, Market Culture and Popular Protest in Eighteenth-Century Britain and Ireland. Liverpool University Press, Liverpool, UK.
Randall, A., Charlesworth, A., 2000. Moral Economy and Popular Protest: Crowds, Conflict and Authority. St. Martin's Press, New York.
Rohrbacher, S., 1993. Gewalt im Biedermeier: Antijüdische Ausschreitungen in Vormärz und Revolution 1815–1848/49. [English Translation: Violence in the Biedermeier Era: Anti-Jewish Riots during Vormrz and Revolution 1815–1848/49]. Campus, Frankfurt, Germany and New York.
Rudé, G., 1964. The Crowd in History: A Study of Popular Disturbances in France and England 1730–1848. Wiley, New York and London and Sydney, Australia.
Schnapp, J.T., Tiews, M., 2006. Crowds. Stanford University Press, Stanford, CA.
Schneider, S., Weinberg, H., 2003. The Large Group Re-visited: The Herd, Primal Horde, Crowds and Masses. Jessica Kingsley, London.
Scott, J.C., 1976. The Moral Economy of the Peasant: Rebellion and Subsistence in Southeast Asia. Yale University Press, New Haven, CT and London.
Tacke, C., 1995. Denkmal im sozialen Raum: Nationale Symbole in Deutschland und Frankreich im 19 Jahrhundert. [English Translation: Monument in the Social Space: National Symbols in Germany and France in the 19th Century]. Vandenhoeck & Ruprecht, Göttingen, Germany.
Thompson, E.P., 1971. The moral economy of the English crowd in the eighteenth century. Past and Present 50, 76–136.
Thompson, E.P., 1991. The patricians and the plebs. In: Thompson, E.P. (Ed.), Customs in Common. Merlin Press, London, pp. 16–96.
Tilly, C., Tilly, L., Tilly, R., 1975. The Rebellious Century 1830–1930. Harvard University Press, Cambridge, MA.
Walter, J., 2006. Crowds and Popular Politics in Early Modern England. Manchester University Press, Manchester, UK.
Wirsching, A., 1999. Vom Weltkrieg zum Brgerkrieg? Politischer Extremismus in Deutschland und Frankreich 1918–1933/39: Berlin und Paris im Vergleich. [English Translation: From World War to Civil War? Political Extremism in Germany and France 1918–1933/39: Berlin and Paris in Comparative Perspective]. Oldenbourg Verlag, Munich, Germany.
Wong, R.B., 1983. Les émeutes de subsistances en Chine et en Europe occidentale. [English translation: Food Riots in China and Western Europe]. Ann Econ Soc Civil 38, 234–258.

Cuban Revolution, The

Manuel R Moreno Fraginals[†]
Teresita Pedraza Moreno, Miami Dade College, Miami, FL, USA

Published by Elsevier Ltd.

Abstract

The Cuban Revolution imbued confidence to reformists and revolutionaries in 1959. Castro's challenge to American hegemony and promises of an independent socialist society came to naught as Cuba became a satellite of the Union of Soviet Socialist Republics (USSR). The breakdown of the Soviet bloc and USSR's collapse ended the subsidies that sustained the system despite the embargo enforced by the United States of America. Cuba's tourist industry, export of Cuban labor, the Internet, and the regime's diminishing coercive power have fostered internal challenges. In 2008, Fidel Castro turned power over to Raúl Castro, his octogenarian younger brother, who started limited economic reforms in an attempt to ensure the regime's survival under the Party's dictatorship. The 2011 economic reforms threaten the safety-net that characterized the revolution.

Introduction

The Cuban Revolution infused confidence throughout the world in 1959. Fidel Castro's challenge to American hegemony and promises of an independent socialist, egalitarian society came to naught by the late 1960s as Cuba became a Caribbean satellite of the Union of Soviet Socialist Republics (USSR). The revolution's draconian political and economic policies soon prompted a massive entrepreneurial and professional middle class exodus despite the government's ban on travel abroad. The dearth of qualified labor had grave consequences for the island's economy. By 1980s, the breakdown of the Soviet bloc and the USSR's collapse ended the magnanimous subsidies through which the regime survived. Failed economic programs and military adventurism caused the citizenry's abject poverty and dependency on remittances sent by relatives in the diaspora. The exploitation of Cuba's natural tourist attractions, the exportation of Cuban labor to obtain foreign exchange, and the regime's diminishing coercive power have promoted internal challenges, as the population comes in contact with the free world. In 2008, Fidel Castro turned power over to Raúl Castro, his octogenarian younger brother, who started economic reforms in an attempt to ensure the regime's survival under the Party's dictatorship.

From Spanish Colony to Republic

From 1895 to 1897 the Cuban war for independence from Spain had become a protracted struggle. US intervention and occupation ended the war in 1898 and although Cubans constituted the land forces facilitating the swift American victory, they were ignored during the Treaty of Paris negotiations. The US aided the war-torn country, improving health conditions and education. American investments and return of Spanish capital contributed to economic recovery. US occupation ended in 1902 after Cuban delegates to the 1901 Constitutional Assembly capitulated to American pressure and appended the Platt Amendment to the Cuban Constitution. The amendment, imposed by the US as a requisite for ending the occupation, gave Americans the power to intervene militarily and restricted Cuba's international relations.

Traditional and pseudo-Marxist historiography reduces US–Cuban relations to events during and after 1898. This perspective marks 1898 as the date when Cuba was no longer a Spanish colony and became a US protectorate. The truth is more complex. Since the mid-nineteenth century, despite being a Spanish colony, the island's economy depended on US trade. Cuba was the world's foremost producer of sugar and the US imported 90% of Cuba's production. The US was a refuge for Cubans aspiring independence and, under José Martí's leadership, Cuba's first political party was created in the US. Martí, although not anti-American, opposed US intervention. Economic ties continued to play a role in US–Cuban relations during the Republican Era.

The Early Republic

Liberalism, electoral fraud, administrative graft, and civil unrest characterized the early republic. Creole elites were nearly destroyed as Spain did not compensate Cubans for property losses. The postindependence concentration of US and Spanish capital in the sugar industry, together with the Spaniards' control of banking and commerce, made access to public office a vehicle for upward mobility for Cubans. Armed struggle was often a means to ensure circulation of elites in government. After intervening militarily in several occasions, the US occupied the island for a second time from 1906 to 1909.

By 1902, Cuba had become the world's second largest sugar producer despite extensive war damages to its agriculture. Although plantations were ravaged by fire, the sugar industrial sector sustained limited damages. Only seven of the 50 largest mills were destroyed and four experienced minor losses (Moreno Fraginals, 2001: p. 485). A Cuban-born entrepreneurial class emerged as sugar production and prices rose during and after the First World War. In the early 1920s, the Cuban banking crash and the collapse of sugar prices plunged

[†] Deceased.

the nation into chaos. By 1923, widespread discontent led to demands for curtailment to foreigners' concessions, abrogation of the Platt Amendment, and an end to sugar dependency.

The 1933 Revolution

Gen. Gerardo Machado's dictatorship (1925–33), US interference, and plummeting sugar exports led to the 1933 Revolution and a provisional government installation. On 14 September declaring Cuba would only honor "spontaneously" signed treaties, Ramon Grau's administration unilaterally repealed the Platt Amendment (Carrillo, 1985: p. 267). Cuba's emissary to the VII International Conference of American States on the Rights and Duties of States ratified Cuba's Platt Amendment's repeal on December 1933 in Montevideo. Opposition to the social, economic, and political reforms enacted by Grau's administration prompted the US Ambassador's request for American military intervention. Instead, the US withheld diplomatic recognition of Grau's government. On January 1934, Sgt. Fulgencio Batista, promoted as army staff chief, deposed Grau with US Embassy support. The 1933 Revolution challenged US hegemony, introduced socioeconomic and political structural changes and established, for the first time in history, a free Cuba. On 29 May 1934, a new US–Cuba treaty was signed, ratifying Cuba's repeal of the Platt Amendment, which US had tacitly agreed to rescind when US Secretary of State Hull signed the Conference's Final Act in Montevideo on December 1933 (Portell-Vila, 1986: p. 431).

The Republic's Failure

Batista ruled by proxy until elected president in 1940. In 1944, the Partido Revolucionario Cubano (Auténtico) came to power when Dr Ramón Grau was elected president. The reformist policies enacted during Grau's previous short presidential term had been incorporated into the 1940 Constitution. Unfortunately Grau's second term and subsequent Auténtico government of Carlos Prío (1948–52), although infused by the 1933 Revolution's agenda, were characterized by public sector graft, corruption, and political gangsterism.

In 1947, Eduardo Chibás, founded the Partido del Pueblo Cubano (Ortodoxo), which embodied the hopes of a generation for whom the 1933 Revolution's promises remained unfulfilled. Honesty in public office was the core of Chibás's movement and his 1951 suicide created a leadership vacuum. On 10 May 1952, 2 months before elections, a coup led by the ubiquitous Fulgencio Batista, ended the constitutional republican experience. On 26 July 1953 in Santiago de Cuba, Fidel Castro led the unsuccessful attack on the Moncada barracks, which drew publicity to him and his Twenty-sixth of July Movement (M-26-7). After Batista refused to concede the demands for free elections by a coalition of leaders seeking a nonviolent solution to the crisis, violence escalated. The Federación Estudiantil Universitaria (FEU) called for a strike. In 1956, Castro and José A. Echeverría, FEU president and leader of the Directorio Revolucionario Estudiantil (DRE), joined forces to depose Batista. By December, Castro had established guerrilla operations in the mountains of Sierra Maestra. On March of 1957, the Auténticos and the DRE, without collaboration from M-26-7, stormed the presidential palace and in a separate incident during the failed assault Echeverría was killed. As government repression mounted, professionals and entrepreneurs joined the fight and other guerrilla fronts emerged. Between 1956 and 1957 conspiracies were discovered in the armed forces, and the urban underground initiated a terrorist campaign. In 1958, the US imposed an arms embargo on Batista's government. A military offensive against the rebels failed and the rebels mounted a counteroffensive. At dawn, on 1 January 1959, Batista fled the country leaving behind a provisional junta. His escape also left an astonished and enraged Castro as he found himself in Oriente province, over a 1000 km away from the center of power while the regime collapsed.

Communist Revolution

After a short period of confusion as anti-Batista groups vied for power and Batista's army surrendered to Castro's rebel forces, the unanticipated communist revolution began. The Cuban Communist Party, Partido Socialista Popular (PSP), barely participated in the struggle against Batista until 1958 and Castro emphatically denied any communist inclinations. Castro's social and economic programs were based on the implementation of the 1940 Constitution's provisions.

On the revolution's eve, Cuba was a developing nation ranked among the top five Latin American countries. The movement to depose Batista had a heterogeneous class base, although the urban underground was primarily comprised by members of Cuba's growing middle class. The movement was neither antibourgeois, nor anti-American. Castro's M-26-7, as well as other anti-Batista groups, focused on issues concerning the middle classes: political freedom, constitutionality, and fiscal integrity.

Origins of the Revolution and Leadership Legitimization

The complexity of Cuban society and of revolutions in general, becomes evident as one surveys the literature on the causes of the Cuban revolution. Draper (1962) claims it was a middle-class phenomenon. Suárez (1967), among others, argues it was the product of Castro's charismatic leadership and political maneuvering. To Aguilar (1972) it was the frustrated 1933 revolution and the absence of reforms. Domínguez (1978) emphasizes that the fragmentation and individual strength of Cuba's civil society in prerevolutionary politics allowed the revolutionary process to crystallize after 1959. Farber (2006) argues the stage for a revolution was laid in the republican period and Castro, once in power, counterbalanced US belligerence with Soviet support.

Cuba's official historiography proclaims a single revolutionary movement beginning in 1868, combining two separate attempts at independence despite their diverse regional and class composition, and ends with Castro's 1959 triumph. Thus both anticolonial nationalist wars, seeking the establishment of an independent democratic republic are portrayed as a single socialist revolution. In 1968, the leadership reinstated the 10th of October, a patriotic date marking the beginning of the Ten Years War, to proclaim themselves the true liberators of Cuba 'after 100 years of struggle." Through symbolic manipulations

Cuba became synonymous with revolution and the revolution synonymous with Fidel Castro (Pérez-Stable, 1999).

Revolutionary Dynamics

The leadership implemented typical totalitarian strategies. Guevara expounded on the need to create a new economic base and a 'new man.' Policies reminiscent of fascist tactics sought to uproot prerevolutionary organizational and interpersonal patterns. The institutions of work and education, radically transformed, became the focus of Cuban life. The government monopolized all means of communication. Slogans with the names of key figures were incessantly repeated through all forms of media (Medin, 1990). Attempts were made to destroy the traditional family structure and institutionalized religion was crippled (Pedraza, 1998). Mass organizations were responsible for transmitting the new ideology, mobilizing the population, and maintaining constant vigilance. The trials, executions, long jail sentences and feelings of constant surveillance unleashed the 'terror' stage. Political parties, elections, freedom of the press, labor's right to strike were eliminated, and universities and unions purged. The 1940 Constitution was proclaimed outdated, and the US became Cuba's enemy. All criticism was counterrevolutionary and political opponents were imprisoned or executed. A pattern of negative integration (Groh, 1973) developed as Cubans that opposed the system and were unable or unwilling to leave, survived by pretending acquiescence.

Economic Crisis and Revolutionary Euphoria (1959–70)

Between 1959 and 1961, Cuba was transformed. Foreign and domestic properties were expropriated and Cuba's foreign relations changed rapidly. Soviet military advisers and weapons started arriving by September of 1960. In October, the US imposed a trade embargo and by January 1961 Cuban–US diplomatic relations were severed. The Bay of Pigs invasion's failure and concomitant detention of approximately 100 000 opponents crushed the underground cementing the revolution. The massive urban middle class exodus intensified with the departure of those who had fought against previous oppressive regimes, leading to the creation of a Cuban diaspora of more than two million exiles. On the invasion's eve, Castro proclaimed, for the first time, the revolution's socialist nature. One month later, Cuba–USSR relations were reestablished. The Soviet connection may have been initiated in the late 1950s by Raúl, Fidel Castro's younger brother through contacts with the international communists while at Sierra Maestra (Bonachea and San Martin, 1974). On 2 December 1961 Castro declared himself a Marxist-Leninist.

A crisis between the superpowers ensued when a US reconnaissance plane discovered Soviet missiles and launching sites in Cuba. The Missile Crisis's outcome altered Cuba's foreign policy. Castro felt betrayed, although all viable external threats disappeared when the US agreed not to invade Cuba in exchange for the missiles' removal, as the compromise was reached without consulting him. Castro reacted by imprisoning the island's pro-Soviet communists and seeking new allies throughout Latin America. The USSR retaliated by decreasing petroleum deliveries to Cuba compounding the catastrophic failure of the revolutionary economic policies. In 1968, to stay in power, Castro affirmed Soviet hegemony by supporting the Soviet invasion of Czechoslovakia. Cuba's ill-conceived strategies of attempting to industrialize, while neglecting sugar production and exchanging the US as trading partner (140 km away) for the USSR (over 4500 km away), resulted in mounting indebtedness. Although the sugar industry had been devastated since 1964, a ten-million ton sugar harvest was programmed to solve the crisis.

As revolutionary euphoria subsided, all those demonstrating disaffection were drafted into Units to Aid Production (UMAP). From 1965 to 1967, priests, seminarians, homosexuals, juvenile delinquents and those deviating from socialist orthodoxy were placed in UMAP concentration camps. In 1968, the 'Revolutionary Offensive' eliminated all vestiges of economic capitalism by closing remaining small businesses. Workforce absenteeism, low productivity, and poor product quality were the population's reaction to moral incentives.

Institutionalization and Foreign Military Adventurism

Material incentives returned by 1970, when Soviet type economic models were introduced. In 1974, Brezhnev's visit strengthened Cuba–Soviet relations. One year later, Cuban African campaigns began with Soviet support. By the 1980s, Soviet subsidies reached $3 billion a year. The USSR supplied 90% of the island's oil needs and an inordinate proportion of basic foodstuffs. Although this aid improved living standards, over 120 000 Cubans abandoned the island during the 1980 Mariel Boatlift. In 1986 Castro launched the 'Rectification of Errors Campaign' eliminating material incentives and market reforms. Mass mobilization and volunteer work were reintroduced. Castro's policies, in conjunction with the Soviet bloc breakdown and subsequent USSR collapse, propelled a severe economic crisis. The US embargo became a weapon for the first time since its inception after Cuba ceased receiving Soviet subsidies. In 1989, General Arnaldo Ochoa and Colonel Antonio de la Guardia were accused of drug dealing and breaching national security. The dismantling and purging of several government agencies followed their trial and execution, as growing dissension could no longer be ignored.

Last Decade of the Twentieth Century

In 1990 the 'Special Period in Peacetime' was declared and policies were adopted to develop an autarkical state. To alleviate food and energy shortages, thousands of urbanites were moved to the countryside to work in food production. Austerity and mass mobilization were expected to solve the country's crisis, until foreign investments, tourism, and an incipient biotech industry could generate the necessary hard currency.

Castro declared his unswerving commitment to socialism in 1991; vowing that neither political nor further economic reforms would be implemented. By 1993 the mounting crisis made economic reforms inevitable. Farmers' markets, duty-free zones and 'dollarization' were instituted. Access to dollars, mainly through exile remittances, tourism, and illicit activities, such as prostitution and black-market participation, provided

economic respite to a segment of the population while contributing to the public's growing awareness of emerging socioeconomic disparities. Since the late 1960s, upper government cadres had been the only groups known to enjoy goods prohibitive to the rest of the population; access to hard currency now allowed other segments to acquire those goods.

As Castro suspected, the government's incapacity to continue its role as sole provider of the nation's needs eroded the state's coercive power. Religious participation increased, critics became more vocal, human rights groups multiplied, and short-lived rebellions erupted. In 1994, another mass exodus occurred. Over 30 000 people, in unique makeshift inner tube rafts, abandoned Cuba.

The embargo was reinforced by the 1996 Helms-Burton Act after Cuban MiGs downed two light planes belonging to 'Brothers to the Rescue,' an exile organization flying over the Strait of Florida searching for rafters. Regime criticism increased as International Labor Organization (ILO), Amnesty International, and the United Nations repeatedly condemned Cuba's human rights violations. Pressures for the US to lift the embargo also proliferated. In 1998 the Christian Liberation Movement in Cuba launched the Varela Project under the leadership of Oswaldo Payá, who opposed both the US embargo and Cuba's human rights violations. The Varela Project, aimed at amending the Cuban constitution, gathered the constitutionally required number of signatures for a referendum. To date, the government has refused to bring the proposal up for discussion at the National Assembly. The twentieth century came to a close with the continued isolation of Cuba, despite Pope John Paul II's bid for Cuba to open to the world and for the US to end the embargo during his 1998 visit.

The Cuban Revolution in the Twenty-First Century

The perpendicular decline of Cuba's floundering economy since the 1980s continued into the next century despite sporadic hiatuses of recovery. The crises provoked by the 9/11 terrorist attacks and the 2008 global depression further exacerbated the island's inherent economic woes.

Old and New Problems

When Castro announced the 'Special Period in Peacetime' in 1990 no one anticipated that the catastrophic conditions forcing such measures would return in the twenty-first century (Gonzalez-Corzo, 2010). Assuming control, Raúl Castro acknowledged both the failure of the socialist economic experiment and that saving socialism would require market-oriented capitalist reforms. A total of 1.8 million public sector jobs were earmarked for elimination by 2014 and more occupations, mostly service categories (*cuentapropistas*) legalized in 2010, were added to the 178 list of self-employments (Scarpaci, 1995). Decriminalized occupations authorized by these guidelines had long been part of the underground economy and the government now hopes to benefit through imposed heavy taxation on earnings. Strict regulations on *paladares* (small restaurants) licensed in 1993 were also relaxed (Henken, 2011).

Foreign Exchange

Although sugar was the foundation of the Cuban economy since the 1800s and the island the "world's sugar bowl," revolutionary policies and technological obsolescence led to the importation of sugar to satisfy internal demand by 2011. Since the late 1970s, numerous efforts had been made to develop other hard currency sources beyond the traditional exports of sugar, rum, tobacco, and minerals. A biotech industry, tourism, oil resales, joint ventures, and the commodification of labor are among those economic strategies. With the exception of biotechnology, the rest achieved some success.

Tourism

For 15 years, tourism, which had been an important revenue source since the 1920s, was eliminated through revolutionary policies and the US–Cuba conflict. In the 1970s, sugar no longer satisfied hard currency needs and the Instituto Nacional de Turismo (INTUR) was established. Tourism was once more a national industry (Aagesen and Paluch, 2000).

Legalization of foreign ventures in the 1980s revitalized tourism. The four S's of Cuban tourism: sea, sun, sand and sex were again internationally recognized. Tourism and prostitution became linked as they were prior to 1959, but tourism presented new dilemmas for the socialist government as capitalism gained a stronghold within the island. The initial solution was to develop tourist enclaves only for foreigners as Cubans were denied access to these amenities (Pérez-Lopez, 1994). The Cuban "apartheid" became a domestic political issue. Economic inequalities also surfaced between those permitted to work in tourist facilities and those barred. After a meteoric rise, the tourist industry stagnated mainly due to the 9/11 terrorist attacks, the 2008 global recession, and growing tourist dissatisfaction with services provided.

Oil Resales and Joint Ventures

As Yanes (2005) indicated, baptizing the 2005 yearly slogan as the "Year of the Bolivarian Alternative for the Americas" served to acknowledge Venezuela's crucial role in the survival of Castro's regime. Venezuelan subsidized oil solved the domestic energy crisis while providing Cuba access to hard currency through its resale; estimated to represent US$1000 million dollars annually. Cuba's indebtedness to Venezuela is evocative of the Soviet–Cuba alliance. The Institute for Cuban and Cuban–American Studies reports that by 2001, after European Union (EU) members' credits extended to Cuba were withheld due to repeated payment defaults, and Venezuela displaced Spain as Cuba's top trading partner (Cuba Transition Project, Institute for Cuban and Cuban–American Studies, 2002). In 2009, Cuba–Venezuela trade reached $10 400 million (Romero, 2011). Oil refineries are being built with Venezuelan and Chinese funds in both coasts and the possibility of finding significant oil deposits in Cuba's north coast is being explored.

Commodification of Cuban Labor

During Castro's 1962 medical faculty inaugural address, Cuba's internationalist proletarian policy was announced (Maury del

Toro, 2010). Revolutionary internationalist solidarity has brought the regime such a political and economic bonanza that medical schools are mushrooming in Cuba. By 2008 revenues from export of labor had tripled tourism receipts. Cuban proletarian exports include physicians, dentists, nurses, medical technicians, teachers, engineers, architects, and other technical personnel. Medical brigades have been sent to Latin America, Africa, Asia, and the Middle East. Sending emergency brigades mostly financed by third party payments to Cuba or by the country receiving the services during disasters, opens the door for government-to-government accords. Developed countries and international organizations provide the lion's share of funds financing Cuban brigades, while the recipient countries mostly provide personnel housing and board. Cuba's missions have been funded by Brazil, France, Germany, Japan, Norway, South Africa, Zimbabwe, and Venezuela. The Pan American Health Organization (PAHO), United Nations Children's Fund (UNICEF) and World Health Organization (WHO) are among major international institutions bankrolling Cuba's health brigades (Werlau, 2011). Cuba's solidarity is humanitarian, but it is not free.

These agreements violate International Labor Organization (ILO) statutes, and the International Declaration of Human Rights (De Cuba, 2011). Cuban personnel do not participate in the accords. Neither their salaries nor working conditions are negotiated by them. Their pay rate is the Cuban $200 a month salary, although funding agencies pay the Cuban government over 10 times the amount. They are forbidden to move outside the deployment area; required to maintain surveillance on all mission members; report host country conditions, and are under Cuban security forces' strict control (Werlau, 2010)

Domestic Politics

After five decades in power, the old and fatigued revolutionaries continue to reaffirm that despite growing demands for freedom there will be no political reforms. Economic reforms are means to build socialism under Party control. Marino Murillo, at 51 one of the youngest recently appointed Politburo members, stated that while Cuba is liberalizing its economy and welcomes foreign investor partners, no concomitant changes in the political sphere will occur. Murillo's declarations were the government's response to Pope Benedict XVI's calls for Cuba to evolve into an open society during his March 2112 visit (CBS, www.cbsnews.com/2102-202_162-57405199.html)

New Challenges

The greatest challenges faced by Cuban authorities are internal corruption and claims from sectors for which the promises of the revolution were not fulfilled. Afro–Cubans, women, and youth are making their demands known through the new media at their disposal. Despite government's efforts to maintain Cubans under the revolutionary rhetoric's spell, a new revolution beyond its control and its borders infiltrated the island. The Internet and cellphones are providing the population with uncensored information and means to make their frustrations known.

Corruption

All governments deal with corruption and centrally planned economies concentrating power in the hands of government bureaucrats are plagued by it. Cuba is no exception, despite Castro's commitment to end corruption. By 1981, 92% of the workforce was employed by the state, making the majority of workers public bureaucrats. During a 20-year period (1968–88), all economic activity came under the state's control, with the exception of agriculture and transportation in which the government exercised 97 and 99% control respectively (Diaz-Briquets and Perez-Lopez, 2006). Chronic consumer goods scarcity experienced since the 1960s turned diversion of state resources, graft, theft, bribing and participation in the black market a common practice.

The danger corruption poses to a revolutionary discourse focused on the creation of an egalitarian society is obvious to regime's supporters. An island academic published an article questioning whether corruption could be the real counterrevolution, as it is entrenched within government structures (Morales-Dominguez, 2010). Citizens' awareness of privileges enjoyed by groups at the top echelons has been documented through interviews of those leaving the island since the 1970s; and, by the 1990s, dollarization and tourism expanded the number of groups perceived as privileged beyond the confines of the revolutionary elite (Clark, 1999). Public accusations of malfeasance and anticorruption campaigns have been made by both Castro brothers since the 1960s. Raúl Castro initiated a crackdown on corruption after succeeding his brother by purging government officials and foreign companies' executives (Acosta, 9 November 2011).

Race Relations

Cuba, as common to other societies in which African slavery was practiced, has a legacy of racial prejudice and discrimination. Cuba's difference is that no other society claims eliminating racism by decree. Continued discrimination charges, including police racial profiling, are widespread (Miami Herald Staff, www.miamiherald.com/multimedia/news/afrolatin/part4). Blacks also claim they are discriminated from hard currency economic sectors, such as tourism.

Despite socioeconomic gains, the proportion of blacks at the top of the power structures is negligible and the advances achieved through policies instituted to improve the poor's living standards, where the majority of the black population concentrated, have been eroding since 1980s economic crisis of the (Morales-Dominguez, 2007). Exile remittances are estimated at $1.5 billion a year, but less than 15% among Cuban exiles in the U.S are black or mestizo (Ennis et al., 2010 Census Brief: p. 15).

Estimates of Cuba's racial composition vary according to the criteria applied to determine racial boundaries. If the North American dichotomy is used, the majority of Cubans would be considered black. Applying the Iberian classification based on phenotype (Hoetink, 1967), which is normally used by Cubans, the population is segmented into three layers in which the white and mestizo groups comprise a statistical majority. Thus, the 2002 Cuban Census reported that the population was: 65.1% white; 24.8% mestizo; and only 10.1% black. The

only segment that, according to Cuban census figures, has unequivocally continued to increase is the intermediate mixed category. Since 1953 the white population decreased by 8% and the black segment dropped 2%. The political realities of a society in which race continues to matter are palpable by the government's attempts at inclusion of more members of the mestizo and black population in the power structure. Raúl Castro promoted some blacks, mostly women, to Council of Minister posts (Partido Comunista de Cuba, www.pcc.cu/eo_secretariado.php). The Council now counts with 31 white, 3 black; and 2 mestizo members. The 14-member Politburo maintained its quota of 2 black males, and added a new female mestizo member. A similar maneuver was implemented at the Party's Secretariat, in which out of the 7 officials that compose it, the sole black male – also a member of the Politburo – was joined by a younger mestizo female (Partido Comunista de Cuba, www.pcc.cu/eo_buro_politico.php). Black dissatisfaction with the state of affairs is evidenced by increasing numbers of blacks among dissidents and the creation of black civil rights groups. Some of the best known figures within the growing dissident movement are black, such as physicians Oscar Elías Biscet and Darcy Ferrer, journalists Guillermo Farinas and the late Orlando Zapata Tamayo who died while on a hunger strike in prison.

Women: The Birth Dearth and Political Protest

A patriarchal ideology always framed women's position in Cuba and the 1970s militarism further eroded the status of women. The initial objectives of the Federation of Cuban Women (FMC), founded under the direction of Raul Castro's late wife, Vilma Espín (1930–2007), were to develop women's socialist consciousness and to organize them in defense of the revolution. Cuban women would be repositories of Cuban-Marxist-Leninist ideology and reproducers of human capital for revolutionary purposes. FMC's membership grew throughout the 1970s, and women were encouraged to do volunteer work and join the workforce. The renewed emphasis on developing an economy based on sugar production forced women to join the urban workforce to fill vacancies created by the transfer of men to agriculture during the Ten-Million-Ton-Sugar Harvest.

The FMC defense of women's rights partly eliminated discriminatory policies proscribing women from hundreds of jobs. Women's government employment rose in tandem with educational levels. Still, patriarchal values continued to undermine women's equality as men refused to share domestic obligations. Women's burden increased exponentially as they had to allocate time to family, work, education, and revolutionary responsibilities. In the late 1980s female academicians addressed Cuban gender disparities by initiating research ranging from domestic violence to women's double-shift (Acosta, 12 October 2011). Interest in gender studies coincided with the alarm over a sustained drop in birth rates.

Cuba's demographic transition started early in the twentieth century due to economic downturns of the 1920s and 1930s and growing women labor market participation. The crude birth rate in 1958 was 28.1 per 1000; well below rates for other developing countries. After a postrevolutionary baby-boom (1959–66), birth rates rapidly declined. Birth rates have been below replacement level since 1978. Several explanations have been advanced for this decline. Alfonso de Armas (2009) enumerates public health care availability, sex health education, increasing gender egalitarianism, and access to legal abortions with few restrictions; Andaya (2007) maintains it is linked to the convergence of modernization and underdevelopment; and Díaz-Briquets and Pérez (1982) believe that unmet rising expectations is the key determinant.

Cuban women claim they would have more children if they could provide for them (Grogg, 2009). They cite lack of resources, low salaries that hardly cover basic needs, chronic food shortages that consume most of their income and day searching from government stores to black market providers; and absence of adequate housing and daycare (Sánchez, Y., 2011). Most unwanted pregnancies end with abortion. Public Health Ministry (MINSAP) 2008 data indicate that there were 23.0 abortions per 1000 women aged between 12 and 49 years old (Gran Alvarez, 2005). Under the euphemism of "menstrual regulation service," vacuum-aspiration abortion is available up to 7 weeks without confirmation of pregnancy. This method, implemented in 1968, is the prevalent procedure for ending pregnancies. The fertility decline coupled with an aging population and the migration of younger cohorts threatens to invert the island's age pyramid (Coyula March 2010).

In 2003, after a group of 75 dissidents including journalists, librarians, and human right activist were tried and condemned to prison terms ranging from 6 to 28 years, their female relatives started peacefully marching to demand their release. The group came to be known as the Ladies in White. Their founder, the late Laura Pollán, was awarded the 2005 Sakharov Prize, which the Cuban government did not allow her to receive by denying her permission to travel abroad. Originating in Havana the group spread to other cities. The men were dispersed throughout Cuba's prison system; isolating them from their local communities. This strategy, designed to also inconvenience relatives by forcing travel in the midst of Cuba's transportation and economic crisis, had the unintended consequence of contributing to the movement's proliferation as the women networked with other convict's female relatives. Despite government's repression and Laura Pollán's death in October of 2011, the Ladies in White continue defending prisoner's rights.

Human Rights

The number of political prisoners in Cuba remains undetermined. The Cuban Criminal Code stipulates that any potential violation of norms is a criminal offense identified under the law of dangerousness, which can lead to detention, trial, and conviction (Amnesty International 2011 Report). Citizens distributing leaflets with the Universal Declaration of Human Rights, gathering signatures for a referendum, or attempting to organize unions have been detained and convicted under this law. Lack of participation in official events, moving outside of one's registered domicile and even unemployment due to disability have been cited as crimes under this code leading in some cases to "official warnings" and in others to years of imprisonment (World Report: 2012. Cuba, Human Rights Watch). Cuba prohibits all political dissent. Cubans that challenge the system face charges and

risk enduring verbal attacks, beatings, recurrent short-term detentions, and years behind bars. The release of political prisoners is usually contingent on forced exile. Although Cuba refuses access to prisons by international human rights organizations, information released by domestic human rights defenders, by ex-prisoners or by their relatives, indicates that Cuban prisons are unsanitary and overcrowded. Prisoners are denied medical care, subjected to beatings, prolonged solitary confinement, and forced to submit to ideological rehabilitation.

In violation of the Universal Declaration of Human Rights, travel within the island without official authorization is restricted by Decree 217 (Human Rights Watch, 1999). Cuban citizens are also forbidden to leave or return to the island without official permission.

The World Wide Web, Cellphones, and Bloggers

Cubans were semiisolated from the world for decades. The revolution in global communications coincided with the socialist camp demise and Cuban authorities' needs for hard currency. The tourist industry demanded a communication infrastructure for hotels and areas linked to the tourist economy; and, as spaces emerged beyond the government's absolute control, the population gradually initiated national and international contacts. Bloggers committed to free expression have ingenuously devised channels to publish abroad.

As the power to coerce the population through its absolute control of the economy diminishes, the state now relies on tactics common to authoritarian regimes, such as public acts of repudiation, intermittent short-term arbitrary arrests and grueling interrogating sessions. Presently, the victim's friends or relatives can report these violations to human right groups abroad mostly through cellphones. Public verbal and physical attacks of dissenters by mobs of government supporters, incited and organized by security police posing as civilians, have been photographed or videotaped by witnesses and transmitted via cellphones to the rest of the world; and, although bloggers are harassed, detained, and subjected to government slander, the government has been unable to silence them. Cuba's best known blogger, Yoani Sánchez was awarded the 2009 Maria Moors Cabot Prize by Columbia University's Graduate School of Journalism and Guillermo Fariñas, independent journalist and member of the 75 dissidents tried, convicted, and imprisoned during the 2003 Black Spring, was conferred the 2010 Sakharov Prize for Freedom of Thought. Both were denied permission to travel and were unable to receive their awards. In a crackdown prior to Pope Benedict XVI's visit, Cuban police detained and intimidated dissidents. While the pope visited Cuba, in an effort to block any uncensored coverage of the events, all calls to dissidents' cellphones were diverted to the Interior Ministry (Tamayo, *Miami–Herald* 2012). The Ladies in White contacted foreign media sources before the pope's arrival to report the ongoing harassment campaign; and, during the papal mass in Santiago de Cuba on 26 March, Andres Carrión was caught on tape as he was dragged by the secret service and beaten by a Cuban Red Cross employee while shouting antigovernment slogans (Leon, Miami-Herald 2012).

US–Cuban Relations

From the 1963 Missile Crisis to the 1980 collapse of the Berlin Wall, Cuba's safety was guaranteed by a standoff between the superpowers. If the Soviets moved on Berlin, the Americans would invade Cuba (Freedman, 2000). At the end of the Cold War, Cuba found itself an orphan in a changing world. Cuba was no longer a prize to be had, nor a threat to Latin American stability. The US embargo began to exact a toll and Washington, despite its uncompromising rhetoric saw no reason to deny American businesses the benefits from a market 140 km from its shores. The Trade Sanctions Reform and Export Enhancement Act (TSRA), signed into law in 2000, provided the legal framework for renewing trade with Cuba. The United States now ranks as Cuba's fifth trading partner, while Cuba ranks among the top 10 importers of American agricultural and food products. The embargo, which only remains as a symbolic sanction, has provided Cuba the excuse for its economic failures during five decades.

In 2008 two major changes in leadership occurred as Castro's resigned in favor of his younger brother Raúl and Barack Obama's election brought the Democratic Party back to the White House. Neither of these changes altered US–Cuban foreign relations significantly. Although the Obama administration seemed to be willing to broaden the dialogue with Havana – the stringent restrictions imposed on travel to Cuba under Bush's administration were eliminated and the dollar amount of remittances sent from the US to the island was increased – the 2009 Cuban government's detention of Alan Gross, an American that regularly traveled to Cuba, hindered further rapprochement (Chavez and Tamayo, 2011). Gross' detention coincided with the 2009 US Supreme Court decision not to hear the appeal of members of a Cuban spy-network operating in the US (Chacón, *Boston Globe*, 2001) Gross, identified by Cuban authorities as an American agent, was subsequently tried, and convicted to 15 years for crimes against the state. US–Cuban relations have reached an impasse as both governments continue to pressure for the release of their respective agents as a prior condition for negotiations.

Conclusion

Cuba entered the second decade of the twenty-first century under the control of a military gerontocracy headed by the Castro brothers and what remains of their loyal clique of 1960s revolutionaries. The revolution's greatest success continues to be this group's ability to stay in power. The regime's coercive power, although diminished, is still able to crush dissent through intimidation or outright violence. The Cuban chronic economic plight and reliance on external support reminds one of Blanche DuBois line "Whoever you are, I have always depended in the kindness of strangers" in *A Streetcar Named Desire* by Tennessee Williams. The system survives through the aid or charity of other nations; with other people's money, be it Soviet, Venezuelan, Canadian, Spanish or the hard currency earned through the exploitation of Cuban medical and technical personnel sent as humanitarian missions; and last, but not least, the remittances sent by Cubans residing in the

constantly growing diaspora. The recent reforms include substantial cuts in health care, education, employment and food subsidies. These austere policies undermine the meager safety net the revolution provided. If a private sector is to satisfy the basic physical and nutritional needs of the citizenry, the reforms would have to go beyond the paltry measures adopted to liberalize agricultural production and real estate sales.

See also: Civil Liberties and Human Rights; Cold War, The; Communism, History of; Diaspora; Discrimination: Racial; Freedom of the Press; Social Movements: A Social Psychological Perspective.

Bibliography

Aguilar, L.E., 1972. 1933 Prologue to Revolution. Cornell University Press, Ithaca.

Alfonso de Armas, M., 2009. Family planning and reproductive and sexual health in Cuba. In: Demographic Transformations and Inequalities in Latin America. Historical Trends and Recent Patterns, pp. 315–319. Retrieved 03.02.12, from: http://www.alapop.org/docs/publicaciones/investigaciones/DemogTransformations_Presentation.pdf.

Andaya, E., 2007. Reproducing the Revolution: Gender, Kinship, and the State in Contemporary Cuba. New York University, 336 pages; AAT 3286472.

Bonachea, R.L., San Martín, M., 1974. The Cuban Insurrection, 1952–1959. Transaction Books, New Brunswick.

Carrillo, J., 1985. Cuba 1933. Estudiantes, Yanquis Y Soldados, p. 267. Instituto de Estudios Interamericanos, University of Miami, Miami.

Clark, J., 1999. Igualdad y privilegio en la revolución de Castro. In: Cordova, E. (Ed.), 40 años de revolución: El legado de Castro. Ediciones Universal, Miami, pp. 219–252.

Coyula March, M., 2010. Havana: aging in an aging city. MEDICC Review 12 (4), 27–29. Retrieved 18.02.12, from: http://www.medicc.org/mediccreview/articles/mr_167.pdf.

De Cuba, P., 2011. La esclavitud moderna: apuntes sobre el caso de los médicos cubanos. In: Cuba in Transition, vol. 21, pp. 213–221.

Díaz-Briquets, S., Perez-Lopez, J., 2006. Corruption in Cuba: Castro and beyond, first ed. University of Texas Press, Austin, TX.

Diaz-Briquets, S., Pérez, L., 1982. Fertility decline in Cuba: a socioeconomic interpretation. Population and Development Review 8 (3), 513–537.

Dominguez, J.I., 1978. Cuba: Order and Revolution. Harvard University Press, Cambridge.

Draper, T., 1962. Castro's Revolution: Myth and Realities. Praeger, New York.

Ennis, S., Rios-Vargas, M., Albert, N., 2011. 2010 Census Briefs. United States Census Bureau.

Farber, S., 2006. The Origins of the Cuban Revolution Reconsidered. University of North Carolina Press, Chapell Hill.

Freedman, L., 2000. Kenned'ys Wars. Berlin, Cuba, Laos, and Vietnam, p. 227. Oxford University Press, New York.

González-Corzo, M.A., 2010. Cuba: ¿hacia otro "periodo especial"?. In: Cuba in Transition, vol. 20, pp. 88–96. Retrieved 16.12.12, from: http://www.ascecuba.org/c/wp-content/uploads/2014/09/v20-gonzalezcorzo.pdf.

Gran Álvarez, M.A., 2005. Interrupción voluntaria de embarazo y anticoncepción. Dos métodos de regulación de la fecundidad. Cuba, 1995–2000. Ministerio de Salud Pública, La Habana.

Groh, D., 1973. Negative integration und revolutionärer Attentismus, 1909–1914. Frankfurt.

Henken, T., 2011. Santiago vs. Havana: Cuba's underground economy. In: Cuba in Transition, vol. 13, pp. 348–353. Retrieved 16.12.11, from: http://www.ascecuba.org/publications/proceedings/volume13/pdfs/henken.pdf.

Hoetink, H., 1967. The Two Variants in Caribbean Race Relations: A Contribution to the Sociology of Segmented Societies. Oxford University Press, London.

Medin, T., 1990. Cuba: The Shaping of Revolutionary Consciousness. Lynne Rienner, Boulder.

Morales-Domínguez, C.E., 2007. Desafíos de la problemática racial en Cuba. Fundación Fernando Ortiz, La Habana.

Moreno-Fraginals, M.R., 2001. El ingenio: Complejo económico-social cubano del azúcar, p. 485. Crítica, Barcelona.

Pedraza, T., 1998. "This too shall pass": religion in Cuba, resistance and endurance. In: Santi, E.M. (Ed.), Cuban Studies, 28. University of Pittsburgh Press, Pittsburgh, pp. 16–39.

Perez-Lopez, J.F., 1994. Islands of capitalism in an ocean of socialism: joint ventures in Cuba's development strategy. In: Perez-Lopez, J.F. (Ed.), Cuba at a Crossroads: Politics. University Press of Florida, Gainesville, pp. 190–219.

Pérez-Stable, M., 1999. The Cuban Revolution. Origins, Course and Legacy. Oxford University Press, Oxford.

Portell-Vila, H., 1986. Nueva historia de la republica de Cuba, pp. 421–422. Moderna Poesía, Miami.

Romero, C.A., 2011. La política, el comercio y la economía entre Cuba y Venezuela. In: Cuba in Transition, vol. 21, pp. 423–434. Retrieved 16.12.11, from: http://www.ascecuba.org/publications/proceedings/volume21/pdfs/romero.pdf.

Sanchez, Y., 2011. Havana real. One Woman Fights to Tell the Truth about Cuba Today (M.J. Porter, Trans.). Melville House, Brooklyn.

Scarpaci, J.L., 1995. The emerging food and paladar market in Havana. In: Cuba in Transition, vol. 5, pp. 74–84. Retrieved 16.12.11, from: http://www.ascecuba.org/publications/proceedings/volume5/pdfs/FILE09.pdf.

Suarez, A., 1967. Cuba: Castroism and Communism, 1959–1966. MIT Press, Cambridge.

Werlau, M.C., 2011. Cuba's business of humanitarianism: the medical mission in Haiti. In: Cuba in Transition, vol. 21, pp. 196–197. Retrieved 27.01.12, from: http://www.ascecuba.org/publications/proceedings/volume21/pdfs/werlau.pdf.

Werlau, M.C., 2010. Cuba-Venezuela health diplomacy: the politics of humanitarianism. In: Cuba in Transition, vol. 20, pp. 143–160. Retrieved 12.03.12, from: http://www.ascecuba.org/publications/proceedings/volume20/pdfs/werlau.pdf.

Yanes, H., 2005. Algunos de los beneficios de la alianza cubano-venezolana para el régimen castrista. In: Cuba in Transition, vol. 15, pp. 470–476. Retrieved 24.02.12, from: http://www.ascecuba.org/publications/proceedings/volume15/pdfs/yanes.pdf.

Relevant Websites

Aagesen, D., Paluch, J., 2000. International tourism trends in Cuba. The Geographical Bulletin 43 (1), 47–58. Retrieved 07.01.12, from: http://www.gammathetaupsilon.org/the-geographical-bulletin/2000s/volume43-1/article5.pdf.

Acosta, D., 12 October 2011. Women's Department Draws Attention to Inequality. Retrieved 21.11.11, from: http://ipsnews.net/print.asp?idnews=105437.

Acosta, N., 9 November 2011. Cuban Official Says Corruption Crackdown to Go on. Retrieved 27.11.11, from: http://www.reuters.com/article/2011/11/09/us-cuba-corruption-idUSTRE7A88F820111109.

Amnesty International 2011 Report. Retrieved 18.03.12, from: http://www.amnesty.org/en/news-and-updates/cuba-urged-revoke-repressive-laws-and-release-prisoners-conscience-2010-03-17.

Chacon, R., 01 January 2001. Trial Turns Spotlight on US-Cuban Espionage. Boston Globe. Retrieved 15.01.12, from: http://www2.fiu.edu/~fcf/trialglobe101.html.

Chavez, J.C., Tamayo, J., 13 December 2011. Castro to Free More Prisoners, but Not Alan Gross. The Miami Herald. Retrieved 15.01.12, from: http://www.miamiherald.com/2011/12/23/2560001/castro-to-free-more-prisoners.html.

Grogg, P., 17 November 2009. Women in Cuba Choosing to Have Fewer Children. Newsroom, New Jersey. Retrieved 26.03.12, from: http://www.newjerseynewsroom.com/international/women-in-cuba-choosing-to-have-fewer-children.

Human Rights Watch, 1999. Impediments to Human Rights in Cuban Law. Retrieved 15.03.12, from: http://www.hrw.org/reports/1999/cuba/Cuba996-03.htm.

Leon, A., 30 March 2012. Protester at Pope's Mass in Cuba Identified. The Miami Herald. Retrieved 07.04.12, Alexandra Leon The Miami Herald. tp://www.miamiherald.com/2012/03/30/2722694/protestor-at-popes-mass-in-cuba.html.

Maury del Toro, J.A., 21 March 2010. [Web Log Message]. Retrieved 12.03.12, from: http://cubalagrannacion.wordpress.com/2010/09/07/el-internacionalismo-proletario-de-la-revolucion-cubana/.

Morales-Domínguez, C.E., 21 April 2010. Corruption: The True Counterrevolution?. Retrieved 23.11.11, from: http://afrocubaweb.com/estebanmorales.htm.

Partido Comunista de Cuba. Buro Político. Retrieved 03.02.12, from: http://www.pcc.cu/eo_buro_politico.php.

Partido Comunista de Cuba. Integrantes del Secretariado del Comité Central. Retrieved 03.02.12, from: http://www.pcc.cu/eo_secretariado.php.

Tamayo, J.O., 30 March 2012. Cuba diverts Dissident's Phone Numbers in Pope Crackdown. The Miami Herald. Retrieved 06.04.12, from: http://www.miamiherald.com/2012/03/30/2723658/cubas-interior-ministry-left-fingerprint.html.

World Report, 2012. Cuba, Human Rights Watch. Retrieved 18.03.12, from: http://www.hrw.org/world-report-2012/world-report-2012-cuba.

Cultural Capital and Education

Susan A Dumais, Lehman College, City University of New York, Bronx, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides the definition(s) of cultural capital and its theoretical context, with a particular focus on the relationship between cultural capital, schooling, and educational outcomes. An overview of Bourdieu's theory of social and cultural reproduction is given, focusing on the role of cultural capital. Empirical research on various operationalizations of cultural capital, both quantitative and qualitative, and various educational outcomes, are reviewed; research from multiple nations is represented. Past and current controversies regarding the cultural capital concept in educational research are discussed, as they are important directions for future research.

Introduction

Cultural capital has been a widely used concept in sociological and educational research in the last several decades. Originated by Pierre Bourdieu to describe educational inequality in France, the concept has now been explored around the world and has taken on a number of different operationalizations in qualitative and quantitative research, from all levels of schooling ranging from elementary school to postsecondary education. Findings regarding the effects of cultural capital are mixed, with some supporting the idea that cultural capital increases educational rewards, others finding a negative or no effect, and others finding different effects based on the characteristics of the group being studied (with variations by gender, race, and social class). Although a popular concept, some critics debate its usefulness or the approaches that researchers use to study it. Nevertheless, the concept is a fruitful one and new directions in using it are developing within sociology and educational research.

Definitions of Cultural Capital

Bourdieu's Definitions

Pierre Bourdieu developed the concept of cultural capital in the late 1960s when studying the French educational system and noticing that middle-class students tended to persist through school, while working-class students had higher drop out rates. Bourdieu argued that although education was supposed to serve as a vehicle for social mobility where students could work to their natural aptitudes, the educational system actually reproduced the existing social class structure; one important mechanism through which this reproduction occurred was cultural capital. In "Cultural Reproduction and Social Reproduction," Bourdieu described cultural capital as "linguistic and cultural competence and that relationship of familiarity with culture which can only be produced by family upbringing when it transmits the dominant culture" (1973: p. 80). The cultural capital recognized and rewarded by teachers and the school system was the kind possessed by the middle and upper classes.

Bourdieu distinguished between three forms of cultural capital: embodied, objectified, and institutionalized. Embodied cultural capital referred to one's tastes, ways of speaking and carrying oneself, and general knowledge of the culture valued by the dominant classes. Objectified cultural capital referred to objects that required embodied cultural capital in order to be appreciated, such as works of art or literature. Institutionalized cultural capital was legitimated cultural capital, often in the form of educational credentials. These forms of cultural capital could be converted to other forms, so that possessing embodied cultural capital enabled an individual to acquire institutionalized cultural capital.

The concept of cultural capital was part of Bourdieu's larger theory of practice, which also included the concepts of habitus (one's worldview or disposition) and field (the context in which social action occurred). Focusing on the educational system, Bourdieu argued that schools operated based on the culture of the dominant class, and expected all students to be knowledgeable of it. While this was true for the middle and upper classes, who were exposed to this culture from birth and throughout childhood in the family setting, the working class and poor students lacked this cultural capital and were at a disadvantage in school. They might not be viewed as favorable by teachers, they might not understand materials or assignments that were based on the dominant culture, and they might opt out of education themselves when seeing the mismatch between their cultural resources and those demanded by the school. Thus, middle and upper class students were more likely to persist in school, attain credentials, and become part of the middle or upper class themselves upon adulthood, reinforcing stratification in society.

However, Bourdieu argued, complete reproduction of the stratification system did not occur. Bourdieu stated that this was a way to keep people believing that success was based on one's ability and effort. In other words, a few students from the lower classes were selected and promoted by teachers, enabling them to succeed, so that others believed that they too had the opportunity to move up the social stratification ladder.

Bourdieu stated that it was possible for members of the lower classes to acquire cultural capital, but it would stand out as different from the cultural capital of the middle classes. For those who were born into privileged families, cultural capital would become second nature, or as Bourdieu described it, embodied, and so it would be used unconsciously by

middle- and upper-class students as they navigated the educational process. On the other hand, individuals whose first acquisition of cultural capital was in the school setting would have more stilted and scholarly cultural capital, and be less able to use it with the ease with which the middle-class students did.

Expansions and Refinements of Bourdieu's Concept and Theory

Bourdieu's definitions of cultural capital varied throughout his works, and so subsequent researchers often had to choose one of his definitions to guide their research, or to come up with their own interpretation of the concept. Paul DiMaggio (1982), one of the earliest empirical cultural capital researchers in the United States, used one of Bourdieu's definitions of cultural capital: "instruments for the appropriation of symbolic wealth socially designated as worth of being sought and possessed." DiMaggio also expanded upon Bourdieu's theoretical argument by bringing in Max Weber's concept of status cultures. DiMaggio, unlike Bourdieu, argued that cultural capital not only served as a tool to reproduce the existing class structure; it also could serve as a vehicle for social mobility for those who were lower in the stratification system. He distinguished between Weber's notion of status group membership and said instead that in the United States, the focus should be on status culture participation. Those who participated in status cultures might be able to ascend the social ladder. Thus, in addition to Bourdieu's cultural reproduction model, DiMaggio introduced a cultural mobility model.

In the late 1980s, Michele Lamont and Annette Lareau published a piece assessing Bourdieu's concept of cultural capital and praised Bourdieu's bringing in of structure to understand the inequalities that occurred in school settings. Based on their readings of Bourdieu's and his colleagues' work, Lamont and Lareau offered a new definition of cultural capital as "institutionalized, i.e., widely shared, high status cultural signals (attitudes, preferences, formal knowledge, behaviors, goods and credentials) used for social and cultural exclusion" (Lamont and Lareau, 1988: p. 156). In a later piece, Kingston (2001) endorsed this definition of cultural capital and added the important qualifier that cultural capital is arbitrary. That is, qualities valued as cultural capital "are not meaningfully related to what schools and work organizations must centrally purport to accomplish, namely, mastery of academic material in schools and productive performance on the job" (90). Thus, the functional skills taught in school were distinct from the elegant turn of phrase or knowledge of obscure literature when teachers evaluated students; the former were the skills that schools were supposed to teach children, while the latter were the mechanisms through which teachers sorted students by cultural capital.

In the early 2000s, Annette Lareau and Elliot Weininger reviewed Bourdieu's writings on cultural capital and also thoroughly examined the 20 years of existing quantitative and qualitative research on cultural capital. Based on their review, they argued that cultural capital should "emphasize(s) micro-interactional processes whereby individuals' strategic use of knowledge, skills, and competence comes into contact with institutionalized standards of evaluation. These specialized skills are transmissible across generations, are subject to monopoly, and may yield advantages or 'profits'" (Lareau and Weininger, 2003: p. 569).

These additional definitions and refinements of the concept of cultural capital are by no means exhaustive, and rely heavily on the United States context. Nevertheless, they share certain traits that point to an underlying coherence in understanding how schools may use cultural differences to perpetuate inequality. The key components of cultural capital thus point to its unequal distribution in society, its ability to provide advantages or open doors for those who possess it and activate it, and its recognition by the dominant institutions of society. The specific ways that cultural capital is measured in empirical research are examined below.

Empirical Cultural Capital Research

Operationalizations of Cultural Capital

Quantitative

Bourdieu's empirical work on cultural capital incorporated many variables on taste, such as respondents' favorite artworks or musical pieces. While the preference for highbrow art was not the only way that Bourdieu measured cultural capital, this approach influenced many subsequent quantitative cultural capital studies, including DiMaggio's work, which in turn set the stage for using arts-related variables in cultural capital research in the United States and other countries.

Quantitative studies are often limited by what is available in the particular data set being used. Some of the more commonly used data sets include the National Education Longitudinal Study (NELS), the Panel Study of Income Dynamics, the Program for International Student Assessment (PISA), the National Longitudinal Survey of Youth, the Family Survey Dutch Population, and the Cultural Capital and Social Exclusion survey from the United Kingdom. Many researchers in different countries also collect their own data via surveys.

Below are some of the more common ways in which quantitative research has operationalized cultural capital:

- Arts lessons: Many quantitative studies (Aschaffenburg and Maas, 1997; DiMaggio, 1982; Dumais, 2002) include whether the student has taken lessons in art, dance, or other forms of the arts. Measures range from simple responses of yes or no, to focusing on the specific ages at which the students took lessons (Aschaffenburg and Maas, 1997). Distinctions are sometimes made between whether the lessons were taken in school or outside of school.
- Attendance at arts events: Researchers have also measured cultural capital using attendance at museums, concerts, or other performances (Roscigno and Ainsworth-Darnell, 1999). Measures range from indicating attendance or not to detailing the frequency of attendance, and again distinctions are sometimes made between school trips and trips outside of school. For some data sets, it is difficult to distinguish whether the two measures described thus far truly tap into highbrow arts or high culture venues, as the survey question may simply ask if the respondent has attended a show or performance, and the questions about music lessons rarely indicate the particular instrument or style of music.

- Reading behavior and books: Reading behavior is used in several quantitative studies. There is some debate over using reading behavior and the presence of books in the home as measures of cultural capital because reading is a skill that is taught in the school setting, rather than a feature of the dominant culture that is inaccessible to the lower classes. Other researchers have argued that reading behavior is a form of cultural capital, particularly when the works being read are of a certain type or when the student is reading books beyond those required for school.
- Educational resources in the home: Jay Teachman (1987) was one of the first American scholars to use educational resources in the home as a form of cultural capital. His measures included the presence of a place to study, reference books, a daily newspaper, and a dictionary/encyclopedia in the home. While Teachman did not explicitly describe his measures as objectified cultural capital, they fit Bourdieu's definition of this concept. Yuxiao Wu (2008) also used educational resources in the home, such as maps and a desk, in studying the relationship between cultural capital and education in China.
- Composite measures: Many researchers have combined different variables into indices of cultural capital. DiMaggio's (1982) operationalization of cultural capital included students' attitudes toward artistic activities and occupations; activities such as creating art, attending performances, reading, or visiting museums; and tests of information about art, literature, and music. More recently, Jun Xu and Gillian Hampden-Thompson (2012), using PISA data, created indices for cultural activities (museum/gallery visits, attendance at ballet or classical concerts, and live theater attendance), cultural communication (discussions of political or social issues, discussions of books, films, or television, and listening to classical music with parents), cultural possessions (works of art, literature, and poetry), and home educational resources (a dictionary, a place to study, a calculator, and textbooks). The authors thus brought together measures of embodied and objectified cultural capital and included parents' influence in the measures.
- Children's vs parents' cultural capital: Some quantitative studies have included cultural capital variables at both the child and parent level. For example, Aschaffenburg and Maas (1997) included measures of cultural training for the respondents and measures of the parents' taste for classical music, as well as whether the parent escorted the child to different cultural events or encouraged reading.

Qualitative

Authors of qualitative studies tend to operationalize cultural capital similarly to the definition given earlier by Lareau and Weininger (2003). Unlike the quantitative studies summarized previously, the qualitative measures of cultural capital rely heavily on parents' behaviors and attributes. For example, Diane Reay (2004) argued that cultural capital was comprised of parents' feelings of confidence and entitlement when interacting with schools, and other qualitative researchers have also analyzed the degree to which parents advocate on behalf of their children in the school and other institutional settings. Possession of information and the ability to use it in a fruitful way has also been described as cultural capital, such as in Patricia McDonough's (1997) study of how parents help their children navigate the college selection and application process. At the student level, some qualitative studies note that cultural capital is demonstrated by the ease with which students interact with strangers or authority figures, and others also mention students' language use as an indicator of cultural capital.

Operationalizations of Educational Outcomes

Quantitative

Much of the quantitative research on cultural capital sets out to accomplish two things: to show that family socioeconomic background is associated with levels of cultural capital, and to show that cultural capital is associated with educational outcomes; often, researchers expect that cultural capital will reduce the direct effect that socioeconomic background has on educational outcomes. Common quantitative educational outcomes are described below:

- Test scores: A common outcome in quantitative studies is students' performance on standardized tests. These are usually tests developed for a specific data set, such as reading or math achievement in NELS, or performance on tests in PISA. In longitudinal studies, prior test scores are often included as controls, so that the outcome being studied reflects gains in achievement.
- Grades: Grades are sometimes preferred over test scores because they include more of a subjective component by teachers, which Bourdieu implied would play a role in social reproduction. Frequently, data sets only offer the students' self-reported grades, but the availability of transcript data in NELS and other data sets allows for an investigation of officially recorded grades.
- Teachers' evaluations of behavior: Some recent research by Susan Dumais (2006) has examined teachers' ratings of students' behavior, including working to one's best ability and one's interpersonal skills, as they are affected by students' cultural capital possession.
- Educational transitions: A number of studies focus on the transitions from lower-level education to higher-level education, particularly in countries where high levels of education are not the norm. In the United States and European countries, many quantitative studies use the transition from high school to college as a dependent variable.
- Degree attainment: The highest degree attained by the respondent is also a measure commonly found in quantitative cultural capital research.

Qualitative

Many of the outcomes detailed in qualitative studies can be categorized as educational advantages. These advantages include a more customized educational experience, placement into gifted programs, help in making the successful transition to the college of one's choice, and a team of parents and other adult authority figures who may advocate on a student's behalf. At the student level, Annette Lareau (2003) focuses on a student's growing sense of entitlement as an outcome of

exposure to cultural capital. This entitlement then enables the student to pursue a variety of opportunities without fear of failure, while students without cultural capital develop a sense of constraint about their lives and future opportunities.

Research Findings

While Bourdieu's research on cultural capital focused on the specific context of the French educational system, researchers throughout the world have applied the concept to their own countries' educational programs, implying that cultural capital is a universal, rather than France-specific, phenomenon. Findings vary by nation and often even within nations, driving more research in order to clarify whether and how cultural capital affects educational success. Many studies also set out to determine the degree to which cultural capital serves to reproduce the existing social structure, or to provide an opportunity for mobility. A few of the major studies from various regions are described below; note that space does not permit all countries to be represented (additional studies not reported here include research on Greece, Kenya, Norway, Russia, and Slovenia, among others).

The Netherlands

Research from the Netherlands on cultural capital has been ongoing since the 1980s and has contributed to the understanding of educational processes in important ways. De Graaf et al. (2000) found that parents' reading behavior during the respondents' childhoods, rather than high arts participation, affected the educational attainment of the adult respondents. They also found that children from less educated families were especially affected by parents' reading behavior, supporting DiMaggio's mobility hypothesis. Kraaykamp and van Eijck (2010) showed that parents' institutionalized cultural capital (years of education) and cultural behavior were positively associated with children's educational attainment, but that the association grew smaller with more recent generations, indicating that the intergenerational transmission of cultural capital may be weakening over time. Finally, while much cultural capital research focuses on family processes, Nagel et al. (2010) examined cultural capital as a required component of an educational program in the Netherlands. They found that there were no long-term effects of participating in the program, either on subsequent cultural participation, or on attitudes toward culture. These findings implied that school interventions may not be effective in enhancing the cultural capital of students beyond what they had already acquired at home.

The United States

Within the United States, DiMaggio (1982) was one of the first researchers to study cultural capital and education. He found both educational and gender differences in the effects of cultural capital on grades for high school students, with males from families with low education benefitting and females from families with high education benefitting. Support thus appeared for both the reproduction and mobility models. In a 1997 piece by Karen Aschaffenburg and Ineke Maas, the authors found that early cultural classes (before age 12) and parents' cultural participation were strongly associated with entering high school, finishing college, and entering college; parents' cultural participation and children's later cultural lessons were associated with finishing college. Additionally, they found some support for the reproduction model, as the transition to college was significantly affected by the interaction of parents' education and child's cultural participation.

Other researchers in the United States have studied the degree to which race and/or gender interact with cultural capital. Susan Dumais (2002) focused on eighth grade students and their self-reported grades, noting differences between males and females. While cultural capital did not have a strong effect overall, it did have a stronger association with grades for females than for males. Roscigno and Ainsworth-Darnell (1999) used data from NELS and examined grade point averages and composite reading and math achievement test scores. They found that not only did the reproduction model work for socioeconomic status, with higher SES students getting greater returns for cultural trips and educational resources, but there was also a reproduction of the racial structure, with white students getting greater returns for their cultural capital than African-American students.

Farkas et al. (1990) were some of the earliest American researchers to employ a mixed quantitative–qualitative approach to understanding cultural capital and schooling. Teachers were given surveys and asked to evaluate students' work habits, disruptive behavior, and appearance and dress. Even after controlling for cognitive performance, students' work habits were an important predictor of course grades, indicating that noncognitive factors affected teachers' perceptions of students.

Additional qualitative research has focused less on academic outcomes, and more on the ways in which cultural capital functions in different settings. Annette Lareau (2003) focused on parenting practices and how they resulted in a sense of entitlement in children. She updated Bourdieu's concepts of cultural capital and habitus to the American context by describing them as a cultural approach to parenting by the middle class, which she deemed as 'concerted cultivation.'

Prudence Carter (2003) coined the term black cultural capital to emphasize that in different contexts, different styles of behavior and dispositions were more advantageous than traditional cultural capital (as defined by Bourdieu) within the black community. Black cultural capital was employed to show authenticity to their peers, while dominant cultural capital was used for success in educational or employment settings.

The research findings for the United States are mixed, but several quantitative studies point to cultural capital as an agent of social reproduction rather than a tool for mobility, and the qualitative studies point to the microlevel interactions that contribute to the reproduction of inequality.

Great Britain

Diane Reay (2004) noted that in England, parental involvement in schooling has become a component of cultural capital. Middle-class mothers in her qualitative study were very confident in debating with and negotiating with teachers, while working-class parents were more apologetic and unsure of themselves, feeling that they did not successfully convey their points of view. Additionally, Reay found that working-class children placed in gifted education often dropped out,

claiming that they did not fit in. In these ways, privilege continued for middle-class parents and their children, while working-class families faced blame because they were less involved or their children did not persist in gifted education. Alice Sullivan (2001) studied four British schools and found that cultural capital, specifically reading activities and watching cultural television programs, was associated with General Certificate of Postsecondary Education (GCSE) attainment. However, these effects disappeared once scores on vocabulary and cultural knowledge tests were included in the model. The findings led Sullivan to conclude that while cultural capital was a useful concept, Bourdieu's overall theory of social reproduction through the schools did not appear to apply to the British context. These two studies (Reay and Sullivan) illustrate the very different findings that can arise from qualitative vs quantitative research on cultural capital.

Asian Countries

Yuxiao Wu (2008) found that cultural capital significantly affected enrollment in junior high school throughout three time periods in China. For senior high school enrollment, cultural capital was not significant during the Cultural Revolution period, and for college enrollment, cultural capital was only significant during the post-Maoist era. The findings thus indicated that the effects of cultural capital were vulnerable to state intervention. Using data from the 2000 PISA, Byun et al. (2012) found an association between socioeconomic status and parental objectified cultural capital, as well as socioeconomic status and children's embodied cultural capital, in South Korea. Only parents' cultural capital, however, was associated with academic achievement. Like Wu (2008), Byun showed that the state's educational system could affect how cultural capital operated.

Yoko Yamamoto and Mary Brinton (2010) studied embodied and objectified cultural capital in Japan and found that fine arts activities were associated both with grades in ninth grade and the quality of high school attended, while objectified cultural capital was associated only with the quality of high school. They argued that fine arts activities built skills that aided in achievement, while objectified cultural capital sent signals to teachers that the students were worthy of being recommended to a more prestigious high school.

International Comparisons

Jun Xu and Gillian Hampden-Thompson (2012) studied 22 countries using PISA data. Despite the increasing educational attainment in different countries and the subsequent opening up of class systems, they found that their measures of cultural capital were moderately correlated with one another as well as with parental education, regardless of the country's regime. Additionally, educational resources at home, cultural communication with parents, cultural participation, and books and works of art in the home were all associated with higher scores on reading-assessment scores. They also found support for the cultural reproduction model in liberal regimes (such as the United States and the United Kingdom), and support for the cultural mobility model in corporatist and Mediterranean regimes. The authors argued that cultural capital remains a salient concept that should continue to be studied across multiple contexts.

Summary of Main Research Findings

Overall, the findings are mixed, although in most of the studies described, cultural capital had a significant association with the educational outcome being studied. However, not all studies found that this translated into the reproduction of inequality through the educational system. Many of the quantitative studies showed a direct effect, or an indirect effect through other variables such as test scores, for cultural capital. The qualitative studies depicted differences in orientations toward schools that resulted in differential rewards for parents and children from different social classes.

Criticisms

There have been several criticisms of the concept of cultural capital, or of its use in research, over the past few decades. One criticism is researchers' use of Paul DiMaggio's cultural mobility model as an explanation for the relationship between cultural capital and school success. As Paul Kingston (2001) argued, Bourdieu's theory of reproduction addressed how the upper classes maintained their position and passed their privileges on to the next generation. To use the mobility model along with the concept of cultural capital could be seen as antithetical to Bourdieu's point about cultural capital and class reproduction. Related to this point is the criticism that Bourdieu's work should not be applied to societies outside of France (and perhaps not outside of the time period during which Bourdieu conducted his study). Kingston (2001) noted that in the United States, there is not an exclusive, dominant cultural group composed of individuals who regularly attend the opera or other highbrow events. Indeed, the rates of cultural participation in highbrow activities are extremely low. How, then, can participation in high culture be argued to represent an exclusive class-based practice that will be used to sort students as they progress through school?

The expansion of the types of variables used to measure cultural capital has also been criticized. As discussed earlier, operationalizations of cultural capital in quantitative research range from arts participation to reading behavior to the presence of a calculator in the home. Not all of these items fit the definitions of cultural capital given earlier by Lamont and Lareau (1988) or Kingston (2001), who respectively emphasized that cultural capital was exclusive and arbitrary. Kingston (2001) argued that even qualitative research has not successfully shown that middle-class parents engaged in exclusive practices that were valued because of their connection to elites. Instead, this research has shown that middle-class parents engaged in practices with their children that were educationally stimulating and thus could be expected to positively affect students' school outcomes.

Annette Lareau and Elliot Weininger (2003) critiqued quantitative cultural capital researchers' tendency to treat cultural capital and ability as conceptually distinct. Reviewing Bourdieu's writing on cultural capital, the authors found no evidence that Bourdieu considered ability to be a separate concept from cultural capital; indeed, his writings indicated that he considered the two to be conjoined. Lareau and Weininger (2003) attributed the frequent separation of cultural capital and ability in quantitative research to a reliance on the status attainment tradition.

Quantitative cultural capital research has been criticized for neglecting to specify the mechanisms by which cultural capital translates into educational outcomes. The lack of mechanisms is often due to the variables available in the data sets analyzed. Furthermore, while Bourdieu's argument implies that teachers play a role in the evaluation and promotion process, many researchers resort to using achievement test scores rather than teacher-based outcome variables. The argument used to justify the use of test scores – that cultural capital enhances students' skills and abilities – relates to the point mentioned above that researchers have tended to consider cultural capital a separate concept from ability.

There have also been some criticisms about the data and methods used to study cultural capital and education. As mentioned earlier, some of the measures in the data sets ask only if the student has ever taken a music lesson, or if the student has attended a concert in the past year. Bourdieu's concept of cultural capital entailed a lifelong exposure to distinctive cultural styles, and so using a single measure of arts attendance may not capture the concept in an adequate way. Additionally, many data sets, such as the Family Survey Dutch Population and the Survey of Public Participation in the Arts, ask retrospective questions about the respondents' childhoods and the behaviors in which their parents engaged during that time. Respondents might not have a clear memory of their parents' cultural participation, or even their own, when it was 15 or 20 years prior. Other data sets are problematic because they are cross-sectional rather than longitudinal, so that measures of cultural capital and of educational outcomes are collected at the same time. It is then very difficult to establish causality. Even longitudinal data may fail to capture the multiple interactions between teachers and students over the course of the school year that result in the teacher's evaluation of the student based on his or her perceived cultural capital.

Future Directions

As research on cultural capital continues, several innovations are occurring that may address the criticisms mentioned in the previous section. For example, S. Michael Gaddis (2013) focuses specifically on the mechanisms through which cultural capital may translate into educational success. His quantitative analyses show that cultural capital affects several psychological variables (which he calls habitus), and that they in turn affect students' grade point averages. Other quantitative studies have also incorporated the concept of habitus, often measuring it as parents' expectations for their children, or the student's expectation for his or her own attainment. Researchers should continue to delve into Bourdieu's work to ascertain what kinds of measures could best explain the link between cultural capital and educational outcomes.

Another important direction, most recently taken by Kraaykamp and van Eijck (2010), is to consider the different components of cultural capital, separately analyzing the effects of the embodied, objectified, and institutionalized states. Kraaykamp and van Eijck's work demonstrates that the intergenerational transmission of cultural capital occurs differently depending on which state is being examined. Yamamoto and Brinton's (2010) work, described earlier, distinguished between two states – embodied and objectified – and found that the mechanisms through which each affected educational outcomes varied, with embodied cultural capital resulting in more skills and objectified cultural capital resulting in better evaluations from teachers. Future research should continue to consider the multiple states of cultural capital and their differential effects on success in school.

The transmission of cultural capital from one generation to another has been a popular research subject, and some scholars are pushing this area forward by analyzing multiple generations, rather than the more common parent–child framework. For example, Josipa Roksa and Daniel Potter (2011) consider the class origins as well as the current class positions of mothers in their analyses of parenting practices and children's academic achievement.

The release of new longitudinal data sets, such as the Early Childhood Longitudinal Study, allows researchers to study the cumulative effects of cultural capital over a period of several years, from the beginning of students' schooling experiences up through their transition to middle school. On the qualitative side, Annette Lareau (2003) conducted a follow-up study with children who were in third grade at the time of the initial study, showing how the cumulative effects of different parenting practices resulted in different pathways toward college or careers for the now young adults.

See also: Art and Socialisation; Class: Social; Culture, Sociology of; Families as Educational Settings; Inequality, Social; Social Inequality and Schooling; Social Inequality in Cultural Consumption Patterns; Social Stratification; Socioeconomic Status and Academic Achievement; Symbolic Boundaries.

Bibliography

Aschaffenburg, Karen, Maas, Ineke, 1997. Cultural and educational careers: the dynamics of social reproduction. American Sociological Review 62, 573–587.

Bourdieu, Pierre, 1973. Cultural reproduction and social reproduction. In: Brown, Richard (Ed.), Knowledge, Education, and Cultural Change: Papers in the Sociology of Education. Tavistock, London.

Byun, Soo-yong, Schofer, Evan, Kim, Kyung-keun, 2012. Revisiting the role of cultural capital in East Asian educational systems: the case of South Korea. Sociology of Education 85, 219–239.

Carter, Prudence L., 2003. "Black" cultural capital, status positioning, and schooling conflicts for low-income African-American youth. Social Problems 50, 136–155.

De Graaf, Nan Dirk, De Graaf, Paul M., Kraaykamp, Gerbert, 2000. Parental cultural capital and educational attainment in the Netherlands: a refinement of the cultural capital perspective. Sociology of Education 73, 92–111.

DiMaggio, Paul, 1982. Cultural capital and school success: the impact of status culture participation on the grades of U.S. high school students. American Sociological Review 47, 89–201.

Dumais, Susan A., 2002. Cultural capital, gender, and school success: the role of habitus. Sociology of Education 75, 44–68.

Dumais, Susan A., 2006. Early childhood cultural capital, parental habitus, and teachers' perceptions. Poetics 34, 83–107.

Farkas, George, Grobe, Robert P., Sheehan, Daniel, Shuan, Yuan, 1990. Cultural resources and school success: gender, ethnicity, and poverty groups within an urban school district. American Sociological Review 55, 127–142.

Gaddis, S. Michael, 2013. The influence of habitus in the relationship between cultural capital and educational achievement. Social Science Research 42, 1–13.

Kingston, Paul W., 2001. The unfulfilled promise of cultural capital theory. Sociology of Education 74, 88–99.

Kraaykamp, Gerbert, van Eijck, Koen, 2010. The intergenerational reproduction of cultural, capital: a threefold perspective. Social Forces 89, 209–232.

Lamont, Michele, Lareau, Annette, 1988. Cultural capital: allusions, gaps and glissandos in recent theoretical developments. Sociological Theory 6, 153–168.
Lareau, Annette, 2003. Unequal Childhoods: Class, Race and Family Life. University of California Press, Berkeley.
Lareau, Annette, Weininger, Elliot, 2003. Cultural capital in educational research: a critical assessment. Theory and Society 32, 567–606.
McDonough, Patricia M., 1997. Choosing Colleges: How Social Class and Schools Structure Opportunity. State University of New York Press, Albany.
Nagel, Ineke, Damen, Marie-Louise, Haanstra, Folkert, 2010. The arts course CKV1 and cultural participation in the Netherlands. Poetics 38, 265–385.
Reay, Diane, 2004. Education and cultural capital: the implications of changing trends in education policies. Cultural Trends 31, 73–86.
Roksa, Josipa, Potter, Daniel Potter, 2011. Parenting and academic achievement: intergenerational transmission of educational advantage. Sociology of Education 84, 299–321.
Roscigno, Vincent J., Ainsworth-Darnell, James W., 1999. Race, cultural capital, and educational resources: persistent inequalities and achievement returns. Sociology of Education 72, 158–178.
Sullivan, Alice, 2001. Cultural capital and educational attainment. Sociology 35, 893–912.
Teachman, Jay D., 1987. Family background, educational resources, and educational, attainment. American Sociological Review 52, 548–557.
Wu, Yuxiao, 2008. Cultural capital, the state, and educational inequality in China, 1949–1996. Sociological Perspectives 51, 201–227.
Xu, Jun, Hampden-Thompson, Gillian, 2012. Cultural reproduction, cultural mobility, cultural resources, or trivial effect? A comparative approach to cultural capital and educational performance. Comparative Education Review 56, 1–27.
Yamamoto, Yoko, Brinton, Mary C., 2010. Cultural capital in East Asian educational systems: the case of Japan. Sociology of Education 83, 67–83.

Cultural Critique, Anthropological

Keith Hart, London School of Economics, London, UK; and University of Pretoria, Pretoria, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Critique is the exercise of judgment. Its roots lie in the Enlightenment's response to the printing revolution. The use of exotic ethnography to shock Westerners out of their complacency is well established. The 1980s saw a vigorous American school of cultural critics at home. They have European, feminist, and Marxist counterparts. Recent developments focus on globalization and cultural diversity. The world culture has become omnipresent and this has diluted anthropologists' attachment to their key concept. It would still pay anthropologists to revisit Rousseau and Kant.

Cultural Critique

Cultural critique, the use of anthropology to draw critical attention to institutions that readers take for granted, is as old as the origins of the modern discipline. It was a major purpose of Boas's students, especially Margaret Mead and Ruth Benedict. Its antecedents range from modernism to the critical theory of the Frankfurt school and the French surrealists between the wars, and the roots of critique lie in eighteenth century philosophy. Cultural critique is a less powerful force today in anthropology than it was in the 1980s, and the lines that separate it from social criticism are more blurred now. As a result, the present article takes a historical view of its subject rather than offering a contemporary survey.

What then is 'cultural critique'? It is to examine the foundations of culture by having recourse to judgment. Judgment in turn is the ability to form an opinion on the basis of careful consideration, beyond that to discern relations linking particulars to more general principles. Although not indifferent to fact and logic, judgment requires consideration of worth. A judge is respected for his or her wisdom and apparent objectivity, that is, for the ability to transcend mere opinion, even by giving expression to universal truth. A cognate expression is critic (which is, after all, derived from the Greek word for judgment). But criticism most often connotes for us the formation of opinions about works of art and the term 'cultural critique' suggests a direct link between anthropology and literary criticism. Indeed, critical anthropology for a time involved a shift in emphasis from life to text, reflecting a trend driven by poststructuralist discourse from the 1970s onward.

The world is currently going through a major transformation, driven by the new digital media, that is social, technological, and cultural in scope. It has fundamental consequences for the human condition and hence for anthropology. The best way to learn about these developments is to take an active part in them. There are analogies between the print revolution and today. For most of human history, information was hard to come by and had to be sought out. With printing, information became omnipresent. This was the origin of critique, since people had to learn how to be selective in what they read. Once choosing not to read something became socially acceptable, the way was open for the mass media. But the relationship between sender and receiver was still asymmetrical. The Internet and especially social media, commonly referred to as Web 2.0, have made available easy-to-use tools that provide a plethora of options for anyone to become engaged with the medium as a communicator in their own right. New social forms adequate to handling this unprecedented freedom of self-expression are at best incipient. They are, moreover, compromised by Web 2.0 being dominated by an outmoded bureaucratic capitalism whose command-and-control system and intellectual property regime continually provoke vigorous demands for more open access to information and for the democratization of its production, distribution, and consumption.

Early Twentieth-Century Ethnography

Ethnography was the dominant form of anthropological enquiry in the twentieth century. It literally means writing about peoples considered to constitute natural units, on the basis of extended participation in and observation of their societies. It arose in opposition to the method that preceded it, evolutionary history. This last sought to explain how the world came to be unified by Western imperialism by postulating a racial hierarchy in which 'our' cultural superiority (having developed a machine technology based on scientific reason) was attributed to biological difference. The First World War undermined the credibility of this claim of Western civilization to be founded on universal reason and opened up a space for anthropologists willing to assert the virtues of ways of life previously considered to be 'primitive' and inferior. Bronislaw Malinowski is rightly considered to be the pioneer of a new ethnographic approach stressing the integrity of contemporary exotic cultures. In *Argonauts of the Western Pacific* (Malinowski, 1961[1922]) he made clear his aim to challenge the assumption that Western economic norms were either universal or inherently superior. Subsequently, Edward Evans-Pritchard drew on liberal myths of Saxon resistance to Norman rule to dispute the identification of political order with the state, finding 'ordered anarchy' among the Nuer of the Sudan (Evans-Pritchard, 1940).

The academic norm of scientific ethnography was made necessary by anthropologists' drive toward legitimacy as a social science discipline. They eschewed fiction of all kinds. The basis of their truth claims was participant observation, the

fact that they had joined the people in some exotic location and shared their life as a means of studying them. Students were encouraged to read the classical monographs as scientific truth. The exercise of judgment was unnecessary, since professional vetting of the empirical object was sufficient guarantee of a monograph's validity. Within this framework of objectivity, ethnography was potentially critical of Western civilization's claim to universality. By stepping out of normal society, ethnographers pointed a sharp light on the assumptions of their readers, showing, for example, that witchcraft accusations made sense or that gifts may be instruments of domination. The general aim of anthropologists at this time seems to have been to discredit the racist assumptions underlying colonial empire. In America, this trend was associated with a position known as 'cultural relativism': every way of life, however barbarous, has a right to exist and it is not for outsiders to judge it as inferior for any reason.

The founder of modern American anthropology, Franz Boas, was himself committed to educating public opinion (Boas, 1929), but his students took this further by setting out to reach a general American audience with the surprising results of their own and others' ethnography. Americans think of their culture as natural, that is, inevitable; by showing them cultural variation elsewhere they are forced to ask if their own culture is malleable. This perhaps accounts for why Ruth Benedict's *Patterns of Culture* (1934), an impressive collation of cultural variety worldwide, was the anthropology bestseller of all time. Its subtitle is particularly instructive: an analysis of *our* social structure as related to *primitive civilizations* (italics added). In *Coming of Age in Samoa* (1928), her colleague Margaret Mead's topic was adolescence. Americans think of teenagers' turbulence as the natural result of sexual development; Samoan youth experienced adolescence more peacefully; so maybe Americans ought to think again about the reasons for teenage behavior. In a comparative work, *Male and Female: a Study of the Sexes in a Changing World* (Mead, 1949), Mead piled example on example to demonstrate that the sexual division of labor thought to be natural in Western societies was a special case. It is not hard to imagine how such a book might enter the campaign for female emancipation. Mead has been subjected to savage criticism of late for the inadequacy of her fieldwork and the exaggeration of her deconstructive cultural contrasts. These attacks have come precisely when the cultural critique she pioneered has sought to undermine the premises of scientific ethnography itself, as she never did.

Cultural Critique in America

More recently, cultural critique in anthropology has been associated with an American school of ethnographers led by George Marcus. With James Clifford he published *Writing Culture: the Poetics and Politics of Ethnography* (Clifford and Marcus, 1986), a book that has come to be seen as marking the postmodern shift in anthropology. At much the same time, Marcus published, with Michael Fischer, *Anthropology as Cultural Critique: an Experimental Moment in the Human Sciences* (Marcus and Fischer, 1999[1986]; see also Marcus, 1999). The growing prominence of cultural critique has been ascribed to a 'literary turn' associated with the work of Clifford Geertz since the early 1970s (Geertz, 1988). Clifford has carved out his own contribution to resolving "the crisis of representation" in a series of works, notably, *The Predicament of Culture* (1988). Paul Rabinow's *Reflections on Fieldwork in Morocco* (Rabinow, 1977) prefigured the critical reassessment of ethnographic method. The journal *Critique of Anthropology* grew out of collaboration between radical anthropologists in Britain, the Netherlands, and the United States, beginning in the 1970s.

The critique launched by *Writing Culture* had as its focus three aspects of the ethnographic enterprise: poetics, politics (both in the subtitle), and epistemology. By 'poetics, attention is drawn to the fact that ethnographers make up what they write. Geertz had been more or less explicit about his ambitions as a creative writer for over a decade before. Language, in such a perspective, is never simply descriptive, but persuasive also. If ethnography is literature, the strident assertion by positivists of its factual (nonfictional) basis must be questioned. The 'politics' of ethnography refers to a triadic relationship between anthropologists, their subjects, and their readers. The political economy of global inequality must be foregrounded in ethnographic descriptions and author/reader relations are more problematic than the original formula allows for. Who are the monographs for and what possibilities for collaboration exist beyond the stereotype of the ethnographer as a lone ranger in a pith helmet? Finally, the status of ethnography as a source of knowledge (epistemology), its methodology, and the configuration of knowledge as power, are called into question. Cultural critique in these terms examines the very intellectual foundations of modern anthropology.

What then have been the poetic and political purposes of the cultural critics and how effective was their epistemological critique? First, they were Americans who wished to engage more directly with American society. That is, they proposed to carry out ethnographic fieldwork in and write about mainstream activities and institutions. This required developing experimental methods, summed up as 'reflexivity, the need for the ethnographer to be continuously self-monitoring in the light of feedback from society. Their aim was similar to Mead's in seeking to disturb conventional assumptions about America's place in the world. Techniques of authorship proved hard to subvert in practice. Although the boundary between fact and fiction became blurred, few anthropologists have openly embraced the novel as a model. The radical rhetoric of the cultural critics, their attempts to develop new constituencies and forms of collaboration ran up against the continuing dominance of the universities as anthropology's institutional home. The ideal of the 'citizen ethnographer, boldly synthesizing fieldwork-based research, critical theory, and political commitment at home, remains a somewhat distant goal. If one can no longer read ethnographies as factual truth, how does one acquire the judgment to make sense of them, especially when cultural relativism has told us one should not judge others as the Victorian imperialists used to? Much postmodern writing adopts an ironic tone and perhaps this is because it straddles unresolved contradictions of this kind. Critics are by definition judgmental, yet convinced relativists have nowhere to stand.

The most controversial aspect of this program is thus its confrontation with the realism, naturalism, and objectification inherent in the positivist tradition of scientific ethnography. The traditional premise of ethnography, the study of other cultures for the sake of our own enlightenment, became more untenable under postcolonial circumstances. It fell to Marcus and his associates to make explicit the faults of a paradigm that had been unraveling for decades. The value of this American movement is much debated. Although the protagonists have sometimes tried to distance themselves from the 'postmodern' tag, preferring to speak of 'high modernism' or 'reflexive modernity, cultural critique has been seen by its positivist detractors as evidence of a postmodern collapse into fragmentation, confusion, and extreme relativism (Gellner, 1992). Others seem to agree that we have entered a postmodern age and that cultural critique is an appropriate response to it.

Cultural Critique in Europe

Cultural critique has not been absent from the other main national traditions of anthropology. Indeed, in France the public's reception of intellectuals has always been readier than was ever the case among the Anglo-Saxons. Moreover, French anthropologists have never embraced a narrow cultural relativism, preferring to keep alive the Enlightenment's legacy of an appeal to universal reason. The publication of Marcel Mauss's political writings (Mauss, 1997) shows how deep seated this public engagement was for him. As his principal successor, Claude Lévi-Strauss wrote a popular classic, *Tristes Tropiques*, and exposed to critical view the central institution of French culture, table manners (Lévi-Strauss, 1976, 1978). Among contemporary anthropologists, Marc Augé has successfully bridged the gap between the academy and popular culture, explaining the symbolic significance of Princess Diana's death and other contemporary events to readers of the serious newspapers.

Augé's contribution to cultural critique has been far-reaching. In *Non-places: Introduction to an anthropology of super-modernity* (1995), the term 'non-place' refers to spaces of transience, such as hotels, airports, motorways, and supermarkets, which are not significant enough to be considered 'places. Starting out as a West African ethnographer, he later turned his attention to Europe and finally to global developments. In the second phase, he applied methods taken from African fieldwork to daily life in Paris. There he drew attention to the absence of a sense of belonging in non-places, the increase in solitude brought about by communication technologies and the erasure of memory. Then, in *An Anthropology for Contemporaneous Worlds* (1998) and *The War of Dreams: exercises in ethno-fiction* (1999), Augé sought to theorize globalization as is lived globally, a reflection in part on his own travels. What is striking about this enterprise is that he exploits the collapse of the fact/fiction pair more fully than the American cultural critics, experimenting with writing techniques such as 'ethnonovels, and he succeeds in renewing anthropology as a whole.

Britain produced the only systematic critique of contemporary civilization by an anthropologist so far in Edmund Leach's prescient British Broadcasting Corporation lectures, *A Runaway World?* (Leach, 1968). There, he identified a world in movement, marked by the interconnectedness of people and things. This provoked the mood of optimism and fear that characterized the 1960s, when established structures seemed to be breaking down. The reality of change could not be understood through conventional cultural categories predicated on stable order. Moral categories based on habits of separation and division could only make the world's movement seem alien and frightening. An ethos of scientific detachment reinforced by binary ideas (right/wrong) lay at the core of society's malaise. Leach called for an intellectual practice based on movement and engagement, connection and dialectic. In short, he was calling for the reinsertion of ideas into life. No one in British social anthropology has come close to this achievement since.

Feminism and Marxism

In the last decades of the twentieth century, feminist anthropology was in the forefront of cultural critique. Indeed, the absence of women from *Writing Culture* provoked a justified response (Behar and Gordon, 1995). After all, the women's movement declared in the 1960s that "the personal is political" and launched a devastating critique of Western institutions on grounds of the invisibility, exclusion, and exploitation of women. These broader criticisms were readily applied to anthropology. An example of the sophistication reached by feminist and postfeminist discourse is Marilyn Strathern's *The Gender of the Gift* (Strathern, 1988), where the confounding of Western gender stereotypes in Melanesian cultures is taken as a point of departure for a much wider critique of such core conceptual pairs as individual/society and nature/culture. The 1980s were a decade of deconstruction, what Hegel called 'negative dialectic, in which the conventional categories of the modern synthesis became confused and discredited. Both inside and outside the academy, this task was performed to a disproportionate extent by women scholars.

Anthropology does not stand alone within an academic division of labor that has recently added Cultural Studies to its ranks. An enhanced focus on the contemporary forms of cultural production has revived interest in the Frankfurt school of Adorno and Benjamin. Their concern with exposing the uses of culture in masking the contradictions of capitalism remains a powerful one. Perhaps even more relevant to critical anthropology is the work of the West Indian Marxist, C.L.R. James (Grimshaw, 1992). Grimshaw and Hart's *Anthropology and the Crisis of the Intellectuals* (1993) placed anthropology's compromised relationship to academic bureaucracy within the general crisis facing modern intellectuals, following James (1993) in *American Civilization*. There he identified a growing conflict between vast anonymous bureaucracies and people's aspirations for an extension of democracy into all areas of their lives, an aspiration that could only be met indirectly through the movies and other forms of popular culture. James saw this as a crisis for modern society as a whole, most acutely revealed in America, and he addressed it through a critique ranging from Hollywood films and comic strips to psychoanalysis and the classical novels of Herman Melville.

For James, the struggle was for individual freedom within new and expanded conceptions of social life (democracy) against a fragmented and repressed subjectivity stifled by coercive bureaucracies (totalitarianism). The intellectuals were caught between the expansion of bureaucracy and the growing presence of people as a force in world society. Unable to recognize that people's lives mattered more than their own ideas, they oscillated between an introspective individualism (psychoanalysis) and service to the ruling powers, whether of the right (fascism) or left (Stalinism). As a result, the traditional role of the intellectual as an independent witness standing for truth had been compromised. The absorption of the bulk of intellectuals into bureaucracy as wage slaves and pensioners not only removed their independence, but also separated their specialized activities from social life.

In Keith Hart's (2005) essay, *The Hit Man's Dilemma* ("Don't take this personal, it's just business") is to be human or inhuman. It is a dilemma shared by kings, generals, presidents, and chief executive officers, when they contemplate the human cost of an action undertaken on behalf of some collective interest. The ability to devise ways of curbing the high-handed behavior of the powerful has been deeply undermined by a legal culture that grants business corporations the rights of living persons. The liberal revolutions against the old regime sought to guarantee citizens equal (and therefore impersonal) rights in society. It meant keeping a clear distinction between individual persons and impersonal institutions. This separation was intrinsic to the rise of modern capitalism. But business corporations sought to collapse the distinction between real and artificial persons in economic law. The impersonal society of the twentieth century flourished on this basis, in the process undermining the idea that ordinary people might exercise personal responsibility in the public sphere. Some intellectuals jumped onto the obvious corruption of liberal ideals to advocate a variety of antiliberal ideologies, drawing on the same confusion of people, ideas, and things that had become normal in law and even in ordinary language.

Ethnicity, Nation-States, and Globalization

John and Jean Comaroff's *Ethnicity Inc.* (2009) highlights a development that seems anomalous in the light of modern social theory, namely, the growing convergence of ethnicity and commerce. They conceive of this phenomenon as a dialectical product of the incorporation of identity and the commodification of culture. Their core case studies are the United States and South Africa, where they live and work. The link to neoliberalism is made explicit, so that the book offers a selective commentary on the subversion of modernity entailed in that movement. Their essay directs attention back to the dual foundation of the nation-state and asks how this proliferation of ethnic brands relates to the nationalist project (and to a parallel commercialization of religion, 'Divinity Inc.'). The Comaroffs thus bring cultural theory up to date by exposing to view a trend that has long historical roots, but is now omnipresent in ways that social science has been slow to recognize because of its own modernist assumptions.

In *Theory from the South* (Comaroff and Comaroff, 2011), they argue that the Global South is rarely seen as a source of theory and explanation for world historical events. Yet the crisis of Northern states encourages a view that they are evolving southward in both positive and problematic ways. They address a range of familiar themes: democracy, law, national borders, labor and capital, religion and the occult, liberalism, and multiculturalism. This view from the South renders key problems of our time at once strange and familiar, giving an ironic twist to the evolutionary pathways long assumed by social scientists.

The line between critique and social theory is increasingly hard to draw in anthropology. But, as this last example shows, much of contemporary critique in the discipline is targeted at globalization. Thomas Hylland Eriksen's (2007) short introduction to the subject is admirably lucid and balanced, but he too has a critical axe to grind. The causes and consequences of globalization are hotly disputed and the book presents a wide range of arguments. Eriksen argues, however, that variation is as characteristic of globalization conceived of as standardization. This dialectical perspective allows him to stress that just as we are being brought closer together by economic and technological developments, the need to differentiate ourselves from each other becomes accordingly that more imperative. This is the reason for an explosion of identity politics of the sort highlighted by the Comaroff and Comaroff (2009), as well as for this often to take overtly commercial or 'neoliberal' forms. Armed with this vision, Eriksen is able to promote some of anthropologists' classical virtues, especially the necessity for a bottom-up, comparative analysis. Distinguishing between the cultural, political, economic, and ecological aspects of globalization, he highlights the implications of globalization for people's everyday lives.

No anthropologist is better known for a cultural approach to the dialectic of nation-states and globalization than Arjun Appadurai. He has formulated concepts that support an anthropological approach to issues such as modernity, globalization, consumption, and public culture. Since these have not traditionally been part of anthropological research, anthropologists have for a long time hesitated to develop tools for dealing with them empirically and theoretically. Appadurai believes that the nation-state is in crisis and hence that current global processes of migration and communication will lead to the deterritorialization of identities in a world that will become culturally hybridized through the growth of diasporic public spheres and the global flow of images, finances, technologies, and ideologies. In *Modernity at Large* (Appadurai, 1996), he suggests one "think beyond the nation, by imagining a form of sovereignty that replaces territoriality with translocal forms. This led his approach to be identified with a 'cosmopolitan' anthropology that was indifferent to local identities.

Appadurai (2006) partially redressed this criticism with the publication of *Fear of Small Numbers* a decade later. Why, in this era of intense globalization, has there been a proliferation of violence, ethnic cleansing, and extreme forms of political violence against civilian populations? Here, he turns his attention to the complex dynamics fueling large-scale, culturally motivated violence, from the genocides that racked

Eastern Europe, Rwanda, and India in the early 1990s to the contemporary 'war on terror. He describes how the nation-state has grown ambivalent about minorities at the same time that minorities, because of global communication technologies and migration flows, increasingly see themselves as parts of powerful global majorities. By exacerbating the inequalities produced by globalization, the volatile relationship between majorities and minorities foments the desire to eradicate cultural difference.

Here is one possible explanation for a contradictory development that has undermined the culture concept and with it cultural critique in recent years. Disciplines outside anthropology turn increasingly to the concept of culture. This is particularly noticeable in certain branches of economics, especially those concerned with development (Rao and Walton, 2004). Joanna Breidenbach and Pál Nyíri (2009) have written engagingly about how culture has become a discursive feature of our world in *Seeing Culture Everywhere*. They conclude with a list of points that they hope will distance them from the more faddish aspects of culture mania. Culture should not be seen as belonging to bounded, homogeneous units. Nor is it unchanging. It is a mistake to imagine that whole cultures could be in conflict with each other or that mixing them up can be dangerous. Uncritical celebration of cultural difference is equally to be avoided. The tendency of neoliberalism to govern through overdrawn stereotypes of 'community' is seen as being a negative development. Cultural pigeonholing is out; cultural creativity at a highly differentiated level is in. One is finally told that ethnography is still a viable method, as long as it is linked to universal notions of 'human development.

Of course, when faced with the popularization of their central concept, anthropologists have started to turn against it. As long as ethnographers reached the remotest corners of the planet, their work could claim to be anthropology in the sense of providing the widest possible context of cross-cultural comparison. It is not easy to see what is anthropological about the indigenous critique of American societies, when the term 'ethnography' is also used by geographers, sociologists, and literary critics. The limitations of cultural critique could be said to lie in two weaknesses. The epistemological critique of ethnographic method has been somewhat introverted, an in-house discourse maintained by a minor academic profession, when the issues involved are potentially of universal significance. And the aspiration to grasp the human condition in general seems to have been abandoned in favor of competing for public attention in the United States. Both these deficiencies might be redressed if critical ethnographers examined their roots in the eighteenth-century philosophy that launched modern anthropology, as well as the notion of critique.

The Roots of Critique in the Enlightenment

This is not the place to explore how Kant developed his original conception of critique, not only most notably in his great *Critique of Pure Reason*, but also, for our purposes, in the *Critique of Judgment* (Cassirer, 1981). In seeking a metaphysical ground between reason and understanding, Kant emulated Copernicus's revolution, showing how, instead of the object revolving around the spectator, the spectator could be seen as revolving, while the object remained at rest. There are so many echoes of Kantian subjectivity in the anthropological moment of cultural critique. What is missing is the as piration to universality of method. This is doubly in excusable in that Kant invented the term 'anthropology' in its modern sense (Kant, 1978). Surely a discipline with the remit of offering knowledgeable guidance on human teleology should not content itself with a jester's role, *épater les bourgeois*.

No one would dispute Jean-Jacques Rousseau's standing as the founder of modern cultural critique. He showed that a critical perspective on society, a refusal to take as inevitable things as they are, requires one to devise new methods of studying and writing about a transitional present. The boundary between fact and fiction has to be blurred if one is to talk about the evolution of possible worlds out of the actual one we live in. In developing a revolutionary critique of politics, education, sexuality, and self-identity (*Le Contrat Social*, *Emile*, *La Nouvelle Héloïse*, and *Les Confessions*), Rousseau came up with a different genre of writing for each topic. Critical anthropologists today might emulate his example, were they not forced to justify themselves to the academic bureaucracy. The two discourses that launched Rousseau's career, however, are the principal sources for a renewal of critical anthropology, combining as they do a critique of corrupt civilization with an anthropology whose aim is to redress global inequality. In particular, his *Discourse on the Origins and Foundation of Inequality among Men* (Rousseau, 1984[1754]), which has inspired anthropologists from Morgan and Engels to Lévi-Strauss, deserves to be seen as the first great work of the modern discipline.

Cultural critique has been an invigorating stimulus to new thinking in anthropology. It is uncertain whether it can flourish while its practitioners remain securely locked up in their academic enclaves. It is time in any case for a return to the universal outlook of the eighteenth-century pioneers. The authors listed toward the end of this article have begun to show how that might be realized.

See also: Center–Periphery Relationships; Critical Theory: Contemporary; Critical Theory; Culture: Contemporary Views; Enlightenment: Impact on the Social Sciences; Enlightenment; Ethnicity: Anthropological Aspects; Ethnography; Feminism in Anthropology; Feminisms and Acculturation around the Globe; Gender, Feminism, and Sexuality in Archaeological Studies; Gender: Anthropological Aspects; Global South; Globalization, Anthropology of; Globalization: Geographical Aspects; Modernity: Anthropological Aspects; Postmodernism in Sociology; Postmodernism: Philosophical Aspects; Reflexivity in Anthropology.

Bibliography

Appadurai, A., 1996. Modernity at Large: Cultural Dimensions of Globalization. University of Minnesota Press, Minneapolis.

Appadurai, A., 2006. Fear of Small Numbers: An Essay on the Geography of Anger. Duke University Press, Durham.

Augé, M., 1995. Non-Places: Introduction to an Anthropology of Supermodernity. Verso, London.
Augé, M., 1998. An Anthropology for Contemporaneous Worlds. Stanford University Press, Palo Alto.
Augé, M., 1999. The War of Dreams: Exercises in Ethno-Fiction. Pluto, London.
Behar, R., Gordon, D. (Eds.), 1995. Women Writing Culture. University of California Press, Berkeley.
Benedict, R., 1934. Patterns of Culture: An Analysis of Our Social Structure as Related to Primitive Civilizations. Penguin Books, New York.
Boas, F., 1929. Anthropology and Modern Life. Allen and Unwin, London.
Breidenbach, J., Nyíri, P., 2009. Seeing Culture Everywhere: From Genocide to Consumer Habits. University of Washington Press, Seattle.
Cassirer, E., 1981. Kant's Life and Thought. Yale University Press, New Haven.
Clifford, J., 1988. The Predicament of Culture: Twentieth-Century Ethnography, Literature and Art. Harvard University Press, Cambridge.
Clifford, J., Marcus, G. (Eds.), 1986. Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley.
Comaroff, J., Comaroff, J., 2009. Ethnicity Inc. University of Chicago Press, Chicago.
Comaroff, J., Comaroff, J., 2011. Theory from the South: Or How Euro-America Is Evolving Toward Africa. Paradigm, Boulder.
Eriksen, T.H., 2007. Globalization: The Key Concepts. Berg, Oxford.
Evans-Pritchard, E., 1940. The Nuer. Clarendon Press, Oxford.
Geertz, C., 1988. Works and Lives: The Anthropologist as Author. Stanford University Press, Stanford.
Gellner, E., 1992. Postmodernism, Reason and Religion. Routledge, London.
Grimshaw, A. (Ed.), 1992. The C.L.R. James Reader. Blackwell, Oxford.
Grimshaw, A., Hart, K., 1993. Anthropology and the Crisis of the Intellectuals. Prickly Pear Press, Cambridge.
Hart, K., 2005. The Hit Man's Dilemma: Or Business, Personal and Impersonal. Prickly Paradigm, Chicago.
James, C.L.R., 1993. Popular Arts and Modern Society. In: Grimshaw, A., Hart, K. (Eds.), American Civilization. Blackwell, Oxford.
Kant, I., 1978. Anthropology from a Pragmatic Point of View. University of Southern Illinois Press, Carbondale.
Leach, E., 1968. A Runaway World? (The 1967 Reith Lectures). BBC, London.
Lévi-Strauss, C., 1976. Tristes Tropiques. Penguin, Harmondsworth.
Lévi-Strauss, C., 1978. The Origin of Table Manners. Jonathan Cape, London.
Malinowski, B., 1961[1922]. Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagos of Melanesian New Guinea. Dutton, New York.
Marcus, G. (Ed.), 1999. Critical Anthropology Now: Unexpected Contexts, Shifting Constituencies, Changing Agendas. School of American Research Press, Santa Fe.
Marcus, G., Fischer, M., 1999[1986]. Anthropology as Cultural Critique: An Experimental Moment in the Human Sciences. University of Chicago Press, Chicago.
Mauss, M., 1997. In: Fournier, M. (Ed.), Ecrits Politiques. Fayard, Paris.
Mead, M., 1928. Coming of Age in Samoa. Mentor Books, New York.
Mead, M., 1949. Male and Female: A Study of the Sexes in a Changing World. Dell, New York.
Rabinow, P., 1977. Reflections on Fieldwork in Morocco. University of California Press, Berkeley.
Rao, V., Walton, M. (Eds.), 2004. Culture and Public Action. Stanford University Press, Palo Alto.
Rousseau, J.-J., 1984[1754]. A Discourse on Inequality. Penguin, Harmondsworth.
Strathern, M., 1988. The Gender of the Gift: Problems with Women and Problems with Society in Melanesia. University of California Press, Berkeley.

Cultural Evolution: Overview

Alex Mesoudi, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by W.H. Durham, volume 5, pp. 3049–3052, © 2001, Elsevier Ltd.

Abstract

This article provides an overview of the field of cultural evolution, in which cultural change is studied as a Darwinian evolutionary process. Although a fringe interest just 20 years ago, cultural evolution is now a major area of interdisciplinary study. This article outlines the key tenets of cultural evolution theory, describes the main methods used, and highlights outstanding debates and issues that are currently being investigated in the field.

The study of cultural evolution is one of the fastest growing fields in the social and behavioral sciences. In the 1970s and 1980s, cultural evolution was studied by just a handful of scholars at the fringes of traditional disciplines such as anthropology, psychology, and biology. From such small beginnings has spawned a major area of contemporary research. As Figure 1 shows, the number of journal articles self-describing their subject matter as 'cultural evolution' has increased from less than 10 per year in the 1970s and 1980s to well over 100 per year at present. The study of cultural evolution is also unusual in its interdisciplinarity: cultural evolution researchers have an equal chance of having backgrounds in anthropology, archaeology, psychology, biology, economics, and linguistics, and inroads have recently been made into other fields such as history and neuroscience. This article provides an overview of the main theoretical tenets of cultural evolution theory, the primary methods used, the topics studied, and the current controversies within the field, supplementing more detailed articles elsewhere in this volume.

What Is Cultural Evolution?

The answer to this question hinges, naturally, on what one means by both 'culture' and 'evolution.' Although 'culture' has been defined in numerous ways across the social and behavioral sciences (Baldwin et al., 2006), cultural evolution researchers adopt a broad definition along the lines of the following:

> Culture is information that is acquired from other individuals via social transmission mechanisms such as imitation, teaching, or language. (Mesoudi, 2011; pp. 2–3)

This may include beliefs, knowledge, attitudes, skills, preferences, words and grammatical rules, and social norms; essentially, anything that is learned from another individual. Culture can be contrasted with information that is acquired genetically, in the form of parentally inherited sequences of DNA, as well as information that is generated *de novo* by a single individual in isolation, with no social transmission. There is no requirement in this definition that culture be restricted to humans; indeed, many cultural evolution researchers study the cultural capacities of nonhuman species in order to illuminate the evolutionary origins of human cultural evolution (*see* Social Learning and Culture in Nonhuman Organisms). It is also possible for culturally transmitted information to be stored exo-somatically, in external storage devices such as books or computer disks, as well as in brains.

'Evolution' has similarly been defined in various ways during the history of biology (*see* Evolution, History of), although Darwin's conceptualization in *The Origin of Species* (Darwin, 1859) is commonly accepted as the foundational definition within biology. Darwin's theory of evolution comprised three key principles, each of which is necessary for evolution to occur (Lewontin, 1970). There must be (1) variation within a population of entities; (2) competition for survival or reproduction of those entities, for example, due to finite resources; and (3) inheritance of the characteristics of those entities, such that offspring resemble parents more than expected by chance. Darwin demonstrated that these principles applied to populations of organisms. The same principles can be documented for human culture (Mesoudi et al., 2004): (1) there is abundant variation in cultural traits, such as the almost 7000 languages spoken, the 7.7 million technological patents issued, or the 10 000 religious beliefs held in the world today; (2) there is inevitable competition between those traits for memory space, as no one could plausibly speak 7000 languages, understand or use 7.7 million patents, or subscribe to just two different religions at the same time; and (3) cultural traits are reliably passed on to other individuals, as people acquire the languages they speak, the technology they use, the religious beliefs they hold, and so on from other people via social learning. This basic idea that human culture comprises an evolutionary process that fulfills Darwin's definition in *Origin* should be relatively uncontroversial and obvious, and indeed soon after Darwin wrote Origin, he himself noted that the same principles apply to cultural change, primarily linguistic change:

> The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel … The survival or preservation of certain favoured words in the struggle for existence is natural selection. (Darwin, 1871; pp. 90–91)

Two important points should be made here. First, this Darwinian notion of cultural evolution is quite different from

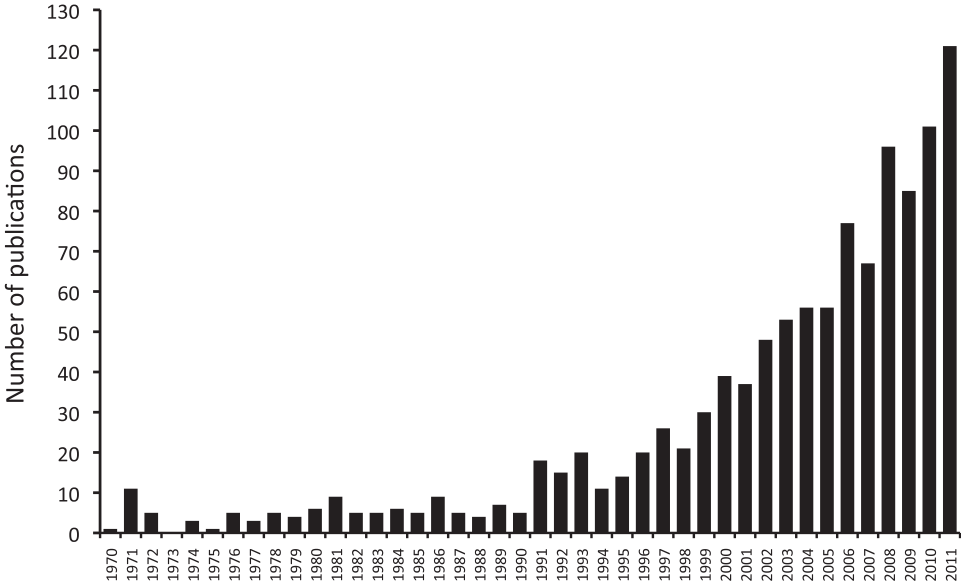

Figure 1 The number of publications with 'cultural evolution' listed as a keyword, 1970–2011. Source: Web of Science.

the theories of social or cultural evolution that were common in anthropology during the late nineteenth century, such as those of Tylor (1871) or Morgan (1877). These theories proposed that entire societies proceed through fixed stages of increasing complexity (e.g., from 'savagery' to 'barbarism' to 'civilization'). These progressive theories are Spencerian, not Darwinian (*see* Evolution, History of). The notion that evolution involves the inevitable progression of species along fixed stages of increasing complexity was long ago shown to be an invalid model of biological evolution (it describes the development of a single organism, not the evolution of a population), and contemporary theories of cultural evolution similarly make no assumptions about the progression of entire societies along developmental stages, focusing instead on how variation changes *within* populations, as did Darwin (Mesoudi, 2011).

The second important point is that while both genetic evolution and cultural evolution are Darwinian processes, they are not identical. As made clear by the founders of contemporary cultural evolution research (Boyd and Richerson, 1985; Cavalli-Sforza and Feldman, 1981), the underlying mechanisms of cultural evolution may be very different from the mechanisms of genetic evolution. For example, while genetic mutation is blind with respect to fitness, cultural innovations may be guided by purposeful agents; while genetic inheritance is particulate (i.e., it involves the high-fidelity transmission of discrete units of inheritance), cultural inheritance may be of much lower fidelity and involve the blending of multiple, continuously varying traits. None of these differences invalidate a Darwinian theory of evolution; there are still mutation and inheritance, even if they are not blind or particulate – they just mean that it must be modeled and understood differently. The failure to appreciate these differences was one of the pitfalls of *memetics* (Aunger, 2000a), which did, for example, require particulate inheritance of gene-like units of inheritance (memes), and which never properly took off as a scientific discipline, perhaps because of this insistence.

Cultural Microevolution

In the 1970s and 1980s, two pairs of researchers laid the foundations for the contemporary study of cultural evolution. Cavalli-Sforza and Feldman (1981) and Boyd and Richerson (1985) constructed mathematical models of cultural evolution using techniques drawn from population genetics (*see* Cultural Evolution: Theory and Models). Their models of cultural evolution were Darwinian, in that they assumed variation in a set of cultural traits and modeled various processes that act to change this variation over time as these traits are transmitted from individual to individual. Yet they also successfully incorporated the differences between genetic evolution and cultural evolution noted above, such as incorporating blending inheritance or guided (rather than blind) variation. Because these models are concerned with the processes that transform variation *within* populations, such models can be described as concerning *cultural microevolution*.

The aim of these models is to identify the population-level signatures of different microevolutionary processes. For example, models suggest that vertically transmitted traits, which are passed from parent to offspring just like genes, will change slower than horizontally transmitted traits, which are passed between unrelated members of the same generation (Cavalli-Sforza and Feldman, 1981), a prediction supported in subsequent empirical research (Hewlett and Cavalli-Sforza, 1986). In another example, Henrich (2001) modeled the diffusion of a novel technological innovation through a society via different transmission biases. Conformist transmission, in which individuals preferentially adopt the most common trait in the population, was shown to generate S-shaped diffusion curves, whereas individual trial-and-error learning generated r-shaped curves. Given that most innovations exhibit S-shaped dynamics (*see* Evolution: Diffusion of Innovations), Henrich (2001) concluded that conformist transmission is the most plausible process responsible for the spread of novel technological innovations.

Cultural Macroevolution

In parallel to these mathematical models of cultural microevolution, another group of researchers has used analytical methods borrowed from evolutionary biology to study *cultural macroevolution*, which refers to large-scale, long-term patterns of cultural change at or above the level of the society. These methods, originally designed to reconstruct the evolutionary history of biological species (*see* Comparative Method in Evolutionary Studies), have been used to reconstruct the evolutionary history of cultural traits such as languages (*see* Evolution and Language: Phylogenetic Analyses; see also Pagel, 2009), written manuscripts (Howe et al., 2001), technological artifacts such as stone tools (*see* Evolution, Technology of; see also Lipo et al., 2006), and behavioral practices such as cattle keeping and rules of wealth inheritance (Holden and Mace, 2003).

One aim of these studies is simply to reconstruct the most likely historical scenario that gave rise to contemporary variation in a trait, in terms of which language, technology, and practice emerged earlier and what came later. However, such methods can go beyond mere historical reconstruction and address specific hypotheses concerning why particular cultural traits successfully spread. For example, Holden and Mace (2003) showed that the adoption of cattle led to the switch from matrilineal (mother-to-daughter) to patrilineal (father-to-son) wealth inheritance, rather than vice versa (inheritance rules influencing the adoption of cattle). Gray and Atkinson (2003) showed that Indo-European languages most likely hitchhiked along with early Anatolian farmers rather than being spread later by nonfarming Kurgan horsemen. A longstanding issue in these analyses is the extent to which cultural macroevolution constitutes a branching process of phylogenetic divergence, thus resembling the evolution of animal species, versus one in which traits readily transfer between lineages, thus resembling the evolution of many bacteria and plant species, in which horizontal gene transfer causes lineages to blend (*see* Cultural Evolution: Phylogeny versus Reticulation).

Methods

As noted in section Cultural Microevolution, the earliest cultural evolution research primarily involved the construction of mathematical models (Boyd and Richerson, 1985; Cavalli-Sforza and Feldman, 1981), and this is still a vibrant area of research (e.g., Boyd and Richerson, 2005; Enquist et al., 2011; Lewis and Laland, 2012; Aoki et al., 2011; Bentley et al., 2004; Mesoudi and Lycett, 2009). Increasingly, the predictions of these models are being tested empirically both in the field and in the lab. The first ethnographic field study to explicitly test the predictions of cultural evolutionary models in a small-scale society was Hewlett and Cavalli-Sforza (1986), and ethnographic methods have continued to be used primarily to determine how cultural traits are transmitted through small-scale social networks (e.g., Aunger, 2000b; Reyes-Garcia et al., 2009; Henrich and Henrich, 2010). More recently, lab experiments have been used to examine, under more controlled conditions, how and when people learn the solutions to semirealistic artificial tasks from other members of small groups (Mesoudi and O'Brien, 2008; McElreath et al., 2005; Toelch et al., 2008; Efferson et al., 2008).

The phylogenetic methods discussed in section Cultural Macroevolution were first applied to cultural datasets by Mace and Pagel (1994), and have since been applied to datasets from linguistics (Pagel, 2009), anthropology (Lipo et al., 2006; Mace et al., 2005), and archaeology (O'Brien et al., 2001; Lycett, 2009). Again, these methods have been supplemented with ethnographic studies of cultural transmission, for example to measure the extent of cross-lineage blending (Tehrani and Collard, 2009), as well as lab experiments (Spencer et al., 2004). In the latter study, artificial manuscripts were passed along a known phylogeny, and then phylogenetic methods were used to reconstruct the simulated phylogeny, in order to test the validity of those methods.

As can be seen by this brief overview, there is an encouraging interplay in the cultural evolution literature between formal models, lab experiments, ethnographic field studies, and phylogenetic reconstruction of historical data. This is the sign of a healthy science, as theoretical models generate predictions that are then tested empirically, with the results of those empirical tests fed back into the models, which generates new predictions to be tested, and so on.

Current Topics and Issues

Domain-Specific versus Content-Free Biases

One ongoing debate concerns the relative importance of domain-specific 'content biases' versus content-independent 'context biases.' The former are emphasized by cognitive anthropologists (e.g., Sperber, 1996; Atran, 2001) who argue that the spread of cultural representations can be predicted from an understanding of the biologically evolved cognitive structure of the brain. This creates 'cognitive attractors' (Sperber, 1996) that bias cultural change toward intrinsically appealing ideas. For example, it is argued that the kinds of supernatural beliefs that are most likely to spread are those that are 'minimally counterintuitive' (Norenzayan et al., 2006), meaning that they violate some of our intuitions but not excessively so, such as ghosts that violate our folk physics by passing through walls but that obey our folk psychology by exhibiting memory and a desire for revenge. In contrast, other researchers (Henrich and Boyd, 2002; Richerson and Boyd, 2005) place more emphasis on content-free transmission biases, such as conformist transmission (copy whatever is most common in your society) or prestige bias (copy whatever the most prestigious member of your society does). This latter camp argues that many cultural adaptations, from canoes to computers, are too complex and novel to have domain-specific cognitive attractors dedicated to them, and are instead the result of chance innovations spreading via content-independent transmission biases (Boyd et al., 2011a). A similar debate is occurring in linguistics between Chomskian nativists who emphasize language-specific cognitive processes and cultural evolution researchers (e.g., Kirby et al., 2007) who emphasize general properties of cultural transmission (*see* Evolution and Language: Cultural Transmission).

This difference in emphasis has led to some spirited debate (Claidiere and Sperber, 2007; Henrich et al., 2008), but the two positions are not mutually exclusive. Some traits, such as food preferences, are likely to be highly influenced by biologically evolved cognitive biases due to the direct biological relevance of their content; others, such as recent and complex technologies, are unlikely to have dedicated cognitive biases and so result from a more content-free, populational process.

The Origin of Cumulative Cultural Evolution

Many species exhibit social transmission (Galef and Laland, 2005), and several exhibit regional traditions in tool use or foraging behavior that appear to have spread culturally (Whiten et al., 1999; Lycett et al., 2007). No other species, however, appears to have the capacity for cumulative cultural evolution (CCE), whereby modifications are built up over successive generations such that a single individual could never have invented it alone (think of quantum physics or a computer tablet). An outstanding question is why only humans have CCE. Some have argued that only humans possess cultural transmission mechanisms of high enough fidelity to support CCE, such as imitation (Lewis and Laland, 2012; Tennie and Over, 2012). Other evidence suggests that our species' capacity for collaboration and cooperation is key (Dean et al., 2012; see also next section). An alternative hypothesis is that there was no cognitive change associated with the emergence of CCE, and instead it was the result of larger population sizes and densities that made the loss of complex cultural traits less likely (Powell et al., 2009). Further comparative studies with nonhuman species, and archaeological study of the origin of complex cultural traits, will shed light on this.

The Evolution of Human Cooperation

Several cultural evolution researchers have drawn a link between our species' unusual capacity for CCE (see previous section) and our similarly unusual capacity to cooperate with nonkin (*see* Human Cooperation, Evolution of; Kin Selection). Whereas other primate species direct help only to close relatives (Jensen et al., 2007), humans frequently help nonkin in situations in which they are unlikely to get any personal return or increase their genetic fitness. Boyd, Richerson, and others (Boyd and Richerson, 1985; Gintis et al., 2003; Boyd et al., 2011b) have argued that a process of cultural group selection has occurred in humans, in which selection has acted on culturally homogeneous (but not necessarily genetically homogeneous) groups to favor group-beneficial traits. This remains a contentious hypothesis given the controversy surrounding group selection within biology (West et al., 2011), and one that is receiving increasing empirical attention (Lamba and Mace, 2011; Mathew and Boyd, 2011).

Culture-Driven Genetic Change

The field of gene–culture coevolution (*see* Gene-Culture Coevolution) has explored, mostly theoretically, the extent to which cultural evolution might interact with genetic evolution (Laland et al., 2010). The conclusion of many of these models is that cultural evolution can often drive genetic evolution to new equilibria that it would not have reached in the absence of culture. For example, the cultural invention of dairy farming led to the spread in some populations of lactose tolerance alleles, due to the fitness benefits of drinking milk in adulthood (Itan et al., 2009). Similarly, sickle cell genes likely spread because they confer resistance to malaria, a disease that spread when agricultural methods created ideal conditions for mosquitoes to breed (Durham, 1991). The extent to which cultural evolution has shaped the human genome, and whether this extends beyond physiological changes in response to the agricultural revolution, are empirical questions that will be addressed as gene-sequencing technology advances.

The Neural Basis of Cultural Transmission

Genetics was revolutionized by Watson and Crick's discovery of the molecular basis of genetic inheritance. An equivalent understanding of the molecular basis of cultural evolution would concern how information is stored neurally in brains, and how that neural information is transmitted from one brain to another. Current understanding of how the brain works is not yet able to provide the answer to this question. One potentially promising line of enquiry involves mirror neurons, cells that fire both when an individual observes a specific action (e.g., a particular hand grip) and when the individual observes another individual doing that same action (Rizzolatti and Craighero, 2004), potentially providing a cellular basis for imitation (Iacoboni et al., 1999). However, how this translates into the transmission of, say, tool-making skills remains unclear (Stout, 2011). As neuroimaging technology improves, a better understanding of the neurological basis of cultural transmission will hopefully emerge.

Conclusions

Cultural evolution provides a rich set of theories and methods for studying cultural change, encompassing the subject matter of several traditional academic disciplines. It provides rigorous, scientific tools for linking individual-level cultural transmission dynamics to population-level patterns of cultural change and variation, solving the 'micro–macro' problem that has beset the social sciences in the past. Previously, I have suggested that just as evolutionary theory served to synthesize the biological sciences in the 1930s and 1940s within an overarching 'evolutionary synthesis' (Huxley, 1942), so too evolutionary theory can synthesize the social sciences (Mesoudi, 2011; Mesoudi et al., 2006). 'Macroevolutionary' disciplines, such as archaeology, comparative sociology, macroeconomics, history, and historical linguistics, would be united with 'microevolutionary' disciplines, such as ethnography, psychology, microsociology, microeconomics, and sociolinguistics, with the historical and comparative patterns identified by the former explained in terms of the individual-level mechanisms studied by the latter. The recent spread of cultural evolutionary theory to disciplines such as history (see

Turchin, 2003), economics (*see* Evolutionary Economics), and archaeology suggests that such a synthesis is, in principle, possible.

Our species' ability to adapt culturally to new environments via cultural evolution, rather than genetically via natural selection as seen in other species, is likely responsible for our rapid colonization of virtually every terrestrial environment on the planet. Many of those environments have now been transformed beyond recognition by our ongoing cultural evolution, both for the better (e.g., the eradication of diseases by medical knowledge and technology) and for the worse (e.g., atmospheric pollution from the processing of fossil fuels). An interesting possibility is whether our increasing understanding of cultural evolution might help to encourage the former and discourage the latter.

See also: Comparative Method in Evolutionary Studies; Cultural Evolution: Phylogeny versus Reticulation; Cultural Evolution: Theory and Models; Evolution and Language: Overview; Evolution and Language: Phylogenetic Analyses; Evolution, History of; Evolution: Diffusion of Innovations; Evolutionary Economics; Evolutionary Epistemology; Gene–Culture Coevolution; Human Cooperation, Evolution of; Kin Selection; Social Learning and Culture in Nonhuman Organisms; Technology, Evolution of.

Bibliography

Aoki, K., Lehmann, L., Feldman, M.W., 2011. Rates of cultural change and patterns of cultural accumulation in stochastic models of social transmission. Theoretical Population Biology 79, 192–202.

Atran, S., 2001. The trouble with memes: inference versus imitation in cultural creation. Human Nature 12, 351–381.

Aunger, R. (Ed.), 2000a. Darwinizing Culture. Oxford University Press, Oxford.

Aunger, R., 2000b. The life history of culture learning in a face-to-face society. Ethos 28, 1–38.

Baldwin, J.R., Faulkner, S.L., Hecht, M.L., Lindsley, S.L., 2006. Redefining Culture: Perspectives across Disciplines. Lawrence Erlbaum, Mahwah, NJ.

Bentley, R.A., Hahn, M.W., Shennan, S.J., 2004. Random drift and culture change. Proceedings of the Royal Society B 271, 1443–1450.

Boyd, R., Richerson, P.J., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago, IL.

Boyd, R., Richerson, P.J., 2005. The Origin and Evolution of Cultures. Oxford University Press, Oxford.

Boyd, R., Richerson, P.J., Henrich, J., 2011a. The cultural niche: why social learning is essential for human adaptation. Proceedings of the National Academy of Sciences 108, 10918–10925.

Boyd, R., Richerson, P.J., Henrich, J., 2011b. Rapid cultural adaptation can facilitate the evolution of large-scale cooperation. Behavioral Ecology and Sociobiology 65, 431–444.

Cavalli-Sforza, L.L., Feldman, M.W., 1981. Cultural Transmission and Evolution. Princeton University Press, Princeton.

Claidiere, N., Sperber, D., 2007. The role of attraction in cultural evolution. Journal of Cognition and Culture 7, 89–111.

Darwin, C., 1859/1968. The Origin of Species. Penguin, London.

Darwin, C., 1871/2003. The Descent of Man. Gibson Square, London.

Dean, L.G., Kendal, R.L., Schapiro, S.J., Thierry, B., Laland, K.N., 2012. Identification of the social and cognitive processes underlying human cumulative culture. Science 335, 1114–1118.

Durham, W.H., 1991. Coevolution: Genes, Culture, and Human Diversity. Stanford University Press, Stanford.

Efferson, C., Lalive, R., Richerson, P.J., McElreath, R., Lubell, M., 2008. Conformists and mavericks: the empirics of frequency-dependent cultural transmission. Evolution and Human Behavior 29, 56–64.

Enquist, M., Ghirlanda, S., Eriksson, K., 2011. Modelling the evolution and diversity of cumulative culture. Philosophical Transactions of the Royal Society B 366, 412–423.

Galef, B.G., Laland, K.N., 2005. Social learning in animals: empirical studies and theoretical models. BioScience 55, 489–499.

Gintis, H., Bowles, S., Boyd, R., Fehr, E., 2003. Explaining altruistic behavior in humans. Evolution and Human Behavior 24, 153–172.

Gray, R.D., Atkinson, Q.D., 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426, 435–439.

Henrich, J., 2001. Cultural transmission and the diffusion of innovations. American Anthropologist 103, 992–1013.

Henrich, J., Boyd, R., 2002. On modeling cognition and culture: why cultural evolution does not require replication of representations. Journal of Cognition and Culture 2, 87–112.

Henrich, J., Boyd, R., Richerson, P.J., 2008. Five misunderstandings about cultural evolution. Human Nature 19, 119–137.

Henrich, J., Henrich, N., 2010. The evolution of cultural adaptations: Fijian food taboos protect against dangerous marine toxins. Proceedings of the Royal Society B 277, 3715–3724.

Hewlett, B., Cavalli-Sforza, L.L., 1986. Cultural transmission among Aka pygmies. American Anthropologist 88, 922–934.

Holden, C.J., Mace, R., 2003. Spread of cattle led to the loss of matrilineal descent in Africa: a coevolutionary analysis. Proceedings of the Royal Society B 270, 2425–2433.

Howe, C.J., Barbrook, A.C., Spencer, M., Robinson, P., Bordalejo, B., Mooney, L.R., 2001. Manuscript evolution. Trends in Genetics 17, 147–152.

Huxley, J.S., 1942. Evolution, the Modern Synthesis. Allen & Unwin, London.

Iacoboni, M., Woods, R.P., Brass, M., Bekkering, H., Mazziotta, J.C., Rizzolatti, G., 1999. Cortical mechanisms of human imitation. Science 286, 2526–2528.

Itan, Y., Powell, A., Beaumont, M.A., Burger, J., Thomas, M.G., 2009. The origins of lactase persistence in Europe. PLoS Computational Biology 5, e1000491.

Jensen, K., Call, J., Tomasello, M., 2007. Chimpanzees are rational maximizers in an ultimatum game. Science 318, 107–109.

Kirby, S., Dowman, M., Griffiths, T.L., 2007. Innateness and culture in the evolution of language. Proceedings of the National Academy of Sciences 104, 5241–5245.

Laland, K.N., Odling-Smee, J., Myles, S., 2010. How culture shaped the human genome: bringing genetics and the human sciences together. Nature Reviews Genetics 11, 137–148.

Lamba, S., Mace, R., 2011. Demography and ecology drive variation in cooperation across human populations. Proceedings of the National Academy of Sciences 108, 14426–14430.

Lewis, H.M., Laland, K.N., 2012. Transmission fidelity is the key to the build-up of cumulative culture. Philosophical Transactions of the Royal Society B 367, 2171–2180.

Lewontin, R.C., 1970. The units of selection. Annual Review of Ecology and Systematics 1, 1–18.

Lipo, C.P., O'Brien, M.J., Collard, M., Shennan, S. (Eds.), 2006. Mapping Our Ancestors: Phylogenetic Approaches in Anthropology and Prehistory. Aldine, New York.

Lycett, S.J., 2009. Understanding ancient hominin dispersals using artefactual data: a phylogeographic analysis of Acheulean handaxes. PLoS One 4, 1–6.

Lycett, S.J., Collard, M., McGrew, W.C., 2007. Phylogenetic analyses of behavior support existence of culture among wild chimpanzees. Proceedings of the National Academy of Sciences 104, 17588.

Mace, R., Holden, C., Shennan, S. (Eds.), 2005. The Evolution of Cultural Diversity: A Phylogenetic Approach. UCL Press, London.

Mace, R., Pagel, M.D., 1994. The comparative method in anthropology. Current Anthropology 35, 549–564.

Mathew, S., Boyd, R., 2011. Punishment sustains large-scale cooperation in prestate warfare. Proceedings of the National Academy of Sciences 108, 11375–11380.

McElreath, R., Lubell, M., Richerson, P.J., Waring, T.M., Baum, W., Edsten, E., Efferson, C., Paciotti, B., 2005. Applying evolutionary models to the laboratory study of social learning. Evolution and Human Behavior 26, 483–508.

Mesoudi, A., 2011. Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences. University of Chicago Press, Chicago, IL.

Mesoudi, A., Lycett, S.J., 2009. Random copying, frequency-dependent copying and culture change. Evolution and Human Behavior 30, 41–48.

Mesoudi, A., O'Brien, M.J., 2008. The cultural transmission of Great Basin projectile point technology I: an experimental simulation. American Antiquity 73, 3–28.

Mesoudi, A., Whiten, A., Laland, K.N., 2004. Is human cultural evolution Darwinian? Evidence reviewed from the perspective of *The Origin of Species*. Evolution 58, 1–11.

Mesoudi, A., Whiten, A., Laland, K.N., 2006. Towards a unified science of cultural evolution. Behavioral and Brain Sciences 29, 329–383.
Morgan, L.H., 1877. Ancient Society. Henry Holt, New York.
Norenzayan, A., Atran, S., Faulkner, J., Schaller, M., 2006. Memory and mystery: the cultural selection of minimally counterintuitive narratives. Cognitive Science 30, 531–553.
O'Brien, M.J., Darwent, J., Lyman, R.L., 2001. Cladistics is useful for reconstructing archaeological phylogenies: Palaeoindian points from the Southeastern United States. Journal of Archaeological Science 28, 1115–1136.
Pagel, M., 2009. Human language as a culturally transmitted replicator. Nature Reviews Genetics 10, 405–415.
Powell, A., Shennan, S., Thomas, M.G., 2009. Late Pleistocene demography and the appearance of modern human behavior. Science 324, 1298–1301.
Reyes-Garcia, V., Broesch, J., Calvet-Mir, L., Fuentes-Pelez, N., McDade, T.W., Parsa, S., Tanner, S., Huanca, T., Leonard, W.R., Martínez-Rodríguez, M.R., 2009. Cultural transmission of ethnobotanical knowledge and skills: an empirical analysis from an Amerindian society. Evolution and Human Behavior 30, 274–285.
Richerson, P.J., Boyd, R., 2005. Not by Genes Alone. University of Chicago Press, Chicago.
Rizzolatti, G., Craighero, L., 2004. The mirror-neuron system. Annual Review of Neuroscience 27, 169–192.
Spencer, M., Davidson, E.A., Barbrook, A.C., Howe, C.J., 2004. Phylogenetics of artificial manuscripts. Journal of Theoretical Biology 227, 503–511.
Sperber, D., 1996. Explaining Culture: A Naturalistic Approach. Oxford University Press, Oxford.
Stout, D., 2011. Stone toolmaking and the evolution of human culture and cognition. Philosophical Transactions of the Royal Society B: Biological Sciences 366, 1050–1059.
Tehrani, J.J., Collard, M., 2009. On the relationship between interindividual cultural transmission and population-level cultural diversity: a case study of weaving in Iranian tribal populations. Evolution and Human Behavior 30, 286–300.
Tennie, C., Over, H., 2012. Cultural intelligence is key to explaining human tool use. Behavioral and Brain Sciences.
Toelch, U., van Delft, M.J., Bruce, Donders, R., Meeus, M.T.H., Reader, S.M., 2008. Decreased environmental variability induces a bias for social information use in humans. Evolution and Human Behavior 30, 32–40.
Turchin, P., 2003. Historical Dynamics: Why States Rise and Fall. Princeton University Press, Princeton, NJ.
Tylor, E.B., 1871. Primitive Culture. John Murray, London.
West, S.A., El Mouden, C., Gardner, A., 2011. Sixteen common misconceptions about the evolution of cooperation in humans. Evolution and Human Behavior.
Whiten, A., Goodall, J., McGrew, W.C., Nishida, T., Reynolds, V., Sugiyama, Y., Tutin, C.E.G., Wrangham, R.W., Boesch, C., 1999. Cultures in chimpanzees. Nature 399, 682–685.

Cultural Evolution: Phylogeny versus Reticulation

Peter Bellwood, Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cultures, languages, and biological populations evolve in time and space through both phylogenetic processes of transgenerational inheritance and reticulate processes of interaction and borrowing (diffusion). Phylogenetic processes in the human context are set in motion by geographic expansion and especially by population migration into new territories. Long periods of nonmigratory interaction can lead eventually to the erosion of phylogenetic signals, but the circumstances under which erosion occurs can differ between material and social aspects of culture, language, and biology. In general, but restricted to the past few millennia, language families retain the most accessible traces of phylogenetic descent.

Why do human cultures and languages form patterns of similarity and differentiation through time and across space, and how do such patterns form? Why do archeological cultures and language families exist? When such patterns have been formed across large areas by processes of expansion, how long can we expect their traces to survive in archaeological and linguistic landscapes? For instance, by comparing living and ethnographic European or Polynesian cultures, languages and genetic profiles, to name two pertinent examples (from many) that are considered again in this article, can we expect to infer ancestral episodes of population migration and ethnogenesis that occurred within the ancestry of these peoples many millennia ago?

These are important questions, given that the major disciplines that investigate human prehistory are divided into two main groups. Some, like comparative linguistics and population genetics, depend on data located in the present or recent past. Others, like archaeology, philology, ancient DNA, and craniometric studies, derive their data directly from the deeper past. The former make inferences from comparison, and the latter from direct investigation of behavioral and biological residues, mostly recovered by archaeological excavation. On some occasions, inferences from these two discipline groups find accord, but sometimes they do not. Broad understanding of the deep human past therefore must be multidisciplinary and also conscious of issues connected with context and scale.

Although this article is concerned mainly with the formation of cultural and linguistic (i.e., behavioral, not genetic) patterns through time and space, the multidisciplinary perspective just proposed demands that we consider genetics and biology too, at least to some extent, because there can be little doubt that cultures and languages as we know them are related to specific populations. Such populations have, or once had, identifiable biological profiles, even if there is constant blurring of boundaries as people communicate, exchange genes and ideas, and sometimes undertake much less pleasant actions against their neighbors across their perceived cultural boundaries. This observation leads to the core concept of this article – cultures, languages, and biological populations must evolve through both descent-based and contact-based processes.

Naturally, total reliance on a biological concept, such as 'phylogeny' (evolutionary history of a genetically related group of organisms), within the anthropological, archeological, and linguistic disciplines is completely unwarranted, because it is apparent that culture is transmitted through both phylogenetic and reticulate processes. Thus, the transmission of cultural knowledge and language from generation to generation cannot be equated precisely, in terms of mechanism, with genetic inheritance alone. People mate and contribute 50% of their genes to each offspring, and offspring cannot acquire genes from any other source, except perhaps through modern transplant technology. But a husband and wife from different ethnolinguistic communities do not create a new language from 50% of the linguistic resources of each drawn at random, or build a house split down the middle to reflect their separate architectural traditions. This is partly why, if we travel in the twenty-first century in Africa, Southeast Asia, and the western Pacific, even Europe and India, we find millions of people who speak closely related languages, but who are as different in biological appearance as western Europeans and northern Indians (Indo-European speakers), Melanesians and Filipinos (Austronesian speakers), or Iraqis and Ethiopians (Afroasiatic speakers). There is nothing terribly mysterious about all of this. It means, quite simply, that we need concepts of both phylogeny and reticulation to understand the creation of historical and ethnographic situations.

The essence here, at any given time, is that the background to any specific cultural and linguistic situation can be considered in terms of inheritance, as recognized between different taxa from patterns in shared and uniquely innovated traits (phylogeny), and in terms of contact and diffusion (reticulation). These two processes always have worked hand-in-hand for all societies, except for those who have existed in utter isolation – not a common occurrence in human history, even for remote Pacific Islands. Any attempt to understand the genesis of a human society purely in terms of one and not the other process, therefore, generally would be futile. Nevertheless, clear awareness of the operating mechanisms of these two processes through time can assist prehistorians and historians to interpret the human past.

Evolution of phylogenetic patterning occurs mainly through linguistic and cultural divergence, following the initial spread of an ancestral form, whereas reticulation proceeds mainly through processes of convergence or borrowing between different entities, that may be related or unrelated in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.81039-9

ancestry. Phylogenetic relationships reflect patterns of common ancestry within language families and sociocultural complexes. They are of great importance because they often can imply the operation of widespread processes of human migration, sometimes rapidly and over large distances. Reticulation alone cannot create a phylogenetic array, just as new genera and species in the animal world cannot be created from unrelated forms that are incapable of breeding together. A new genus, by definition, requires a radiation of ancestral individuals if it is to come into existence. Both language families and cultures that reveal clear signs of a shared phylogenetic descent do so because of a shared basal history of radiation that in many cases occurred millennia back in time and that might have taken centuries or millennia to reach its limits. Indeed, the long-term trends within cultural evolution appear to oscillate between fairly brief and punctuated periods of phylogenetic radiation, often involving population migration, interspersed with much longer periods of reticulate ethnogenesis between relatively stable nodes within networks of communication. As we know from recent recorded history, migration has been a constant yet unevenly distributed process within human affairs.

As an example, since AD 1600, societies have developed in the United Kingdom, North America, and Australasia, which clearly share some degree of common history (phylogeny). This is evident in the almost universal use of the English language, in the direct and well-recorded biological ancestry (via migration) of many inhabitants of these regions from a British source, in the density of surnames of British and Irish origin in telephone books, and in many aspects of material and government culture that originated in ways of doing things widely current in Britain and Ireland between approximately AD 1600 and 1850. These societies in North America and Australasia originated historically via a process of colonization and early political control emanating from the British Isles. Similar generalizations can be made about the spread of Iberian culture, language, and biology through much of Central and South America, although in the early sixteenth century this was assisted by a disease impact on Native American societies far greater than occurred in later centuries (Denevan, 1992; O'Fallon and Fehren-Schmitz, 2011).

Despite these clear indications of shared ancestry, however, it is quite obvious now that by no means all of the cultural and biological patterning found in these regions is of a circumscribed British (or Spanish, or Portuguese) origin. Indigenous peoples have contributed, as have migrants from other regions. Cross-cultural borrowings abound and occur continuously, forming a network of horizontal transmission, or reticulation. Thus, we have the two different conceptual directions of transmission discussed here for the elements of both cultural and biological patterning. Naturally, in a real-life situation, such as the modern United States, it is impossible to untangle the inherited from the borrowed in an overall sense and to present the results in the form of a statistical statement. These are simply concepts that allow us to model the formation of cultural patterns through time, essentially to understand the main trends of history (for discussion, see Durham, 1991; Mace and Pagel, 1994; Bellwood, 1996, 2001, 2005; Dixon, 1997; Mace and Holden, 2005; Richerson and Boyd, 2005; Stark et al., 2008; VanPool et al., 2008).

Phylogeny and Reticulation in Prehistoric Contexts

For prehistoric situations without the benefit of philological materials, there can be no direct linguistic evidence that reflects the deep past, apart from that derived from the application of a comparative methodology to living and recorded languages. Archaeological data survive directly from the past but are subject to varying degrees of erosion from the operation of natural processes in the ground. It might be assumed, however, that material culture combinations in the archaeological record are likely to reflect human populations who shared languages and relatively endogamous breeding networks if they were complex and internally coherent, geographically widespread, and punctuational in appearance, and if they can be shown to have undergone processes of descent-based (phylogenetic) differentiation. Such an assumption might be wrong in certain circumstances, but at least it can serve as a useful working hypothesis.

Demonstration of descent-based differentiation for archaeological categories through space and time is becoming a major focus of research in modern archaeology, just as it has been for a long time in human biology and linguistics (Steele et al., 2010). An excellent recent example of this, among many, is focused on the spread of Neolithic plant economies of Western Asian origin into Europe between 6500 and 5000 BC (Coward et al., 2008). This mathematical analysis of 250 early Neolithic crop assemblages from archaeological sites demonstrates the existence of two separate combinations of crop species moving along two separate routes, one via the Mediterranean and the other up the Danube Valley through central Europe north of the Alps. In both of these regions, such crops were manifestly not present before the Neolithic, and the overall pattern fits the well-established picture, derived from other aspects of archaeology, linguistics, and genetics of a two-pronged Neolithic migration from the Near East into Europe starting around 6500 BC. I refer to the Neolithic spread to the north of the Alps again later in this article.

It is important in this regard, however, to differentiate between spreads of people, languages, and other aspects of culture. The movements of coherent crop complexes alone through Neolithic Europe cannot tell us clearly whether the vectors of movement were migrating human populations or simply an exchange of plant resources between autonomous groups of hunter-gatherers, although ethnographic and historical records of human activity in recent millennia reveal that the latter eventuality would be most unlikely from a comparative perspective (Bellwood, in press). Even so, ethnographic and historical records do offer examples of situations in which patterns of biological, linguistic, and sociocultural variation do not correlate at all, or correlate only slightly. On the other hand, they sometimes can correlate with almost 100% precision, as in the early stages of population expansion into new territories. Whole populations already *in situ* geographically can adopt new languages and cultural forms through contact-induced change (Ross, 1997), whereas at the other extreme whole populations can colonize new territories with closely correlated physical attributes and cultural characteristics.

In terms of human prehistory, it is important to remember the significance of scale. On a small scale, the inhabitants

of a village or small river basin could, and indeed have in many historical and ethnographic examples, adopt a new language and undergo genetic change as a result of intermarriage with members of an outsider group of people, perhaps of higher status. Ethnic groups can also 'fuse' under situations of depopulation or outsider-imposed stress, as has happened frequently in the period of tribal shrinkage, which history offers us in the form of the ethnographic present (for examples, see Atkinson, 1989; Moore, 1994; Nicolaisen, 1977–1978).

On a much larger scale, however, a chain-reactive process of language and cultural shift of this kind will not explain why, for instance, related languages are spoken over vast regions of the Earth in the form of major language families, such as Indo-European, Austronesian, or Uto-Aztecan (see Figure 1 in Language Families, Archeology and History of). These language families had reached their outer geographic limits before the beginnings of written history, and no historical processes such as military conquest or mass media persuasion can hope to explain them. Alexander the Great spread the Indo-European language family only on a ephemeral basis, and over the long term, his conquests had no impact at all on its distribution. Such extensive distributions are records of major human dispersals spreading foundation languages and archeological cultures. Many such dispersals undoubtedly owed their existences to population increase and dispersal consequent on the regional beginnings of agro-pastoral food production (Bellwood, 2005, 2010, in press). But this is not true for all, because many languages, such as English and Spanish, have spread in recent historical times over vast distances through migration, especially at the level of native-speaker populations, and also can be regarded as potential foundation languages for future language families. Comparative linguistics and ethnohistory also reveal some excellent examples of large-scale hunter-gatherer dispersal, especially in the Eskimo-Aleut and Athabaskan language families in North America.

One important observation from comparative linguistics, however, is that the constant evolution of any language in terms of phonological, lexical, and grammatical borrowing and change will ensure that very deep relationships in time, for instance from the Pleistocene epoch (more than 12 000 years ago), will be extremely ephemeral and difficult to trace. Such time depths take us back into the linguistic field of macrofamilies, a field always fraught with controversy.

The Importance of Migration in Spreading Phylogenetic Arrays

Wherever phylogenetic relationships can be demonstrated to exist between the members of an array of languages or cultures (including archeological cultures), and biological populations as well, dispersal of an ancestral complex must be inferred. Whether this ancestral dispersal directly related to a movement of actual people via migration is something that must be decided on a case-by-case basis. In the case of extremely widespread phylogenetic entities, population movement to some

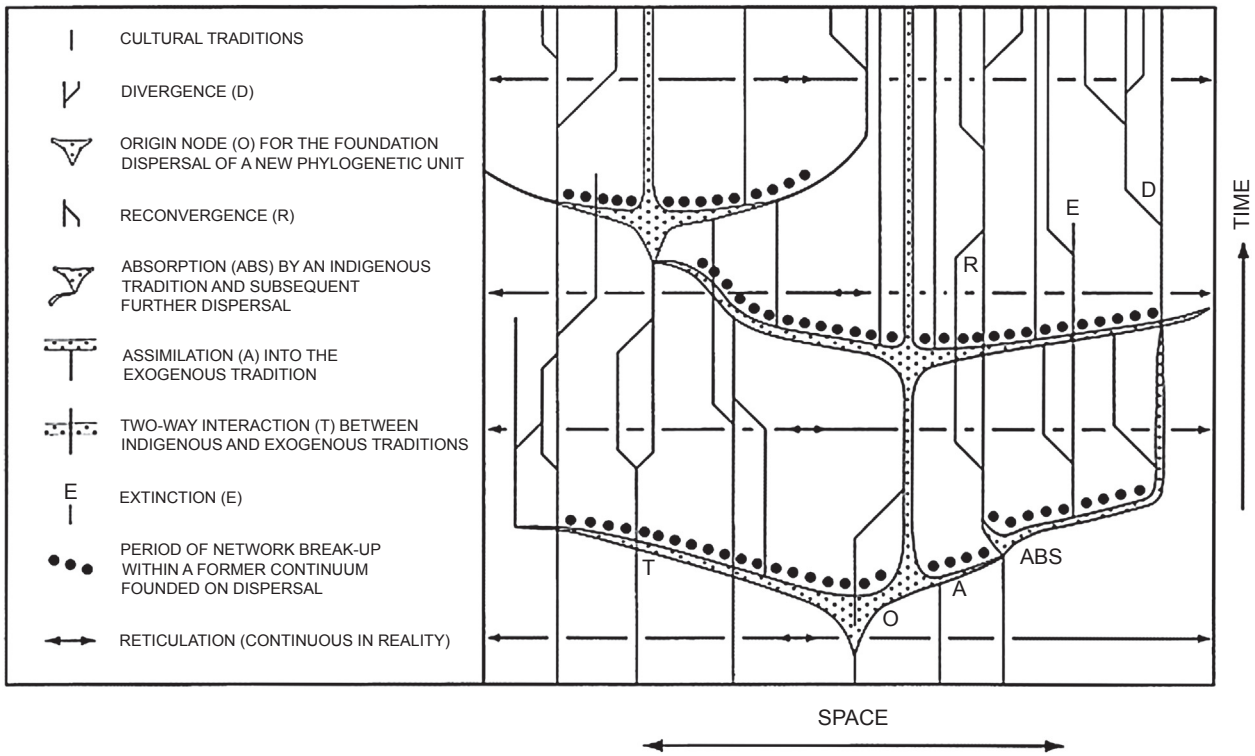

Figure 1 Phylogeny and reticulation in time and space perspective. (NB: A diagram such as this cannot replicate the full complexity of cultural evolution in any real-life situation. It is intended only as a schematic model.)

degree can hardly be denied. In addition, most language families clearly spread relatively quickly in overall chronological terms. From internal evidence and archeological correlations, most reached their geographic limits within 3000 years or less of the commencement of dispersal – for example, between 2000 and 3000 years from origin regions to prehistoric limits in the cases of Indo-European and Austronesian. Speed of dispersal, in this regard, relates directly to the intensity of punctuation.

The comparative linguistic record also can provide clear structural evidence for rapid dispersal, especially when the component language subgroups within a family do not share any unique linguistic innovations that allow them to be subgrouped phylogenetically above the individual subgroup level. In such situations, 'descent trees' created to show the structure within such language families will be extremely rakelike, rather than treelike with many successive branches. In such cases, claimed for instance for several of the major Austronesian language family subgroups of Southeast Asia and the Pacific (Pawley, 1999), there is a strong likelihood that the resulting pattern reflects an initial history of rapid and far-flung dispersal, followed by a formation of independent subgroups, each sharing a common ancestry in the initial dispersal but all reflecting an independent history thereafter.

Direct demonstration of rapid dispersal also can be identified in the archeological record, as for example when a particular combination of artifact style and decoration too complex to reflect functional selection alone is distributed widely in a short time. Examples include the Early Neolithic Linearbandkeramik (LBK) cultural complex, which spread through central Europe about 5400 BC (Shennan, 2009); the Chavin stylistic complex, which spread through the Peruvian Andes after 500 BC (Heggarty and Beresford-Jones, 2010); and the Lapita cultural complex with its distinctive and elaborately decorated pottery, which spread through the western Pacific about 1350–900 BC (Kirch, 1997). In contextual terms, it is apparent that these dispersals of material culture were correlated with dispersals of people and perhaps language, even though the nature of the archeological record does not allow this to be stated as more than a logical possibility rather than a demonstrated fact (we will return to the LBK from a genetic perspective in the following). A parallel with the diversification of language families from a common and relatively homogeneous source also appears in the archeological record when widespread complexes, such as those already mentioned, are followed by a breakdown of widespread homogeneity into regional diversity, with this diversity becoming more marked over time. The later phases of LBK, Chavin and Lapita, illustrate such dramatic fragmentations very clearly.

Clearly, when dealing with migration, it is important to distinguish between populations who migrated into territories devoid of prior human inhabitants and those who moved into already settled areas. In both situations, the inception of a successful trajectory of migration and colonization obviously would have required the acquisition of some form of advantage allowing demographic growth and range expansion – this would have been particularly true in situations in which large native populations already were established. In many of the examples discussed thus far, the advantage seems to have been the acquisition of systematic agriculture, but the whole story is far more complex. The initial human colonizations of empty continents like Australia (50 000 years ago) and the Americas (16 000 years ago) must have led to the dispersal of phylogenetically related languages and cultures, of which some faint traces might have survived in later prehistoric and ethnographic cultural patterns (Bellwood, 2013).

Phylogenetic Studies of Archeological and Ethnographic Cultures: Some Oceanic Examples

The ethnographic and archeological records are replete with examples of cross-cultural contact via trade, warfare, and intermarriage, indeed by 'diffusion' and gene flow in their many forms. Past humans have been just as interested in communicating with their neighbors, whether as friends or foes, as have recent ones. But evidence that ethnographic and prehistoric cultures form phylogenetic arrays indicative of dispersal is not always so obvious once one moves beyond the evidence from comparative linguistics. Normally, teasing out of the nonlinguistic aspects of the pattern requires attention to details from archaeology and biological anthropology (especially genetics), assisted by information from comparative ethnology and, where relevant, historical records and oral tradition (for examples, see Romney, 1957; Flannery and Marcus, 1983; Kirch and Green, 2001; Bellwood, 2005).

As a case study, a debate has emerged in recent years among anthropological scholars working in the islands of Polynesia. Polynesians form a population of many island societies in the great triangle bounded by Hawaii, Easter Island, and New Zealand, including Tonga and Samoa, with a few additional 'Polynesian Outliers' existing to the west among the larger islands of Melanesia. That Polynesian languages share a relatively recent common ancestry is not in doubt, but beyond language, the explanations for the Polynesian cultural pattern are more contested. Two recent perspectives on the Polynesians – the 'phylogenetic' and the 'reticulate' – can be counterbalanced.

The archaeologists Patrick Kirch and Roger Green (1987, 2001) have presented Polynesia as a 'phylogenetic unit' of closely related cultures, languages, and peoples all sharing a common origin derived from the settlement of the Tongan, Samoan, and other nearby Western Polynesian islands by Lapita populations at about 1000–800 BC. Working from the archeological and linguistic records, Kirch and Green suggested that the Polynesian societies of ethnographic times formed a closely related set in terms of material culture, social organization, biological phenotype, and most important, language. Societies so related can be conceived as something akin to the genus of biologists – a group of 'species' defined as a group through possession of certain uniquely shared features (synapomorphies in biological terminology). From this perspective, Polynesian societies all share a single origin, and since that period of origin, they have differentiated because of a wide range of historical processes reflecting both internal change and external contacts with other societies both inside and outside Polynesia. Because Polynesian societies are geographically so clearly bounded by sea, and because archaeology and linguistics lead us to believe that only Polynesians ever inhabited Polynesia before European contact,

the region lends itself ideally to such a phylogenetically oriented historical perspective.

But not all agree entirely with a phylogenetic perspective. The archaeologist John Terrell and his colleagues (Terrell et al., 1997; Terrell and Welsch, 1997; Terrell, 2012) have taken an opposed and rather more reticulated viewpoint, pointing out that most ancient Pacific Island (including Polynesian) societies were never isolated and always would have maintained some degree of contact between different regions. Virtually all archaeologists would agree with this statement in simple form, because the archeological, linguistic, and genetic records carry widespread evidence of trade, borrowing, intermarriage, and so forth between different island populations. But the reticulate perspective is taken further by Terrell and his colleagues to draw the conclusion that interaction will erase any recognizable traces of cultural phylogeny (for a similar statement, also see Moore, 1994). For these scholars, the phylogenetic ancestry of cultures is generally unknowable and thus cannot contribute to an understanding of cultural history.

The perspective offered in this article agrees that a masking of phylogenetic history can occur in cases in which actual mixing between cultural entities of entirely different origin and ancestry has taken place, especially over a long period of time. Such intricate mixing on the level of material culture is believed to have occurred, for instance, during the past 3000 years in some of the Austronesian and Papuan-speaking societies of lowland New Guinea (Welsch et al., 1992; but see also Roberts et al., 1995 for a contrary view). But when 'like meets like,' as generally would have been the case in Polynesia, where all cultures shared a recent common origin, then erasure of all traces of phylogenetic relationship becomes much harder to imagine. Even in the case of lowland New Guinea, the fact that the Austronesian and Papuan languages have not in themselves generally become creolized tells us something about the strength of cultural, especially linguistic, transmission through time, and from generation to generation.

Both the phylogenetic and the reticulate viewpoints therefore are correct to a degree, and disagreement occurs only when they are placed artificially in opposition to each other. Polynesia, and especially the vast region of eastern Polynesia, has a very short time-depth, a relatively high degree of island isolation, and hence superb conditions for the survival of a record of phylogenetic spread and diversification, visible to archaeologists, linguists, and biologists alike (e.g., Buck and Vidarsdottir, 2012 for craniofacial evolution). New Guinea, on the other hand, has a great time depth for human settlement (ca. 50 000 years or more) and a relatively long period of interaction between phylogenetically unrelated Papuan and Austronesian cultures (ca. 3000 years). In many regards, however, New Guinea can perhaps be regarded as more typical than Polynesia for understanding long-term history in continental areas, as we can see from a new and significant genetic contribution to the prehistory of Europe.

A Perspective from the Holocene Prehistory of Central Europe

In the past decade, an important debate, summarized concisely by Ron Pinhasi et al. (2012), has developed within human genetics concerning the biological phylogeny of central European populations. Many regions north of the Alps, from France across to almost the Black Sea, were settled decisively by Neolithic colonists between about 5400 and 5000 BC, carriers of the LBK cultural complex of archaeologists, characterized by large longhouse villages established in regions of fertile loess soil. In archaeological terms, this complex replaced the preceding Mesolithic hunter-gatherer cultures quite sharply, and some also have raised the possibility that this was the period of initial dispersal of ancestral Indo-European languages into Europe (Renfrew, 1987). Genetics, however, steals the show at the moment, especially as a result of recent advances in the extraction of ancient DNA from human bones in prehistoric cemeteries.

Samples of ancient DNA from both Mesolithic and Neolithic skeletons in several regions of Europe, especially the LBK region, reveal that the Mesolithic populations generally were replaced by new farming peoples who came from regions to the east, possibly the eastern Mediterranean or Anatolia. In the case of the LBK, they probably traveled via the Carpathian Basin of what is today Hungary, this being the region of LBK genesis in an archaeological sense. Neolithic replacement of earlier hunter-gatherer populations, in the LBK context, therefore was fairly marked, although some archaeologists have argued for limited continuity in the record of flaked lithic tools. After all, not even the recent colonizations by Europeans of former hunter-gatherer territories in North America and Australia have caused all native peoples to disappear, and we hardly can expect a worse fate for those at interface points deep in prehistory. Many hunter-gatherers must have survived the arrival of farmers in their territories, and many (especially females) probably joined the farming populations and contributed their genes to the mix.

The most interesting observation from the new genetic research, however, concerns the question 'who are the present inhabitants of central Europe?' Surprisingly, they do not appear to descend directly, with little admixture, from either Mesolithic or Neolithic ancestors. We find instead, as discussed by Pinhasi and his colleagues, that their modern genetic characteristics are just as likely to reflect post-Neolithic demographic processes, including the effects of subsequent migrations, natural selection, genetic drift, and so forth, with none actually yet documented genetically in any precise way. There is a lesson here for geneticists in particular, this being that reconstruction of population origins and prehistory at a time depth approaching 7500 years ago cannot be done entirely by comparing the genomes of modern living people. There is no doubting the importance of the Neolithic colonization of Europe north of the Alps, and the cultural developments that occurred there most certainly did not originate entirely in the local Mesolithic. But this does not mean that all Europeans who have occupied that region thereafter, to the present, trace their ancestry only to Neolithic settlers. Migration has been of constant importance in human affairs, and cumulative genetic changes through time indeed might have masked deep levels of phylogenetic descent, as might long-term cumulative changes in culture and language. As noted for Polynesia versus New Guinea, however, the intensity of much of this masking probably will reflect the time depth involved. After all, the 7500 years that take us back to the beginning of the LBK beat the date for the construction of the Egyptian pyramids at Giza by 3000 years. That is a very long period of time.

Conversely, most of the major Polynesian migrations beyond Tonga and Samoa occurred only within the past 1000 years.

What is significant from the new European data on the LBK, however, is that linguistics, archaeology, and genetics throw different pools of light on the deep past, and those pools do not always overlap with 100% precision. Archaeologically, no one could claim that modern European cultures retain artifacts that descend directly, in phylogenetic isolation, from LBK forebears – archaeological phylogeny from 5400 BC to the present effectively has been severed completely. So has that contained within the DNA of modern populations. Skeletally, there might be more continuity, but no systematic study comparing LBK skeletons with those of modern Europeans has been done. The linguistic data, however, if interpreted to indicate an initial spread of ancestral Indo-European languages from Anatolia through Europe commencing around 6500 BC (Bouckaert et al., 2012), reveal phylogenetic relationships very clearly at the whole family level, even if historical expansions of languages, such as Latin, Spanish, English, and Russian, have erased much earlier patterning. The converse is that comparative linguistics cannot easily take us back into the Pleistocene, whereas archaeology and studies of ancient DNA increasingly are doing so.

These observations about Neolithic Europe and Polynesia are reflective of the widespread modern interest in the prospects of combining linguistic, archeological, and genetic perspectives on the human past. A main supporter of a dialogue between these disciplines of prehistoric reconstruction has been the archeologist Colin Renfrew in his writings on Indo-European language prehistory (Renfrew, 1987) and broader relationships (Renfrew, 2002). Renfrew's position, like that of Kirch and Green for Polynesia, and my own (Bellwood, 2011; Bellwood et al., 2011) for Island Southeast Asia and Oceania, and for world populations in general (Bellwood, 2005, in press), is that the prehistories of large-scale ethnolinguistic populations must be approached from the independent, but overlapping, perspectives of genetics, archaeology, and linguistics, and not just from the isolated perspectives of single disciplines.

An attempt is made finally to encapsulate some of the cultural trajectories discussed above in the form of Figure 1. This diagram highlights both divergent and convergent evolutionary situations, which should be regarded as overlapping possibilities rather than as absolutely separate processes. Divergent situations reflect the break-up of communication networks and especially the foundation by dispersal of new phylogenetic units (e.g., the dispersal of the foundation dialect chain for a major language family). Situations of convergence or amalgamation include the periodic cultural reconvergence of recently diverged traditions that already are related phylogenetically, the absorption of dispersing cultural entities by indigenous populations, the converse assimilation of indigenous traditions by exogenous ones, and balanced two-way interaction (the middle road between absorption and assimilation). Also to be taken into account is the possibility of autonomous cultural extinction, perhaps as a result of environmental problems, whether naturally or humanly stimulated.

Four other processes are shown schematically in Figure 1. These are generation-to-generation transmission of cultural traditions through time, a schematic horizontal representation of interaction (reticulation), potential for population growth (rapid during episodes of population dispersal), and chronology. The aim of the diagram is, in part, to indicate that human prehistory cannot be purely regarded as a process of local evolution through all time, but rather that major episodes of dispersal of the foundation cultures for future phylogenetic arrays punctuate the mesh of gradual reticulate ethnogenesis. Without such dispersals there could be no major language families, and none of the rapid and large-scale spreads of prehistoric agricultural economies and cultural complexes that we read from the archeological record in many parts of the world.

See also: Comparative Method in Evolutionary Studies; Cultural Evolution: Overview; Evolution and Language: Overview; Evolution and Language: Phylogenetic Analyses; Language Families, Archaeology and History of; Technology, Evolution of.

Bibliography

Atkinson, R., 1989. The evolution of ethnicity among the Acholi of Uganda: the precolonial phase. Ethnohistory 36, 19–43.

Bellwood, P., 1996. Phylogeny and reticulation in prehistory. Antiquity 70, 881–890.

Bellwood, P., 2001. Early agriculturalist population diasporas? Farming, languages and genes. Annual Review of Anthropology 30, 181–207.

Bellwood, P., 2005. First Farmers. Blackwell, Oxford and Malden, MA.

Bellwood, P., 2010. Language families and the history of human migration. In: Bowden, J., Himmelmann, N., Ross, M. (Eds.), A Journey through Austronesian and Papuan Linguistic and Cultural Space: Papers in Honour of Andrew K. Pawley. Pacific Linguistics, Canberra, pp. 79–93.

Bellwood, P., 2011. Holocene population history in the Pacific region as a model for world-wide food producer dispersals. Current Anthropology 52 (S4), 363–378.

Bellwood, P. 2013. First Migrants: Ancient Migration in Global Perspective. Wiley-Blackwell, Boston and Oxford.

Bellwood, P., Chambers, G., Ross, M., Hung, H-c, 2011. Are "cultures" inherited? Multidisciplinary perspectives on the origins and migrations of Austronesian-speaking peoples prior to 1000 BC. In: Roberts, B., Van der Linden, M. (Eds.), Investigating Archaeological Cultures: Material Culture, Variability and Transmission. Springer, Dordrecht, pp. 321–354.

Bouckaert, R., Lemey, P., et al., 2012. Mapping the origins and expansion of the Indo-European language family. Science 337, 957–960.

Buck, T., Vidarsdottir, U., 2012. Craniofacial evolution in Polynesia: a geometric morphometric study of population diversity. American Journal of Human Biology 24, 776–785.

Coward, F., et al., 2008. The spread of Neolithic plant economies from the Near East to northwest Europe: a phylogenetic analysis. Journal of Archaeological Science 35, 42–56.

Denevan, W., 1992. The pristine myth: the landscapes of the Americas in 1492. Annals of the Association of American Geographers 82, 369–385.

Dixon, R., 1997. The Rise and Fall of Language. Cambridge University Press, Cambridge, UK.

Durham, W., 1991. Coevolution: Genes, Culture, and Human Diversity. Stanford University Press, Stanford, CT.

Flannery, K., Marcus, J., 1983. The Cloud People. Academic Press, New York.

Heggarty, P., Beresford-Jones, D., 2010. Agriculture and language dispersals: limitations, refinements, and an Andean exception? Current Anthropology 51, 163–192.

Kirch, P., 1997. The Lapita Peoples. Blackwell, Oxford, UK.

Kirch, P., Green, R., 1987. History, phylogeny, and evolution in Polynesia. Current Anthropology 28, 431–456.

Kirch, P., Green, R., 2001. Hawaiki: Ancestral Polynesia. Cambridge University Press.

Mace, R., Holden, C., 2005. A phylogenetic approach to human evolution. Trends in Ecology and Evolution 20 (3), 116–121.

Mace, R., Pagel, M., 1994. The comparative method in anthropology. Current Anthropology 35, 549–564.

Moore, J., 1994. Putting anthropology back together again: the ethnogenetic critique of cladistic theory. American Anthropologist 96, 925–948.

Nicolaisen, I., 1977–1978. The dynamics of ethnic classification: a case study of the Punan Bah of Sarawak. Folk 19-20, 183–200.

O'Fallon, B., Fehren-Schmitz, L., 2011. Native Americans experienced a strong population bottleneck coincident with European contact. Proceedings of the National Academy of Sciences 108, 20444–20448.

Pawley, A., 1999. Chasing rainbows: implications of the rapid dispersal of Austronesian languages for subgrouping and reconstruction. In: Zeitoun, E., Li, P. (Eds.), Selected Papers from the Eighth International Conference on Austronesian Linguistics. Institute of Linguistics, Academia Sinica, Taipei, pp. 95–138.

Pinhasi, R., Thomas, M., Hofreiter, M., Currat, M., Burger, J., 2012. The genetic history of Europeans. Trends in Genetics 28, 496–505.

Renfrew, C., 1987. Archaeology and Language. Jonathan Cape, London.

Renfrew, C., 2002. 'The emerging synthesis': the archaeogenetics of language/farming dispersals and other spread zones. In: Bellwood, P., Renfrew, C. (Eds.), Examining the Farming/Language Dispersal Hypothesis. McDonald Institute for Archaeological Research, Cambridge, pp. 3–16.

Richerson, P., Boyd, R., 2005. Not by Genes Alone. University of Chicago Press.

Roberts, J., Moore, C., Romney, A., 1995. Predicting similarity in material culture among New Guinea villages from propinquity and language: a log-linear approach. Current Anthropology 36, 769–788.

Romney, A., 1957. The genetic model and Uto-Aztecan time perspective. Davidson Journal of Anthropology 3, 35–41.

Ross, M., 1997. Social networks and kinds of speech-community event. In: Blench, R., Spriggs, M. (Eds.), Archaeology and Language I: Theoretical and Methodological Orientations. Routledge, London, pp. 209–261.

Shennan, S., 2009. Evolutionary demography and the population history of the European early Neolithic. Human Biology 81, 339–355.

Stark, M., Bowser, B., Horne, L., 2008. Why breaking down boundaries matters for archaeological research on learning and cultural transmission. In: Stark, M., Bowser, B., Horne, L. (Eds.), Cultural Transmission and Material Culture. University of Arizona Press, Tucson, pp. 1–16.

Steele, J., Jordan, P., Cochrane, E., 2010. Evolutionary approaches to cultural and linguistic diversity. Philosophical Transactions of the Royal Society of London, Series B 365, 3781–3785.

Terrell, J., 2012. Polynesians and the Seductive Power of Common Sense. Cultural Geographies. Published online 31 May 2012. http://dx.doi.org/10.1177/1474474011432663.

Terrell, J., Hunt, T., Gosden, C., 1997. The dimensions of social life in the Pacific. Current Anthropology 38, 155–196.

Terrell, J., Welsch, R., 1997. Lapita and the temporal geography of prehistory. Antiquity 71, 548–572.

VanPool, T., Palmer, C., VanPool, C., 2008. Horned serpents, tradition, and the tapestry of culture. In: O'Brien, M.J. (Ed.), Cultural Transmission and Archaeology. Society for American Archaeology, Washington DC, pp. 77–90.

Welsch, R., Terrell, J., Nadolski, J., 1992. Language and culture on the North Coast of New Guinea. American Anthropologist 94, 568–600.

Cultural Evolution: Theory and Models

Laurel Fogarty and Marcus W Feldman, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.W. Feldman, volume 5, pp. 3057–3063, © 2001, Elsevier Ltd.

Abstract

We define culture as those aspects of thought, speech, behavior, and artifacts that can be learned and transmitted socially. Although the cognitive processes that allow cultural transmission may involve the action of genes, transmission of certain types of information may also be purely cultural and may therefore result in evolutionary dynamics different from those of genetic evolution. Cavalli-Sforza and Feldman and Boyd and Richerson developed a quantitative theory for the cultural transmission and evolution of traits within populations. They sought principles that could explain observed levels of variation or estimated rates of change in cultural traits in a manner analogous to, but different from, those that have proven useful in evolutionary biology.

Most models of social learning and cultural evolution have used deterministic approaches. These include models analogous to those of formal population genetics, nonspatial evolutionary game theory, reaction–diffusion systems that assume infinite populations, phylogenetic and cladistic methods, interacting particle systems, network-based models, and cultural demographic models. This article reviews deterministic approaches to cultural evolution and agent-based simulations of social learning strategies. The latter have become increasingly popular and useful as computers became faster and cheaper.

Introduction

More than 150 definitions of 'culture' have been proposed in the literature of anthropology, archeology, sociology, and other disciplines. In the following account, the article takes 'culture' to mean those aspects of thought, speech, behavior, and artifacts that can be learned and transmitted socially. This process of transmission may involve the action of genes through genetic predispositions toward learning certain types of information, for example, but may also be purely cultural and may therefore result in evolutionary dynamics that are different from those of models that allow only genetic transmission. Cultural transmission may involve any or all of imprinting, conditioning, local enhancement, stimulus enhancement, emulation, imitation, or active teaching. Cavalli-Sforza and Feldman (1973a, 1981) and Boyd and Richerson (1985) developed a quantitative theory for the cultural transmission and evolution of traits within populations. Such traits may take either discrete or continuous values. The idea here was to seek principles that could explain observed levels of variation or estimated rates of change in a manner analogous to (but, of course, different from) those that have proven useful in evolutionary biology.

There are differences between the transmission of cultural traits and genes and therefore, between the dynamics expected in cultural and biological evolution. For example, cultural evolution, which in humans (at least) entails that learned traits may be transmitted, allows for a kind of Lamarckian inheritance (Maynard Smith, 1986; Mesoudi, 2011). Genetic models can be expanded so that the members of a population are represented simultaneously by their genotypes and their phenotypes (i.e., their phenogenotypes). Their genotypes are subject to the normal Mendelian rules of transmission but their phenotypes may be learned vertically, from parents or from peers or other older members of the population. The use of methods to describe the transmission and evolution of culture analogous to those used in population genetics was initially met with some resistance. The main reasons were (1) it is difficult to characterize the exact nature of a unit of cultural transmission that could be directly analogous to the gene in genetic evolution, (2) cultural transmission was not considered accurate enough to allow transmission of intact traits (Sperber, 1996; Atran, 2001), and (3) culture was often regarded as a set of ideational properties that could not be subdivided into transmissible components, but could only be valid as a gestalt concept.

Durham (1991) argued that the use of 'cultural traits' as the units of cultural transmission 'pits nature against nurture'; using 'traits' as units sidesteps how the trait is determined. However, by specifying the model of transmission, which may be purely cultural, purely genetic, or some combination of these (Feldman and Cavalli-Sforza, 1976, 1979), and which can in principle be empirically estimated, evolutionary analysis becomes possible. In principle, Durham's alternative 'ideational units' are amenable to the same kind of quantitative analysis as that originated by Cavalli-Sforza and Feldman. In 1976, Dawkins suggested the term 'memes,' which he defined as replicators that could be retained, inherited, and mutated by means that were directly analogous to genetic inheritance. Memes were self-propagating and selected on features like 'stickiness' (propensity to remain in the mind) and 'catchiness' (the propensity to be passed from one mind to another). In this characterization, the human mind was a meme-machine shaped by memes to allow their own propagation and evolution (Dawkins, 1976). This characterization of culturally transmitted entities has led some authors to endow memes with an almost independent existence (Blackmore, 1999).

Henrich and Boyd (2002) attempted to answer critics of cultural evolution using a series of models showing that a discrete 'substrate' and error-free transmission were not in fact necessary for Darwinian cultural evolution to occur. Their

model showed that features unique to cultural transmission such as conformity and prestige bias could compensate for high error rates in transmission and allow the propagation of favorable traits, even in the absence of high-fidelity copying. In another model, assuming continuous rather than discrete representations of cultural traits, Henrich and Boyd (2002) showed that a continuous-trait model with strong 'attractors' (individual cognitive transformation of cultural traits) and weak selection reduced to a discrete-trait model with weak selection. These authors succeeded in showing that, at least with respect to the problems of particulate evolution and copy error, it was reasonable to use appropriately modified genetic models to reconstruct human cultural evolution, by analogy to the way geneticists had done for biological evolution.

Approaches to Cultural Evolutionary Theory

To date, most models of social learning and cultural evolution have used deterministic approaches. These include models analogous to those of formal population genetics (Boyd and Richerson, 1985; Cavalli-Sforza and Feldman, 1981), nonspatial evolutionary game theory (Maynard Smith, 1982), reaction–diffusion systems that assume infinite populations (Cavalli-Sforza and Feldman, 1981; Kandler and Steele, 2009), phylogenetic and cladistic methods (e.g., O'Brien et al., 2001; Mace and Holden, 2005), interacting particle systems (Hauert, 2001; Nowak and May, 1992), network-based models (Watts, 1999; Franz and Nunn, 2009), and cultural demographic models (Fogarty et al., 2013; Fogarty and Feldman, 2011; Powell et al., 2009; Li et al., 2000). Here we focus on deterministic approaches to cultural evolution and agent-based simulations of social learning strategies. The latter have become increasingly popular and useful as computers became faster and cheaper.

Types of Cultural Transmission and Their Consequences

Cavalli-Sforza and Feldman (1981) identified three types of cultural transmission analogous to vertical, horizontal, and oblique disease transmission in epidemiological models. These occur between parents and offspring, peers, and older and younger individuals, respectively. In order to model these processes, we consider individuals in a large population at time t, where t is on a scale of generations, and generation t represents the parents of generation $t+1$. It is common to assume that these generations are nonoverlapping with generation t physically replaced by generation $t+1$. Of course, for culturally transmitted traits, this does not preclude a contribution of generation t to generation $t+2$, for example.

Vertical Transmission

If an offspring's probability of acquiring trait l depends on the traits j and k of its parents, we may represent the set of such transmission probabilities or rates as $\Gamma = \{\beta_{jk,l}\}$. Given the frequencies of each representation of the trait in the parental generation at time t, $\{x_{i,t}\}$ Γ determines $\{x_{i,t+1}\}$. Mating between a pair of parents may occur at random, or we may assume, as is likely for many culturally transmitted traits, that individuals of the same trait value prefer to mate with one another. This is called 'assortative' mating and occurs with probability m, say. An example of a dichotomous trait with forms 0 and 1 (such that $i = 0$ or $i = 1$) is given in Table 1.

Then, if we write x_i' for the frequencies in the offspring, that is, generation $t+1$, we have

$$x_0' = x_0^2 B + x_0 C + b_{11,0} + mBx_0(1 - x_0) \qquad [1]$$

where $B = b_{00,0} - b_{10,0} - b_{01,0} + b_{11,0}$; $C = b_{10,0} + b_{01,0} - 2b_{11,0}$.

Iteration of eqn [1] produces the complete evolutionary trajectory for the population frequencies of this trait. Note that $b_{00,0} = 1$, $b_{01,0} = b_{10,0} = \frac{1}{2}$, $b_{11,0} = 0$ would be the expected transmission rule if 0 and 1 were alleles at a genetic locus in a sexual haploid biological organism. The closer the bs are to these values, the more the dynamics resemble those of purely genetic evolution. But the system shows interesting flexibility; for example, if the probability that an offspring is unlike its parents is high enough, a cyclic trajectory results (Feldman and Cavalli-Sforza, 1976).

Within the framework of vertical cultural transmission, the parameters $\{\beta_{ij,k}\}$ may be disaggregated into separate additive contributions from each parent, although other partitions are conceivable. Estimation of these transmission parameters, for example, under an assumption of equilibrium, is possible from familial correlations such as father–offspring or sib–sib correlations, while m is usually measured as the correlation between mates. However, this may be biased by mating convergence, namely, parents becoming more alike over the time they are together.

Empirical studies designed to assess the importance of vertical transmission of everyday preferences among Stanford University students revealed that only religious and political attitudes were strongly vertically transmitted. Correlations among the designated parents of these students were reasonably high, suggesting a moderate value for m (Cavalli-Sforza et al., 1982). Among the Aka, an African group of hunter-gatherers, vertical transmission of many customs was estimated by Hewlett and Cavalli-Sforza (1986) to be strong. More recent studies suggest that the relative importance of vertical and horizontal learning is age dependent (Hewlett et al., 2011). Generalizations of the transmission system in

Table 1 Cultural transmission with assortment

Mating			Offspring probability[b]	
Mother	*Father*	*Frequency*[a]	*0*	*1*
0	0	$x_0^2 + mx_0x_1$	$b_{00,0}$	$b_{00,1}$
1	0	$x_0x_1(1-m)$	$b_{10,0}$	$b_{10,1}$
0	1	$x_0x_1(1-m)$	$b_{01,0}$	$b_{01,1}$
1	1	$x_1^2 + mx_0x_1$	$b_{11,0}$	$b_{11,1}$

[a]The subscript t for the parental generation is suppressed, $x_0 + x_1 = 1$.
[b]$b_{ij,0} + b_{ij,1} = 1, (i, j \in \{0, 1\})$
After Cavalli-Sforza, L.L., Feldman, M.W., 1981. Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press, Princeton, NJ.

Table 1 to allow for more than two cultural variants were developed by Feldman and Zhivotovsky (1992).

Evolution under long-continued vertical transmission should be conservative relative to that expected under other modes of transmission in the sense that change over generations is restricted by parent–offspring concordance. In particular, the more closely $\{\beta_{ij,k}\}$ resembles a genetic transmission system, the slower will be the evolution. In fact, an exactly genetic transmission structure, in the absence of other forces, results in no temporal change of allele frequencies. These points will be revisited in connection with continuous traits (Guglielmino et al., 1995).

Horizontal and Oblique Transmission

Suppose that after vertical transmission, the frequency of trait variant i in the population at time $t+1$ is $x^V_{i,t+1}$ and suppose that this particular variant, once acquired, cannot be lost. Then, by analogy with epidemic models of disease transmission, we suppose that individuals of type j (different from i) who did not acquire i by vertical transmission contact those who did and are converted to i at rate f_{ji}, the rate of horizontal transmission. It is usual to assume $0 < f_{ji} < 1$. Then, after this transmission, the new frequency of i $x^H_{i,\,t+1}$ can be written as

$$x^H_{i,t+1} = x^V_{i,t+1} + \sum_{j \neq i} x^V_{i,t+1}\left(1 - x^V_{i,t+1}\right) f_{ji} \qquad [2]$$

If the trait has only two variants 0 and 1, for example, then

$$x^H_{0,t+1} = x^V_{0,t+1} + f_{10} x^V_{0,t+1}\left(1 - x^V_{0,t+1}\right) \qquad [2a]$$

which makes the logistic nature of the dynamic obvious (Cavalli-Sforza and Feldman, 1981).

Within the population, horizontal transmission will generally result in more rapid evolution (Cavalli-Sforza and Feldman, 1981; Guglielmino et al., 1995), that is, faster change in the frequencies of trait variants, than under vertical transmission. Horizontal transmission modeled by eqns [2] and [2a] occurs between members of the same demographic cohort. If the transmitters are in generation t (or earlier) and are not the parents of the acquirers of the trait, then we have oblique transmission. In most cases, oblique transmission will have similar dynamic consequences to horizontal transmission, with a slight delay. The consequences of important special cases of oblique transmission for the rate of cultural evolution were investigated by Aoki et al. (2011) and Lehmann et al. (2011). They compared 'one-to-many' or teacher transmission, where one role model is chosen and spreads its trait to the rest of the population; 'many-to-one' transmission where multiple role models may be chosen by one offspring; 'conformist' transmission where learners disproportionately adopt the majority trait; and 'best-of-k' transmission where an offspring will sample k role models from the total population and choose among them. These special cases were shown to differ in both the rate of cultural evolution and the amount of cultural accumulation that results, with the pronovelty 'best-of-k' transmission producing rapid dynamics and the conservative conformist transmission producing slow dynamics.

An important case of horizontal transmission is often called 'cultural diffusion.' The term generally applies to the trajectory of a new cultural variant, that differs from the most frequent one in the population. It can apply to the spread of the new variant within a single population or, if the population consists of several subpopulations between which there is some communication, it can refer to the spread across the collection of subpopulations. A logistic dynamic is often observed for the change in frequency of the new variant as it diffuses (e.g., Rogers, 1995). An important question for many anthropologists is whether this diffusion may occur without migration of the people who carry the trait, and it is this case for which the term cultural diffusion is often reserved. On the other hand, if people move and carry their variant of the trait with them, then this cultural variant diffuses with its carriers, a process described as 'demic diffusion.' Distinction between these is usually difficult to make, but if genetic data are available that suggest a cline or gradient in allele frequencies from a putative source population of the human migrants (Cavalli-Sforza et al., 1993), then it might be reasonable to infer that demic diffusion has occurred. Genetic testing of populations before and after an event such as the advent of dairy farming can also help to distinguish the processes of demic and cultural diffusion. For example, Burger et al. (2007) tested one Mesolithic and eight Neolithic humans and found that prior to, and just following, the innovation of dairy farming practices, the lactase persistence allele was not at high frequency. This suggests that the cultural practice was transmitted to individuals who did not have the lactase persistence allele. This population responded to selection pressures increasing the frequency of the lactase persistence gene over time.

The empirical basis of the conversion parameters $\{f_{ji}\}$ in eqn [2] has been the subject of extensive sociological analysis. Rogers (1995) introduces the notion of awareness-knowledge, which is information that the new variant of the trait exists. Awareness may be followed by an adoption phase, in which the decision to adopt may depend critically on local history of the variant, on the education level or economic status of earlier adopters *vis-á-vis* the decision maker. Late adoption may also be affected by the perceived success of earlier adopters. Clearly, for the dynamics to be interpretable, simplification of the large number of parameters involved in these details is necessary. It is possible to model reinforcement through repeated transmission in both vertical and horizontal cases simply by iterating recursions eqn [1] or [2] a number of times within a generation and taking a census after this number of repeated operations. The rate of change across the new census points will be faster than possible from a single iteration per generation.

Group Effects and Conformity

The pressure on individuals within a population to adopt a belief or perform a task in a certain way may be affected by how many others in that population believe or act in that way. The result is that the rate of transmission of a trait, instead of being constant as in eqns [1] and [2], becomes a function of that trait's frequency (or possibly frequencies). That is, for the dichotomous trait with horizontal transmission, f_{10} in eqn [2a] might be an increasing function of $x^V_{0,t+1}$ when $x^V_{0,t+1}$ is large enough. This would amount to a pressure to conform to the

majority. Boyd and Richerson (1985: Chapter 7) point out that this can accelerate the rate at which a trait spreads once it reaches a threshold frequency. If the frequency-dependent transmission operates in different directions in different places, because of chance effects or environmental differences, this kind of conformity may also augment geographical differences between isolated groups. Applications of such conformity to economic theory in terms of norms and preferences that might be culturally transmitted entities have been initiated by Bowles (1998).

Innovation

In biological evolution, the ultimate source of all variation is mutation, nondirected random changes in the DNA that may or, more often, may not result in functional change in the organism. The equivalent for a cultural trait is the appearance of a new cultural variant, an innovation. The fate of a cultural innovation within a population depends on its rules of transmission, which in turn may depend on a host of structural features of the trait itself, such as the interaction of the innovation with other traits or their rules of transmission. In addition, the general environment in which the innovation occurs may decide its fate; for example, stylistic variation in tool manufacture or behaviors that are coerced economically or by direct force would fall into this category. Finally, chance factors that result from the population size may determine whether and how fast an innovation succeeds.

In terms of the transmission law of **Table 1**, if $b_{00,1} > 0$ or $b_{11,0} > 0$ innovation has occurred in the offspring of the first or fourth matings, respectively. On a more general scale, an analogue to the infinite sites models of population genetic theory (Ewens, 1979: Chapter 3) is more appropriate to describe the stochastic dynamics of innovations, an idea that is explored in detail by Aoki et al. (2011).

Continuous Traits and Linear Gaussian Models

If the trait under study varies continuously, then the transmission scheme is more difficult to represent. It is usual to represent the values of an offspring's trait as X_{t+1} and to specify it as a weighted average of two randomly chosen parental values augmented by a random term:

$$X_{t+1} = b_m X_{m,t} + b_p X_{p,t} + \varepsilon \qquad [3]$$

where b_m and b_p are maternal and paternal weights, and ε is an error random variable with zero mean and variance σ^2. The recursion given by eqn [3] corresponds to vertical transmission, and it is easy to include correlation between the parents (i.e., assortative mating) as well as a kind of oblique transmission. The latter may be represented as a group effect g:

$$X_{t+1} = (1-g)\left[b_m X_{m,t} + b_p X_{p,t}\right] + g\overline{X}_t + \varepsilon \qquad [4]$$

where $\overline{X}_t$ is the mean of the trait at generation t (Cavalli-Sforza and Feldman, 1973b).

Several features of the dynamics resulting from models described by eqns [3] and [4] are worth noting. The existence of a group effect constrains the variance within the population (Cavalli-Sforza and Feldman, 1973b). If a number of groups are considered, however, within each of which transmission occurs as in eqn [4], then the variance, V_t, of the average across groups relative to the original group mean increases linearly with time: $V_t = kt\sigma^2$, where k depends on g, b_m, and b_f.

Linear models analogous to eqn [3] or [4] allow correlations between relatives to be computed rather simply, especially under Gaussian assumptions for the trait distribution. Indeed, the most widely used framework for studying familial aggregation of quantitative characters represents the genetic contribution to the character as a linear factor complementary to a linear environmental factor. With appropriate assumptions on the genetic term, variances and covariances between relatives can be computed that mimic those calculated under the simplest genetic hypotheses (Fisher, 1918). Controversy remains, however, about which joint linear model of genetically transmitted, culturally transmitted, and nontransmitted contributions to traits that aggregate within families is appropriate (Feldman and Otto, 1997), while few attempts to use nonlinear representations have been made.

Natural and Cultural Selection

Most intuition about the effects of natural selection on genotypes comes from an interpretation of fitness as a relative probability of survival from fertilized egg to adult. However, a considerable body of experimental evidence (reviewed by Lewontin, 1974: Chapter 5) suggests that differential fertility, sexual selection, and frequency-dependent effects are more important than differential survival although the former are more difficult to measure experimentally and to analyze theoretically. Despite these difficulties, it is reasonable to view natural selection on biological variation in terms of differential reproduction and survival. More details on the components of selection may be found in Hartl and Clark (1997: Chapter 6).

For cultural variants, natural selection on a trait may be understood in the same Darwinian way. To the extent that the variants of a trait affect the reproduction or survival of their carriers differentially, the frequencies of the variants will change over time accordingly. By contrast, cultural selection refers to differences in the rates of transmission of the cultural variants. As examples, such differences may be manifest in the rates at which individuals become aware of the alternatives or in the decision processes leading to the adoption of one or other of the alternatives. The results of cultural selection will be changes in the frequencies of the cultural variants in a population over time. One obvious difference from natural selection is that cultural selection may act within a generation, rather like an epidemic (Rogers, 1995). The parameters that summarize the process of cultural selection are the vertical, horizontal, or oblique transmission rates described above.

One area of theoretical debate concerns the possibility of conflicts between cultural and natural selection acting on the same variant of a trait. Thus, for the human cultural trait of fertility, the variant whose carriers adopt fertility control (e.g., via contraception) is at a clear disadvantage in natural selection but has spread, especially in developed countries, by virtue of its advantage in cultural selection. Other examples are

discussed by Cavalli-Sforza and Feldman (1981), Boyd and Richerson (1985), and Durham (1991). The debate, which cannot be settled with standard experimentation, centers on whether natural and cultural selection act in the same direction most of the time. Alternatively, did the process of natural selection during human evolution shape cultural transmission in such a way that those variants favored by cultural selection are most likely to be adaptive in the Darwinian sense?

A second contentious topic in debates about selection on a cultural trait concerns the role of group selection in the maintenance of variants, which, within a group, are individually disadvantageous. Two kinds of models have been considered; in one, there is extinction and recolonization of groups according to a stochastic process. If the probability of extinction of a group decreases with increasing frequency of the individually disadvantageous trait, then competition between the groups may preserve the disadvantageous trait. The usual example used for such analyses is altruism (e.g., Boyd and Richerson, 1985: Chapter 7; Uyenoyama, 1979). Maintenance of such a trait by extinction-type group selection is much less likely than by graded group selection, in which extinction does not occur but the expected relative contribution of one deme to the whole set of demes increases with the frequency of the altruist in the former (Uyenoyama, 1979). In the latter formulation, after reproduction, there is redistribution of individuals to subgroups according to a stochastic process whose outcomes depend on the numbers of the individually disadvantageous types (Cohen and Eshel, 1976). In other models, fluctuating environments in each of the subpopulations have been shown to allow group selection to occur under more general mixing conditions (Uyenoyama, 1979). As pointed out by a number of authors, there are great similarities between these theories of group selection, and the shifting balance theory of genetic evolution due to Sewall Wright (1931, 1935, 1948). There is a recent trend toward accepting the efficacy of an appropriately formulated model of graded group selection in maintaining traits that are individually disadvantageous, whether culturally or genetically transmitted. In particular, frequency-dependent cultural transmission, for example, of the conformist type, may amplify differences between groups, thereby increasing the strength of group selection relative to that of natural selection within groups (Boyd and Richerson, 1985; Henrich, 2004).

Some Metatheoretical Considerations

Four main theoretical frameworks have formed the basis for the study of behavioral evolution and, in particular, human behavioral evolution. The first is kin selection theory, the second is evolutionary game theory, the third is gene–culture coevolutionary theory, and the fourth is cultural niche construction.

Kin selection theory, due to W.D. Hamilton (1964), is based on the idea that genetic alleles whose effect is disadvantageous to their carriers may be preserved if a carrier acts to increase the survival rate of others in the population. Other carriers will most likely be relatives of the first carrier, hence the name 'kin selection.' This can result in the spread of the disadvantageous allele if the genetic relationship between the donor and the recipient of the altruism is close enough. By the mid-1970s, this theory had been extended beyond its original insect and animal boundaries to include human behaviors. In this process, the position that a wide array of behaviors would generally be genetically determined and adaptive came to be called sociobiology (Wilson, 1975). The idea that behaviors such as altruism, or preferences for mates, or certain norms or values, have a biological basis, extended sociobiology and became known as evolutionary psychology (Tooby and Cosmides, 1990).

At about the same time as the theory of kin selection originated, the application of game theory, the second framework, to animal and human behaviors also became widespread. Here, in pairwise interactions, for example, each individual was thought to optimize its behavior relative to the alternatives the other could adopt. The notion of optimization became prevalent in discussions of behavioral evolution (see, e.g., Parker and Maynard Smith, 1990). Such treatments did not consider evolutionary dynamics as processes of frequency change but as processes of maximization of functions chosen by the investigator, usually to represent a measure of fitness.

Although traits such as altruism may well have a culturally transmitted basis, they have generally been studied in the theoretical literature using genetic models. Gene–culture coevolutionary theory was introduced as an alternative to the two former frameworks by Feldman and Cavalli-Sforza (1976; see also Feldman et al., 1985). By allowing each individual a phenogenotype, the investigator is free to incorporate transmission mechanisms, selection functions, and dynamics that span the spectrum from purely genetic to purely cultural. Besides the obvious gain in flexibility to describe the evolution of behaviors, this framework also permits a natural statistical analysis of correlations between relatives *á la* Fisher (1918). The advantage here is that nongenetic effects are modeled more realistically than is possible using standard linear regression theory (Feldman et al., 2013). In the coevolutionary dynamics with vertical cultural transmission, an association measure between the cultural and genetic variants is important and the order of magnitude of this association as a function of the comprehensive transmission rule can be derived. More details on the connections between these theories and the following examples can be found in Feldman and Laland (1996).

Gene–Culture Evolutionary Theory

Gene–culture coevolutionary models assume that traits acquired through cultural learning can exert selection pressure on genes. This is typically modeled by assuming that transmission of a cultural trait is vertical and by describing the probability of mating between parents of different phenogenotypes, and the probability that the resultant offspring will possess traits acquired from their parents. Each combination of genes and cultural traits may have a different fitness value. A well-studied example of such gene–culture interaction concerns the evolution of the gene for lactase

persistence after the emergence of dairy farming described mathematically by Aoki (1986). Here the gene in question is an autosomal dominant with lactose absorption the dominant phenotype at this locus. The cultural dichotomy is the presence or absence of a history of dairy farming, which can be realized as the behaviors of drinking or not drinking milk in adulthood. The frequency of lactose absorbers in populations with dairy traditions reaches over 90% in Northern Europe, while in those populations without such a tradition, for example, the Pima of the Americas or the Yoruba of Nigeria, it is typically below 20%. In modeling the joint evolution of the gene and the custom of dairying, both natural selection due to increased nutrition provided by milk and the cultural transmission of milk use are affected by the genotype at the lactose-absorbing locus.

In Aoki's model, we can label the genetic locus **A** with alleles A and a representing lactose absorption and non-absorption, respectively, and label the cultural practice of dairy farming **L** with cultural trait types L and l representing dairy farming and non-dairy farming, respectively. If A is dominant, an individual with at least one copy of the A allele is a lactose absorber. Mating between genotypes is random and the expected frequencies of the phenogenotypes are given in **Table 2**. In **Table 2**, p is the frequency of the A allele, $q = 1 - p$ is the frequency of the a allele, and s is the fitness cost incurred by not drinking milk or drinking milk and failing to absorb lactose. The probability of becoming a dairy farmer through cultural transmission of the trait is given by $f(y)$ for individuals with an A allele and by $g(y)$ for individuals with the aa genotype, where y represents the frequency of the dairy farming trait in the adult population. Thus, the expected frequencies of offspring phenogenotypes are proportional to the product of the gene frequency, the frequency of the cultural trait in the population, and the relative fitness of the phenogenotye.

Analysis of this model suggests that the fate of the absorption allele critically depends on the rate of vertical transmission of milk use. If this rate is very high, then a significant fitness advantage to absorbers can shift the frequency of the dominant allele to high frequencies in as few as 300 generations (see also Feldman and Cavalli-Sforza, 1989).

The spread of agriculture into geographical areas previously occupied by hunter-gatherers has been a fruitful subject for students of genetic and cultural evolution. Aoki and Shida (1996) used a reaction–diffusion process to model the spread of farming. Their analysis included a horizontal transmission process converting hunter-gatherers to farming as well as different intrinsic growth rates for initial farmers, converted farmers, and hunter-gatherers. This analysis produced conditions under which farming would advance at a constant rate and, in addition, a gene frequency cline would simultaneously form. This study significantly extended the earlier analysis by Ammerman and Cavalli-Sforza (1971), which was based on R.A. Fisher's famous model for the advance of an advantageous gene (Fisher, 1937).

A third example of gene–culture coevolution concerns the well-documented cultural bias against daughters in South and East Asia (Das Gupta, 1987; Tuljpurkar et al., 1995). Under the assumption that a preference for sons is vertically transmitted, Kumm et al. (1994) studied how this transmission might affect the evolution of sex-determining genes. They showed that if parents act to increase the proportion of sons, irrespective of the biologically set sex ratio among their children, and if they compensate for lost daughters by having further children, the sex ratio would evolve to be female-biased. On the other hand, if parents try to achieve a desired sex ratio among their children and in so doing have fewer children than impartial parents, the primary sex ratio will become male-biased.

Table 2 Gene–culture model for lactose absorption

	L	l
AA	$p^2 f(y)$	$p^2[1 - f(y)](1 - s)$
Aa	$2pqf(y)(1 - s)$	$2pq[1 - f(y)](1 - s)$
aa	$q^2 g(y)(1 - s)$	$q^2[1 - g(y)](1 - s)$

Aoki, K., 1986. A stochastic model of gene–culture coevolution suggested by the 'culture historical hypothesis' for the evolution of adult lactose absorption in humans. Proceedings of the National Academy of Sciences of the United States of America 83, 2929–2933.

Cultural Niche Construction

Gene–culture coevolutionary theory generally invokes fixed parameters to describe the effects of a cultural trait on natural selection pressures and the frequency of that trait in the population (Laland et al., 2001). Cultural niche construction models allow for more general situations (Ihara and Feldman, 2004; Kendal et al., 2005; Odling-Smee et al., 2003; Laland et al., 2001). A strong example of a less straightforward gene–culture relationship is the prevalence of the *HbS* allele among populations of yam farmers in West Africa (Durham, 1991). In this case, farmers cleared areas of forest for yam cultivation, increasing standing water, which acted as a breeding ground for mosquitoes. The mosquitoes in turn increase the spread of malaria and thus strengthen selection for the *HbS* sickle cell allele, which confers some resistance to malaria in heterozygotes. To model such systems, one must take account of the frequency of the cultural trait, the effect of the trait on the environment, noncultural effects on the environment, and the effect of the environment on the spread of genetic and cultural traits. In the *HbS* example, the cultural trait of forest clearing formed a niche for the spread of malaria. Ihara and Feldman (2004) made the first model of purely cultural niche construction in which one cultural trait (e.g., education of women) forms a 'cultural niche' in which another culturally transmitted trait (e.g., preference for controlled fertility) may evolve. They showed that this kind of context dependence for cultural transmission can have profound effects on the dynamics of a cultural trait, and the demography of the population.

Simulation Approaches

Advances in the availability and power of computational resources including high-level agent-based modeling software packages have made the simulation studies of cultural evolution, gene–culture coevolution, and cultural niche

construction easier and more productive. It is now possible to use simulation to construct both agent-based models of cultural evolution and to use computationally intensive numerical methods to gain insight into models of cultural evolution that may be too complex to solve analytically.

Agent-based simulations have been extended to include spatial and demographic components that were previously difficult or impossible to investigate. One notable recent example is the Social Learning Strategies Tournament (SLST), which ran in 2010 and was designed by Luke Rendell and Kevin Laland at the University of St. Andrews (Rendell et al., 2010b). The tournament followed the example of Axelrod (1984) who ran the famous cooperation tournament. However, in the absence of an appropriately flexible social learning 'game,' the SLST used a multiarmed bandit structure. Participants were invited to suggest social learning strategies that would maximize the reproductive fitness of the individuals using their strategy. The winner out of 114 entries was Discount Machine, a strategy that relied heavily on social learning, innovating only when it was at the very start of a simulation round, and had no available role models to copy. Questions about the effect of these different social learning strategies on the characteristics of the cultures that emerged in the simulations were answered specifically by Rendell et al. (2011a). Smaller scale simulations have attempted to address similar questions, comparing a small number of social learning strategies in detail and assessing their comparative efficacy in an evolutionary framework (e.g., Kameda and Nakanishi, 2002; Rendell et al., 2010a), or testing typical assumptions, such as the faster rate of cultural evolution compared to genetic evolution, made in many analytical and simulation models (Franz and Nunn, 2009).

Geographic effects on models of cultural evolution and cultural niche construction have also been shown through simulation to be important. For example, Rendell et al. (2010a) found that spatial clustering allowed for the evolution of social learning even when this lowered the average fitness of the population, and following a simulation framework proposed by Silver and Di Paolo (2006), Rendell et al. (2011b) showed that spatial structure allowed for runaway cultural niche construction and secondary hitchhiking of costly traits where they would otherwise not occur. Agent-based simulations in this vein tend to invoke a grid of populations or individuals with wraparound boundary conditions. A different approach was taken by Shennan (2001) who, again borrowing from the literature on genetic evolution, assumed a simple vertical learning rule and investigated the effects of innovation and selection on the evolution of culture, an approach extended by Powell et al. (2009) to include the effects of demography and spatial structure.

Extensive computation of deterministic systems can also be used to explore the parameter space of complex transmission models. For example, Creanza et al. (2012) investigated over 25 million parameter combinations of transmission parameters and levels of cultural assortative mating in their model of cultural niche construction and found rare cultural polymorphisms that would have been otherwise difficult to find.

See also: Cultural Evolution: Overview; Cultural Evolution: Phylogeny versus Reticulation; Culture, Sociology of; Evolution: Diffusion of Innovations; Gene–Culture Coevolution; Kin Selection; Social Learning and Culture in Nonhuman Organisms; Technology, Evolution of.

Bibliography

Ammerman, A.J., Cavalli-Sforza, L.L., 1971. Measuring the rate of spread of early farming in Europe. Man 6, 674–688.

Aoki, K., 1986. A stochastic model of gene–culture coevolution suggested by the 'culture historical hypothesis' for the evolution of adult lactose absorption in humans. Proceedings of the National Academy of Sciences of the United States of America 83, 2929–2933.

Aoki, K., Shida, M., 1996. Travelling wave solutions for the spread of farmers into a region occupied by hunter-gatherers. Theoretical Population Biology 50, 1–17.

Aoki, K., Lehmann, L., Feldman, M.W., 2011. Rates of cultural change and patterns of cultural accumulation in stochastic models of social transmission. Theoretical Population Biology 79 (4), 192–202.

Atran, S., 2001. The trouble with memes: inference versus imitation in cultural creation. Human Nature 12, 351–381.

Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York.

Blackmore, S., 1999. The Meme Machine. Oxford University Press, Oxford.

Bowles, S., 1998. Endogenous preferences: the cultural consequences of markets and other economic institutions. Journal of Economic Literature 36, 75–111.

Boyd, R., Richerson, P.J., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago.

Burger, J., Kirchner, M., Bramanti, B., Haak, W., Thomas, M.G., 2007. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. Proceedings of the National Academy of Sciences of the United States of America 104 (10), 3736–3741.

Cavalli-Sforza, L.L., Feldman, M.W., 1973a. Cultural versus biological inheritance: phenotypic transmission from parents to children (a theory of the effect of parental phenotypes on children's phenotypes). The American Journal of Human Genetics 25, 618–637.

Cavalli-Sforza, L., Feldman, M.W., 1973b. Models for cultural inheritance I. Group mean and within group variation. Theoretical Population Biology 4 (1), 42–55.

Cavalli-Sforza, L.L., Feldman, M.W., 1981. Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press, Princeton, NJ.

Cavalli-Sforza, L.L., Feldman, M.W., Chen, K.H., Dornbusch, S.M., 1982. Theory and observation in cultural transmission. Science 218, 19–27.

Cavalli-Sforza, L.L., Menozzi, P., Piazza, A., 1993. Demic expansions and human evolution. Science 259, 639–646.

Cohen, D., Eshel, I., 1976. On the founder effect and the evolution of altruistic traits. Theoretical Population Biology 10, 27–302.

Creanza, N., Fogarty, L., Feldman, M.W., 2012. Models of cultural niche construction with selection and assortative mating. PLoS ONE 7 (8), 42744.

Das Gupta, M., 1987. Selective discrimination against female children in rural Punjab, India. Population and Development Review 13 (1), 77–100.

Dawkins, R., 1976. The Selfish Gene. Oxford University Press, New York.

Durham, W.H., 1991. Coevolution: Genes, Culture and Human Diversity. Stanford University Press, California.

Ewens, W.J., 1979. Mathematical Population Genetics. Springer-Verlag, Berlin.

Feldman, M.W., Cavalli-Sforza, L.L., 1976. Cultural and biological evolutionary processes, selection for a trait under complex transmission. Theoretical Population Biology 9, 239–259.

Feldman, M.W., Cavalli-Sforza, L.L., 1979. Aspects of variance and covariance analysis with cultural inheritance. Theoretical Population Biology 15, 276–307.

Feldman, M.W., Cavalli-Sforza, L.L., 1989. On the theory of evolution under genetic and cultural transmission with application to the lactose absorption problem. In: Feldman, M.W. (Ed.), Mathematical Evolutionary Theory. Princeton University Press, Princeton, pp. 145–173.

Feldman, M.W., Laland, K.N., 1996. Gene–culture coevolutionary theory. Trends in Ecology and Evolution 11, 453–457.

Feldman, M.W., Otto, S.P., 1997. Heritability from twins and other relatives. Science 278, 1383–1384.

Feldman, M.W., Cavalli-Sforza, L.L., Peck, J.R., 1985. Gene-culture coevolution: models for the evolution of altruism with cultural transmission. Proceedings of the National Academy of Sciences 82 (17), 5814–5818.

Feldman, M.W., Christiansen, F.B., Otto, S.P., 2013. Gene-culture co-evolution: teaching, learning, and correlations between relatives. Israel Journal of Ecology and Evolution 59, 72–91.
Feldman, M.W., Zhivotovsky, L.A., 1992. Gene-culture coevolution: toward a general theory of vertical transmission. Proceedings of the National Academy of Sciences 89 (24), 11935–11938.
Fisher, R.A., 1918. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh 52, 399–433.
Fisher, R.A., 1937. The wave of advance of advantageous genes. Annals of Eugenics 7, 355–369.
Fogarty, L., Feldman, M.W., 2011. The cultural and demographic evolution of son preference and marriage type in contemporary China. Biological Theory 6 (3), 272–282.
Fogarty, L., Creanza, N., Feldman, M.W., 2013. The role of cultural transmission in human demographic change: an age-structured model. Theoretical Population Biology 88, 68–77.
Franz, M., Nunn, C.L., 2009. Rapid evolution of social learning. Journal of Evolutionary Biology 22 (9), 1914–1922.
Guglielmino, C.R., et al., 1995. Cultural variation in Africa: role of mechanisms of transmission and adaptation. Proceedings of the National Academy of Sciences of the United States of America 92 (16), 7585–7589.
Hamilton, W.D., 1964. The genetical evolution of social behavior: parts I and II. Journal of Theoretical Biology 7 (1–16), 17–52.
Hartl, D.L., Clark, A.G., 1997. Principles of Population Genetics. Sinauer Associates, Sunderland, MA.
Hauert, Ch., 2001. Fundamental clusters in spatial 2×2 games. Proceedings of the Royal Society of London B 268, 761–769.
Henrich, J., 2004. Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses: the Tasmanian case. American Antiquity 69 (2), 197–214.
Henrich, J., Boyd, R., 2002. On modeling cognition and culture: why replicators are not necessary for cultural evolution. Journal of Cognition and Culture 2 (2), 87–112.
Hewlett, B.S., Cavalli-Sforza, L.L., 1986. Cultural transmission among the aka pygmies. American Anthropologist 88, 922–934.
Hewlett, B.S., et al., 2011. Social learning among Congo Basin hunter-gatherers. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 366 (1567), 1168–1178.
Ihara, Y., Feldman, M., 2004. Cultural niche construction and the evolution of small family size. Theoretical Population Biology 65 (1), 105–111.
Kameda, T., Nakanishi, D., 2002. Cost–benefit analysis of social/cultural learning in a non-stationary uncertain environment: an evolutionary simulation and an experiment with human subjects. Evolution and Human Behavior 23, 373–393.
Kandler, A., Steele, J., 2009. Innovation diffusion in time and space: effects of social information and income inequality. Diffusion Fundamentals 11, 117.
Kendal, J.R., Ihara, Y., Feldman, M.W., 2005. Cultural Niche Construction with Application to Fertility Control: A Model for Education and Social Transmission of Contraceptive Use. Morrison Institute Working Paper Series 102.
Kumm, J., Laland, K.N., Feldman, M.W., 1994. Culturally transmitted prejudices for offspring gender and the evolution of primary sex ratios. Theoretical Population Biology 46, 249–781.
Laland, K., Odling-Smee, J., Feldman, M.W., 2001. Cultural niche construction and human evolution. Journal of Evolutionary Biology 14, 22–33.
Lehmann, L., Aoki, K., Feldman, M.W., 2011. On the number of independent cultural traits carried by individuals and populations. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 366 (1563), 424–435.
Lewontin, R.C., 1974. The Genetic Basis of Evolutionary Change. Columbia University Press, New York.
Li, N., Feldman, M.W., Li, S., 2000. Cultural transmission in a demographic study of sex ratio at birth in China's future. Theoretical Population Biology 58 (2), 161–172.
Mace, R., Holden, C.J., 2005. A phylogenetic approach to cultural evolution. Trends in Ecology and Evolution 20, 116–121.
Maynard, S.J., 1982. Evolution and the Theory of Games. Cambridge University Press.
Maynard Smith, J., 6 November, 1986. Natural selection of culture? New York Review of Books 33.
Mesoudi, A., 2011. Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences. University of Chicago Press, Chicago.
Nowak, M.A., May, R.M., 1992. Evolutionary games and spatial chaos. Nature 359, 826–829.
Odling-Smee, F.J., Laland, K.N., Feldman, M.W., 2003. Niche Construction: The Neglected Process in Evolution. In: Monographs in Population Biology, vol. 37. Princeton University Press, Princeton, NJ.
O'Brien, M.J., Darwent, J., Lyman, R.L., 2001. Cladistics is useful for reconstructing archaeological phylogenies: Palaeoindian points from the Southeastern United States. Journal of Archaeological Science 28 (10), 1115–1136.
Parker, G.A., Maynard Smith, J., 1990. Optimality theory in evolutionary biology. Nature 348, 27–33.
Powell, A., Shennan, S., Thomas, M.G., 2009. Late pleistocene demography and the appearance of modern human behavior. Science (New York) 324 (5932), 1298–1301.
Rendell, L., Fogarty, L., Laland, K.N., 2010a. Rogers' paradox recast and resolved: population structure and the evolution of social learning strategies. Evolution; International Journal of Organic Evolution 64 (2), 534–548.
Rendell, L., et al., 2010b. Why copy others? Insights from the social learning strategies tournament. Science (New York) 328 (5975), 208–213.
Rendell, L., et al., 2011a. How copying affects the amount, evenness and persistence of cultural knowledge: insights from the social learning strategies tournament. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 366 (1567), 1118–1128.
Rendell, L., Fogarty, L., Laland, K.N., 2011b. Runaway cultural niche construction. Philosophical Transactions of the Royal Society of London. Series B. Biological Sciences 366 (1566), 823–835.
Rogers, E.M., 1995. Diffusion of Innovations, fourth ed. The Free Press, New York.
Shennan, S., 2001. Demography and cultural innovation: a model and its implications for the emergence of modern human culture. Cambridge Archaeological Journal 11 (01).
Silver, M., Di Paolo, E., 2006. Spatial effects favour the evolution of niche construction. Theoretical Population Biology 70 (4), 387–400.
Sperber, D., 1996. Explaining Culture: A Naturalistic Approach. Blackwell, Oxford, UK and Cambridge, Mass.
Tooby, J., Cosmides, L., 1990. The past explains the present: emotional adaptations and the structure of ancestral environments. Ethology and Sociobiology 11 (4), 375–424.
Tuljpurkar, S., Nan, L., Feldman, M.W., 1995. High sex ratios in China's future. Science 267, 874–876.
Uyenoyama, M.K., 1979. Evolution of altruism under group selection in large and small populations in fluctuating environments. Theoretical Population Biology 15, 58–85.
Watts, D.J., 1999. Small Worlds: The Dynamics of Networks between Order and Randomness. Princeton University Press, Princeton, NJ.
Wilson, E.O., 1975. Sociobiology. Harvard University Press, Cambridge, UK.
Wright, S., 1931. Evolution in Mendelian populations. Genetics 16, 97–159.
Wright, S., 1935. Evolution in populations in approximate equilibrium. Journal of Genetics 30, 257–266.
Wright, S., 1948. On the roles of directed and random change in gene frequency in the genetics of populations. Evolution 2, 279–294.

Cultural Geography

Kevin Hannam, International Centre for Research in Events, Tourism and Hospitality, Carnegie Faculty, Leeds Beckett University, Leeds, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Pratt, volume 5, pp. 3070–3075, © 2001, Elsevier Ltd.

Abstract

Early conceptualizations of cultural geography by the Berkeley School tended to focus on the morphology of cultural landscapes. Cultural geography has since developed from being a subdiscipline into a dominant critical perspective within the social sciences. Developments in cultural geography in the 1990s focused on questioning various cultural representations (landscape as text). Subsequent work questioned the theoretical and political status of 'culture,' with the effect of blurring taken-for-granted distinctions between culture, nature, and economy. More recently, cultural geography has been developed further through an engagement with 'more-than representational' thinking and the 'mobilities' paradigm.

Cultural geography as a subdisciplinary category has been credited as being "perhaps the most ambiguous term in the discipline's lexicon" (Price and Lewis, 1993), for two reasons. Firstly, the term 'culture' is often not defined; there is simply an "easy coupling with geography in which both refer to a general sensitivity to issues of context, difference, and the local" (Barnett, 1998: 631). Secondly, there has been a proliferation of cultural geographical approaches since the 1980s: "the revival of cultural geography in a radically new form was one of the most striking [disciplinary] developments of the 1980s" (Gregory, 1994: 98).

Cultural geography has developed from being a subdiscipline into a dominant critical perspective within the social sciences. The Berkeley School, with its focus on the relationships between cultural groups and natural environments, dominated Anglo-American cultural geography until the 1980s. Developments in cultural geography in the 1990s focused on various cultural representations (landscape as text, or way of seeing) and on a variety of ways that cultural identities and places are intertwined. This work questioned the theoretical and political status of 'culture,' with the effect of blurring taken-for-granted distinctions between culture, nature, and economy. More recently, this work has been developed further through an engagement with 'ways of telling' in terms of a more narrative approach to landscape (Daniels and Lorimer, 2012) as well as through 'more-than representational' thinking and the emerging 'mobilities' paradigm (Sheller and Urry, 2006; Adey, 2009).

The Berkeley School

The term 'cultural geography' can be traced to nineteenth-century German scholarship, but is thought to have been introduced into American geography in the 1920s by Carl Sauer (Price and Lewis, 1993). Sauer, his colleagues in the geography department at Berkeley, and their students dominated American cultural geography until the 1980s. (This is not to say that this was the only strand of 'traditional' cultural geography. For an overview of a broader range of traditional cultural geographies, see Foote et al., 1994.) Sauer (1925) framed cultural geography in opposition to environmental determinism, which he criticized for subjecting the diversity of cultures to a monocausal, nomothetic theory. Sauer instead advocated historical synthetic accounts of human interaction with natural environments. For Sauer, landscape study was an exercise in historical reconstruction that sought to show how a particular cultural group, working on and through the natural landscape, enacted a cultural landscape. The relevant unit of observation was the cultural region, defined as an area over which a functionally coherent way of life dominated, echoing the French geographer Vidal de la Blache's perspectives on *genres de vie* (ways of life). Members of the Berkeley School explored a range of themes in mostly rural locations: the migration of human groups and the process of adaption of a familiar culture to a new land; the domestication of plants and animals; the development, intensification, and dispersion of agriculture; ancient cultivation systems; and the genesis and diffusion of cultural traits and material culture such as folk housing (for a review see Mathewson, 1998).

In the early 1980s, as new approaches to cultural geography were developed, a (contested) critique of the Berkeley School emerged. Berkeley School cultural geography was charged with conceiving cultural groups and regions as unities, and deploying a superorganic conceptualization of culture that invested agency in culture rather than individuals and ignored power relations (Shurmer-Smith and Hannam, 1994). However, there has been a renewed appreciation of the cultural ecology tradition that grew out of the Berkeley School (Willems-Braun, 1997). Crang (1998: 16) has drawn attention to the ways in which Sauer bound the material and social, viewing crop species and gene pools, for example, as the "material expression and embodiment of social processes and knowledge."

New Cultural Geographies

In the 1990s a new generation of cultural geographers began to engage with a diverse range of philosophical traditions, and social and cultural theories from cultural studies, anthropology, and sociology such that the subdisciplinary – indeed disciplinary – boundaries of cultural geography became blurred. The so-called 'linguistic' turn was felt throughout the social

sciences and humanities; in cultural geography it focused attention on the landscape as text and the politics of the representations that geographers produce with the publication of Peter Jackson's (1989) influential text *Maps of Meaning* and the subsequent establishment of new academic journals such as *Ecumene* which would later evolve into *Cultural Geographies* (Crang and Mitchell, 2002). As Crang and Mitchell (2000: 2) argued, this journal would illustrate research aimed at "the wider deployment and development of cultural geographic imaginations: as part of returns to the 'traditional' humanities – history, literature studies, art criticism and philosophy – and through the emergence of newer interdisciplinary fields of cultural, media, queer, postcolonial, gender, environmental, urban and science studies. It is these transdisciplinary shifts that are placing cultural geographies at the centre of a more extensive intellectual landscape." They then went on to argue that *Cultural Geographies* sought to map out "the intellectual field of cultural geography not in terms of opposing camps – such as 'new' cultural geography and 'old' cultural geography – but by bringing together concerns over the cultural geographies of knowledge, landscape, nature and environment, and space and place" (Crang and Mitchell, 2002: 1). This has been acknowledged as helping to theoretically invigorate cultural geography in the US (Olwig, 2010). It also heralded an emphasis in innovative qualitative methods in cultural geography (Shurmer-Smith, 2003, DeLyser and Rogers, 2010). Nevertheless, a continuing theme has been in terms of mapping and reading cultural landscapes in variously nuanced ways (Mathewson, 1998, 1999; Mitchell, 2002; Crang, 2003; Della Dora, 2009).

There has been a long tradition of 'reading' the landscape in cultural geography. Peirce Lewis (1979: 12), for example, described the landscape as 'our unwitting autobiography.' Since the 1980s, however, the metaphor of landscape as text has been pursued more rigorously through a fuller engagement with literary and cultural theory; for example, in Duncan's (1990) now-classic study of the Kandyan Kingdom in early nineteenth-century Sri Lanka, he describes how the king of Kandy concretized two intertwined discourses on kingship through a massive city-building program, in an attempt to secure his political power; this landscape transformation was then interpreted differently by the king, nobles, and peasants, through the lens of the two key texts. The story, then, is of intertextuality, of the interplay of discourses enacted in landscape and texts.

This textualization of landscape has itself been criticized, less in relation to Duncan's specific empirical application than as a general theory of landscape interpretation. It has been criticized for erasing process (Gregory, 1994), for overemphasizing the coherence of texts and landscapes, and for suppressing traces of nonhuman others: "it treats the landscape as a blank page that only human actors can read and write upon" (Demeritt, 1994: 170). Nevertheless, the metaphor has been reworked, for example, around the notion of theater and the script, to draw out the open, performative possibilities of text – partly structured and partly improvised. Gregory (1999), for example, analyzed the scripting of Egypt in the nineteenth century by a growing tourism industry. European tourists not only read voluminously when traveling down the Nile, the guidebooks also provided stage directions for transforming *dahabeeah* (floating barges) into 'secure viewing platforms' and the remains of ancient Egypt were literally and materially staged as an 'extended exhibition.' It is this attention to the materiality and spatiality of writing, in this case travel writing, that distinguishes the work of cultural geographers (Duncan and Gregory, 1999).

The humanistic geographers of the early 1980s were drawn to the textual metaphor; Cosgrove (1985) entreated them to pay closer attention to the visual, as a way of understanding one of their key, though untheorized, concepts, that of landscape. In Gregory's (1994: 98) assessment, this recognition of the 'conceptuality' of landscape was one of the 'cardinal achievements' of the revival of cultural geography in the 1980s. Cosgrove traced the emergence of landscape as a visual ideology, as a way of seeing, in fifteenth- and sixteenth-century Europe. The discovery of linear perspective enabled a realist illusion of space: "Landscape is thus a way of seeing, a composition and structuring of the world so that it may be appropriated by a detached individual spectator to whom an illusion of order and control is offered through the composition of space according to the certainties of geometry" (Cosgrove, 1985: 55). This way of seeing coincided with and supported a transition from feudalism to capitalism, and new social relations with nature, and land as property. What landscape paintings achieve aesthetically "maps, surveys, and ordnance charts achieve practically" (Cosgrove, 1985: 55); this recognition led to a critical reassessment of another of geographers' representational forms: the map. The assumed link between reality and representation has been broken, and maps are now read as "mechanisms for defining social relationships, sustaining social rules, and strengthening social values" (Harley, 1992: 237), as 'technologies of power' and as 'performances' (Pinder, 2007: 459). Kitchin and Dodge (2007: 331) take this a stage further and urge us to "rethink cartography as ontogenetic in nature; that is maps *emerge* through practices" such that "mapping is a process of constant reterritorialization."

Issues of power, politics, and performances have thus become central concerns for contemporary cultural geographers (Jackson, 1989; Shurmer-Smith and Hannam, 1994). Landscapes not only express social relations, they are also an important means of enacting them. And landscapes, like maps, are such effective technologies of power because they tend to naturalize these relations (Cresswell, 1996; Mitchell, 1996). Critical attention has also been given to the various cultural geographies that constitute colonial relations, in the past and in the 'colonial present' as well as with issues of racism, nationalism, identity, location, belonging, diaspora, and memory (Sidaway, 2000; Nash, 2002, 2003; Price, 2010; Tolia-Kelly, 2010).

Developing from this, cultural geographers have drawn from Bruno Latour and Donna Haraway to interrogate the assumed duality between nature and culture (Demeritt, 1994; Mitchell, 1995; Matless, 1996, 1997). There are two issues at stake. First, the division between nature and culture enforces the view that humans are the sole social agents. An artifactual understanding of nature pluralizes agency; cultural geographers are now willing to consider that we live in a livelier world in which nonhuman actors also have agency. The notion of agency has been reworked, away from that of a conscious and

controlling self, to one of having effects. Second, conceptualizing the boundary between nature and culture as a social construction has opened up a rich set of investigations around both the production of the boundary and slippages across it. In particular, Sarah Whatmore (2006) has been at the forefront of developing materialist concerns, which engage with science and technology studies and issues of performance (see below) to theorize a 'more-than-human world' in which nature and culture are articulated as 'lived.' She further argues that both 'old' and 'new' cultural geographies have sought to "cast the making of landscapes (whether worked or represented) as an exclusively human achievement in which the stuff of the world is so much putty in our hands" (Whatmore, 2006: 603). Instead, she argues drawing upon the philosopher Gilles Deleuze for conceptualizing nature and culture in terms of 'livingness.' Moreover, the lines between economy and culture are no less blurry particularly when we examine issues of consumption, urbanization, and the mobility of capital (Jackson, 2002; Castree, 2004; Gibson and Kong, 2005; Amin and Thrift, 2007; Thrift, 2012). Economic development is increasingly about culture, whether it be in the form of tourism or the redevelopment of urban areas for the purposes of spectacle and consumption. Access to jobs and job performances are increasingly interpreted as cultural phenomena and cultural geographers have begun to examine this in the context of hospitality work (Bell, 2007).

Performances and Performativities

More recently, cultural geographies have taken a particular interest in practices of tourism, which had previously been approached from either anthropological or managerial perspectives. Cultural geographers were at the forefront of the launch of more theoretically orientated and informed academic journals, such as *Tourist Studies* and *Tourism Geographies*, as well as with a more explicit engagement with theory in longer established journals such as *Annals of Tourism Research*. As Franklin and Crang (2001: 3) point out, "tourism has broken away from its beginnings as a relatively minor and ephemeral ritual of modern national life to become a significant modality through which transnational modern life is organized." Rather than analyzing aspects of cultural tourism per se as a somewhat elite activity, they have sought to challenge some of the taken-for-granted dualisms in the analysis of tourism such as the distinction between the everyday and the exotic (Hannam and Knox, 2010). The close ties and connections between what we do at home and what we do on holiday have been explored by Knox (2009) in an account of mass tourism in the Mediterranean. Gale (2009) further backs this up with his study of the 'Paris plage,' the synthetic beach located on the banks of the Seine in Paris on which office workers, tourists, and local leisure-seekers engage in a simulation of beach tourism in the heart of the city.

Latterly, and perhaps more importantly, this has focused attention on the theorization of various embodied practices and experiences, touristic, or otherwise. The classic Cartesian mind/body dualism led many thinkers to consider minds and bodies as separate entities rather than recognizing their interconnectedness in the experience of being human. It is no longer a question of human subjects being considered to be a mind within a body, but about thinking beyond that dualism to see humans as embodied beings. As an example, tourism offers a range of physical and sensual stimuli that reassert the embodied nature of human life: the warmth of the sun on the face, water lapping around your feet, the increased heart rate while riding the roller coaster, the discomfort of the long-haul flight, the upset stomach, the hotel massage, the feelings of being under the influence of alcohol, the bumps of a rickshaw, or even of dancing (Hannam and Knox, 2010).

Some aspects of, and kinds of tourism easily lend themselves to being described as performances. The notion of performance, however, has a much wider and more significant meaning in the social sciences than simply those times when people dress up in costumes and follow scripts, act on a stage, or sing and dance. However, tourism as an industry or cultural activity is however very much tied up with the presentation of place, culture, heritage, or events, and these presentations could be said to take place on stages that are created out of the interaction between destinations or attractions and their staff and visitors. Edensor (2001) took the metaphor of performance further by casting tourists, marketeers, attraction staff and managers, and tour guides as the playwrights, directors, and stage crew.

Building on the notion of performance, since the late 1990s a number of cultural geographers (Nash, 2000; Lorimer, 2008; Wylie, 2010) have sought to develop nonrepresentational forms of theorization by considering the notion of 'performativity.' Performativity is an attempt to "find a more embodied way of rethinking the relationships between determining social structures and personal agency" (Nash, 2000: 654). By nonrepresentational, these writers argue that researchers should not just focus on texts and reading or images and their processing, but also ways of trying to think about and write about physical or emotional behaviors that do not readily lend them to being written about. Nonrepresentational theories challenged the status of social constructionist understandings of the world by highlighting some of the limitations of the representational approaches to studying cultural landscapes outlined above.

Nigel Thrift (1997: 126–127) argues that nonrepresentational theory is about "practices, mundane everyday practices, that shape the conduct of human beings toward others and themselves at particular sites." It is a project not "concerned with representation and meaning, but with the performative 'presentations,' 'showings,' and 'manifestations' of everyday life." It seeks to appreciate the ways in which ordinary people appreciate "the skills and knowledges they get from being embodied beings." Drawing upon Thrift's work, Nash (2000: 655), meanwhile, argues that the notion of performativity "is concerned with practices through which we become 'subjects' decentred, affective, but embodied, relational, expressive and involved with others and objects in a world continually in process. … The emphasis is on practices that cannot adequately be spoken of, that words cannot capture, that texts cannot convey – on forms of experience and movement that are not only or never cognitive." This was further developed into what has been termed 'more-than representational' theory in order to incorporate the insights of landscape analysis with the new emphasis on embodied

encounters (Lorimer, 2005, 2008). Landscapes are retheorized from this perspective as 'traveling landscape-objects' (Della Dora, 2009). Cultural geographers have examined such embodied encounters through, for example, the analysis of nonrational spectro-geographies (Maddern and Adey, 2008), lethargy and tiredness (Bissell, 2009), and touristic activities in slum spaces (Diekmann and Hannam, 2012), but also through the diverse aeromobilities and automobilities discussed below.

Mobilities

Developing from the concern with embodied practices and recognizing that cultures are not static, cultural geographers have also been concerned with everyday practices of movement (Adey, 2009; Price, 2010) and have been active, along with sociologists and anthropologists, in the development of what has been termed the 'new mobilities paradigm' (Sheller and Urry, 2006). In short, proponents of the mobilities paradigm argue that the concept of mobilities is concerned with mapping and understanding both the large-scale movements of people, objects, capital, and information across the world, as well as the more local processes of daily transportation, movement through public space, and the travel of material things within everyday life simultaneously (Hannam et al., 2006).

A great deal of mobilities research has analyzed forms and experiences of so-called 'corporeal' travel. This involves the travel of people for work, leisure, family life, pleasure, migration, and escape, organized in terms of contrasting time–space modalities (ranging from daily commuting to once-in-a-lifetime exile). Using an explicit mobilities perspective, recent work has considered how the technologies of travel have led to new cultural forms, including new cultural identities and subjectivities (see Blunt, 2007). It is recognized that political, technological, financial, and transportational changes have been critical in significantly lowering the mobility barriers for many but not for all. Indeed immobilities become particularly apparent during times of crisis and disaster (Hannam et al., 2006) and social mobilities may be marked by inequalities and injustices (Cresswell, 2010).

The conceptualization of places is important for the study of mobilities in this context (Cresswell, 2004). Often a clear distinction is made between places and those traveling to places; pushing or pulling people to visit. The mobilities paradigm argues against the ontology of distinct 'places' and 'people' (Sheller and Urry, 2006). Rather there is a complex relationality of places and persons. Places are thus not so much fixed but are implicated within complex networks through which 'hosts, guests, buildings, objects, and machines' are contingently brought together to produce certain performances. Moreover, places are also "about proximities, about the bodily co-presence of people who happen to be in that place at that time, doing activities together, moments of physical proximity between people that make travel desirable or even obligatory for some" (Hannam et al., 2006: 13).

From this perspective, cultural geographies have also become concerned with aspects of transport, of mapping the various aeromobilities and automobilities of everyday life (Cresswell, 2010). The sociotechnical system of air travel that has been turned into a form of mass mobility requires an exceptionally extensive and immobile place, the airport city with tens of thousands of workers orchestrating the millions of air journeys taking place each day (Hannam et al., 2006). At airports, travelers have their bodies biometrically scanned or 'read' and are continually monitored on CCTV. That symbol of national allegiance and identity – the passport – is now 'chipped' and used by governments as a tool to monitor the population's movements. The objects people carry, too, are also subject to processes of control through X-rays and searches. Moreover, the very material and immaterial architecture of the modern airport is calculated to control the corporeal movements of the traveler (Adey et al., 2007; Adey, 2008). When this breaks down due to an unexpected environmental event such as the Icelandic ash cloud eruption the consequences can be extreme (Adey and Anderson, 2011).

Through the notion of automobility, personal vehicles may effectively offer the tourist freedom and flexibility other modes of travel do not (*see* Automobility). Cars enable people to tailor their own itineraries and allow them to control the routes they travel between destinations. They also permit the driver (and his or her passengers) to decide when and where to stop, and consequently allow direction and velocity to be controlled. As Urry (2000: 61) has noted, the road has the power to 'set people free' as drivers are offered a range of choices, inaccessible to those using public methods of transport. Cars are therefore identified as important tools that may permit adventure (Sheller and Urry, 2000, 2003; Featherstone, 2004; Huijbens and Benediktsson, 2007; Collin-Lange and Benediktsson, 2011; Butler and Hannam, 2012) and have consequently fostered a 'dynamic culture of individualism' (Jacobsen, 2004: 7) but with significant consequences for the global environment (Vanderheiden, 2006).

Phillip Crang (2010: 192) recently argued that "cultural geography is no longer fashionable" but remains lively, creative, and experimental as the above discussions illustrate. Cultural geography is currently extremely vibrant, no longer operating as a bounded subdiscipline concerned simply with landscapes but as a critical perspective on everyday life.

See also: Automobility; Cultural Critique, Anthropological; Nature and Society in Geography; Postcolonial Geography; Tourism, Geography of.

Bibliography

Adey, P., 2008. Airports, mobility and the calculative architecture of affective control. Geoforum 39 (1), 438–451.

Adey, P., 2009. Mobility. Routledge, London.

Adey, P., Anderson, B., 2011. Anticipation, materiality, event: the icelandic ash cloud disruption and the security of mobility. Mobilities 6 (1), 11–20.

Adey, P., Budd, L., Hubbard, P., 2007. Flying lessons: exploring the social and cultural geographies of global air travel. Progress in Human Geography 31 (6), 773–791.

Amin, A., Thrift, N., 2007. Cultural-economy and cities. Progress in Human Geography 31 (2), 143–161.

Barnett, C., 1998. Cultural twists and turns. Environment and Planning D: Society and Space 16, 631–634.

Bell, D., 2007. The hospitable city: social relations in commercial spaces. Progress in Human Geography 31 (1), 7–22.
Bissell, D., 2009. Travelling vulnerabilities: mobile timespaces of quiescence. Cultural Geographies 16 (4), 427–445.
Blunt, A., 2007. Cultural geographies of migration: mobility, transnationality and diaspora. Progress in Human Geography 31 (5), 684–694.
Butler, G., Hannam, K., 2012. Independent tourists automobilities in Norway. Journal of Tourism and Cultural Change 10 (4), 285–300.
Castree, N., 2004. Economy and culture are dead! Long live economy and culture! Progress in Human Geography 28 (2), 204–226.
Collin-Lange, V., Benediktsson, K., 2011. Entering the regime of automobility: car ownership and use by novice drivers in Iceland. Journal of Transport Geography 19 (4), 851–858.
Cosgrove, D., 1985. Prospect, perspective and the evolution of the landscape idea. Transactions of the Institute of British Geographers 10, 45–62.
Crang, M., 1998. Cultural Geography. Routledge, London.
Crang, P., 2003. Field cultures. Cultural Geographies 10 (2), 251–252.
Crang, P., Mitchell, D., 2000. Editorial. Ecumene 7 (1), 1–6.
Crang, P., 2010. Cultural geography: after a fashion. Cultural Geographies 17 (2), 191–201.
Crang, P., Mitchell, D., 2002. Cultural geographies: an ecumenical journal. Cultural Geographies 9 (1), 1–2.
Cresswell, T., 1996. In Place/Out of Place: Geography, Ideology and Transgression. University of Minnesota Press, Minneapolis.
Cresswell, T., 2004. Place: A Short Introduction. Blackwell, Oxford.
Cresswell, T., 2010. New cultural geography – an unfinished project? Cultural Geographies 17 (2), 169–174.
Daniels, S., Lorimer, H., 2012. Until the end of days: narrating landscape and environment. Cultural Geographies 19 (1), 3–9.
Della Dora, V., 2009. Travelling landscape-objects. Progress in Human Geography 33 (3), 334–354.
DeLyser, D., Rogers, B., 2010. Meaning and methods in cultural geography: practicing the scholarship of teaching. Cultural Geographies 17 (2), 185–190.
Demeritt, D., 1994. The nature of metaphors in cultural geography and environmental history. Progress in Human Geography 18, 163–185.
Diekmann, A., Hannam, K., 2012. Tourism mobilities in India's slum spaces. Annals of Tourism Research 39 (3), 1316–1336.
Duncan, J.S., 1990. The City as Text: The Politics of Landscape Interpretation in the Kandyian Kingdom. Cambridge University Press, Cambridge.
Duncan, J.S., Gregory, D. (Eds.), 1999. Writes of Passage: Reading Travel Writing. Routledge, New York.
Edensor, T., 2001. Tourists at the Taj. Routledge, London.
Featherstone, M., 2004. Automobilities: an introduction. Theory, Culture & Society 21 (4–5), 1–24.
Foote, K.E., Hugill, P.J., Mathewson, K., Smith, J. (Eds.), 1994. Re-reading Cultural Geography. University of Texas Press, Austin.
Franklin, A., Crang, M., 2001. The trouble with tourism and travel theory? Tourist Studies 1 (1), 5–22.
Gale, T., 2009. Urban beaches, virtual Worlds and 'the end of tourism'. Mobilities 4 (1), 119–138.
Gibson, C., Kong, L., 2005. Cultural economy: a critical review. Progress in Human Geography 29 (5), 541–561.
Gregory, D., 1994. Geographical Imaginations. Blackwell, Oxford.
Gregory, D., 1999. Scripting Egypt: orientalism and the cultures of travel. In: Duncan, J., Gregory, D. (Eds.), Writes of Passage: Reading Travel Writing. Routledge, London.
Hannam, K., Knox, D., 2010. Understanding Tourism. Sage, London.
Hannam, K., Sheller, M., Urry, J., 2006. Editorial: mobilities, immobilities and moorings. Mobilities 1 (1), 1–22.
Harley, J.B., 1992. Deconstructing the map. In: Barnes, T.J., Duncan, J.S. (Eds.), Writing Worlds: Discourse, Text and Metaphor in the Representation of Landscape. Routledge, London.
Huijbens, E., Benediktsson, K., 2007. Practising highland heterotopias: automobility in the interior of Iceland. Mobilities 2 (1), 143–165.
Jackson, P., 1989. Maps of Meaning: An Introduction to Cultural Geography. Unwin Hyman, London.
Jackson, P., 2002. Commercial cultures: transcending the cultural and the economic. Progress in Human Geography 26 (1), 3–18.
Jacobsen, J., 2004. Roaming romantics: solitude-seeking and self-centredness in scenic sightseeing. Scandinavian Journal of Hospitality and Tourism 4 (1), 5–23.
Kitchin, R., Dodge, M., 2007. Rethinking maps. Progress in Human Geography 31 (3), 331–344.
Knox, D., 2009. Mobile practice and youth tourism. In: Obrador, P., Travlou, P., Crang, M. (Eds.), Cultures of Mass Tourism. Ashgate, London.
Lewis, P., 1979. Axioms for reading the landscape. In: Meinig, D.W. (Ed.), The Interpretation of Ordinary Landscapes. Oxford University Press, Oxford.
Lorimer, H., 2005. Cultural geography: the busyness of being 'more-than-representational'. Progress in Human Geography 29 (1), 83–94.
Lorimer, H., 2008. Cultural geography: non-representational conditions and concerns. Progress in Human Geography 32 (4), 551–559.
Maddern, J.F., Adey, P., 2008. Editorial: spectro-geographies. Cultural Geographies 15 (3), 291–295.
Mathewson, K., 1998. Cultural landscapes and ecology, 1995–96: of oecumenics and nature(s). Progress in Human Geography 22 (1), 381–389.
Mathewson, K., 1999. Cultural landscape and ecology II: regions, retrospects, revivals. Progress in Human Geography 23 (2), 267–281.
Matless, D., 1996. New material? Work in cultural and social geography, 1995. Progress in Human Geography 20, 379–391.
Matless, D., 1997. The geographical self, the nature of the social and geoaesthetics: work in social and cultural geography, 1996. Progress in Human Geography 21 (3), 393–405.
Mitchell, D., 1995. There's no such thing as culture: towards a reconceptualization of the idea of culture in geography. Transactions of the Institute of British Geographers 20, 102–116.
Mitchell, D., 1996. The Lie of the Land: Migrant Workers and the California Landscape. University of Minnesota Press, Minneapolis.
Mitchell, D., 2002. Cultural landscapes: the dialectical landscape – recent landscape research in human geography. Progress in Human Geography 26 (3), 381–389.
Mitchell, D., 2003. Cultural landscapes: just landscapes or landscapes of justice? Progress in Human Geography 27 (6), 787–796.
Nash, C., 2000. Performativity in practice: some recent work in cultural geography. Progress in Human Geography 24 (4), 653–664.
Nash, C., 2002. Cultural geography: postcolonial cultural geographies. Progress in Human Geography 26 (2), 219–230.
Nash, C., 2003. Cultural geography: anti-racist geographies. Progress in Human Geography 27 (5), 637–648.
Olwig, K., 2010. The 'British invasion': the 'new' cultural geography and beyond. Cultural Geographies 17 (2), 175–179.
Pinder, D., 2007. Cartographies unbound. Cultural Geographies 14 (3), 453–462.
Price, M., Lewis, M., 1993. The reinvention of cultural geography. Annals of the Association of American Geographers 83, 1–17.
Price, P., 2010. Cultural geography and the stories we tell ourselves. Cultural Geographies 17 (2), 202–210.
Sauer, C.O., 1925. The Morphology of Landscape, vol. 2. University of California Publications in Geography, 19–54.
Sheller, M., Urry, J., 2000. The city and the car. International Journal of Urban and Regional Research 24 (4), 737–757.
Sheller, M., Urry, J., 2003. Mobile transformations of 'public' and 'private' life. Theory, Culture & Society 20 (3), 107–125.
Sheller, M., Urry, J., 2006. The new mobilities paradigm. Environment and Planning A 38 (2), 207–226.
Shurmer-Smith, P. (Ed.), 2003. Doing Cultural Geography. Sage, London.
Shurmer-Smith, P., Hannam, K., 1994. Worlds of Desire, Realms of Power: A Cultural Geography. Edward Arnold, London.
Sidaway, J., 2000. Postcolonial geographies: an exploratory essay. Progress in Human Geography 24 (4), 591–612.
Thrift, N., 1997. The still point. In: Pile, S., Keith, M. (Eds.), Geographies of Resistance. Routledge, London, pp. 124–151.
Thrift, N., 2012. The insubstantial pageant: producing an untoward land. Cultural Geographies 19 (2), 141–168.
Tolia-Kelly, D., 2010. The geographies of cultural geography I: identities, bodies and race. Progress in Human Geography 34 (3), 358–367.
Urry, J., 2000. Sociology Beyond Societies. Routledge, London.
Vanderheiden, S., 2006. Assessing the case against the SUV. Environmental Politics 15 (1), 23–40.
Whatmore, S., 2006. Materialist returns: practising cultural geography in and for a more-than-human world. Cultural Geographies 13 (4), 600–609.
Willems-Braun, B., 1997. On cultural politics, sauer, and the politics of citation – reply. Annals of the Association of American Geographers 87, 703–708.
Wylie, J., 2010. Cultural geographies of the future, or looking rosy and feeling blue. Cultural Geographies 17 (2), 211–217.

Cultural Heritage

Christoph Brumann, Max Planck Institute for Social Anthropology, Halle, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cultural heritage includes the sites, things, and practices a society regards as old, important, and worthy of conservation. It is currently the subject of increasing popular and scholarly attention worldwide, and its conceptual scope is expanding. Most social scientists emphasize its functions for supporting ethnic, national, and elite interests but others point to its creative and counterhegemonic sides. The article reviews the relation of heritage with tourism and nostalgia, dissonant/negative heritage, heritage and religion, rural and urban heritage, and heritage institutions, in particular the United Nations Educational, Scientific and Cultural Organization and its conventions. People's personal attachments to heritage deserve further study.

Cultural heritage is usually taken to mean the sites, movable and immovable artifacts, practices, knowledge items, and other things that a group or society has identified as old, important, and therefore worthy of conscious conservation measures, often at the hands of specialized institutions. This invariably comprises only a selection of the total cultural repertory, much of which may not be perceived with similar consciousness. Such a use of 'heritage' is a relatively recent extension of the original lexical meaning of individual heirloom to a collective level, as is also true for corresponding terms in other languages such as 'patrimoine'/'patrimonio.' Cultural heritage overlaps with a number of other phenomena and terms, to the point of interchangeable usage, such as cultural property (which is often more clearly delimited and not always old), tradition (which more clearly points to collective practices and to informal modes of transmission, excluding, for example, the works of individual artists that might nonetheless be seen as heritage), social memory and sites of memory/'lieux de mémoire' (where the focus is on contemporary commemorative practices), and culture (which in its broad anthropological sense is not restricted to old and publicly recognized things). Tradition in the sense of public traditions is particularly close in meaning, and it appears that much that was discussed under this term in the 1980s is now coming back under a new label. In general, cultural heritage is less used as an analytic term in anthropology where alternatives such as culture, material culture, or performative culture are preferred; rather, it points to the public gaze on old cultural things that is of interest as a social phenomenon.

Attention for and reference to cultural heritage is growing, with many commentators diagnosing a 'heritage boom' or even a 'cult of heritage' in contemporary Western society from sometime around the 1960s on. This includes a conceptual expansion where things previously thought unworthy such as petrol stations, hay barns, or everyday dishes qualify for heritage status too. Intangible cultural heritage is establishing itself as a new category, including performances, crafts, and more ephemeral entities such as cultural spaces, and there is even talk of digital heritage now. Not only nation states, local communities, and civil-society organizations participate in this process but also international bodies such as the United Nations Educational, Scientific and Cultural Organization (UNESCO), and they have helped to spread heritage discourses and practices to all corners of the globe. Heritage conservation is universally acknowledged as a moral obligation, and the days when high modernism and socialist cultural revolutions aimed to wipe out large portions of the past in favor of a new society appear remote now.

The bulk of the total textual production on heritage is celebratory, and the bulk of academic research is still done in such fields as art history, architecture, and archaeology on the heritage items proper and on the practical problems of their conservation and presentation, such as restoration techniques, management questions, or juridical aspects. Museum studies have long established themselves as an independent domain. The social sciences contribute modestly to these applied fields, such as when investigating the social messages of museums and exhibitions, when estimating the monetary value of cultural goods, or indirectly when training museum and conservation practitioners (such as curators in ethnographic museums who have degrees in anthropology).

The remainder of this article, however, is concerned with the basic research done by anthropologists, sociologists, geographers, and modern historians on the social, political, and economic contexts of cultural heritage and its conservation *in situ*, which is outside an indoor museum or archival context. Corresponding to the general growth of attention, this is a burgeoning field, with new training centers and interdisciplinary programs in 'World Heritage Studies' and the like mushrooming and book series and journals such as the *International Journal of Heritage Studies* and the *International Journal of Cultural Property* blossoming. The founding conference of the Association of Critical Heritage Studies in Gothenburg (Sweden) in 2012, for instance, attracted 500 paper proposals – five times as many as expected – and a large majority did not employ a narrowly applied perspective.

Dominant in this new strand of research is a critical approach that for specific heritage items or 'heritagization' in general pinpoints the gaps between official, most often celebratory discourses, and the actual historical trajectories, selective amnesia, and true motives – often of commercialization and collective self-elevation – involved. Four common assumptions dominate these analyses, which the author has elsewhere called falsification, petrification, desubstantiation, and enclosure (Brumann, 2009). In an often taken-for-granted way, labeling places, things, or practices as cultural heritage is

assumed to falsify 'real' history; petrify what may no longer evolve as freely as more ordinary, unmarked culture; strip it of its original values and meanings, leaving only the heritage label; and claim it as the possession of a particular community, ethnic or religious group, class, or nation, thereby excluding other such groups. Such assumptions are shared by the founding figures of current critical heritage studies, such as geographer David Lowenthal (1985, 1996) or cultural historian Robert Hewison (1987). But they are also often present in a slightly older interest in the selective appropriation of traditions for contemporary purposes, traditions which in most cases just as well could be labeled heritage. Eric Hobsbawm and Terence Ranger's 'invention of tradition' approach (1983) has proved stimulating across the social sciences and humanities. In the most famous case study of the volume, historian Hugh Trevor-Roper dismantles the myths surrounding the kilt, showing how army regiments, nationalist urban intellectuals, a fraudulent brother pair, and a cloth manufacturer smelling business all contributed to transforming the eighteenth-century creation of an Englishman into the timeless emblem of Scottishness. In addition to Hobsbawm and Ranger's home discipline of history, anthropologists have been particularly receptive to this approach, providing scores of similar case studies. They have also been most active in dismantling the idea of true history lying beneath the beautification of official traditions, as inherent in Hobsbawm's distinction of 'custom' and 'genuine traditions' from the invented ones (1983: pp. 2–3, 8). Instead, Handler and Linnekin argue that "[t]he origin of cultural practices is largely irrelevant to the experience of tradition; authenticity is always defined in the present. It is not pastness or givenness that defines something as traditional. Rather, the latter is an arbitrary symbolic designation; an assigned meaning rather than an objective quality" (1984: p. 286). This leaves socially positioned claims about the past only. Particularly in indigenous rights claims, such a relativist position has proved quite challenging to sustain (cf. e.g., Clifford, 1988: pp. 277–346; Hanson, 1989, 1991).

The four common assumptions are also present in archaeologist Laurajane Smith's (2006) widely quoted formulation of the "authorized heritage discourse" or "AHD." She likewise argues that heritage is not simply there but constituted through social practice, and while the AHD and its claim of salvaging the past for the (nebulously defined) benefit of future generations serve the interests of the middle and upper classes and nationalist aspirations, their main function is to strengthen the control of professional experts and institutions over the heritage in question. This tendency often prevails even against these specialists' best intentions, and it tends to disregard the expertise and deny the rights of nonprofessional heritage carriers, especially subaltern ones such as the working class or indigenous peoples. Even more sinister is heritage in the cultural criticism formulated by British authors who see the boom as a sign of general malaise. "Worship of a bloated heritage invites passive reliance on received authority, imperils rational inquiry, replaces past realities with feel-good history, and saps creative innovation," writes David Lowenthal (1996: p. 12) and sees "[p]rejudiced pride in the past" as "not a sorry consequence of heritage; it is its essential purpose" (1996: p. 122). And archaeologist Kevin Walsh finds that "[h]eritage successfully mediates all our pasts as ephemeral snapshots exploited in the present … to guarantee the success of capital in its attempt to develop new superfluous markets" (1992: p. 149). Several commentators have seen such attacks as amounting to 'heritage-baiting' (Pickering and Keightley, 2006: p. 934; Samuel, 1994: p. 263).

Yet there are also more benign views of the social role of traditions and cultural heritage and caveats against a purely deconstructive analysis. Anthropologist Marshall Sahlins has charged the invention approach with being functionalist. Using the example of sumo wrestling, a famous piece of Japanese cultural heritage whose modern appearance is clearly a nineteenth-century creation, he questions whether the latter can be explained by "some group's quest for power, material gain, resistance or a need of identity" alone (1999: p. 407). He rather argues that "[m]odern sumo is clearly a permutation of older forms and relationships … That it might be suitably reinvented to fit the occasion might better be understood as a sign of vitality rather than of decadence" (1999: pp. 408–409). "[I]nventiveness of tradition" (1999: p. 408) is therefore the more apposite phrase, and Terence Ranger himself would prefer "invention by" – rather than "of" – tradition in hindsight (1993: p. 76). Such creativity itself may be an indicator of traditionality, as argued by anthropologist Christopher Tilley (1997) who, after thoroughly deconstructing a ritual performance that his Wala informants devised for a tourist audience, points out how this bricolage follows a very traditional Melanesian mode of recombining home-grown and imported cultural resources into new forms. The alleged political conservatism of cultural heritage has also been questioned. Historian Richard Samuel emphasizes that preservation is politically chameleonic, "subject to quite startling reversals over very limited periods of time" (1994: p. 303) and "[h]istorically … a cause which owes at least as much to the Left as to the Right" (1994: p. 288). The latter is certainly borne out by the experience of urban residents the world over in their uphill battles against hegemonic development interests transforming old neighborhoods (e.g., Brumann, 2012). All this speaks strongly for attention to the details.

There is, however, a considerable number of empirical studies now that implement such a program, dissecting the multiplicity of voices and interests surrounding cultural heritage and focusing not just on the official spokespeople and discourses but also laypeople's, nonofficial, and subaltern views. Anthropologists in particular have provided in-depth ethnographic studies. In his book on the Cretan town of Rethemnos (Rethymno), Herzfeld (1991) distinguishes between the 'monumental time' implicit in government protection of the medieval and early Renaissance architecture of the historic center, associated with the Venetians, and the 'social time' of the residents and their everyday lives. Conservation requirements imposed by external, mainly national, agencies and the manifold constraints on building alterations and extensions frustrate the residents. Yet in the course of time, many have also discovered the economic opportunities of tourism to the town, leaving them divided about the heritage regime. Joy (2011) describes the challenges that the World Heritage inscription of the mud architecture of the city of Djenné in Mali brings to its

residents. The legacy of French colonial romanticism and a national government intent on gaining a place on the global cultural map imposed a conservation regime that helps to attract foreign tourists as one of the few sources of income. Yet many citizens feel a heavy burden when the regular remudding of houses consumes more resources than the less 'authentic' tiling of facades and when in a poverty-stricken environment, basic amenities such as a canalization are still lacking. An uneven distribution of benefits is also apparent at Chichén Itzá in Mexico, as reported by Breglia (2006). While everyone is proud of the global recognition of the site, almost the entire profits from the intense tourism bypass the local residents, with the single exception of the family that still privately owns the land on which the famous buildings stand. A handful of Mayan families, however – often the descendants of the workers who once excavated the site – retain their moderate privileges as the employed guards and shopkeepers, even when due to the relocation of their houses outside the boundaries, some of their personal ties to heritage, such as their football matches beneath the Great Pyramid, have been cut. More broadly distributed are the benefits of heritage in the rediscovery of the courtyard houses in the old city of Damascus, as studied by Salamandra (2004) prior to the current civil war. Old families, often from the former professional and entrepreneurial elite, had left for the suburbs, renting out the subdivided houses to rural migrants, but now they reclaim their homes, converting them to new purposes such as shops, galleries, and restaurants. Salamandra sees them as asserting symbolic dominance over the new political leadership, which lacks a long-standing connection to the city.

In Damascus, the turn to heritage is mainly an inner-urban phenomenon but in Djenné, Chichén Itzá, and elsewhere, tourism is a major force in the political economy of cultural heritage and a key motive for the conscious rearrangement of cultural practices. In a study of Maasai and Samburu heritage performances and colonial-style high tea on carefully tended lawns at a ranch near Nairobi, Bruner and Kirshenblatt-Gimblett (1994) found that organizers, performers, and tourists, through their scripted interaction, collaborate in creating 'tourist realism.' Volkman (1990) has spoken of 'cultural editing' when analyzing the way Toraja funerary rituals on Sulawesi are shortened, brought into a neat sequence – complete with a running commentary – and cleansed of bloody animal sacrifices. Such modifications aside, however, the importance for status that is enhanced by the presence of domestic and foreign visitors continues to be the dominant factor. Playing to tourists' expectations may be a strategy to gain the resources needed for the continuation of traditional rituals deliberately kept from display, as Martinez reports for the Japanese pearl-divers' village of Kuzaki (1990). Moving beyond instrumental considerations, John and Jean Comaroff (2009) propose an entirely new take, showing how in a range of cases drawn from the indigenous United States and South Africa, commercialization and tourist display is precisely what convinces the protagonists of having an authentic and respectable cultural heritage in the first place. Tourists are not defenseless against the dominant narratives of heritage sites anyway, as Bruner (1994) demonstrates for the open-air museum of New Salem, Abraham Lincoln's erstwhile home. The visitors he talked to value a great variety of often unexpected qualities of the site and also engage with it in playful and ironic ways.

Several authors have identified nostalgia as a major force behind heritagization and heritage tourism, describing it as "an ambivalent longing to erase the temporal difference between subject and object of desire, shot through with not only the impossibility but also the ultimate unwillingness to reinstate what was lost" (Ivy, 1995: p. 10). Such sentiments and the associated visions of lost 'Gemeinschaft' wholeness are seen critically in the social sciences. Yet in a study of Stone Town, the historic center of Zanzibar City, and the range of nostalgias for different phases of the colonial and revolutionary past felt by the residents, Bissel acknowledges the possibility for these expressions to transcend a merely regressive mode (2005: p. 239). And Pickering and Keightley urge us to consider both the melancholic and the utopian potential of nostalgia, as it always entails a critique of current conditions (2006). To do so may presuppose a more fine-grained categorization such as that by Berliner (2012) who distinguishes 'endo-nostalgia' about an actually experienced past from the 'exo-nostalgia' of the Western tourists, conservation experts, and the UNESCO staff who all descend upon Luang Prabang and its Buddhist temples and colonial villas in recent years. He also emphasizes how much nostalgia is actually a transformative force in this former capital of the Lao kingdom, rather than conserving things.

Yet heritage conservation can also be motivated by the very opposite of nostalgia, a wish to preserve the memory of historical events and conditions in order to avoid their recurrence. Variously called dissonant, difficult, or negative heritage, such "places of pain and shame" (Logan and Reeves, 2008) occupy a growing portion of contemporary heritage regimes. UNESCO's World Heritage list, for example, features the Auschwitz concentration camp, the Hiroshima Peace Memorial, or the Bikini nuclear test site. Bruner (1996) studied African-American roots tourism to Elmina Castle in Ghana, the main entrepôt for the North American slave trade and itself a World Heritage site. The tourists' joint walk through the dungeons and the infamous 'door of no return' is an often overwhelming emotional experience, yet all the more disconcerting are encounters with the Ghanaians outside the castle for whom the guests are 'obruni' (whitemen) by virtue of their wealth, rather than long-last brothers and sisters returning 'home.' The meaning of such sites is often subject to constant negotiation, as Macdonald shows in her ethnography of the Nazi Party Rally Grounds in Nuremberg (2009). The site was locally downplayed in the immediate postwar period, not least because many in the (traditionally left-leaning) city felt that they were made to bear an unfair share of the Nazi legacy, but the tide changed around reunification when suspicions of negating the past became the least desirable option. Meskell (2002) has reflected on a special kind of ambiguous heritage, taking up public debates about how to interpret, preserve, and achieve 'past mastering' at Ground Zero and the empty Bamiyan Buddha niches, that is, sites of deliberate, public destruction. Aside from such signal acts, heritage may also provoke petty vandalism, such as in the 'heritage iconoclasm' Chalcraft (2012) observed in Tanzanian, Libyan, and Italian prehistoric

rock art sites and interprets as resistance to the neocolonial aspects of the heritage regime.

Such contestation may arise from the fact that the things and practices recognized as heritage retain their older meanings, contrary to what the assumption of 'desubstantiation' (see above) would let us expect. This may be an entirely harmonious cohabitation, such as in the case of Kyoto where the 'kyô-machiya' houses and the Gion matsuri festival have become celebrated cultural heritage but continue to attract people also as beautiful, atmospheric, and ecological dwellings and as occasions for festive fun, entertainment, and supernatural protection (Brumann, 2009). But there are often also tensions, particularly between heritagization and religious significance. The sacred grove of the Yoruba goddess Osun in Osogbo, Nigeria – the site of a major rejuvenation of religious sculpture in the postindependence period and of an internationally visited annual festival – has over the years strengthened its heritage status, all the way to World Heritage inscription (Probst, 2011). One supporting factor is that traditional political authorities see themselves challenged by devout Christians and Muslims when they side with pagan religion, rather than supporting an emblem of transnational Yoruba heritage. Such an uneasy balance is not always attained, as the 2001 blasting of the stone Buddhas in Bamiyan by the Taliban or the recent destructions of the World Heritage-listed sufi saints' tombs in Timbuktu attest to. Much as seeking the global stage was paramount for the perpetrators, they also challenged heritage primacy, insisting on applying a religious yardstick to what in certain fundamentalist interpretations are pagan and blasphemous sites.

The ramifications of cultural heritage, happy or sad, also depend on where exactly it is found or imagined to belong. In modern nations, the countryside, rural folk, and their traditional customs have often served as a projecting screen for – often themselves rather urban and intellectual – longings for primeval wholeness and uncontaminated values, as has been observed for a variety of countries. In my own work on Kyoto, I have ventured the hypothesis that urban cultural heritage is less easily hijacked by outsiders, given that wealth, power, and intellectual resources usually concentrate in cities, thus giving their inhabitants more defenses. Also, urban tastes for creative and cosmopolitan mixtures do not insist on presumably pristine representations of culture so much (Brumann, 2009, 2012).

Although overviews of the evolution of heritage conservation do exist (Jokilehto, 1999) and steps toward a critical social history of Western heritage institutions since their full-scale emergence in the nationalist 'long nineteenth century' have been taken (Hall, 2011; Swenson, 2013; Swenson and Mandler, 2013), a comprehensive analysis of the rise of modern institutional conservation and its key ideas, social organization, and wider contexts is still lacking. Some work on contemporary national heritage apparatuses has been done (e.g., Bendix et al., 2012; Hauser-Schäublin, 2011), notably on the Japanese system as the internationally most influential non-Western one (Brumann and Cox, 2010; Thornbury, 1997). Lynn Meskell (2012a) deconstructs the role of the governing institutions of Krueger National Park in South Africa, showing how neoliberal pressures and the hegemony of nature tourism and biodiversity lead to the neglect of the many archaeological and rock art sites in the park and also to a suppression of painful histories of racism and eviction. She is one among a growing number of practicing archaeologists who reflect in sensitive ways about the wider societal context of their own work and the involvement of heritage discourses (e.g., Karlström, 2009). Still, more studies like sociologist Nathalie Heinich's in-depth analysis of the organization, underlying values, and hidden logic of the French 'Inventaire' (2009) appear called for, given that the influence of such bodies as the Instituto Nacional de Antropología e Historia in Mexico or the Agency of Cultural Affairs (Bunkachô) in Japan is obvious from studies of cases under their governance (cf. Breglia, 2006; Brumann and Cox, 2010; Thornbury, 1997).

How exactly states deal with their cultural heritage depends increasingly on external actors. These include foundations such as the Getty Foundation, the Aga Khan Foundation, or the World Monuments Fund, the conventions and operations of international organizations such as the European Union, and the funding that bodies such as the World Bank or United Nations Development Programme provide. In particular, UNESCO with its portfolio of conventions is establishing itself as the supreme arbiter of global heritage standards and as an impetus for much national and regional regulation. Still best known and most coveted is the 'Convention Concerning the Protection of the World Cultural and Natural Heritage' adopted in 1972, which now aims to protect close to 1000 cultural and natural sites inscribed on its prestigious list – including many of the above-mentioned cases – as the shared inheritance of humanity. Charges of Eurocentrism have led to a greater receptiveness for vernacular architecture, cultural landscapes, or dissonant heritage and to a self-consciously 'anthropological' perspective. Dissatisfaction with the way the institution represents our world is growing, however, particularly among the heritage latecomers from the Global South that are now as adept at pushing their national interests in this forum as the Northern states (Brumann, 2011; Meskell, 2012b; Schmitt, 2009).

One consequence of this problem has been UNESCO initiatives to honor not only sites but also crafts, performative arts, and the like, first with proclamations of 'Masterpieces of the Oral and Intangible Heritage of Humanity' from 2001 on and then with the adoption of a full-fledged 'Convention for the Safeguarding of Intangible Cultural Heritage' in 2003. More so than World Heritage, which is widely regarded as the province of architects, art historians, and archaeologists, this process has involved anthropologists and folklorists on juries, advisory committees, and decision-making boards. Since national states are rushing to sign the treaty, this is replicated on the national level where the required institutional frameworks are currently taking shape. The governing organs of the new convention have likewise taken to listing, and what gets inscribed is partly informed by a rather conventional folkloristic perspective and partly very general – tango, the French gourmet meal, the Novruz festival. Contemporary anthropologists who are committed to more dynamic views of culture, which would see continuously evolving landscapes rather than neatly delimited and stable packages, have been skeptical of the intellectual merits of such an exercise (Kirshenblatt-Gimblett, 2004) but some also acknowledge the possibilities of

empowerment and improved preservation (e.g., Brown, 2005). Anthropologists, folklorists, and geographers pursue the institutional processes of the convention and their evolution (e.g., Bortolotto, 2010; Rudolff, 2010; Schmitt, 2008), some of them as full participants representing their countries in this arena (Hafstein, 2009; Kuutma, 2007). Several recent interdisciplinary collections attempt to take stock of the intangible cultural heritage (ICH) phenomenon too (Arizpe, 2011; Bortolotto, 2011; Smith and Akagawa, 2009). As this convention is still very much in the formative period, its further fate and growth curve will remain interesting to follow.

All global-level institutionalization and politicization aside, the fact remains that many people feel personally attached and committed to heritage things and practices. The label as such may have limited meaning for them, and their historicities may be alternative ones, such as those of festival participants in Kyoto who are less impressed by the documented great age of some of the paraphernalia but by being a link in a chain of transmission that reaches back centuries (Brumann, 2012: p. 239). Rarely are such attachments explored systematically at the individual level, although McCracken's sketch of a 'curatorial consumer' points the way (1988: pp. 44–56). Beyond all instrumentalization, things that last – resisting change and transcending our own life spans – inspire awe among many people in many societies. This throws up very basic philosophical questions about a panhuman "desire for grounded materiality" (Meskell, 2002: p. 559) and invites further studies relating heritage with the wider context of materiality, performativity, notions of (im)permanence (Karlström, 2009), and emotions in the specific cultural context. As for sentiments, critical heritage researchers may want to reflect on their own, including their positions in and funding by institutions premised on the very commitments to heritage they like to dissect.

See also: Authenticity, Anthropology of; Cultural Resource Management: Conservation of Cultural Heritage; Culture: Contemporary Views; Ethnic Identity and Ethnicity in Archaeology; Ethnicity: Anthropological Aspects; Identity in Anthropology; Materiality and Culture; Nationalism: General; Nostalgia, Anthropology of; Tradition, Anthropology of; Urban Anthropology.

Bibliography

Arizpe, L. (Ed.), 2011. Compartir el patrimonio cultural inmaterial: Narrativas y representaciones. Conaculta, UNAM, CRIM, México.

Bendix, R.F., Eggert, A., Peselmann, A. (Eds.), 2012. Heritage Regimes and the State. Universitätsverlag Göttingen, Göttingen.

Berliner, D., 2012. Multiple nostalgias: the fabric of heritage in Luang Prabang (Lao PDR). Journal of the Royal Anthropological Institute 18, 769–786.

Bissell, W.C., 2005. Engaging colonial nostalgia. Cultural Anthropology 20, 215–248.

Bortolotto, C., 2010. Globalising intangible cultural heritage? between international arenas and local appropriations. In: Labadi, S., Long, C. (Eds.), Heritage and Globalisation. Routledge, London.

Bortolotto, C. (Ed.), 2011. Le patrimoine culturel immatériel: Enjeux d'une nouvelle catégorie. Maison de Sciences de l'Homme, Paris.

Breglia, L.C., 2006. Monumental Ambivalence: The Politics of Heritage. University of Texas Press, Austin.

Brown, M.F., 2005. Heritage trouble: recent work on the protection of intangible cultural property. International Journal of Cultural Property 12, 40–61.

Brumann, C., 2009. Outside the glass case: the social life of urban heritage in Kyoto. American Ethnologist 36, 276–299.

Brumann, C., 2011. Unser aller Kulturgut: Eine ethnologische Annäherung an das UNESCO-Welterbe. Sociologus 61, 19–43.

Brumann, C., 2012. Tradition, Democracy and the Townscape of Kyoto: Claiming a Right to the Past. Routledge, London.

Brumann, C., Cox, R.A. (Eds.), 2010. Making Japanese Heritage. Routledge, London.

Bruner, E.M., 1994. Abraham Lincoln as authentic reproduction: a critique of postmodernism. American Anthropologist 96, 397–415.

Bruner, E.M., 1996. Tourism in Ghana: the representation of slavery and the return of the black diaspora. American Anthropologist 98, 290–304.

Bruner, E.M., Kirshenblatt-Gimblett, B., 1994. Maasai on the lawn: tourist realism in East Africa. Cultural Anthropology 9, 435–470.

Chalcraft, J.M., 2012. Decolonising the Site: The Problems and Pragmatics of World Heritage in Italy, Libya and Tanzania. Conference "World Heritage on the Ground: Ethnographic Perspectives". Max Planck Institute for Social Anthropology, Halle.

Clifford, J., 1988. The Predicament of Culture: Twentieth Century Ethnography, Literature, and Art. Harvard University Press, Cambridge, Massachusetts.

Comaroff, J.L., Comaroff, J., 2009. Ethnicity, Inc. The University of Chicago Press, Chicago.

Hafstein, V.T., 2009. Intangible heritage as a list: from masterpieces to representation. In: Smith, L., Akagawa, N. (Eds.), Intangible Heritage. Routledge, London.

Hall, M. (Ed.), 2011. Towards World Heritage: International Origins of the Preservation Movement 1870–1930. Ashgate, Farnham, Surrey.

Handler, R., Linnekin, J., 1984. Tradition, genuine or spurious. Journal of American Folklore 97, 273–290.

Hanson, A., 1989. The making of the maori: culture invention and its logic. American Anthropologist 91, 890–902.

Hanson, A., 1991. Reply to Langdon, Levine, and Linnekin. American Anthropologist 93, 449–450.

Hauser-Schäublin, B. (Ed.), 2011. World Heritage Angkor and Beyond: Circumstances and Implications of UNESCO Listings in Cambodia. Universitätsverlag Göttingen, Göttingen.

Heinich, N., 2009. La Fabrique Du Patrimoine: De la Cathédrale À la Petite Cuillère. Maison des sciences de l'homme, Paris.

Herzfeld, M., 1991. A Place in History: Social and Monumental Time in a Cretan Town. Princeton University Press, Princeton.

Hewison, R., 1987. The Heritage Industry: Britain in a Climate of Decline. Methuen, London.

Hobsbawm, E., 1983. Introduction: inventing traditions. In: Hobsbawm, E., Ranger, T. (Eds.), The Invention of Tradition. Cambridge University Press, Cambridge.

Hobsbawm, E., Ranger, T. (Eds.), 1983. The Invention of Tradition. Cambridge University Press, Cambridge.

Ivy, M., 1995. Discourses of the Vanishing: Modernity, Phantasm, Japan. University of Chicago Press, Chicago.

Jokilehto, J., 1999. A History of Architectural Conservation. Elsevier Butterworth-Heinemann, Amsterdam.

Joy, C., 2011. The Politics of Heritage Management in Mali: From UNESCO to Djenné. Left Coast Press, Walnut Creek, California.

Karlström, A., 2009. Preserving Impermanence: The Creation of Heritage in Vientiane, Laos. (Ph.D. dissertation). Uppsala University, Uppsala.

Kirshenblatt-Gimblett, B., 2004. Intangible heritage as metacultural production. Museum International 56, 52–65.

Kuutma, K., 2007. The politics of contested representation: UNESCO and the Masterpieces of Intangible Cultural Heritage. In: Hemme, D., Tauschek, M., Bendix, R. (Eds.), Prädikat "Heritage": Wertschöpfungen aus kulturellen ressourcen. Lit, Berlin.

Logan, W.S., Reeves, K. (Eds.), 2008. Places of Pain and Shame: Dealing with "difficult Heritage". Routledge, London.

Lowenthal, D., 1985. The Past Is a Foreign Country. Cambridge University Press, Cambridge.

Lowenthal, D., 1996. Possessed by the Past: The Heritage Crusade and the Spoils of History. The Free Press, New York.

Macdonald, S., 2009. Difficult Heritage: Negotiating the Nazi Past in Nuremberg and beyond. Routledge, London.

Martinez, D.P., 1990. Tourism and the *ama*: the search for a real Japan. In: Ben-Ari, E., Moeran, B., Valentine, J. (Eds.), Unwrapping Japan: Society and Culture in Anthropological Perspective. Manchester University Press, Manchester.

McCracken, G.D., 1988. Culture and Consumption: New Approaches to the Symbolic Character of Consumer Goods and Activities. Indiana University Press, Bloomington.

Meskell, L., 2002. Negative heritage and past mastering in archaeology. Anthropological Quarterly 75, 557–574.

Meskell, L., 2012a. The Nature of Heritage: The New South Africa. John Wiley & Sons, Malden, Massachusetts.
Meskell, L., 2012b. The rush to inscribe: reflections on the 35th session of the world heritage committee, UNESCO Paris, 2011. Journal of Field Archaeology 37, 145–151.
Pickering, M., Keightley, E., 2006. The modalities of nostalgia. Current Sociology 54, 919–941.
Probst, P., 2011. Osogbo and the Art of Heritage: Monuments, Deities, and Money. Indiana University Press, Bloomington.
Rudolff, B., 2010. "Intangible" and "Tangible" Heritage. A Topology of Culture in Contexts of Faith. Scientia Bonnensis, Bonn.
Sahlins, M., 1999. Two or three things that I know about culture. Journal of the Royal Anthropological Institute 5, 399–421.
Salamandra, C., 2004. A New Old Damascus: Authenticity and Distinction in Urban Syria. Indiana University Press, Bloomington.
Samuel, R., 1994. Theatres of Memory. In: Past and Present in Contemporary Culture, vol. 1. Verso, London.
Schmitt, T.M., 2008. The UNESCO concept of safeguarding intangible cultural heritage: its background and *Marrakchi* roots. International Journal of Heritage Studies 14, 95–111.
Schmitt, T.M., 2009. Global cultural governance: decision making concerning world heritage between politics and science. Erdkunde 63, 103–121.
Smith, L., 2006. Uses of Heritage. Routledge, New York.
Smith, L., Akagawa, N. (Eds.), 2009. Intangible Heritage. Routledge, London.
Swenson, A., 2013. The Rise of Heritage: Preserving the Past in France. Germany and England. Cambridge University Press, Cambridge, 1789–1914.
Swenson, A., Mandler, P. (Eds.), 2013. From Plunder to Preservation: Britain and the Heritage of Empire, C.1800–1940. Oxford University Press, Oxford.
Thornbury, B.E., 1997. The Folk Performing Arts: Traditional Culture in Contemporary Japan. State University of New York Press, Albany.
Tilley, C., 1997. Performing culture in the global village. Critique of Anthropology 17, 67–89.
Volkman, T.A., 1990. Visions and revisions: toraja culture and the tourist gaze. American Ethnologist 17, 91–110.
Walsh, K., 1992. The Representation of the Past: Museums and Heritage in the Postmodern World. Routledge, London.

Relevant Websites

whc.unesco.org – United Nations Educational, Scientific and Cultural Organization.
www.unesco.org/culture/ich – United Nations Educational, Scientific and Cultural Organization.

Cultural History

Roger Chartier, Ecole des Hautes Etudes en Sciences Sociales, Paris, France

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 5, pp. 3075–3081, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

It is not easy to define 'cultural history.' The risk is great of failing to draw a clear distinction between cultural history and other histories, such as the history of ideas, the history of literature, the history of art, the history of education, the history of media, or the history of sciences. Should we, consequently, change perspectives and consider that all history, whatever its nature – economical or social, demographic or political – is cultural, insofar as the most objectively measurable phenomena are always the result of the meanings that individuals attribute to things, words, and behavior? In this fundamentally anthropological perspective, the problem is not so much to define the particular sphere of cultural history, differentiated from that of its neighbors, but rather avoid an imperialist definition of the category.

Between these two stumbling blocks, the road is narrow. The course followed in this article consists of marking the shifts that have characterized the historiographical practices designated, in their time or subsequently, as belonging to cultural history.

It is not easy to define 'cultural history' in its specificity. Should it be done by designating objects and practices whose study would constitute the very nature of this history? The risk is great, therefore, of failing to draw a definite and clear line between cultural history and other histories: for instance, the history of ideas, the history of literature, the history of art, the history of education, the history of media, or the history of sciences. Should we, consequently, change perspectives and consider that all history, whatever its nature – economical or social, demographic or political – is cultural, insofar as the most objectively measurable phenomena are always the result of the meanings that individuals attribute to things, words, and behavior? In this fundamentally anthropological perspective, the problem is not so much to define the particular sphere of cultural history, differentiated from that of its neighbors, but rather avoid an imperialist definition of the category.

Between these two stumbling blocks, the road is narrow. From there, the course followed in this article consists of marking the shifts that have characterized the historiographical practices designated, in their time or subsequently, as belonging to cultural history.

The History of Mentalities

The very lengthy genealogy of cultural history generally stops at a certain number of precursors: the historians of the nineteenth century (Michelet, Burckhardt), Voltaire's, *Siècle de Louis XIV*, the legal practitioners of the 'perfect history' during the sixteenth century, even Herodotus himself. The course is not without illusion, measuring the 'anticipations' of certain precursors according to a posterior state of historical science. So as to avoid this trap, it is preferable to limit the research to the twentieth century and to follow the different faces taken by the historiographical projects which intended to focus on the phenomena left aside by the classical forms of political, economical, or social history. The history of mentalities was the first among them.

The Founders of the Annales and the History of Mentalities

Unlike a given idea, the paternity of the category of the 'history of mentalities,' insofar as it indicates a particular area of history, is not to be attributed to the founders of the *Annales*, Lucien Febvre and Marc Bloch. It is the invention of Robert Mandrou and Georges Duby at the end of the 1950s. The expression appears for the first time in the title of a university course in 1956 with the election of Robert Mandrou in the Sixth Section of the Ecole Pratique des Hautes Études to the chair of 'Social History of Modern Mentalities,' while at the same time, Georges Duby was opening a seminar on 'medieval mentalities' at the University of Aix-en-Provence. In 1961, when Robert Mandrou published his book *Introduction à France, Modern*, he presented it as a response to the 'requests formerly made by Lucien Febvre in favor of the history of collective mentalities' (while using another notion as the subtitle of the study, 'psychological history' which was also dear to Febvre). Thus, designated as the inspirer of this new form of history, Lucien Febvre, in fact, had rarely used the exact term 'mentality,' preferring the adjective 'mental' added to words like 'equipment,' 'material,' but above all 'tool,' or even, in the plural form, 'habits' and 'needs.' In 1942, in the *Problème de l'incroyance au XVIe siècle: Religion de Rabelais*, he characterized the 'mental tools' by expressing two essential statements: that the ways to perceive and reason are neither invariable nor universal, and that there is no continuous and necessary progress in the succession of mental tools. The book makes an inventory of the instruments and conceptual categories which are the different supports for thinking: first, the state of language, with its vocabulary and syntactic particularities, the tools and the languages available in the operations of knowledge, and finally the value and the credit attributed to each sense. From there, the conclusion: "So close to us in appearance, the contemporaries of Rabelais are already distant by all their intellectual properties. And even their structure was not ours." Each time, the ways of thinking and feeling outline in a specific way, the limits between nature and the supernatural, or between what is possible and impossible.

Paradoxically, Lucien Febvre, taken to be the father of the history of mentalities, uses the term less often than Marc Bloch who uses it as much in the *Société féodale* (where we encounter 'mental atmosphere' or 'religious mentality') as in *Apologie pour l'histoire ou le métier d' historien*. He prefers 'sensitivity' that defines the subject of 'historical psychology' which seems for him the only capable way of avoiding the culpable anachronism which equips men and women of the past, not only with knowledge and conceptions which were impossible for them, but also with feelings and emotions that were unknown to them. Febvre stigmatizes such an error in *Amour sacré, amour profane: autour de l'Heptaméron* (1944) by concluding: "In fact, a man of the sixteenth century should be intelligible not in relation to us, but in relation to his contemporaries."

In his book of 1961, Robert Mandrou does not dissociate the study of the mental tool from that of sensations, emotions, and passions which make up the mentality. Without separating these two dimensions, he distinguishes the common elements to all men (and women) in a shared time and place and those that are particular to each generation, each profession, each social group, or each class. For him, "all historical psychology, all history of mentalities is social history." We should, therefore, once the common mental tool has been identified, describe 'the mental horizons which are characteristic of the different social groups.' The 'history of mentalities,' or of the 'visions of the world,' another term often used by Mandrou, was thus strongly anchored in the differences between social classes, defined more by the unity of lifestyle and the feeling of belonging than by a strict economical determination.

The Golden Age of the History of Mentalities

From the 1960s, the notion of mentality imposes itself to designate a history whose object is neither ideas nor socioeconomic realities. This 'French' history of mentalities reposes on a certain number of ideas more or less shared by those who practice it (see Le Goff, 1974). In the first instance, the object of the history of mentalities is defined by Le Goff as the opposite of that of classical intellectual history: "The level of the history of mentalities is that of everyday life and the automatic, which is what escapes individual subjects as it shows the impersonal content of their thoughts." These ideas, which result from the conscious elaboration of a singular spirit, are therefore opposed to mentality, always collective, which regulates, without them knowing, the immediate perceptions of social subjects. Such an expression is not very far off the definition of collective representations in the tradition of the sociology of Durkheim as the accent is placed on the contents or the methods of thinking which result from the unconscious incorporation of unknown determinations in each member of a community, which set up their common manner of classifying and judging. Now, the second characteristic underlined by Le Goff: the possibility that the history of mentalities or the historical psychology 'link themselves to another important trend of historical research today: quantitative history.' Having as subjects collective, automatic, and repetitive actions, the history of mentalities should and must be serial and statistical. It is part of the heritage of the history of economies, populations, and societies that, on the horizon of the major crisis of the 1930s, followed by the mutations after the war, constituted the most innovative field of historiography. When, in the 1960s, the history of mentalities and the historical psychology defined a new, promising, and original area of study, they did so often by recapitulating the methods which ensured the conquests of socioeconomical history: the techniques of regressive statistics and the mathematical analysis of series.

From the importance given to series, and therefore to the establishment and treatment of homogenous data, repeated and comparable at temporal regular intervals, two consequences follow. The first is the privilege given to massive sources, widely representative and available over a long period, for example, the inventories after death, wills, library catalogs, legal archives, etc. The second consists of the attempt to articulate, according the Braudelian model of different times (long term, conjuncture, event), the long period of mentalities which often resists to change, with the short period of brutal ruptures or of rapid transfers of belief and sensitivity.

The withdrawal of witchcraft as a criminal act in France during the seventeenth century (see Mandrou, 1968), the transfer of attitudes before childhood or death (see Ariès, 1964, 1977), or the de-Christianization of France during the second half of the eighteenth century (see Vovelle, 1973) illustrate the articulation of the different periods of the history of mentalities. In each case, the problem lies in understanding how, in the long-term stability of mental structure, an essential transfer occurs: thus, the transformation of the representations of the world in the milieu of magistrates, the invention of childhood, and the concealment of death, or the indifference toward devotion practices, moral injunctions, and catholic beliefs.

A third characteristic of the history of mentalities in its golden age lies in its ambiguous way of considering connection to society. The notion seems, in fact, dedicated to erasing the differences in order to find the categories shared by all the members of a same era. "The mentality of a historical individual, a great man, is exactly what he has in common with other men of his time" writes Le Goff (1974), adding, as examples, "it is what Caesar and the last soldier of his legions, Saint Louis and the peasant of his domains, Christopher Columbus and the sailor all have in common." Among all the practitioners of the history of mentalities, Philippe Ariès is no doubt the one who made the greatest attachment to such an identification of the notion as 'common feeling' or 'general feeling.' The recognition of the 'archetypes of civilization,' shared by a whole society, does certainly not signify the cancellation of all differences between social groups or between clerics and laymen. However, these differences are always considered inside a long-term process which produces representations and behavior which become common. Postulating the fundamental unity (at least tendentious) of the 'collective unconscious,' Philippe Ariès reads the texts and images, not like demonstrations of individual peculiarities, but in order to "decipher, beyond the will of authors or artists, the unconscious expression of collective sensitivity," or to "find, beneath the ecclesiastic language, the ordinary set of common representations which are obvious" (see Ariès, 1975). The sensitivity and the collective gestures which are disclosed should be understood at the crossroads of biology and the mental, and at the meeting point of demographic realities (birth, death, etc.) and psychological investments (the forms of self-consciousness, the representations of life

after death, the feeling of childhood, etc.). According to Philippe Ariès, mentality refers to 'currents of the deep' which govern, without them necessarily being aware of it, the most essential attitudes of men and women of the same period.

For other mentality historians, more directly located in the heritage of social history, the essential elements lie in the link between the differences between the ways of thinking and feeling and social differences. Such a perspective organizes the classification of mentality facts into divisions established by the analysis of society and then the superposition postulated as necessary between social boundaries that separate groups or classes and those which differentiate mentalities. This social cutting out is no doubt the most precise trace of the dependence of the 'history of mentalities' in relation to social history in French tradition. It was possible to understand it at a global and macroscopic level – and so in research aiming to characterize a mentality, a religion, or a 'popular' literature, opposed finally to that of the dominant or the elite – or in a more fragmented way, in reference to the hierarchy of conditions and professions. However, in both cases, the study of mental horizons reproduces the divisions proposed by the history of societies.

History of Mentalities or Historical Psychology?

Mentalities, sensitivities, visions of the world: the unstable plurality of vocabulary indicates, at the same time, both the difficulty in defining objects of a new historiographical approach and the will to link, in the same perspective, intellectual and psychological categories.

When Alphonse Dupront proposes, in 1960, at the International Congress of Historical Sciences at Stockholm, to constitute historical psychology as a whole discipline within human sciences, he gives it a maximal extension as it must be "the history of values, mentalities, forms, symbolics, myths." Such a definition reduced the distance established by the founders of *Annales* between mentality and ideas, as the latter participate fully in 'collective mentality' of men of a period. The ideas, perceived through the circulation of words which designate them, situated in their social rooting, considered in their affective and emotional load as much as in their intellectual content, therefore become, just like myths or values, one of those 'collective forces via which men live their time,' one of those elements which Dupront, in words borrowed from Jung, called 'collective psychic.' An expression exists there which, while claiming to be loyal to the project of the *Annales*, surpasses the old oppositions by giving a fundamental psychological definition of mentality and by reintroducing the ideas in the exploration of the collective mental.

Such a perspective (without the word 'mentality') appears in the work of Ignace Meyerson, whose importance, perhaps underestimated today, was central for the renewal of historical studies of antiquity – in particular his book *Les Fonctions psychologiques et les oeuvres* of 1948. A first relationship lies in the assertion of the fundamental historicity of mental categories and the psychological functions. It is this essential historicity of psychological objects which allows Meyerson to define it as a 'historical anthropology': "The psychological functions have a history and they have had different forms throughout this history. Time and memory have a history. Space has a history. Perception has a history. The person has a history." The work of historical psychology does not, therefore, consist of finding different modalities or expressions of functions considered as stable and universal. It attempts to understand, in their discontinuity and their singularity, the emergence and the economy of each of these functions. In this way, by applying the perspective of Meyerson to the question of the person in ancient Greece, Jean-Pierre Vernant writes: "There is not, there cannot be a model person, exterior throughout human history, with differences, variations according to places, transformations due to time. Research should therefore not establish if the person, in Greece, is or is not, but should research what the ancient Greek person is, how he is different, in the multiplicity of his features, of today's person" (see Vernant, 1965).

On the other hand, Ignace Meyerson radically modifies the location and comprehension of the mental and psychological categories. To their immediate, existentialist, phenomenological grasp, he opposes their knowledge based on symbolic forms, works, and acts in which they are objectivized. Analyzing psychological functions via productions (institutional, religious, legal, aesthetic, linguistic) allows for the rupture with the idea of universal and abstract men, and with the universalization of a particular form of the personality.

Success and Criticisms of the History of Mentalities

How can one explain the infatuation, of historians and readers, in France and outside of France, for the history of mentalities, whatever be the designation, in the 1970s and 1980s? No doubt because such an approach allowed for, in its very diversity, the introduction of a new balance between history and social sciences. Contested in its intellectual and institutional superiority by the development of psychology, sociology, and anthropology, history coped by annexing the topics of the discipline which questioned its domination. The focus then moved toward objects (systems of belief, collective attitudes, ritual forms, etc.) which, until then, belonged to the neighbors but which fully entered into the program of a history of collective mentalities.

Adapting to the approaches and analysis methods of socioeconomical history while transforming the historical questionnaire, the history of mentalities (in its widest definition) was able to occupy the front part of the historiographical scene and constitute an effective response to the challenge launched by social sciences.

However, there were many critics of its principles and methods. The first came from Italy. In 1970, Franco Venturi denounced the obliteration of the creative force of new ideas for the benefit of simple 'mental structures' 'lacking dynamism and originality' (see Venturi, 1970). Some years later, Carlo Ginzburg magnified the criticism (see Ginzburg, 1976). He refused the notion of 'mentality' for three reasons: first, for its exclusive insistence on 'elements which are inert, obscure, and unconscious of a determined vision of the world,' which lead to reducing the importance of rationally and consciously expressed ideas; second, for the 'interclass' character, which unduly assumes the sharing by the whole society of the same 'mental equipment'; and third, for the alliance with the

quantitative and serial approach, which, all together, reifies the contents of thought and attaches itself to the most repetitive expressions and ignores singularities. Historians were thus invited to privilege individual appropriations more than statistical distributions, to understand how an individual or a community interpreted, according to its own culture, ideas and beliefs, and texts and books circulating in their society.

In 1990, in the book with the provocative title, *Demystifying Mentalities*, Geoffrey Lloyd, historian of Greek philosophy and science, hardened the indictment once again. The criticism lies within the two essential principles of the 'history of mentalities': on the one hand, allocating to a whole society a stable and homogenous set of ideas and beliefs; on the other hand, considering that all thoughts and all conducts of an individual are governed by a unique mental structure. The two operations are the very condition allowing a mentality to be distinguished from another and permitting the identification within each individual of the mental tool shared with his contemporaries. However, such a way of thinking erases, in the repetitions of the collective, the originality of each singular expression and it encloses within an artificial coherence the plurality of belief systems and ways of reasoning that a same group or a same individual can successively mobilize.

Lloyd therefore proposed to substitute for the notion of 'mentality' that of 'styles of rationality' whose use depends directly on the contexts of discourse and the domains of experiences. Each of them lays down their own rules and conventions, defines a specific form of communication, and supposes particular expectations. This is why it is quite impossible to bring back the plurality of methods of thinking, knowing, and arguing to a homogenous and unique mentality.

The case was well pleaded but is it really justified? On the one hand, the history of mentalities did not only detain the single globalizing definition of the notion, as it inherited it from Lévy-Bruhl (1922), author of *La Mentalité primitive, ou des psychologues* (Charles Blondel, Jean Piaget, and Henri Wallon). If Lucien Febvre surely was tempted by the 'interclass' definition of mentality – in particular in *Le Problème de l'incroyance au XVIe siècle* – and Philippe Ariès after him, Marc Bloch and Robert Mandrou were very attentive to the social differences which command, in a same society, different ways of thinking and feeling, or diverse visions of the world.

On the other hand, French historians have not always ignored the possible presence, within the same individual, of several mentalities, distinct or even contradictory. Le Goff strongly expresses it: "The coexistence of several mentalities at a same time and in a same spirit is one of the delicate but essential elements of the history of mentalities. Louis XI, who, in politics, revealed a modern 'Machiavellian' mentality, in matters of religion revealed a very traditional superstitious mentality."

The critical examination of the contributions and limits of the 'history of mentalities' should neither reduce the diversity of it nor simplify the expressions of it.

From the History of Mentalities to Cultural History

The path of the history of mentalities is, therefore, made up of paradoxes. While it claimed to be clearly differentiated from other historical practices, it was never able to define clearly and unanimously its objects, methods, and concepts. It was during the years when it underwent the most criticism and when expression itself retreated for the benefit of other categories that the multiplication of works which explore the topics it designated were seen. It is no doubt due to the plasticity of its definition and the diversity of its uses that this history, omnipresent and inaccessible, was able to durably characterize a form of *historiographical* work. It is, therefore, no longer possible today, in a time when the notion of 'cultural history' has become dominant, even sometimes supreme. This is what must now be examined.

An Impossible Definition

The difficulty in defining cultural history lies fundamentally in the even larger difficulty in defining what the object 'culture' itself is. The innumerable meanings of the term can be diagrammatically distributed into two families of meaning: one which designates the works and gestures which, in a given society, avoid economical or symbolical urgencies of daily life and are submitted to an aesthetic or intellectual judgment, and one which aims at ordinary practices through which a community, whichever community it may be, lives and thinks out its relations to the world, to others, or to itself.

The first order of meanings leads to building the history of texts, works, and cultural practices like a history with two dimensions, as Schorske suggests: "The historian seeks to locate and interpret the artefact temporally in a field where two lines intersect. One line is vertical, or diachronic, by which he establishes the relation of a text or a system of thought to previous expressions in the same branch of cultural activity (painting, politics, etc). The other is horizontal, or synchronic; by it he assesses the relation of the content of the intellectual object to what is appearing in other branches or aspects of a culture at the same time" (Schorske, 1979). We must, therefore, consider each cultural production in the history of its genre, discipline, or field as well as in context of its relationships with the aesthetic or intellectual productions and the cultural practices which are contemporary to it.

The latter leads to the second family of definitions of culture. It strongly relies on the meaning that symbolical anthropology gives to notion – and in particular Geertz: "The culture concept to which I adhere […] denotes an historically transmitted pattern of meanings embodied in symbols, a system of inherited conceptions expressed in symbolic forms by means of which men communicate, perpetuate, and develop their knowledge about and attitudes towards life" (Geertz, 1973). It is, therefore, the entirety of languages and of the symbolical actions of a community which constitute its culture. From there, for historians, the attention is transferred to collective expressions where a cultural system is expressed in a paroxysmal way: rituals of violence, rituals of passages, carnivalesque festivals, etc. (Davis, 1975; Darnton, 1984).

What the different approaches try to consider today is the paradoxical articulation between a difference – the one by which all societies, in varying methods, separated a field characterized by particular experiences and delights – and subordination – those which make the aesthetic and

intellectual invention possible and intelligible by noting it in the social world and in the symbolical system particular to a time and a place.

Plurality of Practices, Common Questions

According to historiographical traditions and heritage, cultural history favored different objects, fields, and methods. Making an inventory of it is an impossible task – and partially futile as it would cause repetition of the assessments presented in the numerous articles of this encyclopedia. More significant, perhaps, is the pinpointing of the common questions to these very diverse approaches.

A first stake is the necessary articulation between singular works and common representations. There are several ways to conceive this: by concentrating on the particularities of each social space where works develop and circulate (guilds, the court, the academies, the market), by situating them in relation to texts and ordinary practices with which they are in 'negotiation' (to use a term which is dear to 'New Historicism'), or, using Elias's method, by understanding how aesthetic conventions refer to psychological economy and structure of personality in a time and a place.

A second question, very widely shared, lies in the relationships between popular and learned culture. The ways of perceiving them can be dealt with by using two large models of description and interpretation. The first, trying to abolish all forms of cultural ethnocentrism, treats popular culture as a coherent and independent symbolical system, which is organized according to irreducible to logic, that of well-read culture. The second, concerned remembering the existence of relationships of domination and inequalities of the social world, understands popular culture by its subordination and its weaknesses compared to the culture of the dominant. Therefore, on the one hand, popular culture is considered as a symbolical independent system, enclosed within itself, and on the other, it is entirely defined by its distance opposed to cultural legitimacy. For a long time, historians fluctuated between these two perspectives. Then, the work carried out on religion or literature treated as specifically popular and the construction of an opposition, repeated through time, between the golden age of a free and vigorous popular culture and the times of censorship and constraints which condemn it and dismantle it. Distinctions that are so clear are no longer accepted without doubts today, which leads us to consider all the mechanisms which cause internalization by the dominated to be of their own illegitimacy and expressions via which a dominated culture manages to save something of its symbolical coherence. The lesson is valuable for the confrontation between the elite and people in the old Europe (Ginzburg, 1966, 1976) and for the relationships between the dominated and the dominant in the colonial world (Gruzinski, 1988).

A final challenge for cultural history, whatever be the approaches and objects, lies in the articulation between practices and discourse. The questioning of ancient certainties took the form of a 'linguistic turn' which reposed on two essential ideas: (1) the language is a system of signs whose relationships themselves produce multiple and unstable meanings, beyond all intention or all subjective control; (2) 'reality' is not beyond discourse but is always built by discursive practices (Baker, 1990).

Opposed to such a position, numerous are the historians who, following the distinctions proposed by Foucault between 'discursive formations' and 'nondiscursive systems' (see Foucalt, 1969) or by Bourdieu between 'practical sense' and 'scholastic logic' (see Bourdieu, 1997), marked the difference between the logic of practices and that which governs the discursive production and which underlined the irreducibility between the reality which was (or is) and the discourse which intend to organize it, censure it, or represent it.

The fundamental object of a history attempting to recognize the way in which social actors give meaning to their practices and discourse is, therefore, found in the tension between the inventive capacities of individuals or communities and, on the other hand, the constraints and conventions which restrict – more or less tightly according to the position that they occupy in their domination relationships – what is possible for them to think, express, and do. The acknowledgment is valid for well-read works and aesthetic creations as well as for ordinary practices – which is another way of expressing the double definition of the objects of cultural history.

See also: Collective Beliefs, Sociological Explanation of; Collective Memory, Anthropology of; Collective Memory, Psychology of; Cultural Psychology; Culture, Sociology of; Intellectual History; Psychohistory.

Bibliography

Ariès, P., 1964. L'enfant et la vie familiale dans l'Ancien Régime. Plon, Paris.
Ariès, P., 1975. Essais sur la mort en Occident du Moyen Age à nos jours Paris.
Ariès, P., 1977. L'homme devant la mort. Seuil, Paris.
Baker, K.M., 1990. Inventing the French Revolution: Essays on French Political Culture in the Eighteenth Century. Cambridge University Press, Cambridge, UK.
Bourdieu, R., 1997. Meditations Pascaliennes. Editions du Seuil, Paris.
Bourke, J., 2006. Fear: A Cultural History. Shoemaker Hoard, Emeryville, CA.
Burke, P., 2004. What Is Cultural History? Polity Press, Cambridge, UK.
Burleigh, M., 2009. Blood and Rage: A Cultural History of Terrorism. Harper, New York.
Darnton, 1984. The Great Cat Massacre and Other Episodes in French Cultural History. Basic Books, New York.
Davis, N., 1975. Society and Culture in Early Modern France. Stanford University Press, CA.
Figes, O., 2002. Natasha's Dance: A Cultural History of Russia. Metropolitan Books, New York.
Foucalt, C., 1969. L'Archéologie du savoir. Gallimard, Paris.
Geertz, 1973. The Interpretation of Culture. Basic Books, New York.
Ginzburg, C., 1966. I Benandanti. Stregoneria e culti agrari tra Cinquecento e Seicento. Giulio Einaudi editore, Torino.
Ginzburg, C., 1976. Il formaggio e i vermi. Il cosmo di un mugnaio del'500. Giulio Einaudi editore, Torino.
Green, A., 2008. Cultural History. Palgrave Macmillan, Basingstoke.
Gruzinski, S., 1988. La colonisation de l'imaginaire. Sociétés indigènes et occidentalisation dans le Mexique espagnol, XVIe–XVIIIe siècle. Gallimard, Paris.
Halttunen, K., 2008. A Companion to American Cultural History. Blackwell Pub., Malden, MA.
Jacobs, A., 2008. Original Sin: A Cultural History. HarperOne, New York.
Le Goff, J., 1974. Les mentalités. Une histoire ambiguë. In: Le Goff, J., Nora, P. (Eds.), Faire de l'histoire, vol. III. Gallimard, Paris, pp. 76–94.
Mandrou, R., 1961. Introduction à la France moderne, 1500–1640. Essai de psychologie historique. Albin Michel, Paris (re-edited in 1998 with an Afterword by Cottret, M., Joutard, P., Lecuir, J.).
Mandrou, R., 1968. Magistrats et sorciers dans la france du XVIIe Siècle. Plon, Paris.

McLaren, A., 2007. Impotence: A Cultural History. University of Chicago Press, Chicago.
Rietbergen, P.J.A.N., 2006. Europe: A Cultural History. Routledge, London.
Rojas, C., 2010. The Great Wall: A Cultural History. Harvard University Press, Cambridge, MA.
Schorske, C., 1979. Fin-de-siècle Vienna. Politics and Culture. Random House, New York.
Steele, V., 2001. The Corset: A Cultural History. Yale University Press, New Haven.
Venturi, F., 1970. Utopia e riforma nell'Illuminism. Einaudi, Tonno, Italy.
Vernant, J.-P., 1965. Mythe et pensée chez les Grecs. Étude de psychologie historique. Maspero, Paris.
Vovelle, M., 1973. Piété baroque et déchristianisation en Provence au XVIIIe siècle. Plan, Paris.

Cultural Influences on Interpersonal Relationships

Patricia MR Mosquera, Wesleyan University, Middletown, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by K. Takahashi, volume 5, pp. 3127–3130, © 2001, Elsevier Ltd.

Abstract

Individuals relate to others in ways that reflect the values, norms, beliefs, and practices of their cultures. This article provides an overview of research themes and findings on the role of culture in interpersonal relationships. This research is typically guided by theoretical frameworks focused on individualism (and independent selves), collectivism (and interdependent selves), or honor, and is also characterized by the use of multiple methodological approaches (e.g., emotional narratives, experiments, focus groups, surveys). The reviewed research reveals profound cultural differences in various aspects of interpersonal relationships, including beliefs about friends and enemies, relationship formation and dissolution, conflict, social support seeking, and the motivation to maintain a positive image in the eyes of close others.

Culture is a system of values, beliefs, norms, and practices that is shared and maintained by a group of individuals. Individuals form, develop, and dissolve relationships within the social contexts provided by their cultures. This cultural embeddedness of relationships implies that the ways individuals relate to other people reflect the shared values, beliefs, norms, and practices of their culture. Some cultures emphasize independence and autonomy, whereas other cultures emphasize connectedness and interdependence. Yet, other cultures value the protection of reputation and collective honor. Research in social and cultural psychology has systematically examined how culture shapes interpersonal relationships. This article identifies important research themes within the field of culture and interpersonal relationships and provides an overview of this research's most important findings.

Value Orientations and Models of Self and Relationships across Cultures

Research on culture and interpersonal relationships has typically compared individualistic and collectivistic cultures. Individualism and collectivism have been conceptualized as distinctive cultural value orientations (Hofstede, 2001; Kim et al., 1994). Although some of the earlier studies on culture measured individualism and collectivism as opposite ends of the same continuum, contemporary research generally treats them as independent value orientations. This methodological change is important as the latter approach provides a more complex and nuanced view of the influence of culture on interpersonal relationships. The implication of treating individualism and collectivism as opposite ends of the same continuum is that cultures can be either individualistic *or* collectivistic. However, aspects of individualism and collectivism can coexist within the same culture (Jansz, 1991). Individualism implies an independent model of self and relationships, whereas collectivism implies an *inter*dependent model of self and relationships. Independence and interdependence have been defined as distinctive sets of psychological tendencies to view oneself and one's social relationships (Markus and Kitayama, 1991). Cultural differences in independence and interdependence are conceptualized as relative, rather than absolute. In other words, the two models of self and relationships are not mutually exclusive. This means that individuals can endorse both cultural models as it is often the case with bicultural ethnic minorities (e.g., Latino as in the United States or Moroccan-Dutch in the Netherlands), or individuals who have been raised in both individualistic and collectivistic cultures.

Individuals who have a chronic interdependent self-concept (or interdependent selves) tend to define themselves in terms of their social roles (e.g., father, daughter, teacher) and relationships more than in terms of permanent and trans-situational traits (e.g., being extrovert or introvert). In contrast, individuals with a predominantly independent self-concept (or independent selves) tend to think of themselves as having distinct boundaries and traits that make them unique and different from other individuals. In addition, independence and interdependence involve distinctive motivational tendencies in interpersonal relationships (Markus and Kitayama, 1991; Heine and Butchel, 2009). In particular, independent selves are typically more motivated to maintain their individuality, autonomy, and uniqueness in relationships, whereas interdependent selves are typically more motivated to maintain connectedness, cooperation, and harmony in their relationships with others. Interestingly, a motivational emphasis on uniqueness among independent selves creates a psychological need for maintaining a positive view of the self in interpersonal relationships. For example, self-enhancement biases (e.g., engaging in downward comparisons to feel better about oneself) are more prevalent among independent than among interdependent selves. Interestingly, this cultural difference suggests that interdependent selves are likely to engage more freely in self-evaluation and self-improvement in interpersonal relationships.

Furthermore, there is an additional key difference in how independent and interdependent selves view interpersonal relationships. In particular, interdependent selves are more inclined to view relationships as ends in themselves. Thus, interdependent selves tend to have a more communal, collectivistic, or relational approach to interpersonal relationships compared to independent selves. For example, interdependent selves care more about preserving a good image in the eyes of others, being attentive to others' needs

and perspectives, and fulfilling social roles' expectations. In contrast, independent selves view relationships more as interpersonal spaces in which they can express their uniqueness and distinctiveness. As a consequence, independent selves view relationships as more transactional and less relational than interdependent selves.

Friendships and Culture

The Cultural Meaning of Friends and Enemies

Research by Adams and colleagues has revealed intriguing cultural differences in the meaning of friends and enemies. This research has compared European-American culture with Ghanaian culture. The former culture emphasizes individualism and an independent model of self and relationships, whereas the latter culture emphasizes collectivism and an interdependent model of self and relationships (Adams, 2005).

In one of their cross-cultural studies, Adams and Plaut (2003) interviewed European-American and Ghanaian student and nonstudent adults about friendship networks and the meaning of friends. Compared to Ghanaians, European-Americans reported a larger network of friends and emotional support as a key feature of friendships. Interestingly, Ghanaians more often reported that having too many friends could be problematic because friendships have both advantages and disadvantages. For example, Ghanaian participants more often reported that friends might reveal secrets or damage one's reputation. Taken together, these findings reveal three profound differences in the cultural meaning of friendships. First, European-Americans' larger friendship networks and Ghanaians' greater emphasis on the negative side of friendships suggest that European-Americans viewed friendships as less taxing and as involving less interpersonal obligations than Ghanaians did. Second, European-Americans' greater emphasis on emotional support suggests that they valued self-disclosure and self-expression in friendships more than Ghanaians did. Third, Ghanaians' greater emphasis on potential betrayals (e.g., friends revealing secrets) suggests that friends who may want to harm the self are a part of one's social network in Ghanaian culture. This cultural difference motivated Adam and colleagues' research on *enemyship* in interpersonal relationships.

Enemyship is defined by Adams (2005) as an interpersonal relationship marked by hatred and malice. In three studies using multiple methods, Adams (2005) showed that cultural differences in the relational proximity of enemies are grounded in independent and interdependent models of self and relationships. In the first study, Ghanaian and European-American nonstudent and student adults were interviewed about their conceptions of enemyship. Compared to European-American participants, Ghanaian participants reported having more enemies and being more often the target of enemyship. Moreover, Ghanaians described a person who claims to have no enemies as naïve, whereas European-Americans described a person who is suspicious about having enemies as abnormal (e.g., the person may be 'paranoid'). In addition, European-Americans more frequently mentioned that a person could *choose* whether they have enemies or not. These results were replicated in a second study using a survey methodology. In addition, a third experimental study among Ghanaian students examined the cause of Ghanaians' conception of enemyship. This study showed that Ghanaians' enemyship beliefs are based on an interdependent model that views enemies as an unavoidable aspect of interpersonal relationships.

Choice and Friendships: On Relational Mobility

The studies by Adams and colleagues (Adams, 2005; Adams and Plaut, 2003) showed that cultures differ in beliefs about choice in interpersonal relationships (e.g., how much a person can choose friends or enemies). These cultural differences have important implications for how interpersonal relationships are formed and dissolved. Research on relational mobility has examined the formation and dissolution of friendships across cultures.

Relational mobility has been defined as the number of available opportunities to form new relationships in a particular society (Schug et al., 2010). Research on relational mobility has typically examined differences between Japan and the US. Although US society is ethnically and culturally diverse, research on relational mobility has typically made predictions about relationship formation for US society as a whole, and not for specific cultural or ethnic groups within the United States (e.g., European-Americans, African-Americans, Latina/os). However, the majority of US participants in relational mobility studies are of European-American descent (Schug et al., 2009, 2010). US society is typically described as high on relational mobility due to its emphasis on an independent model of self and relationships. This cultural model facilitates the dissolution of friendships and the formation of new friendships because of its emphasis on personal choice. In contrast, Japanese society's emphasis on an interdependent model of self and relationships implies that individuals see friendships as more binding. Although relational mobility has been conceptualized as a characteristic of societies' social networks (i.e., number of available opportunities), it is typically measured as an individual difference variable, i.e., as the extent to which individuals *perceive* that their environment offers them opportunities to form new relationships (i.e., relational mobility scale by Schug et al., 2009).

Schug and colleagues have tested these predictions in several studies with Japanese and US student samples. For example, one study showed that US participants rated the similarity between themselves and their closest friend higher than Japanese participants did. Moreover, US participants also scored higher on relational mobility. In addition, the cultural difference in ratings of self–friend similarity was explained by cultural differences in relational mobility. In other words, US participants perceived a greater self–friend similarity than Japanese participants did because they believed that their society offers them more opportunities to form new relationships. According to Schug et al. (2009), this finding implies that US participants were able to choose friends who are more similar to them because their society is high on relational mobility. Although both Japanese and US participants preferred to have friends who are similar to them, Japanese participants could not realize their preference due to their society's low relational mobility.

In other studies, Schug et al. (2010) examined the relationship between relational mobility and self-disclosure in interpersonal relationships. Although there were no cultural differences in self-disclosure to family members, US participants reported disclosing vulnerable information about the self to friends (e.g., revealing one's greatest worry or secrets) more often than Japanese participants did. Furthermore, greater self-disclosure in friendships was positively associated with perceptions of high relational mobility in one's immediate environment (e.g., neighborhood, school). Moreover, these studies found that US participants – who perceived higher relational mobility in their immediate environments – self-disclosed more to friends because they believed that sharing vulnerable information about themselves would strengthen their friendships.

In sum, societal differences in relational mobility translate into differences in relationship formation and dissolution. In high-relational-mobility societies, friendships are seen as reflecting personal choice and individuals dissolve and form friendships more easily than in low-relational-mobility societies. According to Yuki and Schug (2012), friendships are more fragile in high-relational-mobility societies, which explains the strengthening function of self-disclosure to friends in this type of society. In low-relational-mobility societies, friendships are more stable and firmly embedded in individuals' social networks, which suggests fewer opportunities to form new friendships.

Face in Interpersonal Relationships across Cultures

Psychological research on face and culture is based on Ervin Goffman's seminal work on face (Goffman, 1959, 1967). Goffman was particularly interested in individuals' behavior when they "enter the presence of others" (Goffman, 1959: p. 1). Goffman defined face as how individuals want others to think of them in social interactions. When interacting with others, individuals typically want to maintain a positive face and one that is in line with the personal attributes that are valued in their culture. Face is therefore located in the moment-to-moment interactions between two or more individuals. This implies that face is an emergent and dynamic process in interpersonal relationships as the face that one wants to create, maintain, or protect may change across relationships (Goffman, 1959, 1967).

Importantly, Goffman did not conceptualize the maintenance or protection of face simply as a strategic form of impression management. On the contrary, individuals are attached to the face they wish to maintain or protect. For this reason, facework is necessary to prevent face loss. Goffman defined facework as an individual's verbal and nonverbal behaviors aimed at maintaining or protecting a particular face in a social interaction. According to Goffman, there are two facework orientations or perspectives: a *defensive* orientation focused on the protection of one's face and a *protective* orientation focused on the protection of another person's face. Although some interactions may be dominated by one of the two perspectives, Goffman expected individuals to often take both perspectives simultaneously in their social interactions. In cultural psychology, research based on Goffman's conceptualization of face has examined (1) the factors that affect the perception of face threats and (2) the face in the context of interpersonal conflict.

Threats to Face

The work by Lin and Yamaguchi most clearly follows Goffman's conceptualization of face as an emerging process of social interactions. Indeed, Lin and Yamaguchi defined face as the positive impressions that individuals want to "claim, maintain, or enhance in the presence of others" (2011a: p. 120). Their work provides an in-depth examination of the Japanese indigenous concept of '*mentsu*,' which can be translated as an individual's 'social images in public' (2011a: p. 121). For example, Lin and Yamaguchi recently examined the factors that influence individuals' perception of threats to '*mentsu*' (2011a). A first study among nonstudent and student adults collected 226 descriptions of threats to '*mentsu*.' Interestingly, 97.35% of the participants' narratives referred to others being present in the situations that threaten '*mentsu*.' In addition, 50.44% of the narratives also referred to the participants' failure to live up to social roles' expectations, this category being especially frequent in nonstudent adults' descriptions.

These results were replicated in a second experimental study that manipulated social role salience and publicity in a series of vignettes depicting threats to '*mentsu*.' In particular, a sample of Japanese nonstudent adults reported that the vignettes' protagonist would feel more face loss (1) in public situations and (2) in situations where the protagonist violated social role expectations. Taken together, these findings support Goffman's definition of face as a consequence of social interactions (i.e., public situations). Additionally, these studies show that face is strongly tied to social roles – an idea already advanced by Goffman (1967) – and suggest that this association becomes stronger or more important as individuals get older and acquire more social roles (e.g., when individuals enter the work force).

Furthermore, Lin and Yamaguchi have also examined the emotional consequences of face threats to self and others. In a diary study, they asked Japanese nonstudent adults to answer questions about events related to one's or others' face twice a week during a 10-week period (Lin and Yamaguchi, 2011b). Participants were asked to report how much their own and/or others' face was lost or maintained. Additionally, participants completed a measure of how they generally felt on the days they completed the survey (e.g., happy, upset). Lin and Yamaguchi (2011b) found that the maintenance of one's face in everyday social interactions was associated with more intense joy, increased self-esteem, and less intense feelings of depressiveness. What's more, participants also experienced increased joy when others' face was maintained in their everyday social interactions. Thus, individuals can care about the protection of others' face as much as they care about the protection of their own face.

In a cross-cultural study among Hong Kong Chinese and a diverse sample of US student adults, Liao and Bond (2011) examined the role of power distance in face threat perception (or face loss). Power distance is a cultural value orientation that refers to the extent to which hierarchy and power differences are accepted in a particular culture (Hofstede,

2001). Liao and Bond (2011) proposed that power distance shapes the meaning of perceived power differences in interpersonal relationships. In high-power-distance cultures, power in interpersonal relationships is likely to refer mainly to one's position and associated social role expectations in a relationship. By contrast, in low-power-distance cultures, power is more likely to be seen as reflecting each partner's agency and control in the relationship. These predictions were examined by asking participants to recall an event in which they were harmed by another person with whom they had an ongoing relationship (e.g., a friend). As expected, perceptions of relationship power played a different role in face loss among Hong Kong Chinese and US participants. In particular, US participants associated face loss with a perceived loss of power in the relationship with the person who harmed them to a greater extent than Hong Kong Chinese participants did. Hong Kong Chinese participants, in contrast, associated face loss with a perceived norm violation by the person who harmed them (i.e., this person did not act in accordance with what is expected of them) to a greater extent than US participants did. Thus, the cross-cultural study by Liao and Bond (2011) shows that the antecedents of face threat perception vary across cultures.

Face in Interpersonal Conflict

Research by Ting-Toomey, Oetzel, and colleagues has examined face loss in the context of interpersonal conflict. Based on the work by Goffman (1959, 1967), Ting-Toomey and colleagues developed face negotiation theory (Ting-Toomey and Oetzel, 2001). Although face negotiation theory distinguishes between concerns for one's face, others' face, and mutual face (i.e., a concern for protecting one's and others' face simultaneously, or the face of a relationship), the majority of research on face and conflict has focused to date on a concern for one's and others' face. Empirical studies based on face negotiation theory have typically examined face concerns and facework behavior in the context of autobiographically recalled experiences of interpersonal conflict, or in response to scenarios depicting interpersonal conflict situations with another person of the same biological sex and cultural background as the participant. On the basis of these studies, Oetzel, Ting-Toomey, and colleagues have identified 13 different types of facework behaviors that can be grouped into three types of conflict management styles: dominating (e.g., aggression, defend), integrating (e.g., apologize, respect, private discussion), and avoiding (e.g., avoid, pretend; Oetzel et al., 2000).

Furthermore, face negotiation research has shown that the intensity of a concern for one's face or others' face varies across individualistic and collectivistic cultures. In particular, a concern with the protection of one's face in interpersonal conflict has shown to be more intense among individualistic student adult samples (e.g., US participants, Germans), whereas a concern with the protection of others' face in interpersonal conflict has shown to be more intense among collectivistic student adult samples (e.g., Japanese, Chinese participants; Oetzel et al., 2008). Moreover, Ting-Toomey, Oetzel, and colleagues have also examined the role of face concerns and facework in conflicts with different relationship partners. For instance, Oetzel et al. (2003) examined the role of face concerns and facework in interpersonal conflicts with parents or siblings. Student adult samples from Germany, Japan, Mexico, and the United States (a diverse sample that included European-Americans, Latino/as, African-Americans, Asian-Americans, and Native Americans) were asked to recall a conflict with either a same sex parent or a same sex sibling. Interestingly, participants reported using different facework behaviors in conflicts with parents or siblings. Participants who described conflicts with a parent reported being more likely to be respectful and less likely to use aggression, pretense, and third party compared to participants who reported conflicts with siblings. In addition, this study showed that an independent self-concept was positively associated with a concern with one's face and dominating facework behavior whereas an interdependent self-concept was positively associated with a concern for others' face and integrating and avoiding facework behavior in conflicts with both parents and siblings.

The Cultural Meanings of Social Support

A concern for face has also been examined in the context of social support seeking in interpersonal relationships. This research, however, has not given a central place to face but has rather examined face as one of the variables that can explain cultural differences in *explicit* social support seeking. Explicit social support seeking refers to explicitly asking for advice, emotional comfort, or practical help when dealing with a stressor (Kim et al., 2008).

In a key series of studies, Kim and colleagues manipulated a focus on relationships versus a focus on the self to examine its effects on explicit social support seeking (Kim et al., 2006). Asian-American and European-American student adults were randomly allocated to a self-goals, in-group, out-group, or control condition. In-group and out-group were treated as proxies for close and distant relationships, respectively (Study 2; Kim et al., 2006). Depending on the condition, participants were asked to write about their goals, or the goals of an in-group or out-group. In addition, participants described the most important stressor they were facing and completed measures of coping strategies and expectations of a successful resolution of the stressor. The findings are intriguing. European-Americans sought social support and believed they would be successful in dealing with the stressor to the same extent across conditions. In contrast, Asian-Americans primed with in-group goals sought social support *less* and believed they would be less successful in dealing with the stressor than their counterparts in the self-goals and out-group conditions. A third experimental study replicated these findings and sought to explain these cultural differences. This third study also included a Latino/a student adult sample (Study 3; Kim et al., 2006). Importantly, this study showed that Latina/o and Asian-American participants sought social support less because they were more concerned than European-American participants were about the potential negative effects of seeking social support for their interpersonal relationships. For example, Latino/a and Asian-American participants reported a stronger motivation to

avoid face loss and preserve group harmony. Taken together, these findings indicate that explicit social support seeking has a more negative meaning in interpersonal relationships in collectivistic than in individualistic cultures.

Given the multiple benefits of social support (e.g., faster recovery from health problems, Kim et al., 2008), Taylor and colleagues have also examined the effectiveness of *implicit* social support seeking in collectivistic cultures. In a study among Asian-American and European-American student adults, they compared the effectiveness of explicit vs implicit social support in response to a stressful experimental task (e.g., arithmetic and preparation of a speech; Taylor et al., 2007). The implicit support prime consisted of writing about a group of people the participants felt close to, whereas the explicit support prime consisted of writing a letter to close others asking for advice and support for the stressful tasks. As expected, Asian-Americans in the implicit prime condition experienced less stress (as measured by self-report and cortisol levels) than their counterparts in the explicit prime condition. The reverse pattern of results was found for European-Americans. Thus, thinking about one's close relationships is psychologically and physiologically more beneficial than explicitly asking for support among Asian-Americans. These results therefore suggest that implicit social support seeking has a more positive meaning in interpersonal relationships in collectivistic than in individualistic cultures.

A study by Chang extends research on the cultural meaning of social support by examining differences *between* collectivistic cultures (Chang, 2014). Understanding how social support may be practiced differently in different collectivistic cultures is key to the development of theory on different forms of collectivism. Chang's work focuses on two specific forms of collectivism: familism and filial piety. Familism is a Latino/a value that emphasizes emotional closeness, joint decision making within the family, and loyalty toward family members. Filial piety is especially important among Asian-Americans. Caring and respecting one's elders and the protection of face are key cultural values associated with filial piety. On the basis of these different forms of collectivism, Chang (2014) advanced that self-disclosure of emotional distress may be seen as equally negative among Latino/as and Asian-Americans, but for different reasons. Asian-Americans may see self-disclosure as potentially leading to face loss as family members – especially elders – may form a negative impression of the self. In contrast, self-disclosure among Latino/as may be seen as being emotionally taxing for family members. Chang examined these hypotheses in a qualitative study using focus groups among Latina/o and Asian-American student adults.

In each focus group, the participants were asked to describe several topics related to social support (e.g., who they asked for social support when facing stressors; preferred coping strategies). Results revealed cultural similarities as well as culture-specific patterns of stressors and social support seeking. In terms of cultural similarities, academic problems were the most frequently mentioned stressors by all participants. With regards to cultural differences, both Asian-Americans and Latino/as preferred to be self-reliant than share problems with close others, but for different reasons. Asian-Americans indicated that they wanted to be self-reliant because of being worried about face loss, in particular, giving the impression of being weak and not capable of dealing with problems. Consistent with this finding, Asian-American participants also mentioned that their parents would expect them to show emotional restraint and endurance when dealing with stressors more frequently than Latina/o participants did. In contrast, Latina/o participants reported being self-reliant in the face of stressors because they wanted to maintain harmony and minimize worry within the family. Furthermore, Latina/o participants explicitly mentioned the protection of their family's honor as a reason for their self-reliance (i.e., not disclosing something about the self that may negatively affect the family's collective reputation). Chang's research demonstrates that the in-depth study of specific forms of collectivism provides a more comprehensive and nuanced view of how collectivism and interdependence are practiced in interpersonal relationships.

Honor in Interpersonal Relationships

Honor cultures represent a particular form of collectivism (Triandis, 1989) and are characterized by the existence of honor codes, which include values and norms about appropriate behavior in interpersonal relationships (Rodriguez Mosquera et al., 2012). In honor cultures, an individual has honor when she or he is *known by others* as behaving in accordance with the values and norms of four honor codes: morality-based honor, family honor, masculine honor, and feminine honor (Rodriguez Mosquera et al., 2012). Thus, reputation is a core concern in interpersonal relationships in honor cultures. Morality-based honor and family honor are important for both men's and women's honor, whereas feminine and masculine honor are gender-specific honor codes. Morality-based honor is about maintaining a reputation for being honest and trustworthy in interpersonal relationships. Family honor is about protecting the reputation of one's family. Masculine honor is about maintaining a reputation for being though and strong in interpersonal relationships. Feminine honor is about being known as having sexual propriety (e.g., chastity, modesty). Thus, honor is complex and multifaceted.

Research on honor in cultural and social psychology has shown honor to be important in a wide variety of sociopsychological processes (e.g., in-group relations, sexuality and gender, morality), and has also examined honor across a wide range of countries and cultures (e.g., Afghanistan, Brazil, Finland, Italy, Japan, New Zealand, Russia, United Kingdom; see Rodriguez Mosquera, 2013). In the context of interpersonal relationships, research on honor has mainly focused on the emotional and relational consequences of (threats to) reputation. For example, Cohen and colleagues have examined the protection of masculine honor in interpersonal relationships among White, non-Hispanic males from Southern and Northern regions of the United States. In a series of three experiments, Cohen and colleagues measured how White, non-Hispanic Southern and Northern male student adults responded to being insulted (or not) by a male confederate (Cohen et al., 1996). Results showed that Southern males were more reactive to insults than Northern

males were. In particular, observers rated Southern male participants as more angry and Northern male participants as more amused after being insulted (Study 1; Cohen et al., 1996). Furthermore, Southern male participants were more stressed (as shown by higher cortisol levels) and more prepared for aggression (as shown by increases in testosterone levels) than Northern male participants were after being insulted (Study 2; Cohen et al., 1996). Moreover, Southern males believed to a greater extent than Northern males did that their masculine honor (i.e., their reputation as manly, strong, and tough) was damaged after being insulted (Study 3; Cohen et al., 1996). In other studies, Cohen and colleagues have used other methods (e.g., archival data, survey research) to examine masculine honor's consequences in Southern US culture (for an overview, see Nisbett and Cohen, 1996).

Cohen and colleagues' research has shown masculine honor to have negative effects on interpersonal relationships in Southern US white, non-Hispanic male culture because of its potential association with male aggression and dominance. However, the maintenance of honor among men is not necessarily associated with aggression or dominance in social relationships. Indeed, Mandel and Litt's research has shown that men's behavior in relationships can also be motivated by morality-based honor. Mandel and Litt (2013) studied the relational consequences of honor among a large sample of mostly male Canadian Force members. The participants in this study rated their peers across different ranks on attributes that reflect morality-based honor, for example, integrity, dignity, and loyalty. Interestingly, high-ranking military members were rated as *less* honorable than lower ranking military members were by both low- and high-ranking participants. This result indicates that having high status does not necessarily translate into being seen as honorable in interpersonal relationships. Most importantly, the more the male participants perceived their peers as honorable, the more connected they felt with other members of their group and the more willing they were to engage in *helping behavior*. Thus, the maintenance of honor among men has different interpersonal consequences in different contexts and cultures.

A different line of research on honor in interpersonal relationships has compared emotional and behavioral responses to devaluation across cultures. Rodriguez Mosquera and colleagues collected 149 narratives about insults in interpersonal relationships in a field study among Moroccan-Dutch, Turkish-Dutch, and ethnic Dutch student and nonstudent adults (Rodriguez Mosquera et al., 2008). The protection of honor was more important for Moroccan- and Turkish-Dutch than for ethnic Dutch. Importantly, there were no cultural differences in the types of insult reported by participants (i.e., all cultural groups reported insults to one's competence, or insults to one's place in relationships), or in the person who delivered the insult (all cultural groups reported insults delivered by close others like friends or family members).

Furthermore, there were no cultural differences in feelings of anger about being insulted. In addition, feelings of anger were associated with a motivation to punish the close other who delivered the insult through verbal attack (e.g., criticizing) for all cultural groups. Interestingly, however, cultural differences emerged in the behaviors and motivations associated with feelings of shame. Shame was associated with social withdrawal among ethnic Dutch, whereas it was associated with verbal disapproval of the close other's behavior among Turkish- and Moroccan-Dutch. Most importantly, the study showed that Turkish- and Moroccan-Dutch used verbal disapproval as an attempt to restore a positive reputation (i.e., restore their honor) in their interpersonal relationships (for an overview of other studies on threats to reputation, see Rodriguez Mosquera et al., 2012). Furthermore, this heightened concern for reputation in honor cultures also influences the experience of positive emotions. In particular, the perception that one has a positive reputation in interpersonal relationships has been shown to be associated with increased well-being among honor-oriented Indian, Pakistani, and Bangladeshi student adults, but not among less honor-oriented White British and European-American student adults (Rodriguez Mosquera and Imada, 2013).

Further, two recent studies have revealed that members of honor cultures are also concerned about protecting the reputation of their families in interpersonal relationships. Rodriguez Mosquera et al. (2014) carried out two experimental studies among Pakistani and European-American student adults on emotional and behavioral responses to insults to one's family. Family honor creates a distinctive type of interdependence in interpersonal relationships, one based on shared social image. This type of interdependence implies that individual family members are responsible for protecting the family from potential threats, like insults. In line with their greater emphasis on the protection of family honor, Pakistani participants experienced more intense anger, shame, negative feelings about themselves, and worry about their reputation than European-American participants did in response to an insult to their families. Moreover, Pakistani participants reported that they distanced themselves *from a family member* who insulted the family more than European-American participants did (these cultural differences did not emerge when the person who delivered the insult was not a family member, e.g., a friend). This finding is important as it shows the profound effects that threats to family honor can have on close relationships.

In sum, honor is shared, negotiated, challenged, and maintained in interpersonal relationships. Due to this relational nature, honor is a powerful force in our relationships with others. For example, the reviewed studies show that those who value honor care deeply about their relatives' willingness to protect the family's honor, or about how friends think of them. Moreover, honor can strengthen interpersonal bonds and increase individuals' willingness to help others when relationship partners perceive each other as honorable.

Conclusions

The ways in which individuals relate to other people are grounded in culture. Relationships exist within the contexts and meanings provided by culture. The research reviewed in this article has revealed a great deal of cultural diversity in interpersonal relationships. Cultures differ, for example, in the size of friendship networks, beliefs about what makes a good friend, concerns and behaviors involved in interpersonal

conflict, relationship formation and dissolution, social support seeking, or the role of honor and reputation in relationships. Furthermore, the research reviewed in this article has used a variety of theoretical and methodological approaches – surveys, focus groups, (field) experiments, emotional narratives, diary studies – to examine the depth and breadth of culture's influences on interpersonal relationships. Furthermore, several different cultures have been included in this research. Yet, there are cultures that are still understudied in cross-cultural research on interpersonal relationships. The future of research on culture and interpersonal relationships lies in expanding its theories and methods to include these understudied cultures. Understanding cultural diversity in interpersonal relationships and other sociopsychological processes is key to the advancement of psychology as an inclusive discipline.

See also: Collectivism and Individualism: Cultural and Psychological Concerns; Cross-Cultural Psychology; Cultural Psychology; Culture and Emotion; Values, Psychology of.

Bibliography

Adams, G., 2005. The cultural grounding of personal relationship: enemyship in North American and West African worlds. Journal of Personality and Social Psychology 88, 948–968.

Adams, G., Plaut, V.C., 2003. The cultural grounding of personal relationship: friendship in North American and West African worlds. Personal Relationships 10, 333–348.

Chang, J., 2014. The interplay between collectivism and social support processes among Asian American and Latino college students. Asian American Journal of Psychology. Advance online publication. http://dx.doi.org/10.1037/a0035820.

Cohen, D., Nisbett, R.E., Bowdle, B.F., Schwarz, N., 1996. Insult, aggression and the southern culture of honor: an "experimental ethnography". Journal of Personality and Social Psychology 70, 945–960.

Goffman, E., 1959. The Presentation of Self in Everyday Life. Doubleday, New York.

Goffman, E., 1967. Interaction Ritual. Essays on Face-to-Face Behavior. Anchor Books, New York.

Heine, S.J., Butchel, E.E., 2009. Personality: the universal and the culturally specific. Annual Review of Psychology 60, 369–394.

Hofstede, G., 2001. Culture's Consequences. Comparing Values, Behaviors, Institutions, and Organizations across Nations. Sage, Thousand Oaks, CA.

Jansz, J., 1991. Person, Self, and Moral Demands. DSWO Press, Leiden.

Kim, U., Triandis, H.C., Kagitçibasi, C., Choi, S.-C., Yoon, G. (Eds.), 1994. Individualism and Collectivism: Theory, Method, and Applications. Sage, Thousand Oaks, CA.

Kim, H.S., Sherman, D.K., Taylor, S., 2008. Culture and social support. American Psychologist 63, 518–526. http://dx.doi.org/10.1037/0003-066X.

Kim, H.S., Sherman, D.K., Ko, D., Taylor, S., 2006. Pursuit of comfort and pursuit of harmony: culture, relationships, and social support seeking. Personality and Social Psychology Bulletin 32, 1595–1607. http://dx.doi.org/10.1177/0146167206291991.

Liao, Y., Bond, M.H., 2011. The dynamics of face loss following interpersonal harm for Chinese and Americans. Journal of Cross-Cultural Psychology 42, 25–38. http://dx.doi.org/10.1177/0022022110361774.

Lin, C., Yamaguchi, S., 2011a. Under what conditions do people feel face loss? Effects of the presence of others and social roles on the perception of losing face in Japanese culture. Journal of Cross-Cultural Psychology 42, 120–124. http://dx.doi.org/10.1177/0022022110383423.

Lin, C., Yamaguchi, S., 2011b. Effects of face experience on emotions and self-esteem in Japanese culture. In: Rodriguez Mosquera, P.M., Uskul, A.K., Cross, S.E. (Eds.), The centrality of social image in social psychology. European Journal of Social Psychology 41 (4) (Special issue), 446–455.

Mandel, D.R., Litt, A., 2013. The ultimate sacrifice: perceived peer honor predicts troops' willingness to risk their lives. Group Processes and Intergroup Relations 16, 375–388.

Markus, H.R., Kitayama, S., 1991. Culture and the self: implications for cognition, emotion and motivation. Psychological Review 98, 224–253.

Nisbett, R.E., Cohen, D., 1996. Culture of Honor: The Psychology of Violence in the South. Westview, Boulder, CO.

Oetzel, J.G., Ting-Toomey, S., Yokochi, Y., Masumoto, T., Takai, J., 2000. A typology of facework behaviors in conflicts with best friends and relative strangers. Communication Quarterly 48, 397–419.

Oetzel, J.G., Ting-Toomey, S., Chew-Sanchez, M.I., Harris, R., Wilcox, R., Stumpf, S., 2003. Face and facework in conflicts with parents and siblings: a cross-cultural comparison of Germans, Japanese, Mexicans and U.S. Americans. The Journal of Family Communication 3, 67–93.

Oetzel, J.G., Garcia, A.J., Ting-Toomey, S., 2008. An analysis of the relationships among face concerns and facework behaviors in perceived conflict situations: a four culture investigation. International Journal of Conflict Management 19, 382–403.

Rodriguez Mosquera, P.M., 2013. In the name of honor. on virtue, reputation, and violence. Group Processes and Intergroup Relations 16 (3) (Special Issue).

Rodriguez Mosquera, P.M., Imada, T., 2013. Perceived social image and life satisfaction across cultures. Cognition and Emotion 27, 1132–1141. http://dx.doi.org/10.1080/02699931.2013.767222.

Rodriguez Mosquera, P.M., Tan, L.X., Saleem, F., 2014. Shared burdens, personal costs. on the emotional and social consequences of family honor. Journal of Cross-Cultural Psychology 45 (3), 400–416. http://dx.doi.org/10.1177/0022022113511299.

Rodriguez Mosquera, P.M., Liskow, M., DiBona, K., 2012. Honor and emotion. In-mind 13, 1–5.

Rodriguez Mosquera, P.M., Fischer, A.H., Manstead, A.S.R., Zaalberg, R., 2008. Attack, disapproval, or withdrawal? The role of honor in anger and shame responses to being insulted. Cognition and Emotion 22, 1471–1498. http://dx.doi.org/10.1080/02699930701822272.

Schug, J., Yuki, M., Horikawa, H., Takemura, K., 2009. Similarity attraction and actually selecting similar others: how cross-societal differences in relational mobility affect interpersonal similarity in Japan and the USA. Asian Journal of Social Psychology 12, 95–103. http://dx.doi.org/10.1111/j.1467-839X2009.01277.x.

Schug, J., Yuki, M., Maddux, W., 2010. Relational mobility explains between- and within culture differences in self-disclosure to close friends. Psychological Science 21, 1471–1478. http://dx.doi.org/10.1177/0956797610382786.

Taylor, S.E., Welch, W., Kim, H.S., Sherman, D.K., 2007. Cultural differences in the impact of social support on psychological and biological stress responses. Psychological Science 18, 831–837.

Ting-Toomey, S., Oetzel, J.G., 2001. Managing Intercultural Conflict Effectively. Sage, Thousand Oaks, CA.

Triandis, H.C., 1989. The self and social behavior in differing cultural contexts. Psychological Review 96, 506–520.

Yuki, M., Schug, J., 2012. Relational mobility: a socioecological approach to personal relationships. In: Gillath, O., Adams, G., Kunkel, A. (Eds.), Relationship Science: Integrating Evolutionary, Neuroscience, and Sociocultural Approaches. American Psychological Association, Washington, DC. http://dx.doi.org/10.1037/13489-000.

Cultural Intelligence and Competencies

Soon Ang, Thomas Rockstuhl, and Mei Ling Tan, Nanyang Technological University, Singapore

© 2015 Elsevier Ltd. All rights reserved.

Abstract

With the continuing globalization of the workplace, it is critical to understand why some people excel in intercultural contexts whereas others do not. Cultural intelligence is a person's capability to function effectively in intercultural contexts. In this article, we take stock of the growing stream of research on cultural intelligence. In particular, we review the conceptualization, measurement, and empirical evidence for the nomological network of cultural intelligence. We conclude with an eye toward the future and suggest several exciting research directions to further advance our understanding of cultural intelligence.

Theoretical Conceptualization of Cultural Intelligence

Definition

Cultural intelligence (CQ) refers to a person's capability to function effectively in culturally diverse contexts (Ang and Van Dyne, 2008; Earley and Ang, 2003). This definition of cultural intelligence as a capability emphasizes a person's potential to be effective across a wide range of intercultural contexts. Cultural intelligence differs from the capability to function effectively in a specific culture. Instead, cultural intelligence reflects a general set of capabilities that facilitate one's effectiveness across different cultural environments. In this sense, cultural intelligence is culture-free.

Cultural intelligence also differs from cross-cultural views of intelligence that emphasize the relativity of intelligence definitions depending on particular cultural and ecological contexts (Berry, 1976; Sternberg, 2004). For example, hunter-gatherers require different intelligences than agriculturalists to survive in their respective ecological environments. Therefore, the meaning of intelligence varies in each culture depending on its unique ecological context. While cultural intelligence does not refer to these culturally bound notions of intelligence, knowledge of such culturally bound views of intelligence does reflect high-cultural intelligence (specifically, cognitive cultural intelligence, as we will describe below).

Cultural Intelligence as a Multidimensional Intelligence

The conceptualization of cultural intelligence draws upon the rich history of intelligence research. Cultural intelligence builds on insights from intelligence research suggesting that intelligence is multifaceted. Integrating myriad views on intelligence, Sternberg and Detterman (1986) proposed that intelligence resides in different loci within an individual: the biology, the cognition (including metacognition), the motivation, and the behaviors.

The cultural intelligence model (Earley and Ang, 2003) draws on Sternberg and Detterman's (1986) multiple-loci conceptualization of intelligence and comprises four factors: (1) metacognitive cultural intelligence, which reflects an individual's mental capability to acquire and understand cultural knowledge; (2) cognitive cultural intelligence, which reflects an individual's knowledge about cultures and cultural differences; (3) motivational cultural intelligence, which reflects an individual's capability to direct and sustain effort toward functioning in intercultural situations; and (4) behavioral cultural intelligence, which reflects an individual's capability for behavioral flexibility in cross-cultural interactions. While the initial conceptualization of cultural intelligence did not include biological aspects of intelligence, recent work on cultural intelligence has embraced biological foundations of cultural intelligence (Rockstuhl et al., 2010).

Cultural Intelligence and Other Forms of Intelligences

Cultural intelligence is similar to social and emotional intelligence in that cultural intelligence is a form of interpersonal intelligence. Social intelligence is a broader form of interpersonal or real-world intelligence that refers to the ability to understand and manage others. Emotional intelligence refers more specifically to the ability to deal with emotions of self and others. Cultural intelligence shares similarities with social and emotional intelligence in that cultural intelligence includes the abilities to understand and manage others, as well as to deal with their emotions. However, unlike social and emotional intelligence, cultural intelligence explicitly considers the intercultural context. Understanding culturally different others requires a distinct set of abilities because of cultural variations in how people from different parts of the world express themselves verbally and nonverbally. Hence, a person who is high in emotional intelligence or social intelligence is not necessarily high in cultural intelligence. Empirical studies have shown cultural intelligence to be distinct from emotional and social intelligence. Across multiple studies (for details, see Ang et al., in press), confirmatory factor analyses showed discriminant validity between cultural intelligence and emotional intelligence – correlations between cultural intelligence and emotional intelligence ranged from 0.26 (Ang et al., 2007) to 0.82 (Ward et al., 2009). In addition, one study found cultural intelligence to be discriminant from social intelligence, with the correlation between the two constructs at 0.42 (Crowne, 2009).

Cultural intelligence is also similar to but distinct from general cognitive ability. General cognitive ability is a key predictor of performance across jobs and settings. Similarly, cultural intelligence should predict performance but more specifically in intercultural contexts (Ang and Van Dyne, 2008).

Cultural intelligence is also distinct from general cognitive ability because the latter only includes the cognitive locus of intelligence and excludes the motivational, behavioral, and biological loci. Empirical studies show negative correlations between motivational cultural intelligence and general cognitive ability, and weak correlations ranging from 0.00 to 0.11 for the other three factors of cultural intelligence (Ang et al., 2007; Ward et al., 2009). Empirical evidence also indicates that cultural intelligence has a stronger correlation with task performance in intercultural contexts than does general cognitive ability. Hence, cultural intelligence incrementally predicts performance in intercultural situations beyond cognitive ability (Ang et al., 2007; Rockstuhl et al., 2011).

Cultural Intelligence and Personality Traits

Personality traits refer to stable personal characteristics that lead to consistent patterns of cross-situational behavior (Costa and McCrae, 1992). By contrast, cultural intelligence is a set of capabilities that determine what a person is capable of doing to be effective in culturally diverse settings. Hence, personality traits and cultural intelligence are conceptually distinct. However, given that personality traits affect a person's choice of behaviors and experiences, some personality traits might relate to cultural intelligence. In line with these conceptualizations, Ang et al. (2006) demonstrated discriminant validity between the four factors of cultural intelligence and the Big-Five personality traits. We will elaborate this point below (see Section on Antecedents of Cultural Intelligence).

Cultural Intelligence and Other Cultural Competencies

Cultural competencies are an umbrella term for concepts related to intercultural effectiveness. In a recent review of cultural competence models, Leung et al. (2014) identified more than 30 cultural competence models with over 300 concepts related to cultural competence. These 300 concepts cover a broad range of personal characteristics including intercultural personality traits, intercultural attitudes and worldviews, or intercultural capabilities. Intercultural personality traits describe what a person typically does in intercultural contexts. Examples include tolerance for ambiguity or cultural empathy. Intercultural attitudes and worldviews refer to how a person perceives and evaluates experiences with other cultures. Examples include ethnocentrism or ethnorelativism. Intercultural capabilities describe what a person can do to be effective in intercultural contexts. Examples of intercultural capabilities include self-awareness or global business savvy.

Cultural competence models differ in scope. Some models combine personality traits, attitudes and worldviews, and capabilities. Other models focus on unique domains of characteristics. For example, the cultural intelligence model concerns intercultural capabilities only, while the Global Mindset Inventory (Javidan and Teagarden, 2011) embraces personality traits, attitudes and worldviews, as well as capabilities (see Leung et al., 2014).

The range and scope of different content domains covered by cultural competence models raise questions about the structural relationships between these content domains. Theories of job performance suggest that distal constructs such as personality traits and attitudes exert their effects on job performance via more proximal capabilities (Campbell et al., 1993). Hence, intercultural capabilities are more proximal predictors of performance in intercultural contexts and mediate effects of intercultural personality traits and intercultural attitudes and worldviews.

Therefore, we focus on capability models and the cultural intelligence model in particular, for the remainder of this article. We highlight the cultural intelligence model as a theoretically coherent and parsimonious framework of intercultural capabilities. Based on the multiple-loci conceptualization of intelligence, the concept of cultural intelligence is theoretically precise about what is and is not part of its construct space. The cultural intelligence concept is parsimonious in that it focuses on only four abstract factors (e.g., metacognition) rather than a vast number of narrower dimensions (e.g., self-awareness, cognitive complexity, cognitive flexibility, perspective taking, planning, checking). Capabilities from other cultural competence models can be mapped onto the cultural intelligence model. However, other cultural competence models rarely consider all four factors simultaneously and thus lack the comprehensiveness offered by the cultural intelligence model for describing the capabilities domain.

Measurement of Cultural Intelligence

Individual differences in cultural intelligence are measured using diverse methods. These methods include self-reports, observer-reports, and performance-based measures.

Self-Reports of Cultural Intelligence

Self-reported measures of cultural intelligence present respondents with a list of statements relevant to multiple dimensions of cultural intelligence (e.g., "I check the accuracy of my cultural knowledge as I interact with people from different cultures"). Respondents then rate the extent of their agreement with each statement. As a measure of perceived capability, self-reported measures of cultural intelligence reflect one's self-efficacy in cultural intelligence.

To date, most empirical research uses the 20-item four-factor Cultural Intelligence Scale (CQS) introduced by Ang et al. (2007). Van de Vijver and Leung (2009) advise that measures for use in intercultural contexts should demonstrate both factor structure validity and cross-cultural measurement equivalence. The Cultural Intelligence Scale meets both criteria. Its four-factor structure generalizes across (1) multiple student and executive samples; (2) repeated measurements using time intervals of up to 4 months; (3) multiple countries, including South Korea, Switzerland, Singapore, Turkey, and the United States; (4) culturally diverse samples; or (5) members in multicultural teams (see Leung et al., 2014).

Observer-Reports of Cultural Intelligence

Observer-reports of cultural intelligence are a fundamental source of information about a person's external manifestation of cultural intelligence and reflect a person's cultural intelligence reputation. In observer-reports, acquaintances (e.g., friends,

peers, supervisors, subordinates) summarize their perceptions of someone's cultural intelligence reputation.

Van Dyne et al. (2008) developed an observer-reported measure of cultural intelligence based on the 20-item Cultural Intelligence Scale. In an initial validation study with 142 executive MBAs, these authors found evidence for the convergent validity between self-reported and observer-reported cultural intelligence. Kim and Van Dyne (2012) further supported the predictive validity of observer-reports of cultural intelligence. In a sample of 181 working adults, observer-reports of cultural intelligence from one group of observers predicted international leadership potential as rated by another group of observers.

Performance-Based Assessment of Cultural Intelligence

Ang et al. (2014a) introduced an intercultural situational judgment test (iSJT) to measure cultural intelligence. This test presents respondents with multimedia vignettes of challenging work-related intercultural situations and asks them how they would respond to that situation. Responses are then scored in terms of how effective they resolve the situation in the vignette. The primary appeal of using multimedia over text-based vignettes lies in their greater fidelity (i.e., correspondence to real situations) due to richer portrayals of detailed cultural information (e.g., nonverbal gestures). Rockstuhl et al. (2013b) showed that the iSJT predicted supervisor-rated task performance for offshoring professionals from the Philippines. In a related study, Rockstuhl et al. (2014) showed that the iSJT predicted peer-rated task performance and interpersonal organizational citizenship behavior (OCB) in multicultural teams.

Combining Complementary Measures of Cultural Intelligence

We suggest that different measures of cultural intelligence provide complementary information. In particular, divergence between different measures of the same construct more likely reflects different but theoretically meaningful aspects of a construct (i.e., self-efficacy for self-reports vs reputation for observer-reports) instead of mere bias.

If different measures of the same construct reflect theoretically meaningful aspects instead of bias, then different measures should predict outcomes incrementally over and above each other. Research shows consistently that self-reports predict performance over and above alternative measures of the same construct. For example, in the domain of general cognitive ability, self-reported intelligence predicts academic achievement even after controlling for standardized tests of intelligence (Chamorro-Premuzic et al., 2010). Similarly, meta-analyses show that self-reported emotional intelligence incrementally predicts job performance (Joseph and Newman, 2010) over and above ability-based tests of emotional intelligence. Even in the domain of attitudes, self-reported measures predict actual behavior over and above nonconscious tests of the same attitudes (Greenwald et al., 2009).

Measures of cultural intelligence appear to follow a similar pattern. For example, Rockstuhl et al. (2014) found that self-reported cultural intelligence incrementally predicted task performance in multicultural teams beyond a situational judgment test of cultural intelligence and other intelligence, personality, and experience predictors. This finding suggests that different measures of cultural intelligence reflect theoretically meaningful aspects of a person's cultural intelligence instead of mere bias. Hence, we propose that different measures of cultural intelligence should be used conjointly to provide a complete assessment of a person.

Nomological Network of Cultural Intelligence

The cultural intelligence construct has received worldwide interest. Empirical studies on cultural intelligence have been conducted in North America, South America, Europe, the Middle East, Asia, Australia, and New Zealand. Samples include expatriates, international business travelers, foreign laborers, global domestics, and international students. Altogether, cultural intelligence studies have sampled people from or working in at least 40 different countries.

Across these samples, scholars have studied (1) antecedents of cultural intelligence; (2) outcomes of cultural intelligence; (3) cultural intelligence as a mediator; (4) cultural intelligence as a moderator; and (5) boundary conditions of effects of cultural intelligence. To date, four major reviews have integrated empirical research on cultural intelligence (see Ang et al., 2011, in press; Leung et al., 2014; Ng et al., 2012). Below, we summarize empirical findings.

Antecedents of Cultural Intelligence

Research on antecedents of cultural intelligence has focused on personality traits and international experiences. As noted above, personality traits could relate to cultural intelligence because traits are broad and relatively stable individual differences that affect choices of behaviors and experiences, which in turn can influence the development of cultural intelligence. Among the Big-Five personality traits, openness to experience relates to all four cultural intelligence factors consistently across studies (see Ang et al., in press). Openness to experience refers to a person's tendency to be creative, imaginative, and adventurous (Costa and McCrae, 1992), thus pertaining to novel situations. Similarly, cultural intelligence is a set of capabilities targeted at novel and unfamiliar intercultural situations.

In contrast to the consistent findings for openness to experience, the findings for other Big-Five personality traits are more equivocal. For example, Şahin et al. (2013) found that extraversion was the second most important personality predictor of cultural intelligence, whereas Ang et al. (2007) found that conscientiousness was the second most important predictor of cultural intelligence.

Beyond the broad Big-Five personality traits, studies have also examined narrower traits as predictors of cultural intelligence. For example, of the six subfacets of openness to experience (i.e., intellectual efficiency, ingenuity, curiosity, aesthetics, tolerance, and depth), tolerance and curiosity were the strongest predictors of cultural intelligence (Oolders et al., 2008). Other narrow traits that research has linked to cultural intelligence include: (1) need for cognition and need for closure (Imai and Gelfand, 2010); (2) traits from the Multicultural Personality Questionnaire (i.e., emotional

stability, social initiative, open-mindedness, cultural empathy, and flexibility) (Ward et al., 2009); or (3) traits from the Cross-Cultural Adaptability Inventory (i.e., flexibility/openness, emotional resilience, perceptual acuity, and personal autonomy) (Ang et al., 2007).

Regarding international experiences, scholars have examined both work-related and nonwork-related international experiences. However, findings have not been consistent across the four cultural intelligence factors (for details, see Ng et al., 2012). In one study, the number of countries someone had previously worked in related positively to metacognitive and motivational cultural intelligence. In another study, the same measure of international experience related to metacognitive, cognitive, and behavioral cultural intelligence. Another study found that the length of international work experiences related to cognitive cultural intelligence only.

For nonwork-related international experiences, the number of countries visited related positively to all four cultural intelligence factors in one study, whereas the length of stay predicted metacognitive and cognitive cultural intelligence only. Another study found that the number of educational experiences abroad related positively to both cognitive and behavioral cultural intelligence, but the number of countries visited related only to motivational cultural intelligence.

Other studies have examined the effects of specific programs and interventions on the development of cultural intelligence. For example, participating in a 4-week virtual team project with team members from five countries increased team members' motivational-, metacognitive-, and behavioral-, but not cognitive, cultural intelligence (Shokef and Erez, 2008). Another program designed international experiences based on experiential learning and social contact principles (MacNab et al., 2012). In this program, the time spent interacting with culturally diverse others predicted increases in cultural intelligence for participants. Other studies have replicated the benefits of time spent interacting with culturally diverse others for the development of cultural intelligence (for details, see Ang et al., in press).

Besides personality traits and international experiences, few antecedents of cultural intelligence have been studied. Exceptions include foreign language skills and global identity – both of which relate positively to cultural intelligence.

Outcomes of Cultural Intelligence

Accumulating research shows that cultural intelligence relates to a wide range of cognitive, affective, and behavioral outcomes in intercultural contexts. Research shows that metacognitive and cognitive cultural intelligence predict cognitive outcomes stronger than motivational and behavioral cultural intelligence. An important cognitive outcome is cultural judgment and decision-making, which refers to the quality of decisions regarding intercultural interactions (Ang et al., 2007). Across multiple samples, metacognitive and cognitive cultural intelligence predicted better cultural judgment and decision-making (Ang et al., 2007). In a similar vein, metacognitive and cognitive cultural intelligence related positively with perceived cross-border environment uncertainty, and in turn the accuracy of risk assessments in international business ventures (Prado, 2006).

By contrast, motivational cultural intelligence is the most consistent predictor of affective outcomes in international contexts. To date, the most widely studied affective outcome has been cultural adjustment of sojourners and expatriates, which includes general adjustment (i.e., adjustment to general living conditions in a foreign culture), work adjustment (i.e., adjustment to work in a foreign culture), interactional adjustment (i.e., adjustment to socializing with locals in a foreign culture), psychological adjustment (i.e., feelings of well-being and satisfaction when living in a foreign culture), and sociocultural adjustment (i.e., being able to fit in or negotiate interactive aspects in a foreign culture).

Fourteen studies have examined these cultural adjustment outcomes (see Leung et al., 2014). These studies document consistently the benefits of high-cultural intelligence on all five forms of cultural adjustment. Across these studies, motivational cultural intelligence is the strongest predictor of cultural adjustment.

Beyond cultural adjustment, other studies show that people with higher cultural intelligence (1) experience less culture shock when living abroad (Chen et al., 2011); (2) experience less emotional exhaustion when traveling internationally for business (Tay et al., 2008); (3) report greater intention to complete their expatriate assignments (Wu and Ang, 2011); and (4) report greater satisfaction with their expatriate assignments (Huff, 2013).

Affective trust in culturally diverse others is another outcome that has received growing attention. Rockstuhl and Ng (2008) showed that people are more likely to trust culturally diverse others if (1) they have higher metacognitive cultural intelligence and (2) culturally diverse others have higher behavioral cultural intelligence. Notably, cultural intelligence influences affective trust only in culturally diverse but not in culturally homogeneous dyads. These findings highlight the unique relevance of cultural intelligence in intercultural contexts. At the team level, research also shows that multicultural teams with greater average team member cultural intelligence experience greater cohesion than teams with lower average cultural intelligence (Moynihan et al., 2006). A recent study sheds some light on possible mechanisms behind these effects. In particular, American working adults with greater metacognitive cultural intelligence had greater expectations of cooperative- or relationship-oriented goals – both for themselves and others – when preparing for an interaction with Chinese counterparts (Mor et al., 2013).

Finally, research has studied a wide range of behavioral outcomes of cultural intelligence. Many of these studies use multisource designs and control for a number of alternative predictors, such as general cognitive ability, emotional intelligence, Big-Five personality, and experience-based predictors. Based on levels of specificity, we classify these outcomes broadly into general job performance (including task performance, organizational citizenship behaviors, and adaptive performance), domain-specific performance (including global leadership and negotiation), and specific demonstrated behaviors.

Ten studies show that cultural intelligence predicts task performance in different work contexts, such as global work assignments and work in culturally diverse domestic settings (for details, see Leung et al., 2014). Across these studies,

metacognitive and behavioral cultural intelligence appear to be stronger predictors of task performance than motivational and cognitive cultural intelligence. These studies use self-reported measures of cultural intelligence. Two recent studies show that a situational judgment test of cultural intelligence likewise predicts task performance (Rockstuhl et al., 2013b, 2014). While task performance remains the most widely studied performance outcome to date, studies have also shown that cultural intelligence predicts citizenship behaviors (Rockstuhl et al., 2014; Wu and Ang, 2011) as well as adaptive performance (Oolders et al., 2008). At the team level, empirical evidence suggests that average team members' cultural intelligence predicts performance of multicultural teams (Groves and Feyerherm, 2011) and creative performance both in intercultural dyads (Chua et al., 2012) and multicultural teams (Crotty and Brett, 2012).

A number of studies have related cultural intelligence to global leadership. Several qualitative studies highlight the crucial role that cultural intelligence plays in managing subordinates and offshoring vendors from different cultural backgrounds. Quantitative studies confirm the importance of cultural intelligence for global leaders. Specifically, studies show that cultural intelligence predicts (1) subordinate-rated leader performance in multicultural teams (Groves and Feyerherm, 2011); (2) peer-rated leadership emergence in multicultural teams (Rockstuhl et al., 2013a); (3) peer-rated cross-border leadership effectiveness (Rockstuhl et al., 2011); and (4) peer-rated international leadership potential (Kim and Van Dyne, 2012). At the dyadic level, cultural intelligence of the lower of two intercultural negotiation partners predicts joint profits (Imai and Gelfand, 2010).

Recent studies have also begun to illuminate more specific or proximal behaviors that culturally intelligent people exhibit. For example, nonnative English speakers with higher cultural intelligence tend to interact more frequently with native English speakers, even after controlling for the ability to speak multiple languages (Beyene, 2007). Other studies show that individuals with high rather than low metacognitive cultural intelligence more frequently engage in (1) information sharing with culturally diverse others (Chua et al., 2012), and (2) more cooperative behaviors with culturally diverse others in mixed motive or prisoner's dilemmas (Mor et al., 2013). Similarly, individuals with higher cultural intelligence engage in more information sharing and cooperative-/relationship-management behaviors in intercultural negotiations (Imai and Gelfand, 2010). Perhaps as a result of such communicative and cooperative behaviors, people with higher cultural intelligence also tend to develop larger and more culturally diverse social networks than people with lower cultural intelligence. Mirroring these individual-level results, cultural intelligence at the team-level promotes fusion teamwork behaviors, i.e., teamwork behaviors that encourage meaningful participation and coexistence of different cultures (Crotty and Brett, 2012).

Cultural Intelligence as a Mediator

As noted above, personality and international experience are widely studied antecedents of cultural intelligence. Consequently, research has tested cultural intelligence as a mediator of the effects of these distal predictors on outcomes such as cultural adjustment, job performance, and global leadership (for details, see Ang et al., in press). Empirical studies show that cultural intelligence mediates the effects of personality traits. For example, cultural intelligence mediated the effects of Multicultural Personality traits on general adjustment in a sample of international students in New Zealand. In other studies, cultural intelligence mediated the effect of openness to experience on (1) job performance of expatriates in Malaysia and (2) adaptive performance of undergraduate exchange students in New Zealand.

Cultural intelligence also mediates effects of international experience. In a study of South Korean expatriates, cultural intelligence mediated the effects of previous international experience and predeparture cross-cultural training on cross-cultural adjustment. In a study of culturally diverse MBA students, cultural intelligence also mediated the effects of international experience on international leadership potential.

Finally, one study showed that cultural intelligence mediated the effects of a three-way interaction between home-country identity, host-country identity, and global identity on leadership emergence in multicultural teams.

Cultural Intelligence as a Moderator

Two studies have examined cultural intelligence as a moderator. In a study of senior expatriate leaders in various European countries, higher leader cultural intelligence strengthened the positive relationship between leader's transformational leadership style and organizational innovation (Elenkev and Manev, 2009).

Another study found that cultural intelligence moderated the effects of perceived cultural diversity on voice instrumentality (i.e., perceptions that voicing behaviors will lead to desired organizational changes), which in turn affected actual voice behaviors (Ng et al., 2011). In particular, although cultural diversity lowered voice instrumentality for individuals with low-cultural intelligence, it increased voice instrumentality for individuals with high-cultural intelligence.

Boundary Conditions of the Effects of Cultural Intelligence

More recent studies refine theoretical arguments about effects of cultural intelligence and examine their boundary conditions. Such studies have examined boundary conditions both for effects of international experience on cultural intelligence and for effects of cultural intelligence on outcomes.

In light of the inconsistent effects of international experience on cultural intelligence, scholars have advanced a number of boundary conditions of these effects (for details, see Ang et al., in press). One study found that positive effects of work-related international experience on cultural intelligence were stronger for people with a divergent rather than a convergent learning style. Similarly, another study found that positive effects of nonwork international experiences were strongest when people had high mastery-goal orientations and low performance-avoidance orientations in intercultural contexts. Other studies have found that effects of intercultural contact on cultural intelligence were stronger for (1) people with greater self-efficacy; (2) majority rather than minority members; and (3) people who had their first rather than subsequent intercultural service learning experiences.

Recent research also demonstrates the crucial role of cultural capital (i.e., international education and international experiences of one's parents) in fostering positive relationships between international experience and cultural intelligence. Two related studies found that cultural capital strengthened the indirect effects of international experience on global leadership via cultural intelligence. In particular, international experience related positively to cultural intelligence only when cultural capital was high. Cultural intelligence in turn predicted (1) supervisor-rated international military officer potential, and (2) peer-rated leadership emergence in multicultural teams.

Research on boundary conditions of the effects of cultural intelligence on outcomes has hypothesized and tested moderators that both attenuate and strengthen effects of motivational cultural intelligence. For example, G. Chen et al. (2010) found that subsidiary support (i.e., the extent to which the subsidiary helps expatriates adapt to their assignments and provides them with career and financial support) weakened the effect of motivational cultural intelligence on work adjustment and subsequently performance. Likewise, cultural distance (i.e., the extent to which the culture of the host country of the subsidiary is novel or different from expatriates' home countries) attenuated the effects of motivational cultural intelligence on work adjustment and performance.

By contrast, X.P. Chen et al. (2012) focused on contextual variables that strengthen the effects of individual motivational cultural intelligence. These authors showed that the effect of motivational cultural intelligence on cultural sales (number of sales transactions involving clients from cultures different from the employee's own) was stronger (1) when firm diversity climate (i.e., employees' shared perceptions of the extent to which their firm values diversity within the firm) was stronger; and (2) when firm motivational cultural intelligence (i.e., the firm's capacity to direct attention and energy toward learning about and functioning effectively in cross-cultural situations) was higher. Together, these studies begin to illuminate crucial contextual boundary conditions of the effects of cultural intelligence.

Conclusion and Future Research

Cultural intelligence refers to the capability or potential to function effectively across varying cultural contexts. Cultural intelligence research has demonstrated that cultural intelligence is a distinct capability that accounts for significant performance variance in intercultural contexts. Future research needs to validate recent theoretical extensions, in particular, the subdimensions of cultural intelligence (Van Dyne et al., 2012), neurological correlates of cultural intelligence (Rockstuhl et al., 2010), as well as organizational-level conceptualizations of cultural intelligence (Ang and Inkpen, 2008; Moon, 2010). We also expect to see an increasing diversity in the measurement of cultural intelligence. One such direction includes the development of direct behavioral assessments of cultural intelligence, such as assessment centers. Finally, we note that studies of team-level cultural intelligence remain rare and require more conceptual and empirical work. For example, future work could explore team composition models of cultural intelligence (i.e., how should cultural intelligence within a team be distributed?), as well as processes and norms associated with high team cultural intelligence.

See also: Emotion, Perception and Expression of; Emotional Intelligence and Competencies; Five Factor Model of Personality, Facets of; Implicit Association Test; Intelligence: Assessments of; Intelligence: Central Conceptions and Psychometric Models; Intelligence: Historical and Conceptual Perspectives; Openness to Experience; Personality, Biological Models of; Personality, Trait Models of; Self-Efficacy; Situational Judgment Test; Social Intelligence and Competencies.

Bibliography

Ang, S., Inkpen, A.C., 2008. Cultural intelligence and offshore outsourcing success: a framework of firm-level intercultural capability. Decision Sciences 39, 337–358.

Ang, S., Rockstuhl, T., Ng, K.Y., 2014a. Performance-Based Cultural Intelligence (CQ): Development and Validation of an Intercultural Situational Judgment Test (iSJT). Nanyang Technological University: Center for Leadership and Cultural Intelligence.

Ang, S., Van Dyne, L., Rockstuhl, T. Cultural intelligence: origins, conceptualization, evolution, and methodological diversity. In: Gelfand, M., Chiu, C.Y., Hong, Y.Y. (Eds.), Advances in Culture and Psychology, vol. 5. Oxford University Press, New York, in press.

Ang, S., Van Dyne, L. (Eds.), 2008. Handbook of Cultural Intelligence. M.E. Sharpe, New York.

Ang, S., Van Dyne, L., Koh, C., 2006. Personality correlates of the four-factor model of cultural intelligence. Group & Organization Management 31, 100–123.

Ang, S., Van Dyne, L., Koh, C., Ng, K.Y., Templer, K.J., Tay, C., Chandrasekar, N.A., 2007. Cultural intelligence: its measurement and effects on cultural judgment and decision making, cultural adaptation and task performance. Management and Organization Review 3, 335–371.

Ang, S., Van Dyne, L., Tan, M.L., 2011. Cultural intelligence. In: Sternberg, R.J., Kaufman, S.B. (Eds.), The Cambridge Handbook of Intelligence. Cambridge University Press, New York, pp. 582–602.

Berry, J.W., 1976. Human Ecology and Cognitive Style: Comparative Studies in Cultural and Psychological Adaptation. Sage-Halsted, New York.

Beyene, T., 2007. Fluency as Stigma: Implications of a Language Mandate in Global Work. Doctoral dissertation. Stanford University, CA.

Campbell, J.P., McCloy, R.A., Oppler, S.H., Sager, C.E., 1993. A theory of performance. In: Schmitt, N., Borman, W.C. (Eds.), Personnel Selection in Organizations. Jossey-Bass, San Francisco, CA, pp. 35–70.

Chamorro-Premuzic, T., Harlaar, N., Greven, C.U., Plomin, R., 2010. More than just IQ: a longitudinal examination of self-perceived capabilities as predictors of academic performance in a large sample of UK twins. Intelligence 38, 385–392.

Chen, A.S.Y., Lin, Y.C., Sawangpattanakul, A., 2011. The relationship between cultural intelligence and performance with the mediating effect of culture shock: a case from Philippine laborers in Taiwan. International Journal of Intercultural Relations 35, 246–258.

Chen, G., Kirkman, B.L., Kim, K., Farh, C.I., Tangirala, S., 2010. When does cross-cultural motivation enhance expatriate effectiveness? A multilevel investigation of the moderating roles of subsidiary support and cultural distance. Academy of Management Journal 53, 1110–1130.

Chen, X.P., Liu, D., Portnoy, R., 2012. A multilevel investigation of motivational cultural intelligence, organizational diversity climate, and cultural sales: evidence from U.S. real estate firms. Journal of Applied Psychology 97, 93–106.

Chua, R.Y.J., Morris, M.W., Mor, S., 2012. Collaborating across cultures: cultural metacognition and affect-based trust in creative collaboration. Organizational Behavior and Human Decision Processes 118, 116–131.

Costa, P.T., McCrae, R.R., 1992. Revised NEO Personality Inventory (NEO PI-R) and New Five-Factor Inventory (NEO FFI) Professional Manual. Psychological Assessment Resources, Odessa, FL.

Crotty, S.K., Brett, J.M., 2012. Fusing creativity: cultural metacognition and teamwork in multicultural teams. Negotiation and Conflict Management Research 5, 210–234.

Crowne, K., 2009. Social intelligence, emotional intelligence, cultural intelligence and leadership: testing a new model. Paper presented at the 69th Academy of Management Meeting, Chicago, IL.

Earley, P.C., Ang, S., 2003. Cultural Intelligence: Individual Interactions across Cultures. Stanford University Press, Palo Alto, CA.

Elenkev, D.S., Manev, I.M., 2009. Senior expatriate leadership's effects on innovation and the role of cultural intelligence. Journal of World Business 44, 357–369.

Greenwald, A.G., Poehlman, T.A., Uhlmann, E., Banaji, M.R., 2009. Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. Journal of Personality and Social Psychology 97, 17–41.

Groves, K.S., Feyerherm, A.E., 2011. Leader cultural intelligence in context: testing the moderating effects of team cultural diversity on leader and team performance. Group & Organization Management 36, 535–566.

Huff, K.C., 2013. Language, cultural intelligence and expatriate success. Management Research Review 36, 596–612.

Imai, L., Gelfand, M.J., 2010. The culturally intelligent negotiator: the impact of cultural intelligence (CQ) on negotiation sequences and outcomes. Organizational Behavior and Human Decision Processes 112, 83–98.

Javidan, M., Teagarden, M.B., 2011. Conceptualizing and measuring global mindset. Advances in Global Leadership 6, 13–39.

Joseph, D.L., Newman, D.A., 2010. Emotional intelligence: an integrative meta-analysis and cascading model. Journal of Applied Psychology 95, 54–78.

Kim, Y.J., Van Dyne, L., 2012. Cultural intelligence and international leadership potential: the importance of contact for members of the majority. Applied Psychology: An International Review 61, 272–294.

Leung, K., Ang, S., Tan, M.L., 2014. Intercultural competence. Annual Review of Organizational Psychology and Organizational Behavior 1, 489–519.

MacNab, B., Brislin, R., Worthley, R., 2012. Experiential cultural intelligence development: context and individual attributes. The International Journal of Human Resource Management 23, 1320–1341.

Moon, T., 2010. Organizational cultural intelligence: dynamic capability perspective. Group & Organization Management 35, 456–493.

Mor, S., Morris, M., Joh, J., 2013. Identifying and training adaptive cross-cultural management skills: the crucial role of cultural metacognition. Academy of Management Education & Learning 12, 453–475.

Moynihan, L.M., Peterson, R.S., Earley, P.C., 2006. Cultural intelligence and the multinational team experience: does the experience of working in a multinational team improve cultural intelligence? Research on Managing Groups and Teams 9, 299–323.

Ng, K.Y., Ang, S., Van Dyne, L., August 2011. Speaking Up in the culturally diverse workplace: the role of cultural intelligence and language self-efficacy. Paper presented at the American Psychological Association. Washington, DC.

Ng, K.Y., Van Dyne, L., Ang, S., 2012. Cultural intelligence: a review, reflections, and recommendations for future research. In: Ryan, A.M., Leong, F.T.L., Oswald, F. (Eds.), Conducting Multinational Research Projects in Organizational Psychology. American Psychological Association, Washington, DC, pp. 29–58.

Oolders, T., Chernyshenko, O.S., Stark, S., 2008. Cultural intelligence as a mediator of relationships between openness to experience and adaptive performance. In: Ang, S., Van Dyne, L. (Eds.), Handbook of Cultural Intelligence. M.E. Sharpe, New York, pp. 145–158.

Prado, W.H., 2006. The Relationship between Cultural Intelligence and Perceived Environmental Uncertainty. Doctoral dissertation. University of Phoenix, AZ.

Rockstuhl, T., Ang, S., Lee, Y.-T., Paunova, M., August 2013a. Beyond international experience: effects of cultural capital on cultural intelligence. Paper presented at the Academy of Management Annual Meeting, Orlando, FL.

Rockstuhl, T., Presbitero, A., Ng, K.Y., Ang, S., June 2013b. Metacognitive cultural intelligence and offshoring performance: predictive validity of an intercultural situational judgment test (iSJT). Paper presented at the 3rd International Conference on Outsourcing of Information Services, Mannheim, Germany.

Rockstuhl, T., Ang, S., Ng, K.Y., Lievens, F., Van Dyne, L., 6 October 2014. Putting judging situations into situational judgment tests: evidence from intercultural multimedia SJTs. Journal of Applied Psychology. Advance online publication.

Rockstuhl, T., Hong, Y.Y., Ng, K.Y., Ang, S., Chiu, C.Y., 2010. The culturally intelligent brain: from detecting to bridging cultural differences. Journal of Neuroleadership 3, 1–15.

Rockstuhl, T., Ng, K.Y., 2008. The effects of cultural intelligence on interpersonal trust in multicultural teams. In: Ang, S., Van Dyne, L. (Eds.), Handbook of Cultural Intelligence. M.E. Sharpe, New York, pp. 206–220.

Rockstuhl, T., Seiler, S., Ang, S., Van Dyne, L., Annen, H., 2011. Beyond general intelligence (IQ) and emotional intelligence (EQ): the role of cultural intelligence (CQ) on cross-border leadership effectiveness in a globalized world. Journal of Social Issues 67, 825–840.

Şahin, F., Gürbüz, S., Köksal, O., Ercan, U., 2013. Measuring cultural intelligence in the Turkish context. International Journal of Selection and Assessment 21, 135–144.

Shokef, E., Erez, M., 2008. Cultural intelligence and global identity in multicultural teams. In: Ang, S., Van Dyne, L. (Eds.), Handbook of Cultural Intelligence. M.E. Sharpe, New York, pp. 177–191.

Sternberg, R.J., 2004. Culture and intelligence. American Psychologist 59, 325–338.

Sternberg, R.J., Detterman, D.K., 1986. What Is Intelligence? Contemporary Viewpoints on its Nature and Definition. Ablex, Norwood, NJ.

Tay, C., Westman, M., Chia, A., 2008. Antecedents and consequences of cultural intelligence among short-term business travelers. In: Ang, S., Van Dyne, L. (Eds.), Handbook of Cultural Intelligence. M.E. Sharpe, New York, pp. 126–144.

Van de Vijver, F.J.R., Leung, K., 2009. Methodological issues in researching intercultural competence. In: Deardorff, D.K. (Ed.), The Sage Handbook of Intercultural Competence. Sage, Thousand Oaks, CA, pp. 404–418.

Van Dyne, L., Ang, S., Koh, C., 2008. Development and validation of the CQS. In: Ang, S., Van Dyne, L. (Eds.), Handbook of Cultural Intelligence. M. E. Sharpe, New York, pp. 16–38.

Van Dyne, L., Ang, S., Ng, K.Y., Rockstuhl, T., Tan, M.L., Koh, C., 2012. Subdimensions of the four factor model of cultural intelligence: expanding the conceptualization and measurement of cultural intelligence. Social and Personality Psychology Compass 6, 295–313.

Ward, C., Fischer, R., Lam, F.S.Z., Hall, L., 2009. The convergent, discriminant, and incremental validity of scores on a self-report measure of cultural intelligence. Educational and Psychological Measurement 69, 85–105.

Wu, P.C., Ang, S.H., 2011. The impact of expatriate supporting practices and cultural intelligence on intercultural adjustment and performance of expatriates in Singapore. International Journal of Human Resource Management 22, 2683–2702.

Cultural Mediators and Gatekeepers

Susanne Janssen and Marc Verboord, Erasmus University Rotterdam, DR Rotterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H. Verdaasdonk, volume 20, pp. 13295–13299, © 2001, Elsevier Ltd.

Abstract

The shaping influence of cultural mediators, in particular their legitimizing power, has led cultural scholars to coin them 'tastemakers,' 'gatekeepers,' 'surrogate consumers,' 'reputational entrepreneurs,' or even 'coproducers' of the work of art. Yet, in practice, mediators perform highly different and often distinct activities according to their particular contributions in the (increasingly) vertically differentiated process of cultural production. This article discusses the various roles and activities of cultural mediators, followed by a review of the role and impact of critics and other mediators in the production and consumption of culture.

Introduction

Cultural gatekeepers and mediators – such as publishers, film studios, gallery owners, critics, or reviewers – can be defined as those involved in the mediation between the production of cultural goods and the production of consumer tastes (Bourdieu, 1984).

With the rise of the market as the most important structure for cultural production and distribution, institutions and agents who are in effect mediators have taken on an increasingly crucial role in the development of artistic careers and reputations as well as the formation of cultural tastes and consumption patterns. Cultural mediators became vital agents to artists, not only with respect to the immediate problem of economic survival and reaching an audience, but also for the valuation of their work and the establishment of their reputations.

The formative role of cultural mediators is not limited to traditional 'high' arts worlds by any means. They play an equally important part in the cultural industries, scouting and selecting creative talent (Caves, 2000; Hesmondhalgh, 2012) and determining what creations are turned into marketable commodities and what products eventually reach audiences (Debenedetti, 2006; Hirsch, 1972).

This article reviews key concepts, findings, and developments in the social scientific study of the role and impact of critics and other mediators in the production and consumption of culture. While we start with providing an overview of the various roles and activities of cultural mediators, which are discerned in the literature, our main focus will be on how mediators – in particular reviewers and critics – affect symbolic production in the cultural field (e.g., reputations of artists, legitimacy of genres). Cultural mediators add symbolic value to culture, whether they are involved in the material production (e.g., publishers), distribution (e.g., art galleries), or evaluation (e.g., critics) of culture. Through their selective and evaluative activities they contribute to the "production of the value of the work or, what amounts to the same thing, of the belief in the value of the work" (Bourdieu, 1996: p. 229) by virtue of the symbolic capital or prestige they have acquired as cultural experts. For other work on cultural mediators such as organizational studies or macro-level analyses of power relations in the cultural industries, (see Globalization and World Culture).

We draw liberally upon work inspired by various research perspectives that have been developed by empirically oriented cultural scholars from the 1970s onwards, including the production-of-culture perspective (Peterson and Anand, 2004); art world research (Becker, 1982); field theory (Bourdieu, 1984, 1996; Van Rees and Dorleijn, 2001); artistic classification systems theory (DiMaggio, 1987); and neo-institutional theory (Dowd, 2011). Although these approaches have different points of departure, employ different concepts and methodologies, and focus on different aspects of cultural production and consumption, no clear-cut divisions exist between them. Many individual researchers and studies, with good reason, can be viewed as representative of more than one approach. It lies outside the scope of this article to detail similarities and differences among them.

It is most important in this context that they all have illuminated, in one way or another, the thoroughly social nature of cultural production. This refers not only to the way in which a variety of people are actually involved in the production of a work itself (Becker, 1982), but also to the underlying processes and conditions that make the production possible and those, which determine its subsequent course.

Role of Cultural Mediators in the Lifecycle of Cultural Products

Most studies on cultural mediators have focused on how they impact artistic careers and reputations or influence the preferences and tastes of cultural consumers. The modus operandi of cultural mediators themselves has received somewhat less attention. Cultural mediators generally operate in fields characterized by turbulence, a huge supply of cultural offerings and aspiring artists, (demand) uncertainty, and a lack of unequivocal quality standards. In order to understand how they cope with these challenges and constraints we need to consider the organizational and institutional settings in which they perform their various mediating activities.

The past decades witnessed a flood of research on how mediators in the field of art and culture shape individual works and whole careers. The shaping influence of cultural mediators, in particular their legitimizing power, has led sociologists to coin them 'tastemakers,' 'gatekeepers,' 'surrogate consumers,'

'reputational entrepreneurs,' or even 'coproducers' of the work of art. Yet, in reality mediators perform highly different and often distinct activities according to their particular contributions in the (increasingly) vertically differentiated process of cultural production. Taking the production cycle of cultural products as a starting point, we can distinguish (at least) seven widespread mediating practices, which substantially inform the modus operandi of cultural mediators.

Selecting from the Supply (Gatekeeping)

One of the most studied aspects in the work of cultural mediators relates to what constitutes an essential dilemma in cultural production: of all the artists who turn to cultural organizations or sponsors to seek financial help in getting their products to an audience, only a limited number can actually receive such sponsorship. Agents such as publishers, Artists and repertoire (A&R) staff of record companies, or museum curators are responsible for making a first selection from the artists who are deemed worthy of being produced: they are the gatekeepers of the artistic field.

The term 'gatekeeping' has been applied when the focus is on judgments whether to admit persons or works into a cultural field; it has to do with accepting or rejecting works or their creators and the consequences of these choices for subsequent works and creators. Driving forces behind gatekeepers' decisions range from political and moral concerns, commercial interests, to 'purely' aesthetic motives. In most cases, they consist of a mixture of these.

The concept originates primarily from research into news production (Shoemaker and Vos, 2009), but it can be applied more or less equally to other areas of cultural production.

Gatekeeping refers to the capacity of mediators to exclude or promote. However, cultural industries from book publishing to television regularly make decisions other than choose/reject, altering or recontextualizing works at different stages of the production process. When the focus is on the way that works are changed as they move from being ideas to being finished products, the term decision making is more appropriate (Peterson and Anand, 2004). In the music industry, the original ideas of songwriters are filtered through music publishers' ideas concerning presentation (especially artist and style), which then play a part in promoting the product in several different markets. Processing takes place on the basis of a prediction about what the next 'gatekeeper' in the chain will think, the key being the overall product image.

It should be noted that most 'gatekeepers' are not passively waiting for new talents to submit their work. Research into the literary field has for instance shown how acquisition editors at publishing houses employ various search strategies (e.g., befriended foreign colleagues, agents, scouts, book auctions) to find the authors they want to publish (Franssen and Kuipers, 2013; Thompson, 2010). Similarly, in the music industries A&R managers or talent buyers in the live music scene actively scout for talent (Foster et al., 2011; Negus, 1999).

Cocreating/Editing

The next step in the production cycle is 'cocreating' or 'editing,' all the activities and feedbacks that mediators provide to artists to meet standards of artistic conventions or commercial expectations. These practices generally receive much less academic attention than gatekeeping. One reason is that, as the previous examples showed, the boundaries between gatekeeping and cocreating are often diffuse. As 'coproducers' many gatekeepers guide artists through the production process, shaping the content of cultural products (Becker, 1982; Peterson and Anand, 2004). Another reason, however, appears to be persistent influence of what Bourdieu called the 'charismatic ideology': the tendency to interpret all works of art as outbursts of individual creativity thereby downplaying the contributions of relevant others. Still, most cultural fields contain some form of cocreation: the sound of music records is adapted by producers and technicians, literary texts are fine-tuned by editors or translated by translators, and in film and television, producers and directors exert their influence on various levels: script, actors, photography, special effects, etc.

Connecting/Networking

Within a sector that is as strongly dependent on collaboration as the cultural sector (Becker, 1982), mediators are almost by definition connectors. The gatekeepers who scout for talent via their networks (see above) provide an obvious example. Yet some studies have highlighted how certain cultural mediators are specialized in connecting. We mention some examples. Literary agents have become powerful players in the field of fiction book publishing (Thompson, 2010). They negotiate the deals that authors make with publishers, but at the same time also actively manage their portfolio of author clients to establish their own reputation. This form of cultural brokerage – matching authors and publishers of similar standing – ultimately influences which type of authors are being published in the field (not unlike the influence of song-pluggers in the music industry).

Art dealers, gallery owners, auctioneers, and curators play a similar part in the visual arts, determining which works, artists, and schools come before the public as well as urging artists to produce certain types of work (Moulin, 1987; Quemin, 1997; Velthuis, 2005). Like independent agents spanning organizational boundaries in the recording industry, dealers seek artistic 'talent' in one group and corresponding buyers in a different group. Successful artists are those who recognize the privileged 'entrepreneurial' position of their dealers and listen to their feedback about clients' aesthetic sensibilities.

Selling/Marketing

In the final stage of the production process we find mediators who sell or market the products, and thus establish the contact with the audience. Production organizations in the cultural industries, such as publishing, music, or movie companies, depend for the successful dissemination of their products on media gatekeepers whose 'free publicity' is often more effective than advertising because of its alleged objectivity. Typical strategies to cope with constraints posed by media gatekeepers include the employment of specialized promotion and public relations personnel, overproduction, differential promotion of new products, and cooptation of media gatekeepers (Hirsch, 1972). The latter strategy is often at odds with the interests of media organizations, which are supposed to deploy

independent standards in their selection and assessment of cultural products (rather than promoting items which cultural organization put forward), while at the same time they have to attune their coverage to the target audiences of their advertisers. Hence, the efforts of cultural organizations to influence media gatekeepers vary to the extent that media coverage is vital for generating consumer demand.

Distributing

In most cultural fields, distribution comprises a form of mediation situated outside production. A number of research themes have emerged in the literature. First of all, the continuing competition between producers and distributors for economic dominance due to their interdependence – producers need the network to bring their products to the audience; distributors need attractive products to offer the audience. Fluctuations in the market power of these institutions inform takeovers and subsequent vertical integration. Another, more micro-oriented approach focuses on how distributors shape the production of culture. Thompson (2010), for instance, showed how the rise of book retailers changed the practices of publishers and ultimately the books that are released. Of course, distribution has become an even more timely research theme for mass-produced products as a consequence of digitalization (see later).

Whereas cultural industries studies routinely refer to distribution as the process of bringing products into circulation for sale (e.g., stores, galleries), display (film exhibitors), or performance (theaters), some institutions primarily seek to make culture more accessible (e.g., libraries and art loan organizations). Other ways of disseminating culture impact symbolic production. Festivals and biennales have become events where like-minded audiences gather to experience a special selection of products generally in a relaxed setting with food, drinks, and other entertainment, but whose most successful exponents propel high status (e.g., Cannes film festival or the Venice Biennale).

One can also think of educational institutions and art histories that mediate the relationship between artists and audiences. Curricula and cultural encyclopedias emphasize some works to such an extent that they become institutionalized as canonical art. Schoolbooks tell students whom to read and what kinds of things to say about literary texts. Meanwhile, educators' book orders help to determine which books remain available and enjoy regular updates.

Evaluating, Classifying, and Meaning Making

Many agents and organizations in the cultural field have no immediate interest in the artistic or commercial success of cultural products, but aim to judge the value and/or meaning of cultural endeavors. Most of these mediators (e.g., reviewers, critics, scholars) are situated in the field of media and academia. While Kees Van Rees (1983, 1989) was one of the first to analyze the three-step procedure of critics – description, interpretation, evaluation – and its problematic connection to 'critic's connoisseurship,' many researchers since have been working on cultural evaluation.

We will discuss the work of reviewers and critics in more detail in our Section The Impact of Critics and Other Mediators on Symbolic Production, but would like to point out some other adjacent roads this research has taken. For instance, the traditionally humanistic take on how art can be described and interpreted has been rephrased and readdressed using entirely different methodologies, as the study of cultural meaning by cultural sociologists such as John Mohr (*see* Formal Methods of Cultural Analysis) and Wendy Griswold. Griswold has demonstrated in various publications how mediators in the literary field 'fabricate' cultural meaning from socially shaped presuppositions in close dialogue with the particular characteristics of the cultural object. Another strand of research addresses cultural classification systems. Inspired by the work of Paul DiMaggio (1987) and others, scholars have started to unravel how mediators create, negotiate, and transform genre classifications, thereby impacting audience perceptions and market success (e.g., Zuckerman and Tai-Young Kim, 2003; *see* Classifications in Popular Music).

Censoring, Protecting, and Supporting

An entirely different type of mediating concerns censoring and other policy-related instruments. Political and legal constraints on artists' work and careers belong to the pervasive themes in the sociology of the arts (Alexander and Rueschemeyer, 2005; Peterson and Anand, 2004). Government bodies and laws, either through the carrot of support or the stick of censorship, accomplish exclusion and channeling in the visual arts, literature, and other cultural domains. Many studies have focused on how censorship operates to reduce freedom of expression, but censorship may also provide the stimulus for more complex symbol expression, illusion, euphemism, analogies, and the like.

Policy agents can also protect and support particular types of artists and products. In recent decades, three important research themes received attention. First, cultural globalization studies have highlighted how countries differ in taking legislative measures to protect local artists from competition from foreign artists. Examples are the installment of quota for national music to be played on the radio, or the subsidizing of the national film industry. Second emerging theme, which is directly related to changing patterns in cultural consumption, concerns the tension between highbrow and popular arts. Whereas highbrow art forms such as classical music and theater generally receive the bulk of governmental support, it is increasingly signaled in the media that their audience generally represents just a small, relatively affluent portion of society. A third much debated policy issue is how to retain a meaningful copyright system that protects artists and producers from the unlawful reproduction or distribution of their works while at the same time preserving the innovative power of the Internet and the free flow of information.

The Impact of Critics and Other Mediators on Symbolic Production

Institutional Context of Cultural Evaluation

Cultural mediators are also confronted with challenges and risks of a different nature, which have to do with the institutional setting in which they perform their selective and evaluative

tasks. Their selections and evaluations do not only affect the reputations of cultural products and their makers, but also their own status and authority ('their symbolic capital') in the field.

This applies to all professional experts involved in the selection and valuation of cultural products – talent scouts, agents, publishers, film or music producers, curators, theater programmers, etc. – but it holds a fortiori for critics and reviewers whose selections and evaluations are made publicly available. As Bourdieu (1983: p. 317) points out: "Every critical affirmation contains, on the one hand, a recognition of the value of the work which occasions it, which is thus designated as worthy object of legitimate discourse (…), and, on the other hand, an affirmation of its own legitimacy. Every critic declares not only his judgment of the work, but also his claim to the right to talk about it and judge it."

Like all cultural experts, critics need to gain recognition as connoisseur and to inspire confidence in their capacity to assess the properties and value of a work. However, critics lack a reliable instrument to assess aesthetic qualities in an unequivocal way and no objective agency or procedure can prove their evaluations true or false (Van Rees, 1989). Moreover, they have to perform their job in a context of numerous cultural offerings and limited media space, keeping in mind the audience their media is supposed to serve.

That is why critics take into account various institutional quality indicators in their selection and assessment of new cultural products (Janssen, 1997). Particularly, they consider the reputation of the cultural organization (e.g., the publisher, film company, music label, theater, or gallery) that brings a work or artist on the market and previous critical assessments of an artist's work (i.e., the artist's reputation). They also take due note of the cultural coverage of other media outlets, notably those which are active in the same market segment. Moreover, they strongly rely on genre classifications and genre conventions (Glynn and Lounsbury, 2005; Zuckerman and Tai-Young Kim, 2003), as well as on 'background' and 'inside' information they receive through their contacts with representatives of cultural organizations, peers, and other experts in their professional networks (Bielby and Bielby, 1994).

In considering the institutional 'clues' provided by established reputations, peer opinions, genre distinctions, competitors, and professional networks, critics reduce the uncertainty as to which works deserve their attention and how to express this attention. In doing so, they enhance the odds that their assessments will be taken seriously or even get adopted by their fellow critics and other experts. Ultimately, the authority of critics and other mediators depends on the credits they have earned by making 'successful' evaluations, i.e., evaluations that were adopted by other experts in the field and are trusted by the audiences they serve. This does not imply that critics should always comply with their peers. On the contrary, they have to distinguish themselves in order to gain prestige as a connoisseur. By developing perspectives, which qualify previous critical tenets or by deviating in their evaluations or their choice of works for discussion critics may make their mark as 'independent' experts. However, the institutional framework within which they function poses limits to the scope for dissent (Janssen, 1997; Van Rees, 1989).

The above strategies to minimize uncertainties and risks in critic's assessment of cultural products are also omnipresent in the decision-making processes of other cultural mediators. Like critics, other mediators base their selections and decisions to a large extent on the institutional knowledge and contextual information they have about cultural products and their creators, including their classification into particular genres; the recognition creators have are thus far received from various actors in the field; the reputations of their organizational sponsors; the choices made by competitors; and the information and opinions of members of their professional networks.

Reputation Making and Cultural Consecration

Critics and other experts play a crucial role in the establishment and survival of artistic reputations. Artists and their critic-advocates seeking recognition make claims of creativity, quality, innovativeness, excellence – that is 'talent' or 'genius.' While sociological theory and method are not very useful in identifying 'talent' or 'genius' in works or their makers, they have proved very helpful in demystifying canon formation, the processes by which claims of 'genius' or 'quality' come to be agreed upon and reputations established (Menger, 2014).

Art critics, curators, and gallery owners appear to 'discover' new artistic geniuses all the time in a promotional role that plays a crucial part in the distribution of paintings to sophisticated buyers who would not dream of being influenced by advertising (Moulin, 1987). Studies of the contemporary art market (Velthuis, 2005) underline the promotional capacity of art experts as well as their key role in the construction of artistic value.

How the value and nature of literary works come to be agreed upon by critics and writers' reputations established is illuminated in many studies (Berkers et al., 2014; Janssen, 1997; Van Rees and Dorleijn, 2001). Van Rees (1983), for example, details how the complementary activity reviewers, essayists, and academic critics determine to a great extent which texts are regarded as legitimate forms of fiction; what rank they are supposed to occupy within the hierarchy of literary works; and what statements count as proper and relevant ways of characterizing these texts.

Critics are also key agents in acts and processes of cultural 'consecration,' whereby a few artists and works are identified and set apart from others in their field as exceptionally valuable and gain an exemplary and celebrated status. Cultural consecration refers to the establishment of a 'magical division' between the 'pure' and 'sacred' cultural offerings, on the one side, and the 'facile' or "profane" products on the other (Bourdieu, 1984). Critical attention, peer recognition, and public acclaim all affect the odds that some artists or works are eventually counted among the 'greatest of all time.' Cultural scholars have examined acts and processes of consecration in a variety of cultural fields including literature (Van Rees and Dorleijn, 2001), film (Allen and Lincoln, 2004), visual arts (Ginsburgh and Weyers, 2006), classical music (Dowd, 2011), and popular music (Bennett, 2009). This research has shown that critical recognition is especially important in the consecration of cultural producers and their works.

Cultural Legitimation

'Retrospective' consecration of some works and artists also takes place because art world members seek to raise the prestige of their field by highlighting "elements of their pasts which are most clearly artistic, while suppressing less desirable ancestors" (Becker, 1982: p. 339). In rewriting the history of their field, critics and other experts try to show it has always generated artistically valuable work. Thus retrospective cultural consecration not merely singles out the greatest works or artists in a field but is also a means of bestowing legitimacy upon an entire field of cultural production.

The processes and mechanisms through which cultural forms attain artistic legitimacy have garnered much attention over the past decades. Cultural and organizational scholars have highlighted the role of social change in creating an 'opportunity space' for ascendant art forms (Baumann, 2007; DiMaggio, 1992). Claims of cultural value emerge in a wider social context, which may change over time in terms of the degree of competition, the availability of alternatives or substitutes, public interest, etc., and as such affect whether particular products and genres will gain cultural legitimacy. Cultural legitimation is also contingent upon the institutional and organizational resources (e.g., money, labor, knowledge, networks) that art world members manage to mobilize (Baumann, 2007; Becker, 1982).

A third key factor in processes of cultural legitimation is the development of a legitimating ideology that substantiates claims to artistic value and lends symbolic prestige to a cultural field (Baumann, 2007; Becker, 1982; Bourdieu, 1984; Van Rees, 1983). Artistic value attribution calls for *an explanation* of why particular cultural objects or genres can be deemed artistically important (Baumann, 2007). Media critics and reviewers – drawing on academic criticism – play a central role in developing and disseminating such explanations of artistic worth and are thus crucial agents in elevating the status of cultural forms and genres.

The impact of critics and reviewers may appear more salient in traditional high art worlds where their expert status is institutionally embedded and reinforced through academic programmes and educationally curricula (Baumann, 2007; Bourdieu, 1984; Van Rees and Dorleijn, 2001). However, the impact of media attention and critical discourse is also crucial in the valorization and aesthetic mobility of a variety of popular cultural forms.

Baumann (2007) demonstrates how the intellectualizing discourse of film critics, who over the years increasingly employed terminology (e.g., art, genius) and reviewing techniques (e.g., mentions and comparisons of directors) associated with high culture, contributed to the cultural legitimation of film in the United States. The development of this 'high art discourse' in film reviewing paved the way for the wider acceptance of film as art as it resonated with the cultural capital of higher educated audiences, peer critics and other members of the field. Critics and reviewers have been found to play a similar role in the valorization of a variety of other cultural forms (Janssen et al., 2011), including jazz, popular music, television, and fashion.

Meanwhile, the past decades also witnessed the rise of alternative, 'popular' forms of critical discourse, albeit often in less prestigious settings. Pop music reviews, for example, are often characterized by a popular aesthetic discourse, in which functionality, entertainment and other more audience oriented criteria prevail over high arts standards (Van Venrooij and Schmutz, 2010). Likewise, Kersten and Bielby (2012) find that contemporary film reviews "incorporate aesthetic elements drawn from popular interests as well as elite art considerations, thereby complicating critics' aesthetic systems and analysts' classificatory schemes."

Impact on Cultural Tastes, Consumer Behavior, Commercial Success

Many studies have found an influence of media critics and reviewers on consumer preferences and the commercial performance of cultural products, notably in the field of film and the performing arts (Debenedetti, 2006). When the supply is large, time is scarce, and markets are flooded with commercial information, consumers can draw on critical evaluations to simplify the decision-making process and reduce the basic uncertainty associated with experiential products of which the quality is hard to assess prior to consumption. Nonetheless, whether critics' opinions influence the commercial success of products remains contested and seems to depend on specific features of the product and/or the specific audience at which the product is targeted (Gemser et al., 2007). Individual audience characteristics also play a part. Cultural participants who are more involved or have more expertise are more likely to use reviews than others.

The correlation between critical evaluation and commercial success has been found to be particularly strong for cultural products that have weak signaling properties for the general public (e.g., lack a high-profile cast, special effects, exotic locations, and/or elaborate sets), low-budget works that cannot rely heavily on promotional devices (such as previews, trailers and advertisements) to shape consumer preferences and belong to unfamiliar genres (Debenedetti, 2006).

For blockbuster movies targeting large audiences, the amount of media coverage has a stronger impact on box office performance than whether reviews contain positive evaluations. However, in the case of art house movies and other films that cater for artistically oriented audiences, critical evaluations and recommendations have positive effects on commercial success (Gemser et al., 2007). Thus, whereas 'art' films need positive reviews to gain industry recognition, the commercial success of mainstream films depends on being noticed and creating a 'buzz' (Holbrook and Addis, 2008).

These findings agree with socially embedded patterns of media use and cultural tastes, as well as with the institutional embeddedness of critics. A preference for cultural products of a particular degree of legitimacy often coincides with using media enjoying a similar degree of legitimacy (Bennett et al., 2009; Bourdieu, 1984). The influence of a review thus depends on the critic's or medium's social and cultural proximity to the reader. Cultural participants who have a taste for 'highbrow' cultural products often read the print media in which such products receive extensive coverage, (e.g., 'quality' newspapers) and via processes of self-selection they are (thus) likely to take critics' judgments into consideration (Janssen et al., 2011; Verboord, 2013).

Cross-Cultural Exchange

Cultural mediators are also important when it comes to the international mobility, visibility, and reception of cultural products (Crane et al., 2002). Griswold (1987) considers the export of meaning beyond national borders and the transformations of meaning, which occur through this process. Comparing American, British, and West Indian reviews of the work of the Barbadian novelist George Lamming, she finds stunning cross-cultural divergences. Race is highlighted in American reviews of Lamming's novels, yet virtually ignored in British discussions, which focus instead on the novels' language and literary qualities.

In spite of ongoing 'cultural globalization,' and in contrast to the so-called 'contra-flow' argument, coverage of foreign culture in the European and US elite newspapers has not become much more global between 1955 and 2005 (Berkers et al., 2011; Janssen et al., 2008). The global diversity of cultural coverage has increased, but international art coverage has remained largely confined to a select group of countries belonging to the core of the cultural world system, while domestic cultural products continue to be prominently featured in the cultural coverage of elite newspapers. These findings, which are in line with those of many other empirical studies, underscore how cultural mediators, can enhance the enduring imbalances in international cultural exchanges, and, may promote and sustain local cultural production.

Of course, they can also serve as 'gatemakers' pushing the boundaries of international cultural traffic, as is evidenced, for example, by the foreign literature series that French publishers developed with the help of the French ministry of culture in the 2000s, featuring translations from more than 30 small and peripheral languages into French (Sapiro, 2010). Similarly, Kuipers (2011) shows how television buyers who mediate and maintain relations between the national and transnational television field, play an important role in the diffusion of quality standards, practices, and programs into national fields of television production, and, from there, to consumers.

Developments Impacting the Role and Authority of Cultural Mediators

Increasing globalization has enhanced the risks and uncertainties inherent in the work of cultural mediators, such as publishers, curators, festival programmers, casting agencies, music producers, and television broadcasters. In addition to an already huge supply of domestic products, cultural mediators are confronted with an overwhelming abundance and diversity of foreign cultural offerings, for which the audience interest, market potential, and artistic merits are difficult to assess.

At the same time, due to the advancing commercialization of cultural production fields, cultural mediators face heightened competition, both in national cultural fields and the fast-paced, highly competitive transnational cultural arenas in which the odds of success and recognition are very unequal for players from different countries and places.

The intensified use and increased interactivity of the Internet have transformed the ways in which cultural information is being searched, created, and valued (*see* Internet and Culture). In recent years, cultural consumers across the globe have increasingly turned toward the web: to purchase cultural products at online shops; download and upload music, videos, films, and other content; rate and review cultural products; get advice and recommendations from other consumers; and engage in debates and discussions on social network sites. Many cultural websites prosper through participatory practices by audiences. Peer-to-peer communication enables audiences to create, share and evaluate user-generated content and actively participate in cultural opinion making.

Increasingly, Internet users thus seem to be moving away from institutional gatekeepers and experts and instead refer to information provided by other Internet users. This trend does not merely undermine the authority of these cultural mediators, but also appears to erode the entire underlying system in which institutionally embedded experts decide on the value and legitimacy of cultural artifacts in society (Verboord, 2013). 'Bottom-up' practices of selecting and evaluating culture challenge the existent 'hierarchical' model of cultural valorization, in which critics and other mediators play a pivotal role by legitimizing the culture from the higher status groups through their cultural authority.

However, this development is not just caused by the rise of new media technologies and associated opportunities and practices for cultural consumers, but is deeply rooted in wider societal and cultural changes that occurred since the 1960s in Western societies. The democratization of higher education, the emancipation of minorities (women, youngsters, ethnic, racial, and sexual minorities), increased social mobility and heterogeneity contributed to the erosion of traditional cultural hierarchies and the rise of omnivorous taste patterns, while it also enabled social climbers and minorities to 'import' their tastes into higher circles and to bestow prestige upon their preferred genres. Processes of individualization have made people less prone to subscribe to traditional cultural hierarchies, collective taste patterns, and the judgments of cultural experts, but, instead, increasingly require them to choose individually and to show individual authenticity in their expression of taste (Janssen et al., 2011).

These developments profoundly affect the work of critics and other cultural mediators. As the institutional consensus and institutionally embedded authority through which they used to operate declines, their position becomes more dependent on the uncertainties and vagaries of the market, requiring them to stay more closely attuned to or follow audience interests and consumer preferences.

Future Research

Many avenues for further research relate to the fast changing cultural landscape in which mediators maneuver. One obvious subject for future research already mentioned concerns the rise of the Internet. Many forms of mediating appear to become obsolete as technologies enable artists to bypass mediators. Publishers, record companies, book stores, record stores, newspapers: they are all under pressure, and how their possible disappearance would impact the way culture, and cultural value, is being produced is far from clear.

How increased globalization and the emergence of transnational cultural fields transform the role and practices of cultural mediators, and how this plays out, at the local and transnational level, for agents from different countries and places across the globe, are other issues which definitely call for further research.

Yet also other developments deserve attention: the increasing segmentation and differentiation of cultural work (e.g., the rise of live music industries; literary agents), the advance of oligopolies (in music and film production) and monopolies (online: Google, Amazon; offline: LiveNation), and the decreasing support for the arts among general populations, which impacts not only the work and careers of artists but also the professional lives of the manifold mediators who contribute to the highly differentiated processes of cultural production and distribution.

See also: Artists, Competition and Markets; Classifications in Popular Music; Cultural Policy Regimes in Western Europe; Cultural Production in Networks; Culture, Production of: Prospects for the Twenty-First Century; Formal Methods of Cultural Analysis; Globalization and World Culture; Internet and Culture; Markets: Artistic and Cultural; Symbolic Boundaries; Video Cultures: Television.

Bibliography

Alexander, V.D., Rueschemeyer, M., 2005. Art and the State: The Visual Arts in Comparative Perspective. Palgrave Macmillan, Houndmills, Basingstoke, Hampshire; New York.

Allen, M.P., Lincoln, A.E., 2004. Critical discourse and the cultural consecration of American films. Social Forces 82, 871–894.

Baumann, S., 2007. Hollywood Highbrow. From Entertainment to Art. Princeton University Press, Princeton, NJ.

Becker, H., 1982. Art Worlds. University of California Press, Berkeley, CA.

Bennett, A., 2009. "Heritage rock": rock music, representation and heritage discourse. Poetics 37, 474–489.

Bennett, T., Savage, M., Silva, E.B., Warde, A., Gayo-Cal, M., Wright, D., 2009. Culture, Class, Distinction. Routledge, London.

Berkers, P., Janssen, S., Verboord, M., 2011. Globalization and ethnic diversity in Western newspaper coverage of literary authors. American Behavioral Scientist 55, 624–641.

Berkers, P., Janssen, S., Verboord, M., 2014. Assimilation into the literary mainstream? the classification of ethnic minority authors in newspaper reviews in the United States, the Netherlands and Germany. Cultural Sociology 8, 25–44.

Bielby, W.T., Bielby, D.D., 1994. All hits are flukes: institutionalized decision making and the rhetoric of network prime-time program development. American Journal of Sociology 9, 1287–1313.

Bourdieu, P., 1983. The field of cultural production, or: The economic world reversed. Poetics 12, 311–356.

Bourdieu, P., 1984. Distinction. A Social Critique of the Judgment of Taste. Routledge & Kegan and Paul Ltd, London.

Bourdieu, P., 1996. The Rules of Art: Genesis and Structure of the Literary Field. Stanford University Press, Stanford, CA.

Caves, R.E., 2000. Creative Industries. Harvard University Press, Cambridge, MA.

Crane, D., Kawashima, N., Kawasaki, K. (Eds.), 2002. Global Culture. Media, Arts, Policy and Globalization. Routledge, London.

Debenedetti, S., 2006. The role of media critics in the cultural industries. International Journal of Arts Management 8, 30–42.

DiMaggio, P., 1987. Classification in art. American Sociological Review 52, 440–455.

DiMaggio, P., 1992. Cultural boundaries and structural change: the extension of the high culture model to theater, opera, and the dance, 1900-1940. In: Lamont, M., Fournier, M. (Eds.), Cultivating Differences. University of Chicago Press, Chicago, IL, pp. 21–57.

Dowd, T.J., 2011. Production and producers of lifestyles. Kölner Zeitschrift für Soziologie und Sozialpsychologie 63 (51), 113–138.

Foster, P., Borgatti, S.P., Jones, C., 2011. Gatekeeper search and selection strategies: relational and network governance in a cultural market. Poetics 39, 247–265.

Franssen, T., Kuipers, G., 2013. Coping with uncertainty, abundance and strife: decision-making processes of Dutch acquisition editors in the global market for translations. Poetics 41, 48–74.

Gemser, G., van Oostrum, M., Leenders, M.A.A.M., 2007. The impact of film reviews on the box office performance of art house vs. mainstream motion pictures. Journal of Cultural Economics 31, 43–63.

Ginsburgh, V., Weyers, S., 2006. Persistence and fashion in art Italian Renaissance from Vasari to Berenson and beyond. Poetics 34, 24–44.

Glynn, M.A., Lounsbury, M., 2005. From the critics' corner: logic blending, discursive change and authenticity in a cultural production system. Journal of Management Studies 42, 1031–1055.

Griswold, W., 1987. The fabrication of meaning: literary interpretation in the United States, Great Britain, and the West Indies. American Journal of Sociology, 1077–1117.

Griswold, W., 1992. The writing on the mud wall: Nigerian novels and the imaginary village. American Sociological Review, 709–724.

Hesmondhalgh, D., 2012. The Cultural Industries, third ed. Sage, Los Angeles, CA.

Hirsch, P.M., 1972. Processing fads and fashions: an organization-set analysis of cultural industry systems. American Journal of Sociology, 639–659.

Holbrook, M.B., Addis, M., 2008. Art versus commerce in the movie industry: a two-path model of motion-picture success. Journal of Cultural Economics 32, 87–107.

Janssen, S., 1997. Reviewing as social practice: institutional constraints on critics' attention for contemporary fiction. Poetics, 275–297.

Janssen, S., Verboord, M., Kuipers, G., 2008. Cultural globalization and arts journalism. American Sociological Review 73, 719–740.

Janssen, S., Verboord, M., Kuipers, G., 2011. Comparing cultural classification: high and popular arts in European and US newspapers, 1955-2005. Kölner Zeitschrift für Soziologie und Sozialpsychologie 63 (51), 139–168.

Kersten, A., Bielby, D.D., 2012. Film discourse on the praised and acclaimed: reviewing criteria in the United States and United Kingdom. Popular Communication 10, 183–200.

Kuipers, G., 2011. Cultural globalization as the emergence of a transnational cultural field: transnational television and national media landscapes in four European countries. American Behavioral Scientist 55, 541–557.

Menger, P., 2014. The Economics of Creativity. Art and Achievement under Uncertainty. Harvard University Press, Cambridge, MA.

Moulin, R., 1987. The French Art Market: A Sociological View. Rutgers University Press, New Brunswick, NJ.

Negus, K., 1999. Music Genres and Corporate Cultures. Routledge, London.

Peterson, R.A., Anand, N., 2004. The production of culture perspective. Annual Review of Sociology 30, 311–334.

Quemin, A., 1997. In: Les commissaires-priseurs: la mutation d'une profession. Anthropos, Paris.

Sapiro, G., 2010. Globalization and cultural diversity in the book market: the case of literary translations in the US and in France. Poetics 38, 419–439.

Shoemaker, P.J., Vos, T.P., 2009. Gatekeeping Theory. Routledge, London.

Thompson, J.B., 2010. Merchants of Culture. Polity, London.

Van Rees, C.J., 1983. How a literary work becomes a masterpiece. Poetics 12, 397–418.

Van Rees, C.J., 1989. The institutional foundation of a critic's connoisseurship. Poetics 18, 179–198.

Van Rees, C.J., Dorleijn, G., 2001. The eighteenth-century literary field in Western Europe: interdependence of material and symbolic production and consumption. Poetics 28, 331–348.

Van Venrooij, A., Schmutz, V., 2010. The evaluation of popular music in the United States, Germany and the Netherlands: a comparison of the use of high art and popular aesthetic criteria. Cultural Sociology 4, 395–421.

Velthuis, O., 2005. Talking Prices: Symbolic Meanings of Prices on the Market for Contemporary Art. Princeton University Press, Princeton, NJ.

Verboord, M., 2013. The impact of peer-produced criticism on cultural evaluation. New Media & Society first published on 3 July 2013 as http://dx.doi.org/10.1177/1461444813495164.

Zuckerman, E.W., Tai-Young Kim, T.Y., 2003. The critical trade-off: identity assignment and box-office success in the feature film industry. Industrial and Corporate Change 12, 27–67.

Cultural Participation, Trends in

Henk Roose and Stijn Daenekindt, Ghent University, Ghent, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article focuses on trends in cultural participation. How have the extent and the way in which people consume culture changed? Research shows that participation in highbrow culture remains relatively stable, whereas the consumption of popular, commercial culture has increased since the 1970s. A number of societal developments drive these evolutions in cultural participation. Sociodemographic changes, most notably educational expansion, as well as the rise of the entertainment industry and the Internet boom, greatly affect the field of artistic production, mediation, and consumption. These changes result in the shifting, or, as some argue, the erosion of esthetic boundaries, and the rise of the cultural omnivore – someone who participates in a variety of cultural activities – as provisional endpoints.

Introduction

In this study of trends in cultural participation, we focus on how cultural practices have changed over the past few decades and on how the relationship between practices and position in the social hierarchy has evolved. Our research perspective owes a great deal to Pierre Bourdieu's research on cultural fields, cultural participation, distinction, cultural capital, and education. Bourdieu's work builds on Max Weber's idea of status cultures and on the distinction Durkheim makes between the sacred and the profane. Cultural participation ranges from high to low, from arts participation and consumption of highbrow cultural products (e.g., visiting museums, attending the opera, and reading books) to consumption of lowbrow products (e.g., listening to rock music or going to the movies). Since the publication of Bourdieu's *La Distinction* (1979), research on cultural participation has thrived, and empirical studies have shown recurring patterns across nations and times. Results consistently show that arts participation – defined here as the consumption of highbrow symbolic, artistic goods in the public and/or private sphere – is predominantly situated among the highly educated and the upper and middle classes. In the Bourdieusian paradigm, cultural participation is regarded as a manifestation of cultural capital, as a means of maximizing life chances and opportunities by accumulating and monopolizing scarce economic, social, and cultural resources (Bourdieu, 1986).

Since the 1960s – both in the United States and Europe – a number of evolutions have challenged the monopoly that arts participation has held as a means of ensuring and proclaiming a dominant social position. These include changes in the field of cultural production with the entertainment industry's rise, the advent of television, and the Internet's boom. Now, individuals have unprecedented access to a huge and varied supply of cultural experiences. Changes in cultural supply are paralleled by changes in the field of consumption. Large sociodemographic changes, such as educational expansion/democratization, the increase in women's labor force participation, and the aging of the population have placed additional strain on the monopoly and legitimacy of the prestigious highbrow arts as an essential part of an individual's *Bildung* – the idea that familiarity with the *beaux arts* is essential to a person's upbringing. The deinstitutionalization of high culture (DiMaggio, 1987, 1991), in addition to the waning of cultural hierarchies, may have reinforced these trends. Studying trends in cultural participation is relevant for two reasons. It makes it possible (1) to document how the constitution of cultural capital may have changed, and (2) to probe the implications of this change for the dynamics of social reproduction and inequality.

Trends in Participation: Facts and Figures

How has the frequency of cultural participation evolved over time and can changes be detected in audience composition? Rather unfortunately, one characteristic of research on trends in cultural participation is the scarcity of available data. Notable exceptions to this are found in the Netherlands, where the Social and Cultural Planning Office (SCP) has conducted the Aanvullend Voorzieningengebruik Onderzoek ("Amenities and Services Utilization Survey") every 4 years since 1979; in Denmark (Danish National Centre for Social Research 1975, 1987, 1993, 1998, 2004), and in the United States (Survey of Public Participation in the Arts (SPPA) 1982, 1992, 2002, 2008; and the General Social Survey (GSS) 1993, 2002, 2010). Moreover, the activities recorded in most of these surveys are biased toward legitimate, highbrow genres, especially in the first waves. Additionally, it is unclear to what extent the categories used in questionnaires vary over time and across geographic boundaries (Marsden and Swingle, 1994; Peterson, 2005).

We present trends for four forms of cultural participation: attending classical concerts, visiting museums, reading books, and going to the movies. For Europe we use data on the Netherlands (**Table 1**). Data for the United States are shown in **Table 2**. The figures cannot accurately portray the complexities and specificities of all developments in the different genres. For a more and detailed discussion, we refer to Van Eijck and Knulst (2005) and to DiMaggio and Mukhtar (2004).

The Dutch data reveal three parallel evolutions (Knulst and Kraaykamp, 1998; Van Eijck et al., 2002; Van Eijck and Knulst, 2005). First, the data show that, in general, participation in highbrow cultural activities has not decreased dramatically and that evolutions in participation depend on the particular

Table 1 Trends in arts participation in the Netherlands: relative frequencies for attending classical concerts, reading books, attending museums, and going to the movies

	1979	*1983*	*1987*	*1991*	*1995*	*1999*	*2003*	*2007*
Attended classical concert during the past 12 months.[a]								
−45 years	9.7	10.6	10.8	12.0	10.0	8.3	7.9	7.9
+45 years	15.7	17.5	15.6	17.6	19.0	19.9	16.4	19.6
Total	11.6	13.0	12.5	14.0	13.4	12.9	11.7	12.9
Read at least one book during the past month.								
−45 years	71.8	72.7	68.7	67.6	64.1	63.0	57.5	57.8
+45 years	50.4	54.6	53.5	54.9	50.6	53.6	49.3	52.8
Total	65.0	66.4	63.4	63.0	59.1	59.3	53.9	55.6
Visited a museum during the past 12 months.								
−45 years	29.2	39.4	42.8	43.4	35.4	35.4	36.9	40.9
+45 years	21.1	27.9	32.5	36.7	33.4	39.8	39.4	42.4
Total	26.0	35.4	39.1	41.0	34.7	37.7	38.0	41.6
Went to the movies during the past 12 months.								
−45 years	62.3	62.5	58.1	61.2	64.7	69.7	73.5	72.9
+45 years	18.2	18.0	17.8	17.4	20.3	27.2	30.8	33.0
Total	48.1	47.0	43.8	45.1	48.0	53.0	55.6	55.7

[a]In the 1979 and 1983 rounds, this question includes both opera and classical music.
SCP AVO (Sociaal en Cultureel Planbureau, Aanvullend Voorzieningengebruik Onderzoek) 1979, 1983, 1987, 1991, 1995, 1999, 2003, 2007.

Table 2 Trends in arts participation in the United States: Relative frequencies for attending classical concerts, reading books, and visiting art museums or galleries

	1982	*1985*	*1992*	*2002*	*2008*
Attended a classical concert during the past 12 months.					
−45 years	13.4	12.8	11.0	9.4	7.7
+45 years	12.5	12.5	14.4	13.9	10.9
Total	13.0	12.7	12.5	11.6	9.3
Read at least one book during the past 12 months.[a]					
−45 years	–	–	63.8	57.2	54.0
+45 years	–	–	57.0	56.0	54.5
Total	–	–	60.9	56.6	54.3
Visited an art museum or gallery during the past 12 months.					
−45 years	25.6	24.8	29.4	26.3	24.5
+45 years	17.4	18.0	23.1	26.7	21.1
Total	22.1	21.9	26.7	26.5	22.7

[a]Not available in the 1982 and 1985 waves.
Survey of Public Participation in the Arts 1982, 1985, 1992, 2002, 2008.

cultural genre or domain. Attending classical concerts and going to the opera (not shown), for example, remain stable at around 12 and 5%, respectively, whereas reading books has decreased over the last few decades. Since the 1970s, visiting rates for museums have risen from 26% in 1979 to 41% in 2007. Second, evolutions differ between generations: younger generations, aged 45 or younger, show a decreased participation in these forms of highbrow culture. In particular, reading books declines steeply, although it remains stable in older generations. And third, in addition to these trends in arts participation, participation in the popular, commercial culture in the Netherlands has increased over time, both for younger and older generations. The proportion of those going to the movies has been rising since the 1990s, from 45% in 1991 to 56% in 2007. This trend is corroborated by data from Denmark where participation in popular, lowbrow public culture – such as attending pop concerts and going to the movies – has increased since 1975 (Jæger and Katz-Gerro, 2010; Katz-Gerro and Jæger, 2011).

In the United States, DiMaggio and Mukhtar (2004), using SPPA data (1982, 1992, 2002), show similar evolutions. Participation in highbrow culture has declined since 1982 (see Table 2). This general decline is more substantial among younger cohorts and among highly educated groups in society. Table 2 shows that in 1982, 13% of the US population younger than 45 attended a classical concert the past year, compared with 8% in 2008. Comparable numbers for US citizens aged 45 or older are 13% in 1982 versus 11% in 2008. Two types of cultural participation – visiting art museums and attending jazz concerts – deviate from this general pattern as attendance of

these public practices has either remained stable or increased since 1982. The observation that museum attendance follows a distinct trend compared with other forms of cultural participation parallels findings in Europe, where Knulst has observed increased museum attendance since the 1950s in the Netherlands (Knulst, 1993; see also **Table 1**). DiMaggio and Mukhtar (2004) consider the increased popularity of art museums and jazz to be an indication of the rise of the omnivore, a type of cultural consumer that has eclectic, inclusive tastes and shows interest in a broad variety of cultural practices, mixing lowbrow and highbrow cultural genres.

In the 1990s, Richard Peterson and his colleagues published a series of papers in which they show a marked shift in the cultural preferences of younger generations (Peterson, 1992; Peterson and Kern, 1996; Peterson and Simkus, 1992). Between 1982 and 1992 the preference for folk and pop music increased for people who also enjoy classical music genres. This change is most evident in the younger cohorts. Peterson's thesis on omnivorousness has spurred researchers all over the world to empirically test this supposed shift from snob to omnivore, and that research has met with great success – omnivores are present in the United States, Canada, Europe, and Australia (e.g., Bryson, 1996, 1997; Coulangeon, 2005; Emmison, 2003; Erickson, 1996; Van Eijck, 1999). The cultural omnivore is generally young, highly educated, and from the middle and/or upper middle classes.

Interpretations of Trends in Cultural Participation

To explain these trends, some scholars highlight the importance of changes in the field of cultural production and mediation, the degree of institutionalization of the arts in the educational system and state-subsidized cultural organizations. Other researchers focus on socioeconomic and cultural factors affecting the behavior of cultural participants. Given the macrosociological nature of most of these trends, many of these studies lack decisive or conclusive empirical evidence for most of their hypotheses. The insights presented in the overview therefore remain tentative and explorative.

Changes Affecting the Field of Cultural Production and Mediation

The Rise of the Entertainment Industry

The first and most obvious evolution in the field of cultural production since the 1960s is the spectacular growth of commercial popular culture and the entertainment industry (Turow, 1992). Commercial cultural offerings have become so widespread and pervasive that they challenge the relative share of the arts in people's leisure activity patterns. The fine arts and their nonprofit institutions must compete with offerings from a wide array of commercial music, films, and books produced by a booming industry, and many find it difficult to maintain their cultural centrality and the position as institutionalized cultural products that had been taken for granted in the past (Warde et al., 1999). The competition between the arts and commercial symbolic goods has become acute as the penetration rates of television rise. For example, television programs offer easily accessible at-home entertainment that – given a certain time budget – are able to successfully compete with the entertainment literature for audiences (Knulst, 1991; Van Eijck and Van Rees, 2000; Weibull, 1992). This competition has intensified with the rise of the Internet, which enables producers to constantly offer products in people's homes on demand. This creates a context of cultural abundance both for the general public and niche market audiences (Wright, 2011). Television and the new media also enable previously local cultural products to reach audiences scattered around the globe. Studies of local scenes (e.g., Cohen, 1991; Shank, 1994) show that music-making activities in a small region may spill across regional boundaries and become a translocal or even a global phenomenon (Bennett and Peterson, 2004).

Changing Cultural Hierarchies

A second trend that is inherently related to the first is the deinstitutionalization of highbrow culture (DiMaggio, 1987, 1991). Deinstitutionalization refers to the process whereby social practices formerly taken for granted and formalized in organizations are contested. In this case, the hegemonic position of the fine arts in the cultural hierarchy is being successfully challenged, paradoxically, by artists and by highly educated people themselves. Since the 1960s, the exclusive focus on highbrow arts in school curricula has dwindled; state support for culture is no longer restricted to consecrated genres or organizations staging classical music, opera, or theater. Boundaries between highbrow and lowbrow, between serious art and popular culture, are being challenged and the apparent prestige formerly attached to highbrow culture is now being questioned. The fine arts are further desacralized in postmodern artistic practices that often combine serious art with popular elements, downplaying the artistic aura that institutionalized works of art often have (Schaeffer, 1992). The attention for poetry in mass media has declined whereas the space allocated to pop music in the popular press has increased to the level of classical music. These findings are symptomatic of the change in the institutionalization of the arts and of a redefinition of cultural hierarchies (see, e.g., Janssen, 1999: pp. 337–339). In addition, culture that used to be characterized as commercial or lowbrow is now being taught at colleges and universities and/or being staged in state-subsidized art houses. This institutionalization of nonhighbrow arts further undermines the presence of a strong, universal cultural hierarchy that champions highbrow cultural participation.

Evolutions Affecting the Field of Cultural Consumption

Democratization of Education

A number of social and societal trends related to the field of cultural consumption have also shaped the evolution of cultural participation over the years. The first trend is remarkably paradoxical. Numerous empirical studies have shown a close relationship between an individual's educational level and his arts participation (e.g., DiMaggio, 1996). And yet, despite the significant increase in the overall schooling level – as well as the average level of economic prosperity – of West European and North American populations over the past

decades, levels of arts participation have not increased accordingly. The younger cohorts, especially, benefited from the democratization of education, but their higher educational levels did not lead to higher participation levels. On the contrary, in the 1960s and 1970s, the people who participated most in highbrow arts were in their 30s, whereas in the 1990s the most ardent participants in the arts were between 45 and 54 years of age. By 2000, participation peaked for people aged 55 to 64 (Van Eijck and Knulst, 2005: p. 514). This paradox can be partly explained by the change in the content of school curricula, and hence, by the type of educational regime that no longer champions the supremacy of white, Western elite highbrow culture. Instead, other forms of cultural expression – such as low- and middlebrow cultural products – also find their way into educational practices that increasingly focus on the life-worlds of adolescents, offering pupil-oriented, 'out-of-school' literacies that fall outside of traditionally canonized authors and texts (Vacca, 2002). For example, in literary education, more popular literary genres like thrillers or crime fiction are being incorporated into courses and school curricula. Therefore, educational expansion – the growing number of people with a college or university degree and the inclusion of middle- and lowbrow culture in the curriculum – erodes the rigidity of the cultural stratification system (cf Verboord and Van Rees, 2008).

Increased Rates of Mobility

A parallel evolution that is partly a consequence of educational expansion is the steady increase in rates of social mobility. Upward social mobility, especially, has a profound effect on arts participation and, more importantly, on the symbolic value attached to the arts. When the higher social strata are infused with lowbrow tastes, there are consequences to upward social mobility on both the individual and the aggregate level. On the individual level, upwardly mobile individuals are confronted with the tension between the practices and tastes of their newly acquired social position and the cultural preferences they grew up with. Whereas mobile individuals adapt their cultural behavior to their new social environment, they do not completely dismiss the activities they grew up with (e.g., Daenekindt and Roose, 2013, 2014). On the aggregate level, social mobility causes an influx of individuals originating from different social backgrounds, each with their own characteristic cultural practices. Because of the disproportionally high rates of upward social mobility – compared with downward mobility – this mingling of cultural and artistic tastes is especially tangible among the upper middle classes. The same patterns can be seen in geographical mobility, which mixes individuals from different cultural backgrounds and brings them into contact with a wide variety of formerly local symbolic goods. In that way, the consequences of globalization are tangible in both the field of cultural consumption and the field of cultural production.

The Rise of the Cultural Omnivore

As the rigidity of both the cultural and social stratification system erodes, cosmopolitan values such as being tolerant of diversity and openness are gaining ground (Inglehart, 1990). It is in this context that the cultural omnivore is introduced as a new type of cultural consumer who – in contrast to the snob typical of Bourdieu's Paris in the 1960s – crosses the boundaries between different cultural schemes (Peterson and Kern, 1996). The idea of omnivorousness suggests that cultural capital – and the prestige associated with it – relates to practices from both sides of the symbolic boundary between highbrow and lowbrow practices. The historic shift from snobbishness to omnivorousness, or the rise of the omnivore, is considered to have taken place in the 1970s and 1980s (Peterson, 1992; Peterson and Kern, 1996).

The concept of omnivorousness – and the lack of clarity surrounding the concept, especially – has generated many questions about what it actually means and how it should be measured (Peterson, 2005). This has motivated researchers to make distinctions between different forms of omnivorousness. The best known is the distinction between omnivorousness in composition – which centers on whether or not boundaries between different cultural schemes are crossed – and omnivorousness in volume – which refers simply to the breadth of taste and not the crossing of boundaries per se (Warde et al., 2008). Another question is whether omnivorousness should be studied in a single cultural sector – for example, by focusing on the variety of preferred musical genres – or whether omnivorousness should be considered across different cultural sectors, including preferences in music, films, books, clothing style, food preferences, and so forth (Van Rees et al., 1999).

These theoretical questions pertaining to the conceptual core of omnivorousness obviously have implications for its measurement in empirical research. There is also considerable debate at the methodological level about the ways omnivorousness should be measured. For example, omnivorousness has often been operationalized as a count variable by simply summing the different preferences of respondents (Peterson, 2005). This is especially problematic considering the bias toward highbrow cultural genres in most standardized surveys. If, compared with less legitimate forms, more legitimate forms of cultural participation are characterized by a greater differentiation, it is hardly surprising that individuals from higher social strata are more omnivorous than individuals from a lower social strata. Indeed, the idea that omnivorousness is exclusively situated in the upper middle classes is being contested. Most cultural sociological studies situate omnivorousness among the highly educated, upper middle classes. However, some authors – most notably Bernard Lahire (2004) – argue that omnivorousness is present in every social stratum. His argument draws on the ideas of Berger and colleagues who argue that individuals in contemporary society are subjected to heterogeneity of socializing experiences (Berger et al., 1974; Lahire, 2011 [2001]). Because of the highly differentiated character of contemporary societies, different life-worlds of individuals are now less unified than in the past. This allows individuals to be influenced by different socializing forces. This pluralization of social life-worlds is further reinforced by the eroding rigidity of the social stratification system. Thus, the individual experience of social mobility, parental/partner heterogamy, and so forth also presents the

individual with new forms of cultural activities and preferences.

Moreover, even the depiction of the rise of the omnivore as a recent historic shift has not remained unchallenged. Trends in cultural participation reveal that the presence of the cultural omnivore is actually not that recent. Cultural omnivores were already present in the Netherlands in the mid-1950s. They increased in number until 1983, but have not increased since then (Van Eijck et al., 2002: p. 156). Also, data from Denmark reveal the existence of omnivores in the 1970s (Jæger and Katz-Gerro, 2010; Katz-Gerro and Jæger, 2011). The idea of omnivorousness as a recent phenomenon is further challenged by Lahire (2004), who argues that cultural omnivorousness – or cultural dissonance, as he calls it – is already present in Bourdieu's sample of Parisians in the 1960s. Lahire thus argues that the depiction of the rise of the omnivore as a recent historic shift "confuses a change in the model of reality (the scientific point of view on the world) with a historic change of reality itself" (Lahire, 2008: p. 182).

Implications: Current and Emerging Trends in Theory and Research

What are the implications for sociological thinking of the decreasing centrality of highbrow cultural participation and the rising importance of popular cultural goods in young people's leisure time activity patterns? Certainly, from a practical point of view, if there is a decline in arts participation in younger generations, arts organizations will have to invest in marketing-the-arts strategies to replace existing audiences. Otherwise, the audience for certain performing arts, such as classical concerts, may shrink or even disappear within a couple of decades (Kolb, 2001). Further research is needed to assess why some of the arts are holding firm despite serious competition from commercial cultural goods and why others are losing ground as a form of cultural capital among the younger generations. Is it because some forms of cultural participation – for instance, visiting art museums – appeal to the more cosmopolitan dispositions of open-minded individuals seeking intellectual stimulation more than other forms, such as classical music? Does increased competition with more home-bound, web-related activities keep youth away from ballet and classical concerts? Or are other outdoor activities such as sports and attending pop/rock concerts responsible for this shift? One thing is clear – the arts have to compete with an unprecedentedly large array of other cultural goods and leisure activities; this obviously has implications for the position of arts in society and for the symbolic value or prestige associated with them.

The Remaking of Cultural Capital

The trends outlined above do not, however, indicate that class distinctions in cultural participation are waning, and that we are entering a time of individualism in which the social hierarchy no longer structures cultural practices. Empirical research on cultural participation shows repeatedly that cultural practices are still very much socially structured, and that Bourdieu's original empirical conception of the social space needs to be updated in view of recent societal developments. Bourdieu's stress on the importance of *beaux arts* and participation in highbrow culture as essential to cultural capital may have been overstated and was limited to France in the 1960s. Recently observed patterns and trends suggest that the composition of cultural capital is being remade (Bennet et al., 2009; Hanquinet et al., 2014), and that the distinctive social force of cultural capital is increasingly made up of a passing knowledge of many cultural domains – from film to literature to sports – and of cultural schemes that range from highbrow to lowbrow. *Multicultural capital* is perhaps a more accurate term, as it signals the change from the rather one-sided dependence on the highbrow arts to a broader, more inclusive approach (Bryson, 1997). Perhaps, to differing degrees and within and across different cultural sectors, everybody is becoming omnivorous. This does not falsify Bourdieu's analytic framework as presented in *La Distinction*, as Chan and Goldthorpe (2007), for example, have suggested. Instead, this perspective downplays the role of more snobbish, exclusive forms of cultural participation as socially distinctive and meaningful resources and favors a more open, inclusive, and multicultural attitude toward cultural goods. Some studies are now finding that it is precisely the variety of tastes – instead of an exclusive interest in the arts – and the heterogeneous networks associated with that variety that are related to different life chances, such as the chance of finding a job (Lizardo, 2013). Another related issue that needs further inquiry pertains to the significance of omnivorousness. Does omnivorous taste among highly educated youth indicate a more open and cosmopolitan attitude, as some researchers have suggested (Bryson, 1997; Bryson and DiMaggio, 2000)? Or is omnivorousness only one of the many manifestations – for the young and highly educated – of the underlying trait, openness – manifestations that otherwise vary according to the period one is socialized in (see Roose et al., 2012)?

Symbolic Boundaries and Social Inequality

Future empirical research may focus on whether cultural hierarchies are actually shared between different social groups, as most empirical studies implicitly assume. However, it should also consider whether there are variations. For example, how are competing hierarchies mobilized by various groups in social conflicts and under what conditions are they being institutionalized in the educational system, for example? It is in such institutional contexts especially that symbolic violence is exerted and translated into the perpetuation of social inequality: students whose cultural participation coincides with practices from the top of the cultural hierarchy have a greater chance of educational success (see, e.g., DiMaggio, 1982). If symbolic boundaries between different forms of cultural participation are indeed waning, and if this process is reflected in institutions such as the educational system, the relationship between cultural participation and status boundaries may change. Thus, research on trends in cultural participation sheds a light on the cultural dimension of social stratification.

See also: Aesthetic Education; Art and Socialisation; Art and Socialisation; Cultural Capital and Education; Cultural Capital and Education; Culture, Production of: Prospects for the Twenty-First Century; Globalization and World Culture; Leisure and Cultural Consumption: US Perspective; Social Inequality and Schooling; Social Inequality in Cultural Consumption Patterns; Social Stratification; Symbolic Boundaries.

Bibliography

Bennett, A., Peterson, R., 2004. Music Scenes: Local, Translocal & Virtual. Vanderbilt University Press, Nashville.

Bennet, T., Savage, M., Silva, E., Warde, A., Gayo-Cal, M., Wright, D., 2009. Culture, Class, Distinction. Routledge, London.

Berger, P., Berger, B., Kellner, H., 1974. The Homeless Mind. Pelican Books, Middlesex.

Bourdieu, P., 1979. La Distinction: Critique Sociale du Jugement. Editions de Minuit, Paris.

Bourdieu, P., 1986. The forms of capital. In: Richardson, J.G. (Ed.), Handbook of Theory and Research for the Sociology of Education. Greenwood, New York, pp. 241–258.

Bryson, B., 1996. "Anything but heavy metal": symbolic exclusion and musical dislikes. American Sociological Review 61 (5), 884–899.

Bryson, B., 1997. What about the univores? Musical dislikes and group-based identity construction among Americans with low levels of education. Poetics 25 (2–3), 141–156.

Bryson, B., DiMaggio, P., 2000. Public attitudes towards cultural authority and cultural diversity in higher education and the arts. In: Blake, C. (Ed.), The Arts of Democracy: The State, Civil Society and Culture. Woodrow Wilson Center Press, Princeton, NJ.

Chan, T.W., Goldthorpe, J., 2007. Social stratification and cultural consumption: music in England. European Sociological Review 23 (1), 1–19.

Cohen, S., 1991. Rock Culture in Liverpool. Popular Music in the Making. Oxford University Press.

Coulangeon, P., 2005. Social stratification of musical tastes: questioning the cultural legitimacy model. Revue Française de Sociologie 46, 123–154.

Daenekindt, S., Roose, H., 2013. A mise-en-scène of the shattered habitus: the effect of social mobility on aesthetic dispositions towards films. European Sociological Review 29 (1), 48–59.

Daenekindt, S., Roose, H., 2014. Social mobility and cultural dissonance. Poetics 42 (1), 82–97.

DiMaggio, P., 1982. Cultural capital and school success: the impact of status cultural participation on the grades of U.S. high school students. American Sociological Review 47, 189–201.

DiMaggio, P., 1987. Classification in art. American Sociological Review 52, 440–455.

DiMaggio, P., 1991. Social structure, institutions, and cultural goods: the case of the U.S. In: Bourdieu, P., Coleman, J. (Eds.), Social Theory for a Changing Society. Westview Press, Boulder, CO.

DiMaggio, P., 1996. Are art-museum visitors different from other people? The relationship between attendance and social and political attitudes in the United States. Poetics 24, 161–180.

DiMaggio, P., Mukhtar, T., 2004. Arts participation as cultural capital in the United States, 1982–2002: signs of decline? Poetics 32, 169–194.

Emmison, M., 2003. Social class and cultural mobility: reconfiguring the cultural omnivore thesis. Journal of Sociology 39, 211–230.

Erickson, B., 1996. Culture, class, and connections. American Journal of Sociology 102, 217–251.

Hanquinet, L., Roose, H, Savage, M., 2014. The eyes of the beholder. Aesthetic preferences and the remaking of cultural capital. Sociology 48 (1), 111–132.

Inglehart, R., 1990. Culture Shift in Advanced Industrial Society. Princeton University Press.

Jæger, M.M., Katz-Gerro, T., 2010. The rise of the eclectic cultural consumer in Denmark, 1964–2004. The Sociological Quarterly 51, 460–483.

Janssen, S., 1999. Art journalism and cultural change: the coverage of the arts in Dutch newspapers 1965–1990. Poetics 26 (5–6), 329–348.

Katz-Gerro, T., Jæger, M.M., 2011. Top of the pop, ascend of the omnivores, defeat of the couch potatoes: cultural consumption profiles in Denmark 1975–2004. European Sociological Review. http://dx.doi.org/10.1093/esr/jcr058. Advance online publication.

Knulst, W., 1991. Is television viewing substituting reading? Changes in media usage 1975–1985. Poetics 20, 53–72.

Knulst, W., 1993. The gentrification of a rearguard. An attempt to explain changes in the extent and composition of the arts public in the age of television. In: Rigney, A., Fokkema, D. (Eds.), Cultural Participation—Trends since the Middle Ages. John Benjamins Publishing, Amsterdam, pp. 193–215.

Knulst, W., Kraaykamp, G., 1998. Trends in leisure reading: Forty years of research on reading in the Netherlands. Poetics 26, 21–41.

Kolb, B., 2001. The effect of generational change on classical music concert attendance and orchestra's responses in the UK and US. Cultural Trends 11 (41), 1–35.

Lahire, B., 2004. La Culture des Individus: Dissonances Culturelles et Distinction de Soi. La Découverte, Paris.

Lahire, B., 2008. The individual and the mixing of genres: cultural dissonance and self-distinction. Poetics 36, 166–188.

Lahire, B., 2011[2001]. The Plural Actor. The Polity Press, Cambridge.

Lizardo, O., 2013. Variety in cultural choice and the activation of social ties. Social Science Research 42, 321–330.

Marsden, P., Swingle, J., 1994. Conceptualizing and measuring culture in surveys: values, strategies and symbols. Poetics 22, 269–289.

Peterson, R., 1992. Understanding audience segmentation, from elite and mass to omnivore and univore. Poetics 21, 243–258.

Peterson, R., 2005. Problems in comparative research: the example of omnivorousness. Poetics 33, 257–282.

Peterson, R., Kern, R., 1996. Changing highbrow taste: from snob to omnivore. American Sociological Review 61, 900–907.

Peterson, R., Simkus, A., 1992. How musical tastes mark occupational status groups. In: Lamont, M., Fournier, M. (Eds.), Cultivating Differences: Symbolic Boundaries and the Making of Inequality. The University of Chicago Press, pp. 152–186.

Roose, H., Vander Stichele, A., 2010. Living room versus concert hall. Patterns of music consumption in Flanders. Social Forces 89 (1), 185–208.

Roose, H., Van Eijck, K., Lievens, J., 2012. Culture of distinction or culture of openness? Using a social space approach to analyze the social structuring of lifestyles. Poetics 40 (6), 491–513.

Schaeffer, J.-M., 1992. L'art de l'age Moderne. L'esthétique et la Philosophie de l'art du XVIIIe Siècle à Nos Jours. Gallimard, Paris.

Shank, B., 1994. Dissonant Identities. The Rock 'n' Roll Scene in Austin, Texas. Wesleyan University Press, Hanover.

Turow, J., 1992. Media Systems in Society: Understanding Industries, Strategies, and Power. Longman, New York.

Vacca, R., 2002. Making a difference in adolescents' school lives: visible and invisible aspects of content area reading. In: Farstrup, A.E., Samuels, S.J. (Eds.), What Research Has to Say about Reading Instruction. IRA, Newark, DE.

Van der Voort, T., 1991. Television and the decline of reading. Poetics 20, 73–89.

Van Eijck, K., 1999. Socialization, education, and lifestyle: how social mobility increases the cultural heterogeneity of status groups. Poetics 26, 309–328.

Van Eijck, K., De Haan, J., Knulst, W., 2002. Snobisme Hoeft Niet Meer. De Interesse voor Hoge Cultuur in een Smaakdemocratie. Mens en Maatschappij 77 (2), 153–177.

Van Eijck, K., Knulst, W., 2005. No more need for snobbism: highbrow cultural participation in a taste democracy. European Sociological Review 21 (5), 513–528.

Van Eijck, K., Van Rees, K., 2000. Media orientation and media use: television viewing behavior of specific readers from 1975 to 1995. Communication Research 27 (5), 574–616.

Van Rees, K., Vermunt, J., Verboord, M., 1999. Cultural classifications under discussion: latent class analysis of highbrow and lowbrow reading. Poetics 26 (5–6), 349–365.

Verboord, M., Van Rees, K., 2008. Cultural classifications in literary education: trends in Dutch literary textbooks, 1968–2000. Cultural Sociology 2 (3), 321–343.

Warde, A., Martens, L., Olsen, W., 1999. Consumption and the problem of variety: cultural omnivorousness, social distinction and dining out. Sociology 33 (1), 105–127.

Warde, A., Wright, D., Gayo-Cal, M., 2008. The omnivorous orientation in the UK. Poetics 36, 148–165.

Weibull, L., 1992. The status of the daily newspaper. What readership research tells us about the role of newspapers in the mass media system. Poetics 21 (4), 259–282.

Wright, D., 2011. Making tastes for everything: omnivorousness and cultural abundance. Journal for Cultural Research 15 (4), 355–371.

Cultural Policies in East Asia

Nobuko Kawashima, Doshisha University, Kyoto, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cultural policy research is a relatively new, emerging discipline of academic studies, which has particularly suffered from the lack of literature in the English language in the Asian region. This article discusses some of the common or different trajectories of cultural policy development in East Asian countries, including China, Japan, Taiwan, South Korea, and Singapore, on a few selected issues that are relatively well documented in English. It pays attention to the influence of the economic development style commonly seen in the region on the development of cultural policy in each country. Differences in political histories have also had implications for cultural policy development, particularly at the initial stage. In recent years, the creative and cultural industries policies are similarly pursued across the countries. The transnational flow of popular and media culture in the region is also highlighted. Questions for more research are suggested for future research.

Introduction

Cultural policy research is a relatively new, emerging discipline of academic studies. Although it is still debatable whether cultural policy research is a discrete area of study or a unifying theme of various disciplines, the growth of interest in this area among researchers from a wide range of disciplines such as sociology, history, and economics is manifest in many occasions. To provide concrete examples, the International Journal of Cultural Policy (IJCP) has produced 18 volumes so far, and the number of issues per volume has gradually increased from two to five in response to the growth of submissions. Associated with the above journal is an international conference called the International Conference on Cultural Policy Research (ICCPR). Since 2000, this biannual meeting has been held seven times in various places in the world. Participants get to know about it somehow and travel far for intellectual stimulation and heated discussion during the conference period of 4 days. This is quite an achievement considering that the term (and the concept) cultural policy does not sit particularly well in the English language.

It must be noted, however, that cultural policy research has been so far largely dominated by Anglophone and European thinking, with relatively little involvement of the Asian region. I myself have been one of the few contributors to IJCP from this region, but I must admit that the majority of the papers I have published are on British or Euro-oriented issues (as my postgraduate learning and early career in academia is owed to British universities). ICCPR likewise has attracted very limited numbers of participants from countries in Asia. When we do see people 'originally' from Asia, they often discuss issues of European or American cultural policy.

It is possible to suspect that one major reason why such underrepresentation of Asia in cultural policy studies has continued is related to the lack of source material and research publications accessible in the English language. There might well be cultural policy research developed in their own languages, be it in Japanese, Chinese, or Korean. This may well be the case for Spanish, French, and German as well, so we should not assume that the lack of literature available in English alone means lack of interest or the inactivity of researchers. However, there also is a feeling that Asians do not know enough about each other's cultural policies. It often happens that cultural policy researchers in South Korea are very familiar with, for example, the American way of funding for the arts and that their counterparts in Japan are knowledgeable about different models of cultural policy in Europe but know virtually nothing about cultural policy in Taiwan. In the meantime, however, cultural flows between different nations in Asia are rapidly increasing while researchers of cultural exchange pay little attention to the policies of each country in the production and distribution of culture. Such a state of the study on, and in, Asia is becoming an urgent issue. Nowadays, we may encounter speculative and sporadic comments on some specific aspect of cultural policy of a nation in the region on an ad hoc basis, but they have not yet been established as a systematic, common knowledge.

Given such a state of research, the present article on cultural policies in East Asia cannot be overambitious. It will discuss some of the common or different trajectories of cultural policy development in East Asian countries on a few selected issues that are relatively well documented in English (and in Japanese). However, for the reasons of the limited space and resources to draw on, it will not be able to suggest any analytical 'regime(s)' of the cultural policies in the region. In this way, it will be a very modest contribution of a preliminary nature to the study of cultural policy research with a geographical focus on East Asia.

A few caveats must be noted here. The article will not be able to provide country-by-country descriptions of cultural policy due to the lack of necessary data. It will not be a comprehensive overview of the countries included in the study on specific issues, either. Rather, it will be a discussion of commonalities and general trends with references to information which I have managed to gather. The countries included are Japan, China, South Korea, Taiwan, and Singapore. Geographically, Singapore is a little far from the others and could raise a question of qualification. However, for the purposes of this article, given the relative ease of access to information on Singapore and considering the similarity of economic development patterns and standards of living between Singapore and the other countries (except China), it

makes more sense for the city-state to be included. It is assumed that the similarities in economic and political structures between them will have more implications to cultural policy formation and development than geographical proximity alone.

In the following, the first part will provide some basic facts and backgrounds to cultural policies in East Asian countries. Similar features as well as differences will be discussed. The article proceeds to the second part to discuss developments of cultural policies in these countries over the years with particular attention to rationales of governments' involvement in the area of the arts and culture. The third part of the article will focus on the recent emergence of the creative industries policy. Finally, the fourth part will discuss exchange and flow of media and popular culture in the recent decades among East Asian countries and its implications for cultural policies.

Brief Description of Countries in East Asia

The five countries dealt within this article are widely different in terms of territorial size, ranging from the vast giant China to the smallest 'city-state' Singapore. The same is true of population size, ranging from 1340 million Chinese (even excluding those in Hong Kong and Macau) to a mere 5 million Singaporeans. Such differences in size are well reflected in the size of the economies, with China having recently ascended to the second largest in the world in terms of gross domestic product (GDP) overtaking Japan, and with Singapore for the smallest GDP of the five. A comparison of GDP per capita, however, will show that China lags far behind the rest, with Singapore and Japan forming the wealthy group and South Korea and Taiwan following them (**Table 1**). Overall, however, these nations boast the strongest economic successes and industrial bases in the Asian region as a whole (see Vogel, 1991 on the 'four little dragons' – South Korea, Taiwan, Singapore, and Hong Kong)

What characterizes these countries in terms of economic development is that all of them developed in a relatively short period of time. Japan, as the first of these, turned from an agricultural feudal state to a modern, industrialized economy at around the turn of the nineteenth century. After the loss of World War II, very rapid development occurred again from the 1950s to 1970s to achieve the current state of economic power. South Korea and Taiwan were former Japanese colonies, as was Singapore of Britain. After independence from the 1940s to 1960s, however, these three nations have developed industries and economies, in only three to four decades, transformed from poor, agricultural lands to globally competitive nations of wealth with a large role played by government to guide the development of the economies, and presented a miracle story to the rest of the world. The recent rise of China to the world top league of economic power is a very different story, consisting of a major transformation from a Socialist economy to a more liberal, open economy. The speed of this process is equally short, spanning from the late 1970s to today.

Table 1 Population (2010) and economy size of the five East Asian countries (2011)

Country	*Population (million, 2010)*	*GDP nominal (billion USD, 2011)*	*GDP per capita (USD, 2011)*
China	1341	5739	5 417
Japan	127	5459	45 870
South Korea	48	1014	24 900
Taiwan	23	466	20 083
Singapore	5	334	49 271

World Population Prospects: The 2010 Revision, The United Nations; World Economic Outlook Database 2012, The International Monetary Fund.

Also common to these countries is that they have been situated in the sphere of Chinese civilization and culture for centuries. China has often been a threat to the neighboring countries, who however have historically learned a great deal of civilization from mainland China. Confucian philosophy and values coming from China underline the cultures of these five nations. In turn, the people in these nations are generally characterized as self-disciplined, hard working (not minding working long hours), valuing education for children, meritocracy, and collectivism. These traits, no doubt, have contributed to the state-initiated programmes of industrialization and economic development described above.

Development of Cultural Policy

What relevance and implications do this economic transformation and the general culture of the nations described above have for cultural policy development in these states? One might guess that modernization generally results in an influx of Western culture and causes concerns for the states that their own cultures were under threat. This may well lead to speculation that economic achievements lead to ample funding for the preservation of traditional culture in response to 'over-Westernization.'

To some extent, such an imagined scenario of cultural policy development is reasonable, but the reality is far more mixed. At one extreme is China, with its strong, typically Socialist/Communist cultural policy, whereby art and literature in particular can be produced and supported only when it meets the state's political will and vision. Culture in China, in other words, has been much repressed and survived when it served the ideological cohesion of the party state (Tong and Hung, 2012). In other countries, however, the perceived threat was not unitary, while Western influences were not necessarily viewed with alarm. In Taiwan, for example, the Japanese colonial period (1895–1945) has left a major cultural influence in a wide variety of ways. While some resent it, the Japanese rule is often (and paradoxically) recalled with nostalgia, seen as a productive influence for paving the way for the postwar development that was spurred with American aid. At the same time, the elite class of the Taiwanese are migrants from mainland China who came to Taiwan with the KMT Party in the late 1940s (called the Mainlanders), who do not associate with the legacy of Japanese influence. However, their political relevance to mainland China is complex, prohibiting them from excessively asserting their Chineseness.

Such difficulty in finding a core national identity in Taiwan to develop cultural policy around is no less troubling than the experience of European cultural policy, which was generally

developed around the traditional, White, European culture. Since the 1970s, however, other cultures have started to contest this structure and claimed their places in cultural policy, too. In Taiwan, the Chinese Cultural Revitalization Committee was set up in the 1950s, but it does not seem to have been really about cultural development or Chinese values. Rather, it looks to be for the promotion of nationalistic ideology against the Communism that was sweeping throughout mainland China and for legitimizing the authoritarian rule of Taiwan that continued until 1987 (Chung, 2012: p. 341). In recent years, multiculturalism, or equal treatment of minority cultures, namely, those of the indigenous Taiwanese aborigines (perhaps of Malayo-Polynesian origin) and the Hakka people (minority Chinese who fled to Taiwan around the seventeenth and eighteenth centuries) have seen some development. Although attention to such cultural diversity is laudable, cultural policy is still struggling to ease tensions between diversity and a national identity and to respond to a growing recognition that cultural diversity itself is in constant change (Wang, 2004).

In South Korea, the colonial period (1910–45) has been remembered with hatred, indignation, and deep anger, because of the repressive nature of Japanese dominance and inhumane treatment of Koreans in Japan during and after the war. Thus, the elimination and exclusion of Japanese cultural influence has been one important aspect of cultural policy development in Korea since its independence. For example, Japanese popular culture (e.g., films, popular music, comic books, etc.) was banned in Korea until 1998 when the ban started to be lifted incrementally. Other factors that have shaped Korean cultural policy include the disruption of traditional culture by colonization and rapid Westernization, which has led the Korean government to keenly revitalize it. The cultural policy of South Korea has also developed, particularly in the years after the North–South division of 1945, as a means of building the unified nation against the Communist North Korea (Yim, 2002). More broadly, the authoritarian and militarian rule that remained until 1987 had also created a strong presence of government in the national life of culture.

In Singapore and Japan, in the early decades of struggles to move from the state of poverty (the 1950s in Japan and the 1960s in Singapore), support for the arts and culture was never on the agenda of policy makers. The national preoccupation in both countries was first and foremost to raise the standard of living for citizens by increasing the national wealth and by building the necessary infrastructure for transport, energy production and transmission, communications, and much more. Among welfare state measures, education, housing, and health care were given sufficient attention, as they would serve the quality of labor force necessary for economic development. Quality of life, distinct from the standard of living that suggests physical conditions rather than personal happiness and well being, has come onto the agenda only in the late 1980s in both countries when they acquired confidence and satisfaction with their economic achievements.

A major difference between Singapore and Japan in this context, however, is that Singapore had to build a 'nation,' which was possible with the vision of the charismatic leader Lee Kuan Yew about what the national culture must be like. Singapore as a country played down its link to Chinese culture to keep a distance from mainland China, and instead emphasized the 'local' arts and culture, hoping that cultural policy would nurture patriotism. Not to emphasize Chineseness is related to the multiethnic composition of the nation as well. The country did not choose a free enterprise economy in the Hong Kong style, but followed the course of a 'developmental state,' where strategic industrial policy is vigorously pursued by constraining and harnessing market forces by the state (Öniş, 1991), while a major concern of cultural policy has continued to be the harmonization of different ethnic groups as a nation.

In the case of Japan, similarly, economic objectives were given a high priority supplemented by state welfare only to the extent necessary to achieve these objectives. It had, however, had no need to deliberately define or fabricate national identity, as a relatively homogeneous nation and the project of nation building was finished in the late nineteenth century. The fact that the Japanese have keenly embraced American culture, values, institutions, and lifestyle, shows that there has been little fear that the American influence would destroy or taint the Japanese culture, despite the result of the war fought. Moreover, unlike Taiwan, South Korea, and Singapore, any hint of patriotism has been avoided in Japan in an effort to eradicate the remains of the authoritarian rule of the pre- and interwar years and build a democratic, free country. This was a major policy of the American occupation of Japan postwar years. Thus, whereas Singapore still has measures to monitor the 'appropriateness' of culture of citizens (see Lim, 2012 for the use of a major arts festival in Singapore) and retains censorship, Japan has had no hesitation to sign up for Western ideals and values such as freedom of expression. With little perceived danger of cultural imperialism from the West, Japan has also had the basic attitude of *laissez-faire* for the sphere of culture, a factor that in fact contributed to the weak basis of cultural policy development in the following decades.

Contemporary cultural policies of these countries, however, have tried to follow the European style, with government funding for cultural and artistic institutions such as museums, concert halls, theaters for Western classical arts (such as opera, classical music, and ballet). Issues that have been debated in European cultural policy such as multiculturalism, improving access to culture, and audience development are more or less shared by the East Asians of today, too. It is typical of East Asian nations (except China) to keenly learn from advanced nations of relevance and form similar institutions for economic, political, or social development. Japan, for example, historically (although intermittently) has sent engineers and scholars to China up to the seventeenth century or invited experts from Europe and the United States in the nineteenth and twentieth centuries for her modernization project (Kawashima, 2012: p. 295). However, simplistic emulation or importation of cultural policy from Europe has often failed to work, not necessarily because of a shallow understanding of the model but rather due to the lack of genuine commitment to the purpose of learning, namely, cultural development.

For example, South Korea has recently set up a 'British style' arts council with the aim of instigating cultural policy making at an arm's length. The new institution, however, has failed to create a public funding system independent of political interference. It has been shaped merely as a grant-making organization but its policy-making capacity has been limited

(Lee, 2012). The notion of an arts council (which could actually be anything) has attracted attention in Japan as well. The positive aspects of the British arts council have been introduced for about 20 years to policy makers by cultural policy researchers who advocate for an independent, specialist-led way of cultural policy making. Nonetheless, what has happened in the end (only in 2012) is the introduction of a few 'programme directors' and 'programme officers' (neither of which job titles has been familiar to the Japanese) in a public organization, the largest of its kind in Japan, that provides funds to performing arts organizations. The newly appointed directors and officers are expected to contribute to a strategic, specialist kind of cultural policy making, going beyond the bureaucratic, mechanical allocation of funds.

As another example, cultural policy in Japan has been developed in the late 1980s through the 1990s as an effort to improve her national image in the international arena, changing from the aggressive business people to a respectable nation with sophisticated culture and tradition. The imitation of European-style cultural policy has, however, been only superficial, as investment has heavily been skewed toward capital work, namely, the construction of cultural facilities such as museums and theaters, particularly by local government expenditure. In fact, it can be seen as a policy for public construction work, leaving very little money for creative activities and artists. Even worse, many of the personnel seconded from local authorities to these facilities knew little about the arts or had any skills required of arts management. Thus, although first class, flagship facilities for culture have been constructed all over the country in the two decades, it cannot be said that artistic performances and exhibitions made by professionals have accordingly blossomed. Unfortunately, it cannot be said either that many grass roots activities by local residents have taken place at these venues and significantly contributed to community development (Kawashima, 2012: pp. 301–302).

Recent Trend – Cultural and Creative Industries Policy

With globalization progressing in the economy, technology, information, and culture, despite differences in developments and current forms of policy making amongst themselves and from the Western models (on which cultural policy literature has focused), cultural policies of the East Asian nations have picked up a major trend of cultural policy elsewhere: the cultural and creative industries policy. This is not surprising, considering the transition of their economic bases from manufacturing and service to value added, knowledge creation sectors. As the countries prosper, their labor costs increase and they inevitably lose manufacturing bases to neighboring, developing economies such as Thailand, Malaysia, and Vietnam. Interest in the notion of the creative economy, creative cities, and the creative industries has been shown in policy discourse as well as in academia.

Singapore is also at the forefront of this policy direction. Being close to Australia and internationally minded with good English widely spoken, Singapore seems to be more flexible and quick in adapting to the global trend in policy. It has been very explicit about its ambition to become a global city, in which the cultural and creative industries are given a high profile. In fact, it has a relatively long history of having a cultural policy aligned with the development of tourism (thus economic) as well as cultural policy for noneconomic purposes like creating political unity. More concretely, during the 1980s and 1990s, sizable public funding has gone into the construction of flagship arts venues and museums, with the supporting slogan of Singapore as a global (or at least regional) hub for the arts and culture (Kwok and Low, 2001; Kong, 2012: p. 282). Located at the crossroads of the Southeast, East, and South Asian regions and in close proximity to Australasia, the slogan has not been unrealistic. As an example, long-run musicals 'imported' from Broadway and the West End are sustainable and thriving in Singapore. Together with good restaurants and countless shopping malls, Singapore can attract upper-middle classes and expatriates living in the Philippines, Thailand, and Vietnam for safe, clean, urban, and entertaining weekends. These people obviously go to Hong Kong as well for roughly the same purpose or to Macau for casinos, but Singapore competes with more contemporary artistic events of various kinds throughout the year.

Singapore in the 2000s has gone beyond this stage and articulated a policy for the creative industries with clear targets and implementation plans. Three areas have been identified as the focal sectors of the creative industries policy, namely, arts and culture, media, and design. In order to increase the contribution of these industries to the national GDP from 3% in 2004 to 6% by 2012, a number of measures have also been announced, such as attracting investment and talent to these industries both domestically and from abroad. Enhancement of creative activities in education for children and young people to nurture future creators for long-term development has also been a priority of cultural policy (Kong, 2012).

Likewise, Taiwan, in the 2000s, has been very active in promoting the cultural and creative industries (including fine arts, traditional folk arts, broadcasting and the media, fashion design, video game software as well as advertising, and so on, similar to those conceived as the creative industries by the United Kingdom government since 1997). The rationale is the same as that in Singapore: the need to convert the industrial economy that has imported electronics parts to major American and Japanese companies to a knowledge economy whereby Taiwan is no longer a downstream factory for foreign businesses. The cultural and creative industries policy is thus firmly situated alongside the biotechnology and digital content industries in the overall, long-term plan of economic transformation set out by government in 2002, and furthered in 2008 with another plan called Creative Taiwan. More concretely, the policies have emphasized the development of creative clusters and 'cultural and creative parks' (like Science Parks where research institutes of industries are clustered upon the receipt of financial incentives for relocation). Such vigorous efforts to promote the cultural and creative industries have culminated in the enactment of the Cultural and Creative Industries Development Act of 2010. Cultural policy in Taiwan in this regard, thus, has gone beyond the cabinet-level Council for Cultural Affairs (CCA) and involved ministries for economic affairs, education, transport, and much more (Chung, 2012). In the meantime, the CCA was upgraded to a Ministry of Culture

in 2012, accelerating the government's efforts to make Taiwan competitive in the world of the creative economy.

It is interesting to note that although East Asia may have been late and slow in developing cultural policy, in today's global world, the development of the creative industries policy has been parallel to that in Europe and elsewhere. Creative as an adjective modifying the name of a country or a city has been already repeated everywhere, and the same policy discourse has been heard in the international arena such as UNESCO, European Union (EU), and even in United Nations Conference on Trade and Development (UNCTAD). Despite the common policy discourse, however, there are conflicts and tensions between different interests that have varied from country to country.

In Singapore, artists are dissatisfied and see the government's slogan as being only superficial in light of the paternalistic and restrictive mode of governance in general. The practice of censorship in particular is seen to hamper individual creativity, hence the city-state's vibrancy may well be found only in a bounded, government-controlled territory (Ooi, 2010). The high expectations generally held in Singapore of efficient and quick success with measurable outputs, a feature which one may argue has helped Singapore to achieve economic success so far, may not work in the same way for the creative industries (Kong, 2012: p. 291). Another issue, which then can be linked to cultural policy for social development, is the lack of 'milieu' for cultural debates in Singapore (Kong, 2012: p. 291). It is all very well to count numbers of visitors and arts consumers (many of whom may well be from overseas) at events and celebrate their increases, but what counts in the end is that citizens of Singapore access these cultural events and exchange their views and opinions, admire or criticize the works they see. Arts, in other words, are not just to be consumed, but to be used for self-expression and social development. Such a view of cultural policy is only just emerging in Singapore, but can easily be eclipsed by preoccupation with the economic aspect of cultural policy.

In Taiwan, the cultural and creative industries policy, while predominantly economic in outlook, has started to assume cultural and social aspects as well. As a result, ironically, tension has arisen between the policy for the local and that for the global. In other words, what culture the government should promote has become a more prominent issue over the years in Taiwan with mainland China emerging as a major economic power. On one hand, more emphasis should be given to Taiwan's originality and local creativity that derive from her ethnic and cultural diversity. On the other hand, it seems that opportunities are there for the construction of a greater Chinese culture whereby Taiwanese creative industries can take advantage of the size of the Chinese supply and demand in the market of the Asian cultural economy (Chung, 2012).

While South Korea and China have also been proactive and officially proclaimed their emphasis on the cultural and creative industries policy (see Keane, 2004 on China), Japan is perhaps the least interested in or committed to this area. Like other countries in East Asia, Japan, too, has recognized the potential of the creative industries with regard to their contribution to the national economy and export and in relation to the side effects of their activities such as attracting tourists and producing positive images of the country. Japan has undoubtedly been successful particularly in animation, comic books, and video game industries on a global scale. *Animé*, the Japanese shorthand of the English word animation, has already been established as an English word in its own right, so is *manga* when referring to comic books. For example, the Japan Expo, held annually in a convention center in peripheral Paris, draws over 200 000 visitors over a period of 4 days. Mostly young people, some with their parents, would come and do 'cosplay' (again a Japanese word shortening costume play), being dressed like their favourite animé characters, buying figurines and other goods related to animé and manga, listening to famous creators from Japan talk about their works, and so on. However, the big event, which has been in existence for 13 years attracting people from whole Europe, is a private, commercial venture, and has no particular backing from the Japanese government.

The rhetoric of the creative city, which is the cousin of the creative industries policy, has gained some prominence in Japan among city and prefectural governments. This is because they have increasingly faced financial difficulties as a result of decline in traditional manufacturing and also because of political decentralization whereby local authorities are expected to find their own ways of survival without waiting for subsidies to come from the center. One solution in this context has been the policy of making cities more creative not just in abstract terms but by promoting a range of cultural activities, of not-for-profit and for-profit natures (as the boundary is unclear anyway) and leveraging such creativity for revitalizing the local economy in a sustainable manner and toward the liveability of the places.

However, when it comes to the national level of policy for the cultural and creative industries, it is outside the jurisdiction of the central Agency for Cultural Affairs, whose responsibilities include preservation of heritage, some funding for performing arts, language policy, copyright policy, and matters related to religion. The economic and industrial promotion of the cultural industries has thus largely fallen into the realm of the Ministry of Economy, Trade and Industry (METI) and to a smaller extent among other ministries and departments for local governance, communications, tourism, and the Cabinet Office. Although this is a new area for METI, officials of the Ministry have swiftly come to understand the economic features of the creative industries and identified issues for long-term development. For example, the lack of good business managers with expertise in financing, knowledge of legal matters, and skills in global marketing has been identified as a key issue. However, most of such issues are to be tackled over the long term and are not necessarily amenable to policy intervention of a short-term nature. The projects under the Cool Japan banner are in fact eclectic and arbitrary, ranging from a show of Japanese fashion in Singapore to a promotional event of Japanese food and cuisine in Mumbai (METI, 2012). Although the METI's grasp of the issues involved in the development of the cultural and creative industries is generally sound and good, the policy of Japan for the cultural and creative industries has recently waned, devoid of sustained commitment and seriousness.

Flows of Culture within the East Asia Region

Although there has been little concerted policy of promoting the cultural and creative industries in Japan domestically as well as internationally for export as described above, Japanese

popular culture, including music, television drama, consumer fashion, magazines, animé, and manga, has been widely accepted in East Asian countries, particularly in Taiwan and South Korea despite their official ban in the past in Korea. It must be noted that not all cultural products moving national borders from Japan are through official, lawful distribution. Due largely to the complexity of copyright (and neighboring right), Japanese television programmes are not easy for secondary use, including export. Rights' clearance takes a long time before the programmes from Japan can be aired by overseas television stations. In the meantime, in response to urgent demands for Japanese products by local consumers, unlawful acquisition of the Japanese programmes has been widespread, aided by the development of digital technologies (see Hu, 2004; Davis and Yeh, 2004 on VCD, the media technology widely utilized in Asian nations except Japan).

The background to such a trend toward the expansion of potential markets for Japanese popular culture is the growth in income and rise in educational levels in East Asian countries and the demand of savvy consumers for different cultures. Deregulation of content and transmission as well as technological developments in media have undoubtedly contributed to the increase in transnational flow of culture in Asia. The consumers' desire for imported products of culture has been accompanied by the urban and cosmopolitan lifestyle shared by major cities across the region so that the relatively affluent, elite class, regardless of nationalities, can easily identify with cultural products that they see as interesting and well crafted. Japan in this regard has been advantageous and successful in a wide range of cultural sectors, but since the mid-1990s, South Korea, and China and Taiwan to a smaller extent, have also emerged as cultural exporters for East Asia. At the same time, coproduction of films for theater and television is increasing between Taiwan, Korea, Japan, and China. As a result, the flow of culture within the region is now much more multidirectional, interactive, and intersecting (Iwabuchi, 2011).

Hallyu (Hanryu), or the Korean Wave, has been a successful export from South Korea to Asian countries (and beyond) of television drama and pop music. In Japan, too, starting with the success of *Winter Sonata* (a TV drama), the boom of Hallyu has been phenomenal and surprising, given that one of the most disliked nations for the Japanese has been the Koreans for decades (and vice versa). The Korean government's emphasis on their creative and cultural industries policy is often thought to be a major contributing factor to this rise in export, but there are more reasons for the sudden surge (Choi, 2008). One reason is the restructuring of the invested capital and new talents since the major economic crisis of East Asia in 1997. The lower price of programmes from Korea for importing countries like China than those from Japan and the United States has been another reason. The antagonism of China to capitalistic culture represented by that of the United States and Japan has put Korean products in a privileged position, too.

While such a transregional traffic of cultural products is, in principle, celebrated, it will be important to question whether the flow has contributed to a mutual understanding of different nations and cultures. With the rise of Hallyu, for the first time in postwar Japan interest in and respect for Korean culture is blossoming. There is some evidence as well to suggest that younger Koreans who are big fans of Japanese popular culture have positive images of Japan. Nonetheless, social discrimination against resident Koreans, who were brought to Japan in the war years and their descendants born and raised in Japan, persists. Hallyu thus even bothers people of Korean origin permanently residing in Japan who see the Korean Wave in Japan as being only superficial in the confines of commercial media culture (Iwabuchi, 2011: pp. 270–271; see also Jung, 2009 on Korean popular culture in Asia and the United States). Conventional cultural policy measures for international exchange and mutual understanding have been, however, largely separated from media and commercial culture.

It is possible to speculate on the existence of similar issues of cultural politics elsewhere, complicated by the increased flow of media culture in the East Asia region. As has been mentioned, Taiwan and mainland China experience conflicts over legitimate Chineseness. Within China alone, while her cultural policy emphasizes the promotion of traditional Chinese culture to enhance her image abroad, domestically, it has also made heavy investment into the preservation of cultural traditions of ethnic minorities. The latter policy of cultural preservation has included publishing folk songs, mythologies, and local histories as well as the renovation of architectural symbols of ethnic minorities around the country (Tong and Hung, 2012: pp. 274–275), inviting a criticism from abroad that this policy is only to build theme parks of culture without genuine commitment to and recognition of cultural diversity within the vast land.

Conclusion

This article has explored differences and similarities of cultural policies in East Asian countries. I have discussed the quick development of economies in the 'developmental state' style commonly seen in this region, which has largely left the development of cultural policy behind. The need for nation building, however, has provided a strong basis for cultural policy emergence particularly in Singapore and South Korea, but much less in Japan. Taiwan will need a little more time to work out her policy on cultural diversity while aiming to become a major player in the region's cultural economy. The influence of and historical link to China of the other nations has had different implications for cultural policy development. Taiwan and Singapore at the beginning buried their Chineseness, but the recent rise of China as an economic power is tempting these countries to forge new links with her so as to enlarge their market for the creative industries. The countries in East Asia today pursue the policy to promote the creative and cultural industries, although the degree of commitment and public investment varies. Japan is perhaps the least serious about the value of these industries, but has been very successful in exporting Japanese popular culture to the countries in East Asia. In recent years, however, the transnational flow of cultural products is no longer only from Japan, but multisectional in accordance with cultural, economic, and social globalization in general and in response to the opportunities provided by media deregulation.

As was mentioned at the beginning, given the insufficiency of descriptive data and research material to draw on, the article has only been exploratory and preliminary, leaving much to be developed for more analytical and theoretical research.

Numerous questions remain to be asked. In what ways have non-Western, traditional cultures of the region impacted upon the development of cultural polices? Does the relationship between social class and consumption of culture in East Asian countries, generally different from that of the West, have any bearing on their cultural policies? Is there an East Asian cultural policy model distinctive from those of Europe and the United States that have been much discussed in existing research? What lessons can we learn from East Asia particularly for other countries in Asia? There will be much more work to be undertaken, but all of these questions would help to advance cultural policy research, and further challenge some of the arguments in the literature on global culture that has also been dominated largely by researchers from the West. It may take a while for emerging research to mature, but promisingly, researchers from these countries are beginning to form networks for furthering East Asian cultural policy research.

See also: Artists, Competition and Markets; Cultural Policy Regimes in Western Europe; Culture Policy Regimes in China; Globalization and World Culture.

Bibliography

Choi, Jaz, 2008. The new Korean wave of U. In: Anheier, Helmut K., Isar, Yudhishthir R. (Eds.), The Cultural Economy. Sage, London.

Chung, Hsiao-Ling, 2012. Rebooting the dragon at the cross-roads? Divergence or convergence of cultural policy in Taiwan. International Journal of Cultural Policy 18 (3), 340–355.

Davis, Darrell William, Yeh, Emilie Yueh-yu, 2004. VCD as programmatic technology: Japanese television drama in Hong Kong. In: Iwabuchi, Koichi (Ed.), Feeling Asian Modernities: Transnational Consumption of Japanese TV Dramas. Hong Kong University Press, Hong Kong.

Hu, Kelly, 2004. Chinese re-makings of pirated VCDs of Japanese TV dramas. In: Iwabuchi, Koichi (Ed.), Feeling Asian Modernities: Transnational Consumption of Japanese TV Dramas. Hong Kong University Press, Hong Kong.

Iwabuchi, Koichi, 2011. Cultural flows: Japan and East Asia. In: Bestor, Victoria Lyon, Bestor, Theodore C., Yamagata, Akiko (Eds.), Routledge Handbook of Japanese Culture and Society. Routledge, London and New York.

Jung, Eun-Young, 2009. Transnational Korea: a critical assessment of the Korean wave in Asia and the United States. Southeast Review of Asian Studies 31, 69–80.

Kawashima, Nobuko, 2012. Corporate support for the arts in Japan: beyond emulation of the Western models. International Journal of Cultural Policy 18 (3), 295–307.

Keane, Michael, 2004. Brave new world: understanding China's creative vision. International Journal of Cultural Policy 10 (3), 265–279.

Kong, Lily, 2012. Ambitions of a global city: arts, culture and creative economy in 'Post-Crisis' Singapore. International Journal of Cultural Policy 18 (3), 279–294.

Kwok, Kian W., Low, Kee H., 2001. Cultural policy and the city-state: Singapore and the "New Asian renaissance". In: Crane, Diana, Kawashima, Nobuko, Kawasaki, Ken'ichi (Eds.), Global Culture: Media, Arts, Policy and Globalization. Routledge, New York.

Lee, Hye-Kyung, 2012. Progress without consensus: 'instituting' Arts Council in Korea. International Journal of Cultural Policy 18 (3), 323–339.

Lim, Lorraine, 2012. Constructing habitus: promoting an international arts trend at the Singapore Arts Festival. International Journal of Cultural Policy 18 (3), 308–322.

METI, 2012. Cool Japan Strategy. Ministry of Economy, Trade and Industry, Tokyo.

Ooi, Can-Seng, 2010. Political pragmatism and the creative economy: Singapore as a city for the arts. International Journal of Cultural Policy 16 (4), 403–417.

Öniş, Ziya, 1991. The logic of the developmental state. Comparative Politics 24 (1), 109–126.

Tong, Q.S., Hung, Ruth Y.Y., 2012. Cultural policy between the state and the market: regulation, creativity and contradiction. International Journal of Cultural Policy 18 (3), 265–278.

Vogel, Ezra F., 1991. The Four Little Dragons: The Spread of Industrialization in East Asia. Harvard University Press, Cambridge, Massachusetts.

Wang, Li-Jung, 2004. Multiculturalism in Taiwan—contradictions and challenges in cultural policy. International Journal of Cultural Policy 10 (3), 301–318.

Yim, Haksoon, 2002. Cultural identity and cultural policy in South Korea. International Journal of Cultural Policy 8 (1), 37–48.

Cultural Policy Regimes in Western Europe

Vincent Dubois, Université de Strasbourg, Strasbourg cedex, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

European countries have built cultural policy regimes, which reflect the patterns of national political histories and national cultural fields. These models consist in various ways of establishing culture as a policy domain, in the definition of cultural policy rationales, and in modes of organization and governance compatible with democratic standards. This triple specification process has never been fully achieved. It is being challenged in a way that questions these policy regimes and the notion of cultural policy they had defined.

Introduction

Cultural policy regimes reflect the histories of nation and state building, the institutional configurations, and the modes of government specific to each country. The cultural policy programs that most western European governments have elaborated during the second half of the twentieth century were based on the legacy of the plurisecular history of the relationships between the state and the arts, and between politics and culture. They were also linked with the general renewal of state intervention that consisted in the making of democratic welfare states after World War II. In that sense, contemporary cultural policies cannot be reduced to state patronage of the arts. They have defined their specific rationales, practices, and modes of organization, according to the national political and institutional patterns, and following their changes through time. Cultural policy regimes also reflect the patterns of the national cultural fields, that is to say their socioeconomic structure, their internal hierarchies, and the conceptions of art and culture that prevail within them. The respective importance of heritage and of the cultural industries, the polycentric or centralized organization of the arts, and the social distribution of tastes and cultural participation are part of the evolutive factors that shape specific national cultural policies which, in turn, impact them.

The evolutions of cultural policies result from the interaction between these two levels, the political and the cultural one. They mostly consist in adding new layers of institutions, orientations, and discourses rather than in radical changes, in a way that contemporary cultural policies can be viewed as the sedimentation of previous ones. As a result, a national cultural policy encapsulates the main features of the national political and cultural histories. This is the reason why cultural policy is a fascinating topic for an international comparison that brings together political, cultural, and social aspects. But this sedimentation and multidimensional aspects also complicate a cross-national perspective. Three series of questions help to identify the principles of differentiation between the various cultural policy regimes and to account for their evolution. How is cultural policy defined as a specific policy domain, and what are the issues at stake in this definitional process? What are the main rationales for cultural policy? Finally, what is the relationship between the organizational structure of cultural policy and the cultural and political issues it raises? By addressing these questions, we will shed light on the process of making a specifically 'cultural' policy and see how it is being increasingly challenged by ongoing political and cultural changes in Europe.

Defining Cultural Policy

Historical Legacies

The political uses of the arts as symbols of power are the most common features of the protohistory of modern cultural policies in Europe. Princes, lords, and monarchs all over the continent have commissioned and accumulated works of art; decorated palaces; supported artists; and funded music, theater, and festivals in order to assert their grandeur. During the Renaissance, when the private fortune of the ruling king or queen was distinguished from the public possessions of the state, this personal and private patronage progressively turned into a state affair. Specific services were created to administrate the arts. Beyond the individual pleasures and fame of the monarch, the magnificence, prestige, and moral authority of the state, in competition with the church, was from henceforth at stake.

If they share this common legacy, cultural policy regimes differ according to the various national histories of nation-building and of the making of the modern state. In countries which saw early unification of the nation and building of a powerful state, the arts and culture were directly related to these processes. France certainly is the most obvious example of this link. Linguistic unification was a means to exert power throughout the national territory, and also the condition for the early invention of a national literature, which, in return, contributed to the adhesion to a unified politico-cultural community. In that sense, the creation of the French Academy (*Académie française*) in 1635 was both a literary and linguistic contribution to the making of the French nation-state, and a political contribution to the formation of the national literary field. More generally, political and cultural centralization were joint processes. As a result, a unified state, a national culture, and the subsequent link between state and culture were historically made self-evident. On the contrary, the history of nation and state building in more recently unified countries leads to totally different consequences in terms of cultural policy regime. The history of the relationship between the arts and politics in Italy, for instance, has been the history of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10419-2

the competition between cities for several centuries (Castelnuovo and Ginzburg, 2009). The political and artistic polycentrism of the peninsula rooted in the Italian Renaissance has had long-lasting effects in terms of cultural policy. In other cases, such as Finland, the relationship between culture and politics was defined in the scope of a late and impeded nation-building, in order to promote a nationally representative art after 1860, and in order to create a sense of national community during the first decades of the twentieth century (Sokka and Kangas, 2007).

The political uses of culture can also result in illegitimate legacies that have hindered the founding of cultural policy regimes in the contemporary era. This has occurred in countries in which authoritarian and fascist regimes have manipulated the arts and culture. The main example is Nazi Germany, during which arts censorship and political propaganda through culture was stronger than anywhere else in Western Europe, with a direct impact in terms of reluctance toward a state-driven cultural policy in West Germany after World War II. To a lesser extent, a negative cultural legacy has impeded the making of democratic cultural policies after authoritarian regimes in Italy, Spain, and Portugal. Contemporary cultural policy regimes result from these various national histories, which have led to various notions of viewing culture and its political aspects, and provided various legacies in terms of institutions, types of funding, and modes of organization.

Specifying a Policy Domain

The definition of a cultural policy does not, however, only consist in these historical legacies. Besides 'implicit' cultural policies, namely the wide set of public decisions and programs which have cultural impacts even if they are not conceived and presented as pursuing cultural goals (Ahearne, 2009), explicit and specific cultural policies have been defined as such, in most European countries, from the 1960s onward. National and local governments have created a new policy category (Dubois, 2012), bringing together new functions, previous forms of public support for the arts, and closely related policy domains. However, the idea that culture is per se a legitimate field for public policy is still diversely appreciated from one context to another. If the term 'cultural policy' itself is common in the French political discourse since the 1960s, it has rarely been used at the European Union level, for instance.

What cultural policy includes is also variable in time, and from one country to another. The core commonly includes historical heritage, support of professional artists, and traditional cultural institutions such as libraries, museums, theaters, and concert halls. Quite a few other sectors or issues variably complement it, such as popular, regional, or migrant cultures, language, sports, media, leisure, after school and social activities, cultural education, or the amateur arts. What cultural policy includes or excludes in a specific national and historic context is not only a matter of bureaucratic organization, it partly reflects the boundaries of the legitimate domain for public intervention in a given society. It also reflects the definition of culture which prevails and which is promoted by public authorities.

Cultural public policy participates in the classification process that separates art from nonart, or at least valuable from less valuable art (DiMaggio, 1987). As Heikkinen shows in the case of Northern Europe, the administrative categories used to classify art and artists vary from one country to another according to the interaction between the state and the artistic field specific to each national context (Heikkinen, 2005). The extent to which public authorities determine these classifications or simply follow those defined by the art market is a good indicator of a cultural policy regime, since it reveals the relative weight and symbolic authority of the state and of the market in cultural affairs. With strong national differences, European cultural states gained power in this respect from the 1960s to the 1980s, but the market has become dominant again in the following period.

Beyond classification in art, the definition of culture in cultural policy plays a more general symbolic role in society. The core of the cultural policy domain basically corresponds to the legitimate definition of culture, that is to say the culture of the elite (Bourdieu, 1984). The extension of this domain then consists in the inclusion of the culture of other social groups, which raises the issue of their symbolic status. We can think here of three main cases. In Western Europe, as opposed to the Eastern part of the continent, the culture of the working class has generally been initially excluded from the field of state cultural policy, but promoted by social movements and political organizations, such as the Communist party. Declining traditional aspects from working class cultures have later been taken into consideration, mainly in the perspective of preserving their heritage (Dubois, 2011a). Regional and local subcultures are a second example. They have long been promoted in federalist and highly decentralized countries, where they represent the cultural counterparts of political claims toward the national level. Their partial recognition participates in the highly debated symbolic redefinition of the nation in unitary nation states. The third and currently prominent example concerns the cultures of immigrants. The inclusion of these cultures in the field of public policy serves as a symbolic tool for the inclusion of immigrants in society. This, again, is highly debated in many countries, including Germany, where the *multikulti* orientation of *soziokultur* programs, initially promoted by activists at the local level, has been included in official national policies from the 1980s onward. It is also debated in France, where the reticence emanating from the tradition of a unified Republican political and cultural model limits the development of such programs (Kiwan, 2007).

An Increasingly Uncertain Domain

Although the borders of cultural policies have not been strictly defined in the past, they are probably even less so today. The specification of culture as a policy domain currently faces three principal challenges. The first one is not entirely new and concerns the redefinition of culture. As long as a restrictive and elite notion of culture prevails, the outlines of cultural policy remain quite easy to draw. They broadly correspond to the field of art and heritage, to which are added educational and social aspects. But the invention of new artistic genres or disciplines (such as video art or street art, for instance) and the questioning of the notion of art by postmodern artists resulted in the constant redefinition of the borders of art itself. In addition and sometimes in relation to these redefinitions, the narrow

perimeter of culture as legitimate art and the elitist conception it entails has been contested at least since the end of the 1960s. The spreading of a more comprehensive and anthropological notion of culture and the claims for cultural recognition by minority groups have participated in making the definition of culture (and then of cultural policy) problematic from a social and political point of view. A second challenge results from the changes in the access to culture. As long as this access was mainly provided by institutions, such as theaters, or by media that could be directly regulated by public agencies, such as radio and television, the territory of cultural policy could correspond to the addition of these institutions and media. The advent of the Internet, which now plays a major role in the diffusion of cultural goods, and is far more difficult to regulate, leads to the fact that a substantial part of culture now is out of reach from traditional cultural policies. Finally, new ways of considering the borders of the cultural field have challenged the definition of cultural policy. Since the 1990s, the idea of creative industries and of a creative economy has emerged in the Anglophone world and has experienced a great success in Europe, starting with the United Kingdom (Hughson and Inglis, 2001). The notion of creative industries overlaps with the notion of cultural industries, although it is definitely a broader one. It adds design, advertising, fashion, software, computer games, etc., to the more traditional sectors of publishing, television, film, or performing arts. The notion of creativity tends to replace the notion of arts and culture in policy discourses, at the EU level and to a lesser extent in national and local governments. This evolution tends to replace cultural policy as a specific domain with a policy promoting the new model of a creative economy, in which culture is included, but diluted.

Specifying Cultural Policy Rationales

The Original Trilogy: Heritage, Artistic Creation, and Access

Cultural policy regimes in Western Europe have been built on three principal shared policy rationales. The first one pertains to the preservation of artistic and historic heritage. This is basically acknowledged as a legitimate state function, even in noninterventionist liberal contexts. In most European countries, heritage is generally regarded as national heritage, which is closely related to the national history and identity. Heritage protection corresponds to the model of a regulatory state, which establishes legal rules in order to prevent the loss of what is defined as public goods. It is an example of the previously mentioned classificatory role of cultural policy. In addition to legal protection, costly heritage maintenance requires funding that private individuals or organizations generally cannot afford by themselves. This financial issue has justified public expenses for culture over the two last centuries. In countries such as Italy, for instance, private foundations have nevertheless traditionally played a key role in the preservation of heritage. This is now a general trend in European heritage policies which questions the oldest legitimate basis for the intervention of public authorities in cultural affairs.

A second basic rationale for contemporary public policy is support for artistic creation. Even though there are similarities and continuity between the two, this support differs from the ancient model of state patronage. The manipulation of the arts as symbols of political power is no longer the major aspect of commissioning and subsidizing artists. This support is primarily based on the idea according to which the market is not sufficient to promote artistic creation and innovation. In contemporary visual arts or in literature, we could say that public support of creation is currently ancillary to the art or the publishing markets. But in contemporary art (i.e., experimental music) or to a lesser extent in theater, private demand and consumption could not suffice to fund creation. Cultural policy then aims to provide the conditions for this creation to exist. It can more ambitiously aim at preserving part of the field of artistic creation from the necessity of profitability, which in other words means enabling the implementation of the specific 'rules of art' in lieu of letting the economy rule the art (Bourdieu, 1995). This ambitious rationale for cultural policy raises, however, the issue of the extent to which public bodies should make artistic choices and, by doing so, should promote certain aesthetic and moral values above others. In order to avoid the specter of 'official art,' democratic cultural policy regimes have created forms of delegation of artistic judgment, through experts, commissions, and cultural organizations comanaged by officials and artists. This delegation of judgment is a major distinctive feature of these regimes.

The relationship between democracy and culture is also central to the third main rationale for contemporary European cultural policies, that is, equal access to culture (Vestheim, 2012). The idea of the necessity for the people to access education, arts, and culture so as to gain individual autonomy and to become able to exert citizenship dates back to the Age of Enlightenment. It was translated into a policy goal and has become a guiding principle for cultural policy in the postwar period. We could say that it was the cultural component of the making of European welfare states, which added cultural entitlements to social rights to achieve a more equal and inclusive society. The notions of cultural democratization and equal access to the arts were all the more successful insofar as they could easily bring politicians and cultural actors to a common ground. They were consequently central in the legitimization of modern public cultural policies differentiated from past forms of public support to the arts, with some exceptions such as Greece, where cultural policy focused on heritage and disregarded the democracy dimensions (Zorba, 2009).

These general notions, however, comprehend various if not contradictory orientations. The civic perspective of autonomous individuals able to participate in community life thanks to culture meets educational policies. The humanist and somewhat paternalistic perspective of pushing the lower classes toward greater participation in highbrow culture pertains to a paradoxical form of elitism grounded on the idea that only the upper educated echelon knows what is real culture and what is good for the people. Conversely to this top-down approach, the claim for cultural democracy consists in the desacralization of elite culture and in the achievement of cultural diversity forms that envision democratization from a bottom-up perspective (Bonet and Négrier, 2011; de Jong, 1998; Saukkonen and Pyykkönen, 2008). The balance and combinations between these orientations varies in time and space. A legitimist and elitist approach to the dissemination of

culture in society prevailed until the late 1960s. From then on, this model has been contested, but it has not been entirely replaced by an alternative model. Current cultural democratization policies and discourses actually articulate different strata that may be logically incompatible but are combined in practice. Generally speaking, the legitimist approach remains important, as it reflects the predominance of elite cultural institutions in European cultural policies, but is nuanced by programs that are conceived in a more participatory way, or more oriented toward the promotion of cultural diversity.

From the Diversification of Goals to the Noncultural Rationales of Cultural Policy

This original trilogy sets the basis for policy objectives proper to cultural policy and that aim at specifically cultural purposes. This should not, though, be taken for granted. First, these core rationales have always been combined with other policy goals (Gray, 2009). The search for national prestige and influence at the international level is an example of the link between cultural policy and diplomacy. Among other things, most European governments have established a network of cultural centers abroad, such as the British Council or the Goethe Institut. The making and promotion of collective identity through culture lasts long after the early period of nation-building, as we have seen after the German reunification in 1990, or in the renewal of political uses of history in Spain, France, or even at the EU level during the past two decades. In addition, public cultural policies have a wide range of specifically social policy goals, from urban regeneration to the supervision of idle youth in deprived neighborhoods, from the empowerment of disabled persons to the regulation of ethnic and cultural diversity and to the alleviation of social exclusion (Belfiore, 2002). Last but not least, culture is considered as a means to achieve economic goals. Beside cultural concerns, festivals, museums, or historic buildings are envisioned and publicly funded qua assets in terms of touristic attractiveness. Cultural organizations and training in cultural professions are also supported because the culture is a dynamic sector in terms of employment.

Second, each of the three main objectives specific to cultural policies is challenged in various ways. The protection of heritage is nowadays less than ever a monopoly of the public sector. In addition, the narrow definition of heritage as the legacy of celebrated works of art and buildings from the past has progressively been replaced from the beginning of the 1980s by a much broader one, a potentially never-ending list, which includes industrial sites, folk traditions, or local dialects. This broadening accounts for the changes in the relationship of Western societies to their past. It also transforms the role of heritage policies, which, more than ever, 'invent' rather than only protect heritage. All of this questions the self-evidence of heritage preservation as a public policy function. In times of economic crisis, public support of artistic creation is questioned, especially when this creation hardly meets any demand from the audience. On the other hand, the role of public agencies is weakened by speculation in the art market and the increase in prices. Cultural democratization, that is, the prominent objective and legitimizing discourse in the development of contemporary cultural policies in Europe, has also lost an important part of its strength as a political and cultural guiding principle. The contestation of a unique and universal notion of culture, from countercultural movements in the late 1960s to the current promotion of cultural diversity, has weakened the belief in the legitimist cultural model on which cultural policies have been founded. The mixed results, if not the failure of several decades of public cultural policies in the reduction of social inequalities in cultural participation, have also shaken the founding myth of cultural democratization. Moreover, the huge development of ways to access culture, television, and subsequently the Internet, has increasingly competed with cultural policies, mainly focused on traditional institutions such as libraries or theaters.

Finally, the balance between strictly cultural objectives and noncultural goals has moved toward the latter orientations. Conversely to the specialization and specification processes that started in the postwar period and resulted in asserting cultural policy goals for the sake of culture, public intervention in cultural matters is at present increasingly assigned noncultural purposes, among which economic ones prevail. Public cultural expenses are gradually viewed in terms of investments from which economic impacts are expected. This is the case even in countries with a strong cultural policy and welfare-state tradition, such as France or Northern European countries (Mangset et al., 2008). The successful notion of creativity has become the keyword of current cultural policies. This makes the arts lose their specificity by linking them to entertainment and by viewing them as part of a fuzzy ensemble (the creative sector) defined on an economical rather than on a cultural or aesthetic basis. As a result, national cultural policies change from 'intrinsic' to 'instrumental' (Orr, 2008; Vestheim, 1994), jettisoning part of their cultural rationales in favor of economical ones, and contributing to a new form of 'cultural capitalism' (McGuigan, 2004). This is ironic when we remember that contemporary European cultural policies have partly been conceived to protect arts and culture from the rules of the economic market. At the EU level, instead of developing proper cultural policies, the addition of 'cultural dimensions' to economic and social policies ('mainstreaming') frames and defines the criteria for cultural projects according to the rationales of economic and social programs (Gordon, 2010). This illustrates the cultural and political implications of the organizational structure of cultural policy.

Organizing Cultural Policy in Europe: Structures and Governance

Social, Political, and Institutional Settings

Beyond its technical aspects, the organizational structure of cultural policy raises the question of who is legitimate to define collective cultural orientations. The general public is usually not part of the cultural policy making processes, although there have been some recent attempts to include satisfaction of the citizenry in the policy evaluation criteria. The organized interaction between public officials and professional representatives of artistic sectors is far more important, and there are examples of comanagement, such as cinema in France, or in the Nordic countries. Public intellectuals can be influential, which is also an indicator of cultural policy as an issue in the public debate

(Ahearne and Bennett, 2007; Burns and van der Will, 2006; Dubois, 2011b).

The organizational structure of cultural policy also raises the question of how the sensitive and highly symbolic matters of arts and culture should be governed. In liberal democratic regimes such as Western European states, the mode of governance is required to exclude any possible resemblance to the authoritarian model of arts and culture controlled by the state apparatus. This is a strong political constraint for the institutions that elaborate, implement, and embody national cultural policies. Two models are generally considered to account for Western European cultural policy (Hillman-Chartrand and McCaughey, 1989). The 'architect state' relies on centralized ministries. In this case, a specific department for culture within the national government concentrates a predominant part of the financial, political, informational, and symbolic resources for cultural policy making. Such a department asserts the role of the state in culture and is the main center for institutional and symbolic power in this domain. This type of interventionist state involvement characterizes France, while most governments adopt a less directly activist role. The latter represents the ideal type of the 'patron state,' dispensing funds through quasi-independent institutions. The Arts Councils in Great Britain exemplifies a public body which operates at 'arm's length,' i.e., in relation with the national government but outside its departmental internal structure. This is supposed to protect from a possible excessive political leadership.

While this dual opposition helps to set the landscape of national cultural policy structures, it is, of course, not sufficient. The actual governance systems have never been as simple as that. They also have experienced important changes, such as the relative decline of the French ministry for culture, or the increasing control of the British government and its department for media, culture, and sports over the Arts Councils. In these two countries as elsewhere in Europe, governments increasingly use tax policies to promote private patronage, which refers to the third type, a 'facilitator state,' of which the United States provides an example. There is neither a strong national ministry for culture, nor a strong arts council in Germany or in Italy, whereas there are both in the Nordic countries (Mangset et al., 2008).

Whatever the case, the previously mentioned political constraint in democratic regimes results in the fact that there is no institutional system in contemporary Western Europe in which there is a sole decision and bureaucratic monopoly on culture. As a policy area, culture is shared between infranational, national, and supranational policy levels, with strong connections to the nonprofit private sector (Burns and Van der Will, 2003; Stevenson et al., 2010). This multilevel complexity characterizes all European countries, including those usually regarded as having a high level of state centralism such as France, where the national government is responsible for less than 50% of the total of the public budget for culture, and where all the politico-administrative levels (municipalities, intercommunity structures, departments, and regions) have jurisdiction over culture.

This is, however, far from avoiding jurisdiction conflicts and debates about which level is most legitimate and best suited to take care of cultural affairs (Dubois et al., 2012). While such conflicts and debates occur in any kind of political and institutional system, they are of particular importance in the countries where infranational and national governments compete for political legitimacy. This is clearly the case in Spain, where this competition is a factor in the apparent inconsistency of the multilevel governance of cultural policies, and where diversity results in the nonexistence of a formal Spanish model of cultural policy (Bonet and Négrier, 2010). Although Spain is an extreme case, all European countries experience to various extents comparable dynamics between center and periphery, and decentralization has generally been increasing for the last three decades.

Funding Culture

Cultural policies are often identified with public financial support for the arts and culture. This is obviously simplistic, since they use a wide range of policy tools, from market regulation to organizational reform, but underlines that funding is a key issue. To some extent, cultural policy regimes then consist in cultural funding modes. These regimes are difficult to compare. A rough approximation shows that the public expenditures for culture range from around 0.5 to 2.5% of the total of public budgets of Western European countries in 2010–11 (Council of Europe and ERICarts, 2012). This does not indicate much since all of them mix various financial aspects, which are almost impossible to assess in a single figure (Alexander, 2008). In addition to the most direct forms of funding, that is to say public budgets for state-controlled institutions, such as public libraries, and subsidies granted to individual artists or nonprofit arts organizations, there are many other indirect forms, such as social security benefits paid to artists, support to private companies in the cultural industries, or tax incentives for the patronage of the arts. This strongly varies from one country to another. A general tendency, however, is the predominance of support to art institutions. More precisely, European cultural policies have to a large extent consisted in creating new art institutions in order to extend the supply of culture. This plays a direct role in the current financial crisis of cultural policy whose budget is widely absorbed by the costs to run these institutions. This crisis is all the more acute as that stagnation, if not retrenchment, comes after decades of increase in the cultural public budgets.

This has changed the debate on who is to pay for culture. In the period of development of national cultural policies, which coincided with a general growth of public budgets, national states and local authorities have increased their cultural expenditures, sometimes considerably. Not only because of financial difficulties but also because of neoliberal policies, governments have looked for new financial sources, including the national lottery in the United Kingdom in the mid-1990s. The general trend toward decentralization, which had previously accompanied the global development of cultural policies, is now often related to the withdrawal of the national governments. As a result, cultural policies increasingly appeal to the private sector. This brings European cultural policy regimes in some ways closer to the US model in which cultural organizations have to raise money. Tax incentives for the private patronage of the arts are now a common feature of public policies for culture in Europe, where they sometimes are more favorable than they are in the

United States. This does not amount to the end of European cultural policy regimes, but is part of the redefinition process they are experiencing.

See also: Cultural Policies in East Asia; Culture Policy Regimes in China; Markets: Artistic and Cultural; Nation-State as Symbolic Construct; Organizations and Culture; Social Inequality in Cultural Consumption Patterns; Symbolic Boundaries.

Bibliography

Ahearne, J., 2009. Cultural policy explicit and implicit: a distinction and some uses. International Journal of Cultural Policy 15 (2), 141–153.

Ahearne, J., Bennett, O. (Eds.), 2007. Intellectuals and Cultural Policy. Routledge, London.

Alexander, V., 2008. Cultural organizations and the state: art and state support in contemporary Britain. Sociology Compass 2 (5), 1416–1430.

Belfiore, E., 2002. Art as a means of alleviating social exclusion: does it really work? A critique of instrumental cultural policies and social impact studies in the UK. International Journal of Cultural Policy 8 (1), 91–106.

Bonet, L., Négrier, E., 2010. Cultural policy in Spain: processes and dialectics. Cultural Trends 19 (1–2), 41–52.

Bonet, L., Négrier, E., 2011. The end(s) of national cultures? Cultural policy in the face of diversity. International Journal of Cultural Policy 17 (5), 574–589.

Bourdieu, P., 1984. Distinction. Harvard University Press, Cambridge.

Bourdieu, P., 1995. The Rules of Art. Stanford University Press, Stanford.

Burns, R., Van der Will, W., 2003. German cultural policy: an overview. International Journal of Cultural Policy 9 (2), 133–152.

Burns, R., van der Will, W., 2006. Intellectuals as cultural agenda-setters in the federal republic? International Journal of Cultural Policy 12 (3), 291–322.

Council of Europe and ERICarts, 2012. Compendium of Cultural Policies and Trends in Europe. www.culturalpolicies.net (accessed 02.11.13.).

Castelnuovo, E., Ginzburg, C., 2009. Symbolic domination and artistic geography in Italian art history. Art In Translation 1 (1), 5–48.

de Jong, J., 1998. Cultural diversity and cultural policy in the Netherlands. International Journal of Cultural Policy 4 (2), 357–387.

DiMaggio, P., 1987. Classification in art. American Sociological Review 52, 440–455.

Dubois, V., 2011a. Lowbrow culture and French cultural policy. The socio-political logics of a changing and paradoxical relationship. International Journal of Cultural Policy 17 (4), 394–404.

Dubois, V., 2011b. Cultural capital theory vs. cultural policy beliefs: how Pierre Bourdieu could have become a cultural policy advisor and why he did not. Poetics 39 (6), 491–506.

Dubois, V., 2012. La Politique Culturelle. Genèse D'une Catégorie D'Intervention Publique, second ed. Belin, Paris.

Dubois, V., Bastien, C., Freyermuth, A., Matz, K., 2012. Le Politique, l'Artiste et le Gestionnaire. (Re)configurations Locales et (Dé)politisation de la Culture. Le Croquant, Bellecombe-en-Bauges.

Gordon, C., 2010. Great expectations. The European Union and cultural policy: fact or fiction? International Journal of Cultural Policy 16 (2), 101–120.

Gray, C., 2009. Managing cultural policy: pitfalls and prospects. Public Administration 87 (3), 574–585.

Heikkinen, M., 2005. Administrative definitions of artists in the Nordic model of state support for artists. International Journal of Cultural Policy 11 (3), 325–340.

Hillman-Chartrand, H., McCaughey, C., 1989. The arm's length principle and the arts: an international perspective. In: Cummings, M.C., Schuster, M. (Eds.), Who's to Pay for the Arts? The International Search for Models of Support. American Council for the Arts, New York, pp. 43–80.

Hughson, J., Inglis, D., 2001. 'Creative industries' and the arts in Britain: towards a third way in cultural policy? International Journal of Cultural Policy 7 (3), 457–478.

Kiwan, N., 2007. When the cultural and the social meet: a critical perspective on socially-embedded cultural policy in France. International Journal of Cultural Policy 13 (2), 153–167.

Mangset, P., Kangas, A., Skot-Hansen, D., Vestheim, G., 2008. Nordic cultural policy. International Journal of Cultural Policy 14 (1), 1–5.

McGuigan, J., 2004. Rethinking Cultural Policy. Open University Press, Maidenhead.

Orr, J., 2008. Instrumental or intrinsic? Cultural policy in Scotland since devolution. Cultural Trends 17 (4), 309–316.

Saukkonen, P., Pyykkönen, M., 2008. Cultural policy and cultural diversity in Finland. International Journal of Cultural Policy 14 (1), 49–63.

Sokka, S., Kangas, A., 2007. Intellectuals, nationalism, and the arts. International Journal of Cultural Policy 13 (2), 185–202.

Stevenson, D., McKay, K., Rowe, D., 2010. Tracing British cultural policy domains: contexts, collaborations and constituencies. International Journal of Cultural Policy 16 (2), 159–172.

Vestheim, G., 1994. Instrumental cultural policy in Scandinavian countries. European Journal of Cultural Policy 1 (1), 57–71.

Vestheim, G., 2012. Cultural Policy and Democracy. International Journal of Cultural Policy 18 (5), 493–504.

Zorba, M., 2009. Conceptualizing Greek cultural policy: the non-democratization of public culture. International Journal of Cultural Policy 15 (3), 245–259.

Cultural Production in Networks

Katherine Giuffre, Colorado College, Colorado Springs, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cultural production occurs in a field which is shaped by both cooperation and by conflict and competition. The field is a network of relations among producers and the structure of network ties affects cultural production at multiple levels – from individual to group, industry, and place. Important network characteristics are density, diversity, structural holes, and small world structures.

Relational Thinking vs 'The Lone Genius'

Beginning in the Renaissance, there was a shift in conceptualizing the way in which cultural production occurs. While artists had been considered craftsmen during the early Renaissance – members of guilds, trained as apprentices, producing items to the specifications of detailed contracts, working on a wide variety of objects in a collaborative workshop setting – the later Renaissance saw the emergence in Western culture of the idea of artist as a 'lone genius,' pursuing a singular, individual vision, and working in isolation from others. This conception of the lone genius received a further boost during the Romantic Era when it became the dominant cultural understanding for the way in which art was produced. However, current social scientific research into cultural production shows that art (including both visual and performance art, creative writing, music, fine art, folk art, and popular culture) is produced in social settings and that artists and other cultural producers are embedded in social networks which affect cultural production in a variety of ways.

Social network analysis is a useful method for implementing the relational thinking that is opposed to the romantic 'lone genius' notion of cultural production. The idea that social networks play an important role in individual lives and in social processes has been part of the sociological tradition since at least the mid-1800s, but only in the mid- to late-twentieth century did the metaphor of the 'social network' become transformed into actual social scientific methodology capable of sustaining relational analysis. Harrison White and others developed not only the tools by which network analysis could proceed, but also the theoretical underpinnings for an analysis centered on the social relations among the actors in a system as its primary focus. Pierre Bourdieu's theories (especially of the 'field' (*champ*)) provide key conceptions of relational thinking with regard to cultural production.

Bourdieu's Field as a Space of Relations

Bourdieu argues that cultural production happens on two levels – on the level of the object, performance, etc. and on the level of the understanding of the object as art. "Given that works of art exist as symbolic objects only if they are known and recognized, i.e., socially instituted as works of art and received by spectators capable of knowing and recognizing them as such, the sociology of art and literature has to take as its object not only the material production but also the symbolic production of the work, i.e., the production of the value of the work or, which amounts to the same thing, the belief in the value of the work" (Bourdieu, 1983: p. 318).

Neither the production of the work nor the production of the understanding of the work happens in a vacuum – they happen inside the 'field of cultural production' which is a space of relations and of struggle. The field is "a structured space with its own laws of functioning and its own relations of force … Its structure, at any given moment, is determined by the relations between the positions agents occupy in the field. A field is a dynamic concept in that a change in agents' positions necessarily entails a change in the field's structure" (1993: p. 6).

The field is defined by the objective relations among the actors (which can be persons, groups, institutions, and so on). Actors' positions are determined relative to one another. Moreover, the field is hierarchical with regard to power and the actors are constantly struggling to improve their positions vis-à-vis each other so as to legitimize certain positions (their own, principally) and delegitimize the positions of their rivals. This means they endeavor not only to gain prestige and power for themselves, but also, more importantly, to be able to control the power to define legitimacy in the field.

The simultaneous struggles of all of the actors in the field to achieve positions of legitimacy (dominance) create a dynamic system where any individual position is always dependent on the distribution of all other across the entire field. Successful actors in the field of cultural production accrue symbolic (as opposed to economic) capital and position themselves at the autonomous (as opposed to heteronomous) pole of the field. This creates an 'economic world reversed' where those who amass economic capital lose status relative to those who amass symbolic and cultural capital.

Art Worlds

The concept of the 'art world' also emphasizes the network of the relations among actors. According to Becker, "Art worlds consist of all the people whose activities are necessary to the production of … art. …The same people often cooperate repeatedly, even routinely, in similar ways to produce similar works, so that we can think of an art world as an established network of cooperative links among participants" (1982: pp. 34–35).

As with Bourdieu, Becker recognizes both the material and the symbolic production of art works. The inner circles of different parts of the art world share understandings about meanings and conventions which separate and distinguish them from more peripheral members.

Becker emphasizes cooperation, but there is also conflict and competition, especially around issues of legitimacy and of change. Cultural producers compete for notice and legitimacy. Moreover, those who are most likely to innovate are those whose ties to the established art world are either weakened or never existed in the first place. This fits with more explicitly network analytical conceptions of innovation and creativity having to do with network structures and characteristics, especially ideas about density, structural holes, and small world structures.

Network Structure and Creativity

Network density is a measure of the proportion of possible ties which are actualized among the members of a network. Dense social networks, especially coupled with strong boundaries segregating the group from others, can enforce communal norms so that social pressures for conformity can inhibit creativity, which necessarily contains an element of deviance. Small dense networks may develop 'groupthink' where conformity of ideas is highly valued and normatively enforced. This inhibits creativity within the group.

Social scientists (e.g., Jehn, 1995) have found that conflicts among the members of groups may increase performance of nonroutine tasks. Nemeth and Nemeth-Brown (2003), Milliken et al. (2003), and Strasser and Birchmeier (2003) all similarly argue for the value of dissent and diversity as stimulators for creative thought in group settings. Other research (Ng, 2001) shows that those individuals who conform to group norms tend not to engage in the often divisive deviance necessary for creativity. Moreover, when McLeod et al. (1996) studied groups composed of all Anglo-Americans and groups composed of mixed ethnicities who were asked to brainstorm solutions to a provided problem, they found that groups composed of ethnically diverse members produced more solutions and solutions of higher quality than did the racially homogeneous groups. But this is only true up to a point. Too much diversity can eventually have a negative effect on group functioning when groups are unable to cohere sufficiently to collaborate at all.

Fuchs (2009) makes a case for analyzing cultural production networks with regard to actors and processes in the core of the network vs those in the periphery. Fuchs argues that the core of the network becomes closed and ideas there become the network's 'common sense.' He writes that, "Common sense is slow to learn, resistant to change, and unwilling or unable to make any genuine discoveries" (2009: p. 358). This is in contrast to the actors in the periphery, which are more loosely coupled – that is, less dense – and therefore "accommodate more controversy and contingency; their mood is more playful and their mentality more open to change, alternative interpretations, and innovations" (2009: p. 360).

Density – as a mechanism for normative conformity within a network – is not the only important structural factor in fostering creativity. Small subcultures may develop strong norms, but people who are positioned to bridge the gaps among these groups enjoy the creative benefits of being able to synthesize the information into new forms. This is Burt's (2004) argument about 'structural holes.' He argues that social forces such as homophily and interpersonal balance work to encourage individuals to build network ties with like-minded others. This leads to the formation of groups of tightly knit actors who share similar dispositions, etc. These groups are relatively separated from other groups that have different sets of dispositions, norms, and knowledge. The absence of ties among these groups form structural holes – that is, gaps in the social structure where network ties could exist, but do not.

Burt's main claim is that "people who stand near the holes in a social structure are at higher risk of having good ideas" (2004: p. 349). Because there is homogeneity within cliques, deviants are more likely to exist on the fringes of the cliques and are more likely to have ties with others who are outside the clique. The important factor is the dissimilarity of the others to whom they are connected. Burt argues that "People whose networks span structural holes have early access to diverse, often contradictory, information and interpretations, which gives them a competitive advantage in seeing good ideas" (2004: p. 356). He notes, "This is not creativity born of genius; it is creativity as an import-export business. An idea mundane in one group can be a valuable insight in another" (2004: p. 388). Burt shows that people who act as brokers across structural holes are more likely to have innovative ideas and are more likely to have their ideas recognized as valuable and to be rewarded for those ideas.

The concept of structural holes fits nicely with the idea of 'small world structures.' Small world structures (Watts and Strogatz, 1998) are a further development of Stanley Milgram's small world studies. Milgram (1967) investigated the phenomenon of two people meeting by chance and discovering that they have an acquaintance in common. Milgram found that the chains of interpersonal connections linking randomly chosen individuals in the United States had a much smaller number of steps (about six) than originally hypothesized. This finding allowed Watts and Strogatz to develop a model of social structure where most ties are dense and local, but a few long bridge ties between far-flung people dramatically shortened the chain length of interpersonal ties between everyone in the system. This is the small world structure where small groups of tightly knit actors are connected with each other through a few ties which bridge the gaps between them. With regard to the production of Broadway musicals, Uzzi and Spiro argue that "a small world network governs behavior by shaping the level of connectivity and cohesion among actors embedded in the system. … The more a network exhibits characteristics of a small world, the more connected actors are to each other and connected by persons who know each other well through past collaborations or through having had past collaborations with common third parties. These conditions enable the creative material in separate clusters to circulate to other clusters as well as to gain the kind of credibility that unfamiliar material needs to be regarded as valuable in new contexts, thereby increasing the prospect that the novel material from one cluster can be productively used by other members of other clusters" (2005: p. 449). In small world structures, there are dense local groupings, but also bridge ties

that connect the groups efficiently. Uzzi and Spiro find that the small world structure is beneficial to Broadway musicals (both in terms of financial and artistic success) but only up to a point. Once the entire network becomes too connected, the distinctive individuality of the various groups diminishes and problems of 'groupthink' and lack of diversity set in. Collaboration has to be structured so as to avoid conformity.

Farrell (2001) defines 'collaborative circles' as groups of cultural producers who develop a shared vision of their work. His argument is that the attribute of personal creativity is socially produced. Cultural producers have creative ideas with other people that they would never have on their own because of the social support and critical feedback that producers get from others who share their vision. Farrell argues there are specific ways that successful collaborative groups are structured which make them fertile ground for the flowering of creativity and there are specific roles that group members need to fill to make that happen. Successful collaborative circles are formed in opposition to existing power structures and members are initially drawn from people who are marginal to the mainstream art world. It is this initial opposition that draws the group together. Members often find supportive others within the circle, but also find contentious and combative debates among circle members which function to sharpen the group's vision. They also tend to develop increasingly radical ideas as the members goad each other on and the circle develops into something similar to a deviant youth subculture.

Collaborative circles are often most useful to cultural producers in the earlier stages of their careers when the artists are formulating their distinctive vision. As the artists' careers progress, however, collaborative circles often disband.

Careers

Artists' careers are also built around social networks. Artists (who often work as freelancers connected to specific projects) must make connections with other art world members (such as gallery owners, book publishers, film producers, etc.) in order to make a living from their work. These networks of connections are some of the relations that define the field of cultural production. Artists' relations with others in the field define their place in the hierarchy and their position relative to the autonomous or heteronomous poles of the field. As cultural producers make and break connections with others in the field, the field is continuously re-shaped. An individual's career becomes a trajectory through the field – that is, a series of positions relative to others in the field and demarcated by distance from the poles of the field. Cultural producers engage in strategies (such as succession or subversion) in the field in an attempt to construct career trajectories that are beneficial to themselves. This involves strategically making (and breaking) ties with other actors in the network as that network is continuously changing form over time. Despite this inherent flux in the shape of the field, actors do manage to carve out career paths. Although the specific actors who occupy the positions change, the types of positions available in the field (e.g., dominant or dominated) has some stability.

Faulkner (1983) documents how Hollywood studio composers build careers out of networks by moving from project to project. A composer of music for Hollywood films "competes in a marketplace for projects and in a network for access to owners and controllers of those projects. It was immediately apparent that access breeds success, so there is continual pressure toward monopolization of work" (1983: p. 22). Faulkner found that those composers who could achieve centrality in the network of relations in the industry could maintain a successful career. This happened by strategically making and breaking personal ties with others in the field. This is true for other types of artists as well. Fine art photographers, for example, succeed in the art world by building strategic social networks with other artists and galleries (Giuffre, 1999). Especially useful in this case were broad-ranging networks of weak ties which could connect individual photographers with a variety of others. This ensured them career continuity in a volatile world of many short-lived galleries.

Looking at writers and their connections to literary magazines, de Nooy (2002) pushes this model much further. The prestige of the writers and of the magazines is dynamically linked, each influencing the other and each pursuing strategies in the field through the relations that they make, break, or avoid with other actors. De Nooy found that strategies of 'status preservation' (as opposed to 'status enhancement') were the chief operations in the field. He found that, "the flow of authors toward prestigious magazines seems to be restricted by selective admission, which is probably exerted by the editors of these magazines" (2002: p. 165). This fits with research about the importance of gatekeepers who work to limit access to higher levels of the art world by those actors with lower prestige. Overall, de Nooy's work offers support for the "dynamic and relational character of field theory" (2002: p. 164).

Looking at writers' networks in the early modern period in Korea, Jun and Park (2009) found that writers with high network centrality were more likely to be productive and to be included in the canon. In their research, gender intersects with network form. Jun and Park found that female writers had significantly smaller diameters of their networks – that is, "they belonged to less various groups of writers than their male counterparts" (2009: p. 27) – and were consequently disadvantaged by limited network resources. Becker notes, "You need ability, then, for free-lance success, but it is not enough. Successful free-lancers also need a network of connections, so that a large number of people who might need their services have them in mind, and in their telephone book, to be called when the occasion arises. Reputation helps" (1982: p. 86).

Reputation is built up through networks. Networks act as gatekeepers to desirable parts of the art world. Ridgeway (1989) argues that groups of artists – 'movement circles' – play a crucial role in both supporting and allocating access for young artists. For example, young artists need introductions to galleries that more established artists can provide. This is the role of the movement circle in helping artists developing reputations with other artists – the underground reputation that was critical to getting their first gallery representation. "These young artists perceived the necessity of extending contacts in order to achieve success to the critical centers of the art community and to the market, knowing such acceptance was a crucial link in the process of their own acceptance. The importance of developing an 'underground' reputation was clear. 'Artists initiate styles. (Dealers) bring to the public

guys who are already there. When the Minimalists, say Andre, Judd, LeWitt, you know – put Ryman there – they had been making art for a long time. By the time Virginia (Dwan Gallery) picked them up, they all had underground reputations. All of them'" (Ridgeway, 1989: p. 212).

More established artists acted as gatekeepers by giving younger artists introductions to collectors and galleries. "Few artists ever find their way into a major gallery without such an introduction. The majority of these artists attained their first gallery recommendation from other artists" (Ridgeway, 1989: p. 217). This is especially crucial for art world actors who are attempting a strategy of succession in order to gain a dominant position in the field of cultural production. Moreover, this is true not only in the fine arts, but in culture industries in general, where products, from painting to pop music, must pass through gateways.

Because there is an oversupply of cultural products relative to demand, distributors filter products which make it to the market. This may be a stepwise process where a cultural producer must pass through several successive gatekeepers, risking rejection at each step before eventually reaching the market. Each of these steps involves successfully making network connections with the different gatekeepers and is often facilitated by other network members.

Accominotti (2009) finds that those painters who were part of artistic movements and were embedded in networks of like-minded others had creative peaks in their careers during the time that they were engaged in the movement. Being embedded in an artistic movement network became an important part of a subversive strategy in the field of fine art painting. Accominotti writes, "Over time, organizing into movements increasingly became a strategy for artists eager to supplant older, more established ones. This strategy spread after the success of the Impressionists, once it had become clear that at least some contemporaries regarded movements as the place where significant contemporary art was being produced. As an organizational form, movements thus became appealing to young, ambitious artists" (2009: p. 286). In this way, circles of artists tied to each other in order to establish new artistic schools have become a regular feature of cultural production and play an important role in shaping artists' careers.

Markets and Industries

Hesmondhalgh argues that "Bourdieu misses the importance of the rise of cultural industries [in the 20th century] for understanding the changing social relations of cultural producers" (2006: p. 220). Specifically, the distinction between autonomous and heteronomous production is overly polarized and "restricted production has become introduced *into* the field of mass production" (2006: p. 222), especially as the cultural industries work to address demand uncertainty in order to try to control risk in production. This dovetails with White's arguments (2002) that markets are enduring networks linking producers with consumers in relatively stable patterns across time. These networks attempt to limit uncertainty by monitoring other actors similar to themselves in the market. "So commitment and uncertainty," White writes, "are the twin themes in production markets" (2002: p. 7). Market transactions are repetitive, relatively stable ties among actors in the network so that "the market [is a] joint social construction" (2002: p. 10) from which also arises the socially constructed judgment of 'quality' and of 'reputation' – both important features in the art market.

The institutional logics perspective takes into account "the duality of the material and symbolic aspects of institutions" (Thornton et al., 2012: p. 18) addressing both the structure and the culture of institutions. It allows us to ask *why* actors are engaged in particular network structures. For example, Thornton and Ocasio (1999) argue that the need for new sources of capital (among other pressures) among academic book publishers during the 1970s resulted in a shift of institutional logics from an editorial focus – "with editorial reputation and author-editor networks as key foci of attention in implementing the mission and strategy of the firm" (1999: p. 816) – to a market focus, dominated by "the logic of Wall Street investment bankers and the increasing concern with profitability and market orientation common to other U.S. industries" (1999: p. 816). This affects cultural products.

Sargut (2006) looks at the simultaneous effects of market structure and environmental (e.g., demand) uncertainty on the homogeneity of movie content between 1931 and 1950. Cultural innovation incurs risk for cultural production firms where demand uncertainty plays such an important role in decision making. Sargut found that when market concentration was low, increased environmental uncertainty lead to decreased product homogeneity as firms produced a diversity of products to try out on the public. Sargut argues that increased environmental uncertainty lead to increases in the number of network ties among cultural producers and firms taking cues about product content through their network ties. This in turn had implications for product homogeneity as the uncertainty in the environment is mediated by the linkages among the actors in the network.

Creative Places

Cultural production flourishes in some places and times more than in others. Network structures in those places may facilitate this. Large-scale structures exhibit many of the same network features that characterize individual creativity. While individual creativity is fostered by ties across structural holes which translate into access to diverse others, creative places tend to be those where diverse others are brought into contact with each other on a regular basis and where bridge ties linking small worlds are facilitated. An example of this would be Silicon Valley, where wide-ranging weak ties bridging the gaps between various companies and industries were fostered (Castilla et al., 2000). Other more explicitly cultural examples would be Renaissance Florence or Paris in the 1920s.

Physical environments create an environment for the flourishing of creativity when they facilitate tie formation, provide public spaces where informal meetings can occur, encompass a variety of uses and populations in small areas, provide ease of access at all times of day, and give people informal reasons to go there (such as the existence of cafes, etc.) The design of physical spaces can encourage the building of small world structures for the individuals, groups, and organizations that have access to that space.

In contrast, however, with individual creativity, which may be hampered by high-group density inhibiting deviant thought, the creativity of places is instead enhanced by high density, which allows individual cultural producers easier access to both to a critical mass of like-minded others (who may form artistic movement or collaborative circles) and simultaneously to divergent or even oppositional other groups. Access to these different groups reduces the chance of 'groupthink' and allows for the small world structures of links across structural holes to diverse ideas to form. Fresh information from outside that can generate innovative ideas is then available to the members of the groups.

Cultural production is deeply embedded in social networks at the individual, group, industry, and place level. These networks affect production of ideas as well as the production of cultural products and the ability of those products to reach an audience. These networks are not only sites of cooperation among cultural producers, but are also sites of conflict and competition which shapes the way in which culture is made and experienced.

See also: Artists, Competition and Markets; Artists, Competition and Markets; Avant-Garde Art and Artists; Bourdieu, Pierre (1930–2002); Cultural Mediators and Gatekeepers; Culture and Actor Network Theory; Culture and Institutional Logics; Culture and Institutional Logics; Culture and Networks; Culture, Production of: Prospects for the Twenty-First Century; Markets: Artistic and Cultural; Network Analysis; Networks and Cultural Consumption; Networks and Meaning.

Bibliography

Accominotti, Fabien, 2009. Creativity from interaction: artistic movements and the creativity careers of modern painters. Poetics 37, 267–294.

Becker, Howard S., 1982. Art Worlds. University of California Press, Berkeley, CA.

Bourdieu, Pierre, 1983. The field of cultural production, or the economic world reversed. Poetics 12, 311–356.

Bourdieu, Pierre, 1985. The market of symbolic goods. Poetics 14, 13–44.

Bourdieu, Pierre, 1993. The Field of Cultural Production. Columbia University Press, New York.

Burt, Ronald, 2004. Structural holes and good ideas. American Journal of Sociology 110, 349–399.

Castilla, Emilio J., Hwang, Hokyu, Granovetter, Ellen, Granovetter, Mark, 2000. Social networks in Silicon Valley. In: Lee, Chong-Moon, Miller, William F., Hancock, Marguerite Gong, Rowen, Henry S. (Eds.), The Silicon Valley Edge. Stanford University Press, Stanford, CA, pp. 218–247.

de Nooy, Wouter, 2002. The dynamics of artistic prestige. Poetics 30, 147–167.

Farrell, Michael P., 2001. Collaborative Circles: Friendship Dynamics and Creative Work. University of Chicago Press, Chicago.

Faulkner, Robert, 1983. Music on Demand: Composers and Careers in the Hollywood Film Industry. Transaction Books, New Brunswick, NJ.

Fuchs, Stephan, 2009. The behavior of cultural networks. Soziale Systeme 15, 345–366.

Giuffre, Katherine A., 1999. Sandpiles of opportunity: success in the art world. Social Forces 77, 815–832.

Hesmondhalgh, David, 2002. The Culture Industries. Sage Publications, London.

Hesmondhalgh, David, 2006. Bourdieu, the media and cultural production. Media, Culture & Society 28, 211–231.

Jehn, Karen A., 1995. A multimethod examination of the benefits and detriments of intragroup conflict. Administrative Science Quarterly 40, 256–282.

Jun, Heejin, Park, Kwang-hyung, 2009. The cooperative networks of writers and the making of the "modern" literary canon in early modern colonial Korea. Paper Presented at the Annual Meeting of the American Sociological Association.

McLeod, Poppy Lauretta, Lobel, Sharon Alisa, Cox Jr., Taylor H., 1996. Ethnic diversity and creativity in small groups. Small Group Research 27, 248–264.

Milgram, Stanley, 1967. The small-world problem. Psychology Today 1, 62–67.

Milliken, Frances J., Bartel, Caroline A., Kurtzberg, Terri R., 2003. Diversity and creativity in work groups: a dynamic perspective on the affective and cognitive processes that link diversity and performance. In: Paulus, Paul B., Nijstad, Bernard A. (Eds.), Group Creativity: Innovation through Collaboration. Oxford University Press, Oxford, pp. 32–62.

Nemeth, Charlan Jeanne, Nemeth-Brown, Brendan, 2003. Better than individuals? The potential benefits of dissent and diversity for group creativity. In: Paulus, Paul B., Nijstad, Bernard A. (Eds.), Group Creativity: Innovation through Collaboration. Oxford University Press, Oxford, pp. 63–84.

Ng, Aik Kwang, 2001. Why creators are dogmatic people, 'nice' people are not creative, and creative people are not 'nice'. International Journal of Group Tensions 30, 293–324.

Ridgeway, Sally, 1989. Artists groups: patrons and gatekeepers. In: Foster, Arnold W., Blau, Judith R. (Eds.), Art and Society: Readings in the Sociology of the Arts. SUNY Press, Albany, pp. 205–220.

Sargut, Gökçe, 2006. Linking oligopolistic and organizational explanations of cultural production: an interorganizational approach to measuring product content homogeneity in the golden age of movies, 1931–1950. Paper Presented at the Annual Meeting of the American Sociological Association.

Strasser, Garold, Birchmeier, Zachary, 2003. Group creativity and collective choice. In: Paulus, Paul B., Nijstad, Bernard A. (Eds.), Group Creativity: Innovation through Collaboration. Oxford University Press, Oxford, pp. 85–109.

Thornton, Patrica H., Ocasio, William, 1999. Institutional logics and the historical contingency of power in organizations: executive succession in the higher education publishing industry, 1958–1990. American Journal of Sociology 105, 801–843.

Thornton, Patricia H., Ocasio, William, Lounsbury, Michael, 2012. The Institutional Logics Perspective: A New Approach to Culture, Structure, and Process. Oxford University Press, Oxford.

Uzzi, Brian, Spiro, Jarrett, 2005. Collaboration and creativity: the small world problem. American Journal of Sociology 111, 447–504.

Wacquant, Loic J.D., 1989. Towards a reflexive sociology: a workshop with Pierre Bourdieu. Sociological Theory 7, 26–63.

Watts, Duncan J., Strogatz, Stephen H., 1998. Collective dynamics of 'small world' networks. Nature 393, 440–442.

White, Harrison, 2002. Markets from Networks: Socioeconomic Models of Production. Princeton University Press, Princeton, NJ.

Cultural Psychiatry

Ahmed Hankir, National Institute for Health Research Academic Clinical Fellow in Psychiatry, Manchester University, Manchester, UK
Dinesh Bhugra, Institute of Psychiatry, King's College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Culture plays a crucial role in the crafting of a community's collective character. Moreover, the cultural community that a person is embedded in ineluctably has an influence on that individual's cognitive schema, behavior, occupational and social functioning, and modus operandi. Given the ubiquitous influence of culture on behavior of every sort, it follows that culture is an essential influence on deviant, dysfunctional, or otherwise problematic thoughts and behaviors as well, which carries the further implication that the healing professions, including psychiatry, need to be as attentive to culture as they are to clinical information.

Introduction

> We human beings are, after all, cultural beings…
> Professor Laurence Kirmayer, James McGill Professor and Director of the Division of Social and Transcultural Psychiatry at McGill University, Montreal, Canada

Culture has been defined as the "*Learned, shared and transmitted values, beliefs, norms and life ways of a particular group [of people] that guides their thinking, decision-making, and actions in patterned [stereotypical] ways.*" We know that culture is integrated; people acquire and assimilate culture. Although culture can change (gradually or abruptly), culture also ensures generational continuity. It is society and culture – through the proxy of policy-makers – that determine the provision of physical and mental health-care services (and the parity or *dis*parity of esteem between the two) and those who may receive and use them.

The roles that are played by psychiatry and society, in any given time and place, can influence the dialectic between the two, which can be complex, multifaceted, and constantly shifting due to cultural currents. Culture can mold the meaning of mental health symptoms (to an individual or a group) and can also influence the expression of these symptoms (the so-called 'idioms of distress'). Moreover, culture 'orchestrates' the interaction between the patient and the health-care system that they are situated in. Culture can influence the dynamic between a doctor and a patient, the establishment of a rapport between the two, and, consequently, the development of a therapeutic alliance.

Jung's Archetype, 'The Wounded Healer'

It is noteworthy that each and every individual is susceptible to developing some form of mental or physical illness. Indeed, one could argue that succumbing to illness in any of its many forms is what defines us as human beings. *We* thus have the capacity to 'transform' into treatment-seeking patients.

Carl Jung used the term the wounded healer as an archetypal dynamic to describe a phenomenon that may take place in the relationship between analyst and analysand (Burns and Burns, 2009). Jung discovered the wounded healer archetype in relation to himself; for Jung, "*…it is his own hurt that gives a measure of his power to heal…*" (Jung quoted in Anthony St, 1994). Despite the perception that doctors should be 'invincible' (Harvey et al., 2009), they, too, are human beings, and as such can develop psychopathology and become treatment-seeking patients. Doctors, however, have low levels of help-seeking for their own psychiatric problems, often only presenting to mental health services once a crisis arises. Fear of exposure to stigmatization is a critical factor contributing to secrecy and symptom concealment (Semple and Smith, 2008). The 2008 Stigma Shout survey of more than 4000 people using mental health services and carers revealed that health-care professionals are a common source of discrimination reported by those who have mental illness (http://www.time-to-change).

Just what are the consequences of a doctor with mental health challenges being immersed in a culture in which fellow health-care professionals hold stigmatizing views toward psychiatric illness? The consequences can in fact be quite catastrophic. Dr Daksha Emson, a brilliant psychiatrist with bipolar affective disorder, tragically killed her baby daughter and died following an extended suicide on 9 October 2000. An independent inquiry into Dr Emson's death concluded that she was the victim of stigma within the National Health Service (North East London Strateg, 2003). In other words: a culture of stigma is killing people.

In order to address the pernicious issue of stigma, a starting point would be for each and every health-care provider to be honest with himself or herself in order to develop an insight into their 'soul' and challenge the prejudice that may reside therein. Another way to change negative attitudes toward health-care professionals with psychopathology is to realize that those kinds of experiences may actually be beneficial as opposed to being disadvantageous. For example, a motif in the narratives of doctors who have firsthand experience of psychiatric illness is that they have become more empathetic toward their patients (http://www.rcpsych.ac.uk/). Indeed, in an interview with Laurence Kirmayer, he argues that "*…as a physician, coming to terms with one's own vulnerability and using it to help understand the predicaments of our patients can provide an*

important path to empathy and a way to mobilize their own capacities to heal…" (Hankir, 2014).

Cultural Stigma and Recruitment into Psychiatry

There are some cultures that are more stigmatizing than others toward psychiatry and psychiatric illnesses, and this can have an effect on recruitment into the specialty and the number of psychiatrists there are in a given place. For example, in an interview with a British consultant psychiatrist who rendered voluntary mental health humanitarian services to civilians in post-Arab-Spring Libya, he reported that there was not a single psychiatrist in the entire city of Misrata (Hankir and Sadiq, 2013). A contributory factor to this statistic is the perception, certainly in the Arab world but not confined to it, that psychiatrists are not 'real' doctors and that the specialty does not have a robust scientific basis or 'is not scientific enough.' Medical graduates may thus be reluctant to specialize in this area of medicine, since they may gravitate to what they perceive (and certainly what the media portrays) to be the more 'glamorous' and 'legitimate' specialties such as plastic surgery.

Culture can also determine the amount of resources that a government allocates to the provision of postgraduate psychiatric training and the funding that services receive. For example, culture can contribute to the disparity of esteem between physical and mental illness. From a health-care policy and provision point of view, this can result in the closure of psychiatric inpatient beds, which is what has been happening in the UK, resulting in some service users in a crisis having to travel hundreds of miles from where they live in order to receive emergency inpatient care.

Model of Culture and Individual

Hahn (Hahn, 1988) in his opus discusses and describes the sociocultural ontology of sickness and healing in which he also elaborates on the concepts of 'construction,' 'production,' and 'mediation' and argues that these concepts are interrelated. The formulation of the biopsychosocial model of disease has incorporated the aforementioned concepts. Hahn (Hahn, 1988) proposes that the patterned and stereotypical interactions between members of a group can influence an individual's behavior, which is subject to judgment by a collective entity known as society. These interactions can influence the definitions of health and suffering.

It is important to appreciate that the social definition of illness is the 'construction,' whereas sociocultural influences are the 'mediators.' According to Hahn (Hahn, 1988), 'production' is said to be engineered by socially organized interpersonal relationships and is the least recognized and understood concept out of the three mentioned. It is inevitable that the concepts of 'construction,' 'production,' and 'mediation' are strongly influenced by an individual's beliefs and the relationships that people have with other members of society, both of which may be pathological (i.e., the delusional beliefs of a person who has schizophrenia and the manipulative dependence on other people in patients who have emotionally unstable personality disorder).

'Medically Unexplained Symptoms' and Somatization

The psychosocial aspects of care are often recognized in dealing with common conditions like fibromyalgia syndrome in rheumatology or irritable bowel syndrome in gastroenterology. But understanding the personal and social context of illness is essential not only for categories of medically unexplained symptoms or functional syndromes that are a large part of practice in every medical specialty but for every health problem. We human beings are, after all, cultural beings. The way that we learn to see the world shapes every aspect of experience, including the ways we perceive and cope with illness and disease (Hankir, 2014).

Somatization is becoming increasingly recognized as a problem in patients in both general hospital and primary care settings. However, patients not only *somatize*, but they can also *psychologize*, as is illustrated in the following narrative from a foremost authority on cultural psychiatry (Hankir, 2014):

> …One [experience] that comes to mind was an experience of my own "attributional style." One day, I was on the floor playing with my infant daughter, and I vividly recall feeling so tired that I found it hard to get up from the floor. At the time, I interpreted this fatigue as a sign of depression – though my mood was fine. I saw my family doctor who diagnosed me with asthma (which I had never had before). So it seems I was engaged in psychologising, rather than somatising! Because I am a psychologically oriented practitioner, it was easy for me to devise a psychological explanation for my experience of fatigue. This really drove home the point that the ways we explain symptoms depend on personality, past experience and social context. It is important, however, to say that the division between psychological symptoms and physical symptoms can be quite arbitrary. Illness affects us as whole organisms – involving our bodies, thoughts and feelings. What we focus on – and what we feel we should conceal – is influenced by culture…
>
> Hankir, 2014

The Portrayal of Cultural Interpretations of Disease in 'Popular Culture'

There is a growing perception that science alone provides overall insufficient foundation for the holistic understanding of the interaction between health, illness, and disease (Hurwitz and Dakin, 2009). The health humanities has emerged as a distinct entity in an attempt to improve the provision of health services. The health humanities can broadly be described as the application of art and literature to medicine (Oyebode, 2009).

As mentioned above, culture can influence how an individual and a community perceive and react to physiological and psychological phenomena. Take, for example, the American physician and writer Khaled Hosseini's novel, *A Thousand Splendid Suns*, which focuses on the tumultuous lives of Mariam and Laila, two Afghan women. While Mariam is a teenager, she witnesses her mother having a seizure and believes the cause of this to be religious, i.e., her mother being possessed and persecuted by a malevolent entity (the Islamic belief of *jinn* possession) as opposed to 'medical,' i.e., a disturbance in the electrical activity of the brain such as epilepsy.

Although this is a fictional account, it does have relevance to clinical psychiatry. The explanatory models that a cultural group formulates can influence why they may, for instance, consult a faith healer as opposed to a general practitioner and hence not receive the benefits of early intervention. If, however, an immigrant from Pakistan does make it to, for example, an outpatient clinic in the Royal Bolton Hospital in England, attributing psychological phenomena, such as jinn possession, these patients should not be considered delusional, since the belief they have presented with would be consistent with the cultural norm of countries in that region of the world. This example illustrates both the importance of developing a cultural competence in psychiatry and also how literature and art can be used for educational purposes and undergraduate psychiatric teaching.

The study of the portrayal of mental illness in film also falls under the wide-ranging remit of the health humanities. Film wields a colossal power. Through the journey of a single film, we can take a roller-coaster ride across the spectrum of human emotion. Film can enthrall an entire auditorium heaving with people, or it can silence and even reduce them to tears. Films are extremely popular across the different cultures. India is the country that produces the largest number of films every year. In 2009 alone, the Hindi film industry Bollywood contributed to producing a staggering 1288 Indian feature films. America, Hong Kong, and Nigeria are examples of other countries where film industries are booming.

One could argue that as long as human beings continue to seek entertainment and escapism – for example, as the twentieth-century Noble laureate T.S. Elliot said, "*mankind cannot bear very much reality*" – cinema will remain deeply embedded in our society.

The story lines of films are influenced by the society we live in. Given that one in four of us has a mental illness at some point in our lives (World Health Organisation, 2001), mental illness and the psychiatrists who treat these illnesses play huge roles in our society and on our screens. Movies can provide an insight into what it is like to have a psychiatric illness, portray the role of a psychiatrist, and reveal how society reacts to either or both. In view of this, Professor Dhinesh Bhugra, President-elect of the World Psychiatric Association, actually examined Bollywood films produced since the early 1960s as a means to analyze the changes in Indian society's attitudes toward mental health issues (Deakin and Bhugra, 2012). Professor Bhugra's analyses reveal how in postcolonial India in the 1950s and 1960s, there were many films featuring people with mental illness who were subject to ridicule, but there were also some films that had sympathetic portrayals of sufferers of mental illness. In the 1970s and 1980s, when the country was going through major economic, social, and political crises, the portrayal of mental illness in film was very much of psychopaths who could not rely on the system to provide for the vulnerable, so they were vigilantes taking the law into their own hands. This image of mental illness sufferers transformed in the 1990s, when there were many motion pictures that portrayed morbid jealousy – typically films in which men were trying to control women, and women became a kind of object or property in the same way that economic liberalization allowed them to own objects, so women were seen as property.

Inaccurate portrayals of mental illness can perpetuate stigma and propagate myths, but when correctly presented, they can educate the public, inform employers, and empower service users. Film, therefore, can be used for educational purposes, and the Royal College of Psychiatrists (RCPysch) is not oblivious to this. Indeed, an example of an initiative that the College supports is a monthly blog on their Web site entitled '*Minds on Film*.' According to the Web site, *Minds on Film*, "*…explores psychiatric conditions and mental health issues as portrayed in a selection of readily available films…*"

The international medical film festival *Medfest* is another initiative that operates under the auspices of the Public Education Committee of the RCPsych. The festival in 2012 was entitled "HealthScreen: Understanding Illness through Film," and its aim was "*To stimulate debate of the social, political and ethical implications of portrayals of health and illness on our screens.*"

'Illness' versus 'Disease'

Eisenberg (Eisenberg, 1977) and Kleinman (Kleinman, 1980) argue that in psychiatry, as in the rest of medicine, a line of demarcation must be made between disease and illness. Disease, succinctly put, is about pathology. Illness, however, is something that a person develops when their disease starts to have effects on them and the people who encompass them. Disease literally means dis-ease – a malfunctioning of biological or psychological processes – whereas illness refers to the psychosocial aspects of disease. Illness thus includes the subjective perception of the disease, and its social and interpersonal dimensions and implications.

With respect to the assessment, diagnosis, and management of individuals presenting with physical or psychological symptoms, the culture can be such that doctors, including psychiatrists, are solely trained and interested in recognizing and addressing disease. Patients, however, are by and large interested in the effects that disease has on their functioning, i.e., the implications that illness might have on their social and occupational functioning (for example, the so-called 'illness' role and sick leave from work) and indeed existence in general. Thus, doctors are primarily preoccupied with formulating a differential diagnosis, and although patients are often concerned and interested in what the diagnosis may be, they are more concerned about what the mechanism of the disease is as well as how the disease might impact on their lives. There exists, therefore, a clash between the patient's views and explanatory models with those of his or her doctor. Consequently, the space and discord between the patient and the doctor can lead to tension and poor outcomes. A perceptive and culturally competent clinician must be able to detect this issue and act accordingly to ameliorate it, i.e., ensure that the provision of health-care services is service-user centered.

To make matters even more complex, there is the society-defined concept of sickness, which can determine what is deviant behavior and how much, for example, sick leave is permissible and socially acceptable for a particular illness. (If someone is receiving chemotherapy for breast cancer, society may, by and large, be empathetic toward this patient, whereas if a patient is admitted into a psychiatric unit for electroconvulsive therapy, society, particularly this patient's work colleagues

and employer, may not be as understanding and empathetic as they would be toward the breast cancer patient or even the same patient who may have had breast cancer in the past. In fact, the complete opposite might be true: they may shun and ridicule this person and even blame them and accuse them of having a 'weak' personality. This patient can then become ostracized, and social exclusion can cause them to become isolated, which can exacerbate the psychopathology they are experiencing. In extreme cases, this can even result in suicidal behavior).

A major step in recent years has been the effort to clarify how to collect and organize information about culture and context in mental health. Diagnostic and Statistical Manual (DSM)-5 (APA, 2013) (the recent revision of the diagnostic system of the American Psychiatric Association) introduces a cultural formulation interview. This is a basic approach to exploring the social and cultural context and meaning of illness. It should be part of the tool kit of every physician (Hankir, 2014).

The DSM-5 is an American diagnostic and classificatory system (APA, 2013), and its formulation is influenced by the cultural values in the United States. American culture, therefore, contributes to how disease is defined in DSM. As more and more countries are leaning toward commodification of health, with an ever-increasing remit for private sector providers and pharmaceutical companies, it is important that the sacred contract between medicine and society, as stipulated and enshrined in the Hippocratic oath, is adhered to so as to ensure that the provision of health care is evidence based and service-user centered, and that patients get the best treatments they need and deserve.

Gough's 'Social Contract'

Gough (Gough, 1936) defined and described a 'social contract,' which, in many ways, mimics discussions on cultural observations about what is normal and what system needs to be put in place in order to manage those who are described as 'abnormal.' Although, as Gough (Gough, 1936) had pointed out, in previous generations the original social contract was between the king and his royal subjects, this too was an implicit contract, i.e., obedience to the king was presumed providing the king conducted himself in a judicious fashion. The subjects would promise allegiance to the king, who in return offered them protection and good governance. Both the king and his royal subjects would honor their side of the agreement otherwise the contract would not be tenable. As society has evolved, so too has the application of Gough's social contract, i.e., the social contract between an employer and his or her employees and disciplinary procedures to deal with recalcitrant members of staff.

In the realm of the provision of mental health-care services, recent challenges have meant that psychiatric practice must adapt to the multitude of changes that are taking place in societies across the globe, i.e., migration (both internal and external, for example, the diaspora of Syrian refugees into Turkey consequent to the Arab Spring that spread to that region), seemingly perpetual changes in patient expectations, and advances in health-care systems. All of the aforementioned play a major role in the dynamic between patients and clinicians.

The Stress of Migration

The process of migration has been divided (in what we feel an arbitrary manner) into three stages as premigration, migration, and postmigration, although these may not be entirely distinct stages and a large degree of overlap may exist. Premigration is the period when an individual, having decided to migrate, starts to prepare himself or herself and others around him or her about the potential migration and familiarizes himself or herself with the practicalities needed for this purpose. The stage of migration itself is the actual physical process of relocating to the destination culture. Although the individual may physically be present in the country that they have migrated to, there may be a delay in mentally coming to terms with this. The postmigration phase is seen as the period of adjustment immediately after migration but may last for a protracted period, even years (Bhugra, 2003; Bhugra and Ayonrinde, 2004).

Birthplace is an index of migration, and as such can form part of the definition of a migrant ethnic population. Immigrants, particularly refugees, have long been treated as a marginalized, stigmatized, and socially excluded people from their destination culture. This and other issues that immigrants can encounter (i.e., acculturation stress, poverty, unemployment, and the inability to speak the dominant language to name but a few) can have profound effects on their mental well-being. There are a number of factors that are likely to influence postmigration adjustment and acculturation (Berry, 2007). Several theoretical constructs have been identified and are as follows: loss, fatalism (embedded in a sense of control), expectations of and from the destination culture, negative life events, and lack of social support, along with loss or clash of values and skill-set deficit (Furnham and Bochner, 1986; Ruiz et al., 2011). For those who migrate from rural to urban areas or across countries (or indeed continents), the process of acculturation becomes significant. Acculturation is a complex period of adjustment to the destination culture and may lead to an individual giving up some of the values they acquired from their origin culture. Acculturation itself is not straightforward and may be accompanied by stress.

Gender and other factors such as education, social class, housing, employment, and economic factors will also have an influence on the development of clinical depression. Chandra (Chandra, 2011) points out that the 'feminization of migration' is increasing with globalization and this can have a direct effect on female mental health.

Other vulnerable groups such as older adults, children and adolescents, homosexuals, and transgender individuals are at higher risk of developing depressive illness following migration. Other factors, especially perceived or real racism in the destination culture, may contribute to the genesis of depression by generating feelings of persecution and affecting aspiration and achievement.

Refugee Populations

Globalization has resulted in the mass movement of people as well as commodities and resources. People migrate for a number of reasons such as educational, professional, economic, and political, or any combination thereof. It is

important to differentiate between groups who have migrated voluntarily and those who have migrated through compulsion (i.e., were forced to migrate due to genocide in their country of origin).

Refugees are perhaps the most vulnerable group of all immigrants. According to the Geneva Convention (1951), a refugee is a person who has a "*well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or a particular political opinion, is outside the country of their nationality and cannot, or will not, because of fear, benefit from the protection of that country*." Conflict in the refugee's country of origin may result in the development of severe psychopathology such as post-traumatic stress disorder and major depressive disorder. Severe psychiatric illnesses in refugees may be a prelude to suicidal ideation. This and the way in which they are received by the destination country, dilapidated living conditions, and lack of social support and isolation all contribute and conspire to rendering this immigrant group particularly vulnerable to suicidal behavior (Bhugra et al., 2011).

Some migrant populations (including refugees) are wealthy, while others are in abject poverty. We should therefore realize that the phenomenon of migration is complex, and we should not come to any conclusions about the way in which migration influences social experience and health. Succinctly put, migrants are not a homogeneous group, and we should refrain from making generalizations about them and certainly do our utmost to prevent the perpetuation of negative stereotypes.

The act of migration itself can affect individuals in a diversity of ways. A person might migrate independently or with a group of people. Migration experiences and 'trajectories' can vary considerably. For professionals, such as doctors, the migration process may lead to changes in their status and income with specific issues related to settling down in the destination culture. For example, an immigrant may be a brilliant surgeon but may not be proficient in the language of the destination culture, so despite his ability to, for example, dexterously excise a benign tumor with surgical precision, he does not have the communication skills to 'operate' at the coalface of clinical medicine when trying to explain the procedure and its consequences and possible complications to a patient. Doctors who migrate to the UK have to take exams to prove that they are proficient in the English language and are familiar with colloquial parlance. It is not unheard of that a person who was a competent physician in India ends up working in a restaurant in Edinburgh as a chef because of his inability to speak fluent English. Such a change in status can precipitate psychopathology.

Cultural Psychiatry

Cultural psychiatry has developed as an entity in its own right over the last seven decades, with the term transcultural psychiatry being introduced in the psychiatric lexicon by Eric Wittkower in Montreal (dubbed 'the birthplace of cultural psychiatry') in the 1950s (Hankir, 2014). Byron and Mary-Jo Good from the University of California Davis were also central in a renewed engagement between medical anthropology and psychiatry initiated by the work of psychiatrist/anthropologist Arthur Kleinman.

Cultural psychiatry has focused on health disparities – both globally and locally – in terms of the needs of immigrants, refugees, and ethnocultural minorities. At the same time, it has continued to advocate for an integrative approach to care that challenges mainstream psychiatry. In recent decades, there has been a striking biologization of psychiatry, especially in the US, with the assumption that neuroscience is going to give us the core understanding of the etiology and treatment of illness and disease. To a large extent that has become the dominant view and the perspectives of social science and psychology have been downplayed (Hankir, 2014).

It has been argued, however, that human biology is cultural biology. The brain is the organ of culture, and we use our brains to acquire and adapt through cultural inventions like reading, mathematics, and other complex social practices. Many of the problems that are seen in psychiatry may reflect not structural abnormalities in the brain but the consequences of learning (programming the brain), and the unhealthy environments and social relationships that people must negotiate (Hankir, 2014).

Cultural Identity

All of us have a cultural identity, which is directly related to the cultural framework we carry within us. Cultural identity operates at both individual and group levels. Depending upon a number of factors, an individual may have a tendency to relinquish some aspects of their cultural identity and cling on to others. Individuals can actually have multiple cultural identities. For example, a person may be born in Northern Ireland, be brought up in England, attend university in Lebanon, form a romantic relationship with someone he or she has met in Canada, and work in Australia. Each country's culture will have an effect on the individual and result in the synthesis of multiple cultural identities within a single person.

Cultural Bereavement

As discussed and described above, an immigrant often has to come to terms with loss, and this is often heralded by bereavement. Psychoanalytic explanations of the loss of all that is important for an individual may lead to melancholia and depression and has been termed cultural bereavement (Eisenbruch, 1990, 1991). Stress of migration may also produce physical illnesses, and clinicians must be aware of the interaction between physical and mental illness. For example, a person with depressive illness who also has diabetes may lack the energy and motivation to comply with their medication. This patient may consequently develop life-threatening diabetic ketoacidosis (DKA). When a clinician encounters this scenario, it is paramount not to just provide 'symptomatic control,' i.e., to solely treat the DKA, but rather they must have the insight to address the underlying cause, i.e., to manage the depressive illness.

'Culture Conflict' and 'Culture Shock'

The term 'culture conflict' describes tensions between individuals or even the struggle that an individual has with oneself when they attempt to resolve and integrate perceived or real irreconcilable values (Bhugra and Becker, 2005).

An individual may have with oneself a conflict concerning one's own 'soul.' Take, for example, a single Muslim Saudi Arabian man who has migrated to a Western destination culture where prostitution and alcohol are legal. That very man must, if he cares to abide by the tenets of Islam, attempt to resist these temptations, which are seemingly ubiquitous. Despite the fact that he may be a pious and devout individual, he is, after all, a human being, and as such it is only natural for him to have a desire to succumb to the temptation. He may not be able to resist this powerful impulse and by capitulating to it be engulfed by remorse and pathos for (in his perception) betraying his faith. A possible consequence of this is the development of suicidal ideation and behavior in the context of a severe depressive illness as has been evidenced in clinical case studies (Hankir and Zaman, 2013).

Regarding tensions between individuals, we can use the example of when a younger member of a migrant family is attracted to the destination culture he or she is thrust into. This young immigrant may individuate and deviate from the culture of origin and imitate the prevailing culture. The family may then attempt to reel the perceived maverick youth back into the fold and inculcate him or her with traditional cultural values. Here we have a clash between the ostensibly dichotomous orthodox values of the origin culture and the less traditional values of the destination culture. This can result in the renegade individual feeling melancholic, isolated, ostracized, and, in extreme cases, suicidal. This would be consistent with the findings of research into intergenerational conflict in young people of Asian origin in the United States, which showed that intergenerational conflict increased the risk of suicide by 30-fold, especially in less educated young people (Bursztein and Makinen, 2010).

Culture conflict should be distinguished from 'culture shock,' which is what individuals may experience after migration, as their adjustment to the destination culture may occur over a prolonged and protracted period. There is, however, a degree of overlap between the two. Culture shock has been defined as the "*Sudden unpleasant feelings that violate an individual's expectations of the new culture [they have migrated to] and cause them to value their own culture negatively*" (Bhugra and Ayonrinde, 2004). Oberg identified six aspects of culture shock: strain; a sense of loss or feelings of deprivation; rejection by members of the new culture; role expectation and role confusion; surprise, anxiety, and indignation; and feelings of impotence (Oberg, 1960).

Prevention and Management Strategies for the Development of Psychopathology in Immigrants

Sometimes the clinician has a stereotyped approach, has not had interpreters available, or is unaware of crucial cultural factors when they encounter an immigrant who may be suffering from the symptoms of psychopathology. In order to tackle these issues, many models are available. These include culture brokers and cultural mediators as well as cultural liaison workers. The underlying principle, by and large, is that they will function as interlocutors and communicate between the services and the community the services are catering for.

The lack of translation services is a factor that severely impacts immigrant patients and prevents the completion of a thorough assessment and the offering of therapy. The provision of translation services may enable mental health service providers to detect the early warning signs of suicidal behavior in immigrant patients and consequently facilitate timely intervention to prevent completed suicides from occurring.

There have been further developments addressing the language barrier issue with many organizations developing leaflets in different languages. Indeed, there is evidence that supports the efficacy of leaflets in educating immigrant, cultural, and religious groups about mental illness and distress (Bhugra, 2004).

A Personalized Approach to Managing Mental Illness in Immigrants

> To the world you may be just one person but to one person you may be the world.
>
> Bill Wilson

Dignity is in the heart of everything that we do in the provision of mental health-care services. It can be all too easy to be prescriptive when assessing and treating a service user and the person can be lost in the protocol. By all means we condone and commend adherence to treatment algorithms and guidelines formulated by a robust evidence base in the management of psychopathology in immigrants; however, there can never be a replacement for personalized treatments (and perhaps in no other specialty is this more possible than in psychiatry).

We encourage mental health-care service providers to take a holistic and personalized approach in the assessment and treatment of immigrants who present with psychopathology, to deploy the 'soft skills' of listening and empathy to allow the narratives to unfold and flow, and, consequently, a rapport and therapeutic alliance to become established. "*The voice of the human heart needs no translation*" said Queen Rania of Jordan. Sometimes, just being made to feel that someone is genuinely listening to you with all earnestness and compassion and that someone actually cares can be the antidote that a lonely and forlorn human needs to nourish their soul with. In this scenario, nonverbal (and nonjudgmental) communication and tranquil silence can be more efficacious than any drug and offer a safe sanctuary where deep and seemingly indelible wounds can heal.

Conclusions

Culture is dynamic and affects all aspects of illness and behavior. The 'idioms of distress' that immigrants use may not necessarily match diagnostic criteria, and although culture is being homogenized by globalization, there are no universal

models of psychotherapy or pharmacotherapy for the treatment of mental illness. To summarize succinctly: One size does *not* fit all.

The lack of culturally competent health-care practitioners and culturally sensitive and personalized mental health services can present as major barriers for immigrants to accessing and engaging with the relevant services.

Assessing the mental health needs of immigrants includes openness to exploring cultural formulations to elicit the meaning of symptoms. A presenting complaint has to be explored to understand the suffering behind the symptom. Indeed, Oliver Sacks argues that "*…in order to restore the human subject at the centre, the suffering, afflicted, fighting human subject, we have to deepen a case history into a narrative or tale.*"

As in other areas of medicine, education is key, and training in cultural competency should be part of the curriculum for all mental health professionals, including teachers, nurses, and medical students in attempts to fathom the idioms of distress of immigrant groups, to detect psychopathology that may be a prelude to suicidal behavior, and to offer timely intervention to prevent attempted or completed suicide from taking place.

Lastly, we must be advocates and agents of cultural change in the broader society if we want things to get better for our patients.

See also: Culture-Specific Psychiatric Disorders; Hysteria: History and Critiques; Normality and Mental Health; Psychiatry and Religion; Psychiatry: Anthropological Aspects; Psychology of the Real Self: Psychoanalytic Perspectives; Social Intervention in Psychiatric Disorders.

Bibliography

APA, 2013. Diagnostic and Statistical Manual of Mental Disorders DSM-5. APP1, Washington, DC.

Burns, L., Burns, E.I., 2009. Literature and Therapy, p. 126.

Bhugra, D., 2003. Migration and depression. Acta Psychiatrica Scandinavica, Supplement 418, 67–72.

Bhugra, D., Ayonrinde, O., 2004. Depression in migrants and ethnic minorities. Advances in Psychiatric Treatment 10 (1), 13–17.

Berry, J., 2007. Acculturation and identity. In: Bhugra, D., Bhui, K. (Eds.), Textbook of Cultural Psychiatry. CUP, Cambridge, pp. 169–178.

Bhugra, D., Gupta, S., Bhui, K., et al., 2011. WPA guidance on mental health and mental health care in migrants. World Psychiatry 10 (1), 2–10.

Bhugra, D., Becker, M., February 2005. Migration, cultural bereavement and cultural identity. World Psychiatry 4 (1), 18–24.

Bursztein, L.C., Makinen, I.H., 2010. Immigration and suicidality in the young. Canadian Journal of Psychiatry 55 (5), 274–281.

Bhugra, D., 2004. Migration and mental health. Acta Psychiatrica Scandinavica 109 (4), 243–258.

Chandra, P., 2011. Mental health issues related to migration in women. In: Bhugra, D., Gupta, S. (Eds.), Migration and Mental Health. Cambridge University Press, Cambridge, pp. 209–219.

Deakin, N., Bhugra, D., April 2012. Families in bollywood cinema: changes and context. International Review of Psychiatry 24 (2), 166–172. http://dx.doi.org/10.3109/09540261.2012.656307.

Eisenberg, L., 1977. Disease and illness: distinction between professional and popular ideas of sickness. Culture Medicine and Psychiatry 1, 9–23.

Eisenbruch, M., 1990. The cultural bereavement interview: a new clinical research approach for refugees. Psychiatric Clinics of North America 13, 715–735.

Eisenbruch, M., 1991. From post-traumatic stress disorder to cultural bereavement: diagnosis of Southeast Asian refugees. Social Science and Medicine 33, 673–680.

Furnham, A., Bochner, S., 1986. Culture Shock. Routledge, London.

Gough, J.W., 1936. The Social Contract. Clarendon Press, Oxford.

Harvey, S., Laird, B., Henderson, M., et al., 2009. The Mental Health of Healthcare Professionals. Department of Health, London.

http://www.time-to-change.org.uk/news/stigma-shout-survey-shows-real-impact-stigma-and-discrimination-peoples-lives.

http://www.rcpsych.ac.uk/pdf/Doctors%20Go%20Mad%20Too.pdf.

Hankir, A., 2014. Interview with professor Laurence Kirmayer: director of cultural psychiatry McGill University, Montreal, Canada. World Journal of Medical Education and Research 5 (1). DAUIN 20140039.

Hankir, A., Sadiq, A., September 2013. Lessons from psychiatry in the Arab world–a Lebanese trainee psychiatrist's qualitative views on the provision of mental healthcare services for Palestinian refugees in Lebanon and an interview with a consultant psychiatrist on the effects of the Arab spring on the mental health of Libyans. Psychiatria Danubina 25 (Suppl. 2), S345–S349.

Hahn, R.A., 1988. A socio-cultural model of illness and healing. In: White, L., Tursky, B., Schwartz, G.E. (Eds.), Placebo-theory, Research and Mechanisms. Guilford Press, NY, pp. 167–195.

Hurwitz, B., Dakin, P., 2009. Welcome developments in UK medical humanities. Journal of the Royal Society of Medicine 102, 84–85.

Hankir, A., Zaman, R., July 12, 2013. Jung's archetype, 'the wounded healer', mental illness in the medical profession and the role of the health humanities in psychiatry. BMJ Case Reports. http://dx.doi.org/10.1136/bcr-2013-009990 pii: bcr2013009990.

Jung Quoted in Anthony Stevens, 1994. Jung, Oxford, p. 110.

Kleinman, A., 1980. Patients and Healers in the Context of Culture. University of California Press, Berkley, CA, pp. 259–286.

North East London Strategic Health Authority Report of an Independent Inquiry into the Care and Treatment of Daksha Emson and Her Daughter Freya, 2003. North East London Strategic Health Authority, London.

Oyebode, F., 2009. Preface Mindreadings, Literature and Psychiatry. RCPsych Publications, London.

Oberg, K., 1960. Culture shock: adjustment to new cultural environments. Practical Anthropology 7, 177–182.

Ruiz, P., Maggi, I.C., Yusim, A., 2011. The impact of acculturative stress on the mental health of migrants. In: Bhugra, D., Gupta, S. (Eds.), Migration and Mental Health. Cambridge University Press, Cambridge, pp. 159–171.

Semple, D., Smith, R., 2008. Oxford Handbook of Psychiatry, second ed. Oxford University Press, Oxford. 279–317.

World Health Organisation. The world health report 2001 – Mental Health: New Understanding, New Hope. http://www.who.int/whr/2001/en/ (accessed 01.07.14.).

Cultural Psychology

Richard A Shweder, University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

There is a substantial portion of the psychological nature of human beings that is neither homogeneous nor fixed across time and space. At the heart of the discipline of cultural psychology is the tenet of psychological pluralism, which states that the study of normal psychology is the study of multiple psychologies and not just the study of a single or uniform fundamental psychology for all peoples of the world. Research findings in cultural psychology focus on differences in mentalities by virtue of membership in ancestral groups and raise questions about the value and social function of alternative forms of human subjectivity.

Cultural Psychology: What Is It?

A major aim of the discipline called cultural psychology is to document variations in modes of (and ideals for) normal psychological functioning in different ethnic communities. One assumption of the discipline – the premise of nonuniformity – is that there is a substantial portion of the psychological nature of human beings that is neither homogeneous nor fixed across time and space. Cultural psychology seeks to document the protean cultural aspects of human psychological nature, especially the ways fundamental features of the mental life of all human beings (the various capacities definitive of a subjective life – the abilities to think, know, want, feel, and value things as good or bad) assume different shapes in different cultural traditions. The discipline can be defined as the study of the distinctive mentalities of particular peoples (Balinese Hindus, Satmar Hasidim, Chinese Mandarins, secular coastal elites in the United States) (see Garreau, 1982; Geertz, 1973; Greenfield and Cocking, 1994; Heine, 2011; Kitayama and Cohen, 2010; Levy, 1973; Markus et al., 1996; Menon, 2013; Miller, 1997; Nisbett, 2004; Shweder, 1991; Shweder and LeVine, 1984; Shweder et al., 1998; Stigler et al., 1990).

Cultural psychology can thus be distinguished from the study of general psychology, which is the study of mental structures and processes that are so widely distributed as to characterize the normal psychological functioning of all human beings (or perhaps even nonhuman primates as well). Research in cultural psychology has, for example, systematically corroborated the special status accorded to the defense of female honor in the mentality of many Southern American White males (Nisbett and Cohen, 1995). Research in cultural psychology has authenticated the claim that there is a positive sense of empowerment and a feeling of virtue associated with modesty and the attitude of respectful restraint in the psychology of women in some regions of the contemporary non-Western world (Menon and Shweder, 1998). Such feelings of power and goodness associated with modesty contrast with ideas about (and ideals for) psychological functioning constructed in the contemporary Anglo-American cultural region. Research in cultural psychology has documented factional intragroup and cross-cultural differences in the normative ethical concepts (such as liberty, equality, respect for status, loyalty, and purity) that give shape and meaning to human moral judgments and a personal sense of conscience (Haidt, 2013; Jensen, 2010; Shweder et al., 2003). Research findings in cultural psychology thus raise provocative questions about the integrity and value of alternative forms of subjectivity across cultural groups and within and between countries and regions of the world.

A second assumption of cultural psychology is that many mental states (and some mental processes) are best understood as by-products of the never-ending attempts of particular groups of people to understand themselves and to make manifest their self-understandings through social practices (Bruner, 1990; Geertz, 1973; Wierzbicka, 1993, 1999, 2013). That might be called the premise of self-reflexive social construction. Whether one studies Inuit Eskimos or Anglo-American middle-class conservatives or liberals, the aim in cultural psychology is to spell out the implicit meanings (the goals, values, and pictures of the world) that give shape to psychological processes (Briggs, 1970; White and Kirkpatrick, 1985). The aim is to examine the patchy or uneven distribution of those meanings on a global scale and to investigate the manner of their social acquisition, for example, by means of participation in the symbolic practices, including linguistic practices, of this or that tradition-sensitive cultural group (Haidt, 2013; Menon, 2013; Wierzbicka, 2013).

The Mental States of Others

One of the several aims of cultural psychology is to develop a language for the comparative study of mental states that makes it possible to understand and appreciate the mental life of others. 'Others' refers to members of some different cultural community who by virtue of life long membership in that group ascribe meaning to their lives in the light of wants, feelings, values, and beliefs that are not necessarily the same as one's own.

Cultural psychologists are interested, for example, in cultural variations in the degree to which feelings are constructed as 'emotions.' Anna Wierzbicka (1999) has suggested that while all normal human beings have feelings (e.g., pleasure and pain, arousal and serenity) many of the emotions lexicalized in the English language are not universally available in the mental life of people around the world. There has been

 http://dx.doi.org/10.1016/B978-0-08-097086-8.24053-1

much research in cultural psychology into the character of the particular emotions (Ifaluk 'fago,' American 'happiness') that are salient or important in different types of social worlds (Lutz, 1988; Kitayama and Markus, 1994; Shweder et al., 2010).

Cultural psychologists are interested as well in population-based variations in social cognition, moral judgment, and the sources of personal fulfillment or life-satisfaction in different social groups. For example, they have studied the origin, significance, and place of filial piety and the social motivation to achieve in some East Asian populations (Yang, 1997; also see Kitayama et al., 1997). They have investigated the self-empowering aspects of ascetic denial and other forms of sacrifice among high-caste women in South Asia (Menon, 2000, 2013). They are concerned to document the divergent meanings and distinctive somatic and affective vicissitudes of such experiences as loss (or success or of not getting the things you want, etc.) for members of different cultural communities. It is in the pursuit of such research questions that they have discovered replicable cultural differences in reports about the quality of the experience of loss (and gain). In comparison to majority populations in Northern Europe or the United States, majority populations in Samoa and China are more likely to react to apparent loss with feelings such as headaches, backaches, and other types of physical pain than with feelings such as sadness or dysphoria (Levy, 1973; Kleinman, 1986). Cultural psychologists also seek to document differences in modes of thought (e.g., analytic vs holistic) (Hong et al., 2000; Nisbett, 2004; Nisbett et al., 2001), in self-organization (e.g., interdependent vs independent) (Markus and Kitayama, 1991), and in moral judgment (e.g., reliance on an ethics of autonomy vs an ethics of community vs an ethics of divinity) across different types of groups (Haidt, 2013; Jensen, 2010; Shweder et al., 2003).

In most of its research the field of cultural psychology has been pluralistic in its conception of normal psychological functioning and interdisciplinary in its conception of how to go about studying the origin, meaning, and social role of particular mental states on a worldwide scale. The field draws together anthropologists, psychologists, linguists, biologists, and philosophers in its study of the diverse, yet potentially effective, modes of psychological functioning that have been produced, and socially endorsed, in different cultural traditions, and have made those cultural traditions possible. Indeed, cultural psychology is sometimes described as the study of the way culture and psyche make each other up.

The Tenet of Psychic Pluralism

The premise of nonuniformity when taken together with the premise of self-reflexive social construction sum up to the master principle of the discipline of cultural psychology. This is the tenet of psychic pluralism, which states that the study of normal psychology is the study of multiple psychologies (mentalities, subjectivities), and not just the study of a single or uniform fundamental psychology for all peoples of the world. Cultural psychology is thus the study of the way the human mind (understood to consist of an inherently complex, heterogeneous collection of abstract and/or latent schemata) can be transformed, and made functional, in a number of different ways, which are not equally distributed across time or space. Hence the slogan, "one mind, many mentalities: universalism without the uniformity," which is meant to give expression to goals of a discipline aimed at developing a credible theory of psychological pluralism. A much discussed essay on the 'weirdness' (in the sense of 'atypical' or statistically deviant) of experimental findings in mainstream Western psychology (research typically conducted on subject populations from Western Educated Industrial Rich Democratic (WEIRD) societies) highlights the hazards of rushing to claims about the so-called basic or fundamental or universal psychological processes (Henirich et al., 2010).

Cultural psychology can also be understood as a project designed to critically assess the limitations and incompleteness of all uniformitarian versions of the idea of psychic unity. Alternatively put, cultural psychology is the study of ethnic and cultural sources of diversity in emotional and somatic functioning, self-organization, moral evaluation, social cognition, and human development. It is the study of population differences in the things people know, think, want, feel, and value, and are customarily motivated to do, by virtue of life course membership in some recognizable group that has a history and a conception of its own destiny. "To be a member of a group," the eighteenth-century German romantic philosopher Johann Herder argued "is to think and act in a certain way in the light of particular goals, values and pictures of the world; and to think and act so is to belong to a group" (as represented by Berlin, 1976). Although the field of cultural psychology has many ancestral spirits (Giambattista Vico, Wilhelm Wundt, Wilhelm Dilthey, Edward Sapir, Ruth Benedict, Clifford Geertz) and some very prominent contemporary advocates (Jerome Bruner, Michael Cole, Jacqueline Goodnow, Patricia Greenfield, Shinobu Kitayama, Richard Nisbett, Hazel Markus, Anna Wierzbicka) Johann Herder is justly claimed as one of the original cultural psychologists, although he was probably not the first. For an account of the historical development of the field see Jahoda (1991) and Cole (1996).

Culture and the Custom Complex

One of the contributions of cultural psychology is to revive a conception of culture that is both symbolic and behavioral. In the history of twentieth-century anthropological thought the idea of culture has been variously defined, either behaviorally (as patterns of behavior that are learned and passed on from generation to generation) or symbolically (as the categories, beliefs, and doctrines that organize habitual or customary behavior and can be used to rationalize and justify a way of life). In research on the cultural psychology of a particular cultural community the notion of culture usually refers to community-specific ideas about what is true, good, beautiful, and efficient that are made manifest in behavior. To qualify as cultural those ideas about truth, goodness, beauty, and efficiency must be socially inherited and customary; and they must actually be constitutive of different ways of life, and

play a part in the self-understanding (including the self-criticism) of members of the community.

Alternatively stated, the concept of culture as used in cultural psychology refers to what Isaiah Berlin called goals, values, and pictures of the world that are made manifest in the speech, laws, and routine practices of some self-monitoring group. These are sometimes also called cultural models (D'Andrade, 1995; Shore, 1996). Cultural psychologists thus engage in the interpretive, symbolic, or cognitive analysis of behavior. They assume that actions speak louder than words and that customs and habits are a central unit for cultural analysis. Thus, what John Whiting and Irvin Child (1953) once referred to as the custom complex is a natural unit of analysis or starting point for a study in cultural psychology.

Whiting and Child introduced the idea of a custom complex in 1953 but the basic idea was not really taken up and carried forward until the rebirth of cultural psychology in the 1980s and 1990s. According to Whiting and Child a custom complex "consists of a customary practice [for example, a family meal, arranged marriage, animal sacrifice, a gender identity ceremony involving circumcision] and of beliefs, values, sanctions, rules, motives and satisfactions associated with it." There are many labels that social theorists in different scholarly disciplines have placed on this type of unit of analysis, for example, a custom complex, a life space, a habitus. Whatever the label, this is the type of unit of analysis that makes it possible to conceptualize cultural psychology as the study of the way culture (ideas about what is true, good, beautiful, and efficient made manifest in practice) and psyche (what people know, think, feel, want, value, and hence choose to do) afford each others realization, and thus make each other up.

A Provocative Example of a Custom Complex: Genital Surgeries in Africa

A rather dramatic but highly illuminating example of a custom complex is the circumcision ceremony or genital surgery that is customary for both boys and girls in many East and West African ethnic groups. In the countries of Sierra Leone, Mali, the Gambia, Ethiopia, Somalia, the Northern Sudan, and Egypt genital surgeries are culturally endorsed and popular for both males and females, and receive high approval ratings from the vast majority of men and women in the general population. Human rights advocacy groups in Europe and in the United States have criticized the practice, labeling it 'female genital mutilation.' The expression 'female genital cutting' is also widely used by the mainstream media in Europe and North America; although those are not the ways the custom is labeled by those for whom it is a customary and personally meaningful practice. A recent review of the medical and demographic literature on African genital surgeries by a panel of experts and published in *The Hastings Center Report*, the bioethics journal, suggests that the widely publicized claims about the severe and negative consequences of the custom for health, sexuality, and childbirth have been hyperbolic and rather one-sided and deserve to be viewed with a skeptical eye (Public Policy Advisory Network on Female Genital Surgeries in Africa, 2012; also Obermeyer, 1999). In a detailed description of the cultural psychology of male and female circumcision among the Kono people of Sierra Leone, Fuambai Ahmadu, who is both an insider and an anthropologist, has written as follows (Ahmadu, 2000: 301). "It is difficult for me – considering the number of these ceremonies I have observed, including my own – to accept that what appear to be expressions of joy and ecstatic celebrations of womanhood in actuality disguise hidden experiences of coercion and subjugation. Indeed, I offer that most Kono women who uphold these rituals do so because they want to – they relish the supernatural powers of their ritual leaders over against men in society, and they embrace the legitimacy of female authority and, particularly, the authority of their mothers and grandmothers."

Among the various goals, values, and pictures of the world (the cultural psychology) that make this practice meaningful and satisfying for those men and women for whom it is a custom complex are the following:

1. A culturally shared belief that the body (especially the genitals) of both males and females are sexually ambiguous until modified through surgical intervention. According to this picture of the world the foreskin of a boy is viewed as a feminine element and masculinity is enhanced by its removal. Similarly the slightly protruding visible part of the clitoris is viewed as an unwelcome vestige of the male organ, and Kono females, as described by Ahmadu, seek to feminize and hence empower themselves by getting rid of what they perceive as a dispensable trace of unwanted male-like anatomy.
2. A culturally shared aesthetic standard in terms of which the genitals of both women and men are viewed as ugly, misshapen, and unappealing if left in their natural state. For many African men and women the ideal of beauty is associated with a sexual anatomy that is smooth and cleansed (shaved) and free of all fleshy encumbrances. This aesthetic look has gained some popularity on a global scale and is well represented in the 'vaginal rejuvenation' operations offered to a middle-class clientele by cosmetic surgeons in Europe and North America.

There is of course much more to be said about the beliefs, values, sanctions, rules, motives, and satisfactions associated with female (and male) genital surgeries in Africa. Nevertheless in the light of these and other culturally endorsed reasons a genital surgery is experienced as an improvement of the body in many East and West African ethnic groups (Ahmadu, 2000; also see Shweder, 2013).

The cultural psychology of the customary practices of any particular community is likely to result in a depiction of other minds that is unsettling, astonishing, or at least surprising for those whose goals, values, and pictures of the world are characteristic of the cultural psychology of some other group. Radical divergences in visceral attachments and in the moral evaluation of particular custom complexes (such as male and female genital surgeries) are themselves important topics for research in cultural psychology. Indeed, the cultural psychology of moral evaluation is currently an active research area (Fassin, 2012). Moral concepts (e.g., human rights, justice as equality) privileged in some sections of society and in some regions of the world are not necessarily the moral concepts that are most salient and important in other sectors of society and regions of the world (Haidt, 2013; Jensen, 2010; Shweder et al., 2003).

Cultural Psychology: What It Is Not

The tenet of psychic pluralism and the emphasis on goals, values, and pictures of the world as a source of psychological differences between cultural communities distinguishes cultural psychology from other fields of study (such as cross-cultural psychology and national character studies) with which it should not be confused.

It Is Not Cross-cultural Psychology

Research in cultural psychology proceeds on the assumption that psychological diversity is inherent in the human condition, and that culture and psyche are interdependent and make each other up. It should be noted in passing that any theory of psychological pluralism would lack credibility if it staunchly denied the existence of any and all universals. Indeed, as should be obvious, cultural psychology presupposes many psychological universals (e.g., feelings; wants; goals; ideas of good and bad, of cause and effect, of part-whole relationships, etc.) (see Shweder et al., 1998). However, the search for and the privileging of things that are uniform across all peoples is a project that goes under other names, for example, general psychology or perhaps even cross-cultural psychology.

Some of the goals of cross-cultural psychology (which is here contrasted with cultural psychology) have been described by Segall et al. (1998). One goal is "to generate more nearly universal psychology, one that has pan-human validity" and to attain "a universally applicable psychological theory." A second closely related goal it to "keep peeling away at the onion skin of culture so as to reveal the psychic unity of mankind at its core." It is for that reason that cross-cultural psychology (not to be confused with cultural psychology) can be viewed as a vigilant cousin of general psychology; for they both share the same uniformitarian goals. Such goals give a distinctive character to cross-cultural psychology. And they help explain why all of the following kinds of activities are typical of research by cross-cultural psychologists and distinguish it from research by cultural psychologists.

Thus one of the aims of cross-cultural psychology is to determine the boundary conditions for generalizations generated in the Western labs with Western (mostly college student) subjects, generalizations which, prior to critical examination by cross-cultural psychologists, have been presumptively interpreted as fundamental or natural and universalized to the whole world (see the results of the review of research in cross-cultural psychology by Henirich et al. (2010)). The aim here for the cross-cultural psychologist is not to represent the distinctive cultural psychology of particular peoples. This aim does not result in research focused on differences in the way members of different communities perceive, categorize, feel, want, choose, evaluate, and communicate that can be traced to differences in salient community-based goals, values, and pictures of the world. Rather the aim for the cross-cultural psychologist is to make sure that the hope for universal psychology is truly universal and to throw out any claim that only holds in the Anglo-American world. This is an extremely useful corrective for the tendency of Western psychologists to over generalize their findings, but it is not the same as undertaking a project in cultural psychology.

A second aim of cross-cultural psychology is to establish comparability or equivalence for measuring instruments across different populations. Often the goal here is to try to show that people in different cultures really are alike, and that any reported differences in performance were due to noise, or inappropriate measuring instruments, or bad translations, or misunderstandings about the way to ask and answer questions. The instincts of a cultural psychologist run in quite a different direction. For a cultural psychologist (not to be confused with a cross-cultural psychologist) the 'noise' is interpreted as a signal about true differences in cultural meanings (goals, values, and pictures of the world), and not as something to eliminate or overcome. Indeed, cultural psychologists are likely to worry if one's measuring instruments travel easily and well from university classroom to university classroom around the world and display the same psychometric properties here and there. They may suspect that one has not really landed in a truly different culture at all. This is because "Peeling away the onion skin of culture so as to reveal the psychic unity of mankind at it core" is not what cultural psychology is about.

A third major aim of cross-cultural psychology is to focus on the so-called independent variables of the cultural environment (for example, nucleation of the family, literacy vs nonliteracy) that are thought to either promote or retard psychological development. In such research development is almost always defined in terms of universal norms for promoting progress in cognitive, emotional, or social functioning (e.g., Piaget's notion of 'formal operational' thinking or Ainsworth's notion of healthy 'attachment'). Cultural psychology is primarily concerned with the elaboration and discovery of alternative or plural norms for successful psychological development, which is another way it can be distinguished from cross-cultural psychology (LeVine, 1990).

It Is Not National Character Studies

Attempts to characterize whole populations in terms of generalized dispositions (e.g., authoritarianism, Apollonianism, high need for achievement) went out of fashion in psychological anthropology in the late 1950s and early 1960s. This happened for several reasons but largely because looking for variations in types of personality traits to explain differences in cultural practices or custom complexes (and vice versa) turned out to be something of a dead end. It was discovered that if one tries to describe individuals within and across cultural communities in terms of general dispositions or traits of character, then within group variations typically exceed between group variations. It was discovered that hypothesized 'modal personality' types typically characterize only about one-third of the population in any particular cultural group.

A major insight, although a fragile one, of recent work in cultural psychology is that is better to represent and interpret human behavior the way sensible economists do rather than the way global personality trait theorists do. That is to say, it is better to think about behavior as emanating from agency or the exercise of will by individuals, and to analyze it as the joint

product of 'preferences' (including goals, values, and ends of various sorts) and 'constraints' (including beliefs, information, skills, material and social resources, and means of various sorts). This avoids the hazards of dispositional approaches in which behavior is interpreted as the by-product of mechanical forces pushing both from inside (in the form of personality traits) and outside (in the form of situational pressures). Ultimately, a fully successful piece of research in cultural psychology must avoid nominal dispositional categories such as holistic versus analytic and render behavior intelligible in terms of the particular goals, values, and pictures of the world that motivate and inform the domain-specific behaviors and routine practices of specific intentional agents. To do otherwise is to reify cultural stereotypes and fall into some of the traps of the past.

The Future: Going Indigenous

The field of cultural psychology that has reemerged on the North American and European scene during the past 30 years is quite similar to an intellectual movement that has grown up in the non-Western world and is increasingly known as indigenous psychology. Indeed the Society for Humanistic Psychology of the American Psychological Association has even created an Indigenous Psychology Task Force (http://www.apadivisions.org/division-32/leadership/task-forces/indigenous/index.aspx). One of the most eminent theoreticians of this movement, Kuo-shu Yang, the Taiwanese social psychologist who was originally trained in the United States, lists several ways to 'indigenize' psychological research. Here are four of Professor Yang's virtues for the aspiring indigenous psychologist of China (Yang, 1997):

1. Give priority to the study of culturally unique psychological and behavioral phenomena or characteristics of the Chinese people.
2. Investigate both the specific content and the involved process of the phenomenon.
3. Make it a rule to begin any research with a thorough immersion into the natural, concrete details of the phenomenon to be studied.
4. Let research be based upon the Chinese intellectual tradition rather than the Western intellectual tradition.

Those are some of the virtues that define cultural psychology as well, although it remains to be seen how many of us can live up to such demanding standards. Even today with the rebirth and renewal of cultural psychology not all research actually begins with fieldwork or with "a thorough immersion into the natural concrete details of the phenomenon to be studied." All too often research still starts with a published finding from some Western lab, which is then subjected to critical examination by means of various attempts at replication with populations from other societies. Whether the habits and norms for conducting research in mainstream psychology will change remains to be seen. The recent attention to the 'weirdness' of research findings with restricted subject populations in WEIRD societies has certainly sparked some debate about how to reform the norms for research training in the United States (Henirich et al., 2010). Although fieldwork, local language learning, and the sampling of diverse populations are not yet a standard feature of training in the discipline of psychology, given the increasingly international and interdisciplinary character of collaborative scholarship in cultural psychology one looks forward to more and more research that keeps faith with Kuo-shu Yang's high ideals.

See also: Affect and Emotion, Anthropology of; Body: Anthropological Aspects; Cognitive Anthropology; Colonialism, Anthropology of; Comparative Method in Anthropology; Cross-Cultural Research Methods in Psychology; Cultural Relativism, Anthropology of; Cultural Views of Life Phases; Culture: Contemporary Views; Ethnocentrism; Family: The Anthropology of the Concept and Its History; National Character; Personhood, Anthropology of; Prosocial Behavior, Cultural Differences in; Social Constructionism; Urban Anthropology; Values Across Cultures, Development of.

Bibliography

Ahmadu, F., 2000. Rites and wrongs: an insider/outsider reflects on power and excision. In: Shell-Duncan, B., Hernlund, Y. (Eds.), Female "Circumcision" in Africa: Culture, Change and Controversy. Lynne Rienner, Boulder, CO, pp. 283–312.
Berlin, I., 1976. Vico and Herder. Hogarth Press, London.
Briggs, J.L., 1970. Never in Anger: Portrait of an Eskimo Family. Harvard University Press, Cambridge.
Bruner, J.S., 1990. Acts of Meaning. Harvard University Press, Cambridge, MA.
Cole, M., 1996. Cultural Psychology: A Once and Future Discipline. Harvard University Press, Cambridge.
D'Andrade, R.G., 1995. The Development of Cognitive Anthropology. Cambridge University Press, New York.
Fassin, D. (Ed.), 2012. A Companion to Moral Anthropology. Wiley-Blackwell, Malden.
Garreau, J., 1982. The Nine Nations of North America. Avon, New York.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Greenfield, P., Cocking, R., 1994. Cross-cultural Roots of Minority Child Development. Erlbaum, Hindsdale, NJ.
Haidt, Jonathan, 2013. The Righteous Mind: Why Good People Are Divided by Politics and Religion. Vintage.
Heine, S., 2011. Cultural Psychology. W.W. Norton and Co, New York.
Henirich, J., Heine, S.J., Norenzayan, A., 2010. The weirdest people in the world? Behavioral and Brain Science 33, 61–83.
Hong, Y., Morris, M.W., Chiu, C., Benet-Martinez, V., 2000. Multicultural minds: a dynamic constructivist approach to culture and cognition. American Psychologist 55, 709–720.
Jahoda, G., 1991. Crossroads between Culture and Mind: Continuities and Change in Theories of Human Nature. Harvard University Press, Cambridge.
Jensen, L.A., 2010. The cultural-developmental theory of moral psychology: a new synthesis. In: Jensen, L.A. (Ed.), Bridging Cultural and Developmental Approaches in Psychology. Oxford University Press, New York.
Kitayama, S., Cohen, D., 2010. Handbook of Cultural Psychology. Guilford, New York.
Kitayama, S., Markus, H.R., 1994. Emotion and Culture: Empirical Studies of Mutual Influence. American Psychological Association, Washington, DC.
Kitayama, S., Markus, H.R., Matsumoto, H., Norasakkunit, V., 1997. Individual and collective processes in the construction of the self: self-enhancement in the United States and self-depreciation in Japan. Journal of Personality and Social Psychology 72, 1245–1267.
Kleinman, A., 1986. The Social Origins of Distress and Disease. Yale University Press, New Haven.
LeVine, R.A., 1990. Infant environments in psychoanalysis: a cross-cultural view. In: Stigler, J., Shweder, R., Herdt, G. (Eds.), Cultural Psychology: Essays on Comparative Human Development. Cambridge University Press, New York, pp. 454–474.
Levy, R., 1973. Tahitians: Mind and Experience in the Society Islands. University of Chicago Press, Chicago.
Lutz, C., 1988. Unnatural Emotions: Everyday Sentiments on a Micronesian Atoll and Their Change to Western Theory. University of Chicago Press, Chicago.

Markus, H.R., Kitayama, S., 1991. Culture and the self: implications for cognition, emotion and motivation. Psychological Review 98, 224–253.

Markus, H.R., Kitayama, S., Heiman, R., 1996. Culture and "basic" psychological principles. In: Higgins, E.T., Kruglanski, A.W. (Eds.), Social Psychology: Handbook of Basic Principles. The Guilford Press, New York, pp. 857–913.

Menon, U., 2000. Does feminism have universal relevance? the challenges posed by Oriya Hindu family practices. Daedalus: Journal of the American Academy of Arts and Sciences 129, 77–100 (Special issue on "The End of Tolerance: Engaging Cultural Differences").

Menon, U., 2013. Women Well-being, and the Ethics of Domesticity in an Odia Hindu Temple Town. Springer, India.

Menon, U., Shweder, R.A., 1998. The return of the "white man's burden": the encounter between the moral discourse of anthropology and the domestic life of Hindu women. In: Shweder, R.A. (Ed.), Welcome to Middle Age! (And Other Cultural Fictions). University of Chicago Press, Chicago.

Miller, J.G., 1997. Theoretical issues in cultural psychology. In: Handbook of Cross-cultural Psychology. Theory and Method, vol. 1. Allyn and Bacon, Inc., Boston, MA, pp. 85–128.

Nisbett, R.E., 2004. The Geography of Thought: How Asians and Westerners Think Differently…and Why. Free Press, New York.

Nisbett, R.E., Cohen, D., 1995. The Culture of Honor: The Psychology of Violence in the South. Westview Press, Boulder, CO.

Nisbett, R.E., Peng, K., Choi, I., Norenzayan, A., 2001. Culture and systems of thought: holistic vs. analytic cognition. Psychological Review 108, 291–310.

Obermeyer, C.M., 1999. Female genital surgeries: the known, the unknown, and the unknowable. Medical Anthropology Quarterly 13, 79–106.

Public Policy Advisory Network on Female Genital Surgeries in Africa, 2012. Seven things to know about female genital surgeries in Africa. Hastings Center Report 42, 19–31. http://www.thehastingscenter.org/Publications/HCR/Detail.aspx?id=6059.

Segall, M.H., Lonner, W.J., Berry, J.W., 1998. Cross-cultural psychology as a scholarly discipline. American Psychologist 53, 1101–1110.

Shore, B., 1996. Culture in Mind: Cognition, Culture and the Problem of Meaning. Oxford University Press, New York.

Shweder, R.A., 2013. The goose and the gander: the genital wars. Global Discourse: An Interdisciplinary Journal of Current Affairs and Applied Contemporary Thought. http://www.tandfonline.com/doi/abs/10.1080/23269995.2013.811923#.UsbUQ1OmbyA.

Shweder, R.A., 1991. Thinking through Cultures: Expeditions in Cultural Psychology. Harvard University Press, Cambridge.

Shweder, R.A., Goodnow, J., Hatano, G., LeVine, R., Markus, H., Miller, P., 1998. The cultural psychology of development: one mind, many mentalities. In: Damon, W. (Ed.), Handbook of Child Psychology, vol. 1. John Wiley and Sons, New York, pp. 865–937.

Shweder, R.A., Haidt, J., Horton, R., Joseph, C., 2010. The cultural psychology of the emotions: ancient and renewed. In: Lewis, M., Haviland-Jones, J., Feldman Barrett, L. (Eds.), The Handbook of Emotions, third ed. Guilford Publications, Inc., New York.

Shweder, R.A., LeVine, R.A., 1984. Culture Theory: Essays on Mind, Self and Emotion. Cambridge University Press, New York.

Shweder, R.A., Much, N.C., Mahapatra, M., Park, L., 2003. The "big three" of morality (autonomy, community, divinity), and the "big three" explanations of suffering. In: Shweder, R.A. (Ed.), Why Do Men Barbecue?: Recipes for Cultural Psychology. Harvard University Press, Cambridge, MA.

Stigler, J., Shweder, R.A., Herdt, G., 1990. Cultural Psychology: Essays on Comparative Human Development. Cambridge University Press, New York.

White, G.M., Kirkpatrick, J., 1985. Person, Self and Experience: Exploring Pacific Ethnopsychologies. University of California Press, Berkeley.

Whiting, J.W.M., Child, I., 1953. Child Training and Personality. Yale University Press, New Haven, CT.

Wierzbicka, A., 1993. A conceptual basis for cultural psychology. Ethos 21, 205–231.

Wierzbicka, A., 1999. Emotions across Languages and Cultures: Diversity and Universals. Cambridge University Press, New York.

Wierzbicka, A., 2013. Imprisoned in English: The Hazards of English as a Default Language. Oxford University Press.

Yang, Kuo-shu, 1997. Indigenizing westernized Chinese psychology. In: Bond, M.H. (Ed.), Working at the Interface of Culture: Twenty Lives in Social Science. Routledge, London.

Cultural Relativism, Anthropology of

James W Fernandez, University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses views toward cultural relativism in the history of anthropology, from Franz Boas to the present. It shows the grounding of versions of relativism in the anthropological field experience of cultural diversity, and identifies the tensions between anthropology and more foundationalist philosophies. Yet it notes the openings toward more perspectival views in philosophy as well, and distinguishes between kinds of relativism in anthropological thought and practice. Debates argued over foundationalism and relativism too often draw on stereotyped depictions of the alternatives.

Relativism is a doctrine that, recognizing the importance of the perspectival in experience, offers a skeptical resistance to the philosophical and intellectual interest in universalisms and absolutes. This resistance in the Western tradition dates back to the Sophists, and running through Hume, Kant, Marx, and Nietzsche extends to the modern pragmatists who conceive of knowledge as relative to contexts of power and social solidarity. Anthropological or cultural relativism derives from the awareness produced by ethnographic fieldwork of the multitude of different lifeways and cultural perspectives in the world. It has also given evidence of local awareness of the impact of the perspectival in human affairs. This is found in the folklore of different peoples. In America, there has been a particularly strong strain of cultural relativism deriving from the work of the founder of American anthropology, Franz Boas, and his students. This American school emphasized the shaping force of culture on behavior and understanding. Strong forms of this relativism are found in the work of Benedict and Herskovits. Milder forms are found in the skepticism of Sapir and Geertz about the determined scientific pursuit of universals. In any clarifying discussion, anthropological relativism has to be differentiated into epistemological, descriptive, or methodological, and normative or moral forms. Few anthropological relativists would subscribe to all these forms and are usually partial to one or several only. Cultural relativism, despite the many attacks upon it, continues to inspire interest and commitment to one or another or all of its forms. Since the early 1990s anthropological interest in universal human rights has been a considerable challenge to relativistic thinking within the discipline.

Relativism, which may be taken in most general terms as the awareness of the presence and importance of the perspectival in judgment and understanding, is generally understood as a philosophical problem and a hobgoblin of the pretensions to generalization and universalization of thought characteristic of the intellectual classes. But much if not all that has been taken up and intellectualized by these classes was first found among the folk as part of their lore anchored in the particularities of experience. Anthropologists and folklorists can readily confirm this observation, for in practically any book of proverbs will be found axiomatic observations on particular experiences which are relativistic in import. People are usually aware from their domestic everyday experience of the difference of perspective and the relativity in understanding between men and women, the old and the young, the parent and the child, and the slow and the quick. We might observe that it is philosophers and theologians who have lived among absolutes and have worked to impose them, while in 'the practice of everyday life' the common people characteristically live among relatives. Lately, philosophers have been increasingly taking cognizance of that awareness and its implicit perspectivalism if not relativism.

In philosophy, to be sure, there has been a long tradition of relativistic thought. The Sophists' cultivation of the contestatory rhetoric of community life and community organization is often taken as relativistic in implication and action, and Hume's (1975) skepticism about general observations including any universal moral principles is surely relativistic in spirit. Kant's view of the constructed nature of the metaphysical and Hegel's historicism are instrumental in the development of the more radical relativisms found in Marx and Nietzsche. Modern-day French philosophers continue this perspectivalism whether in the 'archaeology of knowledge' and the 'conditions of understanding' explored by Michel Foucault (1979) or in Jacques Derrida's 'deconstruction' of arguments ordinarily taken as unquestionably foundational (1993). American pragmatism especially has anchored its 'antifoundationalism' in an awareness of the perspectival from William James' *A Pluralistic Universe* (1909) to Richard Rorty's emphasis on the community and cultural context of all understanding (1989, 1999). These scholars focus, respectively, on knowledge as conditioned by power, rhetorical or textual construction, and social solidarity. Such views have steadily and perhaps definitively been undermining overarching claims to metaphysical universals. It has certainly given rise, in respect to moral philosophy, to attempts to relate relativism to reason, that is, to practice a 'reason within relativism.'

Anthropology, a discipline mainly emergent since the mid-nineteenth century, has not been immune to these philosophical developments and the crisis of representation and legitimation they usher in. Indeed, it has become, through its various arguments for cultural and linguistic autochthony, a central player in the relativism/universalism wars. The interest of anthropologists, particularly in the issue of cultural relativism, derives first from ethnographic field experience of the variable proverbial wisdom of local life lived on the ground, as it were, or from the fact that anthropologists usually work comparatively across cultures and, therefore, across

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12047-1

cultural and language perspectives, or from the long interest in anthropology in the difficult problem of universal human rights. Philosophy, understandably, has reacted by exercising pointed critique of the *incoherence* of anthropological relativism – it is inescapably self-refuting and it undermines its own truth claims – and the *perniciousness* of anthropological relativism – it undermines philosophical commitment to the improvement of thought and the discovery of 'truth.'

The anthropology of Franz Boas is generally seen as the fountain head of cultural relativism in American anthropology. Although originally a physicist, Boas was increasingly cautious both about universalization of things human and the hierarchization of human groups which often accompanied such practice. His chief intellectual opponent was classic evolutionism, with its invidious ranking of societies on a scale from savagery to civilization. His chief popular opponent was racism, or in its benign form, the Civilizing Mission. Indeed one of Boas' few popular books, in an otherwise vast oeuvre – *The Mind of Primitive Man* (Boas, 1938) – attacked 'racial determinism,' that is, the association between biology and 'personality' (by which he meant behavior in the broadest sense). He did this by focusing on the shaping force of culture upon behavior and understanding (that is, upon 'mind'). In this book itself, therefore, preceded by any number of lectures and papers dating from the late nineteenth century, Boas laid the early ground work for the 'cultural relativism' of his students, although this was not a term he employed to any extent.

There are several other classic stimuli to relativistic views in the Boas oeuvre. In 'The study of geography' (Boas, 1887), while resisting geographical determinism, he stimulated thinking on methodological relativism by pointing to two different and justifiable approaches to field data, the nomothetic or law seeking and the idiographic or descriptive, each of which differently attuned to local circumstance gave different yet necessary understandings of the field data. It may be said, particularly in the postmodern times, that although increasingly convinced by the 'structuring power of tradition,' Boas in his emerging relativism never abandoned entirely the universalist hope of the scientific mission, and method and the seeking, however cautiously, of lawful generalizations.

Cultural relativism in its strong form is identified with two of Boas' foremost students, Ruth Benedict and Melville J. Herskovits. In her widely read *Patterns of Culture*, Benedict (1934) took particular aim at universalist notions of the normal and abnormal. This argument was made effective through its arresting portraits of three cultures, Pueblo, Dobu, and Northwest Coast Indians, whose characteristic behaviors, abnormal from a Western perspective, are portrayed as normal from the local perspective.

Benedict repeatedly resisted ethnocentric and provincial views of the normal and argued a 'great arc of human possibility' (Benedict, 1934, p. 34) of which no culture could take full advantage and on which every culture had to make its particular choices. It was Melville Herskovits, however, who most directly framed the issue of 'cultural relativism' in his general text *Man and His Works* (Herskovits, 1948), and whose collected arguments could be posthumously published as *Cultural Relativism* (Herskovits, 1972).

None of the Boasians, including Herskovits, were out and out relativists in the epistemological sense that they gave up on the possibilities of discovering universals beyond the fact of cultural relativity. In all their writings there appears the promise of those eventual discoveries. They are all to be considered, rather, methodological and moral relativists whose main aim was to resist ignorance of and moral superiority toward other cultures, and exploitative intolerance of them. The aim was to increase mutual respect and tolerance in the world. Herskovits argued, however, that such cultural relativism was not derived from any kind of moral philosophy, but rather from the empirical anthropological observations on diversity by field-working anthropologists. Nevertheless, he struggled with many dilemmas of philosophical type in his 'cultural relativism': negotiating the universal and the particular, the practical and the ethical, the 'is' and the 'ought,' anthropological science and anthropological humanism, the cross-cultural and the within-cultural view, and the applied and the philosophical. His strategy was to convert these dilemmas, which arise from the fate of the field-working man or woman of science finding himself or herself inescapably situated in the fold of philosophy, into the revelatory paradoxes of anthropological practice (Fernandez, 1990).

The heyday of 'cultural relativism' in the United States – in the sense of seeking to define a durable doctrine in the face of persistent assault from various moral philosophies on the one hand and philosophies of science on the other – were the decades of the mid-twentieth century, the 1930s through the 1960s. Subsequently, however, some notable contributions to this enduring issue have come forth together with a contemporary challenge to relativistic thought: the human rights movement. Marshall Sahlins (1976), although not espousing 'cultural relativism' in its paradoxical sense of moral obligation to defend diverse local moral orders, nevertheless conducted a compelling argument against the culture-free universalization, mainly among economists and political economists, which accepts utility interests as the dominant principle in practical reasoning. In the 1980s, a notable argument developed over relativism and the role of culture in determining reality principally among three anthropologists: Renato Rosaldo, Clifford Geertz, and Melville Spiro.

Rosaldo (1989) addresses the impact on older notions of the stable truths and dispassionate objectivities of classic social analysis made by awareness of both the anthropologist and his or her informants as 'positioned (and repositioned) subjects' recurrently engaged in an ongoing dramatic process of 'relational understanding.' In such a process truth becomes a very moving target only perspectivally to be apprehended, if at all, and in any case best interpreted by narrative analysis firmly set within and not beyond cultural contexts. While not espousing a straight-line cultural relativism, he notes and resists the return of a 'nostalgia' for the authoritative, often imperious, truths of objective social science – a nostalgia most often self-justifying by using a stereotyped notion of relativism.

This strategy of skepticism has also been that of Geertz, the central player in the interpretive approach that puts the play of perspectives at the center of anthropological inquiry. In 1984 he offered an intricate argument, exemplified in the double

negative of the title of the article, 'Anti–anti relativism' (Geertz, 1984). He takes to task the resurgence of absolutist interests in affirming some metaphysical or, what is the same, very physical, that is to say biological, realities behind the influence of cultural and language differences. This search for 'nature's own vocabulary,' and the security of objective biological truths on this side of culture, truncates what anthropology is all about, and Geertz questions the kind of 'hearty common sense' that simply ignores the diversity of lifeways plentifully evidenced in the anthropological archive. While written in defense of that broadly rather than narrowly humanizing knowledge of diversity laboriously acquired by anthropology, Geertz neither directly espouses in any direct way the moral commitments present in an earlier 'cultural relativism' nor denies the probability that on the other side of culture there may be some universals to discover.

In defense of the possibility of scientific, which is to say universal, measurability and explanation across cultural boundaries, Spiro (1986) develops a useful contrast between three types of cultural relativism: descriptive (that there are important empirical differences between cultures), normative (that evaluative judgment of the norms expressed in these differences are always relative to cultural contexts and hence not generalizable), and epistemological (that culture determines what we know and how we know it). Spiro accepts descriptive relativism and the perspectivalism it reflects, and the cautions of field methods it enjoins. He rejects the strong form of normative relativism, which argues against any possibility of cross-cultural judgment of human practices, as well as epistemological relativism in general, which would argue both for ontological incommensurability between cultures and for the irrelevance or perversity of the discovery of biological and psychological commonalities in humankind. It has been the tendency, he argues, in the fin-de-siècle age of textual hermeneutics, to espouse an immoderate form of epistemological relativism; something that the earlier cultural relativists, although committed to descriptive and moderate normative relativism, never did. The consequence is simply to deny the possibilities of an explanatory science of humans anchored in a valid and verifiable logic of discovery, with all the implications for professional authority that such a denial of principled comparative inquiry entails.

In the 1990s, an important area of challenge to cultural relativism arose in the argument for universal human rights. At first glance, cultural relativism would seem to be an impediment to such universalization of a value. Indeed, oppressive states proclaiming state sovereignty over their citizens often claim justification in 'cultural relativism.' This defense begs the question for all concerned of cultural boundaries within states, which is to say the question of the definition of culture and its boundaries in a pluralizing, globalizing world. That the relation between cultural relativism and universal human rights need not be taken as inevitably oppositional is seen in recent efforts on the part of the Human Rights Commission of the American Anthropological Association, to posit a species-wide disposition to create culture and cultural difference, and the corresponding right to such creation and such difference as long as it does not impose on that right in others (American Anthropological Association, 1998; Turner, 1997).

More recently Carrithers (2005) has made a convincing argument that there is a moral universal implicit in anthropology's dedication, through its field working requirement of participation and observation, to obtaining the fullest possible understanding of the other's cultural perspective. The success of such dedication, he argues, is only possible if the anthropologist and his or her interlocutors are persuaded in their interaction by the universal principles of mutual trust, forebearance, and acceptance of the other's worth. These principles, in turn, rest upon the knowledge, nurtured by the pluralism of understandings so evident in the great multitude of human lifeways studied by anthropologists, that the human condition 'could always be otherwise.'

There may remain problems with these enlightenment arguments, as theory and practice, but they move the debate away, as for example does Geertz' argument before them, from a simple and fruitless opposition in which one or another pole of the debate is confined to a stereotyped reading. Also they preserve the practical and methodological consequences of cultural relativism as an open, generous, and locally attuned posture of critical cultural inquiry into the 'other.'

In reviewing this century-long argument over relativism in the social sciences, we may note several principal points. First, there is the tendency for stereotypic views of the opposite argument to emerge, with the consequence that the opponent's views are trivialized and simplified. Second, strong and moderate forms of cultural relativism in its various aspects (descriptive, normative, and epistemological) must be distinguished if the argument is to obtain to any clarifying power. Third, the methodological aspect of cultural relativism and its importance, in its descriptive and normative aspect, as a posture of inquiry must be recognized for those, mostly but not exclusively anthropologists, who have chosen to work in other cultures. Fourth, the discussion of cultural relativism is inevitably related to one's definition of culture and particularly cultural boundedness. Fifth, it must be recognized that the logics of inquiry differ significantly between the sciences and the humanities. Hence the discussion of relativism is itself relative to what disciplinary faculty is anchoring and being evoked in the discussion. Sixth, there is a politics of relativism which is inevitably present – the Boasian relativists, for example, were pronouncedly egalitarian, pluralist, and antiauthoritarian in their values. Relativism has been a recurrent irritation, a valuable provocation, and a useful restraint on imperious inquiry in the academy, and has produced over this long period some of the most interesting arguments about just how individual scholars and researchers with different agendas working in different cultures with differing worldviews are to come together and engage in a productive and fair-minded inquiry. Without the gadfly of relativism, the horsemen of social science, galloping through their research agendas and in pursuit of theoretical confirmation, might well be more headstrong and less empathetic beings.

See also: Benedict, Ruth (1887–1948); Human Rights and Social Work; Mill, John Stuart (1806–73); Postmodernism: Philosophical Aspects; Relativism: Philosophical Aspects.

Bibliography

American Anthropological Association, 1998. Code of Ethics of the American Anthropological Association. Web Document. http://ameranthassn.org/committees/ethics/ethcode.htm.

Benedict, R., 1934. Patterns of Culture. Riverside Press, Cambridge, UK.

Boas, F., 1887. The study of geography. Science 9, 137–141.

Boas, F., 1938. The Mind of Primitive Man. Macmillan, New York.

Carrithers, M., 2005. Anthropology as a moral science of possibilities. Current Anthropology 46, 433–456.

Derrida, J., 1993. Aporias: Dying–Awaiting (One Another at) the 'Limits of Truth' (Mourir–s' Attendre aux 'Limites de la Vérité') (T. Dutoit, Trans.). Stanford University Press, Stanford, CA.

Fernandez, J.W., 1990. Tolerance in a repugnant world and other dilemmas in the cultural relativism of Melville J. Herskovits. Ethos 18, 140–164.

Foucault, M., 1979. Passim. In: Morris, M., Patton, P. (Eds.), Foucault: Power, Truth, Strategy. Feral, Sydney, Australia.

Geertz, C., 1984. Distinguished lecture: anti–anti relativism. American Anthropologist 86, 263–279.

Herskovits, M.J., 1948. Man and His Works. Knopf, New York.

Herskovits, M.J. (Ed.), 1972. Cultural Relativism; Perspectives in Cultural Pluralism. Random House, New York.

Hume, D., 1975/1777. Enquiries Concerning Human Understanding and Concerning the Principles of Morals. Clarendon Press, Oxford, UK.

James, W., 1909. A Pluralistic Universe. Longmans, New York.

Rorty, R., 1989. Contingency, Irony and Solidarity. Cambridge University Press, New York.

Rorty, R., 1999. Introduction: Relativism: Finding and Making, Philosophy and Social Hope. Penguin, New York. pp. xvi–xxxii.

Rosaldo, R., 1989. Culture and Truth. Beacon Press, Boston, MA.

Sahlins, M.D., 1976. Culture and Practical Reason. University of Chicago Press, Chicago.

Spiro, M.E., 1986. Cultural relativism and the future of anthropology. Cultural Anthropology 1, 259–286.

Turner, T., 1997. Human rights, human difference: anthropology's contribution to an emancipatory cultural politics. Journal of Anthropological Research 53, 273–291.

Cultural Resource Management: Conservation of Cultural Heritage

Thomas J Green, Arkansas Archeological Survey, Fayetteville, AR, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cultural resource management is the practice of conserving and preserving cultural resources. In a broad sense, cultural resources include archeological and historic properties, museums, archeological parks, shipwrecks, cultural landscapes, sites of religious and cultural significance, traditional ways of life and languages, and associated artifacts and documents. Modern society values cultural resources for their associative, a esthetic, informational, and economic significance. Such resources are nonrenewable and threatened by economic development, population expansion, and natural forces. Conflicts develop between preservation advocates and development agencies, national governments and local communities, and indigenous peoples and professionals over how such resources should be utilized and preserved. International, national, and local communities have legal and bureaucratic systems to deal with these issues.

Introduction

Cultural resource management (CRM) is the practice of conserving and preserving cultural heritage. Tangible cultural resources include archeological, historical, architectural, and engineering properties, and associated artifacts and documents. Other tangible cultural resources include museums, libraries and archives, archeological parks, shipwrecks, cultural landscapes, religious sites, and locations where plants and animals are gathered by traditional peoples. Intangible cultural resources include traditional lifeways, religious practices, and languages (King, 2009). These resources represent the cultural heritage of individual peoples and nations, and, collectively, they represent the heritage of the world. Cultural resources reflect finite periods of time or ephemeral cultural traditions. Once destroyed, they are gone forever. Economic development, agricultural practices, looting, international trafficking in stolen artifacts, and natural erosion destroy thousands of cultural resources every year. For this reason, the primary management goal is preservation *in situ*. The term *cultural resource management* is used mostly in the United States by archeologists working for federal and state agencies or for private firms contracting with such agencies (Fowler, 1982; Green and Doershuk, 1998). *Heritage management* or *cultural heritage management* are similar terms used in many parts of the world and can be used interchangeably with CRM.

Value of Cultural Resources

People value cultural resources for present and future uses (Lipe, 1984; Torre, 2002). They are tangible links to the past allowing people to experience and relate directly to their history, and they hold an evocative power sometimes lacking in documentary sources. Heritage properties have symbolic values for nations and ethnic groups because they represent a common history and identity, as well as cultural and spiritual norms. The process of deciding which cultural resources to preserve and monumentalize is therefore frequently contentious and politicized (Smith, 1994; Tunbridge and Ashworth, 1996).

The a esthetic value of cultural resources, especially architectural properties, provides incentive for the reuse of historic buildings long after original uses are no longer economically viable. The preservation of the historic core of nineteenth-century business districts in the United States is a priority because of the perceived a esthetic value of early buildings. The a esthetic appeal of architectural ruins around the world is a primary reason for their continued preservation.

Cultural resources have informational value. Archeological sites provide the only record of human achievement for much of history. At best, written documents cover events of only the past 5000 years and only in a few parts of the world. In parts of North America written records represent a scant 200 years. Government and business records, newspapers, diaries, etc., are fragile and subject to loss due to natural events, deterioration, neglect, and intentional destruction. Also, official records provide information primarily about dominant cultures and leaders. For these reasons the physical evidence in historic and archeological sites may be the only record of many past human activities, even in relatively recent times, and their destruction results in an immeasurable loss of information.

Cultural resources have economic value resulting primarily from their associative and a esthetic values. People want to visit heritage sites. Tourism is one of the largest industries in the world, and heritage tourism is one of the largest sectors of the industry. For many countries, such as the United Kingdom and Jordan, tourism is a major part of the economy. In the United States, cultural resources support a billion-dollar industry in the state of New Mexico alone.

Legal Protection

Cultural resources are the tangible links to the cultural heritage of the world. For this reason, UNESCO and the International Council on Monuments and Sites (ICOMOS), a private organization advising UNESCO, have produced charters establishing standards for preservation (US/ICOMOS, 1999). Two foundational documents guiding heritage management at the international level are the 1964 Venice Charter adopted by ICOMOS, and the 1970 Convention Concerning the

Protection of World Cultural and Natural Heritage adopted by UNESCO. UNESCO manages the World Heritage List, which recognizes cultural and natural sites of world significance, and the World Heritage Fund, providing financial assistance for various preservation projects.

Despite international interest in preserving heritage resources, most CRM is conducted under the laws of individual nations or local divisions within nations (McManamon et al., 2008; Messenger and Smith, 2010; Green, 2011). Laws have been established to protect cultural resources threatened by economic development projects, urban sprawl, vandals, looters, and neglect. In addition, nations must make a multitude of complex decisions regarding these issues. Governments must determine the ownership of individual artifacts and collections; oversee excavation permits and arrange for the use and care of resulting artifacts and records; mediate controversial issues such as the excavation of graves and human remains; allocate responsibility for funding archeological excavations or restoration projects in advance of construction projects; and determine the extent of involvement of various interested parties, such as indigenous peoples, in the processes of decision making. Most countries of the world have laws and regulations dealing with these issues, but some issues are more important in one country than another. Legal traditions and history invariably affect the law. For example, in Mexico the state asserts ownership of all prehistoric artifacts. US law reflects strong traditions of private property, and landowners retain ownership of artifacts found on their land.

Management Requirements

Knowledge is crucial in managing cultural resources (Green, 2008). Agencies need to know where the resources are, which are significant, why they are significant, and to whom. Plans for their preservation must be developed. If sites are threatened by dams, roads, water systems, or buildings, planning statutes usually require agencies or various units of government to address the impact of such projects. Nongovernmental corporations or organizations are frequently required to assess the effects of their projects on cultural resources or pay to have this work done. Accurate information is also crucial to protecting sites from looting and vandalism. Cultural resource inventories and evaluation studies should be conducted by appropriately trained specialists, such as archeologists, architectural historians, historians, and anthropologists. Thus, an inventory of significant cultural resources is the primary basis for management decisions.

Many countries have developed criteria to determine which cultural resources are significant, and these criteria vary from country to country. In the United States, properties as young as 50 years old are evaluated, while in Jordan only properties older than AD 1750 need be considered. The range in kind and type of significant properties also varies from place to place. Some countries consider only archeological and historic properties, while others include sacred and traditional cultural sites, and even contemporary expressions of culture such as festivals and dances. Consultation with local communities is essential to understand the values associated with particular cultural resources. Government agencies usually make decisions concerning significance, but even if a country does not have a formal evaluation procedure, someone must decide which resources are ultimately important and worthy of mitigation or preservation.

Once an inventory is obtained, it is necessary to determine which resources are threatened and need immediate consideration. It is a sad fact, but not all sites determined to be culturally significant can be preserved. Developing priorities for action is usually a process of give and take between those with preservation interests and others primarily working toward economic development. When properties cannot be preserved, appropriate mitigation measures must be developed. The type of mitigation adopted depends on the type of cultural resource affected and the values the property holds. Archeological sites are usually excavated, or sometimes buried, for protection. Architectural and engineering properties are documented through photography, photogrammetry, measured drawings, and the salvage of individual architectural features. When feasible, architectural properties can be moved to new locations. It is difficult to develop satisfactory mitigation measures for the religious or sacred properties important to traditional communities. Hard decisions have to be made about whether such controversial projects will be completed as planned or moved to another location.

Cultural resources in national parks and other protected areas may not be threatened by economic development projects, but they still need active management to prevent deterioration. The site of Petra in Jordan is a good example. Petra is famous for massive sandstone architecture and for water-control systems built by the Nabataean Arabs before Roman occupation. These famous architectural elements are continually eroded by wind-driven sand and rain and will disappear if not conserved. Jordanian conservation scientists are actively developing methods to preserve the site. Around the world, even in protected areas, limited financial resources dictate preservation priorities.

Another important element in the management of cultural resources involves the care of archeological and museum collections, archives, and documents. Great expense is involved in providing state-of-the-art, climate-controlled, and safe environments for storage, processing, and use of such collections. The conservation of paper, metal, and other perishable artifacts is a science in itself. Heated debates arise around the world about the appropriate distribution of artifacts. Many countries, such as Guatemala and Turkey, have increasingly strict laws about the transport and use of artifacts, and many European countries with colonial traditions and large archeological collections from the Middle East and elsewhere are facing various questions of repatriation (Herscher, 1998). The establishment and funding of curation facilities are part of this debate, as well as being crucial to the availability of knowledge and resources to future generations.

Traditions and points of view concerning the treatment of human remains and associated funerary objects differ widely among the cultures of the world. Such belief systems must be considered in the formation of CRM policies, as well as in national and international law. In some countries, indigenous peoples have insisted on the repatriation of human remains and associated funerary objects and other objects of cultural

patrimony (Smith, 2004). A satisfactory balance between the interests of the broader society in cultural resources and the needs of individual communities has not yet been reached, and these are among the most controversial issues facing heritage managers today. Conflicts are exacerbated in countries with recent histories of social injustice, and otherwise routine decisions are fraught with emotion. Animosities can be allayed as governments take responsibility for the legal protection of various rights and as people are forced to deal amicably and respectfully with each other's goals and beliefs. Consultation between government bodies and local communities is essential for successful preservation.

Tourism and Cultural Resource Management

The management of archeological parks, museums, and historic districts for tourism is a significant part of CRM. Tourism offers many opportunities for the preservation and restoration of heritage properties around the world, and because of the economic benefits of tourism, many individual communities and national governments are developing properties for tourists. The correct manner of development (Boniface, 1995), the authenticity and quality of interpretative presentation (Jerome, 2008), and the type and use of restoration methods are hotly debated issues. The tourist infrastructure, as well as interpretive facilities like trails and museums, can negatively impact the very cultural resources and landscapes tourists want to visit. In addition, conflicts can be expected when the interests of local communities are not considered.

Public Support and Education

The preservation of cultural resources is dependent on the support of the general public. The UNESCO and ICOMOS charters stress that the success of local and national preservation initiatives is dependent on programs educating the public about the value of cultural resources. Heritage education in the public schools is required in many countries to foster a preservation ethic and pride in national heritage. Governmental organizations and professional societies often cooperate in publicity campaigns to combat looting. In those parts of the world where sites of general interest represent a cultural heritage different from that of the contemporary inhabitants, education about the values and economic uses of these resources is especially important. For example, Roman and Byzantine ruins in modern Islamic countries and Native American sites in the United States are often valued more by foreign visitors than local communities.

The varieties of cultural resources in the world and the multiplicity of values attached to them make cultural resource management a complex practice. The preservation of cultural resources in the face of economic development, population growth, looting, and world conflicts requires integrated legal tools and protection policies at all levels of government, as well as trained professionals to implement them.

See also: Data Bases and Statistical Systems: Archives and Historical Databases; Ecotourism; Environmental Archaeology; Environmental Cognition, Perception, and Attitudes; Indigenous Rights; Tourism, Geography of; Tourism; Tradition, Anthropology of; Tradition: Social.

Bibliography

Boniface, P., 1995. Managing Quality Cultural Tourism. Routledge, London.
de la Torre, M., 2002. Assessing the Values of Cultural Heritage. Research Report. The Getty Conservation Institute, Los Angeles.
Fowler, D.D., 1982. Cultural resources management. In: Schiffer, M. (Ed.), Advances in Archaeological Method and Theory, vol. 5. Academic Press, New York, pp. 1–50.
Green, T.J., 2008. Cultural resource management. In: Bentley, A., Maschner, H., Chippindale, C. (Eds.), Method and Theories in Archaeology. AltaMira Press, Walnut Creek CA, pp. 375–394.
Green, T.J., 2011. International variety in CRM. In: King, T.F. (Ed.), Companion to Cultural Resource Management. Wiley-Blackwell, Malden, MA, pp. 420–438.
Green, W., Doershuk, J.F., 1998. Cultural resource management and American archaeology. Journal of Archaeological Research 6, 121–167.
Herscher, E., 1998. Many happy returns? New contributions to the repatriation debate. American Journal of Archaeology 102, 809–813.
Jerome, P., 2008. An introduction to authenticity in preservation. APT Bulletin: Journal of Preservation Technology 39, 3–7.
King, T.F., 2009. Cultural Resource Laws and Practice, fourth ed. Alta Mira Press, Walnut Creek, CA.
Lipe, W.D., 1984. Value and meaning in cultural resources. In: Cleere, H. (Ed.), Approaches to the Archaeological Heritage. Cambridge University Press, Cambridge.
McManamon, F.P., Stout, A., Barnes, J. (Eds.), 2008. Managing Cultural Resources: Global Context, National Programs, Local Actions. One World Archaeology Series. Left Coast Press, Walnut Creek, CA.
Messenger, P.M., Smith, G.S. (Eds.), 2010. Cultural Heritage Management: A Global Perspective. University Press of Florida, Gainesville, FL.
Smith, L., 1994. Heritage management as postprocessual archaeology. Antiquity 68, 300–309.
Smith, L., 2004. Archaeological Theory and the Politics of Cultural Heritage. Routledge, London.
Tunbridge, J.E., Ashworth, G.J., 1996. Dissonant Heritage: The Management of the Past as a Resource in Conflict. John Wiley & Sons, Chichester.
US/ICOMOS, 1999. ICOMOS charters and other international doctrinal documents. US/ICOMOS Scientific Journal 1 (1), 55–59.

Cultural Rights and Culture Defense: Cultural Concerns

Alison Dundes Renteln, University of Southern California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines debates in political theory, international law, and the social sciences generally regarding the nature and scope of cultural rights. Cultural rights – the most understudied of all human rights – remain controversial, for reasons explored in this article. After identifying several key conceptual questions, the focus shifts to the protection of these rights in legal process and the types of difficulties that may arise when courts must wrestle with the analysis of sometimes unfamiliar claims. When individuals and groups attempt to follow cultural traditions accepted in their countries of origin but that are apparently prohibited in their new homelands, they may raise cultural defenses to explain the motivations for their actions. Defendants invoke the defense in a wide variety of contexts and in many countries. To afford insight into the nature of these disputes, a few of the more famous cases are described here. Although general principles of law and the explicit protection of culture in international human rights treaties appear to require the consideration of culture in legal proceedings, the extent of this protection remains unclear. Ultimately, the question is how much weight cultural rights deserve if their protection risks violating the rights of others, such as women and children.

The scope and nature of cultural rights remains an important topic in the twenty-first century. After interpreting the concept of culture as traditional culture, this article provides an overview of the manner in which individuals and groups invoke 'the right to culture' in order to protect their way of life from governmental intrusion. When members of ethnic minorities and indigenous groups are punished for following their traditions, they sometimes raise cultural defenses during their trials (Renteln, 2004, 2010a,b). This article discusses debates over cultural rights in the scholarly literature, primarily in political theory and international law (Francioni and Scheinin, 2008; Stamatapolou, 2007; Niec, 1999; Levy, 1997; Prott, 1988; Szabo, 1974; Goulet, 1981). One major issue is often who is the proper spokesperson for the group when controversies arise. Examples of litigation are presented from Australia, Canada, France, and the United States. Cultural rights are justified not only by explicit provisions in domestic law and international conventions, but also because their protection is required by general principles such as identity, freedom, and tolerance (Poulter, 1986, 1987, 1998).

Although cultural rights are sometimes regarded as the least developed category of human rights, they are unquestionably a key part of international human rights law (Symonides, 2000; Meyer-Bisch, 1993). Cultural rights have been highly controversial because their protection sometimes requires that other human rights be limited (Shachar, 2001). There is substantial debate about the proper interpretation of cultural rights, about which individuals are entitled to invoke them, and about what obligations governments have to enforce these rights (Stavenhagen, 1995; McGoldrick, 1991; Addis, 1991; Poulter, 1987).

Cultural rights are significant because culture is an intrinsic part of individual and group identities (Kymlicka, 1995). Not only does culture exert a crucial influence on the formation of the identity of individuals, but it also helps perpetuate ethnic groups. It is important to appreciate the status of cultural rights in law and politics (Nafziger et al., 2010).

The Term 'Culture'

In order to interpret cultural rights, it is first necessary to define the term culture. There are many different notions of culture including 'high culture' that refers to museums, operas, and the like; 'popular' or 'mass' culture that means comic books, films, and pulp fiction; and 'traditional' culture that can be construed as a way of life. While some commentators who analyze cultural rights have in mind 'high culture,' most of the discussion focuses on so-called 'traditional culture.' Of the multitude of definitions of culture, a particularly useful one was devised by the Canadian Commission for UNESCO, the UN agency responsible for the protection of culture:

> Culture is a dynamic value system of learned elements, with assumptions, conventions, beliefs and rules permitting members of a group to relate to each other and to the world, to communicate and to develop their creative potential (1977).

culture differs from society inasmuch as culture is an abstraction, whereas society is the collection of individuals in the community.

When culture is analyzed in the context of cultural rights, it is usually used as a synonym for a way of life. Culture encompasses many aspects of life including attire, child-rearing practices, diet, marriage customs, politics, religion, death rituals, and so on. In short, culture is constituted of the many and diverse folkways of ethnic groups. One attempt to delineate the many facets of culture is found in the Human Relations Area Files compiled at Yale University. Although some contend that projects which aim at documenting actual cultural practices risk 'freezing' them in a particular moment in time, cultural theorists invariably acknowledge that cultures are dynamic systems. While subcultures exist and have differing traditions, cultural communities nevertheless often share common characteristics such as language. Although contemporary theories generally emphasize the

fluidity and hybridity of culture, the reality is that many cultural forms persist for centuries (Brumann, 1999).

Cultural Rights

A major question has been whether cultural rights should be conceptualized as individual rights or group rights. Theorists, many of whom come from individual rights traditions, associate group rights with utilitarianism and worry that recognizing group rights may lead to the denial of individual rights. In reality, however, rights that have historically been considered bona fide individual rights such as religious freedom require the existence of a group. While this matter has generated much debate, there is no reason why the right to culture cannot be conceived of both as an individual and as a group right. Insofar as cultural rights include rights to language and religion, it would be difficult to deny the essential group nature of such rights (Das, 1994). Moreover, even if the right to culture were an individual right, it could still come into conflict with other individual rights.

Ordinarily, individuals and peoples exist in cultures without having to think self-consciously about the protection of their traditions, most of which are not controversial. It is only when their way of life is threatened by a dominant political system that ethnic minorities are compelled to assert a 'right to' culture. In this sense cultural rights represent a legal means by which minorities can shield themselves from intervention by the dominant political apparatus, which is normally the state. Governmental intrusion is a manifestation of cultural hegemony, or the imposition of the dominant culture by elites on ethnic minority groups.

Intervention by governments is not an infrequent occurrence as there is a widespread presumption that cultural minorities should become assimilated. When the folkways of a minority group clash with the standards of the dominant system, the expectation is that the members of the minority group should discard their customs. This is sometimes expressed in the form of a proverb: "When in Rome, do as the Romans do," based on the monocultural paradigm. However, shedding their folkways may not be a simple matter as individuals are often strongly motivated to follow their cultural traditions. Sometimes, they may face a dilemma as failure to comply with the folk law can lead to divine retribution or social ostracism. These individuals either violate the state law and face the consequences in the official legal system or else comply with state law and face dire consequences within their customary law systems.

The Cultural Defense

If individuals do, in fact, have cultural rights, this implies that there is a correlative duty on the part of governments to allow their cultural practices. This obligation is not an absolute one, however, and cultural rights will be allowed unless a more important, countervailing reason requires that they not be protected. The challenge, naturally, is to distinguish which are valid reasons for overriding traditions and which are not.

Most of the time governments decide that cultural traditions are not permissible because they threaten public health, safety, or morals. If individuals refuse to discontinue the practices, asserting that they have a right to follow their traditions, they may face criminal prosecution. In these circumstances, ethnic minority defendants may invoke what has been called the 'cultural defense.' Such a defense, while not formally recognized by any legal system, sometimes operates to acquit or to mitigate punishment. The core idea behind the defense is that individuals who act under cultural imperatives should be considered less culpable under the law than those who act in the absence of such an imperative. In order for judges to evaluate the extent to which a defendant's cultural background affected his or her behavior, they must consider cultural evidence relating to the offense. The chief difficulty in many cases is the tendency of courts to exclude cultural evidence as irrelevant. If courts do treat the information as admissible, establishing credentials of experts may be difficult because the nature of the anthropological evidence may not conform precisely to requirements of some rules of evidence.

Supporters of the cultural defense favor its adoption in order to ensure the consideration of evidence and to demonstrate genuine commitment to cultural pluralism. The cultural defense would also guarantee more equal treatment for cultural minorities in a legal system whose standards are based on the values of the dominant culture. Advocates contend that the cultural defense would result in condign punishment for ethnic minorities.

Opponents worry that if culturally motivated offenses go unpunished, this will undermine the deterrent function of the law. They often contend that anarchy will reign supreme if each individual can decide with which particular laws he or she will comply. The defense is criticized for giving special rights to minorities. This puts the victims of culturally motivated crimes at a disadvantage, and is therefore unjust.

The presentation of cultural evidence does not commit the court to accepting the argument. The evidence might not establish the existence of the tradition, that the ethnic group had this tradition, or that the defendant was influenced by the tradition when violating the law. The defense merely ensures that the information is treated as admissible in court. To the extent that cultural rights are guaranteed as part of human rights, the argument is that individuals should be entitled, at the very least, to present evidence about their cultural background in a court of law.

Much of the commentary about the cultural defense emphasizes the tendency of courts to accept oversimplified claims as though they truly represented the communities in question. What analysts fail to point out, however, is that the poor administration of the defense should lead to greater care with its use and not to discarding it altogether.

In the twenty-first century, individuals in Europe and North America raised cultural defenses in drug cases. Rastafarians argued they needed to use ganja (marijuana) in religious ceremonies while individuals from Kenya, Somalia, and Yemen were prosecuted for chewing khat (leaves that have a stimulant effect), a substance widely used in their countries of origin. Although accepted in social life elsewhere, khat was criminalized in Europe and North

America, despite the lack of documented harm, with the United Kingdom banning it in 2014 (The Misuse of Drugs Act (1971) (Amendment) Order, 2014). In Europe there has been litigation concerning whether Article 9, the religious freedom provision of the European Convention permits an exemption from drug laws (Gibson, 2010). Some decisions involved the consumption of hoasca tea as in *Gonzalez v. O Centro Espirita Beneficente Uniao do Vegetal* (2006), the US Supreme Court decision in which the conservative court carved out an exemption for religious minorities.

Another question is whether governments should allow exemptions from noise laws for loud religious sounds (Renteln, 2014c). In some cities the Islamic call to prayer sparks conflict partly because the new sounds are unfamiliar, unwanted, and therefore treated as 'noise.' When the dominant group comes to accept the sound, whether church bells or the call to prayer, the noise ceases to be interpreted as a nuisance and fades into the soundscape.

Minorities also assert their right to erect buildings based on their traditional use of space. For example, numerous controversies surround mosques with one of the most famous being the conflict in Switzerland over the minarets. Differing worldviews clash in local arena over requests for permits which planning authorities may decline to approve.

In many countries we also find cultural concern about which surgeries are consistent with human rights (Renteln, 2014; Sabatello, 2009). For instance, in Europe and North America there have been attempts to stop female genital cutting, also known as 'female genital mutilation' (FGM) by prosecuting those who cut young girls in their new homelands or who take them abroad to have the surgery performed elsewhere. Because family members are reluctant to testify against relatives, prosecutors have had difficulty winning convictions.

In 2014 the UK Home Secretary prosecuted a doctor for performing an episiotomy on a woman after childbirth. Because the statutes appeared to ban any genital cutting, they seem to apply, even though the FGM had occurred in the country of origin decades earlier. This would also mean that doctors seeking to help women by performing reconstructive surgery would be prevented form doing so by badly drafted legislation. Consequently, women encountered difficulty when they sought re-virgination (reconstructive) surgeries because they feared honor-related violence on their wedding nights. This resulted in the adoption of policies by professional medical associations advising surgeons not to perform hymenoplasties in these circumstances; to do so would violate women's rights.

When immigrant parents use corporal punishment to discipline children who misbehave, they also may wind up in court raising a cultural defense (Renteln, 2010). Although many different types of physical discipline are used worldwide, the committee that enforces the Convention on the Rights of the Child issued a policy, General Comment 8, condemning physical discipline in no uncertain terms. Where the United Nations rejects a specific tradition, it may be more difficult for defendants to invoke a cultural defense.

The cultural defense has attracted attention in legal systems across the globe with seminal works by Fabio Basile (2010) and Ilenia Ruggiu in Italy (2013), Pieter Carstens (2009), Tom Bennett (2010), and Christa Rautenbach and Jacques Mattee (2010) in South Africa, and Tamar Tomer-Fishman in Israel (2010). Comparative analysis of the juridical treatment of culture shows that legal systems differ in the extent to which they accommodate expert witnesses who can testify about specific cultural communities and practices. The difficulties associated with evaluating 'cultural' claims raises questions about the criteria by which 'authenticity' of traditions can ever really be determined. Inasmuch as traditions vary and exist in multiple forms, courts have trouble knowing when to accept arguments based on culture (Renteln, 2005).

Scholarship on Cultural Rights

Early scholarship on cultural rights was sponsored by UNESCO through symposia designed to clarify the meaning of culture and cultural rights. In 1950, UNESCO published *Freedom and Culture* which was its first attempt to develop the idea of cultural rights. In 1970, UNESCO sponsored a meeting of experts on cultural rights in Paris to discuss the government responsibilities for sacred places, monuments of cultural significance, and language preservation programs (UNESCO, 1970). As a great deal of emphasis was placed on cultural property, the meeting dealt with the relationship between individuals and cultural objects.

The Interdisciplinary Institute of Ethics and Human Right at the University of Fribourg, Switzerland held a colloquium on cultural rights in 1991. The Council of Europe held the Eighth International Colloquy on the European Convention on Human Rights in Budapest in 1995 on the theme 'Cultural rights: universal, individual, and legally enforceable individual rights.' In 1998 the International Conference on Cultural Policies of Development met in Stockholm, Sweden to evaluate governmental programs designed to ensure that individuals in all communities can exercise the rights guaranteed by Article 27.

In academic circles two parallel scholarly literature emerged on the subject of cultural rights, one in political theory, and the other in international law. Political theorists in the Western tradition analyzed the question of whether cultural rights are compatible with liberalism. For the most part political philosophers worry that the acceptance of cultural rights will lead to 'balkanization,' thereby undermining the unity of the state. The leading proponent of the view that cultural rights can be reconciled with liberalism is Will Kymlicka, a distinguished Canadian scholar. His writings underscore the importance that belonging to a community represents to most individuals.

Despite the fact that his work is generally considered to offer the staunchest support for cultural rights, his defense of cultural rights has been criticized for being too limited. First, Kymlicka advocates the protection of cultural rights only in societies which are said to be structured along liberal lines. His original theory does not provide cultural rights for immigrants who have chosen to move and thereby give up their culture; only minorities long residing in a state and indigenous peoples can legitimately make claims to cultural rights. His interpretation of cultural rights concentrates

mainly on language rights and representation rights. Critics such as Chandran Kukathas have noted that his theory provides minimal protection for cultural rights.

Other leading theorists such as Bhikhu Parekh have attempted to create theories that afford greater protection to cultural rights. In his elaborate scheme advocating 'a principle of dialogical consensus,' Parekh proposes having a minority spokesman, who will engage in a dialogue with representatives of the majority, about cultural practices that clash with 'operative societal values.' It is the duty of the spokesperson to explain how the tradition is authoritative, central to the way of life of the ethnic minority group, and in general, desirable. If it is not feasible to persuade the majority representatives of the value of the tradition under debate and it offends an 'operative societal value,' the ethnic minority must give it up (Parekh, 1996).

Many feminist theorists reject the proposition that cultural rights deserve legal protection because of their concern that protecting these rights will undermine gender equality. Although their work is often framed as a critique of multiculturalism, for example, Susan Okin and Anne Phillips (Phillips, 2009), these analysts are essentially denying the validity of the notion of cultural rights and attacking 'reified' understandings of culture. From the perspective of critical race feminism and postcolonial theory, the use of cultural defense risks 'racializing' disputes by reinforcing false, negative stereotypes about particular groups. Even if some cultural claims are legitimate, they worry that judges are ill equipped to distinguish between valid and illegitimate claims. Generally speaking, most of the commentary on cultural rights in Western political theory is concerned with ways to limit the scope of application of the rights (Kymlicka, 1992; Kukathas, 1992a,b). When cultures do not imprison those who seek to exit, particularly women wishing to flee from oppressive customs, there appear to be fewer objections to the protection of cultural rights. Their objections do not appear to apply to the numerous traditions that have no connection to gender.

In the field of international law, James Anaya, professor of law and special rapporteur on the rights of indigenous peoples has provided a comprehensive analysis of the cultural rights of indigenous peoples and Elsa Stamatopolou a lucid overview of cultural rights jurisprudence within the UN system (Stamatapolou, 2007). A more general theoretical treatment is that of the late Sebastian Poulter whose brilliant scholarship is extraordinarily important (1986, 1998). In Poulter's view, a cultural tradition that violates human rights should not be permitted. According to Poulter, however, if a government, by prohibiting a tradition, is violating human rights, then that tradition should be permitted. His framework has been criticized for presuming the validity of international human rights standards as a basis for determining the validity of traditions and for evading the question of when cultural rights, which are human rights, are superseded by other rights.

Formal Guarantees in Legal Systems

Few national legal systems officially guarantee a right to culture. The extent to which cultural rights are protected depends on whether or not general principles of law such as religious freedom, equal protection, and freedom of association are interpreted to encompass cultural traditions. Because lawmakers are elites from the dominant culture, they tend not to construe abstract principles in such a way as to benefit ethnic minorities. One sign of the preference for monoculturalism is the widespread preference for single official language policies in many countries (de Varennes, 1996; Albanese, 1992).

From time to time, minorities challenge public policies that they regard as unwarranted intrusions into their way of life (Tahzib, 1996). For example, in France the Muslim community objected strenuously to the ban against wearing the *hijab* or headscarf in French public schools (Poulter, 1997). The national controversy was known as *L'Affaire des Foulards* (Gaspard, 1995; Bowman, 2006). The policy was justified because it ensured the separation of church and state, the integration of foreigners into French society, and gender equality. Some women's groups in France thought the Muslim religious garb, the headscarf, symbolized the subordination of women. Eventually, after litigation, the Education Minister relented, allowing the children to wear their religious attire to school.

In the aftermath of L'affaire des foulards, innumerable conflicts occurred over religious garb throughout Western Europe. A vast literature reflects the extent to which visual symbols accentuate differences between groups. In litigation judges in both national courts and the European Court of Human Rights generally accept the state interest in regulating the dress, which is often expressed in terms of protecting national security. The trend toward banning the burqa after 9/11 in Europe and North America sparked heated debate about whether the dress regulations reflected Islamophobia (Bakht, 2012). Maleiha Malik's incisive analysis of these disputes highlights the underlying policy of persecution they support (2014).

In the United States and Canada indigenous peoples have challenged development projects that would undermine their way of life (Eisenberg, 2013). In Lyng v. Northwest Cemetry Protective Association, 1987, tribes argued that the construction of a highway through Chimney Rock would desecrate sacred sites. As part of an environment impact study, the government hired an expert who advised against building the road because of its devastating effect on the tribes. Despite the admonition, the government proceeded with the project, undeterred. The US Supreme Court ruled that the burden on the religion was incidental, not the purpose of the development, and that religious minorities have to accept some burdens.

Difficulties in domestic litigation generally have to do with the question of what counts as the 'authentic' culture or whether a specific practice is 'traditional.' In order to determine the authenticity of cultural traditions, there is the related issue of who is the legitimate representative of the group. Sometimes, indigenous people find offensive the practice of consulting 'experts' or scholars rather than leaders of their communities. Even when the groups themselves are the source of information, it is not always obvious which person from within the group speaks for the group.

Two examples show the practical difficulties in litigating cultural rights claims.

In France the government made an exception to its animal slaughter law to permit Orthodox Jews to slaughter animals in accordance with religious law so that they could consume the meat. France empowered the Central Rabbinical Committee to handle the implementation of the policy. However, one group of Orthodox Jews, concerned that the method employed for the ritual slaughter of animals was not guaranteeing that the meat was sufficiently pure or *glatt*, sued France, challenging its delegation of authority to the Central Rabbinical Committee. After French courts ruled against the group, it appealed to the European Court of Human Rights, where it lost as well.

The Hindmarsh Bridge Island litigation in Australia also shows the dynamics of intragroup tension (Weiner, 1999; Tonkinson, 1997; Lyndall, 1996). A group of aboriginal women objected to a proposed bridge between the Australian mainland and Hindmarsh Island because they claimed it would jeopardize their way of life. The island was ostensibly used for secret women's business, and their customary law supposedly prevented them from divulging information about the nature of the business. During the litigation another group of aboriginal women held a press conference denying the veracity of this claim, saying there was no such 'secret women's business.'

In contrast to domestic litigation that relies on abstract principles interpreted to apply to culture, in international law explicit protections for culture can be found. Among them are the International Covenant on Economic, Social, and Cultural Rights, particularly Article 15, the Declaration on the Rights of Persons Belonging to National or Ethnic, Religious, and Linguistic Minorities (1992), the Framework Convention for the Protection of National Minorities (1995), and the European Charter for Regional or Minority Languages (1992). In addition, there are related documents such as the Declaration on the Rights of Indigenous Peoples finally adopted in 2007 and the Declaration on the Elimination of All Forms of Intolerance and of Discrimination Based on Religion or Belief (1981). Cultural rights are covered not only by human rights treaties, but also by many instruments concerned with safeguarding cultural heritage such as Convention Concerning the Protection of the World Cultural and Natural Heritage (1972), the Convention on Underwater Cultural Heritage (2001), and the Convention on the Safeguarding of Intangible Cultural Heritage (2003).

In the twentieth century the most important cultural rights provision was found in the International Covenant on Civil and Political Rights (ICCPR), a treaty ratified by virtually all countries in the world. This provision is not subject to any reservation (except for France's), that enables a government to avoid obligations under the article. Article 27 provides:

> In those states in which ethnic, religious, or linguistic minorities exist, persons belonging to such minorities shall not be denied the right, in community with the other members of their group, to enjoy their own culture, to profess and practice their own religion, or to use their own language.

However, Article 27 cannot serve as a guide to determine the substantive meaning of cultural rights under international law because it has been interpreted in only a few situations. The provision has been applied mainly to protect sacred land, the right to hunt animals, and language rights.

Although it is widely referred to as a 'minority' rights provision, it is often invoked in cases involving indigenous peoples, many of whom do not consider themselves to be minorities, subordinate to a state. Some members of indigenous groups object to its application in their cases. Article 27 has also been criticized for being formulated as an individual right and as a negative right. These criticisms may be unwarranted: first, the provision can be invoked by individuals or groups; second, if communities only need to invoke a juridical right to culture when governments threaten to halt their practices, it makes sense that the provision was expressed in negative terms: 'governments shall not deny ….' Moreover, because there was some question about the nature of government obligations for the protection of cultural rights, the Human Rights Committee, which enforces the ICCPR, issued a policy statement interpreting Article 27, known as General Comment 23(5), stipulating that governments take affirmative steps to enforce cultural rights.

One of the best-known disputes in international law in which cultural rights were invoked was the celebrated Lovelace decision that was based on Article 27 of the ICCPR. Sandra Lovelace was a Maliseet Indian who married outside the tribe, thereby losing her tribal status (if a male Maliseet Indian married someone not belonging to the tribe, he would not lose his status). After her divorce, she was unable to rejoin the band as a full-fledged member. She sued unsuccessfully in the Canadian courts and subsequently filed a complaint with the Human Rights Committee. Although she challenged the sexist membership rule, alleging that it violated human rights law and the tribe defended itself by invoking the right to self-determination (Article 1 of the ICCPR), the Human Rights Committee decided that she had been denied the right to enjoy her culture in her community. Interestingly, this case stands for the proposition that women also have a right to culture!

The first human rights treaty of the twenty-first century is the Convention on the Rights of Persons with Disabilities, adopted in 2007, and ratified by virtually all states. Article 30 provides a comprehensive elaborate cultural rights provision. The interpretation of the cultural rights of persons with disabilities includes the right to participate in meaningful ways in the arts, education, and science and raises questions about what types of disability culture deserve legal protection as well.

In the final analysis, the main challenge in international human rights jurisprudence is how to evaluate the relative importance of cultural rights vis-à-vis other human rights. This is known as the hierarchy problem. There are rights conflicts between treaty regimes, for instance between Article 27 and provisions of the Convention on the Elimination of All Forms of Discrimination Against Women which require governments to abolish all traditions that interfere with gender equality. The UN appointed Mrs Halima Embarek Warzazi in 1988 as Special Rapporteur on Traditional Practices Affecting the Health of Women and Children. Her studies recommend the elimination of traditions with adverse consequences for women and children including female

genital cutting. For the most part, fundamental rights conflicts within the UN system have not been resolved to date, and the debate centers on what constitutes a harmful traditional practice. In 2009 the UN appointed a special rapporteur on cultural rights, Ms. Farida Shaheed, to identify best practices for the promotion of these rights and complete a series of reports.

Despite modest gains in recognition of cultural rights in domestic and international jurisprudence, this legal protection remained in a precarious position (Renteln, 2014). As the wider public became aware that judges were engaged in application of customary law and the consideration of cultural traditions, a backlash emerged. In some states, for example, voters in the United States passed state referenda, for instance, in Oklahoma banning any consideration of Islamic law. When these measures were subsequently subject to legal challenges, courts sometimes struck them down as unconstitutional (Ceasar, 2012). Not only in violation of fundamental principles of law and international human rights standards, the ban on the consideration of religious law would make it virtually impossible for judges to resolve some legal questions.

Another way of addressing cultural rights claims has been delegating them to religious arbitral boards. Attempts to achieve formal recognition of legal pluralism generated controversy in both England and Canada (Bakht, 2006). Establishing institutional mechanisms to settle particular family law claims in accordance with customary or religious law may facilitate dispute settlements in a manner consistent with national and international law. Yet, because of fear that subnational bodies will fail to adhere to legal principles enshrined in national legal systems, there has been a reluctance to move in the direction of formally authorizing religious arbitration boards.

Theoretical Challenges for the Future

Regardless of whether or not domestic and international legal systems afford adequate protection to cultural rights, the right to culture may be important because abstract principles require it. Appeal can be made to many principles to justify the recognition of cultural rights; among these are identity, freedom, and tolerance. Cultural considerations are likely to influence the disposition of all manner of disputes including criminal and civil litigation as well as a range of conflicts occurring within government agencies and privatized dispute resolution mechanisms.

If culture is intrinsically linked to the formation of identity, then any general notion of individual or group rights requires that culture be protected, at least to some degree. To the extent that democratic systems are committed to the principle of noninterference by the government, intervention in the folkways of cultural communities in the absence of any serious threat would seem illegitimate. The notion of freedom would appear to compel the acceptance of cultural rights in many spheres. Finally, the idea of tolerance is widely considered an important one in liberal democracies. Although tolerance has somewhat pejorative connotations, that is, 'putting up' with practices with which one disagrees, allowing the flourishing of diverse cultural traditions is consonant with this principle.

Even if the right to culture is established as a fundamental right, it remains to be seen how much importance it has vis-à-vis other human rights. The hierarchy question is likely to occupy the attention of scholars interested in the nature and scope of cultural rights. Fortunately, many scholars across the globe have turned their attention to these pressing cultural concerns.

See also: Civil Liberties and Human Rights; Hegemony and Cultural Resistance; Human Rights, Anthropology of; Human Rights, History of; Indigenous Rights; Integration and Multiculturalism: Focus on Western Europe; Rights: Legal Aspects; Transnational Citizenship.

Bibliography

Addis, A., 1991. Individualism, communitarianism, and the rights of ethnic minorities. Notre Dame Law Review 67, 615–676.
Albanese, F., 1992. Ethnic and linguistic minorities in Europe. In: Yearbook of European Law, pp. 313–338.
Anaya, J., 2004. Indigenous Peoples in International Law, second ed. Oxford University Press, New York.
Bakht, N., 2012. Veiled objections: facing public opposition to the Niqab. In: Beaman, L. (Ed.), Defining Reasonable Accommodation. University of British Columbia Press, Vancouver, pp. 70–108.
Bakht, N., 2006. Were muslim Barbarians really knocking on the gates of Ontario? The religious arbitration controversy-another perspective. Ottawa Law Review 69–82 (60th anniversary issue).
Basile, F., 2010. Immigrazione e reati culturalmente motivati. Il diritto penale nelle società. multiculturali, Milano.
Bennett, T.W., 2010. The cultural defence and the custom of Thwala in south African law. Botswana Law Journal 10, 3–15.
Bowman, J., 2006. Why the French Don't like Headscarves. Princeton University Press.
Brumann, C., February 1999. Writing for culture: why a successful concept should not be discarded. Current Anthropology 40 (S1), S1–S27. Special Issue Culture—A Second Chance?
Canadian Commission for Unesco, 1977. A working definition of 'culture'. Cultures 4 (4), 78–83.
Carsten, P., 2009. Cultural defence in criminal law: south African perspectives. In: Foblets, Renteln (Eds.), Multicultural Jurisprudence, pp. 175–195.
Ceasar, Stephen, 11 January 2012. Ruling Upheld Blocking State's Islamic Law Ban. Los Angeles, Times p. A5.
Cha'are Shalom Ve Tsedek vs. France, June 27, 2000. Cha'are Shalom Ve Tsedek vs. France. European Court of Human Rights. Application No. 27417/95. http://www.echr.coe.int/eng/Judgments.htm.
Das, V., 1994. Cultural rights and the definition of community. In: Mendelsohn, O., Baxi, U. (Eds.), The Rights of Subordinated Peoples. Oxford University Press, Delhi, pp. 117–158.
Eisenberg, A., June 2013. Indigenous cultural rights and identity politics in Canada. Review of Constitutional Studies 18 (1), 89.
Foblets, Marie-Claire, Renteln, Alison Dundes, 2009. Multicultural Jurisprudence. Hart, Oxford.
Francioni, F., Scheinin, M., 2008. Cultural Human Rights. Martinus Nijhoff, Leiden.
Gaspard, F., 1995. Le Foulard et la Republique. Editions la Decouverte, Paris.
Gibson, M., 2010. Rastafari and cannabis: framing a criminal law exemption. Ecclesiastical Law Review 1–15.
Gonzalez, v, 2006. O Centro Espirita Beneficente Uniao Do Vegetal, 546 U.S. 418.
Goulet, D., 1981. In defense of cultural rights: technology, tradition, and conflicting models of rationality. Human Rights Quarterly 3, 1–19.
Human Rights Committee, April 26, 1994. general comment No. 23 (5) (art. 27) United Nations, Human Rights Committee, Geneva, CCPR/C/21/Rev.1/Add.5.
James, N., Robert, P., Alison, D.R., 2010. Cultural Law: International, National, and Indigenous (co-authored with James Nafziger and Robert Paterson). Cambridge University Press, New York.

Kukathas, C., 1992a. Are there any cultural rights? Political Theory 20, 105–139.
Kukathas, C., 1992b. Cultural rights again. A rejoinder to Kymlicka. Political Theory 20, 674–680.
Kymlicka, W., 1992. The rights of minority cultures. Reply to Kukathas. Political Theory 20, 140–146.
Kymlicka, W., 1995. Multicultural Citizenship: A Liberal Theory of Minority Rights. Clarendon Press, Oxford, UK.
Levy, J.T., 1997. Classifying cultural rights. In: Shapiro, I., Kymlicka, W. (Eds.), Ethnicity and Groups Rights. New York University Press, New York and London, pp. 22–66.
Lyndall, R. (Ed.), 1996. Secret Women's Business: The Hindmarsh Island Affair. Journal of Australian Studies, vol. 48 (Special Issue).
Lyng v. Northwest Cematry Protective Association, 1987. Lyng v. Northwest Cematry Protective Association 1987 485 US 439.
Malik, M., 2014. The return of a persecuting society? Criminalizing facial veils in europe. In: Brems, E. (Ed.), The Experience of Face Veil Wearers in Europe and the Law. Cambridge University Press.
McGoldrick, D., 1991. Canadian Indians, cultural rights and the human rights committee. International and Comparative Law Quarterly 40, 658–669.
The Misuse of Drugs Act 1971 (Amendment) Order, 2014. http://www.legislation.gov.uk/uksi/2014/1352/contents/made.
Meyer-Bisch, P. (Ed.), 1993. Les droit culturels–une categorie sou-developpee de droits de l'homme. Editions Universitaires, Fribourg.
Nafziger, J., Paterson, R., Renteln, A.D., 2010. Cultural Law: International, National, and Indigenous. Cambridge University Press, New York.
Niec, H. (Ed.), 1999. Cultural Rights and Wrongs. UNESCO Publishing, Paris. Institute of Art and Law, London.
O'Brien, S., 1987. Cultural rights in the United States: a conflict of values. Law and Inequality 5, 268–358.
Parekh, B., 1996. Minority practices and principles of toleration. International Migration Review 30, 251–284.
Phillips, Anne, 2009. Multiculturalism without Culture. Princeton University Press.
Poulter, S., 1986. English Law and Ethnic Minority Customs. Butterworths, London.
Poulter, S., 1987. Ethnic minority customs, English law and human rights. The International and Comparative Law Quarterly 36, 589–615.
Poulter, S., 1997. Muslim headscarves in school: contrasting legal approaches in England and France. Oxford Journal of Legal Studies 17 (1), 43–74.
Poulter, S., 1998. Ethnicity, Law and Human Rights: The English Experience. Clarendon Press, Oxford, UK.
Prott, L.V., 1988. Cultural rights as peoples' rights in international law. In: Crawford, J. (Ed.), The Rights of People. Clarendon Press, Oxford, UK, pp. 92–106.
Rautenbach, Christina, Matthee, Jacques, 2010. South Africa: common law crimes and indigenous customs. Journal of Legal Pluralism 61, 115–144.
Renteln, A.D., 2004. The Cultural Defense. Oxford University Press, New York.
Renteln, A.D., 2005. The use and abuse of the cultural defense. Canadian Journal of Law and Society 20 (1), 47–67.
Renteln, A.D., 2009a. The influence of culture on the determination of damages: how cultural relativism affects the analysis of Trauma. In: Grillo, R., et al. (Eds.), Legal Practice and Cultural Diversity. Ashgate, Aldershot, UK, pp. 199–218.
Renteln, A.D., 2009b. When westerners run afoul of the law in other countries. judicature 92, 238–242. (Special issue on cross-cultural justice).
Renteln, A.D., Spring 2010. Making room for culture in court. The Judges' Journal 49 (2), 7–15.
Renteln, A.D., Spring 2010a. Corporal punishment and the cultural defense. Law and Contemporary Problems 73, 253–279 (special issue).
Renteln, A.D., 2010. Raising cultural defenses. In: Ramirez, L.F. (Ed.), Cultural Issues in Criminal Defense, third ed. Juris Publishing, Inc., New York (second ed., 2007; first ed., 2000).
Renteln, A.D., 2014a. Cutting edge debates: a cross-cultural consideration of surgery. In: Teays, Wanda, Gordon, John-Stewart, Renteln, Alison Dundes (Eds.), Global Perspectives on Bioethics.
Renteln, A.D., 2014b. What do we have to fear from the cultural defense? In: Kymlicka, Will, Lernestedt, Claes, Matravers, Matt (Eds.), Criminal Law and Cultural Diversity. Oxford University Press.
Renteln, A.D., 2014c. The tension between religious Freedom and noise law: the call to prayer in a multicultural society. In: Dagan, Hanoch, et al. (Eds.), Human Rights in Religious Discourse. Israel Democracy Institute, Jerusalem, pp. 375–411.
Ruggiu, I., 2013. Il Giudice Antropologo. Milano.
Sabatello, Maya, 2009. Children's Bioethics: The International Biopolitical Discourse on Harmful Traditional Practices and the Right of the Child to Cultural Identity. Martinus Nijhoff, Leiden.
Shachar, A., 2001. Multicultural Jurisdictions: Cultural Differences and Women's Rights. Cambridge University Press.
Stamatapolou, Elsa, 2007. Cultural Rights in International Law. Martinus Nijhoff.
Stavenhagen, R., 1995. Cultural rights and universal human rights. In: Eide, A., Krause, C., Rosas, A. (Eds.), Economic, Social and Cultural Rights. Martinus Nijhoff Publishers, Dordrecht, The Netherlands, pp. 63–77.
Symonides, J., 2000. Cultural rights. In: Symonides, J. (Ed.), Human Rights: Concepts and Standards, pp. 175–227.
Szabo, I., 1974. Cultural Rights. Leiden. A.W. Nijhoff, The Netherlands.
Tahzib, B.G., 1996. Freedom of Religion or Belief: Ensuring Effective International Legal Protection (Kluwer Law International 1996). Martinus Nijhoff, The Hague.
Tomer-Fishman, T., 2010. "Cultural defense," "cultural offense," or no culture at all? An empirical examination of Israeli judicial decision in cultural conflict criminal cases and of the factors affecting them. Journal of Criminal Law and Criminology 100 (2), 475–522.
Tonkinson, R., 1997. Anthropology and aboriginal tradition: the Hindmarsh Island bridge affair and the politics of interpretation. Oceania 68 (1), 1–26.
UN, Committee on the Rights of the Child, March 2, 2007. General Comment 8 the Right of the Child to Protection from Corporal Punishment and Other Cruel or Degrading Forms of Punishment. CRC/C/GC/8.
UNESCO, 1970. Cultural Rights as Human Rights. UNESCO, Paris.
de Varennes, F., 1996. Language, Minorities and Human Rights. Martinus Nijhoff, The Hague.
Weiner, J.F., 1999. Culture in a sealed envelope: the concealment of Australian aboriginal heritage and tradition in the Hindmarsh Island bridge affair. Journal of the Royal Anthropological Institute 5 (2), 193–210.

Cultural Sociology, History of

Matthias Revers, Department of Sociology, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cultural sociology reaches back to classical thinkers in sociology. In the wake of the 'cultural turn,' cultural sociology constitutes itself drawing on various theoretical resources from the social sciences and humanities: critical theory, cultural anthropology, phenomenology, structuralism, and poststructuralism among others. After its institutionalization in major associations during the 1980s, cultural sociology today represents a vibrant subfield and in some ways a paradigm that permeates other areas of sociological inquiry.

This article looks at cultural sociology in two ways: as a history of thought and process of institutionalization. Cultural sociology will be traced from its early precursors in classical social theory, to its more immediate but still preconscious theoretical foundations of the twentieth century and to the differentiated field it has become since the late 1980s. It discusses some of its central structural characteristics and the final section focuses on how cultural sociology was established as a subfield and explains why this occurred in the 1980s.

Classical Precursors of Cultural Sociology

Karl Marx is certainly not on the forefront of our minds when we think about precursors of cultural sociology. Particularly his early writings, however, were influential for later cultural theorists, most notably Antonio Gramsci, the Frankfurt School, and British cultural studies (subsumed as Cultural Marxists). With the younger Marx (especially *German Ideology* and *The Economic and Philosophical Manuscripts*), these later thinkers share a concern for demystifying ideas (or: ideologies) as expressions and veneers of dominant interests. In other words, this understanding led to analyses of culture as a means of domination.

Emile Durkheim, particularly in his later, less positivist incarnation in *Elementary Forms of Religious Life*, was more directly influential for one particular strand of contemporary cultural sociologists who adopted a structuralist theoretical framework. But even earlier than that, one reading of Durkheim's idea of solidarity in *Division of Labor in Society* suggests that society is the collective conscience of shared moral commitments. More central to recent cultural sociological thinking is his theory of religion in *Elementary Forms*, which is defined by ideas about sacred and profane. Profane refers to the everyday whereas the sacred means matters connoted with awe or fear. The sacred is separated from the profane by taboos and draws power from rituals. Rituals, in turn, create solidarity and thus integrate societies. Although Durkheim mainly talked about religion, he believed that this moral foundation also applies to secular societies. In his earlier book *Primitive Classification*, which he co-authored with Marcel Mauss, he argued that classification is not only cognitive but most importantly moral and emotional.

Max Weber is important for cultural sociology especially (but not exclusively) for his interpretive method of 'verstehende Soziologie' and his emphasis on value-rational ('wertrational') forms of social action. Furthermore, Weber assigned culture a key role in social change but less for social reproduction and stability. His notion of charismatic authority, for instance, which is based on the collective belief in exceptional qualities and powers of a leader, is transitory and relevant for changes of power structures. Although he spent almost the entire book *The Protestant Ethic and the Spirit of Capitalism* on arguing how specific ideas of salvation in Protestantism fostered a capitalist ethos, in essence, religion becomes irrelevant once capitalism is in place and organized by rational systems of authority. Weber's theory of stratification, which distinguished material/economic *class* from symbolic *status*, would become central for later cultural sociologists, particularly Pierre Bourdieu. Of the classical thinkers Weber was the greatest inspiration for multidimensional explanations of historical conditions, accounting for material, social, and ideal factors.

Georg Simmel is certainly the 'most cultural' founding father of the discipline with his detailed and diverse analyses of modern life. However, for lack of a systematic theory he is less important for macro-oriented cultural sociologists than for symbolic interactionists, particularly the Chicago School. Another important precursor, especially for poststructuralist thinkers like Michel Foucault and Gilles Deleuze, was Friedrich Nietzsche; especially, his emphasis on cultural differentiation and cultural relativism as well as an early theory of embodiment.

Of course, none of these thinkers were 'cultural sociologists,' nor did they refer to cultural sociology as an intellectual endeavor. The first sociologist to use the term (i.e., 'Kultursoziologie') was Max Weber's younger brother, Alfred Weber. In contemporary cultural sociology beyond the German-speaking world he is irrelevant, however. Another important thinker for later cultural theorizing in sociology is Talcott Parsons, even though he may be indirectly responsible for a temporary culture-averseness of sociology, at least in the United States. Parson's theory opposed a rational actor model and perceived the cultural system as one subsystem that integrates society. To Parsons, the cultural system was a realm of communicative exchange and coordination of action through stable role expectations that are based on general values. These values and the assumption of a consensus about them were the central

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03059-2

features of his theory and a major point of the critique that followed and culminated in the 1960s. The response of the sociological mainstream in the United States for the better part of two decades had been not to consider culture worth paying attention to at all.

Theoretical Foundations of Cultural Sociology

There are at least six intellectual currents in the twentieth century, which would become important for cultural sociology subsequently. These are (1) a particular reading of Marx (termed Western/Cultural/Humanist/neo-Marxism); (2) (late) Durkheimian social theory; (3) semiotics and French structuralism; (4) poststructuralism; (5) cultural anthropology; and (6) phenomenological sociology:

1. Two scholars were instrumental for promoting Marxist thinking beyond historical materialism and toward a more important role of culture in capitalist domination: Georg Lukács with his work on commodification (and his reconsideration of Marx' *Economic and Philosophical Manuscripts*) and Antonio Gramsci with his theory of hegemony, which emphasized that domination is conditioned by control of ideas and beliefs in society. This Marxian thinking was taken up by the first generation of the Frankfurt School, most notably Theodor W. Adorno, Max Horkheimer, and Walter Benjamin. One of their concerns was the reproduction of popular culture through new technological means and its impact on mass society, which is to impede critical thinking and purvey capitalist ideology. While it played on values of individuality and freedom of choice, popular culture was in fact standardized by capitalist logic, they argued. Louis Althusser established a more cultural reading of Marx' later writings. He focused on how the constituent parts of the superstructure – ideology, state, and the legal system – interact to reproduce capitalism, arguing that they have a 'relative autonomy' from their economic base. Ideas and institutions take shape according to their own logic but the economy defines their boundaries. What unites these theorists is that they assign importance to culture as an object of investigation, though only in terms of ideology. From this viewpoint, culture prevents us from critically assessing circumstances of our lives and traps us in false consciousness.
2. The scholar who directly continued the legacy of Durkheim's later work is his nephew, Marcel Mauss. In his best-known work *The Gift* he argued that gift exchange is central in all societies around the world. Although the exchange often involves commodities, their material worth is secondary to the symbolic value and normative obligations they imply. Maurice Halbwachs would be another Durkheimian scholar of this generation whose work shows that collective memory is a central source of solidarity in societies. In the second half of the twentieth century, the most important intellectual figures advancing Durkheimian cultural theory were cultural anthropologists, particularly Mary Douglas and Victor Turner, as well as sociologist Robert Bellah. Turner offered a theory of cultural change with his concept of liminality, which is a transitional state characterized by a low degree of structuration. As opposed to Durkheim's notion of anomie, the liminal state is defined by the weakening of social distinctions and fostering of creativity. During these moments, Turner argued, communities defined by equality and solidarity are formed (communitas). Mary Douglas was interested in cultural classification and took up Durkheim and Mauss' work on this issue to study symbolic purification and pollution as foundations of social order. What is considered dirty and disgusting may vary across cultures but is not random, she argued. Instead, beliefs about pollution are based on rules of classification and are enforced by rituals of separation of pure and impure. Robert Bellah argued in his famous essay 'Civil Religion in America' that all collective beliefs have a religious dimension, even though they may concern secular issues. Public life, according to Bellah, is structured by religious-like beliefs, which refer not only to Judeo-Christian ideas but national symbols imbued with sacral meaning (e.g., Statute of Liberty in the United States, Brandenburger Tor in Germany, and Marianne in France).
3. Structuralism contends that human experience is rooted in foundational mechanisms, which are patterned by language and call for objective study. Cultural analysis based on structuralism studies culture 'like language.' Human agency, which actualizes language through speech, is irrelevant in such analyses. Structuralism is particularly indebted to structural linguistics of Ferdinand de Saussure. He distinguished *langue* (language) from *parole* (speech), focusing exclusively on the former, and further differentiated between signified (that which is referred to by language) and signifier (that which refers). His central argument is that the relation between signifier and signified is arbitrary but conventional. The task is, therefore, to understand these conventions of signification, which are always structured by opposition. Building on Saussure's structural linguistics, the main reference point for structuralism strictly speaking is Claude Lévi-Strauss. His main objects of study were mythical thought and rituals, which he decoded as cultural classification systems, structured by similarity and difference. Lévi-Strauss' theory of culture thus has resemblance with that of Durkheim, although he hardly referred to him in writing.

   Roland Barthes completed the intellectual move of structuralism to use semiotics for the study of culture. He insisted that the basic linguistic distinctions between langue and parole, signifier and signified, can be applied to all cultural systems. Fixing one often criticized weakness of Lévi-Strauss' theory, which is that it is static, Barthes added a diachronic (he called it *syntagmatic*) element to the analysis of signs, namely by examining narratives. Furthermore, he argued that signifiers do not just denote signifieds but also other signifiers. As opposed to denotation (first order signification), Barthes referred to higher orders of signification as *connotation*. By layering connotative meanings on top of each other, signifiers get a life of their own. To Barthes, signs are not innocuous but part of ideological systems, which he showed most forcefully in *Mythologies*. He stressed that semiotics must be combined with sociological analysis to examine how sign systems, such as myths, help to justify and reproduce social orders. Thus, besides his

many other contributions, Barthes added a critical dimension to structuralist analysis.

4. Partly pioneered by Barthes, poststructuralism is a diverse intellectual movement seeking to refine structuralist thinking. Like structuralism, poststructuralism is concerned with language and symbolic forms, such as narratives, myths, and cultural codes. It is not concerned with the sovereign individual but how subjectivities are affected by semiotic systems under certain historical circumstances (in contrast to structuralism). In opposition to Saussure and continuing Barthes' concept of connotation, poststructuralism rejects the separation between signifier and signified, arguing that signifieds are themselves always signifiers. In contrast to structuralism, poststructuralism is more critical about its own conditions of knowledge production and the notion of the analyst as an objective observer. Related to this, poststructuralism is more sensitive to how cultural systems are affected by power. On the one hand, power is examined as the ability to prise cultural systems open. On the other hand, it asks how power closes cultural orders and makes them appear natural. This is what Foucault's idea of 'dispositif' is concerned with, for example, which subsumes the entanglement of institutional and discursive orders that limit and predetermine what is considered normal and abnormal. A fundamental analytical move of poststructuralism is that it looks at symbolic orders as distinguished from an outside 'Other,' which is necessary to define and sustain its boundaries.

   Michel Foucault made the transition from structuralism to poststructuralism in his own work by increasingly focusing on the power formations generated by expert discourse. In his theory, discourse controls thought and practice and, as in structuralism, human agency is extraneous. Aside from his growing interest in power relationships, his attentiveness to historical processes always distinguished his work from the structuralist tradition. Foucault's analysis of discourse traces the underlying epistemic assumptions and truth claims as well as institutional formations and practices that followed from it. One element that unites his claims about history and power is that there is a transition in modernity from sovereign power (based on hierarchy, physical, public, and ritualized forms of punishment) to disciplinary power (based on surveillance and rationality), which is more profound by altering and controlling behavior, thinking, and bodily functions from within. In contrast to Marxist thinking of ideology, Foucault's theory is free from taken-for-granted assumptions about the basis of discourses and allows more analytic autonomy, which makes it a central reference point for cultural sociology.

   Another central figure in poststructuralism is Jacques Derrida whose work tackled the act of interpretation of texts. In the first instance and in accordance with de Saussure, meanings are always generated through contrast with absent signs ('silences'). Derrida insisted on the multiplicity, contextuality, and endless proliferation of meanings of texts rather than ultimate, fixated meanings asserted by the (structuralist) analyst. Derrida demonstrated that even the most canonical texts can be interpreted in ways that accrue not only alternative but contradictory readings. Texts thus have certain autonomy from their authors' intentions and are always ridden with ambiguity. A search for 'truth' in texts is thus pointless and instead the analyst should engage in deconstruction, that is, in playful hermeneutic interpretation that screens texts from many different angles. Furthermore, the goal of this undertaking is to examine the ways (different) people make sense of texts (in different ways).
5. Besides Mary Douglas and Victor Turner, another cultural anthropologist had significant influence on the cultural turn of the social sciences in general and cultural sociology in particular: Clifford Geertz. Influenced by Max Weber's interpretive method as well as Talcott Parsons emphasis on cultural values and norms, Geertz advanced a hermeneutic approach to culture. During his extensive ethnographic explorations in Indonesia and Morocco, he developed a perspective that considered culture as permeating every aspect of our lives and as worthy of study in its own terms and in an interpretive fashion. The result is a 'thick description,' a densely textured and ethnographic depiction of meanings and experience in local settings, which draws larger conclusions about the culture in question. Geertz, however, shied away from broader theoretical claims beyond the cultural contexts of his studies. Later cultural sociologists would take up this challenge.
6. Phenomenological approaches, such as symbolic interactionism and ethnomethodology, are further influences of cultural sociology. Thinkers associated with these intellectual movements insist on the situational construction of meaning by human agents. For cultural sociologists interested in semiotics, the influence of these works is limited. For others, Alfred Schütz, Peter Berger, and Thomas Luckmann, as also Erving Goffman, Harold Garfinkel, and Harvey Sacks stand as important influences (this is particularly visible in certain strands of science and technology studies, for instance).

Cultural sociologists draw on these intellectual resources with varying degrees. They take from cultural Marxism a concern with critical theory of how culture helps reproduce capitalist orders. They take from Durkheimian thinking a concern with culture as a basis of social cohesion, with rituals as practices to reinvigorate it and a conception of society as a moral accomplishment. They also focus on the binary logic of systems of cultural classification, quite similar to structuralism, which provides more concrete guidelines regarding as to how to go about studying culture. Durkheimian, structuralist and poststructuralist thinking as well as Geertz' cultural anthropology emphasizes culture as an autonomous object of study. Poststructural thinking inspired cultural sociologists to engage in rigorous hermeneutic inquiry and critique of discourses that claim closure and classification (which would become particularly fruitful in queer, gender, ethnic, and postcolonial studies). Phenomenological approaches, finally, have been important to those who focus on cultural practice.

Contemporary Variants of Cultural Sociology

An immediate reference point in cultural sociology is the work of Pierre Bourdieu, which had global impact during the past 20 years of his life and is even more influential now, more than

a decade after his demise. Especially his books *Distinction* and *Rules of Art* as well as the essay collection *The Field of Cultural Production* had a remarkable influence on cultural sociology. His theory provides a comprehensive conceptual repertoire, which lends itself for operationalization in empirical research and examination of culture on different levels of analysis. For cultural sociology, especially helpful are (1) his theory of social stratification and the importance of cultural capital for the attainment and reproduction of social status, (2) his concept of habitus, particularly for ethnography that seeks to make arguments beyond the micro-social and relate structure and agency, and (3) his notion of field as a way to examine the logic of cultural institutions.

British cultural studies are a preceding and parallel intellectual movement, which cultural sociology seeks to distinguish itself from by demonstrative ideological agnosticism and more social scientific empirical methods. Aside from intellectual specificities, cultural studies are defined by a different institutional configuration: it is a more interdisciplinary subfield; it has a higher degree of autonomy (e.g., there are departments of cultural studies); it used to have an organizational center until 2002, namely The Birmingham Centre for Contemporary Cultural Studies (CCCS, founded in 1964). Cultural studies were influenced by cultural Marxism and the New Left, later structuralism and poststructuralism. Important early thinkers were literary critics Raymond Williams and Richard Hoggart (who founded CCCS). The central intellectual figure from the 1970s onward was Stuart Hall. Arguments for the autonomy of culture and interest in power formations promoted by culture are central concerns for cultural studies. Scholars focused on the textual study of political ideology, popular culture, and mass mediation of ideologies as well as ethnographic studies of subcultures. Cultural studies spread from the United Kingdom all over the Anglophone world during the 1980s.

In US cultural sociology, there are two contrary understandings (bracketing 'sociology of culture' for now) of the purpose of culture in sociological analysis, which had wider impact. We may call the first a pragmatist theory of culture and contrast it with a structuralist/semiotic approach.

The paradigmatic formulation of the pragmatist approach to culture is by Ann Swidler who, building on Max Weber, conceptualized culture in an early and often-cited article as a repertoire for constructing strategies of action (also known as the 'tool-kit' approach to culture). In her understanding, symbolic systems are especially important at times of transition and instability ('unsettled times'). Similarly to this conception of culture as a resource, works inspired by Bourdieu treat culture as a means in struggles for power. Works concerned with framing (in David Snow's adaptation of Goffman's original conception of frame analysis) examine social mobilization and claims-making in the public sphere through the strategic construction of meaning (frames). New institutional theory would also belong to this category as it treats culture as a source for legitimacy for organizations, which are competing for market share and state funding. Hence, this approach has wide resonance with works that attribute analytic value but no real explanatory power to culture. One epistemological root for this understanding of culture is a principal division of human experience in social and cultural domains, which is also a common feature of cultural studies.

One key reference point for the structuralist/semiotic approach to cultural sociology is the 'strong program in cultural sociology' (subtitled 'structural hermeneutics') by Jeffrey Alexander, his former students and collaborators. This approach is exemplary for its centrality in the field and programmatic expansiveness. It is institutionally located at the Center for Cultural Sociology at Yale University. Its main theoretical foundations are Clifford Geertz, Max Weber, French structuralism, and late Durkheimian social theory. The strong program is not only a critique of cultural sociology but sociology altogether by arguing that culture is an important constituent of all social relations and that a meaning-centered analysis is thus necessary in all areas of sociological inquiry. The strong program emphasizes cultural autonomy, which requires analytically bracketing out the nonsymbolic, at least in the first instance and before discerning how culture relates to other efficacies. The object of empirical research is typically public discourse, which is subjected to a kind of textual ethnography aiming at thick description of the social realm in question. In accordance with structuralism and Durkheim, this style of cultural sociology emphasizes the binary logic of meaning. It is not static, furthermore, but stresses that meaning is generated sequentially by considering narrative as one key analytical component. Other important theoretical interventions by the strong program were a cultural theory of collective memory (cultural trauma), an approach to study cultural pragmatics (cultural performance) and, most recently, visual culture and iconicity.

The Institutionalization of Cultural Sociology

In the 1980s, the first sections for cultural sociology were founded in disciplinary associations. The cultural sociology section in the German Sociological Society, for instance, was established in 1984 under the name 'Kultursoziologie in der Deutschen Gesellschaft für Soziologie.' At its inception, it was defined by political conservatism and an aspiration to strengthen the German *geisteswissenschaftliche* tradition in sociology. The sociology of culture section in the American Sociological Association followed suit in 1987. Why at that time? The simple answer is that institutional differentiation of the discipline within associations really began only in the 1970s. Besides these associational reasons, differentiation and self-awareness of cultural sociologists occurred in opposition to the sociological mainstream, which was perceived as too scientistic and materialist at the time. In particular, the emergence of cultural sociology was a response to anti-Parsonianism in the sociological mainstream, which prevailed from the late 1960s until the 1980s, not only in the United States but also in Germany. This posture rejected not only the abstract and consensual values envisioned by Parsons but culture in general.

Another important current for the rise of cultural sociology was the 'cultural turn,' which began to take hold in the social sciences in the 1960s. During that turn, disciplines became more susceptible to theorizing in other social scientific fields and especially the humanities. The cultural turn was itself enabled by the 'linguistic turn,' rooted in structural linguistics

of Ferdinand de Saussure and Ludwig Wittgenstein's later philosophy of language, which led scholars in the social sciences and humanities to take language and symbols more seriously.

Besides intellectual developments (but certainly related to them), the historical context for the growing importance of culture as a subject matter of social science was the ascendant postindustrial logic in Western societies of the second half of the twentieth century as well as the counter-cultural and social movements that arose in the 1960s. Although cultural thinking in sociology started much earlier, it seems that it could only really thrive under these historical circumstances, in which we see a rise of 'expressive individualism' (Bellah), an emphasis on consumption and experience and a growing emphasis of identity and symbolic representation in public debate. The perceived dissolution of formerly rigid social categories and a growing sense of ambivalence and contingency necessitated new analytic tools to examine this new condition.

In the United States, the emergent field was structured by internal discussions about the object(s) of study from the beginning. Some sociologists, most notably Howard Becker, Diane Crane, Paul DiMaggio, Wendy Griswold, and Richard Peterson studied culture as an object of production and consumption, which is shaped (and explained!) by other social conditions. This became known as the *sociology of culture* approach. Howard Becker's book *Art Worlds*, for example, argues that art is conditioned by previously reached conventions (and their calculated breach), which shape both the aesthetics of artifacts and the cooperative arrangements necessary for producing them. Many works in this field are influenced by Pierre Bourdieu, focusing on class and status hierarchies in the production and consumption of culture. A well-known exemplar is Richard Peterson's research, which showed that people in privileged positions are better conceived as 'cultural omnivores' than mere snobs, as Bourdieu's homology argument suggests (higher strata exclusively consume high-brow culture). Omnivorousness is not understood as a liberal attitude but a new mode of distinction, which is based on cultural relativism and the ability to criticize a wide spectrum of cultural forms based on knowledge about them.

Opposite sociologists of culture, other scholars considered themselves cultural sociologists. They share a more expansive vision of the field, both in terms of intellectual influences and subject areas of study. They read cultural anthropology, cultural history, cultural studies, and (post)structuralism and seek to make interventions in other areas of sociology from a cultural perspective. Viviana Zelizer is considered as one of the founders of this sociological approach with her early research on the importance of cultural meanings in the insurance business and money economy (published in the 1970s and 1980s). Eviatar Zerubavel published on the temporal structuring of everyday experience as a powerful collective representation around the same time.

The intellectual dynamics of the new subdiscipline in the United States was defined by debates between those two camps, sociology of culture and cultural sociology, which were particularly intense during the 1990s. A central issue of debate was the role of culture in sociological analysis, essentially causality and autonomy of culture. Sociologists of culture treat culture as a dependent variable to be explained by other social or material forces. In cultural sociology, it is the opposite; that is, culture does the explaining or at least a significant part of it.

In Germany, in comparison, the broader understanding of cultural sociology was dominant from the beginning. One reason is that the leading figures, most notably Friedrich Tenbruck and Wolfgang Lipp, set the direction more forcefully in a special issue of the *Kölner Zeitschrift für Soziologie und Sozialpsychologie* where they defined 'Kultursoziologie' as concerned with "general patterns of meaning, which give action stability and significance, explicitly and implicitly, as preconditions and intentions across all areas of life and institutions" (Lipp and Tenbruck, 1979: 395; my translation). Such a defining intervention has not occurred at the inception of the subfield in the United States, or rather: there were multiple interventions, which were subjected to controversy with no immediate resolution. From the 2000s onward, however, cultural sociology became dominant in numbers and directions of the subfield.

Sociology of culture and cultural sociology are now somewhat divided into different journals, *Poetics* preferred by the former, *Cultural Sociology* – an official journal of the British Sociological Association – and the more recent *American Journal of Cultural Sociology* by the latter. Although a cultural studies venue, *Theory, Culture & Society* is also an important outlet for cultural sociological thinking. The field, especially sociology of culture, also enjoys some visibility in flagship journals in US sociology, *American Journal of Sociology*, and *American Sociological Review*. Important theoretical interventions in cultural sociology appeared in *Sociological Theory* and *Theory & Society*. Cultural sociologists have been successful publishing their work with top university presses from the beginning, above all Oxford University Press, Cambridge University Press, and the University of Chicago Press, which have remained important publishing venues for the field. Oxford University Press and Routledge have recently published handbooks on cultural sociology and Blackwell published one on the sociology of culture.

Cultural sociology as it is constituted today is engaged almost in all subfields of sociology, even in subject areas that are most 'material' by definition (economic sociology, criminology, and so on). Disciplinary boundaries are typically rigid in the United States. In Germany, for instance, cultural sociology is embedded in the broad field of 'Kulturwissenschaften,' which takes institutional form in intra- and extra-academic research institutes (similar to cultural studies but less politically oriented). German cultural sociologists are thus subjected to a more interdisciplinary debate to begin with. Cultural sociology in the Anglophone and international context, although it is infused by ideas from other disciplines (as mentioned above), still lacks such branching-out efforts, which could be a direction of the field in the future.

See also: Culture and Actor Network Theory; Culture and Economy; Culture and Institutional Logics; Culture and Networks; Culture, Cognition and Embodiment; Culture, Production of: Prospects for the Twenty-First Century; Embodiment and Culture; Epistemic Cultures; Formal Methods of Cultural Analysis; Organizations and Culture; Symbolic Boundaries.

Bibliography

Alexander, J.C., 2003. The Meanings of Social Life: A Cultural Sociology. Oxford University Press, New York.

Alexander, J.C., Jacobs, R.N., Smith, P., 2012. The Oxford Handbook of Cultural Sociology. Oxford University Press, New York.

Back, L., 2012. Cultural Sociology: An Introduction. Wiley-Blackwell, Chichester, Hoboken.

Crane, D., 1994. The Sociology of Culture: Emerging Theoretical Perspectives. Blackwell, Oxford, Cambridge, MA.

Friedland, R., Mohr, J., 2004. Matters of Culture: Cultural Sociology in Practice. Cambridge University Press, Cambridge, New York.

Hall, J.R., Grindstaff, L., Lo, M.M., 2010. Handbook of Cultural Sociology. Routledge, London, New York.

Jacobs, Mark D., Hanrahan, N.W., 2005. The Blackwell Companion to the Sociology of Culture. Blackwell, Malden.

Lamont, M., Thévenot, L., 2000. Rethinking Comparative Cultural Sociology: Repertoires of Evaluation in France and the United States. Cambridge University Press, Cambridge, New York.

Lamont, M., Wuthnow, R., 1990. Betwixt and between: recent cultural sociology in Europe and the United States. In: Ritzer, G. (Ed.), Frontiers of Social Theory: The New Synthesis. Columbia University Press, New York, pp. 287–315.

Lipp, W., Tenbruck, F.H., 1979. Zum Neubeginn der Kultursoziologie. Kölner Zeitschrift für Soziologie und Sozialpsychologie 31, 393–398.

Moebius, S., 2009. Kultur. Transcript, Bielefeld.

Moebius, S., Papilloud, C., 2012. Einleitung. Trivium 12. http://trivium.revues.org/4377 (accessed 06.11.13.).

Smith, P., 1998. The New American Cultural Sociology. Cambridge University Press, Cambridge.

Smith, P., Riley, A., 2008. Cultural Theory: An Introduction. Wiley-Blackwell, Oxford.

Spillman, L., 2002. Cultural Sociology. Blackwell, Malden.

Swidler, A., 1986. Culture in action: symbols and strategies. American Sociological Review 51 (2), 273–286.

Turner, G., 1996. British Cultural Studies: An Introduction. Routledge, London, New York.

Cultural Studies of Science

Joseph Rouse, Wesleyan University, Middletown, CT, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 5, pp. 3125–3127, © 2001, Elsevier Ltd.

Abstarct

Cultural studies of science, a multidisciplinary field drawing from history, anthropology, feminist theory, sociology, and philosophy of science, treats scientific practices as historically situated, meaningful patterns of interaction with the world. Rejecting many shared assumptions in debates among sociological constructivists and internalist philosophers of science, cultural studies seek interpretive, critical engagement with specific scientific practices rather than global legitimization or critique of science. Their commitment to dynamic, expressive conceptions of meaning, knowledge, and power is primarily articulated through detailed studies of particular practices, performances, or institutions. Their foci include laboratories or field sites; images, models, and metaphors; machines and procedures; and patterns of exchange among scientific work and other sites of meaning-articulation. These studies are situated within the Acultures of science that they investigate, and aim at responsible, mutually transformative interaction with scientific practices rather than explanation of their outcomes.

'Cultural studies of science' marks a loosely amalgamated, multidisciplinary research field drawing from history, anthropology, feminist theory, sociology, and philosophy of science. Its participants emphasize that scientific practices are historically situated, meaningful patterns of interaction with the world. Cultural studies offer an interpretive, critical engagement with scientific practices, rather than an explanation of their outcome. Cultural studies are situated within the 'cultures of science' that they investigate, and are responsive to what is at stake in the ongoing development of scientific practices.

The Place of Cultural Studies within Science Studies

Cultural studies of science first emerged in the late 1980s and early 1990s in response to debates between proponents of internalist history and philosophy of science, and adherents of the 'Strong Program' in the sociology of science. These debates concerned how best to explain the content of scientific knowledge, either by appeal to rational norms or as the outcome of particular social interactions. The stakes in these debates over explanatory strategies seemed to concern the authority of scientific claims, their relation to the natural world, and the normative stance of the explanation itself. With respect to scientific authority, philosophical accounts typically vindicated the acceptance of scientific claims or research programs as rationally warranted or reliable, whereas constructivist sociological work emphasized the contingency of scientific work and the impossibility of rationally foreclosing alternative interpretations. Meanwhile, scientific realists argued that scientific claims were accountable to the world as independent of human language, experience, or social practice, while antirealists (including both sociological constructivists and philosophical empiricists and historical rationalists) insisted that the sciences could never reach beyond our categories, practices, and norms. Finally, classically normative accounts of science aspired to assess the adequacy of scientific practices according to philosophical norms, while philosophical and sociological naturalists proposed primarily to describe (and perhaps explain) how the sciences actually proceed.

Work in cultural studies proceeded differently, in ways that rejected assumptions shared by most participants in these debates between internalists and social constructivists (assumptions that are also implicit in classic works of social history, such as Forman, 1971 or Shapin and Schaffer, 1985, that are otherwise thematically congenial to cultural studies). First, cultural studies are strongly antiessentialist about the sciences: there are no goals, methods, products, or norms belonging to science as such. Such antiessentialism undercuts the explanatory projects of both internalist philosophical histories and of a programmatic social constructivism, which are committed to accounting for the sciences as knowledge-producing activities (where 'knowledge' is a distinctive kind of product, either rationally warranted, causally reliable, or socially authoritative belief). The sciences are historically evolving practices, such that past and present practice cannot fix their nature, significance, aim, or 'content.'

Cultural studies also find no clear boundaries between scientific content and its social context. Both philosophers and sociologists have identified the shared commitments and norms of scientific communities as the locus of their divergent explanatory projects. Cultural studies have instead emphasized both an 'internal' heterogeneity that belies any presumption of community within a scientific field, and an active traffic in meanings and material across any supposed boundaries between scientific communities and their social or cultural contexts. This absence of community-defining consensus also exemplifies a more fundamental issue: cultural studies take the subjectivity of scientific knowers to be part of what needs to be understood rather than as an explanatory resource. Neither the natural body, the rational faculties, or the social identity of a scientist can be taken for granted as the locus or source of agency and belief. Prosthetic embodiments, narrative re-enactments, discursive interactions, and the affordances of material surroundings instead constitute knowing subjects through scientific practices, rather than accounting for practices as the doings of human agents/believers.

The opposing positions in the realism debates thereby dissolve. The question whether human knowers can ever get beyond the confines of their mental or social representations in order to be accountable to natural objects presupposes the

autonomous intelligibility of mental or social representations. Cultural studies take scientific practices to be always already materially engaged and discursively articulated, so that there is no confinement within experience, language, theory, or social world to be transcended. On the contrary, experience, language, theory, and social forms themselves only take on determinate form through ongoing patterns of 'intra-action' in the material world (Barad, 1996 coins this term to avoid the connotation that 'interactions' involve agencies definable apart from their inter-action). But the same is true of 'nature,' whose objects and capacities only acquire determinate articulation in phenomena, ongoing patterns of 'material-semiotic' intra-action (Barad, 1996; Rouse, 1996; Haraway, 1999).

The dispute between normative and naturalistic (descriptive) approaches to scientific practices also turns out to share false presuppositions. Cultural studies agree with naturalists that there can be no standpoint 'outside' of ongoing scientific practices from which to articulate norms to which they are accountable. Yet they disagree that science studies must therefore simply describe (and perhaps explain) the norms that function within those practices. The normative accountability of scientific practices is not a determinate fact or regularity about them (or about scientific 'communities'), but is instead constitutive of their intelligibility to practitioners and interpreters alike.

Scientific practices only make sense in terms of what is at issue and at stake in those practices. Yet these issues and stakes are themselves contested within the practice, and its interpretation. Cultural studies are not neutral descriptions of what scientists do, but interpretive construals of those doings that may thereby partially reshape the field of subsequent practice. In that respect, cultural studies are continuous with scientific work, which likewise reinterprets what is at issue and at stake through re-enacting the practices it participates in (Rouse, 1996).

Cultural studies thereby articulate dynamic, expressive conceptions of meaning, knowledge, and power, which contrast sharply with the standard approaches to these phenomena within philosophy and social theory (Rouse, 1996, 1999). On such accounts, meaning is not a property of utterances or actions; the term 'meaning' instead articulates the ways in which such performances inferentially draw upon and transform the field of prior performances in which they are situated. Knowledge is, likewise, not a theoretically coherent kind of representation; the term 'know' is used expressively in ongoing practices of assessing, attributing, relying upon, and contesting understanding and justification. Power, in turn, is not a substantial component of social interaction, something possessed and exercised by some agents, and lacking for others. The term expresses the ways in which one action aligns with others over time to affect the field of subsequent possible activity: 'power' articulates the connections between causal efficacy and intelligibility. Cultural studies of science thus expressively articulate the semantic, epistemic, and effective roles of scientific practices.

Prominent Themes

The preceding account offers a broadly philosophical interpretation of the aims and commitments embodied in the wide range of historical, anthropological, and sociological investigations that exemplify cultural studies of science. Typically, however, these commitments are expressed through detailed studies of particular practices, performances, or institutions rather than as explicit philosophical theses. A representative list of prominent work in cultural studies of science would include work directed toward particular laboratories, field sites, and other settings (e.g., Haraway, 1989; Heath, 1997; Traweek, 1992; Biagioli, 1993), images, models, and metaphors (e.g., Bono, 1995; Haraway, 1992; Keller, 1995; Martin, 1994; Daston and Park, 1998), and machines and procedures (e.g., Barad, 1998; Galison, 1997; Rheinberger, 1997; Daston and Galison, 1992), all of which involve patterns of exchange between scientific settings and other sites of meaning-articulation. These topics have been marked as productive foci of inquiry not because they are relatively localized or self-contained, but because they have been productive nodes in the circulation and transformation of meaning, knowledge, and power around potent and often contested issues.

'The body' has been among the most prominent of these issues taken up in detailed work by cultural studies scholars, perhaps in part because of their insistence upon the embodied, materially interactive dimension of scientific practices. Attention has been directed toward bodily disciplines and skills (both as constitutive of scientific knowing, and as impositions upon who or what is to be known); toward prosthetic extension or reorganization of bodily engagement with the world (including the immunological internalization of the boundary between bodily self and external others); toward bodies as raced and gendered, and otherwise differentially marked (Heath, 1997, for example, takes up the distinctive bodily markings and significations of Marfan's Syndrome); toward animal bodies, microbial 'bodies,' and their complex material-semiotic relations with human embodiment; and toward various processes and apparatuses of bodily reproduction. Along with 'the body,' other meanings partially constitutive of scientific knowing have also become issues for cultural studies of science. Authorship and authority, objectivity in its multiple contested deployments, the construction of professional and disciplinary identities within and around the sciences, and the affective dimensions of scientific understanding (e.g., wonder, anxiety, love, empathy, and various forms of humor) have been among the epistemically significant issues that have provided a focus for cultural studies of science. Cultural studies have also been attentive to many of the ways in which the objects of scientific understanding have been meaningfully articulated. Genes, organisms, cyborgs, blood, information, life, matter, viruses, origins, secrets, foundations, environment, and ultimately nature in its significant differentiation from culture, society, artifact, and history have been prominent themes in both scientific practices and cultural studies of science.

In rejecting global legitimations or critiques of science as justified knowledge, social consensus, empirically reliable methods, cultural tradition, or approximately true theories, cultural studies of science have attempted to set aside the project of accounting for what science really (essentially) is, or how science is actually done, in favor of participatory engagement with specific scientific practices. The sciences are not regarded as optional practices (or belief-systems), of which it could even make sense to be for or against, credulous or skeptical, as a whole. Scientific practices are integral to the

formation of a field of intelligible possibilities for action and self-understanding, partially constitutive of any conceptualization of who we are and what our situation is. To engage successfully in cultural studies of science is to clarify, and thereby partially transform, that field of possibilities and our involvement within it. Ultimately, the stakes in cultural studies of science are located in the notion of responsibility. To whom, to what, and in which respects are the sciences and science and technology studies responsible? How might such responsibility be exercised, and even partially fulfilled, through scientific research in particular settings, and through cultural studies of specific scientific or technological practices? What criteria or concerns should govern the assessment of responsible and irresponsible science and science studies? With these questions in mind, cultural studies become not simply a specifiable and in-principle exhaustible domain of academic inquiry, but an ongoing project of responsible engagement within the cultures of science.

See also: British Cultural Studies; Cultural History; Cultural Relativism, Anthropology of; Cultural Sociology, History of.

Bibliography

Barad, K., 1996. Meeting the universe halfway. Realism and social constructivism without contradiction. In: Nelson, L.H., Nelson, J. (Eds.), Feminism, Science and Philosophy of Science. Kluwer, Dordrecht, The Netherlands, pp. 161–194.

Barad, K., 1998. Getting real: technoscientific practices and the materialization of reality. Differences 10, 87–126.

Biagioli, M., 1993. Galileo Courtier: The Practice of Science in the Culture of Absolutism. University of Chicago Press, Chicago.

Biagioli, M., 1999. Science Studies Reader. Routledge, New York.

Bono, J., 1995. The Word of God and the Languages of Man. University of Wisconsin Press, Madison, WI.

Daston, L., Park, K., 1998. Wonders and the Order of Nature, 1150–1750. MIT Press, Cambridge, MA.

Daston, L., Galison, P., 1992. The image of objectivity. Representations 40, 81–128.

Forman, P., 1971. Weimar culture, causality and quantum theory 1918–1927. In: McCormmach, R. (Ed.), Historical Studies in the Physical Sciences. University of Pennsylvania Press, Philadelphia.

Galison, P.L., 1997. Image and Logic: A Material Culture of Microphysics. University of Chicago Press, Chicago.

Haraway, D.J., 1989. Primate Visions: Gender, Race, and Nature in the World of Modern Science. Routledge, New York.

Haraway, D.J., 1992. The promises of monsters. In: Grossberg, L., Nelson, N., Treichler, P.A. (Eds.), Cultural Studies. Routledge, New York, pp. 295–337.

Haraway, D.J., 1999. Situated knowledges: The science question in feminism and the privilege of partial perspective. In: Biagioli, M. (Ed.), The Science Studies Reader. Routledge, New York, pp. 72–88.

Heath, D., 1997. Modest interventions. In: Downey, G.L., Dumit, J. (Eds.), Cyborgs and Citadels: Anthropological Interventions in Emerging Sciences and Technologies. School of American Research, Santa Fe, CA, pp. 67–82.

Keller, E.F., 1995. Refiguring Life: Metaphors of Twentieth Century Biology. Columbia University Press, New York.

Latour, B., 1992. One more turn after the social turn. In: McMullin, E. (Ed.), The Social Dimensions of Science. University of Notre Dame Press, Notre Dame, IN, pp. 272–294.

Martin, E., 1994. Flexible Bodies: Tracking Immunity in American Culture from the Days of Polio to the Age of Aids. Beacon Press, Boston.

Rabinow, P., 1996. Essays on the Anthropology of Reason. Princeton University Press, Princeton, NJ.

Rheinberger, H.-J., 1997. Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube. Stanford University Press, Stanford, CA.

Rouse, J., 1996. Engaging Science: How to Understand its Practices Philosophically. Cornell University Press, Ithaca, NY.

Rouse, J., 1999. Understanding scientific practices. In: Biagioli, M. (Ed.), The Science Studies Reader. Routledge, New York, pp. 442–456.

Shapin, S., Schaffer, S., 1985. Leviathan and the Air Pump: Hobbes, Boyle, and the Experimental Life: Including a Translation of Thomas Hobbes, Dialogus Physicus de Natura Aeris by Simon Schaffer. Princeton University Press, Princeton, NJ.

Traweek, S., 1988. Beamtimes and Lifetimes: The World of High Energy Physicists. Harvard University Press, Cambridge, MA.

Traweek, S., 1992. Border crossings. In: Pickering, A. (Ed.), Science as Practice and Culture. University of Chicago Press, Chicago, pp. 429–465.

Cultural Views of Life Phases

M Annette Grove, Utah State University, Logan, UT, USA
David F Lancy, Department of Sociology, Utah State University, Logan, UT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The knowledge base in the study of human development is built primarily from work with children from the modern, global, postindustrial population. This population is unrepresentative in many respects, not least in that childhood and adolescence is dominated by the experience of formal schooling, an experience missing from the lives of most of the world's children until very recently. This entry will examine child development from the perspective of premodern societies as described in the ethnographic, archaeological, and historic records. Specifically, we will review material indicative of cultural or indigenous models of development, phases, and phase transitions, in particular.

Introduction

Our goal is to convey the variability and commonality that exists in culturally constructed views of life phases. When examining these cases it becomes clear that societies differ in their models of development. However, nature provides predictable transitions (e.g., walking, speaking, theory of mind, puberty) upon which most societies build in constructing models of development. We describe these patterns and provide examples that best illustrate the variety and/or similarity found in culturally constructed views of life phases. An anthropological analysis of phases and transitions in children's development takes into account three data sources. First, and most obviously, we consider nomenclature that is used to mark phases or transitions. Second, we take note of particular rites or ceremonies – also known as 'rites of passage' – associated with life-phase transitions. Third, we draw on the ethnographic and historic records for evidence of changes in the behavior of children and their families that signal a shift in the child's status. We utilize these data to construct broadly applicable cultural models of child development. These models coalesce around six phases in the life cycle, which, not coincidentally, correspond to evolutionary biologists' partitioning of the lifespan (Bogin and Smith, 2012: 521 – Lifespan Development: Evolutionary Perspectives) This entry draws on a long-term project designed to develop an anthropological perspective on human development (Lancy, 2007, 2008, 2010, 2012a,b 2013; Lancy and Grove, 2011a,b; Lancy et al., 2010). Our database consists of archival accounts of childhood from nearly 1000 societies, ranging from the Paleolithic to the present and from every area of the world.

Phase Terminology

The Bonerate are representative of the majority of societies that fail to name or identify transitions in the life course. They "cannot define precisely the onset of childhood … they have difficulty describing the boundaries between childhood and adolescence and between adolescence and adulthood" (Broch, 1990:15). When we look at societies that do make such distinctions, we see great variability. At one extreme, the Busulu of E. Kalimantan name 24 distinct phases from *lemaub*, meaning 'to be able to roll over,' to *nysagawan dungai lungud*, meaning 'to carry a small bamboo water tube,' to *timpun sinigod gegalu*, meaning the first appearance of breast buds (Appell-Warren, 2012). By contrast, the Inuit of the Belcher Islands have only three named phases: (1) up to 1 year – *natarak*, (2) 12–18 months – *qitungak*, and (3) from 3 years until they become betrothed they are t*ugusi* (Balikci, 1970: 41).

The basis for named transitions also varies. For example, the Tukano of Brazil take note of physical features in designating phases. They have terms for 12 developmental transitions – birth through puberty. For example, *Soãgõákã* means 'little red one,' noting a change in skin color after birth; *Duhigö* means 'one who sits down'; *Böagö* means 'to creep'; *Sγagö* 'to walk,' and so on (Silva and Lillios, 1962). The Igbo also use terms that note physical maturation: *ino odu* (be–sit) a child who sits at around 3 or 4 months; *igbe igbe* (crawl–crawl) when the child begins to crawl at 7 or 8 months; *iguzo mpe* (stand up) when the child starts to walk; and *ifuteleze* (teething) when the child begins to sprout teeth (Ottenberg, 1989: 20).

The Baining (of New Britain Island) use mode of locomotion as the basis for phase-naming conventions. A newborn baby is carried in an adult's arms or in a sling tied across the chest. In answer to the question 'how old is he (or she)?', a child of this age is described *as ta tal ka (ki)* or 'they carry him (her).' After the age of 5 or 6 months, parents begin to carry their children on their shoulders. This form of transportation requires that the child have some sense of balance and support and take some part in maintaining his or her posture. Children of this age are described as *ka (ki) kalak* or 'he (she) sits on the shoulders.' An older child is identified by the phrase *ka (ki) tit* or 'he (she) goes' which refers to crawling and, then, walking. An older child who has become even more independent (e.g., 7–9 years) is referred to as *ka (ki) tit mas* or 'he (she) goes fully,' meaning that he or she goes for water, firewood, or to gather in the bush (Fajans, 1997: 86–87). In naming practices that focus on physical maturation and basic skills of locomotion and

speech, the first 3 years are partitioned many more times than later periods in the life cycle.

Other societies use terms that call attention to the child's accomplishments: Among the Giriama (Kenya), a two- to three-year-old is labeled, in effect, 'water carrier.' An eight-year-old girl is defined as a 'maize pounder,' a boy of the same age is *muhoho murisa* 'herd boy' (Wenger, 1989: 98). In premodern Russia 'our plowboy,' 'our herd boy,' and 'our nanny girl' were habitual terms parents used to address their children (Gorshkov, 2009: 15). Ages 16–20 years in post-Inkan Peru were collectively designated *cocapalla* or 'coca harvester' because youths of this age were expected to reap the state-owned cocoa crop (Dean, 2002: 43). The Savras (India) recognize five phases in human development and name them according to the primary chores carried out at each age (Mohammad, 1997).

Rites of Passage

Life-phase transitions may also be marked by 'rites of passage.' These celebrations denote the transition between one life phase and the next. Most rites include a *separation* from the previous life phase, a period of *transformation*, followed by a ceremony marking *integration* into the next phase (van Gennep, 1961). As with naming conventions, rites of passage center around milestones in physical maturation or mark achievements or 'firsts.' These rites are rarely tied to precise dates, such as the birthday, because preliterate societies do not keep track of birth dates or the passage of years.

Naming Ceremony

Although birth would seem an obvious candidate for a rite of passage, it is interesting to note just how few societies treat this as worthy of such a celebration. We discuss below how early infancy is characterized, but the main reason that the birth is not seen as cause for celebration is that both mother and baby are in an extremely precarious state. Once it is clear that the infant has a good grasp on life and paternity is acknowledged, the infant is introduced to the community via the naming ceremony, which can occur days, weeks, or even months after birth. During this liminal period, there is a clear sense that the infant is still in the process of becoming a human being (Lancy, 2013). This is conveyed by the timing of the naming practices. For example, in Brazil, Wari babies usually do not receive a personal name until they are about 6 weeks old. Until then, babies of both sexes are called *arawet*, which translates literally as 'still being made' (Conklin and Morgan, 1996: 672). The confirmation of personhood is delayed on Gau Island in Fiji, Melanesia until the infant is 4 days old after which a celebration called the "falling of the umbilical cord (*lutu na nona i vicovico*) occurs" (Toren, 1990: 169). In the Himalayas, the Lepcha of Sikkim, continue to refer to the newborn as a 'rat child' as if it were still inside the mother's womb. After three days and a thorough cleansing, the infant is welcomed into the community with a feast (Gorer, 1967). Among the Masai of East Africa, the first name is given on the fourth day after birth, when the child is brought outside to be shown the sun. The child is named in the presence of the father, mother, and three elders (Huntingford, 1953: 116–17). Among the Azande of north-central Africa, the midwife is not paid and the infant is not named until they are sure that it will survive. When survival seems likely, the whole community takes part in a ceremony in which the infant is removed from the birthing hut and passed through the smoke of a greenwood fire (Baxter, 1953: 72).

First Haircut

The 'first haircut' can occur with the naming ceremony, but usually happens some time after. The first haircut does not necessarily correlate with a particular age as concerns about viability are not easily set aside. Rather, the first haircut seems to commemorate the child's successful survival during a very difficult period. The rite may also be used to signal a weakening of the child's exclusive ties to its mother and the establishment of ties to its father and extended kin network (Fricke, 1994: 133). With the first haircut we also see the beginning of differentiation between the sexes. The Kurds of Rawanduz give the first haircut after the child has survived a year (Masters, 1953: 159). The equivalent Balinese rite occurs at about the same age and until the first haircut the child is not allowed to touch the ground and must be carried at all times (Geertz, 1961: 104). In northwestern Africa, the Bambara of Mali shave only part of the child's head. Some locks are kept in a container, some thrown into the river as a sacrifice, and elders examine the remaining hair to determine the child's *téré* or character (Paques and Turner, 1954: 119). The Navajo see the first haircut as a necessary step on the path to Navajo personhood, attained only after the child speaks its first words demonstrating that it has control over its thought and voice (Schwarz, 1997: 146–47).

Accomplishments

Once the child is free and clear of threats to its existence, the next major milestones are related to puberty and marriage. However, it is not uncommon to make a minor rite of passage out of the child's accomplishments or 'firsts.' These celebrations mark the child's developmental progression as a contributing member of the family. For example the Kaoka of the Solomon Islands allocate small plots of land to their sons and boast of the yams they will grow (Hogbin, 1969: 39). In central Africa a Mbuti boy bringing home his first 'real animal' is immediately proclaimed a hunter and receives scars (cicatrization) incised by one of the 'great hunters' indicating his changed status (Turnbull, 1965: 257). In North America, Hopi girls who have mastered grinding corn are 'shown off' to visitors (Hough, 1915: 63). Among the Saami of northern Europe there are no rites of passage or acknowledgment of any kind for a growing girl until she completes her first pair of reindeer shoes or some other complex needlework (Pelto, 1962). Other examples of the marking of firsts with rites of passage are a Netsilik girl's first caught salmon or a Netsilik boy's first goose (Balikci, 1970: 45); a young Kaoka boy's first pig (Hogbin, 1969: 39); a Wogeo child given his or her first garden plot (Hogbin, 1969: 139–40); the first shepherd's crook given to

a Vlach six-year-old (Campbell, 1964: 156); and the first bow and arrow provided to a Kutenai Indian boy (Grinnel, 1923: 115).

Adolescent Initiation

The timing of the most commonly employed rite of passage is dependent on the appearance of the first signs of puberty. Once the growth spurt, deepening of the voice, menses, breast buds, or pubic hair become visible, a rite of passage may follow. On the island of Vanatinai in Papua New Guinea, at approximately 14 years of age, the first signs of puberty become apparent. "For a girl that is when her breast buds are 'the size of betel nuts,' and for a boy when his voice begins to change" (Lepowsky, 1998: 128). Physiological change and emerging sexuality are often one of the foci of initiation. Among the Tamil of southern India, a girl at first menstruation must avoid 'hot' foods and mature men, for both would inflame her already heightened state of passion. The best cure for the state of heightened passion among nubile young Tamil women is "marriage and frequent sexual intercourse" (Reynolds, 1991: 40). Among the Dogon of north-central Africa, parents keep a watchful eye on the formation of their daughter's breasts and other signs of sexual maturity. They are concerned that, if the necessary puberty rite is not performed before her first menses, when she finally becomes pregnant, her first child may die (Calame-Griaule, 1986). Among the Muria of India, children are initiated (scarification on the chest and upper arms of the body) after the onset of heterosexual relations, to acknowledge this important phase in their transition into adulthood (Elwin, 1943).

Rites for girls emphasize fecundity, subservience to senior women, and obedience to one's future husband (Richards, 1956: 103); and those for boys, subservience to senior men and dominance over women (Tuzin, 1980: 26). The youth is forcibly weaned from the 'bad influence' of the peer group (Rao, 2006: 59). Didactic instruction in the 'lore' of the society is not evident. On the contrary, the initiation rite is an opportunity to impress upon young people their ignorance and powerlessness. "In Kpelle society secrecy … supports the elders' political and economic control of the youth" (Murphy, 1980: 193). Children entering puberty may need to prove to the larger community that they are ready for adulthood and so are tested. They may be sequestered, go without food, and be forced to withstand physically challenging or painful ordeals. Among the Mapuche, at first menses a girl is segregated in a corner of the *toldo*. For the next two days she is made to run long distances as fast as she can. On the third day she is told to go and gather three bundles of firewood, leaving them at three different locations. After she completes these tasks there is a celebration of the girl's newly achieved status (Cooper, 1946). In northwestern New Guinea sexual maturation or menses in Kwoma girls occurs without public recognition, but puberty in Kwoma boys must be induced by imitating menstruation. During their rite of passage, older men repeatedly scrape a boy's penis to induce bleeding. The boys are encouraged to continually bleed themselves after their rite of passage is complete, to ensure proper growth and to keep themselves fit (Williamson, 1983).

Puberty rites may commence at the first signs of puberty or later, and they may be quite short in duration or last several years. Hence the completion of the rites may not automatically confer adult status. There may be further, subtler tests of an adolescent's preparedness for marriage and formation of family. Gusii parents may withhold approval and resources from the aspirant bride, demanding "evidence of *okongainia* … which means … being willing and able to do the work of an adult woman … and perform these duties without having to be ordered" (LeVine and Lloyd, 1966: 167). As we will discuss shortly, there are multiple, cross-cultural pathways, to full adult status. Consequently, marriage is often treated as only a very minor rite of passage.

Life Phases

Our survey of ethnographic literature shows that one cannot rely on semantically or ritually marked transitions in the life course as guides to the cultural construction of child development. This is because phases are rarely specified in full by formal means; rather, one must construct them from descriptions of the child's behavior and the behavior of others toward the child (Mead, 1947: 234). For example, in an earlier and widely cited survey, the authors identified a widespread but unnamed phenomenon they referred to as 'toddler rejection' (Weisner and Gallimore, 1977: 177), which defines conceptually how many societies view the postweaning period. Anthropologists consistently note a significant shift in the status and treatment of the child between 2 and 3 years but this phase is rarely labeled as such.

Phase One: Birth and the External Womb

As noted earlier, the child's birth may be shrouded in secrecy. Pregnancy itself is rarely acknowledged publicly because there are many factors that might adversely affect the outcome. A 'wait and see' attitude is adopted, as the mother may not be physically able to carry the pregnancy to full term due to poor health or she may choose to terminate it due to lack of support from family or the community. This 'wait and see' attitude is carried into the birthing process and postpartum seclusion provides a curtain of secrecy and security behind which the fate of the infant is decided. After birth, an infant may be seen as still intimately linked with its mother for survival and in mortal danger. "[The Somali] conception is that the newborn child for a certain time after birth is still … part of the flesh and blood of the mother" (Cerulli, 1959: 25). "[Wari] mother and infant are treated as a unit; for about six weeks after birth they remain secluded together inside their house … babies of both sexes are called *arawet*, which translates literally as 'still being made'" (Conklin and Morgan, 1996: 672). The newborn is not fully human: it's bones are soft (Helander, 1988: 150); it is pale and lacks the proper human color (Childs, 1949: 120); it has no mind or thoughts other than the breast (Shostak, 1981: 113); it can't speak and is as yet, empty (Du Bois et al., 1944: 76). And these folk beliefs are consistent with most theories of

human ontogeny that suggest the infant is 'born three months too early' (Bogin and Smith, 2012).

The postpartum womb is a key component of the way the infant is conceptualized and cared for. So tight is the bond, that the Japanese infant is looked upon as part of the mother's body (Lebra, 1994: 260). The use of swaddling and severely confining cradles or cradleboards is widespread. In the high Andes, babies are almost constantly confined to a 'manta pouch,' which functions to reduce the baby's metabolism and need for energy (Tronick et al., 1994: 1009–1010). Nurzay women explained that "the newborn baby's flesh is *oma* (unripe) like uncooked meat, and that only by swaddling will it become strong (*chakahosi*) and solid like cooked (*pokh*) meat" (Casimir, 2010: 16). Navajo babies are kept "in the cradleboard to make them straight and strong. Some women let their children lie on sheepskins and roll about, but they are always weak, sick children" (Leighton and Kluckhohn, 1948: 23).

Even more common is the practice of attaching the baby to its mother via a length of cloth or a sling. The nursing mother shifts the baby quickly to the breast at the first sign of movement (Broch, 1990: 31). The infant remains within a cloth envelope throughout and the nursing process passes completely unnoted by others. A baby should be invisible (Lancy and Grove, 2011a: 283) to protect it from many threats, including witchcraft and the supernatural. Even the father may be seen as dangerous (Gray, 1994: 67) and the postpartum taboo prohibiting contact between the nursing mother and her husband is widespread (Lloyd, 1970: 81).

Another component of this phase is the widely distributed notion that the infant is in a liminal state suspended between the human and spirit worlds (Razy, 2007). There is the idea that the soul and body are only loosely connected and that, if the infant is not closely confined, its soul will escape back to the world of spirits and ancestors (Arden, 2011; Leavitt, 1998). Care must be taken to prevent the infant from moving its limbs vigorously or becoming agitated (Nicolaisen, 1998; Lancy, 2013).

Phase Two: Joining the Community

As we have indicated in Phase One, the infant does not exist as a distinct entity. The likelihood of death or infanticide are so great that the infant's passing goes unmarked and there will be no formal funeral, burial, or mourning (Becker, 2007: 282). In effect, gestation continues beyond birth. The infant must exit from this metaphorical womb and enjoy a second birth. This second or social birth (Fabian, 1990) may be marked by a rite of passage such as naming or the first haircut or nail-trimming (Masters, 1953), as discussed earlier. An important element in the construction of personhood is the child's acquisition of kin and linkages to the father, his clan, and extended family (Blanchy, 2007). "Many Hubeer … postnatal practices [involve the] shedding of the symbols for maternal ties [and establishment of] agnatic links" (Helander, 1988: 150). The transition from crawling (which is animal-like) to walking (human) may be highlighted. The child is now acknowledged as human or at least as potentially human.

Another very important attribute of the baby's 'coming out' is a very sharp spike in the involvement of allomothers. Humans are cooperative breeders and, within a few months of birth, the baby is increasingly under the care of grandmothers, older siblings, and fathers (Hurdy, 2005: 65–91), so that mothers can return to their labors full-time. Mothers may be so eager to attract the assistance of allomothers that they market their babies in interesting ways to neighbors and kin (Ochs and Schieffelin, 1984: 279; Gottlieb, 2004). A thriving infant will attract the caring attention of many, as well as threats from those who might be envious of the fertile mother, so steps must be taken to protect the now very public baby from malevolence (Einarsdóttir, 2006: 116–17; Friedl, 1997: 88; Jenkins et al., 1985: 43). The nursling enjoys a kind of honeymoon of affection and care from all sides but as weaning approaches, a change is evident.

Phase Three: Separation

This phase corresponds to our western notion of 'early childhood' (Historical Change and Human Development). As noted earlier, a prominent feature of this phase is 'toddler rejection' (Weisner and Gallimore, 1977: 176). Weaning, sometimes early – long before the child might wean itself – and severe, is widely reported. "A [Luo] woman who is pregnant is supposed to stop breast-feeding, since it is believed that the milk will be poisonous to the nursing baby and will cause it to get the illness *ledho*" (Cosminsky, 1985: 38). Numerous ethnographic accounts show mothers imposing early and abrupt termination of breast-feeding. Extended nursing may be condemned as prolonging the infancy phase, resulting in a 'weak, simpering' adult (Turner, 1987: 107). The mother is also eager to wean the child from her back, which may be just as tearfully resisted as weaning from the breast (Maretzki et al., 1963: 447). The Yoruba are quite typical in averring that "mothers and grandmothers [prefer] wiry and agile babies who learn to walk early" (Zeitlin, 1996: 412). The Nso of the Cameroons believe that: "A standing baby … makes less work for the mother" (Keller, 2007: 124).

Aside from the attentions of allomothers (often a grandmother at this phase), separation from the mother is aided by the child's powerful attraction to the neighborhood playgroup. "With the arrival of the next sibling, *dénanola* (infancy) is over. Now, play begins … and membership in a social group of [Mandinka] peers is taken to be critical to … the forgetting of the breast" (Whittemore, 1989: 92). The locus of the playgroup may be in a space that the Kpelle label the 'mother ground' as it lies in the vicinity of at least a few working but attentive adults (Lancy, 1996: 85). Aside from freeing up the mother for other pursuits, sending toddlers off in the company of sibling caretakers and playmates is seen as an essential component of their socialization. For example, in rural Bengal "Little girls accompany older girls in gathering, and they gradually learn the needed skills" (Rohner and Chaki-Sircar, 1988: 33). Marquesan mothers see toddlers as developing skills because they *want* to hang out with and emulate their older siblings. By imitating their sibling caretakers "toddlers learn to run, feed, and dress

themselves, go outside to urinate and defecate, and help with household chores" (Martini and Kirkpatrick, 1992: 124).

Once the child has accepted its separation from the mother's breast and back, it is readmitted into the family circle, so to speak. The behaviors and conversations of those older constitute a kind of classroom where the child rapidly learns its culture. Matsigenka "infants and young children are embedded in the middle of quotidian activities where they are positioned to quietly observe and learn what others are doing" (Ochs and Izquierdo, 2009: 395)." "At the age of three he chooses his own place at the [Wolof] family meal and here he is encouraged to acquire social norms" (Zempleni-Rabain, 1973: 222).

In addition to learning social graces, this period constitutes the child's induction into the family economy (Lancy, 2012a). Margaret Mead offered one of the earliest descriptions of a phenomenon often recorded since. (On Samoa), "the tiniest little staggerer has tasks to perform – to carry water, to borrow fire brands, to fetch leaves to stuff the pig …. learning to run errands tactfully is one of the first lessons of childhood" (Mead, 1928: 633). Often the child is overeager to emulate those older and they will be reined in. In Botswana, toddlers may be prevented from handling grain for fear they will let it spill on the ground (Bock and Johnson, 2004). Bamana children are prevented from messing up planted rows in the garden (Polak, 2003: 126) and little Inuit boys are kept some distance from the prey during a hunt, so that they would not scare it off (Matthiasson, 1979: 74). Around 5–7 years, a shift in expectations occurs. Tolerance for the child's helpful but clumsy and inconsistent contributions wanes.

Phase Four: Getting Noticed

Around 5 years, children become 'useful.' They are expected to contribute to the household, by helping with chores, taking care of siblings, and acting as examples (Broch, 1990: 28). Although children may participate in household chores from an early age, they may not be trusted with more serious responsibilities until they 'gain sense' (Harkness and Super, 1986). The child's *lack of* sense is also cited to excuse them from misdemeanors that would be chastised in an older child (Maretzki et al., 1963: 481; Read, 1960: 89). Pashtun girls in middle childhood are sent by their mothers to discreetly scout out and give voluminous reports on the latest events in the village or on a recent scandal (Lindholm, 1982: 181). The workload and responsibility grows with the child's size, strength, and competency. Javanese and Nepalese children work about 4 h a day as six- to eight-year-olds; this rises to 10 h or more by age 15 years (Nag et al., 1978). By age 12 years, Aka and Hadza children are already self-supporting in terms of foraging ability (Hewlett and Cavalli-Sforza, 1986: 930; Hill and Hurtado, 1996: 223). Twelve-year-old Bakgalagadi girls can run an entire household (Lancaster, 1984: 86).

Gender differentiation becomes much more pronounced in middle childhood (Lancy and Grove, 2011a). Girls become more closely associated with the domestic sphere. Conversely, boys may actually gain more freedom from the home and their mother's control (Pope-Edwards, 2005: 87), especially if they are relocated to an all-male domicile (Morton, 1996: 112). There are limits placed on interaction with the opposite sex and clothing reflects heightened modesty requirements (Lawton, 2007: 46). At this age, Berber boys will have "only limited contact with their mothers and are given a pair of undergarments (*serwal*) to wear under the *jelloba* and also a skullcap" (Hatt, 1974: 139). The Dusun call boys, 'without loincloth' and girls 'without a skirt' until they are about 5 years old, when they are then called 'child man' or 'virgin' and given appropriate clothing to wear (Williams, 1969: 86). A gap opens between the relative workloads of girls and boys and relative freedom to roam (Nag et al., 1978). Girls are expected to help with the care of siblings and the daily routine work of their mothers (Nerlove et al., 1974: 275). But in an Iranian village, boys are 'turned out in the morning like cows,' coming home for nourishment and sleep (Friedl, 1997: 148; Watson-Franke, 1976: 194).

Phase Five: Youth in Limbo

No phase in the life cycle is variable, cross-culturally, as adolescence. In the most thorough review of the literature, Schlegel and Barry assert that adolescence is discernible in all societies (1991: 18). However, puberty and the onset of adolescence may be evident as young as 12 years in well-nourished populations and as late as 17 years, in societies where nutrition is poor (Eveleth and Tanner, 1990: 170). Marriage, child-bearing, and the establishment of an independent household mark the end of adolescence. In Kau Sai, a traditional Chinese fishing village, marriage and household formation begin at 16 years, immediately following the first menses (Ward, 1970: 115); while Masaai males are not permitted to marry before 30 years (Spencer, 1970: 137). In the first case, development into adulthood proceeds seamlessly as children reliably acquire the skills appropriate to adults of the same gender and are fully competent at an early age. In the second, a rigid hierarchy among males diverts pubertal males from sex to warrior status – a role they can only fulfill by avoiding women and living at the outer perimeter of the group's territory. Pubescent girls join the polygynous households of much older, senior men as junior wives.

For most societies, adolescence, especially if it lasts longer than a year or two, is somewhat problematic. First there is the challenge of adolescent sexuality. Young people are typically interested in sex well before society thinks they are capable of managing a family. We discussed earlier the initiation rite as a means to enforce emotional maturity and deference to one's elders. Other tactics include seclusion. The Khmer seclude daughters after first menses in a state referred to as *coul plup*, or 'entering the shade.' The period varies but a longer period of seclusion provokes a more generous bride price (Smith-Hefner, 1993: 145–146). While the primary purpose of seclusion seems to be to preserve the girl's virtue to insure a successful marriage, a secondary purpose may be to shape the girl's outlook to more closely match that of the older women. A high-ranking Tlingit girl spends 2 years in seclusion during which she is informed of her clan's history and provided with homilies and more forceful reminders

to maintain behavior consonant with her rank, such as avoiding gossip (Markstrom, 2008: 145). Following seclusion, the Tlingit girl is considered marriageable and "prudent … parents took pains to marry her off promptly" (De Laguna, 1965: 21). Even in the absence of seclusion or initiation, pubertal girls are sheltered by the distaff side of the household. Their association with, and close ties to the mother begin quite early as they assist with their younger siblings and with work in the house and garden, with livestock, and craft production. So while adolescent males enjoy a wider compass for their activities, adolescent girls rarely leave their mothers' orbit.

The process for boys – parallel to girls' seclusion – is removal to a dormitory or men's house. An Igbo boy moves out of his natal home to an all-male compound (Ottenberg, 1989: 118). This men's sanctuary is the *haus tambaran* in the Sepik River area of Papua New Guinea (Tuzin, 1980). *Tapirapé* adolescent boys are moved into the *takana* or men's house where their proximity to men affords them the opportunity to observe and replicate typically male behaviors (Wagley, 1977: 149). Boys are sequestered to erase the taint of femininity acquired during years of association (primarily) with women while becoming imprinted with the male ethos. Around the world societies that devote attention to masculinizing youth are often quite warlike, and male adolescence may be synonymous with warriorhood. Creating a cohort of new warriors is not the only means societies have of taming the adolescent male 'gang.' They can be conscripted to serve in work details that benefit the village as a whole. Temne living in the young men's house "*kabankalo* served as a cooperative work group for the chief, their fathers, and other big men of the chiefdom" (Dorjahn, 1982: 41). Not so subtly, they will be reminded of their subservient status vis-à-vis the village elite.

Aside from physical and social maturity and the acquisition of critical survival skills, the portal from adolescence to adulthood is usually opened only when the requisite property is available. This may include payment of a bride price or dowry, the construction of a home for the new couple, the provision of critical resources, such as tools or livestock. When the traditional pathways to adulthood break down through the abandonment of these traditional practices and customs or their suppression by church or government authorities, adolescents can't become social adults (Biersack, 1998). Instead, they become 'insurgents' (Honwana, 2006; Rosen, 2005) or village *bikhets* (Leavitt, 1998).

Phase Six: Adulthood

The most important thing to understand about adulthood is that, outside the contemporary bourgeoisie, societies are organized as gerontocracies (Lancy, 1996: 13, 2008: 11). One's status is governed by age and fertility. Marriage per se may carry little weight and children aged 10 years may already be capable of the full range of adult subsistence tasks (Lancy, 2012a); hence, it may take many years after marriage to become accepted as an adult. The Javanese marry their daughters as early as 9 or 10 years to stave off any hint of premarital sex and illegitimacy. A girl becomes a woman under the guidance of her mother-in-law (Geertz, 1961: 56). A Bagisu bride is given a 'woman's skirt,' but she can't wear it until the birth of her first child (La Fontaine, 1986). In Sumatra, a married Malay woman is dependent on her father, much like her unmarried sisters, until she gives birth to her first child (Swift, 1965: 124). The same criterion is often applied to young males (Leavitt, 1998: 186). For the Sambia, adulthood status is denied to the young couple until they have successfully conceived and birthed at least two children (Herdt, 2001: 164). Progeny are so important to the Akan that a man without children may be called a 'wax penis' and upon death, childless adults may have thorns driven into the soles of their feet (Warren, 1986: 11). A Chukchee man will take an additional wife if his first wife does not give him children; his first wife, if she is a 'good' wife will gladly encourage him to do so (Bogoraz-Tan, 1924: 600).

Conclusion

Cultural models of human development (Strauss, 1992) vary cross-culturally but shared patterns can be detected in the ethnographic record. Clues to discerning such patterns can be sought in milestones of biological development, such as first menses and in indigenous nomenclature and rites of passage. One must also examine child care, dress, patterns of residence, selection of caretakers, assignment of chores, and other culturally sanctioned practices that change reliably over the life cycle. These broadly applicable cultural models coalesce into six life phases in the life cycle. Phase I includes the birth of the infant and early infancy. Unacknowledged by society, the newborn is still seen as part of the mother and therefore still in the process of being made. Its survival not guaranteed, the infant is kept tucked away in some form of an external womb, until its survival is sure. Phase II occurs when it is clear that the infant will survive. The infant is introduced to the community, usually through a rite of passage that may include a first haircut and/or a naming ceremony. Alloparents join the mother in caring for the child. Phase III removes the child from the mother's breast and provides a push toward independence with weaning, walking, and toilet training. Toddler rejection, as this is often called, is mitigated by the lure of the playgroup. In Phase IV children become useful. Younger children are expected to do chores and, as they gain common sense, are entrusted with ever more difficult tasks. Gender differentiation is expressed as girls become tied to the domestic sphere and boys are allowed more freedom from the domestic sphere. Phase V or adolescence is highly variable and elastic across and within cultures. Passage into or out of this phase may be marked by an initiation rite. In some cases, adolescence is short, as youth begins cohabitation leading eventually to family formation. Lengthening this phase often creates a social problem with youth aggregating into gangs and behaving antisocially and we discussed a suite of culturally constructed responses. Phase VI commences with marriage, but marriage is not the key to adulthood – the entry to adulthood is rather the successful siring/bearing and raising of children.

Outlook

The ethnographic record – with well over 1000 discrete reports of childhood from hundreds of societies – has proven a rich source of ideas for reconsidering child development. We have drawn on such sources in constructing this brief overview of life phases. However, we can identify a number of additional issues and questions that might be pursued. For example, in the ethnographic record, infant and child mortality is very high. This tragedy has been largely eliminated in postindustrial society. On the other hand, children in economically developed societies suffer from emotional stress, depression, and disorders like ADHD – all completely unknown in indigenous societies. Are genotypes now preserved that would not have survived prior to the development of modern neonatal medical care? Here is another paradox. Children in small-scale societies are eagerly involved in daily routines and eagerly help out with chores. This contribution to family life continues throughout childhood. Such predilections have almost completely disappeared in contemporary bourgeoisie society and parents lament their children's selfish, self-centered attitudes. Could there be a critical period at 18 months to 2 years where the child's prosocial attempts are rebuffed (because letting them help makes extra work or is too dangerous), thus extinguishing the motivation? Or are contemporary parents so concerned about raising unique individuals that they give little thought to prosociality, whereas traditional economies are dependent on and seek to foster cooperative, sociocentric children? In the ethnographic record, we see that children are rarely pushed to learn or develop; they progress at their own pace. Recognition of achievement or celebrating a child's birthdays and other milestones are rare to nonexistent. Praise and gifts are offered rarely. In contrast, we track age and many other milestones, carefully calibrating progress against a 'standard.' In contemporary societies a child can be identified as failing or 'developmentally challenged' almost from birth. At the same time, we lavish praise on our children, all of whom are 'special.' So, for 1 min they are perfect, the next, flawed. Could this be the source for the epidemic of emotional disorders? In indigenous societies adolescence varies in length, but in most societies children have mastered the full suite of adult subsistence skills and are self-sufficient not long after puberty. In contemporary society, preparations to become employed and self-sufficient and then get established in a career may prolong the adolescent phase for a decade or more. Are there social consequences of this drastic increase in the period of juvenility? Until quite recently, formal education affected a tiny minority. In traditional societies children are primarily social learners and parents rarely see the need to teach. Children learn largely through observation, imitation, and practice. It is our impression that we are witnessing the rapid spread of a form of learned helplessness. Youth in bourgeoisie society seem unable to learn without guidance from a teacher. The belief that children's activities require the careful guidance of an adult extends to play, an activity formerly the exclusive domain of children. Are we undermining our children's very powerful abilities (the basis of human culture) to learn socially? Analysis of the ethnographic record provides numerous examples of the incongruity between how indigenous and bourgeoisie societies view and structure childhood. This contrast sheds light on the lives of all the world's children, enriching conversations about the course of development.

See also: Human Development, Successful: Psychological Conceptions; Human Development, Theories of; Lifespan Development: Evolutionary Perspectives.

Bibliography

Appell-Warren, 2012. Named developmental stages as indicators of personhood and socio-cultural development. Paper presented at the 2012 Meeting of the Society for Cross-Cultural Research held jointly with the Society for Anthropological Sciences and the American Anthropological Association's Children and Childhood Interest Group, Las Vegas, Nevada.

Arden, Traci, 2011. Empowered children in classic Maya sacrificial rites. Childhood in the Past 4, 133–145.

Balikci, Asen, 1970. The Netsilik Eskimo. The Natural History Press, Garden City.

Baxter, Paul T.W., 1953. The Azande, and Related Peoples of the Anglo-Egyptian Sudan and Belgian Congo. International African Institute, London.

Becker, Marshall J., 2007. Childhood among the Etruscans; Mortuary programs at Tarquinia as indicators of the transition to adult status. In: Cohen, Ada, Rutter, Jeremy B. (Eds.), Constructions of Childhood in Ancient Greece and Italy, Hesperia Supplement 410. American School of Classical Studies at Athens, Athens.

Biersack, Aletta, 1998. Horticulture and hierarchy: the youthful beautification of the body in the Paiela and Porgera Valleys. In: Herdt, Gilbert, Leavitt, Stephen (Eds.), Adolescence in Pacific Island Societies. University of Pittsburgh Press, Pittsburgh, pp. 71–91.

Blanchy, Sophie, 2007. Le tambavy des bébés à Madagascar: du soin au ritual d' ancestralité. In: Bonnet, Doris, Pourchez, Laurence (Eds.), Du Soin au Rite dans l'infance. IRD, Paris, pp. 146–166.

Bock, John, Johnson, Sara E., 2004. Subsistence ecology and play among the Okavango delta peoples of Botswana. Human Nature 15 (1), 63–81.

Bogin, Barry, Smith, B. Holly, 2012. Evolution of the human life cycle. In: Stinson, Sara, Bogin, Barry, O'Rourke, Dennis (Eds.), Human Biology: An Evolutionary and Biocultural Perspective, second ed. John Wiley, New York, pp. 515–586.

Bogoraz-Tan, Vladimir Germanovich, 1924. New problems of ethnographical research in polar countries. In: Proceedings of the Twenty-first International Congress of Americanists. The Hague.

Broch, Harald B., 1990. Growing up Agreeably: Bonerate Childhood Observed. University of Hawaii Press, Honolulu.

Calame-Griaule, Geneviève, 1986. Words and the Dogon World. Institute for the Study of Human Issues, Philadelphia.

Campbell, John K., 1964. Honour, Family, and Patronage: A Study of Institutions and Moral Values in a Greek Mountain Community. Clarendon Press, Oxford.

Casimir, Michael J., 2010. Growing up in a Pastoral Society: Socialization among Pashtu Nomads. Kölner Ethnologische Beiträge. Druck and Bindung, Kölon.

Cerulli, Enrico, 1959. How a Hawiye Tribe Used to Live. A Cura dell'Amministrazione Fiduciaria Italiana della Somalia Instituto poligrafico dello Stato P.V, Roma.

Childs, Gladwyn M., 1949. Umbundu Kinship and Character: Being a Description of Social Structure and Individual Development of the Ovimbundu of Angola. Oxford University Press, London.

Conklin, Beth A., Morgan, Lynn M., 1996. Babies, bodies and the production of personhood in North America and a native Amazonian society. Ethos 24 (4), 657–694.

Cooper, John M., 1946. The Araucanians. Smithsonian Institution, Washington, DC.

Cosminsky, Sheila, 1985. Infant feeding practices in rural Kenya. In: Hull, Valeria, Simpson, Mayling (Eds.), Breastfeeding, Child Health and Birth Spacing: Cross-Cultural Perspectives. Croom Helm, London, pp. 35–54.

De Laguna, Frederica, 1965. Childhood among the Yakutat Tlingit. In: Spiro, Melford E. (Ed.), Context and Meaning in Cultural Anthropology. Free Press, New York, pp. 3–23.

Dean, Carolyn, 2002. Sketches of childhood: children in Colonial Andean art and society. In: Hecht, Tobias (Ed.), Minor Omissions: Children in Latin American History and Society. University of Wisconsin Press, Madison, pp. 21–51.

Dorjahn, Vernon R., 1982. The initiation and training of Temne Poro Members. In: Ottenberg, Simon (Ed.), African Religious Groups and Beliefs. Archana, Meerut, India, pp. 35–62.
Du Bois, Cora A., Kardiner, Abram, Oberholzer, Emil, 1944. The People of Alor: A Social Psychological Study of an East Indian Island. University of Minnesota Press, Minneapolis.
Einarsdóttir, Jónína, 2006. Relocation of children: fosterage and child death in Biombo, Guinea-Bissau. In: Christiansen, Catrine, Utah, Mats, Vigh, Henrik E. (Eds.), Navigating Youth, Generating Adulthood: Social Becoming in an African Context. Nordiska Afrikainstitutet, Uppsala, pp. 183–200.
Elwin, Verrier, 1943. Maria Murder and Suicide: Man in India. Oxford University Press, Bombay.
Eveleth, Phyllis B., Tanner, James M., 1990. Worldwide Variation in Human Growth. Cambridge University Press, New York.
Fabian, Stephen M., 1990. Space-Time of the Bororo of Brazil. University Press of Florida, Gainseville.
Fajans, Jane, 1997. They Make Themselves: Work and Play among the Baining of Papua New Guinea. University of Chicago Press, Chicago.
Fricke, Tom, 1994. Himalayan Households: Tamang Demography and Domestic Process. Columbia University Press, New York.
Friedl, Erika, 1997. Children of Deh Koh: Young Life in an Iranian Village. Syracuse University Press, Syracuse.
Geertz, Hildred, 1961. The Javanese Family: A Study of Kinship and Socialization. Free Press, New York.
Gorer, Geoffrey, 1967. Himalayan Village: An Account of the Lepchas of Sikkim. Basic Books, New York.
Gorshkov, Boris B., 2009. Russia's Factory Children: State, Society, and Law, 1800–1917. University of Pittsburgh Press, Pittsburgh.
Gottlieb, Alma, 2004. The Afterlife Is Where We Come from: The Culture of Infancy in West Africa. The University of Chicago Press, Chicago.
Gray, Brenda M., 1994. Enga birth, maturation and survival: physiological characteristics of the life cycle in the New Guinea Highlands. In: MacCormack, Carol P. (Ed.), Ethnography of Fertility and Birth. Waveland Press, Inc, Prospect Heights.
Grinnel, George B., 1923. The Cheyenne Indians: Their History and Ways of Life, Vol. 2. Yale University Press, New Haven.
Harkness, Sara, Super, Charles M., 1986. The cultural structuring of children's play in a rural African community. In: Blanchard, Kendall (Ed.), The Many Faces of Play. Human Kinetics, Champaign, pp. 96–103.
Hatt, Doyle Gordon 1974. Skullcaps and turbans: domestic authority and public leadership among the Idaw Tanan of the Western High Atlas, Morocco (Unpublished Ph.D. Dissertation), University of California, Los Angeles.
Helander, Bernhard, 1988. Slaughtered Camel: Coping with Fictitious Descent among the Hubeer of Southern Somalia. Department of Anthropology, University of Uppsala, Uppsala.
Herdt, Gilbert, 2001. Rituals in manhood: male initiation in Papua New Guinea. In: Brettell, Caroline B., Sargent, Carolyn F. (Eds.), Gender in Cross-Cultural Perspective. Prentice Hall, Upper Saddle River, pp. 162–166.
Hewlett, Barry S., Cavalli-Sforza, Luigi L., 1986. Cultural transmission among Aka pygmies. American Anthropologist 88, 922–934.
Hill, Kim, Hurtado, A. Magdalena, 1996. Ache Life History: The Ecology and Demography of a Foraging People. Aldine de Gruyter, New York.
Hogbin, Ian, 1969. A Guadalcanal Society: The Kaoka Speakers. Holt, Rinehart, and Winston, New York.
Honwana, Alcinda, 2006. Child soldiers in Africa. University of Pennsylvania Press, Philadelphia.
Hough, Walter, 1915. The Hopi Indians. Torch Press, Cedar Rapids.
Hurdy, Sarah Blaffer, 2005. Comes the Child before Man: How cooperative Breeding and Prolonged Postweaning Dependence Shaped Human Potential. In: Hewlett, Barry S., Lamb, Michael E. (Eds.), Hunter Gatherer Childhoods: Evolutionary, Developmental, and Cultural Perspectives. AldineTransaction, New Brunswick, NJ, pp. 65–91.
Huntingford, George Wynn Brereton, 1953. The Southern Nilo-Hamites. International African Institute, London.
Jenkins, Carol L., Orr-Ewing, Alison K., Heywood, Peter F., 1985. Cultural aspects of early childhood growth and nutrition among the Amele of Lowland Papua New Guinea. In: Marshall, Leslie B. (Ed.), Infant Care and Feeding in the South Pacific. Gordon and Beach, New York, pp. 29–50.
Keller, Heidi, 2007. Cultures of Infancy. Lawrence Erlbaum, Mahwah.
La Fontaine, Jean, 1986. Initiation. Penguin.
Lancaster, Jane B., 1984. Evolutionary perspectives on sex differences in the higher primates. In: Rossi, Alice S. (Ed.), Gender and the Life Course. Aldine, New York, pp. 3–28.
Lancy, David F., Grove, M. Annette, 2011a. "Getting noticed": middle childhood in cross-cultural perspective. Human Nature 22, 281–302.
Lancy, David F., Grove, M. Annette, 2011b. Marbles and Machiavelli: the role of game play in children's social development. American Journal of Play 3 (4), 489–499.
Lancy, David F., Gaskins, Suzanne, Bock, John (Eds.), 2010. The Anthropology of Learning in Childhood. Alta-Mira Press, Lanham.
Lancy, David F., 1996. Playing on the Mother Ground: Cultural Routines for Children's Development. Guilford, New York.
Lancy, David F., 2007. Accounting for the presence/absence of mother-child play. American Anthropologist 109 (2), 273–284.
Lancy, David F., 2008. The Anthropology of Childhood: Cherubs, Chattel, Changelings. Cambridge University Press, Cambridge.
Lancy, David F., 2010. Learning "from nobody": the limited role of teaching in folk models of Children's development. Childhood in the Past 3, 79–106.
Lancy, David F., 2012a. The chore curriculum. In: Spittler, Gerd, Bourdillion, Michael (Eds.), Working and Learning among Africa's Children. Lit Verlag, Berlin.
Lancy, David F., 2012b. "First you must master pain:" the nature and purpose of apprenticeship. Society for the Anthropology of Work Review. 33 (2), 113–126.
Lancy, David F., 2013. "Babies aren't persons": a survey of delayed personhood. In: Keller, Heidi, Otto, Hiltrud (Eds.), Different Faces of Attachment: Cultural Variations of a Universal Human Need. Cambridge University Press, Cambridge.
Lawton, Carol, 2007. Children in classical attic votive reliefs. In: Cohen, Ada, Rutter, Jeremy B. (Eds.), Constructions of Childhood in Ancient Greece and Italy. The American School of Classical Study at Athens, Princeton, pp. 41–60.
Leavitt, Setphen C., 1998. The bikhet mystique. In: Herdt, Gilbert, Leavitt, Stephen C. (Eds.), Adolescence in Pacific Island Societies. University of Pittsburg Press, Pittsburg, pp. 173–194.
Lebra, Takie Sugiyama, 1994. Mother and child in Japanese socialization: a Japan-U.S. comparison. In: Greenfield, Patricia M., Cocking, Rodney R. (Eds.), Cross-Cultural Roots of Minority Child Development. Erlbaum, Hillsdale, pp. 259–274.
Leighton, Dorothea, Kluckhohn, Clyde C., 1948. Children of the People. Harvard University Press, Cambridge.
Lepowsky, Maria, 1998. Coming of age on Vanatinai: gender sexuality, and power. In: Herdt, Gilbert, Leavitt, Stephen C. (Eds.), Adolescence in Pacific Island Societies. University of Pittsburgh Press, Pittsburgh, pp. 123–147.
LeVine, Robert Alan, Lloyd, Barbara Bloom, 1966. Nyansongo: A Gusii Community in Kenya. Wiley, New York.
Lindholm, Charles, 1982. Generosity and Jealousy: The Swat Pukhtun of Northern Pakistan. Columbia University Press, New York.
Lloyd, Barbara, 1970. Yoruba mother's reports of child-rearing: some theoretical and methodological considerations. In: Mayer, Philip (Ed.), Socialization: The Approach from Social Anthropology. Tavistock Publications, London, pp. 75–107.
Maretzki, Thomas W., Maretzki, Hatsumi Whiting, Beatrice, B., 1963. Taira: an Okinawan village. In: Whiting, Beatrice B. (Ed.), Six Cultures: Studies of Child Rearing. John Wiley and Sons, New York, pp. 363–539.
Markstrom, Carol A., 2008. Empowerment of North American Indian Girls: Ritual Expressions at Puberty. University of Nebraska Press, Lincoln.
Martini, Mary, Kirkpatrick, John, 1992. Parenting in Polynesia: a view from the Marquesas. In: Roopnarine, Jaipaul L., Carter, D. Bruce (Eds.), Parent-child Socialization in Diverse Cultures. Ablex, Norwood, pp. 199–222.
Masters, William Murray, 1953. Rowanduz: A Kurdish Administrative and Mercantile Center. HRAF Computer File, New Haven.
Matthiasson, John S., 1979. But teacher, why can't I be a hunter: Inuit adolescence as a double-blind situation. In: Ishwaran, K. (Ed.), Childhood and Adolescence in Canada. McGraw-Hill Ryerson, Toronto, pp. 72–82.
Mead, Margaret, 1928. Samoan children at work and play. Natural History 28, 626–636.
Mead, Margaret, 1947. Age patterning in personality development. American Journal of Orthopsychiatry 17, 231–240.
Mohammad, Patel H., 1997. Child rearing and socialization among the Savaras. Man and Life 23, 173–182.
Morton, Helen, 1996. Becoming Tongan: An Ethnography of Childhood. University of Hawai'i Press, Honolulu.
Murphy, William P., 1980. Secret knowledge as property and power in Kpelle society: elders versus youth. Africa 50 (2), 193–207.
Nag, Moni, White, Benjamin N.F., Peet, R. Creighton, 1978. An anthropological approach to the study of the economic value of children in Java and Nepal. Current Anthropology 19 (2), 293–306.
Nerlove, Sara B., Roberts, John M., Klein, Robert E., Yarbrough, Charles, Habicht, Jean-Pierre, 1974. Natural indicators of cognitive development: an observational study of rural Guatemalan children. Ethos 2 (3), 265–295.

Nicolaisen, Ida, 1998. Concepts and learning among the Punan Bah of Sarawak. Acquiring Culture Cross-culture Studies in Child Development, 193–221.
Ochs, Elinor, Izquierdo, Carolina, 2009. Responsibility in childhood: three developmental trajectories. Ethos 37, 391–413.
Ochs, Elinor, Schieffelin, Bambi B., 1984. Language acquisition and socialization: three developmental stories and their implications. In: Shweder, Richard A., LeVine, Robert A. (Eds.), Culture Theory: Essays on Mind, Self and Society. Cambridge University Press, New York, pp. 276–320.
Ottenberg, Simon, 1989. Boyhood Rituals in an African Society: An Interpretation. University of Washington Press, Seattle.
Paques, Viviana, Turner, Thomas, 1954. The Bambara. Presses Universitaires de France, Paris.
Pelto, Pertti, J., 1962. Individualism in Skolt Lapp Society. Suomen Muinaismuistoyhdistys, Helsinki.
Polak, Barbara, 2003. Little peasants: on the importance of reliability in child labour. In: E'Almeida-Topor, Hèlén, Lakroum, Monique, Spittler, Gerd (Eds.), Le travail en Afrique noire: représentations et pratiques é l'époque contemporaine. Karthala, Paris, pp. 125–136.
Pope-Edwards, Carolyn, 2005. Children's play in cross-cultural perspective: a new look at the six culture study. In: McMahon, Felicia F., Lytle, Don E., Sutton-Smith, Brian (Eds.), Play: an Interdisciplinary Synthesis. University Press of America, Lanham, pp. 81–96.
Rao, Aparna, 2006. The acquisition of manners, morals and knowledge: growing into and out of Bakkarwal society. In: Dyer, Caroline (Ed.), The Education of Nomadic Peoples: Current Issues, Future Prospects. Berghan Books, New York, pp. 53–76.
Razy, Élodie, 2007. Naître et devenir: anthropologie de la petite enfance en pays Soninké, Mali. Nanterre. Société D'ethnologie, France.
Read, Margaret, 1960. Children of their fathers. Yale University Press, New Haven.
Reynolds, Holly Baker, 1991. The auspicious married woman. In: Wadly, Susan S. (Ed.), The Powers of Tamil Women. Maxwell School of Citizenship and Public Affairs, Syracuse University, Syracuse, pp. 35–60.
Richards, Audrey I., 1956. Chisungu. Faber and Faber, London.
Rohner, Ronald P., Chaki-Sircar, Manjusri, 1988. Women and Children in a Bengali Village. University Press of New England, Hanover.
Rosen, David M., 2005. Armies of the Young: Child Soldiers in War and Terrorism. Rutgers University Press, New Brunswick.
Schlegel, Alice, Herbert III, Barry L., 1991. Adolescence: An Anthropological Inquiry. The Free Press, New York.
Schwarz, Maureen Trudelle, 1997. Molded in the Image of Changing Woman: Navajo Views on the Human Body and Personhood. University of Arizona Press, Tucson.
Shostak, Marjorie, 1981. Nisa: The Life and Words of a !Kung Woman. Harvard University Press, Cambridge.
Silva, Alcionilio Bruzzi Alves da, Lillios, Ivana, 1962. The Indigenous Civilization of the Uaupés. Centro de Pesquisas de lauareté, Sao Paulo.
Smith-Hefner, Nancy J., 1993. Education, gender, and generational conflict among Khmer refugees. Anthropology and Education Quarterly 24 (2), 135–158.
Spencer, Paul, 1970. The function of ritual in the socialization of the Samburu Moran. In: Mayer, Philip (Ed.), Socialization: The Approach from Social Anthropology. Tavistock Publications, London, pp. 127–157.
Strauss, Claudia, 1992. Models and motives. In: d'Andrade, Roy, Strauss, Claudia (Eds.), Human Motives and Cultural Models. Cambridge University Press, Cambridge, pp. 1–20.
Swift, Michael Godfrey, 1965. Malay Peasant Society in Jelebu. Athlone Press: Humanities Press, New York.
Toren, Christina, 1990. Making Sense of Hierarchy: Cognition as Social Process in Fiji. The Athlone Press, Atlantic Highlands.
Tronick, Edward Z., Thomas, R. Brook, Daltabuit, Magali, 1994. The Quechua manta pouch: a caretaking practice for buffering the Peruvian infant against the multiple stressors of high altitude. Child Development 65, 1005–1013.
Turnbull, Colin M., 1965. The Mbuti Pygmies: An Ethnographic Survey. American Museum of Natural History, New York.
Turner, Diane Michalski, 1987. What happened when my daughter became a Fijan. In: Butler, Barbara, Turner, Diane Michalski (Eds.), Children and Anthropological Research. Plenum Press, New York, pp. 92–114.
Tuzin, Donald, 1980. The Voice of the Tambaran: Truth and Illusion in Ilahita Arapesh Religion. University of California Press, Los Angeles.
van Gennep, Arnold, 1961. The Rites of Passage. University of Chicago Press, Chicago.
Wagley, Charles, 1977. Welcome of Tears: The Tapirapé Indians of Central Brazil. Oxford University Press, New York.
Ward, Barbara E., 1970. Temper tantrums in Kau Sai: some speculations upon their effects. In: Mayer, Philip (Ed.), Socialization: The Approach from Social Anthropology. Tavistock Publications, London, pp. 107–125.
Warren, Dennis M., 1986. The Akan of Ghana: An Overview of the Ethnographic Literature. Pointer Ltd, Accra.
Watson-Franke, Maria-Barbara, 1976. To learn for tomorrow: enculturation of girls and its social importance among the Guajiro of Venezuela. In: Wilbert, Johannes (Ed.), Enculturation in Latin America. UCLA Latin American Center Publications, Los Angeles, pp. 191–211.
Weisner, Thomas S., Gallimore, Ronald, 1977. My brother's keeper: child and sibling caretaking. Current Anthropology 18, 169–190.
Wenger, Martha, 1989. Work, play and social relationships among children in a Giriama community. In: Belle, Deborah (Ed.), Children's Social Networks and Social Supports. Wiley, New York, pp. 91–115.
Whittemore, Robert Dunster, 1989. Child caregiving and socialization to the Mandinka way: toward an ethnography of childhood (Unpublished Ph.D. dissertation), UCLA, Los Angeles.
Williams, Thomas R., 1969. A Borneo Childhood. Holt, Rinehart and Winston, New York.
Williamson, Margaret Holmes, 1983. Sex relations and gender relations: understanding Kwoma conception. Mankind 14 (1), 13–23.
Zeitlin, Marian, 1996. My child is my crown: Yoruba parental theories and practices in early childhood. In: Harkness, Sara, Super, Charles M., Charles, M. (Eds.), Parent's Cultural Belief Systems: Their Origins, Expressions, and Consequences. Guilford, New York, pp. 407–427.
Zempleni-Rabain, Jacqueline, 1973. Food and strategy involved in learning fraternal exchange among Wolof children. French Perspectives in African Studies, 220–233.

Culturally Appropriate Interventions in Social Work

Kish Bhatti-Sinclair, University of Chichester, Chichester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The aim of this article is to define the ideas underpinning culturally appropriate interventions and to introduce related concepts which have contributed to developments in countries such as the United Kingdom. Words such as culture, ethnicity, race, racism, and race relations have played an important role in the development of culturally aware social work practice and will, therefore, be explored for relevance and usefulness.

Introduction

Social work competence (i.e., the incremental development of skills over time) requires understanding the impact of relevant, and sometimes opposing, theoretical developments derived from across the world. Professional interventions may be defined as context-dependent expert interactions used to meet the specific social, health, or psychological needs of individuals, groups, and communities. The aim is to maximize understanding, minimize risk, and take full account of behavior and emotion (Walker, 2002). Cultural competence should be based on advanced knowledge of factors (such as class, race, sex, and income) that intersect to advantage or disadvantage service users (Law, 2010). To be meaningful, professional input should be supported by culturally appropriate training and procedures designed to respond to individuals and groups in relation to their languages, traditions, and customs.

Professionals who intervene in a culturally appropriate manner respond to the welfare needs of individuals and groups from ethnic minority backgrounds seeking help from public, private, and independent sector organizations (Lum, 1997). The methods of intervention may vary according to country and context, but the need to be culturally aware remains dynamic and current, particularly in places with a large and complex population mix. Many countries have developed statutory instruments, which allow individual challenges against discriminatory actions and behaviors. Societies like these are more likely to acknowledge racism as a social construct and have people living together in multiracial communities that demonstrate a clear commitment to improving relations and sharing cultural, religious, and other practices (Aveling, 2007; Walker, 2002).

Social workers are required to use practice methods that are appropriate to the culture and ethnicity of the service user and are in line with the relevant legal and policy requirements (Fook, 2006), which define difference and diversity on the grounds of color, ethnicity, and religion. However, evidence on sensitive practice with indigenous groups has pointed to the shortfall on, for example, the use of aboriginal knowledge, such as the impact of colonization over generations, within aboriginal service developments (Augoustinos and Reynolds, 2001). Critics suggest that approaches, which take little account of how history can influence contemporary experience, are likely to lead to little real change (Taylor et al., 2010).

A great deal of national law and policy is based on the universal conventions on human rights, ethics, and values. Such ideas are used to develop professional practice through organizations such as the International Federation of Social Workers, European Federation of Social Workers, International Association of Schools of Social Work, and national organizations such as the British Association of Social Workers in the United Kingdom. Although standard setting is important in social work (Gray et al., 2008), using global yardsticks to interpret individual cultural interventions can be contextually problematic, that is, it may lead to a professional focus on appropriate behavior rather than on the more politicized, institutional effects of race and racism (Dominelli and Thomas Bernard, 2003; Walker, 2002).

Historical Backdrop

Culturally sensitive services gained prominence in the 1990s alongside the greater professionalization of social work and the move to link discipline-related knowledge to practice outcomes (Fook, 2006; Augoustinos and Reynolds, 2001). During this time, strategies such as the recruitment of same-race social workers and the cultural matching of service users and professional interventions were adopted in countries such as Australia and the United Kingdom. However, critics suggested that tactics such as these sought individual rather than social change and led to the marginalization of particular groups (Patni, 2006). Although some immediate and urgent service needs were met through these methods, the majority remained powerless and outside mainstream services. This was to be addressed through institutional and structural change as a prerequisite to service delivery based on principles of equality and social justice (Sivanandan, 1991; Gilroy, 1995; Ali and Barsamian, 2005; Brah, 1996; Solomos and Back, 1995). However, critical and political approaches to antiracist practice resulted in a major review and change of direction for social work education in countries such as the United Kingdom in the early 2000s (Dominelli, 2002).

Culture describes a system of meaning and custom (Cashmore, 1996) and multicultural implies recognition and approval of groups based on difference and diversity. During the 1990s, simplified notions of culture were often related to the assimilation of newly arrived immigrant groups (Sakamoto, 2007), indigenous peoples (Taylor et al., 2010), and established minorities (Jackson et al., 2010). Within the

context of settlement, nation-states implemented assimilation and integration policies using questionable intervention techniques to persuade newcomers and minorities to fit into majority social norms (Lentin, 2000). This approach led to poor outcomes for the social service users in particular, and for society in general, because it did not always lead to structural change or enhance the customs and traditions of the groups in question (Sakamoto, 2007).

Increasing awareness of assimilation by policy-makers (Ely and Denney, 1987), as a method of dominance, has led to greater attention being paid to other approaches, such as integration, which have gained some support as a middle ground alternative for communities living and contributing to society together, while maintaining their own customs and traditions. However, integration requires consideration of fair and useful service outcomes delivered through systems and structures that allow user-led identification of need and intervention methods. The relationship between ideas such as integration and the delivery and outcome of public sector services has received little attention in social work research and literature particularly in areas such as children's mental health (Walker, 2002).

Attention has, however, also been paid to legal drivers that define key terms (i.e., direct and indirect discrimination) and set out individual and organizational duties and responsibilities. Over time, provisions such as these have played a part in sustaining the enduring legal and policy link between minority groups and culture, religion, and skin color. For example, the 1976 Race Relations Act and the 2000 Race Relations Amendment Act in the United Kingdom define unfair or less favorable treatment of individuals and groups on the grounds of color, race, nationality (including citizenship), ethnic or national origins, and religion.

Although malleable and ambiguous, the terms race, racism, and race relations carry equal weight as legal and social constructs in many countries (LeCouteur and Augoustinos, 2001). In Britain, these have been interpreted to develop antiracist social work practice and fine-tune culturally appropriate services. However, proponents of antiracist practice (Dominelli, 1997; Lentin, 2000; Patni, 2006) believe that racism remains at the heart of the problem and therefore view the focus on cultural perspectives as lightweight and superficial.

Race, Racism, and Culture

Race has a range of meanings derived from history and science, complicated by the division of the human *race* into the subspecies of Negroid, Mongoloid, and Caucasoid (Cashmore, 1996). Within social science literature, however, the term has changed its meaning over time from biological to social and cultural:

> Historically, the meaning of the term 'race' has varied. It once focused on a racialised hierarchy that encompassed the physical attributes of all races – identified by skin colour and depicted in popular parlance as white, yellow, red and black (Gobineau, 1953). Definitions of race propagated by Count Gobineau illustrate a biological theory on racialised skin colour. Later, it focussed on allegedly different intelligence levels that favoured the 'white' race.
> (Hernstein and Murray, 1994; Dominelli, 2008: p. 8)

The contemporary social definition of *race* is commonly used to group people who share common origin or descent, usually originating from different countries. Ideas around the subordination of groups based on the biological definition of *race* have been enshrined in law and social policy in many countries of the world (Ben-Eliezer, 2004), with similar discriminatory consequences for individuals everywhere on an everyday basis (Paradies, 2005).

The notion of *race* is questionable at many levels (Dominelli, 2008). British law and policy illustrates its use to imply cultural and ethnic difference based on physical attributes such as skin color, dress, and linguistic skills. Organizations tend to use these definitions to develop poor procedures and processes that are in turn used by misguided social workers in the daily professional duties. Organizational assumptions and interpretations need to be considered alongside those provided by global and national bodies and checked against evidence-based practice. Service interventions can then better relate to particular circumstances and matched to individual and group capacity, strengths, and resilience (Walker, 2002).

The word racism generates a great deal of feeling and can be defined in many ways. It is an ideology or set of beliefs and its development may be explained by oppressive hostility between two sets of groups, nations, or power blocs. For example, the slave trade was based on the notion that African people were racially inferior and thus open to justifiable exploitation. Cashmore (1996: p. 308) suggests that racism needs to be defined variously depending on context and situation: 'Racism takes different empirical forms in different societies at different points in time.'

The term racial discrimination is used interchangeably with racism in European and international conventions (Bhatti-Sinclair, 2011), but the interchangeable use of racism, racial discrimination, and racial disadvantage does not always address discrimination as the practical criminal result of racist ideology. If acted out to the extreme, racism and discrimination may harm those at the receiving end of racial violence. Confusing violent attack with disadvantage on the grounds of difference and diversity more generally suggests that extreme and moderate outcomes may be the same.

Race relations as a concept has been used widely since the 1960s within law and policy across the world to promote greater understanding and cohesion between minority and majority communities, based on the premise that equal value should be given to all groups. Legal provisions are made and delivered variably across countries and regions, and lack consistency in relation to culturally appropriate welfare services. This is illustrated by the response to the changes in migration patterns within the European Union where large numbers of people are moving between member countries to seek work and education. Nation-states such as the United Kingdom are desperate to stem this flow. To address this, the Government is openly manipulating immigration figures and seeking greater surveillance of students and patients in institutions like the universities and National Health Service (Hutton, 2013). The movement of largely white individuals looking for opportunities across Europe has challenged established ideas of racism defined by color, culture, and religion. Societal fears based on age-old ideas, such as the dilution of indigenous traditions and economic opportunities lost to eager

newcomers, remain close to the surface. Conversely, the real impact of such thinking in the United Kingdom has resulted in the under-recruitment of overseas students to universities and seasonal workers to agricultural industries. Individuals are exercising choice and moving to countries with better entry and education/work permit requirements (Hutton, 2013; Doward, 2013).

Changes in understanding and awareness have led to a greater focus on the more abstract social constructs of culture and racism, or even the combined notion of *culturalist racism* (Lentin, 2000: p. 5). Ideas of biological difference continue to have some meaning, for example, groups being seen to begin from the same reference point and relate to each other purely on the basis of racial difference. Racial difference retains a concrete existence.

The reformulation of terminology has added to the range of words used to describe culturally sensitive practice. Appropriate interventions are recognized as country and context specific and better related to history and geography (Fook, 2006). Evidence from Europe suggests that addressing individual cultural difference is more palatable to the national psyche in countries that are more homogeneous (e.g., Norway). Countries with bigger numbers and a multifaceted population mix (like the United Kingdom) are more likely to define culture in relation to particular groups and target interventions appropriately (Kriz and Skivenes, 2010).

Those who support critical race theory (CRT) suggest that antiracist practice and culturally appropriate responses have failed to provide professionals with the tools to fully engage with oppressive societal structures rooted in past history (Gillborn, 2006). CRT examines the comprehensive power ingrained in structures and systems which were highlighted as institutional racism earlier by writers like Dominelli (1988) and subsequently by the 1999 MacPherson Report on the murder of 18-year-old Stephen Lawrence in the United Kingdom. Although Stephen Lawrence died in 1993, allegations that the police sought to collect evidence through bugged meetings, held by the Lawrence family and friends after his death, continue to be made in 2013 (Syal, 2013). This supports the view that bias and partiality can be used by organizations to exploit minority groups and destabilize race relations.

Discrimination impacts directly on the way some people are perceived, categorized, and responded to within welfare services. For example, groups such as Asians and Hispanics are seen to have above average 'acculturation stressors' (Crunkilton et al., 2005: p. 115) based on strong family bonds (Ceballos and Bratton, 2010) that are likely to result in confused identities and intergenerational conflict.

For culturally sensitive interventions to be meaningful, they must be designed to demonstrate the culturally specific nature of the intervention, the theoretical constructs pertinent to the service user experience, and the specialized professional knowledge required to meet individual or group needs (Jackson et al., 2010). In particular, the link between *race*, migration, and difference needs to be disrupted even when some groups choose social separation based on culture and religion. For example, the majority of established immigrants have a long-term interest in their new country of choice but may still live in exclusive groups where distinctive cultural and religious identities can be located, sustained, and defined (Furness and Gilligan, 2010). Clustering in religious and cultural groups is common among minority and majority groups across the globe. It may occur to protect oneself against external threats such as racism and persecution and/or it may be a means by which families and groups reinforce internal practices.

The Professional Application of Cultural Appropriateness

Social work values have encompassed universal principles and legal duties across the world and nation-states are increasingly required to address social justice by signing treaties such as the International Convention on the Elimination of All Forms of Racial Discrimination, which has been ratified by 168 countries. The Universal Declaration of Human Rights (UDHR), which has also been enshrined in national legislation, provides moral guidance to professional social workers across the world (Banks, 1995).

Several provisions of the UDHR have been incorporated into British law, which has impacted on professional practice. These include the social work Code of Ethics (BASW, 2009), the Professional Capabilities Framework (Social Work Reform Board, 2010), and the Social Work Standards of Proficiency (HCPC, 2012). Social workers registered with professional bodies are signed up to follow both laws and universal principles and are committed to policies that promote difference and diversity (Banks, 2012). However, professional codes of ethics and practice have been criticized for being generalist and universal by nature and too often lacking detailed and specific follow-up guidance on equality procedures, processes, and tasks. The development of cultural competence within these frameworks has been seen as incoherent and disjointed. This has been illustrated by the dilemmas around race-specific vs culturally competent social work practice as highlighted by Patni (2006).

Although rooted in social justice and human rights principles, culturally appropriate interventions have also been used to control and manage some groups. For example, the approach has been used to assimilate newly arrived or established immigrants (Lentin, 2000) and to promote policy and practice based on notions of tolerance (Bhatti-Sinclair, 2011). The emphasis here has been on tolerating newcomers, offering them limited access to an expensive and time-consuming citizenship processes and, if successful, demanding that they take on full civic responsibilities within society.

The theoretical content and specialist knowledge used to intervene in a culturally appropriate manner have been viewed, questioned, and interpreted by social work and social policy scholars across the world for many decades (Augoustinos and Reynolds, 2001; Jackson et al., 2010). For example, cultural racism has been evaluated in the divisive context of contemporary Israel (Ben-Eliezer, 2004).

At the same time, projects which promote pluralistic difference through culturally sensitive responses to particular groups, for example, Latino families and African-American youth (Jackson et al., 2010), have taken place. Initiatives, such as parenting interventions within educational contexts, have played an important role in improving the life chances of school children (Ceballos and Bratton, 2010).

However, the emphasis on family and community has to be considered alongside structural and institutional change as well as meeting fundamental needs, such as income and housing. The assessment of need may require additional consideration of factors such as the impact of immigration on individuals and groups. It is likely that the social worker will need to assess the capacity and resilience of the service user to new experiences and ideas because some people may better retain cultural identities rooted in the past while accepting and adapting to new influences (Sakamoto, 2007). It is the responsibility of the professional to ensure that the behavioral commitment made by the individual to learn and develop is supported by service outcomes that lead to improved life chances, including better employment and education. In an exploration of acculturation relating to immigrant groups, Sakamoto (2007) suggests that service providers are motivated by different goals but that social workers, particularly those committed to antioppressive methods, are likely to understand that social action should lead to structural change. However, they may need guidance on relevant underpinning theory and knowledge framed within specified service objectives.

Culturally Appropriate Interventions

Depending on the context and the service user group, the quality and appropriateness of culturally sensitive intervention can vary. For example, therapeutic responses to substance misuse may follow particular models that are different from traditional therapeutic approaches. Cultural appropriateness requires the use of complex models on alliance building, assessment, collaboration, and sustainability of change (Gallardo and Curry, 2009). Interpretation of culture is likely to be comprehensive and includes wider factors related to immigration and generational influences within the immediate and wider family across countries and continents. Professionals using such methods are more likely to take account of individual and group values, attitudes and behaviors in intervention, monitoring, and evaluation processes (Jackson et al., 2010). Although developments have taken place across the world, the United Kingdom is used to exemplify the complexities of social work practice.

Cultural sensitivity appears to be practiced in various ways, but the most effective approaches are staged to allow time for knowledge-based bonding, cohesion, and sound communication among people (Ceballos and Bratton, 2010; Chand, 2005). This is particularly so in work with vulnerable groups, such as children, who depend on skillful professionals who are able to observe, understand, and influence cultural interactions based on a range of needs rather than a particular difference. Chand (2005) explored the results of the child protection inquiry into the case of Victoria Climbie, an 8-year-old girl from the Ivory Coast living in London (Laming, 2003) whose main carer, when quizzed through a French interpreter, offered particularly poor explanations of childcare responsibilities to social services and the police. Although the professionals had evidence that the carer spoke and understood English, they did not use this knowledge to question, investigate, and act further. Instead of communicating directly with Victoria, they allowed the carer to manipulate and take control of the situation.

The demographic profile of the areas where Victoria lived in London is multicultural. However, the many professionals involved demonstrated little sensitivity or procedural rigor in communication with each other or the carer. The subsequent inquiry found that poor understanding of ethnic influences (i.e., the complex notion of kinship care across international borders) and lack of judgment in the assessment and intervention processes contributed directly to the death of Victoria Climbie (Parton, 2004).

The Climbie case involved many organizations, systems, and professional groups, and highlighted the large amount of poor quality case information (including e-mails, faxes, and paper documents) generated by professionals. The report on the inquiry (Laming, 2003) recommended cohesive, joined-up communication across all child protection agencies. The report was subsequently analyzed at some length. It was acknowledged, for example, that interventions with individuals from a minority background should be carried out by social workers competent in the relevant dialect or language and who have some understanding of relevant words, ideas, and concepts. Professionals also need to be supported by well-trained and managed translation and interpretation services (Chand, 2005).

The Laming Report (Furness and Gilligan, 2010) led to the view that culturally specific communication skills regarding good timekeeping, the use of culturally specific greetings and working with individuals who display humility and deference, should be developed, incorporated, checked out, and acted upon in procedures and processes. Despite the legal and policy influences of the Climbié Inquiry (Parton, 2004), there is little follow-up evidence to suggest that developments on methods that enhance professional engagement have taken place. It may be that insufficient attention to the particularities of culturally sensitive communication and interpersonal skills is a symptom of the national and global development of professional values that, often voluntary, are widely used within social work education (http://www.iassw-aiets.org/global-standards-for-social-work-education-and-training). Social workers may see themselves as competent when working with people originating from countries or regions other than their own. However, professional codes offer general statements that must be supported by context-specific theoretical knowledge and practice evidence (Banks, 2012).

Proponents of reflexive practice in relation to the awareness of self (Fook, 2006) suggest that greater attention needs to be paid to underlying assumptions about how difference is seen, created, and perpetuated in order to improve and develop service interventions. The cultural standards used to provide services for minority groups need to be scrutinized and checked for majority bias. The positive way forward would be to amalgamate and utilize both minority and majority norms within service outcomes (Sakamoto, 2007). This would allow a nuanced, contextualized response to individual needs while taking into account institutional and societal drivers. For example, the use of case files and record keeping by the many welfare organizations involved in the Climbié case received considerable attention (Parton, 2004; Chand, 2005), but not necessarily for cultural appropriateness, cultural currency, or cultural awareness. Chin (2000) recommends a number of ways to develop culturally relevant practice including regular

evaluations of organizational competence and the enhancement of monitoring, access, and quality systems. Detailed scrutiny of population data may also lead to identification of differing needs and highlight factors leading to above average risk for some groups. This may enable service providers to target, prioritize, and respond appropriately. The expectation is that social workers of today are advantaged by well-designed, better-connected systems able to deal with the quality and quantity of electronic communication available to them and to colleagues within and across organizations.

Culturally appropriate or relevant practice can be substantially informed by the use of communication systems such as the Internet search engines and other social media. The knowledge available on cultures is beyond individual reach, but it is likely that contemporary social workers are able to find relevant information quickly and efficiently on a range of issues, such as the impact of the exchange of financial remittances across countries and continents on family income. Social workers safeguarding children and adults have powers to seek knowledge of the individual and his/her ethnicity gained both through face-to-face contact but also through relevant electronically sourced information (e.g., on burial practices). There is limited evidence to suggest that professionals are looking beyond immediate organizational records or using wider media to inform their practice.

Cultural sensitivity suggests the need to base professional judgment on a wide range of information including theoretical, empirical, and practice related (Devore and Schlesinger, 1999). Cultural interventions should be informed by research evidence and data derived from black and ethnic minority individuals, women and other 'ordinary people' Payne (2005: p. 6). The focus should be on what works and what requires further practice and research attention.

Social work policy and organizational change can be bolted on superficially or grounded in well-tested research findings based on consideration of the status quo and the existing cultural norms of the society. Culturally appropriate interventions are multifaceted and require a sound evidence base, but the value of practice wisdom derived from tried and tested methods should not be overlooked. The criteria used should closely relate to the ethnic and cultural understandings of the social problem; for example, the use of indigenous criteria in response to indigenous rehabilitation needs in Australia (Taylor et al., 2010). This approach is likely to acknowledge restrictions and limitations created by situations (Fook, 2003) where the norm is used to benchmark and define other perspectives.

Research evidence from across the world suggests that professionals should not be constrained by procedural constraints and protocols when responding to service user needs. Cultural appropriateness requires the defining, redefining, and adaptation of the social work role and a high level of reflexivity, founded on substantial background knowledge of the issue in question and of the importance of the family and community to the service user. Social workers who develop complex skills in specific expert areas are less likely to pathologize, label, and stereotype.

Gallardo and Curry (2009) explore the importance of personal relationships within Latino groups and how these affect contact with welfare organizations. In a context where sharing information and seeking external support on sensitive topics, such as substance misuse, are seen to be culturally unacceptable by some people, training and support should be available on, for example, the confidential sharing of family history, kinship ties, differences between generations, and the part played by religion and faith in healing processes.

Bernal and Saez-Santiago (2006) suggest cultural centeredness as a better way to understand difference and diversity, as culture can mean many things to many people, including minority groups, who can be as heterogeneous as majority populations. The definition, design, and application of culturally appropriate methods is less problematic if defined in this way within, for example, mental health diagnoses and treatment that can be seen and defined differently by members of the same community.

Social work across the world is keen to develop systemized approaches that better define needs and as a result respond to service users more efficiently and effectively while meeting individual needs in a sensitive manner. The scientific challenge to this is the move in some American states to improve service outcomes through compulsory use of the empirically supported intervention (ESI) guidelines, which promote technological advances to improve service outcomes. ESI has been evaluated for cultural appropriateness by McBeath et al. (2010) who found that the approach is overly clinical and highly procedural. The methods are designed to achieve generic outcomes that do not easily adapt to context, time, and place or follow the particularities of core social work principles, such as client self-determination and respect. The model is a response to the pressure in the United States for large-scale, low-resource interventions, which can and should be replicated in a range of settings and contexts.

Probable Future Directions of Research and Theory

Social workers across the world are engaged on a daily basis in developing theory and practice methods. However, evidence suggests that the focus on culture and ethnicity continues to require professional attention. Knowledge on difference and diversity is still based on majority norms across the globe (Fook, 2006) taking little heed of history and experience relating to factors such as colonization. The literature found for this article on cultural interventions is derived from a range of disciplines but the perspective has in general been that of the professional or academic rather than the service user. This suggests that the views of those receiving services need to be better embedded in practice, research, and theory building.

Another gap relates to the direct links between the social worker's personal history/experience and her/his professional behavior. Although professionals interpret policy on a daily basis and implement targeted, context-specific services, there is little evidence of how they internalize cultural sensitivity.

Families and individuals in need of sensitive interventions are likely to be international by nature and, although movement between countries is important for governments seeking to control welfare, social work should be about recognizing and acting on improving the life chances of such individuals. Too often, the intersections between income, class, and immigrant status are missed leading to poor service outcomes

and institutional improvements. Children, in particular, need confident professionals who are not diverted by color and language and can use human skills to seek and implement change within complex care or family situations.

Attention needs to be paid to the constructs used to define groups in research, in relation to how difference is understood and propagated within professional knowledge. Further evidence is needed on types of interventions, where they originate from and who influenced their developments. While the monitoring and evaluation of processes and methods of intervention are important, practice outcomes and research outputs also require scrutiny for majority norms and cultural bias.

In conclusion, cultural appropriateness suggests that methods and models should vary in range and applicability in line with the populations in question, that is, the combination of scientific methods with individualized and community-based service interventions. Cultural specificity can be of particular concern in one-to-one interactions with service users who may or may not be compliant and/or lack confidence. Within these situations, well-trained social workers can ask pertinent questions, look for information beyond the immediate and find ways around group or family sensitivities in everyday practice. Skills such as these should be fine-tuned rather than overproceduralized. Social work is a relationship-based discipline, which depends on sound intervention strategies, to promote individual change for those in need. This needs to be supported by a societal commitment to macrolevel systemized measures, such as good housing, education, and income, to maximize the life chances of all citizens.

See also: Afrocentric Approaches to Social Work; Antisocial Behavior; Critical Social Work Practice; Social Work Theory.

Bibliography

Ali, T., Barsamian, D., 2005. Speaking of Empire and Resistance. The New Press, New York.

Augoustinos, M., Reynolds, K.J., 2001. Understanding Prejudice, Racism and Social Conflict. Sage, London.

Aveling, N., 2007. Anti-racism in schools: a question of leadership? Discourse: Studies in the Cultural Politics of Education 28 (1), 69–85.

Banks, S., 1995. Ethics and Values in Social Work. BASW/Macmillan, Basingstoke.

Banks, S., 2012. Ethics and Values in Social Work. BASW/Palgrave/Macmillan, Basingstoke.

BASW, 2009. The Code of Ethics for Social Work. Available on: British Association of Social Workers. www.basw.co.uk (accessed 27.05.12.).

Ben-Eliezer, U., 2004. Becoming a black Jew: cultural racism and anti-racism in contemporary Israel. Social Identities 10 (2), 245–267.

Bernal, G., Saez-Santiago, E., 2006. Culturally centered psychosocial interventions. Journal of Community Psychology 34 (2), 121–132.

Bhatti-Sinclair, K., 2011. Anti-racist Practice in Social Work. Palgrave/Macmillan, Basingstoke.

Brah, A., 1996. Cartographies of Diaspora – Contesting Identities. Routledge, London.

Cashmore, E., 1996. Dictionary of Race and Ethnic Relations. Routledge, London.

Ceballos, P.L., Bratton, S.C., 2010. Empowering Latino families: effects of a culturally responsive intervention for low-income immigrant Latino parents on children's behaviours and parental stress. Psychology in the Schools 47 (8), 761–775.

Chand, A., 2005. Do you speak English? Language barriers in child protection social work with minority ethnic families. British Journal of Social Work 35, 807–821. http://dx.doi.org/10.1093/bjsw/bch205.

Chin, J.L., January/February 2000. Culturally Competent Health Care, Public Health Reports, vol. 115, pp. 25–33.

Crunkilton, D., Paz, J.J., Boyle, D.P., 2005. Culturally competent interventions with families of Latino youth at risk for drug abuse. Journal of Social Work Practice in the Addictions 5 (1–2), 113–131.

Devore, W., Schlesinger, E.G., 1999. Ethnic Sensitive Social Work Practice. Allyn & Bacon, New York.

Dominelli, L., 1997. Sociology for Social Work. Macmillan, Baskingstoke.

Dominelli, L., 2002. Anti-oppressive Practice. Palgrave Macmillan, Basingstoke.

Dominelli, L., 2008. Anti-racist Socialwork, third ed. Palgrave/Macmillan, Basingstoke. First published 1988; 1995.

Dominelli, L., Thomas Bernard, W., 2003. Broadening Horizons: International Exchanges in Social Work. Ashgate, Aldershot.

Doward, J., May 25, 2013. Migrant Jobs Squeeze Alarms UK Fruit Farmers, the Observer. London. http//www.guardian.co.uk. (accessed 26.06.13.).

Fook, J., 2003. On the problem of difference… a response to Deborah Padgett's review of Ian Shaw's and Nick Gould's qualitative research in social work. Qualitative Social Work 2 (3), 359–363.

Fook, J., 2006. Uncertainty: the defining characteristic of social work? In: Lumbery, M., Postle, K. (Eds.), Social Work: A Companion to Learning. Sage, London.

Furness, S., Gilligan, P., 2010. Religion, Belief and Social Work: Making a Difference. The Policy Press, Bristol.

Gallardo, M.E., Curry, S.J., 2009. Shifting perspectives: culturally responsive interventions with Latino substance abusers. Journal of Ethnicity in Substance Abuse 8, 314–329.

Gillborn, D., 2006. Critical Race Theory and education: racism and anti-racism in educational theory and praxis. Discourse: Studies in the Cultural Politics of Education 27 (1), 11–32.

Gilroy, P., 1995. There Ain't No Black in the Union Jack – The Cultural Politics of Race and Nation. Routledge, London.

Gobineau, J., 1953. Essai sur l'Inegalite des Races Humains, fourth ed. Firman et Didot, Paris.

Gray, M., Coates, J., Yellow Bird, M. (Eds.), 2008. Indigenous Social Work around the World: Towards Culturally Relevant Education and Practice. Ashgate, Aldershot.

HCPC (Health and Care Professions Council), 2012. Standards of Proficiency: Social Workers in England. London.

Hernstein, R., Murray, C., 1994. The Bell Curve: Intelligence and Class Structure in American Life. Free Press, New York.

Hutton, W., May 26, 2013. Stop Playing Crude Numbers Games with Immigration. We Must be Open. http://www.guardian.co.uk. (accessed 26.06.13.).

IASSW (International Association of Schools of Social Work), Global Standards. http://www.iassw-aiets.org/global-standards-for-social-work-education-and-training (accessed 27.06.13.).

Jackson, K.F., Hodge, D.R., Vaughn, M.G., 2010. A meta-analysis of culturally sensitive interventions designed to reduce high-risk behaviours among African American youth. Journal of Social Service Research 36, 163–173. http://dx.doi.org/10.1080/01488371003697780.

Kriz, K., Skivenes, M., 2010. Knowing our society and fighting against prejudices: how child welfare workers in Norway and England perceive the challenges of minority parents. British Journal of Social Work 40, 2634–2651. http://dx.doi.org/10.1093/bjsw/bcq026.

LeCourteur, A., Augoustinos, M., 2001. The language of prejudice and racism. In: Augoustinos, M., Reynolds, K.J. (Eds.), Understanding Prejudice, Racism and Social Conflict. Sage, London.

Laming, Lord, 2003. The Victoria Climbie Inquiry: Report of an Inquiry by Lord Laming, Cmnd 5730. The Stationery Office, London.

Law, I., 2010. Racism and Ethnicity: Global Debates, Dilemmas and Directions. Pearson, Harlow, Essex, England.

Lentin, A., March 2000. 'Race' racism and anti-racism: challenging contemporary classifications. Social Identities 6 (1), 91–106.

Lum, D., 1997. Culturally Competent Practice. Brook/Cole, Belmont, CA.

McBeath, B., Briggs, H.E., Aisenberg, E., 2010. Examining the premises supporting the empirically supported intervention approach to social work practice. Social Work 55 (4), 347–357.

Paradies, Y., 2005. Anti-racism and indigenous Australians. Analysis of Social Issues and Public Policy 5 (1), 1–28.

Parton, N., 2004. From Maria Colwell to Victoria Climbie: reflections on public inquiries into child abuse a generation apart. Child Abuse Review 13, 80–94. http://dx.doi.org/10.1002/car.838.

Patni, R., 2006. Race-specific vs. culturally competent social workers: the debates and dilemmas around pursuing essentialist or multicultural social work practice. Journal of Social Work Practice 20 (2), 163–174.

Payne, M., 2005. The Origins of Social Work: Continuity and Change. Palgrave/Macmillan, Basingstoke.

Sakamoto, I., 2007. A critical examination of immigrant acculturation: toward an anti-oppressive social work model with immigrant adults in a pluralistic society. British Journal of Social Work 37, 515–535. http://dx.doi.org/10.1093/bjsw/bcm024.

Sivanandan, A., 1991. Black struggles against racism. In: Setting the Context for Change, Central Council for Education and Training in Social Work. CCETSW, Leeds.

Social Work Reform Board, December 2010. Building a Safe and Confident Future. www.education.gov.uk/swrb.

Solomos, J., Back, L., 1995. Race, Politics and Social Change. Routledge, London.

Syal, R., June 26, 2013. Further Stephen Lawrence Inquiry Possible Says David Cameron. http:/www.guardian.co.uk (accessed 28.06.13.).

Taylor, K., Thompson, S., Davis, R., 2010. Delivering culturally appropriate residential rehabilitation for urban indigenous Australians: a review of the challenges and opportunities. Australian and New Zealand Journal of Public Health 34 (S1), 36–41.

Walker, S., 2002. Culturally competent protection of children's mental health. Child Abuse Review 11, 380–393. http://dx.doi.org/10.1002/car.768.

Culture and Actor Network Theory

Ignacio Farías, WZB Berlin Social Science Center, Berlin, Germany
Sophie Mützel, University of Lucerne, Lucerne, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines actor-network theory's (ANT) complex relationships with notions of culture. We start with a discussion of ANT's fundamental works and show how these question common notions of culture. In the second part, we show how current developments in ANT, which empirically focus on arts, markets, care, and democracy, strongly rely on a pragmatist approach to practices. The third part outlines ANT's methodological and theoretical contributions for cultural analysis. In sum, as a perspective ANT suggests a reformatting of what cultural analysis might entail.

Actor-network theory (ANT) as a "theory that maps the social relations between people, objects, and ideas, treating all as agentic entities that form a broad network" (Cerulo, 2009: p. 533) has not only contributed to science and technology studies. Over the last 30 years, ANT has also become an innovative analytical framework for the study of culture. Yet we are confronted with an apparent paradox: while profoundly impacting the study of culture, ANT restrains from referring to the notion of culture, be it as *explanans*, as *explanandum*, or as phenomenon.

This article examines ANT's complex and paradoxical relationships with notions of culture and highlights ANT's implications for cultural sociology. This is a challenging task, not just because culture has been the subject of hundreds of different definitions in the social sciences, but also because ANT remains "an open building site, not a finished and closed construction" (Callon, 2001: p. 65). We start with a discussion of ANT's fundamental works and conceptual repertoires. We show how these works question common notions of culture, including culture as a human realm (as opposed to a natural realm), as a historically constructed mind-set, as semiotic webs of meanings, and as a shared normative orientation to social action. In the second part, we show how current developments in ANT, which empirically focus on arts, markets, care, and democracy, strongly rely on a pragmatist approach to practices. The third part outlines methodological and theoretical contributions for cultural analysis. In sum, as a perspective ANT suggests a reformatting of what cultural analysis might entail.

Networks, Not Culture: A Sociotechnical Approach to Science and Technology

One of the seminal works in the development of ANT is Bruno Latour and Steve Woolgar's *Laboratory Life. The social construction of scientific facts* originally published in 1979. Based on ethnographic research Latour conducted between 1975 and 1977, *Laboratory Life* is an attempt to describe 'the esoteric culture of the scientific laboratory' (Latour and Woolgar, 1986: p. 275). It was inspired by a long anthropological tradition that dealt with the cultural underpinnings of cognition and thereby unveiled the arbitrariness of the Great Divide between the West and the rest. Accordingly, its aim is to apply the same field methods and analytical strategies used to study, for example, Ivory Coast farmers to study first-rate scientists. The book begins to formulate at least three key analytical and methodological principles of what later on would be called 'actor-network theory.'

First, *Laboratory Life* involves elaborating an agnostic description of routine work in the laboratory, without assuming any a priori relationship between the social and the scientific. On the contrary, the intention is to demonstrate how in the actual process of producing scientific knowledge, the very boundary between context and content of science is coproduced. Thus, it problematizes at least two extended types of sociocultural approaches to the study of science. On the one hand, it departs from contextual approaches, which assume that the content of science cannot be the subject of sociological analysis. Such contextual approaches focus instead on institutional frameworks of scientific activity, careers, and status conflicts among scientists or, as Merton (1979) proposes, on the scientific ethos and normative system that would guarantee the strict application of the scientific method. On the other hand, *Laboratory Life* also departs from oversocialized approaches that explain the content of science, its discoveries, facts, and paradigms, by reference to broader societal structures. One example of this is Bloor's 'strong program' (1976) for the study of science, which points out socially and historically produced categories and beliefs that underlie the scientific interpretation of empirical experience. Similarly, *Laboratory Life* presents an alternative to Foucault's (1969) archeology of *epistemes*, understood as the fundamental modes of reasoning that rule the formation and transformation of knowledge for a whole epoch, and the resulting equation of knowledge and power.

This distancing from both types of sociocultural approaches to science eventually led to the erasure of the adjective 'social' in *Laboratory Life*'s subtitle in subsequent editions (1986). In doing so, the authors also repudiate social constructionism. In the mid-1980s, social constructionism was fundamental for the rejuvenation of cultural sociology. Suggesting an alternative to traditional sociological conceptions of understanding culture as values, it understands culture to consist of "symbolic vehicles of meaning, including beliefs, ritual practices, art forms and ceremonies" (Swidler, 1986: p. 273). Culture thus serves as a 'tool kit' people use to solve problems. From a *Laboratory Life*

perspective, this new sociology of culture overestimates the symbolic and assumes context and structure, whereas in fact such divisions do not exist.

Second, *Laboratory Life* reveals the constitutive, agentic role nonhuman entities play in the production of scientific knowledge. This becomes evident in the study of inscriptions (e.g., maps, graphs, illustrations) of 'natural' objects (e.g., a hormone, a soil type, a distant galaxy) produced by technical instruments. Latour and Woolgar (1986) describe laboratory work as the systematic production of experiments and other technical procedures to inscribe or write down the behavior or reactions of 'natural' objects into material and visual media. Furthermore, it also involves translation of visual inscriptions into scientific propositions. Inscriptions thus help to establish comparability, make interpretation possible, and allow for making scientific statements about the natural world. In their ethnographic accounts, they show that nonhuman entities enable and transform chains of reference that constitute scientific facts, and thus argue that distinctions between facts and artifacts need to be overcome.

Michel Callon (1986) condenses these insights into the analytical principle of *generalized symmetry*, according to which the same conceptual repertoires should be used to describe the action of human and nonhuman entities. This radical position, later referred to as 'material semiotics' (Law, 2009), aims to understand phenomena that are completely overlooked by interpretative approaches in sociology and anthropology. Whereas the latter focus on deconstructing cultural texts and webs of meaning (e.g., Geertz, 1985), generalized symmetry describes relationships between heterogeneous, material, and nonsymbolic entities. ANT thus is also incompatible with semiotic readings, for example of systems of objects (Baudrillard, 1996) or fashion (Barthes, 1990).

Third, *Laboratory Life* begins to develop a 'flat' understanding of the circulation of scientific facts and, generally, the social. A key finding was that the robustness of scientific facts depends on the circulation of their inscriptions. Scientific facts move from one laboratory to the next, from there to a journal, then to a policy document, to a ministry office, to an industrial floor, from there to a market, to consumers, to concerned groups, and so on. In their circulation through such heterogeneous local points, scientific facts are constantly tested and contested by other actors. Provided that they endure after different actors have probed into them, they become indisputable scientific facts or, in ANT's terms, 'immutable mobiles' (Latour, 1987), i.e., networks of relationships among heterogeneous elements that remain stable despite their circulation through such diverse local points. The finding that scientific facts exist within such translocal sociotechnical networks led ANT scholars to a radical conclusion: the larger the network, the more objective the scientific fact.

According to this perspective then there is only a 'flat' single-level reality, consisting of larger or smaller networks of always local points. There are no hidden structures, no global levels, or ultimate realities. This leads to a further rejection of common anthropological and sociological understandings of culture(s) as locally bounded entities, knowledge forms, or social realities. For ANT, there is neither the global nor the local, not even a 'glocal society' understood as simultaneously shaped by global and local processes. Instead of a separation of levels, the key question for ANT involves the extension of the socio-material and sociotechnical networks within which social phenomena are constituted.

During the 1980s, Bruno Latour, Michel Callon, and John Law further developed these and other principles collaboratively in what began to be called 'actor-network theory.' This composite notion aims to describe the hybrid and translocal relations that constitute every technological artifact, scientific fact, and social actor. Accordingly, every phenomenon or action is created by a whole network of heterogeneous entities. These entities are called *actants*, since they participate in an actor-network and thus in a collective and distributed action. The notion of *actor-network* was coined to make the tracing of an analytical distinction between an actor and a network impossible. To be sure, a network in ANT's terminology does not define the context in which an actor acts and interacts, as in classic social network analysis and its notion of embeddedness. Rather, any actor is an actor-network since no analytical separation between actor and network exists. Any capacity to act stems from the heterogeneous assemblage of human and nonhuman entities, constituting the actor-network.

One of the key processes ANT describes involves the formation of such actor-networks. In the sociology of *translation*, Callon (1986) describes the sequential processes through which entities and networks cocreate each other enabling new forms of action. It begins with the *problematization* of a situation, i.e., transforming a complex and ambiguous situation in which many entities are involved into a well-defined problem. The second step in the translation process is *interessement*, i.e., generating interest among other entities in the problem defined in the first step, so that other entities recognize this to be their problem as well. Third, Callon speaks of *enrolment* to denote how other human and nonhuman entities are transformed into allies, when their interests and identities are stated in terms of the common problem. Finally, *mobilization* describes the collective action that the translation of these different entities enables. A related concept to describe successful translation is that of *blackboxing*. When a whole network is mobilized in the name of one problem defined by one actor, its heterogeneity and distributed character is blackboxed, so that it seems that one bounded actor carries out the action. Indeed, ANT understands itself as a research program oriented toward opening up blackboxes that hide the hybrid constitution of the social world.

Bruno Latour's *We have never been modern* (1993) most clearly highlights ANT's position toward and its analytical consequences for the study of culture. Drawing upon contemporary anthropological theory, Latour sets out to open the two separate blackboxes of nature and culture, which underlie the constitution of modernity: "the very notion of culture is an artifact created by bracketing Nature off. Cultures – different or universal – do not exist, any more than Nature does. There are only nature-cultures" (p. 104).

In sum, ANT's negative argument characterizes culture as neither contextual nor symbolic, neither explanatory nor autonomous.

After ANT: A Pragmatist Turn in the Study of Practice

With the exception of the explicit rejection of culture in *We have never been modern*, ANT mostly argued implicitly against notions of culture that exclusively refer to the human realm or to a symbolic system providing normative orientation for action. At the same time, ANT's early focus on ethnographic description of sociotechnical practices, highlighting the practical assemblage not just of science and technology, but also of the empirical and the real, positions it within a pragmatist research tradition. While the focus on practices was not central at the time of theory development around the notion of actor-network in the 1980s, it moved center stage in the 1990s. In part, this was a response to the severe criticisms of the notion of actor-network, including that it denies power issues, exhibits male Eurocentricism, and sustains a managerial-functionalist understanding of social phenomena (e.g., Star, 1991). In order to emphasize a new phase of theoretical development that addresses such issues, while leaving behind the conceptual repertoires surrounding the notion of actor-network, Law and Hassard (1999) have even suggested speaking of an 'after-ANT' phase.

Since the mid-and late-1990s, there have been various developments using ethnographic and pragmatist approaches to study the assemblage of objects, spaces, and figures in fields beyond science and technology, while sticking to ANT's analytical principles explained in Section Networks, Not Culture: A Sociotechnical Approach to Science and Technology. As in pragmatism (Dewey, 2004[1916]), these works approach practical activity and experience not as a result of human will or social structure, but rather in terms of its consequences for the constitution of reality. As evident in *Reassembling the social* (Latour, 2005a), ANT staunchly rejects Pierre Bourdieu's (1990) practice theory. The general criticism of Bourdieu hinges upon his idea that 'hidden social structures' predetermine actors' positions in the social world and endow them with a practical sense. Rather than understanding social practices as expressions of internalized social structures, ANT's pragmatist approach studies actual practices in concrete sites and situations, so that the entanglements of human and nonhuman entities enabling certain forms of action can be examined. Four key contributions testify to this interest in practices beyond practice theory. They relate to the notions of attachment, performativity, enactment, and engagement.

In his work on music, Antoine Hennion (1993) studies the practices of the production of aesthetic value and thus the results of artistic production along the entanglements of human actors and material objects. Hennion points out that traditional sociology has assigned a peculiar role to objects in the study of cultural products: while technoscientific objects are often reduced to a technical dimension, cultural objects are conceived of as social intermediaries, i.e., symbols, meanings, or value-laden signs. Thus, traditional sociology views art works as the material production of ourselves as a collective entity. The transformations, resistances, and mediations arising from the material or technical existence of art works are entirely absent. Hennion's work, in contrast, shows that material objects, e.g., instruments, notation systems, recording technologies, reproduction devices, and music halls, are 'mediators.' Such mediators are "neither mere carriers of the work, nor substitutes which dissolve its reality; they are the art itself" (2003: p. 84). Thus, this approach emphasizes the coproduction of cultural products, producers, and audiences. Furthermore, in rejecting Bourdieu's notions of habitus and taste formation, Hennion shows how taste develops as a reflexive practice in situations, in which actors attach themselves to objects and develop abilities to perceive and experience them. Taste, understood from an ANT perspective, is thus a situated performance rather than the result of socially structured preferences (Hennion, 2007).

ANT's criticism of culture as context and its pragmatist turn can also be found in Callon's research program for the study of markets. Arguing against the primary focus on the sociocultural embeddedness of economic action, Callon's (1998b) seminal contribution expands economic sociology to now include the notion that "economics make the economy." He thus coins the idea of the 'performativity of economics.' Accordingly, economics should not be seen as providing true or false descriptions of economic processes. Rather, economics should be understood in terms of constituting economic worlds within which economic models, formulas, and even figures such as *homo oeconomicus*, can be said to exist: no economy without economics. Callon (2007) shows that the performative force of economics does not stem from a purely linguistic or symbolic effect, but rather from the constitution and mobilization of sociotechnical networks or *agencements*. Similar to pragmatist notions, Callon (1998a) points to dynamics of framing/overflowing in order to understand economic goods and agents, prices and transactions as relational and situational achievements. Expanding Goffman's (1974) frame analysis, Callon emphasizes that market frames depend on the sociomaterial arrangements of objects, devices, formulas, and technologies rather than on shared taken for granted assumptions and normative orientations. Such frames then enable economic calculations. However, since overflowing instead of framing is the rule in market dynamics, studying economization practices and processes requires being "acutely attentive to the plurality and open-endedness of 'the economic' as it is brought into being" (Caliskan and Callon, 2010: p. 2).

Another study of practices in concrete sites is Annemarie Mol's (2002) *The body multiple*. In her empirical study on medical practices, she finds that the body, going through different units in a hospital, is a multiple object. Its multiplicity is ontological and involves different practices of bringing bodies into being; this multiplicity does not result from different perspectives, ways of observing, and knowing a single body. Mol proposes the notion of enactment as a means of understanding how objects exist through multiple situated practices. Thus, she also distances herself from the classic ANT approach to immutable mobiles, in which a stabilized result stems from a process of construction. Enactment instead is situational and it reveals the coexistence of multiple stabilizations. Thus, the practical challenge that Mol describes with regards to bodies in hospitals is not one of translation, but rather one of coordination, i.e., how multiple enactments are brought together or kept apart, in order to arrive at a diagnosis.

A further connection to a pragmatist tradition can be seen in works using John Dewey's (1927) ideas on democratic publics. As Dewey points out, the main problem Western democracies confront is that they take the existence of a public

for granted. Instead, publics are formed around complex issues involving unintended consequences, which escape institutions and cannot be managed. According to Latour (2005b), this pragmatist understanding of democratic publics reveals that politics cannot be considered as a purely human matter involving conflicting ideals, norms, values, or interests. Rather, these publics indicate that the hybrid objects of science, technology, medicine, or urban development play a generative role in the constitution of publics and in enabling political activity. Thus, Latour speaks of an *object-oriented politics*. In addition, Marres (2009) explores the role played by practices and devices of public engagement. Her work demonstrates that this involves studying a broad range of practices of issue formation beyond formal participatory events and procedural settings, which are constrained by set issue definitions. Understanding public engagement requires more than just studying discursively constructed frames. Rather, public involvement needs to be studied and explained in terms of the actual attachments between publics and objects that define such public issues.

It is important to note that none of these new lines of ANT research explicitly use the notion of culture. However, as we have shown, they propose pragmatist readings of practices of assembling, enacting, publicizing, attaching, mediating, and performing realities.

Nonhumans, Description, and Ontology: Cultural Analysis after ANT

In conclusion, three general observations regarding ANT's methodological and theoretical contributions for cultural analysis follow:

In general terms, ANT has arguably heightened awareness and sensitivity toward the active role nonhumans, including animals, technologies, objects, or materials play in the phenomena and dynamics studied by social sciences. ANT approaches are thus not just increasingly mobilized for the study of overtly hybrid phenomena, such as carbon transitions, nuclear crises, or the quest for sustainable development, but are also translated as a more general methodological and analytical strategy. One translation can be found in the works of Luc Boltanski and Laurent Thévenot (2006), two pragmatic sociologists. Studying the justifications actors use in disputes and disagreements, they show that the involved actors do not merely rely on words. Rather, in evaluating a situation, actors bring together words, as well as human and nonhuman actors, in tests and thus shape different orders of worth. Economic and cultural sociology have particularly picked up on these insights and are currently analyzing processes of valuation and the constitution of different orders of worth, including nonhuman actors in their analyses (Fourcade, 2011; Stark, 2009). In the field of urban studies, ANT perspectives are particularly helpful to grasp the sociomaterial assemblages of cities and are at the center of current debates (Farías and Bender, 2009; Brenner et al., 2011).

A second important aspect relates to the role of description in ANT accounts. The sensitivity to nonhumans is a result of ANT's methodological commitment to produce detailed descriptions of actual activity, as delineated in the case of the laboratory ethnographies. But descriptions can also be an analytical strategy, allowing social scientists to question traditional accounts based on contexts or causal explanation. Accordingly, the challenge is to describe how actors create contexts for their own action and how some events lead to other events. Thus, contexts and causes become empirical objects to be described. In turn, description becomes a methodological and theoretical device, as opposed to positivist visions of descriptions as simplistic accounts and 'just-so-stories.' Therefore, ANT contributes to a larger sociological movement of 'descriptive sociology' (Savage, 2009), which is also prominent in pattern analyses over time, and in relational methods that make use of visualizations (Mützel, 2009). This usage of description as an analytical tool means that, for example, cultural products, such as art or fashion, should not be explained as a result of the culture of a certain class, city, or nation. Rather, social scientists need to describe and analyze how cultures are being practically assembled through such products (Entwistle and Slater, 2013).

Yet another contribution of such a commitment to meticulous descriptions of whole worlds is a turn toward ontological questions. The strict separation of culture from nature and the social from the material, so strongly underlying social sciences, has confined these questions to the realm of epistemology. While focusing on human beliefs, attitudes, images, and representations of the world, questions about ontology have generally been neglected by both positivist and constructivist social sciences. ANT's contribution is to direct attention toward the study of how reality comes into being.

See also: Actor-Network Theory; Culture and Networks; Culture, Cognition and Embodiment; Networks and Meaning; Pragmatist Social Thought, History of; Science and Technology Studies, History of; Social Ontology.

Bibliography

Barthes, Roland, 1990. The Fashion System. University of California Press, Berkeley, Los Angeles, CA.

Baudrillard, Jean, 1996. The System of Objects. Verso, London, New York.

Bloor, David, 1976. Knowledge and Social Imagery. University of Chicago Press, Chicago.

Boltanski, Luc, Thévenot, Laurent, 2006. On Justification. Economies of Worth. Princeton University Press, Princeton, Oxford.

Bourdieu, Pierre, 1990. The Logic of Practice. Stanford University Press, Stanford.

Brenner, Neil, Madden, David J., Wachsmuth, David, 2011. Assemblage urbanism and the challenges of critical urban theory. City 15 (2), 225–240.

Caliskan, Koray, Callon, Michel, 2010. Economization, part 2: a research programme for the study of markets. Economy and Society 39 (1), 1–32.

Callon, Michel, 1986. Some elements of a sociology of translation: domestication of the scallops and the fishermen of St-Brieuc Bay. In: Law, John (Ed.), Power, Action, and Belief: A New Sociology of Knowledge? Routledge and Kegan Paul, London, pp. 196–233.

Callon, Michel, 1998a. An essay on framing and overflowing: economic externalities revisited by sociology. In: Callon, Michel (Ed.), The Laws of the Market. Blackwell Publishers, Oxford, pp. 244–269.

Callon, Michel, 1998b. The embeddedness of economic markets in economics: introduction. In: Callon, Michel (Ed.), The Laws of the Market. Blackwell Publishers, Oxford, pp. 1–57.

Callon, Michel, 2001. Actor network theory. In: Smelser, Neil, Bates, Paul (Eds.), International Encyclopedia of the Social and Behavioral Sciences, first ed. Pergamon, Oxford, pp. 62–66.

Callon, Michel, 2007. What does it mean to say that economics is performative? In: MacKenzie, Donald, Muniesa, Fabian, Siu, Lucia (Eds.), Do Economists Make Markets? On the Performativity of Economics. Princeton University Press, Princeton, pp. 311–357.
Cerulo, Karen, 2009. Nonhumans in social interaction. Annual Review of Sociology 35, 531–552.
Dewey, John, 2004[1916]. What pragmatism means by practical. In: Dewey, John (Ed.), Essays in Experimental Logic. Dover Publications, Mineola, New York, pp. 192–209.
Dewey, John, 1927. The Public and Its Problems. Holt, New York.
Entwistle, Joanne, Slater, Don, 2013. Reassembling the cultural: fashion models, brands and the meaning of 'culture' after ANT. Journal of Cultural Economy. http://dx.doi.org/10.1080/17530350.2013.783501.
Farías, Ignacio, Bender, Thomas (Eds.), 2009. Urban Assemblages. How Actor-network Theory Changes Urban Studies. Routledge, London, New York.
Foucault, Michel, 1969. Archaeology of Knowledge. Routledge, London.
Fourcade, Marion, 2011. Cents and sensibility: economic valuation and the nature of "nature". American Journal of Sociology 116 (6), 1721–1777.
Geertz, Clifford, 1985. The Interpretation of Cultures. Basic Books, New York.
Goffman, Erving, 1974. Frame Analysis. An Essay on the Organization of Experience. Penguin Books, Harmondsworth.
Hennion, Antoine, 1993. La Passion Musicale. Métailié, Paris.
Hennion, Antoine, 2003. Music and mediation: towards a new sociology of music. In: Clayton, Martin, Herbert, Trevor, Middleton, Richard (Eds.), The Cultural Study of Music: A Critical Introduction. Routledge, London, pp. 80–91.
Hennion, Antoine, 2007. Those things that hold us together: taste and sociology. Cultural Sociology 1 (1), 97–114.
Latour, Bruno, 1987. Science in Action. Open University Press, Milton Keynes.
Latour, Bruno, 1993. We Have Never Been Modern. Harvard University Press, Cambridge.
Latour, Bruno, 2005a. Reassembling the Social: An Introduction to Actor-network Theory. Oxford University Press, Oxford, New York.
Latour, Bruno, 2005b. From realpolitik to dingpolitik or how to make things public. In: Latour, Bruno, Weibel, Peter (Eds.), Making Things Public. Atmospheres of Democracy. ZKM, MIT Press, Karlsruhe, Cambridge, pp. 14–43.
Latour, Bruno, Woolgar, Steve, 1986. Laboratory Life. The Construction of Scientific Facts. Princeton University Press, Princeton.
Law, John, 2009. Actor network theory and material semiotics. In: Turner, Bryan S. (Ed.), The New Blackwell Companion to Social Theory. Blackwell, Oxford, pp. 141–158.
Law, John, Hassard, John (Eds.), 1999. Actor Network Theory and after. Blackwell, Oxford, Malden, MA.
Marres, Noortje, 2009. Testing powers of engagement. Green living experiments, the ontological turn and the undoability of involvement. European Journal of Social Theory 12 (1), 117–133.
Merton, Robert, 1979. The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago Press, Chicago.
Mol, Annemarie, 2002. The Body Multiple: Ontology in Medical Practice. Duke University Press, Durham, NC.
Mützel, Sophie, 2009. Networks as culturally constituted processes: a comparison of relational sociology and actor-network theory. Current Sociology 57 (6), 871–887.
Savage, Mike, 2009. Contemporary sociology and the challenge of descriptive assemblage. European Journal of Social Theory 12 (1), 155–174.
Star, Susan Leigh, 1991. Power, technologies and the phenomenology of standards: on being allergic to onions. In: Law, John (Ed.), A Sociology of Monsters: Essays on Power, Technology, and Domination. Routledge, London, New York, pp. 27–57.
Stark, David, 2009. The Sense of Dissonance: Accounts of Worth in Economic Life. Princeton University Press, Princeton.
Swidler, Ann, 1986. Culture in action: symbols and strategies. American Sociological Review 51 (2), 273–286.

Culture and Economic Development

Eelke de Jong, Radboud University, Nijmegen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by F. Fukuyama, volume 5, pp. 3130–3134, © 2001, Elsevier Ltd.

Abstract

Both historical studies and econometric analyses find that high levels of economic growth are associated with values such as achievement motivation, future orientation, and eagerness to learn. Opinions differ with respect to the direction of the causal relation, if any, and the role of formal institutions. Some regard formal institutions as of more importance than values. Future research should focus on the time pattern of changes in values, institutions, and economic performance and take a close look at the influence of power.

Introduction

Max Weber's the 'Protestant Ethic and the Spirit of Capitalism' is regarded by many as the classical work on the relation between religion and economic development. Often the implicit suggestion is that it forms the start of research on the relation between culture – in the sense of norms and values – and economic growth. In fact, it marks the start of a period in which economic research began to ignore culture. From the time of Adam Smith, who is considered to be the founder of the economic science, until the beginning of the twentieth century, culture was part of economic science, although it was not often explicitly mentioned as such. Weber's work can be seen as the study, which made culture more explicit. From then onward culture was part of anthropology and sociology, and economics developed into a study of human beings optimizing their objective function under the constraints of available resources. After World War II, the discipline was dominated by mathematical and econometric approaches. These quantitative studies left no room for vague concepts such as culture. The subfields of development economics and international business remained exceptions in that they still paid attention to culture.

The dominant post–World War II model of economic growth has capital and labor as inputs. Only the quantities of these production factors matter. Later models of economic growth also take the quality of these inputs into account. In particular the quality of labor can be different between countries depending on the quality of the countries' educational system. This opens the possibility of including as explanatory variables the quality of the educational system and of national institutions. Recently, this has lead to models including culture as one of the determining factors of growth.

The factors that lead to a reemergence of culture in economics can be divided into two groups: developments in the economy and a disappointment with the results of the mathematical models. In the beginning of the 1990s, it became clear that from 1960 onward, some East Asian countries had shown high levels of economic growth. Many were surprised by the success of these newly industrialized countries (NICs). These countries (Taiwan, South Korea, Hong Kong, Singapore, and Malaysia) were not those which in the mid-1950s were mentioned as candidates for success. This unexpected success of the NICs triggered a debate on the Asian Miracle, which ended in the conclusion that so-called Asian values – such as group spirit, mutual assistance, and thrift – were conducive for economic growth (see the literature referred in Khoo Boo Teik, 1999; Chong, 2002).

The fall of the Iron Curtain in 1998 was another event that set off a discussion on the relevance of values for economic growth (Blanchard et al., 1994; Murrell, 1995). The discussion concentrated on the best tactic to introduce a market economy in the former centrally planned economies. Neo-classical economists advised a shock therapy by which free market institutions would be introduced as soon as possible (the authors in the edited volume Blanchard et al., 1994). Others (Murrell, 1995), however, plead for a gradual approach. In their view, the years under communism had led to a culture in which citizens expect the state to take care of and decide on many aspects of life; thus it would take time for the inhabitants of these countries to change their attitude toward a more active stance and to take more responsibility for their own life.

In addition to economic development, developments in economic theory also led to an interest in culture. General equilibrium theory was able to demonstrate that in competitive markets the general equilibrium equations had some solutions. However, it proved to be impossible to show that there is just one (unique) solution toward which the economy converges. Other approaches made unrealistic assumptions in order to obtain any solution. Rational expectation models form an example of such an approach. In these models, economic agents are assumed to foresee developments into the far future and select the one and only path that leads to the stable equilibrium. These unsatisfactory results and assumptions gave rise to an interest among economists in alternative theories and concepts, such as bounded rationality, institutions, and culture in the sense of norms and values. The interest in culture was also stimulated by the fact that databases of cultural dimensions (Hofstede, 2001; Schwartz, 1994), and measures of people's values across the world (European Values Survey and World Values Survey) became available, so that empirical studies on the relationship between values and economic phenomena became possible.

Three Phases of Economic Development

Three phases of economic development provide insight into the relationship between culture and economic change

(Marini, 2004). Marini has defined these phases on the basis of matching findings of socio-psychological research with the literature on the role of culture for economic growth. Each phase is characterized by a particular level of economic growth, income per capita, and values. The preindustrial or antiquity is the first phase. In this phase, income per capita is low and economic growth is almost absent. Extended families and communities are buffers against economic hardship, which correspond with communitarianism and collectivism. The lack of sight on any progress leads to fatalism and an orientation on the past instead of the future. Industrialization is the next phase. During this period, economic growth is high and income per capita increases sharply. Economic growth is stimulated by technical progress and high levels of savings. Success in economic terms is rewarded to the individual, and the high levels of growth enable relatively large groups to escape from poverty. Associated with these events are an orientation on the future, achievement motivation, and an appreciation of hard work and materialism. The third and final stage is the post-modern or postindustrial period. It is characterized by high income per capita and relatively modest levels of economic growth. In principle, the entire population can take survival for granted. With it goes a change in dominant values. Self-expression and quality of life become more important than materialism and physical security. These values are in high esteem "*after* people have attained material security, and *because* they have attained material security" (Inglehart, 1997: p. 35).

Figure 1 summarizes the three phases inspired by the ideas of new institutional economics, which is an economic perspective that extends economics by including social and legal norms and values. In this mode, economic performance is determined by culture, formal institutions (written laws), and governance. The latter refers to the way members of a society normally treat each other without using the formal institutions of court and law: whether one seeks consensus or wants to fight out conflicts. The columns of institutions and governance are empty; we come to that at the end of this contribution.

Economic development is understood as the transition period from the preindustrial (often agrarian) stage to industrial society. The important question is then whether and how culture stimulates or hampers this transition. A related issue is whether it is culture or the formal institutions that are primarily responsible for this transition. These questions have been studied by means of historical and anthropological, in-depth analyses and cross-country regression analysis.

Historical and In-Depth Case Studies

Historical and in-depth studies range from covering a short time period or more than a millennium. Banfield (1958) is a good example of the former. His work describes in detail the society of a small poor village in southern Italy in the 1950s. A culture focused on the interests of the nuclear family is unable to organize activities for the collectivity, even though the latter would benefit all. Anxiety, suspicion, and hate make cooperation burdensome. Banfield regards the region's isolation, its high death rate, and the related fear of orphanage and neglect as important determining factors for this self-interest. It is a good description of what is meant by a backward society and how dominant values seem to consolidate the status quo. At the end of the book, Banfield suggests that some innovations such as formal education and apprenticeships may disrupt the status quo.

In contrast, the role of culture in supporting economic change is illustrated in the studies collected in Harrison and Berger (2006). For example, in a Nigerian town the traditional

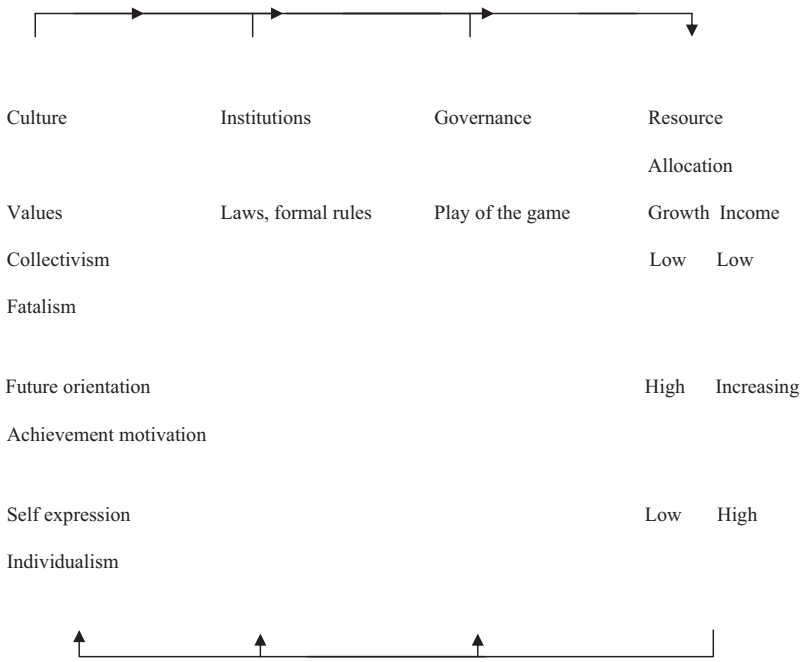

Figure 1 Values, institutions, growth, and income per capita. Source: Adapted from De Jong, Eelke, 2009. Culture and Economic: On Values, Economics and International Relations. Routledge, London (Figure 7.1).

habit was to have many children because a prosperous community was thought of as one with many people. These children were raised by the community. Formal education according to Western standards changed the attitude of young townspeople. They thought fostering as exploitative and as detrimental to the development of the children. Literacy and basic math competence taught in school enabled people to communicate with government's officials. The study on Japan (Kunio, 2006) in the Harrison and Berger's book illustrates the importance of culture for the change from a period of growth to that of an affluent society. According to Kunio, Japanese firms find it difficult to change from the community oriented progrowth strategy to one, which leaves more room for the individual (p. 98). Moreover, Japan is finding it difficult to reduce the government's involvement in the economy.

Historical studies covering a millennium or even a longer time period and a wide range of countries can disentangle the various factors that during several periods have contributed to development and stagnation. An important question dealt with in this type of studies is: why was it that European countries were so successful in the sixteenth and seventeenth century and why not China and countries in the Middle East? At the beginning of the Christian era, the world's scientific center was in the Middle East. Many of the techniques developed by the Europeans during the Industrial Revolution were known to the Chinese. For example, in the thirteenth century China had a power driven spinning machine, whereas in England such a machine was invented about 500 years later. Similarly, by the later eleventh century China knew how to use coal in blast furnaces for smelting iron. About 700 years later, this invention came in Britain, while in the mean time coal smelting had fallen into disuse in China (Landes, 2006: p. 6). Moreover, in the fifteenth century, China had implemented a policy of closure, which meant an abandonment of a program of great voyages. Landes (1998) ascribes the Chinese failure of making a commercial success of existing techniques and their decision to stop exploring the oceans to a lack of curiosity. The Chinese "went to show themselves, not to see and learn. … They were what they were and did not have to change" (Landes, 1998: p. 96). Success was there for those who were open for foreign influences and ideas, and were eager to learn. Moreover, individuals should have the freedom to experiment and be rewarded for their efforts; "freedom first, economy later; because freedom is a necessary if not sufficient condition of development" (Landes, 1998: p. 432).

The importance of freedom for economic development is underscored by Lal (1998), who ascribes a large role to individualism, which distinguished Europe from more ancient civilizations in Asia and the Middle East. This individualism stresses the nuclear family and romantic love instead of the traditional extended household. The move toward individualism was the unintended consequence of measures taken by popes in the sixth and tenth centuries. In the sixth century, Pope Gregory I forbade marriage to close kin, close affines or the widows of close kin, concubinage, and the transfer of children by adoption (Lal, 1998: p. 83). Al these four practices can be regarded as strategies of heirship such as ensuring the inheritance of family property and of the provision of a heir. By forbidding these practices the pope inhibited a family from retaining its property and stimulated its isolation. As a consequence, chances would increase that the church would obtain the property. From the tenth century onward, the enormous accumulation of property by the church raised opposition from within and outside. In 1075, Pope Gregory VII reacted by enforcing the celibacy of the clergy and the extension of the prohibitions to marry, measures which secured the church's ability to secure its property. Moreover, he declared the supremacy of the pope over the church and over secular matters, and the independence of the clergy from secular control. Excommunication was the pope's instrument for enforcing his will. The measures of pope Gregory I enhanced individualism and those of Gregory VII lead to a bureaucratic apparatus, such as a professional judiciary and a treasury needed for making the legal system work. These developments created the institutions of a market economy, which led to high levels of growth in later centuries.

However, institutions established in previous centuries can also hamper economic growth in later periods, as is illustrated by Kuran (2011). He argues that Islamic law and in particular its inheritance law has held back the Middle East. The Islamic inheritance system had very egalitarian rules, which gave women financial security and dampened wealth inequalities. However, combined with the possibility of polygamy it also precluded that a fortune could be inherited by a few successors. Consequently, large firms could not be established. The only means of accumulating wealth was the *waqf*, a form of trust. A *waqf* allowed the use of the return on assets for providing services in perpetuity. However, a *waqf* could not easily change its goals so that it meant a transfer of money from flexible commercial use to an inflexible charitable one. The degree of flexibility depended on the willingness of the local imam. Finally, Islamic contract law lacked any entity shielding, so that a partnership had to be dissolved on if any member or third parties asked for it. The death of a partner terminated the partnership automatically and gave claims to his heirs at the costs of the other partners. All these characteristics of Islamic law precluded the existence of independent, long-lasting large commercial firms. Consequently, trade and other economic activities remained small and were formalized in short-lived partnerships. Laws were changed when in the nineteenth century the disadvantages became too obvious; the Christian and Jewish traders became very rich compared to their Islamic counter partners. However, over the centuries, regulation had stimulated corruption, reduced mutual trust, and undermined civil organizations.

In principle, a historical analysis over a long period enables one to disentangle the causal effects between culture and economic growth. However, it is very difficult to obtain reliable information on dominant values for periods in the distant past, let alone their exact timing. This lack of information hampers studies on the relationship between values and economic performance, in particular when one is interested in the causal direction. As a result, in the end the findings of these historical studies depend to a high degree on the researcher's ability to disentangle the different aspects from the abundance of information provided. The above summary of Lal's study of the Western success and Kuran's investigation of the Middle East's failure illustrate the subjectivity of these studies. Lal tends to emphasize values more than Kuran does.

A way to circumvent this high level of subjectivity might be provided by quantitative studies.

Cross-Country Quantitative Studies

During recent decades, quantitative information has become available on dominant values in a wide range of countries. In 1980 Hofstede was the first to publish national cultural dimensions for a large set of countries. Later initiatives include the dimensions developed by the Israeli psychologist Shalom H. Schwartz, the World Values Survey, the European Values Survey, and the Global Leadership and Organizational Behavior Effectiveness research program (for a review see Appendix 1 in De Jong, 2009). These data sets allow for assessing the role of culture through cross-country regressions, in which growth in per capita income is the dependent variable and culture is among the set of explanatory variables. The latter includes income per capita at the beginning of the period. Its expected coefficient is negative; countries that have relatively low income per capita at the beginning of the period are expected to grow faster over the subsequent period and thus their income converges to that of the richer countries. The other explanatory variables can be derived from the literature on economic growth.

Several variables are suggested as proxies for culture. Inspired by, among others, the work of Putnam et al. (1993), some studies have used trust for explaining economic growth and income per capita. Trust is measured as the percentage of respondents in each country that replied 'most people can be trusted' when asked 'generally speaking, would you say most people can be trusted, or that you cannot be too careful in dealing with people?'. Trust appears to be positively associated with annual growth of income per capita (Knack and Keefer, 1997; Zak and Knack, 2001). The argument is that people who trust each other don't have to rely on costly institutions that ensure the execution of the counterparty's obligation. As a consequence, high trust leads to a more efficient society and growth. Beugelsdijk (2006) argues that this is an argument based on a microlevel concept of trust, whereas the trust question refers to the macrolevel, where it is related to the functioning of the bureaucratic system. People trust each other because they know that in case of misbehavior they can rely on courts to defend their rights. A principle components analysis finds trust in a component with measures of institutional strength. In conclusion, the trust variable seems to measure both values and the quality of institutions.

Strong desires to achieve on the part of at least some people are considered by McClelland (1961) as the cultural factor contributing to economic growth. During other phases of economic development, values as obedience and quality of life are regarded as more important (see **Figure 1**). Granato et al. (1996) and Marini (2004) construct a cultural index representing each of these stages by means of the answers given in the World Values Survey when respondents were asked to mention five qualities they teach their children out of a list of 11 items. Qualities are for example obedience, independence, thrift, and hard work. Both studies find a positive influence on economic growth for cultural indices of achievement motivation vs obedience. Williamson and Mathers (2011) use a similar procedure and explicitly confront the proxies of culture with those of formal institutions. They find that both types of proxies have a positive influence on economic growth. However, institutions enhancing economic freedom appear to be more important and more often significant for explaining economic growth than proxies of culture.

Another group of authors is inspired by Weber's hypothesis that religion is an important explanatory variable for economic growth (e.g., Grier, 1997; Noland, 2005; Barro and McCleary, 2003). Two types of variables are used. The first type measures whether respondents belong to a particular religion. Grier (1997) finds that in a group of 63 former colonies, protestantism is positively and significantly related to real (gross domestic product) GDP growth and GDP per capita is higher, the higher the percentage of protestants in countries. In a similar vein, Noland (2005) studies the influence of religious affiliation on per capita economic growth and on total factor productivity. He finds that Islam is not a drag on growth. In some studies the answers on aspects of belief are representing religion. Barro and McCleary (2003) use the answers on questions about respondents' beliefs in afterlife, heaven, and hell. They find that in particular belief in hell is positively correlated with economic growth, even if they included monthly church attendance as a representation of the costs associated with acquiring theses beliefs.

Individualism vs Collectivism is an important if not the most important cultural dimension. Çigdem Kagitçibasi (1994: p. 52) even labels the 1980s "the decade of I/C". Individualism is often found to be advantageous for innovation. The production process of newly innovated products needs coordination, which Gorodnichenko and Roland (2011) assume to be more easily be arrived at in collectivistic societies, such as many Latin-American countries and countries in Asia (Indonesia, South Korea, and Taiwan). Their model of this trade-off between the innovational advantages of individualism and the production advantages of collectivism, also assumes that the social status reward for developing a better technology is positively associated with individualism. In their empirical part they find that output per capita is positively related to individualism. This relation is robust for including institutions and other measures of culture.

A cross-country regression analysis delivers the factors that explain the *relative* position of a country in a group. This implies that the results depend on the countries selected for the analysis. Beugelsdijk (2006) and Beugelsdijk et al. (2004) show that the positive effect between trust and economic growth found in Zak and Knack (2001) is driven by the inclusion of developing countries in the sample. An analysis restricted to developed countries only does not find any significant positive influence of trust. The coefficient's level of significance increases by the inclusion of countries with lower GDP per capita and lower levels of trust. Williamson and Mathers (2011) make subsamples and find that both culture and institutions are significant in relatively freer countries.

The Role of Institutions

Many authors argue that formal institutions are much more important for economic development than culture (see e.g., Williamson and Mathers, 2011). Institutions such as secure

property rights will give individuals the incentive to invest and thus lead to development. Acemoglu et al. (2001) argue that the current differences in GDP per capita in large parts of Asia, Africa, and America are related to the quality of current institutions that are determined by the institutions brought to the countries by European colonizers. The latter established good institutions in countries with a climate favorable for settlement. Within Europe the countries that grew fastest after the fifteenth century were the countries with access to the Atlantic and with institutions that curbed the influence of the monarchy (Acemoglu et al., 2005). As a consequence, whereas the access to the Atlantic lead to prosperity in Portugal, Spain, The Netherlands, and England, the latter two countries showed a more rapid growth as they were better able to curb the absolutist power of the monarchy.

This discussion has a tendency to set formal institutions against culture as understood as local values and beliefs. As illustrated in **Figure 1**, however, culture and formal institutions are related and, unless they are imposed by foreign forces, culture and formal institutions are complements and might reinforce each other. In the earlier mentioned study of a South Italian village, Banfield (1958: p. 34) notes that "centuries of oppression have left the peasant with a pathological distrust of the state and of all authority". From an outsider's view, villagers seem passive and lack the aspiration for improving their circumstances. This attitude is reinforced or caused by a law passed in fascist times, which prohibits a villager to go to a town and look for work unless he has obtained permission by the provincial authorities (Banfield, 1958: pp. 58, 59). Here one can argue that institutions breed a passive attitude. Another example of the effects of formal institutions on attitudes is provided by Kunio (2006). In his view, the lifetime employment provided by Japanese firms generates a strong sense of identification of Japanese employees with their company.

Another criticism on the literature about the role of institutions in promoting economic development concerns the fact that almost all authors argue that institutions enabling private initiative are progrowth. Some, among whom Chang (2008) is prominent, argue that it is not always promarket institutions that are progrowth. He takes the success of Asian countries as an example to illustrate the possibility for governments to arrange growth. Moreover, promoting promarket institutions in countries, which are not suited for these rules can even do more harm than good. Finally, he notes that the governments who are now advocating free markets (Britain and the United States of America) used protective measures during their own growth phase. This fact is also noticed by Landes (1998: p. 266).

A way of disentangling cultural factors from institutional ones is by studying the culture and performance of districts, which have common national institutions. Putnam's et al. (1993) study of regional patterns of civic engagement and governments' performance in Italy is a good example in this regard. He measures civic engagement by the number of sports clubs per inhabitant, the degree of readership of newspapers, and the turnout in referenda, where voting is not obligatory. A regional government's performance is measured by items as: the number of cabinets during a 10-year period, the speed by which the annual budget is improved, the ability to use nationally allocated funds for the intended purpose. Regional governments appeared to perform better in regions with a high level of civic engagement, which he also labeled as social capital. Moreover, the government is less efficient in regions where local rulers are relatively frequently visited by citizens and a high percentage of these visits are requests for a job or another favor. Putnam concludes that social capital, which includes mutual trust, is crucial for the well functioning of institutions.

Politicized Culture

Scholars have understood culture to be a general trend to which the majority of the population adheres, people's ingrained attitudes toward different aspects of life. In this view, cultural is a neutral process in which no particular party plays a specific role. More recently, however, some scholars emphasize that culture is often consciously created by power holders to justify their actions and privileged position. We have labeled this politicized culture; a purposefully created view on the development and origins of a group (mostly nations) created by political leaders (De Jong, 2009: p. 101). Recent examples are the Asian and Islamic values debate in Malaysia (Maseland, 2006) and the African renaissance under Mbeki in South Africa (Bernstein, 2006).

In the early 1980s when Malaysia was in a recession, Mahathir, the then prime minister of Malaysia, launched the Look East campaign (Maseland, 2006: Chapter 5). Japan and Korea figured in this campaign as models for economic growth. The Eastern work ethic, which was thought to be an important explanatory factor in these countries' success was stressed as an example for the Malays. Later on, during the boom of the late 1980s and 1990s, the campaign stressed that such values already existed in Malaysia, are present in Islam, and hence part of the Malaysian heritage. At that time Mahathir organized an Islamization of public life, in which contents of Islam were promoted, which are in accordance with economic reforms. The pursuit of knowledge, thrift, and hard work were argued to be basic Islamic values. This view was summarized in Vision 2020, the ideological agenda of Mahathir formulated in 1991. Whereas in the 1970s in his The Malaya Dilemma, Mahathir had characterized the Malaya culture as backward, in the 1990s he pointed at the economic success and capabilities of the Malaya, summarized in the slogan "Malaysia Boleh!" (Malaysia can).

The African Renaissance is the invented element of South African politicized culture (Bernstein, 2006). It is promoted by President Thabo Mbeki and a small group of followers. Elements are the attack on African inferiority, and the promotion of African dignity. It shows pride in Africa's heritage but also in its capacity to modernize and take responsibility. African Renaissance "enables its proponents to advocate economic reform and democratic governance … as aspects of an African renaissance and prerequisites for the continent to claim its rightful place in the world. Essentially, it saves face while promoting intrinsically Western ideas about change for the continent" (Bernstein, 2006: p. 27). While the African renaissance can serve as a tool and set of beliefs for the leaders,

it seems to have less impact in the South African people in general (Bernstein, 2006: p. 28).

Causality

The causality between culture and economic development has been debated from the very beginning. Weber's protestant ethics is not a uniform plea in favor of the determining role of culture for economic development. On several places and in particular in the numerous footnotes, Weber shows that he is aware of the reverse causality. For example, the dominant position of protestants in the industry, can in his view, partly be ascribed to historical circumstances 'in which religion affiliation is not a cause of the economic conditions, but to a certain extent appears to be the result of them' (Weber, 1930: 2001, p. 4).

Historical analyses and case studies allow for studying the causality by clearly figuring out the timeline of the events. However, the lack of reliable information on attitudes and dominant values in the past hampers such an analysis. The cross-country regressions make use of values, which unfortunately are often measured only once or twice. As a result, one cannot disentangle the time pattern of the values, so that it becomes almost impossible to investigate the causal relation between culture (values) and economic growth. Culture could be a determining factor of growth, but the causality could also run in the reverse direction; reverse causality. Econometric studies tackle reverse causality by means of instrumental variables (IV) estimators. This estimation process consists of two steps. The first step explains the endogenous values by means of exogenous variables, which are not related to the present level of economic growth. In the second step, the fitted values of the first step regression replaces the original data. Tabellini (2010) is a good example of this approach. In his first step regression he assumes that the literacy rate at the end of the nineteenth century and political institutions in previous centuries explain the level of trust (his proxy for culture) in the 1990s. The fitted values of this regression of trust are found to influence the regional differences in per capita income in the 1990s. In this way the exogenous part of trust (culture) is only incorporated in the regression and the econometric problems associated with the endogeneity problem are solved. However, no information is obtained about the time pattern of culture and economic growth and thus about their mutual influence.

The only way to solve this problem is to measure these phenomena over time. A few studies have done so. Allen et al. (2007) use a quasi-longitudinal approach; in 2002, they used the same questionnaire and a sample similar (in the sense of respondents' age and sex) to the sample used by a study from the beginning of the 1980s for eight Asian countries. They find that GDP per capita is associated with certain value patterns; more submission, embeddedness and hierarchical in relative poor countries and more support for autonomy and egalitarianism in richer countries. The relation between 1982 values and subsequent economic growth was weak, whereas per capita income in 1982 correlated strongly with subsequent change in values. This supports Hofstede's finding that an increase in income per capita leads to an increase in individualism (2001: p. 255).

Ways Forward

As discussed in this article, arguments suggest that a period of economic growth is associated with values such as achievement motivation and future orientation. Research on the direction of causality between values, institutions, and economic growth, however, yields inconclusive results. Historical studies tend to conclude in favor of values driving institutional change and thus economic growth. Recent empirical studies deliver evidence for the opposite causal direction. Consequently, the direction of causality between values, institutions, and growth remains on the agenda for future research.

Such an approach should focus on the time pattern between these three concepts. The current practice of applying IV estimates in cross-section regressions is unsuitable for such an investigation. Analyses based on longitudinal data sets such as the World Values Surveys, along with more detailed historical accounts appear to me as more promising. If possible as many countries and periods as possible are included in these studies in order to investigate under which conditions certain outcomes prevail.

Although certain values are dominant in periods associated with high levels of economic growth, this does not imply that countries or regions in which at this time these values are not important, will not see growth in the future. Evidence suggests that these values are correlated with the phase of economic development (see **Figure 1**). Other values can be correlated with the way a society is organized. Hofstede's cultural dimension: uncertainty avoidance, for example, is often found to be important for explaining cross-country differences in institutions. However, it is uncorrelated with economic growth (De Jong, 2009: Chapters 5 and 6). This suggests that some cultural dimensions are important for explaining differences in the way societies are organized, whereas others are associated with the phase of economic development. In accordance with this is the idea that economic growth does not require a specific set of institutions. This idea is represented by the empty cells under institutions and governance in **Figure 1**. More systematic research is warranted to investigate the claim that different values are associated with the organization vs the growth phase of countries.

By far the majority of studies on values and economic performance are written by Western scholars. A basic principle in cross-cultural research is to avoid a cultural bias. At the end, researchers are also formed by the culture they were born in. Contributions by scholars born in non-Western countries are very welcome to limit the cultural bias that is present in the current literature.

See also: Cultural Critique, Anthropological; Culture: Contemporary Views; Economic Growth; European Religion from the Atlantic to the Black Sea; Globalization and Religion; Institutions; Islam: Middle East; Modernity; Moral Development, Cultural Differences In; Motivational Development: Cross-Cultural Perspectives; Norms; Obedience: Social Psychological Perspectives; Organizations, Sociology of; Religion and Politics in the United States; Religion, Sociology of; Religion: Morality and Social Control; Religious Fundamentalism: Cultural Concerns; Scientific Explanation; Value Pluralism.

Bibliography

Acemoglu, Daron, Johnson, Simon, Robinson, James, 2001. The colonial origins of comparative development: an empirical investigation. The American Economic Review 91, 1369–1401.

Acemoglu, Daron, Johnson, Simon, Robinson, James, 2005. The rise of Europe: Atlantic trade, institutional change and growth. The American Economic Review 95, 546–570.

Allen, Michael W., Ng, Sik H., Ikeda, Ken'ichi, Jawan, Jayum A., Sufi, Anwarul H., Wilson, Marc, Yang, Kuo-Shu, 2007. Two decades of change in cultural values and economic development in eight East Asian and Pacific island nations. Journal of Cross-Cultural Psychology 38, 247–269.

Banfield, Edward C., 1958. The Moral Basis of a Backward Society. The Free Press, Glencoe, IL.

Barro, Robbert J., McCleary, Rachel M., 2003. Religion and economic growth across countries. American Sociological Review 68, 760–781.

Bernstein, Ann, 2006. Culture and development: questions from South Africa. In: Harrison, Lawrence E., Berger, Peter L. (Eds.), Developing Cultures: Case Studies. Routledge, New York, pp. 23–41.

Beugelsdijk, Sjoerd, 2006. A note on the theory and measurement of trust in explaining differences in economic growth. Cambridge Journal of Economics 30, 371–387.

Beugelsdijk, Sjoerd, de Groot, Henri, L.F., van Schaik, Anton, B.T.M., 2004. Trust and economic growth: a robustness analysis. Oxford Economic Papers 56, 118–134.

Blanchard, O.J., Froot, K.A., Sachs, J.D. (Eds.), 1994. The Transition in Eastern Europe, vol. 1 Country Studies and vol. 2 Restructuring. University of Chicago Press, Chicago and London.

Chang, Ha-Joon, 2008. Bad Samaritans: The Myth of Free Trade and the Secret History of Capitalism. Bloomsbury Press, New York.

Chong, Terence, 2002. Asian values and confucian ethics: Malaya Singapore's dilemma. Journal of Contemporary Asia 32 (3), 394–406.

De Jong, Eelke, 2009. Culture and Economics: On Values, Economics and International Relations. Routledge, London.

Granato, Jim, Inglehart, Ronald, Leblang, David, 1996. The effect of cultural values on economic development: theory, hypotheses, and some empirical tests. American Journal of Political Science 40, 607–631.

Grier, Robin, 1997. The effect of religion on economic development: a cross national study of 63 colonies. Kyklos 50, 47–62.

Gorodnichenko, Yuriy, Roland, Gerard, 2011. Which dimensions of culture matter for long-run growth? The American Economic Review 101, 492–498.

Harrison, Lawrence E., Berger, Peter L. (Eds.), 2006. Developing Cultures: Case Studies. Routledge, New York.

Hofstede, Geert, 2001. Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations across Nations, second ed. Sage Publications, Beverly Hills, CA.

Inglehart, Ronald, 1997. Modernization and Postmodernization: Cultural, Economic and Political Change in 43 Societies. Princeton University Press, Princeton, NJ.

Kagitçibasi, Çiğdem, 1994. A critical appraisal of individualism and collectivism: toward an new formulation. In: Kim, Uichol, Triandis, Harry C., Kagitçibasi, Çiğdem, Sang-Chin, Choi, Yoon, Gene (Eds.), Individualism and Collectivism: Theory, Method, and Applications. SAGE Publications, Thousand Oaks, pp. 52–65.

Knack, Stephen, Keefer, Phillip, 1997. Does social capital have an economic payoff?: a cross-country investigation. Quarterly Journal of Economics 112, 1251–1288.

Kunio, Yoshihara, 2006. Japanese culture and postwar economic growth. In: Harrison, L.E., Berger, P.L. (Eds.), Developing Cultures: Case Studies. Routledge, New York, pp. 83–100.

Kuran, Timor, 2011. The Long Divergence: How Islamic Law Held Back the Middle East. Princeton University Press, Princeton, NJ.

Landes, David, 1998. The Wealth and Poverty of Nations. Abacus, London.

Landes, David, 2006. Why Europe and the West? Why not China? Journal of Economic Perspectives 20, 3–22.

Lal, Deepak, 1998. Unintended Consequences: The Impact of Factor Endowments, Culture and Politics on Long-run Economic Performance. The MIT Press, Cambridge, MA.

Marini, Matteo, 2004. Cultural evolution and economic growth: a theoretical hypothesis with some empirical evidence. The Journal of Socio-Economics 33, 765–784.

Maseland, Robbert K.J., 2006. Embedding Economics: The Constitution of Development and Reform in Malaysia and the Philippines (Ph.D. thesis). Radboud University Nijmegen, Nijmegen.

McClelland, David, C., 1961. The Achieving Society. D. van Nostrand Company, Inc., Princeton, NJ.

Murrell, Peter, March 1995. The transition according to Cambridge, Mass. Journal of Economic Literature 32, 164–178.

Noland, Marcus, 2005. Religion and economic growth. World Development 33, 1215–1232.

Putnam, Robert D., Leonardi, Robert, Nonetti, Raffaella Y., 1993. Making Democracy Work: Civic Traditions in Modern Italy. Princeton University Press, Princeton, NJ.

Schwartz, S.H., 1994. Beyond individualism/collectivism: new cultural dimensions of values. In: Kim, U., Triandis, H.C., Kagitçibasi, C., Choi, S.C., Yoon, G. (Eds.), Individualism and Collectivism: Theory, Method, and Applications. Sage, Thousand Oaks, CA, pp. 85–119.

Tabellini, Guido, 2010. Culture and institutions: economic development in the regions of Europe. Journal of the European Economic Association 8, 677–716.

Teik, Khoo Boo, 1999. The value(s) of a miracle: Malaysian and Singaporean elite constructions of Asia. Asian Studies Review 23 (2), 181–192.

Weber, Max, 1930/1992. The Protestant Ethic and the Spirit of Capitalism. Routledge, London.

Williamson, Claudia R., Mathers, Rachel L., 2011. Economic freedom, culture, and growth. Public Choice 148, 313–335.

Zak, Paul J., Knack, Stephen, 2001. Trust and growth. The Economic Journal 111, 295–321.

Culture and Economy

Nina Bandelj and Paul J Morgan, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article begins by providing the definitions of economy and culture, and then reviews classical sociologists' contributions to the topic. Next, we distinguish between two approaches. The first one uses culture as a starting point of inquiry, and examines the production and distribution of cultural objects and corporate and market culture. The second approach focuses on explaining economic phenomena through an examination of the regulatory and constitutive effects of culture on them, including in spheres of production, consumption, and economic development and globalization. In the second of these two approaches, conceptual differences are made between considering culture as part of the context that shapes economic outcomes and examining culture as constitutive of economic processes. Understanding economics as a cultural system and performativity of economic theories is also discussed.

Economy, Culture, Defined

Economy is "that complex of activities which is concerned with the production, distribution, exchange, and consumption of scarce goods and services" (Smelser and Swedberg, 2005: p. 3). Culture can be defined as "symbolic vehicles of meaning, including beliefs, ritual practices, art forms, and ceremonies, as well as informal cultural practices such as language, gossip, stories, and rituals of daily life" (Swidler, 1986: p. 273). DiMaggio (1994: p. 28) calls for an analytic distinction between forms of culture that are "characteristically constitutive (categories, scripts, conceptions of agency, notions of technique) and forms that are predominantly regulatory (norms, values, routines)."

Given the breadth of instantiations of both culture and economy, an inquiry into their relationship implies a great variety of possible venues for research. In this article, we limit our survey to two broad approaches. The first one uses culture as a starting point of inquiry, and examines the production and distribution of cultural objects, cultural industries, corporate cultures, and broad characterizations of how economic principles penetrate contemporary zeitgeist. We review this literature briefly as readers can find more information in the separate articles on these topics. The second approach uses economy as a starting point of inquiry, and focuses on how culture influences economic production, organization, consumption, pricing/valuation, and macrolevel development and globalization. Following DiMaggio's (1994) analytic distinction, we point to conceptual differences in considering culture as part of the context that shapes economic outcomes from that of examining culture as mutually constitutive of the economy, including the role of economic theory and the discipline of economics in this construction process.

Classics on Culture and Economy Connection

Even if sociology classics paid most attention to the changing economic conditions of modern society brought about by the industrial revolution, they have all, to varying degrees, discussed the relationship of these economic developments to culture. Karl Marx emphasized the material relations of capitalist production, or the material base, as determining of "the superstructure," or forms of social consciousness. Weber and Durkheim argued for a greater autonomy of culture and its independent influence on economic outcomes. In *The Protestant Ethic and the Spirit of Capitalism*, Max Weber wrote about the influence of religious worldviews on economic pursuits. For instance, a believer in ascetic Protestantism is driven by a desire to be saved and a belief that meticulous and hard work in the secular sphere represents a means of salvation, and success in the secular sphere symbolizes to others one's place within the elect. In *Economy and Society*, Weber defined economic action as social action to emphasize that economic action is behavior invested with meaning that is oriented to some other actor, emphasizing both meaning making and relationality in economic exchange. Emile Durkheim argued that members of a society share a collective consciousness, or an agreed upon system of morals, values, and symbols. He pointed to the noncontractual basis of contracts as a prime example of how economic behavior relies upon this collective consciousness: the normative expectations of society inform how contracts are conceived and respected and thereby facilitate and legitimate contracts beyond their terms. Further pointing to the moral foundations of economic behavior, in his *Professional Ethics and Civil Morals*, Durkheim placed religious beliefs affirmed through ritual practices at the basis of the inalienability of property and as motivation for upholding contracts. Previewing sociological study of consumption, Veblen used a concept of conspicuous consumption to describe consumption practices of luxury goods, which was wasteful and served as public display of social status, thereby treating consumption more as a signaling process than a fulfillment of needs. Last but not least, Karl Polanyi provided a classic statement on the creation of laissez–faire markets emphasizing not only the importance of state institutions in that process but also the role of liberal creed, or the ideas of economic liberalism, which came to dominate public discussions about how economy should be organized.

Culture in Markets

One major strand of research interrogating a relationship between culture and economy takes culture as the starting

point of analysis and locates it in various aspects of the economy, from production of culture to cultural industries to corporate cultures. We review these in turn.

The production of culture perspective has focused on the following six facets that influence the creation and dissemination of cultural objects (Peterson and Anand, 2004). First, law and regulation determine who controls what can be produced, or whether equal access to producers is assured. These regulations constrain or expand the range of symbolic elements produced. Second, technological innovation allows for communication and offers new openings for expression. Third, industry structure focuses upon processes of institutionalization, specifically at the field level. For instance, some cultural fields are dominated by very few, large firms, which offer a limited range of cultural products, whereas others have many, small firms offering a large range of objects. Fourth, differing organizational structures are variously suited to explore, support, create, and change different systems of symbolic production. Fifth, producers of culture have career trajectories and how their careers are situated within a field matters for their activity. Finally, the production of cultural objects needs to consider the market, mostly in the ways in which consumer tastes are channeled into market demand. Once new cultural goods markets are created, subsequent producers often tailor their creation to resemble already existing popular goods, also using measurement tools such as billboard charts for music or blockbuster ranking of movies that themselves structure the field of cultural production (Peterson and Anand, 2004).

A second way of investigating cultural objects in markets is to focus on cultural industries. Substantial work on this topic stemming from the humanities is linked to the critical theory of Adorno and Horkheimer and the incorporation of cultural industries into the commercial nexus of capitalism to function as ideological instruments with deceptive nature. Sociologist Paul Hirsch's (1972) reformulations reduce the politicized reading of the industries under question and instead focus upon the organizations, central to the processes of the creation, production, and supply of cultural items. Hirsch places emphasis on both inter- and intraorganizational coordination and competition, such as recruiting of new creators via 'contact men.' Ashley Mears' (2011) research on the models in the fashion industry, for instance, highlights the central role of brokers, or gatekeepers, who discover 'the right look,' develop it, and package it as a prized commodity. Patrik Aspers' (2010) study of the global high-fashion garment industry zeroes in on how order is maintained across the various interconnected retailer, producer, and credit market organization. He also suggests an analytical distinction between status markets and standard markets. In status markets, order depends on the identities of participating actors, such as global branded garment retailers, while in standard markets the quality of goods takes precedence.

Another take on culture in markets is to examine corporate cultures or organizational cultures more generally. Edgar Schein (1990: p. 111) defines organizational culture as "(1) a pattern of basic assumptions, (2) invented, discovered, or developed by a given group, (3) as it learns to cope with its problems of external adaptation and internal integration, (4) that has worked well enough to be considered valid and, therefore (5) is to be taught to new members as the (6) correct way to perceive, think, and feel in relation to those problems." Calvin Morrill (1995) shows how even the styles managers use to handle conflict in corporations is a part of corporate culture, which, Morrill finds, is heavily influenced by the hierarchical (or not) nature of the company's organizational structure. Given that scholarly research on corporate culture was accompanied by the focus of business practitioners on building strong cultures, Gideon Kunda's (2006) ethnography of an engineering division of a large American high-tech corporation provides a sobering call. Kunda examined managerial efforts to create corporate culture that promotes informal and flexible work environment to stimulate creativity and personal growth only to find that these pervasive efforts at culture building function as a form of control to which engineers responded with irony and distancing. Reflecting the observation of the increasingly important role of financial markets in contemporary world, several excellent studies also uncover the culture of Wall Street. Mitchell Abolafia's (1996) ethnography of Wall Street traders revealed the socially constructed nature of opportunism, hyperrationality, and heightened materialism, and the structural contexts that sustain them. A decade later, Karen Ho (2009) participant observed with investment bankers to expose how Wall Street culture fashions them to be "readily liquidated in the pursuit of stock price appreciation" as they, personally, "no longer benefit at all (or even suffer) when the corporation makes a profit" (p. 3).

Market Culture

Interested in the culture/economy interplay, a body of research performs a twist on a study of corporate culture to show how its economic logic of rationality, market and, most recently, finance has penetrated society as a whole, to become a prevalent way of how people make sense of their lives. George Ritzer (1993) developed the influential idea of the McDonaldization of society, arguing that the organizational innovations championed by the McDonalds corporation have come to define how contemporary society is run outside of the corporate world, and spilling over, for instance, to health care, education, tourism, and criminal justice. While the emphasis on efficiency, calculability, predictability, and control through automation seems to go hand in hand with Weberian rationalization and bureaucratization of the world, Ritzer, as did Weber, also warns of irrationality of rationality, especially its dehumanizing nature. Similarly interested in pinpointing the shifts in contemporary business culture, Luc Boltanski and Eve Chiapello (2007) examined French management texts from the 1960s and the 1990s to identify a 'new spirit of capitalism,' as labeled in the book's title. This new spirit advocates self-management and versatility in the workplace, and sees managers as coaches who get the workers to do what is needed by convincing them that the firm's interests are also their own interests. The normative order upon which this kind of capitalism rests is that of connections and networks, and a successful actor is one with plentiful contacts, flexibility, and mobility.

An increasingly vigorous line of inquiry has also been preoccupied with the contemporary rise of market fundamentalism, or a belief in "the moral superiority of organizing all

dimensions of social life according to market principles" (Somers and Block, 2005: p. 261). Margaret Somers and Fred Block (2005) coin the term 'ideational embeddedness' to explain how market fundamentalist ideas have radically transformed our knowledge culture. The concept is used to capture that the economy is embedded within the 'ideas, public narratives, and explanatory systems' that serve to create the necessary conditions for certain ways of organizing the economy to be normalized. The authors compare two cases of welfare revolutions in England and in the United States, in different historical periods, which were both preceded by the 'triumph of market fundamentalism as a new ideational authority,' but otherwise exemplified difference on every other significant explanatory factor. In their account, culture in the form of ideas is not only a part of the context in which economic processes are inserted but also a causal influence in constituting market outcomes.

In the wake of financial crises of 2008, scholars also turned to examining financialization as a major driver of societal transformation. Gerald Davis (2009) argues that finance has come to take center stage and reshaped how corporations and individuals behave. In this finance-centered system, the overriding concern of corporations is not productive activity but shareholder value, and individuals come to increasingly rely on financial markets, not employment in corporations, for security and wealth creation. Overall, this and similar research implies that pursuit of financial profit and self-interest is the order of the day. However, a cultural analysis of how 'economic interest' becomes a collective representation is actually lacking. Here, Lyn Spillman's (2012) work makes important headway by studying how economic interest is defined by American trade associations. Unlike typical characterizations, Spillman finds that both market and nonmarket relations are central to the business world. Far from being simply cutthroat, business associations develop shared ideas about collective interests and solidarity, and, overall, make business meaningful.

Cultural Embeddedness of Economic Organization and Production

Much of the understanding of organizational processes of production is influenced by sociological neoinstitutionalism, an early cultural perspective on organizations, which hones in on how people's actions are motivated by cultural norms, habit, and routine. Sociological neoinstitutionalism sees actors embedded in broad-scale contexts of meaning, which shape what is considered rational. Indeed, Meyer and Rowan (1977) argued that formal organizational structures are by-products of responding to commonly held 'myths' and engaging in ceremonial compliance. The neoinstitutional approach has given rise to an inquiry into how organizational fields are structured and a focus on institutional logics of action within industries that provide a guide to how decisions are made within companies. For instance, in her study of print industry, Patricia Thorton (2004) focused on how an institutional logic within the industry shifted from one focused on editorial issues to one focused on market issues. While before talk of profits was shunned and one went into business for the love of authors and books, in the 1980s, the institutional logic began to change with market channels and market positions becoming more of a concern and the basis on which executives were hired and promoted.

Referring to 'conceptions of control,' Fligstein (2001) captures a common understanding among economic actors about what works to make money. Conceptions of control are not just cognitive frameworks but they also define the relationships between producers, consumers, employees, managers, owners, and governments. Fligstein traces the change from the production conception to the sales and marketing conception and then to finance control. The late 1980s have seen a rise of the shareholder conception of control, which focused managerial attention on short-term profits and on making balance sheets look good to financial analysts, which would in turn be more likely to recommend the company stock and therefore enhance the shareholder value.

While studies highlighting the different cultural organization of the economy mostly focus on established markets, in an exemplary cultural analysis of economy, Richard Biernacki (1995) asks how the basic element of productive activity in capitalism, labor, is transformed into a commodity. Biernacki focuses on the diverging cultural definitions of labor as a commodity in Britain and Germany. He traces these understandings to on-the-ground practices by workers and employers in the wool textile mills during the nineteenth century, which were very similar economically and technologically but starkly different in terms of in-the-shop practices. Biernacki points to how the cultural definitions became codified in the political economy writings of Adam Smith and Karl Marx, which solidified them as distinct principles that would come to shape various outcomes, including how workers employed strike and protest, the kind of disciplinary techniques used in factories, and, in fact, susceptibility to Marxist analysis of exploitation.

Providing a cultural perspective on industrial policy formation, Frank Dobbin's (1994) now classic study of the railroad industry in the United States, France, and Britain in the nineteenth century showcases the elective affinity between political culture and economic organization in these countries. Dobbin argues that the principle of political order found in each country was translated into a principle of industrial order. He lays out how the nineteenth-century United States was organized as competing communities under a weak federal superstructure. In France, a strong central state dominated and coordinated private life, and in Britain, the emphasis was put on autonomous individuals with representation in Parliament. These distinct political cultures each left a mark on the economic organization of railroads in these three countries. In the United States, the railroad industrial policy situated economic sovereignty at first in community governments, and later in a market arbitrated by the federal government. In France, the central state oversaw industrial policy, while in Britain the policy placed sovereignty in individual firms.

Culture and Consumption

Sociological investigations of consumer markets have regularly yielded clear insight into how culture and economy intertwine. A classic perspective on the topic is that of Pierre Bourdieu (1984) and his discussion of taste and cultural capital.

Bourdieu has shown that the tastes that people have for different goods and activities, such as food, arts, or houses, are not a result of innate individualistic choices but depend on people's cultural capital, or proficiency in discerning what are considered to be more prestigious cultural goods or high-status signals. A focus on cultural capital suggests that tastes are socially conditioned. Moreover, consumption choices made are not merely to assert who one wants to be but rather to distinguish one from the others, given that consumption reflects a symbolic hierarchy determined and maintained by the dominant classes in society, whose interest is in keeping the social distance or, distinction, from other classes. For Bourdieu, taste is therefore not only determinant of consumption practices but also a social weapon for preserving power.

Viviana Zelizer (2005) develops a more meaning-driven, yet still relational, insight into the processes of culture and consumption within the economy. For Zelizer, the act of consumption in markets is not simply the end process of production and distribution, nor is it merely a momentary site of acquisition. Rather, consumption has within it "the negotiation of meaning, the transformation of relations in the course of economic interaction, and the social process of valuation itself" (p. 351). These processes highlight that the ties between producer, seller, and consumer must be investigated for their content; that these ties are constantly (re)negotiated, and not set in a power hierarchy; and that proper understandings of consumption must also focus on use and not only on acquisition of commodities.

In their review of the culture and consumption topic, Zukin and Maguire (2004) emphasize that large-scale changes in cultural understandings inform consumption practices, especially in relation to processes of individualization and collective identity. This links to an important theme of commodification. From a critical theory perspective, the spread of markets facilitates commodification because it makes everything available for sale, and therefore voids objects of any meaningful content, reducing them to one common denominator, market price. Cultural economic sociologists have countered this position on commodification, which reinforces the assumption that culture and economy are two separate spheres, whereby the introduction of monetization and marketization contaminates meaningful social relations. In contrast, culture and economy are intertwined and function as connected worlds rather than separate spheres. People use money to differentiate social relationships of various kinds and use different types of payments (compensation, gifts, or entitlements) to affirm, rather than corrupt, the meaning of a particular relation (Zelizer, 2010).

Likewise, recent work on exchange of human blood and organs, or other 'contested commodities' (Radin, 2001) asserts that the crux of the problem is not merely to allow or prevent the exposure to the market, but that market relations already coexist with their circulation in various ways. In fact, Kieran Healy (2006) shows that blood and organs are not given everywhere in the same way or to the same extent. That is because, Healy claims, donations do not simply depend on individual gestures of altruism but are structured by the contexts for donations. Specifically, procurement organizations play an important role in sustaining donation by providing structural opportunities to give and by producing cultural accounts of what giving means. Similarly, on examining the medical market for eggs and sperm, Rene Almeling (2011) finds that commodification of bodily goods does not result in uniform meanings, or lack thereof. Even if both men and women are paid for their donations, Almeling finds that clinics encourage sperm donors to think of the money they receive as payment for a job, while women are urged to regard egg donation in feminine terms, as a precious gift that one woman can bestow on another.

Cultural Understandings of Price and Value

A vigorous line of cultural economic sociology focuses on the cultural preconditions of pricing and value. Viviana Zelizer (2010) is the preeminent contributor to this scholarship with her historical analyses of the changing valuation of life and children. In *Pricing the Priceless Child*, a study of the changing economic value of children, Zelizer accounts for the cultural transformation from viewing children as economic contributors to family welfare, to increasing valuation of children as priceless, and sacred, and removed from the economic sphere. In her analysis of putting a price on life as in development of life insurance, Zelizer shows the opposite trend. She begins by describing a strong moral antagonism against life insurance in the nineteenth-century United States, because mixing the sacred value of human life with the profane sphere of money is seen as blasphemous, and these cultural obstacles prevent the market for life insurance to grow. Only when life insurance is reframed as a loving act of caretaking, and therefore aligned with the moral convictions of customers regarding the sacredness of life, can the market proliferate. Cheris Shun-ching Chan (2012) conducts a similar analysis by investigating how the market for life insurance developed in postsocialist China. Chan begins with a puzzle about how companies could sell life insurance in a country where death is a taboo subject, and explores how foreign and domestic companies negotiated local cultural resistance to facilitate market growth.

The cultural perspective on price has also been applied to esthetic markets. Olav Velthuis (2005) analyzed the strategies gallery owners use to price contemporary art in New York and Amsterdam. Far from reducing an esthetic object to a mere number, Velthuis reveals the moral significance of prices and meaning making that goes into their construction. He portrays gallery owners not as profit maximizers but as members and supporters of the art world who follow the conventions of their professional community and price art according to distinct kinds of narratives. A high price may indicate not only the quality of a work but also the identity and status of collectors along with that of the gallery and gallerists. Dealers might raise prices quickly for an artist who they consider to be a superstar, or more gradually in other cases, framing it as prudence in decision making. Velthuis reveals how prices operate as a symbolic system akin to language.

Frederick Wherry (2008) also emphasizes that prices are not culturally neutral but he interestingly points to an intersection of meaning attribution with identity and social location of the purchaser. For Wherry, prices are used to compare the qualities of differently situated people, those near to, as opposed to those far from, the mainstream. Based on this distance

dimension, consumers are characterized as foolish, faithful, frivolous, or frugal. These categories are consequential for societal assessments about what differently characterized consumers deserve. Such considerations preview increasing attention in the literature to the broadly defined notions of worth and value, its social construction, evaluative practices, and regimes of worth (Lamont, 2012). Jens Beckert and Patrik Aspers (2011) bring together a group of economic sociologists who tackle the question of how the worth of goods is established, building on the work of scholars like Charles Smith (1989) who pioneered the study on the social construction of value using auctions as his research site. Beckert and Aspers make a point that "for products to be sold in markets, customers must value them and assess their value in relation to other products. Firms must produce products and position them in the market where customers will consider them valuable. Financial investors must assess the value of assets through judgments of the opportunities and risks involved" (p. 3). Valuing goods is therefore central to all aspects of economic life. Examining the cultural underpinnings of valuation and evaluation helps hone in on how moral values and cultural practices influence economic value.

Culture, Development, and Globalization

Last but not least, culture and economy interplay has been explored at the macrolevel in an investigation of economic cultures that vary across nations. From a social psychological perspective, Geert Hofstede (2001) developed a theory of cultural dimensions, which identifies several dimensions of culture that influence work behavior (and other outcomes). These dimensions include power distance (from low to high), individualism vs collectivism, tolerance or avoidance of uncertainty and ambiguity, masculinity vs femininity, and long- vs short-term orientation. Hofstede developed his model from data compiled in a worldwide survey of employee values in International Business Machines' subsidiaries across the world in the 1960s and 1970s, and his work has been widely influential in social psychology and business and management studies that examine cross-national differences on various economy-related outcomes.

From a constructivist perspective, sociologists have argued that different national contexts are distinguished by different cultural logics of action. For instance, Nicole Biggart and Mauro Guillen (1999) compared auto industry in four nations, South Korean, Taiwan, Spain, and Argentina to propose that economic development depends on how a country's institutionalized patterns of authority and organization – which legitimize certain actors and certain relationships among those actors – are linked to the opportunities in a global market. Different actors, such as states, business networks, small businesses, or families, will be considered as more or less legitimate players in different countries, and this and other collective understanding and cultural practices will contribute to different organizing logics of economy. The power of institutional logic of actions is such that "economic and managerial practices and actions not consistent with the institutional logic of society, even if they are abstractly 'better' or 'more efficient,' are not readily recognized and incorporated" (p. 726). Therefore, the strategies for capital accumulation or any other economic policy that succeed in one country may not be a good fit with the preexisting organization of another country. As such, the course of development and the policies that enable it should build on the institutionalized logics rather than uniform global strategies.

With a different take on the role of culture for development, Nina Bandelj and Frederick Wherry (2011) propose an inquiry into the 'cultural wealth of nations' that exposes how firms located in particular places find themselves at an advantage relative to firms in other places by virtue of the symbolic resources they have at their disposal. These symbolic resources include collective narratives, reputations, status, and ideas. A variety of state and private actors are involved in creating narratives, staging social performances, and managing a country's reputations in a global market so as to, for instance, attract world tourists or inflows of foreign investment and therefore further national economic development. Authors also warn that conversion of cultural capital to economic capital at the level of countries is not a seamless process, and one size does not fit all. Nevertheless, creating and deploying cultural wealth, not only industrial progress, represents a viable strategy of economic prosperity for developing countries in a globalized world.

When considering globalization and the interplay of culture and economy, Ritzer (2003: pp. 193–195) makes a distinction between glocalization and grobalization, where grobalization denotes the 'imperialistic ambitions' of corporations and other entities to see their power and profits *grow* at a global scale, resulting in 'proliferation of nothing,' or 'social forms devoid of substantive content.' This contrasts with glocalization, or 'the interpenetration of global and local, resulting in unique outcomes in different geographic areas,' and proliferation of 'something.' To understand global culture and economy, we also need to take into account a vast literature on world society (Meyer et al., 1997), which claims that, over a century, the world has constituted a single world polity with a corresponding world culture, carried by international nongovernmental organizations promoting universalism, individualism, rational progress, and world citizenship (Boli and Thomas, 1997). While the majority of research in this field examines environmental, educational, and political outcomes, a handful of studies showcase the promise of this type of inquiry for understanding economic phenomena. For example, Henisz et al. (2005) examine how world society theory can help us understand why countries differ in the extent to which they adopt market-oriented reforms and find that international pressures of coercion, normative emulation, and competitive mimicry strongly influence domestic economic policy. More recently, Lim and Tsutsui (2012) found that ties to world culture strongly predicted whether corporations adopt corporate social responsibility initiatives.

Economics and Economic Theory as Cultural Forms

In the final section, we discuss research that intimately couples the economy and culture by understanding the discipline of economics and economic theory as cultural forms. In a major

statement, Marion Fourcade (2009) studies economics as a discipline and profession in the United States, Britain, and France to understand how national cultures inform the understanding of economics as an object of analysis as well as economists as a category of knowledge. Fourcade brings attention to the differing ways in which economists conceptualize problems and practices in economic functioning across the three national contexts. She situates her arguments within three levels of understanding: the political bases of economic order, the cultural bases of institutional orders, and knowledge production dependence on local context. The study is also insightful about the difference in the kind of economic policies favored by economists in different nations. In Britain, the discipline of economics developed out of a political culture that valued small, cohesive societies that exerted much influence over the genesis of national policy. In France, economics was shaped by the administrative exercise of public power. In the United States, it was the dominance of market institutions that contributed to the success of economists in influencing society and policy.

Economists are also increasingly seen as cultural agents who theorize the markets they then subsequently come to create. Donald MacKenzie and Yuval Milo (2003) showed how the now-famous Black-Scholes model for pricing options became incorporated into actual prices of options on the Chicago Board of Options Exchange floor. This work is exemplary of the research in 'performativity' of economics, which, following Michel Callon's (1998) influential statement, seeks to show how economic theory *creates* the reality in its image, as opposed to simply describing – more or less accurately – an external reality that is not affected by the theory. Often the focus of this research, firmly grounded in the social studies of finance, is the role of sociotechnical tools in performing financial markets.

By considering economic objects (including markets, economists and economic theory) as cultural objects, an analysis of the interplay between culture and economy is a far cry from common assumptions that the two are separate spheres with little in common. As we hoped to make evident in this brief review, a growing and vigorous strand of scholarship continually counters the long-held views that some economic phenomena are more cultural than others, and that some more 'standard' economic processes, such as those located in corporations, markets, and finance, leave little or no room for culture. Indeed, relatively recently, cultural economic sociology that envisions culture and economy as mutually constitutive domains has proliferated in a field that before long was dominated by structural analyses. Vigorous interest in the constitutive role of culture in economy points to its promise in answering some vexing questions presented to us during turbulent economic times.

See also: Art and Culture, Economics of; Artists, Competition And markets; Cultural Capital and Education; Culture and Institutional Logics; Culture, Production of: Prospects for the Twenty-First Century; Culture, Sociology of; Economic Sociology; Globalization and World Culture; Markets: Artistic and Cultural; Networks and Cultural Consumption; Organizations and Culture; Sociology and the New Institutionalism.

Bibliography

Abolafia, Mitchel, 1996. Making Markets: Opportunism and Restraint on Wall Street. Harvard University Press, Cambridge, MA.

Almeling, Rene, 2011. Sex Cells: The Medical Market for Eggs and Sperm. University of California Press, Berkeley, CA.

Aspers, Patrik, 2010. Orderly Fashion: A Sociology of Markets. Princeton University Press, Princeton, NJ.

Bandelj, Nina, Wherry, Frederick (Eds.), 2011. The Cultural Wealth of Nations. Stanford University Press, Stanford, CA.

Beckert, Jens, Aspers, Patrik (Eds.), 2011. The Worth of Goods: Valuation and Pricing in the Economy. Oxford University Press, Oxford, UK.

Biernacki, Richard, 1995. The Fabrication of Labor: Germany and Britain, 1640–1914. University of California Press, Berkeley, CA.

Biggart, Nicole W., Guillen, Mauro F., 1999. Developing difference: social organization and the rise of the auto industries of South Korea, Taiwan, Spain, and Argentina. American Sociological Review 64 (5), 722–747.

Boli, John, Thomas, George M., 1997. World culture in the world polity: a century of international non-governmental organizations. American Sociological Review 62 (2), 171–190.

Boltanski, Luc, Chiapello, Eve, 2007. The New Spirit of Capitalism (G. Elliott, Trans.). Verso, London, UK.

Bourdieu, Pierre, 1984. Distinction: A Social Critique of the Judgment of Taste (R. Nice, Trans.). Harvard University Press, Cambridge, MA.

Callon, Michel, 1998. The embeddedness of economic markets in economics. In: Callon, Michel (Ed.), The Laws of Markets. Oxford University Press, Oxford, UK.

Chan, Cheris Shun-Ching, 2012. Marketing Death: Culture and the Making of a Life Insurance Market in China. Oxford University Press, Oxford, UK.

Davis, Gerald F., 2009. Managed by the Market: How Finance Re-Shaped the Market. Oxford University Press, Oxford, UK.

DiMaggio, Paul, 1994. Culture and economy. In: Smelser, N., Swedberg, R. (Eds.), The Handbook of Economic Sociology. Princeton University Press, Princeton, NJ, pp. 27–57.

Dobbin, Frank, 1994. Forging Industrial Policy: The United States, Britain, and France in the Railway Age. Cambridge University Press, New York, NY.

Fligstein, Neil, 2001. The Architecture of Markets: An Economic Sociology of Twenty-first-Century Capitalist Societies. Princeton University Press, Princeton, NJ.

Fourcade, Marion, 2009. Economists and Societies: Discipline and Profession in the United States, Britain and France, 1890s–1990s. Princeton University Press, Princeton, NJ.

Healy, Kieran, 2006. Last Best Gifts: Altruism and the Market for Human Blood and Organs. University of Chicago Press, Chicago, IL.

Henisz, Witold J., Zelner, Bennet A., Guillén, Mauro F., 2005. The worldwide diffusion of market-oriented infrastructure reform. American Sociological Review 70 (6), 871–897.

Hirsch, Paul M., 1972. Processing fads and fashion: an organization-set analysis of cultural industry systems. American Journal of Sociology 77 (4), 639–659.

Ho, Karen, 2009. Liquidated: An Ethnography of Wall Street. Duke University Press, Durham, NC.

Hofstede, Geert, 2001. Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations across Nations. Sage Publications, Thousand Oaks, CA.

Kunda, Gideon, 2006. Engineering Culture: Control and Commitment in a High-Tech Corporation. Temple University Press, Philadelphia, PA.

Lamont, Michele, 2012. Toward a comparative sociology of valuation and evaluation. Annual Review of Sociology 38, 201–221.

Lim, Alwyn, Tsutsui, Kiyoteru, 2012. Globalization and commitment in corporate social responsibility: cross-national analyses of institutional and political-economy effects. American Sociological Review 77 (1), 69–98.

MacKenzie, Donald, Millo, Yuval, 2003. Constructing a market, performing theory: the historical sociology of a financial derivatives exchange. American Journal of Sociology 109 (1), 107–145.

Mears, Ashley, 2011. Pricing Beauty: The Making of a Fashion Model. University of California Press, Berkeley, CA.

Meyer, John W., Boli, John, Thomas, George M., Ramirez, Francisco O., 1997. World society and the Nation-state. American Journal of Sociology 103 (1), 144–181.

Meyer, John W., Rowan, Brian, 1977. Institutionalized organizations: formal structure as myth and ceremony. American Journal of Sociology 83, 340–363.

Morrill, Calvin, 1995. The Executive Way: Conflict Management in Corporations. Chicago University Press, Chicago, IL.

Peterson, Richard A., Anand, Narasimhan, 2004. The production of culture perspective. Annual Review of Sociology 30, 311–334.

Radin, Margaret J., 2001. Contested Commodities: The Trouble with Trade in Sex, Children, Body Parts, and Other Things. Harvard University Press, Cambridge, MA.

Ritzer, George, 1993. The McDonaldization of Society. Sage Publications, Inc., Thousand Oaks, CA.
Ritzer, George, 2003. Rethinking globalization: glocalization/grobalization and something/nothing. Sociological Theory 21 (3), 193–209.
Schein, Edgar H., 1990. Organizational culture. American Psychologist 45 (2), 109–119.
Smelser, Neil, Swedberg, Richard (Eds.), 2005. The Handbook of Economic Sociology, second ed. Princeton University Press, NJ.
Smith, Charles W., 1989. Auctions: The Social Construction of Value. Free Press, New York, NY.
Somers, Margaret, Block, Fred, 2005. From poverty to perversity: ideas, markets, and institutions over 200 Years of welfare debate. American Sociological Review 70, 260–287.
Spillman, Lyn, 2012. Solidarity in Strategy: Making Business Meaningful in American Trade Associations. University of Chicago Press, Chicago, IL.
Swidler, Ann, 1986. Culture in action: symbols and strategies. American Sociological Review 51 (2), 273–286.
Thornton, Patricia H., 2004. Markets from Culture: Institutional Logics and Organizational Decisions in Higher Education Publishing. Stanford University Press, Stanford, CA.
Velthuis, Olav, 2005. Talking Prices: Symbolic Meanings of Prices on the Market for Contemporary Art. Princeton University Press, Princeton, NJ.
Wherry, Frederick, 2008. The social characterizations of price: the fool, the faithful, the frivolous, and the frugal. Sociological Theory 26 (4), 363–379.
Zelizer, Viviana, 2005. Culture and consumption. In: Smelser, N.J., Swedberg, R. (Eds.), The Handbook of Economic Sociology, second ed. Princeton University Press, Princeton, NJ, pp. 345–368.
Zelizer, Viviana, 2010. Economic Lives: How Culture Shapes Economy. Princeton University Press, Princeton, NJ.
Zukin, Sharon, Maguire, Jennifer Smith, 2004. Culture and consumption. Annual Review of Sociology 30, 173–197.

Relevant Websites

http://www.palgrave-journals.com/ajcs/index.html – American Journal of Cultural Sociology.
http://www.tandfonline.com/toc/gcmc20/current – Consumption Markets and Culture.
http://www.tandfonline.com/toc/reso20/current – Economy and Society.
http://www.tandfonline.com/toc/rjce20/current – Journal of Cultural Economy.
http://codeandculture.wordpress.com/2009/08/26/production-of-culture/ – Production of Culture Online Resources.
http://socfinance.wordpress.com/tag/social-studies-of-finance/ – Socializing Finance Blog.
http://ser.oxfordjournals.org/ – Socio-Economic Review.

Culture and Emotion

Batja Mesquita, Nathalie Vissers, and Jozefien De Leersnyder, Center for Social and Cultural Psychology, University of Leuven, Leuven, Belgium

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S. Kitayama, volume 5, pp. 3134–3139, © 2001, Elsevier Ltd.

Abstract

Research on culture and emotion has moved beyond the once central nature–nurture. Evidence suggests that there are universal *constituents* of emotions – the 'emotional potential' – rather than universal *emotions*. Furthermore, the emotional constituents are assembled in culture-specific ways that are meaningful and predictable, resulting in systematic cultural differences in 'emotional practices' (i.e., people's actual emotional lives). Whereas emotions are not universal, the underlying rule of emotional life is: Culturally adaptive emotions are more prevalent and intense. After reviewing empirical evidence for each of these claims, we outline some of the future directions for research on culture and emotion.

Introduction

In the history of emotion research, the pendulum has swung repeatedly between universalist and social constructionist positions on emotions (Kitayama, last edition of the Encyclopedia). Universalist approaches emphasized the evolutionary origin of emotions, and considered them as biologically given and invariant. In contrast, social constructionist approaches emphasized nurture, focusing on the cultural processes shaping emotions, and assuming that emotions were cross-culturally different. Variability in the phenomena was thus taken to be informative about the constitutive processes: The degree of variability of emotion was conflated with their origin.

More recently, the recognition that culture itself is perhaps the greatest evolutionary advantage of all for the human species has shed a different light on the nature–nurture debate. Plasticity – or cross-cultural variability – of emotions in ways that fit, and are adaptive to, the cultural context may itself be an evolutionary advantage. Most scientists would agree that human beings have some universal potential for emotions, but also that, across cultures, emotions are configured in very different, and uniquely adaptive, ways. The questions driving research on culture and emotion have become more nuanced, and focus on the ways the building blocks of emotional life are configured in culture-specific ways.

In the next section we first define the main concepts: culture and emotion. A summary of research on the universality of emotions will follow. The remainder of this review will focus on research that systematically approaches cultural differences in emotions.

Culture and Emotion: Definitions

Culture

Culture is the human-made part of our environment. It forms our reality, and provides the goals, norms, and values that guide our behavior. Culture is not just a varnished layer of meaning onto a core of reality, but rather it shapes reality itself; it includes the ways in which our daily lives are structured, in particular the habitual and normative ways of being, and having interactions and relationships. Culture importantly defines the context of adaptation for each and every individual.

The meanings and practices of self and relationships, or 'cultural models' vary widely between cultures, as illustrated by the contrast between European American and East Asian cultural models. According to middle-class European American models of self and relating, an individual should be both independent and autonomous. To achieve these goals, self-esteem and individual control are important conditions to the American way of independence, as they afford individuals to influence (others in) the world to achieve what is best for themselves. In contrast, the dominant goals of the self in most East Asian cultural contexts are interdependence, relatedness, and harmony, which can be achieved by adjusting to others. In Japan, adjustment is thought to be facilitated by self-criticism, because this leads the way to self-improvements.

Cultural models do not only exist in the head of individuals, but also 'in the world.' The American practice of complimenting and rewarding each other for every level of achievement is an example, and so are the politeness rules and the sessions of critical self-reflection in Japan. Therefore, cultural models invariantly define an individual's context of adaptation, even if not every individual engages in these models in exactly the same way.

Emotion

Emotions are first, and foremost, intentions to act (e.g., Frijda, 2007). Emotions are not just subjective feelings, but having an emotion also means a commitment to act. To take anger as an example: The experience of anger implies an attitude of nonacceptance, and the assessment that others will, or at the very least should, accommodate your wishes, goals, and values. Emotions thus serve important functions in relationships; according to some, the adaptive advantage of emotions is precisely that they help to coordinate and regulate relationships (e.g., Keltner and Haidt, 1999). Moreover, emotions communicate these intentions to act to others: Anger communicates that the person is not likely to accept the current state-of-affairs.

If emotions are commitments to act, it can be seen why they have cultural meaning. Some acts are consistent with the cultural models of self and relationships, whereas others are inconsistent. The philosopher Robert Solomon tried to explain why anger might be so much more rare among the Utku, a group of Inuit living in the Arctic Canada, than in America (2002: 139): Anger "is as basic to American culture as an emotion can be" because it fits with the notion of entitlement and control. On the other hand, anger is quite absent in the lives of the Utku Inuits, because it is incompatible with their 'rational' worldview and includes unjustifiable judgments and structures. As a result, anger plays no socially acceptable role in their society (Solomon, 2002).

In Pursuit of Universal Building Blocks: The Emotion Potential

Much of the psychological research on culture and emotion focuses on the *potential* of emotions: that is, the human capacity to have certain emotions. This research is less concerned with the actual emotions people in different countries have, the so-called emotion *practice*, than with people's ability for these emotions. In the next section, we will discuss two schools of thought on emotional potential that both have inspired the research on culture and emotion.

Universal Emotions

One school of thought that has inspired much of the psychological research on culture and emotions assumes that certain emotions, so-called 'basic emotions,' are universal building blocks of emotional life (e.g., Ekman, 1992). The underlying idea is that basic emotions have evolved for their adaptive value in dealing with fundamental life tasks: Emotions prepare for actions that were particularly useful for recurrent life tasks in our evolutionary past. For instance, anger protects us from exploitation. Most basic emotion theories also assume that each basic emotion (1) corresponds to one particular adaptive value, and (2) is characterized by its own set of unique features (antecedent events, facial and vocal expressions, physiology).

Cross-cultural research designed to prove the universality of basic emotions relied on the observable features characteristic of the basic emotions: facial and vocal expressions, and physiological response patterns.

Universal Emotions? Evidence from Studies on Expressive Behavior

Paul Ekman's research on facial recognition of emotions is a showcase example of the paradigm most commonly used (e.g., Ekman and Friesen, 1969). Ekman and colleagues showed pictures of faces to participants from a large range of different cultures (among them illiterate, remote cultures), and asked them to identify the emotion in the face by choosing one emotion out of a list. Even people from very different, non-Western cultures were able to correctly identify a number of emotions above chance. This was seen as proof for the universality of basic emotions. Similar studies have been reported for vocal expression, and yield comparable results (e.g., Sauter et al., 2010).

To be fair, basic emotion theorists did not deny differences in the *practice* of emotional expression. In fact, basic emotion theorists postulated the notion of 'display rules' to account for these differences. Display rules are the "cultural norms that dictate the management and modification of emotional displays depending on social circumstances" (Ekman and Friesen, 1969). Display rules have been found to differ across cultures. For instance, Matsumoto et al. (2008) asked participants from 32 different countries what they *should do* if they felt each of seven emotions (anger, contempt, disgust, fear, happiness, sadness, and surprise) on a scale ranging from emotion expression to nonexpression or masking (i.e., hide emotion with smile). The strongest differences in display rules were found for high-activation positive emotions, with the level of individualism of a country predicting the extent to which it was considered good to express positive emotions. (In Section Emotional Frequency and Intensity we will return to the finding that individualist cultures seem to value high-activation positive emotions more than collectivist cultures.) There is little research linking display rules to actual expression. Display rules are expected to influence expressive practices, but not the recognition of emotion.

Facial and vocal recognition studies, though designed to find universality, also yielded important and persistent cultural differences in emotions. Granted, people across cultures identified certain facial and vocal expressions *above chance*, but percentages of correct identification largely varied (ranging from about 20 to 95%), depending on both the culture and the emotion considered. Moreover, a systematic in-group advantage occurred, meaning that facial expressions were always best recognized by people from the same culture (for a meta-analysis see Elfenbein and Ambady, 2002). Recognition of facial expressions was also improved for perceivers who had been exposed to the culture of the sender, or whose culture was similar to the sender's culture.

The notion of a biological hard-wired 'language of emotions,' in the form of universal facial and vocal expressions, has itself been challenged. In a series of ingenious experiments, Rachel Jack and her colleagues compared the mental representations of facial expressions of six basic emotions (the same six that were included in the studies by Paul Ekman and colleagues) with a technique called 'reverse correlation.' This technique starts from gray-scale pictures of race-, gender-, and emotion-neutral faces that are covered with a random pattern of black and white dots (white noise). In each trial, one pixel is varied, after which participants choose one emotion out of the six basic emotions identified by Ekman and colleagues; they also have the option to answer 'I don't know.' By aggregating the facial stimuli of all trials that are classified as a certain emotion, the researchers established the mental representation of facial expressions associated with the emotion. In research comparing Western European and East Asian samples, Jack found the representations of facial expressions to be different: the eyebrow and mouth regions were important in Western representations, whereas the eye region, particularly the gaze direction, was important in East Asian representations of facial expressions (Jack et al., 2011). Therefore, the idea of universal facial emotion signals has been challenged by neuropsychological research.

Using a number of behavioral tasks, Maria Gendron and her colleagues also challenged the idea of universal vocal and facial signals of emotions (Gendron et al., 2014 a,b). In their experiments, they tested if facial and vocal expressions would be recognized without providing 'basic emotion' words by people from an isolated culture. Himba participants, an ethnic group living in the remote parts of northwestern Namibia, freely labeled, sorted, or matched emotional expressions. Throughout different tasks, the Himba differentiated between positive and negative emotional expressions, and to some extent between high and low activation expressions, but they did not clearly distinguish between discrete emotions. The studies did not provide support for the idea of a universal language for basic emotions: If anything, people universally distinguish between vocal and facial expressions of positive and negative affect.

Universal Emotions? Evidence from Studies on Physiological Responses

Other research aimed at showing cross-cultural similarity in the patterns of physiological responses associated with several basic emotions (e.g., Levenson et al., 1992). Reliable measures of physiological responding in emotion are hard to obtain, because strong emotions are difficult to induce in the lab, and outside of the laboratory many factors confound measures of emotional responding. Therefore, the scientists relied on a directed facial action task, a task that was built on the assumption that a particular set of different emotion features (facial expression, physiological responding) is linked invariantly to one emotion. In this task, participants' physiological responses were measured (e.g., heart rate, skin conductance, respiration) while they followed instructions to move specific facial muscles in ways that characterize one of the basic emotions. For instance, the instructions for a disgust expression were to (1) wrinkle your nose while keeping your mouth open, (2) pull down your lower lip, and (3) move your tongue forward without sticking it out. A study comparing North American male college students, and men from a remote illiterate culture – the Minangkabau, a relationship-oriented matriarchal society on Western Sumatra, Indonesia – yielded largely similar patterns of physiological response upon the same directed facial action (Levenson et al., 1992). Based on this finding, the authors concluded universality in physiological response patterns associated with the basic emotions.

This interpretation has been challenged. First, it is not clear what the cross-cultural similarity in physiological responses means, given that several meta-reviews of research in exclusively Western cultural settings have failed to find distinctive physiological patterns of presumed basic emotions (see, e.g., Cacioppo et al., 1993). Second, even if the directed action task yielded similarities in physiological responses, it also yielded a very significant difference between the American and the Minangkabau groups: In contrast to the American participants, the Minangkabau did not report *feeling* the matching emotion. One explanation for this difference was suggested by Hazel Markus and Shinobu Kitayama, two prominent cultural psychologists: The Minangkabau conceive of emotions as social events, and did not infer an emotional experience in the context of the directed action task, where the respondents were by themselves (Markus and Kitayama, 1994). In sum, evidence for basic emotions from physiological research is weak at best.

Basic Emotions as Universal Building Blocks

Evidence from research on the measurable features of emotion – expressive behavior and physiological responses – fails to support the basic emotion view. Note that the basic emotions tradition has not generated any research on the subjective experience of emotion; the latter was taken to be unanalyzable. In the next section, we discuss the componential view on emotions that has produced many cross-cultural studies on emotional experience.

Universal Components

Another school of thought, componential theory, focuses on the components of emotional experience (e.g., Frijda, 2007). Componential theory proposes that emotional experience can be further analyzed in terms of more basic and universal emotional components. In addition, the relationship between these components and the emotional experience is universal (Mesquita and Ellsworth, 2001): *If* the same set of components occur cross-culturally, the same emotional experiences will too.

Cross-cultural research in this area has mainly focused on the two core components of emotional experience: appraisals and action readiness. Appraisal refers to the individual's assessment of the meaning and relevance of a certain situation to the individual, and is usually represented as the pattern of outcomes on several dimensions of meaning (e.g., to what extent was the situation either pleasant or unpleasant, either conducive or harmful to self-esteem). Action readiness refers to the motivated goal in the (social) situation, and is usually represented as an urge to act in certain ways (e.g., to what extent did you feel the urge to withdraw from the situation, aggress toward the other person), or as a change in the direction of the relationship with the (social) environment (e.g., moving away, moving against). Appraisal and action readiness have been thought to constitute emotional experience as well as organize all the other components.

The Universal Component–Emotion Link

The paradigm of most componential research consists of self-report studies. Participants report real-life events in which they have experienced a given emotion (and its translations in different cultures), and then rate their experience in that same situation with respect to several emotion components: appraisal dimensions and/or action readiness modes. It is important to mention that the dimensions of meaning (appraisal, action readiness) have typically been theoretically derived by researchers from Western cultures, and that little effort has been made to establish for each culture what the most central dimensions of meaning are.

The largest cross-cultural study using this paradigm was conducted by Klaus Scherer and his colleagues, and included students from 37 different countries from all over the world (Scherer, 1997a,b; ISEAR database by Scherer and Wallbott, see http://www.affective-sciences.org/researchmaterial). The study investigates to what extent the dimensions of appraisal and action readiness mapped onto several emotions (mostly of the

kind that Ekman and colleagues had studied) in cross-culturally similar ways. It yielded substantial support for cross-cultural similarity: A similar core of appraisals and action readiness was associated with each of the seven emotions included. The study yielded some differences in the link between components and emotions as well, to which we will return (Section Emotional Frequency and Intensity).

Evidence for cross-cultural similarities in the set of appraisal (or action readiness) dimensions comes from a slightly different approach to the same data. Several researchers have examined whether the same set of component dimensions cross-culturally discriminated between different emotions, at least to some extent. They found this to be partly the case: Component dimensions accounted at the most for 40% of the variance in emotions, and this held true across different cultures.

Whereas the same set of components meaningfully distinguished between emotions in different cultures, it did not completely account for the differentiation between emotions (40% is still a long way from 100%). Moreover, across cultures, the degree to which individual appraisal and action readiness dimensions (e.g., pleasantness) discriminated between emotions was different.

Differences may have been underestimated, because all studies started from equivalent emotion words in different languages, whose selection may have been based on similarities in the associated meanings to begin with. Thus, finding similarity in the components associated with these emotion words may have merely confirmed the accuracy of translation. To our knowledge, not very many studies have shown a link between emotional components and other features of emotion. An exception is work by Anna Tcherkassof who conducted a cross-cultural recognition study among Burkina Faso and French participants, in which participants identified the action readiness associated with facial expressions. The study was inspired by Nico Frijda's theory of facial expression as the initial manifestation of action readiness. This study found that the rates of recognition, based on action readiness, did not significantly differ between cultures, and were also no worse than recognition rates obtained by studies in which participants judge the emotion expressed by a face (Tcherkassof and de Suremain, 2005). In addition to suggesting that action readiness is an important dimension of emotional meaning, because it differentiates between facial expressions, the findings also show that action readiness is no worse candidate for a universal building block than the basic emotions that were proposed by Ekman.

Understanding Indigenous Emotions from Universal Components

The use of the componential approach is perhaps better illustrated by the way it renders unfamiliar emotions comprehensible. An example comes from a study on *amae*, which is considered a 'culturally unique' emotion for the Japanese culture (Niiya et al., 2006). Amae is defined as the ability to depend and presume upon another's love or bask in another's indulgence, and cannot be translated by any one English emotion word. However, Niiya and colleagues found that North American participants responded with similar appraisals to *amae*-eliciting vignettes (i.e., a friend makes an inappropriate request for help) as were found in Japanese. Therefore, *amae* was rendered comprehensible by studying its associated components. The componential approach has provided a language to compare emotional experience cross-culturally in terms of both the personal relevance of the situation (appraisal) and the actions afforded (action readiness).

Conclusion

Appraisal and action readiness dimensions make unfamiliar emotions comprehensible, which is proof of their usefulness. Yet, a reason to be critical is that much of the research has adopted preconceived dimensions of appraisal and action readiness. Whereas some of these dimensions have a logic appeal, there is no evidence that these dimensions are the ones that are cross-culturally the most salient elements of meaning. Consistently, the combined dimensions adopted in research explain no more than about 40% of the variance in emotions, and the importance of a given dimension differs across cultures.

Toward an Understanding of Cultural Variations in Emotion: the Emotion Practice

Despite marked differences, the commonality between different approaches searching for universals (basic emotions, componential approaches) is that they lack a theory to predict or explain cultural differences in emotions. In this section, we will review research that shows the profound role of culture in the daily emotional experiences of people. The research reviewed generally assumes that emotions, or emotion components, that are 'helpful' or 'functional' in a culture will be more frequent or intense (see for detailed overviews Mesquita, 2003; Mesquita and Leu, 2007). Below, we will illustrate several lines of research that explain variation in emotional practice from central cultural differences.

Emotional Frequency and Intensity

Cultural differences exist with regard to the most frequent and intense emotions. These differences are systematic and meaningful. In one series of studies, Shinobu Kitayama and his colleagues compared emotional experience in the United States and Japan, using a culture-sensitive, bottom-up approach (e.g., Kitayama et al., 2000; Kitayama et al., 2006).

In a first study, participants from both cultures rated the similarity between different emotion words representing not only common American emotion words, but also common Japanese emotion words (and their translations). An emotional space defined by the dimensions of valence (positive, negative) and social engagement (socially engaging, disengaging) emerged from these ratings. Socially engaging emotions, such as friendliness (positive) and shame (negative), reinforce and underline the relationship. In contrast, socially disengaging emotions such as pride (positive) and anger (negative) reinforce and underline the autonomy (and superiority) of the individual.

Many previous studies have established valence as an important dimension of differentiation between emotions, but

the finding of social engagement as a dimension defining of emotions is new. Therefore, the first point to note is that the inclusion of the most common Japanese emotion words (and their translations in English) yielded a different structure of meaning than previous research that started from Western emotion words only. Social engagement is an obvious candidate for a dimension of appraisal that has not been included in most appraisal theories (but see research on the emotional dimension of *Power* or *Dominance* – which may partially overlap with the dimension of Social Engagement – e.g., Fontaine et al., 2013).

Subsequent studies by Kitayama and his colleagues yielded support for the idea that the emotions that are most helpful in coordinating the relationships in a given cultural context are experienced most frequently and intensely. These studies found that socially disengaging emotions (e.g., pride, anger) were most frequent and intense in European American cultural contexts where good relationships are defined by the autonomy and independence between individuals. In contrast, socially disengaging emotions (e.g., pride, anger) were more frequent and intense in East Asian cultural contexts where good relationships are defined by the relatedness and interdependence of the partners. Emotions that fit the cultural models are both more frequent and intense; therefore, there are cultural differences in emotional practice.

Not only the frequency and intensity of emotions as such but also of individual emotional components can be predicted from cultural models (Mesquita and Leu, 2007). For example, agency appraisals (responsibility, control) have been found in many different studies to be more frequent in Western, independent cultures where an individual's agency is deemed more central and important than in nonwestern interdependent cultures. Naturally, the occurrence of certain appraisals may also account for the occurrence of the matching emotions: Claiming responsibility would go with pride (Imada and Ellsworth, 2011)

The Role of Situations

One reason for cultural differences in emotional practice may be that, across cultures, people encounter different types of situations. For example, it is possible that people in some cultures simply receive more compliments than in other cultures, and thus experience more pride. Several studies have suggested that the situations that are likely to occur within a culture are far from random: There will be more opportunities for situations that elicit emotions that are consistent with the prevailing cultural models.

In a series of comparative survey studies from our own lab, we predicted the frequency of a set of emotion-eliciting situations from (1) their potential to elicit a given emotion and (2) the value of that emotion within the particular culture (Boiger et al., 2013; Boiger et al., 2014). The research focused on shame and anger. In a comparison between the United States and Japan, the researchers expected that anger situations in the United States would be afforded, because anger is a hallmark of autonomy and independence. In contrast, anger was expected to be avoided in Japan, where it violates the interdependent model of relatedness and harmony. Conversely, shame was expected to be condoned in Japan, where it is consistent with the ideal of self-reflection and self-improvement, whereas it was expected to be condemned in the United States, because it undermines positive self-regard, which is a cornerstone of the US cultural model. Consistent with these expectations, the researchers found that the higher the potency of a situation to elicit (high intensity) anger, the higher its frequency was rated in the United States, and the lower in Japan; conversely, the higher the potency of a situation to elicit (high intensity) shame, the lower its rated frequency in the United States, and the higher in Japan. More refined analyses of the differences between cultures may thus serve to predict which situations are culturally promoted, and which situations are avoided.

Research by Jeanne Tsai and her colleagues has found that individuals seek out those emotions that afford important cultural tasks: influencing in the United States, and adjusting in East Asian cultures. Consistently, European Americans 'ideally wanted to feel' more high-activation positive states than East Asians, and East Asians 'ideally wanted to feel' more low-activation positive emotional states than European Americans (e.g., Tsai et al., 2006). In one within-culture experimental study, Tsai and colleagues found clear evidence that ideal emotions were associated with the task at hand. American participants valued high-arousal positive states more after being assigned the task to influence others, and low-arousal positive states after they were asked to adjust to others (i.e., suppress personal needs and change their own behaviors to meet others' needs; Tsai et al., 2007).

Across studies, people select situations that will lead to the culturally valued emotions. When given a choice, European Americans in an experimental setting consistently selected artifacts (e.g., CDs) that elicited high-arousal over those eliciting low-arousal activities, whereas the reverse was true for East Asians (Tsai et al., 2007: Study 4). Societal statistics confirmed these experimental results: North Americans consistently preferred recreative drugs, leisure activities, and music preferences that elicited high-arousal positive emotions over those eliciting low-arousal positive emotions, whereas there was a preference for activities that elicit low-arousal positive emotions in East Asian societies. Moreover, within a North American sample, individual differences in ideal affect in fact predicted preferential activities (leisure time, drugs, types of vacations) in ways that mirrored cultural differences between North Americans and East Asians (Tsai, 2007).

The Role of Concern Salience

Why people evaluate a situation as pleasant or unpleasant differs in ways that more or less follow from the important cultural models of independence and interdependence. In experience sampling research, Mesquita and Karasawa (2002) found that in an American group of college students, events were rated as pleasant (or, to the contrary, unpleasant) to the extent they were relevant to independent concerns, such as self-esteem and being in control. In contrast, pleasantness in a Japanese group was judged as much by independent as by interdependent concerns; examples of the latter are feelings of closeness and other people's (im)moral behavior.

Consistently, European Americans and Asian Americans showed enhanced emotional reactivity when the dominant aspects of their self-concept were activated (Chentsova-Dutton and Tsai, 2010). In experimental research, priming *individual* aspects of the self led to greater levels of positive emotions and smiling in European Americans, whereas priming of *relational* aspects of the self was associated with greater emotional reactivity in Asian Americans. Emotions are more intense when they are about culturally focal concerns (including culturally central aspects of the self).

Recent research from our own lab has found that culturally focal concerns do not only affect the intensity of emotions, but also the *types* of emotions (De Leersnyder et al., submitted). Belgian and Turkish respondents indicated how relevant emotional situations were to both self-focused values (e.g., personal success) and other-focused values (e.g., loyalty). Across studies using different methods, we found that situations that self-focused values elicited more disengaging emotions (pride, anger), whereas other-focused values elicited more engaging emotions (friendliness, shame). In addition, we found that the frequency with which these values were perceived as relevant to Belgian students' emotional situations exactly mirrored young Belgians' value hierarchy (i.e., most important values as 'guiding principles in people's life'), as obtained from a national representative sample by the European Social Survey (ESS round5; Norwegian Social Science Data Services, 2012; De Leersnyder and Mesquita, in preparation). Together, these findings suggest that cultural differences in the salience of certain values may be one of the reasons for differences in the rates and intensity of engaging and disengaging emotions.

Feeling Culturally Appropriate Emotions Is Rewarding

If emotions of people from different cultures are different, then, by the same token, emotions of people within one culture should converge. In fact, recent research has suggested that the emotional experiences of immigrants over time grow more similar to the emotions of people from the new culture; and that intergroup contact promotes this emotional acculturation (see, e.g., De Leersnyder et al., 2011).

We have suggested that cultural differences in emotional phenomena can be understood from their functionality to the culture. What is the evidence that this is so? Several studies have found evidence that fitting in, or having the same emotions as most others in your culture, is beneficial; or conversely, that not having the emotions that are culturally aspired can be harmful. For instance, subjective well-being (good feelings) was more closely related to engaging positive emotions (e.g., friendly feelings, respect) in Japan, but to disengaging emotions (e.g., pride, superiority) in the United States (Kitayama et al., 2006). In more recent research from our own lab, we found that emotional fit with the cultural average predicted relational well-being; this was true in three different cultures: the United States, Belgium, and Turkey (De Leersnyder et al., 2014). Finally, Tsai et al. (2007) found that depression, the opposite of well-being, was predicted by the differences between culturally ideal emotions and a person's actual emotions. In the Chinese sample the relative absence of *low*-activation positive emotions predicted depression, but in the American sample it was the relative absence of *high*-activation positive emotions. In sum, individuals have better outcomes when they are more similar to others in their culture.

Emerging Trends

Research on culture and emotion has moved beyond the dichotomous question of whether emotions are nature or nurture, and has started to investigate more nuanced and detailed questions. It is clear that some of the building blocks of emotions may be universal in ways that make it possible to largely comprehend emotional lives in other cultures, even if they are very different from our own. It is equally clear that the emotional lives of people in different cultures vary in systematic, predictable, and meaningful ways. Moreover, the cultural differences are dictated by, what could be considered, universal rules: People have emotions that are helpful in their own cultural context. There is plasticity in emotions, even in later life: When immigrants move to a new cultural context, their emotions slowly adapt.

Whereas our knowledge of cultural similarities and differences in emotions has become much more detailed and systematic, insight into the processes that lead to cross-culturally different emotional lives is still very limited. These processes constitute the challenge of future research on culture and emotion. Two emerging trends will be discussed below.

Emerging Trends: Emotions as Momentary Constructions

Emotions are increasingly seen as constructions in the moment (Barrett, 2006; Mesquita and Boiger, 2014), rather than pre-established entities, with the implication that an emotion concept may be constructed differently across both contexts and cultures. In this regard, the finding from the Directed Action Task that American men and Minangkabau men did not construct their facial expressions to signal a subjective feeling of emotion is interesting (Levenson et al., 1992). The finding suggests that the construction of an emotion is not just bottom-up, as in constituted by appraisals, action readiness, and self-perceived facial expression, but also top-down, guided by the culture's lay theory of emotion. The idea is consistent with psychological constructivist theories of emotion (e.g., Barrett, 2006).

Another consequence of this view that emotions are constructed in the moment is that emotions are differently constituted at different moments in time (even within one culture). To do justice to the complex nature of emotional experience, we recently conducted a study that established the variety of patterns of appraisal and action readiness that were associated with anger and shame in three different cultures: United States, Belgium, and Japan. In each of these cultures, we observed the same three varieties of patterns for each emotion, but not at the same rates. The dominant patterns (of appraisals and action readiness) differed by culture, such that the culture of the participant was accurately predicted in 40–50% of the cases (Boiger et al., in preparation).

Emerging Trends: Emotions as Relational Constructs

A second trend emerging in (cultural) research on emotions is to study emotion in their social and relational context (Butler, 2011; Mesquita, 2010; Mesquita and Boiger, 2014). In particular, researchers have started to study how this social context helps to create and regulate the fit of emotions with cultural norms, and also, in the long run, how it helps to develop culturally adaptive emotional dispositions. The best examples come from developmental psychology. In one cross-cultural study in Germany and Japan, Trommsdorff and Kornadt (2003) compared the emotional interactions between mothers and their children when the child had disobeyed. They found that German mothers readily attributed blame to their children, therefore reacting with anger. An escalation of anger and resistance would ensue. The conflict ended unresolved, leaving both the mother and the child angry and hurt. A very different interaction pattern was observed in Japanese mother–child dyads: Japanese mothers were more likely to interpret disobedience empathetically ("the child is just a child, is too much absorbed in playing, is too tired"; Trommsdorff and Kornadt, 2003: 296), and to react in an accommodating manner. The Japanese mothers remained friendly, though they kept insisting that the child obeyed. Even though repeated failure of the child to give in would get them disappointed, they avoided putting relational harmony in jeopardy. In the end, both parties started making partial concessions, thus protecting the desired feeling of oneness (ittaikan), and reinstating harmony. Both interactions were constructed based on cultural models of self and relationship. German mother–child interaction accomplished the cultural model of independence and autonomy by influencing and asserting oneself, whereas Japanese interactions accomplished harmony and relatedness by adjustment and empathy.

The mother–child interactions had long-term effects on children's relational patterns: In both cultures, the escalation of interactions in early childhood negatively predicted the level of empathy-based altruism and positively predicted aggression 9 years later. These interaction patterns can, therefore, be understood as necessary socialization experiences: Independent German children learn to assert themselves, whereas interdependent Japanese children learn to accommodate and to place relational harmony over individual desires. This example vividly shows how social context and emotions are intricately interwoven in ways that are consistent with the prevalent cultural meanings.

This example shows how engagement in culturally particular social interactions and relationships directly shapes emotions (cf. Kitayama et al., 2009). Therefore, culture's influence on emotions is not merely mediated by beliefs (e.g., Matsumoto, 1990), but rather afforded and scaffolded by social interactions as they evolve in a particular culture. It is the future of culture and emotion research to further probe the ways in which culture is produced and reproduced in emotional interactions, a process known as the 'microgenesis of culture' (Kashima, 2008).

See also: Affect-Regulation Motivation; Collectivism and Individualism: Cultural and Psychological Concerns; Counseling and Psychotherapy: Ethnic and Cultural Differences; Cross-Cultural Psychology; Cross-Cultural Research Methods in Psychology; Cross-Cultural Training in Industrial and Organizational Psychology; Cultural Influences on Interpersonal Relationships; Cultural Psychology; Culture and the Self: Implications for Psychological Theory; Early Emotional Development and Cultural Variability; Emotion and Expression; Emotion in Cognition; Emotion, Perception and Expression of; Emotional Development, Effects of Parenting and Family Structure on; Emotional Intelligence and Competencies; Emotional Regulation; Emotions and Health; Emotions and Work; Emotions, Psychological Structure of; Facial Emotion Expression, Individual Differences in; Immigration: Social Psychological Aspects; Migration, Ethnicity, Aging and Social Work Practice; Mother-Child Communication: Cultural Differences; Self and Emotional Development in Adulthood and Later Life; Selfconscious Emotions, Psychology of; Social and Emotional Development in the Context of the Family; Subjective Wellbeing and Culture; Values, Psychology of.

Bibliography

Barrett, L.F., 2006. Solving the emotion paradox: categorization and the experience of emotion. Personality and Social Psychology Review 10, 20–46.

Boiger, M., De Leersnyder, J., Uchida, Y., Norasakkunkit, V., Ceulemans, E., Mesquita, B., Cultural Variation Is Real: Experiencing Anger and Shame in the United States, Japan, and Belgium, manuscript in preparation.

Boiger, M., Mesquita, B., Uchida, Y., Barrett, L.F., 2013. Condoned or condemned: the situational affordance of anger and shame in the United States and Japan. Personality and Social Psychology Bulletin 39, 540–553.

Boiger, M., Güngör, D., Karasawa, M., Mesquita, B., 2014. Defending honour, keeping face: interpersonal affordances of anger and shame in Turkey and Japan. Cognition and Emotion 28 (7), 1255–1269.

Butler, E.A., 2011. Temporal interpersonal emotion systems: the "TIES" that form relationships. Personality and Social Psychology Review 15, 367–393.

Cacioppo, J.T., Klein, D.J., Berntson, G.G., Hatfield, E., 1993. The psychophysiology of emotion. In: Lewis, R., Haviland, J.M. (Eds.), The Handbook of Emotion. Guilford Press, New York, pp. 119–142.

Chentsova-Dutton, Y.E., Tsai, J.L., 2010. Self-focused attention and emotional reactivity: the role of culture. Journal of Personality and Social Psychology 98 (3), 507–519.

De Leersnyder, J., Mesquita, B., What Are Emotions about? The Role of Cultural Values in Emotion Elicitation, manuscript in preparation.

De Leersnyder, J., Mesquita, B., Kim, H., 2011. Where do my emotions belong? a study of immigrants' emotional acculturation. Personality and Social Psychology Bulletin 37 (4), 451–463.

De Leersnyder, J., Mesquita, B., Kim, H., Eom, K., Choi, H., 2014. Emotional fit with culture: individual differences in emotional concordance predict relational well-being. Emotion. http://dx.doi.org/10.1037/a0035296.

De Leersnyder, J., Mesquita, B., Koval, P., Kuppens, P., Emotions and Concerns: Situational Evidence for Their Systematic Co-occurrence, manuscript submitted for publication.

Ekman, P., 1992. Are there basic emotions? Psychological Review 99 (3), 550–553.

Ekman, P., Friesen, W.V., 1969. The repertoire of nonverbal behavior—Categories, origins, usage and coding. Semiotica 1, 49–98.

Elfenbein, H.A., Ambady, N., 2002. On the universality and cultural specificity of emotion recognition: a meta-analysis. Psychological Bulletin 128, 203–235.

Fontaine, J.R.J., Scherer, K.R., Soriano, C. (Eds.), 2013. Components of Emotional Meaning: A Sourcebook. Oxford University Press, Oxford.

Frijda, N.H., 2007. The Laws of Emotions. Erlbaum, New York.

Gendron, M., Roberson, D., van der Vyver, J.M., Barrett, L.F., 2014a. Cultural relativity in perceiving emotion from vocalizations. Psychological Science 25 (4), 911–920.

Gendron, M., Roberson, D., van der Vyver, J.M., Barrett, L.F., 2014b. Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. Emotion 14 (2), 251–262.

Imada, T., Ellsworth, P.C., 2011. Proud Americans and lucky Japanese: cultural differences in appraisal and corresponding emotion. Emotion 11, 329–345.

Jack, R.E., Caldara, R., Schyns, P.G., 2011. Internal representations reveal cultural diversity in expectations of facial expressions of emotion. Journal of Experimental Psychology, General 141, 19–25.

Kashima, Y., 2008. A social psychology of cultural dynamics: examining how cultures are formed, maintained, and transformed. Social and Personality Psychology Compass 2, 107–120.

Keltner, D., Haidt, J., 1999. Social functions of emotions at four levels of analysis. Cognition and Emotion 13 (5), 505–521.

Kitayama, S., Markus, H.R., Kurokawa, M., 2000. Culture, emotion, and well-being: good feelings in Japan and the United States. Cognition and Emotion 14, 93–124.

Kitayama, S., Mesquita, B., Karasawa, M., 2006. The emotional basis of independent and interdependent selves: socially disengaging and engaging emotions in the US and Japan. Journal of Personality and Social Psychology 91, 890–903.

Kitayama, S., Park, H., Sevincer, H.T., Karasawa, M., Uskul, A.K., 2009. A cultural task analysis of implicit independence: comparing North america, western Europe, and east Asia. Journal of Personality and Social Psychology 97 (2), 236–255.

Levenson, R.W., Ekman, P., Heider, K., Friesen, W.V., 1992. Emotion and autonomic nervous system activity in the Minangkabau of west Sumatra. Journal of Personality and Social Psychology 62 (6), 972–988.

Markus, H.R., Kitayama, S., 1994. The cultural construction of self and emotion: Implications for social behavior. In: Kitayama, S., Markus, H.R. (Eds.), Emotion and Culture: Empirical Studies of Mutual Influence. American Psychological Association, Washington, DC, pp. 89–130.

Matsumoto, D., 1990. Cultural similarities and differences in display rules. Motivation and Emotion 14 (3), 195–214.

Matsumoto, D., Yoo, S.H., Fontaine, J., Anguas-Wong, A.M., Arriola, M., Ataca, B., et al., 2008. Mapping expressive differences around the world: the relationship between emotional display rules and individualism versus collectivism. Journal of Cross- Cultural Psychology 39 (1), 55–74.

Mesquita, B., 2003. Emotions as dynamic cultural phenomena. In: Davidson, R., Goldsmith, H., Scherer, K.R. (Eds.), The Handbook of the Affective Sciences. Oxford University Press, New York, pp. 871–890.

Mesquita, B., 2010. Emoting: a contextualized process. In: Mesquita, B., Barrett, L.F., Smith, E.R. (Eds.), The Mind in Context. Guilford, New York, pp. 83–104.

Mesquita, B., Boiger, M., 2014. Emotions in context: A socio-dynamic model of emotions. Emotion Review 6 (4), 298–302.

Mesquita, B., Ellsworth, P.C., 2001. The role of culture in appraisal. In: Scherer, K.R., Schorr, A. (Eds.), Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, New York, pp. 233–248.

Mesquita, B., Karasawa, M., 2002. Different emotional lives. Cognition and Emotion 16 (1), 127–141.

Mesquita, B., Leu, J., 2007. The cultural psychology of emotions. In: Kitayama, S., Cohen, D. (Eds.), Handbook for Cultural Psychology. Guilford Press, New York, pp. 734–759.

Niiya, Y., Ellsworth, P.C., Yamaguchi, S., 2006. Amae in Japan and the United States: an exploration of a "culturally unique" emotion. Emotion 6 (2), 279–295.

Sauter, D.A., Eisner, F., Ekman, P., Scott, S.K., 2010. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. PNAS 107 (6), 2408–2412.

Scherer, K.R., 1997a. The role of culture in emotion antecedent appraisal. Journal of Personality and Social Psychology 73, 902–922.

Scherer, K.R., 1997b. Profiles of emotion-antecedent appraisal. Cognition and Emotion 11 (2), 113–150.

Solomon, R.C., 2002. Back to basics: on the very idea of "basic emotions". Journal for the Theory of Social Behaviour 32 (2), 115–144.

Tcherkassof, A., de Suremain, F., 2005. Burkina Faso and France: a cross-cultural study of the judgment of action readiness in facial expressions of emotions. Psychologia 48 (4), 317–334.

Trommsdorff, G., Kornadt, H., 2003. Parent-child relations in cross-cultural perspective. In: Kuczynski, L. (Ed.), Handbook of Dynamics in Parent–Child Relations. Sage, Thousand Oaks, CA, pp. 271–305.

Tsai, J.L., Knutson, B., Fung, H.H., 2006. Cultural variation in affect valuation. Journal of Personality and Social Psychology 90 (2), 288–307.

Tsai, J.L., 2007. Ideal affect: Cultural causes and behavioral consequences. Perspectives on Psychological Science 2 (3), 242–259.

Tsai, J.L., Miao, F.F., Seppala, E., Fung, H.H., Yeung, D.Y., 2007. Influence and adjustment goals: sources of cultural differences in ideal affect. Journal of personality and social psychology 92 (6), 1102–1117.

Culture and Institutional Logics

Patricia H Thornton, Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article defines key assumptions of the institutional logics perspective (ILP) particular to the relationship between institutional logics and culture. It discusses central debates in the sociology of culture to highlight affinities between theoretical issues in the ILP and the sociology of culture. Implications are discussed for advances in the sociology of organizations and culture and research in the substantive domains of morality, justification, and symbolic management.

The institutional logics perspective (ILP) is an emerging and rapidly developing research area in organization and management theory (Thornton et al., 2012). Even though it is a meta-theory of cultural systems and micro- and macrobehavior, it remains a relatively untapped framework in the study of culture. This article considers debates among cultural sociologists on the internalization and objectified status of culture and describes the ILP with the goal of suggesting affinities and potential synergies between the two domains.

To lay the ground work for this discussion, this article is organized first to describe and highlight key concepts and mechanisms in the ILP and then to delineate schools and dividing issues in cultural theory. A key theme in discussing these two perspectives is to link discussion of the issues of internalization/externalization in cultural sociology to micro-foundations and exteriority in the ILP. In short the debate in cultural sociology revolves around how much cultural content is attributable to mechanisms internal to the individual as distinct from being external in society. Contemporary cultural sociologists suggest relinquishing the fixation on the internal mechanisms of socialization and cognitive attributions and focusing outside the person and toward the external world in which the individual is embedded (Lizardo and Strand, 2010). The ILP is a scaffolding to understand both the interiority and exteriority of culture. It is the exteriority mechanisms that separate the institutional logic perspective from the bulk of prior neo-institutional theory and provide it with a theory of agency. The sociology of culture work has promised to elaborate and advance these theoretical issues. This article closes with a discussion of the implications for research on the issues of morality, justification, and symbolic management.

The Institutional Logics Perspective

The ILP presents theory and methods for understanding how to embed persons and organizations in their institutional environments (society). It is a multilevel meta-theory useful in studying the role of culture in both traditional and secular societies. "Institutional logics are the socially constructed patterns of cultural symbols and material practices, assumptions, values, beliefs, and rules by which individuals and organizations produce and reproduce their material subsistence, organize time and space, and provide meaning to their social reality" (Thornton and Ocasio, 1999: p. 804). According to this perspective, societies can be conceptualized as an interinstitutional system of institutional orders. In theory these orders are cultural subsystems in which culture is more likely to cohere within than across orders, for example, the family, community, religion, market, state, professions, and the corporation. This framework provides a comparative heuristic to theorize and measure culture as both an independent and a dependent variable. Cultural symbols and practices can emerge, change, and remain stable depending on the levels of contradiction and contestation in logics across institutional orders and how these logics become activated and thus available and accessible to individuals and organizations to use and manipulate.

A key distinction of the ILP from that of Bourdieu's (1990) work is that it does not necessarily see cultural processes as embedded in power struggles and material inequality. The ILP is agnostic on the question of inequality, which in theory could produce stability and cooperation, rather than struggle. For example, material inequality could be relatively culturally acceptable as in the case of the British respecting and preserving their monarchy; whereas citizens of other nations may view such reverence to an extreme standard of inequality with contempt. This example highlights that power struggles are culturally contingent in explaining the coherence of societies.

Stinchcombe (2002: p. 429) has commented on how to theorize the question of power in ways that are compatible with the ILP. He argued that, "an understanding of culture is needed to define the meaning of power and competition and that in building theory this question presents a causal ordering problem. If power is theorized as a first-order construct in explaining institutional change, independently of culture, two problems need to be addressed. First, power is created in the course of action; it does not occur prior to the action that it explains. Second, the decision to use power as an intentional strategic choice may be problematic because it is not always possible for actors to know the cultural framing or menus of available options in advance of any action" (Thornton et al., 2012: p. 65).

It should be stressed that the mental contents of the institutional orders are of both individual and collective origin, hence developmental over time. This contrasts and departs from Durkheim's notion of 'religious phenomena' and Weber's understanding of 'value spheres' in which the cognitive contents were not open to revision in light of the progress of individuals' experience. According to the ILP, actor's practical experience is a source and a building block of the institutional

orders and logics, just as institutional orders and logics have some hand in shaping the minds of individuals and the behavior of organizations through both symbolic and material mechanisms.

There is a significant amount of empirical research to support the development of the ILP (see Thornton and Ocasio, 2008 and Thornton et al., 2012, for reviews). Thornton and Ocasio's (1999) integration of the meso- and macrolevels of culture demonstrated how institutional logics affect organizational actors' attention in defining a problem and recognizing a solution. They showed how a shift in attention to different institutional logics changed the determinants of executive succession in the US higher education publishing industry. Companion articles expanded the scope conditions to other organization decisions (Thornton, 2001, 2002). Using a combination of qualitative and quantitative methods and models, this research partitioned the analysis of cultural meaning from structural effects (see Thornton, 2004: Chapter 8). For example, it showed the number of corporate acquisitions (structural) remained the same under the era of a professional logic (controlled by editors) as compared to a market logic (controlled by corporate headquarters), but the meaning (cultural) of and consequences of acquisition activity changed.

Thornton et al. (2012) discuss how each institutional order is made up of categorical elements that form the categorical elements of institutional content, for example, the sources of legitimacy, norms, values, and practices, which can be expressed in symbolic and material form. Ideal types are used to illustrate society as an interinstitutional system, locating the seven institutional orders of societies on the X-axis (family, community, religion, market, state, professions, and the corporation) and the categorical elements of institutional content on the Y-axis. In theory the interactions between the categorical elements on the Y-axis and the institutional orders on the X-axis constitute an institutional logic. (There are many illustrative examples of these matrices used for analytic purposes in different substantive contexts, see in particular Thornton, 2004; Thornton et al., 2005, 2012; Thornton and Ocasio, 2008.) A key difference between an institutional logic and an institutional order is that logics can be constituted from categorical elements of different institutional orders (hybrids). That is, there is a partial autonomy and near-decomposibility of the interinstitutional system (Thornton et al., 2012: p. 59). This can result from the activities of institutional entrepreneurs and from structural overlap and event sequencing (Thornton et al., 2005). Institutional orders and logics stem from such institutionalizing activities over the course of history.

The ILP assumes the interpretation of the Y-axis elements differ depending on the observational lens – from which X-axes is the camera focused – the interaction effects are represented in the cells of the matrix. Axis or axes may be the referent because singular or multiple institutional orders may be influencing an individual or an organization. In theory ILP does not assume a priori the number of order(s) or logic(s) that are likely to influence behavior in a particular context or whether or not a single logic is more dominant than another or that multiple logics are necessarily operative. In theory no one institutional order has causal primacy a priori, though the origin and prevalence of institutional orders is temporally and contextually dependent, for example, the family is an older institution than the state and religion and family logics in general are more influential in traditional societies. Tracey (2012: p. 118) comments it is a nonissue to frame a contribution to the literature on the basis of the number of institutional logics operating in a given context.

The comparative anchoring of cultural symbols and material practices to the root institutional orders is important in winnowing all of societies' potential influences so as to zoom in on the underlying or more salient meanings of cultural content. For example, family values and practices will differ from market values and practices. The degree of conflict and interdependence among the categorical elements of the institutional orders also differs, for example, the values and practices of the market and the family are likely more opposing whereas those of the state and the professions are likely more interdependent, particularly if the analysis is of Western societies. For example, if an individual's or an organization's environment is more steeped in family values and practices rather than market values and practices, the interpretation of symbols and the enactment of practices is likely to differ. These type of effects have been well demonstrated at the organizational level in qualitative and quantitative large sample controlled studies and are beginning to be demonstrated at the individual-society levels of analysis in experimental research (Glaser et al., 2012).

The microfoundations of institutional logics are less well developed and this is a growth area where new work is emerging and needs to consider not just organization- but individual-level actors. This work draws on attentional theories rooted in cognitive and social psychological research (dynamic constructivism) and the Carnegie School. Dynamic constructivism theory was developed to explain which culture individuals rely on for their behavior. Institutional logics focus the attention of actors through cultural embeddedness, activating a social actor's situated identities, goals, and action schemas for decision making, sense making, and mobilization (Ocasio, 1997, 2011; see Thornton et al., 2012: Chapter 4). Nigam and Ocasio (2010) examined the cross-level effects of how critical events activate and make institutional logics available and accessible to actors for sense making. Thornton et al. (2012) illustrate how entrepreneurs discover new ideas and solve problems in legitimating and justifying their new organizations by switching and recombining institutional logics. When entrepreneurs switch institutional logics they are shifting the sources of analogies and referent categories (think bits of culture) by which people can understand and practice their innovation (Lounsbury and Glynn, 2001; Durand et al., 2013). Examining the creation of banks, Almandoz (2012) reported that founding teams that had functional backgrounds (internal) associated with a community logic, as distinct from a market logic, had higher founding rates and lower rates of founding team dissolution (material practice). Rather than focusing on the divisive debate on internal versus external mechanisms, this research advances theory by measuring dualities of cultural mechanisms.

To solve the problem of a lack of agency in a structural neo-institutional theory (DiMaggio and Powell, 1983), it was necessary to show that individual interests and power were

not institution (culture) free; that interests and power had both a material and symbolic aspect in which actors' interests could be understood independently of actors' understandings (Thornton et al., 2012: p. 41). Meyer and Rowan (1977: p. 45) set the stage for this advance in developing the concept of loose coupling in organizations and their claim that the "building blocks for organizations come to be littered around the societal landscape: it takes only a little entrepreneurial energy to assemble them into a structure." This exteriority of institutions from organizational actors, or in the parlance of cultural sociology, externalization, allowed the development of the ILP as a levels theory of cultural institutions and paved the way to cut loose institutional theory from its overly structural bondage. "This is an important issue because a theory of institutions needs to recognize that individuals and organizations are embedded in social structures, yet also explain how they are externalized or partially autonomous, allowing them to construct institutions socially" (Thornton et al., 2012: p. 72). Loose coupling assumes that internalization can occur through both cognition and practice – it permits identification with symbols rather than learning through practice. Emerging work is theorizing and testing these ideas at the individual-society levels of analysis (Klyver and Thornton, 2014).

Debate in Cultural Sociology

A vibrant debate in cultural theory is raging between the classical and the embodied theorists. Overall the central issue stems from the question of how much culture is internalized in the person (i.e., stored through cognitive mechanisms) versus how much culture is objectified in external publicly accessible formats (i.e., stored in language, vocabularies of motive, discourse, accounts, justifications, rituals, media, artifacts, codes, categories, institutions, and other ways of framing experience) (Lizardo, 2012).

Classical theory focuses on the derivation of individual-level cognition from abstract cultural patterns external to the mind, conceiving of culture as external and organized as an emergent order of symbolic systems (Durkheim, 2005; Parsons, 1972, 1968; Zerubavel, 1999; Alexander, 2003; DiMaggio, 1997; DiMaggio et al., 2013). Variants of the classical approach differ in their attention to cultural content, the functionalists focusing more on norms and values and the phenomenologists on cognition. According to the classic view, meaning derives from collective symbolic categories learned through culture. This requires specifying the mechanisms explaining how an individual's mental representations are linked to material or mnemonically available symbolic systems to create sense making for the individual (Douglas, 1986). In theory persons need symbolic cultural structures to be able to generalize across sensory events and to characterize their present experience so that it can be used to understand future experiences.

But, how do sensory events as a singular example of a larger class of entities get typified in order to be meaningful to the person? Parsons (1964, 1972) argued this occurs because people through their familiarity with a system of generalized cultural symbols learn to typify their experience, i.e., through cognitive socialization, which is a metaphorical learning process for the transmission of culture, for example, in the family environment. Thus, for an individual to be socialized, something external or social has to be converted into something internal or personal (see Lizardo, in this volume).

In contrast, the embodied approach views culture as tightly coupled to peoples' practical action grounded in their relations to the material environment (Lizardo, 2012; Lakoff and Johnson, 1999; Gallese and Lakoff, 2005). It draws its ideas from the intersections of cognitive science, cultural and cognitive anthropology, developmental psychology, robotics, and the philosophy of mind. Embodied views of cognition challenge classical perspectives to explain (1) how internalization occurs and (2) how people use cultural symbols to derive meaning.

In his contribution, Culture, Cognition, and Embodiment Lizardo argues that the problem the classical view raises for embodied theorists is that the mechanisms of internalization are not well specified and that existing knowledge structures may not stem from experience. It is implausible that individuals are subject to such a massive systemic process of socialization into the symbolic cultural system. Classical theorists push back arguing that accounting for only sensory events is tantamount to empiricism, which results in irrationalism. Classical theorists do not deny that sensory stimulation exists, but argue it alone is not a sufficient condition for individuals to derive meaning (see Lizardo's, in this volume, for the elaborated argument).

Embodied theorists reject that a person's cognitive experience is constituted via the internalization of external symbols. Internalization requires the transmission of cultural content from person to person and from the environment to person. It requires strong assumptions about socialization processes through which individuals internalize norms and values and other forms of cultural information. In Lizardo and Strand's (2010: p. 208) words it requires belief "in some predetermined set of internalized cultural objects which are mysteriously downloaded through unseen, undertheorized, underspecified mechanisms from an unseen, undertheorized, underspecified cultural ether." People just can't remember and keep straight all the relevant logical linkages that would need to be connected to make the theory credible (Martin, 2010). It is not analytically parallel to argue that mental experience can be symbolic in the same sense that external cultural symbols are symbolic.

Alternatively, embodied theorists argue that the phenomenological contents of mental experience are directly meaningful and do not require symbolic translation. The mind cannot be thought of as a repository of symbols. Meaning derives from repeated experience that is supported by neural structures that made the experience possible in the first place (Barsalou, 1999). Embodied theorists cite historical evidence that meaning was possible in prelinguistic societies and hence cultural symbols do not need to exist to produce meaningful mental experiences.

In embodied approaches to cognition, cultural symbols do not constitute mental experience, a notion rooted in structuralism, or make sense of it, a legacy from phenomenology, but instead are used to exploit mental experience. "Cultural symbols do not mean anything, instead they serve as prompts for persons to engage in individual or collective acts of meaning construction" (see Lizardo's, in this volume: p. 0110). The potential for meaning construction is based in experience

(Bourdieu, 1990), it is not based on the internal access to cultural symbols. Cultural symbols are necessary but only to coordinate access to delimited chunks of conceptual knowledge in action (see Lizardo's, in this volume: p. 0110).

Swidler (1986, 2001) is credited with igniting the debate between the 'classical' and the 'embodied' approaches by criticizing the reigning Parsonian functionalist view with the introduction of the tool kit perspective. She argued that "culture does not influence how individuals and groups organize action via enduring psychological proclivities implanted in individuals by their socialization. Instead, publicly available meanings facilitate certain patterns of action, making them readily available, while discouraging others" (Swidler, 1986: p. 283). What actors imbibe from the social environment is a set of heuristics, hunches, and shallow practical skills that effectively enable people to interact with the external institutionalized elements such as norms, values, and other cognitive schemata to generate coherent behavior (Swidler, 2003). The key finding motivating tool kit theory is that if actors supposedly internalized culture, then why are they unable to produce a coherent regurgitation of their actions. Thus, Swidler's argument shifts much of actors' cognitive activity to external institutions (Lizardo and Strand, 2010). She argues that persons engage in improvisation, using their cultural competencies to draw from a tool kit of partly internal and partly externalized forms of cultural content. Swidler's tool kit idea initially lacked systematically propositional and empirical support, which more generally plagued cultural sociology (DiMaggio, 1990). However, the tool kit idea gained tremendous appeal particularly in the wake of DiMaggio (1997, 2002: p. 275) arguing its resonance with the findings of controlled studies in social and cognitive psychology.

Bourdieu's (1990, 2000) practice theory argues that both the cognitive and institutional are externalized. By externalized I refer to Zucker's (1977) classic experiment that confirmed culture as social knowledge; once institutionalized culture exists as a fact and therefore part of objective reality, which can be transmitted directly without being internalized by individuals and transmitted through them to other individuals and organizations. However, practice theory emphasizes that individuals are to a significant degree influenced by having a history of recurrent experiences in institutional environments. It is the match between actors' 'socialized' dispositions and external institutional environments that produces coherence in social action and differentiates Bourdieu's practice theory from the 'poverty of cognition,' of tool kit theory. Practice theory argues to look for an explanation of the system of actors in culture through sequences of choices made at the level of practice.

From my reading of their work, two well-known culture scholars, Vaisey (2009) and DiMaggio (1997) have taken a middle road in the classic/embodied debate and in elaborating on Swidler's argument on the externalization of culture. DiMaggio (1997, 2002: p. 276) engages the duality of both internalization and externalization of culture by arguing that persons "store information, opinion, and attitudes organized through schemata." Schemata are images of objects, actions, or events, and the linkages among these represented in stereotyped behavioral routines. Social schemata provide frameworks that people use to interpret new information. People have access to these schemata in varying degrees depending on how central they are to a person's self-image, emotional character, as well as their salience, frequency, and recency with which the environment has activated them (D'Andrade, 1992). "Schemata are themselves organized into relatively independent domains among which there are not necessarily homologic relations" (DiMaggio, 2002: p. 276).

Psychologists have learned that people have two different types of cognition, characterized broadly as deliberate and automatic (Metcalfe and Michel, 1999). DiMaggio (2002) suggests this is evidence of Parsons' belief that orientations to action are variable, which is contrary to ethnomethodology and Bourdieu's (1990) practice theory. DiMaggio (2002) points out that schematically organized knowledge that is internalized by persons is affected by domain independence – that is different classes of life situations (practice), for example, those related to work and family. It is domain independence, which makes people a lot less consistent than expected. Personality resides in interactions between behavioral dispositions and social situations (Mischel and Shoda, 1995). Hence, the lesson here is that cultural effects lie in the interactions of subgroup domain and personal disposition – rather than values, groups may have predilections to act in certain ways in specific situations. DiMaggio et al.'s (2013) recent use of topic models is an attempt to analyze data about the environment of cultural representations with which the person's mental structure (network of schema) interacts (see DiMaggio, 2002: pp. 278–279 for a discussion of the psychological research supporting these ideas).

Vaisey (2009) bridged the duality of socialization and tool kit theories by developing dual process theory to illustrate how culture is a repository of values that motivate individuals' intentions; these values are transmitted to individuals via socialization from institutions such as family and religion (Parsons, 1964). Culture is also a loosely coupled repertoire of rhetorical tools that individuals use to rationalize, justify, and make sense of their life situations and choices – justifications being a form of symbolic management.

Lizardo's view (in this volume) on procedural theory is inspired from practice theory (Strauss and Quinn, 1997; Bourdieu, 2000) and ethnomethodology (Garfinkel, 1967). According to procedural theory, persons are not cognitively capable of internalizing (e.g., learning, memorizing, introjecting) cultural content or symbolic representations (Martin, 2010). Culture can only be embodied as skill and only through practice and experience. With procedural theory Lizardo (2012: p. 82) sets out to disband the presumption that persons carry in their heads cultural content. Citing Zaller (1992) and Barsalou (1987) Lizardo argued, "there is little evidence in the social and cognitive sciences that such structured mental representations as attitudes, opinions, or concepts exist as stored entities in long term memory." Persons can become encultured without internalizing any cultural content. Culture is not embodied in people's heads but in cultural competences embodied in procedures and cultural performances. The building blocks of cultural content are practice-based and experiential – culture is constructed in real time by persons when they encounter situations that socially require this content (Lizardo, 2012).

Procedural theory is consistent with practice theory in that culture is not represented in the heads of social actors in the form of linguistically coded beliefs, nor is it a logically integrated system of beliefs. Lizardo argues, for example, that objectifying

culture is antithetical to the embodiment of culture in persons for several reasons: first, there cannot be a dialectic of internalization/externalization (Berger and Luckman, 1967) – these concepts must be analytically distinct. Second, if cultural embodiment requires direct experience, how then does one explain the existence of codes when persons lack the capacity to use and understand them? Third, there is empirical evidence that artifacts are discovered after the embodied competencies that produced them have disappeared. Last, institutions survive, yet the capacity to apply rules and categories inherent in the institutions has atrophied (Lizardo, 2012).

In sum, tool kit, practice, and procedural theory argue that to attribute too much internal cultural content to the person is likely to run into explanatory difficulties. The solution is to shift much of the cognitive workings that previous approaches inscribed in the mental makeup of the person to outside the person and toward the external world of institutions in which the person is embedded (Lizardo and Strand, 2010: p. 223).

Considering Both Cultural Sociology and Institutional Logics

The ILP has a history of rigorous research specifying the interaction effects between the institutional environment (culture) and organizations (Haveman and Rao, 1997; Thornton and Ocasio, 1999; Marquis and Lounsbury, 2007), but is only beginning to focus on the microfoundations and the relationship between individuals and culture. Incorporating insights discussed from the culture research can help advance the focus on interaction effects at the microlevel. As a meta-theory the ILP has the potential to address the issues of both the classic and embodied theories. One can understand the embodied theorists' argument that people in preliterate societies did not need to derive meaning through internalization in their heads. However, the empirical research in institutional theory suggests that there have been many institutional worker bees since cave man and it is well-known that institutions "take on a life of their own" (Selznick, 1957). Similarly, the functional life of an institutional logic can wax and wane over historical time (Dunn and Jones, 2010).

This article has spanned levels of analysis with perhaps an uncomfortable broad brush shifting of discussion across organizations and individuals. Nonetheless, the punch line is that with the exception of the work of DiMaggio and Vaisey, the sociology of culture debate appears an overly bifurcated representation of the role of culture and its effects on individual cognition and behavior. My own read of the debate is as DiMaggio (1997, 2002) suggests, the weight of sociological and psychological empirical evidence from controlled studies is on the side of exploring the interaction effects between the symbolic and the material, the internal and the external, and the cognitive and experiential. Below I discuss potential topical areas to begin this work.

Morality

Altruism, sympathy for others, the understanding of others' goals, and other aspects of morality is underdeveloped in the ILP. Current theory in the ILP assumes the influences of morality vary across the seven institutional orders. Just as the institutional orders of family and religion influence moral values and practices, so does the market logic, for example, in Hayek's (1944) notions of individualism and self-responsibility that underlie self-regulating free markets. The current typology of the ILP is a representation of the main institutions of societies (cultural subsystems) and it was derived from a reading of Weber's (1978) Economy and Society as well as classic and contemporary organization and management and social science theory and empirical research. This scope allows theorizing the varied influences of self-interest and understanding the empirical observation that winners of legitimacy struggles and justification contests are not always the most moral or worthy or concerned for the common good. Given that contemporary moral thought is much less coherent than it once was (Massengill, 2008), the ILP is suited to theorize and measure the heterogeneity of cultural context that informs different variants of moral thought.

Excepting Greenwood et al. (2010), which shows that corporate management practices were tempered by religion and family logics in the face of market pressures, there is a lack of logics research invoking religion even though it has been a foundry for individual, organizational, and collective identity and behavior (Tracey, 2012). With the exception of Western Europe, much of the world continues to be religious (P.L. Berger, 2001). Tracey (2012) points to micro- and macro-avenues for exploring the moral scope of institutional logics through his critical review of religion, connecting it to behavioral ethics, identity formation, and change and interdependency in the interinstitutional system. Parboteeah et al. (2009) found a negative association between religion and work ethic norms in countries characterized by 'state religion,' where the government regulates religion by appointing religious leaders and collecting taxes for religious organizations. In the institutional logics parlance this is when the boundaries of the logics of religion and state are weak and interdependent. Pearce et al. (2010) found religion related to entrepreneurial orientation and organizational performance. Weaver and Agle (2002: p. 77) found when religious role expectations are 'internalized as a religious self-identity,' it can be a determinant of individual ethical behavior, depending on the importance of religion to identity and whether individuals are intrinsically motivated to participate in religion. Stout and Cormode's (1998) study is of interest in its focus on how the logic of religion is connected to other institutional orders such as the state, the market, and the family.

Justification and Symbolic Management

Because the ILP assumes that what is considered legitimate changes depending upon the cultural context, it has the potential to examine how culture is more than motivating action – it justifies action. It can be used to understand how legitimacy is comparatively negotiated, evaluated, and resolved. The qualitative work on justification struggles spearheaded by the French pragmatists, Boltanski and Thévenot (1991), and research developed by the US cultural sociologists exemplify the approach called justification (Vaisey, 2009).

Conceptualizing the seven institutional orders (family, community, religion, market, state, professions, and the corporation) to represent different repertoires of justification that individuals and organizations use to make sense of and rationalize their choices – is a short step from its current use as a framework for understanding the sources of legitimacy as motivation. For example, one could compare differences in the meaning of case vignettes by switching institutional orders, which suggests different justifications for the same behavior.

Institutional Logics as Culture

In "Markets from Culture …" Thornton (2004) treats culture as an independent variable, to interpret organizational decision-making changes based on how different institutional logics focus actors attention. Institutional logics as culture can be applied to a myriad of unexplored topics such as how objects, spaces, and technologies shape and frame social relations. Hollerer et al. (2013) show the importance of visual material elements of communication in translation. Moreover, culture can be viewed as a dependent variable. Research in the ILP illustrates how alternative meanings of cultural symbols and material practices, as influenced by different institutional orders, wane and wax over time and context.

The major contemporary theoretical schools in cultural, cognitive, and organization theory differ in how they focus on and characterize cultural content – whether it stems from the internalization of tightly coupled values or the externalization of loosely coupled justifications based on categories and codes (Vaisey, 2009; Hannan et al., 2007). The ILP is a scaffolding for fleshing out different parts of a larger story of cultural content – a way to keep the big picture in mind.

In sum, while the ILP is less developed at the microlevel than the literature in both classic and embodied cultural sociology, the two schools of thought share affinities with the ILP in a number of respects of which I highlight three in particular. First, institutional logics are comprised of both the symbolic and the material aspects of actors' culture and experience. Second, as a meta-theory the ILP is the subject of some other theory, those theories could be, for example, socialization, tool kit, practice, dual process, and or procedural theories. Third, the ILP is a method to help specify the internal and external scaffolding of culture and its associated micro- and macrolevel mechanisms. It does not privilege any one cultural context or actor theory. Given these assumptions, the time is ripe for exploiting the affinities between institutional logics and cultural theory.

See also: Culture, Cognition and Embodiment; Embodiment and Culture; Epistemic Cultures; Formal Methods of Cultural Analysis; Organizations and Culture.

Bibliography

Alexander, 2003. The Meanings of Social Life: A Cultural Sociology. Oxford University Press, New York.

Almandoz, J., 2012. Arriving at the starting line: the impact of community and financial logics on new banking ventures. Academy of Management Journal 55 (6), 1381–1406.

Barsalou, L.W., 1987. The instability of graded structure: implications for the nature of concepts. In: Neisser, U. (Ed.), Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization. Cambridge University Press, New York, pp. 101–140.

Barsalou, L.W., 1999. Perpetual symbol systems. Behavioral and Brain Sciences 22, 577–660.

Berger, P.L., 2001. Reflections on the sociology of religion today. Sociology of Religion 62, 425–429.

Berger, P., Luckman, T., 1967. The Social Construction of Reality. Doubleday Anchor, New York.

Boltanski, L., Thévenot, L., 1991. On Justification: The Economies of Worth. Princeton University Press, Princeton, NJ.

Bourdieu, P., 1990. The Logic of Practice. Polity Press, Cambridge.

Bourdieu, P., 2000. Pascalian Meditations. Stanford University Press, Stanford, CA.

Cowan, N., 2001. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behavioral and Brain Sciences 24, 87–114.

D'Andrade, R., 1992. Schemas and motivation. In: Human Motives and Cultural Models. Cambridge University Press, New York, pp. 23–44.

DiMaggio, P.J., 1990. Cultural aspects of economic action and organization. In: Friedland, Roger, Robertson, A.F. (Eds.), Beyond the Marketplace. Aldine de Gruyter, New York.

DiMaggio, P.J., 1997. Culture and cognition. Annual Review of Sociology 23, 263–287.

DiMaggio, P.J., 2002. Why cognitive (and cultural) sociology need cognitive psychology. In: Cerulo, Karen (Ed.), Culture in Mind: Toward a Sociology of Culture and Cognition. Routledge, New York.

DiMaggio, P.J., Powell, W.W., 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. American Sociological Review 48, 147–160.

DiMaggio, P.J., Nag, M., Blei, D., 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of U.S. government arts funding. Poetics 41, 570–606.

Douglas, M., 1986. How Institutions Think. Syracuse University Press, Syracuse, New York.

Dunn, M.B., Jones, C., 2010. Institutional logics and institutional pluralism: the contestation of care and science logics in medical education, 1967–2005. Administrative Science Quarterly 55, 114–149.

Durand, R., Szostak, B., Jourdan, J., Thornton, P.H., 2013. Institutional logics as strategic resources. In: Lounsbury, M., Boxenbaum, E. (Eds.), Institutional Logics in Action, Research in the Sociology of Organizations. Emerald, Bingley.

Durkheim, E., 2005. The dualism of human nature and its social conditions. Durkheimian Studies 11 (1), 35–45.

Friedland, R., Alford, R., 1991. Bringing society back in: symbols, practices, and institutional contradictions. In: Powell, W.W., DiMaggio, P.J. (Eds.), The New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago, IL, pp. 232–263.

Gallese, V., Lakoff, G., 2005. The brains concepts: the role of the sensory-motor system in conceptual knowledge. Cognitive Neuropsychology 22 (3), 455–479.

Garfinkel, H., 1967. Studies in Ethnomethodology. Prentice-Hall, Englewood Cliffs, NJ.

Glaser, V.L., Fast, N.J., Green, S., August 2012. Institutional cues, vocabularies of motive, and automatic behavior: an investigation into the psychological foundations of institutional logics. In: Paper Presented at the Academy of Management. Boston, MA.

Greenwood, R., Díaz, A.M., Li, S.X., Lorente, J.C., 2010. The multiplicity of institutional logics and the heterogeneity of organizational responses. Organization Science 21, 521–539.

Hannan, M., Polos, L., Carroll, G., 2007. Logics of Organization Theory: Audiences, Codes, and Ecologies. Princeton University Press, Princeton, NJ.

Haveman, H.H., Rao, H., 1997. Structuring a theory of moral sentiments: institutional and organization co-evolution in the early thrift industry. American Journal of Sociology 102, 1606–1651.

Hayek, F., 1944. The Road to Serfdom. University of Chicago Press, Chicago.

Hollerer, M.A., Jancsary, D., Meyer, R.E., Vettori, O., 2013. Ways of seeing: the visual re-contextualization of corporate social responsibility. In: Lounsbury, M., Boxenbaum, E. (Eds.), Institutional Logics in Action, Research in the Sociology of Organizations. Emerald, Bingley, UK.

Kahneman, D., 2003. Mapping bounded rationality. American Psychologist 58, 697–720.

Klyver, K., Thornton, P.H., 2014. Who Is More Likely to Symbolically Manage Their Entrepreneurial Intentions? Loose Coupling Effects by Gender and Work Status. Working Paper.

Lakoff, G., Johnson, M., 1999. Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. Basic Books, New York.

Lizardo, O. Culture, Cognition, and Embodiment. International Encyclopedia of the Social and Behavioral Sciences, in this volume.

Lizardo, O., 2012. Embodied Culture as Procedure: Rethinking the Link between Personal and Objective Culture. HELDA The Digital Repository of University of Helsinki.

Lizardo, O., Strand, M., 2010. Skills, toolkits, contexts and institutions: clarifying the relationship between different approaches to cognition in cultural sociology. Poetics 38, 204–227.

Lounsbury, M., Glynn, M.A., 2001. Cultural entrepreneurship: stories, legitimacy, and acquisition of resources. Strategic Management Journal 22, 545–564.

Marquis, C., Lounsbury, M., 2007. Vive la resistance: competing logics and the consolidation of U.S. community banking. Academy of Management Journal 50, 799–820.

Martin, J.L., 2010. Life's a beach but you're an ant, and other unwelcome news for the sociology of culture. Poetics 38, 228–243.

Massengill, R.P., 2008. Prayers of the people: moral metaphors in the right-to-life and faith-based labor movements. Poetics 36, 338–357.

Meyer, J.W., Rowan, B., 1977. Institutionalized organizations: formal structure as myth and ceremony. American Journal of Sociology 83, 340–363.

Metcalfe, J., Michel, W., 1999. A hot/cool system analysis of delay of gratification: dynamics of willpower. Psychological Review, 3–19.

Mischel, W., Shoda, Y., 1995. A cognitive – affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. Psychological Review 102, 246–268.

Nigam, A., Ocasio, W., 2010. Event attention, environmental sensemaking and change in institutional logics: an inductive analysis of the effects of public attention to Clinton's health care reform initiative. Organization Science 21, 823–841.

Ocasio, W., 1997. Towards an attention based view of the firm. Strategic Management Journal 18, 187–206.

Ocasio, W., 2011. Attention to attention. Organization Science 22 (5), 1286–1296.

Parboteeah, K.P., Hoegl, M., Cullen, J., 2009. Religious dimensions and work obligation: a country institutional profile model. Human Relations 62, 119–148.

Parsons, T., 1964. Social Structure and Personality. Free Press, New York.

Parsons, T., 1968. Cooley and the problem of internalization. In: Reiss, A.J. (Ed.), Cooley and Sociological Analysis. University of Michigan Press, Ann Arbor, MI, pp. 48–67.

Parsons, T., 1972. Culture and social systems revisited. Social Science Quarterly 53 (2), 253–266.

Pearce II, J.A., Fritz, D.A., Davis, P.S., 2010. Entrepreneurial orientation and the performance of religious congregations as predicted by rational choice theory. Entrepreneurship Theory and Practice 34, 219–248.

Selznick, P., 1957. Leadership in Administration. University of California Press, Berkeley.

Stinchcombe, A.L., 2002. New sociological microfoundations for organizational theory: a postscript. In: Lounsbury, M., Ventresca, M. (Eds.), Social Structures and Organizations Revisited, vol. 19. Elsevier Science, Amsterdam, The Netherlands, pp. 415–433.

Stout, H.S., Cormode, D.S., 1998. Institutions and the story of American religion: a sketch of a synthesis. In: Demerath III, N.J., Hall, P.D., Schmitt, T., Williams, R.H. (Eds.), Sacred Companies: Organizational Aspects of Religion and Religious Aspects of Organizations. Oxford University Press, New York, pp. 62–78.

Strauss, C., Quinn, A., 1997. A Cognitive Theory of Cultural Meaning. Cambridge University Press, Cambridge.

Swidler, A., 1986. Culture in action: symbols and strategies. American Sociological Review 51, 273–286.

Swidler, A., 2001. What anchors cultural practices. In: Schatzki, T.R., Cetina, K.K., Von Savigny, E. (Eds.), The Practice Turn in Contemporary Theory. Routledge, New York, pp. 74–92.

Swidler, A., 2003. Talk of Love: How Culture Matters. University of Chicago Press, Chicago, IL.

Thornton, P.H., 2001. Personal versus market logics of control: a historically contingent theory of the risk of acquisition. Organization Science 12, 294–311.

Thornton, P.H., 2002. The rise of the corporation in a craft industry: conflict and conformity in institutional logics. Academy of Management Journal 45, 81–101.

Thornton, P.H., 2004. Markets from Culture: Institutional Logics and Organizational Decisions: In Higher Education Publishing. Stanford University Press, Stanford, CA.

Thornton, P.H., Ocasio, W., 1999. Institutional logics and the historical contingency of power in organizations: executive succession in the higher education publishing industry, 1958–1990. American Journal of Sociology 105, 801–843.

Thornton, P.H., Ocasio, W., 2008. Institutional logics. In: Greenwood, R., Oliver, C., Sahlin-Andersson, K., Suddaby, R. (Eds.), Handbook of Organizational Institutionalism. Sage, Thousand Oaks, CA, pp. 99–129.

Thornton, P.H., Jones, C., Kury, K., 2005. Institutional logics and institutional change: transformation in accounting, architecture and publishing. In: Jones, Candace, Thornton, Patricia H. (Eds.), Research in the Sociology of Organizations. JAI Press, Greenwich, CT, pp. 125–170.

Thornton, P.H., Ocasio, W., Lounsbury, M., 2012. The Institutional Logics Perspective: A New Approach to Culture, Structure, and Process. Oxford University Press, Oxford, UK.

Tracey, P., 2012. Religion and organization: a critical review of current trends and future directions. Academy of Management Annals 6, 87–134.

Vaisey, S., 2009. Motivation and justification: toward a dual process theory of culture in action. American Journal of Sociology 114, 1675–1715.

Weaver, G.R., Agle, B.R., 2002. Religiosity and ethical behavior in organizations: a symbolic interactionist perspective. Academy of Management Review 27, 77–97.

Weber, M., 1978. Economy and society: an outline of interpretive sociology. In: Roth, G., Wittich, C. (Eds.). University of California Press, Berkeley.

Zaller, J., 1992. The Nature and Origin of Mass Opinion. Cambridge University Press, New York.

Zerubavel, E., 1999. Social Mindscapes: An Invitation to Cognitive Sociology. Harvard University Press.

Zucker, L.G., 1977. The role of institutionalization in cultural persistence. American Sociological Review 42, 726–743.

Culture and Networks

Ronald L Breiger and Kyle Puetz, University of Arizona, Tucson, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Changing conceptions within the sociology of culture and in the research community of social network analysts have led to the development of a new specialty area, research at the intersection of culture and networks. The new research entails a rethinking both of the kinds of actors and actions that takes place in networks, and of the connections among actors that are relevant. This article emphasizes conversations, objections, and ongoing concerns within this evolving research specialty as well as research accomplishments.

The concept of *culture* has traditionally connoted large-scale phenomena such as national character, the ultimate values of a society, shared stores of knowledge, and people's total way of life. The concept of *social networks* has traditionally referred to the patterning of connections among actors, with analytical power deriving from an emphasis on the formal modeling of system and structure. Over the past 30 years, but especially during the past decade, each of these concepts has undergone significant reorientation. Moreover, the topic of 'culture and networks' has emerged as a research specialty in its own right and became the subject of a number of overviews, reviews, and orienting statements (DiMaggio, 2011; Fuhse and Mützel, 2011; Mische, 2011; Mohr and Rawlings, 2010; Pachucki and Breiger, 2010) in the wake of widely influential calls a decade earlier for the production of new models and syntheses of social structural and cultural analysis (DiMaggio, 1992; Emirbayer and Goodwin, 1994). The focus of this article is to describe the main contours of these changes as well as the most noteworthy of the emerging research questions and approaches guiding the contemporary work.

Culture and Networks as a Developing Research Specialty

We review recent changes in emphasis in the study of both culture and social networks that has led to the new research specialty at their intersection.

Culture

A diverse set of influential anthropologists and sociologists writing in the 1970s pointed the way toward a more activist view of culture than the one that had long dominated these disciplines and which posited an autonomous sphere of clearly known values that were imprinted inside the heads of all members of a society. As against theories of Saussure, Durkheim, Kluckhohn, and Parsons, the new analysts – grouped under the rubric of 'practice theorists' by Ortner (1994) – included Clifford Geertz, Marshall Sahlins, Pierre Bourdieu, and Anthony Giddens. A major contribution of Geertz was his insistence on studying culture from the actor's point of view; however, his most radical theoretical move, as Ortner (1994: p. 129) has emphasized, was to argue that culture is not something locked inside people's heads, but rather is embodied in public symbols. For Geertz, "man is an animal suspended in webs of significance he himself has spun," and Geertz took culture to be those webs (Geertz, 1973: p. 5). Moreover, as argued most sharply by Giddens, the study of practice is not antagonistic to the study of systems or structures, but rather a necessary complement to it (Ortner, 1994: pp. 146–147).

Beginning with Geertz's definition of culture, but seeking to change the emphasis from interpretation to explanation, Ann Swidler (1986) theorized a model alternative to the traditional conception, according to which, rather than values providing the ultimate ends that actors seek, actors' cultivated skills, styles, and habits, bundled together as repertoires or 'tool kits,' provide robust linkages between culture and action. Symbols, stories, rituals, and worldviews constitute repertoires that people use actively to construct diverse strategies of action. The term 'strategy' connoted for Swidler, not a consciously defined plan to attain a goal, but rather a general way of organizing action. Thus, lines of action might for some purposes take shape as trading in a market, and in other instances might refer to gaining resources from existing networks of relatives and friends. This insight provides an opening to subsequent findings on the conditions for within-network exchanges (DiMaggio and Louch, 1998). Many currents of contemporary research in cultural sociology have been strongly influenced by this concept of culture operating at a more local level than in the traditional formulation and emphasizing culture as providing repertoires of capacities and practices.

Networks

Many commentators (including Pachucki and Breiger, 2010) draw attention to the extent to which the 1970s breakthroughs in network modeling seemed explicitly to reject culture as any part of an explanation for network patterning. This is too simplistic a view on two accounts.

Important currents of network analysis have defined themselves, much as is apparent in Swidler's work, as a move from unwarranted abstraction and toward the analysis of contextualized and concrete structures. Harrison White and his collaborators on blockmodeling indeed did explicitly reject rooting network analysis in 'a harmony among abstract norms and values' (a reference to the sociology of Talcott Parsons, which was also the main point of departure for Swidler's

cultural toolkits approach), and in doing so they endeavored to redefine the classic concepts of role and position 'so that they apply to concrete, observable interactions' (White et al., 1976: pp. 731, 733). While there are today, and have always been, analytical movements within network analysis to study abstract 'structure,' in the singular (well epitomized by the search for laws that apply to 'all networks,' as reviewed popularly in Barabási, 2002), White and collaborators set out to ground network analysis in the study of 'structures' in the plural – different networks that might be associated with different strategies or lines of action. Indeed, even a single network, a small college fraternity, was shown to evolve within its first 15 weeks from one pattern (that of in-groups) toward a distinctively different one emphasizing hierarchy and increasingly negative ties aimed at an out-group (White et al., 1976: p. 764).

The second way in which blockmodeling (search for macrostructure) across multiple networks (sets of connections among social actors, with a different network defined for each of several distinct types of relationship, such as affect and economic exchanges) admits an opening to culture is in putting forward an explicit approach to the meaning of ties, albeit an approach that was firmly grounded in structuralism. For Boorman and White (1976), two networks have the same meaning if and only if they have identical pattern of incidence, either at the individual actor level or (more usually) at the macrolevel. Thus, for example, if the network reporting enemies of enemies of each actor (such that there is a tie from each person i to person j if j is an enemy of an enemy of i) produces a pattern of ties among actors that is identical to the network of friendship among the same actors (a tie from i to j in the friendship network indicating that j is a friend of i), then by Boorman and White's (1976) *axiom of quality*, the analyst has a warrant to conclude that, in the context studied, enemies of enemies are friends, and precisely in this way the two networks are said to share the same meaning.

Culture and Networks

The emphases on concrete structure and on structure-as-meaning thus provided some openings toward culture, but the 1970s approaches to social network analysis were justly criticized by Brint (1992) for marginalizing culture, yet smuggling it back in to provide an interpretation of blocks of actors who, for example, might be characterized as the "top dogs … [who] remain largely ignorant of most lesser mortals" (White et al., 1976: p. 749), a phrase expressive of cultural distinctions that the network analysts did not however interpret on a cultural plane. Symbolic interactionists (most recently and astutely, Salvini, 2010) urge a reconceptualization of social networks as emergent effects of the complex interlinking of situated social interactions, never entirely stable, and informing an analysis of how actors work through constraints to formulate courses of action for social interactions.

A significant spur to the creation of joint interest in networks and culture arose in the 1990s with the formation of a working group on 'measurement and meaning' within the 5-year-old Section on the Sociology of Culture of the American Sociological Association. Paul DiMaggio, John Mohr, and Ann Swidler were among the active participants. Meanwhile, Harrison White published *Identity and Control* in 1992 (a largely new second edition appeared in 2008), upending existing network theory and proclaiming that a social network is a network of meanings (see Mische, 2011; Pachucki and Breiger, 2010: pp. 208–209, for expanded discussion). The journal *Poetics*, an interdisciplinary journal of theoretical and empirical research on culture, the media and the arts, began from the 1990s onward to be a vibrant center for research at the intersection of social networks and culture. Within the past decade, the annual meeting of social network analysts has begun to feature sessions specifically on networks and culture, and models of social influence processes have been brought to bear on studying the reception of texts (Childress and Friedkin, 2012).

Significant Research Questions and Controversies

We will address the principal research issues at the interface of culture and networks under two main rubrics, relating respectively to the kinds of actors, and the kinds of connections or ties, that constitute social networks from the perspective of cultural analysis.

What Kinds of Actors Inhabit Social Networks?

Rational vs Practical Network Actors

Much traditional social network analysis models the actors or 'nodes' of network diagrams as black boxes and lacking in purpose, agency, or motivation. DiMaggio (1992: p. 120) referred to it as the effort to create a structural sociology that is culture-free, and urged the formulation of a suitable theory of action and of the cognitive status of role definitions, action scripts, and typifications. Moreover, DiMaggio portrayed the early blockmodeling work as well as the continuing tradition in network analysis as smuggling in such frameworks implicitly. In contrast to both the Parsonsian and the rational choice traditions, DiMaggio (1992: p. 121) saw these forms of network analysis as sharing with ethnomethodology a radical realism focused on concrete social relations, and a view of culture as "a mystifying system of post hoc accounts used by actors to normalize or explain interaction rather than to shape it." Sketching a contrast between rational action and practical action, the latter consisting of a mix of cognitive psychology, social constructionism, the Carnegie School of organizational behavior research, and a conception of practical reason akin to Bourdieu's, DiMaggio (1992: p. 138) urged network theorists to develop the practical actor model, with its emphasis on what actors do not know about social structure and the shortcuts they develop to operate under conditions of uncertainty. Somers (1998: p. 766–767) put forward a 'relational realist and pragmatist ontology' taking as the basic units of social analysis neither individuals nor structural wholes, but rather the relational processes of interaction between and among identities.

Inner-Directed and Tradition-Directed vs Other-Directed Network Actors

Recently, there has arisen something of a counterrevolution against what its proponents see as excessive social constructivism on the part of analysts of the culture–network interface.

With reference to David Riesman's (1950) distinction among action orientations, it is possible that (explicit or, more usually, implicit) models of network actors have tended to portray them as 'other-directed,' implying a muted form of individuality according to which actors who inhabit social networks are incessantly oriented only to the actions of others, rather than holding a firm course fixed by a person's internal gyroscope or moral compass ('inner-directed' actors) or to an orientation emphasizing the value of community norms and a sense of taken for granted purpose ('tradition-directed'). C. Smith (2010: p. 340, italics in original) brings to bear a philosophical standpoint positing that a "*natural drive toward a sustained and thriving personal life broadly … generates*" social structure. Writing on 'network structuralism's missing persons,' he insists that norms cannot be dismissed as causally insignificant. While sympathetic to the network structuralists' rejection of the variables paradigm and Parsonsian theory, Smith portrays the pendulum as having swung too far, resulting in a network theory that is antihumanist and person-annihilating. It is necessary to reject the package of presuppositions that brings together reductionism, constructivism, pragmatism, empiricism, and positivism in order to redress network structuralism's neglect of human dignity, rights, respect, and rational deliberation (C. Smith, 2010: pp. 270, 272).

Vaisey (2009: p. 1687), grounding his approach in the dual-process model of cognitive science, puts forward a conception according to which actors are driven primarily by deeply internalized schemata (akin, he believes, to Bourdieu's habitus or to implicit consciousness), yet are also capable of deliberation and justification (discursive consciousness) when the situational demands of social interaction require it. Vaisey argues that, with their championing of culture as repertoire (which Vaisey identifies with his deliberative pole), analysts of the culture–network nexus have been focusing on only half of the cognitive process, and the less interesting half at that.

Vaisey and Lizardo (2010) applied the dual-process model to survey data relating 'worldviews' (such as whether respondents say they believe in moral relativism) to respondents' network composition at a subsequent time (e.g., whether respondents claim friends who engaged in deviant acts or who, on the other hand, volunteered in community service). Results showed that worldview had a significant and relatively large effect on subsequent network composition, controlling for aspects of network structure and composition. There are questions about whether durable moral worldviews were successfully measured (in that over 50% of the respondents were categorized as having different worldviews in the survey's second wave) and about whether survey data can provide an appropriate measure of implicit cognition (Srivastava and Banaji, 2011: p. 208), but the larger point for present purposes is the desirability of pioneering efforts such as this one to assess the effects of both implicit and discursive cognition on actors' formation of social network ties. Further in this regard, the dual-process model has been brought to the nodes and arcs of full who-to-whom social network data in the study of Srivastava and Banaji (2011). The test for implicit association that they employ is based on the measurement of response times as network actors sort descriptions that might characterize them, thus capturing the phenomenology of implicit cognition. The authors build these cognitive measures into a sophisticated exponential random graph model of network structure.

Network Actors as People Who Remember

T. Smith (2007) models networks in a highly innovative way. The context is extreme ethnic conflict. The location is Istria, the land at the northern tip of the Adriatic where Italy, Slovenia, and Croatia meet. The data concern the stories that people tell about their life histories, as these intertwine with wars and long and internecine conflicts among ethnic Italian and Croatian inhabitants of Istria. Nodes in the network are not persons or groups but narrative elements such as 'superior culture' and 'ethnic cleansing' as told to the researcher in the stories of several dozen people on each side of the conflict in Istria and (subsequently) in New York City, where Smith interviewed members of the immigrant Italian and Croatian immigrant communities. Of special importance to Smith are boundary elements that can act as bridges to otherwise unconnected narratives (those related by Italians and those told by Croatians) and transform the meaning of those story components that overlap the two narratives. At the same time, their ties to the original narratives open up opportunities for importing new meanings into existing narrative structures, which is one meaning of bridging a cultural hole (Pachucki and Breiger, 2010: p. 216).

Network actors have memories, and T. Smith's (2007) study of immigrant Istrians in New York City shows how they can deploy discursive repertoires that are consistent with ethnic reconciliation and enhanced humanity and dignity. When Istrian immigrants of Croation and Italian ancestry came to interact in New York churches, ethnic clubs, and restaurants, the narrative boundaries softened, to the point that all Istrians were now seen as common victims of external forces. Without wanting to be seen as questioning the historical fact of the fierce ethnic conflict, the researcher nonetheless emphasizes the social construction of memories of violence, and in particular their muting, as a potential step toward reconciliation and shared dignity. To be sure, silence about conflict comes at a cost as the suffering of victims is forgotten and perpetrators go unpunished; on the other hand, enshrining the bases of divisions in constitutions or land partitions would deny actors opportunities for conarration (T. Smith, 2007: pp. 43, 44). Mapping individual and collective memory (Cunningham et al., 2010) seems a promising direction for research on cultural strategies of network cognition. The further development of this line of empirical study might begin to address the concerns of C. Smith (2010), reviewed earlier, on the asserted antihumanism of social network analysis.

What Types of Network Tie Comprise Culture?

The connections among actors that are most typically studied by social network analysts are either social relationships or ties of persons to groups. If the locus of culture is 'webs of significance' (Geertz, 1973), then the question is raised as to what types of network tie adequately capture signification. We review answers to this question that have been influential for work on culture and networks.

Beyond Persons and Groups

Building on insights of Simmel, Cooley, Nadel, and Goffman, Breiger proposed in the 1970s a locus for network analysis that differed from the usual emphasis on ties among a set of social actors. Breiger suggested a 'duality' according to which groups are connected to one another on the basis of the members they share in common while, at the same time, the individuals who are members of multiple groups form a dual network. In the one network, persons are the nodes and their joint memberships in groups are the connections among the persons; in the dual network, groups are the nodes and shared persons are the ties among groups. The structure of the intergroup network and the structure of the interpersonal network are distinct but, also, mutually constitutive (reviewed in Mohr and Duquenne, 1997).

Mohr and Duquenne (1997) introduced a crucial move from networks to culture by recognizing that the formalism of this model of dual affiliation well captured a key aspect of the emphasis of Geertz and other practice theorists on the logic of culture as coconstituting the patterning of everyday activities. The ideas of a culture (e.g., 'private property') are constituted by the practices used and recognized to sustain them, even as these practices are shaped by implementations of the idea. In their application, Mohr and Duquenne studied poverty relief in New York City around the beginning of the twentieth century, and in particular the ways in which categories of the poor (such as 'fallen' and 'stranger') were associated with procedural remedies (such as help in providing employment, or putting a person in an almshouse or other long-term custodial care facility) applied to classes of persons, as this information was gleaned from directories of charity organizations. In this way, the researchers were able to deduce a web of meaning among the categories of poor person on the basis of how those categories were used.

The method used by Mohr and Duquenne was dual (Galois) lattice analysis, a technique that represents two orders of information by graphically portraying the composition of elements at each level by those at the other. Mische (2008) applies a form of three-mode lattice analysis (jointly relating persons, organizations, and events) as one among a number of innovative procedures for studying 'civic mediation,' according to which both connections and identities are constructed across social movement organizations.

In addition to algebraic techniques such as lattice analysis, statistical methods have also been used to study duality, notably the method of correspondence analysis. In a departure from usual modes of network analysis, this procedure is capable of application – as in Bourdieu's research – to categories of person (such as occupational categories) and types of cultural object (such as the distribution of preferences of each occupational category for a first communion, a sunset, or a car crash as the subject of a beautiful photograph). Rather than as a lattice, the results of correspondence analysis are often presented in a Cartesian plane, where social positions and cultural dispositions can be captured spatially and with respect to the oppositions as well as affinities that define a social field. Rouanet et al. (2000) provide a more adequate overview of Bourdieu's orientation to correspondence analysis. Further on the subject of methods for the study of duality, Charles Ragin has introduced qualitative comparative analysis (QCA) as a way of understanding cases as constituted by configurations of variables. Breiger (2009), in an application to the mobilization of ethnic subnational groups in Western Europe and the social and cultural properties of those groups, demonstrates how a form of correspondence analysis can be used to analyze the duality of cases (ethnic groups) and variables (social and cultural group properties and behaviors) that is one of the motivating principles of QCA.

Recent work has employed innovative definitions of network affiliation ties to illuminate cultural webs. Using correspondence analysis of survey data, Hedegard (2013) studies relations among racial self-categorizations in Brazil (seven categories spanning black, brown, white, indigenous, and their combinations) on the basis of respondents' tastes for 13 cultural objects (most of which are musical forms). One of the main findings is that, despite incorporation of forms such as samba and *capoeira* into the national identity, and their practice across racial and class lines, these cultural objects continue to associate with divergent racial identities.

A variety of network-analytic concepts and methods may be usefully applied to the domain of culture. In a study of four Washington, DC marches in support of lesbian, gay, bisexual, and transgender agendas and causes in the period 1979–2000, Ghaziani and Baldassarri (2011) take as their basic data structure a matrix indicating the presence or absence of each of 12 themes in each of hundreds of newspaper articles. The authors study this affiliation matrix directly, and they also study the dual theme-by-theme network that it implies. The affiliation matrix is normalized to take account of differing main effects in the popularity of themes. We believe that three kinds of analyses reported in this article will be of interest in subsequent research on culture and networks: (1) the authors apply usual concepts from network analysis (such as centrality and brokerage) to the networks of themes, identifying a stable core of themes as well as several time-varying sets of concerns that are congenial to the authors' argument for the existence of thematic 'anchors' that allow social movements to support differences within the frame of a common core; (2) representing each of the four marches as a network among the 12 thematic elements, the authors compare all pairs of networks by means of a network analysis tool (Quadratic Assignment Procedure correlation), discovering statistically significant coherence over time; and (3) they represent each of the four normalized matrices of thematic cooccurrence as a network graph, thus identifying stable and varying sets of thematic elements over time.

Another method that we believe will prove generally applicable in the study of affiliations of actors with cultural forms is relational class analysis (Goldberg, 2011), which allows the analyst to parse out groups of like-minded individuals, not on the basis of individuals within the groups sharing similar attitudes, but rather on the basis of having similar beliefs about the relations among attitudes. (An example given by the author is that two people within the same relational class might disagree vehemently about abortion, but agree that the debate revolves around women's right to control their bodies; people in a separate class might identify abortion with a complex of issues pertaining to social welfare.) Goldberg's data are in the form of a preference matrix, respondents by musical genres, and he identifies three distinctive logics

combining musical tastes and social class on the basis of his relational class analysis.

At a very different level (but one that turns out to be remarkably relevant), computational scientists and complexity theorists (Ahn et al., 2011) have built on findings of food chemists about the flavor compounds (i.e., chemical compounds) contained in many culinary ingredients. The researchers begin with a matrix of 381 ingredients used in recipes throughout the world, and 1021 flavor compounds, indicating which compound is in which ingredient. A dual projection results in a 'flavor network' in which two ingredients are connected if they share at least one compound, and the weight of each link is the number of compounds shared. Further analysis (not reported here) leads to a visualization of the flavor network. It is found that, on the basis of shared chemical compounds, ingredients used in Western European and North American cuisines are statistically similar. In North American recipes, for example, the more compounds two ingredients share, the more likely they are to appear together. However, the authors found sharply different results for Southern European and East Asian cuisines: in those cases, the more compounds two ingredients share, the less likely they are to appear together in the same recipe. The researchers could identify broad cultural regions of the world on the basis of a network of combinations of chemical compounds. This work reaffirms that only a researcher's imagination limits the kinds of data that can be brought to bear in the study of dual networks of culture and practices. We also note that the study of Ahn et al. connects to actor-network theory and its study of the interpenetration of social and nonhuman networks.

Networks of Narrative Elements

The work reviewed above focuses on affiliations between two orders of phenomena (such as persons and musical tastes). Researchers at the interface of culture and networks have also been providing innovative formulations of one-mode networks (i.e., ties among a single kind of entity, such as persons). We have already reviewed T. Smith's (2007) study of memory and ethnic conflict, in which the nodes are elements of narratives and the ties reflect sequential order in the narrative. DiMaggio (2011: pp. 295–296) reviews a wider range of network studies of narrative events, as well as computational science studies of texts. See also Varga (2011) on measuring the relative degree of fragmentation of scientific fields on the basis of cocitation analysis.

The Role of Negative Ties

Negative network ties may play a distinctively important role in culture, and may even enhance cultural creativity, as Giuffre (2009) found in a study of artists on the Polynesian Island of Rarotonga. In the past decade, the island has earned a reputation across the South Pacific as a hub for the revival of traditional Polynesian arts and as a center of artistic creativity and ferment. The island has more fine arts galleries per capita than London, New York, and smaller arts centers like Santa Fe, New Mexico. Complementing an ethnographic study, the author conducted a network analysis of artists who were asked to name the three best artists on the island, excluding themselves. A number of 'worst artist' references were also volunteered, though they had not been solicited. Among the four positions that emerged from the network analysis were the 'Stars,' consisting of artists who had been born in New Zealand of Cook Islands ancestry and who had recently arrived on Rarotonga with a reverence for its traditional culture along with a Westernized sense of professionalism and an interest in the international fine arts world. By way of contrast, the 'Old Guard' consisted mainly of older artists who had grown up in the Cook Islands and who had pursued their interests in spite of the traditional opprobrium directed against artistic careers; they were more interested in pleasing their local customers than in international standing, and they often painted local flora and landscapes. Although occupying a coherent position in the network relationally, members of the Old Guard directed virtually no esteem toward one another and, moreover, the negative ties were concentrated between the Old Guard and the Stars. A third set of artists endeavored (not too successfully) to mediate between the first two sets, and a fourth were acolytes of the Stars. It is as if the Old Guard and the Stars both served as action ideals (Swidler, 1986) that were taken to be negative by members of the other set. The negativity was heightened by the small size and confined geography of an island setting. Giuffre builds upon work by Labianca and Brass (2006) on negative social network ties in American organizations and the stultifying features of workflow demands and internal hierarchies that might make it difficult or impossible for workers to avoid those who dislike them – much as on an island. Labianca and Brass write that negative ties may be more important than positive ones in their impact on organizations.

Giuffre (2009) suggests that a mix of positivity and negativity may be necessary in social networks that foster creativity, in that the negative ties trigger critical evaluations, discourage groupthink, and (by presenting action ideals that are negative) provide a spur to expanding one's own artistic identity. This study of negative ties and creativity links in a useful way to research on symbolic boundaries. Doreian and Mrvar (2009) provide practicable methods for partitioning networks containing a mixture of positive and negative connections.

Conclusion

In this article, we have reviewed changes in the sociology of culture and in the research community of social network analysts that have led to the development of a new specialty area, research at the intersection of culture and networks. The new research has led to a rethinking both of the kinds of actors and actions that take place in networks, and of the connections among actors that are relevant. We have emphasized conversations, objections, and ongoing concerns within this evolving research community as well as research accomplishments.

DiMaggio (2011) argues that network analysis is the natural methodological framework for developing insights from leading theoretical approaches to cultural analysis. While making use of and offering some further support for that argument, we have tried also in this article to emphasize some of the benefits to network analysis of the opening to culture that has been made possible by a diverse set of researchers who share a commitment to working at the interface of culture and networks.

See also: Cultural Production in Networks; Culture and Actor Network Theory; Culture and Institutional Logics; Culture, Cognition and Embodiment; Formal Methods of Cultural Analysis; Narrative Networks; Networks and Cultural Consumption; Networks and Meaning; Symbolic Boundaries.

Bibliography

Ahn, Y., Ahnert, S.E., Bagrow, J.P., Barabási, A., 2011. Flavor network and the principles of food pairing. Scientific Reports 1 (196). http://dx.doi.org/10.1038/srep00196 (accessed 12.12.04.).

Barabási, A., 2002. Linked: The New Science of Networks. Perseus Pub, Cambridge, MA.

Boorman, S.A., White, H.C., 1976. Social structure from multiple networks. II. Role structures. American Journal of Sociology 81, 1384–1446.

Breiger, R.L., 2009. On the duality of cases and variables: correspondence analysis (CA) and qualitative comparative analysis (QCA). In: Byrne, D., Ragin, C.C. (Eds.), The SAGE Handbook of Case-based Methods. Sage Publications Ltd, London, pp. 243–259.

Brint, S., 1992. Hidden meanings: cultural content and context in Harrison White's structural sociology. Sociological Theory 10, 194–208.

Childress, C.C., Friedkin, N.E., 2012. Cultural reception and production: the social construction of meaning in book clubs. American Sociological Review 77, 45–68.

Cunningham, D., Nugent, C., Slodden, C., 2010. The durability of collective memory: reconciling the "Greensboro Massacre". Social Forces 88, 1517–1542.

DiMaggio, P., 1992. Nadel's paradox revisited: relational and cultural aspects of organizational structure. In: Nohria, N., Eccles, R.G. (Eds.), Networks and Organizations: Structure, Form, and Action. Harvard Business School Press, Boston, pp. 118–142.

DiMaggio, P., 2011. Cultural networks. In: Scott, J., Carrington, P.J. (Eds.), The SAGE Handbook of Social Network Analysis. Sage, London, pp. 286–300.

DiMaggio, P., Louch, H., 1998. Socially embedded consumer transactions: for what kinds of purchases do people most often use networks? American Sociological Review 63, 619–637.

Doreian, P., Mrvar, A., 2009. Partitioning signed social networks. Social Networks 31, 1–11.

Emirbayer, M., Goodwin, J., 1994. Network analysis, culture, and the problem of agency. American Journal of Sociology 99, 1411–1454.

Fuhse, J., Mützel, S., 2011. Tackling connections, structure, and meaning in networks: quantitative and qualitative methods in sociological network research. Quality & Quantity 45, 1067–1089.

Geertz, C., 1973. The Interpretation of Cultures: Selected Essays. Basic Books, New York.

Ghaziani, A., Baldassarri, D., 2011. Cultural anchors and the organization of differences. American Sociological Review 76, 179–206.

Giuffre, K., 2009. The return of the natives: globalization and negative ties. Poetics 37, 333–347.

Goldberg, A., 2011. Mapping shared understandings using relational class analysis: the case of the cultural omnivore reexamined. American Journal of Sociology 116, 1397–1436.

Hedegard, D., 2013. Finding 'strong' and 'soft' racial meanings in cultural taste patterns in Brazil. Ethnic and Racial Studies 36, 774–794.

Labianca, G., Brass, D.J., 2006. Exploring the social ledger: negative relationships and negative asymmetry in social networks in organizations. The Academy of Management Review 31, 596–614.

Mische, A., 2008. Partisan Publics: Communication and Contention across Brazilian Youth Activist Networks. Princeton University Press, Princeton, NJ.

Mische, A., 2011. Relational sociology, culture, and agency. In: Scott, J., Carrington, P.J. (Eds.), SAGE Handbook of Social Network Analysis. Sage, London, pp. 80–97.

Mohr, J.W., Duquenne, V., 1997. The duality of culture and practice: poverty relief in New York City, 1888–1917. Theory and Society 26, 305–356.

Mohr, J.W., Rawlings, C.M., 2010. Formal models of culture. In: Hall, J.R., Grindstaff, L., Lo, M. (Eds.), Handbook of Cultural Sociology. Routledge, New York, pp. 119–128.

Ortner, S.B., 1994. Theory in anthropology since the Sixties. In: Dirks, N.P., Eley, G., Ortner, S.B. (Eds.), Culture/Power/History: A Reader in Contemporary Social Theory. Princeton University Press, Princeton, NJ, pp. 372–411.

Pachucki, M.A., Breiger, R.L., 2010. Cultural holes: beyond relationality in social networks and culture. Annual Review of Sociology 36, 205–224.

Riesman, D., 1950. The Lonely Crowd: A Study of the Changing American Character. Yale University Press, New Haven, CT.

Rouanet, H., Ackermann, W., Le Roux, B., 2000. The geometric analysis of questionnaires: the lesson of Bourdieu's *La Distinction* (Corrected version Sept. 2004). Bulletin de Methodologie Sociologique 65, 5–15.

Salvini, A., 2010. Symbolic interactionism and social network analysis: an uncertain encounter. Symbolic Interaction 33, 364–388.

Smith, C., 2010. What is a Person? Rethinking Humanity, Social Life, and the Moral Good from the Person Up. University of Chicago Press, Chicago, IL.

Smith, T., 2007. Narrative boundaries and the dynamics of ethnic conflict and conciliation. Poetics 35, 22–46.

Somers, M.R., 1998. "We're no angels": realism, rational choice, and relationality in social science. American Journal of Sociology 104, 722–784.

Srivastava, S.B., Banaji, M.R., 2011. Culture, cognition, and collaborative networks in organizations. American Sociological Review 76, 207–233.

Swidler, A., 1986. Culture in action: symbols and strategies. American Sociological Review 51, 273–286.

Vaisey, S., 2009. Motivation and justification: a dual-process model of culture in action. American Journal of Sociology 114, 1675–1715.

Vaisey, S., Lizardo, O., 2010. Can cultural worldviews influence network composition? Social Forces 88, 1595–1618.

Varga, A., 2011. Measuring the semantic integrity of scientific fields: a method and a study of sociology, economics and biophysics. Scientometrics 88, 163–177.

White, H.C., 1992. Identity and Control: A Structural Theory of Social Action. Princeton University Press, Princeton, NJ.

White, H.C., Boorman, S.A., Breiger, R.L., 1976. Social structure from multiple networks. I. Blockmodels of roles and positions. American Journal of Sociology 81, 730–780.

Culture and the Self: Implications for Psychological Theory

Joan G Miller and Malin Källberg-Shroff, New School for Social Research, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.G. Miller, volume 5, pp. 3139–4313, © 2001, Elsevier Ltd.

Abstract

This article presents contemporary theory and research on culture and the self. Cultural psychology, it is shown, includes the assumption of a universal capacity for self-awareness and of culture as a system of meanings and practices that form an essential medium for human development. Discussion is given to research findings pointing to fundamental variation in psychological processes in the areas of self and social attribution, cognitive styles and perception, emotions, motivation, and cultural neuroscience. Finally, consideration is given to challenges in contemporary cultural work on the self and to contributions of this work in culturally broadening psychological theory.

The self may be defined as the individual's understanding and experience of their own psychological functioning. As such, it is basic to human experience, while at the same time, also culturally variable in fundamental ways. The present article examines the processes that contribute to the cultural grounding of self and considers implications for understanding variation in psychological processes and outcomes.

Discussion begins with a conceptual argument for recognizing the self as cultural. This is followed by a brief overview of research highlighting the existence of cultural variation in aspects of the self which has implications for different areas of psychological functioning. Finally, attention turns to a consideration of challenges in future work on culture and the self and to the contributions of this work to psychological inquiry.

Theoretical Premises

The recognition of the self as cultural is based on core assumptions regarding the nature of psychological functioning and culture. These include a conception of the person as self-reflexive and agentic, as well as a view of culture as forming a medium for human development.

This perspective assumes that, universally, individuals maintain some awareness of their own mental activity, of themselves as existing in time and space, and of acting in the world. As Hallowell argues:

> It seems necessary to assume self-awareness as one of the prerequisite psychological conditions for the functioning of any human social order, no matter what linguistic and culture patterns prevail…the phenomena of self-awareness in our species is as integral a part of human sociocultural mode of adaptation as it is of a distinctive human level of psychological structuralization.
>
> 1955: p. 75

However, this capacity for self-reflection and self-awareness give rise not only to common psychological experiences of the self, but also makes it possible for humans, through symbolic means, to formulate culturally variable subjective understandings of the self. Thus, whereas there exists universally an empirical agent or ecological sense of self, the modern Western view of self, with its associated cultural practices, represents only one of many culturally specific forms.

Along with this premise of a universal capacity for self-awareness, theorists assume that individuals are agentic in the sense that they actively contribute meaning to experience. This insight may be traced to the Cognitive Revolution in psychology, an intellectual movement that became prominent in psychology during the later twentieth century as a critique of behaviorism. From this perspective, it is understood that, rather than merely passively absorbing externally given information, individuals go beyond the information given when contributing meaning to experience. To give an example from the domain of food, it is recognized that the edibility of particular animals depends on culturally variable meanings (e.g., is eating dog acceptable or a taboo?) and cannot be explained merely on the basis of objective characteristics alone (e.g., nutritional value).

Also of critical importance to the view of self as culturally grounded is the understanding of culture as a symbolic medium for human development, encompassing shared meanings which are embodied in artifacts and practices (Geertz, 1973; Shweder and LeVine, 1984). Cultures are seen as encompassing knowledge of experience and rules for conduct as well as objects and events that are created, in part, by cultural definitions and practices. This constitutive reality-creating role of culture is recognized to be broad, extending not only to the establishment of institutions (e.g., marriages), roles (e.g., bride), and artifacts (e.g., wedding ring), but also to the creation of psychological concepts and epistemological categories. For example, to gain a deeper understanding of key psychological constructs such as emotions or theory of mind, consideration must be given to culturally variable self-understandings (such as concepts of what constitutes an emotion or of the nature of mind itself).

From the present perspective, it is understood that, whereas involuntary responses, like a reflexive response may proceed without cultural mediation, higher order psychological processes depend, on cultural input for their emergence. This insight is known as the incompleteness thesis and assumes that individuals are unable to acquire higher order psychological processes in the absence of cultural input in the form of

experiences that reflect culturally specific meanings and practices. Wertsch expresses this point:

> Cultural, institutional, and historical forces are 'imported' into individuals' actions by virtue of using cultural tools, on the one hand, and sociocultural settings are created and recreated through individuals' use of meditational means, on the other. The resulting picture is one in which…it is virtually impossible for us to act in a way that is not socioculturally situated. Nearly all human action is mediated action, the only exceptions being found perhaps at very early stages of ontogenesis and in natural responses such as reacting involuntarily to an unexpected loud noise.
>
> 1999: p. 160

Thus, for example, whether a physiological reaction to a situation is experienced in emotional terms depends, in part, on culturally based understandings of the nature, causes, and consequences of emotions, cultural routines for responding to emotions, natural language categories for defining emotions, and a range of other sociocultural processes. This leads to the expectation that qualitative differences in self-conceptions and in related modes of psychological functioning can be expected in individuals from cultural communities characterized by contrasting self-related sociocultural meanings and practices.

Impact of Cultural Variation in the Self on Basic Psychological Processes

The present section considers examples of recent empirical studies that illustrate the cultural grounding of psychological processes, an insight critical to the perspective of cultural psychology (e.g., Cole, 1996; Markus et al., 1996; Miller, 1997; Shweder, 1990). Although the overview here is necessarily highly selective and incomplete, it serves to highlight ways in which contemporary cultural research on the self is offering new process explanations of psychological phenomena as well as identifying variability in basic psychological processes.

Self-Processes and Social Attribution

Cultural research on self-perception reveals that in various Asian populations, the self tends to be experienced in more contextualized and fluid terms than typical among American populations. It has been shown that the self-descriptions of Japanese but not those of Americans tend to vary as a function of the presence of others (Kanagawa et al., 2001). Similar effects have been observed in person perception. For example, Church et al. (2006) observed that, whereas American and Euro-Australian populations tend to view traits as stable over time and generalized over contexts, various East Asian cultural populations viewed traits as more affected by situational factors. Perceptions of oneself as displaying consistency across situations has also been found to show a much weaker relationship to psychological well-being among Korean as compared with American populations (Suh, 2000), with contradictory self-knowledge more accessible among Japanese and Chinese as compared with European-Americans (Spencer-Rodgers et al., 2009).

Cultural influences on the self not only extend to perceptions of one's present behavior but also structure autobiographical memory, affecting its timing, content, and developmental course. Research has shown that the age of earliest-reported autobiographical memories is earlier among American than Chinese populations (Wang and Leichtman, 2000; Wang, 2001). This is a reflection of the emphasis placed on highlighting the self's unique perspective in American socialization practices, which encourage early narration of self-focused cognitions. As a result, the autobiographical memories of American children and adults tend to be focused on the self's unique perspective, while the autobiographical memories of Chinese children and adults tend to focus on everyday social routines and include more information about social relations. To give an example, while an American adult might recall as their earliest childhood memory their subjective feelings associated with a special event like their reactions of joy at receiving their first bicycle, a Chinese adult would more typically report a memory of a recurrent social routine such as the memory of sitting in the kitchen as their mother cooked dinner. As illustrated above, the latter type of autobiographical memory typically includes more references to aspects of the social setting, as well as includes other actors involved. These types of memories also tend to be associated with less intense affective reactions than is the case of the memories typically generated by Western populations.

Recent meta-analyses have documented a tendency for non-Western cultural populations to engage in less self-enhancement than is typical among American populations (Heine and Hamamura, 2007). For example, it has been demonstrated that a wide range of non-Western cultural populations including Chileans (Heine and Raineri, 2009) and Fijiians (Rennie and Dunne, 1994) tend to score lower on measures of positive self-view than do Western populations, with various East Asian populations even showing self-effacing or self-critical biases (Heine and Hamamura, 2007). Recent work also highlights the dynamic nature of self-enhancement tendencies and how these are sensitive to contextual variation. For example, East Asian populations do not tend to show self-effacing tendencies when comparing themselves to the average other (e.g., "How friendly are you compared to most people?"), as compared with absolute terms (e.g., "How friendly are you?"; Hamamura et al., 2007). Research in this area has also utilized a version of the Implicit Association Test. Among Japanese and American populations, findings indicate that the tendency to self-enhance primarily occurs among Japanese populations when assessed implicitly, but not when assessed explicitly, such as through conventional self-report measures (Kitayama and Uchida, 2003). This work then provides an indication that some level of tactical self-enhancement is universal while also supporting claims of marked cross-cultural differences in its prominence.

Cognitive Styles and Perception

The ecological model of cognitive differentiation, developed in early work in cross-cultural psychology, provides a central foundation for contemporary work on cognitive styles in cultural psychology. In groundbreaking research, Berry (1974, 1976) provided evidence to support the claim that the

demands set by particular ecological conditions give rise to cultural practices that are adapted to these demands and that, in turn, lead individuals to develop distinctive patterns of cognitive and perceptual skills. In support of this claim, individuals living in ecological conditions that make hunting an adaptive activity were shown to develop better visual discrimination and spatial skills than developed by individuals whose cultures were adapted to different ecological demands. Integrating this framework with work on cognitive differentiation and testing it in an extensive body of research, Witkin and Berry (1975) forwarded a model of culture and cognitive styles that linked culturally based socialization practices adopted in response to different ecological conditions to the development of field dependent as compared with field independent individual cognitive styles.

Addressing many of the same issues and sharing many assumptions with this earlier tradition, contemporary cultural research is building on claims that East Asian populations tend to emphasize a holistic cognitive outlook that privileges a dialectical approach to thought, whereas the outlooks of Western populations give priority to an analytic outlook (Nisbett, 2003). These outlooks are explained in evolutionary terms as having their origins in different social and cultural systems that existed in Ancient Greek and Chinese society such as the tradition of free debate found in Ancient Greece as compared with the emphasis on social control that characterized ancient China. In terms of psychological implications, the contrasting cognitive styles are seen as giving rise to cultural differences in logic and perception.

In terms of logic, Chinese as compared with American participants have been shown to prefer dialectical over classical Western analytic arguments (Peng and Nisbett, 1999). As part of this research, respondents were presented with scientific conclusions that pointed to opposite empirical relationships such as findings that smoking is associated with being thin and findings that suggested the opposite. When presented with such contradictory evidence, Chinese tended to be moderately accepting both positions, whereas the judgments of Americans became more polarized. Work has also shown not only that preferences for analytic as compared with holistic thinking are culturally variable but that in East Asian cultural groups holistic thinking is normatively preferred and is seen as entailing more sophisticated modes of cognition than the forms of holistic thinking prominent among Western populations (Buchtel and Norenzayan, 2009).

Regarding perception, East Asian populations have been found to be more sensitive than Western populations to contextual information presented in a peripheral way as part of the background to a scene (Masuda and Nisbett, 2001). Thus, this research has demonstrated experimentally that the memory of Japanese but not of American respondents is affected by whether an object is presented in the context of its original setting or in a novel setting. Differences in emphasis on analytic versus holistic cognition have also been demonstrated to affect the phenomena of change blindness. Long considered a basic perceptual capacity that is insensitive to cultural variation, change blindness involves the inability to detect changes in objects or scenes that are being viewed. Research on change blindness among Americans has indicated tendencies for individuals to be more sensitive to changes in focal objects than to objects in the periphery or context, whereas the opposite effect has been shown to occur among Japanese (Masuda and Nisbett, 2006).

Recent work also suggests that cultural variation in context sensitivity relates to the tendencies of East Asians to allocate their attention more broadly than Americans, with each culturally linked cognitive style associated with particular advantages and drawbacks. More specifically, it has been found that East Asians are better than Americans at detecting color changes when a set of colored blocks are expanded to cover a wider region but worse than Americans when the same set of blocks are shrunk in size (Boduroglu et al., 2009). Also, European-Americans tend to be more attentive to using part-object cues in perceptual inference than are Asian-Americans (Ishii et al., 2009). For example, in the above study, participants were asked to guess the identity of an object in a picture based on either a small square of the picture (part-cue) or a blurred version of the whole picture (holistic cue).

Contemporary studies are also exploring the extent to which these cultural differences in allocation of attention may be related to features of the physical and sociocultural environment. For example, a comparative study of landscapes found in comparable US and Japanese cities revealed that Japanese urban scenes include more elements than matched American urban scenes (Miyamoto et al., 2006). These same types of cultural differences have also been found in art, with East Asian paintings characterized by greater context inclusiveness and less object centeredness than Western art (Masuda et al., 2008). Findings in this area provide further evidence that cultural variation in patterns of attention may be explained, in part, by the affordances and constraints of different physical settings in our natural and human-made surroundings.

Emotions

Emotions provide a particularly interesting area for cultural research on the self as they involve behavioral action tendencies and somatic reactions, in addition to cognitive appraisals. Recent research points to culture as a powerful influence on the expression and form of emotional reactions, as well as on their mental health implications.

Culture has been observed to impact on the elaboration and interrelationship of different emotions. For example, whereas all cultures employ both socially engaged emotions (e.g., friendliness, happiness) and socially disengaged emotions (e.g., pride, feelings of superiority), among Americans only socially engaged feelings are linked with general positive feelings, whereas among Japanese both types of emotions have such positive links (Kitayama et al., 2006). Researchers have also documented that cultural groups differ in the valence and degree to which they cross-associate different emotions. Thus, for Americans, the affective experiences associated with positive and negative emotions tend to be negatively correlated, while, for Chinese and Koreans, they tend to be positively correlated (Bagozzi et al., 1999). In general, when asked how they feel, Americans tend to report one dominant, positive or negative, emotion at a given time, while East Asians are more likely to report experiencing positive and negative emotions simultaneously. Likewise, Japanese are more likely than Americans to report experiencing both

positive and negative emotions simultaneously about the same exact experience (Miyamoto et al., 2010; Spencer-Rodgers et al., 2010).

Culture also influences how emotional experiences are linked to different situations and adaptive responses. For example, Americans report stronger self-oriented emotions such as pride than East Asians in success-oriented situations (Imada and Ellsworth, 2011). Similarly, memories of success tend to be more accessible than memories of failure among European-Americans, whereas among East Asians both types of memories tend to be equally accessible. Notably, for East Asians, negative emotions typically do not lead to the same negative consequences for self-regulation observed among Americans. For example, individuals from East Asian cultural backgrounds tend to work harder after experiencing the disappointment of failure, while American populations typically experience a decrement in motivation after experiencing the same disappointment or failure (Zhang and Cross, 2011). Research has also shown that the detrimental consequence of reflecting on negative feelings about oneself is mediated by cultural meanings and practices. For example, self-reflection, or brooding, is associated with fewer depressive symptoms among Russians than among Americans, reflecting stronger tendencies among Russians to self-distance themselves during such processes (Grossmann and Kross, 2010).

Finally, culture also impacts the understanding and experience of emotions themselves. Thus, as compared with Dutch adults, Turkish adults tend to categorize emotions as more grounded in assessment of social worth, more reflective of reality than of inner subjective states of the individual, and as located more within the self–other relationship and less confined within the subjectivity of the individual (Mesquita, 2001).

Motivation

Whereas psychological theories of motivation tend to center on the distinction between internal and external locus of control (e.g., Rotter, 1966), cultural work is broadening the existing theoretical models by demonstrating that motivation may assume socially shared forms. This can be seen in the relative emphasis on secondary control, documented to occur among Japanese populations. Secondary control involves a type of agency that entails striving to accept and adjust to situational demands, as well as allowing others to assume or share some control over one's situation (vicarious control; Morling et al., 2002, 2003). This is in contrast to primary control where the focus is on personal influence. Morling and her colleagues found that Americans were better able to recall primary control/influencing situations than adjusting ones (secondary control), whereas Japanese were better able to recall adjusting situations than influencing ones. As an illustration of this, in coping with stressors, such as pregnancy, Japanese also relied more on secondary control involving family members than did Americans. Research in India, likewise, points to the existence of joint forms of control, in which the agent and their family members are experienced as together agentic in bringing about certain outcomes such as academic achievement (Sinha, 1990).

In related work, research has highlighted the positive affective associations linked to role-related social expectations in family and friend relationships in collectivist cultural groups. For example, Asian-American children tend to experience greater intrinsic motivation for an anagrams task that has been selected for them by their mothers than for one that they have freely chosen (Iyengar and Lepper, 1999). In contrast, European-American children tend to experience greater intrinsic motivation when they have selected the task for themselves. Comparative research has further demonstrated that Chinese children interpret the directive behaviors of teachers in more positive terms than do American children (Zhou et al., 2012). In these cases, the Chinese children perceived the teachers' behaviors as less controlling than did the Americans. They also reported being more motivated in their classes than the American children. Findings like these help explain the apparent paradox of Chinese teachers appearing highly controlling their student's behaviors, and yet their students showing strong academic performance and motivation, a relationship typically not found in American classrooms.

Similar types of findings showing agency as compatible with meeting social expectations are also found in research on attribution, calling into question claims that link agency uniquely with individualism. One such study found that Indian adults associated a greater sense of moral duty with meeting the needs of their friends than did Americans (Miller et al., 2011). Furthermore, compared with Americans they also reported greater feelings of choice or of wanting to help, rather than of feeling pressured to help with meeting the needs of their family and friends. While providing evidence for the universality of choice and of internalization in motivation, this work also highlights the need to broaden current work on self-determination theory (e.g., Chirkov et al., 2003). It demonstrates that cultural variation exists in the affective meanings of role obligation, and in the degree to which acting out of a sense of role-based duty is experienced in agentic ways.

Cultural Neuroscience

In recent years, research in cultural psychology is making increasing use of neuroscience approaches. This, in part, reflects the more general growth of neuroscience work in psychology in recent years, particularly with the advent of new methodological tools such as functional neuroimaging.

A major direction of contemporary work on the self in cultural neuroscience is the examination of neurological correlates of culturally variable psychological phenomena (e.g., Ambady and Bharucha, 2009; Chiao and Ambady, 2007). For example, comparisons of the performance of American and East Asian college students on visuospatial tasks has revealed that in each cultural group activation in the frontal and parietal brain regions is greater during culturally nonpreferred as compared with culturally preferred judgments (Hedden et al., 2008). Since these brain regions are known to be associated with attentional control, such results indicate that greater attention is required to process stimuli that are less culturally familiar. For example, Hedden and his colleagues demonstrated that processing of a task that

involved relative as compared with absolute comparisons of the size of two stimuli involved less activation of the parietal brain regions among the East Asian sample than among the American sample. This reflects the East Asian participants' cultural preference for contextually based information processing, thus requiring less attention, with the reverse observed among the Americans.

Cultural neuroscience also provides evidence that in cases where cross-cultural variation exists in social understanding, different brain mechanisms may be implicated. For example, distinct brain regions are activated among both American and Japanese adults when undertaking theory of mind tasks. These are believed to reflect cultural differences in the relative focus on mental state as compared with contextual information (Naito and Koyama, 2006), Similarly, further research indicates that Japanese respondents show greater brain activation of the inferior frontal gyrus, a brain region linked to emotional processing, than do American respondents (Kobayashi et al., 2008). Also, research in cultural neuroscience is providing evidence that the same brain regions may be recruited in the service of culturally variable modes of self-understanding. One such finding suggests that whereas the medial prefrontal cortex tends to be recruited in representing only information about the individual self among Western populations, among Chinese the same brain region is recruited in representing both the self and close others (Zhu and Han, 2008).

An additional direction of contemporary work in cultural neuroscience is the examination of cultural influences on neural plasticity. Recent work has shown that American young adults engage object-processing areas in the ventral visual cortex more than Chinese young adults when assessed on a visuospatial task (Gutchess et al., 2006). Additionally, findings from research conducted among elderly populations indicate that this cultural difference in brain processing actually increases with age, with elderly Singaporeans displaying larger deficits in object-processing brain areas than elderly Americans (Chee et al., 2006). Such work provides neurological support for a 'use it or lose it' view of cognitive aging, indicating that culture influences not only how brains are used, but also patterns of loss in brain functioning associated with aging (Park and Huang, 2010).

Implications

Challenges

A major challenge in contemporary work on culture and the self is to develop more sophisticated and sensitive approaches to culture (Miller, 2012). Much of the work in cultural psychology to date has been informed by the interdependent/independent distinction or the related East/West or collectivism/individualism dichotomies, distinctions that have been tapped empirically in research by the use of standardized scales (Cross et al., 2011). This perspective on culture has grown in prominence in recent years with the emergence of priming techniques to assess culture (e.g., Oyserman and Lee, 2008). For example, one type of priming involves presenting individuals with a paragraph that describes the behavior of an individual who is referred to either only with the pronoun 'I,' in the independent case, or by the pronoun, 'We,' in the interdependent case (e.g., Gardner et al., 1999). However, these types of approaches are limited by their tendencies to gloss over important distinctions between and within cultures, and by giving limited attention to the impact of context on behavior.

One way that theorists are working to address this challenge is to give more attention to within and between culture variations in outlooks that may be linked to regional, ethnic, or socioeconomic differences. To illustrate, research has identified distinct regional concerns among geographically distinct groups of Americans, with investigators observing such differences as a concern with personal growth and feeling cheerful being emphasized in the US Southwest, compared with a concern with not being socially constrained in New England (e.g., Markus et al., 2004). Researchers are also increasingly attending to the impact of social class. For example, recent findings suggest that individuals from a lower social class as compared with those from a higher social class are more prone to act in a prosocial manner (Piff et al., 2010) and to place greater emphasis on contextual considerations in social attribution (Kraus et al., 2010). An increasing amount of research focuses on expanding comparisons to more cultural groups. This type of stance can be seen in work showing that Central and East Europeans tend to be more holistic than are Americans (Varnum et al., 2008) and, similarly, that southern Italians are generally more interdependent than northern Italians (Knight and Nisbett, 2007).

However, even with this greater attention to diversity, it remains vital to approach work in this area with theoretical understandings that are more culturally specific than the types of dichotomies characterizing present research. It is important to recognize that, while there are similarities between the responses of working-class individuals within the US and those of various collectivist cultural groups from other countries, marked differences also distinguish subgroups of different cultural and ethnic backgrounds from the same country, while simultaneously interacting with influences of social class. It is not only important to extend analysis to more culturally diverse subgroups, but also vital to assess aspects of culture that are more nuanced than the present well-worn cultural comparisons such as the broad distinction between analytic versus holistic thought. For example, forms of holistic thinking are not only found in East Asian cultural communities but are also evident in India as well as various European and South American cultural groups, calling into question theories that explain the origins of these modes of thought in terms of historical traditions linked to Greek and Chinese thought. Instead, it appears likely that preferences for analytic versus holistic thought are anchored in some type of shared features that are highly functional in the communities where they exist. By attending to the distinctive beliefs, practices, and historical traditions of more specific cultural communities, it will be possible to identify more subtle and dynamic insights into cultural and subcultural viewpoints. To make sure future theoretical improvements are even more refined, a further challenge is to make efficient use of the rich traditions of anthropology and sociology, by incorporating their findings into empirical research.

Similarly, as work from a cultural psychology perspective gives greater attention to neurological evidence, it is important to attend to challenges involved in adopting such approaches (e.g., Kagan, 2007). It must be recognized that in many cases neuroscientific evidence merely confirms the findings of research utilizing conventional methods, while adding little theoretical value to existing conclusions. In addition, consideration needs to be given to the extent to which many contemporary neuroscience findings reflect what has been characterized as the 'seductive allure' of neuroscience, thereby supporting unwarranted deterministic conclusions (Gergen, 2010; Miller and Kinsbourne, 2012). This 'seductive allure' is essentially the tendency of neuroscience findings to be favored by the media, funding agencies, and policy makers because they suggest a material cause to something behavioral. The simplicity of a material cause can thus be very appealing, leading to fewer pursuits of alternative explanations. An example of this can be seen in the case of adolescent risk-taking. Findings from neuroscience describe the adolescent brain as fairly immature compared with the brains of adults. These findings have been proposed as a biological explanation for risk-taking tendencies common among adolescents (Steinberg, 2007). Supporters of this view have further suggested that environmental influence on risk-taking is likely limited. This conclusion is sharply deterministic, especially when considering cross-cultural evidence indicating considerable variation in adolescent risk-taking. In particularly, those cultures with norms surrounding adolescence that are distinctly different from those in modern Western societies tend to have less reported occurrences of adolescent risk-taking behavior (Miller and Kinsbourne, 2011).

Contributions

Research on culture and the self is broadening existing psychological constructs and theories through the identification of alternative normative endpoints for human psychological functioning. It is providing insight into ways that cultural meanings and practices form a medium for human development. Dynamic in nature, work on culture and the self is attempting to move beyond the overly static and global tendencies of earlier models. This is seen in recent attempts to transcend traditional East–West comparisons, taking into consideration within-culture variation as well as variation related to social class.

Even while addressing challenges and evolving in new directions, research on culture and the self continues to make major contributions to basic psychological theory. Findings from this line of research serve to make explicit the often implicit cultural processes that underlie existing psychological theory. By continuously working toward identifying models of human psychological experience that are more culturally inclusive than present ones, this type of research forms a necessary part of any truly universal theory of the self.

Finally, one of the most important contributions that research on culture is making is continuing to move beyond the linear approach. That human psychological functioning has different normative endpoints depending on cultural influences is relatively easy to digest. Harder, but all the more interesting, is turning the focus away from the continuum to something akin to parallel, qualitatively different, psychological universes. Evidence from cognition and perception, some of which was highlighted earlier, illustrates this point particularly well. When something as seemingly straightforward as the perception of the physical world is culturally variable, implications for human development are far-reaching. This is not to suggest an extreme form of relativism that precludes all comparison. However, it does require considerably more effort and the challenging of preconceived notions at every corner.

See also: Collectivism and Individualism: Cultural and Psychological Concerns; Cross-Cultural Psychology; Cultural Psychology; Culture and Emotion; Identity and Identification, Social Psychology of; Social Cognition; Social Psychology.

Bibliography

Ambady, N., Bharucha, J., 2009. Culture and the brain. Current Directions in Psychological Science 18 (6), 342–345.

Bagozzi, R.P., Wong, N., Yi, Y., 1999. The role of culture and gender in the relationship between positive and negative affect. Cognition and Emotion 13 (6), 641–672.

Berry, J.W., 1974. An ecological approach to cross cultural psychology. Man-Environment Systems 4 (6), 365–383.

Berry, J.W., 1976. Human Ecology and Cognitive Style. Sage-Halsted, New York.

Boduroglu, A., Shah, P., Nisbett, R.E., 2009. Cultural differences in allocation of attention in visual information processing. Journal of Cross-Cultural Psychology 40 (3), 349–360.

Buchtel, E.E., Norenzayan, A., 2009. Thinking across cultures: implications for dual processes. In: Evans, J.S.B.T., Frankish, K. (Eds.), In Two Minds: Dual Processes and beyond. Oxford University Press, New York, NY, pp. 217–238.

Chee, M.W.L., Goh, J.O.S., Venkatraman, V., Tan, J.C., Gutchess, A., Sutton, B., Park, D., 2006. Age-related changes in object processing and contextual binding revealed using fMR adaptation. Journal of Cognitive Neuroscience 18 (4), 495–507.

Chiao, J.Y., Ambady, N., 2007. Cultural Neuroscience: Parsing Universality and Diversity across Levels of Analysis. Handbook of Cultural Psychology. Guilford Press, New York, NY.

Chirkov, V., Ryan, R.M., Kim, Y., Kaplan, U., 2003. Differentiating autonomy from individualism and independence: a self-determination theory perspective on internalization of cultural orientations and well-being. Journal of Personality and Social Psychology 84 (1), 97–110.

Church, A.T., Katigbak, M.S., del Prado, A.M., Valdez-Medina, J.L., Miramontes, L.G., Ortiz, F.A., 2006. A cross-cultural study of trait self-enhancement, explanatory variables, and adjustment. Journal of Research in Personality 40 (6), 1169–1201.

Cole, M., 1996. Cultural Psychology: A Once and Future Discipline. Harvard University Press, Cambridge, MA.

Cross, S.E., Hardin, E.E., Gercek Swing, B., 2011. The *what, how, why* and *where* of self-construal. Personality and Social Psychology Review 15 (2), 142–179.

Gardner, W.L., Gabriel, S., Lee, A.Y., 1999. "I" value freedom, but "we" value relationships: self-construal priming mirrors cultural differences in judgment. Psychological Science 10 (4), 321–326.

Geertz, C., 1973. The Interpretation of Cultures. Basic Press, New York, NY.

Gergen, K.J., 2010. The acculturated brain. Theory & Psychology 20 (6), 795–816.

Grossmann, I., Kross, E., 2010. The impact of culture on adaptive versus maladaptive self-reflection. Psychological Science 21 (8), 1150–1157.

Gutchess, A.H., Welsh, R.C., Boduroğlu, A., Park, D.C., 2006. Cultural differences in neural function associated with object processing. Cognitive, Affective & Behavioral Neuroscience 6 (2), 102–109.

Hallowell, A.I., 1955. Culture and Experience. University of Pennsylvania Press, Philadelphia, NJ.

Hedden, T., Ketay, S., Aron, A., Markus, H.R., Gabrieli, J.D.E., 2008. Cultural influences on neural substrates of attentional control. Psychological Science 19 (1), 12–17.

Heine, S.J., Hamamura, T., 2007. In search of East Asian self-enhancement. Personality and Social Psychology Review 11 (1), 1–24.

Hamamura, T., Heine, S.J., Takemoto, T.R.S., 2007. Why the better-than-average effect is a worse-than-average measure of self-enhancement: an investigation of conflicting findings from studies of East Asian self-evaluations. Motivation and Emotion 31 (4), 247–259.
Heine, S.J., Raineri, A., 2009. Self-improving motivations and collectivism: the case of Chileans. Journal of Cross-Cultural Psychology 40 (1), 158–163.
Imada, T., Ellsworth, P.C., 2011. Proud Americans and lucky Japanese: cultural differences in appraisal and corresponding emotion. Emotion 11 (2), 329–345.
Ishii, K., Tsukasaki, T., Kitayama, S., 2009. Culture and visual perception: does perceptual inference depend on culture? Japanese Psychological Research 51 (2), 103–109.
Iyengar, S.S., Lepper, M.R., 1999. Rethinking the value of choice: a cultural perspective on intrinsic motivation. Journal of Personality and Social Psychology 76 (3), 349–366.
Kagan, J., 2007. A trio of concerns. Perspectives on Psychological Science 2 (4), 361–376.
Kanagawa, C., Cross, S.E., Markus, H.R., 2001. "Who am i?" The cultural psychology of the conceptual self. Personality and Social Psychology Bulletin 27 (1), 90–103.
Kitayama, S., Uchida, Y., 2003. Explicit self-criticism and implicit self-regard: evaluating self and friend in two cultures. Journal of Experimental Social Psychology 39 (5), 476–482.
Kitayama, S., Mesquita, B., Karasawa, M., 2006. Cultural affordances and emotional experience: socially engaging and disengaging emotions in Japan and the United States. Journal of Personality and Social Psychology 91 (5), 890–903.
Knight, N., Nisbett, R.E., 2007. Culture, class and cognition: evidence from Italy. Journal of Cognition and Culture 7 (3–4), 283–291.
Kobayashi, C., Glover, G.H., Temple, E., 2008. Switching language switches mind: linguistic effects on developmental neural bases of 'theory of mind.' Social Cognitive and Affective Neuroscience 3 (1), 62–70.
Kraus, M.W., Côté, S., Keltner, D., 2010. Social class, contextualism, and empathic accuracy. Psychological Science 21 (11), 1716–1723.
Markus, H.R., Kitayama, S., Heiman, R.J., 1996. Culture and "Basic" Psychological Principles. Social Psychology: Handbook of Basic Principles. The Guilford Press, New York, NY, pp. 857–913.
Markus, H.R., Plaut, V.C., Lachman, M.E., 2004. Well-being in America: Core Features and Regional Patterns. How Healthy Are We?: A National Study of Well-being at Midlife. University of Chicago Press, Chicago, IL.
Masuda, T., Nisbett, R.E., 2001. Attending holistically versus analytically: comparing the context sensitivity of Japanese and Americans. Journal of Personality and Social Psychology 81 (5), 922–934.
Masuda, T., Nisbett, R.E., 2006. Culture and change blindness. Cognitive Science: A Multidisciplinary Journal 30 (2), 381–399.
Masuda, T., Gonzalez, R., Kwan, L., Nisbett, R.E., 2008. Culture and aesthetic preference: comparing the attention to context of East Asians and Americans. Personality and Social Psychology Bulletin 34 (9), 1260–1275.
Mesquita, B., 2001. Emotions in collectivist and individualist contexts. Journal of Personality and Social Psychology 80 (1), 68–74.
Miller, J.G., Das, R., Chakravarthy, S., 2011. Culture and the role of choice in agency. Journal of Personality and Social Psychology 101 (1), 46–61.
Miller, J.G., Kinsbourne, M., 2012. Culture and neuroscience in developmental psychology: contributions and challenges. Child Development Perspectives 6 (1), 35–41.
Miller, J.G., 1997. Theoretical issues in cultural psychology. In: Berry, J.W., Poortinga, Y.H., Pandey, J. (Eds.), Handbook of Cross-Cultural Psychology, Theory and Method, second ed., vol. 1. Allyn & Bacon, Inc, Boston, MA, pp. 85–128.
Miller, J.G., 2012. Bringing culture to basic psychological theory: beyond individualism and collectivism: comment on Oyserman et al. (2002). Psychological Bulletin 128 (1), 97–109.
Miyamoto, Y., Nisbett, R.E., Masuda, T., 2006. Culture and the physical environment: holistic versus analytic perceptual affordances. Psychological Science 17 (2), 113–119.
Miyamoto, Y., Uchida, Y., Ellsworth, P.C., 2010. Culture and mixed emotions: co-occurrence of positive and negative emotions in Japan and the United States. Emotion 10 (3), 404–415.
Morling, B., Kitayama, S., Miyamoto, Y., 2002. Cultural practices emphasize influence in the United States and adjustment in Japan. Personality and Social Psychology Bulletin 28 (3), 311–323.
Morling, B., Kitayama, S., Miyamoto, Y., 2003. American and Japanese women use different coping strategies during normal pregnancy. Personality and Social Psychology Bulletin 29 (12), 1533–1546.
Naito, M., Koyama, K., 2006. The development of false-belief understanding in Japanese children: delay and difference. International Journal of Behavioral Development 30 (4), 290–304.
Nisbett, R.E., 2003. The Geography of Thought: How Asians and Westerners Think Differently – and Why. Free Press, New York, NY.
Oyserman, D., Lee, S.W.S., 2008. Does culture influence what and how we think?: Effects of priming individualism and collectivism. Psychological Bulletin 134 (2), 311–342.
Park, D.C., Huang, C.-M., 2010. Culture wires the brain: a cognitive neuroscience perspective. Perspectives on Psychological Science 5 (4), 391–400.
Peng, K., Nisbett, R.E., 1999. Culture, dialectics, and reasoning about contradiction. American Psychologist 54 (9), 741–754.
Piff, P.K., Kraus, M.W., Côté, S., Cheng, B.H., Keltner, D., 2010. Having less, giving more: the influence of social class on prosocial behavior. Journal of Personality and Social Psychology 99 (5), 771–784.
Rennie, L.J., Dunne, M., 1994. Gender, ethnicity, and students' perceptions about science and science-related careers in Fiji. Science Education 78 (3), 285–300.
Rotter, J.G., 1966. Generalized expectancies for internal versus external locus of control of reinforcement. Psychological Monographs: General and Applied 80. Whole No. 609.
Shweder, R.A., LeVine, R.A., 1984. Culture Theory: Essays on Mind, Self, and Emotion. Cambridge University Press, Cambridge, MA.
Shweder, R.A., 1990. Cultural psychology – what is it? In: Stigler, J.W., Shweder, R.A., Herdt, G. (Eds.), Cultural Psychology: Essays on Comparative Human Development. Cambridge University Press, New York, NY, pp. 27–66.
Sinha, D., 1990. Applied cross-cultural psychology and the developing world. International Journal of Psychology 25 (3), 381–386.
Spencer-Rodgers, J., Boucher, H.C., Mori, S.C., Wang, L., Peng, K., 2009. The dialectical self-concept: contradiction, change, and holism in East Asian cultures. Personality and Social Psychology Bulletin 35 (1), 29–44.
Spencer-Rodgers, J., Peng, K., Wang, L., 2010. Dialecticism and the co-occurrence of positive and negative emotions across cultures. Journal of Cross-Cultural Psychology 41 (1), 109–115.
Steinberg, L., 2007. Risk taking in adolescence: new perspectives from brain and behavioral science. Current Directions in Psychological Science 16, 55–59.
Suh, E.M., 2000. Self, the Hyphen between Culture and Subjective Well-being; Culture and Subjective Well-being. The MIT Press, Cambridge, MA.
Varnum, M.E.W., Grossmann, I., Katunar, D., Nisbett, R.E., Kitayama, S., 2008. Holism in a European cultural context: differences in cognitive style between central and East Europeans and Westerners. Journal of Cognition and Culture 8 (3–4), 321–333.
Wang, Q., Leichtman, M.D., 2000. Same beginnings, different stories: a comparison of American and Chinese children's narratives. Child Development 71 (5), 1329–1346.
Wang, Q., 2001. Culture effects on adults' earliest childhood recollection and self-description: implications for the relation between memory and self. Journal of Personality and Social Psychology 81 (2), 220–233.
Wertsch, J.V., 1999. Sociocultural research in the copyright age. In: Lloyd, P. (Ed.), Lev Vygotsky: Critical Assessments: Future Directions, vol. IV. Taylor & Francis, Florence, KY, pp. 144–163.
Witkin, H.A., Berry, J.W., 1975. Psychological differentiation in cross-cultural perspective. Journal of Cross-Cultural Psychology 6, 4–87.
Zhang, M., Cross, S.E., 2011. Emotions in memories of success and failure: a cultural perspective. Emotion 11 (4), 866–880.
Zhou, N., Lam, S.-F., Chan, K.C., 2012. The Chinese classroom paradox: a cross-cultural comparison of teacher controlling behaviors. Journal of Educational Psychology 104, 1162–1174.
Zhu, Y., Han, S., 2008. Cultural differences in the self: from philosophy to psychology and neuroscience. Social and Personality Psychology Compass 2 (5), 1799–1811.

Culture as a Determinant of Mental Health

Carolyn B Murray, University of California, Riverside, Riverside, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article investigates the cultural determinants of mental health for ethnic minorities (i.e., African Americans, Asian Americans, Latinos, and Native Americans) within the USA. The existing literature indicates that different cultures encourage different worldviews, which can either predispose its members to certain disorders or buffer their immediate effects. In addition, these cultures can shape the interaction between the mental health provider and their clients through symptom presentation, diagnosis, and treatment. This article concludes that, given the complexity of human behavior, it is futile for mental health practitioners to attempt an understanding of ethnic minorities without adequate exploration of each client's individual historical background, societal treatment, subcultural values, and unique conflicts.

Introduction

Parallel to the revolution occurring with regard to the development of psychopharmacological therapies and biological theories of mental disorders, the health sciences and related professions are undergoing a less visible change in theory and practice that promises to radically alter traditional notions about the nature of mental health (Kao et al., 2004; Kleinman and Benson, 2006). The main underlying tenant of this occurrence is the belief that social determinants are critical variables for understanding the etiology, treatment, and prevention of many mental disorders previously attributed to biological causes. Known as 'cross-cultural psychology,' this approach is defined in terms of the study of behavior and experience as these occur in different cultures. Cross-cultural psychology is not in itself a separate branch of psychology, but rather a unique point of view and methods applicable in principle to all other areas of psychology (Ratner and Hui, 2003).

The interplay between culture and mental health is complex and multidimensional with respect to antecedents, perceptions, expressions, and treatments for a variety of mental disorders. Since 1965 the cultural and racial makeup of the United States has changed more rapidly than during any other period in its history. The antecedents of this change were (1) reform in immigration policy in 1965, (2) the increase in self-identification by ethnic minorities, and the (3) reduction in birth rates, especially among non-Hispanic White Americans. America is now a cultural diverse, racial, and ethnic society. Along with this increase in cultural diversity, however, are very serious disparities regarding the mental health status, nosology, and services delivered to cultural minorities (U.S. Department of Health and Human Services, 2001; Bussing and Gary, 2012).

This article investigates the cultural determinants of mental health for nondominant ethnic minorities (i.e., African Americans, Asian Americans, Latinos, and Native Americans) within the USA. The likelihood that different cultures encourage different worldviews that buffer or predispose its members to certain disorders is examined. In addition, we also examine whether these cultures shape the interaction between the mental health provider and their clients through symptom presentation, diagnosis, and treatment. It is proposed that social forces play a critical part in mental health, particularly the attitudes and expectations held by societal authorities (e.g., police, social workers, therapists) toward minorities. This article concludes that, given the complexity of human behavior, it is futile for mental health practitioners to attempt an understanding of ethnic minorities without adequate exploration of each client's individual historical background, societal treatment, subcultural values, and unique conflicts.

The Conceptualization of Culture

In general, culture consists of those aspects of life that people construct and share within a specific reference group (e.g., race, ethnic, social class, age cohort, gender, nationality, and/or profession). The term 'culture' is so widely used and multidimensional in its applications (e.g., learned as opposed to instinctive behavior; tools humans create; and values and traditions) that its precise meaning varies from one situation to another (Kao et al., 2004). Culture can be seen as the collective heritage of a people handed down from generation to generation. It includes language, religious beliefs, customs, rules of etiquette, and ideas used by people to organize and interpret their lives and existence. Thus, the term 'culture' refers generally to the way a people live, the rules of behavior promoted, and the conduct they set for themselves.

As a nation of immigrants, the USA has evolved into a multicultural society. Most researchers, however, address race and ethnicity as demographic characteristics rather than as distinct predisposing cultural and social environmental orientations. This approach has resulted in research and analyses that ignore the contributory role of sociocultural factors to mental health behaviors. Research is needed to understand how these factors contribute to individual and group coping and adaptive mechanisms, particularly in alleviating the distinct socioeconomic and psychological disadvantages of categorical membership. Culture should be conceptualized as a potential individual and group resource, providing psychological, social, and personal identification and group connectedness and worldview (Kao et al., 2004).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.14011-5

Diagnostic and Statistical Manual of Mental Disorders: The Relevance of Culture

In 1952 The American Psychiatric Association, recognizing that practitioners needed a common language and standardized criteria in order to diagnose and communicate effectively with each other, began publishing the *Diagnostic and Statistical Manual of Mental Disorders* (DSM). It evolved from a United States Army manual that reported census data and psychiatric hospital statistics to standardized psychiatric diagnostic categories and criteria. Today, the DSM is the diagnostic tool of choice for identifying psychiatric disorders within the United States and internationally. Relevant to the present topic, during the last three decades, however, the manual has been severely criticized for its superficial coverage of the role of culture in diagnosis and treatment of mental illness (Kao et al., 2004; Mezzich et al., 1999; Sue et al., 2012; Sue and Sue, 1999). In response to this criticism in 1994 the DSM-IV introduced cultural considerations for the use of diagnostic categories and criteria, including a glossary of cultural-bound syndromes and idioms of distress along with an outline for cultural formulation. Leading cross-cultural psychiatrists argued that this update further perpetuated the idea that Western cultural phenomena are universal with nonwestern cultural phenomena being 'culture-bound' aberrant manifestations, and its criteria are rarely applied even when it is appropriate to do so (Kleinman and Benson, 2006).

The most recent attempt to incorporate culture into the treatment of mental disorders has resulted in the vastly improved *DSM*-5 (APA, 2013). For instance, in the previous edition, the DSM-IV-TR the use of the term 'culture-bound' made mental health conditions that were specific to a certain population appear highly localized and confined – a bag of oddities (APA, 2013). The list was also heterogeneous in that some 'syndromes,' including *nervios*, seemed to represent specific situational predicaments, or variations in the way people express their distress, rather than coherent collections of symptoms. Other expressions, such as *ataque de nervios*, are syndromes, but do not always represent psychopathology. Still others (like *shenjing shuairuo* or 'neurasthenia') appeared to be a cover term for several common, but seemingly unrelated, human maladies (e.g., fatigue, dizziness, headache, GI problems, sexual dysfunction, excitability); another group of conditions simply defied DSM categorization, such as *mal de ojo* ('evil eye').

The recently updated DSM-5 includes a 'Glossary of Cultural Concepts of Distress,' and a section on the cultural formulation, featuring a revised version of the outline introduced in DSM-IV as well as an approach to assessment, using the Cultural Formulation Interview. In addition, the chapter includes a section discussing 'Cultural Concepts of Distress.' The notion of 'culture-bound syndromes' has been replaced by three concepts: (1) *cultural syndromes*: "clusters of symptoms and attributions that tend to co-occur among individuals in specific cultural groups, communities, or contexts…that are recognized locally as coherent patterns of experience"; (2) *cultural idioms of distress*: "ways of expressing distress that may not involve specific symptoms or syndromes, but that provide collective, shared ways of experiencing and talking about personal or social concerns"; and (3) *cultural explanations of distress or perceived causes*: "labels, attributions, or features of an explanatory model that indicate culturally recognized meaning or etiology for symptoms, illness, or distress" (APA, 2013). The information provided throughout DSM-5, and particularly in the cultural formulation chapter should help practitioners avoid misdiagnosis, obtain clinically useful information, improve clinical rapport and therapeutic efficacy, guide research, and clarify cultural epidemiology.

Culture and Coping

Culture is transmitted through the process of socialization. Throughout this process, both general and specific behavioral patterns for promoting human survival and growth are conditioned. What are learned are acceptable standards and idealized norms that reflect the hard-won wisdom of previous generations. Through culture, biological adaptive capacities are extended to deal with environmental demands. By understanding the ways that individuals from various cultural traditions cope with mental illness, health researchers and practitioners can gain unique insights into the promotion of human health and prevention of human illness (Chen and Mak, 2008).

For nondominant minority group members, the problems of living are not only the products of biological dysfunctions, aberrant intrapsychic development, and familial dysfunctions, social forces also contribute to the onset and course of mental illness (Kleinman, 2009). Political, social, economic, and historical and present-day relationships (e.g., slavery, legal discrimination and segregation, immigration exclusion and restrictions, internment, poverty, and forced removal to reservations) with the dominant group in the USA, further contributes to mental illness among minorities (Woods et al., 2012). In effect, the larger society has played a major role in the developmental etiology of certain mental health disturbances (Wilson et al., 2000). How do the majority of people of color, in spite of ethnic and racial exploitation, cope and succeed?

To answer this question, culturally defined strengths associated with ethnic minority group life should be illuminated. While racial and ethnic groups of color vary widely in the content of cultural beliefs about time, harmony with nature, reciprocity, uses of traditional medical substances, and community healers, certain common themes emerge. These include deep and abiding community respect for elders, an extended kin network, predominant family commitment, cultural aversion to institutional care, and a clear expectation for respect and dignity in the delivery of health care. Some researchers have specifically commented on the elaborate support networks, extended family relationships, and role flexibility common to African Americans; the strong sense of communalism, generosity, and harmony characteristic of Native Americans; the socialization of children to respect the entire family system and to put family needs over individual needs in Asian American communities; the socializing toward cooperativeness, and the comfort of their Catholic faith in times of stress found in Latino communities (U.S. Department of Health and Human Services, 2001; Wilson et al., 2000). These group characteristics facilitate coping and mental health, and lessen the nature and expression of disease

and chronic conditions, and are necessary to consider for successful care and treatment.

Culture dictates how people cope with everyday problems and more extreme types of adversity. For instance, in comparison to White Americans, many Asian American groups tend not to dwell on upsetting thoughts, believing that reticence or avoidance is better than outward expression. Such Asian subgroups place a higher emphasis on suppression of affect (Chen and Mak, 2008), with some tending to first rely on themselves to cope with distress. Asian Americans are also less likely to seek social support because they are more cautious about potentially upsetting their social network (Kim et al., 2006). This cautious approach to medical health challenges sheds some light on why Asian Americans report an unwillingness to speak to anyone regarding mental health concerns. Interestingly, Asian individuals exposed to and influenced by Western norms and practices hold more positive attitudes toward seeking professional help and actually used more mental health services (Leong et al., 2011). These findings demonstrate the importance of cultural beliefs in shaping help-seeking attitudes and behaviors (Chen and Mak, 2008; Mallinckrodt et al., 2005).

African Americans, on the other hand, tend to take an active approach in facing personal problems, rather than avoiding them. They are more inclined than Whites to depend on handling distress on their own or within familial settings (Conner et al., 2010). They also appear to rely more on spirituality to help them cope with adversity and symptoms of mental illness (Mattis et al., 2007; Ward et al., 2013). African-American older adults, for instance, suffer more psychological distress than their White counterparts due to their exposure to and experiences with racism, discrimination, prejudice, poverty, and violence (Williams et al., 2003) and they tend to have fewer social and financial resources for coping with this stress than their White counterparts (Choi and Gonzalez, 2005). In one study, when asked to identify effective coping strategies for their depression, the most common strategies included handling depression on their own, pushing through the depression, denial, and relying upon God (Conner et al., 2010).

While there are group commonalties, intragroup differences exist in life-course environmental conditions, reflecting differential life experiences and various degrees of acculturation. Thus, there is an indisputable need for mental health care providers to treat each client as an individual, while gathering information on the degree to which cultural coping strategies may facilitate mental health. It is imperative to conduct cross-cultural research on ethnic minorities within the US and within other multicultural societies to identify the relationship between culture-specific coping skills and mental health outcomes.

Cross-Cultural Perception and Expression of Mental Illness

Although most, if not all, cultures have a notion of mental or psychological illness, ethnographic studies demonstrate that concepts of emotions, self and body, and illness categories differ so significantly cross-culturally that it can be said that each culture's beliefs about normal and abnormal behavior is distinctive (Kleinman, 1988). The prevalence and course of mental illness also varies cross-ethnically. Recent research comparing African Americans to Whites indicated no difference between the two groups with regard to the percentage of panic disorders. However, African Americans present a greater frequency of reported tingling in the extremities and a higher mean number of symptoms (Draguns, 2000).

Another study estimated the lifetime prevalence of seven psychiatric affective disorders (i.e., panic disorder, agoraphobia, social phobia, generalized anxiety disorder, post-traumatic stress disorder, major depressive disorder, and dysthymia) for older non-Hispanic White people, African Americans, Caribbean Black people, Latinos, and Asian Americans. The findings indicated that major depressive disorder and social phobia were the two most prevalent disorders among the seven psychiatric conditions. Overall, Whites and Latinos consistently had higher prevalence rate of the seven disorders, African Americans had lower rates of major depression and dysthymia, and Asian Americans were typically less likely to report affective disorders than their counterparts (Woodward et al., 2012). There were also across and within group variations that followed demographic, socioeconomic, and immigration status.

Another classical difference often ignored by Western practitioners is the finding that bodily states and psychological experiences are monitored, perceived, assessed, and reported differently by members of different cultural groups. A consistent finding is that body complaints predominate over psychological complaints in depressive and anxiety disorders among members of non-Western societies, ethnic minorities in the US, and among less educated members of the lower socioeconomic classes (Chen and Mak, 2008).

Culture also influences how depression is experienced and communicated. This may be due to some cultures perceiving the physician as a healer of physical ailments as opposed to psychological ailments. The urgency and sense of legitimacy surrounding physical ailments may vary across cultures, as may the degree to which psychological weakness is stigmatizing. The biology of depression and anxiety disorders underwrites the inner form of these disorders, but cultural beliefs and values shape their expression (Kleinman, 2009).

Etiological Beliefs and Folk Approaches

All healing systems are shaped by the values, worldviews, and self-views of the cultures in which they develop and are utilized. Therefore, all healing systems are in a real sense 'ethnomedical.' Rules and principles guiding the various ethnomedical models form discreet and different genre, making cross-ethnic comparisons difficult. However, while distinctions among the non-Western groups are difficult to delineate, very salient differences exist between various ethnomedical models and traditional Western methods in psychotherapy. For instance, Western psychotherapy is traditionally a one-on-one activity that encourages clients to discuss the most intimate aspects of their lives with health practitioners. In Latino culture, clients rarely describe their complaints as the main step in an intake process since they expect the healer to already know their

complaints (Koss-Chioino, 2000). Due to experience with oppression, African Americans are hesitant to self-disclose. This may serve a protective function against possible physical and psychological harm (Conner et al., 2010), but individuals who do not self-disclose may be seen as resistant, defensive, or superficial by the therapist. Many minorities who are reluctant to self-disclose are judged to be mentally unhealthy and, in the case of African Americans, are diagnosed as 'paranoid' (Murray, 1998; Williams et al., 2003).

In spite of the differences in each group's healing practices, cultural commonalties do exist across the four ethnic groups described in this article. Specifically, the commonalties consist of (1) a widespread belief that the causes of mental illness exist outside the person, (2) preference for a group orientation to healing (i.e., family and/or community) as opposed to an individual orientation, (3) an emphasis on morality as a cause and condition for recovery (Koss-Chioino, 2000), and (4) an emphasis on short-range goals, as opposed to long-range goals (Sue and Sue, 1999). These similar healing practices are common to other non-Western groups but different from most Western therapeutic practices.

Latinos, Asians, Native Americans, and African Americans are all heterogeneous groups, and they have many specific traditions that differ from each other in profound ways (Jimenez et al., 2012). But there are similarities as well. A belief in Divination (i.e., beliefs in the supernatural and practices to connect with the spirit world), for example, was originally part of each of these groups' healing practices. Unlike Western cultural groups who have discarded such practices, divination is still part of the healing practices of many, but not all, of the subcultural groups within each of the broader categories. Among Native Americans, for example, traditional healing practices vary along tribal lines and are largely shamanic (spiritual intervention to heal the sick). One of the major shamanic healing entities is the Native American Church, which is a pan-Indian religion based on a ritual that facilitates direct contact with the Great Spirit (also God) through peyote, pipe and cigarette smoking, drumming and songs, sweat lodge practices, and prayer (Koss-Chioino, 2000). African American healing practices differ in beliefs, practices, and types of healers based on locale (e.g., north vs. south, and places of special tradition such as the Sea Islands of South Carolina). Overall, the underlying belief system of African Americans is a blend of medical beliefs of an earlier day, African beliefs about voodoo, European folklore regarding the supernatural world, and fundamentalist Christianity (Carter, 2002; Mattis et al., 2007).

Etiological notions are divided into natural and unnatural categories, varying from unhealthy lifestyles and nonproductive worry to beliefs in evil influence or God taking punitive action. Physicians or herbalists can cure the first two causes, but the latter two causes demand intercession of a religious healer or a direct contract with God. Among Latinos, divination may involve a healer's use of shells or tarot cards as a guide to see into the spirit world to identify the cause of the client's problem. The healer describes the client's problems (e.g., somatic distress, bad feelings, and persistent interpersonal or social problems), and then the client confirms (most frequently) or denies their origins. Very little epidemiological data are available on the utilization of ethnomedical treatments by Latinos in the USA (Cervantes, 2010). The existing data, however, indicate that the utilization of divination varies across Latino ethnic populations. Within the USA, divination appears to be a flourishing mental health alternative in Miami, Florida, particularly among the Cuban population, and among some Dominican immigrant populations. Estimates among Mexican Americans vary from 7% in San Diego to 54% in South Texas (see discussion in Koss-Chioino, 2000). It is important to note that some studies separate the belief in traditional healing practices from reports of actual use, so discrepancies may not be so extreme.

Variations are especially salient among Asian Americans (e.g., Chinese, Japanese, Vietnamese, Koreans, and tribal people such as the Hmong or Lao) (Gerdner, 2012). Among the Chinese, shamanic healing (e.g., divination) has not been reported in the USA. The Chinese-American ethnohealers emphasize a balanced and disciplined life to maintain physical and psychology health. In addition, physical/bodily manipulation, acupuncture, massage, and breathing exercises are common. In contrast, some Southeast Asians in the USA see shamans (e.g., folk healers, witch doctors) when they are available (see discussion in Koss-Chioino, 2000). This propensity to visit shamans is motivated by the belief that spirits are omnipresent throughout nature and can cause or heal illness. For example, there occurs among adult Southeast Asians, particularly male Hmong refugees living in the USA, a 'Sudden Death Syndrome.' Autopsies produce no identifiable cause of the death: A person in apparently good health went to sleep and died without awakening. Often, the victim displayed labored breathing, screams, and frantic movements just before death. Prior to their deaths, these men reported that they were experiencing spirit visitations, and that the spirits were unhappy and were punishing them. Furthermore, these deaths do not appear to have a primary biological basis, whereas the deaths do appear to involve psychological factors, including a belief in the imminence of death – either by a curse or some form of punishment (Sue and Sue, 1999).

Western-trained therapists too often dismiss these belief systems and impose their own explanations and treatment upon culturally different clients. Without an understanding of the underlying cultural beliefs, there exists the tendency for the therapist to overestimate the degree of pathology. In the case of Southeast Asian men who report spirit attacks, a Western traditional therapist would most likely diagnose these people as paranoid schizophrenic, suffering from delusions and hallucinations (Sue and Sue, 1999). Such a diagnosis would most likely lead the therapist to prescribe a powerful antipsychotic drug and perhaps institutionalization. Studies indicate, however, that shaman techniques cure these symptoms (Sue and Sue, 1999). The effectiveness of such treatment may lie in the client's belief in the power of the doctor (shaman) and/or the treatment (divination) received.

While the four groups presented here vary in their degree of homogeneity concerning beliefs in spiritual attacks, a substantial number of each group do hold these beliefs. Western medicine does not embrace indigenous or alternative healing approaches and often rejects such approaches as unscientific. Using a culturally inappropriate approach may result in harm or even death (e.g., as in the case of Sudden Death Syndrome). In an attempt to insure cultural responsiveness, the mental

health practitioner must either incorporate the legacy of ancient wisdom that may be contained in indigenous models of healing or refer the culturally different client to a culturally appropriate healer (Sue and Torino, 2005).

The Impact of Ignoring Culture in Therapy

Culture thus impacts both mental health and therapeutic intervention. Specifically, culture influences our understanding of the major stages of the clinical process: patient and practitioner explanatory models of sickness; perception, labeling, and presentation of symptoms; health-seeking behavior; doctor–patient communication; curing and healing; and lay and health professional evaluations of therapeutic outcomes. Culture exerts its most fundamental and far-reaching influence through the categories we employ to interpret and respond to sickness. As previously discussed, different societies and ethnic groups within one society often affix different sickness labels to the same syndrome (Kleinman, 1988). When patient and practitioner explanatory modes reflect different cultural backgrounds, the clinical realities they define can be, and frequently are, quite disruptive to the treatment process. When those explanatory models are in conflict, the results may be client dropout, noncompliance and dissatisfaction, and/or practitioner missed diagnosis, inappropriate treatment, and poor quality of care. All of these consequences exacerbate existing problems and may even create new issues for the client (Smedley et al., 2003).

Although the percentage of ethnic minorities is increasing in the USA, and more interest is being paid to establishing cross-cultural criteria for diagnosing and treating clients (APA, 2013), culturally sensitive training is dismal at best (Woods et al., 2012). Furthermore, the practice of substituting a model stereotype (i.e., the most frequent behaviors of middle-class White males) for the real world, while disregarding cultural variations – in a dogmatic adherence to some universal notion of truth – culminates in a technique-oriented definition of the therapeutic process. As a result, a lack of culturally responsiveness forms of treatment is the single most important reason given for ethnic minorities being underserved, inappropriately served, resulting in a significantly higher drop-out rate (U.S. Department of Health and Human Services, 2001).

Another significant aspect of cross-cultural awareness is the acknowledgment that disadvantaged people often experience oppression, discrimination, and racism, as well as limited access to treatment because of their economic or employment status. A lack of sensitivity on the part of mainstream therapists concerning such issues often results in further abuse of the minority ethnic client (Murray, 1998). On every index, minorities have been found to receive fewer and poorer quality services. Specifically, African Americans and Latinos, in contrast to Whites – holding symptoms constant – were less often given individualized psychotherapy, more often jailed instead of hospitalized, and more often treated with pharmacotherapy instead of talk-therapy (Smedley et al., 2003). Many studies have demonstrated that clinicians, given identical test protocols, tend to make more negative prognostic statements and judgments of greater maladjustment when the individual was said to come from a lower rather than a middle-class background, or was an ethnic minority member (Smedley et al., 2003).

When ethnic minority clients are met with insensitivity or outright racism on the part of the therapist, poorer outcomes are likely to result. Thus, the need for changes in the mental health delivery system is real. For mental health professionals to effectively deliver services to members of a negatively stereotyped group, they must first become aware of their own culture and assumptions about human behavior, values, biases, preconceived notions, personal limitations, and so forth. Second, they must actively attempt to understand the worldview of their culturally different clients in terms of the etiology, coping strategies, and course of mental disorders (Murray, 1998). Cross-cultural sensitivity has been accomplished when the therapist has (1) an understanding of his or her own culture and achieved self-awareness; (2) an understanding of the client's culture; (3) an understanding of the client's relationship to the sociopolitical system; (4) an understanding of the ways in which generic counseling and therapy may limit the potential of the client; (5) an array of therapeutic techniques at his or her disposal; and (6) a willingness to exercise institutional intervention on behalf of the client when appropriate (Sue and Sue, 1999).

Conclusions

In conclusion, culture plays a role in every aspect of mental health, from prevention to diagnosis to treatment. Cultural influences vary greatly from individual to individual. Many mental health problems have culture-common, culture-specific, or both components. In the case of members of historically disadvantaged minority ethnic populations, too often the larger society's role in the developmental etiology and maintenance of certain mental health disturbances is ignored. However, the literature does suggest that racial attitudes that poison our society and are internalized by mental health professionals can hamper the efforts of people of color to maintain or regain their mental health (Smedley et al., 2003). In addition, despite the fact that coping is fundamental to our understanding of human health, mental health professionals often do not acknowledge characteristics of minority cultures that are adaptive and facilitate coping within a stressful environment. Every culture represents a unique solution to the myriad forces imposed on its members. Cross-cultural differences exist in perceptions of mental illness in terms of beliefs about the causes, the prevalence and course, and the experience of symptoms. Within the therapeutic environment, therapist/client communication, diagnosis, intervention strategies and techniques, client use of health facilities, and the course of mental disorders have all been impacted by culture. Therefore, mental health care providers must be trained to identify the role that culture plays in mental illness. The culturally sensitive professional goes beyond the qualities of concern, empathy, and credibility. They have self-knowledge and a willingness and ability to take the cultural background of clients into consideration. Without self-knowledge and cultural sensitivity on the part of the therapist, the mental health needs of ethnic minorities will remain underserved and inappropriately served at best, and at worst ethnic minorities will continue to be

misdiagnosed and victimized under the guise of therapeutic intervention.

See also: Doctor–Patient Interaction in the West: Psychosocial Aspects; Environmental Stress and Health; Healing Practices: Alternative and Complementary; Health Literacy; Health and Illness: Mental Representations in Different Cultures; Illness Behavior and Care-seeking; Race: Ethnicity and Health; Social Epidemiology; Socioeconomic Status and Health.

Bibliography

American Psychiatric Association (APA), 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-V), fifth ed. American Psychiatric Association, Washington, DC.

American Psychiatric Association (APA), 1994. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV), fourth ed. American Psychiatric Association, Washington, DC.

Bussing, R., Gary, F.A., 2012. Eliminating mental health disparities by 2020: everyone's actions matter. Journal of the American Academy of Child and Adolescent Psychiatry 51, 663–666.

Carter, J.H., 2002. Religion/spirituality in African-American culture: an essential aspect of psychiatric care. Journal of the National Medical Association 94, 371–375.

Cervantes, J.M., 2010. Mestizo spirituality: toward an integrated approach to psychotherapy for Latina/OS. Psychotherapy (Chicago, IL) 47, 527–539.

Chen, S.X., Mak, W., 2008. Seeking professional help: etiology beliefs about mental illness across cultures. Journal of Counseling Psychology 55, 442–450.

Choi, N.G., Gonzalez, J.M., 2005. Geriatric mental health clinicians' perceptions of barriers and contributors to retention of older minorities in treatment: an exploratory study. Clinical Gerontologist 28, 3–25.

Conner, K., Copeland, V., Grote, N., Rosen, D., Albert, S., McMurray, M., Reynolds, C., Brown, C., Koeske, G., 2010. Barriers to treatment and culturally endorsed coping strategies among depressed African-American older adults. Aging and Mental Health 14, 971–983.

Draguns, J.G., 2000. Psychopathology and ethnicity. In: Aponte, J.F., Wohl, J. (Eds.), Psychological Intervention and Cultural Diversity. Allyn and Bacon, Boston, MA, pp. 40–58.

Gerdner, L.A., 2012. Shamanism: indications and use by older Hmong Americans with chronic illness. Hmong Studies Journal 13, 1–22.

Jimenez, D., Bartels, S., Cardenas, V., Daliwal, S., Alegría, M., 2012. Cultural beliefs and mental health treatment preferences of ethnically diverse older adult consumers in primary care. The American Journal of Geriatric Psychiatry 20, 533–542.

Kao, H.F., Hsu, M.T., Clark, L., 2004. Conceptualizing and critiquing culture in health research. Journal of Transcultural Nursing 15, 269–277.

Kim, H.S., Sherman, D.K., Ko, D., et al., 2006. Pursuit of comfort and pursuit of harmony: culture, relationships, and social support seeking. Personality and Social Psychology Bulletin 32, 1595–1607.

Kleinman, A., 1988. Rethinking Psychiatry: From Cultural Category to Personal Experiences. Free Press, New York.

Kleinman, A., 22 August 2009. Global mental health: a failure of humanity. Lancet 374 (9690), 603–604.

Kleinman, A., Benson, P., 2006. Culture, moral experience and medicine. Mount Sinai Journal of Medicine 73, 834–839.

Koss-Chioino, J.D., 2000. Traditional and folk approaches among ethnic minorities. In: Aponte, J.F., Wohl, J. (Eds.), Psychological Intervention and Cultural Diversity. Allyn and Bacon, Boston, MA, pp. 149–166.

Leong, F., Kim, H., Gupta, A., 2011. Attitudes toward professional counseling among Asian-American college students: acculturation, conceptions of mental illness, and loss of face. Asian American Journal of Psychology 2, 140–153.

Mallinckrodt, B., Shigeoka, S., Suzuki, L., 2005. Asian and Pacific Island American students' acculturation and etiology beliefs about typical counseling presenting problems. Cultural Diversity and Ethnic Minority Psychology 11, 227–238.

Mattis, J.S., Mitchell, N., Zapata, A., Grayman, N., Taylor, R., Chatters, L., Neighbors, H., 2007. Uses of ministerial support by African Americans: a focus group study. The American Journal of Orthopsychiatry 77, 249–258.

Mezzich, J.E., Kirmayer, L., Kleinman, A., Fabrega Jr., H., Parron, D., Good, B., Lin, K., Manson, S., 1999. The place of culture in DSM-IV. The Journal of Nervous and Mental Disease 187, 457–464.

Murray, C.B., 1998. Racism and mental health. In: Friedman, H.S. (Ed.), Encyclopedia of Mental Health. Academic Press, San Diego, CA, pp. 345–357.

Ratner, C., Hui, L., 2003. Theoretical and methodological problems in cross-cultural psychology. Journal for the Theory of Social Behaviour 33, 67–94.

Smedley, B., Stith, A., Nelson, A. (Eds.), 2003. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Institute of Medicine. National Academies Press, Washington, DC.

Sue, D.W., Sue, D., 1999. Counseling the Culturally Different: Theory and Practice. Wiley, New York.

Sue, S., Cheng, J.K.Y., Saad, C., Chu, J., 2012. Asian American mental health: a call to action. American Psychologist. 67, 532–544.

Sue, D., Torino, G., 2005. Racial-cultural competence: awareness, knowledge, and skills. In: Carter, R.T. (Ed.), Handbook of Racial-Cultural Psychology and Counseling: Training and Practice, vol. 2. John Wiley & Sons, Hoboken, NJ, pp. 3–18.

U.S. Department of Health and Human Services, 2001. Mental Health: Culture, Race, and Ethnicity—A Supplement to Mental Health: A Report of the Surgeon General. http://www.surgeongeneral.gov/library/reports/.

Ward, E.C., Wiltshire, J., Detry, M., Brown, R., 2013. African American men and women's attitude toward mental illness, perceptions of stigma, and preferred coping behaviors. Nursing Research 62, 185–194.

Williams, D., Neighbors, H., Jackson, J., 2003. Racial/ethnic discrimination and health: findings from community studies. American Journal of Public Health 93, 200–208.

Wilson, M.N., Kohn, L.P., Lee, T.S., 2000. Cultural relativistic approach toward ethnic minorities in family therapy. In: Aponte, J.F., Wohl, J. (Eds.), Psychological Intervention and Cultural Diversity. Allyn and Bacon, Boston, MA, pp. 92–109.

Woods, V.D., King, N.J., Hanna, S.M., Murray, C. (Eds.), May 2012. We Ain't Crazy! Just Coping with a Crazy System: Pathways into the Black Population for Eliminating Mental Health Disparities. African American Health Institute of San Bernardino County, San Bernardino, CA. http://www.aahi-sbc.org (California Reducing Disparities Project [CRDP] African American Strategic Planning Workgroup, California Department of Mental Health contract #09-79055-006).

Woodward, A., Taylor, R., Bullard, K., Aranda, M., Lincoln, K., Chatters, L., 2012. Prevalence of lifetime DSM-IV affective disorders among older African Americans, Black Caribbeans, Latinos, Asians and non-Hispanic White people. International Journal of Geriatric Psychiatry 27, 816–827.

Culture, Cognition and Embodiment

Omar Lizardo, University of Notre Dame, Notre Dame, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of relationship between the cultural and the mental in the social and human sciences stands at crossroads. The classical approach to this issue attempts to derive individual cognition from abstract cultural patterns external to the mind, and conceives of culture as an external, emergent order of symbols organized as 'systems.' These internalized 'symbols of mind' provide the essential foundation of individual cognition. In contrast to this view, we have a now growing set of perspectives emerging at the intersection of cognitive science, cultural and cognitive anthropology, developmental psychology, robotics, and the philosophy of mind and action. These 'embodied' approaches to cognition reject the classical view of cognition as the organization of sense data via recourse to conventional cultural symbols. Instead, cognition is seen as tightly linked to practical action and as inherently 'grounded' in the nonarbitrary features of human bodies as they relate to the material environment.

Introduction: Beyond the Classical Theory of Culture and Cognition

The study of relationship between the sociocultural and the mental in the social and human sciences stands at crossroads. On the one hand, there is the continued dominance of what the author refers to as the 'classical' approach to the relationship between the cultural and the mental. The classical approach attempts to derive individual cognition from abstract cultural patterns external to the mind, and conceives of culture as an emergent order of symbols organized as 'systems,' linked to their 'referents' by conventional (arbitrary) correspondence relations (Shore, 1996; Saussure, 1966). These internalized 'symbols of mind' provide the essential foundation of individual cognition. The human animal is a thinking animal insofar as it is a social (and socialized) animal (Geertz, 1973a,b,c; Berger and Luckmann, 1966). This is the stance characteristic of cultural theorists in anthropology and cultural and cognitive sociology starting with Durkheim, and continuing on to Parsons, Geertz, Zerubavel, and Alexander (Parsons, 1972, 1968; Zerubavel, 1999; Alexander, 2003).

On the other hand, we have a now growing (but heterogeneous) set of perspectives emerging at the intersection of cognitive science, cultural and cognitive anthropology, developmental psychology, robotics, and the philosophy of mind and action (Clark, 1997; Bloch, 1998a,c; Shore, 1996; Brooks, 1997; Ingold, 2000; Thelen and Smith, 2002; Lakoff and Johnson, 1999). This set of postclassical perspectives, usually glossed under the banner of 'embodied' approaches (e.g., Varela et al., 1991; Lizardo, 2012; Ignatow, 2009; Garbarini and Adenzato, 2004; Lakoff and Johnson, 1999; Gallese and Lakoff, 2005), reject the classical view of cognition as the organization of sense data via recourse to conventional cultural symbols. Instead, cognition is seen as tightly linked to practical action and as inherently 'grounded' in the nonarbitrary features of human bodies as they relate to the material environment (Bourdieu, 2000; Ingold, 2000; Shore, 1996). In this respect, the recent emphasis on both *embodiment* and *grounding* along with the equally resounding rejection of the view that personal experience is *constituted* via the internalization of external symbols represents a recuperation of a theoretical legacy that was emphatically disavowed in twentieth century (Durkheim-inspired) social and cultural theory in anthropology and sociology (Ingold, 2000). This is an approach to cognition that anthropologist Maurice Bloch (1977, 1986) once referred to (provocatively) as a 'psychological' approach to cognition, in contrast to the then dominant (classical) 'anthropological' approach to the question.

In what follows, the author retains the broad thrust of Bloch's argument, but drops the invidious disciplinary labels. The author does this mainly because the issue is no longer one of choosing between 'psychology' (or cognitive science) and 'sociology' as Durkheim (1982) once suggested and as some contemporary theorists of the culture–cognition link – (e.g., Zerubavel, 1999) – continue to believe. In what follows, the author sticks to the 'classical' and 'embodied' labels to refer to these broad sets of perspectives since allegiance to either viewpoint cuts across all disciplines in the social and human sciences. The basic argument is that classical approaches to the question of the relationship between culture and cognition can no longer afford to ignore the challenge posed by embodied perspectives. More strongly, the author argues that embodied approaches to cognition pose such a strong challenge to the theoretical and conceptual core of the classical approach as to render the latter untenable as a future basis for the study of culture and cognition in the social sciences.

The rest of the article is organized as follows: first the basic claims, presuppositions, and parameters of the classical approach are outlined (in Section The Classical Theory of Culture and Cognition). Then the author goes on (in Sections The (Cognitive) Problem of Internalization and The (Cognitive) Problem of Meaning) to show that this approach runs into trouble (in relation to embodied approaches) on two fundamental issues in the study of culture and cognition: what he refers to as the (cognitive) problem of the internalization of culture and the problem of meaning (how persons use cultural symbols for representational purposes). The article closes (Section Conclusion) by outlining the broader implications of the argument.

The Classical Theory of Culture and Cognition

According to the classical view, culture is composed of abstract symbolic representations that require some sort of connection to arbitrary carriers of meaning in order to signify (i.e., to be meaningful). Proponents of this model presuppose that perceptual information that is gathered directly through the senses is 'meaningless' because it is composed of preconceptual, 'raw' stimuli that require processing, transduction, schematization, and linkage to symbolic structures in order to become meaningful. Traditional cultural theory follows Weber (1946), Berger and Luckmann (1966), and Geertz (1973d) in portraying actors as "…*obsessed with sorting out empirical reality* and, typifying from code to event" (Alexander, 2003: p. 124, italics added). Meaning comes only from the collective symbolic categories learnable through the culture. According to this perspective, culture consists of a conventional corpus of symbolic structures (which, following Saussure (1966), are conceptualized as the arbitrary coupling of amodal mental representations) with a material carrier of meaning. These representations organize sensory experience and thus partition the world into arbitrary categories of persons, things, and events (Leach, 1976; Douglas, 1986).

The classical theory follows the Kant-inspired dictum that primary experience is composed of lower order sensory information that carries unique, idiosyncratic, and evanescent content. In order for this sensory content to become generic and durable (and thus part of the cultural repertoire of the social agent), it must be linked to more permanent, mnemonically available symbolic systems (D'Andrade, 1992). That is, sensory events must be 'typified' (e.g., persons, objects, and event tokens must be seen as singular representative of a larger class of entities) in order to be meaningful to the person (Parsons and Shils, 1951; Berger and Luckmann, 1966). Accordingly, persons require symbolic cultural structures in order to generalize across sensory events and to characterize the most relevant features of experience so that it can be available for future use. Under this view, symbolic structures organize experience, but sensory information cannot be stored and encoded in the same format as it is encountered in the world. Instead, sensory input must be 'transduced' (or converted) into a different (symbolic) format distinct in nature from the 'raw' form that constitutes our phenomenological experience (Barsalou, 1999; Ignatow, 2009). The classical theory thus requires that the social agent be at the mercy of culturally acquired symbolic structures in order to classify different sorts of sensory stimulation as belonging to the same category (Durkheim, 1995; Durkheim and Mauss, 1967; Zerubavel, 1999; Douglas, 1966).

The classical theory of culture and cognition implicitly accepts without question the traditional philosophical division of mental contents into those that are directly traceable to the senses and those more abstract representations that have at best an indirect connection to experience (ideas) that we inherit from the rationalist philosophical tradition. The classic philosophical debate deals with the possible existence of mental representations that are not traceable to experience (e.g., Kantian a priori categories). This contrasts with the empiricist (from Hobbes to Locke to Hume) argument that *all* mental contents are ultimately traceable to sensory stimulation and what were called ideas were simply weakened, schematic remnants ('images') of what once was a sensory event. Kant famously argued against British empiricism (and the Humean skepticism regarding the existence of preexperiential knowledge structures) that there had to be an analytical (and ontological) hiatus between abstract ideas (categories) and concrete sensory stimulation. The reason for this is that sensory stimulation in itself was insufficient to produce generalizable concepts with universal validity; empiricism results in irrationalism. The fact that such concepts existed thus required an appeal to a transcendental argument for the existence of knowledge structures that do not come from experience. Thus, the very conditions of possibility of meaningful knowledge required an argument to the best explanation, which clinched the necessary existence of preexperiential abstract conceptual structures that brought order and organization to perceptual experience.

In this respect, the classical sociological theory of culture and cognition can be thought of as a 'sociologization' – which we owe to Durkheim (1995) – of the original Kantian proposal, in which the notion of the a priori is dropped and the categories are given an empirical etiology in society. Durkheim accepted the perception versus high-level cognition binary but deconstructed it by moving the debate away from the realm of abstract epistemological argumentation and remapping the cognition/perception dichotomy to the distinction between the society and the individual (Durkheim, 2005). Even if most people would find the latter assertion questionable (and the Durkhemian argument was immediately assailed as question begging), the Kantian analytical division between sensory stimulation and abstract conceptual structure is kept intact in the modern theory of culture and cognition. The same goes for the Kantian presumption that direct sensory stimulation is inherently meaningless, noncategorical, and nongeneralizable unless it is fitted into a previously existing high-level system of abstract symbolic concepts (Schmaus, 2004).

According to Durkheim, sensory stimulation (and practical action in the world of objects) is evanescent and it is not sufficient for the formation of durable mental categories that generalize from experience. Here, Durkheim rejects both the classical empiricist and the then emerging 'pragmatist' argument for the emergence of general ideas from individual action all the same. Only particular forms of structured interaction in collective ritual occasions or (in differentiated societies) direct instruction from an authoritative representative of such a collectivity could instill in the mind durable categories of the understanding that had that property of generalizing or abstracting from experience. These categories were able to some extent to go beyond evanescent individual experience and had the ability to generate authoritative consent. For Durkheim, the *validity* of the categories was inseparable from their *cognitive authority*. Durkheim equated 'religious phenomena' with those mental contents that were not of individual but of collective origin and that were not open to revision in the light of further experience. Thus Durkheim linked the concept/percept, obligatory/elective, collective/individual, and sacred/profane dichotomies, with

the first and second pair of each opposition corresponding to one another (Durkheim, 1975).

Parsons (1972) accepted this theory of the origins of general ideas, reasoning that these are the components of a *cultural system* that individuals have to internalize in order to learn to typify experience. Parsons' only tweak on the Durkheimian account came in providing a more detailed social mechanism for the internalization of cultural contents. Parsons proposed that this internalization is realized via socialization in the family environment characteristic of differentiated societies that ensure the transmission of the overarching schemes of categorization that make up the culture. Parsons adapted the Freudian notion of 'introjection' to do the requisite explanatory job, but recast Freudianism as an anthropological theory of cognitive socialization. Parsons castigated Freud for proposing a 'simplistic' theory of the reality principle, which presumed that persons simply formed veridical representations of external reality without necessarily having recourse to a socially acquired system of cultural categories. According to Parsons, rather than taking this 'naïve' a-sociological stance, abstract, general representations of the objective environment had to be mediated by a system of collective categories that persons internalized via a learning process. According to Parsons (1964: p. 23), Freud "failed to take explicitly into account the fact that the frame of reference in terms of which objects are cognized, and therefore adapted to, is *cultural* and thus cannot be taken for granted as given but must be *internalized* as a condition of the development of mature ego-functioning" (italics added). Cultural sociologists today might reject the outdated mid-twentieth century psychoanalytic language, but they retain the notion that persons somehow come to internalize culture conceived as "a system of *generalized* symbols and their meanings" (Parsons, 1964: p. 29, italics added).

The (Cognitive) Problem of Internalization

Given this theoretical legacy, the classical theory of culture and cognition provides no way of conceptualizing the enculturation process other than through the metaphorical imagery of the (format-preserving) 'transfer' of objectified cultural contents (e.g., values, beliefs, propositions) *from* the *external* environment *into* the 'internal' environment of the person. Enculturation is thus (naturally) conceptualized as a learning process (Wrong, 1961) and learning as the re-creation of a copy of public culture within the cognitive environment of the actor. Traditional cultural theory thus presumes that persons become encultured subjects through a process of *internalization of objectified cultural contents* and that this internalization requires linguistic 'transmission' of these contents to be effected (Berger and Luckmann, 1966; Parsons, 1964). The major theoretical schools in cultural and cognitive sociology today differ only on the presumed nature of what this content is – whether it is 'values,' 'conceptual schemes,' 'classifications,' or 'codes' (Parsons, 1951, 1964; Zerubavel, 1993; Alexander, 2003). The *content* internalization requirement is never subject to question: all theorists continue to hold on to some modified version of the presumption that the final product of this process is the storage of some (systematic?) set of lingua form representations of external culture.

The basic idea is that in order for an agent to be socialized, something external, or social has to be converted into something internal or personal (on this score, see Wrong, 1961; Parsons, 1964; Bush and Simmons, 1981; Cook-Gumperz, 1973). We can then explain systematic cross-individual and cross-group variations in values, practices, and beliefs as well as the systematic *patterning* of individual behavior within a given society over time, as a result of the differential transmission and acquisition of the relevant cultural and symbolic endowment that characterizes a given collectivity (Cook-Gumperz, 1973: p. 3). Accordingly, one of the key meanings of socialization is essentially, "the [social] transmission of culture" (Wrong, 1961: p. 192; Long and Hadden, 1985: p. 42). This process of socially mediated cultural transmission is conceived as implying the mental modification of the agent who is at the receiving end of the cultural conduit mechanism (Parsons, 1964: p. 23). The production of a set of individuals who share cultural patterns through socialization (i.e., cultural transmission) processes can then be used to answer the question of the origins of systematic patterns of action in social collectivities.

Most American sociologists have recognized – ever since the influential critique of the functionalist penchant to equate socialization with the 'internalization' of values – that there are important limitations to this formulation of the socialization process (e.g., Heritage, 1984; Swidler, 1986). While contemporary sociologists are quick to reject the Parsonian value-internalization account (thus abandoning Parsons account of the emergence of social order via the production of value consensus), they continue to abide by the Parsonian model of *cognitive* socialization. In other words, most sociologists continue to believe that people share cultural contents (e.g., worldviews and beliefs) because they internalize those contents from the larger culture. As we saw above, for Parsons, *both* 'values' and the broader 'conceptual schemes' through which social actors come to know and classify the entire world of objects, agents, and situations have to be internalized (Parsons and Shils, 1951; Parsons, 1964). Any theory that presupposes that persons introject the basic categories with which they make sense of the world from the external environment is still essentially a 'Parsonian' theory of enculturation even if the adjective Parsonian has come to (wrongly) be limited to the 'value-internalization' account.

Accordingly, the Parsonian theory of culture and cognition is (discouragingly) hard to distinguish from contemporary approaches, especially in presuming the wholesale internalization of entire conceptual schemes by socialized actors (Long and Hadden, 1985). That there is something deeply wrong with this model of the process of socialization has been noted repeatedly, even as something pretty close to this model is still implicitly assumed in a lot of contemporary empirical research, especially in cultural and cognitive sociology. For instance, Alexander chides postfunctionalist conflict theory for failing to emphasize "…the power of the symbolic to shape interactions from within, as normative precepts or narratives that carry *internalized* moral force" (Alexander, 2003: p. 16; italics added; see also pp. 152–153 of the same book on the internalization of cultural codes). For his part, Eviatar Zerubavel notes that

> [t]he logic of classification is something we must learn. Socialization involves learning not only a society's norms but also its distinctive classificatory schemas. Being socialized or acculturated entails knowing not only how to behave, but also how to perceive reality in a socially appropriate way. By the time she is three, a child has already internalized the conventional outlines of the category 'birthday present' enough to know that, if someone suggests that the she bring lima beans as a present he must be kidding. (Zerubavel, 1999: p. 77)

Through what specific mechanisms 'the symbolic' or 'logics of classification' are actually learned or internalized by the social agent is never actually specified (Turner, 1994). Lizardo (2012) draws on embodied cognition approaches to propose a solution to the problem of the internalization of culture that rejects the representationalism of the classical account.

The (Cognitive) Problem of Meaning

The continued dominance of a conceptualization of culture as a 'superorganic' system of symbols, coupled with a conceptualization of cognitive socialization as the 'internalization' of those meanings, is inseparable from a view of internal experience as itself composed of symbols. Accordingly, a foundational substantive claim of the classical theory of the culture–cognition link is that cultural symbols either structure or constitute mental experience (Berger and Luckmann, 1966; Geertz, 1973a). From this perspective, humans are symbol-using animals because without symbols they would not be able to *make sense* of their experience. This links with the claim – noted above – that perceptual states are an incoherent, structureless flux that must be fitted into a grid of cultural symbols to be made meaningful (Leach, 1964; Martin, 2011). Stronger versions of this approach propose the substantive hypothesis that mental contents, insofar as they are the product of some form of learning or enculturation (conceptualized as the 'internalization' of external symbols), are themselves symbolic. That is, just as symbols exist 'in the world,' they exist (in similar format) as 'symbols of mind' (Kolers and Roediger, 1984).

Embodied approaches to the same set of issues summarily break with these well-entrenched tenets of classical cultural theory. Two basic observations deserve to be highlighted in this respect. First, from the perspective of embodied approaches to cognition, it is incoherent to suppose that mental experience can be 'symbolic' in the same sense that external cultural symbols are symbolic (Kolers and Smythe, 1984). Instead, there is a conceptualization of mental experience as inherently grounded in perception and action (Barsalou, 1999; Johnson, 1987). In this sense, the 'components' of mental experience do not and cannot share the 'bipolar' structure of cultural symbols, which must couple external vehicles to semantic content in order to be meaningful. The phenomenological contents of our mental experience, on the other hand, are *directly* meaningful (Johnson, 1987; Shore, 1996); they do not require symbolic 'transduction' to be made so (Barsalou, 1999). Precisely because external cultural symbols are bipolar couplings of form and meaning, the mind cannot be thought of as a 'repository' of symbols as naive models of learning and memory implicitly presuppose (Roediger, 1984). Second, it is also incoherent to suppose that perceptual states lack structure or form an incoherent flux waiting for the order provided by linguistic or cultural categories (Gibson, 1979; Merleau-Ponty, 1962). Embodied approaches to language, semantics, and meaning construction show that abstract cultural categories are grounded on image-schematic conceptualizations extracted from perception and action, and not the other way around (Bloch, 1991, 1986; Johnson, 1987).

Proponents of embodied approaches to the culture–cognition link propose that the meaningful component of mental experience is primarily made up of repeated, embodied *simulations* over previous experiences. These simulations are enacted using neural structures that support the same sensory modality that made possible the experience in the first place (Barsalou, 1999). Consistent with so-called dual process models of learning and memory (Smith and DeCoster, 2000; McClelland et al., 1995), embodied cognition theorists propose that these simulations can reconstitute both the rich content of previous experiences as well as its most cross-modally recurrent features, which schematize over that content. Schematization implies the structured 'forgetting' of irrelevant or context-specific detail and the preferential retention only of those features that are correlated across instances (McClelland and Rumelhart, 1985). Accordingly, the conceptual potential accessed by cultural symbols covers both rich experiential content as well as memory for higher order correlations across experiences organized in the form of sensory schemas.

Embodied cognition theorists reject the influential proposal that recourse to linguistic symbols is a necessary component of meaningful experience (Bloch, 1998b). Even if persons did not have recourse to linguistic and cultural symbols, experience would be meaningful in this embodied sense, as it was for our prelinguistic ancestors. In contrast to some well-entrenched postulates of cultural theory, cultural symbols do not exist because they are required to produce meaningful mental experiences. Even if linguistic symbols did not exist, persons would still have a subjective – and even an intersubjective (Gallese, 2003) – grasp of the world as a meaningful structure. This grasp is based on our fundamental capacity to interact in the world and with our conspecifics at a basic embodied level and to extract the basic contours of experience from repeated exposure (Ingold, 2000). It is important to keep in mind that the 'perceptual symbols' constitutive of mental experience are not equivalent to the (external) 'cultural symbols' prominent in the classical theory. In embodied approaches to cognition, the use of cultural symbols is seen as the *exploitation* of *external* forms for purposes of evoking or accessing the rich semantic potential of the perceptual symbols constitutive of mental experience (Shore, 1996; Clark, 1997).

If cultural symbols neither constitute experience nor are they required (in the strong sense postulated by classical cultural theory) to 'make sense' of this experience, then what is their primary function? From an embodied cognition perspective, cultural symbols rather than constituting our subjective semantic potential *serve to evoke or access this preexisting potential in patterned ways*. The semantic pole of a cultural symbol is typically linked to a set of perceptual symbols in their

role as the *simulators* for the concept(s) that constitutes the 'cognitive meaning' of the symbol (Barsalou, 2009, 2003, 1999; Strauss and Quinn, 1997). This *evocation model* of the way that cultural symbols operate contrasts sharply with both the 'constitutive' model inherited from structuralism and with the 'sense-making' model inherited from phenomenology (Bloch, 2006; Shore, 1996).

In contrast to the classical theory, which privileges conceptual structures over primary experience, in embodied cognition approaches experience is primary and 'abstract' concepts must be grounded in bodily experience (Johnson, 1987; Lakoff and Johnson, 1999). The basic presumption is that meaning is not 'symbolic' in the strong sense; instead, the potential to invoke a given semantic content must preexist the establishment of a link between an external symbol and that particular sort of mental experience. In this respect, cultural symbols do not 'mean' anything directly; instead, they serve as prompts for persons to engage in (individual or collective) acts of *meaning construction* (Shore, 1996). The semantic potential of cultural symbols thus depends on the meaning construction potential embodied in persons. This meaning construction potential is irreducibly experiential (Bourdieu, 1990; Ingold, 2000); it does not emerge from the 'internalization' of cultural symbols. Instead, persons use large compendia of (domain general) imaginative, cognitive, affective, and motor-schematic *capacities* to create meaning 'online' (Langacker, 1987; Fauconnier and Turner, 2003). Without the meaning construction potential generated by experience, cultural symbols do not mean much. Rather than providing the person with the capacity to create meaning cultural symbols *afford access* to richly detailed, multimodal forms of mental experience in context.

From the point of view of embodied approaches, the external cultural symbols of the classical theory neither serve as access points to conceptual content nor do they serve exclusively as prompts to initiate acts of meaning construction. In addition to these functions, external cultural symbols also serve as a way to usefully delimit and condense the sort of cognitive meaning that is thereby accessed, while at the same time serving as external coordinators of both evocation and meaning construction episodes. The reason for this is that for embodied cognition theorists, cultural knowledge (as embodied in persons) is not symbolic, amodal, or propositional in the usual sense and therefore not partitionable into delimited units of chunks. Instead, this knowledge is perceptually grounded and continuous, with one packet of knowledge blending onto others. This is what Clark (1987, 1997) has referred to as *superpositional storage*. As such, cognitive meaning as embodied in person has an 'associational' and not a 'rule-based' structure (Sloman, 1996), with one packet of meaning serving as an evocation point for a wide array of associated conceptual content. This implies that the orderly, structured evocation of meaning (as, e.g., observed in highly institutionalized domains) must rely on some sort of external support or 'scaffolding' (Clark, 1997; Lizardo and Strand, 2010), and that support is provided by cultural symbols. Cultural symbols coordinate access to the *delimited* chunks of conceptual knowledge in action, in a way that would be difficult (if not impossible) in their absence. This is particularly important in the case of ritual occasions, where conceptual knowledge must be evoked in time in the right sequence and in the prescribed time and context (Durkheim, 1995).

Conclusion

From the perspective of embodied approaches to cognition, it is becoming increasingly clear that cultural theorists who depart from the classical perspective have mischaracterized the nature, purpose, and structure of human cognition and have thus given misleading accounts of the effects of socially acquired cultural categories on individual experience. Because of this, cultural and cognitive sociology lacks a credible (or coherent) account of what is it that 'culture' is supposed to do in order to influence individual cognition. One major source of incoherence (and embarrassment) for the classical account is that it requires that the person be subject to a massive, systematic process of socialization into the symbolic system that makes up the culture.

Yet, anthropologists became dissatisfied with the traditional culture concept, precisely because as much as they looked, such chronic exposure to socializing institutions and such repeated attempts on the part of socializing agents to impart the precepts that made up the culture were nowhere to be found. Yet persons somehow ended up becoming functioning members of their respective collectivities, while sharing explicit and implicit knowledge structures (Bourdieu, 1990; Palsson, 1994; Toren, 1999; Ingold, 2000). This means that persons somehow acquire culture, but without having recourse to the usual mechanisms of cultural transmission emphasized in the classical account. This phenomenon is similar to the 'poverty of stimulus' situation that Chomsky first noticed was the main weakness of the mid-twentieth century behaviorist account of language acquisition. In order for a coherent theory of the relationship between culture and cognition to get off the ground, the conception of what the raw material of experience is, and by implication both the conception of what 'culture' is and thus of how society manages to constrain thought and action, has to be radically altered.

See also: Culture and Institutional Logics; Embodied Social Cognition; Embodiment Theory; Embodiment and Culture; Human Cognition, Evolution of; Socialization in Infancy and Childhood; Visual Perception, Neural Basis of.

Bibliography

Alexander, J.C., 2003. The Meanings of Social Life: A Cultural Sociology. Oxford University Press, New York.

Barsalou, L.W., 1999. Perceptual symbol systems. Behavioral and Brain Sciences 22, 577–660.

Barsalou, L.W., 2003. Abstraction in perceptual symbol systems. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 358 (1435), 1177–1187.

Barsalou, L.W., 2009. Simulation, situated conceptualization, and prediction. Philosophical Transactions of the Royal Society B: Biological Sciences 364 (1521), 1281–1289.

Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality. Doubleday, New York.

Bloch, M., 1977. The past and the present in the present. Man 12, 278–292.

Bloch, M., 1986. From cognition to ideology. In: Fardon, R. (Ed.), Knowledge and Power: Anthropological and Sociological Approaches. Scottish University Press, Edinburgh, pp. 21–48.
Bloch, M., 1991. Language, anthropology and cognitive science. Man 26, 183–198.
Bloch, M., 1998a. Cognition and ethnography. In: How We Think They Think: Anthropological Approaches to Cognition, Memory and Literacy. Westview Press, Boulder, CO, pp. 39–53.
Bloch, M., 1998b. Language, anthropology and cognitive science. In: How We Think They Think: Anthropological Approaches to Cognition, Memory and Literacy. Westview Press, Boulder, CO, pp. 3–21.
Bloch, M., 1998c. What goes without saying: the conceptualization of Zafimaniry society. In: How We Think They Think: Anthropological Approaches to Cognition, Memory and Literacy. Westview Press, Boulder, CO, pp. 22–38.
Bloch, M., 2006. Teknonymy and the evocation of the 'social' among the Zafimaniry of Madagascar. In: vom Bruck, G., Bodenhorn, B. (Eds.), An Anthropology of Names and Naming. Cambridge University Press, Cambridge, UK, pp. 97–114.
Bourdieu, P., 1990. The Logic of Practice. Polity Press, Cambridge.
Bourdieu, P., 2000. Pascalian Meditations. Stanford University Press, Stanford.
Brooks, R.A., 1997. Intelligence without representation. In: Haugeland, J. (Ed.), Mind Design II. MIT Press, Cambridge, MA.
Bush, D., Simmons, R., 1981. Socialization processes over the life course. In: Rosenberg, M., Turner, R. (Eds.), Social Psychology: Sociological Perspectives. Basic Books, New York, pp. 133–164.
Clark, A., 1997. Being There: Putting Brain, Body, and World Together Again. MIT Press, Cambridge, MA.
Cook-Gumperz, J., 1973. Social Control and Socialization: A Study of Class Differences in the Language of Maternal Control. Routledge/Thoemms Press.
D'Andrade, R., 1992. Schemas and motivation. In: Human Motives and Cultural Models. Cambridge University Press, New York, pp. 23–44.
Douglas, M., 1966. Purity and Danger: An Analysis of Concepts of Purity and Taboo. Routlege & Kegan Paul, London.
Douglas, M., 1986. How Institutions Think. Syracuse University Press, Syracuse, NY.
Durkheim, E., 1975. Concerning the definition of religious phenomena. In: Pickering, W.S.F. (Ed.), Durkheim on Religion: A Selection of Readings with Bibliographies. Routledge & Kegan Paul, Boston, MA, pp. 74–99.
Durkheim, E., 1982. The Rules of the Sociological Method. Free Press, New York.
Durkheim, E., 1995. The Elementary Forms of the Religious Life. Free Press, New York.
Durkheim, E., 2005. The dualism of human nature and its social conditions. Durkheimian Studies 11 (1), 35–45.
Durkheim, E., Mauss, M., 1967. Primitive Classification. University of Chicago Press, Chicago.
Fauconnier, G., Turner, M., 2003. The Way We Think: Conceptual Blending and the Mind's Hidden Complexities. Basic Books, New York.
Gallese, V., 2003. The manifold nature of interpersonal relations: the quest for a common mechanism. Philosophical Transactions of the Royal Society of London B 358, 517–528.
Gallese, V., Lakoff, G., 2005. The brain's concepts: the role of the sensory-motor system in conceptual knowledge. Cognitive Neuropsychology 22 (3), 455–479.
Garbarini, F., Adenzato, M., 2004. At the root of embodied cognition: cognitive science meets neurophysiology. Brain and Cognition 56, 100–106.
Geertz, C., 1973a. The growth of culture and the evolution of mind. In: Geertz, C. (Ed.), The Interpretation of Cultures. Basic Books, New York, pp. 55–83.
Geertz, C., 1973b. The impact of the concept of culture on the concept of man. In: Geertz, C. (Ed.), The Interpretation of Cultures. Basic Books, New York, pp. 32–54.
Geertz, C., 1973c. Person, time, and conduct in Bali. In: Geertz, C. (Ed.), The Interpretation of Cultures. Basic Books, New York, pp. 360–411.
Geertz, C., 1973d. Thick description: towards an interpretive theory of culture. In: Geertz, C. (Ed.), The Interpretation of Cultures. Basic Books, New York, pp. 3–30.
Gibson, J., 1979. The Ecological Approach to Perception. Houghton Mifflin, Boston.
Heritage, J., 1984. Garfinkel and Ethnomethodology. Polity Press, Cambridge.
Ignatow, G., 2009. Culture and embodied cognition: moral discourses in Internet support groups for overeaters. Social Forces 88 (2), 643–669.
Ingold, T., 2000. The Perception of the Environment: Essays on Livelihood, Dwelling and Skill. Routledge, London.
Johnson, M., 1987. The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason. University of Chicago Press, Chicago.
Kolers, P., Roediger, H.L., 1984. Procedures of mind. Journal of Verbal Learning and Verbal Behavior 23 (4), 425–449.
Kolers, P., Smythe, W., 1984. Symbol manipulation: alternatives to the computational view of mind. Journal of Verbal Learning and Verbal Behavior 23 (3), 289–314.
Lakoff, G., Johnson, M., 1999. Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. Basic Books, New York.
Langacker, R.W., 1987. Foundations of Cognitive Grammar. In: Theoretical Prerequisites, vol. I. Stanford University Press, Stanford, CA.
Leach, E., 1964. Anthropological aspects of language: animal categories and verbal abuse. In: Lenneberg, E.H. (Ed.), New Directions in the Study of Language. MIT Press, Cambridge, pp. 23–63.
Leach, E.R., 1976. Culture & Communication. Cambridge University Press, New York.
Lizardo, O., 2012. Embodied culture as procedure: cognitive science and the link between subjective and objective culture. In: Warde, A., Southerton, D. (Eds.), The Habits of Consumption, COLLeGIUM: Studies across Disciplines in the Humanities and Social Sciences, vol. 12. Helsinki Collegium of Advanced Studies, Helsinki, Finland, pp. 70–86.
Lizardo, O., Strand, M., 2010. Skills, toolkits, contexts and institutions: clarifying the relationship between different approaches to cognition in cultural sociology. Poetics 38 (2), 205–228.
Long, T., Hadden, J., 1985. A reconception of socialization. Sociological Theory 3 (1), 39–49.
Martin, J.L., 2011. The Explanation of Social Action. Oxford University Press, New York.
McClelland, J., McNaughton, B., O'Reilly, R., 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychological Review 102, 419–457.
McClelland, J., Rumelhart, D., 1985. Distributed memory and the representation of general and specific information. Journal of Experimental Psychology: General 114 (2), 159–188.
Merleau-Ponty, M., 1962. Phenomenology of Perception. Routledge, London.
Parsons, T., 1951. The Social System. Free Press, New York.
Parsons, T., 1964. The superego and the theory of social systems. In: Social Structure and Personality. Free Press, New York, pp. 17–33.
Parsons, T., 1968. Cooley and the problem of internalization. In: Reiss, A.J. (Ed.), Cooley and Sociological Analysis. University of Michigan Press, Ann Arbor, MI, pp. 48–67.
Parsons, T., 1972. Culture and social systems revisited. Social Science Quarterly 53 (2), 253–266.
Parsons, T.E., Shils, E.A., 1951. Toward a General Theory of Action. Cambridge University Press, Cambridge.
Roediger, H., 1984. Does current evidence from dissociation experiments favor the episodic/semantic distinction? Behavioral and Brain Sciences 7, 252–254.
Saussure, F., 1966. Course in General Linguistics. McGraw-Hill, New York.
Schmaus, W., 2004. Rethinking Durkheim and His Tradition. Cambridge University Press, New York.
Shore, B., 1996. Culture in Mind: Cognition, Culture, and the Problem of Meaning. Oxford University Press, New York.
Sloman, S.A., 1996. The empirical case for two systems of reasoning. Psychological Bulletin 119, 3–22.
Smith, E., DeCoster, J., 2000. Dual-process models in social and cognitive psychology: conceptual integration and links to underlying memory systems. Personality and Social Psychology Review 4 (2), 108–131.
Swidler, A., 1986. Culture in action: symbols and strategies. American Sociological Review 51, 273–286.
Thelen, E., Smith, L.B., 2002. A Dynamic Systems Approach to the Development of Cognition and Action. The MIT Press, Cambridge.
Turner, S.P., 1994. The Social Theory of Practices. University of Chicago Press, Chicago.
Varela, F., Thompson, E., Rosch, E., 1991. The Embodied Mind: Cognitive Science and Human Experience. The MIT Press, Cambridge, MA.
Weber, M., 1946. The social psychology of the world religions. In: Mills, H.H.G., Wright, C. (Eds.), From Max Weber: Essays in Sociology. Oxford University Press, New York, pp. 267–301.
Wrong, D.H., 1961. The oversocialized conception of man in modern sociology. American Sociological Review 26, 183–193.
Zerubavel, E., 1993. The Fine Line: Making Distinctions in Everyday Life. University of Chicago Press, Chicago.
Zerubavel, E., 1999. Social Mindscapes: An Invitation to Cognitive Sociology. Harvard University Press, Cambridge, MA.

Culture: Contemporary Views

Richard A Shweder and Les Beldo, University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.A. Shweder, volume 5, pp. 3151–3158, © 2001, Elsevier Ltd.

Abstract

In 1952 A.L. Kroeber and Clyde Kluckhohn formulated a definition of 'culture' that became a mantra for a generation of cultural anthropologists who came of scholarly age in the middle of the twentieth century. Little did Kroeber and Kluckhohn know that for the next 50 years the idea of 'culture,' in its anthropological sense, would be frequently debated, doubted, distrusted, and scorned and associated with a variety of sins. Nor could they have anticipated that at the beginning of the twenty-first century the idea of 'culture' would be a key concept in many of the social sciences, while cultural anthropology would remain a scene for various kinds of 'anticultural' or 'postcultural' critiques. Nevertheless, a concept of 'culture' very much like the one recommended by Kroeber and Kluckhohn remains useful in social science research today. The concept of 'culture' not only survives; it thrives, and for good reason.

Introduction: Kroeber's and Kluckhohn's Prediction

"…few intellectuals will challenge the statement that the idea of culture, in the technical anthropological sense, is one of the key notions in contemporary American thought." That prediction was made in 1952 by the American anthropologists A.L. Kroeber and Clyde Kluckhohn in the introduction to their monumental book *Culture: A Critical Review of Concepts and Definitions* (1952). Kroeber and Kluckhohn even began their famous treatise proclaiming the idea of culture comparable in explanatory importance to the idea of gravity in physics, disease in medicine, and evolution in biology. They ended by adducing a unified (albeit ponderous) definition that became the mantra for cultural anthropologists who came of scholarly age in mid-century. "Culture," they wrote, "consists of patterns, explicit and implicit, of and for behavior acquired and transmitted by symbols, constituting the distinctive achievement of human groups, including their embodiments in artifacts; the essential core of culture consists of traditional (i.e., historically derived and selected) ideas and especially their attached values; culture systems may, on the one hand, be considered as products of action, on the other hand as conditioning elements of further action" (1952, p. 357).

Kroeber and Kluckhohn were students of intellectual history and brilliant culture theorists, but they were not prophets. Little did they know that during the 50 years following the publication of their book the idea of 'culture,' in its mid-century anthropological sense, would be frequently debated, doubted, distrusted, and scorned, that the discipline of cultural anthropology itself would be 'rethought,' 'remade,' 'recaptured,' and 'reinvented' time and time again. They did not foretell the many types of humanists and social scientists (cognitive revolutionaries, structuralists, poststructuralists, sociobiologists, feminists, skeptical postmodernists, post-colonialists, subalterns, globalization theorists) who would associate the concept of 'culture' with a variety of supposed sins. Sins such as 'essentialism,' 'primordialism,' 'representationalism,' 'monumentalism,' 'reification,' 'idealism,' 'positivism,' 'functionalism,' 'relativism,' 'sexism,' 'racism,' 'ethnic conflict,' 'colonialism,' 'Orientalism,' and just plain old-fashioned stereotyping (see, e.g., Abu-Lughod, 1991; Asad, 1973; Borofsky, 1994; Clifford and Marcus, 1986; Denzin, 1996; Fox, 1991; Freeman, 1983; Hymes, 1972; Kuper, 1999; Marcus and Fischer, 1986; Rabinow, 1983; Reyna, 1994; Rosaldo, 1989; Said, 1978; Sangren, 1988; Scheper-Hughes, 1995; Spiro, 1986; Wikan, 1995).

Nor did Kroeber and Kluckhohn anticipate the ironic fate of the concept of 'culture' at the beginning of the twenty-first century. The irony is that today the idea of culture is once again a key concept in many of the social science disciplines, yet is viewed with great suspicion in some quarters of cultural anthropology. The irony is that after being reviled, pummeled, and rejected by one new wave intellectual movement after another an idea of 'culture' very much like the one recommended by Kroeber and Kluckhohn in 1952 remains useful and defensible in social science research and public policy debates. The concept not only survives, it thrives (see for example, Harrison and Huntington, 2000; Huntington, 1996; Landes, 1998; Prentice and Miller, 1999; Kitayama and Markus, 1994; Wierzbicka, 1997).

It is noteworthy that even in Kroeber and Kluckhohn's time anthropologists outside of the American tradition were skeptical of the concept of culture. The British social anthropologist A.R. Radcliffe-Brown once remarked that he found 'culture' to be a 'vague abstraction' not particularly useful in the study of social life. Continuing in that tradition Adam Kuper has argued that institutions and other 'elements' of society should be studied separately rather than "bound together into a single bundle labeled culture" (Kuper, 1999). But many social anthropologists, including some who have deeply criticized the culture concept or its uses, have nonetheless recognized the importance of attending to the 'ideas and values' or 'cultural content' associated with institutionalized social relationships (see e.g., Beattie, 1964; Malinowski, 1944). Attention to 'cultural' kinds of knowledge and processes allows one to comprehend certain prevalent ideas or values as more than just the particular products of particular institutions.

In this latter sense, 'culture' is not regarded as equivalent to or parallel to other dimensions of social life like economics or politics but as something more basic and pervasive – as a set of

understandings that gives sense and meaning even to, say, acts which might on the surface appear entirely economic or political in character (see Geertz, 1973; Peacock, 2005, p. 54, Rosaldo, 1989; Sahlins, 1976; Sewell, 2005). Roy D'Andrade, for example, has discussed the concept of 'institutionalized values': the values (such as obedience, in-group loyalty, and patriotism) that people in a society agree are important in enacting some role (such as a soldier in the army) and are consensually used to evaluate specific role performance (D'Andrade, 2008). D'Andrade points out that cultural differences seem greatest when one takes 'institutional values' as the unit of analysis in cross-cultural investigations.

The contemporary discipline of anthropology continues to be a scene for various kinds of 'anticultural' or 'postcultural' critiques. Nevertheless, many social scientists rehearse and recite some definition of culture and make good use of it in their scholarship (see, e.g., Appadurai, 1996; Comaroff and Comaroff, 1991; D'Andrade, 1984; Dumont, 1970; Geertz, 1973; Kronenfeld et al., 2011; LeVine et al., 1994; Peacock, 2005; Sahlins, 1995, 1999; Sewell, 2005; Shore, 1996; Shweder, 1991; Shweder and LeVine, 1984). It remains to be seen whether and just how soon the concept regains its former popularity in anthropology.

The 'Standard View' of Culture in North American Anthropology

In the narrower 'humanistic' sense of the term, 'culture' refers to the control of elementary human impulses through the refinement of judgment, taste, and intellect and, by extension, to those activities believed to express and sustain that sophistication (like art or other 'high productions of mind'). What one might call the standard anthropological view of 'culture' is much broader, encompassing, as Renato Rosaldo writes: "the everyday and the esoteric, the mundane and the elevated, the ridiculous and the sublime. Neither high nor low, culture [in the anthropological sense] is all-pervasive" (1989).

But just how broad should the anthropological definition of culture be? Kroeber and Kluckhohn's formulation was cumbersome in part because it was so inclusive. It called on anthropologists to study not just other people's beliefs (their ideas of what the world is like) but also other people's normative standards (their ideas of what is good and what is right). It called on anthropologists to study not just the explicit 'ethnosciences' and doctrinal moral and religious codes of the members of a community but their tacit, implicit, or intuitive understandings as well.

Anthropologists have long disagreed about whether patterns of behavior should be considered a part of 'culture.' Margaret Mead employed a definition that centered around a 'complex of behavior.' David Schneider, on the other hand, went so far as to suggest that even norms for behavior should be excluded from cultural analysis. As Schneider put it: "norms tell the actor how to play the scene, culture tells the actor how the scene is set and what it all means" (1968).

With their definition, Kroeber and Kluckhohn sought a middle course between the *Scylla* of a purely behavioral definition of culture and the *Charybdis* of a purely ideational one. On the one hand, Kroeber and Kluckhohn suggest that culture is more than just social habits or "patterns of behavior that are learned and passed on from generation to generation." On the other hand, it is not just a system of categories, doctrines, propositions, or symbols per se. Thus in the 1952 definition, culture is defined as the ideational side of social action or social practice, and anthropologists are called upon to view cultural analysis as the interpretative study of behavior, although rather little is said about what particular theory of interpretation should guide the analysis.

Useful definitions deserve to be expressed in elegant terms, and Kroeber's and Kluckhohn's definition of culture is cumbersome, to say the least. But it is not the only expression of the standard view. The most exquisite and straightforward formulation is Robert Redfield's definition: "conventional understandings manifest in act and artifact" (Redfield, 1941, p. 133). Another variation, perhaps the most famous definition of culture since the 1950s, is the one proposed by Clifford Geertz (1973, p. 89). He puts it this way: "…the culture concept…denotes an historically transmitted pattern of meanings embodied in symbols, a system of inherited conceptions expressed in symbolic form by means of which men communicate, perpetuate, and develop their knowledge about and attitudes towards life."

The definitions proposed by Kroeber and Kluckhohn, Redfield, and Geertz call out for specification and clarification. Nevertheless, those definitions are a good reference point for understanding current debates about the values and dangers associated with the very idea of 'culture.' One can summarize the standard view by saying that 'culture' refers to community-specific ideas about what is true, good, beautiful, and efficient. To be 'cultural' those ideas about truth, goodness, beauty, and efficiency must be socially inherited and customary. To be 'cultural' those socially inherited and customary ideas must be embodied and/or enacted meanings; they must actually be constitutive of (and thereby revealed in) a way of life. Alternatively stated, the standard North American anthropological view of 'culture' refers to what the British philosopher Isaiah Berlin (1976) called "goals, values and pictures of the world" that are made manifest in the speech, laws, and routine practices of some self-monitoring and self-perpetuating group. These 'goals, values and pictures of the world' or 'ideas about what is true, good, beautiful, and efficient' are sometimes referred to as 'cultural models' (D'Andrade, 1995a; Holland and Quinn, 1987; Shore, 1996).

Fault Lines in Contemporary Anthropology

The 'standard' North American anthropological view of culture was synthesized and defined by Kroeber and Kluckhohn before the discipline went through a series of revolutions and movements that fractured the field and divided it on the basis of somewhat different visions of its mission. Contemporary views of 'culture' reflect those divisions to some extent. Thus, while it may be hazardous to propose a map of the current intellectual camps within cultural anthropology, such a map may also be helpful in understanding the various types of anticultural, postcultural, and procultural positions that have emerged within anthropology over the past 50 years.

It is fair to say that contemporary cultural anthropology is divided into at least these four conceptions of the field.

Identity Politics

The first is a conception of anthropology as a platform for moral activism in the battles against racism, sexism, homophobia, and neocolonialism and as a forum for identity politics in the fight against exploitation, discrimination, and oppression. Advocates of this conception of anthropology have several concerns about the idea of 'culture.' They argue that the idea of culture is an excuse for the maintenance of authoritarian power structures and permits despots and patriarchs around the world to deflect criticism of their practices by saying 'that is our custom' or 'that is the way we do things in our culture' (see Abu-Lughod, 1991; Said, 1978; Scheper-Hughes, 1995; Wikan, 1995). The claim by liberal 'first-world' feminists that 'multiculturalism is bad for women' (Okin, 1999) is an expression of this view, which tends to associate 'culture' with the idea of patriarchal domination (Haynes and Prakash, 1991; Raheja and Gold, 1994). This conception of the mission of anthropology is closely allied with a global human rights movement that has promoted the notion of 'harmful traditional practices' and has a firm sense of what is objectively and universally right and wrong (see Sabatello, 2009).

However, not all moral activists in anthropology want to dump the idea of culture. Some have found ways to put the idea to work in the service of their own political aims. Anthropologists who are active in the identity politics movement find the idea of culture politically and strategically convenient in their egalitarian battles on behalf of 'oppressed peoples.' There are generally three ways to mitigate invidious comparisons between groups (as in, group comparisons of wealth, occupational attainment and success, or school performance): (1) deny that any real differences exist, (2) attribute all differences to a history of oppression or discrimination, or (3) celebrate the differences as 'cultural.' In the identity politics movement, 'culture' has become something of a code word for 'race' and ethnic minority status.

Skeptical Postmodernism

The second conception of the mission of anthropology is a conception of the field as a deconstructive discipline and as an arena for skeptical postmodern critiques of all ethnographic representations and so-called 'objective' knowledge (see for example, Clifford and Marcus, 1986; Latour and Woolgar, 1979; Rabinow, 1983; Foucault, 1973; Rosenau, 1992). Advocates of this conception of anthropology call for a deeply skeptical reading of all anthropological representations of 'others,' especially of those accounts that make claims about some 'primordial' or essential or core cultural identity which members of some group are supposed to share. The skeptical postmodernists raise doubts about the reality and existence of identifiable cultural groups. They are critical of all attempts to draw a portrait of 'others' that represents them with any characteristic face. They are suspicious of the very idea of boundaries and borders and loyalties to one's historical ethical in-group or 'tribe.' They view the idea of a 'culture' as a fiction, the goal of objective representation as misguided, and the products of ethnography as largely 'made up' or constructed in the service of domination.

One of the many ironies of contemporary anthropology is that for a while members of the first two camps of anthropology (the identity politics/moral activists and the skeptical postmodern/deconstructivists) thought they were allies. Indeed, they had an imagined common enemy: the hegemonic heterosexual 'first-world' white males, such as Kroeber, Kluckhohn, Redfield, and Geertz, who historically had defined the mission of cultural anthropology.

For the most part, the alliance was short-lived. Identity politics requires a robust notion of 'identity' and group membership. Moral activism requires a good deal of conviction about the existence of an objective moral charter (such as inalienable human rights) and is typically motivated by the view that some cultural customs or social norms are 'objectively wrong'. Skeptical postmodernism is intellectually incapable of lending support to either of those metaphysical notions and is readily put to use deconstructing the 'woman' of 'Women's Studies,' the imagined common identity of the ethnic group, and all supposed objective moral foundations for any political cause. Paradoxically, continued attempts to reconcile identity politics with postmodern skepticism (e.g., Scheper-Hughes, 1995; Abu-Lughod, 1991) produce a moralizing model in which all forms of generalization and representation are doubted and distrusted *except* for domination and oppression, which are privileged as a priori objective truths or categories of moral condemnation (D'Andrade, 1995b).

Neo-Positivism

A third conception of the mission of anthropology is a conception of the field as a pure 'positive' science (see for example, D'Andrade, 1995a,b, 2008; Kronenfeld et al., 2011; Sperber, 1985; Romney et al., 1986). The positive scientists view anthropology as a value-neutral and nonmoralizing discipline. Their preferred aims for the discipline are to reliably and validly represent the law-like patterns in the world and to develop universal explanatory theories and test specific hypotheses about objectively observable regularities in social and mental life. Advocates of this conception want to protect anthropology from identity politics and skeptical postmodern critique by accurately recording rather than judging and condemning other peoples' practices, and by developing objective or scientific standards for evaluating the truth of ethnographic evidence. This is a laudable aim, although one that has been contested by skeptical postmodernists, and there has been much useful work in neopositivist fields such as 'cognitive anthropology' representing the content, structure, and degree of sharing of 'cultural models' (see, e.g., D'Andrade, 1995a; Holland and Quinn, 1987; Kronenfeld et al., 2011; Romney et al., 1986). Nevertheless, the positive scientists in anthropology thereby tend to beg a critical question close to the heart of all great social theorists. Is this or that social order really a moral order? Is this or that social order a way of living that might appeal to a rational and morally sensitive person, and if not

how can we make it become so? When it comes to evaluating what is truly desirable or really 'good' in social life the neopositivists are very much like the skeptical postmodernists – both turn radically subjective or relativistic and believe there is no scientific or objective foundation for the value judgments made in different traditions of value or historical ethical communities, aside from reporting their potential adaptive value in promoting group cohesiveness and survival.

Romantic Pluralism

A fourth conception of the mission of anthropology is a conception of the field as a romantic discipline designed to test the limits of pluralism. Pluralism is the idea that things can be different but equal, and that diversity can be good. It is a measure of some of the tensions within contemporary anthropology that while the 'ethnography of difference' is viewed with suspicion by some of the anticulturalists, it is universally embraced by romantic pluralists. Anticulturalists worry that any description of cultural difference merely sows the seeds of invidious comparison and ethnic conflict, and thus should be disavowed. For the romantic pluralists, however, the recognition and appreciation of cultural differences is one of the major aims of ethnography in particular and cultural anthropology in general.

The intellectual inheritance of the 'romantic' tradition most relevant to this camp of anthropology is a conception of culture as an extension of the creative imagination of agents who possess the capacity to initiate actions and evaluate their consequences with respect to some conception of what is dignified, divine, moral, or 'good,' which many anthropologists believe are distinctive intellectual capacities of human beings (see Geertz, 1973; Sapir, 1963; Sahlins, 1995; Shweder and LeVine, 1984; Shweder, 1991). According to romantic pluralists, a culture is an historical moral community whose social norms and customs make manifest and give expression to metaphysical notions about what is true and ideas about which values and forms of social organization are of greater value. Those notions and ideas provide members of the moral community with good reasons for action within their own terms, although they are not strictly dictated by logic and do not arise directly from (meaning-free) experience. According to this romantic pluralist conception of culture, there is plenty of room within the limits of logic and experience for cultural variety, and for the historical creation of different lived conceptions of what it means to be a rational and moral human being. According to this view, social and cultural realities are neither logically deduced nor simply found in direct experience but are rather constructed by, and for, more or less rational agents. The human creative imagination has the capacity to fill in, and give definition to, a vast discretionary space that stretches in between the necessary truths of formal logic and the uninterpreted evidence of the senses. Advocates of this conception of anthropology are dedicated not only to the project of accurate ethnographic representation but also to the cognitive and moral defense of different ways of life, frames of reference, and points of view. They write books or articles about Azande Witchcraft (Evans-Pritchard, 1937) or Balinese conceptions of the 'person' (Geertz, 1973) or Oriya Hindu family life and gender relations (Menon and Shweder, 1998) or Egyptian Muslim female conceptions of piety and 'the feminist subject' (Mahmood, 2005) which portray the ideas and practices of others as different but equal to our own, in the sense that such ideas and practices are represented as meaningful and imaginative yet supportable within the broad limits of scientific, practical, and moral reason (see also Haidt, 2012).

A Fifth Camp within Anthropology? The Return of Cultural Developmentalism and the 'First-World's' Burden

Increasingly these days, as the world 'globalizes,' the concept of 'culture' gets used to explain differences in the economic, social, political, educational, and moral accomplishments of nations, groups, or peoples. An 'evolutionary' or 'developmental' view of culture has returned to the intellectual scene. Along with it comes the claim that some groups have the wrong models, the wrong values, the wrong patterns of behavior, and that is the reason that their economies are poor, their governments corrupt, and their people unhealthy, unhappy, and oppressed. The cultural developmental view of cultural differences was quite popular at the very beginning of the twentieth century, and is associated with the 'civilizing project' or the 'white man's burden' to uplift those who are ignorant, superstitious, primitive, savage, and poor. Quite remarkably, the cultural developmental view is increasingly popular at the beginning of the twenty-first century as well, especially outside of anthropology, for example, in economics and political science (Harrison and Huntington, 2000; Landes, 1998; see Shweder, 2011, 2012). For some economists, reference to social and cultural 'capital' provides the basis for a less atomistic methodological individualism still grounded in rational choice theory but sensitive to the role of 'social structures' in everyday decision-making (Coleman, 1988). In development economics, however (e.g., at the World Bank), the view that 'culture counts' or that 'culture matters' is now popular in part because it is a discreet way of telling 'underdeveloped' nations (either rightly or wrongly) that the 'Westernization' of their cultures is a necessary condition for economic growth. Cultural developmentalists want to convert others to some preferred superior way of living. Their aim is to eliminate or at least minimize the differences between peoples rather than tolerate or appreciate them as products of the creative imagination. This viewpoint has returned, at least implicitly, in anthropology as well, especially among moral activists.

Relatively few anthropologists would actually describe themselves as cultural developmentalists. Nevertheless, that stance is far more common in anthropology than many admit, especially when the topic concerns gender relations and family life practices, for example, polygamy, purdah, arranged marriage, burqas, bride-price, female circumcision, and the association of femininity with domesticity and the production of children. So along with the international human rights movement and other agents and agencies promoting Western-style globalization, there are anthropologists these days who now take an interest in other

cultures mainly as objects of moral scorn. The up-from-barbarism theme of (certain versions of) Western liberalism has once again become fashionable on the anthropological scene (see Boddy, 2007; Sabatello, 2009; Thomas, 2003).

Multiculturalism and the Problem of 'Difference'

One tension inherent in all anthropological interpretation is the problem of 'difference,' what to make of it and what to do about it. This is also called the problem of the 'other,' although the term 'other' is used variously in the anthropological literature. It is sometimes used to connote difference per se without any initial judgment of relative worth. It is sometimes used to connote unbridgeable differences. It is sometimes used to connote a solipsistic gap between self-knowledge and a mysterious or 'spectral other' whose identity can never be truly inscribed. It is sometimes used to connote the representation of 'others' as so different as to be less than or other than human, or as different in ways that condemn them to inferior status and/or justify their domination. Here we use the term to connote difference per se.

The problem of 'difference' inherent in anthropological interpretation is not just a problem for anthropology. It arises whenever members of different groups (e.g., Jesuit missionaries and Native North American Indians; British traders and Hindu Brahmans; Western feminist human rights activists and Islamic fundamentalist women) or members of different social categories (e.g., gay men and heterosexual men) encounter each other. Someone finds the encounter disturbing, puzzling, strange, or astonishing because of some apparent difference between self and 'other' and wants to know what to make of it and (if they have the power) what (if anything) to do about it.

In the history of anthropology, the apparent differences mostly concerned differences in the ideas and practices of members of different groups. But in recent decades, globalization and large-scale immigration have intensified the challenge of cultural diversity within liberal democracies (Hechter, 2000). In this context, the term 'multiculturalism' has become a slogan for many rather different and even contradictory sorts of educational and social movements and public policy agendas.

There are multiculturalists who value group differences and want to preserve them. There are other multiculturalists who think group differences are the product of vicious discrimination. There are multiculturalists who think that the word 'multiculturalism' means being a hybrid and actively promoting the erosion of borders or boundaries between groups. There are other multiculturalists who think the word implies autonomy, in-group solidarity, the power to remain separate or pure, and the capacity to maintain boundaries or restore a distinctive way of life.

There are multiculturalists who use the word in an almost ironic sense to commend and promote the mainstreaming, assimilation, integration, or inclusion of people of different colors or ancestries into the society and shared subculture of the American elite; and there are others who use the term to call on the mainstream elite to accommodate themselves to minority-group differences in customs, values, and beliefs.

One suspects that in each of these cases there is a slightly different concept of 'culture' at work. But we cannot avoid the question, what form does and should multiculturalism take in our emerging postmodern society? (see Shweder et al., 2002; Minow et al., 2008; Shweder, 2011).

Globalization

The narrowest definition of 'globalization' refers to the linking of the world's economies (e.g., free trade across borders) with the aim of promoting aggregate wealth and economic growth. Yet it readily expands so that a new cosmopolitan economic order gets imagined, which consists entirely of global economic organizations (the International Monetary Fund, the World Bank), multinational corporations, and multicultural states with open borders. According to this rather utopian vision of a 'borderless capitalism', goods, capital, and labor ought to be freely marketed on a world-wide scale for the sake of global prosperity. For those who adopt such a perspective any desire for an ancestral homeland or for a national identity based on religion, ethnicity, 'race,' or 'tribe' with associated restrictions on residence, affiliation, and trade is viewed as 'illiberal' and disparaged as a form of retrograde or irrational 'apartheid' or 'ethnonationalism.'

Fully expanded, however, the idea of 'globalization' actually becomes a hypothesis about human nature and an imperial call for 'enlightened' moral interventions into other ways of life in order to free them of their supposed 'barbarisms,' 'superstitions,' and 'irrationalities.' This expansive 'globalization hypothesis' makes three related claims: (1) that Western-like aspirations, tastes, and ideas are objectively the best aspirations, tastes, and ideas in the world; (2) that Western-like aspirations, tastes, and ideas will be fired up or freed up by economic globalization; and (3) that the world will/already has and/or ought to become 'Westernized.' Western-like aspirations include the desire for liberal democracy, free enterprise, private property, autonomy, individualism, equality, and the protection of 'natural' or universal 'rights.' They include the modernist notion that all social distinctions based on collective identities (ethnicity, religion, gender) are invidious. They include as well the notion that 'individuals' should transcend their 'tradition-bound' commitments and experience the quality of their lives solely in secular and ecumenical terms, for example, as measured by, wealth, health, or years of life.

The picture of a cosmopolitan world of individuals without groups, in which meanings are detached or abstracted from communities and traded on a free market of ideas, has influenced the thinking of some postcultural theorists. Whether that picture is realistic remains to be seen. It is quite possible that other cultures and civilizations do not need to become just like the United States to materially benefit from participation in an emergent global economy. Modern economic institutions such as private property seem to have effectively served many interests, including the interests of communitarians as well as religious and ethnic groups all over the world (Stolzenberg, 2004). As Comaroff and Comaroff have pointed out, even the outright commercialization of cultural identity – be it in the form of tangible commodities (e.g., souvenirs and

'indigenous crafts'), collective business ventures (e.g., Native American 'tribal' casinos), or 'cultural tourism' – does not necessarily erode genuine difference, but can instead provide a mode of reaffirming and refashioning ethnic identity (Comaroff and Comaroff, 2009).

The true connection between 'globalization' narrowly conceived ('free trade') and 'globalization' expansively conceived (Western values, 'culture', and institutions taking over the world) has yet to be firmly established (see Shweder, 2011). Is this truly the 'end of history' as Francis Fukuyama (1989) once proposed? Or, as in Samuel Huntington's thesis, will great divisions among humankind inevitably spring up along 'cultural fault lines' and lead to a clash of civilizations? (Huntington, 1996). Huntington's description of those fateful cultural differences – 'fundamental,' intractable, the 'product of centuries' – bears striking resemblance to Clifford Geertz's description of 'primordial sentiments' (Geertz, 1963), seen here stubbornly resisting the onslaught of globalizing and 'westernizing' forces. Nevertheless, the idea that the rich nations of North America and Northern Europe have an obligation to use their economic and military power to civilize and develop the world is no less popular today than it was 100 years ago when the empire was British rather than 'neo-liberal' or American.

'Culture': Popular Objections and Common Misattributions

Within anthropology many reasons have been advanced for doubting the usefulness of the culture concept. But are they persuasive or decisive reasons? Those who continue to embrace some variety of the Kroeber and Kluckhohn definition of culture tend to believe that their idea of culture does not carry most of the implications that are the supposed grounds for various anticultural or postcultural critiques.

For example, the Kroeber and Kluckhohn definition of culture does not really imply that 'whatever is, is okay,' nor does it necessarily share in the view expressed by some moral relativists that morality is just a convenient term for socially approved habits (e.g., Benedict, 1934). It is important to recognize that valid social criticism and questions of moral justification are not ruled out by the standard anthropological view of 'culture.' Nothing in the Kroeber and Kluckhohn formulation suggests that the things that other peoples desire are in fact truly desirable or that the things that other peoples think are of value are actually of value. Consensus does not add up to moral truth. In other words, a definition of culture per se is not a theory of the 'good.' From a moral point of view one need not throw out the idea of culture just because some tyrant puts the word 'culture' to some misuse, or because at times some ethnic groups enter into geopolitical conflict (see Shweder, 2012).

The idea of culture also does not imply passive acceptance of received practice or that human beings lack 'agency,' a common claim among anticulture theorists. Indeed, many proculture theorists find it astonishing to see the idea of 'agency' or 'intentionality' used as synonyms for 'resistance to culture' in the discourse of 'anticulture' theorists. Even fully rational, fully empowered, fully 'agentic' human beings discover that membership in some particular tradition of meanings and values is an essential condition for personal identity and individual happiness. Human beings who are 'liberationists' are no more agentic than 'fundamentalists,' and neither stands outside some tradition of meaning and value.

The idea of culture also does not imply the absence of debate, contestation, or dispute among members of a group. Nor does it necessarily imply the existence of within-group homogeneity in knowledge, belief, or practice. Every cultural system has experts and novices; one does not stop being a member of a common culture just because cultural knowledge is distributed and someone knows much more than you do about, for example, how to conduct a funeral or apply for a mortgage. One does not stop being a member of a common culture just because there are factions in the community. The claim that there are between-group cultural differences never has implied the absence of within-group differentiation. The idea of culture does not imply that every item of culture is in the possession or consciousness of every member of that culture. The idea of culture merely directs our attention to those ideas about what is true, good, beautiful, and efficient that are acquired by virtue of membership in some group. Members of a cultural community do not always agree about this or that, but they do take an interest in each other's ideas about what is true, good, beautiful, and efficient because those ideas (and related practices) have a bearing on the perpetuation of their way of life, and what they share is that collective inheritance. Since the standard view does not assume that a culture is a well-bounded, fixed, and homogeneous block, the critique of the concept of 'culture' that starts with the observation of internal variation and ends 'therefore there is no cultural system' should have been a nonstarter.

In its noun form, 'culture' can either be countable (as in *a* culture or 'Japanese culture') or uncountable (denoting culture in the abstract sense as a domain of knowledge or inquiry). The conflation of the two senses is one of the main sources of confusion and misattribution in critiques of the idea of culture. Often, critics purport to take aim at the broader concept of 'culture' while fashioning arguments relevant only to its particular usage in the plural form ('cultures'). The historian of anthropology George Stocking identified the popularization of the plural use of 'cultures' by Franz Boas and his students in the 1930s and 1940s as a crucial turning point, after which the problematic modern notion of cultures as static, bounded entities gained traction. Anticipating contemporary critiques, Kroeber and Kluckhohn (1952) defended the practice of speaking of 'cultures' in the plural as a useful abstraction, pointing out that one could speak at the same time of a Tokyo or a Japanese or an East Asian culture without implying that any of them represented a homogeneous or totalizing way of life.

Just as relevant to contemporary anthropologists' feelings about the concept of 'culture,' however, is the manner of the term's deployment in ordinary language. 'Culture,' as William Sewell tells us, "has escaped all possibility of control by anthropologists" (2005, p. 155). It has become 'compromised' (Clifford and Marcus, 1986). More and more, everyone from policymakers to activists, educators to demagogues seem intent to (affectionately) portray 'cultures' as bounded, uniform, unchanging, and primordial. Such connotations are enough to send cautious anthropologists into flight.

Faced with this misappropriation of what was once their term, anthropologists have generally responded in one of two ways: (1) by substituting terms like 'discourse' or 'habitus' for 'culture' in the hopes of avoiding some of the latter's unsavory connotations (it is never clear if this 'lexical avoidance behavior' avoids the supposed sins of 'culture' or merely reproduces them in a new guise); or (2) by restricting themselves to the adjectival use of the term: instead of 'culture' or worse yet, 'cultures,' anthropologists speak of 'cultural tradition,' cultural practice, 'cultural capital,' and so on (see Appadurai, 1996; Brightman, 1995). The latter tactic is best summed up in the title of Michael Silverstein's (2005) essay, which also does a good job of capturing both the ambivalence and the urgency of contemporary anthropologists' attitudes toward 'culture': "Languages/cultures are dead!" Silverstein exclaims, "Long live the linguistic-cultural!" Nevertheless, many social scientists and public policy analysts continue to look to anthropology for a useful concept of culture and not for no concept of culture at all.

See also: Cultural Psychology; Cultural Relativism, Anthropology of; Culture and the Self: Implications for Psychological Theory; Deconstruction; Ethnography; Globalization and World Culture; Identity Movements; Identity in Anthropology; Pluralism; Positivism, Sociological; Postmodernism: Philosophical Aspects.

Bibliography

Abu-Lughod, L., 1991. Writing against culture. In: Fox, R. (Ed.), Recapturing Anthropology: Working on the Present. School for American Research Press, Santa Fe, New Mexico.
Appadurai, A., 1996. Modernity at Large: Cultural Dimensions of Globalization. University of Minnesota Press.
Asad, T., 1973. Anthropology and the Colonial Encounter. Ithaca Press, London.
Beattie, J., 1964. Other Cultures: Aims, Methods, and Achievements in Social Anthropology. The Free Press, New York.
Benedict, R., 1934. Anthropology and the abnormal. Journal of General Psychology 10, 59–80.
Berlin, I., 1976. Vico and Herder. Hogarth Press, London.
Boddy, J., 2007. Civilizing Women: British Crusades in Colonial Sudan. Princeton University Press.
Borofsky, R. (Ed.), 1994. Assessing Cultural Anthropology. McGraw-Hill, New York.
Brightman, R., 1995. Forget culture: replacement, transcendence, reflexification. Cultural Anthropology 10, 509–546.
Clifford, J., Marcus, G., 1986. Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley, CA.
Coleman, J., 1988. Social capital in the creation of human capital. The American Journal of Sociology 94, S95–S120.
Comaroff, J.L., Comaroff, J., 1991. Of Revelation and Revolution: Christianity, Colonialism, and Consciousness in South Africa. University of Chicago Press, Chicago.
Comaroff, J.L., Comaroff, J., 2009. Ethnicity, Inc. University of Chicago Press, Chicago.
D'Andrade, R., 1984. Cultural meaning systems. In: Shweder, R.A., LeVine, R.A. (Eds.), Culture Theory: Essays on Mind, Self and Emotion. Cambridge University Press, New York.
D'Andrade, R., 1995a. The Development of Cognitive Anthropology. Cambridge University Press, New York.
D'Andrade, R., 1995b. Moral models in anthropology. Current Anthropology 36, 399–408.
D'Andrade, R., 2008. A Study of Personal and Cultural Values: American, Japanese and Vietnamese. Palgrave Macmillan, New York.
Denzin, N., 1996. Epistemological crisis in the human disciplines. In: Jessor, R., Colby, A., Shweder, R.A. (Eds.), Ethnography and Human Development: Context and Meaning in Social Inquiry. University of Chicago Press, Chicago.
Dumont, L., 1970. Homo Hierarchicus: An Essay on the Caste System. University of Chicago Press, Chicago.
Evans-Pritchard, E.E., 1937. Witchcraft, Oracles and Magic among the Azande. Clarendon Press, Oxford, UK.
Foucault, M., 1973. The Order of Things: An Archeology of the Human Sciences. Vintage Books, New York.
Fox, R. (Ed.), 1991. Recapturing Anthropology. School of American Research, Santa Fe.
Freeman, D., 1983. Margaret Mead and Samoa: The Making and Unmaking of an Anthropological Myth. Harvard University Press, Cambridge, MA.
Fukuyama, F., 1989. The end of history? The National Interest 16, 3–18.
Geertz, C., 1963. The integrative revolution: primordial sentiments and civil politics in the new states. In: Geertz, C. (Ed.), Old Societies and New States. The Free Press, New York, pp. 105–157.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Haidt, J., 2012. The Righteous Mind: Why Good People Are Divided by Politics and Religion. Pantheon Books.
Harrison, L.E., Huntington, S.P., 2000. Culture Matters: How Values Shape Human Progress. Basic Books, New York.
Haynes, D.E., Prakash, G. (Eds.), 1991. Contesting Power: Resistance and Everyday Social Relations in South Asia. Oxford University Press, Delhi.
Hechter, M., 2000. Containing Nationalism. Oxford University Press, Oxford.
Holland, D., Quinn, N. (Eds.), 1987. Cultural Models in Language and Thought. Cambridge University Press, New York.
Huntington, S.P., 1996. The Clash of Civilizations and the Remaking of the World Order. Simon and Shuster, New York.
Hymes, D. (Ed.), 1972. Reinventing Anthropology. Random House, New York.
Kitayama, S., Markus, H.R., 1994. Emotion and Culture: Empirical Studies of Mutual Influences. American Psychological Association, Washington, DC.
Kroeber, A.L., Kluckhohn, C., 1952. Culture: A Critical Review of Concepts and Definitions. Random House, New York.
Kronenfeld, D.B., et al. (Eds.), 2011. A Companion to Cognitive Anthropology. Wiley-Blackwell, Malden, MA.
Kuper, A., 1999. Culture: The Anthropologists' Account. Harvard University Press, Cambridge, MA.
Landes, D.S., 1998. The Wealth and Poverty of Nations: Why Some Are So Rich and Some Are So Poor. W.W. Norton, New York.
Latour, B., Woolgar, S., 1979. Laboratory Life. Princeton University Press, Princeton, NJ.
LeVine, R., Dixon, S., LeVine, S., Richman, A., Leiderman, P.H., Keefer, C.H., Brazelton, T.B., 1994. Child Care and Culture: Lessons from Africa. Cambridge University Press, Cambridge, UK.
Mahmood, S., 2005. The Politics of Piety: The Islamic Revival and the Feminist Subject. Princeton University Press.
Malinowski, B., 1944. A Scientific Theory of Culture and Other Essays. The University of North Carolina Press.
Marcus, G.E., Fischer, M.J., 1986. Anthropology as Cultural Critique. University of Chicago Press, Chicago.
Menon, U., Shweder, R.A., 1998. The return of the white man's burden: the moral discourse of anthropology encounters the family life of Oriya women. In: Shweder, R.A. (Ed.), Welcome to Middle Age! (And Other Cultural Fictions). University of Chicago Press, Chicago.
Minow, M., Shweder, R.A., Markus, H.R., 2008. Just Schools: Pursuing Equality in Societies of Difference. Russell Sage Foundation Press, New York.
Okin, S.M., 1999. Is Multiculturalism Bad for Women? Princeton University Press, Princeton.
Peacock, J., 2005. Geertz's concept of culture in historical context. In: Shweder, R., Good, B. (Eds.), Clifford Geertz by His Colleagues. University of Chicago Press, Chicago, pp. 52–62.
Prentice, D., Miller, D., 1999. Cultural Divides: Understanding and Overcoming Group Conflict. Russell Sage Foundation Press, New York.
Rabinow, P., 1983. Humanism as nihilism: the bracketing of truth and consciousness in American cultural anthropology. In: Bellah, R., Rabinow, P., Sullivan, W. (Eds.), Social Science as Moral Inquiry. Columbia University Press, New York, pp. 33–51.
Raheja, G.G., Gold, A.G., 1994. Listen to Heron's Words: Reimagining Gender and Kinship in North India. University of California Press, Berkeley.
Redfield, R., 1941. The Folk Culture of Yucatan. University of Chicago Press, Chicago.
Reyna, S.P., 1994. Literary anthropology and the case against science. Man 29, 555–581.

Rosaldo, R., 1989. Culture and Truth: The Remaking of Social Analysis. Beacon Press, Boston.
Rosenau, P., 1992. Post-Modernism and the Social Sciences. Princeton University Press, Princeton, NJ.
Romney, A.K., Weller, S., Batchelder, W.H., 1986. Culture as consensus: a theory of culture and informant accuracy. American Anthropologist 88, 313–338.
Sabatello, Maya, 2009. Children's Bioethics: The International Biopolitical Discourse on Harmful Traditional Practices and the Right of the Child to Cultural Identity. Brill.
Said, E., 1978. Orientalism. Pantheon, New York.
Sahlins, M., 1976. Culture and Practical Reason. University of Chicago Press, Chicago.
Sahlins, M., 1995. How Natives Think, about Captain Cook, For Example. University of Chicago Press, Chicago.
Sahlins, M., 1999. Two or three things that I know about culture. Journal of the Royal Anthropological Institute 5, 399–421.
Sangren, S.P., 1988. Rhetoric and the authority of ethnography. Postmodernism and the social reproduction of texts. Current Anthropology 29, 405–435.
Sapir, E., 1963. In: Mandelbaum, D. (Ed.), Selected Writings of Edward Sapir in Language, Culture and Personality. University of California Press, Berkeley.
Scheper-Hughes, N., 1995. The primacy of the ethical: propositions for a militant anthropology. Current Anthropology 36, 409–430.
Schneider, D., 1968. American Kinship. The University of Chicago Press, Chicago.
Sewell, W., 2005. The concept(s) of culture. In: Sewell, W. (Ed.), Logics of History. The University of Chicago Press, Chicago, pp. 152–174.
Shore, B., 1996. Culture in Mind: Cognition, Culture and the Problem of Meaning. Oxford University Press, New York.
Shweder, R.A., 2011. Geertz's Challenge: Is it possible to be a robust cultural pluralist and a dedicated political liberal at the same time? In: Sarat, A., Douglas, L., Umphrey, M.M. (Eds.), Law without Nations. Stanford University Press, Stanford, CA.
Shweder, R.A., 2012. Relativism and Universalism. In: Fassin, D. (Ed.), A Companion to Moral Anthropology. Wiley-Blackwell, Malden, MA.
Shweder, R.A., 1991. Thinking through Cultures: Expeditions in Cultural Psychology. Harvard University Press, Cambridge, MA.
Shweder, R.A., LeVine, R.A., 1984. Culture Theory: Essays on Mind, Self and Emotion. Cambridge University Press, New York.
Shweder, R.A., Markus, H.R., Minow, M. (Eds.), 2002. Engaging Cultural Differences: The Multicultural Challenge in Liberal Democracies. Russell Sage Foundation, New York.
Silverstein, M., 2005. Languages/cultures are dead: long live the linguistic-cultural. In: Segal, D., Yanagisako, S. (Eds.), Unwrapping the Sacred Bundle: Reflections on the Discipline of Anthropology. Duke University Press, Durham, pp. 99–125.
Sperber, D., 1985. On Anthropological Knowledge. Cambridge University Press, Cambridge.
Spiro, M., 1986. Cultural relativism and the future of anthropology. Cultural Anthropology 1, 259–286.
Stolzenberg, N., 2004. The culture of property. In: Shweder, R.A., et al. (Eds.), Engaging Cultural Differences: The Multicultural Challenge in Liberal Democracies. Russell Sage Foundation, New York, pp. 177–193.
Thomas, L., 2003. Politics of the Womb: Women, Reproduction and the State in Kenya. University of California Press.
Wierzbicka, A., 1997. Understanding Cultures through Their Key Words: English, Russian, Polish, German, Japanese. Oxford University Press, New York.
Wikan, U., 1995. Culture, power and pain. Female Genital Mutilation in the Western World. Riffi, Stockholm.
WIN News, 22 September 1996. Available online:http://www.highbeam.com/doc/1G1-18907374.html.

Culture of Poverty: Critique

Hervé Varenne, Teachers College, Columbia University, New York, NY, USA
Michael Scroggins, Teachers College, Columbia University, Oakland, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article recounts the historical, theoretical, and empirical basis of the culture of poverty program as it was developed in the writings of Oscar Lewis and examines the anthropological critique of his work that developed immediately upon the heels of Lewis' final publications. Further, this article examines the recent emergence of the culture and poverty program and notes its confluences with and points of departure from the culture of poverty program.

In 1966, the anthropologist Oscar Lewis published an article titled 'The Culture of Poverty.' There, Lewis developed and systematized a way of analyzing what happens to people in poverty. He claimed that many poor people were poor because their parents had passed onto them various traits and habits that keep people in poverty. Also, he suggested that governmental policies might break what was also known as the 'cycle of poverty.' This way of looking at people in poverty, in its moral, intellectual, and political implications, has a very long history, but Lewis' formulation caught the attention of the public and remains as a kind of litmus test in all research on poverty. In 1966, policy makers in Washington were putting the final touches on the set of programs known as the 'Great Society.' They were driven by a complex political imperative. As John F. Kennedy put it in his inaugural address: "Man holds in his mortal hands the power to abolish all forms of human poverty" (20 January 1961). This begged for an overarching narrative grounded in the behavioral sciences. Researchers from all disciplines proposed lists of traits that appeared to make the poor different, and suggested that those traits might be alleviated through various interventions. Lewis brought all this together under the heading of 'culture' taken in what was then the most common sense version of the concept. He appeared to offer a compassionate theory of poverty that confirmed the wisdom of the Great Society programs and suggested how they might be expanded and refined as the United States kept fighting what was also known as the 'War on Poverty.' The anthropological reaction against this theory was swift and severe – but, to this day, new versions keep appearing under new guises.

While the phrase 'culture of poverty' is firmly associated with Oscar Lewis' work, as well as with the policies of the Johnson era, it can easily be placed in a long tradition of conceptualizing the poor and imagining policies to help and control them. In the eighteenth and nineteenth centuries, Malthus and Mayhew had provided the intellectual justification for viewing poverty as a problem to be controlled (Himmelfarb, 1971, 1983). They developed many of the methods and starting points later embraced by Lewis. This included descriptive statistics of the distribution of 'traits' (rate of prostitution, alcoholism, unwed motherhood, etc.) across populations and correlations often interpreted as causations. This paralleled much nineteenth-century social theory purporting to explain other differences in human populations on biological or evolutionary bases.

Lewis' conception of culture also built on a long tradition and it made common sense in an American context heavily influenced by culture and personality studies, as well as by the Parsonian attempt to bring together sociology, anthropology, and psychology in a grand theory of action (Parsons and Shils, 1951). Talcott Parsons was at the time the most powerful sociologist in America who, as the founder of the interdisciplinary Department of Social Relations at Harvard, trained many influential social scientists of the following generation. The Parsonian sense of culture as what is learned, valued, and transmitted from one generation to another may actually marked a significant departure from the Boasian origins of the use of the term in the anthropology he advocated. Boas founded at Columbia University the first modern academic Department of Anthropology in the United States, and he trained many of the most influential anthropologists of the 1930s, 1940s, and 1950s. He had proposed 'culture' as a way to force attention to the processes that distinguish populations from each other even when living in very similar ecological environments with similar technologies (Trouillot, 2003: 91). But many of his students, and even more of his students' students, including Lewis, transformed 'culture' into the possession of a population. It was made to refer to whatever shapes personalities, that is then inscribed in the personality of most member of the population, psychologically, through internalization whether phrased as 'enculturation' or 'socialization.' Many argued that this process could be so complete as to render people unable to imagine other possibilities than those proposed by 'their' culture. The process would lead them to reproduce their conditions in future generations unless external forces intervened. This emphasis on internalization of social and cultural patterns was further systematized by Parsons who argued powerfully that culture is mostly a 'value' concept referring to internal states (Parsons and Shils, 1951: 159–160).

Soon after Lewis' synthesis was published, intellectual opposition to this systematization arose, particularly among cultural anthropologists (Valentine, 1968; Leacock, 1971). By using the preposition 'of' to link culture and poverty, and by arguing that the poor are caught in a culture that they reproduce in their children, Lewis encouraged those who imagined that alleviating poverty had to proceed through reforming the poor, and particularly their children. Soon, the opposition started to argue that accepting culture of poverty framework

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64091-6

amounted to 'blaming the victim' and, in a few years, the phrase became a mark of opprobrium.

The overall argumentation, however, did not disappear, though it is now sometimes couched in different terms. In recent years, William Julius Wilson (2009, 2012[1987]), among others (Furstenberg et al., 1999), has argued for a return to at least some of the lines of investigation Lewis had put forth. The critique of these lines of investigation is also being renewed particularly by anthropologists who have distanced themselves from the concept of culture, as well as by those who are working at recapturing the original Boasian concerns with establishing that, among human beings, like causes, such as 'poverty,' do not necessarily lead to like effects.

This article is organized into seven sections. Beginning with a historical overview of precursors to the 'culture of poverty' program, the sections move chronologically toward the contemporary state of research and critical engagement relating poverty to culture.

Poverty and Policy: Some Historical Background

The Elizabethan Poor Laws of 1601, later carried from England to America (Katz, 2011: 11–16), divided the poor into three categories: those who could not work, those who could work but choose not to, and those who are willing to work. The Poor Laws mark the point at which religious duty toward the poor gave way to legally sanctioned relations. Two centuries later, in his *An Essay on the Principle of Population* (2003 [1798]), Thomas Robert Malthus theorized that there was a relationship between population and economy, but social position between the two was static and thus, whatever happened in the broad economy, the poor would be left mired in 'misery and vice.' Malthus opposed Adam Smith's argument that the spread of capitalism would lead to an improvement in the well-being of all, as national wealth increased (Himmelfarb, 1983). Malthus countered that the population principle would prevent this from happening because population expands to fit available resources. This produces a stress on these resources so that poverty would remain and a class of people would occupy the position. Malthus also inquired into the mechanism by which the population is sorted into successes and failures. His answer is that those at the bottom of the economic ladder remained there because they lacked the qualities of economic superiors; namely moral restraint in the form of vice, drunkenness, and sexual promiscuity, as well as the inability to discipline themselves to foresee and save for the future (Himmelfarb, 1983: 119).

Malthus's work entered the world of nineteenth century public policy in various ways. Mayhew, a journalist, took to the streets of London to observe and describe the types of poor people then living on the streets of London. He did this through a variety of methodological tools such as maps, biographies, and observation. More importantly, he applied the tools of storytelling to exaggerate his observations toward evoking pathos in his readers. Mayhew's (1986[1850]) book is filled with ethnographic vignettes and statistics linked to area maps aimed at showing the intensity of crime, ignorance (illiteracy), and illegitimacy. He provided a further warrant to novelistic accounts of poverty by Dickens and others. All this led to new laws and programs. To quote Himmelfarb: "It is one of the many ironies of the period that just at the time the poor were finally relieved of the stigma of pauperism, when they seemed to be following the model laid down by Adam Smith rather than that of Malthus, Mayhew came along and, with the most laudable intentions and the most generous of sympathies, inflicted upon the poor a new stigma and saddled society with a new problem, the 'culture of poverty'" (Himmelfarb, 1984: 370). Karl Marx himself, if one accepts Rancière's interpretation (2004[1983]), may have contributed to this view of the poor as caught in their position. On the one hand, Marx was suspicious of the labor unions emerging from the poor in England. On the other hand, he reserved for intellectuals a particular place in the enlightenment of the proletariat.

The Culture Concept from Boas to Lewis

The nineteenth century also saw the first definitions of the culture concept as it came to be used by anthropologists. As Edward Tylor put it: "Culture … is that complex whole which includes knowledge, belief, art, law, morals, custom, and any other capabilities and habits acquired by man as a member of society" (2004[1871]: 1). A century later, Lewis issued a definition of culture somewhat consistent with Tylor's definition – though slanted in a more psychological direction. As Lewis put it: "The culture of poverty is not just a matter of deprivation or disorganization, a term signifying the absence of something. It is a culture in the traditional anthropological sense in that it provides human beings with a design for living, with a readymade set of solutions for human problems, and so serves a significant adaptive function" (Lewis, 1966). At about the same time, the most powerful anthropologist of his generation, Clifford Geertz, wrote that 'culture' is both 'model of' and 'model for' behavior (Geertz, 1973[1966]: 93).

This view of culture has now been thoroughly critiqued (Abu-Lughod, 1991; Ortner, 1999) but, even at the time, it represented a significant departure from an earlier version developed by Franz Boas. Boas was trained in physics and geography but moved away from these fields when he started working with the people of the Pacific Northwest and curated ethnographic collections at the Smithsonian and the Museum of Natural History. There, he began to debate with others over the display of ethnographic artifacts and, indirectly, about the human activities that produced these artifacts. J.W. Powell, McGee, and Mason were convinced that primitive people came to adopt more evolved ideas, materials, and processes through a process of social evolution in which the environment strictly determined social development. Following this principle led Mason to argue for the display of ethnographic artifacts in term of evolutionary phases following the logic that, in human affairs, as in nature, like-causes lead to like-events. Boas opposed this. He argued that "due to the intricacy of the acting causes" (Boas, 1887: 485) in human action, it was unlikely that like causes could be established and used for drawing broad generalizations and "the disposition of men to act suitably can be the only general cause." This led him to argue, successfully, for the display of artifacts in the context of their use among a particular people, at a particular time in their history.

After Boas moved to Columbia University, he developed this perspective in his research and argued forcefully and systematically against all those who interpreted culture as the inevitable properties of people caught in various conditions. For him, culture referred to processes of ongoing adaptation to local conditions. Boas emphasized that people of all backgrounds can adapt to any setting if they must, or allowed to (1938[1911]). He demonstrated that people could adapt differently to very similar conditions. This style of argumentation was echoed in a letter, Conrad Arensberg wrote to Oscar Lewis. Lewis had written to Arensberg that "poverty, and all that goes with it, is literally stable, persistent, and passed down along family lines. Can all the many cross cultural similarities I have mentioned been the result of accident or coincidence?" Arensberg answered: "The reactions to poverty are many: Middle Eastern, Arab austerity, Puritan frugality, Chassidism, the lazzaroni of Naples, etc. This shows that poverty, even with urbanization, does not produce similarities of culture …" (Rigdon, 1988: 226) In other words, very similar causes (material deprivation, political discrimination, disease, etc.) lead to very different kinds of adaptations.

Lewis' Intellectual Milieu

Lewis was a product of Columbia's anthropology department, at a time when the department was in transition from the Boasian paradigm as it had evolved to concerns with cultural ecology and multilineal evolution. At Columbia, Lewis studied with Ruth Benedict but also took courses on technology, archaeology, etc. (Rigdon, 1988: 12). In 1939 Lewis, and his wife and frequent collaborator Ruth, were one of four field teams organized by Benedict to carry out a comparative and historical study in the Blackfoot Reservation. This led to Lewis' dissertation on "The Effects of White Contact upon Blackfoot Culture, with Special Reference to the Effects of the Fur Trade" (1943). The dissertation was not based on field research but rather on the analysis of historical documents. But it does echo the tension that characterized Lewis' work throughout his career: the concern with economic interaction between unequal groups, and the consequences of this interaction.

In 1944, and again in 1947, Lewis spent several months in Tepoztlán, a large village in Mexico, the beginning of a series of field investigations in the country. Lewis' subsequent work came to focus more on the psychological consequences of life in difficult economic conditions, and less on the social and political processes that produced difficult conditions. In his first book (1960[1951]), Lewis distinguishes his work from earlier work by Robert Redfield on the same village (1930), emphasizing that life in the village could not be understood except in terms of the political history of the village through the Mexican revolution, the Spanish colonization, and back to pre-Columbian. He lists the governmental agencies that controlled the village, expands on its political organization, and traces differences in wealth and land ownership. He also wrote about the impact of all this on interpersonal relationships, particularly in families. He concludes with comments on the relationship of his work to the discussions of 'culture and personality' that most concerned many influential anthropologists at the time and notes that "I have not written about *the* Tepoztecan personality" (1951: 426). But his interest in the issue remained and led him to focus on families as the unit of analyses and entry point for gaining "greater insight into both the culture and the people" (Lewis cited in Rigdon, 1988: 34). This turn, as well as his admiration for such major figures in the late versions of culture and personality theorizing as Erik Erickson and Jules Henry, took him to the formulation of the culture of poverty for which he remains best known.

The Policy Context of the 'Culture of Poverty'

When Lewis wrote his paper in the Scientific American, a still somewhat obscure Assistant Secretary of Labor, Daniel P. Moynihan, was asked to write a report making the case for federal policies to 'resolve a problem' he summarized as follows: "Three centuries of injustice have brought about deep-seated structural distortions in the life of the Negro American. At this point, the present tangle of pathology is capable of perpetuating itself without assistance from the white world. The cycle can be broken on if these distortions are set right [through a national effort directed toward the question of family structure]" (Moynihan, 1965: 47).

This report quoted extensively from E. Franklin Frazier's *The Negro Family in the United States* (1966[1939]) and used the book as justification. The foreword to the third edition, also published in 1966, was written by the famous sociologist Nathan Glazer who dismissed what he called the "cultural relativism and anthropological romanticization" associated with Allison Davis who had also written about black life in the postreconstruction South (1941). Glazer writes: "Frazier I think was more hardheaded, … as was W.E.B. Du Bois: It was all bad – the abandoned mothers, the roving men, the sexually experienced youth" (1966: xi). Glazer emphasized that much of what Frazier had written was supported by contemporary social science, including anthropology: "The family itself was the support of the social structure. Children were socialized into certain values … and they maintained a society" (Glazer, 1966: ix). This statement may be closer to Parsons than to the anthropologists, Glazer mentions. And he does not mention the Chicago sociologists who had trained Frazier and who would also have encouraged him toward social psychological explanations of individual behavior.

The statement was also close to statements by others hailing from other disciplines. In 1962, Michael Harrington's published a book, *The Other America: Poverty in the United States*, an examination of poverty in the Appalachian region of the United States. Stylistically, Harrington shared with Lewis and Mayhew a depiction of the poor as being both structurally and culturally deprived. Like Lewis, Harrington was moved by a moral and political concern with changing the conditions of the poor that had to proceed through the political activities of the more privileged. Social psychologists (Deutsch, 1960) came to the issue from slightly different perspectives. Linguists also provided evidence (Bereiter and Engelmann, 1966; Bernstein, 1962, 1965). All talked about deprivation, developmental delays, etc., and wrote about the effect of all this on school success, which could then be used as a proxy for the evaluation of the social costs of poverty and the effectiveness of various policies. The method was systematized in the *Equality of*

Educational Opportunity Study (2007[1966]), generally known as the 'Coleman Study,' commissioned by the United States Department of Health, Education, and Welfare in 1966 and parallels the Moynihan report by also emphasizing the role of families over the role of schools in the measure of schooling success. To this day, this method remains the main entry point in the operationalization of poverty and its effects, as well as the means to measure the effectiveness of government policies. The ensemble of political concerns crystalized through the Civil Rights movement, given an opportunity by Lyndon Johnson and the Congressional majorities in the 1960s, and supported by all the most respectable of behavioral scientific theorizing – including by black intellectuals like Frazier – produced a perfect storm.

The Critique

Within a few years, in the mid-1960s, those who hoped that the federal government would get involved in the problems of poverty were encouraged by the creation of programs like *Head Start*, the expansion of older programs like *Aid to Dependent Children*, the funding of such things as *Sesame Street*, etc. But soon, all this slowed down as the mood in Washington changed and the intellectual grounding of these efforts, as well as a reinterpretation of the political motivations that may have made these programs popular, came under withering criticism. In 1967, Lee Rainwater and William Yancey edited a volume summarizing the first wave of reactions to the Moynihan Report. In 1968, Charles Valentine published a critique that remains the best and most systematic presentation of the anthropological response to Lewis.

Lewis' biographer Rigdon (1988: 109–110), in taking a historical look at Lewis' work and the ensuing critiques has summarized the six flaws most often mentioned:

1. Sampling was not from low-income people, but rather from those Lewis believed would exhibit the traits constituting the culture of poverty.
2. Lewis relied on the concept of 'culture trait,' but his departure deviated from common academic usage of the term and he failed to elucidate a definition to illuminate his idiosyncratic use.
3. Lewis put forward no standard to measure the traits typical of the culture of poverty.
4. Lewis assumed that the culture of poverty was a 'way of life' handed down from generation to generation, but conducted no longitudinal studies to demonstrate this point. Instead, he relied upon interviews with consecutive generations of the same family.
5. Lewis' claims about the transmission of the culture of poverty rely upon a causal connection that his data cannot supply.
6. Lewis, much like Mayhew before him, skewed his sample with exceptional cases and eschewed less dramatic cases that might have challenged his assumptions.

Valentine addressed all these matters, and he also emphasized that Lewis' own data, as presented in *La Vida*, for example, did not support his generalizations (1968: 50–66). This interview data actually provides rich portraits of people who are aware of and engaged with concerns and politics beyond their immediate concerns with the problems of family and friends. The empirical Lewis, in contrast to the theoretical Lewis, demonstrates political involvement and activism as well as a cosmopolitanism his theoretical stance denies.

Valentine also criticized Frazier and, by implication, most research in sociology, psychology, linguistics. First, Valentine (1968: 122) argues that Frazier's depictions of the poor are drawn from second-hand case studies and writings made by policemen, social workers, and other observers with their own professional concerns, not from Frazier's first-hand study. Second, Frazier uses census data and other statistical abstracts as evidence of social disorganization, ignoring the effects of the statistical technique in shaping the data.

Alternatives and Follow-Up

Lewis, like his teachers at Columbia University, was concerned with the relationship of individuals to each other and to the institutions they faced. But there were other ways of conceptualizing this relationship that did not veer into studying psychological internalization of 'cultural traits.' The first came directly out of the initial version of the Boasian tradition. The second had its roots in British social anthropology.

Solon Kimball and Alfred Kroeber, at the 1954 Stanford conference on Anthropology and Education, spoke from the Boasian tradition. They called for a powerful turn in the work of the first generation of anthropologists: that any form of human behavior must be seen in its context of use, and approached in a comparative manner.

The second direction was taken by a group of anthropologists working in England. Most notable were Barnes, who studied social like in a Norwegian fishing village (1954), and Bott who wrote about working-class families in London (1958). Both used social network analysis as the means to study the relation of individual to society. Barnes found that individuals were enmeshed in multiple, overlapping networks of social relations stretching into and out of their respective families. The poor in Barnes' study were neither isolated, nor disorganized, nor uninterested in large issues. Similarly, Bott found that the composition of married couples' external social networks in London influenced their roles within the family, and that these networks offered emotional support not offered within the family roles. Hence, Bott found that poor families were not at all isolated. Neither Barnes nor Bott found the lives of the poor or working class to be any less complex or well organized than those of their wealthier contemporaries.

The 1970s and 1980s saw a surge of such studies, mostly by anthropologists, in various settings, through various methodologies, and from the point of view of different theoretical frameworks. Some of the most notable are Carol Stack (1975), Holloman and Lewis (1978), Shimkin et al. (1978) on black families, Ray McDermott's work on the organization of classroom interaction (1977), Paul Willis on working-class youth (1977), Luis Moll on 'funds of knowledge' (1994). Together, they demonstrated more systematically that there are many ways to being poor, that these ways are tied to matters not under the control of the poor, that the poor can strategize in

order to survive, and that they always 'make sense' if one understood the full context of their actions.

A Continuing Debate

It is not the case, however, that all concerned with the poor were convinced by these demonstrations. Glazer's critique of some initial work on blacks in the South by Davis and Havighurst was taken up by sociologists and social pyschologists, who charged anthropologists with romanticization and broad generalization on the basis of very small and nonrandom samples analyzed in unreliable ways that could not be replicated. While many anthropologists criticize the use of statistics, many in the other disciplines criticized their unwillingness to deal with well-established correlations in large data sets. The clarion call may have been sounded by the sociologist William Julius Wilson who, in 1987, affirmed that "liberals will have to propose thoughtful explanations of the rise in inner-city dislocations. Such explanations should emphasize the dynamic interplay between ghetto-specific cultural characteristics and social and economic opportunities" (1987: 18). By casting the debate as one of liberals versus conservatives, Wilson also reaffirmed the moral imperative behind work on the poor that emphasize dislocation.

From a different, though related point of view, the powerful French sociologist Pierre Bourdieu (1977[1972]) wrote many very influential books about what he called the *habitus* that characterize individuals as they are shaped by their conditions ('structured structures') they keep reproducing in themselves and their children ('structuring structures'), as they misunderstand the mechanisms involved in the reproduction of their social positions, particularly those proceeding through schooling and the media. Bourdieu's work has had a major impact on the American social sciences and remains an inescapable reference point when looking at class inequalities and poverty.

In short order, behavioral scientists from a variety of disciplines presented evidence of the dangers a child face when raised in a household led by a single woman with or without step-parents, people like Garfinkel and McLanahan (1986) and Wallerstein and Blakeslee (1989) presented the kind of evidence that led to a reprise of Frazier and Moynihan's take. Linguists presented again the same kind of evidence used in the 1960s to build a case for child deprivation: some children hear more words, in more complex settings, and are thus better prepared for school (Hart and Risley, 2004). This research is once again the foundation of new policies from the 'No Child Left Behind' act to the call for universal preschools, comprehensive education programs, and other broad interventions. Recent research from alternative perspectives points out deficiencies in the use of concepts like 'cultural characteristics,' 'dislocation,' 'deprivation,' by bringing out the evidence that the poor, in the worst of economic circumstances, can still organize themselves to face what is their actual conditions. Evidence has been brought out about the enslaved Africans in the American South who taught themselves how to read (Gundaker, 2007), the Mexican migrant workers who taught themselves English (Kalmar, 2000), fishermen who learn to fish through the experience of catching, or not catching, fish (Pálsson and Helgason, 1998), illiterate parents who teach their children to read (Ranciere, 1991), and the near ubiquitous experience of teaching oneself how to use a computer.

In the long run, it might make the most sense to consider these controversies as a long conversation among mostly well-meaning people seeking to help people in difficulty. The imperative to help will remain, as will the uncertainties of how to do it in such a way that it does not make matters worse. Debates will be continually waged on how to provide the analytic frameworks and empirical studies that might explain the conditions of poverty, and contribute to an assessment of the efficacy of various policies.

See also: Anthropological Approaches to the Study of Education: The United States and Beyond; Boas, Franz (1858–1942); Early Emotional Development and Cultural Variability; First Language Acquisition, Developmental Psychology of; Poverty Law: United States; Poverty Policy; Poverty in History; Poverty: Measurement and Analysis; Special Education in the United States: Legal History; Urban Poverty in Neighborhoods; Values Across Cultures, Development of; Welfare and Education.

Bibliography

Abu-Lughod, Lila, 1991. Writing against culture. Santa Fe, NM. In: Fox, R. (Ed.), Recapturing Anthropology.

Barnes, John A., 1954. Class and committees in a Norwegian Island Parish. Human Relations 7 (1), 39.

Bereiter, Carl, Engelmann, Siegfried, 1966. Teaching Disadvantaged Children in the Pre-school. Prentice Hall.

Bernstein, Basil, 1962/1971. "Social Class, Linguistic Codes and Grammatical Elements". In His Class, Codes and Control. I: Theoretical Studies toward a Sociology of Language, pp. 73–92. Routledge and Kegan Paul, London.

Bernstein, Basil, 1965/1971. "A Socio-linguistic Approach to Social Learning: With Some Reference to Educability". In His Class, Codes and Control. I: Theoretical Studies toward a Sociology of Language, pp. 25–61. Routledge and Kegan Paul, London.

Boas, Franz, 1887. The occurrence of similar inventions in areas widely apart. Science 9 (224), 485–486.

Boas, Franz, 1938[1911]. The Mind of Primitive Man. The Free Press, New York.

Bott, Elizabeth, 1958. Urban Families: Conjugal Roles and Social Networks. Aldine-Atherton.

Bourdieu, Pierre, 1977. Outline of a theory of practice. In: Goody, Jack (Ed.), Cambridge Studies in Social Anthropology, vol.16. Cambridge University Press, Cambridge.

Coleman, James S., 2007[1966]. Equality of Educational Opportunity Study 1966. Inter-university Consortium for Political and Social Research, Ann Arbor, MI.

Davis, Allison, 2009[1941]. Deep South: A Social Anthropological Study of Caste and Class (Southern Classics). University of South Carolina Press.

Deutsch, Martin, 1960. Minority Group and Class Status as Related to Social and Personality Factors in Scholastic Achievement. Society for Applied Anthropology, Ithaca, NY.

Frazier, Edward Franklin, 1966[1939]. The Negro Family in the United States. University of Chicago Press, Chicago.

Furstenberg, Frank F., Cook, Thomas D., Eccles, Jacquelynne, et al., 1999. Managing to Make It: Urban Families and Adolescent Success (The John D. And Catherine T. MacArthur Foundation Series on Mental Health and De), first ed. University of Chicago Press.

Garfinkel, Michael, McLanahan, Sara, 1986. Single Mothers and Their Children: A New American Dilemma. Urban Institute Press.

Geertz, Clifford, 1973[1966]. "Religion as a Cultural System." in His the Interpretation of Culture. Basic Books, New York, pp. 87–125.

Glazer, Nathan, 1966. The Negro Family in the United States. Foreword to the Revised and Abridged Edition of E. Franklin Frazier's. University of Chicago Press, Chicago.

Gundaker, Grey, 2007. Hidden education among African Americans during slavery. The Teachers College Record 109 (7), 1591–1612.
Harrington, Michael, 1997[1962]. The Other America: Poverty in the United States, reprint ed. Scribner.
Hart, Betty, Risley, Todd R., 2004. The early catastrophe. Education Review 77 (1), 100–118.
Himmelfarb, Gertrude, 1971. Mayhew's poor: a problem of identity. Victorian Studies 14 (3), 307–320.
Himmelfarb, Gertrude, 1983. The Idea of Poverty, first ed. Knopf, New York.
Himmelfarb, Gertrude, 1984. The Idea of Poverty: England in the Early Industrial Age. Knopf.
Holloman, Regina, Lewis, Fannie, 1978. The 'Clan': case study of a black extended family in chicago. In: Shimkin, D., Shimkin, E., Frate, D. (Eds.), The Extended Family in Black Societies. Mouton, The Hague, pp. 201–238.
Kalmar, Tomás Mario, 2000. Illegal Alphabets and Adult Biliteracy: Latino Migrants Crossing the Linguistic Border, first ed. Routledge.
Katz, Michael, 2011. The Undeserving Poor. Pantheon.
Leacock, Eleanor Burke, 1971. The Culture of Poverty: A Critique. Touchstone.
Lewis, Oscar, 1943. The Effects of White Contact upon Blackfoot Culture, with Special Reference to the Effects of the Fur Trade (Ph.D. dissertation). Columbia University.
Lewis, Oscar, 1960. Tepoztlán: Village in Mexico (Case Studies in Cultural Anthropology). Harcourt College Publishers.
Lewis, Oscar, 1966. The culture of poverty. Scientific American 215, 19–25.
Malthus, Thomas Robert, Appleman, Philip, 2003[1798]. An Essay on the Principle of Population, second ed. W.W. Norton & Company (Norton Critical Editions). Second ed.
McDermott, Ray, 1977. Social relations as contexts for learning in school. Harvard Educational Review 47 (2), 198–213.
McLanahan, Sara, 1994[1986]. Growing up with a Single Parent: What Hurts, What Helps. Harvard University Press, Cambridge, MA.
Moll, Luis C., Amanti, Cathy, Neff, Deborah, Gonzalez, Norma, 1992. Funds of knowledge for teaching: using a qualitative approach to connect homes and classrooms. Theory into Practice 31 (2), 132–141.
Moynihan, Patrick, 1965. The Negro Family: The Case for National Action. Office of Policy Planning and Research United States Department of Labor.
Ortner, Sherry B., 1999. The Fate of 'culture': Geertz and beyond. University of California Press, Berkeley.
Pálsson, Gísli, Helgason, Agnar, 1998. Schooling and Skipperhood: the development of dexterity. American Anthropologist 100 (4), 908–923.
Parsons, Talcott, Shils, Edward (Eds.), 1951. Toward a General Theory of Action, first ed. Harvard University Press, Cambridge, MA.
Rainwater, Lee, Yancey, William, 1967. The Moynihan Report and the Politics of Controversy: A Trans-action Social Science and Public Policy Report. M.I.T. Press.
Rancière, Jacques, 1991[1987]. The Ignorant Schoolmaster: Five Lessons in Intellectual Emancipation (K. Ross, Trans.). Stanford University Press.
Rancière, Jacques, 2004[1983]. In: Parker (Ed.), The Philosopher and His Poor. Duke University Press Books (John Drury, Corinne Oster, Trans.).
Redfield, Robert, 1930. Tepoztlan: A Mexican Village a Study of Folk Life, first ed. The University of Chicago Press.
Rigdon, Susan, 1988. The Culture Facade: Art, Science, and Politics in the Work of Oscar Lewis, first ed. University of Illinois Press, Urbana.
Shimkin, Demitri, Louie, G., Frate, D., 1978. The black extended family: a basic rural institution and a mechanism of urban adaptation. In: Shimkin, D., Shimkin, E., Frate, D. (Eds.), The Extended Family in Black Societies. Mouton, The Hague, pp. 25–147.
Stack, Carol, 1975. All Our Kin: Strategies for Survival in a Black Community. Harper and Row, New York.
Trouillot, Michel-Rolph, 2003. Global Transformations: Anthropology and the Modern World, first ed. Palgrave Macmillan, New York.
Tylor, Edward Burnett, 2010[1871]. Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Art, and Custom (Cambridge Library Collection – Anthropology), first ed. Cambridge University Press, Cambridge.
Valentine, Charles A., 1968. Culture and Poverty: Critique and Counter-Proposals. Univ of Chicago Press, Chicago.
Wallerstein, Judith, Blakeslee, Judith S., 1989. Second Chances: Men, Women, and Children a Decade after Divorce. Ticknor & Fields, New York.
Willis, Paul, 1981. Learning to Labor: How Working Class Kids Get Working Class Jobs, morningside ed. Columbia University Press, New York.
Wilson, William Julius, 2009. More than Just Race: Being Black and Poor in the Inner City. W.W. Norton, New York.
Wilson, William Julius, 2012[1987]. The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy, second ed. University of Chicago Press, Chicago.

Culture Policy Regimes in China

Yue Zhang, University of Illinois at Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Culture policy regimes in China have gone through remarkable changes in the past few decades as a reflection of the political, economic, and social transformation of the country. This article reviews the major changes in China's culture policy regimes and elaborates the underlying driving forces of change. It addresses the issue of intellectual property rights and reveals that the complex administrative apparatus prohibits the enforcement of patent and copyright laws. It discusses the rapid development of cultural industries as a strategy for the government to enhance China's global image and promote economic growth. Despite the trend of privatization and increased autonomy of cultural professionals, this article demonstrates that the Chinese state still maintains effective control of cultural production, and therefore largely constrains creative freedom in the society.

Culture policy regimes in China have gone through remarkable changes in the past few decades. From the 1950s to the 1970s, culture policy was not a separate policy arena in China, but was instead a subordinate of politics and served ideological purposes. With the shift of the government's focus from class struggle and political campaign to economic development, culture policy has gained an independent status. In recent years, culture policy regimes in China have been characterized by privatization, and the development of cultural industry has increasingly become an important policy agenda for both the national and local governments. Despite the trend of privatization and enhanced autonomy enjoyed by cultural professionals, the Chinese state maintains effective control over cultural production; therefore the tension between political control and creative freedom remains a critical issue in China's culture policy arena.

Culture Policy: From the Subordinate of Politics to a Separate Realm

For nearly 30 years after the Chinese Communist Party (CCP) came to power in 1949, there was no independent culture policy in China. Culture policy was not recognized by the government as a separate policy area, but rather as a subordinate of proletarian politics and a tool for reinforcing communist ideology. Meanwhile, the Chinese Communists adopted the Soviet system for the arts after 1949. Such a system constituted a complete mobilization and total control of the arts allowing no room for private or unofficial art (Croizier, 1999). In the frantic years of the Cultural Revolution (1969–79), culture was used to assist political campaign, and culture policy became extremely ideological. A large number of cultural and art professionals were persecuted because their works were condemned by the authorities as inconsistent with the reigning political ideology (Mao, 2008).

The situation began to change with the end of the Cultural Revolution. From the late 1970s, the government changed its focus from class struggle to economic development. With the launch of economic reform and opening up in 1979, the government began to adopt a more open attitude toward arts and culture. In October 1979, the Fourth Representative Meeting of Cultural and Art Professionals (中国文化艺术工作者第四次代表大会) was held in Beijing. At the meeting, Deng Xiaoping, the leading architect of China's economic reform, gave a speech on behalf of the CCP. In the speech, Deng claims that slogans such as 'culture and art are subordinate to politics' should be abandoned. He emphasizes that CCP's leadership role in cultural affairs should concentrate on providing favorable conditions to cultural and art professionals and thus nurturing the development of culture (Mao, 2008). This meeting is considered the turning point in the development of China's culture policy regime, as it symbolizes the separation of culture from politics and demonstrates the party's recognition of culture policy as an independent realm.

Despite the more open attitude toward art and culture, the Chinese government did not give up its control over the cultural realm. In 1981, the CCP Central Committee announced the 'Decisions on News Media', which emphasizes the CCP's desire to eliminate cultural products that spread 'incorrect thoughts' in society. In the same year, the Department of Propaganda of the CCP Central Committee held a nationwide meeting to criticize the cultural products that were inconsistent with socialist ideology and challenge the leadership of the Party. In 1983, Deng gave a speech titled 'The Urgent Tasks of the Party on the Ideological Front', in which he stresses the importance of reinforcing the leadership of the Party in culture and arts with a particular focus on areas like education, news media, publication, broadcasting, and television. In 1986, he published another speech titled 'Firmly Fighting against Bourgeois Liberalization'. This speech stresses the importance of fighting against bourgeois liberalization and spiritual pollution, warns against the danger of copying Western democracy, and emphasizes the necessity of the CCP's leadership and socialist ideology.

While the Chinese state is wary of Western influence and keeps a close eye on the development of culture and arts, it has separated culture from politics. In connection with the depoliticization of the role of culture, the power over culture policy shifted from the central government to the local government.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10438-6

The first nationwide cultural development plan was proposed in the Eleventh National Five-Year Plan (2006–10). However, this plan does not offer specific solutions to the regional differentiation in China regarding cultural planning. Nor does it provide national level strategies for cultural development. China's culture policy regimes remain largely decentralized and localized. In particular, municipalities have increasingly become the main actor in the making and implementation of culture policy.

Among the various issue areas that the Chinese government is working on, the question of intellectual property rights is one of the most crucial and controversial issues being addressed. China is by far the world's leading producer of pirated goods, including cultural products such as films and books. In recent years, as China becomes a full participant in the international economy, there is increased external pressure on China's enforcement of intellectual property. However, as Mertha (2007) insightfully points out, greater pressure does not lead to better domestic compliance with international norms. Although external pressures may lead to formal agreements in Beijing, resulting in new laws and official regulations, it is China's complex network of bureaucracies that decides actual policy and enforcement. Specifically, the structure of the administrative apparatus that is supposed to protect intellectual property rights makes it possible to track variation in the effects of external pressure for different kinds of intellectual property. As a result, whereas the sustained pressure of state-to-state negotiations has shaped China's patent and copyright laws, it has had little direct impact on the enforcement of those laws. China still has a long way to go in order to better protect intellectual property rights.

Cultural Industry: Between State Monopoly and Privatization

With the depoliticization of culture policy, China launches cultural industries that bring the culture policy regime into a new era of commercialization and privatization. The idea of cultural industries was introduced to China via the epistemic community of international scholars and consultants in the late 1990s, and it has since generated much interest on the domestic policy circuit. In 1998, the Ministry of Culture established the Division of Cultural Industry as the central political agency in charge of cultural industries. In October 2000, the Fifth Plenary Session of the Fifteenth Central Committee of the CCP (zhongguo gongchandang shiwujie wuzhong quanhui 中国共产党十五届五中全会) first mentioned cultural industries in the 'Suggestions for the Tenth National Five-Year Plan'. The document establishes many goals: improving the policies of cultural industries, building the market of cultural products, improving their management, facilitating the development of cultural industries, and promoting the integration between cultural industries and information technology industries. In March 2001, the suggestions were approved by the Fourth Session of the Ninth National People's Congress (quanguo renda jiujie sici huiyi 全国人大九届四次会议) and were included in the Tenth National Five-Year Plan (2001–05). In the same year, the Chinese government published the Blue Book of China's Cultural Industries. It discusses in detail the development of different types of cultural industries, including news media, publishing, movies, and design.

The development of cultural industries is largely driven by two factors. The first is the political consideration associated with the symbolic value of culture, i.e., the desire of the national government to enhance the global image of China and build 'soft power'. The second is the economic concern, i.e., the attempt of both the national and local governments to upgrade labor-intensive manufacturing to knowledge-based industries.

The competition between major world powers is no longer limited to the military or economic realm. In his article entitled *The Rise of China's Soft Power*, Joseph Nye (2005) argues that, in addition to hard power such as military and economic strength, China is also competing with the United States by building its soft sources of power such as culture, political values, and diplomacy. This is demonstrated in the recent government sponsorship of initiatives to teach Chinese language and culture overseas, such as the Confucius Institutes (Ren and Sun, 2012). Nye also points out that China's soft power still has a long way to go in order to match that of the US. For example, China does not have cultural industries like Hollywood that have pervasive global impact. Policymakers in China are fully aware of the imbalance in cultural trade, and they view the development of cultural industries as the right tool to build China's soft power and enhance its global competitiveness.

There is also economic rationale behind the development of cultural industries in China. The widespread support for cultural industries reflects the need for economic upgrading felt by the national and local governments. The Tenth National Five-Year Plan (2001–05) stated that the rationale for promoting 'cultural industries with Chinese characteristics' is to stimulate domestic consumption demand, increase employment opportunities, and restructure the national economy (State Council, 2001). According to Keane (2006, 2007), cultural industries have become a 'supersign' in China, invested with supernatural powers to transform and revitalize the urban economy. Ren and Sun (2012) also point out that many Chinese cities trying to move up the value chain of global production view the development of cultural industries as key to making the leap from 'Made in China' to 'Created in China'. The development of cultural industries would also aid China in the achievement of many other goals such as wealth creation, the reuse of traditional resources, green production, talent renewal, and industrial upgrading.

However, the development of cultural industries poses a problem for the Chinese government. How will the government maintain its control over cultural production while capitalizing on a more market-based cultural economy (Hui, 2006; O'Connor and Xin, 2006)? As Wang (2004) insightfully points out, cultural industries in China comprise at least three subcontexts: the partial privatization of the formerly state-owned cultural sector, the persistence of state monopoly even when it pushes an agenda of commodifying public cultural goods, and the issue of mixed ownership of firms in the cultural sectors. The conflicted role of the state is reflected in the uneasy adoption of the terms 'cultural industries' (wenhua chanye 文化产业) and 'creative industries' (chuangyi chanye

创意产业) in official documents. Shanghai is the only Chinese city that has officially adopted 'creative industries' (Chen and Zhu, 2009), whereas Beijing has clearly rejected 'creative industries', preferring 'cultural industries' (Hui, 2006). O'Connor and Xin (2006) argue that the tensions arising from these terminologies have allowed a renegotiation of the divisions of responsibility from public sector dominated, ideologically charged 'culture' to private sector–oriented 'creativity'. The ambiguous stance of the state can also be glimpsed in key government policies, which invariably push for further privatization of the cultural sector while maintaining the monopoly of state-owned firms (Ren and Sun, 2012).

The tension between state monopoly and privatization of the cultural sector is just one defining feature of the development of cultural industries in China. Other critical issues associated with the development of cultural industries are land acquisition and real estate speculation. Ren and Sun (2012) maintain that the promotion of cultural industries has led to a frantic period of land seizures, which is reminiscent of the development zone fever of the 1990s. Large tracts of land are acquired by local states to develop creative industry parks and clusters. Many such schemes have led to widespread land speculation and involve little in the way of cultural industry development.

A phenomenal example of the development of cultural industry is Songzhuang Town, an eastern suburban district of Metropolitan Beijing (Zhang, 2014). In 2004, the town government proposed a strategy of 'Developing the Town with Culture', and it designated an area of 14.6 km^2 south of the town to be developed into the largest cultural industry district in China. While this plan is helping the local government to increase revenue, it is also bringing about rapid urbanization and drastically changing the spatial and social dimensions of the area. To implement the plan of developing cultural industries, villagers' land is taken over by the town government. The landownership is changed from collective to public and thus becomes eligible for redevelopment. There are two types of landownership in China after 1949, public landownership in urban areas and collective landownership in rural areas. The town government only controls the urban land due to the public landownership. The villagers collectively own the rural land and they often defer to the village committee when making critical decisions about the land. Therefore, in order to make villagers' land eligible for redevelopment, the first step for the town government is to obtain the landownership, i.e., changing it from collective landownership to public landownership. With the change of the landownership from collective to public, the land status is consequently transformed from rural to urban. For more discussion on this topic, see Ho (2000, 2008). Most of the old houses are demolished and villagers are relocated to high-rise apartment buildings elsewhere in the town. A total of 12 villages, 17 995 villagers are affected by the development plan. After the clearance of the site, various types of creative industries and facilities are constructed, including art studios and galleries, animation industry, and movie production. What is happening in Songzhuang raises many concerns regarding the long-term, sustainable development of the area and the well-being of the displaced peasants.

Art Districts: The Control of Art Community and the Creation of New Cultural Space

An important aspect of the culture policy regime is how the government controls the art community and creates cultural space. Artists have long been marginalized in China since their ideas often conflict with the mainstream political ideology. In major Chinese cities, artists often live on the fringe of society in 'artist villages', where they almost always face the threat of being displaced due to political decisions or urban renewal. Yuanmingyuan Artist Village, the first underground artist village in China, was located at the western suburb of Beijing adjacent to major universities. Emerging in the late 1980s, the artist village was dismantled by police in 1995 in the name of improving public safety. In recent years, however, the Chinese government adopted a different policy toward the art community. Instead of demolishing the underground artist villages, the government designated them as art districts. The change is largely driven by the attempt of the government to enhance its image globally and to promote the development of cultural industry. This section explains the formation of the most phenomenal art district in China, Factory 798. While Factory 798 is unique due to its prominent national and global status, the development trajectory of the area is typical among many art districts in China.

Factory 798 is the largest segment of 718 Joint Factory, also known as North China Wireless Joint Equipment Factory, located in Chaoyang District in the northeast of Beijing. This huge, secretive factory complex was built under Chairman Mao in the 1950s in order to produce advanced technology for the Chinese military. The factory complex was designed by Eastern German architects in the Bauhaus style, and the construction was supported by Chinese and Soviet funds (Huang, 2008). As the biggest production site of electronic components in Asia, the factory was considered in the early years of the People's Republic as a showcase of China's achievement in industrialization and modernization as well as a symbol of international socialist cooperation.

Factory 798 began to decline in the 1980s with China's economic reform and the deindustrialization of Beijing. The economic reform was launched in the late 1970s and one of the key agendas was to promote the development of the private sector and reduce the share of state-owned enterprises (SOEs) in the Chinese economy. Deindustrialization took place in Beijing in the 1990s and precipitated the closure of many factories in the city. In 2000, Factory 798 was downgraded from a state-level SOE to a municipal-level SOE, named Beijing Seven Stars Group, specialized in high-tech electronics industry (Huang, 2008). By 2001, Seven Stars had closed several factories and laid off around 15 000 of its 20 000 workers (Currier, 2008). In order to provide pensions to the workers, it began to lease the empty factory space for an additional revenue.

Some independent artists moved into the idle workshops of Factory 798 after the Yuanmingyuan Artist Village was dismissed by the government. They were attracted to the place by the huge space and the low rent. The Bauhaus-styled workshops have enormous, cathedral-like spaces with swooping arched ceilings. The spacious interior, the natural light streams, and the succinctness of this 'form follows

function' design make the workshops a perfect place for art studios and exhibition halls. The legacies from the socialist era remain in the workshops: rusted machines are standing in the corners; gigantic slogans such as 'Long Live Chairman Mao' printed in red are still on the wall. The visual effect is stunning when avant-garde artworks are exhibited in the obsolete Maoist factory setting.

The number of art galleries and studios rapidly increased in Factory 798. A coincidental factor that contributes to the growth of the art community at Factory 798 is the relocation of the Central Academy of Fine Arts (CAFA). To expand its campus, the CAFA moved out of the city center in 1995 and relocated to Wangjing 望京, an area next to the East Fourth Ring Road just west of Factory 798. The proximity of CAFA brings many art professors and students to this area. See Sun (2010). The area began to draw lots of attention from the media and international society. It has become not only an icon of contemporary Chinese arts, but also a symbol of Beijing as a global city. In 2003, *Newsweek* ranked Beijing as the 12th World City, with the existence of the art community at Factory 798 being considered a significant contributor to Beijing's global status. In the same year, *Time Magazine* selected Factory 798 as one of the Top 22 Most Vibrant Art Districts in the world. Many foreign officials have visited Factory 798 over the years. Viviane Reding, the European Union's Minister of Culture, cited the area as evidence that China was opening up, much to the delight of Chinese politicians (Currier, 2008).

Despite the vigorous growth of the art community, Seven Stars planned to raze the area and build a high-tech park in order to make more profit. The leases with most artists lasted for only 3–5 years and they were supposed to be ended in 2005. In 2003, Seven Stars stopped leasing new space to artists and began to prepare for the demolition. To prevent the area from being demolished, the artists launched a campaign, and their activities were widely covered by the media. In 2004, Factory 798-based artist Li Xiangqun, also a professor at the Central Academy of Fine Arts and a representative in the People's Congress of Beijing, raised a motion to the municipal government to preserve the factory as an art district (Sun, 2010). After a lengthy debate and long wait, the motion was eventually supported by the municipal government.

As Zhang (2014) points out, the national government played a pivotal role in preserving the factory from demolition. Li Changchun, one of the nine members of the Politburo Standing Committee, had a direct impact on the decision. As China's 'propaganda chief', Li is in charge of issues related to publicity, including arts, media, and publishing. When asked to offer an opinion on the case of Factory 798, Li said only nine characters – '看一看，论一论，管一管' (kan yi kan, lun yi lun, guan yi guan) – which means in English have a look, have a discussion, and have a regulation. The concise yet vague statement shows that while the government is inclined to protect the art district from demolition, it does not have a clear plan on how to deal with the area. The national government's support for the preservation of the art district is largely due to the Olympics. After the successful Olympic bid, the government felt the urgency of presenting a more open image worldwide, and the art district provided a stage for the fulfillment of this goal.

In 2006, Factory 798 was officially designated by Beijing Municipality as a Cultural Creative Industry Cluster (wenhua chuangyi chanye jijuqu 文化创意产业积聚区). To supervise the activities in the art district, a management office was created in 2006 under the direction of the Beijing Municipal Department of Propaganda. The office monitors and censors art exhibitions and activities taking place in the art district to ensure they are in accordance with the mainstream political ideology. One major action taken by the office was the decision to ban the Dashanzi International Art Festival (Dashanzi guoji yishujie 大山子国际艺术节), the first and only independent art festival in Beijing run by artists in Factory 798 for 3 years. The office launched the 798 Art Festival and claimed it was the only legal art festival in the art district. By depriving the right of the artists to organize their own art festival, the office reduced the chances for the artists to mobilize resources and social forces, thus limiting their potential challenge to the political regime.

The power of land leasing is in the hands of Seven Stars. Through a land leasing office, the SOE exercises its complete discretion on issues like whom to admit as tenants and at what price. Given the fact that Seven Stars gave up the more profitable high-tech park plan to accommodate the city's demand to build the art district, this arrangement is a way for the city to financially compensate the SOE. Without any governmental regulation, land leasing in the art district becomes a means for the SOE to gain profit, so the leasing price rocketed. Accelerated by the unregulated subcontracting, the leasing price at Factory 798 increased sevenfold within 5 years, from 0.65 to 5 yuan m^2 per day in 2001 and 2006, respectively. Many artists, including some of the early settlers, cannot afford the rent anymore and moved out of the district.

Over the years, Factory 798 has rapidly developed into a popular site for art and entertainment. The number of art galleries and studios increased from 22 to 398 in 2003 and 2008, respectively, exceeding many art districts in the world. Taking the art districts in New York City as an example, both SoHo and Chelsea are well-established art districts with much longer history than Factory 798, but the number of art galleries and cultural institutions in those two areas does not exceed that of Factory 798. There are about 300 art institutions in SoHo and 200 in Chelsea. See Molotch and Treskon (2009). Some internationally renowned art institutions, such as the Guy & Myriam Ullens Foundation and Pace Gallery, opened their branches in Factory 798, thus enhancing the global status of the district. Besides art exhibition, a variety of commercial activities take place in the art district, including company opening ceremonies, product advertising and marketing events, fashion shows, and movie premieres. Hosting these activities increases the income for the land leasing office and individual gallery owners, but it causes a decrease in the space available for art exhibitions. Meanwhile, Factory 798 has become a hot tourist spot in Beijing. The number of visitors to Factory 798 has tripled in 3 years, from 450 000 to 1 500 000 in 2005 and 2007, respectively. During the 2008 Olympics, the area drew 10 000 visitors per day. See Sun (2010). Some tourist guide books list it as one of the 'Three Must Sees' in Beijing, along with the Forbidden City and the Great Walls. All sorts of amenities, including restaurants, cafes, bars, and boutique

shops, are booming in the area, transforming Factory 798 into a tourist bubble.

Chinese artists' associational life is a tricky issue. According to the Chinese Constitution, social members are entitled to organize in associations; however, the key personnel in charge of civic organizations are usually governmental officials, either officials who are still in office or those who recently retired. Therefore, instead of promoting civic liberty, the organizations help the government implement its control over civil society. For example, a nongovernmental organization, Songzhuang Association for the Promotion of Arts (Songzhuang yishu cujinhui 宋庄艺术促进会), was created in Songzhuang Cultural Industry District in 2005. The organization is registered as a civic organization whose major mission is to provide social services to artists. However, the key personnel of the association are local governmental officials, and the real function of the association is to monitor the activities of artists. It is under such mechanism that the associational life of Chinese artists does not counterbalance the dominance of the state, but rather reinforces political control.

An Unsolvable Conflict: Political Control vs Creative Freedom

A critical issue facing the cultural realm in China is the tension between political control and creative freedom. As in many authoritarian regimes, political censorship is pervasive in China. Creative imagination and content are subjugated to active state surveillance. According to Wang's report of the censors' latest activities (2004), distributors of 'cultural products' via the Internet have required state licences since 1 July 2003; content censorship of imported video games is to be strengthened; and, finally, only 10 firms approved by the state can now run National Internet Cafe Chains. These state-sanctioned chains will soon push independent operators out of the market. The setback to the Internet in recent years – the medium theoretically most immune to boundary policing and centralized monitoring devices – is highly symptomatic of China's ongoing problems with censorship and creativity.

In 2005, the National People's Congress coined the slogan of 'Building a Harmonious Society' (jianshe hexie shehui 建设和谐社会). As a reflection of New Confucianism, the slogan emphasizes the importance of maintaining balance and harmony at the societal level. However, this slogan is criticized by independent intellectuals and social activists as empty rhetoric that the Chinese government utilizes to justify its suppression of discourse differing from its own and to silence the voices of Chinese society's discontents. See http://www.newcenturynews.com/Article/gd/201107/20110714135748.html. Widely applied to almost every aspect of social life, the slogan is used by the leaders as a principle to evaluate art and cultural production. They contend that art production should reflect the 'bright side of life' and contribute to the construction of a 'Harmonious Society'. By incorporating cultural production into the scheme of 'Building a Harmonious Society', the government not only exercises control over creative freedom, but also employs creative freedom as a device to maintain social and political stability.

Under the scheme of 'Building a Harmonious Society', the government strictly controls the channels of art publication, exhibition, and circulation. In her study of contemporary Chinese arts, Zhang (2014) discovers that Chinese artists seem to enjoy more autonomy in art production than their Eastern European counterparts in the communist era. Chinese artists can address political issues in their works as long as they do not go beyond the 'redline', i.e., not to directly condemn the CCP or its major leaders such as Chairman Mao and President Hu. The secret to the artists' autonomy, Zhang (2014) argues, lies in the fact that the government effectively controls the channels of art publication, exhibition, and circulation. In other words, although artists can produce politically controversial artworks, there are very few chances for them to circulate the works domestically so that they have little impact on Chinese society. Given that the channels of art publication, exhibition, and circulation are under strict state control, many artists have to send their work overseas in order to reach a broader audience (Li and White, 2003; Barmé, 2000). As a result, most politically charged works become 'quarantined domestically' and are intended 'for export only'.

Besides censorship, another approach for the Chinese government to control creative freedom is cooptation. The most important institution to employ cooptation is the China Artists Association (Zhongguo meishujia xiehui 中国美术家协会). Created in 1949, this is a single state-controlled artists' association to which all official artists belonged. This is one of the key features of the Stalinist system for the arts, which the CCP adopted from the beginning of its reign. This association has played a crucial role in organizing and controlling Chinese artists.

A more recent example of how the state uses the strategy of cooptation to influence the cultural professionals is the creation of the Chinese Academy of Contemporary Arts (Zhongguo dangdai yishu yuan 中国当代艺术院). Contemporary Chinese arts burgeoned in the 1980s with China's economic reform and opening up. Given the politically sensitive aspect of much contemporary artwork, contemporary Chinese arts have not been formally recognized by the Chinese government for a long time, and the artists are generally viewed by the political authority as troublesome. In 2009, the Chinese Academy of Contemporary Arts was founded by the government as the first national institute of contemporary arts. It is a branch under the Chinese National Academy of Arts (Zhongguo yishu yanjiu yuan 中国艺术研究院), the only national research institute of art. Supervised by the Publicity Department of the CPC Central Committee and the Ministry of Culture, the Academy of Contemporary Arts functions like a semigovernmental agency. According to the official Web site of the Academy, the mission of the institution is to promote research, education, and production in the field of contemporary Chinese arts. See http://www.zgysyjy.org.cn/newart/lanmuye.jsp?class_id=13_03_08. With the creation of the Academy, 20 leading contemporary Chinese artists (for a list of the names of the artists, see http://www.zgysyjy.org.cn/newart/yishuchuangzuo9.jsp) were appointed as academicians, some of whom are well known for their politically critical works. The creation of the Academy of Contemporary Arts and the entitlement of the artists seem to be a major step taken by the Chinese government to enhance the political and social status of the art community. However,

many artists view the initiatives as a state strategy to co-opt the leading artists and to control the entire art community through them (Zhang, 2014).

Major Characteristics of China's Culture Policy Regimes

Culture policy regimes in China have changed drastically over the past decades as a reflection of the political, economic, and social transformation of the country. The current culture policy regimes in China have three major characteristics. First of all, culture policy is separated from politics as an independent policy arena, but cultural production and creative freedom are not free from politics. They are still subjugated to active state surveillance. Second, the development of a cultural industry and a market-based cultural economy are important goals that China is pursuing at both the national and the local level. However, these efforts are frustrated by the tensions that naturally emerge between the state monopoly and the privatization of the cultural sector. Third, globalization plays a significant role in shaping China's culture policy regimes. Global events, such as the Olympic Games, provide incentives for the Chinese state to adopt a more supportive and open attitude toward the cultural sector in order to enhance its global image. While China has become the most prominent rising power in the world, it still has a long way to go in order to build a vital culture policy regime that truly supports the development of its arts and culture.

See also: Cultural Policies in East Asia; Culture, Sociology of; Leisure and Cultural Consumption: The European Perspective; Leisure and Cultural Consumption: US Perspective.

Bibliography

Barmé, Geremie R., 2000. Artful marketing: who buys it? Contemporary Chinese art at home and abroad. Persimmon 1 (1), 23.

Chen, Shaofeng, Zhu, Jia, 2009. Zhongguo Wenhua Chanye Shinian [Ten Years of Cultural Industries in China]. Great Wall Press, Beijing.

Croizier, Ralph, 1999. The avant-garde and the democracy movement: reflections on late Communism in the USSR and China. Europe-Asia Studies 51 (3), 483–513.

Currier, Jennifer, 2008. Art and power in the New China: an exploration of Beijing's 798 district and its implication for contemporary urbanism. Town Planning Review 79 (2–3), 237–265.

Ho, Peter, 2000. Contesting rural spaces: land disputes, customary tenure and the state. In: Perry, Elizabeth J., Selden, Mark (Eds.), Chinese Society: Change, Conflict and Resistance. Routledge, New York.

Ho, Peter, 2008. Land markets, property and disputes in China. In: Spoor, Max (Ed.), The Political Economy of Rural Livelihoods in Transition Economies: Land, Peasants and Rural Poverty in Transition. Routledge, London.

Huang, Rui, 2008. Beijing 798: zai chuangzao de "gongchang" (Beijing 798: Reflections on "Factory" of Art). Sichuan Fine Art Press, Sichuan.

Hui, Desmond, 2006. From cultural to creative industries: strategies for Chaoyang district, Beijing. International Journal of Cultural Studies 9 (3), 317–331.

Keane, Michael, 2006. From made in China to created in China. International Journal of Cultural Studies 9 (3), 285–296.

Keane, Michael, 2007. Created in China: The Great New Leap Forward. Routledge, Curzon, London.

Li, Cheng, White III, Lynn T., 2003. Dialogue with the West: a political message from avant-garde artists in Shanghai. Critical Asian Studies 35 (1), 59–98.

Mao, Shaoying, 2008. "Zhongguo Wenhua Zhengce Sanshi Nian" [Thirty Years of China's Culture Policy]. http://www.ccmedu.com/bbs35_75790.html.

Mertha, Andrew C., 2007. The Politics of Piracy: Intellectual Property in Contemporary China. Cornell University Press, Ithaca and London.

Molotch, Harvey, Treskon, Mark, 2009. Changing art: SoHo, Chelsea and the dynamic geography of galleries in New York City. International Journal of Urban and Regional Research 33 (2), 517–541.

Nye, Joseph, December 29, 2005. The rise of China's soft power. Wall Street Journal Asia.

O'Connor, Justin., Xin, Gu, 2006. A new modernity? The arrival of 'cultural industries' in China. International Journal of Cultural Studies 9 (3), 271–283.

Ren, Xuefei, Sun, Meng, 2012. Artistic urbanization: creative industries and creative control in Beijing. International Journal of Urban and Regional Research 36 (3), 504–521.

State Council, 2001. Wenhua Chanye Fazhan Dishige Wunian Jihua Gangyao [The Tenth Five-year Plan of Cultural Industries Development]. State Council, Beijing.

Sun, Meng, 2010. The Production of Art Districts and Urban Transformation in Beijing (Ph.D. Dissertation). The University of Illinois at Chicago.

Wang, Jing, 2004. The global reach of a new discourse: how far can 'cultural industries' travel? International Journal of Cultural Studies 7 (1), 9–19.

Zhang, Yue, 2014. Governing art districts: state control and cultural production in contemporary China. The China Quarterly, forthcoming.

Relevant Websites

http://www.caanet.org.cn/ – China Artists Association.
http://www.zgysyjy.org.cn/ – Chinese National Academy of Arts.
http://www.ccnt.gov.cn/sjznew2011/whcys/ – Division of Cultural Industry, Ministry of Culture.
http://www.798art.org/index.html – Factory 798 Art District.
http://www.ccnt.gov.cn/ – Ministry of Culture of the People's Republic of China.

Culture, Production of

Richard A Peterson, Vanderbilt University, Nashville, TN, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 5, pp. 3167–3173, © 2001, Elsevier Ltd.

Abstract

The production perspective focuses on the ways that symbolic elements of culture are shaped by the systems within which they are created, distributed, evaluated, taught, and preserved. The substantive topics have most often been systems in which art works, musical styles, literature, scientific research reports, religious celebrations, legal judgments, media programing, and news reports are made. In addition, the perspective has been applied to situations where the manipulation of symbols is a by-product rather than the goal of collective activity. The perspective is grounded in the constructivist view of culture bounded by the institutionalized systems of production. A widely applicable hypothesis is that the greater the competition between culture producers, the greater will be the diversity of what is produced. Key topics include the differences between 'high' and 'low' forms, cultural mobility between these, the changing criteria of prestige, the autoproduction of culture, cultural entrepreneurship, the fabrication of authenticity, and the shaping of collective memory.

The production of culture perspective focuses on the ways in which the content of symbolic elements of culture are significantly shaped by the systems within which they are created, distributed, evaluated, taught, and preserved. Practitioners of this perspective are most likely to focus on the fabrication of expressive-symbol elements of culture such as art works, scientific research reports, religious celebrations, legal judgments, television programing, and newsmaking which, in turn, embody, modify, and give concrete expression to the 'norm,' 'value,' and 'belief' elements of culture. The perspective has also been successfully applied to a range of quite different situations where the manipulation of symbols is a by-product rather than the goal of the collective activity (Crane, 1992; Peterson, 1994).

Two early studies show the promise of what later came to be called the production perspective. First, in their study of the emergence of impressionist art in nineteenth-century France, Harrison and Cynthia White (1965) found inadequate the then prevalent theories that said changes in art are associated with revolutionary changes in society or with the emergence of persons of genius. Instead they found that the older Royal academic art production system survived the economic turmoil and ideological changes of the French Revolution but collapsed a generation later, because of the emerging art market created by Parisian art dealers and critics who fostered unconventional artists such as the Impressionists.

Second, noting the set of nineteenth-century American novels that traced the battles of lone individuals against nature, novels ranging from the frontiersman stories of James Fenimore Cooper to Herman Melville's saga of the titanic struggle of a man against a white whale, Moby Dick. Literary critics and Americanists had long asserted that these works evidenced America's distinctive culture of individualism and entrepreneurship. They contrasted such works with the novel of manners produced by English authors, which were held to represent English culture. In her careful analysis, Griswold (1981) put this facile generalization in doubt. She found that the English novels of manners sold better in the United States, as well as in England, than the man-against-nature novels. Significantly before 1909, American publishers preferred works by English authors because they could be sold in the US without paying author's copyright fees, while American authors had to be compensated. To sell their work, American authors turned to several specialty topics, including the man-against-nature theme. With the changed copyright law of 1909, American and English authors were put on the same footing, and a rapidly increasing number of American authors successfully published 'English' style novels of manners. This fact that was once read as evidence of 'the closing of the frontier,' Griswold sees as an important result of changes in copyright law and its ramifying influences on the literary production process.

Research in the production perspective draws freely on theories, methods, and concepts developed in other branches of sociology. It is, however, distinctive in focusing on the consequences of social activities for culture. Thus, for example, the 'newsmaking' studies of the 1970s exemplify the production perspective because their authors were not satisfied to simply deconstruct the content of news or to trace the organizational and occupational dynamics of news organizations. In addition, they highlight the consequences of these for the formation and transformation of 'news' (cf Molotch and Lester, 1974; Tuchman, 1978; Gans, 1979).

Establishing the Production Perspective

The production perspective emerged in the 1970s largely in response to the failure of the earlier dominant idea that culture and social structure mirror each other. This 'mirror' or 'reflection' view had been developed by anthropologists who studied small preliterate groups, where it seemed credible that a self-consistent set of social institutions would be reflected in a cognate coherent set of cultural symbols. The idea that there is a symbiotic relationship between a singular functioning social system and its coherent overarching culture was freely embraced by most theorists of contemporary society as well, including in the 1950s and 1960s most Marxists and functionalists such as Talcott Parsons. The former asserted that

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10456-8

those who controlled the means of producing wealth could shape culture to fit their own class interests. The latter believed that a set of monolithic abstract values were more long lasting and determined the shape of social structure.

The reflection view proved to be of limited use, however, in understanding the relationship between social structure and culture in industrial societies and was also being widely questioned by anthropologists studying preliterate societies. Breaking with the older mirror view, the production perspective – like most of the other contemporary perspectives in cultural sociology – views both culture and social structure as an emergent patchwork and the relationship between them as problematic (Peterson, 1979).

Publication in 1976 and 1978 of anthologies entitled *The Production of Culture*, edited by Peterson and Coser, respectively, signaled the emergence of the production perspective as a coherent and self-conscious approach in culture studies, claiming to be applicable across a wide range of specific fields and genres. Many of the early researchers in the perspective had been trained in the sociology of organizations, industry, and occupations and brought their skills in the analysis of material production to the fields of symbol production. Others came to the perspective with a love of or commitment to a particular form of cultural expression such as opera, rock, dance, television, or literature. A number of developments since the 1970s show the progressive establishment in sociology of the production of culture perspective. Two textbooks and an anthology have appeared with the title *The Production of Culture* (Crane, 1992; Ryan and Wentworth, 1998; Gay, 1997), and the perspective receives full chapters in recent texts in the sociology of art (Zolberg, 1990) and in the sociology of culture (Hall and Neitz, 1993).

From the outset, the perspective promised not only a way to better understand the shape and form of any particular cultural expression but also to facilitate three kinds of comparisons as follows: (1) comparisons across broad cultural production realms including art, popular culture, religion, science and law; (2) comparisons across the division between 'elite' and 'common' that tends to emerge in each realm; and (3) comparisons across societal, institutional, or organizational, and microlevels of analysis (Peterson, 1976). Progress toward achieving these goals has been uneven (Ryan and Wentworth, 1998).

We will review progress and problems in working toward these three goals and point to several emerging lines of inquiry that link the production perspective to other approaches in cultural studies, citing studies that exemplify the perspective, even when the authors have not formulated their work in production terms. We begin by showing the utility of the perspective by reviewing a few specific lines of research. This article will also briefly note, as appropriate, cognate work in other realms.

The Production System in the Cultural Diamond

Griswold (1994) has argued that to be complete, the study of any cultural phenomenon, be they books, paintings, legal systems, religious precepts, or news stories, must take into account four classes of phenomena. In addition to a close analysis of the cultural works themselves, they also focus on the situation in which people create works, the nature of the audience, and the relevant aspects of the society at large. She sees these as four points on a diamond with society at the top, the cultural work at the bottom, the creators on the left point, and the audience on the right. With this model in mind, the production system can be seen as a square overlaying the diamond, the square being large enough so that its corners intersect each of the sides of the diamond. This addition makes clear that elements of the production system intervene between each of the four facets of the diamond, so that the influence of each on the others is mediated by the production system.

To illustrate the workings of the production system in shaping culture, Peterson's (1990) analysis of why rock music became the dominant form of popular music will be used. In the early 1950s, the music industry was blind to the large and growing unsatiated demand for greater variety in music and was deaf to the efforts of musicians who might have satisfied that demand. The music industry was financially as well as aesthetically committed to the big band-crooner style of popular music of the time, and, because of its oligopolistic control of the production, distribution, and marketing of new music, was able to thwart the marketing of alternative styles. Yet in the mid-1950s, the power of the music industry oligarchy was broken and the rock of Chuck Berry, Elvis Presley, Bill Haley, and Little Richard became the defining force of music in the United States and much of the world. In the middle of the quiescent Eisenhower years, no major political or economic upheaval could account for such change.

Six production factors working together are identified as making possible the rapid cultural change. These include law and regulation, technology, industrial structure, organizational structure, occupational careers, and the consumer market. In brief, with the transfer of network radio programing to the emerging new medium, television, radio turned to playing records as the cheapest effective form of programing. The granting of many new radio station licenses, the arrival of cheap transistor radios, and the development of the top 40 radio-as-jukebox format meant that the listening audience was exposed to a much larger number and far wider range of music. Using the new durable 45 rpm records, and taking advantage of the developing network of independent record distributors, numerous independent record companies experimented with a wide range of new sounds in an effort to tap the unsatiated market demand. In a matter of 24 months between late 1954 and early 1957, rock was forged in this caldron of entrepreneurial creativity (Peterson, 1990; Ennis, 1992). In the decades since, the large media conglomerates have gradually reconstituted their oligopolistic control of the industry, and the market has been divided into an ever larger number of increasingly homogeneous market segments (Lopes, 1992; Negus, 1999).

Using the same six variables employed in the analysis of the emergence of rock music, the production perspective can bring some order to a welter of different trends taking place in commercial music at the turn of the millennium. On the one hand, a few artists are accounting for an ever larger proportion of diminishing aggregate record sales, and at the same time there is an explosion in the number and diversity of artists who are making and independently distributing CD recordings.

To increase their popularity with the masses of sometime music buyers, megastars such as Michael Jackson, Garth Brooks, Celine Dion, the Spice Girls, and Shania Twain make their work ever more safe and predictable. This aesthetic shift is encouraged by the marketing needs of their record companies who are aggregating into ever larger multinational media conglomerates, dependent for promotion on radio stations whose ownership is rapidly consolidating as well. These large firms can work efficiently only if they sell millions of units of a few recordings. The majors have been able to control the market by effectively controlling the channels of marketing and distribution.

The hegemony of this star-marketing system is being challenged by a new system of promotion and distribution that is coming into being. Since the late 1980s, the technology has been available to make possible the digital recording of music and, therefore, the storage, reproduction, and transmission of sound without any decay in quality. Given the increasing power of Internet technology, its maintenance as a free highway of information, and the availability of inexpensive software, making it easy for any musician to create a 'home page' and transmit sound. In this emerging field, independent music producers can make, market, and distribute their work independent of the major media conglomerates.

If this system develops successfully, it can increase the power of artists, reduce the power and profitability of the large firms, foster an explosion in the number of artists and firms able to reach wide audiences, and significantly increase the levels of experimentation in popular music, creating new markets for an explosion in the range of music regularly on offer. It is as likely that, through the elaboration of copyright and censorship law, together with control of the patents for the technology necessary for the new system, the major firms will keep control of developments in popular music.

The Promise of Comparative Analysis

One of the features of the production perspective is that it facilitates comparative analysis of three sorts: comparisons across symbol-producing realms, comparisons between elite and popular forms within each realm, and comparisons across societal, institutional, and microlevels of analysis. To date, most studies in the production perspective have focused on particular fields, and progress toward fulfilling the promised benefits of comparison has been uneven, as we will show.

Comparisons across Symbol-Producing Realms

Mid-twentieth century functionalists argued that cross-realm comparisons are fruitless because each has its own unique core value. Essentially Parsons (1961) argued that science deals in experimentally established fact, law deals in reason, art deals in aesthetics, and religion concerns faith in the supernatural. The nominalist stance of the production perspective, however, problematizes these differences, making the assertion of uniqueness a residual to be asserted when no similarities can be established.

The expectation that the same range of processes are present in all realms of culture production, be they religion, art, science, law, etc., does not mean that new discoveries in physics are produced in the same way as is new religious doctrine. Rather it means that the differences can be explained, in large part, by the conditions of production in each of the realms (Crane, 1976). For example, knowledge in physics today is largely controlled by a single subsidized and self-perpetuating worldwide 'priesthood' that works to allocate resources and rewards in ways that perpetuate its views; at the same time, religious doctrine varies greatly from country to country, and, particularly in the United States, a wide range of religious specialists mu compete with each other in an open market for the lay consumer's attention and for resources, loyalties, and rewards.

These differences, however, are not inherent in the nature of the subject matter of these two realms, as can be seen by reference to earlier times when the production systems of science and religion were very nearly reversed. In the eighteenth century, English scientists supported themselves and their work by periodic shows of spectacular findings that induced laymen to give them further support, and in the seventeenth century, religious scholars worked with the support and protection of an Established Church orthodoxy and defied it only in mortal peril of their lives. Likewise, eighteenth-century French painters worked with the protection of an all-powerful academy, while today they work in conditions of near perfect competition. In each of these instances, the movement from one sort of production system toward the other has had profound consequences for every aspect of the art world, including even what is considered to be art (White and White, 1965; Becker, 1982).

As recent research on the roles of women in each of the symbol-producing realms has shown, much less dramatic changes in the nature of the production process may have major consequences for a particular class of producers. Systematic comparison across realms will highlight the structural and symbolic underpinnings of such institutional discrimination.

Distinctions between High and Low

An invidious distinction between artefacts, practices, and practitioners considered reverentially and those who are considered to be profane is found in each of the culture-producing realms, and this fact furnishes a great opportunity for comparisons both within and across production realms. To date, however, the distinction has done more to inhibit such comparisons because the truth value vested in high (art, science, religion, law) symbols and the honor status accorded to those in the fields where they are produced is often seen to render these fields incomparable with each other and also with the low counterparts in their own realms (Gans, 1974).

Working from the nominalist stance of the production perspective, one can cut through these problems by asserting that any of the invidious high–low distinctions are not inherent in the cultural objects, but rather they involve a social, or better, a cultural construction of reality. At issue are the questions: who gets to make the designations, on what basis are they made, and within what range are they accepted. There are a number of excellent case studies touching on these issues, but there has been very little thinking about the conditions supporting one system of evaluation or another across all the realms of cultural production. The most systematic efforts in

this direction can be found in the works of Pierre Bourdieu and his colleagues. In the paragraphs that follow, two issues are considered: the mobility between high–low categories, and the differentiation and dedifferentiation of high and low.

Cultural Mobility

In the arts, particular works and artists, and even entire genres, experience aesthetic mobility over time (Becker, 1982). DeVeaux (1997) details the sociopolitical and music industry conditions that caused the aesthetic mobility of jazz from folk to popular to fine art music. In the religious realm in the United States, there has historically been a cycle in which the established church organization becomes increasingly rigid and out of touch with people's felt needs, and a number of reformist charismatic groups form to serve the demand. As the established churches loose power, the surviving reform groups become ever more established and churchlike.

Changing Criteria of Prestige

The distinction between popular culture and high art that was such a fixture of American society by the middle of the twentieth century (Gans, 1974) only began to be clearly articulated in the later decades of the nineteenth century. The differentiation of high–low was driven by the class interests of rising status groups, and the criteria for claiming prestige has changed in recent decades (Crane, 1992). There is evidence that the highbrow exclusiveness that was at the core of cultural capital at mid-century has been replaced by a cultural capital based on cosmopolitan, even omnivorous, tastes. Volume 25, numbers 2 and 3 of *Poetics* are devoted to describing this changing representation of status through taste displays.

Micro, Institutional, and Societal Levels of Analysis

The production of a culture model has been formulated at the level of personal interaction, relations among organizations, and the workings of entire societies. Ideally, work at each of these levels complements work at the other two, and while this may happen in the best individual studies, the theoretical questions and research techniques employed at the three levels are different enough that research has tended to develop separately, and practitioners working at different levels of analysis do not learn all they might from each other.

The great strength of studies that explicitly take into account societal level conditions is their view of the culture industry as an expression and reinforcer of the sociopolitical system of the larger society and the attendant systems of domination and subordination, including social class, ethnicity, religion, and gender. Excellent studies can be cited. The danger in looking from the societal level is that the workings of the production process may be seen to follow automatically from the society-level constraints. This need not be the case, as exemplified in a number of empirical works, which show the way in which societal values and beliefs are mediated by the culture industries.

Production studies that focus primarily on the organizational or institutional level take advantage of the decades of studies of particular organizations, occupations, industries, and the theoretical developments in these research specialties. Because researchers want to remain value-neutral, they tend to bracket their own aesthetic and political views and use uncritically the aesthetic judgments and ideological stances implicit in the symbolic field being studied. Thus the work of these researchers helps to perpetuate the established categories of the systems they study. At the same time, their reports expose for public review the workings of the systems, which create and perpetuate self-serving classifications of a given field (cf Lamont and Fournier, 1992).

Production studies that focus primarily on the microlevel of interaction between persons draw from the rich tradition of research in symbolic interactionism, as well as from the social–psychological elements of the sociologies of occupations and organizations. The most wide-ranging and influential work in this tradition is Howard S. Becker's *Art Worlds* (1982). He shows that by bracketing societal and institutional levels of analysis, one can illustrate how groups of symbol producers create and are shaped by the conventions of aesthetic judgment and collaboration in culture production. In effect, such microlevel studies focus less on the production of culture and more on the culture of production (Fine, 1992).

Some microlevel studies can be criticized for unreflexively assuming the perspective of the group or occupational category under review, and thus obtaining an incomplete and misleading reading of the field. This 'going native' as it were is cognate to the problem of becoming an unwitting apologist for the institution under study, and the research strategy of comparison across subfields is one of the best ways of guarding against unreflexively adopting the perspective of just one of the groups in a production system.

The Production Perspective in Cognate Areas

The production perspective has proved a useful model for organizing ideas and research in a number of areas where the production of culture is itself not the conscious aim of the activity.

Reception and the Autoproduction of Culture

Reception by the reader, listener, viewer, or purchaser has always been a central concern in cultural studies. The term 'reception' suggests the end of a communication process. However, reception is an active process of selection, interpretation, and recombination of elements that can be seen as the production of a symbolic world, which is meaningful for 'consumers.' A prime question here is how people go about selecting cultural elements, and in turn, the meanings that they attach to them. This production of a lifestyle for oneself and one's group is best referred to by the Latin word for self – 'auto' thus the autoproduction of culture. The autoproductions of marginalized groups such as youth, ethnics, eccentrics, and the like have received much scholarly attention, but all individuals and collectivities create lifestyles that display their patterning of cultural choices (Peterson, 1983). A growing number of studies focus on how people of different sorts produce meaning out of mass-mediated symbols.

In the 1940s, a 'mass culture' was seen as the unreflexive response to the seductions of the mass media, dissolving the social class-based hierarchy of class cultures by which wage worker, tradesperson, business person, and farmer could readily be distinguished. While the extreme elitist alarm over massification has fewer proponents now (Gans, 1974), much of the debate since has focused on the question of the degree of fit between social class, defined in relationship to production, and culture class defined in relationship to consumption. Gans (1974) and Bourdieu (1984) assert that the fit between social class and culture class (groupings of people sharing patterns of consumption) is quite close. However, their perceptive analyses of the United States and France, respectively, reveal numerous autoproducing status groups that cannot easily be fit into a linear hierarchy.

More recent studies of autoproduction of culture suggest that most people of high status (in the United States at least) are now little concerned with the old class markers of cultural capital that centered on the appreciation of the fine arts. They no longer shun as inferior all forms of popular or mass culture (Goffman, 1951). In their study of the musical tastes of the full range of occupational status groups, Peterson and Kern (1996) suggest that the cultural symbols of class position are undergoing a profound transformation. While the top strata were earlier characterized by an elitist and exclusive interest in the fine arts and avoidance of other cultural expressions, now persons and groups show their high status as cultural omnivores, consuming not only the fine arts but also showing a knowledgeable appreciation of many, if not all, forms of cultural expression. Near the bottom of the status hierarchy, in sharp contrast to the old conception of the undiscriminating couch potato of the mass audience, they find groups who are univores, that is, people involved with a single form of cultural expression.

The systematic study of the autoproduction of culture is going forward rapidly along several fronts: (1) ethnographic studies of distinctive groups tease out the processes by which group members appropriate and incorporate cultural symbols for their own purposes; (2) other studies show the reappropriation of the mass media-disseminated pop culture by youth and other specific taste groups; and (3) population surveys with increasingly sophisticated culture questions are becoming available, making possible a much richer understanding of patterns of cultural choice.

Tradition, Collective Memory, and Boundary Formation

Early anthropologists, folklorists, and sociologists developed the idea of 'tradition' as the repository of all that is ancient and virtually unchanging. This fitted the needs of both romantics seeking the 'noble savage' and imperialists needing a rationale for ethnic and racial exploitation. In recent decades, this view has been attacked on all sides as factually incorrect and self-serving. For example, Hobsbawm (1983) in a study showing that much of the British monarchial tradition was in fact purposefully created in the middle of the nineteenth century, has coined the felicitous term 'invented tradition' to describe this kind of production of culture.

The term 'collective memory' is now often substituted for 'tradition' in recognition that the past is continually reinterpreted to fit the changing needs of the present. Some studies point to specific individuals or groups of memory 'doctors' and 'cultural entrepreneurs.' One of the most exciting applications of the idea of the mutability of collective memory has been in the study of the manipulation of memory in the service of making status distinctions along the lines of gender, race, class, etc. Gusfield (1963), in one of the earliest works of this sort, studied the role of status politics in the efforts of the (old-family Protestant) Women's Self-Temperance League to stigmatize the alcohol-use norms of immigrant Catholics. A wide array of applications of this sort of culture-production study can be found in Lamont and Fournier's (1992) anthology, *Cultivating Differences*.

Prospects

Research perspectives tend to come and go over the years, so what are the prospects for the production perspective? The numerous emerging developments described above show that the production perspective continues to gain utility for significant research and theory development. The perspective has proved an essential part of any full analysis of cultural phenomena, although some practitioners in cultural studies still do not take into account the influence of production factors. The globalization of production and the high rate of technological development as well as changes in censorship law and intellectual property rights make it all the more imperative in the twenty-first century to take into account the changing influence of production factors. Whether under the distinctive designation 'production of culture' or not, future scholars will find it rewarding to ferret out the many ways in which the system of culture production influences what culture is and what it can become.

See also: Art History; Art and Culture, Economics of; Collective Identity; Leisure and Cultural Consumption: The European Perspective; Leisure and Cultural Consumption: US Perspective; Markets: Artistic and Cultural; Popular Culture; Symbolic Boundaries.

Bibliography

Becker, H.S., 1982. Art Worlds. University of California Press, Berkeley, CA.
Bourdieu, P., 1984[1979]. Distinction: A Social Critique of the Judgement of Taste. Harvard University Press, Cambridge, MA.
Crane, D., 1976. Reward systems in art, science and religion. American Behavioral Scientist 19, 719–734.
Crane, D., 1992. The Production of Culture. Sage, Newbury Park, CA.
Crane, D., Kern, R., 1996. Changing highbrow taste: from snob to omnivore. American Sociological Review 61, 900–907.
DeVeaux, S., 1997. The Birth of Be Bop: A Social and Musical History. University of California Press, Berkeley, CA.
DiMaggio, P., 2000. The production of scientific change: Richard Peterson and the institutional turn in cultural sociology. Poetics 28.
Ennis, P.H., 1992. The Seventh Stream: The Emergence of Rocknroll. Wesleyan University Press, Hanover, NH.
Fine, G.A., 1992. The culture of production: aesthetic choices and constraints in culinary work. American Journal of Sociology 97, 1268–1294.
Gans, H.J., 1974. Popular Culture and High Culture. Basic Books, New York.
Gans, H.J., 1979. Deciding What's News. Pantheon, New York.
Gay, P du (Ed.), 1997. Production of Culture: Cultures of Production. Sage, London.

Goffman, E., 1951. Symbols of class status. British Journal of Sociology 2, 294–304.
Griswold, W., 1981. American character and the American novel. American Journal of Sociology 86, 740–765.
Griswold, W., 1994. Cultures and Societies in a Changing World. Pine Forge Press, Thousand Oaks, CA.
Gusfield, J., 1963. Symbolic Crusade: Status Politics and the American Temperance Movement. University of Illinois Press, Urbana, IL.
Hall, J.R., Neitz, M.J., 1993. Culture: Sociological Perspectives. Prentice-Hall, Englewood Cliffs, NJ.
Hobsbawm, E., 1983. Introduction: inventing traditions. In: Hobsbawm, E., Ranger, T. (Eds.), The Invention of Tradition. Cambridge University Press, Cambridge, UK, pp. 1–14.
Lamont, M., Fournier, M. (Eds.), 1992. Cultivating Differences: Symbolic Boundaries and the Making of Inequality. University of Chicago Press, Chicago.
Lopes, P.D., 1992. Innovation and diversity in the popular music industry, 1969–1990. American Sociological Review 57, 56–71.
Molotch, H., Lester, M., 1974. News as purposive behavior. American Sociological Review 39, 101–112.
Negus, K., 1999. Music Genres and Corporate Structures. Routledge, London.
Peterson, R.A., 1976. The production of culture: a prolegomenon. In: Peterson, R.A. (Ed.), The Production of Culture. Sage, Beverly Hills, CA, pp. 7–22.
Peterson, R.A., 1979. Revitalizing the culture concept. Annual Review of Sociology 5, 137–166.
Peterson, R.A., 1983. Patterns of cultural choice: prolegomenon. American Behavioral Scientist 26, 422–438.
Peterson, R.A., 1990. Why 1955? Explaining the advent of rock music. Popular Music 9, 97–116.
Peterson, R.A., 1994. Culture studies through the production perspective. In: Crane, D. (Ed.), The Sociology of Culture. Blackwell, Oxford, UK, pp. 163–190.
Peterson, R.A., Kern, R.M., 1996. Changing highbrow taste from snob to omnivore. American Sociological Review 61, 900–907.
Peterson, R.A., 2000. Two ways culture is produced. Poetics 28 (2/3), 225–233.
Ryan, J., Wentworth, W.M., 1998. Media and Society: The Production of Culture in the Mass Media. Prentice-Hall, New York.
Tuchman, G., 1978. Making News. Free Press, New York.
White, H.C., White, C.A., 1965. Canvases and Careers. Wiley, New York.
Zolberg, V.L., 1990. Constructing a Sociology of the Arts. Cambridge University Press, New York.

Culture, Production of: Prospects for the Twenty-First Century

Jennifer C Lena, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The production-of-culture perspective focuses on the ways in which the symbolic elements of culture are shaped by the systems in which they are created, distributed, evaluated, taught, and preserved. The perspective has flourished in the last 40 years, and scholars now study such diverse phenomena as the impact of market and industry structures on qualities of artistic objects and experiences; the characteristic features of careers, contracts, and status orders; classification and classification systems; the fabrication of authenticity; and fan cultures, taste, and collective memory.

In Richard A. Peterson's article in this encyclopedia (Culture, Production of), he noted that the production-of-culture perspective focuses on the ways in which the symbolic elements of culture are shaped by the "systems in which they are created, distributed, evaluated, taught, and preserved." In the 13 years since his article was first published, the fulcrum of activity has shifted, meriting a review of contemporary issues in the production of culture.

Briefly, the production-of-culture perspective can be traced back to Harrison White and Cynthia White's (1965) study of the emergence of impressionist art in nineteenth-century France. They argued that this new style emerged as a function of aspects of its context and method of production, of technology and training, and of features of ideology and critical discourse. This perspective has flourished since the 1970s, and scholars in multiple disciplines now use the approach established in this and other early works. Evidence of the consecration of this perspective can be found in the publication of three textbooks and two anthologies with *The Production of Culture* included in their titles (Crane, 1992; Ryan and Wentworth, 1998; du Gay, 1997; Power and Scott, 2004), and the dedication of full chapters in texts in the sociology of art (Zolberg, 1990), and in the sociology of culture (Hall and Neitz, 1993). A special issue assessing the scholarly contributions of Richard A. Peterson offered insight into the application of the approach to science, theory, art and literature, organizations, and consumption (Ryan and Hughes, 2000). An *Annual Review of Sociology* article appeared in 2004 (Peterson and Anand, 2004), a review essay in *Poetics* the following year (Janssen and Peterson, 2005), and a review issue on production perspectives solely within music (Dowd, 2004b). A year later, there was a special issue of the *Journal of Management Studies* (Jones et al., 2005).

As Peterson noted in 2001, the production perspective emerged in the 1970s largely in response to perceived weaknesses in the notion of a strict homology between culture and social structure. Rather than proposing that scholars examine the relationship of symbolic elements to social structures at (only) a macrosocial level, production-of-culture scholars focused on mesolevel structures instead – a transition of perspective broadly associated with the rise of postwar American approaches to sociological research and theory. This mesolevel focus is found in its most prescriptive form in Peterson's (1990) analysis of why rock music became a dominant form of popular music, and why it emerged in 1955. He attributed rock's rise to six production factors: law and regulation, industrial structure, technology, organizational structure, occupational careers, and the consumer market. He argued that a focus on these production factors (with minor adaptations) would allow for similar analyses of other commercial culture industries.

Yet, few studies exist with the ambition of describing a whole production system. The most notable exceptions are the works of Pierre Bourdieu that document the emergence and operation of fields (and fields of fields) of cultural production (Bourdieu, 1993, 1996). As early as 1983, van Rees noted the sympathies between Bourdieu's field theory and the (then) new production-of-culture approach, describing both as 'institutional analyses.' Among American scholars, the group of works that explore field-level phenomena include Peterson's study of rock (1990), Ferguson's study of gastronomy in France (2004), Lena's study of popular music in America (2012), Baumann's study of American art film (2007), and Gary Allen Fine's study of the production of self-taught art (2004). Most scholars focus on one aspect of the production system, the relationship between producers and consumers, or facets of work that impact the culture that is produced. In the next section, I review a few of these specific domains of research.

Dimensions of Cultural Production

To date, most studies in the production perspective have focused on particular issues: the impact of market and industry structures; the characteristic features of careers, contracts, and status orders; classification and classification systems; the fabrication of authenticity; and fan cultures, taste, and collective memory.

Market and Industry Structures

Over the past four decades, production-of-culture scholars have produced important evidence on the role of industry, organizational structure, and markets in the creation of cultural goods. This trend can be traced back to key works on the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10417-9

impacts of social structures on artistic work, including Becker's important *Art Worlds* (1982) and White and White's (1965) study of Impressionists.

For most of the twentieth century, the structure of culture industries (film, music, books, magazines, television) was an oligopoly controlled by large firms with diversified holdings across multiple markets (Hirsch, 1972; Dowd, 2004; Lena, 2006). Most scholars concluded that the culture produced under these oligopolistic conditions was dull and standardized.

Scholars disagree on the market and industry conditions that give rise to diverse music. Some argue that the decentralization of production onto subsidiary firms/organizational units can attenuate the negative effects of oligopolistic markets on musical diversity (Dowd, 2004). Others argue that new or smaller firms produce varied and dissimilar products (Peterson and Berger, 1975; Lopes, 1992).

Careers, Contracts, and Status Orders

Production-of-culture scholars often examine the ways in which careers and contracts produce status orders that influence the character and popularity of different symbolic elements. The culture industries typically employ an administrative model in which professional administrators are located in the managerial subsystem, and 'creatives' are located in the technical subsystems. Our work demonstrates that the structural organization of work weakly determines the sorts of products that can be produced.

The typical division of labor in such systems delegates budgetary decisions to administrative 'suits,' and responsibility for stimulating creativity and diversity (in the face of market uncertainty) to creatives. Since the mid-twentieth century, creative workers in such industries have enjoyed less employment security, and are likely to work under part-time or short-term contracts. As entrepreneurial or contingent labor (Neff et al., 2005), creative workers' reputations do not rest on organizational affiliation or reference to an objective rank of positions. Instead, careers advance iteratively through projects, and workers form career portfolios that ideally reflect variety, the ability to adapt, and high levels of social and economic capital (Faulkner and Anderson, 1987; Menger, 1999). Studies exist that chart careers among actors (Lincoln and Allen, 2004; Rossman et al., 2010; Uzzi and Spiro, 2005), cooks and chefs (Fine, 1992), painters and artists (Fine, 2004; Lang and Lang, 1988), authors (Janssen, 1998; Verboord, 2003; Craig and Dubois, 2010), and celebrities (Gamson, 1994).

The assessment of track records and career stages joins the characteristics known to mediate between financial rewards and peer esteem in art worlds (see Dowd et al., 2005, on gender; Dowd and Blyler, 2002, on race). Social capital and social status are keyed to generating opportunities to compete for contracts. While Becker (1982) argued that the artistic prestige of all artists is degraded by nonartistic work, including building social networks and managing finances, detailed, empirical analyses of artistic labor have revealed that high-status professionals often openly engage in these activities to no ill effect (Janssen, 1998; Craig and Dubois, 2010), and indeed, must become expert in these nonartistic activities in order to fuel the possibility of future contracts.

Status positions are allocated based on the evaluation of 'track records' (Pinheiro and Dowd, 2009: 491) and interpersonal ties (Anheier et al., 1995; Giuffre, 2001; Crossley, 2009). The resulting income distribution is rather more skewed than in other professions, and individual incomes vary more dramatically over the career than in most professions (Menger, 1999: 553). This distribution of rewards holds when the outcome measure is not income, but sales (Dowd, 2004; Lena, 2006), productivity or publication (Anheier et al., 1995; Craig and Dubois, 2010), critical success (Giuffre, 2001; Allen and Lincoln, 2004), or awards (Craig and Dubois, 2010). These 'hot' jobs in 'cool' industries are culturally desirable, but bring with them the paradox of lowered expectations of economic stability (Neff et al., 2005: 331).

There is little consensus on how the structure of career trajectories, contract types, and status orders impacts the work that results. Some maintain that demand uncertainty is so great that even markets with small numbers of elites who hoard contracts (like most contemporary culture industries) still experience stochastic shocks, such that 'all hits are flukes' (Bielby and Bielby, 1994). At least one experimental study has demonstrated that while exceptional quality guarantees at least moderate success and exceptional badness forecloses it, every other outcome is unpredictable (Salganik et al., 2006).

Classification

Interest in the prestige of cultural objects stretches back to and earlier than the start of the production-of-culture perspective. Scholars have taken a particular interest in prestige as a case of the larger class of classificatory behaviors, some of which treat symbolic elements of culture. The broadest of these, and arguably the first one to be studied by production-of-culture scholars, is the distinction between popular culture and high art, a fixture of American society since the middle of the twentieth century. The differentiation of high and low culture was driven by the class interests of rising status groups (Crane, 1992), along with the emergence of the trustee-funded, nonprofit organization (DiMaggio, 1982).

The subsequent puzzle for scholars of 'aesthetic mobility' or 'aestheticization' was to understand the mechanisms that allowed various forms of vernacular culture to be classified among the arts over the twentieth century. These include jazz (Lopes, 2002; Peterson, 2005), film (Baumann, 2001), cuisine (Ferguson, 2004; Johnston and Baumann, 2007), and 'naïve' painting (Fine, 2004). Aesthetic mobility is promoted by cultural capitalists, advantaged individuals who possess varied forms of capital that give them the power to redefine artistic classification systems and situate their tastes in privileged positions (Bourdieu, 1987, 1996). When structural transformations threaten the exclusivity of elite tastes, they simply use their familiarity with, and training in, artistic classification systems to appropriate new works into those systems, a phenomenon discovered in a range of empirical studies (DiMaggio, 1987; Lizardo and Skiles, 2012).

Production-of-culture scholars also address how categorization within industries impacts product sales and critical approval. Genre and market categories facilitate cooperation and interaction between producers and consumers because they represent sets of shared conventions, and thus constrain

producers by sanctioning deviations from these sets (Bielby and Bielby, 1994; Negus, 1999; Lena, 2012). Mixing or transgressing boundaries can inhibit success while 'categorical purity' generally enhances success (Hsu, 2006a; Zuckerman and Kim, 2003). Producers who are situated in multiple categories confuse audiences, and this lowers the appeal of their products (Hsu, 2006b); they are also seen as less able to create good products (Hannan, 2010; Negro et al., 2010). These results are mediated by aspects of organizational identity (Smith, 2011), audience characteristics (Pontikes, 2012), and features of the classification system (Ruef and Patterson, 2009). In particular, if boundaries within the system are in flux, or are weak, boundary crossing may not incur penalties and may result instead in valorization as innovation (DiMaggio, 1987; Rao et al., 2005; Hsu et al., 2012).

A core concern in studies of classification is their link to status; after all, "struggles over definition (or classification) have boundaries at stake … and, therefore, hierarchies" (Bourdieu, 1996: 159). In addition to the studies that focus on boundary drawing by large-status groups, production-of-culture scholars also focus on field-level or disciplinary boundaries. For example, Anand (2000) examines the differences between central-normative fields where professionals wield the power to attribute legitimacy and authenticity, and peripheral-competitive fields where audiences 'vote with their dollars.'

More recent studies of ranking and classification have examined various indexes, lists, and performance algorithms as cultural objects that impact the behavior of those ranked according to them. Rank ordering and other forms of performance evaluation have a tendency to cause individuals to alter their behavior in response to being evaluated, observed, or measured. Research on the insurance industry demonstrates that buying insurance increases our risk-avoidance behavior, rendering our insurance less valuable (Heimer, 1985). The general observation has been made that knowledge of theories of economic behavior shapes how individuals behave economically, a characteristic Callon (1998) calls 'performativity' (see also Latour, 1987). Anand and Peterson (2000) argue that 'market information regimes' like stock prices and bestseller lists define what counts as 'information' about the market, and this impacts what strategies participants use. For example, using market information contained in *Billboard* magazine charts, Anand and Peterson show that music industry participants change their definitions of success and may invest resources in different trends, producing large changes in the culture that is produced. Measures of law school quality are also highly reactive measures, and cause schools to make administrative and curricular decisions in order to increase their rank (Espeland and Sauder, 2007).

Authenticity

Production-of-culture scholars have sought to understand the circumstances under which cultural producers and their works are seen as 'authentic.' Most define authenticity as "the ability of a place, experience, person, or object to conform to an idealized representation of reality, that is, to a set of expectations regarding how something ought to look sound, feel, smell" (Grazian, 2003: 10). Authenticity is a social construction, a set of shared beliefs, and not an objective quality inherent in things, people, places, or experiences. Definitions of authenticity are shared within groups, but are not universalizable.

Based on this approach, authenticity can be viewed as a quality that is produced by teams of coordinated actors within a field. Scholars using the production-of-culture approach focus their attention toward the teams of musicians, managers, agents, producers, and other artistic support staff whose work is to 'fabricate authenticity' (Peterson, 1997; Hughes, 2000; Fine, 2004; Sherman, 2007). To understand how authenticity is fabricated, we often need to find the 'creator of the creator' (Bourdieu, 1996: 168). Attributions of authenticity frequently rest on aspects of particularistic categorical attributes of the creators including their race and gender, and these can be as important during the evaluative process as any of the formal features of the objects or performances they create. In addition to categorical attributes of culture creators, scholars have found that evaluations of authenticity rest on the unique, handmade, or traditional character of an item (Bendix, 1997; Beverland, 2005), its historical existence (Anand and Jones, 2005), and its geographic specificity (Johnston and Baumann, 2007; Phillips and Kim, 2009). The production of identity is the focus of studies of self-taught artists (Fine, 2004), blues musicians (Grazian, 2003), country artists (Peterson, 1997), punk musicians (Lull, 1986), rap musicians (McLeod, 1999; Lena, 2013), jazz musicians (Phillips and Kim, 2009), restaurant cooks and waiters (Lu and Fine, 1995), and foodies (Johnston and Baumann, 2007).

Fan Cultures, Taste, and Collective Memory

Reception by the listener, reader, viewer, or purchaser has been a long-standing concern in the study of culture. As Peterson (2001: 3172) noted in his article in this volume, the term 'reception' suggests the end of a communication process, but most sociologists conceive reception as an active process during which consumers select, interpret, and recombine elements. That is, they conceptualize it as a form of *production*. Instead of identifying consumers who have received the 'correct' message, we observe patterns in the meanings that they attach to them (van Rees, 1987; McDonnell, 2010). Peterson (1983) is among those who argue that this process should properly be referred to as 'autoproduction,' although such work is most often identified as research on 'identity' or 'taste.'

There are many studies that examine the production of identity through cultural consumption and participation in fan communities. The leading light is Tia DeNora, who has used interviews and observation to examine how individuals use music to invoke and manage emotion, relationships, decisions, and the like; she addresses music as a 'technology of the self' (2000).

Recent studies of class differences in taste reveal that most people of high status have dispensed with an exclusive preference for the old class markers of cultural capital (the appreciation of fine arts) and instead have diverse cultural preferences that include popular arts (Peterson and Kern, 1996). This result has been found in multiple countries including Canada, the United Kingdom, the Netherlands, France, Germany, Spain, Austria, and Australia, and in additional analyses of the United

States population (see Peterson, 2005, for a review). These studies document shifting tastes across a range of cultural pursuits including music (Bennett, 2006), reading (Zavisca, 2005), art (DiMaggio and Mukhtar, 2004; Vander Stichele and Laermans, 2006), food (Johnston and Baumann, 2007), and television (Lizardo and Skiles, 2009).

While documenting the rise in omnivorous tastes among elites, scholars emphasized that group boundaries are maintained based not "on *what* is consumed, but on *the way in which items of consumption are understood*" (Peterson and Kern, 1996: 904). It is possible that elites consuming 'middlebrow' and 'lowbrow' culture are making discriminatory judgments at the subgenre level that are missed in surveys. Johnston and Baumann's (2007) work suggests this may be the case among gastronomic omnivores; 'foodies' will eat the same foods, in the same restaurants, as others, but discriminate on the basis of their perceived authenticity. This suggests future research will seek to blend these survey results with more qualitative research and may discard genre in favor of more specific measures of consumption. Some research of this kind is taking place (Sonnett, unpublished manuscript), informed by Bourdieu's approach in *Distinction* (1987; see also Fligstein and McAdam, 2012). The obstacle to research in this area is the resource intensity of consumption at the intragenre level.

Production-of-culture scholars have also investigated the production of reputation (Van Rees and Vermunt, 1996; Barker-Nunn and Fine, 1998) and collective memory. As shared, established images of social actors, reputations emerge from a social process by which information and events are filtered through a selective lens (Lang and Lang, 1988). Certain details acquire relevance and gravity, while others are disregarded or censored out of the stories that concatenate into reputations. Thus, collective memories are reinterpretations of past people and events that harmonize them with current political needs and cultural beliefs (Halbwachs, 1992; Lang and Lang, 1988; Sewell, 1996). Media-studies scholars have offered case studies of celebrity (Gamson, 1994), media frames (Tuchman, 1978), and talk shows on television (Grindstaff, 2002) to illuminate the dynamics of mass reputations.

Prospects

Thirteen years ago, Richard Peterson focused here on the prospect of studying production factors given the globalization of production and the high rate of technological development as well as changes in censorship law and intellectual property rights. While scholars outside the United States have contributed meaningfully to understanding global production and distribution systems, and communications scholars have directed attention toward the impact of regulatory systems in the digital age on content, sociologists have turned their attention elsewhere. In particular, production-of-culture scholars and their works have been producing thousands of studies of omnivorousness, and the growing body of works that focus on product classification within industries.

While the revolution in digital technology did not spark a bumper crop of research on online distribution systems, it did encourage scholars to acquire a set of tools to examine 'big data.' We have an extraordinary capacity to evaluate large amounts of information, but only some of it is unencumbered by legal restrictions on access and use. This means we will likely see research on music and literature replaced by studies of the transformation of social media like Twitter, and scientific collaboration and citation, because data on the latter have thus far been more tractable.

Production-of-culture scholars may also be drawn to study new careers in new or rapidly transforming industries like finance. To maximize their leverage on these cases, they may benefit from collaboration with economic sociologists who already understand a great deal about how such careers work. Relatedly, these subspecialties could productively collaborate to understand new performance metrics and predictive algorithms that have rationalized finance and produced new sources of market volatility. Finally, it is time for production-of-culture scholars to dispense with older (even revised or adapted) categories of consumers and seek to understand tastes by inductively producing market segments from behavior, attitudes, and attributes.

See also: Artists, Competition and Markets; Cultural Mediators and Gatekeepers; Cultural Participation, Trends In; Cultural Production in Networks; Markets: Artistic and Cultural.

Bibliography

Allen, M.P., Lincoln, A.E., 2004. Critical discourse and the cultural consecration of American films. Social Forces 82, 871–894.
Anand, N., 2000. Defocalizing the organization: Richard A. Peterson's sociology of organizations. Poetics 28, 173–184.
Anand, N., Peterson, R.A., 2000. When market information constitutes fields: sensemaking of markets in the commercial music industry. Organization Science 11, 270–284.
Anand, N., Jones, B.C., 2005. Fabricating Authenticity to Produce Organizational Culture: The Case of a British Retail Chain Store. Unpublished Manuscript. London Business School, London.
Anheier, H.K., Gerhards, J., Rono, F.P., 1995. Forms of capital and social structure in cultural fields: examining Bourdieu's social topography. American Journal of Sociology 100, 859–903.
Barker-Nunn, J., Fine, G.A., 1998. The vortex of creation: literary politics and the demise of Herman Melville's reputation. Poetics 26, 81–98.
Baumann, S., 2001. Intellectualization and art world development. American Sociological Review 66, 404–426.
Baumann, S., 2007. Hollywood Highbrow: From Entertainment to Art. Princeton, Princeton, NJ.
Becker, H.A., 1982. Art Worlds. University of California Press, Berkeley, CA.
Bendix, R., 1997. In: Search of Authenticity: The Formation of Folklore Studies. U of Wisconsin Press, Madison.
Bennett, A., 2006. 'Punk's not dead': the continuing significance of punk rock for an older generation of fans. Sociology 40, 219–235.
Beverland, M.B., 2005. Crafting brand authenticity: the case of luxury wines. Journal of Management Studies 42, 1003–1029.
Bielby, W.T., Bielby, D.D., 1994. 'All hits are flukes:' institutionalized decision-making and the rhetoric of network prime-time program development. American Journal of Sociology 99, 1287–1313.
Bourdieu, P., 1987. Distinction: A Social Critique of the Judgement of Taste. Harvard UP, Cambridge, MA.
Bourdieu, P., 1993. The Field of Cultural Production: Essays on Art and Literature. Columbia UP, New York.
Bourdieu, P., 1996. The Rules of Art: Genesis and Structure of the Literary Field. Stanford UP, Stanford.
Callon, M. (Ed.), 1998. The Laws of the Markets. Blackwell, Oxford.
Coser, L.A. (Ed.), 1978. The Production of Culture. Social Research, 45.
Craig, A., Dubois, S., 2010. Between art and money: the social space of public readings in contemporary poetry economies and careers. Poetics 38, 441–460.

Crane, D., 1992. The Production of Culture. Sage, Newbury Park, CA.
Crossley, N., 2009. The man whose web expanded: network dynamics in Manchester's post/punk music scene, 1976–1980. Poetics 73, 24–49.
DeNora, T., 2000. Music in Everyday Life. Cambridge UP, New York.
DiMaggio, P., 1987. Classification in art. American Sociological Review 52, 440–455.
DiMaggio, P., 1982. Cultural entrepreneurship in nineteenth-century Boston: the creation of an organizational base for high culture in America. Media, Culture and Society 4, 33–50.
DiMaggio, P., Mukhtar, T., 2004. Arts participation as cultural capital in the United States, 1982–2002: signs of decline? Poetics 32, 169–194.
Dowd, T., 2004a. Concentration and diversity revisited: production logics and the U.S. mainstream recording market, 1940–1990. Social Forces 82, 1411–1455.
Dowd, T.J., 2004b. Production perspectives in the sociology of music. Poetics 32, 235–246.
Dowd, T., Blyler, M., 2002. Charting race: the success of black performers in the mainstream recording market, 1940 to 1990. Poetics 30, 87–110.
Dowd, T.J., Liddle, K., Blyler, M., 2005. Charting gender: the success of female acts in the U.S. mainstream recording market, 1940–1990. Research in the Sociology of Organizations 23, 81–123.
Espeland, W.N., Sauder, M., 2007. Rankings and reactivity: how public measures recreate social worlds. American Journal of Sociology 113, 1–40.
Faulkner, R.R., Anderson, A.B., 1987. Short-term projects and emergent careers: evidence from Hollywood. American Journal of Sociology 92, 879–909.
Ferguson, P., 2004. Accounting for Taste: The Triumph of French Cuisine. University of Chicago Press, Chicago, IL.
Fine, G.A., 1992. The culture of production: aesthetic choices and constraints in culinary work. American Journal of Sociology 97, 1268–1294.
Fine, G.A., 2004. Everyday Genius: Self-taught Art and the Culture of Authenticity. University of Chicago Press, Chicago, IL.
Fligstein, N., McAdam, D., 2012. A Theory of Fields. Oxford University Press, New York.
Gamson, J., 1994. Claims to Fame: Celebrity in Contemporary America. University of California Press, Berkeley, CA.
du Gay, P. (Ed.), 1997. Production of Culture: Cultures of Production. Sage, London.
Grindstaff, L., 2002. The Money Shot: Trash, Class and the Making of TV Talk Shows. University of Chicago Press, Chicago, IL.
Giuffre, K., 2001. Mental maps: social networks and the language of critical reviews. Sociological Inquiry 71, 381–393.
Grazian, D., 2003. Blue Chicago: The Search for Authenticity in Urban Blues Clubs. University of Chicago Press, Chicago.
Griswold, W., 1994. Cultures and Societies in a Changing World. Pine Forge Press, Thousand Oaks, CA.
Halbwachs, M., 1992. On Collective Memory. University of Chicago Press, Chicago.
Hall, J.R., Neitz, M.J., 1993. Culture: Sociological Perspectives. Prentice-Hall, Englewood Cliffs, NJ.
Hannan, M.T., 2010. Partiality of memberships in categories and audiences. Annual Review of Sociology 36, 159–181.
Heimer, C., 1985. Reactive Risk and Rational Action: Managing Moral Hazard in Insurance Contracts. University of California Press, Los Angeles.
Hirsch, P.M., 1972. Processing fads and fashions: an organization-set analysis of cultural industry systems. American Journal of Sociology 77, 639–659.
Hsu, G., 2006a. Jacks of all trades and masters of none: audiences' reactions to spanning genres in feature film. Administrative Science Quarterly 51, 420–450.
Hsu, G., 2006b. Evaluative schemas and the attention of critics in the US film industry. Industrial and Corporate Change 15, 467–496.
Hsu, G., Roberts, P.W., Swaminathan, A., 2012. Evaluative schemas and the mediating role of Critics. Organization Science 23, 83–97.
Hughes, M., 2000. Country music as impression management: a mediation on fabricating authenticity. Poetics 28, 185–206.
Janssen, S., 1998. Side-roads to success: the effect of sideline activities on the status of writers. Poetics 25, 265–280.
Janssen, S., Peterson, Richard A., 2005. Comparative research on cultural production and consumption. Poetics vol. 33 (5–6), 253–256.
Janssen, S., Kuipers, G., Verboord, M., 2008. Cultural globalization and arts journalism: the international orientation of arts and culture coverage in Dutch, French, German, and U.S. newspapers, 1955–2005. American Sociological Review 73, 719–740.
Johnston, J., Baumann, S., 2007. Democracy versus distinction: a study of omnivorousness in gourmet food writing. American Journal of Sociology 113, 165–204.
Jones, C., Anand, N., Alvarez, J.L. (Eds.), 2005. Manufactured Authenticity and Creative Voice in Cultural Industries. Journal of Management Studies 42 (5), 893–899.
Lang, G.E., Lang, K., 1988. Recognition and renown: the survival of artistic reputation. American Journal of Sociology 94, 79–109.
Latour, B., 1987. Science in Action. Harvard University Press, Cambridge, MA.
Lena, J.C., 2006. Social context and musical content: rap music, 1979–1995. Social Forces 85, 479–495.
Lena, J.C., 2012. Banding Together: How Communities Create Genres in Popular Music. Princeton University Press, Princeton, NJ.
Lena, J.C., 2013. Authenticity and independence in rap music and other genre communities. In: Horsfall, S., Meij, J., Probstfield, M. (Eds.), Explorations in Music Sociology: Examining the Role of Music in Social Life. Paradigm Publishing, Boulder, CO.
Lincoln, A.E., Allen, M.P., 2004. Double jeopardy in Hollywood: age and gender in the careers of film actors, 1926–1999. Sociological Forum 19, 611–631.
Lizardo, O., Skiles, S., 2009. Highbrow omnivorousness on the small screen?: cultural industry systems and patterns of cultural choice in Europe. Poetics 37, 1–23.
Lizardo, O., Skiles, S., 2012. Reconceptualizing and theorizing 'omnivorousness:' genetic and relational mechanisms. Sociological Theory 30, 263–282.
Lopes, P.D., 1992. Innovation and diversity in the popular music industry, 1969–1990. American Sociological Review 57, 56–71.
Lopes, P.D., 2002. The Rise of a Jazz Art World. Cambridge UP, Cambridge.
Lu, S., Fine, G.A., 1995. The presentation of ethnic authenticity: Chinese food as a social accomplishment. Sociological Quarterly 36, 535–553.
Lull, J., 1986. Thrashing in the pit: an ethnography of San Francisco punk subculture. In: Lindlof, T. (Ed.), Natural Audiences: Qualitative Research on Media Uses and Effects. Ablex, Norwood, NJ.
McLeod, K., 1999. Authenticity within hip-hop and other cultures threatened with assimilation. Journal of Communication 49, 134–150.
McDonnell, T.E., 2010. Cultural objects as objects: materiality, urban space, and the interpretation of AIDS media in Accra, Ghana. The American Journal of Sociology 115, 1800–1852.
Menger, P.M., 1999. Artistic labor markets and careers. Annual Review of Sociology 25, 541–574.
Neff, G., Wissinger, E., Zukin, S., 2005. Entrepreneurial labor among cultural producers: 'Cool' jobs in 'hot' industries. Social Semiotics 15, 307–334.
Negro, G., Hannan, M.T., Rao, H., 2010. Categorical contrast and audience appeal: niche width and critical success in winemaking. Industrial and Corporate Change 19, 1397–1425.
Negus, K., 1999. Music Genres and Corporate Cultures. Routledge, London.
Peterson, R.A., Kern, R.M., 1996. Changing highbrow taste: from snob to omnivore. American Sociological Review 61, 900–907.
Peterson, R.A., 1976. The production of culture: a prolegomenon. In: Peterson, R.A. (Ed.), The Production of Culture. Sage, Beverly Hills, CA, pp. 7–22.
Peterson, R.A., 1983. Patterns of cultural choice: prolegomenon. American Behavioral Scientist 26, 422–438.
Peterson, R.A., 1990. Why 1955? Explaining the advent of rock music. Popular Music 9, 97–116.
Peterson, R.A., 1997. Creating Country Music: Fabricating Authenticity. Vanderbilt University Press, Nashville, TN.
Peterson, R.A., Anand, N., 2004. The production of culture perspective. Annual Review of Sociology 30, 311–334.
Peterson, R.A., Berger, D., 1975. Cycles in symbol production: the case of popular music. American Sociological Review 40, 158.
Peterson, R.A., 2005. Problems in comparative research: the example of omnivorousness. Poetics 33, 257–282.
Phillips, D.J., Kim, Y.-K., 2009. Why pseudonyms?: deception as identity preservation among jazz record companies. Organization Science 20, 481–499.
Pinheiro, D.L., Dowd, T.J., 2009. All that jazz: the success of jazz musicians in three metropolitan areas. Poetics 37, 490–506.
Pontikes, E.G., 2012. Two sides of the same coin: how ambiguous classification affects multiple audiences' evaluations. Administrative Science Quarterly 57, 81–118.
Power, D., Scott, A.J. (Eds.), 2004. Cultural Industries and the Production of Culture. Routledge, New York.
Rao, H., Durand, R., Monin, P., 2005. Border crossing: bricolage and the erosion of categorical boundaries in French gastronomy. American Sociological Review 70, 968–991.
Ruef, M., Patterson, K., 2009. Credit and classification: defining industry boundaries in 19th century America. Administrative Science Quarterly 54, 486–520.
Ryan, J., Hughes, M. (Eds.), 2000. The Production and Consumption of Culture: Essays on Richard a. Peterson's Contributions to Cultural Sociology. Special issue Poetics. 28/2–3, 91–233.
Rossman, G., Esparza, N., Bonacich, P., 2010. I'd like to thank the academy: team spillovers and network centrality. American Sociological Review 75, 31–51.
van Rees, K., 1983. Advances in the empirical sociology of literature and the arts. Poetics 12, 285–310.
van Rees, K., 1987. How reviewers reach consensus on the value of literary works. Poetics 16, 275–294.
van Rees, K., Vermunt, J., 1996. Event history analysis of authors' reputation: effects of critics' attention on debutants' careers. Poetics 23, 317–333.

Ryan, J., Wentworth, W.M., 1998. Media and Society: The Production of Culture in the Mass Media. Prentice-Hall, New York.
Salganik, M.J., Dodds, P.S., Watts, D.J., 2006. Experimental study of inequality and unpredictability in an artificial cultural market. Science 311, 854–856.
Sewell Jr., W.H., 1996. Political events as structural transformations: inventing revolution at the Bastille. Theory and Society 25, 841–881.
Sherman, R., 2007. Class Acts: Service and Inequality in Luxury Hotels. University of California Press, Los Angeles.
Smith, E.B., 2011. Identities as lenses: how organizational identity affects audiences' evaluation of organizational performance. Administrative Science Quarterly 5, 61–94.
Sonnett, J. Indifference, ambivalence, and distinction: a configurational analysis of musical boundaries in Mississippi, unpublished manuscript.
Tuchman, G., 1978. Making News. Free Press, New York.
Uzzi, B., Spiro, J., 2005. Collaboration and creativity: the small world problem. American Journal of Sociology 11, 447–504.
Vander Stichele, A., Laermans, R., 2006. Cultural participation in Flanders: testing the cultural omnivore thesis with population data. Poetics 34, 45–64.
Verboord, M., 2003. Classification of authors by literary prestige. Poetics 31, 259–281.
White, H.C., White, C.A., 1965. Canvases and Careers. Wiley, New York.
Zavisca, J., 2005. The status of cultural omnivorism: a case study of reading in Russia. Social Forces 84, 1233–1255.
Zolberg, V.L., 1990. Constructing a Sociology of the Arts. Cambridge University Press, New York.
Zuckerman, E.W., Kim, T.Y., 2003. The critical trade-off: identity assignment and box-office success in the feature film industry. Industrial and Corporate Change 12, 27–67.

Culture Shock

Waud Kracke, University of Illinois, Chicago, IL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 5, pp. 3176–3179, © 2001, Elsevier Ltd.

Abstract

Immersion in a foreign culture, while initially exciting, often leads to a state of emotional distress. This personal disequilibrium, brought about by encountering unexpected or disturbing differences of behavior and social expectations, and by difficulties in communication, has been termed 'culture shock' – a designation introduced in 1950 by Kalervo Oberg. Symptoms of distress may become intense, and may even lead to psychological problems. But in general culture shock is mastered by a process of adaptation to the new culture. Indeed, it may be considered a regular part of learning a culture. Some travelers, surprisingly, experience equal or even greater disorientation on return to their own culture, a condition which shows the depth of personality transformation that is brought about by adapting to a new culture.

Anyone who has spent time in a culture significantly different from his or her own has more than likely experienced a sense of disorientation and perplexity at life in the new culture. After an initial period of excitement about the novelty of the new culture and perhaps pleasure in the absence of some of the constraints of one's own society, the traveler may be plunged into a period of anxiety and frustration at the difficulty of understanding and making himself or herself understood. The visitor may feel confused by differences in expectations and in interpersonal relationships, abhorrence at some cultural practices or perceived dispositions of the people, followed by intense homesickness, depression, and perhaps regression to unaccustomed bouts of annoyance and labile feelings. Frequent manifestations include food concerns – yearning for familiar foods or anxiety about the safety of the drinking water. Such reactions on encounter with a new culture have been referred to in anthropology as 'culture shock,' a term which has been in currency since 1950, when Kalervo Oberg gave a talk of that title to the women's club of Rio de Janeiro – a group of wives of American businessmen and diplomats. Bronislaw Malinowski, although he did not use the term, gave one of the classic descriptions of culture shock in 1922, in the introduction to *Argonauts of the Western Pacific*:

> I well remember the long visits I paid to the villages during the first weeks; the feeling of hopelessness and despair after many obstinate but futile attempts had entirely failed to bring me into real touch with the natives, or supply me with any material. I had periods of despondency, when I buried myself in the reading of novels, as a man might take to drink in a fit of tropical depression and boredom.

Culture shock has been noted frequently in anthropological fieldwork. It has become a regular part of anthropological thinking about culture, especially in the USA. Yet discussion of its manifestations has been comparatively rare in anthropology, only one major book being devoted to it. Some more specialized spinoff concepts, such as Toffler's 'future shock,' have been more extensively written on. Culture shock has been best described, perhaps, in some fictional works by anthropologists and others – Laura Bohannan's *Return to Laughter* (see Bowen, 1954); E M Forster's *A Passage to India* – or in briefer introspective works by anthropologists such as Mary Catherine Bateson (1968) and Jean Briggs (1970). George Devereux's *From Anxiety to Method in the Behavioral Sciences* (1968) still remains one of the best theoretical explorations of its significance.

The earliest writings on culture shock treated it as a kind of pathology, an obstacle to smooth operation in the culture one is entering and to communicating with its people. Kalervo Oberg listed a number of 'symptoms' of this 'transitory disorder,' including: compulsive handwashing, emotional lability and excessive preoccupation with cleanliness. The psychoanalyst Gertrude Ticho, in a 1971 article, expands this list to include

> increased rigidity, fatigue, withdrawal anxiety, depression and elation, distressed obsessive neurotic-like symptoms such as over concern with the cleanliness of food and water, and hysterical reactions.

She goes on to add that

> at times, after a period of enchantment with and enthusiasm for the new culture, a hostile reaction to the people of the foreign culture ensues. This reaction becomes endowed with more global prejudices, such as 'they are lazy and untrustworthy, rigid, treacherous and sabotage progress.' The previously open-minded, outgoing 'foreigner' avoids contact with the people of the host country and withdraws to 'his own people,' if possible, thus creating an enclave within the foreign country.

'If this pattern is not reversed,' she concludes, 'the gap between the different cultures may be widened and result in a vicious circle' (Ticho, 1971: pp. 324–325).

More recently, culture shock has been seen as an essential part of adapting to an unfamiliar culture, of coming to understand ways of living and of seeing the world that are different from one's own. Recent anthropological writers have viewed the process in different ways:

(a) as a loss of one's own culture, with a resulting process of grieving or mourning the loss;

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12049-5

(b) as a 'rite of passage,' a kind of 'initiation' marking the transition from membership in one culture to at least being conversant with another;
(c) as a 'resocialization' into another culture; and
(d) as a process like therapy or psychoanalysis for reorganizing one's personality, resolving old conflicts along new lines (see Kracke, 1987).

The first of these, 'mourning the loss of one's own culture,' along with the old view of culture shock as a 'transitory psychological crisis,' may be grouped together as a 'negative' view of the phenomenon, emphasizing the crisis of moving *out of* one's previous 'culture of orientation.' The last three – 'rite of passage,' 'resocialization' and the analogy with change in psychotherapy – are 'positive' viewpoints, emphasizing the transition *into* the new culture, the learning of new ways of thinking and reacting.

Phases of Culture Shock

While everyone's experience of cultural transition is unique, depending on one's own personality and past experiences, and on the cultures involved – the entry into another culture tends to develop in common stages. Unless the entry into the culture is under duress (capture, exile, or fleeing from a natural or political disaster at home), one's first experience tends to be one of enchantment or delight. One may also experience shock at aspects of the new culture like poverty, dirt (or antiseptic cleanliness), or noise and confusion (for a rural person arriving for the first time in the city); but at first, one is likely to feel curiosity about exploring the new culture. For a while, one may take pleasure in the absence of some of the constraints of one's home culture, at the warmth of emotions people express (if you come from a reserved culture), or the willingness of people to help you. After a time, frustrations begin to mount as one realizes, for example, that the directions people give may be more based on desire to be helpful than on a real knowledge of the information asked for, or that invitations are not to be taken at face value. One becomes aware of not knowing the cues for judging situations or people's intent, and finds oneself in awkward situations, or making gaffes. After a certain period, which varies from a few days to a few months, one begins to feel a certain 'culture fatigue' (as George Guthrie has termed it (Szanton, 1966) – an irritation at the accumulating aggravations which may mount to an exasperation with the culture itself.

At this point, if one does not give in to the inclination to leave, one begins the process of learning the new culture: what William Caudill (1961) calls 'resocialization.' The phases of distress and learning are reiterated in cycles as one is more and more deeply socialized into the culture, gradually approaching a phase of being more at home in it. The process is to some extent that of reconstructing one's identity, or constructing an alternate identity for one's sojourn in the culture. Over years of living in a new culture, the process may culminate in the formation of a new identity, as Garza-Guerrero (1974) suggests, combining patterns and values from one's culture of origin with patterns and values of the new one.

You Can't Go Home Again

An unexpected feature of culture shock is the dis-orientation one experiences on *returning* to one's culture of origin – a phenomenon which has been designated 'reverse culture shock' (Hertz, 1977). Many people feel even more difficulty in readjusting to their *own* culture after returning from abroad than they did adjusting to the foreign culture or cultures they have lived in. Why is this? In part, it is the phenomenon which the novelist Thomas Wolfe identified in *You Can't go Home Again:* You have been changed by your encounter with another culture, and you no longer 'fit' in the same place you did before going. You have an experience you do not share with your compatriots who remained at home. It is perhaps exacerbated by the very unexpectedness of it: you have had dreams of coming back home where everything will at last be unproblematic, and suddenly it is not.

Reverse culture shock reveals something about the nature of the culture shock experience itself: it measures the depth that one is changed by the encounter with another culture. We come back a different person from what we were; we have taken something of the other culture into ourselves, and can never again be fully at home in any one culture.

Aspects of Culture Shock: The Epistemology of Knowing Other Cultures

Culture shock may be an important, if not essential, part of the process of learning another culture, the central task of social or cultural anthropology. In a brilliant article, Mary Catherine Bateson (1968) discusses her own experience of culture shock in a very personal way. Culture shock, she suggests, is, among other things, a useful tool for recognizing where the new culture differs from one's own. Where something about the new culture annoys or angers you, disgusts you, or arouses anxiety or a sense of disorientation, that is a point where one must look for fundamental differences in assumptions about life or in conceptualizing the world. In our western culture, for example, we view death and mourning as something highly private, and avoid talking with the bereaved person about the loss for fear of intruding on their privacy. Filipino culture, which regards bereavement as something that must be worked through with others, so that friends talk openly with the bereaved about the loss, helping the bereaved person to re-experience the loss and the grief in a context of concerned support. Initially, she was shocked at the direct way in which Filipinos ask the bereaved person about the loss, often provoking tears; by the time she came to experience a deep loss of her own in the Philippines, she was prepared to find 'deeply healing' the opportunity to articulate and vent her feelings of loss.

Another view of the nature of cultures leads George Devereux (1968) to formulate a deeper view of the source of culture shock. In a psychoanalytic viewpoint, he suggests that different cultures have different 'distribution of repressions': some feelings and fantasies are acceptable and may be talked about and accepted in oneself, others are repugnant and should not be acknowledged. The latter, in general, will be repressed, or at least disavowed, by most people in that culture. If

something that is regarded as deeply disturbing and repugnant in one's own culture is considered acceptable and openly talked about in the culture you are entering into contact with, that may lead to considerable anxiety on your part. On the other hand, if something you consider normal and not anything to be ashamed of is regarded as repugnant in the culture you are going into, that may lead you to do or say things that will cause great distress to others in that culture; you may inadvertently arouse anxiety and distress, without at first understanding why.

Culture Shock and Mental Illness

Culture shock has received more attention in psychiatric and psychoanalytic literature, where the propensity of cultural estrangement to precipitate emotional problems or mental illness has been widely recognized. A number of articles have been written by psychiatrists and psychotherapists on the need for psychotherapeutic treatment for people whose culture shock has led them to emotional crisis. The Nigerian psychiatrist T. Adeoye Lambo has discussed certain psychiatric problems of African students abroad. Prakash Desai and George Coelho have written about the adaptation problems of Indian immigrants in the USA. Juana Antokoletz and many others have written on problems of Mexican and other Latin American immigrants in the USA. One of the most thorough discussions of culture shock was published by the expatriate Cuban psychoanalyst Garza-Guerrero (1974). Recently, the anthropologist George Spindler has developed a 'cultural therapy' to deal with problems of culture shock and intercultural communication.

Why is the encounter with another culture so unsettling? Perhaps it is because each encounter with a different way of life puts in question our own beliefs and values, confronted as we are with alternative ways of looking at reality or at social relationships, with different rules to guide our conduct. Each culture we come into contact with raises these issues anew, challenging new aspects of our world view and ethical standards. The depth to which these experiences disturb us is an indicator of how deeply the cultural framework we grow up in shapes our view of the world and of relations with others. It may be that this experience, the discomfort we feel in encountering another culture, is what most moves our curiosity about other cultures: the stimulus to anthropological study.

See also: Alienation: Psychosociological Tradition; Assimilation of Immigrants; Diversity and Disagreement in Ethics: Philosophical Implications; Ethnic Identity and Ethnicity in Archaeology; Ethnic Identity, Psychology of; Migration and Health; Multiculturalism, Anthropology of; Multiculturalism; Racial Relations.

Bibliography

Bateson, M.C., 1968. Insight in a bicultural context. Philippine Studies 16, 605–621.
Bowen, E.S., Bohannan, L., 1954. Return to Laughter. Harper & Row, New York.
Briggs, J., 1970. Kapluna daughter. In: Never in Anger: Portrait of an Eskimo Family. Harvard University Press, Cambridge, MA.
Caudill, W., 1961. Some problems in transnational communication: Japan-United States. In: Group for the Advancement of Psychiatry Symposium Number 7: Application of Psychiatric Insights to Cross-cultural Communication.
Devereux, G., 1968. From Anxiety to Method in the Behavioral Sciences. Mouton, The Hague, The Netherlands.
Furnham, A., Bochner, S., 1986. Culture Shock: Psychological Reactions to Unfamiliar Environments. Methuen, London.
Garza-Guerrero, A.C., 1974. Culture shock: Its mourning and the vicissitudes of identity. Journal of the American Psychoanalytic Association 22, 408–429.
Grinberg, L., Grinberg, R., 1989. Psychoanalytic Perspectives on Migration and Exile. Yale University Press, New Haven, CT.
Hertz, D.G., 1977. The problem of 'reverse' culture shock. In: Boroffka, A., Pfeiffer, W.M. (Eds.), Fragen der transkulturell-vergleichende Psychiatrie in Europa. Westfälische Wilhelms-Universität, Münster, Germany.
Kracke, W., 1987. Encounter with other cultures: psychological and epistemological aspects. Ethos 15, 56–79.
Malinowski, B., 1922. Introduction: the subject, method and scope of this enquiry. Argonauts of the Western Pacific, Dutton, New York, pp. 1–25.
Read, K.E., 1965. The High Valley. Scribner, New York.
Szanton, D., 1966. Cultural confrontation in the Philippines. In: Textor R.B. (Ed.), Cultural Frontiers of the Peace Corps. MIT Press, Cambridge, MA.
Ticho, G., 1971. Cultural aspects of transference and counter transference. Bulletin of the Menninger Clinic 35, 313–334.

Culture, Sociology of

Mabel Berezin, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.-L. Harouel, volume 5, pp. 3179–3184, © 2001, Elsevier Ltd.

Abstract

Culture is constitutive of classical sociology and implicit in much postwar European sociology. Postwar American sociology began to consider culture in the 1970s with a focus upon arts. In the 1980s as the work of prominent European sociologists disseminated in translation, a global cultural turn began. Three approaches to cultural analysis have evolved in the last 20 years: first, culture is a quantifiable variable and amenable to the claims of science; second, drawing inspiration from French social theory culture is a practice of boundary formation and evaluation; and third, culture is a narrative read in ritual and performance.

The Problem of Meaning

Culture as an analytic category is a staple of many academic disciplines such as anthropology, literature, and history. Classical sociologists, Emile Durkheim, Max Weber, Georg Simmel, even Karl Marx, recognized culture as fundamental to society and social relations. Max Weber (1968) wrote about the logic of the cultural sciences and located sociology among them. Culture was implicit in mid-twentieth-century European sociology – much of which focused upon political economy in specific national contexts. In the *longue durée* of intellectual life, a concern with culture is relatively new to professional sociology – particularly postwar American sociology with its positivistic and scientific emphasis. Mid-century or postwar sociology, as in international politics, was, to borrow from Henry Luce, the "American Century." It was not until the mid-1970s and 1980s when the work of Jurgen Habermas, Antonio Gramsci, Michel Foucault, and Pierre Bourdieu began to circulate in translation that sociology in general and American sociology in particular became more cosmopolitan and culturally oriented.

Culture was not entirely absent from mid-century sociological studies. Talcott Parsons' work dominated this period and his 'social system' (1951) achieved equilibrium through the 'glue' of common values. Political culture formed the core of modernization studies in the 1950s and 1960s. Almond and Verba's ([1963]1989) study of 'civic culture' and Banfield's (1958) account of a southern Italian village are classic examples. These studies measured the diffusion of western values as an index of democratic dispositions. They seem quaint in the face of the massive political upheavals and cultural complexity of the late twentieth and early twenty-first century. Yet, Almond and Verba as well as Banfield have contemporary counterparts as the essays in Harrison and Huntington's (2000) anthology, *Culture Matters*, illustrate. Samuel Huntington's *Clash of Civilizations* (1996) locates culture in civilizations that are based on shared religions. Though politically conservative, these studies have value in that they point in the direction of a thick conception of culture that owes more to anthropology than to political science or sociology.

Meaning, individual and collective, is constitutive of culture. The 'cultural turn' in social science began in anthropology and to some extent British cultural Marxism in the mid-1970s (Willis, [1977]1981). In *Keywords*, a social history of concepts, British cultural critic Raymond Williams (1976) noted that culture in its original form was an agricultural term that described an action. It was a verb, not a noun. Culture, as Williams points out, became a noun, i.e., the medium in which things grow, as well as a verb in the nineteenth century. Williams' insights are worth revisiting as they underscore the dynamic as well as the sustaining dimensions of culture. It is the tension between change and sustenance (positively referred to as modernity and tradition; negatively as progress versus reaction) that lies at the core of cultural analysis.

Clifford Geertz's essays in *The Interpretation of Cultures* (1973) opened a door to cultural analysis that extended beyond his home discipline of anthropology to sociology. His essay 'Thick Description' lays out many of the core tenets of cultural analysis. Addressing the problem of meaning, Geertz describes culture as "a web of significance that man himself has spun." He argues that culture is a 'template of and for reality' – suggesting that culture is a recognizable model as well as a guide to action. He illustrates this argument with his clever analysis of a twitch and a wink. Both twitching and winking represent the same neurological phenomena – the uncontrolled movement of the muscles in the eyelid. Yet, depending on the context this movement can mean vastly different things. A wink is a playful gesture; a twitch is a sign of intense nervousness or physical disease. Mistaking the difference between a twitch and a wink could lead to awkward situations for a person who was not able to read the cultural message between the two different meanings that the muscular action conveyed. Mary Douglas (1975) in an essay on "Jokes" underscores the importance of cognition and culture. She argues that 'jokes' are subversions of collective norms and practices. Recognizing the 'subversion,' or 'getting' the joke suggests a deep understanding of a culture in its nonsubversive state.

In 1977, Pierre Bourdieu (1977) took a structuralist approach to culture when he introduced the concept of *habitus* in his *Outline of a Theory of Practice*. *Habitus* defined a structured mental space that a collective inhabited that shaped its behavior, beliefs, and practices. The cultural space of a *habitus*

could be thick or thin. However, *habitus* had a steel-trap quality that made it hard for an individual or a collectivity to transcend its limitations and definitions. Sewell (1992) in a seminal article attempted to insert 'agency' into the concept of *habitus*. Arguing that habitus was too rigid a concept, Sewell theorized cultural structures that he labeled 'schemas.' Transposition among schemas was the key to cultural knowledge and flexibility. Schemas were the deep structures of cognition and practice. The ability to 'transpose' a schema or to engage with different schemas suggested a deep cultural knowledge and is somewhat reflective of Douglas' point vis-à-vis the 'joke.' Sewell compares transposability of a schema to language. To really know a language well and the deep culture that it links to, it is necessary to speak as well as read the language. Spoken language transposes deep cultural structures into everyday knowledge.

In sum, the theorizing of these four seminal scholars converges in ways large and small. Cultural analysis, whatever method a scholar deploys, reveals several core assumptions: first, culture is contingent, context dependent, and not fixed either in time or in space; second, culture is manifested in practice – the things that people do together from speaking the same language to common behaviors and repertoires of evaluation; and third, culture is embedded in symbols or material objects or ritualized behaviors all of which provide grist for the sociological mill that form different media of communication.

This article will trace these three analytics through different sociological groups that have developed around them. Each of these groups has had widespread influence and generated voluminous amounts of research and published work. Space constraints dictate that this article will only cover the leaders in each area.

A Social Science of Culture: Comparison and Contingency

The production of culture approach (Peterson, 1976), a subset of organizational sociology that focused principally on the arts, dominated the subfield of cultural sociology in the mid-1970s. While this strain began to diminish in the early 1980s, there remains a sophisticated subset of sociologists who still engage with the arts (for example, Grazian, 2003; Roy and Dowd, 2010; Lena, 2012). In the 1980s, cultural sociologists began to broaden their horizons and to think about ways to contribute to the discussion of meaning while maintaining scientific integrity as sociologists. Cultural sociologists thought long and hard about the relation between methods and meanings, and were acutely aware of positioning themselves via quantitative sociologists.

An emerging generation of cultural sociologists wrote about methods and meaning broadly conceived (Berezin, 2014). Ann Swidler's (1986) article, 'Culture in Action' was a first step in the direction of extending cultural sociology beyond the arts. Her article engaged both Max Weber and Talcott Parsons on the subject of values. How, she asks, do values or culture shape an individual's choice of actions. Swidler develops the concept of a cultural repertoire or 'tool kit' that individuals, or collectivities, turn to pursue courses of action. The culturally competent individual or collectivity understands what repertoire or tool kit to draw upon – depending upon the situation. Contingency enters Swidler's analysis with her distinction between 'unsettled' and 'settled' lives. During periods of stability, it is difficult to entangle culture from action since there is general societal agreement as to what actions to take in any given situation. When societies become unstable due to contingent events and new paths need to be taken, the ties of culture become clearer as what is cultural becomes disentangled from purely structural elements.

Wendy Griswold's (1987a) article "A Methodological Framework for the Study of Culture" is a relevant and compelling exegesis of how cultural sociologists should address method and design. Griswold's discussion of 'parsimony, amplitude, and plenitude' as criteria for validity suggests that her goal was to push cultural sociology in the direction of scientific method. Her insistence on the necessity of comparison embedded in research design remains an important feature of rigorous cultural analysis. Her careful parsing of the elements of cultural production – intention, reception, comprehension, and explanation – still reflected a concern with artistic production. Griswold's own work focused upon the sociology of literature (for example, Griswold, 1987b).

Yet, her article on methodology opens the door to a broader array of subject matter by defining cultural objects as 'significance embodied in form.' Both of these studies reinforced the anthropological point that culture was not fixed and it varied widely by context. Tapping into those various contexts, whether temporal or spatial, was one way of doing a cultural analysis that could be both interpretive and explanatory. The influence of this approach which is broadly Weberian can be found in an array of articles that varied widely as to subject matter – for example, Alexander's work on the Holocaust (2003b) and Jensen's work on commemoration in Central America (2007).

By the 1990s, cultural sociology was becoming a mature subfield – meaning that it was established part of the discipline. What constituted culture widened as the 1980s turned into the 1990s. Cultural sociology expanded to include different subfields and distinctive conceptual approaches to the sociological study of culture began to consolidate. For example, we can identify Lamont's emphasis upon boundaries (Lamont and Molnar, 2002) and national cultural repertoires (Lamont and Thevenot, 2000); Jeffrey Alexander's (2003a) 'strong program' with its emphasis upon cultural performance; the interactional approach of Nina Eliasoph and Paul Lichterman (2003) and a new entry Stephen Vaisey's (2009) 'dual-process model.'

Culture as Practice: Comparative Method and Social Process

Michele Lamont in a series of books consolidated the role of cross-national research in cultural analysis. Lamont's work speaks to the role of culture in the process of social inequality. Emile Durkheim and Pierre Bourdieu broadly influence the theoretical form of Lamont's work. The work has three distinct dimensions; first, the issue of boundaries – how persons draw distinctions among themselves; second, national repertoires of collective practice; and third, rituals of moral evaluation and worth. The drawing of boundaries was the focus of Lamont's initial work (1992). In a series of interviews, Lamont

interviewed a random sample of middle-class men in four locations – two in France and two in the United States, one a cosmopolitan center, and a second a smaller peripheral city, and asked them how they drew distinctions between themselves and others. In America, she found that money and achievement were core values whereas in France style and bearing in the world carried more weight.

She repeated the study with a shift in focus to the working classes (Lamont, 2000). In her later study, she found that while race was salient in the American case, French workers drew heavily upon the Republican values of inclusion that formed the basis of their national political culture. Individualism was the basis of the American discourse even among working-class men. So that if a person was morally responsible, they were valued members of society independently of their race.

From these two initial studies, Lamont joined in collaboration with Laurent Thevenot (Lamont and Thevenot, 2000) and with a group of collaborators developed the concept of a "national cultural repertoire." What they found in this work was that a clearly identifiable repertoire of ideas, meanings, and practices in response to public and private events can be identified across national spaces. Thevenot in collaboration with Luc Boltanski developed another strand of this work that focused upon evaluation and justification (Boltanski and Thevenot, 1999, 2006). The argument is that the modes in which individuals and collectivities evaluate ideas, people, and events are based upon evaluations of worth. These evaluations of worth are embedded in languages of justification which can be identified depending upon the situation that is being evaluated. The language of justification forms the basis of the "normative principles which underlie the critical activity of ordinary persons" (Boltanski and Thevenot 1999: 364).

This foundation work that brings practice and agency to the field of cultural sociology has generated a corpus of work in social inequality as well as national schemes of evaluation. Annette Lareau, a collaborator of Lamont (Lamont and Lareau, 1988), used these ideas in a seminal work on the different ways that working-class and upper-middle-class children are socialized in the United States (Lareau, 2011). As Lareau demonstrates parenting styles 'concerted cultivation' versus 'natural growth' create the different interpretive scheme that children learn from an early age determines their outcomes in later life. In short, upper-middle-class children learn that the world is there for them to manage; whereas working-class children learn that they are largely victims of circumstances and their best solution is avoidance. Marion Fourcade (2011) has recently drawn out the issue of evaluation in an extended study of natural disaster in France and the United States. Focusing upon two oil spills, she demonstrates that Americans cared more about financial adjudication; whereas the French were more concerned with the destruction of the landscape.

In the early 1990s, comparative historical sociologists and political sociologists (for example, the essays in *State/Culture* (Steinmetz, 1999)) began to incorporate culture in their analyses. Multiple influences drove the political turn in cultural sociology. The 1989 revolutions in Eastern Europe reinvigorated an interest in the cultural underpinnings of democracy (Somers, 1993). Anderson's ([1983]2006) work on the nation as an 'imagined community' sparked a renewed interest in nationalism. The English translation of Jurgen Habermas' *The Structural Transformation of the Public Sphere* and the publication of Calhoun's (1992) anthology on Habermas opened new possibilities in the study of culture and politics. A series of review articles (Berezin, 1994; 1997a) written when the linking of politics and culture was novel among political sociologists identified nodal works and mapped the contours and possible trajectory of the field. Politics and culture is no longer a novel coupling. Wide-ranging empirical studies populate the field. My own work on fascist Italy (Berezin, 1997b) and European populism (Berezin, 2009) serves as one example. It represents the tip of a huge iceberg that includes work on family capitalism (Adams, 2005), nationalism (Brubaker, 1992; Kumar 2003), colonialism (Steinmetz, 2007), religion (Gorski, 2003; Zubrzycki, 2006; Lichterman, 2005), social movements (McAdam and Sewell, 2001), memory and identity (Glaeser, 2000; Olick, 2005; Spillman, 1997), as well as ethnographic accounts of American politics (Perrin, 2006; Lichterman, 1996; Eliasoph, 1998).

Culture as Ritual and Interaction

Sociologists such as Collins (2004) and Eliasoph and Lichterman (2003) have introduced the concept of ritual as they analyzed the cultural meaning of social interaction on a microlevel. Collins argues for a microfoundational account of culture based upon the microrituals of ordinary situations. Based upon extensive field work and ethnographic observation of local political groups, Eliasoph and Lichterman have developed a concept of 'group style' that governs collective political action. Alexander (2003a) offers an extended theoretical engagement with ritual in his concept of the 'strong program' in cultural sociology.

Working in the tradition of Emile Durkheim and Clifford Geertz, Alexander argued that a cultural sociology that adequately captures the complexity of social life must establish the 'autonomy' of culture. In contrast to the 'weak program' that Alexander (2003a) sees as dominating American versions of a sociology of culture, he argues, "What is needed here is a Geertzian 'thick description' of the codes, narratives, and symbols that create the textured webs of social meaning. The contrast here is to the 'thin description' that typically characterizes studies inspired by the weak program, in which meaning is either simply read off from social structure or reduced to abstracted descriptions of reified values, norms, ideology, or fetishism" (p. 13). According to Alexander, culture is a dynamic "…text that is underpinned by signs and symbols that are in patterned relationships to each other" (24).

Alexander operationalized the 'strong program' in a series of empirical studies ranging from the Holocaust (2003b) to the 2008 Obama Presidential campaign (2010). Alexander's study of the Holocaust provides an apt example of his method. The Holocaust became the term used to describe the mass murder of the Jews in the Nazi concentration camps across Europe during World War II. Over time, the Holocaust has come to represent a category of 'ontological evil.' Cultural traumas such as the organized genocide of World War II need to be publicly constructed as evil in order to be collectively recognized as such. In the public's mind, the evil of Nazism ended with World War II and the Nuremburg war trials. The significance of

the genocide emerged in a series of cultural events in the postwar period, and it was during this period that the term Holocaust gained public resonance as a bridging metaphor for evil in particular and even in general. It also became a justification for universal human rights.

In a later essay, Alexander (2004) further developed the 'strong program' by explicitly adding 'social performance' to a developing theory of 'cultural pragmatics.' Borrowing from performance theory, Alexander incorporates meaning structures, contingency, power, and materiality into cultural action. He specifies that cultural performances both 'fuse' and 're-fuse' – highlighting the malleability of meaning depending upon context. Cultural performance identifies collective representations, actors, audiences, means of symbolic production, *mise-en-scene*, and social power. These elements act in dynamic interaction with each other to produce an ongoing and changing cultural narrative that includes cultural producers and their audience.

Going Forward

Cultural analysis is decidedly more global today than when it first captured the imagination of sociologists in the 1980s. In addition to being represented in all of the major journals of sociology, there are two new journals *The American Journal of Cultural Sociology* (2013–) based at Yale and the *European Journal of Cultural and Political Sociology* (2014–) based in Europe that have institutionalized the study of culture as a subfield in global sociology. Cultural sociology is reflecting upon its methods (see essays in Berezin, 2014) and pondering its future.

We have now moved to a stage where 'big data' is making inroads in cultural analysis. Mohr and his contributors in a recent special issue of *Poetics* (2013) illustrate the utility of 'big data' to a range of cultural questions. Christopher Bail (2012) used 'big data' to good effect in a recent *ASR* article on how extraordinary terrorist threats, many of which never occur, end up due to media selection as defining the global Islamic community.

The future of cultural sociology is wide open – with a plurality of methods and a multiplicity of venues. Patterson (2014) reminds us that there is still 'sense' to be made of culture. As cultural anthropologists recognized in the postwar period, culture points to differences and similarities among groups in near and far places. The ontological status of culture is global. In sum, cultural sociology is the future.

See also: Collective Memory, Psychology of; Comparative Method in Anthropology; Epistemic Cultures; Formal Methods of Cultural Analysis; Norms; Political Anthropology; Ritual.

Bibliography

Adams, Julia, 2005. The Familial State: Ruling Families and Merchant Capitalism in Early Modern Europe. Cornell University Press, Ithaca, NY.

Alexander, Jeffrey C., 2003a. The Meanings of Social Life: A Cultural Sociology. Oxford University Press, New York.

Alexander, Jeffrey C., 2003b. On the social construction of moral universals: the holocaust from war crime to trauma drama. In: The Meanings of Social Life: A Cultural Sociology. Oxford University Press, New York.

Alexander, Jeffrey C., 2004. Cultural pragmatics: social performance between ritual and strategy. Sociological Theory 22, 527–573.

Alexander, Jeffrey C., 2010. The Performance of Politics. Oxford, NY.

Almond, Gabriel A., Verba, Sidney, 1963. The Civic Culture: Political Attitudes and Democracy in Five Nations. Princeton University Press, Princeton, NJ.

Anderson, Benedict, 2006. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London.

Bail, Christopher A., 2012. The fringe effect: civil society organizations and the evolution of media discourse about Islam since the September 11th attacks. American Sociological Review 77, 855–879.

Banfield, Edward C., 1958. The Moral Basis of a Backward Society. Free Press, Glencoe, IL.

Berezin, Mabel, 1994. Fissured terrain: methodological approaches and research styles in culture and politics. In: Crane, Diana (Ed.), Sociology of Culture: Emerging Theoretical Perspectives. Basil Blackwell, London, pp. 91–116.

Berezin, Mabel, 1997a. Politics and culture: a less fissured terrain. Annual Review of Sociology 23, 361–383.

Berezin, Mabel, 1997b. Making the Fascist Self: The Political Culture of Interwar Italy. Cornell University Press, Ithaca, NY.

Berezin, Mabel, 2009. Illiberal Politics in Neoliberal Times: Culture, Security and Populism in the New Europe. Cambridge, UK and NY.

Berezin, Mabel (Guest Editor), 2014. Theme issue: methods, materials and meanings: designing cultural analysis. Qualitative Sociology 37 (2).

Boltanski, Luc, Thevenot, Laurent, 1999. The sociology of critical capacity. European Journal of Social Theory 2 (3), 359–377.

Boltanski, Luc, Thevenot, Laurent, 2006. On Justification: Economies of Worth. Princeton University Press, Princeton.

Bourdieu, Pierre, 1977. Outline of a Theory of Practice. Cambridge University Press, NY.

Brubaker, Rogers, 1992. Citizenship and Nationhood in France and Germany. Harvard University Press, Cambridge, MA.

Collins, Randall, 2004. Interaction Ritual Chains. Princeton University Press, Princeton.

Calhoun, Craig, 1992. Habermas and the Public Sphere. MIT Press, Cambridge, MA; London, England.

Douglas, Mary, 1975. Implicit Meanings: Essays in Anthropology. Routledge & K. Paul, London.

Eliasoph, Nina, Lichterman, Paul, 2003. Culture in interaction. The American Journal of Sociology 108, 735–794.

Eliasoph, Nina, 1998. Avoiding Politics: How Americans Produce Apathy in Everyday Life. Cambridge University Press, Cambridge, England.

Fourcade, Marion, 2011. Cents and sensibility: economic valuation and the nature of "nature". American Journal of Sociology 116, 1721–1777.

Geertz, Clifford, 1973. The Interpretation of Cultures; Selected Essays. Basic Books, New York.

Glaeser, Andreas, 2000. Divided in Unity: Identity, Germany, and the Berlin Police. University of Chicago Press, Chicago.

Gorski, Philip S., 2003. The Disciplinary Revolution: Calvinism and the Rise of the State in Early Modern State. University of Chicago Press, Chicago.

Grazian, David, 2003. Blue Chicago: The Search for Authenticity in Urban Blues Clubs. University of Chicago Press, Chicago.

Griswold, Wendy, 1987a. A methodological framework for the sociology of culture. Sociological Methodology 17, 1–35.

Griswold, Wendy, 1987b. The fabrication of meaning: literary interpretation in the United States, Great Britain, and the West Indies. The American Journal of Sociology 92, 1077–1117.

Harrison, Lawrence E., Huntington, Samuel P. (Eds.), 2000. Culture Matters. Basic Books, New York.

Huntington, Samuel P., 1996. The Clash of Civilizations and the Remaking of World Order. Simon & Schuster, New York.

Jansen, Robert S., 2007. Resurrection and appropriation: reputational trajectories, memory work, and the political use of historical figures. American Journal of Sociology 112, 953–1007.

Kumar, Krishan, 2003. The Making of English National Identity. Cambridge University Press, Cambridge.

Lamont, Michele, Lareau, Annette, 1988. Cultural capital: allusions, gaps and glissandos in recent theoretical developments. Sociological Theory 6, 153–168.

Lamont, Michele, Thevenot, Laurent (Eds.), 2000. Rethinking Comparative Cultural Sociology: Repertoires of Evaluation in France and the United States. Cambridge University Press, Cambridge.

Lamont, Michele, Molnar, Virag, 2002. The study of boundaries in the social sciences. Annual Review of Sociology 28, 167–195.
Lamont, Michele, 1992. Money, Morals, and Manners: The Culture of the French and American Upper-Middle Class. University of Chicago Press, Chicago.
Lamont, Michele, 2000. The Dignity of Working Men: Morality and the Boundaries of Race, Class, and Immigration. Russell Sage Foundation; Harvard University Press, New York; Cambridge, MA.
Lareau, Annette, 2011. Unequal Childhoods: Class, Race, and Family Life Lareau. University of California Press, Berkeley.
Lena, Jennifer C., 2012. Banding Together: How Communities Create Genres in Popular.
Lichterman, Paul, 2005. Elusive Togetherness: Church Groups Trying to Bridge America's Divisions. Princeton University Press, Princeton, NJ.
Lichterman, Paul, 1996. The Search for Political Community: American Activists Reinventing Commitment. Cambridge University Press, Cambridge.
McAdam, Doug, Sewell Jr., William H., 2001. It's about time: temporality in the study of social movements and revolutions. In: Aminzade, Ronald R., et al. (Eds.), Silence and Voice in the Study of Contentious Politics. Cambridge University Press, New York, pp. 89–125.
Mohr, John W., Bogdanov, Petko, 2013. Introduction – topic models: what they are and why they matter. Poetics 41, 545–569.
Olick, Jeffrey K., 2005. In the House of the Hangman: The Agonies of German Defeat, 1943–1949. University of Chicago Press, Chicago.
Parsons, Talcott, 1951. The Social System. The Free Press, New York.
Patterson, Orlando, 2014. Making sense of culture. Annual Review of Sociology 40 (July): 1–30.
Perrin, Andrew J., 2006. Citizen Speak: The Democratic Imagination in American Life. University of Chicago Press, Chicago.
Peterson, Pete A., 1976. The Production of Culture. Sage, Beverly Hills, CA.
Roy, William G., Dowd, Timothy J., 2010. What is sociological about music? Annual Review of Sociology 36, 183–203.
Sewell Jr., William H., 1992. A theory of structure: duality, agency, and transformation. The American Journal of Sociology 98, 1–29.
Somers, Margaret R., 1993. Citizenship and the place of the public sphere: law, community, and political culture in the transition to democracy. American Sociological Review 58, 587–620.
Spillman, Lyn, 1997. Nation and Commemoration: Creating National Identities in the United States and Australia. Cambridge University Press, Cambridge.
Steinmetz, George, 2007. The Devil's Handwriting: Precoloniality and the German Colonial State in Qingdao, Samoa, and Southwest Africa. University of Chicago Press, Chicago.
Steinmetz, George, 1999. State/Culture: State-Formation after the Cultural Turn. Cornell University Press, Ithaca.
Swidler, Ann, 1986. Culture in action: symbols and strategies. American Sociological Review 51, 273–286.
Vaisey, Stephen, 2009. Motivation and justification: a dual-process model of culture in action. American Journal of Sociology 114 (6), 1675–1715.
Weber, Max, 1968. The Methodology of the Social Sciences; Max Weber on the Methodology of the Social Sciences (Edward Shils, Trans.). Free Press, New York.
Williams, Raymond, 1976. Keywords. Oxford, New York.
Willis, Paul E., 1977/1981. Learning to Labor: How Working Class Kids Get Working Class Jobs. Columbia University Press, New York.
Zubrzycki, Geneviève, 2006. The Crosses of Auschwitz: Nationalism and Religion in Post-Communist Poland. University of Chicago Press, Chicago.

Culture-Specific Psychiatric Disorders

Oyedeji Ayonrinde, Bethlem Royal Hospital, South London and Maudsley NHS Foundation Trust, Beckenham, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W.G. Jilek, volume 18, pp. 12272–12277, © 2001, Elsevier Ltd.

Abstract

Culture influences the presentation and perception of mental distress. It also determines the labeling, idioms of distress, and explanatory models related to mental disorders. Culture-specific or culture-bound syndromes are uncommon, localized symptom patterns of mental distress. The 'exotic' non-Western classificatory histories of most of these disorders are enlightening. As culture is nonstatic, the diagnostic consistency, epidemiology, and long-term stability of culture-specific disorders require further investigation.

Introduction

Humans in all societies and cultures experience mental distress and at times mental disorders or illness. The concept of a troubled mind is universal with recognition given to more disturbing or abnormal mental experiences.

Cultures influence the way this distress is articulated whether by verbal or behavioral means. Cultural presentation of mental distress may share similarity across groups with shared history, or shared concepts of the mind and body. Similarly, these differences may be so localized or unfamiliar as to suggest a notion of a culture-specific or culture-bound disorder.

Sociologists, anthropologists, and psychiatrists have driven investigations into the mind and mental disorders across cultures. Through colonialism, psychiatrists and anthropologists from Western Europe and North America have influenced the early description and dissemination of accounts of alien and exotic presentations of mental distress they have come across or heard of.

Published descriptions of exotic manifestations of mental disorders in indigenous or native groups have fascinated clinicians and nonclinicians alike. Accounts and images of native health practice, unfamiliar idioms of distress, and treatments have permeated psychiatric literature and nosology. The medicalization of cultural expression of distress has led to a broader concept and evolution of 'cultural syndromes' and disorders in social science literature.

Culture has an enmeshed relationship with health as follows:

- Distinguishing between 'normality' and 'abnormality.'
- Etiological influence on some disorders.
- Influencing clinical presentation.
- Influencing interpretation of behaviors.
- Influencing rates and distribution of illness.
- Determining the recognition of disorders.
- Generating labels for disorders.
- Explanatory models of disorders.
- Determining treatment options.
- Determining care pathways.
- Influencing the outcome of treatment interventions.

Short History of Cultural Syndromes

Yap (1962) described culture-bound psychogenic psychoses as a way of harmonizing a wide range of disorders with a complex sociocultural influence on their presentation. In 1969, Yap renamed the concept as 'culture-bound syndromes' with recognition that not all were psychotic presentations.

These syndromes were described as "rare, exotic, unpredictable, and chaotic behaviours at their core among uncivilised people" (Bhugra and Jacob, 1997) They suggested these behaviors were described against a backdrop of Western worldviews and diagnostic systems with little appreciation of socio-environmental influences and the tolerance of symptoms within the culture, hence ethnocentric. In some instances, culturally acceptable behavior was pathologized in non-Western populations. Similarly, a contextual frame of reference was advocated for the full appreciation of unfamiliar clinical presentations (Hughes and Wintrob, 1995).

Society and Disorders

Geopolitical, economic, and health care resource availability influences the rates, distribution, and treatment of common mental disorders. As culture determines what constitutes aberrant behavior or mental disorder, it entirely permeates spheres of life of different populations.

Early descriptions of cultural syndromes emanated from the Far East as sociologists and medical anthropologists explored these parts. Reversing the Western ethnocentric lens, Type A behavior (Hughes, 1996) and bulimia (Littlewood, 1996) have been suggested as Western culture-bound syndromes. Type A personality is a driven, goal pursuant cognitive and behavioral state of cultural salience with significant time pressure and immense frustration when goals are not achieved. Explored in the cultural context, this may highlight the importance of personal achievement in 'individualistic' or egocentric societies with a greater emphasis on 'I-ness.' This would be more difficult to sustain in cultures with greater kinship or strong family links as in collectivist or 'socio-centric' societies which encourage kinship sharing time, financial, and emotional resources (Hofstede, 1980). Likewise, culture plays an important role on perceptions of body image (bulimia, anorexia nervosa, and body dysmorphic disorder) and interacts with socioeconomic factors such as food availability. These cultural values influence individuals and societal acceptance of behaviors.

The historical and contemporary status of culture-bound syndromes in current psychiatric practice is explored in this

section through more detailed reviews of 'windigo,' 'brain fag syndrome,' and 'dhat' – previously documented cultural disorders with folk, colonial, and ancient histories respectively.

Over the last half century, the descriptions and nosology of unfamiliar or exotic mental disorders have evolved in Western scientific literature and diagnostic manuals. A brief nosological list illustrates this:

1. 1962 – Ethnic or exotic psychosis (Yap, 1962)
2. Culture-bound psychogenic psychosis (Yap, 1962)
3. Culture-reactive (ICD, 10)
4. 1969 – Culture-bound syndromes (Yap, 1969)
5. 1992 – Culture-specific disorders (World Health Organization (WHO), International Classification of Diseases 10th edition, ICD 10, 1992)
6. 1994 – Culture-bound syndromes (American Psychiatric Association (APA), Diagnostic and Statistical Manual DSM IV)
7. 2000 – Culture-bound syndromes (American Psychiatric Association (APA), Diagnostic and Statistical Manual, DSM-IV-TR)
8. 2013 – Cultural concepts of distress (American Psychiatric Association (APA), Diagnostic and Statistical Manual, DSM 5)

Stability

Descriptions of some culture-bound syndromes have shown subtle changes over recent years; this would be consistent with core diagnostic features of some psychiatric disorders (e.g., personality disorders, developmental disorders) over recent decades. Furthermore, culture is not a static phenomenon and both *emic* and *etic* contributions enrich our understanding of distress across cultures. The stability of syndromes is explored in greater detail below in the section on – Validity of Culture-Bound Syndromes – Brain Fag, Windigo, and Dhat.

Bound

The concept of a disorder being bound to a specific culture is controversial and may become less relevant with increased global mobility and communication beyond geographical or loose cultural boundaries.

Hughes (1996) noted that diagnostic labels of exotic syndromes and atypical psychoses are ethnocentrically based on the premise that these disorders deviate from recognized and perhaps standard Western diagnoses.

Exotic indicates an unusual, foreign, nonindigenous, alien, or different condition. However, these differences do not ensure marked variation in psychopathology and require emic awareness from within the culture before considering unfamiliar behavior pathological.

Culture-Bound Syndromes in the Diagnostic Manuals (DSM and ICD)

ICD-10 (WHO, 1992)

The 10th edition of the International Classification of Diseases (ICD 10) of the WHO recognizes culturally uncommon symptom patterns and presentations referred to as culture-specific disorders. While diverse in characteristics, culture-specific disorders share two key characteristics, namely:

1. Difficult to accommodate in established international diagnostic categories.
2. Primary description in a particular population or cultural area and subsequent association with this group or culture.

The ICD 10 queries the position of culture-specific disorders, and the international experts argue that the differences observed in cultural disorders are a degree of variation from existing and established familiar disorders. The current associations between cultural syndromes and recognized psychiatric categories in the ICD 10 need a greater evidence base. It is anticipated that a growth in mental health research in less developed countries will contribute to this and influence the salience given to cultural issues and specific syndromes in the future ICD-11.

DSM-IV-TR (APA, 2000)

In this American classification manual, the term culture-bound syndrome denoted "recurrent, locality-specific patterns of aberrant behavior and troubling experience that may or may not be linked to a particular DSM IV diagnostic category."

Characteristics of Culture-Bound Syndromes DSM-IV-TR

1. Indigenously considered illness or afflictions – therefore a recognition within the society as a deviation from normal or healthy presentation.
2. Local names – the ascription of a specific local name to the experience of mental distress. Often in the indigenous or main language of communication and may include features of folk diagnostic categories.
3. Symptoms, course, and social response often influenced by local cultural factors – for instance help seeking, folk healing systems, and remedies based on the explanatory model of the experiences.
4. Limited to specific societies or cultural areas – this may be a geographical region, areas with shared ethnic history or identity – for instance, Ayurvedic health beliefs in Indian communities outside Asia.
5. Localized – do not span different regions therefore within circumscribed areas and not globally recognized.

DSM 5 (APA, 2013) – New Advances

The newer DSM 5 reviewed and redefined the concept of culture-bound syndromes preferring the term *cultural concepts of distress*. These cultural concepts are defined as "ways cultural groups experience, understand, and communicate suffering, behavioural problems, or troubling thoughts and emotions."

This DSM-5 review elaborated on three important cultural concepts, namely:

Syndromes – clusters of symptoms and attributions occurring among individuals in specific cultures.

Idioms of distress – shared ways of communicating, expressing, or sharing distress.

Explanations – labels, attributions suggesting causation of symptoms, or distress.

The APA manual also recognizes the controversy concerning the concept of culture-bound syndromes and the DSM 5 contends the terminology; culture-bound syndromes ignore the impact of culture on more specific experience and expression of distress. This more inclusive viewpoint questions the validity of culturally circumscribed collections of symptoms and syndromes. DSM 5 emphasizes all mental distress is culturally framed and different populations have culturally determined ways of communicating distress, coping, seeking treatment, and explanations of causality.

Both the WHO and the APA classificatory systems of mental disorders review the evidence base for nosological categories with the support of international scientific expert panels incorporating contemporary anthropological, social, and psychiatric research findings as global cultures and societies change.

Not all disorders are likely to endure rigorous scientific reviews as evidenced by changes in which DSM-IV-TR described 25 syndromes in 2000 while the DSM 5 published in 2013 explores only 9 cultural concepts of distress (culture-bound syndromes) in the light of the current research evidence. (See table below.)

For detailed reviews and historical context on presentations the early text by Simon and Hughes (1985) is insightful. The DSM IV and DSM 5 Glossaries (APA; 2000, 2013) also provide details of different culture-bound syndromes and concepts of distress.

Description of Specific Culture-Bound Syndromes/ Concepts of Distress

Ataque de Nervios ('Attack of Nerves')

This is a syndrome predominantly among persons of Latino heritage characterized by intense emotional expression of anger, acute anxiety, screaming, grief, and shouting. In addition, individuals experience heat in the chest spreading to the head and may show physical and verbal aggression. In some cases depersonalization, derealization, and memory impairment may occur with fainting or seizure-like shaking. Suicidal ideation may also manifest.

A shared feature of attacks is a presentation of being out of control.

Context: Stressful life event, for example, the death of a loved one or serious conflict with a close individual often against a backdrop of chronic difficulties.

Associated/differential disorders: Panic attack, dissociative disorder, trauma, or stress-related disorder.

Prevalence: Unknown.

Brain Fag Syndrome ('*Mental Fatigue*')

This is a culture-bound syndrome affecting West African students which is attributed to overwork. It is reported to be characterized by a loss of ability to concentrate, learn,

	Culture-bound syndrome/ cultural concept of distress	*Brief description/ distress*	*Attributed country/ region*	*ICD-10 (12 syndromes) described*	*DSM-IV-TR (25 syndromes) described*	*DSM 5 (9 syndromes) described*
1.	Amok	'Outburst of violence and homicidal behavior'	Malaysia	x	x	
2.	Ataque de nervios	'Attack of nerves'	Latin America/Latino Caribbean	See below	x	x
3.	Bilis and colera	'Anger or rage'	Latino		x	
4.	Boufee delirante	'Sudden agitation and aggression'	West Africa, Haiti		x	
5.	Brain fag	'Mental fatigue'	West Africa	See below for discussion on the colonial history of 'brain fag' syndrome	x	
6.	Dhat	'Semen loss'	India	x See discussion on ancient history of 'dhat' below	x	x
7.	Falling-out (blacking out)	'Sudden collapse'	South United States, Caribbean		x	
8.	Ghost sickness	'Preoccupation with death and deceased'	American Indian (native American)		x	
9.	Hwa-byung	'Anger syndrome'	Korea		x	
10.	Koro	'Genital retraction'	Malaysia	x	x	
11.	Khyâl Cap	'Wind attacks'	Cambodia	See below		x
12.	Kufungisisa		Zimbabwe			x

Continued

	Culture-bound syndrome/ cultural concept of distress	*Brief description/ distress*	*Attributed country/ region*	*ICD-10 (12 syndromes) described*	*DSM-IV-TR (25 syndromes) described*	*DSM 5 (9 syndromes) described*
13.	Latah	'Hypersensitivity to sudden fright'	Malaysia, Indonesia	x	x	
14.	Locura	'Severe chronic psychosis'	Latino, United States, Latin America		x	
15.	Mal de ojo	'Evil eye'	Mediterranean		x	
16.	Maladi moun	'Humanly caused illness'	Haiti			x
17.	Nervios	'Nerves'	Latino, Latin America	x See below	x	x
18.	Pibloktoq	'Dissociative excitement'	Arctic	x	x	
19.	Qi-gong psychotic reaction	'Acute dissociative paranoid reaction'	China		x	
20.	Rootwork	'Witchcraft, voodoo or hex illness'	Southern USA, Caribbean		x	
21.	Sangue dormido	'Sleeping blood'	Portuguese Cape Verde		x	
22.	Shenjing shuairuo	'Weakness of the nervous system'	China	See below	x	x
23.	Shen-k'uei	'Excessive semen loss anxiety'	China	x	x	
24.	Shin-byung	'Anxiety, dissociation, and ancestral possession'	Korea		x	
25.	Spell	'Trance state'	Southern USA		x	
26.	Susto	'Fright' or 'soul loss'	Latino/South America	x See below	x	x
27.	Taijin kyofusho	'Interpersonal fear disorder'	Japan	x See below	x	x
28.	Zar	'Spirit possession'	Ethiopia, Somalia, Egypt, Sudan, North Africa, Middle East		x	
29.	**Others** Ufufuyane, saka Uqamairineq Windigo*	'Cannibal madness'		See discussion below on 'windigo' medicalization of a myth		

x – Indicates a description in the classificatory directory (ICD) or diagnostic manual (DSM).

remember, or think and usually accompanied by sensations of pain, pressure, or tightness around the head or neck and blurred vision (DSM-IV-TR, 2000).

The DSM 5 however limited the reference to brain fag to a brief illustration of symptoms of 'too much thinking' in a West African context. Initially described in Nigeria and later in West Africa and other African countries, this has been reported as a cluster of symptoms characterized by somatic, affective, anxiety, and cognitive complaints in African individuals engaged in educational pursuit or individuals involved in brain work (16).

Contemporary texts report that the term brain fag syndrome was initially coined in Nigeria following the use of the phrase 'brain fag' rooted in a form of pidgin Nigerian-English, and was an idiom of distress colloquially referring to mental exhaustion.

Context: Study, education, and mentally demanding activity.

Prevalence: Unknown.

Associated disorders: Depression, generalized anxiety disorder, somatization disorders.

Dhat Syndrome ('*Semen Loss*')

Coined in South Asia, this describes a range of symptoms attributed to semen loss. It is associated with idioms of distress

of fatigue, anxiety, impotence, weight loss, depressive mood, and multiple somatic complaints.

The core feature is distress regarding the loss of '*dhat*' – a white discharge on micturition or defecation.

Context: The cultural salience is related to the loss of *dhatu* (semen), one of seven health preserving essential bodily fluids in Hindu Ayurvedic medicine.

Prevalence: Up to 64% men attending a clinic in India with sexual complaints and 30% of male attendees of a general clinic in Pakistan (DSM-IV-TR, 2000).

Associated disorders: Depressive disorder, generalized anxiety disorder, erectile dysfunction, sexual dysfunction, and study problem.

Khyâl Cap ('Wind Attacks')

A syndrome found among Cambodian populations characterized by panic attacks, palpitations, breathlessness, dizziness, and coldness of extremities. Additional autonomic features of anxiety may be present. The key belief during episodes is one that a wind-like substance (*Khyâl*) may rise with blood in the body causing compression of the vital organs of the chest; hence respiratory symptoms and cranium with dizziness, blurred vision, or fatal fainting.

Context: Worrisome thoughts, postural hypotension, unique smells, or crowded places.

Prevalence: Unknown.

Associated disorders: Panic attacks, agoraphobia, generalized anxiety disorder, and posttraumatic stress disorder.

Nervios ('Nerves')

This is a common Latino idiom of distress referring to a state of vulnerability to stressful life events and challenging circumstances. The range of symptoms includes emotional distress, somatic complaints such as 'brain aches,' abdominal, paresthesia, tremors, and other autonomic symptoms. Individuals may be tearful and report impaired concentration and sleep difficulty.

Prevalence: Unknown.

Associated disorders: Depressive disorders, generalized anxiety disorder, dissociative disorder, psychosis, and in some cases, no mental disorder.

Shenjing shuairuo ('Weakness of the Nervous System')

A Chinese syndrome characterized by a symptom cluster of weakness such as mental fatigue, emotional expression (feeling vexed), excitement, nervous pain, and sleep disturbance. These are elucidated in the *Chinese Classification of Mental Disorders (CCMD)*.

Context: Feeling vexed involves worry and irritability associated with unfulfilled desires and internal conflict. There may be work or family stressors, unmet expectations, loss of face, and the loss of vitality.

Prevalence: Unknown.

Associated disorders: Depression, generalized anxiety disorder, phobic disorders, and posttraumatic stress disorder; 45% do not meet diagnostic criteria for a mental disorder. The symptom pattern of Shenjing shuairuo bears strong similarity to other conditions such as brain fag syndrome reported in West Africa.

Susto ('Fright')

This is based on a Latino idiom of distress in the Americas associated with misfortune and distress following a frightening event. The explanation of distress is that the soul leaves the body leading to sickness and an unhappy mood. Susuto is characterized by impaired sleep, disturbing dreams, appetite disturbance, impaired motivation, and low self-worth. Somatic complaints such as aches, pains, pallor, or gastrointestinal symptoms may be present.

Three susto subsyndromes have been described: interpersonal, traumatic, and somatic.

Context: The presentation occurs days to weeks or years following a frightening experience. These may be interpersonal situations, supernatural, or extreme natural events.

Prevalence: Unknown.

Associated disorders: Depressive disorder, posttraumatic stress disorder, and somatoform disorders.

Taijin kyofusho ('Interpersonal Fear Disorder')

A Japanese cultural syndrome characterized by anxiety associated with avoidance of interpersonal situations based on the belief that the individual does not appear adequate or is offensive to others in social situations. Some individuals are concerned about their body odor, blushing, inappropriate eye contact, and clumsy body or facial movements. There may also be concerns about body morphology or perceived malformations.

Context: Social situations and the need to maintain appropriate social norms of behavior in interpersonal hierarchies.

Prevalence: Unknown.

Associated disorders: Social anxiety disorder, body dysmorphic disorder, obsessive-compulsive disorder, and delusional disorder.

Stability of Culture-Bound Syndromes: An Exploration of Brain Fag, Windigo, and Dhat syndromes

Brain Fag – A Colonial History

Social science and psychiatric literature indicates that the term 'brain fag' emanated in the 1960s in Africa (Prince, 1960). This etymology has been disputed though, as there is historical evidence of use of the term with analogous connotations in early nineteenth-century Britain (Tuke, 1880; Gillespie, 1926, 1933; Ayonrinde, 2008a,b; 2009) and this consequently spread across the British Empire. In nineteenth-century parlance, to be 'fagged' or 'fag' indicated fatigue, exhaustion, or tiredness.

Concern about the impact of study on education was rife in post-industrial Britain in the 1800s with the growth of 'brain worker' vocations such as lawyers, book-keepers, teachers, and students.

Concern about the hazards of study and risk of nervous diseases, insanity, lethargy, and even death from 'over study' led to a groundswell of anxiety. This spread transatlantically and in the late nineteenth and early twentieth centuries, American news reports raised an alarm about an illness in

England called 'brain fag' striking British 'brain workers' and advancing toward the developing American nation.

Medical lexicons of the time defined brain fag as follows:

Brain fag: exhaustion due to overwork with the brain (American illustrated medical dictionary, 1901).

Brain fag: the effects on the brain of mental overwork (Lippincott's new medical dictionary, 1910).

About 150 years later, observations in Africa of a new cultural syndrome were as follows:

> The distinctive symptoms of the syndrome are intellectual impairment, sensory impairment (chiefly visual), and somatic complaints most commonly of pain or burning in the head and neck. In addition there are a variety of other complaints that often militate against the student's ability to study. The patients have an unhappy, tense facial expression and a characteristic gesture is frequently noted – that of passing the hand over the surface of the scalp or rubbing the vertex of the skull, especially when discussing a question that requires concentration.
>
> Prince, 1960

Symptoms presented during periods of intensive reading or study prior to or following examinations or intensive study.

> 1960s literature stated – "It is clear that the syndrome is related to 'westernization,' for prior to the 19th century when significant European influence commenced, the written word and the student life, which are so intimately bound up with the syndrome, symptoms were unknown to the inhabitants of Southern Nigeria" … "the syndrome is rarely seen in the European." Anthropological thinking of the time proposed "It would appear then that the illness is related in some way to the fusion of European learning techniques with Nigerian personality."
>
> Prince, 1960

Prince suggested that "European learning techniques emphasize isolated endeavor, individual responsibility, and orderliness – unlike the Nigerian by reason of the collectivistic society from which he derives." This historical concept is similar to the concept of 'individualist' and 'collectivist' differences in cultures derived globally by Hofstede (1980) and provides useful insight into the understanding of distinct societies.

Contemporary epidemiological evidence regarding brain fag and other culture-bound syndromes is sparse compared with other psychiatric disorders.

Since the initial description by Prince (1960), Nigerian psychiatrists have debated the relevance of brain fag. Morakinyo (1980) suggested a psycho-physiological and cognitive basis to the disorder among African students compared with nonstudents reinforcing the relevance of brain fag. Ola et al. (2009) describe a clinical tetrad of somatic complaints: cognitive impairment; sleep difficulties; and other somatic impairments with a conceptual history based on traditional medicine, psychoanalysis, biopsychological, and transcultural psychiatry etiologies. They contend brain fag is subjectively valid within a psychiatric, psychological, and sociological framework. However, other Nigerian psychiatrists (Jegede, 1983; Ayonrinde and Obuaya, 2008a,b) question the validity of this syndrome stressing it is nosologically related to depressive and anxiety disorders. All of the psychiatrists agree that the natural history, management, and prognosis of brain fag are unknown.

Syndrome?

A syndrome is a combination of signs and symptoms that form a distinct clinical entity or disorder. Individual symptoms are amalgamated by the healer or clinician based on an understanding of established patterns of presentation. In the case of brain fag, symptoms of depression, somatization disorder, and anxiety disorder occur with notable areas of overlap. With such heterogeneity of symptoms, small sample sizes and the absence of pathognomonic symptoms and signs raise doubts regarding the diagnostic validity of a study-related stress disorder. Impaired mental function is common to many mental disorders.

Culture-Bound?

The concept of culture-bound indicates a syndrome restricted to or associated with a specific culture or peoples. Brain fag has been reported in culturally distinct parts of West, North, South, and East Africa (Harris, 1981; Peltzer et al., 1998). The diverse historical, linguistic, religious, and educational backgrounds of the African countries highlight the heterogeneity of symptoms.

With an etymology rooted in nineteenth-century Britain, current use is arguably not symptom-specific, culture-bound, or unique to any geographical region. Interestingly, *Shenjing shuairuo* symptoms from China are similar to the West African syndrome.

Windigo – The Medicalization of a Cultural Myth

Mythical and historical accounts of a demonic half-man half-beast creature eating the flesh of humans caused anxiety in some regions of the United States and Canada. Folklore in which predominantly male victims developed depressive, homicidal, and suicidal thoughts with a delusional compulsion to eat human flesh were a source of intrigue. Reports indicate sufferers were socially isolated, ostracized, or slain when found. For several years, anthropologists and psychiatrists believed windigo was a culture-bound syndrome but emic understanding about 'windigo' monster myths and the social implication of the narrative led to a questioning of the legitimacy of this syndrome. Interestingly, an indirect outcome was social control and strong negative views regarding homicidal behavior in communities. Windigo is no longer referred to in contemporary literature as a 'living' disorder. Ethnohistorians and anthropologists have on the whole concluded that windigo is not a mental disorder or syndrome but a taboo reinforcing folklore (Marano, 1983).

Dhat or Semen-Loss Anxiety Syndrome – An Ancient Ayurvedic Disorder

Derived from the Sanskrit word *dhatu* (metal), *dhat* is a colloquial term for semen. In 1960, Wig described dhat syndrome

characterized by vague somatic complaints such as fatigue, weakness, anxiety, loss of appetite, guilt, and sexual dysfunction attributed to semen loss following nocturnal emissions, micturition, or masturbation.

Ancient Indian Ayurvedic texts between 5000 BC and the seventh century depict symptoms of semen loss and associated anxiety. Semen production was described as "….food converts to blood which converts to flesh which converts to marrow and ultimately to semen," a process reporting 40 days for 40 drops of food to convert to a drop of blood and 40 drops of blood to convert to a drop of flesh (Bhugra and Buchanan, 1989). The salience of semen across the Indian subcontinent cultures is strongly reinforced by these ancient descriptions and contemporary remedies promoted in newspapers, TV, and by posters on walls.

A study of dhat in North India (Malhotra and Wig, 1975) found higher rates in lower social classes, with the promotion of marriage, diet, and avoidance of masturbation, unscrupulous company, and exotic literature. Labeled the 'sex neurosis of the Orient' vulnerability was linked with beliefs regarding semen loss. Alternative therapies and Western medical interventions were sought for treatment (Malhotra and Wig, 1975).

Chadda and Ahuja (1990) studied patients complaining of dhat in their urine and found that 80% had 'hypochondriacal' symptoms, a view shared that dhat is a form of hypochondriasis. In a sexual dysfunction clinic, Bhatia and Malik (1991) found somatic complaints (physical weakness), sexual difficulties, and depression in half the patients. Janaki et al. (2012) observed that the commonest misconceptions among 24% psychiatric clinic patients was that masturbation was immoral and may lead to loss of potency. Psychosexual dysfunctions associated with dhat include erectile dysfunction (22–62%) and premature ejaculation (22–44%), and unsurprisingly both sexual dysfunction and common mental disorders accompany semen-loss anxiety (Prakash, 2007).

Dhat symptoms were also found outside the Indian subcontinent or Indian diaspora. Jadhav (2007) observed 48% of White British men with depression-linked semen retention with psychological problems and advocated a *semen retention syndrome* characterized by loss of energy, diminished libido, multiple somatic symptoms, depression, and anxiety attributed to retention of semen. Concern about semen in depressed and anxious men is not limited to Asian cultures and shares diagnostic features with other common mental disorders (Udina et al. (2013). Paradoxically, idiomatic use of dhat is a less stigmatizing means of seeking help for sexual dysfunctions from health professionals.

Male potency and virility are entwined with ideas of masculinity, symbolized by semen in some cultures, and dhat through masturbation, nocturnal emission, and the ensuing semen-loss anxiety has connotations of weakness. Historically, such anxiety about semen loss in America and Britain was revitalized with nutritional additives such as cornflakes recommended by Kellogg. In contemporary Indian society, dietary supplements and herbal remedies are offered by traditional healers to enhance virility.

The Future of Culture-Bound Syndromes

Culture-bound syndromes or culture-specific manifestations of distress continue to evolve as knowledge of different cultures increases and the shared features with established disorders are recognized. Their future as a range of disorders appears uncertain. While dhat has endured many centuries of cultural change in the Indian subcontinent, windigo is practically nonexistent, and brain fag is going through substantial nosological decline. Time and further research will determine which syndrome survives.

Impact of Urbanization

Differences in urban and rural life culture could be marked within national and geographical boundaries. In fact, some culturally diverse cities or farming communities may have more in common across borders than within. Folk illness practices are more likely to be preserved in stable rural communities than in bustling cities. The pace of industrialization in countries such as Brazil, China, India, and South Africa often requires rural to urban migration to support this growth with cheaper rural labor. Working age and predominantly male adult migration within these countries may influence family balance and cultural practices across generations.

In Japan, a new syndrome '*hikkikamori*' has been described involving teenagers becoming socially withdrawn as a way of life. It is uncertain whether this is related to social media and a response to associated pressures. This raises the question; can new cultural syndromes emerge as cultures change?

Globalization and the Cross-Fertilization of Cultures

With globalization, the ease of travel and communication has led to a more rapid flow of people, customs, and cultures. Individuals can immerse themselves in far-flung parts of the world within hours or days rather than in weeks or months. In addition, the Internet and mobile phone transmit images and information in real time globally. The advent of 24-hour news TV and real-time use of social media (e.g., Facebook and Twitter) facilitate sharing of beliefs, ideas, and values across geographical and cultural boundaries. Furthermore, large global corporations (e.g., Coca-Cola and McDonalds) transmit lifestyle and new cultural concepts as they penetrate emerging markets and influence the youth. Similarly, variations of Indian meals now have global appeal.

Examples abound on the impact of globalization on ideas of diet, democracy, fashion, sports, education, language, and interpersonal relationships to mention a few. Changes in these ingredients of culture invariably nurture 'mosaic' cultures and diversify societies. Migration trends from conflict or through economic activity often involve individuals retaining their idioms of distress and folk illness categories while acculturizing in the host country. In some countries, cultural or ethnic enclaves (e.g., China town, Asian, and African) may be found in some cities. These migrant groups may preserve health beliefs and systems from countries of origin thereby presenting challenges to health professionals unfamiliar with these.

Conclusion

Culture-bound syndromes are an important part of our understanding of mental distress and disorders around the world. Cultures impact on the pathogenesis and presentation, and therapeutic interventions cannot be underestimated. Certain psychiatric syndromes and idioms of distress are significantly influenced by cultural literacy, values, and attitudes. As cultures develop and evolve terms used to explain distress will change, absorbing new expressions and ignoring others. It is foreseeable that culture-bound syndromes will metamorphosize over time as they become less bound and more culturally influenced.

Some culture-bound or specific disorders will probably decline over time and may well become extinct whereas new ones may well emerge. The history of psychiatry and the classification of mental distress are still in evolution. As cultures change, the boundaries of normality are likely to shift and some symptoms and signs vary in salience. Attitudes to body mass, sleep, and time are shifting already. With a better appreciation of other societies, cross-fertilization of cultures, and ethnocentric shifts the impact of culture on the mind may become clearer.

See also: Anxiety and Anxiety Disorders; Cross-Cultural Psychology; Cross-Cultural Research Methods in Psychology; Cultural Psychiatry; Depression; Developmental Research across Cultures and Nations: Challenges, Biases, and Cautions; Differential Diagnosis in Psychiatry; Equivalence and Transfer Problems in Cross-Cultural Research; Indigenous Psychology; Normality and Mental Health; Nosology in Psychiatry; Panic Disorder; Race and Academic Motivation; Somatoform Disorders.

Bibliography

American Illustrated Medical Dictionary, second ed., 1901. W.B. Saunders and Co., Philadelphia and London.

American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders, fourth ed. DSM-IV-TR, Washington D.C.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. DSM-5, Washington D.C.

Ayonrinde, O., 2008. Brain fag syndrome: new wines in old bottles or old wine in new bottles. Nigerian Journal of Psychiatry 6 (2), 47–50.

Ayonrinde, O., Obuaya, C., May 2008a. Brain Fag Syndrome: The Extinction of a Culture Bound Syndrome? Abstracts of 161st Meeting of the American Psychiatric Association, Washington DC.

Ayonrinde, O., Obuaya, C., July 2008b. "Brain fag": a British culture bound syndrome. In: Abstracts of Annual General Meeting of the Royal College of Psychiatrists. London.

Ayonrinde, O., 2009. A History of "Brain Fag" in the 19th and 20th Century. thesis. Society of Apothecaries, London.

Bhatia, M.S., Malik, S.C., 1991. Dhat syndrome: a useful diagnostic entity in Indian culture. British Journal of Psychiatry 159, 691–695.

Bhugra, D., Buchanan, A., 1989. Impotence in ancient Indian texts. Sexual and Marital Therapy 4, 87–92.

Bhugra, D., Jacob, K.S., 1997. Culture bound syndromes. In: Bhugra, D., Monro, A. (Eds.), Troublesome Disguises. Blackwell, Oxford, pp. 296–334.

Bhugra, D., Sumathipala, A., Siribaddana, S., 2007. Culture-bound syndromes: a re-evaluation. In: Bhugra, D., Bhui, K. (Eds.), Textbook of Cultural Psychiatry. Cambridge University Press, Cambridge, pp. 141–156.

Chadda, R.K., Ahuja, N., 1990. Dhat syndrome: a sex neurosis in India subcontinent. British Journal of Psychiatry 156, 577–579.

Gillespie, R.D., 1926. Fatigue a clinical study. The Journal of Neurology and Psychopathology 7, 26.

Gillespie, R.D., 7, 1933. Psychotherapy in general practice. The Lancet 221 (5706), 1–8.

Harris, B., 1981. A case of brain fag in East Africa. British Journal of Psychiatry 139, 162–163.

Hofstede, G., 1980. Culture's Consequences: International Differences in Work-related Values. Sage, Beverly Hills.

Hughes, C.C., 1985. Glossary. In: Simons, R.C., Hughes, C.C. (Eds.), Culture Bound Syndromes. Reidel, Dordrecht, pp. 465–505.

Hughes, C.C., Wintrob, 1995. Culture bound syndromes and the cultural context of clinical psychiatry. In: Oldham, J.M., Riba, M. (Eds.), Review of Psychiatry. APA, Washington, DC, pp. 565–597.

Hughes, C.C., 1996. The culture bound syndromes and psychiatric diagnosis. In: Mezzich, J., Kleinman, A., Fabrega, H., et al. (Eds.), Culture and Psychiatric Diagnosis: A DSM IV Perspective. APA, Washington DC, pp. 298–308.

Jadhav, S., 2007. Dhis and Dhat: evidence of semen retention syndrome amongst white Britons. Anthropology & Medicine 14 (3), 229–239.

Janaki, R., Vikhram, R., Thiruvikraman, S., Naveen, S., Anjana, R., Srinivasan, B., Ramasubramanian, C., 2012. A study of sexual misconceptions and prevalence of Dhat syndrome in adolescents and young males presenting to psychiatry outpatient department. Indian Journal of Psychiatry 54, S88.

Jegede, R.O., 1983. Psychiatric illness in African students: "brain fag" syndrome revisited. Can J Psychiatry. 28 (3), 188–192.

Lippincott's New Medical Dictionary, 1910. J.P. Lippincott Co., Philadelphia and London.

Littlewood, R., 1996. Cultural comments on culture bound syndromes: I. In: Mezzich, J., Kleinman, A., Fabrega, H., et al. (Eds.), Culture and Psychiatric Diagnosis: A DSM IV Perspective. APA, Washington DC, pp. 309–312.

Maercker, A., 2001. Association of cross-cultural differences in psychiatric morbidity with cultural values: a secondary data analysis. German Journal of Psychiatry 4, 17–23.

Malhotra, H.K., Wig, N.N., 1975. A culture bound sex neurosis in the orient. Archives of Sexual Behaviour 4, 519–528.

Marano, L., 1983. On windigo psychosis. Current Anthropology 24 (1), 121–125.

Morakinyo, O., 1980. A psychophysiological theory of a psychiatric illness (the brain fag syndrome) associated with study among Africans. Journal of Nervous and Mental Disorders 168 (2), 84–89.

Ola, B.A., Morakinyo, O., Adewuya, A.O., 2009. Brain fag syndrome-A myth or a reality. African Journal of Psychiatry 12 (2), 135–143, 199.

Peltzer, K., Cherian, V.I., Cherian, L., 1998. Brain fag symptoms in rural South African secondary school pupils. Psychololgy Reports 83 (3 Pt 2), 1187–1196.

Prakash, O., 2007. Lessons for postgraduate trainees about Dhat syndrome. Indian Journal of Psychiatry 49 (3), 208–210.

Prince, R., 1960. The "brain fag" syndrome in Nigerian students. Journal of Mental Science 106, 559–570.

Simons, R.C., Hughes, C.C., 1985. In: Simons, R.C., Hughes, C.C. (Eds.), Culture Bound Syndromes. Reidel, Dordrecht, pp. 465–505.

Tuke, D.H., 1880. Intemperance in study. Journal of Mental Science 25, 495–504.

Udina, M., Foulon, H., Valdes, M., Bhattacharyya, S., Martin-Santos, R., 2013. Dhat syndrome: a systematic review. Psychosomatics. Journal of Consultation and Liaison Psychiatry 54 (3), 212–218.

World Health Organisation, 1992. ICD 10- International Classification of Diseases. Classifications of Mental and Behavioural Disorder, Geneva.

Wig, N.N., 1960. Problems of the mental health in India. Journal of Clinical & Social Psychiatry (India) 17, 48–53.

Yap, P.M., 1962. Words and things in comparative psychiatry with special reference to exotic psychosis. Acta Psychiatrica Scandinavia 38, 157–182.

Yap, P.M., 1969. The culture bound syndromes. In: Cahil, W., Lin, T.Y. (Eds.), Mental Health Research in Asia and the Pacific. East-West Centre Press, Honolulu, pp. 33–53.

Customer Satisfaction

Michael D Johnson, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp. 3198–3202, © 2001, Elsevier Ltd. with an updated Bibliography section supplied by the Editors.

Abstract

Simply stated, customer satisfaction is a customer's evaluation of their purchase and consumption experience with a product, service, brand, or company. Interest in satisfaction stems from its role in affecting customers' repeat purchase decisions and subsequent company profits. As the link to profitability has become clear, customer satisfaction is now a prominent metric in business accounting and reporting. This article provides a broad overview of customer satisfaction, its meanings, and the streams of research that have emerged in recent years. Research on both transaction-specific satisfaction (a customer's evaluation of their experience with and reactions to a particular product transaction, episode, or service encounter) and cumulative satisfaction (a customer's overall experience with a product or service provider to date) is described. Methodological issues surrounding satisfaction research are also discussed. These include the choice of methods for operationalizing a satisfaction model and how to determine the importance of satisfaction drivers. The article ends with a call to better understand the psychological and behavioral consequences of satisfaction going forward. Specifically, greater attention should be given to how satisfaction, in conjunction with a company's or brand's image (reputation), and strength of relationship with customers explain customer loyalty and retention.

Simply stated, customer satisfactiosn is a customer's evaluation of their purchase and consumption experience with a product, service, brand, or company. Policy researchers have long debated the value of using satisfaction as a subjective measure of just how well a market system is able to meet customer needs (Ölander, 1977). More recent interest in satisfaction stems from its role in affecting customers' repeat purchase decisions and subsequent company profits (Johnson and Gustafsson, 2000). As the link to profitability has become clear, customer satisfaction is now a prominent business performance metric. As a result, satisfaction research has become an important component of an organization's stakeholder management, social accounting, and reporting (Gnan et al., 2013). This article provides a broad overview of customer satisfaction, its meanings, the streams of research that have emerged in recent years, and key methodological issues that surround it. The article is not meant to be a review of what has become a vast literature (see Yi, 1991). The first part of the article describes two distinct yet complementary streams of customer satisfaction research. After a brief overview of methodological issues surrounding customer satisfaction modeling, future research directions are discussed.

Two Views of Satisfaction

Customer satisfaction research has developed around two different types of evaluations: transaction-specific satisfaction and cumulative satisfaction. Transaction-specific satisfaction is a customer's evaluation of their experience with and reactions to a particular product transaction, episode, or service encounter. Early transaction-specific satisfaction research explored the cognitive-psychological antecedents of satisfaction, while more recent research has focused on the effects of positive and negative emotions on satisfaction (Oliver, 1997).

Since the early 1990s, satisfaction research has grown to include an emphasis on cumulative satisfaction, defined as a customer's overall experience with a product or service provider to date. This definition is more consistent with those in economic psychology and welfare economics, where customer satisfaction is synonymous with the concept of consumption utility. It is cumulative satisfaction upon which the prominent national satisfaction, indices, such as the American Customer Satisfaction Index (ACSI; Fornell et al., 1996), are built.

These two streams of satisfaction research are complementary rather than competing. Transaction- specific research focuses on satisfaction with a product or service on a given occasion or transaction, or over a limited period of time. In cumulative satisfaction studies, the period of evaluation is left open. In the ACSI survey, for example, customers consider all of their experiences to date when evaluating their satisfaction with a product or service. The distinction is important as it affects how one views and models the evaluations. Whereas transaction-specific models provide a rich understanding of the dynamics of product and service encounters or episodes, cumulative satisfaction provides a stable basis for determining the drivers of satisfaction and its consequences (customer retention and subsequent economic performance).

Transaction-specific Customer Satisfaction

Oliver's extensive contributions to satisfaction research over the past 20-plus years illustrate how transaction-specific research has evolved over time (see Oliver, 1997). Oliver posited that customers form preconsumption expectations, observe product performance, compare this performance with expectations, and then combine this information with expectation levels to judge satisfaction. When performance exceeds expectations (positive disconfirmation), satisfaction increases. When performance

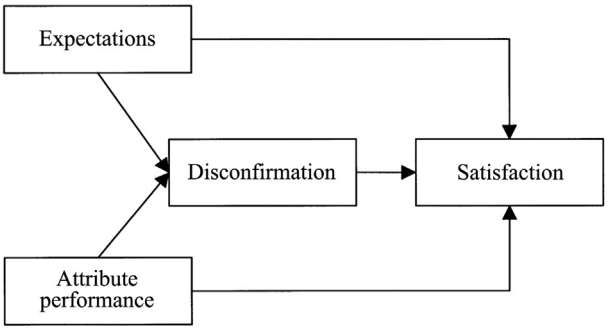

Figure 1 The disconfirmation of expectations model.

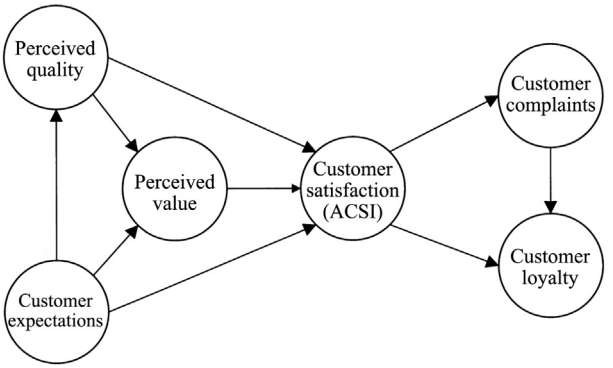

Figure 2 The American Customer Satisfaction Index (ACSI) model.

falls below expectations (negative disconfirmation), satisfaction decreases. The disconfirmation of expectations model has dominated the literature for years; a simple version of it is presented in **Figure 1**. The direct effect of expectations on satisfaction is the result of a framing or assimilation effect; the expectations establish a baseline or anchor that directly influences the resulting satisfaction judgments. Thus attribute performance, expectations of performance, and disconfirmation of expectations may all have direct effects on evaluations of satisfaction.

The disconfirmation model has been augmented over the years to include the effects of equity, attribution, and emotions on satisfaction evaluations. The concept of reciprocity, and more specifically customer perceptions of equity and fairness, lies at the heart of marketing as an exchange process. Thus customers should be more satisfied when they are treated more fairly or equitably in a transaction, especially when there is a prospect for future transactions with the same party. Oliver and Swan (1989) posited and showed that equity is an important driver of satisfaction. Attribution theory has also been used to explain transaction-specific satisfaction from a cognitive perspective. The attribution of favorable outcomes to oneself and unfavorable outcomes to others systematically affects perceptions of performance and satisfaction (Folkes, 1990).

However, Westbrook, Oliver, and others have recognized that these more cognitive perspectives on transaction-specific satisfaction do not completely capture the more affective basis of satisfaction (Oliver, 1997; Westbrook and Oliver, 1991). Customers form both positive and negative affective states that systematically influence satisfaction. For example, on the positive side there is joy and interest, and on the negative side is anger, disgust, and contempt, which affect satisfaction as expected.

Cumulative Customer Satisfaction

Cumulative satisfaction gained momentum with the advent of national customer satisfaction indices.

Johnson and Fornell (1991) speak specifically to the problem of constructing broad-based satisfaction measures. They argue that cumulative satisfaction, as an overall evaluation of the consumption experience, should be viewed as a theoretical or latent variable (similar to an attitude). As a latent variable, satisfaction can be empirically measured and meaningfully compared as a weighted-average or index of satisfaction indicators.

The ACSI model exemplifies the measurement and modeling of cumulative customer satisfaction. The model is presented in **Figure 2** (see Fornell et al., 1996). There are three main drivers of satisfaction in the model: quality, value, and customer expectations. Satisfaction, in turn, should reduce the incidence of complaints and increase customer loyalty. Most of the survey questions are rated on 1 to 10-point scales and the measurement variables are specified as reflective indicators of the latent or abstract constructs (circles) in the model.

Of particular interest is the overall satisfaction index, which is a weighted average of three survey ratings: (1) an overall rating of satisfaction, (2) the degree to which performance falls short of or exceeds expectations, and (3) a rating of overall performance relative to the customer's ideal good or service in the category. The three measures provide a highly reliable and stable index of satisfaction. Notice that factors that may be conceptually and empirically distinct drivers of satisfaction in a transaction-specific satisfaction model become multiple measures of a more overall evaluation in a cumulative satisfaction model. Consider the differential role that disconfirmation plays in transaction-specific vs. cumulative satisfaction research. From transaction to transaction or episode to episode, one would expect some variation between performance and expectations. Service providers have a particularly difficult job of conforming to specifications on any given occasion. When these transactions or episodes are accumulated and evaluated over time, they become easier to predict. Thus significant disconfirmation, when it occurs, is more likely when studying a particular transaction or episode than when studying consumption experience to date. Because cumulative satisfaction is a more overall or molar-level evaluation, disconfirmation of expectations is one of several valuable proxies by which customers evaluate their overall experience (Johnson and Selnes, 2004).

Cumulative satisfaction is systematically related to various financial and business performance measures. Research using the national indices demonstrates the positive impact that cumulative satisfaction has on both market value and accounting returns (Fornell et al., 1996). It is no wonder that customer satisfaction has become an important part of companies' overall accounting, stakeholder management, and business performance measures. Cumulative satisfaction

research has also explored the nature of customer expectations and how satisfaction varies across industries. For example, whereas price expectations tend to be more rational, performance expectations are more adaptive; they do not adjust as quickly to changes in performance. Fornell et al. (1996) report that satisfaction is highest among goods industries, lower among services, and lower still among public and government agencies.

Methodological Issues

In customer satisfaction modeling, satisfaction plays a key role in linking customer perceptions of a product's price and quality with customers' behavioral intentions and subsequent behaviors. Companies use two essential outputs of a satisfaction model to set priorities for improvement. First, how well do products and services perform on their attributes, the benefits they provide, and the customer satisfaction and loyalty that results? Second, what impacts do the attributes and benefits have on satisfaction and subsequent loyalty? That is, how important are the attributes and benefits to customers? This performance and importance information is combined to set priorities for improvement. From a cost-benefit standpoint, the essential benefits to improve are those where importance or impact is relatively high and performance is relatively low.

As latent variables, performance levels on abstract benefits (such as the 'cleanliness' or 'convenience' of a retail store) and overall satisfaction can be constructed using weighted averages or indices of more concrete attributes or proxies. There is a debate, however, as to how to determine attribute and benefit importance measures. Many satisfaction measurement systems rely heavily on customers' direct ratings of importance. The primary advantage of direct importance ratings is that they are easy to collect and analyze. Yet direct measures assume that customers understand what is meant by 'important,' just what attributes are important to them and that they are willing to tell you. As a result, direct importance measures often result in socially acceptable or status quo answers and poor discrimination. More objective estimates of importance or impact can be obtained using statistical estimation. This statistical estimation must accommodate the fact that a satisfaction model is a network of cause and effect relationships (which predicts a pattern of relationships) that contains latent (abstract) variables. Partial least squares (PLS) and principal- components regression (PCR) are particularly well suited to estimating a satisfaction model and deriving importance measures (see Johnson and Gustafsson, 2000).

Future Research

Transaction-specific satisfaction research recognizes that customers may have very complex psychological reactions to a product's performance on a given occasion or over a given time period. Cumulative satisfaction recognizes that, when forming intentions or making decisions regarding their behavior, customers rely on their entire experience to date with a product or service provider. Yet there has been little in the way of an integration of these two perspectives. Going forward, it will be important to better understand just how satisfaction-related constructs and their relationships change when customers move from transaction to cumulative-based evaluations. There is also a need to better understand the psychological and behavioral consequences of satisfaction, specifically customer loyalty (an intention) and retention (a behavior) (Gustafsson et al., 2005).

Company or brand image (reputation) and relationship commitment (involvement or strength), when combined with overall satisfaction, should explain much greater variation in loyalty and retention than does satisfaction alone. Finally, research in the quality improvement and product development areas must strive to 'bridge the gap' between internal (organizational) quality measures and processes and external (customer) perceptions of quality and satisfaction. It is not sufficient to simply understand what attributes and benefits drive satisfaction. Companies must understand how to break customer satisfaction down into its means of accomplishment within an organization.

See also: Advertising: Effects; Consumption, Sociology of.

Bibliography

Folkes, V.S., 1990. Conflict in the marketplace: Explaining why products fail. In: Graham, S., Folkes, V.S. (Eds.), Attribution Theory: Applications to Achievement, Mental Health, and Interpersonal Conflict. Erlbaum, Hillsdale, NJ, pp. 143–160.

Fornell, C., Johnson, M.D., Anderson, E.W., Cha, J., Bryant, B.E., 1996. The American Customer Satisfaction Index: Nature, purpose and findings. Journal of Marketing 60, 7–18.

Gnan, L., Hinna, A., Monteduro, F., Scarozza, D., 2013. Corporate governance and management practices: Stakeholder involvement, quality and sustainability tools adoption. Journal of Management & Governance 17 (4), 907–937.

Gustafsson, A., Johnson, M.D., Roos, I., 2005. The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention. Journal of Marketing 69 (4), 210–218.

Johnson, M.D., Fornell, C., 1991. A framework for comparing customer satisfaction across individuals and product categories. Journal of Economic Psychology 12, 267–286.

Johnson, M.D., Gustafsson, A., 2000. Improving Customer Satisfaction, Loyalty and Profit: An Integrated Measurement and Management System. Jossey-Bass, San Francisco.

Johnson, M.D., Selnes, F., 2004. Customer portfolio management: Toward a dynamic theory of exchange relationships. Journal of Marketing 68 (2), 1–17.

Ölander, F., 1977. Can consumer dissatisfaction and complaints guide public consumer policy? Journal of Consumer Policy 1, 124–137.

Oliver, R.L., 1997. Satisfaction: A Behavioral Perspective on the Consumer. McGraw-Hill, New York.

Oliver, R.L., Swan, J.E., 1989. Equity and disconfirmation perceptions as influences on merchant and product satisfaction. Journal of Consumer Research 16, 372–383.

Westbrook, R.A., Oliver, R.L., 1991. The dimensionality of consumption emotion patterns and consumer satisfaction. Journal of Consumer Research 18, 84–91.

Yi, Y., 1991. A critical review of customer satisfaction. In: Zeithaml, V. (Ed.), Review of Marketing 1990. American Marketing Association, Chicago, pp. 68–123.

Cyber Crime: Identity Theft

Katie A Farina, Cabrini College, Radnor, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Identity theft is considered to be 'one of the fastest growing crimes in America.' Due to the technological advancements and the almost ubiquitous use of computers and the Internet, identity theft is a crime that can occur practically anywhere. This article explores the various types of identity theft including medical identity theft, financial identity theft, and child identity theft. Additionally, this article highlights the prevalence of identity theft and its costs to victims both monetarily and in relation to physical and emotional stress. Finally, this article examines the recent strides of legislation to identify and combat identity theft.

Introduction

Many argue that identity thefts and cybercrimes involving fraud are considered the "fastest growing crime in America" (Cole and Pontell, 2006, p. 125). However, it is difficult for researchers and scholars to truly understand the prevalence, cost, and type of identity theft and fraud that occur, because there is not a universally agreed-upon definition (Allison et al., 2005). This varied and vague definition leads to a lack of reliable data that makes it difficult to capture the true numbers regarding the commission of this crime. This article explores several aspects of identity theft, including how practitioners and scholars define identity theft, as well as the various offenses that are considered identity theft. Additionally, current estimates of both the prevalence and costs of these crimes are considered as well as the target demographics of both victims and offenders. Moreover, a brief discussion of one of the most well-known international identity theft scams is included. Finally, this article explores the evolution of legislation meant to safeguard victims from identity thefts and assist those who are victimized.

What Is Identity Theft?

While there is no singular definition, a common theme of identity theft offenses revolves around "the misuse of another individual's personal information to commit fraud" (The President's Identity Task Force, 2008, p. 2). Crimes that have fallen under the umbrella of identity theft include, but are not limited to, checking and credit card fraud, counterfeiting, forgery, mortgage fraud, human trafficking, terrorism, passport fraud, and social security fraud. According to the United States General Accounting Office (2002), personal identifying information that can be used to commit fraud includes an individual's name, address, social security number, date of birth, alien registration number, taxpayer identification number, passport number, driver's license number, mother's maiden name, or even biometric information such as one's fingerprint, voice print, or retina image. A recent report issued by the Federal Trade Commission (FTC), which has been documenting consumer complaints and experiences for years, highlighted identity theft as the most common reported fraud among consumers. However, nearly two thirds of those who reported the victimization to the FTC did not report to local law enforcement (Federal Trade Commission, 2009).

Types of Identity Theft

According to Allison et al. (2005), the methods employed by identity thieves can best be defined based upon the degree of technology utilized. The authors first distinguish those offenders who utilize 'low technology.' Low-technology methods tend to be the most common mostly due to their relative ease. Examples include theft of wallets or purses and going through someone's trash looking for personal identifying information. A second method used by offenders is termed 'high technology.' Criminals who use high-technology methods must possess a certain degree of expertise and skill set. Their methods include the use of the Internet, skimming, and pretext calling. Skimming is the act of using computers to read and store information that is encoded onto the magnetic strip of a credit card or an automated teller machine (ATM) card. Once the offender has stolen this information, they can then place it onto a blank card thus allowing them to access the victim's bank account. Pretext calling is when offenders make contact with the victim under false pretenses with the purpose of obtaining personal information directly from the victim (Newmann, 1999). A variation of pretext calling occurs when offenders email victims pretending to be their Internet service provider and request an update to their account information and credit card accounts (Allison et al., 2005). Offenders can also obtain personal identifying information from businesses or institutions that maintain customer, employee, patient, or even student records. Identity thieves may work in jobs that allow them access to others' credit records. Offenders may pretend to offer a victim a loan, a job, or an apartment and require that the individual reveal personal information in order to 'qualify.' There are a variety of crimes that identity thieves carry out once they have a victim's information. These include taking out loans, opening credit cards, taking out cash advances and credit applications, establishing a cell phone service in the victim's name, opening a new bank account, and writing bad checks (Allison et al., 2005; Newmann, 1999).

In addition to financial identity theft, there is also medical identity theft. Medical identity theft occurs when an individual uses a victim's information to obtain medical treatment without his/her knowledge or consent. The Identity Theft Resource Center (ITRC) cautions that this type of identity theft is particularly easy as it may only require the use of an individual's name, date of birth, and address. It is also dangerous because hospitals are required to take in any emergency patients and the patient's information may not be verified as accurate before being sent to the billing department. An additional aspect of medical identity theft is related to an individual's medical history. If an offender is using a victim's information for a medical situation, the treatment, prescription drugs, or surgery the offender received will appear on the victim's medical records. As a result, the victim whose medical history has been compromised may now find it difficult to receive prescription medications or receive proper diagnoses and treatment if they require emergency attention. Once a victim of medical identity theft is contacted about these fraudulent charges, it can take months to facilitate changes or updates with insurance providers, Medicare, and/or the hospital's billing department to work out the incorrect billing (Rocha, 2013).

A subset of financial and medical identity theft is child identity theft. Child identity theft is any of the above types of theft or fraud that is perpetrated against a child. This type of theft is particularly troubling, because it often goes undiscovered for many years because children rarely have independent finances (Beltran, 2013). Children are a suitable target for identity thieves for a number of reasons. First, parents rarely check a child's credit history. Because children are unable to enter into a legal contract and should not have established credit histories, parents have little reason to suspect fraud. Second, the use of a child's identity provides offenders with a 'blank slate.' As such, offenders are presented with a wide range of fraudulent activity that they will be able to commit without raising any red flags (Beltran, 2013).

Prevalence of Identity Theft

The use of varied definitions of identity theft by official agencies has made it difficult to calculate the extent and patterns of this crime. However, both the FTC and the Bureau of Justice Statistics (sponsoring the National Crime Victimization Survey, NCVS) have each collected data in an attempt to elucidate the prevalence, costs, and forms of identity theft. Both the FTC (2007) and the NCVS (Bureau of Justice Statistics, 2011) have observed a rise in the victimization of identity theft since the new millennium. Recently, the NCVS announced that 7.0% of all households in the United States, or approximately 8.6 million households, had at least one member aged 12 or older who experienced at least one form of identity theft in 2010. This represented an increase from 5.5% of households, or 6.4 million households, that were victims of identity theft in 2005. According to the NCVS, this increase was due to a rise in the misuse of existing credit card accounts (Bureau of Justice Statistics, 2011). Similarly, the FTC (2009) claimed that identity theft was the most prevalent form of fraud committed in the United States. The most commonly reported identity thefts included credit card fraud, government documents and benefits fraud, employment fraud, phone/utilities fraud, and bank fraud. In fact, the FTC (2009) highlighted that in the past decade, credit card complaints and automobile-related complaints have nearly doubled.

The ITRC tracks data breaches by monitoring and confirming media reports of identity theft. A data breach is defined as "an event in which an individual's name plus Social Security Number (SSN), driver's license number, medical record, or a financial record/credit/debit card is potentially put at risk – either in electronic or paper format" (Identity Theft Resource Center, 2013, p. 95). In 2012, the ITRC reported 470 data breaches in the United States resulting in 17 491 690 records being exposed. It is important to note that this number is low, because not all incidents of identity theft report the amount of records exposed (Identity Theft Resource Center, 2013). Furthermore, it is pertinent to note that a breach does not necessarily indicate that an individual is a victim of identity theft. A breach signifies that an individual's information (name, address, social security number, etc.) has been compromised, exposed, and places the individual at greater risk for identity theft than the average consumer.

AllClear ID Alert Network, a privately funded organization dedicated to providing identity theft protection, released a report detailing the trends of child identity theft. In 2012, they uncovered that 10.7% of children were victims of identity theft, this number corresponding with the FTC's finding that 8% of identity theft complaints involved someone 19 years old and younger (May, 2012). The report found that children are 35 times more likely to become a victim of identity theft than an adult and that the rate of identity theft among victims aged 5 and younger has increased tremendously in the past decade (May, 2012).

Costs of Identity Theft

The inconsistent use of a definition for identity theft has made it difficult for scholars and government agencies to gauge a true count of the prevalence of identity theft. Additionally, these varying definitions hinder the ability to garner the emotional and monetary costs of identity theft to victims. It is difficult to garner a true assessment of costs incurred by victims for a variety of reasons. Usually the individual whose information has been misused is not legally responsible for any costs associated with the theft. Additionally, victims may incur expenses when they are dealing with the ramifications of having their identity stolen. For instance, identity theft victims may need to close existing accounts, open new ones, monitor credit reports, and pay higher interest rates on loans and credit accounts as a result of the theft. The individual may not have suffered a monetary cost as a direct result of the identity theft, but as a means to deal with the victimization. In fact, it took victims an average of 175 h over a span of 2 years to regain sound financial standing following a theft (Copes et al., 2010). According to Pontell et al. (2008), victims who miss work to deal with identity theft lose an average wage of $4000. Furthermore, this process of dealing with the ramifications of identity theft can take a multitude of hours and even years to overcome and consumers often experience

consequences in the interim. For instance, the consumer's credit history may be scarred leading to a denial of mortgages, loans, and bank accounts.

The NCVS and the FTC have attempted to identify both the frequency and the monetary effects of identity theft on victims, sometimes with conflicting accounts. According to the FTC (2007), over 8 million people were affected by identity theft and lost nearly $16 billion dollars. The NCVS (Bureau of Justice Statistics, 2011) surveyed victims directly and uncovered that, on average, victimized households lost $1830, while the FTC (2007) highlighted that businesses lost an average of $4800. It is important to note that this monetary loss may differ based upon the type of victimization. For instance, consumers who experience the misuse of personal information reported an average loss of $5650, while those experiencing fraud of existing accounts lost on average $1140 and finally victims of existing credit card fraud lost about $1300 (Bureau of Justice Statistics, 2011). In addition to the loss of money and time, victims may also experience physical and emotional stress such as anger, helplessness and mistrust, disturbed sleeping patterns, and a feeling of lack of security (Davis and Stevenson, 2004). The ITRC conducted a study of victims who reported mental health consequences of identity theft. In the short term, victims report feeling betrayed, defiled, and powerless; 30% of the sample experienced long-term consequences of feeling distrustful of others and 4% experienced suicidal ideations.

Identity Theft Victims and Offenders

Demographic characteristics are somewhat helpful in determining who is at risk for identity theft. According to the NCVS (Bureau of Justice Statistics, 2011), households headed by an individual 65 years or older reported the lowest rate of identity theft. Additionally, those residing in married households and households with high income levels ($75 000 or above) reported higher rates of victimization. Other studies found that the risk of victimization is highest for consumers between the ages of 25 and 54 (Anderson, 2006). However, age has become a point of contention among researchers. Recently, the NCVS has uncovered an increase in identity theft victimization among those aged 18–24. Households headed by whites, Asians, and those reporting 'two or more races' also reported the highest rates of identity theft victimization. Most victims do not know their offenders and the majority of victims believe that their information was stolen during a purchase or other transaction (Bureau of Justice Statistics, 2011). In addition to these demographic variables, analysis of the FTC's data found that households headed by women with three or more children were more likely to be victims of identity theft (Anderson, 2006). There may also be regional variation among identity theft victims. In 2011, the FTC (2012) reported that Colorado reported the highest per capita rate of fraud followed by Delaware and Maryland, while Florida reported the highest per capita rate of identity theft complaints, followed by Georgia and California.

The lack of reliable data also impacts what we know about identity theft offenders. Demographics about offenders are typically revealed through police reports and closed case files. These data revealed that nearly half of all offenders are between the ages of 24 and 34 and one-third are between 35 and 49 years old. In terms of race and ethnicity, over half were black while whites and Hispanics account for nearly 40% of offenders (Copes and Vieraitis, 2012). While research has made strides in understanding the risks of victimization, scholars have only recently attempted to understand the motivations of identity theft offenders. In one of the first studies that interviewed these offenders in federal prisons, Copes and Vieraitis (2009) examined how an offender's experiences and life circumstances affected their decision to commit identity theft. The primary motivation for committing this type of fraud was a desire for money. Offenders often viewed identity theft as an easy crime that was extremely rewarding. Often, these individuals perceived their risk of capture to be extremely low, because of their supposed advanced skills and the perceived incompetence of law enforcement. Finally, it is of interest to note that many of these offenders justified their crimes by claiming that no real harm was caused to any of their victims.

Identity Theft Legislation

Identity theft officially became a criminal act in the United States when Arizona passed the first state statute in 1996, followed by California in 1997 (Pontell, 2009). Following this legislation, the federal government passed the Identity Theft Assumption and Deterrence Act in 1998. This was the first piece of federal legislation that made identity theft a distinct crime against the individual whose identifying information was stolen. Additionally, the act declared that identity theft was punishable for up to 15 years of imprisonment and a maximum fine of $250 000. Furthermore, the US Sentencing Commission has outlined a sentence of 10–16 months of imprisonment that can be imposed regardless of an offender's previous criminal record or lack of monetary loss suffered by the victim (Copes and Vieraitis, 2009).

As the rate of identity theft continues to rise, crime control agencies have established laws that aim to increase penalties against offenders or to help victims of identity theft. In 2003, the United States Congress passed the Fair and Accurate Credit Transactions Act (FACTA). Specifically, this act was aimed to protect consumers from identity thieves and to help those who report victimization. As a result of this act, FACTA grants all consumers the right to one free credit report each year. Additionally, the act requires merchants to leave all but the last five digits of a customer's credit card number off store receipts and lenders and credit agencies to take action against identity thieves even if a consumer has not realized they have been victimized.

FACTA also mandated the creation of several national systems. The first is a system of fraud detection to increase the likelihood of discovering identity thieves. Furthermore, a nationwide system has been created that allows fraud alerts to be placed on credit files so that consumers may be notified of any unusual activity on their account. As part of this National Fraud Alert System, consumers are now allowed to place three types of fraud alerts on their credit files. The first, an initial alert, is for individuals who suspect they are, or about to become, victims of identity theft. An extended alert may be

placed on an account for an individual who has been a victim of identity theft and has filed a report with law enforcement. This extended alert lasts for 7 years wherein credit agencies may not use the consumer's name for prescreened credit or insurance offers. Finally, military personnel can now place an 'active duty alert' on their credit files when they are either serving active duty or are assigned to service away from their typical station (Fair & Accurate Credit Transactions Act of 2003, 2003). These initiatives serve as a place to assist an individual who has become a victim of identity theft.

As scholars and researchers learn more about identity theft and the various forms it may take, legislators must adjust to this changing arena. In 2004, the Identity Theft Penalty Enhancement Act enacted stiffer penalties for those convicted of an array of identity theft-related crimes as well as codified a new crime known as aggravated identity theft. Aggravated identity theft occurs when an individual "knowingly transfers, possesses, or uses, without lawful authority, a means of identification of another person" during and/or in relation to a list of over 100 felony offenses (Identity Theft Penalty Enhancement Act, 2004, p. 1). These include, but are not limited to, embezzlement, mail, bank and wire fraud, immigration fraud, passport and visa violations, social security fraud, and specified terrorism violations. Convictions of aggravated identity theft mandate enhanced penalties. According to the act, if an individual participates in identity theft during the commission of one of these listed felonies, a minimum of 2-years in prison must be served consecutively with the sentence for the underlying felony. If the identity theft occurs during and in relation to one of the more than 40 federal terrorism-related felonies, then the penalty is a mandatory minimum of an additional 5 years in prison to be served consecutively to the underlying felony (Identity Theft Penalty Enhancement Act, 2004).

In May of 2006, President George Walker Bush established a Presidential Task Force in hopes of coordinating government agencies to more effectively fight identity theft. The hopes of the task force were to identify components to make the government more adept at the awareness, prevention, detection, and prosecution of identity theft. After a 2-year investigation, the task force released a report with 31 recommendations to better assist victims and combat the crime of identity theft. They highlighted four areas where improvements should be made: data protection, avoiding data misuse, victim assistance, and deterrence. In relation to data protection, the task force called for a new culture of security among the private and public sectors. The task force called on the government to do a better job of protecting personal information, especially in relation to social security numbers, while calling on the private sector to do the same. In fact, in 2008, the United States Postal Service delivered mailings to over 146 million residences and businesses with advice on how to protect oneself from identity theft. Second, the task force also worked to make the misuse of consumer data more difficult for potential criminals. In terms of assisting victims, the task force provided identity theft training to over 900 law enforcement officers from over 250 agencies as well as victim assistance counselors to be better equipped to deal with the unique issues related to a victim of identity theft. Additionally, the task force provided grants to organizations to directly help victims while also working closely with the American Bar Association to provide pro bono legal assistance to victims of identity theft. Finally, in relation to detection and punishment, the task force proposed tougher legislation and worked with foreign law enforcement agencies to better detect, prosecute, and punish offenders (The President's Identity Task Force, 2008). Above all, the task force highlighted the ever-evolving nature of identity theft and a need for legislation and protection to evolve along with new techniques of identity theft.

Such policies that are aimed at detecting and deterring identity theft are often varied and complicated due to a lack of resources for investigation and prosecution. Additionally, identity theft is a unique crime in that cross-jurisdictional problems can often arise. For instance, it is often possible for victims and offenders to live in different states due to the use of technology. Someone's identity can be stolen remotely with the use of a computer. The nature of these crimes can lead to issues where law enforcement agencies eschew their responsibility onto another agency (White and Fisher, 2008). Partly in response to this dilemma, a second Presidential Task Force was convened in November 2009 by President Barack Obama. The Financial Fraud Enforcement Task Force (FFETF) is the broadest coalition of law enforcement agencies and regulatory services ever assembled to combat fraud. The FFETF is tasked with improving efforts between the government, state, and local crime control partners to investigate, prosecute, and punish those responsible for financial crime related to fraud. In order to assist victims of identity theft, credit card fraud, and mortgage fraud, the FFETF launched a government-run website (www.stopfraud.gov) to teach the American public about financial fraud and help victims regain any losses. In fact, the task force's Distressed Homeowner Initiative assisted over 73 000 homeowners who were the victims of mortgage and loan frauds and charged over 530 criminal defendants (Financial Fraud Enforcement Task Force, 2010).

When the 2012 legislative sessions came to a close, every state had a law that specifically outlawed identity theft or impersonation. Additionally, 29 states along with Guam, Puerto Rico, and the District of Columbia had specific restitution provisions in place for victims of identity theft, while five had forfeiture provisions set up for crimes of identity theft (Morton, 2012). While legislation to protect and assist victims of identity theft continues to grow, recent research by Piquero et al. (2011) surveyed nearly 30 000 respondents, half of which reported they would be willing to pay an additional tax if it would go toward identity theft prevention. This research finding highlights the public's growing concern with identity theft and their willingness to pay for protection.

Identity Theft: An International Example

With the increasing use of technology in identity theft offenses, this phenomenon is not confined to merely one area. The use of the Internet allows for an individual's identity and private data information to be shared instantly and over any geographical area. In fact, identity theft offenders may take advantage of different levels of law enforcement that exist between local, state, federal, and national agencies (Levi, 2008). One of the most well-known international identity theft syndicates is 'The Nigerian Letter' or '419,' so named after the criminal code against it in Nigeria (Atta-Asamoah, 2009; Levi, 2008).

While these crimes can take various forms, 419s are essentially money scams that affect individuals all over the world. Before the increased use of the Internet, those who perpetrated this offense would send letters to various residential addresses to build a relationship with a potential target and then swindle money. However, the use of technology has led to enhanced 419s wherein the offender uses a variety of tools to discover information about the target online, such as hacking into sites, scanning web pages for email addresses, and targeting victims in online chat rooms (Atta-Asamoah, 2009). Once this information has been procured, the scammer then uses the identity of the victim to scam and exploit him/her. The continued use of this tactic to commit identity fraud, money laundering, and theft has permeated West African nations so much so that it has moved from incidents committed by individuals into loosely organized networks.

Conclusion

Identity theft may not be a new phenomenon, but the prominence of the Internet and our reliance upon technology has not only allowed for an increase in such offenses, but has also changed how this crime is perpetrated. Child identity theft, medical identity theft, social security fraud, credit card fraud, and data breaches affect over 8.6 million American households yearly. The costs of these offenses are staggering. Monetary losses number in the billions and victims report feelings of betrayal, distrust, and powerlessness. The rapid development of technology and increasing use of the Internet has allowed offenders to target victims from anywhere on the globe. Scholars and practitioners alike are grappling with understanding this ever-changing phenomenon. By lacking a common definition of these crimes, it is not only difficult to measure the prevalence and costs of these offenses, but lawmakers are faced with the arduous task of creating effective legislation. Both legislators and academics have turned their interest on these crimes and are seeking to understand not only the crime itself, but how to deter such victimizations.

See also: Crime: Knowledge about and Prevalence; Crime: Whitecollar; Deterrence: Sanction Perceptions; Prevention of Crime and Delinquency; Property Crime: Victimization and Trends; Victimless Crime.

Bibliography

Allison, S.F.H., Schuck, A.M., Lersch, K.M., 2005. Exploring the crime of identity theft: prevalence, clearance rates and victim/offender characteristics. Journal of Criminal Justice 33, 19–29.

Anderson, K.B., 2006. Who are the victims of identity theft? The effect of demographics. Journal of Public Policy and Marketing 25, 160–171.

Atta-Asamoah, A., 2009. Understanding the West African cyber crime process. African Security Review 18, 105–114.

Beltran, G., 2013. Trends in Child Identity Theft. Identity Theft Resource Center. http://www.idtheftcenter.org/Child-ID-Theft/trends-in-child-identity-theft.html (accessed 15.08.13.).

Bureau of Justice Statistics, 2011. Identity Theft Reported by Households, 2005–2010. U.S. Government Printing Office, Washington, D.C.

Cole, S.A., Pontell, H., 2006. Don't be low hanging fruit: Identity theft as moral panic. In: Monahan, T. (Ed.), Surveillance and Security. Routledge, London, pp. 125–147.

Copes, H., Kerley, K.R., Huff, R., Kane, J., 2010. Differentiating identity theft: an exploratory study of victims using a national victimization survey. Journal of Criminal Justice 28, 1045–1052.

Copes, H., Vieraitis, L.M., 2009. Bounded rationality of identity thieves: using offender-based research to inform policy. Criminology & Public Policy 8, 237–262.

Copes, H., Vieraitis, L.M., 2012. Identity Thieves: Motives and Methods. Northeastern University Press, Boston.

Davis, K., Stevenson, A., 2004. They've got your numbers. Kiplinger's Personal Finance 58, 72–77.

Fair & Accurate Credit Transactions Act of 2003, 2003. Bills and Statutes (PL 108-159, December 4, 2003). http://www.gpo.gov/fdsys/pkg/PLAW-108publ159/pdf/PLAW-108publ159.pdf (accessed 08.08.13.).

Financial Fraud Enforcement Task Force, 2010. First Year Report: Financial Fraud Enforcement Task Force. http://www.stopfraud.gov/docs/FFETF-Report-LR.pdf (accessed 08.08.13.).

Federal Trade Commission, 2007. 2006 Identity Theft Report. http://www.ftc.gov/os/2007/11/SynovateFinalReportIDTheft2006.pdf (accessed 08.08.13.).

Federal Trade Commission, 2009. Consumer Sentinel Network Data Book for January–December 2008. Federal Trade Commission, Washington, DC.

Federal Trade Commission, 2012. Consumer Sentinel Network Data Book for January–December 2011. Federal Trade Commission, Washington, DC.

Identity Theft Penalty Enhancement Act, 2004. (PL 108-275, July 15, 2004). Bills and Statutes. http://www.gpo.gov/fdsys/pkg/PLAW-108publ275/pdf/PLAW-108publ275.pdf (accessed 08.08.13.).

Identity Theft Resource Center, 2013. 2012 Breach List. http://www.idtheftcenter.org/images/breach/Breach_Report_2012.pdf (accessed 15.08.13.).

Levi, M., 2008. Organized fraud and organizing funds: unpacking research on networks and organization. Criminology and Criminal Justice 8, 289–419.

May, J., 2012. Child Identity Theft Report 2012. Identity Thieves Target Young Children: What Parents Need to Know to Protect Their Kids. AllClear ID Alert Network. http://www.idtheftcenter.org/Child-ID-Theft/trends-in-child-identity-theft.html (accessed 15.08.13.).

Morton, H., 2012. Identity theft state statutes. National Conference of States Legislatures. http://www.ncsl.org/issues-research/banking/identity-theft-state-statutes.aspx (accessed 08.08.13.).

Newmann, Q.J., 1999. Identity Theft: The Cybercrime of the Millennium. Loompanics Unlimited, Port Townsend, WA.

Piquero, N.L., Cohen, M.A., Piquero, A.R., 2011. How much is the public willing to pay to be protected from identity theft? Justice Quarterly 28, 437–459.

Pontell, H.N., Brown, G.C., Tosouni, A., 2008. Stolen identities: a victim survey. In: Newman, G., McNally, M. (Eds.), Crime Prevention Studies: Identity Theft and Opportunity. Criminal Justice Press, New York, pp. 57–86.

Pontell, H.N., 2009. Identity theft: bounded rationality, research and policy. Criminology & Public Policy 8, 263–270.

The President's Identity Task Force, 2008. The President's Identity Theft Task Force Report. http://www.idtheft.gov/reports/IDTReport2008.pdf (accessed 08.08.13.).

Rocha, K., 2013. Medical Identity Theft: The Basics. The Identity Theft Resource Center. http://www.idtheftcenter.org/Medical-ID-Theft/medical-identity-theft.html (accessed 15.08.13.).

United States General Accounting Office, 2002. Identity Fraud: Greater Awareness and Use of Existing Data Needed. Publication No. USAGAO-02–766. http://www.gao.gov/assets/240/234959.pdf (accessed 08.08.13.).

White, M.D., Fisher, C., 2008. Assessing our knowledge of identity theft. Criminal Justice Review 19, 3–24.

Relevant Websites

http://www.bjs.gov/index.cfm?ty=tp&tid=42 – Bureau of Justice Statistics: Identity Theft.

http://www.consumer.ftc.gov/features/feature-0014-identity-theft – Federal Trade Commission Consumer Information.

http://www.stopfraud.gov/ – Financial Fraud Enforcement Task Force.

http://www.idtheftcenter.org/ – Identity Theft Resource Center.

http://idtheft.gov/ – The President's Task Force on Identity Theft.

https://www.privacyrights.org/Identity-Theft-Data-Breaches – Privacy Rights Clearinghouse.

Cyberbullying

Elizabeth Whittaker and Robin M Kowalski, Clemson University, Clemson, SC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since the turn of the century, researchers, legislators, the media, and public have devoted increasing attention to the topic of cyberbullying. Defined as bullying that uses electronic communication technologies (e-mail, text messages, online gaming, Web sites, chat rooms) to bully others, cyberbullying occurs throughout the world and among people with varying demographic characteristics. This article provides an overview of cyberbullying, beginning with a definition of cyberbullying and an examination of how it compares to traditional bullying. We also discuss prevalence rates of cyberbullying, characteristics of victims and perpetrators, as well as consequences of involvement in cyberbullying. We conclude with a look at prevention and intervention efforts.

Introduction

Compared to the turn of the twenty-first century when few people had heard of cyberbullying, today hardly a week goes by that the topic of cyberbullying does not appear in the media in the United States. Indeed, prior to the shootings at Columbine High School in 1999, no states had legislation even related to traditional bullying and certainly not cyberbullying. As of 2014, 49 states (excluding Montana) have some type of legislation related to traditional bullying and 35 have legislation related to cyberbullying. Since the shootings at Columbine High School, research on the topic of bullying and, subsequently cyberbullying, has surged. While this rise in research has led some to suggest that the prevalence of bullying and cyberbullying has also increased in the recent past, a plausible alternative argument is that scholars, the media, and others are paying more attention to bullying in all its forms.

One reason for the increased media attention is that both traditional bullying and cyberbullying have been shown to be related to a host of negative consequences, including suicidal ideation and suicidal behavior. One such case involved Ryan Patrick Halligan. Ryan had been traditionally bullied at school for several years. Then, after befriending a popular girl whom he liked at school, he realized that she had been sharing his personal instant messaging exchanges meant only for her with some of their classmates. The girl also verbally confronted him in person telling him that her apparent interest in him had all been a joke. Ryan committed suicide in 2003 at the age of 13 (Kowalski et al., 2012b; ryanpatrickhalligan.org).

Although the instances of cyberbullying that end in suicide are relatively infrequent, many have been picked up and sensationalized by the media. Importantly, researchers have increasingly focused on a host of other negative consequences associated with cyberbullying, as explored below. The purpose of this article is to give an overview of cyberbullying, beginning with a definition of cyberbullying (including an examination of how cyberbullying compares to traditional bullying), then exploring prevalence rates, characteristics of victims and perpetrators, the consequences of cyberbullying, and concluding with a discussion of methods of prevention and intervention related to cyberbullying.

Defining Cyberbullying

Cyberbullying, also known as electronic aggression, has generally been defined as bullying that uses electronic communication technologies (e-mail, text messages, online gaming, Web sites, chat rooms) to bully others (Kowalski et al., 2012b). However, this is a very broad definition, and defining the more specific features of cyberbullying has proven somewhat difficult.

Current Issues with Defining Cyberbullying

Researchers do not agree on how best to define cyberbullying. Several factors are involved in creating this lack of uniformity in conceptualizing cyberbullying. One of the factors complicating the development of a precise definition of cyberbullying involves the varied and changing nature of the technologies through which cyberbullying occurs. Narrowing down these electronic communication technologies used in cyberbullying situations has been problematic. Some researchers take a wide view of cyberbullying and include all electronic communication technologies in their definitions, often by simply asking participants whether or not they have experienced cyberbullying without identifying the venue by which the cyberbullying occurred. Other researchers have been more specific in naming the technologies (e-mail, social networking sites, etc.) that must be involved for the instance to be defined as cyberbullying.

This discrepancy in examining the technologies through which cyberbullying occurs may be further compounded by today's rapidly changing technology. The venues through which cyberbullying is perpetrated tend to reflect the current types of technology being used, and the venue found to be the most popular mode for cyberbullying simply reflects which type of technology is the most popular at the time. For instance, Juvonen and Gross (2008) found that message boards and instant messaging were the most prevalent venues for cyberbullying among American students ranging from 12 to 17 years of age. Statistics released by the US Department of Education's Institute of Education Sciences' National Center for Education Statistics in 2010 – just 2 years after Juvonen and Gross's (2008) study – indicated that the most

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64105-3

common venue for cyberbullying among individuals 12–18 years of age was text messaging (Robers et al., 2010). Given the recent rise of social networking sites, these sites will likely emerge as primary venues of cyberbullying sometime in the near future.

The venue by which cyberbullying occurs may also influence the form that cyberbullying takes. For example, how people communicate and interact over text messages is likely different from how they interact in the comments sections of Web sites. This, too, makes it difficult to generate a uniform definition of cyberbullying that applies across situations.

Divergent conceptualizations of cyberbullying are important because they have implications for how cyberbullying is measured, as will be discussed below under prevalence rates. Different definitions of cyberbullying lead to measuring the construct differently, which leads to different prevalence estimates, outcomes, etc. For example, researchers who ask participants whether they have ever been cyberbullied will generate prevalence rates very different from investigators who ask participants whether they have been cyberbullied via instant messaging, e-mail, or online gaming, for example (Kowalski et al., 2014). Typically, prevalence rates are lower in response to global assessments of cyberbullying as opposed to questions asking about specific venues by which the cyberbullying may have occurred.

Categories

Because of all the different forms it may take (e.g., flaming, outing, and trickery; see **Table 1**), there is clearly not one monolithic form of cyberbullying. In many instances, different forms of cyberbullying have been studied together, and the relationship between the perpetrator and the victim not taken into account (see, however, Pyżalski, 2012; see **Table 1**). Several researchers have developed cyberbullying taxonomies to more fully examine the different ways in which cyberbullying occurs, as well as the different relationships between the victim and perpetrator involved in the cyberbullying incident. Two of these taxonomies are those developed by Willard (2007) and Pyżalski (2012). As shown in **Table 1**, Willard's (2007) taxonomy focuses on characteristics of the cyberbullying behavior (see also Kowalski and Whittaker, in press). Pyżalski (2012), on the other hand, developed a classification system that focuses predominantly on the relationship between the victim and the perpetrator as opposed to the behavior itself (see **Table 1**). This taxonomy may be particularly useful in instances where the victim does not know who the perpetrator is, or when the victim and perpetrator do not know each other, which often seems to be the case with cyberbullying (Kowalski and Limber, 2007).

Willard's (2007) taxonomy can be incorporated within Pyżalski (2012). Any tactic listed in Willard's (2007) taxonomy may be used to perpetrate cyberbullying against any of the targets mentioned by Pyżalski (2012). Given the number of ways in which cyberbullying can occur and the multitude of potential victims, these taxonomies provide a useful viewpoint for researchers to begin scrutinizing the question of who does what to whom under what circumstances and with what effect.

Comparison with Traditional Bullying

Conceptualizing cyberbullying has been further compounded by its perceived relationship with traditional bullying. Cyberbullying is often discussed in conjunction with traditional bullying, and the degree to which cyberbullying is similar to and different from traditional bullying, as well as the extent to which knowledge gained from traditional bullying research can be generalized to cyberbullying, are often part of this discussion. Traditional bullying is defined as an aggressive act (1) that is intended to distress or cause harm to the target, (2) that is

Table 1 Cyberbullying taxonomies

Willard (2007)		*Pyżalski (2012)*	
Category	*Definition*	*Category*	*Definition*
Flaming	An online fight	Aggression toward the vulnerable	The victims are 'weaker' people, such as homeless people or the mentally ill, and these people are often unaware of the victimization. Often includes a visual component
Harassment	Repeated, offensive messages sent to a victim	Aggression toward random people known only online	The perpetrator does not know the victim, and vice versa
Outing and trickery	Sharing personal information, without consent of the individual	Aggression toward groups	An entire group of people is victimized, examples of this may include ethnic or religious groups
Exclusion	Blocking the person from friend lists	Aggression toward celebrities	Well-known people, such as celebrities, are victimized
Impersonation	Impersonating the victim online and sending inappropriate to others as if it were being sent from the victim	Aggression toward peers	People known to the perpetrator in real life are victimized
Cyberstalking	Using electronic communications to stalk another person by sending repeated threatening communications		
Sexting	Distributing sexual pictures of others without their consent		

usually repeated over time, and (3) that occurs among people among whom there is a power imbalance. This power imbalance can be physical, social, relational, or psychological (Dooley et al., 2009; Monks and Smith, 2006).

Cyberbullying is believed to share these same three features in common with traditional bullying (Kowalski and Limber, 2007; Kowalski et al., 2012b; Olweus, 1993). Like traditional bullying, cyberbullying is an act of aggression; it occurs among individuals between whom there is a power imbalance; and the behavior is often repeated. However, although the power imbalance with cyberbullying can be of a physical, social, or relational nature, as with traditional bullying, it may also reflect differences in talents or abilities of the victim and perpetrator, such as technological expertise (Dooley et al., 2009). The repetitive nature of cyberbullying can take a number of different forms. A single aggressive text or message sent to multiple people, or posted where many people can see, is repetitive in nature, and a single aggressive post may be viewed repeatedly by the target. Finally, a target may ruminate on cyberbullying that has already been experienced as he or she anticipates when the next derogatory text message or the Internet message will be posted.

Although traditional bullying and cyberbullying share several features in common, there are fundamental ways in which the two types of bullying differ from each other. Although people are rarely as anonymous online as they believe themselves to be, perpetrators of cyberbullying often perceive themselves as anonymous, and research on deindividuation shows that people will say or do things anonymously that they would not say or do in face-to-face interactions (Diener, 1980). This perceived anonymity opens up the pool of individuals who might perpetrate electronic aggression, as those who would never bully another person face-to-face may find it easier to do so online. The perceived anonymity of the perpetrator can also leave the victim feeling powerless, as he or she is unable to identify who is behaving aggressively online. This feeling of powerlessness on the part of the victim plays into the power imbalance between the perpetrator and the victim.

The anonymity inherent in some of these interactions plays another role as well. In instances of traditional bullying, the impact of the bullying behavior on the victim can be directly observed, and, for some perpetrators, the recognition that they have hurt the victim is enough to discourage further instances of bullying. This direct observation of the effects of bullying on the victim is not possible in instances of cyberbullying because of the online nature of the interaction that impairs chances of empathy and remorse on the part of the perpetrator (Kowalski et al., 2014).

Another difference between cyberbullying and traditional bullying involves the accessibility of the victim. Traditional bullying typically occurs at school during the school day (Nansel et al., 2001), but cyberbullying can occur 24 h a day, 7 days a week. At any time, an individual who engages in cyberbullying can create malicious posts, send messages, or create mocking Web sites. Even if the target chooses not to view the messages, that does not mean that the perpetrator is not leaving them or that other people are not viewing them. In addition, because of the nature of the venues through which cyberbullying occurs, the cyberbullying has the potential to reach a much greater audience than instances of traditional bullying do. Thousands of people can view insults posted online, whereas a few dozen at most may be witness to a traditional bullying incident at school.

This may be compounded by the perceived uncontrollability of the Internet. Online communication is rarely regulated by a moderator who will intervene if an interaction becomes aggressive, whereas bystanders in a face-to-face situation *may* be more likely to intervene in a bullying situation (Kowalski et al., 2013b). The relative permanence of online posts also conveys a measure of uncontrollability to the Internet. Messages remain on certain types of media for extended periods of time, or until they are deleted. Even then, there is the possibility that someone has downloaded the message, or reposted it elsewhere without the subject's knowledge. Once something is out there in cyberspace, there is no real way to control what happens to it.

To examine the degree to which traditional bullying and cyberbullying are related to one another, several researchers have empirically tested the relationship between the two types of bullying. Results of these investigations have been mixed. Some degree of overlap has been observed between the victims of traditional bullying and cyberbullying, as well as between the perpetrators of cyberbullying and traditional bullying (Gradinger et al., 2009; Hinduja and Patchin, 2008). However, other researchers have observed weaker relationships between cyberbullying perpetration and traditional bullying perpetration (Varjas et al., 2007). A relationship between traditional bullying victimization and cyberbullying perpetration has been identified as well (Hinduja and Patchin, 2008), suggesting that some individuals may engage in cyberbullying to retaliate for prior traditional bullying victimization. As will be discussed later in this article, this question of the degree of overlap between traditional bullying and cyberbullying is an important one. For example, to the degree that the two types of bullying co-occur together, researchers must ascertain the unique contributions of each to the negative consequences experienced as a result of involvement in both traditional bullying and cyberbullying.

Prevalence Rates

Determining the frequency with which cyberbullying occurs is difficult as prevalence rates are highly variable across studies. Reported prevalence rates differ from study to study as a function of whether or not a definition of cyberbullying is provided, the time frame used to assess prevalence rates (has cyberbullying occurred within the past 2 months, 6 months, 1 year, lifetime?), the stringency of the criteria used to assess prevalence (at least once vs twice to thrice a month or more), the venue by which the cyberbullying is said to occur, demographic characteristics of the participants sampled (gender, age, race), and so on. This variation has led to something of a debate in the literature concerning whether or not the rates of cyberbullying are remaining more or less constant (Olweus, 2013), or increasing with changes in technology (Slonje and Smith, 2008).

Prevalence rates are typically estimated to be somewhere between 10 and 40% (Lenhart, 2010; O'Brennan et al., 2009).

Nonetheless, there have been several instances where much higher prevalence rates have been found. For example, Juvonen and Gross (2008) found that 72% of those sampled reported being victimized online. However, in that study, the term cyberbullying was not used, and participants were instead asked how often they had experienced 'mean things,' defined as "anything that someone else does that upsets or offends someone" (p. 499). This particular definition may be somewhat less specific than other definitions, possibly leading to the findings of increased prevalence rates.

The variation of prevalence rates from study to study may be influenced in part by the venues used to assess cyberbullying. The measurement items used to assess cyberbullying typically reflect recent trends in technology use at the time of data collection. As previously mentioned, the venues through which cyberbullying most commonly occurs tend to reflect the prevailing forms of technology at that particular point in time. Venues that are extremely popular at one point in time may quickly fall out of favor and, as a result, cyberbullying prevalence rates may rapidly decrease on these venues, while simultaneously increasing at equally rapid rates on the new popular venue. For example, as social networking sites, such as Twitter and Facebook have become more popular in use, they will likely become more frequent venues for cyberbullying. Prevalence rates also vary across studies in relation to demographic characteristics of each sample assessed, such as age, gender, and race.

Variations by Age, Gender, Race/Ethnicity, and Rural/Urban Samples

Age

Cyberbullying is often thought to be a problem of middle school students, and there is no doubt that electronic aggression is a problem in middle school, particularly when compared to younger aged samples. For instance, the Fight Crime telephone survey conducted using a nationally representative sample throughout the United States found that younger students were less likely than middle school children to have experienced cyberbullying, as either victim or perpetrator (www.fightcrime.org). However, three points are noteworthy. First, the perception that cyberbullying is a 'middle school' phenomenon is distorted by the fact that the majority of the research to date has been conducted using middle school samples. Second, research shows that cyberbullying is, indeed, alive and well in college and other adult samples (Kowalski et al., 2012a). Third, even among middle school populations, there is significant variation in both the prevalence and venue of cyberbullying (Kowalski et al., 2012b; Tokunaga, 2010). In particular, sixth grade boys appear to lag behind in the use of technology and experiences with cyberbullying. By eighth grade, however, gender differences in cyberbullying experiences have leveled out (Kowalski and Limber, 2007). Unfortunately, age differences in prevalence rates of cyberbullying will likely decrease as children gain experience with technology at younger and younger ages.

Gender

Another demographic factor that has been examined in relation to cyberbullying prevalence rates is gender. Gender differences in experiences with cyberbullying, either as perpetrator or victim, are currently fervently debated. Males are typically thought to engage in more direct forms of aggression, whereas females may prefer to engage in more indirect forms of aggression (Kowalski et al., 2012b). Because cyberbullying has been identified as a form of indirect aggression, the finding in several studies that females are more likely than males to engage in cyberbullying behavior is relatively unremarkable (Kowalski and Limber, 2007; Robers et al., 2010; Tokunaga, 2010).

However, this finding that cyberbullying is more prevalent among girls than boys is inconsistent across studies. Additional studies found no differences between males and females in their *overall* experience with cyberbullying (Hinduja and Patchin, 2008; Slonje and Smith, 2008). Among some of these studies, however, a few found sex differences in the venues through which the cyberbullying was perpetrated. For instance, a study conducted by Hinduja and Patchin (2008) found that females were more likely than males to be targeted via e-mail, even if the overall prevalence rates of cyberbullying among males and females did not differ.

Yet other research has found that males are more likely to perpetrate cyberbullying whereas females are more likely to be the targets of cyberbullying (Sourander et al., 2010). Clearly, at this point in time, evidence regarding gender differences in experiences of cyberbullying is highly variable. This variability may be due in part to measurement differences in the individual studies exploring these differences. Therefore, caution may be in order when it comes to drawing conclusions regarding the relationship between cyberbullying and gender.

Race/Ethnicity

Relative to other demographic variables, race and ethnicity have been vastly understudied in terms of their relationship with cyberbullying. Two studies, however, that did examine racial differences in involvement with cyberbullying found little effect of race (Hinduja and Patchin, 2008; Ybarra et al., 2007). In considering how race and ethnicity relate to cyberbullying experiences, it is important to remember that race and ethnicity are not immediately observable in the virtual world. As such, race and ethnicity may not necessarily be related to cyberbullying behavior with regards to victimization (Kowalski et al., 2012b; Kowalski et al., 2014). If one cannot determine a person's race online, that person cannot be targeted for his or her race. As far as perpetration is concerned, online status may not be as heavily influenced by factors such as racial group membership to the same degree that it is in the nonvirtual world, so race may also be less likely to play a role in perpetrating cyberbullying as well (Kowalski et al., 2012b). Naturally, however, to the degree that cyberbullying develops in relationships formed in the 'real' as opposed to the 'virtual' world, racial and ethnic targeting may be already extant, making overt online identification of race and ethnicity unnecessary. Additional research is needed in this area.

Rural/Urban Samples

Attention has recently been paid to differences in cyberbullying prevalence rates in rural vs urban environments. However, the findings are contradictory. One study, for example, found prevalence rates of cyberbullying to be higher than expected in

three rural Appalachian Schools (Dulmus et al., 2004). They attributed this higher prevalence rate to rural teachers and parents talking to students involved in cyberbullying less frequently than urban teachers. Alternatively, Bauman (2010) found cyberbullying prevalence rates in schools in a rural southeastern Arizona community to be much lower than expected based on previous literature. This dichotomy between findings presented by these studies highlights the need for further research on the topic to fully parse out the relationship between cyberbullying and rural vs urban settings.

Characteristics of Victims and Perpetrators

Although demographic variables are important to consider in relation to involvement in cyberbullying, a myriad of person and situational variables are related to involvement in cyberbullying for both victims and perpetrators. Compared to individuals not involved with cyberbullying, victims score lower on measures of social intelligence (Hunt et al., 2012), they spend more time online, and they engage in riskier behaviors when they are online, such as disclosing more personal information online (Görzig and Ólafsson, 2013). The observation that victims of cyberbullying are more likely to use e-mail, instant messaging, online shopping, blogging, surfing the Internet, personal pages, and gaming than nonvictims suggests that there may be individual differences in lifestyle choices of victims compared to those not involved with cyberbullying (Kowalski et al., 2012b).

Additionally, targets of cyberbullying demonstrate higher levels of depression, anxiety, and suicidal ideation, as well as lower levels of self-esteem, than those uninvolved in cyberbullying (Kowalski et al., 2014; Kowalski and Limber, 2013). However, caution should be exercised when examining these psychological variables as possible predictors of cyberbullying, as it is also plausible that these psychological variables are consequences of previous traditional or cyberbullying victimization that set targets up to be further victimized in the future. Further longitudinal research is currently needed to explicate these effects.

Victimization and perpetration have both been associated with academic difficulties such as disliking school, poor performance in school, increased rates of absenteeism, and concentration problems (Kowalski et al., 2014; Kowalski and Limber, 2013). Perpetrators of cyberbullying also show lower levels of empathy, particularly cognitive empathy, and higher levels of narcissism than those uninvolved in cyberbullying (Kowalski et al., 2014). Perpetrators also typically report higher levels of depression and anxiety, as well as lower levels of self-esteem, similar to what victims of cyberbullying report (Kowalski et al., 2014). Much like targets, perpetrators of cyberbullying are more likely to spend a disproportionate amount of time online when compared to those not involved in cyberbullying in some way (Hinduja and Patchin, 2008). This increased use of the Internet may help explain the increased technological proficiency reported by perpetrators.

In addition to these psychological variables, perpetrators also typically exhibit a multitude of other negative behaviors compared to those uninvolved with cyberbullying. These maladaptive behaviors include higher rates of alcohol and drug use, more run-ins with law enforcement, and a higher rate of physical altercations (Kowalski et al., 2014; Ybarra and Mitchell, 2004). Some of these maladaptive behaviors may stem from the values and attitudes of cyberbullying perpetrators. Relative to those not involved with bullying, perpetrators of cyberbullying demonstrate moral approval of bullying (Williams and Guerra, 2007) and the ability to morally disengage and minimize the negative impact of their behavior on others (Bandura, 1999).

Both cyberbullying victimization and perpetration appear to be inversely related to support from friends and parents, as well as a positive school climate, as would be expected (Fanti et al., 2012). Not surprisingly, youth whose parents more closely attend to their online activities and who communicate with them about the Internet safety show lower cyberbullying victimization rates (Wade and Beran, 2011). As far as cyberbullying perpetration is concerned, youth who report having poor relationships with their parents as well as low levels of parental support tend to exhibit higher levels of cyberbullying perpetration (Wang et al., 2009).

Importantly, many of the variables that have been identified as risk factors or protective factors for cyberbullying are similar to those that are related to involvement in traditional bullying. Thus, practitioners and school personnel can use some of what we know about traditional bullying to inform efforts at preventing and cyberbullying.

Consequences of Cyberbullying

While suicide is one of the most frequently discussed potential outcomes of cyberbullying, particularly in the media, cyberbullying victimization is related to a host of other physical and psychological problems. These problems include suicidal ideation, depression, low self-esteem, anxiety, loneliness, somatic symptoms, school absences, poor school performance, and drug and alcohol use (Hinduja and Patchin, 2008; Kowalski and Limber, 2013; Ybarra and Mitchell, 2004). Victims of cyberbullying are not the only ones to potentially suffer negative effects, and being a perpetrator of cyberbullying has been associated with drug and alcohol use, loneliness, anxiety, depression, low self-esteem, and poor academic achievement, as well as higher levels of loneliness (Ybarra and Mitchell, 2004). These outcomes are similar to the outcomes reported in conjunction with traditional bullying.

To date, most of the literature examining the effects of involvement in cyberbullying has, not surprisingly, focused on victims. The extent to which the effects listed above occur depends on a number of variables including how often the cyberbullying occurs, the venue by which the online aggression is perpetrated (some venues are more public and potentially more damaging than others), and the individual's ability to appraise the situation as nonthreatening or as posing a potential for growth.

In addition, involvement in other types of bullying besides cyberbullying would contribute to determining how negative the effects of cyberbullying would be. Many youth who are involved in cyberbullying are also involved with traditional bullying, as either the perpetrator or the victim. As mentioned previously, the outcomes of cyberbullying have been found to

be similar to the outcomes reported in studies of traditional bullying, and, as a result, it is often difficult to clearly disentangle the negative effects of cyberbullying from the negative effects of traditional bullying (Olweus, 2013). However, a multitude of youth are involved in either traditional bullying or cyberbullying independently of the other form of bullying, and many experience a host of negative outcomes, including suicide in a few unfortunate instances. Hinduja and Patchin (2010) observed that suicide attempts were 1.9 times higher among victims of cyberbullying and 1.5 times higher among cyberbullying perpetrators. Additionally, Kowalski et al. (2013a) found that cyberbullying accounts for unique variance above and beyond traditional bullying in predicting negative outcomes associated with victimization.

Prevention and Intervention

While it is unreasonable to expect that incidences of traditional bullying and cyberbullying will ever be completely eradicated, there are measures that can be taken to reduce the frequency with which these behaviors occur and the harmful effects that they produce. Importantly, however, it is not solely the job of parents or school administrators to accomplish the task of reducing bullying prevalence. Rather, a community-wide effort is needed that includes students, parents, teachers, school administrators, and community members. At the school level, assessments of the prevalence of traditional bullying and cyberbullying need to be conducted. Through these assessments, schools can more readily target their prevention and intervention efforts. Bullying prevention and intervention programs need to be incorporated into the school curriculum so that they are discussed on a regular basis, and the consequences for engaging in the behaviors are both clear to the students and enforced.

Parents, first, need to become aware of the warning signs that accompany both traditional bullying and cyberbullying: depression, anxiety, withdrawal from normal social activities, and poor academic performance. With cyberbullying in particular, parents need to watch for signs that their child(ren) is/are becoming upset after being online or receiving text messages. In addition, parents need to educate themselves about their child's online activities. Whereas today's youth are digital natives, and have spent the majority of their lives thinking nothing of using technology, parents are digital immigrants, sometimes knowing almost nothing about the use of technology. Parents can learn a lot from their children about technology and what their children are doing online by asking them and communicating with their children about what is 'trending' online.

With both traditional bullying and cyberbullying, many students report that they have witnessed bullying behavior within recent months (Kowalski et al., 2013b). These students need to be encouraged to take a more active role as the bystander and intervene on behalf of the victim. Bystanders are, not surprisingly, reluctant to intervene for fear that the cyberbullying behavior will subsequently be directed toward them. However, there are other things the bystanders can do to intervene on behalf of the victim, such as telling someone in a position of authority that the cyberbullying behavior is occurring.

Finally, much more research is needed in the area of cyberbullying to inform these prevention and intervention efforts. Prevention and intervention programs are not a one-size-fits-all model. For example, the programs implemented in rural areas may differ from those that work most effectively in urban areas. Those that work with youth in high school may differ from those that are effective with youth in middle school (Limber et al., 2009a,b). Research that examines the unique cyberbullying experiences of individuals in different settings will inform these efforts.

Future Research

As noted throughout this article, much research remains to be conducted on cyberbullying. A few of these research areas will be mentioned here. As much as remains to be learned about risk and protective factors related to cyberbullying victimization, even more remains to be learned about predictors of cyberbullying perpetration. Additional research is also needed examining prevalence rates of cyberbullying across a wider age demographic than just middle school students. As children at younger and younger ages become exposed to technology, researchers need an understanding of prevalence rates of cyberbullying among preschoolers, for example. They also need to examine the forms that cyberbullying takes place among older populations, particularly within the workplace. Finally, a great deal of research is needed in examining the best way to define and measure cyberbullying. For a detailed discussion of future research suggestions, the reader is referred to Kowalski et al. (2014).

See also: Group Processes, Social Psychology of; Hazing in Fraternities and Sororities; Online Pornography; Oversharing: The Eclipse of Privacy in the Internet Age; Teens, Gender, and Self-Presentation in Social Media; Youth Culture, Anthropology of.

Bibliography

Bandura, A., 1999. Moral disengagement in the perpetration of inhumanities. Personality and Social Psychology Review 3, 193–209.

Bauman, S., 2010. Cyberbullying in a rural intermediate school: an exploratory study. The Journal of Early Adolescence 30 (6), 803–833.

Department of Justice Office of Justice Programs. Available at: http://nces.ed.gov/pubs2011/2011002.pdf

Diener, E., 1980. The Psychology of Group Influence. Erlbaum, New York.

Dooley, J.J., Pyżalski, J., Cross, D., 2009. Cyberbullying versus face-to-face bullying: a theoretical and conceptual review. Zeitschrift für Psychologie/Journal of Psychology 217, 182–188.

Dulmus, C., Theriot, M., Sowers, K., Blackburn, J., 2004. Student reports of peer bullying. Victimization in a Rural School, 1–16. Brunner/Routledge, SocINDEX with Full Text, EBSCOhost, viewed December 20, 2013.

Fanti, K.A., Demetriou, A.G., Hawa, V.V., 2012. A longitudinal study of cyberbullying: examining risk and protective factors. European Journal of Developmental Psychology 9 (2), 168–181.

Fight Crime sponsored studies: Opinion research corporation, 2006. Cyber Bully Teen. Available at: http://www.fightcrime.org/cyberbullying/cyberbullyingteen.pdf.

Görzig, A., Ólafsson, K., 2013. What makes a bully a cyberbully? Unravelling the characteristics of cyberbullies across twenty-five European countries. Journal of Children and Media 7 (1), 9–27.

Gradinger, P., Strohmeier, D., Spiel, C., 2009. Traditional bullying and cyberbullying: identification of risk groups for adjustment problems. Zeitschrift Für Psychologie/ Journal of Psychology 217 (4), 205–213.

Hinduja, S., Patchin, J.W., 2008. Cyberbullying: an exploratory analysis of factors related to offending and victimization. Deviant Behavior 29, 129–156.

Hinduja, S., Patchin, J.W., 2010. Bullying, cyberbullying, and suicide. Archives of Suicide Research 14 (3), 206–221.

Hunt, C., Peters, L., Rapee, R.M., 2012. Development of a measure of the experience of being bullied in youth. Psychological Assessment 24 (1), 156–165.

Juvonen, J., Gross, E.F., 2008. Extending the school grounds? – Bullying experiences in cyberspace. Journal of School Health 78, 496–505.

Kowalski, R.M., Limber, S.P., 2007. Electronic bullying among middle school students. Journal of Adolescent Health 41 (Suppl. 6), S22–S30.

Kowalski, R.M., Limber, S.P., 2013. Psychological, physical, and academic correlates of cyberbullying and traditional bullying. Journal of Adolescent Health 53, S13–S20.

Kowalski, R.M., Whittaker, E. Cyberbullying: prevalence, causes, and consequences. In: Rosen, L.D., Carrier, L.M., Cheever, N.A. (Eds.), The Handbook of Psychology, Technology, and Society. Wiley-Blackwell, Malden, MA.

Kowalski, R.M., Giumetti, G.W., Schroeder, A.N., Reese, H., 2012a. Cyberbullying among college students: evidence from multiple domains of college life. In: Wankel, C., Wankel, L. (Eds.), Misbehavior Online in Higher Education. Emerald Publishing Group, Bingley, UK, pp. 293–321.

Kowalski, R.M., Limber, S.E., Agatston, P.W., 2012b. Cyberbullying: Bullying in the Digital Age, second ed. Wiley-Blackwell, Malden, MA.

Kowalski, R.M., Giumetti, G., Limber, S.P., 2013a. Unique Contributions of Cyberbullying and Traditional Bullying to Negative Outcomes Associated with Victimization. Clemson University, unpublished manuscript.

Kowalski, R.M., Schroeder, A.N., Smith, C.A., 2013b. Bystanders and their willingness to intervene in cyberbullying situations. In: Hanewald, R. (Ed.), From Cyberbullying to Cyber Safety: Issues and Approaches in Educational Contexts. Nova Publishers, Sydney.

Kowalski, R.M., Giumetti, G., Schroeder, A., Lattanner, M., 2014. Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. Psychological Bulletin. Available at: http://dx.doi.org/10.1037/a0035618.

Lenhart, A., 6 May, 2010. Cyberbullying: What the Research is Telling Us. Available at: http://www.pewinternet.org/Presentations/2010/May/Cyberbullying-2010.aspx.

Limber, S., Kowalski, R.M., Agatston, P., 2009a. Cyber Bullying: A Prevention; Curriculum for Grades 6-12. Hazelton, Center City, MN.

Limber, S., Kowalski, R.M., Agatston, P., 2009b. Cyber Bullying: A Prevention; Curriculum for Grades 3-5. Hazelton, Center City, MN.

Monks, C.P., Smith, P.K., 2006. Definitions of bullying: age differences in understanding of the term, and the role of experience. British Journal of Developmental Psychology 24 (4), 801–821.

Nansel, T., Overpeck, M., Pilla, R., Ruan, W., Simons-Morton, B., Scheidt, P., 2001. Bullying behaviors among US youth: prevalence and association with psychosocial adjustment. Journal of the American Medical Association 285, 2094–2100.

O'Brennan, L.M., Bradshaw, C.P., Sawyer, A.L., 2009. Examining developmental differences in the social-emotional problems among frequent bullies, victims, and bully/victims. Psychology in Schools 46, 100–115.

Olweus, D., 1993. Bullying at School: What We Know and What We Can Do. Blackwell, New York.

Olweus, D., 2013. School bullying: development and some important challenges. Annual Review of Clinical Psychology 9, 1–14.

Pyżalski, J., 2012. From cyberbullying to electronic aggression: typology of the phenomenon. Emotional & Behavioural Difficulties 17 (3–4), 305–317.

Robers, S., Zhang, J., Truman, J., Snyder, T.D., 2010. Indicators of School Crime and Safety: 2010 (NCES 2011-002/NCJ230812). U.S. Department of Education.

Slonje, R., Smith, P.K., 2008. Cyberbullying: another main type of bullying? Scandanavian. Journal of Psychology 49, 147–154.

Sourander, A., Klomek, A.B., Ikonen, M., Lindroos, J., Luntamo, T., Koskelainen, M., Ristkari, T., Henenius, H., 2010. Psychosocial risk factors associated with cyberbullying among adolescents. Archives of General Psychiatry 67, 720–728.

Tokunaga, R.S., 2010. Following you home from school: a critical review and synthesis of research on cyber bullying victimization. Computers in Human Behavior 26, 277–287.

Varjas, K., Henrich, C.C., Meyers, J., 2007. Urban middle school students' perceptions of bullying, cyberbullying, and school safety. Journal of School Violence 8, 159–176.

Wade, A., Beran, T., 2011. Cyberbullying: the new era of bullying. Canadian Journal of School Psychology 26 (1), 44–61.

Wang, J., Iannotti, R.J., Nansel, T.R., 2009. School bullying among adolescents in the United States: physical, verbal, relational, and cyber. Journal of Adolescent Health 45, 368–375.

Willard, N.E., 2007. Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress. Research Press, Champaign, IL, US.

Williams, K.R., Guerra, N.G., 2007. Prevalence and predictors of internet bullying. Journal of Adolescent Health 41, 14–21.

Ybarra, M.L., Mitchell, K.J., 2004. Online aggressor/targets, aggressors, and targets: a comparison of associated youth characteristics. Journal of Child Psychology and Psychiatry 45, 1308–1316.

Ybarra, M.L., Diener-West, M., Leaf, P.J., 2007. Examining the overlap in internet harassment and school bullying: implications for school intervention. Journal of Adolescent Health 41, 42–50.

Cybernetics

Andrew Pickering, University of Exeter, Exeter, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews developments in cybernetics from the 1940s to the present, focusing on the most radical threads, those that foreground key questions of agency, performance, and emergence. It explores cybernetic intersections with fields including science and technology studies, brain science, psychiatry and antipsychiatry, biological computing, management, the environment, the arts, architecture, nonmodern spirituality, and the counterculture of the 1960s. It concludes with a discussion of the politics of cybernetics.

After a long prehistory spanning many fields, cybernetics was born in the early 1940s, around the time of World War II and officially named by Norbert Wiener in 1948 in his famous book, *Cybernetics, or Control and Communication in the Animal and the Machine*. Often identified with the Macy conferences held in the United States between 1946 and 1953, cybernetics also flourished in Britain (Pickering, 2010a), Europe, and the Soviet Union. Its visibility peaked in the 1960s to be followed by a gradual decline, and the word 'cybernetics' largely fell out of use – except as the root of words like 'cyberspace' and 'cyborg' – though lines of recognizably cybernetic work have propagated forward to the present, often under other names (Pickering, 2010a). In the third millennium, however, cybernetics is enjoying a renaissance in the social sciences, as well as the natural sciences, the humanities, and the arts.

Cybernetics has always been a hard field to pin down. Wiener presented it a synthesis of work in digital computing, information theory, and feedback control. As such, it represented a new kind of science, devoted to immaterial elements such as 'bits' of information rather than the material substances that define traditional sciences like physics, chemistry, and biology. Again, in contrast to the traditional organization of knowledge, cybernetics was strongly interdisciplinary. The Macy conferences brought anthropologists, sociologists, psychologists, and psychiatrists together with natural scientists, mathematicians, and engineers, and claims were made for cybernetics as a universal superscience capable of accommodating all of the disciplines (Bowker, 1993).

Nonmodern Ontology

The distinctiveness of cybernetics can be expressed in different ways. The most productive and radical, followed here, is in terms of ontology, its vision of what the world is like. The ontology of the modern sciences is that of a world of fixed and specifiable entities in fixed and specifiable relations to one another, and the job of the sciences is to generate knowledge about these entities (quarks, genes, social interests, molecular structures, and dark matter). The ontology of cybernetics is different and in that sense 'nonmodern' (Latour, 1993). Stafford Beer (1959: 18) described cybernetics as the science of 'exceedingly complex systems,' meaning systems we can never fully grasp conceptually and that are always becoming something new. The job of cybernetics is not, therefore, to produce articulated knowledge of specific entities, but rather somehow to conceptualize aspects of the world as built from ultimately unknowable elements, and to explore ways of acting accordingly. Three characteristic features stand out.

1. The cybernetic vision centers on agency and performance – in the sense of worldly actions – rather than on intentions and knowledge. The world is a place of action and reaction among a multiplicity of elements.
2. Agency is emergent in a brutal sense: we cannot know in advance how any entity will act when placed in a novel situation.
3. The analysis is decentered: in the play of performative action and interaction none of the elements acts as a still, controlling (and explanatory) center. Everything is at stake and liable to transformation in 'dances of agency' (Pickering, 1995). As far as the social sciences are concerned, this speaks of a 'posthumanist' decentering of the analysis, an acknowledgment that we ourselves are transformed and become something new in our traffic with people and things.

Performance, emergence, and decentering of the human: these topics are central to cybernetics but marginalized at best in modern thought. We can therefore see cybernetics as instantiating a nonmodern paradigm, in Thomas Kuhn's sense, radically disjoint from the Enlightenment disciplines. This, in turn, explains the resurgence of cybernetics in the new millennium, as indexing a growing conviction that we need to explore the world from new angles. We can now examine some of the important threads of the cybernetic paradigm.

Science and Technology Studies

Though its debts to cybernetics are obscure and sometimes amount to a rediscovery of cybernetic insights, the posthumanist wing of science and technology studies (STS) has worked out a cybernetic analysis close to the heartland of the traditional social sciences. The actor-network theory approach associated with the work of Bruno Latour, Michel Callon, and John Law (Latour, 1987, 2005); Donna Haraway's (1991, 2008) nonessentialist 'cyborg' reconceptualization of the unit of analysis and her parallel analysis of relations between

humans and animals; Katherine Hayles' (1999) discussion of embodiment and disembodiment; and Andrew Pickering's analysis of the history of science as a 'mangle of practice' (Pickering, 1995; Pickering and Guzik, 2008) all share a cybernetic vision of the world as a place of decentered and emergent performative becomings. Posthumanist STS grew out of detailed studies of laboratory practice to encompass, for example, macrotransformations of science, technology, and society in the nineteenth and twentieth centuries, and has articulated principled arguments against the presuppositions of traditional social theory (Pickering, 2010b).

Work in STS has increasingly resonated across the social sciences and humanities, informing important developments in sociology, anthropology, geography, politics, literary studies, and studies of management and organizations. Philosophical inspiration has been sought from pragmatism, especially the writings of William James, from continental philosophy, including Henri Bergson, Alfred North Whitehead, Martin Heidegger, and Gilles Deleuze, and even from premodern traditions, such as Taoism, which likewise envision the world as a place of endless decentered flows and becomings (Pickering, 2013). In contemporary philosophy, much the same ontology is elaborated in works often referred to as the 'new vitalism' and 'speculative realism,' such as those of Karen Barad, Manuel DeLanda, Graham Harman, Brian Massumi, and Isabelle Stengers.

The Brain

The original referent of cybernetics was the brain, but the brain understood as primarily a performative rather than cognitive organ (Pickering, 2010a). While postwar research in artificial intelligence (AI) sought to emulate conscious reasoning processes, cybernetics instead understood the brain as embedded in worldly actions and in preconscious processes of performative adaptation to the unknown (for example, homeostatic processes such as maintaining blood temperature constant in differing environments). While AI had centered on computer programming, cybernetic brain research typically consisted in building adaptive electromechanical systems. Gray Walter's (1953) robot 'tortoises' located and pursued light sources while negotiating obstacles encountered en route. Ross Ashby's (1952) homeostats engaged in transformative dances of agency with one another, randomly reconfiguring their inner circuitry until they achieved collective states of dynamic equilibrium. In the 1980s, tortoise-style robotics was reinvigorated by Rodney Brooks (1999), and this approach to situated robotics is now a major line of research alongside work centered on AI techniques.

Inspired by psychiatric concerns with epilepsy and madness, early cyberneticians were interested in the pathological as well as the normal brain. Walter drove his tortoises mad by contradictory conditioning and then cured them in electromechanical ways, which he argued shed light on the efficacy of sleep therapy, electroshock, and lobotomy, the great and desperate cures of 1950s psychiatry. Gregory Bateson derived a very different approach to madness from Ashby's homeostats. His concept of the double bind as precipitating schizophrenia was one of family members locked into pathological cycles of contradictory communication much like a combination of homeostats unsuccessfully hunting for equilibrium. And he understood the interplay of sufferer and psychiatrist as a symmetric and open-ended process of reciprocal adaptation (Bateson et al., 1956). This view was put into practice by R.D. Laing and colleagues at Kingsley Hall in London in the second half of the 1960s. Kingsley Hall was a centerpiece of the antipsychiatry movement, a commune in which the sane and the mad lived together in nonhierarchical relationships, with the object of sustaining sufferers through their 'inner voyages' and, at the same time, helping the sane to become new sorts of selves (Pickering, 2010a).

At the opposite extreme from madness, the early cyberneticians were interested in altered states and ecstasy. Walter (1953) included within his overall cybernetic framework states such as dreams, visions, telepathy, the achievement of nirvana, and the uncanny performances of Eastern yogis, such as suspending the heartbeat and the breath. He also experimented on the brain, using a strobe light to induce symptoms of epilepsy and discovering in the late 1940s that strobes also induce visions of dynamic geometrical patterns. This work on madness and altered states points to a cybernetic image of the self more encompassing than the rational and controlled image assumed and reproduced by AI and the modern social sciences, and in turn leads in the direction of nonmodern and non-Western understandings of consciousness and spirituality, which share with cybernetics a decentered posthumanist ontology. A common fascination with altered states, explorations of consciousness and Eastern philosophy and spirituality also constituted an important axis of intersection between cybernetics and the counterculture of the 1960s.

Ashby and Walter built machines to emulate the adaptive brain. In their biological computing project of the late 1950s and early 1960s, Stafford Beer and Gordon Pask tried instead to use naturally occurring adaptive systems as substitutes for human brains, attempting to find ways to use pond ecosystems, for example, to replace human managers in factories. The nonmodernity of this project is clear: instead of purposeful and protracted design, it rested on a hylozoist conviction that nature is endlessly lively, and that whatever we need can already be found there (Pickering, 2010a).

Management

The biological computing project was part of the cybernetic concern with theorizing and managing organizations. From the later 1960s onward, Beer's work in management consultancy was organized around what he called the viable system model or VSM (Beer, 1981). Instead of incorporating real biological material into the organization, Beer's idea was now to model information flows within the organization on the most adaptive biological system he could think of: the human brain and nervous system. Beer's formal diagram of the VSM represents information pathways between five levels of the organization modeled on their biological equivalents, running between production units at level 1 and the board of directors at level 5 through various intermediary levels and the firm's environment. Much of Beer's attention focused on level 3, which housed a set of operations research models of the organization,

and level 4, home for a big computer model of the firm's world, the wider economy. These models illustrate the place of articulated knowledge in the overall cybernetic picture. Beer intended them to be continually revised in relation to the organization's performance, and thus as geared into practice and performance, rather than controlling performance as in the modern cognitivist picture.

On the model of Ashby's homeostats, Beer insisted that internal relations between levels within the organization should themselves take the form of a performative dance of agency. The relation between levels was supposed to be a process of 'reciprocal vetoing,' of proposal and counterproposal, until some sort of equilibrium emerged that all of the parties could live with. This constructive give and take contrasts strikingly with the hierarchical command-and-control structures typical of modern management. The VSM has many followers in management today (Espejo and Harnden, 1989). Its most ambitious application was to the entire Chilean economy under the socialist regime of Salvador Allende in the early 1970s. Project Cybersyn, as it was called, went a long way in a couple of years before it was cut off by the Pinochet coup (Beer, 1981; Medina, 2011).

Beer also developed a structured approach to decentered decision making that he called Syntegration. Syntegration is a complex process with many iterations, usually extended over several days, but the basic idea is to assign participants to the edges of a notional icosahedron, and to organize a series of sequential discussions between the parties whose edges end at a common vertex, alternating in steps between the vertices at the end of each edge. In this way, arguments can progressively echo all around the icosahedron, eventually taking an emergent form controlled by no one in particular. Syntegrations have been performed on many topics, running from the reorganization of the British Operations Research (OR) Society (of which Beer was President) up to Israeli–Palestinian relations and world peace, and they continue to be held today (Beer, 1994).

The Environment

In the later years of his life, Gregory Bateson turned his attention to the environment. Further articulating the concerns of the nascent environmental movement, in the late 1960s he developed a cybernetic understanding of our relations with nature (Bateson, 1968). He argued that we should understand nature as itself lively and emergent, as always liable to surprise us, and that attempts to manage and reconfigure the environment should thus be expected to elicit unexpected and possibly catastrophic responses – as in the disastrous effects of pesticides on wildlife described in Rachel Carson's *Silent Spring*. Bateson argued that we need new ways of imagining nature as beyond any linear forms of control. The deep ecology movement, James Lovelock's Gaia hypothesis and work in complexity theory on environmental resilience and tipping points have continued this reimagination up to the present (Lenton et al., 2008). All of these point to an ontological decentering of the place of humanity in nature and evoke echoes of premodern ontologies, especially Taoism (Jullien, 1999). Adaptive approaches to environmental management attempt to put such posthumanist imaginings into practice, exploring nature's tendencies through performative experimentation 'in the wild' – staging artificial floods on rivers, for example, to learn more about how to manage ecosystems (rather than trying to calculate their behavior in advance) – or recognizing, in a hylozoist fashion, that nature is better at absorbing the shocks of flooding than engineers are at containing them, as in the Room for the River project in Holland (Gunderson and Light, 2006; de Groot and Lenders, 2006; Asplen, 2008; Pickering, 2013).

Art

Artists are in the lead in the current cybernetic resurgence, often exploring lines of work opened up in the 1960s (Gere, 2002; Salter, 2010). Grey Walter's tortoises have spawned endless variations and elaborations in robotic artwork. Made for the 1968 *Cybernetic Serendipity* exhibition in London, Gordon Pask's *Colloquy of Mobiles* was an array of five adaptive robots, three designated male and two female, that sought to satisfy each other's desires with light beams, rotating in space and communicating by sounds in pursuit of equilibrium and temporary quiescence. The *Colloquy* was an artistic variant of a long line of adaptive teaching and learning machines built by Pask and his colleagues, beginning with a machine called Musicolour in the early 1950s (Pickering, 2010a). Musicolour turned a musical performance into a light show, but the connection between sound and light was cognitively opaque to the musician. The machine adapted to each performance as it went along; thresholds for activating lights increased, and the machine would eventually 'get bored' and cease to respond to repeated tropes, thus encouraging the performer to adapt to the machine by trying something new. In effect, each Musicolour performance was a decentered and emergent dance of agency between the musician and the machine. The *Colloquy* in its turn was a dance of agency between purely nonhuman participants. Pask elaborated an aesthetic theory around these machines, arguing that we enjoy and derive satisfaction from the process of coming to terms with inscrutable systems, as long as it is possible to achieve some relationship in a finite time (Pask, 1971). Simon Penny's *Petit Mal*, first shown in the early 1990s, is a mobile robot that tracks the motion of humans around it and, again, lures them into open-ended dances of agency (Penny, 2000). Ruairi Glynn's (2008) work, *Performative Ecologies*, updates Pask's *Colloquy* by adding learning algorithms to robots that dance with both human spectators and each other (Glynn, 2008). The *Pask Present* exhibition in Vienna in 2008 brought together a whole range of Pask-inspired artworks (Glanville and Müller, 2008).

On a different line of descent, Walter's explorations of stroboscopic flicker effects crossed over into the art world in the 1960s via the Beats. William Burroughs was a perceptive reader of Walter's *Living Brain*, as is clear from a reading of *Naked Lunch*; Brion Gysin learned of flicker from him and went on to build flicker machines as artworks and putative replacements for TVs in the home (Geiger, 2003). The early years of the twenty-first century have seen a proliferation of flicker artworks, often conjoined with powerful low-frequency sounds, as both material entities and film effects (for example, Matthijs

Munnik's massive 2013 installation *Citadels: Common Structures* (Vimeo, 2013)). One can think of flicker art as a technology of the nonmodern self (Foucault, 1988), continuing Walter's explorations of the performative brain. Works like Chris Salter's *JND* also fall into this category, though operating on different sensory modalities, combining tactile stimulation on the margins of perception in a work that adapts to the responses of participants (another dance of agency) (Salter, 2012).

Another strand of contemporary cybernetic art thematizes the performative agency of nature, matter, and complex systems. Chris Welsby's work in experimental cinema and video since the late 1960s aims in various ways to hand over control in film making to the elements: sun, wind, weather, and the motion of the planet in space. His 1974 movie *Seven Days* is a time-lapse film of a Welsh landscape over a period of a week in which the camera both tracks the sun and oscillates between pointing at the land when the sun is out and at the sky when the sun is obscured by clouds (Welsby, 2011). In Richard Brown's work, *The Electrochemical Glass*, three electrodes of copper, aluminum, and iron were sandwiched in a conducting fluid between two glass plates and the assemblage evolved unpredictably over time into a continuously changing work of art as material was deposited between them. This work again confronts us with the emergent liveliness of nature. Invoking another premodern ontology, Brown (2003: 20) described this work "as a daily reflection of process, decay, transmutation and growth; the slow changes resonating with memory and notions of self – a form of contemplative alchemy."

Various forms of bioart directly thematize the emergent agency of biological systems, including the high-tech and very scientific work of the SymbioticA laboratory in Australia, which aims literally to grow sculptures from living cells (Catts and Zurr, 2002) and, at the everyday level, the ancient East Asian art of bonsai (both of which entail dances of agency between the artist and a living substrate). Ursula Damm's (2010) installation *Greenhouse Converter* (for online video, see Damm, 2014) exemplifies the liveliness of coupled material and biological systems. The state of a mini ecology of algae and water fleas acts to reveal or obscure an array of light-emitting diodes (which spell out the word 'beloved,' a reference to Lynn Margulis' theory of endosymbioisis). The lights change frequency to maintain a dynamic balance of fleas and algae (and spectators can interfere with the system but not control it, by changing the gas composition of the water that flows through it). At the human end of the spectrum, Stelarc's works – beginning with performances in which he was suspended in the air via hooks through his flesh, and running through robotic and surgical augmentations of his limbs and sensory systems – all thematize the performative rather than cognitive aspects of our being and coupling into the world (Stelarc, 2007).

Developments in architecture have paralleled those in art. In the 1960s, Cedric Price, as well as the British architectural collective Archigram, imagined buildings and cities that could change and adapt in response to emergent developments in their use and environment (rather than being designed to meet preconceived needs) (Sadler, 2005; Mathews, 2007). Gordon Pask's first involvement with architecture came in the early 1960s as head of the cybernetics committee involved in planning a public building in London called the Fun Palace. Pask's aim was to design a control system for the building, which would enable it not only to adapt to emergent patterns of use, but also sometimes to get bored along the lines of Musicolour, reconfiguring itself in unpredictable ways and thus inviting its human inhabitants to find novel uses and practices (Mathews, 2007). John Frazer (1995) elaborated on Pask's approach to articulate an evolutionary vision of architecture, in which built structures evolve in relation to their inner and outer environments. Today, adaptive and generative approaches are leading themes in architectural research.

Collectively, these works (and many more) function as ontological theater. They stage aspects of the cybernetic ontology of performative and emergent interactions among humans and nonhumans, and in doing so help us to grasp the cybernetic vision by translating it into nonverbal and performative forms. If orreries can function as icons of the modern ontology of a regular and machinelike cosmos, interactive robots, flicker machines, and movies controlled by the weather are mnemonics for a cybernetic world of endless and unpredictable flows and becomings. Returning to an earlier heading, we can understand these artworks as microenvironments that might, in turn, help to refamiliarize us along Batesonian lines with nature at large and the macroenvironment, and to suggest new approaches to engaging with the latter.

Politics

From its beginning, cybernetics has been politically controversial. On the one hand, it has been seen as a profoundly threatening science of technocratic control and manipulation. The subtitle of Wiener's *Cybernetics* already defined it to be the science of "control and communication in the animal and the machine," and in 1950 in *The Human Use of Human Beings* Wiener himself warned of the dangers of cybernetic automation. Bernard Wolfe's science fiction novel *Limbo* (1952) sketched a horrific picture of a cybernetic world of brain surgery, amputations, and prosthetics. The most extended recent critique is from the Tiqqun group, arguing that "Cybernetics is the police-like thinking of the Empire, entirely animated by an offensive concept of politics … Cybernetics is war against all that lives and all that is lasting" (Tiqqun, n.d.: 8). Against that view, Haraway (1991) opted for the figure of the cyborg, the cybernetic organism, as an icon for her version of a nonessentialist and emergent radical socialist-feminist project, and Pickering (2010a) likewise argues that the history of cybernetics offers us a set of 'sketches of another future,' a future radically different from and preferable to contemporary neoliberalism.

One can understand these divergent political evaluations as referring to different threads in the history of cybernetics and the specific material exemplars from which they grew: the thermostat and the homeostat, respectively. The negative evaluation focuses on cybernetics as a branch of control engineering. Deriving from his work in World War II, Wiener's vision of cybernetics grew out of his wartime work on servomechanisms, simple feedback devices like the domestic thermostat (Galison, 1994). The thermostat seeks to minimize deviations in some specified variable, such as the temperature

in a room. If the temperature exceeds a preset value, the thermostat turns the heating off, and then back on when the temperature falls. The political question is, what would this version of cybernetics look like if played out as a blueprint for a social world? The critical argument is that it would be a disaster, a horror show. In particular, it is easy to see how discussions of the thermostat resonate with visions of overarching social control, with thermostatlike technosocial mechanisms sitting on top of and optimizing the performance of existing organs of control: the police, the army, etc. One thinks of Gilles Deleuze's remarks on 'control society.' And the picture becomes more sinister when one observes that what escape this sort of cybernetic analysis are precisely the goals and purposes of feedback systems. The thermostat's target temperature is something given from outside the system, a number programmed by the user. No great analogical leap is required to get from there to a society governed by a small group of politicians, bankers, and media men setting the operating parameters of society for their own benefit.

But other strands of cybernetics conjure up a quite different political vision. Thermostats are hardwired to maintain control of some variable, but, as discussed above, Ashby's homeostats explored their environments performatively and responded to whatever came back to them, reconfiguring themselves internally until some state of equilibrium was found. And as a blueprint for social organization and action, the homeostat suggests not Tiqqun's vision of a control society, but something very different: a society evolving through transformative open-ended experimentation. One can think here of the contrast between traditional management theory and practice as hierarchical control and Beer's VSM approach, with its symmetric homeostatic linkages between levels of organization; or between traditional authoritarian psychiatry, in which doctors are the only legitimate agents, and the symmetric communal relations of antipsychiatry; or between our usual command-and-control relationship with the environment and the adaptive cybernetic approaches mentioned above.

Philosophically, these political contrasts map onto Martin Heidegger's (1977) distinctions between 'enframing' and 'poiesis' as different modes of being in the world. Enframing is the defining stance of modernity, and grasps the world, human as well as nonhuman, as 'standing reserve' to be put to work in preconceived projects. Poiesis, instead, is a stance of experimental openness to whatever the world emergently has to offer us – finding out, in place of domination. Heidegger looked back to the arts in ancient Greece for examples of poiesis. The positive political evaluation is that the homeostat-inspired wing of cybernetics offers us examples in the present from which to develop ways of thinking and acting radically different from the society of control (Pickering, 2010a).

See also: Actor-Network Theory; Adaptation, Fitness, and Evolution; Environmental Sciences; Human Sciences, History of; Religions of East Asia; Science and Social Movements; Situated Knowledge, Feminist and Science and Technology Studies Perspectives.

Bibliography

Ashby, W.R., 1952. Design for a Brain, second ed., 1960. Chapman & Hall, London.

Asplen, L., 2008. Going with the flow: living the mangle in environmental management practice. In: Pickering, A., Guzik, K. (Eds.), The Mangle in Practice: Science, Society and Becoming. Duke University Press, Durham, NC, pp. 163–184.

Bateson, G., 1968. Conscious purpose versus nature. In: Cooper, D. (Ed.), To Free a Generation: The Dialectics of Liberation. Collier, New York, pp. 34–49.

Bateson, G., Jackson, D., Haley, J., Weakland, J., 1956. Towards a theory of schizophrenia. Behavioral Science 1, 251–264. Reprinted in Bateson, Steps to an Ecology of Mind. Ballantine, New York, 1972, pp. 201–227.

Beer, S., 1959. Cybernetics and Management. English Universities Press, London.

Beer, S., 1981. Brain of the Firm, second ed. Wiley, New York.

Beer, S., 1994. Beyond Dispute: The Invention of Team Syntegrity. Wiley, New York.

Bowker, G., 1993. How to be universal: some cybernetic strategies, 1943–70. Social Studies of Science 23, 107–127.

Brooks, R., 1999. Cambrian Intelligence: The Early History of the New AI. MIT Press, Cambridge, MA.

Brown, R., 2003. Alchemy, Mimetics, Immersion and Consciousness. Paper Presented at Melbourne DAC2003. Available at: hypertext.rmit.edu.au/dac/papers/Brown.pdf (accessed 17.02.14.).

Catts, O., Zurr, I., 2002. Growing semi-living sculptures: the tissue culture and art project. Leonardo 35 (4), 365–370.

Damm, U., 2010. Greenhouse converter. In: Corral, A. (Ed.), Process as Paradigm. LABoral Centro de Arte y Creación Industrial, Gijón, Spain, pp. 76–77.

Espejo, R., Harnden, R. (Eds.), 1989. The Viable System Model: Interpretations, and Applications of Stafford Beer's VSM. Wiley, New York.

Foucault, M., 1988. In: Martin, L.H., Gutman, H., Hutton, P.H. (Eds.), Technologies of the Self: A Seminar with Michel Foucault. University of Massachusetts Press, Amherst.

Frazer, J.H., 1995. An Evolutionary Architecture. Architectural Association, London.

Galison, P., 1994. The ontology of the enemy: Norbert Wiener and the cybernetic vision. Critical Inquiry 21, 228–266.

Geiger, J., 2003. Chapel of Extreme Experience: A Short History of Stroboscopic Light and the Dream Machine. Soft Skull Press, New York.

Gere, C., 2002. Digital Culture. Reaktion Books, London.

Glanville, R., Müller, A. (Eds.), 2008. Pask Present: An Exhibition of Art and Design Inspired by the Work of Gordon Pask (28 June 1928 to 28 March 1996) Cybernetician and Artist. Echoraum, Vienna.

Glynn, R., 2008. Conversational Environments Revisited. In 19th Meeting of Cybernetics & System Research. Graz, Austria.

de Groot, W., Lenders, H., 2006. Emergent principles for river management. Hydrobiologia 565, 309–316.

Gunderson, L., Light, S., 2006. Adaptive management and adaptive governance in the everglades ecosystem. Policy Science 39, 323–334.

Haraway, D., 1991. A cyborg manifesto: science, technology, and socialist-feminism in the late twentieth century. In: Haraway, D. (Ed.), Simians, Cyborgs, and Women. Free Association Books, London, pp. 149–181 (Reprinted from Socialist Review 80, 1985, 65–107).

Haraway, D., 2008. When Species Meet. University of Minnesota Press, Minneapolis.

Hayles, N.K., 1999. How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics. University of Chicago Press, Chicago.

Heidegger, M., 1977. The question concerning technology (W. Lovitt, Transl). In: Heidegger, M. (Ed.), The Question Concerning Technology and Other Essays. Harper & Row, New York, pp. 3–35.

Jullien, F., 1999. The Propensity of Things: Toward a History of Efficacy in China. Zone Books, New York.

Latour, B., 1987. Science in Action: How to Follow Scientists and Engineers through Society. Harvard University Press, Cambridge, MA.

Latour, B., 1993. We Have Never Been Modern. Harvard University Press, Cambridge, MA.

Latour, B., 2005. Reassembling the Social: An Introduction to Actor-Network Theory. Oxford University Press, Oxford.

Lenton, T.M., Held, H., Kriegler, E., Hall, J.W., Lucht, W., Rahmtorf, S., Schellnhuber, H.J., 2008. Tipping elements in the Earth's climate system. Proceedings of the National Academy of Sciences of the United States of America 105 (6), 1786–1793.

Mathews, S., 2007. From Agit-prop to Free Space: The Architecture of Cedric Price. Black Dog, London.

Medina, E., 2011. Cybernetic Revolutionaries: Technology and Politics in Allende's Chile. MIT Press, Cambridge, MA.

Pask, G., 1971. A comment, a case history and a plan. In: Reichardt, J. (Ed.), Cybernetics, Art, and Ideas. New York Graphics Society, Greenwich, CT, pp. 76–99.

Penny, S., 2000. Agents as artworks: agent design as artistic practice. In: Dartenhahn, K. (Ed.), Human Cognition and Social Agent Technology. John Benjamins, Amsterdam, pp. 395–414.
Pickering, A., 1995. The Mangle of Practice: Time, Agency, and Science. University of Chicago Press, Chicago.
Pickering, A., 2010a. The Cybernetic Brain: Sketches of Another Future. University of Chicago Press, Chicago.
Pickering, A., 2010b. Material culture and the dance of agency. In: Hicks, D., Beaudry, M. (Eds.), The Oxford Handbook of Material Culture Studies. Oxford University Press, Oxford, pp. 191–208.
Pickering, A., 2013. Being in an environment: a performative perspective. Natures Sciences Sociétés 21 (1), 77–83. http://dx.doi.org/10.1051/nss/2013067 (accessed 22.06.14.).
Pickering, A., Guzik, K. (Eds.), 2008. The Mangle in Practice: Science, Society and Becoming. Duke University Press, Durham, NC.
Sadler, S., 2005. Archigram: Architecture without Architecture. MIT Press, Cambridge, MA.
Salter, C., 2010. Entangled: Technology and the Transformation of Performance. MIT Press, Cambridge, MA.
Salter, C., 2012. JND: an artistic experiment in bodily experience as research. In: Peters, D., Eckel, G., Dorschel, A. (Eds.), Bodily Expression in Electronic Music: Perspectives on Reclaimed Performativity. Routledge, London.
Stelarc, 2007. Extra Ear: Alternate Anatomical Architectures. Public lecture, Brunel University, West London, UK.
Tiqqun, n.d. The Cybernetic Hypothesis (Online Book). Anarchist Library. http://theanarchistlibrary.org/library/tiqqun-the-cybernetic-hypothesis (accessed 22.06.14.).
Walter, W.G., 1953. The Living Brain. Duckworth, London.
Welsby, C., 2011. Technology, nature, software and networks: materializing the post-romantic landscape. Leonardo 44, 101–106.
Wolfe, B., 1952. Limbo. Random House, New York.

Relevant Websites

ursuladamm.de/treibhauskonverter-venus-v/ – Damm, U., 2014. Greenhouse Converter. (accessed 18.02.14.).
www.ruimtevoorderivier.nl/meta-navigatie/english.aspx – Room for the River, n.d. (accessed 18.02.14.).
vimeo.com/61338326 – Vimeo, 2013. Matthijs Munnik, Citidels: Common Structures. Live recordings from the opening of Sonic Acts 2013 in Stedilijk Museum, Amsterdam, by Tanja Busking. (accessed 18.02.14.).

Cyclical Industrial Dynamics

Hao Tan, University of Newcastle, Callaghan, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In many industries, market dynamics such as sales, price, capacity, and investment are highly cyclical, observed in the form of successive expansions and contractions that affect all firms. These industry-level cycles may have different patterns from the general business cycles at the country level; and they are also distinct from fluctuations due to seasonal effects and those due to random shocks in the industry. The generation of industry cycles is considered from the Schumpeterian perspective to be endogenous, while from the neoclassical perspective they are created by exogenous shocks. Different mechanisms may be seen to account for them, depending on the industry, ranging from effects of macro-business cycles, to mismatch and delay between different market dynamics, and to dynamics of innovations, among others. These cyclical industrial dynamics have profound implications for firms' strategizing.

Many industries have long been characterized by strong cyclicality in their market dynamics, covering such attributes as demand, revenues, prices, profits, and capacity. Those cyclical industrial dynamics, variously termed in different contexts as boom and bust cycles, commodity cycles, durable goods cycles, or simply industry cycles, are of great interest to both economic theoreticians and practitioners. Firms are keen to understand causes of the cyclical dynamics and predict their movements in order to 'stay ahead of the curve.' A better understanding of those cyclical industrial dynamics is also important to the field of social and behavioral sciences. After all, the cyclical dynamics at the industry level arise as a result of behaviors of individual players of the industry; and together the cyclical behaviors of many industries, through superposition, comprise cycles at the macroeconomic level that are of social and industrial significance.

Research into cyclical industrial dynamics, however, has been overshadowed by research concerning two well-known phenomena, namely, business cycles or economic cycles at the country level, and industry life cycles. First, research on business cycles focuses on the general cycles in the aggregate economy. Those general business cycles that are defined by macroeconomic quantities such as GDP and total employment, however, do not usually recognize different cyclical patterns and behaviors existing at the industry level. Moving against the general business cycles, some industries are procyclical (e.g., most of the manufacturing industries), some are acyclical (e.g., the pharmaceutical, education service, and public service industries), and some are even countercyclical (e.g., the health service industry). Many industry-specific factors contribute to the rise of cycles. For high-technology industries, their cycles are affected by the uneven pace of technological changes. For the agriculture sector, the cycles are affected by weather conditions such as droughts and floods. Other sectors such as utilities in many countries are affected by election cycles because of the latter can bring changes to the policy and regulatory environment. Industries also differ in such aspects as nature of technologies and products, industry structure, stage of development, and competition dynamics, which all inevitably affect the emergence and persistence of their own cycles. As a result, the timing, duration, and amplitude of industry-specific cycles may vary widely. Nevertheless, their significance is fundamental.

Second, the term 'industry cycles' is sometimes confused with 'industry life cycles.' The latter, however, is mainly concerned with the long-term evolution during the lifetime of a technology, which typically goes through a sequence of stages such as the exploratory stage, the growth stage, and the mature stage (Klepper, 1997).

This article focuses on cyclical industrial dynamics and provides an overview of the research on the meso-level, medium-term phenomenon. Those cycles, while differing in patterns and mechanisms across industries, are characterized in terms of successive expansions and contractions of sales, prices, capacity, and investment in a given industry, and, in most cases, as recurrent deviations from the long-term trends of the market dynamics in the industry. The research of cyclical industrial dynamics has been concerned with three main questions: What are they? Why do they happen? How can they be utilized by firms themselves as well as by policy makers?

Identification and Measurement

Analysis of cyclical industrial dynamics must start with identification and measurement, especially recognition of the cyclical pattern of industrial activities from other patterns displayed in industrial data series, such as long-term trends, fluctuations due to seasonal effects, and those due to random shocks. In contrast to seasonal fluctuations, cyclical patterns of an industry are not of a fixed period and normally have longer duration (more than a few months) in their phases. Nevertheless, they are more regular than random fluctuation.

One approach to identify and measure cycles in industrial activities is based on the decomposition method. Assuming an industrial time series is underpinned by four basic components, each presenting different driving factors, the data series can be 'detrended,' 'seasonally adjusted,' and 'deflated,' in order to remove the trend, seasonal, and random component parts to derive the cyclical pattern. Figure 1 shows an example of the use of such a method to bring out the cyclical pattern of global semiconductor sales. In particular, the deflated data series is subject to a Hodrick–Prescott (H–P) filter that separates the trend and cycle components.

Similar to the practices of dating business cycles by the National Bureau of Economic Research in the United States or

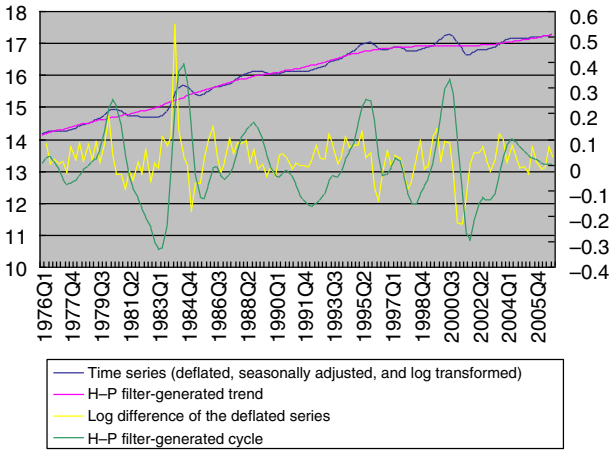

Figure 1 The cyclical pattern based on a decomposition approach. Source: Tan, H., Mathews, J.A., 2010a. Identification and analysis of industry cycles. Journal of Business Research 63 (5), 456. Reprinted with permission of Elsevier via the Copyright Clearance Center.

the Central Statistics Office in the United Kingdom, final turning points of cycles at the industry level can then be determined using preset dating rules or the so-called nonparametric dating algorithm. One such algorithm, as formulated by Bry and Boschan (1971) for dating business cycles, states that a peak (trough) occurs at time t whenever $\{yt > (<)yt \pm k\}$, $k = 1, \ldots, K$, where K is set to five (see Harding and Pagan, 2002). In other words, major peaks and troughs can be identified only if there are no other observations with larger (smaller) values within five periods of time prior to or following the point of concern. Applying modified rules to the cyclical pattern derived from the global semiconductor shipment data, Figure 2 shows the cycles of the industry along with their turning points.

Another approach to identifying and measuring industry cycles heavily relies on econometric models that fit historical data series and describe cycles based on estimated parameters. Models can be based on analysis of a time series in the time domain or in the frequency domain. A widely employed class of models analyzing time series in the time domain involve the so-called autoregressive integrated moving average (ARIMA) method. The method specifies a model by three main parameters representing respectively the possible autocorrelation pattern of the data series, the amount of differencing to make the series stationary, and the order of the moving average. Recently, efforts have been made to employ variants of the ARIMA method such as the Markov regime-switching model to study industry cycles (Liu and Chyi, 2006). However, those models usually do not generate turning points of cycles but are to forecast the future movement of market dynamics. Nor do those models intend to estimate properties of the cycles such as duration and amplitude.

Industrial data series can also be analyzed in the frequency domain, in which most powerful cycles, if any, can be detected. Figure 3 illustrates such an analysis. The analysis transforms the industrial data in the time domain into the frequency domain and generates a frequency spectrum through a Fourier transform (Tan and Mathews, 2010a). The three peaks in the frequency spectrum represent not just one but three pronounced cyclical components underpinning the industrial data, with frequencies of 0.25, 0.44, and 0.97 cycles per year. In other words, the industrial dynamics appear to be driven by at least three types of cycles with average periods of 4, 2.29, and 1.03 years. In this case, it seems that the three cyclical components may correspond to different cyclical effects, such as general business cycles, industry-specific factors, and seasonal effects. Based on a composition of the detected cyclical components, the industrial data series can be modeled and its cyclical dynamics can be more closely analyzed, as shown in Figure 4.

Explanations

In searching the drivers of cyclical industrial dynamics, researchers naturally look into existing theories for business cycles. Early theorists tended to attribute the business cycles to exogenous factors such as the periodicity of sunspots or movement of the planet Venus. Later economists, including Adam Smith, believed that business cycles arose as a result of the self-correcting mechanism of the economy. Contemporary work explains business cycles in terms of the role of governments in executing problematic monetary policies, or random technology shocks as proposed by the Real Business Cycle

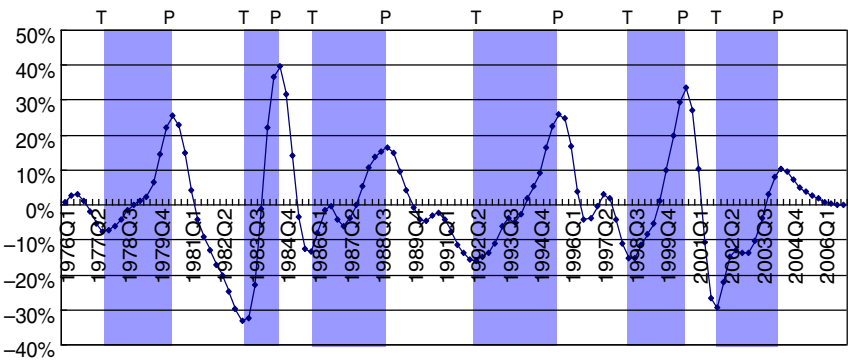

Figure 2 The semiconductor industry cycle defined. Note: T = trough; P = peak. Source: Tan, H., Mathews, J.A., 2010a. Identification and analysis of industry cycles. Journal of Business Research 63 (5), 458. Reprinted with permission of Elsevier via the Copyright Clearance Center.

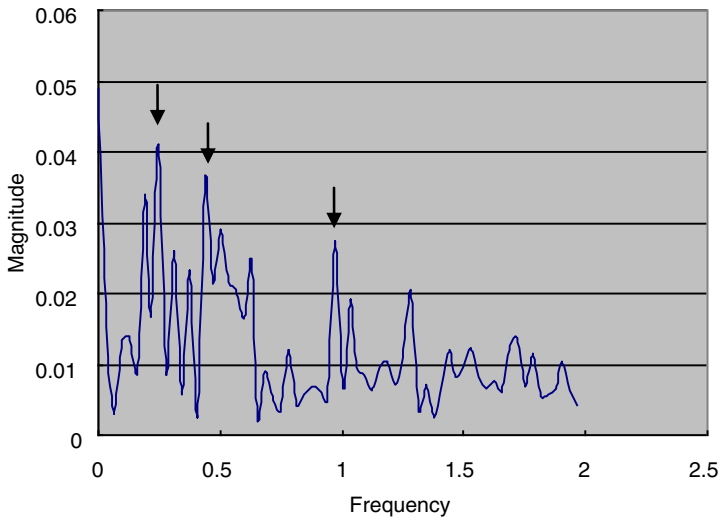

Figure 3 The industrial data in the frequency domain. Source: Tan, H., Mathews, J.A., 2010a. Identification and analysis of industry cycles. Journal of Business Research 63 (5), 460. Reprinted with permission of Elsevier via the Copyright Clearance Center.

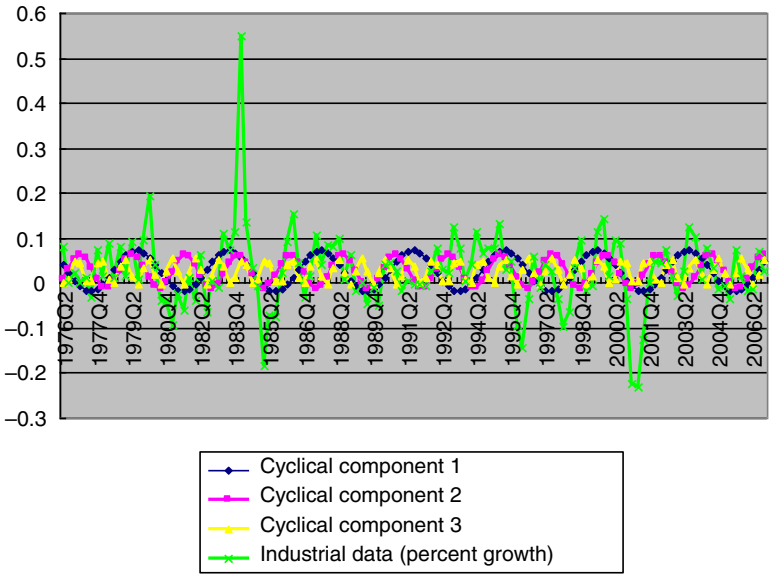

Figure 4 Three cycles composing the industrial data. Source: Tan, H., Mathews, J.A., 2010a. Identification and analysis of industry cycles. Journal of Business Research 63 (5), 460. Reprinted with permission of Elsevier via the Copyright Clearance Center.

theory. By contrast, the school of thought that sees cycles as being endogenously generated takes inspiration from Joseph A. Schumpeter, who distinguished four types of business cycles and attributed them to different mechanisms, ranging from excessive inventory by firms (for the 3–5 year Kitchin cycle), to lags in fixed investment spending (for the 7–11 year Juglar cycle), fluctuations in infrastructure investments (for the 15–25 year Kuznets cycle), and some long-term factors affecting the human society (for the 45–60 year Kondratieff cycle) (Schumpeter, 1939). These are all endogenous mechanisms.

Whereas the majority of contemporary business cycle theories are based on studies of behaviors of industries and firms and indeed are inspiring for industry cycle research, explanations of cyclical industrial dynamics will need to be from a different perspective. As the focus is on individual industries, which are subsystems of a macroeconomy, the general macroeconomic conditions now act as an 'exogenous factor'; and the effects of such factors as monetary policy become indirect through the change of demand of end users. Most individual industries are affected by wider business cycles, but heterogeneity across industries cannot be ignored. Therefore, it seems that more specific explanations are needed for the cyclical dynamics of given industries.

For manufacturing industries, especially those producing durable goods, a profound mechanism in driving the cycles is associated with the time lag of capacity building in response

to demand from downstream industries. From a system dynamics perspective, a substantial time delay in feedback loops consisting of processes of information collection, decision making, and action is largely responsible for industry cycles; and the swings tend to amplify when moving up along the supply chain from downstream to upstream industries (Sterman, 2000). For example, based on an analysis of the feedback structure of the commercial jet aircraft industry involving time lags between travel demand, order backlog, airline capacity, fares, and manufacturer production capacity, Lyneis (2000) established a system dynamics model to forecast the cycles in orders for aircraft, and found that the model describes historical cycles well.

For industries featuring strong competition, some argue that gamelike competitive behaviors of firms may cause cycles of the industries (Dearden et al., 1999). Firms may simultaneously engage in capacity addition or deletion activities, for reasons such as strategic preemption or 'herd' behavior; thus the industry as a whole is likely to end up with excess or insufficient capacity. The succeeding overcapacity or undercapacity behaviors lead to cycles in other market dynamics such as price and sales.

For high-technology industries, sales and price cycles are strongly influenced by technological changes in the industry. It is well-known that technological changes usually occur at an uneven pace, with radical and slow innovations being introduced alternately, a notion that goes back to Schumpeter (1939). In particular, Anderson and Tushman (1990) argue that sales of the industry tend to peak after a 'dominant design' has emerged because uncertainties associated with competing design paths are diminished. An examination of the cyclical dynamics in sales of the global semiconductor industry against the introductions of generations of integrated circuit technology seems to support the argument (Tan and Mathews, 2010a).

In sum, there seem to be multiple mechanisms that trigger and sustain cyclical industrial dynamics. Some drivers of industry cycles may be specific to a given industry. In addition, cycles of sales, price, capacity, and investment of an industry interact with each other and result in lead–lag relationships of the cycles (Tan, 2011).

Implications

In some industries, cyclical industrial dynamics constitute a fundamental dimension of the environment of firms. In particular, the significance of industry-specific cycles to the long-term development of industries lies in the idea of cleansing mechanisms as discovered by Joseph Schumpeter (1939). During industry cycle downturns, resources are released from firms and technologies with poorer performance, and reallocated to those that are more effective and promise to succeed in the market, thus enhancing the overall productivity and competitiveness of the industry. In line with this notion, the recent opportunity hypothesis argues for the positive effects of recessions for economic growth because productivity-improving activities are more likely to occur during the period thanks to the lower opportunity costs (Aghion and Saint-Paul, 1998).

From the perspective of individual firms, cyclical industrial dynamics bring rare opportunities for those that can strategize accordingly. For example, in some industries featuring very high barriers such as the global flat-panel display industry, challengers have been able to enter the industry only during industry cycle downturns because only at those times do expertise and technologies become available (Mathews, 2005). Established players can also benefit from industry cycles. Some recent examples of firms that utilize industry cycles to gain competitive advantages include Nokia (see Carral and Kajanto, 2008) and Intel (see Tan and Mathews, 2010b) in their respective industries. Thanks to their proactive investment and strategic positioning during the industry cycle downturns in the 1990s and 2000s, both came out from the industry downturns in much better shape than their competitors.

See also: Commodity Chains; Economic Growth: Measurement; Entrepreneurship and Economic Growth; Industrial Dynamics: Patenting Perspectives; Innovation and Technological Change, Economics of; Innovation; Strategizing.

Bibliography

Aghion, P., Saint-Paul, G., 1998. Virtues of bad times: interaction between productivity growth and economic fluctuations. Macroeconomic Dynamics 2 (3), 322–344.

Anderson, P., Tushman, M.L., 1990. Technological discontinuities and dominant, designs: a cyclical model of technological change. Administrative Science Quarterly 35, 604–633.

Bry, G., Boschan, C., 1971. Cyclical Analysis of Time Series: Selected Procedures and Computer Programs. National Bureau of Economic Research, New York.

Carral, R., Kajanto, M., 2008. Nokia: a case study in managing industry downturn. Journal of Business Strategy 29 (1), 25–33.

Dearden, J.A., Lilien, G.L., Yoon, E., 1999. Marketing and production capacity strategy for non-differentiated products: winning and losing at the capacity cycle game. International Journal of Research in Marketing 16, 57–74.

Harding, D., Pagan, A.R., 2002. Dissecting the cycle: a methodological investigation. Journal of Monetary Economics 49 (2), 365–381.

Klepper, S., 1997. Industry life cycles. Industrial and Corporate Change 6 (1), 145–182.

Liu, W.-H., Chyi, Y.-L., 2006. A Markov regime-switching model for the, semiconductor industry cycles. Economic Modelling 23, 569–578.

Lyneis, J.M., 2000. System dynamics for market forecasting and structural analysis. System Dynamics Review 16 (1), 3–25.

Mathews, J., 2005. Strategy and the crystal cycle. California Management Review 47 (2), 6–32.

Schumpeter, J., 1939. Business Cycles: A Theoretical, Historical and Statistical, Analysis of the Capitalist Process. McGraw-Hill, New York.

Sterman, J.D., 2000. Business Dynamics. McGraw-Hill, Boston.

Tan, H., 2011. Cyclical industrial dynamics in the global IT sector: origins and sequencing. Industrial and Corporate Change 20 (1), 175–200.

Tan, H., Mathews, J.A., 2010a. Identification and analysis of industry cycles. Journal of Business Research 63 (5), 454–462.

Tan, H., Mathews, J.A., 2010b. Cyclical industrial dynamics: the case of the global semiconductor industry. Technological Forecasting & Social Change 77 (2), 344–352.

D

Dahl, Robert A (1915–2014)

Charles HG Crothers, School of Social Sciences and Public Policy, AUT University, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Robert Dahl (1915–2014) was a prominent political scientist whose academic career was based at Yale University and his studies focused on democracy from the theoretical, empirical, and normative perspectives. He also examined power (or influence) as a central concept and was a leader in the application of behavioral approaches to the study of politics. Although most of his work focused on the USA, he also launched comparative studies of European democracies particularly examining size effects and the role of oppositions. His later work was increasingly concerned with the effects of increasing socioeconomic inequalities on the quality of democracies.

Robert Alan Dahl (1915–2014) was a major theorist and researcher into democracy, who played a prominent role in American and World Political Studies throughout the second half of the twentieth century. Besides the usual book reviews and commentary throughout the relevant literature, Dahl's oeuvre has been complemented by a festschrift (Shapiro and Reeher, 1988) and two volumes of collected works (with bridging section introductions: Dahl, 1997a) and he has written or had recorded several autobiographical accounts (e.g. Dahl, 1997b, 2007).

Personal

Robert Dahl was born in 1915 in the small town of Inwood, Iowa to a family of Norwegian background (who in turn had come from North Dakota) and grew up in the even smaller Alaskan township of Skagway. His paternal grandfather was a farmer holding populist views and with a passionate interest in politics. His father was a medical doctor. His mother was of English-Scottish descent and quietly religious and more politically conservative. His family, which included brothers, was intellectual and engaged with public affairs – his parents are described as 'readers.' Although in his earlier years, the family was often very poor through its reliance on the wildly fluctuating incomes of his father's farming clientele, food (in lieu of money fees) was more plentiful. Within this family context, Dahl always felt he was destined for university education and he developed an interest in a career in politics or public administration.

This small town and remote background affected aspects of Dahl's subsequent work. From the age of 12, he worked on local docks and railroads which set him up for experiencing, and appreciating, the folk wisdom and hard working conditions of ordinary working people, leading to his later cautious advocacy of worker ownership and control. He also acquired a love of mountains and sea and an awareness of racial discrimination in relation to the local Tlingit tribe; despite some everyday links with young tribes-members he never wholly broke through caste barriers. His Norwegian descent also helped lead to the inclusion of small European democracies in his studies and attention to the problem of the effect of scale in the operation of polities.

In line with usual local experience he traveled to the University of Washington (BA in 1936) in Seattle to study. Although this was the closest major university, it still required a horrendous (5-day trip) journey. There, Dahl studied political science and economics as part of a broadly legal education. Once graduated, he applied for postgraduate fellowships, gaining one to Yale (PhD in 1940) – which again required long (but interesting) travel by rail.

Yale University then had a 'Government' department (later renamed 'Political Science' somewhat to Dahl's regret). It was then not as distinguished as Columbia, Harvard, Michigan, or Chicago. But the tiny department had some good staff and Dahl's studies were supplemented by courses in law school on business law and regulation. Political science as it was then conducted involved institutional description and analysis with an emphasis on constitutions, political parties, judiciaries, and executives. Its geographical focus was narrowly Eurocentric. These were all aspects of political science scholarship which Dahl helped to progressively open up during his career. Dahl's dissertation was on socialist programs and their limitations; as with many others of his generation, he had become caught up with the high aspirational horizons of the New Deal. As was usual then, his was a library dissertation. In between coursework and dissertation, he took up an internship in Washington DC. There he met his wife and also worked closely with more cosmopolitan East Coast Jewish socialists, joining the US socialist party.

Dahl later reported that, apart from immediate teachers, his main influences (because he had to struggle with their work) were the giants of past political science such as Plato, Rousseau, Marx, and Schumpeter. Schumpeter's famous 1942 book (*Capitalism, Socialism and Democracy*) provided a major challenge to study democracy in action, e.g., exploring how competition among elites offset the power of each. His views were contrary to those of famous Italian elite theorists Mosca and Pareto but were more in line with those of legal pluralist Leon Duguit and the earlier writings of Harold Laski. The latter were pluralists in their approach, emphasizing that power tended to be spread among elite groups rather than totally concentrated.

Having finished his dissertation, Dahl again returned to Washington DC to work on economic planning agencies. He found that the National Labor Relations Board was an embattled organization. Dahl carried out field investigations but found the response of some employers menacing. After this assignment, he worked in the Department of Agriculture with the role of making the bureaucracy more efficient. Then came a spell in the Office of Civilian Supply, the Office of Price Administration, and the War Production Board doing economic research. These tended to be very routine tasks. Still, his involvement in bureaucracy allowed valuable learning concerning the possibilities and limitations of central economic planning in a democratic system. Another shift was to be gradually converted out of his (still newly found) more idealistic socialism.

Although having a wife and two children, Dahl volunteered for a spell in the US army and served in the European sector for 3 years. He was the leader of a small reconnaissance platoon in an infantry regiment (44th Infantry Div; 71st infantry). This unit carried out night patrols and he earned the Bronze Star with oak cluster for this dangerous work. It was because he had a PhD that he was assigned to military intelligence and this scouting role. His unit fought in the Vogues mountains and after the cessation of hostilities, he was assigned to Paris and then to Frankfurt to de-Nazify German banking system. He did not think he succeeded! Dahl felt that the impact of the war was to increase his feelings of solidarity with working people and also to ignite a passion for the virtues of democracy as opposed to other political systems. It also deepened his thinking on what he wanted out of life: and he now definitely decided he was interested in scholarship rather than a political or administrative career.

On demobilization, Dahl returned to Yale University where the Department of Political Science began to be built up after a period of decline by its chair Jim Fesler. Dahl had sought out his earlier mentors at Yale and was offered a temporary teaching post to teach 'American government' which then became permanent. Indeed, the appointment lasted his whole career – becoming Sterling professor of political science and senior research scientist in sociology and emeritus on his retirement in 1986. The expansion and upgrading of the quality of the department was further pursued by Dahl who was departmental chair from 1957 to 1962. The Yale Department acquired a highly impressive grouping of political scientists including Robert Lane, Karl Deutsch, Nelson Polsby, and Harold Lasswell and was regarded as the top department when ratings began. However, Dahl regarded administrative service as a chore and his continuance at Yale was predicated on not having to take up the chair again. Although such a very long association with a single department might be unusual, Dahl felt his had been a good career choice.

He served as President of the New England Political Science Association in 1951 and of the American Political Science Association in 1966–67. He became a Fellow of the American Academy of Arts and Sciences (AAAS) in 1960, the National Academy of Sciences (NAS) in 1972, and the American Philosophical Society and spent a year at the Center for Advanced Study in the Behavioral Sciences at Stanford University in Palo Alto in 1955/56 and again in 1967. Dahl has received many honorary doctorates and has won many prizes for his work including the Woodrow Wilson Foundation Award (1962) for *Who Governs?* and the James Madison and the Benjamin Evans Lippincott Award (of AMSA) for his *Preface to Democratic Theory*. He was a Guggenheim Fellow twice (1950 and 1978), and the first recipient in 1995 of the Johan Skytte Prize in Political Science, an award given by Uppsala University in Sweden to the scholar who has made the most valuable contribution to political science. He was also a fellow of the Center for Advanced Study in Behavioral Sciences in 1955–56 and 1967. (More background is available at http://politicalscience.yale.edu/news/robert-dahl-sterling-professor-emeritus-political-science-passes-away#sthash.FFlMU035.dpuf.)

Dahl led a 1968 committee that successfully recommended that Yale become one of the first American universities to establish an undergraduate major in African-American culture. But another committee he headed in 1972 recommended that Yale's 4-year course of study be cut to 3 years and that majors be abolished was not successful.

Professor Dahl's first wife (formerly Mary Bartlett) died in 1970 and his second wife was the former Ann Sale. His family included step-daughters Ms Connor and Jane Thery, daughter, Kirsten Dahl; and sons, Peter (deceased), Eric, and Christopher; and four grandchildren.

Work

Dahl's lifelong work has contributed to the academic study areas of democratic theory, American politics, and comparative politics. He has focused on the origins, characteristics, and consequences of democracy, examining democracy from normative, theoretical, and empirical angles. He is often seen as part of a 'Pluralist' school of political thought which was particularly prominent in the Yale Department. In addition, Dahl was a leader of the 'Behavioral revolution' which promoted more rigorous methods and scientific study, and was therefore open to quantitative analysis and survey-based field work. He also contributed theoretical exegeses on the definition of power. Along with many social scientists of his generation whose experience was formed by the New Deal and WW2, his broad moral ethos was to value democracy and other modern achievements but to also recognize their limits and to seek and urge constructive reform.

His life work constitutes a very largely coherent ongoing focus on various aspects of democracy and related issues.

Arguably, he moved somewhat leftward in later life, although this normative direction was also present in his earliest years. Some commentators have argued, though, that there have been two stages in his work: segueing from democrat to democratic socialist, although Dahl was uncomfortable with over-simplifying labels. In contrast to earlier work which emphasized more what was right with democracy, his more recent work encompassed attention to not only why democracy is not working well but also involved a search for alternative solutions. His passion for democracy was accompanied by rather more instrumental views on different forms of economic organization.

Dahl's scholarly style was to be problem orientated within a broad framework. He was as much concerned with raising questions and setting a research agenda as in necessarily answering questions, as his use of the term 'preface' in several important books attests. A self-description was as follows: "… [S]ome question with theoretical and empirical relevance and often for moral importance as well would present itself to me. A piece of work might follow, during which I might reshape the initial formulation of the problem or question." And then questions would be taken up and links would emerge to eventually become a political theory. Because his work was problem orientated no particular approach or methods were preferred, although he always made sure that careful work was advanced. Dahl is often praised for his compact, clear succinct writing. He seldom paused to answer criticism because he preferred to be moving on to his next research project. He wrote some two dozen books (almost all published by Yale University Press) and many (around a hundred) articles.

His influence was particularly promulgated from 1963 onward through the publication of his slim text *Modern Political Analysis* which was reworked through six editions, with the sixth edition produced in 2003 (coauthored with Bruce Stinebrickner). Dahl was a central figure in US and world political science over a long period, with citation count measures of his influence placing him among the top very few political scientists.

Dahl was undoubtedly strongly supported in his work by the microcontext of the Yale Department which apart from its acquisition of many of the leading American political scientists of the time was an intellectually supportive setting. He taught at both undergraduate and more often at postgraduate levels and while very supportive of postgraduate students tended to be relaxed and informal in his teaching style. His work in comparative politics was also strongly supported by his embedding with an enduring collegial friendship network with such authorities as Stein Rokkan (Norway), Hans Daalder (Netherlands), and Val Lowin (American historian of France/Belgium).

Themes that had been raised in his Dissertation were often further developed later on in his intellectual career and included key points such as

- statement of explicit criteria for assessing a democratic political system;
- seeing Marxism (including Leninism) as unsatisfactory;
- being charmed by the potential of small-scale associational and cooperative systems (although these were seen as possibly utopian); and
- stressing the central importance of markets and market socialism in relation to democracy.

Two early studies explored domestic constraints on, and opportunities for, policy formation – *Congress and Foreign Policy* in 1950 and *Domestic Control of Atomic Energy* in 1951. These studies required fieldwork, interviewing officials in Washington DC.

Further studies in the 1950s established his reputation. *Politics, Economics and Welfare* (published in 1952) grew out of contact with fellow-author and fellow-Yale academic Charles Lindblom who was teaching labor economics. They endeavored to restore intellectual analyses in the area of political economy which had tended to lapse as new disciplinary boundaries developed in the previous century. This work first advanced analysis of 'polyarchy' (a term retrieved from a more arcane usage) which endeavored to represent the more attainable approximations to ideal democracy which are possible in large-scale nation-states. A Guggenheim fellowship supported the writing of the book which was a 'sleeper,' taking some time for its influence to emerge.

The book identified the seven conditions that characterize polyarchies:

- control over government policies by elected officials;
- frequent election of those officials without coercion;
- (almost) Universal Right to vote for those officials;
- right of free expression on matters relevant to such elections;
- right to seek alternative sources of information; and
- right to form relatively independent associations/organizations (including political parties and interest groups).

Dahl further developed his general concern with the conceptualization of democracy with his specific argument that it involves the conciliation of numerous minorities and constraints on elite power in his *Preface to Democratic Theory* (1956). Dahl focused on procedural aspects of democracy as opposed to substantive outcomes. Two crucial dimensions are contestation and participation. The book emerged out of a seminar he taught and was then developed further through an invited University of Chicago Walgreen lecture before (after further writing) being converted into the book. Procedural democracy has the ideals of:

- effective participation (adequate and equal opportunities for expressing preferences);
- enlightened understanding (adequate and equal opportunities for arriving at preferences on matters);
- final control of the agenda; and
- inclusion (i.e., widespread sharing of political rights).

In the late 1950s, Dahl participated alongside Paul Lazarsfeld and Mason Haire in the Ford Foundation Program in Economic Development and Administration which endeavored to point up research areas related to problems of business and industry (see Dahl et al., 1959).

His period at Palo Alto exposed Dahl to mathematics and statistical work, including game theory, through other fellows also resident there. This led to his cautious adoption (or at least recommendation) of some of these more formal approaches,

which clearly influenced his 'behavioral' approach to political science research.

After this experience, Dahl felt some need to become more empirically grounded and hence launched a field study of New Haven (the host city for Yale University) published in 1961. Dahl's study was cast as an attempted empirical refutation of the famous study by Floyd Hunter (1953: Community Power Structure) which postulated that communities were controlled by a more or less cohesive and all-powerful elite. Hunter argued that an economic elite dominated the political understructure, composed of civic organizations, trade unions, and political parties. Important policies originated first in the elite to its own benefit; then they were transmitted through the understructure for public approval. The power structure in a community, therefore, is like a pyramid with the economic elite on the top, the political understructure in the middle, and ordinary citizens at the bottom. Hunter's methodology concentrated on a self-selected small group of the most powerful local leaders and so was termed the 'reputational approach.' Dahl's methodology, instead, was to study those who were involved in exerting power on an array of explicit political decisions. He also developed a conceptual critique of elite theorists and the debate rolled on for several decades and extended right across the social sciences (drawing in, for example, C. Wright Mills). Dahl was later to admit that his study was insufficiently sensitive to the limits which external structures imposed on local decision makers.

An academic leave in Rome in 1962–63 inspired Dahl to broaden the geographical scope of his work to encompass small democracies and he also investigated the role and impact of oppositions across a variety of democratic and nondemocratic political systems. This work encompassed working with a network of local experts (*Regimes and Oppositions*, 1973; *Political Oppositions in Western Democracies*, 1966). A further study examined the operations and tradeoffs between efficiency and participation between different sizes of jurisdictions, under the guiding idea that there are significant size thresholds which affect how politics works (*Size and Democracy*, 1973). Following the 1968 student rebellions, Dahl published *After the Revolution* (1970).

In the 1980s, Dahl reworked again the aspects of his theoretical work on democracy with three books (1982, 1985, 1989) on gaps between ideals and actual democracy, and the wider challenges of establishing democratic conditions despite increasing economic inequalities. Dahl became increasingly concerned by the problem of the excessive political power of corporate wealth and the interest in and ability of the wealthy in buying political influence: converting some of their economic capita into political capital. *A Preface to Economic Democracy* (published in 1985) addressed the way in which concentrated wealth can impact the liberal democratic system. He also returned to study policy related to nuclear weaponry. *Democracy and Its Critics* (1989) can be seen as the culmination of Dahl's main work in which he summed up the normative arguments for 'procedural democracy,' explored its limitations, and assessed the empirical conditions for the flourishing of polyarchy. He argues that there has been a 'third transformation' of democracy deepening democratic deliberation and extending it, e.g., to other institutions corporations, unions, and to intentional institutions and governing bodies. Also included is the deployment of new technology. In his *A Preface to Economic Democracy*, Dahl endeavored to show how his model of democracy might be extended to other institutions.

Late in his life, Dahl wrote a summary trilogy of three brief books *On Democracy* (1998), *How Democratic Is the American Constitution?* (2002), and *On Political Equality* (2006a) which could be seen as endeavoring to sum up his legacy. Their message is that democracy is precious but fragile. Having long emphasized defenses of democracy, he turns more to criticism in these more recent works. *On Democracy* builds his strong case for democracy with the criticism following in the further two volumes. His book on the US constitution questions its democratic credentials, while the final book shows concern that US democracy has fallen below an acceptable standard, against a horizon of people's rising expectations and their political disenchantment. Issues facing democracies needing attention include 'globalization,' the rapid development of international organizations, and terrorism. A concern is that the American Constitution is too rigid to allow modernization and the necessary improvement to deal with the mounting difficulties.

In *On Democracy*, Dahl makes the case for democracy generally (i.e., cross-nationally) as it produces demonstrable desirable consequences:

- prevents cruel and vicious rule;
- guarantees citizens certain fundamental rights that undemocratic systems cannot grant;
- ensures its citizens a range of personal freedoms;
- enables them to protect their fundamental interests;
- provides them the opportunity to exercise self-determination and to live under laws of their own choosing;
- gives them the opportunity to exercise moral responsibility;
- fosters human development;
- fosters a relatively high degree of political equality;
- discourages war; and
- encourages prosperity.

These beneficial consequences are good for people because democracy provides the conditions for them to be able to live good, decent, and moral lives. Moreover, it is good for providing fairness and efficiency in other organizations and institutions.

In *How Democratic Is the American Constitution?*, Dahl specifically reviews the American democracy's performance. The US is focused on because it is the oldest, biggest, most powerful, and most self-conscious 'model' modern democracy and is also the one which is most respected and often despised. Dahl argues that, unfortunately, the US Constitution is seen as the repository of eternal political truths and instead Americans need to change the way they think about it. It is too old and was written with no experience of anything resembling modern democracy (e.g., with political parties and a modern press). Judged by today's standards of democracy, Dahl finds at least four serious defects:

- the overwhelming power of the office of the president and the way presidents are elected;
- the inequality of representation, in particular in the Senate;

- the role of the Supreme Court which takes it beyond judicial review to 'judicial legislation;' and
- the dictatorship of history in the near to impossibility of changing and modernizing the Constitution and bringing American democracy up to date with changing expectations.

In *Political Equality*, Dahl turns to the limitations on political equality imposed by economic inequality. He notes that despite attempts by some leaderships to excoriate democracy for their country, everywhere people want democracy and will vote or choose it where possibilities open up – even where there are other societal goals (e.g., development or efficient delivery of government services) which are seen as important.

Cognate More General Analyses

Dahl contributed to further political science discussions closely aligned to his focus on 'democracy.' In the mid-1950s, he contributed to conceptual analysis of the central concept of 'power.' He noted that systematic treatments up to that point were rare and incomplete. Nevertheless he was able to draw on important writers including Hobbes, Weber, and Lasswell. His definition at that time was standard: "A has power over B to the extent that he can get B to do something that B would not otherwise do." In later reflections, Dahl indicated he was dismayed at the continued sloppy use of the term and at the poverty of systematic empirical studies on power which were both conceptually and methodologically sophisticated.

His final version (e.g., 2003: 17ff) involves the definition that influence is

> … a relation among human actors such that the wants, desires, preferences, or intentions of one or more actors affect the actor's actions, or predispositions to act, of one or more actors in a direction consistent with – and not contrary to – the wants, preferences, or intentions of the influence-wielders.

Dahl points out that this definition (2003: 17)

> (1) clearly incorporates the idea of influence-as-causation; (2) excludes relationships between human beings and nature, animals and the like; and (3) avoids the concept of interest and the difficulties inherent in it. Moreover, (4) our definition does *not* exclude trivial or, in the general scheme of things, unimportant instances of influence.

… while recognizing that political scientists prefer to be interested in influence in more formally political settings. Finally, it distinguishes between positive and negative influences, excluding the latter.

Dahl indicates that some further conceptual distinctions can facilitate comparisons of the amount of influence of actors:

- distribution (e.g., extent of political inequality);
- gradations (shape of the distribution); and
- scope and domain of influence (issues over which influence might be wielded).

Differences in amount of influence is explained in terms of three key factors which actors differentially have (2003: 34ff):

- resources/endowments (e.g., money, information, food, ability to make credible threats, jobs, friendship, social standing, ability to make and enforce laws, votes);
- skills (variations in efficiency of using resources to influence); and
- motivations (extent to which actors commit to resources politically).

Finally, some seven forms of influence are identified (in part organized as pairs):

- inducement/power (the influencer can offer something or has the power to deprive the influenced);
- force/coercion (removing the influenced choice – or credibly threatening to);
- persuasion/manipulation (correct or incorrect supply of persuasive information); and
- authority (legitimate orders).

From the 1950s, Dahl was a leader of the 'behavioral revolution' in political science – that is, the use of a sociological approach in the empirical investigation of politics as it happens in reality as opposed to descriptions of ideals set by constitutions or similar institutional analyses. This approach sometimes flirted with formal and mathematical models (e.g., rational choice, games models), and indeed one of the major subsequent developments in political science has been 'rational choice' orientated, although Dahl was more cautious as he wished to retain a central role in political science for examining the influences of wider systems and of historical contexts. Nevertheless, the behavioral revolution required clear conceptualization, careful attention to definitions of key terms, consideration of ways of empirical measurement, and the deployment of appropriate data (often quantitative, such as those derived from surveys) which reported on the views and behaviors of (individual and collective) actors.

Personal experiences such as his neighboring relationship with philosopher Carl Hempel and his time at Palo Alto encouraged this approach but institutionally it was particularly driven through the SSRC committee on Political Behavior on which Dahl served.

Conclusions

Dahl's impact was part of a broad pluralist school whose influence became more limited as political conditions changed after the 1970s (after the era of campus unrest, etc.) and as newer more specialized approaches developed. Arguably pluralism morphed into a broad 'political process' model. More formal rational choice models became a central thrust which leached much for the historical and broader societal context out of analyses. But Dahl's influence remains strong.

In Dahl's considerations of how democracy should work, the distribution of power is central. An underlying assumption is that each person's political voice should have the same weight and that each needs to represent themselves rather than be 'represented' by spokespeople or experts. However, energetic full political participation by everyone is not expected and

might exhaust the political system's capacity. In his view, a somewhat self-selected stratum of 'political man' invests considerable resources in order to effect things and to be involved in the array of associational activity central to the operation of politics. Appropriately designed and maintained institutions will provide facilitative environments for good decision making.

To sum up Dahl's work, he argued that

- democracy is better than nondemocracy both in better delivering fairness and efficiency and fundamentally because it is better for the people living within it;
- established democracies, despite being held out as models, are fragile insufficient and delivering less than they should largely because of unnecessary rigidities of their institutions inherited from the past; and
- democracy is severely challenged by recent trends in capitalism, which have undermined its effective operation and which need to be checked in order to improve the quality of democracy.

Characteristically as a scholar committed to advocating change where it seemed needed Dahl's suggested reforms include

- control of political funding with limit and regulate private contributions to political parties and campaigns;
- social reforms to counter the drift toward ever more economic inequality, mainly in the form of antipoverty policies; and
- in general, nearness and smallness in social and political life to fill the space between citizens and representatives and to keep citizenship control intact.

In this way, Dahl's life project, after adding much complexity and additions, largely returned to its roots when his scholarly work began.

See also: Bounded Rationality; Deliberation and Democracy; Democratic Theory; Hempel, Carl Gustav (1905–97); Majoritarianism and Majority Rule; Party Identification and Nonpartisanship; Party Systems; Pluralism, Political Science and Polling; Politics and Participation; Public Bureaucracies; Rokkan, Stein (1921–79); Social Democracy; Socialism; Voting and Direct Democracy.

Bibliography

Dahl, R.A., 1950. Congress and Foreign Policy. Harcourt, Brace, New York.
Dahl, R.A., 1956. A Preface to Democratic Theory. University of Chicago Press, Chicago.
Dahl, Robert A., Haire, Mason, Lazarsfeld, Paul F., 1959. Social Science Research on Business; Product and Potential. Columbia University Press, New York.
Dahl, R.A., 1961. Who Governs? Democracy and Power in an American City. Yale University Press, New Haven.
Dahl, R.A., 1963. Modern Political Analysis. Prentice-Hall, Englewood Cliffs, NJ. Subsequent editions up to 2003.
Dahl, R.A. (Ed.), 1966. Political Oppositions in Western Democracies. Yale University Press, New Haven.
Dahl, R.A., 1967. Pluralist Democracy in the United States: Conflict and Consent. Rand McNally, Chicago.
Dahl, R.A., 1970. After the Revolution? Authority in a Good Society. Yale University Press, New Haven.
Dahl, R.A., 1971. Polyarchy: Participation and Opposition. Yale University Press, New Haven.
Dahl, R.A. (Ed.), 1973. Regimes and Oppositions. Yale University Press, New Haven.
Dahl, R.A., Tufte, E.R., 1973. Size and Democracy. Stanford University Press, Stanford, CA.
Dahl, R.A., Lindblom, C.E., 1976. Politics, Economics, and Welfare: Planning and Politico-economic Systems Resolved into Basic Social Processes. University of Chicago Press, Chicago.
Dahl, R.A., 1976. Democracy in the United States: Promise and Performance, several ed. Houghton Mifflin Co, Boston.
Dahl, R.A., 1982. Dilemmas of Pluralist Democracy: Autonomy vs. Control. Yale University Press, New Haven.
Dahl, R.A., 1985. A Preface to Economic Democracy. University of California Press, Berkeley.
Dahl, R.A., 1986. Democracy, Liberty, and Equality. Norwegian University Press, Oslo, Oxford; New York.
Dahl, R.A., 1989. Democracy and Its Critics. Yale University Press, New Haven.
Dahl, R.A., 1990. After the Revolution?: Authority in a Good Society, rev. ed. Yale University Press, New Haven.
Dahl, R.A., Abrams, R.M., 1994. The New American Political (Dis)order: An Essay. Institute of Governmental Studies Press, Berkeley.
Dahl, R.A., 1997a. A brief intellectual autobiography. In: Daadler, Hans (Ed.), Comparative European Politics: The Story of a Profession, pp. 68–78.
Dahl, R.A. (Ed.), 1997b. Toward a Democratic Journey: Reflections: 1940–1997, 2 vols. Institute of Government Studies Press, University of California, Berkeley.
Dahl, R.A., 1998. On Democracy. Yale University Press, New Haven.
Dahl, R.A., 2001. How Democratic Is the American Constitution? Yale University Press, New Haven.
Dahl, R.A., Shapiro, I., Cheibub, J.A. (Eds.), 2003. The Democracy Sourcebook. MIT Press, Cambridge, MA.
Dahl, R.A., 2005. After the Gold Rush: Growing Up in Skagway. Xlibris Corp.
Dahl, R.A., 2006a. On Political Equality. Yale University Press, New Haven.
Dahl, R.A., 2006b. A Preface to Democratic Theory, expanded ed. University of Chicago Press, Chicago.
Dahl, R.A., 2007. Normative theory, empirical research and democracy. In: Gerardo, Muck, Snyder, Richard (Eds.), Passion Craft and Method in Comparative Politics. Johns Hopkins University Press, Baltimore, pp. 113–149.
Hunter, F., 1953. Community Power Structure: A Study of Decision Makers. University of North Carolina Press, Chapel Hill.
Lerner, D., Dahl, R.A., 1965. Cause and Effect: The Hayden Colloquium on Scientific Method and Concept. Free Press, New York.
Levi, Margaret, 2009. A conversation with Robert Dahl. Annual Review of Political Science 12. Video of interview at: http://www.deepdyve.com/lp/annual-reviews/a-conversation-with-robert-a-dahl-0jO40gdoSS.
Neubauer, D.E., Dahl, R.A., 1974. Readings in Modern Political Analysis, second ed. Prentice-Hall, Englewood Cliffs, NJ.
Shapiro, Ian, Reeher, Grant (Eds.), 1988. Power, Inequality and Democratic Politics: Essays in Honor of Robert A Dahl. Westview Press, Boulder & London.

Dahrendorf, Ralf (1929–2009)

Hans-Peter Müller, Humboldt-University Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ralf Dahrendorf (1929–2009) was a German-British sociologist, who pondered the question of freedom in modern society. He made a remarkable career in academia, first in Germany (Hamburg, Tübingen, and Konstanz), then in England (Director of the London School of Economics and warden of St Antony's College, Oxford University). Lord Dahrendorf served in the House of Lords and made numerous contributions as a journalist and public intellectual. He developed a sociology of conflict, of life chances, and of civil society. He created a particular concept of the public intellectual, the 'Erasmians,' which resisted the temptations of totalitarianism.

Ralf Dahrendorf was a German-British sociologist preoccupied with the problem of liberty: How is freedom in a modern society possible? His approach to this question was both analytical and empirical, as well as political and normative. In this sense, then, Dahrendorf moved between philosophy and the social sciences. His relentless activism reached well beyond the confines of academia. Rather, as an engaged citizen, critical liberal, and public intellectual he pursued a career in politics and journalism alike. The core of his social and political thought consisted of a sociology of conflict, of life chances, and of civil society. His liberal mind welcomed the pluralism of ideas and ideals, of values and norms as the essential infrastructure of modern society. But this pluralism comes along with different and sometimes antagonistic interests and power differentials, which need rules and regulations. Every society has to be understood as an association of power and domination, necessary to mediate conflicts and the clash of interests. Modern society creates a plethora of life chances for its citizens as a fundamental prerequisite for a liberal conduct of life. Analytically speaking, life chances consist of options and ligatures. Options open up chances of and for free choice. Ligatures help to integrate the individual into this liberal, pluralistic, and often antagonistic order of modern society. Beyond market and state, it is the livelihood of a civil society whose citizens are actively engaged in public affairs, which is the backdrop for liberty.

Life and Career

Ralf Dahrendorf was born on 1 May 1929 in Hamburg. His father, Gustav Dahrendorf made an upward political career as a social democrat, which earned him a parliamentary seat in the 'Reichstag' in Berlin during the years of 1932 and 1933. With the advent of the Nazi regime, he lost his seat and was incarcerated as a member of the German Resistance Movement in 1944. The son, too, began his struggle against the regime by distributing critical leaflets and was taken into custody by the Gestapo. In his autobiography, Dahrendorf (2002) described this experience of solitary confinement as a turning point in his life. His thirst for freedom and the return to a decent society after the cruelty of the Nazi regime determined his course of life.

After the War, he studied philosophy and philology in Hamburg, wrote a dissertation on the concept of justice in Marx's thinking, but soon turned to England for the completion of his academic training. As a postgraduate at the London School of Economics (LSE) in 1952–53, he pursued a PhD under the supervision of T.H. Marshall on "Unskilled Labour in British Industry" (1953). He met Karl Popper, who informed his epistemological thinking in favor of 'critical rationalism,' which helped to transcend both hermeneutics and dialectics in the wake of Kant and Max Weber. Back in Germany, he soon completed his 'habilitation' on *Class and Class Conflict in Industrial Society* (1959) in 1957. In the academic year 1957–58, he followed an invitation of the *Center for Advanced Study in the Behavioral Sciences* in Stanford, where he met fellows like the economist Milton Friedman, the sociologist Talcott Parsons, and the historian Fritz Stern. Upon return to Germany, he started his career as a professor of sociology, first in Hamburg, then in Tübingen, and in Konstanz. In the late 1950s and early 1960s, he published a number of influential books: *Homo Sociologicus* (2010 (1958)), *Über den Ursprung der Ungleichheit unter den Menschen* (On the origin of inequality among men, 1974 (1961)), *Conflict and Liberty* (1962), *Bildung ist Bürgerrecht* (Education is a civil right, 1965), *Society and Democracy in Germany* (1979a (1967)), *Essays in the Theory of Society* (1968), and so on.

From 1968 to 1974, he switched from academia into politics and pursued a political career in the German liberal party, which moved him from a seat in the 'Länder'-parliament in Baden-Württemberg to the German Bundestag and then on to the European commission in Brussels. He returned to academia and served as the director of the LSE from 1974 to 1984. In 1982, he was made a Knight Commander of the Order of the British Empire. From 1987 to 1997, he became the warden of St Antony's College at Oxford University. The most remarkable publications of this middle period were *Life Chances* (1979b), *The Modern Social Conflict* (2007b (1988)), and *Reflections on the Revolution in Europe* (2004 (1990)). Since 1993, Lord Dahrendorf served as a member of the 'House of Lords.' In 1995, he published his history of the LSE – his favorite institution, where he had studied and which he directed for 10 years – and in 2002, his short autobiography *Über Grenzen* (On boundaries). In his last years, he was showered with honorary degrees and prizes from all over the world. When he

died shortly after his 80th birthday on 17 June 2009, the obituaries talked about a sociological classic, a gifted politician, and an influential intellectual we had lost, the manifolded careers he had made, and the multiple boundaries he had crossed in the course of his life.

The Sociology of Conflict I

In *Class and Class Conflict in Industrial Society* (1959), Dahrendorf set out to develop his theory of conflict. He did so by a critical reexamination of Marx's class theory. Why Karl Marx? During the Nazi era, it was the organic concept of 'Volksgemeinschaft' (*literally* 'folk community'), which was said to hold together a modern society like Germany. Since the 'consensus' of the 'organic whole' was, above all, what mattered, every dissensus was persecuted as resistance against the regime. The attempt to conceive of modern society as a close-knit community, therefore, irresistibly led into oppressive totalitarianism. Marx, however, who followed a concept of modern society as an 'antagonistic whole,' had fully understood that the different groupings in society named 'classes' pursue divergent, if not antagonistic interests, which necessarily tend to clash in social life. Social life means 'eternal struggle' over who rules and who gets ruled. Societies under industrial capitalism envisage the confrontation of 'bourgeoisie' and 'proletariat' as the two opposing classes.

Dahrendorf shared Marx's agonistic view of social life, his focus on conflict, and his interest in class formation and class struggle. But he took Marx mainly as a starting point and tried to 'modernize' his theory of conflict. Dahrendorf mobilized the entire corpus of critique which had been brought forward against Marx from Max Weber and Joseph A. Schumpeter to Theodor Geiger. And he used the contemporary approaches like structural functionalism and conflict sociology to conceive of an adequate understanding of conflict in modern liberal and pluralistic societies.

Above all, he criticized the one-sidedness of the economics of class. The means of production are by no means the only basis of class formation. Property is one, but only one source of class formation and domination. Classes are not confined to the economy and the industrial realm, but may evolve in state and society alike. Dahrendorf uses 'class' in an analytical sense and sees it in an intimate relationship to domination. "Social classes and class conflict are present wherever authority is distributed unequally over social positions" (Dahrendorf, 1959: 247). In his eyes, class struggle is an expression of power conflicts. Since modern society is differentiated into a number of plural fields, the division of rulers and ruled can be found in every societal realm. Classes are not confined to the historical configuration of 'bourgeoisie' and 'proletariat,' but emerge wherever power and interest struggles between rulers and ruled take place. Apart from short periods of 'stasis' and stability, which a successful moderation and mediation of interests and conflicts entails, social life evolves by the dynamics of struggles over the distribution of resources and power positions. Society in that sense is in constant flux and social change the overarching feature of modern life. In Dahrendorf's (1959: ix) eyes, "'Classes' are interest groupings emerging from certain structural conditions which operate as such and effect structure changes."

Dahrendorf (1959: xii) sums up his achievement as follows: (1) the retention of the term 'class' for a theory that dispenses with the historical reference to antagonisms between 'bourgeoisie' and 'proletariat'; (2) the 'definition' of the crucial concepts of power and authority; (3) the separation and, in part, the confrontation of industrial and political class conflict; (4) the application of conflict theory to the analysis of contemporary society, especially to its 'ruling classes'; (5) the absence of grounding of this analysis on 'large-scale inquiries of a more "practical" kind.'

His first big and, maybe, best book *Class and Class Conflict in Industrial Society*, which earned him immediate fame, mirrors a discourse centered around three figures: Marx, Parsons, and Weber. Marx is crucial for Dahrendorf, since he underlines the agonistic and changeable character of social reality. In contrast to Parsons, who answered the Hobbesian question of social order with the institutionalization and internalization of values and norms, Dahrendorf puts forward the trias of power, conflict, and change. Not values and norms but power and authority, not consensus but conflict, not stability but change is what characterizes modern society. Although he dubs diplomatically his conflict theory as the necessary complement to the position of stability favored by the structural functionalism of Parsons, he takes boldly the counterposition to the American sociologist. This analytical conflict came to be known as the schism between consensus and conflict theory in the history of sociological theory building. The bridge for the confrontation between Marx ('conflict') and Parsons ('consensus'), however, was Max Weber. The main thrust of Dahrendorf's theory of conflict is Weberian: The agonistic view of social life as struggle, the separation between power and authority, the intimate relationship between class and domination, the conceptualization of classes as 'status groups' of a various kind, the ubiquity of conflicts which call for rules and regulations as means of their moderation and the exceptionality of the 'big conflict' called revolution as well as conflicts as the driving mechanism of social change – all of these arguments can be found in *Economy and Society*.

This Marxian (not Marxist)/Weberian mix was highly welcome since the 1960s among scholars who were neither Marxist nor functionalist. In Britain, sociologists like David Lockwood, John Rex, David Bottomore, John Goldthorpe, and Anthony Giddens followed this legacy by combining the sociology of conflict with race and class. In Germany, it was M. Rainer Lepsius, Walter Müller, and Karl Ulrich Mayer, who pursued the analysis of social structure and social inequality with political sociology. In France, it was above all Pierre Bourdieu, who developed an approach of class and styles of life.

The Sociology of Conflict II

Whereas *Class and Class Conflict in Industrial Society* pertained to macrosociology, Dahrendorf's famous *Homo sociologicus* referred to the microlevel of role analysis. His second classic again was a confrontation with structural-functionalism because role theory became a fashionable approach

particularly able to demonstrate the *systemic*, i.e., societal character of the individual conduct of life. Dahrendorf noted that economics created the concept of 'homo oeconomicus' and psychology the concept of 'psychological man' for their theoretical models, whereas structural-functional systems theory put forward the notion of a 'homo sociologicus.' According to him, this 'sociological man' is nothing but his role-repertoire. The category of role, a concept lent from the world of theater, seems ideal to pinpoint to the crosscutting of individual and society. Where individual and society meet, a role is created.

Interestingly enough and ironic though, Dahrendorf gave such a convincing exposition of this approach that the reception of his book in Germany took Dahrendorf as an ardent proponent of role theory. The contrary was the case. *Homo sociologicus* was written as a serious warning against the totalitarian underpinnings of this approach. If a human being is understood exclusively in terms of his or her role-complex, the dimensions of freedom and individuality get lost. Dahrendorf followed Simmel in that human beings are doubly constituted – as an individual being and as a social being. In Simmel's eyes, we stand with one foot in society, with the other outside of its realm. Otherwise, freedom and individuality would not be possible. Dahrendorf conceptualized the pragmatic paradox of a double human being and showed that role theory captured only one side of this duality: the social one. He demonstrated this one-sidedness in his paradigmatic example of 'Lehrer Dr Hans Schmidt' ('teacher Dr Jack Smith'). Jack Smith plays many parts: man, father, husband, academic, teacher, and so on. In his behavior, he tries to 'function' well because he fulfills all of these roles as best as possible. Roles are conceptualized as expectations with different degrees of bindingness: 'Must' expectations, which are legally binding; 'Should' expectations, which are morally binding; and 'Can' expectations, which are a courtesy. Socially formed roles seem to define a burden on the shoulder of the individual who longs to be free. Dahrendorf talked about 'die ärgerliche Tatsache der Gesellschaft' – 'the annoying fact of society' in order to mark this burden. Though roles conceptualize the crosscutting point of individual and society, there seems to prevail an antagonistic view of individual and society: Society spells constraint, individuality promises freedom. It is in the phrasing of Isaiah Berlin, negative freedom or the 'freedom *of*' society, which both liberal thinkers seem to articulate. Yet, as we will see, this image was corrected by later writings of Dahrendorf in favor of conceptualizations of 'freedom *in*' society.

Since Dahrendorf was influenced by Simmel in his understanding of the interaction between individual and society, he made out conflicting role-obligations as a potential avenue for the liberty of the person. Dr Smith might deliberately miss out on some of his family obligations by overstressing his professional role as a teacher. In this sense, the diligent management of conflicting role-obligations opens up a space for freedom. In Simmel's eyes, this mechanism, the 'crosscutting of social circles' with the result of a conflicting role-repertoire, formed the seedbed of individual freedom. But already Simmel conceded that this liberty amounts just to a 'freedom of' without specifying the 'freedom for' a particular style of life. Yet, again, in this microsociological setting of his sociology of conflict, Dahrendorf stressed the role of freedom and the place of and for individuality in a modern liberal society. He ardently defended the modern person against the encroaching impositions of an ever more demanding society which wants to subdue its members.

The Sociology of Conflict III

Dahrendorf's attempts to develop a theory of conflict culminated in the 1980s in his analysis of the *The Modern Social Conflict* (2007b (1988)). This article presents a synthesis of his thought in that it brings together all the elements of his loosely knit middle range theory of society: his concept of conflict with the twin concepts of power and liberty; his idea of life chances as options and ligatures; his notion of citizenship and civil society beyond state and market; the history of social conflict and before the background of this history the new quality of conflict today. He overhauls his conceptual tool kit and tells the tale of two conflicts and their contemporary 'Gestalt.'

In an earlier book on *Life Chances*, Dahrendorf (1979b) had set out to define them by the twin notion of options and ligatures. "Options are the choices which people have open to them; ligatures are the coordinates which give these choices meaning. Clearly, ligatures are the more complicated notion. I had in mind those deep cultural ties which help people find their bearings in the world" (Dahrendorf, 2007b: 17). In his new book, he continues to talk of life chances, but defines them differently. "Life chances are a product of entitlements and provisions" (Dahrendorf, 2007b: 16). The state gives citizenship rights as entitlements; the market produces goods and services as provisions. A free and liberal modern society needs them both. "The distinctions between politics and economics, chances of participation and standards of living, entry tickets and the choices which await the entrants – in short, between entitlements and provisions – are useful. But despite appearances to the contrary, there is no trade-off relationship between the two" (Dahrendorf, 2007b: 18). Dahrendorf refers to T.H. Marshall's characterization of the evolution of citizenship rights in order to ground his notions in the historical development. Marshall distinguished legal, political, and social rights and used the British case to ascertain that legal rights were won in the eighteenth, political rights in the nineteenth, and social rights in the twentieth century. The twin revolutions – the Industrial Revolution in England (provisions) and the French Revolution (entitlements) – and Marshall's trias of rights both provide the frame for Dahrendorf's reflection on the two thresholds of social change. The class society that Marx described in the nineteenth century was overcome by a society in the second half of the twentieth century, which created many more life chances for all and due to collective actors who took care of the interests of the majority allowed social mobility in a relatively open structure of stratification. Yet, in the 1970s, the social democratic consensus that had informed this model of society of the first threshold came to an end. 'Death by success' is the formula for this process. The world T.H. Marshall envisaged in the 1950s had become true. Ideas that approach their fulfillment lose their momentum and motivating appeal. They are bound to die out. "The rule of the majority class and of the eminently reasonable consensus of social democracy turned out to be unstable" (Dahrendorf, 2007b: 117).

This is not to claim that citizenship rights and the civil society were truly realized. Even the best of all possible worlds was far from being perfect. There were still entitlement barriers or deficient provisions for the lower end of the hierarchy of society.

Yet, societies in the 1980s had crossed the second threshold and faced a new modern social conflict. It turned out to be the gap between the 'majority class,' which comprises two-thirds or three-quarters of a society and a minority or 'underclass' forming the rest. The majority class struggles hard to keep its privileged place. In fact, it employs strategies of closure to keep the mobile parts of the underclass at bay and erects new boundaries against migrants from outside. Closure and exclusion, however, are not an adequate answer of an open society to migration and ethnic diversity. Barriers to entitlement and provisions undermine the growth of life chances for all – the main thrust of a modern liberal and democratic society in Dahrendorf's eyes. Yet, one reason why the strategies of closure and exclusion are successful is the fact that the 'minority class' has lost its power and energy to fight for inclusion. Dahrendorf is taken by surprise about the lack of struggle and resistance: "Why do the long-term unemployed and the persistently poor not join forces and march on their capital cities to demand their full share of citizenship? Why is there not at least an unemployment party or a poverty party? And if this asking too much, why does the underclass not go on a rampage to break up the furniture of the house which the majority class built for itself?" (Dahrendorf, 2007b: 160). The 'Lumpenproletariat' in Marx's times and the underclass today share the same conditions. They are in a state of deep anomie in the Durkheimian sense and a state of alienation in the Marxian sense. "The crucial fact about the underclass and the persistently unemployed is that they have no stake in society. In a very serious sense, society does not need them. Many in the majority class wish that they would simply go away; and if they did their absence would barely be noticed. Those who are in this position know it. Society to them is above all distant. It is symbolized by nothing so much as the police and the courts, and to a lesser extent by the offices and officials of the state" (Dahrendorf, 2007b: 161).

This split between the majority class – furnished with a full set of life chances and firmly established 'inside' of society – and the minority class – confronted with barriers to basic entitlements and provisions and situated 'outside' of society – is the major driving force of anomie and alienation on the one hand, the reason for a lack of ligatures on the other hand. A society that leaves considerable parts of its population behind with no stake in its fabrique will fail to develop a sense of cohesion and belonging, of inclusion and integration.

In a truly liberal vein, Dahrendorf proposes a new social contract which should tackle the central problems of contemporary society. In order to avoid the formation of an underclass in the societies of the OECD world, he is in favor of a basic income for all, realized by a negative income tax system. On a global scale, he makes a strong case for the creation of a world civil society. "Whatever a world civil society may eventually look like, it will be one which, as Kant put it, 'administers law among men.' Freedom under law begins with the universal recognition of human rights. Fighting for such recognition is another task which matters" (Dahrendorf, 2007b: 189).

The Erasmus Intellectuals

The architecture of liberty – the framework of a liberal and democratic society – rests upon the quest for freedom, the creation of life chances, and the power of conflicts as the driving force of social change. This institutional arrangement, which Dahrendorf envisaged in his sociology, would be incomplete without a particular class of people. Intellectuals have to be the watchdog of and for a liberal order. Those high priests of reflection, deliberation, and critique, however, have often failed in their vision and mission. As the history of this class aptly tells, they have often fallen prey to particular ideologies and radical parties. An intellectual who turns partisan loses his freedom to criticize and his independence of judgment – in Max Weber's words, his intellectual probity.

In his late writings *Engagierte Beobachter* (*Engaged Observers*, 2005) and *Versuchungen der Unfreiheit* (*Temptations of Bondage*, 2006), Dahrendorf set out the conditions under which intellectuals keep their independence of judgment and their freedom to critique. Intellectuals in times of crisis need four basic capabilities to fulfill their task as engaged observers:

1. They need to have the capacity to be independent thinkers who manage to resist the pressures of hegemonic ideologies.
2. They must have the force and power to stand antagonisms and conflicts of life and not be seduced by the temptation for 'harmonization' or the 'reconciliation' of antinomies.
3. They have to act as passionate and critical observers who assess political phenomena and social developments with a keen sense of sober realism.
4. They all have to share the position that reason is the foundation of every theory and practice.

Dahrendorf does not claim that the traits of his concept are highly original or new. In fact, the embodiment of them and insofar the paradigmatic figure in the history of intellectuals is Erasmus of Rotterdam. Erasmus in his thought and action displayed this bundle of freedom, independence, and disinterestedness. This is why Dahrendorf names intellectuals able to withstand the temptations of bondage as 'Erasmians.' It is an exclusive club which Dahrendorf, himself a founding member, brings together. Among them are Raymond Aron, Isaiah Berlin, and Karl Popper, who inspired this concept. But he counts in all intellectuals who engendered the revolution in Eastern Europe and withstood the 'totalitarianism' of communist rule.

Conclusion

Throughout his life, Dahrendorf worked at his sociology of conflict. As the three attempts mentioned here demonstrated, he set out to develop his approach in line with Karl Marx's class theory. He modernized Marx with reference to Max Weber in order to find out how the creation of life chances and the possibility of freedom could be realized in the framework of a liberal and democratic order. Much depends on the state and the market to open up life chances as options and ligatures, as entitlements and provisions. T.H. Marshall's model underpinned Dahrendorf's search for citizenships rights and their realization as entitlements guaranteed for all. But a liberal and

democratic society needs active citizens and critical Erasmian intellectuals who as civil society participate in the making of an open society. People as citizens have a stake in their society and fight in multiple conflicts over the course their society and the world is taking.

From the outset, Dahrendorf's antagonist was Talcott Parsons and structural-functionalism, which had overstressed values and norms as integrative mechanisms of society and underemphasized the role of conflict for social change. Dahrendorf's critique helped to form the paradigm of conflict as the counterpart to an alleged paradigm of consensus. But unlike Parsons, Dahrendorf never developed a grand theory. He rather followed Robert K. Merton, who advocated middle range theories. His œuvre displays no coherent and consistent theory but many steps to decipher the logic and dynamic of modern societies. Dahrendorf has no system but only a systematic approach, which he applies to different problems and phenomena. His style of thinking was both analytical and normative in that he always saw the world through the lenses of a liberal worldview.

His liberalism had nothing to do with neoliberalism, market radicalism, and the making of good times for the happy few. His social democratic family background – for him, the world of decency – and the fact that he was a sociologist, not an economist, gave his unconventional liberalism a human face. This character reflects one of his own definitions: "The possibility of progress is a condition of the battle against error. Thus we have almost defined liberalism: misanthropy plus hope; the attempt to link the practical necessity of power as intimately as possible with the greatest life chances of the greatest number; faith in the strength and the right of the individual, tempered by doubt in the perfection of things human; a little morals and a little epistemology" (Dahrendorf, 1979b: 96).

Dahrendorf was a prolific writer as the long list of publications testifies. He wrote with style and grace. In short, he was brilliant. His books did not address first and foremost his fellow social scientists, but a broader and learned public. He was a highly gifted storyteller, and he could approach the most difficult problems in a light tone, yet not superficially. Even dry discussions of concepts were couched in a talelike form and made fun to read. Most of his publications show the author's favorable form: the essay. In the professional world of today, his articles would have no chance to get published: too few footnotes, too few figures, too few publications cited from the professional field of the problem at hand. What they lacked in alleged 'professionality,' Dahrendorf's essays exhibited in intellectual content: A clear sense of the problem – what the French call a 'problématique;' an impressive capability for sober and precise judgment; the framing of the argument by embedding it in the entire cosmos of knowledge since the classics of antiquity; lessons of philosophy, history, and the social sciences woven together into a compelling argument; a sharp and concise analysis of the problem under review; the attempt to offer solutions to the problem, however preliminary and imperfect. In short, Dahrendorf's articles offered to learn something, whereas most of the professional papers by sociologists today offer information which one had not asked for in the first place. This intellectual mind-set and broad horizon he shared with his fellow liberals, Raymon Aron in France and Isaiah Berlin as well as Karl Popper in England. Since he left us, Dahrendorf's position in academia, this curious *mélange* of liberalism, social, and political theory, remains vacant.

See also: Aron, Raymond (1905–83); Civil Society, Concept and History of; Democracy, History of; Freedom, History of; Freedom: Political; Individualism versus Collectivism: Philosophical Aspects; Intellectual History; Justice, Conflicts, and the Justice of Conflict Resolution; Kant, Immanuel (1724–1804); Liberalism: Historical Aspects; Marshall, Thomas Humphrey (1893–1981); Marx, Karl (1818–83); Parsons, Talcott (1902–79); Political Thought, History of; Popper, Karl Raimund (1902–94); Simmel, Georg (1858–1918); Social Inequality in History (Stratification and Classes); Weber, Max (1864–1920); Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Alber, Jens, 2010. Der Soziologe als Hofnarr: Zur politischen und soziologischen Aktualität des Denkens von Ralf Dahrendorf. Leviathan 38 (1), 23–29.

Dahrendorf, Ralf, 1953. Unskilled Labour in British Industy. London School of Economics, London.

Dahrendorf, Ralf, 1959. Class and Class Conflict in Industrial Society. Stanford University Press, Stanford.

Dahrendorf, Ralf, 1962. Conflict and liberty. Some remarks on the social structure of German politics. The British Journal of Sociology 14 (3), 197–211.

Dahrendorf, Ralf, 1965. Bildung ist Bürgerrecht. Nannen, Hamburg.

Dahrendorf, Ralf, 1968. Essays in the Theory of Society. Stanford University Press, Stanford.

Dahrendorf, Ralf, 1974. Über den Ursprung der Ungleichheit unter den Menschen. In: Dahrendorf, Ralf (Ed.), Pfade aus Utopia. Zur Theorie und Methode der Soziologie. Piper, München, pp. 353–379.

Dahrendorf, Ralf, 1975. The New Liberty. Routledge & Kegan Paul, London.

Dahrendorf, Ralf, 1979a. Society and Democracy in Germany. W.W. Norton & Company, New York.

Dahrendorf, Ralf, 1979b. Life Chances. Approaches to Social and Political Theory. University of Chicago Press, Chicago.

Dahrendorf, Ralf, 2002. Über Grenzen. Lebenserinnerungen. C.H. Beck, München.

Dahrendorf, Ralf, 2004. Reflections on the Revolution in Europe. Transaction Publishers, Piscataway.

Dahrendorf, Ralf, 2005. Engagierte Beobachter. Die Intellektuellen und die Versuchungen der Zeit. Passagen Verlag, Wien.

Dahrendorf, Ralf, 2006. Versuchungen der Unfreiheit. Die Intellektuellen in Zeiten der Prüfung. C.H. Beck, München.

Dahrendorf, Ralf, 2007a. Auf der Suche nach einer neuen Ordnung. Vorlesungen zur Politik der Freiheit im 21. Jahrhundert. C.H. Beck, München.

Dahrendorf, Ralf, 2007b. The Modern Social Conflict. The Politics of Liberty. Transaction Publishers, Piscataway.

Dahrendorf, Ralf, 2010. Homo Sociologicus. Ein Versuch zur Geschichte, Bedeutung und Kritik der Kategorie der sozialen Rolle. VS Verlag für Sozialwissenschaften, Wiesbaden.

Kocka, Jürgen, 2009. Ralf Dahrendorf in historischer Perspektive. Aus Anlass seines Todes am 17 Juni 2009. Geschichte und Gesellschaft 35, 346–352.

Matys, Thomas, Brüsemeister, Thomas, 2013. Gesellschaftliche Universalien versus bürgerliche Freiheit des Einzelnen – Macht, Herrschaft und Konflikt bei Ralf Dahrendorf. In: Imbusch, Peter (Ed.), Macht und Herrschaft. Sozialwissenschaftliche Theorien und Konzeptionen. VS Verlag für Sozialwissenschaften, Wiesbaden, pp. 195–216.

Peisert, Hansgert, Zapf, Wolfgang (Eds.), 1994. Gesellschaft, Demokratie und Lebenschancen. Festschrift für Ralf Dahrendorf. DVA, Stuttgart.

Dance, Anthropology of

Helena Wulff, Stockholm University, Stockholm, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Dance as a topic for systematic anthropological investigation was established in the 1960s. As the Western category of dance did not always work in a cross-cultural perspective, bounded rhythmical movements were identified, as well as dance events. Dance is an expression of wider social and cultural situations, often indicating transition or conflict, as well as unity. Dance anthropologists study all forms of dance, Western and non-Western, ranging from ritual dance and social dance to street dance and staged dance performance. Dance and movement are understood in relation to theories of the body and gender, and to ethnicity, nationalism, and transnationality.

Dance has been studied in anthropology for over a century; it has also been studied in classical works, but mostly in passing as an element in ritual and ceremony. This neglect is likely to have come about because of the elusive nature of dance, in combination with the fact that dance has the potential of releasing emotional and erotic forces that are kept at bay most of the time in everyday life. There are many examples of political and religious control of dance, not least in colonial contexts. This has often produced resistance, expressed through the creation or revitalization of ethnic and national dances.

In an article about the beer dance among the Azande, E. E. Evans-Pritchard (1928) was among the first anthropologists to show that the study of dance had a wider analytical potential and that dance needed to be understood within a social and cultural context. More than 30 years later, Gertrude P. Kurath (1960) wrote an article outlining research on dance in American anthropology. A dancer herself, Kurath stressed the importance of dance training for the anthropological study of dance. This article prefigured the emergence of the anthropology of dance as a subdiscipline in the 1960s and 1970s, which was consolidated by Anya Peterson Royce's foundational *The Anthropology of Dance* (2002), first published in 1977. There is also the recent volume (in French) edited by Andrée Grau and Georgina Wierre-Gore (2005) for the early history of dance research in the United States and Europe. At first, the anthropological focus on dance stimulated discussions about the origins and definition of dance that became complicated by the cross-cultural perspective. As Adrienne Kaeppler (1985: p. 93) has remarked, the concept of 'dance' is a Western category that often fails to distinguish among different cultural activities. This concern is also paralleled, however, in the anthropology of music and the arts. Drawing on her research in Tonga, where movement around ceremonial pig presentations, *kava* mixing, and choreographed group speeches revealed hierarchy and social solidarity, Kaeppler suggested 'structured movement systems' as an alternative concept to dance. Another anthropological strategy has been to identify *dance events* as units of study, rather than to look for particular dances in isolation. On the whole, dance anthropologists seem to converge on a consideration of bounded rhythmical movements that are performed during some kind of altered state of consciousness, an elevation or even trance.

The pioneering dance anthropologists had to deal with the ethnocentric assumptions, prevalent among the then mostly autodidactic dance historians, about dance as a universal language and 'primitive dance' (often constructed in terms of a generalized 'African dance'), as an early phase in a Western scheme of dance evolution entailing a hierarchizing where ballet was regarded as a superior dance form. This prompted Joann Kealiinohomoku (1983) to write an article where she argued for a view of ballet as an ethnic dance, as one dance form among others. This is a crucial anthropological contribution to dance scholarship, but Kealiinohomoku's idea that ballet reflects Western cultural traditions is perhaps a bit exaggerated, especially in the long run. Admittedly, Western classical ballet speaks to a limited historically dominant segment of the West, in contrast to the shifting multicultural landscape of contemporary Europe and North America. To state that ballet is a form of ethnic dance also disregards the traditional difference between ethnic dance as a participatory ritual practice and ballet as a stage art performance separated from the audience, although ethnic and folk dance are now frequently moved from secular and sacred ritual to the realm of art in stage dance. The genres are further blurred by contemporary stage dance experiments with the audience taking part in the performance and even being invited to go onstage, as well as by so-called site-specific dance often being performed outdoors in city parks and subway stations, but also in rural places such as in the forest.

The central quest and *raison d'être* for the anthropology of dance have been to find out what dance says about its society, informing about social or cultural circumstances that cannot be sufficiently expressed in any other way. As the contributors in Paul Spencer's *Society and the Dance* (1985) show, this comes about not only through direct reflection but also through social commentary and critique in the form of parody-like (gender) role reversal, suggestions about alternative social scenarios, and even predictions of political events. Dance has increasingly been recognized as having an impact on social change. But there is also the cathartic view of dance as a safety valve, a structural outlet, which will keep the social order and its inequalities intact. Dance is moreover often a feature of socialization. One of the functional aspects of dance is that of transformation: the classical examples are rites of passage where people are moved from one stage in the life cycle to the next one, such as initiation rites, wedding celebrations, and funerals.

Like all subdisciplines, the anthropology of dance has followed main theoretical currents in its parent discipline.

Dance has thus been analyzed through structuralism, symbolic anthropology, semiotics, and linguistic theory, generating studies of the deep structures of dance and dance as nonverbal communication. Taking these theoretical approaches as points of departure, some of the first dance anthropologists worked out their own elaborate models and classificatory schemes for movement analysis. Drid Williams's (1976; see also Williams, 2004) theory of semasiology, the meaning of human movements as action signs, for instance, has been developed further by Brenda Farnell (1995) in a study of Plains Indian sign language. With dance cases as nonverbal communication ranging from rural West Africa to the stage in New York, Judith Lynne Hanna (1987[1979]) showed how dance can be understood through communications theory.

There are basically three kinds of anthropological studies of dance. One kind has dealt with meaning, function, and the cultural context of dance, and another kind has concentrated on the choreographic form. The majority of anthropological studies of dance, however, has combined these approaches by taking both meaning, function, and context *and* choreographic form into account, and has included textual descriptions of the movements and/or illustrations such as photographs or simple movement transcriptions, consisting of circles and arrows that indicate directions and patterns of dance.

In the United States, research on dance in its cultural context has dominated, just as in the United Kingdom where John Blacking was a key figure in the emerging interest in the anthropology of dance. In continental Europe, dance ethnologists, also called choreologists, who came out of a long nationalist folklore tradition, documented folk dances in their own societies for preservation and revival, stressing formal characteristics along the lines of music research there. During the Communist era in Eastern Europe, dance (and other folklore) researchers found themselves in ideological conflict over state manipulations and control of their research (Giurchescu and Torp, 1991). Anthony Shay (2002) has documented how state-sponsored national dance companies from countries such as the former USSR as well as Turkey, Egypt, and Mexico accommodated their repertory and choreography to the prevalent state politics.

In the 1980s, a new generation of dance anthropologists joined the growing transdisciplinary area of dance and movement studies. By engaging with feminist and gender critiques, notions of the body, identity politics, and literary and cultural studies, scholars of dance and movement studies became more visible in the academic world, and also because there was an increasing number of them. With dance historians and scholars such as Helen Thomas (1995), Jane Desmond (1997), Teresa Jill Buckland (2006), and Susan L. Foster (2011), dance scholars have moved toward ethnography and social and cultural theory, redefining their field as critical dance studies, while dance anthropologists have added Western stage dance and culture, as well as European folk dance, to their previous focus on non-Western dance and movement. Most dance scholars, including dance anthropologists, have been trained as modern and/or classical dancers. The anthropological study of dance usually requires participation in the dancing, or at least some kind of bodily knowledge of the steps that are investigated. The fact that dance anthropologists learn sequences of steps and movement that are completely different from their earlier movement experiences is something they often report on in a humble spirit, while it actually deserves recognition not least as an important methodological strategy. Some dance anthropologists have thought of their own dancing in the field as a part of the reflexive turn in anthropology, which has also had an impact on the anthropology of dance. The volume *Dance in the Field* (1999), edited by Buckland, brings together methodological deliberations on a wide variety of dance forms at home and abroad around issues of representation, reflexivity, and ethics.

Moving Bodies, Gender, and Sexuality

In the upsurge of studies in the human sciences on the body as culturally constructed, the *moving* body has nevertheless been marginal, as Susan Reed (1998) notes in her review article of the anthropology of dance, except in recent dance scholarship including dance anthropology where it has been prominent, often in tandem with issues of gender and sexuality. Such aspects were early identified by Judith Lynne Hanna (1988) in a cross-cultural panorama of different dance forms. Also, Helen Thomas (2003) delves into how moving bodies, whether in dance forms as diverse as ballet, contact improvisation, or rave, offer ideas on the politics of identity, while Eduardo P. Archetti (1999) has compared masculinities in football, polo, and tango in Argentina.

By now, a number of monographs that combine attention to moving bodies, gender, and sexuality have appeared. Jane Cowan's (1990) work on the body politics of gender in a town in northern Greece focuses on dance events, including, for example, lively wedding dances that unite a contradictory ideology of inequality and visions of closeness in the dancing bodies. For the women, dance is an ambivalent pleasure since they have to balance their sexual expression carefully between release and control; unlike that of the men, women's dancing demeanor always runs the risk of being interpreted as indecent. This is also a problem for the professional female entertainers, singers, and belly dancers in Cairo who Karin van Nieuwkerk (1995) writes about, who see their job as 'a trade like any other.' Performing at celebrations such as births, engagements, and weddings, these Egyptian entertainers create much happiness and are, in fact, crucial for the success and prestige of these occasions. Nevertheless, they are regarded as dishonorable in the context of Islam and class in Egyptian society because they make a living out of disclosing their bodies in public, thereby tantalizing men. In Cynthia Novack's (1990) ethnography of contact improvisation, the American communal art-sport that emerged in the 1960s, gender relations are quite differently featured in line with the prevailing ideology of egalitarianism. Although the body is constructed as nongendered in contact improvisation, the performers' touching and supporting of each other's weight while moving together, usually in duets, sometimes produce sexual and sensual feelings among both them and the observers.

Novack's study spans more than two decades, showing that the movements of contact improvisation, and thus the construction of these moving bodies, and the arts organizations around them followed changes in American culture so that this dance form gradually diminished together with the

egalitarian agenda of the 1960s and 1970s. A different type of body politics is articulated through Brazilian samba, which Barbara Browning (1995) writes about as a form of political resistance. Body and space are in focus in Sally Ann Ness's (1992) interpretation of the *sinulog* danced in Cebu City in the Philippines. Ness connects dance movements and space to everyday movements like walking and the urban landscape of Cebu City. A Philippine ballet, *Igorot*, which Ness (1997) discusses, produces neoethnic bodies in a structure of postcolonialism and transnationalism. J. Lowell Lewis (1992) also links everyday movement and dance movement in Brazilian *capoeira*, the martial art and dance with roots in Brazilian slavery and, further back, in Africa. Lewis analyzes *capoeira* as a cultural style through the metaphor of play: bodily, musical, and verbal. Learning *capoeira* in Salvador, Brazil, is Greg Downey's (2005) area, as he observes the bodily training having an influence on perception according to a phenomenological perspective. The significance of the black dancing body in the United States is evident in Brenda Dixon Gottshild's (2004) book. In 2005, Bryan S. Turner edited an issue of the journal *Body & Society* on 'the dancing body.' Even though the issue deals with a variety of topics, it was prompted by the ethnography that Steven P. Wainright and Turner (2005) did for their research on injuries and aging among ballet dancers. Applying a broad perspective, Farnell (2012) has pursued her interest in the body, mind, and movement by building an ontological theory of embodiment.

Ethnicity, Nationalism, and Transnationality

Ethnicity and nationalism are often negotiated through dance, but in different ways depending on historical and political circumstances. J. Clyde Mitchell's (1956) small classic monograph on the Kalela dance in colonial northern Rhodesia is an early instance of ethnic dance. The Kalela dance was performed every Sunday by labor migrants who had moved from the countryside to the town of Luanshya. While describing colonial society, the dance was also a marker of a 'tribal' group identity called for in contrast with other groups in the town. The European folk dances that European dance ethnologists have long been documenting can be categorized as ethnic dances. Yet it was the growth in studies of nationalism in the social sciences generally that inspired dance anthropologists to investigate ethnic and national dance. Located in postrevolutionary Cuba, Yvonne Daniel (1995) discusses how the government replaced ballet with rumba as the Cuban national dance. Rumba was associated with the working class and the African heritage, which suited the socialist state ideology. Studies of Javanese court dance exemplify the wider escalating concern with questions of representation, authenticity, and appropriation. In a study of the *devadasis*, female temple dancers in South India, Joan L. Erdmann (1996) points out that it is time to rewrite the history of Indian dance without Orientalism such as the Western label 'Oriental dance,' in order to make room for indigenous categories. More than a decade later, Felicia Hughes-Freeland (2008) explored how the embodied tradition in Javanese colonial court dance has turned into a national dance that influences everyday movement. And Janet O'Shea (2007) shows how the history of the Indian dance *bharata natyam*, which is danced across the world, follows the history of India. Martha Savigliano (1995) applies a transnational perspective in her study of tango in Buenos Aires, Paris, and Tokyo. Her point of departure is the political economy of passion ideas of machismo and postcolonialism. Also, flamenco is connected with passion, as William Washabaugh (1996) points out: this popular culture had a political impact during Franco's reign in Spain. The transnationality of the salsa scene in Montreal, Canada, is scrutinized by Sheenagh Pietrobruno (2006), who contextualizes this in the city's economy. Tango and salsa are among the Latin dances that can be included in ballroom dance, among waltzes and other European social dances. Competitive ballroom dance, so-called dancesport, as a global arena, has been studied by Jonathan S. Marion (2008).

Ritual dance is firstly conceptualized as a conduit for national identity and ethnicity by Zoila Mendoza (2000) in relation to dance in the Peruvian Andes, and secondly by Deirdre Sklar (2001) in relation to a Virgin of Guadalupe fiesta dance in New Mexico in the United States. Francesca Castaldi (2006) writes about the National Ballet and Senghor's Négritude in Senegal. The politics of race and African American culture in the United States is examined by Thomas F. DeFrantz (2004) through the history of the Alvin Ailey American Dance Theater. Helena Wulff (2007) has related Irish dance forms, including traditional folk dance, Irish dance theater, and Riverdance (the Irish dance show), to the development of Ireland as an independent nation. Although most dance anthropologists seem to agree that all dance forms are more or less mixed, with increasing global contacts, ethnic and national dance forms are becoming even more mixed than before. *Igorot*, the Philippine ballet that is a mixture of classical ballet and traditional Philippine dance, has been described as an instance of cultural hybridity (Ness, 1997). Drawing on Howard Becker's (1982) notion of art worlds, Wulff (1998) explores homogeneity and diversity in the transnational world of ballet.

Mediation and Preservation

As dance is elusive, its movements are difficult to document for preservation, distribution, or revival of performances. Dance is mediated to a great extent through text, scholarly as well as journalistic, a fact that is pondered by Sally Banes et al. (2007). To convey movement through the text is an obvious ambition for dance anthropologists. In her little book on tango and the multidimensional national Argentine identity, accurately entitled *Paper Tangos*, Julie Taylor (1998) has inserted small black-and-white photographs in the margin on every page. A quick flipping through the pages thus creates an image of motion, of a couple dancing the tango.

In addition to simple figures illustrating the text, some dance anthropologists have learned advanced standardized notation systems for transcription and preservation of movements, such as Labanotation, which is the most widely used, or the Benesh or Eskhol–Wachmann notation systems. The period of organized training to master any of these systems completely is often said to be like that of learning another language. In the Western dance world, trained notators, or choreologists, document ballet and dance productions through these systems,

and they have also been important in the medical rehabilitation of disabled people. There is, however, a degree of interpretation involved in the notating process, which means that different notators may describe the same movement somewhat differently. In *Choreographics* (1989), Ann Hutchinson Guest discusses in great detail a number of historical as well as contemporary dance notation systems. For notators, dance notation has been available in software, LabanWriter, for some time, and recently as an iPad app. In dance anthropology, there has been a discussion about the applicability of these Western notation systems to non-Western dance and movement forms, although there seems to be agreement that they can be used for such purposes as well. Certain conventions have developed to deal with cultural translation problems, so that the particular use of Labanotation, for example, may be explained in a 'key' for readers.

Photography and film are other, often complementary, media for conveying and preserving dance and movement. There is, however, an element of cultural selectivity in their use, as in film and video recording, which has increased substantially. Video and film have also become significant aids in the mechanical work of drawing the symbols for the scores of Labanotation and the Benesh notation system. The expanded use of technology in dance research has accentuated questions of copyright and ownership, which are further complicated in a transnational cross-cultural context. The area of dance and mediation, also including television, DVDs, and certainly the Internet, is generating an accelerating amount of new research, as exemplified by Susan Kozel's (2008) work with and about digital dance and phenomenology.

Future research in the anthropology of dance is likely to include more systematic reception and audience research. The anthropological inquiry into dance is being extended to popular dance genres such as rave, dance shows, and dance in musical videos, and to world dance, in both Western and non-Western contexts. Street dances, such as breakdance in Mary Fogarty's (2012) research, and flash mobs, the sudden, brief mass dances in public places that are arranged through mobile phones and Facebook, studied by Georgina Gore (2010), are also emerging topics for anthropological dance research; so are health and healing, which Karen Lüdke (2008) has related to the tarantula rite in southern Italy. Also disability, as in Ann Albright Cooper's (1998) work, is attracting scholarly attention. Dance and the senses is yet another emerging topic in the anthropology of dance, as shown by Tomie Hahn (2007) in her account of the traditional Japanese dance, *nihon buyo*. Even issues such as human rights and social movements are expanding areas for dance research, as Naomi Jackson and Toni Samantha Phim (2008) show. Unlike the anthropology of music, which attracts mostly men, the majority of dance anthropologists have been women. This may have played a role in the relatively modest position of the subdiscipline, at least in the past. Its present growth is shown in the volume *Dancing Cultures* (2012), where the editors Hélène Neveu Kringelbach and Jonathan Skinner argue for the importance of dance in social life, in relation to social change. They explore the shift of movements between dance and nondance contexts.

The anthropology of dance keeps contributing to central debates in the human sciences, especially to an understanding of culture in terms of process and diversity. Importantly, dance and movement not only are shaped by society but also shape society. This occurs in increasingly complex ways as the genres of dance and movement in society shift and grow, inviting more studies of the anthropology of dance.

See also: Body: Anthropological Aspects; Cultural Relativism, Anthropology of; Entertainment; Evans-Pritchard, Sir Edward E. (1902–73); Religion: Nationalism and Identity; Ritual and Symbolism, Archaeology of; Ritual.

Bibliography

Albright, A.C., 1998. Strategic abilities: negotiating the disabled body in dance. Issue on "Disability, Art and Culture". Michigan Quarterly Review XXXVII (3). http://hdl.handle.net/2027/spo.act2080.0037.313.

Archetti, Eduardo P., 1999. Masculinities: Football, Polo and the Tango in Argentina. Berg, Oxford, UK.

Banes, S., Harris, A., Acocella, J.R., Garafola, L., 2007. Before, Between, and Beyond: Three Decades of Dance Writing. University of Wisconsin Press, Madison, WI.

Becker, H.S., 1982. Art Worlds. University of California Press, Berkeley, CA.

Browning, B., 1995. Samba: Resistance in Motion. Indiana University Press, Bloomington, IN.

Buckland, T.J. (Ed.), 1999. Dance in the Field: Theory, Methods and Issues in Dance Ethnography. Macmillan, London, UK.

Buckland, T.J. (Ed.), 2006. Dancing from Past to Present: Nation, Culture, Identities. Society for Dance History Scholars Monograph Series. University of Wisconsin Press, Madison, WI.

Dancing from Past to Present: Nation, Culture, Identities. In: Buckland, T.J. (Ed.), Society for Dance History Scholars Monograph Series. University of Wisconsin Press, Madison, WI.

Castaldi, F., 2006. Choreographies of African Identities: Negritude, Dance, and the National Ballet of Senegal. University of Illinois Press, Urbana.

Cowan, J.K., 1990. Dance and the Body Politic in Northern Greece. Princeton University Press, Princeton, NJ.

Daniel, Y., 1995. Rumba: Dance and Social Change in Contemporary Cuba. Indiana University Press, Bloomington, IN.

DeFrantz, T.F., 2004. Dancing Revelations: Alvin Ailey's Embodiment of African American Culture. Oxford University Press, Oxford.

Desmond, J.C. (Ed.), 1997. Meaning in Motion: New Cultural Studies of Dance. Duke University Press, Durham, NC.

Downey, G., 2005. Learning Capoeira: Lessons in Cunning from an Afro-Brazilian Art. Oxford University Press, Oxford, UK.

Erdmann, J.L., 1996. Dance discourses: rethinking the history of the 'Oriental dance'. In: Morris, G. (Ed.), Moving Words. Routledge, London, UK, pp. 288–305.

Evans-Pritchard, E.E., 1928. The dance. Africa 1, 446–462.

Farnell, B., 2012. Dynamic Embodiment for Social Theory: 'I Move Therefore I am'. Routledge, London, UK.

Farnell, B.M., 1995. Do You See What I Mean? Plains Indian Sign Talk and the Embodiment of Action. University of Texas Press, Austin, TX.

Fogarty, M., 2012. Breaking expectations: imagined affinities in mediated youth cultures. Continuum 26 (3), 449–462.

Foster, S.L., 2011. Worlding Dance: Studies in International Performance. Palgrave Macmillan, New York.

Giurchescu, A., Torp, L., 1991. Theory and methods in dance research: a European approach to the holistic study of dance. Yearbook for Traditional Music 23, 1–10.

Gore, G., 2010. Flash mob dance and the territorialisation of urban movement. Anthropological Notebooks 16 (3), 125–131.

Gottschild, B.D., 2004. The Black Dancing Body: A Geography from Coon to Cool. Palgrave Macmillan, New York.

Grau, A., Wierre-Gore, G. (Eds.), 2005. Anthropologie de la Danse: Genèse et Construction d'une Discipline. Centre National de la Danse, Pantin, France.

Hahn, T., 2007. Sensational Knowledge: Embodying Culture through Japanese Dance. Wesleyan University Press, Middletown, CT.

Hanna, J.L., 1987[1979]. To Dance Is Human: A Theory of Nonverbal Communication. University of Chicago Press, Chicago.

Hanna, J.L., 1988. Dance, Sex and Gender Signs of Identity, Dominance, Defiance, and Desire. University of Chicago Press, Chicago.
Hughes-Freeland, F., 2008. Embodied Communities: Dance Traditions and Change in Java. Berghahn Books, Oxford, UK.
Hutchinson Guest, A., 1989. Choreographics: A Comparison of Dance Notation Systems from the Fifteenth Century to the Present. Taylor & Francis, London, UK.
Jackson, N.M., Phim, T.S. (Eds.), 2008. Dance, Human Rights, and Social Justice: Dignity in Motion. Scarecrow Press, Lanham, MD.
Kaeppler, A.L., 1985. Structured movement systems in Tonga. In: Spencer, P. (Ed.), Society and the Dance. Cambridge University Press, Cambridge, UK, pp. 92–118.
Kealiinohomoku, J., 1983[1970]. An anthropologist looks at ballet as a form of ethnic dance. In: Copeland, R., Cohen, M. (Eds.), What Is Dance? Oxford University Press, Oxford, UK.
Kozel, S., 2008. Closer: Performance, Technologies, Phenomenology. MIT Press, Cambridge, MA.
Kurath, G.P., 1960. Panorama of dance ethnology. Current Anthropology 1 (3), 233–254.
Lewis, J.L., 1992. Ring of Liberation: Deceptive Discourse in Brazilian Capoeira. University of Chicago Press, Chicago.
Lüdtke, K., 2008. Dances with Spiders: Crisis, Celebrity and Celebration in Southern Italy. Berghahn Books, Oxford, UK.
Marion, J.S., 2008. Ballroom: Culture and Costume in Competitive Dance. Berg, Oxford, UK.
Mendoza, Z.S., 2000. Shaping Society through Dance: Mestizo Ritual Performance in the Peruvian Andes. University of Chicago Press, Chicago.
Mitchell, J.C., 1956. The Kalela Dance: Aspects of Social Relationships among Urban Africans in Northern Rhodesia. Rhodes-Livingstone Institute by the Manchester University Press, Manchester, UK.
Ness, S.A., 1992. Body, Movement and Culture: Kinesthetic and Visual Symbolism in a Philippine Community. University of Pennsylvania Press, Philadelphia.
Ness, S.A., 1997. Originality in the postcolony: choreographing the neoethnic body of Philippine dance. Cultural Anthropology 12 (1), 64–108.
Neveu Kringelbach, H., Skinner, J. (Eds.), 2012. Dancing Cultures: Globalization, Tourism and Identity in the Anthropology of Dance. Berghahn, Oxford, UK.
Novack, C., 1990. Sharing the Dance: Contact Improvisation and American Culture. University of Wisconsin Press, Madison, WI.
O'Shea, J., 2007. At Home in the World: Bharata Natyam on the Global Stage. Wesleyan University Press, Middletown, CT.
Pietrobruno, S., 2006. Salsa and Its Transnational Moves. Lexington Books, Oxford.
Reed, S.A., 1998. The politics and poetics of dance. Annual Review of Anthropology 27, 503–532.
Royce, A.P., 2002[1977]. The Anthropology of Dance, second ed. Dance Books, Alton.
Savigliano, M., 1995. Tango and the Political Economy of Passion. Westview Press, Boulder, CO.
Shay, A., 2002. Choreographic Politics: State Folk Dance Companies, Representations and Power. Wesleyan University Press, Middletown, CT.
Sklar, D., 2001. Dancing with the Virgin: Body and Faith in the Fiesta of Tortugas, New Mexico. University of California Press, Berkeley.
Spencer, P. (Ed.), 1985. Society and the Dance: The Social Anthropology of Process and Form. Cambridge University Press, Cambridge, UK.
Taylor, J., 1998. Paper Tangos. Duke University Press, Durham, NC.
Thomas, H., 1995. Dance, Modernity and Culture. Routledge, London, UK.
Thomas, H., 2003. The Body, Dance and Cultural Theory. Palgrave Macmillan, London, UK.
Turner, B.S. (Ed.), 2005. Body & Society 11 (4) (issue on 'the dancing body').
van Nieuwkerk, K., 1995. A Trade Like Any Other: Female Singers and Dancers in Egypt. University of Texas Press, Austin, TX.
Wainwright, S.P., Turner, B.S., 2006. 'Just crumbling to bits'? An exploration of the body, ageing, injury and career in classical ballet dancers. Sociology 40 (2), 237–255.
Washabaugh, W., 1996. Flamenco: Passion, Politics, and Popular Culture. Berg, Oxford, UK.
Williams, D., 1976. Deep structures of the dance. Journal of Human Movement Studies 2, 123–144.
Williams, D., 2004. Anthropology and the Dance: Ten Lectures, second ed. University of Illinois Press, Urbana, IL.
Wulff, H., 1998. Ballet across Borders: Career and Culture in the World of Dancers. Berg, Oxford, UK.
Wulff, H., 2007. Dancing at the Crossroads: Memory and Mobility in Ireland. Berghahn Books, Oxford, UK.

Dark Triad: The "Dark Side" of Human Personality

Peter K Jonason and James P Middleton, University of Western Sydney, Milperra, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Most of work in personality psychology has focused on the Big Five traits. However, recent research suggests the Big Five may be insufficient in that it is atheoretically derived and may inadequately assess the 'darker' sides of people's personality. An emerging stream of research has attempted to address this by studying what is called the Dark Triad of personality (i.e., narcissism, psychopathy, and Machiavellianism). In this article we attempt to discuss the Dark Triad traits on their own and as a group, we cover different methodological and theoretical considerations, and end by offering some suggestions for the future.

Personality psychology, also called differential psychology, has a long history. The field is concerned with systematic variation between individuals, what causes those differences, how we can document those differences, and how those differences track with real-world outcomes. The best studied set of personality traits are the Big Five traits of extraversion, agreeableness, neuroticism, conscientiousness, and openness. These traits are based on words identified as adjectives people use to describe one another (viz., the lexical assumption) through the use of various factor analytic techniques. While other factor structures exist (e.g., Cattell's 16 Factors; the PEN Model), the Big Five have dominated the last 60 years of research in personality psychology, finding application in all fields of psychological inquiry.

As popular and useful as the Big Five is and as good as the Big Five is at covering a broad range of personality traits, it tends to be limited in two important ways. The Big Five is a set of atheoretically derived, descriptive adjectives and it tends to better tap 'positive' aspects of people's personality over 'negative' or 'darker' sides. This has created two streams of research, both highly informed by evolutionary or interactionist approaches to understanding within-species variation. One group has focused on expanding the list of basic personality traits; developing what is called the HEXACO. These researchers have focused on the added predictive power of measuring individual differences in people's honesty and humility. However, this approach is still heavily reliant on factor analytic techniques.

The other – perhaps the newest – stream has focused on what are called the 'Dark Triad' of personality traits (i.e., narcissism, psychopathy, and Machiavellianism). Unlike the Big Five or the HEXACO, the Dark Triad traits – as a collective – were not derived through inductive factor analytic techniques, but were, instead, linked based on theoretical observations about the overlap of the behaviors and nature of these traits. Importantly, they also are concerned with 'darker' aspects of human nature.

In this article, we discuss the work on the Dark Triad. In particular, we attempt to highlight the apparently good and bad aspects of these traits. This is essential in that even the worst traits confer some benefits if viewed from the proper perspective. As reviewed elsewhere, the predominant way of seeing these traits may be biased toward prosocial outcomes. That is, if the traits cost the group like in the case of crime, then they are solely seen as bad. We feel this is overly simplistic. Crime is pursued to derive short-term benefits to the criminal at the cost of others. We would argue that research has only focused on the latter half of this last sentence. Apparently costly behavior for the group could have real benefits (e.g., food, mates) for the individual. Even still, the individual may suffer from some disadvantages but, if these traits gear an individual toward maximizing short-term gains as we will argue later, these may be costs he or she is willing to pay.

The Costs and Benefits of the Dark Triad

Narcissism

Narcissism is based on the Greek myth of Narcissus who fell in love with his own reflection and was later identified as a personality disorder. Indeed, most conceptualizations of narcissism have included some dimension of vanity. However, there is a general agreement that narcissism is also composed of other traits like grandiosity, a sense of superiority, and entitlement, although the necessary and sufficient number of factors is a matter of considerable debate. Importantly, narcissism is not simply high self-esteem, though it is positively correlated with it. In fact, narcissism has different effects on aggression depending on the type of self-esteem someone has. In those with stable self-esteem, narcissism is positively correlated with attitudinal aggression, whereas in those with unstable self-esteem, narcissism is negatively correlated with attitudinal aggression. Alternatively, those high in narcissism aggress in response to ego threat as opposed to other threats. Narcissism is a disposition toward an egocentric lifestyle, in which the needs of the self is placed above the needs of others.

One of the hallmarks of narcissism is its inescapable social aspect. Narcissists desire admiration and attention, they wish to feel superior within social circles, and this ties them inextricably to other people. Thus, narcissists employ strategies, which involve improving their relationships with other people. For example, narcissists tend to be overconfident at everything they do, regardless of how good they are, and report being far more competent than others at tasks. While most people tend to overestimate their abilities, narcissists tend to overestimate their abilities more than others. This overconfidence can have

a beneficial effect for the person however, as others are more likely to see them as good leaders, even if their narcissism tends to actually negatively influence group performance. For instance, this overconfidence in their ability leads them to being more resilient in the face of failure.

While it can be difficult for a narcissist to maintain relationships in the long term, they do have a number of advantages in the beginning stages of social relationships. Similarly, narcissism has been linked with other factors with positive social benefits, including higher extraversion, a more affiliative humor style, higher IQ, and a socially charming disposition. Narcissism may even offer some positive effects including self-protective memory biases, the maintenance of one's self-concept across both high- and low-achieving peer groups, increased within-group creativity, and short-term mating success. That said, narcissism remains linked to poor judgment in other areas. Specifically, narcissism has been linked with a higher rate of gambling, as well as an increased amount of money spent on gambling, and these links are mediated by overconfidence, increased risk acceptance, and focus on rewards. Additionally, in tasks where groups of people are asked to harvest a slowly regenerating resource, group performance was worse overall when narcissists were included, as narcissists would tend to overharvest and exhaust resources quicker than others. It seems reasonable to conclude that the Dark Triad may have some short-term benefits and some long-term consequences.

Psychopathy

Like narcissism, psychopathy traditionally was considered a clinical diagnosis and was adapted to measure a nonclinical construct in the general population. This construct includes such things as impulsivity, criminality, aggressiveness, callousness, limited empathy, and low fear. This has resulted in psychopathy being labeled the 'darkest' of the Dark Triad traits. The term carries with it the worst connotations and is a term commonly ascribed to criminals in real life and in the media. In fact, we would contend that the common usage of this term and the tendency to link this term with 'bad' people may actually interfere with clear thinking about this trait more so than the other Dark Triad traits.

Just like narcissism, psychopathy appears to be a mixed bag of some beneficial and some deleterious aspects. One interesting finding from such studies is that successful psychopaths (i.e., those that do not get institutionalized) have heightened cardiovascular stress reactivity and perform better on the Wisconsin Card Sorting Test (i.e., a measure of cognitive function) compared to unsuccessful (i.e., institutionalized) psychopaths. Furthermore, significant neurological deficits have been found in unsuccessful psychopaths, including greater hippocampal asymmetry, and a 22.3% reduction in prefrontal gray matter volume.

Despite being a socially aversive trait, psychopathy is not entirely without benefits. While psychopathy tends to be linked with low conscientiousness and low agreeableness, it is also linked with low neuroticism, an ostensible benefit. Similarly, psychopaths have lowered stress responses in real-world environments such as appearing in court. This fits the general pattern of impulsiveness, shallow affect, and sensation seeking, typical of those high on psychopathy. Additionally, some have argued that psychopathy, while having a number of difficulties in long-term scenarios, has a number of short-term advantages. For example, the aggression and short-term mating success all suggest psychopathy might provide some adaptive benefits. If one thing has been made clear in the literature, it is that psychopathy is not the purely negative, maladaptive, and pathological disposition, it is often considered to be.

Machiavellianism

While the other traits of the Dark Triad were initially pathological disorders that were adapted into measures for nonclinical populations, Machiavellianism was developed for social psychological research. Researchers took inspiration from Niccolo Machiavelli's work *The Prince* to construct a trait characterized by cynicism, manipulativeness, and a pragmatic morality. This construct, being based on Machiavelli's work, was termed 'Machiavellianism.' These traits might be instrumental for princes (and people) to rise in social and political hierarchies in medieval Europe and today. The trait is considered to have an instrumental way of dealing with others, a protean or whatever it takes style of social influence, and an opportunistic approach to economic exchanges or work behavior. For instance, it is linked with the adoption of both *hard* and *soft* tactics of social influence in the workplace. This range of tactics may better equip those high Machiavellianism to influence others to bending to her or his will.

Much like the other factors of the Dark Triad, Machiavellianism has costly aspects. While high Machiavellian individuals tend to rank high on social charm and manipulativeness, Machiavellianism is negatively correlated with self-control, emotional intelligence, and social competency in both intimate and exposed settings. However, seeing these as costs might be overly simplistic. It has recently been argued that limited empathy could itself be advantageous to people motivated to take advantage of others. The idea being that if one understood the feelings of others, it might inhibit their ability to exploit others to their own ends. Indeed, those high on Machiavellianism appear to be skilled liars. Nevertheless, like the other two, the costs and benefits associated with Machiavellianism depend on the context.

Overlap and Distinctiveness

Contemporary research on the Dark Triad has seen a great deal of debate as to whether the Dark Triad should be treated as a single construct, or whether it is best to treat it as a three-factor solution. Essentially, this debate revolves around two questions. First, which explain more variance in people's behavior, attitudes, and cognition, the three individual traits or a collective, composite of the three? Second, might the Dark Triad be tapping some underlying latent personality trait?

Unificationism

Proponents of the view that the Dark Triad can be used as a single construct argue the collective allows for higher-order theoretical predictions, whereas the three traits on their own

show the mechanisms by which the latent construct relates to certain outcomes. The point here is that the three traits are imperfect measures of the same or a similar latent disposition. Taking a logical positivism perspective, they contend that personality traits are convenient fictions that allow us to talk about these traits; the traits themselves are not real themselves.

Evidence for the utility of a unitary construct comes in a number of forms, all implying there is an underlying core behind the Dark Triad. First, there are a wide range of personality traits that appear to link the Dark Triad traits together: Agreeableness, dishonest/immodesty, limited empathy, short-term mating, future discounting, and callousness. Second, factor analyses suggest that the three load on a single factor well and account for over 50% of the variance in each construct. Indeed, at least one measure of the Dark Triad returns better fits for structural models that take into account a latent, global factor. Third, so there is at least one case where the latent construct outperforms the three traits on their own when assessing sociosexuality. Fourth, it might be that these traits are simply reflections of similar a disposition with different researchers making the measures, coming from different perspectives, and focusing on different aspects of the underlying profile. That is, we should not assume that because a given trait as conceptually been about X that the original conceptualization was correct.

Separationism

Their opponents argue that some important relationships can only be found when examining the factors within the Dark Triad, and that it is only through this kind of examination that a fuller, deeper understanding of the construct can be found. They would contend that simply because there are relationships between the three does not mean they are equivalent; after all, the traits of the Dark Triad were chosen *because* they shared a socially aversive component. It should be expected that any set of traits selected to represent the 'dark side' of human personality should, by definition, share a core of being 'dark.' These researchers point the Dark Triad do show some differences in important ways. For example, while both narcissism and psychopathy can predict attitudinal aggression, narcissists tend to aggress toward ego threats, whereas psychopaths tend to aggress toward physical threats. Other factors, such as IQ, are related only to narcissism, while only psychopathy is related to neuroticism, and even within studies showing similar relationships between the three traits, there is often a difference in how strong those relationships are. If some relationships can be isolated to a single trait, and we can see differences in how well each trait can predict other factors, then the claim that they are identical seems inadequate. Indeed, examinations of the various subfactors of narcissism and psychopathy have revealed that it is even suspect to treat the Dark Triad as unified constructs. For example, secondary psychopathy, exploitive narcissism, and Machiavellianism have been linked with a *fast* life strategy, whereas primary psychopathy and grandiose narcissism have been linked with a *slow* life strategy. Additionally, the same cluster of traits related to a 'fast' life strategy (i.e., secondary psychopathy, exploitative narcissism, and Machiavellianism) have been linked with an evening chronotype.

A Resolution?

The most reasonable position might be to just examine all three levels (i.e., the collective, the three traits, and the parts of the traits). Doing this will allow researchers to test the conditions under which each perspective has its utility and to get a high degree of resolution about the loci of the various relationships they investigate. We would contend that the collective is useful for making higher level predictions regarding the nature of the underlying core, whereas the prediction at the trait level may be more mechanistic. This comes down to the quality of the theory one is using to make their predictions. Most personality psychology predictions are on the order of proximal and descriptive. That is, researchers assemble correlates from other studies to build the case that their two pet variables will be linked. We urge caution here in taking an exploratory approach to examine the subfactors. In contrast, if one has a strong theoretical paradigm (e.g., evolutionary psychology), one can make more sensible predictions but those predictions may not be able to focus on any one of the traits if we assume the traits (or at least their measures) are imperfect themselves. Whether the Dark Triad is treated as a single construct, a three-factor construct, or even a construct with more than three factors, should depend on the situation. Ultimately, it is up to the researcher – guided by theory – to determine the best approach to use for her/his project.

The approaches used to study and understand the Dark Triad vary considerably within the literature. This variation stems from disagreements over what the Dark Triad is, how it should be measured, and most importantly, how we can better understand it in future research. What should be clear here is that the factors within the Dark Triad are conceptually important and distinct components, and that blindly forcing them into a single construct in every situation, regardless of their differences, loses an important part of their utility. Likewise, forgetting their similarities and treating the factors as universally separate and isolated concepts can be just as bad. While there is no clean answer to these problems, future research might provide some answers, however, long this might take. Until then, researchers are advised to carefully consider these points before they commit to any one approach.

Measurement and Theoretical Issues with the Dark Triad

Arguably, the first task to making any serious advances in any science is having standard metrics. Doing so allows researchers to all 'speak the same language.' Unfortunately, personality psychology tends to resist this trend with most personality constructs having multiple measures. The Dark Triad is no exception. There are various measures of all three traits. Early work relied mostly on the Narcissistic Personality Inventory, the Self-Report Psychopathy Scale-III, and the MACH-IV. However, in hopes of reducing subject fatigue, standardizing a single Dark Triad measure, reducing bloated specificity, and reducing contamination from related constructs the *Dirty Dozen* was created. However, as a brief measure (viz., 12 items) it cannot maintain the heterogeneity in the longer (i.e., 124 items) measures as has been noted by the scale authors and its

critics. Nevertheless, it remains the most well-validated, short, published, brief instrument to assess the Dark Triad. Despite its brevity, it has proved its value in Classical Test Theory and Item Response Theory assessments, the three traits have different correlates, and the measures appear to cover the range of subtraits (e.g., primary psychopathy, vanity) each of the three constructs contains.

Putting issues of measurement aside, there are some major theoretical considerations. Up until about 2009, nearly all work on the Dark Triad was descriptive. This early work was focused more on accounting for the shared correlation among the Dark Triad. However, descriptive work is extremely limited although it is all-too-common in personality psychology. Thankfully, a number of authors have ventured into different theoretical paradigms to attempt to understand the nature of the Dark Triad. We briefly review some of them here.

A number of theories could be lumped into what we will call interactionist theories (e.g., nature and nurture). Evidence suggests the Dark Triad traits are heritable with each having some degree of susceptibility to environmental contingencies – Machiavellianism, in particular. In addition, the Dark Triad has been linked with many evolutionarily important behaviors, such as a short-term mating style and a 'fast' life strategy leading some to examine the Dark Triad from an evolutionary perspective. But if these traits are heritable, how might we account for the continued persistence of the traits in the population?

Parasites and Hosts

Those high on the Dark Triad do appear to have an exploitive, parasitic, and opportunistic mating style. Might insights from germ theory provide a way to understand the Dark Triad? While reasoning from analogy can have its limitations, we contend that thinking of those high on the Dark Triad as 'parasites' and society as the 'host' may have some value. Parasites operate on different timelines. Genital herpes operates on a slow timeline. It infects the cells of the host and uses the host to increase its spread through infecting others, but little more than mild flu-like symptoms are observed before the first outbreak. In contrast, influenza operates on a faster timeline. It infects the host, quickly creating fever and even causing death. Simultaneously, the host has strategies to avoid infection and minimize its effects through the behavioral and biological immune system. If the Dark Triad is adaptive, those characterized by the traits need to strike a balance in the timeline of their infection and the strategies they use to promote their inclusive fitness and must 'trick' the host. Indeed, people have a heightened amygdala and orbital prefrontal cortex responses to faces of people who score high on a psychopathy measure and those high on the Dark Triad appear to have adaptive strategies to minimize detection. This suggests a coevolutionary arms race between the host and the parasite or the society and the person high on the Dark Triad.

Life History Theory

Research on the Dark Triad has exploded in the last 5 years. One reason for this has its integration within evolutionary psychology, in particular life history theory. From this perspective, personality traits are evolutionarily stable strategies designed by natural selection and shaped by environmental contingencies (e.g., harshness and instability in one's childhood) to enable reproductive fitness. Those high on the these traits might have a 'fast' life history strategy geared toward maximizing immediate returns over delayed outcomes as would be predicted by the tradeoffs expressed in life history theory as seen in future discounting, short-term/exploitive mating and social strategies, and impulsivity. The Dark Triad traits could be maintained in the population if some positive fitness benefit is accrued. What is often disparaged as a maladaptive personality can be conceptualized within life history theory as an adaptive stable strategy on a shorter time horizon.

Niche-Specialization Hypothesis

One of the limitations of evolutionary psychology is that it is too reliant on hypothesized ancestral conditions. As an alternative, one might use behavioral ecology to study within-species variation. In this case, we adopt reasoning by homology. That is, from what we know from primatology we can reason what the role of personality traits might be. Personality traits may drive individuals to select certain environments over others. In order to avoid predation, find food, increase mating success, primates adopt different niches to survive. This might allow for a pairing of trait and niche. That is, because the human socioecology is so varied, individuals can specialize in one or a set of environments to their own benefits. Indeed, those high on the Dark Triad appear to have a preference for the time of the day that would make it easiest for them to take advantages of others and to select friends who serve some function instead of just for enjoyment.

Social Exchange Theory

Social Exchange Theory proposes that behaviors can be thought of as the result of cost-benefit analyses by people attempting to interact with society and the environment. If a person believes that they are able to extract more of a reward through a behavior than they lose by performing it, then the person will perform the behavior. Conversely, when the person feels that the cost will outweigh the benefit, the behavior will not be performed. If someone does not value maintaining a relationship, or is distrustful and does not expect others to follow through with the reward, then the balance would be shifted toward anticipating a lower worth to any social exchange. That is, the relative costs would be higher, and the anticipated rewards would be lower. Such a mind-set could produce a short-term orientation favoring immediate, assured rewards, and would elicit the socially aversive behaviors characteristic of the Dark Triad. However, while we agree that there are costs and benefits, we are suspicious of apparent the Neo-Classical Economics assumption that humans are rational, calculating animals.

Overall

If we are to continue work on the Dark Triad, we need to keep in mind that it provides both benefits and deficits to individuals in terms of both resource and mate management.

An evolutionary approach can be used to understand not only why the Dark Triad exists, but under what circumstances it can be the most beneficial, and the most detrimental. Understanding what triggers the development of the Dark Triad would be an integral part of discovering how we can either prevent, treat, or utilize Dark Triad strategies. It might be that societal and occupational niches could be found for those high on the Dark Triad, and instead of simply trying to change these individuals to fit what we consider to be 'good,' we might find various roles for them to fill in society; much in the same way that we consider killers bad, and yet employ some people (i.e., army infantry) to go out and kill.

Limitations and Final Thoughts

Contemporary research on the Dark Triad comprises mainly self-report measures. While this is often considered standard in psychological research, it raises an interesting problem; in Dark Triad research, we ask naturally deceptive people to be honest about their personality. While every attempt is made to ensure participants feel confident that their responses will be anonymous, it is clear that there is no direct benefit from being truthful. With no direct benefit for their honesty, and a potential serious cost if anonymity is not held, it is hard to find a reason for someone high on the Dark Triad to actually be truthful. In addition, self-report measures are prone to any kind of deceptive behavior, including social desirability influences, self-delusion (e.g., a narcissist refusing to admit that they seek attention for whatever reason), and 'faking.' Nonetheless, studies have reliably found links between self-reported Dark Triad scores and external variables, indicating that the self-report measures are managing to tap the constructs they claim to.

Recently, literature has begun to push further outside social and behavioral correlates of the Dark Triad. For example, a number of studies now have linked psychopathy with various neurological factors. While this represents an important step forward, it is important not to get too far ahead of ourselves. The Dark Triad traits are not *real* constructs; they do not exist within the brain, and they are not neurological structures. The Dark Triad traits are descriptors for behavioral patterns, nothing more. While these behavioral patterns might be the result of specific changes within the brain, this certainly does not make them equivalent.

The dispositional approach itself is not immune to this gradual reification either. Any behavioral trait, no matter how reliably linked to other variables, is a descriptor. That is, all traits are abstract concepts that we use to categorize behavior. While this is useful to help predict behavior across situations and examine the relationships between behaviors, it will not, by itself, explain why the behavior occurs. Thus, while the dispositional perspective can provide important insight into behavior, it can never fully account for all aspects of it. One major problem, for example, is that even in a highly sophisticated trait, you can never use it to fully predict responses in every scenario. People with the same trait in the same environment could very well display different behavior.

The field of personality has come a long way since its inception. Most recently, research has focused on examining behavior in terms of traits, or behavioral descriptors, in which a person's general behavioral trends are summed up into a single constructs. However, preliminary approaches focused too heavily on 'normal' functioning and neglected the more socially aversive aspects of behavior. With the introduction of the Dark Triad, researchers finally had a tool to begin to examine the darker aspects of human personality.

See also: Authoritarian Personality; Five Factor Model of Personality, Assessment of; Five Factor Model of Personality, Personality Disorder; Personality Theory and Psychopathology; Personality, Biological Models of; Personality, Evolutionary Models of; Personality, Trait Models of.

Bibliography

Buss, D.M., 2009. How can evolutionary psychology explain personality and individual differences? Perspectives in Psychological Science 4, 359–366.

Campbell, W.K., Bush, C.P., Brunell, A.B., Shelton, J., 2005. Understanding the social costs of narcissism: the case of the tragedy of the commons. Personality and Social Psychology Bulletin 31, 1358–1368.

Figueredo, A.J., Wolf, P.S.A., Gladden, P.R., Olderbak, S.G., Andrzejczak, D.J., Jacobs, W.J., 2009. Ecological approaches to personality. In: Buss, D.M., Hawley, P.H. (Eds.), The Evolution of Personality and Individual Differences. Oxford University Press, New York.

Furnham, A., Richards, S.C., Paulhus, D.L., 2013. The Dark Triad of personality: a 10 year review. Social and Personality Psychology Compass 7, 199–216.

Glenn, A.L., Raine, A., 2009. Psychopathy and instrumental aggression: evolutionary, neurological, and legal perspectives. International Journal of Law and Psychiatry 32, 253–258.

Hare, R.D., 1993. Without Conscience: The Disturbing World of the Psychopaths among Us. Guilford Press, New York.

Jonason, P.K., Jones, A., Lyons, M., 2013. Creatures of the night: chronotype and the Dark Triad traits. Personality and Individual Differences 55, 538–541.

Jonason, P.K., Koenig, B.L., Tost, J., 2010. Living a *fast* life: the Dark Triad and life history theory. Human Nature 21, 428–442.

Jonason, P.K., Li, N.P., Webster, G.D., Schmitt, D.P., 2009. The Dark Triad: facilitating a short-term mating strategy in men. European Journal of Personality 23, 5–18.

Jonason, P.K., Webster, G.D., 2010. The dirty dozen: a concise measure of the Dark Triad. Psychological Assessment 22, 420–432.

Jonason, P.K., Webster, G.D., 2012. A protean approach to social influence: Dark Triad personalities and social influence tactics. Personality and Individual Differences 52, 521–526.

Jonason, P.K., Webster, G.D., Schmitt, D.P., Li, N.P., Crysel, L., 2012. The antihero on popular culture: life history theory and the Dark Triad personality traits. Review of General Psychology 16, 192–199.

Jones, D.N., Paulhus, D.L., 2010. Different provocations trigger aggression in narcissists and psychopaths. Social Psychological and Personality Science 1, 12–18.

Lee, K., Ashton, M.C., 2005. Psychopathy, Machiavellianism, and narcissism in the five-factor model and the HEXACO model of personality structure. Personality and Individual Differences 38, 1571–1582.

McCrae, R.R., Costa, P.T., 2008. The five-factor theory of personality. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research. The Guilford Press, New York.

McHoskey, J.W., Worzel, W., Szyarto, C., 1998. Machiavellianism and psychopathy. Journal of Personality and Social Psychology 74, 192–210.

Mealey, L., 1995. The sociobiology of sociopathy: an integrated evolutionary model. Behavioral and Brain Sciences 18, 523–599.

O'Boyle, E.H., Forsyth, D.R., Banks, G.C., McDaniel, M.A., 2012. A meta-analysis of the Dark Triad and work behaviour: a social exchange perspective. Journal of Applied Psychology 97, 557–579.

Paulhus, D.L., Williams, K.M., 2002. The Dark Triad of personality: narcissism, Machiavellianism, and psychopathy. Journal of Research in Personality 36, 556–563.

Darwin, Charles Robert (1809–82)

Michael T Ghiselin, California Academy of Sciences, San Francisco, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Charles Darwin (1809–82) became an evolutionist on the basis of biogeographical evidence. His theory of natural selection, based upon individualistic reproductive competition, explained evolution without recourse to teleology or the supernatural. He presented the basic arguments for evolution and natural selection in his book *On the Origin of Species*, published in 1859. He elaborated upon his ideas in a series of books that dealt with a remarkably wide range of evolutionary topics. Darwin was largely responsible for establishing the modern evolutionary point of view and initiating many lines of investigation that continue to the present day.

Darwin's Life

Charles Robert Darwin was born in Shrewsbury, England, on 12 February 1809. He died at Down, England, on 9 April 1882. His father, Robert Waring Darwin (1766–1844), was a prosperous physician. His mother, Susannah Wedgwood (1765–1817), was the daughter of the industrialist Josiah Wedgwood II (1769–1843). Although the young Darwin developed extracurricular interests in the natural sciences, he was considered a mediocre student. At the age of 16 he was sent to the Edinburgh University to study medicine. He interacted with scientists there and furthered his extracurricular interests in natural history. But he decided that he did not want to practice medicine and realized that inherited wealth would make it unnecessary to work for a living. However, his father insisted that he have a respectable profession, and it was decided that he would become a clergyman. In those days, the study of natural history was deemed a suitable activity for clergymen. He therefore enrolled at Cambridge University on 15 October 1827, and remained there until 1831. His academic performance there was respectable, though not outstanding. On the other hand, his scientific interests were encouraged by the faculty.

Darwin was preparing for some scientific travel when he was invited to join a naval surveying expedition on *HMS Beagle* as unofficial naturalist and gentlemanly companion to the captain of the ship. The *Beagle* left England on 27 December 1831, circumnavigated the globe, and returned on 2 October 1836. During the voyage, he made extensive collections of plants, animals, and other natural history specimens. But he was mainly interested in geology, partly as a result of reading *The Principles of Geology* (1830–1832) by Charles Lyell (1797–1875). In that work Lyell discussed the evolutionary ideas of Jean-Baptiste Lamarck (1744–1829) but Darwin did not become an evolutionist until after the voyage, nor did Lyell until many years later. It was Darwin's geological work, especially his theory of coral reefs, that first made him famous.

Early in 1837, expert opinion on his collections, especially those of the birds of the Galapagos Archipelago, convinced Darwin that evolution was a viable idea, and he set out to develop a theory. As part of this research program, Darwin read the *Essay on the Principle of Population* by the economist Thomas Robert Malthus (Malthus, 1826). This led Darwin to appreciate how the scarcity of resources affects reproductive competition, and, around 28 September 1838, he discovered the principle of natural selection. The properties of the most successful competitors are increasingly represented in successive generations, transforming their species.

Darwin married his cousin Emma Wedgwood on 29 January 1839; their first child was born on December 27 of that year. They set up their house in London, but his health began to deteriorate and he suffered from mysterious ailments for the rest of his life. In 1842, the family relocated to a country residence near Down, south of London. Nonetheless, Darwin was highly productive. A book on his voyage (Darwin, 1839) was a popular as well as a scientific success. He edited the zoology of the voyage and, in addition to minor papers, published three volumes on his geological research (Darwin, 1842, 1844, 1846). Development of his ideas on evolution continued, and he wrote a sketch of a book on it in 1842, and a draft in 1844.

In 1846, Darwin began some research on barnacles, which led him to spend 8 years revising the classification of that poorly understood group of marine animals and writing a vast taxonomic monograph (Darwin, 1851, 1854), which in effect was his first book on evolution. He then proceeded to work up his materials on evolution. A large book entitled *Natural Selection* was in the process of composition when, on 18 June 1858, a manuscript from Alfred Russel Wallace (1823–1913) arrived in which the same basic mechanism was proposed. A joint paper was read before the Linnean Society on July 1, and on July 20, Darwin began to write an 'abstract' that ultimately became the *Origin of Species* (Darwin, 1859). The first of six editions was published on 24 November 1859.

Darwin brought out a series of books that dealt with a broad range of evolutionary topics. These books all included results of his ongoing empirical research, much of it on the sexuality, physiology, and behavior of plants. A book on orchids (Darwin, 1862) was followed by one on climbing plants that first appeared as a journal article (Darwin, 1865). What had been a section of *Natural Selection* became *The Variation of Animals and Plants under Domestication* (Darwin, 1868). It was the second of what are considered Darwin's 'major' evolutionary works. The third was *The Descent of Man, and Selection in Relation to Sex* (Darwin, 1871). *The Expression of the Emotions in Man and Animals* (Darwin, 1872) was followed by a series of botanical works (Darwin, 1875, 1876, 1877a, 1877b, 1880). The last of his books dealt with earthworms, especially their

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61021-8

behavior (Darwin, 1881). Darwin died at his home at Down on 19 April 1882. He was buried at Westminster Abbey on April 26.

Darwin's Scientific Contribution

Darwin brought about the most important revolution in the history of the human intellect. For that very reason, it has taken a long time for his ideas to become assimilated and the process is far from complete. Although he was able to convince the scientific community that evolution had in fact taken place, he was far less persuasive with respect to natural selection, which only became the majority view toward the middle of the twentieth century. Furthermore, evolution as Darwin and his successors have come to understand it was something quite different from the 'evolution' of his predecessors and many of his contemporaries. This is true of both evolution and its mechanisms.

Darwin's early geological research was crucial to his intellectual development. It gave him experience in thinking historically, dealing with real events that happened to concrete particular things. The fossil record, however, was still very poorly understood, and one of Darwin's major contributions was showing how incomplete it was. It was biogeography, not paleontology, that convinced Darwin that evolution had occurred and provided his most compelling arguments for it. But that required thinking about groups of organisms not only changing but also having common ancestors. Darwin's predecessor Lamarck (1744–1829) had published a branching diagram, but his basic conception was that of multiple origins of simple organisms by spontaneous generation, and their descendants becoming increasingly complex as independent lineages. It was a more historical version of the traditional idea of a *scala naturae*, with organisms ranked as higher or lower according to how closely they resembled human beings.

When Darwin's predecessors attempted to explain the transformation of species, they invoked the kinds of causation that they were familiar with in other areas of experience. Supernatural influences played a diversity of roles in such thinking, often important ones. Change in lineages was largely thought of in terms of analogies with the development of embryos. A kind of 'evolutionism' was possible in which simple organisms had been created, followed by a long succession of generations before the result that had been foreseen from the beginning – Man – was realized. Alternatively, the Creator had ordained laws of nature, such that matter would organize itself into increasingly complicated forms with Man, again, the preordained result. These unknown laws of nature were often conceptualized as similar to those that govern the formation of crystals.

In Darwin's theory, evolution was subject to the laws of nature, but there was no law or other cause that drove things in any particular direction. Organisms evolved the properties that they did because of historical contingencies, and Darwin's critics often equated such contingency with 'chance.' Attempting to find explanations for evolution in terms of 'law' rather than 'chance' has been a recurrent theme in efforts to find alternatives to Darwinism.

The term 'orthogenesis' is often applied to notions of evolution occurring along predetermined lines, especially when there is some sort of analogy with the formation of an embryo. Such quasievolutionary ideas began to receive attention from social scientists before the publication of the *Origin of Species* (Darwin, 1859). The stage theory of Auguste Comte (1798–1857), with its theological, metaphysical, and positivistic levels, is one important example. Closer to what we moderns think of as evolution was the philosophy of Herbert Spencer (1820–1903). Spencer analogized both biological and socioeconomic evolution with the embryology of Karl Ernst von Baer (1791–1876). In Spencer's 'synthetic philosophy,' the increasing specialization of the parts in developing organisms was analogized with the specialization and diversification that occurs in society.

The 'superorganism' view of social and biological evolution has been firmly rejected by Darwin's followers. However, a more limited form of 'holism' that affirms the reality and importance of supraorganismal groups has become increasingly popular in the philosophy of systematics. Ontologically, species and other taxonomic groups are now considered not to be classes but individuals in the broadest sense: particular, concrete things that have a definite position in space and time. Like individuals in the familiar sense of organisms, they can change indefinitely, they have no definitions, and there are no laws of nature for any one of them.

Species are now understood to be supraorganismal wholes that are held together by sex. Darwin, however, never had a satisfactory species concept. But he did emphasize the interaction between organisms and the reproductive competition among them as the fundamental cause of change. As he saw it, the competition for such resources as food is important only insofar as it affects reproductive success. The crucial role of sexual reproduction is underscored in Darwin's theory of sexual selection, a major topic in *The Descent of Man* (Darwin, 1871). The 'secondary' sexual characters, such as gaudy plumage in birds and the antlers of stags, were interpreted by Darwin as the result of purely reproductive competition, based upon monopolizing mates, and unrelated to survival or procreation pure and simple.

Misunderstandings about competition have always been a major impediment to the assimilation of Darwin's theory. In economics, competition is thought of as a market condition that is antithetical to monopoly. The lay person, however, tends to confuse competition with other phenomena, such as physical combat, and to treat it as if it were the opposite of cooperation. For many, even teachers who grade on a curve, competition is something antisocial, if not downright immoral. Of course, political considerations have added to the confusion but it also reflects a deeper, metaphysical problem. The pre-Darwinian mentality had generally tended to assume that the world was created by God and that, in the words of Alexander Pope, 'Whatever is, is right.' Short of rejecting the basic premise, there are two popular alternatives. One is to endorse the conditions that exist. The other is to deny their existence. Various forms of social Darwinism have opted for the former. More tender-minded social thinkers, from Prince Kropotkin to Stephen Jay Gould, have opted for the latter, portraying the natural economy as one in which competition is unimportant. Darwin of course rejected the basic premise.

This perhaps bothered his contemporaries more than any other aspect of his theory. It allowed the phenomena of life to be explained in purely natural science terms, without recourse to teleology in any but a trivial sense. Everything came into existence as a matter of historical contingency in the course of a vast sex contest.

Partly because of his experience with barnacles, Darwin became increasingly aware of the importance and pervasiveness of sex. The barnacles, although generally hermaphroditic, turned out to cross-fertilize. Turning his attention to plants, Darwin found that they cross-fertilize far more often than anybody had realized. Indeed, there were widespread adaptations furthering outcrossing. He undertook extensive empirical investigations, with striking results that revolutionized floral biology. His first book after the *Origin of Species* (Darwin, 1862) dealt with the pollination biology of orchids. It documented in great detail the often complex adaptations that enabled the orchids to have insects transfer pollen from one flower to another. These adaptations were often quite bizarre and the book can at least be read as a sort of satire on traditional natural theology. It also counteracted the morphological tradition that had tended to dismiss function. The book evoked a debate between Darwin and his supporter the Harvard botanist Asa Gray (1810–88). Darwin firmly resisted Gray's efforts to find God superintending the overall trend of evolution. In the orchid book and other botanical publications, Darwin showed that inbreeding could be highly deleterious and that plants have both anatomical and physiological adaptations that further outcrossing.

In *The Descent of Man, and Selection in Relation to Sex* (Darwin, 1871), Darwin laid the foundation for a great deal of work on the comparative biology of reproduction. His theory of sexual selection followed from his basic idea that evolution results from reproductive competition between individuals. He interpreted such features as the antlers of stags and gaudy plumage in male birds as adaptations used in monopolizing mates, and not as contributing to survival or otherwise giving an advantage in competing with organisms of other species.

Individualistic competition also has a great deal to do with another important topic in that book: the evolution of morals. Darwin was quite clear on the point that organisms generally act in their own interest. He even suggested that an adaptation in one species that existed solely for the good of another species could not be explained in terms of his theory. He likewise ruled out adaptations that would benefit a species as such. There was as he saw it no good evidence that such adaptations exist. However, he still had to account for the existence of self-sacrifice upon the part of social organisms, including apparently altruistic acts by human beings. His discussions of such topics include basically everything that was considered by late-twentieth-century sociobiologists but without the genetics. He seems to have understood kin selection, insofar as he pointed out that in artificial selection traits could be favored by what Darwin called 'breeding from the same stock.' A relative of an animal with a trait such as flavorful meat could be selected for breeding instead of that particular animal. He seems not to have considered kin selection very important, however. Neither did he emphasize reciprocity, though he recognized its existence. Rather, he stressed familial selection, which was important in his treatment of the neuter casts of social insects such as ants and termites. He treated families as individuals, with the relevant competition going on between nests of ants or hives of bees. In dealing with human morality, Darwin invoked selection at the level of the tribe or nation. His reasons for invoking such group selection, rather than, say, reciprocity, are not explicitly stated in his writings.

To what extent did Darwin go along with the 'social Darwinism' of his times? This term is by no means monolithic and merely saying that he endorsed some of it hardly answers the question. Social Darwinism was basically a political doctrine and it generally had a basis that was if not religious, at least teleological. Interfering with the struggle for existence would tend to prevent the desirable end result from happening. It would be like interfering with the growth and maturation of an organism. Darwin did not endorse such claims. However, he did maintain that free competition favored both biological and economic progress, the two being interrelated insofar as the human intellect is a source of innovation. Lessening the struggle would, he thought, tend to be dysgenic.

There was also a psychological aspect to Darwin's thinking about morals, one that often focused upon the 'moral sentiments,' as Adam Smith had called them. He tried to derive those sentiments from a feeling of sympathy. He considered morality highly serviceable in the struggle for existence, but the selective context was crucial. He suggested that in bees, in which the drones are killed by the workers, it would be appropriate for such fratricide to be a matter of conscience. Some of his readers were scandalized. Selection favoring instinctive behavior was important here, and it contrasts with much of the thinking of his contemporaries. Herbert Spencer had attributed moral sentiments to a more Lamarckian mechanism. There was a kind of racial memory of the sufferings of our ancestors. The British empiricist tradition had sought to explain ethics in terms of personal experience. Spencer made it racial experience. Darwin replaced experience with reproductive success.

The Descent of Man, and Selection in Relation to Sex (Darwin, 1871) was followed by its sequel, *The Expression of the Emotions in Man and Animals* (Darwin, 1872), which further elaborated upon Darwin's evolutionary psychology. Darwin rejected mind–body dualism and thought that the functioning of the nervous system and the rest of the body were very tightly linked. He used facial expression and other body movements as a means of gaining access to the inner workings of the nervous system. It also provided examples of evolutionary continuity. As he saw it, emotional expressions might come to function in communication, but in keeping with his general antiteleological position they did not originate that way. Rather, they could be traced back to historical antecedents having a different functional significance. Darwin's interest in psychology can be traced back to his undergraduate years and was a major topic for his empirical research. Two features of this work need to be stressed at the outset. First, although he owed much to the extensive literature on such topics, with which he was quite familiar, much of his contribution came from rejecting the assumptions that had often been made. For example, he rejected the (teleological) notion of Charles Bell that there are anatomical adaptations for expressing emotions. Second,

his work in that area rested upon a solid basis in experience, including laboratory experiments as well as natural history observations, with much attention paid to the underlying physiological mechanisms, this at a time when psychology was just beginning to disassociate itself from philosophy and become a laboratory science.

Much of Darwin's behavioral research was done on plants. One naturally hesitates to call such work 'ethology' or 'physiological psychology,' but the lack of muscles and nerves in plants does not prevent them from moving and responding to stimuli. Bridging the gap between plants and animals was part of Darwin's effort to overturn traditional metaphysics. The book on insectivorous plants does more than just allude to that point. In his work on climbing plants Darwin found a morphological continuity between the organs used in climbing and less modified precursors. Tendrils were the modified petioles of leaves, for example. He also found that he could trace the circling movements with which the plants located objects upon which they climbed to a circling movement that had no such function. That led him to investigate the manner in which plants respond to directional stimuli. His research on that topic was first-rate plant physiology and gave rise to a vast literature. His follower George John Romanes (1848–94) extended that sort of reasoning to simple marine animals and thereby founded an important tradition in the study of animal behavior.

Darwin also concerned himself with the ontogeny of behavior, especially that of human beings. He kept detailed notes on the behavioral maturation of his own children. Some of this material was incorporated in *The Expression of the Emotions in Man and Animals* (Darwin, 1872). A paper containing excerpts from the notes (Darwin, 1877a) is considered one of the classics of child psychology. Darwin had several reasons for studying behavioral ontogeny. Not the least is the light it may cast on the role of inheritance versus learning. If there has been no opportunity for learning, and behavior appears spontaneously in the course of maturation, then a strong case can be made for inheritance.

Another reason for studying the ontogeny of behavior was Darwin's basic position on the relation between development and evolution in general. For him, the modification of ontogeny was a fundamental evolutionary process, and he routinely approached evolutionary problems from the point of view of embryology. Darwin believed that the processes responsible for the development of a fertilized egg into an adult are things that themselves evolve. This belief was in itself an important innovation. Perhaps more importantly, he showed how natural selection could modify the structure and function of organisms by affecting the various stages of development. By implication, the history of life could be approached by asking how the underlying developmental processes had been modified. His great follower Fritz Müller (1821–97) developed Darwin's approach by suggesting that the history of certain crustaceans could be interpreted as the result of stages being added at the end of the life cycle. As a consequence, the most modified animals would repeat, during their life cycles, the series of adult ancestors. Another of Darwin's followers, Ernst Haeckel (1834–1919), took Müller's special case and treated it as the general rule, and enunciated what he called the fundamental biogenetic law, according to which ontogeny recapitulates phylogeny.

Darwin's views on the relationship between development and evolution are discussed at length in *The Variation of Animals and Plants under Domestication* (Darwin, 1868). The variations that play an indispensable role in natural selection are here shown to occur in a more or less orderly fashion, evidently as a result of the modification of existing developmental mechanisms. Variation was thus physiologically constrained, and evolution by natural selection was by no means affected only by external conditions as is so often maintained in caricatures of Darwinism. Because this book contains toward its end an effort to explain heredity as well as development, the work is often misinterpreted as mainly about genetics. Darwin attempted to explain the inheritance of acquired characteristics (and much else as well) in terms of particles that control development. These particles could be affected in the somatic tissues, not just the germ tissues. The particles interested some of Darwin's followers and, with considerable modification in the hands of Hugo De Vries and others, they evolved into the 'genes' of later authors.

Darwin's Relevance for Today

Darwin was a remarkably creative thinker, as attested by the surprise with which his discoveries were so often met. One reason for this was the sort of wide-ranging curiosity and erudition that has become increasingly difficult to emulate as knowledge becomes ever more specialized. He was a very good observer, but insisted (at least in private) that this was because his observations were guided by theory. On the other hand, he was also an empiricist, and did not sympathize with the 'deductive' approach as he called it of such authors as Herbert Spencer. He confronted his theories with a mass of evidence and modified them in the course of investigation.

Darwin was not constrained by academic fashion or pedantry in seeking material that could be put to scientific use. He unearthed a wealth of information in the writings of plant and animal breeders and appealed to gardeners and others to provide additional materials. He also carried out extensive series of observations and experiments, often opening up new areas for investigation. We should mention his studies of competition between plants as pioneering contributions to experimental ecology. Darwin's experimental work is highly regarded by plant physiologists and students of animal behavior. However, his originality so far as methodology is concerned has more to do with comparison than with experiment. In Darwin's hands, comparative biology became linked to the revolution in systematics, a discipline that was now based upon genealogical relationships and therefore fundamentally historical. Hypotheses could be developed and tested by such procedures as plotting data onto a phylogenetic tree. Phylogeny was especially important with respect to adaptation, which was now conceived of as the outcome of a historical process.

The shift to historical narrative is an aspect of the rejection of traditional, teleological modes of thinking in which states of adaptation had implied purposefulness, even conscious design. Although structures might give many of the appearances of having been constructed by an intelligent being, it was no longer legitimate to treat ours as the best of all possible worlds.

By taking a historical approach and understanding the actual processes and events that have brought about the present properties of organisms, we can avoid the teleology. Instead of asking what an organ is 'for,' we ask what has brought it into being. It then becomes apparent that the world abounds in maladaptations and historical vestiges.

Contrary to what has sometimes been maintained, natural selection is not capable of explaining just about any conceivable property of organisms. Indeed the mechanism places strong restrictions upon the actual course of evolutionary change. For that very reason, it provides a powerful tool for evaluating hypotheses, especially where adaptation is involved. There is nothing that might reasonably be considered analogous to foresight when evolution occurs by natural selection. Any 'preparation' for future conditions of existence has to be the result of coincidence, and populations remain adjusted to present circumstances, and organisms repeat that which furthered their ancestors' reproductive success.

Because natural selection depends upon individualistic competition, there are strong constraints upon the kind of adaptation that can occur above the level of the organism. As already mentioned, Darwin clearly recognized that the individuals that reproduce differentially can be families as well as organisms, and on that basis he was able to explain the evolution of neuter castes in social insects. How readily that can happen has been a serious topic of controversy ever since. Darwin did not believe that differential reproduction of species could produce adaptations that exist 'for the good of the species,' though the possibility has subsequently been taken seriously by some authors. 'Group selection' was widely invoked in the early twentieth century. A critical attitude toward it did not become common until the latter half of that century, when in fact there would seem to have been something of an overreaction. From the point of view of Darwin's contribution, however, what matters is not the ultimate answer that may emerge, but rather the manner of framing the question. One needs to ask precisely how the world is structured and how the components might interact, formulate explicit hypotheses, and then test them.

Such issues, of course, are particularly important in the study of social behavior. The modern discussion of what sociobiologists call 'altruism' is a continuation of Darwin's pioneering work in that area. It is important, however, to recognize that 'altruism' in this sense is a technical term without ethical connotations and means self-sacrifice pure and simple. But people habitually slip back into older ways of thinking and tend to attach value judgments to statements of fact on the assumption that what is 'natural' is therefore good. Even where the teleology is repudiated explicitly, it may persist in the form of implicit assumptions. It is clearly an overreaction, however, to reject the study of adaptation in favor of so-called proximate causes. 'Teleology' in the limited sense of homing behavior and the capacity for foresight and understanding on the part of human beings and some other animals is by no means inconsistent with the expectations of selection theory.

Darwin recognized that natural selection is capable of accounting for the phenomena of progress. He reasoned that in a competitive economy, innovations that are effective in dealing with the environment would tend to accumulate. On the other hand, he rejected the views of many of his predecessors and contemporaries whose conceptions of progress remained teleological. Notions about evolution, whether social, economic, or biological, based upon analogies with the growth of crystals or the development of embryos implied a kind of inevitability that was out of line with Darwin's theory. But they appealed to those who were looking for some kind of long-term goal to it all. Darwin witnessed a great deal of technological progress in his day, and like the overwhelming majority of scientists then and now, he obviously believed in scientific progress as well. Although the idea of progress has recently come under attack by philosophers and social scientists, this may just be another episode in the ongoing process of coming to grips with Darwin's contribution.

See also: Darwinism: Social; Darwinism; Emotion: History of the Concept; Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Malthus, Thomas Robert (1766–1834); Organizational Ecology; Sociobiology: Overview; Spencer, Herbert (1820–1903).

Bibliography

Browne, J., 1995. Charles Darwin: Voyaging: Volume One of a Biography. Alfred A. Knopf, New York.
Darwin, C., 1839. Journal of Researches into the Geology and Natural History of the Various Countries Visited by H.M.S. Beagle, under the Command of Captain Fitzroy, R.N. From 1832 to 1836. Henry Colburn, London.
Darwin, C., 1842. The Structure and Distribution of Coral Reefs. Smith, Elder, London.
Darwin, C., 1844. Geological Observations on the Volcanic Islands Visited during the Voyage of H.M.S. Beagle, Together with Some Brief Notices of the Geology of Australia and the Cape of Good Hope. Smith, Elder, London.
Darwin, C., 1846. Geological Observations on South America. Smith, Elder, London.
Darwin, C., 1851. A Monograph on the Sub-class Cirripedia, with Figures of All the Species. The Lepadidae or, Pedunculated Cirripedes. Ray Society, London.
Darwin, C., 1854. A Monograph on the Sub-class Cirripedia, with Figures of All the Species. The Balanidae (or Sessile Cirripedes), the Verrucidae, Etc., Etc., Etc. Ray Society, London.
Darwin, C., 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. John Murray, London.
Darwin, C., 1862. On the Various Contrivances by Which British and Foreign Orchids are Fertilised by Insects, and on the Good Effects of Intercrossing. John Murray, London.
Darwin, C., 1865. On the movement and habits of climbing plants. Journal of the Linnean Society of London (Botany) 9, 1–118.
Darwin, C., 1868. The Variation of Animals and Plants under Domestication. John Murray, London.
Darwin, C., 1871. The Descent of Man, and Selection in Relation to Sex. John Murray, London.
Darwin, C., 1872. The Expression of the Emotions in Man and Animals. John Murray, London.
Darwin, C., 1875. Insectivorous Plants. John Murray, London.
Darwin, C., 1876. The Effects of Cross and Self Fertilisation in the Vegetable Kingdom. John Murray, London.
Darwin, C., 1877a. A biographical sketch of an infant. Mind 2, 285–294.
Darwin, C., 1877b. The Different Forms of Flowers on Plants of the Same Species. John Murray, London.
Darwin, C., 1880. The Power of Movement in Plants. John Murray, London.
Darwin, C., 1881. The Formation of Vegetable Mould, through the Action of Worms, with Observations on Their Habits. John Murray, London.
Ghiselin, M.T., 2009. Darwin: A Reader's Guide. California Academy of Sciences, San Francisco.
Lyell, C., 1830–1832. Principles of Geology. being an Attempt to Explain the Former Changes of the Earth's Surface, by Reference to Causes now in Operation. Murray, London.
Malthus, T.R., 1826. An Essay on the Principle of Population. Murray, London.

Darwinism

Michael Ruse, Florida State University, Tallahassee, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Darwinism is the theory of evolution through natural selection as discussed by Charles Darwin in his *Origin of Species* (1859). Darwin himself established beyond doubt the fact of evolution, and although the mechanism, natural selection, was controversial for many years, now with the backing of modern genetics it is appreciated as the fundamental cause of evolutionary change. There are controversial extensions of Darwinism, especially into the realm of human behavior, and also fruitful new paths, especially in the field of medicine.

Darwinism is the term given to the theory of evolution that originates with the work of the Englishman Charles Robert Darwin (1809–82), particularly as expounded in his work *On the Origin of Species* (1859). It is the theory that claims that the truly significant cause of change is the 'natural selection' of some organisms over others: more organisms are born than can survive and reproduce, some have features that help them in the consequent 'struggle for existence,' and these are winnowed or selected to be the seed stock for all succeeding generations. There is change and the direction is toward 'adaptive' excellence, meaning by this that organisms have features (like the eye and the hand) that enable them to survive and reproduce. These conclusions apply to all organisms, including us humans, consequences that have had major implications for our subsequent thinking about science, philosophy, and also religion. This article discusses these and related claims, paying attention to the reception of Darwin's ideas at the hands of his supporters like Thomas Henry Huxley, and looking especially at those advances arising from the development in the twentieth century of theories of heredity, or genetics. Consideration is given to the students of social behavior, the sociobiologists, and attention is paid to the modern controversies around their claims. Also considered are the claims by the students of paleontology, the punctuated-equilibrium supporters, that the course of evolution is not smooth and gradual but marked by periods of inactivity (stasis), broken by rapid changes. Brief mention is also made of new extensions of Darwinism, particularly into the field of medicine.

'Darwinism' can properly be dated from 1859, the year in which English biologist Charles Robert Darwin (1809–82) published his magisterial evolutionary work, *On the Origin of Species by Means of Natural Selection* (*see* Evolution, History of). To understand the magnitude of what Darwin achieved in that work, and the nature of the movement to which he gave his name, it is perhaps best to recognize that three different kinds of issues are at stake here (Ayala, 1985). First, there is what one might speak of as the very fact of evolution itself: the claim that all organisms, living and dead, are the result of a developmental process governed by natural law. Then, second, is the question of *paths* of evolution, or, as these are known technically, of phylogenies. Here, one is concerned about the routes along which evolution proceeded. Third, there is the question of *causes* or mechanisms. The work of early nineteenth-century French biologist Jean Baptiste de Lamarck (1809), who argued that organisms change due to the inheritance of acquired characteristics, speaks to this particular issue. Here, one is concerned with the driving forces behind the evolutionary process.

Although Charles Darwin was not the first evolutionist – in the eighteenth century, at length in rhyming couplets, his own grandfather Erasmus Darwin had celebrated the joys of organic change – it was in the *Origin* that Darwin established once and for all the fact of evolution. He did this through making appeal to a unifying argument or what he would have thought of as a 'consilience' (Ruse, 1979). Darwin argued simply that if evolution was true, then one could explain facts in a wide range of areas. For instance, one can show why the fossil record is roughly progressive. Likewise, one can show why organisms are distributed as they are around the world. One can explain the facts of embryology, such as why organisms take very different forms when adult (like the human and the dog) but have embryos that are almost identical. Why is it that the facts of anatomy are as they are? Why do we find skeletal isomorphisms (known technically as 'homologies') between the forelimbs of organisms, despite these limbs having different functions? The bat uses its forelimb for flying, the mole for digging, the whale for swimming, and the human for grasping, yet the bones are very similar. One can explain this, argued Darwin, through the fact of evolution. All of these phenomena, and many more, can be explained by evolution. Conversely, they point backward up to the truth of the fact of evolution (Ruse, 2008).

In the *Origin*, Darwin had relatively little to say about the paths of evolution, although it was he who tied evolution firmly to the metaphor of the tree of life. Unlike some early evolutionists, Darwin always stressed the extent to which all life is related and descends 'from one or a few forms,' branching as time unfurls to the vast diversity that exists today. But, paradoxically, in the nineteenth century it was the non-evolutionists – particularly Lamarck's contemporary and fellow countryman Georges Cuvier, who was both the father of comparative anatomy and an accomplished paleontologist – who first started to work out the actual course of life's history (Ruse, 1996). In such labors, Cuvier was backed by German-style biologists – particularly the Swiss-American ichthyologist Louis Agassiz – who saw analogies between life's history and the development of individuals. Darwin himself relied both on the fossil evidence

and on embryology and its supposed analogies when, around 1850, he performed a massive taxonomic study of barnacles. It is clear that although he did not draw explicit inferences about evolutionary paths, he had some strong convictions and suspicions (Richmond, 1988). But none of these found their ways into the *Origin*, and later in the nineteenth century it was Darwin's young supporters like Thomas Henry Huxley who were most interested in particular paths. In the United States, there was much activity in this direction, thanks particularly to the magnificent fossil findings, especially those of the horse and of dinosaurs (Bowler, 1996).

Natural Selection

It was when it came to causes or mechanisms that Darwin himself felt that he really was able to make significant scientific advances. He proposed his well-known mechanism of evolution through natural selection brought on by a struggle for existence, starting from the claims made by the English clergyman Thomas Robert Malthus (1826) that population numbers will always outstrip food supplies and space availability. Generalizing this to the animal and plant world, Darwin suggested that there will always be, at the same time, naturally occurring variation within populations: his studies of barnacles convinced him strongly that this was true. Darwin therefore concluded that in the struggle, success would be at least in part a function of the differences occurring naturally within organisms. There will thus be a natural form of selection, corresponding to the artificial selection practiced so successfully by animal and plant breeders in the first part of the nineteenth century in the United Kingdom.

But, more than just this, Darwin believed that natural selection would lead to change of a particular kind (i.e., toward adaptive advantage). Darwin had read and been much impressed by the teleological arguments expressed by Archdeacon William Paley (1802/1819), a work that Darwin had to read as an undergraduate. Paley argued that the eye is very much like a telescope, and that just as a telescope has a telescope maker, so the eye has an eye maker – that is to say, the Great Optician in the Sky. Darwin rejected Paley's conclusions, but he completely accepted Paley's premises about the design-like, teleological nature of the organic world. It was Darwin's belief that natural selection speaks to this aspect of organic life, providing a naturalistic – that is to say, law-bound – mechanism of explanation (Ruse, 2003).

Huxley's Program

After the *Origin* was published, evolution as fact was never again seriously challenged as a matter of science. But the mechanism of which Darwin was so proud got relatively little attention. Darwin's great supporters, Huxley particularly, had other ends for evolution than a functioning professional science with selection at its core (which had surely been Charles Darwin's own aim). Huxley and his fellows were busy creating a secular society in Victorian times: one that would have open education for all without religious qualifications, and one that would have a professionally trained civil service, not to mention other things like a military in which one's rank did not simply depend on one's social status and personal wealth. With some good reason, the Huxleyites saw the established church as being very much part and parcel of the powers that they wished to overthrow. Hence, they were eager to find some alternative secular substitute (Ruse, 1996, 2005).

Evolution fitted this need entirely. It was not so much that evolution was being pushed as something in direct violation of Genesis or anything like that. Rather, evolution could function as a kind of secular religion or ideology, on which one could hang all sorts of hopes of progress, ethical change, and the like. It is no surprise to find that, as the nineteenth century drew to an end, the evolutionist *par excellence* was less Charles Darwin and more his fellow countryman Herbert Spencer (Richards, 1987). The latter, in his massively written 'Synthetic Philosophy,' was providing a world picture that would be all encompassing, right up to and including our own species. Naturally, as a prospective religion, evolution was used to promote moral and social causes. Notoriously, Spencer and his supporters – prominent in this respect were the American businessmen John D. Rockefeller Sr. and Andrew Carnegie – endorsed a kind of *laissez-faire*, arguing that as with the struggle for existence, survival, and selection in nature, so also should there be such a struggle and selection in society. It is paradoxical that this movement became known as social Darwinism, for although there are certainly hints of such ideas in Darwin's own writings (especially in his work on our species, *The Descent of Man*, 1871), primarily this was a movement to which he contributed little and that altogether took evolutionism in directions he rather regretted.

The Synthetic Theory

Darwin's ideas as science provided only part of the picture. The whole question of mechanisms was rather delayed until the twentieth century. Then came the development of an adequate theory of heredity explaining how and why it is that organic characteristics are transmitted from one generation to another, and why new variations appear occasionally, spontaneously as it were. When this was first done, many thought that the new theory of heredity, genetics, was in opposition to any kind of selection mechanism. But by the 1930s, a number of mathematically gifted biologists – notably R.A. Fisher (1930) and J.B.S. Haldane (1932) in the United Kingdom and Sewall Wright (1931, 1932) in the United States – showed that one can put together Darwinian selection with Mendelian genetics in a smooth synthesis (Provine, 1971; Ruse, 2013). Finally, an adequate mechanism for the evolutionary process was developed and accepted. However, much more work needed to be done. It fell to the next generation of evolutionists, notably people like Julian Huxley (1942) (the grandson of Thomas Henry Huxley) in the United Kingdom and the Russian-born American geneticist Theodosius Dobzhansky (1937), and his associates, to put the empirical flesh on the mathematical skeletons of the 'population geneticists.' Thus, around 1950, a full theory, which came to be known as the 'synthetic theory of evolution' or 'Neo-Darwinism,' had been developed: Darwin's theory of natural selection spliced in with Mendelian genetics, offering the way forward for

a full and satisfying causal picture of evolutionary change (*see* Microevolution).

Sociobiology

The years since around 1950 have been among the most exciting in evolution's history. Although no one now wants to challenge the ultimate picture, there have been both extensions and challenges. Most notably, the former has involved the pushing of a Darwinian selective approach into the realm of social behavior. Darwin himself always thought that such behavior was part and parcel of the evolutionary picture, but he recognized that it was extremely difficult to study. However, by the 1960s evolutionists felt sufficiently confident to extend their work to such organisms as the Hymenoptera (ants, bees, and wasps) as well as to mammals, including our own species. This development of the evolution of social behavior from a Darwinian perspective – so-called sociobiology – has been highly controversial (*see* Microevolution). The leaders in the field, people like William Hamilton (1964) and John Maynard Smith (1982) in the United Kingdom and Edward O. Wilson (1975) in the United States, not to mention the great popularizer Richard Dawkins, author of *The Selfish Gene* (1989), have been subject to a great deal of critical attention (Ruse, 1985). But now I think it is generally realized that one can profitably take a Darwinian selection-based approach into the social realm, dissecting and showing in great detail how the behavior of organisms is just as much a function of their evolutionary past as is their physical nature.

Still highly controversial has been some of the work by human sociobiologists that tries to take a selection-based approach to human nature and behavior (Ruse, 2012). By way of example, one might mention the Canadian researchers Martin Daly and Margo Wilson (1988). They show that, in many respects, family patterns are tightly controlled by evolutionary considerations. Why is it, for instance, that violence in families almost always comes from the hands of men: usually not by biological fathers, but by stepfathers? Remarkably, Daly and Wilson show that stepfathers are 100 times more likely to murder their children than are biological fathers. This they relate to straightforward Darwinian principles, namely, that from a biological point of view, survival and reproduction may be improved if one can clear the way (as it were) for one's own offspring, but that damaging or harming one's own biological offspring is to go completely against the whole Darwinian evolutionary process (*see* Kin Selection).

Critics of Adaptation

On the other side, there are the critics of adaptationism. The most notable is probably the American paleontologist Stephen Jay Gould, who, like Richard Dawkins, was a very popular writer. Gould put forward his theory of so-called punctuated equilibria, arguing that if we look at evolution over long time scales, we see that selection as a force for constant, unbroken change falls away dramatically (Eldredge and Gould, 1972; Gould, 1982). Rather, what we get are long periods of inactivity (so-called stasis) broken or punctuated by very rapid changes from one form or blueprint (Gould likes the German term *Bauplan*) to another. Gould argues that although selection certainly improves organic characteristics, 'tuning them up' as it were, by and large the basic patterns of organisms are part of a general, morphologically constrained, nonselective picture. A vertebrate having four limbs, for instance, rather than six like many invertebrates, has nothing whatsoever to do with natural selection (Sepkoski and Ruse, 2009).

There are many other developments going on, particularly those involving molecular biology. For instance, it is suggested that perhaps at the molecular level, beneath the level at which selection operates, one gets a whole different dimension to evolution: perhaps a dimension where random factors are very significant (this is the theory of 'molecular drift,' Kimura, 1983). In other spheres, it has been documented that viruses can carry units of heredity ('genes') from one organism to another of a very different kind, in what is known as 'horizontal gene transfer' (Doolittle, 1999). Also, there is much work going on at the moment of related questions such as the origin of life. Certainly, no one has a firm solution in this direction, although once again major advances have been made and we know a great deal more than we did in the past (Ruse and Travis, 2009).

Evolutionary Medicine

As Thomas Kuhn (1962) taught us all, one of the marks of a really good area of science ('paradigm') is the extent to which it shows new fields and opens new topics for discussion. This is very true of Darwinism. Thomas Henry Huxley, who was trained as a medical doctor, could see little real applicability to Darwinism; for him, it was always more of a metaphysical approach. Although he was much interested in the reformation of medical education, Darwinism had no role in it. Today, thanks to pioneering work by the evolutionist George C. Williams and the psychiatrist Randolph Nesse, it is realized how much Darwinism might be important in medicine (Nesse and Williams, 1994) (*see* Evolutionary Medicine). This is obvious when dealing with genetic diseases like sickle-cell anemia, where one needs an evolutionary understanding of how these gene-caused illnesses persist (in this case, because the genes confer protection against malaria). It is true also when dealing with the evolution of diseases themselves, for instance the evolution of bacteria in the face of new antibiotics. It is well known that within a year or two of a new medicine's use, the diseases being attacked will respond with new forms. An understanding of the mechanism of evolution can go far to aid in this problem, as does teaching practitioners about the most effective doses and so forth. And, as a third area where Darwinism is important, one might mention the effects of diseases themselves. Many infections bring on elevated temperatures, to which the usual advice is 'Take two aspirin, drink lots of fluids, and go to bed.' But is it the most sensible thing to take aspirin and thus reduce the body temperature? Could it be that the high temperatures are themselves part of (evolved) nature trying to fight the infections itself? Perhaps high temperatures have a purpose! These and other questions have been raised by Darwinian medicine, today one of the most

exciting and challenging areas of application of the theory of the *Origin* (Gluckman et al., 2009).

The Meanings of 'Darwinism'

So, in conclusion, trying to tie together strands from history and the present, how then are we to define and characterize 'Darwinism' or a 'Darwinian,' the latter meaning someone who accepts the doctrine? Darwinism certainly commits one unambiguously to the fact of evolution. The only people today who challenge this seriously are those driven by religious commitments, namely, the extreme Christian evangelicals known as 'fundamentalists' who find that evolution clashes with their prior commitments to the Bible (especially Genesis) taken literally (Johnson, 1991). There are some Darwinians who think that their position also requires belief in some kind of philosophical materialism, claiming that any kind of supernatural belief is false or impossible (Dennett, 1995). Today, the so-called New Atheists are at the head of this particular queue, as it is argued that religion is the source of all evil, and Darwinism is the all-effective cure (Dawkins, 2007; Hitchens, 2007). However, this is surely not a universal view, and there are many (and not all believers) who think that although evolution holds exclusively in the scientific realm, there may be more to life – certainly there may be more to the afterlife – than strict science allows (Ruse, 2001, 2010). Science answers many questions, but even Darwinism does not speak to everything! (This is a good thing. As philosophers like Karl Popper (1976) emphasized, a theory that explains everything is probably a theory that explains nothing.)

As far as evolution as a path is concerned, although it is true that one does not have to be a Darwinian to believe in a tree of life, it is still surely true that Darwinism itself commits one to some kind of tree of life (Ruse, 2013). But Darwinism does not specify the details as such. The much-debated question, for instance, as to whether birds are descended from dinosaurs or from reptiles that were not dinosaurs has no bearing on Darwinism as such. What about the above-mentioned speculations that life's history may have involved such things as the moving of genes from one group to another very different one through the agency of viruses and the like? Hence the tree of life might be a bit more netlike than one would have suspected (Doolittle, 1999). Certainly, this is something that was not envisioned by Darwin himself or indeed by many successors who would have called themselves Darwinians. But I doubt that there would be much concern or felt need to repudiate Darwinism. After all, things like this are more extensions than major denials of the traditional picture.

Evolution as cause is as always the crucial issue. Darwinism has and continues to make natural selection absolutely central. If one does not accept that natural selection is the major cause of evolutionary change and that adaptation is the defining characteristic of organisms, then one is no Darwinian. This applies to humans as well (Ruse, 2012). Interestingly, the very first speculations we have of Darwin about natural selection was a discussion about human evolution and how selection leads to bigger brains and greater intelligence. As far as other points are concerned, then perhaps as with paths there is more flexibility. Darwin himself always stressed that selection works for the benefit of the individual over the group. This is very much the stance of sociobiologists in the twenty-first century, and is indeed the basis of Dawkins's metaphor of the 'selfish gene.' But there have always been those who have yearnings toward a group-based form of selection – one of the first indeed was the codiscoverer of natural selection, Alfred Russel Wallace. Perhaps, therefore, with something like this there is room for difference within the fold (Sober and Wilson, 1997).

It was never the claim of Darwin or any subsequent Darwinian that natural selection is the only mechanism of evolutionary change (Ruse, 2006). Darwin himself accepted Lamarckism. This mechanism is today totally discredited, but there are other putative causes of change, and one can certainly allow them some force and still be a Darwinian. There are many debates about how much of an allowance one can give and still subscribe to Darwinism. To a certain extent, it all depends on how far one is modifying traditional views and how far one is standing outside of, or beyond, them. Controversial is the question of 'genetic drift,' the claim by Sewall Wright that evolution can go by chance, even when selection is operating (Coyne et al., 1997). Really strict ultra-Darwinians like Fisher would have very little tolerance for something like drift, thinking it unimportant at best. Others (including most of the American founders of the synthetic theory) found some drift to be quite compatible with their Darwinism. Kimura himself felt (perhaps with good reason) that he was a good Darwinian because he did not propose drift at the physical level but at the molecular level. Darwin himself had said nothing about molecules, and so in proposing molecular drift there was no reason to think that one was being non-Darwinian.

Social Darwinism

Finally, what about social Darwinism? Few want to use that label given its unfortunate connotations with such things as the ruthless business practices of the twentieth century. But still there is a thriving industry among those who want to derive more from evolutionism than mere science (Ruse, 1999, 2009). Julian Huxley (1927) was one person in the twentieth century who wanted his evolution to be the basis of a 'religion without revelation,' and more recently Edward O. Wilson (1998) is another. In one of his prize-winning books, *On Human Nature* (1978), Wilson is explicit in seeing evolution as more than mere science, as a 'myth,' which will absorb or push aside conventional religion:

> Make no mistake about the power of scientific materialism. It presents the human mind with an alternative mythology that until now has always, point for point in zones of conflict, defeated traditional religion. Its narrative form is the epic: the evolution of the universe from the big bang of 15 billion years ago through the origin of the elements and celestial bodies to the beginnings of life on earth.

And so on and so forth up to humans. Little wonder that 'Theology is not likely to survive as an independent intellectual discipline.'

Wilson even finds a place for an evolution-based morality – not the *laissez-faire* of yesterday but a call to biodiversity and species preservation, lest we humans decline and fall in a world impoverished of living nature.

It is certainly not necessary for a scientific Darwinian to be a social or philosophical Darwinian, but a remarkably large number are. The important conclusion is that, however defined, Darwinism has a rich tradition, and although it has never been without its critics, it continues in the twenty-first century as an important and ever developing testament to human reason and creativity.

See also: Evolution, History of; Kin Selection; Microevolution; Sociobiology, History of; Sociobiology: Overview.

Bibliography

Ayala, F.J., 1985. The theory of evolution: recent successes and challenges. In: McMullin, E. (Ed.), Evolution and Creation. University of Notre Dame Press, Notre Dame, France, pp. 59–90.

Bowler, P.J., 1996. Life's Splendid Drama. University of Chicago Press, Chicago, IL.

Coyne, J.A., Barton, N.H., Turrelli, M., 1997. Perspectives: a critique of Sewall Wright's shifting balance theory of evolution. Evolution 51, 643–671.

Daly, M., Wilson, M., 1988. Homicide. De Gruyter, New York.

Darwin, C., 1859. On the Origin of Species. John Murray, London.

Darwin, C., 1871. The Descent of Man. John Murray, London.

Dawkins, R., 1989. The Selfish Gene, second ed. Oxford University Press, Oxford, UK.

Dawkins, R., 2007. The God Delusion. Houghton, Mifflin, Harcourt, New York.

Dennett, D.C., 1995. Darwin's Dangerous Idea. Simon and Schuster, New York.

Dobzhansky, T., 1937. Genetics and the Origin of Species. Columbia University Press, New York.

Doolittle, W.F., 1999. Phylogenetic classification and the universal tree. Science 284, 2124–2128.

Eldredge, N., Gould, S.J., 1972. Punctuated equilibria: an alternative to phyletic gradualism. In: Schopf, T.J.M. (Ed.), Models in Paleobiology. Freeman, San Francisco, CA, pp. 82–115.

Fisher, R.A., 1930. The Genetical Theory of Natural Selection. Oxford University Press, Oxford, UK.

Gluckman, P., Beedle, A., Hanson, M., 2009. Principles of Evolutionary Medicine. Oxford University Press, Oxford.

Gould, S.J., 1982. Darwinism and the expansion of evolutionary theory. Science 216, 380–387.

Haldane, J.B.S., 1932. The Causes of Evolution. Cornell University Press, New York.

Hamilton, W.D., 1964. The genetical evolution of social behaviour. Journal of Theoretical Biology 7, 1–32.

Hitchens, C., 2007. God Is Not Great: How Religion Poisons Everything. Hachette, New York.

Huxley, J.S., 1927. Religion without Revelation. Ernest Benn, London.

Huxley, J.S., 1942. Evolution: The Modern Synthesis. Allen and Unwin, London.

Johnson, P.E., 1991. Darwin on Trial. Regnery Gateway, Washington, DC.

Kimura, M., 1983. Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago, IL.

Lamarck, J.B., 1809. Philosophie zoologique. Dentu, Paris.

Malthus, T.R., 1826. An Essay on the Principle of Population, sixth ed. John Murray, London.

Maynard Smith, J., 1982. Evolution and the Theory of Games. Cambridge University Press, Cambridge, UK.

Nesse, R.M., Williams, G.C., 1994. Why We Get Sick: The New Science of Darwinian Medicine. Times Books, New York.

Paley, W., 1802/1819. Natural Theology (Collected Works: IV). Rivington, London.

Popper, K.R., 1976. Unended Quest: An Intellectual Autobiography. Open Court, LaSalle, IL.

Provine, W.B., 1971. The Origins of Theoretical Population Genetics. University of Chicago Press, Chicago, IL.

Richards, R.J., 1987. Darwin and the Emergence of Evolutionary Theories of Mind and Behavior. University of Chicago Press, Chicago, IL.

Richmond, M., 1988. Darwin's study of the Cirripedia. In: The Correspondence of Charles Darwin, vol. 4. Cambridge University Press, Cambridge, UK. pp. 388–409.

Ruse, M., 1979. The Darwinian Revolution: Science Red in Tooth and Claw. University of Chicago Press, Chicago, IL.

Ruse, M., 1985. Sociobiology: Sense or Nonsense? Second ed. Reidel, Dordrecht.

Ruse, M., 1996. Monad to Man: The Concept of Progress in Evolutionary Biology. Harvard University Press, Cambridge, MA.

Ruse, M., 1999. Mystery of Mysteries: Is Evolution a Social Construction? Harvard University Press, Cambridge, MA.

Ruse, M., 2001. Can a Darwinian Be a Christian? The Relationship between Science and Religion. Cambridge University Press, Cambridge, UK.

Ruse, M., 2003. Darwin and Design: Does Evolution Have a Purpose? Harvard University Press, Cambridge, MA.

Ruse, M., 2005. The Evolution-Creation Struggle. Harvard University Press, Cambridge, MA.

Ruse, M., 2006. Darwinism and Its Discontents. Cambridge University Press, Cambridge.

Ruse, M., 2008. Charles Darwin. Blackwell, Oxford.

Ruse, M. (Ed.), 2009. Philosophy after Darwin: Classic and Contemporary Readings. Princeton University Press, Princeton, NJ.

Ruse, M., 2010. Science and Spirituality: Making Room for Faith in the Age of Science. Cambridge University Press, Cambridge.

Ruse, M., 2012. The Philosophy of Human Evolution. Cambridge University Press, Cambridge.

Ruse, M. (Ed.), 2013. The Cambridge Encyclopedia of Darwin and Evolution. Cambridge University Press, Cambridge.

Ruse, M., Travis, J. (Eds.), 2009. Evolution: The First Four Billion Years. Harvard University Press, Cambridge, MA.

Sepkoski, D., Ruse, M. (Eds.), 2009. The Paleobiological Revolution. University of Chicago Press, Chicago, IL.

Sober, E., Wilson, D.S., 1997. Unto Others: The Evolution of Altruism. Harvard University Press, Cambridge, MA.

Wilson, E.O., 1975. Sociobiology: The New Synthesis. Harvard University Press, Cambridge, MA.

Wilson, E.O., 1978. On Human Nature. Cambridge University Press, Cambridge, MA.

Wilson, E.O., 1998. Consilience. Knopf, New York.

Wright, S., 1931. Evolution in Mendelian populations. Genetics 16 (2), 97–159.

Wright, S., 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proceedings of the Sixth International Congress of Genetics 1, 356–366.

Darwinism: Social

Alain Drouard, National Center for Scientific Research, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the most general sense, social Darwinism refers to the transposition of Darwinian theories from nature to society. The expression appeared in Europe in the 1880s and later in Anglo-Saxon countries, and was first used by the opponents of these ideas, and particularly by critics of the philosophy of Herbert Spencer.

It is difficult to define social Darwinism. From a theoretical point of view, it is a postulate of an analogy between the laws of nature and the laws of society makes it likely to be confused with evolutionism. On the contrary, if one focuses on the practices and the policies, which it advocates, it is tempting to equate it with eugenics, whereas many eugenicists were not social Darwinists.

Social Darwinism has assumed different forms in different times and in different countries. Between 1850 and 1880, liberal social Darwinism held that 'the struggle for life' between individuals of the same species and 'the survival of the fittest' were the fundamental laws of society.

At the end of the nineteenth century appeared what could be called 'selectionist social Darwinism.' The struggle between nations and races replaces henceforth the struggle between individuals as the driving force of history. States need to intervene to counteract the deleterious effects of civilization.

Introduction

The term 'social Darwinism' was coined in Europe in the 1880s and spread rapidly to England and the United States. It negatively referred to the theories and doctrines deriving social laws from the laws of nature. The term was first used to criticize the Spencerian philosophy of 'the struggle for life' and 'the survival of the fittest' rather than Darwin's own theory of natural selection.

Emile Gautier (1880), a militant anarchist, explained that the law of natural selection does not apply to human societies. He described this extrapolation as 'social Darwinism' and contrasted it with true social Darwinism which for him implied reforming society:

> "Instead of the struggle for existence, help for existence; instead of man, humanity; instead of antagonistic interests, universal solidarity" and added that it would be necessary "to collectivize the means of production" but also "to individualize the means of pleasure".

Emile Laveleye (1885) argued in turn that the application of the 'Darwinian laws' to society would result in the destruction of all institutions.

How Can Social Darwinism Be Defined?

Whatever form it has taken, at different times and in different countries, social Darwinism has been characterized, as clearly shown by Hawkins (1997), by a number of assumptions and common proposals:

- The laws of nature apply to all living beings and thus to man living in society;
- These laws of nature are 'the survival of the fittest,' 'the struggle for life,' and 'the laws of heredity';
- Population growth puts pressure on resources, which leads to a struggle for existence of organisms;
- The physical and mental traits that provide an advantage in the struggle for existence can be propagated throughout the population via hereditary transmission;
- The cumulative effects of selection and hereditary transmission over time explain the emergence of new species and the elimination of others.

More precisely, we can distinguish between a broad and a narrow sense of social Darwinism. In a broad sense, that runs the risk of making it appear similar to evolutionism, social Darwinism includes all doctrines and theories that suppose an analogy between the laws of nature and the laws of society. In a narrow sense, it refers to practices and policies implemented with reference to the laws of nature. Social Darwinism thus intersects with eugenics, without merging with it in the sense that eugenicists have not all been 'social Darwinists' (Drouard, 1999).

In the second half of the nineteenth century, social Darwinism had a very real influence on European social sciences before declining at the beginning of the twentieth century, and then was discredited by the use made of it by the national-socialist ideology.

In recent decades, sociobiology has renewed debates over social Darwinism.

Was Darwin a Social Darwinist?

While social Darwinists use the term 'struggle' in the literal sense and make the survival of the fittest its legitimate sanction, Darwin pointed out that he used the term 'struggle for life' as a metaphor that included different levels of oppositions and dependencies. The word 'struggle,' borrowed from the geologist Lyell, means making efforts in difficult conditions and at the same time facing an opponent. In Darwin, the 'struggle for life' is not defined as a direct clash involving individuals but as a set of solidarity mechanisms and clashes precariously balanced, and taking into account the milieu in which individuals and species live and fight. Darwin only used the term 'survival of the fittest' as late as the fifth edition of *The Origin of Species*, to refer to the process of natural selection and not to the struggle for life. Then in *The Descent of Man* (1871)

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32036-0

he derived the emergence of morality in humans from the social instincts of animals.

The moral sense – an extension of social instincts – leads men to feel sympathy for their fellow men and to act in solidarity for their good:

> The aid we feel impelled to give to the helpless is mainly an individual result of the instinct of sympathy which was originally acquired as part of the social instincts but subsequently rendered in the manner previously indicated, more tender and more widely diffused. Nor could we check our sympathy if so urged by hard reason without deterioration in the noblest part of our nature.

The development of social instincts goes hand in hand with that of civilization and it does not necessarily lead to the 'degeneration of civilized people':

> Although civilisation thus checks in many ways the action of natural selection, it apparently favours by means of improved food and the freedom from occasional hardships the better development of the body…

The same is true for intellectual abilities and moral qualities.

But, ambiguously enough, Darwin agreed to some extent with Galton's eugenic ideas and declared that the development of social instincts should not eliminate competition and the struggle for existence:

> … as Mr. Galton has remarked, if the prudent avoid marriage, whilst the reckless marry, the inferior members will tend to supplant the better members of society. Man like every other animal has no doubt advanced to his present high condition through a struggle for existence consequent of his rapid multiplication, and if he is to advance still higher he must remain subject to a severe struggle … There should be open competition for all men; and the most able should not be prevented by laws and customs from succeeding best and rearing the largest number of offspring.

Unity and Diversity of Social Darwinism

Social Darwinism can be defined as a Weltanschauung, i.e., a philosophy of history or a sociology, which not only postulates an analogy between the laws of nature and the laws of society but also affirms that the laws of nature are 'the survival of the fittest,' 'the struggle for life,' and the laws of heredity, and that there should be no opposition to this for the welfare of society.

At the risk of minimizing the great diversity of its forms, there is a consensus among historians to distinguish two major types of social Darwinism:

1. a liberal and individualistic social Darwinism, influential until the end of the nineteenth century, which sees the struggle between individuals as the fundamental law of society;
2. a selective and interventionist social Darwinism, which can be defined as a sociology of struggle, and which emphasizes the struggle between races. This type of social Darwinism overlaps to a certain extent with eugenics and racism.

Liberal Social Darwinism

The first social Darwinism, that of Herbert Spencer (1820–1903), appeared before Darwin's *The Origin of Species* and *The Descent of Man*. It can be defined as an evolutionist philosophy without any connection to biology. Spencer's concept of 'the survival of the fittest' combines natural selection and economic competition between individuals. It does not deal with struggles between groups and races.

While Spencer (1871) referred to population laws, he did not draw the same conclusions as Malthus: the survivors are those who are endowed with a capacity for self-preservation superior to those who cannot participate in life's banquet. Economic competition is thus part of the laws of nature and nothing should hamper them because the survival of the fittest is the only process capable of ensuring human progress.

Population growth not only stimulates competition but is also the driving force behind civilization:

> This constant increase of people beyond the means of subsistence causes, then, a never-ceasing requirement for skill, intelligence and self-control involves, therefore, a constant exercise of these and gradual growth of them.

Demographic pressure is destined to come to an end when its task has been carried out, bringing the process of civilization to its fulfillment. That was not the case in the last third of the nineteenth century.

Even if the changes in English society and increasing state intervention did not correspond to his expectations, Spencer wrote again about the notion of the survival of the fittest in *The Principles of Sociology*, published from 1876 to 1880. From then on he distinguished between two types of modern societies:

1. militant societies characterized by an increasing role of the state, a growing number of civil servants, and faith in authority; and
2. those he thought to be superior forms of society, characterized by the decline of the state and the blossoming of personal freedom through competition between individuals.

Spencer feared in fact that the British society – the model of an industrial society in his eyes – was regressing to the type of a militant society because of the increasing role of the state. He always remained convinced that social progress could only result from a decrease in the constraints placed on individuals.

Liberal and individualistic social Darwinism spread mainly through the Anglo-Saxon countries. In the United States, the economist William Graham Sumner (1840–1910), the major theoretician of this type of social Darwinism, explained that differences between individuals merely reflect natural inequalities: people are not equal in the natural state, and this should be taken into account.

There is an analogy between the natural selection of the fittest in the animal realm and that of the most talented

citizens. As in nature where the strongest win out over the weakest, in society the most talented and hardworking people win out over the lazy. Wealth is therefore the consecration of the successes of gifted people:

> Millionaires are a product of natural selection, acting on the whole body of men to pick out those who can meet the requirement of certain work to be done … It is because they are thus selected that wealth aggregates under their hands both their own and that entrusted to them … They may fairly be regarded as the natural agents of society for certain work… (1914)

Sumner was opposed to any type of state intervention, protectionism, or form of socialism whose goal is to relieve individuals of any society from the difficulties or hardships of the struggle for life and of competition.

This type of social Darwinism, which was very popular at the time when liberal ideas gained support in the Anglo-Saxon countries, proved less successful on the European continent. Opponents did not fail to underline the contradictions and weaknesses of liberal social Darwinism. In the 1890s, the sociologist Lester Ward explained that economic competition did not result in individual advancement, but in big new corporations whose power might need to be controlled, and he criticized the social Darwinists' "fundamental error" that "the favors of the world are distributed entirely according to merit." This first type of social Darwinism should be distinguished from that of the end of the nineteenth century, which is based on a sociology of struggle between groups, collectivities, and races, and calls for state intervention and the control of human procreation.

Selectionist Social Darwinism and Eugenics

Called initially 'sociology of struggle,' selectionist social Darwinism emphasizes the idea of a struggle between different species in nature, to then arrive at the idea of struggle between races.

One of the first theorists of this form of social Darwinism was Walter Bagehot (1826–77) in England. The 'national character' of a people resulted in his eyes from the selection process of the struggle for life, which opposes nations against each other. But if he considered the struggle between groups as the motor of history, Bagehot did not seek to justify an essential superiority of one race or nation over another.

He classified races and nations a posteriori into a hierarchy according to their success in the survival competition. The strongest nations in each historical period were considered to be best.

Other scientists – biologists or physiologists as well as physical-science specialists – did not hesitate to define an a priori hierarchy, whereby races were considered to inherently possess more or less remarkable basic qualities.

This school of thought was most influential in Germany. Its representatives were convinced of the existence of a natural hierarchy of races in which the Aryans were destined to prevail in the struggle for existence and to be dominant.

One of the most important biologists of the nineteenth century and one of the most committed supporters of Darwin, Ernst Haeckel (1834–1919), declared in *Natürliche Schöpfungsgeschichte* (1868), a book reprinted several times before World War I, that the struggle for life could be applied to humans:

> There is everywhere a *bellum omnium contra omnes*, ruthless and relentless in the highest degree. There is nowhere in nature, wherever you look, the idyllic country the poet speaks of… There are on the contrary struggles and a continuous effort to destroy one's neighbor and enemy. Passion and egoism, conscious and unconscious, are the driving forces of life everywhere. Man, in this regard, does not constitute an exception from the rest of the animal kingdom.

At the same period in Austria, Ludwig Gumplowicz (1893) developed the concept of race struggle: "Under the term race struggle we shall designate the struggles of the most varied and heterogeneous ethnic and social units, groups and communities. They are the essence of the historical process."

Karl Pearson was the major advocate for selectionist social Darwinism in the United Kingdom, a social Darwinism that merged with eugenics. He wrote:

> It is at the opposition of Darwinism and medical progress that eugenics seems to provide a key to this situation. As eugenicists we assert the right to live does not connote the right of each man to reproduce his kind. (1912)

Though he was a socialist, he was opposed to revolution and he was in favor of gradual reforms allowing society to be strengthened and to be able to face international competition. In his eyes, eugenics and socialism could be separated – natural selection must give way to artificial selection in order to prevent the unfit from replacing the fit.

At the end of the nineteenth century, this form of social Darwinism has been considered, especially in the United Kingdom and in the United States, as a eulogy of colonization and imperialism in that the strongest nations were supposed to be the fittest ones.

Selectionist social Darwinism also asserts that the natural selection process is flawed or that it no longer functions properly because of the progress of civilization.

This had already been the opinion of Clémence Royer, who introduced the work of Darwin into France and wrote in the preface to the first translation into English of *On the Origin of Species*:

> But also the law of natural selection, applied to humanity, surprisingly and painfully shows how false our political and civil laws as well as our religious morality have been so far. We thus have come to sacrifice what is very strong to what is weak, the good to the bad, the gifted mind and body to the puny and vicious. What is the result of this unintelligent protection granted exclusively to the weak, the sick, the incurable, the bad, and finally to all of the disgraced beings in nature? The problem is that the suffering they experience tends to perpetuate itself indefinitely; that evil increases steadily instead of decreasing, and is increasingly growing at the expense of good.

One can see from this how selectionist social Darwinism intersected with eugenics, which stresses the need to replace natural selection with social selection. It is often confused with

eugenics because of the similarities and converging views between the two streams of thought concerning both the postulate that races and individuals are unequal and the role of heredity.

Well before the eugenics of Galton, doctors had advocated the selection of progenitors to fight against degeneration. In the midnineteenth century, Proper Lucas (1808–85) defended the idea of the selection of progenitors and of the circumstances most favorable for good reproduction. He identified the people to be excluded: the sick 'whose duty it is to abstain': a moral duty which, in certain circumstances, should be a legal obligation.

Coined by Francis Galton, Charles Darwin's cousin, and derived from the Greek words *eu* meaning good and *genos* meaning stock or race, the term *eugenics* refers to "the science of improving stock, which is by no means confined to questions of judicious mating but which, especially in the case of man, takes cognizance of all influences that tend in however remote a degree to give to the more suitable races or strains of blood a better chance of prevailing speedily over the less suitable than they otherwise would have had" (Galton, 1883).

Above and beyond its utopian component manifested in projects for an ideal city-state, one can see in eugenics a "scientific ideology," as Georges Canguilhem defined it (1977), i.e., a discourse mixing and amalgamating proposals of heterogeneous nature, some of which are or claim to be scientific in that they pertain to the science of human heredity, and others which are arbitrary and unprovable value judgments on human groups, categorized and prioritized as 'inferior' or 'superior,' 'gifted,' 'degenerate,' 'retarded' or 'weak-minded.' Like all ideologies, eugenics provides an explanation of the state of modern society and in particular its tendency to 'degeneration' because of the suppression of the 'natural selection' process. But it also postulates society's 'regeneration' and, to this end, proposes remedies.

Thus an artificial selection, fully recognized, must be substituted for the failing natural selection. The means that are advocated either refer to 'negative' eugenics or to 'positive' eugenics. 'Negative' eugenics aims at preventing people with hereditary diseases and serious deformities from getting married, by resorting to means ranging from segregation to sterilization. 'Positive' eugenics aims at encouraging 'gifted' or 'strong' individuals – the carriers of favorable physical and intellectual characteristics – to reproduce themselves as much as possible.

Georges Vacher de Lapouge (1854–1936) first introduced into the French language the word *eugénique*, which is the French term for "eugenics," created by Galton in 1883. Before designating an applied science and a technique, the term appeared as an adjective ('eugenic races,' 'eugenic families') or as a name for individuals with above-average hereditary skills ('the eugenics'). He contrasted 'the eugenics' to 'the degenerates': "There are families of degenerates … In others, skills such as health, strength, and beauty come from birth … These are the eugenics, and eugenics is the boon of heredity as degeneration is its curse." (1888b)

Eugenics thus refers to a hereditary anthropotechny as opposed to Lamarckism. Indeed, Vacher de Lapouge stressed the hereditary character of inequality:

> Individuals are not only unequal but their inequality is hereditary, classes and nations and races are not only unequal but each of them cannot undergo absolute improvement and the elevation of the average is the consequence of the extermination of the worst elements, of the spread of the best elements, in a word, of unconscious or subconscious selection. Human evolution is the result of this inequality. (1888a)

While Vacher de Lapouge borrowed from Darwin's theory the concepts of selection and the struggle for existence, he freely interpreted them by incorporating them into his problematic of social selection based on a premendelien conception of the 'laws of heredity.'

He explained that in modern societies natural selection no longer played its role and had been replaced by what he called *social selections* leading to the eradication of the fittest and which he qualified as military, political, religious, moral, economical, professional, and urban selection. Consequently, an artificial selection must be substituted for it so as to prevent society and civilization from being weighed down by degenerate and inferior people. He therefore advocated *selectionism* as the only remedy for 'senile democracy' and combined it with socialism:

> Selectionism as a practical doctrine consists in correcting the adverse consequences of natural selection and multiplying types accepted as the most beautiful and the best. There are many analogies with socialism in its purpose which consists in correcting the natural consequences of economic development, according to a determined ideal of social perfection. (1899)

For humans, social selection overrides natural selection. But it is clear that social selections are going in the wrong direction as they lead to the disappearance of the best elements, 'the eugenics.' It is "the mechanism of decadence," a process of "regressive selection eliminating the upper elements."

The analysis of social selections leads "to the most absolutely pessimistic conclusions." However, the dysgenic development of modern societies is not a fatality and the slope of decadence can be stopped "only if a systematic selection is opposed to the destructive and unregulated selection that is putting humanity at risk" (Vacher de Lapouge, 1896).

Otto Ammon (1842–1916) was the main representative in Germany of the anthropological and selectionist school of thought of which Vacher de Lapouge was the leader in France. Like the biologists Ernst Haeckel and Heinrich Ziegler, theorists of German social Darwinism, Ammon criticized socialism in the name of Darwinism (1891). He proposed "to base the foundations of the social order on the natural sciences" (1895) and more specifically on Darwinism:

> It is a sublime feature of the new conception of nature established by Darwin that it is compatible with all dominating historical forces, both of the material and intellectual or religious orders.

In his major work Die Gesellschaftsordnung und ihre natürliche Grundlagen (1895), Ammon defined what he termed 'regressive selection' as resulting in a gradual decrease in 'Germanic elements' or 'Aryans' and therefore in

a 'brachycephalization' of the population. Vacher de Lapouge in *Les sélections sociales* also provided a pessimistic analysis of urban selection, the most dysgenic selection of all, since cities qualitatively depopulate the countryside and destroy the 'eugenic' elements that they attract.

Ammon referred to the work of Galton to argue that inequalities in mental abilities are hereditary. Talent and genius are hereditary, and the social order is there to ensure their perpetuation:

> "The most remarkable result of the evolution of social life" is none other than "the isolation of the privileged classes from the mass of the population".

He feared above all that the hereditarily superior elements would be corrupted by urban life, would no longer be able to reproduce, and would thus disappear after two or three generations.

However, he developed a less pessimistic vision than Vacher de Lapouge of the evolution of society, as he maintained that the prevailing social order expressed and reproduced the natural inequalities between people.

According to Ammon (1895), the social order is inherently unequal, and this inequality is in the order of things:

> The social order is based on inequality, and inequality is not something that can be suppressed: it is as inseparable from the human race as birth and death; it is as fixed as mathematical truths and as eternal as the laws which regulate the functioning of our solar system".

If, due to 'defects associated with an exclusively intellectual culture and a sedentary life' the superior classes are threatened with eradication for the benefit of the inferior classes, the latter can better themselves through education and food, thus contributing to the renewal of the superior classes: "The existence of social classes promotes the working of natural selection, and also restricts 'panmixia,' i.e., the pairing of individuals without preselection, and favors a much more frequent production of highly endowed individuals" (Ammon, 1895).

At the end of the nineteenth century, this form of social Darwinism has been considered, especially in the United Kingdom and in the United States, as a eulogy of colonization and imperialism in that the most powerful nations were supposed to be the fittest ones.

Social Darwinism Today

From the beginning of the twentieth century, social Darwinism has been the subject of harsh criticism; for example, by Novicow (1910), who did not hesitate to denounce "a completely false theory … My criticism tends only to demolish the proposition that collective homicide can result in the progress of mankind."

The decline of social Darwinism, which was already perceptible after World War I, has on the one hand been ascribed to the use that was made of it, especially to justify imperialism, and on the other hand to the atrocities committed by the Nazis in the name of eugenics (mass extermination, euthanasia).

Even if nowadays social Darwinism is not explicitly referred to any more, it is still part of the intellectual scene, as the rise of sociobiology has shown. The latter is a scientific discipline founded by the zoologist Edward O. Wilson, which gave rise to serious ideological controversies in the late 1970s and early 1980s. Wilson (1975) defined it as "the systematic study of the biological basis of all social behavior."

The biological basis is provided by the modern theory of evolution as it has developed from Darwin's work and from contemporary genetics. Social behavior is the result of evolution – it has been selected and is therefore determined. In a sense, sociobiology is a form of biological determinism without being a form of reductionism at the same time, since Wilson contends that humans are the product of the joint influences of heredity and of the environment. To Wilson *biological* only meant that there are genetic predispositions contributing to the shaping of social behavior. To his opponents *biological* meant that social behavior was *biologically determined*. In the wake of the question of animal social instincts dealt with by Darwin in *The Descent of Man*, sociobiology strove to shed new light on the problem of altruism.

In his book *The Selfish Gene* (1976), Richard Dawkins, another representative of this school of thought, asks questions: "While natural selection is a selfish process which endows certain individuals with a superior capacity of reproduction, how can one explain the kind of altruism which is seen among social insects, and how can natural selection favor deeds harmful to those who do them?"

The answer given by sociobiology is that altruistic deeds toward relatives could favor the survival and the dissemination of the altruists' genes, which the parents have in common with them. In the same way, the mother instinct and the care given to the child are explained by social biologists as being the best possible way of arranging gene transmission from one generation to the other.

See also: Darwin, Charles Robert (1809–82); Darwinism; Eugenics as an International Movement; Eugenics, History of; Evolution, Natural and Social: Philosophical Aspects; Evolutionary Social Psychology; Evolutionary Sociology; Human Ecology; Social System; Spencer, Herbert (1820–1903).

Bibliography

Ammon, O., 1891. Der Darwinismus gegen die Sozialdemokratie. Anthropologische Plaudereien, Verlagsanstalt und Druckerei, A.G., Hamburg.

Ammon, O., 1895. Die Gesellschaftsordung und ihre natürlichen Grundlagen: Entwurf einer Sozial-Anthropologie zum Gebrauch für alle Gebildeten, die sich mit sozialen Fragen befassen. Fischer, Jena, Germany.

Bagehot, W., 1867. Physics and Politics. Longmans, Green, London.

Bernardini, J.-M., 1997. Le darwinisme social en France (1859–1918). Fascination et rejet d'une idéologie. CNRS, Paris.

Canguilhem, G., 1977. Qu'est-ce qu'une idéologie scientifique?. In: Idéologies et rationalité dans l'histoire des sciences de la vie. Vrin, Paris, pp. 33–45.

Clark, L., 1984. Social Darwinism in France. University of Alabama Press, AL.

Darwin, C., 1859. On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life. Murray, London.
Darwin, C., 1871. The Descent of man. In: Relation to Sex. Murray, London.
Dawkins, R., 1976. The Selfish Gene. Oxford University Press, Oxford.
Drouard, A., 1999. L'eugénisme en questions L'exemple de l'eugénisme 'français. Ellipses, Paris.
Galton, F., 1883. Inquiries into Human Faculty and its Development. Macmillan, London.
Gautier, E., 1880. Le Darwinisme Social. Dervaux, Paris.
Gumplowicz, L., 1893. La lutte des races Recherches sociologiques. Librairie Guillaumin, Paris.
Haeckel, E., 1868. Natürliche Schöpfungsgeschichte. Georg Reimer, Berlin.
Hawkins, M., 1997. Social Darwinism in European and American Thought (1860-1945) Nature as Model and Nature as Threat. Cambridge University Press, New York and Cambridge.
Hofstadter, R., 1983. Social Darwinism in American Thought. University of Pennsylvania Press, Philadelphia, PA.
Laveleye, E., Spencer, H., 1885. L'Etat et l'individu ou darwinisme social et christianisme, Tirage à part de La Revue internationale. Pellas, Florence.
Lucas, P., 1847. Traité philosophique et physiologique de l' hérédité naturelle dans les états de santé et de maladie du système nerveux. Baillière, Paris tome 1 (tome 2, 1850).
Novicow, J., 1910. La critique du darwinisme social. Alcan, Paris.
Pearson, K., 1912. Darwinism, Medical Progress and Eugenics. The Cavendish Lecture. Dulau, London.
Royer, C., 1862. Préface de la traduction de L'Origine des espèces. Guillaumin, Paris.
Spencer, H., 1868. Social Statics or the Conditions Essential to Human Happiness Specified and the First of Them Developed. Williams and Norgate, London.
Spencer, H., 1871. The Principles of Biology, first ed., vols. 2. Appleton and Cie, New York (Williams and Norgate, London. pp. 1865–1867).
Spencer, H., 1876. Principles of Sociology. Williams and Norgate, London.
Sumner, W.G., 1914. The Challenge of Facts and Other Essays. Yale University Press, New Haven.
Vacher de Lapouge, G., Janvier 15, 1888a. De l'inégalité parmi les hommes. Revue D'anthropologie 3 (III).
Vacher de Lapouge, G., Mars 15, 1888b. L'hérédité dans la science politique. Revue D'anthropologie 3 (III).
Vacher de Lapouge, G., 1896. Les Sélections Sociales. Fontemoing, Paris.
Vacher de Lapouge, G., 1899. L'Aryen, Son Rôle Social. Fontemoing, Paris.
Ward, L., 1907. The establishment of sociology. American Journal of Sociology 12, 581–587.
Wilson, E.O., 1975. Sociobiology the New Synthesis. Harvard University Press, Cambridge, MA.

Data Bases and Statistical Systems: Applied Social Research

Joseph M Gasper, Abt Associates, Bethesda, MD, USA
Kathryn A Henderson, Westat, Rockville, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, we discuss secondary data sets in the applied social and behavioral sciences. First, we provide an overview of the history of secondary data sets. Second, we discuss the wide availability of secondary data sets, including efforts to archive data sets and make them more available to researchers and easier to use. Third, we highlight the benefits of using secondary data sets for applied research. Finally, we provide an overview of eight secondary data sets that we believe have been the most influential in the applied social and behavioral sciences.

Introduction

Secondary data analysis has achieved a prominent place in the applied social and behavioral sciences. *Primary data* are data that were collected by a researcher or team of researchers for the specific study or analysis under consideration. In contrast, *secondary data* are data that were not collected for a specific study or analysis but were collected for the explicit purpose of being analyzed by other researchers. Secondary data sets are usually sponsored by federal agencies or other organizations and often designed to address a broad set of research objectives or questions of interest to the sponsors. Many of the sponsors lack the capability to analyze the data they collect and spend considerable resources on training and funding opportunities to encourage other researchers to use and publish from the data. Researchers can use secondary data sets to answer the research questions that fall within the scope of the secondary data sets or, given the breadth of information available in many secondary data sets, other research questions that sponsors had not considered. Secondary data also include data that were collected by researchers for their own analysis but are later made available to the public to replicate findings and test new theories.

The current popularity of secondary data sets can be explained by several factors, including limited resources for funding primary data collection, demand for scholarly productivity, and the availability of high-quality data sets. Several decades ago, most graduate students and faculty collected and analyzed their own data for a study. However, most graduate students and junior faculty today do not have the financial resources or the time to collect their own data for their dissertation or study. Second, the 'publish or perish' mentality of some universities makes secondary data sets an attractive resource for junior faculty because they can publish an article more quickly than if they had to collect their own data. Finally, there are a large number of high quality, nationally representative, longitudinal secondary data sets publically available that are funded by federal agencies including the U.S. Census Bureau, the Department of Education, the Bureau of Labor Statistics, and the Department of Health and Human Services, to name just a few. The sheer breadth and depth of information that is available in most of these large-scale secondary data sets dwarfs what any individual researcher could collect for a specific study. These nationally representative and longitudinal data sets lend themselves to answering current research questions and using cutting-edge methods demanded by high-quality scientific research. In addition, a large number of data sets originally collected as primary data sets are now made publically available to share outside of the team of researchers that originally collected them for their own use.

In this article, we discuss secondary data sets in the applied social and behavioral sciences. First, we provide a brief overview of the history of secondary data sets in the applied social and behavioral sciences. We show that several simultaneous projects sponsored by federal agencies during the 1960s and 1970s paved the way for the large number of high-quality secondary data sets available today. Second, we discuss the wide availability of secondary data sets, including efforts to archive data sets and make them more available to researchers and easier to use. Third, we highlight the benefits of using secondary data sets for applied research. Finally, we provide a brief overview of eight secondary data sets or federal survey programs that we believe have been the most influential in the applied social and behavioral sciences. These eight data sets highlight the many flavors of secondary data sets and represent a diverse array of substantive research areas, including but not limited to labor, education, marriage and family formation, and health.

History of Secondary Data Sets

The use of secondary data sets is widespread in disciplines such as economics and sociology and has been steadily increasing in psychology. However, the availability of a large number of high-quality secondary data sets is a relatively recent development. Several simultaneous developments in the 1960s and 1970s paved the way for many of the most widely used secondary data sets today. The 1960s saw an expansion of federal government funding of social programs and with this came surveys to monitor the effects of these economic and social investments. In the 1960s and 1970s, several federal agencies launched national studies that would produce the first secondary data sets. The earliest studies were the Occupational Changes in a Generation Surveys of 1962 and 1973 and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10536-7

national surveys of social mobility conducted as supplements to the March 1962 and March 1973 Current Population Surveys (CPS) of the U.S. Census. In the mid-to-late 1960s, one of the most influential federal survey programs – the National Longitudinal Surveys (NLS) of the Department of Labor – first interviewed four groups of young and old men and women to form the original NLS cohorts. These cohorts were selected because each faced important labor market decisions, which were of special concern to policy makers at the time. In 1968, the Office of Economic Opportunity of the Department of Commerce (now the U.S. Census Bureau) sponsored the first wave of the Panel Study of Income Dynamics (PSID) to assess Lyndon Johnson's War on Poverty. The study included a nationally representative sample of over 18 000 individuals living in 5000 families. The PSID is ongoing and is the world's longest running longitudinal study.

In the early 1970s, a number of additional federal survey programs were launched. In 1972, the Department of Education collected the first wave of data for the National Longitudinal Study of 1972 (NLS-72). Participants in the study were selected when they were seniors in high school in the spring of 1972. The data contain information from follow-up surveys in 1973, 1974, 1976, 1979, and 1986 as well as their high school records and postsecondary transcripts (collected in 1984). The NLS-72 is considered the 'grandmother' of the more than 30 surveys conducted by the National Center of Education Statistics in the nearly four decades since its inception. Just a few years later, the NLS launched the National Longitudinal Survey of Youth 1979 (NLSY79), which included a national sample of about 13 000 respondents aged 14–22 when they were first interviewed in 1979. These respondents were interviewed 24 more times with the latest wave in 2010.

Not all secondary data sets start out as major studies funded by federal agencies. Several of the most influential secondary data sets were originally intended to be primary data sets for use only by the researchers who originally collected them. Perhaps the best example of this is the Wisconsin Longitudinal Study (WLS). The WLS began in 1957 as a survey by education professor J. Kenneth Little at the University of Wisconsin (UW) designed to understand the educational plan of Wisconsin high school seniors. Rising demand for college education and concern over technological competition with the Soviet Union was a major concern in the 1950s. In 1962, sociology professor William H. Sewell learned that the surveys were stored in the basement of a UW administrative building. Sewell resurrected the surveys and selected a sample of one-third or 10 317 for further study. The sample has been followed since. The WLS, which started out as a primary data set on high school seniors' college plans, became one of the most prominent secondary data sets used to study health and aging in the United States.

Availability of Secondary Data Sets

Most secondary data sets are funded by federal agencies. These federal agencies contract with private survey research organizations (such as Westat or Research Triangle Institution) or university survey researcher centers (such as the University of Michigan's Survey Research Center or the National Opinion Research Center at the University of Chicago) to assist in the design, collection, cleaning, and preliminary analysis of the data. As each wave of data is collected, it is made available by the contractor to the public. Most often, the contractors maintain Web sites where users can download the data. Many of these Web sites allow users to download the entire data set and, in cases of very large data sets consisting of multiple files, include tools that enable users to select specific cohorts and variables for download. Downloads usually include SAS, SPSS, or STATA code for extracting the data and producing a readable file. Users can save 'tagsets' of variables for download at a later time or for adding variables at a later time or that can be accessed by other members of the research team. Many Web sites also include custom analysis tools that allow users to run their own simple analyses for exploring the data, including means, tabulations, and cross-tabulations. Most of the Web sites also include data documentation, including technical documentation on sampling, copies of the questionnaires, and instructions for analyzing the data, including use of weights and appropriate methods for calculating standard errors in the presence of complex survey design. Technical support is also usually available from the contractor through e-mail or telephone.

The federal government, universities, and private organizations are increasingly taking steps to meet the very real demand for access to high-quality social science data for research. In 2009, the federal government launched data.gov, a Web site that is designed to provide public access to all of the publically available data collected by the federal government. Users can search over 85 000 secondary data sets by topic area or keyword. The Web site provides access to all of the influential secondary data sets collected by federal agencies through external links to the agency Web sites to download the data. The data.gov Web site also includes access to a wealth of data from evaluations and other studies conducted by contractors.

Perhaps, one of the best known efforts to provide easy access to secondary data is the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan. ICPSR is a leader in providing storage of and access to secondary data sets, training in secondary data sets, and education in cutting-edge methods for analysis of secondary data sets. ICPSR hosts 16 specialized collections of data on topics such as education, aging, criminal justice, substance abuse, among others, and archives over 500 000 files related to secondary data sets. ICPSR is well known for its Summer Program in Quantitative Methods of Social Research, available to students and faculty, which includes courses in research design, statistics, and data analysis. Another well-known source of secondary data is the Murray Research Archive at Harvard University, a multidisciplinary secondary data archive. Founded in 1976 with a grant from the Ford Foundation, the Murray Center focuses on acquiring data from studies that emphasize the entire life course. The collection currently contains data from over 52 000 studies available with a user agreement. The Murray Center plays a key role in making primary data available as secondary data to other researchers.

The private sector has also realized the demand for secondary data sets that are easy to use. The newest player in this arena is Sociometrics, a for-profit research and

development firm specializing in making secondary data sets easier for researchers to access. Sociometrics has compiled secondary data sets, documentation, and technical support into one place for researchers to access. The company's signature product is the Social Science Electronic Data Library (SSEDL), which contains nine archives with over 680 data sets from 285 studies available to researchers. Sociometrics offers subscriptions to the data archive that provide unlimited access to all data downloads, supporting documentation, and technical assistance.

There are rich funding opportunities available for researchers who use specific secondary data sets. Many of the federal agencies and organizations that collect secondary data have neither the capacity to nor interest in analyzing them. As such, they seek to encourage other researchers to use the data as much as possible through funding training and workshops and calls for proposals that encourage or even require the use of specific secondary data sets. For example, with funding from the National Science Foundation, the American Educational Research Association (AREA) provides small grants and training for researchers to conduct education policy and practice studies using many of large-scale secondary data sets sponsored by the Department of Education and other federal agencies. The Department of Justice, which funds the criminal behavior questions in the National Longitudinal Survey of Youth 1997 (NLSY97), has recently supported a wide array of research on criminal justice topics that use the NLSY97. This research has focused on the relationship of adolescent employment to delinquency, patterns of criminal recidivism over the life course, and relationships between family structure and delinquency. The National Institutes of Health (NIH) regularly solicits grants under its various programs, including R01, R09, and R21, which call specifically for the analysis of national secondary data sets. These are just a few of the examples of how sponsors of secondary data sets provide support to social science researchers for their analysis.

Use of Secondary Data for Applied Research

Secondary data sets provide a number of opportunities for applied researchers to conduct research that might otherwise be prohibitive due to the cost or time required. Contemporary secondary data sets, especially large-scale data sets sponsored by federal agencies, have several features in common. Most secondary data sets are nationally representative. Estimates from these data, when using the sampling weights provided, represent the U.S. population allowing for national estimates on a range of measures, including labor market participation, educational attainment, family formation behaviors, and public attitudes and opinions. They also allow for comparisons across different communities within the nation. For example, secondary data sets such as the PSID and the CPS enable researchers to examine differences in benefit receipt across states with different eligibility criteria and participation requirements. Finally, because they are large and nationally representative and contain a wealth of information, secondary data sets are sometimes used to derive comparison groups for program evaluations using techniques such as propensity score matching.

Most secondary data sets oversample or make special efforts to track populations of particular interest to researchers or policy makers. These include women, African-Americans, Hispanics, high school dropouts, military personnel, and households under the poverty line. Such oversampling ensures that the sample includes a large enough number of respondents in each of these important groups so that researchers can conduct analyses separately for each group, thereby facilitating comparisons among them.

Many secondary data sets are longitudinal, following the same individuals over several years or even decades. The backbone of many longitudinal secondary data sets is event histories. Secondary data sets often collect detailed event histories about respondents' life events, such as employment, school enrollment, marriage, childbirth, criminal behavior, government assistance, program participation, alcohol and drug use, and residential moves. At each interview, respondents are asked about events that happened since the last interview including the dates of such events. This strategy ensures a complete enumeration of important events even if a respondent missed an interview. Using these data, it is possible to construct monthly, weekly, or even daily histories of important events in respondents' lives. Such event histories allow researchers to examine trajectories over time and to investigate the causes and consequences of different patterns of development. For example, it is possible to know whether a childbirth preceded or followed a spell of dropout from high school.

In addition to collecting information from the primary respondent, many secondary data sets include data collected from other important individuals related to the respondent or influential in the respondents' life. These include surveys of spouses, biological or adoptive children, parents, teachers, school administrators, school counselors, and friends. The availability of social network data is a significant development in secondary data analysis. The National Longitudinal Study of Adolescent Health (Add Health) asked respondents to nominate five male and five female friends to be surveyed. Researchers are able to identify pairs and examine peer influence, pair formation and dissolution, relationship event sequencing, and relationship symmetry. Additionally, secondary data sets often include data collected from sources other than surveys, including school transcripts and medical examinations. Genetic information is increasingly becoming available in secondary data. Secondary data sets are often geocoded to provide rich contextual data such as school or neighborhood characteristics. It is often possible to link secondary data sets to administrative data on respondents, such as Social Security records or Medicare claims data.

Secondary data sets also facilitate comparisons across groups. Data sets that include complex stratified sampling techniques and data collected across wide geographic areas allow researchers to situate individuals within the larger context of neighborhoods, schools, and states through multilevel modeling. Several secondary data sets contain information from different historical cohorts, such as the National Center for Education Statistics (NCES) Surveys, to facilitate comparisons of economic, social, and cultural indicators across generations. Finally, a number of secondary data sets have 'sister' data sets based in other countries specifically designed to

permit comparisons across countries and over time. For example, a number of studies modeled after the PSID have been conducted in other countries facilitating international comparisons. These sister studies include: the German Socio-economic Panel, the British Household Panel Survey, the Household Income and Labour Dynamics Survey of Australia, and the Survey on Household Income and Wealth (Italy).

Despite the many advantages of secondary data sets, potential users should consider a number of drawbacks. First, many secondary data sets can be very challenging to work with and require a steep learning curve. Many of the survey instruments that underlie large-scale secondary data sets use complex skip patterns and all of the items of interest may not be asked of all respondents in all years for analysis. Users must familiarize themselves with the questionnaires, sample design, appropriate use of sampling weights, and calculation of standard errors in the presence of a sample design that often involves stratification and multistage (cluster) sampling. Second, because secondary data sets are available to any researcher, it is possible for other researchers to 'scoop' an idea and publish results prior to you. Finally, at a general level, the reliance on secondary data sets means that much of the scientific knowledge about specific topics is based on a disproportionate number of secondary data sets and is therefore subject to its limitations.

Overview of Secondary Data Sets

In this section, we provide a brief description of several secondary data sets that we believe have been influential in applied social and behavioral sciences. It is not possible to cover the literally hundreds of secondary data sets that are commonly analyzed; therefore we have chosen to highlight eight large-scale data sets that have served as a basis for extensive applied research. These data sets cover a broad array of topical areas and highlight the many strengths and weakness of secondary data sets. Each description includes an overview of the study design, content of the questionnaires, uses of the data set, and information on how to access it.

U.S. Census Bureau Data

The U.S. Census Bureau collects data on American households through two surveys. These are the American Community Survey (ACS) and the CPS.

The ACS is an ongoing statistical survey, sent to approximately 250 000 addresses monthly (or 3 million annually). It regularly gathers information previously contained only in the long form of the decennial census. It is the largest survey other than the decennial census that the Census Bureau administers. The ACS began in 1995 out of a desire to reduce the burden on Americans of the decennial census by shortening the census form and relegating collection of more detailed demographic, housing, social, and economic characteristics to a separate survey administered to a subset of households. The ACS was rolled out gradually and was fully implemented in 2005.

The goal of the ACS is to provide communities the information they need to plan investments and services. Information from the survey helps determine how more than $400 billion in federal and state funds are distributed each year. The survey collects data on demographic characteristics, household composition, income and benefit receipt, access to health insurance, education and employment, and housing. Unlike other surveys discussed here, participation in the ACS is mandatory for households that have been selected. Recipients who do not respond to the mailed request may receive a phone call or visit from a Census staff person. Surveys can be completed either online or by paper.

ACS data are publically available in 1-year, 3-year, and 5-year estimates. The 1-year estimates are released annually and provide data for areas with populations of 65 000 or more. The 3-year estimates provide aggregate information over 36 months on communities with populations of 20 000 or more. The 5-year estimates provide aggregate information over 60 months on all areas in the United States. Users can access information about the survey and methodology and the data and documentation on the ACS Web site at: http://www.census.gov/acs/www/.

The CPS is a statistical survey conducted by the U.S. Census Bureau for the Bureau of Labor Statistics to provide monthly estimates of the number of employed and unemployed people in the United States. Beginning in 1940, data were collected from approximately 60 000 households. A household is interviewed for four successive months, then not interviewed for 8 months, then returned to the sample for 4 months after that. An adult member of each household provides information about the employment status of each member of the household 15 years of age or older as of the calendar week containing the 12th day of the month. The survey also collects data on job characteristics, such as occupation, industry, wages and hours worked, job tenure, union membership, and education level.

Beginning in 1948, the CPS has included supplemental questions (at first, in April; later, in March) on annual income and benefit receipt. Other supplement topics include after-tax money income, benefits that are not cash, displaced workers, job tenure, occupational mobility, temporary work, adult education, volunteering, tobacco use, food availability, fertility, veteran information, and other related topics.

Monthly CPS data are available from January 1994 to the present, along with detailed information about the methodology on the CPS Web site at: http://www.census.gov/cps/data/.

National Longitudinal Surveys

The NLS, sponsored by the Bureau of Labor Statistics, are a set of studies that track the labor market experiences and life events of cohorts of men and women. The National Longitudinal Survey of Young Men is one of the original NLS cohorts. It includes 5225 men who were at ages 14–24 when they were first interviewed in 1966. Interviewers with this cohort stopped in 1981. The National Longitudinal Survey of Older Men also began in 1966 and includes a sample of 5020 men who were at ages 45–59 at the time. Although this survey was also discontinued in 1981, interviews with living respondents and widows or next-of-kin were conducted in 1990. The National Longitudinal Survey of Young Women includes women who were at ages 14–24 in 1968. The National Longitudinal Survey of Mature Women includes women who were at ages 30–44 in 1967. This study was completed in 2003. The NLSY79 is a national sample

of men and women born between 1957 and 1964 and who were at ages 14–22 when they were first interviewed in 1979. The original respondents were followed for 24 waves with the most recent in 2010. Begun in 1986, the NLSY79 Children and Young Adults (NLSY79-C) is a study of the biological children of female respondents of the NLS79. Finally, the NLSY97, sometimes referred to as the 'new NLSY,' is a nationally representative sample of youth born between 1980 and 1984 who were at ages 12–17 when they were first interviewed in 1997. Youth are interviewed annually and this study is still ongoing. Below, we provide a more detailed description of the NLSY97.

Although primarily sponsored by the U.S. Bureau of Labor Statistics to understand labor force participation, education, government program participation, and other topics, the NLSY97 receives funding from several federal agencies that sponsor sets of questions within the survey on other topics. These include the questions on crime and delinquency sponsored by the U.S. Department of Justice, questions on pregnancy and fertility sponsored by the National Institute of Child Health and Human Development, school surveys sponsored by U.S. Department of Education, and cognitive tests sponsored by the U.S. Department of Defense.

The NLSY97 is a nationally representative longitudinal survey of youth who were 12–16 years old as of 31 December 1996. The NLSY97 sample is composed of two independent probability samples: (1) a cross-sectional sample of 6748 youths who are representative of the noninstitutionalized population of youths in the United States, and (2) an oversample of 2236 black and Hispanic youths. The cohort was selected using these two samples in order to meet the survey-design requirement of providing enough black and Hispanic respondents for statistical analyses. Original Wave 1 respondents continue to be interviewed annually. As of this writing, 14 waves (through 2010) of data are available. NLS surveys are well respected due to their relatively high response rates. In the case of the NLSY97, about 90% of the original cohort responded in Wave 10. In addition to annual youth interviews, data were collected from a parent interview in 1997, school surveys in 1996 and 2000, and school transcripts when respondents left high school or turned age 18.

Although primarily designed to collect information on labor and education, the NLSY97 collects fairly detailed information on a vast number of other topics, including delinquency and crime, substance use, fertility and childbearing, marriage and cohabitation, health, income, government program participation, and political participation, to name a few. Detailed monthly event histories are also available for arrests, pregnancies, births, marriages, program participation, and residential moves.

The Center for Human Resources Researcher at Ohio State University has created the NLS Investigator, an online data extraction tool for the NLSY97 data. It can be accessed at: https://www.nlsinfo.org/investigator/pages/welcome.jsp.

National Longitudinal Study of Adolescent Health

Add Health is a longitudinal study of a nationally representative sample of adolescents aimed at understanding causes and consequences of adolescent health and health risk behaviors. A sample of high schools and their feeder middle- or junior-high schools was selected through a stratified, school-based, cluster sampling design.

A random subsample of 20 745 adolescents in grades 7–12 during the 1994–95 school year was selected to complete In-Home Questionnaires for Wave 1. The respondents were followed into young adulthood with three follow-up in-home interviews, in 1996, 2001–02, and 2007–08. Add Health collects data on respondents' social, economic, psychological, and physical well-being with additional data on the family, neighborhood, community, school, friendships, peer groups, and romantic relationships. The fourth wave of interviews included collection of biological data to provide an opportunity to understand the link among social, behavioral, and biological characteristics in health during the transition to adulthood. Other sources of data include questionnaires for parents, siblings, classmates, and school administrators and interviews with romantic partners. The data also include geocode information that allows for links to neighborhood and community level data.

Add Health was developed in response to a mandate from the U.S. Congress to fund a study of adolescent health. Add Health is directed by Kathleen Mullan Harris (UNC, Chapel Hill), with fieldwork done by National Opinion Research Center (Waves I and II) and Research Triangle Institute (Waves III and IV). Add Health is funded by the National Institute of Child Health and Human Development (NICHD) with co-funding from 23 other federal agencies and foundations.

Add Health data can be accessed through the ICPSR Web site at: http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/21600/datasets. Information about the data collection and sampling, codebooks, questionnaires, and lists of publications that use the data can be found at http://www.cpc.unc.edu/projects/addhealth.

The National Center for Education Statistics Surveys Program

The NCES Surveys Program includes 51 surveys and data sets in seven areas related to educational issues. Of particular interest to social scientists who study human development are the Early Childhood Longitudinal Studies (ECLS) Program, the first national study of children from birth through childhood, and the Secondary Longitudinal Studies Program, which include several national longitudinal studies designed to track students from early childhood through postsecondary education and adulthood. The early childhood studies include the Early Childhood Longitudinal Study, and the Kindergarten and Birth Cohorts (ECLS-K and ECLS-B, respectively); the secondary studies include the National Longitudinal Study of the High School Class of 1972 (NLS:72), High School and Beyond (HS&B), the National Education Longitudinal Study of 1988 (NELS:88), the Education Longitudinal Study of 2002 (ELS:2002), and the High School Longitudinal Study of 2009 (HSLS:09). Data collection for the latter two studies is still ongoing. Although the primary focus of these studies is on educational processes and outcomes, they contain a rich set of variables pertaining to nonschool factors of interest to social scientists concerned with a diverse array of topics other than education and warrant inclusion in this article. Below, we briefly highlight

the most recent of these longitudinal studies: HSLS:09 and its connection to the other studies.

HSLS:09 is the latest generation of longitudinal studies designed to track students as they transition from high school to postsecondary education or work and adulthood. HSLS:09 is a nationally representative sample of approximately 24 000 students who were in the ninth grade in 2009. The study uses a two-stage sample design in which 944 public and private high schools were selected in the first stage and 25 ninth-graders per school were selected in the second stage. High school dropouts are identified and followed. This overcomes a known weakness of school-based samples for studying high-risk groups, such as high school dropouts, who tend to be more difficult to capture than in a household survey.

HSLS follows students from high school into post-secondary education and adulthood. Students were surveyed in the fall of ninth grade (2009), in the spring of eleventh grade (2012), and in the spring after high school graduation (2013). Future data collections are planned to take place 2 years after the expected high school graduation year (in 2016) and again in 2021. HSLS also includes surveys of parents, math and science teachers, school administrators, and counselors as well as mathematics assessments.

All NCES data sets can be accessed using the Data Analysis System at: https://nces.ed.gov/das. Users can conduct online analysis or download data sets and extraction programs.

Fragile Families and Child Wellbeing Study

The Fragile Families and Child Wellbeing Study is a study of approximately 5000 children born to 'fragile families' in 20 urban areas in the United States between 1998 and 2000. 'Fragile families' are families composed of unmarried parents and their children, which are typically at greater risk of breaking up and living in poverty than two-parent families. The goal of the Fragile Families data is to address four primary research questions: (1) What are the conditions and capabilities of unmarried parents, especially fathers? (2) What is the nature of the relationships between unmarried parents? (3) How do children born into these families fare? and (4) How do policies and environmental conditions affect families and children?

Using stratified random sampling and oversampling of nonmarital births, baseline data were collected from 1998 to 2000, through interviews with both biological parents shortly after the child's birth. Follow-up interviews were conducted when the child was one, three, five, and nine years old. As of this writing, the sixth wave of data collection, when the child is 15 years old, is ongoing. In addition to parents, the last three waves have included in-home assessments, child care or teacher assessments, and interviews with the child.

The parent interviews collect information on household characteristics, family relationships, parenting behavior, demographic characteristics, parent and child health (mental and physical), economic and employment status, neighborhood characteristics, and benefit receipt. Several collaborative studies provide additional information on parents' medical, employment, and incarceration histories, religion, child care, and early childhood education.

The study is funded by the National Institute of Child Health and Human Development and a consortium of private foundations and other government agencies. Fragile Families is a joint project of Princeton University and Columbia Universities. The principal investigators of the project are Sara McLanahan and Christina Paxson at Princeton; and Irwin Garfinkel, Jeanne Brooks-Gunn, and Ron Mincy at Columbia. The first fives waves of data are available on the web at http://www.fragilefamilies.princeton.edu/public.asp along with detailed information about the study design, data documentation, and publications using the data.

Panel Study of Income Dynamics

The PSID is the world's longest running household survey. The PSID was originally funded by the Office of Economic Opportunity of the U.S. Department of Commerce to monitor the impact of President Lyndon Johnson's War on Poverty and is currently funded primarily by the National Science Foundation, although it also receives support from a number of government agencies, foundations, and other organizations. The original PSID sample consisted of two samples: (1) a national sample, known as the Survey Research Center Sample (SRC), and (2) an oversample of low-income families designed to study issues related to poverty, known as the Economic Opportunity Sample. About 5000 families representing 18 000 individuals were included in the original 1968 sample and have the 'PSID gene.' In 1997, the PSID added the children of more than 2300 PSID families to replenish the sample and facilitate generational research (these children are also said to have the PSID gene). The PSID sample has experienced high retention rates and has grown considerably over time. In 1997, concerns over changing demographics of the United States resulted in the addition of a new immigrant sample to the PSID. Using the sampling weights, the PSID closely resembles the demographics of the U.S. population. The current sample includes 9000 households and 22 000 individuals.

Data are collected from all household members, but more detailed data are collected from the head of household and if, married or cohabitating, the spouse or partner, and 'other family unit members.' The original PSID sample members were followed annually from 1969 through 1997 and biannually beginning in 1997. In 1997, the Child Development Supplement was added in which the biological and adoptive children of PSID were added to the sample. These individuals were interviewed in 1997 when they were 0–12 years old and again in 2002/2003 and 2007/2008 when they were 5–18 and 10 and 18 for the third wave. Once they turned 18, the children continued to be interviewed in the Transition to Adulthood Study. They are interviewed biannually (2005, 2007, 2009, and 2011). In 2009, a supplemental data collection on Disability and Time Use (DUST) occurred for nearly 750 husbands and wives. Information was collected from two randomly selected days. Information was obtained using time diaries about what respondents did, where they were, who did the activities with them, who else was there, how they felt, and for household and care-related activities, for whom the activity was carried out. The survey also included information about well-being for three randomly selected activities per diary.

The original intent of the PSID was to collect data on the economic well-being of Americans. Since that time, additional

content has been added to the survey to reflect changing policy interests. In 2011, the 89-minute survey collected information on the following topics: employment, earnings, income, housing, foreclosure and mortgage distress, geospatial data, health status, adult behaviors, adult and childhood health conditions, health insurance, marriage and fertility, program participation, vehicle ownership, wealth and pensions, debt, and philanthropy.

The majority of PSID data is available on the Internet through the PSID Data Center: http://simba.isr.umich.edu/data/data.aspx. Users can create custom extracts and codebooks and can archive extracts for later use. Supplemental files are not available through the Data Center because they are of a different structure than the main interview files. See the above link for information on use of supplemental files.

Wisconsin Longitudinal Study

The WLS is a long-term longitudinal study of a random sample of men and women who graduated from Wisconsin high schools in 1957. The WLS is funded by the Behavioral and Social Research Program at the National Institute on Aging. The WLS began as a survey of Wisconsin high school seniors' plans for postsecondary education. In the early 1960s, William Sewell, a sociology professor at UW, learned that the surveys were located in the basement of a UW administrative building. He drew a random sample of 10 000 of the graduates and the survey has been following respondents for five decades. The sample was expanded in 1977 to include 2000 randomly selected siblings and again in 1997 to include 5000 siblings.

The sample followed respondents in 1964, 1975, 1992, 2003, and 2011. Data were collected from siblings in 1977, 1993, 2004, and 2011. Data were also collected from the spouse of the respondents in 2004, from the spouse of the siblings in 2006, and from the widows(ers) of graduates and siblings in 2006.

In 1992, when respondents were 35–45 years old, the content began to shift away from careers, education, and family to a sharper focus on health and aging. The survey provides complete longitudinal information on social background, attitudes, and aspirations in young adulthood, high school curriculum, labor market experiences, family formation, and social participation. The survey maintains a focus on understanding the influence of social background and adolescent experiences. The 2004 data collection included the following topics: education, marital history, status of parents, health, employment, pensions, retirement attitudes, income, health insurance, use of medical care, volunteer activity, alcohol history, religious preferences, the Internet use, psychological well-being, end-of-life preparations, values and attitudes, dealing with problems, current psychological states, and use of long-term care insurance. The data set also includes biomarker data, including height, weight, waist and hip circumference, grip strength, lung function, gait speed, and mobility. In the spring of 2006, saliva samples and DNA information was collected from about 4500 of the graduates and 2500 siblings.

Public-use WLS data can be downloaded directly from the WLS Web site at: http://www.ssc.wisc.edu/wlsresearch/data/. Users can download the complete data set or subsets of respondents. Some sensitive variables are not available publically but only to those who complete a restricted use agreement. The WLS Web site contains detailed data set documentation that can be downloaded, including codebooks, questionnaires, and analysis of response rates. The Web site also contains a tool that allows users to search for variables in the data set by name or description.

Health and Retirement Study

Sponsored by the National Institute on Aging and the Social Security Administrative, the Health and Retirement Study (HRS) is a nationally representative longitudinal study of approximately 26 000 Americans over the age of 50. The HRS has its roots in the Retirement History Study (RHS), which was conducted in 1969 and 1979. By the mid-1980s, it was recognized that the RHS had become dated, underrepresented women, blacks, and Hispanics – groups that had become a larger share of the labor market – and did not include questions on physical or mental function. The HRS began as two distinct but closely related surveys and has expanded over time to include new cohorts. The 'original HRS' was administered in 1992 to a sample of Americans at ages 51–61 and their spouses. The Study of Assets and Health Dynamics among the Oldest Old (AHEAD) began in 1993 as a sample of Americans aged 70 and older and their spouses. Both the original HRS and AHEAD were integrated in 1998, at which time two new cohorts were added. The first group was composed of people born between the two cohorts – 1924 and 1930 – and raised during the Great Depression (Children of Depression Age or CODA). The second group was designed to replenish the sample with people in their early 50s, known as the war baby cohort, born between 1942 and 1947.

HRS respondents are interviewed every 2 years. The 1-h interview broadly focuses on three topics: economics, family, and health. Within these broad categories, the survey focuses on the following topics: economic circumstances, occupations and employment, health and health care, cognition, living and housing arrangements, and demographics and family relationships. Special permission was obtained from respondents to link the interview responses to other data sources, including Social Security records on employment and earnings and SSI and SSDI benefit decisions and payments, Medicare claims data, and employer survey data focusing on pension plan provisions. These linked data are only available to researchers under a restricted use agreement. In addition, the HRS has genotyped 12 507 respondents who provided DNA.

Information on downloading the public use HRS data can be found at: http://hrsonline.isr.umich.edu/index.php?p=data. RAND data files are also available for download on this site.

See also: Applied Social Research, History of; Funding of Social Science, History of; Methodology of the History of the Social and Behavioral Sciences; Quantification in the History of the Social Sciences; Social Network Analysis; Statistical Analysis Packages.

Relevant Websites

http://www.census.gov/acs/www/ – American Community Survey Website.
http://www.census.gov/cps/data/ – Current Population Survey Website.
http://www.fragilefamilies.princeton.edu/public.as – Fragile Families and Child Well-being Study Website.
http://hrsonline.isr.umich.edu/index.php?p=data – Health and Retirement Study Website.
https://nces.ed.gov/das – National Center for Education Statistics Surveys Program Website.
https://www.nlsinfo.org/investigator/pages/welcome.jsp – National Longitudinal Surveys Website.
http://www.cpc.unc.edu/projects/addhealth – National Longitudinal Study of Adolescent Health Website.
http://simba.isr.umich.edu/data/data.asp – Panel Study of Income Dynamics Website.
http://www.ssc.wisc.edu/wlsresearch/data/ – Wisconsin Longitudinal Study Website.

Data Bases and Statistical Systems: Archives and Historical Databases

O Dascher, Nordrhein–Westfalen State Archives, Düsseldorf, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 1, pp. 642–646, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article discusses the history of archives from antiquity to the present day. It makes clear that archives follow a legal tradition that reaches back to the beginnings of occidental civilization. The use of archives for research is a relatively recent development: it basically began with the French archives law of 1794. A broad organization of archives, both public and private, has developed since that period and now encompasses all the countries in the world. The scope of this material (i.e., charters, records, registers, maps, plans, audiovisual media, electronic documents, private collections, and archives) and its appraisal pose serious problems for archivists today. Those problems include all questions dealing with preservation and information technologies.

This article discusses the history of archives from antiquity to the present day. It makes clear that archives follow a legal tradition that reaches back to the beginnings of occidental civilization. The use of archives for research is a relatively recent development that basically began with the French archives law of 1794. A broad organization of archives, both public and private, has developed since that period and now encompasses all the countries in the world. The scope of this material (i.e., charters, records, registers, maps, plans, audiovisual media, electronic documents, private collections, and archives) and its appraisal pose serious problems for archivists today. Those problems include all questions dealing with preservation and information technologies.

Terminology and History of Archives

The term 'archive' comes from arché, the Greek word for beginning. In Herodotus' History, Archeion is the town hall or government office. The word was not used to refer to preservation of the written word in ancient Greece, but was used in that manner by Josephus and later by Eusebius during the Hellenistic period. Roman jurists later adapted the Greek Archeion, using it in the Latin Archium and later Archivum. By the end of the fourth century, Archivum refers to the place where public records are preserved (locus quo acta publica asservantur). The archival tradition, therefore, is not based on the concept of preservation for research, but it was preservation for public administration purposes. Elements of this idea that archives only preserve public documents can still be found in present-day British and Dutch archives. It follows that the English language makes a difference between records that possess legal character and papers that possess no such qualities. The official character of such archival material as well as the fact that its preservation is a responsibility of the state, both explain the understanding of public faith and the concept of unbroken custody in the Anglo-Saxon world.

Research done by Johannes Papritz, Ernst Posner, and others indicates that clay tablets from 3000 BC found in the Near East were preserved to facilitate commercial and public administration. Strictly speaking, these tablets are registries, but were not intended for permanent preservation as an archive. Literature that uses the term 'clay tablet archives,' often referring to the royal palaces of Minoan civilization on Crete and the Palace of Nestor at Pylos (Peloponnese), is therefore misleading. While clay tablets are exceptionally durable, even fire-resistant, other material used to write on, such as leather, wooden tablets, or papyrus, proved to be less durable. Finds of such materials are therefore quite fragmentary. In Athens, laws and decrees of councils and assemblies, documents concerning acts of state, and copies of state-commissioned plays were stored at the town hall and, after the fourth century, at the Temple of Kybele. In Rome, senate resolutions were kept in the Temple of Saturn until 78 BC, and after that time senate minutes, state payment demands, census records, and so forth were placed in the 'Tabularium.'

Some remains of this ancient tradition can be found in the registers of the Middle Ages, that is, in the secure documentation of resolutions and records through continuous notation on archival tablets, which were later replaced by papyrus rolls. In AD 538, Emperor Justinian ordered all cities to preserve the commentarii and gesta municipalia, to facilitate jurisdiction, in a separate location, the Archeion or Archivum. Nevertheless, political upheavals in late antiquity and the beginning of the medieval period led to breaks in archival traditions. Roman registers can still be found in the cities of northern Italy and Gallia Cisalpina. Notary registers as compiled by Johannes Scriba of Genoa came into use in the middle of the twelfth century. The Papal Curia reintroduced register-keeping in the eleventh century (on parchment since 1088, though only records from 1198 survived due to a fire in the Lateran). The papal example may have had influence on the creation of royal archives in Aix-la-Chapelle under Charlemagne. The lack of a permanent residence among his heirs prevented the development of an archival organization. The influence of Persian–Arabian legal practice under Norman rule in Sicily could still be felt in France and England in the twelfth century.

Unlike the archives of antiquity, those developed in the medieval period did not serve the public interest in the first place, but rather helped to secure legal positions of individual institutions and groups, i.e., church and monastic repositories, municipal archives, and archives of the nobility.

The refinement of paper in the fourteenth century led to a reorganization of office and chancellery management. Registers, minutes, bills, and so forth became an administrative memory. Older archives were rearranged and concentrated in the fifteenth and sixteenth centuries: the Spanish court archives at Simancas (1548) are one notable example.

Modern Archives since the French Revolution

The French Revolution was an important turning point in the advancement of the archives. It was the revolution that introduced the developments we associate with archives today. The concentration of archival materials in central locations, the organization of an archives administration, the civil administration's jurisdiction over registries, public access, and historical research are all characteristics of this new world of archives.

The National Archives in Paris, established in 1789–90, was initially intended only for the preservation of records produced by the National Assembly. Records from ancien régime administration boards were not incorporated until 1793. The law passed on 25 June 1794 declared the National Archives to be a Centre Commun, giving citizens the right of access to the documents preserved. The departmental archives were created and put under the jurisdiction of the National Archives 2 years later. Much literature refers to 1794 as the 'constitution' of modern archives. Napoleon's 'Universal Archive' of 1810–11, which integrated important artifacts from the conquered German Empire, Austria, Italy, Spain, and the Vatican, lasted only briefly. What made the National Archives revolutionary was the fact that the arrangement of the archives followed the librarians' principle of pertinence. In 1839, initial calls were made, however, to introduce in the departmental archives the principle of provenance, which was codified in a circular on 24 April 1841. This concept of respect de fonds stipulates that all records created by one institution remain together. The paleographer Natalis de Wailly said of it, "La methode est fondée sur la nature des choses." The same principles were introduced on a municipal level in 1857.

In 1861, the principle of provenance was realized in Denmark's ministerial archives. In Germany, it was Prussia that prescribed the principle of provenance with the regulation of 1 July 1881: "Every government board will, once it generates records, receive its own repository." The regulation initially applied only to the Prussian Secret State Archives. Not until 1896 were the provincial archives advised to follow suit. Primary credit for the international success of the principle of provenance in the following decades must be given to the Dutch: In the Netherlands, the principle of provenance was introduced in 1897. The archivists Samuel Muller, Johan Adriaan Feith, and Robert Fruin described the principle of provenance in their handbook of archives, achieving worldwide recognition for Dutch archival theory. Translations into German (1905), Italian (1908), French (1910), and English (1920) followed. The conclusion of this discussion, which dominated the late nineteenth and early twentieth centuries, came at an international conference in Brussels in 1910. Archivists claimed the principle of provenance as fundamental for their profession. The principle of provenance had been accepted by Denmark in 1903, and by Sweden and the USA in 1909. Vienna's Haus-, Hof-, and Staatsarchiv accepted it in 1914. In 1920, Hilary Jenkinson achieved implementation in England, where the law had controlled archives since 1838. A decree by Lenin issued on 1 June 1918, dealing with the reorganizations of archives, gave Russian archives the chance to respect the principle of provenance whenever possible. An international conference in Stockholm in 1993 proved that the principle of provenance continues to play a role in international archival theory and discussion.

Records as Historical Sources: Appraisal and Use

Historical research and the development of critical methods, as exemplified by the German editions of the Monumenta Germaniae Historica since 1826, promoted archival description of older documents. Inventories have been published in England, France, and Belgium since the third decade of the nineteenth century. Editions of charters and records from Prussian archives have appeared since 1878.

While historians in Germany still discussed Karl Lamprecht and his cultural and sociohistorical theory from 1893 to 1898, archivists were already discussing topics that were very modern at the time, such as whether or not to preserve judicial records, which contained material on the administration of justice as well as the moral, financial, and economic circumstances of different classes of society. There were demands for the preservation of records relating to the labor movement and social questions, for census records, and, from 1904 onward, there were tendencies to preserve archives from private firms. It is no coincidence that during the first decade of the twentieth century in Germany the first private business archives were established. These were quite different from Europe's preindustrial trade and merchant registries that survived from the late Middle Ages (e.g., the Datini archives in Prato).

The disintegration of states and society as well as the massive amount of documentation in the twentieth century has led to an international discussion on archival appraisal and preservation. All efforts to come up with a generally applicable model of appraisal have led to dead ends. Obviously, every model of appraisal is in a certain way subjective in that historians' initial questions are formed by their generation: "Every political system considers different documents worthy of preservation" (Papritz, 1983). A free society cannot dictate binding rules for appraisal. The result of this insight was a closer look at Theodore R. Schellenberg's (1903–70) theories. Schellenberg (1964), former administrative head of the US National Archives, first published *Modern Archives: Principles and Techniques* in 1956. In the Anglo-American world at least, Schellenberg's ideas changed appraisal theory fundamentally. Schellenberg called for the appraisal of an entire institution's records. Besides the information in the documents about people, places, and events that was documented (the 'content' of records), the principle of evidence is focused on. Evidence in public records involves the archivist finding out how the originating government bodies functioned and what competence they had. For appraising purposes, it is indispensable that the archivist knows how the records in question were produced. The decision of whether or not to preserve records is then made on both formal and content-based criteria.

Papers issued by the Paris-based International Council of Archivists (ICA) indicate that networks of archives exist worldwide, even in developing nations. Because of its rich cultural tradition, China has the greatest number of archivists. Most countries have laws regulating users' access to archives. Conditions on access vary greatly. The Vatican only allows access to documents originating before 1922, while Scandinavians enjoy freedom of information rights that include access to current registries. In the Anglo-Saxon countries and Germany, there is a limitation on access of 30 years. Records on individuals and private data are regulated separately.

Current Problems: Preservation and Information Technology

Archives deal with two major problems today: the conservation/preservation of archival material and the effects of information technology. The material that is to be preserved does not age well, compared to medieval parchment. It requires the employment of complicated preservation techniques, among them are restoration, microfilm, and digitization.

As far as information technology is concerned, archives are presented with two major challenges: they must develop models for appraising and taking over of electronic media and they wish to provide universal access for researchers as well as for an interested public. The NARA (National Archives and Record Administration) in Washington, DC is the most experienced in this field. As of May 1999, some 100 000 files comprising more than 500 gigabytes had been taken over and stored there. Some 10 000 new files are expected to be added every year. Data that are transferred must first be copied onto separate tapes or cassettes. This procedure then has to be repeated every 10 years (concept of migration). A provisional list of the data available can be found on the Internet. Since data loss can occur in the copying process, Jeff Rothenberg has come up with a concept for emulating software that allows archivists to make copies using the original or related programs. The concept is currently undergoing tests as part of a project in the Netherlands. The goal of the project is the long-term preservation of digitized data in their original, authentic form (concept of emulation). In Germany, the German federal archives, the Bundesarchiv, is the most experienced in the appraisal and storage of digitized data. The Bundesarchiv is involved in a pilot project sponsored by the federal administration office called DOMEA (Document Management and Electronic Archiving in IT-supported operations), which is developing methods of saving, appraising, and preserving electronic files permanently.

For research purposes, the new possibilities that allow the searching of archives both nationally and internationally via the Internet are especially interesting. In Germany, most archives are accessible via the Internet. The state of North Rhine-Westphalia alone has made some 450 guides to state, municipal, church, and business archives accessible for Internet users. Marburg Archive School has presented the prototype of an online search aid especially suited to German descriptive methods, but also with an interface to an international standard of Internet presentation of search aids.

A major undertaking being discussed internationally is called EAD (Encoded Archival Description). In the USA, the Berkeley Art Museum, California State University, and the Library of Congress were instrumental in developing EAD, and have been joined by the Public Record Office in London. EAD has found an ever-growing following in the Anglo-Saxon world since the mid-1990s. It has meanwhile been made compatible with the International Standard Archival Description (ISAD) issued by ICA. Actually implementing these standards on a large scale requires a retroconversion of handwritten and typewritten search aids to a digital format. In Britain, home to some 2000 archives, the National Council on Archives introduced an archival network in 1998, which, after retroconversion of existing search aids, will provide a virtual guide to the archives and an incentive to make increased use of the archives over the Internet. The San Francisco-based Research Library Group (RLG) knows no national boundaries. Founded in 1974, the group comprises some 160 member universities, libraries, archives, museums, and scientific organizations. Databases of libraries and digitized finding aids of archives are made accessible by RLG.

In the 1990s, technological progress has changed archives in a way experts never thought possible. Following the digitization of archival guides, and the retroconversion of existing search aids and the organization of this metadata into databases, the digitization of entire fonds seems at least possible. Most progress has been made in digitizing collections of individual items like photographs and posters from various institutions such as museums, libraries, and archives. The digitization of files, that is, documents with several pages, has not yet begun. Information technology has made archives still more accessible and interesting for the public and has blurred the lines between archives, libraries, and documentary institutions.

See also: Archival Methods; Computers: Impact on the Social Sciences; Historical Archaeology; Historical Demography.

Bibliography

Abukhanfusa, K., Sydbeck, J. (Eds.), 1994. The Principle of Provenance. Report from the First Stockholm Conference on Archival Theory and the Principles of Provenance, 2–3 September 1993. Swedish National Archives, Stockholm.

Archival Legislation 1981–1994, 1995, 1996. 2 vols. Saur, Munich, Germany.

Beiträge zum Workshop im Staatsarchiv Münster July 8–9, 1998. In: Bischoff, F.M., Reininghaus, W. (Eds.), 1999. Die Rolle der Archive in Online-Informationssystemen. Staatsarchiv, Münster, Germany.

Symposion des Hauptstaatsarchivs September 11–13, 2000. In: Black-Veldtrup, M., Dascher, O. (Eds.), 2001. Archive vor der Globalisierung? Hauptstaatsarchiv, Düsseldorf, Germany.

Brennecke, A., Leesch, W., 1953. Archivkunde. Ein Beitrag zur Theorie und Geschichte des europischen Archivwesens. Köhler and Amelang, Leipzig, Germany.

Balloffet, N., Hille, J., Reed, J.A., 2005. Preservation and Conservation for Libraries and Archives. American Library Association, Chicago.

Craven, L., 2008. What Are Archives? Cultural and Theoretical Perspectives: A Reader. Ashgate, Aldershot, England.

Cox, R.J., 2000. Closing an Era: Historical Perspectives on Modern Archives and Records Management. Greenwood Press, Westport, Conn.

Cox, R.J., Wallace, D.A., 2002. Archives and the Public Good: Accountability and Records in Modern Society. Quorum Books, Westport, Conn.

De Lusenet, Y., 2000. Preservation Management. Between Policy and Practice. European Commission on Preservation and Access, Amsterdam.

Eckelmann, S., Kreikamp, H.-D., Menne-Haritz, A., Reininghaus, W., 2000. Neue Medien im Archiv: Onlinezugang und elektronische Unterlagen. Bericht über eine Studienreise nach Nordamerika, 10–21 May 1999 (Veröffentlichungen der Archivschule Marburg No. 32). Archivschule, Marburg, Germany.

Franz, E.G., 1999. Einführung in die Archivkunde. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany.

International Directory of Archives, 1992. Saur, Munich, Germany.

Jenkinson, H., 1937. A Manual of Archive Administration. Lund, Humphries, London.

Ketelaar, E., 1997. The Archival Image. Collected Essays. Verloren, Hilversum, The Netherlands.

Archivwissenschaftlichen Kolloquiums der Archivschule Marburg (Veröffentlichungen der Archivschule Marburg No. 31). In: Metzing, A. (Ed.), 2000. Digitale Archive—Ein neues Paradigma? Beitrge des 4. Archivschule, Marburg, Germany.

Muller, S., Feith, J.A., Fruin, R., 1898. Handleiding voor het ordenen en beschrijven van archieven. Van der Kamp, Groningen, The Netherlands.

Millar, L., 2010. Archives: Principles and Practices. Neal-Schuman Publishers, New York.

O'Toole, J.M., Cox, R.J., Society of American Archivists, 2006. Understanding Archives & Manuscripts. Society of American Archivists, Chicago, IL.

Papritz, J., 1983. Archivwissenschaft, 4 vols. Archivschule, Marburg, Germany.

Rothenberg, J., 1999. Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation. A Report to the Council on Library and Information Resources. http://www.clir.org/pubs/reports/rothenberg/pub77.pdf.

Roe, K., 2005. Arranging & Describing Archives & Manuscripts. Society of American Archivists, Chicago.

Schellenberg, T.R., 1964. Modern Archives. Principles and Techniques. Angus and Robertson, London (A. Menne-Haritz (Trans. and Ed.) 1990 Die Bewertung modernen Verwaltungsschriftguts).

Shepherd, E., 2009. Archives and Archivists in 20th Century England. Ashgate Pub, Farnham, Surrey, England.

Data Bases and Statistical Systems: Communications and Media

Wolfgang Seufert, Friedrich-Schiller-University Jena, Jena, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article presents the most important statistical systems that regularly provide data on communications and media suitable for secondary data analysis in social research. These data from official and nonofficial statistical sources have some essential limitations. Some arise from a lack of data provided by official sources, particularly for data on markets and on the cultural, political, and social dimensions of communications and media. Nonofficial sources fill this 'availability gap' to some extent. However, some severe methodological constraints underlie data from industry associations or commercial data providers, which should be explicitly discussed when used in secondary analyses.

Economic and social research on the impact of communications and media on the society often relies on secondary data analyses, due to the fact that primary data collection is time consuming and expensive. The following article intends to present the most important statistical systems that provide data on communications and media, including their methodological limitations (cf Seufert, 2006).

One of the key characteristics of the communications and media field of research is continuous change in the underlying communication and information (IC) technologies. In the 1980s, 'digitisation' or 'convergence' transferred telecommunication networks, equipment, and media storage and media distribution technologies into another form of data processing technology. The emergence of Internet technology over the last several decades has even altered the meaning of 'communications' and 'media' to some extent. Hence, both terms should be clarified first.

Twenty years ago, it was easy to draw up a frontier between (tele)communications and (mass)media. The term 'communications' was used for various types of private one-to-one communication, based on postal services or telecommunications networks. The term 'media' was used for various types of public one-to-many communication, based either on terrestrial, cable, and satellite distribution networks or on storage technologies (printed paper, sound records, and video records). Internet technology, or more exactly the World Wide Web, now integrates new ways of one-to-one communication (e.g., via email or voice-over-IP) and of mass media communication (e.g., web TV or music-on-demand portals). Furthermore, it also enables a new hybrid form of network-based communication in social networks that is to some extent private and to some extent public. Companies like Twitter, YouTube, Facebook, etc. provide the necessary infrastructure for this new kind of social communication, which is often summarized under the notion of 'social media.'

The following overview covers data suitable for secondary analyses in social research, as provided by the most important official and nonofficial sources on communications, traditional mass media, and new forms of network-based social communication.

Research Questions and Data Needs

What statistical data is appropriate to analyze the communications and media system of a society? The answer to this question differs according to the research interests of scientific subdisciplines. Economic researchers, for example, usually require other data than researchers from other social science disciplines.

Researchers with a focus on telecommunications and mass media economics belong to two scientific communities. Industrial economists focus on the typical economic behavior of actors in telecommunications and media markets and for optimal regulation regimes for those markets in order to maximize economic welfare (cf Albarran, 2010; Doyle, 2002; Picard, 2002; Majumdar et al., 2005). Research in business administration is more interested in finding the best solutions for typical management problems of telecommunication carriers or media companies in order to achieve business success (cf Küng, 2008; Vogel, 2011). Both groups of economic researchers require four types of quantified statistical information to test their hypothesis and their respective underlying economic theories:

1. Statistical information on enterprises and industries: Variables describing the supply side of telecommunications and media markets, and variables which allow for calculations of the macroeconomic relevance of the communications and media sector within the national economy by size, growth, or productivity effects (e.g., sectorial data on gross value added, gross income, number of employees, immaterial investment into intellectual property, etc.).
2. Statistical information on market concentration: Variables describing the market power of different market players by economic concentration indicators (e.g., the number of suppliers, the turnover shares of the biggest companies in the market, etc.).
3. Statistical information on size and structure of market demand: Variables describing market size and price levels of telecommunications and Internet services or media products, including information on the detailed demand structure of households and enterprises for this range of goods (e.g., turnover, sales volume, price indices). It also includes information on the advertising demand of enterprises, as mass media are to a large extent financed by advertising.
4. Statistical information on the technical infrastructure required for telecommunications, Internet or audiovisual media: Variables describing the capacity of telecommunication networks and media distribution networks, the access to these various services, and the inventory of transmitter–receiver equipment in households or enterprises

 http://dx.doi.org/10.1016/B978-0-08-097086-8.41067-6

(e.g., proportion of individuals or small and medium enterprises with access to broadband Internet or mobile telephone services, the average number of TV channels received by households, etc.).

Social scientists, particularly those from the disciplines of sociology, political science, and communications science, are more interested in the impact of communications and media on the society as a whole (e.g., on the social structure, social habits, and common beliefs). Of particular interest is the impact of traditional media content reception, or the participation in social media debates, on political decisions and the political systems (cf Knoche, 1999; McQuail, 1992; Van Cuilenburg, 1999). Researchers who are interested in cultural, political, and social dimensions of communications and media usually request three additional types of quantified statistical information:

5. Statistical information on the use of telecommunications service and the Internet: Variables describing in detail the frequency and type of different telecommunication services or Internet activities including data about the user structure (e.g., sending or receiving email, participating in social networks, selling or purchasing goods or services, streaming or downloading images, movies and other media content, blogging, etc.).
6. Statistical information on media use: Variables describing in detail the demand for different types of media content including data about audience structure (e.g., time spent on radio use per day, the breakdown of TV viewers or cinemagoers by sociodemographic variables, etc.).
7. Statistical information on media content: Variables describing the quality or diversity of content in mass media in detail (e.g., the total number of published newspapers or magazines, the number of feature films of different genres exhibited during a year, the breakdown of TV programs by format categories, the proportion of speech or music content in radio programs, etc.).

All relevant statistical information related to communications and media activities in a society can be analyzed by either microdata or aggregated data. Economists are used to working with macrodata, particularly in time series analyses, using averages or total figures for groups of subjects in countries or other regions, usually defined by political boundaries. Empirical social scientists usually prefer to work with microdata, i.e., by using statistical information from individual enterprises, households, or 'publication units,' e.g., TV channels, newspaper titles, etc.

Official and Nonofficial Data Sources

As in other research fields, two groups of statistical systems provide data on communications and media: official and nonofficial sources.

Official Statistical Sources

'Official statistical sources' on the national level comprise national statistical institutes (NSI) and other government institutions. Central banks, for example, are responsible for the data collection on international financial flows related to the exploitation of intellectual property rights (e.g., income from film licenses). In many countries, national regulatory authorities (NRA) for the telecommunications or media sector also provide regular data on market concentration and other variables relevant for regulatory purposes. Some examples for NRAs are Autorità per le granzie delle communicazioni (AGCOM): www.agcom.it; Federal Communications Commission (FCC): www.fcc.gov; Kommission zur Ermittlung der Konzentration im Medienbereich (KEK): www.kek-online.de; Office of Communications (Ofcom): www.ofcom.org.uk. Most of the economic indicators related to communications and media comply with internationally concerted classifications of production output or consumption activities (c.f. below). This is not the case for most of the indicators describing the cultural, political, or social dimensions of the use of communications and media. One exception is the international classification of activities for time-use statistics (ICATUS), developed by the UN statistics division (UNSD). (UN statistical unit division: UN classifications registry – ICATUS (www.unstats.un.org). It includes 'time spent using mass media' as one of 15 main activity categories.)

Some supranational, intergovernmental institutions partly fill this gap and provide harmonized statistical data on media content, media use, or Internet use, which is suitable for cross-country analyses, namely UN suborganizations (International Telecommunications Union (ITU), UNESCO) or the OECD. One example is Information and Communication Technology (ICT) usage indicators, published by ITU (cf below). In most cases, these supranational official statistical sources do not collect their own data but use data from national sources of their member states. They are usually restricted to use data only from NSIs or other official sources. However, in some cases, they collect additional data from nonofficial sources. This applies particularly for the European Audiovisual Observatory (EAO). Since 1992, the Council of Europe finance the EAO to collect and publish statistical information on audiovisual markets in all member states (at present there are 47 member states: all European countries except Belarus, plus Armenia, Azerbaijan, Georgia, the Russian Federation, and Turkey), North America, and Japan (www.obs.coe.int). The provision of harmonized statistical data on communications and media on EU Community level is also part of the mandate of EUROSTAT, the general directorate of the European Union.

Based on law, official sources can force enterprises or households to participate in statistical surveys, but have the obligation to maintain secrecy to individual household or enterprise data. Therefore, most data is published anonymous on an aggregated level. However, even for industries, enterprise size classes, or household income groups, one can find dots in the published data of official sources. This is particularly the case for aggregated data from smaller regions, if, for example, the total number of enterprises is less than five or if one enterprise of the region has a turnover share of 90% of more (cf Cox et al., 1995). However, there are two exceptions to this rule. Anonymized microdatasets are in some statistical areas available as public-use or scientific-use files, as, for example, with household budget surveys. Individual enterprise data are

available for publicly owned companies, particularly for public broadcasters. It was also available for public telecom carriers until their privatization in the 1990s.

Nonofficial Statistical Sources

The term 'nonofficial statistical source' comprises a heterogeneous spectrum of data providers: nonprofit organizations and companies that provide data as a business activity.

Commercial, nonofficial sources include a broad range of profit-oriented data providers, surveying the economic performance of companies, advertising activities, private consumption of communications and media products, and media use (cf below). However, industry associations publish most of the nonofficial data on communications and media. They serve information needs of member companies but their statistical work is often also a part of lobbying activities. To demonstrate the economic importance of their specific industry, associations have the tendency to include economic activities from neighbor industries, with the result of double counting of market sales or number of employees. Statistical information of industry associations is normally limited to aggregated market data on turnover, number of sold units, average prices, etc. Compared to official sources, it has a much more detailed breakdown by product categories. Some industry associations collect information that is more detailed, particularly data on cost structures. However, to maintain information of individual members confidentially the statistical reports contain only average cost figures for the whole industry or for enterprise size classes.

Several supranational nonofficial sources publish cross-country data on telecommunications and media markets. (The EAO, for example, uses statistical information from the European Broadcasting Union, the European Cable Communications Association, the Group of European Audience Research, the International Federation of Phonographic Industry, and the International Video Federation for its audiovisual yearbooks.) This data is in most cases taken from various national nonofficial sources. Since industry associations or commercial surveying companies from different countries usually work with different classifications and with different data collection methods, there is a general risk to compare apples and oranges. When using this kind of nonofficial statistical sources for cross-country analyses it is essential to proof carefully, how far the data from different countries are harmonized (cf below).

Data on Communications and Media Provided by Official Sources

Governments decide about the scope of a continuous data collection process surveying enterprises, local production sites, households, or other statistical units. These general statistical programs also confine the availability of data on communications and media from official sources. Sometimes new statistical programs improve it. One example is the establishment of a statistical program on audiovisual services and related sectors (AUVIS) by the European Community (EU, 1999a). Another example is a special statistical program on 'information society statistics' by collecting data on the usage of advanced IC technology in enterprises and households, which started because of a G7 summit initiative in the 1990s (ISPO, 2000). Two factors are crucial for decisions about the number and type of collected variables in a statistical program: the 'expressed user needs' for statistical information on the one hand and the 'statistical environment,' i.e., constraints given by internationally agreed classifications for economic statistics, on the other hand.

Statistical Programs and User Needs

A typical implementation process of a new statistical program, for example by EUROSTAT, starts with a consultation process, asking the EU administration (general directorates), NSI and all relevant industry associations organizing enterprises on a European level. This means that the expressed needs of regulators and organized market players will have priority to those needs which might exist in the scientific community or in the general public, for example, journalists who are interested in monitoring the communications and mass media system. In order to reduce the cost burden of data collection for responding enterprises the number of items is minimized, particularly in the case of surveys among small- and medium-sized enterprises.

The expressed data needs of industry associations are normally concentrated on market data and – to a lesser degree – on industry information. They are motivated by a general interest in higher market transparency, which will give members of the association cost-saving opportunities to compare the economic performance of the own enterprise with the industry average, or to analyze structural changes in demand patterns or general trends in market growth. The data needs expressed by the EU administration and by national regulators (via the NSIs) are more related to special regulatory goals pursued in the policy fields of telecommunications, mass media, or IC technology. Three different policies approach this demand:

- Industrial policy, which considers the information and telecommunications (IT) industries and the media industry (nowadays often as a part of 'the creative industries') as key sectors for economic growth with major employment opportunities.
- Competition policy, which aims to avoid monopolistic concentrations of market power in all sectors and their negative impacts on market performance.
- Cultural policy, which wants to avoid negative impacts of media concentration on the diversity and quality of national cultural production or on political pluralism in particular.

The establishment of the above-mentioned AUVIS program, for example, is closely related to an EU industrial policy program in the field of audiovisual services to strengthen the competitiveness of European producers against Hollywood (EU, 1999b). Cultural policy goals are less important. This is demonstrated by the fact that EU member states have to collect data about the country or origin of cinematographic films exhibited in all cinemas or of fictional programs shown on national TV channels, but no data about the genre structure of transmitted TV formats. (Among other actions, a 50% quota for TV programs with European origin was implemented.)

The Statistical Environment: Dominance of Industry Data Instead of Market Data

The general 'statistical environment' also influences the type of data collected by official sources on communications and media. On the one hand, statistical programs on economic indicators are dominated by the concept of national account statistics (System of National Accounts (SNA), established by the United Nations in the 1950s). On the other hand, official sources have to bear in mind internationally agreed statistical classifications of industries, goods, and services.

The SNA standardizes the calculation of growth rates and productivity development in industries or economic sectors (related to the three sector model of agriculture and mining, manufacturing industries, and services industries) across national economies. This means that the variable list for enterprise economic information (also called structural business statistics) is the same for all types of enterprises in all industries and that the opportunity to collect additional variables for communications or media industries is rather limited (EU, 1999a).

Data collection on production output and consumption of goods and services follows two different concepts: data on the supply side of markets is normally collected and aggregated by following the institutional approach, while data on the demand side of markets is collected by following the functional approach. These two approaches are also known as the 'industry approach' and the 'market approach':

- The institutional approach (industry approach) consists of aggregating variables describing the economic attributes of those enterprises, whose primary economic activity characterizes a particular industry within the national economy. Enterprises that produce the same products as a secondary activity are excluded. A TV channel that, for example, also produces its own TV guide belongs to the TV-broadcasting industry but not to the magazine-publishing industry.
- The functional approach (market approach) aggregates economic data by similar products, which characterize a particular economic market. Markets may be defined either from the consumer's point of view, which consists of all functional substitutes serving the consumer's needs (for example 'newspapers'), or by considering the economic or technical attributes of products (for example 'printed newspapers' or 'online newspapers'). Using technical attributes, the following taxonomies are used for telecommunications and Internet services and media products: by type of end users, by type of financial flows (e.g., advertising or direct payments), by type of content (e.g., audio or audio-visual), by degree of interactivity, by type of production technology, or by type of distribution technology.

Business surveys conducted by NSI follow the institutional approach and the data is primarily used to calculate indicators on production output defined in the SNA (e.g., gross value added, gross investment, gross capital formation, or total employment). Since households can be more easily surveyed about the kinds of products they have purchased than about their respective manufactures, economic statisticians follow the functional approach in surveys on household consumption.

The result is that the turnover of a specific industry, for example newspaper publishers, measured by the institutional approach, differs from market sales figures for newspapers, based on the functional approach. There are efforts to minimize this difference by defining goods and services markets generally by the economic origin criterion, i.e., to define products not by technical attributes but by the main economic activities of the enterprises of an industry (the so-called integrated system).

The Statistical Environment: International Nomenclatures and Classification Systems

The UNSD developed two statistical classifications that are obligatory for all economic information published by official statistical sources: The classification of economic activities (defining industries) and that for economic products (defining goods and services markets):

- ISIC Rev.4 (*International Standard Industrial Classification of all economic activities*) is the UN classification for economic activities, designed to categorize production data relating to economic 'statistical units,' like an individual plant or a group of plants constituting an enterprise. It structures the national economy into classes and subclasses of industries. NACE (*Nomenclature statistique des activités économiques dans la Communauté européenne*) Rev.2 is the corresponding classification for EU member states.
- CPC Ver.2 (*Central Product Classification*) is the UN classification for manufacturing and services products. Though closely linked to the ISIC nomenclature, the CPC consists more subclasses. CPA (*Statistical Classification of Products by Activity*) 2008 is the corresponding classification for EU member states.

Another UN classification exists for traded goods, called the *harmonized commodity description and coding system* (HS 2012). The *combined nomenclature* (CN) is the corresponding classification for EU member states. Both have far more subclasses than ISIC and CPC in order to address the special needs of custom offices. (The World Customs Organization originally introduced the HS classification in 1952.) Since some manufactured goods or services are not meant for household consumption, the UN has developed another *classification for individual consumption according to purpose* (COICOP). Its substructure of 14 main purposes is the reference point for all NSI publications on private consumption as well as for the calculation of consumer price indices. ISIC and COICOP are also used to structure production output and consumption data in national account statistics. Since 2008, the national account statistics of EU member states were adapted to the ISIC structure. This means that data on gross output, gross value added, employment, or labor productivity are continuously published for several industries producing 'information and communication services' forming together an 'information and communication sector.' In the North American Industry Classification System, such an 'information sector' was established more than 10 years earlier. Figure 1 illustrates the relationship between ISIC, CPC, HS, and COICOP.

Communications and Media in Official Statistics

Communications and media industries and products have subclasses in all classifications. However, the two main

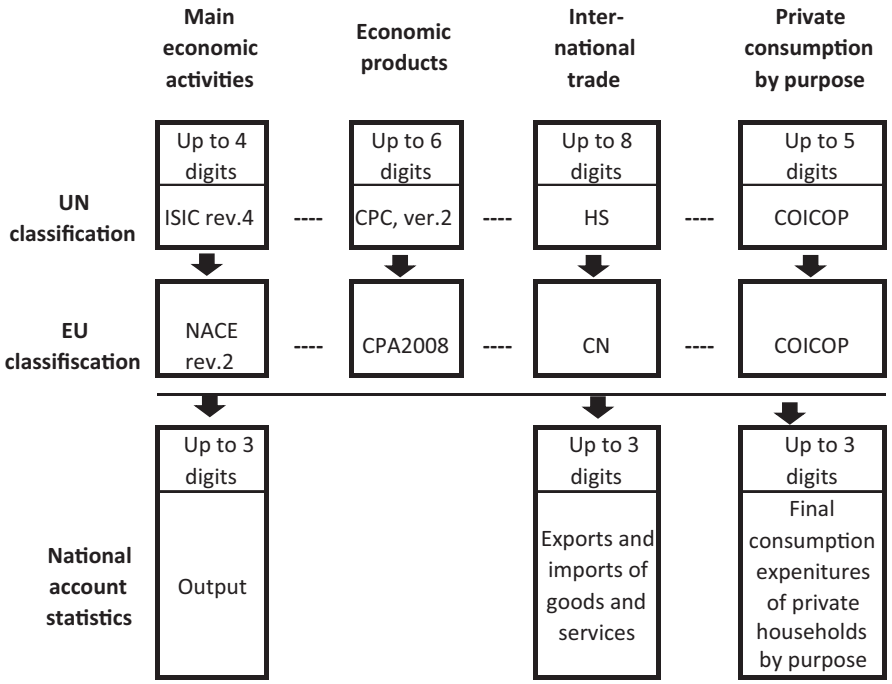

Figure 1 UN and EU activity and product classifications for economic statistics. ISIC, International Standard Industrial Classification; CPC, Central Product Classification; HS, harmonized system; COICOP, classification for individual consumption according to purpose; CN, Combined nomenclature for foreign trade statistics.

economic classifications of official statistical sources ISIC and CPC are restricted to a maximum of subclasses. This means, that the degree of differentiation between communications and media industries and products is rather limited. Figure 2 shows the currently available data on industry output and household consumption of communications and media in SNA statistics.

Since the speed of innovation in ICT is very high, it constantly shifts the boundaries between media industries and creates new media markets. In light of this, data on 'new media' are in high demand from market players and governments. This demand conflicts with the regular revision period of economic classifications: the COICOP, HS, ISIC, and CPC are all revised on a 5-year schedule. This is a consequence of the close relationship between all classifications. If the structure of one classification is changed all others have to be changed accordingly. As a result, new communications and media industries or products (like new Internet services) are usually 'hidden' in official economic statistics for a couple of years, either in one of the 'old' activity or product classes or in the auxiliary category 'not else classified.'

Secondary analyses of cultural, political, or social dimensions of communications and media cannot rely on internationally concerted classifications in the same way. One important exception is indicators for the use of Internet services and IC technology in households and enterprises. Under the heading 'information society statistics,' OECD and ITU developed *Guidelines for statistics on the access to telecommunications infrastructure and for surveys on the use of households of IC technology* in a coordinated action. (International Telecommunications Union (ITU): Manual for Measuring ICT Access and Use by Household and Individuals, 2014 edition (www.itu.int).) Since 2004, EUROSTAT coordinates similar surveys on the use of IC technology by business enterprises in all EU member states, with an emphasis on small and medium scale companies. (EUROSTAT: ICT usage by enterprises. Reference metadata (http://epp.eurostat.ec.europa.eu).)

EUROSTAT has also made an effort to compile cross-national data on cultural activities from EU member states. (EUROSTAT: *Pocketbook on Cultural Statistics* (http://epp.eurostat.ec.europa.eu).) Due to different national traditions in surveying media markets and publicly funded cultural activities, the result is rather small. The *Pocketbook on Cultural Statistics*, published in 2011, contains only a few data on cinema, book, and printed press markets in EU member states.

Data on Communications and Media Provided by Nonofficial Sources

National and Supranational Nonofficial Sources

The bias for industry data instead of market data in official economic statistics, and the high innovation rate in IC technology drive a demand for more detailed statistical information particularly on 'new' communication and media markets. National industry associations often try to provide this missing data. Either they collect turnover and sales data from their member companies or they commission surveying companies to collect it.

In nearly all countries, specialized enterprises collect the following five types of information on communication and media as a commercial business:

- Data on individual communication and media enterprises compiled from balance sheets, profit and loss accounts, or

Communications and media activities (ISIC rev.4/NACE rev.2 - 2/3 digits)	
Manufacturing	**Section C**
Printing and reproduction of recorded media	18
Manufacture of computer, electronic, and optical products	26
Information and communications	**Section J**
Publishing of books, periodicals, and other publishing activities	**581**
Software publishing	**582**
Motion picture, video, and TV program activities	**591**
Sound recording and music publishing activities	**592**
Radio broadcasting	**601**
TV programming and broadcasting activities	**602**
Wired telecommunications activities	**611**
Wireless telecommunications activities	**612**
Satellite telecommunications activities	**613**
Other telecommunications activities	**619**
Computer programming, consultancy, and related activities	**620**
Data processing, hosting, and related activities; web portals	**631**
Other information service activities	**639**
Professional, scientific, and technical activities	**Section M**
Advertising	**731**
Market research and public opinion polling	732
Final consumption expenditure of households by purpose (COICOP - 3 digits)	
Communication	**08**
Postal services	08.1
Telephone and telefax equipment	**08.2**
Telephone and telefax services	**08.3**
Recreation and culture	**09**
Audio visual, photographic, and information processing equipment	**09.1**
Other major durables for recreation and culture	09.2
Other recreational items and equipment, gardens and pets	09.3
Recreational and cultural services	**09.4**
Newspapers, books, and stationery	**09.5**
Package holidays	09.6

Figure 2 Communications and media data on output and household consumption in System of National Accounts. ISIC, International Standard Industrial Classification.

other published materials about ownership structures. User demand primarily stems from financial investors seek to assess the competiveness and financial solidity of enterprises. Such data are also useful in researching and analyzing market concentration in telecommunications and media markets.

- Data on household consumption, particularly on new communications services or media products. Many of these surveys are regularly conducted on behalf of industry associations. Data collection is based on representative samples or panels of households or retailers.
- Data on the advertising strategies of companies across all sectors, including statistical information about total advertising expenditure on various mass media (newspaper, magazines, TV, radio, cinema, etc.). Enterprises wanting to monitor their competitors drive demand for data on advertising. In advertising research, these data are widely used to analyze industry-specific advertising behavior or the relationship between economic growth, the level of advertising spending (compared to GDP), and the financial base of mass media (cf Van der Wurff et al., 2008).
- Data on the audience for different 'publication units' like TV or radio channels or newspaper and magazine titles. Mass media companies commission this data collection as a precondition to get advertising money. It is primarily used for media planning decisions in advertising campaigns. In communication science data on total reach, frequency of use, and breakdown of users is taken to identify different patterns of media use in social groups of the society (cf Webster et al., 2006).
- Data on the structure of audiovisual media content. Copyright collecting societies drive the demand for information on the exploitation of audiovisual or audio copyrights by TV or radio channels. Such data can be used to analyze the genre structure of audiovisual mass media content to some extent (cf Buonanno, 2002).

Though some supranational industry associations publish cross-country data on special media markets on a regular basis, the range of data is often limited to a few key indicators. Examples are:

- The 'World Magazine Trend' annual report of the International Federation of the Periodical Press contains data on the number of consumer and business titles published, the number of copies sold, the average issue readership, as well as data on magazine advertising expenditure (www.fipp.com).
- The 'World Press Trend' annual report of the World Association of Newspapers contains statistical information on the number of newspaper titles published, circulation and reach of the newspapers, as well as data on newspaper advertising expenditure (www.wan-press.org).
- The 'World Advertising Trend' annual report of the World Advertising Research Center (WARC) contains statistical information on money spent for advertising in newspapers, magazines, TV and radio programs, cinemas, or for online advertising (www.warc.com).
- Annual reports on data processing and telecommunications markets in EU member states, the United States and Japan of the European Information Technology Observatory (EITO). The first edition in 1993 was a joint initiative of the EC Commission, the European IT industry association (Eurobit), and several European trade fairs. The EITO uses business enterprise surveys to collect its own primary data (www.eito.com).

Sources Providing Data on Internet-Based Communication

Data on the supply and usage of Internet services and Internet web content is collected by NSI or commercial data providers by using traditional surveying methods, e.g., for the ICT usage indicators published annually by ITU (cf above). However, Internet technology offers opportunities to produce process-generated data on Internet-based communication as well. Software tools, either server- or client-based, are increasingly used for web analytics. The starting point for this development was server-based log file analyses that gave Internet providers information about the search behavior of users on their own Web sites during one Internet session. These data can be used to optimize Web site content structure or promote advertising space (e.g., by counting page visits, click-through rates, etc.). Client-based software enlarged these opportunities. For example, tracking software places little programs (cookies) on Internet user devices to allow for the collection of user-specific data for more than one online session. These individual communication profiles can be used for a broad range of direct marketing activities. There are also tools for semantic content analyses to observe the diffusion and annotation of single news on the web, which show communication links between different Web sites. This type of analyses is frequently used to assess the efficiency of public relation campaigns. Furthermore, social network providers publish own social media metrics to measure the engagement level of users for websites by counting 'likes,' 'shares,' or 'comments.' Big data analyses aggregate and combine all the process-generated data on Internet-based communication for special regions or whole countries, and try to find significant correlations to special dimensions of social behavior, e.g., crime intensity or traffic flows (cf Mayer-Schönberger and Cukier, 2013). Social research on web communication is widely based on published metrics of web analytics tools, particularly of social network providers. However, it should be considered that the underlying software algorithm for most of these tools remains secret.

Methodological Limitations of Statistical Sources on Communications and Media

Data on communications and media provided by official statistical and nonofficial sources have some methodological limitations, which restrict suitability for social research.

Insufficient Data Availability for the Analyses of Cultural, Political, and Social Dimensions of Communications and Media in Official Sources

Due to the dominant user needs of governments and organized industries, official statistical sources mainly provide data about the economic dimension of communications and media. Collecting economic information, NSI, and other official sources are obliged to consider internationally agreed classifications on industries and products and to hold individual data confidential. The consequence is a lack of data, both on the national and the supranational level. The main limitations are as follows:

- lack of microdata on detailed breakdown of turnover, cost structures, and employment of communications and media enterprises, as well as aggregated data for small regions or countries;
- lack of aggregated data on markets, due to the dominance of the institutional approach in production output statistics;
- insufficient aggregated data on 'old' media industries which are not in the focus of policy action programs, e.g., print media;
- a general lack of data on 'new' communications and media industry or products, due to a 5-year revision interval for international activity and product classifications;
- lack of data on nonprofit organizations or individuals producing media content (e.g., so-called 'user-generated content'); and
- finally, a general lack of microdata and aggregated data on media content or media use due to the lack of international classifications for cultural statistics. (The only exception is time budget data, but the main category 'use of mass media' is not very feasible, due to the fact, that about half the time spent on media use is parallel to other main activities.)

To some extent, reports on behalf of governments, political advisory councils, parliament commissions, or expert groups contain data that are suitable for social research on the cultural, political, and social dimensions of communications and media. However, these data are not collected on a regular basis, and unique national statistical categories are used.

Insufficient Data of Nonofficial Sources Arising from a Diversity of Statistical Classifications and Altering Collection Methods

Numerous nonofficial sources provide aggregated data on old and new communications and media markets, as well as

microdata on communications and media enterprises, on individual media use and – on a minor scale – media content. Many of these data providers only offer data on a national level using their own statistical classifications. This leads to severe constraints for secondary analyses in social research.

One methodical problem arises from the tendency of industry associations and commercial data providers to change their classifications and data collection methods very frequently. This means that one has to be careful when using data from nonofficial statistical sources for time series analyses, even on the national level. A break in a statistical trend may be a statistical artifact due to altered statistical categories or data collection methods, and not the result of changing behavior of providers or users of communications and Internet services or media content. In 2000, for example, the interview technique of Germany's largest annual audience survey (media analyze) changed from personal face-to-face interviews to computer-assisted telephone interviews (CATI). A sharp rise in the level of radio listening time per adult of nearly 30 min per day was the result.

The different national statistical categories and collection methods also have a negative effect on the usability of the data for cross-country studies. Supranational nonofficial sources that claim to provide harmonized international data, by altering the original national data, have to be checked thoroughly. One example is data published on the time spent on different media. It is collected by either diary methods, personal interviews, or telephone interviews. There is clear evidence that the chosen method has an impact on the results. Another good example is the above-mentioned annual WARC report on international advertising markets. Some national advertising associations collect the so-called 'net' advertising turnover figures (i.e., without deductions and commission charges) directly from media companies. In other countries, commercial companies measure advertising volume (number of advertising spots or ad pages) and estimate 'gross' advertising expenditures by multiplying the volume figures with the official media tariffs, announced by media companies. However, in reality most of the advertising time and space is sold much cheaper. To harmonize data from different countries, WARC estimates the gross–net difference, but is not very successful in doing so. (In Germany, for example, both methods to collect data on advertising expenditures are in use. In 2012, the gap between published 'net' TV adverting figures and published 'gross' TV adverting spending was 70% while the estimate of WARC was 20%.) To avoid the production of statistical artifacts, all cross-country studies using data from supranational nonofficial sources should demonstrate whether the collection method has no statistically significant effect on the empirical results.

Conclusions

From the perspective of social research, the data on communications and media provided regularly by official and nonofficial statistical systems have some essential limitations. Some arise from a lack of data provided by official sources, particularly for detailed market data and indicators describing the cultural, political, and social dimensions of communications and media. This 'availability gap' is partly filled by nonofficial sources. However, data from industry associations or commercial data providers underlie some severe methodological limitations, which should be explicitly discussed when used in secondary analyses.

To improve this situation, social researchers should exert influence on official statistical sources in two directions – to facilitate the access to microdata on communications and media enterprises, and to develop international classifications not only for the use of advanced IC technology but also for statistical indicators on traditional mass media. The opportunities of researchers to exert influence on statistical classifications or collection methods of nonofficial sources are more limited. Nevertheless, a broader discussion of the methodological limitations, particularly of sources used frequently for cross-country analyses, might help to improve the outcome of these research projects.

See also: Data Bases and Statistical Systems: Culture; Data Bases and Statistical Systems: Economics (Statistical Systems); Data Bases and Statistical Systems: International Comparative Research; E-Communications Platforms and E-Learning; Social Network Algorithms and Software; Social Science and the Media; Time Series: General; Time Use Research Methods.

Bibliography

Albarran, A.B., 2010. The Media Economy. Routledge, New York.
Buonanno, M. (Ed.), 2002. Convergences. Eurofiction Fourth Report. Liguori Publications, Liguori.
Council of Europe, 1992. Resolution (92)70 Establishing a European Audio-Visual Observatory. Adopted by the Committee of Ministers on 15 December 1992.
Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J., Kott, P.S. (Eds.), 1995. Business Survey Methods. John Wiley & Sons, New York.
Doyle, G., 2002. Understanding Media Economics. SAGE, London.
European Commission (EU), 1999a. Council Decision of 26 April 1999 establishing a Community statistical information infrastructure relating to the industry and markets of the audiovisual and related sectors. (1999/297/EC).
European Commission (EU), 1999b. European Audiovisual Policy in the Digital Age. Luxemburg.
Information Society Project Office (ISPO), October 2000. Information Society Indicators in the Member States of the European Union. Brussels.
Knoche, M., 1999. Media economics as a sub discipline of communication science. In: Brosius, H.-B., Holz-Bacha, C. (Eds.), German Communications Yearbook. Hampton Press, Cresskill.
Küng, L., 2008. Strategic Management in the Media Theory to Practice. SAGE, London.
Majumdar, S.K., Vogelsang, I., Cave, M.E. (Eds.), 2005. Handbook of Telecommunications Economics. Elsevier, San Diego, Oxford and London.
Mayer-Schönberger, V., Cukier, K., 2013. Big Data: A Revolution that Will Transform How We Live, Work and Think. John Murrey, London.
McQuail, D., 1992. Media Performance Mass Communication and the Public Interest. SAGE, London.
Picard, R.G., 2002. The Economics and Financing of Media Companies. SAGE, London.
Seufert, W., 2006. Media statistics in Europe. In: Heinrich, J., Kopper, G.G. (Eds.), Media Economics in Europe. Vistas, Berlin, pp. 211–218.
Van Cuilenburg, J., 1999. On competition, access and diversity in media, old and new: some remarks for communications policy in the information age. New Media & Society 1 (2), 183–207.
Van der Wurff, R., Bakker, P., Picard, R.G., 2008. Economic growth and advertising expenditures in different media in different countries. Journal of Media Economics 21 (2), 28–52.
Vogel, H.S., 2011. Entertainment Industry Economics, a Guide or Financial Analyses, eighth ed. University Press, Cambridge.
Webster, J.G., Phalen, P.F., Lichty, L.L., 2006. Ratings Analysis: The Theory and Practice of Audience Research, third ed. Lawrence Erlbaum, Mahwah and London.

Data Bases and Statistical Systems: Crime

Clayton Mosher, Washington State University Vancouver, Vancouver, WA, USA
Terrance Miethe, University of Nevada Las Vegas, Las Vegas, NV, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by T. Makkai, volume 22, pp. 15053–15056, © 2001, Elsevier Ltd.

Abstract

Mass media, criminal justice system officials, and policy-makers all rely on crime statistics to make claims about crime levels and trends; legislators and criminal justice system officials use them to implement policies to address crime; and criminologists and other social scientists use crime statistics to conduct research on the causes of crime, to inform theories of crime, and to evaluate the effectiveness of crime control policies. In this article we discuss the three primary sources of crime statistics: official data (Uniform Crime Reports and the National Incident-Based Reporting System), victimization studies (National Crime Victimization Surveys), and self-report studies, focusing on their basic methodology and the various problems associated with their compilation.

Official Crime Data

Official crime data are derived from several functions of the criminal justice system, including police reports of offenses and arrests, charges filed by prosecutors, criminal complaints and indictments, imprisonment data, and prison releases. The first national crime statistics, based on judicial data, were published in France in 1927. In the United States, despite attempts by the Federal Bureau of the Census to collect crime statistics, as early as 1907, it was not until 1930 that such statistics became available at the national level. Due to space limitations and the fact that they are more widely used in discussions of trends and variations in crime across jurisdictions and over time, we focus here on two major sources of police data on crime – Uniform Crime Reports (UCRs) and the National Incident-Based Reporting System (NIBRS).

Prior to 1930, police reports of crime in the United States were not collected or compiled in any systematic way across jurisdictions. Some cities kept yearly counts of reported crime incidents and persons arrested, while other cities did not formally record such information. There was also widespread variation in the classification of crime across jurisdictions, with different community standards and legal definitions affecting how crimes were defined and whether particular activities were officially recorded as crimes. As such, both comparisons across jurisdictions and estimates of historical crime trends are problematic prior to 1930 due to the lack of uniformity in crime definitions and data collection.

In developing the UCR program in the late 1920s, the International Association of Chiefs of Police recognized that not all crimes were equally important. They therefore focused on seven types of crimes that were prevalent, generally serious in nature, widely identified by victims as criminal incidents, and most likely to be reported to the police. The original seven 'index' crimes (also referred to as 'Part I offenses') are murder and manslaughter, forcible rape, robbery, aggravated assault, burglary, larceny, and motor vehicle theft. Arson was added to the UCR as an index crime in 1979. Twenty-one less serious crimes and status offenses are classified as nonindex or Part II crimes.

The FBI compiles UCR data from law enforcement agencies and publishes annual data on crime and arrests broken down by age, sex, and race of individuals arrested, and the city and region of the country where the crime was committed. Although participation in the UCR program is voluntary, the proportion of law enforcement agencies that submit data to the program is impressive. A total of nearly 18 000 state, county, and city law enforcement agencies, covering more than 288 million US residents, submitted crime reports under the UCR system in 2008. A total of 95% of the US population is covered by this data source, with law enforcement agency participation rates being slightly lower in cities outside metropolitan areas (88%) and in rural areas (90%).

Two essential components of the UCR data system involve the classification and scoring of criminal offense. Classifying crime offenses in the context of the UCR refers to the process of translating offense titles used in particular local and state laws into the standard UCR definitions for Part I and Part II offenses. Scoring of offenses refers to counting the number of offenses after they have been classified under the UCR typology and entering the total count.

Classifying crimes can be a complex process, and there are several sources of ambiguity in the definition and coding of offenses that call into question the uniformity of reporting practices across jurisdictions. The primary sources of variability include differences across jurisdictions in their interpretation of crime incidents, the hierarchy rule, the diligence of record keeping, and the adequacy of follow-up procedures.

To take just one example of this ambiguity, in the specific case of homicide (widely believed to be the most reliably coded crime), the main obstacle to uniform reporting and counting involves the follow-up procedures, the timing of police investigations and UCR filing, and definitional issues in the classification of accidental killings and justifiable homicides. For example, the recording of situations of aggravated assaults that eventually become murders because the victim dies as a result of the assault assumes equal diligence and detailed record keeping across recording agencies in conducting follow-up investigations and correctly adjusting multiple monthly returns. Some less diligent law enforcement agencies may

simply count the aggravated assault and fail to record the subsequent death of the victim as a murder.

An interesting example of the confusion that can occur with respect to the coding of homicides comes from New York City, which, in 2006, saw an apparent increase in homicides from the previous year. However, part of this increase was created by an unusual number of deaths that were classified as homicides because the city's medical examiner determined they were related to crimes that had been committed in earlier years. Of the 25 such reclassified deaths in 2006, 12 were related to injuries that had occurred at least 14 years earlier, including a 72-year-old man who was shot in 1974 and died of pneumonia in 2006 (Vasquez, 2006).

The UCR's hierarchy rule applies to the classification and scoring of crimes when multiple offenses are committed at the same time. When the hierarchy rule is applied, only the most serious offense in the series is reported, and all others are not counted. For example, if an individual breaks into a house, steals items from the house, kills the owner of the house, and makes a getaway in a stolen car, only the murder will be recorded in official statistics, as it is the most serious offense.

National Incident-Based Reporting System

A relatively recent enhancement to the UCR program is the development of the NIBRS. In contrast to the UCR, which uses a summary or aggregate reporting approach, NIBRS categorizes each incident and arrest in one of 22 basic crime categories that span 46 separate offenses. A total of 53 data elements about the victim, property, and offender are collected under NIBRS (Maxfield, 1999). Although NIBRS data have been used in federally published reports on crime, as of 2012, 6115 law enforcement agencies (representing only 33.4% of all agencies) participated in the program, and more than half of the participating agencies were located in cities with fewer than 10 000 inhabitants (FBI, 2013). The low rate of participation in NIBRS by law enforcement agencies in more populous jurisdictions has led to questions about the accuracy of estimates of crime that derive from these data (Addington, 2007). And despite its promise in terms of improving the accuracy of crime statistics, several problems exist with NIBRS data. First and most obvious is the complexity of the coding scheme, which is documented in four volumes published by the FBI. Because many police officials may be less than diligent in following these coding requirements, missing data may be an even greater problem under NIBRS (Maxfield, 1999). In addition, because NIBRS eliminates the UCR's hierarchy rule, its adoption may lead to an increase in reported crime, creating potential public relations problems for law enforcement agencies which are, to at least some extent, evaluated on the basis of crime rates in their jurisdictions.

Problems with Police Data on Crime

Police reports are often considered to be the best official measure of the nature and extent of crime. Compared to prosecutorial, judicial, and correctional data, police reports are more comprehensive in their coverage of types of criminal offenses and include information on criminal incidents even when the offender has not been identified. However, as a measure of the true extent of crime in a jurisdiction, police statistics are inadequate for several reasons. The major problems with police data involve variation in citizen reporting of crime and police recording practices, possible race, and social class biases in the structure of policing, limited coverage of crime types under UCR data, conceptual and methodological factors that affect the classification of crime incidents and estimates of national crime rates, and political manipulation of these data by police departments.

The term 'dark figure' has been used by criminologists to represent the gap between the true extent of crime and the amount of crime known to the police. The primary sources of this gap are the inability of police to observe all criminal activity, the reluctance of crime victims and witnesses to report crime to the police, and variation in the recording of known crime incidents due to police discretion.

Although it is not well known, the vast majority of crime becomes known to the police through citizen complaints or calls for service and many citizens choose not to report crimes they have experienced. In fact, in 2012, only 44% of violent crime victims and 34% of property crime victims reported the incidents to the police (Bureau of Justice Statistics, 2013). The reasons why victims and other citizens do not report crimes to the police include a lack of trust in the police or reservations regarding the ability of police to solve crimes, and fear of retaliation and/or reprisals from offenders for reporting crime. In addition, the victims in some crime situations may also be involved in criminal activities themselves (e.g., drug sellers or prostitutes who are the victims of robbery), which decreases the likelihood of reporting.

Even if a crime is reported by citizens or directly observed by the police, there is no guarantee that such an offense will be recorded in official data. Police discretion, both across and within jurisdictions in recording an incident as a crime is a major source of inconsistency in official crime data. Donald Black (1970) notes that the police are more likely to file an official report if the crime is serious in nature, if the complainant prefers arrest (as opposed to informal action), if the criminal incident involves strangers rather than friends of family members, if the complainant shows deference toward the police, and if the complainant is of higher social status. There is also considerable evidence to suggest that race and social class biases exist in street-level policing, which may explain at least some of the difference between minority and white crime rates.

An additional and important limitation of official crime statistics involves their manipulation and fabrication for political purposes. Because the police have exclusive control over the dissemination of crime data and there is infrequent monitoring of the accuracy of their crime counts, one way to demonstrate effective law enforcement is to distort, manipulate, and fabricate the number and nature of crime reports. Local agencies can 'cook the data' through a variety of procedures, including not reporting all crime incidents, counting separate events as if they occurred in multiple offense incidents, declaring reported crimes as 'unfounded' so they are not counted in annual UCR summaries, and downgrading Part I (index) offenses to minor offenses so that they are not tallied nationally in the UCR summaries.

Numerous examples of fabrication of crime statistics have been identified, including in several major US cities (e.g., Atlanta, Philadelphia, Los Angeles, Detroit, Dallas, Phoenix, Milwaukee, Washington DC, New York, and Chicago). In New York, substantial reductions in crime rates over the 1995–2010 period have been attributed to aggressive and effective law enforcement practices, however, critics have alleged that at least part of this decline in crime rates is due to pressure from police management to reduce officially recorded crime (Silverman and Eterno, 2012). A more recent example of this manipulation comes from Chicago, where, according to data published by the police, the number of index crimes decreased by 56% over the 2010–13 period. However, an audit by the city's Inspector General found that the Chicago Police Department had undercounted aggravated assaults by 24%, and police sources indicated that they had been pressured by their superiors to underreport crime by misclassifying or downgrading offenses, counting multiple incidents as single events, and making it more difficult for people to report crime (Bernstein and Isackson, 2014).

Official Crime Data in International Context

The two major sources of international crime data are from the International Police Agency (Interpol) and the United Nations. Interpol crime reports are published in four languages (Arabic, English, French, and Spanish) and include data on murder, sex offenses, serious assault, theft, aggravated theft (robbery and violent theft and breaking and entering), theft of automobiles, fraud, counterfeit currency offenses, and drug offenses. Although these Interpol data are provided for multiple countries, Interpol reports provide cautions about the appropriateness of using these data for cross-national crime comparisons. This caution is necessary because the statistics do not account for definitional differences in crime categories across various countries, the diversity of methods used by different countries in compiling the statistics, and changes in laws or data collection procedures over time. In addition, as revealed from the International Crime Victimization Survey, the proportion of crime not reported to law enforcement varies substantially across nations and with respect to different types of crime. Similar problems exist in using the United Nations Crime Surveys for cross-national comparisons.

Victimization Surveys

At least partially in response to concerns about the limitations of official crime data, victimization surveys, which ask people about their experiences as victims of crime, were initiated. The first large-scale victimization survey in the United States appeared in the late 1960s, and since that time, they have been widely used to measure the frequency and characteristics of particular types of crimes and the demographic profiles of victims, both in the United States and in other countries. By eliciting information about both crimes that citizens report to the police and those they do not, victimization surveys provide further information regarding the dark figures of crime. Data from these surveys have also been widely used to develop and modify theories of crime.

As implied by the basic definition, these surveys involve self-reports of victimization experiences by victims themselves, and as such, are subject to many of the same problems associated with other forms of survey research more generally. Victimization reports are usually generated from random samples of the general public, and a variety of screening questions are used to identify different types of victims. A number of crimes captured in other data sources (for example, prostitution and alcohol and other drug offenses) are not covered in these surveys because such crimes are considered to be victimless. In addition, in some crime situations, it is not feasible to interview the victim – an obvious example would be in a situation of homicide. Finally, some crimes, such as vandalism, are viewed as trivial and are not covered by victimization surveys; others, such as white-collar and corporate crime, are seen as difficult to measure accurately. The lack of coverage of certain types of crimes makes direct comparisons between victimization and official data problematic.

The National Crime Victimization Survey (NCVS) is the most comprehensive and systematic survey of victims in the United States. The sampling procedure is supervised by the U.S. Census Bureau; the survey is conducted by well-trained staff and interviewers; and changes in the sampling design and format of questions are rigorously evaluated in terms of their effects on estimation of victimization experiences.

With respect to the selection of households to participate in these surveys, with the goal of obtaining a nationally representative sample, the NCVS uses a complex, stratified, multistage cluster sample, with what is known as a rotating panel design. This means that sample households are organized in panels and each panel is divided into six groups so that interviews are ongoing throughout the year, thereby reducing seasonality effects. Residents of sample households are interviewed a total of seven times – once every 6 months for 3 years – and once seven waves of interviews are complete, the panel is rotated out of the sample and replaced by new households.

Crime incident reports in the NCVS involve a series of questions about the particular crime event (that have occurred in the previous 6 months), the offending parties, and the consequences of the crime. For each separate victimization incident identified by the respondent, they are asked whether the crime was reported to the police; whether the offense was completed or merely attempted; whether the offender was identified or known to the victim; the demographic (race, gender, age) characteristics of the offender (if known); whether a weapon was used in the crime; whether the offender was a member of a gang or under the influence of alcohol or other drugs at the time of the incident; whether the victim resisted; and the amount of monetary loss or physical injury (or both) that resulted from the victimization.

Similar to the other two major sources of crime statistics, victimization surveys are susceptible to conceptual and methodological problems that call into question their accuracy as measures of crime. In addition to the fact that they do not cover all crimes that may be of interest to criminologists and policymakers, victimization surveys are subject to sampling bias. For example, particular groups of people are less likely to participate in these surveys than others, and the excluded groups tend

to be more prone to victimization. In the NCVS, homeless persons, young males, and members of minority groups are less likely to be included, and each group has higher rates of victimization than their older, female, and nonminority counterparts. At the other extreme, the very wealthy are likely underrepresented in victimization surveys because of their ability to isolate themselves from interviewers. The exclusion of victimization experienced by businesses and crimes against the government in household victimization surveys can also be viewed as a source of sampling bias that dramatically lowers estimates of national victimization rates.

Another problem with victimization surveys is related to the fact that a variety of factors influence participants' responses, including question wording and reference periods, and basic limitations of human judgments. Both incident rates and subgroup variation in victimization risks are affected by the particular screen questions used to elicit reports of victimization experiences. Short screen questions may cue a respondent's recall of only a small subset of incidents that involve the most serious or frequent violations, whereas longer screens encourage the recounting of a fuller range of experiences. With respect to reference periods, surveys that use a reference period of 1 year or more are susceptible to forward telescoping (i.e., remembering events as occurring more recently than they actually occurred). The 6-month reference period used in the NCVS makes this survey prone to backward telescoping (i.e., remembering incidents as having occurred in a more distant past).

Victimization surveys are also an imperfect measure of crime because of the inherent fallibility of human information processing and judgment. People experience lapses in memory, selectively perceive and misperceive particular actions, and interpret actions and events from their own perspectives. In the case of reports of victimization, people may overestimate or underestimate their experiences through outright deception, exaggeration, embarrassment, or misinterpretation.

Self-Report Studies

Self-report studies of crime were developed in the 1940s and 1950s, largely in response to concerns among criminologists that official measures of crime were systematically biased and provide a distorted picture of the nature of crime and its correlates. Although there are exceptions, self-report surveys are typically administered to youth populations, and among the most prominent and widely used self-report studies in the United States are included the National Youth Survey (NYS), the National Longitudinal Study of Adolescent Health (Add Health), the Monitoring the Future Survey, and the National Household Survey on Drug Abuse (now known as the National Survey on Drug Use and Health). Although there are variations in the specific methodology of each of these studies (and other self-report studies) in order to illustrate the general principles of self-report studies, we briefly discuss the NYS.

First conducted in 1977, the NYS was designed specifically to provide both prevalence and incidence estimates of the commission of delinquent activities by youth. It is a longitudinal survey that uses a national probability-based sample of young people who were 11–17 years old at the time of the first interview. Participants in this study were interviewed in their homes at 1-year intervals through 1981 and at 2–3-year intervals at least through 1995. More than 90% of the original 1725 participants have remained in the NYS over time.

The methodology of the NYS involves confidential, face-to-face interviews to solicit information on the number of times the respondent has engaged in a specific delinquent or criminal activity within the past calendar year, with two different response sets used. If an individual's response to an open-ended question indicates that they have engaged in the particular activity (e.g., destroying property, carrying weapons, physical assault, using or selling drugs) more than 10 times, the interviewer asks the youth to select one of the following responses: (1) once a month, (2) once every 2–3 weeks, (3) once a week, (4) 2–3 times a week, (5) once a day, or (6) 2–3 times a day. Although described in nonlegalistic terms, the 47 activities asked about in the NYS directly parallel offenses listed in the FBI's UCR. Of the Part I offenses, only homicide is excluded; about 75% of Part II offenses are included, along with a wide range of misdemeanors and status offenses.

Collectively, since their inception in the 1940s, self-report measures have consistently revealed dimensions of crime and its correlates that were either at odds with or could not be addressed by official statistics. For example, in self-report surveys, girls and women report being just as delinquent and criminal, although less frequently and intensely, as do boys and men. Whites similarly admit to involvement in a range and number of delinquent and criminal acts closely paralleling that of blacks. Middle- and upper-class youth self-report similar levels of involvement in delinquent activities as lower-class youth.

Self-report data, however, are by no means without their weaknesses and are subject to problems of reliability and validity. These data are constrained by the same elements that affect self-reports of all types of behaviors – respondents to surveys often give inconsistent or incorrect answers. The form of the question, the choice of words in the question, the response categories, and the position of the question in a series of items can affect what respondents think the question means; people may also answer incorrectly because they do not remember certain things. As such, significant underreporting of crime and deviant behavior can occur on self-report surveys, and there are some indications that members of minority groups (in particular, African-American males) are more likely to underreport their involvement in crime (Thornberry and Krohn, 2000). On the other hand, there is evidence that some people will overreport their involvement in criminality and deviance on these surveys (Mosher et al., 2011). With respect to validity, some critics have noted that self-report surveys tend to focus on relatively trivial, noncriminal acts and behaviors (such as skipping school) and as such are not directly comparable to official crime data.

A useful check on the accuracy of self-report data involves comparisons of respondents' self-reports of drug use to the results of urinalysis tests in the Arrestee Drug Abuse Monitoring (ADAM) Program administered at local jails in the United States. Under this program, inmates are asked a series of questions regarding their use of particular substances (among other questions) and are also subject to urinalysis tests. Data from ADAM reveal that there is generally a higher concordance

rate for marijuana use, that is, more arrestees will truthfully report the use of this drug. However, underreporting is much more common for harder drugs such as heroin and cocaine. For example, 1998 ADAM data from New York City indicated that 47% of arrestees reported using cocaine, but 75% had positive urinalysis results for the substance (Golub et al., 2005). The findings of differential honesty in reporting by type of illegal substance is consistent with what is known as 'social desirability theory,' which suggests that the distortion of self-reports, by under- or overreporting, is related to the acceptability of the behavior in question (note that this principle would also apply to behaviors other than drug use) (Edwards, 1957). Because the use of marijuana is less stigmatized than the use of hard drugs, respondents may be more likely to truthfully report using the substance.

What the Data Tell Us about Crime

Official data and victimization surveys reveal somewhat different patterns with respect to trends in crime rates in the United States. Crime rates as measured by the UCR increased substantially between 1960 and 1990, and have declined since the mid-1990s – this pattern has been similar for both violent and property crimes. However, annual crime victimization surveys show trends that are qualitatively different from those based on official police data. Specifically, the victimization data reveal that violent victimization rates remained fairly stable between the early 1970s and the mid-1990s, before declining in a similar fashion to the trend revealed in the UCR data. In contrast to the pattern of an increase followed by a decrease in property crimes shown in the UCR data, victimization surveys indicate that property crime has exhibited an essentially continuous decline since the mid-1970s. This slight contradiction in crime trends based on comparisons of UCR and NCVS data is largely the result of differences across these methods in their coverage of crimes, rules for counting crime incidents, and the population base from which crime rates are calculated.

The convergence of findings across the various methods of measurement is more pronounced with respect to the nature of criminal offenses. Contrary to impressions derived from mass media coverage of the issue, all three sources indicate that most crime in the United States involves relatively minor property offenses (as opposed to serious violent offenses) that occur with considerable frequency. In addition, the vast majority of violent offenses are attempts or threats that involve little or no injury to the victim. It is important to emphasize, however, that none of the three sources of crime statistics provide information on white-collar or corporate crime, organized crime, or crimes committed by governments and their agents.

In terms of the social distribution of crime, all three sources generally indicate that criminal offenders are disproportionately young, male, and members of ethnic or racial minority groups. UCR data indicate that approximately 15% of those arrested are under 18 years of age, 76% are male, and 28% are black. Although the differences across some groups are less pronounced, the NCVS data reveal similar patterns with respect to the demographics of criminal offenders. Smaller differences by age, sex, and race are found in studies of self-reported criminal behavior.

To conclude, all three primary sources of crime statistics – the UCR, victimization, and self-report studies, are characterized by methodological weaknesses. It is important for consumers of these data (including legislators, criminal justice system officials, and mass media) to adopt a healthy skepticism in interpreting crime data.

See also: Immigration, Applied Research on; Immigration, Citizenship, and Integration: Social Work Connections; Immigration: Social Psychological Aspects; Migration: Anthropological Perspectives.

Bibliography

Addington, L., 2007. Using NIBRS to study methodological sources of divergence between the UCR and NCVS. In: Lynch, J., Addington, L. (Eds.), Understanding Crime Statistics: Revisiting the Divergence of the NCVS and the UCR. Cambridge University Press, New York, pp. 225–250.

Bernstein, D., Isackson, N., May 2014. The Truth about Chicago's Crime Rates. Chicago Magazine. From: http://www.chicagomag.com (retrieved 05.05.14.).

Black, D., 1970. The production of crime rates. American Sociological Review 35, 733–748.

Bureau of Justice Statistics, 2013. Criminal Victimization. Washington, DC.

Edwards, A., 1957. The Social Desirability Variable in Personality Assessment and Research. Dryden, New York.

Federal Bureau of Investigation, 2013. National Incident-Based Reporting System. U.S. Government Printing Office, Washington, DC.

Golub, A., Liberty, H., Johnson, B., 2005. The variation in arrestees' disclosure of recent drug use across locations, drugs, and demographic characteristics. Journal of Drug Issues 35, 917–940.

Maxfield, M., 1999. The national incident-based reporting system: research and policy applications. Journal of Quantitative Criminology 15, 119–149.

Mosher, C., Miethe, T., Hart, T., 2011. The Mismeasure of Crime, second ed. Sage Publications, Thousand Oaks, CA.

Silverman, E., Eterno, J., 2012. The Crime Numbers Game: Management by Manipulation. CRC Press, Boca Raton, FL.

Thornberry, T., Krohn, M., 2000. The self-report method for measuring delinquency and crime. In: Duffee, D. (Ed.), Measurement and Analysis of Crime and Justice: Criminal Justice 2000. National Institute of Justice, Washington, DC, pp. 38–84.

Vasquez, E., 29 September, 2006. Despite Continuing Decrease in City Crime, Troubling Signs Emerge. New York Times. From: http://www.nytimes.com (retrieved 29.09.06.).

Data Bases and Statistical Systems: Crime and Delinquency

Mark Warr and David S Kirk, University of Texas at Austin, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by T. Makkai, volume 22, pp. 15053–15056, © 2001, Elsevier Ltd.

Abstract

Criminal and delinquent behavior is inherently difficult to measure because it is surreptitious behavior and most crimes are never reported to the police. This article reviews the primary methods used by criminologists in the United States to measure criminal and delinquent behavior, and describes the advantages and limitations of each method. It provides historical context to the development of crime statistics and identifies some of the major sources of data on crime and delinquency available today.

Introduction

Crime and criminal justice statistics serve many audiences and many purposes. Journalists use crime statistics to report putative levels and changes in crime to the general public. Academics rely on them to investigate the causes of crime, as well as its spatial, temporal, and demographic features. Statistics from law enforcement agencies (e.g., police, jails, courts) are used to assess the effectiveness, efficiency, and fairness of those institutions. Politicians often use crime statistics to frighten or mollify the public to their own political advantage, while popular magazines rank cities according to their ostensible safety. Private organizations like rape crisis centers and domestic violence organizations use them to assess their effectiveness or to justify their existence. As with statistics on unemployment or public health, crime statistics often carry great weight, defining places or periods as dangerous or safe, even when those characterizations may be unwarranted.

Like other social statistics, crime statistics carry an aura of precision and accuracy that is alluring to many. 'Crime rate drops 9% last year.' 'Police clear 56% of homicide cases.' But like all social statistics, crime statistics are the outcome of complex processes that entail many potential sources of measurement error. Intelligent users need to understand and appreciate the uncertainties that inhere in such statistics, and their implications.

This article concentrates on the ways that criminologists measure crime and, in so doing, attempt to circumvent or overcome the inherent difficulties in measuring such behavior. Perhaps the chief difficulty is that criminal behavior is surreptitious behavior. Those who engage in it actively seek to hide their identity and their behavior from public and law enforcement scrutiny. This, along with other problems, makes the measurement of crime particularly difficult and, at times, impossible or unethical.

Official Crime Statistics

Uniform Crime Reports

As part of their public mandate, most police departments and other law enforcement agencies issue regular statistical reports. In the United States, the vast majority of local police departments submit data annually to the Federal Bureau of Investigation (FBI), which compiles and summarizes the data each year in a report entitled 'Crime in the United States: Uniform Crime Reports (UCR).' The UCR strives to standardize definitions of crimes across states and reports the number of offenses known to the police; the number of arrests; and clearance rates for counties, metropolitan areas, states, and the nation as a whole. The UCR covers a wide range of offenses, including violent offenses like assault and rape and property crimes like burglary and larceny, and it has its counterpart in many other countries, from Australia to England to Germany.

Established in 1929, the UCR is the oldest crime reporting system in the United States, and it was for decades the sole source of crime data for journalists, criminologists, and others. Even today it remains widely used by journalists interested in reporting crime trends and related topics. With a few exceptions, however, UCR data are plagued by numerous forms of systematic measurement error. By nature, UCR data are reactive, meaning that they rely on victims and witnesses to report crimes to the police. This requires citizens to both interpret an event as a crime and decide to report it to the police. For a variety of reasons (e.g., fear of reprisal, lack of confidence in the police, fear of social stigma), most crimes that take place in the United States are never reported to the police. This problem is so severe that the UCR itself acknowledges it and cautions users against inappropriate inferences from their data, such as ranking cities and counties by their crime rate. Other problems pertain as well. For example, in criminal events that involve more than one crime (e.g., a residential burglary that includes assault and rape), the UCR records only the most serious offense, thereby systematically undercounting crime. Furthermore, the political pressures associated with being labeled a 'high-crime' city (or other jurisdiction) in the UCR lead some jurisdictions to intentionally underreport crimes or misreport their population size to the FBI in order to lower their crime rate.

In fairness to the UCR, a few crimes (homicide, auto theft, arson) are, by all indications, accurately measured by the UCR (i.e., they can be confirmed by other sources), but they are the exception to the rule, and for other crimes (e.g., rape), UCR estimates are effectively useless for estimating the incidence of the crime.

One traditional weakness of the UCR has been the failure to disaggregate crime statistics by ethnicity (e.g., arrest rates for Latinos). However, the UCR will purportedly begin collecting information on ethnicity soon, thereby allowing analysis of arrests by ethnicity.

Alternative Crime Reporting Systems

In recent years, the FBI has been developing the National Incident-Based Reporting System (NIBRS). Unlike traditional summary (i.e., aggregate) statistics available in the UCR, NIBRS provides detailed data on individual incidents of crime, making these data much more useful to researchers. Nevertheless, NIBRS data pertain only to those offenses that come to the attention of the police.

The phrase 'official statistics' is most commonly used to refer to police or UCR data on crime, but it is also used to characterize data from other governmental institutions, from federal courts to prisons to parole agencies to the US Marshals or the Drug Enforcement Agency. The Bureau of Justice Statistics (BJS), a component of the US Department of Justice, is a massive clearinghouse of data on crime and criminal justice in the United States, and its Website offers ready access to these data. The BJS also sends notifications and reports to subscribers via e-mail. Another compendium of crime and criminal justice statistics is the Sourcebook of Criminal Justice Statistics, compiled by the School of Criminal Justice at the University at Albany, and also available online. Much of the data contained in the Sourcebook are presented as time series, making them unusually helpful and valuable to investigators.

Self-Report Data

As early as the late 1940s, some criminologists were attempting to develop an alternative to the UCR for measuring crime, and the result came to be known as the self-report method (e.g., Porterfield, 1943). According to Thornberry and Krohn (2000: 34), "the development and widespread use of the self-report method of collecting data on delinquent and criminal behavior is one of the most important innovations in criminological research in the 20th century." Setting aside police reports altogether, self-report researchers confidentially interview members of a population (usually a sample) and ask them to report their own criminal or delinquent behavior during some designated period of time (usually a year). By the 1960s, the self-report survey had grown to become one of the dominant methods of measuring crime in the United States, and its adoption by researchers resulted in a number of startling discoveries. For example, the 'dark figure of crime,' or the difference between UCR figures and the true incidence of crime, turned out to be much larger than criminologists had anticipated. Additionally, among juveniles (the age-group with the highest rate of offending), violating the law proved to be group behavior in most instances rather than the action of solitary offenders. The certainty of apprehension and punishment for many crimes was also discovered to be vanishingly low, and criminologists were startled to learn that most adolescent males break the law at least occasionally. Conformity to the law, in other words, is not the rule but the exception.

As with official data, self-report research comes with its own set of issues and problems. For convenience and other reasons, most self-report surveys are conducted on juveniles, usually high school students. Self-report research on adults is rare. Ethical and legal questions are abundant in self-report research. Should investigators ever disclose the information they acquire to the police? Are respondents legally liable for the information they disclose? The accuracy of self-reports is another issue. Do respondents tell the truth? Some respondents misrepresent their behavior in long surveys because of survey fatigue – i.e., they get tired or bored of answering questions. Additionally, are respondents even able to accurately recall their behavior, and, if so, when it occurred? The more frequently individuals engage in criminal activity, the less salient any one incident will be in memory. Surprisingly enough, though, research indicates that the reliability of self-report data is high, and their validity generally falls within acceptable scientific limits when surveys are properly designed and administered (Thornberry and Krohn, 2000).

Once it arrived on the scene, self-report research quickly became the principal means in criminology for testing etiological theories of crime, because the investigator has direct access to the offender and can ask virtually any question imaginable. By contrast, police data are well removed from the offender and the event. Self-report data are also the only means for measuring the prevalence of crime, or the number of offenders in a population. Today, a number of national, longitudinal self-report studies have been funded by the federal government in the US, including the National Youth Survey, the National Longitudinal Study of Adolescent Health, and the Monitoring the Future survey. They contain what are widely considered to be the best available data on topics like drug use and adolescent sexuality.

Victimization Data

Victimization Surveys

As self-report methods were gaining ground, yet another alternative methodology for measuring crime was underway: the victimization survey (Cantor and Lynch, 2000; Mosher et al., 2011). In a victimization survey, researchers survey a population (again, usually a sample) and ask respondents to report incidents of crime committed against them or their household during a specified period of time (usually 6 months or 1 year). Following a series of pilot studies on the feasibility of such surveys, the US Department of Justice joined with the US Census Bureau in 1973 to create what is today called the National Crime Victimization Survey (NCVS), an ongoing sample survey of about 90 000 households and 160 000 individuals each year. This large sample size is necessary because some crimes are statistically rare events and require very large samples to obtain statistically reliable estimates. The NCVS measures fewer offenses than the UCR or most self-report surveys (rape and sexual assault, robbery, aggravated and simple assault, theft, household burglary, and motor vehicle theft), but its estimates of some crimes are considered to be the best available.

Victimology

The NCVS appeared at a time when criminologists were beginning to approach crime from a different perspective: the victim's perspective. This change in focus resulted in a new subfield in criminology named victimology. Rather than asking what kinds of people burgle homes, for example, victimologists ask an equally important question: What types of homes get burgled? This unique perspective generated a new body of knowledge in criminology and helped to align the field with emerging research in social science on risk and risk perception. Victimization surveys also encouraged the growth of routine activities theory, opportunity theory, and rational choice theory (Cantor and Lynch, 2000), and it fueled interest in forms of 'indirect victimization' (Conklin, 1975), that is, impacts of crime that extend beyond those who are direct victims of crime. A prime example is fear of crime, which criminologists discovered to be far more common than victimization itself, and stunning in its effects on American lifestyles (Warr, 2000). For example, a startling proportion of urban women in the United States say that they avoid going out at night or going out alone, and more than half of adult Americans say that there are places in their town they avoid. Americans worry not only about their own safety but also about the safety of others, particularly their partners and children (Warr and Ellison, 2000).

As with any source of crime data, victimization surveys have their own problems and limitations. Some crimes (e.g., so-called victimless crimes like gambling and prostitution) cannot be measured via victims. Victimization surveys can be enormously expensive (particularly, if they rely on face-to-face interviews and large geographic areas) and victim reports, as with self-reports, are susceptible to memory or recall errors on the part of respondents as well as 'telescoping' problems where respondents accurately recall events but move them backward or forward in time. These and other sources of error are relatively well understood in the NCVS, however, and the NCVS uses a sophisticated rotating panel design to 'bound' its estimates in time, as well as high-quality sampling and interview designs.

International Victimization Surveys

Today, nationwide victimization surveys are carried out in many countries. An early program was the British Crime Survey, now known as the Crime Survey for England and Wales. This survey has enhanced the field of comparative criminology and has been more amenable to innovation than its more staid ancestor, the NCVS. The International Crime Victims Survey collects data from 18 European countries with support from the United Nations and the European Commission. Among other contributions, its data on public perceptions and reactions to crime (including fear of crime) have advanced our understanding of the social impact of crime. The International Self-Report Delinquency Study is a self-report survey of seventh, eighth, and ninth graders in 31 countries (including China and Venezuela) that is administered by researchers in the School of Criminology and Criminal Justice at Northeastern University. It promises to increase our knowledge of substance abuse, delinquency, and victimization worldwide using standardized instruments and data collection procedures. Clearly, advances once limited to the United States are now spreading widely across the world.

Conclusion

As we have seen, criminologists have been innovative in devising methods for measuring crime, methods that circumvent its elusive nature. Each advance has led to major discoveries and theoretical advances. One can only imagine where the field of criminology might be if it were still confined to official data. Looking toward the future, the US Department of Justice is currently laying the foundation for a national self-report survey of rape and sexual assault, the first of its kind, and a major undertaking. The FBI continues to expand its NIBRS program, and the list of countries developing their own national victimization surveys continues to grow. These sorts of developments bode well for the future of criminology, both theoretically and methodologically.

This article has concentrated attention on the United States because self-report and victimization methodologies were initially developed there (but certainly not official statistics, which appeared in European countries as early as the nineteenth century).

See also: Crime Trends and Debates; Crime and Class; Crime and Gender; Crime, Race and Ethnicity; Crime: Knowledge about and Prevalence; Data Bases and Statistical Systems: Crime Measurement; Data Bases and Statistical Systems: Crime; International and Comparative Criminology; Organizations and Culture; Property Crime: Victimization and Trends.

Bibliography

Cantor, David, Lynch, James P., 2000. Self-report surveys as measures of crime and criminal victimization. In: Duffee, David (Ed.), Criminal Justice 2000. Volume Four: Measurement and Analysis of Crime and Justice. U.S. Department of Justice. National Institute of Justice, Washington, DC, pp. 85–138.

Conklin, John E., 1975. The Impact of Crime. Macmillan, New York.

Mosher, Clayton J., Miethe, Terance D., Hart, Timothy C., 2011. The Mismeasure of Crime. Sage Publications, Thousand Oaks, CA.

Porterfield, Austin L., November 1943. Delinquency and outcome in court and college. American Journal of Sociology 49, 199–208.

Thornberry, Terence P., Krohn, Marvin D., 2000. The self-report method for measuring delinquency and crime. In: Duffee, David (Ed.), Criminal Justice 2000. Volume Four: Measurement and Analysis of Crime and Justice. U.S. Department of Justice. National Institute of Justice, Washington, DC, pp. 33–83.

Warr, Mark, 2000. Fear of crime in the United States: avenues for research and policy. In: Duffee, David (Ed.), Criminal Justice 2000. Volume Four: Measurement and Analysis of Crime and Justice. U.S. Department of Justice. National Institute of Justice, Washington, DC, pp. 451–489.

Warr, Mark, Ellison, Christopher G., November 2000. Rethinking social reactions to crime: personal and altruistic fear in family households. American Journal of Sociology 106, 551–578.

Data Bases and Statistical Systems: Crime Measurement

Ronet Bachman, University of Delaware, Newark, DE, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, the various sources used to measure criminal victimization will be examined including police report data compiled by the Federal Bureau of Investigation, surveys conducted by the Bureau of Justice Statistics, and other surveys sponsored by the National Institute of Justice and the Centers for Disease Control and Prevention. After the basic differences across these types of data collection are examined, several crimes that are more difficult to measure will be examined including rape and sexual assault, intimate partner violence, and stalking.

Introduction

When we hear the word 'crime,' most of us think instinctively about the stereotypical stranger lurking in the dark, preparing to attack an unsuspecting victim. When we are asked, "How do we know how much crime there is?" most of us probably first think of police reports. In reality, the contextual circumstances of most crimes do not involve strangers and reports of criminal victimization to the police represent less than half of all victimizations, and for some crime, much less. In this article, you will learn how crime is measured in the United States. In addition, the measurement of several specific 'hard to measure' crimes will also be examined. By the end of the article, you will understand that the measurement of crime is a particularly difficult task, and the estimates of victimization you get are only as good as the methods used to measure them.

This article is not meant to exhaustively review all of the sources of information on crime, nor is a summation of the strengths and weaknesses of each method of collecting data provided. There are far too many data sets attempting to measure crime, each encompassing different specific collection methodologies and populations, and each with its own unique set of strengths and weaknesses. Instead, the goal is to provide a brief introduction to a few of the main sources of data on crime and to discuss some of the primary pitfalls and shortcomings commonly found in attempts to accurately identify the scope and magnitude of criminal behavior. Despite the technical nature of this article, measurement is a very important issue. The accuracy of our insights and explanations about crime is largely dependent on the quality of the information about crime that we are able to gather. Moreover, resources and strategies aimed at preventing crime and helping crime victims are also based on these estimates. Bad information can and does result in poor choices being made by policy makers, politicians, activists, and other concerned citizens. Depending on who is gathering the data and what methods they employ to get that information, the results can vary widely (Alvarez and Bachman, 2008).

Reports to Law Enforcement Officials

The most widely used source of statistical information about crime in the United States is the Uniform Crime Reporting Program (UCR), compiled by the Federal Bureau of Investigation (FBI). The UCR has collected information about criminal incidents reported to the police since 1930. The information collected by the FBI is based on the voluntary participation of state, county, and city law enforcement agencies across the United States.

For the crime of homicide, information about both the victim and the offender (e.g., the gender and race of both, the relationship between the victim and offender, the weapon used) is obtained in a separate reporting program called the Supplementary Homicide Reports. Unfortunately, such detailed information is not collected for other crimes in the UCR. To remedy this problem, in 1988, the FBI implemented a change in its collection of crime information to include more characteristics of the incident; appropriately, this is called the National Incident Based Reporting System (NIBRS). NIBRS data are more specific than the UCR and includes many more offenses that local agencies have to report information on. It includes detailed information on crime incidents, including the characteristics of the victim, such as age, gender, race, ethnicity, and resident status. In all, NIBRS categorizes each incident and arrest in one of 22 basic crime categories that span 46 separate offenses. A total of 53 data elements about the victim, property, and offender are collected under NIBRS (Mosher et al., 2002). As you can imagine, it takes a great deal of time and money to make this change and fill out this paperwork at the local police department level. Because of this time and expense, just over half of all states currently use the NIBRS format for collecting information about reported crimes while the others retain the older UCR version.

Both the UCR and the NIBRS data collection methods are problematic when estimating incidence rates of criminal victimization. It's not hard to imagine why. If victimizations are not reported to police, they are never counted in either data collection effort in the first place, and based on comparisons with national survey data, it is estimated that only about 40–50% of crimes become known to police. This is particularly problematic for certain types of violence, such as rape and violence that occurs between intimates, such as spouses and boy/girlfriends. We know that a large percentage of these victimizations are never reported to police. In sum, there is a great deal of evidence documenting the large gap between the true extent of victimization and offending and the amount of crime known to police. The major sources of this gap, according

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45055-5

to Clayton Mosher, and his colleagues (Mosher et al., 2002) are the following: the inability of police to observe all criminal activity, the reluctance of crime victims and witnesses to report crime to the police, and variation in the recording of 'known' crime incidents because of police discretion.

Because of this weakness in police reports, random sample surveys of the population are often used as the social science tool of choice for uncovering more accurate information about criminal victimization. However, as can be imagined, surveys employing diverse methodologies and different definitions of violence have resulted in tremendously diverse estimates. Taking violence against women as an example, survey estimates of how many women experience violence from an intimate partner annually range from 9.3 per 1000 women to 116 per 1000 women (Bachman, 2000). This is a huge range. Further, the methodological differences across survey methodologies often preclude direct comparisons across studies.

The measurement of specific crimes will be reviewed in subsequent sections. Here we will briefly summarize to types of surveys that are designed to more accurately measure crime victimization. The first is the only national ongoing survey that tracks criminal victimization over time. It is sponsored by the Bureau of Justice Statistics of the U.S. Department of Justice and is called the National Crime Victimization Survey (NCVS). Instead of focusing on victimizations, the second survey we will discuss was designed to measure the offending behavior of adolescents and is called the National Youth Survey (NYS).

The NCVS annually interviews over 100 000 individuals of 12 years of age or older and is the second largest ongoing survey sponsored by the U.S. Government. It measures both nonlethal violent and property crime victimizations. In addition, identity theft has also been added to the list of crimes covered by the NCVS along with other questions that are asked on various supplemental surveys periodically. Excluding the identity theft screening questions, the NCVS currently uses the screening questions listed in **Table 1**.

As you can see from **Table 1**, the screening questions rely on behavior-specific wording instead of asking directly about victimizations using crime jargon such as "have you ever been robbed?" (Bureau of Justice Statistics, 2013). This is important. A great deal of research has demonstrated that asking questions using behavior-based wording instead of legally based phrases uncovers a significantly greater number of victimizations, particularly when victims may not self-identify themselves as crime victims. As you might imagine, asking people about their experiences in these numerous behavior-specific ways and with many cues for types of offenders (e.g., a stranger or someone you know well) and types of locations (e.g., in a vacation home, in a laundry room) uncovers many more victimizations than those reported to police.

These types of victimization surveys give us much more accurate estimates of crime victimization, but who is most likely to commit crime? Relying on police reports to estimate characteristics of offenders involves the same problems as using these data to estimate who is most likely to be victimized. Are offenders who are arrested for violent offending actually representative of all offenders? The quick answer is no. Not surprisingly, early self-report surveys of offending behavior in the 1940s revealed that a relatively large number of committed offenses were undetected by the police. Although police report data at the time indicated offenders were more likely to be minorities from low socioeconomic backgrounds, self-report data revealed that a great number of offenses were being reported by people from relatively privileged backgrounds. As you might guess, these offenses rarely came to the attention of the police, and when they did, they rarely resulted in an arrest (Short and Nye, 1958). Based on these early studies, researchers interested in offending behavior – like those interested in victimization – began to rely on survey methodology instead of police reports. That trend continues to this day.

One of the most thorough contemporary surveys of offending behavior is the NYS, which was first collected in 1976 from a national probability sample of 11- to 17-year-olds. These youth were interviewed many times during the following years, with the last interview data collected in 1995, although there are similar surveys that have taken its place. **Table 2** displays some of the questions used to measure the violent offending behavior in the NYS. As you can see, here too researchers have used behavior-specific wording instead of relying on the use of crime categories and labels.

Rape and Sexual Assault

Like all victimizations, estimates of rape and sexual assault victimization can be obtained from several different sources. The important question remains, however, which is the most accurate? Recall that the FBI's UCR and its NIBRS include only those victimizations that the police find out about, either through the victim or someone else. Because rape and sexual assaults remain the least likely forms of violence to be reported to police, our best estimates about how many victimizations there are, and who is most likely to be victimized and by whom, remain problematic. The most reliable information on rape victimization – imperfect and incomplete as it is – comprises the data obtained from social scientific surveys.

Even surveys, however, can have results that vary widely depending on how the questions are worded, the composition of the sample obtained, and the context of the survey. The only survey that monitors rape and sexual assault on an annual basis is the NCVS. However, as presented in **Table 1**, the questions used to uncover rape are probably the least behaviorally specific compared to the other national surveys that have been conducted. The questions used for the NCVS (**Table 1**) are vague and these broad questions lend themselves to different types of interpretations. Perhaps the perpetrator was an acquaintance or even a husband, and the victim may therefore not define it as rape even though it is considered as such by all state legal statutes. Or perhaps there was no overt use of force, so the victim may not define it as an 'attack.' The word rape itself has a variety of meanings that are used differently depending upon the situation and the individuals involved. Because of this, these types of questions about rape are prone to a great deal of underreporting. In

Table 1 Screening questions used for the National Crime Victimization Survey (excluding identity theft questions)

1. I am going to read you some examples that will give you an idea of the kinds of crimes this study covers. As I go through them, tell me if any of these happened to you in the last 6 months.
   Was something belonging to YOU stolen, such as
   a. things you carry, like luggage, a wallet, purse, briefcase, book?
   b. clothing, jewelry, or cell phone?
   c. bicycle or sports equipment?
   d. things in your home – like a TV, stereo, or tools?
   e. things outside your home such as a garden hose or lawn furniture?
   f. things belonging to children in the household?
   g. things from a vehicle, such as a package, groceries, camera, or CDs?
   Did anyone ATTEMPT to steal anything belonging to you?
2. Other than any incidents already mentioned, has anyone
   a. broken in or ATTEMPTED to break into your home by forcing a door or window, pushing past someone, jimmying a lock, cutting a screen, or entering through an open door or window?
   b. illegally gotten in or tried to get into a garage, shed, or storage room?
   c. illegally gotten in or tried to get into a hotel or motel room or vacation home where you were staying?
3. If respondent has vehicle(s) during the last 6 months,
   a. was the vehicle(s) stolen or used without permission?
   b. Did anyone steal any parts such as a tire, car stereo, hubcap, or battery?
   c. Did anyone steal any gas?
   d. Did anyone ATTEMPT to steal any vehicle or parts attached to it/them?
4. Were you attacked or threatened OR did you have something stolen from you
   a. at home including the porch or yard?
   b. at or near a friend's, relative's, or neighbor's home?
   c. at work or school?
   d. in places such as a storage shed or laundry room, a shopping mall, restaurant, bank, or airport?
   e. while riding in any vehicle?
   f. on the street or in a parking lot?
   g. at such places as a party, theater, gym, picnic area, bowling lanes, or while fishing or hunting?
   Did anyone ATTEMPT to attack or ATTEMPT to steal anything belonging to you from any of these places?
5. Other than any incidents already mentioned, has anyone attacked or threatened you
   a. with any weapon, for instance, a gun or knife?
   b. with anything like a baseball bat, frying pan, scissors, or a stick?
   c. by throwing something, such as a rock or bottle?
   d. by grabbing, punching, or choking?
   e. by raping, attempted rape or other type of sexual attack?
   f. by face-to-face threats?
   g. attacking, threatening, or using force by anyone at all?
   Please mention it even if you are not certain it was a crime.
6. People often do not think of incidents committed by someone they know. Did you have something stolen from you OR were you attacked or threatened by
   a. someone at work or school?
   b. a neighbor or friend?
   c. a relative or family member?
   d. any other person you have met or known?
7. Incidents involving forced or unwanted sexual acts are often difficult to talk about. Have you been forced or coerced to engage in unwanted sexual activity by
   a. someone you did not know before?
   b. a casual acquaintance?
   c. someone you know well?

If respondents reply affirmative to one of these latter questions, interviewers next ask *Do you mean forced or coerced sexual intercourse?* to determine whether the incident should be recorded as rape or as another type of sexual attack.

8. During the last 6 months, did you call the police to report something that happened to you which you thought was a crime?

order to avoid this problem, some researchers contend that respondents need to be asked more behaviorally specific questions to uncover rape victimizations that may have occurred. Beginning with a study by Koss, Gydycz and Wisniewski (1987) that examined the rape victimization experiences of college women, most recent surveys attempting to estimate the magnitude of rape have used very specific question wording. For example, the National Violence Against Women and Men Surveys (NVAWMS) asked the questions displayed in **Table 3**. The National Intimate Partner and Sexual Violence Survey (NISVS) conducted by the Centers for Disease Control used these same questions, and further cued respondents to think of unwanted sex that occurred when they were "unable to consent to it or stop it from happening because they were drunk, high, drugged, or passed out from alcohol, drugs, or medications." (Black et al., 2011).

Table 2 Screening questions used by the NYS to uncover self-reported offending

(Introduction) This section deals with our own behavior. I would like to remind you that all your answers are confidential. I will read a series of behaviors to you. Please give me your best estimate of the exact number of times[a] you have done each thing during the last year from Christmas a year ago to the Christmas just past.

1. Carried a hidden weapon other than a plain pocket knife
2. Attacked someone with the idea of seriously hurting or killing him or her
3. Been involved in gang fights
4. Tried to take something from someone with the use of force or with the threat of force

[a]Note: Response options were (1) once a month, (2) once every 2–3 weeks, (3) once a week, (4) two to three times a week, (5) once a day, and (6) two to three times a day. NYS, National Youth Survey.

Table 3 Questions used in the National Violence Against Women and Men Surveys to uncover incidents of rape and sexual assault

- (For female respondents) (1) Has a man or boy ever made you have sex by using force or threatening to harm you or someone close to you? Just so there is no mistake by sex we mean putting a penis in your vagina.
- (The remaining questions are for all respondents) (2) Has anyone, male or female, ever made you have oral sex by using force or threat of force? Just so there is no mistake, by oral sex we mean that a man or boy put his penis in your mouth or someone, male or female, penetrated your vagina or anus with their mouth.
- (3) Has anyone ever made you have anal sex by using force or threat of harm? Just so there is no mistake, by anal sex we mean that a man or boy put his penis in your anus.
- (4) Has anyone, male or female, ever put fingers or objects in your vagina or anus against your will or by using force or threats?
- (5) Has anyone, male or female, ever attempted to make you have vaginal, oral, or anal sex against your will, but intercourse or penetration did not occur?

Importantly, each of the acts described in Table 3 represent an element that constitutes rape in all states as well as federal crime statutes. As with the NCVS, both males and females can be either perpetrators or victims of rape. Asking about each of these acts separately cues respondents to think about all of the events that have happened to them that they may not have considered with the broad question used by the NCVS. In addition, there are other reasons why estimates differ. The context in which the survey is introduced to respondents, along with other factors, also affects how many rapes the survey uncovers. For example, the NVAWMS is introduced to respondents as a survey about issues of safety, unlike the NCVS, which is introduced to respondents as a survey interested in crime. These differences may also affect the estimates obtained. According to the NCVS, the rate of rape has not changed much over time, with approximately 250 000 people of 12 years of age and older becoming the victims of rape annually. The majority of rape victimizations reported to NCVS interviewers were never reported to police. In fact, fewer than 3 in 10 rape and sexual assault victimizations, according to the NCVS, are ever reported to police. The NVAWMS, on the other hand, estimates that over 900 000 men and women over 18 years of age become the victims of rape annually. This is far more than the number estimated by the NCVS. Clearly, the way in which you ask people about the victimization experiences affects the magnitude of estimates obtained by surveys.

Unfortunately, the only estimates of rape and sexual assault available at most local levels, including state and city levels, are from the less reliable police reports. It is very expensive to conduct a survey that is representative of any given population, and most local governments do not have the ability to fund such an endeavor. Unfortunately, we know victims are reluctant to report to police for many reasons, including the personal nature of the victimization, fear of retaliation from the offender, and the treatment they expect to receive from the media and the criminal justice system. Additionally, many victims often believe that they will not be believed by the police if they report the rape, and many also suffer a sense of shame about their victimization (Holmes and Holmes, 2002). It is a sad reality that rape victims still suffer from many stereotypes, misperceptions, and the social stigma and embarrassment associated with this crime.

Intimate Partner Violence

Similar to rape and sexual assault, determining the true magnitude of assaults between intimate partners such as spouses/exspouses, and boy/girlfriends, and exes, has been difficult. While this type of violence also includes verbal and controlling behaviors, this article will focus on measuring the physical nature of the crime. The Centers for Disease Control and Prevention (CDC) have provided a very useful definition when describing intimate partner violence (IPV):

> The intentional use of physical force with the potential for causing death, disability, injury, or harm. Physical violence includes, but is not limited to, scratching; pushing; shoving; throwing; grabbing; biting; choking; shaking; slapping; punching; burning; use of a weapon; and use of restraints or one's body, size, or strength against another person.
>
> CDC, 2013

Of course, we know IPV has been around forever, even though it has only been in the past 30 years that it has finally been acknowledged as a social problem. As recently as the 1970s and into the 1980s, police rarely made arrests when they were called to the scene of an intimate partner assault because it was thought to be a 'private matter' that was not the concern of the criminal justice system. The law, it seemed, did not extend past the front door. Widespread beliefs about the nature of the family and of gender roles contributed to this lack of attention. Fortunately, attitudes and policies have changed – largely as a result of the activism of various victims' and women's rights groups. Today, violence that goes on behind closed doors between intimates is usually treated like any other type of violence, with arrest being the norm rather than the exception. Because IPV is usually more private and hidden than some other forms of violence, its detection and intervention remain somewhat problematic, as does our ability to measure its perpetration.

Similar to the issues that plague measuring rape victimizations, surveys using diverse methodologies and different definitions of IPV have resulted in tremendously dissimilar estimates. For example, estimates of how many women experience violence by an intimate partner annually range from 9.3 per 1000 women to 116 per 1000 women. Furthermore, the differences across survey methodologies often preclude direct comparison across studies. There have been only four large nationally representative surveys that have estimated the annual rates of IPV: the National Family Violence Survey (NFVS), the NCVS, the NVAWMS, the National Intimate Partner and Sexual Violence Survey (NSVIS). Because of their procedural differences, each of these surveys has resulted in quite different estimates of IPV. While most researchers are familiar with the methodological differences between surveys, few policy makers take these into account when making generalizations about how often IPV takes place. This has unfortunately led to widespread confusion and controversy.

The NFVS was the first national survey devoted exclusively to estimating incidents of IPV. Its sample included only married or cohabiting couples and measured violence using an instrument called the Conflict Tactics Scale (CTS) (Straus and Gelles, 1990). The introduction to the CTS asks respondents to "think of situations in the past year when they had a disagreement or were angry with a specified family member (husband, child, etc.)" and to indicate how often they engaged in each of the acts included in the CTS. The list of acts covered in the CTS spans a range of behaviors, including reasoning, verbal aggression, and physical aggression or violence. Physical violence by the CTS index is often subdivided into two categories: minor violence and severe violence. These categories consist of the following acts: *Minor violence*: (1) threw something, (2) pushed, grabbed, or shoved, (3) slapped; *Severe violence*: (1) kicked, bit, or hit with a fist, (2) hit or tried to hit with something, (3) beat up, (4) choked, (5) threatened with a knife or gun, (6) used a knife or fired a gun.

Estimates from the 1985 NFVS indicate that about 11.6% of couples had experienced some type of assault in the previous 12 months. Extrapolated to the total married population, this survey estimates that almost one out of eight husbands carried out one or more violent acts during the year of this study. Further, the rate of severe violence perpetrated by husbands indicates that about 1.8 million women were beaten by their partner that year. The NFVS also estimates that rates of violence perpetrated by wives against husbands are very similar to rates of violence perpetrated by husbands against wives. Herein lies one of the most frequent criticisms of the CTS methodology: that it measures acts of violence in isolation from the circumstances under which the acts were committed. As critics point out, for example, it ignores who initiates the violence, the relative size and strength of the persons involved, and the nature of the participants' relationship (DeKeseredy and Schwartz, 2011). It should be noted here that a new version of the CTS called the CTS2 has been developed to address some of the methodological deficiencies of the original scale, but the new version still does not allow a researcher to account for the sequence of events that precipitated an act of violence. Thus acts of aggressive violence can still not be separated from those assaults that were used in self-defense.

As we noted earlier (**Table 1**), the NCVS asks a number of questions cueing respondents to think of many different kinds of attacks (e.g., with anything like a baseball bat, frying pan, scissors, or a stick), and different types of offenders (e.g., strangers, someone you know well), but it does not ask directly about attacks from intimate partners or other family members. Estimates of IPV from the NCVS indicate that in 2010, the rate of IPV against females was 5.9 per 1000 compared to a rate of only 1.1 for males. This translates into nearly 1 million females of 12 years of age and over being violently attacked by their intimate partners annually, which does not include homicides. Also notice that, unlike the NFVS, the NCVS finds that females have much higher rates of IPV victimization compared with males, which is consistent with homicide data.

The NVAWMS was a nationally representative survey that was introduced to respondents as a survey on personal safety (noted above), and measured assault by using a modified version of the CTS (Tjaden and Thoennes, 2000). Respondents included women and men of 18 years of age and older; they were asked about assaults, which had occurred as children and as adults using the following screening questions:

Not counting any incidents you have already mentioned, after you became an adult did any other adult, male or female ever ...

1. Throw something at you that could hurt?
2. Push, grab or shove you?
3. Pull your hair?
4. Slap or hit you?
5. Kick or bite you?
6. Choke or attempt to drown you?
7. Hit you with some object?
8. Beat you up?
9. Threaten you with a gun?
10. Threaten you with a knife or other weapon?
11. Use a gun on you?
12. Use a knife or other weapon on you?

The NVAWMS obtained higher estimates than the NCVS but lower estimates than the NFVS: 1.3% of women and 0.9% of men reported being assaulted by their intimate partners annually. Importantly, this survey also examined how many women and men experienced violent attacks over the course of their adult lives. Over one in five women (22%) and 7.4% of men had been assaulted by an intimate partner in their lifetimes.

The most recent survey conducted to measure IPV was conducted by the CDC in 2010 and is known as the NISVS. Like the NVAWMS, the NISVS not only asked respondents about victimizations that happened to them in the past 12 months but also about victimizations they had experienced in their lifetimes. And like the NVAWMS, the survey also measured a representative sample of the adult men and women of 18 years of age and older. In addition to questions about coercive control (e.g., tried to keep you from seeing or talking to your family or friends, kept track of you by demanding to know where you were and what you were doing), which are often behaviors that precede physical violence in relationships, the physical violence questions included those from above,

with the addition of "slammed you against something," "burned you on purpose," and "hurt you by pulling your hair." Another addition was that respondents were specifically asked about all intimate partners including same-sex couples. Both the NVAWMS and the NCVS collect data on same-sex couple violence, but do not directly ask respondents their sexual orientation. As such, it is only illuminate when the gender of both the victim and the perpetrator are examined in combination. Estimates from the NISVS reveal that about 4% of women and 4.7% of men had experienced a physical attack by an intimate partner in 2010. However, women were more likely than men to be victimized by severe acts of physical violence, which did not include slapping, pushing, or shoving (Black et al., 2011).

As you can clearly see, the ways in which people are asked about their victimization experiences significantly impact the number of people estimated to have experienced a victimization. Unfortunately, policy makers and researchers alike rely on published estimates of violent victimization from one or more of the surveys without recognizing or understanding the limitations of each. When you ask, "How many men and women are assaulted by their intimate partners each year?" the answer you get depends on which survey you rely on. Is it possible to compare incident rates across surveys? The simple answer is no. Because of the methodological differences between the different surveys, comparisons of their different estimates of IPV are not valid. How accurate can it really be to compare information from the NFVS, which interviewed only married and cohabiting couples, with estimates generated from the NCVS and NVAWMS that relied on national probability of samples of men and women in general, regardless of marital status? However, when the NCVS and the NVAWMS data are restricted to a similar context (i.e., adult victimizations involving lone offenders), research indicates that estimates of IPV still remain higher from the NVAWMS and the NISVS compared with the NCVS. This is undoubtedly due to the more specific screening questions used by the NVAWMS and also the specific questions about acts perpetrated by husbands, boyfriends, and former intimates (Bachman, 2000). In sum, when you are asked how many men and women experience IPV, your answer should be, "That depends on the methodology used to obtain the estimates." Regardless of which data set you rely on, however, all estimates that IPV is a significant problem. For all too many women, their partner poses more of a risk for serious harm and death than a stranger on the street.

Stalking

Before the measurement of stalking is discussed, the conceptualization of stalking should first be addressed. Does stalking involve unwanted repeated attempts at contact or only unwanted repeated attempts at contact that make someone feel fear? If we feel fear, how much fear is necessary for an act of stalking to be considered a crime? A 'great deal of fear' or just 'some fear?' The NISVS used the following definition of stalking:

> Stalking victimization involves a pattern of harassing or threatening tactics used by a perpetrator that is both unwanted and causes fear or safety concerns in the victim. … a person was considered a stalking victim if they experienced multiple stalking tactics or a single stalking tactics multiple times by the same perpetrator and felt very fearful, or believed that they or someone close to them would be harmed or killed as a result of the perpetrator's behavior.
>
> Black et al., 2011

The tactics measured included unwanted phone calls, text messages or hang-ups, unwanted e-mails or messages through social media, unwanted mail or presents, unwanted following from a distance or through the use of devices, such as GPS systems, showing up at the victim's home or workplace, leaving items for victim to find, or doing things that indicate the stalker had been in the victim's home or car.

In 2010, the NISVS found that about 16% of women had been stalked at some point in their lifetimes and 4% had been stalked in the 12 months prior to the survey. The incidents for men was much lower, with about 5% of men experiencing stalking in their lifetimes and just over 1% experiencing it in the past 12 months. For lifetime reports of stalking, the majority of female victims (66%) were stalked by intimate partners, while the male victims were equally likely to be stalked by intimate partners and friend/acquaintances (41 and 40% respectively). The NCVS also conducted a special Supplemental Victimization Survey in 2006 that measured stalking. It found that 2.2% of females of 18 years of age and older compared to 0.8% of their male counterparts had experienced stalking in the past 12 months. Together, it was estimated that over 3 300 000 stalking victimizations occurred annually. These victimization data conform to perpetration statistics, since research indicates that the vast majority of perpetrators are men (Tjaden and Thoennes, 2002).

Conclusion

This article has illustrated that measuring the true magnitude of criminal victimization is not as easy as it sounds. Relying on police report data compiled by the UCR or the NIBRS by the FBI only a tip of the ice burg of victimization. Random sample surveys of the population more accurately measure the victimizations that remain hidden from police detection because of nonreporting, however, even they have measurement error. When a newspaper or news program reports that crime is going up or down in the nation or in various locations, it is important to ask where these numbers came from and how victimization was measured. Of course, this question should also be asked whenever we are provided 'facts' about criminal victimization.

Bibliography

Alvarez, A., Bachman, R., 2008. Violence: The Enduring Problem. Sage, Thousand Oaks, CA.

Bachman, R., 2000. A comparison of annual incidence rates and contextual characteristics of intimate-partner violence against women from the national crime victimization survey (NCVS) and the national violence against women survey (NVAWS). Violence Against Women 6 (8), 839–867.

Black, M.C., et al., 2011. The National Intimate Partner and Sexual Violence Survey (NISVS): 2010 Summary Report. National Center for Injury Prevention and Control, Centers for Disease Control and Prevention, Atlanta, GA.

Bureau of Justice Statistics, 2013. http://bjs.gov/index.cfm?ty=dcdetail&iid=245#Questionnaires.

Centers for Disease Control and Prevention, 2013. www.cdc.gov/ncipc/factsheets/ipvoverview.htm.

DeKeseredy, W., Schwartz, M., 2011. Measuring the Extent of Woman Abuse in Intimate Heterosexual Relationships: A Critique of the Conflict Tactics Scales. U.S. Department of Justice. Sponsored VAW.net at: www.vaw.umn.edu/Vawnet/ctscritique.htm.

Holmes, R.M., Holmes, S.T., 2002. Profiling Violent Crimes: An Investigative Tool. Sage, Thousand Oaks, CA.

Koss, M.P., Gydycz, C.A., Wisniewski, N., 1987. The scope of rape: incidence and prevalence of sexual aggression and victimization in a national sample of higher education students. Journal of consulting and Clinical Psychology 55, 162–170.

Mosher, C.J., Miethe, T.D., Phillips, D.M., 2002. The Mismeasure of Crime. Sage, Thousand Oaks, CA.

Short, J., Nye, I., 1958. Extent of unrecorded juvenile delinquency: tentative conclusions. Journal of Criminal Law, Criminology, and Police Science 49, 296–302.

Straus, M., Gelles, R., 1990. Physical Violence in American Families: Risk Factors and Adaptations to Violence in 8,145 Families. Transaction Press, New Brunswick, NJ.

Tjaden, P., Thoennes, N., 2000. Prevalence, Incidence, and Consequences of Violence against Women. National Institute of Justice, U.S. Department of Justice, Washington, DC.

Tjaden, P., Thoennes, N., 2002. Stalking in America. National Institute of Justice, U.S. Department of Justice, Washington, DC.

Relevant Websites

http://bjs.gov/ – Bureau of Justice Statistics.

http://www.fbi.gov/about-us/cjis/ucr/ucr – Federal Bureau of Investigation, Uniform Crime Reports.

http://www.fbi.gov/about-us/cjis/ucr/nibrs/addendum-for-submitting-cargo-theft-data/nibrsconversion – Federal Bureau of Investigation, NIBRS Data Summary.

http://www.cdc.gov/ – Centers for Disease Control and Prevention.

Data Bases and Statistical Systems: Culture

Gunnar Otte and David Binder, Institute of Sociology, Johannes Gutenberg University Mainz, Mainz, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The measurement of culture in statistical systems is a rather new endeavor. In international comparison, considerable differences exist in classifications and indicators. This article provides an overview of definitional and classificatory approaches as well as transnational attempts of harmonizing cultural statistics. In particular, statistical systems and databases are introduced with regard to the cultural industries, artists, and cultural producers, organizations responsible for the dissemination of culture, and consumers and recipients of culture.

Introduction

'Culture' is one of the most ambiguous terms in the social sciences. Currently, there are at least six scientific usages of the word:

- Culture as everything created by humankind (as opposed to 'nature').
- The anthropological notion of culture as the social organization of whole societies.
- Cultures as historical lineages of world civilizations (e.g., 'Western culture').
- Culture as the value system of a society, group, or individual (e.g., 'postmaterialism').
- Culture as the lifestyle of a society, group, or individual, i.e., a set of typical everyday life practices (e.g., 'hedonistic' or 'sustainable lifestyles').
- Culture in terms of arts.

It is beyond the scope of this article to discuss statistical systems for all these notions of culture – if they exist at all. In particular, we will not address conceptions of measuring religious beliefs, cultural values, language systems, ethnicity, and lifestyles. Rather, we focus on frameworks originating from the arts in a wider sense.

After devoting some more words on definitional issues in Section Defining Culture, we introduce the most important classifications of culture that are used in statistical systems (Section Classifying Culture). Besides, two transnational conceptions of cultural statistics – as proposed by the UNESCO and the European Union (EU) – are presented (Section Cultural Statistics in Transnational Perspective). In Section Cultural Industries, we take up the much debated question of how to delineate the cultural and creative industries. We then follow a process model of cultural products, ranging from their creation and production via their dissemination to their consumption and reception. Specifically, we discuss the problem of the inclusion coverage of artists (Section Artists and Producers of Culture), present research on the organization-based dissemination of culture (Section Dissemination of Culture), and summarize the methodology of cultural participation surveys (Section Consumption and Reception of Culture).

Defining Culture

There is little agreement on definitional issues in statistical systems of culture. What belongs to the 'cultural domain' is often decided ad hoc and on the basis of available databases. Some recent statistical reports refer to the UNESCO universal declaration on cultural diversity, which defines culture as the "set of distinctive spiritual, material, intellectual and emotional features of society or a social group" that "encompasses, in addition to art and literature, lifestyles, ways of living together, value systems, traditions and beliefs" (UNESCO, 2001). Taking up value systems and lifestyles, this definition is considerably broader than arts-centered approaches. When it comes to operationalizing 'culture' in statistical systems, however, the domains and indicators selected refer rather loosely to this definition; 'pragmatic' considerations prevail (UNESCO-UIS, 2009: 22f.). While many dimensions of values and lifestyles are disregarded by official statistics, the arts are either part of most frameworks or even constitute their main focus.

It therefore seems appropriate to engage in definitions of 'the arts.' To begin with, the arts are the social system that is centered around 'works of art.' They can be considered 'products,' comprising (tangible) goods and (intangible) services. Such products vary tremendously in form and content and therefore works of art are hard to determine according to innate characteristics. However, they can be defined by reference to the processes of their production or reception. Reception-based definitions consider a product as a 'work of art' if actors in the art system ascribe this label to it, drawing on quality or beauty criteria. The problem is that value judgments vary historically, spatially, and interindividually among actors and that other products – such as cars – can be judged by aesthetic criteria as well. Production-based definitions avoid these difficulties. They regard products as 'art' if their production is guided primarily by the objectives of aesthetic expression. Although important aesthetic components can be found in automobile production, functional–technical imperatives like roadworthiness, safety, and energy consumption dominate the development of cars. Therefore, they do not constitute artworks. However, children's drawings are primarily aesthetic expressions as well. To exclude such cases, the definition should be restricted to professional producers or artists. An 'artist' is somebody who sends

professional behavioral signals toward the art system, for example, by taking part in contests, becoming a member of professional organizations, or offering his or her artistic products for money (Throsby, 2001). In summary, works of art are all products of primarily aesthetic expression that are produced by artists in their professional practice.

This definition does not restrict artworks and artists to those accepted by art critics, but includes products and producers devalued or ignored by them. It implies treating 'high culture' as well as 'popular,' 'mass,' and 'folk culture' as objects of investigation. Although our definition is rather wide, there are boundary zones of products whose status as art or nonart is unclear, especially in the applied arts (e.g., crafts, architecture, design, advertisement) and in popular entertainment (e.g., vaudeville, circus). Models of concentric circles or core–periphery gradings provide a solution: They treat genuinely artistic products and occupations as the center, or core, of the art system and locate ambiguous products and partially artistic occupations in the fringe, or periphery, of a wider creative sphere (KEA, 2006: 53ff.; Throsby, 2008a).

In the debate on the economic impact of culture, the 'cultural industries' are usually considered a subset of the 'creative industries.' While the former contain artistic core industries like music, film, the visual and performing arts, publishing, and broadcasting, the latter may encompass the advertisement or software industries. For instance, the UK *Creative Industries Mapping Document* defines the 'creative industries' as "those industries which have their origin in individual creativity, skill and talent and which have a potential for wealth and job creation through the generation and exploitation of intellectual property" (DCMS, 2001: 5). Again, the notion of 'creativity' is used rather heterogeneously in this debate. It deviates from the use in psychology, where creativity is understood as "the ability to produce work that is both novel (i.e., original, unexpected) and appropriate (i.e., useful, adaptive concerning task constraints)" (Sternberg and Lubart, 1999: 3). In accordance with this, creativity is likely to be found from the arts and sciences to craftwork and manufacturing and cannot be confined to a specific industry.

Classifying Culture

Process Models of Culture

Definitional difficulties notwithstanding useful models and classifications of culture have been developed. In the sociology of the arts, a process model of cultural products is widely used (Alexander, 2003). In **Figure 1**, a simple three-stage model is presented. It starts with the assumption that cultural producers develop creative ideas and transform them into cultural products. In modern societies, these products are rarely brought to consumers directly, but are evaluated and disseminated by intermediate actors. They are performed or exhibited in concert halls and galleries; distributed by publishing houses and labels; traded via wholesale, retail, and Internet platforms; popularized on television and radio, in magazines and newspapers; filtered and reviewed by critics; archived and preserved; and sometimes canonized in educational curricula and cultural histories. Because of this complex preselection process, cultural products have unequal chances of reaching consumers and being successful. On the demand side, the model distinguishes the processes of 'consumption' in terms of quantitative demand and 'reception' in terms of subjective acquisition. It is worth noting that consumers, or recipients, are to a significant degree amateur artists themselves, when playing instruments, making pottery, painting, and so on. Sometimes process models are presented as a 'cultural cycle' indicating that the creation and production of particular types of products are stimulated by market demand (UNESCO-UIS, 2009: 19f.). We allow for this proposition by the feedback loop in the figure. From an economic perspective, process models of cultural content are often interpretated as production chains with several value-adding stages.

Another model, the three-sector model, also starts with the creative professional, but positions him or her relative to three cultural subsectors that serve as organizational and financial bases of cultural production: the private sector (economy), the public sector (state), and the intermediate sector of nonprofit organizations (civil society). This differentiation is used in some mapping studies of the creative industries, for example, in Switzerland (Weckerle et al., 2008). There is a tendency to confine the creative industries to the private, for-profit sector.

Models of Cultural Industries

A long-standing concern is the classification of the cultural industries. Based on the assumption that cultural products have an economic and a cultural value, Throsby (2008a) proposes a concentric circles model with four layers of cultural industries. The cultural content of their products decreases, relative to their economic value, from the inner to the outer circles. In the center are the 'core creative arts' (literature, music, performing arts, visual arts), followed by 'other core cultural industries' (film, museums/galleries/libraries, photography), the 'wider cultural industries' (heritage services, publishing and print media, sound recording, television and radio, video and

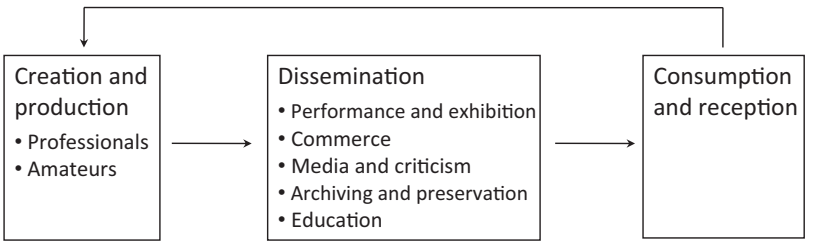

Figure 1 Process model of cultural products.

computer games) and, in the outer circle, 'related industries' (advertising, architecture, design, fashion) (Figure 2). Importantly, Throsby validates his model by classifying single occupations as creative or noncreative and by calculating, for the four layers, the proportion of total industry employment in creative personnel. Using a similar model, the consultancy KEA European Affairs (2006: 56) comes up with a partly different proposal. Because the 'core arts field' is defined by nonindustrial activities whose output are prototypes, it is restricted to the visual and performing arts (including crafts) as well as cultural heritage. Music and literature, together with film, video, television, radio, video games, and publishing, are assigned to the next layer, the 'cultural industries.' They are assumed to be organized around activities aimed at massive reproduction. Obviously, definitional intricacies and decisions on aggregation levels – Throsby separates musical and literary creation from the industries reproducing these works – easily amount to different classification schemas.

Cultural Divisions by Genre

Not only should cultural content be sorted according to product groups and industries, such as 'music,' 'film,' and so forth. It can also be grouped across such divisions. In a major work in Germany, Schulze (1992) identified three 'aesthetic everyday life schemes' that are central for structuring individual tastes. Every consumer exhibits a specific mixture of proximity and distance toward these schemes. The 'high culture scheme' comprises refined products suited for 'contemplation' and an 'antibarbarian' distinction (e.g., classical music, serious literature, museums); the 'trivial culture scheme' includes folk culture and light entertainment emanating 'harmony' and enabling an 'anti-eccentric' distinction (e.g., easy listening, quiz shows, light novels); the 'action culture scheme' includes pop and youth cultural products fit for 'action' and 'anticonventional' distinction (e.g., rock music, science fiction, bars, and clubs). These schemes have shown to be quite stable over time and correlate strongly with age and education. In order to map the schemes, cultural objects have to be classified into artistic genres or stylistic ages. This is rather easy in the domains of music and film, but more difficult in the visual arts, literature, and theater. Moreover, genre classifications have to be adapted to regional cultural traditions.

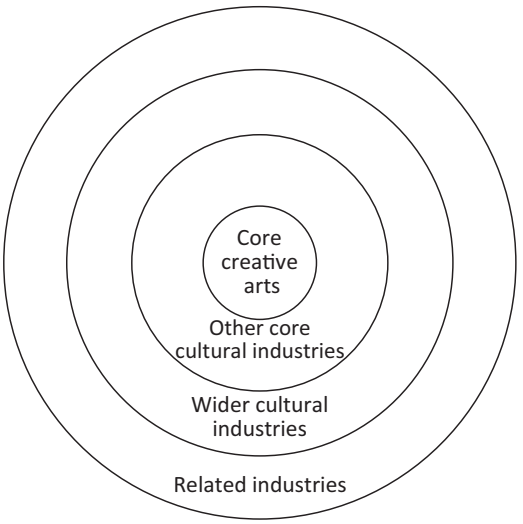

Figure 2 Concentric circles model of cultural industries. Model adapted from Throsby, D., 2008a. The concentric circles model of the cultural industries. Cultural Trends 17, 150.

Cultural Statistics in Transnational Perspective

UNESCO Framework for Cultural Statistics

On the transnational level, two efforts of developing a coherent framework for cultural statistics are particularly worth noting: those of the UNESCO and the EU. Both are meant as tools for organizing cultural statistics in an internationally comparable manner.

After the first initiatives in 1972, it took until 1986 for the first UNESCO *Framework for Cultural Statistics* (FCS) to be enacted (UNESCO, 1986). The central instrument is a set of matrices crossing 10 cultural 'categories' and several subcategories, i.e., product and activity groups in the rows with five cultural 'functions' (taken from a process model like in Figure 1) in the columns. For each cell, quantitative indicators are specified. The framework goes far beyond the arts, as defined above, by taking up cultural heritage, media systems, sports and games, aspects of family, community and organizational life as well as quality of life in the natural and urban environment.

In 2009, the FCS was updated considerably (UNESCO-UIS, 2009). The totality of what is delineated as 'culture' has been enlarged (with design, architecture, advertising, Internet services, and tourism) and the former 'categories' have been regrouped into six 'cultural domains':

- Cultural and natural heritage (museums; archeological and historical places; cultural landscapes; natural heritage)
- Performance and celebration (performing arts; music; festivals, fairs, and feasts)
- Visual arts and crafts (fine arts; photography; crafts)
- Books and press (books; newspaper and magazines; other printed matter; library; book fairs)
- Audiovisual and interactive media (film and video; television and radio; Internet podcasting; video games)
- Design and creative services (fashion, graphic, interior and landscape design; architectural services; advertising services).

Outside of these genuinely cultural domains, two 'related domains' are added:

- Tourism (charter travel and tourist services, hospitality and accommodation)
- Sports and recreation (sports, physical fitness and well-being, amusement and theme parks, gambling).

Linked to each of the above domains are four 'transversal domains,' the first one being considered entirely cultural, the last three partially cultural:

- Intangible cultural heritage (oral traditions and expressions, rituals, languages, social practices)

- Education and training (with regard to cultural values and skills)
- Archiving and preservation
- Equipment and supporting materials.

It is worthy of mention that the current FCS takes up the matrix approach cross-tabulating cultural domains and a five-stage process model – creation, production, dissemination/transmission, exhibition/reception, consumption/participation – but abstains from specifying indicators.

EU Statistical System of Culture

In the EU, an important step toward a statistical system of culture was made by the *Leadership Group on Cultural Statistics* from 1997 to 2000. Its work was taken up by the Statistical Office of the EU (Eurostat, 2011) and the *European Statistical System Network on Culture*, collaborating from 2009 to 2011 (ESSnet-Culture, 2012). For ESSnet-Culture, the conceptual starting point is the notion of 'cultural activities,' i.e., the processing of goods and services whose 'intrinsic or symbolic value' surmounts their 'technical value.' Professionals and amateurs exercise these activities – for-profit or not-for-profit – with regard to specific cultural domains and cultural functions. A 'cultural domain' is defined as "a set of practices, activities or cultural products centred around a group of expressions recognized as artistic ones" (ESSnet-Culture, 2012: 55). Ten equally ranking domains are distinguished:

- Heritage (museums, historical places, archaeological sites, intangible heritage)
- Archives
- Libraries
- Book and press
- Visual arts (plastic and fine arts, photography, design)
- Performing arts (music, dance, drama, circus, cabaret, combined arts, and other live shows)
- Audiovisual and multimedia (film, radio, television, video, sound recordings, multimedia works, video games)
- Architecture
- Advertising
- Art crafts.

Because 'artistic creation' is at the core of the framework, architecture, advertising, and art crafts are included only in their cultural components (e.g., architectural design, but not construction). Explicitly excluded are general system and applications software; information and telecommunication; leisure, entertainment, and tourism; natural reserves, zoos, and botanical gardens; and manufacturing of ornamental products.

Cutting across the domains, six sequenced 'functions' are distinguished: creation (original work), production and publishing, dissemination and trade, preservation, education, and management and regulation. Instead of using a full process model that includes the demand side, ESSnet-Culture builds a third axis into its framework by proposing four 'dimensions' that constitute important research foci. Each of the domain-by-function intersections, i.e., cultural activities or products, is viewed with regard to (1) the employment it entails, (2) the public financing it presupposes, (3) the private consumption it stimulates, and (4) the social participation it generates.

As the 2009 FCS is taken into account by ESSnet-Culture, important similarities appear: The cultural domains overlap to a large extent and the matrix crossing domains and functions is the central organizing tool for sorting statistical indicators. There are also deviations, however. The boundaries of the cultural sector are more restricted in the EU approach: Natural heritage, crafts in general, software industries, tourism, sports, and recreation are not included. The FCS 'transversal domains' are either ranked under 'heritage' (intangible cultural heritage), treated as a cultural domain (archiving), shifted to a function (education and training) or considered a 'related activity' outside the cultural sector (equipment and support). In addition, four 'dimensions' are specified as main research objectives. It should be noted that both frameworks are grounded less in theoretical terms than the concentric circles models by Throsby (2008a) and KEA European Affairs (2006).

Cultural Industries

Since national and local governments strongly discovered the economic impact of the cultural, or creative, industries around the turn of the millennium, fueled by the UK *Creative Industries Mapping Document* (DCMS, 2001) and Richard Florida's (2002) widely recognized book on the 'creative class,' reports on the cultural industries flourish on a worldwide scale (for an overview, see UNESCO-UIS, 2012). The main impetus is to quantify the cultural industries, in comparison with other industries, in terms of structure and size, turnover, gross value added, and employment and to assess their direct, indirect, and induced effects for value creation and regional economic development. A major problem is the lack of comparability of these studies. There are several reasons to this: lacking theoretical foundations, imprecise definitions, deviating classifications, country-specific conventions in the statistical aggregation of industries and occupations, and diverging methodologies and data sources for measuring economic impact. The classificatory models as well as the UNESCO and EU frameworks outlined above might help to achieve greater harmonization in the future (for a comparison of models, see Throsby, 2008b).

A comparison of national practices shows that nearly all European countries number the following domains among their classifications of the cultural industries (UNESCO-UIS, 2012: 51f.):

- Music
- Performing arts
- Visual arts
- Film and video
- Publishing (books, press, journals)
- Broadcasting (radio, television)
- Architecture
- Design.

A majority also includes cultural heritage, advertising, software, and computer games. Other domains like sports, tourism, botanical gardens/zoos, toys, wine/food, cultural education, and artistic training are rarely counted. Countries in the Asia-Pacific region seem to be more heterogenous in their

delineation of the cultural industries, while in most African countries cultural statistics are only beginning to emerge (UNESCO-UIS, 2012: 69ff.).

When mapping the cultural industries, industry and occupational classifications should be considered simultaneously. With their 'creative trident' model, Higgs and Cunningham (2008) argue that within the 'creative industries' there are not only 'creative specialists,' but also sizable numbers of 'business and support workers' such as secretaries. Conversely, members of creative occupations may very well work 'embedded' within other industries, such as designers in the automobile industry. Based on their calculations with the 2001 Australian census, each of the three groups accounts for roughly one-third of the total employment in the 'creative trident.' In **Figure 3**, we adapt this model to the 'cultural industries' and tie it to a center–periphery conception: The 'cultural core' of the workforce consists of people working in cultural occupations within the cultural industries; cultural specialists who are embedded in other industries constitute the 'cultural fringe'; and support personnel within the cultural industries may be termed the 'noncultural fringe.' Both employees and the self-employed are relevant here.

It is beyond the scope of this article to go into details with measurement issues. In fact, procedures depend heavily on the statistical nomenclatures and data sources being in use in the country under consideration. For example, in order to measure cultural employment of the three types in **Figure 3** among EU countries in a comparable way, ESSnet-Culture (2012: 139ff.) proposes to use the EU Labour Force Survey and to combine information on employer's main activity according to NACE ('Nomenclature générale des activités économiques dans les Communautés Européennes') and employee's occupation according to 'International Standard Classification of Occupations.' At world level, NACE corresponds to the 'International Standard Industrial Classification of all Economic Activities.'

Reports on the economic impact of culture rarely separate single components of the creation/production, dissemination, and consumption/reception process of cultural products. As interesting questions arise in each of these stages, we devote single sections to each of them in the remainder of the article.

Artists and Producers of Culture

Various 'status-of-the-artist' studies exist that try to describe and explain the social and economic situation of artists (Karttunen, 1998; for an overview, see Butler, 2000). To study artists and producers of culture, two approaches are feasible (Throsby, 2001). The first is the use of census data or social surveys containing very large sample sizes such as microcensuses that cover the population or labor force in general. Although artists are in principle adequately represented in these data, it can be difficult to identify them all. Census data is being criticized for assigning a single occupation to people based on the job they last performed or from which they derived the main portion of their income. Thus, multiple jobholding and temporary underemployment which are typical for artists get ignored. (Benhamou, 2011; Karttunen, 2001). Moreover, census data are limited to very basic information; important variables such as artists' working hours and income from artistic and nonartistic work are usually missing (Alper and Wassall, 2006). The alternative is to carry out special-purpose surveys on artists. They enable researchers to define who counts as an artist, to use a suitable sampling frame, and to collect information on artists' social background, education, working conditions, earnings, social security, time allocation, and career.

However, defining an artist is challenging. Frey and Pommerehne (1989) identify eight ways of determining artists:

- The amount of time spent on artistic work
- The amount of income derived from artistic activities
- Their reputation as an artist among the general public
- Recognition among other artists
- The quality of artistic work
- Membership in a professional artists' association
- Professional qualification in the arts
- Their subjective self-evaluation of being an artist.

The use of multiple criteria is recommended because only subpopulations are reached by each of them. A major example is the *Economic Study of Professional Artists in Australia* carried out five times since 1983 (Throsby and Zednik, 2010). Specific to these surveys is the definition of artists as professionally recognized in their respective art forms. To arrive at a list of potential artists, artists' unions and associations, published listings, and directories are consulted. From the consolidated list of names, a random sample is drawn for a screening interview. To be identified as 'professional artists,' respondents have to answer a number of screening questions measuring artistic success and involvement (like publishing, exhibitions, full-time training). Irrespective of their income, both full-time and part-time, employed and self-employed artists are incorporated.

Further sampling strategies and research designs do exist. For example, respondent-driven sampling has been suggested to identify 'hidden' artist populations, like jazz musicians (Heckathorn and Jeffri, 2001). Although panel studies of artists are rare, the career success of professionally trained artists can

	Employment within cultural industries	Employment within other industries
Employment in cultural occupations	1. Cultural core ('specialist')	2. Cultural fringe ('embedded')
Employment in other occupations	3. Noncultural fringe ('business and support')	4. Other

Figure 3 Model of the cultural workforce by industry and occupation. Model adapted from Higgs, P., Cunningham, S., 2008. Creative industries mapping. Where have we come from and where are we going? Creative Industries Journal 1, 17.

be traced on the basis of lists of art school alumni, as studies on visual artists in the Netherlands show (Rengers, 2002). In Finland, the register-based census enables researchers to track artists over several census waves by means of a personal code (Karttunen, 2001).

Dissemination of Culture

The dissemination of culture, in large part, takes place on an organizational basis. Hence, organizational and archival data are needed in order to study structures and processes in this part of the cultural process model. However, statistical systems are not well established in this field (for the United States, see Kaple, 2002; for Germany, see Rössel and Otte, 2010).

With regard to nonprofit organizations, arts service organizations – like museum, theater, or orchestra associations – are important bodies that collect data of their members, for example, on revenues, expenditures, and staff structures. These data are usually published in annual reports in aggregated form and rarely allow for differentiated analyses at the organizational level. Moreover, the coverage of these databases tends to be biased in favor of large organizations.

More often than not, rich data sources have to be developed and utilized by researchers in time-consuming collection and coding procedures. For example, archival data of orchestra repertoires can be used to study conformity and innovation in the musical canon (Dowd et al., 2002). Contents of newspapers and school exams can be analyzed to get information about the valuation of different cultural traditions and art forms in education and the public sphere (Bevers, 2005; Janssen et al., 2008). Business data are harder to obtain. Knowing about the artistic offerings and sales figures of publishing houses, labels, and galleries would be helpful to reconstruct the supply side of the market and to explain artistic success. Of course, data can be gathered by survey research. For instance, samples of museum directors and curators as well as art dealers have been surveyed with standardized questionnaires. Since 2001, an archive for data on the arts and cultural policy exists, the *Cultural Policy and the Arts National Data Archive* at Princeton University, offering study descriptions and data sets from the United States.

Consumption and Reception of Culture

Cultural Participation Studies

Research on cultural consumption, reception, and amateur production is conducted in 'cultural participation studies.' It relies on sample surveys of the population or of audiences of concrete cultural institutions. We focus on population surveys because of their potential for representative predications. They are either conducted by academic social research or, more often, commissioned by governmental institutions and official statistics. In the latter case, a main goal is to gain insight into public participation in the cultural infrastructure subsidized by the state. Therefore, cultural domains tend to be constricted to 'high' cultural institutions and cultural heritage.

Quite often, the target population of cultural participation surveys is the population aged 15 years and over living in private households of a country. The treatment of ethnic minorities varies: Sometimes people with limited abilities in the official language or with a foreign citizenship are excluded; sometimes ethnic minorities are oversampled. In Europe, typically multistage, random samples are drawn; data collection is mostly done through face-to-face interviews. The median sample size is about 4500 (ESSnet-Culture, 2012: 243). For the United States, an inventory of 29 studies is provided by Pettit (2000). Morrone (2006) provides insight into the measurement of cultural participation in Bhutan, Thailand, and Uganda.

Questionnaires usually consist of at least four modules:

1. Frequency of attending cultural events and institutions
2. Frequency of media-based cultural consumption at home
3. Frequency of performing artistic amateur activities
4. Sociodemographic characteristics of the respondents.

The set of core items is depicted in **Table 1** (for overviews, see ESSnet-Culture, 2012: 248ff.; Morrone, 2006; Pettit, 2000). Additionally, preferences are often surveyed with respect to

Table 1 Core elements of cultural participation surveys

Domain	*Cultural attendance*	*Home consumption*	*Amateur production*
Music	Live music concerts (by genres)	Listening to radio Listening to recorded music (and genre preferences)	Playing an instrument Singing
Performing arts	Theater Musical theater Opera Ballet/dance performances		Acting Dancing
Literature	Libraries	Reading books (and genre preferences) Reading newspapers/magazines (and arts-related interest)	Writing stories/poetry/novels
Visual arts	Museums (by types) Art galleries		Painting/drawing Producing craftworks Amateur photography
Film	Cinema	Watching TV, watching films on video/DVD (and genre preferences)	
Cultural heritage	Historical monuments Archaeological sites		

music, film, and literature genres as well as museum types. In extended surveys, modules for the socialization in the arts, possession of cultural goods, media use in a wider sense, the social context of attendance, cultural reception modes as well as subjective motives and barriers of attendance are added. In general, the breadth of cultural domains is considerably smaller than the UNESCO-UIS, (2009) FCS recommendations. For example, preferences for architectural or design styles are rarely investigated. It is striking that, despite of the common core of indicators, there is little agreement on the wording of questions, response categories, and other measurement issues (Schuster, 2007). One standard followed by most researchers is to use the reference period of the last 12 months for cultural attendance.

Table 2 lists a selection of large-scale cultural participation survey programs that are organized as repeated cross-sections. The U.S. *Survey of Public Participation in the Arts* is probably the most widely used data set. Although it has been conducted as a supplement to larger surveys and is lacking modules that go beyond cultural consumption, amateur production, and some basic cultural socialization, its sample size and regular repetition with similar questions make it a valuable tool. The Dutch *Amenities and Services Utilisation Survey* and the French *Pratiques Culturelles des Français* reach back in time even longer. A major series of yearly studies was started in England in 2005 with the *Taking Part* program on leisure, culture, and sport. Since 2012, part of it is organized as a panel study. In Sweden, national studies are carried out yearly since 1986. Although differing in focus, cultural attendance is investigated almost every time – resulting in a unique time series. *Cultural Participation in Flanders* is a very detailed and theory-driven study that comprises – beyond all the modules discussed above – cultural attitudes, social network information, and retrospective questions on participation in adolescence. The *Family Survey Dutch Population*, with its distinct life course perspective, is a good example of how to measure the accumulation of cultural knowledge and interests. In some countries, cultural participation modules are repeatedly part of general social surveys (Australia, Canada) or leisure surveys (Finland).

For international comparisons in Europe, the Eurobarometer is a good choice because it is a rare instance of a harmonized survey program. The 2001 pioneer study among 15 EU member states was replicated for the candidate countries in Central and Eastern Europe in 2003. A third study took place in 2007. The culture module was developed by the *Leadership Group on Cultural Statistics*. It contains information on cultural attendance, amateur activities, media use, possession of relevant equipment, and music preferences. Probably due to a self-selection bias of culturally interested respondents, the level of cultural participation is overestimated when compared to national reference surveys (Morrone, 2006). Even in this harmonized survey, comparisons are problematic because of different national meanings of terms like 'concert' and 'opera' (Madden, 2005). Methodological guidelines and a questionnaire for a harmonized survey on cultural participation are formulated by ESSnet-Culture (2012: 256ff.).

Time-Use and Household Expenditure Surveys

Some aspects of cultural consumption can be studied with time-use and household expenditure surveys. In time-use surveys, respondents record, for a few days, their hourly main and secondary activities in a detailed time diary. While such data are instructive for frequent, mostly home-based cultural activities, less-frequent practices cannot be studied adequately. The *Multinational Time Use Study* is an expost harmonization of national data, comprising 60 data sets from 25 countries in 2013. The *Harmonised European Time Use Surveys* program encloses data from 15 national studies and offers a guideline for harmonized data collection (Eurostat, 2009). An exhaustive list of cultural practices is provided by the ICATUS codes (UNESCO-UIS, 2009). In household expenditure surveys, household members' amount of spending on culture is usually included, but not broken down to cultural genres. The *Multinational Household Expenditure Study* and the European *Household Budget Survey* allow for international comparisons.

Conclusion

Although a common core of models and methods is detectable, the current state of cultural statistics is plagued by a multitude of proposals of how to define and classify culture, delineate the cultural industries, capture artists statistically, and measure cultural participation. The UNESCO and EU frameworks will certainly improve harmonized data collection in the future,

Table 2 Major replicative survey programs on cultural participation

Country	*Survey*	*Years*
Australia	Attendance of cultural venues and events (part of Multipurpose Household Survey (MPHS))	1991, 1995, 1999, 2005/06, 2010/11
Belgium	Cultural Participation in Flanders	2003/04, 2009
Canada	Supplement of the General Social Survey (GSS)	1992, 1998, 2005, 2010
England	Taking Part	Since 2005 yearly
Finland	Leisure Survey	1977, 1981, 1991, 2002
France	Pratiques Culturelles des Français	1973, 1981, 1988, 1997, 2008
The Netherlands	Amenities and Services Utilisation Survey (AVO)	Since 1979 every fourth year
	Family Survey Dutch Population	1992, 1998, 2000, 2003, 2009
Sweden	Kulturvanor i Sverige	Since 1986 yearly
The United States	Survey of Public Participation in the Arts (SPPA)	1982, 1985, 1992, 1997, 2002, 2008, 2012
European Union	Eurobarometer	2001/03, 2007

even though the inertia of national statistical systems should not be underestimated. Faced with diverging conceptual approaches, researchers should bear in mind that the utility and validity of concepts can ultimately be evaluated only if these concepts are linked to clearly stated theories, applied to practical research questions, and tested by their empirical performance.

See also: Art and Culture, Economics of; Art and Socialisation; Artists, Competition and Markets; Census Microdata; Classifications in Popular Music; Consumption, Sociology of; Cultural Participation, Trends in; Culture and Economy; Culture, Sociology of; Data Bases and Statistical Systems: Overview; Markets: Artistic and Cultural; Quantitative Cross-national Research Methods; Social Inequality in Cultural Consumption Patterns.

Bibliography

Alexander, V.D., 2003. Sociology of the Arts. Exploring Fine and Popular Forms. Blackwell, Malden.

Alper, N.O., Wassall, G.H., 2006. Artists' careers and their labor markets. In: Ginsburgh, V.A., Throsby, D. (Eds.), Handbook of the Economics of Art and Culture. Elsevier, Amsterdam, pp. 813–864.

Benhamou, F., 2011. Artists' labour markets. In: Towse, R. (Ed.), A Handbook of Cultural Economics. Edward Elgar Publishing, Cheltenham, pp. 53–58.

Bevers, T., 2005. Cultural education and the canon. A comparative analysis of the content of secondary school exams for music and art in England, France, Germany and the Netherlands, 1990–2004. Poetics 33, 388–416.

Butler, D., 2000. Studies of Artists. An Annotated Directory. Working Paper #12. Princeton University, Center for Arts and Cultural Policy Studies, Princeton.

Department of Culture, Media and Sport (DCMS), 2001. Creative Industries Mapping Document 2001. DCMS, London.

Dowd, T.J., Liddle, K., Lupo, K., Borden, A., 2002. Organizing the musical canon: the repertoires of the major U.S. Symphony orchestras, 1842 to 1969. Poetics 30, 35–61.

ESSnet-Culture, 2012. European Statistical System Network on Culture. Final Report. European Union, Luxembourg.

Eurostat, 2009. Harmonised European Time Use Surveys: 2008 Guidelines. European Communities, Luxembourg.

Eurostat, 2011. Cultural Statistics. European Union, Luxembourg.

Florida, R., 2002. The Rise of the Creative Class. And How It's Transforming Work, Leisure, Community and Everyday Life. Basic Books, New York.

Frey, B.S., Pommerehne, W.W., 1989. Muses and Markets. Basil Blackwell, Oxford.

Heckathorn, D.D., Jeffri, J., 2001. Finding the beat: using respondent-driven sampling to study jazz musicians. Poetics 28, 307–329.

Higgs, P., Cunningham, S., 2008. Creative industries mapping. Where have we come from and where are we going? Creative Industries Journal 1, 7–30.

Janssen, S., Kuipers, G., Verboord, M., 2008. Cultural globalization and arts journalism. The international orientation of arts and culture coverage in Dutch, French, German, and U.S. newspapers, 1955 to 2005. American Sociological Review 73, 719–740.

Kaple, D.A., 2002. Current data resources on nonprofit arts organizations. American Behavioral Scientist 45, 1592–1612.

Karttunen, S., 1998. How to identify artists? Defining the population for 'status-of-the-artist' studies. Poetics 26, 1–19.

Karttunen, S., 2001. How to make use of census data in status-of-the-artist studies: advantages and shortcomings of the Finnish register-based census. Poetics 28, 273–290.

KEA European Affairs, 2006. The Economy of Culture in Europe. European Commission, Brussels.

Madden, C., 2005. Cross-country comparisons of cultural statistics: issues and good practice. Cultural Trends 14, 299–316.

Morrone, A., 2006. Guidelines for Measuring Cultural Participation. UNESCO, Montreal.

Pettit, B., 2000. Resources for studying public participation in and attitudes towards the arts. Poetics 27, 351–395.

Rengers, M., 2002. Economic Lives of Artists: Studies into Careers and the Labour Market in the Cultural Sector. Utrecht University, Utrecht.

Rössel, J., Otte, G., 2010. Culture, in: german data forum (RatSWD). In: Building on progress. Expanding the research infrastructure for the social, economic, and behavioral sciences, vol. 2. BudrichUniPress, Opladen, pp. 1153–1172.

Schulze, G., 1992. Die Erlebnisgesellschaft. Kultursoziologie der Gegenwart. Campus, Frankfurt.

Schuster, J.M., 2007. Participation studies and cross-national comparison: proliferation, prudence, and possibility. Cultural Trends 16, 99–196.

Sternberg, R.J., Lubart, T.I., 1999. The concept of creativity: prospects and paradigms. In: Sternberg, R.J. (Ed.), Handbook of Creativity. Cambridge University Press, Cambridge, pp. 3–15.

Throsby, D., 2001. Defining the artistic workforce: the Australian experience. Poetics 28, 255–271.

Throsby, D., 2008a. The concentric circles model of the cultural industries. Cultural Trends 17, 147–164.

Throsby, D., 2008b. Modelling the cultural industries. International Journal of Cultural Policy 14, 217–232.

Throsby, D., Zednik, A., 2010. Do You Really Expect to Get Paid? An Economic Study of Professional Artists in Australia. Australia Council, Sydney.

UNESCO, 1986. The UNESCO framework for cultural statistics (FCS). In: Third Joint Meeting on Cultural Statistics, 17–20 March 1986. Office of Statistics. UNESCO.

UNESCO, 2001. UNESCO Universal Declaration on Cultural Diversity. UNESCO, Paris.

UNESCO-UIS, 2009. The 2009 UNESCO Framework for Cultural Statistics (FCS). UNESCO-UIS, Montreal.

UNESCO-UIS, 2012. Measuring the Economic Contribution of Cultural Industries. A Review and Assessment of Current Methodological Approaches. UNESCO-UIS, Montreal.

Weckerle, C., Gerig, M., Söndermann, M., 2008. Creative Industries Switzerland. Birkhäuser, Basel.

Data Bases and Statistical Systems: Demography

Michaela Kreyenfeld, Hertie School of Governance, Berlin, Germany
Frans Willekens, Max Planck Institute for Demographic Research, Rostock, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article deals with the availability of large-scale data for demographic analysis. The main sources of data that demographers work with are censuses data, microcensus data, population registers, other administrative data, survey data, and big data. Data of this kind can be used to generate demographic rates and study demographic processes, such as family formation, marriage, divorce, residential mobility, and death. The advantages and pitfalls of different types of data are summarized.

Introduction: The Core of Demographic Research

Demography is the study of population dynamics and population structures. At the heart of this discipline lies the investigation of fertility, migration, and mortality processes. These processes determine the size and the composition of a population. Among the related issues that demographers investigate are partnership dynamics (including marriage, cohabitation, separation, divorce, and repartnering), residential mobility, aging, health, and retirement. A general principle that unifies demographic research is the firm conviction that demographic processes need to be located in time. Time includes two prime dimensions, namely, calendar time and age. An important tool used to visualize the age-calendar dimension is the Lexis diagram, which displays age on the horizontal axis and calendar time on the vertical axis (*see* Demographic Techniques: LEXIS Diagram). The diagonal lines in the Lexis diagram represent the aging of different birth cohorts. The evolution of demography as a discipline is strongly connected to the invention of the life-table technique, which originated in the works of John Graunt during the seventeenth century. Till today, life-table techniques are among the main methods demographers use in the study of age-related processes (*see* Life Table).

Demography is a discipline that relies heavily on the availability of large-scale data. During the seventeenth and eighteenth centuries, parish registers were the prime source of demographic analysis (Bengtsson and Dribe, 2006; Knodel, 1988). Genealogies have also been employed as sources in the reconstruction of the demographic behavior of historical populations (*see* Demographic Techniques: Family Reconstitution). With the formation of nation states in Europe during the nineteenth century, national governments developed an increasing interest in the size and structure of their populations. This interest later led to the establishment of national statistical offices. In Prussia, for example, the 'Königlich Preußische Statistische Bureau' was established in 1805. This office collected extremely rich population data, which are still used to study regional fertility dynamics (Lee et al., 1994). The 'modern' census was also created in this era.

A population census is an enumeration of all residents in a country or present in a country at the time of the census. The enumeration of all individuals necessarily implies that not more than a few characteristics can be recorded for each individual. More detailed information is usually collected from a selection of individuals that is representative for the entire population (microcensus). In addition to censuses and sample surveys, administrative processes are sources of information on the population. This data are a by-product of administrative processes. Civil registration is an administrative process by which a government records the vital events of its citizens and residents (e.g., births, deaths, marriages, divorces, migrations). Other administrative processes produce useful data too, e.g., tax collection, pension benefit payments, and issuing residence permits. A new source of data on individuals and their activities are tracking devices in sensors, mobile devices, credit cards, electronic identification cards, social media, Internet searches, and so on. Much of these data are generated in real time and are referred as *big data* because of their massive volume. Mayer-Schönberger and Cukier (2013) situate the emergence of big data in a historical perspective, with the census being the first systematic recording of information about members of a given population. The census could collect limited information only at periodic intervals. The modern tracking devices collect a variety of data in real time. In this article, brief descriptions of the main data sources are presented: census data, microcensus data, population registers, other administrative produced data, survey data, and big data. The processing of these data yields parameters of demographic processes (e.g., demographic rates) and indicators of population dynamics.

Census Data

A population census is a systematic recording of information on all members of a population usually residing in a country (*de jure*) or present at the time of enumeration (*de facto*). The United Nations (UN) recommends that censuses be conducted at least every 10 years, but countries vary greatly in the extent to which they comply with this guidance. The UN has provided a list of core aspects which the national statistical offices are expected to include in the census (United Nations, 2006): sex, age, marital status, employment status, level of education, country of birth, housing conditions, and household composition. Traditionally, censuses have been based on face-to-face surveys. However, in the most recent census, several countries

(such as the Nordic countries, Germany, The Netherlands, Slovenia, and Austria) have turned to register-based censuses in which several registers (e.g., employment, housing, and population registers) are combined. Such a census is sometimes referred to as a *virtual census*. Most other countries conduct traditional censuses in which the total population is interviewed face-to-face. Nonetheless, in some of these countries (e.g., Canada), the traditional census is a subject of political debate. The US Bureau of the Census has considered for decades how government records might help the traditional census-taking process. The use of these records raises concerns about personal privacy.

Censuses are of considerable importance for demographic research because they provide 'official' population counts for generating demographic rates by age and sex at the regional level. A key comparative project that makes use of the available census data is the Integrated Public Use Microdata Series (https://www.ipums.org/). Monitored by the Minnesota Population Center, the project makes harmonized census data available for a large number of countries around the world.

Microcensus Data

Most countries also conduct microcensuses on a regular basis. Microcensuses are surveys that are monitored by the national statistical offices and include a fraction of the total population. Microcensuses include data, which are richer than the data collected by censuses, and they are conducted in relatively short and regular intervals (often on a yearly basis). Among the forms of demographic information that come from microcensus data are, for example, the number of cohabiting unions in a population, the share of single-headed households, or the share of individuals in need of care. Because of their large sample sizes, microcensuses can provide more reliable indicators than social science surveys (see below). In addition, the nonresponse rates are very low in the microcensuses because participation is compulsory in some countries. However, microcensus or census data alone may not be sufficient for generating demographic rates because they often only contain cross-sectional information. However, methods have been developed that allow researchers to exploit microcensus and census data for demographic analysis. The 'own-child method' is a method that uses information on the number of children who live in a household at the time of a survey to reconstruct the fertility histories of women (Cho et al., 1986). This method is generally used for historical populations, but it may also be applied to contemporary societies when order-specific fertility data are not available from the vital statistics. In the 1980s, the UN published a manual of methods for deriving useful demographic indicators from census data. The manual has been updated recently (Moultrie et al., 2013) and is available online (http://demographicestimation.iussp.org/).

Population Register

Several countries, especially in Europe, maintain population registers or other types of registers that contain demographic data (such as birth, marriage or death registers). Individuals are legally required to report births, deaths, and migrations. When a marriage is performed in a civil ceremony involving a legal contract between spouses, marriages are registered and the data becomes part of the register. The dissolution of a marriage, or the end of a marriage contract, is registered too. In many countries, including most European countries, cohabitation has been on the rise since the 1970s. While some countries now provide the possibility to register a cohabiting union (such as France and the Netherlands), other countries (such as Germany) only provide this option for same-sex unions. Registered partnerships may, thus, also be captured in the civil registration system.

Registers that contain information of vital events are the prime form of data with which demographers work. The data generally contain counts of births, deaths, marriages, and residential moves. Births are commonly broken down by the age of the mother, the biological birth order of the child, the duration since the mother's last birth, and the mother's marital status. In most national statistics, birth statistics are only available by the mothers' characteristics. Marriage data are commonly broken down by aspects such as age, marriage order, and citizenship of the partner. Divorce data also fall into the category of demographic data, and are commonly available by age, marriage duration, and citizenship of the partner. Deaths are usually recorded by age, sex, marital status, and cause of death.

Migration data is often less widely available. Not all countries keep registers including data on changes of residence. If they do, internal and international migration statistics are derived from the registers. The quality of register-based migration data depends on reporting practices. If people do not report a change of residence and no system exists to track where people reside, migrations are undercounted. The undercounting is particularly frequent and extensive for emigrations, i.e., migrations to other countries. Some national statistical offices adjust international migration statistics regularly for undercounting. For instance, in the Netherlands, the official emigration statistics consist of emigrations reported and *administrative corrections*. Canada adjusts the number of immigrants. In most countries, though, undercounting goes unnoticed until the next population census. The 2011 census of Germany found 1.5 million people missing. Most of them are likely persons who failed to declare their emigration. The German statistical office had estimated that there were 7.3 million residents of Germany without German passports, but the census counted just 6.2 million. Residence is often not included in the migration registers. For instance, in the UK, the Office of National Statistics (ONS) uses three administrative data sources to produce internal migration statistics: the National Health Service Central Register, the Patient Register Data Service, and the Higher Education Statistics Agency data. To produce international migration data, ONS uses the census and the UK International Passenger Survey. Most official migration statistics tend to underestimate the volumes of immigration and particularly emigration, and to double-count circular and return migrants (United Nations Economic Commission for Europe and Statistical Office of the European Union, 2010). The difficulty of capturing illegal migrants in migration statistics is another source of serious error in the migration counts of some countries.

Differences in data sources, measurement schemes, and definitions of migration explain the lack of comparability of migration data across countries. Despite long-standing efforts to harmonize international migration statistics, the data are not fully comparable across countries (Lemaitre, 2005; United Nations, 1949). These discrepancies in international migration statistics become obvious when the migration flows between sending and receiving countries are directly compared (Raymer and Willekens, 2008).

Other Administrative Data

Administrative data (also called *process-produced data*) are data that are collected as a byproduct of administrative processes. Apart from birth, death, and marriage registers (see Section Population Register), tax, pension, employment, and registers of national health insurance providers may contain valuable demographic information. Administrative data are not subject to many of the problems associated with social science surveys, such as small sample sizes, high nonresponse rates, and the related underrepresentation of certain subpopulations (see below). However, countries vary greatly in terms of the availability of administrative data for demographic research; the Nordic countries of Europe are at the forefront of the development of such data (United Nations, 2007) (see also http://www.suda.su.se/simsam/). For other countries, data protection issues and difficulties in combining different types of administrative data create hurdles for this type of data analysis. In a time of shrinking response rates in social science surveys and the growing capacity of computers to manipulate large-scale data, the exploitation of administrative data is an obvious avenue for future demographic research. Several countries have already taken this route and have replaced their traditional censuses with register-based censuses in the most recent census round. The extent to which register-based demographic research will also be able to replace, enrich, or validate survey-based research is largely contingent on the willingness of the providers of administrative data to make their data available for scientific usage and the willingness of the national governments to promote and allow the linkage of different registers.

Social Science Surveys

Data from large-scale social surveys, which have become increasingly available since the 1980s, have revolutionized demographic research. Previously, demographers were primarily concerned with population data, which were mainly available at the aggregate level. But as a result of the wide availability of social science surveys, demographers have turned increasingly to the analysis of microlevel data. In addition, demographic techniques for microlevel data have advanced. In particular, the development of event history techniques has been paramount for demographic research. These techniques arose from classical life table techniques, but are also related to statistical methodology and modern regression analysis (Hoem, 1978; Mayer and Tuma, 1990). Sequence analysis and methods for displaying demographic behavior across the life course are also related to developments in event history techniques (Raymer and Willekens, 2008; Willekens, 1999).

To generate demographic rates with survey data, longitudinal data that locate the demographic events in time are needed. In this context, it is important to distinguish between retrospectively and prospectively collected data. *Retrospective data* are collected by asking people about their recollections of the timing and occurrence of past events. This information is usually gathered in a modularized way. In different sections of the questionnaire (*modules*), the different *domains* of a respondent's life course (like the fertility, residential, and partnership histories) are surveyed. This means that in these different parts of the questionnaire, the occurrence and timing of past events are recorded, often by relying on so-called *event history calendars*. In *prospective surveys*, life histories are reconstructed by reinterviewing each person at different points in his or her life course. This is done in a follow-up study (*panel survey*) in which the same respondent is reinterviewed on a regular basis (often in a yearly rhythm). By adding together the data gathered over time, it is possible to generate longitudinal information for the person. These data are, however, only partially comparable with the retrospectively collected life histories. A panel study only includes information on a person at different points in time in his or her life, and not for the person's entire life course. However, panel surveys usually also survey retrospective information that covers the last year or the time since the last interview. This type of information allows for the reconstruction of the person's complete life history, or at least the portion of the life history for the period during which the person participated in the panel study.

Retrospective Surveys

Compared to panel surveys, retrospective surveys are far less costly. Furthermore, while it takes several years to build up a panel survey, retrospective surveys can be used immediately for demographic analysis (Beckett et al., 2001). Moreover, for some processes, like cohabitation, a retrospective survey is the only tool for learning about past behavior, because no information exists from other data sources. Retrospective surveys have been important in helping us understand family behavior in contemporary societies. The World Fertility Survey (WFS), which was launched during the 1970s and early 1980s, was one of the largest social science surveys that collected fertility and family data for developing countries (Cleland and Verma, 1989). The WFS was followed by the Demographic and Health Surveys (DHS). DHS have been organized in a large number of developing countries and countries in transition (www.measuredhs.com). The Fertility and Family Surveys (FFS), which focused on developed countries, were conducted during the 1990s. The Generations and Gender Surveys (GGS), which were conducted around 2005, were similar to the FFS, although the GGS has also been conducted as a panel study in some European countries (http://www.ggp-i.org/). Some countries have a tradition of conducting cohort studies, in which the life courses of particular birth cohorts of women and men are surveyed. The British Cohort Study and the German Life History Survey are examples of cohort studies.

There are several advantages associated with the use of retrospective surveys, but also some significant pitfalls. The

largest problem that arises in conjunction with retrospective surveys is recall bias. Even salient events like marriage and childbirth have been shown to suffer from this distortion (Ní Bhrolcháin et al., 2011). For cohabitation, marriage, separation, and divorce, there is convincing evidence that recall bias, or the deliberate disclosure of information during the interview, poses a serious challenge to retrospective surveys (Mitchell, 2012). Another source of bias is related to sampling issues. In a survey, only the people who are residing in the country at the time of the survey are included. This population may have changed over time due to in- and out-migration. Thus, a retrospective survey cannot fully mirror the past behavior of the population resident in the country in the past. *Selection on survival* is another conundrum that is related to the fact that only the surviving population can be interviewed. Selection on survival may affect the analysis of family behavior, if, for example, fertility and family patterns correlate with mortality rates. The older the surveyed cohort, the stronger the bias. Obviously, demographic events like death cannot be studied with retrospective surveys unless proxy informants are used (e.g., information of a death of a person from other household members).

Prospective Surveys

Panel data may be used to overcome many of the shortcomings of retrospective data. A panel follows the same person over a certain period of time. Thus, mortality risks may also be investigated with a panel study, provided the information for the reasons why a person has dropped out of a panel is provided. One of the longest-running panel studies, which has also served as a model for many later surveys, is the Panel Study of Income Dynamics (PSID). Although these data were not surveyed for demographic purposes, they have been employed for various types of demographic analysis (Dowd et al., 2011). In the realm of family and fertility behavior, the National Longitudinal Survey of Youth, which was initiated in 1979, was one of the first large-scale panel projects that focused on family behavior. One recent American initiative is the Fragile Family Child Wellbeing Study. In Europe, the Netherlands Kinship Panel Survey and the German Family Panel are examples of more recent family panels (http://www.pairfam.de/en). They combine a cohort and a *multi-actor design*, meaning that in addition to the 'anchor respondent,' who is selected from specific birth cohorts, the respondent's parents and children are interviewed. One of the first and largest panel studies of the older population is the Health and Retirement Study, which was launched in 1992. For Europe, the Survey of Health, Ageing and Retirement in Europe is one of the landmark projects on the elderly (http://www.share-project.org/).

Panel surveys are indispensable for studying demographic behavior. In the realm of family research, they enable researchers to study how 'soft indicators' (like attitudes and values) affect subsequent fertility and partnership behavior. The relationship between employment and family choices has been studied using data from panel surveys, as there are very few registers that contain information about both employment and fertility careers. The growing availability of household and multiactor panels also provides researchers with the opportunity to study how the couple or household context influences fertility and family decision-making. For mortality research, panel surveys have been regularly employed to study how health-related behavior (like smoking habits) affects old-age disability or mortality. The inclusion of *biomarkers* (like grip strength, blood pressure, or even blood samples) in social science panel studies is an important new area that will help improve our understanding of the interplay of biological and social factors in family and health-related behaviors (McDade et al., 2007).

While panel surveys are very important for demographic research, there are a number of drawbacks to their use. One of the most significant problems associated with panel data is the issue of *attrition* (i.e., respondents dropping out of the study). Attrition is a general problem for any study that draws on a panel survey. But because of the strong correlation of attrition with residential mobility, it is a particularly severe issue for demographic analysis. Marriage, cohabitation, separation, divorce, or childbirth may lead to a residential move, and survey institutes are often unable to keep track of people as they move. A similar issue arises at higher ages, as respondents may be unavailable for an interview because they have moved into an eldercare facility, been hospitalized, or been unable to answer questions for other reasons (such as disability). The analysis of mortality risks using a survey is also contingent on the availability of information on the reasons why a respondent dropped out. Some panel studies (like the PSID and the US Health and Retirement Survey) are able to verify that a respondent dropped out due to death by linking his or her information to official death registers. In other countries, it is not possible to confirm deaths in this way. Another serious issue that affects social science surveys in general is the undercoverage of certain subpopulations. The underrepresentation of disabled individuals and people in eldercare facilities creates a problem for demographic analyses of transitions in the later life course. The undercoverage of young, mobile men, and the inability to properly capture nonresidential fathers, make investigations of family biographies, especially those involving male respondents, challenging (Rendall et al., 1999).

Big Data

Data generated in real time by various types of tracking devices are not used as often in demographic analysis. Zagheni and Weber (2012) studied international mobility using geolocation data generated when individuals log into the computer. IP addresses are mapped to geographical locations, and a person's duration of residence is identified by repeated logins. From that information, short- and long-term migration flows and the typical migration age profile can be produced. A user's IP address is recorded whenever the Internet is accessed. The IP address can be linked to other information on the user, such as cookies containing unique identifiers, purchases paid for by credit card, Internet banking, cameras in combination with image processing software, electronic medical records that include medical and treatment histories, administrative databases, or other databases containing personally identifiable information. From that information, a personalized profile that includes a variety of conditions and activities can be created. Profiling, i.e., the construction of

personalized profiles generated by record linkages and data analysis, is technically possible and offers opportunities for demographic analysis. Linked with Geographical Position Systems (GPS), detailed population profiles can be generated at street level. Whether that technology will be used on a large scale depends on the outcome of the political debate on the protection of personal data. The debate is not entirely new. In the 1980s, popular resentment in West Germany prevented the population census from being conducted. Some felt that the questions to be asked were quite personal and recalled the use of census data to target Jews and other minorities during the Nazi era. The Constitutional Court canceled plans for the 1983 census, and required a revision of the process. The modified census was eventually held in 1987.

Demographic Rates

Structure of Demographic Rates

Demographic indicators are generated by relating the number of events to the population at risk of experiencing an event. The numerator is defined by the number of events (births, deaths, and marriages) recorded in the civil registration system. The denominator is defined by the population count (or the risk population), which commonly comes from census data. Censuses are generally conducted at regular intervals. For the intercensal period, the population size is estimated through updates of the population count using information on the numbers of deaths and births and the volumes of in- and out-migration. As the time since the last census increases, population estimates are subject to increasing error, especially because migration data are often less reliable. In the past, inaccuracies in population estimates have led to inaccurate assessments of the mortality rates of migrants, as well as to erroneous conclusions regarding mortality at higher ages (Kibele et al., 2008).

Availability and Comparability of Demographic Rates

Demographic indicators for various countries are compiled by the United Nations, Eurostat, and the Organisation for Economic Co-operation and Development. These data contain a wide array of demographic measures (such as the share of nonmarital births, marriage rates, divorce rates, life expectancies, migration flows, etc.). However, data across countries are often not fully comparable. Migration data in particular often are not comparable (as was mentioned above). In addition, discrepancies also exist in the way in which births and deaths are documented. One difference concerns the documentation of the age at which an event occurred. Some countries collect data on vital events by the age reached during the year (ARDY method). Other countries record vital events by the age in completed years (ACY method). The Human Mortality Database (HMD) (2002–present) run by the University of California at Berkeley (United States) and the Max Planck Institute for Demographic Research (Rostock, Germany) (http://www.mortality.org/) seeks to harmonize mortality statistics. The database includes death counts, population counts, and life tables for more than 30 countries. The Max Planck Institute for Demographic Research (Rostock, Germany) and the Vienna Institute of Demography (Vienna, Austria) (http://www.humanfertility.org/) offer a comparable database for fertility data: the Human Fertility Database (HFD). The HFD contains birth counts, population data, and age-specific birth rates. A special effort has been made to collect order-specific fertility data, which allows for the generation of indicators such as the age at first childbearing or the share of childlessness by the birth cohorts of women.

Period and Cohort Measures

Demographic indicators may be distinguished by period and cohort measures. Period rates relate the number of events to the average population size in a given year. The *crude birth rate*, for example, relates the number of births to the midyear population in a certain year. The *general fertility rate* uses the number of women aged 15–50 as the denominator, based on the fact that only women are at risk of childbirth. The *total fertility rate* (TFR) is a measure that standardizes for the number of women of a given age. The TFR is conventionally interpreted as the number of children a woman bears over her lifetime if she experiences age-specific fertility rates observed in a given period. The decline in the TFR observed in most industrialized countries since the 1960s has been interpreted as an indication of shrinking family size. However, period measures, like the period TFR, are distorted by so-called 'tempo effects' (Bongaarts and Feeney, 1998). Tempo effects mean that annual birth (or death) rates may decline as vital events are postponed until (or occur) later in the life course (*see* Demographic Measurement: General Issues and Measures of Fertility). For fertility behavior, this is a nontrivial issue, because the age at childbearing is constantly subject to change. The increase in age at first childbearing has risen since the 1960s in most industrialized countries, a process which has suppressed period fertility rates. For mortality development, tempo distortions have resulted in an underestimation of longevity in contemporary societies (Oeppen and Vaupel, 2002). While demographers have developed so-called 'tempo adjustments' to correct for the distortions of demographic period measures, there is so far no consensus regarding which method should be used to correct for tempo effects (Sobotka and Lutz, 2011).

Demographic events, like births or deaths, are events that are situated in the life course of individuals. Period measures, which are by definition focused on a cross section, cannot fully account for this fact. Period measures are essentially 'synthetic' measures of demographic behavior. Cohort measures that give the number of children a cohort of women have had over their lives more accurately describe the fertility behavior of a population. The same applies to cohort mortality rates, which give the life expectancy of a certain birth cohort. The above-mentioned Human Fertility Database (HFD) and Human Mortality Database (HMD) contain cohort indicators that make it possible to display cohort fertility and mortality trends. The disadvantage of using cohort data is that the behavior of a cohort can only be fully assessed for 'completed' cohorts (Ni Bhrolchain, 1992). For fertility behavior, this means that the total number of children can be determined only after the members of a respective cohort have reached the end of their childbearing years. For mortality, this means that the last

member of a given cohort must have died before the cohort's life expectancy can be generated.

Conclusion

Vital statistics continue to be a prime source of data for identifying overall fertility, marriage, divorce, and mortality trends. But because of the small number of items that are usually included in this type of data, the extent to which these statistics can be used to explain the various facets of demographic behavior is limited. In particular, these statistics fall short in their ability to capture the recent changes in family patterns. Most national statistical offices only provide marriage and divorce data, and are unable to provide information about the occurrence and stability of cohabitation and other forms of nonmarital living arrangements, which have become prevalent in recent decades.

Vital statistics are also limited when it comes to generating 'real exposure rates' (or 'occurrence–exposure rates'). Demographic rates are generated by dividing the number of events by the '*exposure population*.' The exposure population is generated by summing up the 'person-years' or 'person-months' that a population has been at risk of experiencing an event. For *absorbing* demographic events like death, exposure may be generated from population counts. For *recurrent* demographic events – like births, marriages, and divorces – it is more cumbersome to generate real exposure rates based on the overall population counts. For second births, for example, only women who have experienced a first birth are at risk of having a second one, and thus only those women should be used to calculate second birth rates. For divorce behavior, only married individuals are at risk of experiencing divorce, and thus only these individuals enter the risk population. Failing to accurately pinpoint the risk population is a frequent distortion of demographic rates. However, most national statistics are unable to provide the types of information that would enable researchers to generate real exposure rates for fertility and nuptiality indicators. Survey data and administrative data must fill this gap. The question of to what extent 'big data' may be used for demographic analysis in the future remains open to discussion.

See also: Abortion, Demography of: United States; Adoption, Demography of; Assimilation of Immigrants; Assortative Mating in the Marriage Market; Biomarkers in Demographic Research; Census Microdata; Censuses, Population: Comparative International Aspects; Censuses: Current Approaches and Methods; Censuses: History and Methods; Cohabitation: United States; Conditional Cash Transfers: Influence on Marriage and Fertility; Demographic Measurement: General Issues and Measures of Fertility; Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Demographic Models; Demographic Techniques: Data Adjustment and Correction; Demographic Techniques: Family Reconstitution; Demographic Techniques: Indirect Estimation; Demographic Techniques: Inverse Projection; Demographic Techniques: Lexis Diagram; Demography: History Since 1900; Dowries and Other Marital Transfers; Ethnic Intermarriage; Event History Analysis in Continuous Time; Evolutionary Demography: A Synthesis of Two Population Sciences; Experimental Methods in Survey Research in Demography; Family Size Preferences; Family Theory: Competing Perspectives in Social Demography; Family Theory: Economics of Childbearing; Family Theory: Economics of Intergenerational Relations; Family Theory: Economics of Marriage and Divorce; Fatherhood; Fertility Change: Quantum and Tempo; Fertility Transition: Cultural Explanations; Fertility Transition: Latin America and the Caribbean; Fertility Transition: Middle East and North Africa; Fertility Transition: Southeast Asia; Fertility Transition: Sub-Saharan Africa; Fertility Trends in the Formerly Socialist Countries of Europe; Fertility and Culture: Anthropological Insights; Fertility and Social Networks; Gender, Time-Use, and Fertility Recovery in Industrialized Countries; Hutterite Demography; Internal Migration in Industrialized Countries; Internal Migration: Developing Countries; International Migration by Ethnic Chinese; International Migration by Ethnic South and Southeast Asians; Life Table; Microsimulation in Demographic Research; Migration, Theory of; Migration: Cultural Aspects; Mortality Crossover; Mortality: The Great Historical Decline; Multilevel Models in Demography; Old Age: Definitions, Theory, and History of the Concept; Organizations and Culture; Period and Cohort Analysis in Demography; Population Cycles, Formal Theory of; Population Dynamics: Classical Applications of Stable Population Theory; Population Dynamics: Mathematic Models of Population, Development, and Natural Resources; Population Dynamics: Momentum of Population Growth; Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems; Population Dynamics: Theory of Nonstable Populations; Population Dynamics: Theory of Stable Populations; Population Forecasts; Repartnering and Stepchildren; Risk Screening, Testing, and Diagnosis: Ethical Aspects; Second Demographic Transition; Sexuality: Cultural Aspects; Twin Studies in Demography.

Bibliography

Beckett, M., Da Vanzo, J., Sastry, N., Panis, C., Peterson, C., 2001. The quality of retrospective data: an examination of long-term recall in a developing country. The Journal of Human Resources 36 (3), 593–625.

Bengtsson, T., Dribe, M., 2006. Deliberate control on natural fertility population: Southern Sweden, 1766–1864. Demography 43 (4), 727–746.

Bongaarts, J., Feeney, G., 1998. On the quantum and tempo of fertility. Population and Development Review 24 (2), 271–291.

Cho, L.-J., Retherford, R.D., Choe, M.K., 1986. The Own-Children Method of Fertility Estimation. University of Hawaii Press for the Population Institute, East-West Center, Honolulu.

Cleland, J., Verma, V., 1989. The world fertility survey: an appraisal of methodology. Journal of the American Statistical Association 84 (407), 756–767.

Dowd, J.B., Albright, J., Raghunathan, T.E., Schoeni, R.F., LeClere, F., Kaplan, G.A., 2011. Income and mortality in the USA over three decades. International Journal of Epidemiology 40 (1), 183–188.

Hoem, J.M., 1978. Demographic incidence rates. Theoretical Population Biology 14 (3), 329–337.

Kibele, E.U.B., Scholz, R.D., Shkolnikov, V.M., 2008. Low migrant mortality in Germany for men aged 65 and older: fact or artifact? European Journal of Epidemiology 23 (6), 389–393.

Knodel, J.E., 1988. Demographic Behavior in the Past: A Study of 14 German Village Populations in the Eighteenth and Nineteenth Centuries. Cambridge University Press, Cambridge.

Lee, R.D., Galloway, P.R., Hammel, E.A., 1994. Fertility decline in Prussia: estimating influences on supply, demand, and degree of control. Demography 31 (2), 347–373.

Lemaitre, G., 2005. The comparability of international migration statistics. OECD Statistics Brief 9, 1–8.

Mayer-Schönberger, V., Cukier, K., 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think. John Murray, London.

Mayer, K.U., Tuma, N.B., 1990. Life course research and event history analysis: an overview. In: Mayer, K.U., Tuma, N.B. (Eds.), Event History Analysis in Life Course Research. University of Wisconsin Press, Madison, pp. 3–20.

McDade, T.W., Williams, S., Snodgrass, J.J., 2007. What a drop can do: dried blood spots as a minimally invasive method for integrating biomarkers into population-based research. Demography 44 (4), 899–925.

Mitchell, C., 2012. Are divorce studies trustworthy? the effects of survey nonresponse and response errors. Journal of Marriage and Family 72 (4), 893–905.

Moultrie, T.A., Dorrington, R.E., Hill, A.G., Hill, K., Timæus, I.M., Zaba, B., 2013. Tools for Demographic Estimation. International Union for the Scientific Study of Population, Paris.

Ni Bhrolchain, M., 1992. Period paramount? A critique of the cohort approach to fertility period paramount? A critique of the cohort approach to fertility. Population and Development Review 18 (4), 599–629.

Ní Bhrolcháin, M., Beaujouan, É., Murphy, M., 2011. Sources of error in reported childlessness in a continuous British household survey. Population Studies 65 (3), 305–318.

Oeppen, J., Vaupel, J.W., 2002. Broken limits to life expectancy. Science 296 (5570), 1029–1031. http://dx.doi.org/10.1126/science.1069675.

Raymer, J., Willekens, F., 2008. International Migration in Europe. Wiley, Chichester, Sussex.

Rendall, M.S., Clarke, L., Peters, H.E., Ranjit, N., Verropoulou, G., 1999. Incomplete reporting of men's fertility in the United States and Britain: a research note. Demography 36 (1), 135–144.

Sobotka, T., Lutz, W., 2011. Misleading policy messages derived from the period TFR: should we stop using it? Comparative Population Studies 35 (3), 637–664.

United Nations, 1949. Problems of Migration Statistics. UN, Lake Success, New York.

United Nations, 2006. Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing. UN, Geneva.

United Nations, 2007. Register-based Statistics in the Nordic Countries: Review of Best Practices with Focus on Population and Social Statistics. UN, Geneva.

United Nations Economic Commission for Europe, Statistical Office of the European Union, 2010. Guidelines for Exchanging Data to Improve Emigration Statistics. UN, Geneva.

Willekens, F.J., 1999. The life course: models and analysis. In: Van Wissen, L.J.G., Dykstra, P.A. (Eds.), Population Issues. An Interdisciplinary Focus. Kluwer Academic/Plenum Publishers, New York, pp. 23–51.

Zagheni, E., Weber, I., 22–24 June 2012. You are where you e-mail: using e-mail data to estimate international migration rates. WebSci, 1–10. http://www.demogr.mpg.de/publications/files/4598_1340471188_1_Zagheni&Weber_Websci12.pdf.

Data Bases and Statistical Systems: Economics (General)

F Thomas Juster, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 14, pp. 9770–9777, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The entry on 'Databases, Core: Economics' gives the readers an overview on the evolution and structure of micro- and macroeconomic databases. In this context, the article summarizes the history and recent development of several economic databases in the United States and beyond, including sample survey data, databases for households and for business establishments, as well as databases of official national accounts. In addition, it revises the distinction between cross-sectional and longitudinal data systems. Finally, the article concludes with a brief assessment of future developments.

Additions to scientific knowledge generally require interactions between theory and measurement. For example, the Hubbel Telescope provides astronomers with information about the universe, biologists have the Human Genome Project to provide them with a mapping of the genetic characteristics of the human species, and physicists have large-scale particle accelerators to allow experiments designed to identify types of matter that could not have been observed in the past because the instrumentation was not powerful enough. It is a common observation among scientists that the rate of scientific progress is a function of the degree to which new measurement techniques enable us to observe events that could not have been observed in the past.

What are the social science counterparts of the Hubbel Telescope, the Human Genome Project, and the particle accelerator? It turns out that, in the social sciences generally and in economics in particular, the evolution of the sample survey provides a close counterpart to the natural sciences' powerful measurement devices.

This article examines, firstly the evolution of sample survey data in economics, often called microdata and then discusses databases that represent households as well as those that stand for business establishments; before it revises the distinction between cross-sectional and longitudinal data systems. Secondly, the entry reflects on annual collections of the official national accounts data; additionally often referred to as macrodata, main databases that deal with economic processes and measurements of economic activity are considered; data systems and the principal features of a selection of the most important data systems both in the United States and in other countries are summarized. The article concludes with a brief assessment of future developments.

History of Economic Databases

The earliest use of quantitative empirical information in the development of economics as a science was probably in the living standards surveys conducted in the United Kingdom during the nineteenth and the early twentieth centuries. These early survey studies were descriptions of income, expenditures, and nutrition in cities or towns, and were designed to help understand poverty. Perhaps the best known studies are those by Charles Booth (1891–1897), and by Seebohm Rountree (1902), which examined York. Earlier work, more in the mode of intensive case studies than surveys, was carried out by Frederic LePlay (1855). Another set of very early studies was the establishment of firm surveys focused on measuring production. Systematic work along those lines was done in the United States during the early part of the twentieth century by Simon Kuznets and his colleagues at the National Bureau of Economic Research (Kuznets, 1941). The Kuznets studies were basically macro descriptions of change over time in the level of production in the U.S. economy, along with a variety of industry subaggregates. The information in these early studies was typically collected from sample surveys (or censuses) of business establishments, conducted by the U.S. Census Bureau.

A distinguishing feature of these early establishment measurements is that the basic microdata were not available to outside analysts. These studies were solely used to produce aggregate estimates for the economic system as a whole, and were not thought of as sources of information about micro-level economic behavior.

Over the twentieth century, there has been an explosion of economic measurements based on sample surveys. In the early part of the century the focus was mainly on business establishments and on the measurement of aggregate economic activity, as previously noted. Later in the twentieth century, starting with the Great Depression of the 1930s, microeconomic databases, particularly those relating to households, began to appear with increasing frequencies. In 1935 and 1936, for example, a massive study of household expenditures was undertaken by the U.S. Bureau of Labor Statistics (BLS), in an attempt to get a better understanding of the impact of the depression on living standards. The Current Population Survey (CPS), which provides the basic measure of unemployment in the United States, began in the early 1940s. The BLS conducted consumer expenditure surveys designed to produce weights for the Consumer Price Index as early as 1890, in 1935–36 as just noted, and in their current form in 1950, the early 1960s, and the early 1970s.

All of the sets of microdata just noted were designed and collected by governmental units. During the latter part of the twentieth century, these governmental efforts began to be supplemented with microdata collections funded by governments but designed and conducted in the private sector,

typically involving academic survey research units, such as the National Opinion Research Center (NORC) at the University of Chicago and the Survey Research Center (SRC) at the University of Michigan. These private sector microdata sets were almost entirely household data, and included the series of studies of consumer net worth titled the Survey of Consumer Finances (begun in the late 1940s), the National Longitudinal Surveys of labor market activity, the Panel Study of Income Dynamics, and the Retirement History Survey (all begun in the late 1960s). During the 1970, 1980, and 1990s there was a dramatic increase in both public and private household microdata collections (*see* Data Bases and Statistical Systems: Sociology). Finally, the past decades of the twentieth century saw the beginnings of publicly available business establishment microdata in the United States.

Establishment Economic Data Sets

Central Statistical Bureaus produce virtually all nationally representative establishment data sets. In the United States, which has a unique decentralized statistical system, establishment databases are typically collected by the Census Bureau and designed (and commissioned) either by U.S. Census Bureau or by the Bureau of Labor Statistics. As noted in the introductory remarks for this article, the major use of establishment databases tends to be macroeconomic rather than microeconomic, in that they are used to produce national estimates and industry distributions of production, wage costs, employment, fringe benefits, work hours, and so on.

The history of establishment databases sponsored by the BLS goes back to the 1800s, starting with occupational pay surveys. The current employment statistics series (CES) began at BLS in 1915, the unemployment insurance series (ES 202) was started in 1935, the Occupational Employment Statistics (OES) survey in 1971, the Employment Cost Index (ECI) series in 1976, the Employee Benefits Survey (EBS) in 1980, and the hours at work survey in 1981. All of these surveys are designed to produce aggregate statistics from which monthly, quarterly, or annual estimates of change can be calculated and used by policy makers and business forecasters. The microdata underlying these aggregates are typically unavailable to the research community, although there are occasional exceptions under tightly controlled circumstances.

Establishment databases that are the responsibility of the U.S. Census Bureau basically include a census of establishments in all industries conducted every 5 years (available starting in 1963), and an annual survey of manufacturing establishments (available starting in 1972). The census data for all industries are designed to provide, among other statistics, benchmark data for total production, price deflators, and product detail. The annual surveys of manufacturing provide data on shipments, wage costs, capital expenditures, materials consumption, and energy use.

These census data are, as with the BLS data described earlier, designed to produce aggregate industry statistics relevant to policy makers and forecasters. However, in recent years Census has tried to increase the use of the microdata in these surveys by creating data enclaves – Longitudinal Research Data Centers.

These enclaves exist not only at the U.S. Census Bureau, but have also been installed at a number of other locations (Boston, Pittsburgh, Los Angeles, and Berkeley). The intent is to encourage researchers throughout the country to use the microdata, with the expectation that such research will not only turn up interesting scientific findings but will also suggest improvements in the basic data.

There is somewhat greater use of establishment microdata in Western European countries than in the United States, largely because there appears to be a bit less concern with privacy (confidentiality) considerations in Europe. In some countries, establishment microdata can be combined with individual microdata, permitting, for example, analyses of the way in which worker wage trajectories vary with industry.

Structure of Microeconomic Databases

Microeconomic databases belong to one of four possible categories, depending on whether they sample households or firms, and whether they are cross-sectional or longitudinal. Whether surveys are cross-sectional or longitudinal depends on whether or not the same household or establishment is followed through time, or whether a new sample is drawn for each study so that the same household or establishment is never in consecutive studies except by chance. Studies can also be focused on households or individuals, often called demographic surveys, or they can examine business firms or establishments, often called economic surveys.

The great bulk of the microdatabases available to the scientific and policy communities for analysis consist of household or individual microdatabases, not establishment or firm databases. These household microdatabases tended to be cross-sectional earlier in the twentieth century, and increasingly longitudinal later in the century.

In part, this is due to the fact that the early databases were likely to be focused on a particular policy variable, where measurement of that variable was the principle focus of the study. Thus, for example, the CPS has a very large monthly sample of cases because its principle objective is to estimate the unemployment rate with great precision, not only for the country as a whole, but also for states and other geographic areas. The CPS has limited information with which to explain employment rates, but the very large sample provides a rich description of the level, change, and regional variation in unemployment rates. In contrast, microdata like the National Longitudinal Surveys or the Panel Study of Income Dynamics were designed as longitudinal microdatabases able to track and eventually model the dynamics of change over time.

These tendencies (for household microdata to be used more than firm or establishment data and for analysis to be increasingly longitudinal) are clearly visible in the published literature (Manser, 1998; Stafford, 1986). Over the period 1984–93, for example, there were over 550 articles in the major U.S. economic journals dealing with household microdata, compared with a little over 100 using firm or establishment microdata. Moreover, the great bulk of the latter were not based on nationally representative establishment data collected by the Census or by the BLS, but were typically local studies with the data collected by the researcher.

The very limited use of establishment data in the published literature compared with household data is probably due to three factors. First, until very recently it was difficult, if not impossible, for researchers to access establishment microdata in the United States, due to the (accurate) perception that privacy/confidentiality considerations were a much more serious problem for establishments than for households; after all, it is impossible to hide GM or AT&T in a microdata set. Second, the theory of household or individual behavior is much better developed than the theory of business behavior, probably due in part to the greater availability of rich household microdata. Third, microdata sets for establishments are likely to be produced in the public sector and to be designed to track change over time in policy-relevant aggregate variables. Thus they will be relatively strong on well-measured policy variables, and relatively weak on a rich set of explanatory variables – a combination that makes analysis of the data set less likely to show up in scientific journals.

The tendency for household studies to be increasingly longitudinal is probably due to two factors. First, it took the economics profession some time to discover the advantages of longitudinal over cross-sectional analysis, and to develop the appropriate statistical tools. Second, the earlier sets of public microdata were strongest at providing careful measurements of the policy-relevant variables that justified the survey, which requires only cross-sectional data; they were typically less strong on including the explanatory variables essential to successful longitudinal modeling.

The principal features of a number of microdatabases are summarized, including databases where the major focus is on measurement of a dependent variable (where the design is apt to be cross-sectional), and others where the focus is on providing a rich explanation of behavior (where the focus is likely to be on careful longitudinal measurement of a very large set of explanatory variables). The series that we cover in this discussion are mainly household data series available in the United States, but some information on microdatabases available from Western Europe, Asia, and the developing world is provided. The series covered includes the Surveys of Consumer Finances (SCF); the Survey of Consumer Attitudes (SCA); the National Longitudinal Surveys (NLS); the Panel Study of Income Dynamics (PSID); the Survey of Income and Program Participation (SIPP); the Consumer Expenditure Survey (CE); the Time Use survey (TU); the Current Population Survey (CPS); the Health and Retirement Study (HRS); the Luxembourg Income Study (LIS); the Living Standards Measurement Study (LSMS); and the British, German, and European versions of the Panel Study of Income Dynamics (the British Household Panel Survey, BHPS; the German Socio-Economic Panel, GSOEP; and the European Household Panel Survey, EHPS). A brief discussion of establishment databases is also included.

Survey of Consumer Finances

The SCF, sponsored by the Federal Reserve Board, was initiated just after World War II to provide some insight into the likely behavior of consumers who had accumulated large amounts of assets during the war, when many types of consumer goods and services were unavailable and thus a large fraction of income was saved. The early surveys contain detailed descriptions of asset ownership and amounts of assets held across a rich spectrum of financial and tangible assets – financial assets like savings accounts, checking accounts, CDs, stocks and bonds, in later years money market accounts, IRAs, and Keoghs, and so on, along with tangible assets like investment real estate, businesses and farms, and vehicles.

Growing dissatisfaction with certain features of the SCF data, in particular with the fact that total financial asset holdings as measured by the survey were much lower than total financial asset holdings as measured in the Federal Reserve Board Flow of Funds accounts, Federal Reserve Board support was substantially reduced after 1960. The series was continued with less financial detail and broader sponsorship during the 1960s, and basically came to an end in its original form in 1969. The series was revived in 1983 with a sampling feature aimed at remedying the substantial underestimate of asset holdings shown by all of the previous surveys. This design feature was the addition of a high-income/high-wealth sample of households selected from statistical records derived from tax files. SCFs of this sort were conducted first in 1983, and have been scheduled on an every-third-year basis since 1989. (The last available data is for 2010.)

In addition to the detailed asset data, the SCF contains a comprehensive income section, has occasionally obtained substantial data on the pension holdings of respondents from the companies providing those pension plans, has a relatively standard set of household demographics and typically contains other variables of analytic interest – savings behavior, subjective health status, inheritances received, intended bequests, and so on. Furthermore, over the 1983–89 and 2007–09 periods, the survey collected panel data to examine how the effects of changes in the value of specific types of assets and debts and other economic disturbances played out at the household level (Bricker et al., 2001: p. 1). (For more information about the SCF, see their website: http://www.federalreserve.gov/econresdata/scf/scfindex.htm.)

Survey of Consumer Attitudes

SCA began in the United States in the late 1940s, originally as a 'soft' introduction to the questions about detailed asset holdings and income sources obtained in the SCF. These attitude surveys moved to an intermittent quarterly basis in the early 1950s, to a regular quarterly basis in the early 1960s, and to a regular monthly basis in 1978.

The basic content of these surveys includes a set of measures designed to reflect consumer assessments of their current financial situation compared with the past, their expectations about future financial conditions, their expectations about business conditions during the next year and the next 5 years, and their assessments of whether the present is a good or bad time to buy durable goods. Three of these core consumer attitude questions are part of the U.S. statistical system of leading economic indicators. The SCA contains substantially more data than the indicator series, including information about expected price change, perceptions of the effectiveness of current economic policy, assessments of buying conditions for houses and cars, expectations about changes in unemployment rates.

Data on consumer attitudes based on U.S. experience are now routinely collected in a number of other countries, including Austria, Australia, Belgium, Canada, China, Czech Republic, Denmark, Finland, France, Germany, Great Britain, Greece, Hungary, Ireland, Italy, Japan, Luxembourg, Norway, Poland, Russia, Spain, South Africa, Sweden, Switzerland, and Taiwan. (Information about the SCA can be obtained from their website: http://www.sca.isr.umich.edu/.)

National Longitudinal Surveys

The Bureau of Labor Statistics at the U.S. Department of Labor sponsors the NLS. The surveys began in 1966 with studies of the age cohorts of men aged 45–59 and of women in the age cohorts of 30–44. Young men and women aged 14–24 were added in the late 1960s, and young men and women in the age cohorts of 14–22 (called NLSY) were added in 1979. The children of the 1979 cohort were added in 1986. Of the six cohorts, data collection on four are continuing, while data collection on the remaining two (the original older male cohort aged 45–59, and the cohort of young men aged 14–24 in 1966) have been terminated.

The NLS surveys combine respondent interviews with a series of separately fielded administrative data collections, especially for the NLSY and the children of the NLSY cohorts. The information collected directly from NLS sample members includes information about work, family, educational experience, home ownership status, pensions, delinquency, and basic demographic information. The administrative data collection includes school characteristics obtained for the young women and young men cohorts and for the NLSY group, as well as for the cohort of children of the NLSY. In addition, school transcripts including coursework and attendance records were collected for the NLSY respondents in the early 1980s, and similar information is being obtained for children of the NLSY. Finally, aptitude and achievement scores from standardized tests were transcribed from school records in the late 1960s for the young men and young women cohorts, in the early 1980s for NLSY respondents, and during 1995 for the children of the NLSY cohort. Summary pension plan descriptions are collected for the mature women respondent and/or their spouses who report pension plan coverage, and death certificate information was collected in the early 1990s for most of the deceased members of the older men cohort. (More information about the NLS can be obtained from their website: http://www.bls.gov/nls/.)

Panel Study of Income Dynamics

The PSID, started in 1968, was based on combining a sample from the Survey of Economic Opportunity, originally conducted by the Department of Health and Human Services in 1966 and 1967, with a probability sample of individuals 18 years of age and older. The sample was heavily overweighed from the beginning with relatively poor households, since the SEO sample was designed to study poverty and had a disproportionate number of poor households.

The PSID has unique design features that have contributed to its status as being (probably) the most widely used micro data set in the world. A critical design feature is the way in which the PSID continues to be cross-sectional representative of the full U.S. population while maintaining the longitudinal characteristic that enables analysts to trace the dynamic of change over time. Representativeness for the U.S. population (except for new immigrants) is maintained by following PSID family members who leave their original sample households and begin their own households. This feature ensures a continued representation of newly formed households in the sample, accurately reflecting how the population changes with the birth of newly formed households. Thus PSID members are born, live their lives, and die just as individuals in the population do, and PSID traces the original sample individuals through their entire lifespan accompanied by whichever family they happen to be attached to – the original family in which they were a child, a new family when they formed their own household, and so on.

The PSID content is directed mainly at labor force experience and economic status, with a standard set of demographics. The study is especially rich in terms of data on jobs, job transitions, and detailed sources of income from work.

In recent years, the PSID has added significant modules on new topics. Health status, health insurance coverage, and functional health began to be collected in the early 1990s; wealth measures were obtained at 5-year intervals starting in 1984; and detailed pension and savings data were first collected in 1989. The wealth measures have been particularly successful. In 1997/1999, a sample of 511 immigrant families was added to enhance representativeness. In addition to the original PSID sample the Child Development Supplement studied the dynamic process of early life experiences in three waves (1997, 2002/2003, and 2007/2008). Finally, the PSID has had a major influence on microdata developments in other countries. Counterparts to the PSID can be found in the British Household Panel Survey (BHPS), the German Socio-Economic Panel (G-SOEP), and the European Household Panel Survey (EHPS). (For more information about the PSID, see their website: http://psidonline.isr.umich.edu/default.aspx.)

Survey of Income and Program Participation

The SIPP was begun in the early 1980s, with the basic objective of providing much richer income information for households at the lower end of the income distribution who might be eligible for various types of government public assistance programs. Thus the survey contains considerable detail on work experience, participation in a variety of public welfare programs, and on sources of income, along with family composition, a few health measures, and a standard set of demographics.

The SIPP basically has a cross-sectional design, although a number of observations are obtained for each respondent household to obtain very detailed information about income flows and work experience. Thus the SIPP is conducted at 4-month intervals, with respondents providing information about the previous 4-month period. Respondents stay in the sample for approximately six quarters, although in recent years the duration of stay has been extended so that a more genuinely longitudinal feature is available. In addition to income, work

experience, program participation, and some health items, the SIPP also has a wealth module that is administered several times to each participating household. In 1996, the SIPP Executive Committee established the Continuous Instrument Improvement Group (CIIG) to improve the SIPP core instrument. For example, selected outgoing rotation groups from the SIPP were followed longitudinally to observe the impact of the 1990s welfare reform on income, labor force participation, and poverty. Results of such new alternative instrumentation (1999–2003) were then compared and analyzed, and the final instruments were delivered for implementation in the 2004 Panel. Recently, examples of SIPP content include questions on medical expenses, child care, retirement and pension plan coverage, marital history, adult and child well-being, and others. The 2014 SIPP Panel uses the SIPP Event History Calendar (EHC) that assists the respondent's ability to recall events accurately over the longer reference period. The aim here is to increase data quality and intertopic consistency for dates reported by respondents. (More information about the SIPP can be obtained from their website: https://www.census.gov/programs-surveys/sipp.html.)

Consumer Expenditure Survey

The Consumer Expenditure Survey (CE), conducted by the BLS, has the principal objective of providing weights for the calculation of the Consumer Price Index. These studies have a long history: the first survey was conducted in 1888–91, the next in 1901 to get a better measure of price inflation in food, the third in 1917–19 to provide weights for a cost-of-living index, the next two during the depression of the 1930s to look at poverty issues, and the sixth, seventh, and eighth in 1950, the early 1960s, and early 1970s, respectively. The current version began in the early 1980s, when the survey was modified to be a continuous study with new samples appearing approximately every other year.

The survey goes into enormous detail on household expenditures across a long list of major product classifications (e.g., clothing, food, recreation, travel, insurance, cars, household durables) and provides estimates of consumer expenditures over past time periods for these categories. In addition, respondents are asked to keep a detailed diary, where they report each item purchased over a 2-week period. The diary information and the expenditure data for the product classifications (e.g., clothing for the respondent, for the spouse, for any children) are then integrated to produce an estimate of total consumer spending. In addition to consumption data, the CES includes detailed information about income, and also has some asset data along with the standard demographics. (More information about the CE can be obtained from their website: http://www.bls.gov/cex/.)

Time Use Survey

Time Use studies have a relatively lengthy history, and actually go back into the very early part of the twentieth century when studies of local communities in the United States, and of cities in the then Union of Soviet Socialist Republics, were conducted. As a major survey effort designed to provide national data on both market and nonmarket activities, the Time Use studies effectively started in 1965, when a cross-national study of urban time use was organized by a set of researchers under the direction of the Hungarian sociologist Alexander Szalai (1972). The countries included Belgium, Bulgaria, Czechoslovakia, France, Federal Republic of Germany, German Democratic Republic, Hungary, Peru, Poland, United States, Union of Soviet Socialist Republics, and Yugoslavia. Subsequently, time use studies in the United States with roughly comparable methodologies were conducted during 1975–76, and during 1981–82. Other time use studies in the United States were conducted in the late 1980s and early 1990s, although some design differences in these later studies make comparability with the earlier 1965, 1975–76, and 1981–82 studies difficult. Most European (and some Asian) countries conduct time use studies periodically, typically every 5 or 10 years, with a common survey design.

The Multinational Time Use Study (MTUS) was created to allow analysis of cross-time and cross-national changes in daily activities. The origins of the MTUS go back to an initiative of Professor Jonathan Gershuny in the early 1980s. The early MTUS mainly European data sets collected in the early 1960s through mid 1980s into a common series of background variables and total time spent per day in 41 activities. The MTUS has currently grown to encompass over 60 data sets from 25 countries, including recent data from the harmonized European time use surveys (HETUS), the American Time Use Survey (ATUS), and other national level time use projects.

One feature of the time use studies that differentiates it from most other microdatabase efforts is the difference in design between the United States and other countries. All studies that focus explicitly on time use collect data for a (usually retrospective) 24-h period – a time diary, as contrasted to a series of questions about the frequency of a set of activities. For the most part, these studies have been used as parts of national accounting systems, in which the major focus is on attempting to estimate the volume of nonmarket work, travel, leisure activities, and so on, as a supplement to the National Income and Product Accounts. Thus the major focus of the MTUS is on the macroeconomic uses of time diary data.

In contrast, the ATUS time diary data in the United States have been designed for use in a microeconomic as well as a macroeconomic framework. For that purpose, data for a single day consist mainly of statistical noise, and multiple time diaries for different days of the week and seasons of the year need to be collected for each sample member. Thus the U.S. design specified that data were to be collected for two weekdays, one Saturday and one Sunday, weighted to produce an estimate of time use during a typical week. In contrast, time diary studies in Europe typically (but not always) collect data for a single day for each respondent; in some recent studies multiple time diaries are collected for each respondent, often getting information for both a weekday and a weekend day in the same survey. (See more on ATUS at: http://www.bls.gov/tus/data.htm.)

Data archives on time use studies are maintained at the Centre for Time Use Research. (For more information, see their website: http://www.timeuse.org/home.)

Current Population Survey

The CPS, probably the most widely reported survey in the United States, if not in the world, is basically designed to produce an estimate of the unemployment rate for the nation as a whole, as well as for geographic subdivisions like states or SMSAs. The CPS, started in the early 1940s, is designed to produce estimates of work status during the prior week, including time spent looking for jobs as well as time spent working, for a very large (about 72 000 cases) monthly sample. The survey is a very simple and very brief instrument that collects virtually nothing else besides the policy-relevant variables – unemployment and work status for the adult population in the United States – from which the unemployment rate for the United States as a whole, or for particular states or SMSAs, can be calculated. The CPS does have additional survey content, since there are a series of annual supplements that include such variables as a detailed assessment of family income, assessments of health status for family members. But the core CPS is a very brief interview focused almost entirely on employment and job search experience. (For more information about the CPS, see their website: https://www.census.gov/cps/.)

Health and Retirement Study

The HRS, sponsored by the National Institute on Aging and the Social Security Administration, was begun in 1992 with the recognition that the U.S. statistical system had virtually no current data focused on the work situation, health status, family responsibilities, and pension status of the very large cohort of Baby Boom individuals (those born between 1946 and 1964) who would be starting to retire in the early twenty-first century. The behavior of this cohort was thought to be critically important for assessing the societal stresses associated with the prospective dramatic change in the proportion of the working population to the older dependent population.

The study, which collects information every other year, was originally focused on the U.S. population between the ages of 51 and 61 (those in the birth cohorts of 1931–41). This design would provide information with which to model the labor force participation decisions of individuals who were (mainly) not yet retired but who would retire shortly. In subsequent years, the survey was modified from one designed to follow an original cohort through its retirement experience, to a continuing study of the U.S. population of age 50 and older. It was designed to follow the retirement experience of successive cohorts of individuals starting with those born between 1931 and 1941 and continuing with those born up through the Baby Boom years. Thus the current study includes the original cohort of those born between 1931 and 1941 (and their spouses if married), and includes other cohorts added in subsequent years (those born before 1923, those born between 1924 and 1930, and most recently those born between 1942 and 1947).

The content areas of the HRS were subject to much more than the usual amount of developmental work. Survey content is decided by a set of researchers organized into working groups and headed by a Steering or Oversight Committee. The criteria that were uniformly adopted were that variables that played well-defined roles in modeling retirement decisions (or other critical decisions, such as saving rates) were eligible for inclusion, but other variables were not. Thus, the working groups pulled together sets of variables measuring job characteristics and work history; family structure and intrafamily transfers; health status and health insurance; economic status, including both income and wealth; several batteries of cognitive tests; a set of probability questions relating to continued work activity; future health status; longevity, bequests, and so on; and detailed information about pensions. The magnitude, duration, and cost of this planning effort are unique and give every indication of having produced a very high quality data set; it is certainly possible that the HRS planning model represents the future of microeconomic database design activities. (For more information about HRS, see their website: http://hrsonline.isr.umich.edu/.)

Luxembourg Income Study

The LIS represents one of the few attempts to collect comparable data on economic status across a large sample of countries. The basic idea is to collect national data on all income sources from as wide a range of countries as feasible, process the data so as to make it as comparable as possible across countries, and then make these data available to the research and policy communities for analysis of differences in economic status and income distribution. The core of the LIS activities is to process the data collected by national statistical agencies, to have a staff of trained analysts interact with their counterparts in these national agencies, and to develop measurements of income and income components that are as comparable as possible across countries.

The most important goal for LIS is harmonization – reshaping and reclassifying the components of income or definitions of household structure into comparable categories. Such harmonization allows the researcher to address important social issues without having to invest countless hours in getting every variable that will be analyzed into a comparable format.

Since its beginning, the LIS project has grown into a cooperative research project with a membership that includes countries in Europe, North America, the Far East, and Australia. The LIS database now contains information for more than 40 countries for one or more years, covering over 90 data sets over the period 1968 to 2010. The countries currently in the LIS database include Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Finland, France, Germany, Hungary, Ireland, Israel, Italy, Luxembourg, the Netherlands, Norway, Poland, Portugal, ROC Taiwan, Russia, Slovak Republic, Spain, Sweden, Switzerland, United Kingdom, and the United States. (For more information about LIS, see their website: http://www.lisdatacenter.org/.)

Living Standards Measurement Study

The World Bank established the LSMS in 1980 to explore ways of improving the type and quality of household data collected by government statistical offices in developing countries. The objectives of the LSMS were to develop new methods for monitoring progress in raising levels of living, to identify the consequences for households of current and proposed

government policies, and to improve communications between survey statisticians, analysts, and policy makers. To accomplish these objectives, LSMS activities have encompassed a range of tasks concerned with the design, implementation, and analysis of household surveys in developing countries. The main objective of LSMS surveys is to collect household data that can be used to assess household welfare, to understand household behavior, and to evaluate the effects of various policies on the living conditions of the population. Accordingly, LSMS surveys collect data on many dimensions of household well-being, including consumption, income, savings, employment, health, education, fertility, nutrition, housing, and migration. Three different kinds of questionnaires are normally used: the household questionnaire, which collects detailed information about household members; the community questionnaire, in which key community leaders and groups are asked about community infrastructure; and the price questionnaire, in which market vendors are asked about prices.

Because welfare is measured by consumption in most LSMS research on poverty, the measurement of consumption is strongly emphasized in the questionnaires. There are detailed questions on cash expenditures, on the value of food items grown at home or received as gifts, and on the ownership of housing and durable goods. A wide range of income information is also collected, including detailed questions about wages, bonuses, and various types of in-kind compensation.

LSMS-type surveys have been conducted in a long list of countries, including Peru, Ivory Coast, Ghana, Mauritania, Jamaica, Bolivia, Morocco, Pakistan, Venezuela, Russia, Guyana, Nicaragua, South Africa, Vietnam, Tanzania, Kyrgyzstan, Ecuador, Romania, and Bulgaria. For more information on the LSMS, see Grosh and Glewwe (1995). (See also the LSMS website at: http://go.worldbank.org/WKOXNZV3X0.)

Macro Databases

Standardized economic indicators at the macro level are central to mentor the economic activity of a country and to compare the outcomes internationally. While there is no known formula for describing economic activity, data can help policy-makers better understand their countries' economic situations and guide any work toward improvement. This part of the article covers the most common databases of economic activity, such as the National Accounts Main Aggregates Database, the OECD Main Economic Indicators (MEI) Database, and the World Economic Outlook (WEO) Database.

The National Accounts Main Aggregates Database

The National Accounts Main Aggregates Database presents a series of analytical national accounts tables for more than 200 countries and areas of the world from 1970 onward. It is the product of a global cooperation effort between the Economic Statistics Branch of the United Nations Statistics Division, international statistical agencies, and the national statistical services of these countries, first initiated by the Statistical Commission in 1947. The aim was to publish regularly the most recent available data on national accounts for as many countries and areas as possible. To enable a coherent, consistent, and integrated set of national accounts, internationally agreed recommendations were established, first stated in the so-called United Nations System of National Accounts 1993 (1993 SNA) and then in the updated version 2008 (2008 SNA). The SNA describes how to compile measures of economic activity accounts in the context of a set of internationally agreed concepts, definitions, classifications, and accounting rules. In addition, the SNA provides an overview of economic processes, recording, among others, how production is distributed among consumers, businesses, government, and foreign nations. More on the National Accounts Main Aggregates Database see online at: https://unstats.un.org/unsd/nationalaccount/madt.asp.

OECD System of Composite Leading Indicators

The OECD system of composite leading indicators (OECD CLIs) was developed in the 1970s to give early signals of turning points of economic activity. This information is of prime importance for economists, businesses, and policy-makers to enable timely analysis of the current and short-term economic situation. OECD CLIs are constructed to predict cycles in a reference series chosen as a proxy for economic activity.

The OECD Main Economic Indicators (MEI) database, which provides the main source for the selected indicators of the OECD CLIs, covers macroeconomic indicators for the following major subject areas: (1) GDP and its components and industrial production, (2) selected commodity output variables (crude steel, crude petroleum, etc.), (3) business and consumer tendency survey series, (4) selected manufacturing variables (deliveries, stocks, new orders, etc.), (5) Construction, (6) domestic trade, (7) labor market series, (8) consumer and producer prices, (9) money aggregates, (10) interest rates, (11) financial variables, (12) exchange rates, (13) international trade, and (14) balance of payments data. For more detailed information on the OECD CLIs see: http://unstats.un.org/unsd/EconStatKB/KnowledgebaseArticle10428.aspx.

OECD Statistics Portal (Including OECD Economic Outlook)

OECD also includes data and metadata for the current 34 OECD countries and selected nonmember economies in its OECD.Stat database and thus enables users to search for and extract data from across many databases under different themes. Among these databases, the OECD database on Economic Outlook allows the analysis of the major economic trends in the past from 1967 and over the coming 2–3 years. It covers the main economic issues and the policy measures required to foster growth in each member country. Forthcoming developments in major nonOECD economies are also evaluated in detail. It provides data on expenditures, foreign trade, output, labor markets, interest and exchange rates, balance of payments, and government debt. For the nonOECD regions, foreign trade and current account series are available. Sources for the historical data are publications of national statistical agencies and OECD statistical publications such as the Quarterly National Accounts, the Annual National Accounts, the Annual Labour Force Statistics, and the Main Economic Indicators. An overview of OECD.Stats' gathering and collation is online available at: http://stats.oecd.org/.

Eurostat's National Accounts

The Eurostat's section on 'National accounts' provides access to a selection of methodological and other background information regarding core national accounts aggregates such as Gross Domestic Product (GDP) and its main components. Data are compiled in accordance with the European System of Accounts 1995 (ESA 95), the EU manual for national accounts. Each Member States' data are first compiled by its own national accounts, through an institution appointed by its government, such as its National Statistical Office or its National Central Bank and then transmitted to Eurostat as agreed in the ESA95 transmission program (ESA95TP), which has legal status in the EU. In addition to the information on the Member States, limited information exists on EFTA countries, EU Candidate countries, the United States, and Japan. This information is obtained from the National statistical authorities. See more information on the Eurostat National Accounts online on their website: http://epp.eurostat.ec.europa.eu/portal/page/portal/national_accounts/data.

IMF's World Economic Outlook

The IMF publishes a range of time series data on IMF lending, exchange rates, and other economic and financial indicators. Manuals, guides, and other material on statistical practices at the IMF, in member countries, and of the statistical community at large are also available.

The WEO Databases and its Outlook Reports were established by the IMF to monitor economic development and policies in its member countries and in the global economic system. They were first created in 1970 and have been publishing since then during the biannual WEO exercise, which begins in January and June of each year and results in the WEO publication, which is released in May and October. The WEO database provides time series data for GDP growth, inflation, unemployment, payments balances, exports, imports, external debt, capital flows, and commodity prices. The World Economic Outlook Reports present IMF staff economists' analyses of global economic developments during the near and medium term. Chapters give an overview as well as more detailed analysis of the world economy; consider issues affecting industrial countries, developing countries, and economies in transition to market; and address topics of pressing current interest. The WEO exercise is coordinated by the World Economic Studies Division in the Research Department. The survey of prospects and policies is the product of a review of world economic developments, which draws primarily on the information IMF staff gathers through its consultations with member countries.

See also the WEO website at: http://www.imf.org/external/ns/cs.aspx?id=28.

World Bank Databases on Economic and Growth

The World Bank's data catalog is a listing of available World Bank data sets, including databases, preformatted tables, reports, and other resources aiming to provide all users with access to World Bank data. Regarding the topic 'economic and growth,' there are currently 26 reports found beginning from the World Development Indicators, which presents the most current and accurate global development data available, and includes national, regional, and global estimates, up to the Global Economic Monitor (GEM), which provides daily updates of global economic developments for exchange rates, equity markets, interest rates, stripped bond spreads, and emerging market bond indices. More information see the World Bank website at: http://datacatalog.worldbank.org/?Topics=Economy-and-Growth.

UNCTAD Statistics

The United Nations Conference on Trade and Development (UNCTAD) provides data on core themes via its online statistical database, UNCTADstat, in nine thematic areas: merchandise trade, trade in services, foreign direct investment, external financial resources, population and labor force, commodities, and maritime transport. Besides, the related annually World Investment Report contains many tables on global and regional, cross-border mergers, and acquisitions and transnational corporations. See more at their website at: http://unctadstat.unctad.org/ReportFolders/reportFolders.aspx?sCS_referer=&sCS_ChosenLang=en.

The Future of Economic Data

Developments at the turn of the century suggest that a continued expansion in the use of both longitudinal household data sets and internationally agreed standard economic data sets is likely to see, along with a vigorous growth in the use of establishment databases for micro- and macro-level analysis. These developments will be the result of three compelling forces: first, the increased ability of economists to understand economic behavior at micro-level and economic activity at macro-level based on the combination of richer theoretical insights and the greater availability of relevant data; second, the increased involvement of academic economists in the design of economic data sets; and third, referring to the micro databases, the development of more effective strategies for safeguarding privacy and confidentiality while still permitting researchers access to microdata files.

See also: Censuses: History and Methods; Data Bases and Statistical Systems: Demography; Data Bases and Statistical Systems: Economics (Macroeconomics); Data Bases and Statistical Systems: Sociology; East Asian Studies: Economics; Economic Education; Microeconomics, History of; Post-Socialism: Political Economy; Quantification in History.

Bibliography

Booth, C., 1891–1897. Life and Labours of the People in London. Macmillan, London and New York.

Bricker, J., 2001. Surveying the Aftermath of the Storm: Changes in Family Finances from 2007 to 2009. DIANE Publishing.

Davis, S.J., Haltiwanger, J., Schuh, S., 1996. Job Creation and Destruction. The MIT Press, Cambridge, MA.

Duncan, G.J., Hofferth, S., Stafford, F.P., 2004. Evolution and change in family income, wealth and health: the panel study of income dynamics, 1968–2000 and beyond. In: House, J.S., Kahn, R.L., Juster, F.T., Schuman, H., Singer, E. (Eds.), A Telescope on Society: Survey Research and Social Science in the 20th and 21st Centuries. University of Michigan Press, Ann Arbor, (Chapter 6).

Grosh, M.E., Glewwe, P., 1995. A Guide to Living Standards Measurement Study Surveys and Their Datasets. World Bank, Washington, DC.

Kuznets, S.S., 1941. National Income and Its Composition, 1919–1938. National Bureau of Economic Research, New York.

LePlay, F., 1855. Les Ou_ners Europeans, 6 Vols.

Manser, M.E., 1998. Existing labor market data: current and potential research uses. In: Haltiwanger, J., Manser, M.E., Topel, R. (Eds.), Labor Statistics Measurement Issues. University of Chicago Press, Chicago.

Rountree, B, Seebohm, 1902. Poverty: A Study of Town Life, second ed. Macmillan, London.

Smeeding, T.M., 1999. Problems of international availability of microdata: the experience of the Luxembourg income study (LIS) infrastructure project. In: Chlumsky, J.B., Schimpl-Neimanns, B., Wagner, G.G. (Eds.), Kooperation zwischen Wissenschaft und amtlicher Statistik—Praxis und Perspektiven. Metzler Poeschel Publishers, Wiesbaden, Germany, pp. 186–198.

Stafford, F.P., 1986. Forestalling the demise of empirical economics: the role of microdata in labor economics research. In: Ashenfelter, O.C., Layard, R. (Eds.), Handbook of Labor Economics, vol. 1. Elsevier, New York, pp. 387–423.

Szalai, A., 1972. The Use of Time. The Hague, Morton.

The health and retirement study, part I—history and overview. In: Woodbury, R. (Ed.), May 2000. Research Highlights, vol. 7.

Data Bases and Statistical Systems: Economics (Macroeconomics)

Thomas P Hill, University of East Anglia, Norwich, UK; and National Accounts and Economic Statistics, OECD, Paris

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Macroeconomic data are built up from the aggregate values, for groups of economic units, of the market transactions associated with economic activities such as production or consumption. All the aggregates are defined and classified consistently with each other using the internationally agreed System of National Accounts. Changes in the values of economic aggregates over time can be factored into their quantity and price components using appropriate aggregate index numbers. Similarly differences between countries in the values of economic aggregates expressed in their own currencies can be split into their price and quantity components using purchasing power parities.

Introduction

Macroeconomic data cover groups of economic units at different levels of aggregation up to and including the economy as a whole, i.e., the entire set of economic units resident in a country. Most macroeconomic data measure the aggregate values of sets of transactions associated with economic activities such as production, consumption, or financing. However, macroeconomic data may also refer to the values, at a point of time, of stocks of physical or financial assets owned by groups of units. The distinction between a stock, which can only be measured at a point of time, and a flow, which can only be measured per period of time, is of course fundamental.

Macroeconomic data are meant to provide a comprehensive picture of the economic activities taking place within an economy. They are used to monitor short- and long-term changes in the level of economic activity. They provide the information needed for purposes of fiscal and monetary policy by governments and central banks and also the data needed for purposes of economic analysis and forecasting by governments, universities, research institutes, or other agencies. For these purposes, it is necessary to measure how much of the change in an economic aggregate is due to changes in the prices of the goods and services concerned and how much to changes in their quantities.

The first part of this article is concerned with the System of National Accounts (SNA) in which the values of macroeconomic aggregates are measured at current prices. The second part is concerned with the associated price and quantity indices. Price and quantity indices are important macroeconomic statistics in their own right. For example, consumer price indices (CPIs), which measure changes in the prices of consumption goods and services purchased by households, have traditionally been used for index linking; that is, adjusting wage rates or social security benefits for the effects of inflation. They are also used as proxies for general measures of inflation for purposes of monetary policy.

Macroeconomic data are inevitably subject to error. Some data are based on surveys of businesses or households conducted by statistical office specifically for the purpose, but others are estimates that draw upon more than one data source, including data collected by government agencies for administrative purposes.

As macroeconomic data need to be as up to date as possible for economic forecasting and policy making, there is considerable pressure from users to obtain the data as quickly as possible. There is inevitably a trade-off between timeliness and reliability when the underlying raw data only become available piecemeal and gradually over a long period of time. The earlier the estimates are made and released, the greater the risk that they may have to be substantially revised later when more data become available. Minor revisions seem be acceptable to users as the price to be paid for receiving data on a timely basis, but major revisions can undermine the credibility and usefulness of the data.

The Main Aggregates

Gross Domestic Product and Gross Value Added

Suppose the objective is to measure the contribution of an enterprise to the total output produced in a country. Measuring the output of a single enterprise on its own is straightforward. Following standard accounting practice, the output of an enterprise is measured by the value of its sales plus the value of the change in its inventories. However, not all of the outputs produced by resident enterprises may be destined for 'final' use (i.e., final consumption by households or government or adding to the stock of fixed assets) because some may be sold to other enterprises for them to use up as inputs into their production. Products that are used up in the production of other products are described as 'intermediate.' In order to measure the contribution of each enterprise to the economy's final output, it is necessary to deduct from the value of its output the value of all the intermediate products consumed in the production process whether the products have been produced by other resident enterprises or by foreign enterprises. The difference is described as 'gross value added' (GVA).

Value added may be measured 'gross' or 'net' of depreciation, i.e., the decline in the value of the stock of fixed assets – machinery, equipment, and structures – used in the production process. However, in practice, because of lack of information about stocks of assets it is often not possible to estimate depreciation, and hence net value added, satisfactorily. GVA is a simpler concept and much more robust measure and it is widely used for this reason.

Gross domestic product (GDP) is defined simply and precisely as the sum of the gross values added of all the enterprises resident in an economy. GDP is probably the most important single macroeconomic aggregate being used throughout the world as a key indicator of the level of economic activity. As GDP does not include the value of outputs sold for intermediate uses, it follows that GDP is equal to the sum of all the outputs sold for final uses, i.e., the sum of household and government consumption expenditures plus gross capital formation plus exports *minus* imports. This means that GDP can be estimated from final expenditure data as well as from production data – one estimate providing a check on the other. It is measured quarterly as well as annually, and usually in real terms as well as at current prices.

From GDP to National Income

The link between measures of production and income is established by the fact that GVA provides the source of the funds out of which all the incomes generated in the course of production can be paid. These incomes are now described as 'primary' incomes although traditionally they used to be called 'factor' incomes. They consist of wages and salaries, rents on land, interest on loan capital, and gross operating profits. Gross profits are defined as the residual amount of value added remaining to the owner(s) of an enterprise after all the other incomes have been paid. The sum of all the primary incomes generated by production including operating profits therefore equals GVA by definition. At the level of the economy, the sum of the incomes generated by production is described as *gross domestic income*, or *GDI*, which is equal to GDP by definition. Net national income is obtained by deducting total deprecation from GDI, although as already noted satisfactory estimates of depreciation are seldom available so that the aggregates have to remain gross.

From National Income to National Accounts

Until the first half of the twentieth century, interest in macroeconomic data was mainly confined to estimating the sizes of a few aggregates at a national level, especially national income. The development of comprehensive macroeconomic data systems dates back to the 1930s and 1940s. It was stimulated by a combination of factors including the great depression and the Second World War. The depression and the switch from peacetime to wartime production both led to more active economic policy making and government intervention, but such policies also require more information. The relevant macroeconomic theory and models were also being developed alongside the macrodata systems in several countries, for example, the UK, the Netherlands, and the USA.

The macroeconomic data systems were described as national accounts. The publication of official estimates of national income aggregates for the USA started in the 1930s, but it was not until the Second World War that proper accounts began to be published. The first set of official national accounts for the UK was compiled by James Meade and Richard Stone in 1941 in the UK Treasury under the guidance of Keynes. Toward the end of the war, very close links were established between economists in the UK and the USA which culminated in the elaboration of the international systems of accounts described later. The development of these systems was a collective achievement by a number of economists from Europe and North America, but the pride of place is generally given to Richard Stone who was awarded the 1984 Nobel Prize in economics for his work in this field. Stone took the leading role in elaborating the international systems published under the auspices of the United Nations and the Organization for European Economic Cooperation or OEEC (later the OECD).

Stone focused on transactions between economic units as the elementary bits of information from which macroeconomic data systems have to be constructed. The objective of economic accounting is to organize, classify, and aggregate the countless transactions taking place within an economy in a way that is informative and useful for purposes of economic analysis and policy making. The challenge that this presents is of a different order of magnitude from that of deciding how to define and measure a single aggregate such as national income.

In 1953, the United Nations published its "*System of National Accounts and Supporting Tables*," or *SNA*, intended for use by all countries in the world, although the Soviet Union and other Socialist countries chose to use their own Marxist version of the system, which they eventually abandoned in the early 1990s. Three major revision updates and extensions of the UN *SNA* have been undertaken and published since 1953 in 1968, 1993, and 2008. The revisions completed in 1993 and 2008 were organized and conducted jointly by all the interested major international economic agencies – *UN*, *IMF*, *World Bank*, the *OECD*, and the *European Union (EU)* – using a group of national experts from around the world. Each revision was approved by the Statistical Commission of the United Nations which consists of a cross section of Heads of National Statistical Offices throughout the world. Almost all National Statistical offices have adopted the SNA as the model or standard for their own accounts.

The International System of Macroeconomic Accounts

The first paragraph of the most recent 2008 version of the International System of National Accounts (SNA) (European Commission et al., 2009; paragraph 1.1) reads as follows:

> The System of National Accounts (SNA) is the internationally agreed standard set of recommendations on how to compile measures of economic activity in accordance with strict accounting conventions based on economic principles. The recommendations are expressed in terms of a set of concepts, definitions, classifications and accounting rules that comprise the internationally agreed standard for measuring such items as gross domestic product (GDP) the most frequently quoted indicator of economic performance. The accounting framework of the SNA allows economic data to be compiled and presented in a format that is designed for purposes of economic analysis, decision taking and policy making. The accounts themselves present in a condensed way a great mass of detailed information, organized according to economic principles and perceptions, about the working of an economy. They provide a comprehensive and detailed record of the complex activities taking place within an economy and of the interaction between the different economic agents and groups of agents, that take place on markets or elsewhere.

In the 1960s, the Statistical Office of the European Communities decided to develop its own variant of the SNA "intended to meet the needs of the economic and social policies of the European Communities." The EU variant is called the European System of Integrated Economic Accounts or ESA. The ESA is meant to be fully consistent with the SNA. It has been revised and updated in parallel with the 1993 and 2008 revisions of the SNA. The EU has also continued to play an active role in the revisions of the SNA in the same way as the other international agencies referred to earlier.

The Sequence of Accounts in the SNA

The accounts of the SNA follow a logical sequence which may be summarized as follows. The first account is a production account that records the values of the various inputs used up and outputs produced. The second records the primary incomes generated in the course of production and their distribution among different economic units. The third records the subsequent redistribution of the incomes as a result of taxation and social insurance schemes. The fourth records the use of disposable income for purposes of final consumption by households and government. The next two accounts record the values of the financial and nonfinancial assets acquired and disposed of during the accounting period. The system is closed by balance sheets that record the values of the stocks of assets and liabilities held at the beginning and end of the accounting period.

The difference between the total uses and total resources recorded on the two sides of an account is described as the 'balancing item.' GVA, profits, disposable income, and saving are all examples of balancing items that can only be measured within an accounting framework. The balancing items are typically the most interesting and important items within the accounts as they encapsulate a great deal of information about the activities in question. As already explained, *GDP* itself is a balancing item equal to the difference between the values of total output and total intermediate inputs in a simplified production account for the total economy.

Real Economic Aggregates and Inflation

In the national accounts at current prices, aggregate flows of goods and services are recorded at the actual prices at which the relevant transactions take place. However, many users want the changes in the values of the aggregates to be factored into their price and quantity components. In the case of GDP, for example, most users are more interested in changes in the quantity component, commonly referred to as the 'real' change in the aggregate or alternatively as the change in the 'real' aggregate, than in changes in the aggregate valued at current prices.

Index Numbers

Changes in prices and quantities at an aggregate level are measured by means of index numbers. The most well-known index is probably the CPI. CPIs pre-date the development of national accounts having been compiled and published for well over a century in some countries. The CPI is designed to measure the average rate of change in the prices of the consumption goods and services purchased by households.

Somewhat surprisingly, until recent years there were no detailed internationally agreed guidelines about the economically most appropriate methods to use to calculate CPIs, but at the end of the 1990s the same group of international statistical agencies that took on the responsibility for revising and updating the SNA decided to collaborate on the production of a new set of international guidelines for CPIs. This initiative lead to the publication in 2004 of a major new reference work entitled *Consumer Price Index Manual: Theory and Practice* (ILO et al., 2004). The manual contains an authoritative, comprehensive, and up-to-date survey of the underlying economic and statistical theory as well as practical advice about how to construct a CPI.

The most common type of price index is one that measures the change in the total value of a specified basket of goods and services between two time periods as a result of changes in their prices, holding the quantities constant. One advantage of this kind of index is that, conceptually, it is easy to see exactly what it measures. The difficult question is to decide which set, or 'basket,' of goods and services to choose. The index that uses the basket of the earlier of the two periods is called a Laspeyres index (Laspeyres, 1871), while the index that uses the basket of the later period is called a Paasche index (Paasche, 1874).

It is easily shown that the Laspeyres index as just defined is equal to a weighted arithmetic average of the ratios of the individual prices in the two periods, using the expenditures in the first period as the weights. (The Paasche is a weighted harmonic average of the same price ratios that uses the expenditures in the second period as weights.) If the price ratios and the corresponding ratios of the quantities in the two periods are negatively correlated, the Laspeyres index will be greater than the Paasche. There are good economic reasons to expect such a negative correlation, as rational utility maximizing consumers will react to changes in relative prices by switching their consumption away from goods and services whose prices have gone up most and replacing them by substitutes whose prices have gone up the least. This is known as the 'substitution effect.' The Paasche index will therefore tend to give less weight than the Laspeyres to goods and services that have experienced the largest price increases. The greater the variability in the prices ratios, the stronger the substitution effect becomes and the greater the 'index number spread' between the Laspeyres and the Paasche indices.

As both the periods compared are equally important, in principle there may be no reason to prefer the Laspeyres to the Paasche or vice versa. In these circumstances, the obvious compromise is to take some kind of average of the two. Several such averages can be envisaged, but the one that has attracted the greatest support is a simple geometric average of the Laspeyres and Paasche, proposed by Irving Fisher (1922). Fisher showed that the resulting index is a well-behaved index that has many desirable statistical properties: see Chapter 16 (by Erwin Diewert) of the *CPI Manual* (2004). It is known as Fisher's 'Ideal Index.'

The Economic Approach to Index Numbers

The economic approach to index numbers seeks to ground them in economic theory. The prime example of an economic

index is the cost of living index (COLI), an index that draws upon the economic theory of consumer behavior (Konüs, 1924). A COLI is defined as the ratio of the minimum expenditures in two periods that would enable consumers to enjoy the same standard of living or level of utility in both periods, other things being equal. Other factors besides price changes, such as environmental or political factors, can affect welfare. For example, a natural disaster, such as a tsunami or simply a very severe winter, will reduce welfare and cause an increase in the overall cost of living but it does not increase a COLI which is meant to measure only the effects of price changes on welfare or utility. A COLI is conditional on nonprice factors remaining constant.

A COLI is not just another index number formula but a different concept of index from one that measures the ratio of the costs of purchasing a fixed basket of goods and services in two periods. It is commonly assumed that CPIs automatically measure changes in the cost of living, but in fact cost of living indices as understood in economics are conceptually quite different from indices based on fixed baskets of good and services.

Consider a COLI based on the level of utility in the first period. It measures the amount by which consumers' expenditures have to be increased, or decreased, between the two periods in order to enable consumers to achieve the same level of utility in the second period as they enjoyed in the first; in other words, to leave consumers feeling as 'well off' in the second period as they were in the first. This is not at all the same as the amount by which expenditures have to be increased in order to enable consumers to continue purchasing the same set of quantities of goods and services in the second period that they did in the first (which is what the Laspeyres price index measures). When confronted with a new set of relative prices, a utility maximizing consumer would not choose to continue to purchase the same set of quantities as before because a higher level of utility could be achieved by adjusting the quantities purchased in response to changes in their relative prices – the substitution effect again. It follows that the Laspeyres index may be expected to be greater than a COLI based on the first period. If therefore the official objective of the CPI is to estimate a COLI, as it is in some countries, the Laspeyres must be viewed as having an upward bias. Conversely, a Paasche index may be expected to have a downward bias as an estimate of the COLI based on the level of utility in the second period.

The problem with a COLI is that the quantities that would be purchased in the second period if the prices were those of the first period are hypothetical. They cannot be observed or used to calculate the index. The COLI is not an operational index that can be calculated directly from the actual prices and quantities observed in the two periods. The best that can be done is to try to approximate to a COLI using indices that can be calculated, such as Laspeyres and Paasche indices.

As just explained, the Laspeyres index places an upper bound on a COLI based on the level of utility in the first period while the Paasche places a lower bound on a COLI based in the second period. However, there are some important special cases that provide useful insights into the interrelationships between the various indexes. In particular, if consumers' preferences are 'homothetic' (that is, the indifference curves at different levels of utility are all uniform expansions or contractions of each other), the COLI is invariant to the level of utility and hence to the choice of base period. In this case, the COLIs based on the first and second periods coincide, and the Laspeyres and Paasche indexes provide upper and lower bound to that COLI.

If consumer preferences can be represented by one particular type of homothetic function (a homogeneous quadratic utility function), the Fisher index is equal to the COLI; in other words, the Fisher index would provide an exact measure of the COLI. Even though it is generally agreed that consumer preferences are most unlikely to conform exactly with this particular functional form, it may be conjectured that the Fisher index is often likely to provide a fairly close approximation to the COLI. The mere fact that it is possible for the Fisher to coincide with the COLI has had a considerable influence on the attitudes of index number users and compilers.

Erwin Diewert (1976) took the analysis a step further by showing that a Fisher index is not the only type of index that may be expected to approximate a COLI. He showed that there is a whole family of indices, which he described as 'superlative' indices, that may reasonably be expected to approximate to a COLI, at least for small changes in the cost of living. A characteristic feature of a superlative index is that it treats both periods symmetrically. Another example of a superlative index is the 'Törnqvist' index, which has been widely used for the measurement of real output in studies of productivity growth.

The intuitive appeal of a symmetric index together with the theoretical underpinning provided by the concept of a superlative index has resulted in a situation in which some form of superlative index is generally regarded by economists as the best choice of index. Which particular superlative index is selected is a secondary consideration as superlative indexes may be expected to approximate each other, given that they all approximate the same underlying COLI. The Fisher seems to be the most popular and is recommended in the 1993 and 2008 SNAs for decomposing changes in the values of macroeconomic aggregates at current prices into their price and quantity components.

GDP and Welfare

In recent years, there has been an increasing tendency to criticize the concept of GDP because changes in real GDP, essentially a quantity index of the economy's output, may not reflect changes in people's welfare. It is worth recalling exactly how GDP is defined and what it actually measures.

At a microlevel, GVA is defined as the balancing item in a simplified production account for an enterprise. It measures the value of the goods and services produced by an enterprise *less* the value of the goods and services produced by other enterprises that it consumes as intermediate inputs in the process: it is the value created by the production carried out within the enterprise. At a macrolevel, GDP is defined simply as the sum of the GVAs of all the enterprises resident in the economy. GDP therefore measures the total value of the outputs produced that are available for final use, i.e., final household and government consumption, gross investment, or

exports. Only part of GDP is destined for consumption by households so that movements in real consumption generally do not coincide with those in real GDP.

As its name suggests, GDP is a measure of output and hence of the level of production taking place in the economy. This information is extremely important in its own right for economic analysts and policy making. For example, levels of employment and unemployment depend directly on the level of real GDP.

As GDP measures enterprise production, there is no justification for using it to measure the welfare of households. The introductory chapters to both the 1993 and the 2008 SNAs make it quite explicit that GDP is designed to be a measure of production and not welfare.

It is important to remember that the purpose of compiling national accounts is not to estimate a single statistic, such as GDP, but to obtain a comprehensive database of interrelated and interdependent macroeconomic statistics that include data on income and consumption as well as production. The SNA is meant to be a multipurpose system that can meet a variety of different user needs. The particular set of statistics chosen by a user depends on the user's purpose.

In the present context, welfare is understood to refer to the standard of living, or more generally the level of well-being, experienced by persons or households. This depends mainly on the volume of goods and services that persons consume to satisfy their own personal needs and wants. If the objective is to track changes in household welfare, the relevant statistics are those for household real consumption.

Changes in real household consumption or preferably changes in per capita real household consumption may be expected to provide a good indicator of changes in household welfare. Given that consumption data are actually available within the national accounts database, choosing real GDP instead, i.e., data on enterprise production, as an indicator of welfare would be illogical.

As GDP may not perform well as an indicator of welfare, many users have reacted by arguing that this shows that the concept of GDP is faulty, whereas the fault actually lies with the users for making an inappropriate choice of welfare indicator.

In recent years, there has been a considerable upsurge in interest in the measurement of economic welfare which has stimulated a great deal of research. It has also lead to the creation of a small number of special high-level expert groups to advise on ways to develop new improved measures of welfare, notably the "*Commission on the Measurement of Economic Performance and Social Progress*" set up in 2008 by Nicholas Sarkozy when he was President of France. The Commission which included two prominent Nobel Prize winners, J.E. Stiglitz and A. Sen, issued its report in 2009.

There appears to be some consensus that series of real per capita household consumption provide a good starting point for measuring household welfare but that such measures are by no means sufficient on their own because they do not take account of all relevant factors, especially the distribution of income and wealth among households. It has always been recognized that the distribution of income and wealth has a major impact on social welfare, but macroeconomic databases, almost by definition, do not include the kind of detailed microdata needed. The development of new micro-databases to complement the macro-databases of the national accounts is likely to remain a high priority for further research.

Internationally Comparable Macroeconomic Data

Purchasing Power Parities and International Comparisons of Real Expenditures and Real GDP

The most important innovation and development in the field of macroeconomic statistics in the last few decades has been the estimation of international purchasing power parities (PPPs), which enable international comparisons to be made of real GDP and other major macroeconomic aggregates.

The first international comparisons were made between a few of the member countries of the OEEC by Milton Gilbert and Irving Kravis in the early 1950s. The Statistical Office of the European Communities started to make comparisons among its own member countries in the late 1960s and early 1970s and has continued to do so on a regular basis ever since. The United Nations embarked on a set of world wide comparisons for 10 countries in 1970 and then repeated them roughly every 5 years thereafter gradually extending the comparisons to more and more countries. The UN program, known as the *International Comparisons Project* or *ICP*, was carried out under the direction of Irving Kravis, Alan Heston, and Robert Summers of the University of Pennsylvania (Kravis et al., 1975). Responsibility for the ICP was taken over by the World Bank in the late 1990s with the support of the other international statistical agencies. The latest sets of world wide comparisons covered 141 countries in 2005 and 200 countries in 2011 (World Bank, 2012). The ICP seems to be the largest and most ambitious program in economic statistics ever carried out in the world. It requires substantial contributions from all the 200 participating statistical offices that collect and transmit all the requisite basic price and expenditure data to the ICP team at the World Bank. It inevitably takes some years to collect and process all the data and finalize the results.

Bilateral comparisons between pairs of countries have much in common with comparisons between two time periods for a single country with the additional complication that prices in different countries are usually denominated in different units of currency. The ratio of the prices in two different countries for a single unit of quantity of some specified product is described as the PPP for that product. It measures the rate at which a unit of currency for one country has to be converted into a unit of currency of the other country in order to ensure that it will purchase the same quantity of the product in both countries. PPPs vary from product to product when the patterns of relative prices in two countries are not the same so that most product PPPs must differ from the market exchange rate as determined on currency markets.

Some kind of average of the individual PPPs is needed in order to estimate the purchasing power for an economic aggregate. In a bilateral comparison, if one country is selected to act as the base country, a Laspeyres type aggregate PPP can be calculated as a weighted average of the individual product PPPs that uses the expenditure shares in the base country as weights. It shows the rate at which the currencies of the two countries would have to be exchanged in order to ensure that the total amount of currency needed in order to buy the basket of

products in the base country would, after being converted into the currency of the second country, be just sufficient to purchase the same basket of products in the second country. By now, there is overwhelming evidence from the international price comparisons that have already been carried out to show that aggregate PPPs tend to differ significantly and systematically from market exchange rates. Market exchange rates cannot be used as proxies for PPPs, whether at the level of an individual product or at higher levels of aggregation.

In international comparison projects, each country is asked to supply both price and expenditure data for the same list of products, making it perfectly feasible to calculate symmetric indices. The results show that Laspeyres PPPs tend systematically to exceed the corresponding Paasche PPPs, at least for consumption aggregates. The substitution effect is at work even more strongly in international comparisons than in intertemporal comparisons as consumers tend to buy relatively more of products in countries in which they are relatively cheap. The index number spread between the Laspeyres and Paasche PPPs can be quite large because patterns of relative prices can vary substantially between countries, not only as a result of differences in income levels but also differences in preferences due to differences in culture, religion, or climate. In these circumstances, symmetric indices, such as Fisher, become particularly attractive, especially if it is felt that equal weight should be given to each country as a matter of principle.

In the late 1970s, the Statistical Office of the European Communities decided to adopt a bilateral approach to multilateral comparisons in which the first step is to estimate the 'best' bilateral indices for every pair of countries in the group; Fisher PPPs were selected for the purpose. The problem that remains is that Fisher PPPs are not transitive, transitivity (or circularity) being generally regarded as an essential property for a set of multilateral PPPs. However, transitivity can easily be imposed on a set of bilateral PPPs by applying the so-called 'EKS' formula or method (*CPI Manual*, 2004: p. 500). The EKS method calculates the set of transitive PPPs that collectively deviate the least from the original bilateral PPPs. Eurostat has continued to use EKS-Fisher PPPs up to the present time.

There has always been controversy about the best method of calculating a set of transitive multilateral PPPs and their associated quantity indices. From the outset, the UN ICP has preferred to use the 'Geary-Khamis' (GK) method (*CPI Manual*, 2004: pp. 500, 501) in which the quantity comparisons are made by revaluing the quantities in different countries at a common set of average international prices. The international prices are weighted averages of the national average prices after they have been converted into a common currency using the aggregate PPPs. The PPPs and the international prices are derived simultaneously by solving a set of simultaneous equations (Kravis et al., 1975: pp. 67–70; *CPI Manual*, 2004: pp. 500, 501).

In the GK method, the group of countries as a whole, rather than an individual country, acts as the base for making the comparisons. The method can be interpreted as being equivalent to one in which a Laspeyres type quantity index is calculated for each country based on the group as a whole, the quantity indices between the individual countries being derived indirectly from their respective quantity indices based on the group (Hill, 1997). Deriving the indices indirectly is a valid procedure because Laspeyres type quantity indices that use a common set of prices are transitive. One advantage of the GK method is that it can be seen how the composition of the group of countries as a whole influences the price and quantity comparisons between individual member countries of the group. Another advantage is that as the aggregates in different countries are all valued at a common set of international prices, they can be summed directly across all the countries in the group to obtain the values of the same aggregates for the group of countries as a whole. National accounts aggregates, including GDP, can be assembled for the group as a whole.

The pros and cons of the two main approaches to international comparisons – the EKS-Fisher and the GK methods – continue to be debated in the literature. An EKS-Fisher PPP is a minimally adjusted superlative index which may be expected to provide an unbiased estimate of the PPP. On the other hand, the GK method does not treat each country equally because the national average prices of economically large countries carry more weight in the average international prices than the national prices of smaller and less developed countries. The pattern of average international prices is inevitably dominated by prices in the larger and richer countries, if it is to reflect the economic realities of the global economy. However, it can be argued that a bias is introduced into world wide comparisons because of the upward biases inherent in Laspeyres type indices. This bias is not uniform from country to country but tends to increase the more the pattern of prices in a country diverges from the pattern of average international prices. The upward bias is likely to be greatest for poor developing countries whose relative prices are likely to diverge the most from the relative international prices. If so, the GK method is likely to introduce a bias into comparisons between developed and developing countries which causes the gaps in living standards and per capita real GDPs between rich and poor countries to be underestimated. The systematic overestimation of the real GDPs of small developing countries relatively to large developed countries has been labeled the 'Gerschenkron effect.' It is viewed as a major problem by many economists as inequalities in per capita income and product at a world level are matters of serious economic and political concern.

The controversy over which of the two methods is better for most purposes is still not finally settled. The differences between the results obtained using the EKS-Fisher and GK methods illustrate the sensitivity of macroeconomic data to the methodology used to compile them. Notwithstanding concerns about the Gerschenkron effect, the ICP is a monumental achievement in the field of economic statistics that has created a major macroeconomic database that is unprecedented in its scope and the sophistication of its methodology.

See also: Economic Growth: Measurement; Income Distribution and Inequality: Measurement Issues; Macroeconomics, History of From 1933 to Present.

Bibliography

Commission of the European Communities, International Monetary Fund, Organisation for Economic Cooperation and Development, United Nations, World Bank, 1993. System of National Accounts 1993. Brussels/Luxembourg, New York, Paris, Washington, DC.

European Commission, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations, World Bank, 2009. System of National Accounts 2008. New York, United Nations.

Eurostat, OECD, 2004. Purchasing Power Parities and Real Expenditures, 2002 Benchmark Year. Eurostat, OECD, Luxembourg, Paris.

Fisher, Irving, 1922. The Making of Index Numbers. Houghton-Mifflin, Boston, MA.

Hill, Robert J., 1997. A taxonomy of multilateral methods for making international comparisons of quantities and prices. Review of Income and Wealth 43, 49–69.

International Labour Office, International Monetary Fund, Organisation for Economic Co-operation and Development, Statistical Office of the European Communities (Eurostat), United Nations, World Bank, 2004. Consumer Price Index Manual, Theory and Practice. International Labour Office, Geneva.

Konüs, Alexander, A., 1924. The problem of the true index of the cost of living. The Economic Bulletin of the Institute of Economic Conjuncture (in Russian): Reprinted in English in Econometrica (1939) 7, 10–29.

Kravis, Irving B., Zoltan, Kenessey, Heston, Alan, Summers, Robert, 1975. A System of International Comparisons of Gross Product and Purchasing Power. The Johns Hopkins University Press, Baltimore and London.

Laspeyres, E., 1871. Die Berechnung einer mittlerer Waarenpreisteigerung. In: Jahrbücher für Nationalökonomie and Statistik, vol. 16, pp. 296–314.

Paasche, H., 1874. Über die preisentwicklung der letzenJahre nach den Hamburger Borsennotirungen. In: Jahrbücher für Nationlökonomie and Statistik, vol. 12, pp. 156–178.

World Bank, 2012. International Comparisons Program, 2011. www.worldbank.org./data.

Data Bases and Statistical Systems: Economics (Statistical Systems)

Jacob Ryten, Federal Government of Canada, London, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article on economic statistics comprises two major subjects. It touches upon the structure of economic statistics and describes how they are related to the system of national accounts. It also examines the major developments brought about in these statistics through the use of government administrative records, principally those related to taxation. The major development in economic statistics, 50 years ago, consisted in replacing cumbersome economic censuses by sample surveys that could be administered more frequently and at less cost. Developments today, however, mostly brought in by massive computer applications may make it possible to return to the census approach by using for statistical purposes what is collected anyway, as the government discharges its legal and operational responsibilities.

'Economic statistics' is used in its narrow definition as statistics produced by official statistical agencies and derived from direct surveys or from government administrative records. Economic statistics concern households, businesses, government and nonprofit institutions, their reactions to changes in the economic environment, and interactions with each other. They are shown as values, prices, and quantities and deal with production, distribution, the use of factors of production, and the ways and circumstances at which they are bought and sold. Economic statistics are essential to compile the national accounts (qv). They are presented typically in terms of the activity of the agents to which they relate or in terms of the product produced, sold, or consumed. These presentations are embodied in international classifications used by most countries and developed at the request of the United Nations and other international and supranational agencies. Economic statistics are derived from either economic censuses or administrative registers. The latter are increasingly popular because their use for statistical purposes does not represent an added cost and the powers of the agencies that compile them tend to ensure better and more thorough coverage. Sample surveys of business have also become increasingly popular particularly when designed to measure the changes in effective demand as a basis for macro stabilization policies. Today's more advanced statistical offices show data relating to the activity of business, their production and balance sheet, the sales and purchases of goods and services to and from countries abroad, and the changes in the price level as well as the participation of households in the national labor force. The consumer price index (definition) and the rate of unemployment have become the two key indicators watched by government policy makers.

Nowadays it is customary to include under economic statistics, information on science and technology, innovation, and estimates of the value and quantity of stocks of natural assets and measurements of the state of the physical environment.

Definitions: The Different Kinds of Statistics

Economic Statistics and Official Statistics

In its strict sense the term economic statistics has a narrower scope, as a distinction is drawn between economic statistics and economic accounts (as in the national or international accounts). For purposes of the present article the term 'economic statistics' excludes the national accounts but includes the statistics that are simultaneously used for economic and social purposes. Some statistics occupy a gray area – between social and economic statistics. They comprise measurements that play a role in describing social conditions and are used intensively in the making of social policies but at the same time relate intimately to the economic fabric of society. Examples are the statistics on income distribution commonly used for the measurement of poverty, the occupational makeup of the labor force, the expenditure patterns of households, and the structure of their income. Most of the references in this article apply to official statistics, i.e., those statistics that are either produced by government statistical agencies or by other bodies but are legitimized by being used by government. The usage adopted corresponds to the way in which government statistical offices are commonly organized.

Economic Statistics and Business Statistics

The kernel of economic statistics is formed by statistics related to business and derived from questionnaires addressed to business, business employees, or from official records, which businesses create in their dealings with government. Such statistics include descriptions of the business activity (what does the business produce, how does it produce it, for how much does it sell, what is the surplus generated by the sale once materials and labor are paid for, etc.) as well as of the nature of the factors engaged in production (structures and equipment used in production by type; labor employed by type of skill and occupation), and in addition statistics on the type of customers who purchase goods and services from businesses.

Data-Gathering Methods

Directly Measured Statistics and Derived Statistics

The term economic statistics is usually applied to statistics based on direct measurement, i.e., either from surveys addressed to economic agents or official administrative procedures that yield numerical data capable of being transformed

 http://dx.doi.org/10.1016/B978-0-08-097086-8.41070-6

into statistics. For example, the statistic on the national output of cement is usually established by directly asking all (or a sample of) cement producers, how much cement they produced over a defined period. The statistic on how many jobs were filled during a given period may be taken from the social security forms that employers are requested to send to the social security administration from time to time.

This is different from the way in which the economic accounts of the nation, the so-called 'derived statistics,' are collected. These are formed by combining all the existing economic (or basic) statistics through an interrelated system of accounting identities and from those inferring identity values that may not be available directly. They are also a means of assessing the strengths and weaknesses of economic statistics by noting how close they come into balance when the system of identities is applied to them. In other words, economic statistics are the raw material on the basis of which the economic or national accounts are compiled. The compilation of the national accounts in addition to summarizing and systematizing economic statistics plays the important role of showing up gaps in the range of economic statistics required to describe the functioning of the economic system (Table 1).

How Economic Statistics Are Measured

In spite of the increasing popularity of using administrative records as sources to generate economic statistics, economic statistics are compiled mostly from surveys addressed to business. Generally speaking, the surveys consist of the following: a list of respondents, a formal questionnaire – which can be used as a basis for a personal interview (or an interview by telephone or the Internet), a procedure followed when dealing with exceptional cases (not finding the right person to whom the questions should be addressed, finding that the business is not in the expected activity, or that the answers are incomprehensible or contradictory) – and a set of steps designed to aggregate individual answers so as to produce statistical totals, averages, or measures of distribution.

Problems of Defining What Is a Business

Part of the process of conducting a business survey consists in locating the addressees – either all the addressees in the survey universe if the survey is of a census type or some of the addressees if it is a sample survey. In either case two issues must be settled beforehand: (1) that all the addressees conduct the activity they are supposed to and (2) that there is no one else who conducts the targeted activity and has been left out. A related question has to do with what is a business. The term 'business' can be defined as the entity that is liable under the law to pay taxes; that has a management capable of deciding its scale of operations and how those operations should be financed; and that inter alia conducts the activity that is germane to the survey. An alternative definition is one that seeks that part of the business (and only that part) that engages in the target activity directly and excludes all other parts of the business. In either case it is important that the respondents to the survey understand the scope of the questions and are capable of answering them out of records available to the business's management.

Business Records

Businesses tend to keep all the records they require to ensure informed management in addition to what they have to keep in order to comply with tax authorities and any other regulatory

Table 1 Types of economic statistics and sources (in italics) from which they are derived

Economic agent	*Production*	*Distribution*	*Resources employed*	*Prices*
Households	Household income/*surveys of household income*	Wages and salaries/*surveys of household income, labor force surveys*	Current population estimates, participation in labor force, employment/*labor force surveys*	Consumer prices and consumer price index/ *surveys of prices*
Business	Value and amount of production; production by type of activity and by goods and services produced, orders and deliveries/*censuses and surveys of business activity*	Value and amount of sales; stocks of finished goods, materials and goods in process; sales by activity of seller and by-product sold; sales by type of customer (exports and imports)/*censuses and surveys of business activity; customs records*	Numbers employed by occupation; remuneration of labor employed; capital assets used; depreciation; profits; loans and equity borrowing/*censuses and specialized business surveys; company balance sheets*	Producer prices; unit values in manufacturing; producer price index/ *surveys of industrial prices, surveys of service activities, current agricultural surveys, customs records*
Government (central and local, public sector corporations, etc.)	Government expenditures and revenues by function and by source/ *public sector accounts*	Public sector sales of goods and services/*public sector accounts*	Public sector employment; public sector wages and salaries/*public sector reports; budget and supporting accounts*	Prices subject to regulation/ *special price surveys*
Other (nonprofit organizations, labor unions, employer confederations, etc.)	Income and expenditure of nonprofit sector/*special surveys*		Employment in nonprofit sector; wages in nonprofit sector/*special surveys*	

bodies that affect them. Usually those records fall into five categories: tax related, social insurance, engineering records on the performance of physical production, personnel and safety records, and environmental records if the productive activity is of sufficient importance and of a kind that may detrimentally affect the physical environment. The efficiency of statistical surveys is measured partly by the ability of the statistical agency to request that information, but mostly by that information which is readily available from the records the business tends to keep anyway.

Household Surveys

Usually the major surveys addressed to households include a survey of households and how their members relate to the labor force and a survey of family income and expenditure. The former is the basis upon which statistics on the percentage of the labor force that is unemployed are calculated. The latter usually produces two key statistics: income distribution and structure of household expenditure. The structure of household expenditure is used to determine the various proportions of expenditure, which in turn are used in the compilation of the consumer price index. The surveys about household behavior – variation in the structure of expenditure and savings according to income, location of residence, family size, and age distribution – also provide key social statistics.

Public Accounts and Administrative Records

There are at least two categories of records that make explicit survey taking redundant: (1) reports on public accounts and (2) administrative registers created as a result of regular and comprehensive administrative procedures. The two have different properties. The accounts of the public sector (and of parastatal enterprises such as nationalized industries or public corporations) are usually made available to the public as a means of displaying transparency and accountability. If the statistical requirement is limited to the displaying of aggregate information, what is published is usually sufficient as a description of the income, expenditure, and balance sheet of all public sector entities. In addition, all operations relative to the collection of taxes generate comprehensive records, susceptible to replace statistics to the extent the administrative registers target the same variables. For example, the collection of income tax generates statistics on the income structure of all households (or individuals) liable to pay tax. The collection of VAT-type taxes generates records on the amount of sales, the nature of the seller, and in many cases the broad economic sector to which the buyer belongs. Social security records tend to generate information on the nature of the workers from whom amounts for social insurance are withheld; their place of work; the hourly, daily, or weekly remuneration that serves as a basis for the estimation of the amount withheld; and so on. And of course, the oldest and most universal administrative operation of all – that connected with the activity of customs administrations – generates comprehensive information on the value, composition, conveyance used, origin, and destination of merchandise that either enters or leaves the national territory. This information, also known as merchandise exports and import statistics, is to be found among the earliest collections of numerical data of relevance to the administration of a country. To the extent that these administrative records are accessible to the compilers of statistics they tend to constitute acceptable replacements for estimates derived from surveys or from censuses.

Economic Classifications and Conceptual Frameworks

Standard Economic Classifications

Economic statistics – like any other kind of descriptive statistics – are usually presented within the framework of standard classifications. In addition to geographic classifications, the most popular of these classifications are the Standard Industrial Classification (the SIC, or in its international version, the ISIC) and the Standard Product Classification (the SPC or, in its international version, the Standard International Trade Classification or currently the Central Product Classification). The SIC refers to the activity (or production technique) used to produce a specified list of goods and services. Similar (or homogeneous) production techniques are grouped into industries. The overall result of these groupings is a classification that views economic activity from the supply side. The SPC groups goods and services by their intrinsic characteristics rather than by the way in which they were produced. It is a classification that views the results of economic activity from the demand side. The use of international classifications is vital to ensure intercountry comparability of economic statistics.

The history of international cooperation in economic statistics is inextricably bound with the history of the two major economic classifications. Thus, the United Nations Statistics Commission, at its first meeting immediately after the end of World War II, commissioned the Secretariat to produce a classification of economic activities and shortly afterward requested the development of a classification of the goods that entered into international exchanges (imports and exports). The Committee on Statistics of the League of Nations had worked on an initial version of such a classification. Almost immediately after the war the Statistics Commission approved the International Standard Industrial Classification of Economic Activities (ISIC) and the Standard International Trade Classification (SITC). Today, both classifications are into their fifth or fourth version (ISIC Rev. 4 and SITC Rev. 3, respectively).

Currently, the economic statistics of the members of the European Union and of many central and eastern European countries are classified by the NACE (Nomenclature des Activités économiques dans les Communautés Européennes). The three North American countries have their own system of economic classifications (North American Industrial Classification System – NAICS). Most other countries use one or another of the various versions of the ISIC.

The classification of goods was taken over by the body created to coordinate the activities of customs administrations worldwide (the Customs Co-ordination Council or since 1994 the World Customs Organization) and ended up as the Harmonized Commodity Description and Coding System (HS). As such it has been adopted for international trade statistics by the overwhelming majority of countries. Indeed, a number of them use the same classification for purposes of presenting their

commodity output in the context of their industrial statistics. The requirement to develop an international classification of services prompted – again at the request of the United Nations Statistical Commission – the development of the Central Product Classification (CPC), which is being increasingly used to compare statistics on international trade in services with the internal production of and commerce in services.

Other Classifications

These apply to a variety of economic statistics but they (with the possible exception of the International Standard Classification of Occupations) are neither as well developed as the ones on activity and product nor are they at this stage the subject of international agreement. There are several examples of such classifications. In the case of capital assets there is still the question of whether a classification of assets ought to be embedded in the standard classifications of goods or whether it should be *sui generis*. There is no suitable classification of intangible assets. In the realm of labor inputs, recent innovations in the contractual relationship between employers and employees have produced a variety of forms of employment, none of which has so far found its way into an accepted classification.

Conceptual Frameworks

The conceptual framework for economic statistics is a hybrid, somewhere between business accounting and the system of national accounts. But, because economic statistics are the prime ingredient for the compilation of the national accounts, an explicit reconciliation between the former and the latter is required. Current trends suggest that because of pressures to keep government-generated paperwork down, economic statistics tend to follow the patterns of business records, where required, supplementary information is collected so that the basic statistics may be reconciled to the requirements of the national accounts compilers. But typically the necessary adjustments are conducted within the statistical agency. This is usually the case for key balances such as profits.

The Evolution of Economic Statistics as a System

The traditional manner in which economic statistics were derived was by the periodic taking of an economic census. Indeed in earlier times the census of population was taken together with the economic census. More recently, the economic census is addressed at businesses and retrieves information on their activity, location, legal form of organization, and on how many and what kind of people they employ. Supplementary information is usually obtained through lengthier questionnaires addressed not to all businesses but rather to a selection, usually above a certain size but sometimes to a randomly selected fraction of all those in the scope of the census.

Censuses and Registers

Economic censuses have become steadily less popular, largely due to their cost, the time it takes to process them, and the fact that, because of rapid changes in the size, makeup, and even activity of businesses, census information loses much of its value shortly after it is compiled (*see* Statistical Systems: Censuses of Population). Instead, there has been an equally steady increase in the popularity of registers of businesses, usually derived from administrative data and used for much the same purposes as censuses but updated far more frequently. Indeed, in a number of cases, the introduction of new administrative procedures connected with tax gathering has been managed bearing in mind possible requirements of importance to the government's statistical agency.

The lists of businesses derived from administrative or tax-gathering procedures, commonly referred to as 'business registers,' are used mostly in two modes: as sampling frames and as a basis for indicators of economic performance such as 'net business formation' (the difference between the numbers of businesses created and the number of businesses closed down in a given period of time). As sampling frames, business registers are invaluable to support ad hoc inquiries into the state of the economy, usually conducted by contacting a randomly chosen group of businesses and gauging the opinion of their chief executive officers.

There are several requirements that must be satisfied for the modern approach to the compilation of business statistics to work. In particular, the source of information for the register must be comprehensive; it must be kept up to date as a result of the administrative procedures on which it is based; and the statistical agency must have unrestricted access to the source. The benefits that accrue to a statistical office from a business register are that: (1) it can dispense with the task of conducting a periodic census; (2) it can conduct a coherent set of sample surveys and link the resulting information so as to form a composite picture of the economy and its performance; and (3) it can control the paper burden that it imposes on individual business respondents through its multiple surveys.

Applications of Business Registers

Business registers are maintained at two distinct levels. There is a list of all businesses through their registration with some government department (usually the tax authority). This list conveys no more than there is a possible business behind a tax account. The physical counterpart of the tax account is usually determined by direct contact once it is established that the business is more than just a legal entity. The relationships between businesses as legal entities and the establishments, mines, mills, warehouses, rail depots, etc., that comprise them are a source of interesting information, but they are also the most difficult part of the management of a business register.

Current Economic Indicators

Under the impulse of macroeconomic stabilization policies and the belief in policy makers' ability to fine tune effective demand, the early 1960s saw a worldwide push to develop current economic indicators of demand by sector. A great deal of the empirical work had been done in the United States, spearheaded by the National Bureau of Economic Research and resulting in a classification of short term economic statistics

according to whether they lead, coincided, or lagged behind the 'reference' economic cycle. Most attention was placed on the performance of manufacturing industry, particularly on the output of equipment for/of fixed investment as well as for export; on the construction of plant and of other nonresidential structures; and, in general, on leading indicators of capital formation. At the same time considerable effort was made to improve the reliability and the timeliness of the key variables for purposes of economic stabilization: the rate of unemployment and the change in the consumer price index. Lastly, the balances in the main markets of the economy were looked at through quarterly national accounts, which in turn demanded major developments in the quality and the quantity of basic economic statistics.

Today's Array of Economic Statistics

In today's array of economic statistics the following would be the expected components in the more advanced statistical agencies: (1) statistics on the structure of the economy derived from either an economic census or from administrative data complemented by a survey on structural data (type of activity, breakdown of production costs, capital stock etc.); (2) exports and imports of goods and services, with the goods part broken down by commodity, country of origin, and destination; (3) producer and consumer prices and indexes thereof; (4) short-term indicators of economic change such as order books, deliveries of manufacturing goods, and stocks thereof; (5) monthly physical output of key commodities and an index of industrial production; and (6) business balance sheets (income, profit and loss, financial) and employment and unemployment as well as household income and expenditure. In addition, a statistical version of the accounts of the public sector would be an expected component.

In order to support these statistics, statistical agencies would maintain an up-to-date version of a classification by economic activity, usually an internationally agreed classification with a national supplement, and a classification of goods and services. They would also have their own or an international version of the classification of occupational employment.

Comparability and the Role of Eurostat

The creation and subsequent development of a statistical agency within the offices of the European Union meant a big step forward in the international comparability of economic statistics. For reasons connected with the administration of European community matters, member countries demand that there be common standards, terminology, classification, quality levels, and above all a verified level of compliance by all member countries with each of these. As a result, the adoption by the Union of new classifications and standards has meant that member countries follow suit albeit with short lags between the acceptance of the standard and its implementation. More importantly, never has so much attention been paid to national compliance with international statistical standards. Moreover, other countries either because they are candidates for admission to the Union or else because they are linked to Union members by economic ties, have also shown willingness to conform to the Union's classifications of economic activities, goods and services, methods of compiling price indices, and other important components of the structure of economic statistics. It would seem that we are about to enter an era in which lack of international comparability is no longer a principal challenge. What is looming large is the need for a capacity to incorporate rapidly in one's system of economic statistics, new ways of engineering and delivering services, new business organizations, and new forms of articulating the demand in one country with the supply in its trading partner.

See also: Data Bases and Statistical Systems: Education, General; Data Bases and Statistical Systems: Labor and Work; Data Bases and Statistical Systems: Sociology; Economic Growth: Measurement; Economic History, Quantitative: United States; Quantification in the History of the Social Sciences; Statistics, History of.

Bibliography

Cox, B., et al., 1995. Business Survey Methods. Wiley, New York. Wiley Series in Probability and Mathematical Statistics.

ISI, Eurostat, BEA, 1996. Economic Statistics, Accuracy, Timeliness and Relevance. US Department of Commerce, Washington, DC.

Statistics Canada, 1997. North American Industry Classification System. Statistics Canada, Ottawa.

Turvey, R., 1989. Consumer Price Indices, an ILO Manual. International Labor Office, Geneva.

United Nations, 1990. International Standard Industrial Classification of All Economic Activities. Statistical Papers Series M No. 4 Rev. 3. United Nations, New York.

United Nations, 1991. A Model Survey of Computer Services. Statistical Papers Series M No. 81. United Nations, New York.

United Nations, 1993. System of National Accounts. United Nations, New York.

United Nations, 1997. Guidelines on Principles of a System of Price and Quantity Statistics. Statistical papers Series M No. 59. United Nations, New York.

Voorburg Group, 1997. Papers and Proceedings. United Nations Statistical Commission, New York.

Data Bases and Statistical Systems: Education, General

Andreas Schröer and Richard B Händel, Evangelische Hochschule Darmstadt, Darmstadt, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by L. Suter, volume 22, pp. 15062–15066, © 2001, Elsevier Ltd.

Abstract

Statistical systems in the field of education provide data on student learning outcomes, student enrollment, education expenditures, and educational equality. The article describes the scope, methods, and basic results of some of the most influential programs developed by the Organisation for Economic Co-operation and Development – OECD (such as Program for International Student Assessment – PISA) and the International Association for the Evaluation of Educational Achievement (such as Third International Mathematics and Science Study – TIMSS and Progress in International Reading Literacy Study – PIRLS), as well as education databases developed by the World Bank and UNESCO. Specific results on educational equality are presented alongside critiques of large-scale student outcome assessments from a methodological and governance perspective.

Introduction

Since the formation of the Programme for International Student Assessment (PISA) by the Organisation for Economic Co-operation and Development (OECD) in 1997, education statistics have gained enormous recognition in science, in the general public, and in the policy world. Different sources for data on education exist on the national level, such as censuses, administrative records, and ministry of education records. But doubts about the effectiveness of education systems in many Western countries, such as the United States in the 1980s (Gardner et al., 1983), have sparked an interest in international education statistics. International organizations have addressed the demand for comparable education data and indicators by developing specialized statistical systems in the field of education. These statistical systems provide information about learning outcomes such as student skills or competencies at different grade levels and school types (PISA, TIMSS, PIRLS). They also report on student demographics (enrollment), education expenditure (World Bank Education Expenditure Database), and education equality in different countries (e.g., Demographic and Health Surveys (USAID), Monitoring the Situation of Children and Women (UNICEF), Living Standards Measurement Study (The World Bank)).

Learning Outcome Statistics

The two most influential organizations involved in the provision of statistical systems on student achievement are the Organisation for Economic Co-operation and Development, founder of the Programme for International Student Assessment (PISA), and the International Association for the Evaluation of Educational Achievement (IEA), best known for its Third International Mathematics and Science Study (TIMSS).

Growing concerns about the effectiveness of educational systems in Western countries, as well as increasing awareness of economic competitiveness in a global market led to a strong interest in comparative education statistics. The OECD responded to this demand by initiating the Indicators of Education Statistics (INES) program in 1987. To improve the quality of international education statistics, countries such as the Netherlands, Scotland, the United States, and Sweden contributed resources and brought together working groups of education practitioners and scholars to develop indicators such as enrollment in education, finance, learning outcomes, labor force outcomes, school functioning, and attitudes toward education. The first report was published as Education at a Glance: OECD Indicators in 1992. The INES working group members quickly became aware of the lack of coherent data for indicators on student achievement, which led to the establishment of PISA. The OECD launched its first PISA report in 2001 with a periodicity of 3 years. PISA covers 34 OECD member countries and 31 partner countries (2012), representing more than 80% of the world economy. PISA 2012 is the program's fifth survey. It assessed the competencies of 15-year olds in reading, mathematics, and science (with a focus on mathematics) in 65 economies. In 44 of those economies, about 85 000 students also took part in an optional assessment of creative problem solving, which measures students' skills to respond to nonroutine situations needed to achieve their potential as constructive and reflective citizens.

Since PISA seeks to understand how well schools prepare students for challenges in the future, it analyses basic knowledge, skills, and attitudes that are important for young adults to take part in social life as responsible citizens. The first of the four identified basic skills is reading literacy, which was the main focus of the PISA 2000 study. The reading literacy test analyzes to what extent adolescents are able to understand and integrate texts they are confronted with in their everyday lives. Reading literacy is defined as "understanding, using and reflecting on written texts, in order to achieve one's goals, to develop one's knowledge and to participate in society" (OECD, 2003: 108). Mathematical literacy is defined as "an individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgements and to use and engage with mathematics in ways that meet the needs of the individual's life as a constructive, concerned and reflective citizen" (OECD, 2003: 24). "Scientific literacy is the capacity to use scientific knowledge, to identify questions and to draw evidence-based conclusions in order to understand and

help make decisions about the natural world and the changes made to it through human activity" (OECD, 2003: 133). And finally problem solving is defined as "an individual's capacity to use cognitive processes to confront and resolve real, cross-disciplinary situations where the solution path is not immediately obvious and where the literacy domains or curricular areas that might be applicable are not within a single domain of mathematics, science or reading" (OECD, 2003: 156).

Around 510 000 students between the ages of 15 years 3 months and 16 years 2 months participated in PISA 2012, representing about 28 million students globally. Students took a paper-based test that lasted 2 hours. The tests were a mixture of open-ended and multiple-choice questions that were organized in groups based on a passage setting out a real-life situation. School principals also answered questionnaires to provide information about the student and teacher body (grade levels, student enrollment, staff size), the school's resources (location, public or private, funding), the school's instruction, curriculum, and assessment (ability grouping, extracurricular activities, forms of student performance assessment), the school climate, the school policies and practices (assessment data, monitoring practices), and the characteristics of the principal (gender). Although most public attention was paid to student performance data, the study also provided data on instructional settings (student perception of class climate and teacher support), on educational service providers (school resources, admittance, and grouping), as well as system-level aggregated data on the educational system as a whole, which were based on the context questionnaires.

According to the results of the latest PISA in 2012, China has the highest scores in mathematics, with a mean score of 613 points, which is 119 points, or the equivalent of nearly 3 years of schooling, above the OECD average. Singapore, Hong Kong-China, Chinese Taipei, Korea, Macao-China, Japan, Liechtenstein, Switzerland, and the Netherlands, in descending order of their scores, round out the top ten performers in mathematics. Of the 64 countries and economies with trend data between 2003 and 2012, 25 improved student performance in mathematics. Boys perform better than girls in mathematics in only 38 out of the 65 countries that participated in PISA 2012, and girls outperform boys in five countries. Across OECD countries, 13% of students are top performers in math.

Shanghai-China, Hong Kong-China, Singapore, Japan, and Korea are the five highest performing countries and economies in reading. Of the 64 countries with comparable data throughout their participation in PISA, 32 improved student performance in reading. Across OECD countries, 8% of students are top performers in reading. These students can handle texts that are unfamiliar in either form or content and can conduct fine-grained analyses of texts.

Shanghai-China, Hong Kong-China, Singapore, Japan, and Finland are the top five performers in science. Between 2006 and 2012, Italy, Poland, and Qatar, and between 2009 and 2012, Estonia, Israel, and Singapore increased their shares of top performers and simultaneously reduced their shares of low performers in science. Across OECD countries, 8% of students are top performers in science. These students can identify, explain, and apply scientific knowledge in a variety of complex life situations.

Students in Singapore and Korea, followed by students in Japan, score higher in problem solving than students in all other participating countries and economies. Boys outperform girls in problem solving in 23 countries/economies, girls outperform boys in five countries, and in 16 countries, there is no significant difference in average performance between boys and girls. The impact of socioeconomic status on problem-solving performance is weaker than it is on performance in mathematics, reading, or science. Across OECD countries, 11.4% of 15-year-old students are top performers in problem solving.

The IEA's TIMSS study has been carried out every 4 years since 1995. Its area of coverage in 2011 included 63 countries, plus 14 benchmarking entities (regional jurisdictions such as Abu Dhabi or Quebec). The TIMSS countries vary greatly in terms of economic development, geographical location, and population size. TIMSS provides an international assessment of mathematics and science at the fourth and eighth grades. The content domains (for example, algebra, geometry, etc. in mathematics, and biology, chemistry, etc. in science) and the topic areas within the domains are described separately for the fourth and eighth grades in the TIMSS 2011 Mathematics Framework and the TIMSS 2011 Science Framework. TIMSS explains its focus on mathematics by stating that "effectiveness as a citizen and success in a workplace are greatly enhanced by knowing and, more important, being able to use mathematics" (Mullis et al., 2012b: 20). Similarly the authors point out that "in today's world, some understanding of science is imperative if citizens are to make informed decisions about themselves and the world in which they live" (Mullis et al., 2012b: 49). Nationally representative samples of students participated in TIMSS. Countries and benchmarking participants could elect to participate in the fourth grade assessment, the eighth grade assessment, or both: Fifty-two countries and seven benchmarking entities participated in the fourth grade assessment, and 45 countries and 14 benchmarking entities participated in the eighth grade assessment. Several of the countries, where fourth and eighth grade students were expected to find the TIMSS assessments too difficult, administered the fourth and eighth grade assessments to their sixth and ninth grade students. In total, more than 600 000 students participated in the 2011 TIMSS.

The TIMSS 2011 mathematics framework is organized around two dimensions: (1) a content dimension specifying the domains or subject matter to be assessed within mathematics (for example, number, algebra, geometry, and data and chance at the eighth grade) and (2) a cognitive dimension specifying the domains or thinking processes to be assessed (that is, knowing, applying, and reasoning). The cognitive domains describe the sets of behaviors expected of students as they engage with the mathematical content (Mullis et al., 2012b: 19).

Fourth graders from East Asian countries continue to lead the world in mathematics achievement. Singapore, Korea, and Hong Kong SAR, followed by Chinese Taipei and Japan, Northern Ireland, Belgium (Flemish), Finland, England, and the Russian Federation were the top 10 performing countries at the fourth grade level.

At the eighth grade level, Korea, Singapore, and Chinese Taipei outperformed all other countries, followed by Hong Kong SAR, Japan, the Russian Federation, Israel, Finland, the United States, and England and were the top 10 performing countries. While small differences appeared from country to

country, there was a substantial range in performance from the top performing to the lower performing countries. The 22 countries assessing fourth grade students and the three assessing sixth grade students displayed average achievement levels below the TIMSS scale centerpoint of 500, as did two benchmarking participants. At the eighth grade level, 27 countries and the three assessing their ninth grade students had average achievement below 500, as did three benchmarking participants.

IEA and the TIMSS and PIRLS International Study Center at Boston College launched another international program, the Progress in International Reading Literacy Study (PIRLS). PIRLS reports on primary school students' literacy achievement in 49 countries every 5 years. Reading literacy is defined as "the ability to understand and use those written language forms required by society and/or valued by the individual. Young readers can construct meaning from a variety of texts" (Mullis et al., 2003: 33). In 2011, PIRLS also examined national policies and practices related to literacy, and included a set of questionnaires for students, parents/caregivers, teachers, and school principals to investigate the experiences that young children have at home and school while learning to read. In 2011 for the first time, pre-PIRLS offered assessment of basic reading skills, working as a bridge to PIRLS. Especially for countries where most children are still developing basic reading skills at the end of the primary school cycle. PIRLS targets fourth grade students. In 2011, 49 countries and 9 benchmarking entities participated in PIRLS and pre-PIRLS. Approximately 325 000 students participated in the 2011 cycle.

Recent studies based on TIMSS and PIRLS data from 34 participating countries in 2011 show positive correlations among reading, mathematics, and science achievement at the fourth grade. More than 50% of the 34 participating countries succeeded in educating 90% or more of the fourth grade students to a basic level of proficiency in all three subjects. The study also demonstrated educational practices of effective schools, teachers, and parents to ensure student achievement (Martin and Mullis, 2013). The TIMSS and PIRLS model of educational practices in effective schools includes being safe and orderly, supporting academic success, having adequate facilities and equipment, being well-staffed with well-prepared teachers, and providing well-resourced classrooms and effective instruction (Martin and Mullis, 2013: 113).

The third international statistical system provided by IEA is the International Civic and Citizenship Education Study 2009, which is the third IEA study (after 1971 and 1999) on civic and citizenship education. The study reported on the achievement of students enrolled in the eighth grade in a test of knowledge and conceptual understanding, along with student dispositions and attitudes relating to civics and citizenship (Schulz et al., 2010). Teacher and school questionnaires collected information about the contexts in which students learn about civics and citizenship, including teaching and classroom management practices, and school governance and climate. Earlier studies (Torney-Purta et al., 2001) highlighted that students who participated in governance and decision-making of their schools score higher in civic knowledge and engagement scales. Therefore the degree of student and teacher involvement in decision-making processes and practices about the way classes are taught, what is taught in classes, teaching and learning materials, timetable, classroom and school rules at their respective schools were part of the questionnaire. National context surveys asked about the provision of civic and citizenship education in each participating country. The data collection took place in 38 countries between 2008 and 2009. The studies provided a concept of civic education, which was not limited to the basic knowledge about the inner workings of political systems, but was aiming at educating citizens which are able to participate in democratic societies. IEA therefore described civic education as follows: "According to the information collected from the ICCS national research centers, civic and citizenship education covered a wide range of topics across the participating countries. These encompassed knowledge and understanding of political institutions and concepts, such as human rights, as well as newer topics covering social and community cohesion, diversity, the environment, communications, and global society. Most teachers and school principals regarded the development of knowledge and skills as the most important aim of civic and citizenship education" (Schulz et al., 2010: 251).

Criticism

Since its formation, the PISA program has sparked international controversy among scholars and policy makers about the use of standardized learning outcome data for evaluating the quality of schools. Critics raise methodological questions on the study's lack of longitudinal data (Fertig, 2003), its lack of comparability (Blum et al., 2001), the nature of the test questions (Prais, 2003; Bonnet, 2002), issues of translating standardized questions in different cultural contexts and languages, and in more general terms a decontextualization of student outcome data in an attempt to compare educational systems worldwide (Wuttke, 2006; Spiegelhalter, 2013). Age differences of up to 2 years are problematic for international data comparability in all grade level studies. The representativeness of national samples and test-wiseness of pupils in different countries can also be highly problematic in biasing differences across countries (Wuttke, 2006). General competency tests in student assessment surveys have strong similarities with intelligence tests. In order to get at cultural competencies, the test would need to include more knowledge-based items (Rindermann, 2007).

Others criticize the policy implications of the study, such as its 3-year cycle, which has shifted policy makers' attention to short-term fixes (Andrews et al., 2014) instead of long-term improvement of national education systems. The narrow focus on measurable education outcomes also undermines the importance of less measurable aspects of education, such as moral, physical, civic, and artistic impacts (Ertl, 2006), which are related to the controversial debate on outcome-centered governance in education (Soquel and Jaccard, 2008). In terms of data validity, scholars criticize the isolation of student assessment studies from other related research such as empirical studies on the intelligence construct or economic human capital research (Rindermann, 2007). Finally scholars find that research on causes for different national competency levels is still in its infancy. There are likely to be complex reciprocal influences between educational, genetic, economic, political, and cultural factors (Rindermann, 2007).

Education Equality, Education Expenditures, and Student Demographics

The big education statistical systems – TIMMS, PIRLS, and PISA – as well as big international surveys by USAID, UNICEF, and The World Bank also provide Education Equality Data. Since 1984, USAID's demographic and health surveys (DHS) have conducted nationally representative household surveys (in over 80 countries) to provide data for a wide range of monitoring and impact evaluation indicators in the areas of population, health, and nutrition, but also on the situation of women and gender differentials in the area of education. Standard DHS sample up to 30 000 households and are conducted about every 5 years. Interim DHS are conducted between rounds of the standard surveys and provide additional, representative data. While clearly focusing on health topics, all demographic and health surveys include DHS indicators about gender differences in literacy, educational attainment, and education based on the idea that education is the key to empowering women.

The second important source of education equality data is the Monitoring the Situation of Children and Women Survey (UNICEF). In 1995, UNICEF began supporting countries for the collection of statistically sound and internationally comparable data through Multiple Indicator Cluster Surveys (MICS). Data are gathered from two sources: enrollment data from administrative records, and attendance data from household surveys. After four rounds of MICS, 240 surveys have been conducted in more than 100 countries. The surveys consist of a household questionnaire, a questionnaire for women aged 15–49, a questionnaire for individual men aged 15–49, and a questionnaire for children under the age of five (administered to the mother or caretaker). Together, they help measure key indicators on households, and life circumstances of women, men, and children. Seen as vital to meeting the Millennium Development Goals, UNICEF collected data about universal primary education, gender disparities at all levels of education, and the empowerment of women. In addition to that, UNICEF gathered administrative data on school participation, youth and adult literacy rates, survey data on primary net attendance rate, primary survival rate, and secondary net attendance rate.

In 1980 the World Bank's Development Research Group (DECRG) established the Living Standards Measurement Study (LSMS) exploring ways of improving the type and quality of household data collected by statistical offices in developing countries. LSMS includes data sets from 37 countries (from Albania to Vietnam), the earliest surveys dating back to 1985. LSMS aimed to help policy makers move beyond simply measuring rates of unemployment, poverty, and healthcare use, by assisting them in their efforts to identify how they could design and improve policies that positively affect outcomes in health, education, economic activities, housing, and utilities. LSMS surveys intend to produce a comprehensive monetary measure of welfare and its distribution, describe patterns of access to and use of social surveys, and allow the study of how households behave in response to changes in the economic environments or government programs. Therefore the LSMS has created a database of harmonized variables from LSMS surveys to facilitate comparative research. LSMS covers various topics. The household questionnaire includes measures of consumption, measures of income, and a variety of sectoral issues, such as health, education, nutrition, and fertility. The community questions describe the economic environment faced by the households in the sample. The prices questionnaires gather information on prices of basic goods in the communities. In some countries special questionnaires are used for schools (Grosh and Munoz, 1996). The LSMS website lists research reports and publications, including LSMS working papers, based on the collected data sets.

Social Background, Gender, and Education Equality

PISA data across OECD countries suggest effects of socioeconomic situation of students on their learning outcome. A student's socioeconomic status is estimated by the PISA index of social, cultural, and economic status, which is based on indicators such as parental education and occupation, the number and type of home possessions that are considered proxies for wealth, and the educational resources available at home. The PISA consortium claims that across OECD countries, 14.8% of differences in performance among students are explained by disparities in students' socioeconomic status, whether the focus is on mathematics, reading, or science. PISA data suggest that a more socioeconomically advantaged student scores 39 points higher in mathematics – the equivalent of nearly 1 year of schooling – than a less-advantaged student (OECD, 2014b: 34). On the other hand, some 6% of students across OECD countries appear 'resilient' to socioeconomic disadvantages and exceed expectations.

Integrating immigrant students into the education system is a challenge for most countries; however a country's success in integrating immigrants' children into society is a key indicator of the efficacy of education policy. In OECD countries, immigration status is still linked to educational disadvantage. While the share of immigrant students in OECD countries increased from 9% in 2003 to 11% in 2012, their performance disadvantage as compared to students without an immigrant background but with similar socioeconomic status shrank by 11 score points during the same period, which indicates a decreasing level of disadvantages for students with immigration status. Overall 'immigrant students' performance in PISA is more strongly (and negatively) associated with the concentration of socioeconomic disadvantage in schools than with the concentration of immigrants per se or the concentration of students who speak a different language at home than the one in which they are taught at school. Reducing the concentration of disadvantage in schools may require changes in other social policy, such as housing or welfare, to encourage a more balanced social mix in schools (OECD, 2014b: 71).

PIRLS data show significant gender differences in reading competencies with the exceptions of Luxemburg and Spain. All participating countries show higher reading competencies for boys (on average a 17 point difference), which holds for both analyzed text categories (literature and information). The gap between girls and boys is higher for upper level reading skills such as interpret, connect, and evaluate than for base level skills such as repetition and simple conclusion (Mullis et al., 2012a). In most countries and economies, boys outperform girls in

mathematics; and among the highest achieving students, the gender gap in favor of boys is even wider.

Education Expenditures and Student Demographics

The World Bank Education Expenditure Database and the UNESCO Institute for Statistics provide data on education expenditures as a percentage of total government spending, as a percentage of GDP and government expenditure per student, and by level of education. In 2002, the World Bank Education Expenditure Database started collecting data from public expenditure documents across 97 countries, using 891 indicators. However, such data are not internationally comparable since there is no standard method of calculation for most of the indicators. As a result, cross-country comparisons are not advisable.

The UNESCO Institute for Statistics also publishes a Global Education Digest, which includes public expenditure data on education based on three questionnaires (UIS, UOE, and WEI), from 204 countries and territories. The main issues regarding student demographics pointed out by the Global Education Digest 2012 are high numbers of about 32 million primary pupils repeating grades, and 31 million students completely dropping out of school. These numbers are alarming because UNESCO asserts that once pupils drop out of school, most never return. Eighty percent of those children live in only three world regions. Sub-Saharan Africa with four, South and West Asia with three, and Latin America and the Caribbean with two dropouts for every 10 children who start school. While girls are less likely to start school in the beginning, boys more often tend to leave school before even completing primary education, and more often repeat grades. The overall gap between boys and girls remains wide. In 2011, an estimated 31 million girls of primary school age and 34 million girls of lower secondary school age were not enrolled in school at all. In addition, pupils from rural households are much more likely to repeat a grade.

Significant regional differences are also demonstrated for reading literacy: the "SACMEQ (2007) assessment conducted in 14 sub-Saharan African countries show that in about one-half of countries pupils at the end of primary school had low performance levels in reading. In Malawl and Zambia. Almost three in four tested pupils were unable to meet minimum requirements for reading. […] A range of data sources show that children with the least opportunities – arising from poverty and compounding disadvantages – are most likely to repeat grades and leave school early without being able to read a basic text." (UNESCO, 2012: 47).

PISA data indicate how disparities in exposure to formal mathematics are more marked in school systems that separate students into different schools based on their performance. The strong relationship between performance and socioeconomic status was again demonstrated in systems where the unintended result of separation by performance is separation by socioeconomic status. PISA data also emphasize the importance of early childhood education, because many inequities that exist within school systems are already present when students enter formal schooling. Earlier entrance into the school system might prepare students better to succeed in formal schooling and therefore might help to prevent some inequities. PISA claims that across OECD countries, students who reported that they had attended preprimary school for more than 1 year score 53 points higher in mathematics (equivalent of more than 1 year of schooling) before accounting for socioeconomic status, and 31 points after accounting for socioeconomic status, than students who had not attended preprimary education (OECD, 2014b: 99). These data suggest that early attendance in preprimary education is positively correlated with socioeconomic status; however it also suggests a positive correlation between preprimary school attendance and performance at age 15 or 16, which is independent of socioeconomic status (OECD, 2014b: 99).

Both PISA and TIMSS data suggest that teacher shortage is an issue in socioeconomically disadvantaged schools. OECD countries allocate at least an equal, if not a larger, number of teachers to socioeconomically disadvantaged schools as to advantaged schools; but disadvantaged schools tend to have great difficulty in attracting qualified teachers (OECD, 2014b: 14).

TIMSS/PIRLS data show that successful schools tend to be well resourced. Ever since the Coleman report in 1966, researchers have recognized that the compositional characteristics of a school's student body can affect student achievement. To provide information on this topic, TIMSS routinely asks school principals to report on their students' economic home backgrounds and home language. While there was variation across countries, higher average mathematics achievement was associated with students attending schools where a greater percentage of students were from relatively affluent socioeconomic backgrounds and spoke the language of the TIMSS assessment as their first language (Mullis et al., 2012b: 205).

Overall successful schools also are likely to have better working conditions and facilities as well as more instructional materials, such as books, computers, technological support, and supplies. TIMSS 2011 created the Mathematics Resource Shortages scale based on principals' responses concerning inadequacies in general school resources (materials, supplies, heating/cooling/lighting, buildings, space, and staff) as well as resources specifically targeted to support mathematics instruction (specialized teachers, computers, computer software, calculators, library materials, and audio-visual resources). Many countries were fortunate to have very few students in schools where instruction was affected a lot by resource shortages. However, this was a crucial problem in some countries. At both the fourth and eighth grades, the one-quarter of students in schools not affected by resource shortages had higher average mathematics achievement than their counterparts in less well-resourced schools. For students at the sixth and ninth grades, there was more impact from lack of resources, with greater percentages of students in schools affected a lot by resource shortages (Mullis et al., 2012b: 229).

Conclusion

Increased public and policy demand in the 1980s for data to compare educational systems among different countries has advanced the field of international education statistics. International organizations such as the OECD and the IEA have created organizational structures for collaboration among

educational professionals, policy makers, and statisticians from different countries, which engaged in the development of shared definitions, concepts, indicators, and methodologies to ensure comparability of results across a wide variety of countries. Statistical systems in education today provide data on student learning outcome for different grade levels and subject areas (including mathematics, reading, science, and civic education), participation in education, education expenditures, and educational equality.

Statistical systems in the field of education not only provide academic knowledge and public information, but have also gained influence in the governance of education. Corresponding to outcome orientation in New Public Management (Nutley and Boaz, 2003), student learning outcome data have become part of public debate on education policy and governance. The indicators in statistical systems in the field of education played an important part in constructing a global policy field through governance by comparison (Martens, 2007). Competitiveness is increasingly used as a strong argument in various national legitimation strategies for education reform. Because the OECD and IEA frame student learning outcome data as indicators of the performance of education systems, they have become a factor in the international trend toward 'governing by numbers' or 'the politics of comparison' in the field of education (Grek, 2009). Yet, despite the effort and resources devoted to the high standards and quality control of those systems, scholars still question the methodological validity of international large-scale surveys on student assessments (Rindermann, 2007).

Bibliography

Andrews, P., et al., 2014. OECD and PISA Tests are Damaging Education Worldwide – Academics. The Guardian.

Blum, A., Goldstein, H., Guerin-Pace, F., 2001. International adult literacy survey (IALS): an analysis of international comparisons of adult literacy. Assessment in Education 8, 225–246.

Bonnet, G., 2002. Reflections in a critical eye: on the pitfalls of international assessment. Assessment in Education 9, 387–400.

Ertl, H., 2006. Educational standards and the changing discourse on education: the reception and consequences of the PISA study in Germany. Oxford Review of Education 32 (5), 619–634.

Fertig, M., 2003. Who Is to Blame? The Determinants of German Students' Achievement in the PISA 2000 Study. Institute for the study of labour, Bonn.

Gardner, D.P., et al., 1983. A Nation at Risk: The Imperative for Educational Reform. An Open Letter to the American People. A Report to the Nation and the Secretary of Education. Department of Education, Washington, DC.

Grek, S., 2009. Governing by numbers: the PISA 'effect' in Europe. Journal of Education Policy 24 (1), 23–37.

Grosh, M.E., Munoz, J., 1996. A Manual for Planning and Implementing the Living Standards Measurement Study Survey. LSMS Working Paper No. 126. The World Bank, Washington, DC.

Martens, K., 2007. How to become an influential actor – the 'comparative turn' in OECD education policy. In: Martens, K., Rusconi, A., Lutz, K. (Eds.), Transformations of the State and Global Governance. Routledge, London, pp. 40–56.

Martin, M.O., Mullis, I.V.S., 2013. TIMSS and PIRLS 2011: Relationships Among Reading, Mathematics, and Science Achievement at the Fourth Grade. Implications for Early Learning. Boston.

Mullis, I., Martin, M.O., Gonzales, E.J., Kennedy, A.M., 2003. PIRLS 2001 International Report. IEA, Chestnut Hill.

Mullis, I.V.S., Martin, M.O., Foy, P., Drucker, K.T., 2012a. PIRLS 2011 International Results in Reading. TIMSS & PIRLS International Study Center, Boston, Chestnut Hill, MA.

Mullis, I.V.S., Martin, M.O., Foy, P., Drucker, K.T., 2012b. TIMSS 2011 International Results in Mathematics. TIMSS & PIRLS International Study Center, Boston, Chestnut Hill, MA.

National Association of State Boards of Education, 1996. A Motion to Reconsider: Education Governance at a Crossroads. National Association of State Boards of Education, Alexandria, VA.

Nutley, S., Boaz, A., 2003. Evidence-based policy and practice. In: Bovaird, Tony, Löffler, Elke (Eds.), Public Management and Governance. Routledge, London.

OECD, 2003. PISA 2003 Assessment Framework. Mathematics, Reading, Science and Problem Solving. Knowledge and Skills. OECD, Paris.

OECD, 2014a. PISA 2012 Results: What Students Know and Can Do. In: Student Performance in Mathematics, Reading and Science, vol. I. PISA, OECD Publishing.

OECD, 2014b. PISA 2012 Results: Excellence through Equity. In: Giving Every Student the Chance to Succeed, vol. II. PISA, OECD Publishing.

Prais, S., 2003. Cautions on OECD's recent educational survey (PISA). Oxford Review of Education 29, 139–163.

Rindermann, H., 2007. The g-factor of international cognitive ability comparisons: the homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. European Journal of Personality 21, 667–706.

Schulz, W., Ainley, J., Fraillon, J., Kerr, D., Losito, B., 2010. ICCS 2009 International Report: Civic Knowledge, Attitudes, and Engagement Among Lower-Secondary School Students in 38 Countries. IEA, Amsterdam.

Soquel, N.C., Jaccard, P., 2008. Governance and Performance of Education Systems. Springer, Dordrecht.

Spiegelhalter, D., 2013. The problems with PISA statistical methods. Opinion Royal Statistical Society. http://www.statslife.org.uk/opinion/1074-the-problems-with-pisa-statistical-methods.

Torney-Purta, J., Lehmann, R., Oswald, H., Schulz, W., 2001. Citizenship and Education in Twenty-Eight Countries: Civic Knowledge and Engagement at Age Fourteen. IEA, Amsterdam.

UNESCO, 2012. Global Education Digest 2012. Opportunities lost: The impact of grade repetition and early school leaving. UNESCO Institute for Statistics, Montreal.

Wuttke, J., 2006. PISA – ein Zufallszahlengenerator. In: Jahnke, Thomas, Meyerhöfer, Wolfram (Eds.), Pisa & Co – Kritik eines Programms, hrsg. Franzbecker-Verlag, Hildesheim.

Relevant Websites

http://datatopics.worldbank.org/education/.
http://eacea.ec.europa.eu/education/eurydice/index_en.php.
http://www.iea.nl/iccs_2009.html.
https://nces.ed.gov.
http://www.oecd.org/edu/eag.htm.
http://www.oecd.org/pisa/.
http://www.oecd.org/edu/school/talis-about.htm.
http://timssandpirls.bc.edu/.
http://www.uis.unesco.org/Education/Pages/default.aspx.

Data Bases and Statistical Systems: Education, Statistical Systems

Tom Snyder, Chantilly, VA, USA

Published by Elsevier Ltd.

Abstract

Globalization of commerce has posed a need for more accurate and comprehensive international education comparisons, and this need has been met by new collections of international data and improvements to traditional education statistics. While many areas of international education comparisons remain problematic, advances in recent years provide important new information for policy makers and break new ground for further work by the research community. This article outlines new opportunities to make contributions to the area of international education statistics, and thus improve the information flow to policy makers and the public to make informed decisions on education processes. The data sets discussed in the article highlight modern education data collections available to the public that enable statistical comparisons among large groups of countries. While these data do not answer many research questions that can only be addressed by more detailed and specialized research studies, today's international data systems provide rich resources for understanding the differences in national education structures and targeting research efforts on specific promising developments.

The evolution of global communications has prompted major shifts in the operations of business and the exchange of information. While heavy and light industries have been affected by international trade for centuries, the intensity of sharing intellectual services across continents is a comparatively recent phenomenon. Corporations now routinely outsource computer programming, web design, accounting work, legal research, and a wide variety of other tasks to well-educated persons in foreign countries. Formerly, these high-paying positions were held by domestic employees. Now, retaining these highly skilled jobs has put pressure on countries and individuals to maintain high education standards to remain competitive in a market for highly skilled labor that is increasingly global.

While traditional metrics of productivity served well in measuring the output of individuals in an industrial setting, these measures are not adequate to measure the potential capacity of a population to engage in processes requiring relatively high intellectual inputs. In order to attract modern factories and support intellectual service enterprises, countries need to demonstrate the ability of their labor force to perform these skills. Countries and localities are finding it necessary to demonstrate the skills and educational backgrounds of their labor force through statistical data to attract new business or retain current ones. These businesses and their employees pay taxes that support government revenue and enable public services. In addition, higher levels of education have been consistently associated with higher income for individuals across countries. While manufacturing firms have long demonstrated the ability to transfer operations to areas with lower-cost labor, intellectual processing tasks now also can be transferred to other countries through simple contracting arrangements and advanced Internet-based communications. Thus, the international demand for intellectual skills and the growing fluidity of labor pools have increased the value and need for international education comparisons.

The need for more accurate and comprehensive international education comparisons has been met by new collections of international data and improvements to traditional education statistics. While many areas of international education comparisons remain problematic, the advances in recent years provide important new information for policy makers and break new ground for further work by the research community. This article is designed to outline new opportunities to make contributions to the area of international education statistics, and thus to improve the information flow to policy makers and the public to make informed decisions on education processes. The data sets discussed in the article highlight modern education data collections available to the public that enable statistical comparisons among large groups of countries. While these data do not answer many research questions that can only be addressed by more detailed and specialized research studies, today's international data systems provide rich resources for understanding the differences in national education structures and targeting research efforts on specific promising developments.

Developing International Comparisons

International education statistics have been available for decades through the efforts of the United Nations Educational, Scientific, and Cultural Organization (UNESCO) Institute for Statistics. Beginning in the 1960s, UNESCO worked on enumerating basic education statistics for all countries. The main efforts were to compile fundamental data on enrollments, degrees and certificates, teachers, and financial statistics. While coverage and timeliness were substantial challenges, the data represented the only comprehensive efforts in those years to compile comparable data for all countries in the world. By the early 1990s, it was clear that more detailed and timely data was needed by economically developed countries to provide statistical indicators on education on an ongoing basis. The Organization for Economic Cooperation and Development (OECD) was encouraged by the United States and other member countries to coordinate new data collection efforts that culminated in the first edition of *Education at a Glance* in 1992 (http://www.oecd.org/edu/skills-beyond-school/49338320.pdf). Later OECD and UNESCO data efforts were coordinated for OECD countries, since the data needs for the two international organizations overlapped to a large extent. Also, the needs for European Union as

represented by Eurostat were integrated into the data collection system. Through systematic efforts of OECD and member countries, as well as support from other international organizations, the collection of education data for OECD countries improved markedly during the following two decades. While data quality challenges remain, there have been notable strides in the coverage of education statistics collections and improvements in data quality and comparability. The administrative components of these data collections today include enrollments; student mobility from country to country; teachers; diplomas and degrees; education finance; labor force characteristics related to education; and system-level data on education structures. In addition, new assessments of students and adults provide information on the skill levels attained by individuals. Taken together, these data provide critical information on the basic characteristics of education systems, and are widely accessible to the public, through the annual *Education at a Glance* indicators report (http://www.oecd.org/edu/eag2013%20(eng)–FINAL%2020%20June%202013.pdf), published by OECD, and the periodic *World Education Indicators*, published by UNESCO. The results from these core data collections are also published by other international and national organizations and components of the raw survey data can be downloaded from UNESCO and OECD.

One of the key challenges in collecting and disseminating international education data is classifying levels of education in ways that are comparable from country to county. The current classification system of levels of education is called the International Standard Classification of Education, 1997, or ISCED 1997. A revised classification system ISCED 2011 has been approved for worldwide use, and was used in data collections beginning in early 2014. The data currently available were collected under the ISCED 1997 system, which is described in the accompanying text box.

International Standard Classification of Education (ISCED 1997)

ISCED Level 0 Education preceding the first level (early childhood education) usually begins at age 3, 4, or 5 (sometimes earlier) and lasts from 1 to 3 years.

ISCED Level 1 Education at the first level (primary or elementary education) usually begins at age 5, 6, or 7 and continues for about 4–6 years.

ISCED Level 2 Education at the second level (lower secondary education) typically begins at about age 11 or 12 and continues for about 2–6 years. Education at the lower secondary level continues the basic programs of the first level, although teaching is typically more subject focused, often using more specialized teachers who conduct classes in their field of specialization. The main criterion for distinguishing lower secondary education from primary education is whether programs begin to be organized in a more subject-oriented pattern, using more specialized teachers conducting classes in their field of specialization. If there is no clear breakpoint for this organizational change, lower secondary education is considered to begin at the end of 6 years of primary education. In countries with no clear division between lower secondary and upper secondary education, and where lower secondary education lasts for more than 3 years, only the first 3 years following primary education are counted as lower secondary education.

ISCED Level 3 Education at the third level (upper secondary education) typically begins at age 15 or 16 and lasts for approximately 3 years. Upper secondary education is the final stage of secondary education in most OECD countries. Instruction is often organized along subject-matter lines, in contrast to the lower secondary level, and teachers typically must have a higher level, or more subject-specific, qualification. There are substantial differences in the typical duration of programs across countries, ranging from 2 to 5 years of schooling. The main criteria for classifications are (1) national boundaries between lower and upper secondary education and (2) admission into educational programs, which usually requires the completion of lower secondary education or a combination of basic education and life experience that demonstrates the ability to handle the subject matter in upper secondary schools.

ISCED Level 4 Education at the fourth level (postsecondary non-tertiary education) straddles the boundary between secondary and postsecondary education. This program of study, which is primarily vocational in nature, is generally taken after the completion of secondary school and typically lasts from 6 months to 2 years. Although the content of these programs may not be significantly more advanced than upper secondary programs, these programs serve to broaden the knowledge of participants who have already gained an upper secondary qualification.

ISCED Level 5 Education at the fifth level (first stage of tertiary education) includes programs with more advanced content than those offered at the two previous levels, and is divided into 5A (academic) and 5B (vocational). Entry into programs at the fifth level normally requires successful completion of ISCED level 3 or ISCED level 4.

ISCED Level 5A Tertiary-type A programs provide an education that is largely theoretical and is intended to provide sufficient qualifications for gaining entry into advanced research programs and professions with high skill requirements. Entry into these programs normally requires the successful completion of an upper secondary education; admission is competitive in most cases. The minimum cumulative theoretical duration at this level is 3 years of full-time enrollment. ISCED 5A programs include not only bachelor's degrees, but also other degrees below the doctor's degree level, such as master's degrees.

ISCED Level 5B Tertiary-type B programs are typically shorter than tertiary-type A programs and focus on practical, technical, or occupational skills for direct entry into the labor market, although they may cover some theoretical foundations in the respective programs. They have a minimum duration of 2 years of full-time enrollment at the tertiary level. In the United States, such programs are often provided at community colleges and lead to an associate's degree.

ISCED Level 6 Education at the sixth level (advanced research qualification) is provided in graduate and professional schools that generally require a university degree or diploma as a minimum condition for admission. Programs at this level lead to the award of an advanced, postgraduate degree, such as a Ph.D. The theoretical duration of these programs is 3 years of full-time enrollment in most countries (for a cumulative total of at least 7 years at levels 5 and 6), although the length of the actual enrollment is often longer. Programs at this level are devoted to advanced study and original research (http://nces.ed.gov/programs/coe/glossary.asp#i, and http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx).

While the ISCED:97 overview serves well as guidance to researchers, more detailed reference documents are provided to country officials who are responsible for reporting national data according to these international guidelines. It should be noted that the ISCED structure was not intended to take into account the rigor of national programs or the curricular content, except in the most basic way (e.g., academic or vocational). While many education programs do correspond clearly into one of the ISCED levels, most countries have at least some programs that contain characteristics associated with two different ISCED levels, and these classification problems present at least some limitations in international comparability. One of the key limitations in the ISCED:97

structure was the lack of a specific classification of master's degrees programs. This presented difficulties for some countries, such as the United States, in the 1990s, but the problem expanded as a result of the 'Bologna Process' in Europe. Through the Bologna Process, many countries adopted the bachelor's/master's/doctor's structure similar to that used in the United States. The introduction of the classification for a master's level of education is one of the important improvements coming through the ISCED:11 classification system (*A New ISCED: Reflecting Today's Education Systems*, UNESCO, Institute for Statistics http://www.uis.unesco.org/Education/Documents/UNSD_newsletter_27e_ISCED.pdf).

The comprehensive international education statistics that are currently compiled on a routine basis are fundamentally based on two different structures of data collection. There are country-level data, primarily based on administrative records that usually are reported on an annual basis by country-level respondents to UNESCO, OECD, and Eurostat. The second type of data collection involves representative samples of individuals within countries. The most frequently encountered sample data of this type are student achievement data collected through the Trends in International Mathematics and Science Survey (TIMSS) and the Programme for International Student Assessment (PISA). There is also a new literacy assessment for adults entitled Programme for International Adult Competencies (PIAAC). These two general structures of international data collection are not completely distinct by original survey method since there are sample data compiled by countries and submitted as national summaries in conjunction with administrative data, and there are country-level data appended to the assessment data sets. Both the country-reported summary data and national samples each contribute important information about education programs and lead to a more complete understanding of national education systems.

Country-Reported Data

Data that are summarized at the country level and reported by county representatives are primarily based on administrative data collections conducted by the national education authorities. The annual data compiled by UNESCO, OECD, and Eurostat are collected through a set of survey instruments called the UOE. The data collected by UNESCO are the more basic and are requested for all UN countries. This data collection includes information on enrollments, staff, graduates, and finance. OECD countries are requested to report more detailed data, particularly on tertiary education, as well as data on educational characteristics of the labor force, and on operations of national education systems. European Union countries are required to report the OECD data, plus additional detailed information for Eurostat. Most of the data collected by UNESCO are publically available from reports and an online data tool. OECD publishes a comprehensive annual report entitled *Education at a Glance*, which highlights these data collections, plus other more specialized reports. Other key data can be downloaded from their web tool. Eurostat primarily releases data through reports, but also makes some data sets available through purchase.

Enrollment

The most basic data characterizing an education system are student enrollments. Since enrollment data are generally a component of education funding formula, they are often among the most accurate and consistently collected data. The enrollment survey is a fundamental component of the OECD UOE data collection. Enrollment data provide fundamental information on the scope of an education system, and trend data provide information on whether the system is growing or contracting. These changes in enrollment are a critical concern of policy makers because they affect budgetary considerations with respect to decisions on such topics as staff hires, school construction and renovations, and labor pool entry. Not only are enrollment data a key statistic on their own, but they are often used as a denominator in conjunction with other education data such as finance and teachers. Statistical measures such as expenditures per student and students per teacher provide metrics that control for the size of the education system. While enrollment data may be thought of as being very definitive in terms of concept and definition, there can be complexities in the classification of students, particularly at the postsecondary level, or between academic and vocational programs. Another element of potential data problems involves full-time and part-time students. The distinction between full-time and part-time enrollment can be important in ratios to finance and other variables that are optimally computed on the basis of full-time-equivalent students. The extent to which the full-time-equivalent of the part-time students can be calculated with precision can affect ratios based on these computations. Countries most often experience difficulties in computing full-time-equivalent enrollment at the preprimary and postsecondary levels. Despite some limitations, enrollment data are considered core information for understanding the scope of education systems and are frequently a key component of long-term trend analyses.

Staff

The annual collections of education data conducted by international organizations (UOE) focus on teachers, rather than on other types of school staff. Teacher counts often are gathered by characteristics, such as age, gender, and full-time/part-time employment status. There is also a separate data collection by the OECD on average salaries for teachers. Data are on other school staff such as teaching assistants and school level managers are collected annually through the UOE, with periodic collections of other staff. Two key metrics of particular interest to the research and policy-making communities are the pupil/teacher ratios and the average salaries for classroom teachers. The pupil/teacher ratio is a key measure of the teaching resources available in schools. The average teacher salary provides an indication of the financial value placed on hiring high-quality staff, particular when referenced to earnings metrics for other professions within the specific country. Thus, teachers' salaries are sometimes compared to GDP per capita, or earnings of other individuals with similar levels of education. Potential challenges in the use of teacher data, particularly in the use of ratios, are the determination of full-time-equivalent teachers based on counts of full-time and

part-time teachers. For many countries at the ISCED Levels 1–3, most teachers instruct full time so this is often not a serious problem. However, some countries do have large numbers of part-time teachers at the ISCED Levels 1–3, and many countries have relatively large numbers of part-time teachers at ISCED Levels 5 and 6. Also, the varying research duties and instructional loads for tertiary faculty can make comparisons at that level somewhat difficult to interpret.

Graduates

The counts of education-level graduates provide essential data on the output of an education system. When compared to enrollment or relevant-aged population groups, these data can provide comparisons of the proportion of students or population who are obtaining diplomas and degrees. Some measures of equity, particularly involving males and females are also feasible in most international data collections; however measures for economically or other types of at-risk populations are generally not available for most international data collections based on administrative data. Completions at the college level can be disaggregated by field of study to measure the relative output of graduates prepared for critical occupational areas. While data on school completions do not involve some of the limitations of full-time versus part-time status that enrollment data can present, the classification of programs at particular ISCED levels can present a problem. The general concept of high school or secondary (ISCED Level 3) is common across many countries, although some programs are longer or shorter in duration. However, the distinctions in academic and vocational programs can pose difficulties in comparisons. For example, Denmark, Germany, Austria, Iceland, and other countries have long duration ISCED Level 3 vocational programs that can last until students are 22, 23 years, or older. Programs at these levels may be classified as postsecondary in other countries. At the postsecondary, there are many types of vocational programs of varying lengths that are currently classified as ISCED Level 5B. Programs that are classified as postsecondary academic may range in length from 3 to 6 years of full-time study. The introduction of the ISCED 2011 classification structures will substantially improve the comparability of international postsecondary data (http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx).

System Level Data

Eurydice, which is coordinated and managed by the European Union Education, Audiovisual and Culture Executive Agency, and the OECD conduct annual and periodic surveys of information describing the structure of education systems. For example, the Eurydice report, *Key Data on Teachers and School Leaders in Europe 2013*, http://eacea.ec.europa.eu/education/eurydice/documents/key_data_series/151EN.pdf provides a range of information on national teaching practices including teacher preparation and support, professional development, and levels of autonomy and responsibilities. Other types of descriptive statistics gathered on a regular basis include instructional time for children at different age levels and teaching hours per year (see Indicators D1 and D4 in *Education at a Glance 2013*). Periodic data collections by the OECD have included such topics as teacher professional development, national assessments, and autonomy of private schools. Descriptive information provides important context for interpreting statistical data such as information on the amount of instructional time in the year can affect interpretation of teacher salaries. Also, system-level information provides important information to policy makers about how the structure of national education systems differ, such as with respect to school governance and autonomy. One of the most frequently encountered challenges in using system-level data is understanding to the degree to which federal countries may have substantively important variations in structures among the various education systems which exist within their countries.

Finance

The education finance comparisons constitute some of the indicators most closely watched by policy makers. Government finance decisions can affect the level of taxation to support government activities as well as the allocation of government funds to education as opposed to other types of government functions. Aside from government decisions, the per capita income of the country can have an important relationship with education funding. Budgetary decisions constitute one of the primary levers that policy makers have on the education system. There is a significant correlation between available national finance resources, as measured by GDP, and expenditures per student at both the primary/secondary and tertiary levels (http://nces.ed.gov/programs/coe/indicator_cmd.asp). Another type of important comparison involves the relationship of total education spending and GDP, which provides an indicator of the priority of education spending relative to other components of the economy. The finance data involve some of the more problematic statistics in international data collections. Accounting systems and principles differ quite substantially from country to country. Most of the country-reported summary data are at least partly estimated to adjust national accounting procedures to international survey definitions. While reasonable estimates can be made for many items, some types of differences are structural. For example, in the United States, the health care system is primarily based on employer/employee funding. Hence, health care costs for the teachers and other school staff are included in the education funding for the United States, while they are included in national health budgets (not education funding) for other countries with national health care systems. Despite limitations, finance data are included in most international comparisons due to the very high policy relevance.

Labor Force

Another set of closely watched data are the labor force outcomes related to education. These data include such measures as employment and unemployment rates, and median incomes, related to level of education completed. These data are collected annually for OECD countries, and there are also surveys conducted by Eurostat. The OECD measures include the labor force participation and unemployment rates, as well as the median annual income for

those who have completed various education levels. Fostering a healthy job market is a key policy arena for all countries. The educational attainment data are considered an important indicator since countries with high levels of education among their citizenry are viewed as more able to adjust to new labor skill demands and technologies. Unlike most of the education data previously described on enrollment, graduates, staff, and finance, the labor force outcomes are collected through labor force surveys based on samples in each country. In general, these labor force surveys are not centrally coordinated by international organizations as with the student assessment surveys described in the following paragraphs. The labor force surveys are nationally designed to meet national reporting requirements. However, the alignment of the definitions for such key concepts as employment, unemployment, and annual earnings are quite good. Because of the relatively strong alignment between national and international concepts for most countries, the comparability of data on labor force statistics is better than for some types of international education statistics. A key limitation is the classification of education levels as noted above on graduates, with the additional complicating factor that many countries have internal classification problems because their education systems have changed over time. Also, although nearly all the labor force data are based on statistical samples, standard errors are only available to a limited extent. While statistics based on large population groups, such as those based on 10-year age bands for ISCED Level 3 completers, are fairly robust for most countries, researchers need to be careful about comparisons involving smaller population groups. Standard error estimates currently available for selected educational attainment and labor force tables in *Education at a Glance* provide some useful guidance on the magnitude of standard errors for various types of comparisons.

Individual-Level Data Collections

Among the most important expansions of international data collections over past two decades has been the increased focus on student assessment data. In contrast to the international comparisons based on summary data reported at the national level by country statistical representatives, these achievement data have been gathered at the student level using comparable assessment survey instruments translated into national languages. The national data collections are coordinated by country representatives who manage such activities as facilitating cooperation from schools who have been selected to have their students participate in the survey. National coordinators are also involved in the design of the survey instruments to ensure that the assessment questions are in contextual and language structures that are appropriate for the students in their country. International authorities require surveys to be conducted within countries according to defined procedures. Also, the sample selections and survey administrations are monitored by international organizations to ensure high quality and consistency with established guidelines.

The international assessment data provide essential information on the ability of students to demonstrate acquired skills in key subjects, such as reading, mathematics, and science. In addition to information on average performance levels of students, the assessment data provide statistics on the range of student performance including information on low socioeconomic and migrant students, compared to other students. The extent to which countries can reduce the range of student scores can provide an indication of the success of programs to provide equity. The student-based information can also be analyzed with respect to relative variations between schools in student performance or other measures such as pupil/teacher ratios. Such information on the range in school characteristics also cannot be addressed by the summary data reported by national authorities. In addition to student assessments, two additional individual-based international education surveys have been conducted recently. A representative sample of teachers in many OECD countries has been gathered through the Teaching and Learning International Survey (TALIS). A new international comparisons milestone was reached with the release of the Programme for International Adult Competencies (PIAAC) in 2013 which provides data on adults and their performance on literacy, numeracy, and computer-related skills.

Trends in International Mathematics and Science Study

The International Association for the Evaluation of Educational Achievement (IEA) coordinates three major survey systems: The Trends in International Mathematics and Science Study (TIMSS); the Progress in International Reading Literacy Study (PIRLS); and the Civic Education Survey (CivEd). TIMSS has measured trends in mathematics and science achievement at the fourth and eighth grades on regular 4-year cycles since 1995. In addition, TIMSS Advanced, which measures trends in advanced mathematics and physics for students in their final year of secondary school, was conducted in 1995 and 2008. http://timss.bc.edu/home/pdf/TP_About.pdf. In 2011, more than 60 countries and other education systems, including the United States, participated in TIMSS. In addition to student assessment data the TMSS survey system includes surveys on student backgrounds, teachers, and schools. National research coordinators provide information on the national curriculum and education policies and practices. http://nces.ed.gov/surveys/international/intl_assessments_table.asp.

Progress in International Reading Literacy Study

The Progress in International Reading Literacy Study (PIRLS) collects information on the reading skills of fourth graders. PIRLS has been administered in 2001, 2006, and in 2011. The PIRLS assessment measures student performance on a combined reading literacy scale. There were also two subscales of purposes of reading: reading for literary experience and reading to acquire and use information and for two subscales of processes of comprehension: Retrieving and Straightforward Inferencing and Interpreting, Integrating, and Evaluation. In 2011, 53 education systems participated in PIRLS. Similar to TIMSS, additional contextual information was gathered through student background questionnaires, teacher questionnaires, school questionnaires, and in many countries, parent questionnaires.

The Programme for International Student Assessment

The OECD coordinates the Programme for International Student Assessment (PISA), which measures 15-year-old students' reading, mathematics, and science literacy. In contrast to the more curricula-based TIMSS assessment, PISA also includes measures of general or cross-curricular competencies, such as problem solving, functional skills that students have acquired as they near the end of compulsory schooling. PISA was first administered in 2000 and is conducted every 3 years. The most recent assessment was in 2012.

Teaching and Learning International Survey

The Teaching and Learning International Survey (TALIS), coordinated by OECD, is an international study of teachers, teaching, and learning environments. TALIS provides internationally comparable data on teachers and instructional practices to highlight current conditions and inform education policy. TALIS was first administered in 2008 in 24 countries. TALIS was administered again in 2013 with new survey instruments that are based on a broad international perspective on policy and research interests. Thirty-four countries and educational systems participated in TALIS 2013 which was released in June 2014.

Programme for International Adult Competencies

The Programme for International Adult Competencies (PIAAC) breaks new ground in assessing the functional literacy and numeracy skills of adults in about 25 countries in 2011–12. This complex survey, coordinated by OECD, also gathered information on the skills of adults in using computers to solve problems in a technological environment. Nationally representative samples of about 5000 adults were surveyed in each of the countries, providing information on the skill levels of adults with various education backgrounds. The background information contained information on labor force participation, occupations, educational activities, and other family and demographic background information.

Using International Data Systems

Over the past decade, there has been a large quantitative and qualitative expansion in the statistical resources available to researchers in the field of international education comparisons. There have been important upgrades to the more traditional data collections based on nationally reported summary data, including many new waves of student assessments, and completely new international data on teachers and on adult functional literacy. The ability to access data has been greatly improved, including not only high-quality publications, but new online data tools that enable researchers to produce custom tabulations and analyses. The *Resources* text box provides helpful starting points to obtain detailed descriptions about the surveys that have been cited in this article and to access data and findings that can serve as a basis for further analyses or references for public presentations.

Despite the expansions and improvements in international education data systems, data limitations remain. Although great efforts are made to adjust data for international comparability, structural differences in education practices, and curricula pose complicating factors for education authorities. Although assessments are designed to reflect curricular content that is common worldwide, there are important differences among all countries and within federal countries. Care must be used in drawing conclusions from educational systems within countries, and the extent to which those subnational units represent the rest of their country. With the array of data now available to the research community, it is possible to validate many data sources with related information from other sources. This not only serves to place conclusions on a more firm basis by aligning similar data from multiple sources, but also can bring together a more complete picture of education across the world.

Resources

The UNESCO Institute for Statistics (http://www.uis.unesco.org/Education/Pages/default.aspx) has a long tradition of collecting education data, and makes information through publications such as *Global Education Digest* (http://www.uis.unesco.org/EDUCATION/Pages/global-education-digest.aspx), and online data tools (http://stats.uis.unesco.org/unesco/TableViewer/document.aspx?ReportId=143&IF_Language=eng).

The OECD publishes the widely publicized annual report Education at a Glance (http://www.oecd.org/edu/eag2013%20(eng)–FINAL%2020%20June%202013.pdf), and supports an extensive program of shorter reports and data tools (http://www.oecd.org/edu/). OECD offers access to microdata (data sets with individual level responses for use by researchers for advanced analyses) such as PISA through data tools and downloadable data sets (http://pisa2012.acer.edu.au/).

Education data for European Union countries are compiled by a number of agencies. The statistical data are compiled primarily by Eurostat (http://epp.eurostat.ec.europa.eu/portal/page/portal/education/introduction). Eurostat prepares a variety of reports on education and their web site contains a large array of statistical tabulations. Access to individual data is provided only research organizations. Other European Union agencies such as Eurydice (http://eacea.ec.europa.eu/education/eurydice%20/index_en.php) and European Centre for the Development of Vocational Training (http://www.cedefop.europa.eu/EN/Index.aspx) provide more specialized information on European education systems.

The TIMSS and PIRLS International Study Center at Boston College provide a comprehensive range of products and documentation for these large survey systems (http://www.timss.org/). These products include methodology and findings reports, data tools, and tables. Microdata can also be downloaded from their web site for more complex analyses.

Within the United States, the National Center for Education Statistics (http://nces.ed.gov/surveys/international/) hosts some convenient summaries of international data http://nces.ed.gov/surveys/international/intl_assessments_table.asp and advanced data tools that enable access to TIMSS, PIRLS, and PISA data for the United States as well as other countries participating in the surveys (http://nces.ed.gov/surveys/international/ide/).

See also: Comparative Research in Education: IEA Studies; Cross-Cultural Study of Education; Cross-National Comparisons in Education: Findings from PISA; Education, History of; Educational Effectiveness, The Field of; Globalization in Education; International Organisations and Education.

Data Bases and Statistical Systems: Environment and Urban Studies

Erika Nóra Regös, Max-Weber-Institute for Sociology, Heidelberg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article on Statistical System: Environment, Urban Studies, and Human Development will demonstrate the importance of statistics on environment, urbanization, and human development to social science. All three fields of this article provide the readers with access to the most reputed national and international databases and frameworks in these fields. Additionally it will enhance the readers' basic understanding of the structure and dynamics of today's rapidly changing global environment, and its impact on living conditions.

Introduction

Statistics on environment, urbanization, and human development are multidisciplinary, their sources are dispersed, and a variety of methods are used for their compilation. Environmental statistics generally aim to describe the quality and availability of natural resources, natural events that affect the environment, the impacts of these resources and events, as well as social responses to these impacts. Urban studies usually compile housing and real estate economies in order to measure infrastructure and quality standards of housing and the degree of urbanization of a given environment. In addition to the description of environmental quality standards, human development studies have put increasing emphasis on the measurements of basic processes and contexts of human life, such as the measurement of human well-being, life expectancy, and life satisfaction. Although from different perspective, all three fields aim to describe the quality standards of life and changes over time according to objective key figures.

Environment statistics, urban and human development studies, and related fields are usually compiled and disseminated by government departments, central statistical services, research institutes, and international organizations. They are often interrelated with statistics in other fields, for example, with data on education, health, and gross domestic product (GDP).

Statistical Systems

Environment

History

During the last four decades, various national and international efforts have been made toward developing a system or framework for environment statistics, either for a planned program of statistics or for presenting available data in a coherent statistical publication. As a conceptual framework, the United Nations Conference on the Human Environment (Stockholm, June 1972) was the first global conference to signal that environmental concerns had increasingly become the subject of mainstream socioeconomic policies. As a more empirical way, the first initiatives for the development of environment statistics at international level stemmed from two meetings of the Economic Commission for Europe (ECE) in 1973, which was followed by pioneered work on standard environment statistics classifications by the UNECE Statistics Division. Environment statistics programs also started at Organisation for Economic Co-operation and Development (OECD) and later at Eurostat, focusing on data collection and indicator development, in the late 1970s and early 1980s.

In the second world conference, the United Nations Conference on Environment and Development (Rio de Janeiro, June 1992), increasing emphasis was put on integrated environmental and economic accounting. One of the main outcomes of this conference was the adoption of Agenda 21, which sets out a series of principles and actions for achieving sustainable development (Hens, 1996).

As a result of the increasing demand for environmental information, the first global collection of environment statistics was launched by United Nations Development Group (UNSD) in 1999. Since then it has been established in a biennial basis. After 1 year, in 2000, 189 countries of the United Nations agreed on 10 globally environmental indicators set up under the Millennium Declaration (2000) by the UNSD. Those indicators are developed to monitor progress, and simultaneously committed the countries to the achievement of the eight declaration's goals and the other targets by 2015.

Most recently, a huge number of documents and programs have been emphasizing the increasing importance of environmental statistics, information, and indicators. Especially, the outcome document, 'The Future We Want' of the United Nations Conference on Sustainable Development (Rio + 20, Brazil, June 2012) highlights the necessary advancement in environmental information. It is also important to mention the European Commission's efforts to integrate seven key environmental indicators in its general framework on 'strategy for sustainable development.' The OECD regularly updates its environmental data and indicators given the need of reliable and harmonized data on environment and including the ways in which it is affected by economic activity.

The Organization of Environment Statistics

Research on environmental data varies from country to country and over time, which makes an international harmonization more difficult, although there have been some successful efforts to organize them internationally.

The Statistical Office of the United Nations Secretariat was the first to identify those common characteristics that could be incorporated into a widely applicable international framework. The results can be divided into four basic approaches: (1) the media approach, (2) the stress-response approach, (3) the accounting approach, and (4) the ecological approach.

The stress-response approach, originally developed by Statistics Canada as a "Structural Framework for the Stress-Response Environmental Statistical System," is the only approach that analyses the interactions between human activities and the environment. Its main focus lies on impacts of human intervention within the environmental (stress – such as waste generation, extraction of natural resources, and the production of hazardous substances) and the environment's subsequent transformation (environmental response). All other three approaches give information on the different environment issues without any deeper consideration of the human intervention in the environmental course: (1) the media approach deals with the major environmental components of air, water, land/soil, and the artificial environment; (2) the accounting approach analyzes the flow of these components from their extraction (harvest) from the environment, through successive stages of processing and final use, to their return to the environment; whereas (3) the ecological approach gives an overview on a variety of models, monitoring techniques, and ecological indices such as the assessment of population diversity and dynamics, of biomass production and its productivity, stability, and resilience of ecosystems.

Despite the presence of these four different basic approaches, the existing national and international approaches indicate a preference for the combination of the media and the stress-response approaches. To operationalize a set of activities that exert stress on the environment, (1) measures of stressors, (2) analysis of environmental response, (3) measurement of collective and individual responses, as well as (4) assessment of stocks of natural resources, artificial structures, and potentially hazardous substances have been developed. (See more on the stress-response at the Web site of the World Bank Group at: https://www.virtualstatisticalsystem.org/themes/theme/31-environment/?no_cache=1.)

Framework for the Development of Environment Statistics

In 1984 the Framework for the Development of Environment Statistics (FDES) was developed and published by the UNSD to systematically present and organize environmental statistics. Its first two technical reports, the 'Concepts and Methods of Environment Statistics: Human Settlements Statistics' (published in 1988) and 'Concepts and Methods of Environment Statistics: Statistics of the Natural Environment' (published in 1991) describe in detail the sets of statistical variables within the FDES.

Recently, the FDES have been updated and enlarged by two aspects: (1) 1984, the FDES has been completely revised, based on improved scientific knowledge about the environment and new statistical requirements created by emerging environmental policies and concerns; and (2) a core set of environment statistics that has been developed aiming to create an agreed, limited set of environmental statistics that are of high priority and relevance to most countries. Actually, it is Tier 1 of a greater (though not exhaustive) basic set of environment statistics composed of three ranks according to the level of relevance, availability, and methodological development of the statistics. The aim of the revision was to emphasize the relationship of human activities and natural events in more detail by creating information categories that can describe a sequence of action, impact, and reaction. According to the UNSD, the revised FDES 2013 aims to guide nation-states in the development process of their own environment statistics programs primarily at early stages. In addition, it aims to be of help for all three institutional levels (national, international, and regional) in order to organize and strengthen the production of environment statistics.

UNSD also cooperates closely with other organizations like the OECD, the Statistical Office of the European Union (Eurostat), United Nations Environment Programme (UNEP), the UN Division of Sustainable Development (DSD), and the regional commissions, as well as with specialized agencies, secretariats of international conventions, NGOs, and others in order to create synergies and to avoid duplication of efforts in data collection.

Organisation for Economic Co-Operation and Development

The OECD has been developing environmental indicators since the early 1990s and has published the first international sets of environmental indicators. These also included the OECD Core Set of Environmental Indicators. By ongoing revision and improvement of its environmental indicators, nowadays those indicators help to integrate environmental considerations into sectoral policies, such as energy, transport, and agriculture, and to measure progress in decoupling environmental pressures from economic growth. Most recently, the OECD has developed a set of indicators to monitor progress toward green growth. (See more at the OECD report 'Environment at a Glance' online at their Web site http://www.oecd-ilibrary.org/docserver/download/9713051e.pdf?expires=1395789956&id=id&accname=guest&checksum=1B9A2B7C9F583B4823DABEA16639F642.)

The Statistical Office of the European Union

Eurostat, The Statistical Office of the European Union, situated in Luxembourg, has also published a number of compendia about environmental statistics at a European level. In addition, and in response to an agreement by the European environment ministries at the Dobris Castle, a pan-European report was completed in 1995 (Dobris Assessment). Currently, the Eurostat environmental accounts include indicators on climate change and greenhouse gas emissions, waste generation and treatment, water resources, abstraction and use, wastewater treatment, forestry and biodiversity, chemicals, material flow accounts, and relevant financial indicators such as environmental protection expenditure and environmental taxes. The indicator data are collected by the Eurostat and the European Environment Agency (EEA) and are available from the EU, its member states, European Free Trade Association countries, and candidate countries. (For detailed data please check the Eurostat Web site at http://ec.europa.eu/eurostat and the data taken from the EEA at http://eea.europa.eu.)

The European Environment Agency EEA

EEA is an agency of the EU established in 1990 to provide sound and independent information on the environment. It is a major information source for those involved in developing, adopting, implementing, and evaluating environmental policy. Currently, the EEA has 33 member countries. EEA's main clients are the EU institutions and the member countries. In addition, EEA also serves other EU institutions such as the Economic and Social Committee and the Committee of the Regions.

Most recently, the implementation of an improved environmental system for Europe, the so-called 'Shared Environmental Information System (SEIS),' has become the main focus of the EEA 2009–2013 Strategy and its daily operations. Originally, it was a proposal of the European Commission (EC) entitled 'Toward a Shared Environmental Information System (SEIS)' in 2008. Since that time, SEIS has become a collaborative initiative of the EC and the EEA and its Eionet (European environment information and observation network), consisting of 38 countries. The SEIS's goal is to create an improved environmental information system for Europe based on a network of public information providers that share their environmental data and information. (For more information on EEA Web site at: http://www.eea.europa.eu.)

The United Nations Environment Programme

The UNEP, founded as a result of the United Nations Conference on the Human Environment in 1972, is an agency of the United Nations that coordinates its environmental activities, i.e., assisting developing countries in implementing environmentally sound policies and practices. The so-called Environmental Data Explorer is the authoritative source for data sets used by UNEP and its partners in the Global Environment Outlook (GEO) report and for other integrated environment assessments. Its online database holds more than 500 different variables, such as national, subregional, regional, and global statistics or geospatial data sets (maps), covering a wide range of topics such as freshwater, population, forests, emissions, climate, disasters, health, and GDP. (See more at their Web site at http://www.unep.org.)

The UN Division for Sustainable Development

The DSD provides leadership in promoting and coordinating the implementation for a sustainable development agenda of the United Nations. The division puts their main focus on five core functions: (1) support to UN intergovernmental processes on sustainable development, (2) analysis and policy development, (3) capacity development at the country level, (4) interagency coordination, and (5) knowledge management, communication, and outreach. Even though there is no special database on environmental indicators provided by the division, its contribution to the sustainable development is highly valuable. By participating in programs such as the Expert Group Meeting on 'Institutionalizing sustainable development indicators for measuring progress of national strategies' in Barbados in 2009, it encourages national governments and a range of regional institutes to conceptualize and implement their own systems of indicators on sustainable development. (See more on the Expert Group Meeting online at: http://sustainabledevelopment.un.org/content/documents/1380REPORT_final.pdf.)

United States Environmental Protection Agency

In 1970 the United States Environmental Protection Agency (EPA) was established to consolidate the variety of federal research, monitoring, standard-setting, and enforcement activities into one agency to ensure environmental protection for the United States. With help from numerous partners, EPA has compiled a set of databases tracking down the following science topics: air, climate change, health risks, pollutants and contaminants, waste, and water. Most of these indicators focus on the United States, but some include global trends to provide a context or a basis for comparison. As EPA points out, their indicators represent a selected set of key measures for the respective scientific area, and do not reflect an exhaustive group of all possible indicators. Moreover, all indicators are based on peer-reviewed data from various government agencies, academic institutions, and other organizations. The selection of these indicators is based on the quality of the data and other criteria.

Additionally, EPA has also developed a data finder to point out data sources that EPA manages and to facilitate access to environmental information (it can be found at: http://www.epa.gov/datafinder/). Similarly, the federal site, Data.gov, helps people to find, download, and use data sets that are generated and conducted by the federal government. (See more details on the EPA's indicators, tools and methodologies here: http://www2.epa.gov/science-and-technology and on the Federal government Web site at: http://www.data.gov/.)

Human Development

Most of the indicators on environmental statistics show changes are occurring throughout the Earth's climate system, including increases in air and ocean temperatures, rising extreme weather events, a rise in sea level, widespread melting of glaciers, and longer ice-free periods on lakes and rivers. Such changes can represent a wide range of challenges to human well-being, the economy, and natural ecosystems. For that and other reasons, many researchers have turned their attention to construct indicators, which study human development best.

Human Development Index

The core insight at the center of the human development approach remains constant and as valid today as two decades ago: "Development is ultimately best measured by its impact on individual lives" (UNDP, 2014: 7).

The Human Development Index (HDI) was developed by the UNDP to rank developed, developing, and underdeveloped countries by their level of human development. It measures the average achievements in a country according to three basic dimensions of human development: (1) a long and healthy life (life expectancy), (2) access to knowledge (mean and expected years of schooling), and (3) a decent standard of living (GINI per capita) (UNDP, 2014: 2). The HDI was created to emphasize that people and their capabilities should be the ultimate criteria for assessing the development of a country, not economic growth on its own. Additionally, the HDI is often used to question national policy choices. (For a full elaboration

of the method and its rationale, see Klugman et al. (2011), for more information on the general use and indicators of the HDI see the following Web site: http://hdr.undp.org/sites/default/files/hdr_2013_en_technotes.pdf.)

Inequality-Adjusted Human Development Index

The Inequality-adjusted Human Development Index (IHDI) was introduced by UNDP in their Human Development Report 2010. It was created to measure the level of human development of people in a society that accounts for inequality. Under perfect equality the IHDI is equal to the HDI. However it falls below the HDI when inequality rises. In other words, the IHDI can be seen as the actual level of human development (taking into account inequality), while the HDI can be viewed as an index of the potential human development that could be achieved if there is no inequality. The index has been applied to 132 countries so far. (For more information and data available see their Web site at: http://hdr.undp.org/en/statistics/ihdi.)

World Population Prospects

The World Population Prospects represents the official United Nations population estimates and projections conducted by the Population Division of the United Nations Department of Economic and Social Affairs of the United Nations Secretariat. It is used widely throughout the United Nations and by many international organizations, research centers, and academic researchers. The 2012 revision of the world population prospects represents the 23rd round to display key demographic indicators for each development country and area within 1950–2100. (Detailed information about indicators on demographic variables such as population, mortality, fertility rates, on data sources used to derive estimates for those indicators, and on method used to derive mortality estimates, age mortality pattern, and child mortality can be found at their Web site at: http://esa.un.org/unpd/wpp/unpp/panel_indicators.htm.)

Housing and Urban Studies

Cities are seen as both the source and solution of today's economic, environmental, and social challenges. For instance, they are home to 75% of the EU's population, they account for about 80% of energy use, and they generate about 85% of Europe's GDP (European Commission, 2013). Therefore, cities are central to achieve sustainable and inclusive growth within an increasingly competitive global context. That is the reason why urban development – economic, social, and environmental – is central to several international frameworks, such as to the Europe 2020.

History

To develop indicators on health of housing, real estate markets, and on cities and their inhabitants, much effort has been undertaken over the past 40 years (Muth, 1969; Mills, 1972; Hochwald, 1961; Tiebout, 1962). Hirsch's (1966), Flax's (1978), and Murphy's (1980) are just a few of early works in urban economics focused on data and indicator issues. The popularity of *Money* magazine's annual rating of cities as well as Savageau and Boyer's annual almanac urban indicators have even become popular (Malpezzi and Mayo, 1997: 3). With the early 1990s, there was hardly any consistent data available for international comparisons of urban development outcomes, with an exception of some urban population figures (see World Bank, 1994). Since the early 1990s, some research on international indicators has started to fill this gap by an awakening interest in comparing housing and urban development indicators across countries. Especially, the United Nations and the World Bank sponsored the collection of several dozen basic urban indicators by individual researchers in a sample of 51 countries. The aim was to document and explain ideal collection procedures and to maximize comparability. The collected indicators could then be used in both a diagnostic and in a comparative way (for example see Mayo, 1993) or for the construction of indexes for housing policies and subpolicy such as regulation and taxation (Angel and Mayo, 1996). These initial efforts of the United Nations and the World Bank toward international comparability resulted in an expansion, both in terms of more countries and a wider range of indicators. In addition to housing and real estate indicators, a huge number of indicators related to socioeconomic development, infrastructure, transport, the urban environment, and local governments have been collected at the turn of the millennium. In the United States, the Department of Housing and Urban Development has taken the lead in preparing an initial set of indicators on 100 large U.S. metropolitan statistical areas (Glickman et al., 1996). In Europe, the Eurostat 'Urban Audit' data collection has been providing information and comparable measurements on the different aspects of urban life quality in the European cities since 2003.

This brief historical paragraph on housing and urban studies relies mostly on the work of Malpezzi and Mayo (1997). In addition to the data bases presented and works cited therein, representative examples include, among many others, the work of Case and Schiller (1989) and Thibodeau (1995) on housing prices; of Holtz-Eakin and Rosen (1989) and Aschauer (1989) on infrastructure; of Ladd and Yinger (1989) on fiscal and governance measures; and of Rose (1989) and Malpezzi (1996) on urban regulations.

Housing and Urban Development

The United States Department of Housing and Urban Development (HUD) is a cabinet department in the executive branch of the United States. Founded in 1965 and assigned to the United States Secretary of Housing and Urban Development, the HUD was created to develop and execute policies on housing and metropolises.

According to the HUD's own mission statement, it aims to create strong, sustainable, inclusive communities and quality affordable homes for all persons within the United States. The HUD is working to strengthen the housing market, to bolster the economy and to protect consumers; to meet the need for quality affordable rental homes; to utilize housing as a platform for improving quality of life; to build inclusive and sustainable communities free from discrimination; and transform the way HUD does business (HUD, 2013: 2).

More recently, several subprograms have been established, such as the 'HUD youthbuild programs' providing "disadvantaged youth with opportunities for employment, education, leadership development, and training in the construction or rehabilitation of housing for homeless

individuals and members of low- and very low-income families." (For more information: http://definitions.uslegal.com/h/hud-youthbuild-programs-hud/.)

The United Nations Centre for Human Settlements (UN-HABITAT)

The United Nations Human Settlements Programme, UN-HABITAT has been a pioneer organization in the collection of urban indicators. In 1978 UN-HABITAT was established through a resolution of the UN General Assembly to promote socially and environmentally sustainable towns and cities with the overall goal of providing adequate shelter for all. In 1991, UN-HABITAT initiated the Housing Indicators Programme, focusing on monitoring the quality of shelters. In 1993 the program was re-developed and became the Urban Indicators Programme, which focuses on a larger range of urban issues. In 1996 and 2001, the Urban Indicators Programme produced the first two global urban indicators databases the Global Urban Indicators Database Version 1 (GUID1) and the Global Urban Indicators Database version 2 (GUID2), presented at the HABITAT II conference and the Istanbul+5. The aim was to establish regional trends in key urban issues in a global perspective. In 2005, the Monitoring Urban Inequities Programme produced the Global Urban Indicators Database (III). In addition to addressing the HABITAT Agenda, this database provided information on the Millennium Development Goals, particularly for the target on achieving a significant improvement in the lives of at least 100 million slum dwellers. Since 2006, the Global Urban Indicators Database has been updated annually. (For the list of UN-HABITAT indicators see: http://ww2.unhabitat.org/programmes/guo/documents/urban_indicators_guidelines.pdf.)

The UN-HABITAT Monitoring System

The UN-Habitat monitoring system was established by UN-HABITAT in response to a decision of the United Nations Commission on Human Settlements. It called for a mechanism to monitor the global progress in implementing the HABITAT Agenda and to monitor and evaluate global urban conditions and trends at the international, regional, national, and local level. This monitoring function is provided through two main instruments: Global Urban Observatory (GUO) and Best Practices and Local Leadership programme (BLP). The GUO has the task to monitor and evaluate the global urban conditions and trends. It includes urban indicators, statistics, and city profiles. The BLP was established to make use of information and networks in support of the HABITAT Agenda Implementation. Both programmes operate under the Monitoring Systems Branch, which has the overall mandate to monitor the progress of the HABITAT Agenda and the Millennium Development Goals. (For more information see the UN monitoring system online at: http://ww2.unhabitat.org/programmes/guo/.)

Eurostat – Urban Audit

The Urban Audit is a statistical data collection by Eurostat providing statistical data and information on living conditions in Europe. After a 'pilot'-audit in 1999, the first full-scale European 'Urban Audit' for the then 15 countries of the European Union took place in 2003. In 2004, the project was extended to the 10 new Member States plus Bulgaria, Romania, and Turkey. For the 2003/2004 data collection exercise 336 variables were brought together, covering most aspects of urban life. The second full-scale data collection for the Urban Audit covered the time period between 2006 and 2007. It involved 321 European cities in the then 27 countries of the European Union along with 36 additional cities in Norway, Switzerland, and Turkey. The basic philosophy was to deviate as little as possible from the concepts used in the 2003/2004 collection. However, some changes were made with the aim to improve comparability, data availability, and quality (Eurostat, 2014). According to the 'EU – Cohesion Policy 2007–2013,' the newest data collection includes 357 European cities providing data with over 330 indicators for the quality of urban life in Europe cities on demography, housing, health, crime, the labor market, economic activity, income disparity, local administration, civic involvement, educational qualifications, cultural infrastructure, and tourism (European Commission, 2013: 1). Data collection has been assembled every 3 years. However according to the Eurostat Web site an annual data collection is being planned for a smaller number of targeted variables. (For more information the indicators and database: http://epp.eurostat.ec.europa.eu/portal/page/portal/region_cities/city_urban/.)

In addition to the GEO and the Urban Audit databases, there are many other databases on indicators related to housing and land in an international basis, such as the World Bank Data & Statistics, UNECE Environment and Human Settlements Division, the World Gazetteer, and the European Mortgage Federation.

The Future of Empirical Data on Environment, Human Development, and Urban Studies

A main task of the environmental and urban statistics work today is the harmonization and standardization of the definitions and classifications of indicators at the international level. Improved tests and amended theory at an international basis are critical to an enhanced understanding of the state of environment, urban markets, and human well-being. A 'standard' list of global indicators can not only help to monitor current global states and trends of living conditions, but it can also represent the impetus for a better understanding of divergences and communalities between regions, countries, and their citizen.

Currently, major international research effort was undertaken in most of the countries with support from numerous academics and other researchers, from the United Nations, the World Bank, the European Commission as well as number of national governments. In the light of these past developments, international data collection, collation, analysis and reporting could help us to monitor past and current trends of environmental and urban indicators and to forecast their possible development in the near future. Continued serious and globally concerted work on environmental, urban, and human developmental indicators will improve and illuminate public policy. Additionally it will enhance our basic understanding of the structure and dynamics of today's rapidly changing global environment, and its impact on living conditions.

See also: Development and Urbanization; Eco-cities and Sustainable Urbanism; Environmental Conservation and Development: Critical Perspectives; Environmental Policy: Protection and Regulation; Planning Issues and Sustainable Development; Urban Sociology; Urbanization.

Bibliography

Angel, S., Mayo, S.K., 1996. Enabling Policies and Their Effects on Housing Sector Performance: A Global Comparison. Paper Presented to the Habitat II Conference. Turkey, Istanbul.

Aschauer, D., 1989. Is public expenditure productive? Journal of Monetary Economics 23 (2), 177–200.

Case, K.E., Schiller, R.J., 1989. The efficiency of the market for single family homes. American Economic Review 79 (1), 125–137.

European Environment Agency, 1995. Europe's Environment. The Dobris Assessment. European Environment Agency, Copenhagen.

European Commission, 2013. EU – Cohesion Policy 2007 – 2013: Urban Development. http://ec.europa.eu/regional_policy/activity/statistics/2007_urban.pdf (accessed 12.01.14.).

Eurostat, 2014. Urban Audit Data Collections. http://epp.eurostat.ec.europa.eu/portal/page/portal/region_cities/city_urban/urban_audit_data_collections (accessed 12.01.14.).

Flax, M.J., 1978. Survey of Urban Indicator Data, 1970–1977. Urban Institute, Washington, DC.

Glickman, N.J., Lahr, M.L., Wyly, E.K., 1996. State of the Nation's Cities: America's Changing Urban Life. U.S. Department of Housing and Urban Development, Washington, DC.

Hens, L., 1996. The rio conference and thereafter. In: Nath, B., Hens, L., Devuyst, D. (Eds.), Sustainable Development. VUB Press, Brussels.

Hirsch, W.Z., 1966. Regional Accounts for Policy Decisions. Johns Hopkins University Press, Baltimore, MD.

Hochwald, W. (Ed.), 1961. Design of Regional Accounts. Johns Hopkins University Press, Baltimore, MD.

Holtz-Eakin, D., Rosen, H.S., 1989. The 'rationality' of municipal capital spending: evidence from New Jersey. Regional Science and Urban Economics 19, 517–536.

HUD, 2013. FY 2013 BUDGET. Housing and Communities Built to Last. http://www.google.hu/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCsQFjAA&url=http%3A%2F%2Fportal.hud.gov%2Fhudportal%2Fdocuments%2Fhuddoc%3Fid%3DCombBudget2013.pdf&ei=8D81U7jJIKnhywPFjIKgBQ&usg=AFQjCNHLrewX8uKNaftmjZwiL9pvFBHOqg&sig2=KYOKL3tv8g4nWI38qdEmNw&bvm=bv.63808443,d.bGQ (accessed 12.01.14.).

Klugman, J., Rodriguez, F., Choi, H., 2011. The HDI 2010: new controversies, old critiques. Journal of Economic Inequality 9 (2), 249–288.

Ladd, H.F., Yinger, J., 1989. America's Ailing Cities: Fiscal Health and the Design of Urban Policy. Johns Hopkins University Press, Baltimore, MD.

Malpezzi, S., 1996. Housing prices, externalities, and regulation in U.S. metropolitan areas. Journal of Housing Research 7 (2), 209–241.

Malpezzi, S., Mayo, S., 1997. Housing and urban development indicators: a good idea whose time has returned. Real Estate Economics 25 (1), 1–11.

Mayo, S.K., 1993. South African Housing Sector Performance in International Perspective. Paper presented to the 21st IAHS Housing Congress. Capetown, South Africa.

Mills, E.S., 1972. Studies in the Structure of the Urban Economy. Johns Hopkins University Press, Baltimore, MD.

Murphy, T., 1980. Urban Indicators: A Guide to Information Sources. Gale Research Company, Detroit, MI.

Muth, R.F., 1969. Cities and Housing. University of Chicago Press, Chicago, IL.

Rose, L.A., 1989. Topographical constraints and urban land supply indexes. Journal of Urban Economics 26 (3), 335–347.

Thibodeau, T.G., 1995. House price indices from the 1884–1992 MSA American Housing Surveys. Journal of Housing Research 6 (3), 439–479.

Tiebout, C.M., 1962. The Community Economic Base Study. Committee for Economic Development, New York.

UNDP, 2013. Human Development Report 2013-Technical Notes. The Rise of the South: Human Progress in a Diverse World. http://hdr.undp.org/sites/default/files/hdr_2013_en_technotes.pdf (accessed 12.01.14.).

United Nations, Department of Economic and Social Affairs, Population Division, 2013. World Population Prospects, the 2012 Revision. United Nations, New York.

United Nations Human Settlement Programme, 2004. Urban Indicator Guidelines. Monitoring the Habitat Agenda and the Millenium Development Goals. http://ww2.unhabitat.org/programmes/guo/documents/urban_indicators_guidelines.pdf (accessed 16.03.14.).

United Nations Development Programs, 2014. Human Development Reports. About Human Development. http://hdr.undp.org/en/humandev (accessed 16.03.14.).

World Bank, 1994. World Development Report 1994: Infrastructure for Development. Oxford University Press, Oxford and New York.

Data Bases and Statistical Systems: Health

Ilona Kickbusch, Graduate Institute of International and Development Studies, Genève, Switzerland
Claudia Stein, World Health Organization Regional Office for Europe, Copenhagen, Denmark

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D. Leadbeter, volume 22, pp. 15066–15072, © 2001, Elsevier Ltd.

Abstract

Health information systems represent a complex series of efforts to collect, analyze, and report information systematically and sustainably and to derive knowledge and evidence serving various purposes with the ultimate aim of improving health. The terminology used in relation to health information lists notions such as *health data, health information or health statistics,* and *knowledge or evidence*. The history of health statistics offers a picture of how the diversity of data has increased over time and how better health information has allowed scientists and health practitioners to reach better health outcomes. *Providing an evidence base* for overall policy and resource allocation, *ensuring access* (especially for vulnerable groups), *monitoring the quality of care* provided by the health system, *understanding overall public health* and the factors affecting it in different settings and *ensuring transparency and accountability* are the main purposes of health information systems. Examining the interactions between data used in national and international contexts is important in view of gaining better efficiency of health interventions. Considering the social determinants of health and their intersectoral nature prompts the need to gather data generated in other areas and sectors other than health. *Registration systems* such as civil registration and vital statistics are an important source of health information. Depending on the context, resources such as *registries, notification systems, health surveys, nonhealth sector sources* are also employed to obtain health statistics. Data integration is instrumental in monitoring health policy, monitoring and understanding of diseases, and better tracking individual health and well-being. However, ethical considerations and policy precautions are vital in data collection, storage, use and destruction, security and consent.

Introduction

Health information systems comprise the actions taken to collect, analyze, and report information systematically and sustainably and to derive knowledge and evidence in the field of health. This requires a dedicated infrastructure, including human resources, technology, software and databases. Many health information systems were born out of long-standing traditions, particularly in industrialized countries. These historic traditions have led to infrastructures and content which is very contextual and often nationally specific resulting in sometimes poorly comparable information elements at the international level. The national focus of health data can also make it difficult to analyze global crossborder flows. The greatest challenge remains the lack of good information systems in many developing countries.

This article will address these issues and aims to define the various aspects, content, and collection systems as well as purposes for health information. The national and the international challenges can no longer be separated in a globalized world, in particular, this relates to issues of ownership, commercialization, and governance of data, as well as the potential of the data revolution for the twenty-first century. Health statistics have gained significant political and commercial relevance.

Health statistics have been used in some form or another to assist with public health practice for hundreds of years, yet no universally agreed definition exists to date. While there have been numerous attempts to do so, the terms 'health statistics,' 'health data,' 'health information,' even 'informatics' are often used interchangeably. To complicate matters, what is collected and reported is information about deaths, diseases, and disabilities rather than information on health, well-being, and their determinants. Only recently have there been increasing attempts to provide measures of well-being by looking at people's material living conditions and quality of life across the population (OECD, 2013).

Terminology (What Are Health Statistics?)

In the context of its European Health Policy *Health 2020*, the WHO Regional Office for Europe (2013a) has recently made an attempt to clarify the definitions used for health information through a support tool for its Member States (WHO Regional Office for Europe, 2013b). The definitions proposed interpret health statistics in the widest sense as the content of health information systems:

- Health *data* (quantitative and qualitative) are the basic units of what is collected, for example, the raw numbers in birth and death registers, on the registrations of births or deaths; raw numbers on volume of medical care provided, numbers on medical care for provided in relation to a specific disease; or the results of a legislative measure banning smoking in public places.
- Health *information* or health *statistics* can be seen as organized data, that is, data that have been processed or arranged in such a way as to provide a clear picture of a situation in a particular place. For example, indicators based on data from various registers that have been collated and checked for omissions or duplications in order to produce complete information on births and deaths; analysis of data from a variety of sources in order to compile complete information about registration; comparisons of long-term trends or between geographic areas; analysis of primary care records to provide a picture of health workers' compliance with guidelines for medical care in relation to specific diseases.
- *Knowledge (or evidence)* can be seen as information in context, that is, information combined with understanding, values,

and a wider, longer-term approach to provide 'actionable information.' For example, an analysis that combines information from various sources on outcomes and their causes used to form the basis of action to reduce premature mortality; the results of an evaluation of a particular medical device that enables an analysis of its cost-effectiveness as the basis for a recommendation on its wider use; an analysis of the impact of implementing specific aspects of antismoking legislation as the basis for a recommendation on their inclusion in national legislation.

Historical Origins of Health Statistics

The historical origins of health statistics are closely related to the development of the modern nation-state and the increased political interest in systematically assessing the vital statistics related to the population (Porter, 1999). For the purpose of this article, four defining phases in the development of health statistics are identified:

Vital Statistics

It is often underestimated how much the history of public health is also the history of vital registration and statistics. The beginnings of death registration can be found in the mid-fifteenth century Italian city-states, in most developed countries they have been gathered for at least a century. A shift occurred when new methods and systems of information gathering were developed in the seventeenth century to better understand the interface of the wealth and health of nations. The groundbreaking work of John Graunt (1620–74) for example, included not only population estimates and (a historic first) a life table which calculated probabilities for survival, but also a proposal to create an early warning system for the onset and spread of bubonic plague in London. Fighting contagion and epidemics became one of the major tasks of states, leading to the establishment of boards of health and strict systems of medical policy, and laying the ground for what was to become one of the fundamental political debates in public health: the balance between individual rights and the demands of society (Baldwin, 2005).

Linking Health and Social Statistics

In the nineteenth century, the industrial revolution led to the deterioration of social and living conditions for large sectors of the population and the improvement of health became an integral part of a new social reform agenda. Enlightenment philosophers had considered health a human right and advocated social reform as a precondition for better health – this argument was increasingly underscored by data. For example the work of Louis René Villermé (1782–1863) and Rudolf Virchow (1821–1902) proved the association between poverty and mortality and Friedrich Engels (1820–95) showed that death rates for workers in large industrial cities were up to four times higher than in the countryside. Precise health statistics became an important political tool for reformers to argue their case in order to quantify the value of investments in better living conditions or investments in public health infrastructure and programs. Here too lie the origins of another long-standing public health debate between the moral value of health and the value of public health investments in cost–benefit terms. In setting policy priorities, health information systems became a governance tool for the state far beyond the initial focus on the control of contagion and disorder.

Sharing and Classifying Health Statistics

In the beginning of the nineteenth century it became critical to share information on disease and outbreaks. This need to collaborate – together with new international sanitary regulations – became a focus of the work of the International Sanitary Conferences as of 1851. At the national level in 1839, William Farr had already called attention to the importance of a uniform statistical classification of causes of death and he proposed a first disease classification for statistical purposes to enable precision in reporting. This was taken up by a proposal at the First International Statistical Congress at Brussels in 1853 to embark on the international statistical comparison of 'Causes of Death.' After extensive work and debate the International Statistical Institute adopted a report in 1893 which became the origin of the International List of Causes of Death, which since then has been regularly reworked and expanded and is now known as the International Classification of Diseases (ICD). In the 1920s, cooperation was developed with the Health Organization of the League of Nations, which had also taken an active interest in vital statistics, today this is one of the constitutional functions of the World Health Organization (WHO) in close cooperation with national statistical offices. The sharing of data – particularly when reporting outbreaks – continues to be a political challenge for global health.

Moving from Mortality to Disease Burden

Today's health information challenges are of a significantly more complex nature: as societies invest in the health of their population they need to have information "that is comparable, reliable and comprehensive across a wide range of conditions and exposures that cause death or ill-health in a population" (Lopez, 2005). In particular, moving beyond mortality and knowing the disease burden of mental health and chronic diseases such as depression or musculoskeletal conditions is critical for decision-makers. The *Global Burden of Disease* (GBD) Study was commissioned by the World Bank in the early 1990s and constituted a breakthrough in the international comparison of health data by using Disability-Adjusted Life Years(-DALYs), an approach which will be discussed further below. Health data have gained increasing political relevance, driving investment decisions, and access to development aid. They have also gained in commercial relevance for a global health industry and insurance industry.

Dimensions of Health Information Systems (for Whom and What Purpose?)

Different uses of health information systems require different kinds and content of information, and produced according to

different timescales and communicated in different ways. For effective learning, feedback needs to be rapid and relevant. For effective policy-making, data should provide feedback on a timescale that is relevant for decision-makers – which for most political systems is linked to election cycles. Health information about population health has different timetables, with slower-moving trends and a more complex process to collect and put together an overall picture. It serves a wide variety of purposes nationally and internationally and for many different stakeholders. Increasingly it brings together data from many different sectors.

Purpose of Health Information Systems

Health information systems serve the following purposes:

- Provide an *evidence base* for overall policy and resource allocation, with key users of such information being health ministries, finance ministries, social affairs ministries, and other parts of the government administration. For information to be used as *evidence base* it needs to be recent; showing comparative information in order to identify potential for improvement; and perhaps linking different data sources such as combining health data with cost or performance-related data in order to assess efficiency and effectiveness; in global health this can be illustrated by the introduction of results-based financing by the World Bank and the Global Fund for AIDS, Tuberculosis, and Malaria. This shift in approaches to aid, by providing funding *after* preagreed results have been achieved puts a high premium on the independent verification of results which requires high quality of data and robust measurement.
- *Ensuring access*, in particular for vulnerable groups, with users both at the overall system level and at the level of those working with the groups concerned. This kind of information has its own challenges, such as collecting data from different types of providers (e.g., special services for impoverished populations), as well as beyond the health system itself precisely in order to identify people falling outside it; this is a major data challenge in the developing world and new approaches to m-health (mobile phone technology used for health reporting in the field) are trying to address this issue.
- *Monitoring the quality of care* provided by the health system in order to ensure high standards of quality and safety, and be aimed at individual providers as well as overall monitoring agencies. This may involve a close link to professional monitoring processes such as audit and continuing professional development, and require fine-grained data down to the level of individual practitioners. Tracking quality of care may also require integrating data about a single patient across different episodes of care and different providers.
- *Understanding overall public health* and the factors affecting it in specific countries and their regions and for particular issues. This requires sources of data on key determinants, such as tobacco use, alcohol, diet, physical activity, social support, income, and employment. A wide range of sources of data from across government is required to gain a full picture.
- *Ensuring transparency and accountability*; this is often cited as a key role of health information, but is used in different senses, such as referring to access to records and information about their care for individual patients; benchmarking of health care providers; or overall transparency about the resources allocated to health systems and how they are spent to support political accountability.

Use at International Level

Health information is used not only at a national level, but is increasingly used to enable international comparisons on progress, set international priorities, provide accountability in relation to international commitments such as the Millennium Development Goals, and fulfill regular reporting requirements to international bodies such as the WHO or to donor agencies. Therefore, there is a myriad of stakeholders who make use of health statistics, most notably: international organizations, intergovernmental bodies at the regional level, development agencies, donors, policy-makers, supervisory and control bodies, the private sector, researchers, health professionals, professionals in other sectors, civil society groups, the media, the public, including subgroups of the population. Each user group will require information in all its forms – data, analytical information, and evidence-based knowledge – in order to play their part in improving health.

Information Generated in Other Areas and Sectors

The bulk of health information is collected within countries primarily for internal use. As such, it is for countries to consider what best meets their needs. The Health Metrics Network Framework (2014) and standards for country health information systems provide guidance on key elements of information systems to put in place, as well as principles for how to go about doing so. It is increasingly accepted that national health information systems should be built not only on demographic characteristics and on information generated within the health system, but also information generated in other areas and sectors. Some key intersectoral issues have been identified through other international work, notably the Commission on Social Determinants of Health (2008), and their report includes recommendations on generating relevant data and information. This has recently been more closely defined for the European Region through the Social Determinants Review for Europe (UCL Institute of Health Equity and Marmot, 2013). The wider environment is also relevant to health, and climate change in particular may bring changes that affect health; health information systems will need to take this into account, such as through including monitoring of changing patterns of threats or warning systems for extreme events.

Comparability at International Level

In order to ensure that information systems can support comparability at international level, however, it would be useful to identify areas of shared interest for international comparison across countries. Agreement on such areas will enable information systems to converge and provide

Table 1 Potential areas of interest for international comparisons

Population health and well-being
● Life expectancy and healthy life years
● Mortality by age group and condition
● Morbidity
● Avoidable/premature mortality
● Population risk factors (including inequalities)
● Subjective and objective well-being
Equity
● Distribution of health status by population/demographic/social groups
● Distribution of access/utilization of health services
● Progressivity of financing system
Health service outcomes
● Performance and quality of different elements (preventative care, primary care, secondary care, long-term care, mental health)
● Health system outcomes
● Health system processes
● Responsiveness
● Financial protection
● Efficiency

comparable information in support of improvement and mutual learning between countries. There are many potential frameworks for such international comparison. A recent review and synthesis of work in this area (Papanicolas and Smith, 2012) has identified a range of dimensions for health and health systems and potential areas of shared interest for comparison, and potential areas of interest for international comparison based on this framework are set out in **Table 1**. This analysis also includes discussion of the main possible indicators for each of these areas, together with their potential uses and limitations.

What Kind of Information?

Countries' health information comes from a huge variety of sources and covers different types of information. Health statistics are often derived from institutionalized *registration systems*, particularly in industrialized countries. This includes *civil registration* (births, marriages, divorces, and deaths) and *vital statistics* (causes of death); all countries should aim to have such registration systems covering the entire population. Ideally, vital statistics registration should operate using the ICD 10th version to permit relevant analyses and international comparisons. In addition, civil registration is key to a person's identity and nationality, human rights, and ability to access benefits. Such registration systems should be fully electronic and integrated and form the absolute basis of a country's health information reporting. Around 80 countries still do not have functioning civil registration systems; most uncounted births and deaths occur in sub-Saharan Africa and Southeast Asia. As a result, the WHO and the Health Metrics Network (HMN) have initiated an action plan 'to make every life count.'

Many countries operate special *registries* including those for cancer, immunizations, communicable diseases (such as tuberculosis), birth defects, disabilities, treatment, and exposure registries (radiotherapy, occupational hazards, etc.) which supplement mortality records with morbidity and disability information. While mortality information tends to be the most reliable where a sustainable system has been established, registries often suffer from underreporting and data quality issues; in addition, registries and vital statistics systems are not always integrated well, making the linkage of information between these systems difficult.

Notification systems are used in many countries for the surveillance of 'notifiable' diseases, most often communicable diseases but also chemical and physical hazards. Such systems are not intended to describe burden of disease but are mostly used for prevention and control, i.e., to identify potential increases in incidence and outbreaks and assist with the control of such events. Notification systems capture only a proportion of cases (often in a delayed manner) and differences in disease definitions used make comparisons over time and space difficult. The responsibilities that come with the adoption of the International Health Regulations require rapid implementation.

In many countries, *health surveys* are a source of health information particularly for nonfatal health outcomes. Such surveys are often population-based, that is, cover a representative portion of the entire population. They may include information and/or measurements of health status and health determinants, and information on access to and use of health services, as well as health service costs. Surveys can also obtain information from private health care providers that are often not available to public sources. Some countries run regular surveys (including the European Union Member States), those that operate the Demographic and Health Surveys and the Multiple Cluster Index Surveys, to name but a few. These surveys are of huge value but should not replace the establishment of civil registration and vital statistics systems in countries.

Information relevant to health may also be obtained from *nonhealth sector sources*. In many countries, the civil registration data are under the governance of ministries of the interior or justice and therefore already outside the health sector. Social services may provide information on social determinants of health (including social support, welfare, child protection), the employment sector information on unemployment rates, the public safety sector on road traffic accidents and violence, the justice system on crime rates, the education sector on school enrollment, education achievement, etc. Such data are becoming increasingly relevant where countries adopt a whole-of-government approach to public health; unfortunately, these information systems are rarely integrated and data linkage is the exception. Accessing and analyzing such information is therefore time consuming and challenging.

Data integration is a key component for the understanding of population health and for successful monitoring of health policy. Individual data and record linkage are practiced in a number of Northern European countries where every individual is given a single identifier at birth which is used in all national information systems throughout this person's life (and beyond). This enables a unique way of monitoring a person's well-being during the life course and longitudinal follow-up in the population. Such linkage does, however, increase the sensitivity of these data and privacy protocols must be adequate to ensure that no misuse occurs.

Institutions Providing Health Statistics

Many public and private players have emerged in recent years to supply health statistics, particularly at the international and regional level. Traditional institutions include large international organizations such as the United Nations, the WHO, UNICEF, the World Bank, the EU, and the European Commission and the Organisation for Economic Cooperation and Development (OECD). These institutions try to act as knowledge brokers and support countries across the three phases of collection, analysis, and use of health information, as well as bring countries together.

Collection

Although the primary focus for collection is by its nature within countries, the WHO supports countries by providing technical expertise and acting as a focal point for sharing technical expertise between countries, including knowledge, good practice, and strategies for eHealth. "eHealth is the use of information and communication technologies (ICT) for health. WHO works with partners at the global, regional and country level to promote and strengthen the use of information and communication technologies in health development, from applications in the field to global governance" (WHO – eHealth, 2014).

At global level, WHO plays a vital technical role in establishing and maintaining international classifications, most notably the ICD (WHO Family of International Classifications, 2014). These classifications provide essential support for the collection of health information within countries, as well as for its comparable analysis at the international level. The WHO Regional Office for Europe also maintains a range of databases, in particular the European Health for All Database as well as others on more specific topics (WHO Regional Office for Europe – Data and evidence – Databases, 2014).

Analysis and Interpretation

The World Health Reports and the European Health Reports, as well as other annual statistical publications of the WHO provide key mechanisms for analysis and interpretation of health information, in particular comparative analysis across the globe and regions. The databases provided its electronic health information portal also constitute a core resource enabling comparative analysis by countries and third parties, and are an international reference point for comparative analysis in health.

Ensuring Evidence-Informed Policy-Making

The WHO can also support the process of bringing knowledge and evidence into policy. WHO policies and publications are key to the dissemination of health information and play a vital role in crystallizing analysis and emerging consensus on health issues. The WHO can provide an authoritative external voice in support of evidence-informed policy discussions within countries. Structures such as the Health Evidence Network (WHO Regional Office for Europe – Health Evidence Network, 2014) and the Evidence-Informed Policy Network (WHO Regional Office for Europe – Evidence-informed Policy Network, 2014) bring capacity building and tools for knowledge translation directly to countries.

The European Commission and OECD are major producers and users of health statistics. While all Directorate-Generals of the European Commission are involved in data collection and analysis, the main thrust rests with Eurostat and the Directorate General for Health and Consumer Protection. This includes regular European Union wide surveys such as European Health Information Surveys, European Health Examination Surveys, and surveys conducted by Eurostat, such as the European Surveys on Income and Living Conditions. The OECD may collect its own data, or use data from WHO Europe, to produce the Health at Glance series. The latter is done to a large extent with WHO and Eurostat through the joint data collections on nonmonetary health outcomes.

What Is Required for Countries to Have Functioning Health Information Systems?

Institutional

An essential requirement is some form of national infrastructure with the capacity to oversee and coordinate a health information strategy to systematize and integrate the information; the independence is to ensure that the information and analyses are accepted as credible and authoritative; and the capacity to ensure that the necessary expertise and skills can be developed and retained. The structure might take the form of a national agency or institute, or the task might be entrusted to an academic institution or partnership of various institutions. Whatever the form, some central, independent, sustainable structure for health information is essential for the long-term success of a health information strategy; it must ensure the adherence to international agreements on data provision.

Legal

The collection of information about health is sensitive and is rightly subject to legal protection. On the one hand, the privacy of personal information must be safeguarded; on the other hand, there is a clear public interest in collecting and disseminating information that can help to improve health now and in the future. Many countries and the European Union have specific legislation on this issue, which ensures sustainability, harmonization, and compliance.

Many countries have a well-established legal framework that balances these different objectives. Changes in data processing brought about by modern information and communication technology, however, have led many countries to review their frameworks to ensure that they are adapted to current circumstances. There is a risk that, in seeking to protect individuals against unduly use of their personal data online, well-founded health information systems could be restricted if the consequences of potential misuse of data are not fully understood. It is essential that countries ensure that their legal frameworks for data collection and processing strike the right balance, allowing health information to be collected, analyzed, and communicated while protecting the privacy and confidentiality of individuals. This is pertinent to many aspects of

data collection (for example, access to patient data), information sharing, (for example, tracking who has access to what data and information, ensuring secure access and transmission) and policy (for example, ensuring that any publicly available information is anonymous).

Health professionals also need protection in the use of data. Monitoring systems for quality and safety rely on accurate, honest reporting of data by professionals, but this may be compromised if honest reporting by well-intentioned professionals could result in punitive sanctions. Here again, an appropriate balance must be found in each country to facilitate open, accurate reporting while ensuring that appropriate corrective action is taken when necessary. These types of data are especially critical for comparative international work on effectiveness and efficiency of health systems.

Technical

Data Collection Systems

The HMN describes the principal sources of health data that should constitute the core of national health information structures. Other less typical sources could also be considered, such as data from other government sectors (for example, social security benefit entitlements, justice system records) or from nongovernment actors, such as purchasing patterns (for example, alcohol sales). Health information strategies should include consideration of what data and information are essential and how to ensure their availability. At the European level, the aim should be to identify a minimum data set for core health information and a set of harmonized indicators to ensure comparability at the regional level.

Analytical Capacity

Additional work is required to turn data into meaningful information and knowledge. An essential part of a health information system is therefore the recruitment, retention, and training of skilled analysts who can do such work consistently over the long periods of time required to analyze health data series, for example. This requires a stable structure for health information. Analysis also depends on continual contact with the wider community of analysts, in particular in the academic world, in a multidisciplinary approach. Making health data available to external analysts (with appropriate safeguards to ensure data protection) can help to ensure up-to-date, independent analysis. International collaboration and comparative analyses can bring valuable additional perspectives and ensure that the analyses reflect international best practice.

Information Technology

One of the major changes in recent years has been the use of information and communication technology to collect, process, and communicate information in new ways. Ensuring that health information systems make good use of this technology is an essential component of health information strategies and their development. An example of using technology in health is the increasing availability of electronic medical records. Such technology may also reduce the administrative burden of health information systems. Interoperability is a central goal within and between systems, to ensure the generation of accurate data and to support good provision of health care.

The omnipresence of technologies such as mobile phones may also allow for the collection and provision of health information in new ways. Realizing the potential of information technology for health information will require that appropriate mechanisms be built in to systems during their development, which will require the active involvement of people responsible for health information systems in wider eHealth initiatives. This has shown particular innovative potential in developing countries.

Resources

Health information systems represent extremely good value for money and health information strategies should acknowledge this. These systems provide the information essential to manage systems, ensure that health system resources are used efficiently and effectively and improve health overall. For a tiny fraction of their total cost; for example, the proportion of national spending on cancer registries in Finland in 2007 was only 0.0037% of total spending on cancer (Commission of the European Communities, 2009). Moreover, health information systems are a prerequisite for good governance, accountability, and transparency. Yet, it can still be difficult to secure the necessary resources for health information strategies and systems particularly in developing countries. The investment is inherently long term when compared with immediate needs for clinical services, for example. Nevertheless, investment in health information is vital if the much larger resources for health care are to be used effectively.

The Dynamics between National and Global Health Actors in Collecting and Sharing of Health Data

Data are playing an ever larger role in global health – some speak of a data revolution. Yet challenges to the comparability of data, the quality of data, and the reliability of data when countries or agencies wish to propose policies based on such reporting still need to be addressed. When the political stakes are high, for example, in reporting on the Millennium Development Goals by country, differences in estimations between various agencies exist and are extremely confusing for users and policy-makers. The WHO, World Bank, UNICEF, and the UN Statistics Division have therefore joined forces and established an Interagency Group on Mortality Estimates to produce joint estimates for infant, child, and maternal mortality. These, however, are not entirely consistent with estimates produced by Institute of Health Metrics and Evaluation (IHME) in Seattle, and it is important for these institutions to collaborate closely to avoid such discrepancies.

WHO, UNICEF, the World Bank, and the UN Statistical Division, among others, provide consistent health information. The regular World Health Reports, issued by the WHO since 1995, the World Development Report of the World Bank (very notably the 1993 report on burden of disease) and the GBD Report 1990 (published in 1996) are major examples of such reporting. These traditional data providers have been joined by many other players in the past decade, particularly those who

operate using complex estimation methods for burden of disease, such as the IHME based in Seattle, USA, which uses the burden of disease approach to estimate DALYs for all countries in the world. Numerous other consortia and commissions, such as the Lancet Commission on Investing in Health (Jamison et al., 2013) have also entered the stage providing and appraising health information.

Countries in the developing world will need international support to build institutions and health management information systems that can produce accurate, unbiased data. For population health this means building national statistical systems and improving not only the measurement of fundamental statistics such as births and deaths and levels of mortality, but also intersectoral data with high relevance to health such as economic growth and poverty, taxes and trade, the environment, and education (Center for Global Development and The African Population and Health Research Center, 2014). As developing countries expand their health systems, metrics that measure quality of care are also needed. Though tools such as the Health Care Quality Indicators project of the OECD and the WHO Performance Assessment Tool for Quality Improvement in Hospitals are helpful, others will need to be developed by making use of new technologies.

Management by objectives has become a key feature of global health programs push for data and health information systems. The development of comprehensive global monitoring frameworks for a range of diseases, which include a set of targets and indicators to track progress toward global and national goals, are becoming a standard procedure. Most recently, a framework was adopted to monitor national and global progress in implementing strategies for the prevention and control of NCDs. This includes monitoring exposures (risk factors and social determinants), outcomes (morbidity and mortality), and health systems capacity and response to health crises (Ijaz et al., 2012). In a similar vein, a wide range of other focused global programs have set targets to measure impact and success. Starting with the goals set for HIV AIDS, tuberculosis, and malaria in the Millennium Development Goals further targets have been set by the global tuberculosis strategy and the global malaria action plan. The global health community has also set goals for the elimination and eradication of a range of diseases, at present for polio eradication. As this program comes to a close, a complex system of surveillance is required in areas at particular risk of missed transmission including targeted community searches, six-monthly active case searches, and case searches during vaccination (Global Polio Eradication Initiative, 2013).

New Developments and Challenges

As the move toward improving the health information systems in the developing world gathers pace, new opportunities, and challenges have emerged. Opportunities consist of the fact that better health information systems can help countries reach critical objectives in health care and the global community to achieve common goals. Better surveillance can enable rapid response to health crises and bigger sets of data have the potential to provide a better understanding of health and its determinants. Challenges arise from concerns that more and more health data are being gathered by corporations, both in developed and developing countries. Consequently, to prevent misuse of data, more policies need to be directed toward the ethics of data collection, storage, use and destruction, security, and consent (Harrison, 2014).

Another new development is the creation of digital open access platforms. Such a platform is the global health research process map by the Global Health Network, which allows for sharing of research and information (The Global Health Network, 2014).

New visualization tools have been a major breakthrough in recent years to display data. The Global Health Observatory at WHO headquarters and the WHO Regional Office for Europe provide numerous atlas-based, and easy-to-use visualization options, including the Atlas of Inequalities in Europe. IHME in Seattle has an altogether different but very effective range of visualization tools, which require a more complex IT setup than those provided by WHO. Scientists, policy-makers, and the general public all have immediate access to these tools. In addition, 'information about the information' or 'data about the data' are given to users to help them appraise the information; these 'metadata' are becoming more and more the standard in health information reporting. However, disaggregation of data by age, sex, vulnerable groups, etc, is still lacking in many countries (and particularly at the subnational level) and is a very important element for the analysis of inequalities in populations.

The provision of 'real-time' data, that is, the collection, analysis, and display of data as they are produced is still some time away from practical application despite the revolution in eHealth and IT solutions. This is largely a reflection of the need for quality assessment of the data, which often requires 'cleaning up' due to data entry errors or misclassification. As a result, time lags or delays invariably occur and are required to ensure that the user only receives correct information.

The fundamental challenge in the current era of global health will not be technological, but rather be better governance and accountability for common health goals by the wide range of public and private actors. The greatest global and national challenge at hand is failing to act on the data that could save lives and improve health.

See also: Evidence-Based Medicine; Functioning, Disability and Health, International Classification of; Globalization and Health; Health Outcomes, Assessment of; Health Surveys; Health: Definitions; Socioeconomic Status and Health.

Bibliography

Baldwin, P., 2005. Contagion and the State in Europe, 1830–1930. Cambridge University Press, Cambridge.

Center for Global Development, The African Population and Health Research Center, 2014. Delivering on the Data Revolution in Sub-Saharan Africa. Final Report of the Data for African Development Working Group [Online]. Center for Global Development, Washington. Available from: http://www.cgdev.org/sites/default/files/data-african-development-working-group.pdf (accessed 08.07.14.).

Commission of the European Communities, 2009. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on Action against Cancer: European Partnership [Online]. COM, Brussels. Available from: http://ec.europa.eu/health/archive/ph_information/dissemination/diseases/docs/com_2009_291.en.pdf (accessed 08.07.14.).

Commission on Social Determinants of Health, 2008. Closing the Gap in a Generation: Health Equity through Action on the Social Determinants of Health. Final Report of the Commission on Social Determinants of Health [Online]. World Health Organization, Geneva. Available from: http://www.who.int/social_determinants/thecommission/finalreport/en/ (accessed 08.07.14.).

Global Polio Eradication Initiative, 2013. "Objective 1: Poliovirus Detection and Interruption" in Polio Eradication & Endgame Strategic Plan 2013–2018 [Online]. World Health Organization, Geneva, pp. 31–48. Available from: http://www.polioeradication.org/Portals/0/Document/Resources/StrategyWork/PEESP_EN_US.pdf (accessed 08.07.14.).

The Global Health Network, 2014. Home page of the global health network [Online]. Available from: https://tghn.org/ (accessed 09.07.14.).

The Health Metrics Network Framework, 2014. Home page of world health organization [Online]. Available from: http://www.who.int/healthmetrics/documents/framework/en/ (accessed 08.07.14.).

Harrison, O., 2014. 'The Data Revolution for Global Health,' Global Health and Diplomacy [Online]. Available from: http://www.ghdnews.com/index.php/global-health-partnerships-and-solutions/technology/61-the-data-revolution-for-global-health (accessed 08.07.14.).

Ijaz, K., Kasowski, E., Arthur, R.R., Angulo, F.J., Dowell, S.F., 2012. International health regulations-what gets measured gets done. Emerging Infectious Diseases 18 (7), 1054–1057.

Jamison, D.T., Summers, L.H., Alleyne, G., Arrow, K.J., Berkley, S., Binagwaho, A., Bustreo, F., Evans, D., Feachem, R.G., Frenk, J., Ghosh, G., Goldie, S.J., Guo, Y., Gupta, S., Horton, R., Kruk, M.E., Mahmoud, A., Mohohlo, L.K., Ncube, M., Pablos-Mendez, A., Reddy, K.S., Saxenian, H., Soucat, A., Ulltveit-Moe, K.H., Yamey, G., 2013. Global health 2035: a world converging within a generation. Lancet 382 (9908), 1898–1955.

Lopez, A.D., 2005. The evolution of the global burden of disease framework for disease, injury and risk factor quantification: developing the evidence base for national, regional and global public health action. Globalization and Health 1, 5. http://dx.doi.org/10.1186/1744-8603-1-5.

OECD, 2013. How's Life? 2013: Measuring Well-being. OECD Publishing. https://dx.doi.org/10.1787/9789264201392-en.

Papanicolas, I., Smith, P.C., 2012. Health System Performance Comparison: An Agenda for Policy, Information and Research [Online]. World Health Organization, Copenhagen. Available from: http://www.euro.who.int/__data/assets/pdf_file/0010/162568/e96456.pdf (accessed 08.07.14.).

Porter, D., 1999. Health, Civilization, and the State: A History of Public Health from Ancient to Modern Times. Routledge, London.

UCL Institute of Health Equity, Marmot, M., 2013, updated reprint 2014. Review of Social Determinants and the Health Divide in the WHO European Region. Final report [Online]. World Health Organization, Copenhagen. Available from: http://www.euro.who.int/en/health-topics/health-policy/health-2020-the-european-policy-for-health-and-well-being/publications/2013/review-of-social-determinants-and-the-health-divide-in-the-who-european-region.-final-report (accessed 08.07.14.).

WHO – eHealth, 2014. Home Page of World Health Organization [Online]. Available: http://www.who.int/ehealth/en/ (accessed 08.07.14.).

WHO Family of International Classifications, 2014. Home Page of World Health Organization [Online]. Available from: http://www.who.int/classifications/en/ (accessed 08.07.14.).

WHO Regional Office for Europe, 2013a. Health 2020. A European Policy Framework and Strategy for the 21st Century [Online]. World Health Organization, Copenhagen. Available from: http://www.euro.who.int/en/health-topics/health-policy/health-2020-the-european-policy-for-health-and-well-being/publications/2013/health-2020-a-european-policy-framework-and-strategy-for-the-21st-century (accessed 08.07.14.).

WHO Regional Office for Europe, 2013b. The Twenty-first Standing Committee of the WHO Regional Committee for Europe (SCRC) Report – Second Session–16–17 December, Floriana, Malta [Online]. World Health Organization, Copenhagen. Available from: http://www.euro.who.int/__data/assets/pdf_file/0006/244275/63s2e00_FinalReport_140006.pdf?ua=1 (accessed 09.07.14.).

WHO Regional Office for Europe – Data and Evidence – Databases, 2014. Home Page of WHO Regional Office for Europe [Online]. Available from: http://www.euro.who.int/en/data-and-evidence/databases (accessed 08.07.14.).

WHO Regional Office for Europe – Evidence-Informed Policy Network, 2014. Home Page of WHO Regional Office for Europe [Online]. Available from: http://www.euro.who.int/en/data-and-evidence/evidence-informed-policy-making/evidence-informed-policy-network-evipnet (accessed 08.07.14.).

WHO Regional Office for Europe – Health Evidence Network, 2014. Home Page of WHO Regional Office for Europe [Online]. Available: http://www.euro.who.int/en/data-and-evidence/evidence-informed-policy-making/health-evidence-network-hen (accessed 08.07.14.).

Data Bases and Statistical Systems: International Comparative Research

Markus Quandt, GESIS Leibniz-Institute for the Social Sciences, Cologne, Germany
Ruud Luijkx, Tilburg University, Tilburg, the Netherlands; and University of Trento, Trento, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In comparative research, obtaining valid and reliable data is more challenging than in most other research techniques in social sciences. Presenting the European Values Study as an example, this article sketches how the field has established comparative databases that increasingly address these challenges. Examples of selected comparative databases at different analytical levels (micro versus macro) and from different source types (academic versus official statistics versus nongovernmental organization) are provided. A brief outlook on remaining and newly arising methodological and organizational issues concludes this article.

Introduction

The Need for Comparative Research

Many research questions in the social sciences are of comparative nature, explicitly or implicitly: investigating the different effects of national welfare regimes on inequality or the effects of particular electoral systems on the extent of participation obviously requires that researchers observe variation across different institutions, and across potential explanatory factors. Sufficient variation across institutional settings can be found either by looking at different countries, regions, and cultural areas, or by looking at changing institutions in the same country over long time spans. Ultimately, the dimensions of time and of space provide the variation needed for empirical testing (Przeworski and Teune, 1970). If both are combined, a space-by-time matrix is obtained. This is useful for assessing developmental hypotheses that predict particular trajectories over time, depending on characteristics that vary on the 'space' dimension (Franklin and Wlezien, 2002). In a nutshell, empirical research at the level of societies almost always requires comparisons across different societies, and across different states of one and the same society.

Macro and Micro Data

Data that describe units at the level of nations or societies are commonly referred to as macro data. Beyond singular questions about particular institutional arrangements etc., analyses of macro data are relevant to testing many 'grand' sociological, political, or economic theories, for example, theories dealing with welfare or political regime types, modernization, and economic development. But data at the individual- or microlevel are often equally relevant in informing 'grand' theories. Value change, secularization, and political processes with citizen involvement all have in common that many of their empirical statements directly refer to the microlevel, therefore making micro data indispensable for empirical work. It is typical for sociological and political theories to touch on both levels, so the question of empirically depicting connections between the macro- and microlevel is one of the most relevant (Coleman, 1990; Erikson, 2002). But even pure microlevel theories (such as psychological theories on personality traits or economic theories on labor-market behavior) benefit from empirical underpinnings in a comparative perspective, because cross-cultural comparisons of within-country microlevel mechanisms speak of the degree of universality that such theories can claim (or not).

Data and methodologies have long not been up to par with the needs of theories that claim an international or even universal scope. Many early comparative studies displayed a certain arbitrariness in the selection of the usually few macrounits they observed and compared, simply because there were no other options for obtaining comparative data but using personal networks or other idiosyncratic options. The independently collected data from these macrounits were then analyzed as a set of parallel case studies (see Kohn (1989) as an example of the genesis of data collection options and the analysis design). The increasing availability of systematically collected comparative data has allowed researchers to match theoretical scope and empirical underpinnings much better, so that certain country-level characteristics can be analyzed for all countries. Often, matching microlevel data are there as well, to build a full multilevel analysis.

Methodology itself has moved from a parallel case study approach to formalized methods looking for statistical confirmation at all levels of observation, with methods such as cross-sectional time series analysis or multilevel regression models. Where still only small numbers of cases are present, the method of qualitative comparative analysis (Ragin, 1994) has become a popular option.

Of course, the development of analytical approaches and the availability of data are tightly wed. Improving the database for comparative research increases the opportunities for such research, and comparative research at the society level in turn contributes to extending the database. Accordingly, both the use of existing databases and the creation of new comparative data have become routine exercises for research communities. This article will therefore only present a limited and intentionally heterogeneous selection from the wealth of existing comparative databases, trying to identify some typical aspects on a few examples that are more deeply elaborated. Readers in need of more complete coverage should consult the relevant cross-references.

Comparative Databases: Definition

By the term *database*, the authors refer to any data file or set of data files that has/have been systematically prepared to support empirical research. A *comparative database* is therefore any database that helps comparative research as described above, by allowing a comparison across the boundaries of at least two units at the macrolevel, regardless of the data itself being micro- or macrolevel data. Under this definition, almost any research project that collects, merges, or edits data to be used in comparative analyses constitutes a new database. Therefore in the remainder of this article, the authors discuss only comparative databases that are intended to serve as research infrastructure, accessible not only to the person(s) that compiled them but also to a broader community of researchers. As the compilation of comparative data almost always requires some form of international collaboration, building such a database usually constitutes a major investment that is more easily justified when wide and enduring usage of the data can be expected. This commonly implies that such databases should be, and often are, used for a variety of purposes and research questions rather than for singular, one-off projects. It also implies that the organizational framework for providing access to the data is geared to a longer time frame.

Starting Ground and Some Examples

Availability of Data

The data that feed into comparative databases come from many sources that can be grouped into academic data collection efforts, national and supranational agencies, and Nongovernmental Organizations (NGOs). Nonacademic agencies such as statistical offices, the Organisation for Economic Co-operation and Development (OECD), different branches of the United Nations (UN), or NGOs provide access to their data in order to meet their legal mission and to support their policy goals. Consequently, macro data from these organizations, e.g., aggregate figures on national economic performance or levels of health care provision, have for decades been accessible to research. If restrictions existed, they mainly lay in the area of lacking documentation, or in data structures that were much less suitable for research than for administrative purposes. Within academic research, in most disciplines, sharing data with third parties has been less common, and only very recently, research funding agencies make the provision of access to the collected data a binding obligation in their grant agreements. However, social sciences have always been among the pioneers with respect to 'data sharing.' On the international level, especially a number of early comparative social science projects were forerunners for establishing research databases as infrastructures, with sociology and political sciences being served by the same institutions and partly the same data (1,2).

Obtaining Comparative Macro Data

With regard to macro data, in the early times, data sharing took the form of providing tables and lists in printed publications, as exemplified in one of the most visible early databases, the data on the stability and capabilities of states coded and compiled by Karl W. Deutsch (1960). In that heavily cited example, only 11 quantifiable characteristics of states are coded that pertain to one single point in time, mostly taken from various UN publications of the 1950s. This already signifies one typical feature of quantitative comparative macro databases: behind the final data outcomes, there are long chains of production, often from surveys administered by national statistical offices, over aggregations at national level, to further aggregations done at supranational level, to compilations and selections done for research projects. In contrast to that, macro data that are not analytical aggregates of micro data can be of a more direct observational and partly qualitative nature. Expert judgments of published documents and media reports inform projects like the Polity database that was founded in the 1960s and is now in its fourth development stage 'Polity IV,' (Marshall and Gurr, 2012). The Polity databases compile data about the political stability or fragility of states, the occurrence of violent conflict, and the state-level factors contributing to both. Both the microaggregation and the expert assessment types of project have in common that – at least in their nascent phase – they can still be within the realm of individual researchers or small teams because the compilation of already published quantitative data or document analysis do not always require systematic international collaboration and large teams.

Obtaining Comparative Micro Data

This contrasts with projects that collect data at the microlevel in a comparative fashion. Conducting comparative surveys has always been a major effort in terms of coordination and collaboration, not the least part of that being raising funds for fieldwork from the respective national funders in a coordinated fashion.

With dispersed funding structures and general organizational overhead being major barriers for the implementation of comparative surveys, it is no surprise that the first examples of comparative micro data collection efforts either used a very loose organizational framework or were spin-offs from other work conducted by sufficiently resourceful agencies. Probably the first instance of a (partly) comparative survey series were the so-called United States Information Agency (USIA) surveys, conducted in France, Italy, the United Kingdom, and West Germany between 1952 and 1976. They were centrally organized and funded, avoiding many problems in international coordination. The carrying agency was the USIA that basically pursued the goal to inform US foreign policy about strategically relevant attitudes of the populations in other countries (the USIA data are available through the GESIS Data Archive). The first truly academic comparative projects are often reported to have been the study on 'Patterns of Human Concerns' (3) in 14 countries across the world, the 'Civic Culture Study' (Almond and Verba), and the 'International Time Use Survey' (4). More complete overviews can be found in Refs. 1, 5, and 2. Even this short list of studies covered a wide scope of topics: the different general expectations that people had for the future, the orientations they had toward political participation, the ways they allocated time to everyday activities. One common driving feature in that heterogeneity was the researchers' interest to gain an initial insight into the degree and quality of international

diversity in the human condition – how different or similar are the lives of humans in different parts of the world? After this need was increasingly satisfied, the following two decades saw a slowly increasing degree of ambition and rigor in the implementation of comparative surveys. The first comparative data collection projects in the 1960s had largely collected data in those places where the opportunities for collaboration happened to be, whereas the next wave saw some efforts that were striving for completeness of coverage in some sense, most notably the European Communities Study, which in 1970 was installed as the forerunner of the Eurobarometers of the European Union. In 1973, it was extended to include all member countries of the European Community at that time, in order to better fulfill its intended main purpose of monitoring the state of European integration in the eyes of the public, and as input for European policy makers. The model of the Eurobarometer – which is repeated in regular intervals and served to provide the empirical base for Inglehart's famous hypothesis of the 'Silent Revolution' – also makes clear how comparative survey series can become tools for the analysis of societal change over time, provided that at least some parts of the questionnaires are standardized (Inglehart, 1977).

Data Archives

With these comparative surveys constituting relatively large investments, the community from the beginning strove to establish mechanisms for wide access by researchers and found these in the emerging survey data archives (Mochmann, 2008). Although institutions such as the Inter-University Consortium for Political and Social Research (ICPSR, University of Michigan, founded in 1961) or the Central Archive for Social Research (in Cologne, Germany, founded in 1960, now GESIS Data Archive for the Social Sciences) were organizations with a primarily national foundation and purpose, they provided basic services for archiving and dissemination that made the wide and continuing use of the comparative data sets from sociology and political science possible. They thus added a dimension of long-term sustainability to the comparative projects without which the cumulative expansion of these would not have been possible. One aspect of this is the built-in institutional stability of the archives, which is achieved either by a network-based funding (as for the ICPSR) or by long-term government funding (as for the GESIS archive and several other European data archives).

The European Values Study

Creation of the European Values Study

The step to data collection projects becoming long-standing infrastructures with an identity of their own is easiest when the data access providers come to play an active role in the overall projects. The authors use the case(s) of the European Values Study (EVS) and the World Values Survey (WVS) as illustrative examples for the trajectories that large infrastructural projects have taken over the past decades, and how they adapted to changing environments. The EVS has its origins in the late 1970s and was started from the Universities of Tilburg and Leuven as a project to study values and norms in European societies. Its first survey was fielded in 1981 in 14 European countries, but also in Canada and the United States. The study eventually evolved into one of the few really long-lasting surveys: the latest survey was fielded in 2008 in all European countries with a population over 100 000.

In the first survey from 1981, the main value domains studied were family, work, religion, politics, society, Europe, and well-being. The survey was coordinated by Gallup from London. Face-to-face interviews were held with random samples of at least a size of 1000. Identical (translated) questionnaires were used. Data were made available for the public through the data archive at the University of Essex (now UK Data Archive).

The second round of surveys was held in 1990 and there was an expansion to many countries in Eastern Europe (partly coordinated by Hans-Dieter Klingemann). The timing of this survey coincided with the fall of the Wall and actually accounted for a 'zero measurement' of the Eastern European societies at the moment of the decline of the communist regimes. The 1990 round was coordinated by the Institut für Demoskopie at Allensbach am Bodensee. From this wave on, National Program Directors got a more important coordinating role. The data were archived at the Steinmetz Archive in Amsterdam (now Data Archiving and Networked Services (DANS)).

Creation of the WVS

During this period, there was also an extension of the (almost equivalent) questionnaire to non-European countries. Under the direction of Ronald Inglehart, surveys were held in many countries. These surveys were the upswing to the creation of the WVS. From 1995 on WVS had its own organization and conducted surveys at 5-year intervals (1995, 2000, 2005, and 2010). Originally, WVS had its data archived at ICPSR in Ann Arbor (Michigan); later the primary data archive became ASEP/JDS in Madrid.

At the turn of the millennium, there were (independent) rounds of surveys of both the EVS and WVS. In total, about 70 countries were surveyed, yielding a big comparable database, because there was still a notable overlap in the questionnaires of EVS and WVS. From this time on, GESIS was the primary data archive for EVS and ASEP/JDS for WVS. One important issue now became the distribution of the data in a unified format. In cooperation EVS and WVS created a common dictionary (codebook) that covered all questions and variables used in any of the values surveys since 1980. This prepared the publication of the first integrated values surveys database that made the joint data set available to the academic community. Through online access platforms maintained by the respective Data Archives, the underlying documentation also became available: the questionnaires used in all the countries with the different languages, the methodological reports containing information on sampling, nonresponse, etc.

Evolution of the EVS

When planning the most recent, fourth round of EVS surveys for 2008, new developments in the data infrastructure were the basis. In the EVS 1999 survey, there already had been an extension of topics for the survey: solidarity (e.g., with the elderly, disabled, or immigrants), social capital (networks, trust,

civism), democracy, and work ethics. An agreement between EVS and WVS was developed that both organizations would take care of the publication of their own data collections, but that they would safeguard the possibility to easily match the two study-specific longitudinal data sets. In practice, this means that there is agreement about the semantic structure of the integrated database and the labeling of the variables in a unique way.

In 2008, the database for EVS contained all (47) European countries. After data collection, analysis-ready data files as well as the original data files and all kinds of documentation on methodological issues were stored in the data archive to be made publicly available. Besides an Integrated Data File for simultaneous analysis of all countries, this includes separate data files for all countries and a new EVS longitudinal data file.

In retrospect, a great development is seen in the questionnaire construction, sampling, fielding, standardizing of variables to match international classifications (NUTS, ISCO, ISCED, ISO-language), and archiving of the data. An event that illustrates the typical shortcomings in the 1980s is how the region information on the respondents was lost in one country's data. In 1981, information had been collected and stored with punch cards and on paper. Later, data was mainly stored on a varied collection of tapes or hard disks, with the documentation being held in unstructured formats. When the data were transferred to newer information carriers, the original respondent numbers that also contained a region identifier were replaced by sequential ID numbers, thus deleting the regional information. Improvements in advance coordination, monitoring, and archiving processes greatly reduced the number of such problems over the consecutive waves of EVS. The last survey of 2008 can now be seen as a state-of-the-art model for survey research data dissemination, since all (*meta*-) data currently regarded as necessary for methodologically sound analyses are easily available through web platforms. Efforts have also been made to integrate the existing data over time, and with the global data of the WVS. Thus, there is a longitudinal data file that maintains all overlapping information from all EVS survey waves, and there are tools to aid the merging of this file with the analogous WVS longitudinal data file. A precondition for this was that both the EVS and the WVS first collated all their survey waves internally within their respective data archives, creating single points of access for the wider research community. Ultimately, global-level data of the WVS and European-level data of the EVS can be analyzed together, providing a wide view over time and space.

The organizational trajectories of EVS and WVS showcase some interesting aspects of sustainable infrastructures. Beginning with the initiative (and funding) of a rather small core group of researchers, both projects established and extended networks of cooperators for national data collection. Importantly, both projects relied on (different, and for some time changing) data service providers to make analysis-ready research data openly available to the scientific community. Parallel to that, both projects of course produced research output in the form of publications, maintained websites, organized conferences, etc. For these projects it was a major building block of their current, well-defined public identity that they managed to produce research output of their own and to provide access to the data for others under the same 'brand names.' In both cases, this was achieved through an increasingly close and permanent relationship between primary researchers and the providers of infrastructural services. Very similar collaborative structures carry other ongoing programs such as the European Social Survey (ESS), the International Social Survey Programme (ISSP), the Comparative Study of Electoral Systems (CSES), and various international 'Barometer'-type programs (see below).

Selected Individual- or Household-Level Databases

General comparative survey programs are probably the most important source of individual-level data for sociological and political science research. Per definition as a 'program,' they collect data in regular intervals, often in (bi-) annual patterns. Being 'general' means that they address multiple research purposes by prescribing a set of basic standard measures for each survey instance (usually also including variables other than mere demographic information), which are then combined with topical modules that vary across survey instances.

International Social Survey Programme

The first truly international survey program – in that it has no proclaimed regional focus but is open to worldwide participation – is the ISSP. The ISSP was started in 1984 as a voluntary network of only four research organizations. As of 2013, ISSP has 48 members across the world and continues to expand. ISSP performs annual surveys organized into rotating topical 'modules,' so that each topical module is repeated after roughly 10 years. As two-thirds of the module questionnaires are replications of the previous module on the same topic, ISSP automatically builds cross-sectional time series over long time spans, creating a tool to monitor long-term societal change on topics such as religion, social inequality, citizenship, and social networks, to name only a few modules. Although countries or regions such as South Africa, Palestine, China, and India are now represented, ISSP is far from achieving universal coverage of the world's nations. With the national surveys being self-funded, only countries with sufficiently developed local social research institutions are able to participate. This explains the sparse coverage of developing countries.

Gallup World Poll

The only general purpose survey program that actually provides almost exhaustive coverage of the world's nations (160) is a primarily commercial endeavor. The 'Gallup World Poll' was created in 2005–06 and is designed to inform political decision makers on international societal trends such as happiness, migration, perception of corruption, and various others, for which it has solid foundations in substantive and methodological research. The data are currently available to academic research under individually negotiated conditions.

Eurobarometer

Other general comparative programs all have a regional focus. The Eurobarometer series is run since the early 1970s by the European Commission. It covers the member countries of the

European Union, sometimes also candidates for the European Union. Thus, it has expanded its geographical horizon along with the expansion of the political union. Thematically, Eurobarometers deal with all citizen attitudes and behaviors of political or societal relevance to EU policies. The 'Standard Eurobarometers,' fielded twice per year, have an ongoing set of questions on attitudes toward European integration, but add changing focus topics. 'Special' and 'Flash Eurobarometers' are conducted in irregular patterns to address short-term issues. All data are in the public domain and made available through the GESIS Data Archive.

European Social Survey

A fully academic project is the ESS, which since 2002 monitors societal and attitudinal change in Europe to inform social research. It is supported as a research infrastructure project by various member states of the European Union. ESS follows the ISSP model of a mix in standard and rotating contents in several respects, but applies a biannual data collection scheme and puts a very strong emphasis on methodological rigor in fieldwork, measurement methods, and international harmonization. The data are openly available through a website maintained by the Norwegian Social Science Data Archive. The site also offers rich material for academic training.

Other Barometer-Type Surveys

Outside Europe, researchers in several world regions have begun to install region-specific attitude monitoring programs, most of which specifically look at topics of democracy, citizenship, governance, etc. The Latinobarometro, the Asia Barometer, the Afrobarometer, and the Arab Barometer all follow a similar model and cooperate under the network of the Global Barometers initiative. Data access is usually possible for nonparticipants in these projects, too, but it is less standardized than with the aforementioned projects.

World Health Organization Surveys

Among the comparative survey programs directed at specific research topics, two large surveys conducted by the World Health Organization (WHO) deserve mentioning, although they are not ongoing programs. The Multi-country Survey Study of 2000–01 and the World Health Survey of 2002–04 tap into individual health issues and health service availability. Although not all microlevel country data sets are accessible, data access is generally transparent, and the inclusion of many third-world countries makes these surveys unique. This is one of the rare cases where an international organization does make at least some of the micro data behind aggregate statistics (see below for WHO aggregate data) available. For Europe, the health topic is now addressed by a specialized ongoing program Survey of Health, Aging and Retirement in Europe (SHARE), with an almost exclusive focus on the population older than 50 years.

Comparative Study of Electoral Systems

Topical programs with a more academic orientation include the EVS and WVS described above. The other specific topic besides values that finds regular coverage in survey programs is elections and democratic development. Especially the CSES constitutes an unprecedented effort to collect data not only on citizens' attitudes but also on several levels of polities, thus providing a basis for understanding the democratic process in a more complete way. Macro data matching specific research questions are collected along with the survey data. Data are collected shortly after national elections, which binds the timing of national surveys to the respective electoral cycles and enforces a wide time bracket for a set of comparative surveys. CSES strives for wide international coverage, but has to rely on the ability of national institutions to participate in one of the survey waves. Since the first wave from 1996 to 2001, CSES has completed two further waves (modules) and is currently fielding a fourth wave of surveys. Similar in their multilevel design, but focused only on elections to the European parliament are the European Election Studies.

Topical Macro Databases from Academic Projects

Historically, most comparative analyses were done on the basis of macro data, comparing societies or states as units. These early analyses were commonly based on published statistics, whereas research questions for which no official data are available – political system attributes, ethnical or political conflict, corruption, and subjective happiness – require that researchers collect their own macro data.

Correlates of War

The area of militarized conflict is thoroughly covered by the Correlates of War (CoW) project, which accumulates close to 20 separate data sets on risk factors of wars (internal and external armed conflicts) such as cultural differences, colonial history, trade data, etc., and, trivially but necessary, the presence of shared borders between states. CoW, founded in 1963 at the University of Michigan, is now maintained by a network of scholars who regularly update the individual databases following central guidelines. CoW data can often be linked to other macro data sets, for example, the Minorities at Risk (MAR) project initiated by Ted Robert Gurr. The MAR database presently contains information such as minority groups' religious orientation, geographical dispersion, degree of distinctiveness from the plurality, and the degree of discrimination on some 280 minority groups involved in political conflict, over a time span from 1945 to 2003.

Transparency International Index

Corruption shall be listed as the final item in this short list of databases on deficiencies of polities. The corruption index produced by Transparency International (a charitable foundation) is based on a mix of expert judgments and surveys and provides corruption rankings of nation-states in regular intervals.

The 'Polity' Projects, and 'IDEA'

Partly matching the individual-level electoral studies described before, political science collects macrolevel data. At the most

general level, the series of Polity projects (I–IV) (again initiated by Gurr) distinguishes states by their levels of democracy versus autocracy, and describes their changes over time. Polity has received funding from US government agencies under the expectation of informing foreign policies, but the data have always been publicly accessible. Looking at democracies alone, core variables of electoral processes – electoral system design, voter turnout, and political finances – are compiled and reported by the International IDEA project. IDEA (International Institute for Democracy and Electoral Assistance) is an intergovernmental organization to support the development of democracy worldwide. Given the variety of electoral systems and the unreliability of certain official statistics, even collecting and harmonizing the apparently straightforward comparative voter turnout data appears simple on the surface only.

Comparative Manifestoes Project

The Comparative Manifestoes Project looks at the content side of political competition in democracies. It collects and codes the texts of political program statements expressed by political parties prior to elections. The coding of manifestoes is performed intellectually, counting positive or negative statements about a predefined list of political topics. This allows for later matching proclaimed party positions with issue preferences in the electorate, or with policy outcomes after the elections.

World Database of Happiness

Although many of the research-driven databases mentioned above are oriented toward political science, other areas are certainly not void of such databases. An example that is equally relevant for economics, sociology, and psychology is the collection on subjective happiness that was compiled under the direction of Ruut Veenhoven. The World Database of Happiness is a special case insofar that it is a compendium of research in the sense of both data and publications. Happiness research necessarily builds on individual-level data, which is not directly available from the Happiness Database website. But the site does provide collected macrolevel aggregates of surveys, systematically ordered to provide time series at the national level. This allows for correlating national happiness levels with other national-level measures, especially such as economic welfare.

World Religion Database

Religion is another area on which official statistical information is available only at national levels and in ways that are not easy to standardize for international comparison. The most complete and continuously maintained source on religions across the world of which the authors are aware, the World Religion Database, grants access to its information on the populations of the world religions, their regional distribution, etc., only under a subscription model. Other databases are available but produce data of lower density.

European Data Center for Work and Welfare

An interesting example of a different kind is the 'European Data Center for Work and Welfare' (EDAC, www.edac.eu). EDAC constitutes a 'meta shell,' i.e., an organized collection of descriptions of topical databases, along with access information. EDAC is a website established in 2006 as an EU-supported initiative at Tilburg University. After this network ended in 2011, EDAC continued as a consortium hosted at CEPS-INSTEAD in Luxembourg and at the University of Leuven. The goal of the website is to supply information and (in)direct access to more than 500 sources in the field of work, care, and welfare. The unique selling point is the structured overview on the start page where one can choose between (1) comparative data, (2) national data, (3) international repositories, and (4) Europe. In the second stage, an overview of the relevant databases or repositories and search engines is provided. It is also possible to do a keyword-based free-text search. The main benefit for end users is obviously that the search for data sources is prestructured and guided. Even though no maintenance of actual research databases is required for this meta shell type of facility, the information collected in the shell itself requires considerable ongoing curation for keeping data and sources up-to-date and in touch with new research developments.

Supranational Organizations and International Official Statistics Data

Supranational organizations, especially the UN with their subdivisions, the World Bank, and the OECD, are major players in the provision of comparative macro data. With evidence-based policy making being part of their mission, they maintain research departments that have continuously improved the access of the wider research community to data collected under their auspices. This nowadays often includes access to micro data of many large official surveys that traditionally would have been reported only as macro aggregates.

One case in point is the International Household Surveys Network (IHSN). IHSN is a joint initiative of UN agencies, World Bank, and others, providing documentation and access for many household-level official statistics surveys, especially those that were used as input for comparative analyses performed at the listed institutions. This has led to a rich catalog of household surveys, with a focus on developing countries. Similarly, the WHO maintains a repository called 'Global Health Observatory.' This holds not only over 50 macro data sets on population-level health-related topics but also numerous individual-level survey data sets.

While supranational bodies with political or administrative objectives have a natural interest in using and producing comparative data, their national counterparts – usually statistical offices – naturally have national priorities when handling data. Thus, international comparability is often a secondary aspect. Still, statistical offices usually possess the most comprehensive and reliable data collections in their respective country, especially as they collect census data. An outstanding achievement is the 'Integrated Public Use Microdata Series, International' (IPUMS International). IPUMS is a long-standing project at the University of Minnesota with multiple sources of funding and support that compiles and harmonizes national, fully-anonymized-level census data and makes them

available for researchers in a uniquely accessible and well-documented way.

A more typical output from official statistics is almost all data that tap into topics of economic activities and developmental status. OECD and the World Bank aggregate national statistics to supranational levels. By virtue of the administrative integration rules of the European Union, the 'European Statistical System' has a special position here, because it obliges national statistical offices of EU countries to supply pre-harmonized data to the central European statistical office Eurostat. Eurostat offers much of these data to researchers, but for micro data access considerable handling fees may apply.

Finally, supranational agencies or NGOs sometimes produce composite indices for which several of the aforementioned sources are selected and joined. A very prominent example is the 'Human Development Index' (HDI), which is a key figure annually produced by the United Nations Human Development Programme (since 1990). It is meant to convey concise information about the economic and societal development at the national level, with the main purpose being to give a general picture of the general living conditions in a given country at a given time. For this, it collates average country-level information on life expectancy, exposition to education, and economic output per capita. Extended versions of the HDI adapted to reflect economic inequality and gender biases have been added in the meantime. Even for the core HDI, computation rules are constantly revised and a fresh, backward-looking calculation for HDI scores is performed with each new publication of the index.

Current Problems and Future Perspectives

Methodological Problems

Typical methodological problems of comparative databases are, naturally, those at the core of all comparative research. Data from different sources were possibly collected using different methods, under different conditions, and in different cultural contexts. With all these differences, there is no certainty that the data describe the same kind of reality across sources. If this is not demonstrably the case, comparisons might not bring out valid substantive differences between countries, but instead produce biased results. The problem of comparability becomes most evident when large amounts of such data are concentrated in one system. The situation is easiest if 'ex-ante harmonization' has been applied, meaning that the data collection process has been intentionally designed to produce comparable results, e.g., by making sure that question formulations in different languages actually convey the same message to respondents in the respective countries. Even though this does not guarantee immediate comparability, ex-ante harmonization is the best base for achieving it.

However, many microlevel databases and, by virtue of their construction, practically all macrolevel databases consist of ex-post harmonized data. Creators of comparative databases using ex-post harmonization are continuously making judgments whether information item 1 from country A is similar enough to information item 2 from country B to appear in the same column of a raw data table without further treatment, or if particular harmonization steps should be taken, or if a comparison should not be allowed at all. Is, for example, an apparently simple property such as the average household size sufficiently comparable between countries A and B if in country A only respondents in an age bracket from 21 to 90 were surveyed, whereas in country B there was no explicit age bracket (but perhaps that survey was limited to citizens of B, whereas the survey in country A included all residents)? What if one survey used a sample at the person level and the other one a sample at the household level? There often is no way of knowing if proper weights were used to correct this before the average figure entered into the database was computed. This kind of problem is prevalent not only in data from academic surveys but also in the data that are aggregated at international organizations, drawing on inputs from national statistical offices. A case in point is the Standardized World Income Inequality Database (SWIID, see link collection). It collates already existing macro data (Gini coefficients from various prestandardized sources) and performs additional harmonization steps to reduce remaining comparability steps, thereby markedly increasing the predictive value of inequality data versus external criterion variables. This kind of preparatory work by field experts is indispensable, as other users of such data are often not aware of the hidden complexities of data, either because they are either not well documented or because exploring the available documentation is deemed too time consuming.

Even beyond the issue of potential harmonization problems, further research on the quality of comparative macro data (including aggregated micro data) would seem in order. Where the measurement quality of microlevel data is extensively researched, very little systematic research exists that looks at the measurement properties of macro data. Two of the few exceptions are Munck and Verkuilen (2002) and Solt (2009).

A problem that limits the usefulness of many comparative databases for certain purposes is lacking coverage of some regions or countries. Such countries either do not have the resources to maintain data collection facilities or abstain from collecting or publishing reliable data for political reasons. For these reasons, academic survey research with microlevel data has almost no footing in Africa and for some regions in Asia, and only started building a database for South America and the former Eastern Bloc since the 1990s. And even in developed countries, there are almost no comparative panel (intra-individual longitudinal) data at the microlevel.

Organizational Challenges

Comparative databases tend to be the more valuable, the longer they have been maintained and extended in a consistent way – a task that has for several decades been managed by dedicated research data archives. With both comparative research and use of the Internet becoming ever more ubiquitous, a large number of diversified and also overlapping databases have become available through various kinds of Internet platforms that are not automatically attached to the traditional infrastructure providers. Potentially, any new comparative project that collects some data of its own, or manipulates existing data could be looked upon as a new database (King, 2011; Mochmann, 2008). It is technically easy to share these

data through the decentral means of the Internet. However, there are challenges arising from this: the publication of new data is only useful if the potential audience is also able to actually find the data, which is unlikely when it is dispersed over many platforms or websites, lacking central search facilities and standardized descriptors. Further, data maintained on platforms without continuous organizational backing may be at risk of loss, change, or manipulation, which would defeat the purpose of replicability that usually is regarded as one of the main benefits from the publication of research materials. Finally, using disparate data preparation and data distribution procedures is likely to make any particular data set less amenable for merging with other data sets to form an integrated database.

Research funders are increasingly aware of the organizational requirements for achieving long-term data availability and support the implementation of adequate national infrastructures. But how this will translate into infrastructures for comparative data is still unclear. The European Union has devised special organizational structures to take care of collaborative projects, the so-called European Research Infrastructure Consortia (ERIC). These are by design international and geared toward the provision of comparative data. The first ERICs in the Social Sciences were installed in 2012 and 2013 (SHARE, the ESS, and the Council of European Social Science Data Archives). Other regions of the world are following different approaches with predominantly national priorities. Presently, the only viable route to ease the building of comparative databases seems to call for voluntary cooperation among the agencies that collect and provide data, as OECD's Global Science Forum does (OECD Global Science Forum, 2013). The main method would be the early standardization of content descriptors and technical formats, as soon as possible in the data production workflow. With both content and technologies being subject to continuous development, this is likely to be a permanent chore.

Links and Link Collections

Overview Collections

- Macro data sources guide at NSD: http://www.nsd.uib.no/macrodataguide/
- Guide to comparative surveys at GESIS: http://www.gesis.org/das-institut/kompetenzzentren/fdz-internationale-umfrageprogramme/data-references/
- EDAC: www.edac.eu
- ICPSR maintains a list of foundational comparative studies, mostly from political sciences, that are available at ICPSR: http://www.icpsr.umich.edu/icpsrweb/ICPSR/international.jsp

Direct Links to Comparative Databases/Programs

- General Programs:
  - ISSP: www.issp.org
  - Gallup World Poll: www.gallup.com/strategicconsulting/en-us/worldpoll.aspx
  - Eurobarometer: www.gesis.org/eurobarometer
  - ESS: www.europeansocialsurvey.org/
  - Asia(n) Barometer: www.asianbarometer.org/
  - Latinobarometer: www.latinobarometro.org/
  - Afrobarometer: www.afrobarometer.org
- Topical programs:
  - WVS and EVS (see above): www.worldvaluessurvey.org/; www.europeanvaluesstudy.eu/
  - Electoral databases, e.g., European Election Studies (EES): eeshomepage.net/
  - CSES: www.cses.org/
  - SHARE: www.share-project.org/

Topical Macro Databases from Academic Projects

- CoW: http://www.correlatesofwar.org/
- MAR: http://www.cidcm.umd.edu/mar/
- Transparency International: http://www.transparency.org/
- Electoral Participation: www.idea.int
- Comparative Party Manifestoes: https://manifestoproject.wzb.eu/
- Polity IV: http://www.systemicpeace.org/polity/polity4.htm
- World Happiness Database: http://worlddatabaseofhappiness.eur.nl/
- World Religion Database: www.worldreligiondatabase.org/
- Social welfare structures and policies: EDAC: www.edac.eu/
- The SWIID: myweb.uiowa.edu/fsolt/swiid/swiid.html

Supranational Organizations and International Official Statistics Data

- Global Health Observatory: http://apps.who.int/gho/data/view.main and http://apps.who.int/healthinfo/systems/surveydata/index.php/catalog/central/about
- UNDP: HDI: http://hdr.undp.org/en/statistics/
- World Bank: http://data.worldbank.org/
- OECD: http://stats.oecd.org/
- IPUMS, International: http://international.ipums.org/
- EUROSTAT: (http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home)

See also: Comparative Studies: Method and Design; Cross-Cultural Research Methods in Sociology; Ethnic Conflict, Geography of; International Social Survey Programme; Pooled Cross-Sectional and Time Series Analyses, in Political Science; Quantitative Cross-national Research Methods.

Bibliography

Coleman, James S., 1990. Foundations of Social Theory.

Deutsch, Karl W., 1960. Toward an inventory of basic trends and patterns in comparative and international politics. American Political Science Review 54, 34–57.

Erikson, R.S., 2002. National election studies and macro analysis. Electoral Studies 21, 269–281.

Franklin, Mark, Wlezien, Christopher, 2002. Reinventing election studies. Electoral Studies 21, 331–338.

Inglehart, Ronald, 1977. The Silent Revolution: Political Change among Western Publics. Princeton University Press, Princeton.

Jagodzinski, Wolfgang, 2006. Comparative survey research and its infrastructure in Europe. Kwansei Gakuin Sociology Department Studies 101, 1–13.

Kaase, Max, 2001. Databases, core: political science and political behavior. In: The International Encyclopedia of Social and Behavioral Sciences. Elsevier, pp. 3251–3255.

King, Gary, 2011. Ensuring the data-rich future of the social sciences. Science 331, 719–721.

Kohn, Melvin L., 1989. Cross-national research as an analytic strategy. In: Kohn, Melvin L. (Ed.), Cross-national Research in Sociology. Sage Publications, pp. 77–102.

Marshall, Monty G., Gurr, Ted Robert, 2012. Polity IV Project: Political Regime Characteristics and Transitions, 1800–2012.

Mochmann, Ekkehard, 2008. Improving the evidence base for international comparative research. International Social Science Journal 59, 489–506.

Munck, Gerardo L., Verkuilen, Jay, 2002. Conceptualizing and measuring democracy – evaluating alternative indices. Comparative Political Studies 35, 5–34.

Norris, Pippa, 2009. The globalization of comparative public opinion research. In: Landmann, Todd, Robinson, Neil (Eds.), The Sage Handbook of Comparative Politics. Sage, Thousand Oaks, pp. 520–540.

OECD Global Science Forum, 2013. New Data for Understanding the Human Condition: International Perspectives. OECD Global Science Forum Report on Data and Research Infrastructure for the Social Sciences. OECD.

Przeworski, Adam, Teune, Henry, 1970. The Logic of Comparative Social Inquiry. John Wiley & Sons, New York.

Ragin, Charles C., 1994. Introduction to qualitative comparative analysis. In: Janoski, Thomas, Hicks, Alexander M. (Eds.), The Comparative Political Economy of the Welfare State. Cambridge University Press, Cambridge, pp. 300–309.

Solt, Frederick, 2009. Standardizing the world income inequality database. Social Science Quarterly 90, 231–242.

Data Bases and Statistical Systems: Labor and Work

Constance Sorrentino, US Bureau of Labor Statistics, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 11, pp. 7511–7517, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The article on 'Statistical Systems: Labor and Work' gives the readers an overview on national and international statistic programs on labor statistics. It comprises information on the most common topics and standard classifications of the social and economic issues of our time; it describes usual institutional frameworks to data collection, presentation and distribution to users. In addition, brief descriptions of comparative labor statistics programs, three international organizations, and a unique international microdatabase are presented.

Labor statistics represent a large body of information relevant to many of the social and economic issues of our time. They are used as indicators of how well the general economy is functioning and also as a basis for designing, implementing, and monitoring employment and social programs and policies. Labor statistics are generally part of a national statistical program and are interrelated with statistics in other fields, for example, with statistics on education, health, and production. International comparisons of labor statistics can help in assessing a country's economic performance relative to other countries (for example, by comparing unemployment rates and employment growth) and in evaluating the competitive position of a country in international trade (for example, by comparing productivity and labor cost trends).

Topics and Classifications

Article 1 of Labor Statistics Convention Number 160 of the International Labor Office (ILO) established in 1985, a minimum set of topics to be covered in official national labor statistics programs (ILO, 2000). The list is as follows:

1. economically active population, employment, where relevant unemployment, and, where possible, visible underemployment;
2. structure and distribution of the economically active population, for detailed analysis and to serve as benchmark data;
3. average earnings and hours of work (hours actually worked or hours paid for) and, where appropriate, time rates of wages and normal hours of work;
4. wage structure and distribution;
5. labor cost;
6. consumer price indices (CPIs);
7. household expenditure or, where appropriate, family expenditure and, where possible, household;
8. income or, where appropriate, family income;
9. occupational injuries and, as far as possible, occupational diseases; and
10. industrial disputes.

ILO convention 160 recommends that each member state (185 of the 193 UN member states are members of the ILO) "collect, compile, and publish basic labor statistics, which shall be progressively expanded in accordance with its resources to cover [these] subjects." The USA and most other developed nations collect and publish data for all subjects listed in Article 1. On the other hand, developing nations often do not yet have the full array of labor statistics. The publication Key Indicators of the Labour Market (KILM), first released in 1999 (Johnson et al., 1999), provides information about country-level data on currently 18 KILM in almost all countries of the world from 1980 to the latest available year. Deficiencies in the availability and timeliness of labor statistics are particularly apparent in Africa and parts of Asia. (See ILO, http://www.ilo.org/empelm/what/WCMS_114240/lang–en/index.htm.)

National Classifications

The presentation of labor statistics requires many distinctions or categories such as occupations and branch of economic activity. In order to facilitate data collection and analysis, these categories are organized into classifications. These are instruments that group together units of a similar kind, often in hierarchies, thus allowing for different degrees of detail to describe the characteristics in a systematic and simplified way. For example, in the USA, the Standard Occupational Classification (SOC) and the Standard Industrial Classification (SIC) have been the structures that define occupations and industries for many years. Such systems need periodic revision to keep up with the changing economy and to accommodate new industries and occupations. Revising the SOC System (2010), provides an account of the 2010 revision to the US occupational classification, a brief history of SOC system as well as a general time line for the 2018 revision.

The North American Industry Classification System (NAICS) represents a complete restructuring of the US SIC system (1997). This system was prepared jointly because the North American Free Trade Agreement of 1994 increased the need for comparable statistics from the USA, Canada, and Mexico. Each country can individualize the new system to meet its own needs, as long as data can be aggregated to the standard NAICS categories.

Other International Classifications

International standard classifications serve to enhance cross-national comparisons. The International Standard Industrial

Classification of all Economic Activities (ISIC) was established under the auspices of the United Nations, and the latest International Standard Classification of Occupations (known as ISCO-08) was established by the ILO's Tripartite Meeting of Experts on Labour Statistics on Updating the ISCO in 2007. Resolution concerning updating these classifications and correspondence tables with ISCO-88 are available on the ISCO Web site in English, French, and Spanish. (See ISCO, http://www.ilo.org/public/english/bureau/stat/isco/index.htm.) Countries often prepare concordances or 'crosswalks' that allow their national classifications to mesh with the international classifications.

International Concepts

The ILO's periodic International Conferences of Labor Statisticians have established resolutions and guidelines for the measurement of labor statistics (ILO, 2000), including recommendations on employment and unemployment statistics, statistics of labor costs, wage statistics, consumer price indexes, and strikes and lockouts. These guidelines serve as a model for countries trying to develop national statistics on the topic concerned and they also provide, to some extent, a benchmark for international comparisons. For example, ILO standards for measuring labor force, employment, and unemployment are followed by many countries. Varying interpretations of these guidelines, however, particularly with regard to different aspects of unemployment, create differences across countries that require adjustments to enhance comparability of unemployment rates (Sorrentino, 2000). The ILO recommendations and guidelines are available online at http://www.ilo.org/stat/lang–en/index.htm.

Institutional Framework

Labor statistics are primarily collected by governments rather than the private sector, although in some countries there has been a movement toward privatizing some data collections. In the USA, which has a decentralized statistical system, over 70 Federal agencies are engaged in the collection of statistics. The Bureau of Labor Statistics (BLS) in the US Department of Labor is the principal data-gathering agency of the US Federal Government in the broad field of labor economics. Most of the Bureau's data come from voluntary responses to surveys of businesses or households conducted by BLS staff, by the U.S. Census Bureau (under contract with BLS), or in conjunction with cooperating State and Federal agencies. Unlike the decentralized system of the USA, other countries tend to have central statistical offices, and the collection of labor statistics is accomplished within these national agencies. However, labor ministries often also play a role in the collection of labor statistics abroad. Several international organizations work in the realm of labor statistics, obtaining and publishing labor information from countries, providing technical assistance to member countries, and establishing conceptual guidelines. National agencies cooperate with the international organizations by transmitting the data requested on a regular basis and also for ad hoc studies. Major international organizations involved in labor statistics are discussed in Section International Comparisons.

Sources of Labor Statistics

Labor statistics are generally collected in censuses or sample surveys of households or business establishments, or processed on the basis of data resulting from administrative procedures of public agencies. Many of the major programs of the US BLS are based on sample surveys. For instance, the Current Population Survey (CPS) is a monthly survey based on a sample of about 72 000 households from 824 sample areas designed to measure demographic and labor force characteristics for the population of the United States; data are collected by computer-assisted personal and telephone interviews. The survey, sponsored jointly by the U.S. Census Bureau and the U.S. BLS, is the primary source of the national statistics on labor force, employment, and unemployment. About 70 countries conduct regular labor force surveys, on an annual, quarterly, or monthly basis. In January 2003, the CPS adopted the 2002 census industry and occupation classification systems, derived from the 2002 NAICS and the 2000 SOC System. (See Chapters 3 (pp. 1–15) and 4 (pp. 1–6) of *Current Population Survey, design and methodology* (Technical paper 66, October 2006 online at https://www.census.gov/prod/2006pubs/tp-66.pdf.)

Another major survey of BLS is the monthly Current Employment Statistics Survey. This survey is a sample of about 144 000 businesses and government agencies, representing approximately 554 000 individual worksites, and it provides employment, hours, and earnings data collected from payroll records of business establishments. Similar surveys are conducted by many other countries.

Labor statistics may also be based on administrative records. Although the primary purpose of administrative records is not the production of statistics, these systems provide a rich information source that has been exploited to varying degrees in different countries. Among the uses in the USA, administrative data from the unemployment insurance system are used in combination with CPS data to produce unemployment data for States, labor market areas, counties, and cities. In several European countries, employment office registrations are the primary monthly indicator of the labor market. This is likely to be the case in countries that do not have a monthly household survey as a source for unemployment statistics.

Some labor statistics are constructed by combining data from different sources. For example, industry labor productivity data (output per employee or per hour) require data on both labor input and business output. The labor input measures are generally constructed from a combination of hours data from labor force and establishment surveys, while the output components typically are based on national income and product accounts, which are derived from business establishment surveys.

Dissemination of Labor Statistics

Disseminating statistics means presenting and distributing the data to users in various forms and preparing analyses of the data. The BLS makes available the information it produces

through a broad publication program that includes news releases, periodicals, reports, and bulletins. Articles in the Monthly Labor Review present analyses of labor data. Some BLS material is available on magnetic tape, computer diskette, CD-ROM, and microfiche. Current BLS data, publications, and articles are also made available through the Internet. Other countries use similar methods of data dissemination.

The publications Major Programs of the BLS (1998) and the Handbook of Methods (1997; see the updated BLS Handbook of Methods online at http://stats.bls.gov/opub/hom/), as well as the BLS Web site noted in the bibliography, provide further information on the Bureau's programs and how to obtain BLS data. Links to the statistical offices of other countries as well as to international organizations are included on the BLS Web site.

In 1996, the International Monetary Fund established a Dissemination Standards Bulletin Board on a Web site to assist countries in following the best practices of dissemination of economic, financial, and sociodemographic data to the public. One of the aspects of this, the General Data Dissemination System (GDDS), guides countries in the provision of comprehensive, timely, accessible, and reliable data, and the site provides information on data produced and disseminated by participating member countries. The GDDS system is built around four dimensions – data characteristics, quality, access, and integrity. Information on the formats and titles of national statistical publications is included. The GDDS is a source of metadata (that is, descriptions of current practices on data concepts, sources, and methods) on labor statistics as well as on other types of economic statistics. The detailed metadata provide users with a tool to better assess the usefulness of the data for their own particular purposes.

International Comparisons

An international perspective can often enrich analysis of national labor markets, but such comparisons must be made with caution because statistical concepts and methods vary from country to country. Simply comparing national data across countries could lead to erroneous conclusions. The BLS and several 'international organizations derive meaningful comparisons by selecting a conceptual framework for comparative purposes; analyzing foreign statistical series; selecting those that most nearly match the desired concepts; and adjusting statistical series, where necessary and feasible, for greater intercountry comparability.

The BLS comparisons generally are based on the US concepts so that they will be on the conceptual basis most familiar to the US data users. Comparisons made by international organizations frequently are based on the international standards recommended by the ILO, discussed in Section 2.3. Brief descriptions of the comparative programs of the BLS, three international organizations, and a unique international microdatabase follow. Further information is available from their respective Web sites, cited in the bibliography.

Bureau of Labor Statistics

From its inception, BLS has published comparative information on labor conditions and developments abroad. Foreign labor research and statistical analyses have been undertaken because: (1) comparisons between US and foreign labor conditions shed light on the US economic performance relative to other nations; (2) comparisons provide information on the competitive position of the USA in foreign trade, which has an important influence on the US economy and employment; (3) information on labor conditions published by a majority of foreign countries is not readily available to the US labor representatives, employers, Government officials, and others, and frequently is not available in English; and (4) often, only an expert can judge the quality and comparability of foreign statistical data.

The BLS International Labor Comparisons (ILC) program has covered five key areas: (1) labor force, employment, and unemployment; (2) hourly compensation costs; (3) productivity and unit labor costs; (4) GDP per capita and per hour; and (5) consumer prices. Only the hourly compensation comparisons have covered 34 countries at various stages of development, whereas all other areas, the labor force (16 countries), productivity comparisons (19 countries), GDP per capita and per hour (20 countries), and consumer prices (16 countries), have given information only on major developed countries. However, according to the current announcement of the Conference Board (see ILC Tables at www.bls.gov/ilc), BLS has eliminated the ILC program due to across-the-board spending cuts termed 'sequestration.' For future release, BLS intends to produce ILC indicators similar to those done previously by BLS for future release. Outside the USA, national statistical agencies generally do not produce their own international comparisons.

They depend on the comparative work of the BLS as well as that of international bodies such as those described below http://www.ipetitions.com/petition/saveilc/.

International Labor Office

The ILO in Geneva plays a major role in the arena of international labor statistics. This agency's role has already been mentioned in terms of promoting the production of labor statistics and setting international standards for statistics. The worldwide scope of the ILO's membership ensures that this organization provides the most comprehensive geographic coverage among the international organizations involved in labor statistics.

The ILO has several publications or programs that provide varying degrees of comparable labor statistics. The annual *Yearbook of Labour Statistics* (ILO, 1999) is the most comprehensive source for worldwide labor statistics. The Yearbook compiles national labor data according to international classifications of industry and occupation, with annotations concerning deviations from these classifications, coverage, and sources. All data are supplied by the national statistical authorities. The Yearbook contains data on economically active population, employment, unemployment, hours of work, wages, labor cost, consumer prices, occupational injuries, and strikes and lockouts. Some statistics are presented as time series (the series usually cover the preceding 10 years), others are provided for only the latest year.

The ILO's program of ILO-Comparable labor force statistics takes the process a step further by making adjustments of the

labor force and unemployment data to ILO concepts for over 30 countries. Data from the ILO-Comparable program are available online (Lawrence, 1999) and are also used in the publication KILM (Johnson et al., 1999). The latter publication is well annotated concerning comparability issues for all of the 18 indicators provided.

Organization for Economic Cooperation and Development

The Paris-based Organization for Economic Cooperation and Development (OECD) comprises 34 member countries, most of which are the advanced countries in North America and Western Europe, as well as Australia, New Zealand, and Japan. In addition, some of the transition countries of Eastern and Central Europe, Mexico, and Korea became members of the OECD in the 1990s. Most OECD labor statistics cover the member countries only, but some series are being expanded to cover nonmember Asian and Latin American countries.

The main labor-related publications of the OECD are the annual and quarterly editions of Labour Force Statistics (1999). The annual publication compiles two decades of statistics, while the quarterly editions show selected data for the past 5 years. These publications provide statistics on key elements of the labor force, rather than on the broader range of labor statistics contained in the ILO Yearbook. Through the OECD's provision of extended time series, it is possible to identify structural changes that have taken place in the labor force of the countries with respect to gender, age, and economic activities. Data on part-time employment and duration of unemployment are included. The detailed statistics on employment are generally in conformity with the ISIC.

Another program of the OECD Harmonized unemployment rates, provides unemployment rates adjusted to ILO concepts for 34 member countries. The Harmonized rates are published in an annex to the aforementioned quarterly editions and they also are available from the OECD Web site. Additional labor statistics are published in the annual OECD Employment Outlook, expanding the range considerably to include special international analyses of training statistics, indicators of the degree of employment protection and regulation, trends in self-employment, income distribution, minimum wages and poverty, the transition from education to the labor market, trends in working hours, and workforce aging, to name some of the special studies. The OECD analyses pay careful attention to comparability and often contain annexes that explain the data comparison issues in detail.

Statistical Office of the European Communities (Eurostat)

Eurostat's main role is to process and publish comparable statistical information for the 28 member countries of the European Union (2014). Eurostat tries to arrive at a common statistical language that embraces concepts, methods, and statistical and technical standards. A significant portion of Eurostat's statistics relates to labor. The European Union Labour Force Survey (2013) is a joint effort by member states to coordinate their national surveys. The result is a detailed set of comparative data on employment, working time, and unemployment. Eurostat also publishes monthly harmonized unemployment rates for all member countries. Comparable data are critically important for the European Union because the Community's social funds are distributed according to a formula that includes unemployment as a significant component. Eurostat's Harmonized Indices of Consumer Prices (HICPs) are designed for international comparisons of consumer price inflation. The HICPs are the result of the collaboration between Eurostat and national statistical offices of member states on harmonizing the different methods used to compile price indices. The harmonization process has resulted in an improved quality of the indices. HICPs are not intended to replace national CPIs. Besides the labor force survey and CPI statistics, Eurostat also publishes results from its harmonized labor cost surveys and conducts a European household panel that allows for analysis of income, employment, social exclusion, health, education, and many other topics.

Luxembourg Income Study

The Luxembourg Income Study (LIS), begun in 1983, is a unique cooperative international research project with a membership that includes 25 countries on four continents. The project is funded mainly by the national science and social science research foundations of its member countries. LIS comprises a collection of household income survey sets of microdata (files of data on individual households) for cross-country research. A companion database established in 1993, is the Luxembourg Employment Study (LES). The data sets of LIS and LES have been harmonized in order to facilitate comparative studies. LIS and LES have become a rich source for comparative research in the areas of poverty and income distribution as well as labor market behavior. Moreover, these databases foster interdisciplinary research in the fields of economics, sociology, and political science. Researchers in member countries may use the LIS and LES databases at no charge, after signing a confidentiality agreement. Copies of the studies based on these databases are available from the LIS Web site.

Disclaimer

All views expressed are those of the author and do no necessarily reflect the views of the Bureau of Labor Statistics.

See also: Data Bases and Statistical Systems: Economics (Statistical Systems); Data Bases and Statistical Systems: Education, General; Data Bases and Statistical Systems: Health; Division of Labor: History of the Concept; Employment and Labor, Regulation of; Labor History; Labor Supply; Labor Unions; Labor, Division of; Work, Labor: History of the Concept.

Bibliography

BLS Handbook of Methods. BLS Bulletin 2490, 1997. BLS, Washington, DC.

Cosca, T., Emmel, A., 2010. Revising the standard occupational classification system for 2010. In: Monthly Labor Review, pp. 32–41.

Employment Outlook, 2000. OECD, Paris, France.

Eurostat (Statistical Office of the European Communities) Web site: http://europa.eu.int//comm//eurostat/.

ILO, 1999. Yearbook of Labour Statistics. ILO, Geneva, Switzerland. The statistics of the yearbook are available, online at: http://laborsta.ilo.org.

ILO, 2000. Current International Recommendations on Labour Statistics. ILO, Geneva, Switzerland. The ILO's conventions and recommendations are also available online at: http://www.ilo.org/public/english/bureau/stat/Standards/index/htm.

ILO, 2004. Current International Recommendations and Guidelines on Labor Statistics. ILO, Geneva, Switzerland. International Monetary Fund Dissemination Standards Bulletin Board Web site: http://dsbb.imf.org/.

Johnson, L., Sorrentino, C., et al. (Eds.), 1999. Key Indicators of the Labour Market. ILO, Geneva, Switzerland. Web site: http://www.ilo.org/public/english/employment/strat/KILM/index.htm.

Labour Force Statistics 1978–1998. Also Quarterly Labour Force Statistics, 1999. OECD, Paris, France.

Labour Force Survey, 1998. Eurostat, Luxembourg. International Labor Office Web site: http://www.ilo.org.

Lawrence, S., 1999. ILO-comparable annual employment and unemployment estimates. ILO Bulletin of Labour Statistics 3, XI–XLIII. The ILO-comparable data are available at the following Web site: http://laborsta.ilo.org.

Luxembourg Income Study (and Employment Study) Web site: http://www.lis.ceps.lu/index.htm.

Major Programs of the Bureau of Labor Statistics. BLS Report 919, 1998. BLS, Washington, DC.

Office of Management and Budget, 1997. North American Industry Classification System. Bernan Press, Lanham, MD.

Organization for Economic Cooperation and Development Web site: http://www.oecd.org. For the Standardized Unemployment Rates: http://www.oecd.org/media/new-numbers.

Revising the Standard Occupational Classification System. BLS Report 929, 1999. BLS, Washington, DC. This publication is available at the following Web site: http://www.bls.gov/soc/socrpt929.pdf.

Sorrentino, C., 2000. International unemployment rates: how comparable are they? Monthly Labor Review 6, 3–20. Web site: http://www.bls.gov/opub/mlr/2000/06/art1full.pdf.

U.S. Bureau of Labor Statistics (BLS) Web site: http://stats.bls.gov. For the BLS international comparisons series: http://stats/bls.gov/flshome.htm.

U.S. Bureau of Labor Statistics (BLS), 2013a. Country Notes and Data Sources. International Comparisons of Annual Labor Force Statistics, 1970–2012. Web site: http://www.bls.gov/fls/flscomparelf/country_notes.pdf.

U.S. Bureau of Labor Statistics (BLS), 2013b. International Labor Comparisons (ILC). Web site: http://www.bls.gov/ilc/#productivity.

Data Bases and Statistical Systems: Law

Ryan Harrington, Yale Law School, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The breadth of information across a variety of jurisdictions makes subscription databases necessary to complete effective legal research. This article analyzes the utility of subscribing to one or more of the leading vendors and also suggests additional databases for anyone doing substantial American legal research. Particular emphasis is paid to the two leading vendors, Lexis and Westlaw.

United States' Legal System and Databases

The Importance of Legal Databases

Effective legal research in the United States' common law system requires mastery of both case law and statutes. To research the law on any given subject, a researcher must be able to identify statutes that address the legal issue. He or she must also locate and synthesize court decisions on the subject. The researcher must then form an argument about how previous court decisions should compel a court to rule on an unresolved legal claim. Although most trial, appellate, and Supreme Court decisions are published for public use, subscriptions to legal databases make researching legal issues easier and more effective. Without the assistance of a legal database, it is likely that a researcher will miss critical cases. Exhaustive searches are particularly important in common law jurisdictions.

Precedents and Stare Decisis

Concepts of 'precedent' and 'stare decisis' might be familiar to lawyers and students in civil law countries, but these issues are paramount in the United States' legal system. Courts reach their decisions based on precedent, or the cumulative body of cases previously decided by similar courts and tribunals. Sometimes a court must follow a precedent. For example, if the Supreme Court has interpreted a statute in a particular way, lower courts must adhere to that rule, even if they disagree with the Supreme Court's decision. The Supreme Court's precedents are thus described as 'controlling' authority. When researching a legal issue, one must be aware of and have access to all cases with precedential value. The influence of precedent is so strong that courts will rarely even reverse themselves. If an issue were to reappear at the Supreme Court, the Court would have the option of reversing its prior decision, but it usually adheres to its own precedent. This tradition of preserving precedents, called stare decisis, rests on the principle that similar cases should be decided consistently. Some precedents are not controlling, but merely persuasive. If a federal district court in Connecticut evaluates the same legal issue that a federal court in New York already addressed, the Connecticut court may elect to follow the prior decision, but it is not obligated to do so, because the prior decision occurred in another jurisdiction. The concept of persuasive authority means that one must be able to find relevant law outside of one's own jurisdiction. Often legal researchers use commercial vendors to search for such cases.

Commercial Vendors of Legal Databases

Traditionally, the commercial vendors Lexis and Westlaw competed for the attention of academics and practitioners; recently BloombergLaw has begun to challenge the supremacy of these two vendors. There are also a host of lower cost databases on the market (Justiss, 2011). Although many legal practitioners have access to only one of the primary vendors, most US law schools provide access to many and allow students and faculty to develop their own preferences. The interfaces of the two leading databases affect how users research (Jones, 2009), but the difference in content is not immediately obvious. Lexis and Westlaw's newest platforms, known respectively as LexisAdvance and WestlawNext, are two databases that demonstrate the utility of conducting legal research using a subscription database to conduct searches for case law, statutes, legislative history, dockets, and secondary sources. (Wheeler and Amrogi provide discussions about the differences between the products' newer models and their previous versions.)

The Value of Databases in Case Law

Organizational Benefits of Legal Databases

The common law system nearly requires a researcher to subscribe to one of the core legal databases to locate precedent. These databases publish decisions and link the user to prior decisions relied upon by a current case. They also provide a feature called the 'Table of Authorities,' which lists all cases cited in the opinion and indicates the depth of treatment each case received. The databases do the opposite as well, by providing access to cases that have cited to the case at hand. Finally, these databases work to distill legal issues so that one can find additional cases that touch upon the same legal issues. With centuries of precedent from a variety of jurisdictions, one benefits from legal databases that attempt to systematize case law and link to related issues and cases.

LexisAdvance

LexisAdvance accomplishes this through the use of several enhancements; two of the most important enhancements are 'headnotes' and 'Shepard's.'

Editors at Lexis do more than describe the subject matter of a case; they curate the materials by drilling down into each

discrete legal issue that appears in the body of a decision (Nevelow Mart, 2013). Editors or algorithms then assign each legal issue a 'headnote' (Nevelow Mart, 2010). Headnotes systematically organize law into categories and make it possible to find cases that are similarly categorized. The headnotes are buried within the text of the opinion, but they also appear at the very beginning of an opinion, which has the effect of creating an outline. Just as an opinion might have numerous headnotes, a particular passage in the opinion might receive several headnotes. Whenever possible, Lexis attempts to classify the headnotes into preexisting categories, which means a headnote will direct the reader to cases which have received similar headnotes.

If a passage receives multiple headnotes, the passage might reflect the convergence of two bodies of law. In other words, when looking at a headnote in a case about Family and Medical Leave, Lexis might indicate that this headnote can be classified under the path 'Labor and Employment > Leaves of Absences > Family and Medical Leaves > Remedies,' but it might also receive the path 'Constitutional Law > Equal Protection > Gender & Sex.' Clicking on the term 'remedies' might provide the user with a host of cases that differ greatly from the cases that would appear after clicking 'Gender & Sex,' but the Lexis editors have determined that the case in question belongs in both categories.

WestlawNext

The WestlawNext platform appears slightly different but also contains headnotes. Similarly, the editors take language from the opinion and attempt to place the language within the company's own proprietary digest, which classifies all points of law into headings and subheadings. Because the systems are unique to each vendor, not every point of law in a published case will receive the same treatment (Nevelow Mart, 2010). For example, there is nothing like 'Labor and Employment > Leaves of Absences > Family and Medical Leaves > Remedies,' which is a Lexis headnote path, on the WestlawNext platform. There is, however, a path of 'Labor and Employment > Time Off; Leave > Actions > Persons Protected and Entitled to Sue' on WestlawNext.

Westlaw created its digest system to bring order to the extraordinary body of case law (Doyle, 1992). Indeed, the terms above 'Labor and Employment > Time Off; Leave > Actions > Persons Protected and Entitled to Sue' are actually equipped with slightly more information, namely 'key numbers,' and direct links to those key numbers. In the WestlawNext system, points of law are classified within the larger framework of a digest. Any legal issue surrounding 'Labor and Employment' will, at a minimum, receive the digest number '231H,' which is how WestlawNext designates all Labor and Employment issues. Clicking on the '231H' link will direct a user toward all cases that Westlaw editors have identified as pertaining to labor and employment (**Figure 1**).

Points of law can be more specific than simply 'Labor and Employment,' as the path demonstrates. A researcher on WestlawNext who thinks a point of law in a case is important will probably be interested in clicking on the final key number that the point has received. Using the above example, a user can click on '231Hk384' which is the designation not just for labor and employment issues, but for cases uniquely described as relating to 'Persons Protected and Entitled to Sue' under the umbrella term of 'Actions' which falls under the umbrella term of 'Time Off; Leave.' WestlawNext refers to the identifier '231Hk384' as a 'key number.'

Clicking the '231Hk384' key brings a researcher into the digest and allows for further manipulation by jurisdiction. Furthermore, the results can be displayed to highlight cases that have received the most citation for that particular legal issue or by cases that have most recently cited that legal issue. One can limit results to require a particular keyword or date range.

Occasionally there will be no cases that have legal issues as specific as what the keycite path indicates. This might mean that a case is the first of its kind in its jurisdiction, or it might mean that the editors created an entirely new headnote path to deal with the case's unique legal issue. One problem with indexing is that editors have to anticipate future legal issues (Dabny, 1986). If the case is unique to one jurisdiction, both databases allow for the user to expand a search into one or more jurisdictions. This will retrieve precedent that might not bind a court, but could be considered persuasive. If there is no similar case in any jurisdiction, the visible digest path can be

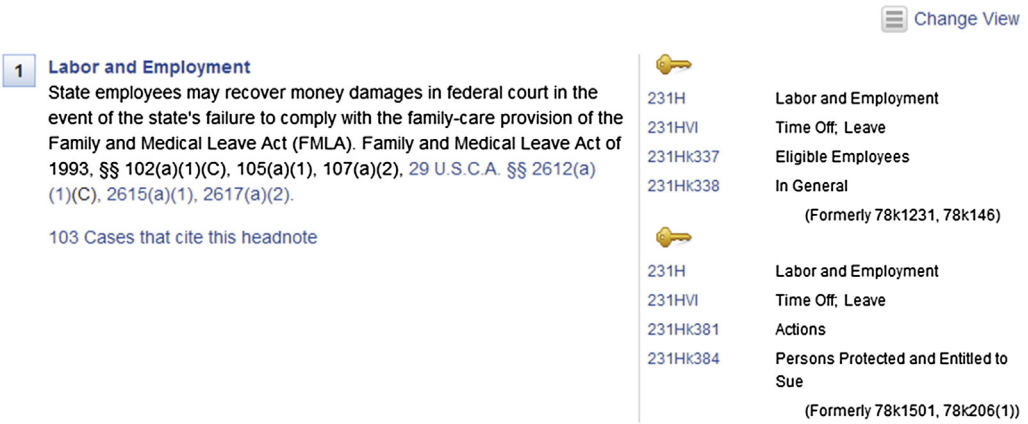

Figure 1 Westlaw digest.

useful because a researcher can click on the key number one step above in the path's hierarchy. In the example above, one could click the key for the term 'Actions' or the key for the term 'Time Off; Leave.' Going up the chain in the digest path means there are more likely to be cases that have received that same key number.

The digest is available to all WestlawNext users by clicking the 'key numbers' link. It can be challenging to locate the particular legal issue by beginning with the digest, because not all users will be familiar with the terms that WestlawNext uses. It can often be easier to find cases by doing traditional keyword searches and looking for the key numbers in the headnotes of a relevant case.

Critical Services Provided by Both Platforms

Aside from headnotes, these databases offer other critical services. The databases' citation systems are particularly important. Citation systems enable researchers to identify whether a case (or a specific point of law within that case) remains good law. It directs the researcher toward cases that cite *to* the original decision. The editors at Lexis and Westlaw attempt to alert the reader to subsequent treatment of a case through the use of warning indicators. A case that has been overruled in its entirety or on a particular legal issue will receive a red stop sign after the case name on LexisAdvance. Additional icons indicate that points of law in a case were emboldened by later decisions or called into question. Sometimes, a later decision will disagree with the result in a particular case and overrule it directly. Because of the concept of stare decisis referenced earlier, these occurrences are rare. More commonly, a later decision will distinguish itself from the case at hand by identifying certain facts or features that might allow an alternative result. In this way, the legal principle of the original case might be limited to a smaller set of circumstances. On LexisAdvance, these cases will receive a yellow triangle, which indicates to the reader that the case might have particular limitations not detectable from the language in the opinion itself.

When looking at a case in LexisAdvance, the 'Shepardize' link accesses a screen that contains several important pieces of information: the Appellate History, Citing Decisions, Citing Law Reviews, Treatises, etc. (This article's discussion of law reviews, treatises, etc. appears in the Section entitled 'Secondary Sources.'; **Figure 2**.) The Appellate History directs

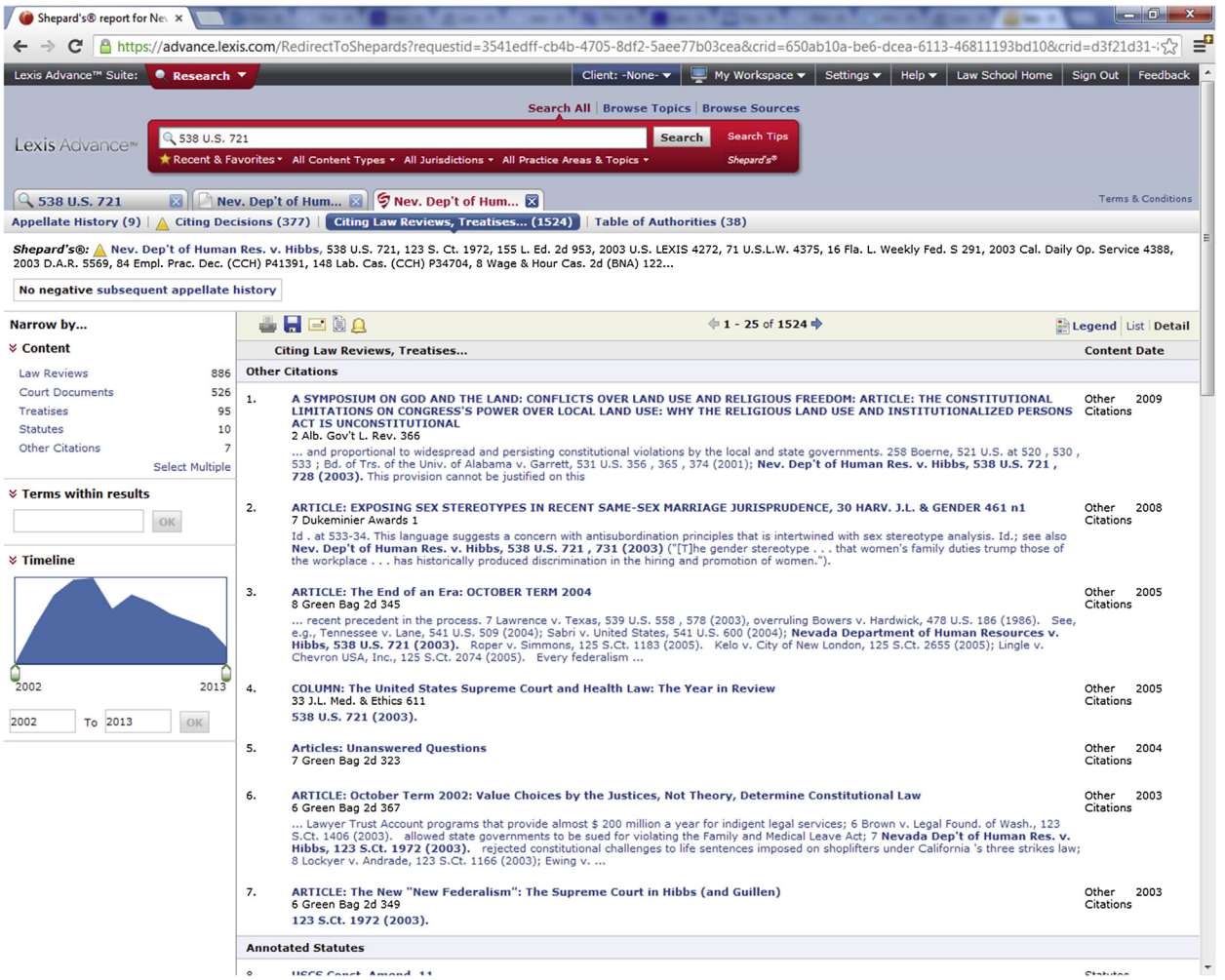

Figure 2 Lexis citing references.

a user toward different stages of a case throughout its appellate process. For example, someone who Shepardizes a Supreme Court case should be able to identify and retrieve decisions from the court of appeals as well as the trial court. In addition, LexisAdvance will point toward subsequent decisions that resulted from the Supreme Court's instructions. One can even visualize the appellate history by clicking on a 'map' function (Figure 3).

Clicking the 'Citing Decisions' tab brings the user into a field of subsequent decisions. It is simple to manipulate the results; for example, one might wish to see only cases that have treated the case negatively or have distinguished the case. Similarly, one might wish to eliminate all subsequent decisions outside of one's own jurisdiction. One field indicates how many opinions have been issued in each circuit, which can indicate that a circuit other than your own might have addressed similar legal issues or might provide a framework for your argument. Another option would be to require a case to have been published within a certain timeframe.

One particularly useful field is called 'discussion.' The discussion field describes subsequent cases by the amount of attention they pay to the case in question. WestlawNext uses one to four stars to indicate the depth of treatment, whereas LexisAdvance presents a bar (Figure 4). Cases that 'analyze' the opinion have all four bars filled, which indicates that the cases will devote nearly a full page of text elaborating on the opinion. Other options include 'Discussed' with three bars, 'mentioned' with two bars, and 'cited' with merely one bar. Discussed indicates an opinion has devoted at least an entire paragraph to the case, while 'mentioned' describes cases that reference the case with slightly more focus than a mere citation.

The WestlawNext model is quite similar in its citator function. WestlawNext calls this function 'KeyCite.' KeyCite has flags (red and yellow) that indicate at least one point of law in a particular case is no longer good or has been called into question. It also features depth of treatment stars, which indicate the amount of attention a citing case will pay toward the case in question. A user can filter by document type (including highest court in a jurisdiction or law review article), jurisdiction generally, depth of treatment, date, and using a keyword search, which can be accessed using the 'locate' feature.

Remarkably, researchers can combine the functionality of headnotes and the citation references to narrowly focus their case law research. This allows a researcher to know immediately

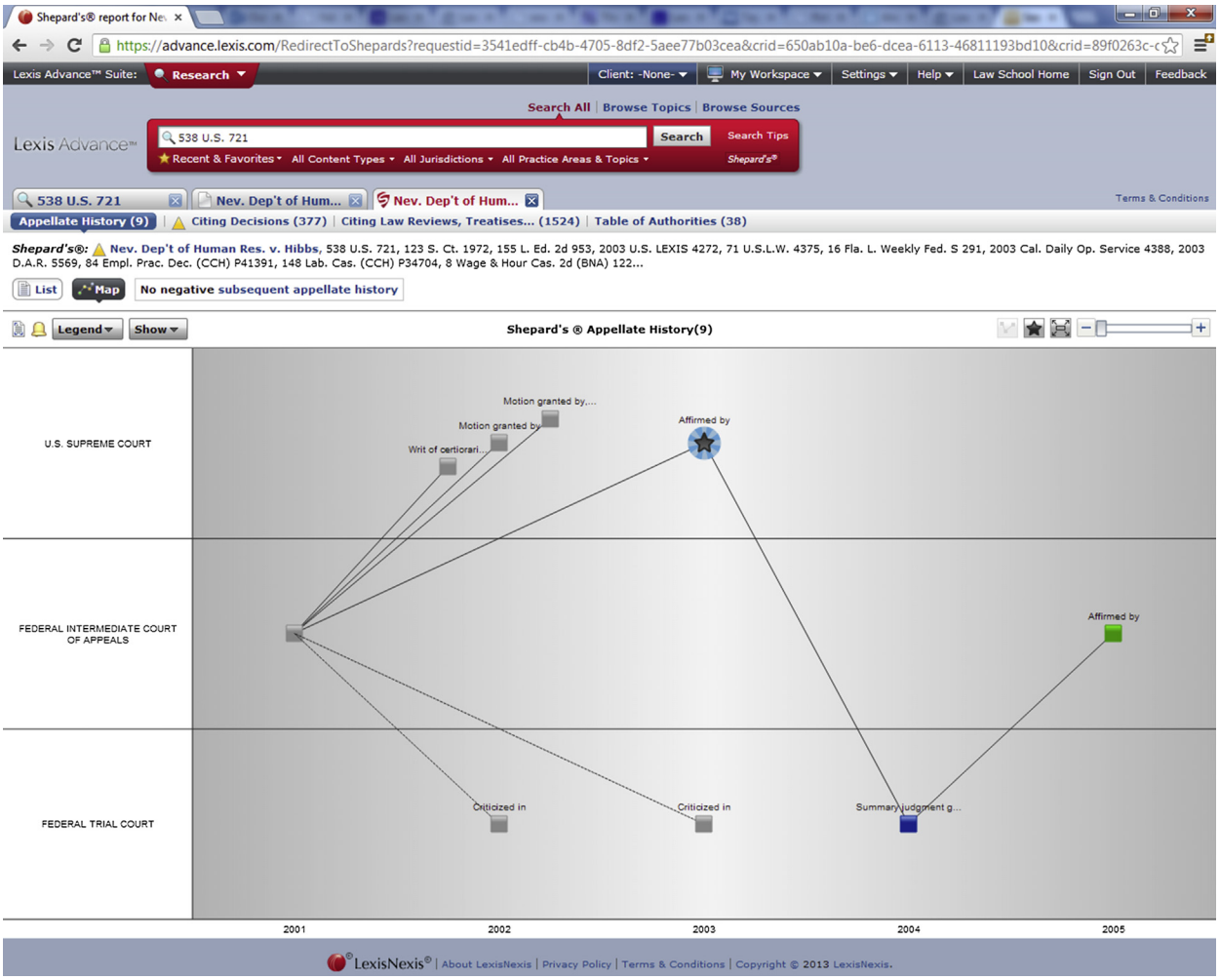

Figure 3 Lexis map.

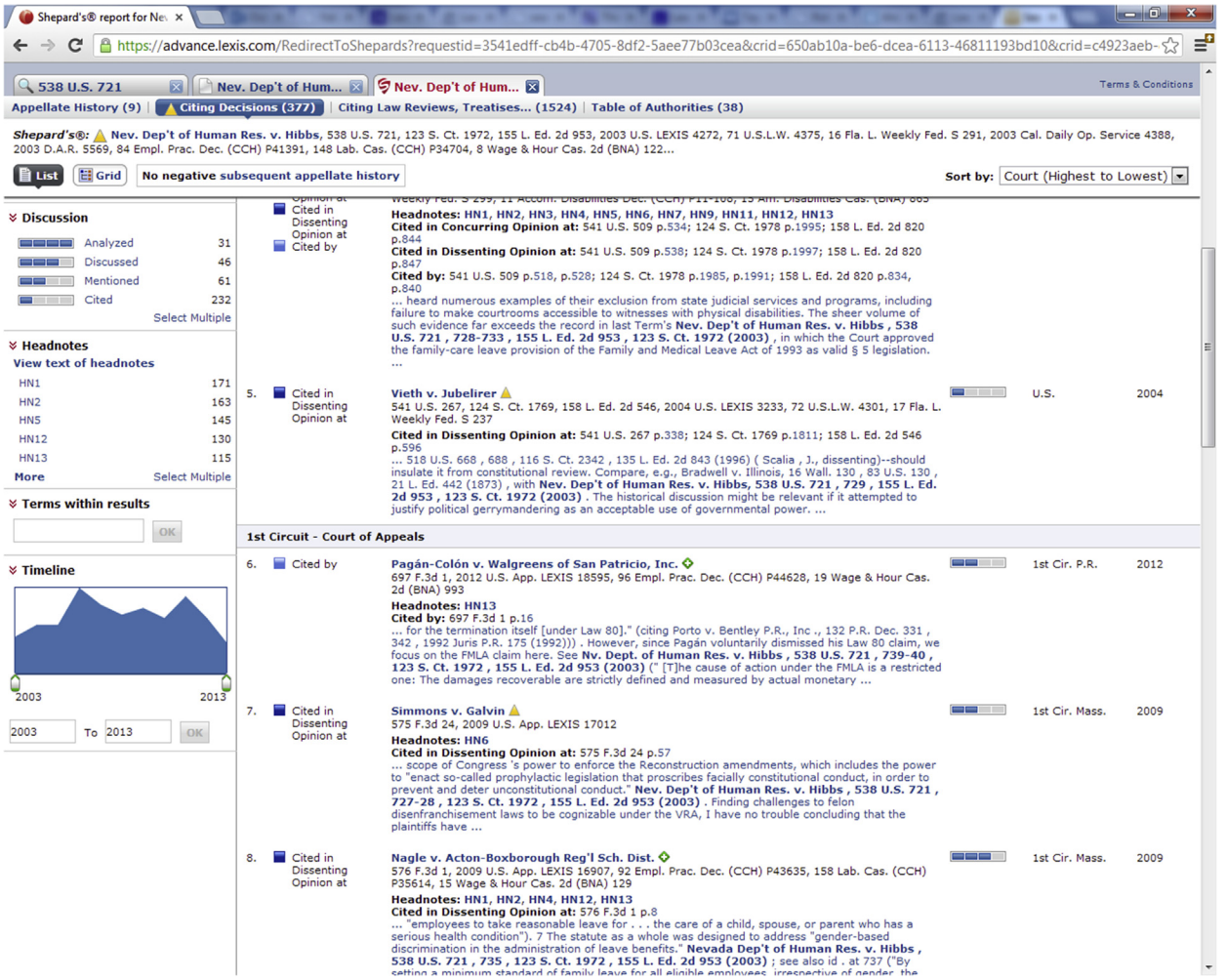

Figure 4 Lexis discussion bar.

whether that point of law is still valid. (Remember that the warning yellow and red indicators for the cases will not specify which point of law has received negative treatment.) Researchers can even require that multiple headnotes appear within the text of a similar case. It is therefore possible to take a case that has thousands of citations and limit the results quickly and reliably to a manageable number. Using the above example for LexisAdvance, it is possible to look for all cases that have received both of the paths 'Labor and Employment > Leaves of Absences > Family and Medical Leaves > Remedies,' and 'Constitutional Law > Equal Protection > Gender & Sex.' This will result in a smaller number of cases, but the cases will be extremely similar.

Organizing Statutes with Legal Databases

Organizational Challenge of Statutes

Statutes are statements of positive law enacted by the legislature at the federal or state level. After the enactment of a federal statute, it is published chronologically in the Statutes at Large. However, Statutes at Large are not generally well indexed so researching them by topic is extremely difficult. In addition, the Statutes at Large are not updated to reflect changes to the text of a law. Laws are therefore divided and organized by subject matter in the United States Code ('USC'). Each volume of the USC covers its own subject matter and these topics are further subdivided into chapters, subchapters, and sections. There are two unofficial annotated versions of the Code: Lexis' United States Code Service ('USCS') and Westlaw's United States Code Annotated ('USCA'). If there were to be a discrepancy in the text of a statute, the language in the USCS would control because it mirrors the language published in the Statutes at Large. West, on the other hand, takes language from the USC. These statutory publications provide more information than the text of the law. The additional information contained in the unofficial versions is crucial to understanding how statutes have changed over time, how statutes interact with other statutes, and how statutes have been interpreted by various courts.

Search Capabilities on LexisAdvance and WestlawNext

When looking for a section of the code, both LexisAdvance and WestlawNext have smart search capabilities. LexisAdvance has a word wheel that brings a researcher into the correct act even if he or she only has a portion of the act name. For example,

typing 'medical leave' will suggest the term 'family and medical leave' and the researcher can quickly retrieve the 'Family and Medical Leave Act of 1993.'

As in print, one can browse the table of contents to the USC or to one of the annotated versions. To access this feature on LexisAdvance, one needs to click the 'browse sources' link above the search bar. After typing 'United States Code Service,' an option for the table of contents will appear. One can search within the table of contents for specific keywords or browse the entire code for relevant sections.

The text of these commercial annotated codes is updated more frequently than the text of the official USC. The USC is purportedly published every 6 years, which means language in the print volume might have been amended, deleted, or struck down by a court. This problem requires print researchers to turn to something called pocket parts: annual publications that supplement the USC. With a subscription service such as LexisAdvance or WestlawNext, changes to the text of the laws are amended directly in the annotated codes. These databases provide links to previous versions of the text as well as the links to the public law which amended the language. The insertion of amendments after the official text enables one to see not only the most current official language, but also changes that have occurred to the text and when the changes occurred.

Annotations

Researchers also benefit greatly from the annotations to the code on these subscription databases. Annotations are helpful references to cases, law review articles, and encyclopedia topics. The annotations are not part of the law itself, but are added to increase understanding of how the law is applied. Annotations group and classify related cases by subject matter. Someone looking at the cases citing to the 'Leave Requirement' section of the Family and Medical Leave Act will be able to identify cases dealing only with the 'expiration' element of the section's language. It is the editorial prerogative of the subscription databases to determine which cases are referenced in the annotations; therefore some annotations might appear in the USCS that do not appear in the USCA.

Shepardize and KeyCite Functions for Statutes

One can also use the Shepardize or KeyCite functions within statutes to create a comprehensive review of cases that have cited a particular statute. Many of the case law citatory filters, such as time and jurisdiction, are available for statutes. On WestlawNext, the 'citing references' filter allows the user to limit by notes of decision. This results in finding cases that have cited the statute, as well as providing the user access to standardized terminology. The standardized terminology in this filter provides users with more accurate results than they could receive by using a keyword search, because editors use the controlled vocabulary to place categorize all similar cases.

Platforms Providing Legislative History

As stated earlier, LexisAdvance and others publish the official text of statutes and provide information on amendments to the text, demonstrating how the text has changed over time. In some instances, one might need to identify why the changes occurred. The process by which one traces a law's development from the time it was a bill is known as legislative history. Students and practitioners can glean important information from changes that occurred in the bill's language, as well as the rationale issued in reports by committees that either endorsed or failed to endorse the bill. There are many important steps in conducting a legislative history, and subscription databases make the process more seamless than print resources. Importantly, many statutes already have compiled legislative histories available, which remove the legwork of tracking down various historical documents.

Conducting legislative history research can be a challenging endeavor when statutes do not already have a compiled legislative history, and there are too many steps to the legislative process to cover in this article. Anyone doing significant research into the legislative history of a statute should consult the commercial database ProQuest Congressional, which provides users with the most comprehensive set of legislative materials. ProQuest Congressional offers full-text congressional publications, including bills and resolutions, as well as both published and unpublished hearings. ProQuest also offers a service called Legislative Insight, which is a federal legislative history service that provides digital compilations of congressional publications leading to US public laws.

Dockets

Dockets are records of court proceedings for individual cases. Most users rely on dockets to research information contained in trials, which may last many years and involve many pleadings and rulings. Many researchers access dockets to find motions, transcripts of a testimony, exhibits, and jury instructions, among other court filings. The primary difficulty in accessing docket information is that dockets vary from jurisdiction to jurisdiction in the type of information recorded and in their availability to the public. Some jurisdictions might place all publicly available information online with open access, while other jurisdictions might require a visit to the courthouse. State court dockets are especially challenging to locate.

At the federal trial court level, the US government has created Public Access to Court Electronic Records (PACER). PACER charges users for each page viewed, so researchers often first try to conduct searches using one of the other databases referenced in this article. LexisAdvance, WestlawNext, and Bloomberg have incorporated PACER results into their own services with dedicated dockets tab. PACER results have their own limitations, and ultimately a trip to a courthouse might be necessary to locate the specific materials that are necessary for research. Most of the commercial vendors will send a runner to a courthouse to retrieve the docket information, for a fee.

Secondary Sources

Law Reviews

Researchers in the United States often consult secondary sources before beginning their legal research (Lihosit, 2009). Secondary

sources, including law reviews, treatises, and encyclopedias, provide critical background information on legal topics and can illuminate a field of law. Secondary sources are not binding authority, but they can sometimes be persuasive to a court. One particularly helpful feature is that they direct readers toward the most relevant cases and statutes. This is especially true of law review articles, which are extensively footnoted.

Academics contribute to the legal field by writing law review articles. Law review articles are one of the places students and practitioners turn to when researching a body of law for the first time, and a strong collection of law reviews are crucial for any library. Every law school has a flagship law review (otherwise known as a law journal) and most schools have secondary journals as well. These law reviews are traditionally edited by a group of students, who choose the content after reviewing submissions.

The legal databases have strong collections of law reviews, although coverage varies by publication. On LexisAdvance, law review articles will appear as a tabulated result under the 'Second' tab after running a keyword search. A keyword search might populate hundreds of thousands of hits, which make filters necessary to focus the results. One useful filter is the 'search within results' box, which narrows the results by introducing another word or set of words. LexisAdvance also provides a 'keyword' box, which suggests a list of potential terms that might be relevant to the search.

The Shepardize function on LexisAdvance works for law reviews as well, enabling users to parse out whether and where the article has been cited by various bodies, including courts and other law reviews. Contentious or influential law review articles will have many articles citing to them, which may indicate a reader should be familiar with the article. This is something one could not easily determine without a legal database such as LexisAdvance or WestlawNext. Google Scholar, for example, provides access to many law review articles but does not provide this additional signal.

As discussed earlier, one can find many articles from law reviews on both LexisAdvance and WestlawNext. These databases have certain limitations in both the breadth and depth of their collections. They also have limitations in their search functions. Because of these limitations, law schools in the United States will subscribe to a number of databases, including Hein and JSTOR, which provide access to additional journals. These full-text periodicals databases contain the actual text of the article in addition to descriptive information. They facilitate browsing and they may also provide enhanced search capabilities to allow for complex searches. Databases such as Hein and JSTOR also usually provide access to scanned pdf versions of the original print periodical, which make them useful for researchers concerned about accuracy. Finally, since many academics are interested in interdisciplinary research, coverage of legal databases is insufficient for their needs. For this reason, universities will subscribe to a host of other databases that provide access to journals in fields ranging from computer science to ethnic studies.

Periodical Indexes

Law schools also subscribe to periodical indexes, which are bibliographic databases that cite articles from a range of publications. These indexes do not always provide direct access to the articles themselves; instead they provide information that makes locating articles easier than locating articles using keyword searches. The indexes contain basic information from the articles as well as subject terms. Indexers assign subject terms to describe the contents of the articles. Articles assigned the same subject term are then cross-referenced to each other, which removes the problems associated with keyword searching (e.g., a keyword search for penitentiary will not retrieve results that use the word jail) and provides more precise results. If one searches for articles on 'self-incrimination' as a subject term, for example, the articles are likely to be narrowly focused on this topic. If one searches 'self-incrimination' as a keyword, on the other hand, results will simply contain the term anywhere in the article's full text. The second benefit of periodical indexes is that they cite to many more articles than are available on any one platform. This prevents the user from having to duplicate searches on different databases.

The two primary legal periodical indexes in the United States are LegalTrac and Index to Legal Periodicals. LegalTrac provides indexing for more than 1400 titles, including major law reviews, legal newspapers, bar association journals, and international legal journals. LegalTrac also contains law-related articles from over 1000 additional business and general interest publications. The Index to Legal Periodicals cites articles from more than 615 legal periodicals. LegalTrac only cites to articles dating back to 1980, whereas the Index to Legal Periodicals includes a Retrospective that cumulates citations from print editions of the Index to Legal Periodicals published between 1918 and 1981. There is also a print version of Index to Legal Periodicals, which goes back to 1908. The vast majority of these publications will not appear on LexisAdvance or WestlawNext.

However, scores of additional secondary materials can be found on LexisAdvance and WestlawNext. Both are particularly rich in treatises, which are books or multivolume sets devoted to particular legal subjects. Many treatises are highly regarded by courts, and courts will cite to treatises in their opinions, particularly when there is no controlling authority or when there are legal issues that have been resolved differently in varying jurisdictions. Courts cite to treatises in far greater numbers than other secondary sources (Merritt and Putnam, 1995). As with most secondary sources, however, the primary function is to illuminate the law for researchers. LexisAdvance and WestlawNext also provide access to American Law Reports, which are articles that provide background and citations to interesting or contentious legal issues; and Restatements, which are distilled summaries of common law principles.

While many secondary sources are published in print, LexisAdvance and WestlawNext provide greater access points to these resources, which can be found by title or keyword search. Perhaps most helpfully, LexisAdvance and WestlawNext often include links to secondary sources after the text of a statute or in the citation references to cases.

Conclusion

Lexis and Westlaw both offer products that make legal research significantly easier and more accurate. Attempting to conduct

serious legal research in this era without the assistance of one of the aforementioned legal databases would be a challenging endeavor. The utility of these products' digests and citation functions as a way to locate similar legal issues cannot be underemphasized. These databases provide access to statutes and legislative history, dockets, and provide a depth of secondary materials that illuminate the law and make conducting legal research significantly easier. While the article focused on benefits, it did not shy away from the handful of instances where LexisAdvance and WestlawNext do not provide adequate coverage for conducting legal research.

Although lawyers and practitioners might only have access to certain databases, the majority of the databases discussed in this article are available to all students enrolled in courses in US law schools. As a result, one can expect a newly graduated US attorney to have experience researching legal issues using more than one database. One downside to the procedure of giving every law student direct access to these databases is that students have unique user accounts. This means there is no institutional access despite the fact that non-law students and academics occasionally need to perform legal research. While non-law students and academics cannot access LexisAdvance or WestlawNext, Lexis and Westlaw both provide an academic platform for universities. However, many of the features described in the preceding pages will not be available to users on these academic platforms.

See also: Common Law; Constitutionalism, Comparative; Courts and Adjudication; Data Bases and Statistical Systems: International Comparative Research; Data Bases and Statistical Systems: Political Science (General); Legal Scholarship.

Bibliography

Amrogi, R., 2012. Law Sites Blog [blog]. www.lawsitesblog.com.

Dabny, D., 1986. The curse of Thamus: an analysis of full-text legal document retrieval. Law Library Journal 78 (1), 5–40.

Daniels, W., 1983. Far beyond the law reports: secondary source citations in United States supreme court opinions October terms 1900, 1940, and 1978. Law Library Journal 76 (1), 1–47.

Doyle, J., 1992. Westlaw and the American digest classification scheme. Law Library Journal 84 (2), 229–258.

Jones, J., 2009. Not just key numbers and keywords anymore: how user interface design affects legal research. Law Library Journal 101 (1), 7–30.

Justiss, L.K., 2011. Survey of electronic research alternatives to LexisNexis and Westlaw in law firms. Law Library Journal 103 (1), 71–90.

Lihosit, J., 2009. Research in the wild: CALR and the role of informal apprenticeship in attorney training. Law Library Journal 101 (2), 157–176.

Mason, D., 2006. Legal information retrieval study—Lexis professional and Westlaw UK. Legal Information Management 6 (4), 246–250.

Merritt, D.J., Putnam, M., 1995. Judges and scholars: do courts and scholarly journals cite the same law review articles? Chicago-Kent Law Review 71 (3), 871–908.

Nevelow, M.S., 2013. The case for curation: the relevance of digest and citator results in Westlaw and Lexis. Legal Reference Services Quarterly 32 (1), 13–53.

Nevelow, M.S., 2010. The relevance of results generated by human indexing and computer algorithms: a study of West's headnotes and key numbers and LexisNexis's headnotes and topics. Law Library Journal 102 (2), 221–250.

Shapiro, F., Pearse, M., 2012. The most-cited law review articles of all time. Michigan Law Review 110 (8), 1483–1520.

Wheeler, R.E., 2011. Does WestlawNext really change everything-the implications of WestlawNext on legal research. Law Library Journal 103 (3), 359–378.

Databases and Statistical Systems: Linguistics

Christopher Cieri, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Linguistic Databases include highly structured paradigms of information about language, such as is found in dictionaries, as well as collections of actual speech and writing. Such collections may include metadata describing how the data were produced or collected and annotation that adds analytic or educational value to the data, for example, transcriptions or translations. In the digital age, linguistic databases are critical components of language research, education, and technology development. As in other empirical sciences, data centers have arisen to address research community needs by creating, distributing, and archiving linguistic databases.

Definition, Examples, and Types

Definition

Linguistic Database is a broad term referring to a set of structured data related to language and, for our purposes, implemented on a computer. The term includes electronic dictionaries and similar databases, as well as the wide range of *corpora* described throughout this article.

Examples

Before defining additional terms it may be useful to give some examples. One of the earliest linguistic corpora is the *Brown Corpus*, created in 1961 from a collection of some 500 text samples totaling more than a million words of American English in 15 categories of writing such as news reports, editorials, book reviews, religious tracts, memoirs periodicals, government documents, technical papers, and fiction. Another example is the *TIMIT* corpus containing recordings of 630 Americans from different dialect regions each reading 10 sentences. The sentences were designed to contain many of the sounds of American English, including some that distinguish different American dialects. For example, all contributors uttered the sentence: "She kept your dark suit in greasy wash water all year," where the absence of the final 'r' in 'dark,' the substitution of 'z' for 's' in 'greasy,' the intrusion of an 'r' in 'wash' and the first vowel of 'water' are characteristic of some dialects. *TIMIT* also includes time stamps showing when in the recording, each word and sound was uttered. This allows researchers to quickly isolate and analyze different occurrences of individual words or sounds as uttered by a single speaker or multiple speakers. The *Penn Treebank* (Marcus et al., 1993) contains more than a million words of news text and conversation marked to show the grammatical structure of each sentence. *WordNet* (Fellbaum, 1998) organizes words into clusters of similar meaning, like a thesaurus, but also creates links among words through other relations, for example, when one word expresses a more specific concept than another (e.g., a cot is a kind of bed), or when one word expresses a part of another (e.g., a mattress is a part of a bed). Created originally for English, there are now WordNets in dozens of languages.

Users of such databases include linguists, or indeed any researcher whose work involves language, language teachers, and developers of human language technologies such as online translation systems, and voice command and response systems now appearing on smart phones.

Types

There are perhaps two main ways of organizing Linguistic Databases. Some contain systematic arrangements of *classes* of related forms and might be labeled *paradigmatic*. Tables of verb conjugations are classic examples of paradigms.

	Singular	*Plural*
First Person	I am	We are
Second Person	You are	You are
Third Person	He, she, it is	They are

However, electronic dictionaries also belong to this type when they contain systematic lists of forms with grammatical and semantic information about each. The second database type is organized around *sequences* of forms that occurred in actual speech or text. One might label this type *syntagmatic*, though the term is not widespread. More commonly, instances of this second type are called *corpora*.

Electronic Dictionaries

Types of Electronic Dictionaries

Electronic dictionaries are the most common and obvious examples of paradigmatic linguistic databases with thousands of instances online. The user experiences them as Web sites or software applications that permit searching and browsing *lexical* information; that is, information about the words of a language. However, behind that interaction is typically a database that contains many, highly structured lexical entries. Each entry will include a *headword*, the form under which information is gathered, and one or more components, e.g., parts of speech, definitions, glosses into other languages, pronunciations, usage notes, related words, time epochs, and genres in which the word appears, example sentences, and frequency counts. Electronic dictionaries may also include multimedia components, such as images that illustrate word meaning, and audio files of native speakers uttering the words.

Headwords and Organization

In Modern English dictionaries, the headword is typically a *citation form*, such as the first person singular of nouns and the infinitive form of verbs. Thus, users may be forced to find the definition of the word *geese* via the entry for *goose* and *broken* via *break*. Entries are also typically ordered alphabetically. However, other approaches are possible. Some older English dictionaries organize entries by meaning, and even modern dictionaries of Arabic may organize entries by root. They gather forms with shared histories and similar meaning or structure, as opposed to those with similar spelling. Electronic dictionaries frequently improve upon the utility of paper dictionaries, whatever their organizing principles, by allowing the user to search the full text as well as the headwords, and by including links to related forms. Finally for languages whose *morphology*, or system of word formation rules, is complex, electronic dictionaries may include a *morphological analyzer* or *stemmer*, a software component designed in this case to identify the probable headword for a given word. Thus a morphological analyzer for French might indicate that the words *mangeons (eat* as in *nous mangeons = we eat)* should be looked up as *manger* (to eat) and even that the word *suis (am* as in *je suis = I am)* should be looked up as *être* (to be).

The electronic dictionaries described so far are digital extensions of traditional paper dictionaries, intended as reference or for learners. Other types exist. *Pronouncing Lexicons*, support work in speech recognition and typically contain entries for all *surface* forms appearing in a corpus along with their pronunciations and frequencies. *Translation Lexicons* may also contain entries for words appearing in a corpus with one or more translations into other languages.

Challenges

Creators of electronic dictionaries face most of the challenges of traditional paper dictionaries plus a few others related to formatting and presentation. Until recently dictionary creators were forced to develop their own methods for storing databases on computers, accepting users' searches and presenting results. Within the past decade, proposals have emerged from different research communities for standards to structure digital dictionaries and permit data exchange. These include the international standard, Lexical Markup Framework, and the Lexicon Interchange Format.

One concern that challenged paper dictionaries publishers, but is effectively absent from their electronic implementations, is space. Publishers of paper dictionaries had to consider the costs and logistical issues associated with the size of dictionaries. However, for electronic dictionaries, space is effectively unlimited. Even a very large dictionary with, say, 100 000 entries plus pictures and sound files demonstrating how to pronounce each word would consume much less storage than a single HD movie today.

Corpora

Corpus (plural corpora), from the Latin for 'body,' refers not only to human bodies and certain bodily structures but also, for at least the past three centuries, to a set of texts collected under some organizing principle. One speaks of the corpus of works by a single author or about a specific subject. More recently the term has also come to mean a collection of texts or speeches used to study language. Here we define corpus as *any body of digital recordings of linguistic behavior developed for a specific, language-related purpose*.

Corpus Components

The form and content of corpora vary with intended uses but generally include: data, metadata, and annotation. The data, sometimes called *raw data*, are the plain recordings of linguistic behavior as audio, text or video files containing spoken, written, or signed language. Some practitioners resist the notion of 'raw' data claiming that any selection of data already encodes decisions about the phenomena to be studied. *Metadata* is information about the data, typically in the form of tables that enumerate various characteristics: the speakers, their demographics, and the circumstances under which recordings were made, etc. *Annotation* refers to the application of human judgment to add value to a corpus, e.g., the transcription of speech, translation into another language, or grammatical analysis. Annotation may involve direct human effort or may be computer-mediated when a computer algorithm annotates data which are subsequently confirmed or corrected by a human. It is worth noting that the algorithm may have also depended upon prior human annotation in such cases.

Users may also categorize corpora by the types of data and annotation they contain, e.g., speech, text, or translation corpora. However, corpora are also categorized by their intended uses.

Uses

Since corpora are developed for specific purposes, before discussing their creation and distribution it may be helpful to enumerate some uses. Language curriculum developers build corpora in order to extract usage and frequency information about words and grammatical structures; they employ this information to decide what aspects of language to introduce and in which order. Dictionary builders use similar information but also examine words in context to identify the multiple *senses* in which they are used. The word *bank*, e.g., may refer to the containing edge of a river or the financial institution. Linguists develop corpora to help document endangered languages and describe the current state of well-established languages that nonetheless change over time. Currently, the largest sources of linguistic data are language technology developer communities that recognize the importance of large, appropriately annotated corpora in their system building efforts. These communities are responsible for thousands of databases containing tens of thousands of hours of audio, billions of words of text, and hundreds of millions of annotation decisions.

Linguistic Databases perform a critical role in both the building and evaluation of human language technologies. A dominant paradigm in modern technology development exploits advances in *machine learning*, a probabilistic approach, in which computers learn to perform a task given enough examples of that same task having been performed on comparable data. For example, a French-to-English machine

translation system may learn from *training data* containing many examples of French documents with appropriate translations into English. Machine learning has led to advances in many language technologies: search engines, online translation services, and smart phone apps that respond to voice command and provide information using spoken natural language. Machine learning approaches also have the advantages that their performance tends to improve as they are given more and better data. Porting them to work on a new language ideally requires only supplying them with appropriate Linguistic Databases in that language.

Evaluating technologies for performance and usefulness also relies on Linguistic Databases. The common approach in human language technologies, as practiced in the evaluation programs administered by the U.S. National Institute for Standards and Technology, is to develop *test data*, new linguistic databases that represent the goal that the technology should achieve. For example, in the French-English translation system, the test set would be a new linguistic database, carefully constructed via human effort, never before seen by the system, containing appropriately selected French documents and their translations into English. During the technology evaluation, an existing system is given the French text and required to produce the translations into English. Those system translations are then compared to the human *gold standard* translations and the differences are scored according to user needs. Some uses may require that the translations contain all of the information in the original. Others may require that sentences in the output are perfectly grammatical. Still others may be able to tolerate some errors but require faster processing.

A few of the many technologies that rely heavily on linguistic databases are

- *Language Identification*: identifies the language(s) of a sample of text or speech.
- *Speaker Recognition*: indicates whether two or more recordings were produced by the same speaker.
- *Speech Recognition* (*Speech-to-Text*): transcribes spoken language indicating each word uttered and typically its speaker.
- *Dialogue Systems*: interact with humans via written or spoken language to accomplish a specific task.
- *Speech Synthesis* (*Text-to-Speech*): produces speech corresponding to input text.
- *Information Retrieval*: identifies documents relevant to a query.
- *Information Extraction*: identifies and categorizes mentions of persons, places, organizations, etc., as well as their activities and the relationships among them.
- *Question Answering*: accepts natural language questions and provides information or complete answers.
- *Summarization*: produces headlines or summaries of various lengths from one or more texts.
- *Machine Translation*: translates a document from one language into another.

Each of these uses requires different types of linguistic databases. For example, language identification from speech requires speech samples in the target languages and some nontarget languages with an indication, for each, of the language spoken. To assure that the technology is truly general, not biased toward a set of speakers, each sample should have only one speaker and each speaker should contribute very few samples. In contrast, speaker recognition requires multiple recordings from each of multiple speakers under different conditions that might include recording in person versus via telephone and speaking casually versus reading a text.

Data Sources

The recordings of linguistic performance that form the data component of corpora come from many sources and fit into four gross categories: text, audio, graphic, and video.

News text, typically acquired in digital form from newswire services or news Web sites, is an important source of well-edited, standard language addressing a broad range of topics, thus providing good coverage of the language's vocabulary. News text is intrinsically interesting and available in great quantities on an ongoing basis from a relatively small number of sources compared with collecting text from individual contributors. Research on specific linguistic genres may rely on corpora built instead from scientific journals or trade magazines or legal or financial documents. Recent work on computer- and phone-mediated textual interactions has required corpora of Web text, newsgroups and other discussion forums, blogs, microblogs, e-mail, chat, and messaging.

Researchers have collected spoken language from broadcast news and talk shows, podcasts, telephone conversations, lectures, interviews, meetings, read and prompted speech, task oriented speech, role playing activities, and Web video. Broadcast news probably makes up the greatest portion of publicly available audio data because it covers a wide range of topics and is available in large and rapidly growing volumes in many languages.

Work on optical character recognition, the creation of computer texts from printed texts, has built upon corpora of digital images of printed and, more recently, handwritten text, hybrid documents, such as printed forms completed by hand and text images appearing within video.

Video sources have typically been broadcast news. More recently, attention has shifted to user-contributed videos now commonplace on the Internet. Other video sources include recordings of interviews and group meetings, and, less frequently, movies and broadcasts that are not news.

Annotation Types

Annotations that researchers may apply to raw data are effectively unbounded; however, we can name some of the commonest. Pustejovsky and Stubbs (2012) treat many of the annotation issues raised here, and others, in more depth and from the perspective of developers of human language technologies.

Before data can be annotated, one must select an appropriate subset. However, data selection is best implemented as a preliminary annotation that marks individual items' appropriateness to subsequent annotation and retains such decisions for future use. To prepare, e.g., a corpus of French newswire with translations into English one might first annotate a large newswire collection for language and text characteristics and

use these annotations to remove documents that are not in French or that contain a predominance of numerical data or names. By retaining the annotations one avoids the duplication of effort that might occur if the corpus were subsequently expanded.

Audio data may be annotated, e.g., for perceived signal quality, language, dialect, speaker and, for broadcast speech, the name of the program recorded. Multiple recordings of the same event using different microphones, sometimes need to be time-aligned so that a researcher may search for words uttered in one recording and hear it in the other aligned recordings. Although there are systems that purport to maintain synchrony between recordings and techniques for recovering time alignment, human effort is still sometimes necessary.

Transcribing speech data makes it much easier to index and search. Transcripts are often *orthographic*, appearing in the writing system used by its native speakers. However, many of the world's languages have no standard writing system. In these cases, and also when the goal of the corpus is to support research in pronunciation, the transcripts may be *phonetic*, ideally using the International Phonetic Alphabet, though such rigor is still relatively uncommon. The time required to create an orthographic transcript of an hour of speech ranges from less than 10 to more than 70 h depending upon the method, expertise of the transcriptionist, degree of care taken, and how disfluent speech, nonspeech events (coughing) and things like acronyms (e.g., NASA, NAACP) are handled.

Researchers have annotated text, including transcripts of speech recordings, to mark the grammatical categories (e.g., nouns, verbs, etc.) and senses of words, to show the boundaries of grammatical structures, and to identify and classify mentions of entities (e.g., persons, organizations), relations, events, time, and locations.

Whole documents may be annotated to indicate their relevance to a search query, or they may be summarized or translated, though it is more useful when individual sentences, not whole documents, of original text are aligned with their translations.

Corpora supporting optical character recognition research have annotated digital pictures of pages to distinguish graphics from printed and handwritten text, to transcribe the text and indicate the order in which it should be read in documents with complex layout.

Metadata

Metadata typically deals with the humans who contribute language to the corpus, and the ways in which their contributions were recorded.

Metadata about talkers and writers tends to focus on personal characteristics that have been shown to covary with differences in language (Yaeger-Dror and Cieri, 2013). Sociolinguists have identified a wide range of these including the speakers' and listeners' sex, age, ethnicity, level of education and occupation, the places where they have lived, and their religious and political affiliations among many others. The recording situation can also impact the language collected. One can identify differences in the speech of a single individual across situations such as speaking to friends and family versus speaking to strangers, communicating in person versus at a distance, and reading versus lecturing. Finally, the speakers' attitudes toward the situation, including attitudes toward the listener and the topic of conversation, also impact the language recorded.

In addition, researchers frequently collect metadata about the method of collection, e.g., the microphones and recording devices used, their placement relative to the speaker, and the dimensions, shape, and characteristics of the recording space. This metadata, covering what are sometimes called *channel characteristics*, are especially important when speech is collected to support technology development.

The reader may have noticed a connection between metadata and annotation in that they are both about the data. The distinction is that *metadata* is the label applied to those characteristics of the contributors and situation that are planned or known in advance, while annotation requires some kind of judgment by humans reviewing the collected data. There are sometimes overlaps, e.g., when a researcher asks contributors to talk about a given topic (metadata) but subsequently reviews the conversation to determine whether they did so (annotation).

Distribution and Archiving

The distribution of a linguistic database to the intended audience in an appropriate form requires significant effort even after the database exists. Many of the issues that are important when the database is planned become critical once it is ready to be distributed.

Standards and Consistency

Appropriate distribution media depend upon data size and the technical capabilities of the intended audience. Current linguistic databases range in size from a few megabytes to nearly a terabyte depending upon whether the raw data is text, audio, image, or video. Appropriate file formats vary independently of media. There exist multiple competing formats for video (e.g., MP4, WMV), audio (WAV/RIFF, MP3), and even images and text, each with its own advantages and promoters. Any digital capture of a continuous signal such as speech or video unavoidably loses information; however, some *lossy* digital formats lose further information in their attempts to reduce file size. For speech where even uncompressed audio is small relative to current desktop storage capacity, it is relatively easy to advise against using lossy compression, such as MP3, for research purposes. However, for video where even a few dozen hours of uncompressed, high-resolution video could fill the average desktop computer, lossy MP4 video is much commoner. However, choosing formats for media files is the simplest of the format challenges. If the researcher has decided on a lossless compression the remaining formats are effectively equivalent and can be converted into each other with freely available tools. Of course conversion to a lossy format truly loses information that cannot be regained by conversion back to a lossless one.

Surprisingly, the situation is not as simple for text even though the adoption of the Unicode Standard for encoding

text has greatly simplified matters. Unicode assigns unique and consistent representations to the characters of the world's writing systems. Before Unicode, creators of electronic texts in languages for which there was not already a standard keyboard and character set were forced to invent their own representations for each new character. Consider the case of someone building a corpus of French, a language that includes *diacritic* marks over some vowels to encode differences in pronunciation, word origin, and meaning. French *mange* means *eats* (*il mange = he eats*) while *mangé* means *eaten* (*il a mangé = he has eaten*). Absent a standard to encode the character *é*, typists invented alternatives such as *e′*, *′e*, *E*, and just plain *e*, ignoring the difference. The lack of standards created inconsistency across and frequently within sources. Even after the introduction of character sets that address this problem for several Western European languages, the inconsistency persisted because the proposed solutions were partial and platform-specific, and achieved insufficient acceptance among users. Unicode has ameliorated this problem to a great degree though there remain numerous writing systems that lack adequate support within applications that purport to be Unicode compliant. *Tifinaghe* is the traditional writing system of languages that outsiders have called Berber or Amazigh and that native speakers call Tamazight and sometimes write ⵜⴰⵎⴰⵣⵉⵖⵜ. Closely related varieties are spoken by more than 15 million people in North Africa, from Morocco in the west to Libya and parts of Egypt and as far south as Niger, Mali, Mauritania, and Burkina Faso. There are also immigrant communities in Western Europe. Although this language has Unicode support, the actual character images of Tifinaghe are not implemented within character sets delivered with most common operating systems. So, writers invent their own representations of Tifinaghe characters. Amazigh has also been written in the Roman and Perso-Arabic scripts and has lacked a standardized representation for many words even within those scripts. One who would compile a corpus of Amazigh must overcome the challenge of normalizing not only the three scripts used, but also the ways those scripts are represented on computer and even the ways a given word is spelled within any one of them. To make this clear, even using the Roman alphabet one can find the name of the language spelled as Tamazight, Tmazight, Tmzight, Tamazirt, and Tamazert, among others. This unwanted variation impedes efforts to build corpora and thus technologies for Amazigh, despite its official recognition in Morocco and Algeria.

There are also numerous competing approaches for encoding annotations, most of which do not enjoy the status of a *de facto* standard. Once the file formats are chosen the distributor must assure that the corpus does in fact conform consistently to those standards. This can be a daunting task with data sets that include, e.g., billions of words of text and millions of annotation decisions.

For databases distributed in fixed formats and media, growth in knowledge about the languages may create a desire to update the databases. On the other hand, if the database is not fixed but allowed to change then its value as a means of evaluating technology is diminished, since even a single system's performance would be unstable if the test set were unstable.

Discoverability

Once the data are in a distributable form, the next challenge is communicating its existence, characteristics, and advantages to potentially interested users, assuring that the database remains discoverable by others who may be interested in the future. The number of researchers interested in linguistic databases changes not only with each graduating class but also with the growing recognition that many fields benefit from language technology. Bird and Simons (2003) deal with discoverability and other issues related to language documentation, especially for endangered languages.

Legal and Regulatory Issues

Distributors of linguistic databases must also face a number of legal and regulatory issues. Within the United States before collecting data directly from humans, even just speech or handwriting samples, one must develop a protocol to assure the protection of subjects and submit it for consideration by an Institutional Review Board regulated by the Federal government. Within most collections of linguistic data, the principal concern is that humans give their informed consent, and that they remain anonymous within the resulting data. However, one can easily imagine data collections with serious health risks. For example, a car manufacturer might wish to develop a voice command and response system that would refuse to start the car's engine if it detected, just by listening to some speech, that the driver were under the influence of alcohol. The data collection necessary to support such a development would presumably include many speakers contributing recordings of voice commands issued while both sober and intoxicated. The administration of the alcohol to subjects would indeed raise significant concerns for their well-being that would need to be addressed under very careful controls. Schiel et al. (2012) describe such a data collection conducted in Germany.

If the raw data in a linguistic database is instead collected from existing sources such as broadcasts, news text, published books, or user-contributed videos then, at least in the United States, a different set of legal constraints come into play. Copyright law gives the creators of such sources the sole right to copy, transfer, broadcast, and commercially exploit their products. This law is designed to encourage innovation by giving content creators the opportunity to be compensated for their efforts. However, the fair use exception to copyright law loosely defines situations in which someone is able to copy another's content, e.g., to build upon it or parody it. Determining which uses are consistent with copyright keeps many lawyers occupied. Complicating matters, laws and regulations differ from country to country and even where international treaties exist to deal with this problem they have not been endorsed universally or enforced consistently.

Reuse

Although we have only touched on a subset of distribution issues, we only have space to discuss one more briefly. Although corpora are developed for a specific purpose, the scarcity of linguistic databases relative to the demand for them

means that many corpora are reused in ways their developers never imagined (Graff and Bird, 2000). Knowing that such reuse occurs, corpus developers need to address quality issues throughout their data and follow best practices beyond those relevant to the initial use of data. Reuse also leads to the development of networks of corpora joined by their reliance on common data and their practice of adding annotation to existing corpora.

Data Centers

Modern desktop computing, storage and networking capabilities allow even individuals and small groups of researchers to create and distribute small-scale linguistic databases. However, developing larger scale databases, and addressing the myriad of technical, managerial, legal, and regulatory issues, as well as attending to challenges of long-term distribution, curation, and archiving, requires specialized equipment and skills. In the early 1990s, human language technology research communities recognized the need for an organization to specialize in this type of distribution, and founded the Linguistic Data Consortium (LDC). During its 22 years of operation the LDC has grown from a distributor and archive of linguistic databases to an organization that collects, creates, and annotates linguistic databases (creating hardware and software systems to assist in those activities). It also coordinates and consults on such activities internationally. Over the past decades other data centers have arisen and followed similar trajectories, such as the European Language Resource Association and the Linguistic Data Consortium for Indian Languages. Although individual researchers, research groups, and commercial organizations still create and distribute linguistic databases, these data centers account for the great proportion of all databases by title and volume and many database creators tend to deposit their work with data centers over time. The reader can visit the sites footnoted for each center to find catalogs totaling thousands of databases.

See also: Corpus Linguistics; Dialectology; Lexicon; Linguistics: Overview; Morphology in Linguistics; Parts of Speech; Phonetics, Articulatory; Quantitative Linguistics; Sociolinguistics; Speech Perception; Writing Systems.

Bibliography

Bird, Steven, Simons, Gary, 2003. The seven dimensions of portability for language documentation and description. Language 79, 557–582.

Fellbaum, Christiane (Ed.), 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.

Graff, David, Bird, Steven, 2000. Many Uses, Many Annotations for Large Speech Corpora: Switchboard and TDT as Case Studies. Proceedings of the Second International Conference on Language Resources and Evaluation. European Language Resources Association, Paris, pp. 427–433.

Marcus, Mitchell, Marcinkiewicz, Mary, Ann, Santorini, Beatrice, 1993. Building a large annotated corpus of English: the penn treebank. Computational Linguistics 19 (2), 313–330.

Pustejovsky, James, Stubbs, Amber, 2012. Natural Language Annotation for Machine Learning. O'Reilly Press, Sebastopol.

Schiel, Florian, Heinrich, Christian, Barfüsser, Sabine, 2012. Alcohol language corpus: the first public corpus of alcoholized German speech. Language Resources and Evaluation 46 (3), 503–521.

Yaeger-Dror, Malcah, Cieri, Christopher, 2013. Prolegomenon for an analysis of dialect coding conventions for data sharing. In: Barysevich, Alena, D'Arcy, Alexandra, Heap, David (Eds.), Methods in Dialectology. Bamberg Series. Lang, Frankfurt am Main, pp. 189–204.

Relevant Websites

http://www.elra.info – European Language Resource Association, a Data Center Providing Many Resources Including an Abundance of Material for Language of Europe.

http://khnt.aksis.uib.no/icame – International Computer Archive of Modern and Medieval English, a Source of the Brown Corpus.

https://www.langsci.ucl.ac.uk/ipa – International Phonetic Association, Includes Charts of the International Phonetic Alphabet.

http://www.lexicalmarkupframework.org – Lexical Markup Framework Website.

http://code.google.com/p/lift-standard – Lexicon Interchange Format Website.

http://catalog.ldc.upenn.edu/LDC93S1 – Linguistic Data Consortium Catalog Page for the TIMIT Corpus.

http://catalog.ldc.upenn.edu/LDC99T42 – Linguistic Data Consortium Catalog Page for the Treebak-3 Corpus.

http://www.ldcil.org – Linguistic Data Consortium for Indian Languages, a Data Center Focusing on the Languages of the Indian Subcontinent.

http://www.ldc.upenn.edu – Linguistic Data Consortium, the Original International Data Center for Linguistics and Source of Hundreds of Corpora and Related Tools and Technical Papers.

http://www.nist.gov/itl/iad/mig – National Institute of Standards and Technology, Multimodal Information Group page. See especially the link for Benchmark Tests.

http://www.unicode.org – Unicode Consortium, Source of Digital Encoding Standard for Many of the World's Languages.

http://wordnet.princeton.edu – WordNet Project, Source of the Original WordNet, in English, Includes a Special Browser for Accessing This Remarkable Database.

Data Bases and Statistical Systems: Overview

Shawna N Smith, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA
Aaron Ponce, Indiana University, Bloomington, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the social and behavioral sciences, the term 'database' refers to collections or aggregations of data sets that can be accessed through a unified repository or archive. Databases often involve data sets that vary over time (e.g., longitudinal data) or space (e.g., cross-national data). Since the 1960s, availability of and access to large databases have increased significantly, spurred by the development of archives across the world willing to store and curate data. As such, an increasing amount of scholarship has been accomplished using preexisting databases. This article also discusses different types of databases within the social and behavioral sciences, and provides considerations for existing challenges and new developments.

A database is any organized collection of information and its supporting structures. Within the social and behavioral sciences, databases tend to refer to collections or aggregations of data sets which can be accessed through a unified repository or data archive. Data sets comprising databases can consist of data (or information) collected at any level of analysis, from the individual (micro) to the organization (meso) to nation-state or country (macro). Indeed databases are often structured to introduce variation at multiple levels of analysis. For example, longitudinal studies, like the Panel Study of Income Dynamics (PSID), may connect changes in social mobility to changes in policy. Cross-national databases, like the European Social Survey (ESS), allow scholars to investigate relationships between country-level and individual-level variables. More broadly, databases frequently focus on aggregating data sets over time and/or space to allow for the exploration of fundamental social science questions, for example, change (or persistence) of social processes, establishing strong institutions, or the impact of macro-level processes or institutions on meso- or micro-actors. As many databases are made available for secondary analysis, or analysis by scholars other than the primary investigators with reference to questions not applicable or of interest at the time of collection, they also provide a repository of already collected data for social scientists to explore either before or in lieu of further data collection. The data archive at the Interuniversity Consortium for Social and Political Research (ICPSR) at the University of Michigan, for example, currently maintains more than 500 000 social science data files.

Databases as a Modern Phenomenon

Although databases are presently seen as central to much of social and behavioral science, the prominence and accessibility of databases are a truly modern phenomenon. Although instances of large-scale data collection (at both micro- and macro-levels) have been recorded dating back to Ancient Greece (Missiakoulis, 2010), databases, and especially those composed of individual-level social and political attitudes, values, and behaviors, are a quite recent phenomenon. Most accounts begin with the rise in importance of social scientists during and following World War II, due to increased government employment of social scientists (Almond, 1996). For example, Stouffer's interviews of more than half a million soldiers for his series *The American Soldier* led to new methodological developments with respect to measuring social attitudes as well as significant revisions to military policy (Stouffer et al., 1949). Around this same time, the Survey Research Center (SRC) at the University of Michigan garnered increased attention for successfully predicting the defeat of Dewey by Truman in the 1948 presidential election. As major pollsters, including Roper-Gallup, had reported a victory by Dewey, the prescient anomaly of the SRC's 662 respondents was enough for the Social Science Research Council (SSRC) to launch an inquiry into the methodological discrepancies (Miller, 1994). The SSRC inquiry led to a grant for the Michigan SRC to conduct another survey of the 1952 presidential election – a continuation which led to the eventual founding of the American National Election Studies (ANES) under the auspices of National Science Foundation (NSF) funding in 1977.

Prior to this point, data collection had been a largely individual pursuit, and scholars generally kept collected data as proprietary, even when funding was public. In discussing the early years of data collection, Warren E. Miller described data as the "scientist's capital," noting further that "they weren't about to share" (Austin, 1982). Following the establishment of the ANES as the nation's premier election survey, however, the scholars behind the study began to look for a cost-effective, 'consortial' way to disseminate the data. In the early 1960s, Miller began approaching political science departments across the country asking them to support an institute at the University of Michigan which would be responsible for permanent dissemination and protection of the ANES studies; in exchange, departments would get access to the full series of data. Although at the time, such a model was largely unheard of, by late 1962 Miller had recruited 21 institutions to join the consortium. Around this same time, Miller wrote a letter to the current director of the University of Michigan's Institute for Social Research (ISR), Rensis Likert, noting that while the SRC had a strong history of collaborations with *individuals*, continual *institutional* collaborations were necessary to ensure the continued development of rigorous survey research and analysis

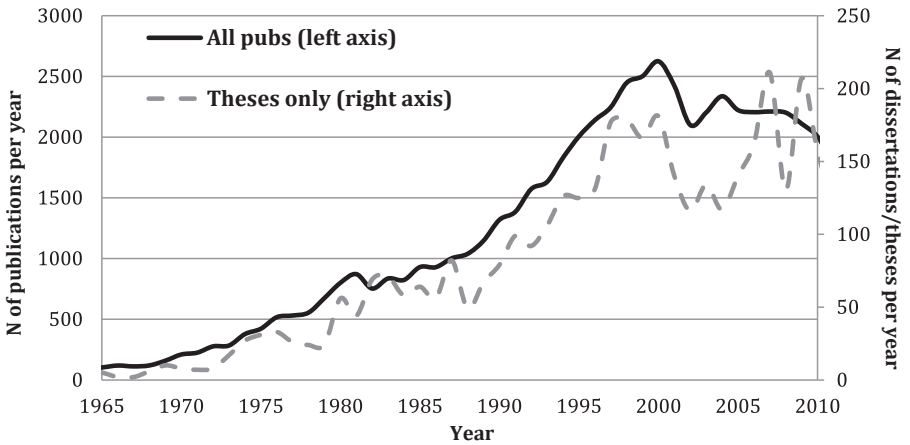

Figure 1 Number of publications per year using data ICPSR databases, 1965–2010.

(Miller, 1962). Miller announced the founding of the Inter-university Consortium on Political Research (ICPR) in a 1963 article in the American Behavioral Scientist, which also noted "over the past year, [ICPR] staff made available to Consortium participants analysis books and codebooks for ten Survey Research Center studies," and listed a number of additional data sets being 'processed,' including Almond and Verba's classic Five-Nation Study, Dahl's New Haven Community Study, and U.S. Census data from 1950 to 1960. At the end of 2013, the ICPSR – which added 'social' to its name in 1975 to reflect growing interest in the Consortium by sociologists – boasted membership from 700 academic institutions and research organizations globally and archives for more than 8500 studies, including nearly 7400 with unrestricted access for members.

Just prior to the founding of the ICPSR at the University of Michigan, the Zentralarchiv für Empirische Sozialforschung (ZA) was founded by the Faculty of Economics and Social Sciences of the University of Cologne, Germany. During the 1960s, one of the ZA's first directors, Erwin K. Scheuch, began to work with Warren Miller and Stein Rokkan to develop some of the earliest infrastructure for comparative research and globalized social science. The Eurobarometer Surveys, for example, were established in 1973 as a joint effort between ZA, ICPSR, and Swedish Data Centers. Both ZA and ICPSR also worked with scholars from other countries to cultivate cross-national and international data archives. The Historical Archive of the ICPSR, established in 1966 began collaborating with French historians in the early 1970s to digitize data from the French census, as well as other industrial, educational, and demographic statistics. In 1986, ZA joined with the Social Science Information Centre in Bonn and the Centre for Survey Research and Methodology in Mannheim to create the German Social Science Infrastructure Services (GESIS), Leibniz-Institute for the Social Sciences. GESIS now serves as the main data archive for several major German studies, including Die Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS), the Eurobarometer, and European Values Survey, and also offers archiving services to scholars looking to preserve and disseminate their data.

Since these early foundings, archives for databases have been established in a number of countries, including a number of countries in Eastern Europe, Asia, and South America. The establishment of archives has been so widespread that several umbrella organizations overseeing data archives have also been established, for example, the Consortium of European Social Science Data Archives, which was established in 1976 to provide access to data across countries, languages, and repositories. (For a full list of data archives by continent and country, see http://www.cessda.net/related_resources/networks_other_archives.html.)

The trend toward databases (and archives), however, has not been entirely supply-side. Since the 1960s, there has been a distinct upward trend in the number of scholars using databases for projects and publications. Figure 1 shows the trend in annual publishing with data available through ICPSR, from 1965 (shortly after ICPSR's establishment) through 2010, as collected by the ICPSR Bibliography of Data-Related Literature, which contains more than 65 000 citations of published and unpublished works using data from the ICPSR archives. The left axis corresponds to the solid black line showing the rise in all publications for a given year using ICPSR data over the period. This line shows a trend of strong yearly growth, starting from the early 1980s and extending into the early 2000s, cresting at more than 2500 works yearly using ICPSR data. Although the plot shows a leveling off starting in the early 2000s in the annual number of papers using ICPSR data, counts still top 2000 articles a year.

A similar trend can be found solely among dissertations and theses, as noted by the gray-dashed line (and corresponding right axis) in Figure 1, in spite of a long-held view in many social and behavioral science fields that PhD candidates should be required to collect their own data (thus demonstrating an understanding of the full scope of the research process). Rather, since the 1980s an increasing number of dissertations and theses have been approved using secondary data. Indeed, an article from the American Psychological Association's *grad*-PSYCH magazine in 2010 quoted one professor as saying it was "a crying shame" for doctoral students to collect new data, given the amount of preexisting, unmined data available from archives like ICPSR (Price, 2010).

Mechanisms behind Change

Although a number of factors have contributed to the increased development, dissemination, and usage of databases, three

mechanisms in particular deserve mention: (1) changes in computing technology that allowed for data to be stored and disseminated at low cost; (2) the establishment of 'gold-standard' research standards that increased the cost of data collection and placed a premium on expertise; and (3) an increased ethos of data sharing, wherein funding providers expect data to be shared with other scholars and preserved in such a way as to provide long-term access to the data. Note that although we refer to these as 'mechanisms,' it is clear that their development may have been concomitant with the establishment and development of databases and archives.

Advances in Computing Technology

Given the fundamental necessity of storage and access for the maintenance and dissemination of databases, advances in computing technology have provided the backbone for increases in database availability and use. Since the 1960s, the cost of data storage has declined from thousands of dollars a megabyte to less than ten-thousandths of a *cent* per megabyte. This decline in storage cost, coupled with the rise of increasingly efficient storage technologies (e.g., magnetic tapes to disk to, most recently, cloud storage) made the possibility of storing large databases, covering multiple countries or many years, possible. The advent of the Internet, further, allowed for efficient and far-reaching dissemination, initially of smaller documents like codebooks, but more recently of full data sets. Scientists no longer have to wait for disks or documentation of data sent via mail; rather they can find and (often immediately) download data sets to their personal computing devices. ICPSR even hosts the Social Science Variables Database, which allows users to search its data holdings for certain variables and compare across multiple databases to find the one(s) that best fit their needs.

One consequence of these computing advances was that new data and documentation are increasingly *created* in electronic format, hastening (if not eliminating) the process of converting data into a shareable electronic format for dissemination. Another consequence, however, was the rediscovery, preservation, and dissemination of 'historical' data. U.S. Census data, for example, can now be accessed online dating back to the 1790s. In 2004, a consortium of universities banded together to create the Data Preservation Alliance for the Social Sciences, established entirely to ensure the preservation and archiving of social science data 'at risk of being lost.'

Establishment of 'Gold-Standard' Research

In his letter to Rensis Likert purporting the establishment of (what would become) the ICPSR, Warren E. Miller noted, "The sample survey is by nature a highly technical and extremely versatile research operation. It requires a formidable combination of specialized personnel, complex methodologies, and organizational apparatus" (Miller, 1962). This early realization that top-notch survey research was both difficult and resource intensive served as a major impetus for forming a consortium of universities. Pooling resources (both intellectual and financial) allowed for the best methods to be used, thus increasing the likelihood of producing the most accurate results. It also allowed for the development of a 'gold standard' for survey research, requiring complete geographical coverage, high response rates, precision sample estimates, and low levels of measurement error (Heath et al., 2005). And while advancements in the areas of coverage, response, and error minimization provided for better and more accurate science, they also had the knock-on effect of making survey data collection an even *more* technical and formidable task. For many scholars considering collecting their own data, the premium for originality, was (perhaps more than) canceled out by the cost of falling short of this gold standard. Scholars without access to the top methodologies and technologies would do better to use 'good' data collected by others than to risk collecting data that might not meet established standards.

This is not to say that all databases currently available meet such 'gold standards.' Indeed, Heath et al. (2005) show that many of the most well-used cross-national databases fall far short of meeting such standards for some, if not most, of the countries included. Other scholars have argued that, with regard to previously unstudied topics, geographic areas, or time periods, that subpar data are preferable to no data. As such, the premium for original data collection has not disappeared entirely; rather those data considered 'original' are simply rarer.

As databases have grown in size and methodological sophistication, they have also been increasingly designed for secondary data analysis, or analysis by scholars not responsible for the study design. For many large surveys, secondary analyses are expected and welcomed. Major longitudinal and cross-national studies, like the ESS, the Survey of Health, Ageing and Retirement in Europe, and the PSID offer training videos and sessions for new users, and conferences for researchers at disparate institutions to connect. Further to this same idea, since 2010, the General Social Survey (GSS) has allowed scholars to submit proposals to add questions to the core GSS battery. In some ways, the 'gold-standard' establishment has led to a bifurcation between those scholars who design surveys and focus publishing on questions of methodology and research design, and those who analyze surveys, with publication focused on broader and more substantive issues.

Increased Ethos of Data Sharing and Curating

In 2009, the journal *Nature* noted, "Research cannot flourish if data are not preserved and made accessible…a research project's success is measured not just by the publications it produces, but also by the data it makes available to the wider community" (145). Although the replication and confirmation of results have been considered an important pillar of the scientific paradigm at least since Kuhn (1970), social and behavioral scientists have generally been slow to make their data available. In many ways, original data are a quintessential public good; although the entirety of the field benefits from more data being made publicly accessible, each individual investigator must be willing to take on the 'cost' that goes in to making their data public. In this case, 'cost' is not only findings potentially lost to other scholars, but also the process of cleaning, documenting, and hosting the data in such a way that it can be accessed and analyzed by other scholars.

Recently, however, other institutions have set about trying to rebalance this equation, and to encourage scholars to share their data, either for purposes of secondary analysis or replication. Data archives, such as those discussed above, are

increasingly offering curation and archival services, reducing both the temporal and financial burden on investigators willing to share their data with others and also providing long-term safeguards for data preservation. More coercively, the major public social and behavioral science funding bodies in the U.S. – namely the NSF and the National Institutes of Health – in recent years have instituted a formal requirement that grant recipients or applicants establish plans for sharing and archiving data. Although these plans are not considered in the proposal review process, they nonetheless require investigators to explicate their intentions for 'managing' their data alongside other well-established ethics of, e.g., participant confidentiality. The UK's Economic and Social Research Council, the country's foremost funding body for social research, and the German Research Council have similarly established guidelines for investigators to discuss 'data management plans' in their proposals.

Recent years have also seen an increase in calls for availability of data for purposes of replication. Several major professional organizations, including the American Sociological Association, American Association of Public Opinion Research, and the American Psychological Association, have addressed issues of replicability in their guidelines for best practice, in all cases noting that scholars should be willing to make their data available for purposes of replication. An increasing number of journals in the social and behavioral sciences have also established policies of 'data availability,' ranging from authors including their data and replication files at submission (e.g., *Journal of Conflict Resolution*, *American Journal of Political Science*, *Political Analysis*) to simply 'expecting' that authors provide replication data and files 'upon request' (Gherghina and Katsanidou, 2013). These changes, alongside calls from well-known scholars like King et al. (1995), King (2006), and Freese (2007), seem to indicate a trend toward data sharing that is unlikely to reverse.

Classifying Databases

Continuing from Campbell (2001), most databases can be thought of as belonging to one of the following categories.

Cross-Sectional Micro-Level Demographic Surveys

Cross-sectional micro-level demographic surveys are typically collected and organized by government agencies as part of their efforts to enumerate their populations while also gathering information about their composition. In the United States, for example, government agencies like the U.S. Census Bureau gather information on the American population periodically by asking individuals basic questions about their households. These data are used for official purposes, such as the apportionment of seats in the legislature and the drawing of district boundaries, and also made available to the public for research purposes. In this sense, the data serve a dual purpose, both influencing official policy as well as increasing our knowledge of population trends through secondary analyses. The U.S. Census Bureau also collects data in important survey programs fielded regularly between census periods: the Current Population Survey (monthly) and the American Community Survey (annually). With respect to international or cross-national research, Census data often provide measures that can be easily understood or compared. The Integrated Public Use Microdata Series (IPUMS), maintained by the Minnesota Population Center, allows scholars access to 1% samples of U.S. Census data from 1850 onward, as well as harmonized Census data from 74 countries, covering more than 544 million people.

Macro-Level Social and Economic Data

Many countries also collect important macro-level data – generally, at the country or state level. For example, the U.S. Department of Agriculture Economic Research Service collects data about agricultural activity in the U.S. as well as data on the prices of important agricultural commodities. The Department of Homeland Security estimates the number of immigrants in the United States and collects the data in its *Yearbook of Immigration Statistics*. These data allow social and behavioral scientists to study large-scale, macro-level trends.

In Europe, many countries have statistical bureaus that collect data on their country's demographic and economic activities. For example, Germany's *Statistisches Bundesamt* (*Destatis*) is responsible for collecting, processing, and analyzing data on the German economy and society. At an even larger aggregate, the European Union's statistical office, Eurostat, collects social and economic statistics for both the EU as a whole and each constituent nation, providing a comprehensive database for study.

International organizations also compile important macro-level data from various countries and offer these data to the public. The World Bank and the International Monetary Fund, for example, provide extensive data on economic development and inequality for most countries. The United Nations Population Division collects data on the demography of UN member countries, including measures of mortality and fertility rates, population growth, and urbanity. The Organization for Economic Cooperation and Development also collects data on social and economic trends for its (largely developed) member countries.

Cross-Sectional Micro-Level Social Surveys

Other sources of cross-national data do not come from government agencies but rather from research programs aimed at documenting public opinions and behaviors across a range of countries. Such programs are typically organized and designed by academics with expertise in specific areas of the social and/or behavioral sciences. Collection of these cross-national data typically involves surveying a nationally representative sample of each country's (or nation's, as appropriate) population, oftentimes in multiple 'waves' over time. The International Social Survey Program (ISSP), for example, is an annual survey program that coordinates the thematic content and fielding of surveys across its member countries. The ISSP has included 53 nations since 1984, developing annual topics that are relevant to researchers in the social sciences and fielding topic-specific modules as 15-minute supplements to regular national surveys (e.g., the GSS in the U.S., or the British Social Attitudes Survey in Great Britain). ISSP surveys give

researchers data on the subjective attitudes and values of national publics with regard to important social, political, and economic issues. The World Values Survey (WVS) is another centrally coordinated cross-national survey project, designed to examine political and sociocultural change. Unlike the annual ISSP, the WVS produces 'waves' on approximately four-year time scales; this longer time scale allows the inclusion of lesser-developed countries not often included in other survey programs.

Nonacademic survey programs with cross-national coverage include the Gallup World Poll and the Pew Global Attitudes Project surveys. While these surveys are guided less by academic research agendas, they are fielded annually (and sometimes semiannually), providing especially up-to-date data on people's attitudes and beliefs from around the world. These programs release data to the public on a release schedule, typically following the release of their own prioritized reports and products.

Other survey programs forgo global coverage to focus on one particular region. The ESS, for instance, has conducted surveys across Europe every 2 years since 2001, exacting some of the highest methodological standards in the field. Other region-specific survey programs include the various 'Barometer' projects (e.g., Afrobarometer, AsiaBarometer, Latinobarómetro, Eurobarometer) and Vanderbilt University's Latin American Public Opinion Project.

Longitudinal Micro-Level Surveys

Distinct from the survey programs discussed above, project-based databases track the same population sample – or panel – over time, typically with the objective of addressing subject-specific research questions. Such panel databases have flourished in recent decades as education, health, and aging and the life course have become important areas of interest to researchers, policy analysts, and government officials. The National Longitudinal Study of Adolescent Health (Add Health), for example, provides data on the general health and well-being of American adolescents by tracking a nationally representative panel of teenagers in grades 7 through 12 into adulthood. The Add Health database has been groundbreaking not only for its design and coverage, but also because it contains data on both respondents' networks (including families, schools, peer groups, romantic relationships, and communities) as well as respondent biomarker data. As such, Add Health data allow investigators to explore the connections between social context and individual outcomes, as well as interactions between biology and the social world. Other longitudinal panel surveys include the National Educational Longitudinal Survey, the pioneering Wisconsin Longitudinal Survey started in 1957, and the PSID, ongoing since 1968.

Ongoing Challenges in Database Construction and Usage

Advances in data storage capabilities combined with the push to collect increasingly more economic, social, and demographic data have led to the establishment of a number of important databases made available by government offices, academic consortia, nongovernmental organizations, and private entities. While these databases offer increasing amounts of information and potential for investigation, these increasingly complex databases also pose a variety of challenges for scholars in the extraction, use, and analysis of data.

Much government data, for example, are often available only in raw formats, which prevent most researchers from easily extracting and using the data. Fortunately, third parties have made efforts to facilitate secondary analysis of government data by researchers. The previously mentioned IPUMS, for example, has collected data from the U.S. Census Bureau dating back to 1850. Prior attempts at providing samples of U.S. Census data for research purposes led to a confusing array of record layouts, coding schemes, and documentation. IPUMS standardizes these by assigning uniform codes and names across all samples and providing coherent documentation of the data. IPUMS also makes available Census samples from a broad array of countries, including the United Kingdom, France, and Switzerland as well as Mexico, Kenya, and Indonesia, with samples data dating back to 1960. To further encourage cross-national work with these data, for a number of measures IPUMS also provides data 'harmonization' – or the integration of data from heterogeneous sources – across countries. Although processes of harmonization may not be ideal for all research purposes, they can help researchers explore similarities and differences across databases that encompass important spatial and temporal variation.

Some government databases can also prove difficult to access. Register databases, for example, while extensive and comprehensive, are often protected by law to ensure individual residents' privacy. Statistical bureaus in many European countries limit register database access to researchers who provide evidence of their need and, thus, are not 'publicly' available in a broad sense. Concerns for respondent privacy also lead to restrictions on a number of survey databases. Often, such databases will offer a publicly available version of the data that exclude certain identifying variables, as well as a full, 'restricted-use' version that requires users to agree to protect confidentiality of respondents, not share data and/or access, and analyze the data only on certain computers.

Even when data are relatively easy to extract and use, there are challenges that arise in the analysis of data. With cross-national data in particular, researchers must determine whether the data available for a set of countries contain equivalent measures for the same concept – that is, whether the data are commensurable. Differences in how survey questions are translated or interpreted among different cultural or linguistic groups, for example, can lead to problems of comparability that even the most advanced of statistics cannot sort out. While those involved in survey planning and administration have become increasingly aware of issues related to the translation of survey instruments (Smith et al., 2011), studies continue to find that cultures using the same language exhibit higher levels of measurement equivalence than cultures using different languages (Davidov and De Beuckelaer, 2010). Issues of cross-national equivalence may persist even as survey products trend increasingly toward the established 'gold standard,' making the role of the analyst in analyzing cross-national data specific and important.

Increased reliance on databases may also raise questions about differences in status between those in the social and behavioral sciences who create and manage databases and those who use databases for secondary analysis. While secondary analysts spend much of their time testing hypotheses and writing up results, those who create and manage databases, in contrast, spend a great deal of their time ensuring that the data – in addition to being scientifically founded – are methodologically rigorous, representative, organized, documented, and accessible. This difference in focus brings about very different final products, which are in turn valued differently. Top-tier journals, for example, often put a premium on first passes at new data or advancement of methodological problems, while scholars retreading old ground (however rigorously) may have a harder time achieving publication. Further, with respect to funding bodies, monies available for the collection of new data far outstrip those available for secondary analyses. As such, and looking forward, we should be aware that disparities in values and resources may risk the creation of a two-tiered system within the social and behavioral sciences.

Big Data and the Future of Databases

The most recent developments in database construction and design reveal a focus on data that do not come from traditional sources such as government agencies, administrative registers, or survey programs. Rather, new sources of data include those produced by unknowing users of everyday technological products, including records kept by companies on the online activities and even physical locations of users. These new sources of data, collected by technology juggernauts such as Google, Twitter, and Facebook, have raised a number of new questions about privacy, consent, confidentiality, and accessibility, and will likely require social and behavioral scientists to consider and codify new ethical and logistical standards for the collection and analysis (or 'mining') of such data.

While improvements in technology have developed largely separate from scholarly data concerns, new opportunities for data collection and database construction nevertheless emerge. With the rise of social media technologies and the digitization of records come traces of raw data left behind by people living their everyday lives. Data from sites like Twitter, cell phone usage data, and digitized spending records compiled by banks and retail merchants all make up this type of data from daily life. Social and behavioral scientists have referred to these organic data as 'big data,' characterizing the scale on which the data are almost effortlessly and consistently produced (Boyd and Crawford, 2012).

Yet, the scale of these databases presents new challenges for scientists. For one, researchers are often not aware of the variety, nor familiar with the quality, of data contained in these records, making the extraction of useful data difficult. Further, the sheer amount and variation within the data found in these data sources prohibit treating them like even the most complex Census database, which tend to be limited and often standardized during any given period. There are also questions concerning the selection of such data as the research standard of representative sampling is difficult to implement and not always practical with such large-N data.

Furthermore, concerns about privacy have stirred controversy. Often, individuals 'producing' big data are not fully aware that they are leaving behind data as they engage with technology and applications. As technology becomes increasingly mobile and 'real time,' individuals are at risk of releasing data from many aspects of their daily lives, interactions, and locations. Such data may be collected by government agencies, private companies, as well as public users; further, modern legal standards pertaining to the expectation of privacy do not apply to noncontent data provided to online sites such as Twitter (Shickich, 2013). As such, big data occupy a legal space somewhere between the public and the private. It is yet to be determined whether the construction and use of such databases should require individuals' notice and/or consent.

See also: Census Microdata; Cross-Cultural Research Methods in Sociology; Data Bases and Statistical Systems: Applied Social Research; Data Bases and Statistical Systems: Demography; Data Bases and Statistical Systems: Economics (General); Data Bases and Statistical Systems: Economics (Macroeconomics); Data Bases and Statistical Systems: International Comparative Research; Data Bases and Statistical Systems: Political Science (General); Data Bases and Statistical Systems: Sociology; International Social Survey Programme.

Bibliography

Almond, G.A., 1996. Political science: the history of the discipline. In: Goodin, R.E., Klingemann, H.-D. (Eds.), A New Handbook of Political Science. Oxford University Press, Oxford, UK, pp. 50–79.

Austin, E., 4 August 1982. Interview of Warren E. Miller.

Boyd, D., Crawford, K., 2012. Critical questions for big data. Information. Communications & Society 15 (5), 662–679.

Campbell, R.T., 2001. Databases, core: sociology. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Elsevier Science, Oxford, UK, pp. 3255–3260.

Data's Shameful Neglect, 2009. Nature 461, 145.

Davidov, E., De Beuckelaer, A., 2010. How harmful are survey translations? A test with Schwartz's human values instrument. International Journal of Public Opinion Research 22 (4), 485.

Freese, J., 2007. Replication standards for quantitative social science why not sociology? Sociological Methods & Research 36 (2), 153–172.

Gherghina, S., Katsanidou, A., 2013. Data availability in political science journals. European Political Science 12, 333–349.

Heath, A., Fisher, S., Smith, S., 2005. The globalization of public opinion research. Annual Review of Political Science 8, 297–333.

King, G., 2006. Publication, publication. Political Science & Politics 39 (1), 119–125.

King, G., Herrnson, P.S., Meier, K.J., Peterson, M.J., Stone, P.M., et al., 1995. Verification/replication. Political Science and Politics 28 (3), 443–449.

Kuhn, T., 1970. The Structure of Scientific Revolutions. University of Chicago Press, Chicago, USA.

Lyngstad, T.H., Skardhamar, T., 2011. Nordic register data and their untapped potential for criminological knowledge. Crime and Justice 40 (1), 613–645.

Miller, W.E., 1962. The Organization of the Inter-university Consortium for Political Research. Correspondence, p. 2. http://www.icpsr.umich.edu/files/ICPSR/fifty/Miller-Likert.pdf.

Miller, W.E., 1994. An organizational history of the intellectual origins of the american national election studies. European Journal of Political Research 25 (3), 247–265.

Missiakoulis, S., 2010. Cecrops, King of Athens: the first (?) recorded population census in history. International Statistical Review 78 (3), 413–418.
Price, M., 2010. More data, less hassle. gradPsych 8 (2), 14.
Røed, K., Raaum, O., 2003. Administrative registers – unexplored reservoirs of scientific knowledge? The Economic Journal 113, F258–F281.
Shickich, D., 2013. What your tweet doesn't say: Twitter, non-content data, and the Stored Communications Act. Washington Journal of Law, Technology & Arts 8 (4), 457–472.
Smith, S.N., Fisher, S.D., Heath, A., 2011. Opportunities and challenges in the expansion of cross-national survey research. International Journal of Social Research Methodology 14 (6), 458–502.
Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A., Williams Jr., R.M., 1949. Studies in Social Psychology in World War II: The American Soldier. Vol. 1, Adjustment during Army Life. Princeton University Press, Princeton, USA.

Relevant Websites

http://www.cessda.net/related_resources/networks_other_archives.html – CESSDA List of Data Repositories.
http://www.ddialliance.org/ – Data Documentation Initiative.
http://www.data-pass.org/ – Data Preservation Alliance for the Social Sciences.
http://www.gesis.org/en/institute/ – GESIS – Leibniz Institute for Social Sciences.
http://www.icpsr.umich.edu/icpsrweb/landing.jsp – Intra-University Consortium for Political and Social Research (ICPSR) at the University of Michigan.
https://www.ipums.org/ – IPUMS at the Minnesota Population Center.
https://www.nsf.gov/bfa/dias/policy/dmp.jsp – NSF Data Management Plan Guidelines.

Data Bases and Statistical Systems: Political Behavior and Elections

Rüdiger Schmitt-Beck, University of Mannheim, Mannheim, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article reviews the global landscape of national and cross-national survey-based infrastructure programs of more or less regularly collecting data on political behavior and citizenship. Two developmental trajectories are distinguished. National election studies were strongly inspired by the American National Election Studies; their global proliferation eventually gave rise to cross-national programs like the Comparative Study of Electoral Systems. The Eurobarometer and the American General Social Survey spurred the development of replicative data collection programs on nonelectoral participation and citizenship. This development culminated in the creation of programs like the World Values Survey and the European Social Survey.

Given the institutional centrality of elections in representative democracies, it cannot come as a surprise that political scientists have always accorded them special attention. Electoral research was the earliest type of inquiry into political behavior to emerge; its pioneer studies date back a full century. Since the mid-twentieth century electoral research has developed into a highly integrated subdiscipline of international political science, and accumulated a massive, mainly survey-based data inventory of national, as well as more recently international, comparative election studies. Electoral research started out to answer two main questions: Who votes in an election, and why? If citizens take part in an election, which parties or candidates do they choose, and again, why? In recent decades, it moved beyond this voter-centric perspective and expanded its scope to broader analyses of elections and their functioning in processes of democratic representation and accountability. Research into other forms of political participation started in the 1960s and opened itself up to internationally comparative perspectives much earlier than electoral research. Comparative research into democratic citizenship more broadly was inaugurated in 1963 with Almond and Verba's famous 'Civic Culture' study. The first programs of replicative national and international representative population surveys on political participation and citizenship started in the 1970s. They developed into a second pillar of survey-based political science data infrastructures alongside electoral studies, with a broader focus than the latter. However, most of these programs did not exclusively cater to the data needs of students of political behavior. They are typically interdisciplinary in orientation, and politics is just one of several mostly sociological themes covered. The data infrastructure generated by these programs is not merely relevant for academic research into political behavior, which seeks to improve scholarly understanding of democracy. It also constitutes a valuable asset for the self-observation of democratic societies and their transformation over time, with powerful potential for enlightening policy makers, journalists, and concerned citizens alike. Its creation and continuous updating is thus an important contribution to the public good of democratic societies.

Data Infrastructures for Electoral Research

From its earliest beginnings, electoral research has defined itself as a quantitative, empirical social science and, as such, it has always been in need of reliable and valid data on citizens' electoral behavior. Qualitative approaches never gained any significance in electoral studies. The pioneer studies of the early to midtwentieth century found a useful source of electoral information in the data collected by statistical offices as part of their administrative duties. Landmark studies like André Siegfried's inquiry of French 'electoral geography' in the nineteenth century (published in 1913), Rudolf Heberle's 'political ecology' of National Socialism in Germany in the early 1930s, Herbert Tingsten's analysis of class voting at the 1932 Swedish election, or V.O. Key's 1955 study on the politics of the American South laid important intellectual and methodological foundations for the discipline by investigating how turnout and vote shares in nation states' territorial subunits, such as electoral districts or counties, covaried with global and aggregate attributes of the same units, ranging from geographic features to cultural and socioeconomic characteristics. However, although profoundly innovative at the time, the scientific scope of aggregate studies such as these was constricted by their inability to observe electoral behavior at the analytical level where it occurred – the microlevel of individual voters. Correlational studies of aggregate data are always vulnerable to the ecological fallacy – the risk of drawing erroneous inferences from aggregate to individual-level relationships. In recent years, statistical techniques have been developed to attenuate this problem, but the heyday of the aggregate approach to analyzing electoral behavior has long since passed. In the 1950s, a trend emerged in American electoral research, which quickly set a new standard in other democracies. Centered around a technique of collecting data on electoral behavior, the new trend became the canonical method of electoral research worldwide to the degree to which it was feasible in the respective countries. The cross-sectional large-N random sample survey, representative of the whole voting age population of a nation, was based on standardized questionnaires and conducted face-to-face. Compared to the data sources dominant during the pioneer era of electoral research, this new approach brought with it at least three major advantages: (1) measurement was located at the level of individual voters; (2) data collection was under the control of scientists themselves, so the core constructs of theories of electoral behavior could be optimally operationalized; and (3) the range of measurable explanatory factors expanded beyond the objective attributes recorded in official statistics to include

 http://dx.doi.org/10.1016/B978-0-08-097086-8.41075-5

also subjective phenomena like values, beliefs, and attitudes. These advantages were traded against a diminished validity with regard to the dependent variable, as research now relied on self-reports by survey respondents instead of records of actual voter behavior (although in the United States, at least for turnout validation is still possible using official voter records).

The new design quickly defined what in time came to be seen as the international 'gold standard' of electoral research. It crowded out not only studies based on aggregate data which retained some significance only in historical election studies but also its greatest competitor, the sociological perspective advocated by Paul Lazarsfeld and his colleagues at Columbia University. This approach conceived of voting as an activity of citizens embedded in, and responding to social groups, and demonstrated its utility during the 1940s in ingenious studies using local multiwave panel surveys. Nonetheless, for a number of reasons, the coming decades were dominated by the intellectual and methodological agenda of what came to be known as the Michigan school of electoral research.

During the 1948 US Presidential election, researchers of the Survey Research Center (SRC) of the University of Michigan led by Angus Campbell created what eventually became the longest uninterrupted series of election studies of the world and a blueprint for election studies worldwide. From 1977 onward, the US National Science Foundation (NSF) funded this project, now relabeled the American National Election Studies (ANES), as a 'national scientific resource'. While data had already been treated as a public good prior to this change, a new governance structure was now established that allowed the scientific community to participate in defining the research agenda. The new structure consists of Principal Investigators as study directors, a Board of Overseers as supervisory and review body, and various ad hoc committees. Project staff is funded to ensure the high methodological quality of data collection.

Regarding its content, the SRC/ANES series' intellectual point of departure is best described by the 'funnel of causality' proposed by its creators in their famous study of 'The American Voter' as a heuristic tool for understanding electoral behavior. It sees voters' psychological orientations as immediate antecedents of electoral choices, most notably beliefs and attitudes concerning issues and candidates, which are assumed to reflect the specific circumstances of particular elections, but also voters' stable partisan identities. Mediated by basic group identities and values, these party affinities can be ultimately traced back to historically rather distant structures of social and cultural conflict. Decades of scholarly debate have transformed, differentiated, and expanded this initial agenda, although the studies' theoretical and methodological focus on independent individuals as core units of electoral analysis remained unchanged over the years.

Despite having created a series of high-quality surveys spanning all presidential and many congressional elections of more than six decades, this extremely successful social science infrastructure program has been placed in jeopardy in recent years, as policy makers have proposed the exclusion of the ANES from the NSF funding remit. This is an ironic development, given that its basic philosophy and design had quickly after its inception turned into an attractive export item, setting in motion, a global proliferation of national election studies trying to emulate its core features: the high methodological standards and long-term continuity; openness to the scientific community with regard to input and output, that is, with regard to instrumentation and design as well as with regard to timely and unrestricted data access; clear and transparent structures of governance and accountability.

Sweden (1956), Germany (1961), the United Kingdom (1963), Norway (1965), and the Netherlands (1971) stand out as early adopters. Many of these new national election studies profited from direct involvement of the Michigan scholars. Today in many democracies, national election studies are conducted more or less regularly at the occasion of parliamentary and, where it applies, presidential elections. Besides the already mentioned countries, the Consortium of European Social Science Data Archives (CESSDA) guide to national election studies in Europe listed already in the mid-1990s studies from Austria, Belgium, Denmark, France, Hungary, Italy, Spain, and Switzerland (Hartmann and Oedegaard, 1997). Outside Europe, similar data collection programs were inaugurated in countries as diverse as Australia, Canada, India, Israel, Japan, New Zealand, and Taiwan. Several studies went even beyond the ANES model, for instance by placing special emphasis on interelection panels (UK) or by adding studies of media and parliamentarians (Sweden).

Despite their *de facto* regularity, these data collection efforts are usually less generously and often also more precariously funded than the American prototype. Only a small number of studies enjoy the security of institutionalization as national science infrastructures with stable funding. Several studies obtain funding by organizations equivalent to the NSF. However, it is often granted only on an ad hoc basis and has to be secured for each election anew. Sometimes funding also comes completely or partly from nonacademic sources, including public agencies like statistical offices, ministries, or parliaments, but also mass media or other private agencies. Securing uninterrupted series of national election studies is no small achievement under such circumstances, and in some countries, such as France, studies are only realized irregularly. Since many programs thus follow incremental trajectories rather than grand designs, data comparability over time is often less than optimal. The degree of openness to the scientific community varies as well. Most studies have subscribed to an ethos of institutionalized accountability to the profession, but some share their data only after an embargo period. Even less well developed are participatory instruments to provide the scientific community with significant influence on the studies' content.

While in recent decades, many countries have established more or less stable programs of national election studies; the field has also substantially expanded its scope in several ways. Traditionally, election studies tended to see individual voters and their attributes as the sole key to understanding elections. This voter-centric paradigm has increasingly given way to a broader perspective that seeks to understand elections as part of multilevel processes of democratic representation. This requires expanding the scope of data collection and analysis so as to refer not only to voters and their reasoning when making up their minds at the polls but also to the manifold interactions between voters, parties, candidates, and incumbents, with traditional and new media functioning as mediating agents. As voter behavior has become more individualized and volatile, election studies have also enriched their designs with

longitudinal components (from short-term panel and rolling cross-section campaign surveys to long-term panels and tracking polls over entire electoral cycles). Some studies also revive the discipline's seemingly outdated interest in aggregate data by combining such information with individual-level voter data for contextual analyses using multilevel modeling.

Another crucial development is the growing interest in internationally comparative studies. Electoral studies were slower at incorporating a comparative perspective than other fields of research into political behavior. National elections are in many respects idiosyncratic affairs (beginning with their dates), and studying them in an internationally comparative framework poses serious challenges in terms of theory, study design, and instrumentation. It seems, therefore, quite natural that the first large-scale comparative election study – the European Election Study (EES) 1979 – was not focused on national elections but on multinational elections to choose the members of a supranational political body: the European Parliament. Since then, EES data have been collected for each European election (partly in connection to Eurobarometer (EB) surveys; see below), although each wave had to face new challenges, especially in securing funds from a variety of sources (Schmitt, 2010). More, importantly, since beginning the EES has sought to move beyond sole reliance on voter surveys by conducting media and party manifesto content analyses, as well as surveys of political elites. During the 2009 European elections, these efforts culminated in a complex pilot study for the creation of an extensive European data infrastructure (Providing an Infrastructure for Research on Electoral Democracy in the European Union (PIREDEU)).

More recently, comparative electoral research also turned to national elections and the question how electoral behavior responds to variations in political systems' institutional architecture. Due to their common intellectual origins, many national election studies appear quite similar with regard to their core constructs, although country-specific measurement solutions were often implemented. Against this backdrop, heroic efforts were undertaken to realize the potential of existing data on European electoral behavior by harmonizing voter surveys from various countries' national election studies and combining them into a single data set. The 'European Voter' project (Thomassen, 2005) initiated this formidable task by merging and analyzing data from six countries over more than four decades (Denmark, Germany, Netherlands, Norway, Sweden and UK). A successor project – 'The True European Voter (TEV)' – began in 2009 with the mission to expand the range of countries to include the rest of Europe. At present, national election studies from 26 countries participate in this collaboration. While these programs aim at output harmonization of existing data, two other networks are building databases using election surveys specifically conducted for the purpose of international comparison. Since 1990, the Comparative National Elections Project (CNEP) has assembled a collection of 25 election studies conducted in 20 countries, ranging from the United States to Mozambique. Two elements are essential for the CNEP: an attempt to revive the neglected research agenda of the Columbia school, which placed an emphasis on institutions and practices of political intermediation (interpersonal communication, mass communication, campaign communication of parties, and other organizations), and a pronounced interest in the electoral politics of new democracies. Due to its focus on communication phenomena, this project has not only generated voter surveys but also media content analyses and data on voters' egocentric networks (Gunther et al., 2007).

The Comparative Study of Electoral Systems (CSESs) is more modest than the CNEP in one respect, which allows it to be more ambitious in several other respects (Klingemann, 2009). It does not require participant countries to run full-fledged election surveys for the sole purpose of integrating them into a cross-national database. It rather asks country collaborators to include a commonly agreed-upon question module of limited size in postelection surveys of their national election studies (and thus accepts variability in survey modes). This makes participation less costly, and as a consequence, the number of countries involved in the CSES is much larger (at present almost 60). Every 5 years, the CSES enters a new cycle of fielding survey modules on a particular theme in as many countries as possible. The first round of the project (1996–2001) focused on the impact of electoral institutions on citizens' political cognition and behavior and the nature of political and social cleavages and alignments, as well as the evaluation of democratic institution and processes, the second (2002–06) on aspects of political representation. The third round (2006–11) is interested in the meaningfulness of electoral choices and the currently running fourth wave (2011–16) has a special focus on electoral mobilization and political knowledge as well as distributional politics and social protection. An important feature of the CSES is that its integrated survey data sets are enriched with system-level data on electoral rules, regime characteristics, and district and overall election outcomes, inviting multilevel analysis. CSES data are treated as a public good. A Planning Committee of elected members is responsible for the program's strategic decisions. A Secretariat, cofunded by the University of Michigan's Center for Political Studies and GESIS – Leibniz Institute for the Social Sciences, takes care of day-to-day management and data work.

Data Infrastructures for Research into Political Participation and Citizenship

The global proliferation of survey-based political data infrastructures beyond electoral studies during the past four decades can be traced back to two crucial initiatives: the gradual implementation of the EB by the European Economic Community since 1962 and the creation of the American General Social Survey (GSS) in 1972. The EB has run biennially as a tool for regularly monitoring public opinion in all member states of the European Union (Reif and Inglehart, 1991) since 1973. About 1000 interviews are conducted in each country per round. The number of countries included in the EB has increased as EU membership has grown. Survey questions primarily cover perceptions and attitudes toward European integration but also more general political, cultural, and social orientations. In the late 1980s, an add-on program of 'Flash' surveys expanded the standard EB. In 1990, the Central and Eastern EB was launched, which would later become the Candidate Countries EB. Although funded and controlled by the European Commission, the EB has, to varying degrees,

been responsive to demands of the academic community. More importantly, the scientific community is granted open access to the data via data archives, which have become an indispensable asset of longitudinal and cross-national comparative research into Europeans' political beliefs, attitudes values, and participatory orientations. The five volumes of the 'Beliefs in Government' series edited by Kaase and Newton in 1995 best exemplify this.

The Global Barometer (GB) is a program that took inspiration from the EB but is independently conducted. In fact, it consists of five successively established but not very tightly integrated, regional programs that include about 75 countries on three continents: the New Europe Barometer (1991) created in a number of recently democratized Central and Eastern European countries, the Latinobarometer (1995), the Afrobarometer (1999), the Asian Barometer (2001), and the Arab Barometer (2005). The GB's thematic focus on orientations toward democracy questions each respective country's social, economic, and political conditions. At the same time, the program is also supposed to assist in improving survey research capacities in countries where such industries are underdeveloped. The various regional programs have achieved different degrees of regularity in data collection, as well as data comparability across time and countries. More importantly, in contrast to most other programs discussed here, not all GB data are available to the scientific public.

Another indirect outgrowth of the EB, the European Values Study (EVS) was first fielded in 1981 in 10 Western European countries. The motivation for creating the EVS came from the sociology of religion, its funding from a private foundation. With basic moral and social values at its core, the EVS focuses on politics only to a limited extent, alongside topics like family, work, leisure time, and religion. The EVS is conducted every 9 years. The last 2008 study covered 47 countries and regions of Europe. The data are freely accessible via the GESIS data archive. While the EVS still goes on, it also branched out into the World Values Survey (WVS), a global investigation of change in social and political values (Esmer and Pettersson, 2007). Since its first wave of 1990–91, the WVS has become the most extensive program of academically driven international survey research. In its five waves, the WVS has covered 80 countries on all continents. What sets the WVS apart from all other academic international programs of comparative survey research is both the high number of participating countries and the fact that it consists of specially conducted surveys. Other programs aiming for worldwide coverage (CSES and International Social Survey Program (ISSP)) are attached to previously existing national programs of survey research. The WVS also seeks to study both democratic and authoritarian countries. Like the EVS, its substantive scope goes beyond political participation and citizen orientations. It also covers topics such as gender roles, work motivations, social capital, tolerance, and well-being. The WVS distributes the data itself. Assemblies of representatives of participating countries govern the EVS and the WVS.

The second strand of data collection programs relevant to the study of political behavior grew out of the American GSS, a study inspired by the social indicators movement of the 1960s and first fielded in 1972 (Smith et al., 2006). Created for the purpose of building an academic social science infrastructure covering a wide range of major research areas of sociology and political science, the GSS adopted a number of basic principles. They include a commitment to the highest methodological standards in every aspect of the survey process, and responsiveness to the data needs of the social science community as well as free, equal, and timely sharing of all data. Another crucial feature is a content philosophy of carefully balancing regularly repeated and innovative instruments. This is to allow for both a continuous monitoring of societal trends on the basis of identical questions and investigating new phenomena respectively new research questions. Great value is placed on detailed and extensive measurement of demographics. The GSS was run annually until 1994 and has been run biennially since then. As a replicative cross-sectional random sample face-to-face survey of the general population, the GSS samples a number of respondents large enough to allow for analyses of important subgroups of society. It is conducted by the University of Chicago's National Opinion Research Center. Like the ANES, it is funded by the NSF as a 'national scientific resource' and directed by Principal Investigators and a Board of Overseers. Similar to the ANES, the GSS quickly gained a reputation as the 'gold standard' in its field. In order to fulfill its promise to become a 'general' instrument, over the years, the GSS has covered numerous fields of interest to social science researchers, including political attitudes and participation, as well as many other topics.

The GSS also quickly became a model for similar programs in other countries. The biennial German ALLBUS was established in 1980 and the annual British Social Attitudes series was founded in 1983. Both studies follow the same principle of combining modules of repeated and new instruments. Similar studies were also created in Australia, Poland, Taiwan, Japan, Korea, and in 2003, even Mainland China. Not unlike the CSES in the field of election studies, networking between various GSS-type national survey programs eventually led to the creation of the ISSP. Since its inception in 1985, the number of member countries has grown from 6 to around 50. The ISSP fields a new survey every year, as a short add-on to a preexisting survey program. The annual modules usually focus on one topic, and most of them are rotated in cycles of around 10 years to allow for long-term comparison. While the national GSS-type studies are conducted via face-to-face interviews, the ISSP combines different survey modes. ISSP data are freely available via data archives like GESIS, both in combination with their national 'carrier' surveys, and as integrated multinational data sets (Haller et al., 2009). The ISSP has a Secretariat but is otherwise rather loosely structured as a network funded by member countries and directed by the assembly of all members. The utility of the ISSP for analyses of political attitudes and behavior depends on the specific research questions. Owing to the add-on principle, which is, of course, vital for its feasibility, its thematic scope is more limited than that of full general social surveys. Alongside themes like social inequality, religion, environment, family, gender roles, or health, past waves of the ISSP have repeatedly included modules on the 'role of government', as well as on national identity and citizenship.

The latest addition to the panoply of internationally comparative survey programs is the European Social Survey (ESS). In everything except coverage (which is not global but

mostly restricted to one continent), it is by far the most ambitious international survey project. In a nutshell, the ESS aims at running a fully fledged comparative survey in a large number of countries in Europe and beyond (Turkey and Israel) with all the substantive and methodological attributes established as hallmarks of the best existing national GSS-type studies (Jowell et al., 2007). In terms of rigor, effort, governance structures, and transparency, the ESS goes to great length to generate data without the substantive and methodological compromises that have to be faced by more modestly funded programs like the ISSP and even the EB. Since 2002, the ESS has completed six biennial rounds of data collection. Its content consists of fixed batteries of questions that are regularly repeated and two rotating modules per wave that are based on competitive proposals from the scientific community. In addition to cumulations of survey data, the ESS also provides users with several kinds of data for context analyses: regionally aggregated data from the survey itself, from various providers of official statistics, and from media-based data on situational circumstances.

While the respective national science foundations have awarded funding for the surveys, the ESS has also received substantial support for central functions of coordination and quality control by the European Science Foundation (ESF) and the European Union. Responsibility for directing the study rests with a Core Scientific Team coordinated by the Center for Comparative Social Surveys at City University London, which obtains advice from a specialized Methods Group and a multinational Scientific Advisory Board composed of representatives of participant countries, the European Commission, and the ESF. In 2005, the ESS received the prestigious Descartes Prize of the European Commission. In 2013, the ESS was officially granted the status of a European Research Infrastructure Consortium (ERIC), which implies its institutionalization as a survey program and will lead to changes in its funding and organizational structure.

Implications and Challenges

Over the last few decades, networks of entrepreneurs in the social sciences have accumulated a wealth of survey data on political behavior and citizens' orientations toward politics, mainly during elections, but also in nonelectoral contexts. To a large extent, this was, in effect, altruistic work, since foremost among its aims was the creation of data infrastructure for the scientific community. In several respects, comparative survey research on political behavior and other key areas of research in the social sciences has reached a new level with the creation of the ESS. This has established a standard against which other programs will be measured in the future but which is extremely difficult to meet because it presupposes enormous investments not only in terms of time and energy on the part of participating researchers but also above all on the part of funding agencies. If there is something like 'big science' in the social sciences, it consists of programs such as this one. Judged by what they achieved for the scientific community with very modest resources, studies such as ISSP and CSES are miracles of scientific and organizational ingenuity and efficiency. The dependency of ambitious national election studies or GSS-type surveys on very significant funds (at least by social science standards) makes them vulnerable to political and administrative interventions, as evidenced by the recent threats to the existence of the ANES. Although the proliferation of academically driven comparative surveys has significantly improved the conditions for research into comparative politics, it also poses many methodological challenges. Minimizing survey error is a daunting task even in quality-conscious single-nation studies, since social, technological, and legal changes constantly alter the conditions under which surveys are conducted. Cross-national research multiplies these difficulties and adds new ones, in particular, the many precautions required to establish cross-national equivalence between surveys with regard to all aspects of their production process. Several comparative survey programs have openly addressed this challenge and contributed significantly not only to substantive but also to methodological research (Harkness et al., 2003, 2010). At least in advanced industrial democracies, it seems, in principle, feasible to aim for fulfilling the strict requirements of the 'best practices' identified by this research field, as the example of the ESS demonstrates. However, it appears doubtful whether these benchmarks are feasible for less developed countries, which are high on the agenda of programs like the WVS. Another foreseeable challenge to both national and international replicative survey programs concerns their relationships to one another. As the number of infrastructure programs increases, calls will emerge demanding more coordination between them. Sooner or later, some of them may even find themselves in competition with one another for scarce funds. Since questionnaire space is limited, finding the right balance between continuity and change in instrumentation has always been a standard problem for replicative surveys of any kind. However, against the background of on-going social differentiation and individualization, a situation may be envisaged where even extensive surveys may no longer be able to carry all instruments needed for adequately measuring social and political phenomena in all their growing complexity. It may become necessary to supplement them with new data collection instruments and to develop new techniques for fusing independently collected data.

See also: Comparative Studies: Method and Design; Cross-Cultural Research Methods in Sociology; Data Bases and Statistical Systems: International Comparative Research; Data Bases and Statistical Systems: Political Science (General); Data Collection: Interviewing; Epistemic Communities and Collaborative Research; International Social Survey Programme; Quantitative Cross-national Research Methods; Replication; Survey Research Centers and Companies; Surveys: Question Wording and Response Categories.

Bibliography

Esmer, Y., Pettersson, T. (Eds.), 2007. Measuring and Mapping Cultures: 25 Years of Comparative Value Surveys. Brill, Leiden and Boston.

Gunther, R., Montero, J.R., Puhle, H.J. (Eds.), 2007. Democracy, Intermediation, and Voting on Four Continents. Oxford University Press, Oxford.

Haller, M., Jowell, R., Smith, T.W. (Eds.), 2009. The International Social Survey Programme 1984–2009. Charting the Globe. Routledge, London.

Harkness, J., Braun, M., Edwards, B., Johnson, T., Lyberg, L., Mohler, P., Pennell, B., Smith, T.W. (Eds.), 2010. Survey Methods in Multinational, Multiregional, and Multicultural Contexts. Wiley, Hoboken, NJ.
Harkness, J., Van de Vijver, F., Mohler, P. (Eds.), 2003. Cross-cultural Survey Methods. Wiley, Hoboken, NJ.
Hartmann, P., Oedegaard, I., 1997. The CESSDA guide to national election studies in Europe 1945–1995. European Journal of Political Research 31, 455–498.
Heath, A., Fischer, S., Smith, S., 2005. The globalization of public opinion research. Annual Review of Political Science 8, 297–333.
Jowell, R., Roberts, C., Fitzgerald, R., Gillian, E. (Eds.), 2007. Measuring Attitudes Crossnationally. Lessons from the European Social Survey. Sage, Los Angeles.
Klingemann, H.-D. (Ed.), 2009. The Comparative Study of Electoral Systems. Oxford University Press, Oxford.
Norris, P., 2009. The globalization of public opinion research. In: Landman, T., Robinson, N. (Eds.), The SAGE Handbook of Comparative Politics. Sage, London, pp. 522–540.
Reif, K., Inglehart, R. (Eds.), 1991. Eurobarometer. The Dynamics of European Public Opinion. Essays in Honor of Jacques-René Rabier. Macmillan, Houndsmill, Basingstoke.
Schmitt, H. (Ed.), 2010. European Parliament Elections after Eastern Enlargement. Routledge, London.
Smith, T.W., Kim, J., Koch, A., Park, A., 2006. Social-science research and the general social surveys. Comparative Sociology 5, 33–43.
Thomassen, J. (Ed.), 1994. The intellectual history of election studies. European Journal of Political Research 25, 239–385.
Thomassen, J. (Ed.), 2005. The European Voter. A Comparative Study of Modern Democracies. Oxford University Press, Oxford.

Relevant Websites

http://www.cnep.ics.ul.pt/index1.asp – Comparative National Elections Project.
http://www.cses.org/ – Comparative Study of Electoral Systems.
http://eeshomepage.net/ – European Election Studies.
http://www.electionstudies.org/ – American National Election Studies.
http://ec.europa.eu/public_opinion/index_en.htm – Public Opinion Analysis sector of the European Commission.
http://www.europeanvaluesstudy.eu/ – European Values Study.
http://www.europeansocialsurvey.org/ – European Social Survey.
http://www.globalbarometers.org/ – Global Barometer Surveys.
http://www.issp.org/ – International Social Survey Programme.
http://www3.norc.org/GSS+Website/ – General Social Survey.
http://www.piredeu.eu/ – Collaborative Project on "Providing an Infrastructure for Research on Electoral Democracy in the European Union".
http://true-european-voter.eu/ – The European Voter.
http://www.worldvaluessurvey.org/ – World Values Survey.

Data Bases and Statistical Systems: Political Science (General)

Max Kaase, University of Mannheim, Mannheim, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes the historical background of developments in political science, which have paved the way for longitudinal comparative surveys and the concept of databases and data archives from the 1950s on.

The initial situation of survey data scarcity in the social sciences has been replaced by a situation of data richness accessible through data archives and other data holding organizations. The article specifically looks at core contemporary projects providing survey data for longitudinal and comparative research with increasing options for micro–macro multilevel analysis as an important element of innovation in empirical political science research.

Data generally give information about units of analyses and/or observations, and are at the center of empirical political science analysis. Data can either come from administrative processes like censuses, or are created by scholars through systematic obtrusive or unobtrusive techniques. Databases are generalized or focused collections of datasets, as held, for example, by data archives. The systematic production of political science data has been stimulated especially through the study of political elections. Based on new survey methodologies, such studies originated in the United States after the Second World War and later spread to Europe and to other parts of the world. International cooperation between leading social scientists improved and facilitated data analysis due to the increasing number of data archives, which permitted free access to datasets. Free access to datasets also encouraged secondary data analysis since existing data could be examined from different aspects and perspectives. At the turn of the twentieth century, the initial scarcity of data was replaced by data abundance not the least through international data archive cooperation which also permitted analyses across time and nations. Thus, empirical political science has been gaining a strong historical dimension. This development is supported by data collections such as the General Social Surveys (GSS) and National Elections Studies nationally, and the International Social Survey Programme (ISSP), the various international Barometer studies and the World Values Surveys (WVS) internationally.

Data and Databases

Data are information on units of analysis or observation. These units can be individuals (micro), organizations like businesses or political parties (meso), or nations (macro). Data can be divided into geographical units like cities, counties, or states, or they can even describe relations like elite networks. Often, units of analysis and units of observation are the same. However, sometimes units of observation, like all family members in a household, can be aggregated if the household is the unit of analysis.

In political science, data are collected through a variety of means. They can be process produced in administrative environments, such as macro or microcensuses for the purposes of societal self-observation and informed policy formation, or they can be created according to (more or less) methodologically sophisticated approaches by scientists, for example, through questioning individuals or other social units, so-called obtrusive techniques. Data can also be derived from unobtrusive techniques like content analyses of printed or video material as well as group observations.

Information on a particular collection of units of observation or analysis placed in a specific time frame and space is called a dataset. Correspondingly, databases are general or focused collections of datasets, as held, for example, by data archives.

The Beginnings

Developments during and after the Second World War

In his account of the history of political science, Almond (1996) pointed out that in the context of the professionalization of the field in the twentieth century, the empirical study of political phenomena blossomed already in the first decades, most noteworthy in the Chicago School, which is associated with such eminent scholars like Charles E. Merriam, Harold Gosnell, and Harold D. Lasswell. He emphasizes, though, that a major breakthrough in the theory, methodology, and practice of empirical political science research was accomplished after the end of the Second World War. During the war, many of the later leading figures in the social sciences were employed in a variety of US government or related public and private organizations with the task to find answers to questions of critical military and civilian importance in the war context. One case in point is the series of four seminal studies of *The American Soldier* directed by Samuel A. Stouffer who surveyed during World War II more than a half million American soldiers to find out about their emotional life, their behavior in combat, and the way they saw their role and position in society.

The Institute for Social Research

The above research was a part of the context, which in the long haul also helped the foundation of the Institute for Social Research (ISR) at the University of Michigan in Ann Arbor by scholars like Angus Campbell, George Katona, Rensis Likert, and Robert Kahn. In 1952, Campbell and Kahn published

 http://dx.doi.org/10.1016/B978-0-08-097086-8.41019-6

a small book on the US presidential election of 1948, which is noteworthy because this first election study by the Michigan Survey Research Center (SRC) was one of the very few, if not the only one which correctly accumulated and analyzed data that proved Truman's win over Dewey. Furthermore, it was innovative in that it was national in scope, used a fairly sophisticated method for selecting respondents, applied a pre–post panel design and operated with a majority of standardized, but also some open-ended questions.

This study is singled out not only because it very well reflects the beginning theoretical and methodological innovations in research on political behavior, but also because it paved the way for one of the most demanding ventures in database construction: the American National Election Studies (ANES).

The American National Election Studies

ANES is a collaborative project producing data on voting and political participation in the United States. Warren E. Miller (1994: pp. 249–250) points out in his recapitulation of the history of ANES that the outcome of the 1948 presidential election resulted in "a public embarrassment to the national community of public opinion pollsters" (because of their falsely picking Dewey as the winner). The successful Michigan study correctly predicting the outcome of that election gained remarkable public and scholarly visibility, resulting in a grant to the Social Science Research Council (SSRC) which in turn commissioned the SRC to also conduct a study of the 1952 presidential election. With the intellectual success of this study (Campbell et al., 1954), the ground was prepared for a continuous series of Michigan-based election studies which finally were transferred into the format of the national political science research resource of the ANES (Miller, 1994: pp. 260–262) funded by the National Science Foundation since 1977.

Influence on European Social Science

The intellectual pull of the Michigan school of electoral research was felt not the least by European political scientists who especially in the 1950s and 1960s came to Ann Arbor to learn about the study of elections and about advanced methods in social research. It is because of this connection that in a fair number of European countries national election studies became institutionalized from the 1970s onward (for an account of this see Thomassen, 1994). What made this development so pertinent is that the Michigan–European interaction helped to prepare the ground for the most important development in creating political science databases: the emergence of academic data archives.

Data Archives – A Core Infrastructure for Political Science

The Necessity of Data Archives

The emergence of academic social science data archives reflects very much the state of empirical political science in the 1950s and 1960s. First, the methodology of social research and particularly survey research had been advanced to a point where it was regarded as a reliable, fruitful way of collecting information on a broad variety of social and political topics found interesting by the research community, but also by social and political elites as well as by the mass media (e.g., see Lazarsfeld and Rosenberg, 1955; Hyman, 1955). However, the time period for which survey studies were then available was still quite short. Therefore, one-shot studies, usually of a cross-sectional nature, meaning the analysis or observation of a unit at a specific point in time, and not of a longitudinal nature, meaning repeated observations over longer periods of time, dominated the field. In addition, high-quality surveys were expensive, funding opportunities were limited, and therefore access of the academic community to such information was difficult. The resulting scarcity of data plus the growing insight that primary investigators were not capable of completely exhausting the analytical potential of a given dataset, paved the road for considerations to make data which were often paid for by public money and therefore in principle a collective property available also to other than the primary investigators after those had their first good shot at data analysis and publication. A further point is also relevant. Often, data collected by primary researchers under a specific conceptual scheme can be reanalyzed under a vastly different scheme and may lead to new insights. Thus, the logic of secondary analysis (Hyman, 1972) of existing data was an important addition to the arsenal of empirical political and social research and triggered an increasing demand for such data. Thus, data archives were the answer to many concerns resulting from the data needs of empirical political science. It is little wonder then that already in 1966, Ralph L. Bisco (1966: p. 93) could speak "of a growing movement within the social sciences: the creation of repositories of social science data."

The Inter-University Consortium for Political and Social Research

There is an agreement in the literature that the oldest general-purpose survey data archive in the world is the Roper Center for Public Opinion Research (Bisco, 1966: pp. 96–97). It was founded in 1947 at Williams College and is now located at the University of Connecticut in Storrs holding data going back to the 1930s. While in many ways it has served as an example for other data archives founded later, the impact of the Inter-University Consortium for Political and Social Research (ICPSR; initially ICPR) at the ISR on the development of data archives in Europe has been more pronounced because it could build on an identity of substantive research interests between American and European scholars.

In his recollection of the foundation of the ICPR, Miller (1989: pp. 152–153) mentions the intersection of two important strands of thought regarding the improvement of the political research infrastructure. The first was the positive impact on the development of at least parts of political science in the United States from the Michigan election studies and, more concretely, from two summer seminars held at the SRC in 1954 and 1958 which were sponsored by the SSRC Political Behavior Committee.

"The attraction of the courses was the ability to analyze the Survey Research Center election surveys with state-of-the-art counter-sorters which were not available at most other campuses" (Weisberg, 1987: p. 282).

Second, in the interaction between the SRC and Stein Rokkan, one of the great European social scientists and comparativists, the SRC took up ideas from a Rokkan report to the Ford Foundation in 1957, proposing the establishment of survey data archives particularly to stimulate comparative research (Rokkan, 1964: pp. 49–62, 1966).

Zentralarchiv für Empirische Sozialforschung

The foundation of the Zentralarchiv für Empirische Sozialforschung (ZA) at the University of Cologne, Germany, preceded the 1962 establishment of the ICPR in time. The ZA was established in 1960 by Guenther Schmoelders, a professor of public finance with many links to the United States and especially George Katona and was later directed by Erwin K. Scheuch, a major figure in international cooperation in the social sciences and in comparative research. Both ICPR and ZA were pioneers in academic data archiving and cooperated closely from their early days on. This innovation owed a lot also to the network of the Committee on Political Sociology of the International Political Science Association and of the International Sociological Association, which brought together scholars like Campbell, Philip E. Converse, Juan Linz, Seymour Martin Lipset, Richard L. Merritt, and Miller from the American and Hans Daalder, Mattei Dogan, S.N. Eisenstadt, Arend Lijphart, Giovanni Sartori, Scheuch, Otto Stammer, and Jerzy J. Wiatr from the European side.

Other European Data Archives

In Europe, other archives followed the Cologne ZA in due time, among them in 1967 the UK Data Archive at the University of Essex in Colchester and in 1971, very much by Stein Rokkans doing, the Norwegian Social Science Data Services (NSD) at the University of Bergen. The latter deserves special mention because its founding was embedded in the establishment of the European Consortium of Political Research in 1970, and was entrusted with the difficult task of stimulating social science data archiving in Europe (Newton and Boncourt, 2010: p. 11). This effort bore fruit when in 1976 the Council of European Social Science Data Archives (CESSDA) was founded (Mochmann, 2009: pp. 491–493; http://www.cessda.org). This organization now (fall 2012) embraces 21 national data archives in Europe and presently is in the difficult process of being transformed into a European Research Infrastructure Consortium (ERIC). This transformation is so important because it promises to put the funding of the national archives on a safe European footing.

It should be added that there is also a worldwide association of social science data organizations (IFDO – International Federation of Data Organizations for Social Science – http://www.ifdo.org). It has more than 30 members with a substantial overlap with CESSDA, but embraces seven members from the United States.

The Historiographic Dimension of Longitudinal Data

In one of his many insightful contributions to the social sciences, Paul F. Lazarsfeld (1972 (1964)) put himself into the role of a 1984 historian and asked about the obligations a 1950 pollster has vis-à-vis this historian. One can carry this torch much farther now in observing that especially data from survey research have opened up a new dimension in historic analysis along the lines of social history. With the advent of representative sample surveys of national population or special subgroups in society, the potential for the *ex post* understanding of microphenomena of an attitudinal as well as of a behavioral nature has been vastly enhanced, and examples of excellent historical analysis based on findings from longitudinal sample surveys are increasing in number (see for instance the analysis by Schildt (1995) on the modernization process in postwar West Germany).

But there is much more to be gained analytically from data, which reach back into the past, and this is even truer when the researcher's interest is in more than one country. In the late 1980s, after heated controversies especially in Europe about the future of democratic government, a stocktaking of public attitudes toward European governments took place between 1989 and 1994 through the Beliefs in Government project directed by Max Kaase and Kenneth Newton (Kaase and Newton, 1995; Abramson and Inglehart, 1998) and solely based on secondary survey data. Obviously, this project would not have been possible without systematic recourse to existing databases held in European data archives. Still, at the same time a number of shortcomings and weaknesses in the availability, equivalence (van Deth, 1998), and quality of the data showed when becoming longitudinal and at the same time comparative in perspective.

Since the 1990s the survey data situation for longitudinal and comparative research has improved to an extent not deemed possible before (Heath et al., 2005; Kittilson, 2007; Norris, 2009). Not all of the data, which are now collected at regular intervals, are directly tuned to political science interests. But all of the collections, which will be addressed later hold information which can be used also for political science topics.

Most of those studies pertain to the individual as unit of analysis, and most of them are done as independent cross sections of the population that is targeted. One important exception is the household surveys conducted in the United States and in some European countries since the 1970s where the unit of analysis is the household and where all members of a given household from a certain age on are included in the study design (Schupp and Frick, 2010). What makes these surveys so complex and expensive, though, is that the studies are conducted as panels where the goal is to interview the same persons across time and thereby measure individual change across time.

Cost is one of the reasons why social researchers are turning away from personal interviews to telephone interviews once the telephone coverage in a given country or population has reached a point where representative sample surveys can be affected. The same situation pertains to Internet surveys, which become increasingly popular. However, whether and when they will become a cost-effective, methodologically satisfactory and probably an altogether superior mode of data collection is an open question (Smyth and Pearson, 2011); to apply them in comparative survey research certainly is a long distance away. And this is even more true for Internet panels capable of assessing individual change across the time dimension. Still,

the Longitudinal Internet Studies for the Social sciences panel in the Netherlands, which started in October 2007, is a project which shows that a representative panel study can be successfully conducted when the necessary resources are invested.

Core Databases in Social Sciences

Reasons for the Abundance of Databases

It was mentioned in the beginning that by the turn of the twentieth century, the initial data scarcity in the social sciences has been replaced by a situation of data richness. There are many reasons for this. With economic growth especially in the Organisation for Economic Co-operation and Development countries since the end of the Second World War, the number and/or size of educational institutions has risen dramatically with larger student populations (measured by shares in birth cohorts), and with this also the number of researchers both within and outside of institutions of higher learning. This development has, of course, embraced the social sciences too, and with more research money around the number of academically driven data gathering research projects has increased substantially. Also, a large number of empirical social science studies is regularly conducted on behalf of and paid for by government agencies, the media, and for private business, often with advice from the academic community. Therefore, quite a few of these data also find their way into academia, and this means mostly through data archives. A specification of which core databases in political science presently exist has to be taken according to explicit criteria. For the purposes of this article, the main selected criteria are political science content, emphasis on microdata stemming from representative sample surveys, longitudinality, international scope, and free data availability through academic data archives. In the following, only the most important such databases will be addressed.

For completeness sake it should be added that there are also macro and aggregate political science databases like the one produced by the Comparative Party Manifestoes Project (Budge et al., 1987; Klingemann et al., 1994). Very pertinent are the databases related to the study of democracies and a more recent emphasis on the quality of democracy (most of these databases are covered in an article by Coppedge et al., 2011). One more recent effort not included in the above discussion is the democracy barometer developed by a group of scholars from the University of Zuerich and the Berlin Social Science Center (Buehlmann et al., 2011).

International Comparisons of National Election Studies

For political scientists the most pertinent data come from the various national election studies, in particular from the ANES. They aim at the heart of the democratic political process – elections – and are longitudinal in nature. The opportunity to look at the electoral process comparatively is provided by these data to the extent that there are enough at least functionally equivalent questions asked across nations and different institutional and political settings to make an international comparison meaningful at all. These claims have been optimally realized by the project Comparative Study of Electoral Systems (CSES; http://www.cses.org). This is a project conceptualized by the International Committee for Research into Elections and Representative Democracy aiming at covering free elections in all parts of the world. This is done since 1996 by attaching a special questionnaire to the post-election studies conducted in the participating countries. Three modules have been fielded in 1996–2001, 2002–06, and 2006–11 emphasizing three different topics. These data covering more than 50 countries are not easy to use because of the widely differing institutional setup of the participating countries, differences in their number and the logical time lags between countries and elections. One big asset is the availability of macro data permitting multilevel analyses (Heath et al., 2005: p. 301; Norris, 2009: pp. 529–530; Klingemann, 2009).

GSS and the ISSP

In 1972, the first GSS was started at the National Opinion Research Center at the University of Chicago. This program was extended through collaboration between research groups from Germany, the United Kingdom, and Australia. Most important, however, was the effort by scholars from these four countries to develop a program of systematic international cooperation, the ISSP (http://www.issp.org). This is a program of comparative survey research, which is defined by drafting groups, which are responsible for the selection of topics to be attached as modules to annual national surveys of the GSS kind and to be repeated for longitudinal analysis in regular intervals (for instance, the module 'role of government' has been run in 1985, 1990, 1996, and 2006). ISSP started with six countries and in 2013 embraced 48 countries from different parts of the world.

Barometer Series of Data Collections

The series began with the Eurobarometers (http://europe.eu.int/comm./public_opinion/index_en.htm). Eurobarometers measure public opinion of certain topics and issues related to the European Union. These are representative sample surveys with usually about 1000 respondents conducted twice a year since 1974 in all countries of the European Union. As the number of EU member countries has grown over the years, so has the number of nations covered by the Eurobarometer. While many of the questions are concerned with special information needs of the European Commission, the data contain enough continuous information on politics to make them very attractive for political scientists (e.g., see Kaase and Newton, 1995). Also, since the demise of communism, comparable surveys have been regularly conducted in Central and Eastern Europe (Eurobarometer East). Furthermore, the Eurobarometer has served as an example for similar surveys around the globe. In the following, the presently (2012) existing barometer data collections are just listed with their web addresses (for details see Heath et al., 2005: pp. 302–314; Kittilson, 2007: pp. 881–887; Norris, 2009: pp. 531–532):

- Afrobarometer (http://www.afrobarometer.org/)
- Arab Barometer (http://www.arabbarometer.org)

- Asia Barometer (http://www.asiabarometer.org/)
- Asian Barometer (http://www.asianbarometer.org/)
- Latinobarometro (http://latinobarometro.org)
- New Baltic Barometer (http://balticvoices.org)
- New Europe Barometer (formerly New Democracies Barometer) (http://www.abdn.ac.uk/cspp/).

Four of the regional barometers (Africa, Asia, Latin America, and the Arabic region) have joined forces to collaborate on substantive and methodological research concerns (http://www.globalbarometers.org).

The European/World Values Surveys

In 1981, the European Values Study Group was founded at the University of Tilburg, the Netherlands. Aim of the study was to examine human values, opinions, and attitudes of European citizens. The group started with surveys in 10 West European societies. The findings of this study were regarded as so interesting that a decision was taken to replicate the study in 1990, but at the same time to enlarge its geographical scope thus becoming the WVS. The second wave of surveys was coordinated by Ronald Inglehart between 1989 and 1993 and already covered 41 nations in all parts of the world. A third wave of surveys was carried out between 1995 and 1998 in 55 countries, the fourth wave of surveys was conducted between 1999 and 2001 in 59 countries, to be followed by a fifth wave between 2005 and 2007 in 87 countries. The sixth wave of surveys conducted between 2010 and 2014 will cover at least 50 countries.

While in total almost 100 countries have participated at least in one survey, the logic of the flexible WVS organization results in a variable geometry regarding country continuity. Details can be found in looking at the project's website also regarding how to access the WVS data (http://www.worldvaluessurvey.org/). There is no question that the WVS is the most encompassing longitudinal and comparative study of values and beliefs worldwide resulting in a plethora of publications. Given the range of countries with very different politicoinstitutional arrangements and value systems, one great asset of the WVS will be its potential for micro–macro multilevel analyses.

The European Social Survey (http://www.europeansocialsurvey.org/)

The Beliefs in Government project (BiG – Kaase and Newton, 1995) mentioned before was the first study, which aimed at analyzing sociopolitical change in postwar democratic Europe and was based exclusively on secondary survey data. The project's vision was to fill a three-dimensional matrix of country, topic, and time, but it soon found out that the available secondary data were not sufficient to reach this goal. This experience was the basis to establish, in the context of the European Science Foundation, a plan to establish a high quality comparative longitudinal survey in Europe (Jowell et al., 2007). Since 2002, the survey is conducted every second year in about 25 European countries, Turkey, and Israel, and applies rigorous methodological standards (Norris, 2009: p. 532). As of now, six rounds of surveys have been conducted (2002, 2004, 2006, 2008, 2010, and 2012). The European Social Survey (ESS) is a combination of a set of core questions to be repeated in every round, and two rotating modules on a variety of different topics (although some rotating modules have already been replicated) which are awarded to European social scientists in Europe at no cost based on a Europe-wide competition. The ESS in all its facets is documented at the highest possible level by the NSD; all data can be directly downloaded from the ESS website at no cost as soon as the surveys are available. The ESS has also from the beginning aimed at opening the road for multilevel analyses, an option which is eased by the availability of contextual data also stored at the NSD. The complex organizational structure of the ESS is, under the auspices of the European Commission, in the process of being transformed into an ERIC which should stabilize its existence.

Perspectives

The abundance of survey data for comparative longitudinal research does not automatically imply that the quality and documentation of the studies is beyond any doubt (Heath et al., 2005: pp. 314–330; Norris, 2009: pp. 535–538). Thus, using such data will require a very considerate approach in conducting analyses, not to speak of new data sources emerging, for example, from facebook which, if becoming available for social science analysis, will require completely new approaches to data analysis. But even more so, the social sciences in general will be experiencing major challenges in many areas, be they conceptual, legal, substantive, methodological, or technological in nature (German Data Forum (RatSWD), 2010).

Summary

Empirical political science over the last 70 years has traveled a long distance from its modest beginnings when ideas of secondary data analysis and for organizations holding, documenting, and making accessible at no or little cost datasets of general interest to researchers began to thrive. Initial situations of data scarcity particularly in survey research have now been replaced by data wealth, creating new and interesting challenges for researchers basing their scholarly work on empirical data from these of sources. In this article, not only has the situation leading to such abundance been sketched out, but also information on the major databases presently in use by the political science research community has been summarized.

Since this information is now not only exploited in research, but also in teaching and policy making, it is now just a question of time to see whether there will be a measurable impact on the better understanding of the dynamics of the social, cultural, and political process in democratic as well as nondemocratic polities, with the concept of nation states still in their center. It will remain to be seen whether the achievements described in this article will keep and enhance their impact on political studies and will remain indispensable instruments of social science analysis or whether they will become standard folklore and lead the way to new frontiers in political science research.

See also: International Social Survey Programme; International Social Survey Programme; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Survey Research Centers and Companies.

Bibliography

Abramson, P.R., Inglehart, R., 1998. Comparing European publics. American Political Science Review 92, 185–190.

Almond, G.A., 1996. Political science: the history of the discipline. In: Goodin, R.E., Klingemann, H.-D. (Eds.), A New Handbook of Political Science. Oxford University Press, Oxford, UK.

Bisco, R.L., 1966. Social science data archives: a review of developments. American Political Science Review 60, 93–109.

Budge, I., Robertson, D., Hearl, D. (Eds.), 1987. Ideology, Strategy and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies. Cambridge University Press, Cambridge, UK.

Buehlmann, M., Merkel, W., Mueller, L., Wessels, B., December 16, 2011. The democracy barometer: a new instrument to measure the quality of democracy and its potential for comparative research. European Political Science. http://dx.doi.org/10.1057/eps.2011.46. Advance online publication.

Campbell, A., Kahn, R.L., 1952. The People Elect the President. University of Michigan, Ann Arbor, MI.

Campbell, A., Gurin, G., Miller, W.E., 1954. The Voter Decides. Row Petersen, Evanston, IL.

Coppedge, M., Gerring, J., et al., 2011. Conceptualizing and measuring democracy: a new approach. Perspectives on Politics 9 (2), 247–267.

van Deth, J.W. (Ed.), 1998. Comparative Politics. The Problem of Equivalence. Routledge, London.

German Data Forum (RatSWD) (Ed.), 2010. Building on Progress. Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences. Budrich UniPress Ltd, Opladen and Farmington Hills.

Heath, A., Fisher, S., Smith, S., 2005. The globalization of public opinion research. Annual Review of Political Science 8, 297–333.

Hyman, H., 1955. Survey Design and Analysis. Principles, Cases and Procedures. Free Press, Glencoe, IL.

Hyman, H.H., 1972. Secondary Analysis of Sample Surveys. Principles, Procedures and Potentialities. Wiley, New York.

Jowell, R., Kaase, M., Fitzgerald, R., Eva, G., 2007. The european social survey as a measurement model. In: Jowell, R., Roberts, C., Fitzgerald, R., Eva, G. (Eds.), Measuring Attitudes Cross-nationally. Lessons from the European Social Survey. Sage Publications, Los Angeles.

Kaase, M., Newton, K., 1995. Beliefs in Government. In: Beliefs in Government Series, vol. 5. Oxford University Press, Oxford, UK.

Kittilson, M.C., 2007. Research resources in comparative political behavior. In: Dalton, R.J., Klingemann, H.D. (Eds.), The Oxford Handbook of Political Behavior. Oxford University Press, Oxford, UK.

Klingemann, H.-D. (Ed.), 2009. The Comparative Study of Electoral Systems. Oxford University Press, Oxford, UK.

Klingemann, H.-D., Hofferbert, R.I., Budge, I., 1994. Parties, Policies and Democracy. Westview Press, Boulder, CO.

Lazarsfeld, P.F., 1972. The obligations of the 1950 pollster to the 1984 historian. In: Lazarsfeld, P.F. (Ed.), Qualitative Analysis. Historical and Critical Essays. Allyn & Bacon, Boston, MA.

Lazarsfeld, P.F., Rosenberg, M. (Eds.), 1955. The Language of Social Research. A Reader in the Methodology of Social Research. Free Press, Glencoe, IL.

Miller, W.E., 1989. Research life as a collection of intersecting probability distributions. In: Eulau, H. (Ed.), Crossroads of Social Science: The ICPSR 25th Anniversary Volume. Agathon Press, New York.

Miller, W.E., 1994. An organizational history of the intellectual origins of the American National Election Studies. European Journal of Political Research 25, 247–265.

Mochmann, E., 2009. Improving the evidence base for international comparative research. International Social Science Journal 59, 489–506.

Newton, K., Boncourt, T., 2010. The ECPR's First Forty Years 1970–2010. ECPR Press, Colchester.

Norris, P., 2009. The globalization of comparative public opinion research. In: Robinson, N., Landman, T. (Eds.), The Sage Handbook of Comparative Politics. Sage, London.

Rokkan, S., 1964. Archives for secondary analysis of sample survey data. An early inquiry into the prospects for Western Europe. International Social Science Journal 16, 49–62.

Rokkan, S. (Ed.), 1966. Data Archives for the Social Sciences. Mouton, Paris and The Hague.

Schildt, A., 1995. Moderne Zeiten: Freizeit, Massenmedien und "Zeitgeist" in der Bundesrepublik der 50er Jahre. Christians, Hamburg, Germany.

Schupp, J., Frick, J., 2010. Interdisciplinary longitudinal surveys. In: German Data Forum (RatSWD) (Ed.), Building on Progress. Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences, vol. 2Budrich UniPress Ltd, Opladen & Farmington Hills.

Smyth, D.J., Pearson, J.A., 2011. Internet survey methods. A review of strengths, weaknesses and innovations. In: Das, M., Ester, P., Kaczmirek, L. (Eds.), Social and Behavioral Research and the Internet Advances in Applied Methods and Research Strategies. Routledge, London and New York.

Thomassen, J., 1994. The intellectual history of election studies: introduction. European Journal of Political Research 25, 239–246.

Weisberg, H., 1987. The silver summer: the ICPSR summer training program's 25th anniversary. PS: Political Science & Politics 20, 281–287.

Data Bases and Statistical Systems: Security and Conflict

Thomas Wencker, Christoph Trinn, and Aurel Croissant, University of Heidelberg, Heidelberg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Conflict and security statistics' refers to empirical data collected and processed for the quantitative and qualitative analysis of the causes, dynamics, and consequences of international and intranational conflicts and for policy making with regard to early warning, risk assessment, as well as conflict and postconflict management. Security in a broader sense refers to the safety from chronic threats to individual integrity and vital freedoms in physical, socioeconomic, and cultural dimensions. Concepts guiding quantitative approaches to conflict and security encompass armed or political conflict, dispute, war, terrorism, human rights violation, genocide, and coup d'état. Research has recently advanced with regard to its empirical scope and data disaggregation. However, challenges related to further disaggregation, the linking of datasets, and the availability and reliability of data remain.

Introduction

'Conflict and security statistics' refer to empirical data collected and processed for the quantitative and qualitative analysis of the causes, dynamics, and consequences of international and intranational conflicts and for policy making with regard to early warning, risk assessment, as well as conflict and post-conflict management. In empirical conflict and security studies, however, the focus is often narrowed to protection from organized violent threats to individuals. Political conflict is the incompatibility between social systems with regard to the security of a population, the integrity of a territory, or the maintenance of a political, socioeconomic or cultural, and national or international order. The concept of political conflict therefore does not include phenomena such as domestic violence or homicide. Conflict and security statistics generally focus on the international and national political systems (systemic levels), and on conflicts either between nonstate and governmental actors, resulting in intranational or transnational conflicts, or among nonstate actors, known as subnational conflicts (configurational levels).

The History of Conflict and Security Statistics

The Origins of Modern Conflict and Security Statistics

The origins of modern conflict and security statistics can be traced back to compilations like P. Quincy Wright's 'A Study of War' (1942) and Richardson's 'Statistics of Deadly Quarrels' (1960). In 1963, the first professional association in the field of statistical analyses of conflicts and security was founded in Malmö, Sweden. Renamed the Peace Science Society (International) in 1973, it is now based at the Pennsylvania State University. Also in 1963, J. David Singer founded the Correlates of War Project (COW) at the University of Michigan (now hosted by Pennsylvania State University). COW was among the first to apply quantitative criteria in defining conflicts, introducing a threshold of war of at least 1000 battle-related fatalities of troops in combat.

In the 1980s, prominent European conflict research institutes were founded. The first one was the Uppsala Conflict Data Program (UCDP), which stands in the tradition of COW's quantitative definition of war. Since 2004, UCDP has closely collaborated in the development of, and data collection for, the 'Armed Conflict Dataset' (ACD) with the Peace Research Institute Oslo (PRIO, founded in 1959). In addition, both PRIO and UCDP have started georeferential conflict data coding: PRIO with its 'Armed Conflict Location and Event Dataset' (ACLED) and UCDP with its 'Georeferenced Event Dataset' (GED). In the early 2000s, new hubs of conflict and security statistics emerged at the ETH Zürich (Lars-Erik Cederman) and the University of Essex (Kristian Skrede Gleditsch).

In contrast to the Scandinavian tradition, two German research institutes have turned away from COW's quantitative approach. Instead, the Arbeitsgemeinschaft Kriegsursachenforschung (AKUF) at the University of Hamburg, established in 1986, and the Heidelberg Institute of International Conflict Research (HIIK), founded in 1991, focus on the quality of conflict processes and structures, broadening the conceptualization and measurement of conflict. Until 1998, the HIIK maintained the Konflikt-Simulations-Modell (KOSIMO) dataset, a spin-off of which is the Conflict Information and Analysis System (CONIAS) at the University of Heidelberg, in existence since 2005. Since 2011, CONIAS and HIIK have moved toward spatial and temporal disaggregation of their conflict data.

Initiated by Ted R. Gurr, in 1994 the State Failure Task Force was established. Renamed Political Instability Task Force (PITF), it is a research panel funded by the US Central Intelligence Agency and concerned with state failure events manifested in revolutionary wars, ethnic wars, contested regime changes, and genocides. It is closely related to the Major Episodes of Political Violence (MEPV) dataset by the Center of Systemic Peace (CSP), an institution founded in 1997 and especially known for the Polity regime datasets.

Among several other datasets on conflict and security are, in particular, the civil war compilation by James D. Fearon and David D. Laitin (2003), the Armed Conflict Database by the International Institute for Strategic Studies in London (since 1958), as well as the International Crisis Behavior Project dataset by the Center for International Development and Conflict Management (CIDCM) at the University of Maryland.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.41076-7

CIDCM is also home of the Minorities at Risk (MAR) Project, established in 1986 by Ted R. Gurr, on politically relevant ethnic groups and organizations and ethnopolitical conflicts.

Extensions toward Other Forms of Political Violence

In addition, more specific forms of political violence such as terrorism, human rights violations, genocide, and coups d'état have also become subject of conflict and security statistics in recent years. With regard to terrorism these include

- the Global Terrorism Database (GTD) by the National Consortium for the Study of Terrorism and Responses to Terrorism (START), founded in 2005 by the US Department of Homeland Security;
- the GTDs I and II as provided by the National Archive of Criminal Justice Data (NACJD);
- the Terrorism and Preparedness Survey Archive, jointly managed by START and NACJD and funded by the US Departments of Justice and Homeland Security;
- the High Casualty Terrorist Bombing Events (HCTB) dataset by CSP;
- the International Terrorism: Attributes of Terrorist Events (ITERATE) database;
- the RAND Database of Worldwide Terrorism Incidents (RDWTI).

Furthermore, the Cingranelli-Richards (CIRI) Human Rights Dataset and the Political Terror Scale (PTS) contain information on government respect for internationally recognized human rights, political violence, and terror. With regard to genocides and other instances of mass killings of civilians, Barbara Harff and Gurr in 1988 developed a comprehensive data collection of genocides and politicides since the Second World War. Benjamin Valentino compiled a list of mass killings of civilians in the twentieth century in 2000. Since 1994, the State Failure Task Force's State Failure Problem Set also includes data on genocides and politicides. Based on these data, Harff published a risk assessment model in 2003, which is regularly updated. Apart from these efforts with a global scope, the Yale University-based Genocide Studies Program (GSP, founded in 1998) hosts case-oriented databases on the genocides in Cambodia, Rwanda, East Timor, Guatemala, and the Former Yugoslavia.

Finally, since the 1960s, data on coups d'état were compiled for the purposes of specific analyses and presented in the appendices of articles or books investigating their causes, types, and consequences. In 2007, Monty Marshall and Donna Marshall of CSP developed a comprehensive coups dataset that is regularly updated. In 2011, Jonathan M. Powell and Clayton L. Thyne compared and merged 14 existing datasets on coups into a single dataset. Their work is also considered in the recent update of the CSP coup d'état events dataset.

Concepts and Measurement

Definitions of Political Conflict

Researchers employ different terms to label the empirical phenomena under investigation. The choice of a specific term is important as it influences academic and public debates that may arise around conflict and security statistics. On the most fundamental level, these terms encompass 'war,' 'armed conflict,' 'political conflict,' 'militarized interstate dispute (MID),' 'major episode of political violence,' or 'ethnopolitical conflict.' Pertaining to similar phenomena, the terms are situated on different levels of abstraction and emphasize specific characteristics of political conflicts. The level of abstraction and the number of defining attributes shape the extension of the concept, i.e., its empirical scope. Some institutions use specific terms such as 'war' or 'crisis,' whereas others employ wider definitions of political conflict and build subtypes. Widespread, although not ubiquitous, is the reference to three dimensions of conflicts: conflict actors, conflict items, and conflict measures. A possible fourth dimension, the intensity of conflicts, is a property derived from the (inter)actions of the conflict parties and can be subsumed under conflict measures.

The COW project's definition of 'war' presupposes three necessary attributes: sustained combat, the involvement of organized armed forces, and a death toll of at least 1000 battle-related fatalities within a 12-month period. The implementation of a threshold of fatalities has deeply influenced subsequent research. It is the single most influential attribute and indicator to identify and differentiate types of violent political conflict.

Similar to COW's definition of 'war,' CSP defines 'MEPV' as periods in which organized groups engage in systematic and sustained use of lethal violence resulting in at least 500 directly related deaths over the course of the period. UCDP employs an even lower threshold in their definition of 'armed conflicts' (synonymously 'state-based conflicts'). 'Armed conflicts' require the occurrence of at least 25 battle-related deaths in a calendar year. In addition, the definition rests on three necessary attributes: (1) the existence of an incompatibility over government or territory, (2) the use of armed force, (3) the involvement of at least one national government. What makes UCDP's definition distinct from other concepts is the inclusion of a specific conflict item as the object of incompatibility; 'armed conflicts' are either about the type of political system or the replacement or change of composition of the central government, or about the status of a territory. In its definition of 'nonstate conflicts,' UCDP drops the attribute of a specific type of incompatibility, but requires both sides of the conflict to be nongovernmental.

Concerning conflict measures, COW's 'war' definition is the most restrictive one as it employs the highest threshold of battle-related fatalities of the aforementioned concepts. The COW project, however, also includes conflicts below this threshold, labeled 'MIDs.' 'Militarized' refers to the threat, display, or use of military force short of war; 'interstate' excludes all those militarized disputes where at least one side is not a state; and 'dispute' refers to the contestation over unresolved issues. An 'MID' is instantiated by at least one 'militarized incident.'

MAR's founder Ted Gurr (1980) defines conflict as "the overt, coercive interactions of contending collectivities," i.e., the observable use of force between state or nonstate groups. CONIAS defines 'political conflict' as a positional difference with three attributes: (1) the difference exists among at least two decisive and directly involved actors; (2) their actions and communications lie outside established regulatory procedures

and threaten core state functions; and (3) the positional difference regards values relevant to a society.

Concepts of Specific Forms of Political Violence

Among the concepts covering threats of physical integrity and security are 'terrorism,' 'human rights violation,' 'genocide,' and 'coup d'état.' 'Terrorism' is a highly contested concept. CSP's HCTB dataset employs three attributes to define a terrorist event: (1) bomb attacks resulting in at least 15 fatalities, (2) civilian and political noncombatant targets, and (3) nonstate actors as perpetuators. START's GTD definition of terrorism rests on three necessary conditions: (1) perpetrators are nonstate actors who (2) use or threaten to use illegal force and violence to create coercion or intimidation in order to (3) attain a political, economic, religious, or social goal. The older GTD II, on which START's GTD is based, used a more elaborate definition. In particular, it stressed the communicative character of terrorist acts and defined illegal force as being outside the parameters of international humanitarian law. Employing four criteria, RDWTI defines terrorism as the use or threat of violence directed against civilian targets by a group or individual who aims at attaining a political goal by inducing fear.

CIRI defines violations of human rights as government disrespect for the physical integrity of individuals, for civil liberties, workers' rights, and gender equality. PTS measures state-sanctioned killings, torture, disappearances, and political imprisonment on a five-level pervasion scale. While both are similar regarding actors, PTS directly refers to the level of indicators.

CSP's compilation of genocides and politicides in the State Failure Problem List rests on the definition of Harff and Gurr, who use three attributes to define the two concepts: (1) the promotion, execution, and/or implied consent of sustained policies by (2) governing elites, their agents, or contending authorities in a civil war that are (3) intended to destroy, in whole or part, a communal, political, or politicized ethnic group. The criterion of defining the victimized group can be based on either communal characteristics (genocide) or as political opposition (politicides). The related phenomenon of mass killings is, according to Benjamin Valentino, defined by (1) a massive number of noncombatants killed, operationalized as 50 000 fatalities in a 5-year period, with (2) casualties caused intentionally and not considered as 'collateral deaths.' This includes the fatal consequences of direct or indirect use of violence, e.g., by creating conditions that cause fatalities.

Wider concepts of political conflict regard coups d'état as a subtype of intrastate political conflict. Accordingly, the concept of coup can be understood as a specification of the three defining attributes of political conflict. While the formulations and specifications vary, coups are usually defined as the use or threat of force by state elites aiming at overthrowing the government. Concerning the item, a coup d'état can be regarded as a conflictive form of regime or government change. Regarding measures, and contrary to more narrow concepts of violent political conflict, the actual use of violence is generally not regarded a necessary attribute of a coup d'état. However, definitions usually refer to the extralegality or the threat of force. Concerning actors, coups are defined as primarily originating from a country's elite, demarcating them from regime changes due to popular uprisings or external intervention. The definition of coup d'états by Marshall and Marshall draws on four attributes: (1) the use of force by (2) an oppositional faction within the country's ruling or political elites that leads to (3) the capture of executive authority and office, and (4) a substantial change in the leadership of the government and its policies. If the new government holds power for at least 1 month, a coup is categorized as successful. Powell and Thyne's new dataset focuses on coup attempts, i.e., (1) the activity is illegal, (2) it is undertaken by any elite, which is part of the state apparatus, and (3) it aims at replacing an incumbent government.

Empirical Scope

Most relevant for conflict and security statistics is the extension of the basic concepts, which derives from the extension of the individual attributes (conflict actors, items, and measures). Concerning conflict actors, 'war' (COW), 'MEPV' (CSP), and 'political conflict' (CONIAS) are the most extensive concepts. They subsume all possible constellations between state and nonstate actors within one term. UCDP differentiates two terms to refer to conflicts with or without state participation, whereas COW's MIDs only comprise disputes between states. MAR covers only conflicts involving ethnopolitical groups, i.e., ethnic identity groups that constitute the focal point of grievances and political action.

UCDP's definition of 'armed conflict' is restricted by two concrete conflict items (government and territory). MAR subsumes under the term 'ethnopolitical conflict' either inter- or intracommunal conflict, protest, rebellion, or government repression. The concepts of 'political conflict' (CONIAS) and 'MID' (COW) are more extensive, as they refer to items in a much more abstract sense, whereas 'war' (COW), 'MEPV' (CSP), and 'nonstate conflict' (UCDP) do not mention conflict items at all. Concerning conflict measures and their consequences, COW's 'war' definition is most restrictive with a necessary 1000 battle-related deaths in a 12-month period, followed by CSP's 'MEPV,' where the number of casualties is not linked to a specific time frame. COW's 'MID' include interstate disputes below the threshold of 1000 deaths. 'Political conflicts' in CONIAS include violent as well as nonviolent forms of conflicts. While MAR's definitions of inter- and intracommunal conflict and rebellion require the presence of (not necessarily fatal) violence, the categories of protest and repression include nonviolent actions, as well.

Whether to employ a broader or narrower definition of political conflict depends primarily on the research interest. While concepts with a larger empirical scope allow subsuming different forms of conflict under a single term, the entity under investigation may become less specific. A narrow definition, on the other hand, allows for more precise and extensive research with a smaller number of cases. Furthermore, concepts and corresponding datasets with identical institutional origin, such as COW's 'war' and 'MID' or UCDP's 'armed conflict' and 'nonstate conflict,' can easily be merged to form wider concepts. Wider conceptual approaches, in contrast, may increase specification by building subtypes through a within-differentiation of the attributes and thereby limit their extension.

Concerning the empirical scope of terrorism databases, HCTB only includes bomb attacks, whereas START and RDWTI include other use of threat of violence. In contrast to START, which does not further specify the target of a terrorist attack, HCTB and RDWTI both include only attacks on civilians.

The empirical scope of the databases on genocides, politicides, and incidents of mass violence varies greatly. Some, such as the CSP list and Valentino's list of incidences of mass violence, have a global focus, whereas the GSP dataset focuses on single cases. The population of cases of mass violence, on the one hand, and of genocides and politicides, on the other, overlap but do not match. By introducing a fatality threshold, Valentino not only excludes some genocides and politicides as coded by Harff but also includes cases that do not meet Harff's criterion of the intention to destroy a specific group. In contrast, the CSP's coup d'état event dataset as well as Powell and Thyne both include all successful and failed coup attempts by national elites.

Conflict and Security Data

The choice of a specific conceptual approach has a large influence on the structure and content of the dataset (see Table 1). For example, COW's 'War' dataset (v.4.0) is not geographically restricted. The dataset differentiates nine different types of wars, which are subsumed under the four categories of inter-, extra-, intra-, and nonstate wars. Some conflict types are also coded regarding conflict parties, but neither conflict-years nor conflict intensities or measures are coded individually. COW's MID dataset (v.3.1) aggregates data on the level of single disputes or on the level of the 5602 dispute participants. In addition, its MIDA and MIDB datasets include also information on specific conflict measures and distinguish seven levels of conflict intensity. The MIDI and MIDIP datasets break down disputes into 512 dyads and contain information on 2119 conflict incidents.

The ACD (v.4-2012) by UCDP has a global scope. The conflict-years and conflicts are differentiated into minor armed conflicts (about two-thirds of all conflict-years) and wars. Furthermore, the data distinguish between different conflict dyads and dyad-years. While the ACD holds information on the type of incompatibility, the other two UCDP datasets – the 'Non-State Conflict Dataset' and the 'One-Sided Violence Dataset' – lack such information.

In contrast to COW and UCDP, the CSP's MEPV dataset (v.2010) includes information on societal effects of warfare, such as population dislocation, infrastructural and environmental damage, the effects on societal networks, and diminished quality of life, and thus goes beyond casualty figures in assessing conflict severity. These data are presented as an aggregated value on a 10-point scale representing the magnitude of an MEPV in an affected country in a given year. Furthermore, the dataset differentiates seven types of conflicts by combining the criteria of conflict constellation (civil-intrastate, ethnic-intrastate, or international) and the aim toward which violence is directed (violence, war, or independence).

The CONIAS country-year dataset (newest version: 2013) comprises data on inter-, intra-, and subnational conflicts. It differentiates two nonviolent and three violent levels of conflict intensity on a yearly basis and includes information on conflict items and types of actors involved. A disaggregated dataset encompasses spatially (i.e., subnational entities) and temporally (month-by-month) disaggregated information on the means and consequences of violent conflict measures for Asia

Table 1 List of conflict and security statistics

Database	*No. of observations*	*Observation period*
COW	654 Conflicts	1816–2007
MIDA	2332 Disputes	1816–2001
MIDB	5602 Disputes at participant level	1816–2001
MIDI/MIDIP	512 Dispute dyads	1993–2001
MIDI	2119 Conflict incidents	1993–2001
MIDIP	4805 Incidents at participant level	1993–2001
ACD	248 Conflicts (2061 conflict-years)	1946–2011
ACD	524 Dyads (2601 dyad-years)	1946–2011
Nonstate conflict dataset	419 Nonstate conflict dyads (646 dyad-years)	1989–2011
One-sided violence	713 Actor-years	1989–2011
MEPV	8244 Episode-years	1946–2008
MEPV	326 Episodes	1946–2012
MAR	1670 Antagonism-years (including conflicts, protests, rebellions, and repressions)	2004–06
CONIAS	985 Conflicts (15 400 conflict-years)	1945–2008
GED	24 381 Events	1989–2010
ACLED	>75 000 Events	1997–2012
GTD	1 04 000 Terrorist events	1970–2011
HCTB	852 Terrorist Events	1991–2012
RDWTI	36 000 Terrorism events	1972–2009
CIRI	6030 Country-years	1982–2010
PTS	6800 Country-years	1976–2011
Mass killing	30 Incidents of mass killings in civil war	1945–2000
CSP coup d'état events	807 Coups (Including attempts, plots, and allegations)	1946–2011
Powell and Thyne (2011)	457 Coups (Including attempts)	1950–2010

and Oceania (2000–09) and globally (2011–12). In addition, event datasets such as UCDP's GED and the ACLED by PRIO code information on single conflict events for limited periods of time and individual countries or regions.

START's GTD contains information on various forms of terrorist attacks (i.e., bombings, assassinations, and kidnappings), whereas HCTB and RDWTI list terrorist events and incidents of terrorism. The ITERATE project provides georeferenced data on transnational terrorist groups, their activities, and the environment in which they operate.

CIRI and PTS both include information on human rights violations in more than 6000 country-years, whereas the dataset on mass violence by Valentino (2004) includes all incidents of mass killings in the twentieth century. The State Failure Problem Set of CSP includes a list of episodes of genocides and politicides and provides an estimation of the annual number of deaths on an 11-point scale. In contrast, the Biographic Database by the Cambodian Genocide Program of GSP contains several thousand biographical records on members of the Khmer Rouge movement and its victims as well as georeferenced data on prisons, mass graves, and memorials. Comparable datasets exist for the genocides in Rwanda, East Timor, Guatemala, and Former Yugoslavia, although the quality and amount of data varies considerably. Finally, the coup d'état events data by CSP includes data on coup leaders and outcomes for 804 successful, attempted, plotted, and alleged coups in the world, whereas the data provided by Powell and Thyne differentiate only between successful and failed coup attempts but exclude plotted or alleged coup d'états.

Contemporary Trends

There are three discernible trends in contemporary conflict and security statistics: an increased focus on individuals and societal groups (individualization), a broader understanding of the conflicts studied (extension), and better spatial and temporal resolution of conflicts (disaggregation).

Individualization

Since the early 1990s, researchers have noted that conflicts within states, and often without direct state participation, are the predominant form of violent disputes in the contemporary world. This has rekindled the interest in the social structures 'beneath' the visible surface of states. This development has been accompanied by a conceptual shift away from a traditional understanding of (military and state) security to a perspective focusing on the security of individuals and groups of individuals ('human security'). As expressed, e.g., in the 1994 Human Development Report of the United Nations Development Programme, the extensiveness of the concept of 'human security' – including 'freedom from fear' and 'freedom from want' – is specified by means of seven dimensions, namely, economic, food, health, personal-physical, cultural, and political security.

Apart from the political, legal, and academic debates that ensued in the wake of this paradigmatic shift (with keywords like 'responsibility to protect' and 'humanitarian intervention') and notwithstanding criticism of the concept of human security as a possible case of conceptual stretching, the new framework induced a quantitative approach to security in the form of the Human Development Index (HDI). While the HDI includes statistics of life expectancy, duration of education and a country's gross national income per capita to provide a ranking of currently 188 countries and territories, the Happy Planet Index (HPI), which was launched by the London-based New Economics Foundation, combines data on life expectancy with life satisfaction and the ecological footprint produced by attaining a certain level of well-being. However, neither the HDI nor the HPI include data on violent conflicts or political repression. However, this is done by the Human Security Index, which includes the Global Peace Index (GPI) and PTS.

Extension

The apparent surge of intrastate conflicts since the end of the Cold War also contributed to a better recognition of the role of nonstate actors in conflicts, especially in conflicts that are less intense than wars ('low-intensity conflicts'). The category of nonstate actors primarily includes armed groups, and also political parties, civil society groups, and the private sector. Armed groups involved in political conflicts may include rebel militias, terrorist groups, paramilitary forces, and private security companies. Furthermore, with the (ostensible) transformation of conflicts and the trend from classical warfare toward 'new wars' and terrorist strategies in the 1990s, many researchers have recognized that wars (and especially civil wars) often do not end with peace treaties or apparent victories. Rather, many conflicts resurface, sometimes repeatedly. In order to map such protracted and oscillating conflict dynamics, it is necessary to cover interwar phases of lesser conflict intensity, as well. In addition, conflicts of high intensity do not appear suddenly but typically undergo antecedent low-level violent and nonviolent stages. Moreover, different types of conflict often occur together with, or consecutive to, other conflict types, e.g., civil war and genocide, repression and revolutions, or subnational conflicts and terrorism.

Research institutions reflected upon these developments by devising more dynamic conflict concepts and extending their empirical scope. Within the framework of COW, the MID dataset was compiled in the beginning of the last decade, which also takes into consideration threats of the use of violence in international conflicts. In 2004, UCDP introduced the additional threshold of 25 battle-related fatalities to define 'minor armed conflicts' in contrast to wars. UCDP's GED dataset further lowered the threshold to one fatality. From its inception, the HIIK covered low-level violent and nonviolent conflicts and conflict phases, employing qualitative assessments to classify conflicts on a four-level scale. Since 2003, the HIIK and later CONIAS have used a five-level scale differentiating between entirely nonviolent disputes, conflicts that make use of threats of violence, low-level violent conflicts, limited wars, and full-fledged wars.

Disaggregation

In contrast to traditional conflict and security statistics, which focused exclusively on the national and international levels of conflicts on a year-by-year base, another recent trend in the research field has been to open the 'black box' of the nation-state by localizing conflicts within the geographical space of

a country. Research on geographical causes of conflict – especially the location of natural resources in a territory – demonstrated the need for a disaggregated form of geographical referencing. Furthermore, due to the increased emphasis on conflict dynamics, the trend has been to disaggregate conflict processes over time into basic units of 'events,' like individual battles or attacks. This development thus continues the conflict event research as represented by, e.g., the Kansas Event Data System, whose roots reach back to the 1970s (Schrodt and Hall, 2006), and Arthur S. Bank's Cross-National Time-Series Data Archive, launched in 1968.

UCDP's GED and PRIO's ACLED datasets combine both trends; here, events form the basis of georeferenced coding. In addition, GED construes geometrically defined conflict areas from the location of the coded events. While GED offers georeferenced data for the African continent for the period from 1989 to 2010, ACLED provides such data for Africa and selected Asian countries (1997–2012). The Event Data on Armed Conflict and Security Project in Berlin has also collected spatially and temporally disaggregated data concerning the use of violence in civil wars in four African countries in the period 1990–2009. In contrast to event data compilation, CONIAS has undertaken a georeferenced coding of intensities since 2011, i.e., of the severity of a given conflict in an administrative unit in a given calendar month, for the region of Asia and Oceania (2000–09). The intensity assessment is based on the violent and nonviolent actions of the conflict participants and is therefore conceptually related to the event-based approach.

In addition to these efforts, the spatial coding of factors relevant to political conflicts was aspired in recent years. In 2005 and 2007, respectively, PRIO published georeferenced information on the location of diamond resources and petroleum fields. In 2010, the International Conflict Research (ICR) group at the ETH Zürich released a dataset on the georeferencing of ethnic groups, containing a total of 929 ethnic groups. In the following year, the ICR published a georeferenced coding of settlement patterns of selected ethnic groups, the Geo-referencing Ethnic Power Relations dataset.

The Role of Conflict and Security Statistics in Research

Conflict and security statistics are primarily employed in research on the causes, dynamics, and structures of political conflicts. In addition, these statistics are relevant for the measurement of political regimes and political fragility as well as for early warning of political conflict and benchmarking of development aid policies.

The most widely used dataset for analyzing the patterns of intranational conflicts are published by UCDP and PRIO. Of central importance are also the datasets by COW, MAR, PITF, as well as by Fearon and Laitin. While some articles refer to the original data, many add self-compiled data to suit their specific explanatory focus. This may be a hint to deficits regarding the main databases on political conflicts as they apparently do not always satisfy the needs of researchers. However, the existence of a multitude of datasets enables robustness tests. In fact, some research articles employ more than one dataset in statistically assessing their hypotheses.

Conflict data are also relevant in regime measurements. For example, indictors of CSP's Polity IV dataset are inherently related to conflict. If conflict leads to a *de facto* separation of countries, this is coded as Polity Fragmentation. Periods where external intervention leads to the occupation of a country or where internal or international conflict triggers the collapse of a political regime are coded as Interruption or Interregnum Periods. Furthermore, internal unrest may lead to the coding of factional competition.

Evaluating political regimes is also associated with the measurement of state fragility. The UCDP dataset has been used by the Country Indicators for Foreign Policy Fragility Index, the GPI, the Harvard Kennedy School Index of African Governance, and the Peace and Conflict Instability Ledger (PCIL). The latter also includes data from the COW dataset. PITF data has been used by PCIL and the Index of State Weakness in the Developing World. CSP's MEPV dataset has been used by CSP's State Fragility Index.

In addition, conflict and security statistics are employed in Early Warning Systems (EWS) that aim at predicting the probability of the occurrence or escalation of violent conflict. Most EWS at least partly rely on or provide quantitative data that are linked to conflict and security statistics. However, as the occurrence of political conflict is the explanandum of EWS, conflict statistics are mostly not employed as explanans due to endogeneity problems, but are rather employed to optimize and test predictive models or as predictors for the escalation of ongoing conflicts of low intensity. In a similar fashion as in EWS, conflict statistics can also be employed in the evaluation of developmental aid. Geographical information on foreign aid projects and violent conflict allow assessing the impact of the former on the latter.

Challenges

Contemporary conflict and security statistics face challenges (1) in the geographical, temporal, conceptual, and material extension of data disaggregation; (2) the collaboration of research institutions; and (3) regarding availability and the reliability of conflict data.

While the disaggregation of conflict and security statistics has made considerable progress in recent years, the coverage of available georeferenced data on conflict events is largely limited to Africa and parts of Asia and the post-Cold War period. Moreover, efforts to disaggregate conflict events mostly focus on conflict measures, whereas conflict actors and conflict items have attracted far less attention. In addition, data on conflicts and conflict periods of no or very low violence, which would be necessary to develop a more adequate understanding of the dynamics of the emergence of conflicts as well as their termination, are mostly lacking. So far, only MID (for international conflicts), MAR (for protests and repression), and CONIAS (for all conflict types) include nonviolent conflicts. A further challenge lies in finding alternatives to the predominant measurement of conflict intensities in terms of battle-related deaths. A comprehensive assessment of the instruments and consequences of violent conflict would have to encompass the use and type of weaponry, the numbers and sociodemographics of persons involved, coverage of fatalities

from indirect consequences, data on displacement, the destruction of the environment or public infrastructure, and the availability of basic goods. MAR and CONIAS have taken steps in this direction. The disaggregation of conflict data may enable researchers to assess conflict dynamics more precisely than ever before, but analyses have to rely on data of explanatory variables that mostly are not yet available on the same level of precision.

Another challenge is the limited connectivity between different conflict datasets. This has resulted in a fragmented picture of the global conflict panorama and to a 'binning' of collected information in separate datasets. So far, only COW and UCDP have made considerable efforts to link their different datasets to each other internally. Externally, the challenge is partly addressed by attempts to interlink data banks of different institutions via shared identification numbers. However, integrating the wealth of information on conflicts into a comprehensive data bank still remains a largely unaccomplished goal. Most progress in this regard has been attained by ICR in Zürich, which hosts the GROW^{up} Research Front-End, integrating data on intranational conflict and ethnic groups. One source of the project is the ACD2EPR Docking Version 1.2 dataset, which links ACD v.4-2010 to the ETH Ethnic Power Relations (EPR-ETH) dataset v.2.0.

Finally, the availability and the reliability of information also pose challenges for the collection and procession of conflict and security data. The more disaggregated the measuring of conflicts is – geographically, temporally, and (as in the case of conflict intensity) conceptually – the more demanding the required information is. However, the quality and availability of information required is especially challenging in those regions affected by (violent) conflict. In addition, the growth of information technology, which has made information more easily available, has created new challenges of data processing. In this regard, a possible solution may be automated data gathering. One example is the Penn State Event Data Project, which covers political events in the Middle East, Balkans, and West Africa. Using information from the public domain such as international and local media reports may create reliability problems as it can be biased, e.g., by attention gaps, state censorship, or attempts of nonstate actors to influence media coverage for their own cause. This challenge can only be overcome by systematically evaluating the sources employed and by training the data collectors in country-specific knowledge. In contrast, attempts by conflict researchers to compensate these challenges by employing personnel in the field should be regarded with caution. Access to events on the ground is frequently restricted, while any direct observation might exacerbate the researcher's problem of objectivity.

See also: Data Bases and Statistical Systems: Crime Measurement; Data Bases and Statistical Systems: Crime and Delinquency; Data Bases and Statistical Systems: Crime; Data Bases and Statistical Systems: International Comparative Research; Data Bases and Statistical Systems: Overview; Data Bases and Statistical Systems: Political Science (General); Genocide: Social and Economic Aspects; Quantitative Cross-national Research Methods; Spatial Dynamics and Crime.

Bibliography

Cederman, L.-E., Wimmer, A., Min, B., 2010. Why do ethnic groups rebel? New data and analysis. World Politics 62 (1), 87–119. The EPR v.1.1 and EPR-ETH v.2.0 datasets are available online at: http://www.icr.ethz.ch/data/growup/epr-eth.

Center for International Development and Conflict Management, 2010. Primary Data Collections (version 10, released July 2010). The ICB data collections are available online at: http://www.cidcm.umd.edu/icb/data/.

Center for International Development and Conflict Management, 2012. Peace and Conflict Instability Ledger. An overview over the PCIL data can be found online at: http://www.cidcm.umd.edu/pc/executive_summary/exec_sum_2012.pdf.

Center for Systemic Peace, 2012. High Casualty Terrorist Bombing (HCTB) Events. The HCTB dataset is available online at: http://www.systemicpeace.org/inscr/inscr.htm.

Chojnacki, S., Ickler, C., Schoenes, K., Spies, M., Wildemann, T., 2012. EDACS Codebook Version 3.7: The Event Data on Armed Conflict and Security (EDACS) Project. Free University Berlin, Berlin, Germany. The EDACS dataset is available online at: http://www.conflict-data.org/edacs/downloads/index.html.

Cingranelli, D.L., Richards, D.L., 2013. The Cingranelli-Richards (CIRI) Human Rights Dataset. The CIRI version 13.03.2013 is available online at: http://www.humanrightsdata.org/.

Country Indicators for Foreign Policy, 2012. Country Indicators for Foreign Policy – Fragility Index. The CIFP-FI data are available online at: http://www4.carleton.ca/cifp/app/ffs_data_methodology.php.

Croicu, M.C., Sundberg, R., 2012. UCDP GED Conflict Polygons Dataset Codebook Version 1.1-2011. Uppsala University, Uppsala, Sweden. The UCDP GED Point Dataset is available online at: http://www.ucdp.uu.se/ged/data.php.

Eck, K., Hultman, L., 2007. Violence against civilians in war. Journal of Peace Research 44 (2). The UCDP One-sided Violence Dataset v. 1.4-2012, 1989–2011 is available online at: http://www.pcr.uu.se/research/ucdp/datasets/ucdp_one-sided_violence_dataset/.

Fearon, J.D., Laitin, D.D., 2003. Ethnicity, insurgency, and civil war. American Political Science Review 97 (1), 75–90.

Ghosn, F., Palmer, G., Bremer, S., 2004. The MID3 data set, 1993–2001: procedures, coding rules, and description. Conflict Management and Peace Science 21, 133–154. The MID v.3.10 dataset and the associated MIDA, MIDB, MIDI, MIDIP datasets are available online at: http://www.correlatesofwar.org/COW2%20Data/MIDs/MID310.html.

Gibney, M., Cornett, L., Wood, R., 2013. Political Terror Scale 1976–2008. The PTS dataset is available online at: http://www.politicalterrorscale.org/.

Gleditsch, N.P., Wallensteen, P., Eriksson, M., Sollenberg, M., Strand, H., 2002. Armed conflict 1946–2001: a new dataset. Journal of Peace Research 39 (5), 615–637. The UCDP/PRIO Armed Conflict Dataset v.4-2012, 1946–2011 is available online at: http://www.pcr.uu.se/research/ucdp/datasets/ucdp_prio_armed_conflict_dataset/.

Gurr, T.R., 1980. Handbook of Political Conflict: Theory and Research. The Free Press, New York, NY.

Harff, B., 2003. No lessons learned from the Holocaust? Assessing risks of genocide and political mass murder since 1955. American Political Science Review 97 (1), 57–73.

Harff, B., Gurr, T.R., 1988. Toward empirical theory of genocides and politicides: identification and measurement of cases since 1945. International Studies Quarterly 32 (3), 359–371.

Harvard University, Kennedy School of Government, 2009. Index of African Governance. The IGA data are available online at: http://www.nber.org/data/iag.html.

Hastings, D.A., 2012. The Human Security Index: An Update and a New Release. Available online at: http://www.humansecurityindex.org/wordpress/wp-content/uploads/2012/02/hsiv2-documentation-report1_1.pdf.

Holmqvist, C., 2005. Engaging armed non-state actors in post-conflict settings. In: Bryden, A., Hänggi, H. (Eds.), Security Governance in Post-conflict Peacebuilding. Lit, Münster, Germany.

International Institute for Strategic Studies, 2013. Armed Conflict Database. The ACD dataset is available online at: http://www.iiss.org/publications/armed-conflict-database/.

LaFree, G., Dugan, L., 2007. Global Terrorism Database, 1970–1997. ICPSR04586-v1, Ann Arbor, MI. The GTD I dataset is available online at: http://www.icpsr.umich.edu/icpsrweb/NACJD/studies/04586/detail.

LaFree, G., Dugan, L., 2008. Global Terrorism Database II, 1998–2004. ICPSR22600-v1, Ann Arbor, MI. The GTD II dataset is available online at: http://www.icpsr.umich.edu/icpsrweb/NACJD/studies/22600/detail.

Marshall, M.G., 2002. Measuring the societal impact of war. In: Hampson, F.O., Malone, D.M. (Eds.), From Reaction to Prevention. Lynne Rienner, Boulder, CO.

Marshall, M.G., 2010. Major Episodes of Political Violence (MEPV) and Conflict Regions, 1946–2008. The MEPV dataset is available online at: http://www.systemicpeace.org/inscr/inscr.htm.

Marshall, M.G., Gurr, T.R., Harff, B., 2012. PITF – State Failure Problem Set. The PITF dataset is available online at: http://www.systemicpeace.org/inscr/inscr.htm.

Marshall, M., Marshall, D., 2007. Coup d'État Events, 1960–2006. The 1946–2011 Coup d'État Events dataset is available online at: http://www.systemicpeace.org/inscr/inscr.htm.

Mata, J.F., Ziaja, S., 2009. User's Guide on Measuring Fragility. Retrieved online from: http://www.die-gdi.de/CMS-Homepage/openwebcms3.nsf/%28ynDK_contentByKey%29/ANES-7W89TW/$FILE/UNDP-DIE%202009%20Users%20Guide%20on%20Measuring%20Fragility.pdf (27.3.2013).

Mickolus, E.F., et al., 2007. International Terrorism: Attributes of Terrorist Events. The ITERATE datasets are available online at: http://library.duke.edu/data/collections/iterate.html.

Minorities at Risk Project, 2009. Minorities at Risk Dataset. The MAR datasets are available online at: http://www.cidcm.umd.edu/mar/.

National Consortium for the Study of Terrorism and Responses to Terrorism, 2012. Global Terrorism Database. The GTD dataset is available online at: http://www.start.umd.edu/gtd.

New Economics Foundation, 2013. Happy Planet Index. The HPI data are available online at: http://www.happyplanetindex.org/data/.

Penn State Event Data Project, Pennsylvania State University, 2010. Penn State Event Data Project. An overview over the PSEDP data can be found online at: http://eventdata.psu.edu/index.html.

Pfetsch, F.R. (Ed.), 1991. Konflikte seit 1945: Daten – Fakten – Hintergründe. Ploetz, Freiburg i. Br., Germany.

Powell, J.M., Thyne, C.L., 2011. Global instances of coups from 1950 to 2010: a new dataset. Journal of Peace Research 48 (2), 249–259.

Raleigh, C., Hegre, H., March 7–8, 2005. Introducing ACLED: An Armed Conflict Location and Event Dataset. Paper Presented to the Conference on 'Disaggregating the Study of Civil War and Transnational Violence'. University of California Institute of Global Conflict and Cooperation, San Diego, CA. The ACLED Dataset (1997–2012) is available online at: http://www.acleddata.com/data/.

RAND Corporation, 2009. RAND Database of Worldwide Terrorism Incidents. The RDWTI dataset is available online at: http://www.rand.org/nsrd/projects/terrorism-incidents.html.

Richardson, L.F., 1960. Statistics of Deadly Quarrels. Boxwood Press, Pittsburgh, PA; Chicago, IL.

Sarkees, M.R., Wayman, F., 2010. Resort to War: 1816–2007. CQ, Thousand Oaks, CA. The COW Wars v.4.0 datasets are available online at: http://www.correlatesofwar.org/COW2%20Data/WarData_NEW/WarList_NEW.html.

Schreiber, W., 2012. Kriege und bewaffnete Konflikte 2012. Ein erster Überblick. AKUF Analysen 11. Retrieved online from: http://www.wiso.uni-hamburg.de/fileadmin/sowi/akuf/Text_2010/AKUF-Analysen-11.pdf. A list of ongoing wars is available online at: http://www.wiso.uni-hamburg.de/fachbereiche/sozialwissenschaften/forschung/akuf/laufende-kriege.

Schrodt, P.A., Hall, B., 2006. Twenty Years of the Kansas Event Data System Project. http://web.ku.edu/keds/KEDS.history.html.

Schwank, N., 2012. Konflikte, Krisen, Kriege: Die Entwicklungsdynamiken Politischer Konflikte Seit 1945. Nomos, Baden-Baden, Germany. The CONIAS country-year dataset is available online at: www.conias.org.

Schwank, N., Trinn, C., Wencker, T., 2013. Der heidelberger ansatz der konfliktdatenerfassung. Zeitschrift für Friedens- und Konfliktforschung 2 (1), 32–63.

Singer, D., Small, M., 1972. The Wages of War, 1816–1965: A Statistical Handbook. Wiley, New York, NY.

Sundberg, R., Eck, K., Kreutz, J., 2012. Introducing the UCDP non-state conflict dataset. Journal of Peace Research 49, 351–362. The UCDP Non-State Conflict Dataset v. 2.4-2012, 1989–2011 is available online at: http://www.pcr.uu.se/research/ucdp/datasets/ucdp_non-state_conflict_dataset_/.

Sundberg, R., Lindgren, M., Padskocimaite, A., 2010. UCDP GED Codebook Version 1.0-2011. Uppsala University, Uppsala, Sweden. The UCDP GED Conflict Polygons Dataset is available online at: http://www.ucdp.uu.se/ged/data.php.

United Nations Development Programme, 1994. Human Development Report. New York.

Uppsala Conflict Data Program, 2012. UCDP Actor Dataset 2.1-2012. Uppsala University, Uppsala, Sweden. The UCDP Actor Dataset v.2.1-2012 is available online at: http://www.pcr.uu.se/research/ucdp/datasets/ucdp_actor_dataset/.

Valentino, B., Huth, P., Balch-Lindsay, D., 2004. "Draining the sea": mass killing and Guerrilla Warfare. International Organization 58 (2), 375–407.

Valentino, B., 2000. Final Solutions: Mass Killing and Genocide in the Twentieth Century. Ithaca, New York.

Vision of Humanity, 2013. Global Peace Index. The GPI data are available online at: http://www.visionofhumanity.org/gpi-data/.

Weidmann, N.B., Rød, J.K., Cederman, L.-E., 2010. Representing ethnic groups in space: a new dataset. Journal of Peace Research 47 (4), 491–499. The GREG dataset is available online at: http://www.icr.ethz.ch/data/other/greg.

Wright, Q., 1942. A Study of War. University of Chicago, Chicago, IL.

Wucherpfennig, J., Weidmann, N.B., Cederman, L.-E., Girardin, L., Duhart, P., Brown, G., Flora, J., Wimmer, A., 2010. GeoEPR Dataset. The GeoEPR v.1.0 dataset is available online at: http://dvn.iq.harvard.edu/dvn/dv/epr/faces/study/StudyPage.xhtml?globalId=hdl:1902.1/14206&studyListingIndex=1_e1e4e12f7976a292445bd043f478.

Data Bases and Statistical Systems: Sociology

Jürgen HP Hoffmeyer-Zlotnik, University of Giessen, Germany
Uwe Warner, Centre d'Etudes de Populations, de Pauvreté et de Politiques Socio-Economiques, Luxembourg

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R.T.Campbell, volume 5, pp. 3255–3260, © 2001, Elsevier Ltd.

Abstract

As a rule, core databases contain quantitative survey data. Traditionally, qualitative data, for example, ethnographic field notes or transcripts of group discussions and interviews, have not been made available for secondary analysis. This is slowly changing, and a number of data archives in Europe, Australia, and the US are working on the archiving of qualitative data. Quantitative social science datasets are usually generated by major surveys, such as general social surveys, household panel studies, thematic surveys, or barometer studies, and made available for analysis.

Methodological Characteristics of Quantitative Surveys

Sampling Techniques

Quantitative surveys are characterized by the fact that they employ a standardized questionnaire that is administered to a group of individuals (subpopulation) selected from the total population being investigated (target population, survey population). Respondents are selected from the target population using a random statistical sampling procedure (Kish, 1994). It must be ensured that each individual in the target population has a chance (greater than zero) of being selected.

Surveys are used to measure demographic, socioeconomic, attitudinal, and behavioral characteristics. With the help of these data, conclusions can be drawn about the attitudes, opinions, and behavior of all the social groups in the sample – and, therefore, in the target population – under the conditions specified in the survey questions. Surveys help researchers to observe society. Because major surveys are repeated at regular intervals, and despite the fact that very few of them have access to the same respondents each time, they provide not only a thematic snapshot but also an insight into attitudinal and behavioral changes in society over time. This is due to the fact that random sampling techniques enable each sample to be drawn under the same conditions and from the same target population so that the subpopulation displays the same structural characteristics each time. The aim of a standardized survey with a random sample is to construct quantitative descriptors of the characteristics of the target population from which the sample is drawn (cf Groves et al., 2004: p. 2).

Methodologies

Once the sample has been drawn, the design of the questions and the questionnaire can begin. Here, the methodological focus lies on choosing a survey method (face-to-face, telephone, mail, etc.), formulating the survey questions, and pretesting the questionnaire. Special methods are used when formulating and structuring questions and response categories in order to reduce response errors (Krosnick and Presser, 2010: p. 263).

First, researchers must familiarize themselves with the cognitive processes underlying responses to survey questions (Schwarz, 2008: p. 374). Bearing this knowledge in mind, they can then begin to formulate the questions and the response categories. Because respondents constitute a heterogeneous group that differs in terms of age cohort, social class, and regional and cultural origins, the wording of questions and response categories must be specific, concrete, and easy to understand (Converse and Presser, 1986: pp. 10, 31; Krosnick and Fabrigar, 2006).

The datasets presented here undergo further processing by the institution that makes them available for secondary analysis, which is the reanalysis of data collected in a survey. While this institution may be the research body on whose behalf the data were collected, more often than not it is an archive for social science datasets. In contrast to a team of researchers that processes its own data and makes them available to others, data archives are experienced in processing data, preparing user-friendly tools for secondary analysis, and ensuring secure, long-term preservation of data. Moreover, major data archives have an international network.

Types of Core National Datasets

A look at the core national datasets generated by surveys and made available for secondary analysis reveals five types of surveys:

1. General social surveys (GSS) conducted at regular intervals. The data from these surveys are made available for research and teaching purposes.
2. Household panel surveys, normally conducted as follow-up surveys, whose main objective is to study the economic and social structure of private households and to monitor and measure changes over time.
3. Longitudinal education surveys (some of which are trend studies, others are panel studies), meaning follow-up surveys that focus on students' competencies and conditions for teaching and learning, as well as passages within the national education system and transitions from the education system to the labor market.
4. Thematic specialized social science surveys that can be regarded either as individual snapshots or as a series of snapshots of a specific theme.

5. And finally, cross-national barometer surveys conducted with more or less standardized questionnaires that furnish an insight into the attitudes, opinions, and behavior of the target populations in the participating countries.

General Social Surveys

GSS are academically driven national surveys carried out at regular intervals. 'Regular' means that the next wave of the survey is conducted 1 or 2 years later. As a rule, the target population comprises all persons of 15, 16, 18 years of age and above, who are residents in private households. An upper age limit is not common. In most countries, the sample is a random one (also known as a probability sample). The questionnaire comprises a standard core of socio-demographic, behavioral, and attitudinal variables collected in each wave, and a topical module that varies from wave to wave. The replicating core enables social developments to be studied over time, thus facilitating time-trend studies. The topical module often focuses on several issues of topical interest. In each country, a renowned national social science institute or a university-affiliated research institute is usually given responsibility for conducting the GSS. However, as the examples from the Australian, New Zealand, and Canadian GSS show, academically driven surveys are not always conducted by an academic research institute or a university. In these three countries, responsibility for the major national surveys lies with the national statistical institute (NSI).

After data collection and processing, datasets are made available for general use. Social science institutes and universities, in particular, employ the data for teaching and research purposes. Because the datasets are often available online, the secondary analysis of the GSS data is not restricted to researchers in the country in question.

The most prominent GSS is the *US General Social Survey*, which has been carried out since 1972 by the National Opinion Research Center (NORC) at the University of Chicago. Although it was conducted annually at first, the US GSS has employed a biennial design since 1994. In 2006, it began the transition from a replicating cross-sectional design – that is, a new sample was drawn for each wave – to a design that uses rotating panels. This rotating panel design means that, in the ideal case, the GSS is administered to each respondent for three successive waves. After three waves, the respondent is replaced by a new recruit. Eventually, about one-third of the panel is replaced after each wave. The advantage of a rotating panel design is that it facilitates the modeling of attitudinal and behavioral changes across three waves. Three waves are the prerequisite for structural equation modeling. Longer participation in the panel is detrimental to the respondent's spontaneity when answering the questions.

What is interesting about the US GSS is that, from its inception, it has not only been a source of data, but has also contributed to survey methodology (Smith, 1979: p. 11). The fact that NORC has fulfilled its mandate in this regard is evidenced by the reports on various aspects of survey methodology to be found on the US GSS Web site (US GSS, 2013).

The idea for, and the concept of, the US GSS has been taken up in numerous countries, for example, Germany, Poland, Russia, Australia, New Zealand, and Canada. As a result, the national GSS is deemed to be the best source of data with which to observe national attitudinal and behavioral trends.

ALLBUS (ALLgemeine BevölkerungsUmfrage der Sozialwissenschaften), the *German General Social Survey* (ALLBUS, 2013), was implemented for the first time in 1980. In 1984, collaboration between the organizers of the GSS in four countries (Germany, the USA, the UK, and Australia) yielded the *International Social Survey Programme* (ISSP) (ISSP, 2013). Fielded for the first time in six countries in 1985 as a supplement to the regular GSS, the ISSP is now conducted annually in 48 countries (as of 2012) either as a supplement to the GSS or another national survey, or as a survey in its own right. To date, 11 topics have been covered by the ISSP, most of which are repeated at regular, but unequal, intervals.

The year 2003 saw the inception of the *Chinese General Social Survey* (China GSS, 2013), which is conducted annually or biennially with different focal themes. The China GSS is based on a stratified random sample representative of the urban and rural population (above 18 years of age). Datasets are not only freely accessible for research and teaching in China, but may also be used by researchers outside China. Hence, the survey provides a valuable opportunity to obtain information about social structure and trends in China.

Selected Household Panel Studies of Income and Living Conditions

Defining Panel Studies

Panel studies are longitudinal surveys that use the same units of observation each time. The advantage of longitudinal (panel) surveys is that they enable researchers to measure and analyze changes over time in socio-demographic and economic situations, as well as the attitudes, opinions, and behaviors of individuals or aggregates of individuals. The unit of observation of household-based panels is the household and its members. Household panels enable researchers to study household change and the changing dynamics of the individuals within the household (cf Stafford, 2010: pp. 765 ff.).

Examples of National Panel Studies

The US *Panel Study of Income Dynamics* is "the world's longest running nationally representative household panel survey" (PSID, 2013). It started in 1968 with a panel of 18 000 individuals living in 5000 families in the US. The panel has since grown to 22 000 individuals living in over 9000 families. Interviews were conducted annually until 1997. Since the 1999 wave a biennial design has been used. Over the years, the panel has been expanded to reflect post-1968 immigration.

Established in 1993, the *Survey of Labour and Income Dynamics* (SLID) is a longitudinal annual household survey conducted by Statistics Canada. SLID comprises two panels, each of which includes some 34 000 individuals living in 17 000 households. "A panel is surveyed for a period of six consecutive years. A new panel is introduced every 3 years, so two panels always overlap" (SLID, 2013). The target population is all individuals resident in private households in Canada with the exception of residents of the Yukon, the Northwest Territories and Nunavut, and persons living on Indian reserves.

The *British Household Panel Survey* (BHPS) started in 1991 with a panel of some 9092 individuals living in 5050 households throughout Great Britain. The panel was topped up for the ninth wave in 1999 by adding two samples of 1500 households each from Scotland and Wales. For the eleventh wave in 2001, 1900 households from the Northern Ireland Household Panel survey were integrated into the BHPS, thereby making it suitable for UK-wide research (BHPS, 2013). The BHPS was conducted annually until 2008. In 2009, it was incorporated into *Understanding Society*, the UK Household Longitudinal Study (Understanding Society, 2013). Understanding Society covers around 40 000 households in England, Scotland, and Wales and 2400 addresses in Northern Ireland. Interviews are repeated annually; the data are available from the Economic and Social Data Service (ESDS, 2013).

The *DNB Household Survey*, formerly known as the CentER Savings Survey, is a Dutch panel survey that studies the psychological and economic aspects of financial behavior (DHS, 2013).

The *Longitudinal Internet Studies for the Social Sciences* (LISS) panel comprises some 8000 individuals living in 5000 households in the Netherlands. LISS began in 2007; it was supplemented with a separate immigrant panel of 2400 individuals living in 1600 households in 2010 (LISS, 2013).

The *Longitudinal Individual Database* (LINDA) is a register-based Swedish household panel that was established in 1960. A nonoverlapping immigrant sample with the same design and covering the same time period as the main sample is attached to LINDA. LINDA stores annual information of about 300 000 individuals. "The core of the data is the annual income registers (Inkomst- och taxeringsstatistiken) and population census data (Folk- och Bostadsräkningen) available every fifth year from 1960 to 1990" (LINDA, 2013). All the variables in these registers are included in the LINDA database. Researchers can access the data through Statistics Sweden.

First implemented in 1984, the *Socio-Economic Panel Study* (SOEP) is a panel survey of private households in Germany. Some 20 000 individuals living in 11 000 households are surveyed annually (SOEP, 2013). The first wave covered 12 245 individuals living in 5921 households. In 1990, after German reunification, 2179 households with 4453 members from the former German Democratic Republic were added to the panel. In 1994, the panel was topped up once again by adding a sample of 1078 immigrants, or persons with a migrant background, living in 522 households. Supplementary samples added in 1998, 2000, 2002, and 2006 ensure that SOEP is representative of the German population as a whole and facilitate more detailed analyses of specific subgroups.

The *Swiss Household Panel* (SHP) is an annual panel study established in 1999 to "observe social change, in particular the dynamics of changing living conditions and representations in the population of Switzerland" (SHP, 2013). The SHP started with a sample of 12 931 individuals living in 5074 households. In 2004 a second sample of 2538 households with a total of 6569 household members was added. A biographical questionnaire aimed at collecting retrospective data on professional development and family life was administered in 2002.

The *Panel Socio-Economique "Liewen zu Letzeburg"* (PSELL) began in 1985 with 6110 individuals living in 2012 households in Luxembourg. In 1995 the original panel was replaced by a new household sample, PSELL2. PSELL3, which started in 2003 with 3500 households, is the Luxembourg component of the European Survey on Income and Living Conditions (EU-SILC).

The *Russia Longitudinal Monitoring Survey* (RLMS) is a household-based panel study launched in 1992. To date, 20 rounds have been conducted. Individual household members who moved away from the address are not followed. In each annual round, the RLMS interview is completed with the household and its members at the original sample dwelling unit (RLMS, 2013). Consequently, the RLMS is not a true household or person panel, but a panel of dwelling units.

The following is a list of further national household panel studies:

- The *Panel Study of Belgian Households* started in 1992. In 2003, it was transformed into the Belgian component of the EU-SILC.
- The *Household, Income and Labour Dynamics in Australia* (HILDA) survey started in 2001 with a panel of 19 914 individuals in 7682 households. In the eleventh wave, a sample of 5477 individuals in 2153 households was added to the panel (HILDA, 2013).
- The *Survey of Family, Income, and Employment* (SoFIE) is a New Zealand panel survey conducted annually between 2002 and 2010.
- The *Mexican Family Life Survey* was first implemented in 2002. The second wave took place in 2005–06 (MxFLS, 2013).
- The *Korean Labor & Income Panel Study* (KLIPS) is an annual survey of households and individuals residing in urban areas. Conducted by the Korea Labor Institute, the first wave was launched in 1998 (KLIPS, 2013).
- The *Hungarian Household Panel Study* (HHP) was a sample of 2600 households surveyed annually between 1991 and 1997.
- The *Bosnia and Herzegovina Household Survey Panel Series – Living in BiH* started in 2001 with over 3000 households. The last wave was implemented in 2004 (LiBiH, 2013).

European-Wide Panel Studies

In 1994, the European Union launched its first panel survey, the European Community Household Panel (ECHP). All member states were obliged to participate.

The ECHP survey was based on a standardized and input-harmonized questionnaire that involved annual interviewing of a representative panel of households and individuals in each member state. In the first wave, carried out in 1994, a sample of some 60 500 nationally represented households with approximately 130 000 adults of 16 years of age and above were interviewed in the 12 member states. Austria and Finland joined the project in 1995 and 1996, respectively. From 1997 onwards, new member Sweden provided data derived from the Swedish Living Conditions Survey. The total duration of the ECHP was 8 years, running from 1994 to 2001 (ECHP, 2013a). Access to the anonymized microdata is provided only by means of research contracts and is restricted in principle to universities, research institutes, NSIs, central banks inside the EU and EEA countries, as well as to the European Central Bank (ECHP,

2013b). The ECHP was discontinued in 2001 and was succeeded by the EU-SILC survey.

The EU-SILC survey was formally launched in 2004 and expanded in 2005 to cover all the 25 member states of the EU. Since 2007, a total of 31 countries – the EU-27 member states together with Norway, Iceland, Switzerland, and Turkey – participate in the survey (EU-SILC, 2013a). Access to EU-SILC microdata is subject to the same restrictions as access to ECHP data (EU-SILC, 2013b).

Education Surveys

International Education Surveys

The most well-known education survey is the OECD *Programme for International Student Assessment* (PISA), which assesses 15-year-old students' competencies in reading, mathematics, and science. The first survey was conducted in 2000 in all OECD member countries and a number of partner countries. Subsequent PISA surveys have followed at 3-year intervals (PISA, 2013). Although PISA is a cross-national comparative survey, it focuses on the weaknesses of students in the individual member countries. By contrast, the focus of the OECD's *Teaching and Learning International Survey* is on the analysis of teaching and learning environments in schools (TALIS, 2013). These education surveys are supplemented by the OECD *Programme for the International Assessment of Adult Competencies* (PIAAC). The first PIAAC survey was conducted in 2011–12, and preliminary results are expected in 2013. PIAAC builds on the OECD's *Adult Literacy and Life Skills Survey*, which was conducted in two main waves between 2002 and 2008, and the *International Adult Literacy Survey* undertaken in three rounds between 1994 and 1998 (PIAAC, 2013).

Developed and conducted by the International Association for the Evaluation of Educational Achievement (IEA), the *Trends in International Mathematics and Science Study* is an international assessment of fourth and eighth grade students' achievements in mathematics and science. First implemented in 1995, it has been carried out every 4 years since then. In 2011, 69 countries participated in the study (TIMSS, 2013).

The *Progress in International Reading Literacy Study* (PIRLS) is also conducted by the IEA. PIRLS measures fourth grade students' reading literacy achievement. First conducted in 2001, it was repeated in 2006 and 2011. The IEA plans to continue testing literacy achievement every 5 years. The data from 2011 are available for 49 countries (PIRLS, 2013).

National Education Surveys

In addition to these international surveys, there are national education studies that provide descriptive data on the educational activities of the national population and conditions of teaching and learning in schools. These studies include, for example, the US *National Household Education Surveys Program*, which provides descriptive data on the educational activities of the US population (NHES, 2013). This study is supplemented by the *Teaching, Empowering, Leading and Learning* surveys that examine conditions of teaching in all the US federal states. In the UK, the *Adult Participation in Learning Surveys* have been measuring adult participation in learning since 1996. The *All India Educational Survey* (AIES), which was implemented for the first time in 1957 and has been conducted at (ir)regular intervals since then, collects information on the various aspects of school education and infrastructure that education planners and administrators require in order to improve conditions for teaching and learning. The eighth AIES was conducted in 2009 (AIES, 2013).

Educational panel studies, such as the *German National Educational Panel Study*, collect longitudinal data on the development of competencies, educational processes, educational decisions, and returns to education in formal, nonformal, and/or informal educational institutions. This facilitates the analysis of educational and training processes throughout the life span.

Transitions from education to employment surveys, for example the *Transitions from Education to Employment* survey in Switzerland and the Canadian *Youth in Transition Survey*, observe the trajectories of school leavers from the educational system to the labor market.

Thematic Social Science Surveys

The Importance of National Archives and Statistical Institutes

The archives for social science data are well stocked with individual thematic social science surveys. Of most interest to social scientists are surveys that are of relevance to their own research questions. Hence, when searching for national manifestations of specific social phenomena, it is worth taking a look at the archives. Their offerings include, for example, national and cross-national comparative surveys on topics such as the labor market, criminality, victimization, and health. These studies range from very specific individual surveys to national or regional surveys that are repeated at regular intervals.

The main individual thematic social science surveys are those conducted by NSIs. The value of these surveys lies in the fact that they are conducted regularly and thereby facilitate the analysis of societal trends. Moreover, to the extent that they are generated in the member countries of the EU, many of these datasets must be submitted to the Statistical Office of the European Union (Eurostat). This means that, in addition to the analysis of national trends, there is the added value of cross-national comparison – at least within the EU.

Labor-Related Surveys

Of interest in this connection is the *Labour Force Survey* (LFS), "a standard household-based survey of work-related statistics" conducted in 200 countries and territories worldwide (ILO INFORM, 2013). The underlying concept of the survey was developed by the International Labour Organization (see ILO, 1982). The LFS employment, labor market, and working conditions data serve to determine the economic potential of the society in question. Since 1992, the LFS is supposed to be implemented on an annual basis. However, in 1998 the Council of the European Union adopted a regulation ruling that the LFS was to be conducted by the EU member states as "a continuous survey providing quarterly and annual results" (Council of the European Union, 1998: p. 3). The quarterly results provide an insight into seasonal fluctuations on the national labor markets.

In addition to measuring respondents' demographic background, data are collected on the following topics: labor status,

employment characteristics of the main job and the second job, visible underemployment, search for employment, education and training, previous work experience of persons not in employment, main labor status, and income (The Council of the European Union, 1998: pp. 4 f.).

Crime and Victimization Surveys

Crime and victimization are topics that are well represented in national surveys as evidenced by the growing number of *crime surveys*, *crime and safety surveys*, and *crime and victimization (victim) surveys*. As the United Nations Office on Drugs and Crime notes: "While in the past only police and criminal justice data were used to measure crime, it is now widely acknowledged that such information alone is not sufficient and should be integrated with victim surveys results" (UNODC, 2013). Irrespective of which of the three labels they bear, these surveys can be found in numerous countries – not only in North America, Europe, Australia, and New Zealand but also in countries such as South Africa and Nigeria. Their primary aim is to study the national situation.

Because some of these surveys have been conducted on a regular basis for several decades, they are a valuable resource when it comes to monitoring trends and identifying changes over time. Examples of such long-standing surveys include the annual *National Crime Victimization Survey*, which has been conducted by the US Census Bureau on behalf of the Bureau of Justice Statistics since 1973 (approx. 75 000 respondents in 2011) and the *British Crime Survey*, which was first implemented in 1982 and has been conducted annually since 2000 (approximately 50 000 respondents in 2011). Even though the questionnaires employed by national surveys are not cross-nationally harmonized, cross-national comparisons are still possible. Comparability is enhanced by the fact that in 1984 the United Nations decided to build a database of crime and victimization survey data. This led to the initiation of the *International Crime Victim Survey* (ICVS) by a group of European criminologists (Van Dijk et al., 1990). The survey has been conducted in some 80 countries since 1989 using a standardized questionnaire and methodology (UNODC, 2013). A consortium of commercial and academic research institutes cofinanced by the European Commission is now responsible for the execution of the ICVS in the member states of the EU. The EU version of the ICVS, the *European Crime and Safety Survey*, was implemented for the first time in 2005 in 18 EU member states using an instrument and methodology based on that of the ICVS (EU-ICS, 2013).

In recent years, crime and safety surveys have been supplemented by economic crime surveys. At present, one main focus of these surveys is cybercrime.

Health-Related Surveys

Health is another important topic of social science surveys. The first *health surveys* go back to the 1950s. Here, too, in addition to once-off surveys, a large number of countries employ a replicating cross-sectional design in order to monitor trends in the health status and behavior of the population (persons living in private households). The health surveys differ from other surveys in that, in some cases, the target population is not the resident population as a whole but rather specific subgroups within the resident population, selected on the basis of age or sex. Moreover, the territorial hierarchy of the investigation areas can range from the world (*WHO World Health Survey*), through national states, to regions, whereby a region can be a federal state, as in the case of the USA or Australia, a province, a district, or even a metropolitan area (for example, the *New York City Community Health Survey*).

The health surveys employ a random sample drawn from the specified target population and a standardized questionnaire. Interviews are frequently supplemented with rudimentary medical examinations entailing the measurement of weight, height, and blood pressure. Some studies also take blood and urine samples. However, datasets from studies of this kind are found less in social science archives than in those, often public, research institutions under whose aegis the survey is conducted. Many of these datasets are classified as public use files and are accessible for research purposes.

Barometer Surveys

Another category of social science survey is the barometer survey. Barometer surveys are cross-sectional surveys of national population samples that are repeated at regular intervals. In addition to socio-demographic variables, these surveys measure respondents' attitudes to the social, economic, and political situation in the country in question. With the exception of Oceania, barometer surveys are conducted on all continents, although they are sometimes limited to metropolitan areas. The following is a brief introduction to the main barometer surveys:

The Eurobarometer

Eurobarometer (EB) is a series of surveys conducted on a regular basis on behalf of the European Commission in all EU member states and the candidate countries (EB, 2013). The aim of the EB program is to monitor the evolution of public opinion in the member states and the candidate countries. Apart from questions about topical issues, respondents are asked for their opinions on topics such as European unification, the social situation of minorities, health, culture, and environmental protection.

The main EB survey is the *Standard EB*. First implemented in 1973, it is now fielded biannually. The Standard EB is supplemented by surveys such as the *Special EB* and the *Flash EB*. The Special EB comprises in-depth thematic studies conducted on behalf of EU institutions on topics such as the European elections (2009) or climate change (2011). The survey data serve as a basis for decision making on the part of the institution in question. The Flash EB are ad hoc thematical telephone surveys aimed at quickly gauging public reaction to current events. Between 2001 and 2004, separate surveys – the *Candidate Countries Eurobarometer* (CCEB), the methodology of which was almost identical to the Standard EB – were conducted in the 13 countries that joined the EU between 2004 and 2007, and in Turkey. The objective of these surveys was to gauge public opinion on European integration and EU institutions and policies. The CCEB replaced the *Central and Eastern*

EB. On the one hand, the EB surveys are a political instrument commissioned by political institutions. On the other hand, however, they yield a wealth of data on the development of living conditions and the attitudes of the citizens of current and potential EU member states, data that are of great interest to social researchers.

The Afrobarometer

The *Afrobarometer* "is an independent, nonpartisan research project that measures the social, political, and economic atmosphere in Africa" (Afrobarometer, 2013). The first Afrobarometer was conducted in 1999 in 12 sub-Saharan African countries. Round 5, conducted by public and private research institutes, got underway in 22 countries north and south of the Sahara in 2011, and data collection was still ongoing in some countries at the time of going to press (2013). It is expected that up to 35 countries will participate. The survey is methodologically well documented and the data are freely available online via the Afrobarometer Web site. Respondents are asked to assess both their own economic and living conditions and the economic and living conditions in the country as a whole. Although a standardized questionnaire is used in all participating countries, the data are more useful for the analysis of the respective national situation than for cross-national comparison purposes.

The Latinobarometer

First fielded in 1995, the *Latinobarómetro* is an annual survey conducted in up to 20 Latin American countries. Between 1000 and 1200 respondents are interviewed in each participating country. The Latinobarómetro focuses on social, economic, and political development. Topics covered include democracy, institutions, law and constitution, politics, political parties, participation, corruption, discrimination, poverty, international relations, trust between countries, values, communication media, social capital, the economy, and trade (Latinobarómetro, 2013).

The AmericasBarometer

The *AmericasBarometer* is conducted by the academic institution, Latin American Public Opinion Project (LAPOP), which is hosted by Vanderbilt University in Nashville, Tennessee. LAPOP began with the study of democratic values in Costa Rica in the 1970s. Between 1991 and 2005 it conducted 41 surveys in Latin American countries. The *AmericasBarometer* was established in 2004, with 11 countries participating in the first round. Since then it has been conducted biennially in a steadily growing number of countries in North, Central, and South America, and the Caribbean. The US, 24 Latin American countries, as well as Canada participated in the latest round, which took place in 2012. The survey focuses on "popular opinions toward the economy, tolerance, corruption, and support for democracy in the Americas" (AmericasBarometer, 2013).

The Asian Barometer Survey

Established in 2001, the *Asian Barometer Survey* (ABS) grew out of the *East Asia Barometer*. It is an "applied research program on public opinion, on political values, democracy, and governance around the region. The regional network encompasses research teams from 13 East Asian political systems, and 5 South Asian countries" (ABS, 2013). The main aim of the ABS is to gauge public opinion on political values, democracy, governance, human security, and economic reforms. The latest wave took place in 2011.

The Arab Democracy Barometer

The *Arab Democracy Barometer* was established in 2005. The second wave of the survey, which was conducted between 2010 and 2012 in eight Arab countries, sought to measure politically relevant attitudes of the resident populations. The organizers attached importance to being able to compare countries beset by political unrest with countries unaffected by such unrest. The second wave sought to measure "citizen attitudes, values, and behaviour patterns relating to pluralism, freedoms, tolerance and equal opportunity; social and inter-personal trust; social, religious and political identities; conceptions of governance and an understanding of democracy; and civic engagement and political participation" (Arab Democracy Barometer, 2013).

The New Democracies Barometer

The *New Democracies Barometer* (NDB) was established in 1991 by the Paul Lazarsfeld Society in Vienna. Between 1991 and 1998, surveys were conducted in 10 Central and Eastern European countries that had a socialist system until 1990 and were undergoing transformation at the time. The aim was to gauge the effect of the experience of democracy and markets on the attitudes and behavior of the people.

Conclusion

Large internationally accessible social science datasets help researchers to classify their own data and may also facilitate comparative analyses. The fact that more and more data holders are providing users with online access to their data and are making their analytical tools/programs freely available to the scientific community has considerably enhanced access to these datasets. The increasing cooperation between the data archives has also considerably enhanced the universal accessibility of their respective data.

However, comparability across datasets is hampered by the lack of standardization of demographic and socioeconomic variables. At present, datasets can be harmonized and compared only according to the sex, age, and education (in three broad and largely undefined categories) of the respondents. Therefore, in order to fully realize the analytical potential of survey data, generally accepted rules for the standardization of demographic and socioeconomic variables are called for.

See also: Cross-National Comparisons in Education: Findings from Pisa; Data Bases and Statistical Systems: Crime; Data

Bases and Statistical Systems: Education, Statistical Systems; Health Surveys; International Social Survey Programme; Panel Surveys: Uses and Applications; Sample Surveys: Cognitive Aspects of Survey Design; Survey Research Centers and Companies; Surveys and Polling: Ethical Aspects.

Bibliography

Converse, Jean M., Presser, Stanley, 1986. Survey Questions: Handcrafting the Standardized Questionnaire. Sage, Beverly Hills, CA.

Groves, Robert M., Fowler Jr., Floyd J., Couper, Mick P., Lepkowski, James M., Singer, Eleanor, Tourangeau, Roger, 2004. Survey Methodology. John Wiley and Sons Inc., Hoboken, NJ.

International Labour Organization (ILO), 1982. Resolution concerning statistics of the economically active population, employment, unemployment and underemployment, adopted by the Thirteenth International Conference of Labour Statisticians. www.ilo.org/public/english/bureau/stat/download/res/ecacpop.pdf (accessed 15.02.13.).

International Labour Organization INFORM, 2013. INFORM Bureau of Library and Information Services. Labour Force Surveys. http://www.ilo.org/dyn/lfsurvey/lfsurvey.home (accessed 15.02.13.).

Kish, Leslie, 1994. Multipopulation survey designs: five types with seven shared aspects. International Statistical Review 62, 167–186.

Krosnick, Jon A., Fabrigar, Leandre R., 2006. The Handbook of Questionnaire Design. Oxford University Press, New York.

Krosnick, Jon A., Presser, Stanley, 2010. Question and questionnaire design. In: Marsden, Peter V., Wright, James D. (Eds.), Handbook of Survey Research, second ed. Emerald Group Publishing Limited, Bingley, pp. 263–313.

McGonagle, Katherine A., Schoeni, Robert F., Sastry, Narayan, Freedman, Vicki A., 2012. The panel study of income dynamics: overview, recent innovations, and potential for life course research. Longitudinal and Life Course Studies 3 (2), 268–284.

Schwarz, Norbert, 2008. The psychology of survey response. In: Donsbach, Wolfgang, Traugott, Michael W. (Eds.), The Sage Handbook of Public Opinion Research. Sage Publications, Los Angeles, pp. 374–387.

Smith, Tom W., 1979. Who, what, where, and why: an analysis of usage of the General Social Survey, 1972–1978. In: GSS Technical Report No. 12. National Opinion Research Center.

Stafford, Frank P., 2010. Panel surveys: conducting surveys over time. In: Marsden, Peter V., Wright, James D. (Eds.), Handbook of Survey Research, second ed. Emerald Group Publishing Limited, Bingley, pp. 765–793.

The Council of the European Union, March 03, 1998. Council regulation (EC) No 577/98 of 9 March 1998 on the organisation of a labour force sample survey in the community. Official Journal L 077, 0003–0007.

Van Dijk, J.J.M., Mayhew, P., Killias, M., 1990. Experiences of crime across the world. Key Findings from the 1989 International Crime Survey. Kluwer Law and Taxation Publishers, Deventer.

Relevant Websites

ABS, 2013: http://www.asianbarometer.org/ (accessed 15.02.13).
Afrobarometer, 2013: http://www.afrobarometer.org/ (accessed 15.02.13).
AIES, 2013: http://aises.nic.in/aboutsurvey (accessed 15.02.13).
ALLBUS, 2013: http://www.gesis.org/en/allbus
AmericasBarometer, 2013: www.vanderbilt.edu/lapop/ (accessed 15.02.13).
Arab Democracy Barometer, 2013: http://www.arabbarometer.org/ (accessed 15.02.13).
BHPS, 2013: https://www.iser.essex.ac.uk/bhps/ (accessed 15.02.13).
China GSS, 2013: http://www.ust.hk/~websosc/survey/GSS_e.html (accessed 15.02.13).
DHS, 2013: http://www.centerdata.nl/en/TopMenu/Projecten/DNB_household_study/ (accessed 15.02.13).
EB, 2013: http://ec.europa.eu/public_opinion/index_en.htm (accessed 15.02.13).
ECHP, 2013a: http://circa.europa.eu/irc/dsis/echpanel/info/data/information.html (accessed 15.02.13).
ECHP, 2013b: http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/echp (accessed 15.02.13).
ESDS, 2013: http://www.esds.ac.uk/ (accessed 15.02.13).
EU-ICS, 2013: http://www.europeansafetyobservatory.eu/ (accessed 15.02.13).
EU-SILC, 2013a: http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/eu_silc (accessed 15.02.13).
EU-SILC, 2013b: http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/documents/EN-EU-SILC-MICRODATA.pdf (accessed 15.02.13).
HILDA, 2013: http://www.melbourneinstitute.com/hilda/ (accessed 15.02.13).
ISSP, 2013: http://www.issp.org/page.php?pageId=216 (accessed 15.02.13).
KLIPS, 2013: http://www.kli.re.kr/klips/en/about/introduce.jsp (accessed 15.02.13).
Latinobarómetro, 2013: www.latinobarometro.org/ (accessed 15.02.13).
LiBiH, 2013: http://www.fzs.ba/LiBiHIntroENG.htm (accessed 15.02.13).
LINDA, 2013: http://www.scb.se/Pages/Standard____38903.aspx (accessed 15.02.13).
LISS, 2013: http://www.lissdata.nl (accessed 15.02.13).
MxFLS, 2013: http://www.ennvih-mxfls.org/en/mxfls.php?seccion=1&subseccion=1&session= (accessed 15.02.13).
NHES, 2013: http://nces.ed.gov/nhes/ (accessed 15.02.13).
PIAAC, 2013: http://www.oecd.org/education/skills-beyond-school/piaacprogrammeforthe internationalassessmentofadultcompetencies.htm (accessed 15.02.13).
PIRLS. 2013: http://www.iea.nl/pirls_2011.html (accessed 15.02.13).
PISA, 2013: http://www.oecd.org/pisa/ (accessed 15.02.13).
PSID, 2013: http://www.psid.org (accessed 15.02.13).
RLMS, 2013: http://www.cpc.unc.edu/projects/rlms-hse (accessed 15.02.13).
SHP, 2013: http://www.swisspanel.ch/?lang=en (accessed 15.02.13).
SLID, 2013: http://www.statcan.gc.ca/pub/75f0011x/75f0011x2012001-eng.htm (accessed 15.02.13).
SOEP, 2013: http://www.diw.de/en/diw_02.c.221178.en/about_soep.html (accessed 15.02.13).
TALIS, 2013: http://www.oecd.org/edu/school/What%20Can%20Be%20Done%20to%20Support%20New%20Teachers.pdf (accessed 15.02.13).
TIMSS, 2013: http://www.iea.nl/timss_2011.html (accessed 15.02.13).
Understanding Society, 2013: https://www.understandingsociety.ac.uk (accessed 15.02.13).
UNODC, 2013: http://www.unodc.org/unodc/en/data-and-analysis/Crime-Victims-Survey.html (accessed 15.02.13).
US GSS, 2013: http://www3.norc.org/GSS+Website/ (accessed 15.02.13).

Data Collection: Interviewing

Albert Hunter, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 5, pp. 3230–3236, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Interviews are fundamental tools in the social sciences for different styles of research such as experimental, survey, fieldwork, and nonreactive research. According to some theories of measurement, all measurement involves interaction (usually between two people), and interviews are particularly noteworthy in this respect. The article relates four theories of social interaction within the interview process: role theory, exchange theory, conflict theory, and symbolic interaction. It concludes by noting the necessity of a multimethod strategy to refine the use of interview techniques in social science research.

Interviewing is a process of social interaction in which a researcher questions subjects or respondents for the purpose of gaining information or data. Interviewing styles may range from informal conversations to semifocused topical interviews to more formal standardized questionnaires that reflect different degrees of control over the interview. As social interaction, interviewing may be analyzed by different theories of social interaction – specifically, role theory, conflict theory, exchange theory, and symbolic interaction. Each of these highlights different aspects of the interview that may introduce bias and affect the validity of the data. Research is presented that demonstrates the degree of bias introduced by race, class, gender, age, and various other aspects of the interview situation.

The social science interview is a situated social interaction, the purpose of which is to transfer information from one person to another. Information or data are obtained from one person, the respondent or subject, as he or she answers questions posed by another, the interviewer or researcher. The interview in its myriad forms is a primary data-gathering tool of social scientists, and therefore, it is important to understand the interview process, since this will affect the nature and quality of the data gathered or generated.

Situating the Interview

Interviews may be conducted in a variety of ways and in a variety of situations. Inevitably, they are socially and methodologically embedded in larger contexts (Kvale and Brinkmann, 2009). Relatively little research has been done either on the impact of different social situations and settings on the conduct of interviews or on the impact of varying situations on the reliability and validity of interview results. The social setting may range from the privacy of the home to the office to the public 'person on the street,' and the social context may range from a highly private, anonymous, or confidential one-on-one interview to one involving the presence of strangers or intimate family and friends as co-respondents. Research has shown that in contrast to the one-on-one interviewer–respondent situation, the presence of others tends to produce more socially normative responses (Lever, 1981). Much more research is needed on the impact of these varying social contexts.

We must also situate interviews in the different methodologies or styles of research used by social scientists. For heuristic purposes, we can identify four major styles of research: experimental, survey, fieldwork, and nonreactive research. The interview, though found in all the four methods, is particularly important in survey and field research, and these may be thought of generally as representing quantitatively verses qualitatively oriented interview styles, respectively. The quantitative–qualitative distinction itself encompasses a variety of comparisons that may range from more philosophical and epistemological issues such as objective verses subjective knowledge to more grounded and specific issues of questionnaire construction and closed verses open answer formats. A central distinction running throughout these comparisons is the question of the nature of control over the interview format and content.

A Note on Control over the Interview – Form and Content

Different interview formats may be arrayed on a number of different continua, but of particular significance is the question of varying degrees of control. At one extreme, the interviewer exerts full control over the situation, and at the other, the interviewer seemingly minimizes control, relinquishing more control over the situation to the respondent. There is a zero-sum quality of control in the relationship between interviewer and respondent. The case of maximal interviewer control is exemplified by the 'forced-choice' survey interview, where not only the format of the question is standardized (stimulus), but the possible answers of the respondent are standardized as well (response). The interviewer ensures a constancy of routinized control in order to generate highly reliable data. Concerns over such control are often associated with surveys where the interview is used as a tool to generate standardized data from a sample of numerous units for the purposes of aggregate analysis that may be generalized to some larger population.

The other extreme, where the interviewer exercises minimal control, is seen, for example, in the Rogerian psychoanalytic interview. Here the respondent is 'in control' as to the content of the interview and the interviewer is trained to be 'nonreactive.' It should be noted, however, that even here the

interviewer is exercising control over the format or form of the interview, and that both the forced-choice survey interview and the nonreactive interview (Rogers, 1945) are interaction situations far removed from the 'natural' form and content of everyday personal interaction. The field interview of participant observation is, by contrast, a situation often closer to the give-and-take of everyday 'normal conversation' (Seidman, 2013). The respondent exercises a degree of control over both the content and the form, but the interviewer, through judicious use of probes, can often guide the interview into topic areas of interest to the researcher. This is the case with the 'focused' or 'topical' interview (Merton, 1956). From these considerations, one may construct a typology of control that the researcher exercises in the interview situation:

typology of interviewer control in different styles of research interviews

	Control over content	
	Hi	Low
Hi	survey questionnaire Control over form	nondirective psychoanalytic
Low	semifocused in-depth	conversational fieldwork

These different interview formats are often associated with survey verses fieldwork, but extend to a variety of methods. A multimethod strategy that emphasizes different strengths and weaknesses of different methods or styles of research nonetheless values the interview as a central data collection tool that can simultaneously provide quantitative and qualitative data, and even be used in multimethod experimental strategies. For example, Schuman et al. (1985) incorporated an experimental design into a survey by having interviewers systematically vary the format of questions asked of different respondents.

The Theory of the Spring

Some have suggested that all measurement involves interaction (Babbie, 2013) – interaction with the external world in some form or another in order that information about that external world may be transferred to an observer. Theories of measurement take into account a detailed exposition of the measurement process itself as constitutive of the data generated. In the prosaic act of weighing a bunch of bananas in the produce section of a grocery store, for example, one is operating on an assumption about how the spring in the scale responds to varying masses of bananas subjected to the law of gravity in moving the dial in a consistent and credible way. One operates on the assumption that some scientist or engineer has worked out in detail the physical elasticity of the steel spring at the heart of the scale's measurement process. In the interview, this interaction is particularly noteworthy, since it is *social* interaction – usually between two people – and the corresponding theories of measurement most applicable are theories of social interaction itself. In this section, we will briefly apply four distinct theories of social interaction to the interview to see the varying perspectives they provide on this all-important data-gathering tool. The four theories of social interaction are (1) role theory, (2) exchange theory, (3) conflict theory, and (4) symbolic interaction.

Role Theory

Role theory begins with a set of normative expectations that are presumed to define particular positions or statuses in social structure and their corresponding roles or behaviors in interaction with others. The roles or statuses most clearly central here are those of 'interviewer' and 'respondent' themselves. How well one knows these roles, that is, knows the normative expectations for behavior associated with the respective roles, is a function of one's prior experience and knowledge gained either first hand or through vicarious observations – in the media, through cartoons, through conversations, or in the classroom. To be 'trained' as an interviewer is to learn a set of normative expectations about how one should interact with a respondent. The simplest norms are 10 of those that are 'taken for granted' such as who asks the questions and who gives the answers. Sometimes there are widely accepted norms defining the ideal interviewer role such as one should not reveal personal information that might thereby bias the respondent's responses (Gordon, 1975). However, even these commonplace, widely shared norms may come into question, as there may be conflict among the different goals for the interview. For example, one goal may be to get as much revelatory information as possible verses another goal to not bias the respondent's answers. The different goals of the interview (more fully revelatory data vs unbiased data) may come into conflict and suggest different contradictory norms, for example, about how much personal information the interviewer should reveal to the respondent in the interview situation.

Role theory asserts that the norms governing interaction are there for the purpose of realizing specified goals out of the interaction. In short, normative behavior is goal oriented. The norms governing the interviewer are likewise oriented to maximizing certain qualities (or values) of the data – for example, norms that call for a common stimulus to achieve validity across respondents or other norms that emphasize supportive interaction to produce full and complete responses.

A common application of role theory to the interview situation is a concern with bias introduced by other social statuses in a person's status set – the most obvious and frequently researched being the ascribed and highly visible social statuses of race, gender, and age. These statuses are often considered to be especially important when questions in the interview touch on topics closely related to them. Much research has been conducted on the effects of these statuses on interview outcomes precisely because they are considered to be master ascribed statuses that are both readily observed and ubiquitous in all interactions. They are often attributed even in telephone interviewing and not just in face-to-face situations.

One of the earliest findings of interviewer effects by race, dating to World War II, was that answers to questions about race relations were strongly impacted by the races of the interviewer and respondent (Hyman et al., 1954). African-Americans reported lower levels of satisfaction with race relations to African-American interviewers than to Caucasian interviewers. Caucasian respondents reported lower levels of

acceptance of African-Americans to Caucasian interviewers than to African-American interviewers. These results have been demonstrated repeatedly since then in a study by Schuman et al. (1985) It is important to note that more limited effects of race of interviewer and respondent have been observed on studies of other topics unrelated to race.

Another interviewer status examined in some detail has been gender. Earlier research has generally shown that the gender of the interviewer has no effect on responses to survey questions except when the content of the questions relates to sexual behavior or gender-related issues (Clark, 1952; Hyman et al., 1954). Research by Kane and Macaulay (1993) found that both male and female respondents express more egalitarian gender-related attitudes or greater criticism of existing gender inequalities to female interviewers. Furthermore, male respondents offer significantly different responses to male and female interviewers on questions dealing with gender inequality.

Age, as the third of the most visible ascribed status characteristics, has also been found to have only limited interviewer effects. School-age subjects have shown different responses to older verses younger interviewers for questions about peer influences. Younger interviewers obtained slightly more peer-oriented and less adult-oriented answers than older interviewers. With respect to age of respondents, research has shown that telephone surveys tend to underrepresent older respondents, but response distributions do not vary by telephone verses face-to-face interviews across age categories (Herzog et al., 1983). As with race and gender, age appears to have little effect on interview responses except when the topic being covered is directly related to the status itself.

Exchange Theory

The central assumption of exchange theory is that all interaction is based on ego's attempt to realize a 'social profit' by minimizing the cost of one's own actions and maximizing the benefits from alter's actions. The operating principle is that for interaction to continue, the norm of reciprocity must be followed such that the ratio of two parties' costs-to-benefits should be equal. If we view the interview as an exchange, it is relatively easy to define the benefits to the researcher: they are the data provided by the respondent that are his or her costs in time and effort in answering the questions. It is often more nebulous to define what the benefits to the respondent are, that is, the costs incurred by the interviewer that directly benefit the respondent. Of course, when respondents are being directly paid by the interviewer for their time, the benefits are clear and unambiguous. Market researchers, for example, sometimes establish panels of randomly selected respondents, who are paid for their time in a variety of repeated surveys. In short, they are 'professional respondents' for whom the exchange is a pure labor market transaction. Exchanges in kind may also be provided such as a lunch or dinner, a free gift, or a chance at a drawing. More often than not, however, the benefits are assumed to be some nebulous internal psychic reward that can include everything from the opportunity to vent one's opinions on a particular topic, that is, make use of a willing listener, to an increased sense of self-worth engendered by volunteering to help someone out, especially when the interviewer may have already appealed to one's 'specialness' by indicating that they have been 'selected' to participate. Altruistic appeals also redound to one's sense of self-worth, as when one is solicited as a respondent to provide data that may be useful in developing new policies and programs or evaluating existing ones even if one is not seen to be a direct beneficiary of them. Even the most abstract altruistic appeal – that by responding one is 'contributing to knowledge' – may at times work as a sufficient psychic reward.

What the exchange perspective highlights is that 'time spent' by the respondent is a very real but variable cost that may influence their willingness to be interviewed in the first place and their loquaciousness even if an interview is granted. The preciousness of 'time spent' doing an interview is differentially valued from one potential respondent to another (for example, the working mother vs the elderly retired nursing-home resident). Such calculations are highly likely to influence differential response rates in a selective or biased way. Not only are these calculations likely to vary from one set of statuses to another, but they will likely vary even for the same individual under different circumstances and in different situations such as time of day or day of week or season of the year, or at home, or at shopping, or work (Pol, 1992).

The cost of an interview is not simply one sided, for the interviewer's time is also a very tangible cost and one that must be taken into consideration often in a direct monetary way as when one is hiring interviewers or relying on a survey supplier. Such costs may have direct bearing on the nature and quality of the data gathered, for they can affect the length of the interview forcing real choices in research design as to sample frame and size, or number of topics explored and questions asked.

In short, from an exchange perspective, the interview is a cost–benefit transaction. The calculus of these varying costs to the interviewer/researcher must be weighed against the value of the data to be generated. In the design of surveys, for example, a distinction is often made between open-ended verses closed-ended questions, and the former generally prove to be more difficult and more expensive to code and analyze than latter. However, Geer (1991) found in an experiment that the open-ended comments are very effective in measuring the important concerns or 'salient' issues of respondents. In short, it appears to be the cost considerations as much as validity issues that have led to the dominance of forced-choice questions in survey research.

Costs and benefits of the interview also form a part of the calculus of the ethical considerations involving human subjects. Costs to the respondent slide over into a variety of 'risks' that may result from the disclosure of certain data and the loss of privacy. Anonymity or confidentiality may be required to forestall those costs. The idea behind 'informed' consent is that this calculus should be made by the subject/respondent based on a full disclosure of risks and benefits. Research in which there are substantial risks can only be justified by substantial benefits that outweigh these costs, and above all, it is up to the respondent to make this determination.

We should be aware that the utilities (value) of a respondent's cost/benefits may differ from those of the researcher and one is confronted with incommensurability, which may work to the advantage of the researcher. For example, an elderly

retired single respondent may have a surplus of free time, which is of low cost to give in answering a lengthy questionnaire, while to the highly trained interviewer this time is a valuable costly commodity in the research budget. Such incommensurability may at times work to the detriment of the interviewer when the inability to reach a common understanding of the relative value of information or time leads to an unwillingness to participate in what is considered to be a 'worthless' activity. (We note that this is one of the points of the symbolic interactionist perspective outlined below, which involves negotiating not only a common value but a common understanding of the meaning as well as the value of various social goods.)

Power Theory

The power or conflict perspective directly addresses the insight that much social interaction – and also the interview – involves differences in power among the parties. To be sure, power or conflict theory may be closely linked to either role theory (as when, for example, one talks about the differences in power that are present between the status of a male and a female in interaction) or exchange theory (as when the costs to participating in the interview are different for the interviewer and the respondent, or the benefits that one can provide to the other are different and one or the other party uses those differences in resources to extract what they want from the other). What the power or conflict perspective stresses is the reality that interviewer and respondent often have different degrees of power relative to one another and these may influence the nature and quality of the data generated in the interview.

Let us first address situations in which the interviewer has more power. We will use power in its broadest meaning to encompass the most subtle uses of prestige to the blatant uses of physical coercion. The latter is rare if ever present, but the former is quite ubiquitous. The imprimatur of an individual's title 'Doctor' or 'Professor,' or an institution's stature 'X University' or 'Y Agency' are often used to impress potential respondents as to the importance of the research and the desirability of their participating. In the actual conduct of the interview itself, power may be reflected in who makes the effort to initiate the contact or who travels to whom to conduct the interview. Being on one's home turf would seem to imply more power and control over the situation. A notable exception is the case in which elites may avoid being interviewed in certain situations (coming to their homes) in order to maintain a heightened degree of privacy. Researchers have sometimes strategically relied on the interviewer shifting out of the role of interviewer to becoming an observer and recording the presence or absence of such things as a piano, types of pictures and paintings hung on walls, and the use of curtains or drapes (Laumann and House, 1970). This is an example of combining methods in a multimethod strategy (Brewer and Hunter, 1990) to provide validity checks surrounding lifestyle and status and class variables.

At times, respondents may sense a threat from the interview situation that reflects a potential power differential between them and their interviewers. Potential threats may apply to divulging information concerning content areas that include socially undesirable and even illegal behaviors and could lead to not reporting or underreporting various forms of deviance and even criminal behavior ranging from drug use to official 'cheating' (Bradburn et al., 1979). Implicit threat could also lead to overreporting socially desirable behavior such as voting or charitable giving. There is some limited evidence that interviewer characteristics are related to threat. Respondents are willing to report socially undesirable behavior such as drug use to interviewers who are perceived to be more like themselves (Johnson et al., 2000).

The most widely used procedure to reduce threat is to eliminate the interviewer entirely or as far as possible from the data gathering process. In some cases, this means group administration of the interview in a public or quasi-pubic setting but where responses are private such as a school classroom setting. Even if the interviewer is present, it is possible that the threatening questions may be self-administered. Maximum privacy is obtained when the threatening questions are on a computer and the respondent reads questions from a screen. The respondent enters the answers directly into the computer and the interviewer is completely unaware of what answers are given (Tourangeau and Smith, 1996). It must be recognized that the use of self-administered methods reduces but does not eliminate respondent misreporting due to the power differentials in the interview situation and the potential threat of disclosure. It is precisely why the use and assurance of anonymity or confidentiality are not only ethical considerations but themselves may have a direct and positive impact on the validity of data generated through the interview.

Power is implicit in the very asymmetry of who asks questions and who answers them. The benefit is seen to be asymmetrical as the powerful ask questions of, glean information from, and acquire data about the less powerful. Much more rarely does information flow in the reverse direction – from the powerful to the powerless. Knowledge is power and its differential distribution is a means of obtaining and maintaining that power (Hunter, 1993).

Interviewing of elites is a situation in which the respondent often has more power than the interviewer. Such control may be reflected first in limiting the ability to gain access to begin with and continue through the interview situation with control over the content, format, recording, and even the use of interview data (Thomas, 1993). Although interviewers in their comportment should, of course, show respect to all respondents, an added degree of deference might be shown to those having more power or higher status than the interviewer. This deference could easily affect the likelihood of the interviewer's continuing to probe in finer and finer detail into the lives of the respondents, thereby limiting the quality and depth of the data.

The power differential enters as well into ethical considerations of the interview, and this is why interviews must be voluntary, since the countercase involving any form of coercion, no matter how mild, may be deemed unethical. For example, if a welfare recipient is being interviewed for research purposes, he/she might agree to participate in the interview if it is closely linked to the intake interview used to determine welfare eligibility, or a person in a public setting may agree to be interviewed out of a sense of peer pressure (e.g., students in a classroom setting). It is imperative that researchers be fully

cognizant of and sensitive to the power relations involved in the interview, as these may not only affect the quality of the data but reflect larger social processes of which the interview itself is a microcosm.

Symbolic Interaction

The interview situation is not one which is necessarily clearly defined and unambiguous, especially on the part of respondents. Therefore, one of the key activities of the good interviewer is to negotiate with the respondent in defining the appropriate behaviors and what can be expected in the interview situation. Drawing on the symbolic interactionist framework (Goffman, 1959), the interview may be considered to be a situation in need of mutual definition through the interaction of the interviewer and the respondent. At a minimal level is the idea that the interviewer is supposed to ask questions and the respondent is to give answers. But the fine nuances of this somewhat 'artificial' interaction are a negotiated reality in which the interviewer has the task of 'educating' the respondent into the appropriate behavior. For example, a too loquacious respondent might be interrupted in the course of an answer and asked to respond more simply – 'strongly agree, agree, neutral, disagree, or strongly disagree.'

The negotiated definition of the situation draws on the past experiences of both interviewer and respondent. The idea of interviewer training focuses on behaviors used to aid in an appropriate definition of the situation such as what to say to gain access, how to present self for gaining agreement to participate, how to develop and maintain rapport and how to terminate an interview, and of course, how to avoid biasing respondents (Billiet and Loosveldt, 1988). Some respondents may be familiar with the interview situation from having participated in them previously, while others may have been exposed to the situation vicariously through the media. Some may in fact be 'professional respondents,' that is, those paid to participate in ongoing panel studies in which they have been more thoroughly trained or educated into the respondent's role.

Expectations of the interviewer about the interview process have also been found to influence the nature of the respondents' participation. Some interviewer expectations are formed before the interviewer conducts any interviews, for example, expectations about how difficult or threatening certain questions may be to respondents, while other expectations may develop in the course of an interview as the interviewer begins to form a picture of the respondent through his/her prior responses, and yet other expectations may emerge as the interviewer completes multiple interviews and begins to form a patterned set of expectations for emerging 'types' of respondents. For example, interviewers who expected that questions about threatening behavior would be easy to ask received higher rates of responding than interviewers who thought the questions would be difficult to ask (Bradburn et al., 1979). It would appear that interviewer behavior in some way reflects expectations and that behavior calls forth different responses on the part of subjects.

Ambiguity in social interaction is a key problematic in the symbolic interactionist framework. Much interaction is seen to be specifically oriented to the reduction of ambiguity in an attempt to arrive at a mutually agreed upon definition of the situation. The issue of ambiguity is also significant in the interaction between interviewer and respondent. If a respondent should ask quizzically, what is meant by a question, often interviewers are simply instructed to repeat the question without further elaboration or clarification of the question so as no to differentially bias one respondent verses another. On the other hand, if a respondent gives an answer that is limited, partial, or ambiguous, the interviewer is often instructed to 'probe' to get greater clarification. In a classic laboratory experiment, Hyman et al. (1954) demonstrated that interviewers faced with an ambiguous response coded the answer on the basis of previous answers in the questionnaire. Respondents were asked a series of questions about whether the United States should actively intervene or should isolate itself from world affairs. When the response was ambiguous, those who had previously given unambiguously interventionist answers were coded as giving an interventionist answer, while those who had previously given isolationist answers were coded as giving an isolationist answer. In this experiment, interviewers were not allowed to probe the ambiguous answer for clarification.

The first phase of an interview involves negotiation in securing the willingness of the respondent to participate. By negotiating a formal 'commitment' from respondents in the form of a signed 'pledge' to answer 'fully' and 'truthfully' has been found to increase the degree of cooperation by respondents and the completeness of their responses (Cannell et al., 1981). A further mechanism for increasing cooperation is to flexibly tailor the appeal to participate to each individual respondent rather than using a standardized approach. Such individually tailored appeals generate higher levels of cooperation and also experienced interviewers are more likely to obtain such higher rates of cooperation (Groves and Couper, 1998).

A Multimethod Strategy

Interviews in their various forms and formats will continue to be a primary tool in the social scientist's tool kit for generating reliable and valid data. Direct comparisons of the interview with other tools for data generating, such as content analysis of media, or conversational analysis, the use of systematic observation of behavior, or the analysis of archives can only further the understanding of the strengths and limits of the interview in a multimethod strategy. Some but far too little research has been concerned with such comparisons. One example is Janet Lever's (1981) multimethod study of children's play. She found that interviews with kids produced more gender-stereotyped responses to the nature of boy's verses girl's play than observations of their behavior by the researcher or diaries kept by the kids themselves. Throughout the above discussion, the multimethod strategy has been mentioned at various points and briefly noted in passing. Bringing the multimethod strategy to the foreground whether narrowly focused on measurement or more broadly on the research process as a whole (Brewer and Hunter, 1990) can only lead to a more refined use of the interview in social science research. The shape

and function of the interview as a tool can more clearly be delineated as it is compared with these other tools in the same research project. The centrality and the ubiquity of the interview require much more systematic research, fully exploring the social, situational, and methodological dynamics of the interview process. Whether viewed from a postmodern perspective of a greater need for reflexivity within the social sciences or a positivist perspective of understanding the sources of bias and the need to reduce measurement error, much more research on the interview itself is needed. A few lines of inquiry and understanding have been developed as noted, but a great abyss of ignorance remains to be plumbed.

At various points throughout this discussion I refer to data being either 'gathered' or 'generated.' I fully recognize the implications of these two distinct concepts, as the former privileges an older classical positivism, while the latter speaks to a more agency-oriented social constructionist view of science. The distinction will itself become part of our nuance discussion of the interview.

See also: Field Observational Research in Anthropology and Sociology; Observational Studies: Overview; Questionnaires: Cognitive Approaches; Questionnaires: Collection of Subjective Expectations Data; Sample Surveys: Cognitive Aspects of Survey Design.

Bibliography

Babbie, E., 2013. The Practice of Social Research, thirteenth ed. Wadsworth, Belmont, CA.

Billiet, J., Loosveldt, G., 1988. Improvement of the quality of responses to factual survey questions by interviewer training. Public Opinion Quarterly 52, 190–211.

Bradburn, N., Sudman, S., et al., 1979. Improving Interview Method and Questionnaire Design. Jossey-Bass Publishers, San Francisco.

Brewer, J., Hunter, A., 1990. Multimethod Research: A Synthesis of Styles. Sage Publications, Newbury Park, CA.

Cannell, C., Miller, P., Oksenberg, L., 1981. Research on interviewing techniques. Sociological Methodology 12, 389–437.

Clark, R.A., 1952. The projective measurement of experimentally induced levels of sexual motivation. Journal of Experimental Psychology 44, 391–399.

DiCicco-Bloom, B., Crabtree, B.F., 2006. The qualitative research interview. Medical Education 40, 314–321.

Geer, J., 1991. Do open-ended questions measure 'salient issues'? Public Opinion Quarterly 55, 360–370.

Goffman, E., 1959. The Presentation of Self in Everyday Life. Doubleday/Anchor, New York.

Gordon, R., 1975. Interviewing: Strategy, Techniques, and Tactics. Dorsey Press, Homewood, IL.

Groves, R., Couper, M., 1998. Nonresponse in Household Interview Surveys. John Wiley and Sons, Inc., New York.

Herzog, A.R., Rodgers, W., Kulka, R., 1983. Interviewing older adults. Public Opinion Quarterly 47, 405–418.

Hunter, A., 1993. Local knowledge and local power: notes on the ethnography of local community elites. Journal of Contemporary Ethnography 93, 36–58.

Hyman, H., Cobb, W., Feldman, J., Hart, C., Stember, C., 1954. Interviewing in Social Research. University of Chicago Press, Chicago.

Johnson, T., Fendrich, M., Shaligram, C., Garcy, A., Gillespie, S., 2000. An evaluation of the effects of interviewer characteristics in an RDD telephone survey of drug use. Journal of Drug Issues 30, 77–101.

Kane, E., Macaulay, L., 1993. Interviewer gender and gender attitudes. Public Opinion Quarterly 57, 1–28.

Kvale, S., Brinkmann, S., 2009. Interviews: Learning the Craft of Qualitative Research Interviewing, second edition. Sage Publications Inc., Thousand Oaks, CA.

Laumann, E., House, J., 1970. Living room styles and social attributes. Sociology and Social Research 54, 321–342.

Lever, J., 1981. Multiple methods of data collection: a note on divergence. Urban Life 10, 199–213.

Merton, R., 1956. The Focused Interview: A Manual of Problems and Procedures. The Free Press, New York.

Pol, L., 1992. A method to increase response when external interference and time constraints reduce interview quality. Public Opinion Quarterly, 356–360.

Rogers, C., 1945. The non-directive method as a technique for social research. American Journal of Sociology 51, 143–152.

Schober, M., Conrad, F., 1997. Conversational interviewing. Public Opinion Quarterly 61, 576–602.

Schuman, H., Steeh, C., Bobo, L., 1985. Racial Attitudes in America: Trends and Interpretations. Harvard University Press, Cambridge, MA.

Seidman, I., 2013. Interviewing as Qualitative Research, fourth ed. Columbia University Teachers College Press, New York.

Suchman, L., Jordan, B., 1990. Interactional troubles in face-to-face survey interviews. Journal of the American Statistical Association 85, 2322–2353.

Suchman, L., Jordan, B., 1991. Validity and collaborative construction of meaning in face-to-face surveys. In: Tanur, J. (Ed.), Questions about Questions: Inquiries into the Cognitive Bases of Surveys. Russell Sage Foundation, New York, pp. 241–267.

Sudman, S., Bradburn, N., 1974. Response Effects in Surveys: A Review and Synthesis. Aldine Publishing Company, Chicago.

Thomas, R., 1993. Interviewing important people in big companies. Journal of Contemporary Ethnography 22, 80–96.

Tourangeau, R., Smith, T., 1996. Asking sensitive questions. Public Opinion Quarterly 60, 275–304.

de Finetti, Bruno (1906–85)

Bruce Hill, University of Michigan, Ann Arbor, MI, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 5, pp. 3260–3263, © 2001, Elsevier Ltd.

Abstract

The contributions of Bruno de Finetti to probability and statistics are discussed. These include his celebrated exchangeability theorem, his theory of coherency, and his theory of finite additivity and adherent mass distributions. De Finetti was perhaps the primary exponent of the subjective theory of probability in the last century, and is to a large extent responsible for the resurgence of the Bayesian approach to statistics.

Bruno de Finetti was born in Innsbruck, Austria, of Italian parents on 13 June 1906, and died on 20 July 1985 in Rome. He enrolled in the Milan Polytechnic in 1923, and then after discovering that his interests lay more in science, switched to the Faculty of Mathematics of the University of Milan.

He first became interested in probability as a result of reading an article in a popular magazine concerning Mendel's laws of genetics, which stimulated him to write an article of his own on the subject. His article attracted the attention of Corrado Gini and eventually led to a position as head of the Mathematical Office of the Instituto Centrale di Statistica. Later he accepted a position as actuary in the Assicurazioni Generali insurance company in Trieste, where he was concerned with automation of all actuarial, statistical, and bookkeeping operations and calculations. He became a lecturer in mathematics, actuarial mathematics, and probability at the Universities of Trieste and Padova, and later full professor in Trieste and finally at the University of Rome, from which he retired in 1976.

In 1957 he spent a quarter at the University of Chicago at the invitation of Professor Leonard Jimmie Savage (1917–71). They became friends and each stimulated the other in the development of his ideas. Although there are differences in the presentations of subjective probability by de Finetti and Savage, their basic outlook was the same, and they stood in sharp contrast to other figures in the field, with the exception of Frank Ramsey (1903–30), whose important article *Truth and Probability* (1926) was very much in the same spirit. De Finetti and Savage were the major figures of the twentieth century in the area of subjective probability, and legitimate heirs of the two giants, Émile Borel (1871–1956) and Henri Poincaré (1854–1912), who also viewed probability theory as a serious subject with numerous important real-world applications, rather than as a mathematical exercise. Both de Finetti and Savage were infuriated by the trivializations of the subject as such an exercise, which dominated the field during most of their lifetimes.

Early in his career de Finetti realized that there was something seriously lacking in the way probability and statistics were being developed and taught. The famous remark 'Everybody speaks about Probability, but no one is able to explain clearly to others what meaning probability has according to his own conception,' by the mathematician Garrett Birkhoff (1940), was indicative of serious problems even with the definition of probability. De Finetti found many of the standard assertions confused and meaningless. In his University of Rome farewell address (1977), concerning the contrast between himself and Savage (a mathematician who had been trained in the standard non-Bayesian theory) he said: 'I, of course, was in quite the opposite situation: a layman or barbarian who had the impression that the others were talking nonsense but who virtually does not know where to begin as he cannot imagine the meaning that others are trying to give to the absurdities that it is imposing as dogma.'

De Finetti then attempted to formulate an appropriate meaning for a probability statement. His knowledge of philosophy and logic as well as mathematics, both pure and applied, enabled him to do so. The areas of his work that will be discussed here are his theory of exchangeability, theory of coherency, and theory of finitely additive probability. For him theory and practice were inseparable, and he emphasized the operational meaning of abstract theory, and particularly its role in making probabilistic predictions.

De Finetti rejected the concept of an objective meaning to probability and was convinced that probability evaluations are purely subjective. In his article *Probability and My Life* published in Gani (1982) he wrote: 'The only approach which, as I have always maintained and have confirmed by experience and comparison, leads to the removal of such ambiguities, is that of probability considered – always and exclusively – in its natural meaning of "degree of belief." This is precisely the notion held by every uncontaminated "man in the street." De Finetti then developed a theory of subjective probability, corresponding to its use as an expression of belief or opinion, and for which the primary feature is coherency.

The starting point for his theory is the collection of gambles that one takes on, whether knowingly or unknowingly. These can be either ordinary gambles or conditional gambles, which are called-off if the contingency being conditioned upon does not occur. The simplest type of gamble is one on an event E, in which one receives a monetary stake S if E occurs and nothing otherwise. If a person values such a gamble at the sure amount $p \times S$, then p is defined to be the probability of E for that person. In this evaluation one uses small stakes S so as to avoid utility complications, which distort probabilities evaluated in this way. One can then take linear combinations of such simple gambles to generate gambles in which there are a finite number of possible outcomes. Similarly, a called-off gamble on the event E given F is such that one receives a monetary stake S if both events occur, nothing if F occurs but not E, and the gamble is called off if F does not occur. The value of such a conditional

gamble for a person is called his or her conditional probability for E given F and is denoted by $P(E\backslash F)$.

In the work of de Finetti (1970, 1972, 1974–75) a primary aim is to avoid the possibility of a Dutch book, that is, to avoid sure loss on a finite collection of gambles. Thus suppose one has taken on a finite collection of gambles of any type whatsoever, including such things as investments, insurance, and so on. De Finetti proved that a necessary and sufficient condition for the avoidance of sure loss of a positive amount greater than some $\varepsilon > 0$ on a finite number of unconditional gambles, consists in consistency with the axioms of finitely additive probability theory. These axioms are the usual ones, apart from the fact that the condition of finite additivity replaces that of countable additivity in the third axiom. Let $P(E)$ denote the numerical probability for the event E. Then the axioms are:

1. $P(W) = 1$, where W is the sure event.
2. For any event E, $0 \leq P(E) \leq 1$.
3. The probability of a finite union of disjoint events is the sum of the probabilities.

Further, if conditional gambles are also included, then to avoid a Dutch book, conditional probabilities must be evaluated in accord with the usual definition:

$$P\{A|B\} = P\{A \textit{ and } B\}/P\{B\},$$

provided that $P\{B\} > 0$. See Hill (1998) for further discussion.

In the subjectivistic conception a probability is essentially a price that one will pay for a certain gamble, with the understanding that this price can and will differ from one person to another, even apart from utility considerations. This conception is very close to that of the common person who speaks of probabilities for various real-world events based upon his or her own experience and judgment. In this case there is no need to confine attention to trials such as arise in games of chance. On the other hand, it is extremely interesting to see what the subjectivist interpretation of such trials might be. By considering this question, de Finetti developed his celebrated exchangeability theorem.

The problem that de Finetti posed and largely solved was that of justifying from a subjective point of view the importance of statistical frequencies in the evaluation of probabilities. Suppose that we have a sequence of events, E_i, for $i = 1, \ldots, N$, each of which may or may not occur, for example where E_i is the result of the ith flip of a coin. For convenience, we can assign the value 1 to heads and 0 to tails. It is natural sometimes to view the order of the flips as irrelevant, so that any permutation of the same sequence of outcomes is given the same probability. This leads to the notion of exchangeability, which is a symmetry condition, specifying that joint distributions of the observable random quantities are invariant under permutations. De Finetti proved a representation theorem which implies that if one regards a sufficiently long sequence of successes and failures as exchangeable, then it necessarily follows that one's opinions can be represented, approximately, by a mixture of Bernoulli sequences, that is, of sequences for which there is a 'true' (but typically unknown) probability P of success, and such that conditional upon $P = p$, the observations are independent with probability p of success on each toss. The proportion of successes, as the number of trials goes to ∞, converges almost surely to P, and the initial or *a priori* distribution for P is also called the de Finetti measure for the process.

This representation in terms of Bernoulli sequences is exact for infinite sequences, and a slight modification holds also for finite sequences, the difference amounting to that between sampling with and without replacement from an urn. It also follows mathematically that provided a sufficiently long sequence has been observed, to a good approximation one will be led to base opinions about a future trial on the past empirical frequency with which the event has occurred, and this constitutes de Finetti's justification for induction. In his work de Finetti emphasized prediction of observable quantities, rather than what is usually called inference about parameters. In other words, he was primarily interested in the posterior distribution of future observable random variables, rather than in the posterior distribution for the parameter P. This point of view has become known as predictivism in modern Bayesian statistics.

The exchangeability theorem of de Finetti allows one to analyze the relationship between subjective probabilities and the frequency theory of probability that developed from the work of Jakob Bernoulli. The fundamental stochastic process is known as a Bernoulli sequence. Such a sequence is by assumption such that on each trial there is a fixed objective probability p for success (1) and $1 - p$ for failure (0), whose value may or may not be known, with the trials independent in the probabilistic sense. It is then a mathematical fact, known as the law of large numbers, that as the number of trials grows large the proportion of successes will stabilize at the 'true' probability p in the sense that the proportion of successes will converge almost surely to p; that is, the sequence converges except for a collection of outcomes that has probability 0.

Critics of this theory, such as de Finetti, argue that there are no real-world applications in which either p or the limiting proportion of successes is known to exist. The empirical support for the objectivist theory is merely that there are certain processes, such as coin flipping, dice throwing, card playing, draws from an urn with replacement, and so on, for which in finite sequences of trials the empirical proportions of successes appear to be in conformity with what the objectivist theory suggests. For example, if there are two similar balls in an urn, one red and one white, and if draws are made with replacement after substantial shaking of the urn, then experience suggests that the proportion of red balls in n draws, when n is sufficiently large, is around .5. Of course it is impossible to perform an infinite number of trials, so the limiting proportion, if it exists at all, can never be known. But in finitely many trials not only does the proportion of red balls drawn often seem to settle down to something like .5, but in addition, if say the n trials are repeated I times, then the proportion of these I trials in which there are exactly j red balls in n trials appears also to settle down to $\binom{n}{j}(1/2^n)$, more or less as the theory suggests. One must, however, exclude situations where such settling down does not appear to happen, and these are usually interpreted as meaning that the shaking was not adequate, or the cards not sufficiently well shuffled, and so on. Nonetheless, it

appears that when sufficient care is taken, then the proportions appear to be roughly in accord with the objectivistic theory.

Does this imply that there exists a true $p = \frac{1}{2}$ which is the objective probability, for example, that the ball drawn is red, and hence also that the proportion of red balls drawn must converge almost surely to .5? No one knows to this day. However, de Finetti's exchangeability theorem provides an alternative way to describe our experiences with such random events. A probability as before is merely a price that one would pay for a certain gamble. Under the assumption of an equal number of red and white balls, and that the urn is well shaken, symmetry suggests a price of .5 for the gamble in which you receive a dollar if a red ball is drawn, nothing otherwise. Similarly, if you were to be offered a gamble in which you would receive the dollar if the first two draws were red, then again symmetry together with the assumption that the urn is well shaken, suggest that a fair price would be .25. Continuing in this way we see that your prices are in accord with the usual probability evaluations for such an urn.

It is then a logical consequence of the exchangeability theorem of de Finetti, that you must evaluate your probability that the sample proportion of red balls converges to .5 as unity. Note that this requires no circular definitions or mysterious conventions. There are, however, two important caveats. First, it is necessary that utility considerations be ignored, which necessitates that the prizes are not too large, and for this reason the stake S was taken to be one dollar. Alternatively, utility can be dealt with after the fashion of Ramsey or Savage. Second, the cases where people take this evaluation most seriously are those in which there have already been many draws in the past, indicating no cheating or violation of the apparent symmetry, so that in effect your probabilities are posterior probabilities based upon Bayes' theorem. In other words: just as dice may be loaded so can other games be fixed, and a prudent bettor would not make the above evaluations merely from apparent physical and time symmetry alone, but rather in addition from experience that the game is not rigged and that the perception of symmetry is not mistaken. At any rate, one sees that it is not necessary to introduce the concept of an objective probability, but merely the experience that suggests an evaluation of .5.

The case of a Bernoulli sequence, whether interpreted according to the objectivist or subjectivist theory, is of course a very special example of an exchangeable sequence of random variables. For a Bernoulli sequence with $p = .5$, that is, a degenerate *a priori* distribution for P concentrated at the value .5, the primary difference between the two theories is that the objectivist theory objectifies the .5 which in the subjectivist theory is merely the opinion of the bettor, and similarly the .25 for drawing two red balls, and so on. In circumstances where prudent bettors have learned from experience to bet in accord with symmetry, the objectivist school states that the .5 and the .25 and so on are the objective or true probabilities, not merely the values that the bettor chooses. Thus the objectivist theory implies that the proportion of successes necessarily converges to .5, while the subjective theory implies that the bettor must believe, with probability 1, that such convergence takes place.

It should be noted that countable additivity, that is, the modification of the third axiom to include also denumerable unions of disjoint events (which corresponds to a denumerable number of gambles) has never been demonstrated within the rigorous approaches to probability theory and, as argued by de Finetti, is merely a particular continuity condition that one may or may not wish to adopt in a particular context, just as with the use of continuous functions in physics. The relaxation of the condition of countable additivity to finite additivity allows one greater freedom in the representation of degrees of belief. For example, the use of adherent mass distributions (which were introduced by de Finetti and are closely related to Dirac functions) allows one to obtain solutions for Bayesian non-parametric inference and prediction. See Hill (1993).

The work of de Finetti played a major role in the resurrection of the classical Bayesian approach to probability, statistics, and decision theory, as founded by Bayes and Laplace and developed by Gauss, Poincaré, and Borel. His work continued the theory of induction of the British empiricists such as David Hume, and provided mathematically sound foundations for the modern-day subjective Bayesian approach.

See also: Probability Theory: Interpretations; Probability and Chance: Philosophical Aspects; Statistics, History of.

Bibliography

Birkhoff, C., 1940. Lattice Theory. American Mathematical Society Publication no. 25.

De Finetti, B., 1970. Teoria delle probabilitá. Einaudi. Turin, Italy.

De Finetti, B., 1972. Probability, Induction and Statistics. Wiley, London.

De Finetti, B., 1974–75. Theory of Probability, vols. 1 and 2. Wiley, London.

De Finetti, B., 1977. Probability: beware of falsifications. In: Aykac, A., Brumat, C. (Eds.), New Developments in the Applications of Bayesian Methods. North-Holland, Amsterdam, pp. 347–379.

Gani, J. (Ed.), 1982. The Making of Statisticians. Springer-Verlag, New York.

Hill, B.M., 1993. Parametric models for A_n: Splitting processes and mixtures. Journal of the Royal Statistical Society, Series B 55, 423–433.

Hill, B.M., 1998. Conditional Probability. Encyclopedia of Statistics. Wiley, New York.

Kolmogorov, A., 1950. Foundations of the Theory of Probability. English translation. Chelsea Press, New York.

Death and Dying, Psychology of

Mary A Varga, University of West Georgia, Carrollton, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Kruse, volume 5, pp. 3263–3267, © 2001, Elsevier Ltd.

Abstract

Death is an inevitable part of life and how an individual approaches the death and dying process is a phenomenon that continues to be examined. Ultimately, how an individual reacts to death and dying is dependent on a variety of factors, including death attitudes, developmental statuses, and cultural considerations.

Death is an inevitable part of life and how an individual approaches the death and dying process is a phenomenon that continues to be examined. Ultimately, how an individual reacts to death and dying is dependent on a variety of factors, including death attitudes, developmental statuses, and cultural considerations. Using Erikson's (1963) psychosocial stages as a framework can help understand how these factors affect individuals coping with loss or their own impending death throughout the various stages of life.

Overview of Death and Dying Perspectives

There is an extensive history associated with the desire to understand death and dying. Research on attitudes around death began in the 1950s when Feifel (1959) laid the foundation for investigating death attitudes and anxiety and establishing the field now known as thanatology, the study of death and dying. Glasser and Strauss (1965) were also early voices in death research, discussing the influence of awareness with dying individuals. Research has continued and expanded, examining various age ranges of individuals and cultures, providing a more comprehensive understanding of death attitudes. Scales measuring death anxiety (Schell and Seefeldt, 1991; Thorson and Powell, 1992), fear of death (Lester, 1990), grief reactions (Hogan et al., 2001), and many others have been developed to provide more scientific understandings of grief and death. Theories centered on coping with grief, such as the five stages of grief and the dual process model of coping with bereavement, have been created to help understand grief trajectories and provide practitioners with resources to support bereaved or dying individuals (Kubler-Ross, 1969; Stroebe and Schut, 1999).

Important consideration in understanding death attitudes and development are psychosocial stages. Psychologist Erik Erikson (1968) categorized the various stages of life according to psychosocial development (*see* Theories of Human Development). At the center of this psychosocial stage theory are ego identity and competences. Ego identity is shaped by social interactions that help develop attitudes, beliefs, and values. Competences refer to individual tasks throughout the stages of life that impact ego identity development. When approaching death, death attitudes and personal wellness are explicitly linked to specific psychosocial stages throughout life.

Death Attitudes

Death attitudes are often defined as the fear and anxiety related to end of life and dying (Neimeyer, 1994; Neimeyer and Hogan, 2001). They are influenced by a number of factors in an individual's life, including age, gender, race/ethnicity, cultural norms, physical and psychological health, religious/spiritual beliefs, previous experience with death, and life satisfaction (*see* Death and Dying, Sociology of). These factors, combined with an individual's age and developmental stage from childhood to adulthood, significantly impact their attitudes and views toward death and dying.

Childhood and Death

Early research on childhood death attitudes indicated three distinct developmental stages: nonfunctionality, irreversibility, and universality (Nagy, 1948). The first stage of nonfunctionality refers to the understanding that specific living functions, such as breathing, end at death. The second stage of irreversibility focuses on the understanding that a human body cannot physically return from death after dying. The third and final stage of universality refers to the understanding that living things will eventually die.

Children specifically under the age of 5 may perceive death as reversible, or the result of magical or medical mediation (Speece and Brent, 1992). Children aged 5–9 often personify death. They also view it as escapable instead of inevitable. Most children aged 9 and above recognize that death is universal, irreversible, and that human functioning ceases; however, research also shows children develop mature understandings of death as early as age 6 (Hunter and Smith, 2008; Mahon et al., 1999; Nagy, 1948). Overall, female children express more resistance to and have negative emotions about death, male children are more likely to accept and confront death (Yang and Chen, 2009) (*see* Infancy and Childhood: Emotional Development).

The attitudes children have toward death vary among gender and are largely influenced by their cognitive development, religions, socialization, and prior experience with death and dying (Mahon et al., 1999; Morgan and Laungani, 2002; Silverman, 2000; Yang and Chen, 2009). An important consideration in understanding death attitudes in children is development and psychosocial stages. Erikson (1963, 1968) characterizes each of the four stages in childhood with

the developmental tasks of trust, autonomy, initiative, and industry. Confrontation with death or dying can significantly impact the developmental fulfillment of these tasks. For example, the developmental task of infants involves the cultivation of trust. This allows children to develop a sense of hope and the ability to rely on others to fulfill their needs. Terminally ill infants with changing care providers and hospital settings can cause disruption to fulfilling this developmental task (Doka, 1996). The psychosocial stage shifts to independence and self-control in later childhood (Erikson, 1963). This can be stressful for young terminal toddlers suffering physical limitations due to terminal illness (Doka, 1996).

During early childhood, children begin balancing their personal desires with their moral responsibility. Children in later childhood are concerned with productivity and mastering skills (Erikson, 1963). These tasks become troublesome for terminal children who are limited by physical barriers, parental apprehension, and safety concerns, thus affecting self-esteem and achieving developmental competencies (Doka, 1996).

Adolescents and Death

Adolescent attitudes toward death are also explicitly linked to developmental issues (Noppe and Noppe, 1996). The concepts of death for younger teens involve parental separation, and concepts for mid-adolescences focus on death impact. Legacy is at the forefront for college-age youth and personal identity is an important development in this psychosocial stage (Erikson, 1963). Struggles with identity development can impede a logical understanding of death and increase death anxiety (Noppe and Noppe, 2004). The highest levels of death anxiety among teens occur among those who possess 'less stable ego pictures' at the peak of identity formation (Alexander and Alderstein, 1958: p. 175).

Identity formation is significantly impacted for terminally ill adolescences (*see* Identity in Childhood and Adolescence). Physical changes associated with puberty may be delayed or replaced by side effects of medical treatment, including hair loss, acne, weight fluctuations, or disfigurements (Freyer, 2004). The need for physical and emotional support and dependence can initiate struggles for independence and affect psychosocial development (Nannis et al., 1978; Freyer, 2004). Isolation can occur affecting identity formation and peer socialization (Freyer, 2004).

Adulthood and Death

Similar to earlier life stages, adulthood attitudes toward death are also influenced by gender, ethnicity, physical and psychological health, religious beliefs, environmental factors, generational differences, experience with death and aging, life satisfaction, and impact death attitudes (DePaola et al., 2003; Fortner and Neimeyer, 1999; Walker and Maiden, 1987). Death anxiety typically decreases as age increases; however, death anxiety peaks for women in the mid-20s age range and then again in the early 50s (Russac et al., 2007). Older women also tend to report more death anxiety than men (Azaiza et al., 2010; Besser and Priel, 2007; Cicirelli, 2001; DePaola et al., 2003; Suhail and Akram, 2002).

Death attitudes also vary among cultures and environment (*see* Health and Illness: Mental Representations in Different Cultures). Older Caucasian adults report fears pertaining to the dying process, while African-Americans experience fears of what proceeds death, such as fear of the unknown, consciousness when dead, and the status of the body after death (DePaola et al., 2003). Older adults with less religiosity exhibit greater anxiety about various dimensions of death (Cicirelli, 2002; Suhail and Akram, 2002; Wu et al., 2002). Older adults living in institutions show more fear of death when compared to older adults living independently (Azaiza et al., 2010; Fortner and Neimeyer, 1999). Older adults with physical and psychological ailments or concerns also experience greater death anxiety (Moreno and Moore, 2008; Wu et al., 2002). Finally, hospice patients who have more experience with death experience lower levels of death anxiety and avoidance (Bluck et al., 2008).

When facing death, the responsibilities and competencies that young adults practice and the intimacy they desire are strained (Cook and Oltjenbruns, 1998; Erikson, 1963). Young adults have established their identity and commitments to friendships, marriage, parenthood, and careers. When impending death or a terminal illness intrudes upon this stage of life, young adults can be resistant of their prognoses (Pattison, 1977). Terminal illness can cause young adults to experience fear about the dying process and alienate their partners, threatening important relationships.

While relationships remain important in the later stage of adulthood, the psychosocial stage of older adults centers on achievements and generativity (Butler and Lewis, 1982; Erikson, 1963; Pattison, 1977). Impending death triggers individuals to reflect on their life accomplishments (Cook and Oltjenbruns, 1998). Generativity is a term that Erikson (1968) refers to when an individual becomes focused on legacy, such as parenting and future generations.

Finally, the concerns among the elderly when facing death center on a life review, which are closely linked to ego identity (Butler, 1963; Erikson, 1963). Life reviews provide the dying elderly with feelings of wisdom, acceptance, fulfillment, and contentment, thus maintaining their ego before death. Unfortunately life reviews that are not validating can generate feelings of hopelessness, disappointment, and an unfulfilled life (Erikson, 1982).

Cultural Considerations

While Erikson's (1963) psychosocial stages provide a solid framework for using attitudes around death, specific cultural factors should also be considered (Parkes et al., 1997). For example, Chinese culture has significantly different views on death and dying, stemming from their beliefs rooted in Buddhism, Confucianism, Taoism, and traditional Chinese medicine. Even speaking of issues around death and dying is believed to bring back luck, thus creating a culture where talking about death is taboo (Hsu et al., 2009).

In the Islam culture, vast differences exist between men and women which affect grief practices. Kobeisy (2004) notes that "the primary role of women has become one of wife and mother within the extended family in a patriarchal society with little concern or awareness of life outside her home" (p. 46). Focusing

specifically on grief and loss, expectations for men are to control their emotions, maintain composure, and express grief quietly while women are allowed to express more emotion, but are not allowed to attend funerals (Rubin and Yasien-Esmael, 2004).

Coping with Death and Dying

One of the most widely known theories surrounding death and dying is the stages of grief model presented by Kubler-Ross (1969). Although this model has not been proven through empirical research, it has been used by bereaved individuals and clinicians to help understand grief and aid with coping. The fallacy with the stages of grief is not only due to the lack of scientific evidence to prove it, but its impractical nature as well. Human beings grieve in individual ways that cannot be confined and defined by five stages. Furthermore, circumstances surrounding death events, such as cause of death, relationship to the deceased, support systems, and cultural considerations, also affect the trajectory of bereavement and preparation for death and are not addressed within the stages (Goldsmith et al., 2008; Wen et al., 2011). Kubler-Ross conducted her research only with patients who were dying, not those who were grieving a loss, therefore the major contributions of her work were opening the conversation about grief and loss and providing a more in depth understanding of the dying process.

Another established model is the dual process of coping with bereavement model (Stroebe and Schut, 1999). Instead of taking a stage approach to the coping process, the dual process model divides bereavement into loss and restoration. Loss includes engaging in grief work including feelings of both acceptance and denial. Restoration involves the simultaneous process of the bereaved learning how to live a new way of life without the deceased, thus restoring their lives. While this model provides a new perspective on coping with loss, the limitation is that it is restricted to individuals coping with a loss of another individual and does not include the dying self.

A more recently developed model is the Holistic Impact of Bereavement (Balk, 2011). This framework was developed to illustrate the multidimensional effects of grief on grieving individuals. Through several years of research with college-aged students, Balk found that grief affects the bereaved physically, cognitively, behaviorally, interpersonally, emotionally, and spiritually and can manifest in various ways. Although they do not refer to the Holistic Impact of Bereavement specifically in their research, numerous research studies have similar findings (Neimeyer et al., 2008; Servaty-Seib and Hamilton, 2006; Walker et al., 2012).

Additional bereavement theories and models, such as the Grief to Personal Growth Model (Hogan and Schmidt, 2002), ATTEND (Attunement (A), Trust (T), Touch (T), Egalitarianism (E), Nuance (N), and Death (D)) (Cacciatore and Flint, 2012), and Meaning Reconstruction Theory (Neimeyer and Anderson, 2002) have also been developed to help dying or bereaved individuals. Each of these approaches provides an important and unique understanding of the coping process. Similar to understanding death attitudes, it is also important to understand developmental and cultural considerations when examining such models.

Conclusions

Death and dying are inevitable experiences in life. They are also complex processes. How a person perceives these phenomena is directly impacted by their stage of life and level of development – and development is significantly impacted by dying as well. Additional factors, such as gender, culture, social support, and others also inform attitudes about death and dying. Therefore combining all these characteristics makes the death and dying process an individualized one that elicits individualized perspectives and experiences.

Further Research

The knowledge gathered by thanatologists, psychologists, medical researchers, and other interested parties has provided a breadth of information on the death, dying, and grieving processes with more yet to be learned. As the world becomes more technologically advanced and ever-changing, research must follow the change and growth. Death and dying research is no exception.

As the number of people turning to the Internet for health-related information and support increases, more attention has been given to how dying and bereaved individuals use the Internet for support (Sofka et al., 2012; Walter et al., 2012). Recent studies have examined the use of support groups to provide grief support, Web memorialization of deceased loved ones; and how social networking impacts the grief process (Dominick et al., 2009; Hieftje, 2012; Roberts, 2004; Tolstikova and Chartier, 2009; Varga and Paulus, 2013). Continued attention must be given to these expanding uses of support to further understand the human response to death and dying.

As the world experiences more contemporary disasters, it is important to understand how these events affect psychology associated with death and dying. New studies have shown that these contemporary disasters are affecting children and instilling new fear (Burnham, 2009). Continued research on other individuals and those in other areas of the world experiencing disaster should be closely examined.

Finally, another area in need of continued research is physician assisted suicide (*see* Suicide: Ethical Aspects). As individuals, particularly in the United States, seek to make their own end-of-life decisions, it is important to understand the outcomes and ramifications of these procedures on individuals, families, and communities (Cerminara and Perez, 2000). As medical technology advances, individuals with decreased quality of life may increasingly turn to this form of accelerated death (Sears and Stanton, 2001). This desired control over the end-of-life process will certainly provide new perspectives on the psychology of death and dying.

See also: Aging and Health in Old Age; Childhood and Adolescence: Developmental Assets; Cognitive Aging; Death and Dying, Sociology of; Early Life Influences on Health and Mortality in Adulthood; Ecology of Aging; Health and Illness: Mental Representations in Different Cultures; Human Development, Theories of; Identity in Childhood and Adolescence; Infancy and Childhood: Emotional Development; Suicide.

Bibliography

Alexander, I.E., Alderstein, A.M., 1958. Affective responses to the concept of death in a population of children and early adolescents. Journal of Genetic Psychology 93, 167–177.

Azaiza, F., Ron, P., Shoman, M., Gigini, I., 2010. Death and dying anxiety among elderly Arab Muslims in Israel. Death Studies 34 (4), 351–364. http://dx.doi.org/10.1080/07481181003613941.

Balk, D.E., 2011. Helping the Bereaved College Student. Springer Publishing Company, New York.

Besser, A., Priel, B., 2007. Attachment, depression, and fear of death in older adults: the roles of neediness and perceived availability of social support. Personality and Individual Differences 44, 1711–1725.

Bluck, S., Dirk, J., Mackay, M.M., Hux, A., 2008. Life experience with death: relation to death attitudes and to the use of death-related memories. Death Studies 32 (6), 524–529. http://dx.doi.org/10.1080/07481180802138860.

Burnham, J.J., 2009. Contemporary fears of children and adolescents: coping and resiliency in the 21st century. Journal of Counseling and Development 87, 28–35.

Butler, R.N., 1963. The life review: an interpretation of reminiscence in the aged. Psychiatry 26, 65–70.

Butler, R., Lewis, M., 1982. Aging and Mental Health. Mosby, St. Louis, MO.

Cacciatore, J., Flint, M., 2012. ATTEND: toward a mindfulness-based bereavement care model. Death Studies 36, 61–82. http://dx.doi.org/10.1080/07481187.2011.591275.

Cerminara, K.L., Perez, A., 2000. Therapeutic death: a look at Oregon's law. Psychology, Public Policy, and Law 6, 503–525. http://dx.doi.org/10.1037/1076-8971.6.2.503.

Cicirelli, V.G., 2001. Personal meanings of death in older adults and young adults in relation to their fears of death. Death Studies 25 (8), 358–366. http://dx.doi.org/10.1080/713769896.

Cicirelli, V.G., 2002. Fear of death in older adults: predictions from terror management theory. Journal of Gerontology 57, 358–366. http://dx.doi.org/10.1093/geronb/57.4.P358.

Cook, A.S., Oltjenbruns, K.A., 1998. Dying and Grieving: Life Span and Family Perspectives, second ed. Harcourt, New York.

DePaola, S.J., Griffin, M., Young, J.R., Neimeyer, R.A., 2003. Death anxiety and attitudes toward the elderly among older adults: the role of gender and ethnicity. Death Studies 27 (2), 335–354. http://dx.doi.org/10.1080/07481180302904.

Doka, K.J., 1996. The cruel paradox: children who are living with life-threatening illnesses. In: Corr, C.A., Corr, D.M. (Eds.), Handbook of Childhood Death and Bereavement. Springer, New York.

Dominick, S.A., Irvine, A.B., Beauchamp, N., Seeley, J.R., Nolen-Hoeksema, S., Doka, K.J., Bonanno, G.A., 2009. An internet tool to normalize grief. OMEGA Journal of Death and Dying 60 (1), 71–87. http://dx.doi.org/10.2190/OM.60.1.d.

Erikson, E., 1963. Childhood and Society, second ed. W. W. Norton & Co, New York.

Erikson, E.H., 1968. Identity: Youth and Crisis. Norton, New York.

Erikson, E.H., 1982. The Life Cycle Completed: A Review. W.W. Norton, New York.

Feifel, H., 1959. The Meaning of Death. McGraw-Hill, New York.

Fortner, B.V., Neimeyer, R.A., 1999. Death anxiety in older adults: a quantitative review. Death Studies 23, 387–411.

Freyer, D.R., 2004. Care of the dying adolescent: special considerations. Pediatrics 113 (2), 381–388.

Glasser, B., Strauss, A., 1965. Awareness of Dying. Aldine, Chicago, IL.

Goldsmith, B., Morrison, R.S., Vanderwerker, L.C., Prigerson, H.G., 2008. Elevated rates of prolonged grief disorder in African Americans. Death Studies 32 (4), 352–365. http://dx.doi.org/10.1080/07481180801929012.

Hieftje, K., 2012. The role of social networking sites in memorialization of college students. In: Sofka, C.J., Cupit, I.N., Gilbert, K.R. (Eds.), Dying, Death, and Grief in an Online Universe. Springer Publishing Company, New York, pp. 31–46.

Hogan, N.S., Greenfield, D.B., Schmidt, L.A., 2001. Development and validation of the Hogan grief reaction checklist. Death Studies 25 (1), 1–32. http://dx.doi.org/10.1080/07481180125831.

Hogan, N.S., Schmidt, L.A., 2002. Testing the grief to personal growth model using structural equation modeling. Death Studies 26 (8), 615–634. http://dx.doi.org/10.1080/07481180290088338.

Hunter, S.B., Smith, D.E., 2008. Predictors of children's understandings of death: age, cognitive ability, death experience and maternal communicative competence. OMEGA: Journal of Death and Dying 57 (2), 143–162. http://dx.doi.org/10.2190/OM.57.2.b.

Hsu, C., O'Connor, M., Lee, S., 2009. Understandings of death and dying for people of Chinese origin. Death Studies 33 (2), 153–174. http://dx.doi.org/10.1080/07481180802440431.

Kobeisy, A.N., 2004. Counseling American Muslims: Understanding the Faith and Helping the; People. Praeger, Westport, CT.

Kubler-Ross, E., 1969. On Death and Dying. Macmillan, New York.

Lester, D., 1990. The collett-lester fear of death scale: the original version and a revision. Death Studies 14 (5), 451–468. http://dx.doi.org/10.1080/07481189008252385.

Mahon, M.M., Goldberg, R.L., Washington, S.K., 1999. Discussing death in the classroom: beliefs and experiences of educators and education students. OMEGA: Journal of Death and Dying 39 (2), 9–121.

Moreno, R.A., Moore, M.K., 2008. Validity and reliability of the multidimensional fear of death scale. In: Neimeyer, R. (Ed.), Death Anxiety Handbook: Research, Instrumentation, and Application. Taylor & Francis, Washington, DC, pp. 3–28.

Morgan, J.D., Laungani, P., 2002. Death and Bereavement Around the World: Major Religious Traditions (Death, Value and Meaning). Baywood Publishing Company, Inc, Amityville, NY.

Nagy, M.A., 1948. The child's theories concerning death. Journal of Genetic Psychology 73, 3–27.

Nannis, E.D., Susman, E.J., Strope, B.E., 1978. The Adolescent with a Life-Threatening Illness: Cultural Myths and Social Realities. Paper presented at the 13th Annual Conference of the Association of the Care of Children in Hospital. Washington, DC.

Neimeyer, R.A., 1994. The threat index and related methods. In: Neimeyer, R.A. (Ed.), Death Anxiety Handbook: Research, Instrumentation, and Application. Taylor & Francis, Philadelphia, PA, pp. 61–101.

Neimeyer, R.A., Anderson, A., 2002. Meaning reconstruction theory. In: Thompson, N. (Ed.), Loss and Grief: A Guide for Human Service Practitioners. Palgrave Macmillan, New York.

Neimeyer, R.A., Hogan, N.S., 2001. Quantitative or qualitative? Measurement issues in the study of grief. In: Stroebe, M.S., Hansson, R.O., Stroebe, W., Schut, H. (Eds.), Handbook of Bereavement Research: Consequences, Coping, and Care. American Psychological Association, Washington, DC, pp. 89–118.

Neimeyer, R.A., Laurie, A., Mehta, T., Hardison, H., Currier, J.M., 2008. Lessons of love: meaning-making in bereaved college students. New Directions for Student Services 121, 27–39.

Noppe, L.D., Noppe, I.C., 1996. Ambiguity in adolescent understandings of death. In: Corr, C.A., Balk, D.E. (Eds.), Handbook of Adolescent Death and Bereavement. Springer Publishing Company, New York, pp. 25–41.

Noppe, I.L., Noppe, L.D., 2004. Adolescent experiences with death: letting go of immortality. Journal of Mental Health Counseling 26 (2), 146–167.

Parkes, C.M., Laungani, P., Young, W., 1997. Death and Bereavement across Cultures. Routledge, London.

Pattison, E.M., 1977. The Experience of Dying. Simon and Schuster, New York.

Roberts, P., 2004. The living and the dead: community in the virtual cemetery. Omega Journal of Death and Dying 49 (1), 57–76. http://dx.doi.org/10.2190/D41T-YFNN-109K-WR4C.

Rubin, S.S., Yasien-Esmael, H., 2004. Loss and bereavement among Israel's Muslims: acceptance of god's will, grief, and the relationship to the deceased. Omega: Journal of Death and Dying 49 (2), 149–162.

Russac, R.J., Gatliff, C., Reece, M., Spottswood, D., 2007. Death anxiety across the adult years: an examination of age and gender effects. Death Studies 31 (6), 549–561. http://dx.doi.org/10.1080/07481180701356936.

Schell, D., Seefeldt, C., 1991. Development of a death anxiety scale for children. OMEGA: Journal of Death and Dying 23 (3), 227–234. http://dx.doi.org/10.2190/NYJ6-5Q3Q-MH4M-7787.

Sears, S.R., Stanton, A.L., 2001. Physician-assisted dying: review of issues and roles for health psychologists. Health Psychology 20, 302–310. http://dx.doi.org/10.1037/0278-6133.20.4.302.

Servaty-Seib, H.L., Hamilton, L.A., 2006. Educational performance and persistence of bereaved college students. Educational performance and persistence of bereaved college students. Journal of College Student Development 47 (2), 225–234. http://dx.doi.org/10.1353/csd.2006.0024.

Silverman, P.R., 2000. Never Too Young to Know: Death in Children's Lives. Oxford, New York.

Sofka, C.J., Culpit, I.N., Gilbert, K.R., 2012. Dying, Death, and Grief in an Online Universe: For Counselors and Educators. Springer Publishing Company, New York.

Speece, M.W., Brent, S.B., 1992. The acquisition of a mature understanding of three components of a death concept. Child Development 55, 1671–1686. http://dx.doi.org/10.1080/07481189208252571.

Suhail, K., Akram, S., 2002. Correlates of death anxiety in Pakistan. Death Studies 26 (1), 39–50. http://dx.doi.org/10.1080/07481180210146.

Stroebe, M., Schut, H., 1999. The dual process model of coping with bereavement: rationale and description. Death Studies 23 (3), 197–224. http://dx.doi.org/10.1080/074811899201046.

Thorson, J.A., Powell, F.C., 1992. A revised death anxiety scale. Death Studies 16 (6), 507–521. http://dx.doi.org/10.1080/07481189208252595.

Tolstikova, K., Chartier, B., 2009. Internet method in bereavement research: comparison of online and offline surveys. OMEGA Journal of Death and Dying 60 (4), 327–349. http://dx.doi.org/10.2190/OM.60.4.b.

Varga, M.A., Paulus, T.M., 2013. Grieving online: newcomers' constructions of grief in an online support group. Death Studies 0, 1–7. http://dx.doi.org/10.1080/07481187.2013.780112.

Walker, A.C., Hathcoat, J.D., Noppe, I.C., 2012. College student bereavement experience in a Christian university. OMEGA Journal of Death and Dying 64 (3), 241–259. http://dx.doi.org/10.2190/OM.64.3.d.

Walker, G., Maiden, R., 1987. Lifespan Attitudes toward Death. Paper presented at the Annual Meeting of the Eastern Psychological Association, Arlington, VA.

Walter, T., Hourizi, R., Moncur, W., Pitsillides, S., 2012. Does the internet change how we die and mourn? Overview and analysis. Omega Journal of Death and Dying 64 (4), 275–302. http://dx.doi.org/10.2190/OM.64.4.a.

Wen, K., McTavish, F., Kreps, G., Wise, M., Gustafson, D., 2011. From diagnosis to death: a case study of coping with breast cancer as seen through online discussion groups. Journal of Computer-Medicated Communication 16 (2), 331–361. http://dx.doi.org/10.1111/j.1083-6101.2011.01542.x.

Wu, A.M.S., Tang, C.,S.,K., Kwok, T.C.Y., 2002. Death anxiety among Chinese elderly people in Hong Kong. Journal of Aging and Health 14, 42–56.

Yang, S.C., Chen, Shih-Fen, 2009. The study of personal constructs of death and fear of death among Taiwanese adolescents. Death Studies 33 (10), 913–940. http://dx.doi.org/10.1080/07481180903251687.

Death and Dying, Sociology of

Tony Walter, University of Bath, Bath, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by H. Abramovitch, volume 5, pp. 3267–3270, © 2001, Elsevier Ltd.

Abstract

Social processes that influence contemporary dying and grieving include secularization, medicalization, professionalization, and sequestration. Sociological research on awareness and trajectories of dying has influenced health care systems toward more open communication styles at the end of life, intended in part to delay the onset of social death to as near as possible to physical death. The recent shift from funerals displaying social status to displaying the deceased's individuality may be understood in terms of postmaterialism. Sociological studies of bereavement are relatively underdeveloped, though the concept of disenfranchised grief has been influential. Recent research links the politics of memorialization to sociological theorizing of collective memory, and examines how the Internet socially relocates the dying, the grieving, and the dead.

Whereas demography measures and seeks causal explanations for variations in mortality rates, the sociology of death and dying is concerned with the role of mortality in social life. Many social thinkers believe that death poses the fundamental threat to the integrity and continuity of society. Social structures must provide a 'symbolic canopy' that guards against the terrors of extinction by providing meaning to human activity (Berger, 1969). Societies respond to this threat in a wide variety of ways ranging from the collectivist, death-accepting, rebirth orientation of much of the East, to the personal, death-defying, hedonism of much of the West. The sociology of death and dying, therefore, provides a unique perspective to investigate the core values of a culture. The discipline has been greatly influenced in content as well as methodology by psychiatry, anthropology, and social history. A main focus has been on changes in experience, attitudes, and roles of the dying, the bereaved, and their professional helpers.

A Broken Connection?

Does society comprise only the living, or does it also include the dead? It has often been asserted that traditional societies, especially those that venerate ancestors, comprise both the living and the dead, while modern Western societies have cut themselves off from the dead. This, however, is debatable. Almost every new communication technology, from writing to the Internet, has expanded the possibilities not only for the living to communicate with the living, but also about or even with the dead. *How* the dead are present in society, and what social institutions they legitimate, is subject to change. In many tribal societies, oral storytelling about the ancestors serves to construct and legitimate membership in an extended family. In the modern world, printing, mass literacy, and universal education (including the teaching of history) have enabled the exemplary life and often heroic death of national heroes to be used to construct national identity (Marvin and Ingle, 1999). Through photography and sound recording, deceased celebrities can continue as economic actors, often grossing more after than before death (Kearl, 2010). The Internet provides intriguing new possibilities for the living to interact with the dead (Walter et al., 2011–2012).

If the dead, or certain of the dead, continue among us as political or economic actors, what about the more everyday dead with whom we were intimate in life? Despite many bereaved people experiencing various kinds of continuing bond with the dead (Klass et al., 1996), Secularism and Protestantism have no ready language with which to describe this, acknowledging that the dead may continue on earth in memory or possibly in heaven as a soul, but cut off from the living and certainly not social agents. Where and what the dead are envisaged to be depends on culture and religion: "In the West, post-death conceptions typically involve the integrity and continuity of one's personal self. In the East, the ultimate goal is often an undifferentiated and impersonal oneness with the universe" (Kearl, 1989: p. 27). Literal and symbolic responses to questions about the dead, therefore, provide key insights into culture.

Concerning personal experience of dying or grief, traditional societies are characterized by widely shared and explicit norms of social behavior in general, and of mortuary ritual in particular. People knew how to behave when faced with death and dying. In many modern societies, individuals no longer have such clear social rules to guide their behavior (Gorer, 1965). They often feel confused or uneasy, and a number of demographic and social trends have been posited as underlying this.

Impact of Recent Trends

There have been significant changes in how death, terminal illness, and mourning are experienced and managed in the West (Blauner, 1966). Until the late nineteenth century, death was common at any age and especially in infancy, childhood, and childbirth; a relatively short illness caused by infectious disease was the most likely precursor to death. Today, death is most commonly caused by the degenerative diseases of old age, and dying can take months or years. Most of those near the end of life have long completed their main roles in work and family, and are hidden away in hospitals and elder care facilities; thus death has become separated from life, and young people do not learn about it from direct observation. In addition to these changing demographics, a number of social

processes such as secularization, medicalization, and sequestration have been implicated in the modern management of death.

Secularization

From a religious perspective, dying was the 'defining moment' in a person's life (Seale, 1998). Typically, a religious authority would assist at this feared rite of passage to 'the next world.' With the decline of religious belief in general, and of fear of damnation in particular, death lost much of its sacred or theological importance. Instead death became a time of separation from loved ones at the ending of a human life, while dying increasingly became a private act. The shift from religious to secular is illustrated by changes in the image of a 'good death.' Many modern secularists and even religious people now wish for an 'unconscious death,' a painless sudden demise during sleep, which in the Age of Belief was considered an 'accursed death' because it left no opportunity for the dying person to prepare to meet his or her maker (Ariès, 1974). Gorer (1965), in a widely cited monograph, argued that death had replaced sex as the modern form of pornography, an embarrassing taboo topic. Becker (1973) claimed that although fear of death was universal, modern society was characterized by a pathological 'denial of death.'

Professionalization

As the majority of people began to die in hospitals, death and dying gradually came under medical control. Some criticized the resulting bureaucratic organization of care as constraining and shaping the behavior of doctors, nurses, patients, and families in ways they often do not understand (Kaufman, 2005). With medicine oriented toward cure, incurable patients risked physical and emotional isolation, characterized by Elias (1985) as "the loneliness of the dying." In response to this, the hospice and palliative care movement has attempted to humanize dying, acknowledging its emotional, social, and spiritual as well as physical aspects, though from a Foucauldian sociology of power perspective, this entails medical surveillance being extended beyond just the patient's body to their entire being (Clark, 1999).

Outside of palliative care, the dramatic rise of life-support technology created situations in which a person was neither fully alive nor completely dead. As a result, medicalized death lost its status as a defining moment, but became instead an arbitrary time when the machines were finally turned off. "Death in the hospital is no longer the occasion of a ritual ceremony, over which the dying person presides amidst his assembled relatives and friends. Death is a technical phenomenon obtained by the cessation of care" (Ariès, 1974: p. 88). This situation also created a need for new definitions of death, the most widely accepted being that of brain death or, more controversially, the demise of the cerebral cortex, associated with the irreversible loss of personhood. Brain death, which opened up the possibility for the transplantation of organs, was widely accepted in the West but remains highly controversial in Japan (Lock, 2002).

The standard narrative in the history and sociology of dying, reflected in the paragraphs above, posits a historical shift from religious dying within a known community to medicalized dying, whether isolated in a bureaucratic hospital or holistically cared for in a hospice. Doctors thus tend to get portrayed as the bad guys – not actually a view held by most modern citizens, who would far rather live to a pain-controlled old age, even if ending in hospitalized social isolation or a technologically induced social death, than to die young and in agony from cholera. Allan Kellehear (2007), however, argues that in a really long-term historical perspective, those who first took dying away from communities were not medical but religious professionals; this occurred with the rise of the world religions in the ancient world. In this view, the key issue is not medicalization but professionalization, and indeed contemporary holistic care of the dying does add a whole new raft of professional groups to the 'multidisciplinary team's' clinical gaze. This has recently led to some radical initiatives attempting to return responsibility for dying to communities, even within complex urban settings (Conway, 2011).

Sequestration

Sociological accounts of death in the modern era often draw on a narrative of a supposed decline of community. The dying and the bereaved become isolated, and grief in particular gets defined as an inner psychological process rather than a shared social experience (Walter, 1999). Whatever merits the Freudian concept of denial may have within a psychological perspective (Becker, 1973), some sociologists have argued that other concepts are needed to make sense of the social position of those personally confronting death (Kellehear, 1984). Chief among these is sequestration, which points to the isolating not only of death but also of other troubling aspects of the human condition such as madness, from everyday life (Mellor and Shilling, 1993). With the dying hidden away in hospital, and the loss of mourning attire rendering the grieving invisible, death disappears from public discourse, yet remains an intense personal experience for those who themselves are dying or grieving. Others, however, have argued that the public/private relationship may be the other way around: with many not having to face significant bereavement until their 30s, yet with images of death ever present in the mass media, death is perhaps more present in public than in private. Either way, relationships between public/private and individual/community have been of considerable interest to sociologists of death.

Sociology of Dying

Sociological studies of dying have focused on care of dying people, rather than the experience of dying. A landmark study was done by Glaser and Strauss (1965, 1968) who developed two frameworks for analyzing the interaction between the dying person and medical staff: awareness contexts and dying trajectories.

Awareness of Dying

'Awareness contexts' referred to the degree of shared knowledge between patient and doctor concerning the fatal prognosis,

which ranged from 'closed,' where only the doctor knew, through intermediate stages – 'suspicion,' 'mutual pretense' – to 'open' when both parties could freely discuss the terminal illness. At the time of the original study, closed awareness was the norm, and arguably the study proved to be one of the most effective ethnographies in any field of sociology in terms of promoting change in policy in practice. By describing hospitalized dying, it caused readers to ask if they wanted to die like that, and many replied in the negative; doctors in individualistic Western societies are now much more open both about diagnoses of cancer, and of its subsequent end stages, though closed awareness remains common in more collectivist and/or family-oriented countries like Japan or Italy.

Dying Trajectories

Dying trajectories are "perceived courses of dying rather than the actual courses" (Glaser and Strauss, 1968: p. 6). The trajectories were based on the "time for dying" (Glaser and Strauss, 1968) as expressed in a pair of dichotomies: expected/unexpected, sudden/lingering. This produced four basic trajectories: sudden unexpected, most commonly seen in the emergency room; sudden expected, common in intensive care units; lingering unexpected as in the uncertain course of AIDS patients and many frail old people; lingering expected, as in the final stages of cancer. Each hospital ward was thus geared to a specific dying trajectory and used specific strategies to maintain its routine or 'sentimental order.' Medical personnel and family members experienced most distress when an actual trajectory deviated from its predicted course.

Social Death

Humans are born as primarily physical beings who then steadily acquire an existence that is increasingly shaped by culture. In many forms of dying, there is a reversal of this process as the dying person 'falls' from culture back into nature (Seale, 1998). "Much of the time I feel I'm just a body to be fed, watered and toileted," a friend, once vibrant but now in frail old age, told me a few weeks before she died, at home. Kept alive in hospital on life-support, the feeding, watering and toileting may be done automatically and impersonally through tubes and pumps.

A similar sociological concept is that of social death (Mulkay and Ernst, 1991). This concept shows that the end of a person's social life may not coincide with the end of their physical life. The person in a coma, or with advanced dementia, may no longer be related to as a person with agency; thus relatives may feel that the person has 'already died.' In non-Western cultures in which relationships are more important than individual autonomy, experiences of dependency and care may differ.

Communication and Control

Many modern Western individuals who value personal autonomy fear ceasing to be a social actor, becoming just a body to be done to by others. To mitigate this fear, contemporary health care agendas promote choice and control toward the end of life; the aim is to enable people to live socially for as long, or almost as long, as they live physically – to continue for as long as possible as cultural beings and social agents, not just physical bodies. Palliative care and euthanasia, though ideologically opposed in some countries, offer different routes to the same goal of reducing the time gap between social and physical death. Palliative care strives to extend social life as long as possible; euthanasia offers the choice to bring forward physical death to before social death sets in. Both value patient choice and control.

Contemporary health care discourse now "redirects incoherence, anxiety, breakdown, diffuse suffering, and any other expression of affect that lacks rationality" into control, choice, and the good death (Kaufman, 2005: p. 17). 'Choice' and 'control' are neat, abstract concepts that assume individual autonomy. But according to Kaufman's American hospital study, choice and control are illusions. As in both consumerism and neoliberal governmentality, patients and families are required to choose, but only at certain times, and only certain choices are available. Patients and families do not understand the system that undergirds this, and so are not prepared. Nor is prognosis an exact art, especially when cancer or AIDS are not involved. Meanwhile, Western medical ethics continue to promote patient autonomy as a key principle in end of life care.

Funerals

In contrast to considerable research into care of the dying, much less attention has been given to the sociology of the dead. Sociological studies of funerals have mainly looked at the funeral industry, whether from the point of view of professionalization (Pine, 1975) or economic sociology (Suzuki, 2001). What is strikingly lacking are sociological theories of the funeral rite itself; the few empirical studies that have been done tend to be through the eyes of professionals such as funeral directors and clergy, very occasionally the bereaved family, and almost never the whole congregation of mourners. Classic anthropological theories of rites of passage have proved difficult to apply to modern Western societies.

It is clear, however, that funeral rites in many Western countries are changing – from looking forward to heaven to looking backward to celebrate the deceased's earthly life; from displaying social status to displaying the deceased's individuality. Such shifts may be understood not only in terms of secularization but also in terms of postmaterialism. According to Ronald Inglehart, postindustrial affluence and security produce a new – postmaterial – cultural formation (Inglehart and Welzel, 2005). As people come to take physical and economic survival for granted, so their values shift toward self-expression, subjective well-being, and quality of life. These are precisely the values promoted by palliative and bereavement care, but the implications are most dramatic for funerals. In many rural societies, funerals were a time for the community to come together. With industrialization, people moved (often across continents) to new industrial towns and became concerned with class and status, that is, with finding their place in the new class-based urban society. Thus funerals became a way of displaying class and status position, as in the nineteenth century British funeral, or the mid-twentieth

century American funeral lampooned by Jessica Mitford (1963); only in death did many mid-century Americans get to ride in a Cadillac! By contrast, today's postmaterialists – secure within a stable middle class position – no longer need to display status, and increasingly use the funeral to display not status, but lifestyle – the unique lifestyle of the deceased.

Postmaterialism has implications also for the funeral's material trappings. Confidently upwardly mobile American baby boomers reject the mid-twentieth century baroque American funeral and instead display their good (postmaterial) taste by abandoning status display and opting for a simpler cremation. In the UK, where a minimalist cremation rite has been the norm for 50 years, postmaterial values are displayed by those choosing instead natural burial. Meanwhile, less secure members of affluent societies, along with recent migrants to cities in developing societies, typically retain a more traditionally 'respectful,' or status-displaying, style of funeral. The status insecurity approach to theorizing funerals, common currency for decades in anthropology and archaeology (Cannon, 1989), thus has considerable potential when combined with the recent thinking about postmaterialism to help explain trends in contemporary funerals.

Bereavement

Psychological studies of bereavement have focused on what modern Western mourners feel, while anthropological studies have focused on what non-Western mourners do, especially through ritual. Reflecting this, a number of sociological studies have looked at the mourning customs of minority groups within Western societies (Kalish and Reynolds, 1981). The impression this uneven research literature gives is that traditional people's feelings are governed by cultural norms and rites, while grief for majority members of Western societies is driven by an inner, natural psychological process and is not a social phenomenon. This is despite Hochschild's (1983) sociological analysis of feeling rules in contemporary capitalism, and historical studies showing how grief norms grief have evolved and been politically manipulated in modern western countries (Jalland, 2010).

A 50-year-old book that still stands out as a beacon in this sociological research desert is Geoffrey Gorer's (1965) national representative survey of grief and mourning in Britain. Gorer identified a number of different cultural norms about grief, one of them being that it should be private, which in the 1990s got picked up in sociological debates about the public/private relationship in dying and grieving, not least in Walter (1999). That grief is governed by social norms, for example, that grief is appropriate for certain kinds of relationships or should last a certain length of time, is also central to Ken Doka's (2002) development of the concept of disenfranchised grief, which identified the circumstances in which grief might not be socially acknowledged. The concept of disenfranchisement is sociologically ambiguous, however, in that although acknowledging the normed nature of grief, it seems to imply that every kind of grief should be enfranchised, that is to say, grief should be naturally expressed and not governed by social norms. To our knowledge, however, no human society has yet managed without grief norms, and it seems unlikely that advanced capitalism's need to control both its workforce and consumers will be significantly different.

Another sociological approach to grief is to draw on the sociology of emotion (Charmaz and Milligan, 2006), and there is considerable potential for developing this.

Remembrance and Collective Memory

Psychological theories of bereavement turned in the 1990s to consider the continuing bonds between the living and the dead (Klass et al., 1996), which helped provide an intellectual climate in which sociologists could explore the 'postselves' (Kearl, 2010) that people construct for themselves in anticipation of their death or have constructed for them after death – although Unruh (1983) had already theorized this phenomenon sociologically in the early 1980s.

Public responses to high-profile deaths (AIDS, 9/11) and the dramatic sudden death of high-profile individuals (e.g., President Kennedy, Princess Diana) may temporarily recreate the solidarity of an 'imagined community' of mourners who vicariously participate in traditional ceremonies via television and the Internet. This recalls the seminal work of the French sociologist, Emile Durkheim (1995), who showed how mourning ritual reinforced the cohesiveness and core values of the social group. Exploring the dialectic of how death can both threaten the basis of society and enhance its solidarity thus continues to be a theme in the sociology of death.

The turn of the century has also witnessed a major memorial boom. Doss's (2010) analysis of memorial politics shows that collective memory is not necessarily consensual, but entails power and conflict. This understanding links the sociology of public mourning to research in collective memory, and indeed several research studies in collective memory focus on memorialization of the dead. Postmodern death seems to have come out of the private closet to which modernity arguably confined it.

Communication Technology

Many people are now using the Internet to socially relocate those who are dying, mourning, or dead (Sofka et al., 2012; Walter et al., 2011–2012). Kin spread around the world and may, through Skype, be present at an old person's hospital deathbed. Close family, left alone by modernity to grieve in private, now log in to the deceased's Facebook site where they encounter posts from a much wider penumbra of the dead person's social network. This reinsertion of the grieving individual into a community of mourning may be welcomed, as when a bereaved parent sees that the deceased's friends have not forgotten an anniversary. However, individuals grieve in different ways – ways that were invisible when grief was private but are now visible online; older kin are not always happy with what they may see as trite comments produced by the deceased's online community. Privacy protects, in death as in life; living in community can be hard work, in death as in life.

People have probably talked to the dead from time immemorial, but in much of the twentieth century such conversation was believed to be irrational, so it came to be conducted in

private, at home or at the graveside when nobody else was around, quietly under the breath, or within the very specific confines of a séance or Spiritualist church meeting. Online, however, twenty-first-century mourners address the dead in the company of other mourners, and are not criticized for so doing; indeed it has become an online convention to post messages to the dead. And unlike snail mail, the online dead often have an address (e.g., their still-extant social network site or a cyber cemetery) to which messages can be posted. Interestingly, even in secular countries such as Britain and Germany, mourners seem more likely online than offline to use spiritual language, referring to the dead as in heaven or as an angel (Jakoby and Reiser, 2013). It is as though heaven, deemed a nonplace once modern astronomy had charted the heavens, becomes plausible again – as cyberspace, a mystical, incomprehensible intermediary between here and there, between now and then, between the living and the dead. And maybe this is why the online dead tend to be addressed not as souls, cut off in heaven from the living, but as angels who maintain some kind of ongoing relationship with the living (Walter, 2011).

See also: Body: Anthropological Aspects; Collective Memory, Anthropology of; Collective Memory, Social Psychology of; Collective Memory, Sociology of; Death, Anthropology of; Disability: Sociological Aspects; Emotions, Sociology of; Euthanasia; Life Course: Sociological Aspects; Medical Sociology; Modernity: History of the Concept; Secular Religions; Shame and the Social Bond; Social Science and the Media; Suicide, Sociology of; Suicide.

Bibliography

Ariès, P., 1974. Western Attitudes toward Death. Marion Boyars, London.
Becker, E., 1973. The Denial of Death. Free Press, New York.
Berger, P., 1969. The Social Reality of Religion. Faber, London.
Blauner, R., 1966. Death and social structure. Psychiatry 29, 378–394.
Cannon, A., 1989. The historical dimension in mortuary expressions of status and sentiment. Current Anthropology 30, 437–457.
Charmaz, K., Milligan, M., 2006. Grief. In: Stets, J.E., Turner, J.A. (Eds.), Handbook of the Sociology of Emotions. Springer, New York.
Clark, D., 1999. 'Total pain', disciplinary power and the body in the work of Cicely Saunders, 1958–1967. Social Science & Medicine 49, 727–736.
Conway, S. (Ed.), 2011. Governing Death and Loss: Empowerment, Involvement and Participation. Oxford University Press, Oxford.
Doka, K.J. (Ed.), 2002. Disenfranchised Grief: New Directions, Challenges, and Strategies for Practice. Research Press, Champaign, IL.
Doss, E., 2010. Memorial Mania: Public Feeling in America. Chicago University Press, Chicago.
Durkheim, E., 1995 [1912]. The Elementary Forms of the Religious Life. Free Press, New York.
Elias, N., 1985. The Loneliness of the Dying. Blackwell, Oxford.
Glaser, B., Strauss, A., 1965. Awareness of Dying. Aldine, Chicago.
Glaser, B., Strauss, A., 1968. Time for Dying. Aldine, Chicago.
Gorer, G., 1965. Death, Grief and Mourning in Contemporary Britain. Cresset, London.
Hochschild, A., 1983. The Managed Heart: Commercialization of Human Feeling. University of California Press, Berkeley, CA.
Inglehart, R., Welzel, C., 2005. Modernization, Cultural Change and Democracy. Cambridge University Press, New York.
Jakoby, N.R., Reiser, S., 2013. Grief 2.0. Exploring virtual cemeteries. In: Benski, T., Fisher, E. (Eds.), Emotions and the Internet. Routledge, London.
Jalland, P., 2010. Death in War and Peace: Loss and Grief in England 1914–1970. Oxford University Press, Oxford.
Kalish, R., Reynolds, D., 1981. Death and Ethnicity: A Psychocultural Study. Baywood, Farmingdale, NY.
Kaufman, S., 2005. And a Time to Die: How American Hospitals Shape the End of Life. University of Chicago University Press, Chicago.
Kearl, M., 1989. Endings: A Sociology of Death and Dying. Oxford University Press, Oxford.
Kearl, M.C., 2010. The proliferation of postselves in American civic and popular cultures. Mortality 15, 47–63.
Kellehear, A., 1984. Are we a death-denying society? A sociological review. Social Science & Medicine 189, 713–723.
Kellehear, A., 2007. A Social History of Dying. Cambridge University Press, Cambridge.
Klass, D., Silverman, P.R., Nickman, S.L. (Eds.), 1996. Continuing Bonds: New Understandings of Grief. Taylor & Francis, Bristol, PA.
Lock, M., 2002. Twice Dead: Organ Transplants and the Reinvention of Death. University of California Press, Berkeley.
Marvin, C., Ingle, D., 1999. Blood Sacrifice and the Nation: Totem Rituals and the American Flag. Cambridge University Press, Cambridge.
Mellor, P., Shilling, C., 1993. Modernity, self-identity and the sequestration of death. Sociology 273, 411–432.
Mitford, J., 1963. The American Way of Death. Simon & Schuster, New York.
Mulkay, M., Ernst, J., 1991. The changing profile of social death. Archives Europe Sociology 23, 172–196.
Pine, V.R., 1975. Caretaker of the Dead: The American Funeral Director. Irvington, New York.
Seale, C., 1998. Constructing Death: The Sociology of Dying and Bereavement. Cambridge University Press, Cambridge.
Sofka, C., Cupit, I.N., Gilbert, K., 2012. Dying, Death, and Grief in an Online Universe: For Counselors and Educators. Springer, New York.
Suzuki, J., 2001. The Price of Death: The Funeral Industry in Contemporary Japan. Stanford University Press, Stanford, CA.
Unruh, D., 1983. Death and personal history: strategies of identity preservation. Social Problems 303, 340–351.
Walter, T., 1999. On Bereavement: The Culture of Grief. Open University Press, Buckingham.
Walter, T., 2011. Angels not souls: popular religion in the online mourning for British celebrity Jade Goody. Religion 411, 29–51.
Walter, T., Hourizi, R., Moncur, W., Pitsillides, S., 2011–2012. Does the internet change how we die and mourn? Omega 644, 275–302.

Death, Anthropology of

Henry Abramovitch, Tel Aviv University, Tel Aviv, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The anthropology of death examines the diversity and commonalties in how human societies respond to the demise of its members. Anthropologists have documented the enormous cultural variation in the methods for disposing of the corpse, the expected behavior of the bereaved, and the ongoing relations between the living and their dead. Most cultures conceptualize death as a transition, or rite of passage, in which the fate of the corpse is linked to the ritual status of the mourners. The social impact of the death is related directly to the social status of the deceased and the type of death. The sudden, violent death of an important leader may paralyze a society, while the lingering death of a marginal stranger may pass unnoticed. Death rituals provide an opportunity to reassert core cultural values at a time of potential demoralization, and as such have important economic, political, and ideological aspects. Contemporary Western attitudes toward death, seen from a cross-cultural perspective, appear decidedly 'deviant.' Future research will focus on rituals and institutions, as well as on the impact of new technologies (medical, computer, space, etc.) on the changing conception of death.

All human cultures struggle to deal with the inevitability and mystery of death. The anthropology of death explores how human societies around the world respond to death. It is concerned with both the conceptual and organizational aspects, that is, what people believe about death and the afterlife, as well as what they actually do when faced with the crisis of death. Two main methodological approaches used are the ethnographic and the comparative. The ethnographic approach examines how single cultures cope with the demise of members, while the comparative approach tries to make sense of the enormous cultural variation in issues such as the disposing of the corpse, the expected behavior of the bereaved, and the ongoing relations between the living and their dead.

Although elaborate death rituals are probably a defining aspect of human culture, comparative studies have revealed few, if any, universal practices. Even the widespread practice of crying at a funeral is actively discouraged, for example, among the Balinese. Eating the dead (endocannibalism) is an exotic and a controversial form of disposing of the human body that was once widespread in the Amazon and Melanesia. Nevertheless, anthropologists have uncovered a number of key metaphors that help to make sense of the enormous diversity of mortuary rituals. The anthropology of death takes as its task to understand the phrase: 'All humans die,' yet in every culture, each dies in their own way.

Centrality of the Death Ritual

In most traditional cultures, death rituals stand at the center of social life. Many societies use expensive and elaborate mortuary ritual as a way of demonstrating status and power, so that the expense incurred is often an enormous financial burden. The atmosphere is not always sad or somber, but may even take on a festive atmosphere so that one may speak of 'celebrations of death' (Huntingdon and Metcalf, 1991).

The impact of the death is related directly to the social status of the deceased. Death is likely to be seen as particularly disruptive when it strikes persons who are most relevant for the functional and moral activities of the social order: "At the death of a chief, or man of high standing a true panic sweeps over the group" (Hertz, 1960 [1907], p. 76). The elaborate funeral rituals and pyramids of Ancient Egypt are outstanding examples of this ethos. Likewise, the death of a spouse often leads to a long period of taboos and restricted activity. "On the contrary, the death of a stranger, slave or child will go almost unnoticed; it will arouse no emotion, occasion no ritual" (Hertz, 1960 [1907], p. 76). In regions of high mortality, the death of an infant who is not yet considered a 'social person' may have no formal ritual or mourning. Bereaved parents often experience their loss privately, without ceremony, in what Scheper-Hughes (1992) called 'death without weeping.' In Northeast Brazil, the death of an infant or young child is often considered a blessing, in which the angel baby attains happiness in heaven. A mother's tears were actively discouraged since they would wet the angel's wings and impede the spirit-child's flight to a better place. The mothers expressed no grief but rather felt that the baby was looking after her from heaven. Scheper-Hughes' work challenges much of the received wisdom concerning the nature of grief, where such mothers would be considered as suffering pathological grief and in need of urgent intervention.

The simple funerals of some hunter-gatherers, such as the Baka Pygmies of the Central African rain forest, are notable exceptions to the widespread pattern of elaborate and complex mortuary rituals. They have no idea of an afterlife, saying "When you are gone, bang! you are gone" (Bloch and Parry, 1982).

Fieldwork Dilemmas

Death is an intensely emotional and often taboo subject, so that studying death raises special dilemmas and emotional challenges for the fieldworker (Palgi and Abramovitch, 1984). In many Western cultures, death and mourning is a time of heightened privacy. The entry into the mourner's space at such times is considered intrusive and inappropriate. Yet the ethnographic imperative demands researching even such acute grief. How fieldworkers manage their own culture's attitude toward death may be considered a case of 'ethnographic countertransference.' One explicit and early example was

described by Hortense Powdermaker, who worked in a matrilineal society in New Ireland. Powdermaker described her own extreme distress when she began taking fieldnotes at her first funeral. She imagined how intrusive such an ethnographic presence would be in her own house of mourning. She, however, discovered to her great surprise that these illiterate people felt no such intrusions, but rather that her writing added prestige to the ritual. They demanded her presence at every subsequent funeral, even long after she had constructed a complete account of the funeral process (Palgi and Abramovitch, 1984).

Many ethnographers have discussed the emotional strain of participating closely in the grief of others (Woodthorpe, 2007). Perhaps, the most moving account is by Rosaldo (2004) who connects how his overwhelming grief at the accidental death of his wife (also an eminent anthropologist) helped him understand more deeply headhunter's rage of the Ilongot in the Philippines. Rosaldo criticized the discipline's focus on death ritual, with its formality and routine, in a way that marginalizes the intense and raw feelings that often accompany actual grief. Similarly, Fabian (1973) argued that anthropologists avoid the intimate encounter with death by focusing on 'how others die.'

Death as an Event vs Death as a Process

In the West, despite recent legal and medical controversies, death is considered as occurring at a specific identifiable moment, symbolized by the 'time of death' on the death certificate. This 'punctual view of death' (Bloch, 1988) in which a person is thus either alive or dead is not shared by many cultures. Physiological demise and social death do not necessarily coincide. Instead, death is seen as part of a long, complex, and even dangerous process, with no sharp boundary between life and death. Cultures express the degree of alive–dead using different metaphors. The Merina of Madagascar, for example, use the image of wet–dry to indicate degrees of alive–dead. A newborn infant is very wet and thus very alive; in contrast, a shriveled elderly person, almost totally dry, is mostly dead. The process of drying, and hence dying, will be only completed long after biological death when the bones are dug up and placed in communal tombs (Bloch and Parry, 1982). Among Sa'dan Toraja of Indonesia, a dying person has the ability to 'see everything' but is not considered dead until the first sacrifice of mortuary rites have taken place (Tsinntjilonis, 2007).

Theoretical Approaches and Key Metaphors

"Few areas of contemporary anthropological inquiry are still so dominated by *fin de siècle* thinking as is the study of death and mortuary ritual" (Palgi and Abramovitch, 1984, p. 387). The three traditional paradigms are French sociological school (Durkheim, Hertz), early British functionalists (Malinowski, Radcliffe-Brown), and 'rite of passage' approach (Van Gennep).

Hertz's Analysis of First and Second Funerals

The single most influential text in the anthropology of death was written by a brilliant disciple of Durkheim, Robert Hertz, who was killed on the battlefield of Verdun in World War I. Many of the central ideas in the anthropology of death: That death is widely perceived as a process and not as an instantaneous event, that death is the starting point of a ceremonial process whereby a dead person becomes an ancestor, and that death is akin to an initiation into a social afterlife and hence resembles a kind of rebirth, all derive from his seminal monograph (Hertz, 1960 [1907]). His study focused on the widespread custom of second funeral, or more correctly, secondary treatment of the remains. One of the main ideological reasons for a double ceremony, often separated by years, has to do with the dual nature of the person as composed of both body and spirit (or soul). Dead bodies or corpses are usually associated with ritual pollution and sorrow. The first funeral, in a literal and symbolic sense seeks to get rid of these polluting and sorrowful aspects. The second funeral is focused on the initiation of dead people's spirits into the realm of the ancestors, where they will continue to serve as a cultural resource, watching over and guiding the living. This structuralist perspective highlighted parallels between the decomposition of the corpse and the fate of the soul.

Death Rituals and Core Values

In many societies, the occurrence of a death disrupts social life severely, so that "it is stricken in the very principle of its life in the faith it has in itself" (Hertz, 1960 [1907], p. 78). A functionalist perspective stressed how mortuary rites served to resolve the disruptive tendencies that operate at times of social crisis. Malinowski demonstrated how ceremonies counteract the centrifugal forces of fear, dismay, and demoralization associated with death, and provide a powerful means of reintegration of the group's shaken morale (Palgi and Abramovitch, 1984). These rituals must provide an answer to the meaning of life for the community at a time when it is most threatened. As a result, the study of such death rituals provides a unique opportunity for studying the core values of any culture. The functionalist perspective emphasized the problem of death for society, and especially the issues of inheritance, redistribution of rights, and statuses, as well as the reintegration of mourners into day-to-day life.

Death as a Journey or Rite of Passage

Most cultures conceptualize death as a transition or rite of passage. In many cultures, this transition is seen as a journey to an ultimate destination that may culminate in rebirth, ancestral abode, reunion with nature or divinity, or indeed total oblivion. Death rituals, like all rites of passage, have a three-part structure, first delineated by Van Gennep: separation, liminality, and reincorporation (Huntingdon and Metcalf, 1991). The spirits of dead people must be separated from their social roles as members in the community of the living and enter an undefined 'in-between' state, finally being reincorporated into a new status as the end of the 'journey.' Often, the fate of the soul is modeled on the fate of the body. Just as the soul of the deceased is formless and repulsive, as in Borneo when the dead are left to ferment in large jars, so the soul of the deceased is seen as homeless and

the object of dread. Only when the bones alone are left can the ritual journey be completed (Huntingdon and Metcalf, 1991). Likewise, the ritual status of the mourners is in turn linked to the fate of the corpse in its process of decay, and the soul in its progress on its journey. Some cultures do have a specific tie-breaking ritual, which symbolizes both the end of mourning for the bereaved, and the end of the journey for the soul or spirit of the deceased. The living are often required to nurture the deceased at each stage along the journey through the afterlife and this conception helps to explain why the dead take up so many economic, emotional, and spiritual resources.

The Danger of the Unincorporated Dead

From this rite of passage perspective, it is possible to understand why the unincorporated dead, trapped permanently in the liminal realm, are often considered as dangerous. These wandering spirits, for whom no rites were performed, may act as hungry ghosts. They yearn to be reincorporated into the world of the living. Since they cannot be, they behave like hostile strangers who consequently must find sustenance at the expense of the living. Many cultures use elaborate strategies to confuse the spirit of the deceased so that it will not return to the realm of the living. Illness, misfortune, and associated healing rituals are often attempts to incorporate these lost souls. The 'Tomb of the Unknown Soldier' in modern states provides a resting-place for unincorporated spirits of the war dead.

Parallels to Other Life Cycle Rituals

Anthropologists have also noticed how death rituals often parallel other life cycle rituals, notably births and weddings. Conceptually, death is often seen as a rebirth into another world, while birth is viewed as the literal rebirth of a deceased ancestor. Likewise, for weddings, just as a bride may leave her family, so too the soul of the deceased departs for another world. In rural Greece, for example, the same lyrics are sung at both weddings and funerals (Danforth, 1982).

Emotional Response of the Mourners

The comparative study of how mourners respond to loss and bereavement has attracted considerable attention (Parkes et al., 1997). Almost all cultures allow, or even encourage, the expression of crying, as well as fear and anger by mourners. Women tend to cry and self-mutilate, while men express anger, or direct aggression away from themselves. Yet most societies have developed mechanisms and institutions to control the anger of the bereaved, especially via the use of ritual experts (Rosenblatt et al., 1976). In many cultures, there is no cultural category of a 'natural death'; rather, each death is akin to murder in which a culturally sanctioned cause of death must be uncovered. Almost all societies concern themselves with the spirits of the dead that probably give expression both to the 'unfinished business' between the living and the dead, and the fact that death does not end a relationship. The more complicated the relationship and unfinished business in which the dead are dispossessed in favor of the living, the greater will be the dread and concern about ghosts.

Good Death vs Bad Death

Recent interest has focused on cultures' conception of what constitutes a 'good death' or a 'bad death' (Aries, 1974; Humphreys and King, 1981). A good death represents a cultural ideal that enacts a symbolic victory over death and the regeneration of life (Bloch and Parry, 1982). A bad death does the opposite, leaving survivors despairing, helpless in the face of meaninglessness, evil, or nothingness. Unpredictability, violence, or intentional harm are widespread attributes of a bad death. Archetypal examples of bad death include suicide, homicide, and death in a traffic accident. A good death is the mirror image of a bad death: An expected, painless death of an elderly individual with a multitude of descendants in attendance. Jewish Talmud compares a good death to the ease of plucking a hair from a bowl of milk and a bad death to the interminable agony of pulling wool from off a thorn bush. But a good death has culture-specific aspects, such as the setting, timing, physical posture, thoughts, and actions at the final moment, and other 'near death experiences,' as well as a preferred cause and manner of death (Abramovitch, 1999). A Hindu should die in the open air, as an act of self-sacrifice, abandoning life to the sound of the chanting of the names of God, for his thoughts at that moment may determine his subsequent fate and rebirth. A reluctance to die at the appropriate time may lead to the death of a young relative in his place (Parry, 1994).

Controversies Over the Beginning and End of Life

Controversies at the beginning and the end of life over abortion, organ donation, and brain death may be analyzed in anthropological terms (Kaufman and Morgan, 2005). The disagreement may be understood as revolving around conflicts concerning the nature of the 'social person.' Each society has different criteria for when an infant becomes a fully human, social person. For example, among Kaliai people of West New Britain, the key distinction is the child's ability to discuss dreams. Children who are unable to do so, will not have death ritual should they die; but dream tellers will (Counts and Counts, 2004). Opponents of abortion identify conception itself as signaling the onset of the social person giving the fetus inalienable rights to life and protection. Supporters of abortion, in contrast, regard the fetus as a part of the woman's body that only becomes social person at birth. Similarly, proponents of brain death and organ transplants identify a social person largely with cognitive function; when that is lost, the person is gone forever and the organs may be harvested. In contrast, in Japan, which long resisted organ transplantation, rejected brain death criteria for death since traditionally they believe that a person's life spirit is diffused throughout the body. The first Japanese surgeon to attempt a transplant, using brain death criteria, was charged with murder. In a different context, Japanese women will often make special ritual offerings to fetal spirits following an abortion.

Changes in Mortuary Ritual

Mortuary rituals have long been considered as one of the most stable aspects of a culture. Recent research has challenged this view. Change may occur rapidly altering the tradition

of millennia. One dramatic case concerns South Korea (Park, 2010a, b). Until the 1980s, death and funerals took place at home, in private ceremony in the family garden (*mandang*), followed by burial in an auspicious site that was consulted in times of misfortune. Cremation was a social taboo that was identified with the Japanese occupiers. In the last decades, traditional burial has almost disappeared. Funerals are held in hospital or private funeral halls, followed by public cremation with the ashes lodged in columbarium or scattered beneath trees in a memorial park. The severe shortage of land for cemeteries, change in design of housing, government policy all played a role in the change, as it did in the rapid introduction of wall burial in Israel.

Future Directions

The anthropology of death traditionally focused narrowly on the death of single individuals in small-scale societies. Future research, in contrast, will need to investigate how local cultural traditions interact with the wider collective and global context. The study of mass death, in the form of epidemics, natural disasters, or political violence, provides one stimulating area of investigation (Robben, 2004b). The diverse impact of new technologies may provide another intriguing area for understanding how conceptions of death are changing. Web pages on the Internet provide for new forms of commemorating the dead. Organ donation allows a 'bad death' to be transformed into a 'good death' as a posthumous act of self-sacrifice. Cloning awakens in some the 'myth of eternal life'; while artificial life support has created a new state in which persons are neither alive nor dead.

Death will always remain a mystery, but the cross-cultural investigation of death will continue to provide unexpected insights into how humans cope with that mystery.

See also: Ancestors, Anthropology of; Anthropology: Overview; Birth, Anthropology of; Collective Memory, Anthropology of; Death and Dying, Sociology of; Ethnography in Applied Social Research; Ethnology; Euthanasia; Fieldwork in Social and Cultural Anthropology; Functionalism in Anthropology; Health: Anthropological Aspects; Holocaust, The; Hunting and Gathering Societies: Anthropology; Liminality; Magic, Anthropology of; Personhood, Anthropology of; Psychological Anthropology; Ritual; Shamanism; Spirit Possession, Anthropology of; Structuralism; Witchcraft.

Bibliography

Abramovitch, H., 1999. 'Good death' and 'bad death'. In: Malkinson, R., Rubin, S., Witztum, E. (Eds.), Traumatic and Non-traumatic Loss and Bereavement. Psychosocial Press, New York.

Ariès, P., 1974. L'Homme devant la Mort. Editions du Seuil, Paris. (Republished, H. Weaver, Trans., 1982. The Hour of Our Death. Knopf, New York.)

Bloch, M., 1988. Death and the concept of person. In: Cederroth, S., Corlin, C., Lindstrom, J. (Eds.), On the Meaning of Death: Essays on Mortuary Rituals and Eschatological Beliefs, Uppsala Studies in Cultural Anthropology, vol. 8. Acta Universitatis Upsaliensis, Uppsala, Sweden.

Bloch, M., Parry, J. (Eds.), 1982. Death and the Regeneration of Life. Cambridge University Press, Cambridge, UK.

Counts, D.A., Counts, D., 2004. The good, the bad, and the unresolved death in Kaliai. Social Science & Medicine 58, 887–897.

Danforth, L., 1982. The Death Rituals of Rural Greece. Princeton University Press, Princeton, NJ.

Fabian, J., 2004 [1973]. How others die: reflections on the anthropology of death. In: Robben, Antonius C.G.M. (Ed.), Death, Mourning and Burial: A Cross-Cultural Reader. Blackwell, Oxford, pp. 49–61.

Hertz, R., 1960 [1907]. Contributions à une étude sur la representation collective de la mort. Année Sociologique 10, 48–137. (Republished, R. Needham, C. Needham, Trans., 1960. Death and the Right Hand. Free Press, New York.)

Humphreys, H., King, H. (Eds.), 1981. Mortality and Immortality: The Anthropology and Archeology of Death. Academic Press, London.

Huntingdon, R., Metcalf, P., 1991. Celebrations of Death: The Anthropology of Mortuary Ritual, second ed. Cambridge University Press, Cambridge, UK.

Kaufman, S.R., Morgan, L.M., 2005. The anthropology of beginnings and ends of life. Annual Review of Anthropology 34, 317–341.

Palgi, P., Abramovitch, H., 1984. Death: a cross-cultural perspective. Annual Review of Anthropology 13, 385–417.

Park, C.-W., 2010a. Cultural Blending in Korean Death Rites: New Interpretive Approaches. Continuum, London & New York.

Park, C.-W., 2010b. Funerary transformations in contemporary South Korea. Mortality.

Parkes, C., Laungani, P., Young, B. (Eds.), 1997. Death and Bereavement across Cultures. Routledge, London.

Parry, J., 1994. Death in Benares. Cambridge University Press, Cambridge, UK.

Robben, A.C.G.M., 2004b. State terror in the netherworld: disappearance and reburial in Argentina. In: Robben, A.C.G.M. (Ed.), Death, Mourning and Burial: A Cross-Cultural Reader. Blackwell, Oxford, pp. 134–148.

Rosenblatt, P., Walsh, R., Jackson, A., 1976. Grief and Mourning in Cross Cultural Perspective. Human Relations Area Files Press, New Haven, CT.

Rosaldo, R., 2004 [1989]. Grief and a headhunter's rage. In: Robben, Antonius C.G.M. (Ed.), Death, Mourning and Burial: A Cross-Cultural Reader. Blackwell, Oxford, pp. 167–178.

Scheper-Hughes, N., 1992. Death without Weeping: The Violence of Everyday Life in Brazil. University of California Press, Los Angeles.

Tsinntjilonis, D., 2007. The death-bearing senses in Tana Toraja. Ethos 72, 173–194.

Woodthorpe, K., 2007. My life after death: connecting the field, the findings and the feelings. Anthropology Matters Journal 9 (1), 1–11.

Death Penalty

Benjamin Fleury-Steiner, University of Delaware, Newark, DE, USA
Ross Kleinstuber, University of Pittsburgh – Johnstown, Johnstown, PA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A. Sarat, volume 5, pp. 3273–3278, © 2001, Elsevier Ltd.

Abstract

Death penalty is the deliberate taking of human life as state-sanctioned punishment. In recent years, a majority of the world's countries have further restricted or abolished executions all together. Abolition of the death penalty in many countries has occurred in a context of momentous political, social, and legal changes. Newly established nations have embraced abolition, moreover, because powerful European institutions (e.g., the European Union) have made abolition a requirement for membership. At the same time, the European campaign has been less successful in China – the world's far and away leading executioner. China's complex politicolegal structure makes death penalty abolition exceedingly difficult.

Introduction

There are currently 58 countries worldwide that retain the death penalty both in law and in practice (Bohm, 2013). Of these, only two – the United States and Japan – are members of the Organisation for Economic Cooperation and Development, the body that represents the world's 36 industrialized nations (Bohm, 2013; Organisation for Economic Cooperation and Development, n.d.), and only 21 conducted an execution in 2012 (Amnesty International, 2013). At one time, the use of the death penalty was widespread, but the practice began to fall into disfavor in the final quarter of the twentieth century as nations began to recognize and emphasize human rights, especially in Europe. Article 5 of the United Nations *Universal Declaration on Human Rights*, which was ratified in 1948, bars the use of 'cruel, inhuman or degrading treatment or punishment,' and in recent years, majority world sentiment has generally concluded that the punishment of death is cruel and degrading.

This article addresses this momentous change in death penalty support from a global perspective. We begin with a discussion of the theoretical justifications for the death penalty, the origins of the death penalty, and its proliferation and subsequent abolition in most of the world's nations. We will then take a more in-depth look at the two retentionist countries: the United States and China. These case studies point to important directions for future research on death penalty abolition and highlight the limitations of privileging one perspective – international, national, or local – over the other (Miao, 2013).

Justifications for the Death Penalty

There are three primary theoretical justifications for the penalty of death: retribution, incapacitation, and deterrence. The idea of retribution is that the offender deserves to die because he or she has harmed society in some way and must sacrifice his or her life to compensate for this harm. He or she owes society a debt; the crime of murder deserves the penalty as just deserts for the pain and suffering caused to the victim(s). The theory of incapacitation is based on the goal of physically preventing offenders from committing future criminal acts. The ultimate incapacitation, therefore, is death. By killing those who have killed, we render them unable to prey upon and kill anyone else. In fact, there is evidence from a program of empirical research from the United States called the Capital Jury Project (Bowers, 1995) that capital jurors use this type of vicarious self-defense to justify their death verdicts. The less time that jurors think a capital defendant not sentenced to death will spend in prison before being released, the more likely they are to vote for death (Bowers and Steiner, 1999). Moreover, jurors who believe that the defendant will be dangerousness in the future are predisposed to vote for death (Vartkessian, 2012). Last, the theory of deterrence suggests that humans are essentially rational actors who weigh the costs and benefits of performing an action. By making the punishment for murder severe enough, people can be convinced not to kill out of fear of being killed themselves. This justification for the death penalty has received the most scholarly attention in the United States. However, an extraordinarily detailed review of more than three decades of deterrence studies of capital punishment in the United States seriously challenges the validity of this body of research as a whole and thereby any empirical claim that the death penalty deters murderers (Nagin and Pepper, 2012).

Historical Context: Reach, Retention, and Rejection

Hood and Hoyle (2008) and William Schabas (2002) have published the most detailed historical and empirical assessments of the death penalty from a global perspective. The purpose of this section is then to briefly highlight the most enduring themes in the use of the death penalty worldwide – namely, its reach, retention, and rejection. Beginning with the Ancient Laws of China, the death penalty was established as a punishment for numerous crimes. In the eighteenth-century BC, the Code of King Hammurabi of Babylon listed over 20 different crimes as punishable by death (Slanski, 2013). Adultery, harboring escaped slaves, and robbery were all punishable by death under the Code but murder was not. The Bible first codified the death penalty's reach as extending to

murder as well as a wide array of offenses such as violation of the Sabbath, adultery, homosexuality, incest, and rape, among dozens of other offenses. The first documented narrowing of crimes punishable by death occurred under Jewish Talmudic Law. Some rabbis called for its very limited use and some even called for abolition (Erez, 1981).

The death penalty had its most unprecedented and brutal expansion during the Middle Ages. Not only were the number of death-eligible crimes expanded but executions were invariably accompanied by torture. A woman convicted of high treason was burned alive and men were hanged or drawn and quartered. In terms of documented numbers, a high water mark for executions may very well have been under the reign of Henry VIII. It is estimated that at this time an astonishing 72 000 executions occurred (Randa, 1997).

Early into the Enlightenment period, Montesquieu was one of the first proponents of limiting the death penalty to murder. However, the modern abolition movement did not begin until Cesare Beccaria published *On Crimes and Punishments*. This revolutionary legal treatise had an enormous influence on legal systems across the globe. In the United States, for example, Pennsylvania adopted degrees of murder with one singular objective: to limit the death penalty's reach (Bedau, 1998). Many other jurisdictions in the Northern United States eliminated the death penalty for rape, robbery, burglary, and arson. By 1860, only treason and murder were death-eligible crimes. Only 4 years later, Michigan became the first US state to abolish the death penalty permanently. The Southern states also limited the number of capital crimes for whites but maintained the death penalty for numerous crimes for both free and enslaved blacks. Under the region's infamous 'Black Codes,' blacks could be executed for offenses such as rebellion, arson, burglary, and assaulting a white woman (Higginbotham, 1980).

In 1863, Venezuela became the first nation in the world to abolish the death penalty; the tiny Republic of San Marino in Southern Europe became the second in 1865. Numerous Latin American nations followed suit in the late nineteenth and the early twentieth centuries, including Costa Rica, Brazil, Ecuador, Uruguay, Colombia, and Panama (Bohm, 2013; Schabas, 2002). However, any momentum toward abolition came to a screeching halt with the rise of totalitarianism, specifically the rise of Nazi Germany. The Nuremberg war crimes tribunals would document Hitler and the Nazi regime's insatiable appetite for the death penalty. In the genocidal aftermath of the Second World War, criticism of the death penalty by human rights experts was a pivotal moment for an already galvanized international abolition movement. In the years leading up to 10 December 1948, the historic date when the United Nations General Assembly adopted the *Universal Declaration of Human Rights*, the abuse of the death penalty was of primary concern. However, the death penalty for Nazi war criminals was authorized as an exception. Continued retention for the so-called most atrocious crimes would be the primary "dialectic that confronted those who first proclaimed, in international law, a 'right to life'" (Schabas, 2002: p. 6).

If the death penalty for the most atrocious crimes created exceptions even in the midst of evidence of extraordinary abuses, then it is not surprising that a majority of countries retained it, at least in some form, for many decades after World War II. Only six nations – Germany, Honduras, Monaco, Dominican Republic, Austria, and Vatican City – abolished the death penalty between the end of World War II and 1970 (Bohm, 2013). However, only a few decades later, the tide would dramatically turn. Indeed, by the twenty-first century, a once historically elusive international consensus against the death penalty became a reality (Hood and Hoyle, 2008).

Rejecting the Death Penalty: A New Global Majority

The final quarter of the twentieth century and the early part of the twenty-first century saw rapid acceleration in the number of abolitionist countries. By the beginning of 1980, only 20 countries had abolished the death penalty, but by the beginning of 2013, this number had nearly *quintupled* – 97 nations had abolished the death penalty for all crimes and a total of 140 nations had ceased conducting executions for at least 10 years (Bohm, 2013). Perhaps the most compelling explanation for this radical decline has been the assimilation of international human rights norms calling for the worldwide abolition of the death penalty as prerequisite for membership to the Council of Europe and the European Union. Given the recent adoption of democratic constitutions and capitalist imperatives in former Eastern and Soviet Bloc nations, membership in these powerful institutions has been an important catalyst for widespread abolition. As Hood and Hoyle (2008: p. 25) explain, quoting the Parliamentary Assembly of the Council of Europe:

> Fundamentally important was the message that had been conveyed: a *principled* (emphasis in original) opposition to the death penalty as a violation of fundamental human rights. Both the Council of Europe and the European Union have declared that 'the death penalty has no legitimate place in the penal systems of modern civilized societies, and its application may well be compared with torture and be seen as a inhuman and degrading punishment.' The language is uncompromising.

Beyond a commitment to abolition as a membership requirement, the European Union has also engaged in aggressive diplomatic efforts with retentionist countries to abolish the death penalty. The European Union has, moreover, funded training for anti–death penalty lawyers and policy makers and has promoted empirical research to study the administration of the death penalty in retentionist countries. In South America and Mexico, the path to abolition has stemmed from a combination of factors, including the fall of authoritarian regimes, rise of antideath penalty religious imperatives, and the creation of national penitentiary systems (Hood and Hoyle, 2008: p. 62).

Retaining the Death Penalty: A New Global Minority

A total of 58 countries – less than a third of the world's nations – retains the death penalty in practice. A majority of these nations are located in Africa, the Middle East, the Caribbean, and Southern Asia. Only one European nation

(Belarus), one South American country (Guyana), and one North American country (the United States) retain the death penalty (Bohm, 2013), and only two retentionist nations – the United States and Japan – are industrialized. When looking at actual executions, the numbers get even more concentrated. Only 21 nations carried out an execution in 2012. China continues to be far and away the world's leading executioner. Although no official data exist, Amnesty International (2013) surmises that China executes thousands of people every year. The reach of its death penalty today cannot be underestimated: "China is only a single entry in the 197-nation scorecard for recording death penalty status, but estimates since the 1990s suggest it probably accounted for *more than 90%* (emphasis added) of executions worldwide during that period" (Johnson and Zimring, 2009: p. 10).

United States

To claim that the United States is a retentionist nation is slightly misleading. Although there were 39 executions in 2013 and 77 death sentences imposed in 2012, bringing the total number of persons on death row in the United States to 3108, the number of death sentences and executions has been in steady decline. While there was a high of 315 death sentences in 1996 and 98 executions in 1999 (Death Penalty Information Center, 2013), the number of death sentences imposed today is at an all-time low (Liebman and Clarke, 2011; Fleury-Steiner et al., forthcoming). Furthermore, 18 of the 50 states in the United States have abolished the death penalty and two more states have not executed anyone since the Supreme Court lifted a nationwide death penalty moratorium in 1976. Three US states – Texas, Virginia, and Oklahoma – account for the majority of executions carried out since 1976 (Bohm, 2013) and 85% of American counties have not been responsible for a single execution since 1976 (Dieter, 2013). In other words, it can be concluded that the death penalty as practiced in the United States today is a phenomenon confined to a decidedly *localized* minority (Liebman and Clarke, 2011).

A closer look at the United States' death penalty as a phenomenon carried out by a small minority of locales is quite revealing. About 15% of US counties account for all executions carried out since 1976, and a miniscule 2% of the nation's more than 3000 counties account for a majority of its executions (Dieter, 2013). Extensive empirical research documents that prosecutors in these locales are more likely to seek death sentences and more likely to find juries willing to impose death sentences (Berk et al., 2005; Paternoster et al., 2003; Poveda, 2006; Pierce and Radelet, 2005). In other words, who gets sentenced to death is more a function of *where* the crime is committed than a function of the facts of the case. But why these counties? For example, one explanation for the dramatic recent spike in death sentences in New Castle County, Delaware, is the sweeping change to the death penalty process that makes a death sentence far more likely to be imposed. Specifically, since 1991, jurors act only in an advisory role in death penalty cases – a move that caused the number of death sentences and executions in the state to skyrocket (Johnson et al., 2012). Importantly, this change to Delaware's death penalty scheme was catalyzed by a racially charged incident involving four black men who escaped execution for the robbery–murder of two white armored car guards after a single juror refused to vote for a death sentence (Fleury-Steiner et al., 2009). Recent empirical research on Delaware death penalty jurors suggests that this 'advisory' role leads capital jurors to describe themselves as less responsible for imposing the death sentence (Kleinstuber, 2013a) and to give only very cursory evaluations of mitigating evidence (Kleinstuber, 2013b).

More recent studies have used county-level data to confirm the centrality of race in maintaining the death penalty. According to Poveda (2006), counties with more black residents and a long tradition of executions are more likely to sentence people to death and more likely to execute them. Liebman and Clarke (2011) have, moreover, concluded that these counties are more insular and thus are characterized by a more visceral fear of nonwhite outsiders. Indeed, Fleury-Steiner et al. (forthcoming: p. 3) argue that these aggressive death penalty locales are driven by 'an enduring local defense of racial boundaries.'

Despite the few remaining active death penalty locales in the United States, its recent decline nationally cannot be underestimated. In the wake of a myriad of scandals, including the imposition of erroneous death sentences (Gould, 2009) and even wrongful executions in Texas (Liebman et al., 2012), it appears that the United States' appetite for the death penalty has diminished greatly and it is moving in the same direction as the majority of the world's nations. Indeed, six states have abolished the death penalty since 2007 and, as of this writing, numerous others are debating abolition. The US Supreme Court has also recently narrowed the scope of who is eligible to be sentenced to death and in the process, relied upon numerous foreign legal judgments (Dennington, 2006). In 2002, the Court outlawed the execution of the mentally challenged, noting that 'the world community … overwhelmingly disapprove[s]' of the execution of mentally challenged offenders (Atkins v. Virginia, 2002: p. 316, n. 21), and in 2005, it barred the use of the death penalty for offenders who were under the age of 18 years at the time of the crime; the decision noted that the United States was the only nation in the world still permitting juvenile executions (Roper v. Simmons, 2005).

China

According to Miao (2013), the use of the death penalty in modern China may be divided into three relevant periods: the first three decades (1949–79), the Strike Hard era (1980–2003), and the Hu-Wen era (2004–12). The high water mark for executions occurred in this first period under Mao Zedong. Drawing on historical records located in Beijing's Central Press of Historical Records, Miao (2013: p. 3) estimates that a massive 3-year (1950–53) crackdown on counterrevolutionaries resulted in approximately 710 000 executions. Over the next three decades, the country experienced a marked decline in executions. As the country began to adopt capitalist market-based reforms in the 1970s, a pronounced fear of disorder led to the now infamous "Strike Hard" Campaigns. While this period resulted in significantly fewer executions than under Mao Zedong, "approximately 24 000 offenders were sentenced to death and around 1 027 000 were convicted of criminal offences" (Miao, 2013: p. 4). China's postcounter revolutionary

history of state-sanctioned executions occurred in a conflicted politicolegal landscape:

> This pattern of ebb and flow in the past can be explained partly by the constant tension between a utilitarian view that the death penalty is an expedient instrument to achieve short-term political goals and a stricter adherence to the rule of law, penal regulation, and due process when dealing with capital cases. In the most recent Hu-Wen era, such an uneasy relationship in penal ideology takes the form of contradictions between the populist impulses of revenge and retribution and a serious commitment to restrain and civilize the use of capital punishment. (Miao, 2013: p. 4)

Similar to strong popular sentiment that has driven pro-death penalty activity in US locales (Simon and Spaulding, 1998), a robust 'penal populism' (Roberts et al., 2003) seems to be present in China today. In response, in 2010, China adopted *Rules to Exclude Illegal Evidence in Criminal Cases* and *Guidelines to Scrutinize and Analyse Evidence in Capital Cases*, the country's first formal attempts to provide defendants in death penalty cases due process protections. By early 2011, China's President, borrowing from the US Constitution's restriction on cruel and unusual punishment, ordered the country's legal system to adopt its own regulations, *The Eighth Amendment to the Criminal Law*. This new law proposed unprecedented steps to shorten the reach of executions by formally limiting the number crimes punishable by death and by narrowing the death eligibility of defendants over the age of 75 years. In a move doubtlessly inspired by the United States, the President also proposed the goal of adopting lethal injection over shooting as the preferred method of execution.

Recent empirical research involving interviews with 18 national and 18 local legal elites closely involved in the administration of the death penalty in China suggests that these reforms have been slow to take hold in practice (Miao, 2013). On the one hand, it appears as though most national and local-level legal elites support greater regulation of China's death penalty, but these reformist sentiments have not translated into institutional reform on the ground. For one, China's top-down reforms have been largely symbolic. Similar to active death penalty locales in the United States (Bright and Sanneh, 2013), the resources needed to enforce greater regulation at the local level – where death sentences and executions take place – have not been provided. Moreover, there has been resistance to death penalty reform by a majority of Chinese lower level judges and government officials. One of the provincial court judges Miao interviewed also expressed fear of overturning death sentences as betraying the wishes of victims' families and catalyzing public suicides and rioting if the reforms of national legal elites were put into practice: "If I sentence someone to death, my verdict may be overturned. It is easier for the people 'at the top' to talk; but it is us who are stuck in a position between the devil and deep blue sea" (quoted in Miao, 2013: p. 11).

Conclusion

Overall, the European influence on global abolition cannot be underestimated. A majority of countries have either dramatically curtailed the use of executions or banned them altogether. However, Asia has been far less receptive to such efforts (Johnson and Zimring, 2009). The limits of international human rights campaigns seem, at least as of this writing, to have been largely symbolic in China. Indeed, despite its very recent adoption of European-inspired reforms intended to limit the reach of its death penalty, thousands of executions are carried out annually in Chinese provinces in secret (Amnesty International, 2013). Although China has made significant efforts at the national level to limit the use of the death penalty, national elites have not provided needed resources to realize such reforms at the local level. Local elites have, moreover, steadfastly resisted such attempts at systemic change. In this way, future empirical research on China's death penalty generally and other death penalty retentionist countries in particular must "take into consideration a complex web of interactions between actors at the international, national and lower levels" (Miao, 2013: p. 16). Attending to jurisdictional complexity and institutional dynamism is clearly important for future research. Any assessment of a nation's future with the death penalty must attend to the particularities (e.g., politics, legal structures, and religious beliefs) of the death penalty as it is carried out on the ground.

See also: Deterrence Theory: Crime; Human Rights, History of; Punishment: Comparative Politics of; Punishment: Social and Legal Aspects.

Bibliography

Amnesty International, 2013. Death Sentences and Executions 2012. Amnesty International Publications, London. Available at: http://www.amnesty.org/en/library/asset/ACT50/001/2013/en/bbfea0d6-39b2-4e5f-a1ad-885a8eb5c607/act500012013en.pdf.

Atkins v. Virginia, 536 U.S. 304, 2002.

Bedau, Hugo A. (Ed.), 1998. The Death Penalty in America: Current Controversies. Oxford University Press, New York.

Berk, Richard A., Li, A., Hickman, Laura J., 2005. Statistical difficulties in determining the role of race in capital cases: a re-analysis of data from the state of Maryland. Journal of Quantitative Criminology 21, 365–390.

Bohm, R., 2013. The Future of Capital Punishment in the United States. Paper presented at the annual meeting of the American Society of Criminology, Atlanta, GA, 21 November 2013.

Bowers, W.J., 1995. Capital Jury Project: rationale, design, and preview of early findings. Indiana Law Journal 70, 1043.

Bowers, W.J., Steiner, B.D., 1999. Death by default: an empirical demonstration of false and forced choices in capital sentencing. Texas Law Review 77, 605.

Bright, S.B., Sanneh, S.M., 2013. Fifty Years of Defiance and resistance after Gideon v. Wainwright. Yale Law Journal 122 (8).

Death Penalty Information Center, 2013. The Death Penalty in 2013: Year End Report. Available at: http://deathpenaltyinfo.org/documents/YearEnd2013.pdf.

Dennington, A.R., 2006. We are the world? Justifying the U.S. Supreme Court's use of contemporary foreign legal practice in Atkins, Lawrence, and Roper. Boston College International & Comparative Law Review 29, 269–296.

Dieter, R.C., 2013. The 2% Death Penalty: How a Minority of Counties Produce Most Death Cases at Enormous Cost to All. Death Penalty Information Center. Available at: http://deathpenaltyinfo.org/documents/TwoPercentReport.pdf.

Erez, E., 1981. Thou shalt not execute. Criminology 19 (1), 25–44.

Fleury-Steiner, B.D., Dunn, K., Fleury-Steiner, R., 2009. Governing through crime as commonsense racism Race, space, and death penalty reform in Delaware. Punishment & Society 11 (1), 5–24.

Fleury-Steiner, B.D., Kaplan, P., Lonazel, J. Racist localisms and the enduring cultural life of America's death penalty: lessons from Maricopa county, Arizona. Studies in Law, Politics, & Society, forthcoming.

Gould, J.B., 2009. The Innocence Commission: Preventing Wrongful Convictions and Restoring the Criminal Justice System. New York University Press, New York.
Higginbotham, A.L., 1980. In the Matter of Color: Race and the American Legal Process. The Colonial Period, vol. 608. Oxford University Press, Oxford.
Hood, R., Hoyle, C., 2008. The Death Penalty: A Worldwide Perspective, fourth ed. Oxford University Press, Oxford.
Johnson, D., Zimring, F., 2009. The Next Frontier: National Development, Political Change, and the Death Penalty in Asia. Oxford University Press, Oxford.
Johnson, S.L., et al., 2012. The Delaware death penalty: an empirical study. Iowa Law Review 97, 1925–1964.
Kleinstuber, R., 2013a. 'Only a recommendation': how Delaware capital sentencing law subverts meaningful deliberations and jurors' feelings of responsibility. Widener Law Review 19 (2), 321–344.
Kleinstuber, R., 2013b. 'We're all born with equal opportunities': hegemonic individualism and contextual mitigation among Delaware capital jurors. Journal of Qualitative Criminal Justice & Criminology 1, 152–180.
Liebman, J.S., Clarke, P., 2011. Minority practice, majority's burden: the death penalty today. Ohio State Journal of Criminal Law 9, 255.
Liebman, J.S., Crowley, S., Markquart, A., Rosenberg, L., White, L.G., Zharkovsky, D., 2012. The Prosecution of Carlos Deluna. Columbia Human Rights Law Review 43, 908–1138.
Lu, H., Zhang, L., 2005. Death penalty in China: the law and the practice. Journal of Criminal Justice 33 (4), 367–376.
Miao, M., 2013. The politics of China's death penalty reform in the context of global abolitionism. British Journal of Criminology 53 (3), 500–519.
Nagin, D.S., Pepper, J.V. (Eds.), 2012. Deterrence and the Death Penalty. National Academies Press, Washington, DC.
Paternoster, R., Brame, R., Bacon, S., Ditchfield, A., Biere, D., Beckman, K., Perez, D., Strauch, M., Frederique, N., Gawkoski, K., Ziegler, D., Murphy, K., 2003. An Empirical Analysis of Maryland's Death-sentencing System with Respect to the Influence of Race and Legal Jurisdiction. Report to the Maryland Office of the Governor.
Pierce, G.L., Radelet, M., 2005. The impact of legally inappropriate factors on death sentencing for California homicides, 1990–1999. Santa Clara Law Review vol. 46, 1–47.
Poveda, T.G., 2006. Geographic location, death sentences and executions in post-*Furman* Virginia. Punishment & Society 8, 423–442.
Randa, L.E., 1997. Society's final solution: a history and discussion of the death penalty. University Press of America, Lanham, MD.
Roberts, J.V., Stalans, L.J., Indermaur, D., Hough, M., 2003. Penal Populism and, Public Opinion: Lessons from Five Countries. Oxford University Press, New York.
Roper v. Simmons, 543 U.S. 551, 2005.
Schabas, W., 2002. The Abolition of the Death Penalty in International Law. Cambridge University Press, Cambridge.
Simon, J., Spaulding, C., 1998. Tokens of our esteem: aggravating factors in the era of deregulated death penalties. In: Sarat, A.D. (Ed.), The Killing State: Capital Punishment in Law, Politics, and Culture, pp. 81–113.
Slanski, K.E., 2013. The law of Hammurabi and its audience. Yale Journal of Law & the Humanities 24 (1), 3.
Steiker, J.M., 2013. The American death penalty: constitutional regulation as the distinctive feature of American exceptionalism. University of Miami Law Review 67, 329–477.
UN Secretary General, 2009. Capital Punishment and Implementation of the Safeguards Guaranteeing Protection of the Rights of Those Facing the Death Penalty. Available online at: www.unodc.org/documents/commissions/CCPCJ_session19/E2010_10eV0989256.pdf.
Vartkessian, E., 2012. What one hand giveth, the other taketh away: how future dangerousness corrupts guilt verdicts and produces premature punishment decisions in capital cases. Pace Law Review 32 (2).

Relevant Websites

http://www.amnesty.org/en/death-penalty/international-law – Death Penalty in International Law.
http://deathpenaltyinfo.org/ – Death Penalty Information Center.
http://www.deathpenaltyworldwide.org/ – Death Penalty Worldwide.
http://www.amnesty.org/en/death-penalty/endorsement-from-regional-bodies-and-civil-society-to-the-call-for-a-moratorium-on-executions – Endorsement from Regional Bodies and Civil Society to Call for a Moratorium on Executions.
http://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx – International Covenenant on Civil and Political Rights.
http://www.amnesty.org/en/library/info/ACT50/001/2006 – International Standards on the Death Penalty.
http://www.oecd.org – Organisation for Economic Cooperation and Development.
http://www.amnesty.org/en/death-penalty/ratification-of-international-treaties – Ratification of International Treaties.
http://www.un.org/en/documents/udhr/index.shtml – Universal Declaration of Human Rights.
http://www.worldcoalition.org/ – World Coalition Against the Death Penalty.

Decision and Choice: Bounded Rationality

Roy Radner, Stern School of Business, New York University, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The 'Savage paradigm' of rational decision-making under uncertainty has become the dominant model of human behavior in mainstream economics and game theory. 'Bounded rationality' refers to the study of how human decision-makers deal with their cognitive limitations that may prevent them from fully applying the Savage paradigm to real problems without entirely abandoning the notion of rationality. This article sketches the historical roots and current developments of this topic, distinguishing between attempts to extend the Savage paradigm ('costly rationality') and the development of more radical departures.

Introduction

Some kind of model of rational decision-making is at the base of most current economic analysis, especially in microeconomics and the economics of organization. Such models typically assume that economic decision-makers have sufficient cognitive capacities to solve the problems they face, and have preferences and beliefs that are consistent in a rather strong sense. In particular, it is typically assumed that decision-makers (1) do not make logical errors, and can solve any relevant mathematical problems; (2) can process all available information in the time required (including computation, storage, and retrieval); and (3) have an adequate understanding of the decision problems they face, which includes having precise beliefs about the relevant uncertainties and precise preferences among the various consequences of their actions. The economic agent described by such models is sometimes stereotyped by title *homo economicus*.

However, this model has been criticized as inadequate from both normative and descriptive viewpoints. Some strands of this critical movement form the topic known as 'bounded rationality.' This article sketches the historical roots and some current developments of this topic, distinguishing between attempts to extend the standard models and the need for more radical departures. Other strands have been grouped under the topic of behavioral economics, which typically goes beyond assumptions of rationality in economic behavior (*see* Behavioral Economics).

Unease with mainstream models of *homo economicus* was already voiced by J.M. Clark (1918) at the beginning of the twentieth century (see Radner, 2000, p. 631), but the work of Jacob Marschak and Herbert Simon provided the stimuli for a more intense level of activity. The term 'bounded rationality' was coined by Simon:

> Theories that incorporate constraints on the information-processing capabilities of the actor may be called *theories of bounded rationality*. (Simon, 1972, p. 162)

(For a review of empirical evidence of bounded rationality, see Conlisk, 1996.)

The current mainstream theory of rational individual decision-making in the face of uncertainty was elaborated by Savage (1954). This will be called here the *Savage paradigm*, and will be the main starting point of this article. Extensions of this theory to describe rational strategic behavior in multiperson situations are the subject of the *theory of games* (discussed later). The notion of *optimizing* is central to all of these models of rationality.

The general concept of *bounded rationality* covers the two rather different approaches of Marschak and Simon. Marschak emphasized that the Savage paradigm could be extended to take account of costs and constraints associated with information acquisition and processing in organizations, without abandoning the notion of optimizing behavior. This approach will be called here *costly rationality*, and is elaborated later in the section Costly Rationality and the Extended Savage Paradigm.

Simon was also concerned with behavior that could not so easily be interpreted, if at all, as optimizing. However, he retained a focus on the notion of 'rationality.' One might paraphrase his study of bounded rationality as 'How do smart people deal with their cognitive limitations?'; this, until further notice, will be the interpretation of this term that is adopted in this article (see section Bounded Rationality).

However, Simon never gave a precise definition of 'smart people.' Recently, even further departures from the Savage paradigm have received increasing attention in the literature, abandoning to some extent the concept of rationality altogether, under the heading 'behavioral economics.' This topic will not be treated here, but see the entry Behavioral Economics.

Uncertainty

Discussions of rational decision-making – unbounded and otherwise – have been closely tied to uncertainty. The very beginnings of formal probability theory were in part stimulated by questions of how to act rationally in playing games of chance with cards and dice. During the first half of the twentieth century, a number of alternative views were developed concerning the nature of uncertainty, the possibly different kinds of uncertainty, and whether, or in what circumstances, it could be measured (Arrow, 1951; Savage, 1954; Chapter 4). One could be uncertain about natural events, the consequences of action, the laws of Nature, or the truth of mathematical propositions. There was general

(but not universal) agreement that, if uncertainty could be measured (quantified), then that quantification should obey the mathematical laws of probability. However, the 'frequentist school' reserved the legitimacy of probabilistic reasoning for 'experiments' (planned or naturally occurring) that were repeated indefinitely under identical conditions. The 'personalist school,' which included a diverse set of methodologies, argued that the concept of probability was applicable to events outside of the frequentist realm. Some personalists went so far as to deny that the frequentist view could be applied meaningfully to any events at all, that is, that all probability judgments were 'personal' (see Arrow's, 1951 and Savage's, 1954, Chapter 4, accounts of the work of Ramsey, Keynes, Jeffries, Carnap, and de Finetti). In some sense, the personalist view might also be ascribed to earlier authors, such as Bayes and Laplace. The personalist view was given a solid foundation by Savage (1954).

Central to the development of thinking about uncertainty was the simple idea that uncertainty about the consequences of an action could (or should) be traced to uncertainty about the 'state of the world' in which the action would be taken. An essential feature of the concept of the state of the world is that it is beyond the control of the decision-maker in question. A further clarification was provided by the *theory of games* (put forward by J. von Neumann and O. Morgenstern in 1944, and further elaborated by J. Nash, J. Harsanyi, R. Selten, and others). In multiperson decision-making situations in which the participants have conflicting goals, this theory distinguishes between two aspects of the state of the world from the point of view of any single decision-maker, namely (1) the 'state of nature,' which is beyond the control of any of the persons involved, and (2) the actions of the other persons, the latter being called 'strategic uncertainty.' (For material on the theory of games, especially noncooperative games, see Game Theory; Noncooperative Games; Game Theory: Relation to Bayesian Theory.)

The remainder of this article concentrates on a critique of the theory of rational behavior as it is applied to single-person decision-making, or to multiperson situations in which the persons do not have conflicting goals. (For discussions of bounded rationality in a game-theory context, see Rubinstein, 1998; Radner, 2000.)

The Savage Paradigm

A sketch of the Savage paradigm is needed here in order to understand the notions of costly and bounded rationality. (For a systematic treatment, see Utility and Subjective Probability: Contemporary Theories.) The essential building blocks of the model are (1) a set of alternative *states of the world*, or simply *states*, which are beyond the decision-maker's control; (2) a set of alternative actions available to the decision-maker, or as Savage calls them, *acts*; and (3) a set of alternative *consequences*. An act determines which consequence will be realized in each state (of the world). Hence a parsimonious way to think about an act is that it is a function from states to consequences.

The decision-maker (DM) is assumed to have preferences among acts. These preferences reflect both the DM's *beliefs* about the relative likelihood of the different states, and the DM's *tastes* with regard to consequences. A few axioms about the independence of beliefs from consequences enables one to impute to the DM two scales: (1) a *probability measure* on the set of states, reflecting the DM's beliefs; and (2) a *utility scale* on the set of consequences, reflecting the DM's tastes. Using these two scales, one can calculate an *expected utility* for each act, in the usual way, since an act associates a consequence with each state. Thus for each state one calculates the product of its probability times the utility of the associated consequence, and then adds all of the products to obtain the expected utility of the act. One proves that expected utility represents the DM's preferences among acts in the following sense: *the DM prefers one act to another if and only if the first act has a higher expected utility*. (This theorem is sometimes called the *Expected Utility Hypothesis*.) The *rational* DM is assumed (or advised) to choose an act that is most preferred among the available ones, that is, has the highest expected utility; this is the assumption of *optimization*.

The simplicity of this formulation hides a wealth of possible interpretations and potential complexities. First, the axioms of the theory enable the DM to infer preferences among complicated acts from those among simpler ones. Nevertheless, the required computations may be quite onerous, even with the aid of a computer.

Second, if the decision problem has any dynamic aspects, then states can be quite complex. In fact, a full description of any particular state will typically require a full description of the *entire history* of those features of the DM's environment that are relevant to the decision problem at hand.

Third, the description of the set of available acts reveals, if only implicitly, the opportunities for the DM to acquire information about the state of the world and react to it. The laws of conditional probability then determine how the DM should learn from observation and experience. *In fact, this is what gives the Savage paradigm its real 'bite.'* An act that describes how the DM acquires information and reacts to it dynamically is sometimes called a *strategy* (plan, policy). In a model of a sequential decision problem, the space of available strategies can, of course, be enormous and complex. This observation will be a dominant motif in what follows.

The formula for learning from observation is sometimes called *Bayes's theorem*, named after the eighteenth century author, Rev. Thomas Bayes. Hence, the method of inference prescribed by the Savage paradigm is called *Bayesian learning*.

Bayesian learning has an important property in a situation that can be informally described as a sequence of independent experiments carried out under 'identical' conditions an indefinitely large (strictly infinite) number of times. Suppose that in each repetition of the experiment the outcome is uncertain, and its probability distribution depends on the values of some parameters, which are also uncertain. Let the set of states of Nature be partitioned into subsets, such that to each element of the partition corresponds a specification of the parameter values. For example, each experiment consists of a single toss of a (not necessarily fair) coin, with a probability p of the outcome 'heads,' and the parameter p has the same value in all of the experiments. Each element of the partition in question, say $B(p)$, corresponds to a specific value of p. The experimenter begins the sequence of experiments with particular beliefs about p, which we shall call the 'prior distribution' of p, and

after each successive toss of the coin uses Bayes's rule to update the experimenter's beliefs about p, which we shall call the corresponding 'posterior distribution' of p. Under certain conditions on the experimenter's model, which express the idea that the experiments are identical and independent, one can prove that, if the state of Nature is actually in the set $B(p)$, then, *as the number of experiments increases without limit, the sequence of posterior distributions will converge to a distribution that assigns probability one to the set* $B(p)$. Savage calls this 'the approach to certainty through experience.' (For a formal treatment, and references to some relevant literature, see Savage, 1954, 1972, Section 3.6.)

The Simon Critique

As Herbert Simon emphasized in his work, the cognitive activities required by the Savage paradigm (and its related precursors) are far beyond the capabilities of human decision-makers, or even modern human/computer systems, except in the simplest decision problems. This led Simon and his colleagues (especially at Carnegie-Mellon University) to investigate models of human decision-making that are more realistic from the point of view of cognitive demands, and yet do not entirely abandon the notion of rationality. This research also had an impact on the emerging field of artificial intelligence (see Simon, 1972, 1981, and the references cited there).

Savage himself was aware of the problem of bounded rationality, but he nevertheless felt that his model was a useful one for thinking about rational decision-making (Savage, 1954; pp. 16, 17).

Costly Rationality and the Extended Savage Paradigm

As just sketched in the previous section, the Savage paradigm does not appear to take account explicitly of the costs of decision-making. However, nothing prevents the DM from incorporating into the description of the consequences of an act the costs – in terms of resources used – of implementing the corresponding actions. The costly activities involved in decision-making include:

1. observation and experimentation;
2. information processing (i.e., computation);
3. memory; and
4. communication.

The last category may be important when the decision-making process is undertaken by a team of individuals.

If the resources used by these decision-making activities are limited, then those limits may impose binding constraints on the activities themselves, constraints that must be taken into account in the DM's optimization problem. If the constraints are on the *rate of resource use per unit time*, then more extensive decision-making activities may cause *delays* in the implementation of the eventual decisions. To the extent that a delay lowers the effectiveness of a decision (e.g., by making it more obsolete), one may think of delay as an 'indirect cost.' Extending the Savage paradigm to incorporate the costs of decision-making may in some cases be natural, and in other cases problematic. Here the first class of cases is called 'costly rationality.'

The notion that observation is costly was implicit in the Neyman–Pearson theory of hypothesis testing, and was made explicit by Abraham Wald in his pioneering studies of sequential statistical procedures (see Wald, 1950, for an influential codification of his general approach). The cost of observation also figures in more classical (nonsequential) statistical problems such as the design of sample surveys and agricultural experiments. Given some model of the costs of observation, the DM chooses the kind and amount of observation, optimally balancing the expected benefits of additional observations against their costs. Such decision problems fit naturally into the Savage paradigm, although taking account of these costs typically complicates the analysis. For example, in the case of clinical trials and similar problems, the calculation of optimal policies quickly becomes computationally intractable for many problems of realistic size.

Even after the information has been collected, it still must be further processed to produce the required decisions. This information-processing task may be quite demanding. Examples include (1) computing a weekly payroll, (2) scheduling many jobs on many machines, (3) managing multi-product inventories at many locations, and (4) project selection and capital budgeting in a large firm. Such tasks are typically too complex to be handled by a single person, even with the aid of modern computers. In such circumstances the required processing of the information is *decentralized* among many persons in the organization. The theoretical study of decentralized decision-making in an organization whose members have identical goals was introduced by J. Marschak in the *theory of teams* (Marschak and Radner, 1972).

Computer science has provided a number of useful models of information processing by both computers and humans, and the decentralization of information processing in human organizations finds its counterpart in the theories of parallel and distributed processing in computer systems. T. A. Marschak and C. B. McGuire (1971) were probably the first to suggest the use of a particular model of a computer (the finite automaton) to represent the limited information-processing capabilities of humans in economic organizations. S. Reiter and K. R. Mount (2002) were early contributors to this line of research, and went further in analyzing economic organizations as networks of computers. (For more recent developments, see Radner, 1992, 2000; Van Zandt, 1998, and the references cited there.) One conclusion from this literature is the *Iron Law of Delay* for networks of processors of bounded individual capacity. This law can be paraphrased in the following way: "As the size of the information-processing task increases, the minimum delay must also increase unboundedly, even for efficient networks and even if the number of available processors is unlimited" (Radner, 1992, and the references cited there).

Memory storage and communication among humans and computers are also resource-using activities, and cause further delays in decision-making. Both the storage and transmission of information and the results of information processing seem to be relatively 'cheap' compared to observation and processing, at least if we consider computer-supported

activities. The proliferation of large databanks, and the flood of junk mail, telephone calls, and e-mail, lend support to this impression. It appears that today it is much cheaper, in some sense, to send, receive, and store memos and papers than it is to process them. (For game-theoretic models of players with limited memory, see Rubinstein, 1998. For models of costly communication in organizations, and some implications for organizational structure, see Marschak and Reichelstein, 1998.)

Bounded Rationality

Many real decision problems present difficulties that prevent the DM from using the Savage paradigm. Among these difficulties are:

1. inconsistency;
2. ambiguity;
3. vagueness;
4. unawareness; and
5. failure of logical omniscience.

As will be seen, these difficulties are somewhat related and overlapping. In particular, it is difficult to distinguish in practice between ambiguity and vagueness.

Regarding inconsistency, Savage (1954, p. 570) wrote:

> According to the personalistic view, the role of the mathematical theory of probability is to enable the person using it to detect inconsistencies in his own real or envisaged behavior. It is also understood that, having detected an inconsistency, he will remove it. An inconsistency is typically removable in many different ways, and the theory gives no guidance for choosing.

Some inconsistencies have been observed so frequently, and have been so 'appealing' – such as the so-called 'Allais paradox' and 'Ellsberg paradox' – that they have been used to criticize the Savage axioms and to form a basis for a somewhat different sets of axioms (*see* Measurement Theory: Conjoint). In other cases, it has been argued that inconsistent preferences arise because the DM is forced to articulate preferences about which he is not 'sure.' (This explanation is related to vagueness, discussed in a moment.) In particular, this unsureness may be related to uncertainty about what are the states of the world, a circumstance that can lead to a preference for flexibility (also discussed in a moment). Finally, it has been observed in experiments that inconsistencies are more frequent the closer the alternatives are in terms of preference. This observation led J. Marschak and others to elaborate models of 'stochastic choice' (Mcfadden, 1999); (*see* Decision and Choice: Random Utility Models of Choice and Response Time).

Allusion has already been made to the DM's possible vagueness about preferences. However, the DM could also be vague about any aspect of his model of the decision problem, and is likely to be so if the problem is at all complex. Vagueness can be about the *interpretation* of a feature of the model, about its *scope*, or both.

The problem of vagueness of probability judgments is related to deeper issues about the meaning of probability itself, and in particular to the debate about objective versus subjective interpretations of probability. Strictly speaking, the Savage paradigm is based on a subjective (or personal) view of probability, although there are extensions that admit both interpretations in the same model. Savage recognized the problem, and proposed a solution he called *minimax regret*. This topic will be discussed in the section Minimax Regret and Related Methods.

At a moment of time, the DM may be unaware of some aspect of the problem; for example, the DM may be unaware of some future contingencies that could arise, or of some actions that are available to the DM. This phenomenon is particularly interesting if the DM is aware of the possibility that he or she may be unaware of something. For example, if the DM anticipates that in the future he or she will become aware of new acts of which he or she is currently unaware, then the DM may prefer present actions that allow for 'flexibility' of choice in the future. This idea was formalized by T. C. Koopmans, D. Kreps, and others (Kreps, 1992). For other theoretical treatments of unforeseen contingencies, see Dekel et al. (1998).

As a consequence of the preceding considerations, decision theorists recognize that it is impossible for a DM to construct a complete model of his or her 'grand decision problem,' that is, for the DM's whole life! A common research strategy is to suppose that the DM can break up the grand decision problem into subproblems that can be solved independently without (much) loss of utility. Savage called this the device of constructing 'small worlds,' but showed that the conditions for this to be done without loss are unrealistically stringent (Savage, 1954; pp. 82–91).

Finally, I come to what is perhaps the most difficult aspect of truly bounded rationality. Up to this point it has been assumed– if only implicitly – that the DM has no difficulty performing mathematical calculations or other logical operations. In particular, having formulated a decision model, the DM will be able to infer what it implies for his or her optimal strategy. As has already been pointed out, this assumption is absurd, even for small-world models, except for 'Mickey Mouse' problems that are constructed for textbooks and academic articles. The crux of the matter is that in any even semirealistic decision problem, *the DM does not know all of the relevant logical implications of what he or she knows*. This phenomenon is sometimes called *the failure of logical omniscience*. Examples of the failure of logical omniscience include the following:

1. A DM who knows the axioms of arithmetic is uncertain about whether they imply that 'the 123rd digit in the decimal expansion of pi is 3,' unless he has a long time to do the calculation and/or has a powerful computer with the appropriate software.
2. Twenty years ago, a DM who knew the axioms of arithmetic was still uncertain about whether they imply Fermat's Last Theorem.

The following examples are closer to practical life, and possibly more intimidating:

1. Given all that a DM knows about the old and new drugs for treating a particular disease, what is the optimal policy for conducting clinical trials on the new ones?

2. Given all that is known, theoretically and empirically, about business organizations in general, and about telecommunications and AT&T in particular, should AT&T reorganize itself internally, and if so, how?

Savage (1954, p. 7fn) commented on this class of problems:

> The assumption that a person's behavior is logical is, of course, far from vacuous. In particular, such a person cannot be uncertain about decidable mathematical propositions. This suggests, at least to me, that the tempting program sketched by Polya of establishing a theory of the probability of mathematical conjectures cannot be fully successful in that it cannot lead to a truly formal theory.

(For further discussion and references, see the 1972 edition of Savage's book.)

In spite of some interesting efforts (see, e.g., Lipman, 1995), it does not appear that there has been significant progress on what it means to be rational in the face of this kind of uncertainty.

Satisficing, Heuristics, and Non-Bayesian Learning

In view of the difficulties posed by the various manifestations of 'truly bounded rationality,' a number of authors have proposed and studied behavior that departs more or less radically from the Savage paradigm. These will be discussed under three headings: satisficing, heuristics, and non-Bayesian learning.

The term 'satisficing' refers to behavior in which the DM searches for an act that yields a 'satisfactory,' as distinct from an optimal, level of expected utility. The target, or satisfactory, level of expected utility is usually called the DM's *aspiration level*. In the simplest model, the aspiration level is 'exogenous,' that is, a given parameter of the model. More ambitious models describe some process whereby the aspiration level is determined within the model, and may change with experience (Simon, 1972; Radner, 1975). Such aspiration levels are called 'endogenous.' In some problems even optimal behavior bears a resemblance to satisficing. One example is the 'secretary problem' (Radner, 2000).

The term 'heuristics' refers generally to behavior that follows certain rules that appear to produce 'good' or 'satisfactory' results most of the time in some class of problems (Simon, 1972; *Heuristics for Decision and Choice*). For example, the calculation of an optimal schedule for assigning jobs to machines is typically intractable if the numbers of jobs and machines are even moderately large. Nevertheless, human schedulers routinely construct 'satisfactory' schedules with such numbers, using various rules-of-thumb that have been developed with experience. Heuristics are central to many artificial intelligence applications. Satisficing plays an important role in many heuristic methods, and also in the processes of their modification.

The discussion of heuristics leads naturally to the consideration of *non-Bayesian learning* (NBL) (Cf. the discussion of Bayesian learning earlier in the article). Many standard statistical methods use NBL. For example, the use of the sample mean to estimate a population mean is typically inconsistent with the Savage paradigm (although in some cases the latter can be shown to be a limit of Bayesian estimates as some parameter of the problem goes to infinity). Most psychological theories of learning postulate some form of NBL. A central question in the theory of NBL is: under what conditions, if any, does a particular NBL procedure converge asymptotically to a procedure that is Savage-paradigm-optimal as the DM's experience increases? (Rustichini, 1999).

Attempts to Define Bounded Rationality

This section briefly examines a few procedures that have been considered by some to provide a 'rational' way to deal with some of the cognitive limits faced by decision-makers, especially in the context of uncertainty.

Minimax Regret and Related Methods

Several approaches are based on estimates of how much is lost by not using an 'optimal' strategy. Using the terminology of the Savage paradigm, let $u(f, s)$ be the utility of an act f in a state s, and define the corresponding regret $r(f, s)$ to be the difference between the maximum utility possible in that state minus the utility of f in that state.

The first strategy we examine is a response to the situation in which the DM is vague about the appropriate prior distribution of the state of Nature – or *prior*, for short. Let $R(f)$ denote the maximum of $r(f, s)$ over all states of Nature, let R^* be the minimum of $R(f)$ over all acts, and let f^* be an act that attains R^*. (Assume that the relevant maxima and minima exist.) The *minimax regret rule* prescribes that the DM should choose an f^*, *provided that* R^* *is not too large*.

The minimax rule is typically expanded to include randomization: (1) the set of states is replaced by the set of probability distributions on the set of states, that is, by the set of possible priors; and (2) the set of acts is replaced by the set of probability distributions on the set of acts. Note that this expansion transforms the decision problem into a zero-sum game against 'Nature.' Savage (1954, pp. 169–171) also argued against using negative utility instead of regret (Radner and Marschak, 1954).

Savage ascribed the minimax rule to Abraham Wald (1950), although Wald did not motivate it in a formal way. Indeed, Savage himself stated: "A categorical defense of the minimax rule seems definitely out of the question… On the other hand, there are practical circumstances in which one might well be willing to accept the rule… It is hard to state the circumstances precisely, indeed they seem vague almost of necessity" (Savage, 1954; Section 9.7).

(Savage used the term 'loss' instead of 'regret.' He also proposed the minimax rule for statisticians, but for a different reason, which is beyond the scope of this article to discuss.)

More recently, several authors have proposed sets of axioms to rationalize the minimax regret rule and variations of it. In some cases, authors use the term 'ambiguity' instead of 'vagueness,' although the distinction is sometimes difficult to discern (Stoye, 2011, 2012 contains recent contributions and previous references). For recent theories of updating such

rules in the light of new information, see Hanany and Klibanoff (2009). In particular, some authors have proposed a mixed model, following ideas of Frank Knight (Knight, 1921), distinguishing 'risk' from 'uncertainty,' with the former objectively quantifiable, and the latter subjective or even not quantifiable. DMs are sometimes modeled as having the analog of two probability distributions and two utility functions, one pair of scales appropriate for risk and the other for uncertainty (ambiguity). However, it can be argued that, in the presence of vagueness about beliefs, it is paradoxical to assume that the DM can articulate two separate pairs of scales as above.

This would leave minimax regret as the most promising candidate for rational behavior in the presence of vagueness of beliefs. Related ideas are strategies for which the DM can compute an upper bound on the regret that is perceived, in some sense, as not too large.

Asymptotic Rationality

Procedures that fall outside of the Savage paradigm are often justified as converging to optimality with the accumulation of experience. For example, NBL statistical estimation procedures are valued if they are *consistent*, meaning roughly that the estimates of a parameter converge to the 'true value' of the parameter as the sample size increases without limit. Under the conditions that guarantee the approach to certainty of Bayesian methods (discussed earlier), this implies that the NBL estimator is asymptotically optimal in the Savage sense. Another example is asymptotically optimal model revision in the case of vagueness about specification of the model of the decision problem (Radner, 2009; Rustichini, 1999). Thus asymptotic rationality is a promising candidate for bounded rationality.

Survival and Evolution

Survival in an individual or evolutionary sense is sometimes taken to be a sign of – or definition of – rationality. For an individual, survival is a matter of preference. A long, miserable life is not necessarily preferable to a short, happy one. Discussion of evolution as a criterion of rationality would take us beyond the scope of this article, which has been focused individual decision problems (see, however, Dutta and Radner, 1999; Smith, 2008).

Conclusions

This incomplete survey has listed a number of ways that limited cognitive abilities prevent the full applicability of the Savage paradigm of rationality in individual and team decision problems under uncertainty. These limitations gave rise to the study of the stimulating but vaguely defined concept of bounded rationality. Some of these limitations could be dealt with under the rubric of costly rationality, without departing from the Savage paradigm itself. Others forced us to go beyond the Savage paradigm, but thus far no clear-cut concept of rationality has emerged from the study of responses to these limitations, although a few look promising. In particular, no persuasive rational response to the failure of logical omniscience seems to have been discovered. Hence the concept of bounded rationality still seems to be in need of a definition.

Bibliographic Notes and References

Many important works on bounded rationality have been omitted from the bibliography because of space limitations. The references cited in the body of the article have historical interest, provide an overview of a topic discussed, and/or provide a key to other literature. The following provide additional references and information on the application of notions of bounded rationality in economics and management: Arrow (1974); Binmore (2009); Gintis (2009); Hansen and Sargent (2007); Radner (2000); Shapira (1997); Smith (2008); and Van Zandt (1999).

Acknowledgments

Some of the material in this article is a condensation of Radner (2000). I thank the Oxford University Press for permission to use this material here.

Bibliography

Arrow, K.J., 1951. Alternative approaches to the theory of choice in risk-taking situations. Econometrica 19, 404–437.
Arrow, K.J., 1974. The Limits of Organization. Norton, New York.
Binmore, K., 2009. Rational Decisions. Princeton University Press, Princeton and Oxford, NJ.
Clark, J.M., 1918. Economics and modern psychology. Journal of Political Economy 26, 1–30.
Conlisk, J., 1996. Why bounded rationality. Journal of Economic Literature 34, 669–700.
Dekel, E., Lipman, B., Rustichini, A., 1998. Recent developments in modeling unforeseen contingencies. European Economic Review 42, 523–542.
Dutta, P.K., Radner, R., 1999. Profit-maximization and the market selection hypothesis. Review of Economic Studies 66, 769–798.
Gintis, H., 2009. The Bounds of Reason. Princeton University Press, Princeton and Oxford, NJ.
Hanany, E., Klibanoff, P., 2009. Updating ambiguity averse preferences. B. E. Journal of Theoretical Economics 9, 1–51.
Hansen, L.P., Sargent, T.J., 2007. Recursive robust estimation and control without commitment. Journal of Economic Theory 136, 1–27.
Knight, F., 1921. Risk, Uncertainty, and Profit. Houghton-Mifflin, Boston.
Kreps, D., 1992. Static choice in the presence of unforeseen contingencies. In: Dasgupta, P., et al. (Eds.), Economic Analysis of Markets and Games. MIT Press, Cambridge, MA, pp. 258–281.
Lipman, B., 1995. Decision Theory without Logical Omniscience: Toward an Axiomatic Framework for Bounded Rationality. Queens University, Canada, Unpublished.
Marschak, J., Radner, R., 1972. Economic Theory of Teams. Cowles Foundation and Yale University Press, New Haven, CT.
Marschak, T.A., McGuire, C.B., 1971. Lecture Notes on Economic Models for Organizational Design. University of California, Berkeley, CA.
Marschak, T.A., Reichelstein, S., 1998. Network mechanisms, informational efficiency, and hierarchies. Journal of Economic Theory 79, 106–141.
McFadden, D., 1999. Rationality for Economists? Journal of Risk and Uncertainty 19, 73–105.
Mount, K., Reiter, S., 2002. Computation and Complexity in Economic Behavior and Organization. Cambridge University Press, Cambridge, MA.
Radner, R., 1975. Satisficing. Journal of Mathematical Economics 2, 253–262.
Radner, R., 1992. Hierarchy: the economics of managing. Journal of Economic Literature, 1382–1415.

Radner, R., 2000. Costly and bounded rationality in individual and team decision-making. Industrial and Corporate Change 9, 623–658.
Radner, R., 2009. 'Bayesian' Model Revision for K'th Order Markov Chains with Unknown K. IOMS Department of New York University, Unpublished.
Radner, R., Marschak, J., 1954. Notes on some proposed decision criteria. In: Thrall, R., Coombs, C., Davis, R. (Eds.), Decision Processes. Wiley, New York, pp. 61–69.
Rubinstein, A., 1998. Modeling Bounded Rationality. MIT Press, Cambridge, MA.
Rustichini, A., 1999. Optimal properties of stimulus-response learning models. Games and Economic Behavior 29, 244–273.
Savage, L.J., 1954. The Foundations of Statistics. Wiley, New York.
Savage, L.J., 1972. The Foundations of Statistics, second ed. Dover, New York.
Shapira, Z. (Ed.), 1997. Organizational Decision Making. Cambridge University Press, Cambridge, UK.
Simon, H.A., 1972. Theories of bounded rationality. In: McGuire, C.B., Radner, R. (Eds.), Decision and Organization. North-Holland, Amsterdam, pp. 161–176. second ed., University of Minnesota Press, Minneapolis, MA, 1986.
Simon, H.A., 1981. The Sciences of the Artificial, second ed. MIT Press, Cambridge, MA.
Smith, V.L., 2008. Rationality in Economics. Cambridge University Press, Cambridge UK.
Stoye, J., 2011. Axioms for minimax regret choice correspondences. Journal of Economic Theory 146, 2226–2251.
Stoye, J., 2012. New perspectives on statistical decisions under ambiguity. Annual Review of Economics.
Van Zandt, T., 1988. Organizations with an endogenous number of information processing agents. In: Majumdar, M.K. (Ed.), Organizations with Incomplete Information. Cambridge University Press, Cambridge, UK, pp. 239–305.
Van Zandt, T., 1999. Real-time decentralized information processing as a model of organizations with boundedly rational agents. The Review of Economic Studies 66, 633–658.
Wald, A., 1950. Statistical Decision Functions. Wiley, New York.

Decision and Choice: Economic Psychology

Stephen EG Lea, University of Exeter, Exeter, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Economic psychology is the interdisciplinary study of the economic behaviors of individuals, and of the impact of the state and nature of the economy, and individuals' positions within it, on individual behavior and mental life. Economic psychologists have focused their attention on questions such as the psychology of saving, debt, and investment; the psychology of money and taxation, both in terms of individuals' compliance with or avoidance of tax, and the psychological impact of tax systems; and the psychological impacts of major economic changes such as the recent financial crisis, the recognition of the ecological unsustainability of some current economic models.

Introduction

Economic psychology is best defined by what economic psychologists do. Typically, they are interdisciplinary social scientists who recognize that the economy is an important field of human behavior and that human beings are important agents in the economy. Those may seem obvious, even banal, propositions. However, a consideration of the most influential theories and schools in psychology would show that they have often ignored economic behavior, in favor of looking at intimate relationships, emotional development, or very general processes. Conversely, it is now almost a cliché that for most of the twentieth century economic theory was dominated by mathematical approaches that abstracted from the living human economic agent, relying on assumptions that people would, to a good enough approximation, do whatever was economically best for them.

Both those approaches have exploded spectacularly in the final decades of the twentieth century and the first decade of the twenty-first century. This has come about partly for academic reasons: social scientists from both economics and psychology have started to take a serious interest in each other's disciplines. More important, however, has been the occurrence or projected occurrence of a series of crises in the economic world: The shift of economic power from North America and Europe toward Asia; the growing realization that the current model of economic policy is unsustainable ecologically; the demographic changes that mean that many, perhaps most, in the traditional developed world cannot rely on having an adequate income for what has been thought of as normal retirement; and, most recently, the prolonged and widespread recession triggered by excessive reliance on credit by governments and financial institutions' promotion of unsustainable loans to individuals. All these are economic changes that have been widely perceived to have profound causes within individual behavior, and profound consequences of individuals' lives, including their mental lives. These developments therefore make it painfully obvious that the economy matters to individuals and individuals matter to the economy.

To put this more formally, economic psychology recognizes that (1) the economy powerfully influences individuals' lives, at the psychological level of feelings, thoughts, and behavior, (2) individuals' feelings, thoughts, and behavior are what make up economic life, and (3) there are many problems, both academic and practical, to which economics and psychology can both contribute despite their very different approaches and explanatory styles.

Relation of Economic Psychology to Other Research Areas

Economic psychology is not a branch of psychology, but the interdisciplinary study of the interface – or, as it tended to be for much of the twentieth century, the gap – between economics and psychology. One of the most powerful and successful ways of looking at that interface has been through the study of economic decisions, particularly in the area of risky choice, limitations on rationality (see Decision and Choice: Costly and Bounded Rationality) and intertemporal choice. So an entry on economic psychology is fittingly placed in this volume, alongside discussions of numerous other aspects of decision theory. However, it is equally strongly linked to articles in this encyclopedia devoted to economics, to developmental, social, personality, and motivational psychology, to organizational and management studies (for example, Customer Satisfaction or Marketing Strategies), and so on.

Even when economics and psychology had the least interest in each other during the mid-twentieth century, a few very distinguished researchers moved freely between the two disciplines, and have been recognized and respected in both disciplines. However, many of them would be more likely to identify themselves as decision theorists than as economic psychologists: examples include Herbert Simon, Amos Tversky (see Tversky, Amos (1937–96)), Richard Thaler, and George Loewenstein. It is true in general that many more people practice economic psychology than call themselves economic psychologists. Very similar approaches are taken not only in decision science but also under other interdisciplinary names, such as behavioral economics, consumer science, or socioeconomics, and within particular branches of both psychology and economics (see Behavioral Economics; Experimental Economics). Individual researchers may designate themselves in different ways within this spectrum, but they mix freely with others who would describe themselves differently, and the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.43029-1

institutions of research such as learned societies and their conferences and journals draw on all of them. Because so many of these areas of research are represented elsewhere in this encyclopedia, this article will focus on the work of those who have explicitly thought of themselves as economic psychologists.

Of these various terms describing interdisciplinary research between psychology and economics, 'economic psychology' is probably the oldest: it was used, probably for the first time, in the title of a book by the influential French social psychologist Gabriel Tarde (1902). However, the modern development of economic psychology started from the work of the Hungarian-American Gestalt psychologist George Katona in the period immediately after World War II. Katona's work is summarized in his book, *Psychological Economics* (Katona, 1975). Recent surveys of current research include Altman (2006) and Lewis (2008); it is significant that both placed research by economic psychologists side by side with work by decision theorists and experimental and behavioral economists.

Although there are certainly no barriers between those who describe themselves in these different ways, there are certainly different nuances associated with the different descriptions. To locate the term more accurately, therefore, it is worth considering what differences exist between people describing themselves as economic psychologists and their colleagues who would describe themselves in other ways. It should be emphasized, however, that these are trends only, and many exceptions could be found.

Unsurprisingly, compared with those who call themselves behavioral economists, those who call themselves economic psychologists are more likely to have had their initial training in psychology, rather than economics, and to have worked in psychological institutes of one sort of another, rather than in economics departments. This is not just a matter of personal history; they are more likely than behavioral economists to make use of general psychological theories and methodologies. At the theoretical level, they are particularly likely to use cognitive social psychological theories such as Fishbein and Ajzen's theories of reasoned action and planned behavior, or more recent ideas such as social identity or self-regulation theory. They may use other theories as well: evolutionary theories have been particularly popular in recent decades, for example, but these are equally popular among interdisciplinary researchers of other orientations. Methodologically, although there has been a huge increase in the use of experimental studies of social dilemmas such as the ultimatum game and trust games, economic psychologists still often use questionnaire surveys or depth interviews followed by qualitative analysis. Furthermore, if they use experiments, these may well involve asking participants to imagine themselves into a scenario, and may well not implement incentive compatibility, rather than the very bare but always incentive-compatible approaches favored by experimental economists.

Compared with those who call themselves decision theorists, economic psychologists are less likely to be concerned with exact or mathematical models of people's choices, and in particular, they are less likely to be concerned with the minutiae of rationality or optimality models, let alone with submitting such models to empirical test. They are typically more concerned with descriptive analyses of actual economic choices, often choices that are part of everyday economic life, than with mathematical analyses of more abstract choices in experimental or other artificial situations. They are also likely to call on a wider range of psychological constructs: decision theory largely draws on and influence concepts within cognitive psychology, and indeed many of its leading exponents would be happy describing themselves as cognitive scientists. Economic psychology, however, is concerned also with the economic role of emotions; recent research has been looking at the economic sources and defects of anger, guilt and shame, and pride and self-esteem.

Quantitatively speaking (that is, in terms of their contributions to household income or expenditure, and the hours people spend engaged in them), the most important kinds of economic behavior are working and buying. These are so significant that they have their own branches of both psychology and (to some extent, at least) economics. Within psychology, the study of work behavior falls within the field variously known as occupational psychology, industrial and organizational psychology, or work psychology, while the study of buying behavior falls within consumer psychology. Within economics, the study of work behavior falls within labor economics. The study of buying behavior offers an interesting example of the problems of twentieth-century economics. In principle, it should fall within microeconomics, and specifically within demand economics. However, for most of the twentieth century, this field was so focused on abstract mathematical theory, and so divorced from the study of actual purchasing behavior, as to be of no use to anyone interested in what people actually buy and why. Empirical economic data on buying behavior therefore have to be sought elsewhere, within econometrics, applied especially to agricultural economics, and various areas of management studies particularly those devoted to retail management.

Logically, all these fields of study are subfields of economic psychology as defined here. It is certainly true that experts in all these fields can be found within the research institutions of economic psychology, and that people calling themselves economic psychologists have carried out much research on both work (e.g., Nyhus and Pons, 2005) and buying (e.g., Dittmar, 2008). However, economic psychology as a discipline would certainly not be so arrogant as to claim these much larger groups of colleagues as its subordinates (cf. Lea, 1992). Furthermore, there are some typical differences in approach to the subject matter of work or buying between people calling themselves economic psychologists and people calling themselves, say, consumer psychologists or labor economists. First, an economic psychologist is more likely to be interested in work or buying as examples of economic behavior in general, and to ask how they illuminate general questions about the relation between the individual and the economy. Second, these more specific fields do typically lie firmly within either economics or psychology: consumer psychologists are clearly psychologists and have little to do with microeconomists or econometricians, and labor economists are clearly economists and have little to do with occupational psychologists. An economic psychologist looking at work or buying, however, is almost inevitably going to do so from an interdisciplinary perspective. Of course, there are some exceptions to these broad

trends: for example, compulsive buying, logically a topic within consumer psychology, has disproportionately been studied by people who would be more likely to describe themselves as economic psychologists, while unemployment, logically a topic within occupational psychology, has also been more of a concern within economic psychology.

Finally, compared with all these other kinds of work that draw on both economics and psychology, economic psychologists have been more likely to be interested in the two-way interaction between individuals and the economy. It is common ground among economic psychology, behavioral economics, and decision theory that the economic decisions of individuals are interesting, that they are at least in part predictable, and that they have serious consequences for the behavior of the economy as a whole. However, it is only within economic psychology that one finds much study of the converse direction of causation – a consideration of the impact the economy as a whole has on the individual. As we turn to a consideration of what, specifically, economic psychologists do, we will do so in two parallel sections: one concerned with the economic behaviors of individuals, the other with the economy as an independent variable acting on individuals' psychology.

Economic Behaviors of Individuals

Economic behaviors include, for example, work, buying, saving, borrowing, giving, and gambling. Several of these have been subjects of intense research within psychology. We have already seen that both work and buying have whole areas of both psychology and economics devoted to them. It is no accident, therefore, that many economic psychologists have concentrated on economic behaviors such as saving, borrowing, giving, and paying (or avoiding) tax, which are left aside by the established subdisciplines.

The study of saving, debt, and investment has been one of the most popular research areas for economic psychologists, for a number of reasons. First, it is macroeconomically important. Katona's consumer surveys focused on how much of their incomes consumers intended to spend rather than save, since quite small changes in this 'propensity to consume' can tilt a modern economy into recession or inflation. Second, the problem of optimal intertemporal choice is theoretically interesting and difficult: economists have often had to appeal to (usually speculative) psychological principles to reconcile their theories with some obvious macroeconomic facts. Third, it is important in the lives of individuals: making adequate provision for future life changes such as children's education or one's own retirement is a difficult and complex matter, while getting involved in serious debt or bankruptcy can be a personal disaster. Finally, the problems have proved tractable to psychological investigation. Wärneryd (1999) has summarized extensive research that has given us a thorough understanding of the psychological processes leading some people to save more than others. Recently, the study of intertemporal choice has become a key field within behavioral economics, from the point of view of investigating individuals' choices in experimental situations and that of seeking ways of encouraging people to make better choices in real-life situations (see Thaler and Sunstein, 2008). However, it has also spread into new areas of psychology, particularly with the involvement of techniques for exploring the areas and processes of those involved in particular choices, using techniques such as electroencephalography, functional magnetic resonance imaging, and most recently pharmacological intervention. These approaches, often collectively referred to as neuroeconomics (see Glimcher, 2010), have been applied to other kinds of choices (for example, those in social dilemmas) as well as intertemporal choice.

The study of debt does not have such a long history in economic psychology as the study of saving, but in recent years it has become more and more prominent. Like the decision to save, the decision to borrow is an intertemporal choice, but from the research that has been done on such fields as serious household debt, bankruptcy, credit card use, and purchase through installment credit (hire purchase), it has become clear that the psychological processes involved are far from being mirror images of each other. In many countries a particular sphere of research has been the study of debt and borrowing by university students, as governments have sought to transfer more of the financing of higher education from taxation to loans. The investigation of the psychology of debt will obviously be given new impetus by the role that unsustainable personal borrowing played in the bank crashes of 2007 onwards, and the higher levels and pervasiveness of debt that are likely to be seen in the ensuing time of economic hardship in many Western economies.

If debt is logically the opposite of saving, investment is its logical consequence: anyone who uses a financial institution for the purposes of saving is also making an investment. The psychology of investment is not well developed. There has been some research on the influence of investor psychology on movements within stock markets; Wärneryd (2001) provides a summary. However, the most thorough research has been on something that is, in terms of the financial markets as a whole, a niche area, namely, ethical investment. This has been of concern partly for practical or political reasons, because it is seen as something that should be encouraged, and partly for theoretical reasons, since it bears on the apparently vexed question of altruistic behavior. This is more fully considered within the study of the psychology of giving.

The psychology of giving has attracted a great deal of popular attention. It is theoretically interesting because of its relation to the whole question of altruistic behavior, which is at least superficially paradoxical within both economic and evolutionary theories. The paradox is at least partly resolved by noting that gifts illustrate in a vivid way a general principle: every gift is both an economic and a social transaction. Within any society there are well-understood, although usually unstated, social rules about what may be given to whom and on what occasions. For example, despite the longstanding protests by economists that money is actually the most efficient kind of gift, so that the practice of choosing objects to give to someone else results in a substantial deadweight loss, money is not a socially acceptable gift in all contexts. For example, in the United Kingdom, young adults typically do not find the idea of giving money to their mothers as a birthday present acceptable. Interestingly, the objections are

felt more strongly by the potential donor than by the potential recipient.

Almost all economic behaviors involve money in some way. A central problem for economic psychologists has therefore been the psychology of money itself. The problem is more acute because modern 'fiducial' money is inherently a psychological creation: it is valuable only because we all believe it is. Economic psychologists have investigated what makes 'good' money, from the kinds of coins that look and feel valuable to people, to the kinds of electronic payment that do not leave people feeling they are out of control of their spending. Transitions to new kinds of money have also been investigated; this concerns not only the Euro but also the rebasing of currencies in response to serious inflation, which on a world scale happens quite frequently. Economic psychologists have also investigated some of the unexpected behaviors people show toward money, for it seems to be a general rule that money attracts a certain amount of behavior that is disproportionate to its actual economic usefulness. Lea and Webley (2006) have reviewed much of the recent literature on the psychology of money, in the process of arguing that some aspects of human behavior toward money cannot be explained by its instrumental (or, as they put it, 'tool-like') properties, but instead require us to see money as a kind of drug, which gives some of the same gratification as biologically relevant stimuli without fulfilling any real biological need.

An important aspect of the psychology of money is the psychology of money management. What makes some people better than others at the task that faces nearly all of us, living in a world where the supply of desirable goods greatly exceeds the income available to buy them? In multiperson households, how are resources shared, and who determines the economic choices that are made on behalf of everyone? In recent years this aspect of family economics has been extended from traditional nuclear families to consider a wider range of household arrangements, such as the blended families that follow divorce and remarriage.

Cutting across all these fields, but commanding ever-increasing attention, are environmental and other ethical concerns. Individual economic behaviors are involved in concerns about the sustainability of the current Western economic model in two ways. On the one hand, there are many individuals who want to adjust their personal economic behavior so as to reduce their impact on planetary ecology (and perhaps to influence other individuals, and the political process, to do the same). They may do this in a variety of ways, all of which have been studied by economic psychologists. Studies of ethical investment have already been mentioned, but there are also investigations of ethical (usually environmentally oriented) workplace behavior, consumption, and giving. It has been the general argument of this section of the article that the economic behaviors of individuals are relevant to the behavior of the economy as a whole. Indeed, from the point of view of the traditional doctrine of consumer sovereignty, they are not just relevant but determinative: as a Ford Motor Company advertisement once put it, "Everything we do is driven by you." For the most part the assumption is that, however economically important individuals' choices are, we neither know nor care about their macroeconomic impact. When people make ethical economic choices, however, they are typically deliberately trying to bring about a particular economic effect.

The Impact of the Economy on Individuals' Psychology

At any moment, within any given economy, different individuals occupy different positions, from millionaire to beggar; even the most stable economy undergoes unpredictable macroeconomic changes, from growth to recession, and sometimes many economies do so in parallel; some economies undergo massive transformations, for example, the 1990s transition from state socialism to market capitalism in Eastern and Central Europe and the former Soviet Union, and the rather different development of a capitalist model within the People's Republic of China; within economies, individual industries and industrial sectors wax and wane; and across the globe there are huge variations between countries in people's average earnings, spending power, and standards of living. These many different dimensions of economic variation must have profound psychological consequences, yet the economy is traditionally ignored as an independent variable within psychology. Economic psychologists have sought to repair this omission.

A popular methodology has been to look at how children learn to function within the particular economy they are born into. This is known as the study of 'economic socialization' (Furnham, 2008). Many earlier studies took a cognitive stance, investigating how children's understanding of economic concepts improves with age. More recent work has tended to look at children's actual economic behavior, and have led us to realize that, as well as interacting directly with the adult economy, children have at least two other kinds of economic life. They are potent influences on the economic behavior of their entire households, both through their unarguable needs and through their ability to influence their parents' consumer and occupational choices. In consequence, children are the target of major advertising and marketing campaigns, and the legitimacy and effectiveness of these has been keenly debated. Second, however, children live in an entire economic world of their own, involving exchange transactions that are at best tangentially related to the adult economy. They may show much more sophisticated economic understanding in these transactions than they reveal in the traditional studies of the development of economic understanding.

Another widespread approach to considering the economy as an independent variable has been comparative study, either cross-sectional or (more rarely) longitudinal. Cross-sectional studies include a series of impressive collaborative cross-national investigations, for example, of tax evasion and of attitudes to the Euro when it was introduced (Pepermans et al., 1998). The many institutional differences between nations mean that such comparisons can give a good insight into the effective economic variables producing differences between people.

Longitudinal studies within single economies or groups of economies are also increasingly important, aided by the increasing availability of data from panel studies, and the increasing number of them that include at least some questions

suggested by economic psychologists. Such studies can throw light on the impact on individual psychology of macroeconomic changes, such as inflation, economic growth, or recession.

Where such formal comparisons are not feasible, 'impact studies' have been popular: people in a particular economic position are interviewed or given psychometric tests to try to understand how they differ from more typical populations. Such studies have, for example, looked at the impact of the tourist economy on its host populations, at the kinds of coping behavior elicited by living in poverty within a wealthy society (Anand and Lea, 2011, introduce a journal special issue devoted to this question), and the impact of unemployment and retirement.

In many such impact studies, the key dependent variable has been individual happiness or life satisfaction. While there are still skeptics around, the measurement of happiness has gone from being a pursuit so 'soft' that no serious social scientist, and certainly no economist, should contemplate it, to being an accepted tool by which governments' policies and performance should be judged. Interest in happiness and its antecedents of course runs much more widely than economic psychology, but it is perhaps the clearest possible example of the proposition, central to economic psychology, that the economic system and the individual's position within it are crucial independent variables in individual psychology.

The impact of the economy on individuals, and the impact of individuals on the economy, of course go together; there is a cycle of causality between the two, which means that the impact of changes will never be entirely predictable from studies that isolate one process or the other. A good example of this occurs in the psychology of taxation, reviewed by Kirchler (2007). On the one hand, economic psychologists have looked at individual behaviors, particularly tax evasion and avoidance; on the other hand, the tax regime in an economy induces changes in such behaviors – its perceived fairness, for example (or the extent to which individuals can plausibly represent it to themselves as unfair), may have a marked effect on tax compliance.

In the first decade of the twenty-first century, the most important economic event within the so-called developed world has been the financial crisis that started in 2007. Research on its psychological impacts has been started in many institutions, but at the time of writing little had found its way to formal publication. It is clear that, despite the fact that behavioral economists are close to several Western governments, many governments' handling of the crisis has been far from ideal from a psychological perspective; not only have steps that seem obvious from the known psychology of previous recessions not been taken but also precisely opposite measures have been put in place in some cases.

From a longer perspective, however, it may turn out that the crisis in the 'Western' economies is not as important in world economic history as the rise of some 'non-Western' countries. It is widely recognized that the so-called BRIC countries (Brazil, Russia, India, and China) have the potential to eclipse the current leading world economies by the mid-twenty-first century; commentators are busily coining terms for other groups, such as MIKT (Malaysia, Indonesia, Korea, Turkey) or CIVETS (Colombia, Indonesia, Vietnam, Egypt, Turkey, and South Africa), that are showing signs of coming to prominence in the wake of the BRICs. There are increasing signs that the long-term economic underperformance of the whole continent of Africa may be ending. A key question is what psychological impact may follow from the explosive growth that such economies are undergoing or may soon undergo. That question is so far unanswered; but every one of the BRIC countries has active groups of economic psychology researchers within them, and there are at least individuals trying to establish economic psychology as a field of research within all the MIKT and CIVETS countries.

Economic Psychology in a New Economic Era

Throughout the twentieth century, economic psychology struggled to establish itself as an interdisciplinary study, in the face of an academic discipline of economics that was dominated by a single theoretical idea, rational self-interest. The focus of twentieth-century economics was theoretical, and much of its theory consisted of working out what a rationally self-interested individual would do in a given situation. Economic psychologists, although not all psychologists by initial training, shared with the discipline of psychology an interest in what people actually do, and a corresponding indifference to any single theory. Furthermore, the empirical bias of psychologists led them, as soon as they started to look at economic behavior, to question the economic assumption of rational self-interest. This means that a great deal of research in economic psychology consisted of investigations by psychologists of the empirical accuracy of theories developed by economists.

In the past two decades, that intellectual environment has changed radically, for two reasons. First, recognizing the various difficulties associated with a frontal assault on rationality assumptions, most economic psychologists ceased testing rationality assumptions directly, but rather set about trying to build empirically valid models of the causation and consequences of economic behaviors. Popular causal variables include personality differences, attitudes, socialization experiences, and psychological disorders. All these variables can potentially cause different people, exposed to apparently identical economic situations, to react in different ways. They thus allow psychologists to account for some of the individual variation in behavior that remains when the obvious economic variables have been taken into account. Cumulatively, of course, such studies can still constitute a powerful attack on the economic theory of rational self-interest. There are many ways of acting rationally in any given situation, depending on one's knowledge, abilities, values, and goals. If most of the variation between people's economic behavior actually depends on which rational behavior people choose, and if this has to be predicted from psychological properties of the person, the role of economic psychology has to increase, while the role of rational theory as such has to decrease.

Second, however, the discipline of economics has changed. Behavioral and experimental economics have

become enormously influential. Within economics, it is now accepted by many – although probably not by most – that economic behavior is not completely described by rational choice models; that deviations from rationality are systematic and substantial; that such deviations can be predicted from known processes in cognitive, social, comparative (and perhaps clinical, developmental, and evolutionary) psychology; that they can be studied in laboratory experiments; and that the results will be predictive for the real economy. The implication is that an economics (and hence an economic policy) that does not take account of psychological processes and data is radically incomplete. This view is widespread within academic economics; it is widely given credence within the consumers of economics, that is, governments and their economic advisors. Of course there are still many economists who would want to defend the assumption of rationality, and to argue that some of the psychological predictions about economic behavior are exaggerated or lack generality. Even they recognize that the defense must now be on empirical grounds; and on such grounds, economics and psychology can always meet.

See also: Behavioral Economics; Decision and Choice: Bounded Rationality; Decision and Choice: Heuristics; Experimental Economics; Money and Finances, Psychology of; Money, Sociology of; Money: Anthropological Aspects; Social Dilemmas, Psychology of.

Bibliography

Altman, M. (Ed.), 2006. Handbook of Contemporary Behavioral Economics. M. E. Sharpe, Armonk, NY.

Anand, P., Lea, S., 2011. The economic psychology of poverty. Journal of Economic Psychology 32, 205.

Dittmar, H., 2008. Consumer Culture, Identity and Well-being. Psychology Press, Hove, UK.

Furnham, A., 2008. The Economic Socialisation of Young People. Social Affairs Unit, London.

Glimcher, P.W., 2010. Foundations of Neuroeconomic Analysis. Oxford University Press, New York.

Katona, G., 1975. Psychological Economics. Elsevier, New York.

Kirchler, E., 2007. Economic Psychology of Tax Behaviour. Cambridge University Press, Cambridge, MA.

Lea, S.E.G., 1992. Editorial: On parent and daughter disciplines: Economic psychology, industrial psychology and consumer science. Journal of Economic Psychology 13, 1–3.

Lea, S.E.G., Webley, P., 2006. Money as tool, money as drug: the biological psychology of a strong incentive. Behavioral and Brain Sciences 29, 161–209.

Lewis, A. (Ed.), 2008. The Cambridge Handbook of Psychology and Economic Behaviour. Cambridge University Press, Cambridge, MA.

Nyhus, E.K., Pons, E., 2005. The effects of personality on earnings. Journal of Economic Psychology 26, 363–384.

Pepermans, R., Burgoyne, C.B., Müller-Peters, A., 1998. Editorial: European integration, psychology and the Euro. Journal of Economic Psychology 19, 657–661.

Thaler, R.H., Sunstein, C.R., 2008. Nudge. Yale University Press, New Haven, CT.

Wärneryd, K.-E., 1999. The Psychology of Saving: A Study on Economic Psychology. Edward Elgar, Cheltenham.

Wärneryd, K.-E., 2001. Stock-market Psychology. Edward Elgar, Cheltenham.

Decision and Choice: Heuristics

Henrik Olsson and Mirta Galesic, Max Planck Institute for Human Development, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P.M. Todd, volume 10, pp. 6676–6679, © 2001, Elsevier Ltd.

Abstract

Heuristics are approximate strategies for decision making. They do not process all the available information and therefore can yield a reasonable solution fast, even in conditions of incomplete or uncertain information. The same, or similar, heuristics can be used in both preference tasks (e.g., choosing between consumer products) and inference tasks (e.g., which of two cities is larger), but some were traditionally studied in the preference, and some in the inference, context. There is also a tradition of studying heuristics that are used primarily in social contexts.

Introduction

Heuristics are approximate strategies or 'rules of thumb' for decision making and problem solving. They do not process all the available information and therefore can yield a reasonable solution fast, even in conditions of incomplete or uncertain information. Heuristics are most commonly studied in psychology within the domains of judgment and decision making, social cognition, and problem solving, and in the fields of artificial intelligence, operations research, and behavioral biology. This article focuses on heuristics proposed as models of how people make decisions and choices.

Historical Overview

The term *heuristic* is of Greek origin and means 'to find out' or 'discover.' Gestalt psychologists proposed heuristic methods to guide searches for information (Duncker, 1935). The mathematician George Pólya (1945) distinguished heuristics from analytical methods: for instance, heuristics are needed to find a proof, whereas analysis is for checking a proof. Pólya's student Allen Newell, together with Herbert Simon, developed models of heuristics to limit large search spaces with applications in artificial intelligence and cognitive science (see Newell Shaw and Simon, 1958). Mathematical psychologists (Luce, 1956; Tversky, 1972; Dawes, 1979) and others studied models of heuristics for choice, such as lexicographic rules, elimination by aspect, and equal-weight rules. Behavioral biologists studied experimentally the rules of thumb that animals use for choosing food sites, nest sites, or mates (see Hutchinson and Gigerenzer, 2005). Psychologists in the 1970s became interested in demonstrating human reasoning errors that they believed are the consequence of using heuristics (Tversky and Kahneman, 1974). They focused on studying when verbally described heuristics break down and yield deviations from classical norms of rationality (Kahneman et al., 1982). The resulting heuristics and biases program has had immense influence, contributing to the emergence of behavioral economics and behavioral law. An alternative approach focused on the adaptive nature of heuristics and investigated how computationally modeled heuristics can exploit structured information to yield fast and accurate decisions (Gigerenzer et al., 1999). Studies conducted in this tradition showed that when used in the appropriate environment, heuristics can surpass the predictive accuracy of more sophisticated and information-greedy tools, including that of regression models or neural nets (Gigerenzer and Brighton, 2009). Recently, there have been attempts to integrate different views of heuristics. For example, Shah and Oppenheimer (2008) proposed that different heuristics share the same goal of effort reduction by examining fewer attributes (also called aspects, cues, or features), reducing the effort of retrieving attribute values, simplifying the weighting of attributes, integrating less information, and examining fewer options.

Evaluating the Performance of Heuristics

Heuristics, as well as any model of decision processes, can be evaluated according to how well they describe people's choices and how well they correspond to normative criteria of accuracy. What normative criteria can be used depends on the nature of the task. In preference tasks, such as choosing between consumer products or monetary gambles, the accuracy of a heuristic can only be determined by assessing how closely its predictions match that of some gold standard for rational preferences, such as maximizing expected utility. In inference tasks, such as which soccer team will win or which of two cities is larger, the accuracy of heuristics is judged by their ability to predict real-world outcomes. The same, or similar, heuristics can be used in both preference tasks and inference tasks, but some were traditionally studied in the preference and some in the inference context. There is also a tradition of studying heuristics that are used primarily in social contexts. Success of these heuristics is also judged by criteria such as fairness, coordination, and equality.

Heuristics for Preferences

The typical normative approach to choice in preference tasks involves weighting and combining all of the available attributes of different options, with the weights determined by a process such as multiple regression. Several heuristic

alternatives to this standard approach have been proposed. Simon's (1955) satisficing heuristic searches through options, stops as soon the first option exceeds an aspiration level, and chooses this option. In Selten's (1998) aspiration adaptation theory, the aspiration level can change throughout the decision process. A family of models called lexicographic heuristics search through attributes (Fishburn, 1974). In their simplest form, attributes are inspected one at a time, and the option with the best value on that attribute is selected. In the case of a tie on the most important attribute, the tied options are searched again according to the second most important attribute, and so on until choosing a single option. Perhaps the best-known member of the lexicographic family is the elimination by aspects heuristic (Tversky, 1972). According to this heuristic, decision makers select and establish a cutoff value for the most important attribute. Decision makers eliminate any options that do not satisfy the criterion value on this attribute and continue to choose attributes in this manner until only one option remains. Payne and colleagues (Payne et al., 1993) provided evidence that the performance of these and other lexicographic heuristics can approach that of a weighted-additive rule while minimizing effort. A recent lexicographic rule for choices between gambles is the priority heuristic (Brandstätter et al., 2006).

Heuristics for Inferences

Many heuristics for inference tasks are also part of the lexicographic family of models, searching sequentially through attributes of different options. The take-the-best heuristic infers which of two options has a higher value on a criterion, based on binary attribute values retrieved from memory where attributes are ordered in descending order of their validity. Lexicographic heuristics can also be applied to categorization (or classification), where the task is to assign an object to one of several categories. These heuristics are called fast and frugal trees (Martignon et al., 2003).

Other heuristics only consider one piece of information. The recognition heuristic (Gigerenzer et al., 1999; Goldstein and Gigerenzer, 2002) considers the single cue of recognition: when one option is recognized and the other is not, this heuristic chooses the former and otherwise guesses. The fluency heuristic, formulated for situations where both options are recognized, chooses the option that was recognized faster (Schooler and Hertwig, 2005).

Some heuristics use all available attributes but follow an equal weighting principle (also called tallying, unit- or equal-weights regression), reducing cognitive demand. Equal weights models can even outperform weighted models such as multiple regression in terms of predictive accuracy (Dawes and Corrigan, 1974; Dawes, 1979). Another variant of the equal weighting principle is the 1/N rule, which is a heuristic for the allocation of resources, such as time or money to N alternatives. This heuristic has been shown to work well in investment situations (DeMiguel et al., 2009).

Several heuristics have been proposed for judgments of probabilities or other quantities (Kahneman et al., 1982; Gilovich et al., 2002). The representativeness heuristic (Kahneman and Tversky, 1972) judges the probability that an object A belongs to a particular category B (e.g., that a person described as a feminist is a bank teller) or that an event A is generated by a particular process B (e.g., that the sequence TTHHHH was generated by throwing a fair coin), based on the extent that A is representative of, or similar to, B. The availability heuristic (Tversky and Kahneman, 1973) judges class frequency or event probability based on how easily instances of the class or event can be mentally retrieved or constructed (e.g., plane crashes may seem like a frequent cause of death because it is easy to recall examples). The anchoring-and-adjustment heuristic (Tversky and Kahneman, 1974) judges quantities by starting with a particular value (the anchor) and adjusting from it (e.g., quick estimates of $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$ are typically higher than quick estimates of $1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$, because the result of the first few steps of multiplication is higher in the former case). These heuristics have been criticized as being vague redescriptions of observed choice behavior that are difficult to falsify. Kahneman and Frederick (2002) strengthened the definition of the representativeness heuristic and proposed that it worked by attribute substitution, that is by replacing a computationally complex attribute (e.g., probability) by one that is more readily available (e.g., similarity). This enabled researchers to implement computational versions of the representativeness heuristic and compare them with other models of how people judge probabilities (e.g., Nilsson et al., 2005). The definition of the availability heuristic has also been strengthened by implementing mechanisms consistent with the original formulation (Sedlmeier et al., 1998).

Heuristics for Social Contexts

Heuristics described so far can also be used to solve problems in social contexts. Attributes used in lexicographic heuristics can be social in nature, for instance, whether one's friends chose one or the other option. Social information can be used to find the best attribute order in the take-the-best heuristic (Garcia-Retamero et al., 2009), and groups can profit from relying on members who can use a particularly useful heuristics (e.g., those who recognize only one of two options and therefore can use recognition heuristic; Reimer and Katsikopoulos, 2004). Rules such as tallying and 1/N can be used to make group decisions and determine how to distribute resources within a group (Hastie and Kameda, 2005; Hertwig et al., 2002).

Other heuristics are used primarily in social contexts. Simple rules such as tit-for-tat, tit-for-two-tats, or win-stay-lose-shift (Axelrod, 1984; Nowak and Sigmund, 1993) can help maintain a high level of cooperation in the group and prevent harmful effect of defectors. To choose whom to learn from, animals including humans use rules such as 'copy the majority,' 'copy if rare,' 'copy successful,' 'copy older individuals,' 'copy kin,' and 'copy friends' (Boyd and Richerson, 1985; Laland, 2004). To decide whether to use social learning in the first place, useful rules can be 'when individual learning is costly,' 'when individual information is not relevant' or 'unreliable' or 'outdated,' and 'when dissatisfied' (Kendal et al., 2005). To make inferences about

broader populations, one can use the social circle heuristic. For example, to judge prevalence of diseases in the general population one can search for occurrences of the diseases in one's immediate social environment (Hertwig et al., 2005). Social environments also guide choices between different socially relevant options. People tend to choose the option that is set as the default in a given society, resulting in a large percentage of organ donors in societies where one needs to opt out of organ donation and a small percentage where one needs to opt in (Johnson and Goldstein, 2003).

See also: Decision Making: Nonrational Theories; Decision and Choice: Bounded Rationality; Decision and Choice: Sequential Decision Making; Heuristics in Social Cognition; Problem Solving and Reasoning, Psychology of.

Bibliography

Axelrod, Robert, 1984. The Evolution of Cooperation. Basic Books, New York.

Boyd, Robert, Richerson, Peter J., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago.

Brandstätter, Eduard, Gigerenzer, Gerd, Hertwig, Ralph, 2006. The priority heuristic: making choices without trade-offs. Psychological Review 113, 409–432.

Dawes, Robyn M., 1979. The robust beauty of improper linear models in decision making. American Psychologist 34, 571–582.

Dawes, Robyn M., Corrigan, Bernard, 1974. Linear models in decision making. Psychological Bulletin 81, 95–106.

DeMiguel, Victor, Garlappi, Lorenzo, Uppal, Raman, 2009. Optimal versus naive diversification: how inefficient is the 1/N portfolio strategy? Review of Financial Studies 22, 1915–1953.

Duncker, Karl, 1935. Zur Psychologie des produktiven Denkens. Springer, Berlin.

Fishburn, Peter C., 1974. Lexicographic orders, utilities and decision rules: a survey. Management Science 20, 1442–1471.

Garcia-Retamero, Rocio, Takezawa, Masanori, Gigerenzer, Gerd, 2009. Does imitation benefit cue order learning? Experimental Psychology 56, 307–320.

Gigerenzer, Gerd, Brighton, Henry, 2009. Homo heuristicus: why biased minds make better inferences. Topics in Cognitive Science 1, 107–143.

Gigerenzer, Gerd, Todd, Peter M., the ABC research group, 1999. Simple Heuristics that Make Us Smart. Oxford University Press, New York.

Gilovich, Thomas, Griffin, Dale W., Kahneman, Daniel (Eds.), 2002. Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge University Press, New York.

Goldstein, Daniel G., Gigerenzer, Gerd, 2002. Models of ecological rationality: the recognition heuristic. Psychological Review 109, 75–90.

Hastie, Reid, Kameda, Tatsuya, 2005. The robust beauty of majority rules in group decisions. Psychological Review 112, 494–508.

Hertwig, Ralph, Davis, Jennifer N., Sulloway, Frank J., 2002. Parental investment: how an equity motive can produce inequality. Psychological Bulletin 128, 728–745.

Hertwig, Ralph, Pachur, Thorsten, Kurzenhäuser, Stephanie, 2005. Judgments of risk frequencies: tests of possible cognitive mechanisms. Journal of Experimental Psychology: Learning, Memory, and Cognition 31, 621–642.

Hutchinson, John M.C., Gigerenzer, Gerd, 2005. Simple heuristics and rules of thumb: where psychologists and behavioural biologists might meet. Behavioural Processes 69, 97–124.

Johnson, Eric J., Goldstein, Daniel G., 2003. Do defaults save lives? Science 302, 1338–1339.

Kahneman, Daniel, Frederick, Shane, 2002. Representativeness revisited: attribute substitution in intuitive judgment. In: Gilovich, Thomas, Griffin, Dale W., Kahneman, Daniel (Eds.), Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge University Press, New York, pp. 49–81.

Kahneman, Daniel, Tversky, Amos, 1972. Subjective probability: a judgment of representativeness. Cognitive Psychology 3, 430–454.

Kahneman, Daniel, Slovic, Paul, Tversky, Amos, 1982. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, New York.

Kendal, Rachel L., Coolen, Isabelle, van Bergen, Yfke, Laland, Kevin N., 2005. Tradeoffs in the adaptive use of social and asocial learning. Advances in the Study of Behavior 35, 333–379.

Laland, Kevin N., 2004. Social learning strategies. Learning & Behavior 32, 4–14.

Luce, R. Duncan, 1956. Semiorders and a theory of utility discrimination. Econometrica 24, 178–191.

Martignon, Laura, Vitouch, Oliver, Takezawa, Masanori, Forster, Malcolm R., 2003. Naive and yet enlightened: from natural frequencies to fast and frugal decision trees. In: Hardman, David, Macchi, Laura (Eds.), Thinking: Psychological Perspectives on Reasoning, Judgment and Decision Making. Wiley, Chichester, pp. 189–211.

Newell, Allen, Shaw, J.C., Simon, Herbert A., 1958. Elements of a theory of human problem solving. Psychological Review 65, 151–166.

Nilsson, Håkan, Olsson, Henrik, Juslin, Peter, 2005. The cognitive substrate of subjective probability. Journal of Experimental Psychology: Learning, Memory, and Cognition 31, 600–620.

Nowak, Martin A., Sigmund, Karl, 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. Nature 364, 56–58.

Payne, John W., Bettman, James R., Johnson, Eric J., 1993. The Adaptive Decision Maker. Cambridge University Press, New York.

Pólya, George, 1945. How to Solve It. Princeton University Press, Princeton.

Reimer, Torsten, Katsikopoulos, Konstantinos V., 2004. The use of recognition in group decision-making. Cognitive Science 28, 1009–1029.

Schooler, Lael J., Hertwig, Ralph, 2005. How forgetting aids heuristic inference. Psychological Review 112, 610–628.

Sedlmeier, Peter, Hertwig, Ralph, Gigerenzer, Gerd, 1998. Are judgments of the positional frequencies of letters systematically biased due to availability? Journal of Experimental Psychology: Learning, Memory, and Cognition 24, 754–770.

Selten, Reinhard, 1998. Aspiration adaptation theory. Journal of Mathematical Psychology 42, 191–214.

Shah, Anuj K., Oppenheimer, Daniel M., 2008. Heuristics made easy: an effort-reduction framework. Psychological Bulletin 134, 207–222.

Simon, Herbert A., 1955. A behavioral model of rational choice. Quarterly Journal of Economics 69, 99–118.

Tversky, Amos, 1972. Elimination by aspects: a theory of choice. Psychological Review 79, 281–299.

Tversky, Amos, Kahneman, Daniel, 1973. Availability: a heuristic for judging frequency and probability. Cognitive Psychology 5, 207–232.

Tversky, Amos, Kahneman, Daniel, 1974. Judgment under uncertainty: heuristics and biases. Science 185, 1124–1131.

Decision and Choice: Luce's Choice Axiom

Timothy J Pleskac, Michigan State University, East Lansing, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.I. Yellott Jr, volume 13, pp. 9094–9097, © 2001, Elsevier Ltd.

Abstract

Luce's choice axiom (LCA) is a theory of individual choice behavior that has proven to be a powerful tool in the behavioral sciences for over 50 years. LCA is grounded in two fundamental properties: choice is probabilistic, and the probability of choosing an option from one set of alternatives is related to the probability of choosing the same option from a different set. This article reviews the basic properties of LCA. It also discusses crucial tests of predictions derived from LCA regarding the role of context. The historical connections of LCA and crucial applications of LCA are also reviewed.

Introduction

A subject in a psychology experiment (animal or human) identifies which of two stimuli are brighter; a diner at a restaurant selects a meal from a menu; an eyewitness determines a face in a lineup. These are all specific examples of a general behavior called 'choice,' where an individual selects one option from a larger set of alternatives.

Owing to the generality of choice behavior, many different areas of the social sciences have sought a mathematical description of choice. Luce's (1959) choice axiom (LCA) is one such description of individual choice behavior. As Estes (1997) describes it, Luce, in developing a theory of individual choice behavior, did not take the standard approach of psychology. Instead of working from the bottom up, accumulating empirical data on a phenomenon of interest and eventually arriving at a general theory of choice behavior, Luce took the approach of theoretical physics and worked from the top down. He employed intuition and reason to find a set of general assumptions that if true would allow the development of a theory and mathematical model useful for interpreting and understanding choice behavior. In short, LCA helped usher in the axiomatic method into psychology and, more broadly, the social sciences. Owing to this new approach is no doubt part of the reason why an early draft of the book and the ideas in it were referred to as the "red menace" (Estes, 1997).

Luce started with two basic tenets of individual choice behavior: (1) it is probabilistic; and (2) the probability of choosing an option from one set of alternatives is related to the probability of choosing the same option from a larger set of alternatives. The tenet that choice behavior is probabilistic is in contrast to assuming choice is deterministic. To understand this distinction, consider a situation where an agent chooses an option x from the set of alternatives T (e.g., possible entrees on a menu). Focusing on the special case of two-alternative forced-choice or paired comparison, deterministic theories assume a binary preference relation $\prec$ that is either true or false for any pair of alternatives, either $x \prec y$, $y \prec x$, or $x \sim y$ for all $x, y \in T$, where $\sim$ is indifference.

A probabilistic choice theory, like LCA, makes the more psychologically plausible assumption that it is only with some probability that an alternative is selected. Focusing again on a paired comparison, a probabilistic choice theory specifies a probability function $P(x,y)$ that maps each pair of alternatives into the closed interval [0,1]. LCA is a multi-alternative choice theory, so it is not limited to paired comparisons but is applicable in the more general choice situation specifying the probability that x will be chosen from the set T, $P_T(x)$. Regardless, as a probabilistic theory, LCA adheres to the axioms of probability theory. These axioms are:

$$1.\ \text{For } S \subset T, 0 \leq P_r(S) \leq 1$$

$$2.\ P_T(T) = 1$$

$$3.\ \text{If } R,\ S \subset T \text{ and } R \cap S = 0,\ \text{ then } P_T(R \cup S) = P_T(R) + P_T(S)$$

The probability axioms constrain each of the measures P_T. The problem is that the axioms themselves offer no connection between measures over different sets composed of some of the same alternatives; yet, such a connection seems necessary for a theory of choice. Anecdotally, the probability that one chooses pan-seared walleye from the dinner menu at a restaurant certainly seems related to the probability that he or she will choose the same entree from the smaller, reduced lunch menu. This observation motivates a second tenet of choice behavior: how decision makers select an option from a smaller set of alternatives is related to how they would choose when the same option is in the context of a larger set of alternatives and vice versa. LCA formalizes this assumption.

The Choice Axiom

The axiom has two parts.

Part 1:

$$\text{If } P(X,Y) \neq 0, 1 \text{ for all } x, y \in T, \text{ then for } R \subset S \subset T \quad P_T(R) = P_S(R) P_T(S) \qquad [1]$$

Part 2:

$$\text{If } P(X,Y) = 0 \text{ for some } x, y \in T, \text{ then for } S \subset T \quad P_T(S) = P_{T-\{x\}}(S - \{x\}) \qquad [2]$$

Working backward, Part 2 is more or less a housekeeping assumption. It allows alternatives that are never chosen in pairwise choices to be deleted from S without impacting the choice probabilities. If salted whitefish is never chosen in pairwise choices with trout, then in choices between salted whitefish, trout, and walleye, salted whitefish can be safely deleted, reducing the choice to trout or walleye.

Part 1 states that the probability of selecting the set of alternatives R (e.g., walleye or trout) from T is equal to the probability of selecting R from S (e.g., walleye, trout, or salmon) multiplied by the probability of selecting S from T. When $S = \{x, y\}$ then the probability measure P_S reduces to a pairwise choice probability mentioned earlier. In other words, Part 1 addresses the issue of formally relating the choice probabilities for subsets of T to those of T itself. Another way to see this is to rewrite eqn [1] as a conditional probability:

$$P_T(R/S) = P_S(R) = \frac{P_T(R)}{P_T(S)} \qquad [3]$$

where $R \subset S \subset T$ and $P_T(S) > 0$. Thus, the probability of selecting R from S equals the conditional probability that R is chosen from S when the full set T is available.

To be clear, LCA is not an additional probability axiom. Rather, it is an empirical hypothesis that has a number of strong consequences. These consequences have been tested, and in some places have been shown not to hold under specific circumstances. Nevertheless, LCA and the mathematical model it implies have proven quite useful in modeling choices whether they are higher level economic decisions or lower level perceptual choices. It has also motivated similar approaches in modeling probability judgments (Tversky and Koehler, 1994).

Implications

Most well known is that LCA implies that there is a numerical ratio scale v over T such that for every $x, y \in S \subset T$ the probability of choosing x can be found with the following mathematical model:

$$P_S(x) = \frac{v(x)}{\sum_{y \in S} v(y)}. \qquad [4]$$

Psychologically $v(x)$ is a response strength associated with the alternative x. One might see from eqn [4] that one measure of response strength is the observed choice probability $v(x) = P_T(x)$. Other scales can be created by multiplying $P_T(x)$ by a positive constant k, $v(x) = k \cdot P_T(x)$. This means that v is a ratio scale of response strength, a provocative idea in psychology that is beset with ordinal scales.

Equation [4] is an intuitive way to model choice probabilities, so much so that others have proposed it as a model of choice. For example, Bradley and Terry (1952) proposed a model consistent with eqn [4] for pairwise choices (see also Clark, 1957). This straightforward way to model choices via response strength also exposes why LCA has proven important in modeling choices in a variety of domains.

Using eqn [4], one can also see that

$$\frac{P(x,y)}{P(y,x)} = \frac{P_S(x)}{P_S(y)}. \qquad [5]$$

This result is because both sides of eqn [5] are equal to $\frac{v(x)}{v(y)}$. This relationship is sometimes called the *constant ratio rule*. The constant ratio rule is a probabilistic form of what is called independence from irrelevant alternatives (Arrow, 1951). It implies that when LCA holds for a set of alternatives T and its subsets, the ratio $P_S(x)/P_S(y)$ is independent of S. In other words, the ratio of the probability of choosing one alternative to the probability of choosing another should be constant regardless of the set of options (or context).

This constant ratio rule is revealing in terms of the strengths and weakness of LCA. A key strength is that pairwise choice probabilities can be predicted from choice probabilities observed in a larger set. A potential weakness is that psychologically the context of the situation (e.g., similarity to other choice alternatives) should impact the probability of choosing an option in some manner. Such a weakness speaks to the validity of LCA, and we will return to this idea in the validity section.

LCA also constrains the possible pairwise probabilities. This constraint is known as the product rule, which states that for alternatives x, y, and z.

$$P(x,y) \cdot P(y,z) \cdot P(z,x) = P(x,z) \cdot P(z,y) \cdot P(y,x). \qquad [6]$$

To see how the product rule works, note that the left-hand and right-hand sides contain complementary pairwise choice probabilities. Consequently, dividing the left-hand side by the right-hand side produces

$$\frac{P(x,y)}{P(y,x)} \cdot \frac{P(y,z)}{P(z,y)} \cdot \frac{P(z,x)}{P(x,z)} = 1. \qquad [7]$$

Each of the ratios of choice probabilities can, in turn, be rewritten as a ratio of response strengths by eqn [4]. This allows each response strength to be canceled off, resulting in the product of the ratios equaling 1. The product rule is another independence condition that implies the response strength for an alternative is the same regardless of the other alternative it is paired with. Just as with the constant ratio rule, the product rule identifies the strengths and potential weaknesses of LCA. The strength is that the product rule allows one to predict the pairwise choice probability $P(x,z)$ if $P(x,y)$ and $P(y,z)$ are both known. The weakness is that again psychologically it seems that the context matters. Again, the similarity between two alternatives seems like it should impact the response strength of an option.

Historical Connections

LCA has a number of connections to other choice models. Perhaps most prominent is the connection to L.L. Thurstone's (1927) law of comparative judgment. This theory assumes that alternatives have numerical psychological values that vary randomly from one choice occasion to another (i.e., the values are random variables). The decision maker is assumed to compare these values in order to make a choice. More precisely, if the psychological value of the alternatives $x_1, \ldots, x_N$ are the random variables $U_1, \ldots, U_N$, the probability of choosing x_i from any set S is the probability that U_i is currently the largest of the random variables $\{U_j : x_j \in S\}$. The

most prominent version of Thurstone's theory is Case V, which assumes the psychological values are normally, independent and identically distributed. This means U_i can be rewritten as $u_i + \varepsilon_i$, where u_i is a real constant (i.e., a scale value) and ε_i is normally distributed with a mean of 0 and variance of 1, $\varepsilon_i \sim N(0,1)$.

Luce (1959) showed that LCA and Case V make nearly identical predictions for pairwise choice probabilities: the numerical predictions were off by at most 0.02. In fact, if the differences between the psychological values have a logistic distribution rather than a normal distribution, then Thurstone's law of comparative judgment and LCA make identical predictions (Adams and Messick, 1957). What Holman and Marley (cited in Luce and Suppes, 1965) subsequently showed is that a logistic distribution of differences arises in Thurstone's (1927) framework if the normal random variables ε_i are replaced by independent random variables that have a double exponential probability distribution.

Yellott (1977), in the spirit of Luce's top-down axiomatic approach, identified an intuitively plausible assumption that would lead one to conclude the random variables ε_i are double exponentially distributed. The assumption is grounded in the property that the asymptotic distribution of the maximum of n independent, identically distributed random variables is the double exponential when the underlying distribution has an upper exponential tail. Using this property, Yellott proposed a thought experiment in which there are three basic alternatives, for example, beverages on a table: coffee, tea, and milk. Instead of one cup of each on the table, consider a situation like a party where there are n identical cups of coffee, n cups of tea, and n glasses of milk. One would expect that the probability of choosing any one of the beverages should be independent of n or how many replicas there are. In fact, if this invariance assumption is met, then the distribution in a Thurstone model must be a double exponential and, as a result, LCA must hold. In the words of Luce (1977), "Yellott's condition is so compelling that this theorem means Thurstone's model … and the choice axiom stand or fall together" (p. 218).

LCA also connects with a number of other models in experimental psychology. This came primarily via what is called the similarity choice model (Estes, 1997). This model is a generalization of the choice model in eqn [4] that Luce (1959) developed to account for response bias and later fully developed in Luce (1963) (see also Shipley, 1960). The similarity choice model decomposes the response strength v into two ratio scales representing two constructs of high interest to psychologists: discriminability and response bias. To see the extension, consider the situation where a participant in a psychology experiment must choose a response R_j upon presentation of a stimulus V_i (e.g., identify a letter upon the presentation of an obscured letter stimulus). The similarity choice model states the response probability is

$$P_{i,j} = \frac{M_{i,j} \cdot b_j}{\sum_k M_{i,k} \cdot b_k}. \qquad [8]$$

The scale value $M_{i,j}$ represents the similarity between stimulus V_i and the stimulus corresponding with the response j, V_j. The other scale value, b_j, measures a subject's bias to choose response j regardless of the stimulus.

It should be noted that the similarity choice model is a generalization of the choice model (eqn [4]), but not necessarily the axiom itself (Luce, 1963). Despite this restriction, the similarity choice model can easily have some of the same context-free assumptions of LCA. The similarity choice model provided researchers with an alternative and more convenient means to model discriminability and response bias than with signal detection theory (Tanner and Swets, 1954). This convenience facilitated the incorporation of the similarity choice model as a choice rule into more complex models of cognitive processes including identification, recognition memory, categorization, and categorization learning (see Estes, 1997).

Validity of LCA

The value of the axiomatic approach that underlies LCA is that it establishes precise and testable conditions that help establish when a theory is applicable and when it is not. In the case of LCA, examples were almost immediately identified that questioned its validity.

One of the first examples came from Gerald Debreu (1960) in his published review of *Individual Choice Behavior*. (A famous adaptation of Debreu's example is the red-bus-blue-bus problem in transportation research (Train, 2003).) The example is as follows. Consider a classical music lover who is choosing between a recording of the Debussy quartet (option D) and two different recordings of Beethoven's Eighth Symphony (option B_1 and option B_2). For this example, assume the two Beethoven recordings are of equal quality (e.g., played by the same orchestra, but under the direction of different conductors). Let us assume the music lover is torn between adding a Debussy or a Beethoven recording to his collection. Therefore, in pairwise choices $P(B_1,B_2) = P(D,B_1) = P(D,B_2) = 1/2$. LCA thus implies that when presented with the set S of all three options $\{D, B_1, B_2\}$ simultaneously that $P_s(D) = 1/3$. However, this conclusion is not quite right intuitively. When the choice is between all three, it seems that the choice is between Debussy or Beethoven. Put differently, $P_S(D)$ should remain closer to 1/2, while choice probability for the two Beethoven recordings should be split equally, moving them closer to 1/4. Empirically, at least in preferential tasks, this is what happens. Tversky (1972) called this the similarity effect. He found violations when subjects were asked to make choices between college applicants and gambles. However, the violations in a perceptual task of judging dot numerosity were not statistically significant.

The similarity effect is inconsistent with the constant ratio rule of LCA (eqn [4]). The implication of the similarity effect is actually stronger in that it is inconsistent with any probabilistic choice theory that assumes choice probability is a monotone function of scale values also known as simple scalability models (Tversky, 1972). Note the similarity between alternatives also impacts pairwise choice probabilities and leads to violations of the product rule and other independence assumptions (Rumelhart and Greeno, 1971; Tversky and Russo, 1969).

Generalizations of LCA can be developed to account for the similarity effect. For example, Tversky (1972) proposed an alternative choice process called elimination by aspects where alternatives are sets of attributes or aspects. Figure 1 depicts

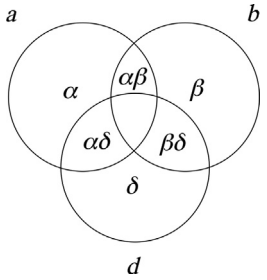

Figure 1 A graphical representation of three alternatives each composed of a set of unique and shared aspects.

a three-alternative set $T = \{a, b, d\}$. The aspects α, β, and δ are unique to alternatives a, b, and d, respectively. The other aspects $\alpha\beta$, $\alpha\delta$, and $\beta\delta$ are shared. Each aspect has a scale value u representing its utility. Because of the elimination by aspect assumption of the theory, aspects shared among all alternatives are not used in this choice rule, hence the central area in Figure 1 is unlabeled.

The basic idea in the theory is that an aspect is attended to with some probability (e.g., Beethoven). This probability is proportional to its utility relative to the sum of all the utilities K. For example, the probability of selecting $\alpha\beta$ is $u(\alpha\beta)/K$. Alternatives that do not have that aspect (e.g., d) are then eliminated and the probability of selecting say a is the pairwise choice probability scaled from the remaining aspects

$$P(a,b) = \frac{u(\alpha) + u(\alpha\delta)}{u(\alpha) + u(\alpha\delta) + u(\beta) + u(\beta\delta)}. \qquad [9]$$

In other words, elimination by aspects is consistent with one's intuitive reaction to Debreu's (1960) Debussy/Beethoven example: after an alternative is eliminated, choice probability gets divided among the remaining alternative.

If an aspect is attended to that is unique to a particular alternative (e.g., a), then the corresponding alternative (e.g., a) is selected with probability 1. Thus, the probability of selecting alternative a from T is

$$P_T(a) = \frac{u(\alpha) + u(\alpha\beta)\cdot P(a,b) + u(\alpha\delta)\cdot P(a,d)}{K}. \qquad [10]$$

Elimination by aspects is one hypothesis of how the similarity between the alternatives (i.e., shared aspects) interacts with the choice process. Using Figure 1, one can also see how elimination by aspects is a generalization of LCA: if alternatives are pairwise disjoint (i.e., no shared aspects between pairs of alternatives), then the predicted choice probabilities are consistent with LCA (Tversky, 1972).

More broadly, the similarity effect demonstrates that context matters: the response strength of an alternative depends on the options that are in the choice set. Over the years, other context effects have been identified and studied that further attest to the importance of context (Rieskamp et al., 2006). A second context effect that questions the validity of LCA (as well as elimination by aspects) is known as the attraction effect. The attraction effect can be illustrated with an example from Simonson and Tversky (1992). They gave one group of subjects a choice between \$6 and a nice Cross pen. The pen was chosen by 36% of the subjects, while the remaining 64% chose \$6. A second group was given a choice between three options: \$6, a nice Cross pen, or another less attractive pen. As one might expect, only 2% chose the less attractive pen. Yet, the mere addition of this asymmetrically dominated alternative boosted the proportion of subjects who chose the Cross pen to 46% (an increase of 33%). This behavior no doubt well known among marketers and salespeople does not seem quite right from the view point of LCA and, for that matter, most probabilistic theories of choice: adding an alternative to a choice set of even minuscule value should reduce choice probabilities for the all the alternatives, not increase them. This condition is known as regularity. Stated formally, for $x \in S \subset T$, $P_S(x) > P_T(x)$. The attraction effect demonstrates that in particular situations, individuals violate regularity.

A third context effect is the compromise effect. This effect can be demonstrated with another example from Simonson and Tversky (1992), where participants made choices between hypothetical cameras. Camera A was high in quality and price; Camera B was low in quality and price; and Cameral C had intermediate quality and price. In pairwise probabilities, $P(A|\{A, C\}) = P(C|\{A, C\})$. However, when Camera B at the other extreme is added to the choice set, $P(C|\{A, C, B\}) > P(A|\{A, C, B\})$. The compromise effect is another violation of independence from irrelevant alternatives. It questions the validity of LCA as well as elimination by aspects (for a proof, see Rieskamp et al., 2006).

These empirical effects identify limitations of LCA as a model of human choice behavior. In particular, they demonstrate a third tenet of choice behavior is that context does matter, and it shapes how individuals value an alternative. There are different psychological explanations for the role of context. Nevertheless, LCA and the choice rule that it implies have persisted as a useful model of individual choice behavior. There are at least two reasons for this. First, the basic idea that choice probability is monotonic with a response strength is intuitively quite plausible and one that has generally persisted even as alternative models have been proposed to account for context effects. Second, the axiom and the choice rule are elegantly simple, allowing for both easy implication and easy explanation.

Applications in Preferential Decisions

As the previous section examining the validity of LCA illustrates, there has been great interest in applying LCA to preferential choices. These are choices made over money and other alternatives (e.g., cameras, records, commuting routes, etc.) where there is no objectively correct answer. Luce (1959) very much anticipated applications of LCA and the choice rule in this domain, but great strides were made when (McFadden, 1976, 2001) developed an economic version of Luce's choice model in the *multinomial logit* model. Formally, the probability of choosing alternative x from set S is

$$P_S(x) = \frac{\exp[u(x)]}{\sum_{y \in S} \exp[u(y)]}, \qquad [11]$$

where $u(x)$ is the utility of alternative x and is a linear function of the attributes of x.

The multinomial logit model proved useful in economic data analysis, but it is also connected to what are called random utility models. This connection perfectly echoes the connection between LCA and Thurstonian choice models discussed earlier. Random utility models akin to Thurstonian models assume the utility of an alternative is a random variable. McFadden (1976), independently from Yellott (1977), showed that if the linear combination of attributes has an error component that is identically and independently distributed following an extreme value distribution (i.e., double exponential), then the result is the multinomial logit model. The multinomial logit model has become one of the most common random utility models (Train, 2003).

LCA has also contributed heavily in the further development of descriptive theories of preferential choice (Rieskamp et al., 2006). In this area, the formal attributes of LCA have proven indispensable in forming testable hypotheses about how individuals make a choice. Doing so has validated the core tenets of LCA (i.e., choice is probabilistic, and there is a connection between the propensity to choose alternatives between contexts). At the same time, as discussed in the previous section, LCA very much helped uncover other tenets such as the context dependence of choice. Psychologists and decision scientists have, in turn, sought to develop other formal theories that better account for these tenets. Examples (among others) include the context-dependent advantage model (Tversky and Simonson, 1993), decision field theory (Roe et al., 2001), and the leaky accumulator model (Usher and McClelland, 2004). The latter two theories perhaps represent the greatest departure from LCA in that they are grounded in a sequential sampling account of choice. Such a move allows for the theories to describe both choice probabilities and the time taken to make a choice. This is an important step as perhaps a fourth tenet of choice behavior is that the process of making a choice takes time, a property LCA is silent on.

One final application of LCA in preferential choices comes in the area of strategic games and game theory (McKelvey and Palphrey, 1995). Strategic games are situations when entities must decide how to act during interactions with other entities. Examples include two tennis players opposing each other in a match, roommates waiting for the other to purchase kitchen supplies, or research and development expenditures by competing pharmaceutical companies. These are all examples of interactions where how one entity decides to act affects the outcomes of the other and vice versa. This interaction is what distinguishes strategic games from the individual decisions where the outcomes an individual receives do not depend on the decisions of others. In game theory, each individual chooses among an action or strategy based on the utility of the outcomes that are expected to come from employing that strategy. Typically, the choice between strategies in game theory is deterministic. McKelvey and Palphrey (1995) adapted the multinomial logit model (eqn [11]) to move away from the deterministic assumption to one where players probabilistically choose among their strategies and assume their opponents do so as well.

The advantage of adapting a probabilistic choice rule in strategic games is twofold. The first advantage is that the probabilistic choice rule provides a more psychologically plausible assumption than a deterministic selection of a strategy. This allows the best equilibrium strategy to be played more often than others (but not always). The second and related advantage is that this can aid in data analysis of games where observed behavior is probabilistic, allowing measurement of interesting phenomena like learning during repeated play.

Applications in Learning

Luce (1959) also discussed LCA's application in modeling learning. Learning has long been of interest in the area of choice whether it be modeling the behavior of rats in a T-maze as they learn a route to food or more recently a participant in a psychology experiment choosing repeatedly between gambles that deliver a reward and/or punishment with unknown probabilities. Luce's idea called the beta model postulated operators that adjust the response strength of each alternative (e.g., $v(x)$) rather than adjusting the choice probabilities. More formally, during a learning experiment on each trial n, each alternative $x \in S$ has a response strength $v_n(x)$. If alternative x was selected on trial n, then the response strength for trial $n+1$ is updated. To make a choice on the trial $n+1$, the individual then uses the choice rule described in eqn [4]. Thus, in the beta model, LCA is assumed to hold on any given trial, but between trials the probability changes via adjustments in response strength.

This same basic idea is implemented in recent reinforcement learning algorithms that have become popular in a variety of areas including machine learning, control theory, game theory, and neuroscience. Perhaps the closest relation is with the Q-learning algorithm. According to this algorithm, after an agent selects an alternative x, the agent uses the experienced reward to update an expectation $Q(x)$ for alternative x (Sutton and Barto, 1998). To make a choice between the available alternatives in the set S, the algorithm uses the following rule sometimes called softmax:

$$P(x) = \frac{\exp[Q(x)/\tau]}{\sum_{y \in S} \exp[Q(y)/\tau]}. \tag{12}$$

The rule is a particular instance of eqn [4]. The parameter τ moderates the degree to which the agent chooses the alternative with the largest expectation. As τ increases, the agent is less likely to choose the option with the largest expectation and instead explore other alternatives. Note unlike the similarity choice model (eqn [8]), τ is not specific to the alternative, so it characterizes general exploration rather than a bias to a specific alternative. In general, this means the probabilistic choice rule provides a mechanism to trade off between selecting an alternative that provides an immediate payoff (exploit) or selecting an alternative that provides information to develop an expectation about future choices (explore). Neuroscientists are beginning to understand the neural basis of these learning mechanisms (Niv, 2009). The softmax expression, in particular, has proved useful in beginning to understand the neural mechanisms underlying the trade-off between exploration and exploitation (Daw et al., 2006). Finally, these learning mechanisms with the softmax rule are being integrated into larger cognitive architectures (Fu and Anderson, 2006).

Future Directions and Conclusion

LCA challenged conventional theories of choice to move from a deterministic account to a probabilistic one. In so doing, LCA has proved to be a power and useful tool in the modeling toolboxes of social and cognitive scientists for over 50 years. A search among scholarly articles is certain to reveal its vast impact. Yet, there are still areas that are left to be explored with LCA and the accompanying mathematical model of choice. Behaviorally, one area that Luce (1959) discussed and has still remained largely unexplored is that of ranking alternatives. Luce (1959) proposed one way to adapt LCA to ranking where the individual essentially makes a series of covert choice first picking, say the best alternative from the set, then the second best alternative from remaining items, and so on. This is called forward ranking. Another procedure, called backward ranking, works the opposite direction: picking the worst, next worst, and so on. Both procedures seem intuitively plausible; however, forward and backward ranking do not lead to the same solution (Luce, 1959). For current work in this area see Marley and Louviere (2005) and Lee et al. (2012).

As decision theorists continue to reconcile with the theoretical implications of probabilistic choice behavior, there are certainly more and more challenges. One challenge is a psychological one, which is to continue to develop a theory of choice behavior that is grounded in some of the core tenets of LCA (i.e., choice is probabilistic and that there is a relationship in how people choose between contexts), but also reflects other important properties like the context dependence and time course of choice. A second challenge is a neural one, which is to understand the neural circuitry underlying choice behavior (Gold and Shadlen, 2007). A final challenge is a statistical one, which is to develop appropriate statistical methodology to test probabilistic theories of choice-appropriate variability in choice behavior (Regenwetter et al., 2011). These are just some of the challenges we face. Regardless, LCA via its simple elegance is certain to continue to be an invaluable tool in helping us understand individual choice behavior.

See also: Decision and Choice: Bounded Rationality; Decision and Choice: Economic Psychology; Decision and Choice: Heuristics; Decision and Choice: Random Utility Models of Choice and Response Time; Dynamic Decision Making; Game Theory; Recognition; Signal Detection Theory.

Bibliography

Adams, E., Messick, S., 1957. An Axiomatization of Thurstone's Successive Intervals and Paired, Comparisons Scaling Models. Technical report. Stanford University.

Arrow, K., 1951. Social Choice and Individual Values. John Wiley and Sons, New York.

Bradley, R.A., Terry, M.E., 1952. Rank analysis of incomplete block designs. I. the method of paired comparison. Biometrika 39, 324–345.

Clark, F.R., 1957. Constant ratio rule for confusion matrices in speech communication. Journal of Acoustical Society of America 29, 715–720.

Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., Dolan, R.J., 2006. Cortical substrates for exploratory decisions in humans. Nature 441 (7095), 876–879.

Debreu, G., 1960. Review of R.D. Luce, individual choice behavior: a theoretical analysis. American Economic Review 50, 186–188.

Estes, W.K., 1997. Some reflections on the role of the choice model in theories of categorization, identification, and learning. In: Marley, A.A.J. (Ed.), Choice, Decision, and Measurement. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 321–329.

Fu, W., Anderson, J.A., 2006. From recurrent choice to skill learning: a reinforcement learning model. Journal of Experimental Psychology: General 135, 184–206.

Gold, J.I., Shadlen, M.N., 2007. The neural basis of decision making. Annual Review of Neuroscience 30, 535–574.

Lee, M.D., Steyvers, M., de Young, M., Miller, B., 2012. Inferring expertise in knowledge and prediction ranking tasks. Topics in Cognitive Science 4, 151–163.

Luce, R.D., 1959. Individual Choice Behavior. John Wiley & Sons, Inc., New York.

Luce, R.D., 1963. Detection and recognition. In: Luce, R.D., Bush, R.R., Galanter, E. (Eds.), Handbook of Mathematical Psychology. Wiley, New York, pp. 103–189.

Luce, R.D., 1977. The choice axiom after twenty years. Journal of Mathematical Psychology 15, 215–233.

Luce, R.D., Suppes, P., 1965. Preference, utility, and subjective probability. In: Luce, R.D., Bush, R., Galanter, E. (Eds.), 1965. Handbook of Mathematical Psychology, vol. 3. Wiley, New York.

Marley, A.A.J., Louviere, J.J., 2005. Some probabilistic models of best, worst, and best-worst choices. Journal of Mathematical Psychology 49, 464–480.

McFadden, D., 1976. Quantal choice analysis: a survey. Annals of Economic and Social Measurement 5, 363–390.

McFadden, D., 2001. Economic choice. The American Economic Review 91 (3), 351–378.

McKelvey, R.D., Palphrey, T.R., 1995. Quantal response equilibria for normal form games. Games and Economic Behavior 10, 6–38.

Niv, Y., 2009. Reinforcement learning in the brain. Journal of Mathematical Psychology 53, 139–154.

Regenwetter, M., Dana, J., Davis-Stober, C.P., 2011. Transitivity of preferences. Psychological Review 118 (1), 42–56.

Rieskamp, J., Busemeyer, J.R., Mellers, B.A., 2006. Extending the bounds of rationality: evidence and theories of preferential choice. Journal of Economic Literature 44, 631–661.

Roe, R.M., Busemeyer, J.R., Townsend, J.T., 2001. Multialternative decision field theory: a dynamic connectionist model of decision making. Psychological Review 108 (2), 370–392.

Rumelhart, D.L., Greeno, J.G., 1971. Similarity between stimuli: an experimental test of the Luce and restle choice models. Journal of Mathematical Psychology 8, 370–381.

Shipley, E.F., 1960. A model for detection and recognition with signal uncertainty. Psychometrika 25, 273–289.

Simonson, I., Tversky, A., 1992. Choice in context: tradeoff contrast and extremeness aversion. Journal of Marketing Research 29 (3), 281–295.

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.

Tanner, W.P.J., Swets, J.A., 1954. A decision-making theory of visual detection. Psychological Review 61, 401–409.

Thurstone, L.L., 1927. A law of comparative judgment. Psychological Review 34 (4), 273–286.

Train, K., 2003. Discrete Choice Models with Simulation. Cambridge University Press, Cambridge.

Tversky, A., 1972. Elimination by aspects: a theory of choice. Psychological Review 79 (4), 281–299.

Tversky, A., Koehler, D.J., 1994. Support theory: a nonextensional representation of subjective probability. Psychological Review 101 (4), 547–567.

Tversky, A., Russo, J.E., 1969. Substitutability and similarity in binary choices. Journal of Mathematical Psychology 6 (1), 1–12.

Tversky, A., Simonson, I., 1993. Context-dependent preferences. Management Science 39 (10), 1179–1189.

Usher, M., McClelland, J.L., 2004. Loss aversion and inhibition in dynamical models of multialternative choice. Psychological Review 111 (3), 757–769.

Yellott, J.I., 1977. The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. Journal of Mathematical Psychology 15, 109–144.

Decision and Choice: Random Utility Models of Choice and Response Time

Hans Colonius, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany
Anthony AJ Marley, University of Victoria, Victoria, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Horse race random utility models for choice and response time are studied, and conditions on the observable choice probabilities and decision times necessary and sufficient for a random utility representation are investigated. We state conditions for stochastic independence of option chosen from time of choice. Context-free accumulator models, like the linear ballistic accumulator model, are presented as horse race models with additional structure. Models with multiplicative and with additive drift rate variability are discussed.

Introduction

The task of choosing a single 'best' option from some available, potentially infinite, set of options A has received considerable study in psychology and economics. *Random utility models* of choice account for the stochastic variability underlying these choices (i.e., the same choice is not necessarily made on repeated presentations of the same set of options) by assuming that there exists a random variable, $U(x)$ for each option, x and a joint probability measure for these random variables such that the probability of choosing a particular option, x from the set of available options is equal to the probability that $U(x)$ takes on a value greater than the values of all other random variables (*see* Decision and Choice: Luce's Choice Axiom). The basic choice paradigm is extended here by considering, in addition to the option chosen, the point in time that a choice is made from the set of available options. The random utility model can be tailored to cover this situation by replacing the utility variable, $U(x)$ by $T(x) = \phi[U(x)]$ (ϕ, some monotonically decreasing transformation of the values of U into the nonnegative real numbers), where $T(x)$ can be interpreted as the decision time for choosing option x, and by replacing the maximum utility rule by a minimum decision time rule: the option chosen is the one that happens to be associated with the minimum choice (or decision) time with respect to all options in the available set. This model will be referred to as a *horse race random utility model*. Over the past 15 years, psychologists and economists have extended the random utility framework to tasks where a person might be asked to select the 'worst' option, or the 'best' and 'worst' option, from some available choice set. For reasons of space, we do not present parallel results for these cases, but see the results in Hawkins et al. (2014) and the summaries in Marley and Flynn (2015) and Marley and Regenwetter, (in press).

A number of questions arise in the study of horse race random utility models for choice and response time, some of which have received partial answers: (1) What conditions on the observable choice probabilities and decision times are necessary and sufficient for a random utility representation? (2) What are the consequences of assuming stochastic independence between the time of choice and the identity of the option chosen, and which assumptions on the joint distribution function imply this independence? (3) What are possible generalizations to other choice paradigms? The presentation here is partly based on Marley (1989) and Marley and Colonius (1992); those papers present and cite related results that are not explicitly referred to in the following.

Horse Race Random Utility Models

Let $A = \{x, y, \ldots\}$ be a *finite* set of potential choice options (for an extension of the theory to infinite choice sets see Resnick and Roy, 1992); the subset X of A containing at least two elements is the currently available choice set. T_X is a random variable denoting the time at which a choice is made; and B_X is a random variable denoting the element chosen from X. For $t \geq 0$,

$$B_X(x;t) = Pr[B_X = x \cap T_X > t] \qquad [1]$$

is the probability that option x is chosen as the best from X after time t, and $\{B_X(x;t) : x \in X \subseteq A\}$, or (A, P) for short, is called a *joint structure of choice probabilities and response times*. (A, P) is called *complete* if it is defined for all subsets of X of A with $|X| \geq 2$.

A complete joint structure of choice probabilities and response times, (A, P) is said to satisfy a *horse race random utility model* if there exists a probability measure, $Pr_X[\ ]$ for the collection of nonnegative random variables, $\{T(x): x \in X\}$ such that, for all $X \subseteq A$ with $|X| \geq 2$,

$$B_X(x;t) = Pr_X[t < T(x) < T(y) : y \in X \backslash \{x\}] \qquad [2]$$

For simplicity, all distributions are assumed to be absolutely continuous, i.e., they must possess a density function; however, as observed in Resnick and Roy (1992), absolute continuity is not necessary for the results cited here. The representation eqn [2] is of the form given in eqn [1] with $T_X = \min\{T(y): y \in X\}$. Also note that with

$$\begin{aligned} B_X(x) &= \lim_{t \downarrow 0} Pr_X[t < T(x) < T(y) : y \in X \backslash \{x\}] \\ &= Pr_X[T(x) < T(y) : y \in X \backslash \{x\}] \end{aligned}$$

the collection of choice probabilities, $\{B_X(x) : x \in X \subseteq A\}$ constitutes a *system of choice probabilities* with

$$B_X(x) = Pr_X[T(x) = \min\{T(y) : y \in X\}]$$

satisfying a *random utility representation* with min replacing the usual max.

In this article the problem of finding conditions on the – at least theoretically – observable choice probabilities and

response time distributions, which are necessary and sufficient for the existence of a particular representation in terms of a collection of underlying random variables, is referred to as a *characterization problem*.

General characterization problem. Given a (complete) joint structure of choice probabilities and response times what conditions on the (survival) functions, $B_X(x;t)$ are necessary and sufficient for the existence of a (possibly dependent) horse race random utility representation for that joint structure?

A complete answer to this problem is still open. However, a number of results in special cases have been found.

Representation by Independent Random Variables

The following is a partial answer to the general characterization problem:

Theorem 1 (Marley and Colonius, 1992).

a. For a collection of choice probabilities and response times, $\{B_X(x;t): x \in X\}$, on a fixed set X with $|X| \geq 2$, there exist independent random variables, $T(x)$, $x \in X$, with unique distributions, such that eqn [2] holds, provided $B_X(x;t)$ is absolutely continuous and positive for all $t \geq 0$.
b. A complete joint structure of choice probabilities and response times (A, P) can be represented by an independent horse race random utility model, with unique distributions, if the conditions of (a) hold and

$$\left[\sum_{y \in X} B_X(y;t)\right]^{-1} (\mathrm{d}/\mathrm{d}t) B_X(x;t) \qquad [3]$$

is independent of $X \subseteq A$ for all $t \geq 0$.

Thus, given the above regularity conditions on the response time distributions, part (a) shows that *any* set of choice probabilities and response times on a *fixed* finite set, X can be represented by *independent* random variables such that the horse race eqn [2] holds; and part (b) gives sufficient conditions for a common representation across all subsets of A. Townsend (1976) develops a related result in the context of parallel–serial processing analysis. Thus, the general characterization problem is still unsolved leaving open the possibility that for some joint structures no independent representation over all subsets X exists.

Some of the distributions, but not all, in the above solutions may be *improper*, i.e., they may have positive measure at infinity implying an infinite decision time for that alternative, even if the distribution of response time, T_X is finite almost surely. Jones and Dzhafarov (2014) present related interpretations and further falsifiability results for horse race and other models of choice and response time.

The above framework has natural extensions to subset choice (Marley and Colonius, 1992) and to the worst and best–worst choice (Marley and Regenwetter, in press). Also, in the theory of competing risks (e.g., David and Moeschburger, 1978) the 'event' associated with each $x \in X$ is reinterpreted as a *cause* (e.g., of failure or death) and thus the selection of $x \in X$ then corresponds to the cause associated with x being the first to occur when X contains the possible causes. This leads to *cause-specific* (or *crude hazard rate*) interpretations of the results and related *nonidentifiability* issues (Marley and Colonius, 1992).

Independence of Option Chosen from Time of Choice

A classic model for best choice is *Luce's choice model* (*see* Decision and Choice: Luce's Choice Axiom), also called the *multinomial logit (MNL) model* (Marley and Regenwetter, in press). That model holds for a system of choice probabilities, $\{B_X(x) : x \in X \subseteq A\}$ provided there is a ratio scale b on A such that for $B_X(x) \neq 0, 1$

$$B_X(x) = \frac{b(x)}{\sum_{y \in X} b(y)} \qquad [4]$$

For simplicity in the following, we assume that all choice probabilities are nonzero. The result can be generalized when this is not the case by adding a connectivity and a transitivity condition (Luce, 1959: Theorem 4, p. 25).

Theorem 2 (Marley and Colonius, 1992)

Consider a (complete) independent horse race random utility model, (A, P) where for each $x \in X \subseteq A$ with $|X| \geq 2$, $B_X(x;t)$ is positive and absolutely continuous for all $t > 0$. If T_X is stochastically independent of B_X, then the choice probabilities satisfy Luce's choice model.

Given the known limited empirical validity of Luce's choice model in empirical preference situations, this result implies that one must study *dependent* horse race random utility models and/or drop the assumption of independence between T_X and B_X. We now present results for such generalizations.

Generalized Stable Survival Functions

A large class of (possibly) dependent horse race random utility models based on extreme value distributions can be generated from the concept of a *generalized stable survival function*. For simplicity, in this section, assume $A = \{x_1, \ldots, x_n\}$. We call

$$P_A(t_1, \ldots, t_n) = Pr[T(x_1) > t_1, \ldots, T(x_n) > t_n)$$

a *(multivariate) survival function*. P_A is called *generalized stable* if there is a strictly monotone decreasing function, η and a constant, $\mu > 0$ such that for all $\alpha > 0$, $t_i \geq 0$, $i = 1, \ldots, n$,

$$(\eta \circ P_A)(\alpha t_1, \ldots, \alpha t_n) = \alpha^{\mu} (\eta \circ P_A)(t_1, \ldots, t_n) \qquad [5]$$

where $\circ$ denotes the concatenation of functions. Letting $G_A = (\eta \circ P_A)$, this amounts to

$$G_A(\alpha t_1, \ldots, \alpha t_n) = \alpha^{\mu} G_A(t_1, \ldots, t_n)$$

i.e., G_A is homogeneous of degree μ. Furthermore, a survival function P_A is called a *strictly monotone transform of a generalized stable survival function*, Q_A if there is a strictly monotone increasing function γ (of the random variables associated with Q_A) with $\gamma(0) = 0$, $\gamma(\infty) = \infty$ such that for $t_i \geq 0$, $i = 1, \ldots, n$,

$$P_A(t_1, \ldots, t_n) = Q_A[\gamma(t_1), \ldots, \gamma(t_n)]$$

Theorem 3 (Marley, 1989)

Any horse race random utility model that is generated by a strictly monotone transform of a generalized stable survival function is such that T_X is stochastically independent of B_X for any $X \subseteq A$.

Resnick and Roy (1992: Section 5) construct an example, which shows that the converse of this result does not hold in general, i.e., independence between T_X and B_X does not lead back to the class of generalized stable survival functions. However, Robertson and Strauss (1981) show that the converse is true if the survival function, P_A (or some strictly monotone transformation of it) belongs to the generalized Thurstone class: for $t_i \geq 0$, $i = 1, \ldots, n$,

$$P_A(t_1, \ldots, t_n) = P(u(x_1)t_1, \ldots, u(x_n)t_n)$$

where $u(x_i) \geq 0$ and P is a survival function that is independent of A.

Marley (1989) uses generalized stable survival functions to extend a large class of frequently applied random utility models for choice to models of choice and response time; these models of choice include McFadden's *generalized extreme value* class, Tversky's *elimination by aspects* model, and Strauss's *general feature* model. However, by the above results, all these models predict the independence of the choice made and the time to make it, and so, in particular, that the response time distribution is the same for all choices. Given that the latter condition is not satisfied by various data (Hawkins et al., 2014), we now turn to related models that do not have this property.

The results above were stated in terms of the *survival* form

$$B_X(x; t) = \Pr(B_X = x \text{ and } T_X > t)$$

If we now let $B_X(x,t)$ denote the probability that option x is chosen *at or before* time t (note the change of the semicolon ; to a comma ,), then those results can be written equally well in the *cumulative* form

$$B_X(x, t) = \Pr(B_X = x \text{ and } T_X \leq t) \quad [6]$$

Since most response time models are written in terms of such cumulative distributions, we will use the latter form in the remaining sections.

Context-Free Linear Accumulator Models

Context-free accumulator models are essentially horse race random utility models with additional structure imposed on the properties of the random variables; later we summarize context-dependent accumulator models. Here, we discuss the *linear ballistic accumulator* (LBA) model (Heathcote and Love, 2012). Like other multiple accumulator models, the LBA is based on the idea that the decision maker accumulates evidence in favor of each choice, and makes a decision as soon as the evidence for any choice reaches a threshold amount. The time to accumulate evidence to threshold is the predicted decision time, and the response time is the decision time plus a fixed offset (t_0), the latter accounting for processes such as response production. The LBA shares this general evidence accumulation framework with many models (Busemeyer and Rieskamp, 2013; *see* Diffusion and Random Walk Processes) but has a practical advantage – namely, the computational tractability afforded by having an easily computed expression for the joint likelihood of a given response time and response choice among any number of options.

In the following sections, we consider two classes of LBA – the first class relates in a natural way to the models already presented, the second class agrees with the assumptions in the original development of the LBA. We first summarize context-independent models, and then summarize context-dependent extensions; we ignore the fixed offset, (t_0).

Multiplicative Drift Rate Variability

Let Δ be a random variable on the positive real numbers; this will be the distribution of the (multiplicative) drift rate variability, where *drift rate* is the rate of accumulation of evidence. Let, for $r > 0$,

$$\Pr(\Delta \leq r) = G(r) \quad [7]$$

where G is a cumulative distribution function (CDF). Let b denote the threshold, x a typical choice option with drift rate $\mathrm{d}(x)$, and Σ a random variable on $[0,A]$ with $A \leq b$; this will be the distribution of the start point of the evidence accumulation. Let, for $h \in [0,A]$,

$$\Pr(\Sigma \leq h) = H(r)$$

where H is a CDF.

For option x, the start point has distribution, Σ and the drift rate has distribution, $\Delta\mathrm{d}(x)$. We assume that the samples of Σ and Δ for different values of x are independent, which we denote by Σ_x and Δ_x; thus, the subscripts do not imply dependence on x beyond the assumption of independent samples for different values of x. Then, with b_x denoting the random variable for the time taken for the LBA with (mean) drift rate, $\mathrm{d}(x)$ to reach threshold, we have

$$\Pr(b_x \leq t) = \Pr\left(\frac{b - \Sigma_x}{\Delta_x \mathrm{d}(x)} \leq t\right) \quad [8]$$

The major LBA applications to date have assumed that the start point variability is given by a random variable, p with uniform distribution on $[b-A, b]$, in which case eqn [8] can be rewritten as

$$\Pr(b_x \leq t) = \Pr\left(\frac{p_x}{\Delta_x \mathrm{d}(x)} \leq t\right) \quad [9]$$

Then the choice and response time representation, $B_X(x,t)$ of eqn [6] holds with $T_X = \min\{b_\gamma: \gamma \in X\}$.

The above representation is a perfectly plausible multiplicative LBA model for *any* cumulative distribution G on the nonnegative reals and H on $[0,A]$ with $A \leq b$. However, a major argument advanced for LBA models is their computational tractability in terms of *probability density functions* (PDFs) and CDFs. In particular, it is usually assumed that the accumulators are independent, and therefore the main need is for the PDF and CDF of eqn [9] to be tractable. We summarize the components of the CDF corresponding to that expression, and then summarize what is known about such forms.

With p, the uniform distribution on $[b-A, b]$, and the CDF of G given by eqn [7], standard calculations show that $Pr(b_x \leq t)$ of eqn [9] can be written in terms of: the constants b and A; the mean of the distribution G when truncated to the interval, $[b/t\ \mathrm{d}(x),\ b-A/t\ \mathrm{d}(x)]$; and the values of $G[b-A/t\ \mathrm{d}(x)]$ and $G[b/t\ \mathrm{d}(x)]$ (Marley and Regenwetter, in press). Closed forms are known for the first quantity for various CDFs on the nonnegative reals, including for Gamma, inverted Beta, Fréchet, and Levy distributions (Nadarajah, 2009). However, one of the major motivations for the LBA framework was to achieve easily computable forms for the joint distribution of choice

probabilities and choice times – i.e., for eqn [6]. In addition to independence, this requires the PDFs corresponding to the CDFs, above, to be easily computable. Ongoing research is exploring the extent to which this is the case for various distributional assumptions. For example, Heathcote and Love (2012) develop and test the tractable *lognormal race*, which has the form: for each option x,

$$\Pr(\mathrm{b}_x \le t) = \Pr\left(\frac{S_x}{D_x} \le t\right)$$

where S_x is the distribution to the boundary and D_x is the drift rate distribution, with each being lognormally distributed.

Relation of Multiplicative LBA Models with No Start Point Variability and Fréchet Drift Rate Variability to Luce (MNL) Models

We now present a multiplicative LBA with no start point variability and Fréchet drift scale variability that leads to the Luce (MNL) model for the best choice; parallel results hold for the worst and best–worst choice (Marley and Regenwetter, in press). Consider the special case of eqn [9] where there is no start point variability, i.e., $A = 0$, and so p has a constant value, b; in this case, without loss of generality, we can set $b = 1$. Then eqn [9] reduces to

$$\Pr(B_x \le t) = \Pr\left(\frac{1}{\Delta \mathrm{d}(x)} \le t\right)$$

We also assume that Δ has a Fréchet distribution, i.e., there are constants α, $\beta > 0$ such that, for $r \ge 0$,

$$\Pr(\Delta \le r) = G(r) = e^{-(\alpha r)^{-\beta}} \qquad [10]$$

Then eqn [6] is (remember that the subscript on Δ_z indicates independent samples, not other dependence on z):

$$B_X(x,t) = \Pr\left(\frac{1}{\Delta_x \mathrm{d}(x)} = \min_{z \in X} \frac{1}{\Delta_z \mathrm{d}(z)} \le t\right) \qquad [11]$$

For $t \ge 0$, let $F(t) = 1 - \exp - (\sum_{z \in X} \mathrm{d}(z)^{\beta})(\alpha t)^{\beta}$. Then routine calculations (paralleling those in Marley, 1989) give: for $X \subseteq A$ and $x, y \in X$,

$$B_X(x;t) = \frac{\mathrm{d}(x)^{\beta}}{\sum_{z \in X}^{n} \mathrm{d}(z)^{\beta}} F(t) \qquad [12]$$

The choice probabilities, $B_X(x)$ are given by this formula in the limit as $t \to \infty$, i.e., as $F(t) \to 1$, which shows that they satisfy Luce's choice model, eqn [4], with $b(z) = d(x)^{\beta}$. Now, for $z \in A$, let $u(z) = \ln \mathrm{d}(z)$, $\varepsilon_z = \ln \Delta_z$. Then the form for the choice probabilities can equally well be written as: for all $x, y \in X \in D(A)$,

$$B_X(x) = \Pr\left\{u(x) + \varepsilon_y = \max_{z \in X}[u(z) + \varepsilon_z]\right\} \qquad [13]$$

However, given that Δ_z and $\Delta_{p,q}$ are generated by Fréchet drift rate variability, i.e., eqn [10], we have that $\varepsilon_z = \ln \Delta_z$ and $\varepsilon_{p,q} = \ln \Delta_{p,q}$ satisfy extreme value distributions (with, in this case: for $-\infty < t < \infty$ $\Pr(\varepsilon_z \le t) = \exp - (\alpha e^t)^{-\beta}$ and $\Pr(\varepsilon_{p,q} \le t) = \exp - (\alpha e^t)^{-\beta}$). Under these conditions, eqn [13] corresponds to the usual representation of Luce's choice model as a random utility model (Marley and Louviere, 2008, used the case $\alpha = \beta = 1$). Note that β is not identifiable from the choice probabilities, but it is, in general, identifiable when response times are also available (though the form predicted for the response time distribution is not suitable for data).

As already shown by eqn [12], this model has the property that the option chosen is independent of the time of choice; that property would not hold for (most) distributions other than the Fréchet for drift scale variability, nor does it hold when the above model is extended to include start point variability.

Additive Drift Rate Variability

In contrast to the multiplicative drift rate model presented above, the *LBA model* assumes *additive* drift rate variability generated by independent normal random variates, truncated at zero – i.e., constrained to be nonnegative (Heathcote and Love, 2012). This model is a horse race random utility model, but does not have the undesirable property that the choices made and the time to make them are independent. Nonetheless, it has been shown to make extremely similar predictions for the best, worst, and best–worst choice probabilities to those made by the multiplicative LBA model with no start point variability and Fréchet drift scale variability (Hawkins et al., 2014). The derivations for this model exactly parallel those for the multiplicative LBA model, above, with $\Delta \mathrm{d}(x)$ replaced by $\mathrm{trunc}[D_z + \mathrm{d}(z)]$, with a major advantage of the additive model being that all formulas for CDFs and PDFs are computationally very tractable when the start point distribution is uniform. Marley and Regenwetter (in press) briefly summarize recent applications of this model to some standard data on the best, worst, and best–worst choice and its extension to handle context effects in the best choice and response times.

Context-Dependent Models of Response Time and Choice

As far as we know, there is no formal definition of what it means for a set of (theoretical or empirical) choice probabilities to demonstrate a *context effect*. The general usage of that term appears to be that the choice probabilities are not consistent with either a *random utility* or a *constant utility* representation as defined in Chapter 10 of Suppes et al. (1989). Busemeyer and Rieskamp (2013) and Rieskamp et al. (2006) summarize standard context effects and the extant models that are, or are not, compatible with those effects holding. They do not discuss the recent *multiattribute linear ballistic accumulator (MLBA) model* (Trueblood et al., 2014) or the *2N-ary tree model* (Wollschläger and Diederich, 2012), to which we now turn.

Trueblood et al. (2014) show that their *context-dependent* MLBA model can produce various standard context effects in choice and various properties of the context effects when participants are under time pressure. These results are obtained by adding a context-dependent 'front end' to a context-independent LBA model. This suggests taking a fresh look at various choice-only models that can handle some or all such context effects in choice, and extending them to response time by, say, adding a context-independent LBA model that is driven by the 'front end' parameters of the choice model (see Marley and Regenwetter, in press).

The *2N-ary choice tree model* for N-alternative preferential choice (Wollschläger and Diederich, 2012) builds on earlier stochastic dynamic models (*accrual-halting models*, Townsend

et al., 2012) and thus includes details of the temporal accumulation of information. The structures, though not the details, of the 2*N*-ary choice tree model and of the context-dependent MLBA model are rather similar – the major differences are between the ballistic accumulators of the MLBA model and the transition probabilities of the 2*N*-ary choice tree model, and the fact that the MLBA model has no decay, whereas the 2*N*-ary choice tree model has decay. Each model well fits the data presented by its authors.

To date, the 2*N*-ary choice tree model has been applied to the best choice, though it includes a mechanism for rejecting options. Thus, it is of interest to extend the 2*N*-ary choice tree model to the best–worst choice and compare its fits to the data from such choice with those of the *parallel best–worst LBA model* (Hawkins et al., 2014).

Discussion and Conclusion

In contrast to the context-dependent MLBA model, the 2*N*-ary choice tree model and related stochastic dynamic models specify details of the temporal accumulation of information, rather than having a front end that incorporates the overall effect of such accumulated information. A significant part of the reason such stochastic dynamic models have not been used in applied areas is the "relatively high difficulty of estimating some sequential sampling models of decision making" (Berkowitsch et al., 2013). This difficulty is being alleviated, for example, by Berkowitsch et al.'s (2013) derivation of a tractable closed form for the choice probabilities given by an asymptotic form of decision field theory and by the various LBA models for (best, worst, best–worst) choice and response times.

The context-dependent models of choice and response time presented in this article involve numerous complex processes. Further empirical study along the lines of Teoderescu and Usher (2013), with follow-up theoretical work, is needed to clarify the conditions under which some or all these, and possibly other, processes are needed.

Nonetheless, the introduction of the notion of response (or decision) time into the random utility approach to modeling choice behavior has been very fruitful. It has brought about new insights into existing stochastic choice models and their characterizations, and it has generated new questions, many of which have only found partial answers up to now.

See also: Decision and Choice: Luce's Choice Axiom; Diffusion and Random Walk Processes.

Bibliography

Berkowitsch, N.A.J., Scheibehenne, B., Rieskamp, J., 2013. Testing multialternative field theory rigorously against random utility models, Manuscript.

Busemeyer, J.R., Rieskamp, J., 2013. Psychological research and theories on preferential choice. In: Hess, S., Daly, A. (Eds.), Handbook of Choice Modelling. Edward Elgar Publishing.

David, H.A., Moeschburger, M.L., 1978. The Theory of Competing Risks. Griffin, London.

Hawkins, G., Marley, A.A.J., Heathcote, A., Flynn, T., Louviere, J., Brown, S.D., 2014. Integrating cognitive process and descriptive models of attitudes and preferences. Cognitive Science 38, 701–735.

Heathcote, A., Love, J., 2012. Linear deterministic accumulator models of simple choice. Frontiers in Psychology 3, 1–19.

Jones, M., Dzhafarov, E.N., 2014. Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. Psychological Review 121, 1–32.

Luce, R.D., 1959. Individual Choice Behavior. Wiley, New York.

Marley, A.A.J., 1989. A random utility family that includes many of the "classical" models and has closed form choice probabilities and reaction times. British Journal of Mathematical and Statistical Psychology 42, 13–36.

Marley, A.A.J., Flynn, T.N., 2015. Best worst scaling: theory and practice. In: Wright, James D. (Editor-in-Chief), International Encyclopedia of the Social and Behavioral Sciences, second ed. Elsevier, Oxford.

Marley, A.A.J., Louviere, J.J., 2008. Some probabilistic models of best, worst, and best-worst choice. Journal of Mathematical Psychology 44, 464–480.

Marley, A.A.J., Regenwetter, M. Choice, preference, and utility: probabilistic and deterministic representations. In: Batchelder, W., Colonius, H., Dzhafarov, E., Myung, J. (Eds.), New Handbook of Mathematical Psychology. Cambridge University Press, Cambridge, in press.

Marley, A.A.J., Colonius, H., 1992. The horse race random utility model for choice probabilities and reaction times, and its competing risks interpretation. Journal of Mathematical Psycholology 36, 1–20.

Nadarajah, S., 2009. Some truncated distributions. Acta Applied Mathematics 106, 105–123.

Resnick, S.I., Roy, R., 1992. On min-stable horse races with infinitely many horses. Mathematical Social Sciences 23, 119–145.

Rieskamp, J., Busemeyer, J., Mellers, B., 2006. Extending the bounds of rationality: evidence and theories of preferential choice. Journal of Economic Literature 44, 631–661.

Robertson, C.A., Strauss, D.J., 1981. A characterization theorem for random utility variables. Journal of Mathematical Psycholology 23, 184–189.

Suppes, P., Krantz, D.H., Luce, R.D., Tversky, A., 1989. Foundations of Measurement, vol. II. Academic Press, San Diego, CA.

Teoderescu, A., Usher, M., 2013. Disentangling decision models – from independence to competition. Psychological Review 120, 1–38.

Townsend, J.T., 1976. Serial and within-stage independent parallel model equivalence on the minimum completion time. Journal of Mathematical Psycholology 14, 219–238.

Townsend, J., Houpt, J., Silbert, N., 2012. General recognition theory extended to include response times: predictions for a class of parallel systems. Journal of Mathematical Psychology 56, 476–494.

Trueblood, J.S., Brown, S.D., Heathcote, A., 2014. The multiattribute linear ballistic accumulator model of context effects in multialternative choice. Psychological Review 121, 179–205.

Wollschläger, L., Diederich, A., 2012. The 2n-ary choice tree model for n-alternative preferential choice. Frontiers in Psychology 3 (189).

Decision and Choice: Sequential Decision Making

Adele Diederich, Jacobs University Bremen, Bremen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sequential decision making describes a situation where the decision maker makes successive observations of a process before a final decision is made. The procedure to decide when to stop taking observations and when to continue, the optimal stopping rule, depends on the specific assumptions made about the situation: the knowledge about the distribution of the observations, the distribution properties of the observations, whether the number of possible observations is bounded or unbounded, the sampling procedure, and the cost function for each observation. Many of these problems have been studied theoretically by mathematicians and experimentally by psychologists.

Sequential decision making describes a situation where the decision maker (DM) makes successive observations of a process before a final decision is made, in contrast to dynamic decision making (*see* Dynamic Decision Making), which is more concerned with controlling the system over time.

Formally, a sequential decision problem is such that the DM can take observations X_1, X_2, … one at a time. After each observation X_n, the DM can decide to terminate the process and make a final decision from a set of decisions D, or continue the process and take the next observation X_{n+1}. If the observations X_1, X_2, … form a random sample, the procedure is called sequential sampling.

In most sequential decision problems, there is an implicit or explicit cost associated with each observation. The procedure to decide when to stop taking observations and when to continue is called the stopping rule. The objective in sequential decision making is to find a stopping rule that optimizes the decision in terms of minimizing losses or maximizing gains including observation costs. The optimal stopping rule is also called the optimal strategy or the optimal policy.

A wide variety of sequential decision problems have been discussed in the statistics literature, including search problems, inventory problems, gambling problems, and secretary-type of problems, including sampling with and without recall. Several methods have been proposed to solve the optimization problem under specified conditions, including dynamic programing (*see* Dynamic Programing), Markov chains, and Bayesian analysis.

Sequential decision making plays an important role in applied research such as operations research, clinical research, computer science, economics, and political science with a variety of topics; for instance, clinical trials, structural learning, search algorithm, tutoring systems, marketing, portfolio, voting, and many more.

In the psychological literature, sequential decision problems are better known as optional stopping problems, optimal stopping problems, and sequential search problems. One line of research using sequential decision making is concerned with seeking information in situations such as buying houses, searching a job candidate, price searching, or target search. The DM continues taking observations until a decision criterion for acceptance is reached. Another line of research applies sequential decision making to account for information processing in binary choice tasks (*see* Diffusion and Random Walk Processes) and hypothesis testing such as in signal detecting tasks (*see* Signal Detection Theory). The DM continues taking observation until either one of two decision criteria is reached. Depending on the particular research area, observations are also called offers, options, items, applicants, information, and the like. Observation cost explicitly includes not only possibly money but also time, effort, aggravation, discomfort, and so on. Contrary to the objective of statisticians or economists, psychologists are less interested in determining the optimal stopping rule but are interested to find variables that affect human decision behavior on sequential decision tasks. Optimal decision strategies are considered as normative models and their predictions are compared to actual choice behavior.

Interestingly, research on sequential decision making with only one criterion received only moderate attention in psychology over the last decade or so, whereas research on sequential decision making with two criteria enjoy great popularity (*see* Diffusion and Random Walk Processes). The main focus here is on the former approach.

Sequential decision Making with One Decision Criterion

In sequential decision making with one decision criterion, the DM takes costly observations X_n, $n = 1, \ldots$ of a random process one at a time. After observing $X_n = x_n$, the DM has to decide whether to continue sampling observations or to stop. In the former case, the observation X_{n+1} is taken at a cost of c_{n+1}; in the latter case, the DM receives a net payoff that consists of the payoff minus the observation costs. The DM's objective is to find a stopping rule that maximizes the expected net payoff.

The optimal stopping rule depends on the specific assumptions made about the situation: (1) the distribution of X is known, not known, or partly known; (2) X_i are identically distributed for all i have the same distribution but with different parameters, or have different distributions; (3) the number of possible observations, n, is bounded or unbounded; (4) the sampling procedure, for example the option to take the highest value observed so far when stopping

 http://dx.doi.org/10.1016/B978-0-08-097086-8.43089-8

(sampling with recall) or only to take the last value when stopping (sampling without recall); and (5) the cost function, c_n, is fixed for each observation or is a function of n. Many of these problems have been studied theoretically by mathematicians and experimentally by psychologists. Pioneering experimental work was done in a series of papers by Rapoport and colleagues (1966, 1970, 1972).

Unknown Sample Distribution: Secretary-Type of Problems

Khan et al. (1967) investigated decision behavior in a sequential search task where the DM had to find the largest of a set of $n = 200$ numbers, observed one at a time from a deck of cards. The observations were taken in random order without replacement. The DM could only declare the current observation as the largest number (sampling without recall), and could compare the number with the previous presented numbers. No explicit cost for each observation were taken, that is $c = 0$. The sample distribution was unknown to the DM. A reward was paid only when the card with the highest number was selected, and nothing otherwise. The experiment describes a decision situation that is known as the secretary problem (a job candidate search problem; for various other names see, e.g., Freeman, 1983), which, in its simplest form, makes explicit the following assumptions (Ferguson, 1989):

1. Only one position is available.
2. The number n of applicants is finite and known.
3. The n applicants are interviewed sequentially in random order, each of the $n! = n\cdot(n-1)\cdot\cdots\cdot 2\cdot 1$ orders being equally likely.
4. All applicants can be ranked from best to worst without ties – the decision to reject or accept an applicant must be based only on the relative ranks of the applicants interviewed so far.
5. An applicant once rejected cannot later be recalled.
6. The payoff is 1 when choosing the best (i.e., the one with absolute rank 1) of the n applicants, 0 otherwise.

The optimal strategy for this kind of problems is to reject the first $s-1$, $s \geq 1$ items (cards, applicants, draws) and then to choose the first item that is best in the relative ranking so far. With

$$a_s \equiv \frac{1}{s} + \frac{1}{s+1} + \cdots + \frac{1}{n-1} \qquad [1]$$

the optimal strategy is to stop if $a_s < 1$ and to continue if $a_s > 1$, which can easily be determined for small n. For large n, the probability of choosing the best item is approximated by $1/e$ and the optimal s by n/e ($e = 2.71\ \ldots$). (For derivations, see, e.g., DeGroot, 1970; Freeman, 1983; Gilbert and Mosteller, 1966.)

Khan et al. (1967) found that about 40% of their participants did not follow the optimal strategy but stopped too late and rejected a card that should have been accepted. The failure of the strategy for describing behavior was assigned to its inadequacy for the described task. Although at the beginning of the experiment the participants did not know anything about the distribution (requirement), they could learn about the distribution by taking observations (partly information). To guarantee ignorance of the distribution, Gilbert and Mosteller (1966) recommended supplying only the rank of the observation made so far and not the actual value. Seale and Rapoport (1997) conducted an experiment following this advice. They found that participants stopped too early (with $n = 40$ and $n = 80$) than prescribed by the optimal stopping rule. They proposed simple decision rules or heuristics to describe choice behavior. Using a cutoff rule, the DM rejects the first $s-1$ applicants and then chooses the next top-ranked applicant, that is the candidate. The DM simply counts the number of applicants and the stops on the first candidate after observing $s-1$ applicants. Under a candidate count rule, the DM counts the number of candidates and chooses the jth candidate after observing $j-1$ candidates. A successive noncandidate rule requires the DM to choose the first candidate after observing at least k consecutive noncandidates following the last candidate. All three heuristics require minimal cognitive effort: merely counting applicants, candidates, or successive noncandidates is needed and comparing this number to some fixed parameter value. Seale and Rapoport (1997) results showed that the cutoff rule, which includes the optimal policy, accounted for the greatest proportion of choices.

The secretary problem has been extended and generalized in many different directions within the mathematical statistics field. Each of the above assumptions has been relaxed in one or the other way and optimal solutions were proposed (see Ferguson, 1989; Freeman, 1983, for reviews). Some of the generalizations have been tested empirically. For instance, Bearden and Rapoport (2005) described a so-called generalized secretary problem (GSP) in which the DM's objective is to find a good and not necessarily the best. The payoff function, 1 when selecting the best applicant and 0 otherwise, is replaced by a more general payoff function $\pi(a)$ for selecting an applicant with absolute rank a where $\pi(1) \geq \ldots \geq \pi(n)$. Experimental results showed that the DM stopped, on average, the search too early than predicted by the optimal decision rule (Bearden et al., 2006). Bearden et al. (2005) proposed a further generalization of the GSP to meet even more realistic decision situations. In the multiattribute secretary problem, the applicants differ along k different attributes. Within a given attribute, the applicants can be ranked from best (1) to worst (n) without ties. The attributes are correlated. The payoff function is $\pi^i(a^i)$, where a^i is the selected applicant's absolute rank on the ith attribute i and $\pi^1(1) \geq \ldots \geq \pi^i(n)$. Experimental results showed again that the DM examined fewer options, on average, than predicted by the optimal decision rule.

Known Sample Distribution

Rapoport and Tversky (1966, 1970) investigated choice behavior when the mean and the variance of the distribution were known to the DM. The cost for each observation was fixed but the amount varied across experimental conditions and the number of possible observations n was unbounded (1966) or bounded and known (1970). Behavior for sampling with and without recall was compared. When sampling is without recall, only the value of the last observation, $X_n = x_n$, can be received and the payoff is this value minus the total sampling cost, that is $x_n - cn$. The optimal strategy is to find a stopping

rule that maximizes the expected payoff $E(X_N - cN)$. When sampling is with recall, the highest value observed so far can be selected and the payoff is $\max(x_1, \ldots, x_n) - cn$ and the optimal strategy is to find a stopping rule that maximizes $E(\max(X_1, \ldots, X_N) - cN)$. In the following, v with subscripts and v^* denote the expected gain from an (optimal) procedure.

Number of Observations Unbounded

If n is unbounded, that is if the number of observations that can be taken is unlimited, and $X_1, X_2, \ldots$ are sampled from a known distribution function $F(x)$, the optimal strategy is the same for both sampling with and without recall. In particular, the optimal strategy is to continue to take observations whenever an observed value $x_j < v^*$ and to stop taking observations as soon as an observed value $x_j \geq v^*$, where v^* is the unique solution of

$$\int_{v^*}^{\infty} (x - v^*)\mathrm{d}F(x) = c \qquad -\infty < v^* < \infty. \qquad [2]$$

When the observations are taken from a standard normal distribution with density functions $\phi(x)$ and distribution function $\Phi(x)$, one has that

$$v^* = \frac{\phi(v^*) - c}{1 - \Phi(v^*)}. \qquad [3]$$

Although sampling with and without recall has the same solution, they seem to be different from a psychological point of view. Rapoport and Tversky (1966) found that the group sampling without recall took significantly fewer observations than the participants sampling with recall. The mean number of observations for both groups decreased with increasing cost c, and the difference with respect to the number of observations taken was diminished. However, the participants of both groups took fewer observations than prescribed by the optimal strategy. This nonoptimal behavior of the participants was attributed to a lack of thorough knowledge of the distributions.

Number of Observations Bounded

If $n, (n \geq 2)$ is bounded, that is, if not more than n observations can be taken, the optimal stopping rules for sampling with and without recall are different. For sampling without recall, an optimal procedure is to continue taking observations whenever $x_j < v_{n-j} - c$ and to stop as soon as $x_j \geq v_{n-1} - c$, where $j = 1,2, \ldots, n$ indicates the number of observations that remain available and

$$v_{j+1} = (v_j - c) - \int_{v_j - c}^{\infty} (x - (v_j - c))\mathrm{d}F(x). \qquad [4]$$

With $v_1 = E(X) - c$, the sequence can be computed successively. Again, assuming a standard normal distribution $v_{j+1} = \phi(v_j - c) + \Phi(v_j - c)$.

For sampling with recall, the optimal strategy is to continue the process whenever a value $x_j < v^*$ and to stop taking observations as soon as an observed value $x_j \geq v^*$, where v^* is as in eqn [2], which is the same as for n unbounded. (For derivations of the strategies, see DeGroot, 1970; Sakaguchi, 1961.)

Rapoport and Tversky (1970) investigated choice behavior within this scenario. Sampling was done with and without recall. The number of observations that could be taken as well as observation cost varied across experimental groups. One-third of the participants did not follow the optimal strategy. Under both sampling procedures and all cost conditions, they took on average fewer observations than predicted by the corresponding optimal stopping rules. There were no systematic differences due to cost as observed in their previous study. They concluded that "the optimal model provides a reasonable good account of the behavior of the subjects" (p. 119).

Different Sample Distributions for Each Observation

Most research concerned with sequential decision making assumes that the observations are sampled from the same distribution; that is X_i are identically distributed for all i. For many decision situations, however, the observations may be sampled from the same distribution family with different parameters, or from different distributions. Especially in economical areas, such as price search, it is reasonable to assume that the distributions from which observations are taken change over time. The sequence of those samples has been called nonstationary series. Of particular interest are two special nonstationary series: ascending and descending series. For ascending series, the observations are drawn from distributions, usually from normal distributions, with increasing mean as i increases; for descending series the mean of the distribution decreases as i increases, i indicating the sample index. For both cases, experiments have been conducted to investigate choice behavior in a changing environment. Shapira and Venezia (1981) compared choice behavior for ascending, descending, and constant (identically distributed X_i) series. In one experiment (numbers from a deck of cards), the distributions were known to the DM; no explicit observation cost were taken; sampling occurred without recall; and the number of observations that could be taken was limited to $n = 7$. The variance of the distributions varied across experimental groups. An optimal procedure was assumed to continue taking observations whenever $x_j < v_{n-j}$ and to stop as soon as $x_j \geq v_{n-1}$, where $j = 1,2, \ldots, n$ indicates the number of observations that remain available. $k = 1, \ldots, n$ indicates the specific distribution for the jth observation. Thus,

$$v_{j+1} = v_j - \int_{v_j}^{\infty} (x - v_j)\mathrm{d}F_k(x). \qquad [5]$$

With $v_1 = E(X_1)$, the sequence can be computed successively. Assuming a standard normal distribution $v_{j+1} = \phi_k(v_j) + \Phi_k(v_j)$.

Across all conditions, 58% of the participants behaved in an optimal way. The proportion of optimal stopping did not depend on the type of the series but depended on the size of the variance. Nonoptimal stopping (24% stopped too early; 18% too late) depended on the series and on the size of the variance. In particular, participants stopped too early on ascending and too late on descending series. A similar result was observed by Brickman (1972). In that study, departing from the optimal stopping rule was attributed to an inadequacy of the

stopping rule taken for the particular experimental conditions (assuming complete knowledge of the distributions). In a secretary problem design (*see* Unknown Sample Distribution: Secretary-Type of Problems), Corbin et al. (1975) were less concerned with optimal choice behavior than with the processes by which the participants made their selections, and with factors that influenced those processes. The emphasis of the investigation was on decision-making heuristics rather than on the adequacy of optimal models. With the same optimal stopping rule for all experimental conditions, they found that stopping behavior depended on contextual variables such as the ascending or descending trend of the inspected numbers of the stack.

In summary, Eriksson and Strimling (2010) argue that suboptimal search behavior is to be expected because optimal strategies differ greatly with subtle details of the search problem setup. People, however, seem to hold general intuitions about the optimal amount of search to be conducted before stopping. Generally, they are not able to foresee how the optimal amount of search depends on details of the problem and therefore, they will be incorrect for most problems.

Search Problems – Multiple Information Sources

In a sequential decision-making task with multiple information sources, the DM has the option to take information sequentially from different sources. Each information source may provide valid information with a particular probability and at different cost. The tasks are to decide not only to stop or to continue the process but also, if continuing, which source of information to consult.

Early experimental studies have been done by Kanarick et al. (1969), Rapoport (1969), and Rapoport and Wallsten (1972). A typical task is to find an object (e.g., a black ball) that is hidden in one of several possible locations (e.g., in one of several bins containing white balls). The optimal search strategy depends on further task specifications, such as whether the object can move from one location to another, how many objects are to be found, and whether the search process may stop before the object has been found. Rapoport (1969) investigated the case when one single object that could not move was to be found in one of r, $r \geq 2$ possible locations. The DM was not allowed to stop the process before the target was found. All of the following were known to the DM: the priori probability p_i, $p_i > 0$ that the object is in location i, $i = 1,2, \ldots, r$ with $\sum_{i=1}^{r} p_i = 1$; miss probability α_i, $0 < \alpha_i < 1$, that even if the object is in location i it will not be found in a particular search of that location ($1 - \alpha_i$ is referred to the respective detection probability); and a cost, c_i, for a single observation at location i. The objective of the DM is to find a search strategy that minimizes the expected cost. For $i = 1, \ldots, r$ and $j = 1,2, \ldots$, let Π_{ij} denotes the probability that the object is found at location i during the jth search and the search is terminated. Then,

$$\Pi_{ij} = p_i \alpha_i^{j-1}(1 - \alpha_i), \quad i = 1, \ldots, r, \quad j = 1, 2, \ldots \qquad [6]$$

If all values of Π_{ij}/c_i for all values of i and j are arranged in order of decreasing magnitude, the optimal strategy is to search according to this ordering (for derivations, see DeGroot, 1970). Ties may be arbitrarily ordered among themselves. The optimal strategy is determined by the detection probabilities and observation costs, and optimal search behavior implies a balance between maximizing the detection and minimizing the observation cost. Rapoport (1969) found that participants did not behave optimally. They were more concerned with maximizing the probability of detecting the target than with minimizing observation cost. Increasing the difference of observation cost c_i among the $i = 1, 2, 3, 4$ locations showed that the deviation from the optimal strategy even increased. Rapoport and Wallsten (1972) varied the search problems by allowing the DM to terminate the search any time; adding a terminal reward, R, for finding the target; and a terminal penalty, B, for not finding the target. Most participants showed a bias toward maximizing detection probability versus minimizing search cost per observation, similar to the previous study.

Sequential Decision Making with Two Decision Criteria

A random sample $X_1, X_2, \ldots$ is generated by an unknown value θ of state of the world, or unknown state of nature, Θ. The DM can take observations one at a time. After observing $X_n = x_n$, the DM makes inferences about Θ based on the values of $X_1, \ldots, X_n$ and can decide whether to continue sampling observations or to stop the process. In the former case, observation X_{n+1} is taken; in the latter case, the DM makes a final decision $d \in D$. The consequences to the DM depend on the decision d and the value θ.

The statistical theory for this situation has been developed by Wald during the 1940s. It has been used to test hypotheses and estimate parameters. In the psychology research, sequential decision making of this kind is usually limited to two decisions $D = \{d_1, d_2\}$ and applied to binary choice tasks (*see* Diffusion and Random Walk Processes; Bayesian Theory, History, Applications, and Contemporary Directions).

The standard theory of sequential analysis by Wald (1947) does not consider observation costs $C(n)$, losses for terminal decisions $L(\theta, d)$, and a priori (subjective) probabilities π about the state of nature as variables. Deferred decision theory generalizes the original theory by including these variables explicitly. The objective of the DM is to find a stopping rule that minimizes expected loss (called risk) and expected observation cost. The form of that optimal stopping rule depends mainly on the assumption about the number of observations that can be taken (bounded or unbounded) and on the assumption of cost per observation (fixed or not) (see DeGroot, 1970). Birdsall and Roberts (1965), Edwards (1965), and Rapoport and Burkheimer (1971) introduced the idea of deferred decision theory as normative models of choice behavior to the psychological community. Experiments investigating human behavior in deferred decision tasks were done by Pitz and colleagues in a series of papers (e.g., 1969) and by Busemeyer and Rapoport (1988). Rapoport and Wallsten (1972) summarize experimental findings.

For illustration, assume the decision problem in its simplest form. Suppose two possible states of nature θ_1 and θ_2 and two

possible decisions d_1 and d_2. Cost c per observation is fixed and the number of observations is unbounded. The DM does not know which of the states of nature, θ_1 or θ_2, is generating the observation but there are a priori probabilities π that it is θ_1 and $(1-\pi)$ that it is θ_2. Let w_i denote the loss for a terminal decision incurred by the DM in deciding that θ_i is not the correct state of nature when it actually is $(i = 1, 2)$. No losses are assumed when the DM makes a correct decision. Let π_n denote the posterior probability that θ_1 is the correct state of nature generating the observations after n observations have been made. The total posterior expected loss is $r_n = \min\{w_1\pi_n, w_2(1-\pi_n)\} + nc$. The DM's objective is to minimize the expected loss. An optimal stopping rule is specified in terms of decision boundaries, α and β. If the posterior probability is greater than or equal to α, then decision d_1 is made; if the posterior probability is smaller than or equal to β, then d_2 is selected; otherwise sampling continues.

See also: Bayesian Theory, History, Applications, and Contemporary Directions; Diffusion and Random Walk Processes; Dynamic Decision Making.

Bibliography

Bearden, J.N., Rapoport, A., 2005. Operations research in experimental psychology. In: Smith, J.C. (Ed.), Tutorials in Operations Research: Emerging Theory, Methods, and Applications, pp. 213–236.

Bearden, J.N., Murphy, R.O., Rapoport, A., 2005. A multi-attribute extension of the secretary problem: theory and experiments. Journal of Mathematical Psychology 49, 410–422.

Bearden, J.N., Rapoport, A., Murphy, R.O., 2006. Sequential observation and selection with rank-dependent payoffs: an experimental study. Management Science 52 (9), 1437–1449.

Birdsall, T.G., Roberts, R.A., 1965. Theory of signal detectability: Deferred decision theory. The Journal of Acoustical Society of America 37, 1064–1074.

Brickman, Ph, 1972. Optional stopping on ascending and descending series. Organizational Behavior and Human Performance 7, 53–62.

Busemeyer, J.R., Rapoport, A., 1988. Psychological models of deferred decision making. Journal of Mathematical Psychology 32 (2), 91–133.

Corbin, R.M., Olson, C.L., Abbondanza, M., 1975. Context effects in optional stopping decisions. Organizational Behavior and Human Performance 14, 207–216.

De Groot, M.H, 1970. Optimal Statistical Decisions. McGraw-Hill, New York.

Edwards, W., 1965. Optimal strategies for seeking information: Models for statistics, choice response times, and human information processing. Journal of Mathematical Psychology 2, 312–329.

Eriksson, K., Strimling, P., 2010. The devil is in the details: Incorrect intuitions in optimal search. Journal of Economic Behavior and Organization 75, 338–347.

Ferguson, T.S., 1989. Who solved the secretary problem? Statistical Science 4 (3), 282–296.

Freeman, P.R., 1983. The secretary problem and its extensions: a review. International Statistical Review 51, 189–206.

Gilbert, J.P., Mosteller, F., 1966. Recognizing the maximum of a sequence. American Statistical Association Journal 35–73.

Kanarick, A.F., Huntington, J.M., Peterson, R.C., 1969. Multisource information acquisition with optional stopping. Human Factors 11, 379–385.

Khan, J.P., Rapoport, A., Jones, L.E., 1967. Decision making in a sequential search task. Perception & Psychophysics 2 (8), 374–376.

Pitz, G.F., Reinhold, H., Geller, E.S., 1969. Strategies of information seeking in deferred decision making. Organizational Behavior and Human Performance 4, 1–19.

Rapoport, A., 1969. Effects of observation cost on sequential search behavior. Perception & Psychophysics 6 (4), 234–240.

Rapoport, A., Burkheimer, G.J., 1971. Models of deferred decision making. Journal of Mathematical Psychology 8, 508–538.

Rapoport, A., Tversky, A., 1966. Cost and accessibility of offers as determinants of optional stopping. Psychonomic Science 4, 145–146.

Rapoport, A., Tversky, A., 1970. Choice behavior in an optional stopping tasks. Organizational Behavior and Human Performance, 105–120.

Rapoport, A., Wallsten, T.S., 1972. Individual decision behavior. Annual Review of Psychology, 131–176.

Sakaguchi, M., 1961. Dynamic programming of some sequential sampling design. Journal of Mathematical Analysis and Applications 2, 446–466.

Seale, D.A., Rapoport, A., 1997. Sequential decision making with relative ranks: an experimental investigation of the "secretary problem." Organizational Behavior and Human Decision Processes 69 (3), 221–236.

Shapira, Z., Venezia, I., 1981. Optional stopping on nonstationary series. Organizational Behavior and Human Performance 27, 32–49.

Wald, A., 1947. Sequential Analysis. Wiley, New York.

Decision Making: Nonrational Theories

Gerd Gigerenzer, Max Planck Institute for Human Development, Berlin, Germany
Wolfgang Gaissmaier, University of Konstanz, Konstanz, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G. Gigerenzer, volume 5, pp. 3304–3309, © 2001, Elsevier Ltd.

Abstract

In a world of known risks, rational theories provide the norms for successful behavior. In a world where not all risks are known and where optimization is not feasible, 'nonrational' tools such as heuristics are needed. In comparison to optimization models, heuristics are robust and can lead to more accurate predictions, while saving time and effort. The study of heuristics addresses the descriptive question of what heuristics an individual or institution has in their 'adaptive toolbox,' as well as the normative question of their ecological rationality, i.e., which heuristics in which situations are most accurate and effective.

Nonrational theories are not theories of irrationality. Rather, they dispense with unrealistic assumptions such as omniscience and optimization in the so-called rational theories. Most important, nonrational theories apply to 'decision-making under uncertainty,' where not all alternatives, consequences, and probabilities are known or knowable (e.g., whom to marry? what job to take? where to invest?). Rational theories, in contrast, are tailored to situations where all risks are known, as in lotteries and roulette ('decision-making under risk'), and where the optimal choice can be calculated. The majority of important problems lie between the two poles of risk and uncertainty, which indicate the necessity of considering both approaches.

Nonrational theories have been denoted by various terms, including *models of bounded rationality*, *procedural rationality*, *fast-and-frugal heuristics*, and *satisficing*. Although there is as yet no consensus on the definition of *nonrational*, these theories typically differ from rational theories on several dimensions, discussed below. The term *decision-making* is used broadly here to include both conscious and unconscious preference, inference, classification, and judgment.

As mentioned, the label *nonrational* should not be confused with irrational behavior: It signifies a type of process, not a type of outcome. In other words, the fact that nonrational theories postulate agents with emotions, limited knowledge, and little time – rather than omniscient 'rational' beings – need not imply that such agents fare badly in the real, uncertain world. We will focus on the theory of heuristics, which is probably the best developed among nonrational theories that do not rely on optimization (Gigerenzer et al., 2011).

Historical Background

A few historical remarks on rational theories help to set the stage. In the mid-seventeenth century, the calculus of probability replaced the ideals of certain knowledge and demonstrative proof (as in mathematics and religion) with a more modest vision of reasonableness. What may be called the first rational theory of decision-making, the maximization of expected value, emerged at this time. According to the theory, an option's expected value is the sum of the product of the probability and the value of each of its consequences; a rational decision-maker chooses the option with the highest expected value.

The notion of defining an option's reasonableness in terms of its expected value soon ran into problems because its prescriptions conflicted in some situations (e.g., the St Petersburg problem) with educated intuition. Daniel Bernoulli therefore proposed replacing the concept of expected value with that of expected utility. For instance, the utility of a monetary gain (say, of $1000) can be defined as a logarithmic function of its dollar value and the agent's current wealth, assuming that the utility of an additional dollar diminishes as the value of the gain and current wealth increase.

The fact that rational decision-making can be defined in more than one way – for example, as maximization of expected value or expected utility – has been interpreted both as the weakness and strength of the program. This ambiguity was one of the reasons why, by 1840, most mathematicians had given up attempting to define a calculus of reasonableness (Daston, 1988). With a few exceptions, rational theories of decision-making largely disappeared until their revival in the 1950s and 1960s. Only then did the major types of rational theories, the maximization of subjective expected utility and Bayesianism, become influential in the social and behavioral sciences. At about the same time, some psychologists and economists – most notably Nobel laureates Herbert Simon and Reinhard Selten – criticized the assumptions about the human decision-maker in modern rational theories as empirically unfounded and psychologically unrealistic, calling for alternative theories.

Optimizing versus Nonoptimizing Theories

Rational theories rest on the ideal of optimization. Optimization entails the calculation of the maximum (or minimum) of some variable across a number of alternatives or values. Nonrational theories, by contrast, apply in the situations where optimization is not feasible. In many real-world situations, no optimizing strategy is known. Even in a game such as chess, whose few rules are stable and well-defined, the optimal strategy (which in fact exists) cannot be computed by a human

or a machine. And even when an optimizing strategy can be calculated, it may demand unrealistic amounts of knowledge about alternatives and consequences, particularly when the problem is novel and time is scarce. Acquiring the requisite knowledge can conflict with goals such as making a decision quickly; in situations of immediate danger, attempting to optimize can actually be deadly. In social and political situations, making a decision at all can take priority over searching for the best option. Most important, for decisions made under uncertainty (i.e., most of our decisions), strategies that do not attempt to optimize (heuristics) can outperform strategies that do. In other words, the concept of an optimizing strategy needs to be distinguished from the concept of an optimal outcome. In an uncertain world, there is no guarantee that 'optimization' will result in a good or optimal outcome. An example is the problem of allocating money to N funds. While Markowitz's Nobel prize-winning mean–variance portfolio is optimal under the assumptions made, in the real world of investment a simple heuristic, 1/N ("allocate your money equally to N assets"), can lead to better performance (DeMiguel et al., 2009).

Normative versus Descriptive Theories

Nonrational theories are descriptive, whereas rational theories are normative – this common distinction is only partly true. Indeed, theories of heuristics are concerned with psychological realism, that is, the capacities and limitations of actual humans, whereas rational theories have little concern for descriptive validity and tend to assume omniscience. But theories of heuristics are normative as well. For instance, in situations where an optimization strategy is nonexistent, unknown, or dangerous because it would impede decision-making, a simple heuristic – such as imitating the behavior of others – can be the better strategy.

Rational theories typically do not assume that agents actually perform optimization or have the knowledge needed to do so. Their purpose is not to describe the reasoning process, but to answer a normative question: What would be the best strategy for an omniscient being to adopt? In economics, psychology, and animal behavior, however, the answer to this question is sometimes used to predict actual behavior. In this way, a rational theory can be descriptive of behavioral outcomes, yet be mute about underlying processes. For instance, optimal foraging theory assumes that animals select and shift between food patches *as if* they had perfect knowledge about the distribution of food, competitors, and other relevant information, without claiming that real animals have this knowledge or perform optimizing computations. Instead, it is assumed that animal behavior has evolved to be close to optimal in specific environments. The question of what proximal mechanisms produce this behavior is a different one. These mechanisms may be simple heuristics such as 1/N and social imitation; that is, the topic of theories of nonrational decision-making.

Ecological Rationality versus Internal Consistency

A classical criterion for rational choice is internal consistency or coherence. Numerous rules of consistency have been formulated: for instance, transitivity and additivity of probabilities. Beginning with in the work of Jean Piaget and Bärbel Inhelder, these rules, which are the building blocks of rational theories, have been used to investigate the development of human thinking. Theories of heuristics, in contrast, place less weight on internal consistency; some heuristics, for instance, can violate transitivity. Instead, their emphasis is on performance in the external world, both physical and social. Measures of external performance include the accuracy, speed, frugality, transparency, and accountability of decision-making. Note that internal consistency does not guarantee that any of these external criteria are met. For instance, although the statement 'there is a 0.01 probability that cigarette smoking causes lung cancer and a 0.99 probability that it does not' is internally consistent in that the probabilities add up to 1, according to relevant research, it is not accurate.

The study of the ecological rationality of heuristics, or strategies in general, is a framework to study performance in the external world: A heuristic is ecologically rational to the degree that it is adapted to the structure of the environment. Heuristics are 'domain-specific' rather than 'domain-general'; that is, they work in a class of environments in which they are ecologically rational. Heuristics provide not a universal rational calculus but a set of domain-specific mechanisms similar to the parts of a Swiss army knife, and have been referred to collectively as the 'adaptive toolbox' (Gigerenzer and Selten, 2001). Thus, the important research question is to specify, for a given heuristic, the structures of environments in which it is faster or more accurate than other strategies in achieving a goal.

For instance, the 1/N heuristic is likely to lead to better returns than the mean–variance portfolio when (1) uncertainty is high (as in the real stock market), (2) sample size is small (typically not more than 10 years of data used by banks), and (3) N is large (because estimation error increases with N). The study of ecological rationality fleshes out Simon's scissors analogy: "Human rational behavior (and the rational behavior of all physical symbol systems) is shaped by a scissors whose two blades are the structure of task environments and the computational capabilities of the actor" (Simon, 1990: p. 7). By looking at only one blade, cognition, it is impossible to understand why and when a behavior succeeds or fails.

Structures of Environments

While some optimization theories treat decision-making as if there were only one tool – maximization of expected utility – the study of decision-making under uncertainty shows that people rely on several tools, not just one. Simple heuristics can succeed by exploiting the structure of information in an environment. In other words, the environment itself can do part of the work for the heuristic. Environmental structures that have been identified to be important in determining the success of heuristics in comparison to other strategies include:

1. Uncertainty: how well a criterion can be predicted.
2. Sample size: number of observations (relative to number of cues).
3. Number of alternatives (N).
4. Redundancy: the correlation between cues.
5. Variability in weights: the distribution of the cue weights (e.g., skewed or uniform).

For instance, heuristics that rely on only one reason, such as *take-the-best* (see below), tend to make more accurate predictions than do strategies such as linear regression in environments with (1) moderate to high uncertainty and (2) moderate to high redundancy. The study of ecological rationality results in comparative statements of the kind "strategy X is more accurate (frugal, fast) than Y in environment E," or in quantitative relations between the performances of strategy X when the structure of an environment changes. Specific findings are introduced in Section Classes of Heuristics.

The concept of ecological rationality should not be confused with the biological concept of adaptation: A match between a heuristic and an environmental structure does not imply that the heuristic evolved because of that environment. Nor does ecological rationality mean that the mental representation mirrors the world: A heuristic is functional, not a veridical copy of the world.

Robustness

A second reason why a simple heuristic can make accurate predictions is robustness. To understand this point, it is necessary to distinguish between data fitting (i.e., determining the best-fitting parameter values for a model given a specific body of data) and prediction (i.e., using these parameter values to predict new data). For data fitting, it generally holds that the more parameters a model has, the better the fit, whereas for prediction there can be a point where less is more. For instance, when recording the air temperature on all 365 days of a year, one can fit the resulting jagged curve increasingly well by adding more free parameters to the model. However, for predicting the daily temperature in the year thereafter, the model that best fits the past data may be less accurate in predicting than a simpler model with fewer parameters and a worse fit. More generally, in noisy environments only part of the available information generalizes to the future. The art of building a good model is to find this part and to ignore the rest. The more noise in the environment, the more models with many free parameters tend to overfit, that is, reflect the noise in a specific sample. Overfitting can become a problem when overly powerful mathematical models, such as neural networks with numerous hidden units and multiple regression with many predictors, are used to fit and then predict behavioral data. Simplicity can reduce overfitting and thereby produce robust decision strategies. A general formulation for when and why one should simplify is the bias–variance dilemma (Geman et al., 1992).

In much research on reasoning, a *bias* typically refers to ignoring part of the information, as in the base-rate fallacy. This can be captured by the equation:

$$\text{Error} = \text{bias} + \varepsilon \qquad [1]$$

where ε is an irreducible random error. In this view, if the bias is eliminated, good inferences are obtained. In statistical theory, however, there are three sources of errors:

$$\text{Error} = \text{bias}^2 + \text{variance} + \varepsilon \qquad [2]$$

where *bias* refers to a systematic deviation between a model and the true state, as in eqn [1]. To define the meaning of *variance*, consider 100 people who rely on the same strategy but have different samples of observations from the same population. Because of sampling error, the 100 predictions may not be the same. Across samples, variance is the expected squared deviation around their mean, while bias is the difference between the mean prediction and the true state of nature.

Consider the problem many companies face in predicting which of their thousands of customers will continue to be active, that is, make purchases in the future. To do so, the Pareto/negative binomial distribution (NBD) model relies on complicated math to estimate its four free parameters from previous data. In contrast, the hiatus heuristic used by experienced managers classifies customers as active only if they have made a purchase within the last 9 months (the hiatus). It was shown to predict customer purchases more accurately than the Pareto/NBD model, despite using less information (Wübben and von Wangenheim, 2008). The bias–variance dilemma provides an explanation for this less-is-more effect. Because the hiatus heuristic does not need to estimate any parameters, it has zero variance, although it probably has a strong bias. In contrast, the Pareto/NBD model likely has a smaller bias but, nevertheless, a larger overall error because it additionally suffers from variance.

Variance decreases with increasing sample size, but also with simpler strategies that have fewer free parameters (and less flexible functional forms; Pitt et al., 2002). Thus, a cognitive system needs to draw a balance between being biased and flexible (variance), rather than simply trying to eliminate bias. Heuristics can be fast, frugal, and accurate by exploiting the structure of information in environments, by being robust, and by striking a good balance between bias and variance. This bias–variance dilemma helps to understand the rationality of simple heuristics and why less can be more (Brighton and Gigerenzer, 2008).

Classes of Heuristics

A heuristic is a strategy that ignores part of the information, with the goal of making decisions more accurately, quickly, and frugally (i.e., with fewer pieces of information) compared to more complex methods. Heuristics are defined by common *building blocks* from which the various heuristics are constructed as an organizing principle. This allows the larger number of heuristics to be reduced to a smaller number of components, similar to how the chemical elements in the periodic table are built from a small number of particles. Three building blocks have been proposed:

1. *Search rules* specify in what direction search extends in the search space.
2. *Stopping rules* specify when search is stopped.
3. *Decision rules* specify how the final decision is reached.

Different classes of heuristics have been specified, all of which can be described in terms of these building blocks. Fast-and-frugal heuristics model decision-making as a dynamic process in which cues or reasons are sequentially searched for – in memory or in the outside world – and inferences are determined by simple stopping and decision rules. The challenge is to understand what the class of heuristics is, how

a heuristic is selected, and in which environments it is successful. The following summary is based on Gigerenzer and Gaissmaier (2011), where further references can be found.

Recognition-Based Decision-Making

The recognition memory literature indicates that a sense of recognition (often called *familiarity*) appears more quickly than recollection in consciousness. This core capacity is exploited by the first class of heuristics. The goal is to make inferences about a criterion that is not directly accessible to the decision-maker, based on recognition retrieved from memory. Prominent heuristics in this class are the recognition heuristic and the fluency heuristic.

Recognition Heuristic

If one of two alternatives is recognized and the other is not, infer that the recognized alternative has the higher value with respect to the criterion (Goldstein and Gigerenzer, 2002). This heuristic is ecologically rational in an environment (reference class) R to the degree the recognition of alternatives a, b ∈ R positively correlates with their criterion value. If the correlation between recognition and the criterion is sufficiently large, a counterintuitive result is observed: Knowing fewer alternatives can lead to more accurate predictions than knowing more alternatives because people who recognize both alternatives cannot use the heuristic.

The recognition heuristic is a model that relies on recognition only and assumes that people will ignore strong, contradicting cues (the so-called noncompensatory inferences). Much research in the past decade was devoted to investigating to what degree people actually rely on the recognition heuristic in such a noncompensatory fashion, and what the boundary conditions are. Additionally, research has identified several environments in which, from a prescriptive perspective, the recognition heuristic is ecologically rational and can compete with well-established forecasting instruments. These environments include the prediction of sports results such as Wimbledon or soccer, election outcomes, university rankings, and geographical properties such as the size of cities or mountains (for a recent overview of the discussion, see Marewski et al., 2011).

Fluency Heuristic

If both alternatives are recognized but one is recognized faster, then infer that this alternative has the higher value with respect to the criterion (Schooler and Hertwig, 2005).

The fluency heuristic is ecologically rational if the speed of recognition is correlated with the criterion, that is, the fluency validity >0.5. Fluency has been shown to predict the performance of stocks, sales figures, and wealth. The validity of fluency is typically lower than that of recognition, but above chance. Fluency also plays a role when alternatives are not given (as in two-alternative choice) but need to be generated from memory. Johnson and Raab (2003) showed, for instance, that experienced handball players can successfully rely on the first alternative that comes to mind when deciding which move to make in a game (a strategy they called *take-the-first*).

One-Reason Decision-Making

While the recognition and fluency heuristics base decisions on recognition information, other heuristics rely on recall. One class looks for only one 'clever' cue and bases its decision on that cue alone, as in the hiatus heuristic (see above). A second class involves sequential search through a number of cues but also bases its decision on just one. Examples include lexicographic rules and elimination-by-aspect (Tversky, 1972). These heuristics were originally developed for preferences; here, the focus is on models of inferences.

One-Clever-Cue Heuristics

What this class of heuristics has in common is that a decision is based on one specific cue only. Many animal species appear to rely on one cue for locating food, nest sites, or mates. For instance, to catch a fly ball high in the air, baseball players rely on the gaze heuristic *"fixate your eye on the ball, start running, and adjust the running speed so that the angle of gaze remains constant."* Variants of this heuristic that rely only on the optical angle have been observed in bats, birds, and fish when pursuing a prey or a mate. Similarly, to choose a mate, peahens have been reported to investigate only three or four of the peacocks displaying in a lek and choose the one with the largest number of eyespots (Petrie and Halliday, 1994). The ecological rationality of one-clever-cue heuristics is not entirely clear at this point in time, but candidates are environments where the variability of cue weights and redundancy is moderate to high and sample size is small (Hogarth and Karelaia, 2007).

Take-the-Best Heuristic

This heuristic is a model of how to infer which of two alternatives has a higher value on a criterion, based on binary cue values retrieved from memory: *Search through cues by their validities, and stop search when the first cue is found that allows for a decision.* A striking discovery was that take-the-best's predictions can be more accurate than those of linear multiple regression models (Czerlinski et al., 1999) and complex nonlinear strategies such as an exemplar-based model (nearest-neighbor classifier), Quinlan's decision-tree induction algorithm C4.5, and classification and regression trees (CART) (Brighton and Gigerenzer, 2008). Research on the ecological rationality of take-the-best suggests two structures of environments it can exploit: moderate to high cue redundancy and moderate to high variability in cue weights (Hogarth and Karelaia, 2007). And it is precisely these features that determine when people's inferences can best be predicted by take-the-best and when they cannot, indicating adaptive strategy use (Bröder, 2012). Variants of take-the-best have been successfully applied to model consumer choice and improve literature search.

Fast-and-Frugal Trees

One way to model classification is in terms of trees. For instance, for m binary cues or attributes, Bayes' rule can be represented as a tree with 2^m leaves. In contrast, a fast-and-frugal tree is defined as a tree with $m + 1$ exits only, with exits at each cue: *Search sequentially through cues, and stop search as soon as a cue leads to an exit* (Martignon et al., 2003). When the number of cues grows, a Bayesian tree becomes computationally intractable or fraught with estimation error because one

typically has too few data points for the thousands of 'leaves' of such a gigantic tree. In contrast, a fast-and-frugal tree needs to estimate fewer parameters and is likely more robust. Fast-and-frugal trees are used by experts in many fields, from emergency medicine to bail decisions.

Trade-off Heuristics

Unlike recognition-based and one-reason decisions, the third class of heuristics weights cues or alternatives equally and thus makes trade-offs (compensatory strategies).

Tallying

Whereas take-the-best ignores cues (but includes a simple form of weighting cues by ordering them), tallying ignores weights and simply counts the number of cues favoring one alternative to others. Dawes (1979) showed that tallying was about as accurate as multiple regression in its predictions and sometimes even better. In a more extensive test across 20 environments, Czerlinski et al. (1999) demonstrated that tallying had, on average, a higher predictive accuracy. The challenge is to figure out *when* this is the case. Einhorn and Hogarth (1975) found that unit-weight models were more successful than multiple regression when the ratio of alternatives to cues was 10 or smaller, the linear predictability of the criterion was small ($R^2 \leq 0.5$), and cues were highly redundant. Successful applications of variants of tallying include predicting strokes with simple bedside rules (which actually outperformed magnetic resonance imaging (MRI) in this regard) and avoiding avalanches.

Mapping Model

How do people arrive at quantitative estimates based on cues? The mapping model assumes that people tally the number of relevant cues with an object's positive values (von Helversen and Rieskamp, 2008). The estimate is the median criterion value of objects with the same number of positive cues. People's judgments were better captured by this model than by a linear regression and an exemplar model when the criterion values followed a skewed distribution, and it was successfully applied to predict sentencing decisions.

1/N Heuristic

Another variant of the equal weighting principle is the 1/N heuristic (also known as the *equality heuristic*), which is a simple rule for the allocation of resources (time, money) to N alternatives. Use of this heuristic has been reported for financial investment (see above), parental investment, and as a tool for fair allocation of money in experimental games.

Aspiration Level Theories

Simon (e.g., 1956, 1982) proposed a heuristic known as *satisficing*, which allows an agent to make a decision without evaluating or even knowing all the alternatives: *Set an aspiration level, search through alternatives sequentially, and stop search as soon as an alternative is found that satisfies the level.* An aspiration level is either a value on a goal variable (e.g., profit or market share) or, in the case of multiple goals, a vector of goal values that is satisfactory to the agent. For instance, agents might set a lower limit on the price at which they would be willing to sell their shares in a company (the aspiration level). In this satisficing model, the agent makes no attempt to calculate an optimal stopping point (here, the best day on which to sell). The aspiration level need not be fixed but can be dynamically adjusted to feedback. For instance, if the investors observe that the share price is monotonically increasing rather than fluctuating over time, they might conclude that there is some stable trend and adjust the limit upward. Thus, aspiration level theories model decision-making as a dynamic process in which alternatives are encountered sequentially and aspiration levels stop search. The challenge is to understand where aspiration levels come from in the first place (Selten, 1998; Simon, 1982).

Summary

In a world of known risks, rational theories provide the norms for successful behavior. In a world where not all risks are known, statistics and logic are not sufficient – additional tools, such as heuristics, are needed. Although heuristics are often classified as nonrational, there is nothing irrational about them. The study of the adaptive toolbox analyzes the heuristics that an individual, a profession, or society relies on to deal with an uncertain world. While this analysis is descriptive, the study of the ecological rationality addresses the normative questions of what heuristic to rely on in what situation. The emerging science of heuristics shows when and why complex problems do not need complex solutions, and why less information can be more.

See also: Apprenticeship and School Learning: Lessons from Germany; Control Behavior: Psychological Perspectives; Health Self-Regulation, Motivational and Volitional Aspects of; Personal Projects; School Achievement: Motivational Determinants and Processes; Self-Regulation in Adulthood; Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model; Vocational Interests, Values, and Preferences, Psychology of.

Bibliography

Brighton, H., Gigerenzer, G., 2008. Bayesian brains and cognitive mechanisms: harmony or dissonance? In: Chater, N., Oaksford, M. (Eds.), The Probabilistic Mind: Prospects for Bayesian Cognitive Science. Oxford University Press, New York, pp. 189–208.

Bröder, A., 2012. The quest for take-the-best. Insights and outlooks from experimental research. In: Todd, P.M., Gigerenzer, G., the ABC Research Group, Ecological Rationality. Intelligence in the World. Oxford University Press, New York, pp. 216–240.

Czerlinski, J., Gigerenzer, G., Goldstein, D.G., 1999. How good are simple heuristics? In: Gigerenzer, G., Todd, P.M., the ABC Research Group, Simple Heuristics that Make Us Smart. Oxford University Press, New York, pp. 97–118.

Daston, L., 1988. Classical Probability in the Enlightenment. Princeton University Press, Princeton, NJ.

Dawes, R.M., 1979. The robust beauty of improper linear models in decision making. American Psychologist 34, 571–582.

DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal versus naive diversification: how inefficient is the 1/N portfolio strategy? Review of Financial Studies 22, 1915–1953.

Einhorn, H.J., Hogarth, R.M., 1975. Unit weighting schemes for decision making. Organizational Behavior and Human Decision Processes 13, 171–192.

Geman, S., Bienenstock, E., Doursat, E., 1992. Neural networks and the bias/variance dilemma. Neural Computation 4, 1–58.

Gigerenzer, G., Gaissmaier, W., 2011. Heuristic decision making. Annual Review of Psychology 62, 451–482.

Gigerenzer, G., Hertwig, R., Pachur, T. (Eds.), 2011. Heuristics: The Foundations of Adaptive Behavior. Oxford University Press, New York.

Gigerenzer, G., Selten, R. (Eds.), 2001. Bounded Rationality: The Adaptive Toolbox. MIT Press, Cambridge, MA.

Goldstein, D.G., Gigerenzer, G., 2002. Models of ecological rationality: the recognition heuristic. Psychological Review 109, 75–90.

Hogarth, R.M., Karelaia, N., 2007. Heuristic and linear models of judgment: matching rules and environments. Psychological Review 114, 733–758.

Johnson, J.G., Raab, M., 2003. Take the first: option generation and resulting choices. Organizational Behavior and Human Decision Processes 91, 215–229.

Marewski, J.N., Pohl, R.F., Vitouch, O., 2011. Recognition-based judgments and decisions: what we have learned (so far). Judgment and Decision Making 6, 359–380.

Martignon, L., Vitouch, O., Takezawa, M., Forster, M., 2003. Naive and yet enlightened: from natural frequencies to fast and frugal trees. In: Hardman, D., Macchi, L. (Eds.), Thinking: Psychological Perspectives on Reasoning, Judgment, and Decision Making. Wiley, Chichester, UK, pp. 189–211.

Petrie, M., Halliday, T., 1994. Experimental and natural changes in the peacock's (*Pavo cristatus*) train can affect mating success. Behavioral Ecology and Sociobiology 35, 213–217.

Pitt, M.A., Myung, I.J., Zhang, S., 2002. Toward a method for selecting among computational models for cognition. Psychological Review 109, 472–491.

Schooler, L.J., Hertwig, R., 2005. How forgetting aids heuristic inference. Psychological Review 112, 610–628.

Selten, R., 1998. Aspiration adaptation theory. Journal of Mathematical Psychology 42, 191–214.

Simon, H.A., 1956. Rational choice and the structure of environments. Psychological Review 63, 129–138.

Simon, H.A., 1982. Models of Bounded Rationality. MIT Press, Cambridge, MA.

Simon, H.A., 1990. Invariants of human behavior. Annual Review of Psychology 42, 1–19.

Tversky, A., 1972. Elimination by aspects: a theory of choice. Psychological Review 79, 281–299.

von Helversen, B., Rieskamp, J., 2008. The mapping model: a cognitive theory of quantitative estimation. Journal of Experimental Psychology: General 137, 73–96.

Wübben, M., von Wangenheim, F., 2008. Instant customer base analysis: managerial heuristics often "get it right." Journal of Marketing 72, 82–93.

Decision Making, Psychology of

Joop van der Pligt[†], University of Amsterdam, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The psychology of decision making combines concepts, models, and methods from a variety of disciplines. It describes and explains the discrepancies between normative models of decision making and actual behavior. It also highlights the adaptive nature of human decision making, the use of heuristics, and simple decision rules that can improve our decision making but can also have detrimental effects on the quality of our judgment. The interplay between affect and cognition is also studied in this tradition, and this has improved our insight in both intuitive and more deliberate decision making.

Human decision making has been studied by a variety of disciplines including economics, philosophy, psychology, biology, and statistics. Behavioral decision making or judgment and decision making (JDM), as the field is generally known in psychology, is being studied in all branches of psychology. The starting point of much work on human JDM is rational choice theory. The most influential explanatory concept in the analysis of decisions under uncertainty is the subjective expected utility (SEU) of available alternatives. SEU theory is a normative theory and is often compared to actual decisions. Deviations that violate the rational principles underlying SEU theory systematically are termed anomalies. Anomalies of decision making will be briefly discussed in this section, followed by descriptive theories of decision making including work on heuristics. Next we turn to a comparison of intuitive versus deliberate decision modes and also pay attention to the role of affect and emotions in decision making.

Normative Theories of Decision Making

Traditionally, choices are assumed to be consequential, and options are evaluated using prior assessments of beliefs and values. Generally, research in this field of research focuses on a *probability* component or a *value* component, or both.

Edwards (1954) provided the first major review of research on human JDM. He argued that normative and prescriptive models based on economic and statistical theory should also be relevant to psychologists interested in human JDM. Edwards introduced SEU theory, which decomposes decisions or choices in probabilities and preferences, and provides a set of rules for combining beliefs (probabilities) and preferences ('values' or 'utilities'). (Decomposes decisions in beliefs (probabilities) and preferences (values or utilities) that are combined to assess the overall utility of an outcome.) Utility maximization is considered an appropriate normative rule for decision making under risk or uncertainty. The theory is normative or prescriptive because it specifies how decisions *should* be made. If one accepts the axioms upon which it is based, then the most rational choice or decision is the one specified by the theory as having the highest expected utility.

The axioms of SEU theory are all based on the work of Von Neumann and Morgenstern. Their work showed that if a person's choices follow certain rules or axioms, it is possible to derive utilities in such a way that a preference for a specific alternative implies that its expected utility is greater than that of the other alternative(s). Their approach assumes that it is possible to associate a real number with each possible consequence of the relevant alternatives. This number (e.g., a number between 1 and 10) represents the utility of that consequence for the individual. The *expected utility* of a specific alternative is then the sum of the numbers associated with each possible consequence weighted by the probability that each consequence will occur. Thus a positive or negative outcome will be weighted by the probability of its occurrence; for example, a certain outcome will be multiplied by 1.0, an outcome with a probability of 0.5 will be multiplied by 0.5, etc.

To illustrate this let us compare the two jobs in terms of three possible outcomes or consequences. These are initial salary, career opportunities, and atmosphere. Initial salary is higher at firm A than at firm B, and receives ratings of 8 and 7, respectively, on a scale ranging from 1 to 10. Career opportunities seem better at firm B, but this element cannot be assessed with certainty. Let us assume that career opportunities get ratings of 8 (firm A) and 9 (firm B), but that you are less certain of your assessment of firm A (subjective probability of 0.7) than of firm B (subjective probability of 0.9). Finally, the atmosphere at firm B seemed to be better than at firm A; your scores are 8 and 7, respectively. You spent quite some time at firm B and are certain about your judgment of the atmosphere (it seemed a great team of colleagues). You are less certain about your verdict of the atmosphere at firm A because you spent less time there (probabilities of 1.0 (B) vs 0.9 (A)). This leads to the following scores: Firm A ($1.0 \times 8 + 0.7 \times 8 + 0.8 \times 7 = 20.0$), Firm B ($1.0 \times 7 + 0.9 \times 9 + 1.0 \times 8 = 23.2$). Thus, according to this analysis, the job offer of firm B is more attractive than that of firm A.

The central idea of the Von Neumann and Morgenstern theory is that it provides an important tool to reach a rational decision. SEU theory and related theories have dominated research on decision making since the mid-1940s, and especially the validity of theories such as SEU theory as adequate *descriptions* of human choice and decision making has been a dominant theme in this research area. This is partly based on the idea that the study of human decision making should focus on the perceptual and cognitive factors that cause human JDM

[†] Deceased.

to *deviate* from the predictions of normative models such as SEU theory.

Violations and Anomalies

Research in the past decades has resulted in many illustrations of systematic violations of principles of rationality. A clear example concerns the violation of the *sure thing principle*: If one prefers one option to another no matter what happens – or has happened – one should prefer that option without knowing what happens, or has happened. Tversky and Shafir (1992) gave an example of violating this principle. Their respondents were asked to consider a gamble with a 50/50 chance of winning $200 or losing $100. Generally, a minority of about one-third of the people offered that gamble accept it. This preference pattern changes dramatically if people are asked to imagine they had already played the gamble and were now being offered the opportunity to play it a second time. In case they were asked to imagine they had won the first time, nearly 70% wanted to play again. If they had been asked to imagine they had lost the first time nearly 60% wished to play it a second time.

Another violation is related to so-called sunk costs. It is Saturday afternoon and you are on the brink of leaving for the railway station in order to get to the football stadium. You bought your ticket for the match some time ago. It is pouring with rain, you do not feel that well and standing in the rain for 2 h is not going to help. In sum, you prefer to spend the afternoon at home. Nevertheless, if you decide not to go, you will have 'wasted' the money you paid for the ticket. The nonrefundable money already paid for the ticket is termed a *sunk cost*. Strictly speaking, there appears to be no rational argument to do things you would prefer not to do in the interest of honoring the sunk cost. However, people often honor sunk costs. Examples include trivial anecdotes such as the one described earlier, but also large-scale enterprises such as the development of the Concorde passenger aircraft (a far too expensive and not very popular aircraft) and the Gulf War (an extremely expensive war with mixed outcomes).

Other violations concern principles such as *transitivity*. If people prefer option A over option B, and also option B over option C, they should also prefer option A over option C. Violations of transitivity may arise from evaluating alternative options in terms of different dimensions. On some dimensions 'A' might be preferred over 'B' and 'C' while on others 'C' might be preferred over 'A.' If the salience of the dimensions varies, preferences could differ and one could violate transitivity. Another way in which transitivity may occur is through regret (see below).

The way in which alternative options are presented can also lead to anomalies. It turns out that people often make different choices as a function of the way alternatives are presented or *framed*, even though the alternatives remain unchanged. One example of framing is to present actions in terms of positive as opposed to negative consequences or vice versa. For example, a 50% chance of a successful outcome of surgery may appear to someone to be a more desirable outcome than a 50% chance of failure. Several studies provide support for this framing effect. (The finding that people often make different choices as a function of how the alternatives are presented and framed.) Framing effects have been extensively investigated in decision problems in which the consequences of options are compared to psychologically salient reference points (e.g., a zero point, the status quo). Kahneman and Tversky (1979) found that when asked to choose between $15 000 for sure or a 50/50 chance of $10 000 or $20 000 most people choose the certain option. If one has received the $20 000 and was asked a second choice between returning $5000 or a 50/50 chance of giving back nothing or $10 000, most people choose the latter option. This pair of choices is contradictory. Similarly, when presented with a choice between a sure loss of $100, or a 50/50 chance of a loss of $200 and a 50/50 chance of no loss at all, most people will prefer the uncertain, risky alternative. This risk-seeking preference is in sharp contrast with preferences when the choice is between a sure gain of $100 or a 50% chance of a gain of $200 and a 50% choice of no gain at all. In this case, most people will prefer the certain, risk-avoiding alternative. According to SEU theory, the two options in each problem are equally (un)attractive. Thus, inducing people to adopt a gain frame when choosing between a gamble and a sure thing tends to result in a risk-avoiding preference (the certain alternative). Inducing people to adopt a loss frame, on the other hand, tends to result in a risk-seeking preference (the risky alternative).

A final anomaly concerns violations of *response invariance*. Choice can also be influenced by the method by which preferences and choices are expressed. An example concerns the effect of choosing versus rejection of alternatives that have both bad and good characteristics. Shafir (1993) argues that a 'choose' orientation leads decision makers to search for good qualities of each option, while a 'reject' orientation makes them focus on bad qualities. Thus, if one alternative has more clear-cut positive *and* negative characteristics, and another alternative is relatively bland, the former could both be chosen and rejected depending on the question. Shafir demonstrated this in the context of awarding custody in which the question was either which parent to *award* custody, or which parent to *deny* sole custody of the child.

Descriptive Approaches to Decision Making

Descriptive theories of human JDM attempt to explain *why* people violate rational choice theory. These descriptive frameworks also view decisions as more reasonable and adaptive than previously thought. Kahneman and Tversky developed *prospect theory* to remedy the descriptive failures of SEU theories of decision making (see Kahneman, 2011). (A theory describing decisions under uncertainty as a function of a value function and a probability function.) Prospect theory attempts to *describe* and explain decisions under uncertainty. Like SEU theories, prospect theory assumes that the value of an option or alternative is calculated as the summed products over specified outcomes. Each product consists of a utility and a weight attached to the objective probability. Both the value function and the probability weighting function are nonlinear. The functions have a number of important features. The most important feature of the probability weighting function is that small probabilities are overweighted and large probabilities are underweighted. The probability weighting function is generally

not well behaved near the endpoints. Extremely low probability outcomes can be exaggerated or ignored entirely. Similarly, small differences between high probability and certainty are sometimes neglected, sometimes accentuated. According to Kahneman and Tversky, this is so because people find it difficult to comprehend and evaluate extreme probabilities.

The value function is defined in terms of gains and losses *relative* to a psychologically neutral reference point. The value function is S-shaped; concave in the region of gains above the reference point, convex in the region of losses (see **Figure 1**). Thus, each unit increase in gain (loss) has decreasing value as gain (loss) increases. In other words, the subjective difference between gaining nothing and gaining $100 is greater than the difference between gaining $100 and gaining $200. Finally, the value function is steeper for losses than for gains. This implies that losing $100 is more unpleasant than gaining $100 is pleasant.

Prospect theory can help to explain the framing effects discussed earlier. The probability function and the value function will confirm the higher attractiveness of the risky option (in case of losses) and the risk-avoiding option (in case of gains).

Other descriptive approaches assume that people rely on a variety of simple decision rules for solving decision problems. Experience will affect the availability of these strategies, and strategy choice will also be affected by the expected advantages (benefits) and disadvantages (costs) of the chosen strategy. For many decisions, an exhaustive analysis such as prescribed by SEU theory simply is not worth the trouble. Thus, for many problems people aim for an acceptable solution and not necessarily the optimal solution due to the costs (time, effort) of 'calculating' the best possible option.

Examples of simplifying decision rules mentioned in the literature are: *the dominance rule* (alternative A should be chosen over B if A is better on at least one attribute and not worse than B on all remaining attributes), *the conjunctive decision rule* (any alternative that does not meet a minimally required value on one or more attributes is dropped from the list of remaining possible alternatives), and *the lexicographic decision rule* (choose alternative which is most attractive on the most important attribute; if two alternatives are equally attractive in terms of the most important attribute move to the next attribute). Finally, the *elimination-by-aspects-rule* is often interpreted as a combination of the lexicographic rule and the conjunctive rule. First, the most important attribute is selected. All alternatives that fail to meet the criterion on this attribute are eliminated. This procedure is repeated for each of the remaining attributes.

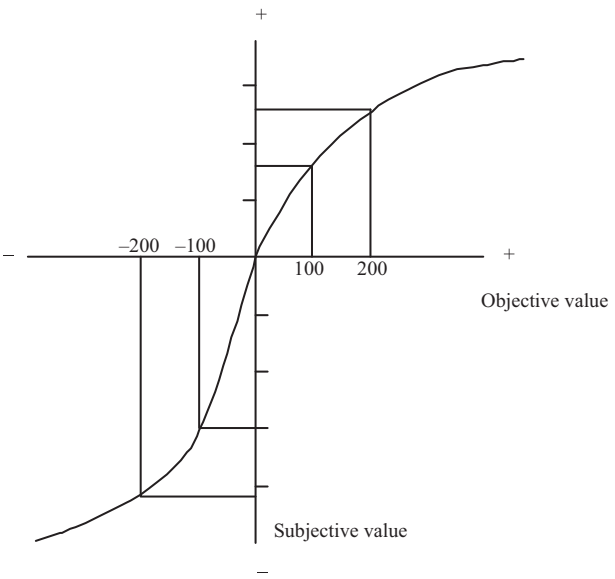

Figure 1 Prospect theory: hypothetical value function.

These decision rules require considerably less cognitive effort than the decision rule required by SEU models (a compensatory decision rule which allows negative scores on one attribute to be compensated by positive scores on another attribute). These simplifying rules can provide adequate shortcuts in complex decision environments. Payne et al. (1992) use the term contingent decision making in which the cognitive costs and benefits of the various strategies people might use determine their choice of strategy. Basic questions in this choice concern the balance between cognitive effort and accuracy, but also factors such as decisional avoidance and accountability influence strategy selection. *Heuristics* can also be seen as simplified decision rules; in the next section, we discuss a variety of heuristics that people use when making decisions.

Heuristics

Heuristics are cognitive shortcuts that can be applied consciously or unconsciously to reduce the complexity of decisions. Initially research emphasized the shortcomings of heuristics. The 'heuristics-and-biases' approach showed that people often rely on simple decision rules but that they would be better off – in terms of accuracy – if they would not. This approach goes back to the seminal work of Kahneman and Tversky. The three heuristics that received most attention were *availability*, *representativeness*, and *anchoring and adjustment*. The *availability* heuristic refers to the tendency to assess the probability of an event based on the ease with which instances of that event come to mind. (The tendency to assess the probability of an event based on the ease with which instances of that event come to mind.) Generally people overestimate the probability of an event if concrete instances of that event are easily accessible in memory. Ease of recall and frequency of occurrence are often correlated. A number of factors that affect memory are, however, unrelated to probability. For example, vivid images are easier to recall than pallid ones. Thus having been involved in a serious car accident is likely to be better remembered than annual statistics about the frequency of (types of) traffic accidents. The former is likely to have more impact on probability estimates than the latter.

The *representativeness* heuristic refers to the tendency to assess the probability that a stimulus belongs to a particular class by judging the degree to which that event corresponds to an appropriate mental model. A well-known example of how ignoring prior probabilities can affect judgment was reported by Kahneman and Tversky in 1973 (see Kahneman, 2011). In their study, respondents were provided with brief personality

sketches, supposedly of engineers and lawyers. They were asked to assess the probability that each sketch described a member of one profession or the other. Half the respondents were told the population from which the sketches were drawn consisted of 30 engineers and 70 lawyers, the remaining respondents were told that there were 70 engineers and 30 lawyers. Findings showed that the prior probabilities were essentially ignored, and that respondents estimated the probability of class membership by judging how similar each personality sketch was to their mental model of an engineer or a lawyer.

Anchoring and adjustment refers to a general judgment process in which an initially given or generated response serves as an anchor, and other information is insufficiently used to adjust that response. The anchoring and adjustment heuristic is based on the assumption that people often start their judgmental process by focusing on some initial value that serves as an anchor. The biases related to this heuristic stem from the use of irrelevant anchors and the tendency to insufficiently adjust up or down from an original starting value or anchor.

Initially these three heuristics were described along with a dozen systematic biases such as nonregressive prediction, neglect of base-rate information, and overconfidence. Both heuristics and biases were seen as causing systematic errors in estimates of known quantities and statistical facts. Kahneman and Frederick (see Kahneman, 2011) revised the early conception of heuristics and biases and proposed a new formulation – highlighting a common process of attribute substitution – to explain how heuristics work. Moreover, their new model also includes an explicit treatment of the conditions under which intuitive judgments are modified or overrated by more elaborate cognitive processes. In this new approach, the word heuristic is used in two senses: The noun refers to the cognitive process, and the adjective in *heuristic attribute* specifies the attribute that is substituted in a particular judgment. Kahneman and Frederick added a family of *prototype heuristics* in which an attribute of a prototype is substituted for an extensional attribute of its category; the original instance of a so-called prototype heuristic is the use of representativeness in category prediction we discussed earlier. (A heuristic in which an attribute of a prototype is substituted for an extensional attribute of its category.)

Gigerenzer and colleagues stressed the *adaptive* role of heuristics. In their view, heuristics are efficient cognitive processes that ignore part of the information: Moreover, how well cognitive heuristics function in decisions under uncertainty is in their view an empirical question. Gigerenzer and Gaissmaier (2011) refer to simple, fast, and frugal heuristics that 'make people smart.' (Various classes of adaptive heuristics that people use to make decisions faster, more frugally, and/or more accurately.) They present an overview of the various heuristics people use in different types of situations, and also attempt to answer the question when people should rely on a given heuristic rather than a more complex decision strategy. Gigerenzer and colleagues stress the conscious and deliberate use of heuristics as strategies that ignore information to make decisions faster, more frugally, and/or more accurately than more complex methods. They also refer to less-is-more effects; i.e., when less information or computation leads to more accurate judgments than more information or computation. They thus disagree with the view of Payne et al. (1992) that people rely on heuristics because information search and computation costs time and effort, and that heuristics save time and effort at the expense of accuracy.

A first class of heuristics introduced by Gigerenzer and colleagues can be related to an important capacity of our memory; i.e., the fact that a sense of recognition appears in consciousness earlier than recollection. The *recognition* heuristic simply states that if one of two alternatives is recognized and the other is not, then one should infer that the recognized alternative has the higher value with respect to the criterion. An example is name recognition of cities; this turns out to be a reasonably valid predictor of their population. The higher the recognition validity (α) for a given criterion, the more ecologically rational it is to rely on the recognition heuristic. For each individual, α can be computed by $\alpha = C/(C + W)$ where C is the number of correct inferences the recognition heuristic would make, computed across all pairs in which one alternative is recognized and the other is not, and W is the number of wrong inferences.

The *fluency* heuristic states that if both alternatives are recognized but one is recognized faster, then infer that this alternative has the higher value with respect to the criterion. The fluency heuristic is ecologically rational if the speed of recognition is correlated with the criterion. Fluency also plays a role when alternatives are not given but need to be generated from memory. Johnson and Raab (2003) introduced the *take-the-first* heuristic: Choose the first alternative that comes to mind. They tested the ecological validity of this approach with experienced handball players when choosing how to proceed in a given situation (e.g., pass the ball to another player, or take a shot). A third class of heuristics are so-called *one-reason* decisions: a class of heuristics that bases judgments on one good reason only, ignoring other cues. Obviously this heuristic is related to the lexicographic decision rule: The latter is defined in terms of the 'most important' attribute. *One-clever-clue* heuristics (an example of one-reason decision making) focuss on the usefulness and predictive value of the selected cue of the overall outcome of the choice. Similarly, the *take-the-best* heuristic is a model of how people infer which of two alternatives has a higher value on a criterion, based on binary cue values retrieved from memory. Take-the-best consists of three steps: (1) search through cues in order of their validity, (2) stop on finding the first cue that discriminates between the alternatives, and (3) the alternative with the more positive cue value has the higher criterion value.

Karelaia (2006) showed that a *confirmatory stopping rule* – stop after two cues are found that point to the same alternative – leads to quite robust results. This heuristic tends to be ecologically valid in situations where the decision maker knows little about the validity of cues. This class of heuristics weighs cues or alternatives equally and thus makes simple trade-offs. Tallying is an example. *Tallying* entails simply counting the number of cues favoring one alternative in comparison to others. In three steps: (1) search through cues, (2) stopping rule: if the number of cues is the same for both alternatives, search for another cue. If no more cues are found, guess. Otherwise (3) decide for the alternative that is favored by more cues.

Basically Gigerenzer's work showed that heuristics are not good or bad and – in some circumstances – can be more

accurate than more complex strategies. More research is needed on how people learn to use heuristics in an adaptive way; i.e., when to use the proper strategy from their adaptive heuristic tool box.

In their review article on heuristic decision making, Gigerenzer and Gaissmaier (2011) also ask whether gut feelings are based on heuristics, and if so, on which? This brings us to the distinction between intuitive versus deliberative decision making.

Intuition versus Deliberation and Reasoning

Kahneman (2003) related the research he and Amos Tversky conducted on heuristics-and-biases to two generic modes of cognitive function: an intuitive mode in which judgments and decisions are made automatically and rapidly (Systems 1), and a controlled mode, which is deliberate and slower (System 2).

An illustration of the differences between System 1 and System 2 is provided by Kahneman in his seminal work *Thinking, fast and slow*, published in 2011. Two characteristics of System 1 are the neglect of duration and the peak-end rule (the global retrospective ratings of pain or pleasure are well predicted by the average level of pain/pleasure reported at the worst moment of the experience and at its end). According to System 2, duration is important. This is not the case in System 1. As a consequence people sometimes favor suboptimal outcomes. An example is that people generally prefer a short period of intense joy over a long period of moderate happiness. Similarly, we fear a short period of intense pain more than a much longer period of moderate pain.

System 2 is characterized by processes that are slow, serial, controlled, effortful, rule-governed, flexible, and neutral. The more intuitive System 1 is characterized by processes that are fast, parallel, automatic, effortless, associative, slow-learning, and emotional. Kahneman argued that highly accessible impressions produced by System 1 control judgments and preferences, unless modified or overruled by more elaborate and deliberate operations in System 2.

Psychologists differ in their definitions of intuition, but there is general agreement that intuition tends to rely on information acquired without conscious and deliberative reasoning. Betsch (2008) puts it as follows:

> "Intuition is a process of thinking. The input to this process is mostly provided by knowledge stored in long-term memory that has been primarily acquired via associative learning. The input is processed automatically and without conscious awareness. The output of the process is a feeling that can serve as a basis for judgments and decisions" (p. 4).

(A more or less automatic process of thinking based on knowledge stored in long-term memory that has primarily been acquired via associative learning.) Deliberation in contrast, is described as a strategy in which "Information is integrated in a serial manner, processing is cognitively demanding and rather slow, and individuals using these strategies are aware of most of the underlying processes and can even verbalize them" (p. 6). There is growing evidence that people can unconsciously integrate large amounts of information, and integrate them to come to summary evaluations or feelings (Betsch and Glöckner, 2010). Intuitions based on these experiences can also be remarkably accurate.

Heuristics (as mentioned in the preceding paragraph) are obviously related to intuition (Gigerenzer and Gaissmaier, 2011; Kahneman, 2003), but some argue that they do not fit the above definition of intuition. For instance, decisions based on heuristics tend to ignore rather than integrate multiple pieces of information and deliberated judgments more often rely on heuristics than judgments made intuitively (see Plessner et al., 2008). Unconscious thought theory (Dijksterhuis and Nordgren, 2006) is also associated with intuition. Unconscious thought theory claims that people can make superior decisions after a period of distraction in which they unconsciously address the judgment task. However, some of its assumptions and claims are controversial.

Affect, Emotion, and Decision Making

The first decades of research on decision making were dominated by cognitive approaches. Individual decision makers were assumed to assess the likelihood and value of the possible consequences of their actions rather dispassionately, as described earlier in this overview. People were assumed to opt for actions that maximize the overall utility of the various possible outcomes; and violations from the principle of utility maximalization were explained in terms of cognitive errors and shortcomings. The various heuristics people employed were seen as cognitive shortcuts that help to cope with the complexity of decisions. As argued by Loewenstein and Lerner (2003), the literature generally ignored the role of emotions in JDM. There are a few exceptions. One of them is *regret theory*. The basic idea underlying this approach is that the value placed on the outcome of a specific choice or gamble also depends on the outcome that *would* have been received if one had chosen an alternative choice or gamble. Thus winning $100 is likely to be affected by the knowledge that one could have won $500 (vs e.g., $10) if one had selected another alternative. Basically, regret theory posits that people *anticipate* the possibility of regret and that this anticipation affects their decision making. People not only want to maximize utility, they also want to minimize the possibility of regret.

Regret is a rather cognitive emotion since it relies on comparing different outcomes, but regret theory was one of the first to explicitly incorporate emotions in explanations of human decision making. It needs to be added that robust and systematic effects of regret have mainly been obtained when the possibility of regretful outcomes are made salient. Another example is the anticipation of *disappointment*. When disappointment is anticipated, it has an effect on judgment and choice; people generally want to minimize the possibility of disappointment. In sum, anticipated emotions can affect decision making but they need to be salient before the decision is made.

Other research focused on the influence of *immediate* affect, as opposed to *anticipated* affect. First, low and moderate levels of affective intensity can serve in an 'advisory' role without altering the perception of the various attributes

underlying a decision (probabilities, values). An example is the *affect-as-information* model in which people rely on subjective assessments of their feelings ('How do I feel about it?') and judge accordingly. Emotions – particularly when they are more intense – can override deliberative cognitive processing. For example 'gut feelings' often determine the acceptability of technological hazards (see e.g., Slovic et al., 2002).

As shown by Loewenstein and others, immediate emotions also have an impact on how people feel about future outcomes. Thus people tend to project their current feeling state on to the future; when people are in a 'cold' state (e.g., not hungry, not tired, not sexually aroused, not angry) they tend to underestimate the impact of these states and their effect on judgment and choice. Loewenstein termed these 'hot/cold empathy gaps.' Emotions not only exert an influence on *how* people judge and weigh expected consequences, but also on the *quality* of information processing. Generally, negative emotions narrow attentional focus, while positive emotions tend to broaden attentional focus. Positive emotions are also generally associated with less thorough information processing.

As argued by Loewenstein and Lerner (2003), emotions are now firmly embedded in the decision-making literature. In the past, emotions were associated with maladaptive strategies and faulty decision making. Now we know that emotions serve an essential function in coordinating cognition and behavior and that ignoring emotions can have detrimental effects on the quality of our decision making. An important benefit of expected emotions is that we employ a broader perspective when assessing the outcomes of a decision. This can be quite functional, if we focus on the right emotions and assess them adequately. Immediate affect that is experienced at the time of making a decision can help to prioritize information processing and make specific consequences of the decision more salient; unfortunately it can also steer decisions in directions that are counter to self-interest. For example, fear may make people more aware of risk, but it may also make them excessively risk averse. Another example is that experimentally induced fear and anger have opposing effects on the perception of risk. Angry people tend to perceive less risk than people who are in a state of fear. All in all, emotions can have a profound impact on decisions.

Conclusions

Nearly seven decades after the seminal work of Edwards (1954), decision-making research is becoming more prominent in psychology textbooks and a clear and separate research area has emerged, generally referred to as *behavioral decision research*, or JDM. An important characteristic of this field of inquiry is that it relies on concepts, models, and methods from a variety of disciplines including economics, statistics, and social and cognitive psychology. Research on decision making pays considerable attention to the discrepancies between normative models and actual behavior. Most of the research attempting to account for these discrepancies focused on the information processing strategies, or heuristics, that people use when making judgments or decisions. A variety of descriptive approaches to human decision making has improved our insight into the cognitive processes underlying decisions and should also help to improve the quality of decision making. Recent advances in this field of research show the important role of anticipated and felt emotions on decision making. The field now generally distinguishes between more analytical and deliberative decision modes, and more intuitive decision modes. Affect and emotion play an important role in the latter. Research on the interplay between cognition and affect is likely to help improve our understanding of human decision making.

See also: Behavioral Economics; Decision Making: Nonrational Theories; Decision and Choice: Bounded Rationality; Decision and Choice: Heuristics; Fear: Psychological and Neural Aspects; Heuristics in Social Cognition; Information and Knowledge, Organizational; Rational Choice and Organization Theory; Risk: Social Psychological Perspectives; Social Cognition; Tversky, Amos (1937–96).

Bibliography

Betsch, T., 2008. The nature of intuition and its neglect in research on judgment and decision making. In: Plessner, H., Betsch, C., Betsch, T. (Eds.), Intuition in Judgment and Decision Making. Erlbaum, Mahwah, NJ, pp. 3–22.

Betsch, T., Glöckner, A., 2010. Intuition in judgment and decision making: extensive thinking without effort. Psychological Inquiry 21, 279–294.

Dijksterhuis, A., Nordgren, L.F., 2006. A theory of unconscious thought. Perspectives on Psychological Science 1, 95–109.

Edwards, W., 1954. The theory of decision making. Psychological Bulletin 51, 380–417.

Gigerenzer, G., Gaissmaier, W., 2011. Heuristic decision making. Annual Review of Psychology 62, 451–482.

Johnson, J.G., Raab, M., 2003. Take the first: option generation and resulting choices. Organizational Behaviour and Human Decision Making Processes 91, 215–219.

Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decisions under risk. Econometrica 47, 263–291.

Kahneman, D., 2003. A perspective on judgment and choice. American Psychologist 58, 697–720.

Kahneman, D., 2011. Thinking, Fast and Slow. Farrar, Straus and Giroux, New York.

Karelaia, N., 2006. Thirst for confirmation in multi-attribute choice: does search for consistency impair decision performance? Organizational Behaviour and Human Decision Making Processes 100, 128–143.

Loewenstein, Lerner, 2003. The role of affect in decision making. In: Davidson, R.J., Goldsmith, H.H. (Eds.), Handbook of Affective Sciences. Oxford University Press, Oxford.

Payne, J.W., Bettman, J.R., Johnson, E.J., 1992. Behavioral decision research: a constructive processing perspective. Annual Review of Psychology 43, 87–132.

Plessner, H., Betsch, C., Betsch, T., 2008. Intuition in Judgment and Decision Making. Lawrence Erlbaum Associates Publishers, Mahwah, NJ.

Shafir, E., 1993. Choosing versus rejecting: why some options are both better and worse than others. Memory and Cognition 21, 546–556.

Slovic, P., Finucane, M., Peters, E., MacGregor, D.G., 2002. The affect heuristic. In: Gilovich, T., Griffin, D., Kahneman, D. (Eds.), Heuristics and Biases. Cambridge University Press, New York, pp. 397–420.

Tversky, A., Shafir, E., 1992. The disjunction effect in choice under uncertainty. Psychological Science 3, 305–309.

Declarative Memory, Neural Basis of

Elizabeth A Buffalo, Washington National Primate Research Center, Seattle, WA, USA
Larry R Squire, Veterans Affairs San Diego Healthcare System, San Diego, CA, USA and University of California, La Jolla, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

An important finding of neuropsychological studies is that memory is not a single entity but is composed of several different forms that rely on separate brain systems. The major distinction is between conscious knowledge of facts and events, referred to as declarative or explicit memory, and nonconscious knowledge systems that provide for the capacity of skill learning, habit formation, the phenomenon of priming, and certain other ways of interacting with the world, collectively referred to as nondeclarative or implicit memory. Declarative memory is dependent on the integrity of the hippocampus and anatomically related structures in the medial temporal lobe and diencephalon. It seems to be especially good at associating the various aspects of a context that are present at a particular time and place, thereby creating a memory of an episode. A particularly good example of declarative memory involves the ability to identify a recently encountered item as familiar, a capacity termed as recognition memory. Declarative memory is the kind of memory impaired in amnesia. Other characteristics of this form of memory have emerged from studies with experimental animals. Declarative memory is fast and is flexible in the sense that it is accessible to multiple response systems. It is especially suited for one-trial learning, and for forming and maintaining associations between arbitrarily different pieces of material.

Introduction

Many important aspects of the neurobiology of memory involve systems-level questions that can be addressed at a relatively global, neuropsychological level of analysis. Two long-standing and interrelated questions have been of central interest since the 1960s. The first is whether there is more than one kind of memory. The second concerns what brain structures and pathways are involved in memory. Scientific work on these issues has introduced the term declarative memory to neuroscience and identified a number of parallels between the neural organization of memory in humans and animals.

It has turned out that the brain is organized such that memory is a distinct and separate cognitive function, which can be studied in isolation from perception and other intellectual abilities. Significant information about how memory is organized has come from two experimental approaches. First, the study of patients with memory disorders (amnesia) has been enormously useful. Second, information about the anatomy of memory has come mainly from the kind of systematic and cumulative work that is possible only in experimental animals.

Amnesia and the Medial Temporal Lobe Memory System

Amnesia refers to difficulty in learning new information or in remembering the past. Amnesia can occur as a strikingly circumscribed memory impairment in the absence of other cognitive deficits (Scoville and Milner, 1957). Patients with this disorder can have intact intelligence test scores, intact language and social skills, and an intact memory for a great deal of information from the remote past, especially childhood. In addition, immediate memory is intact. Indeed, patients can hold information in mind for several minutes, provided there is no distraction and the material can be rehearsed. Their intact immediate memory explains why amnesic patients can carry on a conversation and can appear quite normal to casual observation. The difficulty for individuals with amnesia arises when information must be recalled after a distraction-filled interval, or when recall is attempted after the presentation of an amount of information that exceeds immediate memory capacity (e.g., eight or more digits or four or more objects). In this situation, normal individuals will usually recall more items than amnesic patients.

An animal model of human amnesia was established in the monkey in the early 1980s (Mishkin, 1982; Squire and Zola-Morgan, 1983). Systematic and cumulative work during the subsequent 10 years, using the animal model, succeeded in identifying the system of structures in the medial temporal lobe essential for memory (Squire and Zola-Morgan, 1991) (Figure 1). This system includes the hippocampal region (i.e., the hippocampus proper, the dentate gyrus, and the subicular complex), the entorhinal cortex, and the adjacent perirhinal and parahippocampal cortices. Since that time, a number of studies in monkeys have been carried out, comparing performance on memory tasks in monkeys with damage to different components of the medial temporal lobe. This work has shown that the severity of memory impairment depends on the locus and extent of damage within the medial temporal lobe memory system. Damage limited to the hippocampal region causes significant memory impairment, but additional damage to the adjacent cortex increases the severity of memory impairment. It is important to note that the discovery that larger medial temporal lobe lesions produce more severe amnesia than smaller lesions is compatible with the idea that structures within the medial temporal lobe might make qualitatively different contributions to memory function. This is because anatomical projections carrying information from different parts of the neocortex enter the medial temporal lobe memory system at different points (Suzuki and Amaral, 1994).

Another important brain area for declarative memory is the diencephalon. The important structures include the mediodorsal thalamic nucleus, the anterior thalamic nucleus, the

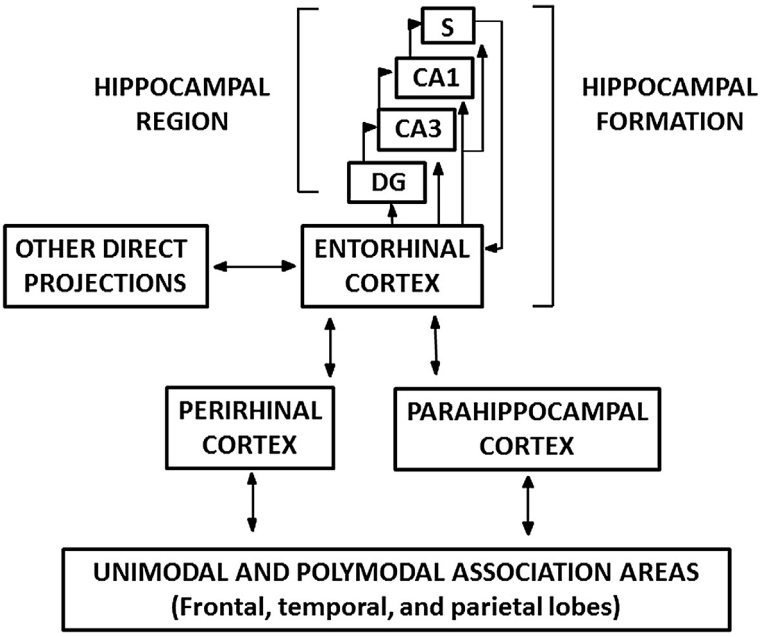

Figure 1 Schematic view of the medial temporal lobe system. The perirhinal and parahippocampal cortices receive projections from unimodal and polymodal areas in the frontal, temporal, and parietal lobes. In turn, the perirhinal and parahippocampal cortices account for nearly two-thirds of the cortical input to the entorhinal cortex the primary source of cortical input to the hippocampal region (i.e., the cell fields of the hippocampus proper, the dentate gyrus, and the subicular complex). The hippocampal region together with the entorhinal cortex makes up the hippocampal formation. The entorhinal cortex also receives other direct inputs from orbital frontal cortex, cingulate cortex, insular cortex, and superior temporal gyrus. All these projections are reciprocal.

internal medullary lamina, the mammillothalamic tract, and the mammillary nuclei. Because diencephalic amnesia resembles medial temporal lobe amnesia in many ways, these two regions likely form an anatomically linked, functional system.

At about the same time that an animal model of amnesia was established, detailed case studies of human amnesia provided extensive neuropsychological information about memory impairment, as well as detailed neuropathological information about the locus and extent of damage in each patient's brain. Recent developments in brain imaging techniques have also made it possible to detect some kinds of neuropathology in living patients. The available neuroanatomical information from humans is consistent with the findings from monkeys. These lines of work have greatly clarified the different forms of memory and their anatomical basis, and have made it possible to pursue parallel work in simpler animals such as rats and mice. As a result, one can now study memory in rodents and have some confidence that what one learns can be relevant to the human condition.

Declarative Memory

An important finding from neuropsychological studies of memory is that memory is not a single entity but is composed of several separate systems (Squire, 1992; Schacter and Tulving, 1994). Figure 2 illustrates how long-term memory can be divided into several parallel memory systems. The major distinction is between conscious knowledge of facts and events, referred to as declarative or explicit memory, and other nonconscious knowledge systems that provide for the capacity of skill learning, habit formation, the phenomenon of priming, and certain other ways of interacting with the world, collectively referred to as nondeclarative or implicit memory. Nondeclarative memory is not itself a brain-systems construct. Rather, this term encompasses several different kinds of memory. Nondeclarative forms of memory have in common the feature that memory is nonconscious. Memory is expressed through performance and does not require reflection on the past or even the knowledge that memory is being influenced by past events.

Declarative memory (explicit memory, relational memory) is a brain-systems construct, referring to the capacity for conscious recollections about facts and events that is dependent on the integrity of the hippocampus and anatomically related structures in the medial temporal lobe and diencephalon. Declarative memory seems to be especially good at associating the various aspects of a context that are present at a particular time and place, thereby creating a memory of an episode. Declarative memory is also well suited for connecting the pieces of information needed to acquire a new fact (e.g., the capital of North Dakota is Bismarck). It is sometimes pointed out that declarative memory allows one to model the external world, and it is either true or false; nondeclarative memory is about performance, and it is neither true nor false.

Declarative memory is the kind of memory impaired in amnesia. However, the concept of declarative memory is not defined solely in terms of what amnesic patients can and cannot learn. Other characteristics of declarative memory have also been identified. For example, work with amnesic patients

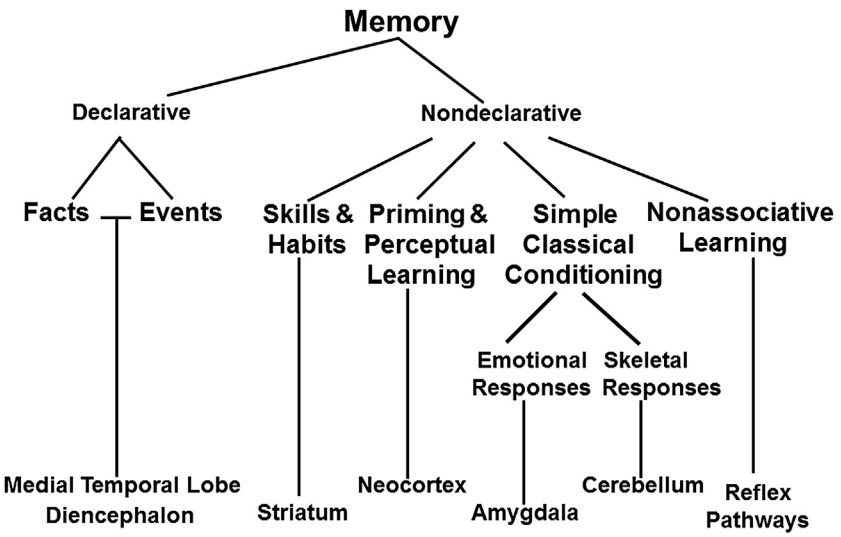

Figure 2 A taxonomy of mammalian long-term memory systems. The taxonomy lists the brain structures thought to be especially important for each form of declarative and nondeclarative memory. In addition to its central role in emotional learning the amygdala is able to modulate the strength of both declarative and nondeclarative memories.

has emphasized the notion of conscious recollection, though this description is unhelpful when considering learning and memory in animals. Other characteristics have emerged from the studies of experimental animals. Declarative memory is fast, it is not always reliable (i.e., forgetting and retrieval failure can occur), and it is flexible in the sense that it is accessible to multiple response systems. It is especially suited for one-trial learning, as well as forming and maintaining associations between arbitrarily different pieces of material.

Recognition Memory

A particularly good example of declarative memory involves the capacity to identify a recently encountered item as familiar, a capacity termed recognition memory. Recognition memory is impaired in amnesic patients, and it is impaired in monkeys and rodents following damage to the hippocampal region. The assessment of recognition memory in animals has depended mainly on two behavioral tasks, delayed nonmatching to sample and novel object recognition. In delayed nonmatching to sample, two objects are presented, a new one and one that was presented earlier. The animal must choose the new object in order to obtain a food or liquid reward. In the novel object task, as adapted for the monkey, a new picture and a recently presented picture are presented side by side, and the monkey's spontaneous tendency to look at the new picture is measured. When this task is administered to rodents, the animal's tendency to explore novel objects over recently encountered objects is measured. Similar to humans, monkeys and rats have an innate preference for novelty, and by showing a preference for viewing the new picture or exploring the new object, the animal is demonstrating that it remembers the old, and now less interesting stimulus.

Studies in humans, monkeys, and rodents with selective hippocampal lesions document impairment in these tasks. Although there have been some studies in which no impairment on these tasks was detected after hippocampal damage, there are considerable data to suggest that the hippocampus is critical for recognition memory performance (Squire et al., 2007 for review, Zola et al., 2000). Recent studies have identified neural signals within the hippocampus in both monkeys and humans that might underlie or support recognition memory performance (Jutras and Buffalo, 2010; Rutishauser et al., 2006). When the hippocampal damage occurs together with damage to adjacent cortices, including the perirhinal cortex, the impairment is more severe. The evidence that the hippocampal region is essential for normal recognition memory is entirely consistent with current ideas about the role of the hippocampus in declarative memory and with the view that the hippocampus is essential for acquiring information about relationships, combinations, and conjunctions among and between stimuli (Manns and Eichenbaum, 2006). The recognition tests ask whether an item that had been presented recently now appears familiar. A recognition (or familiarity) decision requires that the stimulus presented in the retention test be identified as what was presented during learning. At the time of learning a link must therefore be made between the to-be-remembered stimulus and its context or between the stimulus and the animal's interaction with it. It is this associating and the ability to retain relational information across time that many have supposed is at the heart of declarative memory and in turn is the function of the hippocampal region in both humans and nonhuman animals.

Episodic and Semantic Memories

Episodic memory refers to the capacity for recollecting an autobiographical memory of events that occurred in a particular spatial and temporal context. Semantic memory refers to the capacity for recollecting general knowledge and facts about the world. Both types of memory are declarative in the sense

that subjects are retrieving information explicitly, and they are aware that stored information is being retrieved. There is good agreement that episodic memory is impaired in amnesia following damage to the medial temporal lobe memory system. However, there is less agreement about semantic memory. By some accounts, amnesic patients are proportionately impaired in both episodic and semantic memory. By other accounts, the capacity for semantic memory is spared, or partially spared in amnesia relative to episodic memory ability, especially if the damage is limited to the hippocampal region.

Evidence for the second view comes especially from cases where amnesia occurred in early childhood, before much of an individual's semantic knowledge has been acquired (Vargha-Khadem et al., 1997). While these individuals have severe impairment in episodic memory, they show some ability to acquire semantic knowledge, albeit at an impoverished rate. Taken together, the data tend to favor the view that episodic and semantic memories are similarly impaired in amnesia (Squire and Zola, 1998). Other work suggests that episodic memory additionally depends on the frontal lobes, which provide a source of autobiographical involvement with the recollected material. This orientation allows for a fundamental distinction between episodic and semantic memory, without depending on differences in the anatomical basis of the two kinds of memory within medial temporal lobe structures.

Overview

Work with amnesic patients and experimental animals have converged to characterize the organization of declarative memory. Declarative memory depends on the integrity of the medial temporal lobe, which works in concert with the neocortex. Declarative memory may develop gradually from early in life. Human infants 6 months of age successfully perform the novel object recognition task, and performance on this task is impaired by hippocampal lesions in humans and monkeys. The view that some declarative memory is available early in life is also consistent with demonstrations of long-term, recall-like memory abilities in human infants.

Other work with amnesic patients has suggested that declarative memory is equivalently involved in the two fundamental processes of recall and recognition. Work in humans and animals has shown that the medial temporal lobe memory system is required at the time of learning if an enduring and retrievable long-term declarative memory is to be established. Studies of retrograde amnesia in amnesic patients and in experimental animals suggest that the brain system supporting declarative memory has a time-limited function. Medial temporal lobe lesions spare very remote memories. Thus, as time passes after learning, a process of consolidation and reorganization occurs such that permanent storage does not require the medial temporal lobe. Permanent memory is thought to be stored in a distributed way throughout the neocortex where information is received and processed. Information is still accumulating about how memory is organized, what structures and connections are involved, and what functions they support. The disciplines of both psychology and neuroscience continue to contribute to this enterprise.

See also: Amnesia: General; Autobiographical Memory; Episodic Memory; Hippocampal Amnesia; Hippocampus and Related Structures; Learning and Memory, Neural Basis of; Non Declarative (Procedural) Memory; Recognition; Short-Term Memory: Psychological and Neural Aspects; Working Memory, Neural Basis of; Working Memory, Psychology of.

Bibliography

Jutras, M.J., Buffalo, E.A., 2010. Recognition memory signals in the macaque hippocampus. Proceedings of the National Academy Sciences of the United States of America 107, 401–406.

Manns, J.R., Eichenbaum, H., 2006. Evolution of declarative memory. Hippocampus 16, 795–808.

Mishkin, M., 1982. A memory system in the monkey. Philosophical Transactions of the Royal Society of London Series B – Biological Sciences 298, 85–95.

Rutishauser, U., Mamelak, A.N., Schuman, E.M., 2006. Single-trial learning of novel stimuli by individual neurons of the human hippocampus-amygdala complex. Neuron 49, 805–813.

Schacter, D.L., Tulving, E., 1994. Memory Systems 1994. MIT Press, Cambridge, MA.

Scoville, W.B., Milner, B., 1957. Loss of recent memory after bilateral hippocampal lesions. Journal of Neurochemistry 20, 11–21.

Squire, L.R., 1992. Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. Psychological Review 99, 195.

Squire, L.R., Wixted, J.T., Clark, R.E., 2007. Recognition memory and the medial temporal lobe: a new perspective. Nature Reviews in Neuroscience 8, 872–883.

Squire, L.R., Zola, S.M., 1998. Episodic memory, semantic memory, and amnesia. Hippocampus 8, 205–211.

Squire, L.R., Zola-Morgan, S., 1991. The medial temporal lobe memory system. Science 253, 1380–1386.

Squire, L.R., Zola-Morgan, S.M., 1983. The neurology of memory: the case for correspondence between the findings for human and honhuman primate. In: Deutsch, J.A. (Ed.), The Physiological Basis of Memory. Academic Press, New York.

Suzuki, W.A., Amaral, D.G., 1994. Perirhinal and parahippocampal cortices of the macaque monkey: cortical afferents. Journal of Computational Neurology 350, 497–533.

Vargha-Khadem, F., Gadian, D.G., Watkins, K.E., Connelly, A., Van Paesschen, W., Mishkin, M., 1997. Differential effects of early hippocampal pathology on episodic and semantic memory. Science 277, 376.

Zola, S.M., Squire, L.R., Teng, E., Stefanacci, L., Buffalo, E.A., Clark, R.E., 2000. Impaired recognition memory in monkeys after damage limited to the hippocampal region. Journal of Neuroscience 20, 451–463.

Deconstruction

Jonathan Culler, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Deconstruction is a mode of philosophical and literary analysis derived from the work of the French philosopher Jacques Derrida that questions the most basic philosophical categories or concepts. In philosophy, it involves the undoing of fundamental oppositions through a demonstration that they are constructions, not part of the nature of things but categories imposed by discourses that provide the evidence for their deconstruction. Beyond philosophy, deconstruction influenced work in fields such as law, anthropology, psychoanalysis, theology, and gender studies by providing an impetus to treat the categories and the metalanguage of a discipline as interested constructions rather than as a neutral framework, and by encouraging attention to the rhetorical and performative dimensions of texts as well as to their explicit arguments.

Deconstruction is a mode of philosophical and literary analysis, derived from the work of the French philosopher Jacques Derrida that questions the most basic philosophical categories or concepts. In the 1970s, the term was applied to analyses, by Derrida and others, of philosophical and literary writings. In the 1980s, it came to designate more loosely a range of radical theoretical enterprises in diverse areas of the humanities and social sciences, including law, psychoanalysis, architecture, theology, feminism, gay and lesbian studies, historiography, and political theory. What these enterprises shared was a critical dismantling of the conceptual oppositions that had previously been regarded as fundamental to the disciplines in question. The term *deconstruction* was one of the flash points in critical and cultural debates of the last quarter of the twentieth century, a rallying point and term of abuse, the name of a body of difficult, deeply influential theoretical writings, and also the name for a broad movement in twentieth-century thought in which assumptions or presumptions of a millennial philosophical, literary, and critical tradition have been put in question. Although the term primarily designates a mode of analysis derived from the work of Jacques Derrida, *deconstruction* has somewhat surprisingly entered popular usage: in July 2007, when Barry Bonds was on pace to break the record for home runs in Major League Baseball, the sports section of the *New York Times* reported "Bonds spent his free time in Chicago praying with Jesse Jackson, visiting with his family, and deconstructing videotape of his swing."

Deconstruction has thus entered the language, as a synonym for analysis that exposes mechanisms, procedures, or habits, though its connotations do not appear to be settled. Woody Allen's film *Deconstructing Harry* announces a dismantling or undoing that does not bode well for Harry, whereas the *Times* report suggests that Bond's deconstructive analysis provided insights into his swing and enabled him 'to correct some flaws.' In general, the term has come to designate a critical dismantling or undoing: in fashion, for example, a 'deconstructed jacket' is one in which various parts are clearly displayed as parts. However, in polemics about late twentieth-century thought, 'deconstruction' is sometimes used as a pejorative synonym for 'post-structuralism' or 'postmodernism' – one that suggests nihilism or frivolous skepticism.

Deconstruction and Philosophy

Deconstruction undertakes a radical critique of fundamental categories of Western thought by exploring how these categories have been constructed in and by the discourses that rely on them. It is a mode of analysis that examines not only the arguments of philosophical texts but also the rhetorical procedures and devices they employ and the tensions or contradictions between what is claimed or assumed and what the texts themselves must do in order to support such claims. A major focus of deconstruction has been the traditional binary oppositions that have structured Western thought since the time of the Greeks: for example, the oppositions between inside and outside, mind and body, literal and metaphorical, speech and writing, presence and absence, nature and culture, intelligible and sensible, form and meaning. Each of these oppositions is hierarchical, in the sense that one of its terms has been taken to be primary and fundamental, and the other secondary and derivative (Derrida, 1981). Thus, nature is logically prior to culture; writing is seen as merely a way of representing speech, which is taken to be the basic form of language; meaning is what comes first and is then given expression by form.

A deconstruction of these oppositions asks how the philosophical enterprise and thought in general have relied upon the privileging of one term and invites us to consider whether in fact such hierarchies should not be revised. In particular, a deconstruction questions these hierarchical oppositions by showing how they have been constructed and, in the process, undoing that construction and establishing a new relation between the terms. It can be shown, for example, that the idea of 'nature' is a product of culture: conceiving something as prior to culture is a specific cultural operation whose import and function need to be assessed. In the opposition between nature and culture, the primacy of nature cannot, therefore, be taken for granted. What we have, rather, is an opposition elaborated within culture; what counts as nature in any historical moment will be a fact about that culture. To argue in this way is to invert and restructure the opposition (to 'displace' it) in a way characteristic of deconstruction. Instead of a primary nature and a secondary culture, we discover a variable distinction between nature and culture within culture.

For Derrida, the most telling and pervasive opposition has been the one that treats writing as secondary or derivative with respect to speech. According to this opposition, in speech the ideas and intentions of the speaker are immediately present; it is a direct and authentic form of language, whereas writing is merely a graphic representation of the spoken word (a sign of a sign) and hence marked by absence and possibilities of discrepancies between form and meaning (Derrida, 1976). By setting aside writing as a secondary and derivative, a mere representation open to misunderstanding, accounts of language have taken as their object an idealized form of speech, where the linguistic form is a direct expression of what the speaker, as we say, 'has in mind.' But Derrida argues that linguistic forms can function as signs only to the extent that they can be repeated in different contexts, in the absence of any particular speaker's presence or intention (Derrida, 1988). Speech is only possible, in other words, to the extent that it has the qualities assigned to writing, such as absence, difference, and the possibility of misunderstanding. One mark of this, Derrida has shown, is the frequent recourse, in attempts to describe speech, of examples and metaphors drawn from writing. In effect, speech has been described as a form of writing, even when the claim has been that writing is derivative from speech (Derrida, 1976). This deconstruction of the traditional hierarchical opposition between speech and writing argues not that there are no differences between speech and writing but that the traditional opposition is untenable and that both speech and writing are forms of a general writing (*archi-écriture*), which is the condition of possibility for any system of representation whatsoever.

Derrida argues that treating writing as secondary to speech is part of what he calls the 'logocentrism' of Western culture: the presumption of an order of truth or thought prior to its representation by signs (Derrida, 1976). Logocentrism tries to treat representation as inessential rather inextricably involved in the structure of phenomena, but an attentive reading of the texts of the philosophical tradition shows that they tell a different story. Derrida's work attempts to demonstrate that discourses that treat writing as secondary in fact must make use of notions linked to writing when they characterize speech and that thus speech can be seen as a version of a generalized writing, which is the condition of language and thought in general. This is an instance of the deconstruction of concepts seen as fundamental, such as 'presence,' 'truth,' 'origin,' and 'identity.'

Deconstruction may be conceived in the first instance as a critique of Western philosophy. Philosophy, Derrida argues, has been founded on a theory of 'presence,' in which such notions as truth, being, and reality are determined in relation to an ontological center, essence, or origin, based on the repression of absence and difference. But since deconstruction works not by erecting an alternative theory but by exploring how philosophical discourses have produced their concepts through rhetorical stratagems, it entails powerful practice of reading, which illuminates how texts implicitly put in question what they explicitly maintain or what they appear to assume. It has been called "a careful teasing out of warring forces of signification within a text" (B. Johnson 1980: p. 5).

Deconstruction is not, Derrida insists, a school or a method, a philosophy or a practice, but something that happens, as when the arguments of a text undercut the presuppositions on which it relies, or as when the term *deconstruction* (*déconstruction*), which Derrida introduced somewhat casually as a French translation of the terms *Abbau* and *Destruktion* ('unbuilding' and 'destruction') in Martin Heidegger's philosophical writing, takes on a life of its own, escaping the control of the author and coming to refer to a broad intellectual process or movement that the end of the twentieth century by no means exhausted.

Deconstruction arises from the continental tradition in philosophy rather than the analytical tradition. Analytic philosophers have generally been resistant to deconstruction and to Derrida, even to the point of excluding him from philosophy and associating him with literary theorists, but some analytical philosophers have successfully recast Derrida's arguments in the idiom of analytic philosophy (Bearn, 1995; Wheeler, 2000). Other philosophers working in the continental tradition have sought to rescue Derrida from literary theory and to recapture him for the tradition of continental philosophy as a systematic philosopher in a traditional mode (Gasché, 1988). Bennington provides both a systematic description of his thought (Bennington and Derrida, 1993) and analysis of philosophical responses to Derrida's work (2000).

Deconstruction and Literary Studies

In the 1970s and 1980s, deconstruction played a major role in the animation and transformation of literary studies by what was called simply 'theory,' for short. Briefly stated, 'theory' is concerned with questions about the nature of language, the production of meaning, or the relations between literature and the numerous discourses that structure human experience and its histories. In the United States, deconstruction was associated in particular with a group of scholars at Yale University, where Derrida taught regularly as a visiting professor. In the hands of Paul de Man and Barbara Johnson, for instance, deconstruction became a powerful practice of rhetorical reading, attentive to the rhetorical strategies and figurative structures of literary works, to what is at stake in texts of all sorts and to how texts engage those stakes.

Because Derrida does philosophy by reading texts slowly, with virtuoso resourcefulness, attentive to their rhetorical strategies and ideological investments, his work appealed to students and teachers of literature, who found in it close reading that (1) was not subservient to the ideological notion of organic form that underlay the most widespread practice of close reading, that of the New Criticism (the name generally applied to the major school of literary analysis in the United States and United Kingdom from the 1960s on), and (2) showed that philosophically significant issues are at play in these texts: the oppositions that they engage structure thinking about fundamental questions.

Deconstructive literary criticism was often compared, especially by its critics, to an earlier Anglo-American New Criticism that sought to explicate and appreciate verbal works of art as complex constructions and emphasized the role of paradox and irony in these intricate artifacts. However, deconstructive readings, by critics such as those mentioned above, differ in several ways (Culler, 2007). They treat works of art not as the harmonious fusion of form and content but as instances of the

intractable conflicts between modes of meaning. They generally examine the individual work not as a self-contained artifact but a product of relations with other texts or discourses, literary and nonliterary. Finally, they are especially interested in the ways in which the works may offer implicit critiques of the categories critics use to analyze them.

If in the twenty-first century there are fewer critical studies devoted to showing how literary works subvert the premises on which they rely or the meanings they have been presumed by the earlier critical tradition to articulate, deconstruction has come to be broadly diffused within literary studies, so that conceptions associated with deconstruction (the critique of the idea of organic form, for instance, and the notion that writing about literature should explore how it engages the presuppositions on which it relies) have become very widespread.

The most distinctive deconstructive work by critics such as de Man and B. Johnson explores literary works' engagement with major philosophical stakes (de Man 1979). Johnson, author of incisive, elegant deconstructive essays on poets and novelists (Johnson 1980), broadened her engagement with psychoanalysis, women's writing, feminist theory, African-American literature, and cultural studies, in ways indicative of the expansive influence of deconstruction. Two essays that might be cited as exemplary of deconstructive reading are 'Anthropomorphism in Lyric and Law,' and 'Muteness Envy.' The former brilliantly brings together de Man's analysis of the trope of anthropomorphism that takes the human for granted in treating something as human and is central to the functioning of lyric poetry, and the anthropomorphizing operations at issue in legal debates. The determination of what the law will treat as a person – what entities or organizations are deemed to have the rights of a person – has immense consequences for the allocation of social privileges and resources, and an account of the rhetoric of legal reasoning in this realm proves extremely illuminating (Johnson, 2008). 'Muteness Envy' reads Keats's 'Ode on a Grecian Urn,' with its 'still unravished bride of quietness,' against Jane Campion's film *The Piano* and the critical reception of the film, so as to explore the cultural construction and aestheticization of female muteness as a repository of feminine value. Johnson concludes that the work performed by the idealization of women's silence is that it helps culture not to be able to tell the difference between women's pleasure and their violation, a problem at issue in *The Piano* and the debates about it (Johnson, 1998). This is an example of efficient deconstructive reading engaging a variety of cultural texts on issues of great moment.

Derrida himself wrote extensively about literary works (Derrida, 1992), in what one critic calls "the most significant, far-reaching, and inventive exploration of literature for our time" (Attridge, 2004: p. 139); but the mode of these essays is different from the exploration of binary oppositions characteristic of what is called deconstruction. Pursuing particular motifs rather than offering interpretations of a work as a whole, Derrida addresses the work's attempt to be a singular event.

During what were called 'the culture wars' of the 1980s and 1990s in the United States, people combatting pressures to open the canon to works by women, members of minority groups, and Third World cultures or moves to reject high culture in favor of popular culture, sometimes seized upon the term *deconstruction* to label the developments that they saw as threatening Western civilization itself. In fact, deconstructive critics, despite their sympathies with movements to broaden the canon, were deeply engaged with the texts of the high cultural tradition, from Plato to Proust, which they saw as particularly savvy and enlightened, richer, and more subversive than readers had previously realized. They were scarcely agitating to substitute soap operas for Shakespeare: the desire to teach popular culture or to substitute non-Western texts for historically crucial works of the Western tradition came from practitioners of cultural studies, who often saw deconstruction as a dangerous, elitist enemy, committed to high culture, engaged with difficult philosophical texts and their specialized jargon, and valuing difficulty and complexity.

Deconstruction Beyond Philosophy and Literature

Critics of deconstruction have tended to attack it either as obscurantist wordplay, excessively invested in supposedly minor or marginal rhetorical features of texts and cast in language that obfuscates rather than illuminates, or as a nihilistic denial of authority, certainty, and foundations, for which 'anything goes' (Ellis, 1989). But despite attacks, the impact of deconstruction on modern thought, not just literary studies, has been extensive. Most generally, it has encouraged the questioning of the hierarchical oppositions that structure any field of inquiry and an interest in whether and how these fundamental oppositions are subverted by the phenomena that they are used to describe. Deconstruction is thus a powerful version of the general critique of supposedly scientific metalanguages – sets of terms or concepts used to analyze a domain and which are regarded as external to the phenomena they purport to analyze. Deconstructive analyses explore the ways in which the supposedly neutral or purportedly scientific metalanguage is deeply affected by the phenomena it is supposed to analyze. Thus, it might show, for instance, that the terms deployed by theory of metaphor or figurative language generally are themselves scarcely free of metaphor or explore how a psychoanalytic theory is itself affected by the mechanisms of repression and wish fulfillment it purports to analyze (Derrida, 1987). Since within literary and cultural studies, it was above all structuralism (*see* Structuralism) that sought to elaborate metalanguages for analyzing cultural phenomena as a series of languages or signifying systems, the deconstructive critique of the possibility of a neutral metalanguage was seen as a step-beyond structuralism and hence as a case of 'post-structuralism.'

There are two main aspects to the influence of deconstruction outside of literary and philosophical studies. On the one hand, granting primacy not to what a text 'says' but, rather, to how what it 'does' relates to what it says, deconstruction emphasizes that rhetorical structures are centrally at work in language of all sorts and that discourse has an important 'performative' dimension, structuring experience in particular ways. It thus seconds and invigorates the constructivist tendencies of numerous disciplines: the attempt to show how that the phenomena studied by a discipline are not simply given to experience but are produced by conceptual frameworks and discursive practices. On the other hand, as a critical exploration of fundamental oppositions and an attempt to intervene and alter the value attributed to particular terms,

deconstructive thinking has affected not just how texts are read but the sense of what a discipline's goal might be.

Deconstruction and Psychoanalysis

In psychoanalysis, for instance, deconstructive readings of texts by Sigmund Freud and such successors as Jacques Lacan have done several things:

1. focused on the structuring role of language in the formation of the psyche;
2. shown how psychoanalytic case studies are themselves structured by the kinds of psychic mechanisms that they purport to analyze (so Freud's or Lacan's writings are themselves organized by the processes of repression, condensation, and displacement that they theorize); and
3. questioned the complicity of psychoanalysis with a logocentrism that believes in originary principles or an ultimate reality (Derrida, 1987).

Deconstructive thinking in psychoanalysis may seek not the origin of a trauma but the stakes of a belief in origin (Caruth, 1996).

Feminism/Gender Studies/Queer Theory

Although feminists have often been suspicious of deconstruction as a typically male pastime of abstract and universalizing thought that in effect sought to deny, for example, the authority of women's experience (since the appeal to experience always conceals presuppositions), some strands of feminist thinking espouse a deconstruction of the opposition between men and women and a critique of essentialist notions of identity (Elam, 1994). As the translator of Derrida's *Of Grammatology* and an eminent feminist, Marxist and postcolonial critic, Gayatri Spivak has been a powerful voice articulating deconstruction with feminist and other concerns (Spivak, 1988). But the most prominent result of deconstruction has been the work of Judith Butler, who draws on Derrida and Foucault in her theorizing of gender and identity, challenging the claim that a feminist politics requires an identity for woman (for Butler, identity is the product or result rather than the source of action). Butler develops a 'performative' concept of gender and sexual identity, modeled on the way in which linguistic acts (such as promising) work to bring into being the entities (the promise) to which they refer. This perspective has been influential in the realm of gay and lesbian studies or 'queer theory,' as the academic avant-garde linked to movements of gay liberation has styled itself (Butler, 1990, 2004).

Deconstructivist Architecture

The impact of deconstruction has spread beyond the humanities and social sciences to the field of architecture, where deconstructive architecture has come to mean design that flaunts, often in a parodic way, the elements from which it is constructed. In 1985, the architect Bernard Tschumi invited Derrida to collaborate on the design of a section of the Parc de la Villette in Paris, a large park surrounding new museums and exhibition spaces, in which design theory was to be staged in various ways. In Tschumi's plan, a series of large red cube-shaped spaces, the follies, were to be located at specific points; each cube was to be transformed, through 'deviation' into a folly, an explosion or expenditure of structure: the superimposition of different autonomous and logical structures subverted the notion of totality, and the design not only rejected the relation to context by which architecture is often justified but, dislocating and deregulating meaning, it aimed, according to Tschumi, at an architecture that means nothing (Tschumi 1988).

In 1988, Philip Johnson organized an exhibition at the Museum of Modern Art in New York entitled 'Deconstructivist Architecture.' In this perspective, what makes architecture deconstructive is not any derivation from the deconstruction in philosophy but its ability to disturb our thinking about form. This architecture disturbs and subverts by uncovering the instabilities or dilemmas hidden in traditional form, exposing the unfamiliar hidden within the familiar (Johnson and Wigley, 1988).

Religion and Theology

Less surprising is the interaction of deconstruction and the thinking of religion. As a critique of metaphysics, and specifically of the metaphysics of presence or the logocentrism of Western culture, deconstruction seems a resolutely antitheological enterprise, a critique of the theological motifs and structures that continue to undergird our thinking. But this devotion to uncovering the cryptotheological structure of Western secular thought, especially philosophy, gives rise to the idea that deconstruction is a version of negative theology.

Some scholars highlight the role of the concept of the messianic in Derrida's thought; and seek to recover Derridean concepts for a theology, linking the motif of deferral to the waiting for the messiah that marks Christian eschatology and moving from the argument that deconstruction contains religions motifs and a critique of a triumphant secularizing reason to the conviction that in 'religion without religion' or 'the messianic without the messiah' deconstruction gives us not the impossibility of religion but a religion with deniability, religion without the flaws of actual religions (Caputo, 1997). But the distinction Derrida makes between the messianic and messianism is crucial, not a quibble: it is the difference between analyzing a structure of awaiting and deferral and believing in an actual messiah (Derrida, 2002a).

Discussions of deconstruction and religion seem divided between, on the one hand, those who want to bring religion to Derrida and deconstruction, show deconstruction to be, finally, religious in its structure and commitments, giving rise to an ethics that preserves what is most valuable in modern religious commitment, and, on the other hand, those who want to bring deconstruction to theology in order to make it more philosophically complex, sophisticated, and responsible (Taylor, 1982). Must deconstruction assume atheism in order to function or might it work within a theological context and help produce a theology that could 'escape' philosophy? A third approach to deconstruction and religion, using deconstruction

to critique religion and theology, would seem the most obvious, at least based on Derrida's early work, but it was generally absent from the literature until Hägglund's *Radical Atheism* (2008), which decisively shows that deconstruction is not based on any desire for transcendence but rather on the valuing of the mortal, the transient. Derrida's contribution to the book *Religion* that he edited with an Italian philosopher takes on not theology so much as religion itself, as a social and intellectual phenomenon, and shows that deconstruction does indeed have a critical vantage point on matters religious (Derrida and Vattimo, 1998).

Politics, Law, Ethics

What is the relation between deconstruction and politics? Bennington notes that because Derrida's work in philosophy is very radical, there has been an assumption that it should give rise to a politics or political philosophy that is equally radical, so Derrida can be faulted for not living up to the expectations of those for whom politics is the real (Bennington, 2000). While Derrida has taken a position on many political issues and controversies, this sort of commitment has not been satisfying to many, who expect a political theory that will change the world.

Derrida has written extensively immediate political issues, such as apartheid, immigration law, terrorism, the death penalty, and European integration, and about political theory in the broadest sense, approaching politics and democracy through the issues of friends and enemies (Derrida, 2005, 1997). *Specters of Marx* interrogates the legacy of Marxism and its importance for a post-Marxist world (Derrida, 1994). The American Declaration of Independence has been analyzed as exemplary of the violence of acts of foundation, in which performative and constative dimensions of language fail to coincide (Derrida, 2002b). Woven through Derrida's work is a reflection on democracy, deconstruction, and decision: something is a decision only if it cannot be programed but occurs in a situation of undecidability; it must interrupt determination that is nonetheless its condition of possibility. (Bennington, in Royle, 2000). And democracy, a structure based on counting – the counting of singularities – is a concept in the name of which we criticize every determination of democracy; it is in the name of a 'democracy to come' that we deconstruct any given concept of democracy.

There has been considerable investigation of the topic of deconstruction and politics, and of the way in which deconstructive thinking can be marshaled for diverse political ends (Bennington, 1994; Beardsworth, 1996). Scholars/activists have developed a post-Marxist Marxism inflected by deconstruction (Laclau and Mouffe, 1985); and American political scientists have brought the thinking of the paradoxes of difference and the deconstruction of hierarchical oppositions to the discourse of political science (Connolly, 1992; Corlett, 1989; Martin, 1992). One should also place under the rubric of deconstruction and politics much writing in postcolonial studies (Spivak, 1999; Young, 1990).

In the field of law the Critical Legal Studies movement, which has focused on conflicts between principles and counter-principles within the system of legal doctrine has many affinities with deconstruction: critiques of the oppositions such as public vs private, essence vs accident, substance vs as form that are fundamental to the realm of law, and demonstrations that legal doctrine and argument are attempts to paper over contradictions, which nonetheless reassert themselves (Kelman, 1997). Derrida's own writing in this field has something of a different cast: his 'Force of Law: the Mystical Foundations of Authority,' for instance, explores the inextricability of foundational violence and justice (Derrida, 2002a).

The question of justice has been an important one for deconstruction, and leads to debates about deconstruction and ethics. Since deconstruction has been associated with disrespect for established norms and traditions and embrace of a Nietzschean trajectory beyond good and evil, the last thing it might seem to provide is an ethics. If the idea of ethics, with the concepts of law, responsibility, duty, and decision, is derived from metaphysics, how could deconstruction not put these in question? But the very fact that one formulates this as a matter of necessity or obligation draws attention to the possibility of asking what drives deconstruction? What sort of necessity, obligation, or commitment, ethical or not, is at work here, in the drive to deconstruct or in the sense of the importance of attending to deconstructions that happen?

This is a matter of ethics: what sort of value or obligation impels or compels one's actions? In the case of deconstruction, the question of ethics goes to the heart of the method. What impels deconstruction; why bother with it? The answers given often involve the structure of obligation itself, as writers explore the nature of the demand that drives deconstruction, whether the desire for justice, respect for the otherness represented by the text, or obligation to the otherness of the other. Derrida himself had a longstanding engagement with the thought of the philosopher Emmanel Levinas, whom he saw as awakening us to a conception of an unlimited responsibility that exceeds and precedes our own freedom. If one of the oldest ethical injunctions is 'Do unto others as you would have them do unto you,' a modern version, in what is called the 'Ethics of Alterity,' is to respect the otherness of the other: do not presume that the other is like you and thus claim a posture of universality, as if your views and feelings had universal validity. The problem of relation between the singularity of the other and the generality or universality that any ethical proposition entails is unresolved, an ongoing problem that will continue to engage thought of a deconstructive variety, whether in the future it call itself deconstructive or something else.

A concrete ethical issue to which Derrida turned in later years is the relation between the human and the animal. The characterization of the animal as other – Aristotle and Descartes denied reason to animals – has long helped to define the human. What is the nature and the impact of the distinction between the human and the animal? How is this distinction made, on what grounds, and with what values? Pursuing such questions, deconstruction contributes to the burgeoning interdisciplinary field of Human/Animal Studies. After women's liberation and gay liberation, animal liberation seems a further step – or if not animal liberation, recognition that the human treatment of animals for our own convenience is difficult to defend. Some critiques of the human/animal opposition show commonalities and continuities. On the other hand, a powerful strain of recent theoretical work focuses on the

discontinuities, the radical otherness and inaccessibility of animals, whom we cannot presume to understand (especially once we move beyond dogs and horses). Stressing the role that notions of the animal have played in defining the human, this approach demands respect for the otherness of animals and accuses proponents of the first approach of anthropomorphizing, treating animals according to human models.

Derrida seeks to escape the alternatives of a projection that appropriates and an interruption that excludes. He nevertheless emphasizes both the difficulties of understanding the animal's point of view and the human-centered violence of grouping everything from ants to zebra's as instances of 'the animal.' What happens if we move beyond thinking of cute middle-sized mammals and consider insects or birds, for instance? Any attempt to give a single answer concerning the conceptual relation of the human to the animal seems grotesquely human-centered (Derrida, 2008). The effect of recent theory here is to begin exploring the difficulties of maintaining a firm barrier between human and other species and recognizing the irreducibly multiple and reconfigurable relations with animals.

In its diffusion in the humanities and social sciences, following its greatest moment of fame in the 1970s and 1980s, deconstruction in the broadest sense has become a critique of categories taken as natural and a drive to pursue the analysis of the logic of signification in a given area as far as one can, even if the result is disquieting – an exacerbation of questions or problems rather than their resolution. It has thus joined with other strands of postmodern and post-structuralist thinking to inspire suspicion of established categories and canons and skepticism about objectivity. Although opponents of deconstruction have often sought to link its critique of established values to an amoral nihilism, as a critique of scientistic pretensions in the humanities and social sciences (among other things) deconstruction has included a strong element of ethical reflection, particularly about the conditions of possibility of ethics or justice (Critchley, 1992).

See also: Feminism in Anthropology; Feminisms and Acculturation around the Globe; Feminist Political Theory and Political Science; Postmodern Urbanism; Postmodernism in Sociology; Postmodernism: Methodology; Postmodernism: Philosophical Aspects; Structuralism.

Bibliography

Attridge, Derek, 2004. The Singularity of Literature. Routledge, London.
Beardsworth, Richard, 1996. Derrida and the Political. Routledge, London.
Bearn, Gordon, 1995. Derrida dry: iterating iterability analytically. Diacritics 25 (3), 3–25.
Bennington, Geoffrey, 1994. Legislations: The Politics of Deconstruction. Verso, London.
Bennington, Geoffrey, 2000. Interrupting Derrida. Routledge, London.
Bennington, Geoffrey, Derrida, Jacques, 1993. Jacques Derrida. University of Chicago Press, Chicago, IL.
Brunette, Peter, Wills, David (Eds.), 1994. Deconstruction and the Visual Arts: Art, Media, Architecture. Cambridge University Press, Cambridge, UK.
Butler, Judith, 1990. Gender Trouble: Feminism and the Subversion of Identity. Routledge, New York.
Butler, Judith, 2004. Undoing Gender. Routledge, New York.
Caputo, John, 1977. The Prayers and Tears of Jacques Derrida: Religion without Religion. Indiana University Press, Bloomington, IN.
Caruth, Cathy, 1996. Unclaimed Experience: Trauma, Narrative, and History. Johns Hopkins University Press, Baltimore, MD.
Connolly, William, 1992. Identity/Difference: Democratic Negotiations of Political Paradox. Cornell University Press, Ithaca, NY.
Corlett, William, 1989. Community without Unity: A Politics of Derridean Extravagance. Duke University Press, Durham, NC.
Critchley, Simon, 1992. The Ethics of Deconstruction: Derrida and Levinas. Blackwell, Oxford.
Culler, Jonathan (Ed.), 2003. Deconstruction: Critical Concepts in Literary and Cultural Studies, 4 vols. Routledge, London.
Culler, Jonathan, 2007. On Deconstruction: Theory and Criticism after Structuralism, 25th anniversary ed. Cornell University Press, Ithaca, NY.
De Man, Paul, 1979. Allegories of Reading. Yale University Press, New Haven, CT.
Derrida, Jacques, 1976. Of Grammatology. Johns Hopkins University Press, Baltimore, MD.
Derrida, Jacques, 1981. Positions. University of Chicago Press, Chicago, IL.
Derrida, Jacques, 1987. The Post Card: From Socrates to Freud and beyond. University of Chicago Press, Chicago, IL.
Derrida, Jacques, 1988. Limited Inc. Northwestern University Press, Evanston, IL.
Derrida, Jacques, 1992. In: Attridge, Derek (Ed.), Acts of Literature. Routledge, New York.
Derrida, Jacques, 1994. Spectres of Marx: The State of the Debt, the Work of Mourning, and the New International. Routledge, New York.
Derrida, Jacques, 1997. Politics of Friendship. Verso, London.
Derrida, Jacques, 2002a. In: Anidjar, Gil (Ed.), Acts of Religion. Routledge, New York.
Derrida, Jacques, 2002b. In: Rottenberg, Elizabeth. (Ed.), Negotiations: Interventions and Interviews, 1971–2001. Stanford University Press, Stanford, CA.
Derrida, Jacques, 2005. Rogues: Two Essays on Reason. Stanford University Press, Stanford, CA.
Derrida, Jacques, 2008. The Animal that Therefore I Am. Fordham University Press, New York.
Derrida, Jacques, Vattimo, Gianni (Eds.), 1998. Religion. Stanford University Press, Stanford.
Elam, Diane, 1994. Feminism and Deconstruction: Ms en abyme. Routledge, New York.
Ellis, John M., 1989. Against Deconstruction. Princeton University Press, Princeton, NJ.
Gasché, Rodolphe, 1988. The Tain of the Mirror: Derrida and the Philosophy of Reflection. Harvard University Press, Cambridge, MA.
Hägglund, Martin, 2008. Radical Atheism: Derrida and the Time of Life. Stanford University Press, Stanford, CA.
Johnson, Barbara, 1998. The Feminist Difference: Literature, Psychoanalysis, Race, and Gender. Harvard University Press, Cambridge, MA.
Johnson, Barbara, 1980. The Critical Difference. Johns Hopkins University Press, Baltimore, MD.
Johnson, Barbara, 2008. Persons and Things. Harvard University Press, Cambridge, MA.
Johnson, Philip, Wigley, Mark, 1988. Deconstructivist Architecture. Little Brown, New York.
Kelman, Mark, 1997. A Guide to Critical Legal Studies. Harvard University Press, Cambridge, MA.
Laclau, Ernesto, Mouffe, Chantal, 1985. Hegemony and Socialist Strategy: Towards a Radical Democratic Politics. Verso, London.
Martin, Bill, 1992. Matrix and Line: Derrida and the Possibility of Postmodern Social Theory. State University of New York Press, Albany, NY.
McQuillan, Martin (Ed.), 2001. Deconstruction. Routledge, London.
Royle, Nicholas (Ed.), 2000. Deconstructions: A User's Guide. Palgrave, London.
Spivak, Gayatri, 1988. In Other Worlds: Essays in Cultural Politics. Routledge, New York.
Spivak, Gayatri, 1999. A Critique of Postcolonial Reason. Harvard University Press, Cambridge, MA.
Taylor, Mark, 1982. Deconstructing Theology. Crossroad, New York.
Tschumi, Bernard, 1988. Parc de Vilette, Paris. Architectural Design 58 (3/4).
Wheeler, Samuel, 2000. Deconstruction as Analytic Philosophy. Stanford University Press, Stanford, CA.
Young, Robert, 1990. White Mythologies: Writing History and the West. Routledge, New York.

Deductive Reasoning Systems

Wolfgang Bibel, Technische Universität Darmstadt, Darmstadt, Germany
Christoph Kreitz, Institut für Informatik, Universität Potsdam, Potsdam, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by W. Bibel, volume 5, pp. 3346–3349, © 2001 Elsevier Ltd.

Abstract

This article explains the relevance of reasoning for prediction and explanation, i.e., for abilities that are fundamental for human survival. Since humans are prone to mistakes in their reasoning, reasoning systems offer desirable support. Deductive reasoning systems operate on a model of reasoning that results from a number of abstractions applied to human reasoning. These abstractions concern the language, the inferential relationship, and the syntactic nature of inference. By way of these abstractions, other modes of reasoning such as inductive reasoning are also covered. Basic issues concerning research in deductive systems, their performance and applications, and the psychology of human reasoning, are summarized.

Reasoning is one of the fundamental capabilities associated with intelligence as manifested in humans. It is a mental activity by which we generate knowledge not available prior to the act of generation. Experience shows that knowledge generated by reasoning on the basis of already available knowledge enables humans

- To predict states of the world not experienced before and to act accordingly and
- To explain phenomena experienced in the world so that we can understand the reasons why they happened, and make them happen again or avoid them in the future.

Prediction and explanation, and their combination, are fundamental and characteristic ingredients of the human way of life. In consequence, this is the case for reasoning as well, thus demonstrating its fundamental importance.

Experience also shows that humans have a sense for the correctness of reasoning chains and also an understanding that correct reasoning chains do not result in arbitrary predictions or arbitrary explanations. Rather, there is a correlation between the correctness of the underlying chain of reasoning and its usefulness for predictions or explanations in the real world: only predictions and explanations based on correct reasoning chains turn out to be useful in general.

On the other hand, humans are prone to mistakes in their reasoning. In coping with this and other weaknesses lies the motivation for the mechanization of reasoning in the form of *reasoning systems*. They might assist our reasoning, point out mistakes in it, or generate a correct chain of reasoning in cases too complex to be coped with by humans.

This article describes the formal basis for reasoning systems, points out a number of issues of particular importance for the design of such systems, surveys their numerous applications, addresses the philosophy and psychology of reasoning, and offers some literature on the subject.

Abstractions

Reasoning takes place in human minds. The understanding of the processes in the brain involved in reasoning is still in an embryonic state. Therefore, the study of reasoning so far can only be carried out at a higher level of abstraction. For more than two millennia, this study has been carried out at the level of language, thus abstracting from the processes that engage many subsystems in the brain including, for instance, those involved in the mental imagery system (Kosslyn, 1995). Since natural language is still too complex as a basis for the study of reasoning, further abstractions have been used. These are discussed in the present section.

Formal Languages

Any spoken or written text in natural language involves numerous linguistic constructs that carry a variety of informational content. Since reasoning relates assertions, we may for present purposes abstract from most of these linguistic constructs and restrict our focus on propositions (in the sense of logic). "The sun is shining" is a simple example of such a propositional statement.

Because of the complexity, ambiguity, and redundancy of natural language constructs even after the restriction just made, we further abstract from such propositions by use of some formal language. This should be rich enough to be able to express all our propositional knowledge, but simple enough to be useful for the manipulations needed in modeling reasoning, and concise enough to avoid ambiguities and redundancies. There are a number of formal languages meeting these criteria. They differ in the level of abstraction taken for the representation of knowledge and in the emphasis given to the various concepts involved.

The language of *propositional logic* takes the highest level of abstraction by formalizing knowledge at the level of entire propositions. For instance, the statement "The sun is shining" is taken as an atomic element in the language, thus abstracting from its intrinsic structure altogether. In the formalism the statement may therefore be abbreviated by any propositional variable, say S. As seen in section Variants, Specializations and Extensions, it is amazing how far the modeling power of this relatively poor language can already lead us.

The language of *first-order logic* (FOL) is less abstract and hence much more expressive. In our example, it allows the

distinction between the subject "The sun" in the noun phrase from the property "is shining" expressed in the verb phrase (or predicate) as well as the modeling of their relationship in a functional way, e.g., as *is_shining*(*The_sun*). In this way, it lays emphasis on the functional, or - after the philosopher Gottlob Frege - Fregean, relationships among relations, properties, and objects. Many knowledge representation formalisms are first-order formalisms in disguise. Among those are the Aristotelian language, which emphasizes sets rather than functions, the language of description logics (Bader, 2003) emphasizing concepts, associative nets emphasizing the relationships among objects and concepts in a graphical representation, and so forth. So FOL is probably the most popular formal language for knowledge representation, especially in the context of reasoning systems.

Modal, transitional, and higher order languages including lambda calculus extend the expressiveness of first-order languages even further. For the purpose of this article we put the emphasis on FOL, which includes propositional logic as a special case. Most of our considerations generalize in some way to the other more general languages.

The Inferential Relationship

Reasoning takes place with a given body of knowledge, say K, already available. In the *predictive mode* of reasoning we want to determine what facts, say P, could be true in the future. Correct reasoning distinguishes between facts possible on the basis of K and those that are impossible. In other words, there is a relation that associates K and P in the positive case. Following Frege this relation is formally denoted by $\models$. That is, $K \models P$ is short for expressing that P is a correct reasoning prediction on the basis of K. In the predictive mode of reasoning, we thus want to determine P such that $K \models P$ holds, or test for a given P whether it holds.

Human reasoning is mostly done with respect to chronological processes and to a significant extent also concerns knowledge about actions and changes taking place in time. On the other hand, a large part of the inferential relation $\models$ is so to speak "timeless". To simplify further, most logical studies have abstracted from the issues concerning time and change. Only in the last few decades have those issues become of great interest within *Intellectics*, the field studying intelligent behavior and systems within Artificial Intelligence (AI), Cognitive Science, and Neuroscience. The result of those recent studies is that the timeless part of $\models$ is already expressive enough to model the part involving time as well and thus the entire relation $\models$. Insofar this abstraction from time is without restriction of generality. Under this abstraction the predictive mode of reasoning is called the *deductive mode* of reasoning. Because of what we just said, deductive reasoning is timeless reasoning at the surface, but may model general predictive reasoning.

In the *explanatory* mode of reasoning based on K, we start with some observation O and basically look for some explanation E so that $K,E \models O$. This mode of reasoning is called *abductive* or *inductive* depending on the circumstances concerning O and E. Since $\models$ is a relation, the difference to deductive reasoning is minor; instead of the second argument of the relation $\models$ here just (part of) the first one is to be determined. The relationship is even closer. In abductive or inductive reasoning, we employ additional metaknowledge, say M, which guides the process of generalizing from given observations. Taking this into account, inductive reasoning actually boils down exactly to deductive reasoning, namely, in the form $K,M,O \models E$ whereby now the explanation E is deduced from K and O guided by M in some way.

In summary, all prevailing modes of human reasoning may actually be regarded as deductive reasoning. This explains why deductive reasoning systems have such a great importance for modeling inference in all its generality.

Syntactic Inference

The fundamental observation already made more than two thousand years ago by Aristotle and his predecessors is that reasoning (apart from psychological considerations) is a matter of form rather than of content. This became apparent once knowledge was represented in a formalized way. The knowledge "apes are mammals", in the sense of "whenever the property of being an ape holds then also the mammal property holds", or $A \rightarrow M$, and "mammals have hairs", $M \rightarrow H$, allows us to conclude that "apes have hairs", $A \rightarrow H$. The inference is valid independently of the meaning of the words apes, mammals, and hairs, or of the symbols A, M, and H for that matter. Rather it is exclusively the form of the knowledge that sanctions the validity of such an inference. This fundamental property of reasoning enables machines, which to date are ignorant of the meaning of words or symbols, to perform correct inferences.

The example demonstrates a fundamental rule of inference: $A \rightarrow M$ and $M \rightarrow H$ imply $A \rightarrow H$. It is called (a restricted form of) *resolution* or (a slightly generalized form of) *modus ponens*. As we see, it produces new knowledge ($A \rightarrow H$) by eliminating the intermediate *atom* M; thus, the rule may also be seen as implementing a form of transitive closure.

In order to program computers to search for chains of inferences we first need rules of inference such as the one just discussed in addition to the formal languages mentioned before. The combination of a formal language along with a set of rules is called a (formal or logical) *calculus*. If F is some *formula*, i.e., a propositional expression in the formal language under consideration, viz. FOL according to our convention, and G is a formula obtained from F by applying the rules of the calculus, then we write $F \vdash G$. Since $\vdash$ is dependent on the calculus, say c – which, as in the case of formal languages, can be chosen from a great variety of alternatives – we should rather write $\vdash_c$ but omit the index whenever the calculus is clear from the context. We want to design calculi in such a way that $F \vdash G$ holds if, and only if, (shortly, *iff*) $F \models G$, in which case we speak of *correct* (if) and *complete* (only if) calculi. From now on we assume that the rules are chosen such that the resulting calculi are always correct and complete.

We go one step further in the abstraction process described in this entire section. According to the well-known deduction theorem, it holds that $F \vdash G$ *iff* $\vdash F \rightarrow G$. By (iteratively) applying this theorem, we can always achieve that there are no formulas to the left of $\vdash$. Most deductive systems in use are therefore systems that test whether, for some given calculus and some given formula F, it holds that $\vdash F$ (establishing an affirmative proof), or dually, that $\vdash \neg F \equiv$ *false* signaling a contradiction (proof by contradiction).

Let us illustrate the result of this abstraction process by our apes–mammal–hair example. The example asserts that $A \rightarrow M$ and $M \rightarrow H$ implies $A \rightarrow H$. Using the logical symbol $\wedge$ instead of "and" the assertion reads $(A \rightarrow M) \wedge (M \rightarrow H) \vdash (A \rightarrow H)$. Applying the deduction theorem yields $\vdash (A \rightarrow M) \wedge (M \rightarrow H) \rightarrow (A \rightarrow H)$. The formula to the right of $\vdash$ is an example for what was briefly denoted by F above. It could be presented as input to a deductive system, which would easily verify that the formula is valid. If the system works by contradiction, the presented formula would rather be $\neg[(A \rightarrow M) \wedge (M \rightarrow H) \rightarrow (A \rightarrow H)]$.

The field in which such systems are developed, studied, and applied is called *automated deduction* (AD) or *automated reasoning* (AR). It has applications in any field where reasoning of any kind is relevant, which actually is most of science, or most of human endeavors for that matter. Since AD's first application was to mathematics, the field is also known as *automated theorem proving* (ATP).

Issues in Automated Deduction

There are numerous deductive systems that differ in the formal languages or the set of rules making up the underlying calculus. Further distinctions are the strategies used for the proof search, the data structures used for the representation of the formal constructs involved, the programming language used for the code, and so forth. Some of these issues are discussed in the present section.

Clausal Form, Resolution, Tableaux, and Connection Method

A very popular language is a version of the language of FOL, called *clause form*. It uses sets of clauses. A clause is a disjunction of literals. A literal is an atom or its negation. An atom is the simplest propositional construct such as *is_shining(The_sun)* or the A in the ape example, i.e., without involving any logical connections.

Any formula of FOL can be transformed into clause form by a straightforward algorithm based on well-known laws of propositional logic (and techniques discussed in the next section); the algorithm is not explained here in detail. Its application to our formula $\neg[(A \rightarrow M) \wedge (M \rightarrow H) \rightarrow (A \rightarrow H)]$ leads to the following set of four clauses: $\{(\neg A \vee M), (\neg M \vee H), A, \neg H\}$. The meaning of the comma separating the clauses in the set may be viewed as that of a conjunction $\wedge$.

A correct and complete calculus based on clause sets is given by (essentially) the single rule of *resolution*, which was mentioned already in section Syntactic inference. In clause form, it operates on two selected clauses in the given clause set; the two clauses must be of the form $1 \vee A \vee 2$ and $3 \vee \neg A \vee 4$, where the numbers stand for arbitrary, possibly empty clausal parts. The pair $\{A, \neg A\}$ is called complementary, or a *connection*. The result of a resolution step then is $1 \vee 2 \vee 3 \vee 4$, called the resolvent. For example, the resolvent of the first two clauses $\neg A \vee M$ and $\neg M \vee H$ in our example is $\neg A \vee H$. The resolvent (of each resolution step in the derivation) is added to the set of clauses. If the next resolution step selects this last resolvent and the clause A, then the newly added resolvent thereof is H. If this resolvent again is selected together with the clause $\neg H$ then the empty clause results as resolvent. The empty clause represents *false*. Its occurrence in this process is the criterion for success. Once this criterion is fulfilled it is guaranteed that the original formula, from which the deduction process started off, is indeed valid.

A deductive system based on resolution works as a saturation process. It takes the given clause set, and iteratively applies resolution until the empty clause is obtained. The advantage of this type of system is the simplicity of the data structures involved (clause sets) and of the operations performed at each iteration (resolution). The problem is that without additional constraints the system produces relevant and irrelevant intermediate results alike, and for hard problems often is drowned in up to billions of generated clauses. An irrelevant intermediate result in our earlier example might be that "apes have eyes" on the basis of the additional knowledge that "mammals have eyes" when our only interest is in the question about hairs. *Strategies* (like the so-called set-of-support strategy) and reduction rules to some extent help to reduce the amount of irrelevant clauses generated.

An alternative to this *saturation-based* approach are *goal-oriented* tableaux-based deductive systems. Human proofs, i.e., chains of inferences, may be regarded as treelike structures, or (closed) *tableaux*, which in each branch start out from an obviously valid fact. The basic data structure of tableaux-based systems is a (not necessarily closed) tableau. Starting with the tableau consisting of the (denial of the) given formula, the tableau is extended at each iteration with the goal of closing it. The data structure (tableau) and the operations (extension of tableau) are more complicated than those in the saturation-based systems, which complicates the system development. But the approach lends itself to a more goal-oriented fine-tuning of the search for the final proof and can therefore avoid some of the redundancy involved in resolution systems without extra efforts.

Systems based on the *connection method* may be regarded as a compactly coded form of tableaux systems. In this method, all information coded by a tableau is attached in an extremely condensed form to the parts of the given formula as structural information. Starting with an initial structure of this kind, only the structure is modified at each iteration until a certain final structure signaling validity of the formula is reached; in contrast to the previous kinds of methods, the formula itself is not changed at all in the process. As we saw already in the resolution method, connections (i.e., complementary pairs of literals) play a fundamental role in proofs. In the connection method they guide the proof search. In the final structure just mentioned, a so-called spanning set of connections for the formula is identified, which characterizes the validity of the formula. For our example, this structure consisting of three spanning connections is illustrated in the following picture.

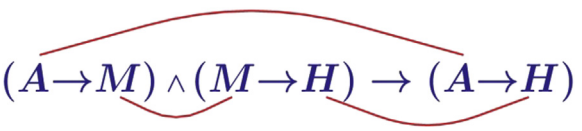

Research in AD pursues all three kinds of approaches, with the focus in each case being on eliminating redundancy. This can be achieved in various ways. One way consists of refined

calculi in terms of the rule sets used. For instance, the application of resolution may be restricted in many ways without sacrificing completeness (e.g., linear resolution). This also includes the application of reduction rules (e.g., subsumption). Another way to reduce redundancy is by applying various strategies that guide the rules application. A third is by performing several small steps as a single global step (e.g., hyperresolution), which again narrows the search space. A fourth involves global steps sanctioned by theories involved (such as equality theory, see section Variants, specializations and extensions). Finally, ever improved software engineering techniques and faster machines contribute to a continuous enhancement of system performance. On top of all, heuristics try to achieve success faster by sacrificing completeness of the method.

First-Order Features

For simplicity, the introduction of some AD methods in the previous subsection has glossed over a number of details. While speaking of FOL, no first-order features beyond the propositional ones were actually taken into account. For instance, the apes example actually talks about objects with certain properties so that its first-order formula would rather read $\forall x\ [(Ax \rightarrow Mx) \wedge (Mx \rightarrow Hx) \rightarrow (Ax \rightarrow Hx)]$. In order to save parentheses, the atoms like Ax are short for the functional form $A(x)$ reading "x is an ape". The statement is asserted for all objects x, formally expressed by the universal quantifier $\forall x$. In addition, FOL features an existential quantifier as in the sentence "any human has a mother", formally $\forall x \exists y\ (Hx \rightarrow Myx)$, whereby Hx is short for "x is human" and Myx abbreviates "y *is* mother of x".

As these examples illustrate, atoms in FOL have an internal structure. It consists of a predicate symbol like M and a certain number of arguments, two in the case of M. The arguments can be variables like x or constants like "*The_sun*", or shortly s. They may even involve functions like "the color of s", or shortly $c(s)$. Altogether we speak of terms serving as arguments.

The occurrence of quantifiers in the formulas to be deduced affects the transformation algorithm into clause form. By a technique called *skolemization* existential quantifiers are eliminated and then all-quantifiers ignored by default. As a result, a set of clauses looks as before except for the more complicated atoms involved. For instance, the apes example would now look like $\{(\neg Ax \vee Mx), (\neg Mx \vee Hx), Ax, \neg Hx\}$.

Connections like $\{A, \neg A\}$, because of the richer structure of atoms, may now become more complicated such as e.g., $\{A(x,f(z)), \neg A(b,y)\}$. Whether or not such a connection can be made complementary and thus will contribute to a proof is subject to a further technique called *unification*, which assigns values (or rather terms) to the variables involved. In the present example, x would be assigned the term b, and similarly y the term $f(z)$; in short, this substitution may be written as $\{x\backslash b, y\backslash f(z)\}$, resulting in the complementary connection $\{A(b,f(z)), \neg A(b,f(z))\}$.

The three types of methods introduced in the previous subsection engage unification in some way, which will not be described here in detail. Resolution, enhanced by unification, allows the elimination of intermediate literals (i.e., connections) even if they are not exactly identical but can be made identical by replacing variables in the atoms by terms as just described. For instance, the resolvent of the first two clauses in the apes example now is $\neg Ax \vee Hx$ and similarly for the two other steps. Except for the additional unification the proof proceeds as before, resulting in the empty clause. In general, the resolution method requires an additional rule of inference, viz. factoring, by which two different literals in a clause by way of unification are factored into a single one.

The proof of our running example in the connection method is unchanged as well:

$$(Ax \rightarrow Mx) \wedge (Mx \rightarrow Hx) \rightarrow (Ax \rightarrow Hx)$$

Note that each connection corresponds to a respective step in the resolution derivation described above. For comparison, recall that resolution in each step adds a new clause to the clause set and thus with it new connections. In contrast, the connection method is concerned exclusively with the connections present in the formula to be proved, although these connections by the nature of FOL may be required in different instances. This allows a more focused search, guided by the structure of the formula in a connection-oriented way. In addition, in the connection method no transformation into clause form is required (Otten, 2011). This is an advantage since this transformation in general introduces redundancy thus increasing the search space. On the contrary, the given formula may even be reduced by factoring, making it more nonnormal, possibly resulting in (much) shorter proofs and less proof search (Bibel, 2006).

In section Syntactic inference we have made the distinction between affirmative proofs and proofs by contradiction. Technically, there is no argument for one or the other form. Any method can be formalized either way without affecting performance. In addition, there is an "affirmative" clause form defined similarly to the one introduced in section Clausal form, resolution, tableaux, and connection method. Resolution by historical accident is usually presented as a proof-by-contradiction method, while the connection method is usually affirmative. The tableaux method is used both ways.

Variants, Specializations, and Extensions

FOL is a rather expressive formalism in which, for instance, all of mathematics as well as more or less all of science can be represented. Typically, deductive tasks arise within a given theory. For instance, if the task F involves an equation then the theory has to contain the equality axioms, say $T_=$. These are, for instance, the axioms of reflexivity and substitutivity (for functions and predicates). The formula to be proved thus has the form $T_= \rightarrow F$, and it is straightforward for deductive systems to attempt a proof for such a formula.

In comparison with the human way to solve equations, however, this approach involving the equality axioms seems pretty awkward. Rather, humans proceed in such cases by substituting terms by equal terms until the equality becomes an identity. This is the motivation for treating equality in rather special ways also in deductive systems. For instance, resolution has a special rule called paramodulation by which equal terms may be substituted without any reference to axioms. Similar

techniques have been developed for the tableaux and the connection method. In fact, there is an extensive research area known as "term rewriting", in which efficient algorithms for treating equality theories have been developed. Appropriate techniques are also available for other special mathematical tools (like mathematical induction) and theories.

The programming language PROLOG (Clocksin and Mellish, 2003) is a restricted FOL language enriched by control features. It restricts the formulas to so-called Horn clauses, which in the contradiction mode means that there is at most one negative literal in each clause. A PROLOG clause is written $A_0 := A_1, A_2, \ldots, A_m$, with the left side of := called the head and the right the body of the clause. For instance, the clause $Ax \rightarrow Mx$ from the apes example here reads $Mx := Ax$, which by comparison of the two versions shows that := can be interpreted as an implication from right to left. In the case $m = 0$ (i.e., no body) the sign := is deleted. A PROLOG program is a set of PROLOG clauses. A clause with no negative literal is written ?– $A_1, A_2, \ldots, A_m$. It is called a query. Except for these superficial syntactic differences a PROLOG system behaves much like a deductive system based on the methods discussed before. One exception is the possibility of allowing negated atoms *not* A_{m+1} in the clause body, which are treated with a special technique called "negation as failure". In addition, with respect to applications in AI the semantics of a PROLOG program differs from that of a corresponding classical formula. The control features allow the programmer to overrule the system's built-in control.

In section Formal languages we already mentioned description logic (DL) as a first-order formalism in disguise. More precisely, DL is actually a family of formal knowledge representation languages which, due to certain restrictions, feature more efficient deductive proof procedures than full FOL (mostly even decision procedures). It models concepts, roles and individuals, and their relationships. For instance, *woman* = *adult* $\sqcap$ *gender*:*female* might be a definitional statement of the concept "woman" in DL. In FOL such a concept is represented as a predicate (like in *Woman*(*mary*)). This syntactic difference is irrelevant from an efficiency point of view; the decisive difference lies in the restrictions imposed on the operations of the language, which lead to less expressiveness of the language but to a higher efficiency of the resulting deductive systems.

Another restricted form of FOL is the formalism of syllogisms, originally invented by Aristotle. It features judgments of the form *all apes are mammals*, or *AaM*, *at least one ape is old*, or *AiO*, and so forth. Along with judgments of these kinds there are inference rules. For instance, the apes–mammals–hair example is an instance of what is known as the valid inference Barbara: *AaM*, *MaH* → *AaH*. The traditional names of these rules are chosen for mnemotechnical reasons: the three *a*'s in the word "Barbara" remind of the three *a*'s in the rule. In Brüning (1996) a strict formalism based on syllogisms has been developed but not yet been realized in a running system.

So far we have discussed restricted forms of FOL. Conversely, there are numerous extensions of FOL, too many to discuss here in any detail. We just mention the great variety of modal and temporal logics that feature additional logical operators (e.g., the necessity operator $\Box$ or the possibility operator $\Diamond$) beyond those of FOL along with corresponding rules of inference. Higher order logic (HOL) is the most powerful extension beyond FOL. It allows the quantification over functions and predicates according to familiar practice, e.g., in mathematics. A formalism for representing HOL is type theory. For all these extensions there are deductive formalisms and systems, analogous to those of FOL, yet more complicated.

All formalisms discussed above test the validity of formulas, i.e., their truth under *any* interpretation. Often it is already sufficient to test the given formula for its satisfiability, i.e., its truth under *one* particular interpretation rather than under all interpretations. This semantic approach has found many applications especially for the restricted language of propositional logic. The satisfiability of propositional formulas is known as the problem class SAT. For instance, the apes–mammal–hair example formula $(A \rightarrow M) \wedge (M \rightarrow H) \rightarrow (A \rightarrow H)$ will turn out true if all propositional variables A, M, H are assigned the value *true* (and in fact under any assignment due to the formula's validity).

There are extremely powerful systems for SAT problems and variants of these, solving problems that may involve millions of propositional variables. One of the popular techniques uses binary decision diagrams, or BDDs (Bryant, 1992). This formalism treats propositional formulas, restricted to the operators conjunction, disjunction, and negation, in a functional way. Here conjunction is written like a product, disjunction as sum, and negation with a bar, e.g., $f(A,M,H) = [A\bar{M} + M\bar{H} + \bar{A} + H] + \bar{A} + H]$ for the formula after elimination of → according to the rules of propositional logic. The efficiency of BDDs is achieved basically by a reduction in the formula by factoring identical subformulas and by choosing a special order of the variables in the formula.

Answer set programming (Gelfond, 2008), or ASP, is a form of declarative programming similar to PROLOG, but with a different semantics and solution technique. The language uses PROLOG clauses with a number of additional features. The semantics is based on so-called stable models. The solution technique includes a grounder and a solver. The grounder in essence transforms the problem into a propositional formula; the solver is something like a SAT solver. Among the additional features are several language extensions like integrity constraints, choice rules, cardinality and weight constraints, and so forth. These language features render ASP a rather expressive representational formalism. ASP is extremely powerful for solving difficult search problems. Due to its declarative nature, it offers a great advantage in comparison to programming such problems in a conventional programming language.

Complexity, Systems, and Performance

From a theoretical point of view, AD is one of the hard problems in computer science. SAT solving, i.e., semantic propositional reasoning, is already NP-complete, which (for the state of art) essentially means no procedure can solve all propositional formulas efficiently, even if the formulas are comparably small. First-order reasoning is even more difficult, since FOL is undecidable: there is no procedure that can decide for all formulas if they are valid or not.

Nevertheless, research in AD has produced a variety of reasoning techniques that can solve most reasoning problems very efficiently. Propositional reasoning techniques like conflict analysis, clause learning, backjumping, adaptive branching, random restarts, and problem reductions have increased the

performance of SAT solvers by many orders of magnitude, which makes it possible to apply them to formulas with millions of clauses and variables. In first-order reasoning, indexing techniques, efficient data structures, and search heuristics have led to efficient proof systems capable of dealing with complex reasoning tasks.

In the past years, benchmark collections have made it possible to evaluate the performance of AD systems, while annual system competitions in SAT solving (The International SAT Competitions Web page) and first-order theorem proving (The CADE ATP System) have resulted in significant improvements. Currently, the best SAT solvers are *3S*, *ppfolio*, and the ASP system *clasp*, while systems like *Vampire*, *Waldmeister*, *Spass*, *E*, and *leanCoP* are the most successful first-order theorem provers. However, the winners of these competitions keep changing, since the development of reasoning techniques makes previously unknown systems even more efficient than the winners of the past years.

Artificial Neural Systems

Human reasoning takes place in brains. The mechanisms occurring in brains are modeled by connectionist systems. So the question naturally arises whether reasoning could be modeled by a connectionist, also termed artificial neural, system. Since we do not yet know how the brain realizes reasoning, such a system could be a very first step toward understanding the brain's processes in reasoning.

Several neural systems have been developed that perform deductive (and abductive and other kinds of) inferences, mostly for propositional logic. The so-called CORE method ("Connectionist mOdel generation using Recurrent networks with feedforward core") achieves such inferences semantically by generating or approximating a (possibly least) model of the premises under which a conclusion can be drawn. In the propositional case, the model can iteratively be computed by a three-layer feed-forward core (i.e., connectionist network) until a stable state is reached. Thereby, the output nodes of the network are recurrently connected back to the input nodes to allow for the iterations. With a number of additional ideas this method has been generalized to the first-order case (Baader et al., 2008). Here, the model can only be approximated, however, with arbitrary precision.

The CORE method bridges the gap between the discrete world of logic programs and the continuous world of neural networks. It integrates distinct functions such as transforming programs (as background knowledge) into networks, learning and optimizing networks, generating or approximating models from networks, drawing inferences by way of the models, and extracting a logic program from a learned or optimized network.

Applications

As pointed out in the introduction, reasoning is a fundamental human capability. Therefore, it pervades many of our activities. In consequence, reasoning systems have numerous and widely varied applications. Some of these are briefly discussed in the present section.

Mathematical Theorem Proving

Originally, mathematics has been the primary field of application for AD. Deductive systems have been applied to virtually all subdisciplines in mathematics, and tens of thousands of theorems have been proved automatically. Automated theorem provers have found new proofs for known theorems as well as proofs of unsolved mathematical problems and conjectures that had never been proved by humans before.

For instance, in 1996, a system called Otter/EQP solved the so-called Robbins problem, which had been open for more than 60 years. The machine-generated proof consists of only 13 equations and can therefore be checked easily by human mathematicians. Finding this proof, however, took more than 7 days of intense machine computation, as several thousand equations had to be generated and combined in order to construct the proof (Kolata, 1996, McCune, 1997).

Another famous example is the proof of the four color theorem, which has been conjectured in 1852 and was the first theorem to be proved by a computer in 1976 (Appel and Haken, 1977). To prove the theorem, a special-purpose computer program had to analyze 1936 different maps. This resulted in a proof that is infeasible for a human to check by hand and is not accepted as a convincing proof by some mathematicians. In 2005, the theorem could be proved with the help of a general purpose theorem prover Coq (Gonthier, 2008). Although even this proof cannot be checked without the help of a computer, it was sufficient to dispel remaining doubt about the validity of the four color theorem.

Hardware and Software Verification

One of the most important applications of AD techniques today is the verification of software and hardware. In the early 1990s, major flaws in hardware design such as the Pentium FDIV bug resulted in hundreds of millions of dollars of damages and caused the industry to look for tools that could validate the design of complex hardware circuits. At the same time, new techniques for eliminating redundancy in propositional reasoning led to significant efficiency improvements, which permitted the use of model checkers and SAT solvers in large-scale applications involving millions of components (Clarke et al., 1999). As a result, model checkers were adopted as the standard technique to verify the design of computer chips before they go into production.

While the verification of hardware involves the analysis of a huge number of possible states, the problem is finite and can therefore be represented in propositional logic. In contrast, software can assume infinitely many different states, which makes software verification much more complex. For specific types of software such as compilers, state-based systems like communication and security protocols, and some commercial applications, it has been demonstrated that automated proof systems and software model checkers are capable of constructing proofs that guarantee the software's validity. But in general these systems do not scale well enough to deal with the complex formulas that are needed for the description of real-world software.

Software and Hardware Synthesis

An alternative approach to software validity is based on the idea that programs and proofs of their correctness should be developed simultaneously instead of verifying software after it has been created. Program synthesis aims at using AD to develop "correct-by-design" software from descriptive specifications. Ideally, a programmer would focus on designing a precise formal specification of a given problem and then call a deductive system to generate the actual implementation. The programming language PROLOG basically works as such a deductive system. However, as a programming language, it also contains nondeclarative control mechanisms and is only a first step toward "programming in logic".

Systems that synthesize programs from descriptive specifications involve strong deductive components to ensure the correctness of the generated program. The systems KIDS and Specware/Isabelle have been used in the automated construction of commercial software and scheduling algorithms that are not only correct by design but also significantly more efficient than hand-crafted versions (Smith, 1991, Smith and Parra, 1993). The higher order theorem prover Nuprl has been used for the semiautomated development and optimization of distributed software with more than 50 000 lines of code (Liu et al., 1999). A key to the success of these systems was the fact that the deductive components were guided by strategies that incorporate knowledge about algorithmic structures and efficiency optimizations.

In hardware design, the standard procedure consists of describing the behavior of circuits in a hardware description language at logic level followed by extensive verification and logic synthesis, a process of turning the circuit descriptions into a design implementation in terms of logic gates. This procedure consumes significant amounts of time and needs a lot of effort. While the idea of synthesizing hardware circuits from logical specifications has been proposed several decades ago, it was deemed impractical because of its high computational complexity. Recent developments (Bloem, 2007), however, have made it applicable to realistic examples, and it is expected that hardware synthesis based on AD will eventually be used in industrial applications.

Knowledge Systems

Reasoning generates new knowledge from given knowledge. Hence, it is an essential feature in any system that incorporates explicitly represented knowledge. Such a system is variably called knowledge system, knowledge-based system, expert system, semantic web, ontologies, and so forth. In the CYC system an entire encyclopedia of knowledge is encoded in this way. Numerous special systems encode knowledge of specific disciplines, allow the manipulation of such knowledge in a deductive way, or can answer questions in complex scenarios (Harmelen et al., 2007). For instance, in 2011, the system WATSON outperformed the two best human players in the US quiz show Jeopardy. Similarly, the Apple iPhone 4S features a system Siri, which understands and answers queries in spoken natural language, not least because of deductive mechanisms. These are just three examples out of tens of thousands of such systems. In the emerging technology of real-world awareness, coping with the masses of information will only be possible if inductive systems will extract knowledge from the data.

Philosophy and Psychology of Reasoning

As seen in the previous section, deductive systems are becoming a part of daily life. We want them to behave rationally in the human sense. Hence, we need to understand human rationality and reasoning in the first place, a goal pursued for more than two thousand years first in philosophy, then in psychology, and more recently in intellectics. In this section, we discuss philosophical and psychological aspects of reasoning.

Can Reasoning Be Mechanized?

Logic and the study of deductive reasoning was a primary topic in philosophy for more than two thousand years (for an exposition of this history see Kneale and Kneale, 1984). Since the nineteenth century, it transformed more and more into a mathematical subject, whereas in the past 60 years it has become a major topic in AI. Therefore, today the major contributions to the subject do not originate in philosophy anymore – with one exception: many philosophers have tried in thousands of papers to argue in favor of the thesis that the mechanization of reasoning has strict limitations and will never be able to reach the human level of performance. Herbert Dreyfuβ, John R. Lucas, Roger Penrose, and John Rogers Searle are some of the philosophical spokesmen in this debate. The debate is a fruitless one, and the arguments in favor of the thesis have proved faulty. One of the papers that have defeated the thesis is LaForte (1998). Although we indeed do not know in detail how the brain realizes our thinking, the computationalism underlying intellectics at present is the best available scientific approach.

Psychology of Reasoning

In many respects, the best available deductive reasoning system is still the human mind. Psychological studies and thus our understanding of (the phenomenological part of) human reasoning can therefore enhance the design of artificial deductive systems (Evans, 1970). The key idea is that humans – in contrast to machines – construct internal representations of the state of affairs described by a logical argument in order to understand it. These "mental models" are similar to the models that people use to perceive the world, but rudimentary and incomplete. Experimental evidence seems to indicate that human deduction depends less on laws of inference but rather on a search for alternative models of the premises that could refute a given conclusion and that a conclusion is accepted as long as there is no refutation for it.

Most humans, for instance, appear to read an implication as bi-implication, or equivalence, and draw conclusions like the following.

- *If John is in New York then Cathy is in New York too.*
- *John is not in New York.*
- *Therefore Cathy isn't in New York either.*

which obviously is not a correct deduction from the point of view of classical logic like FOL.

Negation is even more difficult. The sentence

- *John wouldn't have been in New York if Cathy hadn't been there.*

is logically equivalent to the first one above but is perceived differently.

The theory of mental models aims at capturing the difference between human and formal reasoning. In this theory, the process of deductions depends on a semantic procedure that constructs an internal representation of what has been stated, tries to draw possible conclusions by formulating a more elaborate model based on that representation, and attempts to find a countermodel for that conclusion. If no such countermodel is found, the conclusion is considered valid and the process continues by proposing further conclusions and trying to refute them.

On this basis, cognitive scientists and psychologists have formulated a comprehensive theory of human inference that explains propositional, quantificational, and relational reasoning (Johnson-Laird and Byrne, 1991, Johnson-Laird, 1999). This theory has been used as a foundation for implementing reasoners that appear to be a more appropriate model of the human intellect. There are also theories that view human deduction as a process based on factual knowledge (Anderson, 1993, Newell, 1990) or on formal rules of inference (Rips, 1994) similar to those used in AD. Evidence from a wide range of studies seems to indicate that the theory of mental models provides the most appropriate representation of the mechanisms underlying human deduction. However, in contrast to other areas in psychology such as image processing (Kosslyn, 1995) or "system 1 mental operations" (Kahneman, 2011), reasoning (a system 2 operation) is still poorly understood.

Fitting, 1990 is an excellent textbook on propositional and first-order logic. Bibel, 1993 is a special textbook on all aspects of deductive methods and systems. Bibel and Schmitt, 1998 is a reader covering all modern deductive techniques, a number of very successful systems, and several important applications. Gabbay et al., 1991 is a handbook that includes deductive reasoning systems and other topics. Robinson and Voronkov, 2001 is a handbook covering just deductive methods, techniques, and systems. Plaisted, 2009 is an encyclopedia article similar to the present one. Ongoing research is reported in the *Journal of Automated Reasoning* (Kluwer Academic Publishers) and in the proceedings volumes (published by Springer Verlag) of the International Conferences on Automated Deduction as well as in other journals and conferences.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Knowledge Representation; Nonstandard Reasoning; Organizational Decision Making; Problem Solving and Reasoning: Case-based.

Bibliography

Anderson, J.R., 1993. Rules of Mind. Erlbaum, Hillsdale, NJ.

Appel, K., Haken, W., 1977. Solution of the four color map problem. Scientific American 237 (4), 108–121.

Baader, F., Hitzler, P., Hölldobler, S., 2008. Connectionist model generation: a first-order approach. Neurocomputing 71, 2420–2432.

Bader, S., et al., 2003. The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press, Cambridge, UK.

Bibel, W., 1993. Deduction: Automated Logic. Academic Press, London.

Bibel, W., 2006. Research perspectives for logic and deduction. In: Reasoning, Action, and Interaction in AI Theories and Systems, LNAI, vol. 4155. Springer, pp. 25–43.

Bibel, W., Schmitt, P.H., 1998. Automated Deduction – A Basis for Applications. In: Applied Logic Series, vols. I-III. Kluwer, Dordrecht, NL.

Bloem, R., et al., 2007. Automatic hardware synthesis from specifications: a case study. In: Conference on Design, Automation and Test in Europe, pp. 1188–1193.

Brüning, W., 1996. Grundlagen der Strengen Logik. Königshausen und Neumann, Wurzburg.

Bryant, R., 1992. Symbolic boolean manipulation with ordered binary-decision diagrams. ACM Computing Surveys 24 (3), 293–318.

Clarke, E.M., Grumberg, O., Peled, D., 1999. Model Checking. MIT Press, Cambridge, MA.

Clocksin, W., Mellish, C., 2003. Programming in Prolog, fifth ed. Springer, Heidelberg.

Evans, J.St.B.T., 1970. The Psychology of Deductive Reasoning. Routledge & Kegan Paul, Reading, UK.

Fitting, M., 1990. First-Order Logic and Automated Theorem Proving. Springer, Heidelberg.

Gabbay, D., Hogger, C., Robinson, J.A. (Eds.), 1991. Handbook of Logic in Artificial Intelligence and Logic Programming. Oxford University Press, Oxford, UK.

Gelfond, M., 2008. Answer sets. In: Handbook of Knowledge Representation. Elsevier, Amsterdam, pp. 285–316.

Gonthier, G., 2008. Formal proof – the four-color theorem. Notices of the American Mathematical Society 55 (11), 1382–1393.

van Harmelen, F., Lifschitz, V., Porter, B. (Eds.), 2007. Handbook of Knowledge Representation. Elsevier, Amsterdam.

Johnson-Laird, P.N., 1999. Deductive reasoning. Annual Reviews Psychology 50, 109–135.

Johnson-Laird, P.N., Byrne, R.M.J., 1991. Deduction. Lawrence Earlbaum Associates Ltd, Hillsdale, NJ.

Kahneman, D., 2011. Thinking, Fast and Slow. Farrar, Straus and Giroux, New York.

Kneale, W., Kneale, M., 1984. The Development of Logic. Clarendon Press, Oxford, UK.

Kolata, G., December 10, 1996. Computer Math Proof Shows Reasoning Power. The New York Times. http://www.nytimes.com/library/cyber/week/1210math.html.

Kosslyn, S.M., 1995. Image and Brain – The Resolution of the Imagery Debate. MIT Press, Cambridge, MA, pp. 327–404.

LaForte, G., Hayes, P.J., Ford, K.M., 1998. Why Gödel's theorem cannot refute computationalism. Artificial Intelligence 104, 265–286.

Liu, X., et al., 1999. Building reliable, high-performance communication systems from components. Operating Systems Review 33, 80–92.

McCune, W., 1997. Solution of the Robbins problem. Journal of Automated Reasoning 19 (3), 263–276.

Newell, A., 1990. Unified Theories of Cognition. Harvard University Press, Cambridge, MA.

Otten, J., 2011. A non-clausal connection calculus. In: TABLEAUX 2011, LNAI 6793. Springer, pp. 226–241.

Plaisted, D., 2009. Theorem proving. In: Encyclopedia of Computer Science and Engineering. John Wiley and Sons, New York, pp. 223–244.

Rips, L.J., 1994. The Psychology of Proof. MIT Press, Cambridge, MA.

Robinson, J.A., Voronkov, A. (Eds.), 2001. Handbook of Automated Reasoning, vols. 2. Elsevier Science and MIT Press, Amsterdam, NL.

Smith, D.R., 1991. KIDS – a knowledge-based software development system. In: Automating Software Design. AAAI Press/The MIT Press, pp. 483–514.

Smith, D.R., Parra, E.A., 1993. Transformational approach to transportation scheduling. 8th Knowledge-Based Software Engineering Conference, 60–68.

The CADE ATP System Competition. (http://www.cs.miami.edu/~tptp/CASC)

The international SAT Competitions web page. (http://www.satcompetition.org)

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF
JAMES D. WRIGHT
University of Central Florida, Orlando, FL, USA

VOLUME 6

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser–Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 6

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 6 *xliii*

VOLUME 6

Defense and Security Economics 1
Rémy Herrera

Defiant Behavior During Adolescence across Cultures 7
Leslie Berntsen and Laura A Baker

Defiant Behavior During Infancy and Early Childhood 12
Sheryl L Olson and Adam S Grabell

Definition of the situation: History of the Concept 19
Rainer Egloff

Degrowth 24
Giorgos Kallis, Federico Demaria, and Giacomo D'Alisa

Dehumanization 31
Chiara Volpato and Luca Andrighetto

Deindividuation, Psychology of 38
Tom Postmes and Felicity M Turner

Deinstitutionalisation in Low and Middle Income Countries 42
Julian Leff

Deinstitutionalization, Psychiatry of 47
Julian Leff

Deixis 52
Jürgen Bohnemeyer

Delegation of Power: Agency Theory 58
Arthur Lupia

Deliberation and Democracy 61
Giovanni Boniolo and Giuseppe Schiavone

Delinquency, Sociology of 68
Gary F Jensen

Dementia: Psychiatric Aspects 75
Friedel M Reischies

Democracy, History of 80
James J Sheehan

Democracy: Normative Theory 85
Thomas D Christiano

Democratic Theory 90
Frank Cunningham

Democratic Transitions 97
Gerardo L Munck

Demographic Impact of Disasters 101
Elizabeth Frankenberg, Maria M Laurito, and Duncan Thomas

Demographic Measurement: General Issues and Measures of Fertility 109
Máire Ní Bhrolcháin

Demographic Measurement: Nuptiality, Mortality, Migration, and Growth 116
Máire Ní Bhrolcháin

Demographic Models 123
Yu Xie

Demographic Techniques: Data Adjustment and Correction 126
Heather Booth and Patrick Gerland

Demographic Techniques: Family Reconstitution 138
Cameron D Campbell

Demographic Techniques: Indirect Estimation 143
Tom A Moultrie

Demographic Techniques: Inverse Projection 149
Elisabetta Barbi

Demographic Techniques: LEXIS Diagram 154
Gérard Calot

Demographic Techniques: Small-area Estimates and Projections 162
Stefan Rayer

Demography, Early History of 170
Simon Szreter

Demography: History Since 1900 176
Dennis Hodgson

Dependency in Language 182
Igor Mel'čuk

Dependency Theory 196
Christopher Chase-Dunn

Depression 199
Neel Halder

Depression, Pessimism, and Health 207
Charles S Carver and Emily G Lattie

Deprivation: Relative 214
Mohamed Cherkaoui

Desegregation 218
John Dixon, Kevin Durrheim, and Manuela Thomae

Deserts, Philosophical Concept of 229
John Kleinig

Design and Form: Organizational 235
Naomi Stanford

Determinism: Social and Economic 241
Warren Schmaus

Deterrence 247
Leon V Sigal

Deterrence: Sanction Perceptions 250
Robert Apel and Daniel S Nagin

Deterrence Theory: Crime 255
Mark C Stafford

Deterrent Effect of Police and Prisons 260
Robert Apel and Daniel S Nagin

Deutsch, Karl W (1912–92) 266
Mattei Dogan

Development and Employment: A Mutual Relationship 269
Stefano Scarpetta and Gaëlle Pierre

Development and the State 276
Michelle Williams

Development and Urbanization 282
Diane E Davis and Alexander M Keating

Development Geographies 290
Nina Laurie

Development: Organizational 296
Georg Schreyögg

Development: Social-Anthropological Aspects 301
A Peter Castro and David W Brokensha

Development Studies 307
Su-ming Khoo

Development Theory in Geography 314
Jim Glassman

Developmental Behavioral Genetics and Education 320
Frank M Spinath, Birgit Spinath, and Peter Borkenau

Developmental Research across Cultures and Nations: Challenges, Biases, and Cautions 326
Bettina Lamm

Dewey, John (1859–1952) 332
Helmut Schreier

Diabetes: Psychosocial Aspects 337
Daniela Zahn and Thomas Kubiak

Diagnostic Processes in Clinical Psychology and Psychotherapy 342
Amy Wenzel

Dialectology 348
Jack K Chambers

Diaspora 353
Robin Cohen

Dichotic Listening and Language: Overview 357
Kenneth Hugdahl

Dictatorship 368
A James Gregor

Dictatorship in History 371
Emilio Gentile

Differential Aging 378
Susan K Whitbourne

Differential Diagnosis in Psychiatry 384
Umberto Volpe

Differentiation: Social 391
Uwe Schimank

Diffusion and Random Walk Processes 395
Philip L Smith and Roger Ratcliff

Diffusion: Anthropological Aspects 402
Ronald Stade

Diffusion: Geographical Aspects 406
Lawrence A Brown

Diffusion in Sociological Analysis 411
Alberto Palloni

Diglossia 417
Irene Theodoropoulou

Dilthey, Wilhelm (1833–1911) 424
Timothy D Mooney

Dimensional Comparisons and Their Consequences for Self-Concept, Motivation, and Emotion 430
Jens Möller, Friederike Helm, Hanno Müller-Kalthoff, Nicole Nagy, and Herb W Marsh

Dimensionality of Tests: Methodology 437
John Hattie

Direct-to-Consumer Personal Genetic Testing 440
Sandra Soo-Jin Lee

Disabilism 446
Elizabeth Lightfoot

Disability and Disabling Practices 452
Tania Jivraj and Esther Ignagni

Disability, Geography of 458
Vera Chouinard

Disability Movement 462
Debbie Jolly

Disability: Social and Psychological Aspects 467
Jan D Reinhardt and Marcel WM Post

Disability: Sociological Aspects 472
Nancy G Kutner

Disasters in Social, Cultural and Political Context 478
Julie Drolet

Discipline Formation in the Social Sciences 485
Bjorn Wittrock

Discipline-Building in the Social Sciences: Collective Memory, Biography and Autobiography 491
Charles Crothers

Discourse and Indentity 498
Francisco Yus

Discourse, Anthropology of 503
Charles L Briggs

Discourse Studies 510
Johannes Angermuller

Discovery: Legal Aspects 516
Steven Baicker-McKee

Discrete-State Models of Information Processing 523
William H Batchelder

Discrimination and the Law 530
Deborah C Malamud

Discrimination, Economics of 536
Ronald Oaxaca

Discrimination: Racial 542
Thomas F Pettigrew and Marylee C Taylor

Disgust, Psychology of 546
Paul Rozin

Disposable Youth in America in the Age of Neoliberalism 550
Henry A Giroux

Dispute Resolutions in Economics 557
Henry S Farber

Disputes, Social Construction and Transformation of 560
Lynn Mather and Barbara Yngvesson

Distance Education 567
Natalie B Milman

Distributions, Statistical: Approximations 571
John P Lehoczky

Distributions, Statistical: Special and Continuous 575
John P Lehoczky

Distributions, Statistical: Special and Discrete 580
Cecil D Kemp

Diversification, Relatedness, and the New Logic of Co-Creation 585
David J Teece

Diversity and Disagreement in Ethics: Philosophical Implications 589
Michele M Moody-Adams

Diversity in organizations 595
Jorge A Gonzalez and Azadeh Zamanian

Division of Labor: History of the Concept 601
Richard Sturn

Divorce, Sociology of 606
Ellyn A Steidl and Kelly Raley

Doctoral Education 610
Lynn McAlpine

Doctor–Patient Interaction in the West: Psychosocial Aspects 613
Joshua D Meadors and Carolyn B Murray

Documentary and Ethnographic Film 620
Kizhavana P Jayasankar and Anjali Monteiro

Domestic Violence: Legal Perspectives 625
Yvana LBH Mols

Domestic Violence: Sociological Perspectives 633
Rebecca E Dobash and Russell P Dobash

Dominant Designs 637
Aleksios Gotsopoulos

Dowries and Other Marital Transfers 642
Siwan Anderson

Dramaturgical Analysis 646
Robert D Benford and A Paul Hare

Dreaming, Neural Basis of 651
Lampros Perogamvros and Sophie Schwartz

Drinking, Anthropology of 658
Thomas M Wilson

Drucker, Peter (1909–2005) 663
Winfried W Weber

Drug Control: International Aspects 669
Robin Room

Drugs, Decriminalization of: Cultural Concerns 674
Stephen J Morse

Drugs: Illicit Use and Prevention 679
Wesley G Jennings and Jennifer M Reingle

Drug Use and Abuse: Psychosocial Aspects 685
Catherine Darker and Lucy Whiston

Du Bois, William Edward Burghardt (1868–1963) 690
Paul Gilroy

Dubos, René Jules (1901–82) 693
Jill Cooper

Due Process in Law 696
Samuel Issacharoff

Durkheim, Emile (1858–1917) 700
Steven Lukes

Durkheimians, The 706
Marcel Fournier

Dynamic Decision Making 709
Jared M Hotaling, Pegah Fakhari, and Jerome R Busemeyer

Dynamic Mapping in Geography 715
Doris Dransch and Andrea Unger

Dysexecutive Amnesia 718
Sara B Festini and Patricia A Reuter-Lorenz

Dysexecutive Syndromes 725
Brenda Hanna-Pladdy

Dyslexia, Developmental 728
Usha Goswami

Dyslexias, Acquired and Agraphia 732
Kaylie T Allen and Harry A Whitaker

Early Education and Care in the World Regions 739
Emma Pearson

Early Emotional Development and Cultural Variability 746
Fantasy T Lozada and Amy G Halberstadt

Early Life Influences on Health and Mortality in Adulthood 752
Ken R Smith and Heidi A Hanson

Early Memory Development and Cultural Variations 759
Monika Knopf, Frauke Graf, and Thorsten Kolling

Earnings Inequality 765
Andreas Peichl and Nico Pestel

East Africa: Sociocultural Aspects 773
William Arens

East Asian Legal Traditions 778
Pär Cassel

East Asian Studies: Culture 784
Emiko Ohnuki-Tierney

East Asian Studies: Economics 788
Vivek Chibber and Doug Guthrie

East Asian Studies: Gender 793
Tani E Barlow

East Asian Studies: Politics 798
Haruhiro Fukui

Eastern European Studies: Culture 805
Michael D Kennedy

Eastern European Studies: History 810
Tomasz Blusiewicz

Eastern European Studies: Society 816
Julia Szalai and Jeffrey Hass

Eating Disorders 822
Rebecca J Lester

Eco-cities and Sustainable Urbanism 829
Simon Joss

Ecocultural Theories of Development 838
A Bame Nsamenang

Ecofeminism 845
Susan Buckingham

Ecological Economics 851
Joan Martinez-Alier

Ecological Fallacy, Statistics of 865
Glenn Firebaugh

Ecological Inference 868
David A Freedman

Ecological Social Work: Shifting Paradigms in Environmental Practice 871
Fred H Besthorn

Ecology and Health 878
Margaret Schneider and Daniel Stokols

Ecology of Aging 884
Hans-Werner Wahl

Ecology: Organizations 890
Joel AC Baum

E-Communications Platforms and E-Learning 895
Christopher R Wolfe and Elizabeth M Cedillos

CONTRIBUTORS TO VOLUME 6

Kaylie T. Allen
Central Michigan University, Mt. Pleasant, MI, USA

Siwan Anderson
Vancouver School of Economics at University of British Columbia, Vancouver, BC, Canada

Luca Andrighetto
University of Genova, Genova, Italy

Johannes Angermuller
Centre for Applied Linguistics, The University of Warwick, Coventry, UK; and Centre d'Étude des Mouvements Sociaux (CEMS), École des Hautes Études en Sciences Sociales (EHESS), Paris, France

Robert Apel
Rutgers University, Newark, NJ, USA

William Arens
Stony Brook University, Stony Brook, NY, USA

Steven Baicker-McKee
Duquesne University School of Law, Pittsburgh, PA, USA

Laura A. Baker
University of Southern California, Los Angeles, CA, USA

Elisabetta Barbi
Sapienza University of Rome, Rome, Italy

Tani E. Barlow
University of Washington, Seattle, WA, USA

William H. Batchelder
University of California, Irvine, CA, USA

Joel A.C. Baum
University of Toronto, Toronto, ON, Canada

Robert D. Benford
University of South Florida, Tampa, FL, USA

Leslie Berntsen
University of Southern California, Los Angeles, CA, USA

Fred H. Besthorn
School of Social Work, Wichita State University, Wichita, KS, USA

Tomasz Blusiewicz
Harvard University, Cambridge, MA, USA

Jürgen Bohnemeyer
University at Buffalo, The State University of New York, Buffalo, NY, USA

Giovanni Boniolo
University of Milano, Milano, Italy; and Istituto Europeo di Oncologia (IEO), Milano, Italy

Heather Booth
Australian Demographic and Social Research Institute, The Australian National University, Canberra, Australia

Peter Borkenau
Martin-Luther-University Halle-Wittenberg, Germany

Charles L. Briggs
University of California Berkeley, Berkeley, CA, USA

David W. Brokensha
University of California, Santa Barbara, CA, USA

Lawrence A. Brown†
The Ohio State University, Columbus, OH, USA

Susan Buckingham
Centre for Human Geography, Brunel University, Uxbridge, London, UK

Jerome R. Busemeyer
Indiana University, Bloomington, IN, USA

Gérard Calot†
Observatoire Démographique, Européen, France

Cameron D. Campbell
Hong Kong University of Science and Technology, Kowloon, Hong Kong

† Deceased.

Charles S. Carver
University of Miami, Coral Gables, FL, USA

Pär Cassel
University of Michigan, Ann Arbor, MI, USA

Elizabeth M. Cedillos
Miami University, Oxford, OH, USA

Jack K. Chambers
University of Toronto, Toronto, ON, Canada

Christopher Chase-Dunn
University of California-Riverside, Riverside, CA, USA

Mohamed Cherkaoui
Centre National de la Recherche Scientifique, Paris, France

Vivek Chibber
New York University, New York, NY, USA

Vera Chouinard
McMaster University, Hamilton, ON, Canada

Thomas D. Christiano
University of Arizona, Tucson, AZ, USA

Robin Cohen
University of Oxford, Oxford, UK

Jill Cooper
Rutgers University, New Brunswick, NJ, USA

Charles Crothers
School of Social Sciences and Public Policy, AUT University, Auckland, New Zealand

Frank Cunningham
University of Toronto, Toronto, ON, Canada

Giacomo D'Alisa
Research & Degrowth, Barcelona, Spain; and Institute of Environmental Science and Technology (ICTA), Autonomous University of Barcelona, Barcelona, Spain

Catherine Darker
Trinity College, Dublin, Ireland

Diane E. Davis
Harvard University, Cambridge, MA, USA

Federico Demaria
Research & Degrowth, Barcelona, Spain; and Institute of Environmental Science and Technology (ICTA), Autonomous University of Barcelona, Barcelona, Spain

John Dixon
Open University, Milton Keynes, UK

Rebecca E. Dobash
University of Manchester, Manchester, UK

Russell P. Dobash
University of Manchester, Manchester, UK

Mattei Dogan[†]
National Center for Scientific Research, Paris, France

Doris Dransch
Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, Potsdam, Germany

Julie Drolet
Faculty of Social Work, University of Calgary, Edmonton, AB, Canada

Kevin Durrheim
University of KwaZulu-Natal, Scottsville, South Africa

Rainer Egloff
Ludwik Fleck Center at Collegium Helveticum, University of Zurich and ETH Zurich, Zurich, Switzerland; and Swiss Reinsurance Company Ltd, Zurich, Switzerland

Pegah Fakhari
Indiana University, Bloomington, IN, USA

Henry S. Farber
Princeton University, Princeton, NJ, USA

Sara B. Festini
University of Michigan, Ann Arbor, MI, USA

Glenn Firebaugh
Pennsylvania State University, University Park, PA, USA

Marcel Fournier
Université de Montréal, Montréal, QC, Canada; and Paris Institute for Advanced Study, Paris, France

Elizabeth Frankenberg
Sanford School of Public Policy, Duke University, Durham, NC, USA

David A. Freedman
University of California, Berkeley, CA, USA

Haruhiro Fukui
University of California, Santa Barbara, Goleta, CA, USA

Emilio Gentile
University of Rome La Sapienza, Roma, Italy

Patrick Gerland
Population Division, United Nations, NY, USA

Paul Gilroy
Yale University, New Haven, CT, USA

[†] Deceased.

Henry A. Giroux
McMaster University, Hamilton, ON, Canada; and Ryerson University, Toronto, ON, Canada

Jim Glassman
University of British Columbia, Vancouver, BC, Canada

Jorge A. Gonzalez
University of Texas-Pan American, Edinburg, TX, USA

Usha Goswami
University of Cambridge, Cambridge, UK

Aleksios Gotsopoulos
Sungkyunkwan University, Jongro-gu, South Korea

Adam S. Grabell
University of Michigan, Ann Arbor, MI, USA

Frauke Graf
Goethe-University Frankfurt, Frankfurt am Main, Germany

A. James Gregor
University of California, Berkeley, CA, USA

Doug Guthrie
New York University, New York, NY, USA

Amy G. Halberstadt
North Carolina State University, Raleigh, NC, USA

Neel Halder
Alpha Hospital, Bury, UK

Brenda Hanna-Pladdy
University of Maryland School of Medicine, Baltimore, MD, USA

Heidi A. Hanson
Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA

A. Paul Hare[†]
Ben-Gurion University, Beer Sheva, Israel

Jeffrey Hass
University of Richmond, Richmond, VA, USA; and St. Petersburg State University, St. Petersburg, Russia

John Hattie
University of Auckland, Auckland, New Zealand

Friederike Helm
University of Kiel, Kiel, Germany

Rémy Herrera
Centre national de la Recherche scientifique (CNRS), UMR 8174 — Centre d'Economie de la Sorbonne, University of Paris 1 Panthéon-Sorbonne, Paris, France

Dennis Hodgson
Fairfield University, Fairfield, CT, USA

Jared M. Hotaling
University of Basel, Petersplatz, Basel, Switzerland

Kenneth Hugdahl
University of Bergen, Bergen, Norway; and Haukeland University Hospital, Bergen, Norway

Esther Ignagni
The School of Disability Studies, Ryerson University, Toronto, ON, Canada

Samuel Issacharoff
New York University School of Law, New York, NY, USA

Kizhavana P. Jayasankar
School of Media and Cultural Studies, Tata Institute of Social Sciences, Deonar, Mumbai, India

Wesley G. Jennings
University of South Florida, Tampa, FL, USA

Gary F. Jensen
Vanderbilt University, Nashville, TN, USA

Tania Jivraj
The School of Disability Studies, Ryerson University, Toronto, ON, Canada

Debbie Jolly
Disabled People against Cuts (DPAC), UK

Simon Joss
University of Westminster, London, UK

Giorgos Kallis
Research & Degrowth, Barcelona, Spain; and Institute of Environmental Science and Technology (ICTA), Autonomous University of Barcelona, Barcelona, Spain

Alexander M. Keating
University of Pennsylvania, Philadelphia, PA, USA

Cecil D. Kemp
University of St Andrews, St Andrews, Fife, UK

Michael D. Kennedy
Brown University, Providence, RI, USA

Su-ming Khoo
School of Political Science and Sociology, National University of Ireland, Galway, Ireland

John Kleinig
John Jay College of Criminal Justice, New York, NY, USA; and Charles Sturt University, Canberra, ACT, Australia

Monika Knopf
Goethe-University Frankfurt, Frankfurt am Main, Germany

[†] Deceased.

Thorsten Kolling
Goethe-University Frankfurt, Frankfurt am Main, Germany

Thomas Kubiak
Johannes Gutenberg University Mainz, Mainz, Germany

Nancy G. Kutner
Emory University, Atlanta, GA, USA

Bettina Lamm
University of Osnabrück, Osnabrueck, Germany

Emily G. Lattie
University of Miami, Coral Gables, FL, USA

Nina Laurie
School of Geography, Politics and Sociology, Newcastle University, Newcastle upon Tyne, UK

Maria M. Laurito
Sanford School of Public Policy, Duke University, Durham, NC, USA

Sandra Soo-Jin Lee
Center for Biomedical Ethics, Stanford University Medical School, Stanford, CA, USA

Julian Leff
University College London, London, UK; Kings College London, London, UK; and University of Cape Town, Rondebosch, South Africa

John P. Lehoczky
Carnegie Mellon University, Pittsburgh, PA, USA

Rebecca J. Lester
Washington University, St. Louis, MO, USA

Elizabeth Lightfoot
University of Minnesota, Saint Paul, MN, USA

Fantasy T. Lozada
University of Michigan, Ann Arbor, MI, USA

Steven Lukes
New York University, New York, NY, USA

Arthur Lupia
University of California at San Diego, La Jolla, CA, USA

Deborah C. Malamud
New York University Law School, New York, NY, USA

Herb W. Marsh
Australian Catholic University, Strathfield, NSW, Australia; and King Saud University, Riyadh, Saudi Arabia

Joan Martinez-Alier
After Universitat Autonoma de Barcelona, Barcelona, Spain; and FLACSO, Quito, Ecuador

Lynn Mather
SUNY Buffalo Law School, Buffalo, NY, USA

Lynn McAlpine
University of Oxford, Oxford, UK

Joshua D. Meadors
University of California, Riverside, CA, USA

Igor Mel'čuk
University of Montreal, Montreal, QC, Canada

Natalie B. Milman
George Washington University, Oakton, VA, USA

Jens Möller
University of Kiel, Kiel, Germany

Yvana L.B.H. Mols
Emory University School of Law, Atlanta, GA, USA

Anjali Monteiro
School of Media and Cultural Studies, Tata Institute of Social Sciences, Deonar, Mumbai, India

Michele M. Moody-Adams
Columbia University, New York, NY, USA

Timothy D. Mooney
School of Philosophy, University College Dublin, Dublin, Ireland

Stephen J. Morse
University of Pennsylvania Law School, Philadelphia, PA, USA

Tom A. Moultrie
Centre for Actuarial Research (CARe), University of Cape Town, Rondebosch, South Africa

Hanno Müller-Kalthoff
University of Kiel, Kiel, Germany

Gerardo L. Munck
University of Southern California, Los Angeles, CA, USA

Carolyn B. Murray
University of California, Riverside, CA, USA

Daniel S. Nagin
Carnegie Mellon University, Pittsburgh, PA, USA

Nicole Nagy
University of Kiel, Kiel, Germany

Máire Ní Bhrolcháin
School of Social Sciences, University of Southampton, Southampton, UK

A. Bame Nsamenang
University of Bamenda, Bamenda, Cameroon

Ronald Oaxaca
University of Arizona, Phoenix, AZ, USA

Emiko Ohnuki-Tierney
University of Wisconsin, Madison, WI, USA

Sheryl L. Olson
University of Michigan, Ann Arbor, MI, USA

Alberto Palloni
Center for Demography and Ecology, University of Wisconsin, Madison, WI, USA

Emma Pearson
Universiti Brunei Darussalam, Gadong, Brunei

Andreas Peichl
ZEW Mannheim, University of Mannheim, IZA, ISER and CESifo, Germany

Lampros Perogamvros
University Hospitals of Geneva, Geneva; University of Geneva, Geneva, Switzerland; Geneva Neuroscience Center, University of Geneva, Geneva, Switzerland; and Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland

Nico Pestel
IZA Bonn and ZEW Mannheim, Germany

A. Peter Castro
Syracuse University, Syracuse, NY, USA

Thomas F Pettigrew
University of California, Santa Cruz, CA, USA

Gaëlle Pierre
International Monetary Fund (IMF), Washington, DC, USA

Marcel W.M. Post
University Medical Center Groningen, Groningen, The Netherlands; and Center of Excellence in Rehabilitation Medicine, University Medical Center Utrecht and De Hoogstraat, Utrecht, The Netherlands

Tom Postmes
University of Groningen, Groningen, The Netherlands

Kelly Raley
The University of Texas-Austin, Austin, TX, USA

Roger Ratcliff
The Ohio State University, Columbus OH, USA

Stefan Rayer
College of Liberal Arts and Sciences, University of Florida, Gainesville, FL, USA

Jennifer M. Reingle
University of Texas, Dallas, TX, USA

Jan D. Reinhardt
Institute for Disaster Management and Reconstruction of Sichuan University and Hong Kong Polytechnic University, Chengdu, Sichuan, China; and Swiss Paraplegic Research, Nottwil, Switzerland

Friedel M. Reischies
Freie Universität Berlin, Berlin, Germany

Patricia A. Reuter-Lorenz
University of Michigan, Ann Arbor, MI, USA

Robin Room
University of Melbourne, Melbourne, VIC, Australia; Stockholm University, Stockholm, Sweden; and Centre for Alcohol Policy Research, Turning Point, Melbourne, VIC, Australia

Paul Rozin
University of Pennsylvania, Philadelphia, PA, USA

Stefano Scarpetta
Directorate for Employment, Labour and Social Affairs (ELS) (Organisation for Economic Co-operation and Development (OECD)), Geneva, Switzerland

Giuseppe Schiavone
University of Milano, Milano, Italy; and Istituto Europeo di Oncologia (IEO), Milano, Italy

Uwe Schimank
Universität Bremen, Institute of Sociology, Bremen, Germany

Warren Schmaus
Illinois Institute of Technology, Chicago, IL, USA

Margaret Schneider
University of California Irvine, Irvine, CA, USA

Helmut Schreier
University of Hamburg, Hamburg, Germany

Georg Schreyögg
Freie Universität Berlin, Berlin, Germany

Sophie Schwartz
University of Geneva, Geneva, Switzerland; Geneva Neuroscience Center, University of Geneva, Geneva, Switzerland; and Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland

James J. Sheehan
Stanford University, Stanford, CA, USA

Leon V. Sigal
Social Science Research Council, New York, NY, USA

Ken R. Smith
Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA

Philip L. Smith
The University of Melbourne, VIC, Australia

Birgit Spinath
Ruprecht-Karls-University, Heidelberg, Germany

Frank M. Spinath
Saarland University, Saarbruecken, Germany

Ronald Stade
Malmö University, Malmö, Sweden

Mark C. Stafford
School of Criminal Justice, Texas State University, San Marcos, TX, USA

Naomi Stanford
Capella University, Minneapolis, MN, USA

Ellyn A. Steidl
The University of Texas-Austin, Austin, TX, USA

Daniel Stokols
University of California Irvine, Irvine, CA, USA

Richard Sturn
University of Graz, Graz, Austria

Julia Szalai
Hungarian Academy of Sciences, Budapest, Hungary

Simon Szreter
University of Cambridge, Cambridge, UK

Marylee C. Taylor
Pennsylvania State University, University Park, PA, USA

David J. Teece
University of California, Berkeley, CA, USA; and Berkeley Research Group, Emeryville, CA, USA

Irene Theodoropoulou
College of Arts and Sciences, Qatar University, Doha, State of Qatar

Manuela Thomae
University of Winchester, Winchester, UK

Duncan Thomas
Duke University, Durham, NC, USA

Felicity M. Turner
University of Groningen, Groningen, The Netherlands

Andrea Unger
Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, Potsdam, Germany

Chiara Volpato
University of Milano-Bicocca, Milan, Italy

Umberto Volpe
University of Naples SUN, Naples, Italy

Hans-Werner Wahl
Institute of Psychology, Heidelberg University, Heidelberg, Germany

Winfried W. Weber
Mannheim Institute of Applied Management Research, Mannheim University of Applied Sciences, Mannheim, Germany

Amy Wenzel
American Board of Professional Psychology (ABPP), University of Pennsylvania, Philadelphia, PA, USA

Lucy Whiston
Trinity College, Dublin, Ireland

Harry A. Whitaker
Northern Michigan University, Marquette, MI, USA

Susan K. Whitbourne
University of Massachusetts Amherst, Amherst, MA, USA

Michelle Williams
University of the Witwatersrand, Johannesburg, South Africa

Thomas M. Wilson
Binghamton University, State University of New York, Binghamton, NY, USA

Bjorn Wittrock
Swedish Collegium for Advanced Study in the Social Sciences, Uppsala, Sweden

Christopher R. Wolfe
Miami University, Oxford, OH, USA

Yu Xie
University of Michigan, Ann Arbor, MI, USA

Barbara Yngvesson
Hampshire College, Amherst, MA, USA

Francisco Yus
University of Alicante, Alicante, Spain

Daniela Zahn
Johannes Gutenberg University Mainz, Mainz, Germany

Azadeh Zamanian
University of Texas-Pan American, Edinburg, TX, USA

Defense and Security Economics

Rémy Herrera, Centre national de la Recherche scientifique (CNRS), UMR 8174 – Centre d'Economie de la Sorbonne, University of Paris 1 Panthéon-Sorbonne, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article deals with defense and security, mainly from an economic point of view. It discusses some preliminary conceptual and methodological problems, explains the specific difficulties related to statistical data sources on military spending, and presents a historical contextualization of the theoretical debates. It wonders if defense is a public good and why the army tends to be privatized. Then it examines the effects of defense spending within the neoclassical mainstream, as well as economic studies on defense at the margins of this current. Finally, it moves toward heterodox interpretations in alternative schools of economic thought. It concludes that it is impossible to build a defense and security economics without politics.

Preliminary Conceptual and Methodological Problems

Generally regarded as the sovereign function par excellence, defense is a very special category of public expenditure. A study whose purpose is evaluating military spending in an economy must specify the contours of the diverse phenomena to be investigated and clarify the issues of definition and measurement of these expenditures. It would seem reasonable to define military spending as being the set of human and material resources devoted by a state to its defense and intended to guarantee its national independence, the integrity of its territory and, where appropriate, the respect of the international treaties binding the country to foreign states. This first definition, however, which gives primacy to relations between states and the classic conception of 'war' in the study of conflict in international relations, tends to underestimate the interweaving of internal and international conflicts, in their unstable and constantly evolving interdependence.

Under these circumstances, it would be appropriate to extend the concept of military spending to include all or part of the resources employed by a state to maintain internal security and public order. Actions to preserve order, performed by distinct state forces, justified by reason of state and occurring in extremely varied political contexts, appear difficult to separate from national defense operations in the strict sense. The nature of these actions is often identical and the objective remains the same: to ensure the proper functioning of a nation's society and economy, or basically to guarantee its survival, identity, and power. This broader interpretation thus corresponds to a perception of the state, responsible for allocations of national resources, as a producer of security and *de facto* public goods involving constraint, ensuring the cohesion of social structures, creating the conditions for the continuity and stability of economic life, and guaranteeing the maintenance of an institutional framework for activity.

The defense budget is not necessarily an appropriate indication of a country's real military effort. In some countries, the Ministry of Defense in fact assumes functions that by nature belong primarily to the civilian sector. Conversely, there may be public expenditure from the budgets of civilian ministries, which is mainly for military purposes or which partially benefits the armed forces. In the modern world, there is a growing difficulty of distinguishing between the strictly military and civilian functions of the state. This is the case of 'special activities,' such as drone operations or cyber-attacks.

Problems Related to Statistical Data Sources

Among the most delicate problems to be solved is the determination of the conditions of international comparability of defense spending. It is related to structural elements specific to each country (e.g., type of economy, pricing system practiced), deflators (i.e., the method of deflating time series), or conversion factors (exchange rate used). Here, the methodological difficulties are numerous and highlight the secret nature of such budgets. In addition, there are some specifically statistical complexities, for example, double accounting, removing costs from the central budget, foreign exchange manipulations, calculations of debts associated with military spending, and aggregation methods (Herrera, 1994).

A thorough review of the main statistical data sources that are of better quality than the 'primary' ones (the latter being those issued by the country's central authorities themselves) reveals some shortcomings. We can mention those in relation to the retention of information by the governments and the resulting widespread (legitimate?) opacity existing in this area, and those linked to international balances of power that operate to the detriment of the Southern and Eastern developing countries. Indeed, a few Northern countries occupy dominant positions within the international statistical administrations, and particularly the United Nations (UN, 2013) and the International Monetary Fund (IMF, various years). Frequently, their intelligence services directly support the collection and treatment of indicators, such as in the case of the Central Intelligence Agency (CIA, 2013) and, formerly, the US Arms Control and Disarmament Agency for the United States and the International Institute of Strategic Studies (IISS, 2013) for the United Kingdom.

Despite the well-known experience and excellence of their researchers, the few independent observatories, especially the Stockholm International Peace Research Institute (SIPRI, 2013), have neither the financial resources nor the institutional power to constitute comprehensive and reliable sources of alternative information.

These limitations, greater or lesser depending on the various institutions, prevent this type of research from relying on a single database. The necessary information may be published in national budgetary and fiscal documents or it may be published in the reports disseminated by specialized international organizations, like the United Nations Institute for Disarmament Research (UNIDIR, 2009). Military spending can occur in a wide variety of different parts of the state, and can be included in accounting lines destined for very different purposes.

Historical Contextualization of the Debates

Moreover, it is important to contextualize the current theoretical and empirical debates on the role of defense in the economy. At the worldwide level, military spending grew almost exponentially throughout the twentieth century until the beginning of its last decade. Since the early 1990s, the renewed interest in such issues from authors belonging to the present mainstream (i.e., neoclassical economics) was decisively driven by concerted research programs of major international organizations: IMF, World Bank, Organisation for Economic Co-operation and Development, etc. Mainly oriented toward studying the economic effects of resource reallocation raised by the end of the Cold War and the so-called peace dividends, these programs have coincided – not by accident – with the closure of a series of studies previously carried out on the Third World countries' external debt, whose main policy conclusions were to explain the latter not by the dysfunctioning of the global economic and financial order, but rather by the 'mismanagement' of public finances in these countries. This opened the way to the requirement of 'good governance,' a concept originally developed by the IMF and applied to all aspects of public affairs, including the combat against the financing of terrorism after the 11 September attacks (IMF, 2003).

The pressure exerted on public budgets in the countries of the South (and the East, following the Soviet bloc's collapse, the end of the 'arms race' with the United States, and the launch of 'shock therapy' transitions) then justified recommendations for defense spending reductions and military industries conversions. These changes implied, more often than not, fiscal trade-offs that resulted in cuts in public spending, especially for social purposes, to ultimately constitute the surpluses needed to continuously repay external debt services.

The combined logics of unbalanced disarmament (ensuring the United States its advance in armaments) and accelerated liberalization (with state-owned enterprises' privatization and flexibilization) cumulated their negative impacts and often led to more human insecurity (Mushakoji, 2008). Some countries under structural adjustment even experienced a 'scissors effect' characterized by a drop in defense spending and a simultaneous increase in police expenditures devoted to internal security and public order, and turned against an enemy that is not abroad.

Such reductions in military spending thus allowed the opening of new fiscal margins of maneuver to secure the noninterruption of debt-related transfers to Northern private creditors, along with the opportunity to induce a disarmament of regional powers in the emerging South (especially China), at the very moment when the unipolar hegemony of the United States began.

Is Defense a Public Good?

As in all economic theory, defense economics is currently dominated by the neoclassical mainstream. Its basic goal is to analyze social phenomena exclusively from the behavior of individual agents. In such a theoretical framework, defense represents a typical example of a 'public good,' whose characteristics would theoretically justify the intervention of the state. In neoclassical modeling, the goods that are produced by markets are generally destined for private use, and the interdependencies between individuals are (almost exclusively) mediated by the price system. Public goods can be distinguished from such market goods by the fact that they are subject to a collective consumption, which is not shared but equal for each agent, and benefits all of them – without any access exclusion through the prices. By their nature, these goods cannot be produced and allocated by markets, since price mechanisms affecting rationing of limited goods do not function as usual. In these conditions, neoclassical theory admits that their production can be entrusted to the state, which provides such public goods to the community by levying a tax to finance the associated costs.

Consequently, the justification for the production of collective goods by the public sector lies in the admitted inability of markets to provide them. Two explanations are put forward. First, if the consumption of these goods is not divisible, and if their market price is considered to be zero, private enterprises will have no incentive to produce them. Second, the impossibility of exclusion will lead agents not to reveal their true preferences in order to lower their contributions to paying for them. A problem of free riders exists where agents take advantage of honestly revealed preferences (thus also appropriate payments) of others by falsely underreporting their preferences, and hence underpaying. The result is a risk of a bias in favor of the undersupply of the public good (Stiglitz, 1988).

The main characteristic of defense that makes it a public good is that its supply is indivisible, like the state that supplies it. The indivisibility of its supply, which connects to the concept of 'increasing returns' that potentially can lead to monopolistic market structures that could justify the intervention of the state, is coupled with an indivisibility of use. It is not possible to reserve the exclusive use of this good to some individuals, since every citizen enjoys a right of free access and cannot be denied this protection. Each individual is intended to consume the public good, not because its provision is determined by his/her own decision, but because defense is fully available to the entire community, with a requirement of its use. Furthermore, the addition of a new individual consumer does not alter either the cost of national defense or the level of consumption by the other agents, since all of them consume the good 'completely,' and without rivalry (Samuelson, 1983).

Thus, the neoclassical authors generally admit the state's intervention in the field of defense for reasons that they consider to be less political, even if it is a more 'sovereign function' than economic. Indeed, in the presence of such a public good that markets are unable to produce and allocate

by a standard price system, the state alone is able to ensure the optimal market equilibrium for the entire society. This is a brief presentation in general terms of the academic consensus that exists on this issue in mainstream economics today.

Privatizing the Army?

One of the major difficulties in applying neoclassical public economics to defense, however, comes from the fact that this field combines several types of market imperfections. In order to get an 'optimal' expenditure by the state for this good, corresponding to a level that allows the individuals' willingness to pay to equal the cost of its provision, it is still necessary that the size and the composition of public budgets, as well as the technical characteristics and the fiscal costs associated with defense and other public services, be known by the consumers. Nevertheless, the agents generally face barriers to access to the information needed to measure the exact impact of military spending, especially on their individual well-being, or on the level of homeland security, which determines, among others, the social utility of defense. Therefore, this problem of information asymmetry, related to the confidential nature of the statistical data on military activities, prevents optimal public choices and the associated procedures of intrabudgetary allocation of resources. An additional difficulty stems from the dichotomous and atypical nature of the public good of defense. The latter increases the security, and the level of social utility, of the country that instigates it. But at the same time, internationally it reduces the security level of the foreign countries that are viewed as potential enemies in a noncooperative framework. As a result, the positive consequences generated by the military spending of a country on its national security can be globally offset (and more) by the negative externalities they cause for the agents of the foreign countries.

Some ultraliberal economists have challenged the public good character of defense. Faced with market failures, highlighted by the neoclassical theory itself, the public choice school has never stopped asserting that the functioning of market mechanisms, even when imperfect, remains systematically preferable to the defects of the state's intervention, which can mask individual preferences under the discourse of 'general interest.' Such authors, who come from the 'right' of the mainstream and use utilitarian approaches, assert that state intervention may generate adverse effects. These include pricing practices that are deemed suboptimal, the government's inefficiencies and behaviors that would not get the necessary information to effectively control natural monopolies, the nonappropriability of profits (thereby destroying managerial motivation and incentive), the risks of 'bureaucratic drift,' and eventual abuses of dominant position (Buchanan and Musgrave, 1999). Under these conditions, they wonder why a US citizen, or so-called consumer-taxpayer, living in Seattle or in Washington, DC, for example, would agree to pay a tax whose purpose is to finance the protection of the border of the United States with Mexico? According to them, the corresponding taxation burden should be borne only by the inhabitants of the concerned federal states, from Texas to California.

On the basis of extreme antistate positions, completely in favor of private property and free trade, some 'libertarian' theoreticians have gone so far as to recommend the privatization of the army (Friedman, 1973). For them, defense should be submitted to the laws of competition, like any other private good. There would be markets where enterprises (i.e., private military companies) would offer the goods and services that have been provided so far by the state, and in which agents would 'freely' select according to their security needs and to a rational quality-price calculus, by performing a standard cost-benefit analysis (Squire and van der Tak, 1975). Although these (quite provocative) proposals remain a minority within the theoretical debates, they have come to significantly influence real political practice, and even when their supporters are often firmly against war, because here the main issue is still the role of the state. The result is that while defense is still more often than not conceived of as a 'public good' in theory, it is widely privatized in practice. Did the neoliberal waves associated with Reaganomics not prompt the US government to begin outsourcing defense-related activities already in the early 1980s, a practice that has been continually extended since then?

The Effects of Defense Spending within the Neoclassical Mainstream

Defense economics is the subject of numerous and increasingly technical contributions due to leading mainstream authors (Intriligator, 1990; Arrow, 1992; for example, and even Sen, 1992). Paradoxically, these studies have failed so far to reach clear and sure conclusions.

To begin with, the debate on the externalities of military research and development (R&D) remains highly controversial among neoclassical economists. Many modeled or econometric studies evaluating the effects of military R&D show that a diversion of resources to defense is costly in terms of production. Here various justifications are given: military and civilian R&D would use similar resources, which implies crowding-out effects; the inelasticity of human capital supply would pressure upward R&D input prices; negative price elasticity of demand for inputs in this sector would exert a disincentive to private investment in R&D, etc. But these arguments are disputed and other authors have stated that technical progress incorporated into military equipment has a stimulating impact on private investment. Thus, the eventual tradeoffs are not convincingly demonstrated.

A consensus seems, however, to emerge in the mainstream around the idea that the contribution of military R&D to total factor productivity tends to decrease over time. Often assimilated to externalities, the technological spillovers from defense to civilian sectors (through transfers of innovative products/processes, commercial implementations, organizational and informational effects, and so on), which characterized the post-World War II (WWII) period, may be less important today. Some even believe that at present the net flow of scientific knowledge is reversed and that it is mainly directed from civilian R&D to defense. Under these conditions, the resource diversion to the strategic security of a country could deteriorate its 'economic security' by reducing technical progress in export sectors, thus productivity growth and competitiveness. Such questions are very difficult to solve from the mainstream's

point of view, especially as the effects of military R&D on the economy are contradictory according to the data, samples, specifications, or methodologies that are used. As a consequence, more detailed analyses about dual technologies led on a case-by-case basis and both intra- and intersectorally could be advocated.

The results of empirical studies treating defense as a public good also arouse controversies. The impact of military spending on total factor productivity or gross domestic product (GDP) growth rate is sometimes found to be nonsignificant, by Aschauer (1989) and Barro (1991), for example, but negative by many others, directly or not, through effects on investment, human capital accumulation, or international trade due to the opacity of import procedures or trade restrictions. Some authors, such as Landau (1993), tested a hypothesis of a nonlinear influence of defense on growth, by combining positive effects (enhancing security or incentive to effectiveness of public policies facing external threat) and negative ones (resource diversion from productive uses) and using a quadratic functional form between growth and defense. Estimates would confirm such a nonlinearity: the coefficients of military burden appear positive for its simple value and negative for its squared value. Increases in defense budgets would thus be associated with faster growth up to an optimum, but slower beyond. Deger (1986) argued that the sign and magnitude of the impact of defense would depend on per capita income: in poor countries, military spending would promote growth by their modernizing influence; the relationship is reversed for middle-income countries, where the crowding out of investment would dominate; and when per capita income rises, the impact would be positive again, thanks to technological spillovers. Obviously, the debate remains open.

Derived from multisectoral representations of factorial contributions to growth, supply models with externalities have specified transmission channels of the effects of military spending on the economy (Biswas and Ram, 1986). Growth engines are analyzed in a framework that allows one to decompose and to estimate separately the impact of military spending in two effects: the first one measures intersectoral differentials of marginal factorial productivity; the other captures the external effects of defense disseminated to the private sector. Nevertheless, most econometric estimates using such a methodology give disappointing results, often revealing no significant impact of military spending on GDP growth. These mitigated findings could be partly explained by the persistence of unresolved technical difficulties due to multicollinearity between the variables tested.

In addition, typically neoclassical computable general equilibrium models have been used in this field. Their simulations often exhibit investment and GDP gains consecutive to reduced military spending. However, this type of formalization is strongly questionable. Studying the multifaceted effects of defense requires analyzing more specifically the implications of the accretion of various forms of capital. The enlargement of the approach to long-term growth allowed one to use more sophisticated modeling than econometric specifications, and to capture truly dynamic effects of these expenditures. Endogenous growth theory has placed emphasis on the role of the accumulation of human capital and/or infrastructure in determining the growth rate in the long run. Here, choices of economic policy related to the structure of public budgets may be important in explaining the economies' growth performances, by trade-offs operating between military spending and growth engines. Among the modeled positive effects emanating from the defense sector, one can mention, in the production function, the increase in total factor productivity caused by the influence of military capital on technical progress and, in the social utility function, a strengthening of security, and thereby an improved well-being. The possible negative effects include the mobilization of skilled workers by defense, which reduces human capital available for productive sectors, or fiscal trade-offs to the detriment of other public spending (such as education). Endogenous growth modeling is, however, subject to a series of criticisms (Herrera, 2011).

A deficiency of many neoclassical models is the lack of consideration of security indicators, which prohibits thinking in strategic and geopolitical terms. The analysis of the links between growth and security, resulting particularly in deepening interactions between Pareto optimality and Nash equilibrium can then lead into the field of game theory, a field of mathematical expression (Intriligator and Brito, 1976). But here again, the attempts of incorporating military variables in the neoclassical paradigm, which have certainly brought some very interesting insights, remain often contradictory, uncertain, or unsuccessful (Herrera, 2010).

At the Margins of the Mainstream

Nevertheless, economic theory has experienced consistent developments near or even outside the margins of the mainstream. Such is the case with the interindustry input–output models. Distinct from the standard neoclassical analysis, the latter provide original tools – relevant in our opinion– to measure macroscopic effects of changes in defense spending at the national, regional, and international levels, taking into account resource allocations between civilian and military sectors in disaggregated frameworks. Such a methodology, using projections on short, medium, or long terms, has proved especially suitable for assessing the effects of fiscal choices made in connection with various disarmament plans. Input–output analyses were repeatedly employed by Leontief and his teams to study the consequences of cuts in defense budgets on the US economy, and to simulate scenarios of regional and global disarmament (Leontief and Duchin, 1983). At the margins of the neoclassical mainstream, macroeconomic demand models, which use an equation of national income balance from Keynesian inspiration, postulate a negative impact of military burden on demand – via the multiplier – by distinguishing differentiated influences on nonmilitary public spending and on components of the private sector absorption (Smith, 1977). This approach has the advantage of providing a framework that identifies crowding-out effects following defense spending shocks that are potentially detrimental to consumption, investment, or exports, and may adversely affect short-term activity, in particular through underemployment. However, these models are insufficient: they do not capture the impact of the military sector on the accumulation dynamics; they do not distinguish with the necessary precision effects

operating in long or short term, nor the different contractions of investment (directly productive) or consumption (unproductive); and finally, they do not really clarify the atypical nature of defense, which has complex relations with the rest of the economy, distinct from those associated with other public spending (Benoit, 1973). Moreover, empirical studies based on such specifications remain in a state of uncertainty as to which category of demand may suffer the most from these crowding-out effects (Hartley and Sandler, 2007).

Under these conditions, a path followed by authors of the Keynesian-neoclassical synthesis has been to consider, in a perspective of longer term, a supply component, whose role is frequently neglected in most Keynesian models. Formally speaking, taking into account supply requires using a standard production function. The introduction of military spending in a multisectoral framework with differentiated production functions can consider a final good of the private sector derived from a technology combining usual factors (capital and labor), plus a military public good, to be interpreted as an externality. But it can also go through the disaggregation of the capital stock into civilian and military components in order to estimate econometrically the elasticity of the product to the military capital and the contribution of this stock to growth. However, the economically unproductive (or not directly productive) character of defense still makes this kind of specification difficult to sustain.

Consequently, some economists are led to simultaneously incorporate supply and demand aspects in dynamic equations, which allows them to deduce from the national income balance an expression of capital accumulation and a production function where efficient labor and effectively used capital (with an efficiency parameter characterizing capital underutilization) enter. As growth in the steady state is determined by demographic and technological evolution (assumed exogenous here), military spending has no effect on it in the long term, except if a potential effect of these expenditures on technical progress is considered. But their impact is perceptible on the transitional path leading to the steady state. Such models make the direct impacts of defense spending conceivable: negative supply-side effects on private capital accretion, through the savings rate, by crowding out investment; but positive demand-side effects on the utilization rate of capital, through the mobilization of productive capacities in the short term (Deger and Sen, 1995). More indirect influences may also be highlighted, including on the private sector's propensity to save, the share of civilian expenditures in public budgets, the capital–output ratio, or the balance of payments. Adverse effects may be amplified if, for example, the increase in the military burden is covered by cuts in social expenditures.

Toward Alternative Interpretations

In the manufacture of armament equipment, the economic weight of the private enterprises in the military–industrial complex was already considerable at the end of WWII. In the 1950s, US President Dwight D. Eisenhower felt compelled to warn his fellow citizens against the increasing threat posed by the collusion of interests of this industry, big business, and the army (see Eisenhower's Farewell Address to the Nation on 17 January 1961). During the conflict in Vietnam, many private big corporations were integrated into the war effort; among them were DynCorp, Halliburton, Vinnell, Lockheed Martin, and General Dynamics. Today, the main suppliers of the US Department of Defense are transnational firms involved in shipbuilding, aerospace, and computers, whose military contracts drain off astronomical amounts of federal dollars. Private military companies also provide to the federal states, international institutions, nongovernmental organizations, and other private entities. They provide a very large range of technical services aimed at assisting or supplementing regular armed forces in the protection of strategic sites (military bases, oil fields, mines, nuclear facilities, etc.), and in building, logistics, military training, intelligence, public relations, and the provision of specialized staff (consultants, bodyguards, mercenaries directly involved in combats, etc.).

During the period of the Cold War, the expansion of the productive forces in the United States was partly driven by the 'arms race,' including the development of a number of important associated technological advances, such as computation systems, robots controlled by computers, and, more recently, the Internet. Nowadays, the US defense budget remains large, at nearly 5.0% of GDP and 45% of the world military expenditures. Moreover, the military–industrial complex continues to play a key role, but it now functions under the control of high finance. As a matter of fact, in the military sector in the 2000s, the share of the ownership capital held by institutional investors that were themselves controlled by the biggest financial oligopolies reached approximately 95% of the total assets of Lockheed Martin, 75% of those of General Dynamics, 65% of Boeing, and so on. A similar evolution occurred to the private military companies, with an increasing share of their capital falling under the thumb of high finance. Indeed, Military Professional Resources Inc. (MPRI) was bought by L-3 Communications, Vinnell by Carlyle, and DynCorp by Veritas.

At this stage, one may wonder whether crisis and war are intertwined. From a Marxist point of view, war would be economically integrated into the capitalist cycles as an extreme form of capital destruction (or devalorization). At the same time, and in a context of accentuation of the internal contradictions of capitalism, it would serve politically to reproduce the conditions for maintaining command by the dominant fractions of the ruling classes, that is to say, high finance, over the capitalist world system.

These observations revitalize theories of imperialism, for which the use of armed force is a strategy imposed by finance as the means of reproduction of its power, and where militarization is the essential modality of the existence of capitalism. In such a context, the role of the state becomes absolutely crucial to capital, since it is the state that goes to war on behalf of capital.

Military spending has become a major source of profit for the capitalists, in particular because it can transform productive capital – especially when it is financed by public debt – into fictitious capital. In addition, it should be observed that the wars in Afghanistan and Iraq were launched in a time of crisis (2001–03, as the years 1913 and 1938 were too); a crisis that emerged along with monetary policy changes occurred in connection with the worsening of the internal and external

deficits in the United States, partly because of the need to finance its war effort as a global hegemony.

The question then is whether the United States could revitalize capital accumulation in the center of the world system by wars. Some answer no, insofar as the destruction of (constant and variable) capital caused by these conflicts, which are dramatic for the countries that suffer from them, would be insufficient to promote a new long cycle of capital expansion in the United States, as was the case in the post-WWII period. Such wars would also be insufficient when one takes into account the associated effects of effective demand, mostly active in the short term, and the technological effects, which are positive only for the military–industrial sector. Insufficient, unless modern imperialist wars expand and become permanent.

Under these circumstances, it would definitively be hopeless to try to build, as the neoclassical mainstream is attempting, an 'economics of war,' or even a defense and security economics, without politics (Herrera, 2013).

See also: Alternative Schools of Economic Thought; Behavioral Economics; Data Bases and Statistical Systems: Security and Conflict; Defense and Security Economics; Dispute Resolutions in Economics; Economic Growth; Game Theory: Noncooperative Games; Imperialism; Infrastructure Investment; Innovation and Technological Change, Economics of; Law and Economics; Militarism; Military, War, and Politics; Nation-State and War; National Security Studies and War Potential of Nations; Political Repression; Power; Securitization; War and Democracy.

Bibliography

Arrow, K., 1992. The basic economics of arms reduction. In: Isard, W., Anderton, C. (Eds.), Economics of Arms Reduction and the Peace Process. North Holland, Amsterdam, pp. 57–67.

Aschauer, A., 1989. Is public expenditure productive? Journal of Monetary Economics 23 (2), 177–200.

Barro, R., 1991. Economic growth in a cross section of countries. Quarterly Journal of Economics 106 (2), 407–443.

Benoit, E., 1973. Defense and Economic Growth in Developing Countries. The MIT Press, Cambridge, MA.

Biswas, B., Ram, R., 1986. Military expenditure and economic growth in less developed countries. Economic Development and Cultural Change 34 (2), 361–372.

Buchanan, J., Musgrave, R., 1999. Public Finance and Public Choice: Two Contrasting Views of the State. The MIT Press, Cambridge, MA.

CIA, 2013. The World Factbook 2013–14. Public Affairs, Washington, DC.

Deger, S., 1986. Military Expenditure in the Third World Countries: The Economic Effects. Routledge & Kegan Paul, London.

Deger, S., Sen, S., 1995. Military security and developing countries. In: Hartley, K., Sandler, T. (Eds.), Handbook of Defense Economics, vol. 1. North Holland, pp. 275–307.

Friedman, D., 1973. The Machinery of Freedom. Harper Collophon, New York.

Hartley, K., Sandler, T. (Eds.), 2007. Handbook of Defense Economics. Elsevier, North-Holland.

Herrera, R., 1994. Statistics on Military Expenditure in Developing Countries: Concepts, Methodological Problems, and Sources. OECD Development Centre, Paris.

Herrera, R., 2010. Dépenses publiques et croissance économique. L'Harmattan, Paris.

Herrera, R., 2011. A critique of mainstream growth theory. Research in Political Economy 27 (1), 3–64.

Herrera, R., 2013. Between crisis and war – where is the United States heading. Journal of Innovation Economics and Management 2 (12), 151–174.

IMF, various years. Government Finance Statistics Yearbook. Statistics Department, Washington, DC.

IMF, 2003. Good Governance: The IMF Role. Washington, DC.

IISS, 2013. Strategic Survey 2013 – The Annual Review of World Affairs. London.

Intriligator, M., 1990. On the nature and scope of defence economics. Defence and Peace Economics 1 (1), 3–11.

Intriligator, M., Brito, D., 1976. Formal models of arms races. Journal of Peace Science 2, 77–88.

Landau, D., 1993. The Economic Impact of Military Expenditure Best. World Bank Policy Research Working Paper, No 1138. The World Bank, Washington, DC.

Leontief, W., Duchin, F., 1983. Military Spending – Facts and Figures, Worldwide Implications and Future Outlook. Oxford University Press, Oxford.

Mushakoji, K. (Ed.), 2008. Human (In)security in the Networks of Global Cities. CHS Chubu University, Aichi.

Samuelson, P.A., 1983. Foundations of Economic Analysis (1947), enlarged ed. Harvard Economic Studies, Cambridge, MA.

Sen, A., 1992. Wars and famines. In: Isard, W., Anderton, C. (Eds.), Economics of Arms Reduction and the Peace Process. North Holland, Amsterdam, pp. 219–234.

Smith, R., 1977. Military expenditure and capitalism. Cambridge Journal of Economics 1 (1), 61–76.

Squire, L., van der Tak, H.G., 1975. Economic Analysis Project. Johns Hopkins Press, Baltimore.

Stiglitz, J., 1988. Economics of the Public Sector. W.W. Norton, New York.

SIPRI, 2013. SIPRI Yearbook 2013. Oxford University Press, Oxford.

UN, 2013. United Nations Disarmament Yearbook. UN Office for Disarmament Affairs, New York.

UNIDIR, 2009. Disarmament Forum: Ideas for Peace and Security. Geneva.

Defiant Behavior During Adolescence across Cultures

Leslie Berntsen and Laura A Baker, University of Southern California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A review of cross-cultural studies of defiant and other deviant behavior during adolescence is provided. Race, ethnicity, and nationality are considered as possible aspects of culture. Although deviant behavior is relatively common in the United States and other Western cultures, there are variations across racial and ethnic groups. Cross-nationally, there is little variation across Western cultures, although deviance is somewhat less prevalent in nonindustrialized societies where children gain adult responsibilities immediately after puberty.

Adolescence has long been equated with defiant behavior. In *Rhetoric*, Aristotle declared youth to be as "heated by nature as drunken men by wine." In *Emile*, French Enlightenment philosopher Jean-Jacques Rousseau advised parents to "keep their hand[s] on the helm … [lest] all be lost." And, in 1904, G. Stanley Hall, the first president of the American Psychological Association, coined the now ubiquitous phrase "storm and stress" to refer to the characteristic turbulence of the teenage years. Adolescents, he believed, constituted a separate race that would be properly integrated into normal adult society in their own due time (Arnett, 1999: p. 317).

Hoping to challenge the popular conception that the 'omnipresent and obvious symptoms of unrest' widely observed during adolescence are simply 'characteristics of the period,' anthropologist Margaret Mead set sail for Samoa in 1926. After 2 years of living among young women on the island of Ta'u, she ultimately concluded that "adolescence represented no period of stress or crisis," largely due to the "general casualness of the whole society" (Mead, 1928: pp. 109, 137). Unlike contemporary American culture, authority was diffused over large families, young people could move from family to family within their villages, and girls had the socially sanctioned freedom to sexually experiment. In Samoa, the constraints of a society that demonizes adolescents were conspicuously absent.

In the years since its publication, *Coming of Age in Samoa* has been met with its fair share of controversy. Based on his own fieldwork in Samoa, one of Mead's most vocal critics, Derek Freeman (1983), claims that Mead had not only been deceived by her informants, but that she also deliberately fabricated evidence in support of her ideological leanings. More recently, other anthropologists have come to Mead's defense, claiming that Freeman's criticisms (1) lack ethnographic accuracy and (2) represent an unjustified attack on her reputation driven by his own personal beliefs (see Shankman, 2009). While we will not further discuss the Mead–Freeman debate in this article, we do acknowledge *Coming of Age in Samoa* as the first major work to challenge the universality of adolescent defiance, thus laying the theoretical groundwork for the future study of cross-cultural adolescent behavior.

Defiance as a Form of Deviance

In reviewing cross-cultural studies of defiant behavior, it is useful to consider defiance as a specific form of the more general category of antisocial or deviant behavior. In this article, we thus broadly define 'defiance' to also include various forms of deviant and antisocial behavior. For the sake of uniformity, we will refer to all of these behaviors as 'deviant' for the remainder of the article.

At its most basic level, deviance is principally characterized by a violation of social norms (see Franzese, 2009: pp. 6–7, for a review of contemporary sociological definitions). Since deviance is always defined in relation to a set of specific collective expectations, it is an inherently value-laden, context-dependent term. Beliefs and behaviors that might be viewed by one person as 'deviant' could easily be viewed as acceptable or even virtuous by another (Potter, 2011). In the United States, deviant behaviors that actively impinge upon the rights of others are diagnostically codified in the form of oppositional defiant disorder (ODD), conduct disorder (CD), and antisocial personality disorder (American Psychiatric Association, 2000). Many scholars have argued against the medicalization of deviance on the grounds that a diagnostic label can serve as a powerful tool of social and political control that can be used to pathologize behaviors that are more disagreeable than dangerous (see Conrad, 1992).

For example, during the Cold War, government officials in both China and the former Soviet Union imprisoned 'mentally ill' political and religious dissidents in asylums in order to suppress the opposition they deemed harmful to the rest of society (Bonnie, 2002). An in-depth analysis of how certain acts come to be regarded as deviant or whether they deserve to be considered deviant is outside of the purview of this article. Here, we limit our discussion to deviant behaviors (including defiance), widely associated with adolescence in the Western world. These include drug use and distribution, risky sexual behavior, generalized delinquent (including law-breaking) behavior, and aggression, all of which cause direct economic, emotional, or physical harm to oneself or others.

What Is Culture?

In order to understand how certain cultural factors might encourage or discourage deviant behavior during adolescence, we first must specify what we mean by culture. In the definition often referred to as the 'gold standard' among anthropologists, E.B. Tylor (1871: p. 1) defined culture as "that complex whole which includes knowledge, belief, art, morals, law, custom, and

any other capabilities acquired by man as a member of society." A century later, specifically applying the definition to the study of psychological processes, Marsella (1987: p. 381) proposed that culture also comprises any individual's "values, attitudes, beliefs, cognitive styles, and patterns of consciousness." Although preferred by scholars in different disciplines, both of these definitions converge on the idea that 'culture' cannot be neatly reduced to a single factor. Unfortunately, cross-cultural scientific research often disregards the notion of a 'complex whole' in favor of empirical simplicity. 'Culture' is frequently conflated with racial or ethnic categories, with national (or even continental) borders often used as a proxy for group membership. Although a common identity might be more likely to lead to shared experiences that contribute to common psychological development, an emphasis on group membership above all else can easily obscure within-group heterogeneity and draw misleading distinctions between individuals that actually share more similarities than differences (see Choudhury and Kirmayer, 2009). Not all of the research cited in this review was conducted with these principles in mind, so we urge readers to look beyond labels in order to understand the underlying processes – values, beliefs, practices, etc. – that constitute the ever-evolving notion of 'culture' and that might contribute to adolescent deviance.

With these definitions and concerns in mind, we review here studies investigating the differences in deviant behavior across groups pertaining to *race, ethnicity, and nationality*. We also discuss sociological theories that suggest differences across industrialized vs nonindustrialized societies, and individualism vs collectivism. Although gender and socioeconomic groups might well be considered as cultural entities, we have not reviewed these studies in this article, since they are described elsewhere.

Cross-National Differences

Mixed results emerge from the literature on cross-national differences in deviant behavior during adolescence. One meta-analysis of 25 cross-national studies conducted between 1997 and 2008 revealed that prevalence estimates of both ODD (0.6–11.0%) and CD (0.47–8.7%) did not differ as a function of geographic location (Canino et al., 2010). Rather, methodological differences between studies accounted for a greater proportion of the variability in prevalence rates. However, of the 25 studies that met criteria for inclusion in the final analysis, only four were conducted in non-Western countries (three in Asia and one in the Middle East). In their discussion, the authors note that differences between Western and non-Western countries might have emerged if more non-Western cultures were included in their analysis.

Findings from non-Western cultures do, in fact, suggest some cross-national differences in deviant behavior. Markedly lower rates of adolescent problem behavior (smoking, drinking, and general delinquency) have been found in China compared to the United States (Jessor et al., 2003). In observational studies of children in four non-Western cultures (American Samoa, Belize, Kenya, and Nepal), physical aggression was not an entirely common occurrence, comprising 10% of boys' and 6% of girls' social behavior (Munroe et al., 2000). However, aggressive acts were significantly more common among children in the two patrilineal societies (the Newars of Nepal and Logoli of Kenya) than those in the nonpatrilineal societies. Still other researchers have found adolescent deviance to be quite common in non-Western cultures. Among a random sample of male adolescents in Colombia, Pineda et al. (2006) observed prevalence estimates of CD to range from 10.5 to 35.5%, much higher than independent estimates in the United States (Hinshaw and Lee, 2003).

Despite differences in the incidence of adolescent deviance, it does seem that similar developmental processes are at play in discouraging such behaviors across national borders. Investigating the relationship between adolescent misconduct and the parent–child relationship in Hungary, the Netherlands, Switzerland, and the United States, Vazsonyi (2003) found parental closeness, support, and monitoring to independently predict adolescent problem behaviors (ranging from misbehaving in school to physical assault) in all four countries. Researchers have also found a similar protective relationship between parenting practices and deviant behavior (drug use, nonaggressive antisocial behavior, and physically aggressive antisocial behavior) among adolescents in Canada, France, and Italy (Claes et al., 2005).

In summary, the evidence for cross-national differences in deviant behavior during adolescence appears to be mixed. While few differences have been found between (mostly) Western cultures (Canino et al., 2010), some disparities are evident between Western and non-Western cultures, and when certain non-Western cultures are compared to one another. Rates of substance use and delinquency are lower in China but higher in Colombia compared to the United States, while aggression is not uncommon in other non-Western cultures (American Samoa, Belize, Nepal, and Kenya). Although these findings do not allow us to draw definitive conclusions about the worldwide distribution of adolescent deviance, they do lend support to Mead's initial observation that it is (in its many manifestations) far from a universal phenomenon.

Racial/Ethnic Differences in the United States

To explore whether adolescent deviance might be more consistent within a given country, we now turn to differences between racial and ethnic groups in the United States. Although criminal behavior can be a useful index of deviance, African and Hispanic Americans are disproportionately incarcerated relative to their Asian or European American counterparts (Glaze, 2011). Since there is growing evidence to suggest that these disparities can be largely attributed to racial profiling, aggressive police practices, targeted prosecutions, and sentencing biases (Gelman et al., 2007; Steffensmeier et al., 2006), it is especially crucial to look toward empirical studies that propose cogent explanations for specific differences in deviant behavior, rather than attempt to draw conclusions from crime statistics or any other purely observational data.

Using data from the National Longitudinal Study of Adolescent Health, Blum et al. (2000) found some evidence of racial differences in adolescent high-risk behaviors. Replicating

earlier findings, White adolescents were the most likely to use substances or attempt suicide, Black adolescents were the most likely to have already engaged in sexual intercourse, and Black and Hispanic adolescents together were more likely to participate in acts of violence than White adolescents. However, after controlling for gender, income, and family structure, race accounted for less than 10% of the variance in risk behaviors among both younger and older adolescents. That is, although racial differences were observed, race alone could not accurately predict the likelihood that any one person might engage in those behaviors. An additional study focusing solely on high-risk sexual behaviors during adolescence found similar results. Although higher levels of parental education, a two-parent family, and a European–American background were all independently negatively associated with risky behaviors, neither race nor socioeconomic status could fully explain the rates of pregnancy and sexually transmitted diseases observed in the sample (Santelli et al., 2000). These results further emphasize the heterogeneity that occurs within racial and ethnic groups in the United States, which ultimately should discourage their use as sole explanatory variables in scientific research.

Industrialized vs Nonindustrialized Societies

Contemporary research continues to validate Margaret Mead's assertion that adolescence is neither uniformly experienced nor defined. Schlegel and Barry (1991) have found the usual 'storm and stress' of adolescence to be far less frequent, if at all present, in nonindustrialized societies. Interestingly, many of these societies lack a specific word to describe the period between childhood and adulthood due to various cultural constraints. For some societies that are heavily agrarian, children might work directly alongside adults as soon as they are physically capable. In others, girls might be expected to bear their own children immediately after the onset of puberty. And, in perhaps the most drastic shift from childhood to adulthood, some children are involuntarily recruited to participate in armed conflict. Each of these cases leaves little, if any, room for the developmental middle ground that is adolescence (see Brown and Larson, 2002).

Individualism vs Collectivism

However, as previously mentioned, some industrialized societies (e.g., the United States and China) have been found to show markedly different levels of deviant behavior among adolescents (Jessor et al., 2003). In order to account for this difference, we now turn our attention to two distinct, heavily researched cultural typologies: individualism and collectivism. Much like the name suggests, individualism prioritizes the needs and desires of the individual over those of any given group. Individualists often have loosely defined, dynamic social relationships and generally prefer to exist autonomously of others' influence. In contrast, collectivists tend to define themselves by their interdependent relationships with friends, family, and society at large, placing great emphasis on group needs and social harmony (Markus and Kitayama, 1991; Triandis, 1995). Although collectivism is rooted in traditional Confucian values, it is certainly not geographically bound; collectivistic values have also been widely observed throughout Latin America, Africa, and South Asia (Triandis, 1989).

Might cross-cultural differences in deviant behavior during adolescence emerge as a consequence of different values between individualistic and collectivistic cultures? Tyson and Hubert (2012) have found that adolescents from Australia and various East Asian countries differ in the way they think about delinquent behaviors (e.g., theft, property damage, drug use, drunk driving, etc.). Although participants from both groups agreed which behaviors were more or less severe than others, Australian adolescents rated delinquent behaviors as significantly less wrong than their East Asian counterparts. Especially reflective of the collectivist emphasis on social harmony and deference to authority, 'shouting abuse at your parents' was the single item with the greatest difference between the two groups. Further, as East Asian participants aged, they perceived delinquency as less acceptable, while Australian participants showed no change or even perceived delinquent acts as more acceptable, depending on the specific behavior.

But do differences in beliefs necessarily map on to differences in behavior? One metaanalysis of 185 different cross-national comparisons found that aggression was significantly less frequent in cultures that value collectivist ideals, egalitarianism, and moral discipline (Bergeron and Schneider, 2005). However, this study is only an indirect measure of association between cultural values and aggression. As the authors acknowledge, national indices of values are very useful for understanding cross-national differences in behaviors, but they do little to help us understand how cultural factors might influence individual differences.

When study participants are asked to specifically report the extent to which they endorse individualism or collectivism, similar patterns of results emerge. In a mixed gender sample of Chinese adolescents, self-reported individualism positively predicted, and self-reported collectivism negatively predicted, peer and teacher ratings of both overt and relational aggression (Li et al., 2010). This pattern has also been found to extend to delinquency more generally. Among a sample of Chinese, Cambodian, Laotian, and Vietnamese youth, the frequency of self-reported delinquent behaviors was, again, directly related to individualism and inversely related to collectivism (Le and Stockdale, 2005).

Similar cultural values have been found to exert similar effects among non-Asian adolescents. In research examining the relationship between traditional Latino family values and externalizing behavior, German et al. (2009) found that Mexican–American adolescents with a heightened sense of familial obligation and emotional closeness to their family exhibited fewer externalizing conduct problems according to two independent teacher reports, even when they were exposed to deviant peers. However, a separate study revealed that Mexican adolescents who had acculturated to mainstream American culture reported engaging in more delinquent behaviors than those who had chosen to remain unacculturated, those who had very recently immigrated to the United

States, and White non-Hispanic adolescents (Fridrich and Flannery, 1995).

In short, a wide variety of deviant behaviors appear to be more reliably associated with broad sets of values than with race, ethnicity, or nationality, per se. Regardless of the specific name, belief systems that emphasize the responsibility of the individual to some larger collective entity (whether it be one's family or society at large) seem to protect against adolescent deviance.

Biological Explanations of Adolescence

When it comes to explaining the 'storm and stress' of contemporary Western adolescence, biological, value-free explanations reign supreme. In these accounts, deviant, risk-taking behaviors are framed as unfortunate, yet inevitable, consequences of the many neurodevelopmental changes that accompany puberty. Some of the most profound changes are related to the development of the prefrontal cortex. Often referred to as the brain's CEO, the prefrontal cortex is responsible for a variety of higher-order executive functions, including controlling one's impulses, initiating appropriate behaviors and inhibiting inappropriate behaviors, organizing complex information, making decisions, and empathizing with others. Clinical populations with prefrontal damage often show lapses in moral reasoning, an inability to anticipate future consequences, and an inability to inhibit inappropriate social behavior. In healthy individuals, the prefrontal cortex does not fully mature until the mid- to late 1920s. As it matures, the number of dopaminergic inputs to the region dramatically increases, a change reflected the newfound ability to control one's impulses and make more mature, reasoned decisions (see Weinberger et al., 2005, for a review). As a result, prefrontal deficits are widely cited as the sole underlying cause of adolescent 'storm and stress' (for reviews, see Blakemore and Choudhury, 2006; Casey et al., 2008; Nelson et al., 2005).

Challenges to the Storm and Stress Model

As Jeffrey Jensen Arnett (1999) suggests, the dominant 'storm and stress' model of adolescent deviance discounts individual differences in the adolescent experience within a given society, in addition to the differences observed between societies. Although the teenage years might be the most likely time for storm and stress to occur, not every teenager in the Western world experiences a similar period of angst, rebellion, and high-risk behavior. As a scientific category, adolescence represents the product of a very specific social and historical moment in time in which social Darwinism reflected the dominant ideology. Adolescents, in the view of early theorists like G. Stanley Hall, were simply not fit enough to survive (Choudhury, 2010). As a result of both intracultural differences in the normative adolescent experience as well as its history as a construct, some researchers argue that adolescence is a socially constructed, culture-bound syndrome that has been wrongly medicalized and used as a tool of social control (Hill and Fortenberry, 1992).

Culture and Biology: Mutually Constitutive

Clearly, the brain and the changes it undergoes "cannot be viewed separately from the material, historical, and social circumstances that produced it" (Lock and Nguyen, 2010: p. 109). Even so, it is important to note that the acknowledgment or appreciation of cultural differences does not require an outright repudiation of scientific evidence. As Davis and Morris (2007) argue, "the biological without the cultural, or the cultural without the biological, is doomed to be reductionist at best and inaccurate at worst" (p. 411). Neither a biological explanation of the 'storm and stress' model nor a purely descriptive account of cross-cultural variations is entirely correct. To get a more complete picture, we first need to understand how culture and biology might interact to give rise to psychological outcomes.

In a prime example of culture-gene coevolution at work, Chiao and Blizinsky (2010) found that individuals from collectivist cultures across 29 nations were more likely to carry the short (S) allele of the 5-HTTLPR serotonin transporter gene, which is associated with negative emotion and heightened anxiety. However, these same individuals were also less likely to suffer from mood or anxiety disorders. At first glance, these results seem paradoxical: How could a population with a heightened genetic risk for mood and anxiety disorders show decreased prevalence of those disorders? In their analysis, the authors argue that the collectivistic emphasis on social harmony lessens exposure to the types of chronic stress that put genetically susceptible individuals at risk of developing these conditions. That is, the cultural environment serves as a protective buffer against the harmful effects of various forms of psychopathology.

Conclusions and Future Directions

Deviance is clearly a multidimensional construct. Various forms of deviant behavior during adolescence have been observed in many locations and populations around the world, although there are notable differences and trends. First, adolescent deviance seems to be less common in nonindustrialized societies where children gain adult responsibilities immediately after puberty (see Brown and Larson, 2002). Among Western industrialized nations, there is little overall variability in the prevalence of deviant or antisocial behaviors. And within the United States, there appear to be negligible differences between racial and ethnic groups. Nonetheless, adolescent deviance is generally a common occurrence in the Western world. When compared to Western adolescents, those from non-Western (especially Asian) cultures show significantly lower levels of a wide variety of deviant behaviors. Some researchers suggest that adolescent 'storm and stress' arises as the result of brain development, while others believe that it might be most productively understood as a function of specific cultural constraints and belief systems. Although a single cause remains elusive, it is clear that both biological and cultural explanations are not only useful, but vitally necessary, to understand deviant adolescent behavior.

See also: Aggression, Social Psychology of; Behavioral Problems, Effects of Socio-Economic Status On; Collective Behavior, Social Psychology of; Cross-Cultural Psychology; Cross-Cultural Research Methods in Psychology; Defiant Behavior During Infancy and Early Childhood; Developmental Research across Cultures and Nations: Challenges, Biases, and Cautions; Personality Changes During Adolescence Across Cultures; Self and Identity Development During Adolescence across Cultures; Stress in Adolescence: Effects on Development.

Bibliography

American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders, fourth ed., text rev. Author, Washington, DC.

Arnett, J.J., 1999. Adolescent storm and stress, reconsidered. American Psychologist 54 (5), 317–326.

Bergeron, N., Schneider, B.H., 2005. Explaining cross-national differences in peer-directed aggression: a quantitative synthesis. Aggressive Behavior 31 (2), 116–137.

Blakemore, S.J., Choudhury, S., 2006. Development of the adolescent brain: implications for executive function and social cognition. Journal of Child Psychology and Psychiatry 47 (3–4), 296–312.

Blum, R.W., Beuhring, T., Shew, M.L., Bearinger, L.H., Sieving, R.E., Resnick, M.D., 2000. The effects of race/ethnicity, income, and family structure on adolescent risk behaviors. American Journal of Public Health 90 (12), 1879–1984.

Bonnie, R.J., 2002. Political abuse of psychiatry in the Soviet Union and in China: complexities and controversies. The Journal of the American Academy of Psychiatry and the Law 30 (1), 136–144.

Brown, B.B., Larson, R.W., 2002. The kaleidoscope of adolescence: experiences of the world's youth at the beginning of the 21st century. In: Brown, B.B., Larson, R.W., Saraswathi, T.S. (Eds.), The World's Youth: Adolescence in Eight Regions of the Globe. Cambridge University Press, Cambridge, pp. 1–20.

Canino, G., Polanczyk, G., Bauermeister, J.J., Rohde, L.A., Frick, P.J., 2010. Does the prevalence of CD and ODD vary across cultures? Social Psychiatry and Psychiatric Epidemiology 45 (7), 695–704.

Casey, B.J., Jones, R.M., Hare, T.A., 2008. The adolescent brain. Annals of the New York Academy of Sciences 1124 (1), 111–126.

Chiao, J.Y., Blizinsky, K.D., 2010. Culture-gene coevolution of individualism-collectivism and the serotonin transporter gene. Proceedings of the Royal Society B: Biological Sciences 277 (1681), 529–537.

Choudhury, S., Kirmayer, L.J., 2009. Cultural neuroscience and psychopathology: prospects for cultural psychiatry. Progress in Brain Research 178, 263–283.

Choudhury, S., 2010. Culturing the adolescent brain: what can neuroscience learn from anthropology? Social Cognitive and Affective Neuroscience 5 (2–3), 159–167.

Claes, M., Lacourse, E., Ercolani, A., Pierro, A., Leone, L., Presaghi, F., 2005. Parenting, peer orientation, drug use, and antisocial behavior in late adolescence: a cross-national study. Journal of Youth and Adolescence 34 (5), 401–411.

Conrad, P., 1992. Deviance and Medicalization: From Badness to Sickness. Temple University Press, St. Louis, MO.

Davis, L.J., Morris, D.B., 2007. Biocultures manifesto. New Literary History 38, 411–418.

Franzese, R.J., 2009. The Sociology of Deviance: Differences, Tradition, and Stigma. Charles C. Thomas, Springfield, IL.

Freeman, J.D., 1983. Margaret Mead and Samoa: The Making and Unmaking of an Anthropological Myth. Harvard University Press, Cambridge, MA.

Fridrich, A.H., Flannery, D.J., 1995. The effects of ethnicity and acculturation on early adolescent delinquency. Journal of Child and Family Studies 4 (1), 69–87.

Gelman, A., Fagan, J., Kiss, A., 2007. An analysis of the New York City Police Department's "stop-and-frisk" policy in the context of claims of racial bias. Journal of the American Statistical Association 102 (479), 813–823.

German, M., Gonzalez, N.A., Dumka, L., 2009. Familism values as a protective factor for Mexican-origin adolescents exposed to deviant peers. Journal of Early Adolescence 29 (1), 16–42.

Glaze, L.E., 2011. Correctional Populations in the United States, 2010 (Report No. NCJ 236319). Retrieved from Bureau of Justice Statistics: http://bjs.ojp.usdoj.gov/content/pub/pdf/cpus10.pdf.

Hall, G.S., 1904. Adolescence: Its Psychology and Its Relations to Physiology, Anthropology, Sociology, Sex, Crime, Religion, and Education. D. Appleton and Company, New York.

Hill, R.F., Fortenberry, J.D., 1992. Adolescence as a culture-bound syndrome. Social Science and Medicine 35 (1), 73–80.

Hinshaw, S.P., Lee, S.S., 2003. Conduct and oppositional defiant disorders. In: Mash, E.J., Barkley, R.A. (Eds.), Child Psychopathology. Guilford Press, New York, pp. 144–198.

Jessor, R., Turbin, M.S., Costa, F.M., Dong, Q., Zhang, H., Wang, C., 2003. Adolescent problem behavior in China and the United States: a cross-national study of psychosocial protective factors. Journal of Research on Adolescence 13 (3), 329–360.

Le, T.N., Stockdale, G.D., 2005. Individualism, collectivism, and delinquency in Asian American adolescents. Journal of Clinical Child and Adolescent Psychology 34 (4), 681–691.

Li, Y., Wang, M., Wang, C., Shi, J., 2010. Individualism, collectivism, and Chinese adolescents' aggression: intracultural variations. Aggressive Behavior 36, 187–194.

Lock, M., Nguyen, V., 2010. An Anthropology of Biomedicine. Wiley-Blackwell, Oxford.

Markus, H.R., Kitayama, S., 1991. Culture and the self: implications for cognition, emotion, and motivation. Psychological Review 98 (2), 224–253.

Marsella, A., 1987. The measurement of depressive experience and disorder across cultures. In: Marsella, A., Hirschfeld, R.M., Katz, M.M. (Eds.), The Measurement of Depression. Guilford Press, New York, pp. 376–397.

Mead, M., 1928. Coming of Age in Samoa: A Psychological Study of Primitive Youth for Western Civilization. W. Morrow & Company, New York.

Munroe, R.L., Hulefeld, R., Rodgers, J.M., Tomeo, D.L., Yamazaki, S.K., 2000. Aggression among children in four cultures. Cross-Cultural Research 34 (1), 3–25.

Nelson, E.E., Leibenluft, E., McClure, E., Pine, D.S., 2005. The social re-orientation of adolescence: a neuroscience perspective on the process and its relation to psychopathology. Psychological Medicine 35 (02), 163–174.

Pineda, D.A., Kamphaus, R.W., Restrepo, M.A., Puerta, I.C., Arango, C.P., Lopera, F.J., García-Barrera, M.A., Dowdy, E.T., 2006. Screening for conduct disorder in an adolescent male sample from Colombia. Transcultural Psychiatry 43 (3), 362–382.

Potter, N.N., 2011. Mad, bad, or virtuous? The moral, cultural, and pathologizing features of defiance. Theory & Psychology 22 (1), 23–45.

Santelli, J.S., Lowry, R., Brener, N.D., Robin, L., 2000. The association of sexual behaviors with socioeconomic status, family structure, and race/ethnicity among US adolescents. American Journal of Public Health 90 (10), 1582–1588.

Schlegel, A., Barry, H., 1991. Adolescence: An Anthropological Inquiry. Free Press, New York.

Shankman, P., 2009. The Trashing of Margaret Mead: Anatomy of an Anthropological Controversy. University of Wisconsin Press, Madison, WI.

Steffensmeier, D., Ulmer, J., Kramer, J., 2006. The interaction of race, gender, and age in criminal sentencing: the punishment cost of being young, black, and male. Criminology 36 (4), 763–798.

Triandis, H.C., 1989. Cross-cultural studies of individualism and collectivism. In: Berman, J. (Ed.), Nebraska Symposium on Motivation: Cross Cultural Perspectives. University of Nebraska Press, Lincoln, pp. 41–133.

Triandis, H.C., 1995. Individualism and Collectivism: New Directions in Social Psychology. Westview Press, Boulder, CO.

Tylor, E.B., 1871. Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Art, and Custom. Bradbury, Evans, and Co., London.

Tyson, G.A., Hubert, C.J., 2012. Cultural differences in adolescents' perceptions of the seriousness of delinquent behaviours. Psychiatry, Psychology and Law 10 (2), 316–323.

Vazsonyi, A.T., 2003. Parent–adolescent relations and problem behaviors: Hungary, the Netherlands, Switzerland, and the United States. Marriage & Family Review 35 (3–4), 161–187.

Weinberger, D.R., Elvevag, B., Giedd, J.N., 2005. The Adolescent Brain: A Work in Progress. National Campaign to Prevent Teen Pregnancy, Washington, DC.

Defiant Behavior During Infancy and Early Childhood

Sheryl L Olson and Adam S Grabell, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, we trace normative and clinically significant trends in the development of oppositional, defiant, and aggressive behavior across infancy, toddlerhood, and the preschool years. Defiant, noncompliant, and aggressive behavior is relatively common between the ages of 2–5 years. These problems decline steadily as children develop age-appropriate skills for regulating negative emotions and behavioral impulses. However, some preschool-age children manifest severe behavior problems that persist across later childhood and adolescence. We identify features of early behavior problems that signal clinical relevance and link individual differences in early behavioral adjustment to child–environment transactions that reflect risky patterns of temperament and caregiving.

Learning to establish cooperative social relationships is a critical developmental task for preschool-age children. Getting along with others requires a complex array of skills, particularly sharing attention and play materials with others, inhibiting aggressive and destructive impulses, following directions, responding appropriately to requests, and delaying gratification of immediate desires. Young children who fail to master these challenges, especially those who show chronic high levels of 'externalizing' symptoms such as defiance and aggression, are at risk of later academic failure; conflicted relationships with parents, siblings, peers, and teachers; delinquency; and substance abuse (e.g., Broidy et al., 2003). However, defiant and aggressive behavior is common in early childhood, reflecting normal maturational variations in the development of self-regulatory competence, communication skills, and social understanding (Tremblay, 2010). To understand the significance of early externalizing behavior, we must differentiate children's struggles with normal maturational challenges from precursors of serious problems. In what follows, we examine evidence that helps us differentiate normative expressions of problem behavior from precursors of persistent and serious problems. Next, we define the clinical parameters of early aggressive and oppositional defiant behavior. Finally, we highlight known risk factors and processes associated with persistent early-onset disruptive behavior problems, with an eye on prevention.

Normative Developmental Trends in Early Aggression and Temper Loss

In this section, we trace the early developmental progression of key behavioral markers of early-onset disruptive behavior disorders: aggression, temper loss, and oppositional defiant behavior. Which behavioral patterns differentiate normative from high-risk problem behavior?

Aggression

Relatively little is known about the development of aggressive behavior in the first 2 years of life. Available studies indicate that coercive use of force and frank physical aggression are fairly common behavior in toddler-age children (Alink et al., 2006; Tremblay et al., 1999). For example, in a large longitudinal study of normally developing children in day care (Study of Early Child Care and Youth Development; NICHD ECRRN, 2004), mothers rated the frequency of children's physical aggression at ages 24, 36, and 56 months and through the early school-age years. The most frequent form of early aggression, hitting others, occurred in 70% of the sample at ages 2 and 3, but declined to 20% at ages 4 and 5 and to 12% in the third grade. Other forms of disruptive behavior (e.g., 'destroying others' things') also showed marked declines across the preschool- and early school-age years (NICHD ECCRN, 2004).

Thus, coercive–aggressive behavior is common in early toddlerhood and peaks between ages 2 and 3 years, declining rapidly across the preschool period. By the time children enter kindergarten, overt aggressive behavior is largely inhibited (Tremblay, 2010). However, these normative trends summarize average behavioral tendencies across large groups of children. *Individual differences* in early aggressive behavior remain moderately stable across the preschool to school transition (Keenan et al., 2011). That is, although normally expected levels of aggression decline with development, a child's position in relation to his/her peers is fairly stable, even across major developmental transitions. Moreover, a subgroup of preschool-age children continues to show high levels of problem behavior across the school-age years (e.g., Broidy et al., 2003; NICHD ECCRN, 2004).

Temper Loss and Oppositional Defiant Behavior

Displays of anger and frustration are common among toddlers and preschoolers. Full-blown tantrums appear by the first year after birth, peak around 20 months, then decline across the preschool years (Tremblay, 2010). How do parents discern whether their child's temper outbursts signal the development of serious and long-lasting problems? Several studies have shown that normative temper outbursts are less frequent and prolonged than clinically concerning outbursts, and tend to primarily occur in the home setting (Belden et al., 2008; Osterman and Bjorkqvist, 2010). In addition, there are

important qualitative differences between normative and clinically significant forms of temper loss. For example, hitting, kicking, and/or biting during the tantrum; having a tantrum with no clear precipitating frustration; having a tantrum with adults other than parents; and showing explosive and extreme anger during the tantrum were found to be relatively rare among preschool-age children (Wakschlag et al., 2012). These nonnormative temper loss behavior may indicate serious adjustment problems that persist across the transition to school.

Finally, oppositional behavior appears in the toddler period and is common during the preschool- and early school-age years. Typical problem behavior includes arguing back, refusing to comply with parents' requests, and testing the 'limits' of parental control. Oppositional defiant behavior declines in frequency across the primary grades, then increases again during the transition to puberty (Tremblay, 2010). Clinically significant manifestations of oppositional behavior are marked by the frequency, severity, and persistence of the child's oppositional behavior in relation to normally developing age peers (Keenan and Wakschlag, 2000). Oppositional defiant disorder (ODD) is the most common disruptive behavior disorder in young children. The core feature of ODD is a chronic pattern of negativistic, defiant, disobedient behavior toward authority figures, usually the child's parents (American Psychiatric Association, 2013). Children with ODD tend to show frequent temper tantrums, often refuse to comply with their parent's requests, may make deliberate attempts to annoy others, and often react to frustration with hostility or aggression.

Preschoolers with ODD are at elevated risk for developing conduct disorder (CD) in later childhood. For example, approximately 30–50% of children with diagnoses of ODD progress to later CD (Waldman and Lahey, 2013). CD is defined by a chronic, age-inappropriate pattern of behavior that violates the rights of others and breaks major social rules (American Psychiatric Association, 2013). Children with CD often behave aggressively toward people and animals: they may initiate frequent fights, bully others, or torture animals. In some cases, acts of aggression include rape, assault with a deadly weapon, or homicide. Children with CD also may engage in destructive behavior that results in loss or damage to someone's property. For example, they may vandalize public buildings, set fires, or break up furniture in the family home. Deceitfulness – as evidenced by chronic lying, breaking promises, or stealing – is frequently evident. Finally, children with CD tend to violate important rules set by parents and school officials; for example, they may stay out late without permission or repeatedly fail to attend school without the knowledge of their parents.

Until recently, CD was thought to be rare among preschool-age children. However, studies featuring developmentally appropriate diagnostic assessment tools have shown that CD can be reliably diagnosed in 3- to 5-year-old children (Keenan et al., 2011). Moreover, Keenan et al. found that 80% of preschoolers who met criteria for ODD and 60% of those who met criteria for CD continued to receive the same diagnoses during the next 36 months. Boys were more likely than girls to follow a chronic course. Thus, it is essential to understand the types and combinations of risk factors that increase young children's vulnerability to persistent disruptive behavior problems.

Patterns of Temperament Risk

Early childhood is a time of critical development in self-regulation (Kopp, 1982). Biologically based differences in behavioral tempo and emotional reactivity may have an important impact upon the development of regulatory competence, especially if the child's interaction style leads to conflicts with caregivers and peers (e.g., Rothbart and Bates, 2006). In what follows, we describe two temperament traits that have been linked to individual differences in child externalizing behavior: regulation of attention and inhibitory behavior, and emotion reactivity and regulation, particularly proneness to anger.

Effortful control (Rothbart, 1989) refers to a general executive component of temperament that governs the child's capacity to inhibit a dominant response (e.g., grabbing a toy from a peer) and initiate a subdominant response (asking for a turn to play). The development of effortful control is governed by the maturation of the anterior attention network (Rothbart et al., 2000). Normally, developing levels of effortful control provide a foundation for adaptive socioemotional functioning, including complying with requests (Kopp, 2002), theory of mind understanding (Carlson and Moses, 2001), and normal conscience development (Kochanska et al., 1997). Conversely, toddler- and preschool-age children with deficient effortful control have shown high levels of disruptive behavior problems (e.g., Olson et al., 2005). Thus, deficits in effortful control play a key role in the development of early-onset disruptive behavior problems.

Early-onset externalizing problems also reflect deficits in multiple emotion systems (e.g., Keenan, 2000). The role of negative emotionality is particularly important. Both internalizing and externalizing problems have been linked to global constructs of early negative emotionality such as 'difficultness' or 'proneness to distress' in infancy and toddlerhood (Rothbart and Bates, 2006). However, current models of temperament and psychopathology stress the importance of differentiating distinct emotion systems as potential contributors to specific dimensions of psychopathology (e.g., Eisenberg et al., 2001). Mounting evidence suggests that proneness to frustration and anger is a specific risk factor for externalizing spectrum disorders. For example, young children who show angry responses to frustration also manifest higher levels of aggressive/disruptive peer interactions (Hughes et al., 2001) as well as generalized externalizing problems (Calkins and Dedmon, 2000).

Moreover, distinct *subtypes* of early externalizing problems may reflect contrastive patterns of abnormal emotion processes. A growing body of research suggests that some infants and toddlers at risk of persistent externalizing problems manifest abnormally *low* levels of arousal (Keenan, 2000). Studies of school-age children and adolescents have shown that high levels of 'callous, unemotional traits' such as low empathy, indifference to approval, lack of guilt, superficial affective displays are associated with serious and long-lasting conduct problems (Frick and Morris, 2004). To date, however, relatively little is known about the early developmental precursors of callous unemotional traits.

Individual differences in early regulatory competence have been correlated with distinctive patterns of neurobiological and genetic risk factors. Evidence from event-related potential studies suggests that typically developing preschool children recruit the frontal and anterior cingulate cortex when completing emotion regulation and effortful control tasks (Lewis et al., 2007). Conversely, preschool children with disruptive behavior disorders have shown less right-lateralized frontal lobe activation patterns during a resting state (Baving et al., 1999, 2000), suggesting abnormal cortical activation patterns. In addition, specific genes related to dopamine and serotonin systems have been implicated in externalizing and self-regulation problems, notably the DAT1, DRD4, and 5-HTT genes (e.g., Lavigne et al., 2013), although replication efforts have been inconsistent. Recent evidence suggests that parenting quality plays a critical moderating role in the relation between allele variants and emerging effortful control (Sheese et al., 2012; Smith et al., 2012).

Co-Occurring Developmental Influences

Early childhood is a time of rapid development in all domains of adaptive functioning. Integration across levels of children's cognitive, social, and emotional functioning is a dynamic, reorganizing process underpinned by increasing levels of cognitive maturation (e.g., Lewis, 2000). Hence, we cannot understand how early regulatory competencies contribute to children's behavioral adjustment without also considering these powerful co-occurring developmental influences. In this section, we briefly examine early developments in child social-cognitive and verbal functioning that have been linked with the development of externalizing problems.

Early Deficits in Social Cognition

Delays in social cognitive understanding have been posited to play a major role in the development of early-onset and later childhood aggression (Lemerise and Arsenio, 2000). Furthermore, early self-regulatory competence and social understanding are closely intertwined (Carlson et al., 2004). As with the development of self-regulation, children's understanding of themselves and others changes rapidly across early childhood and individual differences become quite salient (Wellman et al., 1995). One important set of competencies has been labeled 'theory of mind.' Between the ages 3 and 6, children develop an increased awareness that mental states are internal, subjective experiences distinct from behavior and contexts associated with them (Wellman et al., 2001). Examining these developments in children at risk of psychopathology is a natural and important extension of this work, providing a possible link between early mental representations and later impairments in moral understanding and conduct. For example, even after controlling for differences in verbal intelligence quotient, aggressive/disruptive toddlers and preschoolers have been found to show delays in theory of mind understanding (Hughes and Ensor, 2006; Olson et al., 2011a).

As children mature, more complex patterns of social information processing such as hostile attributional biases become important antecedents of aggressive responding (e.g., Dodge, 2006). However, there are major gaps in our knowledge of how social cognitive vulnerabilities combine with self-regulatory processes in the development of aggressive/disruptive behavior (Arsenio and Lemerise, 2004). Also, pathways between early social cognitive understanding and later *distorted* information processing are poorly understood. For example, under what conditions does delayed socioemotional understanding predispose young children to biased (hostile) social information processing in later childhood? Recent findings from our laboratory have revealed that children with relatively advanced theory of mind and emotion understanding skills at the age of 3 years tend to manifest extremely low levels of hostile attribution biases in the early school-age years (Choe et al., 2013). Thus, mature social cognitive skills in early childhood may protect children from developing dysfunctional social cognitive schemata following the transition to school.

Early Verbal Deficits

Individual differences in toddlers' cognitive/verbal maturity have been linked with their emerging abilities to regulate impulsivity (Olson et al., 2002) and strong negative affect (NICHD ECCRN, 2004). These deficits, in turn, place children at risk for conflicted interactions with caregivers and peers. Alternatively, protective social mechanisms may be 'driving' associations between early verbal skills and self-regulation. For example, Olson et al. (2002) found that toddlers who experienced enriched verbal communication with primary caregivers had more advanced cognitive and linguistic skills than others and showed superior behavioral regulation at ages 6 and 8 years. Therefore, the quality of the child's transactions with family and peers is critical to our understanding of *how* early intraindividual assets and vulnerabilities become translated into enduring patterns of social adjustment. In what follows, we outline a model for understanding the development of early regulatory competence through interpersonal transactions.

Interpersonal and Transactional Processes

Most 'difficult' toddlers and preschoolers do not manifest behavior problems in later childhood (Rothbart and Bates, 2006). The quality of the child's early relationships with family and peers is critical to our understanding of why some young children with challenging temperaments show persistent disruptive behavior, whereas others do not. As shown below, the genesis of child behavior problems reflects transactional processes linking parent–child interaction to the development of self-regulation (Sameroff, 2009).

Caregiver–Infant Synchrony

Infants' relationships with primary caregivers have been construed as 'primary learning environments' for the development of self-regulatory competence (Field, 1994). Because infants are wholly dependent upon caregivers, the quality of the caregiving environment is thought to play a critical role

in determining whether early vulnerabilities find expression in chronic self-regulation difficulties (Sroufe, 1996). Through infancy and early childhood, there is a general progression from reliance on caregivers for regulation of arousal to self-initiated modulation of affective reactions and behavioral impulses (Sameroff, 2009). Thus, sensitive caregiving, a major component of secure parent–infant attachment relationships, is thought to help the child modulate states of negative arousal, permitting positive social exchanges (Sroufe, 1996). Dysregulation, marked by the presence of dysynchronous interaction and emotional distress in one or both partners, can result from a variety of factors including insensitive caregiving (e.g., intrusiveness; overstimulation of the infant), physical disruptions in caregiving, and endogenous child characteristics such as developmental immaturity or irritable temperament (Keenan, 2000).

In prospective longitudinal studies, unresponsive mother–infant interaction has been linked to self-regulatory problems across middle childhood and adolescence (Lahey et al., 2008; Olson et al., 2000). Conversely, sensitive, responsive maternal care has been associated with the development of positive self-regulation (Feldman et al., 1999). For example, Feldman et al. (1999) measured individual differences in maternal responsiveness to microshifts in infant affect at 3 and 9 months, and found that synchronous interaction predicted self-regulatory competence at age 2, even when the effects of infant temperament and infant cognitive maturity were controlled.

Early Disciplinary Interactions

Achieving a sense of independent mastery is a major developmental challenge for toddlers. In this period of rapid cognitive, social, and motor growth, it is normal for toddlers to test parental limits and resist control. Thus, toddlerhood marks a critical transition point in the development of self-regulation (Kochanska et al., 2001). The manner in which parents and toddlers resolve these developmental challenges may forecast continuing adjustment problems. During the second year transition phase, regulatory competence is facilitated by caregiver sensitivity to the child's preferred style of engaging the social and physical environment (Bernier et al., 2010). For example, skillful caregivers may be able to channel a highly active toddler's behavior into constructive interactions. Given that, levels of child noncompliance peak between the second and third years of life, the ways in which caregivers interpret and respond to their toddlers' challenging behavior have important implications for understanding children's long-term behavioral adjustment (Olson et al., 2009)

In many previous reports, highly directive, negatively toned parent–toddler interactions have been associated with noncompliance and other early manifestations of self-control problems (e.g., Combs-Ronto et al., 2009). Early conflicts around noncompliance may place children at risk of escalating cycles of coercive interactions with parents and siblings, perhaps eventuating in later externalizing behavior. Intriguingly, however, hard-to-manage toddlers with skilled caregivers may have somewhat better adjustment outcomes than others. For example, Mesman et al. (2009) found that toddlers who were difficult to manage and had responsive, warm caregivers developed *higher than average* levels of behavioral adjustment in later childhood.

What mechanisms underlie associations between early parent–child interaction, child regulatory difficulties, and emerging behavior problems? In a longitudinal study, toddlers who showed evidence of affect dysregulation (strong defiance) with their mothers during laboratory play and forced compliance tasks were found to be at elevated risk of poor cognitive and social functioning in the early school-age years (NICHD ECCRN, 2004). Although early manifestations of emotion dysregulation were associated with multiple child, maternal, and environmental risk factors, these long-term associations held even when controlling for co-occurring early risks. Several findings offered important clues to underlying processes. First, surprisingly, toddlers in the high-risk group had not been identified by mothers as more temperamentally difficult than others in infancy. Second, the authors identified mothers who were in the lowest quartile of caregiving competence between ages 6 and 36 months; only 24% of their children were in the 'dysregulated' group. Finally, mothers in the affect dysregulation group reported feeling less close to their toddlers than others. Clearly, high-risk dyadic patterns reflect a mismatch between parent and child; predictions from child or parenting characteristics alone would be insufficient.

The nature of these social regulatory processes, in turn, may be differentially patterned in boys and girls. For example, previously we reported that low levels of effortful control were strongly associated with early-onset disruptive behavior in young preschool-age boys and girls (Olson et al., 2005). Subsequently, we have shown that poor effortful control is linked with troubled family and peer relationships in both sexes. However, significant pathways between early self-regulatory difficulties, parenting behavior, and school-age externalizing problems were found for boys only, suggesting gender-differentiated patterns of risk and promotive mechanisms (Chang et al., 2009).

In addition, a host of *contextual* factors affect the manner in which the child is perceived and responded to by other caregivers, and also may directly affect the child's ability to develop internal controls. These include factors such as parental depression (Choe et al., 2013), parental sensitivity to rejection (Callender et al., 2012), and high levels of marital discord, especially the presence of physically and emotionally aggressive responses to marital conflict (Cummings et al., 2003). Moreover, sociodemographic risks such as low parental education, poverty, community violence are powerful indicators of the amount of stress and strain on caregiving systems that predict serious long-term antisocial outcomes even before the child is born (Murray et al., 2010). Finally, as shown below, cultural factors are of paramount importance to understanding the development of early behavioral adjustment.

Cultural Foundations of Early Behavioral Adjustment

In this section, we briefly discuss how culturally shaped parental beliefs and disciplinary practices may influence the development of individual differences in young children's

self-regulatory competence, the primary foundation for behavioral adjustment. Cultural norms influence how caregivers define appropriate behavior in young children, their preferred methods of discipline, and the types of causal attributions they make when children violate social expectations, as well as the manner in which children express negative behavior (Bornstein and Lansford, 2009).

Parental beliefs about desirable child behavior reflect implicit meanings that derive in part from folk knowledge about the nature of children and their place in society. Parents' *ethnotheories* about child development have motivational properties and provide a frame of reference for interpreting and responding to child behavior (Harkness and Super, 2002). Most prior research connecting parenting beliefs with children's early behavioral adjustment has been conducted with North American samples. However, studies of countries with contrastive cultural values have strong relevance for understanding the socialization of self-regulation in young children (Tardif et al., 2009). For example, Olson et al. (2001) found that the two most salient categories of child maladjustment in Western populations, externalizing and internalizing problems, did not generalize well to the perceptions of Japanese parents of young children, whereas traits that reflected self-regulatory competence were paramount.

Similarly, Chinese parents have been shown to strongly endorse the value of promoting emotional restraint and impulse control in their children (Chen et al., 2005). Furthermore, because of a strong cultural emphasis upon maintaining social harmony, socially disruptive behavior is viewed negatively and prohibited in family and school settings (Chen et al., 2005). Thus, from young age, Chinese children are taught to suppress impulsivity, anger, and aggression (Wang et al., 2006). Conversely, cultural values concerning the importance of effortful behavior underpin Chinese parents' socialization practices around rigorous, early self-control training (Tardif et al., 2009). Practices described above predict differences in children's behavior across different cultural contexts. However, individual differences in the association between harsh discipline and child behavior *within* Chinese culture have been shown to parallel those of children growing up in the West. For example, studies have shown that Chinese parents' power assertive discipline (e.g., frequent use of prohibitions, direct commands without explanations) with their toddler-age children predicted high levels of child aggression 2 years later, whereas inductive discipline was associated with low levels of aggression (Wang et al., 2006).

Despite the importance of cultural influences on early behavioral adjustment, there has been relatively little research linking parenting and child risk factors with symptoms of externalizing behavior in non-Western preschoolers. A recent study showed that risk factors associated with disruptive behavior problems in Western preschoolers also predicted externalizing problems in children growing up in non-Western countries. Olson et al. (2011b) conducted a comparative cross-cultural study of disruptive behavior problems in preschool-age children growing up in China, Japan, and the United States. In all these three cultures, children who performed poorly on behavioral tests of impulse control and received high levels of corporal punishment showed higher levels of disruptive behavior problems than others. Thus, salient child and parental risk factors associated with the emergence of *individual differences* in disruptive behavior were similar across diverse cultural settings. Given the paucity of comparative research on this topic, these findings should be treated with caution. Much further study is needed.

Another important gap in research concerns the lack of attention given to the behavioral socialization of young children in developing countries. Recently, data from the Multiple Indicator Cluster Study (UNICEF, 2006), a nationally representative and internationally comparable household study of over 30 000 families, have revealed important similarities and differences in child socialization practices (Bornstein et al., 2012). Primary caregivers of 2- to 4-year-old children in 28 developing countries varied widely in caregiving practices known to promote children's executive skills. For example, only one-third of these caregivers read books or told stories to their young children (Bornstein and Putnick, 2012). Moreover, using data from 24 of these nations, Lansford and Deater-Deckard (2012) examined variations in disciplinary practices with 2- to 4-year-old children. Caregivers told interviewers whether they or the members of their household used disciplinary practices that included nonviolent strategies (e.g., reasoning, explaining), psychological aggression (e.g., yelling, screaming, calling the child stupid or lazy), physical violence (e.g., spanking on the bottom or extremities), and severe physical violence (hitting on the head, beating the child with an implement). Across all countries, the most frequently reported disciplinary technique was explaining to the child why certain behavior was undesirable. Use of severe violence showed the highest levels of variability across countries, ranging from 1 to 40%. National education levels were associated with lower levels of violent caregiving practices (Lansford and Deater-Deckard, 2012). Thus, educating parents about positive disciplinary strategies that promote the development of competent self-regulation is an important target for global prevention efforts.

Clinical Implications

A critical challenge for researchers is to understand the nature of the transactional processes that underlie the translation of early child vulnerabilities into stable adjustment problems. Early self-regulatory skills reflect reciprocal influences between the child and his/her social partners as well as the larger social and cultural environment (Sameroff, 2009). The fact that stable disruptive behavior problems emerge during the toddler and preschool years strongly indicates the need for early intervention. Not surprisingly, interventions focused on helping parents respond positively and effectively to provocative child behavior produces better outcomes in early childhood than interventions implemented in later childhood or adolescence (McCart et al., 2006). For example, Shelleby et al. (2012) found that a family-centered intervention (*early steps*) improved parental positive discipline and reduced behavior problems in high-risk toddlers and preschoolers. In future interventions, we may be able to directly target children's emerging effortful control and emotion regulation skills. For example, studies have shown

that we can significantly improve preschool children's executive functioning skills through repeated practice or by teaching them meta-cognitive skills (e.g., Diamond et al., 2007). Relatedly, mothers who employ frequent emotion coaching with their preschool children may protect them from developing externalizing problems when children are exposed to psychosocial risks, including exposure to traumatic stressors (Ellis and Alisic, 2013).

Summary

In summary, we have outlined a complex model of the development of early-onset disruptive behavior problems. Our core thesis is that early externalizing problems reflect failures in the normal establishment of self-regulatory competence, particularly temperament systems governing effortful control and anger regulation. Emerging research highlights the need to examine these intrachild risk markers in much greater detail, paying attention to co-occurring developments in other systems of temperament, social cognition, and communication competence. Perhaps most importantly, we must strive to understand processes underlying the *transformation* of early self-regulation problems into stable patterns of behavioral maladjustment. In our view, this requires examining complex transactions between the child, his/her caregiving system, and a broad range of contextual factors that affect the quality of the child's social experiences and learning.

See also: Behavioral Problems, Effects of Parenting and Family Structure on; Social and Emotional Development in the Context of the Family; Stress: Effects on Neonatal and Child Development.

Bibliography

Alink, L.R.A., Mesman, J., van Zeijl, J., Stolk, M.N., Juffer, F., Koot, H., Bakermans-Kranenburg, M.J., van IJzendoorn, M.H., 2006. The early childhood aggression curve: development of physical aggression in 10- to 50-month-old children. Child Development 77, 954–966.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Association, Washington, DC.

Arsenio, W.F., Lemerise, E.A., 2004. Aggression and moral development: integrating social information processing and moral domain models. Child Development 75 (4), 987–1002.

Baving, L., Laucht, M., Schmidt, M.H., 1999. Atypical frontal brain activation in ADHD: preschool and elementary school boys and girls. Journal of the American Academy of Child & Adolescent Psychiatry 38 (11), 1363–1371.

Baving, L., Laucht, M., Schmidt, M.H., 2000. Oppositional children differ from healthy children in frontal brain activation. Journal of Abnormal Child Psychology 28 (3), 267–275.

Belden, A., Thompson, N., Luby, J., 2008. Temper tantrums in healthy versus depressed and disruptive preschoolers: defining tantrum behaviors associated with clinical problems. Journal of Pediatrics 152, 117–122.

Bernier, A., Carlson, S.M., Whipple, N., 2010. From external regulation to self-regulation: early parenting precursors of young children's executive functioning. Child Development 81 (1), 326–339.

Bornstein, M.H., Lansford, J.E., 2009. Parenting. In: Bornstein, M.H. (Ed.), The Handbook of Cross-cultural Developmental Science. Taylor & Francis, New York, pp. 259–277.

Bornstein, M.H., Britto, P.R., Nonoyama-Tarumi, Y., Ota, Y., Petrivic, O., Putnick, D.L., 2012. Child development in developing countries: introduction and methods. Child Development 83, 16–31.

Bornstein, M.H., Putnick, D.L., 2012. Cognitive and socioemotional caregiving in developing countries. Child Development 83, 46–61.

Broidy, L.M., Nagin, D.S., Tremblay, R.E., Bates, J.E., Brame, B., Dodge, K.A., et al., 2003. Developmental trajectories of childhood disruptive behaviors and adolescent delinquency: a six-site, cross-national study. Developmental Psychology 39 (2), 222–245.

Bugental, D.B., Grusec, J., 2006. Socialization processes. In: Damon, W. (Series Ed), Eisenberg, N. (Volume Ed) (Eds.), Handbook of Child Psychology, Social, emotional and personality development, sixth ed., vol. 3. Wiley, New York, pp. 366–428.

Calkins, S.D., Dedmon, S.E., 2000. Physiological and behavioral regulation in two-year-old children with aggressive/destructive behavior problems. Journal of Abnormal Child Psychology 28, 103–118.

Callender, K., Olson, S.L., Choe, D., Sameroff, A.J., 2012. The effects of parental depression, appraisals, and discipline on later externalizing behavior. Journal of Abnormal Child Psychology 40 (3), 471–483.

Carlson, S.M., Mandell, D.J., Williams, L., 2004. Executive function and theory of mind: stability and prediction from ages 2 to 3. Developmental Psychology 40 (6), 1105–1122.

Carlson, S.M., Moses, L.J., 2001. Individual differences in inhibitory control and children's theory of mind. Child Development 72 (4), 1032–1053.

Chang, H., Olson, S.L., Combs-Ronto, L., Sameroff, A.J., 2009. Child effortful control as a mediator of parenting practices on externalizing behavior: evidence for a sex-differentiated pathway across the transition from preschool to school. Journal of Abnormal Child Psychology 39 (1), 71–81.

Chen, X., Cen, G., Li, D., He, Y., 2005. Social functioning and adjustment in chinese children: the imprint of historical time. Child Development 76 (1), 182–195.

Choe, D.E., Lane, J.D., Grabell, A.S., Olson, S.L., 2013. Developmental precursors of young school-age children's hostile attribution bias. Developmental Psychology 49 (12), 2245–2256.

Choe, D., Olson, S.L., Sameroff, A.J., 2013. Effects of early maternal distress on parenting and the development of children's self-regulation and externalizing behavior problems. Development and Psychopathology 25 (2), 437–453.

Cole, P.M., Teti, L.O., Zahn-Waxler, C., 2003. Mutual emotion regulation and the stability of conduct problems between preschool and school age. Development and Psychopathology 15, 1–18.

Combs-Ronto, L., Olson, S.L., Lunkenheimer, E.S., Sameroff, A.J., 2009. Interactions between maternal parenting and children's early disruptive behavior: bidirectional associations across the transition from preschool to school. Journal of Abnormal Child Psychology 37, 1151–1163.

Cummings, E.M., Goeke-Morey, M.C., Papp, L.M., 2003. Children's responses to everyday marital conflict tactics in the home. Child Development 74 (6), 1918–1929.

Diamond, A., Barnett, W.S., Thomas, J., Munro, S., 2007. Preschool program improves cognitive control. Science 318, 1387–1388.

Dodge, K.A., 2006. Translational science in action: hostile attributional style and the development of aggressive behavior problems. Development and Psychopathology 18 (3), 791–814.

Ellis, B.H., Alisic, E., 2013. Maternal emotion coaching: a protective factor for traumatized children's emotion regulation? Journal of Child & Adolescent Trauma 6 (2), 118–125.

Eisenberg, N., Cumberland, A., Spinrad, T., Fabes, L.R., Shepard, S., Reiser, M., et al., 2001. The relations of regulation and emotionality to children's externalizing and internalizing problem behavior. Child Development 72, 1112–1134.

Feldman, R., Greenbaum, C.W., Yirmiya, N., 1999. Mother-infant affect synchrony as an antecedent of the emergence of self-control. Developmental Psychology 35 (1), 223–231.

Field, T., 1994. The effects of mother's physical and emotional unavailability on emotion regulation. Monographs of the Society for Research in Child Development 59 (2–3), 208–227, 250–283.

Frick, P.J., Morris, A.S., 2004. Temperament and developmental pathways to conduct problems. Journal of Clinical Child and Adolescent Psychology 33, 54–68.

Harkness, S., Super, C., 2002. Culture and parenting. In: Bornstein, M. (Ed.), Handbook of Parenting, second ed. Erlbaum, Majwah, NJ, pp. 253–280.

Hughes, C., Cutting, A.L., Dunn, J., 2001. Acting nasty in the face of failure? Longitudinal observations of "hard-to-manage" children playing a rigged competitive game with a friend. Journal of Abnormal Child Psychology 29 (5), 403–416.

Hughes, C., Ensor, R., 2006. Behavioral problems in two-year-olds: links with individual differences in theory of mind, executive function, and negative parenting. Journal of Child Psychology and Psychiatry 47, 488–497.

Keenan, K., 2000. Emotion dysregulation as a risk factor for child psychopathology. Clinical Psychology: Science and Practice 7 (4), 418–434.

Keenan, K., Wakschlag, L., 2000. More than the terrible twos: the nature and severity of behavior problems in clinical referred preschool children. Journal of Abnormal Child Psychology 28 (1), 33–46.

Keenan, K., Boeldt, D., Chen, D., Coyne, C., Donald, R., Duax, J., Hart, K., Perrott, J., Strickland, J., Davis, B., Hill, C., Davis, S., Humphries, M., 2011. Predictive validity of DSM-IV oppositional defiant and conduct disorders in clinically referred preschoolers. Journal of Child Psychology and Psychiatry 52 (1), 47–55.

Kochanska, G., Coy, K.C., Murray, K.T., 2001. The development of self-regulation in the first four years of life. Child Development 72 (4), 1091–1111.

Kochanska, G., Murray, K., Coy, K.C., 1997. Inhibitory control as a contributor to conscience in childhood: from toddler to early school age. Child Development 68, 263–277.

Kopp, C.B., 1982. Antecedents of self-regulation: a developmental perspective. Developmental Psychology 18, 199–214.

Kopp, C.B., 2002. Commentary: the co-developments of attention and emotion regulation. Infancy 3 (2), 199–208.

Lahey, B.B., Van Hulle, C.A., Keenan, K., Rathouz, P.J., D'Onofrio, B.M., Rodgers, J.E., Waldman, I.D., 2008. Temperament and parenting during the first year of life predict future child conduct problems. Journal of Abnormal Child Psychology 36 (8), 1139–1158.

Lansford, J.E., Deater-Deckard, K., 2012. Childrearing discipline and violence in developing countries. Child Development 83, 62–75.

Lavigne, J.V., Herzing, L.B., Cook, E.H., Lebailly, S.A., Gouze, K.R., Hopkins, J., Bryant, F.B., 2013. Gene × environment effects of serotonin transporter, dopamine receptor D4, and monoamine oxidase A genes with contextual and parenting risk factors on symptoms of oppositional defiant disorder, anxiety, and depression in a community sample of 4-year-old children. Development and Psychopathology 25 (02), 555–575.

Lemerise, E.A., Arsenio, W.F., 2000. An integrated model of emotion processes and cognition in social information processing. Child Development 71 (1), 107–118.

Lewis, M.D., 2000. The promise of dynamic systems approaches for an integrated account of human development. Child Development 71 (1), 36–43.

Lewis, M.D., Todd, R.M., Honsberger, M.J., 2007. Event-related potential measures of emotion regulation in early childhood. NeuroReport 18 (1), 61–65.

McCart, M.R., Priester, P.E., Davies, W.H., Azen, R., 2006. Differential effectiveness of behavioral parent-training and cognitive-behavioral therapy for antisocial youth: a meta-analysis. Journal of Abnormal Child Psychology 34 (4), 525–541.

Mesman, J., Stoel, R., Bakermans-Kranenburg, M.J., vanIJzendoorn, M.H., Juffer, F., Koot, H.M., et al., 2009. Predicting growth curves of early childhood externalizing problems: differential susceptibility of children with difficult temperament. Journal of Abnormal Child Psychology 37, 625–636.

Moffitt, T.E., Caspi, A., Rutter, M., Silva, P., 2001. Sex Differences in Antisocial Behaviour: Conduct Disorder, Delinquency, and Violence in the Dunedin Longitudinal Study. Cambridge University Press, New York.

Murray, L., Arteche, A., Fearson, P., Halligan, S., Crondace, T., Cooper, P., 2010. The effect of maternal postnatal depression and child sex on academic performance at age 16 years: a developmental approach. Journal of Child Psychology and Psychiatry 51 (10), 1150–1159.

NICHD Early Child Care Research Network, 2004. Affect dysregulation in the mother-child relationship in the toddler years: antecedents and consequences. Development and Psychopathology 16, 43–68.

Olson, S.L., Bates, J.E., Bayles, K., 1990. Early antecedents of childhood impulsivity: the role of parent-child interaction, cognitive competence, and temperament. Journal of Abnormal Child Psychology 18, 317–334.

Olson, S.L., Bates, J.E., Sandy, J.M., Lanthier, R., 2000. Early developmental precursors of externalizing behavior in middle childhood and adolescence. Journal of Abnormal Child Psychology 28, 119–133.

Olson, S.L., Bates, J.E., Sandy, J.M., Schilling, E.M., 2002. Early developmental precursors of impulsive and inattentive behavior: from infancy to middle childhood. Journal of Child Psychology and Psychiatry 43, 435–447.

Olson, S.L., Kashiwagi, K., Crystal, D., 2001. Concepts of adaptive and maladaptive child behavior: a comparison of U.S. and Japanese mothers of preschool-age children. Journal of Cross-Cultural Psychology 32, 43–57.

Olson, S.L., Lopez-Duran, N.L., Chang, H., Lunkenheimer, E.S., 2011a. Individual differences in the development of early peer aggression: integrating contributions of self-regulation, theory of mind, and parenting. Development and Psychopathology 23 (1), 253–266.

Olson, S.L., Tardif, T., Miller, A., Felt, B., Grabell, A., Kessler, D., Wang, L., Karasawa, M., Hirabayashi, H., 2011b. Executive function and parenting as predictors of externalizing problems in young children: a cross-national study of U.S., Chinese and Japanese preschoolers. Journal of Abnormal Child Psychology 39, 1163–1175.

Olson, S.L., Sameroff, A.J., Kerr, D.C.R., Lopez, N.L., Wellman, H.M., 2005. Developmental foundations of externalizing problems in young children: the role of effortful control. Development and Psychopathology 17 (1), 25–45.

Olson, S.L., Sameroff, A.J., Kerr, D.C.R., Lunkenheimer, E.S., 2009. Self-regulatory processes in early disruptive behavior: the preschool to school transition. In: Olson, S.L., Sameroff, A.J. (Eds.), Regulatory Processes in the Development of Childhood Behavior Problems: Biological, Behavioral, and Social-ecological Interactions. Cambridge University Press, New York.

Osterman, K., Bjorkqvist, K., 2010. A cross-sectional study of onset, cessation, frequency, duration of children's temper tantrums in a nonclinical sample. Psychological Reports 106, 448–454.

Posner, M.I., Rothbart, M.K., 2000. Developing mechanisms of self-regulation. Development and Psychopathology 12, 427–441.

Rothbart, M.K., 1989. Temperament and development. In: Kohnstamm, G.A., Bates, J.E., Rothbart, M.K. (Eds.), Temperament in Childhood. Wiley, New York, pp. 187–247.

Rothbart, M.K., Posner, M.I., Boylan, A., 2000. Regulatory mechanisms in infant development. In: Enns, J.T. (Ed.), The Development of Attention: Research and Theory. Oxford, North-Holland, England, pp. 47–66.

Rothbart, M.K., Bates, J.E., 2006. Temperament. In: Damon, W., Eisenberg, N. (Eds.), Handbook of Child Psychology, Social, Emotional, and Personality Development, sixth ed., vol. 3. Wiley, New York.

Sameroff, A.J., 2009. Transactional Processes in Development. American Psychological Association, Washington, DC.

Sheese, B.E., Rothbart, M.K., Voelker, P.M., Posner, M.I., 2012. The dopamine receptor D4 gene 7-repeat allele interacts with parenting quality to predict effortful control in four-year-old children. Child Development Research 19, 1039–1046.

Shelleby, E.C., Shaw, D.S., Cheong, J., Chang, H.C., Gardner, F., Dishion, T., Wilson, M., 2012. Behavioral control in at-risk toddlers: the influence of the family check-up. Journal of Clinical Child and Adolescent Psychology 41 (3), 288–301.

Smith, H.J., Sheikh, H.I., Dyson, M.W., Olino, T.M., Laptook, R.S., Durbin, C.E., Hayden, E.P., Singh, S.M., Klein, D.N., 2012. Parenting and child DRD4 genotype interact to predict children's early emerging effortful control. Child Development 83 (6), 1932–1944.

Sroufe, L.A., 1996. Emotional Development: The Organization of Emotional Life in the Early Years. Cambridge University Press, New York.

Tardif, T., Wang, L., Olson, S.L., 2009. Culture and the development of regulatory competence: Chinese–U.S. comparisons. In: Olson, S.L., Sameroff, A.J. (Eds.), Regulatory Processes in the Development of Behavior Problems: Biological, Behavioral, and Social-ecological Interactions. Cambridge University Press, New York, pp. 258–289.

Torpey, D.C., Hajcak, G., Kim, J., Kujawa, A., Klein, D.N., 2012. Electrocortical and behavioral measures of response monitoring in young children during a Go/No-Go task. Developmental Psychobiology 54 (2), 139–150.

Tremblay, R.E., 2010. Developmental origins of disruptive behavior problems: the 'original sin' hypothesis, epigenetics, and their consequences for prevention. Journal of Child Psychology and Psychiatry 51 (4), 341–367.

Tremblay, R.E., Japel, C., Perusse, D., McDuff, P., Boivin, M., Zoccolillo, M., et al., 1999. The search for the age of 'onset' of physical aggression: Rousseau and Bandura revisited. Criminal Behaviour and Mental Health 9 (1), 8–23.

United Nations Children's Fund, 2006. Multiple Indicator Cluster Survey Manual 2005: Monitoring the Situation of Children and Women. United Nations Children's Fund, New York.

Waldman, I.D., Lahey, B.B., 2013. Oppositional defiant disorder, conduct disorder, and juvenile delinquency. In: Beauchaine, T.P., Hinshaw, S.P. (Eds.), Child and Adolescent Psychopathology. John Wiley & sons, NT.

Wakschlag, L., Choi, S.W., Carter, A., Hullsiek, H., Burns, J., McCarthy, K., Leibenluft, E., Briggs-Cowan, M., 2012. Defining the parameters of temper loss in early childhood: implications for developmental psychopathology. Journal of Child Psychology and Psychiatry 53 (11), 1099–1108.

Wakschlag, L.S., Hans, S.L., 1999. Relation of maternal responsiveness during infancy to the development of behavior problems in high-risk youths. Developmental Psychology 35 (2), 569–579.

Wang, L., Chen, X., Chen, H., Cui, L., Miao, L., 2006. Affect and maternal parenting as predictors of adaptive and maladaptive behaviors in Chinese children. International Journal of Behavioral Development 30 (2), 158–166.

Wellman, H.M., Cross, D., Watson, J., 2001. Meta-analysis of theory-of-mind development: the truth about false belief. Child Development 72 (3), 655–684.

Wellman, H.M., Harris, P.L., Banerjee, M., Sinclair, A., 1995. Early understanding of emotion: evidence from natural language. Cognition & Emotion 9 (2–3), 117–149.

Definition of the situation: History of the Concept

Rainer Egloff, Ludwik Fleck Center at Collegium Helveticum, University of Zurich and ETH Zurich, Zurich, Switzerland; and Swiss Reinsurance Company Ltd, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Coined by W.I. Thomas in the early twentieth century in the context of sociological studies on education, deviance, migration and race, the 'definition of the situation' referred to the interactive social psychological process of defining reality as a basis to determine individual and group behavior, particularly in new, unfamiliar, conflictive, or crisis situations. Defining a situation also involved actualization and modification of habits – learned behavior patterns, social values, and individual attitudes. Subsequently, the concept became an integral part of social research in the classical Chicago School of Sociology and later with its inheritors, particularly in symbolic interactionism and in the microtheories of Erving Goffman. Since the Second World War, the concept had also been adopted by functionalist sociology, most importantly through Robert K. Merton who explicitly based his own concept of the self-fulfilling prophecy on Thomas' concept, forging it into a middle-range theory. Subsequently, the concept has been used and differentiated in fields as diverse as ethnography of face-to-face interaction and rational choice theory. The original attempt to generalize and model human action in an overarching and integrative way still serves to bridge the various uses.

The concept of definition of the situation refers to the interactive mental process of determining meaning and subsequently behavior by and for individuals and groups in a given situation. An individual definition of the situation refers to the way in which an individual interprets a situation so that she or he can act (Blumer, 1939: 26). The concept emphasizes the actor's perspective, which in return is shaped by social and other factors. The definition of the situation thus integrates objective and subjective factors. The definition of situation mirrors the diverse motivations and restrictions, behavior patterns and personalities, habits, crises and readjustments, whereby people make sense of their environment and their interaction with one another. The factors to be involved range from biological aspects to biographical and spontaneous experiences, they include individual attitudes, cultural values and group norms, institutional settings and authorities.

The concept has now a history of almost a hundred years in the social sciences, and particularly in sociology. The concept of the definition of the situation was developed in the context of the pragmatist movement in American social science, and it became most influential in the classical Chicago School of Sociology, providing an analytical framework for understanding human behavior in general, and particularly in grasping adjustment difficulties of individuals and groups along their life histories and regarding new environments.

After the Second World War (WWII), functionalist strands within American sociology integrated the 'definition of the situation,' and the concept of the definition of the situation became adopted in mainstream sociology, whereas proponents of microscopic social theory and qualitative research (ethnography, case study, interview, biographical research, etc.) also further developed the concept. Later, the concept was used to bridge and integrate micro- and macrotheoretical approaches. Interdisciplinary at the very outset, the concept has been widely used in sociology and in related disciplines like political science, anthropology, psychology, pedagogy, criminology, and social work. While the concept is still intensely discussed, it has also become adapted so widely, that historical indebtedness to the concept has become less obvious in many cases.

W.I. Thomas and the Situational Approach in the Chicago School of Sociology

William Isaac Thomas (1863–1947), pioneer US sociologist and founding figure of the Chicago School of Sociology, is acknowledged to first have given expression to the concept of the definition of the situation (Perinbanayagam, 1974: 521). Florian Znaniecki, who coauthored the *Polish Peasant in Europe and America* (1918–22) with Thomas, and also Dorothy Swaine Thomas, Thomas's second wife, who wrote *The Child in America* (1928) together with him, are occasionally referenced as cofounders of the concept (Wolff, 1964; Merton, 1995).

Thomas developed the concept of *definition of the situation* in the mid 1910s, in the context of immigration studies and reform pedagogy. It highlighted the interactions between individuals, groups, institutions, and cultures to understand the rationale behind human decisions, actions, and their sanctions. The concept also represents a strong statement for the cultural and social relativity of rationality, values and norms.

The concept of definition of the situation appeared in print in 1917, in a collection of *Suggestions of Modern Science Concerning Education*, which was organized and edited by Thomas. Apart from Thomas' own contribution, the volume comprised essays by behaviorist psychologist J.B. Watson, zoologist and geneticist Herbert S. Jennings, and psychiatrist Adolf Meyer. With reference to Watson's treatment of emotional reactions, Thomas opened his chapter (Thomas, 1917) by stating four wishes – fundamental interests or desires – to represent human behavior. According to Thomas, the wishes for new experience, mastery, recognition, and safety were eventually reducible to food-hunger and sex-hunger, and their expression limited by human society through the formation of a code of behavior, which itself was produced and reproduced by an interactive process Thomas called 'the definition of the

situation': "The code therefore represents the judgment of society on the activities of its members, it dictates the limits within which the desires may find expression, and it is developed by a method which we may call 'the definition of the situation'" (Thomas, 1917: 167f.). Thomas saw the defining of the situation begun by the parents, who order and forbid and inform, "… continued in the community by means of gossip, with its and praise and blame, and […] formally represented by the school, the law, the church" (Thomas, 1917: 168). Thomas emphasized the social relativity of morality and immorality, and organization and disorganization, concluding: "In short, any definition, however arbitrary, that is embodied in the habits of the people is regarded as right" (Thomas, 1917: 169).

The role of the definition of the situation to capture potential tensions between differing perceptions and value orientations was further elaborated in *The Polish Peasant in Europe and America* (Thomas and Znaniecki, 1918–21). In this classic of empirical social research the adjustment process of an immigrant group was documented, mostly through 'life history documents' such as hundreds of letters and a book-long autobiography. The study assessed the ethnic group's potential for resolution of crisis arising from migration and social change through communicative interaction and subsequent reorientation. In the 'methodological note' to the five volumes, Thomas and Znaniecki introduced a model of activity to integrate social values – socially sanctioned objects of activity – on the one hand, and individual attitudes – the individuals' tendencies or potentialities to act. The 'situation' was conceptualized as the interacting set of values and attitudes, and every concrete activity was understood as the solution of a situation (Thomas and Znaniecki, 1918: 68). While the total situation involved (1) 'the totality of values' equaled with 'objective conditions under which the individual or society has to act,' and (2) the 'preexisting attitudes of the individual or the group which at the given moment have an actual influence upon his behavior,' the 'definition of the situation' (3) represented 'the more or less clear conception of the conditions' (1) and 'consciousness of the attitudes' (2): "(…) the definition of the situation is necessary preliminary to any act of the will, for in given conditions and with a given set of attitudes an indefinite plurality of actions is possible, and one definite action can appear only if these conditions are selected, interpreted, and combined in a determined way and if a certain systematization of these attitudes is reached, so that one of them becomes predominant and subordinates the others" (Thomas and Znaniecki, 1918: 68).

The concept of definition of the situation embodied a perspective of social psychology, understood as 'the general science of the subjective side of culture', and it was to be complemented by sociology understood as "theory of social organization" (Thomas and Znaniecki, 1918: 33). While the definition of situation represented the individual and social regulation of behavior and behavior reaction, it was particularly focused at processes of adjusting an attitude, a behavioral code or a value, after former habits had been disturbed and questioned by a crisis demanding a new definition of the situation. The five volumes of the *Polish Peasant* documented migration from one cultural setting to another as representing such crises on individual and group level. Thomas and Znaniecki applied the concept of the definition of the situation to analyze how well the immigrants adjusted their behavior from the traditional norms of rural Poland to the social values of the modern American metropolis.

Throughout Thomas's subsequent work the definition of the situation remained a core concept, whether applied to studies on child development and youth deviance (Thomas, 1923; Thomas and Thomas, 1928) or in comparative ethnology of culture (Thomas, 1937). The concept reflected Thomas's interest in behavior patterns conceived of in terms of stimulus and response, of habit and crisis, in developmental directions and personalities of individuals, groups, societies and cultures. Its dialectics of subjectivism and naturalism points to the roots of the concept in pragmatist philosophy (Joas, 1993) and to a desire to make social science an objective empirical science.

Thomas had originally been a college professor for language and literature, and he was well versed in philological and hermeneutic methods (Abbott and Egloff, 2008). But with his first steps into the new field of social science in the early 1890s, he also adopted theory and research attitudes from natural sciences to study human behavior and psychologies of cultures, groups, and individuals. An admirer of William James and influenced not only by John Dewey and George Herbert Mead, but also by anthropologist Franz Boas, he gradually moved away from an early emphasis on biological factors. Race and sex as prime categories of his research were soon replaced by a preference for the individual in his biography and personality, and he developed his culturalist social psychology embodied in the concept of the definition of the situation. Thomas's 'situational approach' translated the pragmatist stances that reality is always mediated and identities achieved through social interaction into empirical social science. Some tension and ambivalence between hermeneutic foundations and scientific ambitions remained with Thomas's conceptualization, leaving it rather undecided and vague (Blumer, 1939).

Thomas exerted a dominant influence on the Chicago School of Sociology of the 1920s and 1930s, and his concept of the definition of the situation was adopted by Robert Park, Ernest Burgess, and their colleagues and students, who partly had also been students with Thomas himself. In their seminal Chicago School textbook, Park and Burgess stated that "every single act, and eventually all moral life, is dependent upon the definition of the situation" (Park and Burgess, 1921: 764). They highlighted that the plurality of 'universes of discourse' modern life was characterized by, representing diverging definitions of situation – differences in language, meanings and values individuals attach to objects and behavior. They also maintained that "common participation in common activities implies a common, definition of the situation," pointing to the assimilation of immigrants into American society, and to the role of education on the one hand, and propaganda on the other. Park and Burgess were most interested in the adjustment and change of definitions of the situation, and this was always conceived as an interactive or mutual process: "A definition of the situation precedes and limits any possible action, and a redefinition of the situation changes the character of the action. An abusive person, for example, provokes anger and possibly violence, but if we realize that the man is insane this redefinition of the situation results in totally different behavior" (Park and Burgess, 1921: 764).

Immigration studies, educational questions, and criminology remained preferred fields for the Chicago school students' use of the concept of the definition of the situation. The concept of the definition of the situation was highlighted in the educational sociology of Willard Waller, which analyzed interaction in school, and differentiated elements, modes, and closures in the process of defining the situation (Waller, 1932: 292–317). Studies on delinquency by Clifford Shaw and others were centered around case studies, often single autobiographies – 'the delinquent's own story' – providing subjective life histories and following the format from Thomas and Znaniecki's *Polish Peasant* (Shaw, 1930). Frederic M. Thrasher and others related the development of individuals and groups (e.g., youth gangs) to the structures and characters of institutions, communities, and metropolitan areas, advocating 'the study of the total situation' (Thrasher, 1928). Later, Edwin H. Sutherland's differential association theory asserted that a person 'learned' to become delinquent due to social interaction and dominant exposition to definitions favorable to violation of law (Sutherland, 1947). The classic Chicago School's emphasis on the communicative and particularly verbal aspect of the concept is exemplified in Kimball Young's essay "Language, Thought and Social Reality" (Young, 1931) – part of a *festschrift* for Thomas by former students of Thomas on the topic of social attitudes and edited by Young in 1931. Young's article underlined the function of language to convey meaning and thus "both the objective and the subjective types of anticipation" (Young, 1931: 205). The symmetric structure of autistic fantasy associations and reality was underlined by an introductory motto to the essay taken from Thomas: "If men define situations as real, they are real in their consequences" (Young, 1931: 100; Thomas and Thomas, 1928: 572).

Merton's Definition of the Situation

While Thomas's 'definition of the situation' was more a methodological standpoint than a theoretical concept and in its complexity and openness not easy to grasp, it was Robert K. Merton, in his seminal 1948 article 'The Self-Fulfilling Prophecy,' who forged it into a distinct theoretical proposition: the 'Thomas theorem.' "If men define a situation as real, it becomes real in its consequence," quoted Merton – obviously inspired by Kimball Young – from Thomas and Thomas (1928), and made it the core of his own concept of the 'self-fulfilling prophecy.' The more famous the latter became, the more Merton's take also shaped further reception of the concept of the definition of the situation itself.

Whereas the 'definition of the situation' had been an integral part of Thomas's whole approach, Merton singled it out as the centerpiece of Thomas' s work. Merton's 'Thomas theorem' captured the definition of the situation in a most comprehensive way, making it easily understood and communicated. Merton's adoption made the concept famous, but in a peculiar version. This limited and distinctly shaped version became an asset in Merton's search for middle range theory.

Merton adopted the concept in the context of his research on propaganda and mass persuasion, and the self-fulfilling prophecy (opposite to 'suicidal prophecy') came to exemplify a fundamental social problem: ethnic hostility. Formalized in relation with several other concepts, particularly in-group and out-group, Merton emphasizes the "predominant role of the in-group in defining the situation" (Merton, 1948: 208). While putting Thomas on a pedestal for his seminal contribution, Merton saw the concept mirroring a long tradition, and he identified the 'Thomas theorem' with positions set forth by Marx, Freud, and others. By equalizing it with ideology, neurosis, and paranoid behavior, Merton reduced the concept to prejudice and pathology, i.e., to a flawed perception, which confirms itself by effecting confirming behaviors (Merton, 1948, 1995: 384f.). Thomas had treated 'true' and 'false' definitions methodologically equally. Merton's programmatic reception of Thomas's concept was thus selective, implicating that the subjective definition of situation was false or systemically dysfunctional: "The self-fulfilling prophecy is, in the beginning, a *false* definition of the situation evoking a new behavior which makes the originally false conception come *true*. The specious validity of the self-fulfilling prophecy perpetuates a rein of error. For the prophet will cite the actual course of events as proof that he was right from the very beginning" (Merton, 1948: 195).

Merton understood the self-fulfilling prophecy as a basic process of society, which stabilized and reproduced power asymmetries between societal groups. While Merton's concept was explicitly anchored in Thomas's, the earlier emphasis on social *change* in the concept of the definition of the situation was moved to the background in the Mertonian approach.

Qualitative Research and Microtheory after WWII

After WWII, a revived Chicago School Tradition – with exponents like Herbert Blumer, Tamotsu Shibutani, and Anselm Strauss and theoretical programs such as symbolic interactionism – stood in opposition to structural functionalism and reemphasized the individual person and its interaction with smaller groups. Individual interpretation of perception was highlighted as a necessary ingredient to everyday behavior. The so to speak 'neoclassical' reception of the concept of the definition of the situation could not ignore the paramount impact of Merton's interpretation, but it took its chance to reintroduce, refine, and further elaborate some of Thomas's and Park's flavors. A strong interest was dedicated to the organization of change in the definition of the situation, which was analyzed as a phenomenon in time and in physical and social space. Strauss and others elaborated on professional and other reference group related organization of definitive processes. All these attempts put communicative aspects to the core of conceptualization and empirical research.

Most strongly, Erving Goffman (1922–82) emphasized the bodily aspects of the definition of situation. He investigated face-to-face interactions – social situations defined by physical copresence, and which were therefore also limited regarding extension of space and time. Goffman advocated and practiced an ethnography of communication, which researched the cultural rules of individual conduct in social situations (Goffman, 1964). Innovatively, his approach included the senses and emotions as objects of research, and the nonverbal aspects of communicative interaction were taken most seriously. Like Thomas, Goffman worked genuinely interdisciplinary and blended biological methods (ethology)

with linguistic, literature, and arts studies. His argument for the use of theatrical and dramaturgical frameworks to analyze everyday social interactions was directly tied to the concept of the definition of the situation: "A character staged in a theater is not in some ways real, nor does it have the same kind of real consequences as does the thoroughly contrived character performed by a confidence man; but the successful staging of either of these types of false figures involves use of real techniques—the same techniques by which everyday persons sustain their real social situations. Those who conduct face to face interactions on a theater's stage must meet the key requirement of real situations; they must expressively sustain a definition of the situation: but this they do in circumstances that have facilitated their developing an apt terminology for the interactional tasks that all of us share" (Goffman, 1959: 254f.). Goffman's theory sought to generalize the 'framing' of situations, and he opened his book *Frame Analysis: An Essay on the Organization of Experience* (Goffman, 1974) with a discussion of the 'Thomas theorem' insisting that people finding themselves in a situation possibly affected by a crisis would normally not create a definition but rather state what the situation was supposed to be to them. They rely on frames of interaction, reflecting social expectations and identities. Goffman identified the framing with a quasisyntactical structure, which offers stabilizing order. This evoked interpretation of Goffman's approach as rather structuralist than interactionist (Gonos, 1977). The deeper problem revealed was that the focus on instantaneous, actual situations and definitions, microscopic both in terms of space and time, had (again) gained a concentration of the definition of the situation only with limitations of the concept's reach, and with many methodological questions surfacing subsequently. In the late 1960s and in the 1970s programs similar to that of Goffman were upheld by the ethnomethodologists, who derived their theoretical foundation not from Thomas or G. H. Mead, but referred to European authors like Wittgenstein and Schütz. Scholarly feuds developed between the "ethnomethodologists" and the "dramaturgists" (McHugh, 1968; Perinbanayagam, 1974: 521). Debates and attempts also centered on operationalizations of the concept to make it more feasible for empirical research. Modes and phases of the process of defining were differentiated (Stebbins, 1967, 1969; Ball, 1972).

Experimentalization and Transdisciplinary Adaptation

In the late twentieth and early twenty-first century, the contexts of psychiatry, educational studies, psychology and interethnic studies, where the concept of the definition of the situation originated, remain important fields for its use. For its constructivist virtues, also political science and science and technology studies have adopted it. But the concept has spread much wider across the social sciences and beyond. Several factors played together to make the concept of the definition of the situation most pervasive and almost ubiquitous today. One is its original vagueness, subsequently provoking and allowing diverse and highly debated differentiations. These continuing debates gave the concept its classical status. Two other closely related factors are the concept's emphasis on the individual and the small group, and its affinity or applicability to experimental settings. Both factors allowed for direct, questionnaire-based and even scientifically measureable observation in narrowly defined settings. Experimentally induced situational crises and research on resolution was applied by ethnomethodologists and interactionists alike. The concept even made it into pharmaceutical and medical research, where it was found useful to investigate placebo effects (Glass and Barber, 1961). Two further – again closely linked – factors for the proliferation of the concept can be listed here: the integration of objectively given and subjectively perceived structures to define a situation and limit possible action, and its linkage of micro- and macrolevels of the social. These motives have found much interest for situational analysis in the disciplines concerned with decision, from philosophy to economy (Hedström et al., 1998). Conspicuously, rational choice theory has adapted the concept of the definition of the situation. In such reading, actor's rationality is scaled for the definition of the situation, and individual choice of frames provides a particular developmental path and action according to subjectively expected utility. Finally, collective intentions, decisions and actions can be aggregated (Esser, 1996). The use of the concept to bridge social micro- and macrolevels points back to its very origins.

See also: Goffman, Erving (1922–1982); Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Abbott, A., Egloff, R., 2008. The polish peasant in Oberlin and Chicago: the intellectual trajectory of W.I. Thomas. American Sociologist 39, 217–258.

Ball, D.W., 1972. 'The definition of situation': some theoretical and methodological consequences of taking W.I. Thomas seriously. Journal for the Theory of Social Behaviour 2, 61–82.

Blumer, H., 1939. An Appraisal of Thomas and Znaniecki's The Polish Peasant in Europe and America (Critiques of Research in the Social Sciences: 1), Social Science Research Council Bulletin 44. SSRC, New York.

Esser, H., 1996. Die Definition der Situation. Kölner Zeitschrift für Soziologie und Sozialpsychologie 48, 1–34.

Glass, L.B., Barber, T.X., 1961. A note on hypnotic behavior, the definition of the situation and the placebo effect. Journal of Nervous and Mental Disease 132, 539–541.

Goffman, E., 1959. The Presentation of Self in Everyday Life. Doubleday, New York.

Goffman, E., 1964. The neglected situation. American Anthropologist 66, 133–136.

Goffman, E., 1974. Frame Analysis: An Essay on the Organization of Experience. Harper, New York.

Gonos, G., 1977. 'Situation' versus 'frame': the 'interactionist' and the 'structuralist' analysis of everyday life. American Sociological Review 42, 854–867.

Hedström, P., Swedberg, R., Udehn, L., 1998. Popper's situational analysis and contemporary sociology. Philosophy of the Social Sciences 28, 339–364.

Joas, H., 1993. Pragmatism in American sociology. In: Joas, H. (Ed.), Pragmatism and Social Theory. University of Chicago Press, Chicago, pp. 14–51.

McHugh, P., 1968. Defining the Situation: The Organization of Meaning in Social Interaction. Bobbs-Merrill, Indianapolis.

Merton, R.K., 1948. The self-fulfilling prophecy. Antioch Review 8, 193–210. Reprinted in Social Theory and Social Structure, enlarged, third ed. Free Press, New York, pp. 475–490.

Merton, R.K., 1995. The Thomas theorem and the Matthew effect. Social Forces 74, 379–422.

Park, R.E., Burgess, E.W. (Eds.), 1921. Introduction to the Science of Sociology. University of Chicago Press, Chicago.

Perinbanayagam, R.S., 1974. The definition of the situation: an analysis of the ethno-methodological and dramaturgical view. Sociological Quarterly 15, 521–541.
Shaw, C.R., 1930. The Jack-Roller: A Delinquent Boy's Own Story. University of Chicago Press, Chicago.
Stebbins, R.A., 1967. A theory of the definition of the situation. Canadian Review of Sociology/Revue canadienne de sociologie 4, 148–164.
Stebbins, R.A., 1969. Studying the definition of the situation: theory and field research strategies. Canadian Review of Sociology and Anthropology 6, 193–211.
Sutherland, E.H., 1947. Principles of Criminology. Lippincott, Chicago.
Thomas, W.I., 1917. The persistence of primary-group norms in present-day society, and their influence in our educational system. In: Jennings, H.S., et al. (Eds.), Suggestions of Modern Science Concerning Education. Macmillan, New York, pp. 157–197.
Thomas, W.I., 1923. The Unadjusted Girl, with Cases and Standpoint for Behavior Analysis. Little, Brown, Boston.
Thomas, W.I., 1937. Primitive Behavior: An Introduction to the Social Sciences. McGraw-Hill, New York.
Thomas, W.I., Thomas, D.S., 1928. The Child in America: Behavior Problems and Programs. Alfred A. Knopf, New York.
Thomas, W.I., Znaniecki, F., 1918–1921. The Polish Peasant in Europe and America, vols. 1, 2. University of Chicago Press, Chicago. Vols. 3–5, Richard G. Badger, Boston.
Thrasher, F.M., 1928. The study of the total situation. Journal of Educational Sociology 1, 599–612.
Waller, W., 1932. The Sociology of Teaching. Wiley, Hoboken.
Wolff, K.H., 1964. Definition of the situation. In: Gould, J. (Ed.), A Dictionary of the Social Sciences. Free Press, New York, p. 182.
Young, K., 1931. Language, thought and social reality. In: Young, K. (Ed.), Social Attitudes. Holt, Oxford, pp. 100–135.

Degrowth

Giorgos Kallis, Federico Demaria, and Giacomo D'Alisa, Research & Degrowth, Barcelona, Spain; and Institute of Environmental Science and Technology (ICTA), Autonomous University of Barcelona, Barcelona, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Degrowth calls for the abolishment of economic growth as a social objective and signifies a desired direction where societies will use less natural resources and organize to live very differently than today. This article traces the origins of degrowth from theorists of the 1970s, to the French activist movement of *décroissance* in the 2000s, to the contemporary international activist-researcher movement. We identify 10 degrowth theses concerned with the limits to – and of – growth, the relationship between growth, degrowth, and democratic politics, and the everyday and institutional alternatives that can facilitate a degrowth transition.

The Origins of Degrowth

The term '*Décroissance*' was used for the first time in 1972 by French intellectual André Gorz who responded to the Limits to Growth report by Meadows and colleagues (*see* Limits to Growth) by questioning the compatibility of zero growth or degrowth with capitalism. Gorz was inspired by Nicholas Georgescu-Roegen (1971), the intellectual pioneer of ecological economics (*see* Ecological Economics). In 1979, Jacques Grinevald and Ivo Rens edited a collection of articles by Georgescu-Roegen. The title they gave to the book was *Demain la décroissance*, translating with the prior consent of the author as *décroissance* the word 'descent' found in Georgescu-Roegen's 'Minimal Bio-economic Program' (Grinevald and Rens, 1979).

This first phase of the degrowth debate coincided with, and was reinforced by the oil crisis and economic recession in Europe, but was gradually abandoned with the advent of neoliberalism. The second phase of degrowth debates started in the late 1990s to early 2000s, this time framed as a criticism of the then dominant notion of 'sustainable development.' The center of this resurgence was France, and especially Lyon in the wake of protests for car-free cities, communal street feasts, food cooperatives, and campaigns against advertising. In 2002, the Institute for Economic and Social Studies on Sustainable Degrowth was founded in Lyon. Articles and special issues on degrowth were featured in the pages of *Casseurs de pub*, founded by ex-advertisers Bruno Clémentin and Vincent Cheynet in Lyon, *Silence*, a magazine edited by Lyon-based environmental and nonviolence activist Michel Bernard, and since 2004 *La Décroissance, le journal de la joie de vivre*, a magazine with national circulation in France, which today sells around 30 000 copies monthly. A landmark event was a 2002 conference called '*Défaire le développement, refaire le monde*,' which took place at UNESCO in Paris and included 800 participants.

Spreading from France, degrowth became a slogan mobilized by green and antiglobalization activists in Italy in 2004 (as '*Decrescita*') and Catalonia and Spain in 2006 (as '*Decreixement*' and '*Decrecimiento*'). The same year the young researcher-activist Francois Schneider undertook a year-long walking tour on a donkey to disseminate degrowth through France, receiving widespread media coverage. In 2007, Schneider founded in France the academic collective Research & Degrowth, with Fabrice Flipo and Denis Bayon, and promoted a series of international conferences, first in Paris in 2008 and then in Barcelona in 2010. The English term 'degrowth' was 'officially' used for the first time in the Paris conference, marking the birth of an international research community. As the Barcelona group from the Institute of Environmental Science and Technology (ICTA) joined the movement, the degrowth community extended beyond its initial strongholds in France and Italy. ICTA provided links to ecological economics as well as to Latin American networks of political ecology and environmental justice (see also Stonich, 2001; Environmental Justice in the United States). Following the success of the Paris and Barcelona conferences, more conferences were held in Montreal (2011), Venice (2012), and Leipzig (2014), with degrowth continuing to spread to groups and activities in Flanders, Switzerland, Finland, Poland, Greece, Germany, Portugal, Norway, Denmark, Czech Republic, Mexico, Brazil, Puerto Rico, Canada, Bulgaria, Romania, and elsewhere.

Since 2008, the English term has entered academic journals with over one hundred published articles and at least seven Special Issues in peer-reviewed journals (Kallis et al., 2010; Cattaneo et al., 2012; Saed, 2012; Kallis et al., 2012; Sekulova et al., 2013; Whitehead, 2013; Kosoy, 2013). Degrowth signifies first and foremost a critique of the growth economy. Usually, degrowth is associated with the idea that 'smaller can be beautiful.' Ecological economists have defined degrowth as an equitable downscaling of production and consumption that will reduce society's metabolism of energy and raw materials (Schneider et al., 2010). But our emphasis here is on different, not only less. Degrowth signifies a society with a smaller social metabolism but more importantly, a society with a metabolism that has a different structure and serves new functions. Degrowth does not call for doing 'less of the same.' It calls for doing things differently. To use a metaphor: the point is not how to make an elephant leaner, but how to turn an elephant into a snail.

Degrowth should not be confused with negative gross domestic product (GDP) growth. Still, a reduction of GDP, as currently counted, is a likely outcome of actions promoted in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.91041-9

the name of degrowth. A green, caring, and communal economy will be ecologically and socially sustainable, but unlikely to increase GDP by 2–3% each year. A key question then becomes how to make the inevitable decrease of GDP socially sustainable, given the dependence of the current system on growth (Jackson, 2009).

In the minds of the majority of the population, growth is still associated with improvement or well-being. For degrowth advocates, the purpose of using a negation for a positive project is precisely to decolonize an imaginary dominated by a one-way future consisting only of growth. It is the automatic association of growth with 'better' that the word degrowth wants to dismantle (Latouche, 2009).

In what follows, we identify 10 main theses emerging from the degrowth and associated literatures. We divide these into three broad categories, each corresponding to a different intellectual strand: the first with ecological-economic thought emphasizing the limits to – and of – growth; the second with cultural critiques of heteronomy and liberal democracy mainly from anthropologists and political philosophers; and the third with the search for alternatives, in which the degrowth literature meets the social-science literature on post-capitalist politics.

The Ecological-Economic Strand

Thesis 1: Growth Is Senseless

Skidelsky and Skidelsky (2012) argue that the pursuit of endless growth is senseless, since money can only be a means to an end, not an end in and for itself. If the end of a society is the 'good life,' then at some point a society should produce enough wealth to achieve it, making further growth unnecessary. Developed economies have accumulated wealth unimaginable a few generations back, yet they seem compelled to keep growing. A key question the degrowth literature addresses is why is it so.

At the micro-level of individuals, insatiability is driven by the competition for positional goods. Once basic material needs are satisfied, extra incomes are devoted increasingly to positional goods (e.g., a house bigger than the neighbor's). Relative, and not absolute, wealth is what matters. The growth of an economy, however, does not change relative positions. Yet, growth and productivity gains in the material sector increase the price of positional goods. Growth therefore makes positional goods socially scarcer. These are the social limits of growth, that is, intrinsic limits regarding the ability of growth to satisfy well-being (Hirsch, 1976). Such limits are one reason why above a certain level, growth does not increase average happiness (for a review of this literature see chapter on 'Happiness' in D'Alisa et al., 2014).

From an individual perspective it makes sense to pursue more wealth to climb up the positional ladder, but collectively this is a zero-sum game. Aggregate growth is the result of this senseless positional competition between individuals that makes no one better. The promise of growth, on the other hand, is the only way in which the illusion of upward social mobility can be temporarily sustained in the absence of redistribution. Either way, the end result is a senseless pursuit of growth for growth's sake (see chapter on 'Social limits to growth' in D'Alisa et al., 2014).

Thesis 2: Growth Is Uneconomic and Unjust

Growth no longer increases collective welfare. This failure cannot be captured by GDP, because GDP measures growth and not welfare. It does not distinguish welfare-improving from welfare-decreasing economic activity. For example, expenditures to clean an oil spill, build prisons, or wage wars increase the GDP. Independent measurements of well-being, such as self-reported life satisfaction and other happiness indices, or alternative welfare indicators such as the Genuine Progress Index or the Index of Sustainable Economic Welfare, confirm the hypothesis that above a certain level, growth does not improve welfare (Dietz and O'Neill, 2013).

The social costs of growth include bad psychological health, long working hours, congestion, and pollution (Mishan, 1967). The commodification and introduction into the formal GDP economy of ecosystem services and unpaid activities (D'Alisa and Cattaneo, 2013), or the monetization of social relationships such as hospitality or care are eroding sociability, community, and the environment. They also underline moral norms that used to regulate better such social and environmental relations (Hirsch, 1976). This might explain why the relationship between growth and a range of social indicators such as education or safety from crime is gamma-shaped: for low levels of GDP the wealthier a country is, the better its social conditions, but for high levels of GDP, more growth does not make a difference. Some argue that growth is achieved because of increasing costs rather than by a rise in benefits. In other words, growth has become uneconomic (Daly, 1996). Instead, there is a clear correlation between equality and social well-being (Wilkinson and Pickett, 2009).

The continuous growth in the social metabolism of the global economy is also causing social and environmental havoc in the commodity frontiers where energy and raw materials are extracted from, and waste disposed to (*see* Ecological Economics). Growth is sustained by an unequal exchange of resources and materials. This raises questions of environmental justice (*see* Environmental Justice in the United States): the material benefits of extraction and consumption accrue in wealthy areas or among better-off groups (elites) while the costs are shouldered by socially marginalized communities and neighborhoods. The uneven nature of development is a basic insight from political ecology (Stonich, 2001). The implication is that although growth is uneconomic as a whole, it can as well be sustained as long as the benefits accrue to the elites that hold power and the costs shifted elsewhere.

Thesis 3: Growth Is Ecologically Unsustainable

There are strong linkages between the scale of an economy and environmental impacts, from ecosystem loss to greenhouse gas emissions. Some posit the possibility of dematerialization, meaning the decoupling of raw material use and carbon emissions from economic growth. There is no evidence of such dematerialization taking place globally, and where it does it is nowhere near the necessary pace to avoid resource exhaustion or disastrous climate change. Some wealthy economies show signs of slow dematerialization for selected materials or energy sources, but this can also be the

artifact of substituting domestic production with imports from industrializing countries (see chapter on 'Dematerialization' in D'Alisa et al., 2014).

One reason dematerialization is very difficult is because the more efficient an economy becomes, the more resources it might consume (see chapter on 'Jevons' paradox' in D'Alisa et al., 2014). Technological advancements lead to a more efficient use of resources, reducing their cost, but raising the demand for them, making new uses more affordable.

The prospect also of a structural transition to 'weightless,' service-based information economies is unlikely. The emergy (embodied energy) of services is much higher than that of primary activities. Service activities are ecologically subsidized by commodity extraction and industrial production elsewhere (Odum and Odum, 2001).

Thesis 4: Growth Is Coming to an End

While growth beyond a certain threshold does not deliver social or environmental benefits, it also gets harder and harder to sustain over time. There is evidence of a declining long-term rate of growth in developed economies (Gordon, 2012). Some have attributed this to diminishing marginal returns (Bonaiuti, 2014), while others to the exhaustion of major technological innovations, such as the oil-fuelled combustion engine and associated changes (cars, highways, suburbs) that propelled accumulation in the twentieth century (Gordon, 2012). Yet others point to limits in creating effective demand and finding constantly new investment outlets for capital accumulating at the compound rate of 2–3% growth per year (Harvey, 2010).

Degrowthers instead are likely to postulate ecological limitations. A main argument from the bioeconomics of Georgescu-Roegen is that economic growth accelerates entropy. Economic growth degrades high-order (low entropy) stocks and turns them into low-order (high entropy) heat and emissions. Energy cannot be recycled, and materials can be recycled only to some extent. Since the earth has a limited amount of exhaustible stocks (such as raw materials and fossil fuels), in the long run economic activity can only be sustained at the rate of the flow of solar energy (Georgescu-Roegen, 1971).

How long into the future existing energy and material stocks can last is subject to considerable debate. Some treat peak oil (and other peaks in the extraction rates of essential raw materials such as phosphorous) as evidence that stock exhaustion is already limiting growth (*see* Mineral Resources, Limits to: The Case of Peak Oil). Others point to different estimates, or to opportunities for substitution, e.g., of oil by gas or coal. Climate change, a manifestation of what Georgescu-Roegen called 'thermal pollution' from increasing entropy, may restrict economic activity before resource exhaustion does. This will depend on the pace of transition to cleaner sources of energy. Bioeconomists are skeptical of such a transition because they calculate that renewable sources produce much lower energy surpluses (energy returns to energy investment, EROI) than conventional sources (Murphy and Hall, 2010). This means that much conventional energy will need to be expended in the transition to renewables. It also means that the scale of a 'solar economy' will have to be much smaller than the current one, given the actual lower EROI of renewable energies compared to fossil fuels.

Degrowth scholars who analyze the current crisis have postulated that resources have already limited the growth rate of the real economy. Such limits were temporarily superseded with the rise of the fictitious (bubble) economy of finance, which sustained growth for a little while more. In other words, an otherwise unsustainable rate of growth has been sustained through private and public debt (Kallis et al., 2009).

The Culturalist Strand

Thesis 5: The Growth Consensus Has Eroded the Political

Political philosophers have criticized the increasing depoliticization of public debate in Westernized democracies. Some have attributed this trend to the rise of neoliberalism and the 'Washington consensus' that subjugated sovereign political choice to the needs of unregulated capital and liberalized markets. Degrowth scholars agree with this explanation, but trace the origins of depoliticization further back in time and into the imaginaries of development and growth that became hegemonic after the Second World War (see chapter on 'Critiques to Development' in D'Alisa et al., 2014). Since its inception, the notion of economic development was tightly linked to that of growth, referring to the replication in 'underdeveloped areas' of the conditions characterizing industrialized nations, such as specialization, production for markets, agricultural intensification, urbanization, and modern values.

The discourse of 'sustainable development' (*see* Sustainable Development), degrowthers claim, has had a depoliticizing effect too, decoupling environmentalism from its radical critique of modernity and its aspirations to challenge dominant notions of the good life. Degrowth was thrown explicitly as a 'missile word' to *re-politicize* environmentalism and challenge the consensus around sustainable development (Fournier, 2008). With its emphasis on win–win solutions and the utopian goal of perpetuating development without harming the environment, degrowthers charged, sustainable development depoliticizes genuine political antagonisms about the kind of future people want. It evades the core question, which in Bruno Latour's (1998) terms is whether 'to modernize or ecologize.' Ecologizing society, degrowthers argue, is not about implementing an alternative form of development. It is about imagining and enacting alternatives *to* development (see chapter on 'Critiques to Development' in D'Alisa et al., 2014).

Some degrowth advocates trace depoliticization further back than neoliberalism or the post-war ideology of development, focusing instead on the origins of modernity (see chapter on 'Dépense' in D'Alisa et al., 2014). According to the theory of *dépense*, societies always have an excess (surplus) of resources over what is necessary for the mere reproduction of life: politics is about the collective choice of the destination of that surplus. The main innovation of capitalism was the (institutionalized) investment of a significant part of the surplus into new production. This allowed for the growth take-off during the eighteenth century. Consequently, capitalism has disavowed the exercise of political sovereignty *par excellence*, namely collective decisions for the destination of surplus. Older civilizations, for example, defined their purpose through different forms of unproductive surplus expenditure, with members

creating meaning by building a pyramid, supporting a class of monks, or idly philosophizing. This collective pursuit of sense has been lost in capitalist civilization, having been relegated to the individual sphere. Acts of *dépense* have been commodified and individualized. Indeed, the promise of modernity is that the meaning of life is to be found by each individual alone, and that in this pursuit he/she has every right to mobilize all the resources necessary. At the societal level, this translates into the non-negotiable demand for growth. As individuals search for meaning, the genuinely 'political' sphere where meaning is to be collectively constructed has been subordinated to the imperative of growth (see chapter 'From austerity to Dépense' in D'Alisa et al., 2014).

Thesis 6: Limits Liberate

Degrowth thinking in the 1970s associated with names such as Andre Gorz, Ivan Illich, and Cornelius Castoriadis has framed limits not as externally imposed but as a conscious choice that is part of a collective political project. Limits in this vein are 'a social choice and not imposed as an external imperative' (Schneider et al., 2010: 513). This argument for self-limitation in the face of environmental degradation is neither an intrinsic one ('let's save nature for nature's sake') nor a utilitarian one ('let's save nature to save ourselves who depend on it'). It is an argument for a simpler life in more harmony with the nonhuman world, a life that for degrowthers is the one that is worth to be lived.

Ivan Illich (1973) proposes that simplification and self-limitation, which he calls conviviality, is necessary for autonomy, and by extension for genuine democracy and equality. Convivial tools and spaces are those which are understandable, manageable, and controllable by their users. A house built and repaired by its users is convivial; a 'smart' building whose temperature is controlled by software is not. Autonomy means that people are in control of their tools and spaces, meaning that these are not produced or governed somewhere else. Illich took issue with fossil fuels not because of peak oil or climate change. His problem was that high energy use supports complex technological systems that cannot be directly controlled by their users. Such systems by their very nature call for specialized experts to manage them. This division between experts and nonexperts inevitably creates inegalitarian hierarchies and concentrates power. Equality and democracy are unlikely as long as people do not have sovereign and shared control over their means of subsistence and production. From this perspective, degrowth is antithetical to ecological modernization and ecoefficient technologies, such as large-scale renewable projects (industrial wind farms or the Sahara–Europe solar panel project) or high-speed trains, not only because these might be insufficient or not green enough (Thesis 3), but because of their non-convivial complexity.

André Gorz (1982) developed a somewhat similar argument, but with the concept of autonomy he emphasized freedom from wage labor. Gorz defined as autonomous the sphere of nonpaid work where individuals and collectives enjoy leisure or produce for own use, rather than for commodified exchange. The sphere of reproduction and care belongs to the autonomous sphere. Unlike Illich, Gorz was not critical of industry; he envisioned a dual industrial-autonomous society, consisting of an economic sphere where industrial goods and services are produced with the maximum possible efficiency, thereby liberating more and more time to be spent in the autonomous sphere. He also advocated for decentralized micro-industries using a 'digital fabricator' (similar to a 3D printer), similar to contemporary projects such as Open Source Ecology where innovation depends upon open collaboration. Gorz criticized the socialist dream of a collective and democratic control of the means of production; modern production lines are too complex to be controlled by workers without delegates, he argued. Industries should be governed by democratically controlled States and secure the basic material needs of everyone. Socialism will only come through liberation from wage labor, not by a change of wage labor conditions. Despite a slightly different logic from Illich, Gorz therefore also advocated for the expansion of an autonomous sphere of low-intensity, decommodified, frugal activities.

It is in this vein that more recently Serge Latouche (2011) argued that abundance can be found in frugality. For him abundance is a matter of relational, not material goods. It is a matter of autonomy and conviviality. Only a society that decides 'it has had enough' and limits itself from pursuing what can be pursued can solve the problem of scarcity, satisfying itself within what is available. Only a society that is egalitarian and shares what is available can break away from the positional inequalities that sustain insatiability and construct a sense of generalized scarcity.

In this spirit, some degrowthers emphasize that they find inspiration for their theory, not in Malthus but in the Neo-Malthusians anarcho-feminists of Emma Goldman, who advocated conscious procreation in their fight against capitalism and the exploitation of female bodies to produce soldiers and cheap labor (see chapter on 'Neomalthusians' in D'Alisa et al., 2014).

The 'Alternatives' Strand

Thesis 7: A Transition beyond Growth Is a Transition beyond Capitalism

Capitalism here is understood as an ensemble of institutions – private property, the corporation, wage labor, and private credit and money at an interest rate – whose end result is a dynamic of profit in search of more profit ('accumulation'). Degrowth, even in a narrow sense of downscaling the material throughput of an economy, is unlikely to be compatible with capitalism if growth cannot be decoupled from throughput. This is the case even if *in theory*, as some steady-state scholars argue, a universal and declining cap on material and energy throughput could force capitalist economies to work within limits (Lawn, 2011). In practice, however, lack of growth increases workforce exploitation to maintain or increase the rate of profit (Blauwhof, 2012; Harvey, 2011). This in turn renders capitalism and liberal democracy socially and politically unstable under conditions of lack of growth. Growth avoids redistributive conflict and sustains capitalism politically. It is in this concrete sense that growth is necessary for capitalism.

Lawn (2011) advocates a scenario where collective limits (caps) are set, and then social security policies (such as a Job Gurantee to all by the State) are applied to ensure that

downscaling does not threaten the basic well-being of anyone. However, as Blauwhof (2012) argues radical political changes would be necessary to allow such seemingly innocuous policy 'reforms.' In existing capitalist democracies there is inequality in political power, distributed roughly along the lines of economic power. Caps, taxes, or redistributive income/job security programs are likely to harm economically powerful interests with privileged access to governments. A radical rearrangement of power relations would be necessary if these were to be enacted, and after those, capital would not anymore act as the predominant social force.

Thesis 8: Degrowth Alternatives Already Exist

Degrowth has also served as a structuring concept for connecting a diversity of post-capitalist, low-carbon alternatives. In the literature there is an engagement with alternative institutions at both the micro-level of community organization and the macro-level of State economic regulation.

For example, as alternative to the primacy of the institution of private property under capitalism, degrowthers emphasize the reclaiming of – and the making of new – commons, with access to goods and services based on sharing. Illustrations of this are ecocommunities and co-housing projects experimenting with forms of nonprivate housing where users share common space and invest their own labor in the making and maintenance of their shelter. Other examples of new forms of property are pirate programmers and on-line communities of digital commons who break away from private intellectual property, making new 'copy-left,' open digital products (Carlsson, 2008). Cooperatives of workers and users challenge the corporate form of privatized ownership of the means of production and break away from rents and profits. In the field of food production, and unlike corporate agri-businesses, agroecological farmers and back-to-the-landers produce for subsistence and for alternative food networks linked directly to consumer cooperatives without passing through intermediaries. In urban gardens accessed and preserved in common, city dwellers produce their own food and conserve their own recreation spaces, putting nonpaid work into their subsistence and leisure. Degrowthers, in the tradition of Illich, are also concerned with and contribute to various new institutions of education or care, such as parental-cooperative daycare and schools, where parents participate directly in the education of their children.

Such nowtopias (Carlsson, 2008) share five features: first, a shift from production for exchange to production for use (by participants or end-users); second, a partial substitution of wage labor with voluntary participant activity, meaning a decommodification and de-specialization of labor; third, an anti-utilitarian logic, where circulation of goods is set in motion, at least partly, by an exchange of reciprocal 'gifts' rather than the search for profit; fourth, an intrinsic value assigned to connections ('commoning') and relations in and for themselves; and fifth, the lack of a built-in dynamic to accumulate and expand (D'Alisa et al., 2014).

From a sustainability perspective, what is interesting is the relative low-carbon/material content of many such alternatives compared to conventional State or market systems that provide the same services (D'Alisa and Cattaneo, 2013). In some cases, such as, for example, food production, alternative systems may be less efficient than conventional systems per unit of product, given the lower degree of specialization and the limited division of labor (though this would probably be less so if one were to also account for the environmental costs of inputs such as oil). Even so, such alternatives are likely to be more environmentally benign precisely because their unproductiveness limits their scale (a "debound", or inverse rebound or Jevons' effect).

There is an interesting proliferation of new communal institutions and forms of production under conditions of economic crisis, as evidence from Argentina, Catalonia, or Greece suggests. The insight here is that as the conventional institutions of capitalism fail to cater for the basic needs of people, there is a spontaneous rise of new institutions concentrated in the autonomous sphere (Conill et al., 2012). Such institutions do not have to substitute State functions (of care, education, social security, or provision of basic goods). Instead they can offer alternatives that transform the functions of the States without dismantling them, challenging in this way the tendency toward privatization of State assets and services.

At the level of government policies, the degrowth literature has investigated labor and social security institutions that can secure meaningful employment in the absence of growth, supporting where possible the expansion of the autonomous sphere. A job guarantee scheme proposes to make the State an employer of last resort, reducing *de facto* unemployment to 0 and subsidizing care or autonomous activities through State expenditures (see chapter on 'Job Guarantee' in D'Alisa et al., 2014). An unconditional basic income to all citizens, eventually coupled with a ceiling income and financed by progressive taxation on wages and profits, can secure a basic level of subsistence and security for all without access to wage labor. With basic needs covered, the proposition is that there will be more opportunities to dedicate one's time in the autonomous sphere and perhaps share work through part-time employment in the paid sector (see chapter on 'Basic and Ceiling Income' in D'Alisa et al., 2014). Worksharing, achieved through a redistribution of work between the employed and the unemployed via a reduction of working hours in the paid sector, can also reduce unemployment, and redistribute wealth if implemented with income parity (Kallis et al., 2013). The concern addressed by these three institutions is that if technology and productivity keep improving, a lack of growth will render redundant a growing part of the workforce. The State can then redistribute some of the benefits of increased productivity to subsidize liberated time and the autonomous sphere.

Another degrowth strategy for solving the unemployment problem is a State-supported structural shift to a New Economy of social and cooperative enterprises focusing on services (care, health, education, leisure). Such activities have low productivity but are more labor-intensive and provide meaningful employment. Although they have low economic value, they have high social value and can form the basis of a degrowth economy (Jackson, 2009).

Regarding money and the credit system, institutional proposals emerging from the degrowth literature include zero interest rates (Loehr, 2012) and very high reserve requirements for banks to limit credit bubbles that drive fictitious material growth (Daly, 1996). Community currencies, time banks, and barter exchanges have the potential to relocalize economic activity, control its scale, and break away from the

expansionary dynamics of the monetary system. That said, experiences with community currencies suggest that they can at best serve as small-scale complements to mainstream currencies. State money will continue to be a more important locus of change given the need for interlocal exchange in a complex economy, the fact that taxes (a large part of gross economic product) are paid in conventional money, and the vast amounts of financial and organizational resources required for stable large-scale monetary systems.

A related degrowth proposal is for the State to take back control of the money supply from private banks (see chapter on 'Public money' in D'Alisa et al., 2014). Money issued as debt creates a growth dynamic. Currently, money is mostly created through private banks. While private banks can only issue money as debt through loans, the State could also issue money free of debt to meet public needs, for example, by financing a basic income or a job guarantee scheme or by subsidizing cooperatives, care services, or renewable energy projects. By providing money themselves, states would also reclaim seigniorage (the difference between the nominal value of money and the cost of producing it), and save money by not having to pay private banks for borrowing to finance public expenditure.

Concerning debt, degrowthers have argued that the economy cannot be expected to grow at the rate necessary to pay fictitious debts (Kallis et al., 2009). Debt is a social relation, and history is full of examples of societies that forgave debts and started afresh. Western societies have maintained a materially affluent life-style by shifting promises of payment to the future. Debt forgiveness might trigger a decline in the living standards of small creditors and savers. From a degrowth perspective, the goal cannot be to relaunch growth to repay debts but to distribute fairly the costs of such an adjustment. It is legitimate to excuse the debts of those whose basic standard of living is threatened, and not pay those who lent for profit. Some credits are more legitimate than others. The assessment and decision should be democratic. In this vein, degrowthers advocate citizen-run Debt Audits.

Thesis 9: The Politics of a Degrowth Transition Are Open and Plural

How all those alternative property, work and credit, micro and macro institutions could be connected in an alternative form of social and political organization remains subject to considerable debate. There is also no agreement concerning the political strategies needed to replace the current institutions of capitalism with alternatives. Latouche (2009) sees change coming through parliamentary politics, positing degrowth as a program for parties of the Left, though not with a party of its own. Others see a more spontaneous networking and flourishing of post-capitalist alternatives occupying more and more political and economic space as the old system of States and corporations collapses. From this perspective, alternative practices of education, food provision, living, and producing (see Section Thesis 8: Degrowth Alternatives Already Exist) are 'political' practices: they challenge and develop concrete alternatives to the dominant institutions of capitalism. From a Gramscian perspective, everyday practices that rearticulate the common senses of the civil society are as necessary as the struggle for leading the political institutions in *strictu sensu*.

For some degrowthers, a transformation of the political system into a more assemblarian and direct form of democracy is part and parcel of a degrowth transition. They find hope in social movements such as the Indignados (Occupy), whose politics prefigured values and projects similar to those of degrowth. In general, in degrowth debates there is openness to a plurality of possible degrowth strategies and political subjects, ranging from nonwage labor nowtopians with a shared experience of living and producing in the autonomous sphere, to existing social movements, political parties, and Unions. D'Alisa et al. (2013) break down degrowth into civil and 'uncivil' strategies and actors, with the latter being defined as those who refuse to be 'governmentalized.' Organized disobedience features in the repertoire of degrowth activists, spanning from the occupation of abandoned houses ('squats') to sit-ins against megaprojects and coal plants, or financial disobedience such as that of Enric Duran, a prominent advocate of degrowth in Barcelona who 'expropriated,' in his own words, 492 000 Euros via loans from 39 banks in 2008, in order to both denounce the predatory credit system and dedicate that money to social projects (D'Alisa et al., 2013).

Thesis 10: Degrowth in the North Will Let the South Live Well

Degrowth in the Global North is advocated in order to liberate ecological space for improvement of living conditions in the South. This runs against the standard proglobalization argument that growth in the North is necessary so that the South has growing markets to export its products. Degrowthers would point to the historical fact that the Global North did not need growing markets in the South in order to grow, but rather cheap natural and human resources which it exploited. In fact, one could argue that degrowth in the North today will reduce demand for natural resources and industrial goods, making them more affordable and accessible to the South.

Even so, this economistic formulation of North–South relations is unsatisfactory from a degrowth perspective, as it leaves the imaginaries of growth and development intact. The intellectual roots of degrowth can be traced to the post-development literature and the criticism of the imaginary of development for the 'Third world.' The proposition here is that the rejection of growth in the North therefore can also stop the colonization of imaginaries in the South, increasingly streamlined in a one-way future consisting only of growth. There is actually no lack of alternative imaginations about the good life in the South. Examples include Buen Vivir in Latin America (or Sumak Kawsay in Ecuador); Ubuntu in South Africa; or the Gandhian Economy of Permanence in India (D'Alisa et al., 2014). Such visions express alternatives *to* development, alternative trajectories of socioeconomic organization. The degrowth proposition is that these stand to flourish by a retreat of the growth imaginary in the Northern countries that have promoted it, if not forced it to the rest of the world.

The Future of Degrowth

Research-wise, there is much to be done in modifying and challenging each of the 10 theses, producing new ones and further diversifying the field (Kallis et al., 2012). More research,

for example, is necessary to demonstrate that growth is ecologically unsustainable, or that dematerialization is impossible and insufficient. More historic data is necessary to support the thesis that we are facing systemic, and not a periodic, stagnation and that resource limits have something to do with it. The claim also that the abandonment of growth may revive political debate and nourish democracy, rather than animate catastrophic passions, is for the time being speculative. In general, such claims are firmly established *within* the degrowth community, but they are far from being accepted in wider society where dematerialization and green growth continue to be seen not only as possible, but also likely.

More research is also necessary to understand how and why people and nations adapt to the lack of growth, why some alternative institutions succeed and challenge capitalism while others collapse or get incorporated into the mainstream, or how, and under what conditions, institutions such as work-sharing or a basic and ceiling income will work and produce the outcomes their advocates claim they can produce.

Another question concerns the social dynamic, the actors, the alliances, and the processes that will create a degrowth transition. This question is not just intellectual and cannot be answered only by research. Social change is a process of creation and is impossible to predict in advance. What academic studies of degrowth can offer are new narratives to animate the politics of a transition. Degrowth research is not only inter- and transdiscpilinary. It is transparent about its values and politics, and is based explicitly on a model of post-normal science, researchers, practioners, and activists collaborating in novel formats, as, for example, the working groups of the international degrowth conferences (see Cattaneo et al., 2012). The 10 theses presented here are outcomes of this process and already animate politics. But if the degrowth debate is to keep its momentum, there is no reason to be restrained to those.

See also: Ecological Economics; Environmental Justice in the United States; Limits to Growth; Mineral Resources, Limits to: The Case of Peak Oil; Sustainable Development.

Bibliography

Bonaiuti, M., 2014. The Great Transition. Routledge, London.

Blauwhof, F.B., 2012. Overcoming Accumulation: Is a Capitalist Steady-State Economy Possible? Ecological Economic 84, 254–261.

Carlsson, C., 2008. Nowtopia. AK Press.

Cattaneo, C., D'Alisa, G., Kallis, G., Zografos, C. (Eds.), 2012. Degrowth Futures and Democracy. Futures 44 (6), 515–523 (special issue).

Conill, J., Cardenas, A., Castells, M., Servon, L., Svetlana, H., 2012. Otra vida es posible: prácticas alternativas durante la crisis. Ediciones UOC Press, Barcelona.

D'Alisa, G., Cattaneo, C., 2013. Household work and energy consumption: a degrowth perspective. Catalonia's case study. Journal of Cleaner Production 38, 71–79.

D'Alisa, G., Demaria, F., Cattaneo, C., 2013. Civil and uncivil actors for a degrowth society. Journal of Civil Society 9 (2), 212–224.

D'Alisa, G., Demaria, F., Kallis, G., 2014. Degrowth, a Vocabulary for a New Era. Routledge, New York.

Daly, H., 1996. Beyond Growth: The Economics of Sustainable Development. Beacon Press, Boston.

Demaria, F., Schneider, F., Sekulova, F., Martinez-Alier, J., 2013. What is degrowth? From an activist slogan to a social movement. Environmental Values 22 (2), 191–215.

Dietz, R., O'Neill, D.W., 2013. Enough Is Enough: Building a Sustainable Economy in a World of Finite Resources. Routledge, London.

Fournier, V., 2008. Escaping from the economy: the politics of degrowth. International Journal of Sociology and Social Policy 28 (11/12), 528–545.

Gorz, A., 1982. Farewell to the Working Class: An Essay on Post-Industrial Socialism. Pluto Press, London.

Gordon, R.J., 2012. Is U.S. Economic Growth Over? Faltering innovation Confronts the Six Headwinds. NBER Working Paper No. 18315.

Georgescu-Roegen, N., 1971. The Entropy Law and the Economic Process. Harvard University Press, Cambridge, MA.

Grinevald, J., Rens, I., 1979. Demain la décroissance: entropie-écologie-économie. Pierre-Marcel Favre, Lausanne.

Harvey, D., 2010. The Enigma of Capital. Profile Books, London.

Harvey, D., 2011. The Enigma of Capital: and the Crises of Capitalism. Oxford University Press, Oxford.

Hirsch, F., 1976. Social Limits to Growth. Harvard University Press, Cambridge.

Illich, I., 1973. Tools for Conviviality. Harper & Row, New York.

Jackson, T., 2009. Prosperity without Growth. Earthscan, London.

Kallis, G., Martinez-Alier, J., Norgaard, R.B., 2009. Paper assets, real debts: an ecological-economic exploration of the global economic crisis. Critical perspectives on international business 5 (1/2), 14–25.

Kallis, G., Schneider, F., Martinez-Alier, J., (Eds.), 2010. Growth, Recession or Degrowth for Sustainability and Equity? Journal of Cleaner Production 6 (18), 511–606 (special issue).

Kallis, G, Kerschner C., Martinez-Alier, J. (Eds.), 2012. The Economics of Degrowth, Ecological Economics 84, 172–180.

Kallis, G., Kalush, M., O'Flynn, M., Rossiter, J., Ashford, N., 2013. "Friday off": reducing working hours in Europe. Sustainability 5 (4), 1545–1567.

Kosoy, N. (Ed.), 2013. Degrowth: The Economic Alternative for the Anthropocene. Sustainability 5 (special issue). Available at: https://www.mdpi.com/journal/sustainability/special_issues/degrowth.

Latour, B., 1998. To modernize or to ecologize? That's the question. In: Castree, N., Willems-Braun, B. (Eds.), Remaking Reality: Nature at the Millenium. Routledge, London.

Latouche, S., 2011. Vers une société d'abondance frugale: Contresens et controverses de la décroissanc. Fayard/Mille et une nuits, Paris.

Loehr, D., 2012. The Euthanasia of the rentier – a way toward a steady-state economy? Ecological Economics 84, 232–239.

Latouche, S., 2009. Farewell to Growth. Polity Press, Cambridge.

Lawn, P., 2011. Is steady-state capitalism viable? Annals of the New York Academy of Sciences 1219 (1), 1–25.

Martinez-Alier, J., Kallis, J., Veuthey, S., Walter, M., Temper, L., 2010. Ecological distribution conflicts. Ecological Economics 70 (2), 137–452.

Martinez-Alier, 2001. Ecological economics. International Encyclopedia of the Social & Behavioral Sciences, pp. 4016–4023.

Mishan, E.J., 1967. The Costs of Economic Growth. Staples Press.

Murphy, D.J., Hall, C.A., 2010. Year in review – EROI or energy return on (energy) invested. Annals of the New York Academy of Sciences 1185 (1), 102–118.

Odum, H.T., Odum, E.C., 2001. A Prosperous Way Down. University Press of Colorado, Boulder.

Saed, 2012. Introduction to the degrowth symposium. Capitalism Nature Socialism 23 (1), 26–29.

Schneider, F., Kallis, G., Martinez-Alier, J., 2010. Crisis or opportunity? Economic degrowth for social equity and ecological sustainability. Introduction to this special issue. Journal of Cleaner Production 18 (6), 511–518.

Sekulova, F., Kallis, G., Rodríguez-Labajos, B., Schneider, F., 2013. Degrowth: from theory to practice. Journal of Cleaner Production 28, 1–6.

Skidelsky, R., Skidelsky, E., 2012. How Much Is Enough? Other Press, New York.

Stonich, S.C., 2001. Political ecology. International Encyclopedia of the Social & Behavioral Sciences, pp. 4053–4058.

Wilkinson, R.G., Pickett, K., 2011. The Spirit Level. Tantor Media, Incorporated.

Whitehead, M., 2013. Degrowth or regrowth? Environmental Values 22 (2), 141–145.

Dehumanization

Chiara Volpato, University of Milano-Bicocca, Milan, Italy
Luca Andrighetto, University of Genova, Genova, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Dehumanization is the act of denying humanness to other human beings. Throughout history, dehumanization has been an essential ingredient in the perpetration of intergroup violence and atrocities. The first section of this article defines the phenomenon and describes its different manifestations. Theoretical and empirical works on blatant dehumanization are then reviewed. In the third section, the functions fulfilled by dehumanization for groups and individuals are summarized. The fourth section presents more subtle and unconscious forms of dehumanization. In particular, the concept of infrahumanization is introduced with a summary of empirical evidence. The fifth section focuses on phenomena of objectification and self-objectification, and reviews work from the formulation of objectification theory until the most recent studies. Finally, new directions for investigating unexplored aspects of this phenomenon are proposed.

The Course of Dehumanization

Dehumanization involves denying humanness to others, introducing an asymmetry between people who have human qualities and people who are perceived as lacking these qualities. Dehumanization is a radical form of out-group derogation intrinsically linked with conflicts and mass killings. According to Tajfel (1981), the dehumanizing processes stem from intergroup contexts in which group members have undifferentiated perceptions and behaviors toward out-group members; such uniformity leads to depersonalization (in social psychology, depersonalization refers to the stereotypical perception of the self as an example of some defining social groups), which is the prelude to dehumanization. Dehumanization is many sided, multifaced, and flexible. It fits in places, people, and relations, and from time to time, it assumes the forms required by the social context and the *Zeitgeist*. The forms of dehumanization are various: subhuman, superhuman, objectifying, biological, and mechanical metaphors have developed over time. They are described below.

Animalization

Animalistic dehumanization reduces the target to subhuman levels, by denying the qualities that define human primacy over other living beings. Individuals perceived as lacking these characteristics are judged as irrational, immature, coarse, unlearned, or lacking self-control. Furthermore, they are accused of having impulsive behaviors, and of being driven by primitive impulses. People who are dehumanized in the animalistic manner feel degradation and humiliation; those who dehumanize feel disgust and contempt toward the dehumanized targets.

Throughout history, the animalistic metaphor, more than other dehumanizing metaphors, has been used to derogate the out-group to an inferior level. The animal has always been an important comparative example in the definition of human identity. Human beings indeed define themselves through their supremacy over animals, even if, at the same time, they use animals for justifying their supremacy over other human beings. Social groups excluded from 'full humanness' are subjected to treatments, which vary according to how dangerous they are considered. They may be likened to domestic animals, and thus they become targets of paternalism and exploitation, or they may be likened to wild beasts, in which case they suffer contempt, repressions, and mass killings.

Animalization is tightly linked with the Western culture; think about the barbarian definition coined by the classical world, or about the definition of native populations used during the conquest of the New World. European intellectuals discussed at length these definitions and wondered whether these individuals were humans, 'homuncoli,' or even apes. Animalistic dehumanization has been used for sanctioning white supremacy in colonial contexts, for portraying enemies as criminals during intergroup conflicts, and for delegitimizing women and lower classes within nations (Jahoda, 1999).

To date, psychosocial research has primarily focused on the negative side of animalistic dehumanization. In contrast, the positive valence of many animal metaphors has received little attention. In fact, animals have often represented power and high status. For instance, the emblems of many kingdoms portray images of lions, eagles, or griffins. Furthermore, philosophers and literary men refer to animals when talking about humanity and its problems; think about works by Esopo, Fedro, La Fontaine, Melville, Carroll, Collodi, and Orwell, or Dante and Machiavelli's allegories. Thus, the use of animal metaphors is more complex than what psychosocial research has documented, and requires a more articulated investigation that involves both positive and negative meanings.

Demonization

Demonization is a suprahuman metaphor that transforms other people into demons, devils, or witches. Through demonization other people are attributed supernatural powers with the aim of exacerbating their dangerousness, and thus legitimizing their elimination. Such representations stem from the classic concept of *monstrum* (i.e., monster): the one who *monstrat*, that is the one who 'shows.' The *monstrum* indeed shows a gap in the 'natural' norms ruling the relationships

between human beings and animal species. This gap may take the form of excess, defect, or malformation. Within the classical world, the monster was not necessarily considered as a negative character. For instance, the giants were symbols of greatness, rather than dangerousness. However, during Christianity the monster was replaced by the devil, which became the essence of evil (Le Bras-Chopard, 2000). Demonization is still used in the world today, as shown by recent events of 'witch-hunting' in some Central African countries, or by the imprisonment of political opponents accused of using magic in Iran. Notwithstanding the persistence of demonization, psychosocial studies have dedicated little attention to this metaphor.

Biologization

Biological metaphors are rooted in sixteenth-century Spain with the cult of the *limpieza de sangre*. However, they mostly developed during the nineteenth century, concurrently with the rise of scientific racism and social Darwinism.

Biologization employs metaphors linked to illness, hygiene, or purity; biologization involves transforming others with microbes, viruses, pestilence, cancer, dirt, and pollution (Douglas, 2002). Since Enlightenment, germ theory replaced the demon as metaphor of a dark danger. Individuals who were dehumanized in the biological manner had to face drastic measures, similar to those enacted to defeat emergencies and epidemic diseases. In other words, they must be defeated through practices of cleanliness, elimination, disinfection, extirpation, or purification. Most leaders guilty of twentieth-century genocides have been inspired by these practices.

Mechanization

Mechanization is a creature of the modern times. This is a dehumanizing metaphor that is gaining ground in the world today. Mechanization involves perceiving others as mechanical organisms, that is, robots that are incapable of feeling emotions and disclosing to others. Individuals who are denied such characteristics are judged as indifferent, cold, rigid, passive, lacking in curiosity, imagination, and depth; they are perceived as machines; and they do not elicit tender feelings, compassion, or empathy. The most meaningful examples of mechanistic dehumanization may be observed in medical and technological domains, that is, within those contexts in which the standardization of procedures, the myth of efficiency, depersonalization, and forced passivity wear away the richness of interpersonal interactions (Haslam, 2006). The mechanistic metaphor has been most applied in the industrial world, where Taylorism theorized the parceling out of mass production and the subordination of workers. Taylor's ideal was a worker without mental capacities and intellectual autonomy, a worker only capable of carrying out standardized actions: an estranged living being, a programmed robot.

Objectification

When objectified, individuals are perceived as objects, instruments, or goods. The objectification process leads to "a kind of instrumental fragmentation in social perception, the splitting of a whole person into parts that serve specific goals and functions for the observer" (Gruenfeld et al., 2008: p. 111). Objectification is embodied in the figure of the slave. Regrettably, this figure is still present in the world today: a new form of slavery – known as debt bondage – has indeed replaced the old slavery.

The concept of objectification was first introduced by Immanuel Kant. In *Metaphysics of Morals*, he argued that objectification refers to the perception of others as mere sexual instruments. During the past years, this concept has been widely analyzed by feminist thinking, which has focused on the sexual objectification of women (Papadaki, 2007). According to this perspective, objectification implies the perception and evaluation of a woman only on the basis of her sexual parts or functions, which "are separated out from her person, reduced to status of mere instruments, or else regarded as if they were capable of representing her. To be dealt with in this way is to have one's entire being identified with the body …" (Bartky, 1990: p. 35). In Bartky's view, within patriarchal societies, the woman is approached, or not, on the basis of her utility as a sex object, independent of other qualities that might make her attractive as a human being (Fredrickson and Roberts, 1997).

Martha Nussbaum (1995) identified seven dimensions of objectification: instrumentality, denial of autonomy, inertness, fungibility, violability, ownership, and denial of subjectivity. These dimensions are not simultaneously present in all cases of objectification: for instance, slaves are treated as goods, are sold and physically and psychologically violated, but, at the same time, are not considered as inert, because their physical activity establishes the slavery economy. For Nussbaum, instrumentality is the most dangerous dimension: when an individual is perceived as an instrument, he or she becomes useful and therefore interesting for those who intend to exploit him or her. Then, instrumental objectification implies approaching the objectified individual, a characteristic that distinguishes it from other forms of dehumanization, which usually involves the removal of the victim.

These different forms of dehumanization lead to different outcomes. As mentioned above, likening out-group members to animals implicates purposes of exploitation or intentions of mass killings; likening others to devils, monsters, or viruses leads to exacerbating the feelings of threat. Perceiving others as objects, instead, regards the world of commodification and instrumentality. It is also noteworthy that the same group may become the target of more dehumanizing images. For example, during the conquest of North America, the double portrayal of natives as savage beasts (i.e., wolves) and demons (i.e., red devils) led to the ethnic cleansing of the New World (Stannard, 1992).

The Legitimization of Evil: Research on Explicit Dehumanization

Psychosocial studies on dehumanizing processes documented how they are essential ingredients for social exclusion and atrocities: extreme negative behaviors toward a group perceived as hostile are facilitated when their humanity is denied. Thus, in contexts of intergroup conflict identifying dehumanizing attributes is particularly important because their presence is an indication of possible violent behaviors.

In social psychology, first theorizations about dehumanizing processes arose in the 1950s. Allport (1954) conceived

dehumanization as an extreme form of prejudice. In Allport's view, dehumanization concerns those individuals or groups perceived as outside the boundaries defined by ethical and moral values. Inspired by Allport's insights, Kelman (1973) and Opotow (1990) argued that dehumanization involves denying the identity of the victim, which is perceived as excluded by the social community where justice and solidarity are applied. When people are stripped of human qualities, they lose their capacity to evoke compassion and moral emotions, and may be treated instrumentally. When individuals or groups are included inside the boundaries of the moral community, they are treated fairly in resource distribution and in the sharing of sacrifices. On the contrary, the moral exclusion of individuals or groups by such boundaries rationalizes and justifies the perpetration of evil toward them. Thus, moral exclusion makes it normal and acceptable for acts of incredible cruelty toward others to take place.

This theoretical framework led to the first empirical studies. In particular, Bandura et al. (1975), by employing the electric shock paradigm, compared the punishment administered to dehumanized and humanized targets. (Electric shock paradigm was first used in the famous experiments conducted by Stanley Milgram (1963) on obedience to authority. In these studies, the experimenter ordered the participants to administrate electric shocks to another person – collaborator of the experimenter – whenever that person gave a wrong answer on a learning task. The electric shocks were fictitious. However, participants were not aware of this and believed to administrate painful shocks to another person.) Results showed that participants treated more punitively dehumanized targets than targets invested with human qualities. Based on these findings, Bandura (1999) developed the concept of 'moral disengagement.' According to the author, in the course of development people adopt moral standards that serve as guidelines for their behavior. In the face of situational inducements to behave in a reprehensible way, there are different mechanisms that shift the individuals' moral boundaries, and create a version of reality in which reprehensible conduct becomes morally acceptable. Mechanisms of moral disengagement operate, for instance, by cognitively reconstructing the behavior itself, by obscuring the agency of the reprehensible conduct, and by distorting or minimizing the consequences of the detrimental action. Dehumanization operates directly on the recipients of detrimental behaviors, which are blamed and perceived as beings that are not humans. Dehumanization is therefore a powerful process of disengagement from moral self-sanctions. To perceive another person as human arouses empathetic concerns. It is, therefore, difficult to mistreat him or her without feeling personal distress or remorse. If the other person is instead stripped of human qualities and emotions, if he or she is defined without human traits, the empathetic concerns are weakened and inhibited. For this reason, dehumanization is an essential weapon for anyone planning acts of extreme violence toward other individuals or groups. Exterminating, killing, and violating children, women, or men goes against the principles that each community teaches its members for the continuation of its existence, as well as perceiving itself as a complete community. However, at the time that material interests or unhealthy ideologies lead a group to exterminate another group, perceiving the out-group members as subhumans helps to go beyond the moral boundaries, and to enact behaviors that are unthinkable in a 'normal' context.

These assumptions have been corroborated by archival research on influential works of the nineteenth century. This research highlighted the obsessive use of dehumanizing metaphors regarding ostracized groups. For example, by following the theoretical framework of the delegitimization theory by Bar-Tal (1989), Volpato and colleagues (Durante et al., 2010; Volpato et al., 2010) conducted a series of studies analyzing the texts and the images contained in *La Difesa della Razza*, a bimonthly magazine of fascist propaganda, published between 1938 and 1943. Their studies documented the dehumanization of colonized people, who were likened to anthropoid apes. The dehumanization of the Jews was also particularly present in the images. They were portrayed in imaginative ways: they were spiders making their cobwebs on the world, snakes ready to attack, intellectuals writing with pens resembling vipers, chimpanzees posing as human beings, bats, vampires, vultures, parasites infesting the social body, monsters starving the humanity, or devils hiding their horns under a skullcap.

Psychosocial studies have mainly focused on the perpetrators, rather than on the victims. For example, the factors enhancing resistance to dehumanizing practices or the factors leading to acquiescence or internalization are still unknown. Furthermore, self-dehumanization processes remain unexplored. The analysis of Primo Levi's *If This Is a Man* represents the first attempt to investigate such phenomena. The goal of this work (Volpato and Contarello, 1999) was to explore the emotional, cognitive, and behavioral reactions of those who are excluded from humanity. In his book, Levi showed that in concentration camps, the dehumanization process deeply affected the self-perceptions of prisoners, who were forced to interiorize categories and pattern of thoughts imposed by the perpetrators. Throughout the book, Levi refers to a wider concept of dehumanization than the one commonly used in psychosocial studies. According to Levi, indeed, dehumanization concerns both victims and perpetrators because all the involved actors of the concentration camp universe were subjected to a tragic and often final impoverishment of their personality. Not seeing human in the other person destroys the humanity of both perpetrators and victims.

Throughout history, all the different expressions of dehumanization have been effective instruments of oppression and persecution. As discussed in the section below, powerful groups employ dehumanization for exploiting, humiliating, and destroying weaker groups.

The Functions of Dehumanization

Dehumanization mainly serves for (1) justifying group violence, (2) legitimizing the group's status quo, and (3) distancing individuals from potentially stressful events. Regarding the justification of intergroup violence, several studies have shown that dehumanization is an essential antecedent for marginalizing individuals or whole groups, and enacting violent behaviors toward them. Dehumanization decreases, and in some cases completely inhibits, feelings of empathy and compassion that individuals commonly feel

when observing others' sufferings. By doing so, dehumanization becomes useful for intentions of destruction and genocide.

For instance, recent works have revealed that after 9/11 dehumanization has been employed to develop consensus around the *war on terror*, leading Western public opinion to tolerate human rights violations and to neglect the impact of 'collateral damage' on civil populations. Terms like 'fighting enemies' or 'terrorists' have excluded enemies and prisoners from moral boundaries, and thus allowed international forces to treat them in ways forbidden by the Geneva Conventions. Steuter and Wills (2010) analyzed the dehumanizing metaphors used by Western mass media for describing the *war on terror*. They depicted the terrorist enemy through animal and biological metaphors – harmful animals, cancer, metastasis, viruses – that underscore the uncontrolled proliferation. Such metaphors serve for desensitizing the readers, and thus facilitate the approval of political and military decisions based on the denial of enemy humanity.

The second function of dehumanization concerns the legitimization of status quo. This function serves to reassure high-status group members about their social position, and makes them believe that they are immune from the bad luck of lower-status groups. In line with this reasoning, the attitudes exhibited by Western societies toward immigrants are particularly representative. For example, O'Brien (2003) listed an impressive number of dehumanizing metaphors used in the debate around restriction for immigrants entering the United States during the beginning of the twentieth century. The most representative metaphors recalled the objectifying nature of immigrants; they were seen as unassimilable and raw material, trash, and human derelicts. These objectifying metaphors were part of a rhetorical framework of instrumentality and commodification. These past figures of speech are repeated nowadays in many countries for delegitimizing immigrants.

The third and the least explored function of dehumanization serves for protecting those who hold power and control over other lives, and thus have to make potentially dangerous and painful decisions for other human beings. In contrast to the previous one, this function does not necessarily have negative effects on others, but instead it assumes defensive purposes. In this case, dehumanizing serves for distancing oneself from others. For example, it allows health care professionals (e.g., doctors, nurses) to manage and regulate their own negative emotions, above all when they are in particularly stressful and risky contexts (see Haque and Waytz, 2012).

The Hidden Side: Research on Subtle Dehumanization

A growing body of literature is disclosing the hidden side of dehumanization, and has documented the existence of less explicit forms of this phenomenon, which weaken the others' humanity in a subtle and unconscious way. These forms involve perceiving out-group members not necessarily as nonhuman beings, animals, or monsters, but rather as individuals who are only less human than in-group members. These subtle perceptions of lesser humanity occur even in the absence of intergroup conflicts; they pervade daily life without the social actor being aware of them.

Infrahumanization is the most relevant form of subtle dehumanization. Jacques-Philippe Leyens and colleagues (for a review, see Vaes et al., 2012) provided important evidence regarding this process. Infrahumanization is conceived as a subtle and unaware form of ethnocentrism (defined as the universal belief that one's own group is superior to other groups on a variety of dimensions, and at the same time that other groups lack a number of important characteristics, in the same dimensions, compared to the in-group); it does not imply a complete deprivation of others' humanity; it does not require the explicit likening of others to the animal world; and it does not lead to extreme reactions. Infrahumanization involves perceiving others as implicitly less human, without excluding them from social sphere.

Leyens and colleagues showed that people commonly use three characteristics to define human nature: intelligence, language, and sentiments. When perceived as lacking just one of these elements, groups are perceived as less human. The authors focused on the emotional aspect and, in particular, on the distinction between primary and secondary emotions (i.e., uniquely human emotions). This distinction derived from the concepts of 'emotions' and 'sentiments,' typically used in Romance languages. For Italian, French, or Spanish speakers, love is a sentiment and pleasure is an emotion. Emotions are affective states experienced by both human and animals, such as anger, fear, pleasure, and excitement. Sentiments are more complex affective states, which are thought to be uniquely experienced by human beings, such as remorse, nostalgia, pride, and hope. Primary emotions are part of human beings' gene pool, are universally shared, and occur earlier in the course of development. In contrast, secondary emotions are considered unique to human species, occur later in development, are not universally experienced in the same way, and tend to last much longer than the primary ones.

In an impressive series of studies employing different paradigms and instruments, Leyens and collaborators showed that people commonly attribute more secondary emotions to their in-group than to the out-groups, but do not differentially attribute the primary emotions. For instance, in one of the first experiments (Leyens et al., 2001), the researchers presented participants with a list of characteristics including primary and secondary emotions. The participants' task was to choose the characteristics that they considered typical of their in-group or typical of an out-group. The results confirmed the hypothesis that participants selected more secondary emotions for their in-group than for the out-group, whereas primary emotions were not differently associated with in-group and out-group. Importantly, the different association of secondary emotions was independent of the emotions' valence. In other words, it occurred both for positive and negative secondary emotions: participants even tended to attribute negative affective states to their group, in order to preserve its full humanness. Furthermore, such different associations emerged regardless of the group status. Out-group members are thus infrahumanized and perceived as less human than in-group members because they are considered as less capable of experiencing uniquely human emotions. Thus, results confirmed the existence of a 'privileged link' between in-group members and uniquely human emotions. A number of studies have also revealed that

infrahumanization has relevant consequences for intergroup relations, though milder than dehumanization. For example, a research study conducted 2 weeks after Hurricane Katrina documented that infrahumanization reduced the willingness to go to help the victims (Cuddy et al., 2007).

Why do people infrahumanize others? According to the theoretical framework of infrahumanization, the tendency to ascribe lesser humanity to out-group members stems from the universal inclination to essentialism. It refers to the people's tendency to perceive groups as natural entities, rather than as social artifacts built throughout the historical development of human species. Essentialism derives from a reification process whereby the differences between groups – commonly those differences that have important meanings in people's life – are interpreted as natural expressions.

Inspired by research on the infrahumanization-theory, over the past few years a growing body of research has explored the processes of subtle dehumanization by looking at a more general concept of humanity, which goes beyond the distinction between primary and secondary emotions (e.g., Boccato et al., 2008). In particular, Haslam and colleagues (for a review, see Haslam and Loughnan, 2014) have investigated these processes by considering human traits. They have proposed that humanness may be understood as characteristics that are uniquely human (e.g., rationality or morality) and define the boundaries between humans and other living beings, and characteristics that are representative of the core of human beings (i.e., traits of human nature), even though they are shared with other species (e.g., warmth or emotionality). Uniquely, human traits seem to implicate culture, social learning, or complex emotions; they refer to language, higher cognition, and conscientiousness; and they are learned during socialization and vary with the cultural context. Instead, human nature traits are innate, affective, universal, and implicate what is natural (Haslam, 2006). The denial of human nature traits leads to mechanistic dehumanization, while the denial of uniquely human traits leads to animalistic dehumanization. Research conducted in different countries highlighted that people in interpersonal relationships tend to attribute human nature traits of themselves rather than of the other person. In intergroup contexts, instead, people seem to attribute human traits congruent with the self-stereotype, whereas out-groups are attributed the less important human traits for the in-group's image. Further, in intergroup settings, the denial of human nature or uniquely human traits seems to have a distinct impact on willingness to help out-group members (Andrighetto et al., 2014).

Empirical studies on subtle dehumanization are now flourishing. Among others, Goff and colleagues' work (Goff et al., 2008) is noteworthy. It examined the persistence of historical dehumanizing representations (i.e., the 'Negro-ape metaphor'), which seem to have contemporary consequences operating beneath individuals' conscious awareness. Furthermore, the research from the 'ontologization perspective' (Pérez et al., 2007) is very relevant (this concept of ontologization suggests the idea that some ethnic minorities are more discriminated against than others. In particular, 'ontologization' refers to the representation of certain minorities outside the social universe of 'humanity'); it documented the persistent social exclusion of gypsies/Roma.

Objectification and Self-Objectification

Building on the objectification theory proposed by Barbara Fredrickson and Tomi-Ann Roberts (1997), a growing body of research has investigated the objectification processes (for a review, see Gervais, 2013). According to Fredrickson and Roberts, sexual objectification is described as the valuing of a woman exclusively on the basis of her body and on her sexual parts, rather than on her full identity. When a woman is sexually objectified, her body becomes a mere instrument for satisfying sexual male desires.

Although sexual objectification may assume different forms, the outcomes are often the same: women are ascribed restricted attitudes and roles, which reduce them to consumption objects, undistinguishable, interchangeable, and lacking in identity.

Furthermore, when objectified, women tend to adopt the observer's view and to treat themselves as objects to be looked at and evaluated. Self-objectification is the process whereby women and girls learn to think of themselves as instruments for others' desire. Historically, self-objectification derives from the female subordination typical of many societies, and from the fact that physical attraction has traditionally been one of the few means available to women for gaining power and social mobility. However, this strategy leads to focus on the physical aspect, neglecting other possible qualities. If objectification was maybe useful in the past, when women had few possibilities for avoiding the forced roles, nowadays it damages women and girls.

Self-objectification negatively affects psychological and physical well-being because it exacerbates negative emotions, reduces self-awareness, and contributes to mental and physical health problems, such as depression, eating disorders, or sexual dysfunctions. The first consequence of self-objectification is the increase of negative experiences concerning the body. The comparison with unreachable images of physical perfection, which are obsessively shown by the mass media, elicits sentiments of shame, disgust, and distress for one's own inadequacy. Such emotions produce mental stress and obsessive focus on the body, and lead individuals to a state of confusion that does not allow them to be clear-headed in their thoughts and actions. A further consequence of self-objectification is the decrease of the experiences of flow states, conceived as those moments in which a person, while performing an activity, is fully immersed in a feeling of being alive, and feels joy and satisfaction.

Several empirical studies have corroborated the objectification theory. These studies mainly focused on the relation between self-objectification, body image, eating behaviors, and intellectual performance (for a review, see Calogero et al., 2010). In particular, two empirical experiments by Fredrickson and colleagues (Fredrickson et al., 1998) opened the way for other studies. In their first experiment, the researchers manipulated self-objectification by having female participants try on a swimsuit or a sweater (depending on the experimental condition) in front of a full-length mirror. All participants were then asked to fulfill a questionnaire, eat some cookies, and have a drink. The results showed that women in the swimsuit condition reported lower self-esteem and felt more body shame, which in turn predicted restricted eating. The second experiment considered both male and female participants. The effects on body shame and restricted eating were replicated for women only, thus confirming the assumption that

self-objectification particularly concerns the female gender. Additionally, this experiment shed light on a further relevant outcome of self-objectification. After participants tried on the swimsuit or the sweater and completed the questionnaire, they were asked to perform a math test. The results showed that women in the swimsuit performed significantly worse on the math test than those in the sweater condition. No differences emerged instead for men. Such findings confirm the assumption that self-objectification consumes women's attentional resources, which is shown by their diminished mental performance on the test.

More recent works documented the progressive rise of objectification and self-objectification also among men, and highlighted the risk of these processes particularly for adolescents and children, who are increasingly subjected to the objectifying influence of mass media.

To date, psychosocial research has mainly focused on sexual objectification. In fact, in *Economic and Philosophic Manuscripts*, Marx put forward fundamental reflections on objectification and related processes of alienation within the work domain. The philosopher, when analyzing the alienation of worker's labor, argued that in the capitalism system work is not a free conscious activity, and therefore a manifestation of humanity, but rather it is an external imposition, depriving the worker of the product of his or her own work. Building on these reflections, some works are trying to extend the empirical analysis of objectification to the work context. In particular, the purpose of these studies is to explore whether the job, in certain circumstances, may lead workers to self-objectify themselves, and to be perceived by others as mere objects and instruments (see, e.g., Baldissarri et al., in press).

Recent Developments

Beside research on dehumanization and objectification, new research perspectives on denials of humanness phenomena are now emerging. In particular, works on anthropomorphism and on correlated neuropsychology of dehumanized perceptions seem particularly promising.

Anthropomorphism is conceived as the attribution of cognitive, emotional, and behavioral human characteristics to nonhuman entities, belonging to religious, natural, or technological spheres. Anthropomorphism has relevant psychosocial outcomes: perceiving agents as human indeed leads to considering them worthy of moral concern. Anthropomorphizing is a practice of humanization that likely leads to protect the anthropomorphized entities, by starting a process that is opposite to moral disengagement and exclusion (Waytz et al., 2010).

A further promising field of research refers to neuroimaging studies (a relatively new discipline in the field of psychology and employs different techniques or instruments (i.e., fMRI) evaluating functional aspects of neural activity), which explore the neural correlates of dehumanization. For example, Harris and Fiske (2009) found a relative lack of medial prefrontal cortex activation – a brain region hypothesized to be responsible for social cognition – after viewing photographs of social groups eliciting disgust and contempt (i.e., homeless people and drug addicts). Similarly, Cikara et al. (2010) found that pictures of sexualized women induce less activity in brain regions associated with mental state attribution, particularly for men with high ratings of hostile sexism. This is a relevant result suggesting that such out-groups are not primarily processed as human beings, but as objects.

At the present time, research on dehumanization is constantly evolving, and there will certainly be interesting developments in the next few years. Many issues still remain unexplored. For example, 'dehumanization for invisibility' should be explored, as a particular form of dehumanization practiced through silence, indifference, and the use of mere statistical numbers, which nullifies the salience of personal and social identity. This is a form of dehumanization integrating explicit and implicit elements; it stems indeed from the convergence of more explicit forms of dehumanization, which are primarily instilled by institutions, and more implicit ones, which allow society to not take responsibility for the deprivation of humanity of disadvantaged groups.

Another important issue concerns the identification of powerful strategies that inhibit dehumanization. To date, researchers have applied the same means commonly used for defeating prejudice and intergroup conflicts, such as intergroup contact (e.g., Andrighetto et al., 2012). Nevertheless, it is essential to identify specific strategies for contrasting dehumanization, since it is not only an extreme form of prejudice but also implies denying victims' humanity. However, what is clear so far is that dehumanization could be efficiently defeated only through the combined collaboration of institutions, policy makers, educators, and mass media, who must strive to recognize equal human dignity to all inhabitants of the Earth.

See also: Emotions and Intergroup Relations; Essentialism; Intergroup Relations; Mass Killings and Genocide, Psychology of; Racism: Social Psychological Perspectives; Sexism; Social Psychology; Stigma, Social Psychology of; Xenophobia: Social Psychological Aspects.

Bibliography

Allport, G.W., 1954. The Nature of Prejudice. Addison-Wesley, New York.

Andrighetto, L., Baldissarri, L., Lattanzio, S., Loughnan, S., Volpato, C., 2014. Human-itarian aid? Two forms of dehumanization and willingness to help after natural disasters. British Journal of Social Psychology. Advance online publication.

Andrighetto, L., Mari, S., Volpato, C., Behluli, B., 2012. Reducing competitive victimhood in Kosovo: the role of extended contact and common ingroup identity. Political Psychology 33, 513–529.

Baldissarri, C., Andrighetto, L., Volpato, C. When work does not enobble man: psychological consequences of working objectification. TPM, Testing, Psychometrics, and Methodology, in press.

Bandura, A., 1999. Moral disengagement in the perpetration of inhumanities. Personality and Social Psychology Review 3, 193–209.

Bandura, A., Underwood, B., Fromson, M.E., 1975. Disinhibition of aggression through diffusion of responsibility and dehumanization of victims. Journal of Research in Personality 9, 253–269.

Bar-Tal, D., 1989. Delegitimization: the extreme case of stereotyping and prejudice. In: Bar-Tal, D., Graumann, C.F., Kruglanski, A.W., Stroebe, W. (Eds.), Stereotyping and Prejudice. Changing Conceptions. Springer-Verlag, New York, pp. 169–182.

Bartky, S.L., 1990. Femininity and Domination: Studies in the Phenomenology of Oppression. Routledge, New York.

Boccato, G., Capozza, D., Falvo, R., Durante, F., 2008. The missing link: ingroup, outgroup and the human species. Social Cognition 26, 223–233.

Calogero, R.M., Tantleff-Dunn, S., Thompson, J.K., 2010. Self Objectification in Women: Causes, Consequences, and Direction for Research and Practice. American Psychological Association, Washington.
Cikara, M., Eberhardt, J.L., Fiske, S.T., 2010. From agents to objects: sexist attitudes and neural responses to sexualized targets. Journal of Cognitive Neuroscience 23, 1–12.
Cuddy, A.J.C., Rock, M., Norton, M.I., 2007. Aid in the aftermath of hurricane Katrina: inferences of secondary emotions and intergroup helping. Group Processes and Intergroup Relations 10, 107–118.
Douglas, M., 2002. Purity and Danger: An Analysis of Concept of Pollution and Taboo. Routledge, London.
Durante, F., Volpato, C., Fiske, S.T., 2010. Using the stereotype content model to examine group depictions in fascism: an archival approach. European Journal of Social Psychology 40, 465–483.
Fredrickson, B.L., Roberts, T.A., 1997. Objectification theory: toward understanding women's lived experiences and mental health risks. Psychology of Women Quarterly 21, 173–206.
Fredrickson, B.L., Roberts, T.A., Noll, S.M., Quinn, D.M., Twenge, J.M., 1998. That swimsuit becomes you: sex differences in self-objectification, restrained eating, and math performance. Journal of Personality and Social Psychology 75, 269–284.
Gervais, S.J. (Ed.), 2013. Objectification and (De)Humanization. Springer, New York.
Goff, P.A., Eberhardt, J.L., Williams, M.J., Jackson, M.C., 2008. Not yet human: implicit knowledge, historical dehumanization, and contemporary consequences. Journal of Personality and Social Psychology 94, 292–306.
Gruenfeld, D.H., Inesi, M.E., Magee, J.C., Galinsky, A.D., 2008. Power and the objectification of social targets. Journal of Personality and Social Psychology 95, 111–127.
Haque, O.S., Waytz, A., 2012. Dehumanization in medicine: causes, solutions, and functions. Perspectives on Psyhcological Science 7, 176–186.
Harris, L.T., Fiske, S.T., 2009. Social neuroscience evidence for dehumanised perception. European Review of Social Psychology 20, 192–231.
Haslam, N., 2006. Dehumanization: an integrative review. Personality and Social Psychology Review 10, 252–264.
Haslam, N., Loughnan, S., 2014. Dehumanization and infrahumanization. Annual Review of Psychology 65, 399–423.
Jahoda, G., 1999. Images of Savages: Ancient Roots of Modern Prejudice in Western Culture. Routledge, London.
Kelman, H.C., 1973. Violence without moral restraint: reflection on the dehumanization of victims and victimizers. Journal of Social Issues 29, 25–61.
Le Bras-Chopard, A., 2000. Le zoo des philosophes (The Zoo of Philosophers). Plon, Paris.
Leyens, J. Ph, Rodriguez, A.P., Rodriguez, M.T., Gaunt, R., Paladino, M.P., Vaes, J., Demoulin, S., 2001. Psychological essentialism and the differential attribution of uniquely human emotions to ingroups and outgroups. European Journal of Social Psychology 31, 395–411.
Milgram, S., 1963. Behavioral study of obedience. Journal of Abnormal and Social Psychology 67, 371–378.
Nussbaum, M.C., 1995. Objectification. Philosophy and Public Affairs 24, 249–291.
O'Brien, G.V., 2003. Indigestible food, conquering hordes, and waste materials: metaphors of immigrants and the early immigration restriction debate in the United States. Metaphor and Symbol 18, 33–47.
Opotow, S., 1990. Moral exclusion and injustice. Journal of Social Issues 46, 1–20.
Papadaki, L., 2007. Sexual objectification: from Kant to contemporary feminism. Contemporary Political Theory 6, 330–348.
Pérez, J.A., Moscovici, S., Chulvi, B., 2007. The taboo against group contact: hypothesis of Gypsy ontologization. British Journal of Social Psychology 46, 249–272.
Stannard, D.E., 1992. American Holocaust. Columbus and the Conquest of the New World. Oxford University Press, Oxford, UK.
Steuter, E., Wills, D., 2010. 'The vermin have struck again': dehumanizing the enemy in post 9/11 media representations. Media, War & Conflict 3, 152–167.
Tajfel, H., 1981. Human Groups and Social Categories. Cambridge University Press, Cambridge.
Vaes, J., Leyens, J. Ph, Paladino, M.P., Miranda, M.P., 2012. We are human, they are not: driving forces behind outgroup dehumanisation and the humanization of the ingroup. European Review of Social Psychology 23, 64–106.
Volpato, C., Contarello, A., 1999. Towards a social psychology of extreme situations: Primo Levi's *If This is a Man* and social identity theory. European Journal of Social Psychology 29, 239–258.
Volpato, C., Durante, F., Gabbiadini, A., Andrighetto, L., Mari, S., 2010. Picturing the other: targets of delegitimization across time. International Journal of Conflict and Violence 4, 269–287.
Waytz, A., Epley, N., Cacioppo, J.T., 2010. Social cognition unbound: insights into anthropomorphism and dehumanization. Current Directions in Psychological Science 19, 58–62.

Deindividuation, Psychology of

Tom Postmes and Felicity M Turner, University of Groningen, Groningen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by T. Postmes, volume 5, pp. 3364–3366, © 2001, Elsevier Ltd.

Abstract

Deindividuation is a characteristic of the individual in the crowd. It is a psychological state of decreased self-evaluation, causing antinormative and disinhibited behavior. The impact of deindividuation theory in science and society (especially twentieth-century politics) makes it one of social science's more influential contributions. Deindividuation theory is rooted in some of the earliest social psychological theorizing, more than a century ago. It seeks to explain the apparent transformation of rational individuals into an unruly group or crowd. It posits that the group provides an environment in which the individual – submerged and anonymous – suffers from a loss of self-awareness. Deindividuation hinders reflection about the consequences of actions, rendering social norms impotent while increasing suggestibility to random outside influences. Despite its status and impact, empirical support for deindividuation theory is minimal. Recently, this lack of support has been attributed to the faulty assumption that crowds cause a loss of self. Instead, research suggests that 'deindividuation' marks a transition from individual identity to social dimensions of the self. This transition to a social identity may increase responsiveness to social norms particular to the crowd, instead of decreasing responsiveness to generic social norms, as suggested by deindividuation theory.

The Psychology of Deindividuation

According to deindividuation theory, the psychological state of deindividuation is aroused when individuals join crowds or large groups. The state is characterized by diminished awareness of self and individuality. This in turn reduces an individual's self-restraint and normative regulation of behavior. In social psychology, deindividuation is a major theory of group behavior: it provides an explanation of collective behavior of violent crowds, mindless hooligans, and the lynch mob. In addition, deindividuation has been associated with other social phenomena such as genocide, stereotyping, and disinhibition in other settings such as computer-mediated communication (CMC). Later, the theoretical evolution of deindividuation is sketched, followed by a brief impression of the empirical support for this theory. Finally, recent research is discussed, which argues for a reconceptualization of deindividuation: It appears deindividuation is not a loss of individual identity, but may be better construed as a transition to a social identity.

Theoretical Evolution of Deindividuation

Deindividuation theory is rooted in the earliest works of social psychology, in particular Gustave Le Bon's (1895/1995) crowd theory (see Postmes and Spears, 1998; for a review). In his book *The Crowd*, Le Bon vividly describes how the individual in the crowd is psychologically transformed. He proposes that the psychological mechanisms of anonymity, suggestibility, and contagion transform an assembly into a 'psychological crowd.' In the crowd the *collective mind* takes possession of the individual. As a consequence, a crowd member is reduced to an inferior form of evolution: irrational, fickle, and suggestible. The individual submerged in the crowd loses self-control and becomes a mindless puppet, possibly controlled by the crowd's leader, and capable of performing any act, however atrocious or heroic.

Although Le Bon's theory has been extensively criticized, the influence of *The Crowd* in science and society is difficult to underestimate. The sales of this book are among the highest of scientific texts. In science, his analysis has inspired many theories of collective behavior, for example those of Freud, McDougall, Blumer, and Allport. In society, Le Bon's influence has been especially poignant: His theory also informed numerous politicians including Mussolini and Hitler, whose ideologies resonated with Le Bon's emphasis on racial factors (Nye, 1975).

The field of social psychology revived Le Bon's ideas in the 1950s. They were couched in more scientific terms as a theory of 'deindividuation.' Initially, it was argued that deindividuation occurs when individuals in a group are not paid attention to as individuals (Festinger et al., 1952). Thus, being unaccountable in a crowd or group has the psychological consequence of reducing inner restraints, and increasing behavior that is usually inhibited. The causes of deindividuation were gradually extended from anonymity in groups to other contextual factors, such as reductions of responsibility, arousal, sensory overload, a lack of contextual structure or predictability, and altered consciousness due to drugs or alcohol (Zimbardo, 1969).

Thus, although Le Bon's ideas were the blueprint for the deindividuation concept, the situations that cause deindividuation were gradually expanded to include other settings than the crowd. Deindividuation theory differs from Le Bon's theory in one further important respect: The consequence of deindividuation is not that the loss of individuality is replaced by a collective mind that guides the individual's actions. Rather, the loss of individuality leads to a total loss of self-awareness, and this incapacity for introspection releases a person from internalized moral restraints to produce emotional, impulsive, irrational, regressive, and intense behavior.

In the 1970s, deindividuation theory became a popular focus of scientific research, encouraged and inspired in part by the collective disorders characteristic of that period. However,

the empirical support for deindividuation theory was inconsistent. In particular, there was contradictory evidence regarding the influence of anonymity in groups, and there was virtually no evidence for the psychological state of deindividuation (Diener, 1980). Indeed, it appeared that this psychological state, which by all accounts appeared to be easily induced, was more elusive than initially conceived.

In response to these empirical obstacles, deindividuation theory was gradually modified and specified over the course of the 1980s. In particular, the psychological underpinnings of deindividuation have been narrowed down. In the original version of the theory, deindividuation was closely bound up with a lack of accountability due to submergence in the crowd. Thus, the focus was on anonymity in the group as the most important factor in producing deindividuation. But contemporary formulations of the theory shifted the focus to reduced (private) self-awareness as the central feature of deindividuation (Diener, 1980; Prentice-Dunn and Rogers, 1982). When one is self-aware, the focus of one's attention is on the (private) aspects of oneself, such as one's attitudes and norms. This self-focus is assumed to be required in order for people to self-regulate. But irrespective of whether anonymity or self-awareness is the crux of deindividuation, both classical and contemporary perspectives argue that deindividuation leads to antinormative and disinhibited behavior.

In more recent years, the phenomenon of deindividuation has been applied to social behavior on the Internet. Kiesler and colleagues proposed that disinhibited behavior in CMC could be accounted for by anonymity (Kiesler et al., 1984). Based on deindividuation theory, it was argued that visual and auditory anonymity of online interaction partners masked both personality and culture. In a mixture of elements from classical and contemporary deindividuation theory, this was hypothesized to reduce self-awareness and depersonalize one's audience, causing deregulated and antinormative behavior (Kiesler et al., 1984; Siegel et al., 1986).

Deindividuation Research

Several influential studies were conducted to illustrate the force of deindividuation. For example, Zimbardo (1969) reports a study that inspired much subsequent deindividuation research. In one study, participants were rendered anonymous by clothing them in oversized lab coats and hoods, compared with normal clothes and name tags in the control condition. The participants' task was to shock a confederate in a situation similar to the classic Milgram studies on obedience. In a first experiment using groups of female students, Zimbardo demonstrated that anonymous participants shocked longer (and therefore more painfully) than identifiable participants, in confirmation of his theory.

Aware of the limitations of Zimbardo's research in terms of external validity, Edward Diener and colleagues devised more naturalistic contexts in which to test the theory. In some studies, groups of participants were required to 'test' a so-called pacifist, trained to remain nonresponsive, by hitting the pacifist with foam swords and so forth. The researchers claimed this provided a measure of aggression that was more ecologically valid for crowd behavior. Results were inconsistent, however. In one study, for example, isolated individuals displayed more aggression than groups did (see Postmes and Spears, 1998 for a review).

In another study, Prentice-Dunn and Rogers (1982) induced deindividuation by instructing participants repeatedly to focus attention outward. In addition to attentional focus, participants were seated in a dimly lit room with loud rock music playing, verbal interaction was encouraged, and the groups played exciting video games. In contrast, participants in an internal attention focus condition were told not to interact, performed individual tasks, and played nonarousing games in a well-lit quiet room. Results showed that when attention was focused outward, higher levels of aggression were obtained.

More recently, deindividuation research expanded into the realm of CMC. Anecdotal reports of early CMC usage suggested that users were more likely to be disinhibited when they interacted online. Some initial studies appeared to confirm this. Siegel et al. (1986) conducted an experiment studying three-person groups who were presented with a choice-dilemma and asked to reach consensus. These decisions were made face to face, via an anonymous CMC system, and via a nonanonymous CMC system. Consistent with deindividuation theory, interactions via a computer exhibited both more polarized opinions (i.e., more extreme shifts away from members' initial choices) and more disinhibited and aggressive behavior than face-to-face communications.

But in CMC there was also conflicting evidence regarding the fundamental processes thought to promote deindividuation. In their empirical investigation of self-awareness in CMC Matheson and Zanna (1988) showed that, relative to face-to-face interactions, CMC showed significantly higher levels of private self-awareness – a finding opposite to deindividuation expectations. Thus, there was mixed evidence regarding deindividuation in CMC.

Support for Deindividuation Theory

These examples illustrate that empirical investigations of deindividuation theory searched for evidence of deindividuation in ever more exotic social situations. Although several studies show effects that appear to be consistent with deindividuation theory, many studies failed to support it, and many reported contrary results. For example, Zimbardo's (1969) original paper reports that a replication of the study mentioned earlier obtained the exact opposite results. Also in the case of CMC, evidence was inconsistent. In contrast to some positive findings (Siegel et al., 1986), there are other studies showing that online anonymity can sometimes promote more conservative, cautious decisions and prosocial behavior (e.g., Hiltz et al., 1989). Thus, evidence for deindividuation theory appears to be mixed. A meta-analysis of the deindividuation literature confirms this impression. This study examines 60 tests of deindividuation theory and concludes that there is insufficient support for deindividuation theory (Postmes and Spears, 1998). Disinhibition and antinormative behavior are not more common in large groups and (crowded) anonymous settings. Moreover, there is no evidence that deindividuation is associated with reduced self-awareness, or even that reduced self-awareness increases disinhibition.

Overall, then, deindividuation theory does not receive sufficient empirical support.

Reconceptualizing Deindividuation

One likely reason for the lack of support for the deindividuation concept is that the theory is based on Le Bon's characterization of the crowd, which is fundamentally flawed. According to Le Bon, collective behavior is always irrational: the individual in the crowd loses cognitive control. Deindividuation theory accepts this notion. For example, Diener argued that crowd members are "similar to the stimulus-response organism of early behaviorism, with reduced conscious mediation" (1980: p. 230). This negative and disorderly view of collective behavior stands in stark contrast to historical analyses of the crowd, which stress the crowd's capacity for restraint and orderly behavior (e.g., Rudé, 1964). Moreover, these historical analyses accentuate that the crowd often acts on the basis of moral principles and popular consensus (see also McPhail, 1991). Thus, a compelling case can be made that collective behavior is not unrestrained (as argued by deindividuation theory) but rather restrained by normative processes. The implication is that collective behavior is under conscious control, and far from irrational and unrestrained.

Recent social psychological research supports such a normative analysis of collective behavior. In deindividuation studies, for example, people are more likely to follow local group norms in situations where they are anonymous or where self-awareness is low (Postmes and Spears, 1998). The idea that behavior could be the result of local group norms was considered explicitly by Johnson and Downing (1979) who showed that the effects of anonymity depended on the normative cues in the environment. Participants were made anonymous by means of mask and overalls reminiscent of the Ku Klux Klan (similar to Zimbardo's uniforms) or by means of nurses' uniforms. Although compared to the control condition participants shocked somewhat more when dressed in the Ku Klux Klan uniforms, they actually shocked less when dressed as nurses. This finding illustrates that groups are sensitive to normative cues associated with the social context. Similarly, in CMC the phenomenon of online flaming has been interpreted as a locally normative behavior which may be experienced positively within specific subcultures (Postmes et al., 2000).

This finding that crowds act normatively, prompted a reinterpretation of the effects found in deindividuation research. The Social Identity Model of 'Deindividuation' Effects (SIDE; Reicher et al., 1995) uses social identity processes to explain deindividuation phenomena. Central to SIDE is the idea (derived from self-categorization theory) that there is not one unitary self-concept, but rather that the self is multifaceted. Specifically, the self contains not just one's personal identity (i.e., the unique part of the self), but is also composed of many different social identities which the individual shares with groups of others (Reicher et al., 1995). SIDE argues that conditions of anonymity (or CMC) do not cause a loss of self, but that they can shift a person from an individual identity to a collective identity, provided that people are members of the same group or crowd. For example, being in a demonstration will accentuate the unity of participants if they rally behind their cause – seemingly oblivious to strong divisions and disagreements that may exist between them. This common group membership may become accentuated if an opposing group (such as the police) acts upon the crowd as if it were one, for example by blocking its way or attempting to contain it. The strong adherence to this common group membership, which emerges over the course of such events, explains why normative regulation within the crowd may occur (Reicher, 1987).

Thus, whereas deindividuation theory argues that the crowd causes a loss of identity, leading to disinhibition, SIDE argues that the crowd causes a shift from a personal identity to a social identity (shared with the crowd) leading to social regulation of behavior. This may explain the normative restraint within the crowd as observed by historians, as well as the normative restraint observed in experimental studies of deindividuation.

Although the SIDE model originated as an alternative perspective on crowd behavior, it was also applied to normative regulation online. Processes described in the SIDE model revolve around the psychological rather than physical presence of one's group. This means that the SIDE model can be applied to contexts in which others are co-present (such as crowds) just as easily as situations where others are not co-present (such as CMC). In order to understand online 'deindividuation effects' we need to first understand whether communicators think of themselves as an individual or a group member. Online, there are many contexts in which personal identity is likely to be salient. Under such conditions, when the self is thought of as unique and different, the SIDE model argues that online anonymity provides people with the means to follow personal standards and norms. But there are also many online situations in which a clear social identity is available and salient. Under such conditions, anonymity and the physical absence of others can paradoxically amplify the influence of the group and its norms. This is because individual group members (and their idiosyncrasies) are not visible, so all that remains is the unifying idea of 'us.' CMC is therefore an unlikely but fertile forum for the growth and expression of social identities, facilitating increased group cohesion and commitment (and thereby, SIDE foreshadowed the social side of Internet). Thus, according to SIDE, computer-based interactions do not lead to a loss of self, but amplify it. CMC provides the cognitive conditions which allow greater freedom to be true to the personal self, as well as to their group.

In sum, there is no consistent empirical evidence that supports deindividuation theory either in real-life groups and crowds or in online behavior. There is no consistent relation between anonymity and disinhibition. It appears to be more productive to think of deindividuation not as a loss of self, but as a shift to a different aspect of the self and the norms associated with it. In the case of the crowd, this means a shift from personal to a shared social identity. Online, in situations where people interact with others of the same group, a similar accentuation or shift toward a shared social identity can be found. The ability of SIDE to explain diverse phenomena – including behavior in the presence and absence of others – attests to its utility. Ultimately, this reconceptualization emphasizes that explaining deindividuation phenomena as a loss of self is too simplistic. Diverse social and cognitive phenomena come together to influence the norms that people respond to in different contexts. There is no evidence of a loss

of self in the crowd. Rather, the crowd provides individuals with an avenue for the expression of collective identities, directed and inhibited by social norms.

See also: Aggression, Social Psychology of; Collective Action; Collective Behavior, Social Psychology of; Dehumanization; Group Processes, Social Psychology of; Mass Killings and Genocide, Psychology of; Self-Categorization Theory; Social Identity in Social Psychology; Social Movements: A Social Psychological Perspective; Social Psychological Theory, History of; Social Psychology.

Bibliography

Diener, Edward, 1980. Deindividuation: the absence of self-awareness and self-regulation in group members. In: Paulus, Paul B. (Ed.), The Psychology of Group Influence. Lawrence Erlbaum, Hillsdale, NJ.

Festinger, Leon, Pepitone, Angelica, Newcomb, Theodore, 1952. Some consequences of de-individuation in a group. Journal of Abnormal and Social Psychology 47, 382–389.

Hiltz, Starr R., Turoff, Murray, Johnson, Kenneth, 1989. Experiments in group decision making, 3: disinhibition, deindividuation, and group process in pen name and real name computer conferences. Decision Support Systems 5, 217–232.

Johnson, R.D., Downing, L.L., 1979. Deindividuation and valence of cues: effects on prosocial and antisocial behavior. Journal of Personality and Social Psychology 37, 1532–1538.

Kiesler, Sara, Siegel, Jane, McGuire, Timothy W., 1984. Social psychological aspects of computer-mediated communication. American Psychologist 39, 1123–1134.

Le Bon, Gustave, 1995. The Crowd: A Study of the Popular Mind (Original work published in 1895). Transaction Publishers, London.

Matheson, Kimberly, Zanna, Mark P., 1988. The impact of computer-mediated communication on self-awareness. Computers in Human Behavior 4, 221–233.

McPhail, Clark, 1991. The Myth of the Madding Crowd. De Gruyter, New York.

Nye, Robert A., 1975. The Origins of Crowd Psychology: Gustave Le Bon and the Crisis of Mass Democracy in the Third Republic. Sage, Beverly Hills, CA.

Postmes, Tom, Spears, Russell, 1998. Deindividuation and anti-normative behavior: a meta-analysis. Psychological Bulletin 123, 238–259.

Postmes, Tom, Spears, Russell, Lea, Martin, 2000. The formation of group norms in computer-mediated communication. Human Communication Research 26, 341–371.

Prentice-Dunn, Steve, Rogers, Ronald W., 1982. Effects of public and private self-awareness on deindividuation and aggression. Journal of Personality and Social Psychology 43, 503–513.

Reicher, Steven D., 1987. Crowd behavior as social action. In: Turner, John C., Hogg, Michael A., Oakes, Penelope J., Reicher, Steven D., Wetherell, Margaret S. (Eds.), Rediscovering the Social Group: A Self-categorization Theory. Basil Blackwell, Oxford, England, pp. 171–202.

Reicher, Steven D., Spears, Russell, Postmes, Tom, 1995. A social identity model of deindividuation phenomena. European Review of Social Psychology 6, 161–198.

Rudé, George, 1964. The Crowd in History: A Study of Popular Disturbances in France and England, 1730–1848. Wiley, New York.

Siegel, Jane, Dubrovsky, Vitaly, Kiesler, Sara, McGuire, Timothy W., 1986. Group processes in computer-mediated communication. Organizational Behavior and Human Decision Making 37, 157–187.

Zimbardo, Philip G., 1969. The human choice: individuation, reason, and order vs. deindividuation, impulse and chaos. In: Arnold, W.J., Levine, D. (Eds.), Nebraska Symposium on Motivation, vol. 17. University of Nebraska Press, Lincoln, NE, pp. 237–307.

Deinstitutionalisation in Low and Middle Income Countries

Julian Leff, University College London, London, UK; Kings College London, London, UK; and University of Cape Town, Rondebosch, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The progress achieved in transforming a psychiatric-hospital-based service to one based in the community is reviewed, comparing the United States with the United Kingdom. The situation in low- and middle-income (LAMI) countries is considered in terms of the history of psychiatric services, including the role of colonialism. Statutory psychiatric resources in a number of LAMI countries are enumerated. Innovative examples of the provision of psychiatric care in the community in countries with scanty resources are described, augmented by the author's investigative visits to many of them. The lessons to be learned from these experiences are summarized in an overview.

Institutionalization in High-Income Countries and Low- and Middle-Income Countries

While deinstitutionalization in the USA and Western Europe has been relatively successful, with a few exceptions this is not the case in low- and middle-income (LAMI) countries. In England and Wales, the number of functioning psychiatric hospitals has fallen from 130 in 1975 to less than 12 today, (Leff, 2000) having been replaced by a variety of community-based services including community mental health teams, crisis teams, day hospitals, and day centers and a number of alternatives to hospital admission, such as acute day hospitals, and crisis houses (Johnson, 2010). The change in the location of psychiatric care has not been so dramatic in the United States with a less comprehensive network of community services. However, large psychiatric hospitals began to close in the 1950s. In 1955, there was 1 psychiatric bed for every 300 Americans. Currently, there is 1 for every 3000 (Torrey et al., 2010). The situation is quite different in LAMI countries, but varies considerably from country to country, depending on the development of their psychiatric services. The state of mental health services in LAMI countries will be discussed with examples from a variety of countries with different levels of economic development and of service provision. Between them they present a wide range of solutions to the problems of deinstitutionalization, which are instructive.

Ethiopia and Egypt

Ethiopia with a population of 80 million has only a single psychiatric hospital with 250 beds, situated in the capital and staffed by 17 psychiatrists who have received training from a team from Toronto. In the rest of the country, new psychiatrist-led units have been established in four cities, and nurse-led psychiatric units are scattered around the country. Otherwise, the only resource available is the Coptic churches that provide treatment for mental illness with holy water and by beating the patients to drive out evil spirits. By contrast, Egypt, with the same population size as Ethiopia, has 15 state psychiatric hospitals with 7000 patients, 60% of whom have been resident for 5 years or more. Outpatient services are hospital based with no community input. There is 1 psychiatrist per 80 000 population, 1900 psychiatric nurses, and around 200 clinical psychologists. A collaborative agreement between the Egyptian government and the government of Finland initiated a Mental Health Program in 2002, with the aim of moving the locus of care from the psychiatric hospitals to the community. However, this objective has not been achieved to any significant degree. One important outcome has been the ratification of a Mental Health Act in May 2009 (Loza, 2010). In countries without an act of parliament protecting the rights of mentally ill people there can be no recourse for inhumane treatment. Furthermore, such an Act raises public awareness of the plight of mentally ill people in degrading and humiliating institutions.

Chile

Chile after Pinochet has the superficial appearance of a European country with civilized standards. However, it does not have an operational mental health act. Professor Martin Cordero, a senior Chilean psychiatrist who worked in the United Kingdom for 10 years, was commissioned by the Chilean government to draft a mental health act. He prepared a comprehensive document and submitted it several years ago. It has never been enacted by the parliament. As a result, according to World Health Organization (WHO) Chile does not have a mental health policy. The neglect and abuse in psychiatric institutions that persist as a result were conspicuously evident to me on a visit to the University Psychiatric Hospital in the southern city of Conception. As I walked through the wards I noticed piles of sand lying in various places. I was curious about their function and was told that the roofs leaked for lack of funds for repairs and the sand absorbed the rainwater. I also noted that the television sets on the wards were too high up for the patients to reach the controls. Only nursing staff could change the channels. Worse was to come. I entered a ward for female patients, which was adjacent to a ward for long-stay patients with tuberculosis. The women were all nursed in their nightgowns. Their toilet was a very large room with many holes in the floor in place of proper lavatories, and the entrance door was rotting at the bottom leaving a wide

 http://dx.doi.org/10.1016/B978-0-08-097086-8.27081-5

gap. I found it incredible that conditions of this kind could persist in an institution belonging to a university.

Professor Cordero advised me to visit the forensic hospital in Santiago de Chile. The entrance was guarded by two soldiers with rifles. The director proudly showed me his up-to-date leucotomy unit. On questioning he informed me that they performed three leucotomies per week, and that the patients receiving this treatment were suffering from substance abuse, psychopathy, and anorexia nervosa. I asked if the patients gave their signed consent for this procedure, and he replied that it was sufficient for the parents to give their permission. I reported this back to Professor Cordero, who made representations to the Mental Health Department. He was astute enough to base his argument for closing the unit on the large expense of maintaining the unit and the absence of evidence for its effectiveness. This led to the closure of the unit, whereas the human rights issue would almost certainly have failed. Pinochet casts a long shadow.

Mexico

In LAMI countries with faltering economies it is not unusual to find psychiatric hospitals with degrading conditions alongside innovative community initiatives that mobilize local resources with excellent results. Two colleagues and myself were commissioned by the WHO to conduct a survey of mental health services in the administrative region of Mexico City, after the Director of mental health services announced his plan to build 100 new psychiatric hospitals. In one hospital we visited there were two buildings. One was a closed ward for long-stay patients with a solid steel door. The Director proudly informed us that this was a state-of-the-art ward designed by an architect and based on Jeremy Bentham's concept of the panopticon. The ward was circular with the beds arranged around the perimeter of the circle, and a nursing station in the center allowing full visibility of every patient at all times. A consequence of this arrangement was that there was no room between the beds for cupboards to contain the patients' belongings.

Across the way from this immaculate building was a decrepit ward housing patients with psychoses together with patients with severe learning difficulties. At the back of the ward were the lavatories in an open space behind a low brick wall, which allowed visibility of the patients using the facilities. Only the nursing staff had access to the flushing system for the lavatories.

In extreme contrast, our next visit was to the town of Tierra Colorada. We were ushered into the Primary Care Health Center where we were given an enthusiastic welcome by a large group of children and adults of all ages. They presented programs of instruction in mental health care, which are delivered by volunteers from the appropriate age groups. Thus there are child volunteers (Agentes Juveniles en Salud – Young Health Activists), adolescents, adults who address mother–child issues, others who assist parents, and health activists of the third age. The Young Health Activists use puppets in their educational sessions with children. By this means, all age groups receive mental health education from their peers. This was a striking example of what can be achieved with very little expense to reach out to local communities when official mental health services are restricted to inappropriate buildings with institutionalized patients and staff.

Iran

A somewhat similar approach to developing community psychiatric services has been initiated in Iran. A national program has been mounted of training primary health care workers and local community workers. At least one male and one female *behvarz* (a local person from the same village) work in each 'health house.' *Behvarzes* have a general education up to secondary school level and 2 years training in health care, including 1 week of formal training in mental health. This program has been rolled out nationally through effective central control. The expansion of mental health care into primary care has reduced assistance sought from traditional practitioners. Across several areas, first contact with a traditional health practitioner for a mental health problem has shifted from 40% in 1990 to 16% in 2000.

Jamaica

The transformation of psychiatric services in Jamaica was achieved through the vision and dedication of the Professor of Psychiatry in the University of the West Indies, Frederick Hickling. As in many of the Commonwealth countries, the sole psychiatric hospital, Bellevue, was built by the British. It was opened in 1861 and originally named Kingston Lunatic Asylum. By 1945, the number of beds had increased severalfold to 2350, resulting in overcrowding, as happened in the mental hospitals built in the same period in Britain. From 1974 determined efforts were made to shift the locus of care away from the institution by developing a community mental health system. Currently, there is a network of over 300 clinics providing a full range of mental health services, delivered primarily by mental health nurses, supervised by 30 psychiatrists. More than 40 000 patients are treated annually (Hickling, 2010). As a result of this development, the number of beds in Bellevue has dropped to 700. There is a mobile community follow-up service that keeps track of the hundreds of patients who have problems with medication adherence, and provides emergency services, crisis intervention, and acute hospitalization when necessary. More than 60% of those patients requiring hospitalization for acute mental illness are admitted to the medical wards in general hospitals. There is a mountainous region in the center of Jamaica, which renders travel to and from Kingston difficult for patients. During a visit I made to the island, I accompanied a mental health nurse on her regular drive up into the mountains to visit the wooden cabin there, which is used as a dispensary for patients living in that region to collect their psychotropic medication.

South Africa

In 2013, South Africa had its financial aid from Britain discontinued because of its increasing general affluence. However,

there is still a very large section of the population that lives in poverty in makeshift shacks. The psychiatric services exhibit the same patchy quality ranging from sophisticated university departments of psychiatry to no services at all. The history of South African psychiatry begins with Robben Island, where Nelson Mandela was imprisoned. In 1845, the prison colony on the island was converted to a leper colony, and later to a psychiatric hospital. It is interesting that the same change in use occurred on the Finnish island of Sela. In 1910, there were eight government mental hospitals caring for 3624 patients. A Mental Disorders Act was introduced in 1916. By 1955, there were 13 mental hospitals with 17 881 patients. Currently, the population of 52 million is served by 24 public psychiatric hospitals, accommodating 14 000 acute and long-stay patients. There is 1 psychiatrist per 350 000 population, 1 psychologist per 312 000, and 1 social worker per 250 000. However, there are no community psychiatric services in the North West and Northern Cape. The problem of providing any kind of health services to this sparsely populated region has been tackled in a highly innovative way by means of the Health Train. This project began in January 1994 and has brought health care to rural areas where there is 1 doctor for every 4000–5000 patients. The train brings a variety of health specialists to remote areas where most people have no access to specialized services. The train travels around the country for 36 weeks of the year, stopping at least five times at each of the 36 stations in three rural regions. Initially the train had three carriages and was solely financed by Transnet, the railway services company. It expanded to 16 carriages and many companies now act as sponsors. About 40 000 patients receive treatment from the specialists on the train each year and well over 1 million people have been treated to date. The train has an Edu-Clinic, which is equipped to teach groups of 25 volunteers about basic health issues; a medicine clinic with a pharmacy; a dental clinic; and an eye clinic. Initially no psychiatric service was included, but the train now has a psychology clinic with two consultation rooms. This innovative outreach service provides mental health care to the poorest and most remote of the rural poor.

Where there is no possibility of accessing sophisticated psychiatric help, many people consult traditional healers (Behr and Allwood, 1995). It is a mistake to dismiss this source of help as worse than useless. Prince (1960), who worked as a government psychiatrist in western Nigeria, discovered that traditional healers were treating psychotic patients with a preparation from the plant *Rauwolfia serpentina*, which has an antipsychotic effect, and was recommended for mental disturbances in the Ayurvedic writings of India, which date back more than 2000 years (Leff, 1988). Aside from herbal treatments, some healers are acute observers of people and can employ psychological understanding to determine the cause of their problems. Harvey (1976) describes the traditional Korean healers as "keenly intuitive and perceptive persons who make good use of their knowledge of people and human problems in helping their clients make the best of their situations. They give conservative advice, provide clients with outlets for potentially disruptive emotions, and get substantial help from the passage of time in solving clients' problems." I interviewed Mr. Gepe, a traditional healer in Gugulethu, a township in Cape Town, whose abilities accord with Harvey's description. Mr. Gepe told me, "I am sitting in my house when three men appear at my door. They stand there silent. I have to be able to say to them, 'You have come to me to help you find the oxen you have lost.'" There is nothing magical about his intuition. The men are in tribal dress and are obviously herders. As there are three men together it cannot be a domestic problem. That narrows down the possibilities. The South African government is well aware of the yawning gaps in the services, but is also cognizant of the fact that among traditional healers there are quacks and con men as well as people of integrity. Therefore they have initiated the process of registering traditional healers with the aim of excluding the undesirables.

Under the former apartheid government, racial segregation was the rule in all hospitals, including those for psychiatric patients. Nonwhite nurses were forbidden to treat white patients. The enlightened director of Valkenberg psychiatric hospital in Cape Town, Professor Lynn Gillis, subverted this law, taking a risk with his career. That law no longer pertains, but the legacy of the inhumane doctrine of apartheid has left its traces, which are evident in the educational system at all levels; for example Stellenbosch University is training 2000 psychologists, only a handful of whom are nonwhite.

The investment over the years in a relatively large number of psychiatric hospitals and the stultification generated by the apartheid system has severely impeded the transition to community-based psychiatric facilities. For several years I acted as an advisor to the planning committee for psychiatric services in Cape Town. The members listened to my arguments for deinstitutionalization and responded positively. But it was not possible to make plans with any confidence as the government failed to provide any funds for this transition. Instead money was made available to upgrade the acute admission wards. In the event, the building work was so shoddy that it had to be pulled down and the process begun again. Currently, the plans being discussed are to rebuild much of Valkenberg hospital, an unfortunately retrograde decision.

In the absence of government investment in the much-needed community services, voluntary organizations have stepped in to bridge the gap. Rotary has provided funds to create rehabilitation facilities on the site of a small psychiatric hospital. These include the creation of a horticultural center maintained by patients with learning difficulties, and some sheltered residences on the boundary of the hospital. Another nongovernmental organization (NGO), The Cape Mental Health Society, holds regular meetings for staff, patients, and family carers to exchange experiences and to learn from each other better ways of managing daily problems. There is a branch of Fountain House in Cape Town that provides excellent work experience for psychiatric patients in the community.

China

While the transformation of the Chinese economy created a wealthy stratum of society, many people were left in relative poverty. Hence China can still be included among LAMI countries. The investment in health services went mainly to the big hospitals, and primary health care, including care for psychiatric illnesses, was largely neglected (Li et al., 2005). The government's intention to develop community health care was impeded by the lack of training of staff in mental health care,

poorly equipped clinics, and the absence of health insurance cover for community mental health services except for medication and hospitalization. Phillips et al. (2009) estimated that of moderately or severely disabled patients with a psychiatric illness, only 8% have ever sought professional help, and only 5% have ever seen a psychiatric professional. In 2005, the trial known as the '686 Program' started to explore a service model for people with psychoses integrating hospital and community services. It has now spread to 112 sites with a catchment population of nearly 100 million (Ma et al., 2009). The process of transition to a community-based mental health service faces several obstacles: most Chinese psychiatric institutions are isolated from mainstream medicine, both academically and geographically, and the training curriculum for psychiatrists is mainly biological. It is noteworthy that until recently there was no Chinese word for unconscious.

On a visit to China 15 years ago, I was taken by my hosts to a remote rural area to see what was available for psychiatric patients. The small building I entered was dark and poorly furnished with a number of cell-like rooms and had no recreational facilities. On the other hand, the local village pharmacist we next visited raised my spirits. The pharmacist had been trained in the use of three medications, an antidepressant, an antipsychotic, and carbamazepine, which he used both to stabilize the mood of bipolar patients and to treat epilepsy, since it is active in both conditions. I asked him through the interpreter how many patients he sent to the nearest psychiatric hospital each year, and his answer was one. The ingenuity of this solution combined with strong central control, and a massive work force gives hope that China has the potential to create a modern psychiatric service firmly based in the community.

India

At the time of independence from Britain, India had almost no mental health services. There were psychiatric hospitals built by the British, such as the hospital in Agra, staffed by two psychiatrists and serving a population of 4 million. The whole country had 10 000 psychiatric beds for a population of over 300 million. In the first two decades after independence, the number of mental hospital beds was doubled and hospitals were humanized. Dr Vidya Sagar in Amritsar introduced the innovatory practice of encouraging the active involvement of families in the care of their relatives in psychiatric hospitals. The introduction of psychiatric admission beds into general hospitals was the next destigmatizing effort, which took place slowly. In 1975, there was a new initiative to integrate mental health services with general medical services in the community. Four decades later, this integration has been achieved in over 127 districts, representing 20% of the population. In 1982, The National Health Program was formulated to develop a national-level initiative for mental health care based on the community psychiatry approach. This led to the provision of many day care centers, half-way homes, long-stay homes, and rehabilitation facilities.

Despite these positive developments, a large number of people remain untreated, with a greater proportion in rural than urban areas, as was found in China. A major barrier to the use of services is the high level of stigma attached to mental illness, affecting social relationships and even more so employment. In addition, there are many different belief systems, some religious, others pertaining to traditional medical lore, such as Ayurveda and Unani. On visiting the Agra psychiatric hospital, I asked Dr Dube, the director, whether many of his patients consulted traditional healers. He replied that there was a temple of Hanuman, the monkey god, some distance from the hospital. Hanuman is consulted by many people for mental health problems. He went on to say that half the patients admitted to his hospital, had first been to the Hanuman temple. He added, "of course I only see the failures of the temple!" Traditional practitioners are not invariably antagonistic to Western medicine. I interviewed a traditional healer in a village near the Post Graduate Medical Research Institute in Chandigarh, in the Punjab. I discussed with him his treatment of depression. He used mainly psychological methods, sometimes in combination with herbs, but if these failed he told me, "I send the person to the hospital for electroconvulsive therapy (ECT)."

The legacy of the British Raj has been difficult to eradicate. A survey of mental hospitals was conducted in 1998 and found a catalog of the worst aspects of institutional care, including overcrowding, floor beds, complete lack of privacy in the use of toilets, isolation cells, deteriorating structures, and lack of trained staff. The report concluded that the rights of the mentally ill are grossly violated in mental hospitals. The problems are not confined to the institutions. Recent studies have revealed that about half the people with schizophrenia are living in the community without treatment, imposing a heavy burden on their families. Taking India as a whole, there are considerable deficits in the number of trained psychiatrists, psychologists, psychiatric social workers, and psychiatric nurses working in hospitals. These deficits are uneven across the regions with a few having adequate staff, and many with gross deficiencies. The 1998 report stimulated the provision of funds to improve conditions in the mental hospitals. A follow-up report in 2008 documented many positive changes, including reductions in admissions through courts from 70 to 20%, long-stay patients from 90 to 35%, and use of cells from 20 to 8 hospitals. In addition, there had been an increase in hospitals with recreational facilities from 8 to 29, and hospital with rehabilitation facilities from 10 to 23. Despite these progressive changes, the shortage of skilled staff persisted. The government's response was to provide generous funding for educational programs across the mental health care spectrum. In itself this did not tackle the problem of establishing a range of psychiatric facilities in the community. This deficiency was addressed by many NGOs, which have established day care centers, half-way homes, long-stay homes, and school and college mental health programs. Voluntary organizations can provide such services in a user-friendly way ensuring their acceptance by local communities. During the past decade, families have been playing a more active role with the formation of self-help groups, and professionals have been agreeing to work with families in a partnership model (Murthy, 2011).

There is no doubt that the new Indian Government of 1947 faced an enormous task in tackling the inhumane conditions in the psychiatric institutions built by the British, and in providing a comprehensive community psychiatric service for a vast and scattered population with inadequately trained personnel. They have made great progress in improving the conditions in the

outdated institutions, and in this area they are ahead of China. They have attempted to upgrade the specialist workforce, with moderate success, but have yet to mobilize the resources to create a functioning community alternative to institutional care. Instead, as happens in Western countries as well, the gap in services has been partially filled by voluntary organizations. However, the government has an obligation to ensure that they do not shirk their responsibility to consider the needs of the most disadvantaged citizens, leaving it to the goodwill of volunteers with limited resources to make up the deficiency.

Overview

From the examples presented it is obvious that the strategies required to create a well-functioning community psychiatric service depend on the previous history of psychiatric care in each country. In the United States and Western Europe the history of the building of psychiatric institutions is very similar, and by and large so is the availability of trained psychiatric staff. Hence the problems of transition to a community-based service do not differ dramatically. However, if we take the situation in Ethiopia, it might be considered fortunate that the authorities have only to phase out a single psychiatric hospital. However, the fact that only 17 trained psychiatrists are available for a population of 80 million stresses the magnitude of the tasks facing the government of a country still emerging from the effects of the murderous regime of the Derg, and among the 10 poorest countries in the world. I did see one hopeful sign of progress during a visit in 2010 when I was taken to a rural site on which an Italian construction company was building a very well-designed residential rehabilitation unit. However, when I asked the manager how long patients would stay in the unit, his answer was 'indefinitely.' It would be conceivable for Ethiopia to learn from the strategies of Tierra Colorada in Mexico and from Iran and train local people to become educators of their neighbors in mental health issues, a low-cost solution. It would not be feasible to integrate traditional healers into the psychiatric work force as in South Africa, since holy water and beatings are useless as treatments for psychoses, although holy water might exert a placebo effect on mild neuroses.

Egypt was showing signs of positive changes in its psychiatric services until the recent 'Arab Spring' and it is difficult to foresee any further progress until the political situation stabilizes. Fundamentalist governments that discriminate against women are unlikely to take a stand against the stigma of mental illness. The inertia of the Chilean government in taking the first step of humanizing its psychiatric hospitals does not inspire hope of change there. Jamaica is a shining example of how a colonial legacy can be transformed, but needs a charismatic leader like Professor Hickling to see the process through. The ingenious solution of the South African Health Train might be replicable in other LAMI countries with scattered rural populations, but would need the goodwill of the railway companies to kick-start it.

The two population giants, China and India, have different histories and therefore different trajectories. India has made admirable progress in emerging from its situation as a British possession, but its intention to create a well-functioning community-based psychiatric service is a daunting task in a subcontinent with a multitude of different ethnic groups, languages, and systems of treatment, and a caste system with ingrained stigma. China has shown its capacity to successfully transform its social structure, and given the will, could rectify the poor conditions in rural area where half the patients with psychoses have never received any treatment. However, it requires a government that can place the treatment and care of people with psychiatric illnesses high up its list of priorities. The argument presented by the WHO that mental illness incurs a very large cost to a country in terms of lost ability to earn a living and the financial burden on family carers should be an inducement. Voluntary organizations in almost every country have filled the gaps in the services for mentally ill people, and have done this very well. However, the experience in the United Kingdom has demonstrated that the determination of successive governments to carry through a policy of deinstitutionalization is crucial for the success of this humanitarian endeavor.

Acknowledgment

I am deeply indebted to Dr R. Srinivasa Murthy for sending me his paper on the development of psychiatric services in India.

See also: Comparative Health Systems; Culture as a Determinant of Mental Health; Healing Practices: Alternative and Complementary; Health Education and Health Promotion; Health Interventions, Community-based; Hospitals and Health: Psychosocial Aspects; Mental Health; Nonphysician Health Professionals.

Bibliography

Behr, G.M., Allwood, C.W., 1995. Differences between Western and African models of psychiatric illness. South African Medical Journal 85, 580–584.

Harvey, Y.K., 1976. The Korean Mudang as a household therapist. In: Lebra, W.P. (Ed.), Culture-Bound Syndromes, Ethnopsychiatry and Alternative Therapies. University Press of Hawaii, Honolulu, pp. 189–198.

Hickling, F.W., 2010. Psychiatry in Jamaica. International Journal of Psychiatry 7, 9–10.

Johnson, S. (Ed.), 2010. Residential alternatives to standard acute psychiatric wards. British Journal of Psychiatry 197 (Suppl. 534).

Leff, J., 1988. Psychiatry Around the Globe: A Transcultural View. Gaskell Psychiatry Series. Royal College of Psychiatrists, London, p. 139.

Leff, J., et al., 2000. The TAPS project: a report on 13 years of research, 1985–1998. Psychiatric Bulletin 24, 165–168.

Li, Y.-H., Yao, X.-W., Zhang, M.-Y., 2005. Investigation and suggestion on community rehabilitation facilities for psychotic patients in Shanghai. Shanghai Archives of Psychiatry 17 (Suppl.), 35–37 (in Chinese).

Loza, N., 2010. Integrating Egyptian mental health services into primary care: the policy makers perspective. International Psychiatry 7, 5–6.

Ma, H., Liu, J., Yu, X., 2009. Development and interpretations of Chinese mental health policies in the last decade. Chinese Journal of Mental Health 23, 840–843.

Murthy, R.S., 2011. Mental health initiatives in India (1947–2010). The National Medical Journal of India 24, 26–35.

Phillips, M.R., Zhang, J.-X., Shi, Q.-C., et al., 2009. Prevalence, treatment and associated disability of mental disorders in four provinces in China during 2001–05: an epidemiological survey. Lancet 373, 2041–2053.

Prince, R., 1960. The use of Rauwolfia for treatment of psychoses by Nigerian native doctors. American Journal of Psychiatry 117, 147–149.

Torrey, E.F., Kennard, A.D., Eslinger, D., et al., 2010. More Mentally Ill Persons Are in Jails or Prisons than in Hospitals. A Survey of the States Treatment Advocacy Center (TAC) and National Sherrifs Association (NSA). http://www.treatmentadvocacycenter.org/storage/documents/final_jails_v_hospitals_study.pdf.

Deinstitutionalization, Psychiatry of

Julian Leff, University of London, London, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 5, pp. 3366–3371, © 2001, Elsevier Ltd.

Abstract

The spread of asylums throughout Europe and North America is charted and the use of asylums to contain the rejects of society described. Figures are given showing the peak bed occupancy in various countries and the fall in beds thereafter. The turnaround is ascribed to the effects of World War II on military psychiatrists, the introduction of antipsychotic medication, public scandals around the abuse of patients, and sociological critiques of institutions. Political backing for the transition to community psychiatry is discussed with particular reference to the situations in Italy and the UK. Community services developed for the long-stay patients are described and the results of some long-term follow-ups presented. Attention is drawn to the barriers to social integration of patients in the community, and to the lack of job opportunities. The costs of hospital and community care are presented from two studies and the need for cost-benefit analyses stressed. Critiques of deinstitutionalization from professionals, the media, and the public are discussed, with a focus on issues of violence by patients, imprisonment, and homelessness. The distinction is drawn between services for the long-stay and for the acutely ill, and the need for a comprehensive service is emphasized.

The term 'deinstitutionalization' is a somewhat cumbersome shorthand for the process that has occurred in North America and Western Europe of shifting the locus of psychiatric care from the large dedicated psychiatric hospitals to the districts in which people live. In order to appreciate the impetus behind deinstitutionalization, it is first necessary to review the reasons for the establishment of asylums. The rundown and closure of the asylums raises fundamental questions about the attitudes of the public to the mentally ill and the nature of 'the community.'

The Rise of the Asylum

One of the oldest psychiatric institutions in the West is the Bethlem hospital, founded as a priory in 1247 and used to house the mentally ill from 1403 to the present day. Its name, corrupted to Bedlam, attracted the stigma that became an inescapable feature of such institutions and was used to vilify the psychiatrically ill as bedlamites. Similar hospitals were established throughout Europe and North America in the ensuing centuries but the massive expansion of psychiatric hospitals, or asylums as they became known, did not occur until the nineteenth century. In the UK, this was as a result of government policy to provide each county with at least one public asylum, as legislated in the County Asylum Act of 1808. Large sums of money were made available for the asylum building program, of which the Victorians were very proud. Considerable effort went into the design of the buildings, with particular attention to the circulation of air throughout the asylum and to the external decoration.

Asylums were usually sited some miles outside of towns or cities: far enough away to avoid disturbing the public, but near enough to facilitate the transport of patients. Friern Hospital was constructed on a rural site north of London and the nearest railway line was extended to make it accessible. The hospital was opened in 1851 by Prince Albert, the husband of Queen Victoria, and was designed to house a thousand patients. It rapidly filled up and its population continued to grow way beyond the original limit, reaching a peak of 2400 a century after its inauguration. This rapid expansion was a general feature of asylums. Once a place of containment became available, the opportunity was seized to expel from society people with a wide variety of problems who were seen as undesirable. These included people with developmental disabilities, who were sometimes sent to segregated institutions and sometimes mixed indiscriminately with the mentally ill. There were even instances of women bearing illegitimate children who, when pregnant, were sent to such institutions where they spent the rest of their life.

As their populations grew steadily, the asylums became increasingly overcrowded, until they reached the point where the beds were so close together that patients had to climb over others' beds to reach their own. Overcrowding generated squalor, and since the staff were unable to cope with the large number of disturbed patients, it was almost inevitable that abuses occurred. The public scandals that ensued hastened the demise of the asylums, but before charting their decline it is worth considering how they functioned in their heyday.

The typical asylum was either a very large building or a collection of villas set in extensive grounds, which were surrounded by a high wall with a single entrance guarded by a gate lodge. Within the walls were vegetable gardens and often a farm with cows and sometimes pigs. Some asylums had their own artesian wells and many were virtually self-sufficient or even produced a surplus. Powick hospital in the west of England during the 1920s was selling one thousand bushels of wheat and 1800 pounds of beef per year. The patients provided the labor to work the farms, as well as helping in the hospital kitchens and the laundry. The use of unpaid patient labor was viewed by the trade unions in the West as exploitation and it ceased in most hospitals in the mid-twentieth century.

When patient labor was replaced by a paid professional workforce, activities for the patients had to be organized. These took the form of recreational activities, designated as occupational therapy, and work-like activities, often termed industrial

therapy. Industrial therapy ranged from simple repetitive tasks such as assembling darts to complex procedures such as making furniture designed by the staff. The industrial therapy unit at Friern Hospital, at its peak, employed 120 people, some of whom commuted daily from their homes since nothing comparable existed in the community.

The Turning Point

In the UK, the number of psychiatric beds climbed steadily for more than a century, reached a peak of 149 000 in 1954, and since then has fallen progressively. A similar pattern occurred in the USA, where the number of residents in state psychiatric hospitals declined from 186 per 100 000 in 1969 to 33 per 100 000 in 1992. Psychiatric bed occupancy has also been falling in most Western European countries since the 1970s but the rate differs by country. In Italy the number of psychiatric hospital beds rose from 36 000 at the beginning of the century to a peak of 188 000 in 1971. Since then there has been a steady decline to 7500 in 1999, including both state and private hospitals. In Israel the number of psychiatric hospital beds has fallen from more than 11 000 to 6000 in 1999. Four factors can be identified that prompted this change: the introduction of specific antipsychotic drugs, abuse of patients by staff that provoked public scandals, critiques of institutions by sociologists and others, and a change in the attitudes of psychiatrists.

It has been claimed that the key factor was the advent of antipsychotic drugs, but that cannot be the only answer since chlorpromazine, the first of the genre, came into general use in 1955, whereas in some pioneering psychiatric hospitals in the USA and UK the bed numbers began to fall in the late 1940s. By controlling disturbed and unacceptable behavior, the drugs certainly facilitated the process of deinstitution-alization, but they did not initiate it. Similarly, public scandals gave a high profile to the problems of institutions and helped to mobilize a search for alternative forms of care, but did not start the process.

Academic books criticizing psychiatric institutions began to appear at the end of the 1950s and certainly exerted some influence on the attitudes of psychiatrists and other staff. The first was by a psychiatrist, Russell Barton, and appeared under the title *Institutional Neurosis* in 1959. The term he invented was inappropriate since the great majority of patients adversely affected by institutional living suffered from psychotic illnesses. However he accurately described the apathy, inertia, lack of motivation and affective poverty which characterised long-stay patients in institutions. This was followed by a particularly influential book by Erving Goffman (1961) entitled *Asylums*. Goffman spent a year in the late 1950s observing interactions in a US psychiatric hospital. He described the humiliating rituals which were inflicted on new entrants and saw their main purpose as being to sever all links with the everyday world outside the institution. He coined the term 'total institution,' which he defined as 'a place of residence and work where a large number of like-situated individuals, cut off from the wider society for an appreciable period of time, together lead an enforced, formally administered round of life.' Asylums certainly provided accommodation and work, but in addition many supplied every amenity likely to be required in everyday life, including a place of worship, a patients' club, hairdressers, and shops, so that there was no need ever to leave the confines of the hospital.

These books, however seminal, appeared in print too late to explain the fall in beds in the pioneering asylums. For example, Dr. Duncan Macmillan, who occupied the unique position of Medical Officer for Mental Health for Nottingham, UK, established an outpatient clinic in the local general hospital in 1932. Later he became Physician Superintendent of the psychiatric hospital, Mapperley, and by 1957 the senior members of the hospital's medical staff were spending over 50 percent of their time in 34 clinics in the community. As a consequence of these developments, the number of inpatient beds began to fall in 1949.

World War II seems to have provided psychiatrists with experiences that led to innovative experiments in therapy. This entailed the recognition ofthe legitimacy of fear in the face of appalling danger, and affected soldiers were treated with psychological approaches. Military psychiatrists saw healthy young men break down with severe psychiatric syndromes under stress, and then recover when treated in group settings. In these military centers every aspect of the regime was designed to counteract the development of chronicity. At the end of the war, the army psychiatrists, imbued with optimism about the potential for recovery, took up posts in the psychiatric hospitals. They included British pioneers such as Maxwell Jones (1952) who founded the Henderson hospital as a therapeutic community for people with psychopathy, and Tom Main, who took over and transformed the Cassell hospital. As directors of psychiatric hospitals, these charismatic men had the power to make their vision a reality, unimpeded by bureaucratic inertia. However, they had to fight the resistance of other members of staff, who were often more institutionalized than the patients.

In the USA the impetus for attitude change also arose from the war, but from a surprisingly different source. Conscientious objectors were obliged to do public service in psychiatric hospitals. They brought into these institutions a humanitarian attitude which had not been dulled by the hospital regime. Furthermore, as they had no professional training, they treated the patients as equals and as valued individuals. The ferment they introduced into the system worked its effects on the professional staff.

Decline and Fall of the Psychiatric Hospital

Once the antipsychotic drug, chlorpromazine, became available, the handful of pioneering hospitals were joined by the rest in discharging patients in large numbers, and a steady decline in the number of occupied beds occurred in the USA and in England and Wales. This was not observed in Scotland, which had a separate health administration. The decline in England and Wales was so steady over a number of years that Tooth and Brooke (1961) predicted that the existing long-stay population of psychiatric hospitals would be eliminated through discharges and deaths by the year 1986. The reasons their prediction proved to be wrong are of considerable import.

First, some patients admitted to acute wards remained in hospital for over a year. They were termed the 'new long-stay' in distinction to the old long-stay who were a relic of custodial care. Of course the distinction between the two is arbitrary depending on the point in time when observations are made. The significance of this group is that their problems are clearly not created by the institution but are a consequence of their psychiatric illness. Hence closure of the institutions will not abolish the 'new long-stay.' The second group of patients that confounded Tooth and Brooke's prediction are those whose behavior is so disturbed and disturbing that they cannot be moved to homes in the community. These 'difficult-to-place' patients pose a major challenge to deinstitutionalization and will be considered in more detail later.

Despite these problems, the run-down of the asylums has progressed to the point where many of them have been closed. In the USA the number of state psychiatric hospitals decreased from 277 in 1970 to 231 in 1996. There has been an even more dramatic change in England and Wales where 100 of the 130 psychiatric hospitals have closed since 1980. In Italy the closure of the institutions was achieved by revolution rather than evolution. In 1978 the Italian parliament passed Law 180 which prohibited the admission of patients to psychiatric hospitals thenceforth. This legal constraint on admissions did not directly affect the patients already resident in the asylums, and it took many years to improve their conditions of life.

Law 180 reminds us that changes in psychiatric services cannot be carried through by staff alone. They require the backing of the government of the day to provide the financial resources necessary to create and develop new services. The Italian law was pushed through parliament by the Communist party, whose support was recruited by Franco Basaglia, a charismatic psychiatrist and founder of the Democratic Psychiatry movement. In the UK there was no political polarization over deinstitutionalization. The closure of psychiatric hospitals was heralded by the Conservative Minister of Health, Enoch Powell, in a famous speech in 1961 in which he referred to the asylums as 'isolated, majestic, imperious, brooded over by the gigantic watertower and chimney combined, rising unmistakeable and daunting out of the countryside.' Successive governments, whether of the right or the left, continued to endorse the policy, but with different agendas: the left supported patients' rights for ideological reasons, while the right hoped to save money by closing the old asylums.

Alternatives to the Psychiatric Hospital

Where were patients to live when they were discharged from an asylum? Wherever the process of discharge has been studied it has emerged that the least disabled patients were selected to move out first. Many of these patients were able to cope in independent accommodation. As the more dependent patients were considered for discharge, it became evident that they needed supervision in their community homes. For the most disabled patients, 24-hour supervision was required. The closure process for two London psychiatric hospitals, Friern and Claybury, was studied by the Team for the Assessment of Psychiatric Services (TAPS) over 13 years (Leff 1997). TAPS found that 78 percent of the long-stay patients were discharged to staffed homes. In the USA, Eileen Rothbard and her colleagues (1999) monitored the closing of Philadelphia State Hospital and found that 75 percent of the long-stay population were living in residences with 24-hour staff supervision. Intensive staffing invariably carries a high cost, since over 80 percent of the cost of care in the community is contributed by staff salaries.

The obvious choice might seem to be to resettle patients with their families, but this is rarely achieved for the long-stay patients. In the Friern and Claybury programs only 4 percent of them returned to their family home. This is partly due to loss of contact with relatives over the course of decades in an institution, which had occurred to 75 percent of the Friern and Claybury patients. The other reason is that a high proportion of patients are experiencing active delusions and hallucinations which are not controlled by medication and require specialized care.

In many of the resettlement programs, the more disabled patients live in small groups in ordinary houses, indistinguishable from neighboring buildings. This has an obvious advantage in easing reintegration into the community. However there are barriers other than distinctive architecture, which became apparent from the TAPS evaluation. Nearly 700 long-stay patients were followed up for five years after discharge to homes in the community. Although there was no amelioration of their psychiatric symptoms or problems in social behavior, they did acquire skills in caring for their homes and in using community amenities. Compared to the hospital wards, they enjoyed much greater freedom in their homes, which they greatly appreciated. When asked where they wanted to live, 84 percent preferred to stay in their current home, and almost all those who wanted to move wished to live independently. Only a handful desired to return to the hospital.

Beneficial outcomes for long-stay patients discharged from psychiatric institutions have been reported by other researchers in the UK and the USA, although few have studied a large sample of patients over a long period of time. An exception is Courtney Harding and her colleagues (1987) who conducted a 32-year follow-up of 269 patients discharged from Vermont State Hospital, USA in the 1950s as part of an experimental rehabilitation program. The findings were that over half the patients were functioning much better than expected.

One drawback to life outside the asylum that TAPS revealed was the failure of most patients to develop social relationships with their neighbors. Undoubtedly the impairment of social skills by schizophrenia was contributory, but it is not the whole story. Most of the patients were resettled in cities, where there is not a strong sense of community and even healthy neighbors have little to do with each other. Geoffrey Wolff and his colleagues (1996) mounted an experimental study of an education campaign in a street in which long-stay patients from a south London asylum were resettled. The campaign was successful in reducing neighbors' fear of the mentally ill and in increasing their social contact with the new residents.

Another problem is the lack of employment prospects for people with a long history of psychiatric hospitalization. When the national unemployment rate is high, it is almost impossible for someone with such a history to obtain a job. One solution is to establish cooperatives, or 'social firms' as they are sometimes known. These are particularly well-developed in Italy, and a comparison of people with schizophrenia between

Boulder, Colorado, and Bologna, Italy (Warner et al.1998) showed that the Italian patients worked for more hours and earned more than the US patients. The development of employment opportunities for mentally ill people in the community needs far more attention than it has received to date, as it is one of the prime means of social integration.

Is Community Care Affordable?

It is not difficult to provide cheap community care. There have been several low-cost schemes to house discharged patients in the USA and the UK. In the USA, 'board and care' was an inexpensive way to provide patients with accommodation, but very little else. The proprietors of these homes ran them as commercial enterprises and had little or no instruction in the care of the patients. It is hardly surprising that patients were treated extremely badly in some cases, giving rise to public scandals. In the UK, patients were placed in seaside boarding houses, many of which are empty during the winter and would otherwise have closed until the summer season. Hence the offer of permanent residents all the year round was attractive to their owners. However, no provision was made for the patients' work and leisure needs, and television documentaries showed pathetic scenes of patients wandering along deserted wind-swept promenades in the driving rain. Patients placed in these seaside towns were even further from their districts of origin than when in the asylums.

The crucial question is how a *good* system of community care compares with hospital care in terms of cost. There have been very few attempts to make this comparison. The TAPS project involved a collaboration with a team of health economists who carried out a detailed costing of all services used by the long-stay patients when in Friern and Claybury hospitals and then when they had been in the community for one and five years. As this was done prospectively, the costs for specific individuals can be compared over time, adjusting for inflation. For all the discharged patients, the cost in the community after the first year was slightly less than their cost when in hospital, although the community costs rose slightly by five years (Leff 1997). However, not all the long-stay patients were discharged to community homes. A sizeable group were considered to be too disturbed to move to the community. When the hospital closed these difficult-to-place patients were transferred to highly staffed facilities, mostly on other hospital sites. The high staffing levels made their care very expensive, so that when they were added into the cost equation, community care became more expensive than care in the hospital had been.

By contrast, a costing exercise by Eileen Rothbard and her group (1999) for patients discharged from Philadelphia State Hospital showed that community care was considerably cheaper than the estimated cost of hospital care. Cost comparisons on their own are insufficient: what is needed are cost-benefit analyses across sites and countries.

Critiques of Deinstitutionalization

Criticisms have come from within the psychiatric profession and from outside, mainly from the media, from disgruntled family members, and from citizens who object to sheltered homes being established in their neighborhood. One of the most stringent professional critics has been Harold Lamb (1976) in the USA who has pointed to the large number of severely mentally ill people who are in prison. He argues that patients have been transferred from a relatively benign institution, the psychiatric hospital, to a punitive one, prison, and has dubbed this process 'transinsti-tutionalization.' Two related issues are raised by his critique: does community care inevitably lead to the imprisonment of the mentally ill, and are the mentally ill more dangerous than the average citizen? In the TAPS study, of the 670 patients followed up for five years after discharge only five spent time in prison, while 13 (2 percent) committed assaults on others. In the study by Eileen Rothbard and colleagues (1999), 2 percent of the 321 long-stay patients followed up for three years were involved in the criminal justice system. Thus it appears that a well-organized and funded community program does not lose many patients to prison, and furthermore that the great majority of mentally ill people are not dangerous. Unfortunately the media prefer to highlight the few tragic homicides that do occur, hence reinforcing the public stereotype of the 'mad axeman.' It is this irrational fear that fuels the 'not in my backyard' attitude of some members of the public. Geoffrey Wolff and his colleagues (1996) have shown that a localized educational campaign can reassure neighbors about ex-hospital patients coming to live in their street.

The other concern of the media and the public is that many patients discharged from long-term care end up on the streets. While the number of homeless mentally ill has undoubtedly increased, several surveys have found that very few of them have spent much time in psychiatric institutions. In support of this observation, less than 1 percent of the TAPS sample became homeless over five years and only 2 percent of the Philadelphia State Hospital ex-patients over three years. While no mentally ill person should be without a home, the problem stems from the inadequacy of aftercare for those coming out of admission units and not from the long-stay patients who are resettled in sheltered residences.

This raises the question of the balance of resources to be invested in community care programs. The asylums represent a large investment of capital, in the land and buildings, and of revenue, in staff salaries. The value of the site is often considerable since urban expansion has incorporated the once distant asylums. Various financial strategies have been used to release the capital and revenue bound up in the asylum, so that they can be reinvested in providing a good quality service for the long-stay patients. But patients with long-term disabilities continue to arise in a population and need the kind of rehabilitation facilities which used to exist in the psychiatric hospitals but have generally not been re-provided in the new services. Additionally, the care of acutely ill patients is becoming increasingly difficult with the ready availability of street drugs. It seems to be difficult for funders to invest in acute and rehabilitation services at the same time as providing for the long-stay patients coming out of the institutions. Within North America and Europe, only Finland seems to have achieved a reasonable balance between these two sectors. Without such a balance, it is not surprising that family members are discontented with the service offered to their acutely ill

relatives, and that deinstitutionalization has become the target of criticisms that should justly be directed at the failure to provide a *comprehensive* psychiatric service of a uniformly high standard.

Bibliography

Etzioni, A., 1961. A Comparative Analysis of Complex Organizations. Free Press of Glencoe, New York.

Goffman, E., 1961. Asylums: Essays on the Social Situation of Mental Patients and other Inmates. Anchor Books, Doubleday, New York.

Harding, C.M., Brooks, G.W., Ashikaga, T., Strauss, J.S., Breier, A., 1987. The Vermont longitudinal study of persons with severe mental illness, II: Long-term outcome of subjects who retrospectively met DSM-III criteria for schizophrenia. American Journal of Psychiatry 144, 727–735.

Jones, M., 1952. Social Psychiatry: A Study of Therapeutic Communities. Tavistock Publications, London.

Lamb, H.R., 1976. Community Survival for Long-Term Patients. Jossey Bass Publications, San Francisco.

Leff, J. (Ed.), 1997. Care in the Community: Illusion or Reality?. John Wiley, Chichester, UK.

Rothbard, A.B., Kuno, E., Schinnar, A.P., Hadley, T.R., Turk, R., 1999. Service utilization and cost of community care for discharged state hospital patients: A 3-year follow-up study. American Journal of Psychiatry 156, 920–927.

Tooth, G.C., Brooke, E.M., 1961. Trends in the mental hospital population and their effect on future planning. Lancet I, 710–713.

Warner, R., de Girolamo, G., Belleli, G., Bologna, C., Fioritti, A., Rosini, G., 1998. The quality of life of people with schizophrenia in Boulder, Colorado, and Bologna, Italy. Schizophrenia Bulletin 24, 559–568.

Wing, J.K., Brown, G.W., 1970. Institutionalism and Schizophrenia: A Comparative Study of Three Mental Hospitals 1960–1968. Cambridge University Press, Cambridge, UK.

Wolff, G., Pathare, S., Craig, T., Leff, J., 1996. Public education for community care: A new approach. British Journal of Psychiatry 168, 441–447.

Deixis

Jürgen Bohnemeyer, University at Buffalo, The State University of New York, Buffalo, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Deictic signs symbolically denote some aspect or component of the *speech situation* – the extra-linguistic context in which they are used, such as the speaker, time, or place of an utterance. Although they are symbolic as types, their tokens in specific contexts represent the particular component of the speech situation indexically. The semiotic nature of this kind of reference, its exact communicative prerequisites and functions, its acquisition by children, and its processing have long puzzled scholars throughout the cognitive sciences. This article presents an overview of deictic signs and introduces some major research themes.

Introduction

The term *deixis* was introduced by the German linguistic and psychologist Karl Bühler (1879–1963) in his 1934 magnum opus *Sprachtheory* ('theory of language,' or 'linguistic theory'). It refers to the central phenomenon that Bühler dedicated his book to the fact that when we speak (or sign, in the sense of using sign language), our utterances are peppered with expressions/signs that refer to components of the *speech situation* – the extralinguistic (i.e., real-world) context of the utterance. Take, for example, the utterance *She brought this flower for me yesterday*. The pronoun *me* refers to the speaker. It has a different meaning depending on who utters it. This is an instance of 'person deixis.' The demonstrative *this* selects a referent in the interlocutors' proximity – *this flower*, as opposed to *that* one over *there* (at least in its most simple spatial use, exemplifying 'spatial deixis'). The verb *bring* designates transport to a deictically defined location (*here*); this could be the location at which the conversation takes place, or the speaker's home (there are in fact many possibilities). This illustrates another subtype of spatial deixis – 'place deixis.' The past tense of *brought* indicates that the flower arrived at this location prior to the time of utterance, and the adverb *yesterday* restricts this time interval to the day before the day of utterance. Both instantiate 'temporal deixis.' So, in order to know what exactly is meant by *She brought this flower for me yesterday*, and whether this statement is true, one first needs to know who uttered it, on what day, and where. No language has been attested to have a genuine event deictic, but many languages (e.g., Yucatec Maya) have manner indexicals (translating 'like this/that') that are used in reference to events. Further possible types of deictic expressions are discussed in the Section 'Deictoids: Honorifics, Projective Spatial Relators, and Modals.'

Deixis derives from the Greek verb *δείκνυμι* 'show,' 'point,' a metaphor Bühler uses to refer to the fact that deictic signs receive their *referent* – the particular component of the speech situation they refer to – via a relation of spatial and temporal contiguity between the utterance and the speech situations. In this capacity, Bühler likens them to pointing gestures, which often accompany the use of *deictics* – hence the metaphor. Nearly 70 years before Bühler, the American philosopher Charles S. Pierce (pronounced like *purse*) – now often considered one of the most original thinkers of the Modern Age – proposed a classification of signs with roots in medieval philosophy. Peirce treats as signs anything that can be used as a representation of something else – not only linguistic expressions and diagrams, but also cognitive representations or thoughts. Peirce (1867) classified signs based on the relationship between their form and the referent into *icons* (constituted by similarity; e.g., portraits, blueprints, and maps), *indices* (singular *index*; constituted by a causal relation or one of spatiotemporal contiguity; e.g., footprints as a sign of the person or animal who left them), and *symbols* (constituted by convention; e.g., hammer and sickle as the symbol of Communism).

Index of course invokes the same metaphor as *deixis*. However deictic signs are not primarily indices but rather symbols – as are, arguably, all linguistic signs. To see this, consider the fact that the meaning conveyed by the deictic pronoun *I* in English is expressed by *moi* in French, *yo* in Spanish, and *tèen* in Yucatec Maya. Obviously, the combination of the particular string of sounds (or the letters that stand for them in written speech) and this meaning is conventional, in the sense that for any individual speaker, the particular string is associated with this meaning only because they have learned this association from their peers. However, there is a secondary indexical association between the referent of the sign in a given context and, not the sign as an abstract object of the speaker's knowledge and practice, the 'type' in Peirce's terms, but rather the particular instance of the sign used in the particular context – the particular 'token' of the sign, as Peirce puts it. For example, with every 'turn' in a conversation the word *I* refers to a different speaker, identified by the indexical relation between the utterance (and thus the tokens of *I* contained in it) and the person producing it. So deictic signs are symbols as types, but their tokens in any particular utterance bear an indexical relation to their referents (cf also Silverstein, 1976; Tanz, 1980; inter alia).

Linguistic meaning is studied by a number of subdisciplines of linguistics and neighboring sciences, including notably semantics, pragmatics, and linguistic anthropology. Semantics is primarily concerned with the study of the meanings of (simple and complex) linguistic signs, whereas pragmatics is the study and theory of utterance meanings. (Utterances in

a narrow sense may be understood as actions of language use with recognizable communicative intentions.) Linguistic anthropology in turn is dedicated to the study of cultural practices of language use. Because deictic signs are a manifestation of the context-dependence of linguistic meaning sui generis, deixis has received more attention among students of pragmatics (e.g., Levinson, 1983, 2004) and linguistic anthropology (e.g., Levinson, 2004; Silverstein, 1976) than among semanticists (e.g., Schlenker, 2003). Deixis has also been studied extensively elsewhere in the cognitive sciences, notably by philosophers of language (e.g., Barwise and Perry, 1983; Braun, 2012; Kaplan, 1989) and developmental psychologists (e.g., Piaget and Inhelder, 1956; Tanz, 1980).

Deixis, Indexicality, and Context Dependence

While the term 'deixis' has become customary among linguists and psychologists, philosophers of language such as Kaplan (1989) often prefer the Peirce-inspired *indexicality* (alongside *indexicals* for deictics) with roughly the same meaning. This can cause confusion, as token indexicality in Peirce's sense is not restricted to deictics, but characterizes a considerably larger class of linguistic signs.

It is one of the fundamental design features of human language that the interpretation of linguistic utterances may strongly depend both on the linguistic and the nonlinguistic context. It has been argued by philosophers that in fact without some such underlying indexicality in all referring expressions, no successful reference to the world would be possible (Putnam, 1975). Take again the example *She brought this flower for me yesterday*. The noun *flower* has a meaning that can be defined independently of context (e.g., 'the reproductive organs of a plant and their colored (nongreen) envelope'), and it can be used to refer to real or imaginary flowers that fulfill this definition. Yet, in order to refer successfully to a plant as a *flower*, more than mere knowledge of the meaning of *flower* is required: speaker and addressee have to agree on the plant being identifiable by that term. And the conditions on this agreement may in fact vary with context. For instance, if one of the interlocutors is a botanist, they might use the term *flower* differently, not requiring the envelope to be colored.

However, there are expressions whose very meaning inherently refers to the context. Such expressions cannot be used to refer to anything before the relevant information is retrieved from the context. Deictics instantiate this class of expressions – but they do not exhaust it. In the above example, the pronoun *she* takes up a referent of the feminine gender that must have been introduced in the preceding stretch of discourse. If the speaker announced in the preceding utterance that he or she is planning to marry his/her girlfriend Sally, then it will be inferred that it was Sally who brought him/her the flowers. This illustrates *anaphoric* reference. In contrast, *deictic* reference occurs whenever a linguistic sign receives part of its meaning from the extra-linguistic context.

Anaphoric expressions generally admit deictic use, while anaphoric use of inherently deictic terms is usually more restricted (but see Section 'The Semantics of Deixis'). The English demonstratives *this* and *that* can both be used anaphorically, though under slightly different conditions (see Fillmore, 1997), and so can the proximal and distal place adverbs *here* and *there*. In contrast, in Yucatec Maya, only the distal demonstrative forms and adverbs can be used anaphorically (Bohnemeyer, 2012b; Hanks, 1990). The combined impact of the rather weak deictic anchoring of demonstratives in some languages and the pervasiveness of their nonspatial uses has led some researchers to propose nonspatial analyses of the underlying meanings. Similarly, Wilkins and Hill (1995) show that *go* and its equivalents in other languages only acquire a deictic reading pragmatically, through the contrast with a truly deictic *come*.

Another class of indexicals are so-called *definite descriptions* – phrases formed with a definite article; in English, with *the*. For instance, *the chair* refers to a chair that the speaker assumes to be uniquely identifiable to the addressee, perhaps because it has been mentioned before or because there is only one in the room, and so on. Therefore, definite descriptions can be used deictically provided the uniqueness condition is fulfilled in the context: *Move the chair over!* will be felicitous in a situation in which there is only one free chair in the room and the addressee is aware of that. This kind of use extends to other semantically definite expressions as well, such as determinative adnominal possessive constructions (*Floyd's/ my chair*).

It can be argued that reference to any particulars (people, animals, objects, events, places, and times) that lack a proper name requires some form of deictic anchoring. To understand this claim, one may by way of a thought experiment select an object from the environment at random and try to make a statement about it that avoids any form of deixis. If the claim is correct, it follows that language could not be used to talk about the real world (other than in generic statements) without deictic reference.

Between Deictic and Anaphoric Reference: Transposed Deixis and Text Deixis

'Textual' deixis occurs when (part of) an utterance or discourse is itself the referent of a deictic expression, as in *The preceding sentence contains 27 words*. It has been suggested that some form of textual deixis is present in every anaphoric reference, to the extent that anaphors direct the addressee's attention to an earlier mention of the referent. As long as it is clearly discernible as deixis, textual deixis will always have a *metalinguistic* reference. But as soon as any part of the *meaning* of an utterance becomes the target, textual deixis becomes increasingly indistinguishable from anaphoric reference (Levinson, 2004). An intermediate case is constituted by reference to propositions, facts, and events (cf Lyons, 1977).

'Transposed' deixis is constituted by an *imagined* situation replacing the actual utterance context as the 'indexical ground' of a deictic form. When the imagined situation is itself described in discourse, this may have a quasi-anaphoric effect. Consider *He arrived in Paris in early May. Now he finally had the time to explore this great city*. The deictic forms in the second sentence refer to the time and place introduced in the first sentence. However, they are relativized to the character perspective and therefore function as true deictics. This perspectivizing effect was actually one of the primary

concerns of Bühler (1934) (cf also Section 'Introduction; Deictoids: Honorifics, Projective Spatial Relators, and Modals').

Deictoids: Honorifics, Projective Spatial Relators, and Modals

I would like to propose *deictoids* as a cover term for several classes of indexical expressions that are commonly considered deictic because their use is dependent on properties of the extra-linguistic context, but which differ from the prototypical instances of person, spatial, and temporal deictics discussed so far in that the latter, but not the former, semantically denote variables the value of which is either a constituent of the speech situation or, as in the cases of tenses and demonstratives, is defined via a relation to a constituent of the speech situation.

For instance, '*I*' denotes the speaker of the utterance, whoever that may be in any given instance. '*Now*' denotes the time of the utterance, whenever that may be in any given instance. '*This*' denotes an object located near '*here*,' the 'deictic center' (Bühler, 1934), the place occupied by one or all of the interlocutors. And (deictic or 'absolute') tenses denote times in the present, past, or future of utterance times. This is not really the case with *honorifics*. In popular usage, the prototypical 'honorifics' are titles, such as *Ms. (Brown)* or *Her Royal Highness*. But in scientific usage, 'honorifics' encompasses any and all semantic expressions of the social relations between the participants of the speech act the utterance is involved in carrying out – the speaker, addressee, some person referred to, and/or bystanders of the interaction, usually as a means of expressing respect or deference for or solidarity with them. For example, most European languages contrast two addressee pronouns, one that is exchanged among speakers who are both 'solidaries' (friends, relations, or members of the same 'in-group') and status equals, the other taking over when one or both of these relations are absent, with the precise conditions varying from language to language and being subject to change over time (Brown and Gilman, 1960). Other languages (a particularly well-known case is Japanese; e.g., Shibatani, 1990: pp. 374–379) have more complex systems that go beyond binary contrasts and beyond the speaker–addressee relation. In addition, apparently all languages have expressions that do not directly express social relations among the interlocutors, but whose use nevertheless implies them, including a plethora of strategies to convey various forms of politeness (Brown and Levinson, 1987).

All of these expressions are context-sensitive, since their use reflects the speaker's construal, for the purposes of the particular utterance, of the social relations among the interlocutors. By the same token, such expressions are indexicals, and in fact resemble deictics, in that it is specifically the extralinguistic context that they are sensitive to. Fillmore (1997[1975]) in one of the classic texts on deixis, in fact, proposed to treat these under the new category of *social deixis*. However, for, say, an honorific to have a semantics akin to those of prototypical deictic expressions, it would have to encode the relation between – say – the speaker and addressee as a variable, the value of which is picked up from the speech situation. Consequently, one and the same such hypothetical honorific might express deference in one context and solidarity in another. Such honorifics are unattested. Instead of retrieving social relations as values of a variable, honorifics actually classify them. For a recent treatment of the semantics of honorifics and other expressions of 'social deixis,' see Potts (2005).

Bühler (1934: p. 129) noticed that spatial relators such as 'left,' 'right,' 'front,' and 'back' are systematically ambiguous between two types of interpretations, implying different perspectives or, in terms of the more recent literature, 'frames of reference' (Friederici and Levelt, 1990). Thus, *The rock is left of the tree* means that the rock is to the left of the tree as projected from an observer, as the tree has no inherent left, and *The rock is in front of the tree* means it is between the observer and the tree, as the tree has no inherent front. This perspective may be understood as that of the speaker, the addressee, both, or as that of some generic observer. In contrast, when used with respect to people (which do have an inherent left side) or, for instance, buildings (which do have an inherent front), such terms may, but need not, imply the perspective of an observer external to the person/building. In the psychological literature, the former use is known as 'egocentric' and the latter as 'allocentric.' Bühler treated the egocentric, but not the allocentric use as deictic. However, Bohnemeyer (2012a) argues that both types of uses – and in fact all 'projective' (Piaget and Inhelder, 1956) spatial relations, that is, relations depending on reference frames – are indexical, in that the reference frame is always based on one or more axes from some entity or environmental gradient which the interlocutors need to identify in the speech situation. For instance, in order for a 'geocentric' description such as *The rock is upriver of the tree* to be interpretable in a given context, that context must contain a salient river from the course of which the reference frame can be projected.

But while projective spatial relators are thus arguably indexical, and in fact depend on the extralinguistic rather than the linguistic context for their interpretation, they are nevertheless deictoids rather than deictics. This is because they do not actually denote the contextual axes involved in their interpretation. Instead, they denote vectors and spatial regions defined with respect to these axes.

Yet another example of this type of dependence on the extralinguistic context may be found in modal expressions. Consider, for example, the various possible interpretations of '*may*' in *Gusty winds may exist* (an example courtesy of Douglas Adams' unfinished novel *The Salmon of Doubt*) – say, nomological possibility relative to the laws of physics, epistemic possibility relative to the knowledge of the speaker, and meteorological possibility relative to the weather phenomena that have been observed over time in a given location. The third interpretation was presumably the one intended by the New Mexico Department of Transportation, who put this message on a common road sign. This last type is an example of what Kratzer (1981), in a classic theory of the semantics of modal expressions, calls 'circumstantial modality.' According to Kratzer's theory, all of these different uses of *may* share the semantic feature of existential quantification over possible worlds ('There is at least one possible world in which *p* is the case,' where *p* is the proposition inside the scope of the modal operator – in the

example at hand, the proposition 'Gusty winds exist'). Possible worlds are possible contexts or situations relevant to the evaluation of the meaning and truth of propositions. Where they differ is what Kratzer calls the 'modal base,' which spells out the set of possible worlds relevant to the evaluation of a given modal operator – the set of all worlds consistent with the laws of physics; the set of all worlds consistent with the speaker's knowledge; and so on. And the modal base is part of the 'conversational background,' and thus the extralinguistic utterance context. And since the modal base is once again not actually denoted by the modal operator, but merely involved in its interpretation, the modal operator is a deictoid rather than a deictic.

Lastly, even if modals such as *may* are deictoids, this is not to deny the possible existence of truly deictic expressions which in a broad sense may also be considered 'modal.' Prime candidates are moods – grammatical categories that express the distinction between reality and irreality – and also semantically related adverbs such as *actually*. For instance, Kaplan (1989) (see Section 'The Semantics of Deixis') argues that *actually* refers to the utterance world – a possible world that contains the utterance, that is, the 'real world' of the utterance – in a manner straightforward enough to possibly qualify as a genuine deictic.

The Semantics of Deixis

Indexical elements have, arguably, three functions in linguistic utterances. First, they represent the referent in the utterance as a variable. Second, they specify what may be called a *search domain* for the referent in the context (the linguistic context for anaphoric and text-deictic expressions and the extralinguistic context for other deictic expressions). For instance, in the case of the place deictics (*here*, *there*) and demonstratives (*this*, *that*) of English, these search domains are complementary areas of proximity to and the distance from the speaker's and/or addressee's location, at least as far as concrete, literal spatial uses of these expressions are concerned. And third, they direct the addressee's *attention* to the referent (Diessel, 2006).

The search domain can be thought of as a property (formally, a predicate) that possible referents must instantiate. This property is specified as part of the semantics of the indexical. In this respect, it resembles the conventional type meaning of non-indexical expressions, which can be conceived of as a knowledge structure that allows speakers to determine whether and to what extent a given individual or state of affairs (etc.) can be referred to using the term, that is, belongs to its 'extension,' its set of possible referents. This knowledge structure is the 'sense' of the term in Frege's (1892) classic formulation, formalized by the 'intension' in Carnap (1947), a function that assigns to every possible world (see Section 'Deictoids: Honorifics, Projective Spatial Relators, and Modals') the extension of the term in this world. Carnap's intension describes mathematically what Frege's sense does cognitively, allowing us to identify the possible referents of the term in every possible context or 'world.' For example, assuming the sense of *flower* can be paraphrased as 'the reproductive organs of a plant and their colored (nongreen) envelope,' this gives speakers a set of criteria they can rely on to determine whether a given situation contains any possible referents of *flower*.

However, in other respects, the conventional type meaning of indexicals is fundamentally different from the sense or intension of non-indexical expressions. To begin with, the search domain is defined with respect to the utterance, as a property of its context. Hence, the indexicality is irreducible. It is impossible, for example, to paraphrase the meaning of indexicals in strictly non-indexical terms. One might try to paraphrase *I am writing this article* as "The person who writes the sentence that contains the reference to the person at the time of writing the sentence is writing an article that contains the sentence." This in fact eliminates all deictic reference, but it also fails to identify any article or person in particular and, therefore, makes little sense.

Moreover, while an expression such as *flower* can be said to have an extension or set of possible referents that varies with the situations or possible worlds talked *about*, deictic expressions refer to particular individuals, times, or places that vary, not with the situation talked about, but with the utterance itself and specifically its extralinguistic context. For illustration, consider a sentence such as *The flowers were in bloom every time you and I met*. *The flowers* might refer to a different set of flowers with each meeting (perhaps the meetings took place in different years and in different places). But the referents of *you* and *I* clearly cannot vary in this way. Indexicals are generally considered what Kripke (1980) calls 'rigid designators' – expressions associated with a fixed referent across possible worlds.

Kaplan (1989) proposes a two-tier theory of linguistic meaning. One tier is constituted by what Kaplan calls 'content.' Contents map pairs of (extralinguistic) contexts and possible worlds into extensions, thus fulfilling the role of Carnapian intensions. The other tier is a mapping from possible contexts into contents. This is the 'character' of the expression. Non-deictic expressions – especially predicates, which formal semanticists generally assume to represent the meanings of most content words – have contents that do not vary with context, but assign variable extensions to different possible worlds. In this sense, their character is invariable, but their content is variable. Conversely, deictic expressions have different contents in different contexts – for example, the speaker of the utterance in the case of first person pronouns – but then map this same content into all possible worlds. Thus, their character is variable, whereas their content is not. Kaplan's theory is perhaps the most prominent out of an entire line of dual-track approaches to the semantics of indexicals (see Levinson (2004) and Braun (2012) for further discussion).

An interesting prediction Kaplan derives from his theory that has attracted much attention in recent years is that the referent of deictic expressions should not ever be sensitive to the linguistic context in which the expression appears. For example, while the pronoun *I* will be understood to refer to Sally in (1), where the object of *say* is Sally's directly (i.e., verbatim) quoted utterance, it can only refer to the speaker of the matrix utterance in (2), which cannot be Sally (unless she is attempting to conceal her identity or suffering from split-personality disorder):

1. Sally said 'I am right!'
2. Sally said that I was right.

Numerous apparent violations of Kaplan's generalization (often referred to as the 'Monster Prohibition,' after Kaplan's term for hypothetical deictics whose reference is shifted by their linguistic context – 'monsters') have been cited (see e.g., Schlenker (2003) for an overview and discussion). For example, many languages have so-called 'logophoric' pronouns, which are used specifically to represent the quoted speaker in indirect speech (i.e., in place of *I* in contexts such as (2)). Purely logophoric pronouns are a species of anaphoric, rather than deictic, expressions. But some languages allow or require logophoric *uses* of otherwise deictic first and second person pronouns; an example discussed at length in Schlenker's (2003) work is Amharic (Semitic; Ethiopia). Munro et al. (2012) discuss the case of Matsés (Panoan; Peru, and Brazil), where *all* deictics are systematically shifted to the perspective of the quoted speaker in what based on syntactic evidence nevertheless appears to be indirect speech.

The Pragmatics of Deixis

The complexity of deictic search-domain distinctions varies considerably across languages (see Anderson and Keenan, 1985; Levinson, 2004; Weissenborn and Klein, 1982). Where the English demonstratives *this* and *that* make just a two-term distinction of distance relative to the speaker, the Bantu language ChiBemba is said to distinguish proximity to the speaker, proximity to the addressee, proximity to both, distance to both, and relative proximity to the speaker *with respect* to the addressee (other languages have been claimed to make even more fine-grained distinctions, but more recent research has failed to confirm such analyses). Whereas English demonstratives distinguish essentially *relative* distance, such that *this* and *that* may be used *contrastively* within a space close to the speaker, the Yucatec system mentioned earlier operates on an *absolute* proximal–distal distinction (Bohnemeyer, 2012b). Similarly, whereas English tenses merely distinguish past, present, and future with respect to the utterance time, many other languages distinguish, for example, a past of the same day from a 'yesterday' past, a 'tomorrow' future from a more distant future, and so on. ChiBemba distinguishes four such 'degrees of remoteness' in the past and four in the future (Givón, 1972). For an example of a complex pronominal system, see Foley (1997: pp. 112–118) on Tagalog.

The conditions for felicitous deictic reference in space may depend on whether the addressee's attention is already on the referent or not. Some languages even have a formal contrast that matches this distinction. Thus, in Yucatec, the regular demonstrative forms translating 'this' and 'that' are expanded by deictic place adverbs ('this here,' 'that there') when the addressee's attention is not on the referent (Bohnemeyer, 2012b). It has been argued that what is traditionally considered the 'mid-distal' demonstrative of Turkish really has a purely attention-calling function and contrasts with the proximal and distal forms only in this domain, not in terms of distance from the speaker (Özyürek, 1998). Pointing gestures may support attention direction, and in addition serve to narrow down the search domain. Moreover, in case more than one potential referent of the same kind occurs in the search domain selected by the deictic form, pointing gestures may be used to disambiguate the referent. There are, however, search domains that can be referred to without any additional gesture, since they are uniquely identifiable to the addressee. This is true of *here* when used to refer to the speaker's location, and also of *there*, when used to designate the addressee's location, for example, during a phone conversation (Fillmore, 1997). The extent to which the use of spatial deictics requires accompanying gestures seems to vary both with the context of use and with the language-particular term. The attention-directing demonstrative of Turkish cannot be used without a gesture at all. In addition to pointing gestures, iconic and conventional gestures may occur with deictic terms, such as extending the hand, open palm facing up, with presentatives (*Voila!*), or iconically indicating an extension in combination with *this big* or a manner of motion with *like this*. The relationship between such gestures and the linguistic reference act is different from the function of pointing gestures.

Directing the addressee's attention to an object in space is subject to the perceptual accessibility of the object. Many languages exclude the use of certain demonstratives in reference to objects that are not visible, or provide special forms for this purpose. Yucatec has a presentative form for referents of which there is perceptual (e.g., acoustic), but not visual, evidence (Hanks, 1990). In the Wakashan language Kwakwa'la of British Columbia, every noun phrase is marked for whether its referent is visible to the speaker or not (Boas, 1947).

The Genesis of Deixis

It has been argued that pointing gestures are a proto-form of reference, and that accordingly deixis should be the earliest form of verbal reference both in the phylogenetic rise of language in human prehistory (Diessel, 2006; Rolfe, 1989) and ontogenetically in child language acquisition (Clark, 1978; Diessel, 2006). Despite their plausibility, both proposals meet with a major difficulty: the perspectivizing effect of deixis requires highly nontrivial cognitive skills. Thus, there is no conclusive evidence that free-ranging primates use pointing communicatively, and it has been shown that chimpanzees in captivity produce what are commonly considered pointing gestures irrespective of whether they are seen by the addressee or not; in other words, they seem to lack awareness of the interlocutor's mental state of attention (Povinelli et al., 1997). Likewise, it is not the case that the first referring expressions acquired by children are deictics. The first deictic forms acquired by English-learning children are personal pronouns, and the adult-like use of these does not occur before the third year of life. Perspectivizing expressions such as *in front of* are consistently used nondeictically at first, and deictic usage does not come in until the fifth year (cf Piaget and Inhelder, 1956; Tanz, 1980).

Conclusions

Deictic reference plays a particularly important role in language: it serves to glue linguistic representations to the

world, as it were. Deictic expressions directly point the addressee's attention to a referent given in the situation in which the utterance is made, often in combination with gestures. By relativizing the utterance to the particular context in which a particular speaker uses them, they relativize the utterance to this speaker's perspective. The cognitive demands imposed on language processing by this perspectivizing effect are in sharp contrast to the apparent semiotic primitiveness of deictic reference and with its pervasiveness in verbal communication.

See also: Pragmatics, Linguistic; Semiotics.

Bibliography

Anderson, S., Keenan, E., 1985. Deixis. In: Shopen, T. (Ed.), Language Typology and Language Description, vol. III. Cambridge University Press, Cambridge, pp. 259–308.

Barwise, J., Perry, J., 1983. Situations and Attitudes. MIT Press, Cambridge, MA.

Boas, F., 1947. Kwakiutl grammar, with a glossary of the suffixes. Transactions of the American Philosophical Society 37, 203–377.

Bohnemeyer, J., 2012a. Avector space semantics for reference frames in Yucatec. In: Bogal-Allbritten, E. (Ed.), Proceedings of the Sixth Meeting on the Semantics of Under-Represented Languages in the Americas (SULA 6) and SULA-Bar. GLSA Publications, Amherst, pp. 15–34.

Bohnemeyer, J., 2012b. Yucatec demonstratives in interaction: spontaneous vs. elicited data. In: Schalley, A.C. (Ed.), Practical Theories and Empirical Practice. John Benjamins, Amsterdam, New York, pp. 99–128.

Braun, D., 2012. Indexicals. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy, summer 2012 ed. URL: http://plato.stanford.edu/archives/sum2012/entries/indexicals/.

Brown, P., Levinson, S.C., 1987. Politeness: Some Universals in Language Use. Cambridge University Press, Cambridge, UK.

Brown, R., Gilman, A., 1960. The pronouns of power and solidarity. In: Sebeok, T.A. (Ed.), Style in Language. MIT Press, Cambridge, MA, pp. 253–276.

Bühler, K., 1934. Sprachtheorie (Language Theory). G Fischer, Jena.

Carnap, R., 1947. Meaning and Necessity. University of Chicago Press, Chicago.

Clark, E., 1978. From gesture to word. In: Bruner, J.S., Garton, A. (Eds.), Human Growth and Development. Clarendon Press, Oxford, UK, pp. 85–120.

Diessel, H., 2006. Demonstratives, joint attention, and the emergence of grammar. Cognitive Linguistics 17 (4), 463–489.

Fillmore, C.J., 1997[1975]. Lectures on Deixis. CSLI Publications, Stanford, CA. Republication of Fillmore, C.J. Santa Cruz Lectures on Deixis 1971. Indiana University Linguistics Club, Bloomington, IN (1975).

Foley, W.A., 1997. Anthropological Linguistics: An Introduction. Blackwell, Malden, MA.

Frege, G., 1892. Über Sinn und Bedeutung (On sense and reference). Zeitschrift für Philosophie und Philosophische Kritik 100, 25–50.

Friederici, A.D., Levelt, W.J.M., 1990. Spatial reference in weightlessness: perceptual factors and mental representations. Perception and Psychophysics 57, 253–266.

Givón, T., 1972. Studies in ChiBemba and Bantu grammar. Studies in African Linguistics 3 (Suppl.), 1–247.

Hanks, W.F., 1990. Referential Practice: Language and Lived Space among the Mayas. University of Chicago Press, Chicago.

Kaplan, D., 1989. Demonstratives: an essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals. In: Almog, J., Perry, J., Wettstein, H. (Eds.), Themes from Kaplan. Oxford University Press, New York, NY, pp. 481–563.

Kratzer, A., 1981. The notional category of modality. In: Eikmeyer, H.-J., Rieser, H. (Eds.), Words, Worlds, and Contexts. Walter de Gruyter, Berlin, pp. 38–74.

Kripke, S., 1980. Naming and Necessity. Harvard University Press, Cambridge, MA.

Levinson, S.C., 1983. Pragmatics. Cambridge University Press, Cambridge.

Levinson, S.C., 2004. Deixis. In: Horn, L. (Ed.), The Handbook of Pragmatics. Blackwell, Oxford, pp. 97–121.

Lyons, J., 1977. Semantics, vol. 2. Cambridge University Press, Cambridge.

Munro, R., Ludwig, R., Sauerland, U., Fleck, D., 2012. Reported speech in Matses: perspective persistence and evidential narratives. International Journal of American Linguistics 78 (1), 41–75.

Özyürek, A., 1998. An analysis of the basic meaning of Turkish demonstratives in face to face conversational interaction. In: Santi, S., Guaitella, I., Cave, C., Konopczynski, G. (Eds.), Oralité et gestualité: Communication multimodale, interaction (Orality and Gesturality: Multi-Modal Communication, Interaction). L'Harmattan, Paris, pp. 609–619.

Peirce, C.S., 1867. On a new list of categories. Proceedings of the American Academy of Arts and Sciences 7, 287–298.

Piaget, J., Inhelder, B., 1956. The Child's Conception of Space. Routledge, London.

Potts, C., 2005. The Logic of Conventional Implicatures. Oxford University Press, Oxford.

Povinelli, D.J., Reaux, J.E., Bierschwale, D.T., Allain, A.D., Simon, B.B., 1997. Exploitation of pointing as a referential gesture in young children, but not adolescent chimpanzees. Cognitive Development 12, 327–345.

Putnam, H., 1975. The meaning of meaning. In: Gunderson, K. (Ed.), Language, Mind, and Knowledge. University of Minnesota Press, Minneapolis, MN, pp. 131–193.

Rolfe, L.H., 1989. Gesture and deixis. In: Wind, J., Pulleyblank, E.G., De Grolier, E., Bichakjian, M. (Eds.), Studies in Language Origins. John Benjamins, Amsterdam, pp. 25–32.

Schlenker, P., 2003. A plea for monsters. Linguistics and Philosophy 26, 29–120.

Shibatani, M., 1990. The Languages of Japan. Cambridge University Press, Cambridge.

Silverstein, M., 1976. Shifters, linguistic categories and cultural description. In: Basso, K.H., Selby, H.A. (Eds.), Meaning in Anthropology. University of New Mexico Press, Albuquerque, NM, pp. 11–56.

Tanz, C., 1980. Studies in the Acquisition of Deictic Terms. Cambridge University Press, Cambridge, UK.

Weissenborn, J., Klein, W. (Eds.), 1982. Here and There: Cross-Linguistic Studies on Deixis and Demonstration. John Benjamins, Amsterdam.

Wilkins, D.P., Hill, D., 1995. When GO means COME. Cognitive Linguistics 6, 209–259.

Delegation of Power: Agency Theory

Arthur Lupia, University of California at San Diego, La Jolla, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 5, pp. 3375–3377, © 2001, Elsevier Ltd.

Abstract

Delegation is an act where one person or persons relies on other people to take actions on their behalf. It is a critical concept in the study and conduct of politics. Governments use delegation to increase the range of services that they can provide. The most common form of delegation is from lawmakers (e.g., parliaments) to bureaucrats (e.g., cabinet ministries, civil service agencies). Delegation allows government actions to be based on the expertise and abilities of a wide range of people. The downside of delegation, however, is that the people to whom power is delegated can abuse the power they receive. Studies of delegation examine when delegation confers the benefits of expertise to government actors and when it leads to abdication. The theoretical branch of this research is called agency theory. Principal – agent models are mathematical examples of agency theory that provide many studies with logical rigor. Research on delegation is important because it provides people around the world with greater knowledge about how to govern efficiently and effectively. And as nations delegate increasing powers to international organizations such as the European Union and the World Trade Organization, understanding how delegation works is more important than ever.

The Role of Delegation

Delegation occurs when some people ask others to perform tasks on their behalf. For most of us, delegation is an important part of daily life. We delegate to grocers the task of acquiring various foods, we delegate to doctors the task of repairing broken bones, and we delegate to news agencies the task of informing us about local and world events. Delegation is beneficial because we can use it to overcome personal shortcomings. This benefit is important because each of us has limited time, energy, and talents. When the people to whom we delegate devote their time, energy, and talents to our needs, delegation increases the number of tasks that we can accomplish.

In addition to being an important part of daily life, delegation is a central concept in the conduct and the study of politics. Governments great and small use delegation to increase the range of services that they can provide. The governments of large nations, for example, delegate to defense ministries the task of maintaining national security and they delegate to finance ministries the task of managing their nation's economy. Indeed, the modern nation-state could not exist without delegation – for without delegation, lawmakers would be forced to implement and enforce every single law that they pass. With delegation, by contrast, national governments can address a wide range of social issues simultaneously.

In most polities, the most prominent form of delegation is from lawmakers (e.g., parliaments) to government agencies (e.g., economic, security, and public welfare agencies). Such agencies house the persons responsible for enforcing laws and tend to employ the nation's leading experts in the agency's policy area (e.g., economic agencies employ leading economists). In many places, these agencies are known collectively as the bureaucracy. In some places, people who work within these agencies are called bureaucrats or civil servants.

While delegation allows lawmakers to benefit from the expertise and abilities of others, it can also be hazardous. The hazards arise from the fact that delegation entails a transfer of power. For example, every time lawmakers delegate to bureaucrats, they give away a portion of their authority to govern.

The great danger of delegation is that the people to whom power is delegated will abuse the power they receive. This possibility is why some people equate delegation to abdication (e.g., a presidential resignation or a king giving up his throne). If this fear is realized, then those entrusted with the responsibility of governing lose control of what is done in the government's name and anarchy can ensue.

The Study of Delegation

The scholarly study of delegation occurs primarily in the fields of Political Science and Economics. This research examines when delegation confers the benefits of expertise to social decision-makers and when it leads to anarchy. It also helps scholars and practitioners around the world answer one of the most fundamental questions in politics, 'Why do governments do what they do?' For this reason, delegation has been a central focus for social scientists throughout the twentieth century.

Early Studies

In the first half of the twentieth century, scholars sounded a pessimistic tone about the consequences of delegation. Weber's (1946, p. 232) warning is among the best-known:

> Under normal conditions, the power position of a fully developed bureaucracy is always overtowering. The 'political master' finds himself in the position of the 'dilettante' who stands opposite the 'expert,' facing the trained official who stands within the management of administration. This holds whether the 'master' whom the bureaucracy serves is a 'people,' equipped with the weapons of 'legislative initiative,' the 'referendum,' and the right to remove officials, or a parliament, elected on a more aristocratic or more 'democratic' basis and equipped with the right to vote a lack of confidence …

 http://dx.doi.org/10.1016/B978-0-08-097086-8.93029-0

Michels (1915) was also skeptical of delegation. In his iron law of oligarchy, he asserts that democratic organization inevitably becomes subordinated to the interest of their leaders and that the leadership becomes an entrenched oligarchy. Like most of his contemporaries and many of the scholars who followed him, Michels believed in the tendency of people to whom governing power is delegated to use their power against the interests of those who delegated such powers.

Recent Work: The Rise of Agency Theory

Modern scholarship has produced more precise insights about when delegation benefits those who delegate. Many scholars now adopt the language of principal – agent models (i.e., agency theory) to describe the logic of delegation.

The *principal* in principal – agent theories represents someone who delegates. The *agent* represents someone to whom authority is delegated. When a lawmaker delegates authority to an agency, for example, the lawmaker is the *principal* and the agency is the *agent*.

Scholars use principal – agent models to determine when agents do (and do not) act in their principals' interests. *Agency loss* provides a common metric for their distinctions. Agency loss is the difference between the consequences of delegation for the principal and the best possible consequence. Agency loss is zero when the agent takes actions that are entirely consistent with the principal's interests. As the agent's actions diverge from the principal's interests, agency loss increases. When the agent does things that are bad for the principal, agency loss is high.

Research on delegation (see, e.g., Lupia and McCubbins, 1998) shows that agency loss is minimized when two statements are true. The first statement is that the principal and agent *share common interests* (Niskanen, 1971; Romer and Rosenthal, 1978). In other words, the principal and agent desire the same outcomes. The second statement is that the *principal is knowledgeable* about the consequences of the agent's activities. In other words, principals know enough about their agents' actions to determine whether or not these actions serve their interests.

If either of these two statements is false, then agency loss is likely to arise. Agency loss arises when the agent and principal do not have common interests because the agent gains an incentive to act against the principal's interests. If a liberal lawmaker delegates to conservative civil servants, for example, then the agents have less of an incentive to pursue the lawmaker's favored policies than they would if they were also liberals. Agency loss arises when the principal lacks knowledge about an agent's activities because the agent can act against the principal's interests without the principal being aware of the indiscretion. If, for example, lawmakers asks the police to enforce a restriction on public drunkenness, but lack information about agency activities, then the police may not fear retribution if they choose not to enforce the restriction (see Gerber, 2001 for examples).

Agency loss is even more likely if neither of the two statements is true. In other words, when principals have conflicting interests with agents whose activities they cannot know, delegation is likely to be abdication. The threats are greatest in these situations because agents have both an incentive and an opportunity to act against the principal's interests without fear of retribution.

When comparing the main themes of modern and early work on delegation, we see that many early scholars reached their dour conclusions because they presumed that the two statements described above were false. More recent work disputes this presumption, uncovering the many ways in which principals can induce their agents to have common interests and learn about their agents' activities. While the new work does not totally contradict the idea that delegation can become abdication, it is more precise about when this negative outcome occurs.

A central theme in recent literature shows how principals design political institutions to affect the incentives and future actions of their agents. McCubbins et al. (1987), for example, examine how legislative decisions about the structure of political institutions affect the extent of agency loss. Drawing on a wide array of actual administrative procedures, they show that lawmakers can and do set rules for agents that reduce agency loss by ensuring that people who share their policy interests are able to influence what agents do. For example, lawmakers often not only require agents to file extensive reports on their activities, but also invite interested parties from outside the agency to do the same. These interested parties are chosen because they are known to share the lawmakers' policy preferences – including them in the process makes agents more attentive to the lawmakers' policy desires. Many studies also show how lawmakers attempt to reduce agency loss by choosing agents who share their interests (i.e., by making civil service appointments on a strictly political basis: see Epstein and O'Halloran, 1999 and Laffont and Tirole, 1993 for reviews of such findings).

Future Trends

Emerging trends in world politics make understanding delegation more important than ever. Many nations delegate increased responsibility to international organizations such as the European Union and the World Trade Organization. While such actions can produce lower trade barriers and increase social prosperity, they also require national governments to transfer power to others. Many people fear that such transfers will leave nations too weak to combat critical problems. As a result, the future value of research on delegation will be tied to how well it provides political leaders with the theory and evidence they need to make wise choices when delegating to others. Fortunately, today's scholars are responding to this challenge in creative and influential ways.

An emerging trend in delegation research clarifies which ways of delegating are most effective and efficient for political principals. Lupia and McCubbins (1998), for example, use principal – agent models and laboratory experiments to distinguish the types of institutional designs that are most effective in helping principals overcome problems associated with their lack of information about their agents' activities. Their results show when a wide range of political principals (i.e., voters, legislators, and jurors) can (and cannot) learn what they need to know about their agents. Elsewhere, Epstein and O'Halloran, (1999) use substantial theoretical and empirical

research to distinguish when legislatures can (and cannot) benefit from narrowly constraining the agencies to which they delegate (see Huber and Shipan, (2000) for a review of recent studies that adopt this perspective).

Works such as these will likely increase the extent to which real-world politicians learn the most effective and efficient ways to delegate. If they do, then delegation research will continue to give governments large and small the ability to cure broader ranges of social ills – a product of great value in an increasingly complex world.

See also: Authority, Social Theories of; Bureaucracy and Bureaucratization; Organizations, Sociology of; Organizations: Authority and Power; Power in Society; Power; Weber, Max (1864–1920).

Bibliography

Epstein, D., O'Halloran, S., 1999. Delegating Powers: A Transaction Cost Politics Approach to Policy Making Under Separate Powers. Cambridge University Press, New York.

Gerber, E.R., 2001. Stealing the Initiative: How State Government Responds to Direct Democracy. Prentice-Hall, Upper Saddle River, NJ.

Huber, J.D., Shipan, C.R., 2000. The costs of control: Legislators, agencies, and transactions costs. Legislative Studies Quarterly 25, 25–52.

Laffont, J.-J., Tirole, J., 1993. A Theory of Incentives in Procurement and Regulation. MIT Press, Cambridge, MA.

Lupia, A., McCubbins, M.D., 1998. The Democratic Dilemma: Can Citizens Learn What they Need to Know. Cambridge University Press, New York.

McCubbins, M.D., Noll, R.G., Weingast, B.R., 1987. Administrative procedures as instruments of political control. Journal of Law, Economics and Organization 3, 243–279.

Michels, R., 1915. Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracy. Free Press, Glencoe, IL.

Niskanen, W.A., 1971. Bureaucracy and Representative Government. Aldine Atherton, Chicago.

Romer, T., Rosenthal, H., 1978. Political resource allocation, controlled agendas, and the status quo. Public Choice 33, 27–44.

Weber, M., 1946. Economy and society. In: Gerth, H.H., Mills, C.W. (Eds.), From Max Weber: Essays in Sociology. Oxford University Press, New York.

Deliberation and Democracy

Giovanni Boniolo and Giuseppe Schiavone, University of Milano, Milano, Italy; and Istituto Europeo di Oncologia (IEO), Milano, Italy

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.M Bessette, volume 5, pp. 3377–3380, © 2001, Elsevier Ltd.

Abstract

Deliberation is the process of gathering and assessing reasons that ground the preference for some course of action over others. In its political specification, the process is aimed at drafting and issuing collectively binding authoritative decisions. After its initial conceptualization in classical Greece, in the republican Rome, and the fixation of its canons in the Middle Ages, the role of deliberation as a politico-theoretical tool has progressively lost its centrality due to the growing interest modern scholars have developed in democratic representation within large-scale societies. It was not until the 1980s that political theorists restarted discussing the virtues of deliberation in democracy. The idea of democratic deliberation nowadays plays a crucial role in the definition of political legitimacy.

Introduction

Deliberation comes from the Latin verb *deliberare*, meaning to take something off (*de-*) a scale (*libera* or *libra*) after having carefully weighed it. The word is generally taken to refer to the process of gathering and assessing reasons that ground the preference for some course of action over others. In its political specification, the process is aimed at properly drafting and issuing collectively binding authoritative decisions. The history of deliberation in political theory is intertwined with that of democracy. Only recently, however, political theorists openly embraced the so-called deliberative turn in politics and started conceptualizing deliberation as an essential feature of democratic legitimacy in pluralistic societies.

Origins: The Workings of Athenian Democracy

The concept of political deliberation has a long-standing tradition in philosophy. Its original formulation can be dated at least as far back as Aristotle's *Nicomachean Ethics* – which was probably written around 330 BC (Aristotle, 2004a: p. vii). In book III, Aristotle defines deliberation and its scope claiming that it "is concerned with what usually happens in a certain way, where the consequences are unclear, and where things are not definite," adding that – given our somewhat epistemically deficient position – "on important issues, we do not trust our own ability to decide and call in others to help us deliberate" (Aristotle, 2004a: p. 43). The reason why we do not and ought not to deliberate about "eternal things, such as the universe, or the fact that the diagonal is incommensurable with the side" is that none of these things can be effectively brought about. Essentially, we deliberate "about what is in our power" (Aristotle, 2004a: p. 42). A more systematic treatment of the concept of deliberation, specifically in its political signification, features in Aristotle's *Politics* (1998). Chapter 14 in book IV offers an overview of the three elements that Aristotle believes make up all constitutions: these are, roughly, a part dealing with the definition of procedures by means of which a polity would choose their officials, a judiciary part adjudicating controversies among citizens and between citizens and the state, and a deliberative part. The latter "has authority in relation to war and peace, the making and breaking of alliances, and laws; and in relation to death, exile, and the confiscation of property; and in relation to the selection and inspection of officials" (Aristotle, 1998: p. 125). This deliberative function can be interpreted, according to Aristotle, as involving all citizens all the time, only some citizens all the time, or some citizens sometimes and all citizens some other times. The first interpretation, all citizens involved all the time, is the intuitive (Athenian) idea of democratic deliberation. Still, there are several ways of democratically conceiving deliberation with every citizen's participation. Aristotle lists four such ways: (1) citizens might deliberate in turn, rather than all together; (2) citizens might deliberate all together only about the choice of officials, legislation, war and peace, and inspections, leaving any other deliberation to a subset of the citizenry either elected or drawn by lot; (3) the entire citizenry could deliberate about offices and inspections, about war and alliance, leaving a restricted number of issues demanding a high degree of knowledgeability to be dealt with by officials; and (4) every citizen could participate in every decision and appointed officials would only have to prepare issues for decision.

Just like Aristotle's ethics has sometimes been said to be "mere Athenian common sense dressed in philosophical garb" (Aristotle, 2004a: p. viii), so his definition of democratic deliberation is clearly inspired by, if not openly built upon, the configuration of Athenian democratic institutions. The democratic architecture of Athens (see Hansen, 1991; Ober, 1989, 1996) was progressively brought about by a series of reforms, the first of which followed Solon's appointment as *Archon* with extraordinary powers in 594 BC. The *Archons* were rulers chosen among the *eupatridai* (the nobles), and each one was endowed with some powers for the government of the city-state. Solon triggered a series of reforms that progressively turned the Athenian *ekklesia* (the general assembly of citizens), governed by the *boulé* (a council composed of selected citizens), into a democratic institution. Once Cleisthenes, almost 90 years later, managed to assume the leadership of Athens, new reforms further democratized the life of the city-state. *Isegoria*, the right for every citizen to intervene and express his or her own position, became established as one of the most

fundamental rules governing the *ekklesia*. The deliberative practice was based on the competition among different political proposals and arguments publicly presented in their support. Athenian democratic participation could therefore be seen as an attempt (probably historically the first) at including good reasons in the process of public decision-making, where citizens (or at least a subset of them) had an active participative role. Cleisthenes also introduced a draw mechanism for the access both to the *boulé* and to other public offices. In this way, the risk of aristocratic lobby formation was almost ruled out. Moreover, the *dokimasia* was established: access to public offices, regardless of whether one was elected or drawn by lot for that office, was conditional to a thorough examination. This examination was held before the *boulé* or the court, as Aristotle recalled in *The Athenian Constitution* (Aristotle, 1984: p. XLV–XLIX). The candidate was asked whether he was an Athenian citizen, whether he was a devotee of *Apollo Patroos* (the protector of the family) and of *Zeus Herkeios* (the protector of the home), whether he owned a family grave, whether he was good to his parents, whether he was paying taxes, and whether he had previous military experience. Witnesses were asked to support the candidate's claims. Finally, everyone was asked whether they had any objection. In case there was no consensus about the acceptance, a debate would start to either accept or reject the candidate.

The apogee of Athenian democracy was, however, reached just before the death of Pericles (429 BC) when the *misthoforia* was introduced. Before the introduction of this reimbursement, given to anyone who held an institutional office, citizens of the lowest social strata could not afford to comply with their political duties: if they were elected or randomly appointed, they would have suffered too high losses in terms of time and work to sustain themselves. The formal right of access to a public office or of participating in the general assembly was not enough. The *misthoforia* turned this formal right into a substantial right to participation held by all Athenian citizens (slaves, women, children, foreigners, and metics were not considered citizens and therefore could not enjoy these rights).

Medieval Deliberation

If we could indicate the classical Athens as the place and the time where deliberative democracy had its birth, it was in the Middle Ages that the tools and procedures for a good deliberative process were fixated. They were essentially based on the Aristotelian rendering of the art of rhetoric. It is to be recalled that, attaching no negative connotation to the term 'rhetoric,' Aristotle had provided a detailed taxonomy of discourses (Aristotle, 2004b: p. I.3) according to their speaker, their content, and their audience. An audience could, in turn, be conceived either as judge of or as passive spectator to a speech, and in case it acted as a judge, it could do so with respect either to future or to past affairs. Thus, following this categorization, whenever an audience is a collective supposed to judge of future things, i.e., of what course of action we, as a polity, want to pursue, then the speaker is engaging in deliberative discourse (which is contrasted by Aristotle to forensic – the speaker engaging with some audience judging past affairs – and to epideictic – the audience being a mere spectator to the orator's abilities – discourses).

Coming back to the Middle Ages, it was in the newborn universities that the rhetorical canon of deliberation as theorized by Aristotle became crystallized and routinely taught. With the birth of the European universities, around the twelfth century, the Aristotelian rational method to tackle problems worked as a common ground for all teachings in the liberal arts, particularly in the so-called *Trivium* (where logic, the art of good reasoning; rhetoric, the art of good persuading; and grammar, the art of good writing, were taught). This method was based on three main pillars: (1) the clarity of the statement of a problem, (2) the identification of an appropriate solution, and (3) its rational justification (Boniolo, 2012: p. 89). These pillars served (and may still be used) as cornerstones for the construction of proper rhetorical tools for a particular kind of deliberation central to the didactic system of medieval universities – the *disputatio*. Disputes were typically a fundamental part of academic life for medieval scholars. They were chaired by a *magister*, who would introduce a topic asking a specific question (*quaestio*) about it. The *quaestio* would generally be introduced by the formula 'whether or not' (in Latin *utrum*). Two students would then challenge each others' arguments in favor of one answer to the question. Eventually, the *magister* would stop the dispute and propose his view, hence reaching a solution, or *determinatio* (Lawn, 1993).

Although the use of the arts of the medieval *Trivium* was restrained mostly to theoretical issues and hosted by academic institutions, the structure of *disputationes* can be seen, in light of Aristotle's taxonomy sketched above, as a blueprint for proper argumentation in a public arena and therefore deployed as a standard for public deliberation.

At the end of the 1950s, the works of Perelman and Olbrechts-Tyteca (1958) and Toulmin (1958) paved the way for the development of a series of attempts at what has been called informal logic. These authors more specifically kindled a renewed interest in the Aristotelian conception of rhetoric as the art of correct reasoning intended to persuade an audience. This resulted, years later, in the development of a proper argumentative framework for public deliberation.

Republicanism and Democracy

The idea of deliberation as constitutive to proper democratic participation has been substantially underexplored in political theory, essentially until the late twentieth century. Neither the certification or moral competence nor the construction of some institution hosting argumentatively proper debates played any central role in political theorization. Nevertheless, theories of democracy have evolved over time so as to foster the sustainability of political institutions and to better account for contestation. Republicanism is one such attempt at conceiving democracy. This theory, broadly understood, has been revived by a number of contemporary scholars (the so-called "Cambridge School" of history, e.g., Skinner, 1990; Pettit, 1997) who presented and partly reconceptualized the political theories developed by Aristotle and Cicero, in the classical Greece and the republican Rome, respectively, and,

most importantly, those proposed by Machiavelli in the early-sixteenth-century Florentine Republic. These scholars, and Philip Pettit more in particular, developed a definition of liberty that can account for two major facts of political life: on the one hand that a government creates the space for freedom and on the other that the existence of some agency authorized to make use of physical violence to enforce collectively binding rules threatens that very freedom. Accordingly, liberty is not so much conceived of as the absence of interference (as just said one of the assumptions is that the government secures freedom by constraining people's liberty, thus interfering with their lives), but rather as the absence of domination, i.e., the absence of arbitrary interference. This definition is characteristic of the rhetoric of liberty in the Roman republic, where a *liber*, a free person, was more than a mere *servus sine domino* (a slave without a master). The fate of a resourceless slave, set free by his or her master, still did not depend on him or her, but rather on the mercy of others. His or her liberty was constantly threatened by the possibility of interference by others. A *liber*, instead, was necessarily a *civis*, a citizen, whose liberty the state ought to protect against arbitrary interference (Wirszubski, 1968).

Pettit, along with other authors in the Cambridge School of historiography, understands Machiavelli – contrary to the popular reception of his works as providing a toolkit for the enablement of tyranny – as a republican scholar, whose major contribution to political theory lies in the emphasis he puts on the use of proper mechanisms for elite accountability to counteract the risks associated with the unequal distribution of wealth, which in turn could be a major cause of arbitrary interference. However, according to critiques to this account of republicanism – such as McCormick's (2011) – one of the assumptions virtually all these new republicans make is, quite counterintuitively, anti-Machiavellian, meaning that it shares more with the kind of attitude of Guicciardini and other proponents of *governo stretto* (an elitist form of government famously deployed in Venice at the time) than with Machiavelli's take on the inclinations of the *popolo* (the uneducated low social strata) and the *grandi* (the wealthy) laid out in his *Discourses on Titus Livy's First Ten Books*. These two opposing attitudes toward the people can be roughly summarized as confidence as opposed to skepticism, concerning the way in which the masses may participate in and contest public decisions. McCormick provides textual evidence supporting the claim that Machiavelli believed that the *popolo* was able (and actually more inclined to do so with respect to the *grandi*) to participate properly, without menacing the polity. The scarce emphasis put by Pettit, for instance, on the way toward the institutionalization of popular contestation in democracy, seems to imply that he takes Machiavelli to subscribe to the charge that the people have passions and inclinations that can hardly be controlled and their participation ought therefore to be constrained, in order to avoid arbitrary interference, i.e., domination, on the part of the majority on some minority. This is, McCormick argues, a standard elitist assumption instantiated, for example, by Guicciardini's words in his *Considerations*: "The people are too ignorant to discuss important matters freely and ably in assembly" (Guicciardini, 2002: p. I.2) and further developed by modern liberal theorists of representative democracy.

Regardless of the accuracy of McCormick's reconstruction, it is relevant to underscore the underlying argument that elitist theorists of politics may leverage on to make their case for less participatory institutions; participation by the uneducated masses might subject public decision-making to all sorts of irrationality, and this is undesirable because it might affect the output of policymaking. Therefore, it is better to simply restrain the number of people who actively take part in public decision-making.

Liberalism and Modern Representative Democracy

The growth in size of nation-states and the extension of political rights to new subsets of the populaces made representation increasingly the default option for institutional deliberative bodies. Modern political theorists therefore focused almost exclusively on the development of constitutions that would grant deliberative powers to people's representative rather than to people themselves.

As anticipated, the main argument supporting this solution is one fairly elitist in kind: in 1787, in *The Federalist Papers* (Hamilton et al., 1999) – a landmark in the history of representative democracy – Hamilton would alert to the risks linked to deliberation by the people, claiming that when in ancient republics the whole polity assembled to take collectively binding decisions, "a single orator, or an artful statesman, was generally seen to rule with as complete a sway as if a scepter had been placed in his single hand" (Hamilton et al., 1999: p. 328) and this was so because the sheer size of those assemblies made them liable to ignorance and sophistry. This detachment of political decision-making from what we might take as the people's will was equally well embodied in Burke's Speech to the Electors of Bristol from 1774 (Burke, 1949) in which he forcefully makes a case against the responsiveness of political delegates and representatives to the desires and preferences of their constituency. Reason and good judgment demanded that a legislator acted on "his unbiased opinion, his mature judgment, [and] his enlightened conscience" (Burke, 1949: p. 115) rather than on the irrational, uninformed, and potentially inconsistent intuitions of the people. The pursuit of the good of a nation, according to this kind of conception, is a rational endeavor and the more the people are directly involved, the less – allegedly – the outcome will be rational.

A somewhat analogous line of reasoning led to theories of democracy in which direct participation plays a very little role for the legitimacy of deliberative moments of a polity. Joseph Schumpeter famously crafted a definition of democracy that takes the deficiencies of the public (in particular the fact that people cannot possibly hold a rational opinion about every relevant issue) to affect so harshly the capability of an individual to 'directly' rule oneself that the only conceivable way out of this predicament is through open competition for political leadership. He therefore laid out the suggestion that "the democratic method is that institutional arrangement for arriving at political decisions in which individuals acquire the

power to decide by means of a competitive struggle for the people's vote" (Schumpeter, 2004: p. 269).

On the other hand, people who studied how to design institutions in the twentieth century did so from a thoroughly formal analytical angle, trying to envisage mechanisms that would allow for aggregating individual preferences so as to have a collective outcome consistent with those preferences, basically asking to what extent and whether at all some social welfare function, i.e., a function that aggregates individual preferences into societal outcomes, can meet a small set of minimal requirements of acceptability. Kenneth Arrow provided, in the form of an impossibility theorem, reasons why systems for the aggregation of ranked preferences are unable to satisfy a version of intuitive requirements of fairness (Arrow, 1950).

All of these attempts straightforwardly ignored the role that reasoning together about individual preferences might play in forming those preferences and then in making collective decisions.

Late-Twentieth-Century Rediscovery: The Deliberative Turn

For the most part of the twentieth century, political theorists have been occupied with the definition of some proper way of aggregating individual or group preferences so as to have institutions that would reflect the people's will. This favored the choice of bargaining over deliberation as a means for collective decision-making and encouraged the rather formal kind of analysis developed by rational choice theorists.

In the early 1980s, however, a number of theorists of democracy promoted a rediscovery (and a partial redefinition) of the concept of deliberation – trying to voice widespread dissatisfaction with the workings of contemporary representative democracies – contributing to what has been called the 'deliberative' turn in political philosophy. These theorists argued that democratic legitimacy demands inclusive and discursively appropriate participation. As John Dryzek put it, "[o]utcomes are legitimate to the extent they receive reflective assent through participation in authentic deliberation by all those subject to the decision in question" (Dryzek, 2001: p. 651). Even though a host of interpretations of this requirement have been articulated, most of the attempts share at least two main principles: (1) collectively binding decisions ought to be taken by all those affected by the decisions and (2) the process of collective decision-making ought to be the result of a discussion conforming to the standards of public reason among free and equal citizens (Elster, 1998: p. 8). Proper communication among equals is said to be capable of transforming one's uninformed and prejudicial preferences into well-formed preferences, i.e., preferences that fall within some logically consistent architecture of preferences and whose terms are well-defined and shared. This would in turn result in some consensus about decisions that bind the polity. Within this framework, scholars were led to hypothesize and argue what features of a speech situation would bolster the transformational potential of deliberation. German philosopher Jürgen Habermas (1990) compiled a list of rules that, when satisfied, contribute to the construction of what he named an ideal speech situation, one in which every competent speaker is allowed to present arguments and challenge others' assertions.

Principles of Deliberative Democracy

Gutmann and Thompson (1996, 2004) have provided one of the most compelling articulations of the tenets of deliberative democracy. The fact of irreducible moral disagreement, the moral disagreement residual to most collectively binding decisions, in societies truly pluralistic about values poses a tough challenge to democratic institutions. Deliberation, according to Gutmann and Thompson, helps facing at least four main issues that arise in this context. The first one is the paucity of resources that, as a society, we can allocate to specific needs. Someone will always be disappointed with a decision allocating scarce resources and somehow rightly so. He or she will have had less than others did. It becomes therefore essential that this kind of decision is extensively discussed and that everyone has their chance to challenge the arguments supporting it, thus being able to recognize its rationale. Secondly, political actors have very limited generosity. Democratic deliberation helps nourishing people's public-spiritedness, inviting everyone to put themselves in one another's shoes, hence fostering the recognition of common interests. Thirdly, public deliberation about morally sensitive issues allows a significant epistemic result: deliberation helps unveiling the real values behind one's preferences, further allowing for the recognition of possible inconsistencies in one's architecture of values. Finally, deliberating is a way to learn whether and to what extent as a polity we want to allow bargaining special and group interests and when, instead, reasons ought to prevail.

That being said, deliberative democracy, as influentially theorized by Gutmann and Thompson, is both a procedural and a substantive political theory. The values deliberation is supposed to foster do not constrain only how democratic societies should go about discussing issues (the procedural part) but also what can (and cannot) possibly be the result of a democratic deliberation (the substantive bit). There are three principles the authors outline that meant to explain, respectively, what reasons democratic deliberation requires, how these reasons ought to be presented, and who is supposed to provide those reasons.

First, deliberative democrats ought to value reciprocity: In its common understanding, reciprocity is a specific form of proportionality between some good I give away and some other good I receive in exchange for that. In the case of deliberation, one's proportionate concession is the formulation of their claims in terms that are, in principle, acceptable by others. Essentially, anyone properly motivated to find fair terms of cooperation ought to share a disposition to accept other people's arguments as long as the others are mutually bound by the same disposition.

Secondly, the principle of publicity requires that reasons for collectively binding decisions be provided publicly. Whether this entails that reasons ought to always be actually provided in public or that they must always be hypothetically or potentially made publicly available? The demands that the former

interpretation puts on political life are much more burdensome than those implied by the latter. The idea that just any decision must be taken publicly has been widely criticized, mostly with respect to the opportunity of disclosing the process that takes to controversial decisions for which accessible reasons might endanger the very good that the decision is likely to yield (e.g., strategic military decisions).

Thirdly, the principle of accountability would prescribe, ideally, that everyone presenting their claims publicly is responsible for those claims to everyone else. This is what universal accountability would demand. However, this solution looks impractical to say the least. What a realistic understanding of the value of accountability seems to require is that representatives are held accountable for at least some of their decisions (i.e., at least those decisions deemed appropriate for public scrutiny) to at least everyone who is affected by those decisions.

These principles have definitely reshaped the debate in political philosophy. It is worth noting, though, that their theoretical edifice is historically rooted, at least implicitly, in the values and principles embodied in medieval disputes.

Kinds of Deliberative Institutions

A number of institutional design proposals have been produced over the years starting from what we might call a deliberative framework informed by the principles to values just summarized. Most of these proposals have been empirically tested as a way to probe the assumptions that preferences do change thanks to the deliberative process and that the process is itself sustainable (or that it improves upon the moral sustainability of current institutions).

James Fishkin (2011) developed a method, he named 'deliberative polling,' consisting in a solution halfway between an opinion poll and an actual deliberation. Its features are very similar to those of a consensus conference, public engagement initiatives meant to include lay opinions about policy issues in actual policymaking, in which a demographically representative citizens' panel is selected randomly from a group of citizens invited to apply. Deliberative polling draws participants randomly in order to build a representative sample of the population. These people are asked a question about a policy issue. This first question is used as a baseline poll that will serve to measure the impact of the deliberation on people's preferences. Participants will then have to engage in deliberative practices about some policy issue. Briefing materials are carefully drafted and handed to the participants (these materials are also made publicly available). Parts of the events, which extend usually over a weekend, are or can be broadcasted. The deliberation happens in small groups to allow for more reflective and informed preference formation. At the end of the process, people are asked again the same question in order to gauge the impact of the process onto people's preferences. This outcome is supposed to mirror the preference people would actually entertain were they given the chance to be deliberatively informed. This approach has been criticized on the grounds that it can be characterized as a disguised form of aggregative democracy, given that preferences are simply counted at the end of the process and no genuine collectively deliberative outcome results from the polling.

Citizens' juries are minipublics that make no attempt at gathering a sample representative of the population. In fact, participants are drawn by lot in groups of 10–20 people. During meetings that develop over a couple of days, they are exposed to information presented by experts. Questioning of the experts is allowed, and participants might also call upon new witnesses to the panel. Meetings are usually open to everyone willing to attend. Deliberation within the group is directed toward the production of a recommending output, generally in the form of a report to be submitted to some relevant authority, which is typically also the sponsoring institution for these encounters and is bound to reply to the submission by either enforcing its recommendations or saying why it decided not to (Smith and Wales, 2000: p. 55).

The Account of Legitimacy and the Difficulties of Deliberating in the Real World

The account of legitimacy underlying deliberative democracy encounters a series of difficulties. Two of them, in particular, seem to pose a serious challenge to the practice of deliberation in the real world: on the one hand, in fact, deliberation requires that legitimate decisions be taken by the highest number of people that one can possibly hope for, but the more people participate in a debate, the less the likelihood that their participation will genuinely live up to deliberative standards (it will be very unlikely that people actually have time to give and heed reasons). Boniolo suggested a rough calculation concerning the hypothetical feasibility of a genuinely deliberative and democratic Athenian *ekklesia*. Approximately 6000 citizens used to participate in the assembly and assuming "each […] had spoken for, say, 10 minutes, then a full round of opinions would have taken 60 000 minutes, or 1000 hours, which are equal to about 42 consecutive days. Suppose they had slept for 12 hours during the night, then 84 days would have been necessary" (Boniolo, 2012: p. 80). Clearly then, the prospect of having a literal implementation of deliberation would be ill-suited to democratic societies.

On the other hand, even assuming that people will be able to participate properly (i.e., will have the necessary amount of time and skills), it is hard to see why they would do so in practice, meaning there are very few reasons that would succeed in motivating widespread participation. Parkinson calls these two problems the scale and the motivation problems of deliberative democracy (Parkinson, 2006).

These issues seem to affect quite radically the tenability of deliberative democracy as a practicable ideal. Some solutions have been suggested in order to redress these difficulties and ameliorate the ideal of deliberative democracy. In particular, there are competing ways of conceiving deliberation in democracy, and they all can be said to lie along a continuum that goes from micro conceptions to macro conceptions of deliberative democratic institutions. At the former end of the continuum are ways of conceiving deliberative democracy that rely upon microscale deliberations among a very limited number of participants abiding strictly by the procedures of public reasons. At the other end of the continuum, deliberation is conceived as a multithread conversation, involving different people at different times, with less formal rules for participation. With this distinction in mind, Parkinson (2006)

elaborated a new account of legitimacy for deliberative institutions that relies on the further distinction between trusteeship and delegation. Democracy, according to his (plausible) reconstruction cannot do without representation. Representation, however, can be defined either as trusteeship, i.e., the representative can make decisions on behalf of the represented as he/she sees fit, or as delegation, i.e., the representative is bound to be responsive to the instructions of his/her constituency. The preference for one solution to the principal–agent problem over the other, in Parkinsons' account, is thoroughly context-dependent. More specifically, representatives can and ought to act as trustee when they take part into microscale deliberations in which their readiness to be persuaded is essential, whereas they could and should go back to the macro-sphere and be accountable to their constituency.

Latest Developments

The Wisdom of Crowds

A series of publications contributed to a fairly novel approach to knowledge management that goes under the label of 'the wisdom of crowds' (the name comes from James Surowiecki's (2005) book). Surowiecki has tried to leverage empirical evidence to undermine the assumption that a few well-informed people, a so-called expert think tank, would always come up with a better solution than that of some uninformed small group of individuals. More specifically, he argued that disorganized crowds, when their members are allowed to freely participate in some task (given certain additional conditions), are possibly smarter than the smartest person in that very crowd at that very task. The development of infrastructures (particularly information and communication technologies) for the lean and cheap management and replication of knowledge led Weinberger, among others, to put forth a further claim, which can be summarized in his own words concerning the networking of knowledge: "As knowledge becomes networked, the smartest person in the room isn't the person standing at the front lecturing us, and isn't the collective wisdom of those in the room. The smartest person in the room is the room itself" (Weinberger, 2012: loc.132).

These developments led to the revival of some fairly anti-elitist claims against the influence experts have on policy-making, and it also propped the elaboration of some radical projects such as Robin Hanson's 'futarchy,' proposing to deploy prediction markets on policies, i.e., markets of futures in which people bet on the prospective success of policies and the ones deemed more likely to succeed are subsequently implemented (Hanson, 2013).

None of these attempts look deliberative in the sense that citizens participate in policymaking by giving and taking proper arguments, nor the legitimacy of public decisions seems to depend, within this framework, on that process of reason giving and heeding. In fact, these proposals amount to a kind of epistemic proceduralism whereby power is dispersed, or devolved, for the main reason that devolution of decision-making powers will supposedly allow for the best epistemic outcome to emerge, i.e., it would be the best procedure to know what will make a society prosper.

The Middle Ages Reconsidered

As said, the role of the *Trivium* was regarded as pivotal for any course of studies in the liberal arts during the Middle Ages. A profound understanding of rhetoric, logic, and grammar (the three disciplines *Trivium* was articulated in) was seen as a necessary requirement for any scholarly endeavor and therefore taught during the first years of every bachelor degree. The rhetorical canon described above and deployed during disputes in medieval universities demanded that deliberations be conducted following a specific procedure (in particular, the clear statement of a problem had to precede the discussion of one's proposed solution) and that arguments and objections be compliant with standards of public reason, meaning mostly that arguments had to prove logically consistent, i.e., not fallacious or invalid.

Recently, Boniolo and Di Fiore (2010) and Boniolo (2012) developed an approach to deliberation that tries to embed similar epistemic and rhetorical requirements in the decision-making process concerning controversial issues, especially dealing with ethical questions arising in biomedical research. Devising a detailed workflow to be followed for the discussion of ethical issues, these authors propose that every deliberator acquires (or shows he/she masters already) both the relevant knowledge and the appropriate rhetorical skills.

E-Democracy

E-democracy designates the set of institutional practices enabled by new information and communication technologies (ICTs). These practices have been suggested as a potential remedy to the scarce civic engagement of citizens in today's representative democracies. Being able to participate online could lower barriers to participation in a number of politically relevant activities, thus making people more likely to engage in those activities. One of the ways in which ICTs have already had an impact on the shape and modes of political activities is campaign fund-raising. After Howard Dean candidacy to the Democratic Party presidential primaries, which proved particularly successful in recruiting resources from the generally disengaged (young citizens particularly) (Hindman, 2008), political candidates have been relying regularly on online tools such as Meetups and social networks for campaigning purposes. Watchdogging practices have also been incredibly facilitated by the progressive digitalization of public documents, spurring in turn demands for more and more open data.

Attempts at democratic deliberation that rely on the Internet for their implementation, however, are still at a very early stage. Interestingly, the potential for ubiquitous availability of the Web and the increasing actual access to some decent bandwidth across different social strata accompanied by the proper skillset for the use of online tools could remedy at least part of the abovementioned scale problem, lowering costs for mass online deliberations.

See also: Democracy, History of; Democracy: Normative Theory; Democratic Theory.

Bibliography

Aristotle, 1984. Athenian Constitution. Penguin Classics, London.

Aristotle, 1998. In: Reeve, C.D.C. (Ed. and Trans.), Politics. Hackett Publishing Company, Indianapolis/Cambridge.

Aristotle, 2004a. In: Crisp, R. (Ed. and Trans.), Nicomachean Ethics. Cambridge University Press, Cambridge.

Aristotle, 2004b. Rhetoric. Dover Publications, New York.

Arrow, K., 1950. A difficulty with the concept of social welfare. Journal of Political Economy 58 (4), 328–346.

Boniolo, G., Di Fiore, P.P., 2010. Deliberative ethics in a biomedical institution: an example of integration between science and ethics. Journal of Medical Ethics 36 (7), 409–414.

Boniolo, G., 2012. The Art of Deliberating. Springer-Verlag, Berlin, Heidelberg.

Burke, E., 1949. In: Hoffman, R., Levack, P. (Eds.), Burkes Politics: Selected Writings and Speeches of Edmund Burke on Reform, Revolution, and War. Knopf, New York.

Dryzek, J.S., 2001. Legitimacy and economy in deliberative democracy. Political Theory 29 (5), 651–669.

Elster, J., 1998. Deliberative Democracy. Cambridge University Press, Cambridge.

Fishkin, J.S., 2011. When the People Speak: Deliberative Democracy and Public Consultation. Oxford University Press, Oxford.

Guicciardini, F., 2002. Considerations of the discourses of Niccolò Machiavelli. In: Atkinson, James B., Sices, David (Eds. and Trans.), The Sweetness of Power: Machiavelli's Discourses and Guicciardini's Considerations. Northern Illinois University Press, DeKalb, IL, pp. 381–438.

Gutmann, A., Thompson, D., 1996. Democracy and Disagreement: Why Moral Conflict Cannot be Avoided in Politics, and What Should be Done About it. Belknap Press, Cambridge, MA.

Gutmann, A., Thompson, D., 2004. Why Deliberative Democracy. Princeton University Press, Princeton.

Habermas, J., 1990. Discourse ethics: notes on a program of philosophical justification. In: Moral Consciousness and Communicative Action (C. Lenhart, S.W. Nicholson, Trans.). MIT Press, Cambridge, pp. 43–115.

Hamilton, A., Madison, J., Jay, J., 1999. In: Rossiter, C.R. (Ed.), The Federalist Papers. New American Library, New York.

Hansen, M.H., 1991. The Athenian Democracy in the Age of Demosthenes: Structure, Principles, and Ideology. University of Oklahoma Press, New York.

Hanson, R., 2013. Shall we vote on values, but bet on beliefs? Journal of Political Philosophy 21 (2), 151–178.

Hindman, M., 2008. The Myth of Digital Democracy. Princeton University Press, Princeton.

Lawn, B., 1993. The Rise and Decline of the Scholastic "Quaestio Disputata". E.J. Brill, Leiden.

Machiavelli, N., 1997. Discorsi sopra la prima deca di Tito Livio. In: Vivanti, Corrado (Ed.), Opere I: I Primi Scritti Politici. Einaudi-Gallimard, Torino, pp. 193–525.

McCormick, J., 2011. Machiavellian Democracy. Cambridge University Press, Cambridge.

Ober, J., 1989. Mass and Elite in Democratic Athens. Princeton University Press, Princeton.

Ober, J., 1996. The Athenian Revolution. Princeton University Press, Princeton.

Parkinson, J., 2006. Deliberating in the Real World. Problems of Legitimacy in Deliberative Democracy. Oxford University Press, Oxford.

Perelman, C., Olbrechts-Tyteca, L., 1958. The New Rhetoric: A Treatise on Argumentation. University of Notre Dame Press, Notre Dame.

Pettit, P., 1997. Republicanism. A Theory of Freedom and Government. Oxford University Press, Oxford.

Schumpeter, J., 2004. Capitalism, Socialism and Democracy. Routledge, London.

Skinner, Q., 1990. Machiavelli's Discorsi and the pre-humanist origins of republican ideas. In: Bock, G., Skinner, Q., Viroli, M. (Eds.), Machiavelli and Republicanism. Cambridge University Press, Cambridge, pp. 121–141.

Smith, G., Wales, C., 2000. Citizens' juries and deliberative democracy. Political Studies 48, 51–65.

Surowiecki, J., 2005. The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Anchor Books, New York.

Toulmin, S., 1958. The Uses of Argument. Cambridge University Press, Cambridge.

Weinberger, D., 2012. Too Big to Know. Rethinking Knowledge Now That the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room Is the Room, Kindle ed. Basic Books, New York.

Wirszubski, C., 1968. Libertas as a Political Idea at Rome. Oxford University Press, Oxford.

Delinquency, Sociology of

Gary F Jensen, Vanderbilt University, Nashville, TN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the basic features of the sociology of delinquency, including its foundation, evolution, and contemporary forms. Early perspectives are outlined and the contributions of scholars identified. A major methodological innovation, the self-report study, set the stage for scientific tests of theories, and attention shifted from the study of gangs to the study of delinquent behavior. Social control and social learning theories grew in popularity using the method. Recent research has returned to the study of gangs using the method. New sociological theories have been proposed and can be categorized in terms of their links to early perspectives. The article ends with a brief outline of recommendations for the future study of the sociology of delinquency.

Definition of the Field

The contemporary sociology of delinquency is a specialty in the field of criminology, seeking to understand the offense behavior and problems of youth through the use of methods, concepts, and theories characterizing the academic discipline of sociology. Delinquency itself is a legal concept which encompasses a variety of offenses or activities of nonadults. It has varied over time and jurisdiction, originally including a wide range of behaviors that would not be offenses for adults (i.e., status offenses) and vague designations of repeated activity (i.e., incorrigible). The category has narrowed over the years such that the contemporary application by the juvenile or family court is most often limited to offenses that would be criminal if the offender was an adult.

The Link to Criminology

The contemporary field of criminology involves the application of the 'scientific method' to the study of variation in criminal law, the causes of crime, and reactions to crime (Akers, 2000; Jensen and Rojek, 2009). This combination of method and substance evolved over time with a scientific methodology in place by the late nineteenth century, and acceptance of the broad, tripartite, substantive focus established by the middle of the twentieth century. Sociologists played a major role in defining and developing the field of criminology which emerged as an academic discipline housed in sociology programs. However, in the last half of the twentieth century, criminology emerged as a distinct professional field with a broad, interdisciplinary focus and a shared commitment to generating knowledge through systematic research.

Subject Matter

The sociology of delinquency has followed the tripartite path of criminology by seeking to understand variations in laws defining delinquency and other features of juvenile justice, searching for explanations of offense behavior among youth and designing and examining attempts to correct or prevent delinquent behavior. In their attempt to answer questions in these areas sociologists use scientific methods to identify patterns involving (1) relationships among people; (2) values, norms, and beliefs; (3) features of social institutions (family, education, government, religion); (4) features of environments that affect the opportunity to commit offenses; and (5) features of environments that affect opportunities for nondelinquent, conventional lives.

There are several reasons why sociology was the major scholarly home for criminology early in its development. First, delinquency is predominately a form of 'group' behavior. When focusing on youth, delinquent activities occur most often in groups. Some of these groups can be identified as gangs, but most are just groups of friends or acquaintances. Clifford Shaw and Henry McKay (1942) developed a theory of delinquency beginning in the 1920s which they intended to apply to delinquent activities embodied in groups and social organization. Second, the legal concept of juvenile delinquency itself was a social invention which emerged in cities and states experiencing rapid growth and diversification with Chicago leading the way. In Cook County, Illinois, a new type of organization, the juvenile court, was created to be a solution to the growing problem of wayward youth. The new juvenile justice system developed as an alternative to the adult criminal justice system and provided new ways to process and treat juveniles.

Founding Scholars

Three scholars provided much of the early foundation for the sociology of delinquency, Clifford Shaw, Henry McKay, and Earl Sutherland, all associated with the University of Chicago. The most consequential research for the development of sociological criminology and a sociology of delinquency was Clifford Shaw's study of Chicago (1929) which described and sought to explain the distribution of a variety of social problems in the city of Chicago. The observation that certain areas tended to keep high rates despite successive changes in the ethnic groups residing in them suggested that those problems were generated by the social conditions experienced by these groups rather than by any genetic or biological predispositions.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45021-X

In addition to the analysis of the distribution of problems among social areas, the scholars and researchers constituting the Chicago school systematically documented the activities, culture, and organization of distinct groups of people in the city of Chicago, including delinquent gangs (Thrasher, 1927). In his research on gangs, Frederick Thrasher identified regularities in the emergence of gangs that hold true today. In addition, Edwin Sutherland expanded and systematized a sociological explanation of delinquency in editions of his ground-breaking book, *Principles of Criminology*, first published in 1924. His book set the stage for subsequent theoretical debate and close to 100 years of theory and research on the sociology of delinquency.

Social Disorganization

Social disorganization theory was the first sociological perspective attributing variation in crime and delinquency over time and among territories to the absence or breakdown of communal institutions (e.g., family-heads, pastors, school organizations, and local officials) and communal relationships that traditionally encouraged cooperative relationships among people. The concept is defined in terms of the absence or breakdown of certain types of relationships among people, and is intimately tied to conceptions of those properties of relationships that are indicative of social or communal 'organization.' Relationships among people in a given territory are presumed to be especially 'organized' when there are high levels of involvement across age levels in activities coordinated by representatives of the communal.

In later years, numerous sociological criminologists focused on the importance of barriers to delinquent behavior with a lack of barrier comparable to social disorganization. Organized systems have strong barriers or stakes in conformity. In one of the earliest such theories, Albert J. Reiss (1951: 196) proposed that delinquency was "behavior consequent to the failure of personal and social controls." Social control was "the ability of social groups or institutions to make norms or rules effective." Walter Reckless et al. (1956) developed a 'containment' theory, focusing on a youth's self-conception as an 'insulator' against delinquency when outer constraints were weak. Jackson Toby (1957) introduced the concept of 'stakes in conformity' to explain 'candidacy' for gang involvement. F. Ivan Nye (1958) focused on the family and specified different types of control, differentiating between internal, direct, and indirect controls. Gresham Sykes and David Matza (1957) proposed that delinquent youth learned 'techniques of neutralization' that freed them to drift into delinquency. Drifters (Matza, 1964) were depicted as youth who have few stakes in conformity and are free to drift into delinquency. Scott Briar and Irving Piliavin (1965) introduced a similar concept, 'commitments to conformity.' All of these theorists drew on the social disorganization tradition by emphasizing the lack of barriers, low stakes in conformity, or weak social control.

Cultural Conflict (Cultural Deviance)

The depiction of a society as a collection of socially differentiated groups with distinct subcultural perspectives leading some of these groups into conflict with the law is a form of cultural disorganization, typically called cultural conflict or cultural deviance. There may be perfectly consistent messages conveyed within a given subculture, but they may conflict with the views of other subcultures. University of Pennsylvania sociologist, Thorsten Sellin (1938), outlined three ways in which such conflict could occur involving distinct cultural groups. Cultural conflict could occur when (1) groups with different standards overlap or interact in border regions, (2) laws of one group are extended to encompass others, and (3) distinct groups migrate into new territories. This type of theory focuses on cultural or normative conflict as the macro condition generating crime and delinquency.

Behavior acceptable or tolerated in one context may be unacceptable in another. Moreover, because laws are likely to endorse the perspectives of the more advantaged citizens and to be enforced most rigidly when dealing with the disadvantaged, lower class and other disadvantaged categories of youth will experience the greatest conflicts with official authorities. Within their own subcultural environment, they are quite normal products of successful, but subcultural, socialization. Sutherland's differential association theory was intimately linked with the concept of cultural conflict in that variation in 'definitions' among socially differentiated groups in society was the macro condition that allowed associations to be 'differential.' The most prominent cultural conflict theorist in subsequent decades was Walter Miller (1958) who argued that the focal concerns of lower class youth led them into careers of gang delinquency.

Labeling Theory

In 1938, Frank Tannenbaum, a Columbia University sociologist, proposed an age-graded version of conflict theory emphasizing 'the dramatization of evil' where the normal street activities of youth brought them into conflict with their surrounding community. The more attention paid to youthful transgressions, the more serious their involvement in delinquency. This perspective provided the foundation of a theory emphasizing labeling processes. The theory was elaborated by Edwin Lemert (1951), Howard Becker (1963), and Edwin Schur (1971).

The theory focused on the creation of and changes in the legal categories applied to people and their behavior, circumstances affecting the actual application of legal labels, and the individual and societal effects of the labeling process. The focus was not on the causes and correlates of deviant acts but on the social and political construction of the rules defining acts as deviant and the labeling process involved in applying and enforcing those rules.

Anomie, Strain, and Subcultures

Social disorganization, cultural conflict, and the beginnings of the labeling theory had developed by the end of the 1930s. A fourth theory emphasizing a specific type of disorganization as a source of crime and delinquency was elaborated by the Columbia sociologist, Robert Merton, in 1938. Merton argued

that high rates of various forms of deviance are generated in anomic social systems where there is a strong emphasis on economic success coupled with inequality in opportunity to realize success legitimately. The pursuit of success by illegal 'innovative' means is viewed as one adaptation to strain.

Other theorists have followed the same logic introducing other forms of discrepancy between goals and means as a source of frustrated ambitions. In 1955 Albert Cohen took issue with Merton's rather instrumental view of lower class crime and delinquency and introduced emotion and anger into strain theory. He sought to explain male gang delinquency and what he believed to be the hostile, negativistic content of a gang contraculture. The source of anger and hostility was status frustration.

Another elaboration of a strain theory logic was proposed by Richard Cloward and Lloyd Ohlin (1960). They proposed that the outcome of strain varied depending on the nature of illegitimate opportunities available to youth. If adult criminal enterprises were available the outcome would be a criminal subculture of delinquent youth. In settings where organized illegitimate opportunities were not available, delinquent gangs would exhibit the characteristics of a violence prone, conflict subculture. Youth who were unable to join or succeed in conflict or criminal subcultures were double failures and candidates for retreatist subcultures. Their theory was a theoretical integration of elements of Merton's strain theory and Sutherland's theory emphasizing illegal associations and opportunities.

The Self-Report Method

Between the early 1900s and 1960s the focus of the sociology of delinquency was on theoretical development with advocates of different theories criticizing each other's perspectives. Theorists debated over the best theory, but all sought to substantiate their theories by reference to the same bodies of data providing 'facts' about delinquency – police and justice system statistics. Data compiled by 'officials' could provide information on variations by territory, race, home structure, class, and gender variation. There were no direct measures of basic concepts or of delinquency offense behavior.

Sociologists had experimented with the self-report method asking people to report on their own involvement in crime and delinquency in the 1940s. The technique gained popularity in the late 1950s with the work of James F. Short, Jr. testing Sutherland's differential association theory (1957) and F. Ivan Nye's self-report study of the family and delinquency (1958). Nye specified ways to 'operationalize' (measure) control mechanisms and related them to self-reports of delinquent behavior. Not only did these scholars show the feasibility of the self-report method in the study of delinquency, but their findings challenged conventional assumptions about social class and delinquency. Some of the patterns that had dominated discourse appeared to be products of law enforcement rather than variations in offense behavior. The focus shifted from gangs and gang behavior to self-reported offense behavior. The shift to self-reported behavior was accompanied by tests of theories using the new method.

In fact, two theories facilitated by the development of survey methods dominated empirical discourse beginning in the 1960s, Travis Hirschi's social control theory (1969) and Ronald Akers' social learning theory (1964, 1973; Akers et al., 1979). In contrast to earlier theories, neither social control theory nor social learning theory focused specifically on variations among classes or demographic categories of people. Both proposed more immediate variables that affect delinquency regardless of their distribution in the social system. Moreover, the self-report method focused on the offenses reported by random samples of youth and shifted delinquency research away from gangs and gang subcultures.

Social Control Theory

Hirschi's social control theory shared many of the characteristics of earlier social disorganization theories. When applied to delinquency the theory posited that the odds of delinquent behavior were low when youth were emotionally bonded to others (attachment), committed to conventional goals (commitment), accepted laws as morally binding (belief) and were occupied in conventional activities (involvement). While most of these variables had been proposed in earlier theories, Hirschi's elaboration drew considerable attention because he identified crucial differences between this perspective and both strain and cultural deviance theory. Moreover, he presented empirical tests relevant to his view of those distinctions using self-report survey data. His book became the most cited work in criminology in the twentieth century.

Social Learning Theory

Akers' social learning theory focused on learning experiences that could be measured using survey data and most research on that perspective has used that methodology. As a theory of criminality, social learning theory emerged from a combinations of principles derived from behaviorist operant learning and other psychological theories stressing vicarious learning and imitation. Robert Burgess and Ronald Akers reformulated differential association theory in terms of operant learning theory in 1966, and Akers and colleagues elaborated a more general social learning theory in later works (1979). The social learning theory has been subjected to more empirical tests than any other theory of delinquency.

Contemporary Developments

The perspectives discussed above (social disorganization, cultural conflict, labeling and anomie) have not been stagnant and unchanging, and distinct contemporary theories reflect basic themes from the earlier perspectives. Jensen and Rojek (2009) categorize contemporary theories in terms of their answers to three simple questions. "Does the theory assume that it is the absence of constraints or limits on opportunity for crime that are key to the explanation of crime and delinquency?" "Does the theory incorporate or require some form of imbalance, strain, stress, or frustrated pursuit of conventional

or widely shared goals?" "Does the theory focus on normal learning processes, including socially structured schedules of reward and punishment, variable values, norms, and beliefs, and associations with people who both encourage and discourage crime and delinquency?"

Extensions or Reformulations in the Social Disorganization/Social Control Tradition

Six theories can be classified as variants consistent with features of a social disorganization framework: (1) routine activities theory, (2) broken windows theory, (3) self-control theory, (4) informal control/life course theory, (5) social support theory, and (6) collective efficacy theory.

Routine activities theorists (Cohen and Felson, 1979; Cohen et al., 1981; Felson, 1998) propose that variations in crime can be explained by variations in routine activities of everyday life that affect the availability and vulnerability of people and property as targets. Temporal variations in crime rates can reflect increases or decreases in the availability of unprotected targets. The theory does not introduce any form of unusual 'pressure' to explain variation in crime.

'Broken windows' theory was proposed by Wilson and Kelling in 1982 and had a major impact on the operational theories of many police departments in the United States. The theory was based on the idea that there are observable signs of 'social disorder' in an area such as broken windows, grafitti, and abandoned cars and that removing those signs can have an inhibiting affect on crime.

Advocating a 'general theory' of crime, Gottfredson and Hirschi (1990) propose that variations in crime, delinquency, and other forms of rule-breaking can be explained by variations in the sheer possibility of engaging in an activity (opportunity) together with variations in a characteristic of individuals referred to as 'self-control.' In this version of control theory, external forces work in two ways. First, there are structured variations in the opportunities to commit some forms of crime (e.g., teenagers are unlikely to be involved in corporate fraud). Second, external variations in parental supervision and socialization are viewed as the source of variations in self-control.

Another perspective falling within that tradition is Francis Cullen's 'social support' theory (1994). Cullen proposes that 'social support' should be a central 'organizing' concept for criminology and adopts a definition of social support as "the perceived or actual instrumental or expressive provisions supplied by the community, social networks, and confiding partners." Sampson and Laub (1993) propose a 'life course' theory of crime and delinquency that addresses persistence and desistence in offending based on an 'age-graded' theory of 'informal social control.' It is called a life course theory because it focuses on how people change over time. It can be classified as a version of social control because it focuses on the role of conventional institutions and investments in explaining age variation.

Sampson and various colleagues (Sampson et al., 1997; Sampson and Bartusch, 1998; Sampson and Raudenbuch, 1999; Sampson, 2002) and Ralph Taylor (2001: 128–129) introduce a new concept, 'collective efficacy,' (social efficacy for Taylor) as the key to understanding ecological variations in crime and delinquency. When residents do get along, work through local organizations to better the community and take steps to informally control trouble in their neighborhood, they are high in collective efficacy.

Extensions or Reformulations in the Anomie/Strain/Status Frustration Tradition

There are also contemporary twists on anomie/strain/status frustration theories: (1) self-derogation theory, (2) integrated strain-bond theory, (3) general strain theory, (4) institutional anomie theory, and (5) control-balance theory. Such theories are classified here as part of the strain tradition in that they all include an imbalance of some kind that provides the motivation for violating laws.

Howard Kaplan (1975, 1980) and Delbert Elliott et al. (1985) were the first theorists to propose an integrated theory in which variables from strain, social control, and differential association theory all played a role in the explanation of delinquency. However, their perspectives fall in the strain/frustration tradition because they focus on frustrations in goal attainment as the process setting youth in search of a solution. Elliott viewed his perspective as an integration of strain and social control theory while others (Akers and Sellers, 2004) view the theory as quite compatible with differential association and social learning theory.

The title of Robert Agnew's most recent work is *Pressured Into Crime* (2005). Neither social disorganization/control theory nor cultural conflict/differential association theory require a 'pressure' to explain crime. General strain theory is a motivational theory and introduces pressures as a major source of motivation for crime and delinquency. It is called a 'general strain' theory because it is not limited to the specific types of frustrated goal achievement emphasized in the original development of the theory (i.e., failure to achieve positive goals).

Charles Tittle (1995) has developed a very elaborate theory of deviance in general named 'control-balance' theory. Because it includes control in its title it may be tempting to categorize it with social control theory. However, Tittle argues that deviant behavior 'Deviant is a device' that 'helps people escape deficits and extend surpluses.' Early strain theorists emphasized failures to achieve widely shared success goals, and Tittle incorporates problems in achieving more basic human needs or goals as providing motivation or 'predispositions' for deviant action.

In *Crime and the American Dream*, Steven Messner and Richard Rosenfeld (2001) propose yet another version of a strain/frustration theory – an 'institutional anomie' theory in which economic dominance is the cause of high crime rates in the United States compared to other nation. While the theory is clearly anchored in the anomie-strain tradition, as evidenced by its focus on 'anomic cultural pressures,' the mediating mechanisms linking economic dominance to crime are quite compatible with social disorganization/social control theory. Not only do anomic 'cultural pressures' directly facilitate crime, but they also weaken (or prevent the development of) other social institutions as sources of informal social control.

Martin Sanchez Jankowsi's *Islands in the Street* (1991) and Elijah Anderson's book, *Code of the Streets* (1999), fall in the strain tradition because the 'character' traits or 'codes' they

identify as a characteristic of gang youth are not a reflection of a wider subculture passed on through normal processes of socialization from generation to generation. Rather, gang delinquency is depicted as a response to limited resources or the irrelevance of more conventional culture to life on the streets. Based primarily on observational or ethnographic evidence, Jankowski argues that the key to understanding movement into gangs is a type of 'social character' called 'defiant individualism.' When a theorist emphasizes 'defiance' it implies more of a contraculture than a subculture.

Elijah Anderson reintroduced the notion of an oppositional subculture where self-respect requires a demonstration of toughness and maintenance of honor requires physical violence. Although the code is identified as most relevant to black youth, Jankowski's version is relevant to gang members in a wide range of ethnic groups, including whites. It should be noted that codes and characters can become 'subcultural,' if they become an enduring tradition passed on through normal learning processes in an area.

Extensions or Reformulations in the Cultural Conflict/ Differential Association Tradition

The most prominent elaboration of ideas that has been linked with the cultural conflict/differential association theory is Ronald Aker's 'social learning' theory. Social learning theorists believe that although 'definitions' play a role in the learning of conventional and unconventional behavior, such definitions do not capture all ways of learning delinquency. Social learning theory does fit with the normative conflict/differential association theory in that the learning processes are normal and can involve learning of expectations that conflict with the law. Moreover, the perspective has been expanded to include both 'micro' and 'macro' issues and is now called a 'social structure-social learning theory (SSSL).' It has been extended to apply to variations among societies as well as variations among socially differentiated categories of people within societies (see Akers, 1998; Jensen and Akers, 2003; Akers and Jensen, 2003).

Two life course theories can be considered part of this tradition as well. Terence Thornberry (1987) proposed an 'interactionist' theory that he believes integrates social control and social learning theory. It is classified here as a life course theory in the social learning tradition because the factors that influence delinquency and crime are supposed to vary by age and the stage in a delinquent career from initiation to persistence to desistence. Mark Warr (1993, 2002) has elaborated a life course perspective that specifically incorporates ideas, concepts and propositions from differential association and social learning theory.

Extensions in the Labeling/Constructionist Tradition

The labeling/social constructionist perspective has been applied to a wide range of issues using the concept of 'moral panic' to argue that many features of problem behavior are exaggerated or fabricated. Stanley Cohen (1973) is best known for his development of the concept of moral panic to explain how certain groups or episodes come to be viewed as a threat to society – a threat typically amplified by the media. The argument has been extended by Erich Goode and Nacham Ben-Yehuda (2009) to encompass forms of deviance they consider 'imaginary.' Most recently, surges in gang activity have been argued to have properties of a panic in that there may be discrepancies between any real surge in gang activity and media and police responses to it. Constructionists' positions can range from interpretations where the problem is a fabrication to milder forms where variations in some 'measurable' form of 'real world' phenomena are allowed to enter into the social construction of reality.

Marxist Theory

Several theorists (Greenberg, 1979; Schwendinger and Schwendinger, 1976; Quinney, 1980; Young, 1975, among others) have proposed perspectives which draw on ideas developed by the German social philosopher, Karl Marx. A central concept in some Marxist perspectives on delinquency is marginality. Categories of people who exist on the periphery of the economic system as an unemployed or underemployed surplus population or labor force are 'marginal' categories. Marxists attribute the high crime rate of youth to their status as a marginal category, and variations within the adolescent world reflect variations in marginality as well.

There are conflicting ideas derived from Marxist theory. For example, John Hagan et al. (1985) propose a 'power-control theory' and predict that children whose parents control other workers will be raised to take risks and to believe that they are relatively free from the control of others. Hence, children from 'command classes' will engage in more delinquency than lower status youth. Those youth who are least marginal in terms of potential dominance of others and occupational power are predicted to have the highest rates, rather than those on the periphery.

Feminist Theory

Of all the background characteristics of youths studied by criminologists, the most impressive and persistent differences in delinquency across different societies, measures of crime and delinquency, and time are gender differences. Feminist criminologists have been critical of several features of theory and research that they argue reflect a male bias. For one, there has been little research on forms of delinquency where females have rates comparable to or exceeding male rates, such as running away (Chesney-Lind, 1989; Chesney-Lind and Shelden, 1992). Since the most serious female status offenders are more likely than male status offenders to come from families where they have been the victims of male sexual abuse, the role that male dominance over women (i.e., patriarchy) can play in generating female delinquency is slighted.

Recommendations for the Future

This article outlined the history and development of the sociology of delinquency and concludes with recommendations

for the future. For one, although it is well established that variation in cultural learning is an important source of variation in delinquent behavior, the structural and historical foundation as well as the specific cultural form of those influences is totally undetermined at present. Research designed to identify the social locale of any or all of these distinct cultural forms should be a goal for the sociology of delinquency. Second, other basic sociological concepts have been used in the attempt to understand delinquency – social control, socialization, and social learning. At present, there has been no clear definition of the concepts and the differences among them. Clear and precise definitions should lead to new research establishing the independent relevance and overlap of each. Finally, social constructionists and 'realists' should pursue integrated perspectives identifying the objective and subjective foundations of the gang problem. Such research will require the use of survey methods to measure gang behavior coupled with research on media coverage.

Bibliography

Agnew, R., 2005. Pressured into Crime: An Overview of General Strain Theory. Roxbury, Los Angeles.

Akers, R.L., February 1964. Socio-economic status and delinquent behavior: a retest. Journal of Research in Crime and Delinquency 1 (1), 38–46.

Akers, R.L., 1973. Deviant Behavior – a Social Learning Approach. Wadsworth Publishing Company, Belmont, CA.

Akers, R.L., 1998. Social Learning and Social Structure: A General Theory of Crime and Deviance. Northeastern University Press, Boston.

Akers, R.L., 2000. Criminological Theories: Introduction, Evaluation, and Application, third ed. Roxbury Publishing, Los Angeles, CA.

Akers, R.L., 2006. Empirical status of social learning theory: past, present and future. In: Cullen, Frances T., Wright, John Paul, Blevins, Kristie R. (Eds.), Taking Stock: The Status of Criminological Theory, Advances in Criminological Theory, vol. 15. Transaction Books, New Brunswick, NJ, pp. 37–76.

Akers, R.L., Jensen, G.F., 2003. Social Learning Theory and the Explanation of Crime. In: Advances in Criminological Theory, vol. 11. Transaction Books, New Brunswick, NJ.

Akers, R.L., Sellers, C., 2004. Criminological Theories: Introduction, Evaluation, and Application. Roxbury, Los Angeles.

Akers, R.L., Krohn, M.D., Lanza-Kaduce, L., Radosevich, M., 1979. Social learning and deviant behavior: a specific test of a general theory. American Sociological Review 44, 636–655.

Anderson, Elijah, 1999. Code of the Street: Decency, Violence, and the Moral Life of the Inner City. W.W. Norton, New York.

Becker, H., 1963. The Outsiders: Studies in the Sociology of Deviance. Free Press, New York.

Briar, S., Piliavin, I., 1965. Delinquency, situational inducements, and commitments to conformity. Social Problems 13 (1), 35–45.

Burgess, R., Akers, R., 1966. A differential association-reinforcement theory of criminal behavior. Social Problems 14, 128–147.

Chesney-Lind, M., 1989. Girls' crime and woman's place: toward a feminist model of female delinquency. Crime and Delinquency 35, 150–168.

Chesney-Lind, M., Shelden, R., 1992. Girls, Delinquency and Juvenile Justice. Brooks Cole, Pacific Grove, CA.

Cloward, R., Ohlin, L., 1960. Delinquency and Opportunity: A Theory of Delinquent Gangs. Free Press, Glencoe, IL.

Cohen, A., 1955. Delinquent Boys: The Culture of the Gang. Free Press, Glencoe, IL.

Cohen, S., 1973. Folk Devils and Moral Panics. Paladin, St Albans.

Cohen, L., Felson, M., 1979. Social change and crime rate trends: a routine activity approach. American Sociological Review 44, 588–608.

Cohen, L., Kluegal, J., Land, K., October 1981. Social inequality and predatory criminal victimization: an exposition and test of a formal theory. American Sociological Review 46, 505–524.

Cullen, F.T., 1994. Social support as an organizing concept for criminology: presidential address to the academy of criminal justice sciences. Justice Quarterly 11, 527–559.

Elliott, D.S., Voss, H.L., 1974. Delinquency and Dropout. Lexington Books, Lexington, MA.

Elliott, D.S., Huizinga, D., Ageton, S.S., 1985. Explaining Delinquency and Drug Use. Sage, Beverly Hills, CA.

Felson, M., 1998. Crime and Everyday Life, second ed. Pine Forge Press, Thousand Oaks, CA.

Finestone, H., 1976. Victims of Change: Juvenile Delinquents in American Society. Greenwood Press, Westport, CT.

Goode, E., Ben-Yehuda, N., 2009. Moral Panics: The Social Construction of Deviance, second ed. Wiley-Blackwell, Chichester, UK.

Gottfredson, M., Hirschi, T., 1990. A General Theory of Crime. Stanford University Press, Stanford, CA.

Greenberg, D., 1979. Delinquency and the age structure of society. In: Messinger, S., Bittner, E. (Eds.), Criminology Review Yearbook. Sage, Beverly Hills, CA.

Hagan, J., Gillis, A.R., Simpson, J., 1985. The class structure of gender and delinquency: toward a power-control theory of common delinquent behavior. American Journal of Sociology 90, 1151–1178.

Hirschi, T., 1969. Causes of Delinquency. University of California Press, Berkeley, CA.

Jensen, G.F., Akers, R.L., 2003. Taking social learning global: micro–macro transitions in criminological theory. In: Akers, Ronald L., Jensen, Gary F. (Eds.), Social Learning Theory and the Explanation of Crime: A Guide for the New Century, Advances in Criminological Theory, vol. 11. Transaction Books, New Brunswick, NJ, pp. 9–38.

Jensen, G.F., Rojek, D.G., 2009. Delinquency and Youth Crime, fourth ed. Waveland Press, Inc., Long Grove, Illinois.

Kaplan, H., 1975. Self-attitudes and Deviant Behavior. Goodyear, Pacific Palisades, CA.

Kaplan, H., 1980. Deviant Behavior in Defense of Self. Academic Press, New York.

Lemert, E.M., 1951. Social Pathology. McGraw-Hill, New York.

Matza, D., 1964. Delinquency and Drift. Wiley, New York.

Merton, R.K., 1938. Social structure and anomie. American Sociological Review 3, 672–682.

Messner, S.F., Rosenfeld, R., 1997. Political restraint of the market and levels of criminal homicide: a cross-national application of institutional anomie theory. Social Forces 75, 1393–1416.

Messner, S.F., Rosenfeld, R., 2001. Crime and the American Dream, third ed. Wadsworth, Belmont, CA.

Miller, W., 1958. Lower class culture as a generating milieu of gang delinquency. Journal of Social Issues 14, 5–19.

Nye, F.I., 1958. Family Relationships and Delinquent Behavior. Greenwood Press, Westport, CT.

Quinney, R., 1980. Class, State and Crime, second ed. Longman, New York.

Reckless, W., 1961. A new theory of delinquency and crime. Federal Probation 25, 42–46.

Reckless, W., Dinitz, S., Murray, E., 1956. Self-concept as an insulator against delinquency. American Sociological Review 21, 744–756.

Reinarman, C., Levine, H., 1989. The crack attack: politics and media in America's latest drug scare. In: Best, Joel (Ed.), Images of Issues: Typifying Contemporary Social Problems. Aldine de Gruyter, New York, pp. 115–137.

Reiss Jr., A.J., 1951. Delinquency as the failure of personal and social controls. American Sociological Review 16, 196–207.

Sampson, R.J., 2002. Transcending tradition: new directions in community research, Chicago style. Criminology 4, 213–230.

Sampson, R.J., Bartusch, D.J., 1998. Legal cynicism and (subcultural?) tolerance of deviance: the neighborhood context of racial differences. Law and Society Review 32, 777–804.

Sampson, R.J., Laub, J.H., 1993. Crime in the Making: Pathways and Turning Points through Life. Harvard University Press, Cambridge, MA.

Sampson, R.J., Raudenbush, S.W., 1999. Systematic social observation of public spaces: a new look at disorder in urban neighborhoods. American Journal of Sociology 105, 603–651.

Sampson, R.J., Raudenbush, S.W., Earls, F., 1997. Neighborhoods and violent crime: a multilevel study of collective efficacy. Science 277, 918–924.

Sánchez-Jankowski, M., 1991. Islands in the Street: Gangs and American Urban Society. University of California Press, Berkeley.

Schwendinger, H., Schwendinger, J.R. Adolescent Subcultures and Delinquency. Praeger, New York.

Schwendinger, H., Schwendinger, J.R., December 1976. Marginal youth and social policy. Social Problems, 84–91.

Sellin, Thorsten, 1938. Culture Conflict and Crime. Social Science Research Council, NJ.

Shaw, C., 1929. Delinquency Areas. University of Chicago Press, Chicago.

Shaw, C., McKay, H., 1942. Juvenile Delinquency and Urban Areas. University of Chicago Press, Chicago.

Short, J.F., January 1957. Differential association and delinquency. Social Problems 4, 233–239.
Sutherland, E.H., 1929, 1934. Principles of Criminology. J.B. Lippincott, Philadelphia.
Sykes, G.M., Matza, D., December 1957. Techniques of neutralization: a theory of delinquency. American Journal of Sociology 22, 664–670.
Tannenbaum, F., 1938. Crime and the Community. Columbia, New York and London.
Taylor, R.B., 2001. The ecology of crime, fear, and delinquency: social disorganization versus social efficacy. In: Paternoster, R., Bachman, R. (Eds.), Explaining Criminals and Crime. Roxbury, Los Angeles, pp. 124–140.
Taylor, I., Walton, P., Young, J., 1973. The New Criminology. Harper and Row, New York.
Thornberry, T., 1987. Towards an interactional theory of delinquency. Criminology 25, 863–891.
Thrasher, F., 1927. The Gang: A Study of 1,313 Gangs in Chicago. University of Chicago Press.
Tittle, C.R., 1995. Control Balance: Toward a General Theory of Deviance. Westview Press, Boulder, CO.
Toby, J., 1957. Social disorganization and stake in conformity: complementary factors in the predatory behavior of hoodlums. Journal of Criminal Law, Criminology and Police Science 48, 12–17.
Warr, M., 1993. Age, peers, and delinquency. Criminology 31, 17–40.
Warr, M., 2002. Companions in Crime: The Social Aspects of Criminal Conduct. Cambridge University Press, Cambridge.
Wilson, J.Q., Kelling, G., March 1982. Broken windows. Atlantic Monthly, 29–38.
Young, J., 1975. Working-class criminology. In: Taylor, I., Walton, P., Young, J. (Eds.), Critical Criminology. Routledge and Kegan Paul, London.

Dementia: Psychiatric Aspects

Friedel M Reischies, Freie Universität Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 5, pp. 3396–3401, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

In dementia syndromes, the information processing capacity has deteriorated with deficits predominantly in cognitive, but also in emotional and intentional domains. These deficits compromise the social and cultural activities of a person to a critical degree. Dementia is a highly prevalent disorder in old age. Dementia diseases have to be distinguished from dementia syndromes. (1) A dementia disease process destroys certain brain structures, leading to clinically observable signs and symptoms. The most important primary dementia disease is Alzheimer's dementia. However, several other diseases may lead finally to a dementia syndrome, such as the accumulation of vascular brain lesions or strokes in cerebrovascular disease. The primary dementia diseases are progressive. Thus, the clinical syndrome follows a time course of progressive deterioration in different domains. (2) A dementia syndrome is a pattern of regularly observed signs in subjects with dementia diseases. There are several types of dementia syndromes. The dementia syndrome of Alzheimer's dementia is most common. A characteristic feature of the syndrome is the deterioration of cognitive functions – it first presents with a disorder of episodic memory. Then, the language functions, motor performance, and judgment become defective. Personality changes and disorders in the domain of affective and intentional functions are often observed.

Dementia has many aspects, and neurobiological, clinical, and family-related or socioeconomic factors have been investigated. Here neuropsychiatric aspects are presented, i.e., the signs and symptoms of dementia diseases that cause the disturbance in behavior, cognitive functions, and subjective experience of the patients, which are classified as emotional problems or personality change.

Definition of Dementia

In history, the term dementia has been subject to several changes in meaning. Today, it is defined as acquired, severe loss of mental functions due to structural brain diseases.

In dementia, the information processing capacity has deteriorated to a critical degree. Deficits are found predominantly in cognitive, but also in emotional and intentional domains. By definition, these deficits severely compromise the social and cultural activities of a person. There must be a loss of mental functions that had been acquired before (i.e., the long-standing cognitive deficiencies in mental retardation are excluded). The underlying pathology is a destruction or degeneration of brain structures; therefore dementia may not be caused by functional organic disorders such as psychoses or transient brain dysfunction like in delirium. Psychiatry is concerned primarily with the detection and diagnosis of dementia syndromes, and concerned secondly with the identification and treatment of certain dementia diseases. Dementia syndromes and dementia diseases have to be differentiated (Wells and Whitehouse, 1996).

Dementia Syndromes

Characteristic clinical profiles emerge, i.e., patterns of cognitive, affective, and intentional disturbances, because in dementia diseases the destruction or degeneration takes place with a predominance of certain brain regions or brain systems; the disease does not impair all brain structures to the same extent.

Dementia Syndrome of Alzheimer's Type

What is the clinical presentation, the pattern of deficits that characterize the syndrome of a dementia patient who suffers from Alzheimer's disease? This issue has been clarified in large longitudinal studies (e.g., Berg et al., 1990). A second question addresses the reasons for failures in activities of daily living and the need for care. The syndrome can be differentiated into cognitive signs and not primarily cognitive signs.

1. Cognitive neuropsychological signs
   a. *Amnestic signs.* Early in the time course of the dementia syndrome, deficits in episodic memory are noticed (Brandt and Rich, 1995). Recall of word lists, stories, second word of paired associates, or the reconstruction of figures drawings, which were encoded several minutes earlier, becomes increasingly impaired. The immediate recall of digits (digit span), however, remains typically unimpaired. The patients can immediately repeat a list of words but forget the words within several minutes, e.g., after a disruption by intermittent attention to other topics.

Furthermore, errors of orientation for time and space are hallmarks of this dementia syndrome. Demented persons get lost easily – at first in unfamiliar, and then even in familiar surroundings. The forgetting of data of major impact or the forgetting of actions with dangerous consequences affects activities of daily living. These deficits are characteristics of anterograde amnesia and are a consequence of mediotemporal brain damage, including the hippocampus and entorhinal cortex. Degeneration within the basal forebrain, however, may also explain a large part of the symptoms. Later retrograde amnesia starts and the recollection of data that were well known in former times becomes impaired. In later stages of

the disease, demented persons cannot remember anything but episodes from the first decades of their life.

To summarize, a dementia syndrome in persons with Alzheimer's disease presents initially with amnestic signs. Episodic memory and orientation, with respect to time and place, deteriorate, and a retrograde amnesia occurs.

b. *Language disturbance, aphasic signs.* Word finding is impaired early in the time course of the dementia syndrome of Alzheimer's type, but these and other signs of language impairment usually occur later than memory impairment. Difficulties in word finding are the most important feature of this language disturbance. Word finding is clinically observed mostly in the communication situation and in fluency tests.

As demented persons become increasingly incapable of finding correct words, in later stages of the disease, not much content is communicated: a phenomenon that has been called empty speech. Phonematic paraphasias (errors in the phonematic composition of words) or disordered syntax, which characterize various forms of aphasia after cerebrovascular strokes, are, however, not considered prominent signs in this dementia syndrome.

A lowered performance in fluency tasks is often assessed in clinical diagnosis, for example, by asking the patient to name as many animals as possible. It has been argued that fluency is not only a language test, but also a test of retrieval from semantic memory, and it is known that effective retrieval depends on nonlanguage functions. Patients with frontal lobe lesions cannot retrieve words in fluency tasks, presumably because of inefficient retrieval strategies and a defective suppression of overlearned alternative word-finding routines. Furthermore, demented persons cannot retrieve semantic features of perceived objects because of a disturbance in semantic memory (Martin and Fedio, 1983).

Later in the time course, language comprehension difficulties additionally impair the communication with a demented person. In summary, there are important language problems in the dementia syndrome of Alzheimer's type, which are predominantly caused by deficits in word finding and language comprehension.

c. *Apraxia signs.* Persons with the dementia syndrome of Alzheimer's type show errors in performing purposive movements. These errors occur although there are no obvious motor problems, such as weakness of the limb or sensoric deficits in guiding movements. It is thought that the central control of movements is defective. Apraxia in the narrow sense describes errors in specific action sequences, for example, using tools. Clinically, errors in dressing properly are encountered. Tests of apraxia include the call for imitation of limb movements or gestures.

In diagnostic schedules for dementia, constructional apraxia is tested quite often. This apraxia comprises deficits not only of the execution of movement schemata but also of the integration of visuospatial information into appropriate action (Heilman and Rothi, 1997). Copying visual designs or drawings of well-known objects, such as a clock, are defective and are performed laboriously and slowly. In addition, many demented subjects show an apractic type of writing disorder, where the action sequences of writing the individual letters can no longer be carried out correctly.

To summarize, subjects with Alzheimer's dementia show errors in performing motor tasks without weakness or paresis of the limbs, resembling central motor control deficits. Later in the time course of the disease, the impairments lead to inability to perform any practical task in the household or in self-care.

d. *Disorders of judgment.* Judgment errors and inadequate evaluation are especially relevant for the everyday functioning of demented persons and with respect to legal responsibility. In the narrow definition, erroneous or inadequate judgments, with full understanding and appreciation of the problem, can be related to the decision process in itself. This specificity is difficult to ascertain in clinical neuropsychological investigation. Therefore, more generally, inadequate judgments about the own person and the disease are considered for diagnostic evaluation.

Self-related judgments: Demented persons often cannot correctly evaluate the cognitive deficits they suffer from. Some patients describe grossly inadequate future plans. At times, not only disorder of judgment but also the psychopathological phenomenon of denial has to be considered in such situations.

Formal assessment of logical decisions, as part of the capacity for adequate judgment, is done usually by intelligence tests which demand, for example, the understanding of a logical rule within the sequence of symbols, and finding incorrect items.

In summary, inadequate or wrong judgments in self-related or logical problems are found commonly in the dementia syndrome of Alzheimer's type and may lead to legal consequences or intensify the demand of care for the patient.

e. *Other cognitive neuropsychological signs.* Demented persons with Alzheimer's diseases often show deficits in executive functions. The term 'executive functions' denotes the planning and the execution of goal-directed activities, including preparations and the monitoring of already reached partial goals. Some researchers also included self-regulation and volitional processes in the domain of executive functions. Brain structures for executive functions are at least in part localized in the dorsolateral prefrontal lobe.

Mental calculations are often tested in the clinical diagnosis of a dementia syndrome. They do not, however, test calculation function only. For these tasks, the attentional demands, and additionally, in serial calculations, the demands on working memory are high. Therefore, the reason for individual test deficits is difficult to attribute precisely.

Agnosia is mentioned in some of the dementia criteria. The classical form of object agnosia, however, is very rare. In object agnosia, the patient perceives an object – and is even able to copy the form – but is not able to grasp its meaning, to name it, or describe its attributes or functions. Demented persons

often fail to identify faces of well-known friends or relatives in later stages of the disease, but memory problems contribute to this deficit. More often, however, a perceptual deficit concerning visuospatial features of figural information can be documented by neuropsychological tests. Furthermore, the appreciation of the relation of many objects within a complex pictorial scene is often disturbed in dementia.

In addition to cognitive deficits, there are other dimensions of the dementia syndrome of Alzheimer's type.

2. Noncognitive part of the dementia syndrome of Alzheimer's type

Patients with a dementia syndrome of Alzheimer's type suffer from affective symptoms or delusions. Noncognitive signs of dementia are traditionally described by psychopathology. They are not specific for dementia. Several of these signs are commonly observed in other psychiatric diseases, for example, anxiety or restlessness. Although these signs are not as relevant for diagnosis as cognitive deficits, they present major problems for the caregivers (Cummings et al., 1996) and are often the reason for institutionalization of demented patients.

a. *Signs in the affective domain*. Early in the development of the dementia syndrome, depression is observed. Later, affect control becomes impaired. Patients may cry after minor adverse events or be aggressive or inadequately cheerful. Often the emotional state lasts only for a short period of time. The perceived emotional experience may not be as dramatic as the affective expression suggests. This phenomenon has been termed affective lability.

Affective states like aggression may also be expressed as an increased intensity. Such behaviors may often present as behavior that cannot be tolerated and may lead to institutionalization.

b. *Signs in the intentional and attentional domain*. The cerebral information processing for motivation and attention is impaired in late stages of Alzheimer's dementia. A state of 'apathy' is described quite often. A person, for example, sits in a chair for most of the time and does not move, and has no intention to start any action. Many of the noncognitive signs of the dementia syndrome are traditionally described as organic personality change. A further component of this is, as a consequence of frontal lobe disturbance, a disinhibition of action tendencies, which are usually suppressed because of cultural constraints. In addition, paranoid delusional thinking and hallucinations occur quite often within the time course of a dementia syndrome of Alzheimer's type. These also may induce aggressive behavior.

In summary, signs in multiple domains of brain functions are observed. The signs that characterize the dementia syndrome of Alzheimer's type are predominantly from the cognitive domain. However, disorders in the emotional and intentional domains have some impact for the resulting individual syndrome and for caregivers and relatives. The primary sensory and motor functions, however, are relatively spared. Accordingly, the spread of Alzheimer's pathology in the brain does not include the primary sensory and motor areas.

Frontotemporal Dementia Syndrome

The dementia syndrome of frontotemporal type is less common than the dementia syndrome of Alzheimer's type and should be described here in outline only (Brun et al., 1994). The major feature of this syndrome is a change in personality. Features of the dementia syndrome of Alzheimer's type, like predominantly amnestic and apractic signs, or a decrease in intellectual functions, are not detectable during early stages of the frontotemporal dementia syndrome.

The personality change comprises changes in the way of experiencing certain situations, in the way of reacting in certain situations, or in the way a person was used to acts spontaneously.

1. *Lack of expected reactions or actions*. The patient is aspontaneous and does not react as they did before the onset of the dementia syndrome. At times, a lack of drive is noted. As there is no emotional symptom of depressed mood and despair, this has been named pseudodepression.
2. *Disinhibition*. A lack of control of intentions, especially with respect to social norms or cultural restraints, is observable.
3. *Perseveration and stereotyped behavior*. The patient reiterates actions and is inflexible in the way of reacting.
4. *Lack of insight and disorder of judgment*, as described above (see section Cognitive Neuropsychological Signs).

In addition to the psychopathological symptoms described above, formalized neuropsychological testing reveals deficits in planning, in shifting mental set, and in working memory. The frontotemporal dementia syndrome can be identified clinically only in early stages. In late stages, mostly an unspecific terminal dementia syndrome is clinically observed.

As for the dementia syndrome of Alzheimer's type, also in frontotemporal dementia there is a heterogenous etiology with mostly genetically determined underlying diseases.

Dementia Diseases

Dementia has two sides – the individual pattern of neuropsychological or psychopathological impairments on the one side, and the damage or degeneration within the brain on the other side.

A dementia disease is a class of pathophysiological processes that result in structural brain changes that are underlying the clinical signs of the dementia syndromes (Wells and Whitehouse, 1996; McHugh and Slavney, 1998). There are many diseases of the brain that lead to a dementia syndrome.

Primary Dementia Diseases

Alzheimer's dementia disease, Pick dementia disease, or Lewy body dementia are degenerative brain diseases which up to now inevitably lead to a progressive dementia syndrome. They cause no other symptoms except symptoms of the dementia syndromes.

Neurological Diseases with Associated Dementia Syndrome

1. In some diseases the dementia outcome is obligatory. In Huntington's chorea, for example, a movement disorder precedes the progressive dementia syndrome, which regularly develops later in the time course of the disease.

2. In other diseases, the dementia outcome is facultative. These diseases are not dementia diseases per se. Vascular dementia, which is the second most frequent cause of dementia cases, is in some of its variants not associated with a dementia syndrome. Prevention of subsequent strokes, furthermore, may stop the cerebrovascular disease and prevent the development of vascular dementia.

The brain behavior relationship is interindividually variable and even the distribution of pathological changes within one disease is varying. Because in many cases the dementia disease cannot be identified reliably before neuropathological postmortem examination, the clinical dementia syndrome guides the medical decisions about treatment and secondary prevention.

Dementia Disease Etiology

Alzheimer's disease is genetically heterogenous. Many different mutations on several genes are known to cause a presenile Alzheimer's dementia. The same is true for frontotemporal dementia. Additionally, multiple causal factors are identified in senile dementia of Alzheimer's type, and it is not known whether an interaction of several factors is responsible for senile dementia cases. In some dementia cases, etiologically distinct causes are known, like in dementia following traumatic brain lesion or in a dementia patient with a series of strokes within strategic regions of the brain.

Diagnosis of Alzheimer's Dementia

The diagnosis of a dementia disease will be described for the most frequent dementia, Alzheimer's.

1. The confirmation of the disease can only be given by a neuropathological investigation of the brain after death. The microscopic analysis will reveal the amount and distribution of neuropathological features like neuritic plaques and neurofibrillary tangles.
2. There is some uncertainty as to the clinical diagnosis of Alzheimer's dementia. There are several lists of criteria proposed, e.g., by diagnostic systems, such as International Classification of Diseases (ICDs), and by consensus committees (e.g., the classification of probable or possible Alzheimer's dementia according to clinical criteria). However, the criterium lists differ considerably, and the criterium variance results in major differences in the estimated prevalence rates of population-based studies (Erkinjuntti et al., 1997).

Prospective studies of dementia syndromes improved the specificity of clinical diagnostic methods. However, at present technicians cannot distinguish between dementia diseases such as Lewy body dementia and Alzheimer's dementia with sufficient reliability.

Problems of Diagnosis of Mild Dementia in Old Age

The factors mentioned add to the diagnostic uncertainty with respect to differential diagnosis. Most dementia cases are seen in old age. The differentiation from normal aging and organic psychiatric disorders, like mild cognitive impairment or amnestic syndrome, is for many cases not easy or, especially in early stages, not possible at all.

The early detection of dementia is difficult in old age, because age in itself is associated with some cognitive decline (Reischies, 1998). Furthermore, mild cognitive deficits have to be differentiated from low talent domains of a person, and personality changes have to be diagnosed considering the premorbid personality.

Dementia Case

If the severity of the cognitive impairments of the dementia syndrome of Alzheimer's type is sufficiently high, and the severity of impairment in everyday functioning is high as well, a dementia case can be diagnosed. For these decisions, cutoff values of assessment instruments have been applied (e.g., the Mini Mental State Examination and the Global Assessment of Functioning scale).

The diagnosis of a dementia case without specification of the kind of probable dementia disease is often a clinical reality. With the advent of better diagnostic methods, however, the diagnosis of the dementia disease may be made before a clinical dementia syndrome has developed.

Time Course of Alzheimer's Dementia

There is a characteristic time course of Alzheimer's dementia over about 10 years. Interindividual variation of the time course with plateaus and progression phases is common.

Within a preclinical stage of Alzheimer's disease, cognitive decline takes place, which can be assessed by serial neuropsychological testing. Especially memory decline can be found. With memory deficits as the only clinical sign, no formal diagnosis of dementia can be made. The early phase of Alzheimer dementia thus represents an amnestic stage.

In the second stage, a clinical dementia syndrome of Alzheimer's type is detectable, which exhibits more or less the complete features of the dementia syndrome, as described above. In this stage, there are disturbances in activities of daily living, but they may still be minor.

In the stage of a severe dementia syndrome, memory loss, aphasia, apraxia, and the noncognitive symptoms have increased, and global functioning is deteriorated such that normal activities of daily living are critically compromised or no longer possible. Additionally, the noncognitive disorder disturbs the relation between the patient and the environment. Institutionalization is common.

The last stage is a terminal or vegetative stage, where no communication is possible, the patient is mostly bedridden, and has to be cared for completely.

Treatment, Professional Care, and Preventions

An increasing number of new drug treatments for Alzheimer's dementia are being investigated. However, at present only limited relief can be provided. At best, the time course can be stopped or retarded for a short time only. Training of activities

of daily living, with the goal of adaption to the specific environment, can be quite successful in order to aid care of subjects with an advanced dementia syndrome.

Major efforts must be directed to prevention, detection of risk factors, and early detection of dementia syndromes. A certain success of prophylaxis by controlling risk factors for atherosclerosis has been documented in the field of vascular dementia.

Editor's Note

Recent advances in biomedical research on the various dementia disorders are well summarized in David Kennedy (2013). Other useful updates include those by William A. Banks and John E. Morley (2003) and Samantha Loi and Edmond Chiu (2011).

See also: Aging Mind: Facets and Levels of Analysis; Aging and Memory; Alzheimer's Disease, Neural Basis of; Alzheimer's Disease: Behavioral and Social Aspects; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Cognitive Aging; Dementia: Psychiatric Aspects.

Bibliography

Banks, William A., Morley, John E., 2003. Memories are made of this: recent advances in understanding cognitive impairments and dementia. Journal of Gerontology A, Public Policy and Aging Report 58 (4), 214–321.

Berg, L., Miller, J.P., Morris, J.C., Rubin, E.H., Smith, D.S., Storandt, M., Coben, L.A., 1990. Mild senile dementia of the Alzheimer type: 3. Longitudinal and cross-sectional assessment. Annals of Neurology 28, 648–652.

Brandt, J., Rich, J.B., 1995. Memory disorders in the dementias. In: Baddeley, A.D., Wilson, B.A., Watts, F.N. (Eds.), Handbook of Memory Disorders. J. Wiley, Chicester, UK, pp. 244–270.

Brun, A., Neary, D., Englund, B., Gustafson, L., Mann, D.M.A., Passant, U., Snowdon, J.S., 1994. Clinical and neuropathological criteria for frontotemporal dementia. Journal of Neurol Neurosurgery and Psychiatry 57, 416–418.

Cummings, J.L., Fiorello, T., Gornbein, J., Mega, M.S., 1996. The spectrum of behavioral changes in Alzheimer's disease. Neurology 46, 130–135.

Erkinjuntti, T., Østbye, T., Hachinski, V., Steenhuis, R., 1997. The effect of different diagnostic criteria on the prevalence of dementia. The New England Journal of Medicine 337, 1667–1674.

Heilman, K.M., Rothi, L.J.G., 1997. Apraxia – The Neuropsychology of Action. Psychology Press, Hove, UK.

Kennedy, David, 21 May 2013. Recent Advances in Dementia Research. Primary Psychiatry. Retrieved from: http://primarypsychiatry.com/recent-advances-in-dementia-research/.

Loi, Samantha, Chiu, Edmond, 2011. Recent advances in drug treatment for Alzheimer's dementia. Asia-Pacific Psychiatry 3 (3), 115–119.

Martin, A., Fedio, P., 1983. Word production and comprehension in Alzheimer's disease: the breakdown of semantic knowledge. Brain and Language 19, 124–141.

McHugh, P.R., Slavney, P.R., 1998. The Perspectives of Psychiatry. Johns Hopkins University Press, Baltimore, MD.

Reischies, F.M., 1998. Age related cognitive decline and the dementia threshold. In: Lomranz, J. (Ed.), Handbook of Aging and Mental Health. An Integrative Approach. Plenum Press, New York, pp. 435–448 (Chapter 19).

Wells, C.E., Whitehouse, P.J., 1996. Cortical dementia. In: Fogel, B.S., Rao, S.M., Schiffer, R.B. (Eds.), Neuropsychiatry. Williams & Wilkins, Baltimore, MD, pp. 871–889.

Democracy, History of

James J Sheehan, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 5, pp. 3408–3413, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Political authority in a democracy is based on the will of the people. From its origins in ancient Greece, the advocates and opponents of democracy have argued about what this means. Who are the 'people'? How is their will to be expressed? What is the relationship between democratic politics and society? How are the rights and opinions of the minority to be protected? What are the primary threats to effective democracy and how can they be avoided? What are the dangers of democracy? In the modern era, democratic practice began with the great revolutions at the end of the eighteenth century. Throughout the nineteenth century democratic institutions developed everywhere in Europe; at the same time, the principle of popular sovereignty was applied to a range of social, economic, and national issues. Between 1914 and 1945, democracy seemed in danger, threatened on left and right by movements that used democratic methods to establish new sorts of tyrannies. In the second half of the twentieth century, however, democracy seemed to triumph over its opponents until it became, in theory if not in practice, the sole source of political legitimacy in the modern world. This has only been reinforced in the twenty-first century.

In a democracy, political authority is based on the will of the people. From its origins in ancient Greece, the meaning of democracy has been contested and always questioned (Gallie, 1955–56). The history of democracy is the history of these debates, debates between democrats and their opponents and, no less vigorously and persistently, among the advocates of different sorts of democracy. Three overlapping sets of issues have dominated disagreements about democratic theory and practice. First and most obviously is the question of whom to include in the demos. If democracy means government by the people, who are the people? Everyone? All adults? All male adults? All male adults with property or education? Second is the matter of democracy's prerequisites. What rules, institutions, and values are necessary for democracy to succeed? Are correct procedures sufficient to ensure democratic government or does genuine democracy require a particular kind of society or culture? Is there one road to democracy or many? And finally, there is the issue of what might be called democracy's institutional scope. Is democracy primarily a political system of self-government or is it equally important, or equally necessary, in other aspects of life, such as the family, the workplace, and the schoolroom? In order for there to be a 'true' democracy must its principles and aspirations be realized throughout the social order?

Classical Models

The openness and complexity of democracy's history were already apparent in the fifth and fourth centuries BC when a number of Greek-speaking city states scattered across the eastern Mediterranean adopted various forms of popular self-government. From the beginning, there was an ambiguity in the meaning of demos which reflected the Greeks' uneasiness about democratic government. For democracy's admirers, the people's will confirmed legitimacy: decrees passed by popular assemblies often began with the words, 'the demos has decided.' But for democracy's critics, the people were the mob, unreliable and easily swayed by passion and self-interest. This is why Plato, in his famous attack on democracy in the eighth book of The Republic, argues that the excesses of democracy will inevitably lead to tyranny.

The ancient city's distinctive blend of religion and politics, the small scale of its public life, the superiority of the spoken over the written word in its political deliberations, and its emphasis on military service and war, reflected a distinctive set of historical conditions that would never be repeated. Nevertheless, the Greek experience continued to play a role in Europe's political imagination. Pericles's eloquent defense of Athenian democracy, recorded (or perhaps composed) by Thucydides in his history of the Peloponnesian War, was persistently cited as a statement of democratic virtue, just as Plato's and Aristotle's criticisms of democracy were used to underscore its dangers. Like the elements of classical buildings, which architects adapted over and over again to serve their own purposes, Greek ideas about democracy were repeatedly borrowed by European thinkers, who often imbedded them in structures quite distant from their original meaning and purpose.

There is no institutional continuity between ancient Greece and modern politics. The democratic practices that Pericles praised and Plato decried did not survive the destruction of the polis by internal discord and foreign enemies. To be sure, the Roman state based its legitimacy on the popular will: its armies fought in the name of 'The Senate and People of Rome.' But after the fall of the republic, first the substance and then the form of popular sovereignty disappeared from European history for almost a millennium. When they reappeared in the late middle ages, republican institutions were limited to relatively small entities, usually city-states, in which political power was in the hands of a narrow, exclusive elite. There were also a few self-governing rural communities in medieval and early modern Europe, but they were small and scattered. None of

these polities thought of itself as a democracy. The concept, revived with the rest of the Greek legacy in the late middle ages, was a matter for political theory and scholarly debates, not the everyday conduct of public affairs.

The Age of Democratic Revolution

The history of modern democratic practice begins with a series of political revolutions in the final decades of the eighteenth century. But before erupting into the realm of practical politics, democratic ideals were nourished by three currents of enlightenment thought: first was a new concept of the individual, whose identity was defined by his (and for a few, also her) own talents and desires, rather than his place in the social order; second was a new concept of equality, which accorded to each person (definitions of which varied greatly) the same rights and responsibilities; and finally there was a new concept of social and intellectual activity, which imagined a sphere of free competition among people, commodities, and ideas. Many, perhaps most, of the thinkers who formulated these ideas did not think of themselves as democrats; some of them expressly condemned democratic institutions in terms that Plato would have recognized. Nevertheless, enlightenment notions of individualism, equality, and society helped to undermine the ideological foundations of the old regime and to open the way for democracy's long and circuitous march to political power.

Among eighteenth-century political theorists the most significant for the history of democracy was Jean-Jacques Rousseau, whose ideas inspired the French Revolutionaries' assaults on the Old Regime and at the same time reflected democracy's deepest ambiguities. Rousseau is correctly regarded as a spokesman for individual liberty and, no less accurately, as the advocate of the individual's subordination to the 'general will.' He attempted to reconcile these two convictions with a program of education that aimed at freeing individuals from the corruptions of existing society and a program of political reform that was supposed to create institutions appropriate for the products of this education. The result would be a genuinely free society in which individuals want to choose what is best for them and for their community. Because this vision of total harmony between individual desire and social imperatives set a standard that no government could match, among Rousseau's legacies was the temptation to force people to be free, that is, to use political power to shape their desires to match the alleged needs of the community. At the same time Rousseau stated that democracy should never be established as a form of government, being 'a government of gods rather than men' (he preferred an aristocracy of the wise). Rousseau also declared that the sovereignty of the people that was the basis of every legitimate state could only ever be established in small states that were not commercial (he loved the Alpine cantons). Large states like France, he argued, were beyond the pale and could never be reformed. The fact that Rousseau inspired the popular aspects of events in France in the 1790s neatly captures the complicated nature of democracy and the irony that the philosopher in whose name the Revolution was conducted would himself have rejected the very idea of revolution in France.

The first step from democratic theory to democratic practice was taken on the far fringes of the European world, in North America, where the struggle for independence from colonial rule was carried out in the name of 'the people's' natural and inalienable right to self-government. Of course the new American democracy was qualified in a number of important ways: the demos excluded women, African-American slaves, and, in some places, men without sufficient property. Nevertheless, the creation of a republic in the new world established a regime whose institutions and legitimacy depended on the people's will.

The democratic experiment in America aroused widespread interest and some enthusiasm, but no imitators among enlightened Europeans. But when revolution engulfed France, Europe's greatest power and the recognized center of fashion and culture, its wider significance could not be ignored. In the summer of 1789, the French National Assembly solemnly proclaimed 'the rights of man and the citizen,' which promised social equality and political liberty. Sovereignty was now firmly rooted in the popular will: 'No body and no individual may exercise authority which does not emanate directly from the nation.' But who spoke for the nation? At each stage of the revolution there was a different answer to this question as power moved from the nation's elected representatives, to the mobilized people of Paris, to the Committee of Public Safety, to an elite of propertied men intent on restoring order, until it finally came to rest in the hands of an heroic young general, Napoleon Bonaparte. As the revolution moved from moderation to terror to reaction to military dictatorship to imperial monarchy, each successive regime invoked the nation, each claimed to rule in the people's name. Beyond France's borders, the revolution's friends and enemies shared one essential conviction: once the forces of democratic government had been unleashed, the world would never be the same.

For more than a century after the revolution was defeated, its historical memory lived on, a common source of inspiration and anxiety. People drew from the revolution's grand narrative the lessons appropriate for their political hopes and fears. To a democrat like Jules Michelet, the revolution was the truest expression of France's commitment to liberty, the most glorious chapter in the 'people's' struggle to realize their historical destiny. Reactionaries like Hippolyte Taine drew a different message from the revolution's course: popular sovereignty, he argued, will always lead first to anarchy, then despotism. But the revolution's most astute analyst was neither a democrat nor a reactionary: Alexis de Tocqueville saw it as part of a larger historical process through which the ideals and institutions of old-regime Europe were destroyed by the forces of social and political democracy that were deeply woven into the fabric of modern life. Like many of democracy's critics, Tocqueville feared it could easily lead to dictatorship; only in America was there a chance that democracy's true promise could be fulfilled (de Tocqueville, 1988).

The Nineteenth Century

In 1815, many Europeans hoped or feared that democracy had been defeated for ever. The monarchical principle seemed triumphant; republican institutions survived in Switzerland

and a precarious handful of city-states. Although monarchy remained the most common form of government in Europe for another century (France after 1870 was the most important exception), almost everywhere representative institutions gradually began to play an important role in public life. The right to vote, always the most obvious marker of democracy's progress, was granted to an increasing number of European males in the course of the century; political parties, interest groups, trade unions, and a variety of newspapers and other publications provided an infrastructure that supported political participation. By 1900, therefore, the distinction between monarchical and democratic governments, which conservative statesmen had struggled to maintain, was hopelessly blurred.

Every political movement had to come to terms with democratic pressures. During the first half of the century, most European liberals, who were the direct heirs of the Enlightenment's ideals of individualism, equality, and freedom, were suspicious of the masses – all the more so after the collapse of the French republic into terror and dictatorship. But in the course of the century, the left wing of the liberal movement was reconciled to democracy, supported universal suffrage (for men) and tried, with varying degrees of success, to win their share of the newly enfranchised voters. Conservatives, who had begun the century as democracy's most consistent enemies, were also unable to resist being pulled into the orbit of democratic politics. Following the example of the Emperor Napoleon III, who had consolidated his position with popular plebiscites after being elected president of France in 1848, conservative statesmen such as Bismarck in Germany and Disraeli in England, set out to use democratic means in support of conservative ends. No less remarkable than these 'Tory democrats' were the defenders of the Roman Catholic church, another of democracy's former enemies, who turned to electoral politics to defend their religious interests against the intrusions of the secular state. Here, too, we find evidence of the inexorable expansion of democratic ideals and institutions.

Some nineteenth-century democrats insisted that political reform was not enough; authentic self-government, they believed, required social revolution. Far and away the most influential version of this belief was formulated in the 1840s by Karl Marx and Friedrich Engels, who argued that a proletarian revolution was required to establish a truly free and equal society. To Marx, Engels, and their more orthodox disciples, democratic politics was always a means to an end: parties, unions, elections had no value beyond helping the proletariat obtain the power they needed to transform society. In this transformed society, organized political activity would become superfluous because class conflict would disappear. But as so often happens, the socialist parties' democratic means began to shape the ends for which they fought. Eventually the political organizations they built took on their own value, winning elections became an end in itself, democratic politics, a way of life. By the early twentieth century, the European labor movement, most prominently the Social Democratic Party in Germany, had become the most consistent and resolute defender of democratic reform.

In the nineteenth century, the most significant extension of democratic aspirations was not to the social order but rather to the national community. The French revolutionaries had insisted that because people did not belong to a ruler, they could not be traded back forth like pieces of property; nations, like individuals, had the sacred right to determine their own character and destiny. In the course of the century, self-determination came to mean the right of nations to have their own state, a fateful combination of democracy and sovereignty that made the national state the only legitimate way of organizing political space. In the middle decades of the century, Italians and Germans formed new nations; nationalities like the Polish and Irish struggled unsuccessfully to follow their example. In fact, nationality was an elastic and elusive quality, easy enough to invoke, but extraordinarily difficult to use as the foundation for a state. Even in the West, where strong states had had centuries to impose a kind of cultural homogeneity, ethnic minorities remained, while in the multinational empires of the East, the complexities of ethnic settlement meant that any potential nation state would have to contain substantial linguistic and religious minorities. Self-determination, the British political scientist Ivor Jennings remarked, 'seemed reasonable enough: let the people decide. It was in fact ridiculous because the people cannot decide until someone decides who are the people' (Anderson, 1996: p. 40).

As difficult as the question of self-determination might be within Europe itself, its problems became even greater when applied to the rest of the world, much of which remained either directly or indirectly under European control. But if the Germans and Italians had a right to their own nation, if the Irish and the Polish could aspire to nationhood, what about the populations of India, Africa, and Asia? In the nineteenth century, even progressive Europeans would have responded that these peoples were not yet ready for independence because they were barbarous. John Stuart Mill, for instance, believed that 'independence and nationality, so essential to the due growth and development of a people further advanced in improvement, are generally impediments' to uncivilized peoples (Mill, 1859: p. 252). But by the early twentieth century, advocates of national independence had challenged European influence in the ancient empires of the Manchus and Ottomans, while the ideals of democratic self-determination slowly began to attract followers throughout the colonized world. The global movement against colonization, which would be central to the history of the twentieth century, had many sources and took various forms, but its origins and eventual triumph cannot be understood without the democratic tools that had been forged within Europe itself.

The Twentieth Century

The historical period that stretched from the Congress of Vienna in 1815 to the outbreak of World War I in 1914 began with democracy's defeat and ended with its apparently irresistible, if still notably incomplete spread throughout Europe. The period from 1914 to 1945 was characterized by a series of near fatal assaults on the integrity and viability of democratic institutions. There were moments during this dark epoch when thoughtful people could readily imagine that the ideals that had inspired Rousseau and the eighteenth-century revolutionaries, having been abandoned by their advocates and defeated by their enemies, would be consigned to history's dustbin. Who in 1940 could have predicted that, at the century's end, it

would be democracy's enemies who would everywhere be in disarray?

That the years after 1914 would be so full of peril for democracy was not immediately apparent. In fact, the First World War destroyed many of democracy's most obdurate opponents, especially the Russian empire, which collapsed into revolution in 1917. Moreover, in most of Europe and America the war accelerated the extension of voting rights to women, an issue that had become increasingly contentious after the turn of the century. Finally, the war ended with the victory of the Western powers, allied since 1917 with the USA, whose president, Woodrow Wilson, proclaimed that the purpose of the war was to make 'a world safe for democracy.' Wilson's League of Nations, a keystone of the treaties signed in the spring of 1919, was designed to create a society of democratic states, in which self-government at home would nurture goodwill abroad and thus purge the world of the twin evils of tyranny and aggression.

The Wilsonian peace was supposed to be the fulfillment of the belief, held in the eighteenth century by men like Immanuel Kant and in the nineteenth by the followers of Jeremy Bentham, that democracy and peace inevitably went together. 'There are properly no other criminals than the heads of nations,' Bentham had declared, 'the subjects are always honest' (Hinsley, 1963: p. 86). Always a triumph of hope over experience, this belief was especially illusory in the years after 1919, when, having been brutalized by the war and frustrated by its outcome, the peoples of Europe were not necessarily peace-loving, while their elected leaders – including those in the USA – were not prepared to cooperate in a common defense of democratic institutions.

In retrospect, therefore, it is clear that two other events in the spring of 1919 were more significant for democracy's immediate future than Wilson's treaties. In March, Benito Mussolini formed his first fascist organization, an inchoate band of the alienated and discontented, which, within 3 years, was able to take advantage of Italy's postwar crisis and become the basis for a new government. In one state after another, parliamentary democracy collapsed in the twenties and thirties, and nowhere more consequently than in Germany, where Mussolini's admirer, Adolf Hitler, created a dynamic regime based on race, radical nationalism, and anti-communism. Mussolini's fascism and Hitler's National Socialism were at once the enemies and the perversions of democratic politics. Both of them blended dictatorial practice with the trappings of participatory politics – a mass party, rallies and demonstrations, and plebiscites.

A few weeks after Mussolini formed his fascist organization in Milan, Lenin established the Third International in Moscow. Lenin's hopes for world revolution, still very much alive in 1919, were soon disappointed. Contrary to Marxist theory and his own expectations, Lenin was faced with the need to build a socialist society in Russia. This required the violent repression of dissent by the revolutionary regime and then, under Lenin's successor, Joseph Stalin, the ruthless use of state power to create a new social order from above. Tens of millions died in the doomed attempt to turn the USSR into a workers' paradise. Like fascism, which claimed to rule on behalf of the nation, Soviet Communism was a perversion of nineteenth century democratic aspirations which sought to conceal a brutal tyranny behind a facade of cynical democratic rhetoric and empty participatory practice. That so many intelligent people outside of the USSR were drawn to communism in the 1920s and 1930s was a symptom of the political malaise that afflicted democratic regimes everywhere in Europe and America.

The second global war of the twentieth century ended in 1945 with the defeat of the fascist powers. The interwar political spectrum, in which democratic regimes had faced enemies from both the left and the right, was now dramatically simplified; radical rightwing parties were discredited, outlawed, and, in some places, vigorously repressed. Almost all of the regimes that emerged from the war's wreckage declared themselves to be democratic. In the East, 'people's democracies' were formed under the auspices of the USSR and, with the exception of Yugoslavia and Albania, with the assistance of Soviet armies of occupation; in the West, parliamentary regimes were allied with and, to varying degrees, influenced by the USA. This bipolar division of the continent lasted until 1989, when the creation of a reformist regime in the USSR opened the way for democratic movements throughout Eastern Europe. With remarkable speed and, except in Romania, without violence, the communist regimes were driven from power and new, authentically democratic governments were elected to take their place. With the collapse, first of the Soviet Imperium and then of the USSR itself, a new era in the history of democracy – with a new set of problems and possibilities – clearly began.

Among the casualties of World War II was European colonialism. The principle of national self-determination, proclaimed by Wilson at the end of World War I and reaffirmed by Franklin Roosevelt and Winston Churchill at the beginning of the second, was not applied to the non-European world without great delay and difficulty. Obviously Japan lost its overseas possessions when it lost the war. While neither Britain nor France, who were on the winning side in 1918 and 1945, wanted to abandon their colonies, they lacked both the ability and the will to maintain them in the face of indigenous resistance, the hostility of both the USA and the USSR, and the reluctance of their own peoples to support a vigorous defense of their empires. Sometimes peacefully, sometimes after violent wars of liberation, Europe's former colonies achieved independence; by the 1980s virtually no residue of direct imperial rule remained. Although the struggle against colonialism was carried out in the name of national self-determination, it proved to be extremely difficult to plant democratic institutions in the new states. With a few notable exceptions, most significantly India, the world's largest democracy, most of Europe's former colonies, especially in Africa, fell under authoritarian regimes, often based on the power of the military. For many of the world's population, therefore, the promise of democracy remained unfulfilled.

Conclusion

Despite democracy's disappointing record in much of Africa and Asia, there are more functioning democracies in the world at the beginning of the twenty-first century than at any other time in history. Furthermore, alternative sources of political authority have apparently disappeared; even dictatorial regimes usually feel obliged to exercise power in the people's name. Elections, even when they are manifestly fraudulent,

have become the essential rituals through which authority is legitimized.

Although the alternatives to democracy have largely disappeared, at least in theory, debates about democracy's meaning continue. Some of these debates are new versions of old contests about the composition of the demos, the procedures best suited to guarantee democratic representations, the social and cultural preconditions of popular rule, and the proper scope of democratic politics. Some of the contemporary debates over democracy's meaning reflect new sources of conflict, such as the question of how – or whether – ethnic minorities should be shielded from the impact of majority rule? Others concern the extension of democratic principles into families, schools, religious institutions, health care systems.

These debates, as painful and occasionally destructive as they sometimes are, reflect the continued vitality of the democratic tradition. A more serious danger to democracy is a growing skepticism about whether democratic institutions can solve the most significant problems people face in their everyday existence. To the rich, democracy sometimes seems superfluous, an unnecessary distraction from the challenges and opportunities of private life. In the USA, for instance, less than half of those eligible turn out to vote in national elections. To the poor, democracy can seem irrelevant, a woefully inadequate antidote to the poisonous power of poverty, violence, and political corruption. Over the past two centuries, the democratic vision has survived defeat and outlived its historic enemies; if it is to survive and flourish in the new century, it must recapture – and in some places, attain for the first time – its hold on ordinary men and women's political imagination. In the twenty-first century the additional issue is whether democracy can adapt to theocratic conceptions of the state and civil society, which are seen to be a significant challenge to democratic ideals in many parts of the world. In the aftermath of failed democratic revolutions in the Arab world and elsewhere, the question of democracy as a global form of government in once again being questioned.

See also: Citizen Participation; Democracy: Normative Theory; Democratic Theory; Democratic Transitions; Political Parties, History of; Political Representation; Politics and Participation.

Bibliography

Anderson, M., 1996. Frontiers. Polity Press, Cambridge, UK.
Birch, A.H., 2002. The Concepts and Theories of Modern Democracy. Routledge, London.
Blokland, H.T., 2011. Pluralism, Democracy and Political Knowledge: Robert a Dahl and His Critics on Modern Politics. Ashgate, Burlington, VT.
Bobbio, N., 1989. Democracy and Dictatorship: The Nature and Limits of State Power. Polity Press, Cambridge, UK.
Bryce, J., 1921. Modern Democracies. Macmillan, New York.
Dahl, R.A., 1989. Democracy and its Critics. Yale University Press, New Haven, CT.
Duverger, M., 1974. Modern Democracies. Dryden Press, Hinsdale, IL.
Finley, M.I., 1985. Democracy Ancient and Modern. Rutgers University Press, New Brunswick, NJ.
Gallie, W.B., 1955–56. Essentially contested concepts. Proceedings of the Aristotelian Society New Series 56, 167–198.
Gunther, R., Montero, J.R., Botella, J., 2004. Democracy in Modern Spain. Yale University Press, New Haven, CT.
Held, D., 1996. Models of Democracy, second ed. Stanford University Press, Stanford, CA.
Hinsley, F.H., 1963. Power and the Pursuit of Peace. Cambridge University Press, Cambridge, UK.
Israel, J.I., 2010. A Revolution of the Mind: Radical Enlightenment and the Intellectual Origins of Modern Democracy. Princeton University Press, Princeton, NJ.
Kornbluh, M.L., 2000. Why America Stopped Voting: the Decline of Participatory Democracy and the Emergence of Modern American Politics. New York University Press, New York.
Marshall, T.H., 1950. Citizenship and Social Class. Cambridge University Press, Cambridge, UK.
Michels, R., 2001. Political Parties a Sociological Study of the Oligarchical Tendencies of Modern Democracy. Batoche, Kitchener, ON.
Mill, J.S., 1859. A few words on non-intervention, 1873. In: Dissertations and Discussions, vol. 3. Holt, New York.
Mirsepassi, A., 2010. Democracy in Modern Iran: Islam, Culture, and Political Change. New York University Press, New York.
Ober, J., Hedrick, C. (Eds.), 1996. Demokratia. A Conversation on Democracies, Ancient and Modern. Princeton University Press, Princeton, NJ.
Ostrogorski, M., 1902. Democracy and the Organization of Political Parties. Macmillan, New York.
Putnam, R.D., 1993. Making Democracy Work. Princeton University Press, Princeton, NJ.
Rasanayagam, A., 2003. Afghanistan a Modern History: Monarchy, Despotism or Democracy?: The Problems of Governance in the Muslim Tradition. Tauris, London, IB.
Siedentop, L., 2001. Democracy in Europe. Columbia University Press, New York.
de Tocqueville, A., 1988. Democracy in America. Perennial Library, New York.
Worden, N., 2007. The Making of Modern South Africa: Conquest, Apartheid, Democracy. Blackwell Pub., Malden, MA.

Democracy: Normative Theory

Thomas D Christiano, University of Arizona, Tucson, AZ, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 5, pp. 3413–3419, © 2001, Elsevier Ltd.

Abstract

Normative democratic theory deals with the ethical foundations of democracy and democratic institutions. It does not offer in the first instance a scientific study of those societies that are called democratic. It aims to provide an account of when and why democracy is ethically desirable as well as ethical principles for guiding the design of democratic institutions. It is inherently interdisciplinary and must call on the results of political science, sociology, and economics in order to give this kind of concrete guidance. This brief outline of normative democratic theory focuses attention on three distinct issues in recent work. First, it outlines some different approaches to the question of why democracy is morally desirable at all. Second, it explores the problem of the rationality of participation in large democratic societies given the inevitably small impact of participants on the outcomes. It also discusses blueprints of democratic institutions for dealing with this. Third, it surveys different accounts of the proper characterization of equality in the processes of representation. The last two parts display the interdisciplinary nature of normative democratic theory.

Foundational Questions

To fix ideas, the term 'democracy' refers very generally to a method of group decision making characterized by a kind of equality of input among the participants. That equality may be the mere formal equality of one person, one vote in an election for representatives to an assembly, or it may be more robust, including equality in the processes of deliberation and coalition building. 'Democracy' may refer to any of these political arrangements. Normative democratic theorists attempt to determine which one of these kinds of arrangements is desirable, ethically speaking, and in what contexts. For instance, Joseph Schumpeter (1956) argues, with some force, that the first highly formal kind of democracy is highly desirable while the second is dangerous. On the other hand, Jean-Jacques Rousseau (1967) is apt to argue that the first kind of democracy is akin to slavery while only robustly egalitarian democracies have political legitimacy. To evaluate their arguments we must decide on the merits of the different principles and conceptions of humanity and society from which they proceed.

The Justification of Democracy

We can evaluate democracy ethically along at least two different dimensions. We can evaluate a group decision-making method in terms of the outcomes of using it and compare the values of those outcomes with those of other possible methods of political decision making. We can also evaluate democracy in terms of qualities that are inherent in the method itself. We can ask whether there is something inherently fair about making democratic decisions on matters on which people disagree.

Instrumental Values

The instrumental values often attributed to democracy can be broadly classed into two categories: the quality of the resulting legislation and the impact of democracy on the characters of the participants. John Stuart Mill (1991) argued that democratic ways of deciding tend to be superior to nondemocratic ways on account of the fact that they produce superior legislation in three distinct ways. Strategically, democracy is thought to produce superior outcomes to other forms of rule just because it forces decision makers to take into account the interests, rights and opinions of most people in society. Since democracy gives some political power to each, more people are taken into account than under aristocracy or monarchy.

Epistemologically, democracy is thought to be the best decision-making method on the grounds that it is generally more reliable in helping participants discover the right decisions. Since democracy brings a lot of people into the process of decision making, it can take advantage of many sources of information and critical assessment of laws and policies. Democratic decision making tends to be more informed than other forms about the interests of citizens and the causal mechanisms necessary to advance those interests. Furthermore, the broad-based discussion typical of democracy enhances the critical assessment of the different moral ideas that guide decision makers.

Regarding its effects on character, many have noted with Mill and Rousseau that democracy tends to make people stand up for themselves more because it makes collective decisions depend on them more than monarchy or aristocracy do. Hence, in democratic societies individuals are encouraged to be more autonomous. In addition, democracy tends to get people to think carefully and rationally more than other forms of rule because it makes a difference whether they do or not. Finally, some have argued that democracy tends to enhance the moral qualities of citizens. When they participate in making decisions, they have to listen to others, they are called upon to justify themselves to others and they are forced to think in part in terms of the interests of others. Some have argued that when people find themselves in this kind of circumstance, they come genuinely to think in terms of the common good and justice. Hence some have argued that democratic processes tend to enhance the autonomy, rationality, and morality of participants. Since these beneficial effects are thought to be

worthwhile in themselves, they count in favor of democracy and against other forms of rule.

Some argue in addition that the above effects on character also tend to enhance the quality of legislation. A society of rational, autonomous, and moral decision makers is more likely to produce good legislation than a society ruled by a self-centered person or small group of persons who rule over slavish and unreflective subjects.

More detailed knowledge of the effects of political institutions can be used to discriminate in favor of particular kinds of democratic institutions or modifications of them. For instance, in the United States, James Madison (1788) argued in favor of a fairly strong federal government on the grounds that local governments are more likely to be oppressive to minorities. Of course, the soundness of any of the above arguments depends on the truth or validity of the associated substantive views about justice and the common good as well as a causal theory of the consequences of different institutions.

It is also the case that many theorists argue against any kind of democracy on instrumental grounds. Hobbes (1968) argues that democratic institutions are inferior to monarchical government on the grounds that they are more likely to breed destabilizing dissension among subjects. Plato (1974) argues that democracy is inferior to various forms of monarchy, aristocracy, and even oligarchy on the grounds that democracy tends to undermine the expertise necessary to properly governed societies. Some recent thinkers have argued in favor of giving over complete control over society to the market, on the grounds that democracy tends to produce serious economic inefficiencies. More modest versions of these arguments have been used to justify modification of democratic institutions.

Instrumentalism

Instrumentalists argue that these instrumental values of the democratic process are the only bases on which to evaluate democracy or compare it with other forms of political decision making. The standards one uses depend on one's conception of a just and good society. For example, libertarian instrumentalists such as Friedrich Hayek (1960) argue that democracy is desirable only to the extent that it protects liberty and private property.

Noninstrumental Values

Few theorists deny that political institutions must be at least in part evaluated in terms of the outcomes of having those institutions. Some argue, in addition, that some forms of decision making are morally desirable independent of the consequences of having them. A variety of different approaches have been used to show that democracy has this kind of intrinsic value. The most common of these come broadly under the rubrics of liberty and equality.

Liberty

Some argue that the basic principles of democracy are founded in the idea that each individual has a right to liberty. Democracy, it is said, extends the idea that each ought to be master of his or her life to the domain of collective decision making. First, each person's life is deeply affected by the larger social, legal, and cultural environment in which he or she lives. Second, only when each person has an equal voice and vote in the process of collective decision making will each have control over this larger environment. Thinkers such as Carol Gould (1988) conclude that only when some kind of democracy is implemented will individuals have a chance at self-government. Since individuals have a right to self-government, they have a right to democratic participation. This right is established at least partly independently of the worth of the outcomes of democratic decision making. The idea is that the right to self-government gives one a right, within limits, to do wrong. Just as an individual has a right to make some bad decisions for himself or herself, so a group of individuals have a right to make bad or unjust decisions for themselves. Here we can see the makings of an argument against instrumentalism. To the extent that an instrumentalist wishes to diminish a person's power to contribute to the democratic process for the sake of enhancing the quality of decisions, he or she is committed to thinking that there is no moral loss in the fact that our power has been diminished; but if the liberty argument is correct our right to control our lives is violated by this.

One major difficulty with this line of argument is that it appears to require that the basic rule of decision making be consensus or unanimity. If each person must freely choose the outcomes that bind him or her, then those who oppose the decision are not self-governing. They live in an environment imposed on them by others. So only when all agree to a decision are they freely adopting the decision.

The trouble is that there is rarely agreement on major issues in politics. Indeed, it appears that one of the main reasons for having political decision-making procedures is that they can settle matters despite disagreement. Hence it is hard to see how any political decision-making method can respect everyone's liberty.

Democracy as Public Justification

One distant relative of the self-government approach is the account of democracy as a process of public justification defended by, among others, Joshua Cohen (1998). The idea behind this approach is that laws and policies are legitimate to the extent that they are publicly justified to the citizens of the community. Public justification is justification to each citizen as a result of free and reasoned debate among equals. Citizens justify laws and policies to each other on the basis of mutually acceptable reasons. Democracy, properly understood, is the context in which individuals freely engage in a process of reasoned discussion and deliberation on an equal footing. The ideas of freedom and equality provide guidelines for structuring democratic institutions.

The aim of democracy as public justification is reasoned consensus among citizens. However, a serious problem arises when we ask about what happens when disagreement remains. Two possible replies have been suggested to this kind of worry. It has been urged that forms of consensus weaker than full

consensus are sufficient for public justification and that the weaker varieties are achievable in many societies. For instance, there may be consensus on the list of reasons that are acceptable publicly but disagreement on the weight of the different reasons. Or there may be agreement on general reasons but disagreement about particular interpretations of those reasons. What would have to be shown here is that such weak consensus is achievable in many societies and that the disagreements that remain are not incompatible with the ideal of public justification.

Equality

Many democratic theorists have argued that democracy is a way of treating persons as equals when there is good reason to impose some kind of organization on their shared lives, but they disagree about how best to do it. On one version, defended by Peter Singer (1973), when people insist on different ways of arranging matters properly, each person in a sense claims a right to be dictator over their shared lives. However, these claims to dictatorship cannot all hold up. Democracy embodies a kind of peaceful and fair compromise among these conflicting claims to rule. Each compromises equally on what he or she claims as long as the others do, resulting in each having an equal say over decision making. In effect, democratic decision making respects each person's point of view on matters of common concern by giving each an equal say about what to do in cases of disagreement.

One difficulty is that this view relies on agreement much as the liberty views described above. What if people disagree on the democratic method or on the particular form democracy is to take? Are we to decide these latter questions by means of a higher order procedure? And if there is disagreement on the higher order procedure, must we also democratically decide that question? The view seems to lead to an infinite regress. Another egalitarian defense democracy asserts that it embodies the equal consideration of the interests of the citizens of a society when there is disagreement about how best to organize their shared life. On this account, equal consideration of interests requires that individuals' judgments be taken into account equal when there is disagreement. Respect for each citizen's judgment is grounded in the fallibility of everyone's judgment as well as the generally recognized tendency of the moral judgments of persons to be biased toward their own interests. Each citizen would have good reason to think that his or her interests were not being given the same weight as others if he or she had less decision-making power than the others. On this account, since each person has a preeminent interest in being recognized as an equal member of the community, only equality in decision-making power is compatible with equal consideration of interests. The principle of equal consideration of interests also implies limits to what can be up for democratic control and so the infinite regress noted above is avoided.

Democratic Citizenship

A vexing problem of democratic theory has been to determine whether ordinary citizens are up to the task of governing a large society. There are three distinct problems here. First, Plato argued that some people are innately more intelligent than others and that those persons ought to rule. Second, others have argued that a society must have a division of labor. If everyone were engaged in the complex and difficult task of politics, little time or energy would be left for the other essential tasks of a society. Conversely, if we expect most people to engage in other difficult and complex tasks, how can we expect them to have the time and resources sufficient to devote themselves intelligently to politics?

Third, since individuals have so little impact on the outcomes of political decision making in large societies, they have little sense of responsibility for the outcomes. Some have argued that it is not rational to vote since the chances that a vote will affect the outcome of an election are nearly indistinguishable from zero. Worse still, Anthony Downs (1957) has argued that almost all of those who do vote have little reason to become informed about how best to vote. On the assumption that citizens reason and behave roughly according to the Downsian model, either the society must in fact be run by a relatively small group of people with minimal input from the rest or it will be very poorly run.

These observations pose challenges for any robustly egalitarian or deliberative conception of democracy. Without the ability to participate intelligently in politics one cannot use one's votes to advance one's aims, nor can one be said to participate in a process of reasoned deliberation among equals. So, either equality of political power implies a kind of self-defeating equal participation of citizens in politics or a reasonable division of labor seems to undermine equality of power.

Plato and Hobbes argued against democracy in any form on these grounds. Some modern theorists of democracy, called elite theorists, have argued against any robustly egalitarian forms of democracy on these grounds. They argue that high levels of citizen participation tend to produce bad legislation designed by demagogues to appeal to poorly informed and overly emotional citizens. They look upon the alleged apathy and uninformedness of citizens evidenced in many empirical studies in the 1950s and 1960s as highly desirable social phenomena. Joseph Schumpeter's assertion that the "democratic method is that institutional arrangement for arriving at political decisions in which individuals acquire the power to decide by means of a competitive struggle for the people's vote" (Schumpeter, 1956: p. 269) still stands as a concise statement of the elitist view. In this view, the emphasis is placed on responsible political leadership. Political leaders are to avoid divisive and emotionally charged issues and make policy and law with little regard for the fickle and diffuse demands made by ordinary citizens.

An approach that is in part motivated by the concerns above but that attempts to avoid elitism is the interest group pluralist account of politics. Robert Dahl's early statement of the view is very powerful. "In a rough sense, the essence of all competitive politics is bribery of the electorate by politicians… The farmer… supports a candidate committed to high price supports, the businessman… supports an advocate of low corporation taxes… the consumer… votes for candidates opposed to a sale tax" (Dahl, 1959: p. 69). In this conception of the democratic process, each citizen is a member of an interest group with

narrowly defined interests that are closely connected to their everyday lives. On these subjects citizens are supposed to be fairly well informed and interested in having an influence. On this account, democracy is not rule by the majority but rather rule by coalitions of minorities.

The account ensures that individuals can participate roughly as equals to the extent that it narrowly confines the issues with which each individual is concerned.

A third approach inspired by the kinds of problems above may be called the neoliberal approach to politics favored by public choice theorists such as James Buchanan and Gordon Tullock (1965). Against elite theories, they contend that elites and their allies will tend to expand the powers of government and bureaucracy for their own interests and that this expansion will occur at the expense of a largely inattentive public. They argue against the interest group pluralist theorists that the problem of participation occurs within interest groups more or less as powerfully as among the citizenry at large. Only some interest groups are likely to succeed in organizing to influence the government. Hence only some interest groups will succeed in influencing government and they will do so largely for their own benefit while spreading the costs to others.

Neoliberals argue that any way of organizing a large and powerful democratic state is likely to produce serious inefficiencies. They infer that one ought to transfer many of the current functions of the state to the economic market and limit the state to the enforcement of basic property rights and liberties. These can be more easily understood and brought under the control of ordinary citizens.

But the neoliberal account of democracy must answer to two large worries. First, citizens in modern societies have more ambitious conceptions of social justice and the common good than are realizable by the minimal state. The neoliberal account thus implies a very serious curtailment of democracy of its own. More evidence is needed to support the contention that these aspirations cannot be achieved by the modern state. Second, the neoliberal approach ignores the problem of large private concentrations of wealth and power that are capable of pushing small states around for their own benefit and imposing their wills on populations without their consent. The assumptions that lead neoliberals to be skeptical about the large modern state imply equally disturbing problems for a neoliberal society.

A considerable amount of the literature in political science and the economic theory of the state are grounded in the assumption that individuals act primarily and perhaps even exclusively in their self-interest narrowly construed. The problem of participation and the accounts of the democratic process described above are in large part dependent on this assumption. While these ideas have generated interesting results and have become ever more sophisticated, there has been a growing chorus of opponents. Against the self-interest axiom, defenders of deliberative democracy and others claim that citizens are capable of being motivated by a concern for the common good and justice. They also claim, with Mill and Rousseau, that such concerns are not merely given prior to politics but that they can evolve and improve through the process of discussion and debate in politics. They assert that much debate and discussion in politics would not be intelligible were it not for the fact that citizens are willing to engage in open-minded discussion with those who have distinct morally informed points of view. Empirical evidence suggests that individuals are motivated by moral considerations in politics in addition to their interests. Accordingly, many propose that democratic institutions be designed to support the inclination to engage in moral and open-minded discussion with others.

Once we approach the idea of citizenship from a moral point of view and we recognize the importance of a division of labor, the question arises of what the appropriate role for a citizen in a democracy is. If we think that citizens are too often uninformed, we should ask two questions. What ought citizens have knowledge about in order to fulfill their role? What standards ought citizens' beliefs live up to in order to be adequately supported? Some have proposed that citizens know about their particular sectors of society and not others. We have seen that this view has a number of difficulties. Christiano (1996) proposes, along with others, that citizens must think about what ends the society ought to aim at and leave the question of how to achieve those aims to experts. This kind of view needs to answer to the problem of how to ensure that politicians, administrators, and experts actually do attempt to realize the aims set by citizens. And it must show how institutions can be designed so as to establish the division of labor while preserving equality among citizens. But if citizens genuinely do choose the aims and others faithfully pursue the means to achieving those aims, then citizens are in the driver's seat in society.

It is hard to see how citizens can satisfy any even moderate standards for beliefs about how best to achieve their political aims. Knowledge of means requires an immense amount of social science and knowledge of particular facts. For citizens to have this kind of knowledge generally would require that we abandon the division of labor in society. On the other hand, citizens do have first-hand and daily experience with thinking about the values and aims they pursue. This gives them a chance to satisfy standards of belief regarding what the best aims are. Still the view is not defensible without a compelling institutional answer to the question of how to ensure that others are genuinely pursuing the means to achieve the aims specified by citizens.

Legislative Institutions

A number of debates have centered on the question of what kinds of legislative institution are best for a democratic society. What choice we make here will depend heavily on the kind of underlying ethical justification of democracy, our conception of citizenship, and our empirical understanding of political institutions and how they function. The most basic types of formal political representation available are single-member district representation, proportional representation, and group representation. In addition, many societies have opted for multicameral legislative institutions. In some cases, combinations of the above forms have been tried.

Single-member district representation returns single representatives of geographically defined areas containing roughly equal populations to the legislature and is present most prominently in the USA and the UK. The most common form of proportional representation is party-list proportional representation. In a simple form of such a scheme, a number of

parties compete for election to a legislature that is not divided into geographical districts. Parties acquire seats in the legislature as a proportion of the total number of votes they receive in the population as a whole. Group representation occurs when the society is divided into nongeographically defined groups such as ethnic or linguistic groups or even functional groups such as workers, farmers, and capitalists and returns representatives to a legislature from each of them.

Many have argued in favor of single-member district legislation on the grounds that it has appeared to them to lead to more stable government than other forms of representation. The thought is that proportional representation tends to fragment the citizenry into opposing homogeneous camps that adhere rigidly to their party lines and that are continually vying for control over the government. Since there are many parties and they are unwilling to compromise with each other, governments formed from coalitions of parties tend to fall apart rather quickly. The post-War experience of governments in Italy appears to confirm this hypothesis. Single-member district representation, in contrast, is said to enhance the stability of governments by virtue of its favoring a two-party system of government. Each election cycle then determines which party is to stay in power for some length of time.

Charles Beitz (1989) argues that single-member district representation encourages moderation in party programs offered for citizens to consider. This results from the tendency of this kind of representation toward two-party systems. In a two-party system with majority rule, it is argued, each party must appeal to the median voter in the political spectrum. Hence they must moderate their programs to appeal to the median voter. Furthermore, they encourage compromise among groups since they must try to appeal to a lot of other groups in order to become part of one of the two leading parties. These tendencies encourage moderation and compromise in citizens to the extent that political parties, and interest groups, hold these qualities up as necessary to functioning well in a democracy.

In criticism, advocates of proportional and group representation have argued that single-member district representation tends to muffle the voices and ignore the interests of minority groups in the society. Minority interests and views tend to be articulated in background negotiations and in ways that muffle their distinctiveness. Furthermore, representatives of minority interests and views often have a difficult time becoming elected at all in single-member district systems, so it has been charged that minority views and interests are often systematically underrepresented. Sometimes these problems are dealt with by redrawing the boundaries of districts in a way that ensures greater minority representation. The efforts are invariably controversial since there is considerable disagreement about the criteria for apportionment. In proportional representation, by contrast, representatives of different groups are seated in the legislature in proportion to citizens' choices. Minorities need not make their demands conform to the basic dichotomy of views and interests that characterize single-member district systems so their views are more articulated and distinctive as well as better represented.

Another criticism of single-member district representation is that it encourages parties to pursue dubious electoral campaign strategies. The need to appeal to a large, diverse, and somewhat amorphous sector of the population can very often be best met by using ambiguous, vague, and often irrelevant appeals to the citizens. Thus, instead of encouraging reasonable compromise, the scheme tends to support tendencies toward ignorance, superficiality, and fatuousness in political campaigns and in the citizenry. It encourages political leaders to take care of the real issues of politics in back rooms while they appeal to citizens by means of smoke and mirrors. Of course, those who agree in the main with the elitist-type theories will see nothing wrong in this; indeed, they may well champion this effect. Proportional representation requires that parties be relatively clear and up front about their proposals, so those who believe that democracy is ethically grounded in the appeal to equality tend to favor proportional representation.

Advocates of group representation such as Iris Marion Young (1993) have argued that some historically disenfranchised groups may still not do very well under proportional representation. They may not be able to organize and articulate their views as easily as other groups. Also, minority groups can still be systematically defeated in the legislature and their interests may be consistently set back even if they do have some representation. For these groups, some have argued that the only way to protect their interests is legally to ensure that they have adequate and even disproportionate representation.

One worry about group representation is that it tends to freeze some aspects of the agenda that might be better left to the choice of citizens. For instance, consider a population that is divided into linguistic groups for a long time, and suppose that only some citizens continue to think of linguistic conflict as important. In the circumstances, such schemes may tend to be biased in an arbitrary way that favors the views or interests of those who do think of linguistic conflict as important.

See also: Appeasement: Political; Law and Democracy; Politics and Participation.

Bibliography

Beitz, C., 1989. Political Equality: An Essay on Democratic Theory. Princeton University Press, Princeton, NJ.
Buchanan, J., Tullock, G., 1965. The Calculus of Consent: Logical Foundations of Constitutional Democracy. University of Michigan Press, Ann Arbor, MI.
Christiano, T., 1996. The Rule of the Many: Fundamental Issues in Democratic Theory. Westview Press, Boulder, CO.
Cohen, J., 1998. Deliberation and democratic legitimacy. In: Bohman, J., Rehg, W. (Eds.), Deliberative Democracy. MIT Press, Cambridge, MA.
Dahl, R., 1959. A Preface to Democratic Theory. University of Chicago Press, Chicago.
Downs, A., 1957. An Economic Theory of Democracy. Harper and Row, New York.
Gould, C., 1988. Rethinking Democracy: Freedom and Social Cooperation in Politics, Economics and Society. Cambridge University Press, New York.
Hayek, F., 1960. The Constitution of Liberty. University of Chicago Press, Chicago.
Hobbes, T., 1968. In: MacPherson, C.B. (Ed.), Leviathan. Penguin-Harmondsworth, UK.
Madison, J., 1788. The federalist no 10. In: Hamilton, A., Jay, J., Madison, J. (Eds.), The Federalist Papers. Random House, New York, pp. 59–62.
Mill, J.S., 1991. Considerations on Representative Government. Prometheus, Buffalo, NY.
Plato, 1974. The Republic (Lee D. Penguin, Revised Trans.), second ed. Harmondsworth, UK.
Rousseau, J.-J., 1967. The Social Contract and Discourses (G.H.D. Cole, Trans.). Dent, London.
Schumpeter, J., 1956. Capitalism, Socialism and Democracy. Harper and Row, New York.
Singer, P., 1973. Democracy and Disobedience. Oxford University Press, Oxford, UK.
Young, I.M., 1993. Justice and the Politics of Difference. Princeton University Press, Princeton, NJ.

Democratic Theory

Frank Cunningham, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S. Fishkin, volume 5, pp. 3422–3425, © 2001, Elsevier Ltd.

Abstract

The responses to core questions of democracy by classic theorists, including Aristotle, John Locke, Jean-Jacques Rousseau, Alexis de Tocqueville, John Stuart Mill, and Joseph Schumpeter, are surveyed by way of introduction to contemporary theories that draw on them. These include participatory democracy, catallaxy and collective choice theories, deliberative democracy, power-political pluralism, and radical-democratic pluralism. Also discussed is a selection of persisting problems addressed by democratic theorists.

The two generic questions to which answers are either given or presupposed by all democratic theorists are: What is democracy? and Why (if at all) ought democracy to be valued? The responses of Aristotle (384–322 BC), in this as in many other domains of Western theorizing, usefully set the stage for summaries of the main alternatives.

Aristotle

With typical attention to nuance, Aristotle (1986[*c*.320 BC]) distinguishes between ideal conceptions of how a political society might be governed and ways that they are governed in actual practice. In either case, rule might be in the hands of one person, of a few people, or of many people. In the ideal world, the latter situation is called by him a 'polity' to be distinguished from a 'royalty' or rule by one and rule by a few, called 'aristocracy.' What makes all of these conceptions ideal is that the one, the few, or the many rule with a view to promoting the common good of the entire society. Such rule is almost never attained in actuality, where it is exercised instead in accord with the specific interests of the rulers. Rule by and for one person is 'tyranny' and by a few people, 'oligarchy.' Aristotle reserves the term 'democracy' to refer to the remaining category, where the many rule in their own interests.

As to his evaluations, Aristotle holds that the best rule in an ideal world would be royalty, followed by aristocracy, and then polity. In the real world, however, Aristotle anticipated by over two millennia the often-cited opinion of Winston Churchill that democracy is the worst form of government except for all the others. Aristotle's main ground for viewing democracy as the most tolerable realistic choice is that at least the largest number of people profit from its self-serving rule. Also, there is something to be gained from pooling many people's experiences in formulating policies, and the dangers of majority discontent confronted by undemocratic societies are averted.

Throughout the history of democratic theory, the boundary between it and political-scientific considerations is often blurred. Indeed, a case can be made that each of political science and political theory gains from their interaction. Aristotle's distinction between real and ideal rule is a case in point. His view that ideal rule is rarely enacted is based not on speculation but on study of actual political societies in the ancient world, where Aristotle saw very few approximations in practice: royalty during the reign of certain generals in Sparta, polity for a time after the rule of Solon in Athens. Another example is Aristotle's equation of rule of the many with rule of the poor. While admitting of theoretical explanation for him, this assimilation is largely an empirical claim. Also empirical are his views about the preconditions for a stable and virtuous citizenry. These include a large middle class, a universal – for boys – public educational system, and an appropriate size.

Locke and Rousseau

Historicistic accounts of democratic theories see them as expressing, rationalizing, or otherwise being embedded in the social and economic circumstances of their times. An example is C.B. Macpherson's interpretation of British political theorizing in the seventeenth century as reflection of and support for the nascent capitalism of that time (2011[1962]). This article abstracts from such historical accounts while recognizing that there is much to be learned from them. Still, it should be noted that democratic theorists themselves differ in how much attention they pay to actual practices called democratic.

Virtually all approaches to democratic theory address all three of the meaning of democracy, the value of democracy, and actual democratic practices, but they differ depending on which of these is the lead consideration. Aristotle generalizes a view about the meaning of democracy from his surveys of democratic practices, which also motivate his evaluations of its various forms. An alternative strategy is to begin with the value of democracy. This approach is illustrated in the political theory of John Locke (1632–1704). Like other political philosophers in the 'social contract' tradition, Locke is primarily concerned to identify the conditions that make political institutions and practices legitimate.

Political society for the social contract theorists, in modern times stretching from Thomas Hobbes through John Rawls, should be thought of as the result (or as if it were the result) of people in a prepolitical state of nature agreeing to transfer some or all of their individual powers to a political authority. To the extent that the authority comports itself in accord with the original intentions of the contractors, it is legitimate. Thus for Hobbes (1588–1679), the intent was to preserve the peace, and

failure to do so challenges the legitimacy of a sovereign power (1968[1651]).

The intent of contractors in a state of nature for Locke is to preserve their individual freedoms along with their lives and property. Unlike Hobbes, for whom democratic or autocratic forms of political authority were a matter of choice for his entirely self-interested contractors (his recommendation being for monarchal autocracy), Locke maintained that only a certain form of democracy could be legitimate, namely majority rule issuing in representative government. The majority embodies the will of the entire community, and the representatives it selects to govern are charged with acting in accord with this will, which includes preservation of individual liberty, life, and property. Thus for Locke, the value of democracy is its protection of these things, and democracy itself is majority rule and representative government (1962[1690]).

The approach of Jean-Jacques Rousseau (1712–78) illustrates how following a similar methodology by different theorists can lead to alternative or even opposed viewpoints on democracy. Like Locke, Rousseau appeals to a social contract. He shares Hobbes' conception of original contractors as self-interested, but he maintains that the very act of coming together to seek consensus on a desirable form of political society changes people's nature to ones valuing collective goods (1968[1762]). The resulting prescriptions regarding political society differ from Locke's in two major respects.

For Locke, some matters must be outside of governmental control, in particular those things that impact individual freedom, life, or property. By contrast, for Rousseau, all aspects of human interaction are subject to the power of the body politic. What is supposed to keep this from being oppressive is that its members, in virtue of the character transformation they have undergone, will act to promote the common good, and, in a second departure from Locke, their power is not exercised by representatives but directly by the populace itself.

Like Aristotle, Rousseau recognized that this is an ideal seldom fully realized. Thus, he employed the term 'democracy' to refer to a situation only approximating ideal self-rule. This is representative democracy where representatives are strictly delegated by a majority that, for its part, is motivated by concern for common interests (famously called by him 'the general will'). Such rule is possible in relatively small communities where legislative officials are selected by lot. Rousseau's views are almost always referred to by subsequent defenders of what came from around 1960 to be called 'participatory democracy.'

In its strongest form, participatory democracy merges with anarchism, where a self-governing population dispenses with government altogether. In a weaker form, participatory democrats can accommodate representative government, provided representatives are strictly delegated by the citizens who select them and as many of the functions of a state as possible are carried out directly by citizens (Pateman, 1970; Mansbridge, 1983; Barber, 1984). One value claimed for participatory democracy is that it draws on the expertise of ground-level citizen experiences in formulating and executing policies. Participatory democrats also echo a 'civic republican' theme in Rousseau's view about the character-building dimension of the social contract. This is the hypothesis that collective experience in self-government promotes socially virtuous citizenship. For modern-day civic republicans, this is important for promoting values that both respect individual autonomy and encourage the pursuit of common goods (Maynor, 2003).

de Tocqueville and Mill

In an extended study tour of America in the early 1830s, the French aristocrat Alexis de Tocqueville (1805–59) came to appreciate the vibrancy of that early democracy, which he attributed in large part to the experiences of its citizens in local, town governance unconstrained by European traditional conservatism and inequalities. His conception of democracy is an example of one focused on the literal meaning of the word, namely 'rule by the people' or 'popular sovereignty.' "The people," he concluded, "reign over the American political world as God reigns over the universe" (1969[1835–40]: p. 60).

The advantages de Tocqueville saw in this democracy are not unlike those extolled by participatory democrats – an engaged and energized population making valuable contributions to governance. However, not sharing the Rousseauean image of a homogenous public, he identified as a major problem for the American or any other democracy that a minority might be oppressed by a majority, and he coined the phrase 'the tyranny of the majority' to describe this possibility. Majoritarian democracy not only makes such oppression possible but also disguises it because the majority, as Locke claimed, is supposed to embody the legitimate will of the entire society. This, in turn, is facilitated by the fact that while rulers in a monarchy are identifiable individuals, in a democracy, ultimate rule is supposed to be exercised by the abstract entity of 'the people.'

Mill

John Stuart Mill (1806–73) furthered Locke's defense of representative democracy, though he thought that it should be supplemented when feasible by opportunities for direct citizen participation. Mill also endorsed de Tocqueville's warnings about majority tyranny. Not only can a majority oppress a minority but it can also suppress freedom of speech and the like in ways that rebound negatively on democracy itself. Mill's solution was to prescribe combining representative democracy with legally binding protections of individual liberties, thus providing what is generally taken to be the classic expression of the theory of liberal democracy (1991[1861]).

As to the value of democracy, one finds two considerations. As a utilitarian in ethics (2002[1863]), Mill saw democracy as aggregating preferences in such a way as to yield the greatest amount of citizen satisfaction. Additionally, in connection with his endorsement of direct participation, Mill valued democracy for developing people's personally enriching and socially beneficial potentials. His nuanced deployment of utilitarian theory makes these two goals of democracy compatible in his eyes, but subsequent theorists typically divide on whether democracy can or should achieve more than the aggregation of preferences.

Schumpeter

Joseph Schumpeter (1883–1950) is often credited with inaugurating the self-described 'realist' school of democratic theory. He took the strategy of generalizing a theory of democracy from the example of existing practices called democratic to its limit. If one looks at such practices, he held, one sees that they mainly involve politicians and political parties competing with one another for the popular vote. This, Schumpeter concluded, is all that democracy is (1962[1942]: p. 269). He took the idea of popular sovereignty to be a myth because even people voting for the same candidates often do so for different reasons. Similarly, he denied that democracy could be construed as promoting the common good, since for him there is no such thing.

The principal value of democracy for Schumpeter is that leaders cannot act in entirely irresponsible ways, depending on elections to remain in power. Contrary to participationists, he thought citizen engagement in governing should be minimized so as not to hamper professional government bureaucracies.

Contemporary Variations

Main theories of democracy today present a large array of approaches to the primary questions of democratic theory (surveyed in Cunningham (2002) and Held (1996)), but in their basic orientations, they are all variations or combinations of the classic alternatives. For example, 'associative democracy,' where voluntary self-governing organizations – of workers, consumers, producers, social movements, and the like – have special access to a central government or wherever possible replace it, is informed by the values of participatory democracy (Hirst, 1994). Macpherson strives to combine liberal and participatory democracy guided by neo-Aristotelian values and a Marx-like critique of capitalism (2012: Chapter 5). Michael Sandel draws on the civic republican dimension of the thought of Rousseau and others to advocate an anti-individualistic, communitarian version of democracy (1996). It is also worth noting that while not seeing the new electronic media as offering a unique form of democracy, many now look to it to enhance democratic practices and institutions variously conceived (see the contributions to Jenkins and Thorburn, 2004).

Two current approaches stand out for drawing the largest contemporary following – catallaxy – especially as deployed by collective choice theories and deliberative democracy.

Catallaxy

Also in the mode of thinking of Schumpeter was a fellow Austrian, Friedich von Hayek (1899–1992), who popularized the idea of 'catallaxy' or the application of economic methods to the study of politics (1976: Chapter 10). The views of Schumpeter and von Hayek found their most enthusiastic proponents among certain theorists of democracy in the United States. Anthony Downs took as his main thesis that parties in democratic politics are primarily concerned to formulate "whatever policies they believe will gain the most votes, just as entrepreneurs produce whatever product they believe will gain the most profits." (1957: p. 295).

Primary analytic tools of those in this tradition are taken from rational choice theory in which it is ascertained what means, given available resources, it is rational for individuals to take in realizing their goals, where the goals are not themselves evaluated. The species of the approach variously called social or public or collective choice theory employs this method in situations where collective decisions are being made such as voting to select a government. (Useful summaries are by Harden (1982) and Ordeshook (1992).)

Early deployment of these methods challenged some core tenets of democratic theory, as in the conclusion of James Buchanan and Gordon Tullock, that sometimes majority rule is an irrational decision procedure (1962). It also posed seemingly intractable problems for democratic decision-making. At the core of some voter 'paradoxes' described by John Kenneth Arrow (1951, and Riker (1982) is the possibility of stalemate due to a cyclical majority when more than two choices confront voters.

Downs addresses the problem of voter abstention not as a matter of figuring out how to get people to the polls but as a problem of collective rationality. Because the likelihood of any one vote making a crucial difference is so low and because there are personal costs to taking the time and effort to vote, the rational citizen will not vote at all. Subsequent collective choice theorists have drawn on increasingly sophisticated rational choice tools, such as game theory, to explicate democratic decision-making and flag problems it confronts. (An example of very many is Strom, 1992.) Insofar as theory in this mode has practical application, it is to analyze and recommend legislative and voting models and methods. These include questions about the relative superiority of presidential or parliamentary systems of government, the virtues or otherwise of multiparty systems, first-past-the-post voting rules versus proportional representation, ranked preference voting, runoff elections, and so on.

Like almost all other liberal-democratic theorists, those in this tradition think of democracy exclusively in formal representative-democratic terms. Activities within religious organizations, social movements, interest groups, or other organizations of civil society are seldom, if ever, addressed. Formally structured majority voting is seen as the core of democracy and is to be valued (if it can overcome the voter paradoxes) as a way of making collective decisions. In this, the catallactic theorists differ from participatory democrats, many of whom agree with Benjamin Barber's view of voting as among "the thinnest features of democracy" and majority rule as "a tribute to the failure of democracy" (1984: p. 198).

Deliberative Democracy

Deliberative democrats also strongly resist an exclusive focus on voting and simple majority rule. As carried out in actual democratic politics, they see such behavior as deeply flawed. Democracy on their explicitly normative orientation ought centrally to encourage reasoned discussion and debate on the part of democratic decision makers in an open-minded spirit and equally possessing the information and other resources

productively to enter into deliberative forums before voting, if voting is needed at all. They do not deny that there are differences of interest and opinion among people but hold, in the optimistic manner of Rousseau, that given the opportunity to freely and equally discuss policy alternatives, they will sincerely strive to reach consensus.

Among the forums for deliberation are legislatures and courts, political parties, work places, unions and professional associations, social movements, and the media (Benhabib, 1996: p. 75; Gutman and Thompson, 1996: pp. 358–359). Deliberative democrats differ on what sorts of forums are the most important. They also differ on whether they see deliberative democracy primarily as strengthened liberal democracy or as a version of participatory democracy and civic republicanism, though many agree with Jürgen Habermas (1998: pp. 295–302) that deliberative democracy can bridge the divide between these orientations.

While placing different emphases on democratic deliberation as a procedure for making collective decisions or on the outcomes of deliberation, all deliberative democrats see the value of democracy undertaken in accord with their prescriptions as superior to simply aggregating preferences by voting. It at least leads to reasoned decisions, and it has the potential to shape the values of deliberators, in the manner of Rousseau's contractors, to ones aimed at promoting the common good. Democracy pursued in this way also is supposed by them to confer legitimacy on resulting policies lacking in what is often perceived by citizens as democratically unresponsive government (Benhabib, 1996; Manin, 1987).

Power-Political Pluralism

Deliberative democrats set themselves not just against the catallactic theorists but also against a kindred school also especially strong in U.S. political-scientific circles, the 'interest group,' or 'power-political' pluralists. The aspects of actual practices in modern democracies that these theorists take as their point of departure include voting, but they primarily focus on competition among the interest groups (labor, business, advocacy groups, etc.) that they see as the major actors in civil society. Robert Dahl, one of the leading representatives of pluralism in this sense, presented its basic principles in the form of a commentary (1956) on the Federalist Paper No. 10, written by James Madison (1751–1836).

According to Madison, the new American democracy was challenged by conflicts among those with different religious and political opinions and especially by class divisions based on "a landed interest, a mercantile interests, a moneyed interest," which have "inflamed them with mutual animosity" and "impeded cooperation for their common good" (1987 [1788]: p. 124). The check and balance system of federal government is defended by Madison as a way of preventing conflicts from becoming destructive of the whole society. Later pluralists added the encouragement of interest-group overlap, the nurturing of appropriate leadership, and reform of political institutions to regulate and contain conflict (Truman, 1951; Polsby, 1963). The main virtue pluralists see in democracy is its potential to keep the peace.

Radical-Democratic Pluralism

Whether deliberative-democratic measures could better address persisting conflicts is a matter of continuing debate. However, the pluralists, themselves, are in any case less concerned about conflict than their deliberative-democratic critics, since they view conflict as the natural human condition or even as 'democracy's lifeblood' (Lipset, 1960: p. 83). These views are echoed by a later version of pluralism based not on Hobbes-like political science but on contemporary, especially French poststructuralist philosophical views.

This is the orientation of 'radical-democratic pluralism.' Claude Lefort focuses on de Tocqueville's observation that in a democracy, rule is by 'the people as such' to argue that this makes the 'space' of democratic rule empty of identifiable rulers and opens the possibility for demagogues or populists to fill the space, claiming that they represent the will of the people considered in Rousseauean fashion as an ideal, homogenous entity (1988). Conflict in this perspective has no place in the space of popular sovereignty. Radical pluralists refer to this notion in criticizing the overriding value that deliberative democrats put on seeking consensus (Mouffe, 1993).

Societies are always composed of groups standing in relations of 'agonic' conflict with others. The challenge of democracy is not just to contain this conflict within peaceful limits, as is the goal with interest-group pluralists, but to prevent social agents (called 'subject positions' in keeping with the post-structuralist view that individual identities are always group-embedded and subject to change) from justifying domination over others on the grounds that they represent the true interests of a society. Rather than trying to replace conflict with consensus, the prime task is to recognize persisting conflict but to keep the space of democratic rule open. This, the radical-democratic pluralists maintain, is best accomplished by finding ways to integrate otherwise conflict-supporting identities with commitment to liberal democracy.

Issues and Problems

Most democratic-theoretical scholarship presupposes some view about the nature and value of democracy and concerns itself with specific issues and problems, three groupings of which are noted below.

The Boundary Problem

If democracy, whatever else it is thought to be, involves the idea that people have or ought to have some input to decisions that affect them, then it faces what Dahl saw as the insolvable 'embarrassment of democracy' that no matter where one draws a boundary around those who have input to democratic decision-making, some who are affected by the result will be outside this boundary (1982: pp. 97–99). For most of the history of democratic politics, this has been the case for women and usually also for those without property, that is, for a very large majority of those subject to democratic rule. In recent centuries, these exclusions have been eliminated as a matter of law, although feminist theorists and in general critics of systemic disadvantages argue that they and other exclusions

informally persist in all modern democracies (for instance, Young, 2000).

While almost every theorist agrees that exclusions based on such criteria as gender, wealth, or race are democratically illegitimate, there is less accord regarding other boundaries. Voting age varies from jurisdiction to jurisdiction, but no matter what it is fixed at, unless newborns are to have the vote, some people will be excluded. Similarly, noncitizen immigrants typically cannot vote, as this is seen as a privilege just of citizenship; but in some cities, recent immigrants not yet national citizens are granted the municipal franchise.

The most extensive current debate about democratic boundaries concerns states. Citizens in democratic states have access to means of intervention within those they inhabit, but, especially since the era of globalization, much of what affects their lives takes place outside of these states. So the problem is whether or how democracy can be extended across state boundaries. The strongest advocates of such extension are 'cosmopolitan' theorists who recommend trans-state institutions such as the European Union (Beck, 2000; the contributions to Held and McGrew, 2000).

Of course, such institutions do not completely solve the boundary problem unless the entire world is under single democratic rule. But aside from this consideration, other theorists maintain that it is unrealistic to extend democracy beyond state borders (Dahl, 1999) or that it is even undesirable if democratic institutions and practices are to be sensitive to the needs and values specific to people in individual states (Zolo, 1997). In response, yet other theorists look to increasing cross-state democracy, whether by coordination of associations of popular or social movement among them (Walker, 1993) or by a variety of international institutions short of world government (Saward, 2000).

Preconditions for Democracy

Most democratic theorists, beginning with Aristotle, devote at least as much attention to preconditions for achieving or sustaining democracy as they do to more abstract topics.

Culture

One question has to do with whether some sort of cultural 'glue' is required to hold citizens together in a democracy, and if so, of what kind. Mill maintained that democracy requires shared nationhood (1991[1861]), while disagreeing with him was Lord John Acton, who feared that this prompts national chauvinism (1995[1862]). Echoing Machiavelli and Hobbes, Rousseau thought that any enduring political regime, including a democracy, requires some form of religion to bind together its citizens (Beiner, 2011).

The general question here is whether there are specific cultural values – national, religious, or otherwise – that must be shared by a population if it is to sustain democracy. As approached in recent democratic theory, this issue was foreshadowed in the work of de Tocqueville, who saw cultural values particularly suited to democracy in the early American republic. The theme was later revived in empirical studies by Gabriel Almond and Sidney Verba about values that provide the sort of social cohesion they saw as vital for democracy (1963). Robert Putnam coined the phrase 'social capital' to describe cohesive cultural values, maintaining that paucity in social capital currently undermines U.S. democracy (2000).

Consociationalism

Consociational democratic theorists concern themselves with societies where there are major, largely regionalized religious or national divisions within a country (for instance, between Catholics and Protestants in the Netherlands or between Flemish- and French-speaking peoples in Belgium). Rather than seeing these divisions as insuperable threats to democracy, the consociationalists maintain that they can be contained by partial segmentation of political authority among the major groups while encouraging joint authority over matters of common concern and effective dialogue among segment leaders (Lipjhart, 1977). They therefore deny that thoroughgoing social cohesion is a condition for democracy, provided there are institutional means for regulating major divisions.

Democratic Consolidation

The issue of preconditions for democracy with respect to the developing world is central to theories of 'democratic consolidation.' Consolidation theorists usually take Dahl's concept of 'polyarchy' as definitive of democracy, where polyarchies are representative democracies with minimal restrictions on who can run for office; leadership depends on a universal vote; and citizens enjoy enforced rights of expression, access to politically relevant information, and freedom to join or form political or other associations (1989: p. 233). A democracy in this sense is consolidated when it is in no danger of reverting to a nondemocratic situation, when, as consolidationists put it, democracy in a country is 'the only game in town.'

Assuming that long-standing democracies in the developed world are safely consolidated, these theorists engage in empirical research with respect to developing countries and to those emerging from Communism to ascertain where and to what extent each is progressing toward consolidated democracy. The theoretical challenge of these endeavors is to identify the necessary or, in the most ambitious studies, also sufficient conditions for consolidation. Many consolidationists put primary or even exclusive emphasis on economic development (for example, Przeworski et al., 2000), while others look to blends of economic, political, cultural, and institutional factors (Diamond, 2008; Linz and Stepan, 1996).

One theoretical question raised in evaluating consolidationism is whether, as its advocates maintain, there should be a single list of characteristics of democracy, such as the components of Dahl's polyarchy, that pertain to all democratic societies no matter what their economic, cultural, and political features and histories. A related question is whether democracy should be seen as a matter of degree such that almost any society exhibits a mixture of democratic and undemocratic elements and there is no 'cutoff' point below which a society is simply undemocratic or above which it is as democratic as it can be. Orientations at variance with the consolidationists on these are expressed, among others, by the pragmatist, John Dewey (1927), and by Ian Shapiro (1999, Ch. 3).

Economic Systems

Related to the concerns of consolidationists, but of longer heritage and more prevalent before the demise of Communism in the Soviet Union than now, are debates over broad economic structures, specifically capitalism and socialism, as candidates for being conditions for democracy. Early perspectives on this question involved blanket judgments and had ideological rings to them. One example is von Hayek's assertion that "only within [capitalism] is democracy possible. When it becomes dominated by a collectivist creed, democracy will inevitably destroy itself" (1944: pp. 69–70). A contrasting claim is V.I. Lenin's announcement that socialism is "a million times more democratic than the most democratic bourgeois republic" in virtue of its being explicitly partisan in the interests of workers, while democracy in capitalism, according to Lenin, always serves the interests only of capitalists (1963–80[1918]: pp. 248–249).

There is a tendency in the extreme expressions to see capitalist or socialist economic structures not as empirical preconditions for democracy but to make them so by definition, thus cutting off nonideological debate. Less extreme treatments explore questions of how large-scale egalitarian policies by socialist or social-democratic governments can coexist with economic markets or avoid oppressive bureaucratization (examples are Bowles and Gintis, 1986; Schweickart, 1996). In recent years, approaches to these questions have been more empirical and have focused on the effects of economic inequalities in capitalist societies on democracy (e.g., by Gilens, 2012; Piketty, 2014).

Value Pluralism

Some theorists argue that, especially in an era of global-cultural interpenetration and multiculturalism, commitment to value pluralism is an essential component of democracy. Some think that democracy is at least severely stressed by multiculturalism (Glazer, 1983), while others see multiculturalism and democracy as compatible or even mutually reinforcing (Kymlicka, 1995). This debate joins a broader one about the desirability and the feasibility of value pluralism. Value pluralists champion toleration with respect to people's differing values. This orientation pertains for the most part to specifically liberal theory and practice, but it bears on democratic theory insofar as it prescribes against state paternalism and against the concomitant idea that a democratic state should itself promote a particular value-infused worldview.

Theorists in the tradition of civic republicanism not only challenge the notion that commitment to value pluralism is an essential part of democracy but see it as a threat to what they consider a properly conceived democracy. They hold instead that it is a primary obligation of a democratic state to imbue its citizens with the values of civic virtue (Beiner, 1992). Others, including some who place themselves within the liberal tradition, also hold that a democratic society ought to embody visions with values specific to them.

Mill announced the principle embraced by nearly all liberals since that "the only freedom which deserves the name is that of pursuing our own good in our own way" (2002 [1859]: p. 14). His rider is that restrictions are justified when some exercising this ability interfere with others doing likewise. (He also thought that the principle did not apply to 'backward states of society,': p. 12). Much liberal theory has subsequently occupied itself with the 'paradoxes' that arise from such value pluralism, as when deciding whether or how to be tolerant of intolerant values or how to deal with preferences about what the preferences of others should be (examples of contrasting approaches, in a vast literature, are Galston (1991) and Larmore (1987)). Democratic theorists confront similar problems. If a democracy does not legally enforce some moral visions or guiding moral norms and pursue compatible educational policies, then it could find itself sanctioning activities motivated by antidemocratic visions and norms and even in the end being undone by them.

John Rawls (1921–2003) attempts to solve this problem by distinguishing between 'political' values, such as support for civil liberties, and the 'comprehensive' values embodied in general worldviews, arguing that the first merit democratic state enforcement, while full toleration should reign in respect of the latter (1996). It is questionable whether this approach can deal with the problem posed to democracy and democratic theory by theocracy. This problem is to ascertain whether or how democracy, where ultimate authority at least over political matters rests in secular hands, can be reconciled with the theocratic principle that ultimate authority on all matters is divine. Debates on this question are currently most extensive among scholars of Islam over whether or how an Islamic state can be democratic. So, against the view of Samuel Huntington (1996) and others, Khaled Abu El Fadl argues that there is room for democratic values and politics within Islam (2002). (A survey of positions on this topic is in Hashemi (2009).)

See also: Citizenship: Political; Democracy, History of; Democracy: Normative Theory; Democratic Transitions; Dewey, John (1859–1952); Global Governance; Majoritarianism and Majority Rule; Mill, John Stuart (1806–73); Pluralism; Politics and Participation; Rousseau, Jean-Jacques (1712–78); Schumpeter, Joseph A. (1883–1950); Tocqueville, Alexis de (1805–59); Value Pluralism; von Hayek, Friedrich A. (1899–1992).

Bibliography

Acton, Lord John, 1995[1862]. Nationality. In: Himmelfarb, G. (Ed.), Essays on Freedom and Power. Meridian Books, Cleveland, pp. 141–170.

Almond, Gabriel, Verba, Sidney, 1963. The Civic Culture. Sage Publications, New York.

Aristotle, 1986[c.320 BC]. Aristotle's Politics. The Peripatetic Press, Grinnell, Iowa.

Arrow, John Kenneth, et al., 1951. Social Choice and Individual Values. John Wiley, New York.

Barber, Benjamin, 1984. Strong Democracy: Participatory Politics for a New Age. University of California Press, Berkeley.

Beck, Ulrich, 2000. What Is Globalization? Polity Press, Cambridge, UK.

Beiner, Ronald, 1992. What's the Matter with Liberalism? University of California Press, Berkeley.

Beiner, Ronald, 2011. Civil Religion: A Dialogue in the History of Political Philosophy. Cambridge University Press, New York.

Benhabib, Seyla (Ed.), 1996. Democracy and Difference: Contesting the Boundaries of the Political. Princeton University Press, Princeton.

Bowles, Samuel, Gintis, Herbert, 1986. Democracy and Capitalism: Property, Community, and the Contradictions of Modern Social Thought. Basic Books, New York.

Cohen, Joshua, Rogers, Joel (Eds.), 1995. Associations and Democracy. Verso, London.
Cunningham, Frank, 2002. Theories of Democracy: A Critical Introduction. Routledge, New York.
Dahl, Robert, 1956. Preface to Democratic Theory. University of Chicago Press, Chicago.
Dahl, Robert, 1982. Dilemmas of Pluralist Democracy. Yale University Press, New Haven, CT.
Dahl, Robert, 1989. Democracy and Its Critics. Yale University Press, New Haven, CT.
Dahl, Robert, 1999. Can international organizations be democratic? A skeptic's view. In: Hacker-Cordon, C., Shapiro, I. (Eds.), Democracy's Edges. Cambridge University Press, Cambridge.
Dewey, John, 1927. The Public and Its Problems. Alan Swallow, Denver.
Diamond, Larry, 2008. The Spirit of Democracy: The Struggle to Build Free Societies throughout the World. Holt & Co, New York.
Downs, Anthony, 1957. An Economic Theory of Democracy. Random House, New York.
El Fadl, Khaled Abu, 2002. The Place of Tolerance in Islam. Beacon Press, Boston.
Galston, William, 1991. Liberal Purposes: Goods, Virtue and Diversity in the Liberal State. Cambridge University Press, New York.
Gilens, Martin, 2012. Affluence and Influence: Economic Inequality and Political Power in America. Princeton University Press, Princeton.
Glazer, Nathan, 1983. Ethnic Dilemmas: 1964–1982. Harvard University Press, Cambridge, MA.
Gutman, Amy, Thompson, Dennis, 1996. Democracy and Disagreement. Harvard University Press, Cambridge, MA.
Habermas, Jürgen, 1998. Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy. The MIT Press, Cambridge, MA.
Harden, Russell, 1982. Collective Action. The Johns Hopkins University Press, Baltimore.
Hashemi, Nader, 2009. Islam, Secularism, and Liberal Democracy: Toward a Democratic Theory for Muslim Societies. Oxford University Press, Oxford.
Hayek, Friedrich von, 1944. The Road to Serfdom. University of Chicago Press, Chicago.
Hayek, Friedrich von, 1976. Law, Legislation and Liberty. In: The Mirage of Social Justice, vol. 2. University of Chicago Press, Chicago.
Held, David, 1996. Models of Democracy, second ed. Stanford University Press, Stanford.
Held, David, McGrew, Anthony (Eds.), 2000. The Global Transformation Reader. Polity Press, Cambridge, UK.
Hirst, Paul, 1994. Associative Democracy: New Forms of Economic and Social Governance. Polity Press, Cambridge, UK.
Hobbes, Thomas, 1968[1651]. Leviathan. Penguin Books, Harmondsworth, UK.
Huntington, Samuel, 1996. The Clash of Civilizations and the Remaking of World Order. Simon & Schuster, New York.
Jenkins, Henry, Thorburn, David (Eds.), 2004. Democracy and New Media. The MIT Press, Cambridge, MA.
Kymlicka, Will, 1995. Multicultural Citizenship. Clarendon Press, Oxford.
Larmore, Charles E., 1987. Patterns of Moral Complexity. Cambridge University Press, Cambridge, UK.
Lefort, Claude, 1988. Democracy and Political Theory. Polity Press, Cambridge, UK.
Lenin, V.I., 1965[1918]. The proletarian revolution and the Renegade Kautsky. In: Lenin, V.I. (Ed.), Collected Works, vol. 28. Progress Publishers, Moscow, pp. 226–325.
Linz, Juan, Stepan, Alfred, 1996. Problems of Democratic Transition and Consolidation: Southern Europe, South Africa, and Post Communist Europe. Johns Hopkins University Press, Baltimore.
Lipjhart, Arendt, 1977. Democracy in Plural Societies. Yale University Press, New Haven.
Lipset, Seymour M., 1960. Political Man. Doubleday, Garden City.
Locke, John, 1963[1690]. Two Treatises of Government. Cambridge University Press, Cambridge, UK.
Macpherson, C.B., 2011[1962]. The Political Theory of Possessive Individualism Hobbes to Locke. Oxford University Press, Toronto.
Macpherson, C.B., 2012[1977]. The Life and Times of Liberal Democracy. Oxford University Press, Toronto.
Madison, James, et al., 1987[1788]. The Federalist Papers. Penguin Books, London.
Manin, Bernard, 1987. On legitimacy and political deliberation. Political Theory 15 (3), 338–368.
Mansbridge, Jane J., 1983. Beyond Adversary Democracy. University of Chicago Press, Chicago.
Maynor, John, 2003. Republicanism in the Modern World. Cambridge University Press, Cambridge, UK.
Mill, John S., 2002[1859]. On liberty. In: The Basic Writings of John Stuart Mill. The Modern Library, New York.
Mill, John S., 1991[1861]. Considerations on representative government. In: John Stuart Mill on Liberty and Other Essays. Oxford University Press, Oxford.
Mill, John S., 2002[1863]. Utilitarianism. In: The Basic Writings of John Stuart Mill. The Modern Library, New York.
Mouffe, Chantal, 1993. The Return of the Political. Verso, London.
Ordeshook, Peter C., 1992. A Political Primer. Routledge, New York.
Pateman, Carole, 1970. Participation and Democratic Theory. Cambridge University Press, Cambridge, UK.
Piketty, Thomas, 2014. Capital in the Twenty-first Century. Harvard University Press, Cambridge, MA.
Polsby, Nelson, 1963. Community, Power, and Political Theory. Yale University Press, New Haven.
Przeworski, Adam, et al., 2000. Democracy and Development. Cambridge University Press, New York.
Putnam, Robert, 2000. Bowling Alone: America's Declining Social Capital. Simon & Schuster, New York.
Rawls, John, 1996. Political Liberalism, expanded ed. Columbia University Press, New York.
Riker, William, 1982. Liberalism against Populism: A Confrontation between the Theory of Democracy and the Theory of Social Choice. W.H. Freeman & Company, San Francisco.
Rousseau, Jean-Jacques, 1968[1762]. The Social Contract. Penguin Classics, New York.
Sandel, Michael, 1996. Democracy's Discontent: America in Search of a Public Philosophy. Harvard University Press, Cambridge, MA.
Saward, Michael, 2000. A critique of held. In: Holden, B. (Ed.), Global Democracy: Key Debates. Routledge, London, pp. 32–46.
Schumpeter, Joseph, 1962[1942]. Capitalism, Socialism, and Democracy. Harper & Row, New York.
Schweickart, David, 1996. Against Capitalism. Westview Press, Boulder.
Shapiro, Ian, 1999. Democratic Justice. Yale University Press, New Haven, CT.
Strom, Gerald S., 1992. The Logic of Lawmaking: A Spatial Theory Approach. The Johns Hopkins University Press, Baltimore.
Tocqueville, Alexis de, 1969[1835–40]. Democracy in America. Harper & Row, New York.
Truman, David B., 1951. The Governmental Process. Knopf Publishers, New York.
Walker, R.B.J., 1993. Inside/Outside: International Relations as Political Theory. Cambridge University Press, Cambridge.
Young, Iris M., 2000. Inclusion and Democracy. Oxford University Press, Oxford.
Zolo, Danilo, 1997. Cosmopolis: Prospects for World Government. Cambridge University Press, Cambridge.

Democratic Transitions

Gerardo L Munck, University of Southern California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Theories of democratic transitions focus mainly on domestic political and economic factors, though cultural and international factors are also addressed. Many hypotheses have been tested, increasingly using quantitative methods. Yet, the quest for robust positive results has proved elusive. Thus, current knowledge supports a negative funding that there is more than one path to democracy and that democracy can emerge under all sorts of conditions.

The study on democratic transitions began to take shape as a distinct field when researchers sought to make sense of the numerous cases of democratization that, from 1974 on, led to fundamental political change in all corners of the world. This literature, in contrast to the broader literature on democratization and democracy, focuses squarely on the critical step in the history of democracy when a country passes a threshold marked by the introduction of competitive elections with mass suffrage for the main political offices in the land. The seminal texts of this field are Rustow (1970), O'Donnell and Schmitter (1986), Przeworski (1991), Przeworski et al. (2000), and Linz and Stepan (1996).

Theories

Various families of theories of democratic transitions can be distinguished according to their emphasis on political, economic, or cultural factors – all understood as domestic factors – or on international factors. The most novel aspect of theorizing on democratic transitions starting in the 1980s is the attention given to *political* explanations, and the strategic choices of political actors, in contrast to the economism and sociologism of the dominant theories in the 1960s. Central to these new theories is the idea that splits within the regime or, more precisely, among the rulers who held power, lowers the cost of dissent, and opens the possibility of bargaining between reformist sectors of the ruling elite and moderate elements of the opposition. Relatedly, these theories reject the previously widespread view that there are some economic and cultural prerequisites of democratization, and hence that certain countries can be deemed to be not ready for democracy (O'Donnell and Schmitter, 1986; Przeworski, 1991).

A related strand of theorizing focuses on the political sphere but is more macro in orientation. Many scholars seek to explain democratic transitions in terms of features of nondemocratic regimes. This literature initially drew on broad typologies of twentieth-century regimes (Linz and Stepan, 1996: Chapters 3 and 4). But it has become more strictly institutional (Geddes, 1999; Gandhi and Przeworski, 2007; Teorell, 2010: Chapter 6), placing much attention on the so-called 'hybrid regimes' characterized by nondemocratic elections (Schedler, 2013); in some cases, though, the preferences of political actors are highlighted (Bellin, 2012). In addition, rekindling an old idea (Mazzuca, 2010), some scholars seek to theorize the impact of state formation and state capacity on the prospects of democratic transitions (Bunce, 1999; Wantchekon, 2004; Bratton and Chang, 2006; Tilly, 2007).

Alongside the theories spelled out in strictly political terms, scholars continue to pay attention to *economics*, focusing on economic actors, such as social classes, and on economic factors such as economic development, inequality, and performance. But unlike the classics of the 1960s (Lipset, 1960; Moore, 1966), the new economic literature does not consider politics only as an outcome and thus contributes to a political sociology and a political economy of democratic transitions. The political sociology of democratic transitions largely takes Moore (1966) as its point of reference but brings in politics through its attention to elites, the state, and social movements (Rueschemeyer et al., 1992; Collier, 1999; Tilly, 2007). In turn, the political economy literature draws on the Meltzer and Richard's (1981) median voter model of redistributive politics, which is used to derive predictions about the preferences of elites and masses for democracy in light of expected pattern of taxation and redistribution.

Some of the key debates take place within the political economy literature. Indeed, its common reliance on the median voter model notwithstanding, the political economic literature presents some divergent hypotheses. Some scholars see democratization as a concession by elites when the costs of democracy to the elites have been sufficiently reduced (Boix, 2003), while others see democratization as driven by elite efforts to curtail revolutionary threats from below and as entailing a real conquest for the masses (Acemoglu and Robinson, 2006). As a consequence, while some authors hypothesize that inequality and democracy are inversely related, others expect to find an inverted-U relationship between inequality and democracy. Relatedly, though scholars have emphasized that resources such as oil contribute to the endurance of nondemocratic regimes by reducing the demand for democracy while increasing the repressive capacity of autocrats (Ross, 2001), others suggest that oil raises the prospects of democracy in very unequal societies by reducing elite resistance to democracy (Dunning, 2008).

Beyond the literature that seeks an explanation of democratic transitions in the political and economic spheres, two less-developed literatures focus on cultural and international factors. Much of the literature that emphasizes *culture* takes Almond and Verba's (1963) classic as its point of reference and, as that work, usually sees politics as an outcome of culture,

understood in terms of attitudes and values, and ignores or downplays the possibility that politics shapes culture (Inglehart and Welzel, 2005). The literature on the links between religion, especially Islam, and democracy fits the standard cultural approach to democracy. One of the few exceptions to this approach is the literature on nationalism, which breaks new ground in that it emphasizes the various ways in which political choices and arrangement can make democracy viable in the context of multinational societies (Linz and Stepan, 1996: Chapter 2; Stepan et al., 2011). Overall, thus, much of the literature on culture and democracy can be questioned for offering a cultural approach to politics rather than a political cultural approach and for unjustifiably positing that culture is a prerequisite of democracy (Muller and Seligson, 1994; Przeworski, 2003).

Finally, a more recent literature addresses the impact of *international* factors on democracy. An older literature provides some clues about the impact of the geopolitical context on the origins of representative government and citizenship – though not on democracy per se – mainly through its effect on state formation (Hintze, 1975; Tilly, 1990: Chapter 4). But the literature of the 1960s and 1970s on economic dependency and world systems hardly contributed to a theory of democratization (Almond, 1989). Thus, an obvious gap is filled by a new literature that pays attention to the possible impact on democracy of the global economy, international diffusion, and external political influences. Some of the better-developed arguments link international factors to political economic explanations. For example, globalization is seen as contributing to democratization by making assets more mobile and hence reducing the threat of democracy to elites (Boix, 2003). But this literature, especially the strand on the array of activities that have as a stated aim the international promotion of democracy, is quite atheoretical, and there is little basis for the strong prescriptions offered by some authors regarding what Western governments should do and not do in the name of democracy (Schmitter, 2008).

Tests

Researchers have tested theories of democratic transitions in a variety of ways. But a marked novelty, since the late 1990s, is the increase of quantitative tests, usually using data on nearly all countries of the world from 1945 or 1972 onward but in some cases going back to 1800. All in all, scholars have devoted much time to assessing whether theories can postdict or explain the known course of events. Nonetheless, one overall conclusion of this research is that the quest for robust positive results has proved elusive.

The most thoroughly tested hypothesis, derived from modernization theory, concerns the impact of economic development, usually operationalized in terms of per capita income, on the prospects of democratic transitions as distinct from the endurance of democracy. Przeworski et al. (2000: Chapter 2) provide a key disconfirmation of this aspect of modernization theory, and Przeworski's finding is backed up by other statistical analyses (Acemoglu et al., 2008). But other authors challenge this finding (Boix and Stokes, 2003). Similarly, though some research shows that oil has a negative effect on democratization (Ross, 2001), other studies suggest that such a pattern holds only after 1980 and not in Latin America (Dunning, 2008; Ross, 2012: Chapter 3), and some researchers do not find an association between resource reliance and authoritarianism, and challenge the idea of a resource curse (Haber and Menaldo, 2011). Much the same can be said with regard to the inequality–democracy link. Some qualitative and quantitative studies support the idea that distributive politics is central to democratization (Tilly, 2004: 255, Przeworski, 2009) and hence question the theory that posits an inverse relationship between levels of inequality and democratization. But others suggest that the role of distributive politics is a common but far from universal feature of democratic transitions (Collier, 1999; Haggard and Kaufman, 2012). And yet others suggest that there is neither an inverse nor an inverted-U relationship between inequality and democracy (Houle, 2009).

Two comprehensive quantitative studies reinforce the view that robust positive results are hard to come by. A sensitivity analysis that considers over 50 possible explanatory factors and is based on some 1.7 million regressions with data from 165 countries from 1976 to 2002 concludes that only two determinants of democratic transitions are robust: GDP growth, which has a negative effect, and past transitions, which have a positive effect (Gassebner et al., 2013). In other words, most of the tested hypotheses are not confirmed. In turn, a quantitative study that tests a broad range of hypotheses using data on 165 countries from 1972 to 2006 concludes, "when the entire range of putative determinants is taken into account, [the] statistical models at best explain 10 percent of the yearly change in democracy" (Teorell, 2010: 10, 146–148). That is, the range of theories that are tested perform very poorly (for more detailed overviews of research findings, see Geddes, 2009; Schmitter, 2010; Teorell, 2010; Coppedge, 2012).

The failure to uncover global patterns does, however, provide indirect but strong support for a negative conclusion. Indeed, the accumulated empirical research does support the view (1) that there is more than one path to democracy (Rustow, 1970: 345, Collier, 1999; Tilly, 2004: 217) and (2) that democracy can emerge under all sorts of conditions (Przeworski et al., 2000: Chapter 2; Wantchekon, 2004: 17; Gassebner et al., 2013: 171). This is not a meaningless conclusion. It runs counter to common assumptions in the literature and oft-repeated arguments about democratic transitions. Moreover, it suggests that with regard to prediction – a challenge that is different from postdiction and explanation – the well-known failure of scholars to predict the collapse of communism in 1989 and the Arab Spring that began in the Middle East and North Africa in late 2010 is not just a reflection of the current state of knowledge. In all likelihood, the failure to predict these events is also due to the uncertainty that is inherent in political changes as momentous as transitions to democracy.

Challenges

Given the current state of knowledge, some suggestions can be ventured regarding challenges that need to be tackled, so as to ensure the continued development of research on democratic

transitions. First, because democratization entails a political change, it makes sense to look for explanations within the political sphere. However, because politics is not fully autonomous, it is also important to avoid a politicist approach and to consider interactions between politics and other spheres. This has been one of the strengths of the literature on democratic transitions. Yet, it is also clear that the strongest theory either focuses on politics or bridges the political and economic spheres. Thus, while it is critical to address the potential impact on democratic transitions of cultural and international factors, it is advisable to go about this task by building on the more developed theorizing in the political and political economy literature.

Second, though theorizing on democratic transitions has been fruitful, it has also been somewhat unstructured. Indeed, the fact that a comprehensive effort to test theories addresses some 60 hypotheses, each considered as equally important (Gassebner et al., 2013), suggests that something is wrong with the approach to theorizing used in this literature. Thus, future theorizing should be geared less to adding yet more hypotheses and more to embedding hypotheses in broad theoretical frameworks, developing theoretical syntheses, and formulating theory in a way that lends itself to more definitive testing.

Third, though many scholars have taken important steps to theorize political processes, most theories are largely a matter of macro-associations and most quantitative tests address conditions as opposed to actions. Thus, inasmuch as researchers seek to offer explanations, it is crucial that more emphasis is placed on causal mechanisms. And what this means is that theorists should focus on events and processes and formally propose testable hypotheses about causal mechanisms.

Finally, though the strengths of current tests should be recognized – the amount and rigorousness of tests has improved considerably over the past 15 years – so should the limits of these tests be noted. Many theories have not been tested. Thus, it is quite easy to make a case that most tests suffer from omitted variable bias; similarly, the possibility of reverse causation has not been thoroughly addressed. And many tests rely on poor data or poor data analysis. Yet, little has been done to rely on explicit criteria that would distinguish more credible from less credible results – the credibility of results is currently based on vaguely specified, informal criteria – and hence to distinguish theories that deserve to be further developed from theories that should be discarded. Indeed, in light of the explosion of empirical research, the need for criteria to identify credible results has taken on added urgency.

See also: Authoritarianism; Democracy, History of; Revolutions of 1989–90 in Eastern Central Europe; Russian and Post-Soviet Studies: Politics.

Bibliography

Acemoglu, D., Robinson, J.A., 2006. Economic Origins of Dictatorship and Democracy. Cambridge University Press, New York.

Acemoglu, D., Johnson, S., Robinson, J.A., Yared, P., 2008. Income and democracy. American Economic Review 98 (3), 808–842.

Almond, G., 1989. The international–national connection. British Journal of Political Science 19 (2), 237–259.

Almond, G., Verba, S., 1963. The Civic Culture: Political Attitudes and Democracy in Five Nations. Princeton University Press, Princeton, NJ.

Bellin, E., 2012. Reconsidering the robustness of authoritarianism in the Middle East: lessons from the Arab spring. Comparative Politics 44 (2), 127–149.

Boix, C., 2003. Democracy and Redistribution. Cambridge University Press, New York.

Boix, C., Stokes, S.C., 2003. Endogenous democratization. World Politics 55 (4), 517–549.

Bratton, M., Chang, E.C.C., 2006. State building and democratization in Sub-Saharan Africa: forwards, backwards, or together? Comparative Political Studies 39 (9), 1059–1083.

Bunce, V., 1999. Subversive Institutions. The Design and the Destruction of Socialism and the State. Cambridge University Press, New York.

Collier, R.B., 1999. Paths toward Democracy: Working Class and Elites in Western Europe and South America. Cambridge University Press, New York.

Coppedge, M., 2012. Democratization and Research Methods. Cambridge University Press, New York.

Dunning, T., 2008. Crude Democracy: Natural Resource Wealth and Political Regimes. Cambridge University Press, New York.

Geddes, B., 1999. What do we know about democratization after 20 years? Annual Review of Political Science 2, 115–144.

Geddes, B., 2009. Changes in the causes of democratization through time. In: Landman, T., Robinson, N. (Eds.), Sage Handbook of Comparative Politics. Sage, London, pp. 278–298.

Gandhi, J., Przeworski, A., 2007. Authoritarian institutions and the survival of autocrats. Comparative Political Studies 40 (11), 1279–1301.

Gassebner, M., Lamla, M.J., Vreeland, J.R., 2013. Extreme bounds of democracy. Journal of Conflict Resolution 57 (2), 171–197.

Haber, S., Menaldo, V., 2011. Do natural resources fuel authoritarianism? A reappraisal of the resource curse. American Political Science Review 105 (1), 1–26.

Haggard, S., Kaufman, R.R., 2012. Inequality and regime change: democratic transitions and the stability of democratic rule. American Political Science Review 106 (3), 495–516.

Hintze, O., 1975[1902]. The formation of states and constitutional development: a study in history and politics. In: Gilbert, F. (Ed.), The Historical Essays of Otto Hintze. Oxford University Press, New York, pp. 157–177.

Houle, C., 2009. Inequality and democracy: why inequality harms consolidation but does not affect democratization. World Politics 61 (4), 589–622.

Inglehart, R., Welzel, C., 2005. Modernization, Cultural Change, and Democracy. Cambridge University Press, New York.

Linz, J.J., Stepan, A., 1996. Problems of Democratic Transition and Consolidation: Southern Europe, South America and Post-communist Europe. Johns Hopkins University Press, Baltimore, MD.

Lipset, S.M., 1960. Political Man: The Social Bases of Politics. Doubleday/Anchor Books, New York.

Mazzuca, S., 2010. Macrofoundations of regime change: democracy, state formation, and capitalist development. Comparative Politics 43 (1), 1–19.

Meltzer, A.H., Richard, S.F., 1981. A rational theory of the size of government. Journal of Political Economy 89 (5), 914–927.

Moore, B., 1966. Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World. Beacon Press, Boston.

Muller, E.N., Seligson, M., 1994. Civic culture and democracy: the question of the causal relationships. American Political Science Review 88 (3), 635–652.

O'Donnell, G., Schmitter, P., 1986. Transitions from Authoritarian Rule. Tentative Conclusions about Uncertain Democracies. Johns Hopkins University Press, Baltimore, MD.

Przeworski, A., 1991. Democracy and the Market. Political and Economic Reforms in Eastern Europe and Latin America. Cambridge University Press, New York.

Przeworski, A., 2003. Why do political parties obey results of elections? In: Maravall, J.M., Przeworski, A. (Eds.), Democracy and the Rule of Law. Cambridge University Press, New York, pp. 114–144.

Przeworski, A., 2009. Conquered or granted?: a history of franchise extensions. British Journal of Political Science 39 (2), 291–321.

Przeworski, A., Alvarez, M.E., Cheibub, J.A., Limongi, F., 2000. Democracy and Development: Political Institutions and Well-being in the World, 1950–1990. Cambridge University Press, New York.

Ross, M.L., 2001. Does oil hinder democracy? World Politics 53 (3), 325–361.

Ross, M.L., 2012. The Oil Curse: How Petroleum Wealth Shapes the Development of Nations. Princeton University Press, Princeton, NJ.

Rueschemeyer, D., Stephens, E.H., Stephens, J.D., 1992. Capitalist Development and Democracy. University of Chicago Press, Chicago.

Rustow, Dankwart, 1970. Transitions to democracy: toward a dynamic model. Comparative Politics 2 (3), 337–363.

Schedler, A., 2013. The Politics of Uncertainty: Sustaining and Subverting Electoral Authoritarianism. Oxford University Press, Oxford.
Schmitter, P.C., 2008. International democracy promotion and protection: theory and impact. In: Teixeira, N.S. (Ed.), The International Politics of Democratization: Comparative Perspectives. Routledge, New York, pp. 26–52.
Schmitter, P.C., 2010. Twenty-five years, fifteen findings. Journal of Democracy 21 (1), 17–28.
Stepan, A., Linz, J.J., Yadav, Y., 2011. Crafting State–Nations: India and Other Multinational Democracies. Johns Hopkins University Press, Baltimore, MD.
Teorell, J., 2010. Determinants of Democratization: Explaining Regime Change in the World, 1972–2006. Cambridge University Press, New York.
Tilly, C., 1990. Coercion, Capital, and European States, AD 990–1990. Basil Blackwell, Oxford.
Tilly, C., 2004. Contention and Democracy in Europe, 1650–2000. Cambridge University Press, New York.
Tilly, C., 2007. Democracy. Cambridge University Press, New York.
Wantchekon, L., 2004. The paradox of "warlord" democracy: a theoretical investigation. American Political Science Review 98 (1), 17–33.

Demographic Impact of Disasters

Elizabeth Frankenberg and Maria M Laurito, Sanford School of Public Policy, Duke University, Durham, NC, USA
Duncan Thomas, Duke University, Durham, NC, USA

Published by Elsevier Ltd.

Abstract

The frequency and magnitude of large-scale disasters in recent years has prompted increased interest in better understanding how major disruptive events alter key demographic processes. This article summarizes evidence establishing that disasters have significantly impacted mortality, health, fertility, and migration. While these processes are intimately interrelated, there have been relatively few integrative analyses that draw the evidence together, in large part because of inadequate data. Investment in population data collection systems to provide scientific evidence in the wake of disasters will broaden the depth and scope of disaster research, advance understanding of demographic changes, and inform policy interventions.

Introduction

The past decade has seen multiple large-scale natural disasters. These events, in combination with predictions that such events will increase in frequency as a result of global warming and rising population densities in vulnerable areas, have catalyzed interest in how affected populations respond to disasters, the effectiveness of emergency response programs, and whether emergency or longer-term assistance programs alter the life course trajectories of those affected by disasters.

Both disasters and recovery efforts have the potential to affect many outcomes of interest to social and health scientists. Accordingly, a wide array of disciplines has contributed to the study of disasters. We thus begin with a brief review of definitions of disasters and approaches to their study. This article focuses on the relationship between disasters and demographic phenomena. We therefore focus on mortality, fertility, and migration. These are three key demographic outcomes that are affected not only by disasters themselves but also by policies and programs implemented both before and after disasters.

The Study of Disasters

Disciplinary Approaches

Efforts to define disasters and use these definitions to shed light on various social, health, and demographic phenomena have a long history in the social sciences, particularly in sociology, geography, and epidemiology. Early sociological work on disasters focused on understanding collective behavior under high-stress conditions, taking the approach of dispatching researchers to stricken communities to observe postdisaster dynamics as they unfolded (Fritz and Marks, 1954). Although areas of emphasis have varied over the years, understanding how disasters strain social systems has remained a focus of this research (Quarantelli, 1989; Dynes et al., 1994; Klinenberg, 2002; Browning et al., 2006; Tierney, 2007). Within geography, efforts have centered more on the vulnerability of human physical systems to hazards (Gray et al., 2014) and on how to increase resilience of human systems, most frequently by studying one or a small number of communities. Disaster epidemiology provides timely assessments of the short- and long-term impact on health, broadly defined, and seeks to improve prevention and mitigation strategies (Guha-Sapir and Hoyois, 2012).

In many other disciplines, disasters are viewed through a broader lens. For example, in the economics literature, disasters are often treated as part of a more general class of 'shocks,' or unexpected changes. Shocks have been used to make inferences regarding the extent to which individuals are able to protect themselves, their families, and their communities from unexpected events with potentially adverse consequences (Townsend, 1994). A related literature exists in demography, although the focus is on demographic outcomes, and larger-scale macroevents with the potential to affect whole populations have received more attention (see, for example, Heuveline's innovative reconstruction of the Cambodian population after the Khmer Rouge regime) (Heuveline, 1998, 2001; Heuveline and Poch, 2007). Economic and demographic analyses have tended to prioritize representativeness at the population level and typically draw comparisons on a larger scale than geographic or purely sociological approaches. Finally, a large body of work in psychology and epidemiology examines the implications of disasters for various mental and in some cases physical health outcomes, most frequently through the analysis of survey data of varying degrees of representativeness (Armenian et al., 1998; Cao et al., 2003; Norris and Elrod, 2006).

Definitions of Disaster

Many definitions of disaster have been proffered over the years, generating a great deal of debate (Quarantelli, 1989, 2005). Most definitions recognize disasters as phenomena that arise from an unusually extreme precipitating event that is concentrated in time and space and that overwhelms local or, in some cases, larger-scale systems. Over time, definitions have increasingly recognized the importance of social construction on whether an event is interpreted as a disaster.

From a perspective based in sociology, Kreps (2001) defines disasters as "non-routine events in societies that involve conjunctions of physical conditions with social definitions of human harm and social disruption." Several important elements of this definition warrant highlighting. First,

according to this definition, disasters do not occur without people and the social systems they create. Second, chronic conditions, such as poverty, are not considered disasters. In some contexts, wars, genocides, famines, disease outbreaks, and weather-related events like floods, hurricanes, and tornadoes fit naturally into this definition. However, it is important to be precise about 'nonroutine.' In some parts of the world, these types of events have become all too routine. In those cases, it is necessary to appeal to an argument that the incidence may be routine but the precise timing or the severity of the event is not routine in order to apply this definition of a disaster. Alternatively, studies have treated the onset of such events as a disaster.

Disasters typically involve significant loss of resources and threats to – or actual loss of – life. The combination of these factors may result in a humanitarian crisis. Some studies have tried to classify precipitating events into various typologies, for example, whether they arise from environmental phenomena such as geophysical forces or weather, or from technical or willful actions. Sociologists' emphasis on understanding shifting social processes caused by disasters, beyond studying the catastrophic event itself, has resulted in the evolution of the set of events that are included within the rubric of a disaster, in particular as new findings related to social impacts have emerged (Herring, 2013). It is widely recognized that the interaction of the physical events with characteristics of the affected societies, such as the nature of housing and patterns of land use, plays a fundamental role in how disasters unfold.

Other important and distinguishing features that vary across disasters are the magnitude, scope and duration of the event, the degree of advance warning available to the population at risk, and the precision with which the event can be predicted. As an example, the 2004 Indian Ocean tsunami that hit the island of Sumatra, Indonesia, and Hurricane Katrina (that affected the U.S. gulf coast) were very different with respect both to whether the event was (or could have been) predicted and the amount of time that those likely to be affected had to prepare for the onslaught of the water. The tsunami hit some coastlines of Indonesia within minutes of the earthquake, whereas residents of New Orleans were warned about the approaching storm several days before Katrina made landfall (Munasinghe, 2007). Moreover, whereas hurricanes have been a part of the history of New Orleans since it was settled in the early eighteenth century, geological evidence indicates that there have been no tsunamis along the coast of Sumatra for over six centuries.

A Framework for Considering Disasters' Demographic Impacts

Demographers have examined disasters largely when their consequences occur at a scale with the potential to affect regional or national populations. Key parameters of interest are typically mortality, health, fertility, and migration, which, in turn, have implications for population size and composition.

The most dramatic link between disasters and demography arises in those disasters that cause substantial numbers of deaths. The risk of death in a disaster may vary by age and sex, reflecting differences in vulnerability across these dimensions as a result of physical differences or likelihood of exposure. Socioeconomic status may also be associated with risks of exposure such as when the poorest live in particularly vulnerable areas (such as flood plains) or when damage caused by the disaster depends on the quality of housing (as might be the case for earthquakes or tornadoes). In addition to immediate deaths, disasters may have longer-term or indirect effects on mortality through their effects on health status or its drivers.

Disasters also have the potential to displace people, either because people move preemptively or because the disaster affects their property or source of livelihood in ways that make remaining in the affected area unattractive or impossible. After disasters, displaced individuals may return, and others may move into the area, attracted by opportunities of various forms.

Though the mechanisms are less immediately obvious, disasters can also change fertility patterns. Disasters and the associated stress can affect coital frequency and potentially the ability to conceive a child or carry a pregnancy to term or they can change the demand for children. Disruption of services could also affect access to contraception. Disasters that result in large numbers of death may be followed by increased fertility. Finally, if the disaster changes the age and sex composition of the population as a result of mortality or migration, patterns of union formation may change in ways that alter fertility trends.

Several implications of the preceding discussion are worth drawing out in a little detail. First, although disasters typically begin with a precipitating event that can be pinpointed precisely in time, and often in space, temporal and spatial end points of disasters' impacts are not so cleanly delineated. Effects may be felt for a far longer time scale and over a far greater geographic area than that within which the disaster causes physical damage. A corollary is that the longer-term effects of a disaster may well reach far beyond those who were directly. For example, a disaster that affects the local area population composition may also affect the labor and marriage markets on a wider scale. Furthermore, a disaster that is followed by in-migration from other areas clearly affects not only the affected population and the migrants but also the communities from which the migrants have moved.

Second, the demographic processes we have discussed earlier are connected in myriad ways. To provide just a few examples, decisions to move are likely to be affected by household structure and decision-making processes which are, themselves, directly altered by mortality. Loss of kin and migration can affect stress levels, health, the desire for children, and the ability to conceive and bear children. Relatively few analyses consider these interlinkages, in part because few studies collect information on a broad set of potentially relevant behaviors and outcomes; fewer studies have successfully followed the trajectories of individuals, families, and communities affected by disasters over time.

Third, although we have emphasized the negative impacts of disasters, it is important to note that to the extent that disasters change opportunities, some of these changes may have positive consequences for individuals and communities. As disasters destroy land, housing, or infrastructure in one area, the value of comparable land, housing, or infrastructure is likely to increase. Disasters are often followed by reduced access to food and shelter: those who can provide food or

shelter are potential beneficiaries of this misfortune. In many cases, these are provided by governments or nongovernmental organizations as assistance. Indeed, assistance programs, themselves, may provide opportunities that change outcomes in positive ways as, for example, the programs buy food from local farmers. Other policy changes may affect longer-term outcomes and diminish the potential for known hazards to turn into large-scale disasters in the future. More generally, the policy environment can affect demographic phenomenon associated with disasters in significant ways.

High-Impact Disasters of the Last Decade

Before turning to evidence regarding demographic change in the context of disasters, we review the scope of disasters that have occurred over the last decade based on two indicators of impact, the death toll and the estimated value of property damage.

Table 1 summarizes impacts for the five disasters that are estimated to have had the highest number of deaths over the past decade (Panel 1) and the five disasters that are estimated to have had the greatest damage measured in financial terms (Panel 2). Of course, financial estimates of damage are, at best, rough; the table is intended to provide an illustration of the magnitudes that are involved.

With respect to mortality, the deadliest disaster was the 2004 Indian Ocean tsunami, which cost over 225 000 people their lives in Southeast and South Asia. The 2010 Haiti earthquake killed over 222 000, followed by Cyclone Nargis (138 366). Two other earthquakes, in China and Pakistan, caused tens of thousands of deaths as well.

With respect to property damage (measured in dollars of the year of the event), the earthquake and tsunami that hit Japan tops the list. Estimated property damages were approximately $US 210 billion, and Hurricane Katrina caused damages of $US 125 billion, according to data from the Center for Research on the Epidemiology of Disasters (CRED) at the University of Louvain. The 2008 earthquake in China is the only disaster that appears in both panels, which points to an interesting feature of disasters: the deadliest are not necessarily the costliest in terms of property damage. Instead, the highest costs have accrued in high- or medium-income countries, and in three of those disasters the numbers of deaths were quite low.

How such a table will look in a decade is unknown. On the one hand, if the various hazards mitigations and emergency preparedness policies that have reduced death tolls in industrialized countries can be implemented in a greater number of developing countries, death tolls from disasters may diminish. On the other hand, the size and density of populations in areas vulnerable to extreme geophysical or meteorological events is growing, which increases the number of people at risk of exposure.

The table is based on estimates aggregated to the national level. While useful as summary evidence regarding scale, the numbers shed little light on the full demographic impact of these events, let alone processes and mechanisms linking disasters to demographic change over the longer term.

Indeed, a major barrier to studying the demography of disasters is the relative scarcity of appropriate data. Many studies are based on relatively small-scale unrepresentative samples of individuals clustered into camps or other highly visible housing arrangements. Other studies use interviews of those who have remained behind in the affected areas. Typically, interviews are conducted once, shortly after the disaster. It is difficult to draw firm conclusions about the impact of the disaster on processes of population change with these types of data. The complexities of collecting population-representative data that characterize the periods both before and after the disaster, that span a continuum of destruction, and that follow movers to destinations other than camps have severely limited the representativeness of

Table 1 Death and property destruction

Country	*Year*	*Type of disaster*	*Deaths*	*Affected population*	*Damage[a] ($US million)*
1. Costliest for lives					
South and Southeast Asia	2004	Earthquake/tsunami	225 841	2 273 723	7791
Indonesia			165 708	532 898	4452
Sri Lanka			35 399	1 019 306	1317
India			16 389	654 512	1023
Thailand			8 345	67 007	1000
Haiti	2010	Earthquake	222 570	3 700 000	8000
Myanmar	2008	Cyclone Nargis	138 366	2 420 000	4000
China	2008	Earthquake	87 476	45 976 596	85 000
Pakistan	2005	Earthquake	73 338	5 128 309	5200
2. Costliest for property					
Japan	2011	Earthquake/tsunami	19 846	368 820	210 000
United States	2005	Hurricane Katrina	1 833	500 000	125 000
China	2008	Earthquake	87 476	45 976 596	85 000
United States	2012	Hurricane Sandy	154	8 500 000[b]	65 000
Chile	2010	Earthquake	562	2 671 556	30 000

[a]The estimated damage is dollars of the year of occurrence.
[b]Estimate based on affected power consumers in the US, Tropical Cyclone report, Hurricane Sandy, National Hurricane Center, 2013 – http://www.nhc.noaa.gov/data/tcr/AL182012_Sandy.pdf.
Source: EM-DAT: The OFDA/CRED International Disaster Database – www.emdat.be, University of Louvain, Brussels, Belgium. NOAA: Billion-Dollar Weather/Climate Disasters – www.ncdc.noaa.gov/billions/events.

study samples, sample sizes, and follow-up periods of available data (Sastry and VanLandingham, 2009; Galea and Maxwell, 2009).

Demographic Processes and Outcomes

In this section we review the evidence for linkages between disasters and various demographic outcomes.

Mortality and Morbidity

As shown in **Table 1**, disasters can cause loss of life on a massive scale. The extent of mortality caused by disasters varies depending on factors such as type of disaster, location, and timing. Counts of the dead and missing may not be accurate, and in some cases these rough estimates may not be finalized into an official death toll. Whether deaths have a meaningful impact on population size depends, in part, on the extent to which mortality is spatially concentrated. Mortality rates can help clarify impact, but their calculation requires information on the number of individuals exposed as well as the number dead – information that is not always available.

Risk of mortality may vary by age and sex in ways that differ across disasters. Several studies provide evidence that the Indian Ocean tsunami was associated with higher mortality for women than for men, and for the young and old. For example, based on a survey representative of pre-tsunami population, Frankenberg et al. (2011) show that mortality was lowest among men aged 20–44 years. Mortality rates were significantly higher for same-aged women, and higher still for children and teenagers, and for older individuals. It is likely that these differences reflect differences in strength and ability to swim since exposure to the tsunami is unlikely to vary by demographic group. Similar results are also reported for Sri Lanka following the 2004 Indian Ocean tsunami, where women and children experienced higher mortality compared to adults aged 20–29 years (Nishikiori et al., 2006).

These mortality patterns are unusual. They are not observed for the tsunami in Japan, where mortality was lower for school-age children than for individuals of other ages, and there are no significant differences in tsunami-related mortality between men and women (Nakahara and Ichikawa, 2013). Nor do they describe mortality in Haiti following the 2010 earthquake, where mortality was highest among those aged 18 and younger. In Haiti, about two-thirds of deaths were children younger than 12 years. Not only did children suffer higher mortality rates during the earthquake, they were also more likely to die from injuries and illness in the aftermath of the disaster (Kolbe et al., 2010). In contrast, mortality rates associated with Hurricane Katrina were highest among older adults (those 75 years or older), mostly due to these people not leaving the affected areas prior to the hurricane and ultimately drowning in the water (Brunkard et al., 2008).

Beyond mortality, disasters can affect health, with the nature of the impacts contingent on the characteristics of the event. Earthquakes often cause multiple injuries, including lacerations, fractures, contusions, chest and neurological problems, and potential cardiovascular issues (Bartels and VanRooyen, 2011). Health problems triggered by a disaster can continue in the long term. Residents of Louisiana and Mississippi experienced an increase in headache, nausea and digestive ailments, and respiratory and cardiac problems after Katrina (Adeola and Picou, 2012). Longer-term effects may be particularly important when the disaster creates additional health and environmental hazards, such as chemical, oil, or radiation spills.

The psychological, physical, and economic stresses associated with disasters can also affect birth outcomes for children born to women who are exposed while pregnant, and potentially these children's longer-term health outcomes. Evidence from offspring of women pregnant during the Dutch Hunger Winter indicates that extremely low caloric intake during the pregnancy results in reduced weight at birth, shorter stature as an adult, elevated risk of cardiovascular disease in midlife, and premature mortality. These results have been interpreted as being driven by the impact of the *in utero* nutritional insults through plausible biological pathways and have also implicated the postnatal nutrition environment (Stein et al., 1975; Ravelli et al., 1998, 1999; Roseboom et al., 2001; Rooij et al., 2010).

Several studies have explored the impact of natural disasters such as droughts or floods that occur *in utero* or in early life and established that these events can have effects that reach well into adulthood (Foster, 1995; Maccini and Yang, 2009). Other studies have investigated the impact of large-scale outbreaks of infectious diseases, most prominently, the 1918 influenza pandemic. Arguing the pandemic was unanticipated, Almond (2006) concludes that *in utero* exposure to the pandemic resulted in worse socioeconomic outcomes in adulthood relative to those surrounding cohorts that were not exposed. However, Brown and Thomas (2014) establish that the parents of the cohorts exposed to the pandemic were also of lower socioeconomic status relative to the parents of the surrounding cohorts, which complicates interpretation of the evidence. This highlights the general issue of the centrality of establishing whether a disaster is anticipated when interpreting the evidence. If the disaster is anticipated (or anticipatable), estimated effects likely reflect a combination of the impact of the disaster and potential behavioral responses in anticipation of the disaster.

Other studies have sought to investigate the relationship between stressful experiences during pregnancy and the physical health of offspring. These studies use quasiexperimental designs, drawing on some unexpected source of stress, such as acts of terror or war (Lauderdale, 2006; Camacho, 2008), an epidemic of crime (Brown, 2014), earthquakes (Torche, 2011), or hurricanes (Currie and Rossin-Slater, 2012). Using birth record data, they compare outcomes at birth of children *in utero* at the time of the event with outcomes of comparison children who were not exposed *in utero*. While the findings are consistent with the hypothesis that stress exposure affects birth outcomes, the estimated impacts are generally very small in magnitude and clinically unimportant. Moreover, many of the studies use large samples and the statistical significance of the estimated effects is not clear.

Morbidity after disaster extends beyond physical health. Studies in psychology and epidemiology have documented a number of systematic patterns of postdisaster mental health.

Studies assess a variety of specific and nonspecific psychological problems, but posttraumatic stress, depression, and anxiety are the most common ones regardless of the nature of the disaster (Norris et al., 2002; Norris and Elrod, 2006).

Relatively few studies are able to examine changes in psychological symptoms over time, but among those that do, most find that symptoms improve (Norris et al., 2002; Pietrzak et al., 2012). Data from the Galveston Bay Recovery Study (GBRS) show that individuals exposed to Hurricane Ike had symptoms consistent with PTSD, anxiety, and panic but that prevalence of these symptoms declined with time (depression remained stable across waves) (Pietrzak et al., 2012). Similar evidence has been reported for Indonesians affected by the Indian Ocean tsunami (Frankenberg et al., 2014).

Migration and Relocation

Migration, whether voluntary or impelled, is a potentially strong factor in reshaping populations after a disaster. In the context of disasters, population movements encompass temporary and permanent migration, both out of and into disaster-affected areas. Destinations can include unaffected areas (where the displaced may or may not have family or friends) as well as locations relatively nearby the impacted area where formal and informal camps may be established. But residents may also choose to stay where they were living at the time of the disaster, even if the area was badly affected, and new groups may move in, particularly if they perceive opportunities (most likely work-related) associated with relocating.

Migration decisions are affected by damage to the built and natural environment, among other things. Massive destruction displaces victims from their homes, potentially for extended periods of time. The 2010 earthquake in Haiti destroyed an estimated 24.4% of homes and damaged 41.5%. About 1.24 million Haitians were relocated to temporary settlements (Kolbe et al., 2010). Overall, it is estimated that Port-au-Prince lost 23% of its permanent residents.

Using an unusual methodology that relies on mobile phone data to study migration patterns in Haiti following the earthquake, Lu et al. (2012) analyze movements of 1.9 million phone users for up to 1 year after the earthquake. They show that shifts in location were more predictable than originally anticipated. Movements were closely related to patterns that predated the earthquake and were associated with location of social support networks.

Hurricane Katrina is another disaster that caused an almost unprecedented level of relocation of people in the United States, with some 1.5 million people emigrating from affected areas (Groen and Polivka, 2010). For residents of heavily damaged and low-income communities, opportunities to return to New Orleans were highly constrained by resource availability and the degree of destruction to their homes. Using a representative sample of pre-Katrina residents of New Orleans, Fussell et al. (2010) show that black residents experienced higher rates of housing damage, which often delayed their return to the city and contributed to reshape demographic characteristics postdisaster. Relying on data from the Current Population Survey, Groen and Polivka (2010) show that age, family income, and magnitude of damage at the county level significantly conditioned return of displaced populations even a year after the hurricane.

As discussed briefly earlier, some outcomes linked to disasters may have unexpected positive elements. Hurricanes Katrina and Rita caused massive migrations of students from low-performing to higher-quality schools. After an initial drop in median test scores, performance quickly recovered, and by 2009 median tests scores for movers rose by 0.18 standard deviations (Sacerdote, 2012).

Overall, Hurricane Katrina served to accelerate an ongoing reduction of the population of New Orleans (Zaninetti and Colten, 2012). In some cases, however, disasters do not result in significant population shifts. Using data collected from 291 tornado victims in Bangladesh, Paul (2005) shows that for the most part victims of this event decided to remain in affected areas, where they were likely to receive recovery assistance – an example of how postdisaster programs can generate their own demographic impacts.

Gray et al. (2014) compare mobility strategies among individuals living in small local areas that sustained different degrees of damage caused by the 2004 Indian Ocean tsunami in Indonesia. Those analyses call into question several common assumptions in the literature. For one thing, many individuals from heavily damaged areas did not move to IDP camps. More than half of the population either relocated to private homes or stayed in their community weathering adverse living conditions. Outside the heavily damaged zone, there was movement among individuals from areas relatively untouched by the tsunami, and regardless of origin area, mobility was higher for the better educated. These data underscore the likely value-added of studies that are designed to be population representative and capture both immediate impacts of disasters as well as responses to disasters over time.

Disasters can also attract new residents to damaged areas, especially if reconstruction efforts create employment opportunities. Given data limitations, it can be difficult to differentiate between returning and new migrants. Analyses of population structure of New Orleans after Hurricane Katrina show an increase in the percentage of Hispanic residents compared to predisaster levels (Groen and Polivka, 2010). Using census data from 2000 to 2010, Zaninetti and Colten (2012) estimate that the Hispanic share of the population in New Orleans increased by 57%, while the hurricane accelerated the decline of the black population. Latino immigrants took advantage of high demand of construction labor in the aftermath of Hurricane Katrina to relocate to New Orleans. Using a sample collected within 5 and 7 months after Katrina, Fusell (2009) shows that Hispanic immigrants moving to the area were younger and had weaker social networks and were often of Brazilian and Mexican origin.

Fertility, Reproductive Health, and Family Change

Disasters have the potential to affect fertility, reproductive health, and family change through a variety of mechanisms. Most proximately, high mortality disasters may change family composition. Adult deaths can leave young children without parents and create widows and widowers, who may or may not remarry. The 1958–61 famine in China is an example of an event that had significant consequences for family

demography. During the famine period, fertility and marriage rates declined, while divorce rates and family division increased (Zhao and Reimondos, 2012). It is not clear that the entire 3-year famine can be treated as unanticipated, which complicates interpretation of the results.

However, the Indian Ocean tsunami caused similar disruption to families. Using a population-based survey for Indonesia, Frankenberg et al. (2011) estimate that about 10% of children lost a parent and 10% of survivors lost a spouse. Deaths of children may leave parents with fewer offspring than they planned for and desire. On the other hand, when disasters cause high mortality, they are also likely to destroy assets and livelihoods, which may leave unmarried adults feeling less prepared to enter a union and couples feeling less prepared to provide for another child.

Even when disasters do not result in substantial mortality, they may prompt family change. Using a theoretical framework guided by work in psychology on stress and attachment, Cohan and Cole (2002) analyze rates of marriage, birth, and divorce before and after Hurricane Hugo, in affected and unaffected South Carolina counties. In counties declared disaster areas, rates for each of these outcomes rise and then fall. These changes do not occur in other counties, leading the authors to suggest that exposure to a life-threatening event prompted significant actions and measurable changes with respect to close relationships.

Potential changes to fertility can also be considered through a proximate determinants framework. Stress-affected populations may experience changes in coital frequency, rates of conception, miscarriages, and stillbirths. With respect to pregnancy outcomes, the Chinese famine resulted in a significant increase in stillbirths and miscarriages. By the end of the famine period, only 92% of pregnancies resulted in live births, down from more than 95% in 1958 (Zhao and Reimondos, 2012). Deterioration of physical health in the case of the famine also contributed to delays in reproductive development of young girls and reduced rates of breastfeeding (Zhao and Reimondos, 2012). Evidence from the 2004 Indian Ocean tsunami suggests that in the months following the disasters, rates of miscarriage rose for women from badly damaged areas (Hamoudi et al., 2012).

On the other hand, destruction of facilities and services may reduce access to contraception, leading to unintended births. Just as disasters destroy homes, shocks caused by disasters can create significant disruptions in infrastructure and complicate access to contraceptive supplies. Following the 2004 Indian Ocean tsunami, there were significant issues with health infrastructure (public and private health centers and midwifery). Frankenberg et al. (2014) show that for many communities disruptions in health facilities extended for at least a month and their reopening did not necessarily imply resumption in contraceptive supply. Midwifery services also suffered interruptions, in particular in communities where midwifes were among tsunami casualties.

Literature exploring empirical links between contraceptive availability after disasters and behaviors is limited. Some evidence shows disasters may result in the substitution from more- to less-effective family planning alternatives, with a resulting increase in unplanned births. In a retrospective study of 450 women in the aftermath of the 2006 earthquake in Yogyakarta, Indonesia, Hapsari et al. (2009) uncovered evidence of increased usage of less-effective contraceptive methods, which was accompanied by a subsequent rise in unplanned pregnancies in the year following the earthquake. Because the earthquake caused damage to over 200 health facilities, the authors conclude that shortage in access likely explains substitution of contraceptive methods. This evidence is consistent with studies that document increased fertility as a response to disruptions (availability, price) in contraceptive supply (Salas, 2013; Potter et al., 2013; White et al., 2012). In the case of service disruptions in Indonesia following the tsunami, Frankenberg et al. (2014) show lower uptake and higher discontinuation of family planning services among married women in heavily damaged areas. Changes in contraceptive usage were also predominant among women who desired more children; however, the authors do not find conclusive evidence of an increase in unintended fertility triggered by issues in supply of contraception.

Change in contraceptive use after a disaster may reflect changes in demand rather than supply. Loss of resources at the community and individual level may discourage couples from pregnancy and childbirth under precarious conditions (Carballo et al., 2005). On the other hand, child deaths may encourage parents to increase fertility, which, in turn, would reduce demand for family planning options (Preston, 1978; Zhu et al., 2013; Nobles et al., forthcoming).

In terms of linkages between demographic processes prompted by a disaster, the evidence base in the demography literature does not fully characterize fertility responses to large-scale mortality caused by disasters (Hill, 2004). Some of the best evidence comes not from disasters, but from wars and conflict. Decreases in fertility, either overall or for more- relative to less-affected subgroups, during conflicts accompanied by major social upheaval have emerged in a number of studies (Caldwell, 2004; Agadjanian and Prata, 2002; Blanc, 2004; Heuveline and Poch, 2007). In some instances, the end of the conflict is accompanied by a fertility increase. Famines are characterized by a similar temporal fertility pattern, as evidenced by studies from the Netherlands, China, and Bangladesh.

Relying on retrospective birth histories, Heuveline and Poch (2007) show a significant decrease in fertility during the years of the Khmer Rouge regime (1975–78), a sharp increase afterward (1978–80), and a subsequent decline. Based on this fertility pattern, the authors conclude that the significant increase in fertility was a direct response to heightened mortality during the conflict.

Isolating proximate mechanisms and disentangling whether fertility increases represent fundamental shifts in fertility desires or simply the realization of deferred reproduction are complicated when the precipitating events occur over multiple years and involve shifting spatial boundaries.

Nobles et al. (forthcoming) document similar behavior following the 2004 Indian Ocean tsunami in Indonesia. Using data from a population representative longitudinal study that includes observations before and after the 2004 Indian Ocean tsunami in Indonesia, they provide evidence of an increase in fertility between 2006 and 2009 in areas affected by the tsunami relative to unaffected areas. Two other studies explore fertility in the aftermath of high-mortality disasters. Finlay

(2009), using cross-sectional surveys, considers fertility for three earthquakes, each with death tolls of 15 000 or more. Comparing fertility before and after the earthquake for residents of areas affected by the earthquake with fertility of residents of areas that were not affected reveals greater postdisaster increases in fertility in affected areas. The same approach is adopted, with census data, to examine the impact of the 2003 Bam earthquake in Iran. The authors document a fertility decline in 2004, followed by a rise in 2005–07 (Hosseini-Chavoshi and Abassi-Shavazi, 2013).

Discussion

As discussed in the preceding sections, disasters have implications for the demographic processes of mortality, migration, and fertility, as well as for the linkages among them. The potential for feedback mechanisms among these processes is clear from a theoretical perspective, but relatively little empirical work has attempted to examine the interconnections and thereby test hypotheses about and provide a better understanding of the complex behaviors that underlie demographic processes. An exception is Heuveline's work on mortality and fertility in Cambodia during and after the decade marked by war and genocide (Heuveline, 1998, 2001; Heuveline and Poch, 2007). These investigations require data on multiple domains and behaviors of individuals, their families and, in many instances, communities that are collected over a time frame that is sufficiently long after the disaster for the behaviors to be revealed.

Unfortunately, such data are rarely available. Moreover, in the ideal world, the data would also be collected prior to the disaster although, as we have emphasized, disasters are in general unanticipated. However, creative investment in data infrastructure that exploits recent technological advances is likely to have a substantial payoff for science and policy. With the availability of high-quality satellite imagery, extensive administrative data sources, and population-representative longitudinal surveys being conducted on a regular basis across the globe, much of the key information needed to evaluate is already in place. In the case of the 2004 Indian Ocean tsunami, our international team of collaborators worked closely with Statistics Indonesia and followed up people who had been interviewed as part of the annual socioeconomic survey, SUSENAS, conducted about 10 months before the tsunami in all the districts along the west coast of Aceh and North Sumatra on the island of Sumatra. The 2004 SUSENAS served as the baseline for the Study of the Tsunami Aftermath and Recovery, which followed up the respondents annually for 5 years after the tsunami and then reinterviewed them again 10 years after the tsunami.

The demographic outcomes we have considered, as well as their interconnections, can be affected by policies and programs related to disaster preparedness and response. The development and enforcement of building codes, the creation of warning systems, evacuation plans, and temporary shelters, and well-trained first responders can alter the immediate consequences of disasters for mortality, health, and migration. Emergency assistance and longer-term recovery efforts can shape how outcomes unfold in a disaster's aftermath. The intersection of these policy and planning efforts with demographic phenomena has received relatively little attention in the literature to date. The questions are important, however, given the continued occurrence of major disasters and projected increase of people exposed to risk as a function of climate change.

See also: Afrocentric Approaches to Social Work; Community Resilience to Disasters; Environmental Stress and Health; Hazards and Disaster Research in Contemporary Anthropology; Mass Trauma: Psychopathological Effects across the Life Span; Natural Disasters: Health-Related Aspects.

Bibliography

Adeola, F.O., Picou, J.S., 2012. Race, social capital, and the health impacts of Katrina: evidence from the Louisiana and Mississippi gulf coast. Human Ecology Review 19 (1), 10–24.

Agadjanian, V., Prata, N., 2002. War, peace, and fertility in Angola. Demography 39 (2), 215–231.

Almond, D., 2006. Is the 1918 influenza pandemic over? Long-term effects of in utero influenza exposure in the post-1940 U.S. Population. Journal of Political Economy 114 (4), 672–712.

Armenian, H., Melkonian, A.K., Hovanesian, A.P., 1998. Long term mortality and morbidity related to degree of damage following the 1998 earthquake in Armenia. American Journal of Epidemiology 148, 1077–1084.

Bartels, S., VanRooyen, M.J., 2011. Medical complications associated with earthquakes. The Lancet S0140-6736 (11), 60887–60888.

Blanc, A.K., 2004. The role of conflict in the rapid fertility decline in Eritrea and prospects for the future. Studies in Family Planning 35 (4), 236–245.

Brown, R., April 2014. The Intergenerational Impact of Terror: Does the 9/11 Tragedy Reverberate into the Outcomes of the Next Generation? Working Paper.

Brown, R., Thomas, D., 2014. On the Long Term Effects of the 1918 Influenza Pandemic. Mimeo.

Browning, C.R., Feinberg, S.L., Wallace, D., Cagney, K.A., 2006. Neighborhood social processes, physical conditions, and disaster-related mortality: the case of the 1995 Chicago heat wave. American Sociological Review 71 (4), 661–678.

Brunkard, J., Namulanda, G., Ratard, R., 2008. Hurricane Katrina deaths, Louisiana, 2005. Disaster Medicine and Public Health Preparedness 2 (4), 215–223.

Caldwell, J.C., 2004. Social upheaval and fertility decline. Journal of Family History 29 (4), 382–406.

Camacho, A., 2008. Stress and birth weight: evidence from terror attacks. American Economic Review, Papers and Proceedings 98 (2), 511–515.

Cao, H., McFarlane, A.C., Klimidis, S., 2003. Prevalence of psychiatric disorder following the 1988 Yun Nan (China) earthquake – the first 5-month period. Social Psychiatry and Psychiatric Epidemiology 38, 204–212.

Carballo, M., Hernandez, M., Scheider, K., Welle, E., 2005. Impact of the tsunami on reproductive health. Journal of the Royal Society of Medicine 98, 400–403.

Cohan, C.L., Cole, S.W., 2002. Life course transitions and natural disaster: marriage, birth, and divorce following Hurricane Hugo. Journal of Family Psychology 16 (1), 14–23.

Committee on Increasing National Resilience to Hazards and Disasters; Committee on Science, Engineering, and Public Policy; The National Academies, 2012. Disaster Resilience: A National Imperative. http://www.nap.edu/catalog.php?record_id=13457 (accessed 18.06.14.).

Currie, J., Rossin-Slater, M., 2012. Weathering the storm: hurricanes and birth outcomes. NBER Working Paper 18070.

Dynes, R., Tierney, K., Fritz, C., 1994. Foreword: the emergence and importance of social organization: the contributions of E.L. Quarantelli. In: Dynes, R., Tierney, K. (Eds.), Disasters, Collective Behavior, and Social Organization. University of Delaware Press, Newark.

EM-DAT: The International Disaster Database, Center for Research on the Epidemiology of Disasters. www.emdat.be/database.

Finlay, J.E., 2009. Fertility Response to Natural Disasters: The Case of Three High Mortality Earthquakes. Policy Research Working Paper #4338. World Bank.

Foster, A., 1995. Prices, credit constraints and child growth in rural Bangladesh. Economic Journal 105 (430), 551–570.

Frankenberg, E., Nobles, J., Thomas, D., 2014. Contraception in the Aftermath of Natural Disaster. Population Association of America Meetings.

Frankenberg, E., Gillespie, T., Preston, S., Sikoki, B., Thomas, D., 2011. Mortality, the family and the Indian Ocean tsunami. The Economics Journal 221 (554), F162–F182.

Fritz, C.E., Marks, E.S., 1954. The NORC studies of human behavior in disaster. Journal of Social Issues 10 (3), 26–41.

Fusell, E., 2009. Post-Katrina New Orleans as a new migrant destination. Organization and Environment 22 (4), 458–469.

Fussell, E., Sastry, N., VanLandingham, M., 2010. Race, socioeconomic status, and return migration to New Orleans after Hurricane Katrina. Population and Environment 31, 20–42.

Galea, S., Maxwell, A.R., 2009. Methodological challenges in studying the mental health consequences of disasters. In: Neria, Y., Galea, S., Norris, F. (Eds.), Mental Health and Disasters. Cambridge University Press, New York.

Gray, C., Frankenberg, E., Gillespie, T., Sumantri, C., Thomas, D., 2014. Studying displacement after a disaster using large scale survey methods: Sumatra after the 2004 tsunami. Annals of the Association of American Geographers 104 (3), 594–612.

Groen, J.A., Polivka, A.E., 2010. Going home after Hurricane Katrina: determinants of return migration and changes in affected areas. Demography 47 (4), 821–844.

Guha-Sapir, D., Hoyois, P., 2012. Measuring the Human and Economic Impact of Disasters. Government Office for Science, Brussels.

Hamoudi, A., Frankenberg, E., Sumantri, C., Thomas, D., 2012. Impact of the December 2004 Tsunami on Birth Outcomes in Aceh, Indonesia.

Hapsari, E.D., Widyawati, Nisman, W.A., Lusmilasari, L., Siswishanto, R., Matsuo, H., 2009. Change in contraceptive methods following the Yogyakarta earthquake and its association with the prevalence of unplanned pregnancy. Contraception 79 (4), 316–322.

Herring, A., 2013. Sociology of disasters. In: Bobrowsky, P.T. (Ed.), Encyclopedia of Natural Hazards. Springer, Netherlands, pp. 926–935.

Heuveline, P., Poch, B., 2007. The phoenix population: demographic crisis and rebound in Cambodia. Demography 44 (2), 405–426.

Heuveline, P., 2001. The demographic analysis of mortality crises: the case of Cambodia, 1970–1979. In: Forced Migration and Mortality. NAP, Washington, DC, pp. 102–129.

Heuveline, P., 1998. 'Between one and three million': towards a demographic reconstruction of a decade of Cambodian history (1970–79). Population Studies 52 (1998), 49–65.

Hill, K., 2004. War, Humanitarian Crises, Population Displacement, and Fertility: A Review of Evidence. National Resource Council, Washington, DC.

Hosseini-Chavoshi, M., Abbasi-Shavazi, M.J., 2013. Demographic Consequences of the 2003 Bam Earthquake. International Conference on the Demography of Disasters. Australian National University.

Kamel, N., 2012. Social marginalization, federal assistance and repopulation patterns in the New Orleans metropolitan area following Hurricane Katrina. Urban Studies 49 (14), 3211–3231.

Klinenberg, E., 2002. Heat Wave: A Social Autopsy of a Disaster, first ed. University of Chicago Press, Illinois.

Kolbe, A.R., Hutson, R.A., Shannon, H., Trzcinski, A., Miles, B., Levitz, N., Puccio, M., James, L., Noel, J.R., Muggah, R., 2010. Mortality, crime and access to basic needs before and after the Haiti earthquake: a random survey of Port-au-Prince households. Medicine. Conflict and Survival 26 (4), 281–297.

Kreps, G.A., 2001. Sociology of disasters. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences, first ed. Elservier, Ltd, Oxford, pp. 3718–3721.

Lauderdale, D., 2006. Birth outcomes for Arabic-named women before and after September 11. Demography 43 (1), 185–201.

Lindstrom, D.P., Kiros, G.-E., 2007. The impact of infant and child death on subsequent fertility in Ethiopia. Population Research and Policy Review 26, 31–49.

Lu, X., Bengtsson, L., Holme, P., 2012. Predictability of population displacement after the 2010 Haiti earthquake. Proceedings of the National Academy of Sciences of the United States of America 109 (29), 11576–11581.

Maccini, S., Yang, D., 2009. Under the weather: health, schooling and economic consequences of early-life rainfall. American Economic Review 99 (3), 1006–1026.

Munasinghe, M., 2007. The importance of social capital: comparing the impacts of the 2004 Asian Tsunami on Sri Lanka, and Hurricane Katrina 2005 on New Orleans. Ecological Economics 64, 9–11.

Nakahara, S., Ichikawa, M., 2013. Mortality in the 2011 tsunami in Japan. Journal of Epidemiology 23 (1), 70–73.

National Hurricane Center, 2013. Tropical Cyclone Report: Hurricane Sandy. http://www.nhc.noaa.gov/data/tcr/AL182012_Sandy.pdf (accessed 24.06.14.).

Nishikiori, N., Abe, T., Cost, D.G.M., Dharmaratne, S.D., Kunii, O., Moji, K., 2006. Who dies as a result of the tsunami? – risk factors of mortality among internally displaced persons in Sri Lanka: a retrospective cohort analysis. BMC Public Health 73 (6), 1–8.

NOAA: Billion-Dollar Weather/Climate Disasters. www.ncdc.noaa.gov/billions/events.

Nobles, J., Frankenberg, E., Thomas, D. The effects of mortality on fertility: population dynamics after a natural disaster, forthcoming.

Norris, F.H., Elrod, C.L., 2006. Psychosocial consequences of disaster. A review of past research. In: Norris, F.H., Galea, S., Friedman, M.J., Watson, P.J. (Eds.), Methods for Disaster Mental Health Research. The Guildford Press, London, pp. 20–42.

Norris, F.H., Friedman, M.J., Watson, P.J., Byrne, C.M., Diaz, E., Kaniasty, K., 2002. 60,000 disaster victims speak: Part I. An empirical review of the empirical literature, 1981–2001. Psychiatry 65 (3), 207–239.

Paul, B.K., 2005. Evidence against disaster-induced migration: the 2004 tornado in north-central Bangladesh. Disasters 29 (4), 370–385.

Paul, S.K., Paul, B.K., Routray, J.K., 2012. Post-cyclone Sydr nutritional status of women and children in coastal Bangladesh: an empirical study. Natural Hazards 64, 19–36.

Pietrzak, R., Tracy, M., Galea, S., Kilpatrick, D.G., Ruggiero, K.J., Hamblen, J.L., Southwick, S.M., Norris, F.H., 2012. Resilience in the face of disaster: prevalence and longitudinal course of mental disorders following Hurricane Ike. PLoS ONE 7 (6), e389664.

Potter, J.E., Hopkins, K., White, K., Grossman, D., 2013. Evaluating the impact of abortion restrictions and drastic budget cuts for family planning in Texas. Working Paper, IUSSP.

Preston, S.H., 1978. The Effects of Infant and Child Mortality on Fertility. Academic Press, New York.

Quarantelli, E.L., 1989. Preliminary Paper No.136, the NORC Research on the Arkansas Tornado: A Fountainhead Study. Disaster Research Center University of Delaware, pp. 1–20.

Quarantelli, E.L., 2005. A social science research agenda for the disasters of the 21st century: theoretical, methodological and empirical issues and their professional implementation (Chapter 20). In: Perry, R., Quarantelli, E.L. (Eds.), What Is a Disaster? New Answers to Old Questions. International Research Committee on Disasters, pp. 325–395.

Ravelli, A.C.J., van der Meulen, Jan H.P., Michels, R.P.J., Osmond, C., Barker, D., Hales, C.N., Bleker, O.P., 1998. Glucose tolerance in adults after prenatal exposure to famine. Lancet 351, 173–177.

Ravelli, A.C.J., van der Meulen, Jan H.P., Osmond, C., Barker, D., Bleker, O.P., 1999. Obesity at the age of 50 years in men and women exposed to famine prenatally. American Journal of Clinical Nutrition 70, 811–816.

Rooij, W.H., Yonker, J.E., Painter, R.C., Roseboom, T.J., 2010. Prenatal undernutrition and cognitive function in late adulthood. Proceedings of the National Academy of Sciences 107, 16881–16886.

Roseboom, T.J., van der Meulen, Jan H.P., Osmond, Clive, Barker, David J., Blecker, Otto P., 2001. Adult survival after prenatal exposure to the Dutch famine 1944–1945. Paedeatric and Perinatal Epidemiology 15, 220–225.

Sacerdote, B., 2012. When the saints go marching out: long-term outcomes for student evacuees from Hurricanes Katrina and Rita. American Economic Journal: Applied Economics 4 (1), 109–135.

Salas, J.M.I., 2013. Consequences of withdrawal: free condoms and birth rates in the Philippines. Working paper.

Sastry, N., VanLandingham, M., 2009. Prevalence of and disparities in mental illness among pre-Katrina residents of New Orleans one year after Hurricane Katrina. American Journal of Public Health 99 (S3), 725–731.

Stein, Z., Susser, M., Saenger, G., Marolla, F., 1975. Famine and Human Development: The Dutch Hunger Winter of 1944–45. Oxford University Press, Oxford.

Tierney, K.J., 2007. From the margins to the mainstream? Disaster research at the crossroads. Annual Review of Sociology 33, 503–525.

Torche, F., 2011. The effect of maternal stress on birth outcomes: exploiting a natural experiment. Demography 48, 1473–1491.

Townsend, R.M., 1994. Risk and insurance in village India. Econometrica 62 (3), 539–591.

Wang, Y., Chen, H., Li, J., 2012. Factors affecting earthquake recovery: the Yao'an earthquake of China. Nat Hazards 64, 37–53.

White, K., Grossman, D., Hopkins, K., Potter, J.E., 2012. Cutting family planning in Texas. New England Journal of Medicine 367, 1179–1181.

Zaninetti, J.M., Colten, C.E., 2012. Shrinking New Orleans: post-Katrina population adjustments. Urban Geography 33 (5), 675–699.

Zhao, Z., Reimondos, A., 2012. The demography of China's 1958–61 famine: a closer examination. Population-E 67 (2), 281–308.

Zhu, H., Lei, H., Huang, W., Fu, J., Wang, Q., Shen, L., Wang, Q., Ruan, J., Liu, D., Song, H., Hu, L., 2013. Fertility in older women following removal of long-term intrauterine devices in the wake of a natural disaster. Contraception 87 (4), 416–420.

Demographic Measurement: General Issues and Measures of Fertility

Máire Ní Bhrolcháin, School of Social Sciences, University of Southampton, Southampton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article explains the rationale for demographic measures and the principles of their construction. The difference between crude and specific rates is outlined. The main indicators used in the demographic analysis of fertility, including measures of both level and timing, are defined and set out in tabular form. Synthetic cohort indicators of fertility routinely used to summarize sets of period rates are described, and their advantages and disadvantages discussed. The arguments for and against tempo adjusted fertility indicators are outlined.

Introduction

Demographic measures are designed to quantify the level, timing, and distribution of demographic phenomena. Their purpose is to allow comparison across time and space, and between subpopulations. The need for measures arises particularly because the frequency of demographic events and states varies substantially by age; and because populations vary in size and composition with respect to age and other factors influencing the frequency of events and states. Measures may be crude – events per 1000 population, taking no account of population composition – or specific by one or more factors. Rates specific by individual attributes – age, marital status, urban/rural residence – are generally preferable to crude rates. Specific rates may be used to describe e.g., the age profile of demographic event rates, or may be summarized into a standardized single-figure index – such as a total fertility rate (TFR) or period life expectancy – using several forms of standardization.

General Overview

Demographic analysis employs measures of events, states, attributes, and cumulative experience. Classically, a core demographic activity consists in estimating the frequency of vital events, births and deaths, and of related events – migration, marriage and divorce, and, in more recent times, the formation and dissolution of coresidential unions. Other events studied in demography include conception, pregnancy, spontaneous and induced abortion, the use of contraception, leaving the parental home, household formation and dissolution, family changes such as transition to and from lone parenthood, and widow(er)hood. States of demographic interest include marital status, whether legal or *de facto*, type of family or household to which individuals belong, enrolment in education or training, migrant status, and so on. The classic attributes of demographic interest are age and sex, along with ethnicity, country of birth, and socioeconomic characteristics such as educational attainment and social class. Cumulative experience is examined using proportions of ever having experienced an event such as marriage, or estimates of the average number of a specified event experienced in a lifetime, or by a specific age, for example, the mean number of births by a particular age.

Events can be thought of as changes in state and, correspondingly, states can be considered to result from the occurrence and nonoccurrence of events. For example, the event of first marriage changes a person's marital status from single to married, and a person who has married and has not dissolved their marriage is currently married. In addition, the concepts of 'cumulative experience' and 'state' are to some extent interchangeable, for example, a woman who has had n live births (cumulative experience) is said to be of parity n (state) (the term *parity* relates to *women*). Measures of cumulative experience that do not refer to individual states are also used in demography, and these will be considered presently. The current entry does not cover methods of collecting and classifying demographic data, but is concerned with the use of data once it has been collected and processed, ready for analysis.

Demography is concerned in general with aggregate phenomena at the population level, where populations vary from global to local in extent. The units whose events and states are aggregated can be individuals, couples, families, or households. Some demographic measures are, however, strictly aggregate in type. These include the population growth rate, population density per unit area, or population dependency ratios, all of which are meaningful at the aggregate level only. Demographic measures focus on level, timing, and distribution: the frequency of events or prevalence of states (level), the age at or time between events (timing), and distributional aspects such as age structure or the geographical distribution of a population.

The need for demographic rates and other measures is readily stated. Counts of population or events are of demographic interest, particularly in an applied context, and are frequently the key input or output in a policy context. But absolute numbers, whether population size or numbers of vital events, are not sufficient for most analytical purposes. Population analysis is conducted in terms of the underlying phenomena – mortality, fertility, migration, and so on – that determine population change. For example, population projections are usually carried out by the component method, applying assumptions about fertility, mortality, and migration rates, rather than by extrapolating births, deaths, or population numbers. Rates, proportions, and other indices are required also for comparative purposes, whether tracking time-trends or making comparisons between (sub)populations. Spatially and temporally, populations vary in size and also in structure, and so measures that abstract from size and structure are required to compare the demographic metabolism, so to speak, of different populations.

Demographic rates and measures are numerous. The variety stems both from the widely varying forms in which data are

available and from the nature of the phenomena themselves. Data may come from vital registration systems, from population registers, from censuses and surveys, from parish registers, and other historical sources; or from administrative records. These sources may differ in the kinds of information recorded, in questions asked, and in details published. Fertility is the area in which measures are probably most numerous, a result of the complexity of the phenomenon: births are repeatable events, they can occur to women inside and outside of marital or cohabiting unions, their order in a birth history can be of significance, and they can be associated with several dimensions of personal time – age, union duration, duration since previous birth, and so on. (The term order relates to births. 'Order' is the order of a birth in a woman's fertility history. 'Parity' is the number of births a woman has already had. First births (births of order 1) occur to childless women, i.e., women of parity 0. Second births (births of order 2) occur to women who have had one birth previously, i.e., women of parity 1, etc. It is incorrect in English to refer to a 'parity 1 birth,' though such mistaken usage can be found in the literature.)

Rates, Probabilities, Proportions, and Ratios

A demographic rate expresses the number of events occurring relative to person years at risk of the event in a defined population for a specified time and place. It may be expressed per person year, per 1000 person years, per 100 000 person years at risk, and so on. The denominator, person years at risk or person years lived, is often estimated by means of the mid-year population or by the average of two end-year population figures, each of these being an approximation to the average population during the year. Francophone demography has specific terms to distinguish between two types of rates. In type 1 rates, the denominator is confined to those who are at risk of the event represented in the numerator (i.e., those who have not yet experienced the event but who could do so in principle). In type 2 rates, no such restriction is applied: the denominator may include people who have already experienced the event. Type 1 rates are known in English as occurrence-exposure rates. There is no specific term in general use in English for type 2 rates, though they are occasionally referred to as incidence rates or accumulation rates. A probability is the likelihood that an individual will experience an event. It is estimated by the number of events occurring during a defined period or at a particular age, to a specified group divided by the numbers of individuals present at the start of the period or age. A proportion is defined in the usual way, as the number with a given attribute at a given time point divided by the total population in question at that time. Finally, demographic measures also take the form of ratios, for example, the sex ratio at birth is the ratio of the numbers of male to female births. Ratios generally refer to aggregates, although anthropometric measurements in ratio form relating to individuals are also found in the medical demographic literature.

Crude Rates and Degree of Specificity

Demographic measures vary in their level of detail. The simplest are crude rates, expressed as the number of events per 1000 (or other multiple) of the total population, without any further specificity. The two most common are the crude birth rate (CBR) – the number of births per 1000 population in a year – and the crude death rate – the number of deaths per 1000 population in a year. They give basic information but do not allow refined analysis. Crude rates are of value in setting out the basic demographic parameters of a population and are used for descriptive purposes, particularly when the data needed for more detailed measures are either unavailable or unreliable.

Demographic rates and other measures are influenced by the composition of the population in respect of any factor by which the frequency of the phenomenon under study varies. Most importantly, crude rates are influenced by the age structure of a population, because of the association between age and demographic event rates, and also by sex composition, since demographic event rates usually vary by sex. Accordingly, specific rates may be calculated by placing conditions on the numerator or the denominator, or on both. Specificity may be introduced for one or more factors. For example, age-specific rates may be influenced by composition in respect of marital status, duration in a particular state, parity (number of children already born to a woman), urban/rural residence, and so on. So, the analyst might choose to calculate rates specific for a number of dimensions. For some purposes, variation according to specific factors may itself be the focus of interest. In other contexts, variation with respect to one factor (e.g., age) is taken as given and is not the subject of study. If so, the analyst will wish to remove the compositional effect from the comparison to be made. Traditionally, one of two procedures is used to remove the influence of such (nuisance) factors. Rates may be standardized for the factors concerned or may be disaggregated progressively so as to arrive at rates specific for internally homogeneous groups. Standardization may be through the conventional direct or indirect methods or by calculating a synthetic indicator of some kind (see Section More Complex Indicators below).

Both procedures have disadvantages. Standardization, whether by the conventional direct or indirect methods, or by constructing synthetic indicators, is well known to be valid only where there are no interactions between the factors for which the standardization is carried out, or when there are no interactions between them and the categories, populations, etc. to be compared. Since such interactions are often found, straightforward standardization is inappropriate in many instances. Progressive disaggregation of rates has the disadvantage of producing potentially large numbers of rates that are not readily summarized and perhaps not readily interpretable. Modern methods of model fitting can in many instances provide a more general and rigorous solution to the routine need for standardization in demographic analysis, and can offer a considerable advance on the progressive disaggregation approach. Hoem (1987) and his predecessors have shown that indirect standardization can be seen as a first step in an iterative estimation procedure for intensity regression based on a multiplicative hazard model.

Measures of Fertility

Basic Measures

Level of Fertility

The definition of the most common fertility measures is set out in Table 1. All of them may be calculated either on a calendar

Table 1 Selected measures of the level of fertility

Measure and time reference[a]	*Definition*[b]	*Notes*[c,d]
Crude birth rate (CBR) Period; cohort version also possible but rare	$\frac{B}{P} \times 1000$	B = number of births P = total person-years lived
General fertility rate[e] (GFR) Period; cohort	$\frac{B}{\sum_{15}^{49} W_x} \times 1000$	W_x = number of woman years lived at age x or in age group x
Age-specific fertility rate (ASFR) Period; cohort	$F_x = \frac{B_x}{W_x} \times 1000$	B_x = births to women aged x or in age group x W_x = woman years lived at age x or in age group x
Total fertility rate[e] (TFR) Period; cohort	$\sum_{15}^{49} F_x$	F_x is the age-specific fertility rate defined above. When rates are for 5-year age groups, the sum is multiplied by 5.
Age-specific marital fertility rate Period; cohort	$\frac{B_{x,m}}{W_{x,m}} \times 1000$	$B_{x,\ m}$ = births to married women at age x or in age group x $W_{x,\ m}$ = woman years lived while married and at age x or in age group x
Age-specific non-marital fertility rate Period; cohort	$\frac{B_{x,u}}{W_{x,u}} \times 1000$	$B_{x,\ u}$ = births to unmarried women aged x or in age group x $W_{x,\ u}$ = woman years lived while unmarried at age x or in age group x
Proportion of births that are non-marital[f] Period; cohort	$\frac{B_u}{B} \times k$	B_u = births to unmarried women during a year or period. The multiplier k may be 100 or 1000. The indicator may also be specific for age.
Age-parity specific fertility rate (type 1 rate) Period; cohort	$\frac{B_{x,i}}{W_{x,i-1}} \times 1000$	$B_{x,\ i}$ = births of order i to women aged x or in age group x $W_{x,\ i-1}$ = woman years lived while of parity $i-1$ and aged x or in age group x
Age-order specific fertility rate (type 2 rate) Period; cohort	$F_{x,i} = \frac{B_{x,i}}{W_x} \times 1000$	
Parity progression ratio Cohort; period version differs: see text	$a_i = \frac{\sum_{j=i+1}^{j=m} N_j}{\sum_{j=i}^{j=m} N_j}$	a_i = *parity progression of order i = probability of progressing from i-th to i + 1-th birth* N_j = number of women who have j births m = maximum number of births occurring
Gross reproduction rate[e] (GRR) Period; cohort	$\sum_{15}^{49} F_x^f$	F_x^f is The age-specific rate of female births. The sum is multiplied by 5 if 5-year age groups are used.
Net reproduction rate[e] (NRR) Period; cohort	$\sum_{15}^{49} F_x^f L_x^f$	L_x^f is The average number of person-years aged x in the stationary female life table with specified mortality, with a radix of 1.

[a]All measures given here relate either to a calendar year, time period, or cohort. For clarity, subscripts denoting the time reference are omitted, as most of the measures can be specified in either period or cohort form.
[b]The measures defined in female terms here can be defined analogously for men, but the male versions are rarely used. While summation is to 44 or 49 in a female index, summation could extend to later ages in male indices.
[c]In the specification of demographic rates, person-years lived, whether overall, or at a particular age or in a given state, is the true denominator. Where the rate relates to a year or period, events and person-years lived are those occurring during the year or period in question, at age x, where specified. A cohort rate is based on events experienced and person-years lived by the cohort concerned, at age x, where specified. Person years lived are normally estimated by the mid-year population.
[d]Throughout the table, age x denotes age x in completed years or, equivalently, age x at last birthday.
[e]When summation is over the reproductive ages, the lower limit of the summation is usually 15, but may also be 10, and the upper limit may be 44 rather than 49.
[f]This indicator was formerly known as the illegitimacy ratio.

period or on a cohort basis, though a CBR for a cohort is rarely seen. The article assumes throughout that period or cohort, data on births and population estimates are available. Where this is not the case, indirect methods for estimating population fertility parameters are available (*see* Demographic Techniques: Indirect Estimation).

The CBR is widely used in cross-national comparison, particularly since it requires relatively simple data – an estimate of total births and of total population. The general fertility rate is more specific, and restricts the denominator to those at risk of experiencing a birth, that is, women of childbearing age, usually 15–44 years or 15–49 years. Age-specific fertility rates introduce a further level of detail by age and are key indicators – they reveal both the level and the pattern of childbearing by age and are also the basis for calculating the total fertility rate, a widely used summary indicator described below. Age-specific fertility rates may be disaggregated further – by marital status, by duration of partnership or marriage, in place of or jointly with age, and by order of birth.

Fertility rates that are specific by order of birth are more precise and for many purposes more useful than the more widely available age-specific fertility rates. They are more informative about the nature of change through time because time-trends often vary by order of birth, particularly in modern contracepting populations or where the use of contraception is on the increase. Birth rates by order can be either type 1 or type 2 rates (see Section Rates, Probabilities, Proportion, and Ratios above). A type 1 (or occurrence-exposure) rate is more precise and restricts the numerator to births of a specified order, say order i ($i = 1,2, 3, \ldots$), and the denominator to women of parity $i-1$ ($i = 1, 2, 3, \ldots$), that is, only those women at risk of a birth of order i. Where age-specific also, this is known as an age-parity specific fertility rate: it is specific both by the age and the parity of the woman. Age-parity specific rates are conditional in the sense that they are conditional on the event not having

previously occurred. A type 2 order-specific fertility rate restricts the numerator to births of order i, but includes women of all parities in the denominator. For an age group, this gives the order-specific fertility rate for the particular age. To illustrate, suppose we have 1000 women aged 25–29 years in a particular year, 500 of whom are of parity 0 (childless) and 500 of parity 1 and above (women who are already mothers), and that 50 first births occur to these women during the year. The age-parity specific first birth rate at age 25–29 years is 50/500*1000 or 100 first births per 1000 childless women aged 25–29 years, per annum – a type 1 rate. The age-order specific first birth rate at age 25–29 years is 50/1000*1000 or 50 first births per 1000 women aged 25–29 years, per annum – a type 2 rate.

Both order-specific and parity-specific rates may also be specific by duration since marriage or start of union, or since some other event such as a previous birth (for second and higher order births), or since leaving education, or after arrival in a country of destination. A duration scale can either replace, or be combined with, age. Fertility rates specific by order of birth based on vital registration data will be inaccurate wherever, as sometimes occurs, the order of birth as recorded in a particular system is not based on the woman's complete birth history; population registers or survey retrospective fertility histories are an alternative source.

Fertility rates may vary by parity, and in modern contracepting populations always do so. Because any given overall level of fertility may be reached through differing patterns of parity-specific fertility, measures known as parity progression ratios (PPRs) were developed in the early 1950s. They represent the probability of making each transition in the family building process. The PPR of order i, a_i, is defined as the probability of proceeding from parity i to parity $i+1$. It can be obtained as the proportion of women of final parity i and above, who, by the end of childbearing, have had at least $i+1$ births. Here $i = 0, 1, 2 \ldots m-1$, where m is the maximum number of births to any woman in the population. These ratios express, in other words, the probability that a woman who has had a given number of live births, will have at least one further birth. In their original form, PPRs are calculated on a cohort basis but they are meaningful for any group of women aged 45 years plus. Methods for estimating these for women who have not reached the end of childbearing have also been proposed (Brass and Juarez, 1983).

Timing of Fertility

The principal measures of fertility timing are set out in Table 2. They are somewhat less numerous than indicators of fertility level. The mean and median age at birth are the most basic factors. Somewhat more precise is the mean or median age-specific by order of birth; mean or median age at first birth; at second birth; and so on. Mean and median ages may be obtained from either period or cohort data, and as with measures of the level of fertility, time series of these will generally differ. Whether or not specific by order, the period mean or median age at birth can be calculated in crude or standardized form, and published sources do not always indicate which of these is presented. The period crude mean age at birth (or at first birth, second birth, etc.) is obtained simply as the arithmetic mean age of all women having a (first, second, etc.) birth in a given period. The crude mean age at birth is widely used but has the disadvantage, in

Table 2 Selected measures of fertility timing

Measure and time reference[a]	*Definition*[b]	*Notes*[c]
Mean age at birth (crude) Period; cohort	$\frac{\sum_{15}^{49} xW_x^b}{\sum W_x^b} + 0.5$	W_x^b = the number of women aged x last birthday having a birth Where age is grouped, x = mid-point of each age group
Standardized mean age at birth Period	$\frac{\sum_{15}^{49} xF_x}{\sum_{15}^{49} F_x} + 0.5$	Where age is grouped, x = mid-point of each age group
Median age at birth (crude) Period; cohort	Median age of women having a birth in a period or, within a cohort, median age of women at childbirth	
Standardized median age at birth Period	Median of the age-specific fertility distribution	
Mean age at i-th birth (crude) Period; cohort	$\frac{\sum_{15}^{49} xW_{x,i}^b}{\sum_{15}^{49} W_{x,i}^b} + 0.5$	W_x^b = women aged x last birthday having a birth of order i Where age is grouped, x = mid-point of each age group
Standardized mean age at i-th birth Period	$\frac{\sum_{15}^{49} xF_{x,i}}{\sum_{15}^{49} F_{x,i}} + 0.5$	$F_{x,i}$ = age-order specific birth rate of order i, during a year or period (defined in Table 1)
First birth interval		
Cohort; for synthetic period version, see text	Duration from marriage (or start of union) to first birth	Less useful where moderate to high proportions of first births occur outside of a union
Second, third, etc. birth interval Cohort; for synthetic period version, see text	Duration from first to second birth, second to third birth and so on	Caution: Cannot be calculated from the mean ages at births of each order, unless specifically for women with $i+$ births; see text.

[a]All measures given here relate either to a calendar year, time period, or cohort. For clarity, subscripts denoting the time reference are omitted, as most of the measures can be specified in either period or cohort form. In a period index, events relate to those occurring in the year or period, at age x where specified. In the cohort case, events are those occurring to members of the cohort concerned, at age x where specified.
[b]The measures defined in female terms here can be defined analogously for men, but the male versions are rarely used.
[c]Throughout the table, age x denotes age x in completed years or, equivalently, age x at last birthday.

a period context, of being influenced by the age structure of women of childbearing age, which can vary through time and from one country to another. Far better for the purposes of period analysis, although more demanding of data since population denominators are required to calculate age-specific rates, is the standardized mean age at birth, or at first birth, second birth, and so on. Standardization is unnecessary in a cohort framework, which is self-standardizing, although care should be taken in comparisons of cohorts widely differing in mortality during the childbearing years. The standardized mean age at birth is obtained by weighting by the relative period age-specific fertility rates at each age (see **Table 2**) and so is, in fact, the mean of the age-specific fertility distribution. Similarly, the standardized mean age at *i*-th birth uses as weights, the relative age-order-specific fertility rates (type 2, not type 1 rates). The standardized mean age at birth is not a pure measure of fertility timing since it is influenced not only by timing but also by the overall level of fertility. If populations A and B have identical (standardized) mean ages at each order of birth but if proportionately more women in A than in B have births of higher orders, the standardized mean age at birth will be older in A than in B. The mean age at last birth is occasionally used in historical demographic literature as an indirect indicator of the presence of birth control. It can be calculated, without bias, only for women who have reached the end of reproduction (age 45 or 50 years).

Fertility tempo is also measured by the mean or median duration of birth intervals. The first birth interval is the time from marriage, or start of informal union, to the first birth. This interval may be of limited utility where a large proportion of births occur outside of formal or informal marital unions. Later intervals are obtained as the duration from first to second birth (second birth interval), from second to third birth (third birth interval), and so on. Some care is needed in measuring such intervals. Mean birth intervals are sometimes estimated erroneously from the ages of women at births of successive orders of birth. This is incorrect because the ages at successively higher orders of birth are based on women of differing ultimate family sizes. The difficulty arises because women who ultimately have larger families are usually younger at births of any given order than are those with smaller completed families. The *i*-th birth interval is calculated instead as (1) the time from birth $i-1$ (or start of union where $i = 1$) to birth i among women with i or more births, or equivalently as (2) the age at birth i minus the age at birth i-1 (or at the start of union when $i = 1$) among women with i or more births. Where individual level data are available, means or medians may then be calculated. The median may be preferred to the mean birth interval, since birth interval distributions are typically positively skewed. Synthetic or hypothetical mean or median birth intervals for time periods may be obtained from the life tables used in constructing PPPRs (see Section Period Parity Progression Ratios below), thus giving a period measure of birth timing analogous to, and supplementing, the standardized mean age at birth (Ní Bhrolcháin, 1987).

The mean length of a generation is of importance in stable population theory. It is defined as the time it takes a stable population to grow by the factor net reproduction rate (NRR). It is represented in stable population theory by the symbol T and is given by $T = (\ln NRR)/r$, where r is the intrinsic growth rate of a stable population (*see* Preston et al., 2001; Population Dynamics: Theory of Stable Populations).

More Complex Indicators

The TFR and Associated Indices

Since a set of specific fertility rates for a period can be numerous, single-figure indices that summarize them are convenient and often a practical necessity. The classic method of condensing them is the synthetic or hypothetical cohort indicator. A period synthetic cohort indicator is obtained in one of two ways: by summing age-specific rates (type 2) across ages, or by combining occurrence-exposure rates (type 1) multiplicatively across ages in a life table calculation. The most widely used synthetic indicator in the fertility arena is the classic period TFR, obtained by adding the age-specific fertility rates of a given year or period across ages 15–49 years. The most familiar example of the multiplicatively produced synthetic indicator is period life expectancy (*see* Demographic Measurement: Nuptiality, Mortality, Migration, and Growth). The TFR is routinely interpreted as representing the mean family size in a hypothetical cohort of women who experienced the age-specific fertility rates of the period in question throughout their childbearing years.

Closely related to the classic TFR are the gross reproduction rate (GRR) and the NRR. The GRR is simply the TFR confined to female births, and can be seen as the average number of daughters a woman would have if she experienced the age-specific female birth rates of a particular period. The NRR is obtained by modifying the GRR to take account of female mortality up to the end of the reproductive age range (see **Table 1**). It can be interpreted as the average number of daughters who would be born to a newly born cohort of girls in a stable population in which age-specific female fertility and mortality rates remain fixed (see Population Dynamics: Theory of Stable Populations). In a theoretical, stable population, the NRR represents the extent to which generations replace themselves. Under stable conditions, an NRR of <1, 1, or >1 means that successive generations are, respectively, declining, stationary, or growing in numerical size. The GRR is roughly half the size of the TFR. The NRR is less than the GRR, the gap between them depending mainly on the level of female mortality at ages under 50 years: the higher the level of mortality, the greater the disparity between the GRR and the NRR. The TFR, GRR, and NRR can also be specified on a male basis, using male age-specific fertility rates and male survivorship, and the results will in general differ from the female values. Female rates are, however, more widely available and more reliable and are used universally; however, male rates are used when a specific focus is on male fertility or in theoretical work.

Difficulties with the TFR

A key advantage of the TFR is that it is standardized for age and that it is relatively undemanding of data. But it has shortcomings as a measure. A theoretical difficulty with the TFR and with synthetic cohort indicators in general, is that they express period events in terms of lifetime experience. Mean family size is conceptually meaningful for a cohort but the mean family size of a period has no concrete meaning (Ní Bhrolcháin, 1992). The synthetic cohort interpretation can be avoided by thinking of

the TFR as a convenient summary of the level of a phenomenon in a particular period. Nevertheless, the difficulty remains that time-trends in fertility are not homogeneous among those at risk, but can vary by parity, by age, and by duration. No single-figure indicator can accurately represent the multiple dimensions of change. Finally, the TFR is highly subject to variations in the timing of childbearing. When the pace of childbearing is accelerating, the period TFR is above the corresponding cohort mean family sizes, and when the pace is decelerating, the period TFR falls below the associated cohort values. The TFR is more volatile than the comparable cohort series for two main reasons: it takes no account of the past fertility experience of the women at risk at a particular time and, relatedly, it is not standardized for the parity distribution of the population at risk. The TFR of a given period is influenced both by the behavioral parameters of the period and by the impact of the fertility of previous years on the distribution of exposure in the period.

One time-honoured solution, due originally to Hajnal (1947), is not to rely exclusively on period analysis to interpret trends but to use cohort fertility in conjunction with period information: cohorts, he suggested, are the best guide to long-run trends and parity-specific period indicators better describe recent and current trends. Further approaches to handling the shortcomings of the classic TFR are described in the two sections that follow.

Period Parity Progression Ratios

The period parity progression approach to fertility measurement is an alternative to the classic TFR, and goes a long way toward addressing the sensitivity of the TFR to changing tempo. The period parity progression ratio (PPPR) is defined for each order of birth. It is a synthetic indicator representing the PPR that would result if a cohort were to experience during their lifetime the parity-specific, duration-specific, and/or age-specific birth probabilities of a particular period. An indirect method of estimating these is due originally to Henry (1953, translated 1980) and a direct method to Ní Bhrolcháin (1987). The PPPR is calculated by cumulating parity specific rates (type 1) multiplicatively, in life table format, for each order of birth. The parity specific rates may be specific by duration, by age for first birth and duration for later births, by age alone, or by both age and duration (Ní Bhrolcháin, 1987; Murphy and Berrington, 1993; Rallu and Toulemon, 1994). PPPRs largely correct for differences in composition resulting from past experience and have the advantage also that they do not produce results that would be impossible in a true cohort. The PPPR approach improves on the classic TFR not only in standardizing for parity and in removing tempo effects, but also in measuring multiple facets of the fertility of a given period, and in providing multiple indicators of time-trends.

The PPPRs can be used to obtain an estimate of the period TFR that is alternative to the standard TFR, generated as

$$\mathrm{TFR(PPPR)} = a^p_{0,t} + a^p_{0,t} \times a^p_{1,t} + a^p_{0,t} \times a^p_{1,t} \times a^p_{2,t} + \cdots + a^p_{0,t} \times a^p_{1,t} \times a^p_{2,t} \times \cdots a^p_{m-1,t}$$

where $a^p_{i,t}$ is the PPPR of order i for period t, and m is the maximum number of births occurring. The TFR (PPPR) has the advantage over the conventional TFR that it is standardized for parity and also removes much of the timing influence to which the classic TFR is subject. Like the classic TFR, however, it is a one-dimensional measure.

Tempo Adjustment

The decades since the 1970s in developed countries have been marked by very substantial shifts to later ages at childbearing. The result is a sizable downswing in the TFR that has made it even less trustworthy than usual as a guide to longer-run fertility levels. And while the period parity progression approach provides a useful counterweight to tempo shifts, it is fairly demanding of data, requiring parity specific rates by age and/or duration. One methodological response is to attempt to remove the influence of tempo change from the TFR and so produce a tempo adjusted TFR, denoted as TFR* (Bongaarts and Feeney, 1998). The approach has been much discussed, and has been developed further in a number of directions (Kohler and Philipov, 2001; Kohler and Ortega, 2002; Ediev, 2008). Demographic opinion remains divided on the case for and utility of tempo adjustment. Difficulties with tempo adjustment include lack of parity specificity of the TFR*, erratic behavior of TFR* under realistic theoretical scenarios of change, lack of evidence validating tempo adjusted measures as estimators of underlying cohort values, and that the tempo component is integral to period fertility measures where the aim is to explain either period change or cross-national differences (Lesthaeghe and Willems, 1999; Van Imhoff and Keilman, 2000; Schoen, 2004; Ní Bhrolcháin, 2011). If the objective is to seek accurate and efficient estimators of the final fertility of cohorts that have not yet completed their childbearing, relying on tempo adjustment alone is limiting. Recent approaches move away from using tempo adjustment to characterize a particular period's fertility, and instead adopt a broader framework aiming explicitly to estimate the completed fertility of cohorts (Cheng and Lin, 2011; Cheng and Goldstein, 2012). This alternative strategy appears to be promising.

See also: Demographic Techniques: Data Adjustment and Correction; Demographic Techniques: Indirect Estimation; Life Table; Period and Cohort Analysis in Demography; Population Dynamics: Theory of Stable Populations.

Bibliography

Bongaarts, J., Feeney, G., 1998. On the quantum and tempo of fertility. Population and Development Review 24, 271–291.

Brass, W., Juarez, F., 1983. Censored parity progression ratios from birth histories. Asia and Pacific Census Forum 10, 5–13.

Cheng, P.C.R., Goldstein, J.R., 2012. On the Possibilities of Predicting Cohort Fertiltiy Measures from Period Fertility Measures: Theory and Empirical Evidence. Population Association of America, San Francisco.

Cheng, P.C.R., Lin, E.S., 2011. How Well Can Period Fertility Measures Help to Forecast Cohort Fertility Measures? Cross-country Evidence. Population Association of America, Washington, DC.

Ediev, D., 2008. On the Theory of Distortions of Period Estimates of the Quantum Caused by the Tempo Changes. European Demographic Research Paper. Vienna Institute of Demography, Vienna.

Hajnal, J., 1947. The analysis of birth statistics in the light of the recent international recovery of the birth-rate. Population Studies 1, 137–164.

Henry, L., 1953, translated 1980. Fertility of Marriage: A New Method of Measurement (fécondité des mariages. Nouvelle méthode de mesure. Ined, Paris 1953) UN/ESCAP Population Translation Series. UN ESCAP.

Hoem, J., 1987. Statistical analysis of a multiplicative model and its application to the standardization of vital rates: A review. International Statistical Review 55, 119–152.
Kohler, H.P., Ortega, J.A., 2002. Tempo-adjusted period parity progression measures, fertility postponement and completed cohort fertility. Demographic Research 6 (7), 145–190.
Kohler, H.P., Philipov, D., 2001. Variance effects in the bongaarts-feeney formula. Demography 38, 1–16.
Lesthaeghe, R., Willems, P., 1999. Is low fertility a temporary phenomenon in the European union? Population and Development Review 25, 211–228.
Murphy, M., Berrington, A., 1993. Constructing period parity progression ratios from household survey data. In: Ní Bhrolcháin, M. (Ed.), New Perspectives on Fertility in Britain. HMSO, London.
Ní Bhrolcháin, M., 1987. Period parity progression ratios and birth intervals in England and Wales, 1941–1971 – a synthetic life table analysis. Population Studies 41, 103–125.
Ní Bhrolcháin, M., 1992. Period paramount? A critique of the cohort approach to fertility. Population and Development Review 18, 599–629.
Ní Bhrolcháin, M., 2011. Tempo and the TFR. Demography, 841–861.
Preston, S., Heuveline, P., Guillot, M., 2001. Demography. Measuring and Modeling Population Processes. Oxford University Press, Oxford.
Rallu, J.L., Toulemon, L., 1994. Period fertility measures: the construction of different indices and their application to France, 1946–89. Population: An English Selection 6, 59–93.
Schoen, R., 2004. Timing effects and the interpretation of period fertility. Demography 41, 801–819.
Van Imhoff, E., Keilman, N., 2000. On the quantum and tempo of fertility: comment. Population and Development Review 26, 549–553.

Demographic Measurement: Nuptiality, Mortality, Migration, and Growth

Máire Ní Bhrolcháin, School of Social Sciences, University of Southampton, Southampton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This is the second of two articles on demographic measurement, the first article being entitled *Demographic Measurement: General Issues and Measures of Fertility*. This second article presents the principal ways in which nuptiality, mortality, migration, and population growth are currently measured in demography. The section on nuptiality covers indices of both level and timing of marriage, divorce, and cohabitation. Crude and specific rates are presented along with synthetic or hypothetical cohort measures. The definitions of the main types of nuptiality and mortality measures are given in tabular form.

Introduction

In the article *Demographic Measurement: General Issues and Measures of Fertility*, we saw that demographic measures are required for comparisons through time and space, and between subpopulations. Well-specified measures are needed because the rate of demographic events varies both with age and with other dimensions of personal time, and because populations vary in size and in composition with respect to age and other factors influencing demographic event rates. Type 1 and type 2 rates were distinguished. In type 1, or occurrence-exposure, rates the denominator is confined to people who are at risk of the event in question. In type 2 rates, no such restriction is placed on the denominator. Synthetic or hypothetical cohort measures were introduced, their purpose being to summarize an array of specific rates pertaining to a period and to estimate the hypothetical lifetime consequences of the set of rates obtaining in a calendar period. Some demographic indicators measure 'flow' – the movement into and out of the population or into particular states – whereas others measure the 'stock,' or cross-sectional state of the population at a given time. Some further types of demographic behavior are considered in this article.

Marriage, Cohabitation, and Union Disruption/Divorce

Indicators of marriage and divorce are much less numerous than those of fertility. Measures of both level and timing are considered here.

Level

The most basic of the 'flow' indicators are the crude marriage rate, defined as the number of marriages per 1000 population, and the general marriage rate, defined as the number of marriages per 1000 population of marriageable age, usually 15 years of age and above, but the legally allowable minimum age at marriage varies between countries (**Table 1**). These rates are influenced both (a) by the structure of a population by age and sex, since marriage rates vary by age and sex, and (b) by the proportions currently married, since those already married are not at risk of a further marriage (though would be so if they were to divorce). To deal with the second of these, a general marriage rate may be defined per 1000 unmarried persons. Age–sex variation in rates is handled in the conventional way by defining marriage rates specific by age and sex. Age–sex specific marriage rates may be refined by restricting the denominator to the unmarried. Further detail may also be introduced by calculating marriage rates specific by order of marriage. Thus, in an occurrence-exposure rate of first marriage (type 1), the numerator of the age–sex specific first marriage rate is confined to first marriages and the denominator is restricted to people who are single (i.e., never married). A type 2 rate of first marriage has the same numerator as the type 1 rate, but the denominator includes all persons in the age–sex group in question. Remarriage rates are generally of type 1, i.e., they express the number of remarriages, by age and sex, per 1000 divorced and widowed persons, but they too can be calculated in type 2 form – remarriages per 1000 persons in an age–sex group, irrespective of marital status. All rates may be calculated on either a period or a cohort basis. The crude and general marriage rates can be standardized for age, using conventional standardization methods.

As an adjunct to, or sometimes in place of, period or cohort marriage rates, the cross-sectional distribution of marital status by age and sex may be used to depict the nuptiality of a population. These are 'stock' figures and have the advantage that they can be obtained from any census or survey in which individuals' current marital status is recorded and so do not require vital registration information or annual population estimates by age and sex. They reflect the prevalence of marriage, divorce, and widow(er)hood resulting from past rates of marriage and marital dissolution, and so do not represent marriage rates, etc., in the population concerned. However, marriage rates in preceding years can be reconstructed from cross-sectional surveys if information on retrospective marriage histories is collected.

Measuring the frequency of nonmarital cohabitation and of other types of informal partnerships is in principle similar to marriage but less straightforward. This is both because vital registration data are not available and because cohabitation and marriage behavior may need to be considered jointly, for a full picture (see Murphy, 2000 for a discussion of some methodological issues). Because informal partnerships are not recorded in vital registration systems, and that population estimates by *de facto* marital status are not usually available, the traditional approaches to measuring marriage often cannot be adopted. Current status data and retrospective questions in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31002-9

Table 1 Selected measures of marriage

Measure and time reference[a]	*Definition*	*Notes*[b,c]
Level		
Crude marriage rate		
Period	$\frac{M}{P} \times 1000$	$M =$ number of marriages during a specified year or period. $P =$ total person-years lived.
General marriage rate		
Period	$\frac{M}{P_{15+}} \times 1000$	$P_{15+} =$ person-years lived at ages 15 and above
Age–sex specific marriage rate (type 2)		
Period; cohort	$N_x^f = \frac{M_x^f}{P_x^f} \times 1000$	$M_x^f =$ women marrying at age x or in age-group x. $P_x^f =$ woman-years lived at age x or in age-group x. Analogous definition for men.
Age–sex specific first marriage rate (type 2)		
Period; cohort	$\frac{M_{x,1}^f}{P_x^f} \times 1000$	$M_{x,1}^f =$ single women marrying aged x or in age-group x. Analogous definition for men.
Age–sex specific first marriage rate (type 1)		
Period; cohort	$\frac{M_{x,1}^f}{P_{x,u}^f} \times 1000$	$P_{x,u}^f =$ woman-years lived while single and aged x or in age-group x Analogous definition for men.
Total marriage rate		
Period	$\sum_{15+} \frac{M_x^f}{P_x^f}$	Where age is in 5-year groups, the sum is multiplied by five. Analogous definition for men.
Total first marriage rate		
Period	$\sum_{15+} \frac{M_{x,1}^f}{P_x^f}$	Where age is in 5-year groups, the sum is multiplied by five. Analogous definition for men.
Gross nuptiality		
Period; cohort	$1 - \prod_{x=15+} (1 - q_x^f)$	$q_x^f =$ probability that a single woman aged x exactly will marry before $(x+1)$th birthday. Age range used may vary. Analogous definition for men.
Net nuptiality		
Period; cohort	Cumulative proportions ever marrying in a double-decrement life table with marriage and mortality as decrements	Age range used may vary. Defined for males and for females.
Timing		
Mean age at marriage (crude)		
Period; cohort	$\frac{\sum_{15+} xM_x^f}{\sum M_x^f} + 0.5$	Analogous definition for men.
Mean age at marriage (standardized)		
Period	$\frac{\sum xN_x^f}{\sum N_x^f} + 0.5$	Analogous definition for men.
Singulate mean age at marriage		
Period	$15 + \frac{\sum_{15}^{49} S_x^f - 35S_{50}^f}{1 - S_{50}^f}$ Assuming no marriage before age 15	$S_x^f =$ the proportion of women aged x who are single. S_{50}^f can be estimated by the mean of the proportions single in age-groups 45-49 and 50-54. Where 5-year age-groups are used, the first term in the numerator is multiplied by 5. Analogous definition for men.

[a]All measures given here relate either to a calendar year, time period, or cohort. For clarity, subscripts denoting the time reference are omitted.
[b]In the specification of demographic rates, person-years lived is the true denominator. Where the rate relates to a year or period, events and person-years lived are those occurring during the year or period in question, at age x where specified. A cohort rate is based on events experienced and person-years lived by the cohort concerned, at age x where specified. Person-years lived are normally estimated by the mid-year population.
[c]Throughout the table, age x denotes age x in completed years or, equivalently, age x at last birthday.

cross-sectional censuses and surveys are therefore a primary source of information. The absence of good quality information on informal unions limits assessment of levels and trends in *de facto* unions particularly in developed nations, where the prevalence of cohabitation has increased substantially since the 1970s, and in some countries of the Caribbean and Latin America, where visiting or consensual unions have long been common (Fussell and Palloni, 2004). Among the indicators used to evaluate the frequency of informal unions are: the proportion currently cohabiting, the proportion who have ever cohabited, the proportion cohabiting with their marriage partner prior to marriage, and the proportion of current unions that are cohabiting unions. The frequency of informal unions evaluated by all of these indicators tends to vary with age, and social differentials can vary both with age and the measure used (Ni Bhrolchain and Beaujouan, 2013).

The crude divorce rate is analogous to other crude rates, being defined as the number of divorces per 1000 persons in a population. It is influenced in a compositional sense by the proportions married by age and sex, and by the distribution of marriages by duration, since divorce rates are closely tied to duration of marriage. Where data are available, a slightly more detailed measure is preferable: divorces per 1000 existing marriages in a population. As with other events, divorce rates are more informative if specific by age and sex: divorces in an age–sex group per 1000 married persons in that age–sex group. However, divorce rates specific by duration of marriage are usually preferred to age-specific rates since divorce risk varies substantially by marriage duration. Divorce rates specific by both age at and duration of marriage, as well as by sex, are especially informative since the risk of divorce is strongly related to age at marriage, being higher among those marrying at younger ages. The divorce rate is sometimes expressed, mistakenly, as the number of divorces in a year per 1000 marriages in that year. This is erroneous because the divorces in a given year stem from marriages that took place in preceding years, and so divorces should be related in some way to the stock of marriages that are at risk of divorce, that is past marriages rather than those of the current year. Divorce rates expressed per 1000 married population are a crude way of achieving this, and rates by marriage cohort or duration of marriage improve on this. A further method is given in the section on synthetic indicators. It is separation rather than divorce that marks the end of informal unions, and measures such as the proportion of cohabitations separating within, e.g., 5 or 10 years are used instead. For many practical purposes, separation is also of greater interest as marking the *de facto* end of a marriage. Information on separation is obtained from retrospective histories of partnership collected in demographic surveys.

Where birth cohort information is available, the cumulative proportions ever having married by specified ages are useful indicators of both the level and the pattern of marriage within and across generations. Birth cohort proportions ever having divorced or ever having remarried can be obtained for the same purpose, though in this case the figures may be influenced by the cohort proportions ever having married or ever having divorced, respectively. The proportion ever having divorced or separated calculated by year of marriage or partnership, specific by duration of marriage/union, overcomes this issue. Similarly, the proportions ever having remarried may be calculated for divorce cohorts, or alternatively the proportions in a new partnership by duration since separation.

Synthetic or Hypothetical Cohort Measures

The total marriage rate is often used to summarize the overall level of marriage for each sex in a population. It is obtained by adding the (type 2) age–sex specific marriage rates and represents the hypothetical average number of marriages in a cohort of women or men if they experienced the age–sex specific marriage rates of a particular year throughout their lifetime. The total first marriage rate is defined similarly as the sum of the age–sex specific (type 2) first marriage rates and represents the proportion of women or men who would eventually marry, if they were subject through their lifetime to the age–sex specific first marriage rates of a given period. Both of these summary measures have the same strengths and weaknesses as the analogous conventional total fertility rate (TFR; *see* Demographic Measurement: General Issues and Measures of Fertility). On the one hand, they are standardized for age and so are an improvement on the crude marriage rate. However, like the conventional TFR they are influenced by timing changes and they do not take account of past history, since they are not standardized for marital status. As a result, a period-specific total first marriage rate can, and has been observed to, exceed one first marriage per person – an impossible result that could not apply in a real cohort. Like the TFR, however, the total marriage rate can be considered as a simple indicator of the level of marriage in a period; like the TFR, it is deficient as a summary indicator whenever change in marriage rates differs between age-groups.

As in the case of fertility, multiplicative summary indicators are available to overcome some of the deficiencies of additive measures: gross nuptiality and net nuptiality. Gross nuptiality is so called because it assumes that all survive to the latest age considered – usually to age 50. It is obtained, for each sex, by combining the age-specific first marriage probabilities of the unmarried (type 1) multiplicatively and computing the proportions ever marrying by 50 years of age (*see* Life Table). Net nuptiality, on the other hand, takes into account that death may intervene before marriage and is obtained by constructing a double-decrement life table, with marriage as one source of decrement and death while single as a second source of decrement (*see* Life Table). Gross nuptiality may be interpreted as the proportion of a cohort who would marry at a given period's age-specific rates if there were no mortality before 50 years of age, and net nuptiality as the proportion who would marry by 50 years of age, at a given period's age-specific marriage and mortality rates. Both gross and net nuptiality require more detailed information than is needed to obtain the total marriage rate – namely first marriages by age and sex together with population estimates by age, sex, and marital status in the case of gross nuptiality, and additionally, mortality rates by sex, age, and marital status in the case of net nuptiality. As with life table quantities in general, they are standardized for age structure.

The level of divorce in a population may also be summarized from period rates using hypothetical cohort indicators. Analogous to the total marriage rate is the male or female total divorce rate, defined as the sum of age–sex specific type 2

divorce rates and representing the average number of divorces per person if the age–sex specific divorce rates of the period in question were to be experienced throughout a person's lifetime. As before, a multiplicative life table-based indicator is preferable to the additive measure. The most common indicator used for the purpose is a life table estimate of the proportions who would divorce by a specified duration of marriage, at current duration-specific rates. As with analogous multiplicative indicators, this not only standardizes for the distribution by marriage duration but also takes some account of past history. It is probably the best available summary measure of the divorce propensity implied by current divorce rates, but is also subject to tempo effects. If successive marriage cohorts are moving to an earlier pattern of divorce, the high divorce rates of early divorcing couples may be combined with the high divorce rates of later-divorcing couples who had lower divorce rates at early durations, to produce an overall hypothetical proportion ever divorcing that is higher than that of any real cohort. Again, the measure can be specified ignoring mortality or incorporating mortality as a competing risk (Wilson and Smallwood, 2008).

Timing

Age at marriage may be calculated from period data in either crude or standardized form, though sources do not always indicate which measure is presented and, as with births, the terms 'crude' and 'standardized' are not in universal use in this context. The crude mean age at marriage is simply the mean age of those marrying in a period. The standardized mean age at marriage is analogous to the standardized mean age at birth and is the mean of the schedule of age-specific marriage rates (type 2). A crude and standardized median age at marriage may also be obtained. It is often useful to distinguish marriage age by the order of the marriage since the proportion of remarriages among all marriages can vary substantially through time or cross-nationally. Thus, we have the mean or median age at first marriage or at remarriage, either crude or standardized. Cohort mean ages at marriage can also be calculated and, as before, do not need to be standardized. Analogous indicators can be obtained for informal unions; the most useful are the mean age at first cohabitation and the mean age at first union. Mean and median ages at marriage can also be obtained from either gross or net nuptiality tables. Care is needed since the e_x column of a nuptiality table does not give the expected age at or years to marriage. Rather the column gives the expected number of years lived while single at and after age x. Since some never marry, this is not the same as the expected number of years to marriage among those who marry (*see* Life Table).

Where direct data are not available, an indirect procedure developed by Hajnal (1953) can be used to estimate the average age at marriage from cross-sectional proportions single (never married) by age. It gives what is termed the singulate mean age at marriage. Based on the hypothetical cohort principle, the procedure is appropriate where marriage rates by age have been reasonably stable and assumes that neither migration nor mortality rates are associated with marital status. Should the proportions single be found to increase over any part of the age range, this indicates that one or more of the assumptions does not hold and so the calculation is not valid. Brief details are given in **Table 1**; see also Preston et al. (2001: Section 4.6).

Mean and median ages at divorce may be calculated. Of particular interest in relation to divorce is the mean/median, etc., duration of marriage at which the divorce occurs. A distinction should be made between the duration of marriages in general – those terminated by either divorce or death – and the duration of those marriages that end in divorce. Crude or standardized versions of these indicators may be obtained; for most purposes, standardized indicators are to be preferred.

Mortality

Table 2 sets out selected mortality measures. These require data on deaths and population estimates. Where such data are not available, indirect methods for estimating mortality are available (*see* Demographic Techniques: Indirect Estimation).

The crude death rate (CDR) – deaths per 1000 population – is widely used as a basic measure of the level of mortality. It is strongly influenced by the age composition of a population because of the wide variation in death rates by age. Age-specific mortality rates – deaths at a particular age per 1000 population of that age – are at the next level of detail, usually specific by sex. Age is usually in 5-year groups and for refined analysis, in single year of age rates. Rates specific by single years of age are of particular importance at the younger and older parts of the age range, since in these age ranges differences in rates between adjacent single years of age may be sizable. Age may be grouped in a variety of ways. However, even where quinary age-groups are used, ages 0 and 1–4 are often distinguished, because of the relatively high death rate of infants, even in low-mortality countries.

The measurement of mortality is simplified somewhat by the fact that death occurs ultimately to all and occurs only once. These two together mean that crude and age-specific death rates are necessarily of type 1 (*see* Demographic Measurement: General Issues and Measures of Fertility). Death rates are so closely and systematically associated with age, and to a lesser extent sex, that specificity by age and sex is adequate for many demographic purposes. Other dimensions of personal time are sometimes used: e.g., duration since diagnosis, or since treatment, in epidemiological studies. Mortality rates may also be specific by social class, race/ethnicity, income, area of residence, marital status, and other socioeconomic characteristics. Analysis by such factors is of interest in examining social and economic variations in health and for actuarial and public health purposes.

The infant mortality rate (IMR) is reported and published extensively, partly as a demographic indicator but also as a proxy for socioeconomic development. It is defined as the number of deaths of infants (children under 1 year of age) per 1000 births in a given year. As the denominator is not an estimate of the person-years at risk of the event, it is, strictly, not a true rate. It has a correctly specified counterpart in the infant death rate, the number of deaths less than 1 year of age per 1000 person-years lived under age 1. The IMR is a convenient measure since it can be obtained from simple counts of vital events and does not require population estimates by age. But it

Table 2 Selected mortality measures

Measure and time reference[a]	*Definition*	*Notes*[b,c]
Crude death rate		
Period; cohort also possible but rare	$\frac{D}{P}$	$D =$ the number of deaths occurring during a particular year/period. $P =$ person-years lived.
Age-specific death rate[d]		
Period; cohort	$M_x = \frac{D_x}{P_x}$	$D_x =$ deaths at age x. $P_x =$ person-years lived at age x or in age-group x.
Infant mortality rate		
Period; cohort	$\frac{D_0}{B}$	$D_0 =$ deaths of infants under 12 months. B = births.
Infant death rate		
Period; cohort	$\frac{D_0}{P_0}$	$P_0 =$ person-years lived aged under 1 year.
Maternal mortality ratio (sometimes referred to as the maternal mortality rate)		
Period; cohort	$\frac{D_m}{B}$	$D_m =$ maternal deaths.
Maternal mortality rate		
Period; cohort	$\frac{D_m}{P^f_{15-49}}$	$P^f_{15-49} =$ woman-years lived at ages $15-49$ (age limits may vary).
Directly standardized death rate[e]		
Period	$\sum_x M^i_x c^s_x$	$M^i_x =$ age-specific death rate in index population i. c^s_x is the proportion of the standard population who are aged x.
Standardized mortality ratio[e] *(SMR)*		
Period	$\mathrm{SMR}^i = \frac{D^i}{\sum_x n^i_x M^s_x}$ or equivalently $\mathrm{SMR}^i = \frac{CDR^i}{\sum_x c^i_x M^s_x}$	$D^i =$ total deaths in index population. $n^i_x =$ numbers aged x in the index population. $M^s_x =$ the age-specific death rate in the standard population. $\mathrm{CDR}^i =$ crude death rate in the index population. $c^i_x =$ proportion who are aged x in index population. The SMR is often multiplied by 100.
Indirectly standardized death rate[e]		
Period	$\mathrm{SMR}^i \times \mathrm{CDR}^s$	CDR^s is the crude death rate in the standard population used in obtaining the SMR.

[a]All measures given here relate either to a calendar year, time period, or cohort. For clarity, subscripts denoting the time reference are omitted.
[b]For person-years lived, see **Table 1**, note 2.
[c]Throughout the table, age x denotes age x in completed years or, equivalently, age x at last birthday.
[d]Age-specific death rates, and associated mortality measures such as life expectancy, are generally also specific by sex, as mortality rates differ between the sexes.
[e]Each of these standardized indices can be compared meaningfully only with the standard population or with others standardized on the same standard population.

will approximate less well to the infant death rate when births are appreciably seasonal or that there are sizable year-on-year fluctuations in births. Several mortality rates are distinguished during the first year of life, both because of the steep decline in death rates over the first 12 months of life and because of the changing role of endogenous (genetic, intrauterine, perinatal) and exogenous (environmental, external) causes of death. Endogenous causes predominate in the earliest postpartum period, with exogenous factors growing in importance thereafter. The perinatal mortality rate is defined as the number of late fetal deaths plus the number of deaths within 1 week of birth per 1000 total late fetal deaths plus live births. The neonatal mortality rate is defined as the number of deaths within 1 month (28 days) of birth per 1000 live births, and the postneonatal mortality rate as deaths between 28 days of birth and 1 year of age per 1000 live births. The probability of dying between birth and the fifth birthday is widely referred to in the current demographic literature as the child mortality rate, though the latter is more correctly defined as the number of deaths of children under 5 years of age per 1000 person-years in the age-group. The probability definition (also sometimes termed child mortality risk) has become widespread because under-five mortality is of particular interest in high-mortality, less-developed societies, for which the indirect methods used to evaluate child mortality estimate a cohort probability rather than a rate (*see* Demographic Techniques: Indirect Estimation). High-mortality societies do not usually have good vital registration statistics and it can be difficult to formulate accurate assumptions about the average number of years lived by those who die, and thus to obtain an estimate of person-years lived.

Maternal deaths are those that occur during pregnancy or within 42 days of the end of the pregnancy, due to a cause related to the pregnancy or a condition aggravated by pregnancy, including abortion-related deaths. Maternal mortality

is most often represented by the maternal mortality ratio: the number of maternal deaths per 100 000 live births in a period. This indicator is referred to by many authors and in many sources as the maternal mortality rate, although it is not in fact a true rate. Strictly, the maternal mortality rate is the number of maternal deaths per 100 000 women of reproductive age in a period, and some authors employ the term in this sense. It should be clear from the context which definition is in use. Indirect methods are available to estimate maternal mortality.

Mortality specific by cause of death, of which maternal mortality is an instance, is the principal way in which mortality indicators are disaggregated beyond age and sex. The International Classification of Disease, produced and revised from time to time by the World Health Organization, is the standard scheme by which cause of death is classified. Measures of mortality by cause of death are of two kinds. Cause-specific mortality rates refer to the number of deaths from a particular cause during a period per 100 000 population, and may be specific also by age and/or sex. A multiplier larger than 1000 is normally used for the purpose, since deaths from any one cause will usually be relatively small in number. Cause-specific mortality ratios, on the other hand, relate to the percentage of all deaths that result from a particular cause, either overall or by age and sex, and are of value where information on population denominators is inaccurate or unavailable. Proportional mortality analysis is used particularly in the study of occupational mortality.

Age Standardization

Because of the substantial variation in death rates by age, and in age structure across time and place, cross-national, areal, or time comparisons of the level of mortality require that death rates are standardized for age. The objective is to remove the effect of age structure from the comparison. Direct standardization, in which a standard age distribution is employed, yields a directly standardized death rate. Indirect standardization, based on a standard set of age-specific mortality rates, yields two indicators: the standardized mortality ratio (SMR) and the indirectly standardized death rate. The SMR gives an index figure expressing the level of mortality in the index population relative to that of the standard population, set to 1 or to 100: it is simply the ratio of observed to expected deaths in the index population, with expected deaths based on the standard population's age-specific rates. Multiplying the SMR of the index population by the CDR in the standard population gives the indirectly standardized death rate in the index population. Direct standardization requires that the age-specific mortality rates are known and precisely estimated in the populations to be standardized. Where the age-specific rates are either unknown or subject to sizable errors, the indirect method must be used. Statistical theory singles out a version of indirect standardization as optimal when a multiplicative model holds for the force of mortality.

Life Expectancy and Life Tables

Life expectancy at birth, e_0, is a summary of the overall level of mortality in a population. Because death occurs universally and removes an individual permanently from the population, any increase in death rates at any age is necessarily reflected in a shorter life expectancy, while a decline in death rates at any age is necessarily accompanied by an increase in life expectancy, all else equal. The expectation of life is influenced both by the level of mortality in a population and by the pattern of mortality by age. As a result, two populations with the same overall life expectancy may differ in their patterns of age-specific death rates. Life expectancy is calculated from a life table (*see* Life Table), and so is an age-standardized measure that can be compared fairly between populations and time periods. Life tables are more commonly based on period mortality rates, but are also compiled for birth cohorts. Where based on a period life table, the expectation of life and other life table functions are synthetic cohort measures and do not, strictly, represent the experience of any particular birth cohort. In practice, however, where mortality improvement is gradual, as in developed countries such as the US and Sweden, period life expectancy has been found to approximate to that of the cohort born about 40–50 years before (Goldstein and Wachter, 2006). That mortality change tends to be steady and unidirectional distinguishes it from period fertility, for which an approximate empirical correspondence between lagged synthetic period and ultimate cohort parameters cannot be established. Adjustment of period life expectancy for changing tempo has been proposed and has been subject to much debate (see especially Barbi et al., 2008).

The life table is fairly demanding of data since it requires information on age-specific death rates and hence on both deaths and population estimates by age. Because detailed data of this kind are often unavailable for historical populations and for less-developed nations, sets of empirically based model life tables have been developed corresponding to varying levels and patterns of mortality (Murray et al., n.d.). Such compilations allow estimates of the level and/or age pattern of mortality to be made from information that is insufficient in itself to determine these (*see* Demographic Techniques: Indirect Estimation). For further details of the measurement of mortality see Preston et al. (2001), Rowland (2003), and Siegel (2012).

Migration

Along with births and deaths, migration is the final component of population change. Migration is classified according to whether it is into an area (in-migration) or out of an area (out-migration). The sum of in- and out-migration is termed gross migration and the difference between them, in-migration minus out-migration, is known as net migration. For many purposes it is the absolute values of these quantities that are of interest. Crude and specific rates are also used. The crude in-migration (or immigration) rate is the number of in-migrants per 1000 population and the crude out-migration (or emigration) rate the number of out-migrants per 1000 population. The crude gross migration rate is the sum of these and the crude net migration rate is the difference between them.

There are, however, difficulties with these indicators. In particular, in-migrants are from populations outside the destination area, and a rate of in-migration per 1000

population in the destination area clearly does not link in-migrants with the true population at risk of such migration; it is therefore not a true rate. In addition, net migration is somewhat contested as a category. There is no such thing as a 'net migrant,' and looking for the influences on net migration rather than on the component in-flows and out-flows may be misconceived (Rogers, 1990). On the other hand, it is argued that net migration is useful where information on gross flows is lacking (Smith and Swanson, 1998).

Further indicators based on transition data – having changed residence between specified two time points – are in use. Cumulative migration experience can be represented by whether a person has ever migrated from his or her place of birth or during a specified time, e.g., the last year, the last 5 years. Synthetic summary indicators of lifetime migration experience are also in use, analogous to the fertility equivalent – the gross migraproduction rate, which ignores mortality, and migration expectancy or the net migraproduction rate, which takes mortality into account. Like other demographic events, migration is patterned by age and measures standardized for age can be constructed. For a detailed exposition and discussion of migration measurement, see Bell et al. (2002).

Population Growth

The difference between the crude birth rate (CBR) and CDR is known as the crude rate of natural increase and represents the rate at which a population is growing or declining in a year purely as a result of vital events. Since migration also influences population growth, the overall, crude population growth rate is CBR − CDR ± the crude net migration rate. Where annual crude rates are not available, the population growth rate can be inferred from population size at two time points, say t years apart. If we assume that the population has been growing continuously at a fixed rate over that period, this implies an exponential model of growth, in which $P_t = P_0\, e^{rt}$, where P_0 is the initial population and P_t is population size t years later, and r is the annual growth rate. Under these assumptions, the annual growth rate can be estimated as $r = \ln(P_t/P_0)/t$. Note that the exponential model of growth does not necessarily represent, as is often assumed, extremely rapid growth. Whether exponential growth is rapid depends on the growth rate. For example, after 50 years a population growing exponentially at the rate of 1% per annum would have increased by 65% but the increase would be just 5% if it was growing exponentially at 0.1% per annum.

The growth rate of a stable population – one with a fixed set of age-specific mortality and fertility rates, and closed to migration – is known as the intrinsic rate of natural increase. It is a theoretical quantity. It can be calculated from the fertility and mortality rates of an empirical population. It should be interpreted with care since it is the hypothetical growth rate that would apply in a theoretical stable population with the vital rates assumed, and not the growth rate of the real population from which the vital rates are drawn. Further indicators relevant to the theoretical analysis of population growth are outlined in the article Demographic Measurement: General Issues and Measures of Fertility: the gross and net reproduction rates (GRR and NRR). While the GRR and NRR may be calculated for real populations, the same proviso must be borne in mind: the NRR reflects growth prospects only under strict hypothetical conditions. The same caution is necessary in relation to the TFR: in low-mortality countries, a TFR of 2.1 is widely referred to as 'replacement level fertility,' that is, the level of fertility necessary for population replacement in the long run. But this is potentially misleading. Low-mortality populations with a TFR of below 2.1 can and do continue to grow because of in-migration, declines in mortality, changing fertility, and/or population momentum. Equally, low-mortality populations with a TFR of 2.1 or more can and do decline because of out-migration, changes in fertility and/or mortality, and/or population momentum.

See also: Adult Mortality in Developing Countries; Adult Mortality in Industrialized Societies; Demographic Techniques: Indirect Estimation; Infant and Child Mortality in Industrialized Countries; Infant and Child Mortality in the Less Developed World; Internal Migration in Industrialized Countries; Internal Migration: Developing Countries; International Migration; Life Table; Partnership Formation and Dissolution in Western Societies.

Bibliography

Barbi, E., Bongaarts, J., Vaupel, J.W. (Eds.), 2008. How Long Do We Live? Springer, Berlin.

Bell, M., Blake, M., Boyle, P., Duke-Williams, O., Rees, P., Stillwell, J., Hugo, G., 2002. Cross-national comparison of internal migration: issues and measures. Journal of the Royal Statistical Society (A) 165, 435–464.

Fussell, E., Palloni, A., 2004. Persistent marriage regimes in changing times. Journal of Marriage and the Family 66, 1201–1213.

Goldstein, J.R., Wachter, K.W., 2006. Relationships between period and cohort life expectancy: gaps and lags. Population Studies 60, 257–269.

Hajnal, J., 1953. Age at marriage and proportions marrying. Population Studies 7, 111–136.

Murphy, M., 2000. The evolution of cohabitation in Britain, 1960-95. Population Studies 54, 43–56.

Murray, C.J.L., Ahmad, O.B., Lopez, A.D. Salomon, J.A., n.d. Who system of model life tables. GPE Discussion Paper Series. Geneva: EIP/GPE/EBD, World Health Organization.

Ni Bhrolchain, M., Beaujouan, E., 2013. Education and cohabitation in Britain: a return to traditional patterns? Population and Development Review 39, 441–458.

Preston, S., Heuveline, P., Guillot, M., 2001. Demography. Measuring and Modeling Population Processes. Oxford University Press, Oxford.

Rogers, A., 1990. Requiem for the net migrant. Geographical Analysis 22, 283–300.

Rowland, D.T., 2003. Demographic Methods and Concepts. Oxford University Press, Oxford.

Siegel, J.S., 2012. The Demography and Epidemiology of Human Health and Aging. Springer, London and New York.

Smith, S.K., Swanson, D.A., 1998. In defense of the net migrant. Journal of Economic and Social Measurement 24, 249–264.

Wilson, B., Smallwood, S., 2008. The proportion of marriages ending in divorce. Population Trends 131, 28–36.

Demographic Models

Yu Xie, University of Michigan, Ann Arbor, MI, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 5, pp. 3449–3452, © 2001, Elsevier Ltd.

Abstract

Demographic models can have two meanings, one broad and one narrow. In its broad meaning, demographic models refer to all mathematical, statistical, forecast, and microsimulation models that are applied to studies of demographic phenomena. In its narrow meaning, demographic models refer to empirical regularities in age patterns of demographic events. This article is concerned with demographic models in the narrow definition. Demographic models are widely used (1) to improve data quality and (2) to compare demographic outcomes and processes across populations or subpopulations. Both parametric and semiparametric specifications have been proposed for modeling age patterns of demographic events, giving rise to parametric models and semiparametric models. Successful applications of both types of models are found in research on mortality, nuptiality, and fertility. As an integral part of formal demography, demographic models have been linked closely to mathematical demography. In recent decades, however, statistical demography has played an increasingly important role in demographic models.

The term 'demographic models' can have two meanings, one broad and one narrow. In its broad meaning, demographic models refer to all mathematical, statistical, forecast, and microsimulation models that are applied to studies of demographic phenomena. In its narrow meaning, demographic models refer to empirical regularities in age patterns of demographic events. Demographic models in the broad definition can be found in various entries related to demography (*see*, among others, Multistate Transition Models in Demography; Event History Analysis: Applications; Population Dynamics: Theory of Stable Populations; Population Forecasts; Microsimulation in Demographic Research). This article is concerned with demographic models in the narrow definition.

Age Patterns of Demographic Events

In a classic statement, Hauser and Duncan (1959) defined demography as 'the study of the size, territorial distribution, and composition of population, changes therein, and the components of such changes' (p. 2). In the Hauser–Duncan definition of demography, the study of population changes goes hand in hand with the study of the components of population changes. This is necessitated by the need to decompose a population into components and then study changes in the components before arriving at an overall understanding of the changes in the population. The most elementary, and also the most important, form of population decomposition is by sex and age. Since men are only indirectly involved in reproduction, demographic analysis is often simplified by focussing on women. Such simplification is called 'one-sex' modeling (*see* Population Dynamics: Two-Sex Demographic Models), as is commonly seen in fertility and nuptiality models. In mortality models, however, men and women are always kept distinct, and there are substantial mortality differentials in favor of women.

The treatment of age is of significant concern in demographic research. Without exception, the occurrence of all demographic events is age-dependent. Here, the correct interpretation of age-dependency is one of life-course, that is, the likelihood of the occurrence of an event changes as a person (or a cohort) ages. This is true even though most demographic methods and models use cross-sectional data, capitalizing on age-gradients of vital rates in any given population. The use of period-based data usually is necessitated by the lack of cohort-based data.

Demographic models of age schedules are developed on the observation that age patterns of demographic events often show some regularity. Two cautionary notes are in order. First, age regularity is not universal, either across space or over time. Second, all that is assumed is empirical regularity; theoretical reasons behind such regularities are typically neither well established nor well understood. Nonetheless, demographic models capitalizing on empirical regularities are very useful in practice and may provide the basis for theoretical work. A brief discussion of the main uses of demographic models is given as follows.

Use of Demographic Models

Demographic models are intended to summarize empirical regularities in age patterns of demographic events, ideally in simple mathematical formulas. Such models can prove very useful in demographic research. One type of use of demographic models concerns data quality. For example, demographic models may be used to detect and correct faulty data, impute missing data, and allow researchers to infer from partial data. Not surprisingly, this use of demographic models is often found in historical demography and research in less developed countries, where quality data are scarce.

The second type of use is substantive and commonly it is found in research with a comparative focus, be it over time (trends) or across societies, regions, or subpopulations. Age-

schedule models are also useful for actuarial calculations of life-insurance premiums and premium reserves and for the computation of national or regional population forecasts (e.g., Lee and Tuljapurkar, 1994). In demographic applications, cross-group variations in age patterns are typically parameterized as functions of two components: a shape component to capture age effects and a modification component reflecting group membership. Ideally, we would like the second component to be as parsimonious as possible. When demographic models are fully parametric, the two components are integrated, with the functional form being the shape component and the parameter values being the modification component. However, parametric models are not the norm in demography. Parametric models and semiparametric models will be discussed separately below, with concrete examples taken from research on mortality, nuptiality, and fertility. Coale and Trussell (1996) give a more detailed account of some of the models.

Parametric Models

Gompertz (1825) is accredited for the discovery that mortality rates in human populations increase nearly as an exponential function of age. This regularity holds true only after early adulthood (ages 25–30). According to 'Gompertz' law,' the force of mortality (or hazard) is a parsimonious loglinear function of age

$$\mathrm{In}(\mu_x) = A + Bx \qquad [1]$$

where μ_x denotes the force of mortality at age x. Naturally, A is the parameter characterizing the level of mortality at early ages and B is the parameter characterizing the rate of increase in mortality with age. While very parsimonious, the Gompertz model does not always fit empirical data well. Other researchers have modified the model either by adding additional terms or by changing the functional form. For example, Makeham (1867) altered the simple loglinear relationship of eqn. [1] by adding a time invariant term intended to capture cause-specific 'partial forces of mortality.' Heligman and Pollard's (1980) three-component model is another such extension.

In studying age patterns of first marriages among women, Coale (1971) found a basic form that is common in different populations. Through trial-and-error, Coale was able to characterize the typical age pattern of first marriage by a double-exponential curve

$$r_x = (0.174)\exp[-4.411\exp(-0.309x)] \qquad [2]$$

where r_x denotes the hazard rate (or risk) of first marriage at age x. This mathematical function was shown later to be approximated closely by the density function corresponding to the sum of distributions representing demographic processes (Coale and McNeil, 1972).

To fit observed age patterns of fertility, Brass (1960) proposed a model using polynomial functions. Although Brass' parametric model is very flexible, it requires the estimation of four unknown parameters. Since fertility rates typically are given in a limited number of 5-year intervals, this method leaves very few degrees of freedom for evaluating goodness-of-fit. An evaluation of various parametric models based on empirical data is given by Hoem et al. (1981).

There are three notable problems with parametric models. First, parametric models often do not fit observed phenomena. Demographers have dealt with such empirical deviations through (1) restricting the applicable age range and (2) allowing for further parameterization. Both solutions have surfaced in research using the Gompertz model.

The second major disadvantage associated with parametric models is the lack of behavioral interpretations for key parameters. This problem arises from the fact that almost all parametric models have resulted from exercises of curve-fitting. Parametric models may reproduce observed age patterns of demographic events, but theoretical interpretation of involved parameters is often unclear.

Finally, parametric models are not always convenient to use for comparing populations or subpopulations, even though this was one of the motivations for developing them in the first place. This problem is apparent, for example, in the case of a polynomial model. When several parameters in a polynomial function differ between two populations, it is difficult to characterize one population as having higher or lower rates than the other population.

Semiparametric Models

In response to these problems with parametric models, semiparametric models have been developed. Semiparametric models are similar to parametric models in specifying parsimonious mathematical functions but differ from parametric models in allowing age-dependency to be unconstrained and subject to empirical estimation, that is, semiparametric models do not impose any global constraint limiting the age pattern to the rigidity of a parametric mathematical function. Instead, the age pattern is estimated freely and empirically from observed data or calculated from external sources. One manifestation of the semiparametric approach is the use of model schedules or model tables. While allowing for the flexibility in a common age function, model tables place constraints on the variations in the age pattern across populations or subpopulations. Such constraints are often motivated by substantive knowledge of demographic phenomena.

In the area of mortality studies, for example, model life tables have been in wide use, primarily as a tool for correcting faulty data and estimating missing data. In essence, a model life table allows for a typical age schedule of mortality shared by a set of populations that differ mainly in their levels of mortality. The age pattern is flexible and empirically determined over all ages but constrained across populations; and cross-population variability lies in the overall level of mortality. The earliest model life table was developed by the United Nations (1955) for all national populations. Subsequently, Coale and Demeny (1966) added more flexibility by identifying four regional model life tables based on distinctive age patterns of mortality and refined the technique for constructing model life tables.

As an alternative, Brass (1968) developed a system of relational life tables, in which a standard mortality schedule is modified by two parameters through a logit equation

$$\text{logit}(p_{xj}) = \alpha_j + \beta_j \text{logit}(p_x^*) \quad [3]$$

Here p_{xj} is the probability of survival to age x in population j, p_x^* is the probability of survival to age x in a selected standard table, and α_j and β_j are parameters characterizing population j. In the Brass system, α determines the level of survival, whereas β changes the shape of survival parametrically.

In fertility research, Henry (1961) posited the existence of natural fertility with a peculiar age pattern. According to him, for physiological and social factors, different populations not exercising fertility control may have different levels of fertility but exhibit the same age pattern. Coale and Trussell (1974) later formalized this idea into the *Mm* method for studying age patterns of marital fertility. Let the marital fertility rate for the xth age of the jth population be denoted by r_{xj}. The Coale–Trussell *Mm* method parsimoniously describes the cross-population variation in age patterns of marital fertility with

$$r_{xj} = n_x M_j \exp(m_j v_x) \quad [4]$$

Here n_x is the standard age pattern of natural fertility; v_x is a typical age-specific deviation of controlled fertility from natural fertility; and M_j and m_j measure the jth population's fertility level and control, respectively. Equation [4] then states that marital fertility can be modeled as a product of natural fertility and a factor reflecting fertility control. The former is represented by $n_x M_j$; the latter by $\exp(m_j v_x)$. Coale and Trussell estimated n and v and suggested that these age schedules be taken as known in all applications. This approach has been superseded by Xie (1990) and Xie and Pimentel (1992), who suggested that n and v be regarded as parameters to be estimated along with M and m from any set of empirical fertility rates r_{xj}.

Statistical Demography versus Mathematical Demography

The construction and derivation of parametric and semi-parametric demographic model age schedules have been accomplished through mostly mathematical and sometimes graphical methods. From the outset, demographic modeling has been regarded as part of mathematical demography. Since the 1980s, however, a mainline statistical approach has played an increasingly important role in the development of demographic models (e.g., Clogg and Eliason, 1988; Hoem, 1987; Xie, 1990, Xie and Pimentel, 1992) for a number of good reasons.

First, the advancement of demography has brought with it more, richer, and better data in the form of sample data, and the use of sample data requires statistical tools. Treating sample data as exactly known quantities induces the danger of contamination by sampling errors. Second, while the method of disaggregation commonly used in mathematical demography may easily lead to inaccurate estimation due to small group sizes, statistical methods more efficiently utilize covariates to examine group differences. Third, because observed data may sometimes be irregular or simply missing, statistical models can help smooth or impute data. Conversely, the strength of empirical techniques developed around demographic models (e.g., indirect estimation, age patterns of fertility, model life tables, etc.) is that they provide descriptions of age patterns that can be utilized to improve statistical analysis.

See also: Demographic Measurement: General Issues and Measures of Fertility; Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Demography: History Since 1900.

Bibliography

Brass, W., 1960. The graduation of fertility distributions by polynomial functions. Population Studies 14, 148–162.

Brass, W., 1968. The Demography of Tropical Africa. Princeton University Press, Princeton, NJ.

Clogg, C.C., Eliason, S.R., 1988. A flexible procedure for adjusting rates and proportions, including statistical methods for group comparisons. American Sociological Review 53, 267–283.

Coale, A.J., 1971. Age patterns of marriage. Population Studies 25, 193–214.

Coale, A.J., Demeny, P., 1966. Regional Model Life Tables and Stable Populations. Princeton University Press, Princeton, NJ.

Coale, A.J., McNeil, D.R., 1972. The distribution of age of the frequency of first marriage in a female cohort. Journal of the American Statistical Association 67, 743–749.

Coale, A.J., Trussell, T.J., 1974. Model fertility schedules: variations in age structure of childbearing in human populations. Population Index 40, 185–258.

Coale, A.J., Trussell, T.J., 1996. The development and use of demographic models. Population Studies 50, 469–484.

Gompertz, B., 1825. On the nature of the function expressive of the law of human mortality and on a new mode of determining the value of life contingencies. Philosophical Transactions of the Royal Society of London 115 (Part II), 513–585.

Hauser, P.M., Duncan, O.D., 1959. The Study of Population: An Inventory and Appraisal. University of Chicago Press, Chicago.

Heligman, L., Pollard, J.H., 1980. The age pattern of mortality. Journal of the Institute of Actuaries 107, 49–80.

Henry, L., 1961. Some data on natural fertility. Eugenics Quarterly 8, 81–91.

Hoem, J.M., 1987. Statistical analysis of a multiplicative model and its application to the standardization of vital rates: a review. International Statistical Review 55, 119–152.

Hoem, J.M., Madsen, D., Nielsen, J.L., Ohlsen, E.M., Hansen, H.O., Rennermalm, B., 1981. Experiments in modeling recent Danish fertility curves. Demography 18, 231–244.

Lee, R.D., Tuljapurkar, S., 1994. Stochastic population forecasts for the United States: beyond high, medium, and low. Journal of the American Statistical Association 89, 1175–1189.

Makeham, W.M., 1867. On the law of mortality. Journal of the Institute of Actuaries 13, 325–367.

United Nations, 1955. Age and Sex Patterns of Mortality: Model Life Tables for Underdeveloped Countries. United Nations, New York.

Xie, Y., 1990. What is natural fertility? The remodelling of a concept. Population Index 56, 656–663.

Xie, Y., Pimentel, E.F., 1992. Age patterns of marital fertility: revising the Coale-Trussell method. Journal of the American Statistical Association 87, 977–984.

Demographic Techniques: Data Adjustment and Correction

Heather Booth, Australian Demographic and Social Research Institute, The Australian National University, Canberra, Australia
Patrick Gerland, Population Division, United Nations, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

The views and opinions expressed in this article are those of the authors and do not necessarily represent those of the United Nations.

This article is a revision of the previous edition article by H. Booth, volume 5, pp. 3452–3456, © 2001, Elsevier Ltd.

Abstract

Demographic data suffer from sampling errors and from biases arising from coverage and content errors that may be systematic and noncompensating. Common and problematic errors for demographic estimation are those affecting the reporting of age, parity, and deaths. Age misreporting affects population counts and vital rates. Techniques of data evaluation and correction relying on individual-level analysis include the postenumeration survey, imputation, capture–recapture methods, and statistical analysis. Techniques using aggregate-level data employ digit preference indices, sex and age ratios, and smoothing; demographic accounting and internal consistency; and parametric functions, relational models, statistical modeling, and time-based methods.

Demographic data suffer from sampling errors and from nonsampling errors related to coverage and content. Nonsampling errors are often systematic and noncompensating, introducing biases in the same direction rather than random. Such errors are more common or more serious in populations with lower levels of education and economic development. Errors in the reporting and classification of age, parity (number of live-born children), and deaths are the most common and most problematic for demographic measurement and estimation. Techniques of data evaluation and correction may rely on re-enumeration and matching of individual records, data editing, and individual-level statistical analysis and modeling. Alternatively, analytical methods based on population-level aggregated data may be applied, using mathematical formulae, statistical modeling, and the relationship between demographic processes (birth, death, and migration) and age distribution.

Types of Errors

Sampling Errors

Issues such as sampling errors, sample size, and representativity apply to data collected through sample surveys. This increasingly includes data traditionally collected through the census, because many countries (e.g., USA, France, China) are turning to sampling techniques to avoid the cost of complete coverage.

One of the critical factors in determining sampling error is the sample size: the larger the sample, the smaller the sampling error. But the overall size of the population surveyed, cost, and logistical issues often limit the sample size. In a simple random sample, doubling of the sample size will decrease sampling error by 30%. To double the accuracy requires quadrupling the sample size, incurring associated costs. Stratification and more complex sample designs can also help to improve the accuracy of the results (ICF International, 2012b; United Nations, 2005; United Nations, 2008a).

Sampling errors need to be accounted for when analyzing the data. Differences in means or proportions between subgroups, for instance, should always take into account standard errors to determine whether differences are statistically significant within a margin of error. The increasing availability of statistical software able to handle complex sample designs makes it easier to undertake analyses that appropriately take standard errors into account (Deaton, 1997).

The overall size of the population under study is also relevant for the computation of standard errors of vital rates by age (and sex). Random fluctuations in events in small populations (often due to disaggregation by particular characteristics or by small areas) can be addressed by pooling multiple years to ensure a sufficient number of person-years of observation. This is often necessary when computing life tables for small areas or subpopulations (Eayres and Williams, 2004; Scherbov and Ediev, 2011; *see* Life Table; Demographic Techniques: Small Area Estimates and Projections). Pooling of data is also useful for sample data when sample sizes are small (Dwyer-Lindgren et al., 2013; Pedersen and Liu, 2012).

More general issues related to sampling theory and sample design, such as the computation of sampling errors and design effects, are beyond the scope of this section and are addressed in sampling statistics textbooks (Kish, 1965; Levy and Lemeshow, 2013; Lohr, 2009; Som, 1996).

Errors of Coverage

Coverage errors are nonsampling errors and result in bias, affecting the representativity of the data. Coverage errors may occur in data collected through both censuses and sample surveys.

In censuses, errors of coverage comprise the underenumeration or (less commonly) the overenumeration of individuals in the population. Underenumeration occurs when areas, settlements, or households are erroneously omitted from the census, and when individuals within (otherwise covered) households are not enumerated. Outdated or incomplete maps or housing listings, a shortage of trained and experienced interviewers, inadequate field supervision, extreme remoteness or inaccessibility, political or security problems, conflict zones,

difficulties in locating nomadic people and those living in slums or refugee camps, refusals to be enumerated, and loss of census records after enumeration may all contribute to the underenumeration of households. Short-term migrants and infants are among household members most likely to be omitted from enumeration. Overenumeration may arise from double counting, which can occur as a result of poor planning and management of fieldwork leading different interviewers to enumerate the same household or as a result of migrants and more mobile persons (e.g., household members who are temporarily absent, traveling, studying or working elsewhere) being reported both at their household and at their place of visit, work, or study (United Nations, 2008b). Overenumeration also occurs when ineligible persons (e.g., visitors or nonresident aliens) are enumerated as if eligible. Sex-selective underreporting of girls, especially when reporting is by proxy, may also occur in settings favoring sons (e.g., parts of South Asia), especially in the context of restrictive child policy (e.g., China).

Surveys are affected by similar problems. Sample underrepresentation of the population may arise from use of an outdated or inadequate sampling frame for the selection of primary sampling units (usually households), particularly in countries lacking a census in the past decade or where population change has been rapid. Similar sampling bias occurs when a sampling frame systematically excludes some population members (e.g., when a reproductive health survey samples only family housing units, thereby excluding women living in institutions or collectives) or where the probability of selection is lower for certain population members (e.g., because of cultural differences in defining who is a household member (Randall et al., 2013, 2011)). In reproductive health surveys, these issues can lead to single women being omitted (Hull and Hartanto, 2009), or to the overrepresentation of women with more children (Avery et al., 2013), upwardly biasing fertility. The extent of the resulting bias will depend on the difference in the variable of interest between those omitted from and included in the sample, and on their fractions of the total population.

Moreover, the representativity of the sample can be compromised by selection biases during fieldwork (United Nations, 2008a). For example, interviewers may predominantly visit households that are easier to reach (along main roads, for instance) especially in conflict zones (Johnson et al., 2008), or they may miss households that have been recently dissolved due to mortality or selective out-migration of the population at risk in zones experiencing conflicts or crises (Walker et al., 2012), leading to biased mortality estimates, *see* Selection Bias, Statistics of. In addition, many surveys restrict some questions to subgroups of individuals based on eligibility criteria such as sex and age (e.g., females aged 15–49 for questions on reproduction) or duration (e.g., vital events in the last 5 years). Interviewers often need to assess eligibility when respondents either do not know (even after probing) or when values are close to the eligibility limits (ICF International, 2012a), because of the implications in terms of extra workload (i.e., additional sets of questions to collect on supplementary topics). The interpretation by interviewers of selection filters can lead to the omission of potential respondents or events and to shifting of dates or ages (ICF Macro 2009; Pullum, 2006; Silva, 2012; Sullivan, 2008).

Incomplete coverage is also a common deficiency of vital registration data. Further, since year of registration rather than year of event is the usual basis of compilation, late reporting adds to the inaccuracies involved. Births are more affected by late reporting because registration is often not required until children reach school age. Death registration, where it occurs, is usually fairly timely since registration is often a legal requirement for burial; though in some populations such requirements are often ignored, notably in rural areas, leading to underregistration rather than late registration.

Nonresponse is a further source of bias and nonrepresentativity (Groves and Couper, 1998). Item nonresponse occurs when only partial information is collected for an individual. For most variables, nonresponses are omitted from analysis, only their extent being reported; this implicitly assumes that respondents and nonrespondents do not differ with respect to the item in question. For age, nonresponses are often distributed on a *pro rata* basis for each sex because of the need to maintain population size in demographic measurement; this also assumes no difference in the age distributions of respondents and nonrespondents. These assumptions may be erroneous and care is needed in the interpretation of results, especially if nonresponse is substantial.

Errors of Content

Misclassification and content error may affect all demographic variables. These errors arise from incorrect substantive reporting on the part of the respondent and/or interviewer, or from mistakes during data processing. Some errors may be caused by specification error, including poorly phrased questions or instructions, ambiguity of interpretation, and respondent misunderstanding or inability to distinguish key concepts, for example, respondent's biological children from children of brothers or sisters living in the same household. Other errors may be related to the mode of data collection (e.g., length and place of interview, order of questions, whether through interview or self-administered questionnaire, etc.) and interviewer effects (e.g., interviewer errors in phrasing questions or transcribing responses). Errors on the part of the respondent may include digit preference, telescoping and memory lapse related to long recall periods, deliberate misreporting, and errors due to proxy responses. Finally, data processing errors may occur during coding, data entry, editing, analysis, or dissemination (United Nations, 1982; United Nations, 2005).

Assessment and Adjustment of Individual-Level Data

The evaluation and adjustment of individual-level data may be achieved through the use of postenumeration checks, data editing, and individual-level statistical analysis and modeling.

Both coverage and content of surveys and, in particular, censuses are assessed through the postenumeration survey

(PES). The PES involves the recollection of data for a representative selection of the original population or sample. The two sets of data are matched for each individual enumerated in the PES (Alho and Spencer, 2006; United Nations, 2010b; United States. Bureau of the Census. 1985). Similarly, event capture–recapture and matching methods, using two or more independent data sources for a representative sample of cases, are used to estimate the completeness of birth and death registration or of retrospective reports from censuses or surveys (Amstrup et al., 2005; Krótki, 1978; United Nations, 1991). This general approach is regarded as the gold standard, but its cost and complexity often limit its usage. Dual-registration systems (e.g., India's sample registration system (Mahapatra, 2010) and China's disease surveillance point system (Yang et al., 2005)) rely on this approach to derive adjusted sample vital statistics using the Chandrasekaran–Deming formula (1949).

Methods employed to address coverage issues due to unit nonresponse include the use of longitudinal data. Survival analysis and other longitudinal methods such as random- and fixed-effects modeling can deal with issues such as loss to follow-up and censoring (*see* Survival Analysis: Whether, and If So, When?; Panel Surveys: Uses and Applications). In addition, even with cross-sectional data, various multivariate statistical methods can be used to correct for sample selection biases (Stolzenberg and Relles, 1997; Winship and Mare, 1992).

Item nonresponses affecting content can be addressed during data editing or analysis. Missing responses, for instance, may be imputed based on responses from other respondents with similar known characteristics (United Nations, 2010a). In recent years, the greater availability of microdata sets and more advanced multivariate statistical software have made it increasingly easy to apply multiple imputations based on extensive sets of household or individual characteristics (Little and Rubin, 2002; Schafer, 1997), thus allowing missing values to represent multiple sets of plausible values and to analyze data without assuming that nonresponses are missing at random (*see* Multiple Imputation).

Content errors such as digit preference can also be analyzed at the individual level through the use of multinomial logistic regression. An example is Pullum's (2005, 2006) adaptation of the Myers' blended method (Hobbs, 2004; see also below section Evaluation and Adjustment of Population Distribution by Age) for the estimation of interpretable misreporting parameters based on covariates (with standard errors and confidence intervals corrected for sample design). In addition, general patterns of digit preference can be modeled statistically, and the age distribution can potentially be adjusted without imposing any other constraint than a smoothness assumption for the underlying true distribution. For example, Impicciatore and Billari (2011) proposed a general method to estimate smoothed age profiles and relative risks for time-fixed and time-varying covariates using generalized additive models; and Camarda et al. (2008) proposed a method to estimate the unobserved latent distribution, free of the effects of misreporting, using a composite link model with a penalized likelihood.

Until recently, the evaluation of individual-level data had been the prerogative of national statistical offices or other agencies charged with data collection and custody. The increasing access to microdata enables a wider range of users to directly apply and develop methods of evaluation and correction, greatly improving capability in this area.

Assessment and Adjustment of Aggregate-Level Data

Many analytical methods of assessment and adjustment are based on aggregated data. These methods use mathematical formulae, statistical modeling, and the relationship between demographic processes (birth, death, and migration) and the age distribution.

Assessment and Adjustment for Completeness

The completeness of a census enumeration may be assessed, at the aggregate level, through analytical demographic methods using successive censuses and intercensal estimates of the components of population change (Alho and Spencer, 2006; United Nations, 2010b; United States Bureau of the Census, 1985). Similarly, the completeness of birth and death registration (or of retrospective reporting in censuses and surveys) may be assessed at the population level using demographic relationships (*see* Demographic Techniques: Indirect Estimation). For example, vital rates can be compared across independent data sources for similar periods or cohorts, and analyzed for consistency in age-specific trends in rates and sex ratios; such analyses may also be disaggregated by the characteristics of the deceased or, in the case of fertility, of the mother, or of the household (United Nations, 1991). Adjustment factors are usually based on the reciprocal of the estimate of the completeness of vital registration for the total or disaggregated population.

In addition to these approaches, the completeness of adult death registration (or reporting in censuses or surveys) can be evaluated using death distribution or growth-balance methods. These methods require the availability of death data and at least one census, and preferably two or more censuses (Hill et al., 2009; Moultrie et al., 2013; United Nations, 2002). Death distribution methods assume that age is accurately reported, and therefore do not correct for age misreporting, which at older ages usually involves exaggeration in age at death leading to downwardly biased mortality rates (prior to the open-ended age group). These methods address only relative under- and overreporting of registered age-specific deaths against intercensal estimates, which are potentially biased by differential accuracy of census enumeration, and by migration in the intercensal period (Bhat, 2002). In addition, they assume that both the proportion of deaths registered and the completeness of the census enumerations are constant over adult ages. More robust results are usually obtained by combining the results of growth-balance methods and synthetic extinct generations methods (Hill et al., 2009), and using various age combinations (e.g., 20+/50+, 5+/50+, 5+/65+) to minimize the perturbating effects of violations of underlying assumptions in each method (Murray et al., 2010). The resulting estimates of completeness are used to adjust registered deaths.

These estimates of mortality and others obtained from indirect methods (e.g., intercensal cohort survival,

orphanhood, siblings, widowhood – *see* Demographic Techniques: Indirect Estimation) can be compared across methods using different data sources. The most reliable and robust estimates can be used as a baseline reference to potentially adjust more deficient data sources (Hill et al., 2005; Moultrie et al., 2013; Timaeus, 1991; United Nations, 2002).

Evaluation and Correction of Parity Data

Parity data, or number of live-born children a woman has borne, are collected through censuses and surveys or at birth registration. Errors in the reporting of parity often involve understatement due to the omission of births, especially births of girls. Analysis of sex ratios of births by parity (of women) can reveal systematic sex-selective understatement due to gender preference (*see* Missing Girls: A Globalizing Issue). Where adoption is common, overstatement may occur due to both the natural and adoptive mother reporting the birth (Moultrie, 2013a). The inclusion of stillbirths as live births also results in slight parity overstatement.

Omissions of births may be detected and corrected by indirect estimation (*see* Demographic Techniques: Indirect Estimation). Further error arises when childless women are recorded as 'parity not stated' due to the ambiguous use of the dash symbol (–) by interviewers. If these apparent not-stated cases are excluded, average parity is overestimated. If all not-stated cases are included and treated as childless, average parity is underestimated. The El-Badry correction estimates the proportion of women with parity not stated, who are in fact childless (El-Badry, 1961; Moultrie, 2013b). Based on the observed high correlation between true proportions of childless, $Z^*(i)$, and proportions of parity not stated, $NS(i)$ where $i = 1$ to 8 represents the age groups 10–14 to 45–49, El-Badry assumes $NS(i) = \alpha\ Z^*(i) + \beta$, where α is the proportion of truly childless women who are erroneously recorded with parity not stated and β is the constant true proportion of parity not stated. Since the reported proportions of childless, $Z(i)$, equal $(1 - \alpha)\ Z^*(i)$, then $NS(i) = [\alpha/(1 - \alpha)]Z(i) + \beta$, and $\alpha/(1 - \alpha)$ and β can be estimated by ordinary least squares based on the first four or five age groups. Then, β is used to estimate the number of women in each age group for analytical purposes, and true proportions childless are estimated by $Z^*(i) = Z(i) + NS(i) - \beta$. This correction is based on all women. If data on parity are available only for ever-married women (in which case all never-married women are generally assumed childless), the procedure can be applied on the assumption of a linear relation between the proportions of ever-married women who are childless and who have parity not stated.

Evaluation and Adjustment of Population Distribution by Age

Errors in the reporting of age can affect population counts and vital events. Age misreporting includes digit preference (or age heaping) and age under- and overstatement, which may occur together. Digit preference occurs when real age is not known and its estimation involves the 'rounding' of age (or year of birth) to end in a preferred digit such as 0 or 5. Digit preference is best detected graphically. **Figure 1**, for example, shows persistent age heaping across the last five censuses of India.

Various indices have been devised to measure the extent of digit preference in population counts, but they do not offer a means of correction (Hobbs, 2004). Whipple's index measures the preference or avoidance of the digits 0 and 5 over

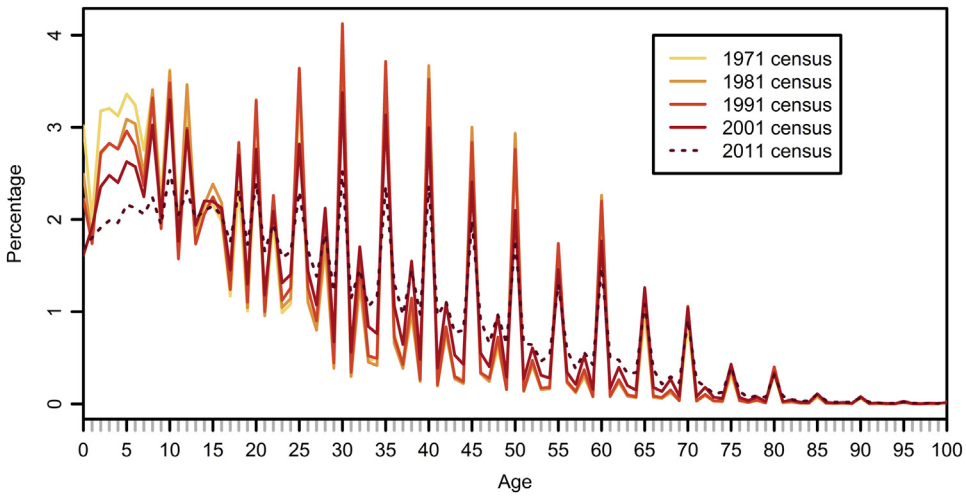

Figure 1 Percentage distribution by single age of the population (both sexes combined) enumerated in the 1971, 1981, 1991, 2001, and 2011 censuses of India. Source: Data from (a) United Nations Statistics Division, 2014. Demographic Yearbook. Population Censuses' Datasets, Population by Age, Sex and Urban/Rural Residence Tabulation. http://unstats.un.org//unsd/demographic/products/dyb/dybcensusdata.htm (accessed 20.01.14.) and (b) Government of India, Office of the Registrar General & Census Commissioner, 2014. 2011 Census Final Results, India. Single Year Age Data – C13 Table (India-States-UTs). http://www.censusindia.gov.in/2011census/Age_level_data/Age_level_data.html (accessed 09.01.14.). Computations by authors. Figure made using R 3.0.2.

the age interval 23–62, though other multidecennial intervals can be used. Myers' blended index measures preference or avoidance of all 10 digits. Since neither index is theoretically precise, Bachi derived a composite index without this deficiency. Whipple's and Myers' indices assume rectangularity or linearity within quinquennial age groups. Later indices, developed by Carrier and by Ramachandran, take account of age structure and variable digit preference over the age range. In practice, the Bachi, Carrier, and Ramachandran indices are laborious and are used less frequently. Given the focus of Whipple's index on 0 and 5, its value in reflecting improvements in the quality of age reporting diminish as digit preference becomes less common (Spoorenberg, 2009). The Myers' blended index and the total modified Whipple's index provide better assessments of the overall spatial and temporal variability of age reporting (Spoorenberg, 2007).

Age misstatement generally involves a directional tendency often related to cultural factors (Ewbank, 1981; Pullum, 2006). Age understatement may occur among women of lower parity than average for their age, especially in cultures with high fertility. Age overstatement often occurs at older ages, particularly among males, in cultures where status is gained with age. Other sex-specific age misstatement may also occur. For example, young female children generally are reported as younger than their male peers. Inaccuracies in the reporting of age tend to be more accentuated in vital events than in population counts. In the case of deaths, age exaggeration may partly be related to proxy reporting. Overstatement in age at death can lead to substantial underestimation of mortality at older ages (Coale and Li, 1991; Condran et al., 1991; Preston et al., 1999) and even to abnormal crossovers and paradoxes between population subgroups when such errors affect more systematically some subgroups than others (Elo and Preston, 1997; Elo et al., 2004; Preston et al., 1996; *see* Mortality Crossover).

In many cases, it is impossible to distinguish between age misstatement, differential coverage by age, and digit preference. Further, one error may offset another. For population counts, the pattern of overall error may be seen by examination of age ratios and sex ratios for 5- or 10-year age groups (Hobbs, 2004; Pullum, 2006). **Figure 2** illustrates these deviations for the 1991 census of India. Panel A shows the age distributions by sex. Panel B shows how sex ratios by age differ from sex equality (sex ratio = 100), indicating distortion due to age misreporting (given sex differences in migration and mortality). Panel C shows age ratios by sex, contrasting over- and understatement of age by age group. These sex and age ratios may also be compared with ratios from suitable model life table populations or smoothed age distributions (as seen in **Figure 2**).

An indication of the overall accuracy of age reporting is given by the United Nations age-sex accuracy index (or 'joint score') (United Nations, 1955). This index assumes no significant irregularities by age (i.e., limited change in past birth cohort size) and sex parity at all ages (implying equal numbers of male and female births and equal mortality). Since the index assumes that all irregularities are errors, it should be interpreted with caution in populations affected by historical events (e.g., baby booms or baby busts, age–sex-specific migration, or mortality shocks or crises).

Errors may also be assessed through comparison with expected counts and distributions based on external data using demographic relationships (Pollard et al., 1990; Preston et al., 2001b). When accuracy permits, a previous enumeration of a population can be 'updated' using data on births, deaths, and net migration to derive an expected age distribution using a cohort-component population projection (Himes and Clogg, 1992; United States Bureau of the Census. 1985). Consistency between the enumerated population at older ages (e.g., age 80 and over) and recorded deaths can be enforced by using the (almost) extinct generation method when death registration records are considered more reliable than census information (Andreev, 2004; Machemedze and Dorrington, 2011; Wilmoth et al., 2007). Comparison of the cumulative age distribution with model stable distributions may identify distortions due to misreporting, in the absence of substantial migration and assuming no significant changes in past fertility and mortality. A suitable model stable population can also be used to derive the structure of the open-ended age group when it is unknown (United Nations, 1983; *see* Population Dynamics: Theory of Stable Populations). When fertility and mortality have been declining in recent decades (as in many countries), the availability of two or more censuses can help to assess the accuracy of age distributions by examining cohort consistency between censuses, and by allowing the assumption of stability to be relaxed by iteratively fitting projected population structures to the observed numbers (Bhat, 1990). Using successive censuses, cohorts can be compared at different points in time: allowing for mortality, the (smoothed) age distributions enable the detection of possible enumeration inconsistencies. **Figure 3** shows the detection of missing infants and young adults in the 1991 census of India.

The effects in demographic measurement and estimation of both digit preference and age misstatement can be reduced by collecting exact date (or at least year) of event rather than age and by using wider (than standard 5-year) age intervals at the analysis stage, though this is not always practicable and involves greater approximation. Similarly, the effects on standard quinquennial age groups can be reduced by adopting nonstandard age groups, the optimum choice of which is based on Myers' index (Hobbs, 2004), but these are generally inconvenient for analytical and comparative purposes.

For many purposes it may be appropriate to smooth or graduate the age distribution. Methods of smoothing include the use of graphs (manual smoothing), moving averages, curve fitting, and statistical modeling. Mathematical graduation is used to smooth narrower age groups within broader age groups. The Carrier–Farrag and Karup–King–Newton formulae treat all age groups the same, whereas the Arriaga's formulae treat the first and last differently (Arriaga et al., 1994). In the following formulae, ${}_nP_x$ represents the population aged x to $x + n - 1$, x and n being integers. The Carrier–Farrag ratio method defines the relationship of a 5-year age group to the 10-year age group of which it is a constituent as

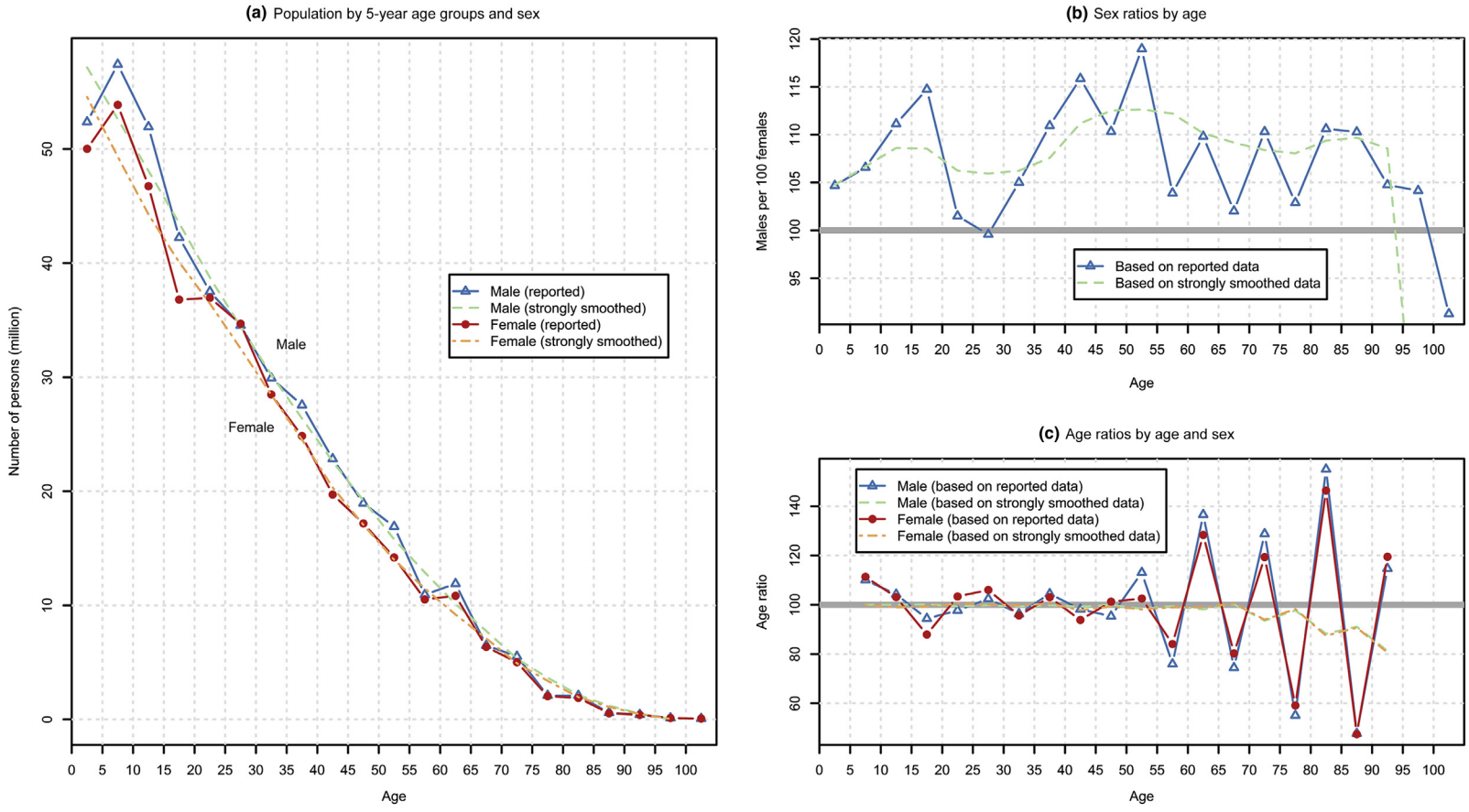

Figure 2 Analysis of the distribution by 5-year age group and sex of the 1991 population census of India. Panel (a) presents for each sex and 5-year age group the reported (solid line with symbol) and heavily smoothed (dashed line) distributions. Panel (b) shows corresponding sex ratios by 5-year age group in comparison with assumed sex equality (sex ratio = 1). Panel (c) shows corresponding age ratios for each sex, calculated as the ratio of one age group to the average of the two adjacent age groups. Source: Data from United Nations Statistics Division, 2014. Demographic Yearbook. Population Censuses' Datasets, Population by Age, Sex, and Urban/Rural Residence Tabulation. http://unstats.un.org//unsd/demographic/products/dyb/dybcensusdata.htm (accessed 20.01.14.). Computations by authors based on U.S. Bureau of the Census, International Programs Center, 2013. Population Analysis Spreadsheets (PAS): AGESEX and AGESMTH. http://www.census.gov/population/international/software/pas/ (accessed 30.01.14.). Figure made using R 3.0.2.

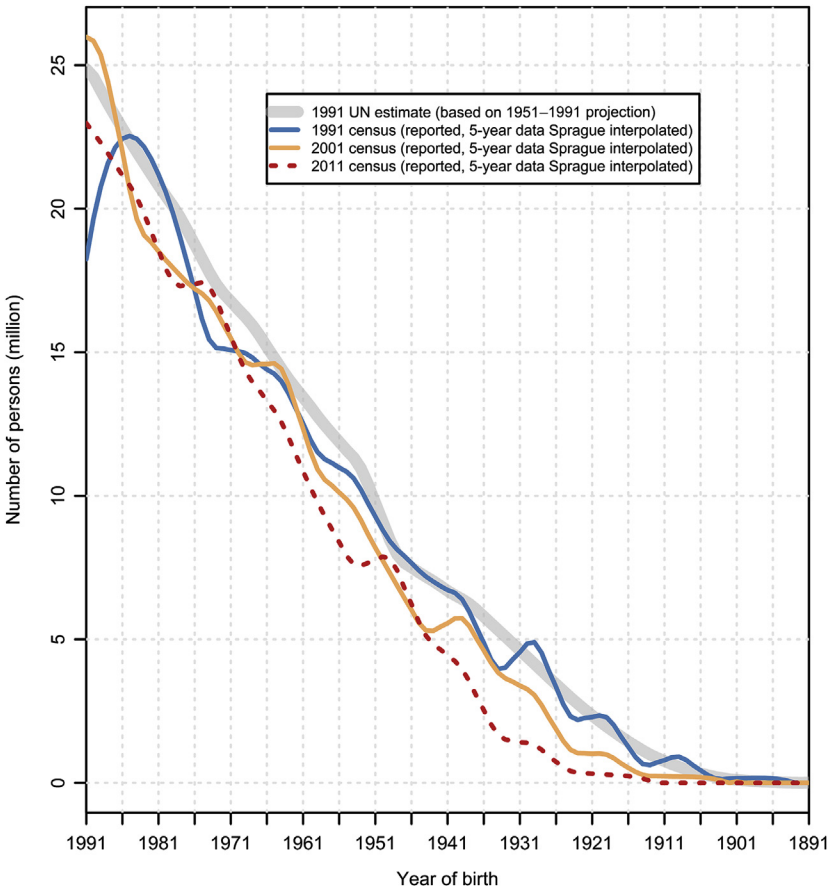

Figure 3 Evaluation of the distribution by age of the 1991 census population of India (both sexes combined). Census data are smoothed using Sprague interpolation with 5-year age groups. The 1991 age distribution (blue solid line) is compared with the UN estimate for 1991 based on a 1951–1991 cohort-component population projection (gray bold solid line), and with the corresponding birth cohorts enumerated in the 2001 census (orange solid line) and 2011 census (red dotted line). Source: Data from (a) United Nations Statistics Division, 2014. Demographic Yearbook. Population Censuses' Datasets, Population by Age, Sex, and Urban/Rural Residence Tabulation. http://unstats.un.org//unsd/demographic/products/dyb/dybcensusdata.htm (accessed 20.01.14.) and (b) Government of India, Office of the Registrar General & Census Commissioner, 2014. 2011 Census Final Results, India. Single Year Age Data – C13 Table (India-States-UTs). http://www.censusindia.gov.in/2011census/Age_level_data/Age_level_data.html (accessed 09.01.14.). Computations by authors. Figure made using R 3.0.2.

$$ {}_5P_{x+5} = {}_{10}P_x \Big/ \left[1 + \left({}_{10}P_{x-10} / {}_{10}P_{x+10}\right)^{1/4}\right] \quad \text{and} $$
$$ {}_5P_x = {}_{10}P_x - {}_5P_{x+5} $$

The Karup–King–Newton formula assumes a quadratic relationship between three consecutive 10-year age groups, that is, ${}_5P_x = \frac{{}_{10}P_x}{2} + \frac{\left({}_{10}P_{x-10} - {}_{10}P_{x+10}\right)}{16}$. The Arriaga's formulae are based on second-degree polynomials passing through the midpoints of three consecutive 10-year age groups. Then

$$ {}_5P_{x+5} = \frac{\left(8\,{}_{10}P_x + 5\,{}_{10}P_{x+10} - {}_{10}P_{x+20}\right)}{24} $$

when the 10-year age group is the first,

$$ {}_5P_x = \frac{\left(-\,{}_{10}P_{x-20} + 5\,{}_{10}P_{x-10} + 8\,{}_{10}P_x\right)}{24} $$

when it is the last, and

$$ {}_5P_{x+5} = \frac{\left(-\,{}_{10}P_{x-10} + 11\,{}_{10}P_x + 2\,{}_{10}P_{x+10}\right)}{24} \quad \text{otherwise} $$

In all cases, the second 5-year age group is derived by subtraction with ${}_5P_x = {}_{10}P_x - {}_5P_{x+5}$

When age misreporting is more severe, the totals of all age groups need potential modification. The United Nations formula lightly smooths 5-year age groups, based on five consecutive 5-year groups, and gives the smoothed population aged x to $x + 4$ as

$$ {}_5P'_x = \frac{\left(-\,{}_5P_{x-10} + 4\,{}_5P_{x-5} + 10\,{}_5P_x + 4\,{}_5P_{x+5} - {}_5P_{x+10}\right)}{16} $$

For heavier smoothing, three consecutive 10-year age groups are used to derive the central smoothed age group as

$$ {}_{10}P'_x = \frac{\left({}_{10}P_{x-10} + 2\,{}_{10}P_x + {}_{10}P_{x+10}\right)}{4} $$

Neither formula smooths the youngest and oldest age groups. Since smoothing changes overall population size, the smoothed age groups are adjusted proportionately to achieve

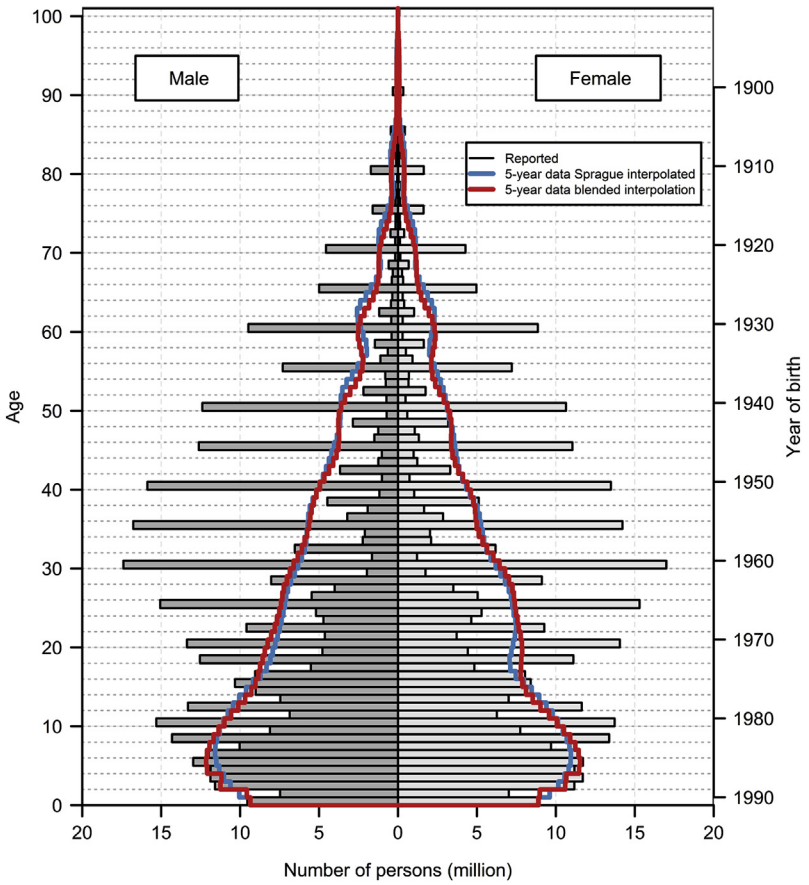

Figure 4 Analysis of the distribution by single age and sex of the 1991 census of India. Reported data by age and sex (solid gray bars) show digit preference. Smoothed data based on 5-year age groups using Sprague interpolation (solid blue line), and a blended osculatory interpolation using the five possible age groupings (red solid line) are included as reference. Source: Data from (a) United Nations Statistics Division, 2014. Demographic Yearbook. Population Censuses' Datasets, Population by Age, Sex, and Urban/Rural Residence Tabulation. http://unstats.un.org//unsd/demographic/products/dyb/dybcensusdata.htm (accessed 20.01.14.) and (b) Government of India, Office of the Registrar General & Census Commissioner (2014). 2011 Census Final Results, India. Single Year Age Data – C13 Table (India-States-UTs). http://www.censusindia.gov.in/2011census/Age_level_data/Age_level_data.html (accessed 09.01.14.). Computations by authors. Figure made using R 3.0.2 and Pyramid-package by Tim Riffe. https://github.com/timriffe/Pyramid/ (accessed on 30.01.14.)

the enumerated total. The Arriaga's (or other such) formulae are then applied to smoothed and adjusted 10-year age groups to derive smoothed 5-year age groups. An example of heavier smoothing is seen in Figure 2.

A single-year age distribution may be derived from grouped data by adopting the within-group distribution of a suitable unabridged life table population, but discontinuities at age-group junctions are inherent in this approach. A smoother distribution is obtained by mathematical interpolation, such as osculatory interpolation, which is especially useful for irregular data because of the constraint of continuity at junctions. Sprague multipliers employ fourth- and fifth-difference osculatory interpolation to smooth single-year data within 5-year age groups without changing the totals. They are less satisfactory for the first and last two age groups and where there is marked curvature (see Figure 4). Sprague multipliers can be employed with any continuous data including rates, for any age partition (e.g., tenths), and with unequal age intervals. A blended distribution can be derived by averaging smoothed distributions obtained by osculatory interpolation within the five possible age groupings with terminal digits 0–4, 1–5, 2–6, 3–7, 4–8 (Figure 4).

Spline interpolation is a more sophisticated type of osculatory interpolation that has become more widely used with increasing computing power. Spline interpolation involves extensive iteration resulting in a highly smoothed distribution (Judson and Popoff, 2004). The approach is very flexible, can be used on counts, proportions, rates, and ratios, and is applicable to any age groupings including those of unequal length. Monotonicity can be imposed using penalized splines (Ligges, 2013; Schmertmann, 2012; Smith et al., 2004); and weighted splines allow for unequal variance by age (Booth et al., 2014). More robust results may be obtained by application to cumulative distributions, with differencing of the interpolated results (Wilmoth et al., 2007).

Adjustment of Age Patterns in Demographic Processes

In addition to spline interpolation of rates by age (Section Evaluation and Adjustment of Population Distribution by

Age), specific analytical methods are available for the smoothing and adjustment of rates or age distributions. These methods include relational models and predefined mathematical functional forms.

Relational models are used when the data suffer from biases due to reporting error. An empirical standard or reference distribution is used to adjust the data. The observed distribution is related to the standard through a two-parameter linear model after applying a suitable transformation (e.g., Logit) to both distributions. Thus, the adjusted distribution is the best-fitting (over the age range employed) linear function of the standard.

For mortality, the Brass Logit life-table system involves fitting observed data to a standard model life table (Brass et al., 1968; *see* Demographic Models; Stochastic Models; Population Forecasts). A larger number of parameters help to address mortality deviations from the standard at young and old ages (Ewbank et al., 1983; Rachet et al., 2008; Zaba, 1979). For fertility, the Brass Relational Gompertz model is used with an appropriate standard, such as the Booth Standard for high fertility (Booth, 1984; Brass, 1974; Moultrie, 2013c; *see* Demographic Models). The adjustment of immigration and emigration data can also be achieved through relational models (Zaba, 1985).

Relational models are flexible and easy to apply, but the adjusted distribution is influenced by the choice of the reference standard (Moultrie, 2013d). Ideally, a more context-specific standard is used (De Beer, 2012; INDEPTH Network. 2004; Murray et al., 2003; Wang et al., 2012; Wilmoth et al., 2011). To reduce dependence on the chosen standard, the relationship between observed and standard fertility or mortality may be modeled by splines (De Beer, 2011, 2012).

Mathematical functions are used in demographic modeling and data smoothing. A parametric function is fitted to the empirical observations, imposing a functional form on the data. This approach is advantageous when data availability is limited to certain ages, or when the data are too noisy for nonparametric smoothing to provide a plausible functional form. Adult mortality rates, especially at older ages, are often smoothed and potentially adjusted using the Gompertz or Makeham functions (Heuveline and Clark, 2011), typically fitted to adult mortality rates and used to smooth and extrapolate mortality at older ages. For observed mortality at the oldest ages (e.g., 85 and over), additional smoothing or adjustment may be achieved using the Kannisto model of old-age mortality (Thatcher et al., 1998) fitted to data at age 75 and over. Adjustments to old-age mortality ensure consistency (Condran et al., 1991) by age (e.g., monotonic increase), over time (e.g., monotonic decline) and between the sexes (e.g., male rates greater than or equal to female rates). Where mortality rates must be smoothed and adjusted at all ages, additional terms are required to model infancy and childhood, and young adulthood. Several parametric functions serve to model the whole age pattern of mortality (e.g., Heligman and Pollard (1980), Kostaki (1992), Siler (1983), Carriere (1992); see reviews by Gage and Mode (1993) and Booth and Tickle (2008)).

Parametric functions are also used to model the age patterns of nuptiality and fertility (Preston et al., 2001c); see review by Booth (2006). The Coale–McNeil (Coale and McNeil, 1972) parameterization of nuptiality applies to first marriage (Coale, 1971; Hernes, 1972) and can be used to model first births (Bloom and Trussell, 1984). The parsimony of these models and the functional shape they impose on the data are useful to smooth and adjust poor-quality data. However, their ability to appropriately model empirical observations depends on their flexibility to capture the range of situations experienced by various countries over time (e.g., the emergence of bimodal fertility distributions by age in low fertility populations (Bermúdez et al., 2012)).

Migration is modeled by the multiexponential function (Rogers and Little, 1994), which is useful in imposing a smooth age pattern on data that are often highly defective; see also Booth (2006). Rogers et al. (2010) demonstrate how to combine smoothing to remove randomness with the use of the multiexponential and empirical schedules to adjust immigration and emigration data.

Adjustments over Time

Demographic series for two or more points in time can be used to evaluate trends, growth, and structural population change. Time series facilitate interpolation and extrapolation. In particular, time series of population counts and the demographic processes of birth, death, and migration jointly enable the derivation of robust estimates.

One of the most common adjustments in demography is shifting the population to refer to a specific date, such as 30 June. This may be achieved in two ways: first by linear or exponential interpolation of each age group between two time points (Arriaga et al., 1994); and second by adjusting the total population exponentially (the growth rate is required) and assuming the same age–sex structure as at the nearest enumeration. This last option is valid only for shifts of short duration (i.e., a few years). For longer extrapolations, a cohort-component population projection with assumptions about fertility, mortality, and migration trends should be used.

Age-specific growth rates can also be employed to assess and potentially adjust age-specific data through demographic relationships in nonstable populations using the variable-r approach (Cai, 2008; Preston and Coale, 1982; Preston et al., 2001a; Schmertmann, 2002; *see* Population Dynamics: Theory of Nonstable Populations). This approach, which may also use duration-specific data, assumes no significant changes in patterns of age misreporting and completeness of coverage between enumerations.

When demographic events in a reference time period are reported retrospectively, errors of timing may occur (e.g., displacement and recall lapse). For births, such errors are often corrected through indirect estimation (*see* Demographic Techniques: Indirect Estimation). Consistency between successive censuses and surveys for overlapping periods can be analyzed visually, as well as through robust statistical modeling of time trends (Pullum, 2006). Inconsistencies can be resolved by the adjustment of indices over age or over time using splines (or locally weighted regressions), with the optional use of covariates and hierarchical statistical modeling to take account of likely biases associated with factors such as method of data collection or length of recall period (Alkema and New, 2013; Alkema et al., 2012). The

increased availability of both microdata and internationally comparable public use data sets have made it increasingly possible to perform pooled analyses of multiple datasets for a country, and to analyze all countries simultaneously to detect potential systematic biases and to adjust them if possible. Such statistical modeling approaches are being increasingly used to derive robust time series estimates (with uncertainty bounds), including for child mortality (Alkema and New, 2013; Hill et al., 1999), maternal mortality (Lozano et al., 2011; Wilmoth et al., 2012), adult mortality (Masquelier et al., 2013; Obermeyer et al., 2010; Wang et al., 2012), fertility (Alkema et al., 2012; Machiyama, 2010; Schoumaker, 2011, 2013), and contraception (Alkema et al., 2013). Furthermore, simultaneous estimation of population counts and vital rates (taking into account the underlying uncertainty in each data source) can be used to ensure internal consistency between all demographic components (Luther et al., 1987; Luther and Retherford, 1988; Wheldon et al., 2013a, 2013b; 2013c).

See also: Demographic Models; Demographic Techniques: Indirect Estimation; Demographic Techniques: Small-area Estimates and Projections; Life Table; Missing Data; Nonsampling Errors; Panel Surveys: Uses and Applications; Period and Cohort Analysis in Demography; Population Dynamics: Momentum of Population Growth; Population Dynamics: Theory of Nonstable Populations; Population Dynamics: Theory of Stable Populations; Selection Bias, Statistics of; Survival Analysis: Introduction; Survival Analysis: Whether, and If So, When?.

Bibliography

Alho, J.M., Spencer, B.D., 2006. Errors in census numbers. In: Statistical Demography and Forecasting. Springer, New York, NY, pp. 296–326.

Alkema, L., Kantorova, V., Menozzi, C., Biddlecom, A., 2013. National, regional, and global rates and trends in contraceptive prevalence and unmet need for family planning between 1990 and 2015: a systematic and comprehensive analysis. Lancet 381 (9878), 1642–1652.

Alkema, L., New, J.R., 2013. Global Estimation of Child Mortality Using a Bayesian B-Spline Regression Model. Arxiv.org.

Alkema, L., Raftery, A., Gerland, P., Clark, S.J., Pelletier, F., 2012. Estimating trends in the total fertility rate with uncertainty using imperfect data: examples from West Africa. Demographic Research 26 (15), 331–362.

Amstrup, S.C., McDonald, T.L., Manly, B.F.J. (Eds.), 2005. Handbook of Capture-Recapture Analysis. Princeton University Press, Princeton.

Andreev, K.F., 2004. A method for estimating size of population aged 90 and over with application to the U.S. census 2000 data. Demographic Research 11 (9), 235–262.

Arriaga, E.E., Johnson, P.D., Jamison, E., 1994. Age and sex composition. In: Arriaga, E.E. (Ed.), Population Analysis with Microcomputers, Presentation of Techniques, vol. I. US Bureau of the Census, Washington, DC, pp. 11–56.

Avery, C., St Clair, T., Levin, M., Hill, K., 2013. The 'own children' fertility estimation procedure: a reappraisal. Population Studies. A Journal of Demography 67 (2), 171–183.

Bermúdez, S., Blanquero, R., Hernández, J.A., Planelles, J., 2012. A new parametric model for fitting fertility curves. Population Studies 66 (3), 297–310.

Bhat, P.N.M., 1990. Estimating transition probabilities of age misstatement. Demography 27 (1), 149–163.

Bhat, P.N.M., 2002. General growth balance method: a reformulation for populations open to migration. Population Studies 56 (1), 23–34.

Bloom, D.E., Trussell, J., 1984. What are the determinants of delayed childbearing and permanent childlessness in the United States? Demography 21 (4), 591–611.

Booth, H., 1984. Transforming Gompertz's function for fertility analysis: the development of a standard for the relational Gompertz function. Population Studies 38 (3), 495–506.

Booth, H., 2006. Demographic forecasting: 1980 to 2005 in review. International Journal of Forecasting 22 (3), 547–581.

Booth, H., Hyndman, R., Tickle, L., 2014. Prospective life tables. In: Charpentier, A. (Ed.), Computational Actuarial Science with R. CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 319–344.

Booth, H., Tickle, L., 2008. Mortality modelling and forecasting: a review of methods. Annals of Actuarial Science 3 (1–2), 3–43.

Brass, W., 1974. Perspectives in population prediction: illustrated by the statistics of England and Wales. Journal of the Royal Statistical Society. Series A (General) 137 (4), 532–583.

Brass, W., Coale, A.J., Demeny, P., Heisel, D.F., Lorimer, F., Romaniuk, A., Van de Walle, E., 1968. The Demography of Tropical Africa. Princeton University Press, Princeton, NJ.

Cai, Y., 2008. An assessment of China's fertility level using the variable-r method. Demography 45 (2), 271–281.

Camarda, C.G., Eilers, P.H.C., Gampe, J., 2008. Modelling general patterns of digit preference. Statistical Modelling 8 (4), 385–401.

Carriere, J.F., 1992. Parametric models for life tables. Transactions of the Society Actuaries 44, 77–99.

Chandra Sekar, C., Deming, W.E., 1949. On a method of estimating birth and death rates and the extent of registration. Journal of the American Statistical Association 44 (245), 101–115.

Coale, A.J., 1971. Age patterns of marriage. Population Studies 25 (2), 193–214.

Coale, A.J., Li, S., 1991. The effect of age misreporting in China on the calculation of mortality rates at very high ages. Demography 28 (2), 293–301.

Coale, A.J., McNeil, D.R., 1972. The distribution by age of the frequency of first marriage in a female cohort. Journal of the American Statistical Association 67 (340), 743–749.

Condran, G.A., Himes, C.L., Preston, S.H., 1991. Old-age mortality patterns in low-mortality countries: an evaluation of population and death data at advanced ages, 1950 to the present. Population Bulletin of the United Nations 30, 23–60.

De Beer, J., 2011. A new relational method for smoothing and projecting age-specific fertility rates: TOPALS. Demographic Research 24 (18), 409–454.

De Beer, J., 2012. Smoothing and projecting age-specific probabilities of death by TOPALS. Demographic Research 27 (20), 543–592.

Deaton, A., 1997. The Analysis of Household Surveys: A Microeconometric Approach to Development Policy. World Bank Publications, Washington, DC.

Dwyer-Lindgren, L., Gakidou, E., Flaxman, A., Wang, H., 2013. Error and bias in under-5 mortality estimates derived from birth histories with small sample sizes. Population Health Metrics 11 (1), 13.

Eayres, D., Williams, E.S., 2004. Evaluation of methodologies for small area life expectancy estimation. Journal of Epidemiology and Community Health 58 (3), 243–249.

El-Badry, M.A., 1961. Failure of enumerators to make entries of zero: errors in recording childless cases in population censuses. Journal of the American Statistical Association 56 (296), 909–924.

Elo, I.T., Preston, S.H., 1997. Racial and ethnic differences in mortality at older ages. In: Martin, L.G., Soldo, B.J., National Research Council Committee on Population (Eds.), Racial and Ethnic Differences in the Health of Older Americans. National Academy Press, Washington, DC, pp. 10–42.

Elo, I.T., Turra, C.M., Kestenbaum, B., Ferguson, B.R., 2004. Mortality among elderly hispanics in the United States: past evidence and new results. Demography 41 (1), 109–128.

Ewbank, D.C., 1981. Age Misreporting and Age-Selective Underenumeration: Sources, Patterns, and Consequences for Demographic Analysis. National Academy Press, Washington, DC.

Ewbank, D.C., Leon, J.C.G.D., Stoto, M.A., 1983. A reducible four-parameter system of model life tables. Population Studies 37 (1), 105–127.

Gage, T.B., Mode, C.J., 1993. Some laws of mortality: how well do they fit? Human Biology 65 (3), 445–461.

Groves, R.M., Couper, M.P., 1998. Nonresponse in Household Interview Surveys, Wiley Series in Probability and Statistics. Survey Methodology Section. Wiley, New York.

Heligman, L., Pollard, J.H., 1980. The age pattern of mortality. The Journal of the Institute of Actuaries 107 (434), 49–80.

Hernes, G., 1972. Process of entry into first marriage. American Sociological Review 37 (2), 173–182.

Heuveline, P., Clark, S.J., 2011. Model schedules of mortality. In: Rogers, R.G., Crimmins, E.M. (Eds.), International Handbook of Adult Mortality. Springer, New York, pp. 511–532.

Hill, K., Choi, Y., Timaeus, I., 2005. Unconventional approaches to mortality estimation. Demographic Research 13, 281–300.
Hill, K., Pande, R., Mahy, M., Jones, G., 1999. Trends in Child Mortality in the Developing World: 1960 to 1996. UNICEF, New York, NY.
Hill, K., You, D., Choi, Y., 2009. Death distribution methods for estimating adult mortality: sensitivity analysis with simulated data errors. Demographic Research 21 (9), 235–254.
Himes, C.L., Clogg, C.C., 1992. An overview of demographic analysis as a method for evaluating census coverage in the United States. Population Index 58 (4), 587–607.
Hobbs, F.B., 2004. Age and sex composition. In: Siegel, J.S., Swanson, D.A., Shryock, H.S. (Eds.), The Methods and Materials of Demography. Elsevier/Academic Press, Amsterdam; Boston, pp. 125–174.
Hull, T.H., Hartanto, W., 2009. Resolving contradictions in Indonesian fertility estimates. Bulletin of Indonesian Economic Studies 45 (1), 61–71.
ICF International, 2012a. Demographic and Health Survey Interviewer's Manual, Demographic and Health Survey's DHS Toolkit of Methodology. ICF International, Calverton, Maryland, USA.
ICF International, 2012b. Demographic and Health Survey Sampling and Household Listing Manual, Demographic and Health Survey's DHS Toolkit of Methodology. ICF International, Calverton, Maryland, USA.
ICF Macro, 2009. Training Field Staff for DHS Surveys, Demographic and Health Survey's DHS Toolkit of Methodology. ICF Macro, Calverton, Maryland, USA.
Impicciatore, R., Billari, F., 2011. MAPLES: a general method for the estimation of age profiles from standard demographic surveys (with an application to fertility). Demographic Research 24 (29), 719–748.
INDEPTH Network, 2004. INDEPTH Model Life Tables for Sub-Saharan Africa. Ashgate, Aldershot, Hants, England; Burlington, VT.
Johnson, N.F., Spagat, M., Gourley, S., Onnela, J.P., Reinert, G., 2008. Bias in epidemiological studies of conflict mortality. Journal of Peace Research 45 (5), 653–663.
Judson, D.H., Popoff, C.L., 2004. Appendix C. Selected general methods. In: Siegel, J.S., Swanson, D.A., Shryock, H.S. (Eds.), The Methods and Materials of Demography. Elsevier/Academic Press, Amsterdam; Boston, pp. 677–732.
Kish, L., 1965. Survey Sampling. Wiley, New York, NY.
Kostaki, A., 1992. A nine-parameter version of the Heligman-Pollard formula. Mathematical Population Studies 3 (4), 277–288.
Krótki, K.J. (Ed.), 1978. Developments in Dual System Estimation of Population Size and Growth. University of Alberta Press, Edmonton.
Levy, P.S., Lemeshow, S., 2013. Sampling of Populations: Methods and Applications. Wiley.
Ligges, U., 2013. Package 'signal': pchip (Piecewise cubic hermite interpolation) and interp1 (Interpolation methods, including linear, spline, and cubic interpolation). In: R Core Team (Ed.), R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data, Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, second ed. Wiley, New York, NY.
Lohr, S., 2009. Sampling: Design and Analysis. Cengage Learning, Boston, MA.
Lozano, R., Wang, H., Foreman, K.J., Rajaratnam, J.K., Naghavi, M., Marcus, J.R., Dwyer-Lindgren, L., Lofgren, K.T., Phillips, D., Atkinson, C., Lopez, A.D., Murray, C.J.L., 2011. Progress towards millennium development goals 4 and 5 on maternal and child mortality: an updated systematic analysis. Lancet 378 (9797), 1139–1165.
Luther, N.Y., Gaminiratne, K.H.W., Silva, S.D., Retherford, R.D., 1987. Consistent correction of international migration data for Sri Lanka, 1971–81. International Migration Review 21 (4), 1335–1369.
Luther, N.Y., Retherford, R.D., 1988. Consistent correction of census and vital registration data. Mathematical Population Studies 1 (1), 1–20.
Machemedze, T., Dorrington, R., 2011. Levels of mortality of the South African aged population using the method of extinct generations. Journal of African Population Studies 25 (Suppl. 1), 63–76.
Machiyama, K., 2010. A Re-examination of Recent Fertility Declines in Sub-saharan Africa. Demographic and Health Surveys Working Papers No. 68. Macro International, Calverton, MD.
Mahapatra, P., 2010. An overview of the sample registration system in India. In: Prince Mahidol Award Conference & Global Health Information Forum.
Masquelier, B., Reniers, G., Pison, G., 2013. Divergences in trends in child and adult mortality in sub-Saharan Africa: survey evidence on the survival of children and siblings. Population Studies, 1–17.
Moultrie, T.A., 2013a. Assessment of parity data. In: Moultrie, T.A., Dorrington, R.E., Hill, A.G., Hill, K., Timæus, I.M., Zaba, B. (Eds.), Tools for Demographic Estimation. International Union for the Scientific Study of Population, Paris.
Moultrie, T.A., 2013b. The el-Badry correction. In: Moultrie, T.A., Dorrington, R.E., Hill, A.G., Hill, K., Timæus, I.M., Zaba, B. (Eds.), Tools for Demographic Estimation. International Union for the Scientific Study of Population, Paris.
Moultrie, T.A., 2013c. The relational Gompertz model. In: Moultrie, T.A., Dorrington, R.E., Hill, A.G., Hill, K., Timæus, I.M., Zaba, B. (Eds.), Tools for Demographic Estimation. International Union for the Scientific Study of Population, Paris.
Moultrie, T.A., 2013d. Relational standards used in this project. In: Moultrie, T.A., Dorrington, R.E., Hill, A.G., Hill, K., Timæus, I.M., Zaba, B. (Eds.), Tools for Demographic Estimation. International Union for the Scientific Study of Population, Paris.
Moultrie, T.A., Dorrington, R.E., Hill, A.G., Hill, K., Timæus, I.M., Zaba, B. (Eds.), 2013. Tools for Demographic Estimation. International Union for the Scientific Study of Population.
Murray, C.J.L., Rajaratnam, J.K., Marcus, J., Laakso, T., Lopez, A.D., 2010. What can we conclude from death registration? Improved methods for evaluating completeness. PLoS Med 7 (4), e1000262.
Murray, C.J.L., Ferguson, B.D., Lopez, A.D., Guillot, M., Salomon, J.A., Ahmad, O., 2003. Modified logit life table system: principles, empirical validation, and application. Population Studies 57 (2), 165–182.
Obermeyer, Z., Rajaratnam, J.K., Park, C.H., Gakidou, E., Hogan, M.C., Lopez, A.D., Murray, C.J., 2010. Measuring adult mortality using sibling survival: a new analytical method and new results for 44 countries, 1974–2006. PLoS Med 7 (4), e1000260.
Pedersen, J., Liu, J., 2012. Child mortality estimation: appropriate time periods for child mortality estimates from full birth histories. PLoS Med 9 (8), e1001289.
Pollard, A.H., Yusuf, F., Pollard, G.N., 1990. Testing the accuracy of demographic data. In: Demographic Techniques. Pergamon Press, Rushcutters Bay, NSW, pp. 152–163.
Preston, S.H., Coale, A.J., 1982. Age structure, growth, attrition, and accession: a new synthesis. Population Index 48 (2), 217–259.
Preston, S.H., Elo, I.T., Rosenwaike, I., Hill, M., 1996. African-American mortality at older ages: results of a matching study. Demography 33 (2), 193–209.
Preston, S.H., Elo, I.T., Stewart, Q., 1999. Effects of age misreporting on mortality estimates at older ages. Population Studies 53 (2), 165–177.
Preston, S.H., Heuveline, P., Guillot, M., 2001a. Demographic relations in nonstable populations. In: Demography: Measuring and Modeling Population Processes. Blackwell Publishers, Malden, MA, pp. 171–190.
Preston, S.H., Heuveline, P., Guillot, M., 2001b. Methods for evaluating data quality. In: Demography: Measuring and Modeling Population Processes. Blackwell Publishers, Malden, MA, pp. 211–223.
Preston, S.H., Heuveline, P., Guillot, M., 2001c. Modeling age patterns of vital events. In: Demography: Measuring and Modeling Population Processes. Blackwell Publishers, Malden, MA, pp. 191–210.
Pullum, T.W., 2005. A statistical reformulation of demographic methods to assess the quality of age and date reporting, with application to the Demographic and Health Surveys. In: Annual Meeting of the Population Association of America. Population Association of America, Philadelphia, PA.
Pullum, T.W., 2006. An Assessment of Age and Date Reporting in the DHS Surveys, 1985–2003. Demographic and Health Surveys Methodological Reports No. 5. Macro International, Calverton, MD.
Rachet, B., Maringe, C., Ellis, L., 2008. Ewblft (creation and smoothing of life tables). In: Cancer Research UK Cancer Survival Group (Ed.), Tools for Cancer Survival Analysis. London School of Hygiene & Tropical Medicine, University of London, London, UK.
Randall, S., Coast, E., Compaore, N., Antoine, P., 2013. The power of the interviewer. Demographic Research 28, 763–792.
Randall, S., Coast, E., Leone, T., 2011. Cultural constructions of the concept of household in sample surveys. Population Studies. A Journal of Demography 65 (2), 217–229.
Rogers, A., Little, J., Raymer, J., 2010. The Indirect Estimation of Migration: Methods for Dealing with Irregular, Inadequate, and Missing Data. In: The Springer Series on Demographic Methods and Population Analysis. Springer, Dordrecht.
Rogers, A., Little, J.S., 1994. Parameterizing age patterns of demographic rates with the multiexponential model schedule. Mathematical Population Studies 4 (3), 175–195.
Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data, Monographs on Statistics and Applied Probability. Chapman & Hall, New York, NY.
Scherbov, S., Ediev, D., 2011. Significance of life table estimates for small populations: simulation-based study of estimation errors. Demographic Research 24 (22), 527–550.
Schmertmann, C., 2012. Calibrated Spline Estimation of Detailed Fertility Schedules from Abridged Data. MPIDR Working Papers. Max Planck Institute for Demographic Research, Rostock, Germany.

Schmertmann, C.P., 2002. A simple method for estimating age-specific rates from sequential cross sections. Demography 39 (2), 287–310.
Schoumaker, B., 2011. Omissions of births in DHS birth histories in sub-Saharan Africa: measurement and determinants. In: Annual Meeting of the Population Association of America. Population Association of America, Washington, DC.
Schoumaker, B., 2013. Reconstructing long term fertility trends with pooled birth histories. In: XXVII IUSSP International Population Conference. Republic of Korea, Busan. XXVII IUSSP International Population Conference.
Siler, W., 1983. Parameters of mortality in human populations with widely varying life spans. Statistics in Medicine 2 (3), 373–380.
Silva, R., 2012. Child mortality estimation: consistency of under-five mortality rate estimates using full birth histories and summary birth histories. PLoS Med 9 (8), e1001296.
Smith, L., Hyndman, R., Wood, S., 2004. Spline interpolation for demographic variables: the monotonicity problem. Journal of Population Research 21 (1), 95–98.
Som, R.K., 1996. Practical Sampling Techniques, Statistics, Textbooks and Monographs, second ed. M. Dekker, New York, NY.
Spoorenberg, T., 2007. Quality of age reporting: extension and application of the modified Whipple's index. Population (english ed.) 62 (4), 729–741.
Spoorenberg, T., 2009. Is the Whipple's index really a fair and reliable measure of the quality of age reporting? An analysis of 234 censuses from 145 countries. In: XXVI IUSSP International Population Conference,Marrakech, Morocco. XXVI IUSSP International Population Conference.
Stolzenberg, R.M., Relles, D.A., 1997. Tools for intuition about sample selection bias and its correction. American Sociological Review 62 (3), 494–507.
Sullivan, J.M., 2008. An Assessment of the Credibility of Child Mortality Declines Estimated from DHS Mortality Rates. Working Draft; Revision 1. UNICEF, New York, NY.
Thatcher, A.R., Kannisto, V., Vaupel, J.W., 1998. The Force of Mortality at Ages 80 to 120, Odense Monographs on Population Aging Series. Odense University Press, Odense, Denmark.
Timaeus, I.M., 1991. Estimation of adult mortality from orphanhood before and since marriage. Population Studies 45 (3), 455–472.
United Nations, 1955. Manual II. Methods of Appraisal of Quality of Basic Data for Population Estimates, Population Studies. Series a: Manuals on Methods of Estimating Population. United Nations, New York, NY.
United Nations, 1982. Non-sampling Errors of Household Surveys: Sources, Assessment and Control, National Household Survey Capability Programme, Prelim. version. ed. United Nations, New York, NY.
United Nations, 1983. Manual X: Indirect Techniques for Demographic Estimation, Population Studies. United Nations, New York, NY.
United Nations, 1991. Handbook of Vital Statistics Systems and Methods. In: Legal, Organization and Technical Aspects, Studies in Methods Series F, vol. I. United Nations, New York, NY.
United Nations, 2002. Methods for Estimating Adult Mortality. United Nations, New York.
United Nations, 2005. Household Surveys in Developing and Transition Countries, Studies in Methods, Series F. United Nations, New York, NY.
United Nations, 2008a. Designing Household Survey Samples: Practical Guidelines, Studies in Methods. United Nations, New York, NY.
United Nations, 2008b. Principles and Recommendations for Population and Housing Censuses, Revision 2, Statistical Papers Series M, rev. second ed. United Nations, New York, NY.
United Nations, 2010a. Handbook on Population and Housing Census Editing: Revision 1, Studies in Methods Series F. United Nations, New York,NY.
United Nations, 2010b. Post Enumeration Surveys: Operational Guidelines (Technical Report). 2010 World Population and Housing Census Programme. United Nations, New York, NY.
United States. Bureau of the Census, 1985. Evaluating Censuses of Population and Housing, Statistical Training Document. U.S. Department of Commerce, Bureau of the Census, Washington, DC.
Walker, N., Hill, K., Zhao, F., 2012. Child mortality estimation: methods used to adjust for bias due to AIDS in estimating trends in under-five mortality. PLoS Med 9 (8), e1001298.
Wang, H., Dwyer-Lindgren, L., Lofgren, K.T., Rajaratnam, J.K., Marcus, J.R., Levin-Rector, A., Levitz, C.E., Lopez, A.D., Murray, C.J., 2012. Age-specific and sex-specific mortality in 187 countries, 1970-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 380 (9859), 2071–2094.
Wheldon, M.C., Raftery, A.E., Clark, S.J., Gerland, P., 2013a. Bayesian reconstruction of past populations and vital rates by age for developing and developed countries. In: Paper Presented to the Annual Meeting of the Population Association of America, New Orleans, LA.
Wheldon, M.C., Raftery, A.E., Clark, S.J., Gerland, P., 2013b. Bayesian Reconstruction of Two-sex Populations by Age: Estimating Sex Ratios at Birth and Sex Ratios of Mortality arXiv preprint arXiv:1312.4594.
Wheldon, M.C., Raftery, A.E., Clark, S.J., Gerland, P., 2013c. Reconstructing past populations with uncertainty from fragmentary data. Journal of the American Statistical Association 108 (501), 96–110.
Wilmoth, J., Zureick, S., Canudas-Romo, V., Inoue, M., Sawyer, C., 2011. A flexible two-dimensional mortality model for use in indirect estimation. Population Studies 66 (1), 1–28.
Wilmoth, J.R., Andreev, K., Jdanov, D., Glei, D.A., Boe, C., Bubenheim, M., Philipov, D., Shkolnikov, V., Vachon, P., 2007. Methods Protocol for the Human Mortality Database. University of California, Berkeley and Max Planck Institute for Demographic Research, Rostock. http://mortality.org (version 31.05.07.).
Wilmoth, J.R., Mizoguchi, N., Oestergaard, M.Z., Say, L., Mathers, C.D., Zureick-Brown, S., Inoue, M., Chou, D., 2012. A new method for deriving global estimates of maternal mortality. Statistics, Politics and Policy 3 (2).
Winship, C., Mare, R.D., 1992. Models for sample selection bias. Annual Review of Sociology 18, 327–350.
Yang, G., Hu, J., Rao, K.Q., Ma, J., Rao, C., Lopez, A.D., 2005. Mortality registration and surveillance in China: history, current situation and challenges. Population Health Metrics 3 (1), 3.
Zaba, B., 1979. The four-parameter logit life table system. Population Studies 33 (1), 79–100.
Zaba, B., 1985. A parameterized procedure for projecting population. In: International Population Conference. International Union for the Scientific Study of Population, Florence, Italy; Liege, Belgium, pp. 137–150.

Relevant Websites

http://demographicestimation.iussp.org/ – IUSSP Tools for Demographic Estimation.
http://www.census.gov/population/international/software/pas/ – U.S. Bureau of the Census, International Programs Center: Population Analysis System (PAS).
http://unstats.un.org/unsd/demographic/standmeth/ – UN Statistics Division: Standards and Methods.
http://www.un.org/en/development/desa/population/publications/manual/ – UN Population Division: Demographic Manuals.
http://www.demog.berkeley.edu/~eddieh/toolbox.html – Applied Demography Toolbox.

Demographic Techniques: Family Reconstitution

Cameron D Campbell, Hong Kong University of Science and Technology, Kowloon, Hong Kong

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E.A. Hammel, volume 5, pp. 3456–3461, © 2001, Elsevier Ltd.

Abstract

Family reconstitution is the technique of linking records of demographic events, usually of an ecclesiastical nature (baptisms, marriages, burials), within and between individual lives to recreate individual life histories and the histories of families. Methods of family reconstitution include purely manual linkage, computer-assisted manual linkage, and linkage solely by computer. Family reconstitution has given scholarship a remarkable window on the demographic past. In so doing it provides information for much longer spans of time than contemporary data and often for periods before any effects of industrialization or modernization.

Family reconstitution is the technique of linking records of demographic events, usually of an ecclesiastical nature (baptisms, marriages, burials), within and between individual lives to recreate individual life histories and the histories of families. It originated with studies that linked by name the originally separate parish records of baptisms, burials, and marriages (*nominal data linkage*) of family members. It revolutionized the study of population processes in the past, allowing estimation of demographic rates that were virtually impossible to obtain by other means. Extensions of the technique have been applied to related sources such as genealogies, *Ortssippenbücher* (local books of family histories), civil vital event registers, civil population registers, *Libri status animarum* (*Book of the State of Souls*), and *censuses*. Methods of family reconstitution include purely manual linkage, computer-assisted manual linkage, and linkage solely by computer. Some limitations of the technique include data quality, coverage, and representativeness; statistical issues; selection by quality of data; selection by migration, truncation, and censoring; and bias from delays in recording. Despite difficulties in quality and coverage of the data and the possibility of biased inference, family reconstitution has given scholarship a remarkable window on the demographic past. In so doing it provides information for much longer spans of time than contemporary data and often for periods before any effects of industrialization or modernization.

The Technique

Family reconstitution is the technique of linking records of demographic events within and between individual lives, in order to recreate individual life histories and the histories of families. Early studies largely made use of ecclesiastical records, but recent studies make use of a wider variety of data. While genealogists have always pursued such linking, the intent of demographers is not simply to record chains of descent and marriage but rather to compile information on the *demographic rates* pertaining to the *population* of which the individuals and families were a part.

Family reconstitution has allowed estimation of demographic rates in populations for which population statistics were not available. Demographers ordinarily estimate vital rates by comparing counts of events in vital records to counts of the population at risk in a census. For example, the crude birth rate is estimated by dividing the number of births in a year by the total size of the population. However, accurate censuses contemporary with records of vital events are rare in the past. Family reconstitution overcomes this difficulty by counting both events and persons. Application of family reconstitution has led to important insights. For example, it showed (in contrast to the expectations of Notesteinian theory) that the decline of fertility in some places occurred well before economic modernization and before any decline in mortality.

History and Scope of the Technique

Family reconstitution as a demographic technique was developed initially by Swedish demographers (Edin, 1915) and advanced by the French (Henry, 1967), then by English scholars (Wrigley, 1966; Wrigley et al., 1997). It linked by *names of persons* the parish records of baptisms, burials, and marriages of individuals (hence the technical phrase, *nominal data linkage*). Genealogists had always used the linking of disparate records by names of individuals. Its large-scale use by historical demographers was an innovation. Nominal data linkage is of course now commonplace, especially using unique identifiers, in the management of medical, insurance, taxation, police, and other data, especially by computer. In the hands of historical demographers, the ecclesiastical ceremonies noted in parish records served as *proxies* for the underlying births, deaths, and onsets of cohabitation. It must always be remembered that the events in parish records are not the same as the underlying demographic events, e.g., a baptism is not a birth. With that caveat, a family reconstitution should provide information on the dates of birth, marriage, widowhood, remarriage; dates of the birth of children; and dates of death for individuals. It is possible in principle to continue such linkages across succeeding generations, or across collateral lines. Historians and demographers can examine family reconstitutions for patterns of fertility, mortality, and nuptiality, and to a limited extent those of migration in historical populations. The technique can be extended to the study of any other type of social or biological phenomena supported by the particular

dataset or other data such as medical, notarial, probate, or tax records that can be linked to it, e.g., heritability of cause of death, marriage patterns between social groups, fertility by social class, etc.

The parish registers on which family reconstitution relies are most commonly available in Catholic and Protestant Europe, and former European colonies that are predominantly Catholic. While many studies using parish registers focus on single communities, there have been some more ambitious efforts, most notably the study by Wrigley et al. (1997) of a large sample of parishes in English, and ongoing, major efforts to reconstitute large portions of the population of Quebec from its settlement in the seventeenth century up to the present (Bouchard, 1992; Desjardins, 1998). Parish registers have also been used in historical demographic studies of the Philippines (Doeppers and Xenos, 1998) and Latin America.

Related Methods and Expansions

Parish registers are by no means the only source that allow for the reconstitution of individual life histories and family pedigrees by record linkage. In recent decades, researchers have identified a variety of other sources from other societies that also allow for family reconstitution. In some cases, they provide information not normally available from parish registers, e.g., on household composition.

Genealogies

Genealogies are records of the linked lives of persons related by consanguinity and affinity and can, like reconstituted families, be used to extract demographic information. They were especially important for early historical demographic studies of China and Korea, because many lineages maintained them for ritual and other use. These retrospective family genealogies are often very selective, may not be representative of the broader population (Campbell and Lee, 2002). Aristocratic and wealthy families and lineages are more likely to be recorded than those more characteristic of the population. Within genealogies, persons of high status are more likely to be remembered, and those who die early, never marry, or otherwise do not leave descendants are more likely to be forgotten. In East Asian genealogies, males are more likely to be remembered than females. Females may be less well remembered if they have fewer male children of consequence with progeny, so that events linked through females are likely to be lost (especially if the females were in informal liaisons with the males of the lineage).

Certain databases of reconstituted families produced from genealogies nevertheless have proven very valuable. One of the largest and best known is the Utah Population Database (Bean, 1990). It was originally constructed from family genealogies compiled by Mormon families in Utah and neighboring areas from the nineteenth century onward. Another prominent genealogical source are the *Ortssippenbücher* (local books of family histories) of some German regions, compiled by local genealogists from parish registers in the early twentieth century primarily to demonstrate the ethnic purity of families. These are genealogies of the common folk (but of course selective as their purpose indicates), and have been used for demographic estimation (Knodel, 1988). In China, the Qing Imperial Lineage genealogy was compiled by a government office and covered all the descendants of the great-grandfather of the founder of the Lineage from the seventeenth century to the beginning of the twentieth century. Because the data were compiled prospectively, not retrospectively, they do not suffer from the limitations of other Chinese family genealogies. These data have been used extensively for demographic estimation (Lee et al., 1993). Compared to other Chinese genealogies, these data record daughters much more completely, and provide much more detail on the social and economic status of adult males.

Civil Vital Event Registers

Civil registers of vital events have the same potential as parish registers. They are intended to cover all faiths in a civil jurisdiction, but they may underreport some population segments such as vagrants or the poor. In some regions, civil registers can serve as a continuation of earlier parish registers and are especially valuable if secularization progresses and observation of ecclesiastical ceremonies declines.

Population Registers

Historically, civil and ecclesiastical population registers were listings of the inhabitants of residential units. They have the advantage of preserving the identity of the residential unit (usually 'households') within which the demographic events occurred, although just what that unit was is itself often problematic, since actual and legal residence might be confused, and even long-term out-migrants might sometimes be continued in the listings. (Such continuation is not unknown in some parish records as well, e.g., births to out-migrant former inhabitants.) Some registers were compiled at fixed intervals. Other registers were compiled on a continuous basis, with entrances into and exits from the household records as annotations of a single entry. Registers accomplish what is most difficult in family reconstitution, namely the nominal data linkage within and between individual lives, because local recorders presumably knew the identities of the persons and would not ordinarily confuse one individual with another who had the same name. Thus, the families discernible in population registers are not so much reconstituted as preconstituted. The most ambitious use of population registers has been in comparative studies of Europe and Asia (Bengtsson et al., 2004; Tsuya et al., 2010). The longitudinal, demographic, and household data collected from population registers are perhaps best understood as the historical analogue of data generated by contemporary Demographic Surveillance Sites like the ones that are part of the INDEPTH network (Sankoh and Byass, 2012). The ongoing release of data from population registers and INDEPTH sites raises the prospect of comparison between historical and contemporary populations.

Population Registers Updated at Fixed Intervals

Population registers updated at fixed intervals resemble annual or triennial censuses in terms of organization and content. Each new update described the population at a specific point in time.

While such registers can be treated like cross-sectional data and analyzed with the same methods used for census data, they differ from censuses in that individuals and families may be followed from one to the next. The occurrence of vital events in the interval between updates is reflected in an annotation, or sometimes implied through changes in household membership or individual characteristics. The longer the interval between updates, the more uncertain are the precise dates of the events. These registers are typically analyzed by application of discrete-time event-history techniques.

The best-known such registers are from East Asia. The first such historical population registers to be used for demographic analysis are the annual household registers that have been preserved for a number of communities from Tokugawa Japan *c*.1650–1868 (Hayami, 1969; Smith et al., 1977). Triennial registers covering a large number of communities in eighteenth- and nineteenth-century northeast China have also been transcribed into databases, analyzed (Lee and Campbell, 1997), and publicly released (Lee et al., 2010). Triennial registers have also been transcribed and analyzed for several Korean communities before the twentieth century (Lee and Park, 2008)

Continuously Updated Population Registers

In household registers compiled on a continuous basis, the timing of vital events and movements into and out of the household are recorded precisely. The best known of these registers are the ones compiled in Japanese-occupied Taiwan *c*.1904–45 (Wolf and Huang, 1980), but registers were also kept in some European countries, e.g., Belgium (Gutmann and van de Walle, 1978) and Italy (Kertzer and Hogan, 1989). Data from these registers are typically analyzed using continuous-time event-history techniques, most commonly Cox regression.

In some European contexts, linkage of multiple sources has produced demographic databases that resemble these continuous registers. For example, ecclesiastical population registers such as Swedish catechetical records and Italian *Libri status animarum* (literally, 'Book of the State of Souls') that recorded household and individual characteristics every year have been linked to parish registers that provide dates of baptisms, marriages, and burials that took place between annual updates.

Censuses

Construction of large demographic databases from historical censuses and advances in automated record linkage now make possible the linkage of individuals across censuses, and the reconstitution of families. Until recently, efforts to follow individuals from one census to another were mostly unproductive because censuses were decennial or at best every 5 years, transcribed samples only covered small geographic areas, and geographic mobility was high. Transcription of the entire 1880 United States census, availability of name indexes for censuses up to 1940, and advances in automated nominative linkage, however, have recently led to renewed efforts to create longitudinal data from United States censuses (Ferrie, 1996; Ruggles, 2006). For the 1850 through 1930 United States censuses, samples of individuals linked to the 1880 census are already available at the Integrated Public Use Microsample Series (IPUMS) Web site.

Methods of Family Reconstitution

The work of family reconstitution was from its inception strongly dependent on the paleographic and historical skills and intuition of the scholar. The earliest efforts were done by hand, starting with a marriage record; then attempting to find the burial and any remarriage records of the spouses; then the baptismal records of the children; and then the marriages and deaths of those children. Such data were entered on a 'marriage fiche' or 'family reconstitution form' for later collation and analysis. Account had to be taken of the literacy of scribes, variations in spellings of the same names, variations in standardization of surnames, and of course the completeness of the records themselves. Later attempts were assisted by computers, but final judgments often still depended on intuition in deciding between two or more equally plausible linkages. Efforts have been made to completely automate the linkage procedure, including the use of mathematical programming methods to resolve ambiguities, although it is not clear whether the algorithms themselves may not simply substitute new for old biases (Wrigley and Schofield, 1973; Bouchard, 1992). It is difficult to know when the correctness of the nominal links in a dataset has been maximized, especially when data are scanty or contain errors. For example, if surnames are often omitted in the parish documents, and especially if the stock of first names is small (e.g., 'John married Mary, 1 June 1712'), the number of alternative linkages can be very large because many Johns may have married Marys on that date. In principle linear programming methods can be used to obtain the best solution, one that maximizes the density or quality of linkage, but that solution may not be historically correct. Estimates of central tendency (e.g., the general fertility rate) will not be biased by such presumably random error, but variances will be, as will be any associations with other variables if guesses are more likely to be bad for some classes of persons. The results of nominal record linkage may be compared to population registers, but population registers also have problems of accuracy. Such problems of optimizing linkages are of course less when data are detailed (e.g., Swedish parish and population register records) or contain ethnographically validated clues to linkage (e.g., interfamilial patterns of godparenting in some South Slavic contexts wherein children of a marriage tend to have the same godparents, who also serve as marriage witnesses, etc.).

Limitations of the Technique

Although a powerful tool, family reconstitution has important limitations caused by missing data. Some stem from the nature of the sources.

Data Quality and Coverage

While some parish records (e.g., Florence, Istria) go back to the fifteenth century, most do not commence until the sixteenth, and many are not reliable until the eighteenth century. Civil event and population registers are generally no older than the nineteenth century, but ecclesiastical registers are often as old as the corresponding event registers. *Libri status animarum*

(Book of the State of Souls) were only irregularly kept in many areas, and all ecclesiastical records were subject to variations in quality depending on the abilities and dedication of the recording priest, attempts by higher authorities (both ecclesiastical and civil) to collect sufficient and standardized information, and the vagaries of copying and preservation. Ecclesiastical records are known for Catholic and Protestant populations but seldom for Orthodox Christians (not to mention Muslims, Jews, or others). Thus in areas of mixed religion, e.g., in nominally Catholic Croatia or parts of Bosnia, large non-Catholic proportions of the population would not be included in the data and investigation of regional patterns would be hampered. Remarkable discontinuities in the data can occur in periods of shifts in religious adherence, as e.g., sharp swings between conformity and nonconformity in English parishes, or the widespread conversion of Catholics to Islam in the western Balkans. Similar problems would arise in using missionary records outside of Europe and the Americas. *Civil* event and population registers are by their nature relatively independent of religious bias, although they may miss subpopulations whose members for other reasons are excluded or otherwise underregistered.

Statistical Biases Caused by Sample Selection

Probably the most serious problems for family reconstitution are selection bias and truncation, issues addressed early by Henry (1967).

Selection by Data Quality

Where data are in general of acceptable but varying quality, there is an inherent tendency to want to analyze records that are only of the best quality. The socioeconomic condition and demographic regime of a part of a population that has good records may be quite different from those of another part that has poor records. Analysis limited to only the best data may give a misleading picture of the whole, either with respect to entire parishes or to subpopulations within parishes. This kind of selectivity bias is similar to that experienced in the analysis of genealogical data.

Selection by Migration

Because the work of linkage is so tedious, even with computer assistance, most reconstitutions have been limited to single parishes. Out-migration from such parishes will ordinarily go unremarked in parish records, leading to the truncation and censoring problems described in Section Truncation and Censoring. Since the kinds of persons who complete their lives within a parish may be quite different from the kinds who leave it, building a demographic picture of a parish based on completed lives may be misleading. These and similar problems (see below) can be partly overcome by analysis of a set of contiguous parishes, especially if migration tended to be short range. Migration effects may be powerful where migration flows were strong (Kasakoff and Adams, 1995). Reconstitution data have been used to examine migration per se (Bideau and Brunet, 1993). Since population registers typically provide dates of entry and exit into a population for migrants, they can also be used to study the characteristics of in- and out-migrants.

Truncation and Censoring

In the absence of linkage of parish registers to other sources during family reconstitution, the time actually spent in a parish by a person who was not born in it or did not die in it may be unknown, and thus the time at risk of some demographic event may not be known. Many demographic measures cannot be properly estimated if time at risk is not known. For example, a woman may have had recorded births but no recorded death. The amount of time she was actually at risk of having the next birth cannot then be accurately estimated. Thus, the fertility of women at her achieved parity or at ages higher than that of her last recorded birth cannot be estimated accurately. For example, estimating the total fertility rate for the population would demand knowledge of the age-specific fertility rates by the usual 5-year age groups or by single years, and the loss of information on some women who died without record within such spans would deflate the denominators of the rate calculations.

In general, observation on individuals at risk of some event is said to be truncated if the size of the cohort at risk to which the individuals belong is unknown and censored if the number of individuals at risk is known but the events are unobserved or if observed events are recorded only in time intervals instead of accurately. Examples are (1) death records lacking age so that date of birth is unknown and the number of persons in the birth cohort cannot be recovered from birth records (truncated on the left); (2) death records with age recorded but birth records unavailable so that information is obtained only on persons who died at each age, not on those who survived that age (truncated on the right); (3) no marriage records but death records with age information showing survival and marital status so that the birth cohort size is known but the date of marriage is not (censored on the left); (4) birth and marriage records but no information on death so that observation ceases on the date of marriage (censored on the right). If statistical methods take censoring into account, censoring induces no bias in the computation of rates or probabilities of occurrence of an event, or in event-history analyses. By contrast, truncation does induce such bias. As a general rule, the onset or cessation of observation of an individual at risk should not be correlated with the event of interest.

Solutions to the problem of missing data in sets of life histories are various but problematic. They include using the occurrence of another event to the same individual (e.g., next births if interest is in mortality), random censoring, occurrence of an event to a close family member (thus implying continued residence and at least potential registration of an event), or statistical imputation of the censoring point, etc., (Hammel, 1993; Trussell and Guinnane, 1993). Another alternative is to make use of population registers or other sources that record entrances and exits into the population, thereby allowing for the individuals at risk of an event at each point in time to be identified with some precision (Bengtsson et al., 2004; Tsuya et al., 2010).

Bias from Delays in Recording

Bias may also be created by delays in the recording of demographic events or in the ceremonies that recognize them. An individual may die or move before an event is recorded and therefore be missed. Such delay is especially problematic for

births and infant deaths. If the date of recording is accepted as the date of birth, long delays will make infants seem younger than they are, biasing any age-related computations. If an infant dies before it can be baptized, there will be no record of the birth, and often there will be no record of the death since unbaptized infants often received no formal burial. If the death is recorded and an age at death is given, the date of birth can be imputed and thus recovered. If death is unrecorded, mortality may be underestimated. Where both birth and death are unrecorded, infant mortality will also seem to be lower than it really is. Since in many historical populations up to a third of infants died in their first year, and half of those died in their first month, such statistical biases can be very important.

Prospects

The most promising outlook for family reconstitution and related techniques includes the extension and elaboration of computer-based programming methods for rigorous and testable nominal data linkage; the combination of multiple data sources such as parish event registers, population registers, tax and notarial records, and most recently historical censuses, to increase the accuracy and scope of data; and the use of event-history techniques to overcome problems of missing event dates.

Summary

Despite difficulties imposed by the quality and coverage of the historical data and the possibility of biased inference, family reconstitution has given demographic and historical scholarship a remarkable window on the demographic past, unachievable by any other means. In so doing it provides information for much longer spans of time than do modern data and often for periods before any effects of industrialization or modernization have been felt. It thus allows examination of longer term trends under broader socioeconomic conditions, and like all historical enterprises allows scholarship to escape the constraints imposed by modern conditions.

See also: Data Bases and Statistical Systems: Overview; Demographic Measurement: General Issues and Measures of Fertility; Demographic Techniques: Lexis Diagram; Event History Analysis in Continuous Time; Event History Analysis: Applications; Historical Demography.

Bibliography

Bean, L.L., 1990. The Utah population database: demographic and genetic convergence and divergence. In: Adams, J., Lam, D.A., Hermalin, A., Smouse, P.E. (Eds.), Convergent Issues in Genetics and Demography. Oxford University Press, New York, Oxford, pp. 231–244.

Bengtsson, T., Campbell, C., Lee, J., et al., 2004. Life under Pressure: Mortality and Living Standards in Europe and Asia, 1700–1900. MIT Press, Cambridge.

Bideau, A., Brunet, G., 1993. The construction of individual life histories: application to the study of geographical mobility in the Valserine Valley in the nineteenth and twentieth centuries. In: Reher, D., Schofield, R. (Eds.), Old and New Methods in Historical Demography. Clarendon Press, Oxford, UK, pp. 111–124.

Bouchard, G., 1992. Current issues and new prospects for computerized record linkage in the province of Quebec. Historical Methods 25, 67–73.

Campbell, C., Lee, J., 2002. State views and local views of population: linking and comparing genealogies and household registers in Liaoning, 1749–1909. History and Computing 14 (1+2), 9–29.

Desjardins, B., 1998. Le Registre de population du Quebec ancien. Annales de demographie historique 2, 215–226.

Doeppers, D., Xenos, P. (Eds.), 1998. Population and History: The Demographic Origins of the Modern Philippines. Ateneo de Manila University Press, and University of Wisconsin-Madison: Center for Southeast Asian Studies, Quezon City.

Edin, K.A., 1915. Studier i Svensk fruktsamhetsstatistik. Ekonomisk Tidskrift 9, 251–304.

Ferrie, J., 1996. A new sample of males linked from the public-use-microdata-sample of the 1850 US Federal Census of Population to the 1860 US Federal Census Manuscript Schedules. Historical Methods 29, 141–156.

Gutmann, M.P., van de Walle, E., 1978. New sources for social and demographic history: the Belgian population registers. Social Science History 2, 121–143.

Hammel, E.A., 1993. Censored intervals in family reconstitution: a sensitivity test of alternative strategies with historical Croatian data. In: Reher, D., Schofield, R. (Eds.), Old and New Methods in Historical Demography. Clarendon Press, Oxford, UK, pp. 125–144.

Hayami, A., 1969. Aspects démographiques d'un village japonais, 1671–1871. Annales Economies, Sociétés, Civilisations 3, 617–639.

Henry, L., 1967. Manuel de démographie historique. Droz, Paris.

Hyrenius, H., 1958. Fertility and reproduction in a Swedish population without family limitation. Population Studies 12, 121–130.

Kasakoff, A.B., Adams, J.W., 1995. The effect of migration on ages at vital events: a critique of family reconstitution in historical demography. European Journal of Population 11, 199–242.

Kertzer, D., Hogan, D.P., 1989. Family, Political Economy, and Demographic Change: The Transformation of Life in Casalecchio, Italy, 1861–1921. University of Wisconsin Press, Madison, WI.

Knodel, J.E., 1988. Demographic Behavior in the Past. A Study of Fourteen German Village Populations in the Eighteenth and Nineteenth Centuries. Cambridge University Press, New York, Cambridge, UK.

Lee, J., Campbell, C., 1997. Fate and Fortune: Social Organization and Population Behavior in Liaoning, 1774–1983. Cambridge University Press, Cambridge, UK.

Lee, J., Campbell, C., Chen, S., 2010. China Multi-Generational Panel Dataset, Liaoning (CMGPD-LN) 1749–1909. User Guide. Inter-University Consortium for Political and Social Research, Ann Arbor, MI.

Lee, J., Campbell, C., Feng, W., 1993. The last emperors: an introduction to the demography of the Qing (1644–1911) lineage. In: Reher, D., Schofield, R. (Eds.), Old and New Methods in Historical Demography. Clarendon Press, Oxford, UK, pp. 361–382.

Lee, S., Park, H., 2008. A survey of data sources for studies of family and population in Korean history. The History of the Family 13 (3), 258–267.

Reher, D., Schofield, R. (Eds.), 1993. Old and New Methods in Historical Demography. Clarendon Press, Oxford, UK.

Ruggles, S., 2006. Linking historical censuses: a new approach. History and Computing 14, 213–224.

Sankoh, O., Byass, P., 2012. The INDEPTH network: filling vital gaps in global epidemiology. International Journal of Epidemiology 41 (3), 579–588.

Smith, T.C., Eng, R.Y., Lundy, R.T., 1977. Nakahara: Family Farming and Population in a Japanese Village, 1717–1830. Stanford University Press, Stanford, CA.

Trussell, J., Guinnane, T., 1993. Techniques of event-history analysis. In: Reher, D., Schofield, R. (Eds.), Old and New Methods in Historical Demography. Clarendon Press, Oxford, UK, pp. 181–205.

Tsuya, N., Wang, F., Alter, G., Lee, J., et al., 2010. Prudence and Pressure: Reproduction and Human Agency in Europe and Asia, 1700–1900. The MIT Press, Cambridge, MA.

Wolf, A.P., Huang, C.S., 1980. Marriage and Adoption in China, 1845–1945. Stanford University Press, Stanford, CA.

Wrigley, E.A. (Ed.), 1966. An Introduction to English Historical Demography from the Sixteenth to the Nineteenth Centuries. Basic Books, New York.

Wrigley, E.A., Schofield, R.S., 1973. Nominal record linkage by computer and the logic of family reconstitution. In: Wrigley, E.A. (Ed.), Identifying People in the Past. Edward Arnold, London, pp. 64–101.

Wrigley, E.A., Davies, R.S., Oeppen, J.E., Schofield, R.S., 1997. English Population History from Family Reconstitution, 1580–1837. Cambridge University Press, New York, Cambridge, UK.

Demographic Techniques: Indirect Estimation

Tom A Moultrie, Centre for Actuarial Research (CARe), University of Cape Town, Rondebosch, South Africa

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by K. Hill, volume 5, pp. 3461–3464, © 2001, Elsevier Ltd.

Abstract

This article describes the principles and characteristics of techniques of indirect estimation used to derive estimates of demographic parameters from limited and/or defective data. The two broad classes of estimation techniques are described, giving examples. While the hope was that eventually these methods would be made redundant by improved data collection procedures, the suite of methods is still in widespread use half a century after they were first proposed.

'Indirect estimation' refers to the range of methods used by demographers to estimate rates of fertility, mortality (and, to a much lesser extent, migration) from data that are limited or defective, or both. The origin of the term stems from the pioneering work of the demographer William Brass in Africa in the 1960s. With the first censuses conducted after independence it was apparent, due to the observed deficiencies in the census data and the substantial incompleteness of vital registration systems of births and deaths, as well as the high levels of illiteracy and innumeracy in those populations, that direct estimation of rates of fertility and mortality, based on an analysis of the number of events occurring in a defined period of time and the exposure to risk of the population over that same period, could not produce robust and reliable estimates. Brass' important insight was that answers to simple questions, which might be better collected by enumerators and reported by respondents, could be manipulated by means of demographic models of fertility and mortality in conjunction with established mathematical relationships between demographic quantities to produce better estimates of the parameters of interest, and be robust to common patterns of error in the data.

A second category of indirect methods are those that seek to correct directly derived estimates of demographic parameters for commonly encountered errors in demographic data. Methods in this category use established relationships between demographic quantities to evaluate the consistency of the data, highlight possible sources of error, and adjust the directly observed rates accordingly.

Although these categories are rather different in their development and application, they nonetheless share the important characteristic of being based either on mathematically or logically robust relationships between different aspects of demographic data, or use model patterns of fertility, mortality, or migration to estimate or correct the parameters. To facilitate the process of demographic estimation by users who may not have extensive training in demographic estimation, parsimony of data requirements and simplicity of application by the end user are other important hallmarks of these methods.

While the methods derived by Brass and his collaborators have been developed, extended, and refined by many other demographers over the last 50 years, the approaches now in use are rooted in Brass' work, and still find use across the developing world. The material presented in this article cannot hope to provide even a summary description of all the methods that have been developed. Most of them were first written up comprehensively in *UN Manual X* (UN Population Division, 1983), the first systematic attempt to document the methods of demographic estimation from limited and defective data. Aspects of this work were revised and updated by Sloggett et al. (1994) and again (in the case of methods for estimating adult mortality) by the UN Population Division (2002). More recently, a project to revisit, revise, and update the canon of indirect methods of demographic estimation has been completed (Moultrie et al., 2013).

Indirect Estimation

The first category of methods of indirect demographic estimation seeks to produce more reliable and robust estimates of demographic parameters indirectly; that is, using information only loosely related to the parameters desired, and relying heavily on mathematical manipulation using demographic models to produce estimates of the desired parameters. These methods, all of which require very limited data as inputs, have been described by Ansley Coale – a collaborator with Brass in his work on African demography in the 1960s – as being akin to one of "making bricks with very little straw" (Coale, 2000: 55). The foundational logic of several of these methods is outlined below.

Estimation of Child Mortality

The classic, and still most widely applied, indirect method is what has come to be known as the Brass children surviving–children ever born method of estimating child mortality. The method, first set out in Brass (1964) and elaborated on in the seminal *Demography of Tropical Africa* (Brass et al., 1968), makes use of very simple summary data collected from women: how many children had they ever given birth to and how many of those were still alive, and transforms these inputs into estimates of the probability of a child dying between birth and a defined set of ages. Brass' insight was that the proportion of children dead of those ever born to women is closely associated with the age of the mother and the overall level of child mortality. Accordingly, using mathematical models of fertility and mortality, which exhibit strong regularities across a variety of demographic settings, Brass showed

Table 1 Brass' rubric for determining the ages to which the cumulative probabilities of mortality should apply, based on mother's age

Mother's age	*Best estimate of the age, x, to which the cumulative probability of mortality, q(x), applies*
15–19	1
20–24	2
25–29	3
30–34	5
35–39	10
40–44	15
45–49	20

how to convert these proportions of dead, controlling for the age of the mother, into cumulative probabilities of child mortality by age x, $q(x)$. Specifically, he showed that these proportions of dead could be mapped with some robustness (considering the poor quality or limited nature of the original data used as inputs) onto cumulative probabilities of death as shown in Table 1.

While Brass' approach has been refined and improved upon several times since its first exposition, the single most important contribution to the method was Feeney's (1980) observation that when mortality cannot be assumed to be constant over time, approximate reference dates could be put on the estimates of cumulative probabilities of mortality from different age groups of mother by assuming that changes in mortality are approximately linear over time.

A further refinement was to assume that the age pattern of mortality in a given setting follows a constant shape, with the level of mortality changing over time. This assumption, in conjunction with the Brass relational logit system of life tables, allows those different estimates of child mortality to be further converted into estimates of a single parameter, for example, $q(5)$, the probability of death before age 5, and for those estimates of the same parameter to be located in time using Feeney's approach, thereby allowing the identification of an approximate trend in mortality over time.

An example of the application of the method is presented in Figure 1, and shows the estimates of under-5 mortality in Malawi from the 1987 Malawi Census. Estimates of the same parameter from the United Nations Population Division are shown for comparison. A commonly encountered bias in the indirect method is indicated by the last data point. This estimate is derived from the reports of children ever born and children surviving from women aged 15–19 at the time of the census. The children of women who give birth at very young ages are more likely to die than children of women who give birth at older ages, resulting in the suggestion that child mortality in the country may have been increasing in the period just before the census. In most practical applications, this data point is ignored.

As Hill (2001: 3462) has noted, "[t]he Brass method and the developments of it have been very widely applied. It revolutionized knowledge of mortality conditions, particularly in sub-Saharan Africa but also in Latin America. It also gave rise to a concerted effort to find indirect estimation methods for adult mortality and migration. As such, it has proved one of the most influential ideas of demography." More recently, however, the method's applicability has been compromised by the spread of the HIV epidemic in sub-Saharan Africa. Most crucially, the epidemiology of HIV means that one of the central underlying assumptions of the method – namely the independence of the mortality of mothers and their children – is violated. Significant effort has been expended in recent years to further modify Brass's approach to accommodate the effects of HIV, largely without success. Even with the most promising approach developed so far, proposed by Ward and Zaba (2008), difficulties are encountered if the level of HIV prevalence in the population has changed over the preceding 15 or so years.

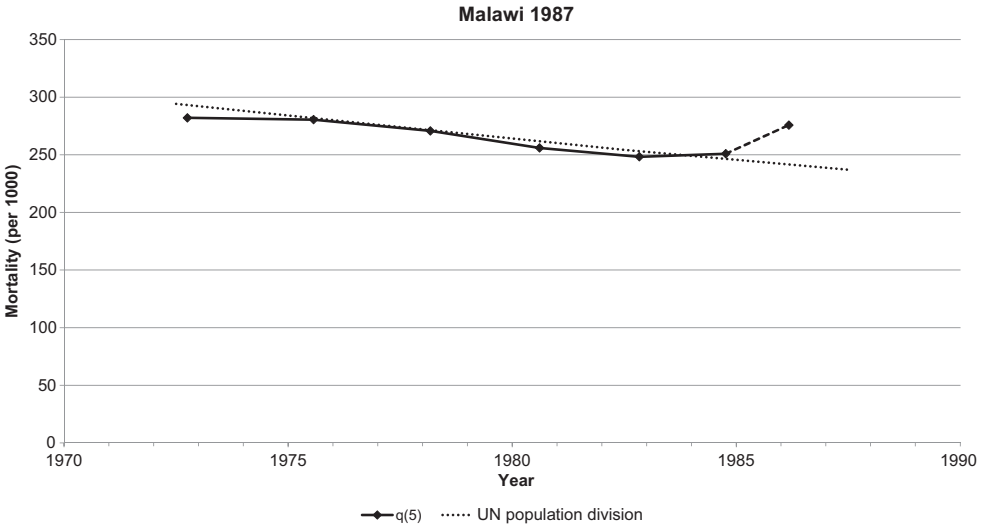

Figure 1 Estimates of under-5 child mortality for Malawi from the 1987 Malawi Census using the Brass children ever born-children surviving technique. Data from the 1987 Malawi Census held by IPUMS (ipums.org). Estimates derived using the Princeton 'North' Model Life table and the spreadsheet available at demographicestimation.iussp.org.

Estimation of Adult Mortality

Following the success of the Brass method for estimating child mortality, similar approaches were developed by Brass and his colleagues to estimate adult mortality using responses on the proportion – by respondents' age – of their mothers and fathers, spouses, and siblings who are reported to have died (respectively, the maternal and paternal orphanhood methods, the widowhood method, and the siblinghood method). The last of these has found particular use in the estimation of maternal mortality, and in the estimation of adult mortality from data collected in surveys where the sample size is typically too small to produce reliable estimates of adult mortality but, by asking about the survival of (typically in developing countries) relatively large numbers of siblings, allows mortality rates to be estimated.

As with the Brass children ever born–children surviving method, the robustness of the widowhood and orphanhood methods has been somewhat compromised by the effects of HIV/AIDS, although a simple adjustment has been proposed by Timæus and Nunn (1997) to compensate for this. In addition, problems with the application of the orphanhood method have been identified in many countries where adult death rates are high and fostering of the children of dead adults is common; in these situations, it is not uncommon for respondents to report on the vital status of their foster, rather than their biological, parents, resulting in an underestimation of mortality.

Estimation of Migration

Attempts have been made (Hill, 1987b; Zaba, 1985, 1987) to devise simple questions that can be asked in censuses that can be used to estimate migration in a similar fashion to the methods outlined above. These methods have not found widespread use, and doubts persist as to their robustness.

Estimation of Fertility

Estimates of trends in fertility can also be derived using indirect methods. The age structure of a population contains information that allow past trends in fertility to be measured. The population enumerated at a point in time of a given age, x to $x + 1$, represent the survivors of births that must have occurred x to $x + 1$ years previously. If the population can be assumed to be closed to migration, or if migration is negligible, an estimate of the mortality of successive cohorts of births allows one to estimate the number of births in preceding years. Furthermore, since it is possible to back-project the entire population, or that of women of reproductive age, to each preceding year using the same approach (or more simply, by assuming that the population has grown exponentially at a specified rate), it is possible to derive an estimate of either the crude birth rate or the general fertility ratio for approximately 15 years before the enumeration. (The limit of 15 years is placed to avoid the potential for in-or out-migration of young adults to bias the back-projected estimates). Despite, or perhaps because of, its simplicity, the method as outlined above has not been particularly useful, not least because it cannot produce estimates of age-specific fertility rates or of the total fertility rate.

This last limitation was addressed with the development of the 'own-child' method of fertility estimation (Cho et al., 1986). The method, further refined by Avery et al. (2013), relies on the ability to link children enumerated in a census in a household to their mothers present in the same household, and to thereby establish the mother's age at the time the child was born. Complex algorithms have to be employed to perform the matching, and those children whose mother cannot be identified in the same household have to have an appropriate mother's age imputed to them. Where children almost invariably live with their parents, the extent of the imputation required may be quite small; where fostering of children, or cohabitation in extended family groups, is widespread, the effects may be rather greater.

Recently, an alternative approach for estimation of fertility using reverse survival techniques has been proposed by Timæus and Moultrie (2013). This method has the advantage of providing the metrics desired (age-specific and total fertility rates) but avoids the issues associated with having to link children to their mothers by assuming that the shape of the age-specific fertility changes in a systematic and gradual fashion over time.

Consistency Checks

The second category of indirect techniques are those that rely on comparisons of two slightly different measures related to the same demographic parameter, and using the results to derive more robust estimates of that parameter.

Fertility

The Brass P/F ratio method, also developed by William Brass, is one such example. Brass had hypothesized (and evidence suggests that the hypothesis is generally correct) that, with census data on fertility, two specific errors frequently manifest themselves. The first relates to women's reports of lifetime fertility. Brass argued that older women's reports of lifetime fertility would tend to be biased systematically downward, particularly in very high fertility, and less literate or numerate environments, as a result of women neglecting to report dead children as such, or as a consequences of a failure to accurately estimate the exposure to risk. Conversely, he argued, younger women's reports of their lifetime fertility would be more accurate.

The second error hypothesized by Brass was that while women's recent reports – that is, collected on the basis of the number of births in the preceding year, or women's reports of fertility in a defined period of time – are frequently underreported, this underreporting is mostly independent of the age of the respondent. Consequently, while the level of fertility implied by direct estimates from the data may be too low, the shape of the fertility distribution would remain substantially correct.

Brass observed that if fertility had been constant, the cumulated period fertility up to a given age should be equal to the lifetime fertility of women of that age. Further, since

the lifetime fertility of younger women is presumed to be reasonably accurate, the ratio of lifetime to cumulated period fertility at younger ages provides a basis for adjusting the observed fertility schedule. The trend in these ratios by age of respondent also provides an important diagnostic of recent trends in fertility and possible errors in the data (Brass, 1975).

While the method has been in use almost without modification since it was first proposed, a significant refinement of the model by Zaba (1981) and Booth (1984) in the form of the relational Gompertz model (described in detail by Moultrie (2013)) allows the assumption of constant fertility to be relaxed. While this model is computationally rather different from the Brass *P*/*F* method, it is based on the same underlying logic of comparing lifetime and recent fertility to adjust the observed fertility schedules.

A number of other methods based on the *P*/*F* approach, allowing for estimation of fertility from two sets of census data have been derived (UN Population Division, 1983; Zlotnik and Hill, 1981). While these methods are highly sensitive to differential errors between censuses in recording lifetime fertility, an advantage is that they allow the assumption of constant fertility to be relaxed. These have all been recast as variants of the relational Gompertz method in the updated manual of indirect techniques, *Tools for Demographic Estimation* (Moultrie et al., 2013).

Adult Mortality

Indirect methods based on consistency checks have played an increasingly important role in the estimation of adult mortality. While the denominators used to estimate adult mortality rates can be sourced from census data, the numerator (that is, the number of deaths by age) comes from either a vital registration system, or – quite commonly – answers to questions in the census about the number of deaths that occurred in each household over a defined period before the census. Both sources of the numerator are prone to underreport the number of deaths, either because the vital registration system is incomplete or because not all adult deaths occurring in a household in the defined period are reported accurately.

Two similar, but nonetheless distinct, methods have been developed to correct the recorded or reported numbers of deaths for incompleteness. The first is the generalization of the Growth Balance method, first described by Brass (1975). Brass' original formulation assumed that the growth rate in the population had been stable for an extended period of time and that migration was negligible. Under these conditions, the rate at which people enter any open-ended age interval ($a+$) is equal to the sum of the (stable) population growth rate and the death rate of those in that age interval. However, this death rate is usually understated as a result of not all deaths being accurately recorded or reported. Brass' insight here was that if the completeness of death recording or reporting was constant with respect to age, the observed death rate in the open-ended interval $d^{o}(a+)$ could be expressed as the product of the actual death rate in the interval, $d(a+)$ and the proportion of deaths, c, that were correctly recorded or reported.

Hence, mathematically,

$$b(a+) = r + (1/c)d^{\circ}(a+)$$

where, in addition to the terms defined above, $b(a+)$ is the rate at which people enter the open interval, $a+$, which can be derived from the stable population age structure alone, and r is the stable population growth rate. In this formulation, then, there is a linear relationship between b and d, with the intercept being given by r, and the slope by the reciprocal of the completeness of death registration, c. With this estimate of c, the true estimate of the death rate in the open interval can be derived, and hence the reported or recorded deaths can be adjusted for the extent of underreporting.

In a major advance, Hill (1987a) extended Brass' idea and showed how it could be generalized to populations with unstable, or variable, growth rates using data from two censuses. Hill's generalization, termed the General Growth Balance method, further allows for the possibly differential completeness of the two censuses to be taken into account in estimating the level of completeness.

The second approach, developed by Bennett and Horiuchi (1981, 1984) and termed the 'Synthetic Extinct Generations' method, relies on the observation that, in a population closed to migration, the number of people alive in the population aged a and over must – since everyone must die – equal the number of deaths that will occur in the future of people aged a and over. In a stationary population (that is, one that has a constant birth rate equal to a constant death rate, and hence with a growth rate of zero), this number of future deaths will also be the number of annual deaths of those aged a and over. Bennett and Horiuchi further showed that this reasoning and their method could be generalized to nonstable populations where the growth rate in the population is not zero, using data from two censuses.

An example of the application of this method is presented by Dorrington (2013) using data from the 2001 Census and the 2007 Community Survey in South Africa. The results are shown in **Figure 2**. Although the model life table used to smooth the results allows for the effects of AIDS on mortality, the fit is not particularly good, but nonetheless is much better than those produced when using a life table that does not incorporate the effects of AIDS.

There is some debate as to which of these two consistency-based approaches to estimating adult mortality is preferable. Hill et al. (2009) suggest, based on the results of a series of simulations, that combining the two methods, using the General Growth Balance method to estimate the change in completeness of coverage of the data from the two censuses, and then using this as an input in the Synthetic Extinct Generations method produces more robust results than either method applied separately. However, Dorrington and Timæus (2008) observed that this was not the case in populations severely affected by AIDS.

Future Directions

Indirect estimation methods were not intended as a long-term replacement for direct estimation of demographic rates from census and vital registration data; the methods were derived as

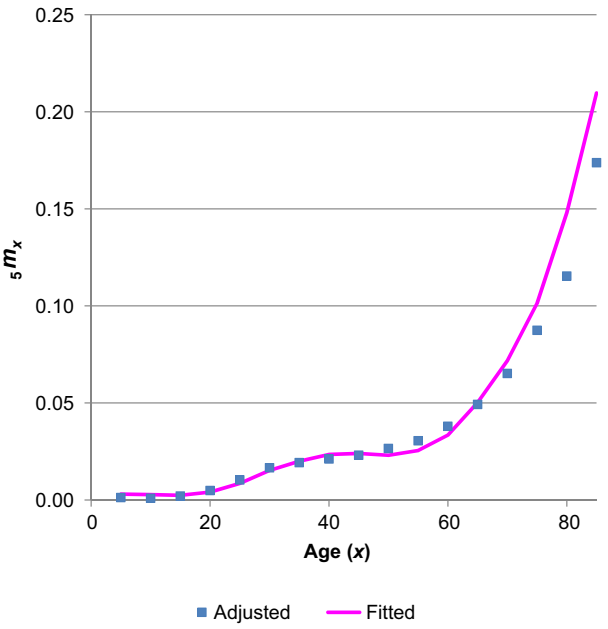

Figure 2 Estimation of male adult mortality in South Africa, applying the synthetic extinct generations method to data from the 2001 South African census and the 2007 Community Survey. Source: Dorrington, R.E., 2013. The synthetic extinct generations method. In: Moultrie, T.A., Dorrington, R.E., Hill, A.G., Hill, K., Timæus, I.M., Zaba, B. (Eds.), Tools for Demographic Estimation. International Union for the Scientific Study of Population, Paris. http://demographicestimation.iussp.org/content/synthetic-extinct-generations-methods. Material on this site is licensed under a CC BY-NC-SA license.

a stopgap way of estimating rates from limited and defective data, with the hope that, in time, improvements in census data collection and improving completeness of vital registration systems would render the methods obsolete. Unfortunately, progress in this regard in developing countries, and especially in sub-Saharan Africa, has been slow. Consequently, the canon of indirect methods still plays an important role in demographic estimation where the quality of data remains poor, or – increasingly – in situations such as in the demographic estimation of parameters for small areas where the data are too sparse to permit direct calculation. With much greater quantities of data now available, even from developing countries, work still continues on validating, refining, and improving the suite of indirect methods using simulation techniques as well as comparison with gold standard data. A recent example is the work of Masquelier (2013) in his investigations of the nature of biases in the estimates of adult mortality produced by the siblinghood methods.

See also: Demographic Measurement: General Issues and Measures of Fertility; Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Demographic Models; Demographic Techniques: Data Adjustment and Correction; Demographic Techniques: Inverse Projection; Demographic Techniques: Small-area Estimates and Projections; Microsimulation in Demographic Research; Population Dynamics: Theory of Nonstable Populations.

Bibliography

Avery, C., St. Clair, T., Levin, M., Hill, K., 2013. The 'own children' fertility estimation procedure: a reappraisal. Population Studies 67 (2), 171–183. http://dx.doi.org/10.1080/00324728.2013.769616.

Bennett, N.G., Horiuchi, S., 1981. Estimating the completeness of death registration in a closed population. Population Index 47 (2), 207–221. http://dx.doi.org/10.2307/2736447.

Bennett, N.G., Horiuchi, S., 1984. Mortality estimation from registered deaths in less developed countries. Demography 21 (2), 217–233. http://dx.doi.org/10.2307/2061041.

Booth, H., 1984. Transforming Gompertz' function for fertility analysis: the development of a standard for the relational Gompertz function. Population Studies 38 (3), 495–506. http://dx.doi.org/10.2307/2174137.

Brass, W., December 1964. Uses of census or survey data for the estimation of vital rates. In: Paper Prepared for the African Seminar on Vital Statistics, Addis Ababa 14–19 1964. Document No. E/CN.14/CAS.4/V57. United Nations, New York.

Brass, W., 1975. Methods for Estimating Fertility and Mortality from Limited and Defective Data. Carolina Population Centre, Chapel Hill, NC.

Brass, W., Coale, A.J., Demeny, P., Heisel, D.F., et al. (Eds.), 1968. The Demography of Tropical Africa. Princeton University Press, Princeton, NJ.

Cho, L.J., Retherford, R.D., Cho, M.K., 1986. The Own-Child Method of Fertility Estimation. University of Hawaii Press, Honolulu.

Coale, A.J., 2000. Ansley J. Coale: An Autobiography. American Philosophical Society, Darby, PA.

Dorrington, R.E., 2013. The synthetic extinct generations method. In: Moultrie, T.A., Dorrington, R.E., Hill, A.G., Hill, K., Timæus, I.M., Zaba, B. (Eds.), Tools for Demographic Estimation. International Union for the Scientific Study of Population, Paris. http://demographicestimation.iussp.org/content/general-growth-balance.

Dorrington, R.E., Timæus, I.M., 2008. Death distribution methods for estimating adult mortality: sensitivity analysis with simulated data errors, revisited. In: Paper Presented at Annual Meeting of the Population Association of America. Louisiana, New Orleans, 17–19 April 2008.

Feeney, G., 1980. Estimating infant mortality trends from child survivorship data. Population Studies 34 (1), 109–128. http://dx.doi.org/10.1080/00324728.1980.10412839.

Hill, K., 1987a. Estimating census and death registration completeness. Asian and Pacific Census Forum 1 (3), 8–13, 23–24. http://hdl.handle.net/10125/3602.

Hill, K., 1987b. New approaches to the estimation of migration flows from census and administrative data sources. International Migration Review 21 (4), 1279–1303. http://dx.doi.org/10.2307/2546515.

Hill, K., 2001. Demographic techniques: indirect estimation. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Pergamon, Oxford, pp. 3461–3464. http://www.sciencedirect.com/science/article/pii/B0080430767020945.

Hill, K., You, D., Choi, Y., 2009. Death distribution methods for estimating adult mortality: sensitivity analysis with simulated data error. Demographic Research 21 (Article 9), 235–254. http://dx.doi.org/10.4054/DemRes.2009.21.9.

Masquelier, B., 2013. Adult mortality from sibling survival data: a reappraisal of selection biases. Demography 50 (1), 207–228. http://dx.doi.org/10.1007/s13524-012-0149-1 (Online: 2 October 2012).

Moultrie, T.A., 2013. The relational Gompertz model. In: Moultrie, T.A., Dorrington, R.E., Hill, A.G., Hill, K., Timæus, I.M., Zaba, B. (Eds.), Tools for Demographic Estimation. International Union for the Scientific Study of Population,, Paris. http://demographicestimation.iussp.org/content/relational-gompertz-model.

Moultrie, T.A., Dorrington, R.E., Hill, A.G., Hill, K., Timæus, I.M., Zaba, B. (Eds.), 2013. Tools for Demographic Estimation. International Union for the Scientific Study of Population, Paris. Available at: demographicestimation.iussp.org.

Sloggett, A., Brass, W., Eldridge, S.M., Timæus, I.M., Ward, P., Zaba, B. (Eds.), 1994. Estimation of Demographic Parameters from Census Data. Statistical Institute for Asia and the Pacific, Tokyo.

Timæus, I.M., Moultrie, T.A., 2013. Estimation of fertility by reverse survival. In: Moultrie, T.A., Dorrington, R.E., Hill, A.G., Hill, K., Timæus, I.M., Zaba, B. (Eds.), Tools for Demographic Estimation. International Union for the Scientific Study of Population, Paris. http://demographicestimation.iussp.org/content/reverse-survival-methods.

Timæus, I.M., Nunn, A.J., 1997. Measurement of adult mortality in populations affected by AIDS: an assessment of the orphanhood method. Health Transition Review 7 (Suppl. 2), 23–43. http://www.jstor.org/stable/pdfplus/40652324.pdf.

UN Population Division, 1983. Manual X: Indirect Techniques for Demographic Estimation. United Nations, Department of Economic and Social Affairs, New York. ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html.

UN Population Division, 2002. Methods for Estimating Adult Mortality. United Nations, Department of Economic and Social Affairs, New York. ESA/P/WP.175. http://www.un.org/esa/population/techcoop/DemEst/methods_adultmort/methods_adultmort.html.

Ward, P., Zaba, B., 2008. The effect of HIV on the estimation of child mortality using the children surviving/children ever born technique. Southern African Journal of Demography 11 (1), 39–73.

Zaba, B., 1981. Use of the Relational Gompertz Model in Analysing Fertility Data Collected in Retrospective Surveys. Centre for Population Studies Research Paper 81–82. Centre for Population Studies, London School of Hygiene & Tropical Medicine, London.

Zaba, B., 1985. Measurement of Emigration Using Indirect Techniques: Manual for the Collection and Analysis of Data on Residence of Relatives. Liège, Belgium. Ordina Editions.

Zaba, B., 1987. The indirect estimation of migration: a critical review. International Migration Review 21 (4), 1395–1445. http://dx.doi.org/10.2307/2546519.

Zlotnik, H., Hill, K., 1981. The use of hypothetical cohorts in estimating demographic parameters under conditions of changing fertility and mortality. Demography 18 (1), 103–122. http://dx.doi.org/10.2307/2061052.

Relevant Website

demographicestimation.iussp.org – The methods described here are all written up, together with spreadsheets that implement the methods, in the recently published *Tools for Demographic Estimation*.

Demographic Techniques: Inverse Projection

Elisabetta Barbi, Sapienza University of Rome, Rome, Italy

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R. McCaa, volume 5, pp. 3464–3471, © 2001, Elsevier Ltd.

Abstract

Inverse projection, designed in the late 1960s by Ronald Lee, is a method for reconstructing populations where vital registration data are available but age details are scarce and population census is lacking or unreliable, as it is often the case of historical populations. Since its introduction, several modifications have been made to Lee's original procedure in order to address some of its shortcomings and to make the procedure suitable to the existing sources. Inverse projection and its developments have produced important innovations in the field of population reconstruction.

Origins

Inverse projection (IP) is a method for reconstructing populations where vital registration data are available but age details are scarce and population censuses are lacking or unreliable, as it is often the case of historical populations. The technique has been devised by Ronald Lee in the late 1960s as part of his PhD dissertation, which was a study of the broad macro determinants and consequences of aggregate population change in preindustrial England (Lee, 2004). It then appeared in an article published in 'Population Studies' in 1974. IP estimates accurate demographic indicators – life expectancy, gross and net reproduction ratios – and population age structures just using time series of deaths and births and an estimate of the initial population size. The projection is hence logically inverted because instead of using age-specific rates to estimate totals of births and deaths and population age structures, as conventional projection techniques do, it derives age-specific rates and population age structures from counts of births and deaths.

The inverted projection routine offers a solution to the problem of analyzing the demographic characteristics of past populations in the case of lack and incompleteness of sources. As in the case of family reconstitution – the microdemographic technique devised by Louis Henry in the 1960s – IP is suitable to the available documentation for past populations because it uses parish registration data, which for European countries before the nineteenth century are often more available than data on the size of the population and its age and sex structures. However, unlike Henry's nominative technique, IP requires only aggregate data thus overcoming some important limits that result from the family reconstitution and that were at the center of a lively debate among scholars of the time. 'Moving families', which may differ from 'stable families' for social and demographic characteristics, are difficult to detect and thus they can escape the family reconstitution. Hence given the difficulty of extending the reconstruction to all households in the village studied, results from the nominative technique may result unrepresentative. Other obstacles refer to the inability to aggregate results related to single villages and to consider these as representative of wider areas. Furthermore, although researchers today can avail themselves of computer programs for linkage, it must be recognized the enormous amount of energy needed to reconstruct a village made up of only a few families. Instead, with little effort, IP is able to reconstruct the demographic history of entire, even large populations over long periods of time, producing surprisingly accurate demographic indicators, by requiring only aggregate data. As Lee has pointed out (2004) "Some good ideas just don't work out in practice, while other ideas that seem to be based on very questionable assumptions, surprise us by working better than one could reasonably expect. Inverse Projection certainly falls in this second category". However, it must be said that historical demographers today are broadly in agreement in considering the aggregative technique of IP and the nominative reconstitution of families as two different techniques to be used profitably in a quite complementary way.

Method

In the study of past populations, the lack of complete data inevitably leads to the use of techniques that rely on models. This is also the case of IP that, in fact, presupposes the use of age schedules of mortality, fertility, and migration. To initiate the projection, IP requires the total population at the beginning of the period, an initial age structure of the population (obtained from either a census or a model), and counts of births and deaths for each interval. Under the hypothesis of closed population, to project the initial population at the subsequent instant, deaths occurred in the time interval must be subtracted for each age. But, since only the total number of deaths is known, this total is distributed by age through a series of probabilities of death that, applied to a given age structure, gives precisely the number of deaths observed in the considered time interval. This particular series of probabilities is determined by linear interpolation between two adjacent age schedules of mortality chosen from a single-parameter family of life tables. An adjustment factor, k_t, which is proportional to the difference of the two age schedules chosen as standard, locates the current series of probabilities of death. This variable can be viewed as an index of the force of mortality.

By removing the hypothesis of closed population, the procedure requires the knowledge of the total population not only at the beginning, but also at the end of the period considered, or, better, at instants subsequent to the initial. The

projected total, derived just adding the counts of births and subtracting the counts of deaths, is subtracted from the observed total population to obtain an estimate of net migration. Net migration is then apportioned equally among the intervening periods. In case of long periods, net migration is apportioned among the intervening periods through a more sophisticated solution based on the amount of the population and the number of intervals included in the period. Once the number of migrants in each interval is determined, they are distributed by age according to a standard pattern of age-specific rates.

Once the age distribution of the population is estimated, it is then possible to estimate the most important measures of mortality and fertility. The gross reproduction ratio is estimated from the number of birth, the age structure of the population, and a normalized age pattern of fertility which changes over time by a proportional factor, f_t, at all ages. This variable can be viewed as an index of the force of fertility. Models of fertility, therefore, affect only the estimates of the relative measures, since they are not used to estimate the number of births that are already known at the beginning of the procedure.

Theoretically, IP can generate estimates of nuptiality measures. Of course, in order to obtain these estimates, it is necessary to consider as inputs an age schedule of marriage, the marital status of the initial population, and the time series of marriages. But apart from the objective (and not negligible) difficulty to find these information for historical populations, IP that, also in this case, as for mortality and fertility, assumes a model of marriage rates with proportional changes over time at all ages, is not suitable to the analysis of a complex phenomenon such as nuptiality (Brunborg, 1976 cited in Lee, 2004; Breschi, 1990).

Over time, some extensions have been added to the original IP formulation. The first enhancement has been the extension of the number of age groups beyond age 55. The final open-ended group was in fact originally fixed at age 55 because computer was slow and expensive but it has been shown that this limit produces bias when life expectancy improves (Brunborg, 1976 cited in McCaa and Barbi, 2004). Another important enhancement was the implementation of annual projections in place of 5- or 10-year periods (Biraben and Bonneuil, 1986; McCaa and Brignoli, 1989). Also, the IP method is not well suitable to capture age patterns of mortality for particular conditions of mortality such as epidemics, famine, civil war etc. Mortality crises were frequent in historical populations and they often affected selectively the population, thus determining substantial changes in the age profile of mortality. These changes, however, are not captured by IP since the underlying hypothesis in the IP algorithm is that mortality varies over time only by an age-independent factor, which is proportional to the difference between two adjacent age schedules of mortality. Thus IP will produce estimates of the probabilities of death disproportionately high at younger and older ages, leading hence to a distorted reconstruction. A solution was to set a mortality ceiling considered normal for the study period. The number of deaths below this threshold is apportioned according to the IP method while the exceeding part is apportioned proportionally to the size of the population of each age group (Biraben and Bonneuil, 1986; McCaa, 1993).

Many applications have highlighted the extreme adaptability of the procedure to the available documentation. Its robustness with respect to the choice of age patterns of vital events and the initial age structure, guarantees a good stability of the estimates and therefore it is a great advantage for those who, due to lack of information, are forced to make arbitrary choices. Many studies have shown that any wrong choice of the age structure and the mortality model have only a negligible impact on the estimates of the various demographic indicators (Lee, 1974, 1985; McCaa, 1989; McCaa and Vaupel, 1992; McCaa, 1993). The differences in the estimates obtained with two different initial age structures are quite strong in the early cycles of the projection, but diminish rapidly as the projection continues. Wachter has shown (1986) that the weak ergodicity theorem is extendible also to IP. In the long run, vital rates determine the population, not its initial state. The population hence quickly forgets its past age distribution. With regard to the mortality model, the effects of different hypotheses are more modest than those caused by different initial age structures, but appear to be more persistent over time. On the other hand, IP requires good vital registration data. As shown by Brunborg (1976), and McCaa and Vaupel (1992), quality of data is the most important factor for successfully applying the method (McCaa and Barbi, 2004).

Thanks to McCaa (1989) and McCaa and Brignoli (1989), who devised the user-friendly program 'populate', the IP could be readily applied by several researchers around the world.

Developments

Back projection was developed by Wrigley and Schofield for their massive reconstruction of the population of England, 1541–1871, as a means of dealing with the closure problem (Wrigley and Schofield, 1981; Oeppen, 1993b). Most populations, over the long run, are not closed, and England is no exception. While forward IP takes into account authentic migration data, Lee's method cannot produce migration flows (Lee, 1993). Back projection seeks to generate migration estimates from a terminal age structure by backcasting the population against the flow of births and deaths. While it might seem commonsensical to begin with the best, i.e., most recent, data and project backwards, Lee argues that this ignores the weak ergodicity theorem. In Lee's words (1993), "the problem originates with the attempt to resurrect people who have died into the oldest age group, an attempt that is, in practice, hypersensitive to error". The Cambridge team used iteration procedures to settle upon a single series of best estimates. However, beyond the Cambridge team, back projection has attracted few proselytes.

Generalized inverse projection (GIP) responds to Lee's criticisms of back projection and broadens the method into an analytical system which exploits whatever data are available as well as a broad range of assumptions or constraints, including components derived from back projection (Oeppen, 1993a,b). Lee's IP and back projection may be seen in fact as members of a same class of constrained projection models. In the worst situation of data scarcity – i.e., in the back projection frame, when only series of births and deaths and the final population age structure are known – generalized inverse

projectionists argue that the method can endogenously estimate migration flows as well as population totals and age structures by using a standard nonlinear optimization algorithm which maximizes the consistency between assumptions and data. The starting age structure can be also estimated using a stable population assumption. Since there are more parameters to be estimated (migration and mortality) than empirical observations (death totals and age groups of the final population), additional assumptions and/or data are required. Despite its flexibility, GIP has to solve an 'ill-posed' problem because of the uncertainty of the uniqueness and stability of the final solution. In Oeppen's words (1993b), "The selection of extra targets is somewhat arbitrary, which is in part a virtue since it allows a researcher to use whatever is available, but it also makes the estimates dependent on these extra targets". Lee (1993) remains unpersuaded, arguing that "…one can easily construct an infinity of radically different trajectories of population size, net migration, fertility and mortality which exactly match the given data".

Trend projection (Biraben and Bonneuil, 1986; Bonneuil, 1993) reconstructs past populations using series of births and deaths but in the absence of any census count or age structure. Given a birth and death series, the method finds a single value for which the trend in life expectancy after the initial date of the burial series is equal to the assumed value before. The method works well for closed populations but is ill-suited to open ones.

The flexibility and the potentiality of the IP method could not be ignored by Italian demographers given the rich availability of historical record of parish data of exceedingly high quality. In Italy, the first attempt to reconstruct a past population using the IP method is the work by Marco Breschi (1990) who tried to trace the demographic evolution of Tuscany from 1640 to 1940. The case of Tuscany, where demographic-statistical documentation is particularly detailed, allows Breschi to fully exploit the method and to develop a first and experimental version of two-sex projections. Results from Breschi's reconstruction demonstrated that the growth of the population of Tuscany starting from the middle of the eighteenth century was probably due, as in the case of England, to a rise in fertility rather than to a fall in mortality.

Since Breschi's work, new flavors of the method were developed, on the one hand, to exploit all available information and, on the other, to overcome some limits of Lee's IP.

The *differentiated IP* (Rosina, 1993, 1996; Rosina and Rossi, 1993) arises from the richness of data available in historical-demographic sources of some Italian regions where the age at death is faithfully reported. The procedure is based on the distribution of death by big age classes or, more simply, by only two age classes: infant and adult ages. In this case, the age-differentiated version of IP takes into account two life tables which, following Lee's procedure, have to fulfill two constraints: the first life table has to be such as the probabilities of death, applied to the young population, yield the total number of deaths recorded for infants and children and the second one has to be such as the probabilities of death, applied to the adult population, yield the total of deaths recorded at adult ages.

The aim of the age-differentiated procedure is to overcome possible shortcomings, already mentioned in Section Method, deriving from Lee's original assumption about the estimation of the age structure of the mortality function which, as said, varies over time only by a proportionality factor, the k_t variable, which does not depend on age but refers to the time-independent differences between two life tables from the same family. This assumption can lead, in exceptionally low or high mortality conditions, to probabilities of death greater than one or less than zero! This matter is particularly important in the past when age-selective mortality crises shifted not only the level but also the profile of mortality. It has been shown (Rosina and Rossi, 1994) that the distribution of deaths by as few as three age classes leads to an improvement of the fit to the observed data of about 80%. The differentiated IP has more explanatory power than the original IP as demonstrated by Rossi and Rosina in the study on the transition in the Venetian area (1998). However, since the age structure of the initial population has to be consistent with the information on deaths by age, the differentiated version of IP is, on the one hand, more rigorous, on the other, due to the larger number of constraints, less flexible with respect to Lee's procedure. This feature, especially in single-year projections, with the occurrence of a heavy mortality crisis, may lead to impossible probabilities of death (Barbi, 1997). The features of the method and an overview of various applications are described by Rosina (2004).

To make the profile of the mortality function more sensitive to possible changes over time and to avoid estimates of 'negative' population or 'births' at adult or old ages, Salvatore Bertino (1995) suggested some others modifications to Lee's method. Of particular interest is a model that estimates the mortality function at each time through a recursive procedure. The adjustment factor at each time t, k_t, is still determined through Lee's technique but the mortality function at each time t is estimated through a linear combination of the mortality functions estimated at time $t-1$ and $t-2$ with coefficients respectively equal to $1 + k_t$ and $-k_t$. Furthermore, the procedure imposes some convenient limits to the probabilities of death in the extreme ages (besides the obvious limits of 0 and 1) and then the adjustment factor is determined in an iterative way. Unfortunately, the robustness of Bertino's idea cannot be assessed in practice because, up to now, no software has been devised to implement his formulas.

An innovative more recent algorithm is the *stochastic inverse projection* (SIP) devised by Salvatore Bertino and Eugenio Sonnino in (1995) and appeared in an article published on Mathematical Population Studies in 2003. This procedure reconstructs past populations through a microsimulation approach. First, similar to Lee's procedure, a linear variation of the mortality function between two life tables relative to the beginning and the end of the studied period is hypothesized. Additionally, it is possible to enter two additional models for dealing with years where unusual conditions of mortality prevailed. Then, it is assumed that each individual in the population follows, over the entire life span, a nonhomogeneous Poisson process with intensity equal to the force of mortality obtained by approximation from step 1. Assuming that all individuals behave independently from each others, the

Poisson theory allows one to calculate the probability that a mortality event occurring at time t is to be attributed to an individual of age x. Generating a realization of the random variable 'age at death', according to the probability distribution just determined, it is possible to assign, by simulation, an age of death to the individual. The procedure is replicated for all recorded deaths so as to obtain the distribution of deaths by age, which is necessary to project the population. Similarly, fertility statistics are also determined by simulation. Although methodologically different, both forward and backward projections are possible with the stochastic method.

Instead of a single deterministic solution as with usual reconstruction methods, the stochastic approach offers a set of different reconstructions for the same time t. Then, the average scenario and the variability in a population's evolution are also evaluated. The results of all these alternative simulations differ by chance only and are equally coherent because they are governed by the same data and models. A comparison between the behavior of the IP technique, the differentiated version, and the stochastic method in various simulated scenarios of exceptional mortality (Barbi, 1997) shows that the stochastic approach proceeds in a way, which closely reflects empirically observed patterns. The estimated probabilities of death are based not only on theoretical risks (the intensity of the Poisson process) but also on the proportion of the population at each age. This is a key point, especially in the reconstruction of small populations. The stochastic approach recognizes the eventuality that very few mortality events or even none may be attributed in some old age classes where individuals are few although their theoretical mortality risks are high.

Barbi and Oeppen (2004) compared the results from the GIP and the SIP in the back projection frame (when the terminal age structure of the population is specified) and showed that both the techniques reconstruct demographic scenarios coherently with recorded data. However, Lee (2004), while appreciates the stochastic approach because it provides an estimate of the uncertainty in the IP results, and Oeppen's generalized version because of its flexibility to incorporate demographic information that are available, does not see any advantage to running SIP backward and continues to be critic with respect to GIP when running in the back projection frame in the absence of input data on migration.

IP and its variations have produced important innovations in the field of population projection. Their strategy is to exploit specific existing documentation so as to lead to coherent demographic reconstructions. Desirable developments should look at an even more flexible approach that takes benefit of any bit of demographic information, a general model able "to cover situations in which one did not have full time series of births and deaths, and perhaps had more information of other kinds, such as censuses" (Lee, 2004). Toward this direction is the approach devised recently by Bonneuil and Fursa (2011, 2012). Combining the Lotka–McKendrick population model with stochastic optimization, the authors reconstruct the demographic dynamics, even by civil status, from partially missing or flawed demographic data. (Large part of this Section is reprinted from McCaa and Barbi, 2004, © Springer-Verlag Berlin-Heidelberg 2004, with kind permission of Springer Science + Business Media.)

See also: Demographic Techniques: Family Reconstitution; Population Dynamics: Mathematic Models of Population, Development, and Natural Resources; Population Forecasts.

Bibliography

Barbi, E., 1997. Tecniche di proiezione inversa nello studio della mortalità infantile e giovanile. In: Proceedings of "Il Convegno dei Giovani Studiosi dei Problemi di Popolazione." Roma June 25–27, 1996, pp. 193–210.

Barbi, E., Oeppen, J., 2004. Comparing the results from generalized inverse projection and stochastic inverse projection. In: Barbi, E., Bertino, S., Sonnino, E. (Eds.), Inverse Projection Techniques. Old and New Approaches, Demographic Research Monographs. Springer, Heidelberg, pp. 95–106.

Bertino, S., 1995. Osservazioni sui modeli di mortalità e di fecondità per la proiezione inversa. In: Serie A-Ricerche, vol. 4. Dipartimento di Statistica, Probabilità e Statistiche Applicate, University of Rome, La Sapienza.

Bertino, S., Sonnino, E., 1995. La proiezione inversa stocastica: tecnica e applicazione. In: Le Italie Demografiche. Saggi di demografia storica. Dipartimento di Scienze Statistiche, University of Udine, Udine.

Bertino, S., Sonnino, E., 2003. The stochastic inverse projection and the population of Velletri (1590–1870). Mathematical Population Studies 10 (1), 41–73.

Biraben, J.N., Bonneuil, N., 1986. Population et société en Pays de Caux au XVIIe siècle. Population 6, 937–960.

Bonneuil, N., 1993. The trend method applied to English data. In: Reher, David S., Schofield, Roger (Eds.), Old and New Methods in Historical Demography. Clarendon Press, Oxford, pp. 57–65.

Bonneuil, N., Fursa, E., 2011. Optimal population path fitting for flawed vital statistics and censuses. Journal of Optimization Theory and Applications 148, 301–317.

Bonneuil, N., Fursa, E., 2012. Optimal marriage fitting for imperfect statistics. Journal of Optimization Theory and Applications 153, 532–545.

Breschi, M., 1990. La popolazione della Toscana dal 1640 al 1940. Un'ipotesi di ricostruzione. Dipartimento statistico, University di Florence, Florence.

Brunborg H., 1976. The inverse projection method applied to Norway, 1735–1974, unpublished manuscript.

Lee, R.D., 1974. Estimating series of vital rates and age structures from baptisms and burials: a new technique with applications to pre-industrial England. Population Studies 28, 495–512.

Lee, R.D., 1985. Inverse projection and back projection: a critical appraisal and comparative results for England, 1539–1871. Population Studies 39, 233–248.

Lee, R.D., 1993. Inverse projection and demographic fluctuations: a critical assessment of new methods. In: Reher, David S., Schofield, Roger (Eds.), Old and New Methods in Historical Demography. Clarendon Press, Oxford, pp. 7–28.

Lee, R.D., 2004. Reflections on inverse projections: its origins, development, extensions, and relation to forecasting. In: Barbi, E., Bertino, S., Sonnino, E. (Eds.), Inverse Projection Techniques. Old and New Approaches, Demographic Research Monographs. Springer, Heidelberg, pp. 1–10,.

McCaa, R., 1989. Populate: a microcomputer projection package for aggregative data applied to Norway, 1736–1970. Annales de Démographie Historique, 287–298.

McCaa, R., 1993. Benchmarks for a new inverse population projection program: England, Sweden, and a standard demographic transition. In: Reher, David S., Schofield, Roger (Eds.), Old and New Methods in Historical Demography. Clarendon Press, Oxford, pp. 40–56.

McCaa, R., Brignoli, H.P., 1989. Populate: from Births and Deaths to the Demography of the Past, Present, and Future. University of Minnesota, Minneapolis.

McCaa, R., Vaupel, J.W., 1992. Comment la projection inverse se comporte-t-elle sur des données simulées? In: Blum, Alain, Bonneuil, Noël, Blanchet, Didier (Eds.), Modèles de la démographie historique. Institut National d'Etudes Démographiques, Paris, pp. 129–146.

McCaa, R., Barbi, E., 2004. Inverse projection: fine-tuning and expanding the method. In: Barbi, E., Bertino, S., Sonnino, E. (Eds.), Inverse Projection Techniques. Old and New Approaches, Demographic Research Monographs. Springer, Heidelberg, pp. 11–28.

Oeppen, J., 1993a. Back projection and inverse projection: members of a wider class of constrained projection models. Population Studies 47 (2), 245–267.

Oeppen, J., 1993b. Generalized inverse projection. In: Reher, David S., Schofield, Roger (Eds.), Old and New Methods in Historical Demography. Clarendon Press, Oxford, pp. 29–39.

Rosina, A., 1993. Una generalizzazione dell'inverse projection. In Per una storia di metodo della popolazione italiana: problemi di metodo. Dipartimento di Scienze Statistiche "Paolo Fortunati", University of Bologna, Bologna, pp. 73–80.

Rosina, A., 1996. IPD 3.0: Applicazione automatica dell'inverse projection differenziata (passo annuale e quinquennale). Technical Report. Dipartimento di Scienze Statistiche, University of Padova.

Rosina, A., 2004. Using information on the age distribution of deaths in population reconstruction: an extension of inverse projection with applications. In: Barbi, E., Bertino, S., Sonnino, E. (Eds.), Inverse Projection Techniques. Old and New Approaches, Demographic Research Monographs. Springer, Heidelberg, pp. 29–38.

Rosina, A., Rossi, F., 1993. Una estensione dell'Inverse Projection con mortalita' differenziata per eta'. Statistica 4, 619–631.

Rosina, A., Rossi, F., 1994. Ricostruzioni aggregate dei processi evolutivi delle popolazioni. Cleup, Padova.

Rossi, F., Rosina, A., 1998. Il Veneto tra Sette e Ottocento. Bollettino di Demografia Storica 28, 89–114.

Wachter, K.W., 1986. Ergodicity and inverse projection. Population Studies 40, 275–287.

Wrigley, E.A., Schofield, R.S., 1981. The Population History of England, 1541–1871: A Reconstruction. Harvard University Press, Cambridge, MA.

Demographic Techniques: LEXIS Diagram

Gérard Calot[†], Observatoire Démographique, Européen, France

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 5, pp. 3472–3480, © 2001, Elsevier Ltd.

Abstract

Demographic rates may be computed for a calendar period or for a birth cohort, by age attained at the time of occurrence, or at the end of the calendar year of occurrence. Demographers keep track of this plethora of rates by means of a simple graphical device that has become known as the Lexis diagram. The purpose of this article is to explain this diagram and show how it can be of help in describing usage of rates in each particular analysis. To enable analysts to improve comparability, it also provides conversion formulas that can be used to compute selected types of rates from the ingredients originally intended for the computation of other rates. Methods to eliminate biases produced when standard procedures are applied to unsteady flows of births caused by major political events like wars are discussed.

The quintessential building block of demographic analysis is a set of occurrence/exposure rates computed by dividing the number of occurrences of a selected demographic event in a (sub) population into the corresponding number of person-years under risk. The event may be repeatable, which means that it may occur more than once in the life course of an individual, such as geographical moves and the birth of children. Alternatively the event occurs only once in a lifetime, such as a first birth or a first marriage, or at the end of life, namely death. For repeatable events, the occurrences recorded for each age group are divided by the total number of person-years observed in the same age group, and one gets quantities like births per 1000 woman-years at ages 30–34. For non-repeatable events, the occurrences in the age group are divided into the number of person-years observed among population members who have not yet experienced the index event, and one gets quantities like first births per 1000 person-years among childless women. Demographic rates may be computed for a calendar period or for a birth cohort, by age attained at the time of occurrence or at the end of the calendar year of occurrence. Demographers keep track of this plethora of rates by means of a simple graphical device that has become known as the LEXIS diagram. The purpose of the present Encyclopedia article is to explain this diagram and show how it can be of help in describing usage of rates in each particular analysis.

Demographic rates are computed according to different conventions in different countries, and some countries have changed their practice over time. This makes it difficult to compare series of rates across countries and sometimes over time. To enable analysts to improve comparability, we will also provide conversion formulas that can be used to compute selected types of rates from the ingredients originally intended for the computation of other rates. We also provide methods to eliminate biases produced when standard procedures are applied to unsteady flows of births caused by major political events like wars. Such biases only affect certain birth cohorts, but in the disturbed cohorts all ages are involved.

[†] Deceased.

The LEXIS Diagram

The LEXIS diagram provides a visual representation of occurrences and exposures across age and time. Since age and time increase in lockstep, a basic ingredient in the diagram is the lifeline of a person, showing age plotted against time on a 45-degree line starting at the time t_0 of birth and at age zero. If an event occurs to the person at a given time t, the age $t-t_0$ of that person at the time of occurrence is the ordinate of abscissa t on the lifeline (Figure 1).

The LEXIS diagram pertains to the residents of a given territory. The start of a lifeline may be the birth on the territory. In that case, the lifeline begins at the horizontal axis. When a person immigrates at age x at time t, the lifeline starts at the point (t, x). The end of the lifeline corresponds to the death or out-migration by the person.

Major vital events (births, deaths, marriages, divorces) are typically recorded through a vital registration system, which provides breakdowns of events by the age of the persons concerned. Residents are periodically enumerated through

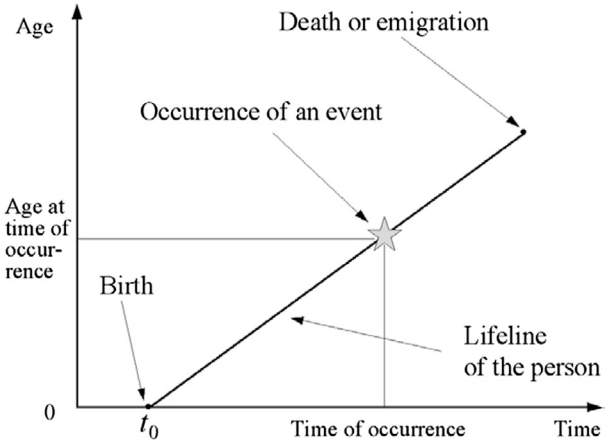

Figure 1 Lifeline of a person born at time t_0.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31014-5

censuses, which record the date of birth (or the completed age on the census day) of the respondents. At least at the national level, censuses are sometimes updated every year on a fixed date, most often January 1, so that the distribution of residents by single year of (completed) age is usually available annually. The number of persons whose completed age at time t is i is the number of lifelines that cross the vertical segment $(t, [i, i+1])$. These persons were born between times $t-i-1$ and $t-i$.

Two events that occur simultaneously are located on the same vertical line in the LEXIS diagram. Two non-simultaneous events that occur to persons having the same age at the time of occurrence are located on a horizontal line. Two non-simultaneous events that occur to persons born at the same time are located on a 45-degree line.

The events that occur during a given period (e.g., a given calendar year) are located in a vertical strip, between the vertical lines corresponding to the beginning and the end of the period (Figure 2). Those that occur to persons whose age at the time of occurrence belongs to the same age group are located in a horizontal strip. Those that occur to persons born during the

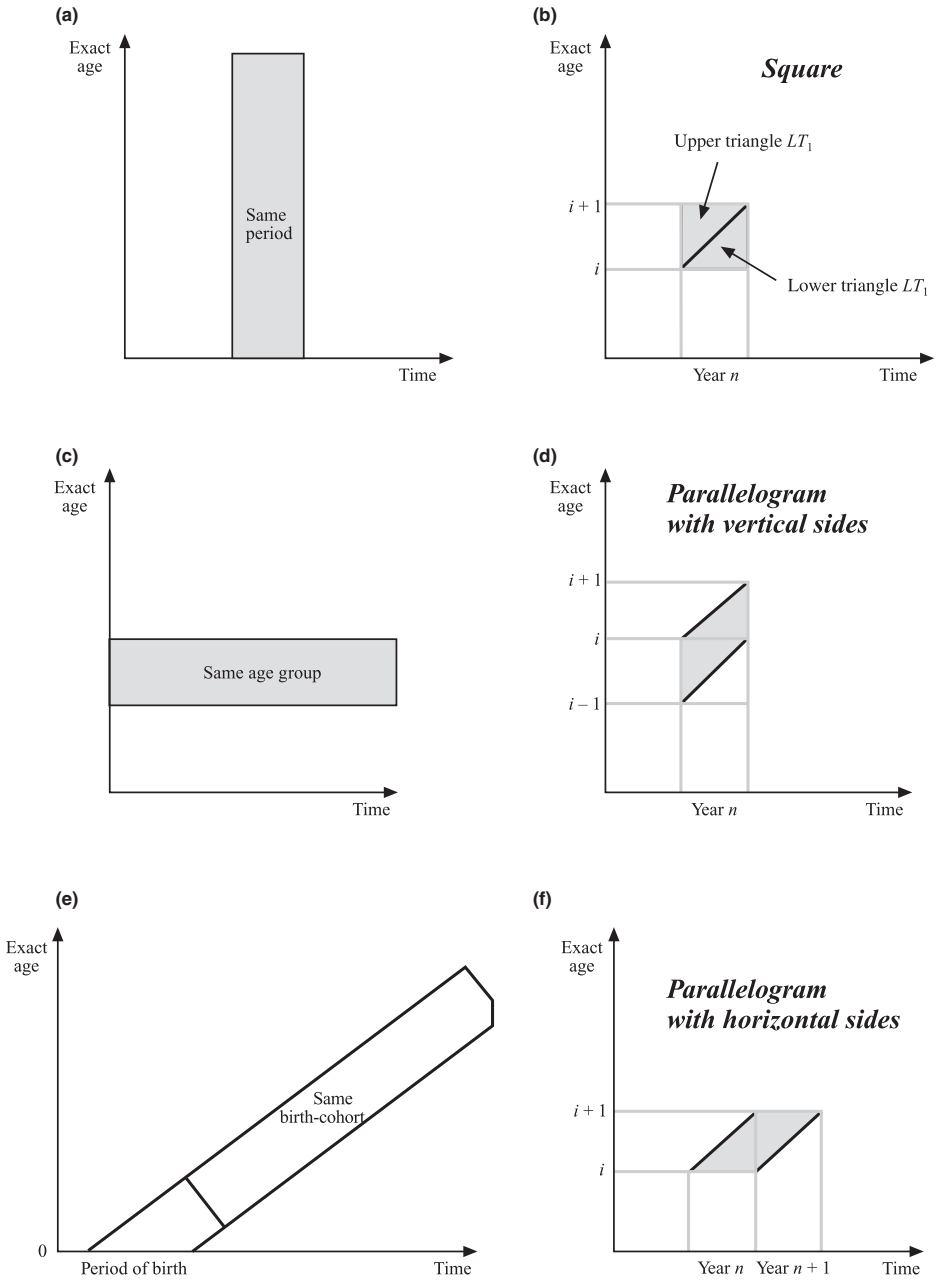

Figure 2 Regions in the LEXIS diagram. (a) Events occurring during the same period to persons of various ages. (b) Events occurring during the same year to persons who had the same age in completed years at the time of occurrence (members of two cohorts). (c) Events occurring to persons whose age was in the same age group at the time of occurrence. (d) Events occurring during the same year to persons who attained the same birthday during that year (members of the same cohort). (e) Events occurring to persons who belong to the same cohort. (f) Events occurring during two consecutive years to members of the same cohort who had the same age in completed years at the time of occurrence.

same period (i.e., who belong to the same birth cohort), are located in a 45-degree strip.

Age

The exact age of a person on a given date t is a continuous variable. It is equal to the exact duration from the date of birth to date t. In most connections and in common language, exact age is grouped to take only integer values, however, and this makes age a discrete variable. We then refer to it as age at last birthday, or age in completed years, or completed age. The distribution of residents by age on a given date is a breakdown of the population by completed age. If the date is January 1 of year n, persons then aged i in completed years are the residents born during the year $n-i-1$.

Tables classifying the events that occurred to residents during year n by age, may be presented either by completed age at the time of occurrence or, based on the year of birth, by the age that has been or will be attained during year n. The events of year n that correspond to completed age i (as measured at time of occurrence) are located in a square of the LEXIS diagram. Those that correspond to the same year of birth i.e., to the same age i attained during the course of year n or, in other words, to the same completed age i on December 31 of year n, are located in a parallelogram with vertical sides. A third possibility that combines the two consists in retaining completed age but distinguishing persons according to whether the birthday they celebrated in year n fell before or after the occurrence of the event. When the latter distinction is made, the numbers of events refer to triangles in the LEXIS diagram, the lower triangle of the square concerns persons whose birthday in year n was before the occurrence of the event, the upper triangle those for whom it was after. Note that persons for whom an event was recorded during year n at completed age i were half a year older than those who had it recorded at attained age i on 'average' at the time of occurrence. The former had an 'average' exact age equal to $i+\frac{1}{2}$, the latter an 'average' exact age equal to i. (What we mean by 'average' here will be made clear later.) This half a year difference may be of some importance. For example, teenage fertility, defined as fertility below age 20 years, may differ by as much as 25% if age is defined in terms of completed age instead of age attained during the year.

Rates of the First and Second Kind

Suppose that one's interest focuses on a given demographic (or other) event. For each region in the LEXIS diagram where the data are available, the analyst may record the number of occurrences of the event in question. For repeatable events, one may also add up the lengths of all lifelines that intersect the given LEXIS region (or their projections on the time axis) to get the total person-years of exposure to the risk of occurrence of the event in question. The ratio between the number of occurrences and the latter exposures is the corresponding occurrence/exposure rate.

For non-repeatable events one can compute two kinds of rates. The rate of the first kind (or occurrence/exposure rate) is computed using lifelines interrupted at the time of event occurrence. Their denominator is again the 'total person-years exposed to risk.' A corresponding rate of the second kind (or occurrence/population rate) uses the same occurrences, but the denominator is computed in the same manner as for repeatable events i.e., lifelines are not interrupted at the time of event occurrence. If the non-repeatable event always terminates a lifeline, the rates of the first and second kind coincide and there is a single occurrence/exposure rate for the given region. This is of course the case if the event is death or out-migration.

Rates of the second kind are sometimes computed because the aggregate data do not provide the person-years exposed to risk but only the total number of person-years for the LEXIS region selected i.e., the person-years at risk plus the person-years for members of the population who are no longer at risk. A case in point is when first marriages observed in the relevant population segment are referred to the total resident population in the same segment, instead of only to the person-years for the never-married, perhaps because the latter exposures are not available by sex and single years of age on an annual basis. Another example arises when births of a given order are divided by the total number of person-years lived in the same population segment instead of only by the corresponding number of person-years for persons of the previous childbearing parity. This article is concerned with rates of the first kind. Rates of the second kind are discussed in a companion article, as is the connection between the two kinds of rates.

To avoid using many decimals, the rates are normally multiplied by a convenient factor like 1000, 10 000, or 100 000 before tabulation. If the factor is 100 000, such a rate is said to be computed 'per 100 000 person-years,' where for rates of the first kind the latter are years of exposure to risk.

A Mathematical Framework of the Study of Occurrence/Exposure Rates

Repeatable Events

Let us consider a person exposed between ages α and ω to the occurrence of a *repeatable* event E, and let us reason as if the person is not exposed to any other event (in particular, no migration and no death). Let $\lambda(x)\,\mathrm{d}x$ be the probability that E occurs to this person between exact ages x and $x+\mathrm{d}x$. Outside the interval (α, ω), $\lambda(x)$ is zero. Then function λ is called the risk function of the repeatable event E.

By the properties of a Poisson process, the expected total number of occurrences of E to this person is $\int_\alpha^\omega \lambda(x)\mathrm{d}x$. If we let $f(x) = \lambda(x)/\int_\alpha^\omega \lambda(u)\mathrm{d}u$, we can regard this as the density of a probability distribution with mean $m = \int_\alpha^\omega xf(x)\mathrm{d}x$ and variance $V = \int_\alpha^\omega (x-m)^2 f(x)\mathrm{d}x$. If the field of study is fertility and the event E is the birth of a child to a woman, the integral $\int_\alpha^\omega \lambda(x)\mathrm{d}x$ is called the cohort-completed fertility and is the average number of children per woman. The mean age at occurrence is called the mean age at childbirth.

Non-Repeatable Events

If the event E can happen only once in a lifetime, then $\lambda(x)\,\mathrm{d}x$ is the probability that E occurs to a given person between exact

ages x and $x + \mathrm{d}x$, provided it has not occurred by exact age x. According to standard single-decrement life-table theory (*see* Life Table), the probability that the event E does not occur to a given person before age $x > a$ is $\ell(x) = \exp\left(-\int_\alpha^x \lambda(u)\mathrm{d}u\right)$. The probability that event E occurs between ages x and $x+t$, given that it has not occurred before age x, is

$$ {}_tq_x = 1 - \exp\left(-\int_x^{x+t} \lambda(u)\mathrm{d}u\right) $$

The probability of ever experiencing E is $1 - \ell(\omega) = 1 - \exp\left(-\int_\alpha^\omega \lambda(x)\mathrm{d}x\right)$. Note that this probability is less than 1 if and only if $\int_\alpha^\omega \lambda(x)\ \mathrm{d}x < \infty$.

The density of the age at which E occurs, given that it ever does, is $\lambda(x)\ell(x)/[1 - \ell(\omega)]$. The mean age at occurrence is

$$ m = \alpha + \int_\alpha^\omega \frac{\ell(x) - \ell(\omega)}{1 - \ell(\omega)}\mathrm{d}x $$

The higher moments of this distribution are defined similarly. If the event E is death, considered from birth (which means that $\alpha = 0$ and ω is the maximum human life span) i.e., if the field of study is overall mortality, then the mean age at occurrence is called the life expectancy at birth. The quantity we have denoted $\lambda(x)$ is called the force of mortality at age x and $\ell(x)$ is the survival function at age x. If the event E is death considered only after a given age $a > \alpha$, the difference between the mean age at occurrence and a is called the life expectancy at age a; it is the average number of years remaining to a person still alive at age a.

If event E is first marriage below completed age 50, $1 - \ell(50)$ is the proportion ever married by age 50 and $\ell(50)$ is called the final celibacy. The mean age at occurrence is the mean age at first marriage (below 50).

If the event E is the birth of a child of order k ($k \geq 2$) to a woman and time is recorded not from the birth of the woman but from the date of birth of her preceding child, $1 - \ell(\omega)$ is the proportion who reach parity k among women reaching parity $k - 1$. This is also called the parity progression ratio from order $k - 1$ to order k. For $k = 1$, time is recorded from the birth of the woman or from any other suitable origin (e.g., the moment of marriage if the field of study is marital fertility) and $1 - \ell(\omega)$ is the proportion of women who have at least one child. The frequency of childlessness is $\ell(\omega)$.

Annual Record of Events for Both Time of Occurrence and Age at Occurrence

Rates

Tables classifying events that have occurred during one calendar year, say year n, by single year of age of the person concerned, may be established by squares or by vertical-sided parallelograms in the LEXIS diagram. We let the current value of age completed be the integer i. In both cases, we have to estimate the probability of occurrence of event E within a doubly annual figure. In the mathematical framework presented above, we face three difficulties. *First*, we do not have a single person to consider but a discrete cohort born during calendar year $n-i$ in the case of a vertical-sided parallelogram, or two consecutive discrete cohorts (born in years $n-i-1$ and $n-i$) in the case of a square. In the parallelogram, each member of the cohort is exposed to risk during exactly 1 year (unless the person dies or migrates), but in the square the members of the two cohorts are exposed to risk during periods of variable duration ranging from 0 (for those born close to January 1 of year $n-i-1$ or close to December 31 of year $n-i$) to 1 year (for those born close to January 1 of year $n-i$). Furthermore, the risk function $\lambda(x)$ may now well vary with the time of birth g, in which case we should rewrite the function as $\lambda(x, g)$.

Second, the statistical data available may not be the full individual-level information, which would consist of the date of birth of each person and the various dates of occurrence of event E if it happens to occur to that person. We typically just know the *total number* of events inside the parallelogram or the square and the corresponding exposures.

Third, since our mathematical theory was developed for a single event E, migration and death (or only migration if event E is itself death) complicate the situation. The total number of events recorded includes those occurring to in-migrants or to persons who die or out-migrate subsequently to an occurrence inside the region in the LEXIS diagram; migration and deaths are duly taken into account in the total number of residents classified by age at both year-ends. But in aggregate data the exact date of death, in-migration or out-migration of each person exposed to risk in the territory considered is not known. For these reasons, we will now reformulate our mathematical framework. We do it in two steps: first a simple one and second by a more sophisticated one.

First Simple Assumption: Constant Risk Function

Our simple reformulation holds for any kind of event, whether repeatable or not, and is as follows: Let us assume that the risk function λ is constant inside the age bracket concerned (exact ages i to $i+1$ in the case of a square, $i-1$ to $i+1$ in the case of a vertical-sided parallelogram). The probability that event E occurs is $\lambda\ \mathrm{d}g\ \mathrm{d}x$ for a resident who is born between times g and $g+\mathrm{d}g$ and whose exact age is between x and $x+\mathrm{d}x$, for every g and x inside the LEXIS-diagram region. Then the occurrence/exposure rate (of the first kind) is the maximum likelihood estimate of λ in the case of squares as well as of parallelograms. Let us furthermore assume that the number of entries due to in-migration, and exits due to death and out-migration, lead to a uniform balance: the net number of entries of persons born between g and $g+\mathrm{d}g$ when their exact age is between x and $x+\mathrm{d}x$ is $s\ \mathrm{d}g\ \mathrm{d}x$, where s is a constant.

As we already stated, net 'in-migrants' are supposed to have the same risk λ as the non-migrants. The quantity s can be approximated by $s_i^n = P_i^{n+1} - P_{i-1}^n$ inside a parallelogram with vertical sides on the basis of the numbers of residents at year-ends by single year of age. Here p_i^n is the number of residents of completed age i on January 1 of year n. (If event E is death, s_i^n is equal to $P_i^{n+1} - P_{i-1}^n - D_i^n$, where D_i^n is the number of deaths occurring during year n of residents born in year $n-i$.) If events are observed in squares, then s_{i+1}^n is an approximation to s_U in the upper triangle of the square, while s_i^n is an approximation to s_L in the lower triangle.

Vertical-Sided Parallelograms

When events are observed in vertical-sided parallelograms, the P_{i-1}^n residents on January 1 of year n, born in year $n-i-1$, who attain their ith birthday during year n, are all at risk during one year each, and the s net 'in-migrants' are at risk during half a year on average, so that the sum of exposures to risk is

$$R_i^n \approx P_{i-1}^n + \frac{1}{2}s = \frac{1}{2}\left(P_{i-1}^n + P_i^{n+1}\right)$$

The estimate $\widehat{\lambda}$ of λ is the observed occurrence/exposure rate $\widehat{\lambda} = VSP_i^n/R_i^n$, where VSP_i^n is indicated in Figure 3(a).

Squares

When events are observed in squares, we estimate the exposures to risk in each triangle of the square. We consider the upper and the lower triangle, one at a time. In the upper triangle, there are p_i^n persons in the population on January 1, or rather $P_i^n + \frac{1}{2}s_U$ persons who attain their $(i+1)$th birthday during year n and who are exposed between 0 and 1 year. If we introduce $m(G)$ as the average location inside year n of the birthdays attained during that year by persons born in year G, their average exposure to risk is $m(n-i-1)$. Total exposures equal $(p_i^n + \frac{1}{2}s_U)m(n-i-1)$. However, the $s_U/2$ net entries have an average duration at risk of $\frac{1}{3}$ year. The corresponding exposures must be deducted since they are unduly included in the previous total. Thus the total exposure in the upper triangle is

$$\left(P_i^n + \frac{s_U}{2}\right)m(n-i-1) - \frac{s_U}{6} \approx P_i^n m(n-i-1) + \frac{s_U}{12}$$

Since $m(G)$ is anyhow close to $\frac{1}{2}$ and s is small compared to p_i^n. In the lower triangle, $P_i^{n+1} - \frac{1}{2}s_L$ persons attain their ith birthday during year n. They are exposed during an average period of $1-m(n-i)$, and there are $s_L/2$ net entries with an average duration of $\frac{1}{3}$ year, leading to total exposures of

$$\left(P_i^{n+1} - \frac{s_L}{2}\right)[1 - m(n-i)] + \frac{s_L}{6} \approx P_i^{n+1}[1 - m(n-i)] - \frac{s_L}{12}$$

Adding up the exposures in the triangles, we get total exposures in the square of

$$P_i^n m(n-i-1) + P_i^{n+1}[1 - m(n-i)] + \frac{s_U - s_L}{12}.$$

If the two consecutive cohorts born in years $n-i-1$ and $n-i$ are similar in size, if the mean locations of birthdays in each year for these cohorts are both close to $\frac{1}{2}$, and if migration balances are similar in the two triangles, then the exposures are close to $(P_i^n + P_i^{n+1})/2$ and the constant risk function λ is estimated by

$$\widehat{\lambda} = \frac{SQ_i^n}{\left(P_i^n + P_i^{n+1}\right)/2}$$

where the occurrences SQ_i^n are indicated in Figure 3(b).

On the other hand, if cohorts born years $n-i-1$ and $n-i$ are quite dissimilar in size or if the mean locations of birthdays are quite different in these two cohorts, the latter formula may prove inaccurate. This is the case when a major disruption in the birth rate such as that provoked by the outbreak of a war occurs during the course of either of the 2 years. The consequence is then a multiplicative bias of approximately

$$\frac{B(n-i)\left[\frac{1}{2} - m(n-i)\right] - B(n-i-1)\left[\frac{1}{2} - m(n-i-1)\right]}{[B(n-i) + B(n-i-1)]/2}$$

in the above estimate of λ. Here $B(G)$ denotes the absolute number of births recorded in year G. The bias is approximately a constant percent of λ. For example, rates calculated for Belgian and French cohorts born in 1915 and 1916 are systematically biased upwards by almost 15%. This bias has almost the same magnitude for any kind of event (fertility, nuptiality, or mortality) and for any age. Rumanian cohorts born in 1966 and 1967 show a bias of a similar size caused by the huge increase in births subsequent to a change in the legislation on abortion in October 1966.

To avoid such biases, one should observe demographic events in vertical-sided parallelograms rather than in squares when the age-pyramid shows irregularities due to major historical events. Contrary to common language, 'statistical' age at the time of occurrence is preferably age attained during the calendar year rather than completed age at the occurrence of the event.

Note that if E is a non-repeatable event, the probability of occurrence of E during a 1-year period, given that it has not occurred before the beginning of the period, is $1-e^{-\lambda}$ under the assumption of a constant risk function equal to λ, and if the

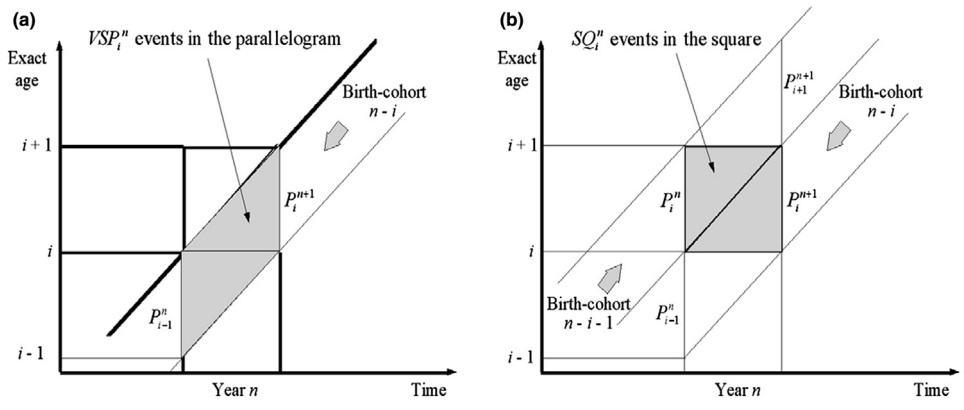

Figure 3 (a) Events observed in a vertical-sided parallelogram. (b) Events observed in a square.

observed rate r is small, the probability is estimated by $1-e^{-r} \approx 1/(1+\frac{1}{2}r)$.

The value of $m(G)$ hardly changes over years n for a given cohort G; mortality and migration do not alter substantially the characteristics of the distribution of birthdays of the cohort through time, even when these characteristics are exceptional. These characteristics can then be approximated by their value at the time of birth i.e., on the basis of the monthly distribution of births that occurred during year G. They can also be estimated from census data i.e., on the basis of a breakdown of the resident population by year and month of birth.

More Sophisticated Assumption: Variable Risk Function

In our second step we now turn to more refined approximations. We still assume that the risk function is independent of the time g at birth but we now allow (soft) variations with age x. The risk of occurrence of the event E at an exact age comprised between x and $x+\mathrm{d}x$ among the $N(g)\mathrm{d}g$ residents born between g and $g+\mathrm{d}g$ is $\lambda(x)\mathrm{d}x\mathrm{d}g$. The risk function λ will be assumed to be a quadratic polynomial inside several adjacent regions of the LEXIS diagram in the case of a repeatable event, and a linear polynomial in the case of a non-repeatable event. In the latter case, especially if E is death, we may alternatively assume that it is $\log[\lambda(x)]$ that varies linearly over an interval of a few years of age. This corresponds to the Gompertz law of mortality, which is generally valid at higher ages (say at ages over 30 or 40) below the oldest old. The density s of net entries due to migration (or to deaths if E is not death itself) is still assumed to be constant inside each vertical-sided parallelogram in the LEXIS diagram.

For the following reason we need to estimate the integral of $\lambda(x)$ over a 1-year interval, say between a and $a+1$, where a is either i or $i-\frac{1}{2}$. For repeatable events the mean number of occurrences per person between ages a and $a+1$ is $\int_a^{a+1}\lambda(x)\mathrm{d}x$. For non-repeatable events the probability of recording E between ages a and $a+1$ among those who have not recorded it before age a is $1-\exp(-\int_a^{a+1}\lambda(x)\mathrm{d}x)$. In both cases the integral plays a central role.

If the derivatives λ' and λ'' are not too large (i.e., if λ indeed has 'soft' variations), the integral can be estimated on the basis of available estimates of λ, λ', and λ'' as follows. We get

$$\int_i^{i+1}\lambda(x)\mathrm{d}x \approx \lambda\left(i+\frac{1}{2}\right)+\frac{\lambda''(i+\frac{1}{2})}{24} \approx \lambda(i)+\frac{\lambda''(i)}{2}+\frac{\lambda''(i)}{6},$$

and

$$\int_{i-1/2}^{i+1/2}\lambda(x)\mathrm{d}x \approx \lambda(i)+\frac{\lambda''(i)}{24} \approx \lambda\left(i+\frac{1}{2}\right)-\frac{\lambda'(i+\frac{1}{2})}{2}+\frac{\lambda''(i+\frac{1}{2})}{6}$$

Vertical-Sided Parallelograms

Under the above assumptions, the number of events occurring in the vertical-sided parallelogram for a repeatable event is

$$VSP_i^n = N\left[\lambda+\lambda'\left(m-\frac{1}{2}\right)+\frac{\lambda'}{2}V\left(\frac{1}{3}-m+m^2+V\right)+s\frac{\lambda'}{6}\right]$$

and for a non-repeatable event it is

$$VSP_i^n = N\left[1-e^{-\lambda-\frac{\lambda'}{2}(1-2m)}\right]e^{\lambda m}-\frac{\lambda'}{2}\left(m^2+V\right)+\frac{\lambda^2}{2}V + \frac{s}{2}\left[2-e^{\frac{\lambda}{3}-\frac{\lambda'}{6}+\frac{\lambda^2}{36}}-e^{\frac{\lambda}{3}-\frac{\lambda'}{6}+\frac{\lambda^2}{36}}\right]$$

In these formulae, N is the number of persons who reach their ith birthday during year n. (This is estimated by the mid-value of P_{i-1}^n and P_i^{n+1} in the case of a repeatable event, and by the mid-value of P_{i-1}^n and P_i^{n+1} plus half the difference between the number of occurrences of E in the upper triangle and in the lower triangle of the parallelogram, in the case of a non-repeatable event.) Furthermore, λ, λ', and λ'' are the values of the risk function and of its two first derivatives at exact age i, s is the density of net entries inside the parallelogram, and m and V are the mean and variance of the locations of birthdays in the cohort born in year $n-i$. (If the distribution of birthdays in the cohort is *uniform*, m is equal to $\frac{1}{2}$, V to $\frac{1}{12}$, and the third moment μ is zero. When the actual distribution of birthdays differs from uniformity only due to usual seasonal variations in the number of births, there is no need to take care of non-uniformity because m is then close to $\frac{1}{2}$, and V to $\frac{1}{12}$.)

If we know N, m, and V, and if we estimate λ' and λ'' on the basis of the values of λ at neighboring ages, the appropriate formula above yields an estimate for λ. This is done iteratively and the iteration process is stopped when λ has stabilized for all i. In the first iteration, λ' and λ'' are set equal to zero. In subsequent iterations, $\lambda'(i)$ may be estimated by $\frac{1}{2}[\lambda(i+1)-(i-1)]$ and $\lambda'(i)$ by $\lambda(i+1)+\lambda(i-1)-2\lambda(i)$, where λ values are those obtained in the previous iteration at adjacent ages $i-1$, i, and $i+1$.

Squares

The corresponding equations in the case of squares are for a repeatable event

$$\begin{aligned} SQ_i^n = {} & N_1\left[\lambda m_1+\frac{\lambda'}{2}\left(m_1-m_1^2+V_1\right)\right] \\ & +\frac{\lambda''}{2}\left(\frac{m_1}{4}+\frac{m_1^2+V_1}{2}+\frac{m_1^2}{3}+m_1V_1+\frac{\mu_1}{3}\right) \\ & +N_2\left\{\lambda(1-m_2)-\frac{\lambda'}{2}\left[m_2-m_2^2-V_2\right]\right. \\ & \left.+\frac{\lambda''}{2}\left[\frac{1}{12}-\frac{m_2}{4}+\frac{m_2+V_2}{2}-\frac{m_2^3}{3}-m_2V_2-\frac{\mu_2}{3}\right]\right\} \\ & +\frac{s_2-s_1}{6}\left(\lambda+\frac{\lambda''}{40}\right) \end{aligned}$$

and for a non-repeatable event it is

$$\begin{aligned} SQ_i^n = {} & N_1\left[e^{\lambda m_1+\frac{\lambda'}{2}[m_1(1-m_1)-V_1]+\frac{V_1}{2}\lambda^2}-1\right] \\ & +N_2\left[1-e^{-\lambda(1-m_2)+\frac{\lambda'}{2}[m_2(1-m_2)-V_2]+\frac{\lambda^2}{2}V_2}\right] \\ & +\frac{s_1}{2}\left[1-e^{\frac{\lambda}{3}+\frac{\lambda^2}{36}}\right]+\frac{s_2}{2}\left[1-e^{-\frac{\lambda}{3}+\frac{\lambda^2}{36}}\right] \end{aligned}$$

In the two latter formulae, $_1$ refers to the upper triangle (or to the cohort born in year $n-i-1$), $_2$ to the lower triangle (or to the cohort born during year $n-1$), N_1 is the number of persons reaching their $(i+1)$th birthday during year n, N_2 the number of persons reaching their ith birthday during year n, λ, λ', and λ'' are the risk function and its two first derivatives at exact age $i+\frac{1}{2}$, s_1 and s_2 the densities of net entries inside the parallelogram containing the upper and the lower triangle, respectively, m_1, m_2, V_1, V_2, μ_1, and μ_2 the means, variances, and third centred moments of the locations of birthdays in the cohorts born in years $n-i-1$ and $n-i$, respectively.

If we know N_1, m_1, V_1, μ_1, N_2, m_2, V_2, and μ_2, and estimate λ' and λ'' from the values of λ obtained in the previous iteration at neighboring ages (with λ' and λ'' set equal to zero in the first iteration), it is possible to derive λ from the number of events in the square using the appropriate formula. The iteration process is again stopped when λ has stabilized for all i.

On the basis of the above formulae, it is possible to estimate the integral of λ over a 1-year range (between exact ages i and $i+1$ and also between exact ages $i-\frac{1}{2}$ and $i+\frac{1}{2}$), whatever the LEXIS diagram region in which occurrences were observed (squares, vertical-sided parallelograms, or triangles). This leads to an estimate of the risk regime at all relevant ages during any year of observation. These formulae make it possible to reach international comparability in spite of the diversity of national practices. (Some countries observe events in vertical-sided parallelograms, others in squares.) For countries where the national statistical office has changed from one system to the other, it also becomes possible to compute a complete time-series with comparable terms.

The Special Case of Mortality at Ages 0 and 1

The above formulae hold for fertility, nuptiality, and mortality at any age, except for mortality at age 0 and, to a lesser extent, at age 1. The reason is that the risk function $\lambda(x)$ of mortality decreases very sharply from birth i.e., from exact age 0. This makes the polynomial approximation invalid over the age-interval (0, 1) and even the interval (1, 2). We now provide a solution even for infant mortality.

Let us denote by D_L^0 and D_U^0 (with $D_L^0 + D_U^0 = SQ_0^n$) the numbers of deaths in the lower and the upper triangles of the first square (i.e., between exact ages 0 and 1), D_L^1 and D_U^1 the similar numbers of deaths in the triangles of the second square (i.e., between exact ages 1 and 2), D_L^2 and D_U^2 those in the triangles of the third square (i.e., between exact ages 2 and 3).

Under the assumption that there is no migration inside the first square and that distributions of birthdays within cohorts born in year $n-1$ and year n are identical (in practice: both departing from uniformity only due to 'usual' seasonal variations), the infant mortality rate $IMR^n = 1 - \exp\left(-\int_0^1 \lambda(x)\mathrm{d}x\right)$ can be estimated by

$$IMR^n = 1 - \left(1 - \frac{D_U^0}{P_0^n}\right)\left(1 - \frac{D_L^0}{B^n}\right)$$

where P_0^n is the population of completed age 0 on January 1 of year n and B^n is the number of live births of the corresponding sex during year n. If net migration is not zero inside the first square, we rewrite the preceding formula as

$$IMR^n = 1 - (1-a)(1-b)$$

Two estimates of each term a and b can be provided. The term a refers to the upper triangle. The first estimate assumes that migration during year n within the cohort born in year $n-1$ takes place exclusively on January 1 of year n. The second estimate assumes that migration during year n within the cohort born in year n takes place exclusively on December 31. We then get the following estimates:

$$a_1 = \frac{D_U^0}{P_1^{n+1} + D_U^0 + D_L^1} \text{ and } a_2 = \frac{D_U^0}{P_0^n}$$

The term b refers to the lower triangle, and our first estimate assumes that migration during year n within cohort born during year $n-1$ takes place exclusively immediately after birth. Our second estimate assumes that migration during year n within cohort born in year n takes place exclusively on December 31 This gives

$$b_1 = \frac{D_L^0}{P_1^{n+1} + D_L^0}, \quad b_2 = \frac{D_L^0}{B^n}$$

It follows that a set of lower and upper bounds of IMR^n are

$$1 - [1 - \min(a_1, a_2)][1 - \min(b_1, b_2)]$$

and

$$1 - [1 - \max(a_1, a_2)][1 - \max(b_1, b_2)]$$

If migration balances are small during year n both inside the vertical-sided parallelogram corresponding to the cohort born in year $n-1$ ($s_U = P_1^{n+1} - P_0^n + D_U^0 + D_L^1$) and inside the lower triangle ($s_L = P_0^{n+1} - B^n + D_L^0$), the two bounds are naturally close to each other. For example, the lower and upper bound of the estimated infant mortality rate were 5.30 and 5.36 per 1000 male live births and 4.16 and 4.13 per 1000 female live births, respectively, in France in 1997. For most practical purposes, the differences are negligible in contemporary Europe.

A satisfactory estimate of the infant mortality rate requires that the numbers of deaths are available for triangles in the LEXIS diagram. If the information available is limited to the total number of deaths SQ_0^n inside the first square, a rough estimate of the infant mortality rate is made possible nonetheless via a preliminary estimate of the breakdown of SQ_0^n into D_U^0 and D_L^0. In contemporary Europe, D_L^0 is approximately 90% of SQ_0^n.

The same procedure can be applied to the estimate of the probability of death between exact ages 1 and 2. The lower and upper bounds of that probability are

$$1 - [1 - \min(a_1, a_2)][1 - \min(b_1, b_2)]$$

and

$$1 - [1 - \max(a_1, a_2)][1 - \max(b_1, b_2)]$$

with

$$a_1 = \frac{D_U^1}{P_2^{n+1} + D_U^1 + D_L^2}, \quad a_2 = \frac{D_U^1}{P_1^n}$$

$$b_1 = \frac{D_L^1}{P_1^{n+1} + D_L^1}, \text{ and } b_2 = \frac{D_L^1}{P_0^n - D_U^0}$$

The lower and upper bounds of the probability of death between exact ages 1 and 2 are usually even closer to each other than the two estimates of the infant mortality rate. For example,

the lower and upper bounds were 4.839 and 4.843 per 10 000 male live births and 4.102 and 4.106 per 10 000 female live births, respectively, in France in 1997. In contemporary Europe, the share of the lower triangle in the second square (between exact ages 1 and 2) is only slightly over 0.5 (around 52–55%).

Literature

The LEXIS diagram got its name from the German statistician Wilhelm Lexis, who published a treatise on the subject some 125 years ago (Lexis, 1875). Like so many methodological innovations, the device is really older (Vandeschrick, 2001; Keiding, 2001). Most straightforward computational formulas above have a similar ancestry, but the more sophisticated ones were developed by the author in a series of publications between 1984 and 1999.

See also: Age Structure; Event Sequence Analysis: Methods; Period and Cohort Analysis in Demography.

Bibliography

Calot, G., et al., 1998. Two Centuries of Swiss Demographic History. Swiss Federal Statistical Office, Neuchatel.

Calot, G., et al., 1999. Un siglo de Demografía en Andalucía–La población desde 1900. Instituto de Estadística de Andalucía, Sevilla, Spain.

Calot, G., 1984. La mesure des taux en démographie, Cahier No. 104, Travaux et Documents de l'INED. Presses universitaires de France, Paris.

Calot, G. (1999). L'analyse démographique conjoncturelle. In A. Kuijsten, H. de Gans, & H. de Feiter (Eds.), Amsterdam: Thela Thesis Edited in honor of Dirk J. van de Kaa.

Calot, G., 1999. Statement on fertility made at the Round Table organised in Wiesbaden on the 25th anniversary of the BIB, Heft 93. Bundesinstitut für Bevölkerungsforschung, Wiesbaden, Germany.

Calot, G., 2000. Explanatory note on concepts and calculation methods. In: Recent Demographic Developments in Europe, 2000 issue of the Council of Europe Report. Council of Europe Publishing, Strasbourg, France.

Keiding, N., 2001. Graphical Representations in Mortality Measurement. Knapp, Zeuner, Becker, Lexis. University of Copenhagen, Denmark.

Lexis, W., 1875. Einleitung in die Theorie der Bevölkerungsstatistik. Trbner, Strasbourg, France.

Sardon, J.-P., 1992. La fécondité: Mesure et Évolution dans la Communauté Européenne. Collection Population et conditions sociales. Série D, Etudes et Analyses. Office for Official Publications of the European Community, Luxembourg, p. 169.

Vandeschrick, C., 2001. The Lexis diagram, a misnomer. Demographic Research 4, 97–124.

Demographic Techniques: Small-area Estimates and Projections

Stefan Rayer, College of Liberal Arts and Sciences, University of Florida, Gainesville, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.M. Alho, volume 5, pp. 3483–3486, © 2001, Elsevier Ltd.

Abstract

Estimates and projections for small areas are used extensively in the public and private sectors, and demand for them has been growing. Because of population size and data availability issues, estimates and projections for small areas face methodological challenges not commonly encountered at larger geographical scales. This article defines 'smallness' for estimation and projection purposes, describes the types and sources of data used, discusses the methods for estimating and projecting small area populations, and assesses recent methodological developments, in particular the impact of the geographic information system and spatial techniques.

Population estimates and projections play a critical role in market analysis, facility planning, environmental planning, and the allocation of public and private funds. Estimates and projections for small areas, in particular, are used extensively in the public and private sectors, and demand for them has been growing. Because of population size and data availability issues, estimates and projections for small areas face methodological challenges not commonly encountered at larger geographical scales. While most population estimation and projection techniques can be used at any scale, the focus of this article is on those approaches that are particularly suited for small areas.

The article is organized as follows: After a brief description of what distinguishes population estimates from projections, 'smallness' is defined and put in context. This is followed by a discussion of the data and sources typically used when preparing small area estimates and projections. Next, various types of estimates are defined. The central sections address in more detail the methods used for small area estimation and projection. The article concludes assessing recent methodological developments, in particular the impact of geographic information system (GIS) and spatial techniques.

Estimates, Projections, and Forecasts

Demographers consider information about a past or present population not based on a census or population register an *estimate*, while information about the future is referred to as a *projection* or a *forecast*. The terms projection and forecast are often used interchangeably, but can be differentiated according to the expected likelihood of their outcomes: a projection is a numerical outcome of a particular set of assumptions regarding the future population, whereas a forecast is the projection believed most likely to provide an accurate prediction of the future population (George et al., 2004: 561). In addition to the time dimension, a key distinction between estimates and projections involves the type of data used. Since projections refer to the size of the population at a future point in time, they cannot be based on actual data comprising the components of population change; rather, they must be based on the extension of either current or expected population trends into the future (Raymondo, 1992).

Smallness

Small area demography refers to demographic applications and analyses executed at local and regional scales (Smith and Morrison, 2005). While no universally accepted definition exists for what constitutes a small area for population estimation and projection purposes, it commonly involves administrative units below the national and state level. Examples include counties, cities, municipalities, townships, wards, local government areas, postal areas, school districts, census tracts, and census blocks; these can vary greatly in area and population size, from less than one to thousands of square kilometers, and from a handful of residents (or none) to over a million. Alternatively, a small area can refer to any subpopulation or domain for which direct estimates of adequate precision cannot be produced (Rao, 2003). Subpopulations or domains are subsets of larger populations defined by criteria other than geographic affiliation such as age, sex, income, educational level, and health status. Smallness can thus be understood in terms of size or in terms of data availability (Alho, 2001: 3484).

Small area analyses face a number of unique challenges not commonly encountered at larger geographic scales: (1) the boundaries of small areas often change over time making time-series analyses challenging; (2) many types of data, especially those covering more detailed population characteristics, are not tabulated for smaller areas, necessitating the use of proxy variables; (3) because of a paucity of data, there are often no discernible past patterns of change that can serve as a basis for estimation or projection, which may require the application of model rates based on areas for which data are available but which may not be directly comparable; (4) even when data are available for small areas, they may be less reliable because of smaller sample sizes and greater sampling variability; and (5) location-specific factors such as institutional populations, seasonal populations, facility closings, or changes in zoning have a greater impact on population changes in small areas (Smith and Morrison, 2005; Murdock et al., 2012). Because of

these challenges, population estimates and projections for small areas necessitate different sets of tools than those applied for larger areas such as states and nations.

Data and Sources

Two general categories of data can be distinguished: *direct data* and *indirect* or *symptomatic data* (Bryan, 2004b: 526). The former measure population and population change directly and can be obtained from censuses, administrative records, and surveys. Indirect data provide information indirectly related to, or symptomatic of, the population being estimated or projected; examples include school enrollment, tax returns, vital statistics, employment statistics, voter registration, electrical hook-ups, and housing counts. There is no clear dividing line between the two categories of data; the same data can be direct for one type of estimate and indirect for another. For example, birth and death counts are direct data when used to estimate natural change in a population but indirect data when used in the censal-ratio method (see later) to estimate total population (Swanson and Tayman, 2012: 43). The usefulness of indirect data depends on the extent to which factors other than population size and distribution influence them.

The primary data sources used in small area demography include censuses, administrative records, and surveys (Smith and Morrison, 2005). *Censuses* generally provide accurate and comprehensive data, regular repetition, fine geographic detail, and relative ease of data access; potential downsides include cost, enumeration difficulties, infrequent updates, falling response rates, differential underenumeration, and concerns about disclosure control (Martin, 2006). While censuses are still generally considered the 'gold standard,' in many countries there has been a movement toward greater utilization and application of administrative data sources and sample surveys to provide more timely data.

Administrative records, which are collected by federal, state, and local government agencies for registration, licensing, and program administration purposes, can provide ongoing information on a variety of demographic events and characteristics (Smith and Morrison, 2005). The most important administrative records for small area estimates and projections are vital statistics (especially birth and death data), which tend to be widely available and generally quite accurate, at least in the developed countries, and which form crucial inputs in many estimation and projection models. Other sources of administrative records come from a variety of fields and cover data related to the economy, education, health, social services, safety and security, community resources and participation, housing, and the environment (Coulton, 2008).

A special case of administrative records are *population registers* in which population characteristics are continually recorded. Population registers can be divided into *universal registers*, which attempt to include the entire population, and *partial registers*, which are established for specific administrative purposes and cover only those persons directly affected by the particular program (Bryan, 2004a: 31–35). Universal registers are maintained by the Scandinavian countries, the Netherlands, Japan, and a few other countries; partial registers are more common and include social insurance and welfare, military service, voter, school enrollment, and judicial data. A continuously updated enhanced master area file has been proposed for the United States; it has the potential to deliver timely, cost-effective, and precise population estimates even for small geographical units, but a number of challenges – ranging from confidentiality and privacy, up-front costs, to accuracy and technical issues – must be overcome for the system to become established (Swanson and McKibben, 2010). To be most useful, administrative records and population registers must be accurate and up to date, allow for linking from one source to another via a personal identification number, and be embedded in a supportive legislative framework (Smith and Morrison, 2005).

The third major source of data for small area estimates and projections is *surveys*. While surveys vary, their central features are the use of a fixed design, the collection of data in standardized form from individuals or organizations such as schools or businesses, and the selection of representative samples from known populations (Robson, 2011: 238). Surveys are often employed to collect data on variables not covered in a census or by administrative records. Surveys come from governmental/official statistics, academic/social research, and commercial/advertising/market research (O'Muircheartaigh, 1997: 1–2). Small sample sizes often limit the usefulness of survey estimates for small areas. This can be the case even for surveys designed specifically to provide accurate and timely demographic, social, and economic data on an ongoing basis for large and small areas, such as the American Community Survey in the United States (Swanson and Hough Jr., 2012). To combat the challenges posed by small sizes, small area estimation techniques have been developed that apply indirect estimators which 'borrow strength' by using values of the variable of interest from related areas and/or time periods, thus increasing 'effective' sample sizes (Rao, 2003: 2).

Types of Population Estimates

Population estimates can be divided – based on their time reference and method of derivation – into *intercensal estimates*, which relate to a date between two censuses and take the results of these censuses into account, and *postcensal estimates*, which relate to a date following a census that take that census into account, but not later censuses (Bryan, 2004b: 523). The former can be regarded as interpolations and the latter as extrapolations. Although postcensal estimates are sometimes made with extrapolative techniques, more commonly symptomatic indicators of population change are applied (see later). Estimates can also be made for dates prior to census taking (*precensal*), which are of interest to historical demographers in particular. The estimation methods discussed in the following section are primarily used for making postcensal estimates, which are the most common type of population estimation. Intercensal and precensal estimates can be made with some of these methods but usually require a different approach (Swanson and Tayman, 2012: 331–355). Estimates can further be divided whether they are made for a legally resident *de jure* population or for a physically present *de facto* population. Most population estimates follow a *de jure* definition, and the

methods described later are primarily applicable to them. *De facto* population estimates, which are useful for estimating daytime, visitor, seasonal, homeless, and disaster-impacted populations, for the most part utilize different sets of techniques (Swanson and Tayman, 2012: 313–330).

Methods for Small Area Population Estimates

The type and quality of data available are crucial determining factors when choosing a method for population estimation (Bryan, 2004b: 526). Estimation accuracy, and how different methods account for uncertainty, are important factors to consider; additional criteria include the provision of necessary detail, face validity, plausibility, costs of production, timeliness, and ease of application and explanation (Swanson and Tayman, 2012: 267–302). There are various ways to classify estimation methods for small areas. A recently developed classification scheme by Swanson and Tayman (2012: 106–107) combined earlier approaches and proposed the following categories: extrapolation, ratio, symptomatic, regression, component, sample based, and other methods. It should be noted that although the estimation methods discussed here have varying data requirements and can be applied across different areas, many less-developed countries have underdeveloped statistical systems and are better served by indirect estimation methods which utilize model life tables and model stable populations (Popoff and Judson, 2004). For an overview of data and methods used for at-risk subnational population estimation, see National Research Council (2007).

Extrapolation techniques, which range from simple linear change to complex ARIMA models, rely solely on the pattern of past population changes to estimate postcensal population, and assume that trends in the postcensal period will be similar to historical trends. Simple extrapolation models are most useful for postcensal estimates close to the last census, when resources are limited, and for very small areas or demographic subgroups; complex models allow the construction of probabilistic intervals around the estimates but are more difficult to implement and not necessarily more accurate (Swanson and Tayman, 2012: 115–127). *Ratio extrapolation methods*, such as share-of-growth, shift-share, and constant-share, express the population of a subgroup as a proportion of a larger population. Like simple extrapolation techniques, they have small data requirements and are easy to apply, but share the general shortcomings of extrapolation methods in that they do not account for differences in demographic characteristics or the components of growth, and they ignore potentially relevant information related to postcensal population changes (Swanson and Tayman, 2012: 127–135).

Symptomatic estimation techniques include the housing unit and censal-ratio methods. The *housing unit method* is one of the most widely used techniques for making small area population estimates (Bryan, 2004b: 550). It relies on the assumption that nearly everyone in a population lives in some type of housing structure. The housing unit method calculates the population of an area as equal to the number of occupied housing units (households) times the average number of persons per household plus the number of persons living in group quarters (e.g., prisons, college dormitories, military barracks, nursing homes). The number of households can be estimated using measures of construction activity such as building permits or certificates of occupancy, using utility data such as a residential electric or telephone customers, from property tax records, and from aerial photographs; the average number of persons per household can be taken from the most recent census, extrapolated as a trend from the two most recent censuses, or estimated using postcensal data in combination with data from the last census; and the group quarters population can be obtained either directly from the facilities or from a past census (Smith, 1986). The housing unit method has a long and successful track record, is flexible in terms of data sources, and can be applied at most levels of geography, but it also requires a major commitment of time and resources as well as sound professional judgment to yield accurate estimates (Smith and Cody, 2004). Recent research that models housing unit change for small areas using a density-dependent growth model, as well as approaches that integrate spatial factors derived from remote sensing and GIS datasets, have shown promising results and hint at future refinements to the housing unit method (Baker et al., 2008; Deng et al., 2010).

The *censal-ratio method*, introduced by Bogue (1950) as the 'vital rates method,' is another symptomatic population estimation technique used for small area estimation. It is related to the ratio extrapolation methods, but is based on ratios of symptomatic data to total population rather than proportions of national, state, or regional totals. The technique involves computing the ratio of symptomatic data to total population at the time of the last census, extrapolating the ratio to the estimate date, and dividing the estimated ratio into the value from the symptomatic series for the estimate date (Bryan, 2004b: 546–547). Symptomatic data that can be used in the censal-ratio method include birth and death statistics, school enrollment data, tax returns, number of electric, gas, or water meter accounts, number of building permits issued, bank receipts, motor vehicle registrations, and voter registration rolls. To be useful, accurate and comparable symptomatic data must be available at frequent intervals, including the census date, the annual number of cases should be high in relation to population size, and the ratio should be stable or change in a regular fashion in order to be projected to the estimate date (Bryan, 2004b: 547).

Regression techniques derive population estimates by means of symptomatic indicators of population change. The *ratio-correlation method* introduced by Schmitt and Crosetti (1954) is the most widely used regression technique for population estimation. It involves relating changes in several symptomatic indicators to population changes – expressed in the form of ratios to totals for geographic areas – by a multiple regression equation (Bryan, 2004b: 548). The symptomatic variables that have been used are similar to those discussed earlier for the censal-ratio method. The *apportionment*, *ratio change*, and *additive change methods* frequently used for making small area population estimates in Great Britain can be considered as simplified versions of the ratio-correlation method (Simpson et al., 1997). Regression techniques for population estimation have a firm foundation in statistical inference, which allows for the construction of meaningful measures of uncertainty (Swanson and Beck, 1994). The ratio-correlation method, in particular, has been used, evaluated, and refined

for over half a century, giving it a large measure of quality control and contributing to its widespread use and satisfactory performance (Swanson, 2004). Regression techniques rely for their accuracy on the validity of the assumption that the relationship between the independent and dependent variables observed in the past will persist in the postcensal period; they also require judgment with regard to the reliability and consistency of coverage of the symptomatic indicators. Shortcomings include potential time lags in the availability of the symptomatic indicators; the use of multiple and differing variables, which make decomposition of error and comparability of estimates between areas difficult; and limitations in producing postcensal estimates by age and sex (Bryan, 2004b: 549–550).

Component methods are based on the demographic balancing equation in which the population at the end of the time period (here, the estimate date) is expressed as the population at the beginning of the time period (commonly the most recent census), to which the number of births and in-migrants that occurred over the time period are added, and from which the number of deaths and out-migrants that occurred over the time period are subtracted. All component methods generally employ birth and death data but vary in how migration is estimated. Of the various component methods, the *component method-II* and the *tax returns/administrative records method* developed by the U.S. Census Bureau are widely used (Murdock et al., 1995; Starsinic et al., 1995). The component method-II employs changes in school enrollments to estimate migration, whereas the tax returns/administrative records method uses address changes on federal income tax returns. If migration data of sufficient quality are available, the *cohort-component method* can also be applied. It divides the population at the launch date into age-sex groups (cohorts) and accounts separately for the fertility, mortality, and migration each cohort experiences until the estimate date (Swanson and Tayman, 2012: 195–206). Although useful when estimates by age and sex are required, the cohort-component method is data- and computationally intensive and more commonly applied for population projections than for estimates. Component methods are attractive, because they specifically account for the three components of population change, births, deaths, and migration. Limitations include that they assume the continuation of historical patterns in symptomatic data in measuring migration, that they can be resource-intense and require data which may not be available for all small areas, and that they may not be appropriate for areas with substantial 'special' populations (Bryan, 2004b: 556; Rives et al., 1989: 30–31).

Sample-based methods of population estimation, which include *synthetic methods*, *structure preserving estimation*, the *ranked set samples method*, and *Bayesian methods*, are more commonly employed by statisticians than by demographers (Swanson and Tayman, 2012: 207–218); for an overview and review of small area estimation methods of this kind, see Ghosh and Rao (1994), Pfeffermann (2002), and Rao (2003). Other methods that are sometimes used for creating small area estimates include *structural models*, *economic-demographic models*, *dual system estimation*, *microsimulation models*, *neural networks*, the *grouped answer method*, *social network analysis*, and various methods related to *spatial demography* (Swanson and Tayman, 2012: 219–242). A *local census*, which provides a direct count of small area populations independent of any other data source, can also be used, though it is expensive and accuracy dependent on the response rate achieved (Simpson et al., 1996; Lunn et al., 1998).

In addition to individual techniques of population estimation, one can also estimate the population by the composite method and by combining estimates. The *composite method*, which was developed by Bogue and Duncan (1959), is a 'portfolio' of separate estimation methods that are tailored to particular segments of the population such as age groups (Bryan, 2004b: 550–551). For example, one could apply the censal-ratio method using births to develop an estimate for the population aged 0–4, the component method using school enrollment data for the population aged 5–17, the censal-ratio method using births and deaths for the population aged 18–64, and administrative records using Medicare data for the population aged 65 and older. The estimates for each age group are then summed up to yield an estimate of total population. Many combinations of methods and data are available, though there exists little empirical guidance in terms of which to choose (Siegel, 2002: 416–417). Another approach is to *combine estimates* made with two or more different estimation techniques. In its most simple way, this involves averaging or weighting the estimates directly; alternatively, one can join estimates from sample surveys and estimates based on demographic analysis as independent variables in a regression model. There is evidence suggesting that combining estimates made with different methods, data, and assumptions improves accuracy and reduces bias (Hoque, 2012; Simpson et al., 1998).

In recent years, much research has been devoted to small area population estimation methods that utilize remote sensing and GIS technologies (Wu et al., 2005; Wang and Wu, 2010). Dasymetric mapping – the redistribution of areal enumeration data using ancillary information for the display of statistical surface data – once confined to visualizing population distribution, is also increasingly being used for estimating the population of small areas (Mennis, 2009; Petrov, 2012). These methods are attractive since they are not tied to the administrative boundaries associated with census or survey data and allow population estimates to be made for user-defined areas. They also show potential for areas where census data and administrative records are unavailable, unreliable, or out of date, as is the case in many countries in the developing world (Hardin and Shumway, 2008; Sutton et al., 2001). While promising, there currently exists a disconnect between academic research and the actual adoption and use of remotely sensed technologies for population estimation purposes, with lack of expertise, technical, application, and financial issues preventing more widespread adoption by practitioners (Hoalst-Pullen and Patterson, 2011).

Methods for Small Area Population Projections

Many of the methods used for producing population estimates can also be applied when preparing population projections, and the sources of data and factors to be considered when choosing among methods are similar as well. There are important differences, though, since estimates are concerned with giving an accurate population count for a past or present

date, whereas projections require that assumptions about future population changes be made. Projections involve more uncertainty, and projection methods vary widely in how they attempt to model it. As was the case with estimates, there are various ways to classify projection methods. For small areas, Smith et al. (2013) propose four types: trend extrapolation, cohort-component, structural, and microsimulation models.

Trend extrapolation models determine future population values solely by their historical values; they range in complexity from simple methods such as linear, geometric, and exponential growth to complex polynomial, logistic curve fitting, and ARIMA time-series models. *Ratio methods* – in which the population or population change of a small area is expressed as a proportion of population or population change of a larger area in which the smaller area is located – is another type of extrapolation that is often used (Smith et al., 2013: 185–213). Recent research developments involve the application of ratio methods in combination with GIS techniques to produce small area projections based on grids (Hachadoorian et al., 2011). Trend extrapolation projection models have a number of shortcomings: on the one hand, they do not account for differences in demographic composition or for differences in the components of growth, provide little or no information on the projected demographic characteristics of the population, have no theoretical content, and generally cannot be related to theories of population change. On the other hand, the small data requirements, ease of application and explanation, timeliness, and low cost of simple trend and ratio models make them an attractive option particularly for small areas where data availability and reliability issues may preclude the application of more complex projection models (George et al., 2004: 571). Furthermore, simple trend models often provide reasonably accurate projections for short and even long horizons, and the empirical evidence suggests that more complex and sophisticated models do not offer more accurate projections, at least for total population (Armstrong, 1984; Chi, 2009; Smith, 1997).

Cohort-component models are the most commonly used population projection method. Going back to the work of Cannan (1895), the cohort-component method accounts separately for the three components of population change: births, deaths, and migration. Most models subdivide the population into age and sex groups, and each demographic subgroup is projected separately; further subdivisions by race, ethnicity, or other demographic characteristics are possible. While assumptions about the future need to be made for all three components of population change in a cohort-component model, migration tends to be the most significant, most volatile, and hardest to predict for small areas. Approaches for projecting migration include methods which use only base period migration data, methods which require some additional qualitative or quantitative information, and methods based on quantitative projections of independent variables (Wilson, 2011: 33–36). Cohort-component models vary further in whether they apply gross or net migration data. Gross migration models range from complex migration pool and multiregional models that are primarily useful for larger areas to simpler biregional models that are easier to apply for small areas. Net migration models can follow a top–down or bottom–up approach; a simplified version proposed by Hamilton and Perry (1962) that treats mortality and migration as a single unit is also sometimes used (Smith et al., 2013: 155–183). The cohort-component method is a very popular projection method because it can incorporate many different data sources, assumptions, and application techniques. It is well suited to provide projections of demographic characteristics in addition to projections of total population. However, the method is also very data intensive and relatively expensive to apply, and the potential lack of complete and reliable data can be a serious obstacle when projecting the population for small areas (Smith and Morrison, 2005).

Structural models form the third major group of projection methods. In a structural model, population change is related to changes in one or more explanatory variables, and the focus is generally on modeling migration. There are two major types of structural models: *economic-demographic* and *urban systems models*. Examples of the former include econometric models, models that balance labor supply and demand, models based on population/employment ratios, and regional economic models. Economic-demographic models generally focus on economic variables to project migration, though some include amenities as well; most models are complex, resource intensive, expensive to develop and apply, and primarily used for relatively large areas such as counties, metropolitan areas, states, and nations (Smith et al., 2013: 216–228). For smaller areas, urban systems models are more commonly applied; these typically include land use and transportation characteristics in addition to economic variables and rely heavily on GIS techniques. Like most economic-demographic models, urban systems models tend to be very resource intensive and often require a substantial degree of technical expertise (Smith et al., 2013: 228–237). While not necessarily more accurate than other types of projection methods, structural models are useful in that they can address a wide range of theoretical, planning, and policy questions. They can make important contributions to the planning and decision-making process and are particularly well suited for simulation and scenario purposes (George et al., 2004: 586).

While most population projections for small areas are made with trend extrapolation, cohort-component, or structural methods, a number of additional approaches are worth mentioning. The *housing unit method*, though more commonly applied to prepare population estimates, also shows potential for small area population projections (Foss, 2002). The method takes into account local housing supply and, either explicitly or implicitly, residential land availability, and can also be used to constrain cohort-component projections. It promises to be particularly useful for growing urban areas and areas earmarked for residential development, though projecting its various components is challenging (Wilson, 2011: 20–22). *Microsimulation models* differ from traditional demographic modeling as follows: they use a sample rather than the total population; they work on the level of the individual or the household rather than with grouped data; and they rely on repeated random experiments to derive a projection (van Imhoff and Post, 1998). By modeling individual behavior, spatial microsimulation models avoid aggregation bias, are internally consistent, provide very detailed projection outputs, and allow for scenario and what-if analyses; on the downside, the models tend to be extremely complex, require extensive data and staff resources, and there have been few attempts to

validate the assumptions and outcomes of these models (Birkin and Clarke, 2011; Wilson, 2011: 37–38). Another avenue of recent research involves the incorporation of spatial effects into the projection model. Originating in the geographic concept of spatial diffusion – the spread of a particular phenomenon over space and time – *spatial diffusion* and *spatio-temporal projection models* incorporate population spillovers and legacy effects arising from neighboring growth rates and characteristics into the model (Thrall et al., 2001; Chi and Voss, 2011). While these methods address a major shortcoming of most projection models used for small areas – which treat each unit of geography as an independent, stand-alone entity rather than as an entity surrounded by other areas with which they interact – they have not been found to outperform forecasts made with simple extrapolation methods (Chi and Voss, 2011; Chi et al., 2011).

As was true with estimates, *combining projections* using simple averages, weighted averages, trimmed means, and composite approaches, is another option. Combined projections are ideally derived from methods that differ substantially from one another and that draw from different sources of information; combining has been found to improve forecast accuracy under many conditions (Ahlburg, 1995; Armstrong, 2001, 2006; Clemen, 1989; Rayer, 2008). Finally, in addition to projections of total population and by demographic characteristics, many types of planning, budgeting, and analysis also require projections of households, school enrollment, employment, health, poverty, or other population-related characteristics. While structural models are often applied for these purposes, the *participation-ratio* and *cohort-progression methods* are worthwhile to consider. In the former, socioeconomic characteristics are related to demographic characteristics through the use of ratios; in the latter, projections are developed by 'surviving' people with particular socioeconomic characteristics (George et al., 2004: 586–590).

Among the most important advances in recent years has been the development of *probabilistic projection methods* (Wilson and Rees, 2005). Probabilistic methods acknowledge that forecasting the population entails a significant amount of uncertainty, especially for longer time periods, for places with small or rapidly changing populations, for certain age groups, and for the demographic components of population change. Conventional deterministic projections provide a single population number, or high and low variants that do not attach probability to the high–low ranges. Probabilistic methods, in contrast, assess and communicate uncertainty by providing forecasts that come with probability distributions or that are fully probabilistic, i.e., generated by probabilistic population renewal (Lee, 1998). Predictive intervals are commonly made with three distinct approaches: (1) time-series analyses; (2) expert-based probabilistic projections; and (3) ex-post analysis, i.e., the use of past forecast errors (Bongaarts and Bulatao, 2000: 200–204). Most of the research on probabilistic projection methods has focused on large geographic areas such as nations, states, and regions. Yet a probabilistic approach is arguably even more important for small areas, given the greater uncertainty of future population changes at the subnational level (Tayman, 2011; Wilson and Rees, 2005). Cameron and Poot (2011) discuss some of the challenges of providing probabilistic projections for small areas and offer directions for further methodological development.

From Administrative Units to User-Defined Areas?

In the first edition of this article, Alho (2001: 3484) opined that "with the development of the GIS, small area estimation and forecasting will no longer be limited by administrative boundaries." Arguing that administrative boundaries are frequently not well suited to many types of social and medical research, and that domains of interest may intersect several areas or be too small for their effect to be discernible in small area data, the introduction of GIS techniques was viewed as a major step forward for small area estimation and forecasting (Alho, 2001: 3485–3486). Smith et al. (2001: 365–367) also predicted a bright future for GIS in small area analysis, but argued that while GIS techniques may be particularly useful when combined with remote sensing to develop small area population and housing estimates, for projections the main value of GIS would likely be in developing databases rather than in constructing projections per se. Now, more than a decade later, one can ask: to what extent have small area estimates and projections incorporated spatial techniques, and how much progress has been made to overcome the limitations of estimating and projecting populations based on administrative units?

As discussed earlier, in recent years much research has focused on population estimation using remote sensing and GIS techniques. Not only have these techniques created new and alternative means of acquiring population estimates for small areas previously estimated with traditional methods, they also allow estimates to be made for user-defined areas that were hard to come by in the past. On the other hand, the literature on projections made using GIS techniques remains sparse. Projections require assumptions about the future, and while GIS and spatial techniques can add to our understanding of past changes and present population distributions, they do not make forecasting them easier. It seems likely that for the foreseeable future GIS and spatial techniques will remain most useful as tools for distributing populations projected with established demographic models to areas smaller than, or different from, those they were originally made for, rather than providing an alternative projection approach.

Most small area estimates and projections are still made for administrative units, and a major shift toward user-defined areas is unlikely to happen soon. Most of the required input data are still collected for administrative units, and methods that allow for customized bottom–up aggregations – such as microsimulation models – or methods that transcend administrative boundaries – such as remote sensing and GIS techniques – have not yet replaced estimates and projections made with established demographic models. Administrative units can be inconvenient to work with, but they often represent meaningful entities for which knowledge about past and present population changes exists. While small area population estimates and projections will always contain an element of uncertainty, this professional knowledge, when combined with a careful selection of data and methods, still

provides the best foundation for satisfactory estimation and forecasting outcomes.

See also: Censuses: Current Approaches and Methods; Demographic Models; Demographic Techniques: Indirect Estimation; Geographic Information Systems; Microsimulation in Demographic Research; Population Forecasts; Population Geography; Remote Sensing; Urban Planning: Methods and Technologies.

Bibliography

Ahlburg, Dennis A., 1995. Simple versus complex models: evaluation, accuracy, and combining. Mathematical Population Studies 5, 281–290.

Alho, Juha M., 2001. Demographic techniques: small-area estimates and projections. In: Smelser, Neil J., Baltes, Paul B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Elsevier, Amsterdam and New York, pp. 3483–3486.

Armstrong, J. Scott, 1984. Forecasting by extrapolation: conclusions from 25 years of research. Interfaces 14 (6), 52–66.

Armstrong, J. Scott, 2001. Combining forecasts. In: Armstrong, J. Scott (Ed.), Principles of Forecasting: A Handbook for Researchers and Practitioners. Kluwer Academic Publishers, Norwell, MA, pp. 417–439.

Armstrong, J. Scott, 2006. Findings from evidence-based forecasting: methods for reducing forecast error. International Journal of Forecasting 22, 583–598.

Baker, Jack, Ruan, Xiaomin, Alcantara, Adelamar, Jones, Troy, Watkins, Kendra, McDaniel, Michael, Frey, Margaret, Crouse, Nathan, Rajbhandari, Ruji, Morehouse, Jana, Sanchez, Jeremy, Inglis, Mike, Baros, Shirley, Penman, Shawn, Morrison, Susan, Budge, Tom, Stallcup, Wes, 2008. Density-dependence in urban housing unit growth: an evaluation of the Pearl-Reed model for predicting housing unit stock at the census tract level. Journal of Economic and Social Measurement 33, 155–163.

Birkin, Mark, Clarke, Martin, 2011. Spatial microsimulation models: a review and a glimpse into the future. In: Stillwell, John, Clarke, Martin (Eds.), Population Dynamics and Projection Methods, Understanding population trends and processes, first ed. vol. 4. Springer, Dordrecht, The Netherlands, pp. 193–208.

Bogue, Donald J., 1950. A technique for making extensive population estimates. Journal of the American Statistical Association 45 (250), 149–163.

Bogue, Donald J., Duncan, Beverly, 1959. A composite method for estimating post-censal population of small areas by age, sex, and color. In: Vital Statistics – Special Reports. National Office of Vital Statistics, Washington, DC.

Bongaarts, John, Bulatao, Rodolfo A. (Eds.), 2000. Beyond Six Billion: Forecasting the World's Population. National Academy Press, Washington, DC.

Bryan, Thomas, 2004a. Basic sources of statistics. In: Siegel, Jacob S., Swanson, David A. (Eds.), The Methods and Materials of Demography, second ed. Elsevier Academic Press, San Diego, CA, pp. 9–41.

Bryan, Thomas, 2004b. Population estimates. In: Siegel, Jacob S., Swanson, David A. (Eds.), The Methods and Materials of Demography, second ed. Elsevier Academic Press, San Diego, CA, pp. 523–560.

Cameron, Michael P., Poot, Jacques, 2011. Lessons from stochastic small-area population projections: the case of Waikato subregions in New Zealand. Journal of Population Research 18, 245–265.

Cannan, Edwin, 1895. The probability of a cessation of the growth of population in England and Wales during the next century. The Economic Journal 5 (20), 505–515.

Chi, Guangqing, 2009. Can knowledge improve population forecasts at subcounty levels? Demography 46 (2), 405–427.

Chi, Guangqing, Voss, Paul R., 2011. Small-area population forecasting: borrowing strength across space and time. Population, Space and Place 17, 505–520.

Chi, Guangqing, Zhou, Xuan, Voss, Paul R., 2011. Small-area population forecasting in an urban setting: a spatial regression approach. Journal of Population Research 28, 185–201.

Clemen, Robert T., 1989. Combining forecasts: a review and annotated bibliography. International Journal of Forecasting 5, 559–583.

Coulton, Claudia J., 2008. Catalog of Administrative Data Sources for Neighborhood Indicators. The Urban Institute, Washington, DC.

Deng, Chengbin, Wu, Changshan, Wang, Le, 2010. Improving the housing-unit method for small-area population estimation using remote-sensing and GIS information. International Journal of Remote Sensing 31 (21), 5673–5688.

Foss, Wayne, 2002. Small area population forecasting. The Appraisal Journal 70 (2), 163–172.

George, M.V., Smith, Stanley K., Swanson, David A., Tayman, Jeff, 2004. Population projections. In: Siegel, Jacob S., Swanson, David A. (Eds.), The Methods and Materials of Demography, second ed. Elsevier Academic Press, San Diego, CA, pp. 561–601.

Ghosh, M., Rao, J.N.K., 1994. Small-area estimation – an appraisal. Statistical Science 9 (1), 55–76.

Hachadoorian, Lee, Gaffin, Stuart R., Engelman, Robert, 2011. Projecting a gridded population of the world using ratio methods of trend extrapolation. In: Cincotta, Richard P., Gorenflo, Larry J. (Eds.), Human Population: Its Influences on Biological Diversity, Ecological Studies, vol. 214. Springer, Berlin, Germany, pp. 13–25.

Hamilton, C. Horace, Perry, Josef, 1962. A short method for projecting population by age from one decennial census to another. Social Forces 41 (2), 163–170.

Hardin, Perry J., Shumway, J. Matthew, 2008. Intraurban population estimation with remotely sensed imagery: a developing world perspective. Geography Compass 2 (3), 744–770.

Hoalst-Pullen, Nancy, Patterson, Mark W., 2011. Applications and trends of remote sensing in professional urban planning. Geography Compass 5 (5), 249–261.

Hoque, Nazrul, 2012. Evaluation of small area population estimates produced by housing unit, ratio-correlation, and component method II compared to 2000 census counts. Canadian Studies in Population 39 (1–2), 91–108.

van Imhoff, Evert, Post, Wendy, 1998. Microsimulation methods for population projection. Population: An English Selection 10 (10), 97–138.

Lee, Ronald D., 1998. Probabilistic approaches to population forecasting. Population and Development Review 24, 156–190.

Lunn, David J., Simpson, Stephen N., Diamond, Ian, Middleton, Liz, 1998. The accuracy of age-specific population estimates for small areas in Britain. Population Studies 52 (3), 327–344.

Martin, David, 2006. Last of the censuses? The future of small area population data. Transactions of the Institute of British Geographers 31 (1), 6–18.

Mennis, Jeremy, 2009. Dasymetric mapping for estimating population in small areas. Geography Compass 3 (2), 727–745.

Murdock, Steve H., Cline, Michael, Zey, Mary, 2012. Challenges in the analysis of rural populations in the United States. In: Kulcsar, Laszlo J., Curtis, Katherine J. (Eds.), International Handbook of Rural Demography. Springer, Dordrecht, The Netherlands, pp. 7–15.

Murdock, Steve H., Hwang, S., Hamm, R.R., 1995. Component methods. In: Rives, N.W., Serow, W.J., Lee, A.S., Goldsmith, H.F., Voss, P.R. (Eds.), Basic Methods for Preparing Small-Area Estimates. Applied Population Laboratory, University of Wisconsin, Madison, Wisconsin, pp. 10–53.

National Research Council – Committee on the Effective Use of Data, Methodologies, and Technologies to Estimate Subnational Populations at Risk, 2007. Tools and Methods for Estimating Populations at Risk from Natural Disasters and Complex Humanitarian Crises. National Academy Press, Washington, DC.

O'Muircheartaigh, Colm, 1997. Measurement error in surveys: a historical perspective. In: Lyberg, Lars, Biemer, Paul, Collins, Martin, de Leeuw, Edith, Dippo, Cathryn, Schwarz, Norbert, Trewin, Dennis (Eds.), Survey Measurement and Process Quality. Wiley, New York, pp. 1–25.

Petrov, Andrey, 2012. One hundred years of dasymetric mapping: back to the future. The Cartographic Journal 49 (3), 256–264.

Pfeffermann, Danny, 2002. Small area estimation – new developments and directions. International Statistical Review 70 (1), 125–143.

Popoff, Carole, Judson, D.H., 2004. Some methods of estimation for statistically underdeveloped areas. In: Siegel, Jacob S., Swanson, David A. (Eds.), The Methods and Materials of Demography, second ed. Elsevier Academic Press, San Diego, CA, pp. 603–641.

Rao, J.N.K., 2003. Small Area Estimation. Wiley, Hoboken, NJ.

Rayer, Stefan, 2008. Population forecast errors: a primer for planners. Journal of Planning Education and Research 27, 417–430.

Raymondo, James C., 1992. Population Estimation and Projection: Methods for Marketing, Demographic, and Planning Personnel. Quorum Books, New York.

Rives, N.W., Serow, W.J., Lee, A.S., Goldsmith, H.F., 1989. Small Area Analysis: Estimating Total Population. U.S. Department of Health and Human Services, Rockville, Maryland.

Robson, Colin, 2011. Real World Research, third ed. Wiley, Chichester, United Kingdom.

Schmitt, Robert C., Crosetti, Albert H., 1954. Accuracy of the ratio-correlation method for estimating post-censal population. Land Economics 30 (3), 279–281.

Siegel, Jacob S., 2002. Applied Demography: Applications to Business, Government, Law, and Public Policy. Academic Press, San Diego, CA.

Simpson, Stephen, Diamond, Ian, Tonkin, Pete, Tye, Rachel, 1996. Updating small area population estimates in England and Wales. Journal of the Royal Statistical Society, Series A 159, 235–247.

Simpson, Stephen, Diamond, Ian, Middleton, Elizabeth, Lunn, David, Cossey, Rachel, 1998. How can a national set of small area population estimates gain acceptance? Lessons from the estimating with confidence project. In: Arnold, Richard, Elliott, Paul, Wakefield, Jon, Quinn, Mike (Eds.), Population Counts in Small Areas: Implications for Studies of Environment and Health. The Stationary Office, London, England.

Simpson, Stephen, Middleton, Liz, Diamond, Ian, Lunn, David, 1997. Small-area population estimates: a review of methods used in Britain in the 1990s. International Journal of Population Geography 3 (3), 265–280.

Smith, Stanley K., 1986. A review and evaluation of the housing unit method of population estimation. Journal of the American Statistical Association 81 (394), 287–296.

Smith, Stanley K., 1997. Further thoughts on simplicity and complexity in population projection models. International Journal of Forecasting 13 (4), 557–565.

Smith, Stanley K., Cody, Scott, 2004. An evaluation of population estimates in Florida: April 1, 2000. Population Research and Policy Review 23 (1), 1–24.

Smith, Stanley K., Morrison, Peter A., 2005. Small-area and business demography. In: Poston, Dudley, Micklin, Michael (Eds.), Handbook of Population. Kluwer Academic/Plenum Publishers, New York, pp. 761–785.

Smith, Stanley K., Tayman, Jeff, Swanson, David A., 2013. A Practitioner's Guide to State and Local Population Projections. Springer, Dordrecht.

Smith, Stanley K., Tayman, Jeff, Swanson, David A., 2001. State and Local Population Projections: Methodology and Analysis. Kluwer Academic/Plenum Publishers, New York.

Starsinic, D.E., Lee, A.S., Goldsmith, H.F., Sparr, M.A., 1995. The census bureau's administrative records method. In: Rives, N.W., Serow, W.J., Lee, A.S., Goldsmith, H.F., Voss, P.R. (Eds.), Basic Methods for Preparing Small-Area Estimates. Applied Population Laboratory, University of Wisconsin, Madison, WI, pp. 54–69.

Sutton, P., Roberts, D., Elvidge, C., Baugh, K., 2001. Census from heaven: an estimate of the global human population using night-time satellite imagery. International Journal of Remote Sensing 22 (16), 3061–3076.

Swanson, David A., 2004. Advancing methodological knowledge within state and local demography: a case study. Population Research and Policy Review 23 (4), 379–398.

Swanson, David A., Beck, Donald M., 1994. A new short-term county population projection method. Journal of Economic and Social Measurement 20, 25–50.

Swanson, David A., Hough Jr., George C., 2012. An evaluation of persons per household (PPH) estimates generated by the American Community Survey: a demographic perspective. Population Research and Policy Review 31 (2), 235–266.

Swanson, David A., McKibben, Jerome N., 2010. New directions in the development of population estimates in the United States? Population Research and Policy Review 29, 797–818.

Swanson, David A., Tayman, Jeff, 2012. Subnational Population Estimates. Springer, Dordrecht, The Netherlands.

Tayman, Jeff, 2011. Assessing uncertainty in small area forecasts: state of the practice and implementation strategy. Population Research and Policy Review 30, 781–800.

Thrall, Grant I., Sidman, Charles F., Thrall, Susan E., Fik, Timothy J., 2001. The cascade GIS diffusion model for measuring housing absorption by small area with a case study of St. Lucie County, Florida. The Journal of Real Estate Research 8 (3), 401–420.

Wang, Le, Wu, Changshan, 2010. Preface: population estimation using remote sensing and GIS technologies. International Journal of Remote Sensing 31 (21), 5569–5570.

Wilson, Tom, 2011. A Review of Sub-regional Population Projection Methods. Queensland Centre for Population Research, The University of Queensland, Australia.

Wilson, Tom, Rees, Phil, 2005. Recent developments in population projection methodology: a review. Population, Space and Place 11, 337–360.

Wu, Shou-sheng, Qiu, Xiaomin, Wang, Le, 2005. Population estimation methods in GIS and remote sensing: a review. GIScience and Remote Sensing 42 (1), 58–74.

Demography, Early History of

Simon Szreter, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Demography has developed since the seventeenth century as a core set of analytical techniques for evaluating observed change in population aggregates in relation to the varying processes of mortality, fertility, and migration. It has been dependent on the apparatus of nation-states to generate its essential data of population age- and sex-structures and flows of vital events. Crucial technical developments have been the perfection of life tables and of stable population theory. The necessarily contentious interpretation of the causes of demographic change involves a wide range of ecological, political, and cultural factors, with the seminal work of Thomas Malthus (1766–1834) remaining highly influential.

Demography is the study of processes of change in human populations. Demography consists of three interrelated intellectual activities: the collection and validation of appropriate evidence; analysis using an expanding and sophisticated body of mathematical and statistical techniques, which are designed to avoid a range of pitfalls; and, finally, the interpretation of the relationship between the revealed demographic processes and the encompassing epidemiological, natural, economic, and social environments. To a limited extent these relations can be analyzed quantitatively but interpretation also requires engagement with wider issues relating to culture, religion, ideology, politics, and international relations, as exemplified in the seminal work of the field's most influential founding thinker, Thomas R. Malthus (1766–1834).

The gradual emergence of the modern discipline of demography between the seventeenth and nineteenth centuries was associated with an increasingly exclusive focus on mathematically tractable properties of population aggregates. Foucault has identified this schematically with an increasing focus of sovereign powers – typically nation-states – on their populations as aggregates, which came to be conceived within analytical demography analogous to biological entities of the natural world, with processes of fertility and mortality of paramount concern for explaining change over time. Hence knowledge of the age- and sex- structure of the population has been necessary in order to measure these processes accurately, requiring appropriate official census and registration systems to provide the necessary data. Kreager points out, however, that since at least Aristotle, there has also always been a second approach to the study of human population change, which addresses sociocultural and political questions of what determines mutually acknowledged membership of a defined population, independently of the biological facts of life of birth and death. As Bashford has recently shown, toward the close of the nineteenth century the most economically advanced nation-states became aware of the practical importance of such issues of membership of their national populations as their imperialist aspirations bumped up against each other in a newly 'closed world' lacking any further uncharted and unclaimed open frontiers. Such ecological competition was also informed by increasingly dominant racist and social Darwinist conflated perceptions of the physical and cultural differences between human groups. Attempts to control and guide or prevent certain population movements, to secure desired imperialist patterns of settlement across the globe, became a new object of state policy, with the proliferation of the technology of the international passport *c.*1880–1930, as both symbol of an individual's membership of a national population and tool for attempting to control migration.

As an institutionalized academic discipline, demography is a relatively recent arrival in the social sciences. Before World War II there were only a handful of established academic posts in demography, mostly privately funded in the US universities, a couple at the London School of Economics. Yet, officials performing at least some key population counting and accounting tasks have existed for millennia, by virtue of the obvious importance to political leaders of enumeration of populations, or at least sections of them, for fiscal, military, administrative, and colonial purposes. This dates back to the world's ancient civilizations, for instance, in Mesopotamia, China, and Egypt. Furthermore, compilation of family genealogies and maintenance of vital registers of births, marriages, and deaths both for religious and legal purposes have venerable histories, too; and the data they generated are now proving to be of value for demographic historians of early modern Europe, North and South America, Japan, and China, for instance.

Modern demographic study of population aggregates requires two kinds of information. First, on the current stock of the defined population (numbers of each sex alive at each year of age) usually supplied through accurate and regular census details (now taken decennially in most countries); and second, information on the continuous flow of vital events, the numbers of each sex born alive, and the numbers dying at each exact age. This permits calculation of the age- and sex-specific rates of fertility and mortality necessary for genuinely comparable measures of a population's principal dynamic demographic characteristics. Allowing for the effects of net in- or out-migration requires an additional source of information for accurate demographic measures of fertility and mortality, with various forms of undocumented migration always providing a challenge for demographers to estimate in order to produce unbiased estimates of population change. Many further kinds of demographic and epidemiological analysis can be undertaken provided that the appropriate detailed information is routinely collected, as can occur in advanced

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03031-2

universalist, health care systems, such as details on ages and causes of death, or on mothers' ages and parity (i.e., how many previous live births).

The study of populations of any size presupposes an official apparatus of data collection of some considerable sophistication and demography is therefore unique among the social sciences in the extent of its intimate involvement with the priorities of governments and officials and the wider influences of politics upon the state. Until recently only government bureaucracies could command the clerical resources required for the statistical analysis of national demographic data. Hence demography before 1945 tended to be the preserve of state officials or those with trusted official access. It was in fact government census bureaux, which provided the initial demand for automated punch-card tabulating machines and later for the production of the first nonmilitary electronic computers (the US Census Bureau's UNIVAC, commissioned for the 1950 census). The subsequent proliferation of cheap computing resources has undoubtedly been a necessary precondition for the expansion of demography as an academic discipline with numerous independent, nonofficial practitioners now able to participate in the field of study.

Demography has therefore had a close historical association with the rise of nation-states throughout the entire modern period. But demography's earlier development has also been strongly influenced at times by the needs of smaller, subnational polities, such as the municipalities of Britain since the seventeenth century or the federated states of eighteenth-century Germany. Furthermore, the institutional and intellectual character of demography's rapid disciplinary growth since World War II has been much influenced by explicitly transnational institutions such as the UN and the World Bank and major independent (though mostly American) philanthropic organizations, such as the Rockefeller and the Ford Foundations, which funded the postwar institutional expansion of demography and its associated theory of demographic transition through the UN Population Council and the first centers of training in Latin America, Asia, and Africa. Nevertheless, national governments remain the principal collectors of continuous series of verified, authentic demographic data, so that the nation remains the typical reference 'population' addressed.

Modern demography's origins are conventionally traced to the practitioners of 'political arithmetic' in Restoration England. From the fifteenth century onward, initially among some Italian city states, it was increasingly the practice of European cities to compile bills of mortality – weekly death records – as an early warning system for epidemics, particularly of plague. In London, the existence of this data provided the immediate context for the first recorded efforts at demographic analysis by John Graunt (1629–74), 'the father of modern demography.' In his *Natural and political observations made upon the bills of mortality* (1662), he also used records of London christenings to develop a rudimentary life table. This is the fundamental technique for the correct analysis of mortality because it divides the population into age sections and gives statistics for the probability of dying at each age, which can then be arithmetically summed to a single average life expectancy at any age, enabling populations with vastly different age structures or mortality regimes to be compared. Graunt had thus grasped the two fundamental demographic insights: that the age- and sex- composition of a population powerfully influences the incidence of vital events; and this means that population processes are all significantly interdependent.

Life table techniques developed very gradually over the ensuing century, with notable work by Sir Edmund Halley (1656–1742), the Dutchman Willem Kersseboom (*c.*1690–1771), the German Johann Süssmilch (1707–67), the Frenchman Antoine Deparcieux (1703–68), the Swiss-born Leonhard Euler (1707–83), and a host of British contributors to the so-called 'population controversy' (over whether or not the kingdom's population had increased since the Glorious Revolution of 1688–89, when the civil servant Gregory King had compiled, from tax returns, the only detailed estimate of a European nation's population made in the early modern period). Finally, the first technically correct empirical life table was published in 1815, for commercial actuarial purposes, by Joshua Milne (1776–1851), using accurate census and vital registration data collected for public health purposes in Carlisle by Dr John Heysham (1753–1834) during the years 1779–87.

During the nineteenth century the principal institution promoting the further technical development of demographic analysis was the General Register Office (GRO) in London. This British lead was an unlikely development, in that the British census was not even established on a permanent basis (it had to be invoked by a separate Act of Parliament every 10 years), whereas the Scandinavian countries of Iceland (1703) and Sweden (1749) had already instigated regular census-taking, while the new, expanding federation of the United States of America had also committed itself to regular censuses, as an integral part of its new process of democratic political representation. Meanwhile the French and Belgians, with the Marquis de Condorcet, Pierre-Simon Laplace, Joseph Fourier, Adolphe Quetelet, Alexandre Parent-Duchâtelet, and Louis Réné Villermé, initially led the world in the relevant mathematical fields of probability and the application of statistics to human biomedical affairs. The British lead in 'vital statistics' that opened up in the mid-nineteenth century was due to a chain of chance occurrences. The GRO was created by statute to oversee from 1837, a national civil registration system of births, deaths, and marriages, which was created not for public health or demographic purposes but as a liberal reform granting dissenters, a record of their vital events independent of the long-standing parish registers of the Established Church, in order to ensure nonconformists' legal rights in common with the rest of the population in relation to records for property-holding and inheritance. In 1840, as a temporary expedient this small department of government, the GRO, then became vested with responsibility for taking the nation's fifth census of 1841 (the first occurred in 1801). The Registrar General's recently appointed statistical assistant, William Farr, was an unknown man of humble origins, a trained medical doctor, and a self-taught statistician (partly through a visit to Paris, in 1830s – a leading center of application of statistics to medicine).

A representative member of the European-wide generation enthused by the promise of a statistical science of society inspired by the Belgian astronomer, Adolphe Quetelet (1796–1874), William Farr (1807–83) seized the opportunity presented by his responsibility for processing the two essential

sources required for demographic analysis. He launched a four-decade long career of innovative demographic and comparative epidemiological work. Britain was at this time in the throes of a profound mortality crisis in its industrial cities and Farr saw his work analyzing the death rates as guiding a national public health campaign to reduce 'preventable' mortality from the many infectious and sanitary diseases, which his studies demonstrated to be so variable in their incidence in different local government jurisdictions across the land. Of course rural Ireland also suffered the calamity of its great famine in the late 1840s but Farr's remit did not extend there. Although Ireland was an integral part of the United Kingdom from 1801, there was to be no vital registration system for the property-less, Catholic peasantry until 1864. Thus, while it was in a Parisian publication that the term 'demography' was first coined in 1855 – by Quetelet's Belgian countryman, Achille Guillard – it was in Britain that it was most intensively practiced and developed as 'vital statistics.' Significant work was certainly also done in other official statistical offices. Indeed, the most important subsequent innovation following the perfection of the life table in the development of demographic techniques of analysis was the Lexis diagram invented by the German statistician, Wilhelm Lexis (1837–1914), in 1875 for representing the relationship between cohorts (or generations) and period effects (events).

The GRO's lead was surprising in another sense in that the single English name most widely associated with the origins of demography is not the civil servant William Farr with his primary medical focus on preventable mortality, but the cleric and academic, Thomas Robert Malthus, whose focus was on the economy and preventable fertility. Malthus's *An essay on the principle of population, as it affects the future improvement of society*, published anonymously in 1798, famously enunciated on page 14 of its opening chapter that "Population, when unchecked, increases in a geometrical ratio. Subsistence increases only in an arithmetical ratio." He argued that all policy to enhance a population's welfare was conservatively constrained by this limitation principle. Malthus, who held the first ever Chair in Political Economy in Britain and regarded Adam Smith as his principal mentor, thereby applied the fundamental economic problem of scarcity to demography and produced an ecological analysis, which, as is well-known, subsequently independently stimulated first Darwin and then A.R. Wallace to each apply a similar logic to the problem of the origin of species, thereby coauthoring in 1858 the theory of evolution through survival of the fittest.

Malthus was therefore, at once, a seminal influence on evolution, ecology, and demography. In relation to demography this was especially in terms of the third intellectual activity outlined above, interpretation. His forthright axioms and their policy implications demanded a response from the intellectuals of every country and secured a permanent place for the discussion of population problems in the social and biological sciences and in the corridors of government and power. In the five-times revised and extended editions of his *Essay* in his own lifetime he continued to feed the debate, further analyzing and exploring the manner in which aggregate population change, mortality and migration, and individuals' and families' reproductive behavior is integrated into the encompassing natural, social, and economic environment, while utilizing an ever-increasing range of examples drawn from across the world to show how profoundly they were also influenced by customs, beliefs, and politics. He notoriously drew the wrath of Marx for his support for the radical reduction of the nation's welfare net which resulted in the amendment of the Poor Laws in 1834, on the grounds that its 'generosity' was discouraging prudence among the prolific poor who married too young and so had too many children. However, as Bashford points out, Malthus cannot be caricatured simply as a heartless apologist for bourgeois capitalism and European expansion – his moral concerns for the future of the indigenous inhabitants of North America and their way of life contrast interestingly with the contemporary views of Benjamin Franklin, for instance. While Mao Zedong's initial policy for Communist China, 'more people, more power,' was, following Marx, consciously anti-Malthusian, China retained allegiance to communism when it radically reversed this policy with the hyper-Malthusianism of the one-child policy from the late 1970s under the influence of the 'Song group' of technocrats, whose rise to policy preeminence has been documented by Greenhalgh.

Throughout the remainder of the nineteenth century, however, most of those studying demographic phenomena were, like Darwin, gentlemen scholars in local and national statistical societies; also medical men, municipal, and some state officials scattered all across Europe and North America. From mid-nineteenth century they met periodically at International Congresses of Statistics, then of 'Hygiene and Demography' and eventually of 'Eugenics.' But it was not until the interwar decades of the twentieth century that demography began to emerge with an independent disciplinary and professional identity, marked initially by the formation in 1928 of the International Union for the Scientific Study of Population, followed in 1931 by the foundation of the Population Association of America.

However, this formative era from the late nineteenth century onward had also witnessed the rise of nationalist and imperialist rivalries and associated racist, social Darwinist ideologies. This came to exert a dominant (though not uncontested) influence over the practice of demography during the interwar years. French demography was perhaps the first to experience this overtly nationalist and xenophobic current, in the form of a defensive pronatalist popular reaction to defeat in the Franco–Prussian war of 1870–71. This was partly ascribed to her extremely low birth rate, the French being the first nation in the world to have practiced widespread family limitation. The titles of the works of France's leading demographers, such as Arsène Dumont, Jacques Bertillon, and Adolphe Landry repeatedly referred to 'Dépopulation' and 'decadence' while her great novelist, Zola, writing in exile in London after the Dreyfus affair, mourned France's lost *Fécondité* (1899). The study of population problems in the interwar period became increasingly subsumed within this social Darwinist agenda. With its continuing necessary association with the state, demography was inevitably heavily implicated in the policies of the Fascist and Nazi regimes with their totalitarian ideologies. Under the Nazi regime the potential for both demographic and economic forms of analysis to be applied as analytic systems of logic devoid of humane or moral considerations became tragically manifest. Nazi Bevölkerungswissenschaft mechanically

extended the notion of an economically optimal density of population (relative to capital and living standards, rather than simply land), which had underpinned Hitler's racist notion of Lebensraum, to arrive at apparently logical but murderous conclusions that population densities should be reduced in inefficient circumstances, where Jews and Poles were concerned, for instance.

International demography conferences in both Rome and Berlin in the late 1930s were partially boycotted and disrupted due to concern at the emerging demographic policies of Fascist regimes. But even in more liberal Britain, USA, or social democratic Sweden, demographic analysis became obsessed with class-differential fertility, 'problem families,' and fears of a 'flight from parenthood' associated with a succession of illiberal movements for national efficiency, eugenics, and 'social hygiene.' Only in the ideologically separatist USSR, was population analysis relatively free from this association (until Stalin's own pronatalist drive during the war). But this was primarily because of the absence of Western-style demographers, due to Marx's historic demonization of Malthus as the ultimate apologist of the bourgeoisie and their 'dismal science' of political economy. In the massive, unmanageable continental empire of the new Soviet Union, Hirsch shows how the alternative (Aristotelian) 'compositional' and membership approach to population accounting was dominant in the official practice of the revolutionary people's commissars: a preoccupation with mapping the geographical distribution of ethnic nationals (*narodnosti*) and with the control of movement of categories of workers in the economy through the use of an internal passport system in fact inherited from the Tsarist regime, the Propiska, similarly to the Chinese Communist state's subsequent adapted usage of the ancient Hukuo system of household registration.

In the non-Communist world in the immediate aftermath of World War II and Nazi excesses, in Europe, it was only in France and Britain that institutions for demographic research positively flourished. L'Institut National d'Études Démographiques under Alfred Sauvy (1898–1990) and the Population Investigation Committee of the London School of Economics under David Glass (1911–78) both founded their respective journals in 1946, *Population* and *Population Studies*, respectively. However, the decade of the 1940s in fact marks a significant shift in which the center of gravity of the discipline of demography migrated from the colonial powers of Europe to USA. It was symbolic of this coming change that in work published between 1907 and 1939 a European-born but American-naturalized mathematician, Alfred Lotka (1880–1949), devised the proofs for the equations of renewal processes, which underlie stable population theory (showing that populations always come to exhibit a constant age structure and growth rate if their fertility and mortality rates are unchanging). This has proved to be the most important mathematical advance in demography since the perfection of the life table technique (which applied to nonrenewal processes). Intriguingly, unknown to Lotka, Leonhard Euler (1707–83) had already solved these equations in the eighteenth century. Indeed, Lotka's work itself remained largely unexploited until after World War II, both facts which demonstrate that technical advances within social science disciplines do not in themselves necessarily cause change but are dependent on the institutional context and on 'demand' factors external to the discipline.

After World War II, demography rapidly responded to the 'demand' of the growing global preoccupation with planned economic growth, propelled in this direction by the US strategic interests in the context of the Cold War rivalry with the USSR for client states among the many decolonizing peoples of the less-developed world. Consequently, the period since 1945 has become a relatively clearly demarcated third era in the history of demography when, with racist sociobiology discredited, in many senses it returned to the original Malthusian economic agenda and a policy preoccupation with fertility restraint, especially in the 'developing' countries where most of the world's poor reside. The model of 'demographic transition' has dominated the discipline's approach throughout this era. This is a general historical model, promoted and adapted by Frank Notestein (1902–83), the highly influential and well-connected first Director, 1936–59, of the Princeton Office of Population Research (OPR), the most important institutional center for the new American-led demography. Demographic transition 'theory' proposes that in order for a nation to achieve sustained long-term economic growth it must reduce both its mortality and, especially, its fertility levels to those attained by the 'modernized,' liberal democratic Western countries. The model has surprisingly little consistent support in either the historical or contemporary record but it has survived repeated revisionist challenges and has undoubtedly served the discipline well as a model that is compliant with an overarching ideological position endorsing the goals of promoting economic development of liberal market economies round the world. This has successfully elicited a continuous flow of funds for both research and family planning programs from Western governments and philanthropic foundations.

It was in this policy context that external demands increased for various applications of stable population theory. Within the dominant postwar liberal consensus of 'modernization theory,' economic growth was seen as a generic process, such that less-developed countries could be engineered to repeat the experience of the developed countries. Fertility restraint and mortality reduction were important population policy inputs to permit the occurrence of incremental capital–output ratios, considered essential for industrialization. This would be jeopardized by too great a demographic dependency burden (too many young and invalid old in the population) diverting national income into current consumption instead of savings and capital accumulation. Stable population theory enabled model life tables to be devised by A.J. Coale (1917–2002) (Notestein's successor as Director of OPR) and Paul Demeny, and indirect estimation techniques to be formulated by W. Brass (1921–99) of the London School of Hygiene, which permitted the measurement of demographic parameters in many of the less-developed countries where primary information, both census and, especially, vital registration, was defective. A further application, for the same reason, has been its use in back projection and generalized inverse projection in historical demographic work. Along with the application of ground-breaking work in the 'family reconstitution' of historical communities by Louis Henry (1911–91), this has enabled Sir E.A. Wrigley (b.1931) and his colleagues at the Cambridge Group for the History of Population and Social Structure to utilize the parish registers of

baptisms, marriages, and burials originally established by the Tudor monarchy to reconstruct the detailed demographic record of the first industrial nation, England, back to the mid-sixteenth century, thereby, incidentally, demonstrating the invalidity of the demographic transition model in this important case.

Demography has achieved a considerable degree of disciplinary and institutional success, in many senses enviable by the standards of the other social sciences. Since 1945, training institutions have proliferated, primarily under the aegis of the Rockefeller-funded Population Council located in New York, so that demographers are now trained in many centers in all the continents. By plausibly reducing, for certain purposes, the complexity of human affairs in matters of reproduction, morbidity, mortality, migration, and mobility, to tractable mathematical models, demography has achieved considerable analytical power. Like classical and neoclassical economics, conventional demographic analysis embraces a positivist epistemology. It also starts from the liberal and 'democratic' foundational fiction that each individual can, in principle, be legitimately treated as an equivalent accounting unit, a presumption that could only make sense in a post-Lockean world, where a working man could be deemed to count for the same as a prince, hence demography's first text by John Graunt came from the hand of an exact contemporary of John Locke's, another survivor of the modern world's first overthrow of a supposedly divinely appointed sovereign. The resulting accounting system has provided the basis for an increasingly elaborate range of useful technical developments, such as, recently, multistate and multiregional demography.

Inevitably, however, there is an intellectual price to be paid for the preservation of demography's analytical rigor. Rather like the practice of so-called 'mainstream economics,' this, along with its close relationship with the policy needs of governments and other agencies involved in planning and development practices, has tended to insulate demography somewhat from the discipline of full engagement with the rest of the social sciences and humanities. Like all scientific activities, demography socially and ideologically constructs its objects of study but there has been little explicit acknowledgment of the implications of this. Thus, for a discipline, which has devoted much energy to refining, for instance, calibration of the proximate (biological and physiological) causes of fertility variation, there has been much less interest in understanding how the demographers' own concept of fertility itself has arisen, and this applies to other social categories and their context dependence, such as 'population,' 'family,' or 'household.' This disinclination to engage in critical reflexive practice has undoubtedly had significant and, for such an applied and policy-oriented discipline, real consequences. For many decades, vast resources have been devoted to 'family planning' programs aiming to transplant 'modern,' 'Western' contraceptive practices into an enormous variety of diverse cultures. More policy effectiveness might have been forthcoming if both fertility and birth control had from the start been approached as culturally variable entities, whose meanings had to be researched in any population before attempting to modify them.

The future of demography, established as it now is in all the world's major cultures and with departments of practitioners in most governments, is likely to include much more active discussion of the relationship between its inherited inventory of 'Western' social categories and those of other peoples, something also likely to follow from the increasing sophistication of historical and anthropological demography. It is likely to be less exclusively consumed with the problems of engendering economic growth and more with social equity and public health issues, including those of longevity and 'aging,' along with – another Malthusian inheritance – ecological problems of the uses and distribution of material wealth, plus the major issues of the rise of international migration and mobility in a more culturally interconnected ('globalized') but also more unequal world economy. This would, to some extent, represent a return, on a global scale, to the issues, which so preoccupied the mid-nineteenth century demographic observers of the human disruption wreaked in Britain and Europe by the world's first great wave of rapid economic growth. At the same time greater reintegration of demography into the intellectual concerns of both the biosciences and also the wider humanities is likely to be mutually beneficial, with the recent advances in study of the human genome, greater appreciation of gender and sexual diversity, and also of the changing planetary environment as three salient and diverse examples of major recent intellectual developments likely to impact on the ways demography is used and studied in the future.

See also: Data Bases and Statistical Systems: Demography; Demography: History Since 1900; Families and Households, Formal Demography of; Historical Demography; Malthus, Thomas Robert (1766–1834); Population Aging: Economic and Social Consequences; Population Forecasts.

Bibliography

Anderson, M.J., Citro, C.F., Salvo, J.J., 2012. Encylopedia of the U.S. Census: From the Constitution to the American Community Survey, second ed. Sage Copress, Los Angeles, CA.

Bashford, A., 2014. Global Population: History, Geopolitics, and Life on Earth. Columbia University Press, New York.

Breckenridge, K., Szreter, S. (Eds.), 2012. Registration and Recognition. Documenting the Person in World History. Proceedings of the British Academy, vol. 182. Oxford University Press, Oxford.

Desrosières, A., 1998. The Politics of Large Numbers: A History of Statistical Reasoning. Harvard University Press, Cambridge, MA.

Dupâquier, J., 1984. Pour la démographie historique. P.U.F., Paris.

Dupâquier, M., Dupâquier, J., 1985. Histoire de la démographie. Perrin, Paris.

Eyler, J.M., 1979. Victorian Social Medicine. The Ideas and Methods of William Farr. Johns Hopkins University Press, Baltimore, MD.

Foucault, M., 1991. Governmentality. In: Burchell, P., Gordon, C., Miller, P. (Eds.), The Foucault Effect: Studies in Governmentality, with Two Lectures by and an Interview with Michel Foucault (R. Pasquino, Trans.). Harvester, Hemel Hempstead, UK, pp. 87–104.

Girard, A., 1986. L'institut national d'études démographiques. Histoire et developpement. INED, Paris.

Glass, D.V., 1973a. The Development of Population Statistics. D C Heath, Farnborough, UK.

Glass, D.V., 1973b. Numbering the People. The Eighteenth-Century Population Controversy and the Development of Census and Vital Statistics in Britain. D C Heath, Farnborough, UK.

Gotz, A., Heim, S., 2002. Architects of Annihilation. Auschwitz and the Logic of Destruction (A.G. Blunden, Trans.). Princeton University Press, Princeton, NJ.

Greenhalgh, S., 1996. The social construction of population science: an intellectual, institutional and political history of twentieth-century demography. Comparative Studies in Society and History 38 (1), 26–66.
Grrenhalgh, S., 2008. Just One Child: Science and Policy in Deng's China. University of California Press, Berkeley, CA.
Hacking, I., 1990. The Taming of Chance. Cambridge University Press, Cambridge, UK.
Hauser, P.M., Duncan, O.D. (Eds.), 1959. The Study of Population: An Inventory and Appraisal. University of Chicago Press, Chicago, IL.
Hirsch, F., 2005. Empire of Nations: Ethnographic Knowledge and the Making of the Soviet Union. Cornell University Press, Ithaca, NY.
Hodgson, D., 1983. Demography as social science and policy science. Population and Development Review 9 (1), 1–34.
Hodgson, D., 1991. The ideological origins of the population association of America. Population and Development Review 17 (1), 1–34.
Ipsen, C., 1996. Dictating Demography: The Problem of Population in Fascist Italy. Cambridge University Press, Cambridge, UK.
Kreager, P., 1988. New light on Graunt. Population Studies – Journal of Demography 42 (1), 129–140.
Kreager, P., 2008. Aristotle and open population thinking. Population and Development Review 34 (4), 599–629.
Kreager, P., 2009. Darwin and Lotka: two concepts of population. Demographic Research 21, 469–502.
Le Bras, H., 2008. The Nature of Demography. Princeton University Press, Princeton, NJ.
Lee, R., 2004. Official statistics and demography in the Third Reich. In: Mackensen, R. (Ed.), Bevölkerungslehre und Bevölkerungspolitik im "Dirtten Reich". Leske + Budrich, Opladen, pp. 101–124.
Notestein, F.W., 1982. Demography in the United States: a partial account of the development of the field. Population and Development Review 8 (4), 651–687.
Ryder, N.B., 1964. Notes on the concept of a population. American Journal of Sociology 69, 447–463.
Schweber, L., 2006. Disciplining Statistics: Demography and Vital Statistics in France and England, 1830–1885. Duke University Press, Durham, NC.
Soloway, R.A., 1990. Demography and Degeneration. Eugenics and the Declining Birthrate in Twentieth-Century Britain. University of North Carolina Press, Durham, NC.
Szreter, S., 1991. Special issue: the general register office and the public health movement 1837–1914. Social History of Medicine 4, 435–463.
Szreter, S., 1993. The idea of demographic transition and the study of fertility change: a critical intellectual history. Population and Development Review 19 (4), 659–701.
Szreter, S., 1996. Fertility, Class and Gender in Britain, 1860–1940. Cambridge University Press, Cambridge, UK.
Szreter, S., Sholkamy, H., Dharmalinghan, A., 2004. Categories and Contexts. Anthropological and Historical Studies in Critical Demography. Oxford University Press, Oxford.
Teitelbaum, M., Winter, J., 1985. The Fear of Population Decline. Academic Press, New York.
Von Glahn, R., 2012. Household registration, property rights and social obligations in imperial China: principles and practices. chapter 1. In: Breckenridge, K., Szreter, S. (Eds.), Registration and Recognition. Documenting the Person in World History, Proceedings of the British Academy, vol. 182. Oxford University Press, pp. 39–66.
Wrigley, E.A., Davies, R.S., Oeppen, J.E., Schofeld, R.S., 1997. English Population History from Family Reconstitution 1580–1837. Cambridge University Press, Cambridge.

Demography: History Since 1900

Dennis Hodgson, Fairfield University, Fairfield, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

For an intellectual activity to become a discipline, certain things are needed: a significant interaction among those performing the activity; an organization that directs its investigation and teaching; and entities outside itself who value and use its findings. Using this definition, demography lacked the status of a discipline for the first half of the twentieth century. It became a discipline in the second half of the twentieth century largely as a result of a confluence of events that took place at mid-century, the most notable of which was the rise of an international population control movement that expanded vastly the resources available for demographic research on problems associated with population growth in developing countries. These resources helped to establish demography within the academy and to shape its contours, but also worked to politicize the discipline for several decades. Currently, demography possesses a disciplinary cohesiveness that arises from its well-defined subject matter and methodology, and from a common set of policy concerns. Certain twenty-first century trends, however, are likely to erode that cohesiveness.

Achille Guillard first used the word 'demography' in 1855 to describe a new science of population dynamics, by which he meant the statistical study of the size, condition, and movement of populations. Malthus, a half century earlier, had initiated a period of speculation on population growth's constraining effect on prosperity, but only by mid-century had the collection of population statistics become extensive enough in Europe and the United States for a 'science' of population to be imagined. The rise of industrial capitalism in Europe and its overseas extensions increased the pace of social, demographic, and economic change, stimulated the development of state structures aimed at directing these changes, and made obvious the usefulness of accurate population statistics. As India's first census in 1868 reflects, such usefulness extended to the administration of colonies as well. In 1882 the fourth International Conference on Hygiene (Geneva) included a section on demography. By its next meeting in 1884 (The Hague) the conference included demography in its title and sections on "the science of statistics applied to the social well-being of people" were presented. Although the growing body of population statistics was being discussed at international meetings, as the nineteenth century came to an end a unified discipline of demography, with a distinct body of theory and method, had yet to emerge.

Shils notes that for an intellectual activity to become a discipline, certain things are needed: a significant interaction among those performing the activity; an organization that directs its investigation and teaching; and entities outside itself who value and use its findings. Using this definition, demography lacked the status of a discipline for much of the twentieth century. The century began with a diverse group of individuals considering themselves population specialists: statisticians, public health officials, biologists, economists, sociologists, geographers, historians, lawyers, political activists, and politicians. There was haphazard interaction among this group and the teaching of demography had yet to find a clear place within the academy. While basic population statistics had an undeniable utility for those engaged in pragmatic planning activities, there had yet to develop a consensus about the larger significance of modern population changes. In fact, while consensus was developing about the causes of modern population changes, the twentieth century began with growing dissension about their societal consequences. Those engaged in examining population trends warned of very different population problems and recommended conflicting population policies. With such disparate voices speaking about population, a harmonizing core discipline was difficult to detect. The lack of a unifying theoretical perspective that explicated why population mattered, and the presence of a fractured population policy agenda were two factors that obstructed the development of a distinct discipline of demography; these factors continued to inhibit disciplinary development throughout much of the twentieth century.

During the nineteenth century, social movements aimed at influencing population trends arose that attempted to provide the unifying theoretical perspective and clear policy agenda for the study of population statistics. Neo-Malthusianism, based on the belief that growing populations are a major cause of poverty and that lowering fertility by making contraception more accessible will facilitate prosperity, originated in Great Britain early in the nineteenth century and spread gradually throughout Europe. The British Malthusian League was established in 1877, followed by neo-Malthusian leagues in Holland (1885), Germany (1889), and France (1895). But by the time the leagues united at the first International Neo-Malthusian Conference (Paris) in 1900, fertility decline had so spread among urbanizing and industrializing societies that the crisis aura surrounding population growth and large families largely had abated. For countries like France, which were beginning to face the specter of depopulation in the context of an increasingly quarrelsome Europe, it had become less clear that population growth was a social evil or that small families were a social good.

The fertility decline experienced in more developed societies generated new population movements as the twentieth century began. The upper and middle classes were embracing the small family norm more quickly and more fully than other classes. The rediscovery of Mendel's research and the power of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31107-2

Weismann's work on an inviolable germ plasm had convinced many that heredity played a transcendent role in molding human destiny. The educated of the period were much inclined to use social class position as a surrogate measure of biological quality, and they bemoaned the 'race degradation' that resulted from the 'less fit' lower classes contributing an increasing proportion of future generations.

As early as 1891 Sir Francis Galton, addressing the demography section of the International Conference on Hygiene and Demography in London, was urging that the issue of heredity and fertility be dealt with in the 'practical politics' of member states. As the twentieth century began, eugenics movements were forming in a number of countries. In the United States an active immigration restriction movement allied itself with those expressing eugenic fears. Restrictionists railed against the flood of 'inferior' Eastern and Southern European immigrants who were 'degrading' the high quality of the largely Anglo-Saxon and Northern European 'stock'; they even claimed that the increased number of the 'inferior' was a major cause of fertility decline among the 'superior.'

The eugenists and restrictionists feared changes in population composition and quality, not population quantity and growth. The trend that neo-Malthusians praised, fertility decline, was the same one that they bemoaned. Most eugenists fought to keep access to contraceptives restricted, contending that only the 'more fit' higher classes were sufficiently disciplined to limit their family size. Into this amalgam of population movements Emma Goldman, Margaret Sanger, and Marie Stopes attempted to add 'birth control' movements in the United States and Great Britain. During a period when high-ranking politicians were reminding educated women publicly of their patriotic duty to marry and have children, they began mobilizing citizens to legalize a woman's access to contraception.

With most academic population experts at the time producing works that fed the eugenic-based fears of elites, the task of birth controllers was to develop convincing counterarguments as to why legalizing contraception was socially beneficial. The socialist arguments that Emma Goldman presented – that working-class women could improve their class' negotiating position with capitalists by restricting the production of new workers – had limited public appeal.

Margaret Sanger attempted to attract the allegiance of experts and a wider public by deftly weaving together eugenic and neo-Malthusian themes: restrictive laws do not keep contraception out of the hands of the educated classes and only serve to slow the adoption of birth control among the less motivated 'inferior' classes, an adoption that would benefit both the individual and the society. When during the 1920s the pace of fertility decline increased in developed societies, fears of actual depopulation began to grow and additional controversy swirled around population trends.

Many consider demography a discipline that is either defined by its specialized techniques or delimited by its particular area of study: the causes and consequences of population change, especially the mortality and fertility transitions that have so dramatically swept the world since the 1750s. But for the first third of this century, for a variety of reasons, those engaged in both areas of research found themselves enmeshed in population controversies that made them less able to focus their energies on disciplinary development. For instance, when Louis Dublin and Alfred Lotka developed 'intrinsic' vital rates in 1925 that controlled for the age structure's influence on crude birth and death rates, they applied their technique to the study of US vital rates and dramatically announced that the average American woman in 1920 was having only half a child more than was needed to maintain a stationary population. Dublin then was moved to write articles identifying excessive birth control as a national problem and attacking Margaret Sanger for fostering a one-sided socially harmful movement to lower fertility. In the late 1920s, the technical demographer was hard to distinguish from the movement advocate.

Similarly, experts studying the causes and consequences of modern population movements found the 1930s to be an era in which nationalistic chauvinism made discussion of these trends fraught with controversy. The International Union for the Scientific Investigation of Population Problems was launched in 1928, an outgrowth of the 1927 World Population Conference sponsored by Margaret Sanger in Geneva. The IUSIPP (later to become the IUSSP) planned to hold its first meeting in Rome in 1931. Mussolini's rise to power and hints that the meeting would be used to promulgate his racial theories caused the leadership of the IUSIPP to 'disestablish' the Rome meeting and convene a hastily planned counter-conference in London. The IUSIPP's next official meeting was held in Berlin in 1935, and was even more controversial. Several national committees, correctly fearing that the conference would be used as a platform from which to spread Nazi racial theories, boycotted the meeting. The first two IUSIPP meetings were exercises in divisiveness, and certainly not venues for establishing a new discipline.

Even the academic study of the causes of modern population dynamics proved controversial during the interwar period. By the early 1930s a number of students of population (Warren Thompson in the United States, Adolphe Landry in France, and A.M. Carr-Saunders in Great Britain) arrived at a quite similar summation of modern population dynamics: as agrarian societies become industrial there is a shift from high to low vital rates during which mortality declines more quickly than fertility, producing a period of population growth. Northern and Western Europe and the United States had largely experienced this 'demographic revolution' (Landry's term), Eastern and Southern Europe and Japan were in the middle of their expansion stage, and much of the rest of the world had just begun the revolution.

Although demographic transition theory was to play a central role in the institutionalization of demography 20 years later, its development in the 1930s proved awkward. Both Germany and Japan were using *Lebensraum* rationales for their imperialist moves into their neighbors' land, claiming that their growing populations needed room to expand. Any summation of modern population dynamics that seemed to confer scientific legitimacy on these claims lost attractiveness to a number of Western policy makers. Explicating the causes of modern population changes during this era did little to establish a firm foundation on which to build a disciplinary infrastructure for demography.

The end of World War II, however, began a most propitious period for establishing such a foundation. War-torn nations, both allied and axis, entered into a period of state-planned

national reconstruction. The production and analysis of population statistics were vital for such planning. The National Institute of Demographic Studies (INED) was established in France in 1947 as a planning arm of the Ministry of Social Affairs and National Solidarity. Although not engaged in the teaching of demography, INED did much to enhance the reputation of the discipline, as did the publication of the first volume of *Population* (Paris) in 1946. One year later *Population Studies* made its debut in London. The establishment of two scholarly journals that focused exclusively on publishing analyses with a distinct demographic focus did much to solidify the boundaries of demography as an academic subject. On a more fundamental level, the Population Division of the UN Secretariat was established in 1946 and worked to standardize demographic measures and promulgate a common set of techniques for demographic analysis.

The war's conclusion also resolved many of the more fractious issues that had fueled earlier population controversies, and the link between population discourse and divisive controversy weakened considerably. Knowledge of the enormity of the holocaust led to the eugenics movement dissipating. A widespread postwar baby boom dispelled many countries' worries of depopulation. The removal of German and Japanese military threats broke the association between transition accounts of modern population dynamics and population-pressure rationales for imperialism. In fact, in the postwar world the transition framework would provide Western policy makers with a needed policy tool: a way of interpreting the unprecedented demographic changes arising in the world's underdeveloped regions during the 1950s and 1960s.

The rapid decolonization of the world that followed World War II produced a 'Third World' of newly independent governments that became a Cold War battleground where the United States and the Soviet Union fought for supremacy. Unprecedented mortality decline resulting from the use of newly developed antibiotics and the application of effective methods for combating malaria produced rapid population growth throughout much of the developing world. Such growth was especially problematic when considered from the perspective of the postwar versions of demographic transition theory promulgated by US demographers working at Princeton University's Office of Population. If the economic strains associated with rapid population growth prevented the transformation of traditional agrarian societies into modern industrial ones, then rapid population growth might forestall the very socioeconomic changes that would induce fertility decline. Without fertility decline, the Third World's period of population expansion would come to an end with mortality rising as starvation and disease increased. Starvation, economic stagnation, and growing poverty were judged propitious for the spread of communism. Inducing fertility decline in still agrarian societies came to be seen as the only way to humanely resolve the Third World's population dilemma and, additionally, it would protect the geopolitical interests of the Free World.

This vision of the postwar global population situation was promulgated by Princeton demographers during the decade from 1945 to 1955, especially by Frank W. Notestein and Kingsley Davis. During the early 1950s, John D. Rockefeller III and the leadership of the Ford and Rockefeller foundations accepted the validity of this vision and worked to establish a neo-Malthusian movement with a global focus. Their goal was to set up family planning programs throughout the Third World, lower fertility, and lessen population growth. They recognized that only governments could implement effective family planning programs, and their immediate task became converting government leaders, in both developed and Third World countries, to the position that high fertility was a major social problem in need of state intervention. They determined that a dramatic increase in academic research on international population issues was a necessary first step in this conversion process. During the next two decades they expended millions of dollars to develop demographic research centers that focused on international population issues.

These expenditures on demographic research had a profound impact. In the United States in 1950, for example, courses in demography could be found at the graduate level in only three universities. Between 1951 and 1967 major population research centers were established at 16 US universities; all owed their existence to foundation funding, largely from the Ford Foundation. Similar expenditures helped to establish internationally oriented population centers at a number of major universities in Europe and Australia. Funding from the Population Council, a foundation established by John D. Rockefeller III in 1952 to provide a leadership role for the international population control movement, established UN regional centers for demographic training and research in Bombay, India (1957), Santiago, Chile (1958), and Cairo, Egypt (1963).

Additionally, its fellowship program brought hundreds of Third World students to major population research centers in developed countries for graduate training in demography. The research generated by the rapidly expanding number of demographers found its way into print in the many demographic and family planning journals that foundation funds were instrumental in launching: *Studies in Family Planning* (1963), *Demography* (1964), *Family Planning Perspectives* (1969), *Population and Development Review* (1975), and *International Family Planning Perspectives* (1975). This conscious cultivation of demographers engaged in addressing Third World population issues had the desired effect. While no government leader thought that lowering the fertility of couples was a legitimate state function in 1950, a substantial number did so 20 years later. As governments adopted fertility control policies and implemented family planning programs, additional funds for demographic research became available, as did additional employment opportunities for the alumni of population research centers.

The vast expansion in resources made available for demographic research by leaders of the international population control movement from 1950 to 1975 was largely responsible for establishing demography as an academic discipline worldwide. All the prerequisites for an intellectual activity becoming a discipline – a significant interaction among those performing the activity, an organization that directs its investigation and teaching, and entities outside itself that value and use its findings – were deliberately assembled by this leadership. The decision to fund graduate-level demographic training programs at leading Western research universities assured that there

would be a significant and continuing interaction among those engaged in demographic research. It also assured that academic programs in demography would be established and that standards for the teaching of the discipline would be developed. The many private and government contracts for fertility control research generated a demand for this new discipline's findings by outside entities, a demand that no doubt impressed university administrators.

These resource flows, however, did not simply impact the disciplinary development of demography in positive ways. There were costs as well. The neo-Malthusian impetus behind the institutionalization of demography caused the discipline itself to be tinged with an ideological aura. During the 1950s and 1960s, Soviet and Chinese population specialists constantly questioned the objectivity of Western analyses of Third World population issues. The major premise of the international population control movement was that rapid population growth had deleterious effects on development, and this was certainly the premise that guided the vast bulk of orthodox demographic research. It was, however, more hypothesis than fact. Although it is not uncommon among the social sciences to have certain perspectives represented within a discipline that are based on premises not subject to empirical verification, the discipline itself is usually composed of a multiplicity of such perspectives. Except in France, where a pronatalist tradition among French demographers made many suspicious of neo-Malthusianism, significant counter-perspectives did not develop within Western demography until a stream of 'revisionist' literature began to develop in the late 1970s. Until that lively debate over the actual relationship between population growth and economic development developed, demography's close association with the international population control movement worked to politicize it. At occasions, such as the World Population Conference at Bucharest in 1974, demographers found not only the validity of their research findings questioned but also their motives.

The neo-Malthusian impetus for much foundation and government funding of demographic research also made it subject to the shifting fortunes of that movement. Population research centers were established on the basis of grants with time limits and which lacked departmental status. Many of their positions were 'soft money' nontenured ones, although all had core faculty with permanent positions in established departments, usually sociology, economics, or public health. During the 1950s and 1960s few universities were willing to subsidize population research centers internally, and the survival of demographic training programs required the continual influx of outside funds. Foundation support for the international population control movement remained strong during the 1950s and 1960s but declined dramatically during the 1970s.

The end of the Cold War dissipated a powerful source of Northern concern over problems associated with Southern population growth. Northern government support of the movement was still significant but tended to ebb and flow with shifts in domestic political currents. Academic demography responded to these funding uncertainties by becoming leaner, more embedded within the university framework, and less oriented to an international fertility control agenda. In the North, where the discipline still finds its center, every country faced a significant aging of its population and many faced problems associated with long term below replacement fertility. These domestic population concerns increasingly became the disciplinary focus of Northern demographers and population research centers. A similar transformation occurred in many Southern countries where fertility was falling to replacement levels and below. Only in countries with still high fertility levels did demographic research remain synonymous with fertility reduction research. The one population trend that attracted the attention of demographers in all regions was international migration. A globalizing economy accompanied by great disparities in national wealth was causing a persistent mass movement from the South to the North.

As the twenty-first century commenced, the most notable development in the field of population was that a discipline of demography much like that envisioned by Achille Guillard in 1855 finally arrived. The study of the causes and consequences of population dynamics had become a widely recognized intellectual activity that was investigated and taught within an academic setting by individuals with academic credentials who were paid for their specialized knowledge and who reproduced themselves by producing a new generation of demographers. The institutionalization of this discipline in the second half of the twentieth century was largely the result of a confluence of events happening at mid-century, the most notable of which was the rise of an international population control movement. Sufficient resources were available for a long enough period to establish an academic discipline that became self-sustaining.

Twenty-first century demography, however, still lacks some characteristics of a fully mature discipline. Its disciplinary cohesiveness arises more from having a well-defined subject matter, a distinctive set of data sources, and a focus on common policy issues rather than from having a unifying theory capable of linking societal and individual demographic behavior. Because most populations have completed or are near completing their mortality and fertility transitions they experience common trends: population aging in nearly every region; below replacement fertility in much of the North; declining population growth in much of the South. For the moment these common trends have produced a policy environment with few ideological divides and the discipline's immediate twenty-first century prospects appear secure.

New disciplinary challenges, however, can be seen on the horizon. Demography traditionally has used its distinctive data sources and analytical tools to define itself. Those who analyze census data and vital statistics using projection modeling or life table techniques are 'demographers.' Countries, however, are increasingly using surveys, not censuses, to generate demographic data. Studies of contemporary changes in sexual behavior, marriage, fertility, household composition, morbidity, mortality, education, income, and employment increasingly analyze household survey data using standard statistical techniques. The boundaries separating demography from economics, sociology, and other social sciences are blurring as more of these studies are undertaken by individuals who do not self-identify as 'demographers.' With theory playing such a limited role in defining demography, it is unclear if a distinct discipline will remain as its data and methods become less unique.

It is also unclear how long the tranquil ideological environment now surrounding most contemporary demographic research will remain in place. The discipline evolved in a funding environment in which population specialists focused on the demographic trends that troubled Western policy makers and devised policy interventions aimed at solving their population problems. With much of the world proceeding on a course of demographic change similar to that initially experienced by the West, the discipline's general policy relevance gradually expanded to encompass wider territories. But the economic and demographic convergence occurring in the twenty-first century is already giving birth to a multipolar political world in which countries and regions are developing distinctly different and oftentimes competing population interests. The OECD estimated that the US, Japan, and Europe produced 37% of the world's GDP in 2011, while India and China produced 24%. It projects that by 2060 the US, Japan, and Europe's share of the world's GDP will drop to 28%, whereas India's and China's share will increase to 46%. Less developed economies are growing faster than more developed economies, and by century's end their average resident is predicted to have a higher per capita income than is found today in more developed economies. A new economic pyramid is already emerging and significant shifts in countries' economic rankings are likely to become more noticeable as the century progresses.

Demographic changes are only magnifying the significance of the economic rise of the South. The UN estimates that the 'more developed' world contained 18% of the world's population in 2011 but predicts that its share will fall to 13% by 2060 as Europe's share slips from 11 to 7% while Africa's share increases from 15 to 28%. Although the UN foresees a general global convergence of fertility by century's end to around two children per woman, since the South is experiencing its fertility transition at a later time, the result will be a future world in which Southern citizens constitute the great majority of the population. In the twentieth century the overwhelming economic dominance of the West lessened the importance of relative human numbers. A country in which a large majority of citizens were impoverished had little political clout even if its population was quite large. But the economic and demographic rise of the South in the twenty-first century is producing a new political environment in which democratic ideals and political clout are much more closely aligned. In a world that privileges democratic norms and in which each individual's economic weight is becoming more equal, relative human numbers assume a greater political significance.

The long-term prospects of demography, therefore, will depend on how skillfully it can negotiate the new multipolar political world that is emerging as the North's demographic and economic importance recedes. Demography always has had to prove itself in terms of its policy relevance, but it no longer has an easily defined principal client. If pursuit of national interests simply is assumed to be demography's raison d'être, with US demographers working to help craft the best migration policies for the United States while Mexican demographers do the same for Mexico, then population policy conflicts will become the norm and demography will lose some of its scientific luster. In truth, the new physical world that is emerging as a result of twenty-first century economic and demographic changes requires a new policy focus for the discipline. An enlarged human population consuming natural resources at increasing per capita rates has already generated disruptive climate changes that recognize no national boundaries. Shortages of vital resources, especially of water, appear likely as the century progresses. If demography is to have the capability of offering humankind a way of negotiating the population/environment policy challenges of the twenty-first century, then it needs to adopt the world as its principal client. Then demography might be the source of objective policies to move populations away from rising coastlines or to resettle those who find themselves living in increasingly inhospitable areas. It is not clear that demography as presently constituted is equipped for such tasks.

Editor's Note

Plausible cross-references for this article include virtually the entire Section on Demography, a number of entries in Sociology, and even quite a number of entries in Environmental and Ecological Sciences. From this rich array, the most directly pertinent cross-references are listed in *See also* section below.

See also: Age Structure; Carrying Capacity of the Environment; Censuses: History and Methods; Data Bases and Statistical Systems: Demography; Demographic Measurement: General Issues and Measures of Fertility; Demographic Models; Demography, Early History of; Eugenics as an International Movement; Eugenics, History of; Families and Households, Formal Demography of; Family Planning: Program Effects; Fertility Theory; Galton, Sir Francis (1822–1911); Historical Demography; Malthus, Thomas Robert (1766–1834); Population Dynamics: Mathematic Models of Population, Development, and Natural Resources.

Bibliography

Barrett, D., 1995. Reproducing Persons as a Global Concern: The Making of an Institution (Ph.D. Thesis). Stanford University.

Caldwell, J.C., 3 November 1996. Demography and social science. Population Studies 50, 305–333.

Caldwell, J.C., Caldwell, P., 1986. Limiting Population Growth and the Ford Foundation Contribution. Pinter, London.

Chasteland, J.-C., Roussel, L. (Eds.), 1997. Les Contours de la Demographie: Au Seuil du XXIe Siècle. Institut National d'Etudes Demographiques, Paris.

Davis, K., 1945. The world demographic transition. Annals of the American Academy of Political and Social Science 237, 1–11.

Demeny, P., 3 September 1988. Social science and population policy. Population and Development Review 14, 451–479.

Donaldson, P.J., 1990. Nature against Us: The United States and the World Population Crisis. University of North Carolina Press, Chapel Hill, NC.

Finkle, J.L., Crane, B.B., 1975. The politics of Bucharest: population, development, and the new international economic order. Population and Development Review 1 (1), 87–114.

Greenhalgh, S., 1 January 1996. The social construction of population science: an intellectual, institutional, and political history of twentieth-century demography. Comparative Studies in Society and History 38, 26–66.

Harkavy, O., 1995. Curbing Population Growth: An Insider's Perspective on the Population Movement. Plenum Press, New York.

Hauser, P.M., Duncan, O.D. (Eds.), 1959. The Study of Population: An Inventory and Appraisal. University of Chicago Press, Chicago.

Hodgson, D., 4 December 1988. Orthodoxy and revisionism in American demography. Population and Development Review 14, 541–569.

Lorimer, F., 4 October 1971. The role of the international union for the scientific study of population. Milbank Memorial Fund Quarterly 49, 86–96.
McNicoll, G., 3 September 1992. The agenda of population studies: a commentary and complaint. Population and Development Review 18, 399–420.
Notestein, F.W., 1971. Reminiscences: the role of foundations, of the Population Association of America, Princeton University and the United Nations in fostering American interest in population problems. Milbank Memorial Fund Quarterly 49 (4), 67–84.
OECD, 2012. Looking to 2060: Long-Term Global Growth Prospects. OECD. Economic Policy Paper Series No. 3.
Preston, S.H., 4 November 1993. The contours of demography: estimates and projections. Demography 30, 593–606.
Sauvy, A., 1966. Théorie Générale de la Population. Presses Universitaires de France, Paris. General Theory of Population. Basic Books, New York, (1969).
Shils, E., 4 Fall 1970. Tradition, ecology, and institution in the history of sociology. Daedalus 99, 760–825.
Szreter, S., 4 December 1993. The idea of demographic transition and the study of fertility change: a critical intellectual history. Population and Development Review 19, 659–701.
United Nations, Department of Economic and Social Affairs, Population Division, 2013. World Population Prospects: The 2012 Revision.

Dependency in Language

Igor Mel'čuk, University of Montreal, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The notion of linguistic dependency is introduced with three major types: semantic, syntactic, and morphological dependencies. Semantic dependency [= Sem-D] is a relation between a predicate semanteme 'σ_1' and its argument semanteme 'σ_2' or between two lexemes L('σ_1') and L('σ_2'). Syntactic dependency [= Synt-D] is a relation between two lexemes defined by three groups of formal criteria: (A) Criterion for the *presence* of Synt-D; (B) Criteria for the *direction* of Synt-D; and (C) Criteria for the *type* of Synt-D. Morphological dependency [= Morph-D] is a relation between two lexemes defined by the imposition of a grammeme by one lexeme upon another; this dependency type appears either as agreement or as government. The three types of dependency can combine, producing 14 possible combinations of dependencies between two given lexemes in an utterance. Dependency and constituency are cursorily compared; several linguistic phenomena whose description is more conveniently done in terms of dependency are considered (valence and voice, lexical functions, phraseology, paraphrasing, word order).

Abbreviations and Notations

-A Actant
C Constituency
CAT Particular inflectional category (inflectional)
D- Deep
D Dependency
DirO Direct Object
L Particular language
L Particular lexeme
LF Lexical function
Morph- Morphological
r Particular syntactic relation
-Rel Relation
-S Structure
S- Surface
Sem- Semantic
Synt- Syntactical
W Wordform
W(L) Wordform of lexeme L

Notion of Linguistic Dependency

Consider an English Sentence, for Instance, (1):

(1) *The boys care for their mother*

As any sentence in any language, it is organized by syntagmatic relations that hold between its words and are involved in (i) arranging words in linear strings, (ii) supplying them with appropriate prosody, and (iii) giving them a necessary morphological form. (Syntagmatic relations are opposed to paradigmatic relations, holding between words in the language – not in a sentence – and involved in choosing semantically/syntactically appropriate words to be used in sentences.)

The main type of syntagmatic relation in any language is dependency: in a sentence of language L, lexical unit L_2 (syntactically) depends on lexical unit L_1 if and only if L_1 and L_2 can form a phrase of L and L_1 determines L_2's presence and linear position in the sentence under examination (for details, see Section Syntactic Dependency [= Synt-D], Criterion A); we also say that L_1 is the syntactic governor of L_2. Thus, in (1), the noun *boys* depends on the Main Verb [= MV] *care*: *boys care* is a phrase of English, and, as the syntactic subject, *boys* must precede the MV; *for* also depends on *care*: if MV were *thinks* or *likes*, the preposition should be *of*, respectively, nothing, and *for* must follow *care*; *mother* depends on *for*, since *mother* is positioned with respect to it; etc. Dependency D is symbolized with an arrow pointing from the governor to the dependent: *boys*←*care*, *care*→*for*, *for*→*mother*. D is an antireflexive and antisymmetrical binary relation (= 'X depends on Y'):

- no L can depend on itself: $\neg L \leftarrow L$ (antireflexivity);
- if L_2 depends on L_1, then L_1 cannot depend on L_2: $L_1 \rightarrow L_2$ entails $\neg L_2 \leftarrow L_1$ (antisymmetry).

Transitivity of D varies for different types of dependency; it will be characterized later.

Some syntagmatic relations in language are not dependencies, e.g., coreferentiality: in (1), *boys* and *their* have the same referent; this relation is reflexive, symmetrical, and transitive.

Three Major Types of Linguistic Dependency

Speaking about dependency, people often mean just syntactic dependency. Thus, the sketchy analysis of example (1) presented above illustrates syntactic dependency. However, there are three different types of linguistic dependency: semantic, syntactic, and morphological dependencies (Mel'čuk 1981, 1988: 105ff., 2009: 8–66).

Semantic Dependency [= Sem-D]

The lexeme L_2 semantically depends on the lexeme L_1 in the semantic structure of a sentence, i.e., L_1–**sem**→L_2, if and only if [= iff] the meaning of L_1 is a predicate and the meaning of L_2 is an argument of it; in other words, L_1–**sem**→L_2 means '$L_1(L_2)$.'

A Sem-dependent of a lexeme is called its semantic actant [= SemA] (for actants, see Mel'čuk 2004a,b).

Properties of Semantic Dependency

Like all Ds, Sem-D is antireflexive; like Synt-D, it is antisymmetrical. As for transitivity, Sem-D is neither transitive nor antitransitive: in most cases, L_1–**sem**→L_2 and L_2–**sem**→L_3 entail neither L_1–**sem**→L_3 nor ¬(L_1–**sem**→L_3). Figure 1 shows possible transitivity of Sem-D, and Figure 2, its possible antitransitivity.

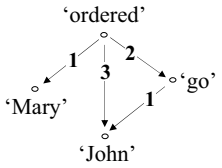

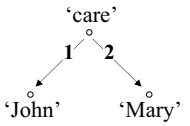

The structural words TO and FOR are not represented in the SemS since they are semantically empty – their role is to mark syntacic links, not carry meaning.

Figure 1 The semantic structure of the sentence *Mary ordered John to go*

Figure 2 The semantic structure of the sentence *John cares for Mary*

Sem-Ds are labeled: a Sem-D arc carries a distinguishing symbol identifying the corresponding Sem-actant.

Sem-D does not presuppose the uniqueness of the governor: a lexeme can semantically depend simultaneously on more than one lexeme, cf Figure 1: 'John.'

Sem-D is linguistically universal: it is present in all languages, appears in all sentences of a language, and embraces all full lexemes of a sentence. In a sentence, Sem-Ds form a connected structure (there is a Sem-D 'path' between any two lexemes of the sentence); the semantic structure [= SemS] is a network.

Syntactic Dependency [= Synt-D]

Sem-D is mentally perceived by the speakers, and Morph-D (see below) is perceived physically; but Synt-D, which mediates between them, is not directly perceivable. It is an abstract construct that must be specified by some formal criteria.

Criteria for Synt-D

To establish a Synt-D relation between lexemes L_1 and L_2 in a sentence, three groups of criteria are needed – namely, A: Criterion for Synt-connectedness (= for the presence of a Synt-D); B1–3: Criteria for the Synt-dominance (= for the orientation of the Synt-D); and C1–3: Criteria for the specific type of the Synt-D.

Criterion A: Synt-Connectedness

For L_1 and L_2 to be syntactically directly linked (that is, for there to be L_1–**synt**–L_2), Criterion A must be satisfied.

Potential Prosodic Unity and Linear Disposition

In a given sentence of L, lexemes L_1 and L_2 have a direct Synt-D link iff both conditions 1 and 2 are satisfied:

1. a. **General case**: Either L_1 or L_2 can form a phrase in L.
   b. **Special case**: Or L_1 and L_2 cannot form a phrase in L, but L_1, L_2, and L_3 can, such that in the phrase L_1–L_2–L_3 the Synt-head is L_1 and L_2 and L_3 also form a phrase in which the Synt-head is L_2 (for the concept of Synt-head, see immediately below).
2. The linear position of one of the lexemes, L_1 and L_2 or, more precisely, of the corresponding wordforms, w(L_1) and w(L_2) must be specified with respect to the other: w(L_2) either precedes or follows w(L_1), or else can precede or follow it – freely or under some conditions.

Examples

General case: N + V (*John snores*), V + ADJ (*was sick*), V + N (*see John*), PREP + N (*for John*), ADV + ADJ (*extremely intelligent*), etc.

Special case: For example, *escape* [L_1] *from* [L_2] *problems* [L_3], where *escape from problems* is a phrase of English, with *escape* being its Synt-head; *from problems* is also a phrase in which *from* is the Synt-head; therefore, *escape* and *from* have – within the phrase *escape from problems* – a direct syntactic link: *escape*–**synt**–*from*. Note that mentioning (in Criterion A) the syntactic head, which was not yet defined, does not violate logical consistency: first, the head in the phrases L_1–L_2–L_3 and L_2–L_3 is established in due course, and only then are these phrases used to establish the prosodic unity for L_1–L_2.

Criteria B1–3: Synt-Dominance

In a phrase L_1–**synt**–L_2, one of the two lexemes is the Synt-governor of the other (and the Synt-head of the phrase): the Synt-governor is the lexeme that determines, to a greater extent than the other (its Synt-dependent), the syntactic, morphological, and semantic properties of the phrase.

Criterion B1: The Passive Synt-Valence of a Phrase

The passive Synt-valence of an expression *X* is a set of syntactic constructions of L in which *X* can appear as an addition to the corresponding construction. To put it differently, you take a correct utterance of L and add *X* to it, so that the correctness of the utterance is not violated and its meaning is not deformed, but enriched. For instance, an ADJ(ective) such as *white* can be added to a N(oun), say, *white leaf*, or to a copula verb, producing *is/seems white*; the passive valence of an ADJ is a set including N, $V_{(copula)}$, etc. The passive valence of a phrase ADJ + N is that of the N, not that of the ADJ, since this phrase can be used in all the contexts where an N can, but not in all the contexts where an ADJ is usable.

In a phrase L_1–**synt**–L_2, the lexeme L_1 is the Synt-governor of L_2 if the passive valence of the whole phrase is determined to a greater extent by the passive valence of L_1 than by that of L_2. L_1 is also called Synt-head of the phrase.

The Synt-head of a phrase determines the possible external syntactic links of the phrase more than any other of its wordforms.

(2) a. The passive valence of the phrase *for decades* is fully determined by the preposition (= it can be used where this preposition is allowed); therefore, *for*–**synt**→*decades*.
   b. A phrase like *has escaped* or *does not escape* plays the same Synt-role as *has/does* rather than that of *escaped* or *escape*; therefore, *has*–**synt**→*escaped*, *does*–**synt**→*escape*.

In case L_1 and L_2 have the same passive syntactic valence, Criterion B1 does not supply an answer, so that Criterion B2 has to be used.

Criterion B2: Morphological Links between the Lexemes of a Phrase and Its External Context

In a phrase L_1–**synt**–L_2 (in which the passive Synt-valence does not allow us to establish the Synt-governor), L_1 is the Synt-governor of L_2 if L_1 controls the inflection of lexemes external to the phrase or if L_1's own inflection is controlled by such lexemes.

(3) The phrase Rus. *jubka-štany* lit. 'skirt-pants' does not allow for the application of Criterion B1; but Criterion B2 singles out *jubka* as the Synt-governor: *èt* + ***a*** FEM.SG *jubka-štany byl* + ***a*** FEM.SG ... 'this skirt-pants was...,' where the agreement is with *jubka* FEM.SG, and not with *štany* PL ⟨**èt* + ***i*** *jubka-štany byl* + ***i***...⟩; therefore, *jubka*–**synt**→*štany*.

In case L_1 and L_2 have the same morphological properties, Criterion B2 does not supply an answer and Criterion B3 has to be used.

Criterion B3: The Semantic Content of the Phrase

In the phrase L_1–**synt**–L_2 (in which neither the Synt-valence nor the morphology allow us to establish the Synt-governor), L_1 is the Synt-governor of L_2 if the denotation of L_1–**synt**–L_2 is a subset of the denotation of L_1 rather than a subset of the denotation of L_2.

By denotation of L is understood the set of all its possible referents. In other words, L_1–**synt**–L_2 means 'a kind/an instance of L_1,' and not 'a kind/an instance of L_2.'

(4) a. In the phrase *jam sandwich*, the Synt-governor is *sandwich*, because "*jam sandwich* refers to a kind of sandwich, rather than to a kind of jam" (Hudson, 1990).

b. In the Mandarin Chinese phrase [*shí*] *bàng ròu* '[ten] pounds [of] meat,' *ròu* 'meat' is the Synt-governor: [*shí*] *bàng ròu* refers to meat, not pounds; therefore,

[*shí*←**synt**–]*bàng*←**synt**–*ròu*.

Both lexemes *bàng* and *ròu* are nouns – that is, they have the same passive syntactic valence, and Chinese has no inflection.

The Synt-governor is thus the syntactic class determinant, or (if there is no syntactic determinant) the morphological behavior determinant, or (in case both syntactic and morphological determinants are absent) the semantic content, i.e., denotation, determinant (1993: 295-296).

Although most approaches dealing with Synt-Ds require concord between these three properties, i.e., between Criteria B1–3, such a concord should not be expected. Only Criterion B1 is genuinely syntactic; B2 is morphological; and B3 semantic. It is well known that orientations of Sem-D, Synt-D, and Morph-D between two wordforms can differ (cf Section Possible Combinations of the Three Types of Linguistic Dependency below); therefore, these criteria are often in conflict. Criteria B1–3 form a hierarchy: B1 > B2 > B3. If B1 is applicable, its indication is sufficient. Only if it is not applicable, B2 applies – in L having inflection and for L_1 and L_2 with different morphological properties. Otherwise, B3 applies. Therefore, Criteria B1–3 are never applied simultaneously and thus cannot contradict each other.

For Criteria B to be satisfied, at least one of Criteria B1–3 must be satisfied.

The criteria for the orientation of Synt-D ('Head vs Dependent' problem) are discussed in Zwicky (1985, 1993), Hudson (1987, 1990: 106–107), Mel'čuk (1988: 132ff., 2009: 25–40), and Corbett et al. (1993).

Criteria C1–3: Labeled Synt-Dependencies

For each phrase L_1–**synt**→L_2, the type of the Synt-D relation must be established, because to represent the SyntSs of sentences, different types of Synt-D are needed. Thus, I←**synt**–LOVE–**synt**→YOU does not distinguish between 'I love you' and 'You love me.' In all such cases, different types of Synt-D, or different syntactic relations [= SyntRels], are needed: I←$\mathbf{r_1}$–LOVE–$\mathbf{r_2}$→YOU, $\mathbf{r_1}$ = **subjectival**, and $\mathbf{r_2}$ = **direct-objectival**. (Recall that the nodes of a SyntS are not linearly ordered and the markers of syntactic roles – such as grammatical cases – are not shown in it, exactly because word order and cases are **exponents** of Synt-Ds, which a SyntS represents.)

SyntRels are labeled, and the label **r** of a SyntRel denotes a family of syntactic constructions that implement the SyntRel **r**.

NB: As shown in Section Deep- versus Surface-Syntactic Structure and Deep- versus Surface-Syntactic Relation below, two syntactic levels are distinguished: a deep-syntactic and a surface-syntactic level, so that in fact there are two corresponding families of Synt-Ds – namely, deep- and surface-syntactic dependencies. To keep things as simple as possible, this section considers only the surface-syntactic dependencies.

Thus, the SSynt-label «**subjectival**» just mentioned identifies a set of SSynt-rules that (in English) linearly position the noun L_2 with respect to the verb L_1 – in a neutral context, before it – and make L_1 agree with L_2 in person and number (L_2 is of course the syntactic subject).

In the phrases L_1–**r[?]**→L_2, the SSynt-D that links the two wordforms can be labeled **r** only if it satisfies all of the following criteria C1–3; if one of them is not satisfied, the presumed SyntRel **r[?]** must be split into different SSyntRels $\mathbf{r_1}$ and $\mathbf{r_2}$.

Criterion C1: Absence of Semantic Contrast [w(L) Stands for 'a Wordform of Lexeme L']

A SSyntRel must not describe two phrases $\mathbf{w_i}(L_1)$–**synt**→$\mathbf{w_j}(L_2)$ and $\mathbf{w_m}(L_1)$–**synt**→$\mathbf{w_n}(L_2)$, which:

i. contrast semantically – that is, '$\mathbf{w_i}(L_1)$–**synt**→$\mathbf{w_j}(L_2)$' ≠ '$\mathbf{w_m}(L_1)$–**synt**→$\mathbf{w_n}(L_2)$' and

ii. differ formally by purely syntactic means of expression, i.e., only by word order, syntactic prosody, or syntactic grammemes.

(5) The Russian construction PJAT'←**r[?]**–DOLLAR has two implementations with different meanings: *pjat' dollarov* '5 dollars' versus *dollarov pjat'* 'approximately 5 dollars';

the formal difference here is purely syntactic: word order; therefore, the presumed SSyntRel **r[?]** must be split into two different SSyntRels:

PJAT′ ← **quantitative** – DOLLAR ⇔ *pjat′ dollarov*
PJAT′ ← **approx-quantit** – DOLLAR ⇔ *dollarov pjat′*

Criterion C1 corresponds to the Minimal Pair Test, widely used in linguistics.

Criterion C2 deals with the substitutability of syntactic subtrees. The use of this property as a criterion for establishing SyntRels was proposed in Kunze (1975: 235ff.); hence the name: the Kunze property. Here, however, a weaker version of it is used: the quasi-Kunze property.

The quasi-Kunze property [$\Delta_{(X)}$ is a subtree with the top node of the syntactic class X]:

A SSyntRel **r** has the quasi-Kunze property iff there exists in L a syntactic class X such that for any SSynt-configuration L-**r**→$\Delta_{(Y)}$, replacing $\Delta_{(Y)}$ by $\Delta_{(X)}$ (but not necessarily vice versa!) in any SSynt-structure does not affect its syntactic well-formedness.

In other words, for a SSyntRel **r** having the quasi-Kunze property it is always possible to indicate such a prototypical dependent that it will be admitted by any eventual governor.

Criterion C2: Syntactic Substitutability

Any SSyntRel **r** in L must possess the quasi-Kunze property.

(6) In the phrases *have*–**r[?]**→*been* and *be*–**r[?]**→*going* the presumed SSyntRel **r[?]** does not possess the quasi-Kunze property:
have–**perfect-auxiliary**→*been* and *be*–**progressive-aux**→*going*, but not
have*–synt**→*going* or **be*–**synt**→ *been*.

Criterion C2 is a particular case of the Substitution Test: it deals with the substitution of subtrees that hang on the same SSyntRel **r**.

Criterion C3 takes into account the fact that a SSyntRel of language L can be either nonrepeatable or unlimitedly repeatable.

Nonrepeatable SSyntRels

A SSyntRel **r** is nonrepeatable iff no more than one branch labeled **r** can start from any governor.

Actantial SSyntRels, which describe the Subjects, the Objects, and the Complements, are nonrepeatable. (In some languages, an actant can/must be 'duplicated' by a resumptive pronoun, but such duplication is not considered as repeatability: it has a completely grammaticalized character.)

Unlimitedly Repeatable SSyntRels

A SSyntRel **r** is (unlimitedly) repeatable if and only if several branches labeled **r** can start from a governor.

For a repeatable SSyntRel **r**, the number of branches labeled **r** starting from a governor is theoretically unlimited, although practically it is, of course, limited by pragmatic considerations or the properties of governors. For instance, the **modificative** and **circumstantial** SSyntRels in English are unlimitedly repeatable.

Criterion C3: Repeatability with the Same Synt-Governor

Any SSyntRel **r** must be either unlimitedly repeatable or non-repeatable – that is, it cannot be limitedly repeatable.

(7) In Persian, expressions of the following type are extremely widespread: *Raminra bedar kärd* '[He/She] awoke Ramin' lit. '[He/She] Ramin awakening made.' This sentence has the Synt-structure as follows:
Ramin+ra ← **r[?]**–*kärd*–**r[?]**→*bedar*
DirO made awakening [Noun]

A verbal collocation of the type *bedar kärd* 'awakening make' = 'awake' or *därs däd* lit. 'lesson give' = 'teach' seems to contain a DirO (*bedar* or *därs*), but behaves as a normal transitive verb and takes a 'genuine' DirO (with the suffix **-ra**). The presumed SSyntRel **r[?]** here would be limitedly repeatable – just twice. Therefore, there are two different SSyntRels:

RAMIN ← **dir-obj**–KÄRD–**quasi-dir-obj** → BEDAR.

The nominal base of the collocation in question is a quasi-direct object.

Criterion C3 corresponds to the Cooccurrence Test: an element of a grammatical category is

- either nonrepeatable, like tense or number in English;
- or unlimitedly repeatable, like the causative in Turkish: *uyu-* 'sleep' ~ *uyu-**t**-*'cause to sleep' (e.g., by rocking) ~ *uyu-**t-dur**-* 'cause to cause to sleep' (e.g., by telling somebody to make the baby sleep') ~ *uyu-**t-dur-t**-* 'cause to cause to cause to sleep' (e.g., by reminding a person to tell somebody to make the baby sleep).

Notion of Synt-D and Dependency Tree as the Formalism for Representing Syntactic Structures

Now the definition of SSynt-D can be formulated.

Surface-Syntactic Dependency [= Synt-D]

In a given sentence, the lexeme L_2 syntactically depends on the lexeme L_1 via SSyntRel **r**, i.e., L_1–**r**→L_2, iff Criteria A, B, and C are satisfied for this pair of lexemes and **r**.

The lexeme L's SSynt-dependent that satisfies some particular conditions (roughly speaking, it corresponds to L's SemA), is L's SSynt-actant [= SSyntA] (Mel'čuk, 2004b).

For a different definition of Synt-D, see Gerdes and Kahane (2011) (where, among other things, dependency is generalized to represent the structure of discourses as well as the internal structure of wordforms).

Properties of Syntactic Dependency

Like Sem-D, Synt-D is antireflexive and antisymmetrical: ¬(L_1–**synt**→L_1); L_1–**synt**→L_2 entails ¬(L_1 ← **synt**–L_2). It is also antitransitive: within one sentence, L_1–**synt**→L_2 and L_2–**synt**→L_3 entail ¬(L_1–**synt**→L_3). Thus, Synt-D presupposes the uniqueness of the governor: a wordform either can syntactically depend on only one other wordform or be syntactically independent, as is the top node of a syntactic structure [= SyntS]. This requirement is based on a simple consideration: Synt-D ensures linear ordering of lexemes in the sentence, and a lexeme can be ordered – before or after – with respect to only one other lexeme.

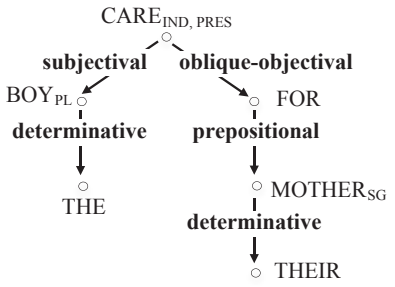

Figure 3 Surface-syntactic structure of sentence (1).

Synt-Ds are labeled: to properly represent *Mary loves John*, in the phrases *Mary*←$\mathbf{r_1}$–*love* and *John*←$\mathbf{r_2}$–*love* two different SSyntRels, $\mathbf{r_1} \neq \mathbf{r_2}$, are needed.

Synt-D is linguistically universal: it is present in all languages, appears in all sentences of a language, and embraces all wordforms of a sentence. In a sentence, Synt-Ds form a connected structure, which formally is a dependency tree: a connected graph where (i) each node, except one, depends on only one node (= the uniqueness of the Synt-governor) and (ii) one and only one node does not depend on anything (= the uniqueness of the top node, or the root, of the tree).

Thus, SyntSs of sentences are described by dependency trees. The linear order of the nodes in a SyntS is not defined; the D-description consistently separates the Synt-links between wordforms and their linear order. Elements of a D-tree are labeled: nodes with the names of lexical units supplied with grammemes, and branches with the names of Synt-relations – that is, particular Synt-Ds. See the Surface-Syntactic Structure of sentence (1) in Figure 3:

Deep- versus Surface-Syntactic Structure and Deep- versus Surface-Syntactic Relations

At the syntactic level of linguistic representation (as well as at the morphological and phonic levels), two sublevels are distinguished: Deep-Syntactic [= DSynt-] versus Surface-Syntactic [= SSynt-]. DSynt-structure is geared to meaning: it reflects explicitly all syntactically relevant semantic distinctions and ensures maximally smooth transition between the SemS and itself; SSynt-structure is oriented toward the physical shape of the sentence: it presents explicitly all formal contrasts and ensures smooth transition between itself and the linearly ordered Deep-Morphological structure of the sentence. Both Synt-structures are D-trees differing by the labeling of nodes and branches: deep lexical units versus surface lexical units with corresponding grammemes; DSyntRels (language-universal, a dozen) versus SSyntRels (language-specific, a few dozen).

Deep-Syntactic Relations

A DSynt-relation is a generalization over several actual syntactic constructions found in natural languages; all DSyntRels taken together are expected to cover all syntactic constructions known at present. The set of DSyntRels is determined by the following five DSynt-oppositions:

1. *Coordination versus Subordination*: coordinate versus subordinate constructions; the former are represented by the DSyntRel COORD [*honey*–COORD→*and cinnamon*] and QUASI-COORD [*yesterday*–QUASI-COORD→*at 3 o'clock*]. (The QUASI-COORD DSynt Rel represent several constructions that formally resemble coordination: *in Spain, on the South Coast, in a small village; one summer, in June; 10 dollars* (*and*) *55 cents.*)
2. *Weak versus Strong Subordination*: constructions with weak versus strong structural links between sentence elements; the former are represented by the DSyntRel APPEND(itive) [*He is,*–APPEND→*unfortunately, absent*].
3. *Modification versus Complementation*: modification is a type of Synt-D such that in the phrase L_1–**synt**→L_2 the Sem-D goes in the opposite direction: 'L_1←**sem**–L_2'; complementation is a type of Synt-D such that in the phrase L_1–**synt**→L_2 the Sem-D goes in the same direction: 'L_1–**sem**→L_2.' The DSyntRels of the first class are ATTR(ibutive) [*He works*–ATTR→*hard*]; the DSyntRels of the second class are actantial.
4. *Restrictive versus Descriptive Modification*: the former is represented by the DSyntRel ATTR (which by default is understood as restrictive): *He reads only* ***interesting*** ⟨***Spanish***⟩ *books*; the latter is represented by the DSyntRel $\text{ATTR}_{\text{descr}}$: *These three students,* ***who just returned from Europe****, were selected to represent the department.*
5. *Different Actantial Roles*: **I**, **II**, …, **VI**, $\mathbf{II}_{\text{dir.sp}}$. Constructions with actantial DSyntRels are divided into seven classes: the empirically determined maximal number of DSyntAs that a lexical unit in natural language can have is six, plus a special DSyntRel for Direct Speech, as in '*Never*!'←$\mathbf{II}_{\text{dir.sp}}$–*whispered*–**I**→*Alan*.

The full inventory of DSyntRels is represented in Figure 4:

Surface-Syntactic Relations

In sharp contrast to DSyntRels, Surface-Syntactic relations describe real syntactic constructions of a particular language; they have to be empirically established for each language, just like its phoneme inventory or the set of its grammatical cases. The methodology is, generally speaking, the same as that used for establishing any linguistic units (Iordanskaja and Mel'čuk, 2009): (i) determining a list of relevant syntactic properties of lexical units for the given language L (such as being obligatory in any clause, being combinable with a finite verb only, being cleftable, being relativizable, being dislocatable, being the

coordinate DSyntRels		subordinate DSyntRels									
		weak subordinate DSyntRel	strong subordinate DSyntRels								
			modification: attributive DSyntRels		complementation: actantial DSyntRels						
COORD	QUASI-COORD	APPEND	ATTR	$\text{ATTR}_{\text{descr}}$	I	II	III	IV	V	VI	$\text{II}_{\text{dir.sp}}$
1	2	3	4	5	6	7	8	9	10	11	12

Figure 4 Language-universal inventory of DSynt-relations.

controller of agreement, being the controller of reflexivization, being sensitive to negation, etc.); (ii) partitioning all syntactic constructions of L into families according to the similarity of their relevant syntactic properties; (iii) associating each family thus obtained to a SSyntRel; (iv) checking the set of established SSyntRels against the above criteria; (v) cyclically repeating the procedure – modifying the starting list of properties, changing the distribution of constructions between families, etc., until an acceptable set of SSyntRels is obtained.

Here is a tentative list of 54 SSyntRels of English, compiled on the basis of Mel'čuk and Pertsov (1987: 85–156) and Mel'čuk (2009: 52–58), and also taking into consideration Apresjan et al. (1989: 71–121).

English Surface-Syntactic Relations

Square brackets isolate the fragments of the text that should be ignored to clearly see the dependency arrow; the symbol '|' separates full sentences.

I. Subordinate SSyntRels: 1–52

Clause-Level (= Clausal) SSyntRels: 1–22

These SSyntRels link between themselves the elements of the sentence.

Valence-controlled SSyntRels: Complementation

Actantial SSyntRels

1. Subjectival:
   I ← **subj**–*am old*; ***Intervention*** ← **subj**–*seems out of the question*. | ***Smoking*** ← **subj**–*is dangerous*.
   That ← **subj**–[*Alan can do that*]–*is clear*.
   | ***It*** ← **subj**–*is clear that Alan can do that*.
2. Quasi-subjectival:
   It ← **subj**–*is*–[*clear*]–**quasi-subj** → ***that*** *Alan can do that*.
3. Direct-objectival:
   sees–**dir-obj** → ***me***; *to have written*–**dir-obj** → ***novels***; *Helen wants*–**dir-obj** → ***Alan*** *to read*.
   worth–[*a*]–**dir-obj** → ***trip***; *prefer*–[*her*]–**dir-obj** → ***staying*** *home*
   explain–[*to him*]–**dir-obj** → ***that*** *Alan was absent*
   make–**dir-obj** → ***it*** *possible to neutralize the consequen-ces*.
4. Quasi-direct-objectival:
   make–[*it possible*]–**quasi-dir-obj** → ***to*** *neutralize the consequen-ces*.
5. Indirect-objectival:
   gives–**indir-obj** → ***Alan/him*** *some money*
   gives–[*some money*]–**indir-obj** → ***to*** *Alan/him*
6. Oblique-objectival [= prepositional-objectival]:
   depends–**obl-obj** → ***on*** *Alan*
   my respect–**obl-obj** → ***for*** *Alan*
   translation–**obl-obj** → ***from*** *Lushootseed into Polish*
   translation–[*from Lushootseed*]–**obl-obj** → ***into*** *Polish*
   convince–[*Alan*]–**obl-obj** → ***that*** *he should work less*
   This is an easy–[*book*]–**obl-obj** → ***to*** *translate*.
7. Infinitival-objectival:
   can–**inf-obj** → ***read***; *want*–**inf-obj** → ***to*** *read*
   Helen wants–[*Alan*]–**inf-obj** → ***to*** *read*. |
   Helen makes–[*Alan*]–**inf-obj** → ***read***. |
   her desire–**inf-obj** → ***to*** *come home*
8. Copular [= copular-completive]:
   be–**copul** → ***easy***; *be*–[*a*]–**copul** → ***teacher***
   be–**copul** → ***without*** *a hat*
   be–**copul** → ***at*** *home*
   The problem was–**copul** → ***that*** *these components had to be interconnected*.
9. Subject-completive:
   seem–**subj-compl** → ***easy***; *appears*–**subj-compl** → ***as*** *a victim*
   become–**subj-compl** → ***independent***
   was declared–**subj-compl** → ***dead***
10. Object-completive:
    find–[*this*]–**obj-compl** → ***easy***
    consider–[*him*]–**obj-compl** → ***as*** *a victim*
    make–[*it*]–**obj-compl** → ***possible***
    make–[*Helen a good*]–**obj-compl** → ***wife***
11. Agentive:
    written–**agent** → ***by*** *Alan*; *arrival*–**agent** → ***of*** *Alan*
    decipherable–**agent** → ***by*** *Army's cryptologists*
    shooting–**agent** → ***of*** *the hunters* [= 'the hunters shoot']
    a translation–**agent** → ***by*** *Alan*
    I like ***for*** ← **agent**–[*Alan to*]–*play cards*.
12. Patientive:
    translation–**patient** → ***of*** *this text*; *shooting*–**patient** → ***of*** *the hunters* [= 'the hunters are shot']

Copredicative SSyntRels

13. Subject-copredicative:
    Alan returned–**subj-copred** → ***rich***.
14. Object-copredicative:
    Alan likes–[*Helen*]–**obj-copred** → ***slim***.
    Alan hammered–[*the coin*]–**obj-copred** → ***flat***.

Comparative SSyntRel

15. Comparative:
    older–**compar** → ***than*** *Leo*
    less–[*important*]–**compar** → ***than*** *Leo*
    as–[*important*]–**compar** → ***as*** *Leo*
    Alan loves Helen more–**compar** → ***than*** *Leo*.

Non-Valence-controlled SSyntRels: Modification

Absolutive SSyntRel

16. Absolute-predicative:
    His first attempt–[*a*]–**abs-pred** → ***failure***, *Alan …*
    He went out, his anger–**abs-pred** → ***gone***.
    He ran, his gun–**abs-pred** → ***in*** *his left hand*.

Circumstantial SSyntRels

17. Circumstantial:
    walk–**circum** → ***fast***; *delve*–**circum** → ***deeply***
    He works–**circum** → ***there*** ⟨***in*** [*this office*]⟩.
    [*will*] *write*–[*next*]–**circum** → ***week***
    He ran,–[*his*]–**circum** → ***gun*** [*in his left hand*].

With ← **circum**–[*her paper finished, Helen*]–*can afford this trip*.
walk–[*three*]–**circum** → ***miles***
Leo is coming–[*next*]–**circum** → ***week***.

18. Modificative-circumstantial:
Elegant ← **modif-circum**–[*as always, Alan*]–*walked away*.
19. Appositive-circumstantial:
An old ***man***, ← **appos-circum**–[*Alan*]–*works less*.
20. Attributive-circumstantial:
Abroad, ← **attr-circum**–[*Alan*]–*works less*.

Sentential SSyntRels

21. Parenthetical:
Oddly, ← **parenth**–[*Alan*] *works less*. |
Alan, ***naturally***, ← **parenth**–*accepted the offer*.
As ← **parenth**–[*we have known for some time, Alan*]–*works less*.
To ← **parenth**–[*give an example, I*]–*consider now nominal suffixes*.
22. Adjunctive:
OK, ← **adjunct**–[*I*]–*agree*. | ***John***, ← **adjunct**–[*I*]–*agree*.

PHRASE-LEVEL (= Phrasal) SSYNTRELS: 23–52

These SSyntRels function within elements of the sentence.

General Phrase SSyntRels

Non-valence-controlled SSyntRels: Modification

23. Restrictive:
still ← **restr**–*taller*; ***most*** ← **restr**–*frequent*
not ← **restr**–*here*; *does*–**restr** →
Alan has ***just*** ← **restr**–*arrived*.

Noun Phrase SSyntRels

Valence-controlled SSyntRels: complementation

24. Elective:
the poorest–**elect** → ***among*** *peasants*; *the best*–[*ones*]–**elect** → ***of*** ⟨***from***⟩ *these boys, five*–**elect** → ***of*** *these books*
the most–[*expensive car*]–**elect** → ***in*** *France*

Mixed Type SSyntRels = Valence-controlled/Non-Valence-controlled: Modification

25. Possessive:
Alan's ← **poss**–*arrival*; ***Alan's*** ← **poss**–*bed*
Alan's ← **poss**–*garden*
26. Compositive:
man ← **compos**–[-*machine*]–*interaction*;
car ← **compos**–*repair*; ***noun*** ← **compos**–*phrase*
test ← **compos**–*fire*; ***color*** ← **compos**–*blind*
Chinese- ← **compos**–*engineered*

Non-Valence-controlled SSyntRels: Modification

27. Determinative:
my ← **determ**–*bed*; ***your*** ← **determ**–*departure*
a ← **determ**–*bed*; ***those*** ← **determ**–*beds*
28. Quantitative:
three ← **quant**–*beds*
[*three* ← **num-junct**–]–***thousand*** ← **quant**–*people*
29. Modificative:
comfortable ← **modif**–*beds*; ***visible*** ← **modif**–*stars*
French ← **modif**–*production*
30. Post-modificative:
stars–**post-modif** → ***visible*** (vs *visible stars*)
31. Descriptive-modificative:
these beds,–**descr-modif** → ***comfortable*** *and not expen-sive, …*
The blast–[*came,*]–**descr-modif** → ***sharp*** *and deafening*.
32. Relative:
the paper–[*that I*]–**relat** → ***read*** *yesterday*
the paper–[*I*]–**relat** → ***read*** *yesterday*
the girl–[*who*]–**relat** → ***came*** *first*
name of the file–[*where*]–**relat** → ***to*** *output these data*
33. Descriptive-relative:
this paper–[*which I*]–**descr-relat** → ***read*** *yesterday*
Alan,–[*who*]–**descr-relat** → ***loves*** *her so much, should…*
34. Appositive:
Alan–[*the*]–**appos** → ***Powerful***; ***General*** ← **appos**–*Wanner*
the term–**appos** → ***'suffix'***
Chapter–**appos** → ***XIX***
It happened–[*on the day Spangler disappeared, a*]–**appos** → ***fact*** mentioned in the note.
35. Descriptive-appositive:
This term–**descr-appos** → (***'suffix'***) *will be considered later*. |
You forget about me,–[*your*]–**descr-appos** → ***mother***.
36. Sequential:
man–**sequent** → ***machine*** *interaction*
fifty–**sequent** → ***to*** *seventy dollars*
37. Attributive:
learners–**attr** → ***with*** *different backgrounds*
dress–**attr** → ***of*** *a beautiful color*
years–**attr** → ***of*** *war*; *the bed*–**attr** → ***of*** *Alan*
a man–[*the same*]–**attr** → ***age***
early–**attr** → ***in*** *October*; *elsewhere*–**attr** → ***in*** *the chapter*
38. Descriptive-attributive:
Professor Wanner,–**descr-attr** → ***from*** *Stuttgart, was also present*.

Prepositional Phrase SSyntRels

Valence-controlled SSyntRels: Complementation

39. Prepositional:
in–**prepos** → ***bed***
without–[*three hundred*]–**prepos** → ***dollars***
from–[*pale*]–**prepos** → ***yellow*** *to*–**prepos** → ***red***
from–**prepos** → ***under*** *the bench*; *until*–**prepos** → ***recently***

This SSyntRel is also used for postpositions:
a ***year*** ← **prepos**–*ago*
John's ***disapproval*** ← **prepos**–*notwithstanding*

40. Prepositional-infinitival:
to–**prepos-inf** → ***go*** *to bed*

Verb Phrase (= Analytical Form) SSyntRels

Non-valence-controlled SSyntRels: Ancillary

41. Perfect-analytical:
has–**perf-analyt** → ***written***; *has*–**perf-analyt** → ***been*** *beaten*
42. Progressive-analytical:
was–**progr-analyt** → ***writing***
43. Passive-analytical:
was–**pass-analyt** → ***written***

44. Do-analytical:
    do–**DO-analyt**→***write***; *does*–[*not*]–**DO-analyt**→*write*

Conjunction Phrase SSyntRels

Valence-controlled SSyntRels: Complementation

45. Subordinate-conjunctional:
    Suppose that–[*Alan*]–**subord-conj**→***comes***
    so as–[*not*]–**subord-conj**→***to*** *irritate Leo*
46. Coordinate-conjunctional:
    Alan and–**coord-conj**→***Helen***
47. Comparative-conjunctional:
    than–**compar-conj**→***Helen***; *as*–**compar-conj**→***always***
    closer than–[*John*]–**compar-conj**→***thought***
48. Absolute-conjunctional:
    If–[*a*]–**abs-conj**→***pronoun***, *the grammatical subject may…*
    while–**abs-conj**→***in*** *bed*

Wordlike Phrase SSyntRels

Non-valence-controlled SSyntRels: Ancillary

49. Verb-junctive:
    give–**verb-junct**→***up***; *bring*–**verb-junct**→***down***
50. Numeral-junctive:
    fifty←**num-junct**–*three*; ***fifty***←**num-junct**–*third*
51. Binary-junctive:
    if–[…]–**bin-junct**→***then***…; *the*–[*more*…]–**bin-junct**→***the*** *more*…; *till*–**bin-junct**→***after***
    from–[…]–**bin-junct**→***to***…; *either*–[…]–**bin-junct**→***or***…
52. Colligative:
    is dealt–**collig**→***with*** [stranded prepositions] *later*

II. Coordinate SSyntRels: 53–54

Non-valence-controlled SSyntRels: Coordination

53. Coordinative:
    Alan–**coord**→**and** *Leo*
    rich,–**coord**→***intelligent***–***coord***→***and*** *beautiful*
54. Quasi-coordinative:
    He was abroad–**quasi-coord**→***without***–[*a penny*]–**quasi-coord**→***in*** *a desperate situation.* |
    These moneys we keep hidden under–[*a loose board*]–**quasi-coord**→***under***–[*the floor*]–**quasi-coord**→***under***–[*a chamber pot*]–**quasi-coord**→***under*** [*my friend's bed*] [T. Capote, 'A Christmas Memory']

Morphological Dependency [= Morph-D]

The lexeme L_2 morphologically depends on the lexeme L_1 according to the inflectional category CAT, i.e., L_1–**morph**/CAT→L_2, iff at least one grammeme G of L_2, such that G ∈ CAT, is determined by L_1.

L_1 is called the controller of L_2, and L_2 the target of L_1.

Two major types of Morph-D are known: agreement (including a particular case: congruence of a pronoun with its antecedent) and government.

Properties of Morphological Dependency

Like Sem-D and Synt-D, Morph-D is antireflexive: the inflection of a lexeme in a sentence cannot be determined by the lexeme itself. Like other Ds, Morph-D is also antisymmetrical, although in a more complex sense than in case of Sem-D and Synt-D: one has to take into account the inflectional category whose grammeme is imposed by the controller. Thus, L_1–**morph**/CAT→L_2 entails ¬(L_1←**morph**/CAT–L_2). However, for two different inflectional categories $CAT_1 \neq CAT_2$, the following situation is possible: L_1–**morph**/CAT_1→L_2 and, at the same time, L_1←**morph**/CAT_2–L_2; here is an example:

(8) Rus. *dv*+***a*** *stol* +***a*** 'two tables'
    two MASC.NOM table$_{(masc)}$ SG.GEN
    The noun form *stola* morphologically depends for its singular and genitive case on *dva*, while *dva* morphologically depends on *stola* for its masculine gender; cf *pjat' stol* + ***ov*** PL.GEN 'five tables' or *dvadcat' odin stol*+***Ø*** SG.NOM lit. 'twenty-one table,' but *dv*+***e*** FEM.NOM *krovati*$_{(fem)}$ 'two beds.' (In Russian, a quantified N after a numeral Num in the nominative or the inananimate accusative is in the singular genitive, if this Num ends in 2, 3 or 4, and in the plural genitive, if the Num ends in 5 or a bigger number; if the Num ends in 1, the N is in the singular.)

Like Sem-D and Synt-D, Morph-D is labeled in the sense that, in order to fully specify a Morph-D, we have to indicate the inflectional category of L_2 whose grammeme is imposed by L_1.

Morph-D does not presuppose the uniqueness of the governor: a wordform can morphologically depend simultaneously on several wordforms (but, as was indicated, for different inflectional categories).

Morph-D is not linguistically universal: it is not present in many languages at all – languages that have neither agreement nor government (e.g., Chinese or Vietnamese); in a language where it exists, it is not present in all sentences; and in a sentence where it is present it does not necessarily embrace all wordforms. Morph-Ds do not form, generally speaking, a connected structure in a sentence.

Possible Combinations of the Three Types of Linguistic Dependency

The semantic, syntactic, and morphological dependencies are reciprocally autonomous and can co-occur in 14 logically possible combinations (Mel'čuk, 1981, 1988: 118–128, 2009: 58–65; Nichols, 1986). Here are illustrations of some of these combinations.

CASE 1: L_1 L_2, i.e., no syntagmatic dependency between two wordforms.

Example: *the* and *mother* in (1).

CASE 3: L_1 **synt**→L_2, i.e., Synt-D without any Sem-D or Morph-D.

(9) a. In Japanese, a quantifier, while bearing semantically on the Subject as in '**Five** people were injured,' depends syntactically on the verb, with which it has neither semantic nor morphological links (semantically, *gonin*

bears on *keganin*, while morphologically *gonin* depends on it for its classifier):

Sono ziko +de keganin +ga
this accident LOC injured.people SUBJ(ective)
go +nin *deta*
five CLASS(ifier) emerged
'In this accident, five people were injured.' = lit. '…injured.people five-ly emerged.'

b. In English, a measure noun used as a DirO depends syntactically on the verb, but has no semantic or morphological link with it (semantically, the verb dominates the noun quantified):
John bought–[*five*]–**synt**→***kilos*** [*of potatoes*].

CASE 6: $L_1 \begin{smallmatrix}\text{—sem}\rightarrow\\ \leftarrow\text{synt-}\end{smallmatrix} L_2$, i.e., Sem-D and Synt-D oriented the opposite ways, with no Morph-D.
Example: an adjective and the modified noun in an ADJ + N phrase in a language with invariant adjectives, cf *new* and *areas* in *new areas*.

CASE 8: $L_1 \begin{smallmatrix}\text{—sem}\rightarrow\\ \leftarrow\text{morph-}\end{smallmatrix} L_2$, i.e., Sem-D and Morph-D, oriented the opposite ways, with no Synt-D.
Example: an attributive adjective in a language with adjective agreement.

(10) Fr. *Marie semblait* ***fatiguée*** 'Marie seemed tired.'
Semantically, *Marie* depends on *fatiguée* ['fatigué–**sem**→Marie'], but morphologically, *fatiguée* depends on *Marie* for its singular and feminine; syntactically, the two lexemes are not directly linked.

CASE 13: $L_1 \begin{smallmatrix}\text{—sem}\rightarrow\\ \text{—synt}\rightarrow\\ \leftarrow\text{morph-}\end{smallmatrix} L_2$, i.e., Sem-D and Synt-D oriented the same way, while Morph-D is opposite.
Example: a Main Verb and its nominal actants in a caseless language with polypersonal agreement of the MV.

(11) Abkhaz *Sara Nadš'a* ***i+l*** *+əs+teit'* *aš*^{w}*q*^{ʷ}*aə*
I it her I gave book
'I gave Nadsha [a] book.'

CASE 14: $L_1 \begin{smallmatrix}\text{—sem}\rightarrow\\ \leftarrow\text{synt—}\\ \leftarrow\text{morph-}\end{smallmatrix} L_2$, i.e., Synt-D and Morph-D are oriented the same way, while Sem-D is opposite.

Example: an adjective and the modified noun in a language with adjectival agreement, where the adjective bears semantically on the noun, but syntactically and morphologically depends on it, e.g., Lat. *bon* + ***a***$_{\text{FEM.SG.NOM}}$ *urb*$_{\text{(fem)}}$ + ***s***$_{\text{SG.NOM}}$ 'good city' ~ *bon* + ***arum***$_{\text{FEM.PL.GEN}}$ *urb*$_{\text{(fem)}}$ + ***ium***$_{\text{PL.GEN}}$ 'of.good cities.'

Syntactic Dependency and Syntactic Constituency

General Remarks

Before the appearance of the dependency approach to syntax (early 1960s), the mainstream syntactic investigations were based on constituency, or phrase structure – the staple diet of Bloomfieldean structuralism and then of Chomskyan generative-transformational grammar. At the beginning, it was believed that dependency and constituency were two different legitimate ways of describing the syntactic structure of sentences. Padučeva (1964) and Gaifman (1965) studied formally the mathematical equivalence of D-approach with constituency, or C-, approach, and formulated conditions under which they can in fact be made equivalent. For systematic comparison of both approaches, see Vennemann (1977), Hudson (1980a,b), Dahl (1980), Matthews (1981: 71–95), Mel'čuk (1988: 13–17, 2009: 89–92), Sgall and Panevová (1988–89), Rambow and Joshi (1997), and Langacker (1997). However, as time was passing and more studies and descriptions were published, it became increasingly clear that there are serious difficulties with constituency, for instance:

- Constituents, which essentially are groups of words with particular properties, function in language as a **means** used to express the syntactic structure of a sentence; therefore, they cannot be part of the syntactic structure, let alone the main formalism for its description.
- Constituents are not sufficient to univocally express any syntactic structure. The specification of the head element in a constituent and, in some cases, the type of relation between it and other elements have to be added. But as soon as this is done, we obtain, in point of fact, a dependency structure (with some useless 'architectural excesses').
- Constituents do not allow one to distinguish, in a natural way, the purely **syntactic** hierarchy of sentence elements and their **linear** hierarchy (word order).
- Constituents are not convenient for the description of the transition from a semantic structure thought of as semantic network (an equivalent of predicate logic formulas) to a syntactic structure written in terms of constituents.
- Constituents fail in the treatment of several important linguistic phenomena: a stumbling block that warrants a more detailed consideration, see below, Section Linguistic Advantages of syntactic Dependency.

By the end of the past century, constituency representation in syntax was in rapid decline. Langacker (1997) was a *coup de grâce*: the paper explicitly shows that constituents are good only for the (mor)phonological surface, where they take form of familiar prosodic units – phrases, and for conceptual level, where they form conceptual groupings, candidates for unified symbolization; however, constituents are not good for syntax.

Because of these and other similar problems, linguists started very early looking for a hybrid between D- and C-representations (e.g., Vennemann, 1977). The incentive for integration comes from such phenomena as coordination (see Section Syntactic Dependency in the First Half of the Twenty-First Century), extractions (*I know* ***which girl*** *you told my wife Alan was going out with*), analytical forms (*has been detected*), collocations (*make headway*, *pay a visit*, *launch an attack*), and the like. The idea is to introduce for any of these constructions a special type of subtree that is allowed to occupy as a whole one node of a D-tree. The first formal proposal for a mixed D-/C-representation known to me – by means of syntactic groups – was advanced in Gladkij (1966, 1985). The most recent one is Kahane (1997), which introduces the concept of *bubble* (a subset of nodes of a D-tree which is treated as a node, while having inside a completely specified D-structure of its own, including other bubbles).

However, all the difficulties of description that hybrid D-/C-structures try to address are more efficiently overridden by distinguishing deep and surface sublevels of representation, see Section Deep- versus Surface-Syntactic Structure and Deep- versus Surface-Syntactic Relation.

Linguistic Advantages of Syntactic Dependency

Among linguistic phenomena that are quite adequately described in terms of syntactic dependency, but cannot be in terms of constituency, the following five stand out.

Valence, Diatheses, and Voices

A diathesis of a lexeme L is a particular correspondence between L's Sem-actants [= SemAs] and its Deep-Syntactic actants [= DSyntAs]. For instance, the verbs FOLLOW and PRECEDE have inverted diatheses: $X_{\mathbf{I}}$ *follows* $Y_{\mathbf{II}} \equiv Y_{\mathbf{I}}$ *precedes* $X_{\mathbf{II}}$; or, symbolically, X ⇔ **I**, Y ⇔ **II** for FOLLOW and X ⇔ **II**, Y ⇔ **I** for PRECEDE. Such a formulation, as well as the notion itself of actant – on three different levels (SemAs, DSyntAs, and SSyntAs, see Mel'čuk, 2004a,b) – is possible only within a dependency framework.

This description of diathesis leads to a clear definition of voice: a voice is a particular modification of the basic diathesis grammatically marked on the verb. Among other things, the correlation between the active and the passive voices can be represented in the same way: $X_{\mathbf{I}}$ *follows* $Y_{\mathbf{II}} \equiv Y_{\mathbf{I}}$ *is followed by* $X_{\mathbf{II}}$. A calculus of voices can be developed by combining permutations of DSyntAs of L with respect to its SemAs, DSyntAs' suppression, and SemAs' referential identification (see Mel'čuk, 2006: 181–262).

The set of L's actants forms L's active valence, semantic, and syntactic (for passive valence, see above, Section Criteria for Synt-D, Criterion B1 for the orientation of Synt-D). The idea of valence turned out to be a correct metaphor, which spawned many useful solutions. Under the C-approach, this phenomenon has to be described in terms of *subcategorization frame*, thus emphasizing the classification of valence-controlled entities.

Lexical Functions

To describe regular collocations of the type *wield authority*, *pursue a policy* or *honor a commitment*, an inventory of a few dozen Lexical Functions [= LFs] is proposed (Mel'čuk 1996 and 2007); cf Real_1(AUTHORITY) = *wield* [~], Real_1(POLICY) = *pursue* [ART ~], Real_1(COMMITMENT) = *honor* [ART ~], where Real_1 is a verb that means 'to do with N what one is supposed to do with N.' Similarly, for *empty promise*, *poor example* or *pipe dream* we have: AntiVer(PROMISE) = *empty*, AntiVer(EXAMPLE) = *poor*, AntiVer(DREAM) = *pipe* [~], where AntiVer is an adjective that means 'such that it should not be.' An LF is applied to the base of a collocation (in small caps above) and returns the corresponding collocate. LFs, specified for a lexeme in its lexical entry, allow for correct lexical choices under text generation or automatic translation, as well as for efficient paraphrasing, equally necessary for these tasks. No less important is their role in lexicography, as well as in language, teaching, and learning.

The base of a collocation and its collocate are always linked by a particular Deep-Synt-D, specific for a given LF: Real_1(L)–**II**→L, L–ATTR→AntiVer(L), etc.

The LF formalism is only possible based on syntactic D-representation.

Phrasemes

Numerous phrasemes – idioms as well as collocations – do not easily lend themselves to a C-description. Thus, any idiom one of whose Sem-actants depends syntactically on an 'internal' element of the idiom raises a problem: what could a C-structure be of such idioms as LOOSE [X's] COOL, PULL [Y's] LEG, CAT GOT [X's] TONGUE, [X's] GOOSE IS COOKED, DRAG [X's] FEET, etc.? The same problem occurs with all collocations that accept free (= nonidiomatic) modifiers: *kick the* ***filthy*** *habit*, *pull* ***many new*** *strings*, *make an* ***unbelievable*** *headway*, etc. A D-structure for all such expressions is straightforward. O'Grady (1998) said that much, although apologizing for using something that looks like dependencies, but which is 'in fact' just an abbreviation for, as the author says, "more conventional syntactic structure" (p. 284).

Paraphrasing

Expressing the syntactic structure in terms of DS opens the way for powerful paraphrasing – that is, the calculus of sets of semantically equivalent DSyntSs. Such paraphrasing proves to be necessary in translation because of lexical, deep-syntactic, and morphological mismatches between sentences of different languages that translate each other (Mel'čuk and Wanner, 2001, 2006, 2008). An example of mismatches can be seen in the translation of the English sentence (12a) into Russian (and vice versa):

(12) a. *It seems to have stopped raining.*
b. *Dožd,' kažetsja, perestal* lit. 'Rain, [it] seems, stopped.'

The DSyntSs of these sentences and relevant lexical equivalences are given in (12c):

c.

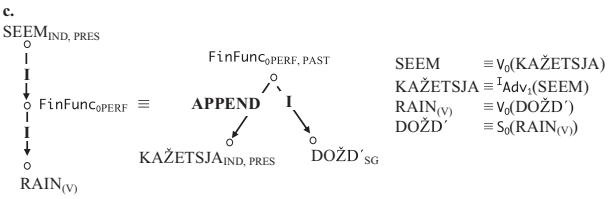

$\mathsf{FinFunc}_0$ in both trees is an LF meaning roughly 'cease to take place'; $\mathsf{FinFunc}_0(\mathrm{RAIN_N})$ = *stop*, while $\mathsf{FinFunc}_0$(DOŽD') = *končit'sja, perestat,' prekratit'sja, projti*.

The DSynt-paraphrasing rules necessary for the transition presented in (12c) are as follows (with serious simplifications):

d. Head-switching Synonymous substitution

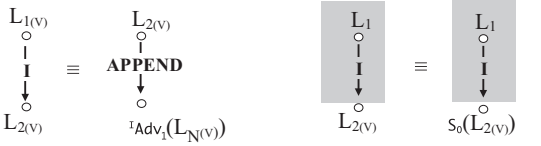

The shaded zones represent the context – that is, the elements that are not affected by the given rule, but control its applicability.

Given the limited number of LFs and of DSyntRels, on the one hand, and the fact that all DSynt-transformations are binary, it is possible to develop an exhaustive set of DSynt-

paraphrasing rules, which are cross-linguistically universal (see Mel'čuk, 1992, 2012, 2013; Milićević, 2007).

Word Order

Synt-D is especially convenient for the description of word order: it forces strict separation of the hierarchical (= genuinely syntactic) order from the linear order, the latter being a surface means for the expression of the former. The main task of syntax – linearizing a two-dimensional Synt-structure (Tesnière, 1959: 19–20) – is solved more easily with Synt-Ds:

- Word order rules can be naturally formulated in terms of positioning a Synt-dependent with respect to its Synt-governor.
- One of the most universal properties of word order – so-called projectivity – can be identified and described only in terms of dependency.

Projectivity

(The word order in) the sentence *S* is projective iff in the projection of the SSyntS(*S*) on *S* no dependency arrow crosses another dependency arrow or a projection perpendicular.

Thus, sentence (1) is projective:

(13)

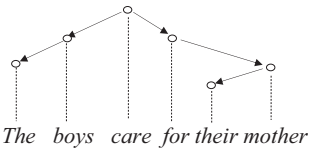

A majority of sentences in actual texts are projective. (Although how big a majority depends of course on the language; according to Havelka (2007): 614, several European languages have, on the average, between 23% and 36% of nonprojective sentences: for instance, Czech – 23.15%, Dutch – 36.44%, German – 27.75%, Slovene – 22.16%, while Chinese has no nonprojective sentences at all, Spanish has only 1.72% and Japanese – 5.29%.) Projectivity allows for simpler and more general word order rules and facilitates computer processing of texts: except for particular cases, only projective sentences must be produced from a given SyntS, and only SyntSs that guarantee projectivity must be associated with a given sentence. However, establishing and describing cases where projectivity violations are permitted or even obligatory is also highly instructive.

For instance, Latin sentence (14a) is nonprojective; cf its SSyntS projected on it in (14b):

(14) a. *Tu solebas meas esse aliquid putare nugas*
[Catullus]
thou used my be something believe trifles
'You used to believe my trifles to be something.'

b.

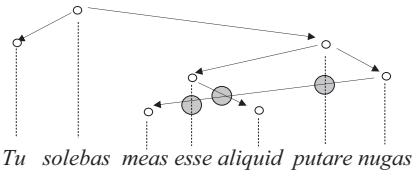

Shaded circles indicate "crime scenes" – that is, the spots of projectivity violations.

These violations happen because, instead of a 'normal' word order in the sentence *Tu solebas putare meas nugas aliquid esse*, Catullus chose to place the phrase *esse aliquid* before the infinitive *putare* and move the modifier *meas* before this phrase.

Projectivity can be violated in many special cases, for instance:

(15) a. English b. French

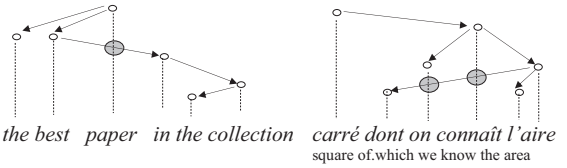

The source of nonprojectivity in the English example is the phrase 'in the collection,' which is the actant of the superlative BEST, but must be positioned after the modified noun PAPER; in the French example, it is the relative pronoun DONT 'of which': it depends on AIRE 'area,' but has to be positioned at the beginning of the clause.

Within D-approach, nonprojective constructions pose no problem whatsoever (while C-approach cannot represent nonprojective constructions without additional machinery – e.g., transformations).

- D-approach has the inherent ability to accommodate long-range syntactic links. The perturbations introduced into the word order of a sentence by its Communicative Structure – Frontings, Extractions, Postponings, etc. – can wreak havoc on a C-structure, since even the closest-knit phrases can be torn apart and permuted. But a D-structure, completely free of linearity and contiguity, is insensitive to such permutations; the discontinuities simply do not exist in a D-structure.

Linguistic Insufficiency of Syntactic Dependency: A Case of Coordination

If we use two levels of representation for syntactic structures (Deep-SyntS and Surface-SyntS) plus Communicative Structures on each level, then a pure D-representation in syntax seems to be sufficient for all syntactic phenomena, except for some types of coordination (cf Hudson, 1990: 97ff.). A wordform W may relate either to a whole conjoined phrase or to its Synt-head alone, such that the two constructions are formally distinguished – by syntactic means – and have different meanings; however, within the strict D-approach, both types of structure can be represented only by the direct Synt-D of W on the Synt-head of the conjoined phrase. Consequently, one SyntS corresponds in such cases to two different meanings:

(16) The SyntS *old*←*men*→*and*→*women* can correspond to
– either a joint reading, where the adjective semantically bears on both nouns:
'old {men + women}' = '{old men} and {old women}';

– or a disjoint reading, where the adjective bears on one noun only:
'{old men} and women [who are not necessarily old].'

Different surface implementations that formally distinguish intended meanings are in principle available, for instance, prosody: *old men and women* (without a pause) versus *old men* || *and women*. The semantic contrast accompanied by a formal contrast requires that the semantic distinction be maintained in the SyntS (Criterion C1: p. 3). The only natural way to do so is to complement D-approach with syntactic groupings – specification, within the SyntS, of D-subtrees relevant in such cases (Mel'čuk, 1974: 214–216, 1988: 28–33): *old*←**modif**–*men*→*and*→*women* (without grouping) stands for the disjoint reading, but *old*←**modif**–[–*men*→*and*→*women*] (with a grouping indicated by square brackets) for the joint reading. Note that a grouping is not a constituent: there is no higher node to represent it as a whole, there is no embedding, and it does not participate in D-links as such. The use of groupings makes D-formalism sufficient for all syntactic purposes.

Phrases and Phrase-Like Units in Syntactic Dependency Approach

To sum up: constituents in the strict sense of the term have no place in a dependency description. Nonetheless, dependency syntax uses four types of constituent-like entities.

- Syntactic groupings are closer to constituents than anything else used within D-approach; they are shown in the SyntS. Their deployment is, however, limited to one very specific coordinate construction (and, as just stated, they are not constituents).
- SSynt-rules make use of standard subtrees [= Δ] Mel'čuk and Pertsov, 485–491). For instance, $\Delta_{(NP)}$ is a subtree equivalent to the tree of a noun phrase; it includes, among other things, the trees of the phrases FROM Num_1 TO Num_2 N (*We use* ***from ten to fifteen tons*** [*of concrete*]) or BETWEEN Num_1 AND Num_2 N (*We use* ***between ten and fifteen tons*** [*of concrete*]), etc., which have the distribution of noun phrases; $\Delta_{(VP)}$ is a nonbranching subtree equivalent to a verb (e.g., *might have been sure to cease to handle*); etc. These standard subtrees are convenient abbreviations that allow for the generalization and, at the same time, simplification of SSynt-rules. A standard subtree is of course not a constituent in the classical sense of the term.
- Minimal phrase linear patterns indicate rigid linear arrangements of wordforms within minimal phrases (a minimal phrase is a phrase that does not consist of other phrases); here is, for instance, such a pattern for noun phrases in English: Standard subtrees are used exclusively in word-order rules, not in the syntactic structures.

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Coordinate conjunction	Preposition	Demonstrative	Numeral	Possessive adjective	Adjective	Noun	'Formula'
OR	ABOUT	THESE	FOUR	OUR	INTERESTING	EXAMPLES	(11)

or about these four our interesting examples (*11*)

- Actual phrases in the Morph-ProsS of a sentence very closely remind one of constituents: there are continuous segments of speech, separated by pauses (actual and potential) and characterized by specific stresses and intonational contours. For instance:

(17) The DMorph and DMorph-Prosodic Structures of sentence (1):

↗ ↗ ↘
THE BOY_{PL} (|) $CARE_{IND, PRES}$ || FOR THEIR MÓ$THER_{SG}$ |||

Yet they are not constituents, since there is no hierarchy with embedding.

Syntactic Dependency in the First Half of the Twenty-First Century

Dependency representation in syntax used to be the staple diet of grammarians – since the Antiquity through the Middle Ages and up to modern times. But then, in the 1930s, Leonard Bloomfield introduced constituents to the English-speaking linguistic community, and the second half of the last century saw a genuine boom produced in and around C-approach by Chomskyan linguistics. Pushed into a corner of the linguistic stage, D-approach, however, not only survived, but almost immediately started taking back lost territories – slowly but steadily: Tesnière (1959), Lecerf (1960), Hays (1960, 1964), Mel'čuk (1962, 1963), Iordanskaja (1963), Robinson (1970), Kunze (1975), Mel'čuk (1988), Sgall and Panevová (1988/1989), etc. Several new linguistic frameworks where practically based on dependencies, even though this is not always explicitly stated, such as Case Grammar (Ch. Fillmore), Lexical-Functional Grammar (J. Bresnan), Relational Grammar (D. Perlmutter), Word Grammar (R. Hudson), or Cognitive Grammar (R. Langacker). In the first quarter of the twenty-first century, the position of D-approach in syntax seems to be rather strong. For a general review of main topics related to Synt-D, see Heringer (1993), Hudson (1993a,b), and Mel'čuk (2009); a description of German syntax in terms of Synt-D is offered in Engel (1977) and Heringer (1996), of English Syntax, in Mel'čuk and Pertsov (1987) and Apresjan et al. (1989), of Russian Syntax in Mel'čuk (1974) and Apresjan et al. (1992a, 2010). In Natural Language Processing, dependency syntax has

also been advanced by, for instance, Maxwell and Schubert (1989), Apresjan et al. (1992b, 2003). Nearly all state-of-the-art parsing proposals are today based on dependency: see Bohnet (2011) and the references therein. Furthermore, a significant number of corpora annotated with dependency structures are available (see the references in Burga et al. (2011). A good perspective on Synt-D in linguistics and Natural Language Processing is offered by Gerdes et al. (2011); more specifically, on Text Generation, see Wanner and Lareau (2009).

Acknowledgments

The ideas about linguistic dependencies presented in this article have been discussed more than once with and profited a lot from D. Beck, S. Kahane, and N. Pertsov; the final version of the text underwent the scrutiny of L. Iordanskaja, J. Milićević, L. Wanner, and R. Zangenfeind. I express my deepest gratitude to all of them for their fair criticisms and constructive suggestions.

See also: Grammatical Relations; Syntax (General); Valency and Argument Structure in Syntax.

Bibliography

Apresian, J.u., Boguslavsky, I., Iomdin, L., Lazursky, A., Sannikov, V., Sizov, V., Tsinman, L., 2003. ETAP-3 linguistic processor: a full-fledged NLP implementation of the MTT. In: Kahane, S., Nasr, A. (Eds.), First International Conference on Meaning-Text Theory (MTT 2003). École Normale Supeerieure, Paris, pp. 279–288.

Apresian, J.u., Boguslavsky, I., Iomdin, L., Lazursky, A., Mitjušin, L., Sannikov, V., Cinman, L., 1992a. Lingvističeskij processor dlja složnyx informacionnyx sistem [Linguistic Processor for Complex Information Systems]. Nauka, Moskva.

Apresian, J.u., Boguslavsky, I., Iomdin, L., Lazursky, A., Mitjušin, L., Sannikov, V., Tsinman, L., 1992b. ETAP-2: the linguistics of a machine translation system. META 37 (1), 97–112.

Apresian, J.u, Boguslavsky, I., Iomdin, L., Lazursky, A., Pertsov, N., Sannikov, V., Cinman, L., 1989. Lingvističeskoe obespečenie sistemy ÈTAP-2 [Linguistic Software for ÈTAP-2 System]. Nauka, Moskva.

Apresian, J.u, Boguslavsky, I., Iomdin, L., Sannikov, V., 2010. Teoretičeskie problemy russkogo sintaksisa. Vzaimodejstvie grammatiki I slovarja [Theoretical Problems of Russian Syntax. Grammar-lexicon Interaction]. Jazyki slavjanskix kul'tur, Moskva.

Bohnet, B., 2011. In: Gerdes, et al. (Eds.), Comparing Graph-Based and Transition-Based Dependency Parsers, pp. 282–289. See also: http://depling.org/proceedingsDepling2011/papers/bohnet.pdf.

Burga, A., Mille, S., Wanner, L., 2011. In: Gerdes, et al. (Eds.), Looking behind the Scenes of Syntactic Dependency Corpus Annotation: Towards a Motivated Annotation Schema of Surface-Syntax in Spanish, pp. 104–114. See also: http://depling.org/proceedingsDepling2011/papers/burgaMilleWanner.pdf.

Corbett, G., Fraser, N., McGlashan, S. (Eds.), 1993. Heads in Grammatical Theory. Cambridge University Press, Cambridge.

Dahl, Ö., 1980. Some arguments for higher nodes in syntax: a reply to Hudson's "Constituency and dependency". Linguistics 18 (5/6), 485–488.

Engel, U., 1977. Syntax der deutschen Gegenwartssprache, fourth rev. ed.: 2009. Schmidt, Berlin.

Gaifman, H., 1965. Dependency systems and phrase-structure systems. Information and Control 8, 304–337.

Gerdes, K., Hajicová, E., Wanner, L. (Eds.), 2011. International Conference on Dependency Linguistics. Depling 2011, Barcelona, September 5–7, 2011. Exploring Dependency Grammar, Semantics, and the Lexicon. Barcelona. http://depling.org/proceedingsDepling2011/papers/.

Gerdes, K., Kahane, S., 2007. Phrasing it differently. In: Wanner, L. (Ed.), Selected Lexical and Grammatical Issues in the Meaning-Text Theory. John Benjamins, Amsterdam/Philadelphia, pp. 297–335.

Gerdes, K., Kahane, S., 2011. In: Gerdes, et al. (Eds.), Defining Dependencies (and Constituents), pp. 17–27. See also: http://depling.org/proceedingsDepling2011/papers/gerdeskahane.pdf.

Gladkij, A., 1966. Lekcii Po Matematičeskoj Lingvistike. Novosibirsk. [French translation: Leçons de linguistique mathématique, fasc. 1, 1970, Paris.].

Gladkij, A., 1985. Sintaksičeskie Struktury Estestvennogo Jazyka V Avtomatizirovannyx Sistemax Obščenija [Syntactic Structures of Natural Language in Automated Communication Systems]. Nauka, Moskva [A new extended edition: 2007.].

Havelka, J., 2007. Beyond projectivity: multilingual evaluation of constraints and measures on non-projective structures. In: Proceedings of the 45th Meeting for the Association for Computational Linguistics. Association for Computational Linguistics, Prague, pp. 608–615. See also: http://aclweb.org/anthology-new/P/P07/P07-1077.pdf.

Hays, D., 1960. Basic Principles and Technical Variations in Sentence Structure Determination. RAND Corporation, Santa Monica, CA.

Hays, D., 1964. Dependency theory: a formalism and some observations. Language 40 (4), 511–525.

Heringer, H.J., 1993. [Dependency theory] basic ideas and the classic model. Formalized models. In: Jacobs, J., von Stechow, A., Sternefeld, W., Vennemann, T. (Eds.), Syntax. An International Handbook of Contemporary Research, vol. 1. W. de Gruyter, Berlin/New York, pp. 298–328.

Hudson, R., 1980a. Constituency and dependency. Linguistics 18 (3/4), 179–198.

Hudson, R., 1980b. A second attack on constituency: a reply to dahl. Linguistics 18 (5/6), 489–504.

Hudson, R., 1987. Zwicky on heads. Journal of Linguistics 23, 109–132.

Hudson, R., 1990. English Word Grammar. Basil Blackwell, Oxford.

Hudson, R., 1993a. In: Corbett, et al. (Eds.), Do We Have Heads in Our Minds?, pp. 266–291.

Hudson, R., 1993b. In: Jacobs, et al. (Eds.), Recent Developments in Dependency Theory, pp. 329–338.

Iordanskaja, L., 1963. O nekotoryx svojstvax pravil'noj sintaksičeskoj struktury [On some properties of the well-formed syntactic structure]. Voprosy Jazykoznanija 4, 102–112.

Iordanskaja, L., Mel'čuk, I., 2009. In: Polguère, Mel'čuk (Eds.), Establishing an Inventory of Surface-Syntactic Relations: Valence-Controlled Surface-Dependents of the Verb in French, pp. 151–234.

Kahane, S., 1997. Bubble trees and syntactic representations. In: Becker, T., Krieger, H.-U. (Eds.), Proceedings of the 5th Meeting of Mathematics of Language, pp. 70–76. Saarbrücken.

Kunze, J., 1975. Abhängigkeitsgrammatik. Akademie-Verlag, Berlin.

Langacker, R., 1997. Constituency, dependency, and conceptual grouping. Cognitive Linguistics 8 (1), 1–32.

Lecerf, Y., 1960. Programme des conflits, modèle des conflits. Traduction Automatique 1 (4), 11–18, 1 (5), 17–36.

Maxwell, D., Schubert, K. (Eds.), 1989. Metataxis in Practice. Dependency Syntax for Multilingual Machine Translation. Foris, Dordrecht/Providence.

Mel'čuk, I., 1962. Ob algoritme sintaksičeskogo analiza jazykovyx tekstov (obščie principy i nekotorye itogi) [Algorithm for syntactic analysis of texts (general principles and some results)]. Mašinnyj perevod i prikladnaja lingvistika 7, 45–87.

Mel'čuk, I., 1963. Avtomatičeskij analiz tekstov (na materiale russkogo jazyka) [Automatic Analysis of Texts (On the Basis of Russian)]. In: Slavjanskoe Jazykoznanie. Akademija nauk SSSR, Moskva, pp. 477–509.

Mel'čuk, I., 1974. Opyt teorii lingvističeskix modelej «Smysl™ Tekst» [Outline of a Theory of Meaning-Text Linguistic Models]. Nauka, Moskva [Reprinted by «Jazyki russkoj kul'tury» in 1999.].

Mel'čuk, I., 1981. Types de dépendance syntagmatique entre les mots-formes d'une phrase. Bulletin de la Société linguistique de Paris 76 (1), 1–59.

Mel'čuk, I., 1988. Dependency Syntax. SUNY Press, Albany, NY.

Mel'čuk, I., 1992. Paraphrase et lexique: la théorie Sens-Texte et le Dictionnaire explicatif et combinatoire. In: Mel'čuk, et al. (Eds.), Dictionnaire explicatif et combinatoire du français contempo-rain. Recherches lexico-sémantiques III. Les Presses de l'Université de Montréal, Montréal, pp. 9–58.

Mel'čuk, I., 1996. Lexical functions: a tool for the description of lexical relations in the lexicon. In: Wanner, L. (Ed.), Lexical Functions in Lexicography and Natural Language Processing. John Benjamins, Amsterdam/Philadelphia, pp. 37–102.

Mel'čuk, I., 2004a. Actants in semantics and syntax I: actants in semantics. Linguistics 42 (1), 1–66.

Mel'čuk, I., 2004b. Actants in semantics and syntax II: actants in syntax. Linguistics 42 (2), 247–291.

Mel'čuk, I., 2006. Aspects of the Theory of Morphology. W. de Gruyter, Berlin/New York.

Mel'čuk, I., 2007. Lexical functions. In: Burger, H., Dobrovol'skij, D., Kühn, P., Norrick, N. (Eds.), Phraseology. An International Handbook of Contemporary Research. W. de Gruyter, Berlin/New York, pp. 119–131.

Mel'čuk, I., 2009. In: Polguère, Mel'čuk (Eds.), Dependency in Natural Language, pp. 1–110.

Mel'čuk, I., 2012. Semantics: From Meaning to Text, vol. 1. John Benjamins, Amsterdam/Philadelphia.
Mel'čuk, I., 2013. Semantics: From Meaning to Text, vol. 2. John Benjamins, Amsterdam/Philadelphia.
Mel'čuk, I., Pertsov, N., 1987. Surface Syntax of English. A Formal Model within the Meaning-Text Framework. John Benjamins, Amsterdam/Philadelphia.
Mel'čuk, I., Wanner, L., 2001. Towards a lexicographic approach to lexical transfer in machine translation (illustrated by the German-Russian language pair). Machine Translation 16 (1), 21–87.
Mel'čuk, I., Wanner, L., 2006. Syntactic mismatches in machine translation. Machine Translation 20 (2), 81–138.
Mel'čuk, I., Wanner, L., 2008. Morphological mismatches in machine translation. Machine Translation 22, 101–152.
Milićević, J., 2007. La paraphrase. Modélisation de la paraphrase langagière. Peter Lang, Berne.
Nichols, J., 1986. Head-marking and dependent-marking grammar. Language 62 (1), 56–119.
O'Grady, W., 1998. The syntax of idioms. Natural Language and Linguistic Theory 16, 279–312.
Padučeva, E., 1964. O sposobax predstavlenija sintaksičeskoj struktury predloženija [On ways for representing the syntactic structure of a sentence]. Voprosy Jazykoznanija no. 2, 99–113.
Polguère, A., Mel'čuk, I. (Eds.), 2009. Dependency in Linguistic Description. John Benjamins, Amsterdam/Philadelphia.
Rambow, O., Joshi, A., 1997. A formal look at dependency grammars and phrase-structure grammars. In: Wanner, L. (Ed.), Recent Trends in Meaning-Text Theory. John Benjamins, Amsterdam/Philadelphia, pp. 167–190.
Robinson, J., 1970. Dependency structures and transformational rules. Language 46 (2), 259–285.
Sgall, P., Panevová, J., 1988–89. Dependency syntax – a challenge. Theoretical Linguistics 15 (1), 73–86.
Tesnière, L., 1959. Éléments de syntaxe structurale. Klincksieck, Paris.
Vennemann, T., 1977. Konstituenz und Dependenz in einigen neueren Grammatiktheorien. Sprachwissenschaft 2 (3), 259–301.
Wanner, L., Lareau, F., 2009. Applying the meaning-text theory model to text synthesis with low- and middle-density languages in mind. In: Nirenburg, S. (Ed.), Language Engineering for Lesser-Studied Languages [= NATO Science for Peace and Security Series – D: Information and Communication Security, Volume 21]. IOS Press, Amsterdam, pp. 207–242.
Zwicky, A., 1985. Heads. Journal of Linguistics 21 (1), 1–30.
Zwicky, A., 1993. In: Corbett, et al. (Eds.), Heads, Bases and Functors, pp. 292–315.

Dependency Theory

Christopher Chase-Dunn, University of California-Riverside, Riverside, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H.R. Sonntag, volume 5, pp. 3501–3505, © 2001, Elsevier Ltd.

Abstract

Dependency theory emerged in the 1960s in reaction to modernization theories of development, arguing that international inequalities were socially structured and that hierarchy is a central feature of the global system of societies. It sought to explicate the institutional structures by which powerful core states continued to exploit and dominate less powerful states even after decolonization and the establishment of official sovereignty in peripheral nations. Ignoring the core/periphery hierarchy is a mistake not only for reasons of completeness, but also because the ability of core capitalist states to exploit noncore resources and labor has been a major factor in deciding the winners of global competition. A key insight of dependency theory is that capitalist globalization has occurred in waves and that waves of integration are followed by periods of globalization backlash. Although industrial production has largely moved from the core to the noncore, rather than flattening the world this trend has been accompanied by the extension and reorganization of modes of control and exploitation based on financial transactions and foreign investment.

Dependency theory emerged out of Latin America in the 1960s in reaction to modernization theories of development. *Dependistas* attributed the difficulties of development in the global South to the legacies of the long history of colonialism as well as contemporary international power relations. This approach suggested that international inequalities were socially structured and that hierarchy is a central feature of the global system of societies. It was a reformulation of earlier theories of imperialism but it focused on the institutional structures by which powerful core states continued to exploit and dominate less powerful states even after decolonization and the establishment of official sovereignty in the noncore. It also focused on the continuing consequences of the global hierarchy for the development of national societies in Latin America (Cardoso and Faletto, 1979). This approach was eventually extended to Africa and Asia and it inspired the world-systems perspective.

In the 1970s, the world-systems perspective emerged out of dependency theory when Samir Amin, Andre Gunder Frank, and Immanuel Wallerstein began to formulate the concepts and to narrate the analytic history of the modern world system. A whole global system was conceived as a stratified structure composed of economically, culturally, and militarily dominant core societies (themselves in competition with one another) and dependent peripheral and semiperipheral regions. Some dependent regions have been successful in improving their positions in the larger core/periphery hierarchy, while most have simply maintained their peripheral or semiperipheral positions. This perspective on world history also examines certain cyclical features of social change and long-term patterns of development in historical and comparative perspective. The development of the modern world system is alleged to be driven primarily by capitalist accumulation and geopolitics in which businesses and states compete with one another for power and wealth. Competition among states and capitals is conditioned by the dynamics of struggle among classes and by the resistance of peripheral and semiperipheral peoples to domination and exploitation from the core. The contemporary semiperiphery is composed of large and powerful countries in the global South (e.g., Mexico, India, Indonesia, Brazil, China) as well as smaller countries that have intermediate levels of economic development (e.g., the East Asian NICs, South Africa, Israel).

Dependency theory requires that we think structurally. We must be able to abstract from the particularities of the game of musical chairs that constitutes uneven development in the system to see the structural continuities. The core/periphery hierarchy is reproduced, though some countries have moved up or down. The interstate system remains, though the internationalization of capital has further constrained the abilities of states to structure national economies. States have always been subjected to larger geopolitical and economic forces in the world system, and as is still the case, some have been more successful at exploiting opportunities and protecting themselves from liabilities than others.

Dependency theory emerged to contest modernization theory, which contended that development would diffuse from the developed to the underdeveloped societies by means of certain kinds of contact between the two. Andre Gunder Frank proposed the notion of "the development of underdevelopment" to explain how certain kinds of hierarchical interaction based on economic exploitation and neocolonial forms of political/military control continued to reproduce the inequalities between the developed and the less developed national societies. Other dependistas examined the conditions under which some developing societies were able to somewhat successfully overcome the larger structures of power–dependence relations and to develop core-like economies and polities. Cardoso and Faletto (1979) explored the notion of 'dependent development' and this notion was examined in the work of Peter Evans (1979, 1995). Other researchers used data on large numbers of national societies to test hypotheses about the effects of international dependence on national development and within-country inequality, often finding that the negative effects on development and the increasing of within-country inequality predicted by dependency theory were confirmed (e.g., Beer and Boswell, 2002).

From the perspective of dependency theory, many of the phenomena that have been called 'globalization' are the continuation of trends that have long existed. The expansion of international trade, financial flows, and foreign investment by transnational corporations and banks in recent decades are similar to the great nineteenth-century wave of globalization. Much of the globalization discourse assumes that until recently there were separate national societies and economies and that these have now been superseded by an expansion of international integration driven by information and transportation technologies. Rather than a wholly unique and new phenomenon, globalization is primarily international economic integration, and as such it is a feature of the world system that has been oscillating as well as increasing for centuries.

The Great Chartered Companies of the seventeenth century were already playing an important role in shaping the development of world regions. Certainly the transnational corporations of the present are much more important players, but the point is that 'foreign investment' is not an institution that only became important since 1970 (nor since World War II). Giovanni Arrighi (1994) has shown that finance capital has been a central component of the commanding heights of the world system since the fourteenth century. The current floods and ebbs of world money are typical of the late phase of very long 'systemic cycles of accumulation.'

Dependistas have long contended that ignoring the core/periphery hierarchy or treating the noncore as inert are grave mistakes, not only for reasons of completeness, but also because the ability of core capitalists and their states to exploit noncore resources and labor has been a major factor in deciding the winners of the competition among core contenders. Moreover, the resistance to exploitation and domination mounted by noncore peoples has played a powerful role in shaping the historical development of world orders for centuries. Thus, world history cannot be properly understood without attention to the continuities and changes in global power–dependency relations.

Phillip McMichael (2000) has studied the 'globalization project' – the abandoning of Keynesian models of national development and a new (or renewed) emphasis on deregulation and opening national commodity and financial markets to foreign trade and investment. This approach focuses on the political and ideological aspects of the recent wave of international integration. The term many prefer for this turn in global discourse is 'neo-liberalism' but it has also been called 'Reaganism/Thatcherism' and the 'Washington Consensus.' The worldwide decline of the political left predated the revolutions of 1989 and the demise of the Soviet Union, but it was certainly also accelerated by these events. The structural basis of the rise of the globalization project is the new level of integration reached by the global capitalist class. The internationalization of capital has long been an important part of the trend toward economic globalization, and there have been many claims to represent the general interests of business before. Indeed every modern hegemon has made this claim. But the real integration of the interests of capitalists all over the world has very likely reached a level greater than at the peak of the nineteenth-century wave of globalization.

Dependency theory has always been compatible with an analysis of the global class structure as a whole. Samir Amin's (1980) important study of the global class structure and the North/South wage gap that is far greater than differences in productivity is a valuable contribution that was made long before the global stage of capitalism school discovered the global class structure. But dependistas can agree that the transnational segments of both capitalists and workers have grown and become more important in both production and politics, and so the analysis of transnational class relations is more important than it has been in the past. Of course, the big question is whether or not this new level of transnational integration will be strong enough to prevent competition among states for world hegemony from turning into warfare, as it has always done in the past, during a period in which a hegemon (now the United States) is declining.

The insight that capitalist globalization has occurred in waves, and that these waves of integration are followed by periods of globalization backlash, has important implications for the future. Capitalist globalization increased both intranational and international inequalities in the nineteenth century and in the late twentieth and early twenty-first centuries within-country inequalities mostly increased, while between-country inequality has remained at the same high level that it attained in the nineteenth century (O'Rourke and Williamson, 2000; Bornschier, 2010). Dependency theorists note that industrial production has largely moved from the core to the noncore, but rather than flattening the world this has been accompanied by the extension and reorganization of modes of control and exploitation based on financial transactions and foreign investment. The economic rise of China and India has not much reduced by the overall amount of global inequality, especially if we take into account the growth of inequality within countries as well as between them.

Dependency theory is alive and well. Understanding the changing nature of global power is not a task whose time has ended. Penetrating the fog of globalization, including the rise of the BRICS and the rise of transnational social movements that are contesting the basic logic of capitalist accumulation, can be well understood only using the past insights of dependency theory and extending it to the present and the future.

See also: Capitalism: Global; Economic Sociology; Labor, Division of; Marxism in Contemporary Sociology; Modernization, Sociological Theories of; World Systems Theory.

Bibliography

Amin, Samir, 1980. The class structure of the contemporary imperialist system. Monthly Review 31 (8), 9–26.

Arrighi, Giovanni, 1994. The Long Twentieth Century. Verso, London.

Beer, Linda, Boswell, Terry, 2002. The resiliency of dependency effects in explaining income inequality in the global economy: a cross-national analysis, 1975–1995. Journal of World-Systems Research viii (1), 30–59. Winter.

Bornschier, Volker, 2010. On the evolution of inequality in the world-system. In: Suter, Christian (Ed.), Inequality beyond Globalization. Transaction Publishers, Piscataway, NJ, pp. 39–64.

Cardoso, Fernando Henrique, Faletto, Enzo, 1979. Dependency and Development in Latin America. University of California Press, Berkeley.

Evans, Peter, 1979. Dependent Development. Princeton University Press, Princeton, NJ.
Evans, Peter, 1995. Embedded Autonomy: States and Industrial Transformation. Princeton University Press, Princeton, NJ.
Hirschman, Albert O., 1945. National Power and the Structure of Foreign Trade. University of California Press, Berkeley.
McMichael, Philip, 2000. Development and Social Change: A Global Perspective. Pine Forge, Thousand Oaks, CA.
O'Rourke, Kevin H., Williamson, Jeffrey G., 2000. Globalization and History. MIT Press, Cambridge, MA.
Wallerstein, Immanuel, 2000. The Essential Wallerstein. New Press, New York.

Depression

Neel Halder, Alpha Hospital, Bury, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Depression, a term often used synonymously with being fed up or unhappy, is much more than this. It is a serious mental illness characterized by low mood, loss of interest or pleasure in nearly all activities, feelings of utter hopelessness, and can lead to suicide. It is currently ranked the third most prevalent moderate and severe disabling condition globally by the World Health Organization. This article explores the history of depression, diagnosis, etiology, evidenced-based management, and scales used in depression.

Introduction

> That's the thing about depression: A human being can survive almost anything, as long as she sees the end in sight. But depression is so insidious, and it compounds daily, that it's impossible to ever see the end. The fog is like a cage without a key.
>
> Elizabeth Wurtzel

Depression can mean different things to different people, which is why it is vital that one clarifies what they mean when taking a history from them. The word comes from the Latin 'deprimere' which translates as 'press down.' Indeed, many do describe it as a downward pressure, and 'feeling down' is colloquially used often. *The Oxford Dictionary* (2010) describes it as "a mental condition characterized by severe feelings of hopelessness and inadequacy, typically accompanied by a lack of energy and interest in life." However, one overall sentence does not do this complex and serious issue justice; depression has its own personal meaning for each individual suffering from it. Winston Churchill used to call it his 'Black Dog,' and I recall one individual describing his depression as a "mental and physical equivalent of walking through treacle, because everything is an effort and slowed down." One chapter in a book cannot also do it justice, but what follows is a brief overview of some important aspects of depression.

History of Depression

Depression was one of the earliest conceptualized mental health disorders. Hippocrates was the first physician to recognize melancholia, as he termed it, being a 'brain disorder' in the fifth century BC (Angst and Marneros, 2001). Despite this astute observation, the observation was superseded by the humoral theory, which held that it was caused by excessive black bile. Asclepiades, a first-century BC Greek physician, prescribed 'bathing, exercise, massages, and wine' to help with the mood disorder (Alexander and Selesnick, 1966). In the nineteenth century, French doctors started thinking of mania and depression as being on the same continuum: 'la folie double forme' (Baillarger, 1854). It was around this time that van Gogh produced his paintings, and future historians would comment on how they could tell when he was in his depressed stages of his illness from his paintings alone (see **Figure 1**). Note the use of dark colors in his palate in **Figure 1**, and the use of dark skies and black crows often interpreted as a reflection of his dark mood and suicidal thoughts. He shot himself in the wheat fields within a month of this painting. Others who advanced the knowledge and made significant contributions to the field included Kraepelin (the first disease model of psychiatry, and included social and psychological aspects), Freud, and Bleuler (Goodwin and Ghaemi, 2000). This ultimately culminated in the contemporary classifications given below.

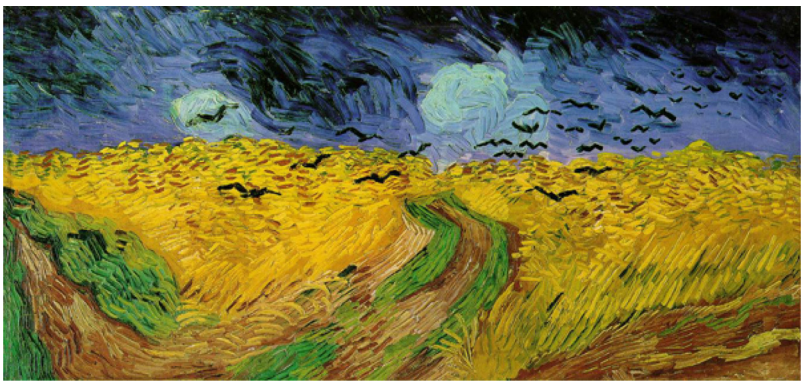

Figure 1 Wheat field with crows. Vincent van Gogh, (1853–1890). Wheatfield with crows, 1890 oil on canvas, 50.5 × 103 cm. Reproduced with permission from Van Gogh Museum, Amsterdam (Vincent van Gogh Foundation).

Diagnosis (Signs and Symptoms) and Phenomenology

There are two widely recognized contemporary classification systems in use worldwide for depression: ICD-10 (WHO, 1992) and DSM-IV (APA, 1994), both of which have new versions due out in the near future.

A duration of at least 2 weeks of symptoms is required for the diagnosis. In ICD-10, depression is divided into a depressive episode (one) or depressive disorder (more than one), both having subcategories of mild, moderate, or severe. Severe episodes can be with or without psychosis. The difference between categories is a clinical judgment that depends on the number, type, and severity of symptoms present.

A diagnosis is made following an extensive history, obtaining collateral history from close family members and health care professionals where possible, possibly aided by rating tools (see below) and following careful exclusion of other disorders and illnesses that can be causing the problem. The other issues (differential diagnoses) that need to be screened for include:

- Substance abuse causing depressed mood (e.g., drugs and alcohol)
- Prescribed medications causing depressed mood (e.g., centrally acting antihypertensives, lipid-soluble beta blockers, benzodiazepines, and progesterone contraceptives)
- Medical illness causing depressed mood
- Medical illness presenting in a similar way to depression (e.g., hypothyroidism)
- Other psychiatric disorders: mania, hypomania, bipolar, schizoaffective, schizophrenia, and so on
- Bereavement unless symptoms persist for more than 2 months or show marked functional impairment, morbid preoccupation with worthlessness, suicidal ideation, psychotic symptoms, or psychomotor retardation (DSM-IV, APA, 1994).

Table 1 outlines the symptoms for both classification systems. The ICD-10 has ten symptoms compared to DSM-IV's nine symptoms for measuring the severity of depressive states.

Functional impact needs to be taken into account. Even with the above guidelines, assessing mood can be problematic, especially regarding the phenomenology of the terms used. Andrew Sims (1995), in what is often regarded as the definitive text on phenomenology, provides the following clarifications. "Feelings" are described as the "passive subjective experience of emotion … [they are] personal convictions, predictive forecasts and social sensibilities." They are the reaction to an experience and are transient, as opposed to 'affect,' which is 'differentiated specific feelings toward objects,' and 'mood,' which is the 'prolonged prevailing state.' 'Emotion' can be regarded as the physiological and psychosomatic concomitants of mood' (Sims, 1995). There is of course a huge variation in mood that is not pathological, and clinicians must make a collaborative judgment based on the information given to them, the classification given above, and the social impact, and the extent of the suffering. Clinicians can also make use of validated rating scales given in the following paragraphs.

Apathy, the absence of feeling or interest, and anhedonia, the inability of experiencing pleasure in previously enjoyed experiences, are common in depression. Anhedonia has been described as 'probably the best clinical marker predicting response to treatment (Sims, 1995). Many also experience depersonalization, which is where the thoughts and feelings seem unreal or not to belong to oneself.

Table 1 Acute episode as defined in DSM-IV and ICD-10

	Symptom of depression	*DSM-IV*	*ICD-10*
1	Depressed mood most of the day, nearly every day	√	√
2	Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day	√	√
3	Loss of energy or fatigue nearly every day	√	√
4	Loss of confidence or self-esteem		√
5	Unreasonable feelings of self-reproach or excessive or inappropriate guilt, nearly every day	√	√
6	Recurrent thoughts of death, suicide, or suicidal behavior	√	√
7	Diminished ability to think or concentrate or indecisiveness, nearly every day	√	√
8	Psychomotor agitation or retardation nearly every day	√	√
9	Insomnia or hypersomnia nearly every day	√	√
10	Change in appetite (decrease or increase with corresponding weight change	√	√

√ indicates the symptom is included.
Source: Gelder, M.G., Andreasen, N.C., Lopez-Ibor, J.L. Jr., Geddes, J.R. (Eds.), 2000. Acute episode as defined by ICD-10 and DSM-IV. In: New Oxford Textbook of Psychiatry. Oxford University Press, Oxford.

There are critics to the diagnostic validity using the classification systems described, using both US and UK data (Kendler and Gardner, 1998; Wainwright et al., 1997), suggesting that the current systems are rather arbitrary, and should be replaced with a likelihood measurement on a continuum rather than a binary decision of whether someone is depressed or not.

In the UK, general practitioners and other primary health care professionals are urged to actively seek out depression in these individuals and others suspected of having depression using the two screening question test (if either are positive, confirmatory evidence should be sought):

1. During the last month have you often been bothered by feeling down, depressed, or hopeless?'
2. During the last month have you been bothered by having little interest or pleasure in doing things?'

For those with a chronic health problem you need to ask three further questions (NICE, 2009):

During the last month have you often been bothered by:

a. Feelings of worthlessness?
b. Poor concentration?
c. Thoughts of death?

In rare and severe cases, psychosis can occur in depression. The psychotic delusions usually take the form of what is known a 'nihilistic' delusion, where there is a belief that oneself, a part of one's body, or the real world does not exist or has been destroyed.

Etiology

There are several models of depression looking at the disorder from different perspectives. What follows is a brief overview of some of them.

Psychological and Behavioral

Aaron Beck in 1979 proposed a triad of thoughts that served as a cognitive theory for depression. The triad involves negative thoughts about:

- The self (i.e., the self is worthless)
- The world/environment (i.e., the world is unfair)
- The future (i.e., the future is hopeless)

The cognitive–behavioral model (CBT) is well established and has a wide evidence base for it (Whitfield and Williams, 2003). Indeed in the UK, it is recommended as the first-line approach for mild to moderate depression (NICE, 2009).

Negative automatic thoughts occur immediately in response to a situation or event, which are outside the person's control. CBT aims to work collaboratively with the current situation and attempts to generate alternative hypotheses and more logical explanations. Cognitive distortions are common (see **Figure 2**), and may stem from underlying 'core beliefs' or 'schemas.' These core beliefs can refer to the self-worth of an individual and take a negative form in depression (e.g., I am unlovable).

Some have expanded on this work and propose a 'five areas approach' (Williams, 2001) for categorizing the symptoms: the theory being that a change in any of the other four areas (social, cognition, behavioral, and social) can lead to a change in the fifth (mood).

Seligman's learned helplessness theory (1975) has been used to explain depression being caused by the perceived lack of control of a situation's outcome (such as abuse for example).

Psychodynamic

According to the self-discrepancy theory (Higgins, 1987), depression occurs when there is a consistent discrepancy between the ideal representation of the self, and the actual representation. This leads to hostility directed toward the actual self. Largely unconscious conflicts between the id (primal desires), superego (a mental representation of the rules and norms of society), and ego (that serves to bridge the other two) lead to regression, repression, and anger turned inward on the self. One needs to resolve early developmental and interpersonal conflicts to overcome depression.

Arbitrary inference

Selective abstraction

Over-generalization

Personalization

Magnification and minimization

Dichotomous thinking

Figure 2 Cognitive distortion examples.

Social

Lack of good-quality social support (actual or perceived) is a risk factor for depressive relapse. Brown and Harris' (1978) often-quoted study found the following vulnerability factors for depression:

- Having three or more children under the age of 14 years
- Not working outside the home
- Lack of a confiding relationship
- Losing the mother before the age of 11 years

'Life events' which are major social events or changes for an individual are thought to play a part in depression. Holmes and Rahe (1967) 'score' life events by calculating 'life change units' that apply to events in the past year of an individual's life which are added, and the final score will give a rough estimate of how stress affects health (**Table 2**). One needs to bear in mind that the way someone responds to any life event can be perceived differently.

An example of a formulation for the social model of depression is given in **Figure 3**.

Genetic

The concordance rate for major depression is greater in monozygotic twins (46%) compared to dizygotic twins (20%) (McGuffin et al., 1996). Other authors think heritability is at about 40–50% for depression and probably more in severe depression (Levinson and Nichols, 2014). Those with first-degree relatives having major depression or recurrent depressive disorder (e.g., parent or sibling) have an increased risk of

Table 2 Holmes and Rahe stress scale

Life event	*Life change units*
Death of a spouse	100
Divorce	73
Marital separation	65
Imprisonment	63
Death of a close family member	63
Personal injury or illness	53
Marriage	50
Dismissal from work	47
Marital reconciliation	45
Retirement	45
Change in health of family member	44
Pregnancy	40
Sexual difficulties	39
Gain a new family member	39
Business readjustment	39
Change in financial state	38
Death of a close friend	37

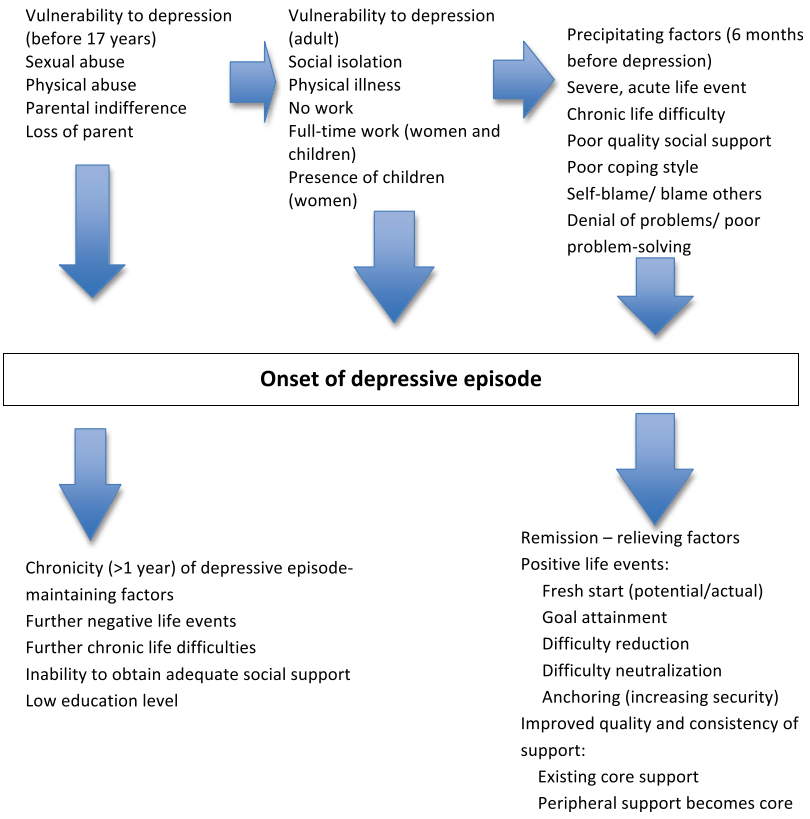

Figure 3 Social model of depressive disorder. Source: Moriss, R., 2003. Mood disorders. In: Young A.H., Ferrier, I.N. (Eds.), Psychiatry, vol. 2 (4). The Medicine Publishing Company, p. 2.

developing depression, by up to five times (Levinson and Nichols, 2014).

Even though twice more females than males are thought to be developing depression, the inheritance of the disorder is thought to be from the combined action of several genes exerting a small or modest effect. Several genes are thought to play a role in developing depression. These genes confer a vulnerability to affective disorder, but the genes must interact with nongenetic (environmental) factors in order to precipitate the illness. This is the basis of the stress-vulnerability theory.

Hormones and Neurotransmitters

In depression, there are reports of a disturbed hypothalamic–pituitary–adrenal axis. Specifically, there appears to be a disrupted normal circadian rhythm of cortisol secretion. Elevated cortisol levels have been found in depressed individuals. Also the normal nocturnal rise in thyroid stimulating hormone (TSH) is blunted in depression.

Those with depression have been shown to have early morning wakening, a diurnal mood variation (usually worse in the mornings), and a short REM latency (which is the time from falling asleep to the onset of REM sleep). Melatonin may, therefore, have a part to play (Puri and Hall, 2004).

In 1965, Schildkraut discussed the monoamine hypothesis of mood disorders, suggesting there is a depletion of central monoamine in depression. The monoamines implicated include serotonin, norepinephrine, dopamine, glutamate, and brain-derived neurotrophic factors (Dunlop and Nemeroff, 2007). The following findings are in favor of the hypothesis: many antidepressant drugs work by increasing the availability of monoamine, such as tricyclic antidepressants (that increase the availability of monoamines in the synaptic cleft), monoamine oxidase inhibitors, and selective serotonin reuptake inhibitors (that increase the availability of serotonin). Also, reserpine, a drug that depletes monoamines such as serotonin, can lead to severe depression and suicide. Serotonin especially has been found to be reduced in depression and can be brought up to normal levels with treatment. However, evidence against the monoamine hypothesis includes that despite the antidepressants causing rapid biochemical changes as described, the clinical improvement observed only starts much later. Also not all drugs that increase the central availability of monoamines have an antidepressant effect (Puri and Hall, 2004). **Figure 4** depicts a neurobiological model of depression (Gelder et al., 2001).

Cultural

Pathological low mood can be defined differently in different cultures. The tools for diagnosis have mainly been developed in Europe or America. The prevalence across countries can, therefore, vary widely. The rates may be influenced by those seeking help too. For example, Punjabi women do not see depression as a medical condition requiring treatment, but

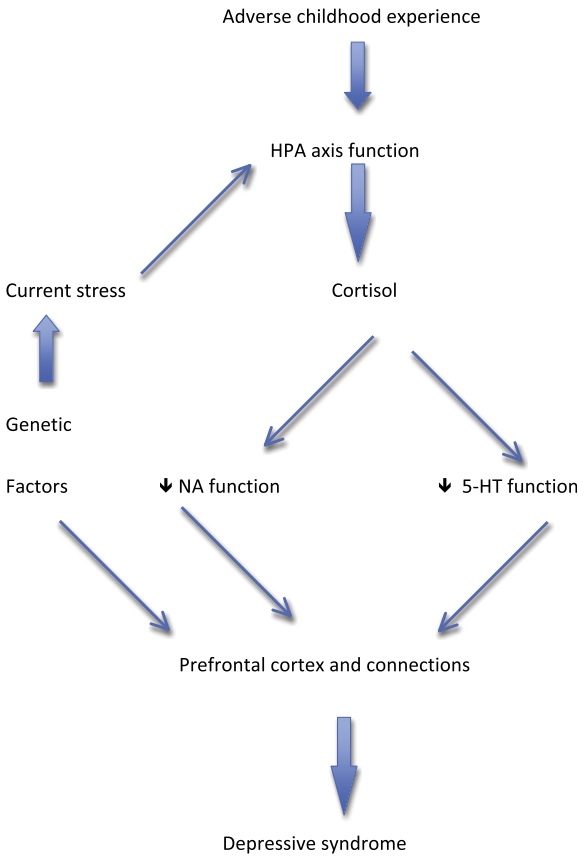

Figure 4 Simplified neurobiological model of depression. In Gelder, M., Mayou, R., Cowen, P., 2001. Shorter Oxford Textbook of Psychiatry. Part 1. Oxford University Press, Oxford, p. 303.

merely a normal reaction to life stressors (Bhugra et al., 1997). In China, many patients receive a diagnosis of 'neurasthenia,' that when reassessed using detailed tools, can be reclassified as major depression in over 80% of cases (Kleinman, 1982). It could be that the diagnosis of neurasthenia is made to avoid the stigma associated with a diagnosis of depression. In Japan, 'shinkei-suijaku' is treated with Morita therapy involving mandatory rest and isolation, followed by progressively more difficult work, and a resumption of a previous social role (Tsung-Yi Lin, 1989). From a study covering 14 countries worldwide, and using the General Health Questionnaire (described in the following section), the prevalence rates of depression varied widely from 1.6% in Nagasaki, to 26.3% in Santiago. The highest levels of correlation were found in symptoms of feeling guilty or worthless, and the lowest correlations were on the biological symptoms of sleep disturbance and altered appetite (Simon et al., 2002).

Treatment and Management

The following is largely based on the current National Institute for Health and Clinical Excellence (NICE) guidelines in the UK for people over 18 years with depression (NICE, 2009). NICE provides national guidance and advice to improve health and social care, and the recommendations are based on the best available evidence, also taking into account the cost–benefit ratio. **Table 3** depicts the stepped-care model for treating depression (NICE, 2009).

For those on step 2, subthreshold depression means fewer than five symptoms from the DSM-IV criteria above. Also, for this group, sleep hygiene is recommended. This essentially includes establishing regular sleep and wake times, avoiding excess eating, smoking, or drinking alcohol before sleep, creating a proper environment for sleep, and taking regular physical exercise. Clinicians are advised to offer active monitoring (to discuss concerns, provide information about depression, reassess within 2 weeks, and to contact the person if they do not attend follow-up appointment). The low-intensity psychological and psychosocial interventions referred to include individual self-help based on CBT principles, computerized CBT, group CBT, and group physical activity programs. Antidepressants are not routinely advised in this group because the risk–benefit ratio is poor, unless they have a past history of moderate–severe depression or they present with subthreshold symptoms that have been present for 2 years or more, or they have subthreshold symptoms for less than 2 years but they do not respond to other interventions.

Herbal remedies such as St John's wort are not recommended by NICE (2009) due to the lack of standard dosage and there are potentially quite serious interactions with other drugs such as oral contraceptives, anticoagulants, and anticonvulsants.

For those on step 3 (**Table 2** above) the recommended management is medication (usually SSRI), high-intensity psychological interventions (e.g., CBT, interpersonal therapy,

Table 3 The stepped-care model

Focus of the intervention	*Nature of the intervention*
STEP 4: Severe and complex depression; risk to life; severe self-neglect	Medication, high-intensity psychological interventions, electroconvulsive therapy, crisis service, combined treatments, multiprofessional and inpatient care
STEP 3: Persistent subthreshold depressive symptoms or mild to moderate depression with inadequate response to initial interventions; moderate and severe depression	Medication, high-intensity psychological interventions, combined treatments, collaborative care, and referral for further assessment and interventions
STEP 2: Persistent subthreshold depressive symptoms; mild to moderate depression	Low-intensity psychosocial interventions, psychological interventions, medication, and referral for further assessment and interventions
STEP 1: All known and suspected presentations of depression	Assessment, support, psychoeducation, active monitoring and referral for further assessment, and interventions

Reproduced with permission from National Institute for Health and Clinical Excellence (2012). Adapted from 'CG90/91.' NICE, London. Available from www.nice.org.uk/CGX90, www.nice.org.uk/CG91.

behavioral couples therapy if relationship contributing to or maintaining depression, or partner can be therapeutically treated), or combined treatment (antidepressants and psychological intervention, e.g., CBT).

For those with severe and complex depression the focus there is on maintaining a safe environment and combined, often multifaceted interventions are required, including medication, high-intensity psychological interventions, electro convulsive therapy, crisis service, and inpatient care. Suicidal ideation and risk assessments should be carried out at every stage and consultation.

There is another chapter devoted to pharmacotherapy for depression and, therefore, it is not covered further here.

Scales Used in Depression

The Beck Depression Inventory (BDI) is a 21-question self-report inventory and is one of the most widely used tools for assessing and screening depression severity (Beck et al., 1961). It looks at both mental and somatic components of depression, but has as its drawback the same problems as any other self-reported tool.

The Hospital Anxiety and Depression Scale (HADS) is a 14-item scale; seven of the items relate to anxiety and seven relate to depression (Zigmond and Snaith, 1983). Each item on the questionnaire is given a score from 0 to 3, and this means that a person can score between 0 and 21 for either anxiety or depression. The items that relate to depression are:

- I still enjoy the things I used to enjoy
- I can laugh and see the funny side of things
- I feel cheerful
- I feel as if I am slowed down
- I have lost interest in my appearance
- I look forward with enjoyment to things
- I can enjoy a good book or radio or TV program

The shortened version of the Patient Health Questionnaire (PHQ-4) is just as good as the longer PHQ9 or HADS or BDI for those who have previously had depression (Thapar et al., 2014).

PHQ-4 involves four questions: "Over the last 2 weeks how often have you been bothered by the following problems: Feeling nervous, anxious, or on edge? Not being able to stop or control worrying? Little interest or pleasure in doing things? Feeling down, depressed, or hopeless?"

For each question, there are four answers: not at all (score 0), several days (score one), more than half the days (score two), and nearly every day (score three). The maximum score is 12. Scores can be interpreted using the following cutoffs: score 0–2 = normal, score 3–5 = mild, score 6–8 = moderate, score 9–12 = severe.

Suicide

Most people (90%) who commit suicide have a psychiatric disorder, and of these 50% have depression (in Puri and Hall, 2004). People who commit suicide are more likely to be:

- Male
- Elderly
- Suffered loss or bereavement
- Unemployed
- Retired
- Childless
- Living alone in a densely populated town
- Broken home in childhood
- Mental or physical illness
- Loss of role
- Drug and alcohol misuse
- Social disorganization, including overcrowding and criminality.

Protective factors against suicide include:

- Religious devoutness
- Lots of children
- Times of war

The above list is taken from Puri and Hall (2004).

During any risk assessment, certain features of a suicide attempt would indicate a higher degree of suicidal intent. These would include whether there was any planning of the event, if the person did not seek help afterward, whether precautions were taken to avoid discovery, whether any dangerous method was used, and the presence of a will or a suicide note.

The National Confidential Inquiry, which produces annual reports on suicide and homicide in the UK, reported in 2014 that particularly the first 3 months after discharge remain a time of high suicide risk – this is especially true in the first 1–2 weeks, and the most common method for suicide remains by hanging (National Confidential Inquiry Annual Report, 2014).

Availability and access to items that can be used for suicide make a difference. Several occupations with the highest suicide rates (per 100 000 population) during 1979–80 and 1982–83, including veterinarians (ranked first), pharmacists (fourth), dentists (sixth), doctors (tenth), and farmers (thirteenth), have easy occupational access to a method of suicide such as pharmaceuticals or guns (Roberts et al., 2013). Literature reviews suggest gun availability increases the rates of suicide in the USA (Miller and Hemmingway, 1999, 2001). More recently, the high-risk occupations appear to have changed, and a breakdown of the 10 highest risk occupations for suicide for men and women is given in **Table 4** below (from Roberts et al., 2013).

Prognosis

With appropriate treatment, 70–80% of individuals with major depressive disorder can achieve a significant reduction in symptoms (Halverson and Bienenfield, 2014). For mild depression, spontaneous recovery is likely. For major depression, four out of five who have had psychiatric care for an episode will have at least one more episode in their lifetimes (with a median of four episodes). The average length of an episode of depression is 6–8 months (NICE, 2013), but the outlook obviously varies with the severity of the episode.

The risk factors for increased risk of depression recurrence include (Timonen and Liukkonen, 2008):

- Three or more episodes of major depression
- High prior frequency of recurrence

Table 4 Suicide rates ranked by occupation

Rank by rate of suicide	*Occupation (for men 2001–05)*	*Occupation (for women 2001–05)*
1	Beauticians and related occupations	Sports players
2	Coal miners	Artists
3	Merchant seafarers	Veterinarians
4	Laborers in building trades	Gardeners and groundswomen
5	Artists	Librarians
6	Window cleaners	Lorry drivers
7	Plasterers	Pharmacists
8	Market and street traders	Taxi drivers
9	Musicians	Photographers and camerawomen
10	Refuse and salvage workers	Prison service officers

Reproduced from Roberts, S.E., Jaremin, B., Lloyd, K., June 2013. High-risk occupations for suicide. Psychologie Medicale 43 (6), 1231–1240.

- An episode in the previous 12 months
- Residual symptoms during continuation treatment
- Severe episodes, for example, 'suicidality' and psychotic features
- Long previous episodes
- Relapse after drug discontinuation

Summary

Depression is the most common psychiatric disorder, and has a huge impact in the sufferer as well as for those close to him or her. It is ranked the third most prevalent moderate and severe disabling condition globally by the World Health Organization, and with time, this is predicted to worsen, to a ranking of second leading cause of disease burden by 2020 (WHO, 2004). It is clear that more funding and resources are needed to tackle a problem that affects so many people across the world, especially when treatment strategies are available and effective. It could literally be a case of life or death.

See also: Antidepressant Drugs; Bipolar (Manic-Depressive) Disorder; Cultural Psychiatry; Mania; Mental Disorders, Epidemiology of; Psychopharmacotherapy: Side Effects; Psychotherapy and Pharmacotherapy; Suicide.

Bibliography

Alexander, F., Selesnick, S.T., 1966. The History of Psychiatry. Harper and Row, New York.

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV), fourth ed. APA, Washington, DC.

Angst, J., Marneros, A., 2001. Bipolarity from ancient to modern times: conception, birth and rebirth. Journal of Affective Disorders 67, 3–19.

Baillarger, J., 1854. Note sur un genre de folie. Bulletin de la Academie Imperiale de Medicine (Paris) 19, 340–352.

Beck, Aaron T., Rush, A. John, Shaw, Brian F., Emery, Gary, 1979. Cognitive Therapy of Depression. The Guilford Press, New York, pp. 11.

Beck, A.T., Ward, C., Mendelson, M., 1961. Beck depression inventory (BDI). Archives of General Psychiatry 4 (6), 561–571.

Bhugra, D., Desai, M., Balwin, D., 1997. Focus groups: iplications for primary care and cultural psychiatry. Primary Care Psychiatry 5, 45–50.

Brown, G.W., Harris, T.O., 1978. Social Origins of Depression: A Study of Psychiatric Disorder in Women. Tavistock, London.

Dunlop, B.W., Nemeroff, C.B., 2007. The role of dopamine in the pathophysiology of depression. Archives of General Psychiatry 64 (3), 327–337.

Gelder, M.G., Andreasen, N.C., Lopez-Ibor Jr., J.L., Geddes, J.R. (Eds.), 2000. Acute episode as defined by ICD-10 and DSM-IV. In: New Oxford Textbook of Psychiatry. Oxford University Press, Oxford.

Gelder, M., Mayou, R., Cowen, P., 2001. Shorter Oxford Textbook of Psychiatry. Part 1. Page 303. Oxford University Press, Oxford.

Goodwin, F.K., Ghaemi, N.S., 2000. An introduction to and historical review of mood disorders. In: Gelder, M.G., Lopez-Ibor, J.J., Andreasen, N. (Eds.), New Oxford Textbook of Psychiatry, vol. 1. Oxford University Press, Oxford, pp. 677–682.

Halverson, J., Bienenfield, D., Updated April 23, 2014. Depression, Prognosis. Medscape. Available at: http://emedicine.medscape.com/article/286759-overview#aw2aab6b2b6 (last accessed 10.08.14.).

Higgins, E.T., 1987. Self-discrepancy; a theory relating self and affect. Psychological Review 94, 319–340.

Holmes, T.H., Rahe, R.H., 1967. The social Readjustment rating scale. Journal of Psychosomatic Research 11 (2), 213–218.

Kendler, K.S., Gardner, C.O., 1998. Boundaries of major depression: an evaluation of DSM-IV criteria. American Journal of Psychiatry 155, 172–177.

Kleinman, A., 1982. Neurasthenia and depression: a study of somatization and culture in China. Culture Medicine and Psychiatry 6 (2), 117–190.

Levinson, D.F., Nichols, W.E., 2014. Major Depression and Genetics. Department of Psychiatry and Behavioural Sciences. Available on line at: http://depressiongenetics.stanford.edu/mddandgenes.html (last accessed on 01.08.14.).

Lin, Tsung-Yi, 1989. Neurasthenia revisited: Its place in modern psychiatry. Culture Medicine and Psychiatry 13.2, 105–129.

Miller, M., Hemenway, D., 1999. The relationship between firearms and suicide: a review of the literature. Aggression and Violent Behavior: A Review Journal 14, 59–75.

Miller, M., Hemenway, D., 2001. Gun prevalence and the risk of suicide: a review. Harvard Health Policy Review 2, 29–37.

McGuffin, P., Katz, R., Watkins, S., Rutherford, J., 1996. A hospital-based twin register of the heritability of DSM-IV unipolar depression. Archives of General Psychiatry 53, 129–136.

National Confidential Inquiry Annual Report, 2014. University of Manchester. Available at: http://www.bbmh.manchester.ac.uk/cmhr/centreforsuicideprevention/nci/reports/Annualreport2014.pdf (last accessed on 02.08.14.).

National Institute for Health and Clinical Excellence (NICE). Clinical guideline 90, 2009. Depression in adults: the treatment and management of depression in adults. Available at: http://www.nice.org.uk/guidance/cg90/resources/guidance-depression-in-adults-pdf (last accessed 02.08.14.).

National Institute for Health and Clinical Excellence (NICE), August 2013. Clincal knowledge Summaries. Depression. Available at: http://cks.nice.org.uk/depression (last accessed 10.08.14.).

Puri, B.K., Hall, A.D., 2004. Revision Notes in Psychiatry, second ed. Arnold, London.

Roberts, S.E., Jaremin, B., Lloyd, K., June 2013. High-risk occupations for suicide. Psychologie Medicale 43 (6), 1231–1240.

Stevenson, Angus (Ed.), 2010. Oxford Dictionary of English, third ed. Oxford University Press.

Seligman, M.E.P., 1975. Helplessness: On Depression, Development, and Death. W.H. Freeman, San Francisco.

Simon, G.E., Goldberg, D.P., von Kroff, M., Ustun, T.B., 2002. Understanding cross-naitonal differences in depression prevalence. Psychologie Medicale 32, 585–594.

Sims, A., 1995. Symptoms in the Mind. An Introduction to Descriptive Psychopathology, second ed. W.B. Sauders Company Ltd, London.

Thapar, A., Hammerton, G., Collishaw, S., Potter, R., Rice, F., Harold, G., Craddock, N., Thapar, A., Smith, D.J., January 2014. Detecting recurrent major depressive disorder within primary care rapidly and reliably using short questionnaire measures. British Journal of General Practice 64, e31–e37.

Timonen, M., Liukkonen, T., February 23, 2008. Management of depression in adults. British Medical Journal 336 (7641), 435–439.

The World Health Organization, 2004. The Global Burden of Disease. Available at. http://www.who.int/healthinfo/global_burden_disease/GBD_report_2004update_full.pdf (accessed on 03.12.14.).

Wainwright, N.W.J., Surtees, P.G., Gilks, W.R., 1997. Diagnostic boundaries, reasoning and depressive disorder, I. Development of a probabilistic, morbidity model for public health psychiatry. Psychological Medicine 27, 835–845.
Whitfield, G., Williams, C., 2003. The evidence base for cognitive-behavioural delivery in busy clinical settings. Advances in Psychiatric Treatment 9, 21–30.
Williams, C.J., 2001. Overcoming Depression: A Five Areas Approach. Arnold, London.
World Health Organization, 1992. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines. WHO, Geneva.
Zigmond, A.S., Snaith, R.P., 1983. The hospital anxiety and depression scale. Acta Psychiatrica Scandinavica 67 (6), 361–370.

Relevant Websites

http://www.allaboutdepression.com/.
http://www.mentalhealthamerica.net/conditions/depression – Mental Health America.
http://www.nimh.nih.gov/health/topics/depression/index.shtml – National Institute of Mental Health: Depression.
www.nice.org.uk – National Institute for Health and Clinical Excellence.
http://www.rcpsych.ac.uk/healthadvice/problemsdisorders/depression.aspx – Royal College of Psychiatrists Depression leaflet.
http://www.patient.co.uk/health/depression-leaflet.

Depression, Pessimism, and Health

Charles S Carver and Emily G Lattie, University of Miami, Coral Gables, FL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C.S. Carver, volume 5, pp. 3516–3522, © 2001, Elsevier Ltd.

Abstract

Depression is a syndrome involving sadness and doubt, loss of motivation, and the ability to experience enjoyment. Pessimism (the negative pole of the optimism dimension) is a trait consisting of the generalized expectancy of bad outcomes in important life domains; the term hopelessness conveys a similar sense of anticipating bad outcomes. All these characteristics have been implicated as contributors to adverse health outcomes. Pessimism has been related to low persistence at health promotion efforts and to activities that reflect disengagement from life. Pessimism has also been linked to adverse reactions in certain medical events (e.g., coronary artery bypass surgery, heart transplantation) and to cardiovascular disease. Depression has long been related to elevated likelihood of subsequent cardiac death after initial myocardial infarction, and to heart disease more generally. Both pessimism and depression have also been linked to inflammation, which may be a precursor to several health problems. There is evidence that some of these effects are larger among younger than older persons. Future research must focus more on the pathways by which these variables create their influence.

Depression is a psychological syndrome encompassing a variety of symptoms, including loss of energy and motivation; loss of the capacity to experience happiness or enjoyment; feelings of sadness, despair, and hopelessness; and sometimes a sense of guilt, shame, or worthlessness. Depression is a debilitating problem in its own right for people who experience it. Depression can also have serious adverse implications for physical health and well-being.

Depression is an episodic condition. After first manifestation it can be long-lasting or repeated, but it does not usually last for the person's entire lifetime. On the other hand, there are also personality traits that have strong overtones of depression. Consider the trait dimension of optimism versus pessimism. Optimists are people who expect good things to happen to them; pessimists expect bad things to happen. Pessimism thus resembles aspects of depression. Unlike depression, however, pessimism is an enduring trait. The dimension of optimism versus pessimism has been studied a good deal in Health Psychology. Some of that research has found that this aspect of personality also relates to physical well-being.

The concept of hopelessness has links to both depression and pessimism. Sometimes hopelessness is used to refer to a state and sometimes it is used to refer to a personality trait. One might view the concept of hopelessness as a conceptual bridge between the (usually temporary) state of depression and longer lasting traits such as pessimism. People experiencing depression often feel hopeless about the future; people who are pessimistic are more broadly and consistently lacking in hope for the future.

This article describes some of the research that links depression and pessimism to health problems. As background to that research, the next section describes a conceptual basis for the hypothesis that these variables may be involved in health.

Pessimism, Hopelessness, and Disengagement

The concepts of optimism and pessimism are rooted in centuries of folk wisdom about effects of people's expectations for the future. Scientific research on optimism and pessimism also is rooted in the principle of expectations for the future. By taking expectancies as a starting point, this research makes a connection to more than a century of psychological theory (Carver et al., 2010; Carver and Scheier, 2014). That is, the expectancy construct relates to a broad collection of theories of motivation and action called expectancy–value theories. This body of theory suggests a basis for examining how the traits of optimism and pessimism are expressed in motivation, behavior, and feelings. These various expressions of optimism and pessimism turn out to also represent a set of potential pathways by which this personality disposition may relate to health.

Expectancy–value models begin with the idea that behavior is aimed at attaining desired goals (values). Without a valued goal, there is no reason to act. People also have varying degrees of confidence or doubt about being able to attain the goal. Without confidence, again there will be no action. Only with sufficient confidence do people exert efforts and remain engaged in those efforts. These principles pertain to specific values and focused confidence, and they also apply to the more general traits of optimism and pessimism. In optimism and pessimism the 'confidence' involved is simply broader in scope.

These expectancy–value principles provide the basis for many predictions about optimists and pessimists. When confronting a challenge, optimists should be confident and persistent, even if progress is difficult or slow. Pessimists should be more doubtful and hesitant. This divergence should be further amplified under conditions of adversity. Optimists are likely to believe that the adversity can be handled successfully, whereas pessimists are likely to anticipate disaster. These differences in orientation can lead to differences in behavioral persistence, differences in efforts to take precautions, differences in many decisions and actions that pertain to health risks.

People's behavioral responses to adversity are important, but overt behavior is not the only response that occurs when people confront adversity. People also experience emotions in such situations. Difficulties elicit a mixture of feelings, and variables such as optimism influence the balance among the

feelings. Again, predictions are easily derived from theory. Optimists expect good outcomes, even when things are difficult. This confidence should yield a more positive mix of feelings, for example, enthusiasm, excitement, and challenge. Pessimists expect bad outcomes. This is related to more negative feelings, such as anxiety, sadness, or despair. There is good evidence for such emotional differences between optimists and pessimists (Carver et al., 2010; Segerstrom, 2006).

The experiences of intense distress, anxiety, hopelessness, and despair have physical concomitants, including changes in the neuroendocrine and cardiovascular systems. There is some evidence to suggest that these physiological changes play an important role in health outcomes, although research on the precise pathways of influence is at an early stage. Such processes may influence a person's probability of getting a disease in the first place, they may influence the progression of diseases, and they may even influence mortality. There is a good deal of evidence that personality differences in optimism versus pessimism relate to differences in health-related outcomes (Rasmussen et al., 2009).

As stated earlier, depression has conceptual connections to pessimism, that is, a sense of negativity about the future. This negativity about the future means that the reasoning just outlined about expectancy–value theories also applies to the syndrome of depression. Depression should be associated with relative lack of persistence in trying to maintain health, and it should also be associated with the sorts of physiological responses to challenges and adversity that pessimism is associated with. The result in each case should be a potential for adverse influences on health.

Pessimism and Health

Health is a complex phenomenon with many contributors (for a broader discussion see Carver and Vargas, 2011). Health is partly a matter of the presence versus absence of disease. Health is partly ensuring that your body is in a condition that will help it fight off any potential intruders. Health stems partly from engaging in health-promoting activities and avoiding health-damaging activities. Health also comes partly from not having intense and potentially damaging physiological arousal states too often and for too prolonged a period. There thus are several ways in which either traits such as pessimism or states such as depression can influence health.

Pessimism and Health-Related Actions

Sometimes being persistent in goal-related efforts is an important determinant of health and well-being. This would be the case when persistence is expressed as active efforts at health promotion, that is, to be proactive in staying healthy. A reasonable hypothesis is that optimists, who are confident about attaining and maintaining the goal of good health, should be more likely than pessimists to continue the active pursuit of health-promoting activities. Evidence from several studies supports this hypothesis.

One study looking at potential individual differences in health promotion followed a group of heart patients who were involved in a cardiac rehabilitation program (Shepperd et al., 1996). Optimism among participants was related to increases in exercise levels across the study period. Optimism also related to greater success in lowering levels of saturated fat, lowering body fat, and lowering an index of global coronary risk.

Another study followed coronary artery bypass patients during the period immediately after surgery (Scheier et al., 1989) and reexamined them 5 years later (described in Scheier and Carver, 1992). The early part of the project focused mostly on patients' responses to the surgery as an event in their lives; the later part of the project investigated their lifestyles 5 years postsurgery. In the early part of the project, optimists gave several indications of focusing on getting back on their feet quickly. They were viewed by the hospital staff as recuperating faster, and they reported concentrating on the rehabilitation process. Pessimists, in contrast, were more likely to be avoiding even thinking about the future. Five years later, optimists were more likely than pessimists to be taking vitamins, to be eating low-fat lunches, and to be enrolled in a cardiac rehabilitation program. Thus, both initially and later on, pessimists were less persistent than optimists in a wide range of health-promoting behaviors.

Another health risk that is related to behavior, though in a different manner, is HIV infection. Infection can occur from engaging in sexual activities with partners who are not known to be disease free. By taking steps to avoid such sexual practices – for example, avoiding sex with unknown partners – people reduce their risk of infection. One study of HIV-negative gay men found that optimists reported having fewer anonymous partners than pessimists (Taylor et al., 1992). This finding suggests that optimists were making efforts to reduce their risk, thereby safeguarding their health.

Another example comes from a study of hospital employees at risk for skin cancer (Friedman et al., 1995). Optimists were significantly more likely than pessimists to report the intentions to engage in skin cancer-relevant health prevention behaviors (e.g., regular sunscreen use). In addition, among those identified at screening as having suspicious lesions, those who were more optimistic were also more likely to comply with recommended follow-up care.

Other studies have examined health-related habits reported by groups of people with no particular salient health concerns. At least two such projects found that optimists reported more health-promoting behaviors than did pessimists (Robbins et al., 1991; Steptoe et al., 1994). In sum, these results suggest that optimism is associated with behaviors aimed at promoting health and well-being and reducing health risk.

It is important to note, however, that while optimists definitely do not stick their heads in the sand and ignore threats to their well-being, neither do they always jump immediately into action. Rather, they display a pattern of attending to risks selectively, reacting to risks that are clearly applicable to them and also are related to potentially serious health problems (Aspinwall and Brunhart, 1996). If the potential problem is minor, or if it is unlikely to bear on them personally, they do not display elevated vigilance. This fits the idea that optimists scan their surroundings for threats to their well-being, but save their active coping efforts for threats that are truly meaningful.

Pessimism and Disengagement

The hypothesis discussed in the preceding section derives from the idea that optimism leads to continued effort and pessimism leads to giving up. In that section the focus was on the continued effort. What about the giving up? Several kinds of health-relevant behaviors seem to reflect a giving up tendency. One of them is alcohol consumption, which at high levels is a health risk. Excessive alcohol consumption is often seen as occurring in the service of an escape from one's problems. If so, it seems reasonable that pessimists should be more vulnerable to alcohol abuse than optimists.

Results from at least two studies fit this picture. One study of women with family history of alcoholism found that pessimists were more likely than optimists to report drinking problems (Ohannessian et al., 1993). Another study (Strack et al., 1987) examined people who had already been treated for alcoholism and were now entering an aftercare program. This study found that pessimists were more likely to drop out of the aftercare program and to return to drinking than optimists. These two studies converge in showing that pessimists display a form of disengagement – alcohol consumption – more than optimists.

People can give up in many ways, of course. Alcohol dulls awareness of failures and problems. People can also turn their backs on their problems by distracting themselves. Even sleeping can help people escape from situations they do not want to face. Sometimes, though, giving up is more complete. Sometimes people give up not just on specific goals, but on all the goals that define their lives, by killing themselves. Pessimism has been found in several studies to be a key indicator of suicide risk (Beck et al., 1985; Fawcett et al., 1987; Petrie and Chamberlain, 1983).

Physiological Reactivity and Disease Susceptibility

The examples just described all concern how people varied in their efforts to protect themselves against health risks, or how hard they tried to maintain good health, or the extent to which they engaged in behaviors that may reflect a disengagement from such efforts. Such factors are important contributors to disease susceptibility. That is, people who take steps to avoid health risks will generally be healthier in the long run than people who do not. But there are also more direct associations, cases in which people's susceptibility to diseases is not so much a function of their overt actions as it is of their body's autonomic reactions.

This pathway to disease follows from overuse of the body's normal reactions to stress. Many kinds of experiences produce involuntary physiological changes. Although these changes are generally functional and adaptive in the short run, some of them can also disrupt important health-protective functions. In such cases, a person's stress and coping processes affect physiological reactions in a way that can render the person more vulnerable to promotion or progression of disease. This is particularly likely to be the case if the stress reaction is prolonged or repeated.

This sort of effect is generally assumed to begin with the experience of emotional distress. Intense emotional states are associated with a variety of cardiovascular responses and neuroendocrine and immune changes. Given the broad array of responses that are associated with negative emotions, there again are diverse specific potential influences on health. Some of them follow from cardiovascular responses per se, others follow from endocrine responses.

As described earlier, optimists and pessimists differ in the patterning of emotions they experience when confronting adversity. This means that – given identical experiences – they should experience different degrees and durations of physiological changes that in the long run can create problems. An example of a study that seems to support this notion was reported by Räikkönen et al. (1999). This study examined associations of pessimism with blood pressure elevations. This study involved 3 days of ambulatory monitoring of blood pressure during normal activities. Results indicated that pessimists were more likely to have elevations in blood pressure during those activities than optimists. This study does not provide evidence of disease per se, but it suggests that further examination of the role of pessimism in the development of hypertension is warranted.

The role of stress on neuroendocrine and immune functioning has been extensively examined in recent decades. Research shows that psychological stress can lead to dysregulation of the neuroendocrine system (i.e., disruption of typical ranges of regulation) both through direct physiological pathways and also through behavioral changes intended to aid in survival (Tsigos and Chrousos, 2002). Stress and dysregulated neuroendocrine functioning can also prompt immune responses, which may result in an increase in inflammation (Chrousos, 1995). Elevated inflammation is an important predictor of heart disease and several other disorders (Cesari et al., 2003; Dentino et al., 1999; Luc et al., 2003; Miller et al., 2002).

As people who are more pessimistic may experience more psychological stress, with greater feelings of anxiety, sadness, or despair, it follows that pessimistic people may also have higher circulating markers of inflammation. One large-scale study looked at over 5000 men and women from the Multi-Ethnic Study of Atherosclerosis, all of whom were confirmed to have no history of clinical cardiovascular disease at initial measurement. Nonetheless, pessimism was associated with higher levels of several key markers of inflammation (Roy et al., 2010).

It is hard to study initial vulnerability to disease, because doing so requires recruiting participants who are all completely healthy, and waiting until some of them become ill. An alternative strategy is to recruit persons who are all in the same category of risk for a disease, or are at the same early stage of a disease, and following them over time. One study that used that strategy examined women's carotid intima-media thickness (Matthews et al., 2004). Carotid intima-media thickness is an index of atherosclerosis in the carotid artery, a physical marker of the development of heart disease. In this study, intima-media thickness was measured among middle-aged women at baseline and again 3 years later. Greater pessimism at the initial assessment predicted increases in intima-media thickness at follow-up. Optimists experienced almost no increase over the 3-year period.

A good deal of research has examined differences in well-being that follow the initial diagnosis or treatment of disease. Some of this work looks at responses to medical procedures.

For example, the study of men undergoing coronary bypass surgery (described a little earlier) by Scheier et al. (1989) included measures of the patients' progress through the surgery itself. Two kinds of indirect evidence suggested that pessimists fared more poorly than optimists even on the operating table. Specifically, pessimists were more likely to display two markers during surgery that are widely taken as indicators of myocardial infarction (MI). Thus, pessimists may have been at greater physical risk during the bypass surgery itself.

Scheier et al. (1999) collected data on a second sample of bypass patients. The first analyses of data from that project focused on a problem that often arises after major surgery: the need for rehospitalization, due either to infection from the surgery or to complications from the disease underlying the surgery. Scheier et al. (1999) found that optimistic persons were significantly less likely to be rehospitalized either for problems related to postsurgical infection or for problems related to the coronary artery disease, even after controlling for medical variables and other personality measures.

Another research group has explored how positive expectations before heart transplant surgery related to later medical adherence and health (Leedham et al., 1995). In this case, optimism was assessed at several levels of abstraction. Patients completed measures of optimism prior to surgery; they were then evaluated at discharge and at 3 and 6 months postsurgery. Initial scales assessed patients' confidence about the efficacy of treatment, their expectations about their future health and survival, and broader expectations for the future. Confidence about the future predicted better adherence to the postoperative medical regimen, and strongly predicted nurses' ratings of the patients' physical health at 6 months after surgery. There was also a tendency for positive expectations to predict longer delays before development of infection (a near-universal side effect of heart transplantation).

Perhaps the most compelling study to date on optimism and cardiovascular disease grew out of the Women's Health Initiative, a large-scale project designed to study changes in and predictors of quality of life, chronic disease, morbidity, and mortality among women across America. Using participants in this project, Tindale et al. (2009) studied over 95 000 women across an 8-year period. All of the women were free of cancer and cardiovascular disease at study entry. The results were clear and striking. Optimists were less likely than pessimists to develop coronary heart disease (CHD) and to die from CHD-related causes, and had lower total mortality due to all causes, across the 8 years of study. The advantage due to optimism ranged from 9% for incident cases of CHD to 30% for CHD-related mortality.

Associations between Depression and Health Outcomes

Most of the foregoing has dealt with pessimism or hopelessness. What about depression and health? Links between depression and adverse health outcomes have been examined extensively in recent years (Lustman et al., 2000; Rutledge et al., 2006; Seifert et al., 2012). There is, of course, a great deal of evidence that depression follows from adverse health events. Depression also cooccurs with a variety of serious medical conditions of widely varying types (e.g., Freedland et al., 2003; Gavard et al., 1993; Kessler et al., 2008; Minden and Schiffer, 1990; Thombs et al., 2006; Sadovnick et al., 1996). In most of these cases, of course, the depression is most readily interpreted as a reaction to the physical illness, rather than a contributor. But that does not mean that depression is therefore unimportant to illness.

Again, there appear to be both behavioral and medical pathways to poorer health. Recall that one consequence of pessimism is less effort and persistence in health-promoting behaviors. The same is true for depression. Increased depressive symptoms often result in poorer self-care, including missed medical appointments and failure to take medicines as prescribed. Depression has also been linked to a variety of behaviors indicating an unhealthy lifestyle, such as physical inactivity, poor diet, and elevations in use of tobacco (Bonnet et al., 2005; Steptoe et al., 1997).

Depressive symptoms can also be a barrier to self-care behaviors aimed at repairing damage already done. For example, Anda et al. (1990) noted that depressive symptoms were negatively related to the ability to quit smoking over a 9-year period. A large meta-analysis found a substantial relationship between depression and medical noncompliance (DiMatteo et al., 2000). Thus, some proportion of the poorer health among depressed persons may be attributed to simply not following the doctor's orders.

These failures of adequate self-care can create disease-specific outcomes such as poor glycemic control in diabetes (Bane et al., 2006; Lustman et al., 2000), and a more general vulnerability to whatever challenges are represented by the illness. Thus, depression can act through behavioral mechanisms to make an existing illness worse.

Links between depression and subsequent health are best documented for cardiac-related events. For example, Ladwig et al. (1991) gathered psychological data from over 500 male survivors of MIs shortly after the MI, including a measure of depression. By 6-month follow-up, 12 patients had died from new cardiac events. Death was predicted by post-MI depression.

Stronger findings relating depression to post-MI mortality were reported by Frasure-Smith et al. (1993). Depression was assessed by interview during the week after an MI. In addition, a large number of medical variables were collected. By 6 months after the initial interview, 12 cardiac-related deaths had occurred, and depression was a significant predictor. Patients meeting criteria for clinical depression after their MI were five times more likely than nondepressed patients to have died by the 6-month follow-up.

Frasure-Smith et al. (1999) have more recently reported a larger and longer term study of men and women who had been hospitalized for MI. Both men and women who were depressed while in the hospital were three times as likely to die of cardiac causes than were patients who were not depressed.

It is tempting to view this evidence as reflecting physical manifestations of a loss of desire to continue living. Perhaps the depressed MI patients in the studies reviewed above simply gave up the struggle to go on. This would represent a kind of disengagement response, albeit one that is quite different from the overt actions represented by suicide or alcohol abuse.

Depressive symptoms that emerge after the onset of physical illness generally reflect the stress of adapting to the illness.

In some cases, though, the disease itself may induce depression by physiological pathways. For example, approximately half of persons with multiple sclerosis (MS) experience at least one major depressive episode (Minden and Schiffer, 1990; Sadovnick et al., 1996). Structural brain changes due to the MS activity contribute to the development of the depressive symptoms (Pujol et al., 1997; Bakshi et al., 2000). Here is a case in which effects go both ways. That is, in one study reductions in depression by psychotherapy led to reductions in MS-related disability and fatigue (Mohr et al., 2007).

It is no surprise that depression can follow from illness. But evidence suggests that depression can exacerbate the illness. Indeed, recent research suggests that depression can be an independent risk factor for the initial development of certain kinds of disease. The best evidence for relationships of this kind is in cardiovascular disease. One meta-analysis of 11 longitudinal studies revealed that clinical depression predicts the subsequent development of CHD in initially healthy people (Rugulies, 2002). Depressive symptoms also increase the risk of ischemic stroke (Majed et al., 2012). Depression appears to be a stronger risk factor for mortality in patients with CHD than it is for the general population (Barth et al., 2004). Indeed, Rozanski et al. (2005) noted that depression is as strong an independent predictor of cardiac events as are traditional risk factors such as high blood pressure and tobacco use.

Several cross-sectional studies have found that depressed persons have elevated levels of inflammatory markers that themselves are predictive of developing cardiovascular disease (Cesari et al., 2003; Dentino et al., 1999; Luc et al., 2003; Miller et al., 2002). More recently, depressive symptoms were found to lead to inflammation in healthy adults (Stewart et al., 2009) and in healthy adolescents (Copeland et al., 2012). These findings suggest that inflammation may be a pathway by which depression fosters development of cardiovascular disease independent of other traditional risk factors, and that the pathway develops throughout adolescence. It is important to note that these studies examined samples of healthy individuals rather than those already diagnosed with cardiovascular disease.

A meta-analysis of studies on depression and type 2 diabetes found that depression is also associated with a 60% increased risk of type 2 diabetes (Mezuk et al., 2008). Given the role of inflammation in insulin resistance (Shoelson et al., 2006), this raises questions about whether these inflammatory mechanisms may contribute to relationships between depression and health outcomes in a variety of other medical conditions.

Potential Moderators

The focus of this article has been on the direct associations between health and the psychological variables of depression and pessimism. However, some attention should also be paid to limitations on – or moderators of – such effects. One limitation may be the seriousness of the disease. It is possible that the effects of the psychological variables under discussion (and others) are greater among persons with disease that is less severe than among persons whose disease is advanced. That is, perhaps advancing disease affords less opportunity for psychological variables to gain leverage over the impact of the disease.

Another potential moderator is the age of the person whose health is under study. In one earlier review of the literature, Scheier and Bridges (1995) noted that variables such as personality and depression sometimes had stronger influences on health among younger than older persons. Indeed, every age difference they encountered was one in which stronger associations between personality and health outcome occurred among younger persons. This observation led Scheier and Bridges to suggest that biological factors may become more important in determining disease outcomes as people grow older, and psychological factors correspondingly less important.

Closing Comment

This article reviewed some associations that have been observed between pessimism, depression, and health. Evidence of associations between these psychological variables and health is fairly robust. On the other hand, it would be wrong to conclude that the mechanisms underlying these effects are well understood. In some cases the mechanism is obvious: persisting in a health-promoting exercise activity, taking steps to control one's weight, acting to lower one's health risks – all these are behavioral ways in which optimism and hopefulness can enhance health outcomes through overt behavior. In other cases, though, the mechanism is far more speculative, involving potential neuroendocrine and immune responses as mediators. The role of such mediators is presently under active investigation, but there remain many uncertainties. Despite the many questions remaining, though, the pattern seems consistent with a picture in which continued optimistic engagement with life goals promotes better health outcomes, and in which the loss of such engagement – giving up – promotes poorer outcomes.

See also: Cardiac Disease, Coping with; Chronic Illness, Psychosocial Coping with; Co-morbidity of Mental and Physical Conditions; Control Beliefs: Health Perspectives; Coronary Heart Disease: Psychosocial Aspects; Depression; Explanatory Style and Health; Mental Health; Personality and Health; Psychoneuroimmunology; Rehabilitation (Cardiac): Psychosocial Aspects; Religion and Health: Clinical Considerations and Applications; Resilience; Self-Efficacy and Health; Stress, Coping and Health.

Bibliography

Anda, R.F., Williamson, D.F., Escobedo, L.G., Mast, E.E., Giovino, G.A., Remington, P.L., 1990. Depression and the dynamics of smoking. Journal of the American Medical Association 264, 1541–1545.

Aspinwall, L.G., Brunhart, S.N., 1996. Distinguishing optimism from denial: optimistic beliefs predict attention to health threats. Personality and Social Psychology Bulletin 22, 993–1003.

Bakshi, R., Czarnecki, D., Shaikh, Z.A., Priore, R.L., Janardhan, V., Kaliszky, Z., Kinkel, P.R., 2000. Brain MRI lesions and atrophy are related to depression in multiple sclerosis. Neuroreport 11, 1153–1158.

Bane, C., Hughes, C.M., McElnay, J.C., 2006. The impact of depressive symptoms and psychosocial factors on medication adherence in cardiovascular disease. Patient Education Counseling 60, 187–193.

Barth, J., Schumacher, M., Herrmann-Lingen, C., 2004. Depression as a risk factor for mortality in patients with coronary heart disease: a meta-analysis. Psychosomatic Medicine 66, 802–813.

Beck, A.T., Steer, R.A., Kovacs, M., Garrison, B., 1985. Hopelessness and eventual suicide: a 10-year prospective study of patients hospitalized with suicidal ideation. American Journal of Psychiatry 142, 559–563.

Bonnet, F., Irving, K., Terra, J.L., Nony, P., Berthezène, F., Moulin, P., 2005. Anxiety and depression are associated with unhealthy lifestyle in patients at risk of cardiovascular disease. Atherosclerosis 178, 339–344.

Carver, C.S., Scheier, M.F., 2014. Dispositional optimism. Trends in Cognitive Sciences 18, 293–299. http://dx.doi.org/10.1016/j.tics.2014.02.003.

Carver, C.S., Scheier, M.F., Segerstrom, S.C., 2010. Optimism. Clinical Psychology Review 30, 879–889.

Carver, C.S., Vargas, S., 2011. Stress, coping, and health. In: Friedman, H.S. (Ed.), The Oxford Handbook of Health Psychology (165–191). Oxford University Press, New York.

Cesari, M., Penninx, B.W., Newman, A.B., Kritchevsky, S.B., Nicklas, B.J., Sutton-Tyrrell, K., Pahor, M., 2003. Inflammatory markers and onset of cardiovascular events. Circulation 108, 2317–2322.

Chrousos, G.P., 1995. The hypothalamic–pituitary–adrenal axis and immune-mediated inflammation. New England Journal of Medicine 332, 1351–1363.

Copeland, W.E., Shanahan, L., Worthman, C., Angold, A., Costello, E.J., 2012. Cumulative depression episodes predict later C-reactive protein levels: a prospective analysis. Biological Psychiatry 71, 15–21.

Dentino, A.N., Pieper, C.F., Rao, K.M.K., Currie, M.S., Harris, T., Blazer, D.G., Cohen, H.J., 1999. Association of interleukin-6 and other biologic variables with depression in older people living in the community. Journal of the American Geriatrics Society 47, 6–11.

DiMatteo, M.R., Lepper, H.S., Croghan, T.W., 2000. Depression is a risk factor for noncompliance with medical treatment: meta-analysis of the effects of anxiety and depression on patient adherence. Archives of Internal Medicine 160, 2101–2107.

Fawcett, J., Scheftner, W., Clark, D., Hedeker, D., Gibbons, R., Coryell, W., 1987. Clinical predictors of suicide in patients with major affective disorders: a controlled prospective study. American Journal of Psychiatry 144, 35–40.

Frasure-Smith, N., Lesprance, F., Juneau, M., Talajic, M., Bourassa, M.G., 1999. Gender, depression, and one-year prognosis after myocardial infarction. Psychosomatic Medicine 61, 26–37.

Frasure-Smith, N., Lesprance, F., Talajic, M., 1993. Depression following myocardial infarction. Journal of the American Medical Association 270, 1819–1825.

Fredrikson, M., Matthews, K.A., 1990. Cardiovascular responses to behavioral stress and hypertension: a meta-analytic review. Annals of Behavioral Medicine 12, 30–39.

Freedland, K.E., Rich, M.W., Skala, J.A., Carney, R.M., Davila-Roman, V.G., Jaffe, A.S., 2003. Prevalence of depression in hospitalized patients with congestive heart failure. Psychosomatic Medicine 65, 119–128.

Friedman, L.C., Weinberg, A.D., Webb, J.A., Cooper, H.P., Bruce, S., 1995. Skin cancer prevention and early detection intentions and behavior. American Journal of Preventive Medicine 11, 59–65.

Gavard, J.A., Lustman, P.J., Clouse, R.E., 1993. Prevalence of depression in adults with diabetes. An epidemiological evaluation. Diabetes Care 16, 1167–1178.

Kessler, R.C., Gruber, M., Hettema, J.M., Hwang, I., Sampson, N., Yonkers, K.A., 2008. Co-morbid major depression and generalized anxiety disorders in the National Comborbity Survey follow-up. Psychological Medicine 38, 365–374.

Krantz, D.S., McCeney, M.K., 2002. Effects of psychological and social factors on organic disease: a critical assessment of research on coronary heart disease. Annual Review of Psychology 53, 341–369.

Ladwig, K.H., Kieser, M., Konig, J., 1991. Affective disorders and survival after acute myocardial infarction. European Heart Journal 12, 959–964.

Leedham, B., Meyerowitz, B.E., Muirhead, J., Frist, W.H., 1995. Positive expectations predict health after heart transplantation. Health Psychology 14, 74–79.

Luc, G., Bard, J.M., Juhan-Vague, I., Ferrieres, J., Evans, A., Amouyel, P., Ducimetiere, P., 2003. C-reactive protein, interleukin-6, and fibrinogen as predictors of coronary heart disease the PRIME study. Arteriosclerosis, Thrombosis, and Vascular Biology 23, 1255–1261.

Lustman, P.J., Anderson, R.J., Freedland, K.E., de Groot, M., Carney, R.M., Clouse, R.E., 2000. Depression and poor glycemic control: a meta-analytic review of the literature. Diabetes Care 23, 934–942.

Majed, B., Arveiler, D., Bingham, A., Ferrieres, J., Ruidavets, J.B., Montaye, M., Empana, J.P., 2012. Depressive symptoms, a time-dependent risk factor for coronary heart disease and stroke in middle-aged men: the PRIME study. Stroke 43, 1761–1767.

Matthews, K.A., Räikkönen, K., Sutton-Tyrrell, K., Kuller, L.H., 2004. Optimistic attitudes protect against progression of carotid atherosclerosis in healthy middle-aged women. Psychosomatic Medicine 66, 640–644.

Miller, G.E., Cohen, S., Ritchey, A.K., 2002. Chronic psychological stress and the regulation of pro-inflammatory cytokines: a glucocorticoid-resistance model. Health Psychology 21, 531.

Mezuk, B., Eaton, W.W., Albrecht, S., Golden, S.H., 2008. Depression and type 2 diabetes over the lifespan: a meta-analysis. Diabetes Care 31, 2383–2390.

Minden, S.L., Schiffer, R.B., 1990. Affective disorders in multiple sclerosis. Review and recommendations for clinical research. Archives of Neurology 47, 98–104.

Mohr, D.C., Hart, S., Vella, L., 2007. Reduction in disability in a randomized controlled trial of telephone-administered cognitive-behavioral therapy. Health Psychology 26, 554–563.

O'Donovan, A., Lin, J., Dhabar, F.S., Wolkowitz, O., Tillie, J.M., Blackburn, E., Epel, E., 2009. Pessimism correlates with leukocyte telomere shortness and elevated interleukin-6 in post-menopausal women. Brain, Behavior, and Immunity 23, 446–449.

Ohannessian, C.M., Hesselbrock, V.M., Tennen, H., Affleck, G., 1993. Hassles and uplifts and generalized outcome expectancies as moderators on the relation between a family history of alcoholism and drinking behaviors. Journal of Studies on Alcohol 55, 754–763.

Petrie, K., Chamberlain, K., 1983. Hopelessness and social desirability as moderator variables in predicting suicidal behavior. Journal of Consulting and Clinical Psychology 51, 485–487.

Pujol, J., Bello, J., Deus, J., Marti-Vilalta, J.L., Capdevila, A., 1997. Lesions in the left arcuate fasciculus region and depressive symptoms in multiple sclerosis. Neurology 49, 1105–1110.

Räikkönen, K., Matthews, K.A., Flory, J.D., Owens, J.F., Gump, B.B., 1999. Effects of optimism, pessimism, and trait anxiety on ambulatory blood pressure and mood during everyday life. Journal of Personality and Social Psychology 76, 104–113.

Rasmussen, H.N., Scheier, M.F., Greenhouse, J.B., 2009. Optimism and physical health: a meta-analytic review. Annals of Behavioral Medicine 37, 239–256.

Robbins, A.S., Spence, J.T., Clark, H., 1991. Psychological determinants of health and performance: the tangled web of desirable and undesirable characteristics. Journal of Personality and Social Psychology 61, 755–765.

Roy, B., Diez-Roux, A.V., Seeman, T., Ranjit, N., Shea, S., Cushman, M., 2010. Association of optimism and pessimism with inflammation and hemostasis in the Multi-Ethnic Study of Atherosclerosis (MESA). Psychosomatic Medicine 72, 134–140.

Rozanski, A., Blumenthal, J.A., Davidson, K.W., Saab, P.G., Kubzansky, L., 2005. The epidemiology, pathophysiology, and management of psychosocial risk factors in cardiac practice: the emerging field of behavioral cardiology. Journal of the American College of Cardiology 45, 637–651.

Rugulies, R., 2002. Depression as a predictor for coronary heart disease—A review and meta-analysis. American Journal of Preventive Medicine 23, 51–61.

Rutledge, T., Reis, V.A., Linke, S.E., Greenberg, B.H., Mills, P.J., 2006. Depression in heart failure—A meta-analytic review of prevalence, intervention effects, and associations with clinical outcomes. Journal of the American College of Cardiology 48, 1527–1537.

Sadovnick, A.D., Remick, R.A., Allen, J., Swartz, E., Yee, I.M., Eisen, K., Farquhar, R., Hashimoto, S.A., Hooge, J., Kastrukoff, L.F., Morrison, W., Nelson, J., Oger, J., Paty, D.W., 1996. Depression and multiple sclerosis. Neurology 46, 628–632.

Scheier, M.F., Bridges, M.W., 1995. Person variables and health: personality predispositions and acute psychological states as shared determinants for disease. Psychosomatic Medicine 57, 255–268.

Scheier, M.F., Carver, C.S., 1992. Effects of optimism on psychological and physical well-being: theoretical overview and empirical update. Cognitive Therapy and Research 16, 201–228.

Scheier, M.F., Matthews, K.A., Owens, J.F., Magovern, G.J., Lefebvre, R.C., Abbott, R.A., Carver, C.S., 1989. Dispositional optimism and recovery from coronary artery bypass surgery: the beneficial effects on physical and psychological well-being. Journal of Personality and Social Psychology 57, 1024–1040.

Scheier, M.F., Matthews, K.A., Owens, J.F., Schulz, R., Bridges, M.W., Magovern Sr., G.J., Carver, C.S., 1999. Optimism and rehospitalization following coronary artery bypass graft surgery. Archives of Internal Medicine 159, 829–835.

Segerstrom, S.C., 2006. Breaking Murphy's Law. Guilford, New York.

Seifert, C.L., Poppert, H., Sander, D., Feurer, R., Etgen, T., Ander, K.H., Purner, K., Bronner, M., Sepp, D., Kehl, V., Forstl, H., Bickell, H., 2012. Depressive symptoms and the risk of ischemic stroke in the elderly-influence of age and sex. PLoS One 7, e50803.

Shepperd, J.A., Maroto, J.J., Pbert, L.A., 1996. Dispositional optimism as a predictor of health changes among cardiac patients. Journal of Research in Personality 30, 517–534.

Shoelson, S.E., Lee, J., Goldfine, A.B., 2006. Inflammation and insulin resistance. The Journal of Clinical Investigation 116, 1793–1801.

Smith, T.W., Ruiz, J.M., 2002. Coronary heart disease. In: Christensen, A.J., Antoni, M.H. (Eds.), Chronic Physical Disorders: Behavioral Medicine's Perspective. Blackwell, Oxford, UK, pp. 83–111.

Steptoe, A., Wardle, J., Vinck, J., Tuomisto, M., Holte, A., Wichstrm, L., 1994. Personality and attitudinal correlates of healthy lifestyles in young adults. Psychology and Health 9, 331–343.

Steptoe, A., Wardle, J., Fuller, R., Holte, A., Justo, J., Sanderman, R., Wichstrøm, L., 1997. Leisure-time physical exercise: prevalence, attitudinal correlates, and behavioral correlates among young Europeans from 21 countries. Preventive Medicine 26, 845–854.

Stewart, J.C., Rand, K.L., Muldoon, M.F., Kamarck, T.W., 2009. A prospective evaluation of the directionality of the depression-inflammation relationship. Brain, Behavior, and Immunity 23, 936–944.

Strack, S., Carver, C.S., Blaney, P.H., 1987. Predicting successful completion of an aftercare program following treatment for alcoholism: the role of dispositional optimism. Journal of Personality and Social Psychology 53, 579–584.

Taylor, S.E., Kemeny, M.E., Aspinwall, L.G., Schneider, S.G., Rodriguez, R., Herbert, M., 1992. Optimism, coping, psychological distress, and high-risk sexual behavior among men at risk for Acquired Immunodeficiency Syndrome (AIDS). Journal of Personality and Social Psychology 63, 460–473.

Thombs, B.D., Bass, E.B., Ford, D.E., Stewart, K.J., Tsilidis, K.K., Patel, U., Fauerbach, J.A., Bush, D.E., Ziegelstein, R.C., 2006. Prevalence of depression in survivors of acute myocardial infarction—Review of the evidence. Journal of General Internal Medicine 21, 30–38.

Tindale, H.A., Chang, Y., Kuller, L.H., Manson, J.E., Robinson, J.G., Rosal, M.C., Siegle, G.J., Matthews, K.A., 2009. Optimism, cynical hostility, and incident coronary heart disease and mortality in the Women's Health Initiative. Circulation 120, 656–662.

Tsigos, C., Chrousos, G.P., 2002. Hypothalamic-pituitary-adrenal axis, neuroendocrine factors and stress. Journal of Psychosomatic Research 53, 865–871.

Deprivation: Relative

Mohamed Cherkaoui, Centre National de la Recherche Scientifique, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The theory of relative deprivation aims at accounting for puzzling phenomena. Why, when agents belong to a group in which promotion or social mobility is scarce, they are more satisfied than those who belong to a group where promotion or social mobility is more frequent? Why inequality is more easily tolerated in a society where it is greater than in another? The theory was extended to explain social movements and revolution.

Following Runciman (1966), we can "roughly say that A is relatively deprived of X when (1) he does not have X, (2) he sees some other person or persons, who may include himself at some previous or expected time, as having X (whether or not this is or will be in fact the case), (3) he wants X, and (4) he sees it as feasible that he should have X." Relative deprivation and the notions that are directly or indirectly linked to it constitute a strongly structured conceptual set, what Merton terms a 'middle range theory,' that is, a set of propositions that combines hypotheses and empirical regularities proper to a specific domain of validity and without which these regularities would be neither accounted for nor cemented.

The concept of relative deprivation was first used by Stouffer et al. (1949) in *The American Soldier*. However, this idea can be found in the writings of several classical sociologists such as Tocqueville (1852), Durkheim (1893, 1897), and Weber (1904). It also features in the works of contemporary nonsociologists such as Duesenberry (1949) and Easterlin (1973). These authors underscore the importance of the notion of relative income compared to absolute income and thereby that of the comparison between two statuses of the same individual at two different time intervals. This accounts for certain paradoxical phenomena that the standard economic theory is unable to explain. Starting with the well-known study of Stouffer et al. (1949), the theory of relative deprivation has been gradually developed, particularly by Merton and Kitt (1950), Runciman (1966), Boudon (1977), and Hirschman (1981). Its application expands to various areas, mostly to those where the relations between individual contributions and retributions are crucial, just like the problems of justice and equity or revolution. This theory is also interpreted from different standpoints that are discussed later.

Themes and Theory of Deprivation

Relative deprivation is often used in *The American Soldier* to account for puzzling results whose exhaustive study was conducted by Merton and Kitt (1950). One of these results is that the most favored groups seem to be inclined to criticize the system of promotion. Thus, Military Police soldiers belonging to a unit where promotion is scarce are satisfied with the promotion system governing their military career. In contrast, men in the Air Corps, although belonging to a body where promotion is frequent, are dissatisfied with their opportunities for promotion. To explain this paradox, Stouffer et al. (1949) introduced the notions of relative deprivation and reference group: "if at the same age, level of education, and seniority, the majority of my colleagues and I have been promoted, I will consider my promotion as normal. If, however, I have not been promoted, I will go through a feeling of injustice, which can translate into criticism of the mobility system. If the promotion is relatively scarce in the group to which I belong or refer to in my evaluations, I will be very pleased with my promotion. However, if I am not promoted I will not experience any sense of injustice since all my peers have not been promoted." Deprivation therefore occurs only under specific conditions:

1. The individual retains some knowledge of the system of social promotion, mobility, or retributions.
2. He/she does not get an expected good, for example, promotion.
3. This expectation is based on an evaluation and a comparison that the individual conducts between his/her position, contributions, and retributions on the one hand, and those of a large number of his/her peers, who make up his/her reference group on the other. This condition expresses inter alia the principle of interdependence among social actors.
4. He/she has a feeling of inequity, which translates into criticism of the mobility system.

Thanks to particular models illustrating the idea of deprivation, several heterogeneous paradoxical phenomena have been accounted for by the founders of sociology as well as by contemporary sociologists. As explained and formalized by Cherkaoui (2005), Tocqueville (1852) underscores the following surprising positive ecological correlation: The higher the economic prosperity, the quicker the laws of the old regime disappear and political liberalization gains ground in certain regions, the stronger the support of these regions to the revolution. He points out that: "While prosperity (…) in France is growing, general attitudes seem to be less content and more anxious; public dissatisfaction is turning sour; hatred for all the old institutions is increasing. The nation is clearly heading towards a revolution. Moreover, the parts of France which should have been the major sources of this revolution are precisely those parts where progress is being shown most…. It is not always by going from bad to worse that we fall into revolution."

Durkheim (1893, 1897) also develops two partly complementary ideas linked to deprivation. In the *Division of Labor*, he maintains that happiness is not an increasing monotonic function of wealth. The relation between satisfaction and abundance

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32039-6

of wealth is a parabolic function with a maximum corresponding, according to him, to the happy medium, to the *aurea mediocritas* of the common saying: "happiness lies between the two extremes of wealth." Besides, the French sociologist devotes the chapter on 'the division of constrained labor' of the same book to the analysis of the relation between deprivation resulting from the violation of meritocratic rules and the questioning of social order. Equal opportunities and congruence between distribution of talents and jobs are, in his view, the conditions of legitimacy and stability of social order. In *Suicide*, it is the pure model of anomie that expresses deprivation. When aspirations and expectations grow faster than the means of realization and when these are not set back or brought to an end, a state of anomie occurs with numerous individual and social consequences. The absence of social regulation is a source of social pathology. With no clear limits to the legitimate needs, and free from any social pressure, men like Goethe's Faust will experience the infinite dimension of desire, frustration, and torment and can go as far as committing suicide.

Hirschman (1981), as an example of contemporary authors, explains tolerance to inequality in the developing countries by his 'tunnel effect', which is directly tied to the theory of deprivation and reference group. Let us take the example of a car driver in a two-lane tunnel with two lanes going in the same direction. Suppose that, after a short period of time during which no car in either lane has moved forward, the cars on the other lane start moving. He will feel relieved and hopeful that the cars on his lane will also move forward. However, upon realizing that only the cars on the other lane keep moving, he will feel disappointed and later furious. He might end up tempted to readjust the inflicted injustice by switching to the other lane, thus violating traffic regulations. The same applies to inequalities. An improvement of the social positions of my neighbors or my peers gives me hope. In case of disappointment, I will feel deprived and try to change the order of things.

Generating Mechanisms of Deprivation

One can wonder, following Boudon (1977), if all these phenomena are not likely to be generated by social mechanisms. He sets up simple models stemming from game theory, which bring about and explain relative deprivation. In his view, this deprivation can be interpreted as the product of the structures of interaction in which individuals end up. Deprivation entails the notion of competition for good and that of objective opportunities for promotion (Stouffer et al., 1949).

Let us suppose a group of N players who can win the lot B1 for a stake C1 or the lot B2 for the stake C2 with

$$\mathrm{B1} > \mathrm{C1},\ \mathrm{B2} > \mathrm{C2},\ \mathrm{B1} > \mathrm{B2},\ \mathrm{C1} > \mathrm{C2},\ \mathrm{B1}\ \text{and}$$
$$\mathrm{C1} > 0,\ \mathrm{B2}\ \text{and}\ \mathrm{C} > 0$$

$n1$ and $n2$ are the number of winners of the lots B1 and B2, respectively, with $n1 + n2 = N$.

Being rational, the players will bet on C1 rather than C2 if the expected net profit by betting on C1 is superior to that corresponding to C2.

If $x1$ players ($n1 < x1 < N$) bet on C1, the expectation of net profit E1 ($x1$) of a player betting on C1 is

$$\begin{aligned}\mathrm{E1}(x1) &= (\mathrm{B1} - \mathrm{C1})n1/x1 + (\mathrm{B2} - \mathrm{C2})(x1 - n1)/x1 \\ &= (\mathrm{B1} - \mathrm{B2})n1/x1 + \mathrm{B2} - \mathrm{C1}\end{aligned}$$

On the other hand, the expectation of net profit E2 of a player betting on C2, when $x1$ ($>n1$) players bet on C1 is,

$$\mathrm{E2}(x1) = \mathrm{B2} - \mathrm{C2}$$

Thus, $x1(>n1)$ players bet on C1, a player is advised to bet on C1 rather C2 if

$$\mathrm{E1}(x1) > \mathrm{E2}(x1)\ \text{or}\ (\mathrm{B1} - \mathrm{B2})n1/x1 + \mathrm{B2} - \mathrm{C1} > \mathrm{B2} - \mathrm{C2}$$

that is,

$$(\mathrm{B1} - \mathrm{B2})/(\mathrm{C1} - \mathrm{C2}) > x1/n1$$

One can analyze the relationship between the game structure and deprivation, which is the proportion of individuals betting on C1 (the highest price) but winning only B2 (the lowest price). For that matter, one should examine the different values of the ratio (B1 – B2)/(C1 – C2) that is the ratio of the possible additional gain and the additional cost on one hand, the percentages of players 100 $x1/N$ and of the winners 100 $n1/N$ on the other hand.

The simulation shows how complex the nature of relations is between deprivation and the objective opportunities of promotion. Let us consider the following general cases:

1. The rate of deprivation increases with the values (B1 – B2)/(C1 – C2). This rate increases with the marginal rate of profitability or return on investment {(B – B2) – (C1 – C2)}/(C1 – C2), as shown by Kosaka (1986) and Yamaguchi (1998) who systematize the propositions of Boudon (1977) and deduce new ones, and make 'critical points' emerge.
2. The highest levels of deprivation correspond to cases where a small number of players win the lot B1, the highest one.
3. Deprivation rate is moderated as long as the moderated value lots are distributed to players in a limited number (the case of the Military Police in Stouffer's study). This is no longer the case when the number of winners increases (the case of men of the Air Corps).
4. We get equal deprivation rates when only a small number of players win important lots or when a large number of players win less important ones.
5. In general, the curves corresponding to identical relative deprivation levels are complex with different plateaus (ceilings) and critical points.

The simulation confirms the idea according to which the game structure determines the deprived players and their number. It is equally possible to simulate the situations in which the players are not equal, that is, when the group is stratified and each social class has its different resources and behaviors toward risk. These simple game models particularly show the conditions in which the mere fact of increasing everyone's promotion opportunities can result in further increasing promotion expectations. These modes also demonstrate that the propositions of Tocqueville, Durkheim, and Stouffer are not unconditionally valid: when equal opportunities increase

among individuals, deprivation tends to grow under certain conditions, but it can decrease under others. The trend of the relation between promotion and social mobility depends on the structure of interdependence among individuals.

Manzo (2009, 2011) enriches our understanding of relative deprivation by two new proposals. First, he suggests introducing a distinction between relative deprivation frequency, that is, the proportion of agents who do not have what they want, and relative deprivation intensity, that is, the strength of the feeling agents associate with this discrepancy. The distinction suggests that different mechanisms generate these two proprieties. Second, he carries on agent-based simulation by introducing a new component. He analyses the level of satisfaction of the agent by taking into account neighborhood based interindividual comparisons. The model with interpersonal comparisons that inversely link individual dissatisfaction and the diffusion of deprivation situation leads to two paradoxically apparent propositions: 'more opportunities, more dissatisfied yet less intensely dissatisfied agents' and 'more opportunities, fewer dissatisfied yet more intensely dissatisfied agents.'

Deprivation and Collective Violence: A Case Study

Revolution theorists generally agree on the assumption that revolutions occur during a period of economic growth and political liberalization. This paradox seems to cast doubt on the hypothesis of the rationality of human behavior. One can wonder why people revolt at the very moment of prosperity. Deprivation theorists respond that an improvement in objective living conditions tends to generate higher expectations. As the expectations grow faster than the improvement in objective conditions, a wider gap comes up and translates into a higher degree of deprivation. This model is similar to the Durkheimian theory of anomie, as shown by Besnard (1987) and Coleman (1990). The revolution theory developed particularly by Davies (1962) and Gurr (1968) is different. In their view, revolutions occur when a long period of socioeconomic prosperity is immediately followed by a sudden downturn that is a sharp decrease in gratification accompanied by a constant increase in expectations. The rapidly widening gap between expectations and gratification generates revolution. This proposition commonly translates in an inverted J curve.

Even though the notion of deprivation in this theory is not as precisely and explicitly defined as with Stouffer, it is easy to demonstrate that it also supposes a meaningful difference deems unacceptable between the expectation and the realization at the individual level and an interpersonal comparison. First, with regard to intrapersonal comparison, the difference is at the same time relative and absolute. It is absolute if at the $t1$ moment we compare expectation with realization. It is relative if at the $t2$ moment, the difference between expectation and realization deems unacceptable or unbearable compared to what it was at the $t1$ moment. Besides, individuals with threatened social positions or subject to downward social mobility compare themselves to those with stable positions or upward mobility.

Several varieties of deprivations can be constructed. The first one relies on the theory of relative deprivation and shows up in the works of Gurr (1968) particularly for whom "a psychological variable, relative deprivation, is the basic precondition for civil strife of any kind, and that the more widespread and intense deprivation is among members of a population, the greater the magnitude of strife in one or another form." The relation between deprivation and civil strife is, however, mediated by societal variables such as coercive potential that can deter strife, institutionalization (social and political association may provide nonviolent means for expressing discontent), facilitation (i.e., means that facilitate civil strife), and finally legitimacy of the regime.

The second variety is that of the status inconsistency that can be found in the founding fathers works such as those of Tocqueville (1852) and Weber (1904, 1924), and in such contemporary research as that of Lenski (1954). The rich bourgeois and the well-off farmer, who, thanks to the liberalization of the rules governing the acquisition of land, end up as owners of vast lands, remain for the French aristocrat mere plebeians carrying visible stigma of their original social class. If, as Tocqueville states, the bourgeois and the aristocrats become unknown to each other, it is mainly because they evaluate each other according to values located on different scales. The bourgeoisie has economic quasi monopoly and looks forward to assuming political power and enjoying honors, whereas the bourgeoisie cannot gain access to these institutions because the noblesse get in the way. The bourgeoisie feels compelled to opt for the revolutionary solution. Conversely, in the nineteenth century, as Weber points out, while the Prussian Junker maintains his monopoly over the army and the bureaucracy, the industrialists buy lands to acquire nobility titles and gain access to political positions. Politically, as Weber states, this type of industrialist is reactionary. These two examples illustrate two special cases of the multiple effects of the status inconsistency whose complete taxonomy Cherkaoui (1995) attempts to construct.

The collective action theorists, especially those of mobilization, question the theory of revolution based on deprivation. As revolution is a collective action seeking to realize a public interest, the rational actor will face problems specific to this type of action, particularly the temptation of free riding. Following Olson (1965), it is believed that the rational individual belonging to a latent group as a social class seeking to satisfy collective goals would rather turn to see the sacrifices accepted by others. He would not be interested in participating in a revolutionary enterprise whose costs can be heavy. Eventually, as pointed out by several analysts of revolutionary movements as Cochin (1924) and Brinton (1965), the revolutions are the product of some elites of organized and almost unanimous conspirers toppling weak governments at periods of social disorganization. For the theorists of mobilization and social movements as Oberschall (1973) and Tilly (1978), a psychological explanation of the revolutionary phenomenon is unable to account for the shift from disenchantment to the mobilization against authorities and in particular against the state. Besides, the theory of rational choice reduces the moral and solidarity incentives, which play a key role in the emergence of collective action. This is sometimes a moral obligation for the individual who belongs to collectivities where solidarity bonds are predominant. From this perspective, it is rather the lack of participation that will be costly. Oberschall shows that the lack of bonds and the dismantling of social networks are major setbacks to mobilization. Cherkaoui (2007, 2013) has recently

tested this theory in his studies on social movements in Morocco and on the Arab spring uprisings by taking into consideration other central variables such as the legitimacy of governing power, the homogeneity and solidarity of elites, and the role of the army.

However, by trying to turn the theory of relative deprivation into a maid of all work, it is put through risks that hang over any theory whose domain of application has been excessively expanded. But despite some criticism, its underlying principles remain solid. To mention but one, the principle of interdependence is the center of its axiomatic structure that the formalization put forward by Boudon and others has clearly highlighted. To explain and understand, inter alia, the inflationary spiral of demands, the perception of inequality, the individual conception of justice and of equity, or the degree of satisfaction, the theory of relative deprivation reminds us of an obvious fact that the sociological tradition has taken for granted since the days of Montesquieu (1748) whereby the structures of interdependence between social actors are often decisive.

See also: Anomie; Collective Behavior, Social Psychology of; Durkheim, Emile (1858–1917); Groups, Sociology of; Intergroup Relations; Mobility, Social; Organizations and Culture; Social Comparison, Psychology of; Social Justice and Social Inequalities; Social Psychology, Sociological; Social Psychology; Social Simulation: Computational Models; Status and Role, Social Psychology of; Suicide, Sociology of; Tocqueville, Alexis de (1805–59); Violence: Public.

Bibliography

Besnard, Ph, 1987. L'Anomie, first ed. Presses Universitaires de France, Paris.
Boudon, R., 1977. Effets Pervers et Ordre Social. Presses Universitaires de France, Paris [The Unintended Consequences of Social Action. Macmillan, London, 1982].
Brinton, C., 1965. The Anatomy of Revolution, Rev. and Expanded Edn. Vintage Books, New York.
Cherkaoui, M., 1995. Mobilité soiciale et comportement politique: taxinomie des modèles de relations. Revue Française de Sociologie 36, 171–183.
Cherkaoui, M., 2005. Invisible Codes. Essays on Generative Mechanisms. Bardwell-Press, Oxford.
Cherkaoui, M., 2007. Morocco and the Sahara. Social Bonds and Geopolitical Issues. Bardwell-Press, Oxford.
Cherkaoui, M., 2013. Incertitude et Ellipse des Crises, to be published by the Laval University Press.
Cochin, A., 1924. La Révolution et la Libre Pensée. Copernic, Paris. 1979.
Coleman, J.S., 1990. Foundations of Social Theory. The Belknap Press of Harvard University Press, Cambridge, MA.
Davies, J.C., 1962. Towards a theory of revolution. American Sociological Review 27, 5–19.
Duesenberry, J.S., 1949. Income, Saving and the Theory of Consumer Behavior. Harvard University Press, Cambridge, MA.
Durkheim, E., 1893. De la Division du Travail Social. Alcan, Paris [The Division of Labor in Society, Free Press, New York, 1947].
Durkheim, E., 1897. Le Suicide. Alcan, Paris [Suicide, Free Press, New York, 1951].
Easterlin, R., 1973. Relative economic status and the American fertility swing. In: Sheldon, E.B. (Ed.), Family Economic Behavior. Lippincott, Philadelphia, PA.
Gurr, T., 1968. A causal model of civil strife. American Political Science Review 62, 1104–1124.
Hirschman, A.O., 1981. Essays in Trespassing: Economics to Politics and Beyond. Cambridge University Press, Cambridge, UK.
Kosaka, K., 1986. A model of relative deprivation. Journal of Mathematical Sociology 12 (1), 35–48.
Lenski, G., 1954. Status crystallization: a non-vertical dimension of social status. American Sociological Review 19, 405–413.
Manzo, G., 2009. Boudon's model of relative deprivation revisited. In: Cherkaoui, M., Hamilton, P. (Eds.), Raymond Boudon: A Life in Sociology, vol. III. Bardwell-Press, Oxford, pp. 91–121.
Manzo, G., 2011. Relative deprivation in silico: agent-based models and causality in analytical sociology. In: Demeulenaere, P. (Ed.), Analytical Sociology and Social Mechanisms. Cambridge University Press, Cambridge, pp. 266–308, ch, 13.
Merton, R.K., Kitt, A.S., 1950. Contributions to the theory of reference group behavior. In: Merton, R.K. (Ed.), 1968 Social Theory and Social Structure. Free Press of Glencoe, New York.
Montesquieu, C.L., 1748. L'esprit des lois. Gallimard, Paris. Bibliothèque de la Pléiade, 1965 [The Spirit of the Laws, Cambridge University Press, 1989].
Oberschall, A., 1973. Social Conflict and Social Movements. Prentice-Hall, Englewood Cliffs.
Olson Jr., M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.
Runciman, W.G., 1966. Relative Deprivation and Social Justice. Routledge and K Paul, London.
Stouffer, S.A., Suchman, A., DeVinney, L.C., Star, S.A., 1949. The American Soldier. Princeton University Press, Princeton, NJ.
Tilly, C., 1978. From Mobilization to Revolution. Addison Wesley, Reading, MA.
Tocqueville, A de, 1852. L'Ancien Régime et la Révolution. In: Œuvres Complètes. Gallimard, Paris [The Old Regime and the Revolution].
Weber, M., 1904. In: Weber, M. (Ed.), Agrarstatistische und sozialpolitische Betrachtungen zur Fiedeikommissfrage in Preussen (1924).
Weber, M., 1924. Gesammelte Aufsätze zur Soziologie und Sozialpolitik. JCB Mohr, Tübingen, Germany.
Yamaguchi, K., 1998. Rational-choice theories of anticipatory socialization and anctici-patory non socialization. Rationality and Society 10, 163–199.

Desegregation

John Dixon, Open University, Milton Keynes, UK
Kevin Durrheim, University of KwaZulu-Natal, Scottsville, South Africa
Manuela Thomae, University of Winchester, Winchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Desegregation is a process through which members of formerly separated groups are brought together, often through the removal of institutional barriers to interaction. Two recurring arguments have been presented in favor of desegregation. The first holds that the process promotes intergroup harmony and tolerance; the second holds that it promotes social justice and equality. Present in the earliest work in the field, these arguments continue to set the agenda for future research, inspiring recent debates, developments, and controversies.

Introduction

Desegregation is a deceptively simple concept. It can be defined as a process through which members of formerly separated groups are brought together, often through the removal of formal barriers to interaction. The concept acquired historical significance largely through its association with political movements such as the Civil Rights Movement in the United States and the antiapartheid movement in South Africa, which, among other goals, fought for the abolition of legal systems of racial segregation.

A simple definition of desegregation belies its complexity however. Desegregation is generally not best conceived as a discrete *event*: it is more typically an unfolding and incomplete *process* that takes multiple forms and operates across multiple spatial and temporal scales. In the United States, for example, the phrase 'educational desegregation' denotes a wide range of processes, including the abolition of Jim Crow laws, open enrollment in formerly exclusive schools or colleges, quota systems, bussing programs, the realignment of district school boundaries, and the establishment of 'magnet schools' in African-American neighborhoods. Writing this article is challenging because these different forms of desegregation are associated with different historical trajectories, lived experiences, and social consequences. Moreover, educational desegregation is itself distinct from residential or employment desegregation, which are in turn distinct from the desegregation of spaces of leisure, recreation, or public facilities.

Our aims in this article, then, must be modest. We do not pretend to offer a systematic review of the complex literature that has addressed the varying forms of segregation. Instead, we sketch a few wider themes, issues, and debates in the field emphasizing, in particular, how desegregation affects the social psychology of relations between groups. Much of the work we discuss has been conducted in the United States and has focused on racial segregation, and this emphasis reflects a wider bias in the historical development of the field. We focus on two recurring arguments in favor of desegregation. The first holds that desegregation is essential to the promotion of intergroup harmony and tolerance; the second holds that desegregation is essential to the struggle for social justice and equality. As we discuss in our conclusion, these arguments continue to set the agenda for future research, inspiring recent debates and controversies. They were present too in the earliest work in the field.

The Early Social Science Case for Desegregation

In order to understand the emerging social science case for desegregation, we need first to understand what it was directed against, namely the assumption that the separation of 'races' is a social, political, or even natural imperative. Nightingale (2006) holds that a number of 'transnational' arguments in favor of race segregation were predominant in the late-nineteenth and the early twentieth centuries, gaining both scientific credibility and popular appeal. One suggested that different 'races' were naturally adapted to different environments and indeed that those environments were intimately related to their characteristics. Livingstone (1991) has shown, for example, how geographers at this time believed that place, climate, and group morality were intimately connected. The torrid equatorial regions, they argued, had fostered a 'lower' race of Africans distinguished by their sensuous, spontaneous, emotional, and childlike character, while Europeans could thank their more temperate environments for their virtues of reason, restraint, thrift, and endeavor. Each race had its proper place in the world to which it was adapted and from which its moral character had arisen. As such, the entry of one group into the environment of another was deemed inherently dangerous, as personified by the figure of the colonist who had 'gone native.' However anachronistic this line of argument seems now, it was taken seriously at the time. (Read, for example, Nightingale's (1996) account of the measures taken by the British Army to prevent their soldiers succumbing to the degenerative effects of the Indian summer.)

A second argument in favor of segregation is perhaps more familiar, for it lingers as a residual explanation even in societies where *de jure* segregation has been abolished. It holds that contact between members of different ethnic and racial groups produces negative consequences. Along with the 'threat' of miscegenation, such contact creates friction between groups, arising as a result of cultural, intellectual, and moral differences, as well as instinctive aggressions. Such friction can in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.24055-5

turn produce intergroup conflict that destabilizes society. In the early years of the twentieth century, in a period Pettigrew (2008) brands the 'nadir' of social science research on race relations, this kind of argument found support, among other sources, from the psychological literature on race differences in intelligence.

As the century unfolded, such arguments lost favor. Social scientists set about the task of 'puncturing the biological myths of race' (Fairchild and Gurin, 1978), and in what Samelson (1978) called an 'abrupt reversal,' racial inequality and conflict were attributed increasingly to the prejudices of the advantaged rather than the inferiorities of the disadvantaged. Desegregation emerged as a potential solution to range of ethnic and racial 'problems,' and many social scientists became political advocates within a struggle that was to ramify across the globe. In the United States, a cradle of this emerging vision, the social science case for desegregation was encapsulated by a contrast between two legal landmarks. If the 'Plessy versus Ferguson Ruling of 1896' had upheld the legal status of 'separate but equal' racial segregation of public facilities in the antebellum southern states and other parts of the United States, then the 'Brown versus the Board of Education Ruling of 1954' famously overturned this decision and set in motion a process that culminated in the end of *de jure* segregation in the United States.

Although their contribution has been much debated, that social scientists played a significant part in the *Brown* judgment is indisputable, as the outpouring of work that accompanied its 50-year anniversary reaffirmed (e.g., see the special issue edited by Pickren, 2004). Handing down their judgment on the legality of educational segregation, the Justices echoed the language of a statement prepared for the courts by a team of experts that included Isidor Chein, Stuart Cook, and Kenneth B. Clark (Since this landmark statement, several further Amicus Curae on the effects of segregation have been submitted to courts. For a recent example, see Grutter et al., Brief of the American Psychological Association as Amicus Curiae. http://www.apa.org/about/offices/ogc/amicus/grutter.pdf. Retrieved 24.11.13.). This statement argued that black children subjected to legal segregation would suffer a range of negative consequences, including self-hatred, reduced ambitions, 'defeatist attitudes,' and weaker educational aspirations and achievements. Conversely, white children 'benefiting' from such segregation were likely to acquire an unrealistic sense of personal worth, guilt feelings associated with their unfair advantage, and misplaced fears and hatreds of minority groups. Regarding relations between groups, the statement noted in addition that "Segregation leads to a blockage in the communication and interaction between the two groups. Such blockages tend to increase mutual suspicion, distrust, and hostility" (Clark et al., 1952/2004: p. 497).

Clark (1953) developed this emerging social science case in a landmark paper published in the *Journal of Social Issues*. Entitled 'Desegregation: an appraisal of the evidence,' the paper elaborated a model of social change that challenged the received wisdom that desegregation was dangerous policy that, if implemented, would lead to widespread conflict. The latter view, he argued, relied upon an outdated, pseudopsychology that presupposed that racial prejudices are so emotionally deep-seated, so intractable, that they inevitably thwart interventions to promote change. Clark insisted that the available evidence supported the opposite view. Desegregation could proceed smoothly, efficiently, and without violent resistance if implemented under the right conditions. Challenging the more cautious, 'gradualist' model of change recommended by some commentators (e.g., Ashmore, 1954), Clark proposed a sweeping transformation of American society. Desegregation should be implemented rapidly, decisively, and with the full weight of institutional sanctions against those who resist. Stateways could change folkways, and in due course, their associated patterns of racial attitudes would fall in line.

Of the early evidence supporting this argument, troops' reactions to the desegregation of the US armed forces during the Second World War is perhaps the most poignant. This evidence focused on the experiences of white soldiers brought together with their 'Negro' counterparts in order to fight in a war in which many of them would die. Desegregation in this context happened suddenly, with no questions asked, and as a matter of military expediency rather than for some noble ideal such as 'racial equality.'

Consider, for example, Star et al.'s (1949/1958) mixed-methods study of white soldiers' responses to the amalgamation of Negro platoons in infantry companies. Their quantitative survey demonstrated how the concrete experience of desegregation actually increased rather than decreased soldiers' support for its basic principles. Most infantrymen who had experience of soldiering in mixed platoons, for example, were either broadly accepting or positively supportive of further amalgamation with 'Negro' platoons (around 60%) and few (<5%) reported that they would actively 'dislike' this outcome. By contrast, most infantrymen soldiering in segregated platoons (>60%) reported they would actively dislike the desegregation of platoons and vanishingly few (around 2%) said they would like this process. Desegregation, in short, seemed to produce tolerant acceptance rather than resentful conflict, a process clarified by the interview accounts Star et al. (1949) gathered:

> Relations are very good. They have had their pictures taken together, go to church services, movies, play ball together. For a time there in combat our platoons got so small that we had to put a white squad in the colored platoon. You might think that wouldn't work well, but it did. The white squad didn't want to leave the platoon. I've never seen anything like it.
>
> Company Commander from Nevada; Star et al., 1949: p. 600

> When I heard about it, I said that I'd be damned if I'd wear the same shoulder patch they did. After the first day when we saw how they fought, I changed my mind. They're just like any of the other boys to us.
>
> Platoon Sergeant from South Carolina, 1949: p. 598

As these accounts illustrate, desegregation does not inevitably lead to more negative racial attitudes. Indeed, the opposite often proves true. Desegregation may improve ethnic, racial, and other intergroup attitudes by allowing participants to get to know one another better, form bonds of emotional affiliation and attachment, and even create a sense of shared

identity, as captured in the phrase, 'they are just like any of the other boys to us.' Later developed as 'the contact hypothesis,' this idea was to become a dominant theme in desegregation research.

Desegregation and the Reduction of Prejudice

In the conclusion of their paper, Star and colleagues voiced a note of caution. Researchers should be wary, they warned, of generalizing their findings to other types of intergroup relations. "Far from being a 'test case' in ordinary Negro-white relations, the combat setting may be regarded as a special case for making good relationships, for the sense of common danger and common obligation was high, the need for unity was at a maximum, and there was great consciousness of shared experience of an intensely emotional kind" (Star et al., 1949/1958: p. 600). They noted that these 'special' circumstances not only increase the likelihood of interracial contact having positive effects, but also promote what Gaertner and Dovidio (2000) would later term 'common identification.'

In his classic statement of the 'contact hypothesis,' which consolidated earlier work in the field (e.g., McKenzie, 1948), Gordon Allport (1954) was similarly cautious. He observed that desegregation can produce various consequences and that its nature and surrounding context was crucial. Indeed, paradoxically, contact may either worsen or improve relations between formerly segregated groups. If contact is superficial, marked by competition, of unequal status, or lacking institutional support, then it may perpetuate negative stereotypes and feelings about members of another group. However, if contact is intimate, cooperative, of equal status, and supported by institutional norms and structures, then it may reduce prejudice and dampen the potential for future intergroup conflict. Allport's classic statement of the contact hypothesis, then, was not a naive call for more interaction between groups, but a measured statement of the conditions under which such interaction yields positive outcomes.

In the intervening years since the publication of Allport's seminal work, many hundreds of studies have been conducted on the contact hypothesis. There is now a vast literature on how, when, and why contact is effective. Early reviews (e.g., Amir, 1969) of this burgeoning literature were no more than cautiously optimistic. Recent reviews present a far more confident picture of the field's contribution from which we wish to extract four themes (see Pettigrew and Tropp, 2006, 2011; Brown and Hewstone, 2005; Dovidio et al., 2003).

First, most research has confirmed that contact correlates negatively with prejudice. The average strength of this effect is modest ($r = 0.22$), but it emerges consistently across different historical periods, types of intergroup relations, and outcome measures. Early research focused mainly on race relations unfolding under conditions of institutional desegregation in domains such as housing (e.g., Wilner et al., 1952), employment (e.g., McKenzie, 1948), and education (e.g., Mann, 1959). Later work has encompassed a far wider array of social relations, including relations between Catholics and Protestants (Schmid et al., 2008), Arabs and Jews (Maoz, 2011), gays and heterosexuals (Herek and Capitanio, 1996), students and people with HIV/Aids (Werth and Lord, 1992), Dutch people and Muslims (Savelkoul et al., 2011), doctors and patients (Revenson, 1989), the young and the elderly (Caspi, 1984), and people with and without mental illness (Desforges et al., 1991). More exotic entries in the contact literature include work on attitudes toward divorcees, transvestite comedians, and nude sunbathers (Ray, 1983; Schiappa et al., 2005).

Second, although the majority of research has focused on the effects of contact on intergroup attitudes of the historically advantaged (Perhaps because the prejudices of the dominant group members have historically been perceived as the main problem in issues such as racism, the attitudes of the disadvantaged have often been neglected in the literature.), several studies have confirmed that contact also works in the opposite direction, reducing the 'prejudices' of the disadvantaged. For example, in postapartheid South Africa, positive contact with white South Africans has been associated with a range of attitudinal shifts among black South Africans, including reductions in negative emotions and social distance and increases in forgiveness and support for reconciliation (e.g., Gibson, 2004; Swart et al., 2011; Dixon et al., 2010). Even in the society that implemented 'the highest stage of white supremacy' (Cell, 1982), contact is softening the hearts and minds of the disadvantaged.

Third, the theoretical mechanisms through which contact produces positive effects have been progressively clarified. Early work tended to describe rather than explain such effects or to fall back on the simplistic assumption that contact dispels false stereotypes by providing participants with more accurate information about one another. Since the 1980s, building on advances in work on social categorization and identification, researchers have explored the role of cognitive processes such as decategorization (Miller and Brewer, 1984), mutual differentiation (Hewstone and Brown, 1986), and common identification (Gaertner et al., 1996) in mediating contact effects, treating such processes as either competing or sometimes as complementary explanations (e.g., see Pettigrew, 1998). For example, the common identity model holds that positive contact promotes a 'recategorization' process through which 'us' and 'them' become viewed as 'we' and that this cognitive shift then produces associated improvements in intergroup attitudes (e.g., Gaertner et al., 1989). More recently, researchers have also emphasized the role of affective processes in explaining the impacts of contact on prejudice. As **Figure 1** below illustrates – taken from Pettigrew and Tropp's (2008: p. 928) recent analysis of European survey data – the roles of anxiety reduction and empathy promotion have been attributed particular importance.

Fourth, emerging research suggests that contact works not only directly, via immediate face-to-face interactions between members of historically divided groups, but also indirectly, via social networks of connection, vicarious observation, mass-mediated communication, or even 'imagined' interactions. The so-called extended contact effect, for instance, indicates that being embedded in a diverse social network (i.e., having friends who have friends of another group) can in itself be sufficient to produce positive attitude changes. Originally posited by Wright et al. (1997), this idea has gathered momentum in recent years (Dovidio et al., 2011). Along similar lines, several recent studies suggest that

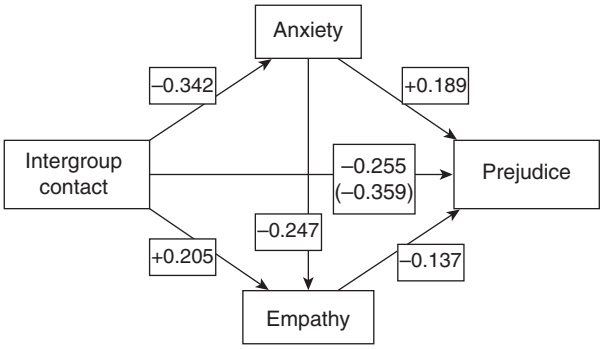

Figure 1 Path model of mediation of the contact–prejudice association by anxiety reduction and empathy. Reproduced from Pettigrew, T.F., Tropp, L.R., 2008. How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. European Journal of Social Psychology 38, 922–934.

simply imagining having an encounter with a member of another group improves our attitudes toward that group (Crisp and Turner, 2009), especially if participants also imagine that the encounter is happening under cooperative conditions (Kuchenbrandt et al., 2013).

In our opinion, this recent proliferation of work on indirect and imagined forms of contact exemplifies a broader shift of emphasis in the literature on intergroup contact and desegregation. There has been a gradual move away from research on how to promote sociopolitical and institutional change in real world settings and toward a more contextually abstracted study of prejudice reduction as an ideal in its own right (Durrheim and Dixon, forthcoming). Certainly, work on 'imagined contact' is at best tenuously connected to Clark and colleagues' original project, which sought to understand the conditions under which desegregation can become an effective mechanism of social change.

Desegregation and Social Justice

That project, it is worth reminding ourselves, was never simply about getting black kids to sit next to white kids in the classroom. It was never simply about getting members of different groups to hold nicer thoughts and warmer feelings about one another. Rather, it was a political project that sought to dismantle the forms of inequality that racial segregation had created. Dreger (1985) reminds us of what was at stake here and why desegregation was seen as such a vital a mechanism of sociopolitical change:

> Desegregation, led by school desegregation, has revolutionized the South. Let me describe what segregation was like when I came to the South in 1949. A black gas station attendant could pour the gas, but the white man who supervised the black attendant handled the money. There were *no* black cashiers in grocery stores, except in stores in deep black neighborhoods and even in the latter white owners or managers were constantly supervising closely. There were *no* black salespersons in major or even minor stores. There were *no* black bus drivers except for rickety buses devoted entirely to transporting black workers. There were some black school bus drivers for buses that carried black children. At the bus station, blacks stood back as whites got on first; if there was no room left, too bad. At the bus stops, blacks cleverly made it appear that they were occupied with other things in order to make it seem that they preferred to get on after whites. And of course, all the obvious marks were there, seats in the back of the bus, seats in the theater galleries, rear door entrances to white homes or establishments, and so forth …
>
> p. 124–125

The 'obvious marks' referred to here are, of course, the marks of subordination. They are the marks of a system of inequality that pervaded even the most banal scenes of everyday life, such as queuing for a bus, going to the cinema, or buying groceries, and acted as a constant reminder of one's place in the world. Early social science researchers were concerned fundamentally with the material and social psychological effects of this system on historically disadvantaged communities. They conceived desegregation as a vehicle of social justice and not only, or even primarily, as a vehicle of social harmony.

Evidence from research on the racial (re)organization of institutions of education and residence attest to the continuing importance of this political project. In the next section, we take the relationship between residential organization and social justice as an example. Readers interested in related work on educational desegregation can weigh the diverse perspectives offered by Orfield and Lee (2005), Schofield and Eurich-Fulcer (2001), and Rossell et al. (2002).

The (re)organization of residential space: Clark's (1965) classic the *Dark Ghetto* revealed the power of urban segregation not only to divide the disadvantaged from the advantaged, but also to divide them from the material resources and opportunities with which to build better lives, including education, employment, bank loans, and prospects for social mobility. Confined to poor, inner city neighborhoods, he argued, black residents also became trapped in a cycle of deprivation and indignity: "… the dark ghetto's invisible walls have been erected by the white society, by those who have power, both to confine those who have no power and to perpetuate their powerlessness" (p. 11).

It is largely because of this intimate relationship to poverty and exclusion that residential (de)segregation remains of such vital significance. Indeed, the organization of residential space has been described as a 'structural lynchpin' of race relations in the United States (Pettigrew, 1979) and elsewhere (Durrheim and Dixon, 2005). It is thus widely viewed as a *sine qua non* unlocking other forms of desegregation. *Ceteris paribus*, the 'dark ghetto' can be read as an analogy for urban relations between historically advantaged and disadvantaged groups in many societies.

In the decades since *the Dark Ghetto* was published a lot has changed in the United States. Partly because of the civil rights movement, many of the practices that helped to sustain urban segregation have now been outlawed by legislation such as the Fair Housing Act (1968) and the Equal Credit Opportunity Act (1974); it is now illegal, for example, for bankers or realtors to engage in practices of the so-called 'redlining' (e.g., refusing mortgages to clients who live in predominantly minority neighborhoods or channeling property purchases or lets along

racial lines). As the last century came to an end, there were also signs that some American cities were becoming less segregated or at least that 'hypersegregated' cities (Massey and Denton, 1988) might be downgraded to merely 'segregated.' It became possible to speak of 'stable racially integrated communities' (Ellen, 1998, 2000) who had not succumbed to the 'white flight' and resegregation, but instead established spaces of diversity. Certainly, there is sufficient variation – both within and between American cities – to enable systematic comparative work on the consequences of living in more or less racially diverse neighborhoods. This work confirms how and why desegregation remains a matter of social justice.

Consider, for example, **Figure 2** below, which depicts the racial demography of Chicago using an innovative technique in which different colored dots are used to represent members of different race groups living in different areas of the city (see Rankin, 2009). This kind of visual representation has the advantage of capturing the gradients and blurred boundaries between racially differentiated areas; the spaces of mixing and transition as well as the sharp territorial divisions. It thus provides a more subtle image of the nature of segregation than the standard modes of visual representation (e.g., see the inset image in **Figure 1**). Nevertheless, **Figure 1** also shows why Chicago is a city that regularly tops the segregation league table of American cities in what remains, unfortunately, a highly competitive field. It illustrates *in extremis* a pattern that is repeated in many other cities, as captured by Farley et al.'s (1978) memorable phrase 'Chocolate city, vanilla suburbs.'

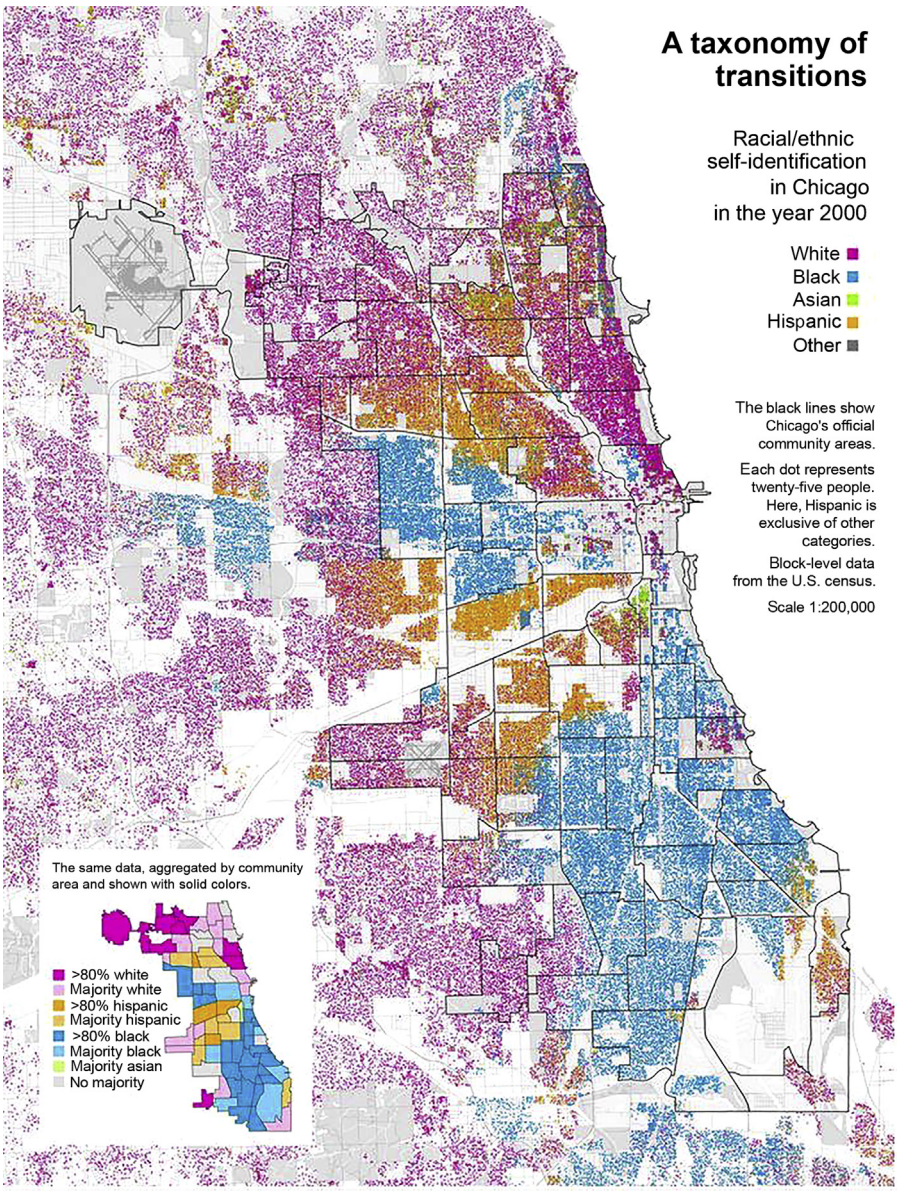

Figure 2 Mapping residential distribution by race in Chicago, 2000. Taken from Rankin, B., 2009. http://www.radicalcartography.net/index.html?chicagodots (retrieved 20.10.13.). Note each dot represents 25 residents who self-identified as white, black, Hispanic, Asian, or other.

African-American residents, Hispanic, and other minority group residents are clustered mainly in the central city and immediately surrounding areas. The wider suburbs are overwhelming populated by white residents. In Chicago, this pattern has further ossified into a racialized division between northern and the southern parts of the city: the 'southside' having a high concentration of African-American residents, the 'northside' having very few.

Figure 3 provides another representation of Chicago, which displays the relationship between variations in racial demography of the city and household income. As is apparent, areas with a higher proportion of African-American residents tend to have lower average-income levels. This pattern is most prominent in the central areas of the city, where segregation and poverty starkly coincide. However, what does this kind of association mean? Is it coincidence, a historical artifact, the product of a 'third variable' such as class or cultural preferences? If not, does poverty lead to segregation or vice versa?

In the early 1990s, Douglas Massey and Nancy Denton (1993) addressed these kinds of questions in a book provocatively entitled 'American Apartheid.' This book presented compelling new evidence on the role of segregated environments in creating an African-American 'underclass' who had effectively become trapped in the self-same cycle of deprivation described by Clark some 30 years earlier. It showed that segregation served as an active cause and not merely a passive reflection of inner city poverty and that, above and beyond dynamics of class, race played a role *sui generis* in this

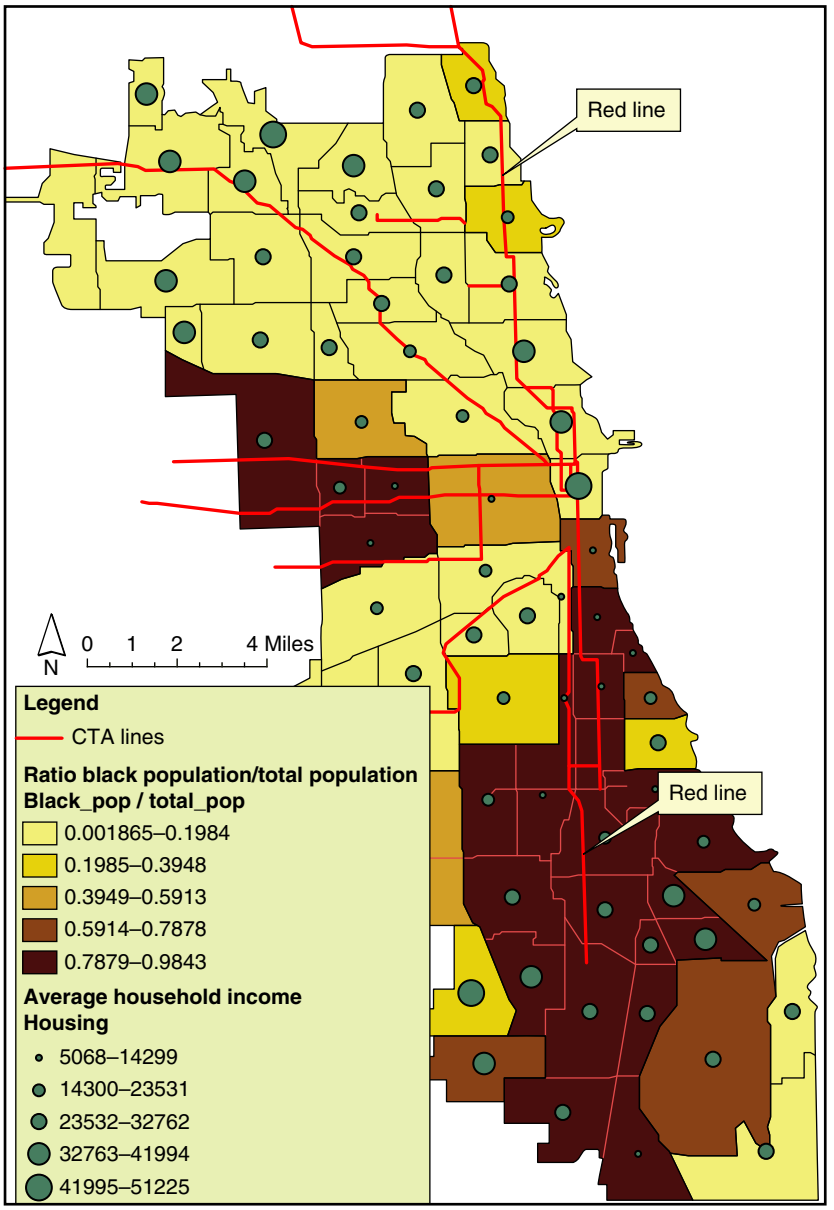

Figure 3 The relationship between proportion black in residential areas and mean household income in Chicago. Taken from Swyngedouw, E., 2013. The segregation of social interactions in the red line L-train in Chicago. Symbolic Interaction 36, 293–313.

process. It showed that African-American segregation is maintained largely through institutionalized practices of discrimination in housing and mortgage markets, which effectively prevents upwardly mobile blacks from moving to wealthier, more racially integrated, neighborhoods. It showed, too, how in periods of economic recession problems such as unemployment often became concentrated within – indeed sometimes confined to – such communities, with often devastating social and economic consequences. Above all, it showed that dismantling urban segregation was a necessary component of for any meaningful policy of social change.

American Apartheid reinvigorated a tradition of research that had become somewhat moribund in period between the 1960s and the late-1980s. It brought questions about the relationship between urban segregation, desegregation, and social justice back into razor-sharp focus. Since its publication, there has been an explosion of work on the deleterious consequences of living in racially segregated environments or, to put the matter more positively, on the benefits of living in more diverse environments. Renewed interest in the links between race, residential organization, and poverty (e.g., Charles, 2001; Massey and Fischer, 2000) has been complemented by work on closely related problems. We want to extract two examples from a more complex and sobering literature.

The first example concerns the effects of racial diversity and segregation on the physical well-being of members of disadvantaged groups. Growing evidence indicates that living in more diverse communities is literally good for your health. Relative to living in segregated environments, it leads to decreases in infant mortality, increases in longevity, and general improvements on other indicators of physical and mental well-being. There are debates about the underlying causes of the association between residential demography and indicators of health and mortality (e.g., Hearst et al., 2008; Williams and Collins, 2001). It is plausible, for example, that this association can be explained by economic factors, with more segregated neighborhoods tending to have higher levels of poverty, poorer access to resources (e.g., health care, employment, and education), and greater exposure to negative environments (e.g., poor housing quality, exposure to pollutants and smog). Social problems such as violent crime and drug use also tend to concentrate more in segregated than in diverse communities. Exploring the additive and interactive effects of these kinds of variables on health and health-related behaviors is an ongoing project. Acevedo-Garcia et al. (2003) offer a systematic review of the field and some useful suggestions for directions in which it might develop.

The second example concerns the effects of racial diversity and segregation on residents' achievement. It is well-documented, for example, that highly segregated residential environments are also environments in which economic and educational achievement is retarded (e.g., Massey and Denton, 1993), even when controlling for the more direct impact of school segregation. Card and Rothstein (2006) explored the effects of residential segregation on the racial gap in the SAT scores of black and white students and reported: (1) that this gap is higher in more segregated cities, (2) that neighborhood segregation has a more consistently negative effect on the gap than school segregation, and (3) having controlled for variance contributed by factors such as family background, moving from a 'fully segregated' to a 'fully integrated' city narrows the achievement gap by around 25%.

Similarly, in a book on the consequences of being trapped in poor, racially segregated environments, Sharkey (2009) has recently presented an important summary of research on what happens if residents either escape the ghetto via residential mobility or if their neighborhoods become more diverse and prosperous through economic investment. The results of this research are complex. Nevertheless, they offset simple, pessimistic accounts of the role of neighborhood level factors in shaping achievement. **Figure 4** depicts, for instance, the result six studies of the effects of moving out of impoverished racial ghettos on children's reading and language skills, expressed as standard deviations of the effect size of improvement (or decline) in test scores. As is apparent, the majority of studies showed some improvement in children's skills as a result of mobility and in only one case was there evidence of decline (See Sharkey (2009) for a far more nuanced interpretation of these data and also for an important discussion the role of

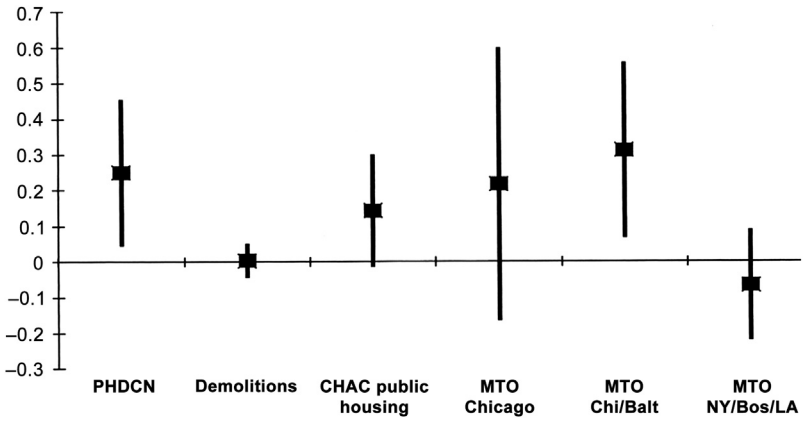

Figure 4 Effects of leaving high-poverty neighborhoods on children's reading/language skills. Note: the *x*-axis lists the name of the study; the *y*-axis shows the estimated effect of neighborhood change on test scores, expressed as an effect size.

neighborhood change as a mechanism for achieving economic and educational mobility.). In at least some circumstances, it seems, leaving the 'dark ghetto' is beneficial not only for residents' wealth and health, but also for their cognitive skills and attainment.

Conclusions: Controversies, Debates, and Future Directions

> At best, their writings reflect their identification with the pervasive racism of the society within which they were socialized and with which they are now identified. They share the values and perspectives of their political and educational peers. Many of them are 'practical' and 'realistic' as they accept and seek to justify the racial rationalizations of the society of which they are a part.
>
> Clark, 1979: p. 481

Clark's (1979) indictment of social scientific resistance to desegregation captures a defining feature of the field. Because desegregation has such significance for issues of social justice and intergroup harmony, as we hope to have shown in this brief article, research inevitably blurs the line between scholarship and advocacy. The field is characterized by heated political debates, polarized positions, and ongoing controversies. Desegregation is itself an essentially contested concept, with little agreement about what forms it should take in particular settings. In this concluding section, we outline two current areas of controversy and debate, which also highlight some directions in which research is currently developing.

Understanding the Persistence and 'Adaptability' of Segregation

As should be apparent already, research on the role of desegregation in promoting social equality and harmony is haunted by an inconvenient truth. In many historically divided societies, despite state interventions and well-meaning policies to promote integration, segregation has never truly disappeared. It is alive and well. The well-documented retreat from Brown and the continuing – and in some cases increasing (Reardon et al., 2012) – high levels of racial segregation of American schools provides a symbolic case in point that is, of course, difficult to disentangle from the tortoise-slow process of residential desegregation in many American cities (Pettigrew, 2004).

The persistence of segregation in the United States is by no means unique. For example, the first and second authors of this article have lived and worked for most of their lives in two historically divided societies, namely South Africa and Northern Ireland. In our own lifetimes, both societies have undergone remarkable transitions to democratic government that brought about the end of armed conflict, the creation of new constitutions, and sweeping changes to promote social equality. Those transitions were also expected by many observers, including ourselves, to bring about wide-scale processes of desegregation, and in some contexts they did. However, even the most optimistic analyst would concede that South Africa and Northern Ireland, like the United States, remain deeply segregated societies. Moreover, research in both societies reveals how segregation is able to adapt and metamorphisize during periods of social change, reemerging in new forms. As the Group Areas Act and similar apartheid legislation was swept aside in South Africa, for instance, whites increasingly retreated into fortified enclaves and security villages and to the safety of shopping malls and other semi-'privatized' spaces of leisure and consumption (see Hook and Vrdoljak, 2002 for a dramatic example). Their current levels of residential segregation from other groups remain high and continue to correlate strongly with their wealth and privilege. Similarly, as Northern Ireland made the transition from sectarian violence to a postconflict society, the peace walls and interface barriers that historically separated protestant and catholic communities actually proliferated in number and size. Nearly a decade and a half after the signing of the Good Friday Agreement, Belfast, the country's capital city, remains organized along starkly sectarian lines (e.g., see Jarman, 2012).

For those interested more specifically in the role of desegregation in promoting new forms of intergroup contact, the adaptability of segregation is further illustrated by recent research on its so-called 'microecological' expressions within everyday 'activity spaces' (see **Figure 5**). This work is beginning to show how the spatial organization of relations in supposedly desegregated contexts such as university lecture theaters and dining halls, school classrooms and playgrounds, public seating areas, public transport, and spaces of leisure may reproduce racial segregation at a microinteractional scale (e.g., Dixon and Durrheim, 2003; Tredoux and Dixon, 2009; Orr et al., 2012; Swyngedouw, 2013). Drawing on methodological advances in studying human mobility in cities, it is also beginning to show how segregation is reproduced via the everyday timing and spacing of ordinary activities (e.g., Palmer et al., 2014; Wong and Shaw, 2011; Wang et al., 2012); that is, how recurring patterns of division and integration emerge as citizens travel each day through the city, selecting some pathways but avoiding others, and timing their itineraries so that they encounter the 'right' kinds of people, while distancing themselves from the 'wrong' kind.

Such research is deeply troubling. It seems that even in contexts where desegregation is actively promoted and where legal segregation has been abolished, intergroup contact may occur infrequently. For this reason, in their state-of-the-art review of contact research, Pettigrew and Tropp (2011) have called for a better understanding of the 'leading a horse to water' problem. As illustrated by our earlier discussion, we now know a lot about the nature and consequences of contact between groups and about the conditions under which it facilitates prejudice reduction. We know far less about why, even in the face of institutional pressures to integrate, individuals continue to act in ways that reproduce segregation.

Writing as social psychologists, we suspect that emerging forms of segregation cannot be explained entirely by distal economic, historical, legal, or institutional factors, however important those factors remain. As research on its microecological expressions in everyday activity spaces reveals, segregation is also, at least in part, a reflection of individuals' behavioral choices in their day-to-day lives. Understanding what underlies those choices is nowadays a matter of

Figure 5 Mapping the microecology of segregation on a South African beach (see Durrheim, K., Dixon, J., 2005. Racial Encounter: The Social Psychology of Contact and Desegregation. Psychology Press, London for more detailed discussion). Note: each black circle = one black person; each white circle = one white person; each black triangle = one Indian person.

considerable controversy. There is a fairly broad consensus, for example, that residential segregation in the United States is substantively driven by choices made by white home-owners and renters about where and with whom they want to live (e.g., see Charles, 2003; Dawkins, 2004). However, there is intense debate about the nature of the motives and beliefs that underpin those choices. For some, they express a kind of 'benign ethnocentricism' in which – here researchers often use Schelling's classic simulations as a touchstone – a little bit of in-group favoritism leads to an awful lot more segregation than anyone could anticipate (e.g., Clark et al., 2004). For others, this is a rather 'innocent' interpretation that misses the impetus behind preference driven segregation (e.g., Krysan et al., 2009). As Zubrinsky and Bobo (1996) observe: "… it is not merely empirically untenable to assert that simple economics or a morally innocent mutual ethnocentrism are major components of the process of racial residential segregation: these are egregiously mistaken analyses given our results and those of recent studies… Race matters. And it matters not merely because members of any group prefer 'their own kind,' but because everyone is aware of and must adapt to the historically developed, structurally rooted, and psychologically unavoidable American racial order or hierarchy" (p. 372).

Let Them Eat Harmony? Some Ironic Consequences of Intergroup Contact and Desegregation

As we have seen, a central benefit of desegregation is that it enables interaction between formerly divided groups and that this process, in turn, leads to more harmonious intergroup attitudes and relationships. Although research in some institutional contexts (e.g., schools) has produced mixed findings, there is a growing consensus that intergroup contact generally reduces prejudice, particularly when it occurs under Allport's (1954) optimal conditions. Recent debates, however, have raised questions about the model of social change that underpins the contact hypothesis (and similar prejudice reduction interventions), which is based on the idea that discrimination and inequality can be reduced if we can get the historically advantaged to have more positive thoughts and feelings about others (Dixon, 2012a,b; Reicher, 2007; Wright and Lubensky, 2009; Reynolds et al., 2012).

The limits of this perspective are evidenced in particular by the research on the effects of intergroup contact on the political attitudes of the disadvantaged, a relatively neglected area of research. Dixon and colleagues found, for example, that positive intergroup contact improved blacks' emotional responses toward whites in postapartheid South Africa, in line with the predictions of the contact hypothesis, but also shifted their political attitudes in a conservative direction (Dixon et al., 2007, 2010). That is, contact reduced their perceptions of own-group relative deprivation and discrimination and decreased support for policies designed to combat racial inequality in the postapartheid era (e.g., affirmative action). Researchers working in other countries with other kinds of intergroup relations have reported similar findings (e.g., see Glasford and Calcagno, 2011; Saguy et al., 2009; Sengupta and Sibley, 2013). The overall message of this emerging work has been relatively consistent across research programs. Intergroup contact encourages us to like one another more and reduces negative intergroup stereotypes. However, these benefits come at the cost of decreasing the political awareness

of the historically disadvantaged, fostering acceptance of systems justifying beliefs, and producing a 'sedative effect' (Cakal et al., 2011) on collective action to promote social change.

In the light of this kind of critique, researchers have been exploring several possible routes forward. Along these lines, Dixon et al. (2013) have recently outlined the so-called incommensurablist, compatibilist, and contextualist responses to arguments about the limits of a prejudice reduction model of social change. The overarching point to be taken from this debate is not, in their view, that the creation of intergroup harmony through contact and desegregation is somehow inherently problematic. To the contrary, there are many contexts in which getting people to like one another more is a valuable end in itself. Rather, this debate raises important questions about the conditions under which different forms of contact may produce different kinds of social, psychological, and political change. It thus invites us to reflect anew on about what kind of model of social change could or should be informing the process of desegregation and what we wish to accomplish by removing barriers to interaction between groups.

See also: Collective Action; Collective Behavior, Social Psychology of; Diversity in organizations; Emotions and Intergroup Relations; Ethnicity and Migration in Europe; Friendship During Adolescence and Cultural Variations; Group Processes, Social Psychology of; Immigration: Social Psychological Aspects; Intergroup Relations; Organizational Justice; Pay, Compensation, and Performance, Psychology of; Racism: Social Psychological Perspectives; Social Psychology; Stereotypes in Social Psychology; Xenophobia: Social Psychological Aspects.

Bibliography

Acevedo-Garcia, D., Lochner, K.A., Osypuk, T.L., Subramanian, S.V., 2003. Future directions in residential segregation and health research: a multilevel approach. American Journal of Public Health 93, 215–221.

Allport, G.W., 1954. The Nature of Prejudice. Doubleday, Garden City, NY.

Amir, Y., 1969. Contact hypothesis in ethnic relations. Psychological Bulletin 71, 319–342.

Ashmore, H.S., 1954. The Negro and the Schools. University of North Carolina Press, Chapel Hill.

Brown, R., Hewstone, M., 2005. An integrative theory of intergroup contact. In: Zanna, M. (Ed.), Advances in experimental social psychology, vol. 37. San Diego, CA, pp. 255–343.

Cakal, H., Hewstone, M., Schwär, G., Heath, A., 2011. An investigation of the social identity model of collective action and the "sedative" effect of intergroup contact among Black and White students in South Africa. British Journal of Social Psychology 50, 606–627.

Card, D., Rothstein, J., 2006. Racial Segregation and the Black-White Test Score Gap. Working Paper Genealogy, in Anthropology, National Bureau of Economic Research. Massachusetts, Cambridge.

Caspi, A., 1984. Contact hypothesis and inter-age attitudes. A field study of cross-age contact. Social Psychology Quarterly 47, 74–80.

Cell, J., 1982. The Highest Stage of White Supremacy. Cambridge University Press, Cambridge.

Charles, C.Z., 2001. Processes of racial residential segregation. In: O'Connor, A., Tilly, C., Bobo, L.D. (Eds.), Urban Inequality: Evidence from Four Cities. Russell Sage Foundation, New York, pp. 217–271.

Charles, C.Z., 2003. The dynamics of racial residential segregation. Annual Review of Sociology 29, 167–207.

Clark, K.B., 1953. Desegregation: an appraisal of the evidence. Journal of Social Issues 9, 1–77.

Clark, K.B., 1979. The role of social scientists 25 years after Brown. Personality and Social Psychology Bulletin 5, 477–481.

Clark, K.B., 1965. The Dark Ghetto: Dilemmas of Social Power. Harper & Row, New York.

Clark, K.B., Chein, I., Cook, S.W., 1952/2004. The effects of segregation and the consequences of desegregation. American Psychologist 59, 495–501.

Crisp, R.J., Turner, R.N., 2009. Can imagined interactions produce positive perceptions? Reducing prejudice through simulated social contact. American Psychologist 4, 231–240.

Dawkins, C.J., 2004. Recent evidence on the continuing causes of black-white residential segregation. Journal of Urban Affairs 26, 379–400.

Desforges, D.M., Lord, C.G., Ramsey, S.L., Mason, J.A., van Leeuwen, M.D., West, S.C., Lepper, M.R., 1991. Effects of structured cooperative contact on changing negative attitudes toward stigmatized groups. Journal of Personality and Social Psychology 60, 531–544.

Dixon, J., Durrheim, K., 2003. Contact and the ecology of racial division: some varieties of informal segregation. British Journal of Social Psychology 42, 1–23.

Dixon, J., Durrheim, K., Tredoux, C., 2007. Intergroup contact and attitudes towards the principle and practice of racial equality. Psychological Science 18, 867–872.

Dixon, J., Durrheim, K., Kerr, P., Thomae, M., 2013 'What's so funny 'bout peace, love and understanding'? Further reflections on the limits of prejudice reduction as a model of social change. Journal of Social and Political Psychology 1, 239–252. http://dx.doi.org/10.5964/jspp.v1i1.234.

Dixon, J., Durrheim, K., Tredoux, C., Tropp, L.R., Clack, B., Eaton, E., 2010. A paradox of integration? Interracial contact, prejudice reduction and black South Africans' perceptions of racial discrimination. Journal of Social Issues 66, 403–418.

Dixon, J., Levine, M., Reicher, S., Durrheim, K., 2012a. Beyond prejudice: are negative evaluations the problem and is getting us to like one another more the solution? Behavioral and Brain Sciences 35, 411–425.

Dixon, J., Levine, M., Reicher, S., Durrheim, K., 2012b. Beyond prejudice: relational inequality, collective action and social change revisited. Behavioral and Brain Sciences 35, 451–466.

Dovidio, J.F., Eller, A., Hewstone, M., 2011. Improving intergroup relations through direct, extended and other forms of indirect contact. Group Processes and Intergroup Relations 14, 147–160.

Dovidio, J.F., Gaertner, S.L., Kawakami, K., 2003. Intergroup contact: the past, present and the future. Group Processes and Intergroup Relations 6, 5–21.

Dreger, R.M., 1985. On Gerard on "school desegregation". American Psychologist 40, 124–125.

Durrheim, K., Dixon, J., 2005. Racial Encounter: The Social Psychology of Contact and Desegregation. Psychology Press, London.

Durrheim, K., Dixon, J. Intergroup contact and the struggle for social justice. In: Hammock, P.L. (Ed.), Oxford Handbook of Social Psychology and Social Justice. Oxford University Press, Oxford, forthcoming.

Ellen, I.G., 1998. Stable racial integration in the contemporary United States: an empirical overview. Journal of Urban Affairs 20, 27–42.

Ellen, I.G., 2000. Sharing America's Neighbourhoods: The Prospects for Stable Racial Integration. Harvard University Press, Harvard.

Fairchild, H.H., Gurin, P., 1978. Traditions in the social psychological analysis of race relations. American Behavioral Scientist 21, 757–778.

Farley, R., Schuman, H., Bianchi, S., Coslasanto, D., Hatchett, S., 1978. "Chocolate city, vanilla suburbs": will the trend towards racially separate communities continue? Social Science Research 7, 319–344.

Gaertner, S.L., Dovidio, J.F., 2000. Reducing Intergroup Bias: The Common Identity Model. Psychology Press, USA.

Gaertner, S.L., Mann, J.A., Murrell, A., Dovidio, J.F., 1989. Reducing intergroup bias: the benefits of recategorization. Journal of Personality and Social Psychology 57, 239–249.

Gaertner, S.L., Dovidio, J.F., Bachman, B.A., 1996. Revisiting the contact hypothesis: the induction of a common ingroup identity. International Journal of Intercultural Relations 20, 271–290.

Gibson, J.L., 2004. Does truth lead to reconciliation? Testing the causal assumptions of the South African truth and reconciliation process. American Journal of Political Science 48, 201–217.

Glasford, D.E., Calcagno, J., 2011. The conflict of harmony: intergroup contact, commonalty and political solidarity between disadvantaged groups. Journal of Experimental Social Psychology 48, 323–328.

Hearst, M.O., Oakes, J.M., Johnson, P.J., 2008. The effect of racial residential segregation on black infant mortality. American Journal of Epidemiology 168, 1247–1254. http://dx.doi.org/10.1093/aje/kwn291.

Herek, G.M., Capitanio, J.P., 1996. 'Some of my best friends': Intergroup contact, concealable stigma and heterosexuals attitudes towards gay men and lesbians. Personality and Social Psychology Bulletin 22, 412–424.

Hewstone, M., Brown, R., 1986. Contact is not enough: an intergroup perspective on the 'contact hypothesis'. In: Hewstone, M., Brown, R. (Eds.), Contact and Conflict in Intergroup Encounters. Basil Blackwell, Oxford.

Hook, D., Vrdoljak, M., 2002. Gated communities, heterotopia and a 'rights' of privilege: a 'heterotopology' of the South African security-park. Geoforum 33, 195–219.

Jarman, N., 2012. Belfast Interfaces: Security Barriers and Defensive Use of Space. Belfast Interface Project (BIP), Belfast.

Krysan, M., Couper, M.P., Farely, R., Forman, T., 2009. Does race matter in neighbourhood preferences? Results from a video experiment. American Journal of Sociology 115, 527–559.

Kuchenbrandt, D., Eyssel, F., Seidel, S.K., 2013. Cooperation makes it happen: imagined intergroup cooperation enhances the positive effects of imagined contact. Group Processes and Intergroup Relations 16, 635–647.

Livingstone, D.N., 1991. The moral discourse of climate: historical considerations on race, place, and virtue. Journal of Historical Geography 17, 413–434.

Mann, J.H., 1959. The effect of inter-racial contact on sociometric choices and perceptions. Journal of Social Psychology 50, 143–157.

Maoz, I., 2011. Does contact work in prolonged asymmetrical conflict? Appraising 20 years of reconciliation-aimed encounters between Israeli Jews and Palestinians. Journal of Peace Research 48, 115–125.

Massey, D.S., Denton, N.A., 1988. The dimensions of residential segregation. Social Forces 67, 281–315.

Massey, D.S., Denton, N.A., 1993. American Apartheid: Segregation and the Making of the Underclass. Harvard University Press, Cambridge, MA.

Massey, D.S., Fischer, M.J., 2000. How segregation concentrates poverty. Ethnic and Racial Studies 23, 670–691.

McKenzie, B.K., 1948. The importance of contact in determining attitudes towards Negroes. Journal of Abnormal and Social Psychology 43, 417–441.

Miller, N., Brewer, M.B., 1984. The social psychology of desegregation: an introduction. In: Miller, N., Brewer, M.B. (Eds.), Groups in Contact: The Psychology of Desegregation. Academic Press, New York.

Nightingale, C.H., 1996. The transnational contexts of early Twentieth-Century American urban segregation. Journal of Social History 39, 667–702.

Nightingale, C.H., 2006. The transnational contexts of early twentieth century American urban segregation. Journal of Social History 39, 667–702.

Orfield, G., Lee, C., 2005. Why Segregation Matters: Poverty and Educational Inequality. The Civil Rights Project, Harvard University.

Orr, R., McKwoen, S., Cairns, E., Stringer, M., 2012. Examining non-racial segregation: a micro-ecological approach. British Journal of Social Psychology 51, 717–723.

Palmer, J.R., et al., 2014. New approaches to human mobility: using mobile phones for demographic research. Demography 50, 1105–1128.

Pettigrew, T.F., Tropp, L., 2006. A meta-analytic test of intergroup contact theory. Journal of Personality and Social Psychology 90, 751–783.

Pettigrew, T.F., 1979. Racial change and social policy. Annals of the American Academy of Political Social Science 441, 114–131.

Pettigrew, T.F., Tropp, L.R., 2008. How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. European Journal of Social Psychology 38, 922–934.

Pettigrew, T.F., Tropp, L.R., 2011. When Groups Meet: The Dynamics of Intergroup Contact. Psychology Press, Philadelphia, PA.

Pettigrew, T.F., 1998. Intergroup contact theory. Annual Review of Psychology 49, 65–85.

Pettigrew, T.F., 2004. Justice deferred: a half century after Brown V. Board of Education. American Psychologist 59, 521–529.

Pettigrew, T.F., 2008. The social scientific study of American race attitudes in the twentieth century. Personality and Social Psychology Compass 2, 339–367.

Pickren, W.E. (Ed.), 2004. Fifty Years on: Psychology and Brown V. Board. American Psychologist, vol. 59 [Special issue].

Rossell, Christine H., Armor, David J., Walberg, Herbert (Eds.), 2002. School Desegregation in the 21st Century. Praeger Publishers, Westport, CT.

Rankin, B., 2009. http://www.radicalcartography.net/index.html?chicagodots (retrieved 20.10.13.).

Ray, J.J., 1983. Racial attitudes and the contact hypothesis. Journal of Social Psychology 119, 3–10.

Reardon, S.F., Grewal, E., Kalogrides, D., Greenberg, E., 2012. Brown fades: the end of court-ordered school desegregation and the resegregation of American public schools. Journal of Policy Analysis and Management 31, 876–904.

Reicher, S., 2007. Rethinking the paradigm of prejudice. South African Journal of Psychology 37, 820–834.

Revenson, T.A., 1989. Compassionate stereotyping of elderly patients by physicians: revising the social contact hypothesis. Psychology and Ageing 4, 230–234.

Reynolds, K.J., Haslam, S.A., Turner, J.C., 2012. Prejudice, social identity and social change: resolving the Allportian problematic. In: Dixon, J., Levine, M. (Eds.), Beyond Prejudice: Extending the Social Psychology of Conflict, Inequality and Social Change. Cambridge University Press, Cambridge, pp. 48–69.

Saguy, T., Tausch, N., Dovidio, J., Pratto, F., 2009. The irony of harmony: intergroup contact can produce false expectations for equality. Psychological Science 20, 14–121.

Samelson, F., 1978. From "race psychology" to '"studies in prejudice": some observations on the thematic reversal in social psychology. Journal of the History of the Behavioral Sciences 14, 265–278.

Savelkoul, Scheepers, P., Tolsma, J., Hagendoorn, L., 2011. Anti-Muslim attitudes in The Netherlands: tests of contradictory hypotheses derived from ethnic competition theory and intergroup contact theory. European Sociological Review 27, 741–758.

Schiappa, E., Gregg, P.B., Hewes, D.E., 2005. The parasocial contact hypothesis. Communication Monographs 72, 92–115.

Schmid, K., Tausch, N., Hewstone, M., Hughes, J., Cairns, E., 2008. The effects of living in segregated vs. mixed areas in Northern Ireland: a simultaneous test of contact and threat effects in the context of micro-level neighbourhoods. International Journal of Conflict and Violence 2, 56–71.

Schofield, J.W., Eurich-Fulcer, R., 2001. When and how school desegregation improves intergroup relations? In: Brown, R., Gaetner, S.L. (Eds.), Blackwell Handbook of Social Psychology: Intergroup Processes. Blackwell, Oxford.

Sengupta, N.K., Sibley, C.G., 2013. Perpetuating one's own disadvantage: intergroup contact enables the ideological legitimation of inequality. Personality and Social Psychology Bulletin 39, 1391–1403.

Sharkey, P., 2009. Stuck in Place: Urban Neighborhoods and the End of Progress toward Racial Equality. University of Chicago Press, Chicago.

Star, S.A., Williams, R.M., Stouffer, S.A., 1949/1958. Negro infantry platoons in white companies. In: Maccoby, E.E., Newcomb, T., Hartley, E.L. (Eds.), Readings in Social Psychology. Holt, Rhinehart and Winston, New York.

Swart, H., Hewstone, M., Christ, O., Voci, A., 2011. Affective mediators of intergroup contact: a three-wave longitudinal study in South Africa. Journal of Personality and Social Psychology 101, 1221–1238.

Swyngedouw, E., 2013. The segregation of social interactions in the red line L-train in Chicago. Symbolic Interaction 36, 293–313.

Tredoux, C., Dixon, J.A., 2009. Mapping the multiple contexts of racial isolation: some reflections on the concept of scale in segregation research. Urban Studies 46, 761–777.

Wang, D., Li, F., Chai, Y., 2012. Activity spaces and socio-spatial segregation in Beijing. Urban Geography 33, 256–277.

Werth, J.L., Lord, C.G., 1992. Previous conceptions of typical group member and the contact hypothesis. Basic and Applied Social Psychology 13, 351–369.

Williams, D.R., Collins, C., 2001. Racial residential segregation: a fundamental cause of racial disparities in health. Public Health Reports 116, 404–416.

Wilner, D., Walkley, R.P., Cook, S.W., 1952. Residential proximity and intergroup relations in public housing projects. Journal of Social Issues 8, 45–69.

Wong, D., Shaw, S., 2011. Measuring segregation: an activity space approach. Journal of Geographic Systems 13, 125–145.

Wright, S.C., Aron, A., McLaughlin-Volpe, T., Ropp, S.A., 1997. The extended contact effect: knowledge of cross-group friendships and prejudice. Journal of Personality and Social Psychology 73, 73–90.

Wright, S.C., Lubensky, M., 2009. The struggle for social equality: collective action vs. prejudice reduction. In: Demoulin, S., Leyens, J.P., Dovidio, J.F. (Eds.), Intergroup Misunderstandings: Impact of Divergent Social Realities. Psychology Press, New York, pp. 291–310.

Zubrinsky, C.L., Bobo, L., 1996. Prismatic metropolis: race and residential segregation in the city of the angels. Social Science Research 25, 335–374.

Deserts, Philosophical Concept of

John Kleinig, John Jay College of Criminal Justice, New York, NY, USA; and Charles Sturt University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article reviews the structure of desert claims – the deserving, the deserved, and the desert bases – as well as their pre-institutional and institutional forms. It reviews the distinctions among desert, merit, and entitlement, considers their moral force, and engages with rejections of desert claims.

In everyday discourse, the language of desert is prominent and pervasive. It is common to speak of certain conduct (or persons) as deserving of praise or blame, reward or punishment, and of goods or ills that are visited upon people as deserved or undeserved. The subjects of desert claims may vary: most commonly they are persons or person-related entities (such as paintings that deserve prizes or cities that deserve their reputations), but insofar as inanimate or natural objects (such as rock formations) are evaluable in some way, there may be no barrier to speaking of them also as having deserts of certain kinds. The objects of desert claims are almost always seen as goods or ills. Inter alia a person may deserve success or failure, a good or bad reputation, good or bad luck, a greater or lesser portion of some social resource, compensation, or fairness.

Desert and Merit

The terms 'desert' and 'merit' are often used interchangeably. Some writers, however, have insisted on their conceptual distinction. If, for example, the ancient tie between diligence and desert (*deservire*: to serve zealously) is focused upon, the language of merit may be used to capture those outcomes warranted by other forms of value. Milne, for example, argues that:

> a natural beauty spot is *worthy of*, *merits*, or *rates* inclusion in a published guide to places of natural beauty, but it does not make sense to say that the inclusion in the guide is something *deserved* in the way of reward. The classification of human beauty (intelligence, or skill) is a parallel case: the naturally beautiful person merits a certain rating, but he or she does not deserve it any more than Niagara Falls is *rewarded* by its rating as a tourist spectacle. Only insofar as effort is expended in producing the achievement or contribution, or in acquiring a skill, is there the basis of a desert claim: a person cannot be deserving of praise or reward without trying. (Milne, 1986: p. 240)

The point is expressed generally by Pojman. Merit (or demerit), he claims, is the more inclusive term, referring to "any feature or quality that is the basis for distributing positive [or negative] attribution," whereas desert is a species of merit, "typically or paradigmatically connected with action, since it rests on what we voluntarily do or produce" (Pojman, 1999: p. 86). He then distinguishes a secondary, negative, compensatory, or rectificatory use of desert, to accommodate cases in which – relatively faultlessly – a person suffers some evil and recompense is appropriately called for.

Although a distinction of sorts can thus be made, it is at best a soft one, since a great deal of current debate about desert (see Section The Rejection of Desert) has concerned the enhancement of arbitrarily possessed advantages through diligent effort or the exacerbation of disadvantages, including some relating to effort.

Preinstitutional and Institutional Deserts

In recent philosophical discussion, a distinction has frequently been drawn between preinstitutional and institutional deserts – between deserts that are said to be 'naturally' implicated by virtue of some characteristic/activity of their subject, and those that are attributable only within the framework of some institutionalized arrangement. A person is said – preinstitutionally – to deserve his good fortune, or – institutionally – to have deserved to win a particular prize. The former is often said to be prior to the latter: "desert is a 'natural' moral notion (that is, one which is not logically tied to institutions, practices, and rules)" (Feinberg, 1963: p. 70). In many of its attributions to humans, preinstitutional desert is tied to reactive attitudes – the sense we have of others' agency, of their sharing with us at the deepest human level (Strawson, 1962). And the rejection of preinstitutional desert is often associated with a deterministic naturalism (Scheffler, 1992), more recently manifested in neuroscientific inquiry (Greene and Cohen, 2004).

Those who contend that only institutional desert claims have a legitimate place in the understanding assert that talk of preinstitutional deserts assumes, contrary to fact, that the characteristics on the basis of which such claims are made are fairly distributed. But – such critics of preinstitutional desert argue – since people naturally and arbitrarily differ in the talents, capacities, and social environments that ground what preinstitutional desert claims would allocate to them, they cannot really be said to deserve such things, or their fruits (cf Rawls, 1971: pp. 104–106, 312–314).

Leaving these latter claims aside (see Section The Rejection of Desert), it may nevertheless turn out that the distinction between preinstitutional and institutional deserts is one of degree, more than one of kind. The good weather that – it might be said – *A* deserves for vacation makes sense in part because the social institution of providing vacations has a point (recreation) to which 'good weather' intelligibly relates. Where no social institution is in place, it is much more difficult to make a specific attribution of desert. Thus, given the institution

of punishment, it may be easier to argue that those who deserve it are wrongdoers, than to argue – *ab initio* – that what wrongdoers deserve is punishment. It may well be that doing good has positive deserts, and doing ill has negative deserts, but the specific character of those deserts, whether good or ill, may not be easily determinable independently of particular social institutions (that, admittedly, are themselves open to scrutiny). It is – in part – the difficulty in establishing what, preinstitutionally, people deserve that creates some of the moral space for contemporary debates about restorative justice. Certainly, wrongdoing deserves some form of negative response, but should that response take the form of punishment (narrowly understood as fines, imprisonment), or instead as confrontation, blame, reparation, apology, penitence, and so forth? Perhaps, given that deserts are not morally determinative, punishment itself has a preinstitutional form that need not always translate into its institutional form (Kleinig, 2011).

Desert and Entitlement

But even if the distinction between preinstitutional and institutional deserts is soft, it should not lead to a conflation of desert with entitlement. '*A* deserves *p*' is not equivalent to '*A* is entitled to *p*'. Deserts seem to function comparatively or incrementally – as more or less; entitlements tend to be absolute – either we are or we are not entitled to something. Because of his or her talent and the effort put into training, *A* may be the most deserving when it comes to winning the race. But if *A* pulls a muscle and has to withdraw, *A* is not entitled to the prize, or even to win. The purpose of a race may be to determine who is the fastest runner, and *A* may be the fastest runner and therefore deserve to win. But the rules of the race – which formalize the process of determining who is the fastest runner – will determine who is entitled to win. The entitlement rules may not always succeed in establishing the most deserving, and sometimes they may even serve to exclude the most deserving.

Moral and Nonmoral Deserts

Desert often functions as a moral notion, especially where fair distributions of social benefits and burdens are involved. But it need not do so. If a person deserves success, or deserves to win a prize, or deserves good weather, one need not be talking about moral desert. True, it may seem to align with a certain cosmic fairness, but this need not be moral in character. Even if deserts are paradigmatically moral, they shade off into the nonmoral. In asserting that the Niagara Falls is deservedly famous or that the Mona Lisa deserves its reputation or even that a person's efforts deserved a better result, one is not making a moral claim. Although desert is a normative notion, it is not necessarily a moral one.

The Structure of Desert Claims

It is generally agreed that desert claims have the implicit structure '*A* deserves *p* in virtue of *x*,' where *A* is the subject of the desert claim, *p* its object, and *x* its basis.

For the most part, the subjects of desert claims will be persons/personal conduct (individual or collective) or people-related artifacts/constructs (such as art objects, cities, and so forth), although common parlance sometimes extends the language of desert to environmental and other natural objects (the sequoia, for example, may be said to deserve preservation or its reputation).

The range of deserved objects is vast: to those mentioned earlier may be added inter alia jobs, wages, grades, respect, wins, and losses. Linking all such objects is their perception as goods or ills, as benefits or burdens, as valued or disvalued. Although there may be marginal cases in which one deserves to be treated neutrally rather than well or ill (Sverdlik, 1983: pp. 320–321), such cases are parasitic on a context in which the distribution of goods and ills is primary. Sometimes – as in the case of *A*'s good weather – what seems to be deserved is a state of affairs rather than some mode of treatment by others or some achievement of one's own.

There is an ambiguity in the negative claim '*A* does not deserve *p*'. It may mean, '*A* does not deserve *p* (but deserves *q* instead)' or '*A* does not deserve *p* (or *q*)'. In the former case, we may speak of *A* being nondeserving, in the latter case of *A* being undeserving.

Desert Bases

Desert claims are more than expressions of personal assessment. They are subject to public scrutiny and criticism. In particular, some proportionality between the desert basis and deserved outcome is presupposed, although the principle of proportionality stands in need of clearer articulation than it is generally given. So, although one might be inclined to say that a person who was injured in an automobile accident – because he or she foolishly refused to use a safety belt – deserved the injuries he or she suffered (although not to be killed), we would be reluctant to say of a person who failed to wear a safety belt that he or she deserved to get injured (although we might want to say that if the person gets injured he or she will deserve it).

Most writers on desert consider that desert bases must not lie in the future, but are instead constituted by some past behavior or characteristics already possessed by the deserving object (Kleinig, 1971: p. 73; Sher, 1987: pp. 150–193). Appeals to future consequences as bases for distributing benefits and burdens may sometimes be justified, but cannot be justified as deserts. Or, at least these consequences cannot be directly justified as deserts. As indirect desert bases, reference to future-oriented considerations may be more plausible. For example, it may be argued that, although a person may deserve *p* only in virtue of his past diligence, diligence may be valued as a virtue (in part) because of its consequential/productive significance.

A few writers, however, have argued that a desert base might lie directly in the future (New, 1992, 1995; Feldman, 1995: pp. 69–71). The examples used to sustain this thesis refer to some known future facts (say, death and suffering), on the basis of which its subjects are said to deserve some recognition now. It is not clear, however, that the examples show what they are intended to (Smilansky, 1994): either there is some past or present commitment that is the basis for the desert or the

language of desert is better replaced by that of distributional fairness.

Locating Desert Bases

It has been suggested above, and argued at length by others, that the nature of a desert base can be established only by going outside the concept of desert to external goals and values that inform the activity in relation to which *p* is said to be deserved (Cummisky, 1987; Lamont, 1994). That is, merely by looking at the concept of desert, we cannot tell – except in the most general terms – what a particular act or course of conduct deserves. Moreover, because the goals and values of the activities in question may differ, depending on the context, disagreements about deservedness can become intractable if this wider context is forgotten. Disagreements may also be chronically unresolvable if there is no clarity about the goals and values implicit in the activities or institutions in relation to which attributions of desert are made. In addition, of course, there may be a further debate about the appropriateness of the goals and values to the activities or institution in question, and that debate may also impact on a discussion of the appropriateness of a particular desert basis or on the deservingness of those who are the recipients of its benefits or burdens.

Consider, for example, the institution of paying wages or a salary for work. It may not be determinable – given only a person's expenditure of productive effort outside the context of an employment contract – what the person deserves: whether he or she should receive further opportunities, praise, good fortune, sustenance, free housing, or a wage/salary. But, given that we have in place the work-related institutions that we do, we might wish to argue that *A*'s work deserves – at a minimum, assuming that it is 'full time' – a living wage. If the work that *A* does is especially demanding and/or highly valued, we might want to say that it should be more highly rewarded. Even so – to simplify the illustration – we might view *A*'s work in at least two ways, and these may be considered separately or in combination. We may focus on the demandingness of the work, on its productivity, or on some combination of the two. If the former constitutes our focus, work that manifests great effort, even if minimally productive, will be seen as most deserving of remuneration (whatever we do about entitlements). If the latter is more significant for the particular institution, then productivity will provide the basis for remunerative desert, even if the achievement reflects native talent more than effort. Given that, for many social institutions, productivity and effort will tend to be correlated, we may decide to take both into account in determining what a person's remunerative deserts should be – although if we are not sure for how much each is to count, there may be interminable disputes about a person's deserts. In such cases, rules governing entitlements will enable the institutions to function.

Consider now the salaries earned by sports and pop stars. At one level, we might wish to claim that LeBron James and Lady Gaga deserve – in terms of the institutions within which they operate – the status they possess and the remuneration they receive. But it is also open to us to debate the social value our society has accorded to sporting and singing prowess and, if as a result we conclude that they are overvalued, we will also consider that they have received more than they deserved (albeit no less than that to which they were entitled).

Where, as in many human activities, effort and achievement tend to be highly valued as desert bases, it is sometimes asserted that the desert basis – whether it focuses on effort or achievement – must be something for which the deserving person is responsible or over which he or she must have a significant degree of control. So, for example, a person cannot plausibly be said to deserve punishment unless that in virtue of which punishment is said to be deserved is something for which the person can be held responsible.

It is, however, more likely that the issue of responsibility/control is to be settled by reference to the goals and values of the institution to which the desert is related rather than to some feature of desert itself. In the case of impositions such as punishment, the idea that they could be deserved without the triggering behavior being under the person's control would be deeply problematic – although some proponents of hard determinism have argued that punishment could still be justified on consequentialist grounds. On the other hand, there is no great problem involved in thinking that those who deserve to win beauty contests are trading on genetic endowments more than on personal effort. And if inanimate objects – such as the Grand Canyon – can be said to be deservedly famous, the question of responsibility does not even arise.

There are, moreover, institutional contexts in which attributions of desert seem to require that one not have had control over one's circumstances. Consider, for example, deserved compensation. Here the deserving party is not usually seen as being in any way responsible for the misfortune in virtue of which he or she is said to deserve the compensation. Suggesting that the person who caused the misfortune is responsible puts the responsibility in the wrong place, and it is not adequately put in the right place by suggesting that the person who is deserving had to be nonresponsible for putting himself or herself in the way of the misfortune. What is more, the compensation may be deserved even if the cause of the misfortune cannot be held responsible for bringing it about.

Thus, the issue of responsibility too is determined by reference to the activity or institution in relation to which the desert claim is made. Prizes may be distributed to reward achievement or some combination of achievement and effort; wages may be given to reward productivity or some combination of productivity, talent, social importance, and effort. Some of these are more closely associated with personal control than others.

Among those who have wished to assert a tight relationship between desert and responsibility – or at least between desert and certain kinds of objects, such as punishments – several subsidiary discussions have ensued.

1. One arises out of debates about personal identity. It has been argued (notably by Derek Parfit) that since our personal identity is constituted by our beliefs, a significant change in our beliefs will reconstitute our identity as other than it was, with implications for our deserts. Parfit considers the case of a person who, at 90 years, after

receiving the Nobel Peace Prize, confesses to injuring a police officer when he was 20 years. Given the changes he has undergone, he no longer deserves punishment for what he did: "just as someone's deserts correspond to the degree of complicity with some criminal, so his deserts now, for some past crime, correspond to the degree of psychological connectedness between himself now and himself when committing the crime" (Parfit, 1984: p. 326). Although Parfit's account of identity has come in for criticism, his discussion draws attention to an important assumption about the subject of desert claims (see Fields, 1987).

2. A problem has also surfaced in discussions of moral luck (Richards, 1986; or, more generally, Rosebury, 1995). If it is simply a matter of good or bad luck that certain consequences come about through *A*'s actions, should this bear on the matter of desert? Given that luck refers to something beyond one's control, it might seem as though it should not. Nevertheless, we often act as though it does. If *A* drives recklessly through the streets and – as a matter of bad luck – an innocent pedestrian happens to step out just as he or she passes, we will be inclined to see the injury or death as exacerbating the recklessness and making the person more deserving of whatever penalty might have been due to him or her. Richards disagrees. Deserts are, he thinks, reflective of character, and the character of an unlucky and lucky reckless driver is the same. He suggests that the difference is largely an epistemic one. We are to judge on the basis of what good reason we have to know about their deserts, and the person whose recklessness (or other failing) luckily results in no harm to others does not provide us with the epistemic basis for judging the person as severely as the person who – unluckily, but not surprisingly – injures the other person. "Luck … affects how we ought to treat him, not by changing what he deserves, but by changing the grounds on which we are obliged to judge" (Richards, 1986: pp. 200–201). "To a very large extent people are lucky only in our verdicts, not in what they deserve, and their luck must affect our verdicts because it affects our grounds for reaching them" (Richards, 1986: p. 206). Richards' position has not gone unchallenged: recklessness may not deserve what is normally attributed to the driver in the event of an accident (Adler, 1987: p. 247); moreover, we may be in an epistemically better position to judge the lucky driver's character because of his record of speeding tickets (Adler, 1987: p. 247). Furthermore, Richards fails to give appropriate regard to the way in which the actual harm exacerbates the situation.

Enumerating Desert Bases

In the *Nicomachean Ethics*, Aristotle asserts that "everyone agrees that in distributions the just share must be given on the basis of what one deserves, though not everyone would name the same criterion of deserving: democrats say it is free birth, oligarchies that it is wealth or noble birth, and aristocrats that it is excellence" (Aristotle, 1962: pp. 118–119). For a modern reader, this well illustrates the dependence of desert bases on the goals and values of the activities or institutions within which desert attributions are made. In a world that is now characterized more noticeably by institutions and values that give greater weight – on the one hand – to individual responsibility and effort and – on the other – to achievement and social contribution, the bases for desert tend to be similarly characterized. As part of the tension between individual and collective values, a tension is also seen between desert bases that focus on individual responsibility, diligence, effort, and those that are more outcome directed.

Effort/Diligence

Writers who emphasize the importance of effort as a distributive desert base usually limit it in various ways. The effort must be purposeful, in the sense of being directed to the ends of the activity that grounds the desert; it must display a certain level of competence, in the sense of bearing some reasonable relation to achieving the ends of the activity in question; and the activity itself must be normatively acceptable: "There is no question of rewarding murderers and rapists, no matter how much effort has been expended by them" (Milne, 1986: p. 240).

If two people put in the same amount of effort, but one succeeds, whereas the other does not, then it might be said that they are equally deserving in virtue of their efforts. If, however, I offer a reward (or, more correctly, a prize) to the person who succeeds, then only the person who succeeds deserves it. Here the desert has been institutionalized. Our language is blurred between prizes and rewards, reflecting an overlap that sometimes occurs in competitive situations – but it should not lead us to the conclusion that achievement is 'the' proper basis for award claims.

Achievement

Those who focus on the importance of achievement draw attention to the inadequacy of effort unless it embodies competence – that is, is likely to be productive. If brilliant *A* can effortlessly produce the goods that diligent *B* cannot, then *A*, and not *B*, is likely to deserve whatever rewards or benefits are associated with that achievement. However, if it is sheer luck that results in the outcome for *A*, it is not so clear that *A* is more deserving than *B*. Even in cases in which the focus is on achievement, there seems to be some gesture in the direction of purposive effort.

The Force of Desert Claims

What follows from the claim *A* deserves *p*? Several possibilities have been canvassed: (1) *B* ought to ensure that *A* gets *p*, (2) *A* ought – all things considered – to get *p*, (3) *A* has a right to *p*, (4) There is a reason for giving *p* to *A*, and (5) *A* would be presumptively justified in getting *p*.

Most of these options are implausible. (1) It would be a rare desert claim that imposed a duty on some particular party. If *A* deserves to be punished, it does not follow that the state or some particular person should inflict it; if *A* deserves to succeed, it is unlikely to fall to any particular person to ensure that he or she does so. (2) Even if desert claims imply that someone ought to get something, they do not provide an overriding reason for it. Suppose, for example, that *A* deserves

to be punished. There may, nevertheless, be countervailing – and overriding – reasons why *A* ought not to be punished: perhaps mercy ought to be shown, or the social consequences of punishment would be disastrous. The same would hold of good deserts: other considerations of social justice may take priority. (3) Is also much too strong. (4) Seems too weak – not just because there may be many other considerations (besides deserts) that constitute reasons for giving *p* to *A* – but because deserts seem to be quite strong reasons for giving *p* to *A*. (5) Represents most closely the force of a desert claim. Such claims are presumptively strong enough to justify *A*'s getting *p* even if, in a particular case, some countervailing consideration would defeat the presumption. In practice, the situation is more complicated. If, for example, *p* refers not to some preinstitutional generality, but legal punishment or a competitive prize, other conditions may be necessary if the *p* is to be justified. In such cases, a formalized desert-determining procedure (an entitling procedure) will take precedence over the nonformal assessment of deserts.

Desert and Justice

One way of understanding desert claims is claims about justice. John Hospers, for example, writes, "Justice is getting what one deserves. What could be simpler?" (Hospers, 1961: p. 433). Desert claims then would have the force of claims about justice. But many have thought that this way of connecting desert with justice is too simple (Feinberg, 1963: p. 90; Sher, 1987: p. 49; Slote, 1973: p. 333). Even so, these writers have retained some sort of conceptual link between desert and justice. After all, when people do not get what they deserve, our immediate response is to think that they have suffered unfairly or unjustly. It is one thing to say, however, that a person's deserts constitute a factor – albeit an important one – in deciding what justice requires, and another to say – with Hospers – that deserts are the sole determinant of justice. The justice of social institutions may affect what people deserve no less than their deserts may constitute a factor in determining how they are to be justly treated (Sher, 1987: pp. 206–210).

The Rejection of Desert

Although I have argued that desert bases need not refer to conduct for which people are held responsible; nevertheless, for some distributions desert bases need to be factors for which people can be held responsible. Punishment, I noted, can hardly be said to be deserved if the conduct for which it is said to be deserved is not conduct for which the person to be punished can be held responsible.

Rawls turned this point into a general thesis against the justice of desert-based distributions. He argued that the factors that are usually adduced as desert bases in the distribution of social resources are not really things for which people can be held responsible. Those who have greater abilities and hence greater achievements have them because of their 'greater natural endowments,' not because of something for which they can be held responsible. Even a person's "character depends in large part upon fortunate family and social circumstances for which he can claim no credit" (Rawls, 1971: p. 104). It is "one of the fixed points of our considered judgments that no-one deserves his place in the distribution of native endowments, any more than one deserves one's initial starting place in society" (Rawls, 1971: pp. 103–104). Not even effort will get around this. "Even the willingness to make an effort, to try, and so to be deserving in the ordinary sense is itself dependent upon happy family and social circumstances" (Rawls, 1971: p. 74).

Although many have interpreted Rawls as taking a deterministic position on effort and achievement – seeing all effort and achievement as the inevitable outcome of native endowment and social opportunity – Rawls's rejection of desert is actually less categorical. He does not argue that we are only what we are born and bred to be, but that the task of separating out the various factors that have contributed both to our effort and to our achievement – those for which we are responsible from those that represent our endowment and opportunities – and thus enable an assessment of desert, is 'impracticable' (Rawls, 1971: p. 312). Judgments about deserts should, therefore, be excluded from assessments of social justice. Curiously, Rawls does not draw out the same implications for his briefer discussion of retributive justice.

If we allow that deserts need not depend on things for which we can be held responsible, it might not seem so problematic to maintain both that the traditional desert bases can be used as one possible ground for distribution, and also that these are not things for which we can be held – in any strong sense – responsible. Nozick has argued that "it needn't be that the foundations underlying desert are themselves deserved, all the way down" (Nozick, 1974: p. 225). If I win a race in part because of my natural talent, I may plausibly claim to deserve the prize. My personal desert does not presuppose that my talent was also deserved.

Nozick's critique goes further. Rawls infers from the arbitrariness of natural talent that such talents should be seen as common assets. Nozick does not see how the arbitrariness and undeserved nature of their natural distribution serves to disqualify those who have them from benefiting from them. Since the holdings that come into the world are already held, is it not plausible to assume that those who fortuitously have the holdings are 'preinstitutionally entitled' to the benefits that flow from them? Plausible, but not self-evident. We may consider it appropriate to grant the naturally talented and socially advantaged the benefits that flow from their fortuitous position, just as we may consider inherited wealth an entitlement of those who benefit from it. But it is hardly more obvious in the former case than in the latter that this is a matter of desert. If it appears so, this is because it is usually the case that the naturally talented and socially advantaged have to make something of their own of the endowments and opportunities they have received.

In its own way, the contemporary debate about personal desert reflects a much larger social debate about the extent to which our individuality – and that which might be said to flow from that by way of our share of social goods and ills – is viewed simply through the lens of our efforts and achievements, and the extent to which those efforts and achievements are considered as elaborations of a social

environment that has fostered and sustained them. Because the answer is unlikely to lie at either extreme, the defense of desert will inevitably be qualified.

See also: Attributional Processes: Psychological; Justice and Law; Justice, Conflicts, and the Justice of Conflict Resolution; Justice: Philosophical Aspects; Justice: Social Psychological Perspectives; Punishment: Comparative Politics of; Punishment: Social and Legal Aspects; Social Justice and Social Inequalities.

Bibliography

Adler, J.E., 1987. Luckless desert is different desert. Mind 96, 247–249.
Aristotle, 1962. Nicomachean Ethics (M. Ostwald, Trans.). Bobbs-Merrill, Indianapolis, IN.
Cummisky, D., 1987. Desert and entitlement: a Rawlsian consequentialist account. Analysis 47, 15–19.
Feinberg, J., 1963. Justice and personal desert. In: Friedrich, C.J., Chapman, J.W. (Eds.), Nomos VI: Justice. Atherton Press, New York, pp. 69–97.
Feldman, F., 1995. Desert: reconsideration of some received wisdom. Mind 104, 63–77.
Fields, L., 1987. Parfit on personal identity and desert. Philosophical Quarterly 37, 432–441.
Greene, J.D., Cohen, J.D., 2004. For the law, neuroscience changes nothing and everything. Philosophical Transactions of the Royal Society of London B, (Special Issue on Law and the Brain) 359, 1775–1785.
Hospers, J., 1961. Human Conduct: An Introduction to the Problems of Ethics. Harcourt, Brace & World, New York.
Kleinig, J., 1971. The concept of desert. American Philosophical Quarterly 8, 71–78.
Kleinig, J., 2011. What does wrongdoing deserve? In: Tonry, M. (Ed.), Retribution Has a Past: Has It a Future? Oxford University Press, New York, pp. 46–62.
Lamont, J., 1994. The concept of desert in distributive justice. Philosophical Quarterly 44, 45–64.
Milne, H., 1986. Desert, effort and equality. Journal of Applied Philosophy 3, 235–243.
New, C., 1992. Time and punishment. Analysis 52, 35–40.
New, C., 1995. Punishing times: reply to Smilansky. Analysis 55, 60–62.
Nozick, R., 1974. Anarchy, State, and Utopia. Basic Books, New York.
Parfit, D., 1984. Reasons and Persons. Oxford University Press, New York.
Pojman, L.P., 1999. Merit: why do we value it? Journal of Social Philosophy 30, 83–102.
Rawls, J., 1971. A Theory of Justice. Belknap Press of Harvard University Press, Cambridge, MA.
Richards, N., 1986. Luck and desert. Mind 95, 198–209.
Rosebury, B., 1995. Moral responsibility and 'moral luck'. Philosophical Review 104, 499–524.
Scheffler, S., 1992. Responsibility, reactive attitudes, and liberalism in philosophy and politics. Philosophy & Public Affairs 21, 299–323.
Sher, G., 1987. Desert. Princeton University Press, Princeton, NJ.
Slote, M.A., 1973. Desert, consent and justice. Philosophy & Public Affairs 2, 323–347.
Smilansky, S., 1994. The time to punish. Analysis 54, 50–53.
Strawson, P.F., 1962. Freedom and resentment. Proceedings and Addresses of the British Academy 48, 1–25.
Sverdlik, S., 1983. The logic of desert. Journal of Value Inquiry 17, 317–324.

Design and Form: Organizational

Naomi Stanford, Capella University, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by I. Palmer, R. Dunford, volume 5, pp. 3535–3538, © 2001, Elsevier Ltd.

Abstract

This article opens with a short discussion about organizational design and form arguing that it is not just about structure although that is a commonly held belief. Then follows a history of the three main theories (rational, natural, and open systems) that impact organizational design and suggest that this systems theory of organizations is giving way to one based in complexity sciences. Because these theories are both competitive and complementary, the author argues for taking a multiple perspectives approach to organizational design and form and shows why this is useful by drawing links to other subjects that influence it. Finally three major directions for developing new knowledge on the topic are presented.

> From the beginning, the forces of light and the forces of darkness have polarized the field of organizational analysis, and the struggle has been protracted and inconclusive. Perrow, 1973: pp. 2–15

This article opens with a short discussion about organizational design and form arguing that it is not just about structure although that is a commonly held belief. Then follows a history of the three main theories (rational, natural, and open systems) that impact organizational design and suggest that this systems theory of organizations is giving way to one based in complexity sciences. Because these theories are both competitive and complementary, the following section argues for taking a multiple perspectives approach to organizational design and form and shows why this is useful by drawing links to other subjects that influence it. Finally, three major directions for developing new knowledge on the topic are presented.

About Organizational Design and Form

Organizational design and form is still often taken to be the 'structure' of an organization. To improve organizational performance, there is a tendency to 'reorganize' or 'restructure' the reporting relationships between internal functional departments as shown on an organization chart and the associated coordination and linking mechanisms that facilitate the flow of information/resources between and among the business units. This view that 'structure' was synonymous with 'design' and that a 'right' structure could be developed for an organization may be traced to organization theories prevalent during the mid-/late 1970s, and based in mechanistic systems theory. At this point, Bob Duncan was able to say confidently: "Organization design is a central problem for managers. What is the 'best' structure for an organization? … This article discusses the purposes of organization structure and presents a decision tree analysis approach to help managers pick the right organization structure" (Duncan, 1979: p. 59). He went on to state that: "In discussing organization design … Organization theorists describe structure as more or less formalized, centralized, specialized, or hierarchical" (p. 61).

But consider a more recent definition of 'organization' as "a social arrangement for achieving controlled performance in pursuit of collective goals" (Huczynski and Buchanan, 2013). To make that definition come alive in the day-to-day work of an organization means thinking about design and form as much more than the 'structure' described by Duncan. Design and form emerges from making conscious decisions about explicit things like key partners, key activities, key resources, value propositions, customer relationships/segments, channels, cost structure, and revenue streams.

Yet design and form is not only about these explicit, easily described, aspects. It is also about recognizing (if possible) the power and impact of the tacit aspects of an organization – the ones that frequently defy conscious decisions made in relation to them. Such aspects include the cultural norms, social behaviors, management styles, organizational politics, language and tone, intuition, insights, decision-making processes, conflicts, emotions, and all the dimensions of work that are hard to pin down (Nonaka and Takeuchi, 1995).

These tacit aspects are integral to organizational design and form as becomes clearer when the idea that organizations are born from conversations is presented. Conversations are undirected exchanges with participants drawing meaning from each other's speech, building on what has gone before, and in the course of the discussion deriving a collective understanding of possibilities. Hewlett–Packard (HP), for example, established in 1939 developed from conversations between original founders David Packard and William Hewlett as they developed an audio oscillator in a garage. Their first customer was Walt Disney Studios who ordered eight of these. Over decades the company has developed into a global provider of software, hardware, and associated services. An organization does not spontaneously spring into being. It grows from interactions between people who share thinking and interests and focus toward a common goal (e.g., fulfilling a first order). Founder stories are rich sources of information on how organizations emanate from conversations.

These explicit and tacit organizational 'ingredients' are not working in an enclosed 'container.' They are influenced and mixed by constantly moving external conditions: shifting politics, rising environmental concerns, changing patterns of society/demography, increasing types of information and communication technologies, and demanding new legal/compliance requirements to name just a few. These too influence an organization's design and form (and are influenced by it).

Thus to see organizational design and form only in terms of structure is to get a very partial view of an exceptionally rich

arena. Certainly, the one element to be considered in 'designing' is an organization, but perhaps it is no longer the one to be considered dominant or even hugely important. As research on 320 companies on the relationship between structure, functioning, and effectiveness reported: "no direct relationship between structure and effectiveness was found. When functioning is conceptualized as a mediating factor no direct causality between structure and effectiveness is implied. … To improve effectiveness re-organizing is probably not the first option to consider" (Andersen and Jonsson, 2006: p. 257).

Theories of Organizational Design and Form

Over the years, organizational researchers and commentators have theorized from a variety of perspectives on the most efficient and effective way of organizing people to work collectively to achieve a specific purpose or outcome. Theorists broadly fall into one of the two schools: the systems (rational) school that focuses attention on the explicit elements of organization and the human relations (natural) school that focuses attention on the tacit elements of organization (Gouldner, 1959). The distinction is somewhat artificial as each school recognizes its interdependence with the other, but the 'angle of approach' as it were is what differentiates one from the other. Additionally, each of those schools is itself fragmented, offering often competing, contested, and conflicted perspectives and theories on organizations and each is continuing to advance its theories.

The rational systems school traces its roots to the early 1900s with the publication of Frederick Winslow Taylor's Principles of Scientific Management. At this point, organizations were observed as more or less closed systems operating independently of their context. They "exhibited a high degree of formalization. The co-operation among participants is 'conscious' and 'deliberate'; the structure of relations is made explicit and can be 'deliberately constructed and reconstructed.' … A structure is *formalized* to the extent that the rules governing behavior are precisely and explicitly formulated and to the extent that roles and role relations are prescribed independently of the personal attributes and relations of individuals occupying positions in the structure" (Scott, 2003). In this conception of organizations, their form and design could be consciously determined, sculpted, and shaped to achieve a specific goal or goals.

The natural (human relations) school has origins in the work of Chester Barnard and Elton Mayo in the late 1930s. Mayo, in contrast to the scientific view of management promulgated by Taylor, was of the view that "So long as commerce specializes in business methods which take no account of human nature and social motives, so long may we expect strikes and sabotage to be the ordinary accompaniment of industry" (Harvard Business School and Baker Library Historical Collections, n.d.). The focus of this school of thought was on the individuals working in the organization; their social relationships and their motivations and rewards for contributing their effort to meet organizational goals. "It is their energy, their ideas, their conformity and non-conformity that constitutes and shapes the structure of the organization and carries on its functions" (Scott, 2003).

By the mid-1960s, in signs of some partial unity between the two schools, theorists in both camps were examining the notion that design starts with the business strategy, and were describing organizations as open systems responsive to their environment. In *Differentiation and Integration in Complex Organizations* (Lawrence and Lorsch, 1967: p. 1), for example, the authors explain: "This is a comparative study of six organizations operating in the same industrial environment. The subsystems (sales, research, and production) in each organization were differentiated from each other in terms of subsystem formal structures, the member's goal orientation, member's time orientations, and member's interpersonal orientations."

From this and similar research (e.g.,.Thompson, 1967) contingency theory was developed. Briefly put, this suggests that organizations vary in design and form contingent on their operating environment and the types of products and services they are offering. Theorists agreed that there is no one right design and form for organizations or even an organization. What is 'right' depends on a number of variables and the way these interact to shape and be shaped by their operating environment.

Even so, until the early 2000s, there was a prevailing view that in order to function effectively and achieve performance success, organizations needed 'fit,' 'congruence,' 'alignment,' and/or 'equilibrium' between various organizational components and subsystems. Nadler and Tushman, talking about their congruence model, for example, commented that "strategy drives organizational architecture, a term that describes the variety of ways in which the enterprise structures, coordinates and manages the work of its people in pursuit of strategic objectives. … The more closely each component of the organization is aligned with the others – and with the strategy – the more effective the organizational performance" (Nadler and Tushman, 1999: p. 46).

Organizational systems models (showing the components for 'alignment') were developed in line with this thinking. Ones in common use from the early 1970s (and still used 40 years later) include the Star model (Galbraith, 1973), 7-S model (Pascale and Athos, 1981), Congruence model (Nadler et al., 1992), Burke Litwin model (Burke and Litwin, 1992), Weisbord's six box model (Weisbord, 1976), and Leavitt's Diamond model (Leavitt, 1965).

These models are useful as shorthand for thinking about an organization's design and form from more than one angle, but limited in that they cannot convey the interactions, emotions, and experiences of organizational functioning from either a rational or a natural or open systems perspective. Anyone who has worked in or with an organization will be clear that thinking dogmatically about it solely in terms of 'components' would not work to introduce, maintain, or develop a design and form that resulted in 'a social arrangement for achieving controlled performance in pursuit of collective goals.'

Additionally, getting to 'alignment,' which all the models aim toward, implies reaching a point of stability which could never happen: all organizations are in a constant state of flux (people leave or join, competitors emerge, other external factors come into play, and so on). So even without conscious design work, an organization's design and form is constantly changing in unpredictable ways: design works with paradox – design and

form is planned and emergent, explicit and tacit, systematic and conversational, enabling and constraining, differentiated and integrated (Rodgers, 2010).

Recognizing the inherent paradoxes in organizational design and form fosters a more spontaneous, potentially self-organizing view of organizations where order emerges from disorder. This finds voice in management researchers who are, in a departure from systems thinking, looking for insights from complexity sciences 'as a source domain for analogies' (Stacey et al., 2000). Stacey, among others, has come to view organizations as complex patterns of interactions that involve conversations, power dynamics, social cuing, cultural norms, and so on between individuals as they seek to explain and make sense of something, agree a course of action, or simply work together. Even with formal controlling structures like performance management systems, grade levels of staff, and rewards and penalties, the outcome of the informal interactions is not predictable or certain. This uncertainty results in ongoing interactions to create an organizational direction and future (Stacey, 2012).

Taking Multiple Perspectives

Although some theorists and academics stoutly defend their perspectives, others, for example Hatch and Cunliffe (2006), observe that it makes sense to commit to not a single perspective on organizational theory but a multiple perspective. And another academic who has done just that – committed to multiple perspectives from which to consider design and form – is Gareth Morgan. In his book *Images of Organization* he presents eight images (metaphors) of organization: machines (the systems perspective), organisms, cultures (the human relations perspective), brains, psychological prisons, instruments of domination, flux and transformation (complexity sciences), and political systems (Morgan, 2006).

Each one of these offers a multiplicity of ways in which to assess, interpret, and design an organization and each metaphor is associated with a large body of theory drawing from diverse disciplines that illustrate the many links that organizational design and form has to other subject areas: anthropology, architecture, strategy, communication, social and industrial psychology, culture, biosciences, economics, history, computer and information sciences, decision sciences, political science, and sociology to name but a few.

Additionally each metaphor offers a path to further theory construction and paves the way for the development of practitioner tools and intervention approaches – i.e., the 'how to' design methods. In presenting the eight metaphors, Morgan reinforces the notion that there is no 'right way' *to* design and form an organization and neither is there one right design or form *of* an organization. As he explains "Every metaphor is presented as a framework that generates both strength and limitations … Think 'culture' and you'll see all kinds of cultural dimensions. Think 'politics' and you'll find politics. Think in terms of system patterns and loops, and you'll find a whole range of them" (Morgan, 2011).

Taken as a set, it is easy to see in the eight images the competing and complementary theories of organizational design and form. There is practical value in recognizing that there is no unifying theory (or practice) related to it because the absence of an integrated theory 'shows the inherent incompleteness of any particular point of view.' As Morgan remarks holding only one perspective as "a way of seeing becomes a way of not seeing; and that any attempt to understand the complex nature of organizations (as with any other complex subject) always requires an open and pluralistic approach based on the interplay of multiple perspectives" (p. 466). It thus behooves leaders and organizational designers to consider a diversity of options, opinions, and possibilities as they strive to achieve 'a social arrangement for achieving controlled performance in pursuit of collective goals.'

New Directions for Knowledge

New directions for knowledge on organizational design and form are not easy to pinpoint. But there are three starting points to look for:

1. Bridging theory and practice in a way that develops actionable knowledge.
2. Looking at trends and fads that are 'out there' and potentially learning/developing knowledge from them.
3. Scanning, listening, and looking out for items of information that are not a trend or a fad but could have a big impact on 'connecting the dots,' or getting a completely new insight or perspective that developed new knowledge.

Bridging Theory and Practice in a Way that Develops Actionable Knowledge on Organizational Design and Form

Whether holding a single perspective of organizational design or form or a multiple one, in general academics tend to fight shy of discussing what will make the theories actionable in day-to-day organizational life: a gap exists between theory and practice attributed to a number of possible causes including academic 'fragmentation,' 'self-absorption,' and 'living in a self-indulgent state' (Starbuck, 2003). So although it is obvious that "there is a very large body of [theoretical] knowledge about organizations and organizing, examples of effective applications of this knowledge in designing real organization are few and far between" (Meyer, 2013: p. 16).

In the day-to-day (and as somewhat of a generalization), many organizational managers are distrustful of academic 'theory' and 'ivory tower thinking.' They just want something practical they can do to improve performance or achieve whatever it is that they are aiming for. And from the other side of the fence, academics are rather dismissive of fads and quick fixes that are not underpinned by rigorous and validated research.

Meyer, reporting on a 2013 organizational design community conference with both academic and practitioner participants, observed that "most established design efforts are rooted in a rational model of action … scholars should understand organizations, consultants should translate scholars' understandings and practitioners should take action based on understanding" but noted that several conference

participants "offered support for a model of action that accumulates knowledge through feedback from experience instead of through analysis and anticipation. … In this model action becomes the basis for understanding" (Meyer, 2013: p. 20). He presents a table of assumptions on the relationship between knowledge and action (Table 1).

The inherent danger of moving from one set of assumptions toward another is that the emerging assumptions become the new 'way it is,' leaving no room for competing and equally valid approaches. The risk of that happening could be minimized (and yield new knowledge) by testing each of the emerging assumptions individually and collectively in a variety of situations and by a variety of means: generating new metaphors, developing and testing scenarios, and working as a multidisciplinary team on real organizational design issues to get to a solution and simultaneously develop theory around it are some methods for doing this.

Table 1 Assumptions on the relationship between knowledge and action

Established assumptions	*Emerging assumptions*
Model of action: Understand → Act → Evaluate	Model of action: Evaluate → Act → Understand
Designers must understand organizations before they attempt to change them.	Designers cannot understand organizations until and unless they try to change them.
Organizational design should be undertaken in an episodic or periodic fashion.	Organizational design should be an ongoing, continuous process.
Designing should be decomposed into specialized roles that are invoked sequentially. Scholars should understand, consultants should translate, and practitioners should implement.	Designing should unfold as an iterative sequence of experiments in which scholars, consultants, and practitioners collaborate in acting, evaluating, and designing.

Looking at Trends and Fads that Are 'Out There' and Potentially Learning/Developing Knowledge from Them

Trends and fads are not the same thing. A trend is month-by-month or year-by-year movement. Trends are often shown graphically as 'trend lines' drawn from quantifiable metrics collected over time. For example, the graphic below shows trend data for U.S. adults' Internet adoption between 1995 and 2013:

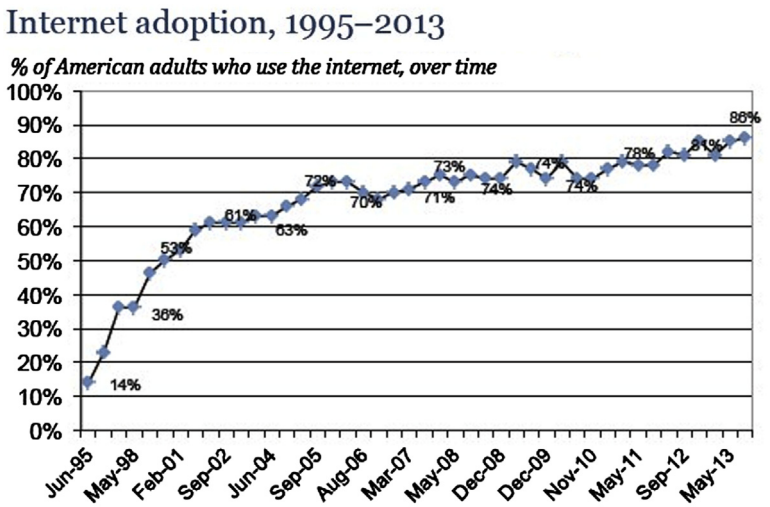

Source: Pew Internet & American Life Project Surveys, March 2000–September 2013. All surveys prior to March 2000 were conducted by the Pew Research Center for People & the Press.

A fad is a "transitory collective beliefs, disseminated by the discourse of management-knowledge entrepreneurs that a management technique is at the forefront of rational progress" (Williams, 2004) and because there is a gap between theory and practice, the field is left open for management fad adoption (a topic that is researched in its own right).

In the day-to-day, fads have a much heavier impact on organizational design and form than the impact of academic research and theory. As one academic notes, "the influential management fads-such as Japanese management, downsizing, reengineering, teamwork, Quality Circles, Six-Sigma quality management, the Learning Organization, outsourcing, knowledge management – have originated from managers and consultants" (Starbuck, 2003). In a challenge to the first emerging assumption in the table above, the adoption of the fads has been, in many cases, followed by academic research on their effectiveness (see, for example, Zairi et al., 1994) – so what is presented as an emerging assumption could also be an established assumption.

Managers in organizations are likely to be attracted to a fad if it is:

1. Simple to understand, easy to communicate, and associated with buzzwords and catchphrases.
2. Prescriptive in its approach – it tells manager what to do in things like seven steps, or five phases.
3. Encouraging of the successful outcomes by raising hopes even if there is little in the way of evaluative process attached to the sales pitch.
4. Universally relevant or one-size-fits-all as shown by exemplars who have already adopted the fad.
5. Easy to apply in practice or even partially apply by taking some elements of the method – this makes it easier to graft on to existing operations without whole scale change.
6. Able to speak directly to business issues of the day, e.g., to downsizing in a recession.
7. Interesting because of its novelty but not so radical as to disturb the underlying status quo.
8. Given legitimacy by consultants or and their successful devotees – an endorsement by a management celebrity and/or their followers goes a long way even without any evidence of true results (Miller and Hartwick, 2002).

Over time, some things that appear fads may become trends and then mainstream. In any event, adopting a fad or following a trend often has a learning value in its own right that makes it worth the investment assuming, of course, that there is conscious effort and reflective activity to generate the new knowledge perhaps through research and then theory generation as suggested above.

Scanning, Listening, and Looking Out for Items of Information

The third area for development of new knowledge is in what might be dismissed as random horizon scanning. Nevertheless activities such as following what futurists are saying (see the journal *Futures*, published by Elsevier which is concerned with the futures of international business), subscribing to a range of newsletters from the multiversity of disciplines mentioned above, and following blogs/tweets offering differing perspectives and world views are ways of contributing to the development of new knowledge. Two examples serve to illustrate: the first is an extract from an article on Bitcoin: "Peek into the future, and it's possible to envisage this sort of technology being used to cut intermediaries out of trades of many kinds – beginning with payment systems such as Visa and moving on to banks, real estate and more" (New Scientist, 2013) – this is exactly the type of snippet that an organizational designer working in the financial sector should seize on.

The second is the 'submit a hack' MIT: The Unlimited Human Potential Challenge inviting people to present case examples of the future of business. The challenge explains that "The Social Web and other digital technologies and platforms offer a robust set of alternatives to industrial-era forms of organizing. We now can imagine organizations

- Where coordination happens without centralization,
- Where power is the product of contribution rather than position,
- Where the wisdom of the many trumps the authority of the few,
- Where novel viewpoints get amplified rather than ignored or suppressed,
- Where communities form spontaneously around shared interests,
- Where opportunities to 'opt-in' blur the line between vocation and hobby, where titles and credentials count for less than value-added,
- Where performance is judged by your peers,
- Where influence comes from sharing information, not from hoarding it" (Management Innovation Exchange, 2013).

In fact, there are growing numbers of examples of organizations exhibiting these characteristics. Four that are gaining ground are 'sharing organizations,' 'crowd-sourced organizations,' 'virtual organizations,' and 'self-organizing networks.'

Sharing organizations are online services based on a principle that enable people to share the assets they own: cars, accommodation, bicycles, household appliances, and other items. Their role is to connect owners of underused assets with others willing to pay to use them. These types of collectivities are disrupting established companies in the sphere who are in turn starting to embrace the model themselves.

Crowd-sourced organizations are those that draw on society for input to their operation. Governments are increasingly looking to this as their budgets are reduced and their citizens are demanding more. Many are seeing opportunities to use digital technologies that give access to 'the crowd' to change the way they interact with citizens, save money, and provide services in new ways. As one commentator noted, "Many policymakers are beginning to look to society to help address social problems they know government cannot solve alone. The funding gap at the front-line is increasingly being filled by innovative financing mechanisms that blend private investment with public goals and a social return. Social media and open data are already forcing inward-looking bureaucracies to share decision-making and open up service delivery" (Cooper et al., 2013).

Virtual organizations comprise employees who are geographically distributed but who work effectively in groups or teams using available technologies. Call center and tech companies are increasingly known for this organizational form and others – including governments are moving toward a distributed workforce. The positives of a virtual organization include being able to access a very wide labor market, and offer employee benefits such as home working, or no requirement to relocate. The negatives are those associated with managing a distributed team and creating a sense of community and purpose necessary to achieve a common goal.

Self-organizing networks are the ones able to respond continuously to context changes without being hampered by restrictions of a rigid or well-established organizational infrastructure that might include policies, grading systems, and management controls. In these self-organizing networks, the resources and the individuals coalesce for a temporary period to achieve something that individually they could not do. Grassroots neighborhood organizations that form to protest against the closure of a manufacturing plant, or the siting of a power station, are typical of self-organization where leaders emerge 'in the moment.' Such organizations often disband when an outcome has been realized. Self-organizing like this is also seen within organizations – for

example, if there is a will to change a policy or develop a new direction. Some organizations, Google among them, actively encourage self-organizing interest groups to form with the intention that they will explore and develop ideas that could lead to innovative products and services. This self-organizing within an established enterprise is often the precursor or initiator of a new organizational form. All these four types of new organizational forms rely on information and communication technologies and have the potential to upend established organizational forms. It remains to be seen whether these and other new types of organizations arising in emerging circumstances with new technologies will displace the established organizations, whose design and form originated in different circumstances. Research appears to confirm that it is more often the case than not that they do. "With a few exceptions, the only instances in which mainstream firms have successfully established a timely position in a disruptive technology were those in which the firms' managers set up an autonomous organization charged with building a new and independent business around the disruptive technology" (Christensen, 2013 reprint edition).

Conclusion

This article opened with a quote that although said in 1973 holds true 40 years on. The field of organizational design and form is still polarized and struggling and that is what makes it fascinating. There is room for more research, deeper learning from practice, new theories, and much more.

See also: Corporate Culture; Development: Organizational; Ecology: Organizations; Modularity and Organizations; Narrative Approaches to Organizations; Network Paradigm: Applications in Organizational Science; Organization: Overview; Organizational Path Dependence; Organizations, Sociology of.

Bibliography

Ahuja, A., 2013. Neuroscience on trial. RSA Journal 159 (5556), 41.
Andersen, J.A., Jonsson, P., 2006. Does organization structure matter? On the relationship between structure, functioning and effectiveness. International Journal of Innovation and Technology Management 3 (3), 237–263.
Burke, W.W., Litwin, G.H., 1992. A causal model of organizational performance and change. Journal of Management 18 (3), 523–545.
Christensen, C., 2013 reprint edition. The Innovator's Dilemma. Harvard Business Review Press, Boston.
Cooper, T., Robinson, M., Kippin, H., September/October 2013. Delivering Public Service for the Future: Harnessing the Crowd. Retrieved 8 February 2014 from Ivey Business Journal: http://iveybusinessjournal.com/topics/strategy/delivering-public-service-for-the-future-harnessing-the-crowd#.UvZmOkeYb-e.
Duncan, R., 1979. What is the right organization structure? Organizational Dynamics 7 (3), 59–80.
Galbraith, J., 1973. Designing Complex Organizations. Addison Wesley, New York.
Gouldner, A., 1959. Organizational analysis. In: Merton, R.K., Broom, L., Cottrell, L.S. (Eds.), Sociology Today: Problems and Prospects, vol. II. Harper Torchbook, New York, pp. 400–428.
Harvard Business School, Baker Library Historical Collections, n.d. Retrieved 3 February 2014 from Enter Elton Mayo: http://www.library.hbs.edu/hc/hawthorne/04.html#four.
Hatch, M.J., Cunliffe, A.L., 2006. Organization Theory, second ed. Oxford University Press, Oxford.
Huczynski, A., Buchanan, D., 2013. Organization Behaviour, eighth ed. Pearson Education Ltd, Harlow.
Lawrence, P.R., Lorsch, J.W., June 1967. Differentiation and integration in complex organizations. Administrative Science Quarterly 12 (1), 1–47.
Leavitt, H.J., 1965. Applied organizational change in industry: structural, technological and humanistic approaches. In: March, J.G. (Ed.), Handbook of Organizations. Rand McNally, Chicago, pp. 1144–1170.
Lindebaum, D., Zundel, M., June 2013. Not quite a revolution: scrutinizing organizational neuroscience in leadership studies. Human Relations 66 (6), 857–877.
Management Innovation Exchange, 2013. The Unlimited Human Potential Challenge. http://www.mixprize.org/sites/default/files/features/uhp/sap-uhp-cheat-sheet.pdf?3.
Meyer, A., 2013. Emerging assumptions about organization design, knowledge and action. Journal of Organization Design 2 (3), 16–22.
Miller, D., Hartwick, J., October 2002. Spotting management fads. Harvard Business Review, 26–27.
Morgan, G., 2006. Images of Organization. Sage Publications Inc. Thousand Oaks.
Morgan, G., 2011. Reflections on images of organization and its implications for organizations and environment. Organization and Environment 24 (4), 459–478.
Nadler, D.A., Tushman, M.L., January 1999. The organization of the future: strategic imperatives and core competences for the 21st century. Organizational Dynamics 28 (1), 45–60.
Nadler, D., Tushman, M., Shaw, R., 1992. Designing organizations that have good fit. In: Nadler, D., Gerstein, M. (Eds.), Organizational Architecture. Jossey Bass, San Francisco, pp. 39–56.
National Science Foundation, 14 October 2010. Virtual Organizations as Sociotechnical Systems. Retrieved 8 February 2014 from National Science Foundation: http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503256.
New Scientist, 30 November 2013. The b-commerce revolution. New Scientist 3.
Nonaka, I., Takeuchi, H., 1995. The Knowledge Creating Company. Oxford University Press, Oxford.
Pascale, R.T., Athos, A.G., 1981. The Art of Japanese Management. Simon & Schuster, New York.
Perrow, C., 1973. The short and glorious history of organizational theory. Organizational Dynamics 2 (1).
Rodgers, C., 11 March 2010. Paradox, Complexity and Organizational Dynamics. Retrieved 10 February 2014 from Informal Coalitions: http://informalcoalitions.typepad.com/informal_coalitions/2010/03/paradox-complexity-and-organizational-dynamics.html.
Scott, W.R., 2003. Organizations: Rational, Natural and Open Systems, fifth ed. Prentice Hall, Upper Saddle River.
Stacey, R.D., Griffin, D., Shaw, P., 2000. Complexity and Management: Fad or Radical Challenge to Systems Thinking? Routledge, London.
Stacey, R., 23 November 2012. Trust in Organizations. Complexity and Management Centre. http://complexityandmanagement.wordpress.com/2012/11/23/trust-in-organisations/.
Stanford, N., 2013. Organizational Health: An Integrated Approach to Building Optimum Performance. Kogan Page, London.
Starbuck, W.H., 2003. Shouldn't organization theory emerge from adolescence? Organization 10 (3), 439–452.
Thompson, J.D., 1967. Organizations in Action. McGraw-Hill, New York.
Weisbord, M.A., December 1976. Organizational diagnosis: six places to look for trouble with or without a theory. Group & Organization Studies, 430–447.
Williams, R., 2004. Management fashions and fads. Management Decision 42 (6), 769–780.
Zairi, M., Letza, S.R., Oakland, J.S., 1994. Does TQM impact on bottom-line results? The TQM Magazine 6 (1), 38–43.

Relevant Websites

Organization Design Communities and Membership Organizations

http://organizationdesignforum.org/.
http://www.organisationdesignforum.eu/.
http://orgdesigncomm.com/.
There are several organization design LinkedIn groups

Consultancies

Most of the big consultancies have an organizational design practice (Accenture, Capgemini, Deloitte, Booz, PwC, Tata Consultancy Services, etc).

Business Magazines

Harvard Business Review
McKinsey Quarterly
Strategy+Business
MIT Sloan Management Review

Determinism: Social and Economic

Warren Schmaus, Illinois Institute of Technology, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social and economic determinism is the thesis that human beliefs and practices are to be explained by social and economic factors. However, sociologists rarely mean 'determine' in any mechanical, causal sense. Those such as Marx and Durkheim who have been considered determinists have typically argued that our beliefs and practices are adaptations to, rather than effects of, social and economic conditions. More recent versions of this thesis include the 'strong programme' in the sociology of scientific knowledge of David Bloor and his associates and the feminist standpoint epistemology of Sandra Harding and others.

Social and economic determinism is the thesis that human beliefs, practices, and actions are to be explained in terms of social and economic factors. Few social theorists have ever maintained this as a strictly causal, mechanistic thesis. However, critics will often interpret social and economic determinism in this way for rhetorical effect. When one looks more closely at major social theorists such as Karl Marx and Emile Durkheim who have been charged with economic or social determinism, they can be read as having offered functional rather than causal explanations of people's views. Karl Mannheim understood Marx this way. On this reading, people's religious or political beliefs are not caused by their social or economic positions, but rather have the effect of helping to maintain people in these positions. Explanations that appeal to social and economic interests, which the Strong Programme in the sociology of scientific knowledge have recently invoked to account for scientists' theoretical commitments, are functional explanations of this sort. Similarly, Michael Mann's accounts of the sources of social power are closer to interest or functional than causal explanations. Feminist standpoint epistemology borrows the Marxian or Mannheimian notion that a person's point of view may be explained in terms of their social location within a hierarchy of power relations, but allows for other contributing factors as well and is not fully deterministic.

Deterministic philosophies of one sort or another of course date back to the pre-Socratic philosophers, but the universal determinism defended by some philosophers and religious thinkers does not appeal to specifically social or economic causes. Nor would theories that explain social phenomena in terms of biological and physical causes be considered forms of social or economic determinism. Thus, Social Darwinists such as Herbert Spencer are not social or economic determinists. Social and economic determinism has its historical antecedents in sociologists like Auguste Comte and economists like David Ricardo, who held that social and economic phenomena are to be explained by specifically sociological and economic laws. Although Comte regarded sociological laws as deterministic, he did not give any empirical evidence to support this position. Indeed, he not could have. As David Hume argued, experience can show us only that one thing has always followed another; it can never show us that these things *must* follow in this way. Like Spencer and other nineteenth-century social theorists, Comte appears to have unwittingly introduced a metaphysical assumption into his social science.

Causality and Determinism

This idea that human beliefs, practices, and actions are determined is often confused with the claim that these results are inevitable and predictable. However, phenomena may not be predictable if the conditions on which they depend are either unknown or, even when they are known, are so complex as to make it humanly impossible to deduce their consequences. Also, one may believe in inevitability for mystical or spiritual reasons. Nevertheless, for whatever reason one holds the thesis that the course of human events is inevitable, it is a metaphysical claim that cannot be empirically tested.

Causal deterministic laws have the form "under conditions C, effect E will result," in which C is said to be a sufficient condition for E. That is, in a deterministic law, the relationship between cause and effect is not merely probabilistic or statistical. It is invariable and unconditional. However, even those such as Ludwig Gumplowicz who maintain that social phenomena are governed by deterministic laws rarely, if ever, provide examples of such laws. Their putative causal claims and the arguments put forth in defense of them typically fall short of providing a set of conditions sufficient for determining the beliefs, practices, or actions to be explained. Frequently, at best merely necessary conditions are given. Sometimes, even these conditions are necessary only under a certain set of circumstances.

Most often a purported causal relationship between certain social or economic conditions and the beliefs and practices to which they are supposed to give rise turns out to be a functional relationship instead. Functional accounts of beliefs and practices show them to be adaptations to, rather than results of, these and some additional social and economic conditions. A functional explanation differs from a causal explanation of some phenomenon by providing, instead of a sufficient condition for its occurrence, one of its effects that help to maintain that phenomenon in existence. Such an effect is only a necessary and not a sufficient or determining condition for that phenomenon. When providing a functional explanation, one must avoid the trap of simply pointing to what appear to

be the obvious beneficial consequences of that phenomenon and adduce actual evidence that these consequences actually helped to maintain that phenomenon in existence.

Marx and Economic Determinism

There are two different but related senses of determinism in Marxist thought. The first is historical determinism or materialism, according to which the course of history is to be explained through material or economic causes. The second is Marx's theory that what he calls the economic 'base' of a society, which consists in the modes and relations of economic production, somehow 'determines' a 'superstructure' of laws, politics, religion, and other ideological beliefs. Marx claimed to be offering not just another social philosophy or philosophy of history, but an actual science with deterministic laws that made it possible to predict the coming proletarian revolution in the most advanced capitalist economies.

Historical Materialism

'Historical materialism' is actually Friedrich Engels', not Marx's term. According to the materialist conception of history, there is an inevitable class struggle between the bourgeoisie and the proletariat. This belief has antecedents in earlier thinkers, most notably David Ricardo. For Ricardo, class conflict is the inevitable result of a population increase leading to the bringing of marginal lands under cultivation, which leads to wages dropping to a subsistence level. For Marx, on the other hand, class conflict was not a result of population growth but of changing modes of economic production. As new modes of production develop, they come into conflict with the social relations that existed under the previous economic system. The older social relations are encumbrances on these new modes of production. Eventually these older relations give way to new ones that are more in line with the new form of production. Each major change in the social relations of production, from slavery to feudalism, from feudalism to bourgeois capitalism, and from capitalism to socialism, is accompanied by class conflict.

Base and Superstructure

Marx appears to have thought that there are irreconcilable differences in thought between the bourgeoisie and the proletariat due to their different economic positions in society. According to Marx's famous 'Preface' to his *A Contribution to the Critique of Political Economy* (1859):

> In the social production of their existence, men inevitably enter into definite relations, which are independent of their will, namely relations of production appropriate to a given stage in the development of their material forces of production. The totality of these relations of production constitutes the economic structure of society, the real foundation, on which arises a legal and political superstructure and to which correspond definite forms of social consciousness. The mode of production of material life conditions the general process of social, political and intellectual life. It is not the consciousness of men that determines their existence, but their social existence that determines (*bestimmt*) their consciousness.

Although this passage has been read as suggesting that individuals are mere puppets pulled by economic and social strings, it is not at all clear that Marx meant 'determines' in any causal or mechanistic sense. The German word '*bestimmt*' means 'determine' in the sense of decide, define, set limits or boundaries to, or fix something, such as a price or time. Furthermore, the very building metaphor implied in the language of base and superstructure suggests that the form of economic life is at best a necessary condition and not a sufficient cause for that which rests upon it. On the mechanistic reading of 'determines', Marx would have been guilty of an inexcusable non sequitur from premises about necessary conditions to a conclusion about sufficient conditions. Finally, Marx objected to the practice of taking the causal powers of human beings and projecting them onto mere abstractions such as god, the state, or capital. In his *Theses on Feuerbach* (1845), he had criticized the materialists for doing this, and yet this is precisely what the mechanistic reading of the 'Preface' does.

If one reads 'determine' in the sense of deciding, Marx appears to have said that social and economic conditions choose or select rather than cause certain forms of thought. Marx would then have been using a form of explanation that Darwin, whom Marx much admired, had introduced in the *On the Origin of Species*, which was published the very same year Marx wrote this preface. Perhaps the clearest example of such a selectionist explanation in Marx is provided by his appeal to class interests in his account of the struggle between bourgeoisie and proletariat.

Class Interests and Functional Explanation

Under the capitalist mode of production, the bourgeoisie must constantly strive to maintain or increase their rate of profit to stay competitive. Since, to put it simply, the rate of profit is equal to the value of a product divided by the cost of its production, perhaps the best way to increase the rate of profit is to drive down the cost of production. However, part of these costs, such as raw materials and machinery, is fixed, according to Marx. Hence, the bourgeoisie must try to decrease the variable costs of labor, either by reducing wages or by replacing workers with machinery. However, as a greater and greater proportion of the costs of production go into the fixed costs of machinery and a decreasing proportion go into the variable costs of labor, it becomes increasingly difficult for the capitalist to find ways to cut costs. The capitalist is forced continuously to reinvest his profits in machinery in order to replace even more workers with machines and the remaining workers are reduced to a subsistence wage. According to Marx's law of capitalist accumulation in *Capital* (1867), the greater the wealth invested in fixed capital, the greater the misery of the proletariat. Class conflict is the inevitable result.

Economic class interests are not necessarily consciously thought by social actors, but rather are ascribed to these actors by the social scientist observing the situation. This is what Marx meant when he referred to them as 'objective' class interests. When he said that the class interests of the bourgeoisie and the proletariat conflict, what he meant is that a member of the bourgeoisie who acts in accordance with his objective class interests will thereby act in a way that directly conflicts with the

objective interests of the working class. Marx did not mean to say that being a member of the bourgeoisie actually causes one to think in a certain way, which is different than the way in which the proletariat thinks; indeed, it is not all that important or relevant what individuals may actually be thinking. For example, to say that it is in the class interest of the bourgeoisie to reinvest their profits in new machinery is to say that those capitalists who thus reinvest their profits will remain in the bourgeoisie while those who do not act in this way will not remain competitive and will eventually drop down into the proletariat. Thus, to say that the capitalist replaces workers with machines because it is in his interest to do so is to give a functional rather than a causal explanation.

Indeed, the alternative, causal determinist interpretation of Marx makes him out to be an inconsistent thinker. If our ideas were really determined by our class membership, what would be the point of writing manifestoes? How could he have hoped to change people's ideas through rhetoric and persuasion? One might object that his writings were addressed only to the working class. However, in the *Communist Manifesto* (1848) he described – with obvious reference to people such as himself and Engels – how a portion of the bourgeoisie, "who have raised themselves to the level of comprehending theoretically the historical movement as a whole", "cuts itself adrift" and joins with the proletariat. On the determinist reading of Marx, these claims appear embarrassingly ad hoc. A functionalist reading suggests merely that people like Marx have chosen a path in life that will not make them rich.

However, as G. A. Cohen (1978) points out, orthodox Marxists tend to resist the interpretation that Marx was offering functional explanations due to their tendency to associate such explanations with the functionalist tradition in the social sciences that has its roots in thinkers such as Durkheim and Bronislaw Malinowski. According to this tradition, a society or culture forms an organic unit, every element of which serves a function that helps to sustain the whole. Marxists regard functionalist social science as resistant to social change and thus as politically conservative. However, as Cohen argues, the use of functional explanations does not necessarily commit one to this brand of functionalism.

Economic and Social Determinism after Marx

Some would say that Engels, Karl Kautsky, Antonio Labriola, Lenin, Nikolai Bukharin, and others posthumously transformed Marx's theories into deterministic ones. Whether or not this is true of all these thinkers, Marx's theories were certainly understood deterministically in their role as the official ideology of the Soviet Union. They were incorporated into an entire deterministic worldview that included all of the natural as well as the social sciences. Thus Stalin suppressed theories in genetics and quantum mechanics that entailed that nature was not completely deterministic in its operations. How could the triumph of proletarian communism be determined if even the course of nature were not? Even mathematical theories of probability and statistics were prohibited in the Stalinist Soviet Union (Huxley, 1949).

Although Ludwig Gumplowicz was not a Marxist, he affirmed both social and economic determinism in *The Outlines of Sociology* (1885). For Gumplowicz, phenomena such as customs, languages, religions, and laws arise from the interaction of social groups. A person's thoughts, feelings, and tastes are all products of her social environment. Modes of living and thinking vary not only among what he considered the three major classes – the aristocracy, the bourgeoisie, and the peasantry – but even among the different trades and professions of the bourgeoisie. However, there is no rigorous use of evidence to support any of these claims or to demonstrate the existence of causal relationships and not merely a correlation between, say, class membership and a style of thought. Economic and social determinism had metaphysical implications for Gumplowicz: he argued that if there are universal laws governing social as well as physical and psychological phenomena, then monism is true and dualism is false.

Karl Mannheim's Sociology of Knowledge

Mannheim maintained a more subtle, sophisticated version of social and economic determinism than Gumplowicz. As he explained in the English version of *Ideology and Utopia* (1936), when he spoke of the 'existential determination of knowledge', he did not have a causal, mechanistic determinism in mind. The German term is *Seinsverbundenheit des Wissens*, which carries more the sense of connection, agreement, relationship, association, bond, or tie than one of causal determination between existence and knowledge. What Mannheim meant by the expression is that thought does not develop solely through some sort of inner logic, but is influenced by extratheoretical or existential factors that play more than a peripheral role in shaping our ideas.

Mannheim's essay "The Problem of a Sociology of Knowledge" (1952) presents perhaps the clearest exposition of his views. According to Mannheim, the sociology of knowledge, which investigates the existential determination of knowledge, has its origins in Marx's theory of ideology, which is characterized by an 'unmasking turn of mind'. To unmask an idea is not to refute it or deny its truth, but to ascertain its extratheoretical function, the social and economic interests that it serves, and thereby to undermine its practical effectiveness. For Mannheim, however, a person's ideas or beliefs do not depend on their social existence in any piecemeal fashion, but only as interdependent parts of an entire superstructure or worldview. The sociology of knowledge investigates the existential determination of this entire worldview, unmasking not single ideas but entire ideologies.

Mannheim also referred to these worldviews as styles of thought, standpoints, or world postulates. These are not static systems of ideas, but rather ways of thinking that are continuously developing and changing. At any given moment of history, there may be several different styles of thought. Marxism itself is simply one theoretical standpoint that corresponds to a particular phase of history. For Mannheim, the relationship between styles of thought and social factors is not a straightforward, causal one. The economic base and the superstructure are both mental realms for Mannheim, and both parts of the same evolutionary process. The sociology of knowledge investigates this evolutionary process and the functional relationships between intellectual standpoints and

social locations. Ideas change their meaning when their social function changes. One cannot identify once and for all time any given intellectual standpoint with any particular social stratum or class, as everything is constantly changing. Thus, for example, there is no such thing as a specifically bourgeois or proletarian style of thought.

According to Mannheim, Marxists recognize only one sort of correlation between social reality and intellectual phenomena: that is, that the latter are dictated by social interests. However, Mannheim thought that to attempt to explain anything other than political and economic theories in this way would only be to expand the notion of interest in such a way that it would lose all meaning. Nevertheless, economic theories are embedded in larger intellectual universes in such a way that someone with a direct interest in an economic theory may be indirectly committed to the artistic, philosophical, and other intellectual components that are part of the same universe as that economic theory.

Like Gumplowicz, Mannheim also recognized that social groups other than just the bourgeoisie and the proletariat – such as occupational groups, sects, and generations – play a role in the existential determination of ideas. Within the proletariat, for example, there are different groups each sharing a different worldview and belonging to a different political party, reflecting different social strata within the proletariat. The sociologist of knowledge investigates the functional dependence of each intellectual standpoint on the social group that stands behind it and traces the evolution of these standpoints.

Durkheim and Social Causes

Durkheim's works have been subjected to a variety of interpretations from materialistic and deterministic to idealistic. Deterministic readings are due in large part to the ambiguity of his sociological explanations: it is often not clear whether he was offering a causal explanation, functional explanation, or both. In *The Rules of Sociological Method* (1895), he insisted that social facts are to be explained by social causes and that these social causes are to be carefully distinguished from psychological and other sorts of causes as well as from social functions. However, he did not always follow these methodological prescriptions in his empirical sociological treatises.

The Causes of the Division of Labor

For instance, Durkheim's *The Division of Labor in Society* (1893) is often read as having made the deterministic claim that increasing labor specialization results from increasing population density. However, on a closer reading, he actually said that specialization results from an increase in what he called 'social density', or the degree to which individual human beings form social relations with one another. He also seems to have suggested that population density, along with the rise of cities and means of communication, is merely a symptom or sign of social density. However, because he made population and social density each the necessary and sufficient condition for the other, his interpreters have been confused about which is supposed to be the cause and which the effect. Hence, Durkheim was taken to have advanced the social determinist thesis that rising population density causes increased social density which in turn results in the division of labor in society.

Nevertheless, Durkheim explained that specialization is not caused by population or social density but is an adaptation to it, making explicit reference to Darwin's account of how a divergence of characteristics is naturally selected in the struggle for existence. Durkheim even mentioned alternative responses to population pressure, such as emigration. Specialization takes place only when the social bonds among individuals are so strong that emigration is out of the question.

Furthermore, nothing Durkheim said rules out the possibility that specialization could be a successful strategy for coping with other sorts of conditions besides social and population density. For example, hunter-gatherer tribes living under low-density conditions have been known to specialize in the natural products available in their territory, trading with other tribes for other things that they need. Hence, population density does not appear to be even a necessary condition for specialization. At best, it may be what J. L. Mackie (1974) has called an 'inus' condition, or an insufficient but necessary part of an unnecessary but sufficient condition, or, less technically, a condition that is only necessary under a certain set of circumstances.

The Social Causes of the Categories

Later in his career, Durkheim, along with Marcel Mauss and others, advanced a sociology of knowledge according to which our basic categories of thought, including space, time, causality, substance, and genus or class, are to be explained by social causes (Durkheim, 1912; Durkheim and Mauss, 1903). For example, they claimed that a culture's units for measuring time reflect the rhythm of its social life and its system of spatial directions is patterned on the distribution of clans at tribal campsites. The very practice of classifying things in nature by a nested hierarchy of genera and species reflects an elementary social structure consisting of clans subsumed under phratries.

However, Durkheim's sociology of knowledge is ambiguous between the causal claim that the categories result from social life and the functional claim that these concepts make human social life possible. He argued that every society needs for instance some method of indicating directions in space and measuring time, in order to organize social life and make communication possible. A society also requires some way of assigning causal responsibility, since law and morality depend on it. The evidence that the Durkheimians have marshaled in support of their theory actually shows that cultures have different ways of doing these things, which amount to different ways of representing what are more or less the same basic concepts, and that these cultural representations of the categories have social causes or at least social models. However, Durkheim was not entirely clear about whether all cultures shared the same basic concepts, because although he thought that these concepts were functionally necessary to society, he also claimed that it was possible for a culture to have a different system of categories, which would imply that concepts like space, time, and causality are not really necessary for social life after all.

Despite the ambiguity of Durkheim and Mauss's sociology of knowledge, it provided part of the inspiration, along with Edward Sapir and Benjamin Lee Whorf's linguistic determinism, for the sociological determinism of such British social anthropologists as Max Gluckman, Edmund Leach, and Mary Douglas. According to this view, because the concepts or categories through which people perceive the world are the product of their cultures, the very contents of their experiences will be socially and culturally determined. These social anthropologists, however, did not draw a clear distinction between mere classificatory concepts and the more fundamental categories of thought, such as the very idea of having a classification in the first place. Gluckman, for instance, included concepts such as shapes and colors among the categories that determine perceptual reality for us:

> From infancy, every individual is moulded by the culture of the society into which it is born. All human beings see, but we know, for example, that how they see shapes and colours is to some extent determined by this process of moulding. More than this, their ability to describe their perceptions depends on the categories contained in their respective languages.
> (Gluckman, 1949–1950: pp. 73–74)

This claim rests on dubious assumptions about the psychology of perception and is not supported by more recent anthropological research.

More Recent Determinism

The Strong Programme in the Sociology of Scientific Knowledge

In the 1970s, a group calling themselves the "Strong Programme in the Sociology of Scientific Knowledge" defended a determinist approach with roots in Durkheim and Marx. David Bloor (1976/1991), a principal spokesperson for the group, laid out four tenets for this program. They are (1) that explanations of belief and knowledge should be causal; (2) that sociology should be impartial with respect to the truth or falsity, or the rationality or irrationality, of the belief to be explained; (3) that explanation should be symmetrical, in the sense that the same kind of causes would explain both true and false beliefs; and (4) that the sociology of knowledge should be reflexive, in the sense that the same principles of explanation should apply to sociology itself. The Strong Programme regarded accounts of the beliefs and practices of scientists in terms of their social interests as satisfying the demand for causal explanation, although, as mentioned earlier, interest accounts are functional rather than causal explanations. Their concept of social interests included not only economic or class interests but political, religious, professional, and other interests as well. Thus, ironically, while criticizing Mannheim for failing to include the natural sciences within the explanatory scope of the sociology of knowledge, they also neglected his warning about expanding the notion of social interests to such an extent as to empty it of meaning.

Philosophers and other social scientists sharply criticized the Strong Programme. Philosophers insisted that scientists held the views that they did on the basis of reason and evidence. They furthermore argued that the tenets of impartiality and symmetry prevent the Strong Programme from arguing for their position on epistemic grounds: that is, these tenets prohibit them from saying that their beliefs are any more well founded than the alternatives. Among the social scientists who criticized them, some felt that in plumping for causal explanations, the Strong Programme was trying too hard to model the social on the natural sciences. Others criticized them for taking the existence of social structure and social classes as given and not recognizing that these are constantly changing social phenomena brought about through complex human interaction. Like Mannheim, these critics recognize that social structure and class exist in and through the beliefs and practices of the members of society and are to be explained in the same way as any other beliefs and practices. To explain a person's scientific beliefs in terms of social interests is thus in effect to privilege one sort of belief over another.

Feminist Standpoint Epistemology

More recently, a type of feminist social theory called 'standpoint epistemology' reflects the influence of Marx's social and economic determinism. Feminist standpoint epistemology is characterized by two related claims: the situated knowledge thesis and the privileged location thesis.

According to the situated knowledge thesis, what one knows or experiences reflects one's social, cultural, and historical location. As Alison Wylie puts it:

> What we experience and understand, the differential strengths and liabilities we develop as epistemic agents, are systematically shaped by our location in hierarchically structured systems of power relations, by the material conditions of our lives, by the relations of production and reproduction that shape our social interactions, and by the conceptual resources we rely on to interpret and represent these relations.
> (Wylie, 2007: p. 574)

The privileged location thesis says that some social locations are privileged in the sense of being more likely to produce true or reliable knowledge. According to Sandra Harding, the privileged location thesis has its roots in Hegel and Marx (1986: p. 26).

For Hegel, the slave enjoyed this privileged position relative to his or her master; for Marx, the proletariat occupies an epistemologically privileged location relative to the bourgeoisie. Feminist standpoint theory says that this privileged location belongs to women, especially feminist women.

Marx and Engels, recall, claimed that they were not putting forth a new social philosophy, but establishing a new science, based on observation, which revealed the direction in which history was determined. They also held that the proletariat was in a privileged position to observe the social conditions that determined this direction. Marx and Engels had assumed something akin to Francis Bacon's model of the methods of science, according to which scientific objectivity consisted in individual scientists freeing their minds of certain biases, or 'idols' as he called them, in order to observe and classify the

facts that could serve as the basis for general laws. For Marx and Engels, those who are in positions of power have more biases that interfere with gathering objective knowledge from experience than those who are socially and economically marginalized. This was because the bourgeoisie saw things only through the veil of an ideology that made possible their higher social and economic status. Members of the proletariat were encumbered by fewer such ideological biases than the bourgeoisie and thus enjoyed an epistemologically privileged location.

Similarly, for the feminist, men's minds may be clouded by the very ideology through which they maintain their power over women. However, the feminist standpoint theorist does not think that someone automatically enjoys an epistemologically privileged location simply in virtue of being a woman. Nor do all Marxists think that all members of the proletariat automatically enjoy an epistemologically privileged location. It is not always obvious to someone that he or she is being oppressed and it is possible for someone to succumb to the very ideology that helps to maintain them in an inferior social position. For the feminist standpoint theorist, an epistemologically privileged location is achieved as the result of an intellectual struggle. A woman may come to realize her oppression only through discussion with others and an analysis of her personal situation.

However, lacking a clear way to define social location, the privileged location thesis is difficult to test. Marx had assumed a simple linear hierarchy of power relationships. In contemporary society, one's social location is a function of gender, race, ethnicity, religion, sexual orientation, wealth, level of education, marital status, and other factors, not all of which vary together. Feminist standpoint theorists may reject the simple linear hierarchy of the Marxists, but the idea of a hierarchy of power relations persists in the notion of a center of power surrounded by a periphery of oppressed groups. But how does one measure and thus compare the distances from the center of these oppressed groups? And how can one test the claim that members of these groups see things more clearly? Even members of marginalized groups have their prejudices. Nevertheless, one could argue that marginalized social locations may be epistemologically privileged just to the extent they allow the prejudices of the dominant group to become more visible, regardless of the prejudices of the marginalized group. That is, the victims of oppression are in a better position to observe its effects, and perhaps to question the assumptions of the dominant ideology.

It is ironic that a social theorist like Marx would hold the Baconian notion that epistemological objectivity depends on the elimination of biases, since it assumes that the knowledge producer is the individual, rather than the community. From a social point of view, objectivity depends on critical interaction among of group of people, rather than each individual cleansing her mind of prejudice. The more that different people with various experiences and points of view contribute to this critical discussion, the more objective will be the result. Thus, regardless whether the privileged location thesis can be demonstrated, the situated knowledge thesis still implies that members of marginalized and oppressed groups should be given a hearing in the collective production of knowledge for epistemological and not just ethical reasons.

See also: Causation: Physical, Mental, and Social; Causes and Laws: Philosophical Aspects; Class: Social; Darwinism; Division of Labor: History of the Concept; Durkheim, Emile (1858–1917); Field Theory; Functional Explanation: Philosophical Aspects; Functionalism in Anthropology; Historical Explanation, Theories of: Philosophical Aspects; Inequality, Social; Interest: History of the Concept; Knowledge, Sociology of; Labor, Division of; Mannheim, Karl (1893–1947); Marx, Karl (1818–83); Marxism in Contemporary Sociology; Marxism–Leninism: The Ideology of Twentieth-Century Communism; Mauss, Marcel (1872–1950); Objectivity, History of; Objectivity: Philosophical Aspects; Scientific Knowledge, Sociology of; Situated Knowledge, Feminist and Science and Technology Studies Perspectives; Social Constructivism; Sociological Theory; Sociology, History of; Standpoint Theory, in Science; Strong Program, Sociology of; Whorf, Benjamin Lee (1897–1941).

Bibliography

Bloor, D., 1976/1991. Knowledge and Social Imagery. University of Chicago Press, Chicago, IL.

Bottomore, T.B., Rubel, M. (Eds.), 1956. Karl Marx: Selected Writings in Sociology and Social Philosophy. McGraw-Hill, New York.

Cohen, G.A., 1978. Karl Marx's Theory of History: A Defense. Princeton University Press, Princeton, NJ.

Durkheim, E., 1893. De la division du travail social. Alcan, Paris [1984. The Division of Labor in Society. The Free Press, New York].

Durkheim, E., 1895. Les Règles de la méthode sociologique. Alcan, Paris [1982. The Rules of Sociological Method. The Free Press, New York].

Durkheim, E., 1912. Les formes élémentaires de la vie religieuse. Alcan, Paris [1995. The Elementary Forms of Religious Life. The Free Press, New York].

Durkheim, E., Mauss, M., 1903. De quelques formes primitives de classification. L'Année Sociologique 6, 1–72 [1963. Primitive Classification. University of Chicago Press, Chicago].

Gluckman, M., 1949–1950. Social beliefs and individual thinking in primitive society. Memoirs and Proceedings of the Manchester Literary and Philosophical Society 91, 73–98.

Gumplowicz, Ludwig, 1885. Grudriss der Soziologie. Manz'sche k. k. Hof-Verlags- und Universitäts-Buchhandlung, Vienna [1899. The Outlines of Sociology. American Academy of Political and Social Science, Philadelphia].

Harding, S.G., 1986. The Science Question in Feminism. Cornell University Press, Ithaca, NY.

Huxley, J., 1949. Heredity, East and West. Henry Schuman, New York.

Mackie, J.L., 1974. The Cement of the Universe: A Study of Causation. Oxford University Press, Oxford, UK.

Mannheim, K., 1936. Ideology and Utopia. Routledge, London, UK.

Mannheim, K., 1952. The problem of a sociology of knowledge. In: Essays on the Sociology of Knowledge. Oxford University Press, New York.

Marx, K., 1848/1955. The Communist Manifesto. Harlan Davidson Inc., Arlington Heights, IL.

Marx, K., 1859. Preface to a Contribution to the Critique of Political Economy. http://www.marxists.org/archive/marx/works/1859/critique-pol-economy/preface.htm (accessed 25.06.13.).

Marx, K., 1867/1906. Capital. Modern Library, New York.

Marx, K., 1845. Theses on Feuerbach, appendix. In: Engels, F. (Ed.), 1888. Ludwig Feuerbach and the End of German Classical Philosophy. http://www.marxists.org/archive/marx/works/1845/theses/theses.htm (accessed 24.06.13).

Wylie, A., 2007. The feminism question in science: what does it mean to 'do social science as a feminist'? In: Hesse-Biber, S.N. (Ed.), Handbook of Feminist Research. Sage, Thousand Oaks, CA, pp. 567–577.

Deterrence

Leon V Sigal, Social Science Research Council, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although the idea of threatening war to prevent war is as old as the balance of power, deterrence is not just a new word for that old idea but a new idea spawned by the nuclear age. Deterrence means threatening with punitive retaliation to prevent a foe from attacking. It differs fundamentally from defense, which means threatening to fight back if attacked and deny an attacker its objectives. To deter nuclear attack, it is deemed necessary to threaten nuclear retaliation. The prohibitive cost of a nuclear war keeps either side from risking it. Yet that rational calculus overlooks the irrationality of retaliating for a nuclear first strike with a nuclear second strike against a well-armed attacker. Moreover, military plans in practice do not resemble deterrence in theory. The enormous advantage of striking first has prompted the United States and the Soviet Union, now Russia, to keep forces on hair-trigger alert, raising fear of loss of nuclear control in a crisis. As a consequence, while some saw mutual deterrence as a source of stability, others saw it as extremely unstable, leading them to renewed appreciation of the security dilemma, where attempts to enhance or demonstrate military were self-defeating, and safety lay in cooperation between enemies.

Deterrence means threatening with punitive retaliation to prevent a foe from attacking. It differs fundamentally from defense, which means threatening to fight back if attacked and deny an attacker its objectives. The heart of the distinction is whether it would make sense to carry out the threat.

Although the idea of threatening war to prevent war is as old as the balance of power, deterrence is not just a new word for that old idea. The idea of deterrence originated with the atomic bomb. It was first formulated in 1946 by Bernard Brodie, who tried to capture what was revolutionary about the advent of the nuclear era. First, a nuclear warhead was not a weapon, in Brodie's view, but a terrorist device, whose blast and radiation effects made it inherently indiscriminate. Second, the atomic bomb brought home the fact of mutual vulnerability in a way that aerial bombardment had not. Before the nuclear era, a state's armed forces had to be defeated before the state and its centers of population could be held hostage or destroyed at will. Once nuclear warheads were mated to missiles in the late 1950s, the state and its populace were defenseless. The balance of terror differed in fundamental respects from the prenuclear balance of power, according to Brodie. "The first and most vital step," he wrote, "for the age of atomic bombs is to take measures to guarantee ourselves in case of attack the means of retaliation. Thus far the chief purpose of a military establishment has been to win wars. From now on its chief purpose must be to avert them." These prescient insights became a matter of considerable scholarly discourse. Regrettably they had little influence on nuclear policy.

Deterrence in Theory

The logic of deterrence, as Brodie and others recognized, is inherently paradoxical. To deter nuclear attack, it is deemed necessary to threaten nuclear retaliation. Inasmuch as the cost of a nuclear war is prohibitive, neither side would risk it. Yet that rational calculus conveniently overlooks the irrationality of retaliating for a nuclear first strike with a nuclear second strike. Retaliation makes no sense unless the second strike could eliminate the other side's nuclear forces and prevent it from launching another strike of its own. That deadly logic led Brodie to the paradoxical conclusion that what made deterrence work was the possibility that it might fail, making it prudent for neither side to tempt fate. The same logic led other strategic thinkers to conclude that for deterrence to work, it could not be mere bluff. The threatener had to be willing to carry out the threat. For that to happen, nuclear retaliation had to be automatic or mad. Herman Kahn (1960) captured the sense of that line of reasoning in a phrase, the 'rationality of irrationality.' Yet that line of reasoning only posed another paradox: if the deterrer was hotheaded, why would the attacker be prudent?

Even if the threat of retaliation ruled out nuclear war, that seemed to leave the world safe for war waged by other than nuclear means. That was a source of gnawing anxiety for the USA, which was committed by a treaty to deter any attack by the USSR on its allies in Western Europe. It did not have the capability to do that by conventional military means alone, at least in the 1950s, so it chose to rely on the threat to initiate nuclear war. That threat was potentially suicidal.

Some strategists insisted that 'extended deterrence,' as that threat came to be called, would work if the USA had nuclear superiority. Thomas Schelling (1966) questioned whether superiority assured that the weaker side must yield to the stronger. It was impossible, he insisted, to raise the risk of a nuclear war for the other side without raising it for oneself. He spoke, instead, of a 'competition in risk-taking.' Extended deterrence still might work if the Soviet leaders could not be sure the USA would shrink from first use of nuclear weapons in the event of war. Schelling (1960) called it 'the threat that leaves something to chance.' Such a cosmic bluff prompted a preoccupation with 'credibility' and a willingness to run seemingly irrational risks to shore up an incredible threat. McGeorge Bundy (1983) came up with a less demanding alternative, 'existential deterrence.' So long as a state possessed nuclear arms, Bundy argued, it had a latent ability to use them,

even if it did not explicitly threaten to do so. The mere existence of these arms and the incalculable costs of a nuclear war exerted a cautionary effect on potential foes.

Other strategic thinkers took a more forceful tack. British strategist Basil Liddell Hart (1946) first formulated the idea of graduated deterrence, carefully calibrated to threaten greater costs to an attacker without generating all-out war. Still others debated whether military forces or civilians were the most appropriate targets for nuclear attack. This line of thought culminated in the baroque escalation ladder of Herman Kahn (1960) and the contention of some of his followers that the USA could deter any aggression if it somehow had 'escalation dominance,' or military superiority at every level of violence, short of an all-out nuclear war. Yet the idea of waging limited nuclear war and controlling escalation raised the question of how a nuclear war could stop, short of a dead end.

Deterrence in Practice

Deterrence, in short, did not work very well in theory. What happened in practice was in many ways worse. Deterrence did not govern the production, deployment, or military plans for nuclear arms (Ball, 1980). The number of the Soviet and the US warheads vastly exceeded what was needed for deterrence. The warheads themselves were initially deployed in ways that made them vulnerable to attack. Dispersing them on land and at sea reduced, though did not eliminate their vulnerability, and made command-and-control all the more precarious. Because of the chance that the cosmic bluff could be called, each side drew up detailed war plans to destroy the other. Those war plans fully reflect the deadly logic of deterrence: in the event that war breaks out, nuclear arms confer enormous, some say decisive, advantage on the side that strikes first. The most urgent targets were the other side's nuclear forces and especially their command-and-control. Concern about a disarming first strike led the USA to keep some of its bombers airborne at all times and to put others on alert, ready to take off at a moment's notice. It did the same with its missiles. This hair-trigger posture, and the delegation of authority that it necessitated, raised the risk of loss of nuclear control in a crisis. Schelling (1960) called this predicament 'the reciprocal fear of surprise attack.' As first one side and then the other began mobilizing forces, the very steps taken to deter a nuclear war might provoke one. Far from exerting a cautionary effect on preparations for war, that fear led both sides to a massive buildup of arms, conventional as well as nuclear, which did little to calm the fear.

That had important implications for theorizing about deterrence. Although some theorists saw mutual deterrence as a source of stability in international politics, others saw it as potentially unstable in the extreme. The first group emphasized the need to demonstrate the capability and will to wage war, lest a potential aggressor doubt a state's strength or resolve. By this way of thinking, conciliation was dangerous because it might be mistaken for weakness or irresoluteness. A second group of theorists were led to a renewed appreciation of the security dilemma, an idea dating back to Thucydides, who noted that attempts by one side to enhance or demonstrate its military might prove self-defeating. By alarming its rival and leading it to respond in kind, these measures would leave both sides less secure. Such a vicious circle could generate an arms race, or worse, trigger preemptive war. These theorists saw the need to couple coercion with conciliation, lest the interaction spiral out of control. To some scholars, the difference between the deterrence and spiral logic depended on each side's perception of the other's intentions.

Where perception matters, so does the danger of misperception. The fear that war is imminent is sometimes irrational. Once it takes hold, it is not amenable to a cool calculation of costs. The possibility of preemptive war calls into question two major premises of most theorizing about nuclear deterrence, that statesmen will behave rationally in the heat of the moment and that they can control the operations of their armed forces. The questioning seems all the more pertinent with the proliferation of nuclear arms to states like Pakistan and North Korea or the risk that terrorists could get their hands on a nuclear weapon. This led to critiques of deterrence theory on the grounds that it was psychologically naive and operationally uninformed. Robert Jervis (1976), John Steinbruner (1974), and Scott Sagan (1993) are leading exemplars of these approaches.

One side's strategy depended on the other's. So did its security. Their interdependence was the central feature of *game theory*. Whether the game of chicken or the prisoner's dilemma was the appropriate analogy, many scholars who formulated the logic of deterrence drew heavily on game theory for their insights.

One consequence is that deterrence is largely deductive and only weakly grounded empirically. Another reason for its weak evidentiary base is that nuclear history everywhere remains cloaked in secrecy. Without knowledge of the details of nuclear planning and operations, it is difficult to determine how much nuclear strategy is informed by deterrence theory, if at all. A more fundamental reason why the evidence for deterrence is less than compelling is the epistemological difficulty of proving why something did not happen in order to demonstrate that deterrence worked. The obvious counter to the contention that nuclear deterrence kept the USSR out of Western Europe is that it never intended to invade in the first place.

Because the side that is losing could still destroy the side that is winning, safety in the nuclear era lay in cooperation between enemies. That interdependence made nonsense of traditional strategic thought. It also called into question the very idea of strategy as a rational relationship between means and ends. War, in theory, if not always in practice, "is controlled by its political object," Clausewitz had reasoned. Consequently, "the value of this object must determine the sacrifices to be made for it in magnitude and also duration." But what objective could possibly sustain the full measure of sacrifice in a nuclear war? As Lawrence Freedman (1981) concludes, "The position we have reached is one where stability depends on something that is more the antithesis of strategy than its apotheosis – on threats that things will get out of hand, that we might act irrationally, that possibly through inadvertence we could set in motion a process that in its development and conclusion would be beyond human control and comprehension." Nuclear strategy, understood this way, was a contradiction in terms.

See also: Genealogy in Anthropology; Wars among Nation-States: Patterns and Causes.

Bibliography

Ball, D., 1980. Politics and Force Levels. University of California Press, Berkeley, CA.
Brodie, B., 1946. The Absolute Weapon. Harcourt, Brace, New York.
Bundy, M., 1983. The bishops and the bomb. New York Review of Books, vol. 10, p. 30.
Freedman, L., 1981. The Evolution of Nuclear Strategy. St. Martin's Press, New York.
Jervis, R., 1976. Perception and Misperception in International Politics. Princeton University Press, Princeton, NJ.
Kahn, H., 1960. On Thermonuclear War. Princeton University Press, Princeton, NJ.
Liddell Hart, B.H., 1946. The Revolution in Warfare. Faber and Faber Ltd, London.
Sagan, S.D., 1993. The Limits of Safety. Princeton University Press, Princeton, NJ.
Schelling, T., 1960. The Strategy of Conflict. Harvard University Press, Cambridge, MA.
Schelling, T.C., 1966. Arms and Influence. Yale University Press, New Haven, CT.
Steinbruner, J.D., 1974. The Cybernetic Theory of Decision. Princeton University Press, Princeton, NJ.

Deterrence: Sanction Perceptions

Robert Apel, Rutgers University, Newark, NJ, USA
Daniel S Nagin, Carnegie Mellon University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article conducts a survey of theory and research concerning the determinants of an individual's perceptions of the risk of formal sanctions as a consequence of criminal behavior. The specific questions considered are as follows: (1) How accurate is people's knowledge about criminal sanctions? (2) What is the impact of risk perceptions on the likelihood of criminal behavior? (3) What is the impact of criminal experience on the modification of risk perceptions?

Overview

Decisions about participation in criminal behavior involve forming expectations about the future. These expectations concern the potential rewards as well as the potential harms of criminal actions. The potential rewards associated with crime are largely pecuniary (i.e., the 'loot'), but they may also entail reputational rewards (e.g., peer approval) and psychic rewards (e.g., thrill). The potential harms of crime encompass bodily harm (e.g., injury if a victim resists), medical harm (e.g., disease from sharing drug needles), and psychic harm (e.g., pangs of conscience). Most importantly for present purposes, the potential harms also include legal harms, which involve restrictions on personal freedom, such as a fine, an arrest, or confinement in a jail or prison. Moreover, because the legal consequences of crime are uncertain, would-be offenders must form assessments of the probability and magnitude of criminal sanctions. Sanction perceptions refer to a would-be offender's estimate of the likelihood of being detected, arrested, convicted, and incarcerated upon committing an offense (perceived sanction certainty), as well as the length of the custodial sentence or the strictness of the supervisory conditions imposed upon conviction (perceived sanction severity).

Modern criminological interest in the study of sanction perceptions arose in the 1970s and 1980s as researchers sought to more clearly delineate the link between criminal punishments and individual criminal behavior. From this interest arose a tradition of research in perceptual deterrence. Scholars observed that criminal deterrence is fundamentally a process of information transmission. It is intended to discourage law violation by communicating to potential law violators the punishments that might arise when they violate legal proscriptions. Indeed, Nagin (1998: p. 18) observes that "behavior is immune to policy manipulation" to the degree that there is no link between policy and perceptions. Three distinct research strategies can be discerned in the perceptual deterrence tradition, and these are the subjects of this article (for more in-depth review, see Apel, 2013). These include the accuracy of sanction risk perceptions, the deterrent effect of risk perceptions on criminal behavior, and the experiential effect of criminal behavior on risk perceptions. Each will be reviewed in more detail below.

Accuracy of Sanction Risk Perceptions

One major perceptual deterrence research tradition entails estimating the degree of correspondence between area-level measures (usually the county) of criminal punishment and individual perceptions of sanction risk. These are known as calibration studies. In order for a sanction policy to influence behavior, individual perceptions of the certainty and severity of sanctions must have some grounding in reality. Consequently, the mean perceived sanction from a sample of respondents should be closely calibrated to the objective sanction of interest.

In 1967, the California Assembly commissioned a survey to determine how knowledgeable different groups of respondents were of the prevailing statutory penalties for criminal behavior (Assembly Committee on Criminal Procedure, 1968). The sample included over 1500 subjects and comprised noninstitutionalized civilians, college students, high school youth (stratified by low- and high-delinquency schools), institutionalized youth, and institutionalized adults. Of six offenses examined – assault with a deadly weapon, second degree burglary, first degree robbery, forcible rape, joy riding, fraudulent check writing – the percentage of respondents who identified the correct maximum penalty was 26% (these are manual calculations made by the authors based on information tabulated in the original report). The worst calibration was attained by the noninstitutionalized civilians and low-delinquency high school youth (22 and 23%, respectively), while the best calibration was attained by the institutionalized adults (62%). Among those who misjudged the maximum sentence, the predominant tendency was to underestimate the penalty.

The committee also inquired about public knowledge concerning legislative activity for three offenses whose penalties had recently increased (rape, robbery, and burglary) and two offenses whose penalties had not changed (marijuana possession and driving under the influence). The percentage of respondents who answered correctly about whether or not the penalties had recently changed was 28%. Again, the most knowledgeable subjects were the institutionalized adults, among whom 52% responded correctly. Among respondents who misjudged the legislative changes, the tendency was to report that the severity of the penalty had decreased or else

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45041-5

to understate the magnitude of the penalty increase. The committee's conclusion concerning recent legislative changes was blunt: "While the Legislature had supposedly responded to public appeal and increased the penalties for crime of violence to victims, this was not known by the public" (Assembly Committee on Criminal Procedure, 1968: p. 14).

Kleck et al. (2005) conducted a telephone survey of about 1500 adults residing in 54 large urban counties in 1998, inquiring about respondents' perceptions of criminal punishment for four offenses in their county (homicide, robbery, burglary, and aggravated assault). The mean perceived arrest rates for homicide and assault underestimated the objective arrest rates, while the perceived arrest rates for robbery and burglary overestimated the objective arrest rates. Perceived incarceration probabilities (conditional on conviction), on the other hand, were consistently underestimated and considerably so.

Ross (1973) reported on the findings of a public opinion survey conducted before and after passage of the British Road Safety Act (BRSA) of 1967. The BRSA included a number of provisions: a fixed blood alcohol level was adopted as the legal definition of impairment, a field breath test followed by laboratory analysis of the motorist's blood or urine was to provide the evidence of impairment, and motorists were required to submit to a breath test when an officer had reasonable cause to suspect drunkenness or else face the same penalties as if they had failed the test (e.g., loss of driver's license and arrest). The British government also embarked on a large-scale campaign to inform the public about the provisions of the BRSA. In a public opinion survey intended to gauge the level of public awareness of the BRSA, the percentage of respondents who were aware of the blood alcohol limit increased from 22 to 39%. Furthermore, 99% of respondents knew that a breath test was the basis for prosecution and 95% knew that refusal of a breath test could lead to arrest.

MacCoun et al. (2009) pooled data from the National Survey on Drug Use and Health, limiting their attention to about 135 000 adults 18 years or older. Respondents were asked to report whether, to their knowledge, the maximum penalty in their state for first-time possession of an ounce or less of marijuana was a fine, probation, community service, possible prison sentence, or mandatory prison sentence. The authors then compiled information on each state's laws, classifying states by the exact statutory penalties assessed (e.g., fine amount, jail time, diversionary, and expungement provisions). Residents of states that decriminalized marijuana possession, compared to states in which marijuana possession was still criminalized, did indeed report significantly more lenient sanctions for violation, but the actual differences were surprisingly modest in substantive terms. For example, among residents of states with a decriminalization policy, 44% still reported that the maximum penalty included the possibility of a jail or prison sentence compared to 49% among residents of states without a decriminalization policy.

To summarize calibration studies, scholars have shown that survey respondents have a general tendency to overestimate arrest probabilities (sanction certainty) but to underestimate maximum penalties (sanction severity). Yet when respondents are asked about the statutory applicability of specific criminal penalties, there tends to be much stronger correspondence between perceptions and reality. Therefore, while the average citizen does a reasonable job of knowing what criminal penalties are statutorily allowed, she does a poor job of estimating the probability and magnitude of the penalties.

If people are only vaguely aware of criminal punishments, then the deterrence rationale of punishment is seriously undermined. Yet there is still much to be learned about how individuals calibrate their risk perceptions with respect to objective punishment likelihoods. For example, most people are not 'in the market' for the kinds of serious criminal offenses that have been the subject of many studies. In the language of deterrence, most survey respondents are not 'on the margin' for serious law-violating behavior – they are committed law abiders. Population-based studies are likely to be composed largely of committed law abiders, especially when serious crimes are under consideration, and consequently, there is no a priori reason to expect that their perceptions will be highly accurate. It is noteworthy that studies of criminal statutes concerning drunk driving and marijuana possession tend to provide more consistently favorable evidence for perceptual deterrence principles. One plausible reason for this is that these studies are concerned with offenses (alcohol and marijuana use) and populations (young people) that are actually 'deterrable' in the sense that many individuals are on the margin for these kinds of behaviors.

Furthermore, experience with drunk-driving legislation or enforcement highlights two perceptual deterrence weaknesses in practice. For one, many sanction changes concern 'downstream penalties' that increase sanction severity rather than sanction certainty. This distinction is relevant because perceptual deterrence research uniformly suggests that would-be offenders are more strongly deterred by the certainty of punishment than by the severity of punishment (see Nagin, 1998) not to mention that downstream penalties can be heavily discounted because of the inevitable delay in their implementation. In addition, the failure of criminal justice officials to carry through on punishment threats can undermine perceptual deterrence. Drunk-driving evaluations demonstrate that while downstream penalties are often harsher, as intended, 'upstream enforcement' is largely unchanged (e.g., Ross, 1973). Specifically, changes in drunk-driving policies do not materially alter police behavior and thus result in little change in the actual risk of apprehension. Even when citizens are knowledgeable about criminal penalties, therefore, the deterrence potential of sanctions can be significantly eroded when experience indicates that crime detection remains so improbable as to be ignorable, even when the sanctions are quite severe.

Deterrent Effect of Risk Perceptions on Criminal Behavior

Research on the deterrent effects of perceptions on unlawful behavior has proceeded along three distinctive methodological lines: cross-sectional surveys, panel studies, and vignette research. Cross-sectional surveys inquire about individuals' current perceptions of sanction risks (certainty and severity) and either their behavior within some reference period prior to the interview or alternatively their intentions to commit crime in the future. Panel studies are designed to untangle

temporal priority for the perceptions–behavior link that vexes cross-sectional research by relating perceptions in one period to actual behavior reported in a later period. While cross-sectional surveys consistently show that risk perceptions are inversely correlated with offending behavior, panel studies tend to reveal correlations that are substantially diminished (Paternoster, 1987).

Vignette research provides respondents with a detailed hypothetical crime scenario and then asks them about their perceptions of the certainty and severity of punishment for the crime as well as their own behavior if they found themselves in the same situation. A unique feature of this design is that situational characteristics can be experimentally manipulated in order to study how subjects respond to a variety of incentives, disincentives, and opportunity structures. For example, Nagin and Pogarsky (2001) presented to university students a scenario describing an incident of drunk driving in which potential punishments were experimentally manipulated. Respondents given a scenario with a longer length of license suspension (higher severity) reported a significantly lower likelihood of driving drunk, although this effect appeared to be diminished among individuals who were more present oriented or impulsive (i.e., those with a higher 'discount rate'). The authors also inquired about the subjects' own estimated likelihood of being apprehended and convicted for drunk driving (a measure of certainty) under the conditions described in the scenario and found that perceived certainty was a robust predictor of intentions to drive drunk.

Research on the perceptual deterrence has produced two other sets of findings that are important to understanding the sources of deterrence. The first concerns the comparative deterrent effects of the certainty and severity of punishment. The perceptual deterrence literature finds more consistent evidence of the certainty of punishment as a crime deterrent relative to the severity of punishment. The second concerns the role of informal sanctions in the deterrence process. Zimring and Hawkins (1973: p. 174) observe that formal punishment may best deter when it sets off informal sanctions: "Official actions can set off societal reactions that may provide potential offenders with more reason to avoid conviction than the officially imposed unpleasantness of punishment." This research has consistently found that individuals who report higher stakes in conventionality are more strongly deterred by their perceived risk of punishment for law breaking.

A salient example of research supporting the 'certainty principle' as well as untangling the link between formal and informal sanctions concerns tax evasion. In the United States, civil enforcement actions by tax authorities are a private matter unless the taxpayer appeals the action. Because tax authorities are scrupulous about maintaining the confidentially of tax return information, for civil enforcement actions, noncompliers are gambling only with their money and not their personal reputations. In Klepper and Nagin (1989), a sample of generally middle-class adults were posed a series of tax noncompliance scenarios. The scenarios laid out the essential features of a tax report – income from different sources, number of exemptions, and various deductions. They then experimentally varied the amount and type of noncompliance (e.g., overstating charitable deductions or understating business income) across tax return line items and found that a majority of respondents reported a nonzero probability of taking advantage of the noncompliance opportunity described in the scenario. Plainly, the respondents were generally willing to consider tax noncompliance when only their money was at risk. They also seemed to be calculating; the attractiveness of the tax noncompliance gamble was inversely related to the perceived risk of civil enforcement.

The one exception to the rule of confidentially of enforcement interventions is criminal prosecution. As with all criminal cases, criminal prosecutions for tax evasion are a matter of public record. Here, Klepper and Nagin found evidence of a different decision calculus – seemingly all that was necessary to deter evasion was the perception of a nonzero chance of criminal prosecution. Stated differently, if the evasion gamble also involved putting reputation and community standing at risk, the middle-class respondents were seemingly unwilling to consider taking the noncompliance gamble.

This finding helps explain why the certainty of punishment may be a greater deterrent than severity. If the social and economic costs of punishment are strictly proportional to the punishment received; for example, if the cost to the individual of a 2-year prison term is twice that of a 1-year sentence, certainty and severity will equally affect expected cost. This is because expected cost is simply the product of certainty, P, and severity, S. The value of the product, $P \times S$, is equally affected by proportional changes in P or S. For example, the impact on expected cost of a 50% increase in P is the same as a 50% increase in S. However, the Klepper and Nagin study suggests that people do not perceive that costs are proportional to potential punishment. Instead, it seems that they perceive there is fixed cost associated with merely being convicted or even apprehended if it is public record.

While Klepper and Nagin (1989) did not pin down the specific sources of these costs, other research on the impact of a criminal record on access to the legal labor market suggests a very real basis for the fear of stigmatization. For example, a prison record depresses the probability of employment by about 15%. More recent studies reinforce these earlier findings. Western (2002) estimates the wage reduction effect of incarceration to be about 16%. He also finds that incarceration deflects individuals onto a flatter wage trajectory in which wage growth is slowed by 31% relative to comparably high-risk men who were not incarcerated. Pager (2003) reports that employers advertising entry-level job openings were less than half as likely to call back applicants who reported a criminal history (a felony cocaine trafficking conviction with 18-months prison time). Her conclusion is that "criminal records close doors in employment situations" (p. 956).

Experiential Effect of Criminal Behavior on Risk Perceptions

An early conclusion of perceptual deterrence researchers was that people's own experiences with crime and punishment were more salient determinants of their risk perceptions than external sources of information about sanction risk. Perceptual deterrence research also found that individuals who have experience as criminal offenders tended to have substantially lower risk perceptions, and therefore, more accurate risk

perceptions, compared to individuals who lack such experience. This begs the question of whether offenders' lower risk perceptions might be a consequence of their criminal behavior as well as a cause. Panel studies of the deterrent effect of perceptions were conducted to speak to this question. The first wave of such studies indicated that there is a pronounced experiential effect of delinquent or criminal behavior on risk perceptions (Paternoster et al., 1982). Furthermore, among active offenders, the experience of being sanctioned (and sanctioned more severely) itself contributes to an increase in risk perceptions (Apospori and Alpert, 1993). There are also compelling reasons to believe that individuals change their risk perceptions, in part, on the basis of the crime and punishment experiences of their friends and family members. Stafford and Warr (1993) refer to such effects as indirect or vicarious experiences.

Another important determinant of risk perceptions in addition to personal and vicarious punishment experience, according to Stafford and Warr (1993), is personal and vicarious punishment avoidance. There is evidence from a number of studies that, other things equal, a higher level of unsanctioned offending lowers one's perceptions of formal sanction risk. For example, in a study by Horney and Marshall (1992), offenders with higher 'arrest ratios' (i.e., more reported arrests per reported offenses) had higher perceptions of the risk of detection. Stated differently, offenders with more unsanctioned offenses – that is, more successful criminal careers, with success defined as avoidance of arrest – had lower risk perceptions.

These findings on the experiential effects of punishment and behavior on risk perceptions have spawned a second-generation research tradition concerned with the development and testing of formal models of change in risk perceptions as a consequence of personal and vicarious experiences with, and avoidances of, arrest and punishment (e.g., Anwar and Loughran, 2011; Matsueda et al., 2006). Many of these studies appeal to a Bayesian model of learning. A Bayesian model of risk perceptions and criminal behavior begins with an individual's initial assessment of the likelihood of apprehension for criminal conduct. This is known as the 'prior probability' of the risk of arrest (among other sanctions). The perceptual deterrence literature strongly suggests that the prior probability estimates of individuals without offending experience systematically and often substantially overstate the true risk of apprehension. Over time, the individual will accumulate personal and/or vicarious experiences as a successful or unsuccessful criminal offender. Note the parallel with Stafford and Warr's (1993) conceptualization of personal and vicarious punishment experience and punishment avoidance. The expectation is that the individual will then 'update' his or her assessment of the risk of apprehension based on these experiences. The resulting 'posterior probability' of the risk of arrest is a weighted sum of the individual's prior probability and the new information.

Anwar and Loughran (2011) have significantly advanced empirical testing of the Bayesian updating model by looking for more fine-grained predictions of it. First, they observe that the effect of arrest within a reference window should depend on the number of crimes committed during that period. For example, being arrested once should matter more for an offender who committed only one crime as opposed to an offender who committed 10 crimes. In the case of the former, the 'experienced arrest certainty' is 100%, whereas in the case of the latter, it is only 10%. Second, Anwar and Loughran observe that the effect of arrest should diminish as individuals gain more offending experience. In other words, since posterior risk perceptions are a function of prior risk perceptions and new information in the form of their arrest ratio (what they refer to as a 'signal'), experienced offenders should place more weight on their prior risk perceptions in the sense that they should update less in response to new information simply because they have more experience upon which to draw. Inexperienced offenders, on the other hand, should update more to bring their risk perceptions closer in line with their actual punishment risk. Third, Anwar and Loughran observe that the effect of arrest should be crime specific or at least specific within a class of criminal behaviors. That is, an arrest for a violent crime should influence risk perceptions of violent crime only and not risk perceptions of property crime. All three of these predictions were tested and supported from a sample of serious juvenile offenders.

Conclusions

Research on criminal sanctions indicates quite clearly that punishment has a substantial deterrent effect on criminal behavior (Apel and Nagin, 2011). However, devising sensible deterrence-based crime policies requires much better knowledge of the determinants of sanction risk perceptions. In recent years, greater attention has been devoted to analyzing the linkage of sanction risk perceptions to policy and to personal and vicarious experiences with the criminal justice enforcement apparatus. The Bayesian learning model provides a valuable theoretical structure for organizing and building on this research. Given the theoretical significance of sanction risk perceptions, and the policy significance of planful efforts to manipulate them, filling the existing empirical gaps should be a high priority for criminological scholars.

See also: Bayesian Theory, History, Applications, and Contemporary Directions; Data Bases and Statistical Systems: Crime and Delinquency; Decision and Choice: Heuristics; Deterrence Theory: Crime; Prevention of Crime and Delinquency; Questionnaires: Collection of Subjective Expectations Data.

Bibliography

Anwar, Shamena, Loughran, Thomas A., 2011. Testing a Bayesian learning theory of deterrence among serious juvenile offenders. Criminology 49, 667–698.

Apel, Robert, 2013. Sanctions, perceptions, and crime: implications for criminal deterrence. Journal of Quantitative Criminology 29, 67–101.

Apel, Robert, Nagin, Daniel S., 2011. General deterrence: a review of recent evidence. In: Wilson, James Q., Petersilia, Joan (Eds.), Crime and Public Policy, fourth ed. Oxford University Press, New York, pp. 411–436.

Apospori, Eleni, Alpert, Geoffrey, 1993. The role of differential experience with the criminal justice system in changes in perceptions of severity of legal sanctions over time. Journal of Research in Crime and Delinquency 39, 184–194.

Assembly Committee on Criminal Procedure, 1968. Deterrent Effects of Criminal Sanctions. Assembly of the State of California, Sacramento, CA.

Horney, Julie, Marshall, Ineke Haen, 1992. Risk perceptions among serious offenders: the role of crime and punishment. Criminology 30, 575–593.

Kleck, Gary, Sever, Brion, Li, Spencer, Gertz, Marc, 2005. The missing link in general deterrence research. Criminology 43, 623–659.

Klepper, Steven, Nagin, Daniel, 1989. Tax compliance and perceptions of the risks of detection and criminal prosecution. Law and Society Review 23, 209–240.

MacCoun, Robert, Pacula, Rosalie Liccardo, Chriqui, Jamie, Harris, Katherine, Reuter, Peter, 2009. Do citizens know whether their state has decriminalized marijuana? Assessing the perceptual component of deterrence theory. Review of Law and Economics 5, 347–371.

Matsueda, Ross L., Kreager, Derek A., Huizinga, David, 2006. Deterring delinquents: a rational choice model of theft and violence. American Sociological Review 71, 95–122.

Nagin, Daniel S., 1998. Criminal deterrence research at the outset of the twenty-first century. In: Tonry, Michael (Ed.), Crime and Justice: A Review of Research, vol. 23. University of Chicago Press, Chicago, IL, pp. 1–42.

Nagin, Daniel S., Pogarsky, Greg, 2001. Integrating celerity, impulsivity, and extralegal sanction threats into a model of general deterrence: theory and evidence. Criminology 39, 865–891.

Pager, Devah, 2003. The mark of a criminal record. American Journal of Sociology 108, 937–975.

Paternoster, Raymond, 1987. The deterrent effect of the perceived certainty and severity of punishment: a review of the evidence and issues. Justice Quarterly 4, 173–217.

Paternoster, Raymond, Saltzman, Linda E., Chiricos, Theodore G., Waldo, Gordon P., 1982. Perceived risk and deterrence: methodological artifacts in perceptual deterrence research. Journal of Criminal Law and Criminology 73, 1238–1258.

Ross, H. Laurence, 1973. Law, science, and accidents: the British Road Safety Act of 1967. Journal of Legal Studies 2, 1–78.

Stafford, Mark C., Warr, Mark, 1993. A reconceptualization of general and specific deterrence. Journal of Research in Crime and Delinquency 30, 123–135.

Western, Bruce, 2002. The impact of incarceration on wage mobility and inequality. American Sociological Review 67, 526–546.

Zimring, Franklin E., Hawkins, Gordon J., 1973. Deterrence: The Legal Threat in Crime Control. University of Chicago Press, Chicago, IL.

Deterrence Theory: Crime

Mark C Stafford, School of Criminal Justice, Texas State University, San Marcos, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Deterrence theory can be traced to the early utilitarian philosophers, Cesare Beccaria and Jeremy Bentham, who believed that people are motivated to obtain pleasure and avoid pain. Crime, then, can be deterred by increasing the certainty (likelihood), celerity (swiftness), and severity (amount) of legal punishment for committing it. That idea continues to form the basis for contemporary deterrence theory, and that theory is subsumable under the more general social learning and rational choice theories. Contemporary deterrence theorists have recognized several important complexities, including distinctions between specific and general deterrence, absolute and restrictive deterrence, and actual and perceived punishments. Moreover, they have recognized that the deterrent effects of legal punishment may be contingent on a host of other factors, including the threat of extralegal punishment.

Introduction

The word *deterrence* has been linked to different topics, including nuclear deterrence (see *Oxford English Dictionary*), but the focus here is on how legal punishments deter crime. Defined simply, criminal deterrence is the omission or curtailment of crime out of fear of legal punishment.

An early version of deterrence theory was proposed by the utilitarian philosophers Cesare Beccaria (1738–94) and Jeremy Bentham (1748–1832), who believed that people will seek to maximize the pleasurable consequences (rewards) and minimize the painful consequences (punishments) of their behavior (principle of hedonism). To the extent that people anticipate pleasure from crime, they can be deterred by making its consequences more painful. This can be accomplished by making legal punishments more *certain*, *celeritous*, and *severe*. *Certainty* refers to the likelihood of being caught and legally punished for a crime; *celerity* is the swiftness of legal punishment; and *severity* is the amount of legal punishment (Durlauf and Nagin, 2011; Pratt et al., 2006). The ideas proposed by Beccaria and Bentham can be reduced to one concise statement: "The rate for a particular type of crime … [should vary] inversely [negatively] with the celerity, certainty, and severity of punishments of that type of crime" (Gibbs, 1975: p. 5).

Deterrence depends on the interplay of certainty, celerity, and severity. For example, "no level of severity will deter when no prospect of actual punishment is present … [i.e., certainty must be greater than 0] just as certain punishment … will not deter without severity [i.e., severity must be greater than 0]" (Durlauf and Nagin, 2011: p. 17). However, this does not mean that certainty and severity are equally important in deterring crime. There is considerable evidence that the deterrent effects of legal punishment are stronger for certainty than severity (Nagin and Pogarsky, 2001; Paternoster, 2010), and the reason may have to do with the low celerity of severe legal punishments. They often commence long after a crime has been committed and are experienced over a lengthy period (e.g., long prison sentences). According to Beccaria and Bentham, delayed legal punishments reduce punishment severity. "In economics, the concept of discounting is used to explain why future [delayed] consequences might be given less weight than present consequences" (Durlauf and Nagin, 2011: p. 17), and criminal offenders might be prone to greater discounting than other people.

Complexities

Several complexities need to be considered, one of which has to do with the distinction between *specific* and *general* deterrence (Stafford and Warr, 1993). Whereas *specific deterrence* refers to the deterrent effects of directly experiencing a legal punishment, *general deterrence* refers to the deterrent effects of indirectly experiencing it. Thus, when a person is imprisoned for committing a burglary and later is released, the imprisonment deters *specifically* to the extent the released burglar refrains from burglary (and possibly other types of crime as well) out of fear of additional imprisonment. If people other than the imprisoned burglar learn about the imprisonment and refrain from committing the same crime (or other types of crime) out of fear of it happening to them, they have been *generally* deterred. The distinction is important because a given legal punishment, such as imprisonment, could deter only *specifically* or only *generally*, instead of both. The exception is the death penalty, which can deter only generally, if at all. Executed offenders do not refrain from crime out of fear of legal punishment; they are incapacitated rather than specifically deterred.

It may seem that the commission of a crime is itself negative evidence for deterrence. For example, it sometimes is argued in reference to U.S. criminal justice policy that threatens illegal drug users with imprisonment that continued drug use signals the threat of imprisonment did not deter. However, the argument ignores the distinction between *absolute* and *restrictive* deterrence, which are notions that correspond with the words *omission* and *curtailment*, respectively, in the definition of criminal deterrence above (Gibbs, 1975: pp. 32–34). To the extent that people refrain completely from committing a crime (i.e., omission of crime) out of fear of legal punishment, they are deterred *absolutely*. However, they may not be deterred absolutely, but, instead, be deterred *restrictively* by curtailing or limiting the amount of crime they commit, as when people

limit their illegal drug use to reduce their risk of legal punishment, possibly out of a sense of cumulative risk. With restrictive deterrence, the key question is how much crime would occur if there were either no or less threatened legal punishment for it, not whether crimes would occur at all.

Three Central Premises of Deterrence Theory

According to Gibbs (1975), there are three central premises of deterrence theory:

1. The greater the actual certainty, celerity, and severity of legal punishment, the greater the perceived certainty, celerity, and severity of legal punishment.
2. The greater the perceived certainty, celerity and severity of legal punishment, the less the likelihood of crime.
3. The greater the actual certainty, celerity, and severity of legal punishment, the less the likelihood of crime.

Or stated diagrammatically: AP < - > PP > - < CR, where AP denotes actual punishments, PP denotes perceived punishments, CR denotes crime, the symbol < - > denotes a positive relationship, and > - < denotes a negative relationship.

In contrast to the actual certainty, celerity, and severity of legal punishment in a given time and place for a given type of crime, perceived punishments pertain to people's perceptions of the certainty, celerity, and severity of legal punishment for the same crime in the same time and place (Pratt et al., 2006). For example, if the actual certainty of arrest for burglary in a particular U.S. state in a given year is 25% (police make 25 arrests for every 100 burglaries committed), people may perceive the certainty of arrest for burglary to be 25%, or they may perceive it to be higher or lower than 25%. Although the example allows for the possibility of error (over- and underestimation) in people's perceptions of punishment (Paternoster, 2010: p. 785), deterrence theorists since Beccaria and Bentham have assumed there is a strong positive relationship between actual and perceived punishments (Premise 1), and that assumption is important for two reasons. First, "to the extent that changes in actual penal policies do not alter potential offenders' beliefs about the likelihood or severity [or celerity] of punishment, they cannot generate any … deterrence" (Von Hirsch et al., 1999: p. 6). Second, demonstration of a positive relationship between actual and perceived legal punishments is crucial in distinguishing deterrence from other possible preventive consequences of legal punishment. To illustrate, suppose there is a strong negative relationship among U.S. states between the actual certainty of imprisonment for auto theft and the auto theft rate (as would be predicted by Premise 3 above). It may not be fear of legal punishment (deterrence) that accounts for the relatively low auto theft rate in states with a high certainty of imprisonment for auto theft. There are few, if any, opportunities to commit auto theft in prison, so the imprisonment of large numbers of auto thieves, especially high-rate auto thieves, could cause a low auto theft rate through *incapacitation* rather than deterrence. There would be evidence of deterrence only to the extent that people in states with a high actual certainty of imprisonment for auto theft perceived it to be high (Premise 1) and there were low auto theft rates in states with high actual certainty of imprisonment for auto theft (Premise 3) and high perceived certainty of imprisonment for auto theft (Premise 2).

Conditions

There are many conditions under which the certainty, celerity, and severity of legal punishments deter to a greater or lesser degree, and any discussion of deterrence theory must consider them (Tittle et al., 2010: p. 1042). One such condition already has been briefly considered – that the deterrent effects of a particular property of punishment, say certainty, depend on the level of another property of punishment, say severity. In one of the earliest empirical studies of deterrence, Tittle (1969: p. 7) examined the relationship among U.S. states between the actual certainty and severity of imprisonment and rates for seven crime types (e.g., homicide) and found that the actual severity of imprisonment "acted as a deterrent only when there was a high [actual] certainty of imprisonment." In contrast, actual certainty was negatively associated with the crime rates (i.e., the greater the actual certainty, the less the crime rate) regardless of the level of actual severity. Soon after publication of the Tittle study, Grasmick and Bryjak (1980) found in a study of the relationship among individuals between the perceived certainty and severity of punishment and self-reported crime that perceived certainty had stronger deterrent effects at higher levels of perceived severity and also that perceived severity had stronger deterrent effects at higher levels of perceived certainty.

The Grasmick–Bryjak study was particularly noteworthy at the time because of its focus on *perceived* certainty and severity. Recognizing problems in separating the effects of deterrence from incapacitation in studies like that of Tittle, scholars after the early 1980s increasingly believed that deterrence is mainly a psychological theory about individuals' perceived punishments. The prevailing belief became that actual "properties of punishment are important … only because they affect crime through individual perceptions" (Paternoster, 2010: p. 780).

Compared to research on perceived certainty and severity, there has been less research on the perceived celerity of legal punishments, and this includes whether it conditions the deterrent effects of perceived certainty and severity on crime. Nagin and Pogarsky (2001) reported no evidence that threat of delayed legal punishments (low celerity) conditioned the deterrent effects of either perceived certainty or severity on the likelihood of drinking and driving. Both perceived certainty and severity were negatively associated with drinking and driving, but there was no systematic association between it and celerity.

There is some evidence that the deterrent effects of the perceived certainty of legal punishment depend on the level of perceived certainty itself, that there is a nonlinear relationship between perceived certainty and crime. Loughran et al. (2012) have studied youth between ages 14 and 17, who were convicted of a serious crime either in juvenile or adult court. At 6-month intervals over a 36-month period, they were asked how likely was it was "they would be caught and arrested for … : fighting, robbery with gun, stabbing someone, breaking into

a store or home, stealing clothes from a store, vandalism, and auto theft" (p. 723). They also were asked to self-report whether they committed any of 17 crimes that were roughly equivalent to those in the perceived certainty questions. There was clear evidence of a nonlinear relationship. There was no deterrent effect of perceived certainty on a composite scale of the 17 crimes when it was below a tipping point – a perceived certainty value of about 0.3. When perceived certainty was between 0.3 and 0.9, a "10 percent increase in [perceived] risk [was] associated with a … 3.2% percent reduction in the probability of offending" (Loughran et al., 2012: p. 731). However, when perceived certainty was 0.9 or higher, reduction in the probability of offending was over four times greater – 13.3%. Thus, it appears that potential offenders are inattentive to the perceived certainty of legal punishment when it is below a certain threshold, and they may be very attentive when it seems a near certainty they will be caught and punished for committing a crime.

Additional Conditions

A long-held belief about deterrence is that some people are more deterrable than others because of their individual or group characteristics, and there is mounting evidence about what these characteristics are (Jacobs, 2010). Matthews and Agnew (2008: p. 96) have reported that high perceived certainty of detection by the police was less likely to deter crime among high-school students with delinquent peers because they were "more likely to (1) be rewarded for engaging in crime, (2) be sanctioned for *not* engaging in crime, (3) view the consequences of detection as less threatening, (4) temporarily reduce their perceived certainty of detection, and/or (5) give less thought to the likelihood of detection." Moreover, Tittle et al. (2011) have found that legal punishments are more likely to deter citizens of some nations than others – Russians and Ukrainians more than Greeks, seemingly because Greeks are less likely to contemplate committing crime.

There are many reasons why people do not contemplate committing crime, but one of the most important is morality (Tittle et al., 2010). Many conform to the law because they believe it is the morally right thing to do; for them, threats of legal punishment and deterrence are irrelevant. Empirical research has borne this out; for example, Bachman et al. (1992) indicated that perceived punishments for sexual assault reduced the likelihood of committing it, but only among those whose moral condemnation of sexually aggressive behavior was weak.

One of the most widely researched questions about deterrence is whether the deterrent effects of legal punishments depend on self-control (Piquero et al., 2011). People with low self-control have been characterized as impulsive risk takers with a present rather than a future orientation and, consequently, should be less deterrable than persons with high self-control (Gottfredson and Hirschi, 1990). While plausible, there is considerable negative evidence. Wright et al. (2004) found greater deterrent effects of punishment among persons with *low* self-control, and neither Tittle and Botchkovar (2005) nor Gallupe and Baron (2010) reported any moderating effect of self-control on deterrence.

Role of Extralegal Punishments in Deterrence

Although criminal deterrence, by definition, involves imposition of legal punishments, Williams and Hawkins (1986) have pointed out that deterrence also can involve extralegal punishments. To illustrate, while people may be deterred from committing a crime out of fear of being fined (a legally imposed economic loss) or imprisoned (a legally imposed deprivation of liberty), they also may refrain from crime because they fear extralegal punishments that may result from being caught as when some people lose their jobs after criminal convictions. Hence, extralegal punishments may be important to deterrence theory.

Nagin's (1998) study of tax evasion further illustrates the point. He found that there was at least some chance that middle-class persons would cheat on their taxes if enforcement was limited to threat of civil fines. However, there was no chance of cheating if enforcement put their "reputation and community standing at stake" (p. 352). Hence, legal punishments deterred best when tied to extralegal punishments.

In some cases, extralegal punishments may prevent crime only because they are linked to threat of legal punishments. In that connection, Williams and Hawkins (1992) examined the extent to which self-reported wife assault was affected by (1) perceived assault costs and (2) the perceived indirect costs of arrest. Perceived assault costs were measured by men's perceptions of the likelihood of losing (1) a spouse/partner, (2) respect of friends/relatives, and (3) self-respect for committing wife assault. Perceived assault costs were linked only to assault itself, not arrest. The perceived indirect costs of arrest included the same types of losses (e.g., losing a spouse/partner), but they could be triggered by arrest (e.g., survey respondents were asked about the chances of losing the respect of their friends if they were arrested for hitting their spouses or partners). A composite scale of the perceived indirect costs of arrest had a significant negative effect on men's self-reported wife assault, but there was no such effect for perceived assault costs that were not linked to arrest.

The deterrent effects of legal punishment also may depend on the possibility of extralegal punishments. Legal punishments may deter best among persons with the most to lose in the way of extralegal punishments for crime (Piquero et al., 2011). Williams and Hawkins (1986: p. 566) have offered this prediction: "the greater the perceived costs of … [a crime], the greater should be the negative effects of perceptions of certainty [of arrest] on the incidence of crime."

That prediction was supported by results from the famous domestic violence experiments conducted by Sherman and his colleagues. In one such experiment in Milwaukee, Sherman et al. (1992) randomly determined whether 1200 suspected domestic batterers would be arrested or warned by the police without an arrest. They, then, compared rates of new domestic violence by the two groups. Overall, arrest was no more effective than a police warning in decreasing new violence. However, absence of an overall arrest effect concealed a more complex pattern. While arrest decreased new violence among married and employed persons, it increased new violence among the unmarried and unemployed. These findings suggested to Sherman and his colleagues that "the effectiveness of legal sanctions rests on a foundation of informal control

[i.e., extralegal punishments]" (p. 688). Married and employed persons had more to lose in the way of extralegal punishments for domestic violence. Unmarried and unemployed batterers did not perceive police intervention as legitimate and "displaced [their anger about being arrested by the police] onto their present or future romantic partners or other citizens" (Sherman, 1993: p. 465). According to Sherman (1993: p. 460), legal punishments provoke "defiance" (i.e., persistent or more serious crime) to the extent offenders (1) perceive the punishment as unfair, (2) are weakly bonded to other community members, (3) perceive the punishment as stigmatizing, and (4) deny the shame associated with it.

The findings of Sherman et al. (1992) have been partially corroborated by Wooldredge and Thistlethwaite (2002) in a nonexperimental study of persons arrested for misdemeanor partner violence in Ohio. Consistent with the Sherman et al. findings, prosecuted offenders undergoing counseling had lower rearrest rates than arrested offenders who had no charges filed against them if they were more educated and residentially stable (same residence for five or more years). Moreover, probation or jail produced lower rearrest rates if sentenced offenders lived in neighborhoods with greater residentially stability. More educated and residentially stable people have more to lose than other sentenced offenders in the way of extralegal punishments, as do people who live in residentially stable neighborhoods "because of the possible stigma faced" for being sentenced (Wooldredge and Thistlewaite, 2002: p. 49). More educated and residentially stable offenders were not more deterred if they were sentenced to probation or jail without counseling.

Links between Deterrence Theory and Other Theories

Deterrence theory can be linked to other, more general theories of human behavior. Akers (1990) has argued that deterrence theory is derivable from more general *social learning* or *behavioral* principles. "Threat of legal punishment is one source or indicator of aversive stimulus under the general concept of differential reinforcement (balance of rewarding and aversive stimuli)" (p. 658). Differential reinforcement refers to the overall balance of rewards and punishments for behaviors, including past, present, and future rewards and punishments. General deterrence is equivalent to observational/vicarious learning in social learning theory, which involves observations of other people and how their behavior is rewarded and punished; and specific deterrence is equivalent to experiential learning, which involves direct, personal experience with the rewards and punishments for behaviors (Stafford and Warr, 1993).

Both deterrence theory and social learning theory can be linked to *rational choice theory* in that both address the question of why people in certain situations *decide* or *choose* to engage in particular behaviors (Akers, 1990; Paternoster, 2010). However, the meaning of "rational" is ambiguous. To some deterrence researchers, rationality implies conscious and deliberate calculations about the rewards and punishments for crime (e.g., Bachman et al., 1992; Wikstrom, 2008). Others deny that overt mental calculations of any kind are necessarily involved (e.g., Casey and Scholz, 1991). Moreover, rational choices are sometimes conceived as willful rather than determined (e.g., Bachman et al., 1992), but Cornish and Clarke (1986: p. 13) have argued that the "rational choice perspective is neutral with respect to the free will-determinism debate."

Disagreement about the meaning of the word "rational" stems from the existence of different versions of rational choice theory: old rational choice theory and new rational choice theory, each with different assumptions about rationality (Felson, 1993). Old rational choice theory, as set forth by such utilitarian philosophers as Beccaria and Bentham assumed that people engage in conscious and deliberate cost–benefit analysis to maximize the rewards and minimize the punishments for their behavior. The new version of rational choice theory entails weaker assumptions about rationality – that people intuit the rewards and punishments for a behavior and that they are imperfect processors of information and consequently must behave within the limits of their abilities to pursue what they perceive as most satisfying (e.g., Cornish and Clarke, 1986). The assumption of "maximizing" is rejected and replaced by alternative assumptions, such as "satisficing" and "bounded" or "situated" rationality. As for the issue of free will vs determinism, new rational choice theory posits that "people make choices, but they cannot choose the choices available to them" (Felson, 1986: p. 119).

Like new rational choice theory, deterrence theory does not require a strong assumption about rationality. In particular, crime need not be preconceived or carefully planned for legal punishments to deter, only that potential offenders choose between crime and conformity to the law on the basis of the rewards and punishments involved. Also, like new rational choice theory, deterrence theory does not require that people have free will. Indeed, a negative relationship between properties of legal punishment (e.g., the perceived certainty of legal punishment) and crime is as compatible with a strictly deterministic conception of human behavior as it is with free will (Gibbs, 1975: p. 24).

See also: Bentham, Jeremy (1748–1832); Criminology: Development and Current Status of the Discipline; Death Penalty; Decision and Choice: Bounded Rationality; Deterrence: Sanction Perceptions; Deterrent Effect of Police and Prisons; Punishment; Rational Choice Theory in Sociology.

Bibliography

Akers, R.L., 1990. Rational choice, deterrence, and social learning theory in criminology: the path not taken. Journal of Criminal Law and Criminology 81, 653–676.

Bachman, R., Paternoster, R., Ward, S., 1992. The rationality of sexual offending: testing a deterrence/rational choice conception of sexual assault. Law and Society Review 26, 343–372.

Casey, J.T., Scholz, J.T., 1991. Beyond deterrence: behavioral decision theory and tax compliance. Law and Society Review 25, 821–843.

Cornish, D., Clarke, R., 1986. Introduction. In: Cornish, D., Clarke, R. (Eds.), The Reasoning Criminal: Rational Choice Perspectives on Offending. Springer-Verlag, New York, pp. 1–16.

Durlauf, S.N., Nagin, D.S., 2011. Overview of 'imprisonment and crime: can both be reduced''? Criminology and Public Policy 10, 9–54.

Felson, M., 1986. Linking criminal choices, routine activities, informal controls, and criminal outcomes. In: Cornish, D., Clarke, R. (Eds.), The Reasoning Criminal: Rational Choice Perspectives on Offending. Springer-Verlag, New York, pp. 119–128.

Felson, M., 1993. Book review. American Journal of Sociology 98, 1497–1499.
Gallupe, O., Baron, S.W., 2010. Morality, self-control, deterrence, and drug use: street youths and situational action theory. Crime and Delinquency 20, 1–22.
Gibbs, J.P., 1975. Crime, Punishment, and Deterrence. Elsevier, New York.
Gottfredson, M.R., Hirschi, T., 1990. A General Theory of Crime. Stanford University Press, Stanford, CA.
Grasmick, H.G., Bryjak, G.J., 1980. The deterrent effect of perceived severity of punishment. Social Forces 59, 471–491.
Jacobs, B.A., 2010. Deterrence and deterrability. Criminology 48, 417–441.
Loughran, T.A., Pogarsky, G., Piquero, A., Paternoster, R., 2012. Re-examining the functional form of the certainty effect in deterrence theory. Justice Quarterly 29, 712–741.
Matthews, S.K., Agnew, R., 2008. Extending deterrence theory: do delinquent peers condition the relationship between perceptions of getting caught and offending? Journal of Research in Crime and Delinquency 45, 91–118.
Nagin, D.S., 1998. Deterrence and incapacitation. In: Tonry, M. (Ed.), The Handbook of Crime and Punishment. Oxford University Press, New York, pp. 345–368.
Nagin, D.S., Pogarsky, G., 2001. Integrating celerity, impulsivity, and extralegal sanction threats into a model of general deterrence: theory and evidence. Criminology 39, 865–892.
Paternoster, R., 2010. How much do we really know about criminal deterrence? Journal of Criminal Law and Criminology 100, 765–823.
Piquero, A.R., Paternoster, R., Pogarsky, G., Loughran, T., 2011. Elaborating the individual difference component in deterrence theory. Annual Review of Law and Social Science 7, 335–360.
Pratt, T.C., Cullen, F.T., Blevins, K.R., Daigle, L.E., Madensen, T.D., 2006. The empirical status of deterrence theory: a meta-analysis. In: Cullen, F.T., Wright, J.P., Blevins, K.R. (Eds.), Taking Stock: The Status of Criminological Theory. Transaction Publishers, New Brunswick, pp. 367–395.
Sherman, L.W., 1993. Defiance, deterrence, and irrelevance: a theory of the criminal sanction. Journal of Research in Crime and Delinquency 30, 445–473.
Sherman, L.W., Smith, D.A., Schmidt, J.D., Rogan, D.P., 1992. Crime, punishment, and stake in conformity: legal and informal control of domestic violence. American Sociological Review 57, 680–690.
Stafford, M.C., Warr, M., 1993. A reconceptualization of general and specific deterrence. Journal of Research in Crime and Delinquency 30, 123–135.
Tittle, C.R., 1969. Crime rates and legal sanctions. Social Problems 16, 408–423.
Tittle, C.R., Botchkovar, E.V., 2005. Self-control, criminal motivation and deterrence: an investigation using Russian respondents. Criminology 43, 307–353.
Tittle, C.R., Antonaccio, O., Botchkovar, E., Kranidioti, M., 2010. Expected utility, self-control, morality, and criminal probability. Social Science Research 39, 1029–1046.
Tittle, C.R., Botchkovar, E.V., Antonaccio, O., 2011. Criminal contemplation, national context, and deterrence. Journal of Quantitative Criminology 27, 224–249.
Von Hirsch, A., Bottoms, A.E., Burney, E., Wikstrom, P.H., 1999. Criminal Deterrence and Sentence Severity: An Analysis of Recent Research. Hart Publishing, Oxford, UK.
Wikstrom, P.H., 2008. Deterrence and deterrence experiences: preventing crime through the threat of punishment. In: Shohan, G.S., Beck, O., Kett, M. (Eds.), International Handbook of Penology and Criminal Justice. CRC Press, Boca Raton, pp. 345–378.
Williams, K.R., Hawkins, R., 1986. Perceptual research on criminal deterrence: a critical review. Law and Society Review 20, 545–572.
Williams, K.R., Hawkins, R., 1992. Wife assault, costs of arrest, and the deterrence process. Journal of Research in Crime and Delinquency 29, 292–310.
Wooldredge, J., Thistlethwaite, A., 2002. Reconsidering domestic violence recidivism: conditioned effects of legal controls by individual and aggregate levels of stake in conformity. Journal of Quantitative Criminology 18, 45–70.
Wright, B.R.E., Caspi, A., Moffitt, T.E., Paternoster, R., 2004. Does the perceived risk of punishment deter criminally prone individuals? Rational choice, self-control, and crime. Journal of Research in Crime and Delinquency 41, 180–213.

Deterrent Effect of Police and Prisons

Robert Apel, Rutgers University, Newark, NJ, USA
Daniel S Nagin, Carnegie Mellon University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D.S. Nagin, volume 5, pp. 3544–3550, © 2001, Elsevier Ltd.

Abstract

This article reviews the evidence on the deterrent effect of police and imprisonment. Studies of changes in police presences, whether achieved by changes in police numbers or in their strategic deployment, consistently find evidence of deterrent effects. Studies of the deterrent effect of increases in already long prison sentences find at most a modest deterrent effect. Three high-priority areas for future research are identified: developing and testing an integrated model of the effects of the threat and experience of punishment, measuring perceptions of sanction regimes, developing and estimating the deterrent effect of shorter prison sentences and identifying high-deterrence policies.

Overview

The criminal justice system dispenses justice by apprehending, prosecuting, and punishing individuals who break the law. These activities may also prevent crime by three distinct mechanisms – incapacitation, specific deterrence, and general deterrence. Convicted offenders are often punished with imprisonment. Incapacitation refers to the crimes averted by their physical isolation during the period of their incarceration. Specific deterrence and general deterrence involve possible behavioral responses. Specific deterrence refers to the reduction in reoffending that is presumed to follow from the experience of actually being punished. However, there are many sound reasons for suspecting that the experience of punishment might instead increase reoffending. The threat of punishment might also discourage potential and actual criminals in the general public from committing crime. This effect is known as general deterrence and is the subject of this article.

Key Concepts of Deterrence

Deterrence is a theory of choice in which would-be offenders balance the benefits and costs of crime. Benefits may be pecuniary in the case of property crime but may also involve intangibles such as defending one's honor, expressing outrage, demonstrating dominance, cementing a reputation, or seeking a thrill. The potential costs of crime are comparably varied. For example, crime can entail personal risk if the victim resists, and it may also invoke pangs of conscience or shame. The theory of deterrence is predicated on the idea that if state-imposed sanction costs are sufficiently severe, criminal activity will be discouraged, at least for some. Thus, one of the key concepts of deterrence is the severity of punishment. The authors' review of severity effects focuses on research findings concerning imprisonment.

Severity alone, however, cannot deter. There must also be some possibility that the sanction will be incurred if the crime is committed. For that to happen, the offender must be apprehended, usually by the police. He must next be charged and successfully prosecuted, and finally sentenced by the judiciary. None of these successive stages in processing through the criminal justice system is certain. Thus, another key concept in deterrence theory is the certainty of punishment. In this regard, the most important set of actors are the police – absent detection and apprehension, there is no possibility of conviction or punishment. For this reason, the authors discuss separately what is known about the deterrent effect of police.

One of the key conclusions that emerged from the 1960s- and 1970s-era deterrence literature was that the certainty of punishment was a more powerful deterrent than the severity of punishment. The analyses of this era generally used cross-sectional data on states and involved testing the effects on the statewide crime rate of the certainty and severity of punishment, along with other demographic and socioeconomic control variables. The certainty of punishment was measured by the ratio of prison admissions to the number of reported crimes, while the severity of punishment was measured by median time served of recent prison releases. The basis for the 'certainty not severity' deterrence conclusion was that punishment certainty was consistently found to have a negative and significant association with the crime rate, whereas punishment severity generally had no significant association.

This conclusion at the time was probably based on faulty statistical inference. Two primary criticisms were leveled. The first was that the negative association between the certainty measure and crime rate was an artifact of the number of crimes appearing in the denominator of the certainty measure and the numerator of the crime rate. It can be mathematically demonstrated that errors in the measurement of number of crimes, of which there are many, will force a negative, deterrent-like association between the crime rate and certainty even if, in fact, the certainty of punishment had no deterrent effect on crime. The second involved the use of theoretically indefensible statistical methods for parsing out the cause–effect relationship between sanction levels and the crime rate. After all, sanctions may deter crime, but crime may also affect sanction levels. For example, perhaps overcrowded prisons might reduce the chances of newly caught offenders going to prison. However, subsequent findings from the so-called perceptual deterrence literature and economic studies of the effects of contact with

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86021-3

the criminal justice system on access to legal labor markets provide a far firmer empirical and theoretical basis for the 'certainty principle.' In order to focus on more recent research findings, this article will not cover these earlier research traditions.

The Deterrent Effect of Imprisonment

There have been two distinct waves of studies of the deterrent effect of imprisonment. As already noted, studies in the 1960s and 1970s examined the relationship of the crime rate to the certainty of punishment, measured by the ratio of prison admissions to reported crimes, and the severity of punishment as measured by median time served. These studies suffered from a number of serious statistical flaws that are detailed in Blumstein et al. (1978). In response to these deficiencies, a second generation of studies emerged in the 1990s. Unlike the first-generation studies, which primarily involved cross-sectional analyses of states, second-generation studies had a longitudinal component in which data were analyzed not only across states but also over time. Another important difference is that the second-generation studies did not attempt to estimate certainty and severity effects separately. Instead, they examined the relationship between the crime rate and rate of imprisonment as measured by prisoners per capita.

A review by Donohue (2007) identifies six such studies. All find statistically significant negative associations between imprisonment rates and crimes rates, implying a crime prevention effect of imprisonment. However, the magnitude of the estimate varied widely – from nil for a study that allowed for the possibility of diminishing returns, to an elasticity of −0.4. (By elasticities of 0.0 and −0.4, the authors mean that 10% growth in the imprisonment rate had either no effect on the crime rate or resulted in a substantial decline of 4%, respectively.) It is important to note that these studies are actually measuring a combination of deterrent and incapacitation effects. Thus, it is impossible to decipher the degree to which crime prevention is occurring because of a behavioral response by the population at large or because of the physical isolation of crime-prone people.

Donohue (2007) goes on to show that the small elasticity estimates imply that the current imprisonment rate is too large, while the high-end estimates imply the rate is too small. He lists a variety of technical shortcomings of these studies that, in our view, make it impossible to distinguish among the widely varying effect size estimates. The most important is the degree to which the studies were successful in separating cause from effect. While imprisonment prevents crime through a combination of deterrence and incapacitation, crime also generates the prison population. This is an example of what is called the "simultaneity problem," whereby one wants to ascertain the effect of one variable (the imprisonment rate) on another variable (the crime rate) in a circumstance where one knows or suspects that reverse causation is also present, namely, that the crime rate simultaneously affects the imprisonment rate. Thus, statistical isolation of the crime prevention effect requires properly accounting for the effect of crime on imprisonment. The Levitt (1996) study is arguably the most successful in this regard. It uses court-ordered prison releases as an instrument for untangling the cause-and-effect relationship. However, even the Levitt analysis suffers from many of the technical limitations detailed by Donohue.

More fundamentally, this literature suffers from more than just technical shortcomings that future research might strive to correct. It also suffers from important conceptual flaws that limit its usefulness in devising crime-control policy. Prison population is not a policy variable; rather, it is an outcome of sanction policies dictating who goes to prison and for how long, namely, the certainty and severity of punishment. In all incentive-based theories of criminal behavior, the deterrence response to sanction threats is posed in terms of the certainty and severity of punishment, not in terms of the imprisonment rate. Therefore, to predict how changes in certainty and severity might affect the crime rate requires knowledge of the relationship of the crime rate to certainty and severity as separate entities, which is not provided by the literature that analyzes the relationship of the crime rate to the imprisonment rate. The studies are also conducted at too global a level. There are good reasons for predicting differences in the crime reduction effects of different types of sanctions (e.g., mandatory minimums for repeat offenders versus prison diversion programs for first-time offenders). Obvious sources of heterogeneity in offender response include factors such as prior contact with the criminal justice system, demographic characteristics, and the mechanism by which sanction threats are communicated to their intended audience.

Three studies nicely illustrate heterogeneity in the deterrence response to the threat of imprisonment: the Weisburd et al. (2008) study on the use of imprisonment to enforce fine payment finds a substantial deterrent effect, the Helland and Tabarrok (2007) analysis of the deterrent effect of California's third strike provision finds only a modest deterrent effect, and the Lee and McCrary (2009) examination of the heightened threat of imprisonment that attends coming under the jurisdiction of the adult courts at the age of majority finds no deterrent effect.

Weisburd et al. (2008) report on a randomized field trial of alternative strategies for incentivizing the payment of court-ordered fines. The most salient finding involves the 'miracle of the cells,' namely, that the imminent threat of incarceration is a powerful incentive for paying delinquent fines. The miracle of the cells, the authors believe, provides a valuable vantage point for considering the oft-repeated conclusion from the deterrence literature that the certainty rather the severity of punishment is the more powerful deterrent. Consistent with the 'certainty principle,' the common feature of treatment conditions involving incarceration was a high certainty of imprisonment for failure to pay the fine. However, the fact that Weisburd and colleagues label the response the 'miracle of the cells' and not the 'miracle of certainty' is telling. Their choice of label is a reminder that certainty must result in a distasteful consequence, namely, incarceration in this experiment, in order for it to be a deterrent. The consequences need not be draconian, just sufficiently costly to deter proscribed behavior.

Helland and Tabarrok (2007) examine whether California's "Three Strikes and You're Out" law deters offending among individuals previously convicted of strike-eligible offenses. The future offending of individuals convicted of two previous

strike offenses was compared with that of individuals who had been convicted of only one strike offense but who, in addition, had been tried for a second strike offense but were ultimately convicted of a nonstrike offense. The study demonstrates that these two groups of individuals were comparable on many characteristics such as age, race, and time in prison. Even so, it finds that arrest rates were about 20% lower for the group with convictions for two strike offenses. The authors attribute this reduction to the greatly enhanced sentence that could have accompanied conviction for a third strike offense.

For most crimes, the certainty and severity of punishment increases markedly upon reaching the age of majority, when jurisdiction for criminal wrongdoing shifts from the juvenile to the adult court. In an extraordinarily careful analysis of individual-level crime histories from Florida, Lee and McCrary (2009) attempt to identify a decline in the risk of offending at age 18 years, the age of majority in Florida. However, they were unable to detect any significant change in crime risk.

In combination, these three studies nicely illustrate that the deterrent effect of the threat of punishment is context specific and that debates about whether deterrence works or not are ill posed. Instead, the discussion should be in terms of whether the specific sanction deters or not and if it does, whether the benefits of crime reduction are sufficient to justify the costs of imposing the sanction. To illustrate, while Helland and Tabarrok (2007) conclude that the third-strike effect in California is a deterrent, they also conclude, based on a cost-benefit analysis, that the crime-saving benefits are likely far smaller than the increased costs of incarceration. The Helland and Tabarrok study is an exemplar of the approach that should be taken in evaluating different sanctioning regimes.

The Deterrent Effect of Police

The police may prevent crime through many possible mechanisms. Apprehension of active offenders is a necessary first step for their conviction and punishment. If the sanction involves imprisonment, crime may be prevented by the incapacitation of the apprehended offender. The apprehension of active offenders may also deter would-be criminals by increasing their perception of the risk of apprehension and thereby the certainty of punishment. Many police tactics such as rapid response to calls for service at crime scenes or post-crime investigation are intended not only to capture the offender but also to deter others by projecting a tangible threat of apprehension. Police may, however, deter without actually apprehending criminals because their very presence projects a threat of apprehension if a crime were to be committed. Indeed, some of the most compelling evidence of deterrence involve instances where there is complete or near-complete collapse of police presence. In September 1944, German soldiers occupying Denmark arrested the entire Danish police force. According to an account by Andenaes (1974), crime rates rose immediately but not uniformly. The frequency of street crimes like robbery, whose control depends heavily upon visible police presence, rose sharply. By contrast, crimes like fraud were less affected.

The Andenaes anecdote illustrates two important points. First, sanction threats (or the absence thereof) may not uniformly affect all types of crime and more generally all types of people. Second, it draws attention to the difference between absolute and marginal deterrence. Absolute deterrence refers to the difference in the crime rate between the status quo level of sanction threat and a complete (or near) absence of sanction threat. The Andenaes anecdote is a compelling demonstration that the absolute deterrent effect is large. However, from a policy perspective, the important question is whether, on the margin, crime deterrence can be affected by incrementally manipulating sanction threats.

Research on the marginal deterrent effect of police has evolved in two distinct literature. One has focused on the deterrent effect of the aggregate police presence measured, for example, by the relationship between police per capita and crime rates. The other has focused on the crime prevention effectiveness of different strategies for deploying police. The authors review these two literature separately.

Aggregate Police Presence and Crime

Studies of police hiring and crime rates have been plagued by a number of impediments to causal inference. Among these are cross-jurisdictional differences in the recording of crime, feedback effects from crime rates to police hiring, the confounding of deterrence with incapacitation, and aggregation of police manpower effects across heterogeneous units, among others. Yet the challenge that has received the most attention in empirical applications is the simultaneity problem referred to in the previous section – in the present case, the feedback from crime rates to police hiring.

The two studies of police manpower by Marvell and Moody (1996) and Levitt (1997) are notable for their different identification strategies. The Marvell and Moody (1996) study is based on an analysis of two panel data sets, one composed of 49 states for the years 1968–93 and the other of 56 large cities for the years 1971–92. In order to overcome the problem of a bidirectional relationship between police manpower and the crime rate, they examine the effect of police manpower in the prior year on the crime rate in the current year. The strongest evidence for an impact of police hiring on total crime rates comes from the city-level analysis, with an estimated elasticity of -0.3, meaning that 10% growth in police manpower produces a 3% decline in the crime rate the following year, representing a modestly large effect.

However, regression analyses of this type do not generally provide a valid basis for making causal claims. But other forms of analysis can provide such a basis. One is an approach that exploits pseudorandom variation in policies that are only correlated with the crime rate through their effects on police manpower, in what is known as an instrumental variables (IVs) regression. Levitt (1997) performs this type of analysis from a panel of 59 large cities for the years 1970–92. Reasoning that political incumbents have incentives to devote resources to increasing the size of the police force in anticipation of upcoming elections, he uses election cycles to help untangle the cause–effect relationship between crime rates and police manpower. Levitt's model produces a very large elasticity of about -1.0 for the violent crime rate and a more modest but still meaningful elasticity of -0.3 for the property

crime rate. Following Levitt's use of the electoral cycle as an instrument for the number of sworn police officers, other studies have employed alternative IVs and reported comparable elasticities.

In recent years, a number of more targeted tests of the police–crime relationship have appeared. These studies investigate the impact on the crime rate of reductions in police presence and productivity as a result of massive budget cuts or lawsuits following racial profiling scandals. Each of these studies concludes that increases (decreases) in police presence and activity substantially decrease (increase) crime. By the way of example, Shi (2009) studies the fallout from an incident in Cincinnati in which a white police officer shot and killed an unarmed African-American suspect. The incident was followed by 3 days of rioting, heavy media attention, the filing of a class action lawsuit, a federal civil rights investigation, and the indictment of the officer in question. These events created an unofficial incentive for officers from the Cincinnati Police Department to curtail their use of arrest for misdemeanor crimes, especially in communities with higher proportional representation of African-Americans out of concern for allegations of racial profiling. Shi demonstrates measurable declines in police productivity in the aftermath of the riot and also documents a substantial increase in criminal activity. The estimated elasticities of crime to policing based on her approach were −0.5 for violent crime and −0.3 for property crime, both of which have policy significance.

The ongoing threat of terrorism has also provided a number of unique opportunities to study the impact of police resource allocation in cities around the world. The study by Klick and Tabarrok (2005) examines the effect on crime of the color-coded alert system devised by the US Department of Homeland Security (in the aftermath of the 11 September 2001, terrorist attack) to denote the terrorism threat level. Its purpose was to signal federal, state, and local law enforcement agencies to occasions when it might be prudent to divert resources to sensitive locations. Klick and Tabarrok (2005) use daily police reports of crime for the period of March 2002–July 2003, during which time the terrorism alert level rose from 'elevated' (yellow) to 'high' (orange) and back down to 'elevated' on four occasions. During high alerts, anecdotal evidence suggested that police presence increased by 50%. Their estimate of the elasticity of total crime to changes in police presence as the alert level rose and fell was −0.3.

To summarize, aggregate studies of police presence conducted since the mid-1990s consistently find that putting more police officers on the street – either by hiring new officers or by allocating existing officers in ways that put them on the street in larger numbers or for longer periods of time – has a substantial deterrent effect on serious crime. There is also consistency with respect to the size of the effect. Most estimates reveal that a 10% increase in police presence yields a reduction in total crime in the neighborhood of 3%, although studies that consider violent crime tend to find reductions ranging from 5 to 10%. Yet these police manpower studies speak only to the number and allocation of police officers and not to what police officers actually do on the street beyond making arrests. The next section proceeds from here by reviewing recent evaluations of deployment strategies used by police departments in order to control crime.

Police Deployment and Crime

Much research has examined the crime prevention effectiveness of alternative strategies for deploying police resources. This research has largely been conducted by criminologists and sociologists. Among this group of researchers, the preferred research designs are quasi-experiments involving before-and-after studies of the effect of targeted interventions as well as true randomized experiments, which are more difficult to implement and thus rarer, but yield the most internally valid results. The discussion that follows draws heavily upon two excellent reviews of this research by Weisburd and Eck (2004) and Braga (2008). As a preface to this summary, we draw the theoretical link between police deployment and the certainty and severity of punishment. For the most part, deployment strategies affect the certainty of punishment through its impact on the probability of apprehension. There are, however, notable examples where severity may also be affected.

One way to increase apprehension risk is to mobilize police in a fashion that increases the probability that an offender is arrested after committing a crime. Strong evidence of a deterrent as opposed to an incapacitation effect resulting from the apprehension of criminals is limited. Studies of the effect of rapid response to calls for service (Spelman and Brown, 1981) find no evidence of a crime prevention effect, but this may be because most calls for service occur well after the crime event, with the result that the perpetrator has fled the scene. Thus, it is doubtful that rapid response materially affects apprehension risk. Similarly, because most arrests result from the presence of witnesses or physical evidence, improved investigations are not likely to yield material deterrent effects because, again, apprehension risk is not likely to be affected. A series of randomized experiments were conducted to test the deterrent effect of mandatory arrest for domestic violence. The initial experiment conducted in Minneapolis by Sherman and Berk (1984) found that mandatory arrest was effective in reducing domestic violence reoffending. However, findings from follow-up replication studies (as part of the so-called Spouse Assault Replication Program) were inconsistent.

The second source of deterrence from police activities involves averting crime before it occurs, that is, crime prevention. In this circumstance, there is no apprehension because there was no offense. In the authors' view, this is the primary source of deterrence from the presence of police – the proverbial 'cop on the corner.' If an occupied police car is parked outside a liquor store, a would-be robber of the store will likely be deterred because apprehension is all but certain. Thus, measures of apprehension risk based only on arrest rates and reported crimes are seriously incomplete because such measures do not capture the criminal opportunities that were not acted upon by potential offenders because the risk was deemed too high. Consequently, there is reason to believe that deterrent effects of police presence are actually larger than what are commonly reported, because of the challenge posed by the measurement of behavior which never occurs.

Two examples of police deployment strategies that have been shown to be effective in averting crime are 'hot spots'

policing and 'problem-oriented' policing. A much broader review of these practices is provided in a recent National Academy of Sciences report (Committee to Review Research on Police Policy and Practices, 2004). Weisburd and Eck (2004) propose a two-dimensional taxonomy of policing strategies. One dimension is 'level of focus' and the other is 'diversity of focus.' Level of focus represents the degree to which police activities are targeted. Targeting can occur in variety of ways, but Weisburd and Eck give special attention to policing strategies that target police resources in small geographic areas (e.g., blocks or specific addresses) that have very high levels of criminal activity, so-called crime hot spots. Just as in the liquor store example, the rationale for concentrating police in crime hot spots is to create a prohibitively high risk of apprehension and thereby to eliminate the opportunity to commit crime at the hot spot in the first place.

Braga's (2008) informative review of hot spots policing summarizes the findings from nine experimental or quasi-experimental evaluations. The studies were conducted in five large US cities and one suburb in Australia. Crime-incident reports and citizen calls for service were used to evaluate impacts in and around the geographic area of the crime hot spot. The targets of the police actions varied. Some hot spots were generally high-crime locations, whereas others were characterized by specific crime problems like drug trafficking. All but two of the studies found evidence of significant reductions in crime. Further, while not all of the studies examined crime displacement, among those studies that did, no evidence was found of material displacement to immediately surrounding locations. On the contrary, some studies found evidence of crime reductions, not increases, in the surrounding locations – a 'diffusion of crime-control benefits' to non-targeted locales.

The second dimension of the Weisburd and Eck taxonomy is diversity of approaches. This dimension concerns the variety of approaches that police use to impact public safety. Low diversity is associated with reliance on time-honored law enforcement strategies for affecting the threat of apprehension, for example, by dramatically increasing police presence. High diversity involves expanding beyond conventional practice to prevent crime. One example of a high-diversity approach is problem-oriented policing. Problem-oriented policy comes in so many different forms that it is regrettably hard to define.

One of the most visible examples of problem-oriented policing is Boston's Operation Cease Fire (Kennedy et al., 2001). The objective of the collaborative operation was to prevent intergang gun violence using two deterrence-based strategies. One was to target enforcement against weapons traffickers who were supplying weapons to Boston's violent youth gangs. The second involved a more innovative use of deterrence. The youth gangs themselves were assembled (and reassembled) to send the message that the response to any instance of serious violence would be 'pulling every lever' legally available to punish gang members collectively. This included a salient severity-related dimension – vigorous prosecution for unrelated, nonviolent crime such as drug dealing. Thus, the aim of Operation Cease Fire was to deter violent crime by increasing the certainty and severity of punishment but only in targeted circumstances, namely, if the gang members were perpetrators of a violent crime. Just as important, Operation Cease Fire illustrates the potential for combining elements of both certainty and severity enhancement to generate a targeted deterrent effect. Further evaluations of the efficacy of this and other problem-solving strategies should be a high priority.

Conclusions

This article has reviewed the evidence on the general deterrent effect of the use of imprisonment as a sanction for criminal behavior, as well as the size of police departments and selective policing strategies. Evidence of substantial deterrent effects is considerable. Therefore, a well-balanced portfolio of strategies and programs to prevent crime must necessarily include deterrence-based policies. That being said, policy makers should be sensitive to the possibility that not all deterrence policies will be effective in reducing crime or, if effective, that the crime-reduction benefits may fall short of the social and economic costs of the sanction. A 'one-size-fits-all' crime policy is hard to justify.

Future research on the utility of prisons and police for crime control will be most useful for policy evaluation if it moves closer to a medical model. Medical research is not organized around the theme of whether medical care cures diseases, the analog to the question of whether criminal justice agencies prevent crime. Instead, medical researchers address far more specific questions. Is a specific drug or procedure effective in treating a specific disease? Does the drug or procedure have adverse side effects for certain types of people? Furthermore, most such research is comparative – is the specific drug or procedure more effective than the status quo alternative? The analogous questions for deterrence research are whether and in what circumstances are the use of police and prisons the most effective.

See also: Crime: Knowledge about and Prevalence; Data Bases and Statistical Systems: Crime and Delinquency; Empirical Legal Studies.

Bibliography

Andenaes, Johannes, 1974. Punishment and Deterrence. University of Michigan Press, Ann Arbor.

Blumstein, Alfred, Cohen, Jacqueline, Nagin, Daniel, 1978. Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates. National Academies Press, Washington, DC.

Braga, Anthony A., 2008. Police Enforcement Strategies to Prevent Crime in Hot Spot Areas. Office of Community Oriented Policing, U.S. Department of Justice, Washington, DC.

Committee to Review Research on Police Policy and Practices, 2004. Fairness and Effectiveness in Policing: the Evidence. National Academies Press, Washington, DC.

Donohue, John, 2007. Assessing the Relative Benefits of Incarceration: the Overall Change over the Previous Decades and the Benefits on the Margin. Working paper. Yale Law School, New Haven, CT.

Helland, Eric, Tabarrok, Alexander, 2007. Does three strike deter? A nonparametric estimation. Journal of Human Resources 42, 309–330.

Kennedy, David M., Braga, Anthony A., Morrison Piehl, Anne, Waring, Elin J., 2001. Reducing Gun Violence: the Boston Gun Project's Operation Ceasefire. U.S. National Institute of Justice, Washington, DC.

Klick, Jonathan, Tabarrok, Alexander, 2005. Using terror alert levels to estimate the effect of police on crime. Journal of Law and Economics 48, 267–279.

Lee, David S., Justin McCrary, 2009. The Deterrent Effect of Prison: Dynamic Theory and Evidence. Unpublished manuscript. Princeton University, Princeton, NJ.

Levitt, Steven D., 1996. The effect of prison population size on crime rates: evidence from prison overcrowding legislation. Quarterly Journal of Economics 111, 319–352.

Levitt, Steven D., 1997. Using electoral cycles in police hiring to estimate the effect of police on crime. American Economic Review 87, 270–290.

Marvell, Thomas, Moody, Carlisle, 1996. Specification problems, police levels, and crime rates. Criminology 34, 609–646.

Sherman, Lawrence W., Berk, Richard A., 1984. The specific deterrent effects of arrest for domestic assault. American Sociological Review 49, 261–272.

Shi, Lan, 2009. The limits of oversight in policing: evidence from the 2001 Cincinnati riot. Journal of Public Economics 93, 99–113.

Spelman, William, Brown, Dale K., 1981. Calling the Police: a Replication of the Citizen Reporting Component of the Kansas City Response Time Analysis. Police Executive Research Forum, Washington, DC.

Weisburd, David, Eck, John, 2004. What can police do to reduce crime, disorder, and fear? Annals of the American Academy of Political and Social Science 593, 42–65.

Weisburd, David, Einat, Tomar, Kowalski, Matt, 2008. The miracle of the cells: an experimental study of interventions to increase payment of court-ordered financial obligations. Criminology and Public Policy 7, 9–36.

Deutsch, Karl W (1912–92)

Mattei Dogan, National Center for Scientific Research, Paris, France

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 5, pp. 3553–3555, © 2001, Elsevier Ltd.

Abstract

Karl W Deutsch remains in the history of political science as a leading figure in three domains: comparative politics, quantitative analysis, and international relations. He had a great capacity to interpret statistical data with an intimate knowledge of the historical and social contexts. Several concepts designed or developed by him, like social communication, rate of political mobilization, or spread of national consciousness, belong today the common patrimony of social sciences. In the field of international relations, he has emphasized the importance of transnational trade for political integration. Deutsch is author, co-author, editor, or co-editor of 16 books and more than 200 articles in academic journals. His most important books are *Nationalism and Social Communication*, *Political Community at the International Level*, and *The Nerves of Governments*. As an interdisciplinary scholar, Deutsch has co-edited a book on *Advances in the Social Sciences, 1900–80*.

Karl Wolfgang Deutsch belongs to the cohort of the 'brain drain' which, by emigrating in the 1930s from Europe, contributed to the development of the social sciences in the United States. Born in Prague on 21 July 1912, Deutsch escaped from the imprisoned continent in 1939, after having received a law degree from Charles University. Son of the first woman parliamentarian in Czechoslovakia, he moved from one privileged social stratum to another, debarking with a fellowship at Harvard University. He worked for the Office of Strategic Studies during World War II, began teaching at MIT in 1945, then at Yale University in 1957 and, from there, taught at Harvard University until retiring in 1985. At that time, he was appointed as one of the directors of the Wissenschaftzentrum in Berlin. As is often the case with those who learn a new language after puberty, Karl Deutsch kept his distinctive German accent. He died at the age of 80 in Cambridge, MA, on 1 November 1992.

Karl Deutsch's contributions to the advancement of social sciences can be summed up in five domains:

1. Building empirically grounded concepts and theories, such as nation-building, social mobilization, national and international integration, center–periphery, distribution of power, and understanding macrosocial processes. His definition of integration of a society is still in use in the literature: the capacity to receive and transmit information on wide ranges of different sectors with little delay. One of his writings appears retrospectively as predictive: "Cracks in the monolith, possibilities and patterns of disintegration in totalitarian systems" (Deutsch, 1954).
2. Pioneering, cross-national research. Deutsch remains a leading figure of the first post-War generation of comparativists. In the late 1950s and 1960s, he did comparative research, using aggregate data at a time when other great comparativists, such as Gabriel Almond and Sidney Verba, preferred the survey method for comparing countries.
3. Quantitative methodology. Karl Deutsch advanced the introduction of statistical analysis in political science. Such a vocation could be explained by the fact that, as the son of an optician, early in life, Deutsch considered the possibility of becoming one himself, and consequently studied mathematics. He did quantitative analysis on the size of government, growth of governments, foreign trade as a percentage of GDP, global modeling, two-way channels of communication between élites and mass. Deutsch was in the forefront of the so-called 'data movement.' His contribution to *Quantitative Ecological Analysis in the Social Sciences* (Deutsch, 1969) was considered as a research agenda.
4. Work on communication theory, particularly cybernetics. He developed Wiener and Neumann's research, focusing on politics and society. The subtitle of *The Nerves of Government*, 'Political Communication and Control,' (Deutsch, 1963) paraphrases Wiener's definition of cybernetics: 'Control and Communication in the Animal and Machine.'
5. Bridges between disciplines, by importing in political science concepts and theories from various disciplines (anthropology, social psychology, economics, and biology). *Nationalism and Social Communication* (Deutsch, 1953) is an eminently interdisciplinary work, where Deutsch's intimate knowledge of European history and geography appears in all chapters. Deutsch was a leading contributor to *Advances in the Social Sciences* 1900–80 (Deutsch et al., 1986).

Retrospectively, it can be said that Deutsch's most eminent book is his first, *Nationalism and Social Communication*: *An Inquiry into the Foundation of Nationality* (1953). This is his most cited work, and has been so since its publication. The subject of this book is related to his personal experience. Born to a Sudeten German family in the Austro-Hungarian empire, a multinational state, "Deutsch put his emotional and intellectual energy into scholarship focused upon nationalism and formation of large-scale political communities" (Merritt and Russett, 1981: 6). As a teenager, Deutsch witnessed the collapse of three empires and the emergence of many new independent nation-states in Eastern Europe and the Balkans.

The success of this book can be explained by Deutsch's capacity to interpret statistical data with an intimate knowledge of the historical and social contexts. There are social scientists who write about countries they have only read about in books and who, in spite of their statistical virtuosity, too often offer superficial explanations. They should contemplate the following: "We will always need the clinician who knows the problem from having seen many cases, who knows the

countries, the cultures, the areas where the problem takes place, and who knows the particular syndromes. It is by equipping the clinician with the work of the technicians, by having clinical judgement and technological and scientific evidence side by side, that we are likely to make the most progress" (Deutsch, 1980: 19).

In a survey of the most commonly cited political scientists, Deutsch's name ranked eighth between 1945 and 1960, fourth from 1960 to 1970, and tenth between 1970 and 1976. After 1980 his name still appears on the list of the three dozen most-cited political scientists. Between 1970 and 1979, his name was cited 1870 times. Between 1996 and 1999 he has been cited 60–90 times per year. Most of them concern the following three books: *Political Community at the International Level* (Deutsch, 1954), *Nationalism and Social Communication* (Deutsch, 1953), and *The Nerves of Government* (Deutsch, 1963).

It is interesting to note that Deutsch's two most successful books were conceived and written simultaneously, and were published only one year apart, in 1953 and 1954. Undoubtedly the fundamental ideas for these two works were already present in his mind when he escaped from the totalitarian regime. *The Nerves of Government* does not appear to be Deutsch's most popular book. It has been cited less frequently than the other two and, when it has been cited, the citations were in many cases perfunctory, the content not being commented upon. This book was published in 1963, at a time when Deutsch had already been at the front of the stage. Such perfunctory practice has been described by Robert Merton in his *Sociology of Science*.

The high 'mortality' rate of articles in political science, particularly in the field of international relations, is a well-known phenomenon. In several other disciplines, the life expectancy rate of articles is longer. Karl Deutsch published as author, co-author, editor, and co-editor a total of 16 books. In addition, between 1943 and 1980, he published 240 articles in academic journals and chapters in collective works, approximately half of them in collaboration with other scholars. According to the SSCI, which covers as sources only journals, excluding books, some of these 240 contributions were still being cited during the 1990–95 quinquennium. The content of some of Deutsch's articles has become chapters of his books, particularly in *The Nerves of Government*. Other articles, among the best *Towards an Inventory of Basic Trends and Patterns in Comparative and International Politics* (1960) and *Social Mobilization and Political Development* (1961), have become standard information and common knowledge, and, consequently, do not need to be cited. What had been previously innovative has been superseded by more recent studies based on more precise data, analyzed with more refined methods. When writing about microbes, Pasteur need not be cited. When using social indicators, Karl Deutsch, Gabriel Almond, and Daniel Lerner no longer need to be cited. The concept of social mobilization already belongs to the common patrimony of the social sciences.

In the late 1960s and 1970s, Yale's Department of Political Science was one of the most attractive for graduate students, many of whom converged in Deutsch's seminars. One of these seminars inspired the conception of an important book entitled *World Handbook of Political and Social Indicators*, which users of quantitative data had named 'the Yale bible,' by Bruce M. Russett in collaboration with H. R. Alker, K. Deutsch, and H. Lasswell (Russett et al., 1964).

The following are among Deutsch's intellectual progeny: Hayward Alker, Jorge Dominguez, William Foltz, Michael Hudson, Peter Katzenstein, Manfred Kochen, Arendt Lijphart, Richard Merritt, Donald Puchala, Bruce Russett, Charles Taylor, these and others recognized as outstanding political scientists. They honored their mentor in a collective book *From National Development to Global Community*, edited by Richard L. Merritt and Bruce M. Russett (1981).

Stein Rokkan's tribute to Deutsch's intellectual leadership deserves to be cited at length: "Karl Deutsch taught my generation of social scientists to develop models and to test them against the data of history. My own early research was heavily influenced by Deutsch's ideas. My later work on dimensions of state formation and nation-building was directly inspired by Deutsch's path-breaking study *Nationalism and Social Communication*; he taught me to look out for the decisive characteristics of center-building networks and to study the functions of the printed media" (Rokkan, 1981: 70).

Deutsch was an interdisciplinary scholar, comfortable in many fields. In his latter days he became interested in 'the conditions favoring major advances in the social sciences.' In 1986, he co-edited a volume focusing on this subject in collaboration with Andrew Markovits and John Platt (Deutsch et al. 1986).

Those who encountered Deutsch were impressed by his encyclopedic erudition. In his book *Nationalism and Social Communication* (Deutsch, 1953) some 2000 authors are mentioned. His historical knowledge appears in a particular light in an article published in 1945 on 'Anti-semitic ideas in the Middle Ages,' based on the reading of very old literature (Deutsch, 1945). This article merits to be included in any modern anthology of historical sociology.

Soon after the publication of *The Limits to Growth* by the Club of Rome in 1972, a new trend appeared in political science, imitating economists, demographers, and ecologists. Karl Deutsch had alerted the specialists: "Computers can assist thought, but not replace it. Mathematical thought produced new ideas and techniques, such as game theory and coalition theory, scaling, latent structure analysis, matrix analysis and a succession of increasingly sophisticated models of social and political structures and processes ... Some of us have tried to deal with this by over-specialization, risking to fall into the old trap of learning more and more about less and less" (Deutsch, 1978). Nevertheless, in the 1980s, he was seduced by the siren of 'world modeling.' Under his directorship at the Wissenschaftzentrum in Berlin, modeling became a priority. The virtue of Deutsch's effort in Berlin was to combine mathematical modeling and solid empirical referent. Abstract mathematical modeling has attracted the attention of a limited number of political scientists outside the circle of the builders of reductionist models.

World trade was one of Deutsch's favored topics, which he discussed in several writings. He measured the ratio of foreign trade to national income for almost half a century. He presented an impressive amount of statistical data demonstrating a decline in international trade between 1913 and 1957 (Deutsch and Eckstein, 1961). But this period was marked by World War I, the Great Depression, protectionism, and autarchy practiced by several great powers in the 1930s, World War II, and the post-War reconstruction of the

devastated European countries and Japan. In the 1950s, only 6% of the American GNP depended on foreign trade. Deutsch did not explicitly predict a continuing decline of international transaction. More recent developments do not invalidate Deutsch's historical observation.

One of Deutsch's main ideas was the importance of intertwining economies for political integration. This thesis is still validated: half of France's foreign trade takes place with Germany, and reciprocally. Economic integration has preceded political integration of the European Union, even for Britain.

Since the publication of *Nationalism and Social Communication*, the nations of Western Europe, with the exception of Greece and Ireland, have experienced a rapid decline of nationalist tendencies, reversing a historical trend (Dogan, 1993), at a moment when, in contrast, many developing countries in Asia, the Middle East, and Africa have been mobilized by movements of national independence, developing an exacerbated ethnic nationalism. The concepts designed by Deutsch and other scholars – social communication, rapidity of assimilation, rate of mobilization, spread of national consciousness, population clusters, linguistic cleavages, center–periphery – have been transplanted from the continent to which they were first applied to many new nation-states, where Deutsch's theories are more relevant.

Karl Deutsch's professional career was one of exceptional brilliance. He was elected President of the American Political Science Association in 1969 and of the International Political Science Association (IPSA) in 1976, playing an important role in the development of the research committees within IPSA. In collaboration with Richard Merritt, as program director, he organized the IPSA World Congress in Moscow in 1979, building a bridge between East and West for political scientists. A member of the American Academy of Sciences, he lectured in many European universities. Deutsch received honorary doctorates from seven American and European universities and the Commander's Cross Order of Merit from the German Federal Republic.

See also: Comparative Studies: Method and Design; Nationalism: General; Quantification in the History of the Social Sciences; Rokkan, Stein (1921–79); State Formation, Theory of.

Bibliography

Deutsch, K.W., 1945. Anti-semitic ideas in the Middle Ages. Journal of History of Ideas 6, 239–251.

Deutsch, K.W., 1953. 1966 Nationalism and Social Communication: An Inquiry into the Foundations of Nationality, second ed. MIT Press, Cambridge, MA.

Deutsch, K.W., 1954. Political Community at the International Level, Problems of Definition and Measurement. Doubleday, Garden City, NY.

Deutsch, K.W., 1957. Political Community and the North Atlantic Area. Princeton University Press, Princeton, NJ (in collaboration with seven other scholars).

Deutsch, K.W., 1963. 1966 The Nerves of Government: Models of Political Communication and Control, second ed. Free Press, New York.

Deutsch, K.W., 1969. On methodological problems of quantitative research. In: Dogan, M., Rokkan, S. (Eds.), Quantitative Ecological Analysis in the Social Sciences. MIT Press, Cambridge, MA, pp. 19–39.

Deutsch, K.W., 1978. Major Changes in Political Science 1952–1977. Research paper. Internationales Institut für Vergleichende Gesellschaftsforschung, Berlin, 28.

Deutsch, K.W., 1980. On the utility of indicator systems. In: Taylor, C.L. (Ed.), Indicator Systems for Political Economic and Social Analysis. Oelgeschlager, Gunn, and Hain, Cambridge, MA, pp. 11–23.

Deutsch, K.W., Alker, H., Stoetzel, A. (Eds.), 1973. Mathematical Approaches to Politics. Jossey-Bass, San Francisco.

Deutsch, K.W., Eckstein, A., 1961. National industrialization and the declining share of the international economic sector 1890–1959. World Politics 13, 267–299.

Deutsch, K.W., Edinger, L., 1959. Germany Rejoins the Powers: Mass Opinion, Interest Groups and Elites in Contemporary German Foreign Policy. Stanford University Press, Stanford, CA.

Deutsch, K.W., Markovits, A.S., Platt, J., 1986. Advances in the Social Sciences 1900–1980. University Press of America, New York.

Deutsch, K.W., Platt, J., Senghaas, D., 1971. Conditions favoring major advances in the social sciences. Science 171, 450–459.

Dogan, M., 1993. Comparing the decline of nationalisms in Western Europe: the generational dynamic. International Social Science Journal 136, 177–198.

Dogan, M., Rokkan, S., 1969. Quantitative Ecological Analysis in the Social Sciences. MIT Press, Cambridge, MA.

Jodice, D.H., Taylor, C.L., Deutsch, K.W., 1980. Cumulation in Social Science, Data Archiving. A study of the Impact of the Two Handbooks of Political and Social Indicators. Wissenschaftzentrum, Berlin.

Merritt, L.R., Russett, B.M. (Eds.), 1981. From National Development to Global Community. Allen and Irwin, London.

Rokkan, S., 1981. Variations within Western Europe. In: Merritt, L.R., Russett, B.M. (Eds.), From National Development to Global Community. Allen and Irwin, London, pp. 115–144.

Russett, B.M., Alker, H.R., Deutsch, K., Lasswell, H., 1964. World Handbook of Political and Social Indicators. Yale University Press, New Haven, CT.

Taylor, C.L., 1968. Aggregate Data Analysis: Political and Social Indicators in Cross-national Research. Mouton, Paris.

Development and Employment: A Mutual Relationship

Stefano Scarpetta, Directorate for Employment, Labour and Social Affairs (ELS) (Organisation for Economic Co-operation and Development (OECD)), Geneva, Switzerland
Gaëlle Pierre, International Monetary Fund (IMF), Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The past three decades have witnessed major changes in the economic and social fabric of developing and emerging economies, bringing new opportunities but also challenges. Addressing the latter requires a complex set of factors that shape the investment climate and promote sustainable growth and job creation in the formal sector. Labor market, social policy, and institutions can not only facilitate the required reallocation of workers toward more rewarding and productive jobs but also provide adequate support to those affected by these changes. This requires improving the institutional fit of labor institutions with the reality of the different countries and devoting adequate resources to effective social protection programs. Strong economic development is key for more but also productive and rewarding jobs, and better jobs are key to ensure the benefits of growth are widespread and conducive to improvements in the well-being of the population.

Introduction

Job creation is often at the top of the development agenda, for policy makers as well as for social partners and civil society. This is because labor is more than a factor of production; it is an essential element of any development process, of well-being and social cohesion. Labor is the main asset of people; the type of work people hold is one of the most powerful determinants of well-being, as, over their lifetime, most people spend a substantial part of their time at work. Moreover, it is an essential factor of production for firms of all types and sizes. Even in high-tech manufacturing firms, workers bring new skills and competencies to enhance the organization of production or to improve the quality and variety of products. The way in which employers' demand for labor and households' supply of it interact is crucial in determining the conditions in which firms find the proper incentives for engaging in growth enhancing activities and workers participate in, and benefit from, these activities.

Labor markets have been profoundly affected by global developments in the past decades. The transition of many developing countries and former centrally planned economies to market-based systems and the greater integration of countries in the global economy have changed the rules of the game within which businesses, governments, and workers interact. The opening of domestic markets has exposed previously protected firms to greater competition but also to new, more productive technologies. The scale of production has often changed, from large oligopolies protected by trade barriers, to smaller, but often more technologically advanced, production units. Overall, firms are facing greater pressure to innovate in order to survive and expand.

These changes have created new economic opportunities, but also exposed firms and workers to new challenges. Greater integration in world markets have enhanced opportunities for entrepreneurs to launch new activities and for existing firms to expand, both contributing to a faster job creation in expanding sectors. But technological advances and expanding international interactions have also changed comparative advantages of individual countries and threatened firms and jobs in declining sectors. More generally, pressure has increased to make work arrangements more adaptable to enable firms to respond to changing economic and technological conditions. In many developing and emerging economies, labor market policy and social security systems are often underdeveloped, and many workers are increasingly exposed to the volatility of a more competitive, open but also more risky economic environment.

The nature of jobs is varied. As recently indicated (World Bank, World Development Report 2013), of the three billion people at work, more than half are employed and receive regular wages or salaries; the rest work in farming and small household enterprises, or in seasonal or casual day labor. And while there are about 210 million people unemployed worldwide, two billion working-age adults, the majority women, are neither working nor looking for work, even if a number of them are eager to have a job. Understanding the nature of work relations and how they are evolving as part of the development process is crucial to design adequate policies and institutions that promote strong but also sustainable and job-rich economic growth.

Fostering Job Creation and the Efficient Utilization of Labor

Recent Developments

Efficient utilization of labor is essential for sustainable long-term economic growth. This entails high levels of employment, enhancement of human capital, and labor mobility. A simple decomposition of the sources of economic growth over the past decades suggests that the combined effect of greater utilization of labor and enhancement in human capital have generally made a significant contribution to improvements in output per capita. Many countries that have seen improvements in their growth performance over the past decades have also seen improvement in the quality of the workforce and in some cases an increase in the employment rate (the share of the total working age population involved in productive activities).

The views expressed herein are those of the authors and should not be attributed to the their organisations of affiliation.

Even more importantly, efficient utilization of labor is critical for poverty reduction. Household living standards, especially amongst the poor, are closely tied to their income from work, and a labor market that leads to high labor utilization, especially in formal activities, is a key vehicle to reduce poverty and exclusion. Economic development has led to significant reductions in poverty. For example, an estimated 447 million workers in emerging G20 economies were extremely or moderately poor in 2013, a reduction by half since 1991 (World Bank, 2013). Yet, more remains to be done; working poverty and near poverty remain large and widespread challenges for these countries. Taking into account these near-poor workers, some 837 million workers, equal to just over half of the workforce in emerging G20 countries, were either poor or just above the poverty line in 2013 (Kapsos and Bourmpoula, 2013).

Some countries – especially in East and South Asia – have been successful in combining strong economic growth with a rising labor utilization and enhancement of human capital. However, many countries face weak labor market conditions, which have been compounded by the recent financial crisis. Poor labor market outcomes have been driven in many countries by private job creation not keeping pace with the increase in the working age population and/or the increase in female participation in the labor market. In a number of countries in the Middle East and North Africa, the state has been a leading and generous employer but as the fiscal space for continued expansion in the public sector is shrinking, 'queuing' for public sector jobs has become more prevalent, leading to informality, a devaluation of educational credentials, and forms of social exclusion and unrest in a number of them (see Nabli et al., 2008; Assaad, 2012). In the former Soviet block, after a period of practically full employment, unemployment rates reached two-digit levels in the early phases of the transition to a market economy. Many of those countries still face relatively high rates of joblessness and falling employment (see inter alia, Rutkowski and Scarpetta, 2005). In Latin America, many countries, including those that had sustained growth prior to the global financial crisis, have seen unemployment and inequality increase, even if recent reforms and effective transfers to the poor have contributed to reduce household income inequality (e.g., Brazil, Chile, and Mexico) (see Pagés, 2010; Pagés et al., 2009). While measured unemployment has remained relatively low in South Asia, the share of working poor is still around 40% on average in the region. Advanced economies have been hit by the global financial and economic crisis of 2008–09, which has led to a large increase in unemployment and underemployment in a number of economies, often concentrated on the most vulnerable groups, including the youth, the low skilled, and migrants (see OECD, *Employment Outlook*, 2013; see also OECD-ILO, 2014; OECD, 2010b).

These unsatisfactory outcomes are both due to poorly functioning labor markets, as well as failures of other markets to promote strong and sustainable, and thus job-rich, growth. High and persistent unemployment in some countries, a large informal sector or low participation in others partly reflect the inability of labor market to facilitate labor mobility and promote firms' investment, because of restrictions in wage or labor adjustments or low incentives for investment in human capital. At the same time, failures in other markets, by curbing firms' developments, also limit the process of job creation and wage improvements. Some deterioration in labor market conditions may also be a temporary side effect of structural reforms aimed at improving the investment climate. For example, macroeconomic reforms, trade liberalization, or large-scale privatization all require large reallocation of workers and changes in skills. Even in the best of all possible labor markets, such changes may lead to temporary increases in unemployment and loss of income, as workers need to change job, acquire new skills, and sometimes change location.

In practice, it is difficult to disentangle these different elements as labor markets interact closely with other markets and reforms impose pressure on all of them also affecting these interactions. Bearing these caveats in mind, the remainder of this section will shed some light on these linkages in turn.

How Adaptable Are Labor Markets in Developing Countries?

Different factors contribute to the adaptability of the labor market and thus its ability to contribute to promote investment, job creation, and growth. First, firms should be allowed to adjust wages and employment to changing demand conditions and to adopt new technologies. Second, market economies require a continuous reallocation of labor away from declining firms and industries toward expanding ones, and labor markets have a key role in facilitating this process and thus contributing to productivity gains and better prospects for workers (see, e.g., Bartelsman et al., 2013). Developing and emerging economies differ with respect to actual labor market adaptability, although lack of relevant information often makes it difficult to depict a comprehensive picture.

Evidence suggests a significant – and often growing – response of wages to changes in labor market conditions in some countries. At the same time, in many countries a substantial number of jobs are created and destroyed every year pointing to a continuous process of labor reallocation. In particular, structural reforms have sometimes changed the responsiveness of wages to labor market conditions. (For example, in Chile there was no evidence of the wage curve during the period of inward-led development (1957–73), but the curve emerged along with the opening of the economy and regulatory reforms (post 1974). See Berg and Contreras, 2002.) However, it should be stressed that the flexibility of real wages in many developing countries has often been achieved in the past by the lack of full indexation of nominal wages during periods of high inflation in downturns. In a low inflation environment and given the natural tendency for workers to resist nominal wage reductions, the ability of wages to accommodate demand fluctuations may be reduced. (For example, wage elasticity in Argentina in the 1980s – a period of high inflation – was about 10 times higher than wage elasticity in the 1990s – a period of lower inflation – and employment elasticity in the 1990s was twice as large as employment elasticity in the 1980s. See IADB, 2003.) In any event, wage adjustments cannot be the only factor to accommodate changes in labor demand, especially in countries undergoing major structural changes, where new firms replace obsolete ones, new sectors emerge, and new technologies have to be adopted to maintain competitiveness.

Most low- and middle-income countries have experienced significant shifts in employment, away from low-productive agricultural activities to manufacturing and, especially, new service activities (see World Bank, World Development Report

2013 for an up-to-date review). This secular trend away from agriculture and toward manufacturing and services hides significant differences in the ability to match labor demand and supply. In countries for which data are available, gross rates of job creation and destruction each range between 5 and 20%, adding up to a total job turnover of up to 40% (see Haltiwanger et al., 2014). A significant fraction of this job turnover (often 30–50%) is due to the entry and exit of firms, an important factor for output and productivity growth. (It should also be noted that in all countries, worker turnover is even larger than job turnover, because workers not only move directly from one job to another, but also between employment and unemployment and inactivity as a result of their own personal decisions. See OECD, *Employment Outlook*, 2012.)

This sizable process of creation and destruction of jobs in most countries does not necessarily mean that the reallocation of labor has been efficient. The latter depends on whether job flows lead to the allocation of workers to most productive use and thus are associated with productivity growth and wage increases.

In particular, job destruction may be inefficient if, for example, capital markets constrain the ability of firms to cope with negative shocks and retain valuable workers. Credit constraints are a severe problem for small firms in many developing countries and it is thus not surprising that these firms tend to have higher levels of job reallocation. (See Caballero et al. (2004) for a discussion of the ability of different firms to adapt to negative shocks.) Moreover, certain labor market policies may inflate job flows, without necessarily leading to better outcomes.

In addition, labor reallocation may go in the wrong direction, especially if dismissed workers cannot properly search for jobs because they cannot afford to remain without income. For example, in Latin American countries dismissed workers often go to the informal sector because the lack of unemployment benefits prevent them for properly searching for another (formal job) (see IADB, 2003; Pagés et al., 2009).

Moreover, a major consequence of labor reallocation is large flows of workers across regions. Traditional models have emphasized the role of rural–urban migration as a key engine of development (see Fields, 2004 for a review). But whether these large domestic flows of workers smooth labor market pressures and contribute to a better allocation of resources depend on the characteristics of those who migrate and on policies that may influence the decision at the individual or household levels. For example, large flows of often unskilled workers are likely to increase unemployment in receiving urban areas if labor demand does not respond quickly. Immigrants in urban areas may be low skilled compared to the urban labor force, but they are generally young with relatively better education than the local workforce. Hence, emigration may reduce the potential for rural development if remittances toward rural areas do not compensate for the loss of human capital.

International migration may also be considered as a factor in promoting efficient allocation of resources, but, at the same time, may deprive a country of its main human capital assets. Recent research also point to the benefits of emigration, in particular, through the potentials offered by remittances and the role that returning emigrants may play for local development. Remittances have become an important aspect of global development finance (see World Bank, Migration and Remittances Briefs (various issues)). They are often used to support family members and can stimulate domestic demand as well local investment. Many relatively high-skilled emigrants also return to their home countries and establish their own business. Several East Asian economies have benefited from the return of migrants to the home country.

The Policy Challenges of Promoting More and Better Jobs for Development

Although the idea that a well-functioning investment climate is essential to generate higher demand for labor and for enabling wages to rise is well accepted, some of the key reforms that are necessary to promote the investment climate are often perceived as harmful by workers. Evidence points to significant long-term benefits from a comprehensive reform of the investment climate, in terms of both access to better jobs and reduced unemployment and underemployment. Most of these reforms, however, imply major changes which have short-term costs, especially for workers in firms previously protected, for example, by trade barriers, state support or lack of competition, and for workers with low skills or in declining areas. Often, these costs are exacerbated by the fact that reforms are implemented in the aftermath of a major economic crisis when the choice of the reform package or its implementation occurs when the economy is already undergoing painful adjustments. (This is indeed the case of a number of reforms undertaken during the recent financial and economic crisis in a number of advanced economies – especially in Southern Europe. For a review, see OECD, *Employment Outlook*, 2014.)

Adequately designed reforms of the investment climate – be it via enhancements in product market competition, greater transparency, or more secure property rights – have all the potentials for improving workers' welfare over time. Successful reforms are expected to bring about higher wages and better working conditions, as well as higher employment and lower levels of unemployment and informality in the long run (see Bourguignon and Goh, 2003; De Ferranti et al., 2000; Gill et al., 2002). Empirical studies that have linked different structural reforms to labor market outcomes seem to confirm these points, although they also strongly underline the importance of a comprehensive and well tailored approach that is consistent with underlying economic conditions in each country. For example, countries which have opened their economies to greater foreign trade have enjoyed higher economic growth and higher wages compared to those which have persisted in the inward-oriented strategy (see amongst others; Dollar and Kraay, 2001; Rodrik, 1997; Rama, 2003).

Trade liberalization is an important reform to improve the competitiveness of a country, which can lead to better labor market outcomes. Greater exposure to foreign competition strengthens incentives for firms to invest in new technologies and in enhancement of the human capital and skills of their workforce. Moreover, within each individual country, sectors exposed to stronger competitive pressure tend to have higher labor productivity and pay higher wages than those concentrated on the internal markets, even for workers of similar characteristics. Similarly, there is evidence that foreign

owned firms in developing countries tend to pay premium wages and provide better working conditions, which reflect their better productivity performance associated with foreign ownership and especially easier access to foreign markets and technologies (see Brown et al., 2003).

Despite often being singled out as the main cause behind greater economic volatility and greater income and job security for workers, trade liberalization does not appear to increase wage disparity in the long run (see, e.g., OECD, 2011). Open economies tend to be more exposed to fluctuations in global demand and to changes in relative prices. But in the long run, there is little evidence that trade liberalization has led per se to an increase in the year-to-year fluctuations in earnings, nor are wages more volatile in sectors more exposed to foreign competition – a result which applies to a wide range of countries in Asia and Latin America (Bourguignon and Goh, 2003; De Ferranti et al., 2000; Pagés, 2010). At the same time, there is no strong evidence that, by fostering competition amongst domestic firms, countries have increased the elasticity of labor demand with respect to wages: no such effect was found in India, Turkey, and a sample of Latin American countries (Epifani, 2003; Fajnzylber and Maloney, 2000).

It must be recognized that many of the reforms aimed at improving the investment climate have short-term and long-term impacts on the structure of the labor market. They may well imply changes in the nature of jobs, with a decline in stable jobs in sectors sheltered from competition and lower job tenure. They are often associated with short-term adjustment costs – in terms of employment losses and major changes in relative wages. These costs can be large and protracted if the reallocation of resources associated with most structural reforms is sluggish, because relative prices do not adjust, because labor adjustment costs are high, or because old jobs are destroyed before new jobs are created. For example, empirical studies suggest that wages may well decline in the wake of trade liberalization, before they begin to rise again, usually at a higher pace (see, e.g., Rama, 2003). Moreover, structural reforms prompt large reallocation of resources, including labor. Job destruction and job creation might not be fully synchronized thereby giving rise to some temporary increase in unemployment, or the expansion of informal activities especially in countries where workers cannot afford to remain without a job.

The adjustment costs associated with most structural reforms tend to be borne by the most vulnerable groups in the labor market. These groups include individuals with low levels of education, or with a specialization which is no longer demanded in the market, and those with little work experience or in geographical locations dominated by backward activities. In developing and industrial countries alike, low-skilled workers, youth, and often prime-age women with little work experience tend to face a high incidence of unemployment, informality, or low pay. (See OECD, *Employment Outlook*, 2013, for a review of the impact of the global financial crisis of 2008–09 on different sociodemographic groups.) These unfavorable conditions for vulnerable groups tend to deteriorate even further during the initial phases of large-scale structural reforms. Workers with little experience tend to be the first to be dismissed in case of downsizing and the low skilled face the greatest difficulties if they lose their jobs. These difficulties can be particularly severe if the adjustment process follows a crisis that has plunged the economy into a recession or if the renewal of growth takes longer than originally expected. For all these reasons it is vital to adopt a comprehensive reform strategy that includes adequate labor market and social protection policies to support workers and households affected by the structural transformations.

Labor Market Institutions: Governing Worker–Firm Relations

Policy makers intervene in the worker–firm relations on two main fronts. They set regulations for working conditions and the labor contract between workers and employers, and they fix wage floors and set the legal framework in which social partners operate. These interventions are theoretically justified by the (perceived or effective) inability of laissez-faire conditions to deliver efficient and equitable outcomes. Efficiency arguments stress information problems and a need to promote matching between labor demand and supply. There are also equity arguments if there is unequal bargaining power between employers and workers, potential discrimination against vulnerable groups, or incomplete or imperfect insurance of workers against risks. The policy priorities for improving labor market outcomes vary greatly across countries, depending on their level of development, institutional settings, and social preferences. (See the World Development Report 2013 of the World Bank for a review of the typology of job challenges.)

Intervening in the Wage-Setting Process

The success of fostering opportunities for firms to invest productively largely depends on their ability to remunerate factors of production, including labor, according to their productivity and to adjust input prices to accommodate changes in demand. Properly compensating workers also fosters their effort in the production process and strengthens their incentives to invest in human capital. Governments intervene in the wage-setting process by establishing rules for wage bargaining and for industrial relations.

Setting Wage Floors

The main objective of setting wage floors is to promote decent jobs and reduce poverty among workers. Wage floors also tend to control market power that companies may have over individual workers who lack information, bargaining power, or the ability to seek better-paid jobs. This is particularly the case in monocultural areas, where a single company may account for the bulk of labor demand. The effectiveness of wage floors in many low- and middle-income countries is, however, questionable. Minimum wages tend to be fairly high compared with market wages in a number of these countries, and any further increase would shift the entire wage distribution upward. This concentrates the disemployment effects among those they intended to support – young, low skilled, and female workers. When enforcement is weak, a hike in the minimum wage stimulates more underreporting of wages or further incentives for firms and jobs to remain in the informal economy. However, a number of emerging economies have recently raised the minimum wage relative to the average wage (often from very low levels) as a way to fight working poverty and boost

household income, without sizeable disemployment effects (see OECD-ILO-WBG 2014 – *G20 labour markets: outlook, key challenges and policy responses*).

Setting Workplace Regulations

Promoting better health and safety conditions in firms, regulating working time, and encouraging paid annual leave have been major achievements in all societies, and well-designed regulations can help to achieve this goal. The beneficial impact of such regulations may, however, be limited if they, or other features of labor regulation, have the effect of keeping firms and workers in the informal economy, where workers usually lack any statutory protection. Stronger enforcement can help in some cases. However, when regulations are out of step with local realities, there will be trade-offs between providing a high level of protection to workers in regulated employment and reaching a broader group of workers.

Balancing Employment Stability with Firms' Need to Adjust the Workforce

Probably the most controversial government intervention in the labor market is setting rules for hiring and firing workers. By affecting the cost of workforce reorganization, employment protection legislation strongly influences the cost of doing business, but especially the incentives and opportunities for firms to exploit new technologies and expand. The protection offered to regular workers and the conditions for temporary employment vary considerably across and within regions. Countries in Latin America, Eastern Europe, and Central Asia tend to offer the most employment protection for regular workers. By contrast, 'common law' industrial countries and East Asian economies have the lowest statutory protection. Beyond ensuring full respect for the core labor standards, setting adequate regulations on hiring and firing procedures can benefit both workers and firms. For example, by reinforcing job security, they can enhance productivity, as workers will be more willing to cooperate with employers (see Martin and Scarpetta, 2012 for a review of the literature, as well as Addison and Teixeira, 2001; Heckman and Pagés, 2004; Betcherman, 2012). To the extent that job protection leads to long-lasting work relationships, it may encourage employers to provide training. A better skilled workforce may also increase internal flexibility and thus lead to a better functioning of production activity. Job protection may also be a way to internalize the social costs of dismissals by moving the social burden of reallocating a worker to another job closer to the firm's profitability criteria.

Given the high level of protection granted to workers in the formal sector in many developing countries, it is not surprising that managers often consider such regulations as an obstacle for the expansion of their firm. In fact, overly stringent regulations – those at odds with the international experience and with the stage of development of the economy – affect different aspects of firms' performance, including spending on innovation, the entry of new firms, their average size, and the incidence of informality.

Strict regulations in industrial countries, where enforcement is high, tend to promote job stability for prime-age males, but tend to reduce job opportunities and lengthen unemployment spells for youths, women lacking work experience, and those with low skills. Firms may become reluctant to hire new workers from these groups for fear of incurring expensive dismissal costs in the future.

When enforcement is weak, as it is in many low- and middle-income countries, stringent regulations do not reduce the size of labor reallocation, but they do change its nature and reduce its effectiveness. Women, youths, and the unskilled – facing greater difficulties in obtaining a job in the formal sector – are more frequently unemployed or engaged in informal activities.

Active Labor Market Policies: Promoting Workers' Adaptability and Mobility

Many countries face the challenge to equip their future labor force with skills and competencies that are relevant for the labor market. This involves important reforms in the education and training systems to improve the coverage and quality of education and enhance learning outcomes. Such reforms should be started now and will bear fruits for future generations of workers.

At the same time, there are a number of measures that governments can use to improve the ability and willingness of current workers to move to more productive and rewarding jobs. Such policies include supporting job intermediation and training. The effectiveness of these programs has been challenged especially in countries with limited capacity, but when properly targeted they can be a complementary strategy to skill enhancements and income support.

Job search assistance is aimed at promoting transparency and information in the labor market so as to facilitate job matching. In this respect, it offers a service to both employers who have otherwise to go through a lengthy and costly selection process and workers who may lack the information on suitable jobs in the market. To be effective, job search assistance requires close links between employers, the intermediation office – generally public employment services and, increasingly in industrial countries, private agencies – and the job seekers. (See Immervoll and Scarpetta (2012) for a review of the effectiveness of active labor market policies in developing and emerging economies and Card et al. (2010) for a meta-analysis of the empirical evidence.)

Labor market training includes publicly supported programs, usually through either direct provision (in public training institutes) or financial support (funding training costs and/or subsidizing trainees). In many countries, governments are moving away from the role of direct provision of training and focusing more on addressing market failures in information and financing, while leaving more of the delivery to private providers. Job skills training have been shown to be quite successful in improving the chances of beneficiaries in securing employment (see, e.g., OECD, 2010a).

Training programs for youths, even when well-targeted, tend to have a poor track record. Interventions typically include some combination of on-the-job training, vocational (classroom) training, job readiness training, and/or internships. However, earlier interventions at the schooling stage are likely to be more effective than trying to remedy education failures (Betcherman et al., 2003) The experience of some Latin American countries offers some interesting insights. The

'Jovenes' programs in Argentina, Chile, Peru, and Uruguay are targeted at disadvantaged youth – combining training and work experience with other services including psychological development and vocational assessment (see OECD, 2010 for a review and OECD-ILO, 2014).

Reinforcing Self-Insurance among Formal Workers

In most low- and middle-income countries, mandatory severance pay provisions are the prevalent form of insurance against unemployment for workers in the formal sector. Generally easy to administer, the provisions exchange resources, in the event of unemployment, for an 'insurance premium.' Whether the premium translates into lower wages affects labor costs for firms and their incentive for hiring. But the schemes do offer a limited pooling of unemployment risk because they are firm-specific and because the premium generally evolves with tenure and not with the risk of unemployment. (De Ferranti et al. (2000) argue that, when Latin American countries had little exposure to foreign competition, the effective pooling of unemployment risk offered by severance pays was spread over a greater population, as consumers often times subsidize potentially bankrupt firms through higher prices. However, this possibility has declined rapidly as countries have embarked on trade liberalization and reforms aimed at fostering domestic competition.)

Severance pay provisions suffer from noncompliance in many countries, increasing workers' resistance to leaving a job. The risk of nonpayments tends to increase when financial resources are lacking because the firm is experiencing difficulties, or simply not be available if the firm goes bankrupt. Noncompliance also creates a burden on labor courts and government budgets. To tackle these shortcomings, some countries have introduced prefunding or reduced the generosity of payments to bring them more in line with international experience. Colombia moved toward a funded system under individual savings accounts in 1990, and Chile introduced a social insurance component to its system in 2002.

Workfare Programs to Combine Income Support with Improvement in Local Conditions for Investment

Workfare programs have traditionally been the most widely used form of intervention to support poor households and promote local infrastructure development. They generally transfer income to poor households, by providing unskilled manual workers with short-term employment on projects such as road construction and maintenance, irrigation infrastructure, reforestation, and soil conservation. In most developing countries, workfare programs have been used to smooth consumption and keep poor people in contact with the labor market (see Grosh et al., 2008 for a review). Well-designed workfare programs build much-needed infrastructure and thus minimize the trade-off between public spending on income transfers and on development. The National Rural Employment Guarantee Scheme, introduced in India in 2005, which provided 100 days of guaranteed paid employment to every household whose adult members volunteer to do unskilled manual work, while expensive has contributed to the local development of rural communities (see Dutta et al., 2012 for an assessment). Workfare programs have also helped many small private contractors to emerge and grow.

Conditional Cash Transfers to Fight Poverty while Supporting Human Capital and Health

Conditional cash transfers are another policy tool used to combine income support with local development. They belong to a family of transfer programs that combine close targeting with capital accumulation by making income support conditional on either basic needs triggers, such as utility offset payments (in some transition economies), or behavioral changes, such as the continued school enrollment of children or attendance at health clinics. They typically address chronic poverty rather than idiosyncratic risks associated with job loss. Their focus on human capital formation makes them suitable to address poverty and local development at the same time. In Mexico (Progresa, which reached 2.3 million families in 1999), Brazil (Bolsa Familia and PETI), and Jamaica (PATH), conditional cash transfers are largely used to promote health and human capital of children. The programs also tend to be better targeted than general subsidies by the use of proxy means testing and geographical targeting. They are highly transparent about who receives the transfers, and the level of benefits and the number of beneficiaries are easily adjusted in times of crisis. As with any transfer program, conditional cash transfers have problems, especially when the increased demand for services is not met by increased supply (schools or clinics) or when the targeting is not sufficiently resistant to outside influence.

Concluding Remarks

Over the past three decades, developing and emerging economies have undergone major changes in their economic and social systems; in many countries, economic growth has strengthened with significant effect on poverty reductions and improvements in the well-being of populations. These major changes have also been reflected in the labor markets; jobs have been transformed, skill demand has evolved dramatically, and more generally labor markets have been exposed to greater pressure to modernize and increase their adaptability.

No doubt, better labor market outcomes depend on a complex set of factors that shape the investment climate in each country and promote sustainable growth and job creation in the formal sector. Adequate labor market and social policies and institutions can play an important role to foster better outcomes for workers and their families. While many low- and middle-income countries have made significant progress in setting rules governing working conditions and the employer–worker relationships, the enforcement of these provisions remains often weak and uneven across sectors and firms. In this context, many firms opt out of the formal sector and remain uncovered to maintain profitability, facing great uncertainty in their labor relations and constraints in their growth and job creation potentials. Improving the institutional fit of labor regulations with the economic reality of the different countries – and in line with the international experience – is one of the main challenges of labor reforms in developing and transition countries.

Reforms of labor regulations will face large resistance and may not succeed if they are not accompanied by improvements in social protection mechanisms that cushion adjustment costs for affected workers. In many developing countries, this task is currently hampered by a narrow tax base. Most social protection schemes only cover formal sector workers – that is to say the non-poor – and do not offer adequate protection against job losses even amongst those with a formal job. Rural and informal workers are often under the threat of unexpected illness, job loss, or poverty in old age. There are clear opportunities for improving the insurance component in income support schemes for formal workers and the pooling of risks across individuals. And innovative programs can also reach out to poor and informal workers who cannot be covered by broader insurance schemes. For example, in a number of countries, some new forms of income support and job creation have tried to shelter poor workers from major income losses by providing targeted cash transfers and workfare programs. While greater resources and better targeting of social protection interventions toward the neediest people are warranted to promote labor market adaptability, a better prospect for the poor requires a comprehensive intervention in the investment climate that promotes growth and job creation.

See also: Culture and Economic Development; Development Theory in Geography; Development: Social-Anthropological Aspects; Global Agenda for Social Work and Social Development; Internal Migration: Developing Countries; Labor Market Institutions and Labor Market Outcomes in Transition Economies; Migration and Development; Mismatch on the Labor Market; Planning Issues and Sustainable Development; Sustainable Development and Social Work; Sustainable Development: An Economic Perspective.

Bibliography

Addison, J., Teixeira, P., 2001. The Economics of Employment Protection. Institute for the Study of Labour Discussion Paper 381, Bonn.

Assaad, R., 2012. The MENA paradox: higher education but lower job quality. In: Moving Jobs to the Center Stage. BMZ (Bundesministerium fuer Wirtschaftliche Zussamenarbeit), Berlin Workshop Series, Berlin.

Bartelsman, E., Haltiwanger, J.S., Scarpetta, S., 2013. Cross-country differences in productivity: the role of allocation and selection. American Economic Review 103 (1), 305–334.

Berg, J., Contreras, H., 2002. Political-Economic Regime and the Wage Curve: Evidence from Chile, 1957–96. CEPA Working Paper 2002-10.

Besley, T., Burgess, S., 2004. Can labour regulation hinder economic performance? Evidence from India. Quarterly Journal of Economics 119 (1), 91–134.

Betcherman, G., Olivas, K., Dar, A., 2003. Impacts of Active Labour Market Programs: New Evidence from Evaluations with Particular Attention to Developing and Transition Countries. Social Protection Network Paper Series 402. World Bank, Washington, DC.

Betcherman, G., 2012. Labor Market Institutions: A Review of the Literature. Background Paper for the World Bank, World Development Report 2013, Washington, DC.

Bourguignon, F., Goh, C., 2003. Trade and Labour Market Vulnerability in Korea, Indonesia, and Thailand: Evaluating Vulnerability from Repeated Cross-Sections. World Bank, Washington, DC.

Brown, D., Deardorff, A.V., Stern, R.M., 2003. The Effects of Multinational Production on Wages and Working Conditions in Developing Countries. National Bureau of Economic Research Working Paper 9669, Cambridge, MA.

Caballero, R., Engel, E., Micco, A., 2004. Microeconomic Flexibility in Latin America. National Bureau of Economic Research Working Paper 10398, Cambridge, MA.

Card, D., Kluve, J., Weber, A., 2010. Active labour market policy evaluations: a meta-analysis. Economic Journal 120 (11), 452–477.

De Ferranti, D., Perry, Guillermo E., Gill, Indermit S., Guasch, L., Schady, N., Maloney, W., Sanchez-Paramo, C., 2003. Closing the Gap in Education and Technology. World Bank, Washington, DC.

De Ferranti, D., Perry, G.E., Gill, I.S., Serven, L., 2000. Securing Our Future in a Global Economy. World Bank, Washington, DC.

Dutta, P., Murgai, R., Ravallion, M., van de Walle, D., 2012. Does India's Employment Guarantee Scheme Guarantee Employment? Policy Research Working Paper Series 6003 The World Bank.

Dollar, D., Kraay, A., 2001. Trade, growth, and poverty. Finance and Development 38 (3), 16–19.

Epifani, P., 2003. Trade Liberalization, Firm Performance and Labour Market Outcomes in the Developing World. What Can We Learn from Micro-Level Data? World Bank Policy Research Working Paper 3063, Washington, DC.

Fajnzylber, P., Maloney, W., 2000. Labour Demand and Trade Reform in Latin America. World Bank Policy Research Working Paper 2491, Washington, DC.

Fields, G.S., December 2004. Dualism in the labor market: a perspective on the Lewis model after half a century. Manchester School, University of Manchester 72 (6), 724–735.

Freeman, R., 2009. "Labor regulations, unions, and social protection in developing countries: market distortions or efficient institutions? In: Rodrik, D., Rosenzweig, M. (Eds.), Handbook of Development Economics, vol. 5. Elsevier, Amsterdam, pp. 4657–4702.

Gill, I.S., Maloney, W., Sanchez-Paramo, C., 2002. Trade Liberalization and Labour Reform in Latin America and the Caribbean in the 1990s. World Bank, Washington, DC.

Grosh, M., del Ninno, C., Tesliuc, E., Ouerghi, A., 2008. For Protection and Promotion: The Design and Implementation of Effective Safety Nets. The World Bank, Washington, DC, p. 587.

Haltiwanger, J., Scarpetta, S.S., Schweiger, H., 2014. Cross country differences in job reallocation: the role of industry, firm size and regulations. Labour Economics, Elsevier 26 (C), 11–25.

Heckman, J., Pagés, C., 2004. Law and Employment: Lessons from Latin America and the Caribbean. National Bureau of Economic Research, Cambridge, MA.

ILO, OECD, World Bank, September 2014. G20 Labour Markets: Outlook, Key Challenges and Policy Responses. Melbourne.

Inter-American Development Bank, 2003. Good Jobs Wanted: Labour Market in Latin America. Johns Hopkins University Press, Washington, DC.

Immervoll, H., Scarpetta, S., 2012. Activation and employment support policies in OECD countries. An overview of current approaches. IZA Journal of Labor Policy 1, 9.

Kapsos, S., Bourmpoula, E., 2013. Employment and Economic Class in the Developing World. ILO Research Paper No. 6, Geneva.

Martin, J., Scarpetta, S., June 2012. Setting it right: employment protection, labour reallocation and productivity. De Economist, Springer 160 (2), 89–116.

Nabli, M.K., Silva-Jáuregui, C., Faruk Aysan, A., 2008. Authoritarianism, Credibility of Reforms, and Private Sector Development in the Middle East and North Africa. Working Paper Series 443. Economic Research Forum, Cairo.

OECD, 2010a. The Employment Outlook – 2012. OECD, Paris.

OECD, 2010b. Off to a Good Start: Jobs for Youth. OECD, Paris.

OECD, 2011. Divided We Stand: Why Inequality Kepps Rising. OECD, Paris.

OECD, 2012. The Employment Outlook – 2012. OECD, Paris.

OECD, 2013. The Employment Outlook – 2013. OECD, Paris.

OECD, 2014. The Employment Outlook – 2014. OECD, Paris.

OECD, ILO, September 2014. Promoting Better Labour Market Outcomes for Youth: A Report for the G20 Meeting of Labour and Employment Ministers. Melbourne.

Pagés, C., Scarpetta, S., Pierre, G., 2009. Job Creation in Latin America and the Caribbean: Recent Trends and Policy Challenges. Palgrave MacMillan and the World Bank, Washington DC.

Pagés, C. (Ed.), 2010. The Age of Productivity: Transforming Economies from the Bottom up. Palgrave Macmillan, New York.

Rama, M., 2003. Globalization and Workers in Developing Countries. World Bank Policy Research Working Paper Series 2958, Washington, DC.

Rodrik, D., 1997. Has Globalization Gone Too Far? Institute for International Economics, Washington, DC.

Rutkowski, J., Scarpetta, S. (Eds.), 2005. Enhancing Job Opportunities: Eastern Europe and the Soviet Union. World Bank, Washington, DC.

World Bank, 1995. World Development Report 1995: Workers in an Integrating World. Oxford University Press, New York.

World Bank, 2013. World Development Report 2013: Jobs. Oxford University Press, New York.

World Bank. Various Issues, Migration and Remittances Briefs. World Bank, Washington, DC.

World Bank, 2004. Minimum Wages in Latin America and the Caribbean: The Impact on Employment, Inequality and Poverty. World Bank Office of the Chief Economist for LAC, Washington, DC.

Development and the State

Michelle Williams, University of the Witwatersrand, Johannesburg, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The state has long played a vital role in countries' development trajectories. Institutionally oriented economic historians (such as Polanyi and Gerschenkron), development economists (such as Hirschman, Amsden, and Sen), and development sociologists (such as Evans) have assumed development required active state involvement in constructing new institutions to facilitate trade and investment, and mobilizing resources in ways that accelerate the growth of new productive capacities. The record of East Asian developmental states from the 1960s through the 1980s vindicated this position. Neoliberal theorists saw the state as extracting rents in a self-serving manner, giving entrepreneurs the wrong incentives and impeding development. Predatory states were the empirical referents for their theories. Changes in the global economy and contemporary criticisms of development as insufficiently oriented toward the well-being of ordinary citizens, neglecting the importance of democracy, and failing to protect the natural environment present new challenges to states that are trying to be 'developmental.'

Over the course of the twentieth century, states' roles have broadened from the classical concern of making war and enforcing internal order to include creating the conditions for 'development.' The term 'development' became synonymous with the process of speeding up socioeconomic change, especially in the industrial sector, and registering macroeconomic results, particularly for countries in Africa, Asia, and Latin America. In other words, for much of the twentieth century, development meant the transformation of agriculture, the growth of industry, and a host of accompanying social, political, and economic changes. The role of the state in developmental success has been vital. However, not all states have played a developmental role, leading to the question, 'When and how does state behavior facilitate or impede development?' This question becomes even more acute today as the conditions faced by developmental states have undergone important changes in the last 20 years, making the tasks for states seeking development even more complex.

Historical Perspectives on the State's Role in Economic Transformation

Calling the poor nations of Asia, Africa, and Latin America 'developing countries' implies that they are undergoing economic and social transformations that mimic the experiences of the industrialized countries of the North. Understanding the role of the state in earlier periods of industrialization is, therefore, important to thinking about development and the state today.

Three streams of thinking are worth highlighting. First, there is the perspective of economic historians, such as Douglass North (1981), who see the provision of predictable norms and laws, particularly those that define and protect property rights, as the state's primary contribution to development.

The second view sees states as actively constructing new institutions to facilitate trade and investment and mobilizing resources in ways that accelerate the growth of new productive capacities. This view is epitomized in Karl Polanyi's (1957 [1944]: p. 140) classic statement: "The road to the free market [in England] was opened and kept open by an enormous increase in continuous, centrally organized and controlled interventionism." Alexander Gerschenkron (1962), looking at 'late developers' in Europe, reinforced and extended Polanyi's analysis by suggesting that countries like Germany and Russia faced a more difficult economic problem than England. They had to 'catch up.' By the time they began to industrialize, existing production technology required a scale that exceeded their societal capacities. Lacking both individual capitalists able to assume risks at the scale required by modern technology and private institutions able to spread risk, the state had to serve as investment banker or bear the risks itself, bringing together the necessary funds and encouraging their application in productive activities.

The third strand of thinking emphasizes the centrality of struggles over conflicting interests to the dynamics of development, the vital role of exploitation, immiseration and repression in the functioning of advanced capitalist economies, and the fundamental transformation of economic and political institutions globally. Within this Marxist critical political economy perspective, dependency theorists argued that late developers face historical disadvantages and unequal relations among nations within the world capitalist system (e.g., Frank, 1967; Dos Santos, 1970; Wallerstein, 1974; Cardoso and Faletto, 1979). Building on this framework and integrating it with elements of the institutionalist view above, scholars of 'dependent development' went beyond the structural determinism that negated the possibility of development in the Third World to suggest that states have a pivotal role as an agent of social control through three primary tasks: (1) it ensures a quiescent working class to facilitate industrialization; (2) it enlists multinational corporations to play a driving role in industrialization; and (3) it develops entrepreneurial initiative (Evans, 1979). (The term Third World emerged during the Cold War to distinguish between the First World (e.g., the capitalist West), Second World (e.g., communist East), and Third World (e.g., developing countries). Today, the term

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10101-1

Global South is used as it refers to a country's geopolitical position in the global capitalist economy, and not its geographic location. You can have countries geographically in the northern hemisphere, but part of the Global South (e.g., India), and countries geographically in the southern hemisphere as part of the Global North (e.g., Australia). I will use the corresponding term according to the historical period being discussed.) In short, states play the crucial role in coordinating and promoting national economic development.

Analyzing the State's Role in Third World Development

Perspectives, like those of Polanyi, Gerschenkron, and Evans in which development was not the spontaneous result of the activities of private economic actors but had to be explicitly fostered by the state, resonated in the Third World. In Asia, Africa, and Latin America, the gap between local capacities and the institutional scale required to be competitive with firms in the industrialized North was much bigger than the gap Russia and Germany had faced a century earlier. Local elites were too often wedded to maintaining agrarian structures that preserved their privilege at the expense of national productivity. The fact that the United States and Europe were just emerging from a period in which substantial increases in state intervention, first to combat the Great Depression and then to mobilize for World War II, made this perspective even more plausible.

Development Planning, Import Substitution, and Entrepreneurial Initiative

The conviction that Third World countries could not develop unless they managed to escape their traditional dependence on exports of agricultural products and raw materials was a key impetus to increased state involvement. Manufacturing products locally rather than importing them was seen as a central part of development, and emerging local manufacturers needed protection from established rich-country firms, if they were going to succeed at this process of 'import-substituting industrialization.' They also needed public investments in energy production and transportation. Deciding which of these investments made sense required planning. Private investors also faced a collective action problem. Investment in manufacturing would make more sense if other local entrepreneurs were making complementary investments that would provide needed inputs. Active state involvement was a way of giving entrepreneurs confidence that they were part of a general, mutually reinforcing set of investment decisions.

Without advocating expansion of the state's role, Albert Hirschman was one of the most creative contributors to thinking on these issues. Hirschman carried Gerschenkron's argument a step further, arguing that a dearth of entrepreneurship in the simple sense of "the perception of investment opportunities and transformation into actual investments" (Hirschman, 1958: p. 35) stood in the way of "late late industrialization" more than the lack of capital. The state could help induce private capital to become more entrepreneurial by providing incentives that made decisions harder to avoid. In Hirschman's view, inventing a national project of entrepreneurship is the key to development and the state can be an important participant in the process of invention.

Neoliberal Critiques of State Intervention

Import-substituting industrialization and the state involvement that went with it resulted in impressive rates of economic growth during the 1950s and 1960s in countries ranging from India to Brazil, but by the end of the 1970s development strategies in Africa, Latin America, and South Asia were in disarray. Imports were growing faster than exports, creating balance of payment problems, and state expenditures were increasing more rapidly than revenues, creating fiscal problems. One of the responses to this disarray was a renewed interest in the state on the part of neoclassical political economists.

Like Marxists, neoclassical political economists saw the state as the instrument of special interests rather than universal ones. Unlike Marxists, they applied the utilitarian logic of individual maximization to actors within the state itself. They saw the state extracting rents and giving entrepreneurs the wrong incentives rather than solving collective action problems and inducing more entrepreneurship. Theorists like Anne Krueger (1974) applied this logic to strategies of import-substituting industrialization, arguing that the desire to gain protection from foreign competition led entrepreneurs to focus on the unproductive quest for state favors rather than on competitiveness. Studies like Robert Bates' (1981) analysis of the perverse effects of state intervention on the development of African agriculture helped build the case that there were serious problems with earlier optimistic visions of the state's role in development.

These critiques had some merit. Development planning had indeed presumed an unrealistically benign state, ignoring the self-seeking behavior of state elites. Corruption was rife in developing states. Entrepreneurs in industries like textiles did indeed continue to demand that indulgent states provide them with protection long after they should have become mature and internationally competitive. This negative neoliberal image was best characterized by predatory states with their self-interested elites who extract society's surplus wealth and income, use it for their own benefit, and provide so little in the way of collective goods in return that development becomes impossible. Zaire under the rule of Mobutu in the 1970s and 1980s was a classic example. Under Mobutu, the income of Zaire's citizens shrank at over 2% a year, while state officials acquired multiple European mansions and Swiss bank accounts. Nonetheless, these neoliberal theorists lacked good answers as to how developing countries could overcome the weaknesses of their private entrepreneurial classes and self-serving state elites. Ironically, the policy prescriptions of the neoliberal paradigm actually reinforced the dynamics of predation as state and economic elites were released from the obligation of national development. The state's role was now focused on unleashing and deregulating markets, dismantling industrial and social policies, privatizing state assets, and removing barriers for imports. For the predatory elite, these were ideal conditions, as they were now free to pursue profits

unencumbered by the responsibility of providing collective goods and pursuing national economic development. Moreover, neoliberal theorists also failed to acknowledge that there were important variations, even among poor countries, in the behavior and capacities of states.

The Developmental State

For most states, the neoliberal view is partial and misleading. The state apparatus may extract, sometimes corruptly, but it also provides essential collective goods without which development could not proceed. Most interesting is the subset of states that are able to foster long-term entrepreneurial perspectives among private elites, help solve collective action problems, and supply essential public investments in education and infrastructure. The elites that control these states may not be immune from using society's surplus for personal ends, but the nations they govern enjoy faster rates of economic development because of the collective goods they supply. This category of 'developmental states' is best illustrated by looking at East Asia from the 1950s through the 1980s.

The East Asian 'Miracles'

The rise of East Asian economies further illuminated the role of the state in achieving development. In his study of Japan's Ministry of International Trade and Industry (MITI), Chalmers Johnson (1982) coined the term 'the developmental state,' which was almost immediately picked up and used in seminal studies of other newly industrializing economies (e.g., Amsden, 1989; Wade, 1990; Woo-Cumings, 1991; Evans, 1995). Japan's successful challenge of the industrial supremacy of Europe and the United States in the 1950s and 1960s was followed by the developmental success of South Korea, Taiwan, and Singapore. These countries grew six times faster than the advanced industrial countries had grown in the nineteenth century during the industrial revolution. They are unique in the Third World in the degree to which they have managed to change their position in the world economic hierarchy. These states are considered developmental, not just because the nations they governed grew but because the transformation involved in that growth was directly related to the behavior of the state.

The Analytical Characteristics of the 'Developmental State'

Starting with Johnson's (1982) analysis of the Japanese case, a variety of analysts have tried to specify what makes states developmental (see, for example, Amsden, 1989; Wade, 1990; Chang, 1994; Evans, 1995; Akyüz, 1999; Chibber, 2003; Kohli, 2004). While there has been much debate, there are surprising elements of consensus that extend into mainstream development economics (see Meier and Rauch, 2000: Chapter 9) and the international policy community (see World Bank, 1997). Among the essential structural features is a capable and coherent state bureaucracy that is connected to, but still independent from, various social forces and in particular the economic elite. An effective and independent state bureaucracy is necessary in order to sagaciously pursue economic policies in the national interest. Links to the industrial class are equally essential in order to persuade them to invest in key sectors of the economy and in collective goods. At the same time, the autonomy from the economic elite is vital so that the state can pursue national interests unencumbered, or without being captured, by private interests. Maintaining 'embedded autonomy' is pivotal for developmental success (Evans, 1995). The roots of such capable bureaucracies lie, in turn, in mechanisms familiar to students of Max Weber such as meritocratic recruitment and public service careers that offer long-term rewards commensurate with those available in the private sector.

The policies of developmental states mirrored the behavior Gerschenkron and Hirschman argued was needed in order to promote late industrialization. Investments in essential modern infrastructure, such as telecommunications, helped to set the stage for industrial growth. Entrepreneurs were induced to make long-term investments that raised national productivity. Subsidies and selective protection against external competition led to the creation of industries like semiconductors, steel, and shipbuilding that private entrepreneurs would never have embarked on without state support. Linked to this is the state's capacity and willingness to discipline capital if it fails to deliver on its end of the agreement (i.e., the state provides subsidies and protection in return for productive investment in targeted sectors). Exports, overall output, and real wages all grew at unprecedented rates. Public investments made rapid rises in education levels possible and health measures, such as infant mortality levels, improved. In short, essential features of the developmental state include a coherent and effective bureaucracy with the capacity to discipline capital, extract revenues through taxes for investment in public goods, provide an environment of rule of law, create appropriate policies, attract investment, protect industries, and encourage exports.

Contemporary Issues and Future Directions

Beginning with the East Asian financial crisis of 1997–98, the efficacy of the developmental state was brought into question. Current thinking about development and the state raise at least four different challenges. First, the evolution of the global political economy has brought into question (1) the capability of individual states to shape developmental trajectories and (2) the focus on manufacturing-based industrialization. Second, economic development saw the concomitant rise of powerful and militant labor movements that placed increasing pressure on states to democratize. State–business linkages that were the hallmark of the twentieth-century developmental state have to be reconceptualized to include relations with wider society. Third, in an even more fundamental challenge, the unidimensional understanding of development as economic growth has been broadened to include understandings of human well-being. Linked to this, the validity of traditional strategies of development is increasingly contested. And finally, the ecological crisis has put a fundamental challenge to fossil-fuel-based industrial development. In combination, this questioning suggests a redefinition of the role of the developmental state.

Globalization, Economic Restructuring, and the Role of the State

The reorganization of production, and more crucially finance, around global networks, in combination with the increasing hegemony of market-oriented Anglo-Saxon ideology and legal norms, makes traditional industry-focused state strategies more difficult, even when these strategies combine import substitution with export promotion, as in the case of developmental states. Even though global organizations, such as the World Trade Organization, continue to depend on the institutional capacities of individual states to implement global governance rules, the leeway that states have in making their own rules to increase local entrepreneurship has been restricted. Even developmental states have felt the effects of globalization.

In addition, in recent years there has been a transformation of the world economy from the predominance of mining and manufacturing to the high-value, information-intensive knowledge and service sectors. The knowledge economy is an infinitely more complex terrain for economic policy-makers than an economy based on manufacturing. For countries in the Global South that seek development, a shift to knowledge and service sectors will be a necessary ingredient for their success. This means that the traditional emphasis on machine production and a manufacturing-based economy (i.e., the physical production of tangible goods) must be supplemented by a turn to the more complex and innovative sectors of knowledge and service. Growth is increasingly found in 'bit-driven' information, and technology such as software codes, images, and formulas, and the state's role is even more important but vastly more complicated (Evans, 2014; Ó Riain, 2014; Wang, 2014).

The Crisis of the Developmental State and Expanded Embeddedness

The dominant tenor of first reactions to the 1997–98 Asian financial crisis was that either globalization has turned the developmental state into an anachronism or it was a hoax to begin with. More sober reflection produced a different diagnosis. One of the lessons of the crisis was that states cannot afford to relinquish controls over international financial flows before constructing appropriate mechanisms for regulating national financial markets. Abandoning its previous role, not clinging to it, was the developmental state's failing. This failing did not result simply from the pressures of globalization. It also reflected a change in the politics and policies of the developmental state itself. Having achieved impressive gains in global markets, local corporate elites were less likely to believe that a coherent, capable state bureaucracy was relevant to their success. Global ideology favoring the dismantling of the state's regulatory capability also became more popular within the state apparatus. Even though financial stability and growth were restored with remarkable rapidity, the crisis revealed that, precisely because of its economic success, the state's alliance with private capital was no longer likely to produce the same developmental results.

At the same time, the exclusive focus on state–business relationships and states' repressive stance toward organized labor unleashed protests that were increasingly difficult to restrain (see, for example, Lee, 2014 for a discussion of China). Indeed, various country experiences suggest that economic growth has brought about demands for more democratic participation, or an 'expanded embeddedness' beyond the developmental state's narrow alliance with the business elite. Sen (1999) captures this shift by highlighting that democracy is both a goal of and means to development. While both representative and direct democracy are vital components of democracy, Sen highlights the importance of direct democracy for ordering public interests and promoting effective citizenship (see Kerstenetzky, 2014; Isaac, 2014). Since the 2008 economic crisis in the Global North, calls for greater participation of ordinary citizens and increased state involvement in regulating economies have intensified. In short, expanded embeddedness and deepening democracy seek to increase society's power in relation to the state and capital. At the same time, the 'accumulation-oriented' definition of development that had formed the basis of the state's alliance with capital was increasingly criticized.

The Contested Definition of Development

The strategy of the developmental state, like the strategy of import-substituting industrialization before it, was built on the premise that investment in local manufacturing would produce jobs and that the incomes from these jobs would generate widespread increases in well-being. There have always been critics of this view, but recently they have become more vociferous (e.g., Evans, 2014; McMichael, 2007; De Sousa Santos and Rodriguez-Garavito, 2006; Escobar, 1995; Sachs, 1992). Critiques suggest that as manufacturing becomes more capital intensive, manufacturing investment consumes a disproportionate share of the social surplus relative to the number of jobs that results. Even when the money incomes of some ordinary workers do increase, improved overall well-being is not necessarily the result. Their ownership of more consumer goods does not compensate for the environmental degradation, increasing inequality, and destruction of intangible social and cultural assets that accompany the accumulation of productive capacity. For too large a proportion of the citizens of poor countries, the new global cornucopia of heavily advertised consumer goods remains beyond economic reach but still devalues prior cultural and social satisfactions.

While such perspectives may ignore the benefits of investment in productive capacity, they have contributed to the challenges of defining development exclusively in terms of accumulation and economic growth. As a result, the ideational understandings of development have experienced revision from a singular concern with economic growth to a more encompassing goal of social and human development. Thus, development has become more complex and multifaceted to include quality of life issues and noneconomic indicators (Sen 1995, 1999, 2001). (The idea of a 'happiness index,' first pioneered by the Kingdom of Bhutan, is also an effort to move beyond economic indicators of development.) A new generation of theories have emerged that include the 'basic needs approach,' which looks at social services (e.g., adequate food, clean drinking water, decent shelter, education, and security of livelihood) as a means for poverty reduction (Burki and Ul Haq, 1981; Melkote, 1991), the 'new growth theory' and the

human capabilities approaches both of which highlight investment in education, ideas, and democratic processes as a means for economic growth (Evans, 2014; Helpman, 2004; Sen, 1999; Romer, 1990, 1994). In short, the accumulation-oriented definition of development has been fundamentally challenged, both empirically and ideationally, leading to more encompassing and complex understandings of development.

Ecological Crisis and Development

In addition to the above challenges, ecological issues were not even defined as a problem at the start of the twentieth century, but now are considered to be one of the most important challenges states face (Harriss-White, 2014; Satgar, 2014; Castells, 2010; Peet and Watts, 1996). Development has been predicated on fossil-fuel-driven industrialization, using enormous resources and creating vast amounts of pollution and waste that most severely affect citizens in the Global South. Would-be developmental states of the twenty-first century have to find new ways of developing their economies that do not intensify the destruction of the natural world, including investing in renewable forms of energy. There are also calls not just for the protection of the environment but also to ensure that resources are used more equitably for the benefit of all and not just an elite few both within countries and among countries (Cock, 2010). The ecological question will certainly be one of the most important challenges for states seeking development for the foreseeable future.

Redefining the State's 'Developmental' Role

The changes in the goals of development and the conditions under which it occurs have posed fundamental challenges to the state's role in development. If development is really about well-being rather than about accumulating industrial capital, then what is the role of the state? A 'well-being-centered' redefinition of development implies a more aggressive focus on the state's traditional role in delivering collective goods and services, like health, education, clean air, water, and sewage disposal, and in preserving the collective environment, both natural and cultural, as well as promoting democratic participation of citizens. The kinds of state behavior and structures required by this redefined developmental role are surprisingly consistent with early pre-neoliberal thinking. Hirschman's notion of inducing decision-making among private economic agents becomes an even bigger challenge if the focus is on the provision and preservation of collective goods. Finding ways to convince entrepreneurs that they should invest in clean production techniques or take into account the social impact of investment decisions requires much more skill and capacity than helping them increase their profits through subsidies and protection. A would-be developmental state today must have the capacity to do the tasks of the twentieth century, but also add a commitment to human well-being, democratic participation, protection of the environment, and greater equality. At the same time, the conditions of the global economy have created a more complex environment under which states must pursue these goals. The greater magnitude of the challenge makes the emphasis on a capable, coherent, yet innovative and responsive to citizen engagement, state bureaucracy, which is central to thinking about the developmental state, more important than ever.

The biggest single change implied by this redefinition is already foreshadowed by the crisis of the developmental state. State–society relations dominated by close ties to industrial elites at the expense of other social groups provide neither the political will nor the societal collaboration necessary to execute a 'well-being-oriented' strategy of development that includes care for the natural environment. A more broadly defined set of state–society relations is imperative. There are some studies of the role of more broadly based state–society ties in promoting development (e.g., Isaac, 2014; Kerstenetzky, 2014; Heller, 1999; Evans, 1996), but extending and deepening this kind of analysis is the most obvious challenge confronting contemporary work on development and the state.

Acknowledgments

This article draws on and updates the earlier article on 'Development and the State' written by Peter Evans for the 2001 edition of International Encyclopedia of the Social and Behavioral Sciences. I would like to thank Peter Evans, Vishwas Satgar, Roger Southall, and Devan Pillay for their comments and feedback.

See also: Development Theory in Geography; Environment and Development; Environmental Conservation and Development: Critical Perspectives; Planning Issues and Sustainable Development; Science and Development; State, Geography of; State, Sociology of the; Sustainable Development: An Economic Perspective.

Bibliography

Akyüz, Y. (Ed.), 1999. East Asian Development: New Perspectives. Frank Cass, London.
Amsden, A., 1989. Asia's Next Giant: South Korea and Late Industrialization. Oxford University Press, New York.
Bates, R.H., 1981. Markets and States in Tropical Africa: The Political Basis of Agricultural Policies. University of California Press, Berkeley, CA.
Burki, S.J., Ul Haq, M., 1981. Meeting basic needs: an overview. World Development 9 (2), 167–182.
Cardoso, F.H., Faletto, E., 1979. Dependency and Development in Latin America. University of California Press, Berkeley, CA.
Castells, M., 2010. Power of Identity: The Information Age: Economy, Society, and Culture, vol. II. Wiley and Blackwell, West Sussex.
Chang, H.-J., 1994. The Political Economy of Industrial Policy. Macmillan, London.
Chibber, V., 2003. Locked in Place: State-Building and Late Industrialization in India. Princeton University Press, Princeton, NJ.
Cock, J., 2010. Green Capitalism, unpublished discussion paper.
De Sousa Santos, B., Rodriguez-Garavito, C., 2006. Introduction: expanding the economic canon and searching for alternatives to neoliberal globalization. In: de Sousa Santos, B. (Ed.), Another Production Is Possible: Beyond the Capitalist Canon. Verso, London and New York.
Dos Santos, T., May 1970. The structure of dependence. American Economic Review 60 (2), 231–236.
Escobar, A., 1995. Encountering Development: The Making and Unmaking of the Third World. Princeton University Press, Princeton, NJ.
Evans, P.B., 1979. Dependent Development: The Alliance of Multinational, State, and Local Capital in Brazil. Princeton University Press, Princeton, NJ.

Evans, P.B., 1995. Embedded Autonomy: States and Industrial Transformation. Princeton University Press, Princeton, NJ.
Evans, P.B. (Ed.), 1996. Government action, social capital and development: creating synergy across the public-private divide. World Development 24 (6), 1033–1132 (special section).
Evans, P.B., 2014. The developmental state: divergent responses to modern economic theory and the twenty-first century economy. In: Williams, Michelle (Ed.), The End of the Developmental State? Routledge, London and New York.
Frank, A.G., 1967. Capitalism and Underdevelopment in Latin America. Monthly Review Press, New York.
Gerschenkron, A., 1962. Economic Backwardness in Historical Perspective. Belknap, Cambridge, MA.
Heller, P., 1999. The Labor of Development: Workers and the Transformation of Capitalism in Kerala, India. Cornell University Press, Ithaca, NY.
Helpman, E., 2004. The Mystery of Economic Growth. Harvard University Press, Cambridge, MA.
Harriss-White, B., 2014. Development in an anti-developmental state: the market politics of renewable energy in an advanced country and its implications for the environment. In: Williams, Michelle (Ed.), The End of the Developmental State? Routledge, London and New York.
Hirschman, A., 1958. The Strategy of Economic Development. Yale University Press, New Haven, CT.
Isaac, T.M., 2014. Politics of democratic decentralization and the developmental state: a study of the Kerala experience. In: Williams, Michelle (Ed.), The End of the Developmental State?. Routledge, London and New York.
Johnson, C., 1982. MITI and the Japanese Miracle: The Growth of Industrial Policy, 1925–1975. Stanford University Press, Stanford, CA.
Kerstenetzky, C.L., 2014. The Brazilian social developmental state: a progressive agenda in a (still) conservative political society. In: Williams, Michelle (Ed.), The End of the Developmental State? Routledge, London and New York.
Kohli, A., 2004. State-Directed Development: Political Power and Industrialization in the Global Periphery. Cambridge University Press, New York.
Krueger, A.O., 1974. The political economy of the rent-seeking society. American Economic Review 64 (3), 291–303.
Lee, C.K., 2014. The enigma of Chinese capitalism. In: Williams, Michelle (Ed.), The End of the Developmental State? Routledge, London and New York.
McMichael, P., 2007[1996]. Development and Social Change, fourth ed. Pine Forge Press, Thousand Oaks, CA.
Melkote, S.R., 1991. Communication for Development in the Third World. Sage, New Delhi.
Meier, G.M., Rauch, J.E., 2000. Leading Issues in Economic Development, seventh ed. Oxford University Press, New York.
North, D.C., 1981. Structure and Change in Economic History. Norton, New York.
Ó Riain, S., 2014. Liberal globalization, capabilities, and the developmental network state in Ireland. In: Williams, Michelle (Ed.), The End of the Developmental State? Routledge, London and New York.
Peet, R., Watts, M. (Eds.), 1996. Liberation Ecologies: Environment, Development, and Social Movements. Routledge, London and New York.
Polanyi, K., 1957[1944]. The Great Transformation: The Political and Economic Origins of Our Time. Beacon Press, Boston.
Romer, P., 1990. Endogenous technological Change. Journal of Political Economy 98, S71–S102.
Romer, P., 1994. The origins of endogenous growth. Journal of Economic Perspectives 8, 3–22.
Sachs, W. (Ed.), 1992. The Development Dictionary: A Guide to Knowledge as Power. Zed Books, London.
Satgar, V., 2014. South Africa's emergent 'green developmental state'? In: Williams, Michelle (Ed.), The End of the Developmental State? Routledge, London and New York.
Sen, A., 1995. Rationality and Social Choice. American Economic Review 85, 1–24.
Sen, A., 1999. Development as Freedom. Knopf, New York.
Sen, A., 2001. What development is about. In: Meier, G., Stiglitz, J. (Eds.), Frontiers of Development Economics. Oxford University Press, New York.
Wade, R., 1990. Governing the Market: Economic Theory and the Role of Government in Taiwan's Industrialization. Princeton University Press, Princeton, NJ.
Wallerstein, I., 1974. Dependence in an interdependent world: the limited possibilities of transformation within the capitalist world economy. African Studies Review 17 (1), 1–26.
Wang, J., 2014. Developmental state in transition: the state and the development of Taiwan's biopharmaceutical industry. In: Williams, Michelle (Ed.), The End of the Developmental State? Routledge, London and New York.
Woo-Cumings, M., 1991. Race to the Shaft: State and Finance in Korean Industrialization. Columbia University Press, New York.
World Bank (IBRD), 1997. World Development Report: The State in a Changing World. Oxford University Press, New York.

Development and Urbanization

Diane E Davis, Harvard University, Cambridge, MA, USA
Alexander M Keating, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D.E. Davis, volume 6, pp. 3560–3566, © 2001, Elsevier Ltd.

Abstract

This article examines the interrelationship between economic prosperity and the growth of cities, tracing the field from its original preoccupation with overurbanization and underdevelopment in the 1950s and 1960s to its current fixation on dynamic global cities with redeveloped property markets that showcase new forms of wealth creation. The historical change in emphasis chronicled here is understood to be combined product of three different causalities. The first is the shift from industrialization to financial and other services as the principal source of wealth creation in the post-1980s era. The second is the changing territorial scale of accumulation, reflected in the shifting importance of global markets vis-à-vis national markets and in the increasingly key mediating role that cities play in facilitating this transition. The third is the rescaling of state power, seen not just in terms of decentralization but also in the declining capacities of national states to discipline global investors in an era of intensifying economic liberalization. The article ends with a discussion of the emergent social and spatial problems that accompany these shifts, ranging from the rise of urban informality to dispossession and displacement to newfound struggles over urban property rights.

Economists, sociologists, political scientists, and anthropologists as well as demographers and planners have long theorized a direct connection between national development, understood economically and/or politically, and the growth of cities. But there has been little consensus about the nature and direction of the relationship, the desirability of analytically linking these processes, and how to define both terms. Reigning ideas have shifted in the face of new theoretical paradigms and as the territorial scope and essential character of markets, states, and cities themselves transform. Unresolved issues include which urban dynamics most enable or constrain developmental patterns, and whether the global phenomenon of rapid urbanization, intensifying economic globalization, and worldwide population concentration in cities have changed our theoretical or empirical understanding of these processes.

Historical Antecedents

Most scholarship on the relationship between development and urbanization traces its intellectual origins to writings by nineteenth- and early twentieth-century sociologists and economic historians who sought to account for the development of capitalism, the nation-state, and social modernization in early Europe, and who posited a direct and positive role for cities in these fundamental transformations (*see* State Formation, Theory of; Modernization, Sociological Theories of; Capitalism).

Cities and Development in the Early Modern World

In early modern Europe, cities emerged from and fueled processes of capital accumulation, by virtue of their role as exchange nodes in local, national, and international trade (Pirenne, 1936; Weber, 1927). Their growth also was tied to the demise of feudal or absolutist orders (Weber, 1958) and the rise of the modern nation-state (Tilly, 1975). By the late nineteenth century, cities also served as the sites in which the social relations of modernity materialized (Durkheim, 1933; Simmel, 1950), in turn transforming markets and states.

Twentieth-Century Developments

In the twentieth century when nationalism, war, empire, trade, democracy, and fascism required European attention, many US scholars turned to the study of cities. This was a time when American society demographically tipped its balance to become more urban than rural, and when new waves of postwar black migration combined with renewed streams of international migration to spark interest in the social relations of urban life. Initial research concentrated on the social process and experience within cities, ranging from ghetto life, immigrant social clubs, and gang organization to criminality. Few social scientists chose to examine the larger political and economic processes that engendered patterns of urbanization in the US, as occurred among European urban scholars; and even fewer tried to link urbanism or urbanization to patterns of national development, understood either politically or economically. It was only when US sociologists began studying cities comparatively and historically, starting in the 1950s and during a period of post-WWII expansionism, that debates over the relationship between development and urbanization garnered a central place. When the March 1955 (Vol. 60, No. 5) edition of the *American Journal of Sociology* published an entire volume on 'World Urbanism', the importance of these themes was firmly established.

The desire to examine other parts of the world was partly grounded in a passion for testing ideas drawn from the European-derived grand theoretical narratives of sociology. Practically the only counterfactual cases available for confirming prevailing propositions about the connection between urbanization and developmental progress – whether understood in terms of democratic state formation, capitalist development, and/or 'modern' social values – were the countries of

the so-called underdeveloped world of Latin America, East Asia, and Africa. These were countries of the world racked with poverty, limited industrialization, and governed primarily under oligarchic pacts or by despotic or authoritarian governments (*see* Attitudes Political: Authoritarianism and Tolerance). To the extent that their urbanization patterns also differed across country contexts, scholars were inspired to revisit general claims about development and urbanization, as well as to ascertain how these two processes were articulated in practice.

Working under the influence of the Chicago School, initial scholarship on what were then called Third World cities focused primarily on the folk-urban continuum and urbanization as a social process, including the ways that urbanization produced individualism and severed kinship bases of social organization (see, e.g., Hauser, 1957; Redfield, 1941, 1953). Over time, interest grew in urbanization as a demographic and spatial process, with scholars examining patterns of population concentration and the growth of cities in a wide variety of comparative and historical contexts (Berry, 1973; Hauser and Schnore, 1965), occasionally linking them to general propositions about national development. In addition to Hauser and Schnore's seminal *The Study of Urbanization* (1965), the Chicago-based journal *Economic Development and Cultural Change* served as an important outlet for new ideas on the topic.

Twentieth-Century Dilemmas: Questioning the Nature and Direction of Causality

It was one thing to quantify urban population patterns across time and place (Davis, 1969; Hoyt, 1962), to identify the extent to which urbanism, as a social attribute, materialized in different localities around the world (Breese, 1966; Mangin, 1970), and to elucidate growth processes in these cities (Gugler, 1978; McGee, 1971). American social scientists were quite good at these tasks, especially the first; and their successes helped catapult the field of demography onto the disciplinary map in the 1960s and 1970s (see, e.g., Goldscheider, 1971; Goldstein and Sly, 1979). It was quite another to demonstrate a direct relationship between urbanization and patterns of economic development, especially within the confines of the nation-state, the unit that most social scientists were then using for assessing such issues. It was on these latter counts that scholars were least successful, and that longstanding propositions about the positive relationships between urbanization and national development began to generate widespread skepticism.

As Third World cities burgeoned over the 1960s and 1970s, scholars faced the fact that even with high urbanization rates, developmental gains in most of the late industrializing world remained minimal, not just in comparison to the modal patterns established in Europe and the United States, but also as measured in per capita income, gross national product, and practically all other standard macroeconomic indicators (Friedmann, 1967). With dependency theory taking center stage at about the same period of time, certain attributes of both urbanization and national development were thus traced to colonial or mercantile relations that spanned national boundaries rather than to political, social, and economic conditions within the formal confines of the nation-state (*see* Dependency Theory). These transnational relations modified urbanization and development processes, at least when compared to early modern Europe, even as they raised new questions about the best unit of analysis to study the nature and direction of the proposed relationship. Did urbanization husband development, or vice versa; and was it possible that in the late developmental context the relationship between the two sets of processes was more likely to be negative than positive, with one impeding rather than facilitating the other? The answers to these questions shifted over time.

Development and Urbanization *c.* 1950: A Positive Synergy

The predominant argument in the 1950s and early 1960s was that the plight of developing nations and their low rates of urbanization owed to their 'backward' social and economic nature, which not only explained why industrialization was minimal, but also why tradition-bound peasants demographically outbalanced 'modern' urban folk. Analytical attention was directed toward how to facilitate one so the other would follow suit, a normative preoccupation pervasive in American social science and quite consistent with US foreign policy goals of the times.

There were disagreements about which domain deserved intervention first. Some urban scholars argued that the best way to jump-start Third World development was to shift population and resources out of the countryside and into the city, where industry could expand and modern values flourish (Herrick, 1965; Field, 1970). Others called for more foreign aid or national investment to buttress urban economies, primarily through industrialization (Richardson, 1973), in the hopes that such macroeconomic policies would stimulate employment opportunities and perhaps even produce individuals inculcated with rationality and enough achievement drive to sustain further urban and national gains (*see* Industrialization). Despite policy disagreements, both schools of thought expressed enough faith in the idea of a positive relationship between development and urbanization to keep it alive, albeit embodied in a variety of paradigmatic forms.

Over the 1960s and 1970s, one particularly popular variant was demographic, as those interested in urbanization as a social process looked for linkages between urban growth and national development in fertility patterns as well as family structure and employment. During this period ideas about the demographic transition, at times recast in terms of the 'urban transition' (Friedmann, 1975), generated considerable interest among economists and sociologists, who developed expert quantitative skills that sustained a field of research broadly understood as 'population studies'. Those less preoccupied with the social or economic behavior of individuals and more interested in overall patterns of city growth and distribution were more likely to concern themselves with ideas about the proper balance between urban and rural populations (El-Shakhs, 1974; Friedmann and Alonso, 1964).

During this period, scholars who assumed that the US served as a model of successful economic development paid

considerable attention to the extent to which urban systems in developing countries achieved a lognormal rank-size distribution, as opposed to being dominated by one or two large 'primate' cities (Berry, 1961), basing their interest on the assumption that rank-size lognormality correlated with economic development. In counterattack, these ideas were criticized as inherently Western-biased, if not drawn directly from the American experience, and thus inapplicable in the Third World (El-Shakhs, 1972; McGreevey, 1971). While there was no resolution on this contentious issue, in the process the concept of 'overurbanization' gathered widespread attention, since a city could only be 'too' urbanized if it grew beyond its expected and economically efficient size, which often was calculated on the basis of a rank-size distribution logic.

To the extent that most Third World countries fell into the latter camp, with their cities ballooning in size beyond their fiscal and infrastructural capacities, scholars began to study the internal social and demographic processes that fueled patterns of overurbanization, and what Michael Lipton (1977) called 'urban bias', including rural–urban migration and the national investment decisions that over privileged a few urban centers at the expense of struggling provincial towns and the impoverished countryside. Still, concerns about city systems died hard, rematerializing in slightly new packaging in the 1980s in the form of arguments about the importance of 'secondary cities' to sustaining national development (Hardoy and Satterthwaite, 1986; Rondinelli, 1983). This claim, like the rank-size distribution argument, was built on the understanding that (smaller) size and (more equal) distribution of cities mattered; and both arguments confirmed the preoccupation with a relationship between patterns of urbanization and development that had initiated and sustained the field.

But far from merely confirming a connection between cities and national development, these findings laid the foundation for an immanent critique of long-standing claims. Indeed, to the extent that the burgeoning growth of Third World cities was soon seen as an obstacle to economic well-being, there was now evidence that a positive relationship between urbanization and development was effectively reversed on substantive grounds, even if analytically the connection remained. To be sure, there were scholars who conceived of large cities as economically efficient localizers of economies of scale, and they continued to claim a positive relationship between large-scale urbanization and development (Richardson, 1973). But by and large, promoters of this idea were arguing against a tide of contrary evidence.

Development and Urbanization, *c.*1975: A Destructive Relationship

The ascendance of ideas about a negative relationship between urbanization and development owed partly to the fact that between 1950 and 1975 conditions in the developing world changed dramatically. This was the period when most Third World cities began to grow enormously, outpacing the capacities of their governments to invest in infrastructure or sustain industrial output in ways that contributed to national development. With large cities peopled by new migrants and un- or underemployed folk with limited education and skills for a newly industrializing economy, they soon were seen as a drain on national coffers. The increasing urban inequality and economic polarization accompanying these trends also sustained urban social movements and overall political coalitions that themselves limited national developmental prospects by bringing to power governments with restrictions on foreign investment, populist economic policies, and/or protectionist measures that frequently undermined short-term efficiency goals. These social and political problems became so salient in the urban landscape of the developing world that eventually scholars turned attention directly to them (see Abu-Lughod and Hay, 1977; Gugler, 1988; Mangin, 1970), partially eclipsing the concern with national development. Among the most documented issues were urban housing and employment scarcities, rural–urban migration flows, illegal settlements and urban 'marginality,' and migrant or informal sector politics (Cornelius, 1973; Drakakis-Smith, 1987; Perlman, 1976).

This is not to say that all those who looked internally to cities over the 1970s and 1980s failed to examine the larger relationships between urbanization and national development that had consumed their predecessors. Economists at the World Bank and other macroeconomic policymakers loyal to free market ideals continued to study urban economies of scale and the ways urban growth interfaced with industrialization, still seen as the principal route toward prosperity; although they too concerned themselves with questions about housing, illegal settlements, and urban service scarcities more generally (Linn, 1983). Likewise, but on an entirely different front, sociologists and political scientists began to examine the national political conditions that made these urban problems so pervasive (Eckstein, 1977; Gilbert and Ward, 1985; Rabinowitz, 1973; Roberts, 1978). But this also meant that urban scholars were focusing as much on political as economic developments, linking urbanization patterns to national politics as much as markets (*see* Populism).

When dependency theory emerged as a dominant paradigm among both resident and foreign scholars of the developing world over the 1970s and 1980s, fewer scholars began to pose questions about the relations between urbanization and national development unless they were framed within a focus on international economic conditions (Castells and Velez, 1971; Portes and Walton, 1981; Walton, 1985); and even fewer saw the articulation of urban and national development in positive terms, as had their predecessors. If anything, when urbanization was linked to capital accumulation, the social effects were seen as relatively negative (Armstrong and McGee, 1985). By the late 1980s, the main idea in good currency was that capitalist development on a global scale produced and/or reinforced mercantilist relations that fueled the growth of cities, even as they disadvantaged the countryside or the nation as a whole.

Using a global framing, scholars focused not just on the ways that unequal, semi-imperial relations between core and periphery brought a few large cities, themselves parasitic with respect to their hinterlands. They also examined the ways these relations produced a form of 'internal colonialism' predicated on extreme income and investment polarization between city and countryside, which further fueled migration,

contributed to overurbanization, and reinforced income and other inequalities within cities themselves (see Timberlake, 1985). The argument for a negative relationship between urbanization and sustained or autonomous national development thus reached new heights of acceptance (*see* World Systems Theory).

Development and Urbanization, *c.*1990: Disarticulated Processes

The growing popularity of world-system theory over the 1980s and 1990s may have strengthened scholars' understandings of the strong connections between urbanization and (under) development (Smith, 1996); but in the long run it contributed to the analytic disarticulation of cities and nations, thereby casting a mortal blow to studies of relations between these two domains. As scholars joined the world-system bandwagon and bypassed the nation-state as a unit of analysis, few examined national development as analytically distinct from global processes. Those development scholars who did concern themselves with the domestic dynamics of macroeconomic development rarely reserved a place for cities in their analytic frameworks, turning instead to states, classes, and the global economy. The result was that scholars who analyzed cities in the developing world did so increasingly in the context of global economic conditions (Friedmann, 1986; Abu-Lughod, 1989), even as most scholars of national development ignored cities by and large.

This analytic disarticulation of urbanization and national development was further fueled by intellectual and political dynamics in the developing world. With many countries in Latin America, Africa, and Asia becoming less democratic and more authoritarian throughout the 1960s and 1970s, their states became stronger and more repressive. These political regimes tended to privilege capital over labor and cities over the countryside, using macroeconomic policy, national investments, and targeted subsidies to accomplish these aims. Resident and foreign observers saw little autonomy for cities, with urban development mostly dictated by powerful (if not militarized) national states. This was most evident in the largest cities, home to their nation's most significant industries and investments; but it also was seen as a pattern more generally. The result: in studies of the Third World, urban dynamics were frequently subsumed under national ones. This conceptual sleight of hand may have reinforced views about the direct relationship between city and nation to some extent; but it also chipped away at the analytic distinction between urban and national dynamics. And with these two processes increasingly collapsed into each other, the tendency was to examine one or the other, but not the interactive relationship between the two.

Among the few still fervently examining the direct connections between urbanization and national economic development were neoclassical economists, along with World Bank staff and their clients. They worked under the supposition that by facilitating and/or managing cities, national actors could ensure productivity, efficiency, and the flow of people and goods sufficiently to maximize overall development. Yet in most of the nonapplied fields, urbanists and developmentalists hardly spoke to one another.

A New Articulation (or, Development and Urbanization in the Age of Economic Globalization)

By the turn of the millennium scholars began to widely accept the premise that globalization ought to be viewed as a central point of entry for studying changes in the contemporary world, thus bringing about renewed interest in the urbanization-development nexus. Initially, scholars became fixated with cities of enormous size and importance, whose growth and character owed to the role they play not just within their national borders, but also within global networks of production and consumption (Henderson and Castells, 1987; Lo and Yeung, 1998; Sassen, 1991, 2000). This realization also produced a shift in terminology. Few scholars referred to Third World cities anymore, preferring instead 'cities of the global south' (or north) or the concept of world or global cities, both in recognition of the unequal world distribution of production and consumption wealth. The idea of world cities was not new of course (Friedmann and Wolff, 1982); and much of the work drew inspiration from earlier writings on overurbanization in Third World cities in which the inflated size and rapid growth of primate cities was linked to unequal relations of exchange between so-called core and peripheral economies. But as economic globalization continued to bring rich and poor countries into the same orbit, scholars were more willing to think about cities in global rather than national context, hence the popularity of world or global city nomenclature. Similarly, world cities were less likely to be seen as fetters on the national development of their host countries, and more likely to be perceived as the mechanisms through which global economic integration took root.

Such views not only challenged the dependency era understandings of the destructive impact of global dynamics on cities and nations, they were also accompanied by a new geographic focus among urbanists. Over the period 1960–2000 most scholars interested in the urbanization-economic development nexus examined the poorer countries of the global south; but by the 1990s the US and Europe were back in the center of the conceptual map (King, 1990; Sassen, 1991), with cities like New York, London, Tokyo, and Paris taking center stage as global cities where urban employment patterns, shifts in sectoral character, and spatial forms were seen as products of the globalization of capital and labor (Fainstein et al., 1992; Sassen, 1991). The analytical focus also produced a renewed interest in the changing 'locations' or economic roles these cities played in a regional, national, or international hierarchy of urban places (see Knox and Taylor, 1995).

But as the global system of world cities itself began to transform, new questions began to emerge as to whether those cities with the greatest potential to generate national or even international developmental gains were indeed sited in the advanced capitalist world (see Robinson, 2002; Gugler, 2004). Such queries have arisen in response to rapid and concentrated urbanization of cities in the emerging economies of India, China, Brazil, and other formerly poor regions where rapid and concentrated urbanization has been accompanied by massive wealth creation for individuals, urban authorities, and national accounts. As some scholars turned

their attention to these urban transformations, chronicling the new cadre of city builders responsible for jumpstarting formerly lagging economies through global city branding (Ren, 2011), others began to criticize the rush to find new ways of categorizing cities in terms of their paradigmatic status (Beauregard, 2003; Brenner, 2003; Roy and Ong, 2011). Among those most willing to jettison superlatives and theorize cities in more 'ordinary' terms (Robinson, 2006) were those who also were willing to recognize that even the most prosperous cities in the late developing world still hosted high degrees of poverty, inequality, and extraordinarily high degrees of economic informality. As such, even with focus shifting to formerly poor regions of the world, questions persisted about nature of the relationship between development and urbanization, and whether the latter could really be seen as a route to prosperity if the scale of the city remained the unit of analysis.

Development in the 'Urban Century': Are Cities Leaving Nations Behind?

Though exact projections differ, it is widely accepted that the global population reached and surpassed the demographic tipping point to become majority urban in or around 2008 (Satterthwaite, 2007). Within the contemporary study of urbanization and development, the more significant trend may be that nearly all population growth projected to take place globally over the next 30 years will be in developing-world cities throughout Asia and Sub-Saharan Africa and that this growth is unlikely to slow significantly before 2030 (United Nations, 2012). Due to the high level of variation in national and subnational development trajectories throughout these regions, most notably between the comparatively high-performing economies of countries such as China, India, Singapore, and South Korea, and economically stagnant, but rapidly urbanizing countries of Sub-Saharan Africa, two distinct analytical frameworks have emerged within the current study of the development-urbanization nexus: one continues to articulate the role of cities and agglomeration economies as the principle drivers of economic and social development, while the other calls attention to the significant social and environmental costs associated with rapid and often unplanned urbanization.

Those who argue that cities are the privileged sites for economic development have brought new analytical rigor to the debate by focusing on how urbanity unleashes creative potential through agglomeration and the spillover effects of knowledge economies in developed and developing world cities alike. These ideas, gaining popularity among economists in particular, view the proliferation of urban slums in the developing world not as markers of systemic weakness, but rather of 'the city's' ability to attract rural poor through the promise of improved livelihoods and health (see Glaeser, 2011). In the developed world, attention is shifting to economic clusters in key cities or areas within them, with research suggesting that such experiments can support new, creative economies in previously declining, postindustrial cities and regions of the United States and Europe, thus reenergizing the study of the urban-economic development nexus at a scale smaller than the nation (see Porter, 2003).

As a sign of the globalization of ideas, cluster theory also migrated to the developing world, particularly in East Asia where bottlenecks among export-led industrializers pushed their national governments to invest in industrial clusters outside major cities. But the theoretical and practical significance of cluster theory in the developing world has been slow to gain a foothold in the face of the increasing national obsession with producing global cities that can showcase high-end real estate, global finance, and ICT activities, all considered the principal avenues through which cities and their nations were to prosper economically (Olds, 2011).

The preoccupation with finance, services, communications, and other tertiary sector activities can be explained by international capital's search for new investment outlets in the face of declines in industrial and agricultural output. Yet it also captures the imagination of investors – both public and private – in emerging economies that seek to generate national economic gains through more than export-led industrialization or agriculture. The cases of China and India are exemplary in these regards. In these and a selected array of other emerging economies, urbanization has become associated with major transformations of the urban built environment, often in the form of iconic architectural projects and massive mega-project developments that regularly coexist side by side with high overall levels of urban poverty and informality.

New Governance Arrangements

The move to transform urban landscapes into the incubators and showcases of innovation and development is greatest in cities where decentralized governance has limited the local state's fiscal capacities. In the search for fiscal incentives to attract the innovative classes, authorities have turned to the private sector for resources and ideas, thus increasing state willingness to support the building of global city landscapes. Reliance on local public–private partnerships for urban development is seen as enabling regional and national progress outside the typically slow moving and bureaucratic mechanisms of national-level governance, drawing strong comparisons with the introduction of Special Economic Zones, traditionally oriented toward industrial uses, in developing countries throughout the 1980s and 1990s (see Fuller and Romer, 2010).

In the neoliberal era of decentralization and state downsizing, experiments that enable free market zones to guide development usually involve both public and private sector actors, as both share incentives to generate and capture revenues. But when local states begin to ally with real estate developers around such massive urban transformations, and when they join private investors in displacing or dispossessing residents from properties or locations, they often face political opposition. While similar challenges in the 1970s placed the analytic focus on the failure of the state to keep pace with increasing urban populations, contemporary scholars, and development agencies have broadened the focus to include the failure of the market, private sector investment, and neoliberal structural adjustment policies to adequately account for the needs of the urban poor (see Davis, 2004; Beall,

2002; Harvey, 2008). Thus, to the extent that neo-liberal restructuring of urban space within the context of continued globalization produces extreme spatial fragmentation, exclusionary urban development, and local conflict, the route to global city status via transformation of urban land uses and accelerated real estate development rather than manufacturing and industrialization continues to be paved with social and political obstacles.

For this reason, much of the recent work produced by scholars and high-profile multilateral development agencies continues to highlight the many negative externalities associated with the dispossession and dramatic transformations in urban land use deployed in order to pave the way for these new projects, including decreasing access to shelter and secure tenure for existing and incoming populations of urban poor and displacement of vulnerable populations to the urban periphery. Even among those scholars whose concern is less about land use change and more about the seemingly inexorable process of urbanization that draws ever larger proportions of a nation's population into a circumscribed spatial orbit, assessments underline serious challenges. From inadequate infrastructure to environmental degradation to significant increases in urban poverty and rising rates of violence and insecurity (see Martine et al., 2008), rapidly urbanizing cities are seen as hosting a plethora of pathologies that persist in the midst of accelerating prosperity.

Urban Informality: Where Citizens, Urbanization, and Development Collide

In the most rapidly urbanizing regions of the world, the management of urban informality, most clearly manifested through the increasing occurrence of slums and informal settlement patterns, has emerged as one of the most consistent lens through which scholars and practitioners now address the costs and benefits of rapid urbanization across the developing world (Roy and AlSayyad, 2004). While the cities of Sub-Saharan Africa have the highest rates of informality and slum growth as a percentage of total urban expansion, Asian cities are now home to the largest total number slum-dwellers, placing the issue at the center of a multidisciplinary debate over the role of urban informality as a driver or inhibitor of economic development. From an economic standpoint alone, the informal economy in developing nations, operating primarily in the urban periphery and 'slum' settlements, is estimated to account for upwards of 50% of all urban employment across the developing world (see Becker, 2004). Despite the observable increases in per capita economic productivity that accompany rural–urban migration, others point out that the lack of formal tenure and property rights which has accompanied the relative lack of affordable housing provision by local governments and the private sector alike, alongside the inadequate provision of educational opportunity, likely account for tens of billions of dollars of unrealized capital in slums (see de Soto, 2001; Yamauchi et al., 2009).

Renewed interest in the forms and functionality of informal settlements is also closely related to the continued consolidation of human rights at the heart of development discourse, most clearly articulated through the formulation and adoption of the Millennium Development Goals (MDGs) as primary measures of development and progress in the twenty-first century. The demographic shift to a majority urban global population, along with the expansion of urban footprints at rates that more than doubles that of urban population growth (see Angel et al., 2011) mean that the environmental, economic, and social impact of urbanization has a progressively more meaningful impact on national and global outcomes. In this capacity, cities have emerged as one of the few truly cross-cutting arenas in which to effectively implement policies and measure progress toward meeting the MDG targets (see Hasan et al., 2005).

Beyond the Urban: Rescaling the Territorialities of Development

Though urbanization is likely to remain widely regarded as the primary source of development, employment, and capital accumulation in the decades to come, advances in data collection and analysis methods, along with renewed questioning of what spatial and demographic characteristics define the 'urban' as opposed to 'periurban' and 'rural', have led many to begin questioning the utility of these categorical indicators and whether they can truly measure relationships between urbanization and development outcomes. This is particularly true of comparative urban studies, which, more often than not, rely on census data that is markedly inconsistent internationally. Levels of urbanization in both India and China, for example, can be characterized well above or below 50% depending on the classification of 'urban' employed, posing serious questions about how to effectively utilize this data in policy formulation and planning efforts (see Satterthwaite, 2010; Zhang, 2004). The continued dissemination of information technologies has also lead many to question traditional differentiations between 'rural' and 'urban' given the fact that many populations living outside the traditional limits of cities can now earn the majority of their income via nonagricultural employment.

Moving beyond the demographic and spatial inconstancies in characterization of the 'urban', some headway is now being made in bridging the traditional, rigid differentiation between the literal and theoretical territories in which urban and 'nonurban' development articulations and trajectories play out. In her groundbreaking study of urban transformation in China, for example, You-Tien Hsing (2010) provides a strong historical case for a more holistic understanding of the development-urbanization nexus through her description of how regional, rural, and urban land-management policies and practices have combined to profoundly impact everything from human rights and land tenure, to agricultural development and ecological sustainability, to the nature of state formation and capital accumulation in modern China. Further broadening the theoretical landscape of development, the analytical framework of 'Planetary Urbanization' (Brenner and Schmid, 2011) aims to actively integrate the various spatial scales and diverse social, political, and economic feedback loops through which urbanization is transforming the human experience. It seems likely that in the coming decades those questions

emerging at center of the multidisciplinary study of development and urbanization will require increasingly dynamic understandings of the interconnected and fluid relationships between the city and its hinterlands, as well as the between regional, national, and global transformations that will accompany the continued process of urbanization.

See also: Attitudes, Political: Authoritarianism and Tolerance; Capitalism; Dependency Theory; Development Studies; Development: Social-Anthropological Aspects; Environment and Development; Industrialization; Migration and Development; Modernization, Sociological Theories of; Population Dynamics: Mathematic Models of Population, Development, and Natural Resources; Populism; State Formation, Theory of; Urbanization in Africa; Urbanization in China; Urbanization; World Systems Theory.

Bibliography

Abu-Lughod, J., Hay, R. (Eds.), 1977. Third World Urbanization. Maaroufa Press, Chicago, IL.

Abu-Lughod, J., 1989. Before European Hegemony: The World System AD 1250–1350. Oxford University Press, Oxford.

Angel, S., et al., 2011. Making Room for a Planet of Cities. Lincoln Institute of Land Policy, Cambridge, MA.

Armstrong, W., McGee, T.G., 1985. Theatres of Accumulation: Studies in Asian and Latin American Urbanization. Methuen, London.

Beall, J., 2002. Globalization and social exclusion in cities: framing the debate with lessons from Africa and Asia. Environment and Urbanization 14, 41–51.

Beauregard, R., 2003. City of superlatives. City & Community 2, 183–199.

Becker, K.F., 2004. The Informal Economy: Fact Finding Study. Department for Infrastructure and Economic Co-Operation, Sida.

Berry, B.J.L., 1961. City distribution and economic development. Economic Development and Cultural Change 9, 573–578.

Berry, B.J.L., 1973. The Human Consequences of Urbanization. St.Martin's, New York.

Breese, G., 1966. Urbanization in Developing Countries. Prentice-Hall, Englewood Cliffs, NJ.

Brenner, N., 2003. Stereotypes, archetypes and prototypes: three uses of superlatives in contemporary urban studies. City & Community 2, 205–218.

Brenner, N., Schmid, C., 2011. Planetary urbanisation. In: Gandy, M. (Ed.), Urban Constellations. JOVIS Publishers, Berlin.

Castells, M., Velez, P. (Eds.), 1971. Imperialismo y urbanizacion en America Latina. Gustavo Gili, Barcelona, Spain.

Clark, D., 1996. Urban World/Global City. Routledge, London.

Cornelius, W., 1973. Political Learning Among the Migrant Poor. Sage Publications, Beverly Hills, CA.

Davis, D.E., 2004. Discipline and Development: Middle Classes and Economic Prosperity in East Asia and Latin America. Cambridge University Press, Cambridge, UK.

Davis, D.E., 2005. Cities in global context: a brief intellectual history. International Journal of Urban and Regional Research 29 (1), 92–109.

Davis, K., 1969. World Urbanization, 1950–1970. University of California Press, Berkeley, CA.

Drakakis-Smith, D., 1987. The Third World City. Methuen, London.

Durkheim, E., 1933. The Division of Labor in Society. Macmillan, London.

Eckstein, S., 1977. The Poverty of Revolution: The State and the Urban Poor in Mexico. Princeton University Press, Princeton, NJ.

El-Shakhs, S., 1972. Development, primary, and system of cities. Journal of Developing Areas 7, 11–35.

El-Shakhs, S., 1974. Urbanization, National Planning, and Regional Development in Africa. Praeger Publishers, New York.

Fainstein, S., Gordon, I., Harloe, M., 1992. Divided Cities: New York and London in the Contemporary World. Oxford University Press, Oxford, UK.

Field, A. (Ed.), 1970. City and Countryside in the Third World: Issues in the Modernization of Latin America. Schenkman, Cambridge, UK.

Friedmann, J., 1967. A General Theory of Polarized Development. Ford Foundation, Santiago, Chile.

Friedmann, J., 1975. The Urban Transition: Comparative Studies of Newly Industrializing Countries. UCLA School of Architecture and Urban Planning, Los Angeles.

Friedmann, J., 1986. The world city hypothesis. Development and Change 17, 69–83.

Friedmann, J., Alonso, W., 1964. Regional Development and Planning. MIT Press, Cambridge, MA.

Friedmann, J., Wolff, G., 1982. World city formation: an agenda for research and action. International Journal of Urban and Regional Research 6, 309–344.

Fuller, B., Romer, P., 2010. Cities from scratch: a new path for development. City Journal 20 (4).

Gilbert, A., Ward, P., 1985. Housing, the State and the Poor: Policy and Practice in Three Latin American Cities. Cambridge University Press, Cambridge, UK.

Glaeser, E., 2011. Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier. Penguin Press, New York.

Goldscheider, C., 1971. Population, Modernization, and Social Structure. Little Brown, Boston, MA.

Goldstein, S., Sly, D., 1979. Patterns of Urbanization: Comparative Country Studies. Ordina Editions, Dolhain.

Gugler, J., 1978. Urbanization and Social Change in West Africa. Cambridge University Press, New York.

Gugler, J. (Ed.), 1988. The Urbanization of the Third World. Oxford University Press, New York.

Gugler, J. (Ed.), 2004. World Cities beyond the West: Globalization, Development and Inequality. Cambridge University Press, New York.

Hardoy, J., Geisse, G. (Eds.), 1972. Politicas de desarrollo urbano y regional en America Latina. SIAP, Buenos Aires, Argentina.

Hardoy, J., Satterthwaite, D., 1986. Small and Intermediate Centers: Their Role in Regional and National Development in the Third World. Westview Press, Boulder, CO.

Hasan, A., et al., 2005. How to meet the millennium development goals (MDGS) in urban areas. Environment and Urbanization 17 (1).

Hauser, P., 1957. Urbanization in Asia and the Far East. UNESCO Research Centre. India, Calcutta.

Hauser, P., Schnore, L. (Eds.), 1965. The Study of Urbanization. Wiley, New York.

Harvey, D., 2008. The right to city. The New Left Review 53.

Henderson, J., Castells, M., 1987. Global Restructuring and Territorial Development. Sage Publications, Beverly Hills, CA.

Herrick, B., 1965. Urban Migration and Economic Development in Chile. MIT Press, Cambridge, MA.

Hoyt, H., 1962. World Urbanization: Expanding Population in a Shrinking World. Urban Land Institute, Washington, DC.

Hsing, Y., 2010. The Great Urban Transformation. Oxford University Press, Oxford, UK.

Jones, G., Corbridge, S., 2010. The continuing debate about urban bias: the thesis, its critics, its influence and its implications for poverty reduction strategies. Progress in Development Studies 10, 1–18.

King, A., 1990. Global Cities: Postimperialism and the Internationalization of London. Routledge, London.

Knox, P., Taylor, P., 1995. World Cities in a World System. Cambridge University Press, Cambridge, UK.

Linn, J., 1983. Cities in the Developing World: Policies for Their Equitable and Efficient Growth. Oxford University Press, New York.

Lipton, M., 1977. Why Poor People Stay Poor: Urban Bias in World Development. Harvard University Press, Cambridge, MA.

Lo, F., Yeung, Y., 1998. Globalization and the World of Large Cities. United Nations University Press, Tokyo.

Mangin, W. (Ed.), 1970. Peasants in Cities: Readings in the Anthropology of Urbanization. Houghton Mifflin, Boston, MA.

Martine, G., et al. (Eds.), 2008. The New Global Frontier: Urbanization, Poverty and Environment in the 21st Century. Earthscan, London.

McGee, T.G., 1971. The Urbanization Process in the Third World. Bell, London.

McGreevey, W.P., 1971. A statistical analysis of primacy and lognormality in the size distribution of Latin American cities, 1950–1960. In: Morse, R. (Ed.), The Urban Development of Latin America: 1750–1920. Stanford University Press, Stanford, CA.

Olds, K., 2011. Globalization and the Asia-Pacific Region: Captial, Culture, and Pacific Rim Mega Projects. Oxford Univerity Press, Oxford, UK.

Perlman, J., 1976. The Myth of Marginality: Urban Poor and Politics in Rio de Janeiro. University of California Press, Berkeley, CA.

Pirenne, H., 1936. Economic and Social History of Medieval Europe. Harcourt, Brace, Jovanovich, New York.

Porter, M., 2003. The economic performance of regions. Regional Studies 37 (6–7), 549–578.

Portes, A., Walton, J., 1976. Urban Latin America: The Political Condition from Above and Below. University of Texas Press, Austin, TX.
Portes, A., Walton, J., 1981. Labor, Class, and the International System. Academic Press, New York.
Rabinowitz, F., 1973. National-local Linkages: The Inter-relationships of Urban and National Politics in Latin America. Sage Publications, Beverly Hills, CA.
Redfield, R., 1941. The Folk Culture of the Yucatan. University of Chicago Press, Chicago, IL.
Redfield, R., 1953. The Primitive World and its Transformations. Cornell University Press, Ithaca, NY.
Ren, Y., 2011. Building Globalization: Transnational Architectural Production in Urban China. University of Chicago, Chicago, IL.
Richardson, H., 1973. The Economics of Urban Size. Saxon House, Westmead, UK.
Roberts, B., 1978. Cities of Peasants: The Political Economy of Urbanization in the Third World. Arnold Press, London.
Rondinelli, H., 1983. Secondary Cities in Developing Countries. Sage Publications, Beverly Hills, CA.
Robinson, J., 2002. Global and world cities: a view from off the map. International Journal of Urban and Regional Research 26 (3), 531–554.
Robinson, J., 2006. Ordinary Cities: Between Modernity and Development. Routeledge, London.
Roy, A., AlSayyad, N., 2004. Urban Informality: International Perspectives from the Middle East, Latin America, and South Asia. Lexington Books, Langham, MA.
Roy, A., Ong, 2011. Worlding Cities: Asian Experiments and the Art of Being Global. Blackwell, Malden, MA.
Sassen, S., 1991. The Global City: New York, London, Tokyo. Princeton University Press, Princeton, NJ.
Sassen, S., 2000. Cities in a World Economy. Pine Forge Press, Thousand Oaks, CA.
Satterthwaite, D., 2007. The Transition to a Predominantly Urban Worlds and its Underpinnings. International Institute for Environment and Development, Human Settlements Discussion Paper Series. Theme: Urban Change – 4.
Satterthwaite, D., 2010. Urban myths and the mis-use of data that underpin them. In: Beall, J., et al. (Eds.), Urbanization and Development: Multidisciplinary Perspectives. Oxford University Press, UK.
Simmel, G., 1950. The Sociology of Georg Simmel. Free Press, New York.
Smith, D., 1996. Third World Cities in Global Perspective: The Political Economy of Uneven Urbanization. Westview Press, Boulder, CO.
de Soto, H., 2001. The mystery of capital. Finance and Development 31 (1).
Stren, R. (Ed.), 1995. Urban Research in the Developing World. Centre for Urban and Community Studies, University of Toronto, Canada.
Tilly, C., 1975. The Formation of National States in Western Europe. Princeton University Press, Princeton, NJ.
Timberlake, M. (Ed.), 1985. Urbanization in the World Economy. Academic Press, New York.
United Nations Department of Economic and Social Affairs, 2012. World Urbanization Prospects, the 2011 Revision. United Nations, New York.
Walton, J. (Ed.), 1985. Capital and Labor in the Urbanized World. Sage Publications, Beverly Hills, CA.
Weber, M., 1927. General Economic History. Greenburg Press, New York.
Weber, M., 1958. The City. Free Press, New York.
Yamauchi, F., et al., 2009. Human Capital Accumulation, Migration, and the Transition from Urban Poverty: Evidence from Nairobi Slums. International Food Policy Research Institute, Washington, DC.
Zhang, L., 2004. China's Limited Urbanization under Socialism and beyond. Nova Science Publishers, New York.

Development Geographies

Nina Laurie, School of Geography, Politics and Sociology, Newcastle University, Newcastle upon Tyne, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article outlines key strengths in development geography arguing for a plural understanding of development geographies rooted in different traditions, schools of thought and generated by place and history-specific interests and diverse policy and interdisciplinary forms of engagement. It highlights development geographies' contribution to debates on participatory methods and argues that current research funding regimes are likely negatively to affect how development research knowledges are being produced. This article outlines three productive interfaces within geography around postcolonial, urban, and political agendas in development which are helping to make sense of the rapidly changing global contexts of development and inequality.

What is Development Geography?

Development geography produces debate and analysis on international development and global inequality. The hallmark of work in this area is the elaboration of theory from sustained field engagement in specific Global South settings. Because of this it makes an important contribution to the wider discipline of Human Geography and other social sciences by emphasizing the need to bridge North–South divides and international distances in knowledge production about geographical themes. It is a well-established research subfield and a very popular part of the undergraduate curricula. Geography departments often house introductory courses for students from other disciplines as part of foundation training.

Like scholars working in the policy-focused discipline of Development Studies, many development geographers produce "issue-driven research concerning the analysis of global to local processes of cultural, demographic, economic, environmental, political, technological and social change in developing and emerging parts of the world." (Definition of Development Studies research taken from the Anthropology and Development Study Research Excellence Framework of the UK Higher Education Funding Council. www.ref.ac.uk/media/ref/content/pub/.../01_12.doc (accessed 13.08.13.).) They also make a distinct contribution to Development Studies with the introduction of concepts such as scale, connectivity, political ecology, socionatures, and postcoloniality, as I elaborate below. Many individual scholars also engage with specific Area Studies audiences, participating in international learned societies (e.g., Latin American Studies Association, African Studies Association, Association of South East Asian Studies) and publishing in their journals. The International Geographical Union (http://www.igu-online.org/site/) also provides an important global network for geographers that cuts across North–South–East–West. The presence of feminist scholars, in particular, in this network has shaped gender and development research since the publication of the first edited collection by geographers (Momsen and Townsend, 1987). The presence of development geography has also been institutionalized within the two largest national-based learned societies: the Developing Areas Research Group of The Royal Geographical Society with the Institute of British Geographers and the Development Geographies Specialty Group of the Association of American Geographers. While development geography is published in journals associated with these learned societies, the Singapore Journal of Tropics Geography, founded in 1953, represents a longer tradition of publishing research dedicated to the 'tropical world.' While the term 'tropical geography' is an 'amorphous' category associated with colonial geographical imaginaries (Savage and Sidaway, 2008: 236), the more generic label 'development geography' also comes with intellectual and political baggage which shapes how it is defined and received by wider academic communities. Linked to geography's own history as a discipline, which emerged from British colonial interests in 'distant' places and people, this complex story has shaped how development geography is seen by different audiences and in part explains its often ambivalent position as a subdiscipline in wider intellectual, institutional, and policy landscapes.

As the world changes and new global processes and power reconfigurations emerge, this history and context continues to provoke questions about what development is and whose development geography we are talking about. Below I outline some of the main strengths of development geography arguing for a plural understanding of development geographies rooted in different traditions, schools of thought and generated by place and history-specific interests. I point to the role that interdisciplinarity and a long track record in conducting engaged research have played in forging diverse development geographies. In the section on methodologies I highlight development geographies' contribution to debates on participatory methods and suggest some potential threats that current research funding regimes present for the ways in which development research knowledge is being produced. I end by highlighting three productive interfaces within geography around postcolonial, urban, and political agendas in development which are helping to make sense of the rapidly changing global contexts of development and inequality.

Development Geographies

Treated with caution because of geography's colonial origins and characterized in the past, by some, for producing

a theoretical case study research, development geography has undergone a transformation over the last decade. It is now recognized as an intellectual home for fruitful cross-discipline collaborations that responds to the diverse nature of contemporary development challenges. As a result, current work is shaping development debates on the 'big issues' such as climate change (Pelling et al., 2001), poverty (Rigg et al., 2009), and the changing world order (e.g., Mawdsley, 2012). In this context different 'development geographies' are emerging to capture the breadth of this field both in terms of epistemological and methodological approaches as well as thematic and empirical focus. Existing strengths in development geography are being deepened and new research directions are emerging encompassing geographers across all geography subdisciplines including economic geography (see Peck and Sheppard, 2010). Academics based in Development Studies, Planning and Sociology increasingly publish on development topics in Geography journals, while a Scopus search indicates that articles by geographers ranked in the top 50 downloaded papers in leading Development Studies journals for the period 2006–12 (see **Table 1**). Papers by geographers ranked first and second place in World Development (Blaikie, 2006; Rigg, 2006). Together with Bebbington (1999) these papers appear in the list of top 25 downloads in World Development over the last decade.

Calls for more critical thinking about what precisely is meant by 'development geography' as opposed to 'geography in developing countries' (Simon, 2011) have been at the heart of moving geography and wider social science debates away from thinking about the Global South/North as separate. From the 1990s onwards this agenda harmonized particularly well with the growing interest in globalization. As a result, a strong body of research emerged around globalization-related themes of transnationalism and mobility that built on long-established interests in the connections between places. In particular research has examined the development implications of boundaries becoming more fluid with globalization and the repercussions of increased mobility for how development is understood and practiced. Attention has been drawn to the limits of crisis-driven narratives in the 'age of migration' (Silvey, 2009: 508) to focus more positively on the role of circular migrants (Potts, 2010) and diasporas (Mercer et al., 2008; McIlwaine, 2011) in brokering and imagining development in specific ways that emphasize the myriad social, cultural, and economic connections between places across distance. This includes work on migration as a form of adaptation to climate change (Black et al., 2011) while related research points to how Information and Communications Technologies (ICTs) for development (ICT4D) are potentially able to bridge digital divides and promote sustainable development (Kleine and Unwin, 2009).

Geographical research on globalization has also been characterized by interrogating the relationship between globalization and neoliberalism thereby providing a robust critique of the global power relations of development. As a result, research on structural adjustment and its key characteristic elements (privatization and the switch to export commodity production) is a hallmark of much development geography. Particularly strong scholarship has emerged on Latin America and the Caribbean (Kohl, 2004; Perrault and Martin, 2005; Klak, 1998). Critical attention has also been paid to the rise of neoliberal ideas about how to 'manage' development with research focusing on the growing audit culture, professionalization processes as well as scrutiny of the means through which such ideas travel and become consolidated globally (Townsend et al., 2004; Larner and Laurie, 2010).

Work on neoliberal natures is also currently shaping agendas in development geography building on a rich body of work in political ecology, one of the most significant contributions to development research more generally in the last few decades. Bill Adams' land mark Green Development (2008) is in its third edition, Peet and Watts' (2004) highly influential Liberation Ecologies is in its second, and the foundational of work of Piers Blaikie has been profiled in journal special issues (e.g., Geoforum ('In honor of the Life Work of Piers Blaikie' Special Issue in Geoforum 39:2, March 2008. with nine articles including 'Epilogue: Toward a future for political ecology that works' (765–772) by Piers Blaikie)). Research in this area brings livelihoods together with political ecology in a way that helps rethink rural development in Asia (Rigg, 2006), Africa and Latin America (Bassett and Zimmerer, 2003) and provides new frameworks for thinking about natural resources (Watts, 2004). Indicating the reach of this work in geography more widely, the first two annual keynotes of the Cultural and Political Ecology Specialty Group at the annual conference of the Association of American Geographers were given by Tony Bebbington and Piers Blaikie in 2010 and 2011.

Engagement, Influence, and 'Impact'

The nonacademic influence of development geography research has changed policy making, development practice, and debate. Landmark books on Political Ecology (see above) and Participatory Development (see below), for example, have changed the way in which development issues are imagined and tackled, influencing nonacademic development stakeholders at all scales. Illustrating strengths in political ecology, urban research, and gender and development, examples of recent high-level engagements by development geographers include working with different United Nations agencies. Development geographers have authored and acted as commissioning editors for a number of recent UN HABITAT 'state of the world' reports (e.g., UNHABITAT, 2010 (*State of African Cities 2010* Professor David Simon, Royal Holloway, University of London State of); UNHABITAT, 2012 (*Women in Cities: Gender and the Prosperity of Cities 2012–13* Professor Sylvia Chant, London School of Economics)); Moser and McIlwaine's (2006) work on gender and violence has shaped the World Bank thinking by providing a typology of economic, social, and political violence perpetrated at different scales – individual, community, city etc.; Bebbington's work on civil society has also been very influential in wider international donor policy circles, especially in terms of highlighting the relationships between different types of civil society organizations and the state. A joint publication with World Bank colleagues on this work (Bebbington and Dharmawan et al., 2006) is one of the top downloaded articles over the last decade in world development.

Table 1 Articles by geographers in leading international peer-reviewed development journals 2006–12

Journal title	*No. articles*	*Top ranking*	*Top ranking article title*
Development and Change	5	12	Jønsson, J. B., Bryceson, D.F., 2009. Rushing for gold: mobility and small-scale mining in East Africa. Development and Change 40 (2), 249–279.
Journal of Development Studies	7	2	Bennett, R., Morse, S., et al., 2006. The economic impact of genetically modified cotton on South African smallholders: yield, profit and health effects. Journal of Development Studies 42 (4), 662–677.
Progress in Development Studies	12	2	McFarlane, C., 2006. Knowledge, learning and development: A post-rationalist approach. Progress in Development Studies 6 (4), 287–305.
Third World Quarterly	6	5	McFarlane, C., 2006. Crossing borders: development, learning and the north–south divide. Third World Quarterly 27 (8), 1413–1437.
World Development	15	1	Blaikie, P., 2006. Is small really beautiful? Community-based natural resource management in Malawi and Botswana. *World Development* 34 (11), 1942–1957.
Gender and Development	2	22	Vullnetari, J., King, R., 2011. Gendering remittances in Albania: a human and social development perspective. *Gender and Development* 19 (1), 39–51.
Development in Practice	3	3	Oslender, U., 2007. Violence in development: the logic of forced displacement on Colombia's Pacific coast. *Development in Practice* 17 (6), 752–764.
Bulletin of Latin American Research	6	2	Gilbert, A., 2006. Good urban governance: evidence from a model city? Bulletin of Latin American Research 25 (3), 392–419.

Source: SCOPUS search conducted by author in March 2012 analyzing the top 50 cited articles in the named journals for this period.

Influence on bilateral donor organizations has also been achieved in a number of ways with geographers contributing to 'national flagship' areas of international development expertise. In Europe for example this includes the Norwegian Agency for Development Co-operation's profile in post conflict humanitarianism and development and Denmark's Development Organization's expertise in youth and development. (Drawing on the work of scholars such as Professor Kristian Stokke (accessed 13.08.13.) http://www.sv.uio.no/iss/english/people/aca/stokke/and Professor Kate Gough: http://www.lboro.ac.uk/departments/geography/staff/gough-k.html (accessed 13.08.13).) In the UK the Department for International Development's (DfiD) sustainable livelihoods framework draws on the political ecology research of Blaikie and others which linked concepts of vulnerability and risk to natural disasters and rural poverty (Wisner et al., 1994). Stuart Corbridge et al.'s work on India (2005) also influenced the UK government thinking on state formation through engagement with the Crisis States Group at DfID. While these examples illustrate how individuals have shaped policy through advisory roles and research dissemination, specific funding streams have also been created in recent years to enhance the academic–policy interface. The UK's 'ESRC-DfiD joint fund for poverty alleviation research' scheme launched in 2009 is an emblematic case through which a number of geography-based projects have been funded. (For more details of the ways in which UK development geographers have engaged with development donors and stakeholders see the review of UK development geography for the ESRC-AHRC international benchmarking review of Human Geography: http://www.rgs.org/OurWork/Research+and+Higher+Education/International+benchmarking+review+of+UK+human+geography.htm.)

While engagement with large donor organizations produces policy-focused development geographies that speak to the global development industry, diverse development geographies are also produced through research engagement with a plethora of grassroots organizations and NGOs operating in the Global South. This engagement often brings direct benefit to some of the poorest members of society. For example, Tony Bebbington's research on mining has been credited by the Peru Support Group for ensuring a UK company paid compensation to farmers for human rights violations. Dorothea Kelin's research on Fair Tracing (www.fairtracing.org), which used smartphones to give consumers more traceability information, persuaded the UK's leading ethical consumer magazine to move their online ethiscore subscription service for shoppers to online phones. This type of impact emerges from a politics of knowledge production that aims to be about working 'with' and not 'on' the Global South. A wide range of innovative forms of working, partnership, participatory practices, and coproduction with local organizations underpins such an approach to research, as we discuss in more detail below. Nevertheless it is important to recognize that the term 'impact' itself is controversial, for its implied unilinear, one-way effect which often makes invisible the coproduction of development knowledges and privileges development outcomes that can only be measured in restricted ways. The recent rollout of an impact agenda by research funding bodies in many settings continues to generate debate among development geographers (Williams, 2012).

Methodology: Participation, Collaboration, and Coproduction

The book *Participation: the New Tyranny* (Cooke and Kothari, 2001) and its follow-on companion text *Beyond Participation: Strategies for Deeper Empowerment* (Mohan, 2004) shifted how participation issues are framed and taught in contemporary

development theory and practice globally. This discussion developed in part from the valuable cross-over between Geography and Development Studies outlined above and from strengths in reflexive methodological research practice in development geography. (Both Professor Uma Kothari and Professor Giles Mohan are geographers based in Development Studies Units in the UK.) This background served as a springboard for both books' critique of the ways in which participation has entered development planning and practice, often reflecting 'tick box' approaches that are more akin to consultation with 'aid recipients' rather than genuine participatory engagement with grassroots development actors around empowerment agendas.

Feminist development geographers, in particular, have been long-time champions of grassroots empowerment and participatory approaches in development geography, resulting in the coproduction of knowledge with marginalized groups and Global South NGOs in many cases. Laurie et al.'s (forthcoming) work with an NGO founded by trafficked women in Nepal (www.shaktisamuha.org.np (accessed 13.08.13.)) provides a recent example of such an approach (see http://www.posttraffickingnepal.co.uk/ (accessed 13.08.13.)). For feminist development geographers based in North America the coproduction of knowledge with Global South feminist activists has often been facilitated by strong links between Geography and Women's Studies. Examples include research with the women's organization 'Red Thread' (https://www.facebook.com/RedThreadWomenGuyana/info (accessed 13.08.13.)) in Guyana (Peake and Trotz, 1999) and the work of the Sangtin Writers with Richa Nagar (2006) focusing on feminist thought and activism in India. These sets of works are based on long-term working partnerships and collaborations with specific NGOs and a commitment to a feminist politics of knowledge production. In these ways feminist geography makes important contributions to development debates on empowerment through collaborations with Global South partners including in non-English-speaking environments. The feminist geography journal *Gender Place and Culture*, which publishes much development geography, was a pioneer among geography journals in introducing multilanguage abstracts, indicating a commitment to making research more accessible internationally. Feminist-engaged research on development also influences international policy making through strategic collaboration with specific international donors resulting in the sorts of coproduced policy outputs on gender and violence identified in the previous section.

While there is much to praise in development geography's emphasis on participatory methods and the coproduction of knowledge through collaborations that cross North–South, current shifts in research funding represent a number of threats to the ways in which development geographies are produced. This includes the ways in which full economic cost models potentially undermine strengths in ethnographic research and the fact that low academic wages in Global South universities generate consultancy driven systems which provide little time and space for academics to reflect on their field experiences. While it is argued that ICT4D offer opportunities to bridge the digital divide, scholars and activists based in the Global South continue to face structural biases and inequalities in terms of publishing in international peer-reviewed journals and in accessing journals generally, whether online or in hard copy. It remains unclear the extent to which current debates about open-access publishing protocols will resolve or deepen these problems. Together with the increasing costs of postgraduate fees, especially for international students and the poor funding for thesis completion at undergraduate and postgraduate levels in many countries in the Global South, these factors mean that the capacity to build long-term, engaged research collaboration across difference and distance is diminishing.

New Development Actors and Emerging Academic 'Interfaces'

Contemporary changes in the economic world order are demanding new conceptual tools for understanding development. While development geography has traditionally focused on grassroots organizations, NGOs, the state, international agencies and, to a lesser extent, religious organizations, research has recently shifted to embrace a range of emerging development actors and processes. These include for example, BRIC economies as donors and investors (Mawdsley, 2012) as well as children and youth (Jones, 2011) and indigenous people (Andolina et al., 2009) as authors of 'alternative' development logics. Social justice movements (Cumbers et al., 2008), international volunteers (Baillie Smith and Laurie, 2011), celebrities (Brockington, 2009), and ethical consumers (Goodman, 2010) have also become new research subjects. A growing body of work examines how they act as development brokers and interpreters when they engage with and performing diverse forms of global citizenship.

Much of this work on new actors has been inspired by the lead that development geography has taken in theorizing postcolonialism in relation to different aspects of development (Power et al., 2006). Research has examined postcolonial development at its interfaces with culture (McEwan, 2001), economy (Pollard et al., 2009), and (geo)politics (Noxolo, 2006). In this way innovative research has interrogated identity making and the performative construction of (diverse) economies, politics, and knowledges including ethical, alternative, and fair trade. As a result, development geography has forged new agendas and innovative interfaces with urban and political geographies that have sought to 'provincialize' (Chakrabarty, 2000) the so-called metropolitan thinking.

For urban geographers writing of and from contexts like South Africa, where the analyses of urban space has led to compelling arguments to reformulate the Global South as comprising 'Ordinary Cities' (Robinson, 2006), the relevance of development labeling is something to call into question. Raising such questions, however, can provoke strong reactions, as Sue Parnell makes clear in some personal reflections on her forthcoming book (Parnell and Oldfield, 2014). Here she argues that the current global context, in which neoliberal policies are affecting cities globally, requires a rethinking of urban theory:

> I have spent much of the past 7 months presenting talks on (re) theorizing cities from the Global South in intellectual centers of the Global North - London, Geneva, New York, Paris, Manchester and Edinburgh. I have been surprised at the often intense emotional

reaction that the suggestion, premised on the facts of the global reconfiguration of population and economy (to urban Africa, Latin America and Asia) and the distortion of academic knowledge production, that there is an imperative for theoretical and empirical recalibration in urban studies.

In its crudest form the push back or negation of the call for a more Southern urban studies is an objection to the suggestion that work undertaken in Europe or North America is no longer cutting edge or universally relevant. There is also an austerity tinged objection, based on the erroneous conflation of the term 'the Global South' with that of poverty and exclusion, that objects to the suggestion that cities in the Global North have no serious problems and that inequality and deprivation are the preserve of the South.

Speaking at a panel session 'Theorising from the global South: Inverting and challenging geographies of knowledge production,' RGS-IBG annual conference, Edinburgh, July 3–5, 2012, session organized by Charlotte Lemanski and Katherine V. Gough.

Her interpretations of what the reactions to her analysis imply about geographical imaginaries of development, poverty, and knowledge production are important. At another level, the response that her call for new ways of thinking has provoked can also been seen as indicative of how changing global circumstances are generating ontological insecurities within the discipline. The realignment of global power and changing economic and political configurations are pushing geographers to rethink their geographical maps (Sidaway, 2012) and to reexamine 'the politics – personal and professional' of "parachuting into new and unfamiliar terrain armed with little understanding of the specificities of 'other' places and with only Western theoretical toolkits" (Pollard et al., 2009: 138). In development geography these types of anxieties have led to a reevaluation of the contribution of empirically rich, historically grounded ethnographic analyses of development (Silvey and Rankin, 2012) inspired by postcolonial approaches which aim to speak across North–South (McFarlane, 2006) and disciplinary differences (Rankin, 2003).

Conclusion

Development geography has transformed itself over the last decade. Responding to calls for more critical thinking about what development is and means in the contemporary world, it is providing an important analysis of the key challenges the world is currently facing. It is also making visible the geographies and politics of knowledge production that shape why we see these challenges in particular ways.

What constitutes development geographies and their relevance to wider debates about international development and global inequality depend very much on where a researcher is looking from: what their engagement is with that place, its connections, its people, different schools of thought, and their own biographies (academic/activist/personal). It is only possible to understand the ambivalence and sometimes tension around the use of specific ideas and labels, if these factors are taken into account and given critical consideration. Unequal power relations mean that exclusions, namings, and silences play an important role in knowledge production around development. These factors influence which academics, from where, become labeled as what, and listened to by whom. Development geographers aim to make these issues visible in their work. In this way they help ensure that important questions about whose development geography continue to be forefront in analyses of global inequality.

Acknowledgments

I am grateful to Tony Bebbington, Ruth Pearson, Rhiannon Redpath, Jonathan Rigg, Katie Willis, and members of the Developing Areas Research Group of the Royal Geographical Society with the Institute of British Geographers, who helped gather information and provided feedback on some of the ideas developed here.

See also: Development Theory in Geography; Gender and Feminist Studies in Geography; Globalization: Geographical Aspects; Political Ecology: Anthropological Perspectives; Poverty, Geography of.

Bibliography

Adams, W.M., 2008. Green Development: Environment and Sustainability in a Developing World. Routledge, London.

Andolina, R., Laurie, N., Radcliffe, S., 2009. Indigenous Development in the Andes: Culture, Power and Transnationalism. Duke University Press, Durham, NC.

Baillie Smith, M., Laurie, N., 2011. International volunteering and development: global citizenship and neoliberal professionalisation today. Transactions of the Institute of British Geographers 36 (4), 545–559.

Bassett, T.J., Zimmerer, K., 2003. Political Ecology: An Integrative Approach to Geography and Environment – Development Studies. Guilford Press, New York.

Bebbington, A., 1999. Capitals and capabilities: a framework for analyzing peasant viability, rural livelihoods and poverty. World Development 27 (12), 2021–2044.

Bebbington, A., Dharmawan, L., et al., 2006. Local capacity, village governance, and the political economy of rural development in Indonesia. World Development 34 (11), 1958–1976.

Black, R., Bennett, S.R.G., Thomas, S.M., Beddington, J.R., 2011. Climate change: migration as adaptation. Nature 478, 477–479.

Blaikie, P., 2006a. Is small really beautiful? Community-based natural resource management in Malawi and Botswana. World Development 34 (11), 1942–1957.

Blaikie, P., 2006b. Epilogue: towards a future for political ecology that works. Geoforum 39 (2), 765–772.

Brockington, D., 2009. Celebrity and the Environment. Fame, Wealth and Power in Conservation. Zed Books, London.

Bryceson, D.F., Fonseca, J., 2006. Risking death for survival: peasant responses to hunger and HIV/AIDS in Malawi. World Development 34 (9), 1654–1666.

Chakrabarty, D., 2000. Provincialising Europe: Postcolonial Thought and Historical Difference. Princeton University Press, Princeton, NJ.

Cooke, B., Kothari, U. (Eds.), 2001. Participation: The New Tyranny?. Zed Books, New York.

Corbridge, S., Williams, G., Srivastava, M., Verón, R., 2005. Seeing the State: Governance and Governmentality in India. Cambridge University Press, Cambridge.

Cumbers, A., Routledge, P., Nativel, C., 2008. The entangled geographies of global justice networks. Progress in Human Geography 32 (2), 183–202.

Goodman, M.K., 2010. The mirror of consumption: celebritization, developmental consumption and the shifting cultural politics of fair trade. Geoforum 41 (1), 104–116.

Jones, G.A., 2011. Children and development III: making space for young people. Progress in Development Studies 11 (2), 145–149.

Kleine, D., Unwin, T., 2009. Technological revolution, evolution and new dependencies: what's new about ICT4D? Third World Quarterly 30 (5), 1045–1067.

Klak, T. (Ed.), 1998. Globalization and Neoliberalism: The Caribbean Context. Rowman & Littlefield Publishers, Lanham, MD.

Kohl, B., 2004. Privatization Bolivian style: a cautionary tale. International Journal of Urban and Regional Research 28 (4), 893–908.

Larner, W., Laurie, N., 2010. Travelling technocrats, embodied knowledges: globalising privatisation in telecoms and water. Geoforum 41, 218–226.

Laurie, N., Richardson, D., Poudel, M., Shakti Samuha, Townsend, J.G. Anti-trafficking activism: collaborating on transforming citizenship. Development in Practice, forthcoming.

Momsen, J.H., Townsend, J.G. (Eds.), 1987. Geography of Gender in the Third World. Hutchinson Education, London.
Mawdsley, E., 2012. From Recipients to Donors: Emerging Powers and the Changing Development Landscape. Zed Books, London.
McEwan, C., 2001. Postcolonialism, feminism and development: intersections and dilemmas. Progress in Development Studies 1 (2), 93–111.
McFarlane, C., 2006. Crossing borders: development, learning and the north–south divide. Third World Quarterly 27 (8), 1413–1437.
McIlwaine, C., 2011. Cross-Border Migration Among Latin Americans: European Perspectives and Beyond. Palgrave Macmillan, New York.
Mercer, C., Page, B., Evans, M. (Eds.), 2008. Development and the African Diaspora: Place and the Politics of Home. Zed Books, London.
Mohan, G., 2004. Beyond Participation Strategies for Deeper Empowerment. Mendely, London.
Moser, C.O.N., McIlwaine, C., 2006. Latin American urban violence as a development concern: towards a framework for violence reduction. World Development 34 (1), 89–112.
Noxolo, P., 2006. Claims: a postcolonial geographical critique of 'partnership' in Britain's development discourse. Singapore Journal of Tropical Geography 27 (3), 254–269.
Parnell, S., Oldfield, S., 2014. Handbook on Cities of the Global South. Routledge, London.
Peck, J., Sheppard, E., 2010. Worlds apart? Engaging with the World Development Report 2009: reshaping economic geography. Economic Geography 86 (4), 331–340.
Peake, P., Trotz, A., 1999. Gender, Ethnicity and Place: Women and Identities in Guyana. Routledge, London.
Peet, R., Watts, M.J. (Eds.), 2004. Liberation Ecologies, second ed. Routledge, London.
Pelling, M., Uitto, J.I., 2001. Small island developing states: natural disaster vulnerability and global change. Environmental Hazards 3 (2), 49–62.
Perreault, T., Martin, P., 2005. Geographies of neoliberalism in Latin America. Environment and Planning A 37 (2), 191–201.
Pollard, J.S., McEwan, C., Laurie, N., Stenning, A., 2009. Economic geography under postcolonial scrutiny. Transactions of the Institute of British Geographers 34, 137–142.
Potts, D., 2010. Circular Migration in Zimbabwe and Contemporary Sub-Saharan Africa. James Currey, Oxford.
Power, M., Mohan, G., Mercer, C., 2006. Postcolonial geographies of development: introduction. Singapore Journal of Tropical Geography 27, 231–234.
Rankin, K.N., 2003. Anthropologies and geographies of globalization. Progress in Human Geography 27 (6), 708–734.
Rigg, J., 2006. Land, farming, livelihoods, and poverty: rethinking the links in the rural south. World Development 34 (1), 180–202.
Rigg, J., Bebbington, A., Gough, K.V., Bryceson, D.F., Agergaard, J., Fold, N., Tacoli, C., 2009. The World Development Report 2009 reshapes economic geography: geographical reflections. Transactions of the Institute of British Geographers 34, 128–136.
Robinson, J., 2006. Ordinary Cities: Between Modernity and Development. Routledge, London.
Sangtin, W., Nagar, R., 2006. Playing with Fire: Feminist Thought and Activism through Seven Lives in India. University of Minnesota Press, Minneapolis (Also published as a South Asian edition in 2006 by New Delhi: Zubaan Books with all authors identified as Sangtin Writers and translated into Turkish 2011. Turkish edition published by Ankara: Ayizi Kitap as Ateşle Oynamak: Hindistan'da Yedi Yaşam Üzerinden Feminist Düşünce ve Eylem).
Savage, V., Sidaway, J., 2008. The Singapore journal of tropical geography: 1953 onwards. Scottish Journal of Geography 124 (4), 236–240.
Sidaway, J., 2012. Geographies of development: new maps, new visions? The Professional Geographer 64 (1), 49–62.
Simon, D. (Ed.), 2011. Symposium: Geographers and/in Development. Environment and Planning A, vol. 43, pp. 2788–2800.
Silvey, R., 2009. Development and geography: anxious times, anemic geographies, and migration. Progress in Human Geography 33 (4), 507–515.
Silvey, R., Rankin, K., 2012. Historical ethnography and critical development studies in geography: a dialogue. Annals of the Association of American Geographers 102 (2), 499–500.
Tan-Mullins, M., Rigg, J., Law, L., Grundy-Warr, C., 2007. Re-mapping the politics of aid: the changing structures and networks of humanitarian assistance in post-tsunami Thailand. Progress in Development Studies 7 (4), 327–344.
Townsend, J.G., Porter, R.E., Mawdsley, E.M., 2004. Creating spaces of resistance: development NGOs and their clients in Ghana, India and Mexico. Antipode 36 (5), 871–899.
UNHABITAT, 2010. The State of African Cities. UNEP, Nairobi.
UNHABITAT, 2012. State of Women in Cities: Gender and the Prosperity of Cities 2012–13. UNEP, Nairobi.
Williams, G., 2012. The disciplining effects of impact evaluation practices: negotiating the pressures of impact within an ESRC–DFID project. Transactions of the Institute of British Geographers 37 (4), 489–495.
Wisner, B., Blaikie, P., Cannon, T., Davis, I., 1994. At Risk: Natural Hazards, People's Vulnerability and Disasters, second ed. 2004. Routledge, London.
Watts, M.J., 2004. Resource curse? Governmentality, oil and power in the Niger Delta, Nigeria. Geopolitics 9 (1), (Special issue).

Relevant Websites

www.fairtracing.org – Engineering and Physical Sciences Research Council (EPSRC) of the UK funded project to support Ethical Trade by implementing IT Tracking and Tracing Technologies in supply chains to provide consumers and producers with enhanced information. (accessed 13.08.13.).

http://www.posttraffickingnepal.co.uk – Economic and Social Research Council (ESRC) of the UK funded project on post-trafficking livelihoods and development (accessed 13.08.13.).

http://www.igu-online.org/site/ – The International Geographical Union (accessed 13.08.13.).

http://www.rgs.org/OurWork/Research+and+Higher+Education/International+benchmarking+review+of+UK+human+geography.htm – Joint Economic and Social Research Council (ESRC) and Arts and Humanities Research Council (AHRC) International Bench Marking Review of Human Geography (accessed 13.08.13.).

www.shaktisamuha.org.np – NGO founded by trafficked women in Nepal (accessed 13.08.13.).

https://www.facebook.com/RedThreadWomenGuyana/info – NGO targeting gender, race and violence in Guayana (accessed 13.08.13.).

www.ref.ac.uk/media/ref/content/pub/.../01_12.doc – Research Excellence Framework of the UK Higher Education Funding Council's definition of Development Studies (accessed 13.08.13.).

Development: Organizational

Georg Schreyögg, Freie Universität Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Organizational Development (OD) is a field of applied social sciences devoted to understanding and facilitating change processes in organizations. Techniques to overcome resistance to change and methods for managing the renewal process in organizations effectively are at the very core of this discipline. While OD interventions address various organizational levels (individual, group, department, organization), they usually aim at changing the target system on both dimensions: structure and behavior. Various OD programs have been developed, among them the most salient are Survey–Feedback and Process Consultation. OD has been criticized for ignoring the power relations in organizations, for misreading resistance to change behavior, and for misconceiving change processes as exceptional episodes. More recent approaches reconceptualize change as a continuous feature in organizational life, thereby exaggerating the fluidity of social institutions. Concluding remarks advocate a view of mastering the delicate balancing of stability requirements and change challenges.

Historical Overview

In classical organization theory, the problem of coping with organizational change was not of much concern. Change basically meant reorganization and the critical activity was to develop the best organizational solution for mastering new situations brought about by diversification, mergers and acquisitions, internationalization, factory closure, etc. Implementing the new design, if regarded as a problem at all, was conceived as a problem of careful preparation and of giving the right instructions. After a brief period of tolerance, all employees were supposed to work according to the new organizational guidelines.

This traditional model, however, rarely proved to be a success: organizational members notoriously resisted the new organizational design, unforeseen events rendered the new design worthless, or the change process came to nothing.

The typical reaction to these irritations was to blame the plans for having been imperfect and to push harder for renewal. In most cases this reaction pattern actually made the situation even worse; the resistance became stronger, the rejections more sophisticated, etc. Although the problems of organizational change made themselves strongly felt in practice, for a long time they did not gain the status of a critical issue in classical theory.

It was the behavioral sciences, especially the Human Resources Approach, that put organizational change on the research agenda and developed approaches for understanding the texture of change processes.

The impetus to establish organizational change as a special research field came from empirical studies finding that success in change critically depends on the attitudes of organizational members as well as on the general mood toward the new design. This insight was provided by studies on 'resistance to change' (Lawrence, 1954).

Studying resistance to change, as well as approaches on how to overcome it, dates back to Lewin's (1958) experiments. Most salient are his seminal experiments on changing deep-seated food aversions and avoidance. Group meetings and active involvement of group members proved to be the best way to change food habits. The foundations for what were later to become the 'golden rules' of successful organizational change have been laid in these studies:

1. Broad information at an early stage and active participation in the change process.
2. The group as an important medium (facilitator) for change: change processes are perceived as less worrying in groups and generally are accepted more easily.
3. Mutual cooperation among organizational members.
4. A settling-down period at the end of the change process designed to institutionalize the changes that were carried out.

Lewin (1958, p. 210) later elaborated on these insights and conceived the well-known 'triadic episode' of successful change. The model builds on the idea of homeostasis and advocates a three-stage process: unfreezing, moving, and refreezing (see Figure 1).

The most important insight of this model is that organizations need an initial period of irritation and relaxation to enable them breaking up their habits and routines. Or as Lewin put it, organizations have to undergo a stage of 'unfreezing'; otherwise, they are not likely to leave their equilibrium position and to move to a new level. Changes tend to be short lived if there are no efforts to bring them to permanency. The new patterns, therefore, need 'refreezing'; that is, the system has to settle down again and find a new equilibrium.

Organizational Development as a Discipline

Lewin's group experiments and his three-stage model brought about a totally new form of behavioral training (Laboratory Method) designed to cope with habits and changes. This new technology of group intervention and feedback soon became very popular and built the core of the newly founded National Training Laboratories in Group Development. A series of further related approaches emerged. As a result, facilitating and supporting organizational change processes increasingly became a management consulting profession. All these studies and initiatives eventually gave rise to establishing a new field, special discipline within behavioral sciences, called Organizational Development (OD).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.73079-0

Figure 1 Lewin's triadic episode of organizational change.

Among the other theoretical and practical streams of OD, the 'Survey–Feedback' approach features prominently. This approach makes use of attitude surveys and data feedback to organizational members as a means of facilitating change processes (Bowers and Franklin, 1976). All organizational members are invited to take part in the change process including data collection and analysis as well the design of change initiatives (shared power approach). The intention is to stimulate organizational members' interest in change and to create a motivation for overcoming the current state and its weaknesses. The general aim is to initiate a joint problem-solving process. Research on the effectiveness of Survey–Feedback intervention found that the greater the involvement of organizational members in the survey process, the greater was the awareness of problems in the organization.

Another closely related OD building stream was the work of The Tavistock Institute (UK) which is deeply rooted in psychotherapy and group therapy (Bion, 1961). Tavistock started to transfer clinical methods to organizational change settings in the early 1950s. The doctor–patient interaction pattern and the corresponding high professional demands put on the change agents stand for this conception. Among the different methods used in this context the so-called Tavistock Conferences feature prominently. These are leaderless group discussions designed to experience and to explore leadership, authority, and obedience.

A further impetus in building OD came from approaches to redesign jobs and to the improvement of the quality of working life. Here again, The Tavistock Institute (of Human Relations) played a major role by conducing the sociotechnical systems approach. That is, a conceptual framework and an action research program which is devoted to meet both individual needs and technical demands (Rice, 1958). The emphasis is on changing traditional Tayloristic organizational design toward group-oriented work designs providing organizational members more variety, meaning, and feedback in their work.

Later, the scope of OD and the repertoire of OD interventions were extended significantly. Among them the conceptions of organizational learning, corporate culture, and sensemaking play a major role. Also, OD has been linked to strategic management and strategy renewal. All the additional approaches and new issues considerably enriched the field but made it at the same time more diverse. At the core of OD, however, there still remain theories of change and interventions devoted to effectuate change at different levels of social systems.

OD: Conception

OD is truly interdisciplinary in both focus and methods. Even though different perspectives exist on what OD is and should be, some overarching characteristics can nevertheless be identified (Cummings and Worley, 2008):

1. *Holistic:* OD is a holistic approach; it seeks to change both structure and behavior. While OD aims at changing the entire system, the interventions address different levels: individual, group, intergroup relations, department, division, and interorganizational relations.
2. *Planned organizational change:* OD conceives organizational change as a planned, systematic activity building on a longer-term perspective.
3. *Based on social sciences:* OD methods and procedures apply knowledge from social science research; it is explicitly science based.
4. *Full cycle:* OD defines change as a project and aims at including all stages of a change project: its initiation (unfreezing), its realization (moving), and the subsequent stabilization (refreezing).
5. *Change agent:* OD projects are conducted by external professional consultants who are trained in applying sophisticated intervention techniques and clinical methods. Organizations are assumed to often misread critical signals and to stick closely to their traditional mind-sets which easily amount to invisible barriers to change. They therefore need professional help from outside for reanalyzing the problems at stake and redefining the setting in innovative ways (reframing).
6. *Top management support:* Commitment of the top management group is considered a precondition of any successful OD project. A commitment is expected in both paying attention to the change problems and giving the OD intervention a sense of urgency.
7. *Shared power approach:* OD processes are designed to involve all parties affected by the change project and to encourage participants to jointly develop new solutions. Authority is still present but there is a willingness to share power. Organizational members have a voice in the process and are expected to become committed to the change project. They share their knowledge.
8. *Integration:* OD projects aim at integrating the individual and the organization by reaching both organizational effectiveness and fulfillment of individual needs.

From time to time, new themes were additionally included. The inclusion of these parts did, however, not result in a paradigm shift or an essential theoretical revision. Up to now, the overall framework has basically remained the same.

OD Models

Various authors and consultants have developed comprehensive OD programs; among them four models are salient: (1) Survey-guided Development, (2) Process Consultation, (3) Grid Organization Development, and (4) Systemic Intervention.

The Survey-Guided Development

This model is based on a multistep data feedback procedure. After an orientation stage a survey is conducted using a standardized questionnaire to gain a comprehensive diagnosis of the client organization's situation, its major problems, and

their possible causes. The diagnosis is based on a theoretical model of organizational functioning as a frame of reference (Likert, 1967). The emphasis is on gaining a solid and rigorous database, which allows for developing systematically further activities. Starting points are discrepancies between the model of organizational functioning and the observed and measured status quo (the conception comes close to a classical medical treatment). The subsequent data feedback is a further cornerstone of this model. All organizational levels are involved. It is essential that the persons involved understand the data and the underlying problem diagnosis. Feedback is typically provided in the context of the work groups. Supervisors are assigned an important role in the feedback process. Apart from process facilitating change agents are primarily conceived as knowledge brokers, they should link theoretical knowledge and the world of action. The aim of the feedback sessions is to determine what changes seem to be required, how they should be reached, and what interventions are appropriate. In general the data feedback aims at unfreezing participants' perception patterns and encouraging their willingness to take part in the change process. Follow-up surveys serve to monitor the progress of the OD process and to determine whether the change efforts did work. A somewhat similar abbreviated form of this model is the *Confrontation Meeting* originally developed by Beckhard (1967). Data collection and evaluation are conducted directly and in a condensed form (1 day only). A meeting of all those organizational members involved is held to identify the major problems of the system. After a joint ranking of the problems, the unit develops an action plan and establishes a system of periodical reports on the progress of implementation.

Process Consultation

This is a nondirective model for supporting the client organization in carrying out change processes; it was originally developed by Schein (1988). The process consultant helps the client organization to figure out what should be done to improve the situation. A Process Consultation project typically includes the following stages: entering and defining the relationship, choice of setting and work methods, data gathering, clarifying and confronting interventions (e.g., through the use of feedback or coaching) by the consultant, evaluation of results, and consultant's disengagement. Due to its indirect character, the results of Process Consultation are not predictable at the beginning. Opposed to the Survey-guided Development model, Process Consultation does not rely on a specific theory of organizational functioning as a basis for identifying problems and formulating solutions. The consultant is expected to 'go with the flow' and to realize that the client organization is the 'owner' of the problem and in charge for finding a solution. The consultant is conceived as enabler and not as problem solver.

Grid Organization Development

This is a highly standardized intervention program building on one basic conception; that is, an ideal managerial style with high concern for both people and productivity (Blake and Mouton, 1969). The development process is designed to unfold along six distinct phases: Grid Seminar, teamwork development, intergroup development, developing an ideal model for the whole organization, implementing the model, and examination of the organization's progress toward the plan.

Systemic Interventions

This is a more recent OD model that builds on family therapy and systems theory. The major focus is on fixed paradoxical interaction patterns as invisible barriers to organizational change, thereby drawing on the Double Bind Theory of Schizophrenia (Watzlawick et al., 1967). The transformation of these ideas into an OD approach originally was initiated by the Italian family therapist Selvini-Palazzoli (1995) and her Milan School. In a first step this approach aims at surfacing hidden rules and games in organizations. These patterns (among management teams, between groups, etc.) are assumed as being possibly vicious in character; their circular reinforcement is likely to consume a lot of the system's energy and to evoke strong aversions against any attempt to change them. Due to the circular involvement of the organizational members, this approach refrains from participative change designs. Rather, it addresses the deeply embedded hidden agenda and its paradoxical architecture. After having identified the self-reinforcing logic of fixing the paradoxical dynamics consultants set out to dissolve it (Kilmann, 1988). OD interventions target the core dynamics of the vicious circle; they are carefully planned in terms of a counter paradox. Typical paradoxical interventions are prescription of the circumstances which are causing the symptom, positive connotation of the symptom (e.g., resistance to change), increasing intensity of reinforcement, etc. Systemic Interventions, if well targeted, are likely to result in 'explosions'; that is, the whole architecture of pathological rules collapses like a house of cards, thereby stopping the fatal circles of recursion and paving the way for reopening the system. The general aim is to regain the scope for necessary change. The form of the resulting transformation is, however, hard to predict; new patterns emerge and get reinforced. The result of the process is, however, not fully under the control of the actors.

OD and Beyond

In the beginning the OD approach was received enthusiastically in both theoretical and practical fields but later became increasingly criticized. The most common objection is to the underlying premise of harmony of interests; that is, promising OD projects meet always both employee needs and the organization's efficiency goals. Critics doubt whether there is always such a win–win solution for interest conflicts when responding to change requirements (under fierce competition). This argument leads into a broader criticism accusing OD of showing a general naïveté toward organizational power structures. It is said that OD actors can never claim neutrality because they always operate within power structures and are therefore likely to become part of micropolitics. Also, the financial compensation of the change agents is emphasized. It does not allow them to play the alleged neutral role in the change process. Instead of sharing power OD is criticized of increasing the power of management.

Another objection relates OD to hidden agendas. It is suspected that OD consultants often provide legitimation for a predetermined change program. Management wants them to

skillfully confirm and realize what they have already fixed (Zand, 2010). Agreement is reached through a process design which actually does not leave any scope for members' influence.

In addition to these objections a lack of solid theoretical foundation has been diagnosed. The focus is on techniques and instruments but there is no compelling overarching conceptual framework.

Recently, a new debate on the nature of change came into the foreground of academic and practical interest (Brown and Eisenhardt, 1997). It goes beyond the criticisms mentioned above and fundamentally challenges the paradigm of organizational change underlying OD models. The focus is on the conceptualization of change; episodic change is contrasted with continuous change (Weick and Quinn, 1999). OD clearly stands for episodic change.

Episodic Change

An in-depth analysis of the OD approach reveals three assumptions on the nature of change.

Organizational Change is a Clear-Cut Process

In OD and related approaches, organizational changes are conceived as clear-cut processes with a clearly defined beginning and end, processes that are separated from ordinary organizational operations. This view misconceives the character of organizational processes. Problem-solving processes largely overlap in terms of participants, subjects, and interests and penetrate each other (March and Olsen, 1976). This results from the fact that organizations are, after all, to be managed in a complex and ambiguous context. Change processes are an integral part of the system's general dynamics; they are not separate entities or islands. The definition of a change period is always an artificial construction within a flow of events. The system observes the OD activities and reacts to them, thereby involving the project into the everyday dynamics.

Organizational Change is an Exception in Organizational Life

In OD change is given a special status, that of an exception. It is conceived as an extraordinary phase, which tends to be infrequent. Organizational members perceive them threatening and somewhat chaotic. In this view, changes come close to disturbances likely to confuse the rational order and system of rules. The guiding idea is the stable state of equilibrium. Stability stands for efficiency, rationality, and harmony. Organizational members are assumed to be strongly in favor of order and stability and to feel helpless and frightened when facing change requirements. Furthermore, change is conceived as replacement: Model A does no longer work and therefore has to be replaced by a new model. The replacement thinking favors an either-or logic which (implicitly) excludes multidimensional views of problems and solutions (Weick and Quinn, 1999).

Organizational Change is an Area for Specialists Only

The strong psychological or even psychotherapeutic orientation of OD led to the implicit assumption that organizational change should be restricted to specially trained professionals only. The organization or department is put into the role of a client or even a patient who badly needs a 'doctor' to master the challenge 'illness' of change.

This perspective puts managers into a paradoxical situation. They are expected to rationalize and improve procedures and structures, but the important function of putting improved systems into action is said to be an exercise of which they are not capable. This corresponds with the episodic view, the rare event logic. But can such view match the actual situation in modern organizations where change has become such a prominent feature? The recent debate brought a counter paradigm to the fore: the image of continuous change. Modern organizations are reconceived as *inherently restless* and fluid in nature.

Continuous Change

The alternative view portrays change processes in different terms, the emphasis is on ongoing processes, emerging and located in everyday practices. They unfold simultaneously across units and levels and may eventually accumulate to a far-reaching reorientation of the system. Some models to portray the reverse logic have been developed.

Weick (1977), for example, discusses a radically contrasting model, the 'chronically unfrozen' organization. This refers to an organization that has the unfreezing stage as the rule and refrains from any refreezing. More recent organizational concepts follow the same line of reasoning by emphasizing continuous processes of coming into being, adaptations, modifications in organizational practicing and relentlessly changing forms of cooperation (Orlikowski, 2002).

Other authors (e.g., Brown and Eisenhardt, 1997) conclude from their empirical studies that the ability to change continuously has become a critical success factor and the basis for competitive advantage. These studies show that for many firms relentless continuous change is at the heart of their character, nothing but 'business as usual.' They are constantly redesigning their practices. There are striking similarities with the entrepreneurial discussion in economics. In particular, it is reminiscent of Schumpeter's notion of permanent 'creative destruction' as a critical element of any efficient market economy. In his view, the basic role of entrepreneurs is to constantly irritate the market, so that it can therefore hardly settle down.

More recently, the focus has shifted from structural flexibility to behavioral features, such as absorptive capacity and competencies. In this context, the notion of *dynamic capabilities* features most prominently, stressing the role of organizational capacities for relentless change (Helfat et al., 2007). The idea is that organizational capabilities themselves have to become fluid to enable organizations to continuously create new (combinations of) resources. As a result, organizations are in a continuously unstable state. Dynamic capabilities are seen as a means as well as an outcome of ever-changing decision-making processes concerning resource reconfiguration, integration, and acquisition.

While these models differ on many dimensions, they all have a common core, that is, the critical emphasis on fluidity.

The ideal of equilibrium is replaced by a new ideal, the state of restless change. High-performing organizations are seen as constantly redesigning and reinventing themselves, with increasingly fuzzy and eventually dissipating boundaries. The distinguishing characteristic of these types of organizations is that their activities are no longer bound to a 'deep structure' and organizational routines built in the past. Rather, the organization is steadily revising its cognitions and changing its expectations. Departmentalization, routines, and replicable practices are seen as features representing solutions to old problems, developed for an older industrial environment which valued stability, bureaucracy, and formality as a source of building competitive advantages.

Outlook

At first glance convincing but the model of continuous change is troubled by several drawbacks as well. Put bluntly, the fluidity model radicalizes the right insight at the wrong point: change models should refrain from throwing the baby out with the bathwater. One should not forget that social systems basically need sensemaking and cognitive patterns in order to come to terms with their complex environment. Organizational practices, routines, and capabilities provide a set of problem-solving patterns that enable the system to master tasks in a complex environment and to draw a boundary. Organizations are simply not conceivable without workable identities and boundaries which distinguish them from their environment. By implication, the concept of fluid and relentlessly changing systems ignores an essential feature of any system building. The ideal of fluidity advocates the logic of reacting to any environmental event in a new (not patterned) way. As modern systems theory has clearly shown (Luhmann, 1995), organizations cannot exist and act without guidance from an organizational frame of reference, informed by past learning and experience. Reacting in a turbulent environment to any new event by improvisation and without any pattern implies giving up the distinction between system and environment (Schreyögg and Sydow, 2010).

For further advancing OD is therefore an approach needed which on the one hand avoids a too episodic view of change, and a blurring conception of fluid systems on the other hand. A new approach would have to balance the contradictory functions of flexibility and efficient replication. Change requirements can only be properly understood when put in the context of the interplay between these contradicting functions. When considering change initiatives one should not forget that organizations always also have the option to stay with the established rules, interpretation patterns, and boundaries built in the past. The system has to learn not only to adapt its practices and patterns but also to make use of intentional nonadaptation. Organizations are not relentlessly changing systems; they are partially invariant systems in a complex and dynamic environment.

While the basic OD concept of change seems to be questionable, it is not necessary to give up all those well-developed OD methods and experiences. OD methods can be reframed. The mission of OD could be reformulated to help identify communication disturbances and to eliminate learning blockades (Argyris, 1985).

See also: Boundaries and New Organization Forms; Corporate Culture; Design and Form: Organizational; Entrepreneurship and Organizations; Innovation; Learning: Organizational; Strategic Management.

Bibliography

Argyris, C., 1985. Strategy, Change and Defensive Routines. Pitman, Boston.
Beckhard, R., 1967. The confrontation meeting. Harvard Business Review 45, 149–155.
Bion, W.R., 1961. Experiences in Groups. Basic Books, New York.
Blake, R.R., Mouton, J.S., 1969. Building a Dynamic Organization through Grid Organizational Development. Addison-Wesley, Reading, MA.
Bowers, D.G., Franklin, J.L., 1976. Survey-Guided Development: Data Based Organizational Change. Institute for Social Research, Ann Arbor, Michigan.
Brown, S.L., Eisenhardt, K.M., 1997. The art of continuous change: linking complexity theory and time-paced evolution in relentlessly shifting organizations. Administrative Science Quarterly 42, 1–34.
Cummings, T.G., Worley, C.G., 2008. Organization Development and Change, nineth ed. South-Western College Publishing, Cincinnati, OH.
Helfat, C.E., et al., 2007. Dynamic Capabilities: Understanding Strategic Change in Organizations. Wiley-Blackwell, Chichester, UK.
Kilmann, R., 1988. Paradoxical interventions for top managers and troublemakers. In: Quinn, R.E., Cameron, K.S. (Eds.), Paradox and Transformation: Toward a Theory of Change in Organization and Management. Ballinger, Cambridge, pp. 229–235.
Lawrence, P.R., 1954. How to deal with resistance to change. Harvard Business Review 32, 49–57.
Lewin, K., 1958. Group decision and social change. In: Maccoby, E.E., Newcomb, T.M., Hartley, E.L. (Eds.), Readings in Social Psychology. Holt, Rinehart and Winston, New York, pp. 197–211.
Likert, R., 1967. The Human Organization: Its Management and Value. McGraw-Hill, New York.
Luhmann, N., 1995. Social Systems. Stanford University Press, Stanford, CA.
March, J.G., Olsen, J.P., 1976. Ambiguity and Choice in Organizations. Universitetsforlaget, Bergen.
Orlikowski, W.J., 2002. Knowing in practice: enacting a collective capability in distributed organizing. Organization Science 13, 249–273.
Rice, A.K., 1958. Productivity and Social Organization: The Ahmedabad Experiment. Tavistock, London.
Schein, E.H., 1988. Process Consultation, second ed. Addison-Wesley, Reading, MA.
Schreyögg, G., Sydow, J., 2010. Organizing for fluidity? Dilemmas of new organizational forms. Organization Science 21, 1251–1262.
Selvini-Palazzoli, M., 1995. The Hidden Games of Organization. Pantheon, New York.
Watzlawick, P., Beavin, J.B., Jackson, D.D., 1967. Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies, and Paradoxes. Norton, New York.
Weick, K.E., 1977. Enactment processes in organizations. In: Staw, B.M., Salancik, G.R. (Eds.), New Directions in Organizational Behavior. St. Clair, Chicago, pp. 267–300.
Weick, K.E., Quinn, R.E., 1999. Organizational change and development. Annual Review of Psychology 50, 361–386.
Zand, D., 2010. An OD odyssey: in search of inward light. The Journal of Applied Behavioral Science 46, 424–435.

Development: Social-Anthropological Aspects

A Peter Castro, Syracuse University, Syracuse, NY, USA
David W Brokensha, University of California, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D. Brokensha, volume 6, pp. 3587–3592, © 2001, Elsevier Ltd.

Abstract

Anthropology and development as institutional practices share a common origin in the unequal power relations among peoples that shaped the modern world. Humanitarian and reformist impulses existed from the discipline's earliest days regarding the societies that served as its focus of study. Colonial administrations provided the initial setting for applied anthropology concerned with social betterment. International development was launched after World War II, prioritizing economic growth, with quality of life, equity, and sustainability later emphasized as well. Neoliberal policies and globalization once again stress growth. The rise of China, India, and other newly industrializing countries is challenging conventional views of a Western-dominated world.

Antecedents

Anthropology seeks to understand humanity past and present, including our biological and sociocultural dimensions, in a holistic and comparative manner. As a field of study, it draws on the perspectives and approaches of both science and the humanities. Social and political issues also shaped the discipline from its earliest day. The global expansion of states and markets served as the context for the profession's emergence. While its early practitioners were attracted by scholarly questions, some addressed the predicament of the societies who furnished the focal point of their investigations. Dispossession and depopulation were occurring along the world's colonial frontiers. Lewis Henry Morgan (1851, p. ix) wrote a pioneering ethnography on the Iroquois in the mid-1800s as a means of remedying the 'universally unjust' view of American Indians that existed due to 'imperfect knowledge' and 'prejudice.' Morgan aimed 'to encourage a kinder feeling toward' them, emphasizing their 'capacities for future elevation.' From his reformist but ethnocentric and paternalistic perspective, this meant assimilation into the dominant society through schooling and conversion to Christianity. Contemporary critics of development anthropology contend that current efforts constitute little more than a continuation of such ideas. As will be seen, this interpretation is too simplistic. Yet it bears emphasis that both anthropology and development share a common origin in the unequal relations of power among peoples that shaped the modern world.

Colonialism and Its Legacy

As late as 1945, the distinction between 'colonial and non-colonial areas, dependent and independent nations' constituted 'the major and most well-marked division of the lands and peoples of the world' (Kennedy, 1945, p. 306). Indigenous and ethnic groups also held a quasicolonial status within many nation states. Racial and other cultural barriers, along with political and economic inequalities, often set apart the rulers from the ruled. Within this context colonial officials from Britain, Holland, and, to a lesser extent, France, Italy, Belgium, the United States, and other countries called on anthropologists to provide information on local societies, troubleshoot particular problems such as land tenure, and offer training to administrators. Anthropologists were usually supported through institutes or universities, with a few hired directly as members of the colonial administration.

Felix Keesing (1945, p. 392), surveying applied colonial anthropology during the 1940s, identified what has remained as a central tenet of development anthropology: "the existing culture of a people must be made the constant point of reference if administrative and welfare work is to be intelligently planned and effective." He observed that officials overlooked or ignored it because they used their 'own cultural standards, or some distant goal of future improvement, or some overall policy which will apply to every group.' More than six decades later, top–down, universal, and standardized approaches are still the bane of development practice (Banerjee and Duflo, 2011). The widespread, perpetually repeated nature of this 'error' is in itself a testament to the entrenched inequalities and bureaucratic rigidness that typify development situations (Chambers, 1997).

Controversy existed whether anthropologists (who were mainly academically based) should be involved in applied colonial activities. Some saw this as an opportunity to reduce cultural misunderstanding and serve the interests of both local populations and administrators (Gulliver, 1985). Most of the information generated by anthropologists with the assistance of colonial governments was scholastic in nature, lacking direct applicability to official action. Yet, many anthropologists feared that applied work might further imperial exploitation or social engineering. It seemed at odds with the discipline's guiding principles of cultural relativism and scientific objectivity. Some officials also doubted whether 'academic people' had much to offer in terms of practical guidance (Keesing, 1945). Julian Steward (1969) claimed that their skepticism was well founded because most anthropologists of that era were ill-equipped by training and interests to deal with applied issues and cultural change. Even

decades later, anthropologists who took up development work frequently reported feeling poorly prepared by their academic training for it (Green, 1986). As will be discussed below, well into the new millennium some mutual suspicion still exists between anthropologists and other development practitioners, as well as between academic and applied anthropologists (Horowitz, 1999; Lewis and Mosse, 2006).

In assessing colonialism, emphasis in anthropology and development studies has been on its exploitative and hegemonic nature. Because of colonialism's disrepute, other aspects of its complex legacy have tended to be overlooked or dismissed. Jack Goody (1995) has argued, for example, that the relationship between anthropologists and officials was more differentiated and oppositional than typically portrayed in recent studies. Experiences from colonial agricultural, health, and other interventions formed part of local social memories. A collection of studies from East and West Africa, for example, recorded that these projects often mirrored today's decentralized, community-based efforts (Castro, 1998). Relevant lessons from this experience have been missed by planners, either because of its colonial association or due the broader problem of ignoring local knowledge.

Development and Anthropologists: Early Encounters

Post–World War II reconstruction, decolonization (including nationalist struggles), and the onset of the Cold War were the immediate context for the rise of aid programs aimed at promoting development. The Allies produced a new global institutional framework to facilitate problem solving: the United Nations (UN), including its specialized technical assistance agencies such as the World Health Organization (WHO), Food and Agriculture Organization (FAO), and the UN Educational, Scientific and Cultural Organization (UNESCO). Keynesian theories about managed economies shaped the International Monetary Fund (IMF) and World Bank Group to ensure stability and growth. American President Harry Truman's announcement of the 'Point Four Program' in 1949 is often cited as the key moment in launching development as a worldwide enterprise. From its start international aid combined humanitarian motives with political, economic, and other strategic concerns, a situation that has never altered.

Public administrators, engineers, and economists generally guided the new development programs, whose missions were to promote growth and improved living conditions by transferring technology. Anthropologists, when they were involved, usually served as experts on local societies. The new technical assistance agencies such as UNESCO, WHO, and FAO included some anthropologists early on in their activities. Margaret Mead (1953) and others, for example, produced for UNESCO a volume on *Cultural Patterns and Technical Change*. In contrast, international financial institutions took longer to engage them. The World Bank hired its first on-staff 'sociologist' in 1974, and anthropologists did not enter financial institutions in significant numbers until the 1990s (Horowitz, 1999; Marsden, 2010).

Anthropologists entered US aid in 1951, when a small group under George Foster were seconded from a Smithsonian training program in Latin America to evaluate community health care in the region. Their presence was not sustained, as the United States Agency for International Development (USAID) had only one full-time anthropologist on its staff in 1974 (Hoben, 1982). The early anthropologists were often assigned troubleshooter roles, seeking to explain local resistance to projects. Charles Erasmus (1954), one of the Smithsonian anthropologists, soon documented the same problem identified earlier by Keesing: projects often reflected the priorities and assumptions of planners, instead of taking into account local needs, capacities, and preferences. Such findings were not often welcomed, and anthropologists sometimes held the reputation as 'spoilers' or as romantic protectors of local communities. The dwindling of anthropology reflected both a lack of interest on the part of USAID and anthropologists. The latter were drawn away by the boom in academic jobs during the 1950s and 1960s, as well as the controversies of Cold War politics, including the exposing of a few anthropologists who engaged in clandestine research for the US military.

'New Directions' and the Return of Anthropologists

By the early 1970s, disillusionment emerged among the public, politicians, and some development practitioners with the approaches of the previous two decades. These efforts had been characterized by huge projects and centralized planning guided by theories of evolutionary modernization and the 'trickle-down effect.' As Allan Hoben (1982, p. 350) wrote, "the ethnocentric tech-fix orientation of the dominant development paradigm" was challenged. In many countries, indices on the quality of life – infant mortality, life expectancy, per capita income, food production – had failed to improve or had even fallen. The gap between rich and poor, both between and within countries, was widening. In response, the World Bank and the USAID announced in 1973 their intention to direct interventions more specifically at the poor. Under its 'New Direction' guidelines, USAID required all proposed projects to undergo a social soundness analysis, identifying probable impacts on the supposed 'beneficiaries.' In essence, consideration of social variables was brought earlier into the project cycle, influencing project design and even policy. To facilitate these tasks, USAID hired anthropologists, with 50 on staff and at least 100 as consultants by 1980 (Hoben, 1982). Their presence reflected a growing desire by many anthropologists to use their skills and knowledge to address global social issues. A depressed academic job market also fostered a reassessment of desirable employment possibilities.

Anthropologists were pivotal in promoting social- and participatory-oriented approaches to development policy and practice. In doing so, they drew on cutting-edge knowledge, theory, and methods, demonstrating the vital link between academic and applied fields. For example, the World Bank and other development agencies long supported projects such as dams that displaced large populations, yet they lacked guidelines for dealing with resettlement. Drawing on pioneering research by Thayer Scudder, Elizabeth Colson, and others, anthropologists at World Bank led by Michael Cernea (its founding sociologist) spearheaded during the 1980s the formulation of comprehensive policies. Both policy and

practice were also addressed by the Community Forestry Unit (CFU) of FAO's Forestry Department, which was directed by anthropologists Marilyn Hoskins and Katherine Warner from the mid-1980s to 2002. The CFU aimed to strengthen the capacity of the public, resource users, technical personnel, and policymakers to engage in participatory natural resource management by disseminating knowledge of key topics, including gender analysis, indigenous knowledge systems, common property systems, conflict analysis, and livelihoods analysis. It also generated manuals and guides for participatory planning, evaluation, and specialized tasks such as income-generation projects.

Not everyone in development institutions and recipient countries welcomed these innovations. Resistance occurred among those who viewed them as superfluous, wasting time, and increasing costs. Social and participatory approaches also threatened bureaucratic fiefdoms and patronage networks. Evaluative studies by the World Bank and USAID challenged these claims, showing that projects that incorporated social analysis generally proved more effective and attained a higher economic rate of return (Kottak, 1991). Obtaining official compliance with institutional policies and guidelines, as well as with collaborative agreements, still remains a source of contention. The World Bank Group, for example, set up a permanent independent body in 2000 to handle such complaints in a thorough manner.

The Era of Sustainability, Neoliberalism, and Globalization

What constituted development practice shifted yet again during the 1980s. Agencies retained their nominal 'pro-poor' commitment, with emphasis also placed on 'sustainability' in the wake of the 1987 Brundtland Commission report, the 1992 Earth Summit in Rio. New measures, such as the Human Development Index, introduced by the UN Development Programme, highlighted quality of life aspects. Practitioners and activists sought to connect the meaning of development to rights, human capacities, and freedom. The institutional scope of development activities widened, as nongovernmental organizations (NGOs), indigenous peoples, social movements, and other activists increasingly appeared on the scene. Many NGOs followed the funding and pursued donor-driven agendas, but others, both international and local, pressed forward with their own concerns. The rise of Grameen Bank, a grassroots movement for local empowerment, especially focusing on women and families, through microcredit that spread worldwide demonstrated the new creative possibilities of bottom–up development (Yunus, 2007). International donors, for a variety of reasons, endorsed decentralization, creating new political spaces for participation by communities. These trends generally fit well with the campaign by anthropologists and others within and outside international development institutions to 'put people first.' Anthropologists still strove to influence both policy and specific interventions in a wide range of sectors and institutional settings (Grillo and Rew, 1985; Cernea, 1991; Gardner and Lewis, 1996; Nolan, 2002). The Institute for Development Anthropology (IDA), an NGO founded by David Brokensha, Michael Horowitz, and Thayer Scudder in 1976, used its engagement with a large number of multilateral and bilateral aid agencies to promote equity, human rights, and poverty reduction (Brokensha, 2007). IDA operated until 2005.

Yet broad political forces simultaneously asserted another agenda: putting money first, with economic growth as the main and perhaps only goal and measure of development. Neoliberal ideology was in ascendancy in the West, seeking to roll back the Keynesian-managed economy in favor of 'free markets' (which were always selectively defined) and global trade. China's economic reforms and the Soviet Union's collapse heightened the sense of the market's inevitability – some pundits even claimed that the 'end of history' was at hand. While some anthropologists attacked foreign aid as a tool of capitalist domination (Ferguson, 1990; Escobar, 1995), the most significant attacks in terms of policy came from economists who viewed it as supporting bloated, corrupt bureaucracies that choked markets (Moyo, 2009). In general, public support for long-term foreign aid declined, as media coverage of relief and development issues fostered 'compassion fatigue' (Moeller, 1999).

The IMF and World Bank (which retained its guidelines and proponents for socially oriented development) now promoted strict neoliberal agendas in debt-laden countries, forcing them to accept rigid and rapid market liberalization as a condition for aid. The debt crisis surfaced by the late 1970s, the result of mismanagement by both borrowers and lenders, rising energy costs, and falling commodity prices. The 'structural adjustment' program eliminated barriers for trade and investment, reduced state spending on social services, and promoted privatization. Development assistance was increasingly directed away from governments to NGOs (including faith-based enterprises). Donors also favored decentralization, though resources and capacity were often lacking to ensure that it could be done effectively. Economic shocks and declining services often resulted, with devastating impacts, especially on the poor. For example, anthropologists and health professionals documented the declining access of low-income populations to medical care (Kim et al., 2000; Farmer, 2003; Pfeiffer and Chapman, 2010).

An unprecedented expansion in world trade took place after 1980, as multinational corporations outsourced operations to low-wage areas. Some countries, such as South Korea, China, Brazil, and India, managed market integration in a manner that raised incomes for a large number of their citizens. But anthropologists and others documented that neoliberalism and globalization also increased vulnerability worldwide, though in different ways (Comaroff and Comaroff, 2001; Edelman and Haugerud, 2005). Declining social services, removal of food and other important subsidies, privatization of land and other resources, displacement of people due to megaprojects, the spread of the weapons trade, and other trends made life more difficult for people in the world's developing areas. Vulnerability also increased in developed countries due to outsourcing of jobs, the stagnation of working class wages, reduced public services, and other trends. Instability and chaos arising out of neoliberal practices and its politics of inequality, particularly in the 2008 crisis, further extended the sense of insecurity. For anthropologists, these tumultuous changes reduced the

distinction between development and other research contexts. As Chris Hann and Keith Hart (2011, p. 142) observed, "We all seemed to be living in one world unified by capitalism, so anthropologists now study that." Physician and anthropologist Paul Farmer (2003), noted that similar 'pathologies of power' operated across and within societies, generating a kind of structural violence against the poor, who were excluded, and vulnerable.

The New Millennium and Development

In a perceptive article, Curt Grimm (1998) observed USAID's declining interest in, and capacity for, socially oriented development. USAID had not rejected the lessons of its 'New Directions' agenda but reframed it into an emphasis on market-oriented 'customer service.' A contrast can be made with British foreign aid, the Department for International Development (DfID), which had build up its social development capacity, while promoting holistic concepts such as 'sustainable livelihoods analysis' in their planning. By the new millennium, DfID had assumed a leadership role in international development once held by their American counterparts. Even as World Bank had promoted severe structural adjustment programs, its anthropologists and others had continued to press its leadership to acknowledge equity issues, including promoting local empowerment, respecting indigenous rights, and so on (Davis, 1998). Their efforts demonstrated that an institution such as World Bank was not a monolith, but composed of competing interests. In fact, anthropologists widened their scope of study to probe the agencies, examining them ethnographically as sites of contestation rather than as 'totalizing' entities of capitalism (Lewis and Mosse, 2006; Mosse, 2011).

National leaders meeting at the UN in 2000 adopted the Millennium Declaration, which committed them to pursuing eight development goals through 2015. The Millennium Development Goals (MDGs) offered a reaffirmation of quality of life issues, focusing on poverty/hunger, education, gender equity, child mortality, maternal health, disease, environmental stability, and 'partnership' issues. Ironically, given the latter goal, the MDGs were mainly shaped by members of the Organization for Economic Co-operation and Development, a group of 30 countries, most of them among the most prosperous in the world. Critics, including anthropologists, expressed skepticism, often portraying it as yet another example of a donor-driven process imposed on developing countries (Black and White, 2004). With their emphasis on target delivery, the MDGs appeared to some anthropologists as a return to social engineering. There also seemed to be a disconnect from the broader context of neoliberalism, with its destabilizing inequalities, ongoing financial chaos, development-related environmental change (including climate change, deforestation, and biodiversity loss), and other issues.

The Global War on Terror in the wake of 11 September 2001 added to the confusion. Policymakers increasingly sought to link development activities to counterinsurgency 'hearts and minds' campaigns, a practice dating back to anticolonial struggles. The military were asked to fulfill 'humanitarian interventions,' a practice long regarded as controversial. The open recruitment of anthropologists and other social scientists into the Human Terrain Project and related 'security' endeavors sparked heated debate on political and ethical grounds within the discipline (Price, 2011). The 'War' also helped draw attention to another manifestation of globalization: the presence of sizable transnational immigrant communities in developed countries who kept up financial and political, as well as social, links to their home areas. Remittances by international migrants now represented a substantial economic force, totaling more than $125 billion in 2005, exceeding the contributions of donor countries to international development (Cohen, 2011).

In a provocative analysis, Scudder (2010), drawing on his five decades of experience as an applied anthropologist, explored the possibility that current crises foreshadow a coming era of declining global living standards. To cope, 'transformational changes' must occur in terms of values, political economy, and social relations. Some anthropologists and activists already portrayed 'postdevelopment' communities and social movements, where people possess values opposed to the hegemonic discourse of development and consumerism (Escobar, 1995). Others assert that most people worldwide still seek to be a part of the global consumer society, thus adhering to, as well as pushing forward, the 'old premise of development' (Hann and Hart, 2011, p. 120). Goody (2010) and others suggest that the economic rise of East Asia and India also offer a significant challenge to the view of development as synonymous with Westernization.

The Anthropology of Development

No singular subdiscipline of development anthropology has ever emerged in the sense of 'a coherent or distinctive body of theory, concepts, and methods' (Hoben, 1982, p. 349). What constitutes an 'anthropology of development' is simply the sum total of the diverse perspectives and experiences of those who engage in, write about, or teach on policies, practices, and discourse dealing with development and topics such as globalization. In this regard, anthropological approaches and activities related to development, as well as colonialism and globalization, closely reflect what is happening in the wider discipline as a whole. At the same time, anthropological engagement with development has shaped to some extent thinking and practice within the discipline. It also influenced, both for better and worse, how other fields understood development and globalization.

As mentioned earlier, applied anthropologists were often initially assigned the task of explaining local resistance to development projects. Given the discipline's predominant ideas in the 1950s and 1960s, they often did so by emphasizing local values and cultural patterns. Academic studies by Clifford Geertz, Oscar Lewis (who devised the concept of the 'culture of poverty'), and others dealing with 'modernization' and development themes covered similar ground. However, unintentionally, these analyses reinforced prevailing ethnocentric ideas about the people and processes involved in development. Anthropologists eventually moved on (see later text), but

specialists in other fields clung to the idea of culture as the primary barrier to the development. Lawrence Harrison and Samuel Huntington (2000) heralded these views in an influential collection, with Daniel Etounga-Manguelle (2000), for example, arguing that Africa needed a kind of 'cultural structural adjustment program' to mend its feckless ways. Ironically, the World Bank (purveyor of structural adjustment programs), issued a volume featuring anthropologists such as Mary Douglas and Lourdes Arizpe, as well as others, refuting such arguments (Rao and Walton, 2004). The debate about culture's role in development still continues.

Anthropological concerns about development have shifted in the past three-plus decades as the field itself has undergone substantial alteration. Diverse theoretical currents have flown through it, including variants of Marxism, feminism, post-modernism, and practice theory, inspiring innovative forms of analyses, including actor-oriented approaches and discourse analysis. Anthropologists have put together these concepts and tools for very different ends. Arturo Escobar (1995) used discourse analysis to expose development as a discourse and institutional practices that perpetuate domination of the Third World. He argued that anthropologists should not be involved with development agencies. In contrast, Katy Gardner and David Lewis (1996) agreed with many of the criticisms of the development industry, but reached a different conclusion: contending that the insights of discourse analysis, combined with ethnographic research, offered new possibilities for supporting effective participatory efforts. Gardner and Lewis felt that the involvement of anthropologists from the global 'South' especially offered innovative avenues for change. In some countries, such as Mexico, anthropology has a long, if contentious, involvement with development issues (Nahmad Sittón, 2011). Other countries are at a relatively early stage of building their capacity in anthropology, though applied issues still often loom large (Hill and Baba, 2006).

Overall, the rise of neoliberalism and globalization has narrowed the gap between development and other domains within the discipline, as convincingly displayed in Marc Edelman and Angelique Haugerud's (2005) edited volume. In addition, anthropologists are increasingly seeking to make their work 'relevant' to a wider public, including the populations that they are studying (Low and Merry, 2010). The discipline's strength has always been in the ability of analysts to illuminate site-specific conditions through ethnographic research. The shrinking difference between academic and applied studies is especially evident in studies dealing with fields and topics such health, AIDS, migration, postsocialist transitions, social movements, conflict, livelihoods, and pastoralism.

Conclusion

The appointment of anthropologist and physician Jim Yong Kim as president of the World Bank in July 2012 may signal a significant shift in the practice of international development. Taking on a position always held by economists, Kim has been a critic of policies that promoted growth yet undermined the health status of the poor. In many ways, his new role reflects the outcome of a long, contentious history of anthropological involvement with development issues. Dating back to Morgan's Iroquois study and no doubt beyond, anthropologists have sought to use their skills and knowledge to assist societies marginalized by market and state expansion. Those interested in applied anthropology have not only faced skepticism and skepticism from colonial and development officials, but also from their own academically based colleagues. The experience of anthropologists in development, especially in the era of neoliberalism and globalization, has narrowed, though not eliminated, such tensions. As anthropology faces growing uncertainty in terms of future academic job growth, greater attention will have to be given to prepare students who may seek positions in the wide range of organizations and groups now involved in development-related activities.

See also: Development and the State; Ethnic Cleansing, History of; Globalization, Anthropology of.

Bibliography

Banerjee, A.V., Duflo, E., 2011. Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty. Public Affairs, New York.
Black, R., White, H. (Eds.), 2004. Targeting Development: Critical Perspectives on the Millennium Development Goals. Routledge, New York.
Brokensha, D.W., 2007. Brokie's Way: An Anthropologist's Story. Amani Press, Fish Hoek.
Castro, A.P. (Ed.), 1998. Special Section on Historical Consciousness and Development Planning. World Development 29, 1695–1784.
Cernea, M.M. (Ed.), 1991. Putting People First: Sociological Variables in Rural Development, second ed. Oxford University Press, New York.
Chambers, R., 1997. Whose Reality Counts? Putting the First Last. Intermediate Technology Development Group, London.
Cohen, J.H., 2011. Migration, remittances, and household strategies. Annual Review of Anthropology 40, 103–114.
Comaroff, J., Comaroff, J. (Eds.), 2001. Millennial Capitalism and the Culture of Neoliberalism. Duke University Press, Durham.
Davis, S.H., 1998. Bringing culture into the development paradigm: the view from the World Bank. Development Anthropologist 16, 25–31.
Edelman, M., Haugerud, A. (Eds.), 2005. The Anthropology of Development and Globalization: From Classical Political Economy to Contemporary Neoliberalism. Blackwell, Malden.
Erasmus, C.J., 1954. An anthropologist views technical assistance. The Scientific Monthly 78, 147–158.
Escobar, A., 1995. Encountering Development: The Making and Unmaking of the Third World. Princeton University Press, Princeton.
Etounga-Manguelle, D., 2000. Does Africa Need a Cultural Adjustment Program? In: Harrison, L.E., Huntington, S.P. (Eds.), Culture Matters: How Values Shape Human Progress. Basic Books, New York, pp. 56–64.
Farmer, P., 2003. Pathologies of Power: Health, Human Rights, and the New War on the Poor. University of California Press.
Ferguson, J., 1990. The Anti-politics Machine: Development, Depolitization and Bureaucratic Power in Lesotho. Columbia University Press, New York.
Gardner, K., Lewis, D., 1996. Anthropology, Development and the Post-modern Challenge. Pluto Press, London.
Grillo, R., Rew, A. (Eds.), 1985. Social Anthropology and Development Policy. Tavistock, London.
Grimm, C.D., 1998. Anthropology and the U.S. Agency for International Development: are the best years behind? Development Anthropologist 16, 22–25.
Goody, J., 1995. The Expansive Moment: Anthropology in Britain and Africa, 1918–1970. Cambridge University Press, Cambridge.
Goody, J., 2010. The Eurasian Miracle. Polity, Cambridge.
Green, E. (Ed.), 1986. Practicing Development Anthropology. Westview, Boulder.
Gulliver, P.H., 1985. An applied anthropology in East Africa during the colonial era. In: Grillo, R., Rew, A. (Eds.), Social Anthropology and Development Policy. Tavistock, London, pp. 37–57.

Hann, C., Hart, K., 2011. Economic Anthropology. Polity, Cambridge.
Harrison, L.E., Huntington, S.P. (Eds.), 2000. Culture Matters: How Values Shape Human Progress. Basic Books, New York.
Hill, C.E., Baba, M.L. (Eds.), 2006. The globalization of anthropology. NAPA Bulletin 25.
Hoben, A., 1982. Anthropologists and development. Annual Review of Anthropology 11, 349–375.
Horowitz, M.M., 1999. The sustainability of anthropology and development. Development Anthropologist 17, 10–24.
Keesing, F.M., 1945. Applied anthropology in colonial administration. In: Linton, R. (Ed.), The Science of Man in the World Crisis. Columbia University Press, New York, pp. 377–398.
Kennedy, R., 1945. The colonial crisis and the future. In: Linton, R. (Ed.), The Science of Man in the World Crisis. Columbia University Press, New York, pp. 306–346.
Kim, J.Y., Mullen, J.V., Irwin, A., Gershman, J. (Eds.), 2000. Dying for Growth: Global Inequality and the Health of the Poor. Monroe. Common Courage Press.
Kottak, C.P., 1991. When people don't come first: some sociological lessons from completed projects. In: Cernea, M.M. (Ed.), Putting People First: Sociological Variables in Rural Development, second ed. Oxford University Press, New York, pp. 431–464.
Lewis, D., Mosse, D., 2006. Development Brokers and Translators. Kumarian, Bloomfield.
Low, S.M., Merry, S.E., 2010. Engaged anthropology: diversity and dilemmas. Current Anthropology 51 (Suppl.), S203–S224.
Marsden, D., 2010. W(h)ither anthropology? Opening up or closing down. Anthropology Today 26, 1–3.
Mead, M. (Ed.), 1953. Cultural Patterns and Technical Change. UNESCO, Paris.
Moeller, S., 1999. Compassion Fatigue: How the Media Sell Disease, Famine, War and Death. Routledge, New York.
Morgan, L.H., 1851. League of the Ho-de-no-sau-nee. Sage & Brother, Rochester.
Mosse, D. (Ed.), 2011. Adventures in Aidland: The Anthropology of Professionals in International Development. Berghahn Books, Oxford.
Moyo, D., 2009. Dead Aid: Why Aid Is Not Working and How There Is a Better Way for Africa. Farrar, Straus and Giroux, New York.
Nahmad Sittón, S., 2011. The role of anthropology with the changes and challenges of the 21st century in Mexico and the world. Human Organization 70.
Nolan, R., 2002. Development Anthropology: Encounters in the Real World. Westview, Boulder.
Pfeiffer, Chapman, 2010. Anthropological perspectives on structural adjustment and public health. Annual Review of Anthropology 39, 149–165.
Price, D.H., 2011. Weaponizing Anthropology. CounterPunch, Petrolia.
Rao, V., Walton, M. (Eds.), 2004. Culture and Public Action. Stanford University Press, Stanford.
Scudder, T., 2010. Global Threats, Global Futures: Living with Declining Living Standards. Edward Elgar, Cheltenham.
Steward, J., 1969. Limitations of applied anthropology: the case of the American Indian New Deal. Journal of the Steward Anthropological Society 1, 1–17.
Yunus, M., 2007. Creating a World without Poverty. Public Affairs, New York.

Development Studies

Su-ming Khoo, School of Political Science and Sociology, National University of Ireland, Galway, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This contribution identifies development as an *emergent* and *contested* key concept in the social and environmental sciences. 'Development studies' describes a field that includes multi-, inter-, and transdisciplinary approaches to economic growth, basic needs, inequality, human development, rights, resources, livelihoods, and sustainability, focusing mainly, but not exclusively, on 'developing' countries. The scope and content of development studies is described and the main schools of thought, theoretical debates, concepts, and indicators are discussed in historical context, ranging from the colonial appropriation of natural resources to modernization and its critiques, to current agendas for sustainable and rights-based human development.

Introduction

'Development' has been a core concept for the social sciences since the Enlightenment and has been named as 'the central organizing concept of our time' (Cowen and Shenton, 1995: p. 27). Development can be understood as an *emergent* and *contested* concept – different meanings have emerged through continuing critical debates. In mid-eighteenth-century European thought, 'development' signified a general process of unfolding or becoming. It acquired evolutionary and biological meanings as these ideas took shape in the later eighteenth and nineteenth centuries. Ideas of progressive development became highly controversial as evolutionary theory turned into a major battleground for the contending forces of science and religion. The twentieth-century politics of decolonization brought further contention as neoimperial and national interests competed to claim the project of 'development.'

'Development studies' constitutes a relatively young field (Haddad et al., 2011). The term has been used, since the late 1940s, to describe a multidisciplinary field within the social sciences, addressing the economic, social, and political concerns of the 'developing' countries. A looser practical definition of development studies describes it as "all forms of writing and talking about development, as well as forms of action (such as training and even project work) which grow out of these processes" (Edwards, 1989: p. 117). 'Development studies' covers a spectrum of academic and nonacademic theory, policy, and practice. It is cross-disciplinary, involving a spectrum of multidisciplinarity, interdisciplinarity, and transdisciplinarity (Sumner and Tribe, 2008: pp. 35–37) across subject areas including (but not limited to) economics, politics, sociology, anthropology, agriculture, health, education, law, geography, and area studies. Hettne defines development studies as "…a problem-oriented, applied and interdisciplinary field, analyzing social change in a world context, but with due consideration to the specificity of different societies in terms of history, ecology, culture, etc." (1995: p. 4). This definition encapsulates three paradoxes: the field of development studies is cross-disciplinary, yet rooted in disciplinary knowledge; global in scope, yet embedded in local contexts; generalizing and theoretical, yet practical and problem oriented.

Development studies' emergence in the 1940s was influenced by changing colonial policy, and also by a move toward social democracy and welfare states within the imperial core which became defining characteristics of 'developed' countries. Dominant theories about development appear to follow a cyclical pattern, alternating between *laissez-faire*, growth-oriented approaches and more interventionist, redistributionist, and equity-oriented ones (Cooper and Packard, 1997: pp. 29, 30).

The ideas and institutions of development studies were strongly influenced by the Cold War between the 1940s and the end of the 1980s, and from then on by globalization, neoliberalism, the rise of new, nongovernmental development actors in the 1990s, and recent social market and social entrepreneurial approaches. Current debates in development studies also highlight the role of faith (Tadros, 2011), philanthropy (Edwards, 2011), and the influence of economically powerful, large Southern states such as Brazil, India, and China on development (UNDP, 2013). Conventionally, development studies has been more concerned with the 'Third World,' 'Global South,' or 'developing countries' of Africa, Asia, and Latin America and less concerned with the problems of the 'Global North' – the European, North American, and Antipodean 'developed countries.' The term 'Third World' derives from the French demographer Alfred Sauvy's expression '*Tiers Monde*.' Sauvy likened the masses or 'Third Estate' of the 1789 French Revolution to the countries that had assembled in Bandung in 1955 to declare their 'nonalignment' with the non-Communist 'First World' or the Communist 'Second World' (Sauvy, 1952; see also Hettne, 1995). These distinctions between the 'developed' or richer and 'underdeveloped' or poorer countries of the world, and between the three 'worlds' of development became problematic with the end of the Cold War, the disappearance of the 'Second World,' the spread of globalization, and the rise of middle-income countries by the end of the 1980s.

Globalization has surfaced shared crises and challenges such as climate change, financial instability, resource depletion, and food crises, and made the need for coordinated responses and alternative approaches more salient. The global economic turmoil of 2008 stimulated a degree of critical self-reflection in the North and swung debates back to the questions that

emerged following earlier historical crises. Current critiques of growth (e.g., Jackson, 2011) revive earlier debates on the 'affluent society' and the downsides of economic growth (Galbraith, 1958; Mishan, 1967). Environmental concerns that emerged in the 1960s (Carson, 1962/2000) have become accepted by the mainstream (e.g., Collier, 2011). The questions leveled by nonorthodox development economists at their own discipline about inequality and unemployment (e.g., Seers, 1963, 1969) have been revisited by critical commentators on globalization and inequality (Kaplinsky, 2005), while the fundamental assumptions and disciplinary primacy of economics, the dominant discipline in development studies since the 1950s, have been similarly questioned (Krugman, 2009).

The questions of economic, social, and political progress addressed by development studies have been, from the outset, accompanied, heightened, and complicated by political processes of decolonization and the interplay between national self-determination, human rights, democratization, and citizen politics. Since 2000, development studies has taken a more technical and less political track, focusing on the Millennium Development Goals (MDGs, 2000–15), and questions of 'effectiveness,' financing, and 'good governance.' However, questions about ecological damage, depletion of natural resources, and the fundamental role of the natural environment in sustaining human life, health, and livelihoods have not, so far, been fully addressed. The integration of environmental and natural resource dimensions, and their connection to problems of inequality, continue to pose a central challenge to development studies in the twenty-first century.

Historical Origins of Development Studies

Current patterns of global development have been profoundly shaped by the history of capitalist and colonial power relations. 'Development' is conventionally understood as a post-Second World War artifact, oriented toward national states' pragmatic concerns with planned economic growth and international cooperation. However, this postwar agenda was not completely new, but built upon historical patterns of colonial and imperial rule which fundamentally structured global and local economies, societies, and cultures over several centuries. Imperialism and colonialism profoundly reshaped the relationships between people, nature, and natural resources (Woodhouse and Chimhowu, 2005: p. 180). The colonial concept of development referred to the *development of economic resources*. 'Developing' a territory or region meant 'opening up' natural resources for the colonial economy, mobilizing labor, possibly by promoting large-scale labor migration, and building infrastructure such as railways, ports, and roads to facilitate resource exploitation (Arndt, 1981: p. 462). Colonial regimes secured their access to resources such as land, forests, water, and minerals by codifying laws, procedures, and institutions of colonial administration, displacing the traditions and norms that underpinned social relations and people's access to livelihood resources (Randeria, 2003: pp. 40–43). Imperialists generally viewed their 'development' activities to be humanitarian and benevolent, a view supported by the expansion of both missionary activity and social science in the late nineteenth and early twentieth centuries.

Social welfare became an explicit consideration with the advent of a new mode of colonialism, the '*dual mandate*,' which cast the colonizers in a tutelary, protective role, even as they sought out resources and profits from their colonies (Arndt, 1981: p. 463). This ambiguous rationale continued into the era of decolonization under the provisions of 'trusteeship.' The British Colonial Development and Welfare Act of 1939 defined trusteeship in terms of a colonial responsibility for 'minimum standards of nutrition, health and education' (Hancock, 1942 cited in Arndt, 1981: p. 463). After the Second World War, 'development' was presented as the solution to a global problem of 'underdevelopment' in a new era of decolonization and international cooperation. An international 'New Deal' was promised by the new global power, the United States. At his 1949 inaugural address, President Truman declared that "[t]he old imperialism – exploitation for foreign profit – has no place in our plans. What we envisage is a program of development based on the concepts of democratic fair-dealing" (cited in Rist, 1997: p. 71). Nevertheless, the political and economic self-interest of leading powers remained important drivers of development in the former colonies, and geopolitical competition influenced the establishment and emergence of development studies programs (Cooper and Packard, 1997; Gilman, 2007). While many consider development studies to be largely an 'asymmetric business' involving Northern social scientists studying Southern countries' development problems (Forster, 1997), important exceptions include the Latin American 'dependency school' and Islamic critiques of western modernity (Connell, 2007). Dependency theorists emphasized the systemic links between colonial resource appropriation, unequal exchange, and global inequality, while Islamic critics highlighted the undesirability of western materialism and consumerism, while emphasizing responsibility and stewardship as spiritual values. The current combination of economic and social crisis in the North and the 'rise of the South' (UNDP, 2013) have deepened skepticism about the 'Eurocentric' model based on 'transferring' knowledge from the 'North' to the 'South,' favoring a more equitably global, interdependent, and partnership-based approach to development as a shared concern of all countries – rich, poor, or 'transitional.'

Major Theoretical Approaches and Debates: From Modernization to Dependency Critiques to Human Development

Development studies' formative period – from the late 1940s to the early 1970s – could be described as the 'golden age' of welfare capitalism (Marglin and Schor, 1990) characterized by coordinated state action for national development. National accounting was established as the basis for planned national development. Advances in 'modernization' and industrialization could be measured in terms of aggregate formal economic production, gross domestic product (GDP) and retained national income, and gross national product (GNP, also termed gross national income). National

accounting techniques developed in the USA and Europe in the 1930s and 1940s in response to the Great Depression, the rise of Keynesian macroeconomics, and demands for resources for major war efforts. However, despite the increasing sophistication of national economic accounting techniques, they also presented certain limitations. Critiques emerged as the tensions between differing social concerns, theoretical issues, and data constraints became increasingly evident (Vanoli, 2008).

Modernization theory, originating in the 'modernization school' of the 1950s and 1960s, predominated in development studies up to the 1990s, with one notable modification – the neoliberal turn against state developmentalism and toward free-markets at the end of the 1970s. Modernization theorists had incorporated the basic assumptions of nineteenth-century evolutionary and imperial social science, but shaped them according to the emerging geopolitics of the Cold War (Gilman, 2007). An example is Walt Rostow's *Five Stages of Economic Growth* which bore the subtitle 'A Non-Communist Manifesto' (Rostow, 1960). Like their nineteenth-century intellectual forebears, modernization theorists assumed that all societies began as 'primitive societies' that would gradually make the transition to cultural modernity and 'advanced' economic status, defined as high and growing national income, maximizing the commodification and exploitation of natural resources, industrial production, and the transition to a consumer society. The need to assist the economic growth process came from the perceived threat of Communist revolution; hence Cold War superpower rivalry drove programs of international assistance for 'development' in the Third World. More concerned with political stability than social justice, modernization theorists assumed that the benefits of economic growth would trickle down to reach the poor. However, statistics that were now being systematically collected showed that poverty, inequality, and deprivation had not diminished, but persisted and even increased in many countries.

The postwar international development regime's claim to fair-dealing under the *Pax Americana* was tested by Structuralists who noted the problem of asymmetry in the global economy. In 1950, Hans Singer and Raul Prebisch came up with the 'Singer–Prebisch thesis' which postulated that the primary product exporting 'developing' economies experienced declining terms of trade vis-à-vis advanced economies exporting manufactured goods. The Structuralist findings led neo-Marxist dependency theorists like Paul Baran and Andre Gunder Frank to argue that the structure of the global economy was essentially neocolonial, presenting intrinsic difficulties for developing countries (e.g., Frank, 1966). National economies were therefore advised to 'delink' their economies and shift their economic base away from primary commodity exporting and toward industrial production with national tariff protection. The 'import substitution industrialization' (ISI) strategy gained political support in the 1950s and 1960s, but declined in favor, partly due to a lack of success of ISI policies, but also because transnational corporations countered developing country protectionism with the new strategy of 'foreign direct investment' – this became the predominant approach by the later 1970s.

The new discursive forums convened under the auspices of the United Nations, such as the UN Economic Commission for Latin America (ECLA or, in Spanish, *CEPAL*) and the UN Conference on Trade and Development, enabled a developing awareness and solidarity among developing countries, based on shared critical analyses of unjust structural tendencies and aggregating developing country demands for global reforms.

In 1966, Andre Gunder Frank popularized the perspective of the Latin American 'dependency theorists' with his essay on '*the development of underdevelopment*,' which highlighted the unequal, neocolonial nature of global development (Frank, 1966). Hettne (1995) characterizes dependency theory as principally a Latin American contribution – 'the voice of the other' (see also Connell, 2007).

In 1972, the Senegalese judge Keba M'Baye connected development with the emerging human rights agenda of the United Nations, through the idea of a Right to Development (Barsh, 1991). Developing countries articulated a collective demand for a New International Economic Order (NIEO) at the UN General Assembly in 1974 (United Nations, 1974). The demands included fairer financing and terms of trade; more controls over multinational corporations; and greater equality and influence for developing countries within the United Nations. The UN Human Rights Commission began to examine development from a human rights perspective in 1977, resulting in the 1986 UN Declaration on the Right to Development (United Nations, 1986). However, like the NIEO, the Right to Development came to be seen as little more than 'an expression of South frustration in the face of intransigent North power' (Barsh, 1991: p. 322). The Right to Development was viewed with suspicion by orthodox (and predominantly Northern) human rights advocates, who believed that the 'first-generation' individual civil and political rights (which were supported by the capitalist 'First World' powers) were more workable and 'justiciable' than 'second-generation' economic, social cultural rights (which were attributed to the socialist 'Second World'). Third-generation 'solidarity rights,' which were identified with the Third World nations, were thought to be even less likely to achieve justiciability. Nevertheless, theoretical bridges between human rights and development continued to strengthen, particularly through human development and capability approaches. The human development approach was adopted by the United Nations Development Programme (UNDP), which published the annual Human Development Index and Human Development Reports from 1990 onwards (see http://hdr.undp.org/en/). Amartya Sen's idea of 'development as freedom' (1999) set out an ethically grounded, philosophically liberal, and individualist approach to human development.

The British development economist and founder of the UK Institute of Development Studies, Dudley Seers, directly challenged mainstream development theory, which was dominated by the concern for national economic growth:

> Economic growth will not merely fail to solve social and political difficulties; certain types of growth can actually cause them. Now that the complexity of development problems is becoming increasingly obvious, the continued addiction to the use of a single aggregative indicator [economic growth] … begins to look like a preference for avoiding the real problems.
>
> Seers, 1963: p. 77

The later 1960s saw the development of significant innovations in conceptualizing and measuring development. Seers (1969: pp. 3, 4) argued that the three major problems of development were poverty, unemployment, and inequality: "if one or two of these central problems have been growing worse, especially if all three have, it would be strange to call the result 'development,' even if per capita income soared." Morris's 'Physical Quality of Life Index' (PQLI) offered another alternative approach, evaluating development on three dimensions: infant mortality rates, life expectancy at age one, and literacy (Estes, 2003). Morris devised the PQLI to reflect the results of development efforts, instead of relying on notionally available economic resources that might not be mobilized for desirable human outcomes. Comparisons of PQLI data across countries showed that there was no obvious correlation between increasing aggregate income and welfare outcomes. This suggested that quality of life was only indirectly determined by high economic growth, and that wise government policies and spending priorities were probably more important factors. Morris was aware of the need to avoid ethnocentric assumptions, since not all countries might wish to develop in the same way. However, he insisted that the measure should be internationally comparable, settling on indicators of human survival and functioning to reflect basic conditions "from which further forms of social satisfaction could be determined" (Morris, 1978: pp. 225, 230–231).

In practical terms, the 1974 NIEO demands for global economic restructuring saw expanding developing country expectations at a time when developed country support was shrinking. In 1976, the International Labour Organization (ILO) pushed forward the concept of 'basic needs,' arguing that "it is no longer acceptable in human terms or responsible in political terms to wait several generations for the benefits of development to trickle down until they reach the poorest groups." Basic needs were defined as "minimum food, shelter, clothing, access to essential services such as transport, education, sanitation, health care, an adequately paid job, healthy environment and popular participation in decision making" (ILO, 1976). The expansive definition of 'basic' needs reflected the content of the two major global human rights treaties agreed in 1966, after almost two decades of contention – the International Convention on Civil and Political Rights and the International Convention on Economic, Social and Cultural Rights.

Despite the fractiousness that had characterized North–South relations in the decolonizing period from 1945 to 1973, these decades delivered sustained economic progress with improving welfare for large parts of the world. However, the end of the 1970s saw the introduction of neoliberalism, fiscal austerity, welfare cutbacks, and intensified militarization, all contributing toward a darker and more pessimistic context of 'de-development.' Global recession, debt crisis, and neoliberal structural adjustment policies caused the 1980s to become known as the 'impasse' or 'lost decade' of development (Schuurman, 1993).

The 1990s saw many critical development theorists taking a 'postdevelopment' turn, declaring 'development' to be 'a ruin in the intellectual landscape' (Sachs, 1999). Development theory incorporated postcolonial critiques of Eurocentrism, taking a 'cultural turn' toward nativist and populist conceptions of community and grassroots empowerment. Some political economy, class-based, and feminist theorists remained skeptical of postdevelopment's reactionary, conservative, and identitarian aspects (Brass, 2007). Development reformists, on the other hand, stuck to critically examining the dilemmas of doing development and continuing to strive for progress in difficult circumstances. Reformists continued to seek improvements in international cooperation, to alleviate suffering, and protect the basic needs of the most vulnerable, within the context of constrained development (Cornia et al., 1987). The basic needs agenda was also expressed through rights-based approaches as some attempted to define a 'minimum core' of socioeconomic rights (Young, 2008). Official development assistance declined after 1989, as Cold War geopolitical motives for assistance waned. This led to the creation of the Millennium Development Goals (MDGs) to revitalize commitments to development and focus international action on poverty alleviation. Since their introduction in 2001, the MDGs have become an accepted framework for international development as even those who initially hesitated or opposed the MDGs ended up using them (Fukuda-Parr, 2012).

Amartya Sen, Mahbub ul-Haq, and Martha Nussbaum are the best-known exponents of the 'human development paradigm.' The human development approach, launched in 1990 under the auspices of the UNDP, incorporated aspects of the PQLI and responded to the question about how economic growth related to other human needs or goals. Economic growth and income are seen as instrumental means to achieve development, but not its ultimate goal. Sen's 'capabilities approach' informed the design of the Human Development Index (HDI), a composite measure of welfare and income integrating social and economic dimensions of well-being. The HDI components are health, indicated by infant mortality and life expectancy; knowledge, indicated by school enrollment and adult literacy; and income, indicated by GNP. The HDI reunited welfare components with income measures, because the PQLI on its own "misses the synergy between social and economic progress" (Ul-Haq, 1995: p. 128). Like the PQLI, the HDI was intended to widen the scope beyond GDP/GNP as the focus of development, and support investments in health and education, without rejecting economic growth. The human development approach aspired to be "the most holistic model that exists today … a practical reflection of life itself" (Ul-Haq, 1995: p. 21). Since 1990, new dimensions have been added, incorporating expanding understandings of poverty, gender, and inequality. Capability theory has expanded into a large body of work on concept, measurement, theory, and policy, focused on particular areas such as health, hunger, gender equality, and education.

Further Directions: Sustainable Human Development and Rights-Based Approaches

The concept of 'sustainable development' entered into the development mainstream with the publication of the Brundtland Commission's report in 1987 (WCED, 1987). This concept brought the 'Southern' critique of unequal, neocolonial, and 'dependent' development together with other systemic

concerns about population growth, pollution, and ecological limits to growth, concerns that were global in scope, but largely Northern in their origin. The debate was influenced by the publication of *The Population Bomb* (Ehrlich, 1968) which revisited Malthus's controversial 1798 *Essay on the Principle of Population*:

> …that it can be a very bad thing to have more than a certain number of people alive at the same time, that Earth has a finite carrying capacity, and that the future of civilization was in grave doubt.
> Ehrlich and Ehrlich, 2009: p. 63

Malthus had opposed the provision of famine relief in the nineteenth century because he regarded famine, war, and disease as functional 'natural checks' on population. Almost two centuries later, Ehrlich suggested that the only measures that could prevent civilizational collapse were either mass birth control ('the birth rate solution') or allowing the poor to die ('the death rate solution') (Ehrlich and Ehrlich, 2009). The burden was heaviest for the poorer countries, which had not yet undergone the demographic transition (Dyson, 2010). At its extremes, Malthusianism coincided with eugenicist views that the 'least-fit' individuals must perish in order that the 'fitter' survive (Desrochers and Hoffbauer, 2009: p. 39). Malthusian pessimism was countered by 'cornucopians' who viewed population growth optimistically, as providing labor and 'talent' for development, more than compensating for the burden they represented in terms of food consumption (Dyson, 2010). Cornucopian thinking has continued to develop in the form of 'ecological modernization' and 'weak sustainability' theory, which suggests that increased consumption burdens will be compensated ('substituted') by increased knowledge and improved technology (Neumayer, 2003).

Human rights and development connected in the 1972 'Right to Development' proposal to the United Nations, articulating the human right to "a particular process of development, in which all human rights and fundamental freedoms can be fully realized." The 1986 Declaration on the Right to Development announced a humanist revolution in the meaning and purpose of development, in stating that "the human person is the central subject of development and should be the active participant and beneficiary of the right to development" (United Nations, 1986). Governments were tasked to empower citizens and guarantee them equal benefits from development as rights. In the 1970s and 1980s, the right to development tended to be claimed by developing country governments, not individual citizens, as a way to assert the 'developing country position' in international negotiations, especially after the NIEO reforms proposed in 1974 failed to materialize. However, by the 1990s more attention began to be paid to participation and benefits in more specific contexts, especially in the three 'immediate action areas' of hunger and food, health, and education (CDHR, 2004).

A recurrent theme throughout the theoretical debates in development studies has been the insufficiency of income measures and the turn to alternatives. *Sustainable human development* has been proposed as the ultimate goal; however, human development and capability theory's expansive and freedom-based conceptions of human progress may come up against environmental limits (Neumayer, 2010). Some theorists have returned to the basic needs concept, and attempted to define a 'minimum core' of human needs or values, that can be used as the basis of substantive, justiciable rights claims, defining national and international responsibilities (Young, 2008).

Advocates of sustainable human development hope that the divergence between economic and human development can be resolved by putting economic development 'back in its place' as a means of human development. However, economic growth remains a fundamental component of human development that may be impossible to sustain. From an ecological perspective, recognition for absolute planetary carrying capacity is essential. For development studies to deal with these limits, it must find new ways to integrate ethics, economics, and politics, recognizing that development problems cannot be solved by the application of scientific or technical solutions on their own. Concepts of 'environmental space' and 'new economics' try to bring together these concerns to inform and achieve equitable and sustainable development (Khoo, 2013).

Conclusion: Development Studies, Complexity, and Uncertain Futures

The critical debates in development studies have consistently questioned the meaning, relevance, and legitimacy of 'development' and asked whether development studies is asking the right questions about the right problems, whether it is justifiable as an interdisciplinary subject for study, research, and practice, and whether such a subject deserves support. Such questions are unsurprising, given the starting point for this entry is that 'development' is an emergent and contested concept. We should expect the legitimacy and relevance of 'development studies' to be contested, given the power, ideology, and resources that are at stake. Complexity and uncertainty are intrinsic to the development effort, and development studies, as a field, has been paradoxically constituted by both fierce disagreement and pragmatic consensus. Development studies' cross-disciplinary character is vindicated by the complex and multifaceted character of pressing global challenges and problems. The durable big topics and concerns include poverty, inequality, demographics, ecology and sustainability, democratization, gender equity, globalization, and rights. A rights-based approach suggests a substantive focus on the three 'immediate action areas' of the 1986 Right to Development (CDHR, 2004): food (including hunger and agriculture), health, and education – all are ultimately bound by considerations of 'environmental space,' social justice, and basic needs.

Current debates within development studies have focused on a need to 'rethink' and reimagine development studies (Haddad et al., 2011) in a world of growing complexity, increasing population, and limited resources. Although much criticized, the MDGs offered an achievable approach for development studies (Fukuda Parr, 2012) because they reduced the complexity of development challenges down to eight goals with concrete global targets, to be achieved between 2000 and 2015. The MDGs prioritized poverty, some communicable diseases, and gender equity and improved coordinated action, but their narrow emphasis detracted from other important concerns such

as inequality, or broader conceptions of health or ecology, and they oversimplified the policy debates. Haddad et al. (2011) indicate that development studies has to move away from the historical distinction between 'developed' and 'developing' countries and beyond the limited policy mindset of official development assistance to a broader global vision. The academic field of development studies has been challenged by the policy and practice community to become less 'problem focused' and more 'solution oriented' and provide answers. The theory-applied practice tension has also pushed development studies to embrace a new organizational approach that moves away from the North-dominated, 'expat-expert' driven and top-down approaches that have predominated since the 1940s and 1950s. Development studies has moved slowly toward a 'co-constructive' approach that recognizes and involves different actors on a more truly global and equal basis, through methods of research and decision-making that link bottom-up grassroots and top-down policy processes (see Haddad et al., 2011; UNDP, 2013). The UNDP (2013) suggests that development studies must learn more from the 'rise of the South,' identifying four important areas for future development studies to concentrate on – enhancing equity, enabling voice and participation, confronting environmental pressures, and managing demographic change. The UNDP suggests that future development studies could focus on issues surrounding the provision of global public goods and concentrate on finding ways to unlock stalemated global issues (UNDP, 2013: p. 20).

However, Severino and Ray (2009) argue that greatly increased funding for 'international public policies' has coincided with a lack of clarity and unity of purpose, due to the 'triple revolution' which has multiplied development objectives, players, and instruments in recent times. Another important question for development studies is how 'development' interacts with humanitarian theory, policy, and practice. Humanitarianism and development have tended to occupy separate and parallel worlds; however, current trends have seen a swing toward the integration of humanitarianism and development, with a practical focus on conflict reduction, postconflict reconstruction, social protection, and 'governance' issues (Barnett, 2013).

'Development' denotes a collective striving, a concern to envision, and programmatically pursue, societal progress. Yet 'progress' is difficult to define, and its meaning is shaped by both long-term trends and short-term crises and fads (Eade and Cornwall, 2010) and by significant contests over power, ideology, and resources. As a cross-disciplinary, contextually grounded but global field, development studies continues to be a lively, contentious, and open field, addressing necessary and relevant dilemmas of economy, society, politics, and environment, while facing fundamental questions about humanity, justice, and the responsibility for collective action.

Bibliography

Arndt, H.W., 1981. Economic development: a semantic history. Economic Development and Cultural Change 29 (3), 457–466.

Barnett, M., 2013. Empire of Humanity: A History of Humanitarianism. Cornell University Press, New York.

Barsh, R.L., 1991. The right to development as a human right: results of the global consultation. Human Rights Quarterly 13 (3), 322–338.

Brass, T., 2007. Weapons of the week, weakness of the weapons: shifts and stasis in development theory. The Journal of Peasant Studies 34 (1), 111–153.

Carson, R., 1962/2000. Silent Spring, first ed. 1962/new ed. 2000. Houghton Mifflin, Boston. Penguin Classics, London.

CDHR (Centre for Development and Human Rights), 2004. The Right to Development: A Primer. CDHR, New Delhi.

Collier, P., 2011. The Plundered Planet. Oxford University Press, Oxford.

Connell, R., 2007. Southern Theory: The Global Dynamics of Knowledge in Social Science. Allen and Unwin, Sydney.

Cooper, F., Packard, R., 1997. International Development and the Social Sciences: Essays on the History and Politics of Knowledge. University of California Press, Berkeley, CA.

Cornia, A., Jolly, R., Stewart, F., 1987. Adjustment with a Human Face: Protecting the Vulnerable and Promoting Growth. Unicef/Oxford University Press, Oxford.

Cowen, M.P., Shenton, R.W., 1995. The invention of development. In: Corbridge, S. (Ed.), Development: Critical Concepts in the Social Sciences, vol. 1. Routledge, London, pp. 27–45.

Desrochers, P., Hoffbauer, C., 2009. The post war intellectual roots of the population bomb. Fairfield Osborn's 'our plundered planet' and William Vogt's 'road to survival' in retrospect. Electronic Journal of Sustainable Development 1 (3), 37—61.

Dyson, T., 2010. Population and Development: The Demographic Transition. Zed Books, London.

Eade, D., Cornwall, A., 2010. Deconstructing Development Discourse: Buzzwords and Fuzzwords. Practical Action Publishing, Rugby.

Edwards, M., 1989. The irrelevance of development studies. Third World Quarterly 11 (1), 116–213.

Edwards, M., 2011. The Role and Limitations of Philanthropy (Commissioned paper for The Bellagio Initiative on The Future of Philanthropy and Development in the Pursuit of Human Wellbeing).

Ehrlich, P.R., 1968. The Population Bomb. Ballantine Books, New York.

Ehrlich, P.R., Ehrlich, A.H., 2009. The population bomb revisited. Electronic Journal of Sustainable Development 1 (3), 63–71.

Estes, R.J., 2003. Global change and indicators of social development. In: Weil, M. (Ed.), The Handbook of Community Practice. Sage Publications, Thousand Oaks, CA.

Forster, J., 1997. The new boundaries of international development co-operation. In: King, K., Buchert, L. (Eds.), Changing International Aid to Education. UNESCO, Paris.

Frank, A.G., 1966. The development of underdevelopment. Monthly Review 18 (7), 27–37.

Fukuda-Parr, S., 2012. Should Global Goal Setting Continue, and How, in the Post-2015 Era? DESA Working Paper No 117 United Nations Department of Economic and Social Affairs, New York. Un Document ST/ESA/2012/DWP/117.

Galbraith, J.K., 1958. The Affluent Society. Houghton Mifflin, New York.

Gilman, N., 2007. Mandarins of the Future: Modernization Theory in Cold War America. Johns Hopkins University Press, Baltimore.

Haddad, L., Hossain, N., McGregor, J.A., Mehta, L., 2011. Introduction: time to reimagine development? IDS Bulletin 42 (5), 1–12.

Hancock, W.K., 1942. Survey of British Commonwealth Affairs Volume 2 – Problems of Economic Policy 1918–1939, (Part 2). Oxford University Press, London.

Hettne, B., 1995. Development Theory and the Three Worlds: Towards an International Political Economy of Development. Longman, London.

ILO (International Labour Organization), 1976. Employment, Growth and Basic Needs: A One World Problem. ILO, Geneva.

Jackson, T., 2011. Prosperity Without Growth. Routledge, London.

Kaplinsky, R., 2005. Globalization, Poverty and Inequality: Between a Rock and a Hard Place. Polity, Cambridge.

Khoo, S., 2013. Sustainable development of what? Contesting global development concepts and measures. In: Fahy, F., Rau, H. (Eds.), Methods of Sustainability Research in the Social Sciences. Sage, London, Thousand Oaks, CA and New Delhi, pp. 91–113.

Krugman, P., 6 September 2009. How did economists get it so wrong? New York Times. http://www.nytimes.com/2009/09/06/magazine/06Economic-t.html?_r=0 (accessed 31.07.13.).

Marglin, S., Schor, J., 1990. The Golden Age of Welfare Capitalism. Oxford University Press, Oxford.

Mishan, E.J., 1967. The Costs of Economic Growth. Pelican, Harmondsworth.

Morris, M.D., 1978. A physical quality of life index. Urban Ecology 3, 225–240.

Neumayer, E., 2003. Weak versus Strong Sustainability: Exploring the Limits of Two Opposing Paradigms, second ed. Edward Elgar, Cheltenham.

Neumayer, E., 2010. Human Development and Sustainability. Human Development Research Paper 2010/05. United Nations Development Programme, Geneva.
Randeria, S., 2003. Cunning states and unaccountable international institutions: legal plurality, social movements and rights of local communities to common property resources. Archives of European Sociology 44 (1), 27–60.
Rist, G., 1997. The History of Development: From Western Origins to Global Faith. Zed Books, London.
Rostow, W.W., 1960. The Stages of Economic Growth: A Non-Communist Manifesto. Cambridge University Press, Cambridge.
Sachs, W., 1999. Planet Dialectics: Explorations of Environment and Development. Zed Press, London.
Sauvy, A., 1952. Trois Mondes, Une Planète. L'Observateur politique, économique et littéraire, 118, 14. http://www.homme-moderne.org/societe/demo/sauvy/3mondes.html (accessed 31.07.13.).
Schuurman, F. (Ed.), 1993. Beyond the Impasse: New Directions in Development Theory. Zed Books, London.
Seers, D., 1963. The limitations of the special case. Bulletin of the Oxford Institute of Statistics and Economics 25 (2), 77–98.
Seers, D., 1969. The meaning of development. International Development Review 11 (4), 3–4.
Severino, J.M., Ray, O., 2009. The End of ODA: Death and Rebirth of a Global Public Policy. Center for Global Development Working Paper Number 167. Center for Global Development, Washington, DC. http://www.cgdev.org/files/1421419_file_End_of_ODA_FINAL.pdf (accessed 31.07.13.).
Sumner, A., Tribe, M., 2008. International Development Studies: Theories and Methods in Research Practice. Sage, London.
Tadros, M., 2011. The faith factor in reimagining development. IDS Bulletin 42 (5), 63–67.
Ul Haq, M., 1995. The birth of the human development index. In: Fukuda-Parr, S., Shiva Kumar A.K. (Eds.), 2003, Readings in Human Development: Concepts, Measures and Policies for a Development Paradigm. Oxford University Press, Oxford, pp. 127–137.
United Nations, 1974. Declaration on the Establishment of a New International Economic Order 1 May 1974. Geneva: United Nations Document A/RES/S-6/3201.
United Nations, 4 December 1986. Declaration on the Right to Development. Geneva: United Nations Document A/RES/41/128.
UNDP (United Nations Development Programme), 2013. The Rise of the South: Human Progress in a Diverse World. Human Development Report 2013. UNDP, New York.
Vanoli, A., 2008. National accounting, history of. In: Durlauf, S.N., Blume, L.E. (Eds.), The New Palgrave Dictionary of Economics, second ed. http://www.dictionaryofeconomics.com/article?id=pde2008_N000160&edition=current&q=History (accessed 21.07.12.).
WCED (World Commission on Environment and Development), 1987. Report of the World Commission on Environment and Development: Our Common Future. UN Document A/42/427. http://www.un-documents.net/a42-427.htm (accessed 30.07.13.).
Woodhouse, P., Chimhowu, A., 2005. Development studies, nature and natural resources: changing narratives and discursive practices. In: Kothari, U. (Ed.), A Radical History of Development Studies. Zed Books, London, pp. 180–199.
Young, K., 2008. The minimum core of economic and social rights: a concept in search of content. Yale Journal of International Law 33 (1), 113–175.

Development Theory in Geography

Jim Glassman, University of British Columbia, Vancouver, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Development geography since the end of World War II has been transformed from a subdiscipline that largely borrowed from major themes in development theory to one that helps develop major themes, this change reflecting the increasing salience of sociospatial issues in an era of globalization.

Introduction

The issues which geographers have placed on the agenda of development theory in the post-World War II period fall broadly under the heading of what can be called the 'spatial problematic.' This problematic includes issues such as reasons for developmental patterns of diffusion and/or geographic difference, as well as the relationships between development processes in different locations. The spatial problematic can also be taken to embrace issues of human–environment interaction in the development process, including patterns of resource use, food production and consumption, and generation of environmental hazards. By the end of the Cold War 'development decades' (Sheppard et al., 2009), in the 1980s and early 1990s, neoliberal globalization generated a context in which geography as a field, and as an aspect of social reality, had become central to the formulation of theoretical positions surrounding development issues, a position of theoretical centrality that geography has maintained into the twenty-first century.

The Development of Post-World War II Development Geography

Given the obvious significance of topics falling within the spatial problematic, it is surprising that until the last two decades of the twentieth century the contributions of geographers to development theory in the English-speaking world were comparatively limited. A 1960 report by leaders in the field estimated that only two dozen studies had been published on the issue of economic development between 1949 and 1959, and only half of these were on development in 'underdeveloped areas' (Ginsburg, 1960, p. ix). Similarly, a 1967 article by D.E. Keeble showed that between 1955 and 1964, the journal *Economic Geography* published 251 major articles, only 10 of which (under 4%) were concerned in whole or in part with issues of economic development, while the topic was addressed in only 6 out of 242 articles (2.5%) in the *Annals of the Association of American Geographers* (Keeble, 1967, p. 243). A review of research published in *The New Zealand Geographer* cited a similar neglect of development theory (Chapman, 1969).

The theories of development employed in this early postwar literature were fundamentally those that had been developed in disciplines such as economics, political science, and sociology. In particular, most work by geographers in the 1960s fell more or less explicitly into the category of 'modernization' studies, adopting the view that all countries could and should follow more or less the same course of development as the advanced industrialized capitalist countries, passing through 'stages of growth' which lead from premodern, preindustrial, and relatively stagnant to industrial, modern, and dynamic. Geographers gave modernization approaches a spatial bearing through examination of issues such as urbanization patterns (Berry, 1961), regional development (Friedmann, 1966), and the development of transportation systems (Gould, 1960; Taafe et al., 1963), as well as through statistical analyses of the diffusion of modernization (Gould, 1964).

By the time development geographers had embraced modernization approaches, however, modernization theories had already been under serious attack for a decade or more by various 'core–periphery theorists' – including structuralist and dependency authors – who placed the claims of modernization theory up against the realities of development in peripheral regions of the global economy such as Latin America, and found the theory wanting. Structuralists argued that the historical legacies of colonialism and imperialism – such as technological backwardness, maldistribution of wealth and income, and a large, underemployed rural labor force – made the development of modern industry in the periphery a much more problematic process than modernization theorists had suggested (Sheppard et al., 2009, pp. 84–86). Dependency authors more explicitly identified the process of development in core industrialized capitalist countries as an obstacle to development in countries of the global periphery, and more radical dependency and underdevelopment perspectives which became popular by the late 1960s went so far as to suggest the impossibility or unlikelihood of capitalist development in the global periphery, precisely because of its relations with the global core (Sheppard et al., 2009, pp. 86–89).

These structuralist and dependency perspectives, even though not originally articulated by geographers, helped bring greater attention to geographic differences in patterns of development and to relationships between development processes in different places. This spatial dimension was naturally attractive to many geographers and spurred a significant amount of core–periphery analysis, including theories regarding revolutionary change in Third World urban centers (Armstrong and McGee, 1968), general theories of imperialism (Blaut, 1970), and critiques of Eurocentrism in geographical research (Connell, 1971). By the middle of the 1970s there were a large number of works adopting either

structuralist (e.g., Brookfield, 1975) or dependency approaches (e.g., Santos, 1974). This burst of core–periphery analyses opened a period in which attention to development issues in geography became much more sustained. Between 1974 and 1983, for example, articles which addressed development issues rose to 16% of all articles in *Economic Geography* and 13% in the *Annals*, while articles on development in the more recently founded *Antipode* reached almost 40% (Corbridge, 1986, pp. 4–5).

Even with this more intensive theorization of development in the Global South and deployment of the spatial problematic, however, development geography remained more a theoretical borrower than a theoretical leader. By the mid-1970s, dependency approaches were already coming under increasing fire from Marxist scholars for fetishizing geographical relations and miscomprehending class issues, and were thus being forced into refinements that could take account of increasing complexity in the 'world-system' (Sheppard et al., 2009, pp. 81, 89–92), including the possibility of 'dependent development' in what had formerly been peripheral areas. Moreover, the 1970s had seen the emergence within development studies of important work on gender issues (Sheppard et al., 2009, pp. 92–94), an emphasis which was at that point absent in work by most development geographers.

By the late 1970s and early 1980s, a more fully fledged turn in the direction of Marxist theory was evident in the work of many development geographers (Slater, 1977), while others had taken note of the importance of world-systems theory (Taylor, 1981). Eventually, development geographers also paid greater attention to the claims about dependency put forward by the partisans of 'dependent development' approaches (Forbes, 1984; Corbridge, 1986), and relatively complex arguments about the interaction of 'external' and 'internal' factors began to appear in works on issues such as regional development (Gore, 1984) and urbanization (Armstrong and McGee, 1985).

Within development theory in general, however, the issues raised by both dependency theory and its Marxist critics had largely faded from view in the 1980s. For neoclassically oriented economists, the dramatic growth of East Asian newly industrialized countries (NICs) in the 1970s and 1980s both undermined claims of blocked development on the periphery and vindicated assertions that unfettered markets performed better than those in which the state intervened (Sheppard et al., 2009, pp. 97–104). These arguments were countered by the claims of an emerging 'neo-Weberian' school, which presented evidence for the important role of the state in guiding the economic growth of the Asian NICs (Glassman and Samatar, 1997).

In spite of the fact that theorization of Third World states by geographers has been limited (Glassman and Samatar, 1997; Glassman, 1999), emphasis on the political and economic role of the state was by no means foreign to development geography. Indeed, by the mid-1980s a number of development geographers had already begun to take up the theme of state activism in the East Asian growth story (Grice and Drakakis-Smith, 1985; Browett, 1986). Moreover, development geographers also foregrounded the spatial problematic in discussions of the state by highlighting state roles in creating and maintaining regional disparities within developing countries (Watts, 1983; Samatar, 1989; Slater, 1989).

Work on issues such as gender and development began to gain more attention in development geography by the end of the 1980s (e.g., Momsen and Townsend, 1987). These included analyses of the gendering of immigration processes (Radcliffe, 1986) and of the gendered environmental outcomes of development (Rocheleau, 1987).

By the 1980s and 1990s, geography as a field had begun to make what were arguably leading contributions to the formulation of theoretical positions surrounding contemporary development issues. To some extent, this was less a matter of geography as a field catching up with development studies than it was a matter of development studies coming around to some of the issues which are part of the spatial problematic – in particular, transnational geographic expansion of economic processes (neoliberal globalization), human–environment interactions, and difference and spatiality.

Yet, even as development geography and the rest of the social sciences began to increasingly speak a 'spatial' language, further permutations in the social sciences began to transform the ways development geographers worked and the ways they talked about spatiality. In particular, in the era of neoliberal globalization, and especially from the 1990s to the present, a combination of feminist theory, neo-Foucauldian influences, postcolonial theory, and postdevelopment approaches has generated new research agendas that in part build on the previous spatial problematic and in part transform it (Sheppard et al., 2009, pp. 104–112; Slater, 2004). The general move toward such 'post-prefixed' approaches has been enabled by a growing sense that in a post-Cold War world globalizing capitalist development defines the immediate future of virtually all countries. As such, development theory has tended to evolve past debates of the Cold War era that posed socialism as a short-term possibility within much of the Global South. Instead, a series of postdevelopment perspectives have emerged, calling into question more radically than had dependency, Marxist, or neo-Weberian perspectives the very project of development, whether in capitalist or any other form.

Lines of postdevelopmental research and theorization that extend to the present have generated much unresolved debate, including over some of the basic historical and theoretical contentions of the postdevelopmentalists (Cowen and Shenton, 1996; Mitchell, 2002) and the degree to which development can or should be opposed (Gidwani, 2008; Wainwright, 2009). More specifically, development geographies of the neoliberal globalization era have placed more firmly on the agenda issues such as the gendering and racialization of development (e.g., Perrault, 2003; Bakker and Silvey, 2008; Andolina et al., 2009), the imperial and statist violence that has been integral to actually existing development (e.g., Harvey, 2003; Gregory, 2004; Nevins, 2005; Retort, 2005), and the geographical imaginaries that are part of globalization and regionalization (e.g., Slater, 1997; Sidaway, 2002, 2012; Glassman, 2010). The spatial problematic has been enriched by the emergence of these issues within development geography, with debates in the field – for better or worse – becoming in some ways both more enlivening and less focused than in the earlier post-Cold War period.

Contemporary Themes in Development Geography

While modernization approaches had been generally optimistic about the prospects for capitalist development in developing countries, much of the literature by dependency theorists had been highly pessimistic. Since at least the 1980s, however, many development theorists have been inclined toward a more contextual approach that recognizes both possibilities for and limitations to capitalist development. In place of the earlier emphasis on either diffusion or blocked development, there is now much greater emphasis on the general geographical expansiveness of capital and the unevenness with which its expansion proceeds – at all scales from the global, transnational, and national to the subnational, local, and bodily. The important theoretical work of David Harvey (1982/2006; 2003) has enriched the conceptual foundations for analyses highlighting uneven internationalization of capital – analyses which have been carried forward by a variety of development geographers (see Smith, 1984/2008).

Much work on uneven development by geographers has focused on uneven development at the national scale, especially uneven economic growth between regions (Watts, 1983; Samatar, 1989; Slater, 1989; Park, 2003). The most basic point to emerge from such studies has been that capitalist development – whether or not it is relatively dynamic in a given context – occurs as the result of differential power relations that have concrete spatial expression. These differential power relations lie behind the spatial unevenness of capitalist development and tend to work at odds with efforts to create greater sociospatial equality in the development process. Indeed, in countries that have been relatively successful at 'late' industrialization, the production of disparity is sometimes accelerated rather than slowed (Glassman, 2004, 2010).

Directly related to the phenomenon of national unevenness in the development process are issues surrounding urbanization in developing countries. While it is less fashionable now than it once was to speak of 'over-urbanization,' there is still much emphasis on how regional disparities in growth relate to the preeminence of primate urban centers – these often being former centers of colonial administration and/or commercial centers of long standing. Instead of gradual equilibration of growth rates between these centers and outlying regions, capitalist development has generally increased the dominance of primate urban centers. Indeed, it has been suggested by development geographers studying Asia that spill over from expansion of primate urban centers has led to a historically unique form of extended metropolitan growth, which blurs preconceived boundaries between urban and rural, industry and agriculture (McGee, 1991).

At the same time, the explosive growth of particular urban centers in the context of global economic transformation has led to increasing investigation of the ways uneven development detaches the trajectories of urban centers from their 'hinterlands' while creating global networks of cities through which there is an intensified flow of money, goods, and information (Friedmann, 1986; Taylor, 2004). On this view, the spatial unevenness of development increasingly manifests itself as a bifurcation between the major nodes through which capital moves internationally (large urban centers) and much of the rest of the world, with nonurban areas and slums of the developing world being particularly marginalized (Davis, 2006). Thus, as many contemporary core–periphery theorists see the matter, in the era of neoliberal globalization the global core–periphery structure has been transformed from a global hierarchy of states shot through by global commodity chains (Wallerstein and Hopkins, 1986) to a global city network in which capitalist monopoly power is increasingly centered in networks of cities housing crucial advanced producer services (Brown et al., 2010). In this global city network, much of the rest of the world is seen as 'yoked' to this city network as a peripheral supplier of resources and labor (Taylor, 2007). In the neoliberal world of globalization, the ultimate expression of global uneven development is thus less the global hierarchy of states – though this still exists – than the relationship between dominant cities (and the social classes residing in them) and what Mike Davis calls the 'planet of slums' – the vast populations all over the world, still predominantly in the Global South, that increasingly face a dismal future of insecure informal sector employment, substandard housing, noxious living and working environments, social violence, and generally difficult, truncated lives (Davis, 2006).

Both the emphasis on networks of global cities and the emphasis on slum settlements can threaten a one-sided and monolithic vision of capitalist globalization, a vision contested by postcolonial geographers arguing for more attention to 'ordinary cities' (Robinson, 2006). Moreover, even with the transformations of the geographical imagination of development pursued by world systems and world cities theorists, to present the unevenness of capitalist development in too monolithic a fashion risks reification of the spatial dimensions of development at the expense of a spatially grounded analysis of social relations. Geographers have attempted to guard against such 'spatial fetishism' – that is, the view that geographical relations have their own independent dynamics, rather than being the spatial expressions of social processes – by foregrounding issues that include the spatial dynamics of class struggles and gender relations. Uneven development thus implies not only uneven development between places but uneven power relations and opportunities for groups of people, dependent upon such characteristics as their class, gender, race or ethnicity, sexual orientation, and age, with considerable variability in social outcomes, contingent on specific histories and contemporary social struggles (Hart, 2003; Roy, 2003).

Among development geographers, particular attention has been paid to the spatialization of disparity in the development process. For example, feminist geographers have examined how transformations in Mexico's accumulation process have involved breaking down the hegemony of the family wage-based, male-dominated, and Mexico City-centered labor regime while substituting a lower-wage and geographically more decentralized regime in which women are employed in much larger numbers, while being increasingly regarded as disposable (Cravey, 1998; Wright, 2006). They have also examined the migration of women to export processing zones, which have been the key to economic growth in newly industrializing countries of Southeast Asia, China, and Latin America (Chant and Radcliffe, 1992; Chant and McIlwain, 1995; Wright, 2006). Labor geographers have examined the

strengths and weaknesses of attempts to create various forms of solidarity between labor and other popular organizations in developed countries such as the USA, and less developed countries such as those of Latin America (Herod, 1995; Johns, 1998). Authors writing about 'new social movements' and 'popular development' (Brohman, 1996) have examined how such movements attempt to forge international linkages ('jumping scale') in order to strengthen their position (Cumbers et al., 2008). Geographers examining state power have noted both how it is transformed by processes such as 'globalization' and how it continues to play a crucial role in relation to various kinds of social struggles through which 'globalization' occurs (Agnew and Corbridge, 1995; Yeung, 1998; Glassman, 1999).

All of these types of analyses bring an explicitly geographical dimension to social critique of globalization. What is either implicit or explicit in such critique is an emphasis on the way general processes, such as capitalist development, can have different types of manifestations in different locations, because of the relationships between different social groups within and across those locations (Nagar et al., 2002). Thus, the study of development geographies requires balancing an adequately supple general theory with examination of the specifics of development in particular contexts. Theories of capitalist development must be able to explain why the features of capitalism which are posited as general can lead to different kinds of outcomes – e.g., 'late industrialization' in one setting and 'blocked development' in another; stagnant real wages in one setting and rising real wages in another; increasing marginalization of certain women workers in one setting and new opportunities for empowerment of women in another.

The variegated outcomes of global capitalist development include its varied environmental impacts. Within development geography, an important literature on issues of political ecology emerged in the 1990s and has since then taken up the task of theorizing environmental change in the context of globalization (Peet and Watts, 1996). Political ecology can be seen as combining a focus on details of local ecosystems and social structures with attention to global forces. In addition, political ecologists have shown a strong interest in what Peet and Watts (1996) term 'liberation ecologies' – varied struggles to transform environmental conditions in developing countries, which are often approached by political ecologists through post-prefixed conceptual frameworks, emphasizing the multiclass and frequently community-based forms of local environmental resistance (e.g., Tsing, 2004; Wolford, 2010). A number of political ecologists have brought gender and race into the center of the research agenda, highlighting, for example, the importance of gender-differentiated control over local and household resources to the economic and environmental outcomes of development projects, and the racialization of different forms of agricultural development (e.g., Carney, 1993, 2001). More generally, feminist political ecologists have noted how processes of transformation which are occurring in the global economy may have varied effects on women and men in specific localities, depending upon the social position in which they start and the sorts of resources to which they have (or are denied) access (Rocheleau et al., 1996).

Political ecologists have paid special attention to issues of food and agriculture. Michael Watts' study of uneven development and the food crisis in Nigeria (1983) helped bring the issue of food and access to it firmly onto the agenda, and subsequent research by development geographers has continued to analyze access to and control over agricultural resources (Samatar, 1989; Gidwani, 2008). Topics examined in relation to such issues of access and control include the impact of contract farming on local food production (Little and Watts, 1994), the impact of new agricultural technologies on poverty and food security (Yapa, 1993, 1996), and the ways varied global processes enable or disable forms of struggle, not only over resources but over the conceptualization of resources and social space (Wainwright, 2009), including its relationship to indigenous identities. Given the monumental social transformations underway in China, it is not surprising that a significant and robust literature has emerged charting varied transformations of land use in China, in both urban and rural settings (e.g., Lin, 2009; Hsing, 2010; Webber, 2012). Likewise, given the enormous importance of petroleum and other extractive industries to global capitalism, it is not surprising that such industries have begun to receive sustained and detailed attention from development geographers (e.g., Bebbington, 2012; Bridge and LeBillon, 2013).

Conclusion

The various but often overlapping and interconnected strands of research identified here, while they all exemplify the spatial problematic in development geography, may defeat any attempt at any simple synthesis – though attempts are made to present development geography as a unity in diversity (e.g., Sheppard et al., 2009). Such presentations of development geography's unity in diversity connect the traditions of political economy of development and political ecology with postmodern and postcolonial approaches to development studies. It is at the conjunction of these approaches, particularly as they address various issues surrounding neoliberal globalization, that much of the most interesting work is currently being conducted. In the twenty-first century, the centrality of globalization, uneven development, environmental challenges, and popular responses to these processes by diverse groups seems clear; and much of the specific research carried out by development geographers is related to this central problematic. At the same time, the pace and scale of global change itself makes attempts to predict where development theory in geography will head in the future a daunting task.

See also: Agroecology and Agricultural Change; Cities: Capital, Global, and World; Cultural Geography; Development Theory in Geography; Economic Geography; Economic Transformation: From Central Planning to Market Economy; Environment and Development; Feminist Political Theory and Political Science; Food Security and 'Green Revolution'; Foreign Investment: Direct; Gender and Feminist Studies in Geography; Gene–Environment Interactions in Health and

Well-Being; Globalization: Geographical Aspects; International Trade: Commercial Policy and Trade Negotiations; Marxist Geography; Political Ecology: Anthropological Perspectives; Postcolonial Geography; Postmodernism in Sociology; Science and Social Movements; Social Movements, Geography of; Space and Social Theory in Geography; Spatial Equity.

Bibliography

Agnew, J., Corbridge, S., 1995. Mastering Space: Hegemony, Territory and International Political Economy. Routledge, London.

Andolina, R., Laurie, N., Radcliffe, S., 2009. Indigenous Development in the Andes: Culture, Power, and Transnationalism. Duke University Press, Durham and London.

Armstrong, W.R., McGee, T.G., 1968. Revolutionary change and the Third World city: a theory of urban involution. Civilisations 18 (3), 353–377.

Armstrong, W.R., McGee, T.G., 1985. Theatres of Accumulation: Studies in Asian and Latin American Urbanization. Methuen, London.

Bakker, I., Silvey, R., 2008. Beyond States and Markets: The Challenges of Social Reproduction. Routledge, Oxon and New York.

Bebbington, A., 2012. Social Conflict, Economic Development, and Extractive Industry: Evidence from South America. Routledge, Oxon and New York.

Berry, B.J.L., 1961. City-size distributions and economic development. Economic Development and Cultural Change 9, 573–587.

Blaut, J.M., 1970. Geographic models of imperialism. Antipode 2 (1), 65–85.

Bridge, G., LeBillon, P., 2013. Oil. Polity Press, Cambridge and Malden.

Brohman, J., 1996. Popular Development: Rethinking the Theory and Practice of Development. Blackwell, Oxford.

Brookfield, H.C., 1975. Interdependent Development. University of Pittsburgh Press, Pittsburgh, PA.

Browett, J.G., 1986. Industrialization in the global periphery: the significance of the Newly Industrializing Countries of East and Southeast Asia. Environment and Planning D. Society and Space 4 (4), 401–418.

Brown, E., Derudder, B., Parnreiter, C., Pelupessy, W., Taylor, P.J., Witlox, F., 2010. World City Networks and Global Commodity Chains: towards a world systems' integration. Global Networks 10 (1), 12–24.

Carney, J., 1993. Converting the wetlands, engendering the environment: the intersection of gender with agrarian change in the Gambia. Economic Geography 69 (4), 329–348.

Carney, J., 2001. Black Rice: The African Origins of Rice Cultivation in the Americas. Harvard University Press, Cambridge.

Chant, S.H., McIlwaine, C., 1995. Women of a Lesser Cost: Female Labour, Foreign Exchange, and Philippine Development. Pluto Press, London.

Chant, S., Radcliffe, S. (Eds.), 1992. Gender and Migration in Developing Countries. Belhaven, London.

Chapman, M., 1969. Geography and the study of development. Journal of Developing Areas 3, 319–338.

Connell, J., 1971. The geography of development. Area 3 (4), 259–265.

Corbridge, S., 1986. Capitalist World Development: A Critique of Radical Development Geography. Rowman & Littlefield, Totowam, NJ.

Cowen, M.P., Shenton, R.W., 1996. Doctrines of Development. Routledge, London.

Cravey, A.J., 1998. Women and Work in Mexico's Maquiladoras. Rowman & Littlefield, Lanham, MD.

Cumbers, A., Routledge, P., Nativel, C., 2008. The entangled geographies of global justice networks. Progress in Human Geography 32 (2), 183–202.

Davis, M., 2006. Planet of Slums. Verso, London and New York.

Forbes, D., 1984. The Geography of Underdevelopment: A Critical Survey. Croom Helm, London.

Friedmann, J., 1966. Regional Development Policy: A Case Study of Venezuela. MIT Press, Cambridge, MA.

Friedmann, J., 1986. The world city hypothesis. Development and Change 17, 69–84.

Gidwani, V., 2008. Capital, Interrupted: Agrarian Development and the Politics of Work in India. University of Minnesota Press, Minneapolis, MN.

Ginsburg, N.S. (Ed.), 1960. Essays on Geography and Economic Development. University of Chicago Press, Chicago.

Glassman, J., 1999. State Power Beyond the 'Territorial Trap': The Internationalization of the State. Political Geography 18 (6), 669–696.

Glassman, J., 2004. Thailand at the Margins: Internationalization of the State and the Transformation of Labour. Oxford University Press, Oxford and New York.

Glassman, J., 2010. Bounding the Mekong: The Asian Development Bank, China, and Thailand. University of Hawai'i Press, Honolulu, HI.

Glassman, J., Samatar, A.I., 1997. Development geography and the third-world state. Progress in Human Geography 21 (2), 164–198.

Gore, C.G., 1984. Regions in Question: Space, Development Theory and Regional Policy. Methuen, London.

Gould, P.R., 1960. The Development of the Transportation Pattern in Ghana, vol. 5. Northwestern University Studies in Geography.

Gould, P.R., 1964. A note on research into the diffusion of development. Journal of Modern African Studies 2, 123–125.

Gregory, D., 2004. The Colonial Present: Afghanistan, Palestine, Iraq. Blackwell, Malden, Oxford, and Carlton.

Grice, K., Drakakis-Smith, D., 1985. The role of the state in shaping development: two decades of growth in Singapore. Transactions (Institute of British Geographers) 10 (3), 347–359.

Hart, G., 2003. Disabling Globalization: Places of Power in Post-Apartheid South Africa. University of California Press, Berkeley, CA.

Harvey, D., 1982/2006. The Limits to Capital. University of Chicago Press, Chicago.

Harvey, D., 2003. The New Imperialism. Oxford University Press, Oxford.

Herod, A., 1995. The practice of international labor solidarity and the geography of the global economy. Economic Geography 71 (4), 341–363.

Hsing, Y.-T., 2010. The Great Urban Transformation: Politics of Land and Property in China. Oxford University Press, Oxford and New York.

Johns, R.A., 1998. Bridging the gap between class and space: US worker solidarity with Guatemala. Economic Geography 74 (3), 252–271.

Keeble, D.E., 1967. Models of economic development. In: Chorley, R.J., Haggett, P. (Eds.), Models in Geography. Methuen, London.

Lin, G.C.S., 2009. Developing China: Land, Politics, and Social Conditions. Routledge, London.

Little, P.D., Watts, M.J. (Eds.), 1994. Living under Contract: Contract Farming and Agrarian Transformation in Sub-saharan Africa. University of Wisconsin Press, Madison, WI.

McGee, T.G., 1991. The emergence of desakota regions in Asia: expanding a hypothesis. In: Ginsburg, N., Koppel, B., McGee, T.G. (Eds.), The Extended Metropolis: Settlement Transition in Asia. University of Hawaii Press, Honolulu, HI, pp. 3–25.

Mitchell, T., 2002. Rule of Experts: Egypt, Techno-politics, Modernity. University of California Press, Brekeley, CA.

Momsen, J.H., Townsend, J.G. (Eds.), 1987. Geography and Gender in the Third World. Hutchinson, London.

Nagar, R., McDowell, L., Lawson, V., Pratt, G., 2002. Locating globalization: feminist (re)readings of the subjects and spaces of globalization. Economic Geography 78 (3), 257–284.

Nevins, J., 2005. A Not-So-Distant Horror: Mass Violence in East Timor. Cornell University Press, Ithaca, NY.

Park, B.-G., 2003. Politics of Scale and the Globalization of the South Korean Automobile Industry. Economic Geography 79 (2), 173–194.

Peet, R., Watts, M. (Eds.), 1996. Liberation Ecologies: Environment, Development, Social Movements. Routledge, London.

Perrault, T., 2003. Changing Places: transnational networks, ethnic politics, and community development in the Ecuadorian Amazon. Political Geography 22 (1), 61–88.

Radcliffe, S., 1986. Gender relations, peasant livelihood strategies and mobility: a case study from Cuzco, Peru. Bulletin of Latin American Research 5 (2), 29–47.

Retort, 2005. Afflicted Powers: Capital and Spectacle in a New Age of War. Verso, London and New York.

Robinson, J., 2006. Ordinary Cities: Between Modernity and Development. Routledge, Oxon and New York.

Rocheleau, D., 1987. Women, trees and tenure: implications for agroforestry research and development. In: Raintree, J. (Ed.), Land, Trees and Tenure. ICRAF, Nairobi and Land Tenure Center, Madison, WI.

Rocheleau, D., Thomas-Slayter, B., Wangari, E. (Eds.), 1996. Feminist Political Ecology: Global Issues and Local Experiences. Routledge, London.

Roy, A., 2003. City Requiem, Calcutta: Gender and the Politics of Poverty. University of Minnesota Press, Minneapolis, MN.

Samatar, A.I., 1989. The State and Rural Transformation in Northern Somalia 1884–1986. University of Wisconsin Press, Madison, WI.

Santos, M., 1974. Geography, Marxism, and underdevelopment. Antipode 6, 1–9.
Sheppard, E., Porter, P., Faust, D., Nagar, R., 2009. A World of Difference: Encountering and Contesting Development. Guilford Press, New York.
Sidaway, J., 2002. Imagined Regional Communities: Integration and Sovereignty in the Global South. Routledge, New York and London.
Sidaway, J., 2012. Geographies of Development: New Maps, New Visions? Professional Geographer 64 (1), 49–62.
Slater, D., 1977. Geography and underdevelopment—II. Antipode 9, 1–31.
Slater, D., 1989. Territory and State Power in Latin America: The Peruvian Case. St. Martin's Press, New York.
Slater, D., 1997. Geopolitical imaginations across the North-South divide: issues of difference, development and power. Political Geography 16 (8), 631–653.
Slater, D., 2004. Geopolitics and the Post-Colonial: Rethinking North-South Relations. Blackwell, Malden, Oxford, and Carlton.
Smith, N., 1984/2008. Uneven Development: Nature, Capital, and the Production of Space. University of Georgia Press, Athens.
Taafe, E.J., Morrill, R.L., Gould, P.R., 1963. Transport expansion in underdeveloped countries: a comparative analysis. Geographical Review 53, 503–529.
Taylor, P.J., 1981. Political geography and the world economy. In: Burnett, A.D., Taylor, P.J. (Eds.), Political Studies from Spatial Perspectives. Wiley, Chichester, UK.
Taylor, P.J., 2004. World City Network: A Global Urban Analysis. Routledge, London.
Taylor, P., 2007. World City Network and Planet of Slums. Access and Exclusion in Neoliberal Globalization, at http://www.lboro.ac.uk/gawc/rb/rb239.html.
Tsing, A., 2004. Friction: An Ethnography of Global Connection. Princeton University Press, Princeton, NJ.
Wainwright, J., 2009. Decolonizing Development: Colonial Power and the Maya. Blackwell, Malden, Oxford, and Carlton.
Wallerstein, I., Hopkins, T., 1986. Commodity chains in the world economy Prior to 1800. Review 10 (1), 157–170.
Watts, M., 1983. Silent Violence: Food, Famine & Peasantry in Northern Nigeria. University of California Press, Berkeley, CA.
Webber, M., 2012. Making Capitalism in Rural China. Edward Elgar, Cheltenham and Northampton.
Wolford, W., 2010. This Land Is Ours Now: Social Mobilizations and the Meanings of Land in Brazil. Duke University Press, Durham.
Wright, M., 2006. Disposable Women and Other Myths of Global Capitalism. Routledge, New York and Oxon.
Yapa, L., 1993. What are improved seeds? an epistemology of the green revolution. Economic Geography 69 (3), 254–273.
Yapa, L., 1996. What causes poverty?: a postmodern view. Annals of the Association of American Geographers 86 (4), 707–728.
Yeung, H.W.-C., 1998. Capital, State, and Space: contesting the Borderless World. Transactions of the Institute of British Geographers 23 (3), 291–309.

Developmental Behavioral Genetics and Education

Frank M Spinath, Saarland University, Saarbruecken, Germany
Birgit Spinath, Ruprecht-Karls-University, Heidelberg, Germany
Peter Borkenau, Martin-Luther-University, Halle-Wittenberg, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by P. Borkenau, volume 6, pp. 3604–3607, © 2001, Elsevier Ltd.

Abstract

Behavior genetics is the study of genetic and environmental influences on individual differences in behavior. Over the past decades, this rapidly evolving field of research has provided compelling evidence for substantial genetic contributions to all important domains of human behavior. Therefore, human development cannot be fully understood without comprehension of the logic of behavior genetics research and knowledge about its main findings. The present article will focus on the quantitative-genetic approach and primarily on twin studies constituting the bulk of the pertinent evidence. We will begin by describing the logic of twin studies and defining the major concepts of quantitative genetics such as heritability as well as shared and nonshared environmental effects. Furthermore, the assumptions underlying the classic twin designs as well as more recent methodological developments are surveyed and findings in the areas of ability and education with a focus on developmental behavioral genetics will be presented. One of the most important findings is that not only individual prerequisites of learning, such as intelligence, are highly heritable, but that this is also true for educational outcomes, such as school achievement. Further counterintuitive findings are that the influence of genes on behavior becomes stronger with increasing age and that it varies in accordance with socioeconomic status. In the final section, we discuss the common misconceptions about the implications of these findings for education.

Introduction

Behavior genetics, the study of genetic and environmental origins of individual differences in behavior, is an interdisciplinary area combining the behavioral sciences and genetics. Today's field of behavior genetics encompasses diverse research strategies – such as twin and adoption studies – called *quantitative genetics*, which investigate the relative importance of genetic and environmental influences, as well as strategies to identify specific genes associated with individual differences in behavior – such as allele association studies – called *molecular genetics* (Plomin et al., 2008). Behavior genetics is arguably one of the most rapidly evolving scientific fields today, characterized by a breathtaking pace of discoveries and methodological advances that continues to accelerate. A generation ago, it seemed that the outcome of the nature–nurture debate depended on the degree to which twin and adoption studies revealed substantial genetic contributions to important domains of human behavior. This research has yielded almost unanimous results indicating that individual differences in all behaviors are subject to genetic influence. It is now generally accepted that genetic influences on behavior are ubiquitous, so much so that Turkheimer (2000) summarized this fact as "the first law of behavior genetics". This general prevalence of genetic influences is of critical importance for our understanding that major sources of behavior lie in individual differences intrinsic to the individual. Therefore, it is a logical next step trying to identify specific genes that account for this genetic variance. However, despite marked advances in molecular biology and genetics during the past decade that have included techniques that make it possible to search the entire genome for genetic variants, which may be related to disease states or psychological traits (see also Strobel and Brocke, 2011), our knowledge of specific genes that are responsible for, and provide a sufficient explanation of, specific behavioral outcomes is rather limited (e.g., Johnson et al., 2011) and it seems unlikely "that most of the genes responsible for the heritability for any complex trait will be identified in the foreseeable future" (Haworth and Plomin, 2010: p. 784). The present article will therefore focus on the quantitative-genetic approach and primarily on twin studies constituting the bulk of the pertinent evidence, although we acknowledge adoption designs as alternative or complementary approaches, which are discussed in detail elsewhere (Plomin et al., 2008). We will begin by defining the major concepts of quantitative genetics and then summarize findings in the areas of ability and education with a focus on *developmental behavioral genetics*. This is the study of genetic and environmental influences on individual differences in behavioral development, and it combines a behavior-genetic approach with a developmental approach. A topic for developmental behavioral genetics that can be studied longitudinally as well as cross sectionally is how genetic and environmental influences are related to the age of the samples under study.

Concepts

Quantitative Behavior Genetics and the Classical Twin Method

The development of the twin method is usually ascribed to Galton (1876) although it is uncertain whether Galton was aware of the distinction between monozygotic (MZ) and dizygotic (DZ) twins (Rende et al., 1990). The classical twin study compares the phenotypic resemblances of identical or MZ and fraternal or DZ twins. MZ twins derive from the splitting of one fertilized zygote, and therefore they inherit identical genetic material. DZ twins develop from two separately fertilized eggs and are first-degree relatives: They share on

average 50% of the genes that differ among humans – as do full singleton siblings. It follows that a greater within-pair similarity in MZ compared to DZ twins suggests that genetic variance influences the trait under study. Comparing the phenotypic or observable resemblance of MZ and DZ twins for a trait or measure under study offers a first estimate of the extent to which the phenotypic variation of that trait reflects genetic variance. This is typically measured by using a statistic called heritability.

Heritability, Shared Environmental, and Nonshared Environmental Effects

Heritability is defined as the proportion of phenotypic differences among individuals that can be attributed to genetic differences in a particular population. The *heritability* (h^2) of the trait can be estimated by doubling the difference between MZ and DZ correlations, that is, $h^2 = 2(r_{MZ} - r_{DZ})$ (Falconer, 1960), with r_{MZ} and r_{DZ} as the correlations between MZ and DZ cotwins, respectively.

Shared environmental effects index the extent to which people are similar independently of shared genetic effects, because they share some environmental influence such as parental socioeconomic status. The proportion of the variance that is due to *shared environmental effects* (c^2) can be derived by calculating $c^2 = r_{MZ} - h^2$, because MZ similarity can be conceptualized as h^2 (similarity due to genetic influences) + c^2 (similarity due to shared environmental influences). Substituting $2(r_{MZ} - r_{DZ})$ for h^2 offers another way of calculating shared environmental effects, resulting in the formula $c^2 = 2r_{DZ} - r_{MZ}$. In other words, the presence of shared environmental influences on a certain trait is suggested if DZ twin similarity exceeds half the MZ twin similarity for that trait. A DZ twin correlation, which is less than half the MZ twin correlation, indicates the presence of nonadditive genetic effects being shared 100% by MZ twins but 25% or less by DZ twins. In contrast to additive genetic effects, which occur when the effects of alleles at a particular locus and across loci act independently of each other and thus additively, nonadditive genetic effects occur when the effects of alleles depend on the presence of other alleles either at the same locus (dominance) or at other loci (epistasis). Nonshared environmental effects (e^2) are all nongenetic sources of differences between family members. Contributing factors include having different peers, attending different classes, participating in different leisure activities, and receiving different parental treatment. Their combined effects can be estimated from $e^2 = r_{tt} - r_{MZ}$, where r_{tt} is the reliability of the measure. If $1 - r_{MZ}$ is used to estimate e^2 instead, the resulting nonshared environmental influences are confounded with measurement error.

For example, in studies of more than 10 000 MZ and DZ twin pairs on general cognitive ability (*g*), the average MZ correlation is 0.86, which is near the test–retest reliability of the measures, in contrast to the DZ correlation of 0.60 (Plomin and Spinath, 2004). On the basis of these data, application of the above formulae results in a heritability estimate of $h^2 = 0.52$, shared environmental estimate of $c^2 = 0.34$, and nonshared environmental influence/measurement error estimate of $e^2 = 0.14$. Thus, given that MZ and DZ twin correlations are available, this straightforward set of formulae can be used to derive estimates of genetic and environmental influences on any given trait. Instead of the more widespread Pearson product–moment correlation coefficient, twin similarity is typically calculated by using intraclass correlation ICC1.1 (Shrout and Fleiss, 1979) because the latter eliminates any bias due to nonrandom ordering of twins within pairs.

Requirements and Assumptions of the Twin Method

Interpretation of twin correlations in the described manner relies on a number of assumptions: the absence of assortative mating for the trait in question; the absence of G(enotype)–E(nvironment) correlation (rGE) and interaction (G × E); and the viability of the equal environments assumption (EEA). Each of these assumptions is addressed briefly below.

Assortative mating describes nonrandom mating that results in similarity or dissimilarity between spouses. If spouses are more similar than randomly paired couples and a trait shows genetic influences, their offspring show higher correlations due to increased genetic similarity resulting from the transmission of similar genes from both parents. Assortative mating can result from social stratification for similar traits, and thus from simple propinquity (people tend to select their spouses from those around them socially), and/or from active social selection, as when people of a given religious faith in a religiously mixed society select their spouses specifically from among those of their own faith. Assortative mating tends to be low for some psychological traits (e.g., personality), but can be substantial for others (e.g., intelligence, with average spouse correlations of about 0.40; Jensen, 1998). In twin studies, unrecognized assortative mating can result in underestimates of heritability, because it raises the DZ correlation but does not affect the MZ correlation, thus mimicking shared environmental influences.

Gene–environment correlation (rGE) describes the phenomenon that genetic propensities can be correlated with individual differences in experience. This occurs when either there is genetic control of exposure to environmental circumstances, or there is environmental control of gene frequency, or a third variable like parents' socioeconomic status shows accordant genetic and environmental effects.

Three kinds of gene–environment correlation are typically distinguished (Plomin et al., 2008). *Passive* correlation occurs when parenting practices that foster the development of heritable traits in children are correlated with those traits. The children need play no active role in the gene–environment correlation they experience. But parenting is clearly a reciprocal function: parents respond to what their children do and appear to need. *Evocative* gene–environment correlation expresses the extent to which parental responses (and those of anyone else in the child's environment) are tailored to traits that are genetically influenced in the children. And children make active choices about their activities, whether they involve other people or not. *Active* gene–environmental correlation takes place when people's selections of activities that relate to trait maintenance or development are themselves genetically influenced. The distinctions among these kinds of gene–environment correlation are not clear-cut. For example, parents may respond to children's active expression of interest in specific activities by giving them lessons in those activities, mixing

active and evocative gene–environment correlation. And parents with certain characteristics may be more likely than others to take their children to museums, but some of the children who have these experiences may learn much more from them than others, mixing passive and active gene–environment correlations (see also Spinath and Johnson, 2011). Research indicates that genetic factors often contribute substantially to measures of the environment, especially the family environment (Plomin, 1994). In the classic twin study, however, rGE is assumed to be zero.

Genotype–environment interaction ($G \times E$) is often conceptualized as the genetic control of sensitivity to the environment (Kendler and Eaves, 1986) or as environmental control of differential gene effects. Heritability that is conditional on environmental exposure can indicate the presence of $G \times E$. The classic twin study is based on the assumption that there is no $G \times E$.

The classic twin model is based on the assumed equality of pre- and post-natal environmental influences within the two types of twins. In other words, under the *equal environments assumption* we assume that environmentally caused similarity is roughly the same for both types of twins reared in the same family. Violations of the EEA because MZ twins experience more similar environments than DZ twins inflate estimates of genetic influences. The EEA has been tested in a number of studies. Even though MZ twins appeared to experience more similar environments than DZ twins, these differences did not seem to be responsible for the twins' differential similarity (Borkenau et al., 2002; Bouchard and Propping, 1993).

Advances: Beyond the Classic Twin Study

The comparison of intraclass correlations between MZ and DZ twins can be regarded as a reasonable first step in our understanding of the etiology of particular traits. This approach, however, cannot accommodate the effects of variations in the sizes of MZ and DZ samples, or mean or variance differences between types of twins. Nor can it accommodate the effect of gender on variances and covariances of opposite-sex DZ twins. To model genetic and environmental effects as the contribution of unmeasured (latent) variables to phenotypic differences, structural equation modeling (SEM) is required. Analyzing univariate data from MZ and DZ twins by means of SEM offers numerous advances over the mere use of correlations; these advances include a test of overall statistical fit of the model for the greater MZ twin compared to DZ twin similarity, tests of parsimonious submodels, maximum likelihood confidence intervals for each latent influence included in the model, and use of maximum likelihood estimates to incorporate information on participants with incomplete data. The most important strength of SEM, however, lies in its application to multivariate and multigroup data. During the last two decades, powerful models and programs to estimate these models efficiently have been developed (Neale et al., 1999). Recently, an even more flexible tool-kit was provided to the scientific community. OpenMx is a free and open source software for use with R that allows estimation of a wide variety of advanced multivariate statistical models (Boker et al., 2011; http://openmx.psyc.virginia.edu/). Extended twin designs, that is studies including relatives of twin pairs, and multivariate genetic analyses, that is the simultaneous analysis of correlated

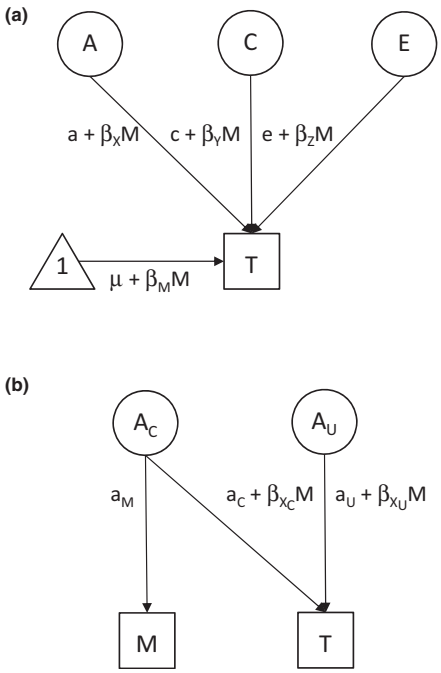

Figure 1 (a) Partial path diagram for one twin demonstrating the incorporation of $G \times E$ in a univariate ACE model by partitioning genetic effects into a mean part, which is independent of the environment, and a part which is a linear function of the environment. (b) Partial path diagram for one twin depicting an extended $G \times E$ model allowing for gene–environment correlation (rGE). The moderator features twice in the model – as a dependent variable to be modeled as well as a moderator variable to define the paths to the trait. The main effect in the means model has been replaced with a path indicating shared genetic effects. The trait is now influenced by two sources of genetic influence: that which is shared with the moderator, and that which is not (common and unique paths a_C and a_U). A = additive genetic effects, C = shared environmental effects, E = nonshared environmental effects, M = environmental moderator variable, T = behavioral trait, μ = intercept. Adapted from Purcell, S., 2002. Variance component models for gene–environment interaction in twin analysis. Twin Research 5, 554–571.

traits, are among the most important developments that go beyond the classic twin designs, yet still use the information inherently available in twins (Boomsma et al., 2002). Among the most exciting advances in quantitative behavior genetics is the possibility of analyzing Gene–Environment correlations (rGE) and Gene × Environment ($G \times E$) interactions within the framework of extended twin designs (e.g., Johnson et al., 2006; Purcell, 2002) and multigroup analyses (Hahn et al., 2012). Figure 1 illustrates how rGE and $G \times E$ can be incorporated in variance components twin-analyses.

Findings

Quantitative-Genetic Findings in the Domains of Ability and Education

Twin, adoption, and family studies have convincingly shown that individual differences not only in intelligence (for an overview, see Deary et al., 2006, 2009; Plomin et al., 2008) but

also in educational success (like school achievement and educational attainment) as well as in noncognitive predictors (e.g., personality) are driven by both genetic and environmental effects (Bouchard, & Loehlin, 2001; Johnson et al., 2008). There is now ample evidence that the heritability of intelligence shows a monotonous increase from childhood to adulthood (for a recent meta-analysis, see Bergen et al., 2007; see also Haworth et al., 2010). In addition, longitudinal behavior genetic studies have found evidence for genetic change (e.g., the onset of genetic effects which were not in effect at the previous measurement occasion) at two important developmental transitions: the transition from infancy to early childhood (marked by rapid language development) and the transition from early to middle childhood (around 7 years of age, marked by the start of formal schooling). A number of studies found genetic influence on facets of educational success like years of school attendance (Behrman et al., 1980; Behrman and Taubman, 1989), scores in specific school achievement tests (see Plomin et al., 2008), and broader measures of school achievement (Bartels et al., 2002; Johnson et al., 2005, 2006, 2007; Sacerdote, 2002). Only a handful of studies has included multiple cognitive and noncognitive predictors (intelligence, motivation, and personality) as well as an assessment of environmental and family influences in a genetically informative design. Johnson et al. (2005, 2006) found that about 70% of the variance of a school achievement measure (based on school grades) could be attributed to genetic influence of which 56% was shared with genetic influences on intelligence. In addition, engagement and motivation, family risk, conscientiousness, and disruptive behaviors also share genetic and environmental influences with intelligence and/or school achievement. Johnson et al. (2007), for instance, separated the relative influences of genetic and environmental effects on the covariance between individual traits (e.g., IQ), and family influences (e.g., parental expectations) on school grades. They found significant genetic influences common to parental variables and IQ as well common to parental variables and school grades, which indicate genetic generational transmission. Genetic and environmental contributions to the association between self-perceived ability (SPA) and school achievement (Greven et al., 2009) were studied recently and analyzed in a longitudinal genetic cross-lagged analysis (Luo et al., 2010). Results revealed that genetic influences contributed moderately to the association between earlier SPA and later school achievement and substantially to the association between earlier school achievement and later SPA.

Environmental Influences and Intellectual Development

Studies with a stronger focus on environmental factors influencing intellectual development have often targeted processes related to the family. Phenotypic research in this area has consistently shown a relationship between deprived family environments (e.g., low-quality home environment, poor nonmaternal care, etc.) and poor cognitive functioning (e.g., Feinstein and Bynner, 2004; NICHD Early Child Care Research Network, 2001; Petrill et al., 2004). More specifically, children from low-socioeconomic status (SES) groups are more likely to experience academic problems (Ellis and Bonin, 2003; Gutman et al., 2003; Jimerson et al., 2000). Beyond SES, researchers have begun to identify characteristics of the home and family environment (e.g., environmental confusion) that account for individual differences in cognitive functioning (van den Oord and Rowe, 1999; Hart et al., 2007).

Gene–Environment Correlation and Interaction

Family risk factors cluster as they do, in part, because they both have common causes and are mutually influential. Thus the roles of processes reflecting gene–environment (G–E) correlation and interaction (G × E) need to be examined (Johnson et al., 2006). For example, poor parental education and mental health or single parenthood may contribute directly to parental poverty, poor parental involvement and support, increasing the risk of low achievement in offspring of poor parents. These risk factors do occur systematically because they reflect characteristics of family members to some degree: low parental intelligence contributes directly to low parental education (Ceci and Williams, 1997), and certain personality traits contribute to poor parental mental health (e.g., Krueger et al., 2000) and to single parenthood (Jockin et al., 1996). The impact of genetically influenced risks can on the other hand be buffered by specific forms of social support (social capital) in the positive sense of social control (Shanahan et al., 2008), whereas restrictive educational policies (Heath et al., 1985) as well as poor social environments (Guo and Stearns, 2002) can severely limit the degree to which capabilities lead to intellectual performance and academic achievement. The study by Shanahan et al. (2008) is particularly noteworthy because it combines a candidate gene approach with the demonstration of a gene–environment correlation and interaction. Drawing on the National Longitudinal Study of Adolescent Health, Shanahan and colleagues studied the effects of social capital and the DRD2 dopamine receptor gene on educational continuation beyond secondary school. For boys, they found that DRD2 risk was associated with a decreased likelihood of school continuation. Parental socioeconomic status, high parental involvement in school, and a high-quality school compensated for this negative relationship, consistent with environmental contingency. However, the DRD2 risk allele was also less commonly observed in settings that were rich in social capital, suggesting that boys who were at genetic risk had a lower probability of benefitting from environmental compensation. Like all studies, this one clearly requires replication, but if it reports a robust phenomenon, it documents an important example of the kinds of gene–environment correlations and interactions that are likely to pervade the phenomena of interest to behavioral scientists, and the fusion of molecular and quantitative genetics that is going to be required to understand them.

Recent quantitative behavior genetic studies have also begun to include G × E interaction into their analyses. For example, the effect of the family environment on the heritability of intelligence was investigated in a couple of G × E interaction studies (Harden et al., 2007; Rowe et al., 1999; Turkheimer et al., 2003). Results suggest the importance of parental education and further socioeconomic indicators as

they modified the relative contribution of genetic and environmental effects on individual differences in intelligence.

Implications for Education

Behavioral genetic research and its findings are often perceived as a threat to the aims of education. This resentment is mostly based on misconceptions about what these findings mean and what genetically sensitive designs can tell us.

The most widespread misconception in this context is that heritability means that a characteristic is predetermined and fixed. To illustrate that this is not true, intelligence makes a good example. Although intelligence is substantially heritable within cohorts, massive gains in average IQ have been observed in industrialized countries since World War II (Flynn, 1987), most probably reflecting secular changes in environmental conditions. Heritability refers to differences between individuals, i.e., relative scores, not mean levels. Therefore, despite its heritability, a characteristic such as intelligence can undergo significant mean-level changes such as documented by Flynn (1987) and by meta-analyses on the effects of intelligence trainings (Klauer and Phye, 2008). In view of the distinction between rank order and mean level, heritability, and malleability no longer present a contradiction.

How can genetically sensitive designs inform education? It is counterintuitive, but behavioral genetic studies provide the best available evidence that environment does matter. This is the case because most studies aiming at demonstrating the importance of environmental factors neglect the fact that similarities between children and parents might in part be explained by shared genes, not only shared environments. Without controlling for the effects of shared genes, an association between parental behavior and children's outcomes cannot be interpreted as evidence of the effects of parental behavior. Nevertheless, the overwhelming majority of educational studies follow such a design.

There are many examples of important findings from studies with genetically sensitive designs that serve the aims of education. The fact that for intelligence heritability has been shown to increase over the life course is in line with the notion that education can achieve its best effects early in life (Bergen et al., 2007). However, it has often been shown that the effects of early interventions on intelligence do not transfer to later developmental stages (e.g., Schweinhart et al., 2005). The concepts of behavioral genetics, such as gene–environment correlation, can help to understand such disappointing findings. The older children and adolescents become, the more they create and choose their environments. According to the concept of active gene–environment correlation, this happens in such a way that the fit between environments and genetically influenced dispositions increases. For example, more intelligent children tend to be more successful in intellectually demanding tasks and thus tend to actively seek such challenges, whereas less intelligent children experience more failure in those tasks and therefore turn to less demanding activities. These insights make it clear that for long lasting changes, intervention programs need to continue beyond early childhood and should ideally take individual differences in a broader range of intervention-relevant variables on the side of the program participants into account.

Behavior genetics research can not only help to understand findings relevant to education but it can also show educational scientists and practitioners where to look for windows of malleability. To learn that shared environmental influences, such as parental educational styles, are probably less important than traditional educational theories claim, whereas nonshared environmental factors, such as peers, seem to be more important, might redirect the attention of researchers and educators interested in effective ways to optimize developmental processes.

Finally, behavioral genetic findings can guide educators in defining the goals of education. Because individual differences in educational prerequisites are substantially heritable, it is unrealistic that all learners will acquire the same level of educational outcomes. The aim of education cannot be to level out individual differences. The aim should rather be to enable each individual to make the best out of their potentials. A good school produces different educational outcomes for its students, but on a high level. Another consequence implied by behavior genetic findings is that the aim of realizing the potential of each student will not be reached by providing equal environments for all. Because of their individuality, learners need individualized environments and differential treatments to fully realize their potentials. That, of course, makes education not less important but more demanding.

See also: Anthropology, Genomics, and Human Variation: National Roots; Behavioral Genetics: Psychological Perspectives; Genetics and Anthropology; Genetics and Society; Genetics and Sociology; Intelligence, Genetics of: Cognitive Abilities; Intelligence: Central Conceptions and Psychometric Models; Intelligence: History of the Concept; School Achievement: Motivational Determinants and Processes; Twin Studies and Epigenetics; Twin Studies in Demography.

Bibliography

Bartels, M., Rietveld, M.J.H., van Baal, G.C.M., Boomsma, D.I., 2002. Genetic and environmental influences on the development of intelligence. Behavior Genetics 32, 237–249.

Behrman, J.R., Hrubec, Z., Taubman, P., Wales, T.E., 1980. Socioeconomic Success: A Study of the Effects of Genetic Endowments, Family Environment, and Schooling. North-Holland, New York.

Behrman, J.R., Taubman, P., 1989. Is schooling "mostly in genes"? Nature-nurture decomposition using data on relatives. The Journal of Political Economy 97, 1425–1446.

Bergen, S.E., Gardner, C.O., Kendler, K.S., 2007. Age-related changes in heritability of behavioral phenotypes over adolescence and adulthood: a meta-analysis. Twin Research and Human Genetics 10, 423–433.

Boker, S., Neale, M.C., Maes, H.H., Wilde, M.J., Spiegel, M., Brick, T.R., Spies, J., Estabrook, R., Kenny, S., Bates, T.C., Mehta, P., Fox, J., 2011. OpenMx: an open source extended structural equation modeling framework. Psychometrika 76, 306–317.

Boomsma, D.I., Busjahn, A., Peltonen, L., 2002. Classical twin studies and beyond. Nature Reviews Genetics 3, 872–882.

Borkenau, P., Riemann, R., Angleitner, A., Spinath, F.M., 2002. Similarity of childhood experiences and personality resemblance in monozygotic and dizygotic twins: a test of the equal environments assumption. Personality and Individual Differences 33, 261–269.

Bouchard, T.J., Loehlin, J.C., 2001. Genes, evolution, and personality. Behavior Genetics 31, 243–273.
Bouchard, T.J., Propping, P., 1993. Twins as a Tool of Behavior Genetics. Wiley, Chichester, UK.
Ceci, S.J., Williams, W.M., 1997. Schooling, intelligence, and income. American Psychologist 52, 1051–1058.
Deary, I.J., Johnson, W., Houlihan, L., 2009. Genetic foundations of human intelligence. Human Genetics 126, 215–232.
Deary, I.J., Spinath, F.M., Bates, T.C., 2006. Genetics and IQ. European Journal of Human Genetics 14, 690–700.
Ellis, L., Bonin, S.L., 2003. Genetics and occupation-related preferences. Evidence from adoptive and non-adoptive families. Personality and Individual Differences 35, 929–937.
Falconer, D.S., 1960. Introduction to Quantitative Genetics. Ronald Press, New York.
Feinstein, L., Bynner, J., 2004. The importance of cognitive development in middle childhood for adulthood socioeconomic status, mental health and problem behavior. Child Development 75, 1329–1339.
Flynn, J.R., 1987. Massive IQ gains in 14 nations: what IQ tests really measure. Psychological Bulletin 101, 171–191.
Galton, F., 1876. The history of twins as a criterion of the relative powers of nature and nurture. Royal Anthropological Institute of Great Britain and Ireland Journal 6, 391–406.
Greven, C.U., Harlaar, N., Kovas, Y., Chamorro-Premuzie, T., Plomin, R., 2009. More than just IQ: school achievement is predicted by self-perceived abilities – but for genetic rather than environmental reasons. Psychological Science 20, 753–762.
Guo, G., Stearns, E., 2002. The social influences on the realization of genetic potential for intellectual development. Social Forces 80, 881–910.
Gutman, L.M., Sameroff, A.J., Cole, R., 2003. Academic growth curve trajectories from 1st grade to 12th grade: effects of multiple social risk factors and preschool child factors. Developmental Psychology 39, 777–790.
Hahn, E., Spinath, F.M., Siedler, T., Wagner, G., Schupp, J., Kandler, C., 2012. The complexity of personality: advantages of a genetically sensitive multi-group design. Behavior Genetics 42, 221–233.
Harden, K.P., Turkheimer, E., Loehlin, J.C., 2007. Genotype by environment interaction in adolescents' cognitive aptitude. Behavior Genetics 37, 273–283.
Hart, S.A., Petrill, S.A., Deater Deckard, K., Thompson, L.A., 2007. SES and CHAOS as environmental mediators of cognitive ability: a longitudinal genetic analysis. Intelligence 35, 233–242.
Haworth, C.M.A., Plomin, R., 2010. Quantitative genetics in the era of molecular genetics: learning abilities and disabilities as an example. Journal of the American Academy of Child and Adolescent Psychiatry 49, 783–793.
Haworth, C.M.A., Wright, M.J., Martin, N.W., Martin, N.G., Boomsma, D.I., Bartels, M., et al., 2010. A twin study of the genetics of high cognitive ability selected from 11000 twin pairs in six studies from four countries. Behavior Genetics 39, 359–370.
Heath, A.C., Berg, K., Eaves, L.J., Solaas, M.H., Corey, L.A., Sundet, J., et al., 1985. Education policy and the heritability of educational attainment. Nature 314, 734–736.
Jensen, A.R., 1998. The G-factor: The Science of Mental Ability. Praeger, London.
Jimerson, S., Egeland, B., Sroufe, L.A., Carlson, B., 2000. A prospective longitudinal study of high school dropouts: examining multiple predictors across development. Journal of School Psychology 38, 525–549.
Jockin, V., McGue, M., Lykken, D.T., 1996. Personality and divorce: a genetic analysis. Journal of Personality and Social Psychology 71, 288–299.
Johnson, A.M., Vernon, P.A., Feiler, A.R., 2008. Behavioral genetic studies of personality: an introduction and review of the results of 50+ years of research. In: Boyle, G.J., Matthews, G., Saklofske, D.H. (Eds.), The Sage Handbook of Personality Theory and Assessment. Sage, Los Angeles, pp. 145–173.
Johnson, W., McGue, M., Iacono, W.G., 2005. Disruptive behavior and school achievement: genetic and environmental relationships in 11-year-olds. Journal of Educational Psychology 97, 391–405.
Johnson, W., McGue, M., Iacono, W.G., 2006. Genetic and environmental influences on academic achievement trajectories during adolescence. Developmental Psychology 42, 514–532.
Johnson, W., McGue, M., Iacono, W.G., 2007. Socioeconomic status and school grades: placing their association in broader context in a sample of biological and adoptive families. Intelligence 35, 526–541.
Johnson, W., Penke, L., Spinath, F.M., 2011. Understanding heritability: what it is and what it is not. European Journal of Personality 25, 287–294.
Kendler, K.S., Eaves, L.J., 1986. Models for the joint effects of genotype and environment on liability to psychiatric illness. American Journal of Psychiatry 143, 279–289.
Klauer, K.J., Phye, G.D., 2008. Inductive reasoning: a training approach. Review of Educational Research 78, 85–123.
Krueger, R.F., Caspi, A., Moffitt, T.E., 2000. Epidemiological personology: the unifying role of personality in population-based research on problem behaviors. Journal of Personality 68, 967–998.
Luo, Y.L.L., Haworth, C.M.A., Plomin, R., 2010. A novel approach to genetic and environmental analysis of cross-lagged associations over time: the cross-lagged relationships between self-perceived abilities and school achievement is mediated by genes as well as the environment. Twin Research and Human Genetics 13, 426–436.
Neale, M.C., Boker, S.M., Xie, G., Maes, H., 1999. Mx: Statistical Modeling, fifth ed. Department of Psychiatry, VCU Box 900126, Richmond. p. 23298.
NICHD Early Child Care Research Network, 2001. Before head start: income and ethnicity, family characteristics, child care experiences, and child development. Early Education and Development 12, 545–576.
Petrill, S.A., Pike, A., Price, T., Plomin, R., 2004. Chaos in the home and socioeconomic status are associated with cognitive development in early childhood: environmental mediators identified in a genetic design. Intelligence 32, 445–460.
Plomin, R., 1994. Genetics and Experience: The Interplay Between Nature and Nurture. Sage Publications, Thousand Oaks, CA.
Plomin, R., DeFries, J.C., McClearn, G.E., McGuffin, P., 2008. Behavioral Genetics, fifth ed. Worth Publishers, New York.
Plomin, R., Spinath, F.M., 2004. Intelligence: genetics, genes, and genomics. Journal of Personality and Social Psychology 86, 112–129.
Purcell, S., 2002. Variance component models for gene–environment interaction in twin analysis. Twin Research 5, 554–571.
Rende, R.D., Plomin, R., Vandenberg, S.G., 1990. Who discovered the twin method? Behavior Genetics 20, 277–285.
Rowe, D.C., Jacobson, K.C., van den Oord, E.J.C.G., 1999. Genetic and environmental influences on vocabulary IQ: parental education level as moderator. Child Development 70, 1151–1162.
Sacerdote, B., 2002. The nature and nurture of economic outcomes. The American Economic Review 92, 344–348.
Schweinhart, L.J., Montie, J., Xiang, Z., Barnett, W.S., Belfield, C.R., Nores, M., 2005. Lifetime Effects: The HighScope Perry Preschool Study Through Age 40 (Monographs of the HighScope Educational Research Foundation, 14). HighScope Press, Ypsilanti, MI.
Shanahan, M.J., Vaisey, S., Erickson, L.D., Smolen, A., 2008. Environmental contingencies and genetic propensities: social capital, educational continuation, and dopamine receptor gene DRD2. American Journal of Sociology 114 (Suppl.), S260–S286.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86, 420–428.
Spinath, F.M., Johnson, W., 2011. Behavior genetics. In: Chamorro-Premuzic, T., von Stumm, S., Furnham, A. (Eds.), Handbook of Individual Differences. Wiley-Blackwell, Chichester, UK, pp. 271–304.
Strobel, A., Brocke, B., 2011. Molecular genetic aspects of personality. In: Chamorro-Premuzic, T., von Stumm, S., Furnham, A. (Eds.), Handbook of Individual Differences. Wiley-Blackwell, Chichester, UK, pp. 309–333.
Turkheimer, E., 2000. Three laws of behavior genetics and what they mean. Current Directions in Psychological Science 9, 160–164.
Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B.M., Gottesman, I.I., 2003. Socioeconomic status modifies heritability of IQ in young children. Psychological Science 14, 623–628.
van den Oord, E.J.C.G., Rowe, D.C., 1999. A cousin study of associations between family demographic characteristics and children's intellectual ability. Intelligence 27, 251–266.

Developmental Research across Cultures and Nations: Challenges, Biases, and Cautions

Bettina Lamm, University of Osnabrück, Osnabrueck, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cultural models inform developmental pathways. Therefore the study of development requires the consideration of the cultural context, which, however, poses methodical challenges, since the research methods are mainly developed in the Western middle-class context. The present article describes those challenges exemplarily considering sampling, conceptions of knowledge and time, as well as situational demands and stimulus material. In the discussion, research strategies that combine naturalistic observations of everyday life and standardized psychological assessments are proposed.

Introduction

It is an interaction between biology and culture that constitutes human development (Keller, 2002, 2007). Culture selects and amplifies biological predispositions and thus affects developmental pathways. Therefore, the understanding of development requires cross-cultural research. However, the vast majority of studies on infant or early child development have been conducted in highly educated, industrialized, urban, Western middle-class communities, representing less than 5% of the world's population (Henrich et al., 2010). Usually, these samples of very privileged children, whose highly educated parents have time, resources, and motivation to carry their children to university laboratories, serve as the basis for claims about universal human behavior (Fernald, 2010).

This ethnocentrism in developmental research influences not only the selection of the research samples but also the research questions and methods (e.g., Bennis and Medin, 2010; Baumard and Sperber, 2010). Although constructs mostly evolved from Western psychology, a precondition of cross-cultural research is that the studied construct is incorporated in the psychological functioning in all included cultural groups. Even though this call for equivalence seems comprehensible, it is difficult to prove. While elaborated methods can statistically test structural or metric equivalence of scales, construct equivalence stays a theoretical assumption (Poortinga and van de Vijver, 2013).

From a methodical perspective, the general challenge of research with infants or young children arises from the fact that the participants are neither able to express their opinion verbally nor to react adequately to verbal requests. Therefore, researchers need to ask the caregivers, observe spontaneous behavior, or create situations and stimulus material, which prompt observable or measurable behavior providing information about the processes in question. As the field of infancy research has steadily grown since the 1950s and especially during the past decades, sophisticated research methods were developed in order to assess infant competencies. However, all research paradigms were developed in Western university laboratories reflecting the cultural model of psychological autonomy and the everyday experiences of infants and young children in highly educated Western middle-class families, therefore posing challenges in cross-cultural research.

The cultural model of psychological autonomy acts on the assumption that human behavior and experience is governed by personal preferences and independent mental states (Keller and Kärtner, 2013). Non-Western, traditional, subsistence-based communities constitute a context that differs tremendously from the typical Western research context. Usually, families in these traditional communities have only a low level of formal school education and limited access to other resources. The prototypical cultural model of this context is the model of hierarchical relatedness, which is based on an emphasis on communal goals, situational demands, and social role obligations (Keller and Kärtner, 2013). It is important to mention that the contrasting of the two cultural prototypes does not imply bipolarity and one-dimensionality of the concepts. These prototypes describe extreme forms that appear mutually exclusive, because the respective cultural model concordantly influences virtually all domains of behavior and experience. Of course, there are alternative combinations of the foci on autonomy and relatedness possible, since both constitute basic human needs. However, the comparison of these extreme, almost pure forms facilitates understanding the cultural influences on developmental processes.

Based on their ecocultural model, Keller and colleagues have argued and empirically documented that psychological autonomy and hierarchical relatedness organize caregivers' socialization strategies and thus infants' learning environments (Keller, 2007; Keller and Kärtner, 2013). The socialization goals of parents in Western middle-class contexts reflect the emphasis on psychological autonomy. Parents want their children to learn to express their individual preferences clearly and to develop personal talents, interests, and uniqueness. The main socialization goals of parents in rural traditional contexts can be described as obedience and respect to parents and elderly people and maintenance of social harmony. Pursuing these goals, parents prefer different interactional styles according to their cultural model (Keller, 2007; Keller et al., 2005a,b). Western middle-class parents set value on face-to-face interactions, object-related play, and elaborated talking about mental states. On the one hand, Western middle-class infants spend most of their time in exclusive dyadic interaction with

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23159-0

their mothers or alone in their separate rooms. Infants in rural traditional families on the other hand, are cared for by multiple persons in their extended family system and are practically never alone. Most of the time, they experience close body contact and rhythmic body stimulations accompanied by rhythmic vocalizations.

In the following, it is discussed how these different cultural models and associated socialization goals and parenting behaviors challenge research routines in cross-cultural developmental research.

Sampling in Different Cultures

Sampling is always a demanding task in developmental research, because it is not only the Western educated middle-class that is overrepresented in psychological research, but within this group, college students are the most studied sample (Henrich et al., 2010). However, if one is interested in other populations, like infants, children, or older people, there will be several questions: How is the best way to get in contact with the study population? How should one approach them and convince them to take part in the study? What kind of reimbursement for study participation would be appropriate? Of course, there are no universal answers to these questions; they rather call for culture-sensitive solutions. Some cultural specifics will be illustrated using the example of infancy research.

To conduct an infant study in a Western middle-class context, one would probably first obtain a clearance from the university's or national ethics committee. Then the appropriate authority should be contacted to get access to the register of births in order to get a list of children of the relevant age. Alternatively, the potential participants could be approached via maternity or children's clinics, independent pediatricians, midwives, prenatal or baby classes, newspaper advertisements, or word of mouth. Once in contact, an individual appointment is arranged, informed consent is established, and the investigation can start. Afterward, the participants usually receive an individual allowance, even though most middle-class families participate in studies because of personal interest in the respective topic or the wish to support research in general.

However, in rural traditional contexts this is a different matter. Although an official ethical clearance would not be amiss, it is far from being sufficient for full acceptance of the project in the community. It is necessary to introduce the research team and the project to the local authorities or traditional rulers. They are the ones to decide what is going on in their villages, to represent the community's interest, i.e., to allow access to potential participants and demanding return service for the community. After reaching agreement with the village head, the actual recruitment process starts. Yet, in cultural contexts where birthdays do not play a major role in a child's life and chronological age is far less important than the social age, which is defined by assumed responsibility and social role obligations, sampling children of a certain age proves to be very difficult. If there are no birth registers at local health centers or clinic cards available, it is almost impossible to age a child reliably. Having found the right-age children it is not the mother (or parents) but rather the family head (usually an older male) who decides about the participation. Accordingly, the mothers or children are not in a position to cancel participation, in case they do not feel comfortable with the study requirements. Only high sensitivity of the researcher can secure to notice unexpressed escapes from the situation that would bias the results dramatically. Written consent as requested by ethical boards is unreasonable for illiterate participants.

Conceptions of Knowledge

Western psychology considers knowing as a mental activity of an individual. However, Greenfield (1997) argues that there is a different view of knowing as group process in many cultural communities. Referring to her experiences when interviewing Zinacantecan Maya girls and their mothers in Chiapas, Mexico, about weaving and embroidery, Greenfield describes the discrepancy between her perspective of individual participants with independent viewpoints and her participants' expectation of the more knowledgeable mothers answering for their daughters and answering being a cooperative effort to produce the most valid information (Greenfield, 1997). This view reflects the epistemological assumption of knowledge as collaborative construction (cf Duranti, 1988) and thus influences data collection procedures in several ways.

The author's experiences when assessing parental ethnotheories about infant care among rural Nso mothers point to the same direction. Many young mothers felt discomfort in the dyadic interview setting. They did not understand why other family members were excluded or at least requested not to intervene in the conversation. Furthermore, the restraint communication that is expected from the objective Western researcher was at odds with their expectation about appropriate communication. Thus, cooperative construction of knowledge was not permitted and diminished information might have been the consequence.

On the other hand, knowledge that is too obvious can also cause problems. Some interviewees felt embarrassed to answer questions about the best maternal care for infants. According to them, even young children know how to care for infants and thus, this is not worth talking about. Besides the discomfort in the interaction, this brings methodical problems when using rating scales. Asking for matters that everybody knows, i.e., shared knowledge within a cultural community, increases the probability of extreme responding and decreases the intracultural variation. These effects are usually seen as critical response styles and thus methodical artifacts. However, an alternative interpretation suggests understanding extreme response behavior and intracultural variation as meaningful aspects of cultural models (of parenting) (Lamm and Keller, 2007).

Affordances of Situations

Shweder (2010) argued that there is no research situation independent of the interpretation of the participant. Differential interpretation of experimental situations or assessment

settings by the participants might cause different behavior, even though the participants do not differ with respect to the psychological concept in question. Thus, it is impossible to interpret study results without knowing the situational interpretation of the participants (Rai and Fiske, 2010). However, infants or young children cannot tell how they experienced the situation or why they reacted in a certain way. Yet, detailed observations of infant behavior and emotional expressions can facilitate valid interpretations of results. This also implies looking at the infants' behavior from a holistic perspective rather than just assessing target behavior and analyzing the behavior of those infants who did not complete a task.

Using the example of working memory assessments with an A-non-B task illustrates the different situational interpretations of 9-month-old German middle-class and rural Nso infants (Lamm, 2013). Although the A-non-B task is an established method (e.g., Diamond, 1985), it did not work with the rural Nso infants. About two-thirds of the 9-month-old rural Nso babies did not search at all for the hidden objects (Lamm, 2013). The first possible explanation could be the absence of object permanence, which is a precondition for searching behavior. Given the scarce object experience of the Nso babies and the parental focus on relational developmental goals (Keller, 2007) this interpretation does not seem absurd. Possibly, the development of object permanence, i.e., an understanding that objects continue to exist even if they cannot be seen, heard, or touched (Piaget and Inhelder, 1987) is delayed compared to Western middle-class children. However, an alternative explanation for the not searching of the rural Nso infants is based on the situation itself. The hide-and-seek game might have been ambiguous for some Nso babies. Unless Western middle-class babies, who are familiar with this kind of peekaboo, the Nso babies occasionally seemed uncertain about the expected reaction. To them it was not always clear, whether they were allowed to fetch the toy that had been taken from them and hidden under the cloth. Even the question, where the object was, was not an explicit invitation to search for the object. These differences in the interpretation of the situation and the expected response between German middle-class and rural Nso infants reflect previous social experiences and cultural values and norms. Obedience and respect for elderly people are highly valued in the Nso socialization. Touching an object that is not mine or given to me marks a kind of disobedience, which caregivers would not authorize and which would be punished even in very young children.

Another example illustrating the effect of situational demands, even though not as much for interpretative reasons than mere familiarity with the situation, constitutes the application of the so-called mobile task in the rural Nso context (Graf et al., 2012). The mobile conjugate reinforcement task (e.g., Rovee-Collier and Cuevas, 2009) assesses contingency learning in infancy. Therefore, the baby is lying under a mobile, one leg of the baby is connected to the mobile, and it is observed whether the infant realizes the contingency between the leg kicking and the visible movement of the mobile. The cross-cultural comparison between three- and 6-month-old German middle-class and rural Nso infants using the mobile task revealed a remarkable difference in the spontaneous kicking activity during baseline phase, when the kicking had no effect on the mobile (Lamm, 2013; Graf et al., 2012). A considerable percentage of Nso infants (more than two-thirds at the age of 6 months) did not kick at all or showed an extremely low kicking activity. These infants had no chance to recognize the association between their movement and the movement of the mobile. Thus, the mobile conjugate reinforcement task is not considered an appropriate procedure to assess contingency learning in a sample of 6-month-old Nso infants. However, even those Nso infants, who showed at least a minimum of kicking behavior, kicked less than German middle-class infants. This difference in baseline activity is a methodical problem as it predicts the achievement of the learning criterion, which is following Rovee-Collier (1997) defined as a kick rate that exceeds the baseline rate by 1.5 times in 2 of 3 consecutive minutes during reinforcement. Across cultural groups, those infants who kicked less were more likely to reach the learning criterion, which implicates that more rural Nso infants were classified as learners (Graf et al., 2012). Yet, the interpretation of this result is anything but clear. It is not known whether the Nso infants learn the contingency more easily, because they kick less, or whether the German infants learn as well, but are not able to increase their high baseline kicking rate sufficiently to reach the learning criterion. Apparently, rural Nso and German middle-class infants start from different preconditions as regards the baseline activity level, which reflects the divergent everyday experiences of infants in these cultural contexts. German middle-class babies are used to the (half) lying position, as under the mobile. At the age of 3 months, they spend more than 50% of the day without body contact to anybody (Keller et al., 2005a,b) and consequently in a lying position. In contrast, 3-month-old Nso infants spend more than 80% of the time of a day in close body contact with another person, i.e., they are on the arm, back, or lap of somebody. If they are not in body contact, they are usually seated in plastic bowls or buckets padded with clothes. They are trained to sit from early on and at 6 months of age, virtually all Nso infants are able to sit without support (Lohaus et al., 2011). Thus, rural Nso infants are hardly ever lying and experience daily routines that are not conducive to kicking or independent leg movements. Hence, the task setting is not equally familiar to the infants from both cultural contexts. Lying and kicking are anything but typical activities for Nso babies. In order to avoid false conclusions on contingency learning, this has to be considered when interpreting the results.

A general characteristic of most assessment situations in infancy research is the dyadic nature. With respect to knowledge construction, the communicational implications of this dyadic isolation have already been discussed in the last section (cf *Conceptions of knowledge*). However, dyadic versus multiple person situations also influence the requested mode of attention. One-to-one situations between researcher and child are characterized by a researcher paying exclusive attention to the child and expecting the child to pay exclusive attention to the task. This constellation is atypical for children in rural traditional contexts. Cultures differ substantially with respect to norms and strategies for distributing caregiver's attention. In Western middle-class contexts, one of the major factors defining high-quality

parenting is the availability of a caregiver's undivided or exclusive attention within a dyadic context. Experts (cf sensitivity concept, Ainsworth et al., 1978) as well as mothers (Keller et al., 2005a,b) concordantly consider exclusive attention to an infant as absolutely essential to optimal child development. Mothers should be completely involved with their infants, see the world from their infants' point of view, and respond even to their infants' most subtle and understated cues, which is not possible without directing full and exclusive attention toward the infant. Despite the popularity of this view in Western middle-class contexts, in most of the world's child-rearing contexts, caregiving is conceptualized as a co-occurring activity (Saraswathi, 1994). Exclusive attention is a resource that caregivers in rural subsistence-based communities cannot afford to provide to large degrees. These caregivers provide their infants with constant physical proximity, while sharing their attention and engaging in extradyadic activities like household or subsistence work at the same time, thus contributing to family sustenance. But the mode of attention is not just a question of workload. Awarding exclusive attention encourages the development of a concept of self as a distinct individual and unique agent, while the shared attentional pattern supports the development of close social relations and the experience of being a coagent in a communicative system (Keller, 2007). Thus, the experience of exclusive attention of an adult in the dyadic research situation is unfamiliar, if not unsettling for children in rural traditional contexts. Many children in those contexts react with extreme shyness to this exceptional situation. Moreover, a cultural community's normative style of attention management is also acquired very early in life (Chavajay and Rogoff, 1999; Rogoff et al., 1993). Guatemalan Mayan toddlers from San Pedro showed remarkable ability to attend to several events simultaneously, for example, working with their caregiver on a task while monitoring another ongoing conversation, whereas Euro-American toddlers from Salt Lake City mostly focused on one event at a time, alternating exclusive attention between different events (Rogoff et al., 1993). Sibling caregivers from a rural Nso community in Northwest Cameroon also preferred a co-occurring care strategy (Lamm, 2008). In interviews, the 4- to 8-year-old Nso children mentioned the importance of accomplishing household tasks (e.g., fetching water or firewood) or the wish to play with their friends while caring for their infant siblings. Hence, the expectation of exclusive attention toward the test is also biased. The exclusion of possible distracting factors from the research situation reflects the Western middle-class idea of best possible results with exclusive focus. However, there is no empirical evidence that this presupposition holds true for children in rural traditional contexts. There are reasonable doubts based on the observation that rural Nso infants showed higher imitation rates when tested outside the laboratory room on a green area in front of the local health center as compared to the controlled laboratory setting (Teiser, 2010). Yet, it is too early for a clear conclusion and there is more empirical work on the effects of research settings needed, since the inside and outside testing differed in several other aspects (e.g., sitting position).

Affordances of Materials

Besides translation, adaptations of test materials and stimuli are the most obvious challenges in cross-cultural developmental research. There are several reports on the influence of familiarity with test materials on test results. Serpell (1979) observed considerable cultural differences in pattern reproduction as a function of test material: whereas Zambian children showed much better reproduction skills in the medium of wire modeling, British children performed better when reproducing the patterns by drawing. Yet, there were no group differences concerning the performances in the clay modeling task. Comparable experiences were reported from the Zambian Early Childhood Development Project (Fink et al., 2012) applying the Pattern Reasoning subscale of the Kaufman Assessment Battery for Children (Kaufman and Kaufman, 2004). During piloting it was noticed that Zambian preschoolers struggled to recognize the patterns. Based on the assumption that minor experience with things drawn on paper had complicated the tasks for them, an object-based version of the reasoning test was developed. The same items displayed through objects rather than printed on paper were much easier for the Zambian preschoolers. Thus, they were able to see the patterns, if they were presented in three-dimensional format and scored higher in the Tactile Pattern Reasoning scale in comparison to the Kaufman Pattern Reasoning scale (Fink et al., 2012). A third example for the effect of test material on test results comes from a study with Zinacantecan Maya children (Greenfield and Childs, 1977). These participants without or with very little school experience did not succeed in continuing weaving patterns correctly when they were drawn on paper. However, they were able to continue weaving patterns constructed with colored sticks in a wooden frame.

With respect to stimuli the necessity for adaptations to cultural contexts has been generally accepted. Malda et al. (2008) proposed a systematization of adaptation processes related to cognitive tests. They distinguished five types of adaptations that they consider sufficient to make cognitive instruments adequate to new cultural contexts: construct-driven adaptations reflect different definitions of psychological concepts across cultures; language-driven adaptations are necessary in the case of structural differences between languages or unavailability of semantically equivalent words across languages; culture-driven adaptations arise from cultural differences with respect to norms, values, customs, or practices; theory-driven adaptations describe modifications that are needed for theoretical reasons; and familiarity/recognizability-driven adaptations are caused by diverging familiarity with item characteristics or stimulus materials (Malda et al., 2008).

However, even visual stimuli that have been considered universally relevant provoke culturally different reactions. Pictures of human faces constitute preferred visual stimuli for infants, which was interpreted as evolutionary adaption (Keller, 2000). They caused the best and earliest memory performances, especially if the person was highly socially relevant and familiar (Bushnell, 2001). Therefore, they are highly qualified for studying processes of learning and memory in infancy. Nevertheless, the use of face stimuli from

the respective cultural context in a study on visual habituation revealed that rural Nso and German middle-class infants perceive different affordances in pictures of human faces. The Nso infants seemed much more animated to initiate an interaction with the faces than the German infants (Lamm, 2013). They smiled, vocalized, or tried to touch the face in order to get in contact. Additionally, Nso infants looked longer at the faces before achieving the habituation criterion. Yet, this difference in the total habituation time is not necessarily reflective of different information processing speed. In consequence of unsuccessful attempts to initiate social interaction infants reacted with frustration and crying causing dropouts from the study. Reasons for the different interpretations of the stimuli might be manifold. The rural Nso infants are not as used to the two-dimensional world of pictures as German middle-class infants. Therefore, they might have misinterpreted the pictures as real persons. Furthermore, Nso babies experience face-to-face positions and direct eye contact in mother–infant interaction less frequently than German middle-class babies (Keller, 2007). However, the photographs used as stimuli showed young women directly looking into the camera. Thus, the quasieye contact of the stimuli faces paired with the not reacting to social approaches might have confused, scared, or distracted the Nso infants from effective information processing.

Conceptions of Time

The pace of life differs tremendously across cultural contexts. What travelers experience and report frequently is reflected in prejudices (e.g., German punctuality or Mediterranean life-style), but has also been assessed systematically. Based on observations of three indicators (walking speed, work speed, and accuracy of public clocks) in 31 countries around the world, Levine and Norenzayan (1999) found that the pace of life was significantly faster in colder climates, economically productive countries, and in individualistic cultures. Thus, the pace of life is intertwined with different aspects of the ecocultural context as well as cultural norms and values. Accordingly, these pace differences influence and complicate cross-cultural psychological assessments.

To begin with, differences in pace of life result in organizational difficulties. It is challenging for a research team as well as in contact with study participants if one person is used to follow a tight schedule and fixed appointments and the other one does not even have a clock. Many Westerners name it among their first lessons from Africa that a great deal of the time is spent waiting. However, cultural openness, respectfulness, and reflection of those differences can resolve initial communicational difficulties. In the author's research team in the Northwestern grassfields of Cameroon, for instance, it has become a running gag to ask for each appointment whether it is set in white men's time or black men's time.

More serious than these anecdotes of intercultural communication is the question of time limits within test items or psychological assessments. Considering the described differences in pace of life or local observations that children are not trained to manage their time or perform quickly, some researchers relax original time limits of tests when using them in rural, non-Western contexts (e.g., Malda et al., 2008). Yet, this practice is problematic. It appears absolutely reasonable at first glance, but there is no conversion factor between different tracks of time. It is an arbitrary decision of the researcher to consider different time limits in different contexts equivalent. Not enough is known about cultural influences on information processing speed. The developing field of cultural neuroscience demonstrates that cultural experiences modify neurocognitive connectivity (Kitayama and Uskul, 2011). Thus, cultural differences in information processing speed might be reflective of experienced differences in the pace of everyday life. However, observed differences in the working speed might also be biases resulting from different familiarity with materials and settings.

Conclusion

In this article, it was argued that the only way to understand developmental processes implies the consideration of the ecocultural context. Systematic assessments and observations in divergent contexts are essential to gain insight into the interplay between biology and culture that constitutes development (Keller, 2002). However, cross-cultural developmental research turns out to be a great methodical challenge. Mostly, research constructs and assessment methods have been developed in Western university laboratories and are adapted to the highly educated, urban, Western, middle-class context. This bias becomes obvious in research questions, recruitment procedures, as well as the design of research materials and situations. There is no culture-free stimulus constellation and accordingly no culture-free test. Social conventions and cultural routines interfere with expected response behavior and familiarity with the materials and test situation influence achievement in a test. Hence, the development of methods and test procedures that are adequate in different cultural contexts is very demanding. The intended goal that a test reveals equivalent scores for members of different populations given the same skill (culture-fair test) has been proven untenable (e.g., Rosselli and Ardila, 2003). Hence, the efforts aim at reducing the cultural bias as far as possible (culture-reduced tests) and alternatively controlling the bias statistically (Poortinga, 1995). Throughout this article several examples of challenges in cross-cultural developmental research have been discussed. Even though the answers to these problems must always be culture-specific, general suggestions for culture-sensitive methods are derived from these examples. The basic prerequisite for cross-cultural research is the reflection of the own cultural values and norms as well as openness to experiences and a multicultural team. Methodically, psychologists can seek inspiration from ethnography. Field observations, detailed interviews, and participating in cultural routines give necessary information in order to adapt a psychological assessment instrument to a new cultural environment or develop a new instrument for the use in different cultural groups. This approach will help to overcome the ethnocentric bias in developmental research and increase the knowledge about cultural pathways of development.

See also: Cross-Cultural Research Methods in Psychology; Equivalence and Transfer Problems in Cross-Cultural Research.

Bibliography

Ainsworth, M.D.S., Blehar, M.C., Waters, E., Wall, S., 1978. Patterns of Attachment: A Psychological Study of the Strange Situation. Erlbaum, Hillsdale, NJ.

Baumard, N., Sperber, D., 2010. Weird people, yes, but also weird experiments. Behavioral and Brain Sciences 33, 84–85.

Bennis, W.M., Medin, D.L., 2010. Weirdness is in the eye of the beholder. Behavioral and Brain Sciences 33, 85–86.

Bushnell, I.W.R., 2001. Mother's face recognition in newborn infants: learning and memory. Infant and Child Development 10, 67–74.

Chavajay, P., Rogoff, B., 1999. Cultural variation in management of attention by children and their caregivers. Developmental Psychology 35, 1079–1090.

Diamond, A., 1985. Development of the ability to use recall to guide action, as indicated by infants' performance on AB. Child Development 56, 868–883.

Duranti, A., 1988. Intentions, language, and social action in a Samoan context. Journal of Pragmatics 12, 13–33.

Fernald, A., 2010. Getting beyond the "convenience sample" in research on early cognitive development. Behavioral and Brain Sciences 33, 91–92.

Fink, G., Matafwali, B., Moucheraud, C., Zuilkowski, S.S., 2012. The Zambian Early Childhood Development Project – 2010 Assessment Final Report. Retrieved from: http://developingchild.harvard.edu/activities/global_initiative/zambian_project/ (23.05.12.).

Graf, F., Lamm, B., Goertz, C., Kolling, T., Freitag, C., Spangler, S., Fassbender, I., Teubert, M., Vierhaus, M., Keller, H., Lohaus, A., Schwarzer, G., Knopf, M., 2012. Infant contingency learning in different cultural contexts. Infant and Child Development 21, 458–473. http://dx.doi.org/10.1002/icd.1755.

Greenfield, P.M., 1997. You can't take it with you: why ability assessments don't cross cultures. American Psychologist 52, 1115–1124.

Greenfield, P.M., Childs, C.P., 1977. Weaving, color terms, and pattern representation: cultural influences and cognitive development among the Zinacantecos of Southern Mexico. Interamerican Journal of Psychology 11, 23–48.

Henrich, J., Heine, S.J., Norenzayan, A., 2010. The weirdest people in the world? Behavioral and Brain Sciences 33, 61–135.

Kaufman, A.S., Kaufman, N.L., 2004. Kaufman Assessment Battery for Children. Manual, second ed. AGS Publishing, Circle Pines, MN.

Keller, H., 2000. Human parent-child relationships from an evolutionary perspective. Special Issue of the Journal "American Behavioral Scientist" with the topic "Evolutionary Psychology: Potential and limits of a Darwinian Framework for the Behavioral Sciences" 43, 957–969.

Keller, H., 2002. Development as the interface between biology and culture: a conceptualization of early ontogenetic experiences. In: Keller, H., Poortinga, Y., Schölmerich, A. (Eds.), Between Culture and Biology. Cambridge University Press, Cambridge, pp. 215–240.

Keller, H., 2007. Cultures of Infancy. Erlbaum, Hillsdale, NJ.

Keller, H., Abels, M., Lamm, B., Yovsi, R.D., Völker, S., Lakhani, A., 2005a. Ecocultural effects on early infant care: a study in Cameroon, India, and Germany. Ethos 33 (4), 512–541.

Keller, H., Völker, S., Yovsi, R.D., 2005b. Conceptions of parenting in different cultural communities: the case of West African Nso and Northern German women. Social Development 14, 158–180.

Keller, H., Kärtner, J., 2013. Development – the cultural solution of universal developmental tasks. In: Gelfand, M., Chiu, C., Hong, Y. (Eds.), Advances in Culture and Psychology, vol. 3. Oxford University Press, New York, pp. 63–116.

Kitayama, S., Uskul, A.K., 2011. Culture, mind, and the brain: current evidence and future directions. Annual Review of Psychology 62, 419–449.

Lamm, B., 2008. Children's ideas about infant care: a comparison of rural Nso children from Cameroon and German middle class children (Doctoral dissertation). University of Osnabrück http://d-nb.info/989904156/34.

Lamm, B., 2013. Challenges to finding similarities and differences in cross-cultural infancy research. In: Tchombe, T., Nsamenang, B.A., Fülop, M., Keller, H. (Eds.), Cross-cultural Psychology Research: An Africa-centric Perspective. Design House, Limbe, Cameroon, pp. 248–261.

Lamm, B., Keller, H., 2007. Understanding cultural models of parenting. The role of intra-cultural variation and response style. Journal of Cross-cultural Psychology 38 (1), 50–57.

Levine, R.V., Norenzayan, A., 1999. The pace of life in 31 countries. Journal of Cross-cultural Psychology 30 (2), 178–205.

Lohaus, A., Keller, H., Lamm, B., Teubert, M., Fassbender, I., Freitag, C., Goertz, C., Graf, F., Kolling, T., Spangler, S., Vierhaus, M., Knopf, M., Schwarzer, G., 2011. Infant development in two cultural contexts: Cameroonian Nso farmer and German middle class infants. Journal of Reproductive and Infant Psychology 29 (2), 148–161.

Malda, M., Van de Vijver, F.J.R., Srinivasan, K., Transler, C., Sukumar, P., Rao, K., 2008. Adapting a cognitive test for a different culture: an illustration of qualitative procedures. Psychology Science Quarterly 50, 451–468.

Piaget, J., Inhelder, R., 1987. The construction of reality. In: Oates, J., Sheldon, S. (Eds.), Cognitive Development in Infancy. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 165–169.

Poortinga, Y.H., 1995. Cultural bias in assessment: historical and thematic Issues. European Journal of Psychological Assessment 11, 140–146.

Poortinga, Y.H., van de Vijver, F.J.R., 2013. Der Umgang mit methodischen Stolperfallen in der kulturvergleichenden Stressforschung [Dealing with methodical pitfalls in cross-cultural stress research]. In: Genkova, P., Ringeisen, T., Leong, F.T.L. (Eds.), Handbuch Stress und Kultur [Handbook Stress and Culture]. Springer VS, Wiesbaden, pp. 153–171.

Rai, T.S., Fiske, A., 2010. ODD (observation- and description-deprived psychological research). Behavioral and Brain Sciences 33, 106–107.

Rogoff, B., Mistry, J., Göncü, A., Mosier, C., 1993. Guided participation in cultural activity by toddlers and caregivers. Monographs of the Society for Research in Child Development 58 (8) v-179.

Rovee-Collier, C., 1997. Dissociations in infant memory: rethinking the development of implicit and explicit memory. Psychological Review 104, 467–498.

Rosselli, M., Ardila, A., 2003. The impact of culture and education on non-verbal neuropsychological measurements: a critical review. Brain and Cognition 52, 326–333.

Rovee-Collier, C., Cuevas, K., 2009. The development of infant memory. In: Courage, M.L., Cowan, N. (Eds.), The Development of Memory in Infancy and Childhood. Psychology Press, Hove, pp. 11–41.

Saraswathi, T.S., 1994. Women in poverty context: balancing economic and child care needs. In: Borooah, R., Cloud, K., Seshadri, S., Saraswathi, T.S., Peterson, J.T., Verma, A. (Eds.), Capturing Complexity: An Interdisciplinary Look at Women, Households and Development. Sage, New Delhi, India, pp. 162–178.

Serpell, R., 1979. How specific are perceptual skills? A cross-cultural study of pattern reproduction. British Journal of Psychology 70, 365–380.

Shweder, R.A., 2010. Donald Campbell's doubt: cultural difference or failure of communication? Behavioral and Brain Sciences 33, 109–110.

Teiser, J., Imitationsverhalten kamerunischer Säuglinge: Beeinflusst das Alter des Modells die Nachahmung? [Imitational behavior of Cameroonian infants: does the age of the model influence the imitation rate?]. University of Osnabrück, Unpublished Diploma thesis.

Dewey, John (1859–1952)

Helmut Schreier, University of Hamburg, Hamburg Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 3624–3628, © 2001, Elsevier Ltd.

Abstract

Dewey's work provides the foundation for a framework for social thought. Participatory democracy aims at 'the great community,' liberating the capabilities of each individual to promote the common good and thus improve on the quality of experience through application of intelligence to practice. Intelligence is the grasp of possibilities in a situation at hand. Democracy is contingent upon education, and education at its best is modeled for democracy ('embryonic society'). Philosophy, once it turns away from the futile quest for certainty, is seeking principles to solve the problems of men, and may be understood as education in a generic state. Dewey sought to expose dualisms as results of faulty thinking. Dewey's studies of various fields and disciplines are all contributions to the foundation of a framework for social life that facilitates the liberation of human potential. Against the separation of experience and dualistic views, he promotes the unifying moments in experience and the resolution of dualism; against concepts of certain truth and the eternal, he emphasizes the contingent and the present in the situation at hand; instead of the closed-down (e.g., the 'law of the laws of history') and shut-off, he points to the continuous and the transactional; against reifications of language and reason, he demonstrates their instrumentality; and instead of leaning to romantic pessimism, he puts a premium on cautious optimism in matters concerning human beings and the universe.

Biography

John Dewey was born in Burlington, Vermont, on 20 October 1859. By the time of his death (1 June 1952 in New York City), he was regarded by many as America's foremost philosopher. During his boyhood and youth in Vermont, Dewey had been an average pupil. In 1864, Dewey's mother moved with her four boys to Northern Virginia to be close to her husband, a quartermaster with the Union army during the civil war. They did not return to Vermont before 1867, and John as a young boy saw the carnage that war brings.

As a student at the University of Vermont in Burlington, he found a mentor, H. A. P. Torrey, who nourished his interest in philosophy. After graduation in 1879, Dewey spent 2 years teaching at the high school at South Oil City in Pennsylvania, and for one year taught at an elementary school in a small town in Vermont. He continued studying philosophy in private tutorial, and in 1882, enrolled at Johns Hopkins University in Baltimore. There, he gravitated quickly to G. S. Morris, a neo-Hegelian. Hegel's ideas had a liberating effect on Dewey's thought. The religious teachings that he was brought up with were transformed into the Hegelian view of cultural institutions as agents of an objective mind that through them shapes the mental life of individuals. Morris let Dewey teach his undergraduate class, secured a fellowship for him, and offered him a position as instructor at the University of Michigan at Ann Arbor where he had moved in 1884, the year of Dewey's PhD exam. During the following decade, Dewey stayed at the University of Michigan (except for 1888, the year spent at the University of Minnesota), where he was promoted to full professor, and wrote articles on psychology and ethics. In 1886, he married Alice Chipman, a strong personality with a secular humanist background. A formative influence – next to Morris' and to that of William James that mainly came by correspondence – was his friendship with G. H. Mead.

The exchange began in Ann Arbor and continued during the decade in Chicago, lasting until Mead's death in 1931. Mead's scholarship in physiology surpassed that of Dewey, and Mead's concept of psychology – based on a view of the organism that is interacting with its environment – was integrated by Dewey into his own transactional theory.

He accepted the offer to head the department of Philosophy, Psychology, and Pedagogy at the University of Chicago in 1894. The city of Chicago, teeming with European immigrants, full of destitute people as well as of people of means and goodwill, appeared to Dewey like one big social experiment. He worked closely with Jane Addams of Hull House, a social settlement; with Ella Flagg Young, District Superintendent of Chicago Schools; and with Francis Parker, principal of the Cook County Teachers' Training School. In 1896, at the request of parents who were willing to pay for most of the school's maintenance, he founded the Laboratory School that quickly became known as the Dewey School. The school's curriculum centered on 'occupations' that engaged pupils' minds and hands in activities such as sewing, weaving, cooking, building, growing corn, etc. A debate society complemented the syllabus for fifth- and sixth-graders. What made the school unique was not so much the emphasis upon the priority of the manual to the conceptual as the attempt to realize the idea of school as a community, in which both teachers and students were stakeholders.

Dewey's book *School and Society* (1899) – probably his most widely read book, with translations into a dozen languages – consists of talks he gave to raise money for the Laboratory School. In 1897, he had published a brief text, *My Pedagogic Creed*, that not only reflects his idea of the Laboratory School but holds much of his later more explicit writing on education in a nutshell.

Dewey's time in Chicago came to an abrupt end in 1904 when the University's president merged the Laboratory School with the Chicago Institute for Teacher Training, dismissing Dewey's school's teachers and his wife, who acted as principal, without prior notice or consultation with Dewey. He resigned immediately and found the position as professor of philosophy

at New York's Columbia University that he would hold from 1904 through his retirement in 1930, until his final resignation of all duties in 1939, at age 80.

Courses he had taught in Chicago were published in New York (*How we Think*, 1910; *Human Nature and Conduct*, 1922), and his major work on philosophy of education, *Democracy and Education* (1916) is a fruit of Dewey's involvement with schools and philosophy of education in Chicago.

In New York, Dewey was engaged with philosophy at Columbia and education at Teachers' College, and quickly he became involved in various political activities, not only as president of the American Association of University Professors and functions in the American Federation of Teachers, but also as leading contributor to *The Republic*, a magazine promoting liberal views, and as president of the People's Lobby and Chairman of the League for Independent Political Action. His travels and consultant activities in Japan (1919), China (1919–21), Turkey (1924), and Mexico (1926) extended his influence, and a series of lectures on a variety of topics resulted in major publications (Tokyo University 1919: *Reconstruction in Philosophy*; Carus lectures at Columbia 1923: *Experience and Nature;* Gifford lectures in Edinburgh 1929: *The Quest for Certainty*; Terry lectures at Yale 1934: *A Common Faith*).

John and Alice Dewey had six children, three boys, one of whom had died in 1895 on a sojourn in Italy, and three girls. In 1904, on their second trip to Europe, another boy died, and even though the Deweys adopted later, on that same tour, an Italian orphan who was eight years old like their lost son, they never came to accept these deaths.

Alice died after long illness in 1927; John Dewey was married again in 1946, at age 87, to Roberta Grant, and the couple adopted two children. In 1949, when he was 90 years old, Dewey published a book that he had written together with Arthur Bentley, *Knowing and the Known*, a study in epistemology.

Experience the Connecting Agent of Dewey's Life and Work

One of the more unusual features for a philosopher's biography is the close connection between Dewey's intellectual and actual life.

In politics, he tried his hand at founding a third party, at editing a newspaper devoted to 'thought news,' and at heading the Outlawry of War Movement ('That war be outlawed and declared a crime'). In his 50s and 60s, he traveled widely not only to spread his ideas, but also to help install systems of education in Turkey, Mexico, the Soviet Union; his sojourn in China lasted over two years. At age 78, he chaired the committee to investigate Leon Trotsky (an investigation at Trotsky's own request, in Mexico, into accusations by Stalin's courts in Moscow).

Looking back at 80, he wrote: "I reached fairly early in the growth of my ideas a belief in the intimate and indissoluble connection of means used and ends reached. I doubt if the force of the idea in the theory of social action would have come home to me without my experience in social and political movements, culminating in events associated with my membership in the Trotsky Inquiry Commission" (Jane Dewey, 1939).

As for education, having been a schoolteacher himself and having run the Laboratory School in Chicago, he points to the interaction between thoughts and actions: "The idea that lay back of my educational undertaking was a rather abstract one of the relation of knowledge and action. My school work translated this into a much more vital form" (Jane Dewey, 1939).

In the field of aesthetics and the fine arts, he had worked with Albert C. Barnes, the millionaire collector who befriended him and introduced him to the art of French postimpressionism; Dewey dedicated his book *Art as Experience* (1934) to him.

As for metaphysics, his concept of unity of the ideal and the actual that he named 'God' (*A Common Faith*) appears to have been backed by personal experience. During his sojourn as a young man in South Oil City, he had experienced a mystic oneness with the universe, and he would later declare: "I've never had any doubts since then, nor any beliefs. To me faith means not worrying ..." (Dykhuizen, 1973).

His first years as a young instructor at the University of Michigan in Ann Arbor had impressed on him the actual attainability of a cooperative community and the experience of personal growth within a democratic atmosphere, with encouragement of freedom and individual responsibility, and the immediate acceptance as an adult responsible member of the faculty. This experience may well have contributed to the formulation of his later visionary concept, presented in *The Public and Its Problems* (1927), of democracy as 'the great community.'

He himself summed up the mutual enforcement of idea and action throughout his experience in the autobiographical notes that he added to his daughter Jane's account of his life: "My belief in the office of intelligence as a continuously reconstructive agency is at least a faithful report of my own life and experience" (Jane Dewey, 1939).

Contribution

The upshot of Dewey's work can be described as the subordination of philosophy to life, a contribution to the project of intellectualization of practice. Since intelligence – the assessment of the possibilities inherent in a given situation – is a tool for the solution of practical problems, Dewey referred to his brand of pragmatism as 'instrumentalism': like the wings of a bird and the fins of a fish, human intelligence is an instrument for survival, but unlike the other creatures' physical features, the human mind is capable through the powers of community to promote life's quality itself, not unlike a process of evolution directed by intelligent action.

One of the most obvious principles in Dewey's philosophy is his exposure of dualisms as artificial alternatives that have become barriers to the kind of intelligence needed to promote the process of experience, e.g., the separation between science and morals appeared scandalous to him, a roadblock on the path to progressive application of intelligence. The method of scientific inquiry developed and employed successfully in modern times provides a model

for inquiry into social affairs to the degree that it is an application of reason against that of myth; analogously, intelligence must be applied in the social realm to develop a methodology for the resolution of the problems of men and to facilitate social reconstruction.

His dictum, that philosophers should turn away from the problems of philosophy and instead should turn to the problems of men, is another way of saying that traditional philosophy is concerned with questions that result from views framed in terms of alternatives which do not relate. As he states in an essay, *The Influence of Darwinism on Philosophy* (1909): "But in fact intellectual progress usually occurs through sheer abandonment of questions together with both of the alternatives they assume – an abandonment that results from their decreasing vitality and a change of urgent interest. We do not solve them: we get over them. Old questions are solved by disappearing, evaporating, while new questions corresponding to the new attitude of endeavor and preference take their place" (Middle Works 1973–83).

With regard to politics, Dewey maintained that the majority of men and women could and should become active participants in the decision-making processes that govern and shape their lives. He urged the development of a society that facilitates individual growth through participation in community life. Integration of self-realization and the social to form a common good is the major ethical postulate entwined with such an instrumentalization of the political. In *The Public and Its Problems* (1927), Dewey answered an analysis of the political situation in the USA in the years after World War II by Walter Lippmann, *The Phantom Public* (1925). Lippmann, a democratic realist, had shown that self-government and the idea of participatory democracy fell short of the reality of a population largely manipulated by propaganda, and he proposed the establishment of an elite ruling class to identify and serve the public interest. Dewey, while accepting the factuality of Lippman's analysis, did not subscribe to his proposal, but suggested instead the continued pursuit of the idea of a great community, especially through the means of education. His idea of the fullest possible liberation of the potentials of human experience allowed for no other social framework than that of participatory democracy. And for Dewey, the school as the 'church of democracy' is to provide the actual experience for achieving the 'great community' through its model as 'embryonic society.'

Probably, Dewey is known best for his contributions to educational thought. At the core of his philosophy of education is the concept of experience as an ongoing process with no externally identifiable aim. " ... education is a constant reorganizing or reconstructing of experience. It has all the time an immediate end, and so far as activity is educative, it reaches that end – the direct transformation of the quality of experience" (*Democracy and Education* 1916, Middle Works 1973–83). This concept was opposed to teleological concepts of education dominant in countries with an idealistically inclined tradition such as Germany or Japan, and at the same time provided the foundation for an interchange of participatory democracy and education as its exercise in embryonic form.

In *Experience and Education* (1938), Dewey emphasizes the need for a theory of education. Such theory is not a system or an application of static principles. It is entirely subservient to practice and should, in Dewey's words, be 'a light to the eyes and a lamp for the feet'; it might be described as a framework to facilitate a plan of action, or as a set of principles to interpret an experience's educational function.

Dewey elaborates two such principles of experience, that of continuity and that of interaction. He reminds educators to avoid the opposite pitfalls of overindulging children and of subjecting them entirely to external demands. Proponents of progressive education who promoted a child-centered approach had claimed Dewey as provider of the philosophical foundation for their sentiment, and W. H. Kilpatrick, an influential representative of that school, had declared the project method as Dewey's brainchild. But Dewey, while sympathetic to the idea of a participatory approach to education, would not want to see the curriculum content in terms of knowledge and competence buried underneath project-learning for its own sake. Continuity of experience, as he explained in 1938, requested the educator to transfer 'the vested capital of civilization': "The mature person, to put it in moral terms, has no right to withhold from the young on given occasions whatever capacity for sympathetic understanding his own experience has given him" (*Experience and Education*, Later Works 1981–91).

In *The Quest for Certainty* (1929), Dewey gives an account of cultural history that exposes the tradition of bad metaphysics in philosophy which seeks to establish a realm of certainty protected from the precariousness of human experience in a world of unavoidable uncertainty. While early religion attempted an alignment with the powers of destiny, philosophy – its inheritor – sought access to knowledge of the immutable, to the certainty provided by everlasting values and principles. Both religion and philosophy distinguished between skills and technical knowledge that eased the circumstances of everyday existence, and the superior knowledge connected to the higher order of absolute certainty in terms of truth, goodness, and beauty. The sciences that developed in the seventeenth century effectively replaced the qualitative view of nature with one describing it in terms of mathematical relationships. But instead of using that knowledge as an instrument in the service of liberating experience, scientists engaged in a quest for certainty on their own by redescribing reality in scientific terms, as if 'the findings of science are a disclosure of the inherent properties of the ultimate real' (*The Quest for Certainty*, Later Works 1981–91). Thereby, another dualism was generated: that between scientific procedure with the resulting quantitative description of relationships in nature, and that of the appreciation of qualities and values in experience. Dewey demonstrated how these dualisms dissolve as one turns away from the fixation on certainty in the name of reason and turns instead to intelligence as a means of assessing the possibilities inherent to a situation.

With *The Quest for Certainty*, Dewey aimed at a destruction of the history of philosophy in order to gain plausibility for democracy as the social arrangement with the greatest potentials for the full realization of society's capacities.

In *Art as Experience* (1934), Dewey stresses the continuity of all experience. The aesthetic is that part of experience that is identical with the form that separates an experience from the

ongoing flow of ordinary experience. It is of a satisfying quality for its consistent movement toward an intended end, which makes it self-sufficient and lifts it out of the stream of the ordinary, at the same time. An agreeable meal, an intelligent conversation, a talk professionally delivered, a well-conducted concert, a well-played game, etc. all qualify as an experience. As for art, the artist clarifies and intensifies an object's traits in such a way that a new experience results. He does so in a process that requires his attention to the medium he is working with, and the beholder of the finished work appreciates it only to the degree that he goes through the same processes that the artist went through when making the object. By describing art this way, by identifying the aesthetic as part of experience accessible to all, Dewey claims that the removal of the arts from the ordinary lives of the common people, its relegation to museums and to private collections of the wealthy has to be understood as symptomatic of the dualism that separates the aesthetic from the quotidian and is at the roots of the divisions of society.

Like the aesthetic, the religious offers an aspect of consummatory experience. Dewey had hinted in the final chapter of *Reconstruction in Philosophy* (1920) at a 'natural piety' that transcends the denominational and liberates the potentials of religious belief. In 1934, he delivered a series of lectures to extend and clarify his position, which were published under the title *A Common Faith*. He draws a distinction between the religious as disposition and religions as sets of rituals and dogmatic beliefs; he shows the realm of the religious to be the kind of belief that is not in competition with the kind of knowledge achieved through application of research and reason; organized religion, however, has interfered with the progressive expansion of knowledge. Dewey sketches the contour of a natural piety that takes humans to be part of nature and seeks its cooperation not without a just and dignified attitude. Experience of the religious culminates in the unification of the ideal and the actual, i.e., the realization of our most inclusive moral convictions in a given situation, through which an all-pervasive unity may be glimpsed momentarily that aligns human purpose and nature.

Dewey's studies of various fields are all contributions to the foundation of a framework for social life that facilitates the liberation of human potential. Against the separation of experience and dualistic views, he promotes the unifying moments in experience and the resolution of dualism; against concepts of certain truth and the eternal, he emphasizes the contingent and the present in the situation at hand; instead of the closed-down (e.g., Marxist 'law of the laws of history') and shut-off, he points out the continuous and the transactional; against reifications of language and reason, he demonstrates their instrumentality; and instead of leaning to romantic pessimism, he puts a premium on cautious optimism in matters concerning human beings and the universe.

Relevance for Contemporary Social Sciences

A renewed interest in Dewey is indicated by the number of doctoral dissertations, which doubled during the 1990s from an average seven at US universities in the early years to an average 15 for each of the later years. At the same time, the tone of studies on Dewey appears to have changed. An observation that might be symptomatic: Westbrook (1991) follows Dewey for the most part, but is still critical of his emphasis on education to reach the goal of participatory democracy; Eldridge (1998) attacks Rorty for misreading Dewey on scientific method, Niebuhr for misconstruing Dewey's metaphysics, and Westbrook for prioritizing political action over cultural development.

The explanation of Russel's and Horkheimer's reproach of Dewey's pragmatism as curious misunderstanding has become a common feature in the literature of the 1990s; the fierce Marxist–Leninist critique of Dewey's philosophy during the Soviet era is no longer an issue, other than a footnote, and the neoThomist repudiation of Dewey's secular philosophy is no longer even mentioned.

At the same time, several tendencies in the development of philosophy have helped facilitate a renewed interest in Dewey's philosophy. The rejection of 'great tales' in French postmodernist philosophy, the promotion of pluralistic views by proponents of constructivism, has exposed the untenability of philosophy's claim of a relationship with the immutable, for their part. One of the most influential thinkers of the 1980s and 1990s, Richard Rorty, has combined these sentiments and applied them in his reformulation of the contributions of philosophy. His *Philosophy and the Mirror of Nature* (1979) can be read as mirroring Dewey's *Quest for Certainty*. He claims to be 'Deweyan' and comments explicitly on Dewey's writings. While this claim has been disputed by Westbrook (1991) and Eldridge (1998), Rorty's influence on contemporary thought has contributed to a renewed interest in Dewey's work.

There is renewed interest in Dewey's theory of democracy. Westbrook, in his seminal study (1991), has made explicit the relevancy and actuality of the Lippmann–Dewey debate for American democracy at the end of the twentieth century, and the time of that study's publication and its wide reception coincided with a similar appeal for the systematic pursuit of community by a group of philosophers (e.g., M. Dancer) and sociologists (e.g., R. N. Bellah) referred to as 'communitarians.' Perhaps it is indicative of a renewed influence of Dewey's concept of 'great community' that J. Habermas closed a review of the publication in German, after 70 years, of *Quest for Certainty*, expressing the wish: "He (Dewey) would be the better patron for the Berlin Republic."

See also: Darwinism; Democratic Theory; Educational Philosophy; James, William (1842–1910); Mead, George Herbert (1863–1931); Pragmatism: Philosophical Aspects; Pragmatist Social Thought, History of.

Bibliography

Boydston, J.A. (Ed.), 1973. The Early Works of John Dewey, 1882–1898, vol. 5. Southern Illinois University Press, Carb-ondale, IL.

Boydston, J.A. (Ed.), 1973-1983. The Middle Works of John Dewey, 1899–1924, vol. 15. Southern Illinois University Press, Carbondale, IL.

Boydston, J.A. (Ed.), 1981-1991. The Later Works of John Dewey, 1925–1953, vol. 17. Southern Illinois University Press, Carbondale, IL.

Boydston, J.A. (Ed.), 1974. Guide to the Works of John Dewey. Southern Illinois University Press, Carbondale, IL.

Dewey, J., 1939. Biography of John Dewey. In: Schilpp, P.A. (Ed.), The Philosophy of John Dewey. Tudor, New York.

Dykhuizen, G., 1973. The Life and Mind of John Dewey. Southern Illinois University Press, Carbondale, IL.

Eldridge, M., 1998. Transforming Experience. John Dewey's Cultural Instrumentalism. Vanderbilt University Press, Nashville, TN.

Hickman, L.A., 1990. John Dewey's Pragmatic Technology. Indiana University Press, Bloomington, IN.

Rorty, R., 1979. Philosophy and the Mirror of Nature. Princeton University Press, Princeton, NJ.

Thomas, M.H., 1962/1972. John Dewey: A Centennial Bibliography. Chicago University Press, Chicago.

Thomas, M.H., Writings of John Dewey to October 1939. In: Schilpp, P.A. (Ed.), The Philosophy of John Dewey. Tudor, New York.

Westbrook, R.B., 1991. John Dewey and American Democracy. Cornell University Press, Ithaca, NY.

Diabetes: Psychosocial Aspects

Daniela Zahn and Thomas Kubiak, Johannes Gutenberg University Mainz, Mainz, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by K. Chamberlain, volume 6, pp. 3628–3631, © 2001, Elsevier Ltd.

Abstract

Diabetes mellitus refers to a group of chronic diseases characterized by a disorder of carbohydrate metabolism. The most common forms are type 1 diabetes mellitus, type 2 diabetes mellitus, and gestational diabetes. Diabetes requires lifelong intense self-management behavior on a daily basis, and prognosis is heavily dependent on the individual's ability to successfully perform diabetes self-management. Thus, behavioral and psychosocial aspects play an important role in diabetes. Understanding how psychosocial risk factors and protective factors affect self-management and subsequently prognosis in individuals with diabetes constitutes a major concern in improving diabetes care and maintain the patients' quality of life.

Diabetes Mellitus

Diabetes mellitus is a group of metabolic diseases characterized by chronic hyperglycemia (high blood glucose levels) due to an abnormality in insulin production, its utilization, or both (Davidson et al., 2004). Currently, more than 371 million people worldwide are suffering from diabetes (International Diabetes Federation, 2012). With about 90–95% of all cases, the most common form is type 2 diabetes mellitus (T2DM), which results from a progressive insulin deficiency caused by insulin resistance. In addition to genetic factors, lifestyle factors such as obesity, unhealthy diet, and physical inactivity play an important role in the etiology of insulin resistance in T2DM. Type 1 diabetes (T1DM) comprises about 5–10% of all cases and results from a destruction of β-cells in the pancreas leading to an absolute insulin deficiency. The destruction of β-cells is often caused by progressive autoimmune processes. Gestational diabetes occurs in about 1–14% of all pregnancies and – in most cases – ends with giving birth. However, women with gestational diabetes are at higher life-time-risk of developing T2DM.

There are also other rare specific types of diabetes caused by genetic defects or drug- or chemical-induced forms.

T2DM is mostly diagnosed at 45 years of age or later. With respect to age range, prevalence rates increase continuously from 50 years, with the highest prevalence of T2DM in individuals aged 70–79 years (about 20%). An increase for T2DM in obese children and adolescents has been observed in the last years. T1DM can occur at any age but incidence rates are elevated in children and adolescents aged 10–14 years. Incidence rates decrease after puberty and stabilize between ages 15 and 29 years. About one quarter of all T1DM cases are diagnosed during adulthood (see Maahs et al., 2010, for an overview).

Prognosis

Diabetes may lead to short- and long-term diabetes complications: Short-term complications include diabetic ketoacidosis (diabetic coma) and hypoglycemia. Diabetic ketoacidosis refers to a state characterized by hyperglycemia and elevated levels of ketones in the bloodstream resulting from a breakdown of fatty acids. Ketoacidosis can lead to unconsciousness, diabetic coma, and is fatal if untreated. Hypoglycemia refers to low blood glucose levels, mostly caused by too high insulin dosage, inaccuracies in estimating one's carbohydrate uptake, or increased physical activity. Low blood glucose levels are associated with symptoms such as trembling, sweating, blurred vision, or difficulties to concentrate. If blood glucose levels are very low and brain glucose metabolism is compromised, unconsciousness, coma, and seizure may occur (severe hypoglycemia).

Long-term complications are caused by chronic hyperglycemia, leading to damage at small and large blood vessels (microangiopathy and macroangiopathy) as well as damage to nerves (neuropathy). Frequent clinical syndromes resulting from microangiopathy are diabetic nephropathy and diabetic retinopathy, which can lead to renal failure and loss of vision. Macroangiopathic complications include atherosclerotic processes associated with increased risk for coronary artery disease, periphery artery disease, and cerebrovascular diseases. Diabetic neuropathy can affect peripheral nerves, leading to symptoms such as pain, paresthesia, or loss of sensation in the affected regions and constitutes a risk factor for diabetic foot-ulcers, which is the major cause for amputations in diabetes. Diabetic neuropathy of the autonomic nervous system can impair the function of various organs such as the cardiovascular system, gastrointestinal tract, or genitourinary tract. For instance, a common clinical outcome of autonomic neuropathy in diabetes is erectile dysfunction. Diabetes is associated with increased mortality and reduced lifeexpectancy: In 2011 about 8% of all deaths in people aged 20–79 years were attributable to diabetes, and half of the people dying from diabetes were younger than 60 years of age (International Diabetes Federation, 2012). Good glycemic control, i.e., glucose levels as close to normal (nondiabetic) levels as safely possible, and avoiding hypoglycemia and hyperglycemia, is necessary to minimize the risk for acute as well as long-term complications and to improve prognosis. As glucose binds to hemoglobin in the red blood cells, glycosylated hemoglobin (HbA1C) can be used as a marker of glycemic control. Due to the average lifetime of a red cell of about 120 days, HbA1C reflects average blood glucose levels of the last 2–3 months.

Self-Management in Diabetes

Good glycemic control and, hence, prognosis in diabetes are highly dependent on the individual's skills to successfully perform diabetes self-management behavior in everyday life. Diabetes self-management is complex and differs between types of diabetes. In patients with T1DM, insulin treatment is mandatory from the onset due to the destruction of the insulin producing β-cells. Patients have to measure their blood glucose levels multiple times per day, estimate carbohydrate content of meals, calculate insulin dosages, and inject insulin. This requires knowledge about blood glucose metabolism and its external and internal influences, self-management skills as well as motivation and self-efficacy beliefs with respect to the treatment.

In patients with T2DM, the first-line treatment approach consists of lifestyle modification including healthy diet, physical activity, and weight reduction to improve insulin sensitivity. Oral antidiabetics are recommended in addition to lifestyle-modification or if lifestyle modification alone has not been sufficient. During the course of diabetes, insulin treatment is also necessary in most patients with T2DM. If target blood glucose levels are not reached, insulin treatment should be considered as soon as possible to avoid long-term complications. Lifestyle modification and adherence to oral antidiabetics also require sufficient knowledge, self-efficacy, and motivation from the person affected.

Thus, diabetes self-management is a daily burden that requires life long adherence and is influenced by behavioral and psychosocial aspects.

Psychosocial Aspects in Diabetes

With respect to psychosocial aspects, diabetes-specific problems such as self-management related problems, impaired hypoglycemia awareness (IHA), diabetes-specific fears, and mental disorders such as depression, anxiety, and eating disorders can be distinguished.

In the following passages diabetes-specific problems will be outlined first; in the second part prevalence, etiology, and impact of mental disorders in diabetes patients will be summarized.

Questionnaires that might be helpful in screening for diabetes-specific problems are displayed in **Table 1**. Mental disorders should be assessed using established diagnostic instruments based on criteria of the *International Classification of Diseases* (*ICD-10*) or the *Diagnostic and Statistical Manual of Mental Disorders* (DSM V), respectively.

Table 1 Questionnaires to assess diabetes-specific psychosocial problems

Problem area	*Questionnaire*
Treatment satisfaction	Diabetes Treatment Satisfaction Questionnaire (DTSQ; Bradley, 1994)
Diabetes-related emotional distress	Problem Areas In Diabetes survey (PAID; Polonsky et al., 1995)
Psychological insulin resistance	Barriers to Insulin Treatment questionnaire (BIT; Petrak et al., 2007)
Hypoglycemia awareness	Structured questionnaire to assess hypoglycemia awareness (Clarke et al., 1995)
Fear of hypoglycemia	Hypoglycemia Fear Survey (HFS; Cox et al., 1987; Gonder-Frederick et al., 2011)

Diabetes-Specific Psychosocial Aspects

Self-Management

Similar to other chronic diseases, quality of life may be reduced in patients with diabetes. Coping with the fears of acute and long-term complications and integrating self-management in daily life are important issues in diabetes. Results from the cross-national Diabetes Attitudes, Wishes and Needs (DAWN) study (Peyrot et al., 2005) showed that psychological well-being was poor in almost half of the study participants and diabetes-related worries such as the fear of deterioration of diabetes or worries about hypoglycemia were common, albeit a broad range between 25 and 50% across the different countries was observed. Only about 39–46% of all patients reported complete success in adherence in at least two thirds of their self-care domains. Self-reported adherence to diet and exercise was significantly lower than adherence to medication or appointment keeping. Most providers reported that they were aware of these psychosocial problems, but they indicated difficulties with respect to treatment such as a lack of confidence in their skills to manage psychosocial problems. These results highlight the need to identify patients with diabetes-related distress and to develop interventional tools for providers. Thereby, providers might able to help their patients in overcoming these problems.

Directly related to treatment and adherence in T2DM is psychological insulin resistance (PIR). PIR is a cross-national phenomenon, and refers to the reluctance to start insulin treatment due to various attitudes, beliefs and expectation with regard to insulin and diabetes. Studies examining PIR in insulin-naive T2DM patients indicate that feelings of personal failure, low self-efficacy, expected hardship from insulin therapy, negative outcome expectancy with respect to insulin, worries about increased risk of hypoglycemia, or fear of injections among others may contribute to PIR. PIR can lead to a delay in necessary insulin treatment with the subsequent risk of poor glycemic control. With respect to the DAWN-study population, more than 50% of insulin-naive T2DM patients reported worries about starting insulin treatment. Interestingly, this reluctance is observed in both patients and providers, with providers reporting similar concerns as their patients. Expected problems with adherence to insulin treatment and extra-time required for training insulin-self-management skills may also be among the reasons for a delay of insulin therapy (see Peyrot et al., 2010, for an overview). But as most of the studies were cross-sectional, it remains unclear, whether patients' and providers' psychological insulin resistance does actually delay insulin initiation and worsen glycemic control via non-adherence to treatment. However, these results indicate the need to address potential barriers when discussing insulin treatment with the patient.

Hypoglycemia Awareness

Hypoglycemia awareness refers to the ability to perceive the onset of hypoglycemia. In case of IHA, it becomes difficult to

detect the warning signs of hypoglycemia, and timely treatment of low blood glucose is often compromised. As a consequence, the risk for severe hypoglycemia and hypoglycemia-induced complications increases. IHA is associated with an up to six times higher risk for severe hypoglycemia compared to those with good/normal hypoglycemia awareness. IHA also constitutes a psychosocial burden for affected patients as well as for their relatives, because the fear to experience a hypoglycemia and not being able to detect hypoglycemia warning symptoms in time may lead to a feeling of helplessness and loss of control. IHA affects about 19–25% of T1DM patients and is less frequent in insulin-treated T2DM patients with incidence rates about 8–10%. Additional risk factors for IHA in insulin-treated patients are age, diabetes duration, strict glycemic control, and low diabetes knowledge (see Graveling and Fisher, 2007, for an overview).

The development of IHA has physiological and behavioral components: If hypoglycemic episodes are experienced frequently, a physiological adaptation to lower blood glucose levels occurs. As a consequence, the hormonal counterregulation that causes the typical first symptoms of hypoglycemia is altered, i.e., the threshold for the counterregulation processes shifts toward lower blood glucose. Thus, warning symptoms occur at lower blood glucose levels, and the possibility of a timely treatment of decreasing blood glucose is reduced. As a consequence, moderate and severe hypoglycemic episodes are more likely, leading to further shift of the hypoglycemia symptoms threshold. A vicious circle can develop. On the behavioral side, the individual decision to treat decreasing blood glucose plays an important role. If the patient decides not to treat hypoglycemic levels timely, hypoglycemic episodes are more likely and symptom threshold shifts further. Possible reasons not to treat hypoglycemia include underestimation of the individual risk of severe hypoglycemia, poor interoception, fear of diabetes complications, or situational factors complicating a timely treatment. Treatment approaches to IHA comprise the rigorous avoidance of hypoglycemic episodes to reverse the adaptational processes that led to lower thresholds for hypoglycemic symptoms, educational programs consisting of symptom perception training such as the blood glucose awareness training (BGAT, Cox et al., 2006), or more empowerment-based approaches with a focus on personal goal-setting, self-management, and coping skills to prevent and treat hypoglycemia (HyPOS, Hermanns et al., 2010).

Studies examining hypoglycemia mostly conclude that the experience of frequent episodes of severe hypoglycemia often occurs in a high risk subgroup of patients, which may also be at increased risk for IHA. Identifying these patients and the physiological, behavioral, and psychosocial influences that contribute to heightened risk in these patients should be of special importance.

Diabetes-Specific Fears

The most common forms of diabetes-specific fears are fear of hypoglycemia and fear of diabetes complications.

Fear of hypoglycemia refers to a distinct fear to experience hypoglycemia that is associated with avoiding behaviors in order to minimize the possibility of hypoglycemia. Patients often develop fear of hypoglycemia after experiencing a severe hypoglycemic episode, which can be a very aversive experience with strong feelings of helplessness and lack of control. Additionally, individuals might feel embarrassed when they experienced hypoglycemia in public, and outside help from stranger was required. Typical avoidance behaviors include aiming at higher blood glucose levels than recommended, excessively monitoring blood glucose to avoid an undetected decrease in blood glucose, omitting or reducing insulin doses, eating more carbohydrates, and avoiding physical activity because it could decrease blood glucose level. Some patients even try not to leave their home due to the fear to experience hypoglycemia in public. Thus, fear of hypoglycemia can be a major burden for people affected. Consequences with respect to glycemic control can also be severe: The above-mentioned avoidance behaviors can lead to higher average blood glucose levels, and risk of ketoacidosis and late complications are increased. The prevalence of fear of hypoglycemia is difficult to determine. Due to the fact that more T1DM patients than T2DM patients are treated with insulin, hypoglycemic episodes and in consequence fear of hypoglycemia are more common in T1DM.

Fear of diabetes complications refers to the constant and excessive fear to develop late complications such as loss of vision or diabetic foot syndrome in the course of diabetes. Patients suffering from this fear often have unrealistic or pessimistic beliefs about the association of elevated blood glucose levels and risk of diabetes complications or overestimate their individual risk of diabetes complications. Subsequently, avoiding behaviors can develop such as excessively monitoring of blood glucose, aiming rigidly at meeting strict target levels for blood glucose, or setting themselves lower target levels than recommended. Psychosocial burden in patients affected by fear of diabetes complications is high and consequences for glycemic control can also be severe. Trying to meet target levels in a rigid way can increase the risk for hypoglycemic episodes even in patients without fear of diabetes complication. However, as patients suffering from this fear may be less likely to treat mild hypoglycemic episodes, risk for severe hypoglycemia, and hypoglycemia-induced complications such as unconsciousness or seizure is increased. Furthermore, frequent episodes of hypoglycemia can also lead to an elevated risk of developing an IHA.

Treatment approaches depend on the severity of the experienced fears: With respect to severe forms of fear and distinct avoiding behaviors, established psychotherapeutic approaches for anxiety disorders, best conducted by a psychologist with knowledge about diabetes or in close collaboration with the diabetologist, are indicated. In case of milder forms, discussing the fears with the diabetologist might be sufficient.

Mental Disorders in Diabetes

Depression

There is a large body of evidence that prevalence of depression in diabetes patients is about twice as high as in individuals without diabetes. Prevalence rates vary depending on criteria (diagnosis of depression vs depressive symptoms) and assessment of depression (questionnaires vs structured diagnostic interviews). Depressive disorders as well as subclinical levels of depression are associated with poorer adherence to treatment, worse glycemic control, elevated risk for diabetes complications, cardiovascular disease, and increased all-cause mortality

as well as increased cardiovascular mortality compared with diabetes patients without depressive symptoms. Depression risk is increased in females with diabetes, directly after experiencing severe hypoglycemia, and at first occurrence of long-term complications. An increasing number of diabetes complications is also associated with increased depression risk. However, as most studies examined individuals with T2DM, the results are clearer for T2DM patients than T1DM patients.

With respect to T2DM, recent studies indicate that the relationship between diabetes and depression may be bidirectional – with an increased risk for depression in patients diagnosed with T2DM and vice versa an increased risk for T2DM in individuals with depression. The pathways relating both diseases are still unclear and range from behavioral factors such as poor health behaviors and reduced adherence to pathophysiological explanations such as increased inflammation or alterations in hypothalamic–pituitary–adrenal axis function. A recent meta-analysis on the role of hyperglycemia for depression risk in diabetes concluded that prevalence of depression was only elevated in individuals diagnosed with T2DM, but not in patients with undiagnosed diabetes or impaired glucose tolerance (Nouwen et al., 2011). These results indicate that the psychosocial burden of knowing to have a chronic disease like diabetes is not to be neglected with respect to the development of depression.

Established approaches in treating depressive disorders should be used in patients with diabetes. However, with respect to antidepressive medication, selective serotonin reuptake inhibitors should be recommended because of fewer metabolic side effects. Interestingly, current research indicates that psychotherapy and antidepressant medication both improve depression outcomes in diabetes, whereas diabetes outcomes such as glycemic controls could not be improved in most studies.

To conclude, although it is almost common knowledge that depression is frequent in diabetes and adversely affects prognosis, specific pathways linking both diseases remain still unclear. In addition to studies with regard to the currently discussed psychophysiological pathways, further research examining alterations in cerebral activation and neuronal connectivity using imaging techniques might contribute to a better understanding of this relationship.

Anxiety

With respect to prevalence of anxiety disorders in diabetes there is less research than for depression and diabetes. Results from the World Mental Health Surveys (Lin et al., 2008) are the first to indicate that anxiety disorders are about 20% more frequent in patients with diabetes relative to those without diabetes. Differences in prevalence of specific anxiety disorders could not be proven. Anxiety disorders can also have diabetes-specific content such as fear of hypoglycemia or fear of diabetes-complications. Although these diabetes-specific fears are mostly subclinical, in some cases criteria of anxiety disorders may be met. For example, fear of hypoglycemia can present as a specific phobia, or fear of diabetes complications can be part of a generalized anxiety disorder. In these cases, patients may benefit most from a combination of psychotherapeutic interventions established in the treatment of anxiety disorders by a psychotherapist with background medical knowledge about diabetes mellitus and its self-management. With respect to the consequences of anxiety disorders in diabetes not much is known – but it has to be assumed that anxiety disorders might adversely affect prognosis in diabetes.

Further studies are necessary to explore the relationship of anxiety and diabetes.

Eating Disorders and Disordered Eating Behaviors

With respect to prevalence of eating disorders and disordered eating behavior (DEB) in diabetes, research results are heterogeneous. Whereas some studies indicate elevated rates of eating disorders and DEB in diabetes patients, other studies do not find any differences (see Young-Hyman and Davis, 2010, for an overview). In addition, methodological issues hamper the interpretation of these results: Many studies did not include weight-matched controls or focused on women. The use of existing measures to assess DEB may not be adequate for patients with diabetes: Self-management skills such as carbohydrate counting and attitudes toward food and food intake that may reflect functional behaviors and cognitions in diabetes may be wrongly diagnosed as dysfunctional by these instruments.

However, various studies indicate that DEB such as binge eating, restrained eating, self-induced vomiting are common in diabetes. Insulin purging, a diabetes-specific form of DEB, is of particular importance and occurs more frequently in T1DM. Insulin purging refers to the omission of insulin to avoid weight gain or to compensate for eating binges. Omission of necessary insulin leads to high blood glucose levels, and in consequence, body fat is metabolized. In addition to weight loss, this increases the risk of diabetic ketoacidosis, and long-term complications. Cross-sectional studies indicate that HbA1C levels tend to be higher in diabetes patients with subclinical or clinical eating disorders, and diabetes complications such as retinopathy or neuropathy as well as hospitalizations and diabetic ketoacidosis are more common in these patients. Due to the negative consequences of DEB, it is important to consider the possibility of DEB in patients with diabetes. Possible indicators of DEB include underweight, large fluctuations in HbA1C and/or body weight, or decompensation of glycemic control. With respect to treatment of eating disorders, established treatment approaches similar to those without diabetes are recommended. For cases with insulin purging, a close collaboration between psychotherapists and diabetes care specialists is indicated. Further studies should aim at addressing the methodological issues mentioned above and at developing measures that can discriminate between diabetes-specific functional occupation with eating and dysfunctional forms of eating behavior that result from unrealistic beliefs about body image or thinness.

Conclusion and Further Research

As a chronic disease requiring lifelong self-management, behavioral and psychosocial aspects play an important role in diabetes. Diabetes-specific problems such as the psychosocial burden of self-management, fears of acute or long-term diabetes complications, or IHA can compromise diabetes

management and glycemic control. In addition, mental disorders such as depression or anxiety are more prevalent in diabetes compared with healthy controls. Even subclinical forms of mental disorders such as disordered eating behaviors may adversely affect glycemic control. To ensure successful diabetes-management addressing psychosocial issues is a key element in the treatment of diabetes. More research is necessary to identify patients at risk for psychosocial problems and to develop treatment approaches that can improve psychosocial as well as medical outcomes.

Due to the increase of T2DM in children and the high prevalence of diabetes in the elderly, new challenges for psychosocial research in diabetes emerge: As the role of behavioral and psychosocial factors for prognosis in these age groups is not well understood yet, future research addressing these aspects could help improving diabetes care and maintain these patients' quality of life.

See also: Chronic Illness, Psychosocial Coping with; Doctor–Patient Interaction in the West: Psychosocial Aspects; Health Education and Health Promotion; Health-Related Support and Self-help Mutual Aid Groups; Patient Adherence to Health Care Regimens.

Bibliography

Bradley, C., 1994. The diabetes treatment satisfaction questionnaire: DTSQ. In: Bradley, C. (Ed.), Handbook of Psychology and Diabetes: A Guide to Psychological Measurement in Diabetes Research and Practice. Harwood Academic Publishers, Chur, pp. 111–132.

Clarke, W.L., Cox, D.J., Gonder-Frederick, L.A., Julian, D., Schlundt, D., Polonsky, W., 1995. Reduced awareness of hypoglycemia in adults with IDDM. A prospective study of hypoglycemic frequency and associated symptoms. Diabetes Care 18, 517–522.

Cox, D.J., Irvine, A., Gonder-Frederick, L., Nowacek, G., Butterfield, J., 1987. Fear of hypoglycemia: quantification, validation, and utilization. Diabetes Care 10, 617–621.

Cox, D.J., Gonder-Frederick, L., Ritterband, L., Patel, K., Schächinger, H., Fehm-Wolfsdorf, G., Hermanns, N., 2006. Blood glucose awareness training: what is it, where is it and where is it going? Diabetes Spectrum 19.

Davidson, M., Harmel, A.P., Mathur, R., 2004. Davidson's Diabetes Mellitus, fifth ed. Saunders, Philadelphia.

Gonder-Frederick, L.A., Schmidt, K.M., Vajda, K.A., Greear, M.L., Singh, H., Shepard, J.A., Cox, D.J., 2011. Psychometric properties of the hypoglycemia fear survey-II for adults with type 1 diabetes. Diabetes Care 34, 801–806.

Graveling, A.J., Fisher, M. (Eds.), 2007. Hypoglycaemia in Clinical Diabetes, second ed. John Wiley & Sons, Chichester.

Hermanns, N., Kulzer, B., Krichbaum, M., Kubiak, T., Haak, T., 2010. Long-term effect of an education program (HyPOS) on the incidence of severe hypoglycemia in patients with type 1 diabetes. Diabetes Care 33, e36.

International Diabetes Federation, 2012. IDF Diabetes Atlas, fifth ed. Update 2012, International Diabetes Federation, Brussels. http://www.idf.org/diabetesatlas

Lin, E.H., Korff, M.V., Alonso, J., Angermeyer, M.C., Anthony, J., Bromet, E., Bruffaerts, R., Gasquet, I., de Girolamo, G., Gureje, O., Haro, J.M., Karam, E., Lara, C., Lee, S., Levinson, D., Ormel, J.H., Posada-Villa, J., Scott, K., Watanabe, M., Williams, D., 2008. Mental disorders among persons with diabetes – results from the world mental health surveys. Journal of Psychosomatic Research 65, 571–580.

Maahs, D.M., West, N.A., Lawrence, J.M., Mayer-Davis, E.J., 2010. Epidemiology of type 1 diabetes. Endocrinology and Metabolism Clinics of North America 39, 481–497.

Nouwen, A., Nefs, G., Caramlau, I., Connock, M., Winkley, K., Lloyd, C.E., Peyrot, M., Pouwer, F., 2011. Prevalence of depression in individuals with impaired glucose metabolism or undiagnosed diabetes: a systematic review and meta-analysis of the European Depression in Diabetes (EDID) research consortium. Diabetes Care 34, 752–762.

Petrak, F., Stridde, E., Leverkus, F., Crispin, A.A., Forst, T., Pfützner, A., 2007. Development and validation of a new measure to evaluate psychological resistance to insulin treatment. Diabetes Care 30, 2199–2204.

Peyrot, M., Rubin, R.R., Lauritzen, T., Snoek, F.J., Matthews, D.R., Skovlund, S.E., 2005. Psychosocial problems and barriers to improved diabetes management: results of the cross-national diabetes attitudes, wishes and needs (DAWN) study. Diabetic Medicine 22, 1379–1385.

Peyrot, M., Rubin, R.R., Khunti, K., 2010. Addressing barriers to initiation of insulin in patients with type 2 diabetes. Primary Care Diabetes 4 (Suppl. 1), S11–S18.

Polonsky, W., Anderson, B., Lohrer, P., Welch, G., Jacobson, A., Aponte, J., Schwartz, C., 1995. Assessment of diabetes-related distress. Diabetes Care 18, 754–760.

Young-Hyman, D.L., Davis, C.L., 2010. Disordered eating behavior in individuals with diabetes: importance of context, evaluation, and classification. Diabetes Care 33, 683–689.

Relevant Websites

www.diabetes.org/ – American Diabetes Association (ADA).

www.diabetespsychology.nl – Diabetes Psychology Research (DPR) group (The Netherlands).

www.dawnstudy.com – DAWN-Study – Diabetes attitudes wishes and needs.

http://uvtapp.uvt.nl/fsw/spits.ws.frmShowpage?v_page_id=7840016642395120 – The European Depression in Diabetes (EDID) Research Consortium.

www.deutsche-diabetes-gesellschaft.de/ – Evidence-based guidelines of the German Diabetes Association "Psychosocial issues and diabetes."

www.idf.org/ – International Diabetes Federation (IDF).

www.psad-easd.eu/ – PsychoSocial Aspects in Diabetes (PSAD) Study Group.

Diagnostic Processes in Clinical Psychology and Psychotherapy

Amy Wenzel, American Board of Professional Psychology (ABPP), University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A person is assigned a diagnosis of a mental health disorder if he or she meets the minimum number of criteria for that diagnosis according to an established classification system like the Diagnostic and Statistical Manual or the International Classification of Diseases. Diagnoses are determined through the process of psychological assessment, which, at minimum, includes a clinical interview, but can also include data from self-report inventories, behavioral observation, assessment of intellectual functioning, thought sampling, and projective testing. Although many established assessment tools have adequate reliability and validity, the diagnostic process also relies upon the clinical judgment of the clinician, which at times can be subject to bias. Current issues and trends in diagnostic processes include the development and evaluation of (1) clinical interviews that correspond to recently revised diagnostic criteria, (2) user-friendly screening tools, (3) computerized assessments, and (4) alternatives to categorical diagnostic systems.

Definitions

According to the Diagnostic and Statistical Manual of Mental Disorders – fifth Edition (DSM-5; APA, 2013), "a *mental health disorder* [italics added] is a syndrome characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behavior that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning" (p. 20). In order to be assigned a *diagnosis* of a mental health disorder, a person must meet an established number of criteria that are indicative of a particular diagnosis. Mental health professionals, such as clinical psychologists and psychiatrists, are trained to conduct assessments of the presence and severity of psychiatric symptoms, as well as the degree to which psychiatric symptoms are associated with life interference and subjective distress, in order to determine the diagnosis of mental health patients.

There are currently two comprehensive systems that guide the diagnosis of mental health disorders. The *International Classification of Diseases* (ICD) is a health care classification system spearheaded by the WHO that categorizes both medical and mental health illnesses. It is currently in its 10th edition, with an 11th edition in development that is to be submitted to the WHO's World Health Assembly for official endorsement in by 2015. The DSM is a classification system specifically for mental health disorders that was developed by the American Psychiatric Association in 1952, the fifth edition of which was released in 2013. In the United States, health care professionals use ICD codes to designate diagnoses of all diseases and illnesses other than those that are considered mental health disorders, and they use the DSM to designate diagnoses of mental health disorders. Although the ICD is an international classification system, many clinicians outside the United States use the DSM. The WHO formed a DSM–ICD Harmonization Coordinating Group in 2009 with the aim of making the DSM-5 and ICD-11 diagnoses as uniform as is possible, which is a formidable task because only 1 of the 176 sets of diagnostic criteria present in previous versions of both classification systems was identical (First, 2009).

Psychological assessment is the process by which a clinician administers one or more assessment tools in order to gain comprehensive knowledge about a patient's current and past cognitive, emotional, and behavioral experiences, as well as about the patient's current and past life circumstances. An assessment tool can be a clinical interview, a psychological test, on observation, or even a physiological measurement. Among the many purposes of psychological assessment are the identification of acute and chronic life stressors that are exacerbating emotional distress, an understanding of the patient's level of functioning, the construction of a patient's personality profile, the assessment of the patient's strengths, and the development of a treatment plan that targets the most salient aspects of the patient's clinical presentation. However, perhaps the most fundamental purpose of psychological assessment is to arrive upon one or more mental health disorder diagnoses.

Current Knowledge and Schools of Thought

An array of assessment tools is available to assist mental health providers in determining accurate mental health disorder diagnoses. Some of these assessment tools measure the frequency and severity of psychiatric symptoms and cut across different theoretical orientations, whereas other assessment tools were developed within one particular theoretical orientation and yield data that advance a conceptualization on the basis of that theoretical orientation.

Clinical Interviews

According to Groth-Marnat (2009), "Probably the single most important means of data collection during psychological evaluation is the assessment interview" (p. 65). Clinical interviews can range from being unstructured and general to

 http://dx.doi.org/10.1016/B978-0-08-097086-8.21030-1

those that are structured or semistructured and that specifically target DSM or ICD symptoms of mental health diagnoses. The advantages of using a structured or semistructured clinical interview are that they (1) include carefully crafted questions that cover each criterion that contributes to the diagnosis of a mental health disorder, (2) have research demonstrating their strong psychometric properties (see Section Methodological Issues), and (3) are considered the 'gold standard' method of establishing mental health disorder diagnoses.

Structured and semistructured clinical interviews typically included several layers of questions. *Screening questions* are those that are first administered to respondents to assess the likelihood of meeting diagnostic criteria for a mental health disorder. Those who respond negatively to one or more screening questions for a mental health disorder are assumed not to meet diagnostic criteria for the disorder, and no further questions are asked. Those who respond affirmatively to one or more screening questions are asked an additional set of questions that correspond to the diagnostic criteria for that particular mental health disorder. Clinicians who administer the clinical interview are trained to assess for the presence or absence of each criterion, often assigning an indicator of whether the criterion is absent, present at a subthreshold level of clinical significance, or present at or above the threshold of clinical significance. Mental health disorder diagnoses are assigned to respondents when they meet the minimal number of diagnostic criterion at or above the threshold of clinical significance. Diagnoses can be assigned according to two time frames: (1) lifetime and (2) current (typically defined as occurring within the previous 12 months).

Structured and semistructured clinical interviews were developed mainly for research purposes, but they have been increasingly incorporated into clinical practice (Groth-Marnat, 2009). They are used most typically by clinicians who valued evidence-based practice and who have received training in the administration of these instruments, often in the context of research conducted in psychology and psychiatry departments that house large grant-funded clinical research. In the United States, perhaps the most commonly used structured or semistructured clinical interview is the Structured Clinical Interview for DSM Disorders (SCID), most recently developed to assess DSM-IV-TR (APA, 2000) diagnoses (First et al., 1996), and currently being adapted to assess DSM-5 diagnoses (see www.scid4.org). In addition, the WHO Composite International Diagnostic Interview (CIDI; Kessler and Üstün, 2004) yields diagnoses according to the 4th edition of the DSM and the 10th edition of the ICD. Administration of either of these instruments yields data to diagnose the major mental health disorders, including mood disorders, anxiety disorders, alcohol and substance use disorders, eating disorders, and other impulse control disorders.

In contrast, other clinicians, often who practise from a client-centered or psychodynamic orientation, use unstructured formats that allow patients to talk freely about their lives and 'tell their story.' Advantages of an unstructured approach include the opportunity to develop a strong sense of rapport, flexibility, and the potential to gather information about unique aspects of clients' lives (Groth-Marnat, 2009). The main disadvantage is that the reliability and validity (see Section Methodological Issues) of diagnoses that are made from this approach are unclear.

Self-Report Inventories

Self-report inventories, or *questionnaires*, are tools in which respondents answer a series of questions or make ratings of statements associated with cognitive, behavioral, and emotional characteristics of mental health disorders. Self-report inventories can ask respondents to indicate the presence or absence of these characteristics; the severity of the characteristics, often on a Likert-type scale; or the frequency of the characteristics. Many clinicians administer self-report inventories in addition to conducting a clinical interview because (1) they are usually easy to administer and complete, and (2) data obtained from them can be compared with normative data to yield information about where the patient stands relative to a people representative of the general population.

There are literally thousands of self-report inventories. Many of them are referenced in the Buros Mental Measurements Yearbook, a compilation of test reviews and evaluations that is now in its 18th edition (Spies et al., 2010), with a 19th schedule to be published in early 2014, or in the two volume compilation, *Measures in Clinical Practice and Research* by Corcoran and Fischer (2013a,b).

Many clinicians administer self-report inventories that assess a range of psychiatric symptoms in order to obtain a general sense of areas of concern for patients, with the idea that they will follow up on these areas using clinical interviewing in order to establish a diagnosis of a mental health disorder. For example, the Symptom Checklist-90-R (Derogatis, 1994) is a 90-item list of symptoms that requires respondents to rate their severity on a 5-point Likert scale (0 = not at all; 4 = extremely). The measure yields nine scores representative of specific clinical presentations (e.g., obsessive compulsive, phobic anxiety), as well as three global indices (e.g., Global Severity Index). Another commonly used measure to assess symptoms associated with an array of mental health disorders is the Patient Health Questionnaire (PHQ) (Spitzer et al., 1999), a self-report inventory that was developed from the Primary Care Evaluation of Mental Disorders (Spitzer et al., 1994). It assesses 11 major domains of symptoms, with a number of items corresponding to specific symptoms in these domains. It yields information to assist in the diagnosis of depressive, anxiety (particularly panic), somatoform, alcohol use, and eating disorders.

Other self-report inventories assess symptoms associated with one specific psychiatric disorder or clinical condition. One of the most common self-report inventories used in clinical research and practice to assist in arriving upon a diagnosis of major depressive disorder is the Beck Depression Inventory-II (BDI-II; Beck et al., 1996). The BDI-II is a 21-item inventory that requires respondents to endorse the severity of depressive symptoms on a 0–3 scale. It classifies respondents as having no or minimal, mild, moderate, or severe depressive symptoms, and research has shown that it can identify cases of depression as effectively as more costly and time-consuming structured interviews (Stukenberg et al., 1990). The lead developer of the BDI-II,

Dr Aaron Beck, has also developed similar inventories to assist in the diagnosis and conceptualization of other psychiatric disorders, including anxiety (Beck Anxiety Inventory; Beck and Steer, 1990), hopelessness (Beck Hopelessness Scale; Beck and Steer, 1988), and cognitive insight in schizophrenia (Beck et al., 2004). The PHQ, described previously, has also been adapted to assess symptoms of particular disorders, such as depression (i.e., PHQ-9; Kroenke et al., 2001) and generalized anxiety disorder (Spitzer et al., 2006).

In addition to these inventories that assess symptoms of major mental health disorders, there are also self-report inventories whose data yield important insights into the diagnosis of personality styles and disorders. The Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher et al., 2001) is a 567-item self-report inventory that yields scores for 10 clinical scales (e.g., hypochondriasis, depression), 10 validity scales (e.g., lie, infrequency), and 15 content scales (e.g., anxiety, fears). Rather than using responses on individual scales to facilitate the diagnosis process, clinicians evaluate the clients' elevations on some scales relative to other scales in order to identify particular personality profiles. The Millon Clinical Multiaxial Inventory-III (MCMI-III; Millon, 1997) is a 175-item self-report inventory that yields scores for 11 clinical personality patterns (e.g., avoidant, dependent), 3 severe personality pathologies (e.g., borderline), 7 clinical syndromes (e.g., anxiety, somatoform), and 3 severe syndromes (e.g., thought disorder), as well as 4 validity scales (e.g., disclosure). Scores from the MCMI yield important data that can assist clinicians in making personality disorder diagnoses (i.e., Axis II diagnoses according to the DSM).

Other Assessment Tools

Assessment tools need not be limited to those in which clients respond to straightforward questions about their behavior and psychological experiences, either via clinical interview or self-report. For example, *behavioral observation* is a tool in which a clinician establishes, a priori, behavioral manifestations of a mental health disorder, as well as their antecedents and consequences, and then observes the frequency of the behaviors in a particular individual within a particular time frame. It is used most frequently by clinicians with a behavioral orientation in order to establish specific targets for behavior therapy. *Assessment of intellectual functioning* (e.g., Wechsler Adult Intelligence Scale-IV (WAIS-IV; Wechsler, 2008a) and Wechsler Intelligence Scale for Children (Wechsler, 2008b)) is done to establish an index of a person's overall cognitive ability, as well as specific cognitive functions such as verbal ability and short-term memory. It is administered most frequently to aid in the diagnosis of intellectual disability and learning disorders, as well as in the diagnosis of dementia-related disorders. *Thought sampling* is a tool in which respondents list their cognitive contents, or what is running through their minds, at predetermined intervals or in response to a particular stimulus. It is used most frequently by clinicians with a cognitive orientation in order to develop a case conceptualization and establish specific targets for cognitive therapy. *Projective tests* are tools that allow respondents to interpret ambiguous stimuli, with the idea that their responses provide insight into unconscious impulses or drives (e.g., The Rorschach (Rorschach, 1941) and Thematic Apperception Test (TAT; Morgan and Murray, 1935)). They are used most frequently by clinicians with a psychodynamic orientation.

Clinical Judgment

Clinical judgment is defined as a clinician's attempt to assimilate objective and subjective data gathered from a psychological assessment and arrive upon one or more diagnoses of a mental health disorder (cf Groth-Marnat, 2009). Clinical judgment affects psychological assessment on many levels. For example, clinicians use clinical judgment when they decide how to conduct their psychological assessment and the particular domains to be assessed. When administering clinical interviews, they use their judgment as to the kind of degree of clarification they ask of the patient in order to accurately determine whether they meet the diagnostic criterion under consideration at a level of clinical significance.

Although results from structured clinical interviews like the SCID and CIDI are usually considered the 'gold standard' in identifying diagnoses of mental health disorders, if additional assessment tools are included in a psychological assessment, then the clinicians use clinical judgment in determining the amount of weight they put on the data they obtained from various assessment tools and the manner in which they reconcile inconsistent data. Research has shown that clinicians often make errors in judgment when assimilating the results of psychological assessment in order to formulate a diagnosis of a mental health disorder. For example, some clinicians do not account for the base rate of the diagnosis in the general population (Hawkins and Hastie, 1990); others do not receive feedback regarding the accuracy and predictive ability of their diagnoses (Garb, 2005). In addition, research shows that information gathered earlier in an assessment is often given more importance than information gathered later in the assessment, resulting in a confirmatory bias in which clinicians do not thoroughly seek out data that will disconfirm their original impressions (Haverkamp, 1993). Groth-Marnat (2009) made eight recommendations for reducing bias in clinical judgments, including: (1) using comprehensive structured or semistructured clinical interviews to avoid missing important information; (2) considering evidence that is inconsistent with initial hypothesis; (3) paying careful attention to ICD or DSM criteria; (4) relying on careful note taking rather than on memory; (5) considering base rates of diagnoses being considered; (6) seeking feedback regarding the accuracy of clinical judgments; (7) staying abreast of the theoretical and empirical literature on the diagnosis under consideration; and (8) staying abreast of the literature on clinical judgment to be aware of new trends.

Treatment Planning and Psychotherapy

When a patient presents for treatment, he or she will typically undergo some sort of psychological assessment using the approaches described previously so that his or her mental health provider can arrive upon one or more diagnoses. The mental health provider and the patient then collaboratively

develop a treatment plan in which specific targets for treatment are identified. The treatment plan is shaped, at least in part, by the mental health disorder diagnosis(es) that the patient has been assigned. There are many manualized treatments that have been developed for specific mental health disorder diagnoses, and many mental health providers will administer the treatments described in these manuals to alleviate their patients' mental health disorder. A caveat to this statement, however, is that the majority of published treatment manuals describe cognitive, behavioral, or cognitive behavioral treatments, and clinicians who practice from other theoretical orientations may choose other approaches to treatment even if a treatment manual is available for the specific diagnosis that they have assigned to their patient.

It is also important to acknowledge that many psychological processes cut across diagnostic categories. Critics of manualized treatments observe that the number of treatment manuals that mental health providers must learn is unwieldy (Beutler, 2000). Thus, there is movement toward developing psychotherapeutic approaches that are transdiagnostic in nature and that are based on evidence-based psychological principles (e.g., Barlow et al., 2004) rather than on entire treatment packages that are limited to one specific mental health disorder diagnosis. In their manual on the transdiagnostic treatment of emotional disorders (e.g., depression, anxiety), Barlow et al. (2011) indicate that the core features of emotional disorders are inappropriate emotional responding, lack of skill in emotion regulation, and a perception that one does not have the ability to manage emotional responses, and they describe corresponding strategic interventions that target these cognitive, behavioral, and emotional processes regardless of the specific diagnosis that is assigned to the patient.

Methodological Issues

For psychological tests to be useful, they must have adequate psychometric properties. The two most important psychometric properties are reliability and validity. *Reliability* refers to the stability and consistency of data obtained by the psychological assessment. Reliability is demonstrated when the assessment tool yields the same results if the person takes the test on a different occasion or if the test was administered by a different clinician. The higher an assessment tool's reliability, the less chance there is that measurement error influenced the results. An assessment tool's reliability can be calculated in four ways. An estimate of *test–retest reliability* is determined by correlating the scores yielded by an assessment tool administered on two separate occasions. An estimate of *alternative forms reliability* is determined by correlating the scores yielded by two supposedly equivalent versions of an assessment tool administered on two separate occasions. In contrast, an estimate of *internal consistency* can be obtained in a single administration of the assessment tool. It is calculated by obtaining correlations between responses to specific items on the test, with the idea that the correlations should be high if the items are all measuring the same construct. Finally, an estimate of *interrater reliability* is determined by correlating the scores of an assessment tool, which is either administered on two separate occasions with two different test administrators, or it is administered on one occasion and scored by two different individuals. Because estimates of reliability are based on correlations, they always range from 0.0 to 1.0, with higher numbers indicative of higher reliabilities.

Validity refers to the extent to which the assessment tool is measuring what it is supposed to measure. There are several ways to establish the validity of an assessment tool. *Content validity* refers to the degree to which the items on an assessment tool are representative of and relevant to the construct being measured. *Criterion validity* is an estimate of the degree to which data obtained from an assessment tool correlate with performance on other measures assessing similar constructs. There are two types of criterion validity. *Concurrent validity* refers to the correlation between measurements taken at the same time, such as two assessment tools measuring similar constructs that are administered in a single sitting. *Predictive validity* refers to the correlation between an assessment tool administered on one occasion and data obtained at a later-time, such as a measure of psychotic symptoms upon admission to the hospital and the number of days the patient stayed in the hospital. *Construct validity* is the degree to which an assessment tool measures a theoretical construct or trait. It is often obtained by examining the degree to which data obtained from an assessment tool correlate with data obtained by measures of a similar theoretical construct (i.e., *convergent validity*), as well as the degree to which data obtained from an assessment tool do not correlate with data obtained by measures of a dissimilar theoretical construct (i.e., *discriminant validity*). *Face validity* refers to the degree to which the items on the assessment tool are sensible to the test taker, which can affect the manner in which the test taker approaches the assessment. Finally, *incremental validity* refers to the degree to which the results from any one assessment tool yield information above and beyond that obtained by other assessment tools, which has implications for the efficiency of the psychological assessment process and its cost-effectiveness.

These psychometric constructs have important implications for the process of achieving a mental health diagnosis. Clinicians should use assessment tools with adequate reliability and validity in order to ensure the accuracy of the mental health diagnosis. In addition, they should choose their assessment tools carefully so that the data obtained from each assessment tool yield unique information and so that the assessment battery is not unwieldy, as a lengthy assessment battery is associated with patient fatigue and boredom.

Another methodological issue to consider in using assessment tools is the degree to which they accurately identify patients with the particular mental health diagnosis under consideration. *Sensitivity* refers to the percentage of true cases of the mental health diagnosis that are identified by the particular assessment tool, whereas *specificity* refers to the percentage of true negatives that are identified by the particular assessment tool. Ideally, an assessment tool would be characterized by high levels of both sensitivity and specificity. However, in reality, sensitivity and specificity covary, such that specificity often decreases when sensitivity increases, and sensitivity decreases when specificity increases. The optimal ratio of sensitivity to specificity depends on the costs and benefits

of identifying a large number of false positives and of missing true negatives.

Change in Focus or Emphasis Over Time

Psychological assessment is central to the training, and even to the identities, of clinical psychologists. However, the amount of time that clinicians devote to psychological assessment has decreased substantially as the role of psychologists has expanded beyond test administration (Groth-Marnat, 2009). Surveys of psychological test usage have shown that the 10 most frequently used tests have remained remarkably stable over the past several decades and include many of the assessment tools mentioned previously, such as the BDI, MMPI, WAIS, Rorschach, and TAT (Camara et al., 2000). At the same time, there has been a decrease in the use of projective tests (Piotrowski, 1999), which could be due to multiple factors such as their questionable norms and validity, the extensive time required to learn their administration, and the availability of more cost-effective assessment tools (Groth-Marnat, 2009). Another change that has occurred over the second half of the twentieth century and into the twenty-first century is that the pool of available diagnoses has greatly expanded, requiring modern clinicians to be cognizant of and potentially assess for many mental health disorders that had not been considered by clinicians in the past.

Current and Emerging Directions in Theory and Research

Four emerging directions in theory and research associated with diagnostic processes deserve note. First, new versions of structured clinical interviews are being developed, which assess mental health disorders that correspond to the newest versions of the DSM and ICD. A host of empirical research will emerge from these developments, such as research that establishes the psychometric properties of these interviews, as well as epidemiological research that establishes the prevalence and incidence of mental health disorders according to these most recent diagnostic criteria. Second, there is a pressing need to identify effective, yet quick and user-friendly, screening tools that identify individuals that likely meet criteria for mental health disorders in settings such as primary care, obstetrics and gynecology, and pediatrics. Not only does research on this topic focus on the psychometric properties (e.g., sensitivity and specificity) of these screening tools, but it also examines the degree to which identified individuals follow through with referrals for treatment and barriers to implementing screening programs in nonmental health settings. Third, the era of managed care has prompted research into cost-effective mental health procedures. Relevant to diagnostic processes is the increasing reliance on computerized assessment. Research on this topic not only establishes the psychometric properties of these instruments, but also their correspondence with diagnoses arrived upon by face-to-face contact with a clinician. Finally, there has been long-standing debate about the best approach to diagnosing mental health disorders, and particularly personality disorders, in the DSM, with one side arguing for a categorical approach, and the other side arguing for a dimensional approach (Simonsen, 2010). In fact, a dimensional approach to diagnosing personality disorders (i.e., Axis II disorders) was considered for the DSM-5 (Skodol, 2010), but ultimately not adopted. However, research will continue to examine alternative approaches to categorical diagnosis and the degree to which their utility exceeds categorical approaches.

See also: Intellectual Functioning, Assessment of; Personality Assessment; Projective Methods in Psychology.

Bibliography

American Psychiatric Association [APA], 2000. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR), fourth ed. – text revision. American Psychiatric Association, Washington, DC.

American Psychiatric Association [APA], 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5), fifth ed. American Psychiatric Association, Washington, DC.

Barlow, D.H., Allen, L.B., Choate, M.L., 2004. Toward a unified treatment for emotional disorders. Behavior Therapy 35, 205–230.

Barlow, D.H., Ellard, K.K., Fairholme, C.P., Farchione, T.J., Boisseau, C.L., Ehrenreich-May, J.T., 2011. Unified Protocol for Transdiagnostic Treatment of Emotional Disorders. Oxford University Press, New York.

Beck, A.T., Baruch, E., Balter, J.M., Steer, R.A., Warman, D.M., 2004. A new instrument for measuring insight: the beck cognitive insight scale. Schizophrenia Research 68, 319–329.

Beck, A.T., Steer, R.A., 1988. Manual for the Beck Hopelessness Scale. The Psychological Corporation, San Antonio, TX.

Beck, A.T., Steer, R.A., 1990. Beck Anxiety Inventory Manual. The Psychological Corporation, San Antonio, TX.

Beck, A.T., Steer, R.A., Brown, G.K., 1996. Beck Depression Inventory-II Manual. The Psychological Corporation, San Antonio, TX.

Beutler, L.E., 2000. Empirical based decision making in clinical practice. Prevention & Treatment 3. Article 27.

Butcher, J.N., Graham, J.R., Ben-Porath, Y.S., Tellegen, A., Dahlstrom, W.G., Kaemmer, B., 2001. Minnesota Multiphasic Personality Inventory: Manual for Administration and Scoring, second ed. University of Minnesota Press, Minneapolis, MN.

Camara, W.J., Nathan, J.S., Puente, A.E., 2000. Psychological test usage: implications in professional psychology. Professional Psychology: Research and Practice 32, 289–297.

Corcoran, K., Fischer, J., 2013a. Measures for Clinical Practice and Research: A Source Book, Volume 1, Couples, Families, and Children, fifth ed. Oxford University Press, New York.

Corcoran, K., Fischer, J., 2013b. Measures for Clinical Practice and Research: A Source Book, Volume 2, Adults, fifth ed. Oxford University Press, New York.

Derogatis, L.R., 1994. SCL-90-R: Administration, Scoring, and Procedures Manual, third ed. National Computer Systems, Minneapolis, MN.

First, M.B., 2009. Harmonisation of the ICD-11 and DSM-V: opportunities and challenges. The British Journal of Psychiatry 195, 382–390.

First, M.B., Spitzer, R.L., Gibbon, M., Williams, J.B.W., 1996. Structure Clinical Interview for Axis I DSM-IV Disorders: Research Version – Patient Edition (SCID-I/P, Version 2.0). New York State Psychiatric Institute, Biometrics Research Department, New York.

Garb, H.N., 2005. Clinical judgment and decision making. Annual Review of Psychology 1, 67–89.

Groth-Marnat, G., 2009. Handbook of Psychological Assessment, fifth ed. Wiley, Hoboken, NJ.

Haverkamp, B.E., 1993. Confirmatory bias in hypothesis testing for client-identified and counselor self-generated hypotheses. Journal of Counseling Psychology 40, 303–315.

Hawkins, S.A., Hastie, R., 1990. Hindsight: biased judgments of past events after the outcomes are known. Psychological Bulletin 107, 311–327.

Kessler, R.C., Üstün, T.B., 2004. The world mental health (WMH) survey initiative version of the world health organization (WHO) composite international diagnostic interview (CIDI). International Journal of Methods in Psychiatric Research 13, 93–121.

Kroenke, K., Spitzer, R.L., Williams, J.B.W., 2001. The PHQ-9: validity of a brief depression severity measure. Journal of General Internal Medicine 16, 606–613.
Millon, T., 1997. Millon Clinical Multiaxial Inventory-III Manual, second ed. National Computer Systems, Minneapolis, MN.
Morgan, C.D., Murray, H.A., 1935. A method for investigating fantasies. AMA Archives of Neurology and Psychiatry 34, 389–406.
Piotrowski, C., 1999. Assessment practices in the era of managed care: current status and future directions. Journal of Clinical Psychology 55, 787–796.
Rorschach, H., 1941. Psychodiagnostik (Hans Huber Verlag, Trans.). Bircher, Bern, Switzerland (Original work published 1921).
Simonsen, E., 2010. The integration of categorical and dimensional approaches to psychopathology. In: Millon, T., Krueger, R.F., Simonsen, E. (Eds.), Contemporary Directions in Psychopathology: Scientific Foundations of the DSM-V and ICD-11. The Guilford Press, New York, pp. 350–361.
Skodol, A.E., 2010. Dimensionalizing existing personality disorder categories. In: Millon, T., Krueger, R.F., Simonsen, E. (Eds.), Contemporary Directions in Psychopathology: Scientific Foundations of the DSM-V and ICD-11. The Guilford Press, New York, pp. 362–373.
Spies, R.A., Carlson, J.F., Gesinger, K.F., 2010. The Eighteenth Mental Measurements Yearbook. The University of Nebraska Press, Lincoln, NE.
Spitzer, R.L., Kroenke, K., Williams, J.B.W., For the Patient Health Questionnaire Primary Care Study Group, 1999. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. JAMA 282, 1737–1744.
Spitzer, R.L., Kroenke, K., Williams, J.B.W., Löwe, B., 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. Archives of Internal Medicine 166, 1092–1097.
Spitzer, R.L., Williams, J.B.W., Kroenke, K., Linzer, M., deGruy, F.V., Hahn, S.R., Brody, D., Johnson, J.G., 1994. Utility of a new procedure for diagnosing mental disorders in primary care: the PRIME-MD 1000 study. JAMA 272, 1749–1756.
Stukenberg, K.W., Dura, J.R., Kiecolt-Glazer, J.K., 1990. Depression screening scale validation in an elderly, community-dwelling population. Psychological Assessment 2, 134–138.
Wechsler, D., 2008a. Wechsler Adult Intelligence Scale, fourth ed. NCS Pearson, Inc., San Antonio, TX.
Wechsler, D., 2008b. Wechsler Intelligence Scale for Children. Administration and scoring manual, fourth ed. NCS Pearson, Inc., San Antonio, TX.
Wedding, D., Faust, D., 1989. Clinical judgment and decision making in neuropsychology. Archives of Clinical Neuropsychology 4, 233–265.

Relevant Websites

http://www.med.upenn.edu/suicide/beck/scales.html – Beck Scales.
http://www.hcp.med.harvard.edu/wmhcidi/index.php – CIDI.
http://www.ebscohost.com/academic/mental-measurements-yearbook – Mental Measurements Yearbook.
http://www1.umn.edu/mmpi/ – MMPI.
http://www.phqscreeners.com/ – PHQ.
http://www.scid4.org/ – SCID.

Links to Databases and Software

www.unl.edu/buros/ – Mental Measurement Database.

Dialectology

Jack K Chambers, University of Toronto, Toronto, ON, Canada

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 3637–3642, © 2001, Elsevier Ltd.

Abstract

Dialectology, the systematic study of language variation, had its effective beginnings in Europe after 1850, but it has its roots in an age-old fascination with accent and dialect. Dialectology first evolved as a classificatory discipline that concentrated mainly on the speech patterns of the most conservative speakers, called NORMs, an acronym for nonmobile, older, rural males. Its consummation came with the *Atlas linguistique de France* (10 Vols., 1901–13) directed by Jules Gilliéron at the Sorbonne, which provided an influential precedent for 75 years. The rise of linguistics after 1916 brought with it initiatives for integrating dialectology, but these were mostly resisted. As social changes in the second half of the twentieth century diminished the rural, uneducated, isolated populaces that had been the dialectologists' object of study, the discipline declined. However, it was revitalized by two concurrent forces, one technological and the other methodological. The first is computerization, which provides dialectologists with the means of managing the superfluity of data that had been both its strength and its burden. The second is sociolinguistics, the study of language variation on social-science principles, which defines its domain more inclusively in terms of populations and social settings. Dialectology, thus reformed, has again become a thriving intellectual enterprise.

Dialectology, the systematic study of linguistic variation, began as an autonomous discipline in the second half of the nineteenth century and became subsumed as a branch of linguistics in the second half of the twentieth century, though not without difficulty. Today it has worldwide currency and hundreds of active participants engaged in diverse research.

Beginnings

Its beginning is conventionally dated as 1876, when the German professor Georg Wenker mailed a list of sentences written in standard German to schoolmasters in the north of Germany and asked them to return the sentences to him transcribed into the regional dialect of their school districts. Wenker made further mailings until 1887 by which time he had blanketed the entire nation. He ultimately sent his list to nearly 50 000 schoolmasters, and he received completed forms from 45 000 of them, The schoolmasters, with few guidelines and no precedents, provided Wenker with thousands of regional variants of all kinds: lexical, pronunciation, phonological, morphological, and syntactic. Overburdened by data, Wenker was forced to limit his analyses to a few variants found in a closely circumscribed area of north and central Germany, which he then painstakingly transcribed by hand onto two sets of maps bound as *Sprachatlas des Deutschen Reichs* and deposited in Marburg and Berlin in 1881 (Wenker, 1881).

Thirty years before Wenker, the Finnish polymath Anto Warelius (1821–1904), approached the study of dialect from an entirely different perspective in a scholarly undertaking that remains virtually unknown outside his own country, but which has a prior claim over Wenker as the first truly systematic study. In 1846, Warelius set out on foot from Hamina in southeastern Finland and walked in a northwesterly line, eliciting and transcribing words and expressions from the villagers he encountered across almost 400 km (Rapolo, 1969; Chambers, 2000).

Warelius remains an obscure figure in the history of dialectology, but this method of eliciting dialect forms in face-to-face interviews with villagers foreshadowed the predominant method used by professional dialectologists for the next 150 years. The influential precedent for the method of direct fieldworker elicitation was set in 1896 by Jules Gilliéron, a Swiss professor at the Sorbonne in Paris, in the dialect survey of France. Gilliéron trained an obscure grocer, Edmond Edmont, in the craft of phonetic transcription, and dispatched him into the French countryside on his bicycle. Edmont has become a legendary figure in dialectology for his unflagging devotion as he pedaled into 639 villages and interviewed 700 *paysans* in less than 4 years. He sent his transcriptions to Gilliéron in Paris at intervals, where they were incorporated into a cumulative mapping enterprise, and publication of the maps began in 1901 and ended in 1910 with the thirteenth volume of the *Atlas linguistique de France* (ALF). Gilliéron and Edmont's atlas (1901–10) stands as an early consummation of dialect geography and became the influential precedent for future projects for almost a century.

Impetus

The motivation for these methodical, large-scale studies in the second half of the nineteenth century is conventionally held to be a reaction to Neogrammarian philology, one of the great intellectual currents of the day. The Neogrammarians demonstrated law-like regularities in the historical descent of modern languages from classical ones. One of their tenets was *Ausnahmslosigkeit der Lautgesetze* 'sound changes are exceptionless.' Such a hypothesis, bold and admirably empirical, invites refutation, and a likely source would come from dialect data gathered from highly differentiated, relatively isolated villages.

The view that the first dialectologists rose to the Neogrammarian challenge has romantic resonance, but it is not strongly supported by facts. Neither Wenker nor Gilliéron nor

any of their successors actually marshaled their data for elucidation, revision or refutation of any overriding linguistic principles, Neogrammarian or otherwise. In practice, the main results of their scholarship were descriptive rather than analytical, and their expressed goals aspired to no more. The *Sprachatlas* became the preferred medium for dissemination of results, with linguistic tokens carefully arrayed on geographic representations, usually with commentary on secondary variants which had to be omitted because they would overcrowd the map, and occasionally with isoglosses, lines drawn between locations where different variants tended to occur.

Rather than Neogrammarianism, the main impetus for dialectology as a systematic discipline appears to have been Linnaean taxonomy, another forceful current of nineteenth century intellectual history. Dialect data lends itself quite naturally to the taxonomic bent, with numerous gradations subcategorizable as alloforms of -emic categories roughly analogous to genera and their species.

Though taxonomy dictated the form that dialect studies took, the ultimate impetus has a much deeper source in human nature. History is replete with observations about dialect, going back at least three millennia in the Western tradition to the myth of Babel in *Genesis* 11: 1–9, whereby the Lord punished his ambitious subjects by 'confusing their speech, so that they will not understand what they say to one another.' Since Babel there has been a continuous tradition of astute comment on dialect differences which makes a long prelude to the institution of dialectology as a systematic classificatory enterprise. The English scholar Alexander Ellis remarked in 1875 that 'collecting country words is looked upon as an amusement, not as laying a brick in the temple of science' (quoted by Chambers and Trudgill, 1998, p. 15). Ellis's spirit thrives, and is evident today in word columns in newspapers, tourist pamphlets on the local 'twang,' stand-up mimicry in theaters, news features on slang, and in dozens of other ways. The discipline of dialectology appears to be the scholarly manifestation of this natural human curiosity about language variation.

Methods

Until recently, dialectology gave primacy to recovering the oldest extant forms of language. This deliberate bias colored its methodology at every point. Fieldworkers chose their subjects as a judgment sample of NORMs, an acronym for a prototype who was nonmobile, older, rural, and (predominantly) male. For instance, Edmont's 700 subjects, all ofwhom were rural, included only about 200 educated people and 60 women. The NORM prototype was purposeful. Lack of mobility guaranteed that their speech would be characteristic of their village. Harold Orton, director of the Survey of English Dialects (SED), England's nationwide survey on Gilliéronian principles for which data was gathered from 1950 to 1961, said, 'dialectspeakers whose residence in the locality had been interrupted by significant absences were constantly regarded with suspicion' (Orton, 1962, 15–16). Old age guaranteed that their speech would reflect a bygone era. Hans Kurath, director of the *Linguistic Atlas of New England* (LANE), part of the Linguistic Atlas of the USA and Canada that was begun in 1931 and still continues, declared that 'since most of the informants … are over 70 and not a few over 80, it would seem that we shall be able to establish the regionalism of the pre-industrial era of New England' (Kurath, 1949, p. 27). Rurality eliminated mobile urban dwellers who were likely to be exposed to other dialects and accents (Chambers, 2000). Males were targeted because their speech is less class-conscious than women's. As Orton put it, speaking about England, 'In this country, men speak vernacular more frequently, more consistently, and more genuinely than women' (1962, p. 15).

The basic methodology of traditional (Gilliéronian) dialect geography is to send trained fieldworkers to the locations where the subjects live in order to elicit specific items from the subjects guided by a standard questionnaire. The questionnaires are lengthy (Gilliéron's no less than 1500 questions, the SED about 1200, LANE about 700), with questions usually ordered by subject matter (flora and fauna, the barnyard, chores, etc.). Its purpose is to guarantee comparable data from all subjects. The fieldworker is expected to transcribe only the key word or phrase from the subject's response, no matter how long it might be. Labov summarized the method as 'a long question from the interviewer and a short answer from the subject' (Labov, 1971, p. 113).

These methods have been subject to criticism from commentators versed in social-science methodology, especially with respect to the reliability of the population sample and the validity of the data (for instance, Pickford, 1956). Obviously, NORMs constitute a rarefied minority when the majority population is exactly the converse – mobile, younger, urban, and female (Chambers, 1993, p. 134). But criticisms like these largely miss the point about the antiquarian bias that motivates the methods. Raven I. McDavid, a venerated fieldworker on the American project and an articulate apologist, defended the traditional methods with this succinct capitulation: 'There is a deliberate bias in the choice of communities – in the direction of smaller and often isolated places; there is a deliberate bias in the selection of informants, insistence on local roots and a statistically disproportionate sample of the older and less educated; there is a deliberate bias in the choice of items for the questionnaire, in the direction of the old fashioned rural society, the horsedrawn vehicle, the mule-powered plow and homemade bread. All of these biases are essential to the purpose of the investigation, to push the historical records as far back in time as possible' (McDavid, 1981, p. 71). Even as he articulated this defence, he inadvertently exposed the limits of the tradition at a time when modern technology had eliminated horsedrawn vehicles and social forces were inexorably diminishing the ranks of the uneducated and, indeed, rural society.

Structuralist Analyses

With the advent of Saussurean structural linguistics in 1916, language studies took a decisive turn in the direction of synchronic analysis. Taxonomy fell out of favor and intensionality gained adherents, at least in theory. However, dialectology clung to its autonomy, no more moved by structuralism than it had been by Neogrammarianism. Over the years, linguists made several attempts at integrating dialectology into linguistics. For instance, Uriel Weinreich (1954) proposed a level of analysis called the diasystem, which in

effect posited a level above the phonemic where 'diaphones' could be segmented into the actual phonemes in two (or more) specific dialect systems by means of interpretive rules. Weinreich speculated that speakers must use something comparable to a diasystem for comprehension when they interact with speakers of other dialects.

Later, with the development of generative frameworks, there were similar attempts at integrating dialectology. For example, Keyser reanalyzed traditional atlas data from the Atlantic coast of the USA by demonstrating that several vowel distinctions found in the painstaking tabulations for dozens of speakers could be generated simply by positing two phonological rules and 'accounting for the formation of new dialect groups through the geographical dissemination of rules' (Keyser, 1963, p. 313). Efforts like these gained few adherents, and they remained peripheral to the linguistic mainstream and had little impact on dialect studies.

More successful were the uses of atlas data in elucidating problems of historical linguistics. For instance, Hall (1943) used data collected in the *Sprachund Sachatlas Italiens und der Südschweiz* ([AIS] Jaberg and Jud 1928–40) to demonstrate the persistence of the old political boundary of the Papal States on patterns of dialect differentiation in twentieth-century central Italy. The purposes of historical linguistics were of course consonant with the intentions of traditional dialectology, although the exploitation of the latter by the former offered much greater potential than was actually realized. It did, nevertheless, establish a vital, if minor, tradition.

Decline and Rebirth

As post-Saussurean language studies continued to be dominated by synchronic linguistics and entered a period of considerable currency stimulated by the ideas of Noam Chomsky, the failure of dialectology to become integrated caused it to slip ever further into the periphery of language studies. By the middle of the twentieth century, scholarly activities had slowed and become relatively inconspicuous. By 1980, even dialectologists began to realize that reform was necessary. In the first book-length summation of the field, Chambers and Trudgill stated: 'one might well wonder what the future holds for dialect geography. There is no doubt whatever that the financing of research and publishing for large-scale surveys is a great problem, and there is also no doubt that capable and enthusiastic practitioners of the calibre of those who went before are difficult to find; but it seems most likely that both of these facts follow from another fact, namely, that dialect research has taken a new direction … It is probably safe to say that the future of dialect geography depends upon the ability of its practitioners to embrace and incorporate the concerns and perhaps the methodology of urban dialectology' (Chambers and Trudgill, 1980, p. 23).

Many aspects of dialect geography came under criticism when they were noticed at all, including the utility of the *Sprachatlas* as a medium for disseminating results. 'For whom are these expensive volumes designed?' Macaulay asked, and then he said: 'Far from being a visual aid the maps are probably a barrier to grasping the pattern of distribution and it seems very unlikely that anyone nowadays would attempt this kind of mapping. Given the sorting power of computers, basic information is much better stored in a form that is suitable to machine processing than in conventional map form' (Macaulay, 1985, p. 180).

With the founders gone, the rearguard reaction to these criticisms was muted. Starting in the mid-1980s, dialectology underwent a renaissance from which it would emerge as a vital and vigorous discipline with greater worldwide currency than ever before and, not coincidentally, greater intellectual depth. The sources were the antidotes identified in the two critiques above: the embracing of powerful new technology and the adaptation of more inclusive research methods.

Computerization

From its inception, dialectology had been burdened by an overabundance of data. Dialectology aspired to be, in Saussurean terms, a science of *parole*, that is, the study of language as a communal effusion, apprehended in actual speech events, imbued by and indeed mediating what Sapir called 'the apparent lawlessness of social phenomena' (Sapir, 1929, p. 166). But the profusion of data proved daunting, and the methods used in classic dialect geography came into being partly as devices for delimiting the effusion and subjugating its lawlessness, by focusing on a narrow, linguistically conservative population, restricting elicitation to a closed set of items, and transcribing only individual tokens from the flow of speech.

The development of tape-recording as a reliable, affordable technology in the 1950s provided a means for making a permanent record of the flow of speech. However, the immediate impact of tape-recording on dialectology was negligible, coming into use merely for recording traditional elicitation sessions and thus serving mainly to provide archival documentation of the practices of McDavid and other fieldworkers. The most innovative of all regional surveys, the *Linguistic Atlas of the Gulf States* (Pederson et al., 1986–92 [LAGS]), directed by Lee Pederson, made a significant advance by recording all its interviews so they could be transcribed later under laboratory conditions. Meanwhile, beyond the formal confines of dialectology, tape-recording was embraced as a liberating medium by a few linguists genuinely interested in studying vernacular speech. In this way, tape-recording stimulated the rise of sociolinguistics, an urban dialectology with methods and goals that bore little resemblance to dialect geography.

More significant technology for the revitalization of dialectology came with the development of computers as reliable data-handling devices in the late 1960s, and especially as affordable and accessible universal processors in the 1980s. As Kretzschmar and colleagues put it: '… the development of dialect studies, whether geographical or sociolinguistic, has always been hampered by a superfluity of data…. Even smaller surveys have had to settle for selective analysis of their data because the wealth of possibilities for analysis overran the editors' time and the human capacity for holding in the mind only so much information at once. Computers can help overcome these problems: they are wonderful tools for quickly sorting and matching pieces of information and for performing complex calculations on the results, and…. they are practically unlimited in their ability to store data' (Kretzschmar et al., 1989, p. v).

Many dialectologists recognized the computers' potential for relieving the perpetual problem of overabundance. No less important, many scholars recognized the enormous advantages that would accrue from the computerization of the classic databases of the ALF (Heap, 1997), SED (Ramisch, 1997; Elmer and Rudin, 1997), AIS (Goebl, 1993), and the others. Generalizations buried in lists and tables for many decades could be disinterred by judicious computation.

Dialectometry

A significant development directly attributable to computational processing is dialectometry, literally, the 'measure of dialect.' The word *dialectométrie* was coined by Pierre Séguy, director of the *Atlas Linguistique de Gascogne*, one of the regional surveys comprising the *Nouvel Atlas de la France*. Séguy hoped to replace subjective judgments on the occurrence of dialect regions by quantifying the Gascony data in such a way that the regions would be revealed objectively – in effect, they would reveal themselves. In the sixth and final volume of the Gascony atlas (Séguy, 1973), Séguy and his associates – *mes élévescamarades*, he called them – published a set of maps, which linked each of the survey sites to its nearest neighbors by a number representing the percentage by which the two sites differed with respect to the forms elicited there. Séguy conceived the percentages as an index for measuring the 'linguistic distance' between the two places. The analyst, by inspecting the distance indices, could draw inferences about dialect regions based on degrees of similarity and difference.

Though Séguy's methods now look primitive, restricted as they were at the time to univariate comparisons and handmade cartography, the numerical gradations across the maps rightly impressed many dialectologists as the first approximation of a visual analogue for the dynamics of dialect gradation in space. Since then, dialectometry has advanced to multivariate statistical programs and computer generated multidimensional displays that simultaneously compare all sites to one another. Dialectometry has become a generic term for all kinds of statistical dialect geography linked, as Goebl says, by a common goal: 'The preponderant aim of dialectometry consists in discovering, by the numeric combination of many low ranking patterns, higher ranking patterns which have remained hitherto hidden, in order to obtain a systematic insight into the problem of the basilectal management of space by HOMO LOQUENS' (Goebl, 1993, p. 277). Practitioners of dialectometry can be found around the world from Japan to Canada. The numerous research programs remain largely uncoordinated and partly incompatible, but in the next decades the consolidation of methods and goals will undoubtedly crystallize into a powerful branch of dialectology unimaginable in the first century of the discipline.

Sociolinguistic Dialectology

The other revitalizing impulse for dialectology is sociolinguistics, which had its effective beginning in the 1960s with the dissemination of William Labov's fountainhead studies in Martha's Vineyard and New York's Lower East Side (collected in Labov, 1972). Linguists began studying language variation with freer data-gathering methods, using larger and more representative subject samples, and they sought to understand its social significance with quantitative, acoustic and statistical analyses. The intellectual forebears of sociolinguistics, far removed from Linnaean taxonomy, were post-structural linguistics and the social sciences.

Dialect geographers, not surprisingly, did not immediately embrace the new currents, and it took a generation schooled primarily in sociolinguistics to recognize that the two disciplines were both, in essence, dialectologies. With hindsight, the rise of sociolin guistics can be viewed as a natural response to global social forces that have arisen since dialectology came into being, including urbanization, mobility of all three kinds – geographical, occupational and social – embourgeoisement, mass education and mass literacy. Those forces shrank the traditional domains of dialect geographers by reducing geographically isolated regions and socially isolated subjects to statistical insignificance in all the developed nations and beyond.

Observations of linguistic variation became refocused on social class, network, age, sex, ethnicity, and other social subgroups as well as region, and the systematicization of those observations required new sets of independent variables and new methods for tracking them. In the social context that came into existence a century after Warelius and Wenker, the remaking of dialectology as sociolinguistic dialectology was inevitable. As Chambers and Trudgill said, 'A decade or two ago, it might have been possible to think that the common subject matter of dialectology and sociolinguistics counted for next to nothing. Now we know it counts for everything. Dialectology without sociolinguistics at its core is a relic. No serious perspective on dialectology can grant urban research and variation theory less than a central role' (Chambers and Trudgill, 1998, p. 188).

One of the historical consequences of dialectology becoming sociolinguistic is that it has become an integral branch of linguistics. Like other branches of linguistics, its seminal works and formative investigations have been impelled and largely defined by specific issues, that is, by the need to develop or refine or refute particular hypotheses. The defining issues include these (discussed in Chambers, 1993), among others: (1) How do the changes dialectologists observe in progress reflect the changes that occur in the history of languages? (2) How, if at all, do the grand concepts of the sociology of language – concepts such as di glossia, language shift, heteronomy, linguistic legislation and planning – affect social interaction, or become encoded emblematically in sociolinguistic variants? (3) How do dialect features diffuse throughout regions, and how does such diffusion affect the linguistic behavior of people in the community? (4) What is the adaptive function of dialect and accent? This last question, in particular, can be posed as a kind of ultimate question for all dialect studies.

There is a long tradition that presupposes dialect and accent to be dysfunctional, starting with the Babelian myth. Dialectology in all its methodological and intellectual guises shows that, on the contrary, in every human society linguistic variation prevails and thrives. It is a profoundly human attribute that is essentially irrepressible. The large, open-ended questions of dialectology can only begin to find their answers in fine-grained, minute research into the social significance of linguistic

variation. The answers to those questions will ultimately elucidate the human language faculty. With that as its goal, dialectology constitutes a vital branch of modern linguistics.

See also: Language and Ethnicity; Linguistic Fieldwork; Sociolinguistics.

Bibliography

Chambers, J.K., 1993. Sociolinguistic dialectology. In: Preston, D. (Ed.), American Dialect Research. John Benjamins, Amsterdam/Philadelphia, pp. 133–164.

Chambers, J.K., 2000. Region and language variation. English World-Wide 21, 169–199.

Chambers, J.K., Trudgill, P., 1980. Dialectology. Cambridge University Press, Cambridge, UK.

Chambers, J.K., Trudgill, P., 1998. Dialectology, second ed. Cambridge University Press, Cambridge, UK.

Elmer, W., Rubin, E., 1997. The survey of English dialects as an electronic database for research in areal and variationist linguistics. In: Thomas, A.R. (Ed.), Issues and Methods in Dialectology. Department of Linguistics. University of Wales, Bangor, Wales, UK, pp. 234–246.

Gilliéron, J., Edmont, E., 1901–10. Atlas Linguistique de la France. vols. 13. Champion, Paris.

Goebl, H., 1993. Dialectometry: a short overview of the principles and practice of quantitative classification of linguistic atlas data. In: Kohler, R., Rieger, B.B. (Eds.), Contributions to Quantitative Linguistics. Kluwer Academic, Amsterdam, pp. 277–315.

Hall Jr., R.A., 1943. The papal states in Italian linguistic history. Language 19, 125–140.

Heap, D.J., 1997. La Variation grammaticale en géolinguistique: les pronoms sujet en roman central. Department of Linguistics. University of Toronto: Toronto Working Papers in Linguistics.

Jaberg, K., Jud, J., 1928–40. Sprach- und Sachatlas Italiens und der Südschweiz. Ringier. Zolfingen [AIS].

Keyser, S.J., 1963. Review of Kurath and McDavid, Pronunciation of English in the Atlantic States. Language 39, 303–316.

Kretzschmar, W., Schneider, E., Johnson, E. (Eds.), 1989. Computer Methods in Dialectology. Journal of English Linguistics, 22, special edition.

Kurath, H., 1949. Word Geography of the Eastern United States. University of Michigan Press, Ann Arbor, MI.

Labov, W., 1971. Some principles of linguistic methodology. Language in Society 1, 97–120.

Labov, W., 1972. Sociolinguistic Patterns. University of Pennsylvania Press, Philadelphia.

Macaulay, R.K.S., 1985. Linguistic maps: visual aid or abstract art? In: Kirk, J., Sanderson, S., Widdowson, J.D.A. (Eds.), Studies in Linguistic Geography. Croom Helm, London, pp. 172–186.

McDavid, R.I., 1981. On obtaining spontaneous responses. In: Warkentyne, H.J. (Ed.), Papers from the 4th International Conference on Methods in Dialectology. University of Victoria, British Columbia, pp. 66–84.

Orton, H., 1962. Survey of English Dialects: Introduction. E. J. Arnold, Leeds, UK (SED).

Pederson, L., et al., 1986–92. The Linguistic Atlas of the Gulf States. vols. 1–7. University of Georgia Press, Athens, GA.

Pickford, G., 1956. American linguistic geography: a sociological appraisal. Word 12, 211–233.

Ramisch, H., 1997. Dialectological and cartographical features of the computer developed linguistic atlas of England. In: Thomas, A.R. (Ed.), Issues and Methods in Dialectology. DepartmentofLinguistics. UniversityofWales, Bangor, Wales, UK, pp. 224–233.

Rapola, M., 1969. Johdatus Suomen Murteisiin. Suomalaisuuden Kirjallisuuden, Seura.

Sapir, E., 1929. The status of linguistics as a science. Language 5, 207–214.

Séguy, J., 1973. Atlas linguistique de Gascogne. Centre national de la recherche scientifique, Paris. vol. 6.

Weinreich, U., 1954. Is a structural dialectology possible? Word 10, 388–400.

Wenker, G., 1881. Sprachatlas des Deutschen Reichs. MS folio. Marburg University.

Diaspora

Robin Cohen, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article considers the origin and changing use of the notion of diaspora, showing how, despite its specific reference to the history of the Jews, it has now come to be used in understanding wider implications of international migration and transnationalism. How modern nation-states respond to the challenges and opportunities posed by diasporas is discussed. The different types and forms of diaspora, and the diverse ways in which the concept is now deployed are also identified.

Origins and Development

The word 'diaspora' is derived from the Greek word *speiro* (to sow) and the preposition *dia* (over). When applied to humans, the ancient Greeks thought of the word as signifying expansion through settlement. 'Diaspora' soon acquired a more brutal and catastrophic meaning. In the Greek translation of the Old Testament it was used to describe the forcible dispersion of the Jews. The destruction of the Temple in Jerusalem in 586 BC led to the enslavement and displacement of the key military, civil, and religious leaders of Judah and their exile in Babylon. This fate was held to be predicted in Deuteronomy (28: 58–68) where it is stated that people who disobeyed God's law would be scattered to all ends of the earth. There they would 'find no peace,' while they would additionally suffer 'an unquiet mind, dim eyes and a failing appetite.' 'Babylon,' then the capital of the Mesopotamian empire, became the code word for Jews (and later for Africans in the New World) signifying a place of loneliness and estrangement, where they were held in captivity.

This catastrophic meaning of diaspora, which may be described as the 'victim tradition,' was one that informed the historical and popular use of the term for nearly two millennia. Gradually, other groups saw themselves as victims of forcible dispersion in a manner similar to the ancient Jews and used the expression 'diaspora' as a self-description of their fate. Armenians, for example, allude to the period 1894–96 when their national movement was violently opposed by the Ottoman Sultan Hamid and some 300 000 Armenians were killed in Turkish Armenia. In 1915–16 'close to one million people' (Melson, 1992: p. 146), about half the Armenian nation, were either killed or died of starvation. Armenian scholars and activists have accused the Ottoman authorities of genocidal intentions, an interpretation that is generally accepted, although one that has been hotly disputed by successive Turkish governments. The subsequent dispersal of the Armenians to the United States, Iran, France, Lebanon, and other countries is now commonly described as a diasporic experience. The brutal Atlantic slave trade for New World Africans and the potato famine for the Irish are similarly considered to be the crucial traumatic events that triggered the creation of these other victim diasporas. For Palestinians, the birth of Israel is *al-Nakba* (the catastrophe), which impelled their dispersal and exile and continues to fuel their struggles for national self-determination.

The above examples have an apparent theoretical and experiential coherence. A tragedy perpetrated by a hostile exogenous force generates movement to a number of countries. But because the displacement was forced and the traumatic event so encompassing, collective folk memories are developed, the social group remains psychologically unsettled and socially marginalized, and a return movement gains some following. A number of conventional historians would wish to go no further with the idea of diaspora. However, from the 1980s, this relatively well-established meaning began to collapse. Many ethnic groups of varying histories defined themselves as 'diasporas,' or were so recognized by scholars or journalists. The decisive intellectual recognition of this new heterogeneity was marked by the launch of a journal called *Diaspora*, issued by the New York office of Oxford University Press in 1991 and edited by a US-based literary scholar of Armenian provenance. His opening manifesto was an impressive call for openness. The journal would be concerned with "the way in which nations, real but imagined communities, are fabulated, brought into being, made and unmade, in culture and politics, both on the land people call their own and in exile" (Tölölyan, 1991: p. 3). By 1998, the journal had published articles centrally identifying 34 ethnic groups as diasporas. Many hundreds of subnational, national, transnational, and religious groups are now more casually referred to as 'diasporas.'

Types of Diaspora

Natural scholarly caution would invite rejection of this profligacy. Yet, arguably, the pluralists have a good case. The paradigmatic example of the Jews, for example, shows that the communities abroad often arose from voluntary, not forced, movement, while the Babylonian experience can be characterized far more positively – as a place where new creative energies were released and challenging cultural encounters arose over a long period in a largely peaceful setting. Moreover, as has been mentioned, the original Greek meaning of diaspora allowed voluntary movement or even colonial settlement. Drawing on this etymology, it would be consistent, if unconventional, to refer to the British colonies in New Zealand, Australia, and Canada as comprising part of 'the British diaspora' – of an imperial, rather than victim, type.

In addition to the victim and imperial types, scholars have identified labor, trade, business, professional, religious, cultural, deterritorialized, and refugee diasporas, as well as intermediate forms such as incipient, asylum, and 'new' diasporas (Safran, 1991; Sheffer, 1995; McDowell, 1996; Van Hear, 1998; Cohen, 2008). These subtypes describe the predominant characteristic or characteristics of each diaspora, but cannot represent adequately the full complexity of each group in terms of its history, mixed motivations for emigration, or class and regional diversity. It is apparent, for example, that the 50 million 'Chinese overseas,' who have only recently been described as a diaspora, include people from different ethnic and linguistic backgrounds, whose origins lay in semifree 'coolie labor,' as well as those who migrated as entrepreneurs, traders, and skilled professionals.

Despite the new pluralism, there remain key differences between a diaspora and migrants in general. First, cultural, linguistic, religious, historical, and affective ties with the place of origin remain strong. Second, at least in theory, while emigration from 'an old country' involved a one-way ticket, assimilation to the 'new country,' the adoption of a local citizenship and language, and the public acceptance of local ways and customs, these common features of migration do not apply to a diaspora. As studies of a number of ethnic groups in the United States and elsewhere have shown, in culture and endogamous practices the old ways are more resilient than public ideology proclaimed. In these cases, group boundaries endure while strong and complete assimilation does not take place. Third, families and diasporic activists in the different places to which groups have been dispersed effect links to one other, thereby forming a transnational community.

Forms of Diaspora

In addition to typologizing diasporas by examining their broad characteristics, scholars have sought to classify the different 'forms' in which diasporas appear and how these map on to the understanding of the concept in different academic disciplines. A generally successful fourfold scheme has been proposed by Vertovec (1997) and modified in Vertovec and Cohen (1999). A fifth form has been added below:

1. The first category, usually advanced by economists and sociologists, is 'diaspora as social form,' alluding to the maintenance and development of transnational social organizations and networks, a collective identity and particular kinds of economic strategies. These strategies are closely linked to the idea of social capital, which allows those with a network of trust and social cohesion to fund investments, support enterprise, and promote collective social mobility through education and other means (Kotkin, 1992; Cohen, 2008).
2. The second form is 'diaspora as consciousness.' Cultural studies theorists as well as those working in the fields of ethnic and race relations found in the notion of diaspora a means of breaking free from the 'essentialist' discourse that previously informed the subject matter. Notions of what constituted ethnicity became more fluid, hybrid, and negotiated, and 'diaspora' was a way of representing how complexity and difference arouse as cultures traveled, interacted, and mutated (Clifford, 1992; Brah, 1996).
3. The third is 'diaspora as mode of cultural production.' Although closely related to the second, this form is used by anthropologists to show how social and cultural formations, as well as cultural objects, images, and meanings have leaped from their old locales and crossed national lines (Hannerz, 1996; Basch et al., 1994).
4. Fourth, 'diaspora as political orientation' has been considered by political scientists and international relations theorists. As noted below, such specialists concentrate on how the attachment to the politics of homeland can negatively and positively impact on the countries at both ends of the migration chain.
5. We need now to add a fifth form, namely, 'e-diasporas as connected migrants,' a notion advanced by digital specialists. Diminescu (2012: p. 452) argues that an e-diaspora "organizes itself and is active first and foremost on the Web." In so doing, she argues, an unstable collective emerges (as people switch on or exit) but, despite this constant renegotiation of the extent and nature of an e-diaspora, the connection is sufficiently dense and meaningful to disrupt the traditional rupture between 'home' and 'away' and between the different points (or edge nodes) of settlement.

Diasporas and Public Policy

For many countries, the goal of full integration was at the heart of the nation-building project. During the period of intense nation-building in the core European, South, and North American states a diasporic consciousness was either not sanctioned or firmly suppressed. The US state's motto 'Ex Pluribus Unum' provided an example of the prevailing preference for assimilation or the creation of a new national culture. However, under the impact of diverse and accelerating immigration a number of important nation states (Canada, the United States, the Netherlands, Australia, and the United Kingdom might be cited) lost their appetite and perhaps the capacity to assimilate immigrants to a single dominant political culture. By recognizing ideas of multicultural, multifaith, and multiracial societies, their political leaderships reversed their previous disapprobation of the retention of preimmigration ethnic identities.

Many policy makers and social commentators are relaxed about this development and see no incompatibility between a national and an ethnic affiliation. Others, including those from the progressive side of the political spectrum, are anxious that the painfully achieved democratic and secular order in societies such as the United States will be undermined by the recognition of diasporas. For the liberal historian, Arthur Schlesinger Jr. (cited Hu-DeHart, 1993: pp. 7–8), "The US escaped the divisiveness of a multi-ethnic society by a brilliant solution: the creation of a brand new identity. The point of America was not to preserve old cultures but to forge a new, American culture." Diasporas are seen as the reason why this idea has not been fully realized. For some commentators, diasporas foster religious zealotry, the international drugs

trade, political instability, and cross-border terrorism. The involvement of exile groups, like the Tamils, Kurds, Sikhs, or Afghans, in the political life of their homelands or in movements to create nation-states, has often created violent conflicts or embarrassing and difficult diplomatic and legal conflicts (Sheffer, 1986). By contrast, Shain (1999) argues that the democratic and liberal ways learned in the place of settlement are more likely to spread back into the countries of origin, with benign political effects for all parties.

The policies of the countries of origin to their diasporic communities abroad has only recently gained systematic discussion by diaspora scholars, in recognition that these states are increasingly important players in retaining and amplifying the links between 'home' and 'away.' Historically, countries like China and India were hostile to their overseas communities, seeing them as painful reminders of colonial days when their people could be seduced by labor recruiters offering indentures on European-run tropical plantations or, in the contemporary period, as 'brain-drain' migrants. India, China, and many other migrant-exporting countries now recognize that their diasporas can provide an important political lobby for advancing their countries' interests, a bridgehead for economic marketing of the home countries' goods, a means for acquiring transferable skills (e.g., in information technology) and, above all, a rich source for remittances and investment income.

So important have diasporas become to origin states that many have initiated a battery of initiatives to establish, reestablish, or maintain ties to homelands. As Gamlen (2008: pp. 840–856) shows, these include recognizing diaspora associations or founding them, granting a variety of civil, social, and political rights to nationals abroad, issuing specialized passports, allowing or encouraging dual nationality, providing emergency consular support, creating incentives to invest 'at home' and permitting tax holidays for the remittance of income and profits. Many countries have now created ministries, departments, or other units to service the needs of their diasporas and represent their concerns to other government departments. Examples of such units are the Institute for Mexicans Abroad, the Irish Abroad Unit, and Delhi's Ministry of Indian Overseas Affairs. The Philippines also has an expertly run, state-supported set of agencies (notably the Commission on Filipinos Overseas) that promote labor export on a vast scale, help to channel return income, and seek to protect their workers from exploitative employers abroad. The Commission additionally sponsors a 'Global Summit of Filipinos in the diaspora,' where 'D2D' (Diaspora to Development) initiatives in the arts, culture, education, philanthropy, medicine, and agriculture are encouraged and sustained.

Future of the Concept

The concept of diaspora has decisively escaped its historical confines in the Jewish experience and is now used in a variety of settings and for a variety of purposes. The enduring logic is that the term conceptually connects home and abroad. The conflicting claims on group loyalties have been given an added emphasis in an age of global networks and the changing role, some would argue the relative decline, of the nation-state as a focus for group identification. 'Diaspora' will remain as an important means of interrogating and exemplifying this tension. It will also provide a useful means of describing the enhanced role for cyclical, transversal, and intermittent international migration. The arrival of the Internet will both enhance the historical ties found within diasporas and, to some degree, change their character. A striking possibility is that e-diasporas might create new virtual homes, without the need for a territorial homeland.

Beyond these significant changes are other, more adventurous uses of the expression. For example, the use of 'diaspora' to describe all transnational religious communities and nationalities abroad is not always apposite. In some cases, such as the Jews, Parsis, and Sikhs, religion, ethnicity, minority status, and peoplehood are so closely interpenetrated that the description 'diaspora' is appropriate. However, 'Muslim diaspora,' which yields many hits on a Google search, undervalues the ethnic and national diversity concealed by the expression. Is the *Ummah*, one may further ask, coterminous with 'a Muslim diaspora'? Again, 'Indian diaspora,' an expression that is now in common and official use, elides the differences between the three principal faith communities (Hindu, Muslim, and Sikh) and ignores well-organized national minorities (like the Kashmiris), with potentially confusing effects. In addressing such problems, there may be some virtue in using the wider expression 'transnational communities' to include a number of different postnational and nonnational forms of sociality and association. Examples so categorized can be virtual Internet communities, business networks, Christianity, Islam, transnational social movements, and international professional associations. Within the rubric of transnational communities can shelter 'diaspora' as a particular form of ethnically defined community.

Despite this suggestion for self-limiting the term, it is clear that a number of scholars, particularly in the humanities, have abandoned the confines of ethnicity in their enthusiasm for the expression 'diaspora.' At the heart of this appropriation of the term is the idea that social marginalization (seen earlier in the failure of ethnic diasporas to fully assimilate) is a sufficient precondition for a formation of a diaspora. Motley and Henderson (2008) argue the global hip-hop movement is a diaspora as it transcends ethnic, linguistic, and geographical boundaries, while Patten and Sánchez-Eppler (2000) suggest that queer diasporas are formed as people move across literal and figurative boundaries. These are but two examples of the argument that diasporas are formed by real or symbolic marginality, chosen lifestyles, or mutually recognized cultural preferences. It is common ground that the Internet has facilitated some degree of cohesion within such groups.

There is little doubt that the elastic band holding the meaning of 'diaspora' together is getting slacker and slacker. As Brubaker (2005) has suggested, this can be seen as a conceptual dispersal, a movement into uncharted semantic, conceptual, and disciplinary spaces. He amusingly calls this tendency a 'diaspora diaspora.' To avoid this attenuation of meaning, some degree of self-restraint may prove necessary if the term is to retain its conceptual purchase. In other words, the future of the term may be imperiled by having been commandeered to serve too many discrepant purposes and its use to describe cognate, yet importantly different, phenomena.

See also: Assimilation of Immigrants; Creolization: Sociocultural Aspects; Hybridity; Jewish Diaspora; Migrations, Colonizations, and Diasporas in Archaeology; Multiculturalism.

Bibliography

Basch, L., Glick Schiller, N., Szanton Blanc, C., 1994. Nations Unbound: Transnational Projects, Post-colonial Predicaments, and Deterritorialized Nation-states. Gordon & Breach, Langhorne, PA.

Brah, A., 1996. Cartographies of Diaspora. Routledge, London.

Clifford, J., 1992. Traveling cultures. In: Grossberg, L., Nelson, C., Treichler, P.A. (Eds.), Cultural Studies. Routledge, New York, pp. 96–116.

Cohen, R., 2008. Global Diasporas: An Introduction. Routledge, London.

Diminescu, D., 2012. Introduction: digital methods for the exploration, analysis and mapping of e-diasporas. Social Science Information 51 (4), 452.

Gamlen, A., 2008. The emigration state and the modern geopolitical imagination. Political Geography 27 (8), 840–856.

Hannerz, U., 1996. Transnational Connections: Culture, People, Places. Routledge, London.

Hu-DeHart, E., 1993. Rethinking America: the practice and politics of multiculturalism in higher education. In: Thompson, B.W., Tyagi, S. (Eds.), Beyond a Dream Deferred: Multicultural Education and the Politics of Excellence. University of Minnesota Press, Minneapolis, pp. 3–17.

Kotkin, J., 1992. Tribes: How Race, Religion and Identity Determine Success in the New Global Economy. Random House, New York.

McDowell, C., 1996. A Tamil Asylum Diaspora. Berghahn, Oxford, UK.

Melson, R., 1992. Revolution and Genocide: On the Origins of the Armenian Genocide and the Holocaust. University of Chicago Press, Chicago.

Motley, C.M., Henderson, G.R., 2008. The global hip-hop diaspora: understanding the culture. Journal of Business Research 61, 243–253.

Patten, C., Sánchez-Eppler, B. (Eds.), 2000. Queer Diaporas. Duke University Press, Durham, NC.

Safran, W., 1991. Diasporas in modern societies: myths of homeland and return. Diaspora 1 (1), 83–99.

Shain, Y., 1999. Marketing the American Creed Abroad: Diasporas in US and Their Homelands. Cambridge University Press, Cambridge, UK.

Sheffer, G. (Ed.), 1986. Modern Diasporas in International Politics. Croom Helm, London.

Sheffer, G., 1995. The emergence of new ethno-national diasporas. Migration 28, 5–28.

Tölölyan, K., 1991. Preface. Diaspora 1 (1), 3–7.

Van Hear, N., 1998. New Diasporas: The Mass Exodus, Dispersal and Regrouping of Migrant Communities. University of Washington Press, Seattle, WA.

Vertovec, S., 1997. Three meanings of 'diaspora' exemplified among South Asian religions. Diaspora 6 (3), 277–299.

Vertovec, S., Cohen, R. (Eds.), 1999. Migration, Diasporas and Transnationalism. Edward Elgar, Cheltenham, UK.

Brubaker, R., 2005. The 'diaspora' diaspora. Ethnic and Racial Studies 28 (1), 1–19.

Dichotic Listening and Language: Overview

Kenneth Hugdahl, University of Bergen, Bergen, Norway; and Haukeland University Hospital, Bergen, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, I provide a selective review of the history of dichotic listening and its use as an application in neuropsychology, both with regard to research and clinical work. The review traces the concept of presenting simple speech sounds, such as syllables or numbers dichotically, from its original use by Broadbent in studies of attention in the early 1950s, to the pioneering work by Kimura a decade later in her studies of hemispheric asymmetry, and to the present-day use of the dichotic listening procedure in research and clinical practice. The article reviews studies on both healthy individuals and psychiatric and neurological populations. In addition to behavioral studies, dichotic listening has also been used in functional neuroimaging studies to reveal the underlying neuronal circuitry of language at the brain systems level.

Brief History of Dichotic Listening Research

The British psychologist Donald Broadbent is credited as having been the first to systematically use a dichotic setup in his studies of how air traffic controllers managed to maintain attentional focus when they had to process flight bearings from several airplanes at the same time (Broadbent, 1956; see also Bryden, 1988; Kimura, 2011; Hugdahl, 2011 for historical overviews of dichotic listening research). Broadbent used a dichotic procedure with different auditory messages being sent to the right and left ear, requiring the subject to focus attention on either the left or right-ear message, as a way to simulate the real-life situation facing the air traffic controllers. It should be noted, however, that although Broadbent used a dichotic setup, he did not systematically investigate whether there was an asymmetry in how subjects were able to process the different messages in the right or left ear. In fact, if there was any interest in dichotic listening outside of an attentional framework it was not in asymmetry of the brain, but rather implications for the understanding of short-term memory. Bryden (1988) writes that Donald Hebb had suggested that Broadbent's findings of subjects' tendency to first report the set of stimuli in one ear and then proceed to the other ear, must involve some kind of short-term memory mechanism. The discovery that dichotic presentations of verbal stimuli produced an increased number of correct reports for items presented in the right ear is credited to Doreen Kimura, who in a seminal paper in the Canadian Journal of Psychology (1961a) found that patients with temporal-lobe pathology reported more items from the right ear than from the left ear. She followed up the 1961a findings on patients with a study on right-handed neurologically intact individuals published the same year (Kimura, 1961b; in fact the two papers appeared back-to-back in the same issue). Again the results showed that there were more correctly reported items from the right ear than from the left ear. In a personal recollection of the pioneering 1961-studies, Kimura (2011) writes: "I knew from animal research that the crossed pathways from the ear to auditory cortex predominate over the uncrossed, and in fact might occlude the uncrossed input where there was overlap (Rosenzweig, 1951; Tunturi, 1946). I concluded that the right-ear effect was due to the fact that in people also, the crossed auditory pathways were more effective than the uncrossed. This gave the right-ear input an advantage in accessing areas in the left hemisphere critical for speech perception" (p. 214). **Figure 1** shows a drawing of the crossing of the auditory pathways and the underlying neuroanatomical basis for the right-ear effect (adapted from Kimura, 2011; originally published in Kimura, 1967). The original electrophysiology work by Rosenzweig (1951) and the anatomical model proposed by Kimura in 1967 to explain the REA were later corroborated in both lesion studies (e.g., Sparks and Geschwind, 1968; Pollmann et al., 2002) and with advanced electrophysiology (e.g., Brancucci et al., 2005) and finally with neuroimaging techniques (Hugdahl et al., 1999). The study by Sparks and Geschwind (1968) was the first study to also show the importance of the corpus callosum for the REA. They tested a patient with commissurotomy in which the two

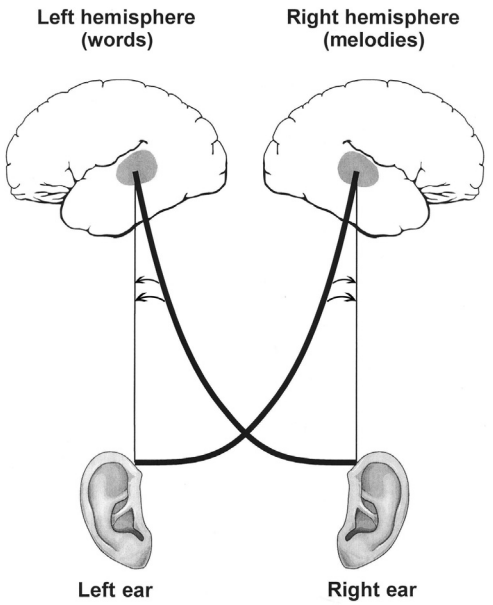

Figure 1 Schematic presentation of the auditory pathways and the neuroanatomy of the dichotic stimulus presentation as originally published in Kimura (1967). Reproduced from Kimura, D., 2011. From ear to brain. Brain and Cognition 76, 214–217, with permission from the publisher.

hemispheres were surgically separated by sectioning the corpus callosum. This patient showed an almost total REA, and a corresponding left-ear extinction effect, which then demonstrated the direct pathways for the right-ear stimulus and the indirect (callosal) pathways for the left-ear stimulus in the dichotic situation. This was later corroborated by Westerhausen et al. (2008) using diffusion-tensor imaging (DTI) technique. Others have, however, claimed that a neuroanatomical model does not take into account cognitive factors, e.g., attention that may also contribute to the REA (e.g., Hiscock and Kinsbourne, 2011; see also below under top-down modulations of the REA).

The right-ear effect that Kimura mentioned was called a 'right-ear advantage (REA)' by Shankweiler and Studdert-Kennedy (1967), and later abbreviated REA (e.g., Bryden et al., 1983; Hugdahl and Andersson, 1984). The next researcher to take up the thread of the right-ear effect was P.M. Bryden who suggested that the right-ear effect observed by Kimura could have been caused by an 'order-effect' since subjects were instructed to report the right-ear items first. He therefore counterbalanced the order of presentation, but the REA still prevailed despite controlling for any 'order-effects.' Bryden (1963) concluded that "The findings suggest that the auditory system is better organized for the perception of verbal material presented to the right ear" (Abstract, p. 103).

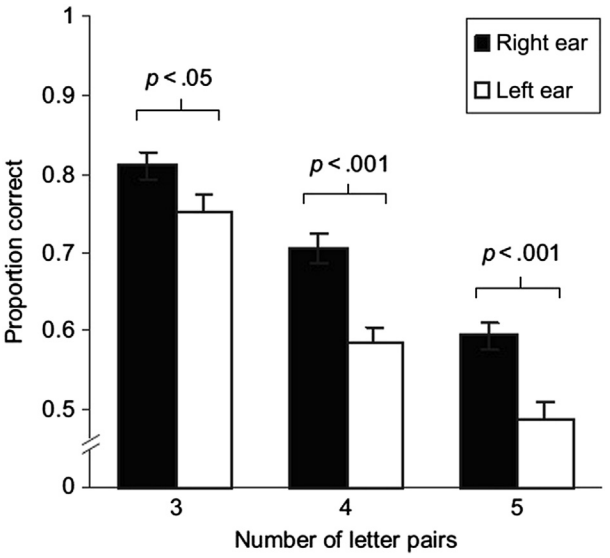

Figure 2 Results for the effect of working memory load on the right-ear advantage, as a function of increasing the number of letter pairs to be held in the working memory buffer. From Penner, I., Schlaefli, K., Opwis, K., Hugdahl, K., 2009. The role of working memory in dichotic-listening studies of auditory laterality. Journal of Clinical and Experimental Neuropsychology 31, 956–966, reproduced with permission from the publisher.

Asymmetry for Working Memory

It should be noted, however, that the original studies by Kimura and Bryden were not exclusively about laterality for speech perception, but also contained elements of working memory. This was evident from the stimulus paradigms used in these studies. In both the Kimura and Bryden original studies, series of 3, 4, or 5 pairs of digits were presented dichotically, i.e., one digit in the right ear and another digit simultaneously in the other ear. The instruction to the subject was that he/she had to later recall which digits had been presented. Thus, the situation was to first listen to 6, 8, or 10 digits in sequence, and immediately after report as many digits that was possible, while the experimenter noted how many were reported that had been presented in the right ear and how many were reported that had been presented in the left ear. However, this situation requires the subject to keep the nonreported, but recalled, items in the working memory buffer for its turn to be reported. Thus, the original studies also pointed to the fact that working memory for verbal stimuli interacted with the perception of the same stimuli to produce a laterality effect. Much later, Penner et al. (2009) explicitly studied the interaction of laterality for speech sound perception and working memory by having subjects to listen to three-, four-, or five-letter pairs i.e., using the same working memory load as in the original studies, but with different stimuli (spoken letters rather than spoken digits). The results are shown in **Figure 2**, which reveals that as the working memory load increased from three to four stimulus-pairs the REA also increased, while it leveled off when the working memory load increased further to five-letter pairs. The results in the study by Penner et al. (2009) followed the original working memory effect in dichotic listening reported by Bryden (1963) when controlling for the order of report from either the right or left ear.

The Consonant-Vowel Syllables Paradigm

In 1967, Donald Shankweiler and Michael Studdert-Kennedy at the Haskins Laboratories (see also Studdert-Kennedy and Shankweiler, 1970) asked the question at which level of speech processing the asymmetry the right-ear effect occurred, and if the REA was modified by the phonetic features of the stimuli. To be able to study these questions empirically they introduced a new paradigm, the consonant-vowel (CV) syllables paradigm. This paradigm consisted of pairwise presentations of CV-syllables based on the six stop-consonants /b/, /d/, /g/, /p/, /t/, /k/ that were paired with the vowels /a/, /e/, /i/, /æ/, and /u/. With this paradigm, it was possible to study the relative contribution of the consonants and vowels separately, by analyzing the different phonetic contrasts made up of combinations of the consonants and vowels. The instruction to the subject was to verbally report which syllable he/she had perceived after each trial. Shankweiler and Studdert-Kennedy (1967) included 15 right-handed members of the laboratory staff, and included also a test with only the steady-state vowels that were presented dichotically, a different vowel in each ear. A significant REA was observed only for the consonants, but not for the vowels (later replicated by Hugdahl & Andersson, 1984), indicating that the REA may be more of a perceptual effect for the identification of the initial phonological information in a syllable. Shankweiler and Studdert-Kennedy (1967) summarized the strong laterality effects they observed for the superiority for the right-ear items such that: "This strongly suggests that left hemisphere dominance in speech perception operates at the level of speech sound structure, [and that] the effect can be demonstrated when only a single pair of

syllables is presented on each trial indicating that it pertains to the registration of the stimuli and not only their retention." (p. 62). Thus, the study by Shankweiler and Studdert-Kennedy (1967) set the record straight that the REA was not a memory retention artifact, but a direct consequence of the phonetic features of the speech sound. Moreover, the effect was based on the energy release of the initial consonant segment of the CV-syllable, but relatively insensitive to the manipulation of the following vowel sound. The CV-syllables paradigm soon became very popular in research on hemispheric asymmetry and a search through PubMed and ISI Web of Knowledge data bases show that more than 2000 articles have used a CV approach in studies of auditory laterality. A paradigm that has been as popular as the CV-syllables paradigm is the fused-rhymed paradigm (Wexler and Halwes, 1983), which is similar to the CV-syllables paradigm except for the important distinction that the two sounds presented at the ears fuse into a coherent perceptual unit. A characteristic of the fused-rhymed paradigm is that although subjects subjectively report that they hear 'only one sound,' they report this to be the right-ear item.

Before proceeding with other studies that followed in the wake of the pioneering studies by Kimura, Bryden, and Shankweiler and Studdert-Kennedy, two factors that have been related to hemispheric asymmetry and dichotic listening should be mentioned, namely effects of handedness and sex differences.

The Effect of Handedness

The CV distinction as reported by Shankweiler and Studdert-Kennedy (1967; Studdert-Kennedy and Shankweiler, 1970) was later corroborated by Hugdahl and Andersson (1984) who contrasted the six stop-consonants with the vowels /a/, /i/, /u/. When analyzing the errors made for the different consonants and vowels respectively, they found significant differences for the consonants, depending on the ear of presentation, but not for the vowels. Interestingly, Hugdahl and Andersson (1984) also included a group of left handers, and found a reduction of the REA in the left-handed group, measured as fewer errors for the left-ear input compared to the left-ear input. Hugdahl and Andersson (1984) concluded that the CV-syllables dichotic listening paradigm "allows for the separation of left- and right-hemisphere dominance for right- and left-handed subjects" (p. 141). Later research has however concluded that the CV-syllables dichotic listening paradigm does not discriminate right- from left handers (e.g., Sommer, 2010 for a recent review), which would be expected considering that about 70% of left handers have language in the left hemisphere equal to right handers (Satz, 1973; Springer and Deutsch, 1989).

The Effect of Sex

Studdert-Kennedy and Shankweiler did not report on the sex of their subjects, only that they were right handed, and Hugdahl and Andersson (1984) only included males. Thus, it was not clear from the early studies if sex had an effect on performance on the CV-syllables paradigm and on hemispheric asymmetry in general. Although several early studies on hemispheric asymmetry and laterality, using different tasks and paradigms, both auditory and visual, had shown that males were superior on visuospatial tasks, while females were superior on verbal tasks (e.g., Harris, 1978; Levy, 1971), there were also studies failing to report significant sex differences (e.g., Hannay and Boyer, 1978; see also McGlone, 1980 for review). The first systematic review of sex differences on tasks measuring hemispheric asymmetry (Voyer, 1996) showed modest overall effects of sex, accounting for a small proportion of the variance. When it comes to sex differences in performance on a dichotic listening task, the data essentially show the same picture, the sex difference is either marginal or absent. Two recent meta-analyses (Voyer, 2011; Sommer et al., 2008; Hiscock and MacKay, 1985) both showed that sex differences in dichotic listening performance accounted for only about 1% of the total variance, with no significant differences between the sexes for the REA. These findings were recently followed up by Hirnstein et al. (2012) in the largest sample of males and females so far ($N = 1782$) where they analyzed the ear advantage scores separately for males and females, including children, young adults, and old adults. The results showed that sex differences in dichotic listening are age-dependent such that male young adults had a greater REA than young female adults. The results from the Hirnstein et al. study are shown in **Figure 3**.

Handbook of Dichotic Listening, 1988

After the initial studies which showed that the REA was a robust empirical phenomenon, replicable in many contexts and situations, there was a surge of studies applying dichotic listening to almost every phenomenon thinkable in research on hemispheric asymmetry. This produced a well of interesting findings, summarized in the volume *Handbook of dichotic listening*, Edited by Kenneth Hugdahl, and Published by Wiley and Sons, UK in 1988. These studies also created an optimism that dichotic listening could provide answers to outstanding questions regarding how the two hemispheres of the brain functioned, not the least in the clinical domain. For example, it was believed that dichotic listening could replace invasive methods, like the Sodium Amytal test (Wada and Rasmussen, 1960) for the identification of side of language in patients with epilepsy undergoing surgical treatment.

Clinical and Developmental Applications

The different chapters in the Handbook of dichotic listening (Hugdahl, 1988) showed that dichotic listening at that time had diversified into understanding not only the functioning of the two cerebral hemispheres when it came to laterality for speech and speech perception, but also applications to a wide range of disorders where these functions would be expected to be compromised. This was particularly true for studies of children with various forms of language disorders, like dyslexia (Bakker and Kappers, 1988), learning and reading disability (Obrzut and Boliek, 1988), and cognitive development in children (Hiscock and Decter, 1988). There were also chapters on the application of dichotic listening to psychiatric and

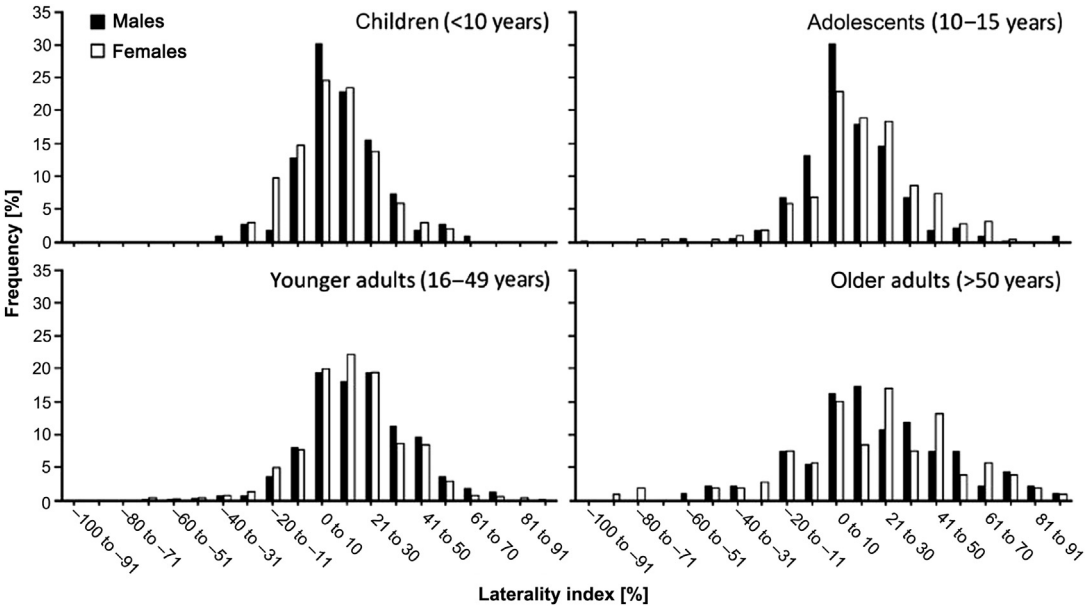

Figure 3 Effects of sex on the right-ear advantage (REA) in dichotic listening based on a large sample of healthy individuals ($N > 1780$). The figure shows the frequency of males and females split for different age groups and for magnitude of the laterality index (LI). An LI of '0' indicates no difference for the right- and left-ear scores, an LI of +100 indicates 100% REA and an LI of −100 indicates 100% left-ear advantage. From Hirnstein, M., Westerhausen, R., Kosrnes, M.S., Hugdahl, K., 2012. Sex differences in language asymmetry are age-dependent and small: a large-scale, consonant- vowel dichotic listening study with behavioral and FMRI data. Cortex, Epub ahead of publication, with permission from the publisher.

neurological disorders, like depression (Bruder, 1988), schizophrenia (Nachshon, 1988), Sodium Amytal studies of patients with epilepsy (Strauss, 1988), split-brain patients (Sidtis, 1988), and correlates with brain lesion findings (Eslinger and Damasio, 1988). A general conclusion from reading these applications was that dichotic listening was a valuable noninvasive measure of abnormality of the laterality of speech. However, these early clinical findings also prompted a surge of interest in both hemispheric asymmetry in general and dichotic listening in particular, with very high expectations of what a noninvasive measure could achieve of data. When dichotic listening results did not completely match invasive data, as e.g., when matching the size of REA with aphasia data, people started losing interest in the field, not considering that no neuropsychological measure or test ever match lesion data on that level of matching.

Bottom-Up Modulation of the REA

An important question that was asked after the initial studies was the degree to which it could be, and was, modulated by other factors than laterality, and how hemispheric asymmetry interacted with those factors to modulate the REA. The modulation of the REA can be studied from two perspectives, one is the modulation as a consequence of the characteristic of the subject being tested, e.g., whether children show different degrees of the REA than adults, or if certain disorders affect the REA in a specific way, e.g., if dyslexic individuals perform different on a dichotic listening task than normal-reading individuals. The other perspective on how the REA is modulated is as a consequence of the stimulus characteristics, e.g., voicing and timing of the syllable sounds that are presented to the subject. With regard to the first kind of modulation, Hugdahl et al. (2001) showed that the REA is present at an early age, actually before reading, and is as such a feature of hemispheric asymmetry for language in a very basic sense. When it comes to studies of dyslexia and reading disorders, Obrzut and colleagues (see Obrzut and Boliek, 1988; Obrzut and Mahoney, 2011) used a CV-syllable dichotic listening paradigm to study the phonetic aspects of reading impairment in children with dyslexia and other language-related problems. This is in line with what Obrzut and Boliek already in 1986 reported of a reduction of the REA, or a tendency for a left-ear advantage (LEA) in learning- and reading-disabled children, see also Satz and Sparrow (1970) for an even earlier study in reading-disabled children. Such findings would then support a notion that reading disability is related to impaired phonological processing and phonological awareness (Høien and Lundberg, 2000), or reduced ability for low-level auditory discrimination (cf Tallal and Piercy, 1973). Another finding was from studies being conducted by Bakker and his colleagues in the Netherlands (e.g., Bakker, 1970; Bakker and Kappers, 1988), which showed that the development of the ear advantage and the REA was delayed in dyslexic children compared to normal-reading children. In still another series of studies Cohen et al. (1992) found that the REA was modulated not only by reading impairment and dyslexia in general, but children with auditory as compared to visual dyslexia showed greater reduction of the REA. A later study by Helland et al. (2007) found that the REA in dyslexic children was also modulated by how severe the disorder is, such that children

diagnosed with a reading disorder requiring special education training showed a greater reduction in the REA than children following regular classes. See **Figure 4**, for example, of the effect of how the need for special education training affects the REA. As can be seen in **Figure 4**, children with a dyslexia diagnosis and in need of special education training compared to children where such training was not considered necessary revealed a significant reduction of the REA. The REA may therefore have clinical value as a complementary diagnostic tool.

All in all, studies of dyslexia and reading impairment have shown that children with reading impairment and dyslexia show an aberrant laterality displayed as reduced phonetic discrimination and abnormal REA.

With regard to modulations of the REA as a consequence of the stimulus characteristics, Berlin and colleagues reported a series of intriguing studies in the 1970s (summarized in Berlin, 1977) on the effects of amplitude (intensity) and phase (time) shifts of the dichotic signal in the right versus the left ear. Berlin manipulated the two signals of the dichotic pair by increasing the intensity of, e.g., the right-ear syllable relative to the left, and vice versa, and noted how performance on the task varied. He also varied the relative timing of the onset of the two signals such that either the right or the left syllable trailed the other by a few ms, and noted how time differences affected the REA. The results by Berlin et al. were that when the right-ear signal lagged behind the left-ear signal by 15–30 ms, the REA was increased, and vice versa if the left-ear signal lagged behind the right-ear signal, thus showing a kind of feed-forward effect such that the signal arriving later had a greater effect on the ear advantage. From these studies, it is clear that the REA is modified and modulated by both amplitude and phase shifts of the two syllables in the dichotic pair. The early studies by Berlin and colleagues have later been followed up by Hugdahl and Westerhausen (e.g., Hugdahl et al., 2008; Westerhausen et al., 2010; Falkenberg et al., 2011) where they gradually increased the intensity of the syllable in either the right or left ear in steps of 3 dB at a time (Hugdahl et al., 2008). The results are shown in **Figure 5**, and revealed a linear increase in the magnitude of the REA when the right-ear signal was increased, and a corresponding linear decrease when the left-ear signal intensity was increased.

What is interesting from **Figure 5** is that the REA was shifted to LEA when the intensity of the left-ear signal was increased 6 dB above the level of the right-ear signal. Thus, the REA could therefore be said to withstand a sound pressure intensity difference of 6 dB, making it possible to express a cognitive concept (REA) in physical terms (dB). In later experiments, the same research group has found that the REA withstands up to a 9 dB difference favoring the left-ear signal, depending on the context in which the test is conducted.

A third kind of stimulus manipulation is voice-onset-time (VOT), also called voicing, which means comparing syllable pairs with voiced (/ba/,/da/,/ga/) versus unvoiced (/pa/,/ta/, ka/) syllables, corresponding to short (~20 ms) versus long (~70 ms) VOTs. The REA is sensitive to the VOT features of the dichotic pair such that having an unvoiced syllable in the right ear and a voiced syllable in the left ear produces a stronger REA than having equally voiced syllables in either ear. Interestingly, a voiced syllable in the right ear and an unvoiced syllable in the left ear produce LEA in adults (Rimol et al., 2006). Thus, it is obvious that the REA is modulated by both amplitude and phase shifts of the signals making up the dichotic pair. Özgören et al. (2012) put the two features together to investigate the relative contribution of amplitude- and phase shifts when the two features are in conflict, such as when increasing the intensity of the right-ear signal, but having the left-ear

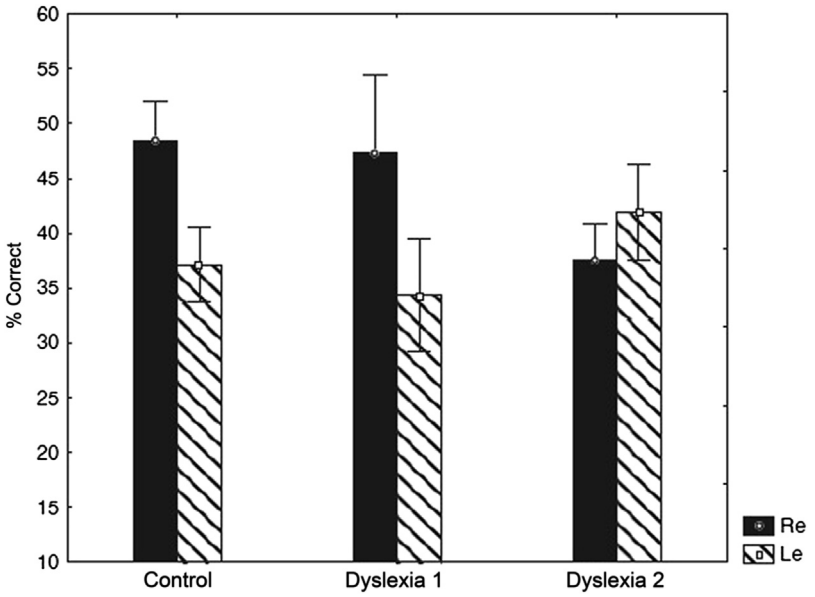

Figure 4 Difference in direction and magnitude of the ear advantage in dichotic listening between a normal-reading control group (control), a dyslexia group with no further assessment, training (dyslexia 1), and a dyslexia group, which was further referred by the local psychological school services to a specialized regional resource center (dyslexia 2). Small bars indicate SD. From Helland, T., Asbjørnsen, A., Hushovd, A.E., Hugdahl, K., 2007. Dichotic listening and school performance in dyslexia. Dyslexia 14, 42–53, reproduced with permission from the publisher.

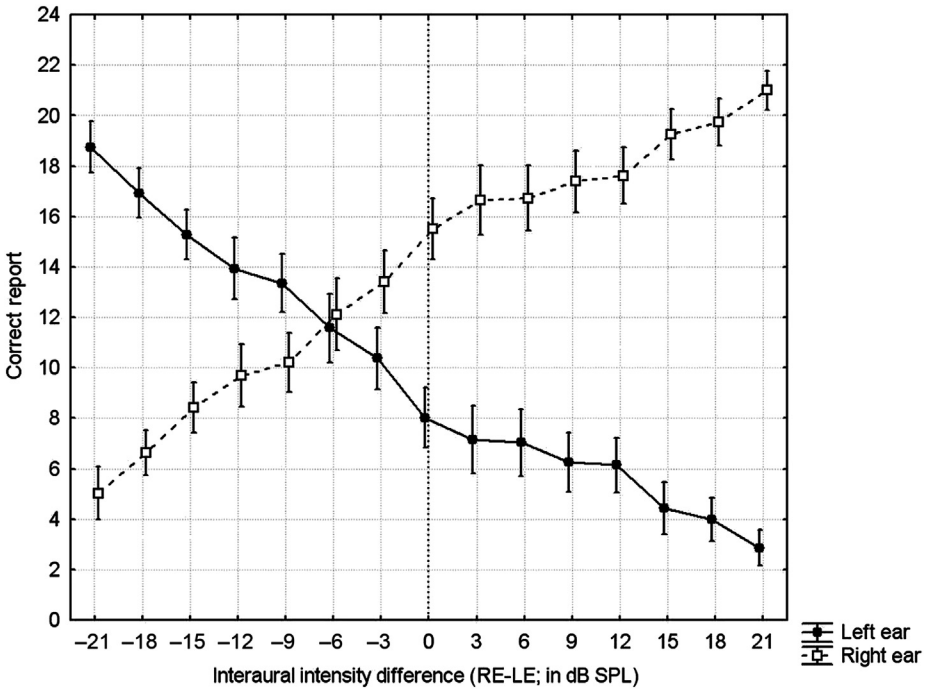

Figure 5 Effects of systematic variation of the intensity in one ear relative to the other ear in steps of 3 dB at a time. '0' on the *x*-axis indicates equal intensity for the right and left-ear signal. +21 means an increase of 21 dB of the signal in the right ear, −21 means an increase of 21 dB of the signal in the left ear. From Hugdahl, K., Westerhausen, R., Alho, K., Medvedev, S., Hämäläinen, H., 2008. The effect of stimulus intensity on the right ear advantage in dichotic listening. Neuroscience Letters 431, 90–94, reproduced with permission from the publisher.

signal presented 10 ms before the right-ear signal. Overall, the results of the Özgören et al. (2012) study showed that amplitude shifts favoring the right ear had a greater effect on the REA than corresponding phase shifts, while phase shifts favoring the left ear were found to have a greater effect on the LEA than corresponding amplitude shifts. In addition, it was found that phase shifts favoring the left ear had a greater effect on the LEA than corresponding phase shifts favoring the right ear had on the REA. The results of simultaneously manipulating amplitude and phase of the syllables in the dichotic pair is seen in **Figure 6**.

In contrast to the findings by Berlin and colleagues, Özgören et al. (2012) found that the REA was increased for syllable pairs where the right-ear syllable had an earlier onset, and vice versa for the LEA to syllable pairs where the left-ear syllable had an earlier onset. Recall that Berlin et al. (reviewed in Berlin, 1977) found that the REA was increased to syllable pairs where the right-ear syllable trailed the left-ear syllable,

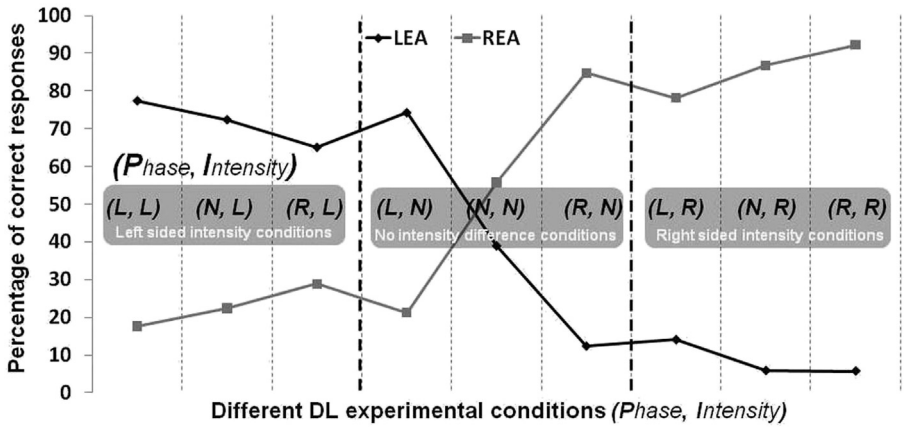

Figure 6 Effects on the right-ear advantage (REA) by simultaneously manipulating intensity (amplitude) and time (phase) of signal arrival to the ear. L,L = phase favoring left ear, intensity favoring left ear, N,L = no intensity difference, phase favoring left ear, R,L = phase favoring right ear, intensity favoring left ear, etc., for the abbreviations. From Özgören, M., Bayazit, O., Oniz, A., Hugdahl, K., 2012. Amplitude and phase-shift effects on dichotic listening performance. International Journal of Audiology 51, 591–596 with permission from the publisher.

and vice versa. The difference in results could be due to that the subject in previous studies was instructed to report both syllables in the dichotic pair, while Özgören et al. (2012) instructed their subjects to only report one syllable of the pair (the one perceived most clearly).

Top-Down Modulation of the REA

A completely different kind of modulation of the REA is from a top-down, or cognitive, perspective, e.g., when instructing the subject to focus attention to either the right- or left-ear stimulus of the dichotic pair (Bryden et al., 1983; Hugdahl and Andersson, 1986), or have another cue-stimulus presented in either the right or left ear a few hundred ms before the CV-syllables are presented, and instruct the subject to report the syllable in the same ear as the cue was presented in (Mondor and Bryden, 1991). In both of these situations, the aim is to induce a top-down cognitive strategy to focus attention on either syllable and then for the Experimenter to record the effect this has on the REA. The two procedures differ, however, in that in the first example, attention is shifted to one side for a block of trials, until the subject is instructed to shift attention to the other side. This was labeled as 'the forced-attention' paradigm by Hugdahl and Andersson (1986). In the second example, attention is shifted to either the right or left side on a trial-by-trial basis, since the subject cannot know which side he/she is expected to shift attention to on a give trial until the cue-stimulus is given (typically a brief tone), see e.g., Mondor and Bryden (1991). Top-down procedures go back to the use of shadowing approaches where the subject is instructed to shadow a continuous message in one ear while another message simultaneously is presented in the other ear (Cherry, 1953).

The 'Forced-attention' Paradigm

In the following, I will describe studies using the 'forced-attention paradigm' as it was developed by Hugdahl and Andersson (1986) since this is one of the most widely used variants of top-down REA modulation. A detailed description of the 'forced-attention' paradigm is at the end of manuscript for anyone interested in using the paradigm in empirical research, including instructions how to obtain a copy of the paradigm. Top-down manipulations were introduced into asymmetry research as a result of laterality studies by Marcel Kinsbourne (e.g., 1970; see also Hiscock and Kinsbourne, 2011 for an updated account of the attentional model in research on auditory laterality) that the processing differences for language seen between the right and left hemispheres could be due to a default attentional bias to process events happing in one half field of sensory space. Kinsbourne suggested that the REA may be caused by a kind of hemisphere preactivation mechanism of the left, language dominant, hemisphere in a language processing situation caused by attention bias to that side. Part of this suggestion came from studies of hemispace neglect where patients with localized lesions to the right parietal lobe revealed a deficit in processing events occurring the left hemispace (see Heilman and Valenstein, 1972; Heilman and Van den Abel, 1980; see also Heilman, 1995 for a review). It may be interesting to note that Oxbury et al. already in 1967 in a paper in Nature suggested that the ear advantage in dichotic listening may be caused by an attentional bias toward the right ear, and that the effect is not because of an effect of hemisphere dominance. Satz (1968) wrote a response letter to Nature where he presented new data and argued that Oxbury et al. (1967) were mistaken, and added: " …this conclusion is misleading, that it ignores central nervous system mechanisms which are probably associated with the asymmetry, and that the conclusion is inconsistent with experiments which have controlled for order-effects and have found a right ear superiority" (p. 277). Others have later (e.g., Pollmann, 2010) also pointed out that attention could not explain the findings in the 'forced-attention' paradigm introduced by Hugdahl and Andersson (1986) since in this situation the subject is asked also to pay attention to the nondominant left ear, which would not be part of an in-built bias to always shift attention to the left side in hemispace. Instructing the subject to pay attention to either the right or left syllable, and to only report from that side has nevertheless profound effects on the ear advantage. As expected, reporting only from the right ear increases the REA while reporting only from the left ear produces LEA in healthy adult individuals. In the initial studies, Bryden et al. (1983) and Hugdahl and Andersson (1986) it was believed that instructing the subject to half of the time report from the right ear and half of the time to report from the left ear would control for any unwanted effects of attention. Thus, attention was seen as an artifact, confounding a true laterality effect. Asbjørnsen and Hugdahl (1995) and Hiscock et al. (1999) found, however, that the different results for the right and left ear focus could be due to a more general cognitive factor, i.e., that attention was an integrated aspect of auditory laterality. At this time, it was however still believed that the active factor in modulating the REA as a consequence of instructions of side of reporting focus was attention. Later both Hugdahl (e.g., Hugdahl et al., 2009) and Hiscock and Kinsbourne (e.g., 2011) suggested a two-stage model with bottom-up and top-down processing. Hiscock and Kinsbourne (2011) argued that top-down manipulations only help in providing information about localization of signals, but not for the processing of the signal. Hugdahl, on the other hand, argued that the top-down manipulation, depending on which side is attended to, provides information about two fundamentally different modes of cognitive processing. Hugdahl et al. (2009) reviewed the dichotic listening literature on clinical patients, particularly patients with schizophrenia, and found it puzzling why these patients could not perform the condition with instruction to focus attention on the left-ear syllable (labeled 'forced-left' condition), while they performed normal when instructed to focus attention on the right-ear syllable (labeled 'forced-right' condition). It was moreover clear that healthy individuals also struggled more with the 'forced-left' instruction condition than with the 'forced-right' instruction condition, although they could switch to LEA, which the patients could not. Performance data for healthy individuals and patients with schizophrenia are shown in **Figure 7**.

From the schizophrenia and other clinical data, Hugdahl et al. (2009) describe the process that led them to the conclusion that the 'forced-right' and 'forced-left' instruction

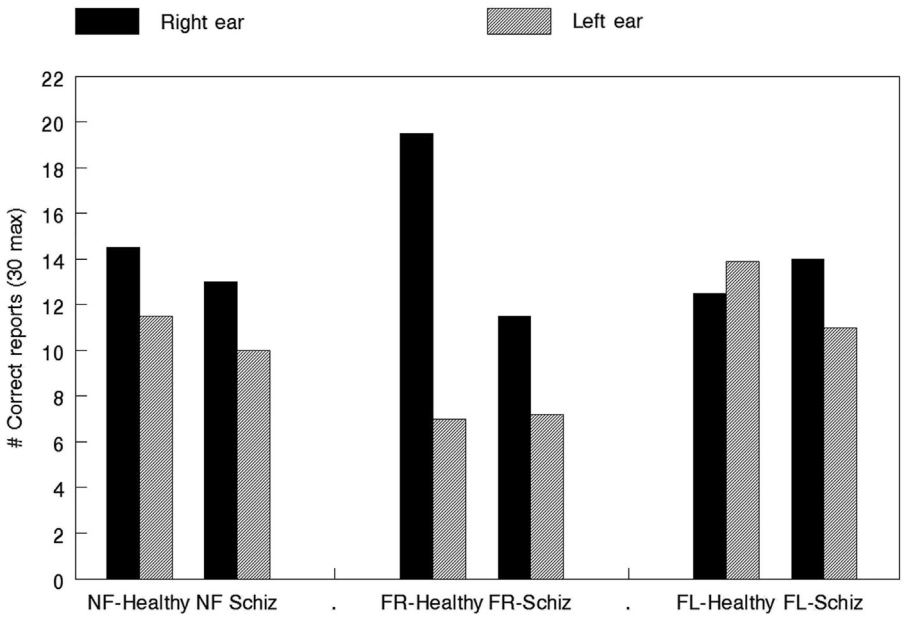

Figure 7 Mean correct reports for the right and left-ear signal in the three 'forced-attention' instruction conditions, nonforced (NF), forced-right/(FR), and forced-left (FL), split for healthy individuals and schizophrenia patients.

conditions reflects different cognitive processes, or different degrees of different cognitive processes.

> We now believe that instructions to focus attention on the right or left ear stimulus induces different degrees of cognitive conflict and a corresponding need for cognitive control strategies. We have below described the process that brings us to this conclusion. The puzzle occurred when interpreting the results in two previous studies from our laboratory (Hugdahl et al., 2003; Løberg et al., 1999) with the 'forced-attention' dichotic listening paradigm (Hugdahl and Andersson, 1986) where it was found that patients with schizophrenia could not modulate the REA when instructed to focus attention on and report only the left ear stimulus, while they were able to modulate the REA when instructed to focus on the right ear stimulus. Healthy control subjects could easily perform both tasks, that is, they increased the REA when instructed to focus attention on and report the right ear stimulus, and reduced the REA, shifting to a LEA when instructed to focus attention on the left ear stimulus. A clue to an explanation of the selective FL effect seen in the patients as an example of cognitive control may be the fMRI study by Thomsen et al. (2004; see also Hugdahl et al., 2000; Jäncke et al., 2001) where it was shown that healthy subjects uniquely showed activation in the left prefrontal and anterior cingulate cortex areas when the BOLD activation images acquired during the FL task were contrasted with images obtained during the FR task, while the reversed contrast showed no remaining significant activations…. Considering that the prefrontal and anterior cingulate cortex have been linked to increased cognitive load, as seen in executive and other cognitively demanding tasks (Duncan and Owen, 2000), it could be argued that the FR and FL attention situations differ in the degree to which they reflect cognitive control (p. 3–4).

Brain Correlates of the REA

The underlying neuronal circuitry for the REA, which would also validate any theoretical model of the REA (e.g., Kimura, 1967; Pollmann et al., 2002) could be studied with functional neuroimaging methods like positron emission tomography (PET using radioactive ^{15}O, or functional magnetic resonance imaging (fMRI)). The first PET imaging study with CV-syllables was published by Hugdahl et al., 1999, using the ^{15}O-PET method. Subjects listened to the standard CV-syllables (see Appendix) while in the PET scanner and had to press a button placed on the chest whenever they detected a predetermined target syllable, thus the paradigm in the PET scanner deviated somewhat from the standard paradigm in that a target-detection procedure was used (see Clark et al., 1988 for a description of dichotic target-detection variants). As control stimuli, musical chords were presented, also dichotically, from three different instruments. The results are shown in **Figure 8**, and reveal significant bilateral activations in the superior temporal gyrus, with a significant leftward asymmetry when directly comparing the left minus the right side.

The corresponding activation for the musical chords showed a rightward asymmetry in the same region in the temporal lobe. The results thus confirmed previous behavioral results, and also the theoretical assumptions regarding the underlying neuronal pathways for the REA. The findings by Hugdahl et al. (1999) were replicated by van den Noort et al. (2008) who used fMRI, with superior spatial resolution and where the subject was instructed to give a verbal response after each syllable-pair presentation. Thus, the Van den Noort et al. (2008) study is so far the neuroimaging study that best resembles the standard behavioral paradigm, using verbal responses. The results replicated the basic asymmetry in the superior temporal gyrus reported in the Hugdahl et al. (1999) study, again confirming the specialization for processing of a phonetic stimulus (CV-syllable) in the left hemisphere. Van den Noort et al. (2008) also applied a conservative statistical procedure in that they used a probability mapping statistics,

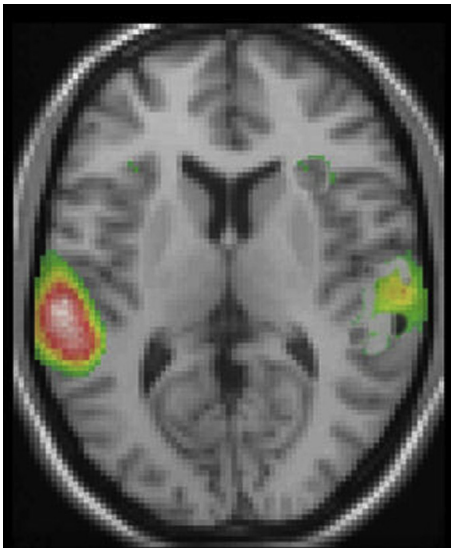

Figure 8 Changes in blood flow in the left and right superior temporal gyrus in healthy individuals, measured with ^{15}O-PET. Data from Hugdahl, K., Brønnick, K., Kyllingsbæk, S., Law, I., Gade, A., Paulson, O.B., 1999. Brain activation during dichotic presentations of consonant-vowel and musical instruments stimuli: a ^{15}O-PET study. Neuropsychologia 37, 431–440 with permission from the publisher.

showing only activated areas that succeeded a threshold of 80% (or 60%) of all subjects showing significantly activated in the same region of interest in the superior posterior region of the left temporal lobe.

Although the study by Hugdahl et al. (1999) was the first functional neruroimaging study using a dynamic imaging method, and with CV-syllables, Coffey et al. (1989) used a blood flow technique with inhalation of a radioactively tagged nobel gas, ^{133}Xenon, which would be distributed with the blood flow in the brain. The stimuli in the Coffey et al. (1989) study were however a pitch-discrimination task where the subject had to discriminate between the high and low pitch of tones presented dichotically, e.g., a high pitch tone in the right ear and a low pitch tone in the left ear. Thus, the task in the Coffey et al. study was not really a study of dichotic listening and hemispheric asymmetry for language, but for auditory processing and pitch processing. The results showed an asymmetry-effect in that activation was higher in the temporal lobe contralateral to the target stimulus of the dichotic pair.

Summary and Conclusions

In this article, I have summarized some trends in the history of dichotic listening research from the original studies by Kimura (1961a,b) to more recent studies using PET (Hugdahl et al., 1999) and fMRI (Van den Noort et al., 2008) to reveal the underlying neuronal circuitry. I have also shown how the REA is modulated and influenced by both bottom-up perceptual and top-down cognitive factors, and that this make the dichotic listening paradigm a much more dynamic and wide-ranging measure for the study of complex perceptual-cognitive asymmetry interactions with consequences for the understanding of cognitive deficits and impairments in clinical groups, like schizophrenia and brain lesion patients. I also show how dichotic listening may be an important vehicle to understand language development and language-related disorders, like dyslexia and other forms of learning disabilities. It may be relevant in this context to remind oneself of the conclusion in the introduction paper in the 50 years anniversary special issue in the *Journal Brain and Cognition* (Hugdahl, 2011), that "auditory laterality and dichotic listening will be an important field of study and method in neuropsychology and related fields of research and in the clinic…also in the next 50 years" (p. 213). Thus, despite unwarranted criticism that dichotic listening has outplayed its role in neuropsychology, recent advances with applications to the clinical domain, and with the introduction of top-down manipulations to study ever more complex brain-behavior interactions, dichotic listening is "still going and going and…" (Hugdahl, 2011, p. 211).

Appendix

In the following the 'forced-attention' paradigm is described as it is used in our laboratory, for the reader who wants to implement the paradigm for own research. The paradigm runs as an application in the E-prime programming platform (Psychology Software Tools, Inc., http://www.pstnet.com/eprime.cfm). A copy of the E-prime dichotic listening application and a manual describing how to use the paradigm and score the data can be obtained by sending an e-mail to Kenneth Hugdahl (hugdahl@psybp.uib.no). The manual also contains normative data for comparisons based on 1800 healthy subjects, aged 7–89 years, males and females, right- and left handers.

The 'Forced-Attention' Dichotic Listening Paradigm

The paradigm consists of series of presentation of CV-syllables via headphones to the subject. The stimuli are the six stop-consonants /b/, /d/, /g/, /p/, /t, /k/ that are paired with the vowel /a/ to form six dichotic six CV-syllable pairs of the type /ba/ – /ga/, /ta/ – /ka/ etc. The syllables are then paired with each other for all possible combinations, thus yielding 36 dichotic pairs, including the six homonymic pairs. The stimuli are natural speech, spoken by a male. The homonymic pairs are not included in the statistical analysis, but just as catch trials to ensure that the subject has understood the instructions. The maximum number of correct reports is thus 30 for each ear/instruction condition (nonforced, NF), with no instruction, forced-right, FR, instruction to focus on the right ear, and forced-left, FL with instructions to focus on the left ear. Mean duration is 350–400 m and the intertrial interval is ~4 sec, during which the subject orally respond, which syllable he/she has perceived on the previous trial. The syllables are read through a microphone and digitized for later computer editing on a standard PC using state-of-the-art audio editing software (SWELL, Goldwave, CoolEdit). The syllables are recorded with a sampling rate of 44 000 Hz and amplitude resolution of 16 bit. After digitization, each CV-pair is displayed on the PC screen and synchronized for simultaneous onset at the

first identifiable energy release in the consonant segment between the right and left channels. The stimuli are played to the subject using digital play-back equipment, connected to high-performance headphones, with intensity between 70 and 75 dB. The subject is told that he/she would hear repeated presentations of the six CV-syllables (ba, da, ga, pa, ta, ka), and that he/she should report which one he/she heard from the six possible syllables after each trial.

Acknowledgment

The present research was funded by a grant from the European Research Council (ERC) advanced grant to Kenneth Hugdahl.

See also: Functional Brain Imaging of Language processes; Lateralization of Language as Demonstrated by Brain Imaging Procedures.

Bibliography

Asbjørnsen, A., Hugdahl, K., 1995. Attentional effects in dichotic listening. Brain and Language 49, 189–201.

Bakker, D.J., 1970. Ear-asymmetry with monaural stimulation: relations to lateral dominance and lateral awareness. Neuropsychologia 8, 103–117.

Bakker, D.J., Kappers, E.J., 1988. Dichotic listening and reading (dis)ability. In: Hugdahl, K. (Ed.), Handbook of Dichotic Listening: Method, Theory, and Research (pp. 513–526). Wiley and Sons., Chichester, UK, pp. 103–117.

Berlin, C.I., 1977. Hemispheric asymmetry in auditory tasks. In: Harnad, S., Doty, R.W., Goldstein, L., Jaynes, J., Krauthamer, G. (Eds.), Lateralization in the Nervous System. Academic Press, New York, pp. 303–324.

Brancucci, A., Babiloni, C., Vecchio, F., Galderisi, S., Mucci, A., Tecchio, F., Romani, G.L., Rossini, P.M., 2005. Decrease of functional coupling between left and right auditory cortices during dichotic listening: an electroencephalography study. Neuroscience 136, 323–332.

Broadbent, D.E., 1956. Successive responses to simultaneous stimuli. Quarterly Journal of Experimental Psychology 8, 145–162.

Bruder, G.E., 1988. Dichotic listening in psychiatric patients. In: Hugdahl, K. (Ed.), Handbook of Dichotic Listening: Theory, Methods, and Research. Wiley and Sons., Chichester, UK, pp. 527–564.

Bryden, M.P., 1988. An overview of the dichotic listening procedure and its relation to cerebral organization. In: Hugdahl, K. (Ed.), Handbook of Dichotic Listening: Theory, Methods, and Research. Wiley and Sons, Chichester, UK, pp. 1–44.

Bryden, M.P., Munhall, K., Allard, F., 1983. Attentional biases and the right-ear effect in dichotic listening. Brain and Language 18, 236–248.

Bryden, M.P., 1963. Ear preference in auditory perception. Journal of Experimental Psychology 65, 103–105.

Cherry, E.C., 1953. Some experiments on the recognition of speech with one and two ears. Journal of the Acoustical Society of America 25, 975–979.

Clark, C.R., Geffen, L.B., Geffen, G., 1988. Invariant properties of auditory perceptual asymmetry assessed by dichotic listening. In: Hugdahl, K. (Ed.), Handbook of Dichotic Listening: Theory, Methods, and Research. Wiley and Sons., Chichester, UK, pp. 71–84.

Coffey, C.E., Bryden, M.P., Schroering, E.S., Wilson, W.H., Matthew, R.J., 1989. Regional cerebral blood flow correlates of a dichotic listening task. Neuropsychiatry 1, 46–52.

Cohen, M., Hynd, G., Hugdahl, K., 1992. Dichotic listening performance in subtypes of developmental dyslexia and a left temporal lobe brain tumour contrast group. Brain and Language 42, 187–202.

Duncan, J., Owen, A.M., 2000. Common regions of the human frontal lobe recruited by diverse cognitive demands. Trends in Neurosciences 10, 475–483.

Eslinger, P.J., Damasio, H., 1988. Anatomical correlates of paradoxic ear extinction. In: Hugdahl, K. (Ed.), Handbook of Dichotic Listening: Theory, Methods, and Research. Wiley and Sons, Chichester, UK, pp. 139–160.

Falkenberg, L.E., Specht, K., Westerhausen, R., 2011. Attention and cognitive control networks assessed in a dichotic listening fMRI study. Brain and Cognition 76, 276–285.

Hannay, H.J., Boyer, C.L., 1978. Sex differences differences in hemispheric asymmetry revisited. Perceptual and Motor and Motor Skills 315–321.

Harris, L.J., 1978. Sex differences in spatial ability: possible environmental, genetic, and neurological factors. In: Kinsbourne, M. (Ed.), Asymmetrical Functions of the Brain. Cambridge University Press., Cambridge, UK, pp. 405–522.

Heilman, K.M., Valenstein, E., 1972. Auditory neglect in man. Archives of Neurology 22, 32–35.

Heilman, K.M., Van den Abell, T., 1980. Right hemisphere dominance for attention: the mechanims underlying hemispheric asymmetries on inattention (neglect). Neurology 30, 327–330.

Heilman, K.M., 1995. Attentional asymmetries. In: Davidson, R.J., Hugdahl, K. (Eds.), Brain Asymmetry. MIT Press, Cambridge MA, pp. 217–234.

Helland, T., Asbjørnsen, A., Hushovd, A.E., Hugdahl, K., 2007. Dichotic listening and school performance in dyslexia. Dyslexia 14, 42–53.

Hirnstein, M., Westerhausen, R., Kosrnes, M.S., Hugdahl, K., 2012. Sex differences in language asymmetry are age-dependent and small: a large-scale, consonant-vowel dichotic listening study with behavioral and FMRI data. Cortex. Epub ahead of publication.

Hiscock, M., Kinsbourne, M., 2011. Attention and the right-ear advantage: what is the connection? Brain and Cognition 76, 263–275.

Hiscock, M., Decter, M.H., 1988. Dichotic listening in children. In: Hugdahl, K. (Ed.), Handbook of Dichotic Listening: Theory, Methods, and Research. John Wiley and Sons, New York, NY, pp. 431–474.

Hiscock, M., Inch, R., Kinsbourne, M., 1999. Allocation of attention in dichotic listening: differential effects on detection and localization of signals. Neuropsychology 13, 404–414.

Hiscock, M., MacKay, M., 1985. The sex difference in dichotic listening: multiple negative findings. Neuropsychologia 23, 441–444.

Høien, T., Lundberg, I., 2000. Dyslexia: From Theory to Intervention. Kluwer Academic Publishers, Dordrecht, NL.

Hugdahl, K., 2011. Fifty years of dichotic listening research – still going and going and going…. Brain and Cognition 76, 211–214.

Hugdahl, K., Andersson, B., 1984. A dichotic listening study of differences in cerebral organization in dextral and sinistral subjects. Cortex 20, 135–141.

Hugdahl, K., Andersson, L., 1986. The "forced-attention paradigm" in dichotic listening to CV-syllables: a comparison between adults and children. Cortex 22, 417–432.

Hugdahl, K., Brønnick, K., Kyllingsbæk, S., Law, I., Gade, A., Paulson, O.B., 1999. Brain activation during dichotic presentations of consonant-vowel and musical instruments stimuli: a ^{15}O-PET study. Neuropsychologia 37, 431–440.

Hugdahl, K., Carlsson, G., Eichele, T., 2001. Age effects in dichotic listening to consonant-vowel syllables: interactions with attention. Developmental Neuropsychology 20, 449–457.

Hugdahl, K., Law, I., Kyllingsbæk, S., Brønnick, K., Gade, A., Paulson, O.B., 2000. Effects of attention on dichotic listening: an ^{15}O-PET study. Human Brain Mapping 10, 87–97.

Hugdahl, K., Rund, B.R., Lund, A., Asbjørnsen, A., Egeland, J., Landrø, N.I., Roness, A., Stordal, K., Sundet, K., 2003. Attentional and executive dysfunctions in schizophrenia and depression: evidence from dichotic listening performance. Biological Psychiatry 53, 609–616.

Hugdahl, K., Westerhausen, R., Alho, K., Medvedev, S., Hämäläinen, H., 2008. The effect of stimulus intensity on the right ear advantage in dichotic listening. Neuroscience Letters 431, 90–94.

Hugdahl, K., Westerhausen, R., Alho, K., Medvedev, S., Laine, M., Hämäläinen, H., 2009. Attention and cognitive control: unfolding the dichotic listening story. Scandinavian Journal of Psychology 50, 11–22.

Hugdahl, K. (Ed.), 1988. Handbook of Dichotic Listening: Theory, Methods and Research. Wiley and Sons, Chichester, UK.

Jäncke, L., Buchanan, T.W., Lutz, K., Shah, N.J., 2001. Focused and nonfocused attention in verbal and emotional dichotic listening: an fMRI study. Brain and Language 78, 349–363.

Kimura, D., 1961a. Some effects of temporal-lobe damage on auditory perception. Canadian Journal of Psychology 15, 156–165.

Kimura, D., 1961b. Cerebral-dominance and the perception of verbal stimuli. Canadian Journal of Psychology 15 (3), 166–171.

Kimura, D., 1967. Functional asymmetry of the brain in dichotic listening. Cortex 3, 163–178.

Kimura, D., 2011. From ear to brain. Brain and Cognition 76, 214–217.

Levy, J., 1971. Lateral specialization of the human brain: behavioral manifestation and possible evolutionary basis. In: Kirger, J.A. (Ed.), The Biology of Behavior. Oregon State University Press, Corvallis, pp. 159–180.

Løberg, E.M., Hugdahl, K., Green, M.F., 1999. Hemispheric asymmetry in schizophrenia: a "dual deficits" model. Biological Psychiatry 45, 76–81.

McGlone, J., 1980. Sex differences in human brain asymmetry: a critical survey. Behavioral and Brain Science 3, 215–227.

Mondor, T.A., Bryden, M.P., 1991. The influence of attention on the dichotic REA. Neuropsychologia 29, 1179–1190.

Nachshon, I., 1988. Dichotic listening models of cerebral deficit in schizophrenia. In: Hugdahl, K. (Ed.), Handbook of Dichotic Listening: Theory, Methods, and Research. Wiley and Sons, Chichester, UK, pp. 565–598.

Obrzut, J., Boliek, C., 1986. Lateralization characteristics in learning disabled children. Journal of Learning Disabilities 19, 308–314.

Obrzut, J., Mahoney, E.B., 2011. Use of dichotic listening techniques with learning disabilities. Brain and Cognition 76, 323–331.

Obrzut, J.E., Boliek, C.A., 1988. Dichotic listening: evidence from learning and reading disabled children. In: Hugdahl, K. (Ed.), Handbook of Dichotic Listening: Theory, Methods, and Research. Wiley and Sons, Chichester, UK, pp. 475–512.

Oxbury, S., Oxbury, J., Gardiner, J., 1967. Laterality effects in dichotic listening. Nature 214, 742–743.

Özgören, M., Bayazit, O., Oniz, A., Hugdahl, K., 2012. Amplitude and phase-shift effects on dichotic listening performance. International Journal of Audiology 51, 591–596.

Penner, I., Schlaefli, K., Opwis, K., Hugdahl, K., 2009. The role of working memory in dichotic-listening studies of auditory laterality. Journal of Clinical and Experimental Neuropsychology 31, 956–966.

Pollmann, S., 2010. A unified structural-attentional framework for dichotic listening. In: Hugdahl, K., Westerhausen, R. (Eds.), The Two Halves of the Brain. MIT Press, Cambridge, MA.

Pollmann, S., Maertens, M., von Cramon, D.Y., Lepsien, J., Hugdahl, K., 2002. Dichotic listening in patients with splenial and nonsplenial callosal lesions. Neuropsychology 16, 56–64.

Rimol, L.M., Eichele, T., Hugdahl, K., 2006. The effect of voice-onset-time on dichotic listening with consonant-vowel syllables. Neuropsychologia 44, 191–196.

Rosenzweig, M.R., 1951. Representations of the two ears at the auditory cortex. American Journal of Physiology 167, 147–214.

Satz, P., 1968. Laterality effects in dichotic listening. Nature 218, 277–278.

Satz, P., 1973. Left-handedness and early brain insult. Neuropsychologia 11, 115–117.

Satz, P., Sparrow, S.S., 1970. Specific developmental dyslexia: a theoretical formulation. In: Bakker, D.J., Satz, P. (Eds.), Specific Reading Disability: Advance in Theory and Method. Rotterdam University Press, Rotterdam, pp. 95–110.

Shankweiler, D., Studdert-Kennedy, M., 1967. Identification of consonants and vowels presented to left and right ears. Quarterly Journal of Experimental Psychology 19, 59–63.

Sidtis, J.J., 1988. Dichotic listening after commissurotomy. In: Hugdahl, K. (Ed.), Handbook of Dichotic Listening: Theory, Methods, and Research. Wiley and Sons, Chichester, UK, pp. 161–184.

Sommer, I.E., Aleman, A., Sommers, M., Boks, M., Kahn, S., 2008. Sex differences in handedness, asymmetry of the Planum Temporale and functional language lateralization. Brain Research 1206, 76–88.

Sommer, I.E.C., 2010. Sex differences inhandedness, brain asymmetry, and language lateralization. In: Hugdahl, K., Westerhausen, R. (Eds.), The Two Halves of the Brain. MIT Press, Cambridge, MA, pp. 287–312.

Sparks, R., Geschwind, N., 1968. Dichotic listening in man after section of neocortical commissures. Cortex 4, 3–16.

Springer, S., Deutsch, G., 1989. Left Brain, Right Brain. Freeman Publishers, San Francisco, CA.

Strauss, E., 1988. Dichotic listening and sodium-amytal: functional and morphological aspects of hemispheric asymmetry. In: Hugdahl, K. (Ed.), Handbook of Dichotic Listening: Theory, Methods, and Research. Wiley and Sons, Chichester, UK, pp. 117–139.

Studdert-Kennedy, M., Shankweiler, D., 1970. Hemispheric specialization for speech perception. Journal of the Acoustical Society of America 48, 579–594.

Tallal, P., Piercy, M., 1973. Developmental aphasia: impaired rate of non-verbal processing as a function of sensory modality. Neuropsychologia 11, 389–398.

Thomsen, T., Specht, K., Rimol, L.M., Hammar, Å., Nyttingnes, J., Ersland, L., Hugdahl, K., 2004. Brain localization of attentional control in different age groups by combining functional and structural MRI. Neuroimage 22, 912–919.

Tunturi, A.R., 1946. A study on the pathway from the medial geniculate body to the acoustic cortex in the dog. American Journal of Physiology 147, 311–319.

Van den Noort, M., Specht, K., Rimol, L.M., Ersland, L., Hugdahl, K., 2008. A new verbal reports fMRI dichotic listening paradigm for studies of hemispheric asymmetry. Neuroimage 40, 902–911.

Voyer, D., 1996. On the magnitude of laterality effects and sex differences in functional lateralities. Laterality 1, 51–83.

Voyer, D., 2011. Sex differences in dichotic listening. Brain and Cognition 76, 245–255.

Wada, J., Rasmussen, T., 1960. Intracarotid injection of sodium amytal for the lateralization of cerebral speech dominance – experimental and clinical observations. Journal of Neurosurgery 17, 266–282.

Westerhausen, R., Hugdahl, K., 2008. The corpus callosum in dichotic listening studies of hemispheric asymmetry: a review of clinical and experimental evidence. Neuroscience and Biobehavioral Reviews 32, 1044–1054.

Westerhausen, R., Moosmann, M., Alho, K., Belsby, S.O., Hämäläinen, H., Medvedev, S., Specht, K., Hugdahl, K., 2010. Identification of attention and cognitive control networks in a parametric auditory fMRI study. Neuropsychologia 48, 2075–2081.

Wexler, B.E., Halwes, T., 1983. Increasing the power of dichotic methods: the fused rhymed words test. Neuropsychologia 21, 59–66.

Dictatorship

A James Gregor, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Before the modern period, the term 'dictatorship' was employed to signify a temporary suspension of a constitution, allowing one person to assume unlimited power in order to effectively respond to some crisis that threatened the integrity or the political survival of a community. By the Middle Ages in Europe, and the institutionalization of hereditary dynasties in East and South Asia, rulers who exercised unlimited powers were identified as 'emperors,' 'monarchs,' 'despots,' or 'tyrants.' By the beginning of the modern period, the term 'dictatorship' was reintroduced and applied to those systems in which constitutional constraints on governmental powers were suspended and unlimited power was transferred, usually, to a single individual. After the termination of the First World War, a new class of modern dictatorships emerged, characterized by 'vanguard parties,' 'charismatic leaders,' formal ideologies that served to legitimate self-selected minority rule, and security and control infrastructures capable of dominating entire populations. Out of that class of dictatorships a subset of regimes appeared that has been identified as 'totalitarianism.' Those regimes, because of their exclusive control over entire nations, and their conviction in their own imperative moral responsibilities, have undertaken the mass-murder or genocide of millions of innocents in the pursuit of political ends, for example, 'the purification of the race,' or the success of the international 'proletarian revolution.' Some of these regimes may have survived into the twenty-first century.

The term 'dictatorship' is one of Roman derivation, and originally signified the temporary suspension of constraints on governmental power and its unlimited transfer to one person. Among the Romans, there was provision for a temporary suppression of citizen rights during times of crises, and the equally temporary transfer of unlimited power, to a person chosen by the Senate and appointed by the Consul. At the end of a specified period – usually 6 months – the person so endowed, the dictator, surrendered power, and citizen rights were restored. The period of emergency stewardship was then duly reviewed, in accordance with constitutional requirements. During instances of emergency, over three centuries of republican constitutional rule, Rome appointed 88 dictators in conformity with law.

Following the rule of Julius Caesar, who exercised dictatorial powers for an unconstitutional length of time, the institution disintegrated and rulers assumed extra-constitutional powers in permanence. Rather than a traditional dictatorship, such rule took on the properties of what might more appropriately be called 'autocracy,' 'despotism,' or, perhaps, 'tyranny.'

Throughout the Middle Ages and more modern times in Europe, as well as during the institutionalization of dynastic rule in East and South Asia, 'autocracies' tended to be regularized through the agency of hereditary right – providing for succession without recourse to violence. Emperors, kings, and sultans could, and frequently would, allocate powers to 'noble' subordinates through a relatively well-articulated, rule-governed process. At times, as was the case in the rights embodied in the Magna Carta, some substantial rights were accorded to subordinates.

The Modern Period

With the decline of the Middle Ages, it became increasingly common that adventurers would seize power in emerging city-states and principalities, and rule as despots or benevolent autocrats. The term 'dictatorship' was rarely, if ever, employed to identify such rule until the commencement of the modern period.

In regions such as Latin America, the term came into increasing vogue after the liberation from Spanish colonialism, irrespective of the fact that the ideal, throughout Central and South America – influenced by the North American model – was constitutional democracy and political stability. The basic charters of most of the nations of Central and South America had incorporated broad versions of the civil, political, and economic rights of citizens to the point of extravagance (Davis, 1961). Unfortunately, governments in Latin America, throughout the remainder of the nineteenth and the beginning of the twentieth centuries, regularly came to power through violence and were commonly dislodged by revolution. During a 25-year period, from 1930 to 1955, for example, more than 50 violent, mainly military, revolutions took place in the region – to provide an impressive roster of 'military dictatorships' that ruled for variable periods of time (Lieuwen, 1960).

In Africa, military dictatorships became the norm in the decades after the decolonization of European holdings that followed the conclusion of the Second World War. The African military – having a monopoly of coercion at its disposal, and often the only institution having logistical reach over the entire territory – acceded to absolute political power over inhabitants. Control, unconstrained by constitutional limitations, was more frequently than not secured by violence or by the threat of violence.

In fact, the term 'dictatorship' (as distinct from terms like *caudillismo* used in Latin America) came into increased currency in Europe long before its employment elsewhere. At the commencement of the modern period, it came to be used, for example, to designate what was initially understood to be a temporary assumption of full power by what was called the Directorate under Cromwell in England, and in 1793, by the Committee of Public Safety, during the French Revolution.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.93034-4

After the conclusion of the First World War, the term was invoked to distinguish a pattern of rule by 'dictators' such as Benito Mussolini in Italy, Primo de Rivera in Spain, or Mustafa Kemal in Turkey. Actually, what such dictatorships signaled was a development that was far more portentous than simply the identification of the exercise of political power by a single ruler in any particular country. What was emerging was a singular political system, having a relatively specific syndrome of properties that can be legitimately identified as 'modern dictatorship' (Talmon, 1960).

Modern Dictatorship

The dislocations that attended the end of the First World War, gave rise to a wave of revolutionary ferment that involved Eastern and Southern Europe (Lee, 1987). In Russia, revolutionaries first overthrew the Romanov monarchy, and then the provisional democratic government of Alexander Kerensky, in order to impose, in 1917, a 'dictatorship of the proletariat' under the leadership of V.I. Lenin.

According to the tenets of classical Marxism (the Marxism of Karl Marx and Friedrich Engels), the dictatorship of the proletariat was to be a transitional form of minority rule that marked the distance between 'bourgeois democracy' – what Marxists considered to be, in reality, a 'dictatorship of the bourgeoisie' – and communist democracy. Lenin assumed full power over the emerging political system that followed the 'proletarian' revolution (*see* Revolution) by legitimizing it through an appeal to an ideological rationale (Furet, 1999).

At approximately the same time, in 1922, Benito Mussolini led a collection of revolutionaries against the parliamentary and monarchial government of peninsular Italy (De Felice, 1965). Initially what appeared to be a transitional government, calculated to deliver the nation from the crises that followed the First World War, the government of Mussolini quickly took on the features of a modern dictatorship. Through reforms commencing as early as 1925, Mussolini's Fascist regime assumed the properties of permanent minority rule under single-party auspices (De Felice, 1968).

Lenin's proletarian dictatorship and Mussolini's Fascism provide a syndrome of traits that specifically characterize modern dictatorship (Johnson, 1991). In both cases, a charismatic leader exercises almost total control over followers as well as the national community out of which they emerge. Leadership in both cases is epistemarchic in character. That is to say, leadership, in the absence of popular elections, is legitimatized by the leader's putative possession of an ideology that is an inerrant guide to policy, satisfying the needs of the collectivity (respectively the proletariat, or the nation). The inerrant ideology informs the belief system of an organization of believers, usually identified as a 'vanguard' of the revolutionary truth. In the course of time, that vanguard becomes the 'unitary party' of the political system – to the exclusion of alternatives. The inerrancy of the ideology and its attendant policy, in turn, is used to justify control over any and all political dissidence (which, under the circumstances, could only be in error) and any initiatives independent of the vanguard leadership (Manoilescu, 1941).

In general, modern dictatorships maintain a military adjunct to unitary party rule. In the Bolshevik case that adjunct was the Red Army. In the case of the Fascist revolution, that adjunct was the Blackshirt militia. In the case of Adolf Hitler's National Socialism, that adjunct was the "Schutz Staffel" (SS). While military force remains an ultimate recourse of the system, modern dictatorships tend to be populist in disposition. They seek plebiscitary confirmation through controlled elections. Leadership, in modern dictatorships, because it is characteristically 'charismatic,' enjoys life tenure. The Democratic People's Republic of Korea (North Korea), as a case in point, exemplifies such a system. After the Second World War, Castro's Cuba, and Mao's China constituted illustrative instances. The system under the rule of Gamal I Abd' El-Nasser in Egypt, and Ghana under Kwame Nkrumah took on many of the same features. They took on the properties of a modern dictatorship however much their actual practice failed to fully satisfy the criteria for admission into the class.

In most cases, modern dictatorships, whatever their initial character, took on the features of an emphatic nationalism. Having begun as an ideology predicated on class, Bolshevism, as an instantial case, very quickly made recourse to 'patriotism' to sustain popular commitment. All classes and segments of classes in the Soviet Union were expected to be supportive of what were unmistakably national purposes, however 'proletarian' they were supposed to be.

Similarly, given the traits exemplified by such a class of polities, they all failed to allow for what parliamentary and constitutional systems understand to be voluntary association, formal opposition, free elections, the separation of legislative, judicial and executive powers, as well as popular freedom of expression and individual movement. In almost all cases, the charter constitutions of such systems change with some regularity (the constitutions of the People's Republic of China, for example, have changed with every change of leadership) and their provisions are subject to unconstrained interpretation (the provisions of the constitution of Fascist Italy, for example, were subject to the regular reinterpretation of Mussolini and the Grand Council of Fascism).

Modern Dictatorship and Totalitarianism

By the beginning of the 1930s, a subset of modern dictatorships began to evince characteristics that led political scientists to assign them a specific identity: totalitarianisms (*see* Totalitarianism). Possessed of an elaborate security and control infrastructure, supported by a factitious or real consensus, and animated by an irrepressible conviction of their own aggressive moral superiority, such politically repressive systems destroyed millions of innocents, either as a measure of control, or in the pursuit of utopian intention (Horowitz, 1997).

The numbers killed by Stalin, Hitler, Mao, and the Khmer Rouge will probably never be accurately known – but the victims of Marxist–Leninist (*see* Marxism–Leninism: The Ideology of Twentieth-Century Communism) regimes probably exceed 100 million (Courtois et al., 1999). 'Mild' totalitarianisms – like those of Mussolini's Italy, Chiang Kai-shek's

China, or Castro's Cuba – neither particularly homicidal nor genocidal – employ incarceration and exile, together with the characteristic denial of voluntary association and freedom of expression, as means of control.

Whatever the subsequent qualifications on the use of the term 'totalitarianism,' throughout the twentieth century an imprecise subset of modern dictatorships has shared a number of reasonably well-defined traits that qualify them for entry into a specific class of modern political systems (Schapiro, 1972; Soper, 1985; Gleason, 1995). How far into the twenty-first century exemplars of the totalitarian subset of modern dictatorships will persist remains a matter of conjecture.

That such systems can evolve into something like a democratic alternative has been established by the transformation of Chiang Kai-shek's Kuomintang on Taiwan into something very much like a vital democracy. The massive reforms on the mainland of China, that followed the death of Mao Zedong, suggest that such systems have the potential to undergo systemic change. How such change is effected remains uncertain.

Editor's Note

This article is a classic statement on dictatorship by a 50-year veteran scholar of the topic and is being reprinted here as a "Legacy Entry."

See also: Authoritarianism; National Socialism and Fascism; Political Parties: Communist; Socialist Societies: Anthropological Aspects; Spatial Data Infrastructures.

Bibliography

Courtois, S., Werth, N., Panne, J.-L., Paczkowski, A., Bartosek, K., Margolin, J.-L., 1999. The Black Book of Communism: Crimes, Terror, Repression. Harvard University Press, Cambridge, MA.

Davis, H.E., 1961. Latin American Social Thought since Independence. University Press, Washington, DC.

De Felice, R., 1965. Mussolini il rivoluzionario 1883–1920. Giulio Einaudi, Turin, Italy.

De Felice, R., 1968. Mussolini il fascista: L'organizzazione dello Stato fascista 1925–1929. Giulio Einaudi, Turin, Italy.

Furet, F., 1999. The Passing of an Illusion: The Idea of Communism in the Twentieth Century. University of Chicago Press, Chicago.

Gleason, A., 1995. Totalitarianism: The Inner History of the Cold War. Oxford University Press, New York.

Horowitz, I.L., 1997. Taking Lives: Genocide and State Power. Transaction, New Brunswick, NJ.

Johnson, P., 1991. Modern Times: The World from the Twenties to the Nineties. Harper, New York.

Lee, S.J., 1987. The European Dictatorships 1918–1945. Methuen, London.

Lieuwen, E., 1960. Arms and Politics in Latin America. Praeger, New York.

Manoilescu, M., 1941. Die einzige Partei als Institution der neuen Regime. Otto Stollberg, Berlin.

Schapiro, L., 1972. Totalitarianism. Praeger, New York.

Soper, P.S., 1985. Totalitarianism: A Conceptual Approach. University Press of America, Lanham, MD.

Talmon, J.L., 1960. The Origins of Totalitarian Democracy. Praeger, New York.

Dictatorship in History

Emilio Gentile, University of Rome La Sapienza, Roma, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Originally, dictatorship was an extraordinary magistrature of the ancient Roman Republic. During the nineteenth and, particularly, the twentieth century, the term has been applied to a variety of political system, whose common characteristic was the unlimited concentration of power in the hands of one individual (or of a group of people, a party or institution) holding it for an indeterminate time. According to specific historical circumstances, modern dictatorships have been defined differently, such as Bonapartism, caudillism, authoritarianism, and totalitarianism.

Introduction

Originally, the word dictatorship defined an extraordinary magistrature of the ancient Roman Republic. During the nineteenth and, particularly, the twentieth century, the meaning of the term dictatorship has been substantially modified. In contemporary parlance, the term dictatorship is used to define *any means of governing and any type of political regime based upon the 'unlimited concentration of power' in the hands of one individual (or of a group of people, a party or institution) holding it for an indeterminate time and exercising it without being controlled by either the State's institutions or by its citizens*.

This article will trace the transformation of the concept of dictatorship within the main instances of government to which it has been applied and with regards the principal interpretations of the phenomenon, that is the ensemble of the governmental types defined as dictatorships produced during the twentieth century when the term first became the object of specific historical research and theoretical analysis.

Roman Dictatorship

Origins and Functions

Dictatorship was probably introduced into the constitutional order of the Roman Republic at the beginning of the fifth century BC, soon after the end of the monarchy. Today, the original interpretation of dictatorship as attributed to the survival of the monarchy tradition has been abandoned. A more plausible interpretation is that which sees Roman dictatorship as being derived of an already existing institution in other cities in the province of Latium (Sterling, 1965). The role of dictator could be conferred for the performance of various public duties, the most important being that of the defense of the city and of its rule in situations of serious danger, in the case of external war (*belli gerendi causa*) or internal turmoil (*seditionis sedandae causa*). In these circumstances, full civil and military powers (*imperium maximum*) were conferred upon the dictator for a maximum period of 6 months. The nomination was made by one of the two consuls in office (who could not, however, nominate themselves) upon the request of the senate, which assessed the gravity of the situation. The dictator stepped down at the same time as the consul who had nominated him. All other magistrates were subordinate to the dictator. He himself, however, did not have the power to repeal them nor to pass or repeal laws, modify the constitutional order or nominate his successor. Once the duty for which he had been nominated to fulfill was completed, even before the passing of 6 months, the dictator resigned power to the ordinary magistrates.

The Transformation and End of the Roman Dictator

The dictator remained in force until the third century BC. The last dictator to have maximum civil and military powers was nominated in 216 BC during the second Punic war between Rome and Carthage. When reinstated in the first century BC by L. Cornelius Sulla and later by Julius Caesar during the time of the civil wars that led to the end of the Republic, the dictator's function was radically transformed. The dictatorships of Sulla and Caesar, despite the outward appearance of legality, originated in the force of the army, were conferred for an indeterminate length of time and were used to eliminate political adversaries and to change the order of the Roman State. Roman dictatorship was legally abolished by Mark Anthony in 44 BC, directly following the death of Caesar (*lex Antonia de dictatura tollenda*) and was not reinstated again during the Empire. The term dictator also was not used for many centuries and was not in any case adopted again until the beginning of contemporary times in order to define new exceptional forms of personal power that appeared in the medieval and modern ages in the form of the 'Podestà' of the Italian communes, the Renaissance 'Lords' or Oliver Cromwell's 'Protectorate' (*see* Middle Ages, The; Renaissance; English Revolution, The).

From Roman Dictatorship to Modern Dictatorship

The Model of Roman Dictatorship in the Modern Ages

The term dictatorship, referring exclusively to the Republican magistrature, appeared once again in political writings of modern times such as those by Jean Bodin and Thomas Hobbes who cited it as an example of the fulfillment of the principle of Reason of State and of absolute sovereignty. Niccolò Machiavelli made a decisive contribution to that which we may term the 'apology conception' of dictatorship, understood as exceptional personal government for the purpose of saving the Republic and defending its liberty against internal and external

enemies. The value and the benefit of dictatorship, according to Machiavelli, consist only in the brevity of its duration, otherwise it was harmful to the Republic (*Discourses*, *I*, *10*, *33*, *35*).

This apologetic conception of dictatorship was again proposed in the seventeenth century with the idealization by democratic thinkers of the Enlightenment of 'republican public spirit' (*see* Enlightenment). Jean-Jacques Rousseau argued that in periods of danger the democratic State could nominate a 'supreme chief who silences all the laws', fixing a relatively short period for the fulfillment of his role without the capacity for its being prolonged because only in this way would it be possible to avoid the transformation of dictatorship into tyranny (*Social Contract*, 1762, IV (6): 138).

Dictatorship and Revolutionary Power

The term dictatorship reentered political language at the end of the eighteenth century with the democratic revolutions in America and Europe (*see* American Revolution, The; French Revolution, The). Following the experience of the Public Health Committee, nominated by the Convention (1793–95) both a new conception and practical experience of dictatorship emerged. This was identified by the advent of 'revolutionary power' as a form of absolute government, exceptional and temporary yet personal, which claimed its legitimacy on the basis of the principle of popular sovereignty, the wish to create a new constitutional order and a more just society. After the fall of the Jacobins, François-Noël Babeuf theorized the necessity of dictatorship by a small revolutionary minority in order to seize power by means of insurrection with the end of realizing a 'society of equals'. After the failure of his insurrectionary attempt (1796) the concept of revolutionary dictatorship was taken up by Filippo Buonarroti, a disciple of Babeuf.

Dictatorship in the Nineteenth Century

'Bonapartism' and 'Caesarism' in Europe

In the course of the nineteenth century, the term dictatorship was used in its original meaning by the military leaders of the national revolutions. Simon Bolivar proclaimed himself dictator in 1813; Giuseppe Garibaldi assumed the power of dictator in 1860 during the expedition of the Thousand. The idea of dictatorship was often associated, within nationalism, with the idea of exceptional government for national liberation. Most important, for the history of dictatorship and the transformation of the concept were the new forms of personal regimes founded in Europe by Napoleon Bonaparte (1799–1814) and later by his grandson Louis Napoleon Bonaparte, becoming Napoleon III (1851–70). Their dictatorship originated in a coup d'état. However, they attempted to legitimize their personal power later with a popular plebiscite and with the transformation of the monarchy without formally reneging on the principles of the French Revolution.

These two experiences of government gave rise to the term 'Bonapartism', from then on used to define the new type of dictatorship imposed by a coup d'état but claiming a democratic legitimacy by means of plebiscites. The introduction of the term 'Caesarism' was also used in the definition of this new type of dictatorship, in that Julius Caesar was considered the prototype for the modern dictator, emerging victorious after a period of serious social and civic conflict and imposing his personal power using military strength but also with the support of the masses, presenting himself as the pacifier of the classes, above parties, restorer of stability and of the unity of the State, the incarnation and the interpreter of the national will.

Dictatorship in Latin America

A particular demonstration of a South American version of 'Caesarism' can be found in the presidential dictatorships installed in the majority of the Latin American Republics emerging after independence. These dictators were politicians or military men who elected themselves president and perpetuated their personal power for decades while, nevertheless, paying homage to liberal ideology, allowing for the survival of the constitutional order and, above all, imposing the centralization of state power against the *caudillos*, the local leaders who, in the first years following independence, instituted personal control with the help of the army, the personal fascination with virility and the support of the indigenous and métisse populations. Other president-dictators, followers of positivism, justified their dictatorship with the formula 'Order and Progress' as a necessary means for reinforcing the unity and cohesion of the State, as a pedagogical tool for the education of the people and as a principal factor in the renovation of society and progress. This new version of the apologetic conception of dictatorship remained one of the principal formulae of populist self-legitimation for the dictators that successively dominated South America (*see* Populism).

Apologists for and Critics of Dictatorship

Other versions of the apologetic conception of dictatorship were elaborated in romantic culture, in Catholic reactionary thought, and in positivism: the dictator was variously exalted as the supreme guide of the people, the interpreter of God's will, or the political genius who actuated the laws of history and of progress. Liberal culture, on the other hand, held that the same possibility to confer absolute power upon a single individual was incompatible with the concept of the modern State founded upon the liberty of citizens, on the sharing of power and on the principle of majority rule. The *Dictionnaire Politique* (Duclerc and Pagnerre, 1848) explicitly refuted the treatment of dictatorship as 'an element of modern politics' because 'invoking dictatorship means invoking violence' and annihilating 'the most sacred principle of democracy, that of majority rule' (*see* Liberalism, Anthropology of; Crowds in History).

The Dictatorship of the Proletariat

In socialist thinking, the concept of dictatorship as revolutionary power was developed. Louis-Auguste Blanqui theorized the necessity of the dictatorship of a revolutionary minority for the realization of Communism. This conception was criticized by Karl Marx and Friedrich Engels because their idea of revolutionary dictatorship had to be the *impersonal power* of the working class. The Paris Commune of 1870 represented for Marx, the first concrete attempt at proletarian dictatorship.

In that in a society divided into classes every type of State is no other than the dictatorship of the dominant class, according to the idea sketched out by Marx in the *Critique of the Gotha Programme* (1875), 'proletarian dictatorship' was understood as the organization of the working class at the level of the State following the conquest of power in the transition period from capitalist to Communist society.

The concept of 'proletarian dictatorship' was first elaborated with the conquest of power in Russia in 1917 by the Bolshevik Party led by Lenin. Lenin was a convicted and tenacious advocate of the necessity of dictatorship intended as an unlimited, inflexible, and implacable power in the repression of enemies of the proletariat because, in the absence of this power, the transition from Capitalism to Socialism would be impossible. However, against Marx, Lenin claimed that dictatorship of single individuals, endowed with unlimited power for the realization of the socialist revolution was not in fact incompatible with proletarian dictatorship, exercised by the Bolshevik Party as the vanguard of professional revolutionaries and interpreters of the historical role of the proletariat.

The Leninist conception of dictatorship and its practical activation were criticized by other revolutionary Marxists such as Rosa Luxemburg who claimed that proletarian dictatorship should consist of a means of applying democracy, not of its abolition, and the work of the *class*, and not of a small minority of leaders in the name of the class.

The Twentieth Century: A Century of Dictatorships

The Epidemic of Dictatorship in Europe between the Two World Wars

"In the Victorian Era", wrote Winston Churchill in 1930 in the introduction to one of the first attempts at the historical and theoretical analysis of dictatorship (Forst de Battaglia, 1930), "dictatorship was at once a monstrosity and an anachronism". The triumph of 'parliamentary democracy' in Europe in the aftermath of the Great War was ephemeral. In 1922, the 'crisis of democracy' and the 'crisis of parliamentarism' were already being talked of, while with the affirmation of absolute power by the Bolshevik Party in Russia, a new type of dictatorship as a regime by a single party was being affirmed for the first time. The fascination with the Bolshevik experiment excited masses of European workers and prompted the birth of Communist parties with the aim of violent conquest of power for the establishment of 'proletarian dictatorship'. Nevertheless, no attempt at Communist conquest of power was successful. On the contrary, the fear of Bolshevism generated reaction on the part of the Bourgeoisie and the middle classes in many countries, which led to the creation of authoritarian, nationalistic, and anti-Bolshevik governments. After 1922, almost every year, more or less durable dictatorial nationalist governments emerged in Italy, Spain, Turkey, Poland, Greece, Portugal, Austria, Germany, the Baltic States, and the Balkans.

The Model of Fascist Dictatorship

At the end of 1922, with B. Mussolini's coming to government in Italy (28 October), for the first time a nationalist revolutionary movement organized around a new form of party, the 'party militia', conquered power in a country governed by a liberal democracy with the declared object of destroying it. This gave rise to the establishment of a regime that antifascists were the first to call a 'one-party state', 'totalitarian system', or 'totalitarianism', defining a new type of dictatorship founded upon the predominance of a single party, the centralization of power in the hands of the party leader, the imposition of a singular ideology with the aim of realizing the subordination of society to the State by means of control by the singular party and its organizations (Gentile, 2002).

Until the middle of the 1920s, Fascism, and not Bolshevism, was considered the real threat to parliamentary democracy. Italian fascism effectively became the model for other nationalist movements that proclaimed their dictatorial vocation, openly evoking the ideals of the French Revolution. At the beginning of the 1930s, Europe was invaded by an 'epidemic of dictatorship' (Sforza, 1931). Initially, the dictatorial disease was considered to be exclusive to the backward countries. This interpretation was proved wrong after the rise of National Socialist dictatorship in Germany (*see* National Socialism and Fascism). In the second half of the 1930s, however, the 'seduction of Fascism' seemed to grow (Drabowitch, 1934). During the Second World War, the majority of Southern, Central, and Eastern Europe were governed by dictatorial systems.

Totalitarian Regimes

In the period between the two world wars a new apologetic myth of the dictatorial figure as the 'strong man' who emerges in times of crisis to rescue the State and reestablishes order, discipline, and efficiency was diffused. At the same time, the first attempts at comparative analysis aiming at elaborating a unitary perspective on the phenomenon of dictatorship were began. Already, by the second half of the 1920s, several analogies between the dictatorship of Bolshevism and that of Fascism came to prominence, later compared also to the dictatorship of National Socialism. The classification of these new types of dictatorship under the category of 'totalitarian states' appeared at the end of the 1920s and was already consolidated by the 1930s and 1940s when the first comparative analyses of Fascist, Communist and National Socialist regimes were carried out (see Kohn in Stanton Ford, 1935; Hoover, 1937; Sturzo, 1938 in Sturzo, 1972; Aron 1939 in Aron, 1993; Borkenau, 1939; Cobban, 1939; Proceeding of the American Philosophical Society, 1940; Neumann, 1942). These studies analyzed the new form of dictatorship from the point of view of its most original aspects, neatly distinguishing it from both Roman and nineteenth century dictatorship. These aspects included (1) the peculiarities of the original phenomena of dictatorship, attributed to the legacy of war, the embitterment of class war, the radicalization of politics, the incapacity of the parliamentary system to adapt itself to the tumultuous growth of mass society, and to provide an effective solution to social and economic conflicts; (2) the interest of the masses living through the turmoil of a society undergoing transformation and looking for simple and reassuring solutions by subordinating themselves to the command of a 'strong man' who promised to resolve the

crisis with brutal yet efficient methods; (3) the novelty of contemporary dictatorship as a one-party regime, created by a mass movement that used parliamentary channels, violence, and terror to conquer and maintain power, rendering permanent its own dictatorship and institutionalizing it as a fundamental pillar of a new political system; (4) the personalities of the dictators brought to power by these parties such as Stalin, Mussolini, and Hitler, that is 'new men', emerging from anonymity and rising to power due solely to their political capacity, exalted and revered like demigods by their followers and by the masses; (5) the fanaticism and missionary spirit of the new dictators who endowed their ideology and action with the character of a *secular religion*; (6) the use of methods of technological and organizational modernization for the mobilization of the masses and the development of an intense and constant activity of collective indoctrination; and (7) the ambition of subordinating the collectivity and society to the permanent and capillary control of a *single party* with the aim of creating a new species of human being.

After the Second World War, the end of the Fascist and Nazi regimes, the experience of the Holocaust, the victory and expansion of Stalin's Russia, and the foundations the new Communist regimes in Eastern Europe and in Asia, brought about new interpretations of totalitarianism, highly influenced by the political orientation of Western anticommunism during the Cold War, identifying Nazism and Stalinism as the extreme experiences of the totalitarian dictatorship (Gleason, 1995). To define totalitarianism as a totally new of modern dictatorship, scholars emphasized mass terror as the essence of totalitarianism (Arendt, 1951), the total control over the State and society (Friedrich and Brzezinski, 1956), or the monopoly of political power by a single party (Aron, 1998[1995]).

Totalitarianism: A Controversial Phenomenon

For over 60 years, historians, social scientists, philosophers, theologians have examined the nature of totalitarianism and the effects it has had on the history of humanity (Söllner et al., 1997). However, as often happens when theoretical problems collide with the meaning of human existence, the problems sparked by the debate on totalitarianism are more numerous than the answers given by the experts. There are conflicting opinions even in regards to the definition of totalitarianism, as well as in the application of this definition to political movements and regimes of the 1900s and earlier centuries, starting from the French Revolution. For example, there are those who refuse to identify Communism, Fascism, and National Socialism as totalitarian, and those who support the historical validity of this identification. There are those who believe that the totalitarian theory has a scientific value and those who believe it is only a residue of the anti-Soviet propaganda of the Cold War. Not few misunderstandings on the historical interpretation of phenomena such as Fascism, Communism, and National Socialism have resulted from the issue of totalitarianism. The major objections to the validity of totalitarian theories, stem from the doubt of the legitimacy of adopting a single theoretical model to define such profoundly different historical experiences, favoring similarities, and analogies in action procedures, mentality, behavior, and organization and in political power. Because of this criticism, there was an eclipse of the concept of totalitarianism during the 1970s and 1980s. But, since the end of Communist regimes in Eastern Europe in 1991, there has been a 'revival' of the debate on totalitarianism (Siegel, 1998) as a modern type of dictatorship peculiar of the twentieth century. A new elaboration of the concept has been put forward by scholars who consider that totalitarianism is better understood as a method and a technique of political domination in order to achieve a specific goal, than as a goal per se to be achieved, by stressing the differences as well as the affinities among the totalitarian regimes (Gentile, 2005, 2006).

The Dictatorial Phenomenon after the End of the Second World War

In the aftermath of the Second World War, while parliamentary democracy was being reestablished in Italy and Germany, liberal democracy was definitively consolidated in Western Europe and, after 1945, there were no new dictatorships in this part of the continent. From among the nationalist dictatorships emerging from the interwar period, only the Salazar regime in Portugal and that of Franco in Spain remained standing, both of which returned to democracy after the death of the two dictators in the second half of the 1970s. The only new dictatorship was that of the Colonels' regime in Greece (1967–74).

The phenomenon of dictatorship, nevertheless, enjoyed a new and intensified period of expansion in almost all continents. Only North America and Australia remained uncontaminated by the new dictatorial epidemic. Directly after the end of the Second World War, East Germany, all of Eastern Europe and the Balkans were put under the control of the new Communist dictatorial regimes, the so-called popular democracies, completely subordinated to the political and military hegemony of Stalin's Soviet Union from which only the Communist regimes of Yugoslavia and Albania escaped. From 1949, with the victory of the Communist revolution led by Mao in China, the world's most populous country became governed by a dictatorial regime with its own characteristics, different to the European Communist dictatorships and to those of the Soviet Union. During the 'Cold War', the Communist model, with variations original to each country, inspired the construction of several dictatorial regimes, both civilian and military, in Central America, Asia, and Africa with the help of the two largest Communist countries, the Soviet Union and China, becoming, since the 1950s, rival and antagonistic states both for ideological reasons and for those concerning territorial conflict (*see* Revolutions; Chinese Revolutions: Twentieth Century; Cuban Revolution, The; Vietnamese Revolution, The).

The common characteristics of the Communist dictatorships were (1) the adoption of Marxist–Leninism as their official ideology; (2) the supremacy of the Communist Party as the single and dominant party; (3) the exaltation of the party leader resulting in the institutionalization of the 'cult of personality'; (4) the policed repression of any opposition or dissidence, including the recourse to mass terror; (5) the nationalization and planning of the forces of production and of the economy; (6) the submission to control by the Communist Party of all political and social activities; and

(7) the development of a particular form of 'Communist patriotism', centered upon the myth of the historical mission of the proletariat (Kautsky, 1962).

During the 'Cold War', anti-Communism was the motive and principal ideological justification for the establishment of right-wing, civilian and, more often, military dictatorships sustained by the United States. These dictatorships, especially the military ones (Nordlinger, 1977), when not based exclusively on terrorist repression and the arbitrary use of power in the interests of the dictator, his family or acolytes, attempted to relate policed repression to a demagogic politics of nationalist populism all the while linked to programmes of modernization and development (*see* Modernisation and Modernity in History). In the second half of the twentieth century, dictatorship in the Third World became 'the normal form of government' (Roy, 1977). Almost all the new states emerging in Africa and Asia after the fall of the colonial empires were governed by dictatorial regimes, generally military with a single party created from above, a charismatic leader and a populist ideology of National Socialism with programmes of 'secularization' and 'modernization'. In the second half of the 1970s, in several Islamic countries, religious 'fundamentalism' became the matrix and ideology for new dictatorial governments.

The tendency toward the proliferation of dictatorships came to a halt in the last two decades of the twentieth century, when the number of dictatorships – or governments considered as such –notably diminished (Hamon, 1982). In Latin America, parliamentary democracy was reinstated in the grand majority of states previously governed by dictators. In the late twentieth century, numerous states in Africa were involved in a process of democratization. The fall of the Soviet Union brought about also that of the Communist regimes of Central and Eastern Europe that it sustained (*see* Revolutions of 1989–90 in Eastern and Central Europe). There was a 'third wave' of democratization, as it has been called, to distinguish it from the first 'long' wave of democratization in 1828–1926 and the second, 'short' wave in 1943–62 (Huntington, 1991).

Although it would be too optimistic to declare the definitive death of dictatorship in the contemporary world and the victory of the democracy of human rights (Hermet, 1993), the promotion of human rights has acquired new determination, albeit not always effective, in contesting antidemocratic, repressive, and terrorist governments and in the defeat of the various apologetic conceptions of dictatorship (*see* Human Rights, History of). The breakdown of longstanding dictatorships in some Muslim countries (Tunisia, Egypt, and Libya) during the so-called Arab Spring, in 2011, has paved the way for democracy, and some other autocratic regimes in Muslim, such as Yemen and Syria, are crumbling under the attack of popular insurgence. However, at the present (2012) the countries involved are still in the throes of change. Meanwhile, in the last decade democratic regimes have been under pressure in many parts of the Western world because of longstanding economic crisis, high unemployment rate, growing social inequality, xenophobia, political, and/or religious extremism, corruption in State administration, and the unscrupulous political influence of media tycoons. In most regions the average democracy score for 2011 is lower than in 2010, including the developed countries of North America and Western Europe; there was a decline in the average score for Eastern Europe and small declines for both Asia and Latin America, although these were offset by increase in average scores in the Middle East and North Africa and Sub-Saharan Africa (Economist Intelligence Unit Index of Democracy, 2012). At the beginning of the second decade of the Third millennium, a 'fourth' wave of democratization may be possible, but not for sure.

Contemporary Dictatorship: Problems of Definition and Interpretation

Dictatorship, 'Tyranny', 'Despotism'

In contemporary speech, dictatorship is often associated with tyranny or despotism. Terms such as dictator, tyrant, or despot are often used as synonyms or alternated between. The first students of the phenomenon of dictatorship in the interwar period, while maintaining that the concept of dictatorship is linked to the Roman model of an extraordinary, yet constitutional and temporary, magistrature, preferred to speak of a 'new age of despots', or of a new 'ère des tyrannies' (Halèvy, 1938). In more recent studies also the term 'tyrant' is preferred over that of dictator for defining the numerous ranks of those holding absolute personal power, dominating, in the second half of the century – and still in some cases dominating – in many countries of the so-called Third World (Rubin, 1987; Chirot, 1994). In fact, contemporary dictatorship evokes some of the characteristics of Greek tyranny, but there are also substantial differences. Above all, the origin of contemporary dictatorship is not always illegitimate. For example, Hitler came to power in 1933 through legal channels and many presidents who became dictators in Latin America were legally elected. Furthermore, the majority of modern dictatorships have a very complex structure of organization and exercise of power due to the mass dimension of the societies in which they operate, the methods they adopt and the scopes they propose. All the above makes contemporary dictatorship substantially new and different to tyranny. Therefore, the thesis that holds that the difference between tyranny and contemporary dictatorship as a form of absolute personal power is not substantive but regards solely the form, the technique and method of government should be forcefully rejected.

Dictatorship and the Dictator

Several scholars consider the figure of the 'dictator' an essential and fundamental element in the definition of dictatorship (Cobban, 1939). A large part of the general history of dictatorship consists of collections of biographies that sometimes link, in more or less chronological succession, Greek tyranny and the Roman dictators, Cola of Rienzi and Girolamo Savonarola, Cromwell and Robespierre, the two Napoleons, the 'caudillos' and the authoritarian and totalitarian dictators of the twentieth century (Hallgarten, 1957). Some students maintain that governments in which unlimited power is not concentrated in one individual cannot be defined as dictatorships in the proper sense of the term. The identification of dictatorship with the figure of the dictator is certainly justified in the majority of the cases in history. In almost all

contemporary dictatorships, the dominant personality of one man emerged at the outset or was subsequently imposed, often surrounded by a 'charismatic' and 'sacred' fascination with the cult of the personality (Schweitzer, 1984). At the same time, dictatorial governments exist without the figure of the dictator. For example, the Soviet Union following Stalin's death and the dictatorial regimes governed by military juntas. Moreover, not even in totalitarian regimes that appear to be internally controlled by the personal power of a dominant and charismatic figure, can the nature and character of the dictatorship as a system of power founded upon the single party be identified in the personage of the dictator (Schapiro and Lewis, 1970; Tucker, 1972).

Definitions and Typologies

Contemporary dictatorships are extremely heterogeneous: civilian or military, personal or collective, conservative or revolutionary, traditionalist or modernizing, bureaucratic or charismatic, authoritarian or totalitarian. Moreover, not only do contemporary dictatorships have nothing in common with their ancient predecessor, apart from the concentration of power, they also differ extensively between them according to the geographical, historical, and cultural contexts in which they emerge and due to the economic, political, and social situations that give rise to them. Furthermore, they differ as regards the social origins of their protagonists as well as the latter's political formation, the modes in which power and methods of action are organized, the ideologies they profess, and the scopes they propose. Finally, differences exist relative to the historical development of their emergence, the results they have achieved, and the legacies they have left.

Given the variety and complexity of the phenomenon of dictatorship in the twentieth century, it is very difficult to elaborate a general, unitary theory of dictatorship. Several fundamental problems therefore remain unsolved, particularly those that concern the specific nature of contemporary dictatorship with respect to other forms of absolute personal power. There are various definitions and typologies for classifying the numerous forms of unlimited concentration of power appearing during the course of the twentieth century, commonly adopting the term dictatorship albeit with a meaning varying widely from the original (see especially Schmitt, 1928; Friedrich and Brzezinski, 1956; Neumann, 1957). There are some scholars, however, who prefer to use the term dictatorship in reference only to its original meaning, finding the term 'antidemocratic regime' more appropriate for defining the modern forms of unlimited concentration of power, be they authoritarian or totalitarian (Linz, 2000). The identification of contemporary dictatorship with antidemocratic regimes is, however, only one aspect of a much more ambiguous relationship between contemporary dictatorship and democracy, understood as the affirmation of the principle of popular sovereignty and mass participation in political life. Contemporary dictatorship searches for its own formula of legitimation within the principle of popular or national sovereignty, and this is particularly noticeable in certain states in Africa, in the Middle East, and across Asia. The 'third' wave of democratization has not resulted everywhere in stable democratic regimes, but in pseudo-democratic new autocracies, even if they claim their democratic legitimacy because of plebiscitary support in national elections. A variety of new concept has been elaborated such as 'semidictatorship', 'semidemocracies', 'hybrid regimes', to define the regimes that combine democratic and authoritarian features (Ottaway, 2003).

Conclusion: General Characteristics of Contemporary Dictatorship

The diversity in the experiences of power defined as dictatorship in the contemporary world, the variety of definitions and typologies and the problems remaining controversial regarding the nature of dictatorship do not, nonetheless, impede us from being able to identify at least some common characteristics enabling the definition of the specificity of the phenomenon of dictatorship in the modern era in their formative or stabilizing phases (Ezrow and Frantz, 2011).

According to the definition furnished at the beginning of this article, all types of government and regime in which a single person, a group of people, a party, a military or religious institution (1) conquers the monopoly over political power, by legal or illegal means, and maintains it for an indeterminate length of time; (2) suppresses, reduces, or impedes citizens' enjoyment of their civil or political rights; (3) forbids or represses the activities of dissidents or of the opposition with the use of force and with recourse to terrorist means; and (4) excludes *de facto* the possibility for the legal and pacific repeal of the rulers by the ruled even while allowing for electoral consultations and plebiscites to take place may be included under the category of contemporary dictatorship. Contemporary dictatorship is, thus, the antithesis of democracy understood as a representative political system based upon the division and balance of power, a respect for the personal and public rights of citizens, on freedom of speech and of association, on the periodic election of the government by the electorate, by means of a free and peaceful political contest and according to legal procedures guaranteed by the constitution.

See also: Absolutism, History of; Democracy, History of; Historical Thought and Historiography: Christian Tradition; National Liberation Movements; National Socialism and Fascism; Nations and Nation-States in History.

Bibliography

American Philosophical Society, 1940. Symposium on the totalitarian state. From the standpoints of the history, political science, economics and sociology. In: Proceeding of the American Philosophical Society. Philadelphia, PA.

Arendt, H., 1951. The Origins of Totalitarianism. Harcourt, Brace and Company, New York.

Aron, R., 1993. Machiavel et les tyrannies modernes. Fallois, Paris, pp. 173–191.

Aron, R., 1998[1995]. Democracies and Totalitarianism. Weidenfeld & Nicholson, London.

Borkenau, F., 1939. The Totalitarian Enemy. Faber&Faber, London.

Brooker, P., 2000. Non-Democratic Regimes. Pallgrave Macmillan, London.

Chirot, D., 1994. Modern Tyrants: The Power and Prevalence of Evil in Our Age. The Free Press, New York.

Cobban, A., 1939. Dictatorship: Its History and Theory. Jonathan Cape, London.

Collier, D., 1979. The New Authoritarianism in Latin America. Princeton University Press, Princeton, NJ.

Drabowitch, W., 1934. Fragilité de la liberté et seduction des dictatures. Mercures de France, Paris.
Duclerc, E., Pagnerre, 1848. Dictionaire politique de Langage et de la Sience Politiquee rédigé par une réunion de députes, de publicistes et de journaliste avec une introduction par Garinier-Pagès publiè. Pagnerre éditeur, Paris.
Duverger, M. (Ed.), 1982. Dictatures et légitimité. Presse Univeristaire de France, Paris.
Ezrow, N.M., Frantz, E., 2011. Dictators and Dictatorships: Understanding Authoritarian Regimes and Their Leaders. The Continuum International Publishing Group, New York.
Ford, S.G., 1935. Dictatorship in the Modern World. The University of Minnesota Press, Minneapolis, MN.
Forst de Battaglia, O. (Ed.), 1930. Dictatorship on its Trial. George G.Harp, London.
Friedrich, C.J., Brzezinski, Z., 1956. Totalitarian Dictatorships and Autocracy. Harvard University Press, Cambridge, MA.
Gentile, E., 2002. Fascism in power: the totalitarian experiment. In: Lyttelton, A. (Ed.), Liberal and Fascist Italy. Oxford University Press, Oxford, pp. 139–174.
Gentile, E., 2005. The Origins of Fascist Ideology. 1918–1925. Enigma, New York.
Gentile, E., 2006. Politics as Religion. Princeton University Press, Princeton, NJ.
Gleason, A., 1995. Totalitarianism. The Inner History of the Cold War. Oxford University Press, Oxford and New York.
Griffin, R., 2005. Fascism, Totalitarianism and Political Religion. Routledge, London and New York.
Halèvy, E., 1938. L'ère des tyrannies: études sur le socialisme et la guerre. Gallimard, Paris.
Hallgarten, W.F., 1957. Dämonen Oder Retter. Europäische Verlagsanstalt, Frankfurt a.m.
Hamon, L. (Ed.), 1982. Mort des Dictatures? Economica, Paris.
Hermet, G., 1993. Les désenchantements de la libertè: La sortie des dictatures dans les années 90. Fayard, Paris.
Hoover, C.B., 1937. Dictators and Democracies. The Macmillan Company, New York.
Huntington, S.P., 1991. The Third Wave: Democratization in the Late Twentieth Century. University of Oklahoma Press, Norman, OK.
Kautsky, J.H. (Ed.), 1962. Political Change in Underdeveloped Countries. Nationalism and Communism. J. Wiley & Sons, New York and London.
Linz, J., 2000[1975]. Totalitarian and Authoritarian Regimes. Boulder, London.
Neumann, S., 1942. Permanent Revolution: The Total State in a World at War. Harper & Brothers, New York.
Neumann, F.L., 1957. Anxiety and Politics. In: Marcuse, H. (Ed.), The Democratic and the Authoritarian State: Essays in Political and Legal Theory. Free Press, Glencoe, IL, p. 284.
Nordlinger, E.A., 1977. Soldiers in Politics: Military Coups and Governments. Prentice Hall Inc., Englewood Cliffs.
Ottaway, M., 2003. Democracy Challenged: The Rise of Semi-Authoritarianism. Brooking Institution Press, Washington, DC.
Perlmutter, A., 1981. Modern Authoritarianism: A Comparative Institutional Analysis. Yale University Press, New Haven and London.
Roberts, D.D., 2006. The Totalitarian Experiment in Twentieth Century Europe. Understanding the Power of Great Politics. Routledge, New York and London.
Roy, M.-P., 1977. Les regimes politiques du Tiers monde. Libraire générale de droit et de jurisprudence, Paris.
Rubin, B., 1987. Modern Dictators: Third World Coup Makers, Strongmen, and Populist Tyrants. McGraw-Hill Book Company, New York.
Schapiro, L., Lewis, J.W., 1970. Party Leadership and Revolutionary Power in China. Cambridge University Press, Cambridge.
Schmitt, C., 1928[1921]. Die Diktatur: Von del Anfängen des modernen Souveränitätsgedankens bis zum proletarischen Klassenkampf, second ed. Duncker & Humbolt, Munich und Leipzig.
Schweitzer, A., 1984. The Age of Charisma. Nelson-Hall, Chicago.
Sforza, C., 1931. Dictateurs et dictatures de l'après-guerre. Gallimard, Paris.
Siegel, A., 1998. The Totalitarian Paradigm after the End of Communism. Toward a Theoretical Reassessment. Rodopi, Amsterdam and Atlanta.
Söllner, A., Walkenhaus, R., Wieland, K. (Eds.), 1997. Totalitarismus. Eine Ideengeschichte des 20. Jahrhunderts, Akademie Verlag, Berlin.
Stammer, O., 1955. Demokratie und Diktatur. Gehlen, Berlin.
Sterling, E., 1965. Der unvollkommene Staat: Studien über Diktatur und Demokratie. Europäische Verlagsanstalt, Frankfurt a.m.
Sturzo, L., 1972. Politica e morale (1938). Coscienza e politica (1953). Zanichelli, Bologna.
Tucker, R., 1972. The dictator and totalitarianism. In: The Soviet Political Mind. Stalinism and Post-Stalin Change. Norton & Company, New York and London, pp. 178–179.
References used in translation:
Rousseau, J.J., 1997. The Social Contract and Other Later Political Writings. In: Gourevitch, V. (Ed.). Cambridge University Press, Cambridge, p. 138.

Differential Aging

Susan K Whitbourne, University of Massachusetts Amherst, Amherst, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Differential aging refers to the principle that the rate of functional changes in various life domains shows inter- and intra-individual variability. The study of differential aging may be traced to large-scale investigations identifying different patterns of aging. Current models of successful aging provide insight into the differential aging process as experienced by individuals showing optimal age-related changes. A biopsychosocial model provides a useful framework for organizing these models. As translated into successful brain aging, such approaches identify those individuals who maintain superior levels of neurological integrity which, in turn, translates into superior cognitive performance. Personality, health, and sociocultural factors further interact with age-related changes in the brain and cognitive functioning. Future research will continue to expand the knowledge base in this important, and growing, field within life span psychology.

Differential aging refers to the principle that the rate of functional changes in various life domains shows inter- and intra-individual variability. Interindividual variability is shown by the fact that there are differences across individuals in the rate at which changes occur over time. Intraindividual variability is demonstrated by the finding of differences in the rates of change over time in functions within the same individual. The principle of differential aging stands in sharp contrast to the previously held views in developmental psychology that aging is a uniform process of decline, also known as dedifferentiation. Although it is true that death is the inevitable outcome of the aging process, this fact does not rule out the possibility that growth can occur in psychological functions even at the end of life. Plasticity is another component of differential aging, meaning that the potential exists throughout later adulthood for growth and adaptation.

Several distinctions within the aging population already exist, which make it clear that individuals show important differences in their patterns of change over time. One is the division of the 65 and older population into subdivisions of young–old (65–74), old–old (75–84), and oldest–old (85 and older). The second set of divisions of older adults takes patterns of change over time into account. Individuals may show primary (normal) aging and secondary (infirmed) aging. At the very end of life, as their health rapidly decreases, tertiary aging becomes the norm as multiple functions simultaneously decline.

According to the principle of differential aging, the appropriate study of development in later life is on providing an understanding of why people vary in the ways they age. In the area of brain functioning, where much of the differential aging research is now taking place, neuroimaging technology is making it possible to correlate variations in abilities between and within older adults with variations in the rate of brain aging. Statistical methods for modeling individual change over time are becoming the standard for analytic methods in longitudinal and sequential research. The goals of much of this research are not only to reflect more accurately the nature of the aging process, but also to provide the basis for effective interventions to improve the daily lives of older adults.

The Origins of the Principle of Differential Aging

The term differential aging was introduced in the literature by Hans Thomae, one of the original collaborators in Bonn Gerontologic Longitudinal Study (BOLSA). BOLSA was started in 1965 with two cohorts born in 1890–95 and 1900–05, following survivors for 15 and 19 years, respectively. BOLSA data linked patterns of aging defined by health, cognitive and social abilities, coping, and personality with social structural variables including socioeconomic status, gender, and other social variables. Findings from BOLSA, which included predicting longevity from psychosocial variables, reinforced the view that the study of aging requires a multidimensional approach.

The concept of differential aging became a central component of the Berlin Aging Study (BASE), a comprehensive investigation begun in the late 1980s of over 500 adults between 70 and 100 years of age living in Berlin, Germany. This study continued the mission of BOLSA to understand patterns of aging and provide insight into why some people age successfully and some do not. BASE also recognized that successful agers do not show uniformly positive trajectories of change. Thus, the goal of this literature was to understand how individuals balance their gains and losses with age in such a way as to maintain their physical and mental health.

The selective optimization with compensation (SOC) model was developed in the context of BASE as a way to understand the factors that influence successful aging. Selection occurs when individuals actively or passively reduce the number of goals and domains in order to free up energy for investing in new or more important goals. Individuals engage in compensation when they search for and use alternate means to reach their goals once they can no longer use their previous means. BASE showed that older individuals most likely to achieve successful aging by the criteria of maintaining an active lifestyle across 21 indices of everyday functioning were those rich in sensorimotor–cognitive and personality–social resources (Baltes and Lang, 1997).

A follow-up to the original BASE tracked the trajectories of groups defined on the basis of scores on constructs measuring cognition, personality, self-rated functioning, and

social integration (Gerstorf et al., 2006). These scores, obtained over three measurement occasions, were used to define three clusters among a subset of the original sample: an overall positive profile; an overall average profile; and a high cognition, low self, and social profile. Impressively, over the 6-year period, the majority of individuals remained in their relative subgroups. Regardless of which group they were in, though, individuals who changed group membership showed the highest rates of mortality. In a subsequent follow-up conducted 7 years later, however, 'dedifferentiation' occurred. There was less evidence in favor of correlates of differential change as all groups showed a cascadelike descent prior to mortality. Nevertheless, given that individuals vary in dying trajectories, some differentiation may still exist even toward the very end of life.

Differential Aging: Perspectives from Successful Aging

The initial BASE findings make clear the fact that aging progresses at differential rates in part, at least, due to the differential availability of resources as well as the individual's ability to be able to maximize his or her use of those resources. In addition, as was clear in BASE, the concept of differential aging assumes an integrated biopsychosocial perspective that brings together biological, psychological, and sociocultural factors as they impinge upon, and change, across the life span.

The most well-known approach to successful aging was, for many years, the Rowe and Kahn model (1998). This model identifies the three components of successful aging as lack of disease and disability, high cognitive and physical functioning, and engagement with life. The Rowe and Kahn model would imply that adults who age successfully do not have physical or cognitive limitations and are vitally involved with their social context.

Critics of the Rowe and Kahn model argue that it is overly focused being free from infirmities that inevitably accompany aging, even in the healthiest older adults. Moreover, as BASE showed, older adults have differential access to the social resources that can support their ability to age successfully. The model does not account for these contextual factors. Older adults with social support, particularly support from family, are most likely to have high ratings of successful aging. Such support would make it more likely that the older adult would meet the criteria of remaining engaged with life (Reichstadt et al., 2007).

Perhaps for these reasons, researchers employing the Rowe and Kahn criteria find that only a small minority of older adults age successfully. In comparison, the large majority of older adults in the U.S. report high feelings of subjective well-being. The National Health Interview Survey, which tracks the incidence of serious mental illness, reports year after year that adults of 65 years of age and older report lower rates of serious mental illness within the past year than do younger age groups of adults (Substance Abuse and Mental Health Services Administration, 2012).

Thus, the Rowe and Kahn criteria may be too strict in requiring lack of disability to be a criterion of successful aging. In the Health and Retirement Study (HRS), a longitudinal investigation of adults of 51 years of age and older, successful aging was defined as involving no major disease, no limitation in activity of daily living, ability to perform a variety of physical tasks (such as walking one block), high levels of cognitive function, and active engagement in work, volunteer activity, or interactions with family, neighbors, and friends (McLaughlin et al., 2010). The HRS measures were assessed over four time intervals from 1998 to 2004.

The percentage of the HRS sample who met all four criteria for successful aging decreased over the 6 years of the study from about 12 to 11%. Roughly half of the sample met the criteria for successful aging across the entire 6-year period in terms of active engagement, high cognitive functioning, and high physical functioning. Approximately 80% met the criterion of no disability, even by the end of the period. However, only 43% reported having no disease at the beginning of the study, and even fewer (37%) did by the point of the final data collection.

The failure of Rowe and Kahn model to give sufficient attention to social context is supported by the fact that in the HRS, successful agers were more heavily represented among Whites and among those with a college education or higher (McLaughlin et al., 2010). Results from the Survey of Health, Ageing and Retirement in Europe (SHARE) on 22 464 men and women in 13 European countries averaging 63 years of age further reinforces this point (Brandt et al., 2012). In SHARE, early childhood conditions including parental social class and income predicted late life successful aging as measured by the Rowe and Kahn criteria. These findings reinforce the point that differential aging reflects not just current, but also lifelong influences of multiple factors, including favorable or unfavorable childhood conditions. Thus, although Rowe and Kahn count social engagement as a criterion for successful aging, social context conceived more generally needs to be given greater attention.

Further evidence for the role of social context comes from a study of life satisfaction across the adult years in two studies of large national samples showing different age-related patterns. Baird et al. (2010) used self-rated happiness data from the German Socio-Economic Panel Study, begun in 1984 with a full sample of over 40 000 participants who were followed over time, along with new samples added at each testing occasion. The final sample available for analysis consisted of nearly 21 000 adults who had participated an average of eight times. After controlling for cohort, time of measurement, and number of testing occasions, the picture that emerged was of near stability in life satisfaction until the very oldest ages, beginning at the age of 70. However, when they conducted similar analyses on the British Household Panel Study, a different pattern emerged. Life satisfaction among this sample decreased early in adulthood, increased from mid to late adulthood, and then declined at the very end of life. Thus, although the end points were similar in that a decline in life satisfaction emerged in both samples, there were differing trajectories for adults in the two countries.

Social class may play a role in affecting successful aging through its relationship with physical health. In analyses of 12-year longitudinal data from the German Aging Survey (Schöllgen et al., 2012), scores on a measure of negative emotions predicted change in physical health in individuals with lower levels of education. In contrast, physical health was related to change in positive emotions over time independently of educational level.

Taking social context into account is the model of active aging put forth by the World Health Organization (WHO) (2002). The WHO model specifies roles for social, health care, and economic determinants as well as pointing to the importance of the physical environment, thus incorporating into its definition recognition of the fact that older adults may be operating under a number of constraints outside of their own personal control. Noteworthy is the fact that WHO places greater emphasis on the individual's ability to get around in the environment, rather than on whether or not the individual needs physical accommodations due to disability, thereby further broadening the definition of successful aging.

However, neither the Rowe and Kahn and WHO definitions take into account psychological factors such as personality and identity. Vahia et al. (2012), using data from the San Diego site of the Women's Health Initiative, developed a dimensional model of successful aging that went beyond cognition and physical functioning to include psychological resilience, optimism, and self-efficacy. Similarly, Ouwehand et al. (2007) proposed that greater attention is to be paid by successful aging models to proactive methods taken by older adults to adapt to life stresses. Critiquing the SOC model, they note that older adults can in fact achieve their desired goals, rather than having to pare them down, through proactive coping in which they prepare for potential stressors, anticipating events that they believe represent a threat to their well-being and then engaging in mental preparation to reduce the impact of the stress when it occurs. When the stressful event does occur, it will drain less of their coping resources.

Thus, both social context and self-perceptions seem to be important in assessing the factors that influence resilience in the form of successful aging in later life. Models of differential aging must therefore go beyond objective, within-person focused criteria and incorporate the strategies that older adults use to navigate the biopsychosocial effects of aging that might otherwise thwart their abilities to achieve high levels of subjective well-being.

Differential Brain Aging and Successful Cognitive Aging

The concept of successful aging is becoming a central organizing theme in research on brain and cognitive aging. It has long been known that individuals differ in the rate of central nervous system aging which, in turn, is reflected at least in part in differences across and within older adults in their levels of performance on cognitive tasks. Researchers are now attempting to identify the factors that contribute to successful brain aging and, by extension, successful cognitive agers.

An overarching framework in differential brain and successful cognitive aging is provided by scaffolding theory (Park and Bischof, 2013), which provides an elaborate model indicating the interaction among aging, compensatory activities, and the possibility for neural growth in later adulthood. According to this model, normal aging brings with it amyloid accumulation, shrinkage, changes in white matter such as development of hyperintensities, cortical thinning, and depletion of dopamine, which can affect prefrontal function. These structural challenges combine with age-related deterioration of functioning such as losses in the visual area, decreased ability to recruit the medial temporal lobe, involved in memory, and increased default activity, which leads to an increase in interference from background thoughts during effortful cognitive activity.

Counteracting these deleterious changes is scaffolding, a process that allows older adults both to compensate for and even enhance cognitive functioning. Compensatory scaffolding includes recruiting the anterior or frontal regions of the brain to increase its activity to make up for decreased activation of the posterior regions ('PASA'). New neurons can even be generated to replace those lost through cell death ('neurogenesis'). Older adults also seem to show more distributed processing of information and greater bilaterality in which the opposite hemisphere takes over when the original hemisphere suffers losses (Hemispheric asymmetry reduction in older adults, 'HAROLD') (Cabeza, 2002). Older adults may also use the process of scaffolding enhancement in which mental and physical activity as well as engagement contribute to maintain or build upon the individual's existing cognitive resources. For example, older adults with higher levels of physical fitness have enhanced hippocampal volume and spatial memory (Erickson et al., 2009). All of these factors combine to produce the older individual's level of cognitive function. With so many variables at play, it is clear that a number of alternative pathways are possible in the process of brain, and hence cognitive, development in later life.

The search for successful cognitive agers includes several lines of investigation in which objective criteria are used to define what are now being called 'SuperAgers'. One way to identify such individuals is to use statistical criteria defining these high-functioning older adults as being above the average for an individual's age group (Fiocco and Yaffe, 2010). Another approach is to define successful cognitive agers in an absolute sense. Harrison et al. (2012) identified a group of 'SuperAgers', who were individuals of 80 years of age and older with memory performance equivalent to that of middle-aged adults.

The SuperAgers identified by Harrison et al. (2012) were found, on MRIs, to have thicker cerebral cortex volume. Their memory performance, in turn, was correlated positively with the cortical thickness. The SuperAgers also showed greater volume of the cingulate cortex, an area that may be involved in preventing episodic memory loss.

However, the sheer size of a brain region may not be as important as the way that brain region is organized or functions. In a review of 80 studies comparing brain function and cognitive performance in older adults, Eyler et al. (2011) looked for relationships between MRI and PET scan activity in the cortical lobes and subcortical areas involved in movement, sensory processing, and the insula, located deep within the cortex. In 61% of the studies reviewed there was either a mixed relationship, a negative relationship, or no relationship at all between brain aging and cognitive aging. More of the positive associations between brain activity and cognitive performance showed up for the frontal cortex (35%). In the medial temporal lobe, which includes the hippocampus, there were far fewer significant relationships (20%). Consistent with the HAROLD and PASA compensation models, it is possible that the successful brain agers are most likely to show greater abilities to take advantage of brain compensatory scaffolding.

Differential aging also occurs within regions of the brain with association areas, neostriatum and cerebellum more

affected by aging than the sensory cortices and the subcortical area of the pons. Subcortical white matter and hippocampus also show losses in later life, rates of shrinkage that may be accelerated in individuals suffering from hypertension (Raz and Rodrigue, 2006). Differential changes also occur in myelinization versus coherence of fiber organization within the white matter of the brain (Bennett et al., 2010). Variations according to gender are evident in many areas of health and physical functioning, and there is evidence that such differential aging in the brain occurs in the putamen, a structure in the basal ganglia involved in motor function (Nunnemann et al., 2009).

A number of investigations of sensory, perceptual, and cognitive functions and their connection to brain activity, similarly, point to the importance of assessing not only mean differences between age groups but also the existence of intra-individual and interindividual variability. Such functions include, for example, executive functioning (Goh et al., 2012), involved in planning and organizing cognitive operations. Aging has differential effects on regions of the brain, across and within individuals, as well as differential relationships to behavioral performance indicators.

These findings on brain aging are consistent with the work of Schaie and colleagues in the continued investigations based on the Seattle Longitudinal Study (SLS), a program of research using a sequential research design to identify age, cohort, and time of measurement influences on intellectual performance. Overall, the findings reveal a picture of intelligence in later adulthood as one of stabilities until the 50s or 60s, followed by decline through the oldest ages tested. However, there are important qualifications on this conclusion. First, although some individuals may show declines in intelligence by the mid-50s, significant losses do not occur overall until the decade of the 70s. More to the point, changes in functioning do not occur across all abilities, even into the decade of the 80s. There are distinctly different patterns of age changes across the primary abilities. Vocabulary, a measure of crystallized intelligence, showed the least amount of change, holding up fairly steadily until 74 years of age. The largest drop in scores occurred on quantitative knowledge. Moreover, health status also contributes to individual differences in intelligence. Arthritis, cancer, and osteoporosis are health conditions associated with intelligence test scores (Schaie, 1996). These effects are not limited to Western countries; similar findings were reported among a study of more than 3200 Chinese of 60 years of age and older, even after controlling for age, smoking, hypertension, and diabetes (Liu et al., 2009).

Supporting these findings of differential aging patterns are analyses from the University of Manchester Longitudinal Study, conducted over a 20-year period on over 6000 middle-aged and older adults (Ghisletta et al., 2012). In this study, tests measuring fluid intelligence declined at a faster rate than did those of crystallized. However, confirming findings from the SLS, there were large individual differences, particularly during the midlife years. There is also increasing evidence showing a relationship between measures of brain activation and intelligence test performance with fluid intelligence fluctuating to a larger degree among older adults with higher levels of abilities (Ram et al., 2005). Perhaps the individuals at the lower end of the spectrum are more vulnerable to the physiological changes that, in turn, can further erode their fluid abilities in a cascading effect.

Individual differences in intelligence test scores are related to social and cultural factors as well. People with higher levels of education receive higher scores on fluid intelligence measures and, in fact, when controlling for education, even cross-sectional differences in middle adulthood disappear (Ronnlund and Nilsson, 2006). These findings support the WHO model of active aging and its inclusion of social and cultural factors as playing a role in differential aging.

Application to Cognitive Functioning

Replacing the focus on mean differences between younger and older adults on tests of cognitive functioning is increasing emphasis on variations within older groups of adults studied cross-sectionally and variations over time in longitudinal studies of cognition and aging. Why older adults show these variations also is becoming the basis for new theories of aging and cognition. For example, Salthouse in refining his general slowing hypothesis, incorporates more complex models of the effects of aging on speeded performance and verbal intelligence (Bowles and Salthouse, 2008).

Differential aging also becomes relevant to understanding association memory and age. Visual and spatial working memory also appear to show differential age patterns (Oosterman et al., 2011) as does learning of emotional and nonemotional stimuli (Monti et al., 2010), voluntary and involuntary autobiographical memory (Schlagman et al., 2009), and prospective and retrospective memory (Kvavilashvili et al., 2009). Older individuals also differ in their use of strategies in working memory tasks (Bailey et al., 2009). Across the life span, top-down and binding processes in memory show differential patterns of change, with top-down processes showing a greater contribution to working memory in older adults, whereas processes that bind working memories are more subject to age-related declines (Sander et al., 2012).

These are just a few examples of the paradigm shift taking place in research on aging and cognition. Instead of recording differences between groups of younger and older adults, researchers are now examining variation and possible causes for variation within older samples.

Application to Personality and Social Functioning

As noted above, identity and personality must be taken into account in any model of differential aging. Not only do individuals possess differing degrees of resilience and sense of control, but also their life experiences contribute to their overall sense of self, experiences best understood from a bio-psychosocial perspective.

First, identity influences and is influenced by the challenges associated with negotiating the aging process (Whitbourne et al., 2002). Individuals show differing degrees of incorporating age-related physical and cognitive changes into their identities through their use of identity processes. They may use identity assimilation, in which they attempt to minimize or even deny the extent to which they are undergoing these changes. In part, identity assimilation may be a defensive way to ward off the negative stereotypes of aging present in Western society (Sneed and Whitbourne, 2003). Conversely, through identity accommodation, older adults incorporate these

changes into their identities, potentially to the point of believing that they cannot ever live the life to which they have become accustomed. The process of identity accommodation may itself create further problems, such as when older adults begin to believe that they do not have control over their own aging (Jones et al., 2009). Ideally, however, older adults achieve a sense of balance between these processes in which they are able to acknowledge the reality of the fact that they are getting older without becoming despondent. Cross-national studies show differences in the extent to which older adults use these processes reflecting, in part, the social and economic environment in which they are aging (Westerhof et al., 2012). Moreover, they may change in their relative use of these processes in a dynamic way as they confront each new age-related change in functioning (Whitbourne and Collins, 1998).

Looking next at personality, studies of aging based on the scales of the Five Factor Model show a high degree of consistency over time throughout adulthood, with greater consistency among increasingly older groups of adults. This means that people maintain their relative positions along the traits in comparison to their age peers. However, the consistency between scores obtained at different measurement points becomes smaller as the time interval between them becomes longer. Although people may maintain their relative positions along each of the traits, their mean scores may undergo change across adulthood. Roberts et al. (2006), based on a meta-analysis of 92 longitudinal studies, concluded that individuals increase in social dominance, conscientiousness, emotional stability, social vitality, and openness to experience through 40 years of age; after that point, social vitality and openness decreased after this point, and agreeableness increased.

Personality may also play a role in the maintenance or decline of cognitive abilities. Individuals who feel confident in their abilities are liberal and autonomous in attitudes, and have an open-minded approach to new experiences, thoughts, and feelings also have higher intelligence test scores in later adulthood (Schaie et al., 2004). In contrast, older adults who are higher on trait anxiety perform more poorly on a variety of fluid and crystallized tasks (Wetherell et al., 2002).

People higher in openness to experience when young, may benefit from a lifetime of curiosity and willingness to entertain new ideas. In the Aberdeen Scotland Birth Cohort of 1936, longitudinal study begun when participants were children, those high in this trait had higher scores on measures of reading ability in the decade of the 60s (Hogan et al., 2012). Similarly, in an Australian study on adults of 74–90 years of age, participants higher on one aspect of openness to experience (enjoyment of fantasy) had superior intelligence and everyday functioning test scores (Gregory et al., 2010). However, in both of these studies, personality alone did not serve to predict intellectual functioning. In the Aberdeen study, childhood intelligence was a stronger predictor of cognitive ability in later adulthood than personality openness. The Australian study showed that fluid intelligence interacted with personality openness in influencing cognitive abilities. It is possible that people with higher intelligence are higher in personality openness, which allows them to achieve superior cognitive performance.

These studies raise the question of directionality in the cognition–personality relationship. The intellectually more able may seek out more stimulating environments, or being in an environment rich in cognitive stimulation may lead to greater preservation of cognitive abilities. Perhaps the successful cognitive agers purposefully search for ways to maximize their abilities by seeking complexity in their daily activities. In research from the Zurich Longitudinal Study on Cognitive Aging, a factor formed by items assessing reading, problem solving, abstract thinking, and intellectual curiosity in everyday life predicted higher crystallized intelligence scores (Dellenbach and Zimprich, 2008).

Suggestions for Future Research

Despite the existence of the principle of differential aging, many researchers within psychological gerontology and life span psychology continue to focus their efforts on the search for consistent patterns of age differences that characterize older adults as a homogenous group. However, as indicated by the studies reviewed here, the principle of differential aging is starting to take hold as a framework for understanding inter- and intraindividual variation in a range of abilities. Moreover, new models that take into account the broad range of factors influencing well-being in older adults, such as WHO's focus on active aging, will broaden and deepen scientific understanding of the contribution of biopsychosocial processes or successful aging.

Summary

The principle of differential aging originated from broadscale investigations of intra- and interindividual variations in patterns of aging. Operationalizing differential aging as successful aging links this principle to efforts to identify contributions to successful aging which, in turn, is increasingly being understood as a biopsychosocial process.

Within the past 10 years, studies in the fields of aging within neuroscience, cognition, and personality are identifying patterns of differential changes across the adult age span as well as differential changes within individuals in particular types of abilities. Moreover, these studies either directly provide linkages among biological, psychological, and sociocultural factors or allow for interpretations that bring multiple disciplines into the equation.

It is clear that differential aging will continue to serve as an organizing principle and provide a rich conceptual model to inform future research and theory in gerontology and life span development.

See also: Aging and Health in Old Age; Early Life Influences on Health and Mortality in Adulthood; Mortality, Biodemography of; Old Age: Definitions, Theory, and History of the Concept.

Bibliography

Bailey, H., Dunlosky, J., Hertzog, C., 2009. Does differential strategy use account for age-related deficits in working-memory performance? Psychology and Aging 24, 82–92. http://dx.doi.org/10.1037/a0014078.

Baird, B.M., Lucas, R.E., Donnellan, M.B., 2010. Life satisfaction across the lifespan: findings from two nationally representative panel studies. Social Indicators Research 99, 183–203. http://dx.doi.org/10.1007/s11205-010-9584-9.

Baltes, M.M., Lang, R.F., 1997. Everyday functioning and successful aging: the impact of resources. Psychology and Aging, 12, 433–443.

Bennett, I.J., Madden, D.J., Vaidya, C.J., Howard, D.V., Howard Jr, J.H., 2010. Age-related differences in multiple measures of white matter integrity: a diffusion tensor imaging study of healthy aging. Human Brain Mapping 31, 378–390.

Bowles, R.P., Salthouse, T.A., 2008. Vocabulary test format and differential relations to age. Psychology and Aging 23, 366–376. http://dx.doi.org/10.1037/0882-7974.23.2.366.

Brandt, M., Deindl, C., Hank, K., 2012. Tracing the origins of successful aging: the role of childhood conditions and social inequality in explaining later life health. Social Science & Medicine 74, 1418–1425. http://dx.doi.org/10.1016/j.socscimed.2012.01.004.

Cabeza, R., 2002. Hemispheric asymmetry reduction in older adults: the HAROLD model. Psychology and Aging 17, 85–100. http://dx.doi.org/10.1037/0882-7974.17.1.85.

Dellenbach, M., Zimprich, D., 2008. Typical intellectual engagement and cognition in old age. Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition 15, 208–231. http://dx.doi.org/10.1080/13825580701338094, 781708432 [pii].

Erickson, K.I., Prakash, R.S., Voss, M.W., Chaddock, L., Hu, L., Morris, K.S., Kramer, A.F., 2009. Aerobic fitness is associated with hippocampal volume in elderly humans. Hippocampus 19, 1030–1039. http://dx.doi.org/10.1002/hipo.20547.

Eyler, L.T., Sherzai, A., Kaup, A.R., Jeste, D.V., 2011. A review of functional brain imaging correlates of successful cognitive aging. Biological Psychiatry 70, 115–122. http://dx.doi.org/10.1016/j.biopsych.2010.12.032.

Fiocco, A.J., Yaffe, K., 2010. Defining successful aging: the importance of including cognitive function over time. Archives of Neurology 67, 876–880. http://dx.doi.org/10.1001/archneurol.2010.130.

Gerstorf, D., Smith, J., Baltes, P.B., 2006. A systemic-wholistic approach to differential aging: longitudinal findings from the Berlin Aging Study. Psychology and Aging 21, 645–663. http://dx.doi.org/10.1037/0882-7974.21.4.645.

Ghisletta, P., Rabbitt, P., Lunn, M., Lindenberger, U., 2012. Two thirds of the age-based changes in fluid and crystallized intelligence, perceptual speed, and memory in adulthood are shared. Intelligence 40, 260–268. http://dx.doi.org/10.1016/j.intell.2012.02.008.

Goh, J.O., An, Y., Resnick, S.M., 2012. Differential trajectories of age-related changes in components of executive and memory processes. Psychology and Aging 27, 707–719. http://dx.doi.org/10.1037/a0026715.

Gregory, T., Nettelbeck, T., Wilson, C., 2010. Openness to experience, intelligence, and successful ageing. Personality and Individual Differences 48, 895–899. http://dx.doi.org/10.1016/j.paid.2010.02.017.

Harrison, T.M., Weintraub, S., Mesulam, M.M., Rogalski, E., 2012. Superior memory and higher cortical volumes in unusually successful cognitive aging. Journal of the International Neuropsychological Society 18, 1081–1085. http://dx.doi.org/10.1017/s1355617712000847.

Hogan, M.J., Staff, R.T., Bunting, B.P., Deary, I.J., Whalley, L.J., 2012. Openness to experience and activity engagement facilitate the maintenance of verbal ability in older adults. Psychology and Aging 27, 849–854. http://dx.doi.org/10.1037/a0029066.

Jones, K.M., Whitbourne, S., Whitbourne, S.B., Skultety, K.M., 2009. Identity processes and memory controllability in middle and later adulthood. Journal of Applied Gerontology 28, 582–599. http://dx.doi.org/10.1177/0733464808330823.

Kvavilashvili, L., Kornbrot, D.E., Mash, V., Cockburn, J., Milne, A., 2009. Differential effects of age on prospective and retrospective memory tasks in young, young-old, and old-old adults. Memory 17, 180–196. http://dx.doi.org/10.1080/09658210802194366.

Liu, C.Y., Zhou, H.D., Xu, Z.Q., Zhang, W.W., Li, X.Y., Zhao, J., 2009. Metabolic syndrome and cognitive impairment amongst elderly people in Chinese population: a cross-sectional study. European Journal of Neurology 16, 1022–1027. http://dx.doi.org/10.1111/j.1468-1331.2009.02640.x. ENE2640 [pii].

McLaughlin, S.J., Connell, C.M., Heeringa, S.G., Li, L.W., Roberts, J.S., 2010. Successful aging in the United States: prevalence estimates from a national sample of older adults. The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences 65B, 216–226. http://dx.doi.org/10.1093/geronb/gbp101.

Monti, J.M., Weintraub, S., Egner, T., 2010. Differential age-related decline in conflict-driven task-set shielding from emotional versus non-emotional distractors. Neuropsychologia 48, 1697–1706. http://dx.doi.org/10.1016/j.neuropsychologia.2010.02.017.

Nunnemann, S., Wohlschläger, A.M., Ilg, R., Gaser, C., Etgen, T., Conrad, B., Mühlau, M., 2009. Accelerated aging of the putamen in men but not in women. Neurobiology of Aging 30, 147–151. http://dx.doi.org/10.1016/j.neurobiolaging.2007.05.016.

Oosterman, J.M., Morel, S., Meijer, L., Buvens, C., Kessels, R.P.C., Postma, A., 2011. Differential age effects on spatial and visual working memory. The International Journal of Aging & Human Development 73, 195–208. http://dx.doi.org/10.2190/AG.73.3.a.

Ouwehand, C., de Ridder, D.T.D., Bensing, J.M., 2007. A review of successful aging models: proposing proactive coping as an important additional strategy. Clinical Psychology Review 27, 873–884. http://dx.doi.org/10.1016/j.cpr.2006.11.003.

Park, D.C., Bischof, G.N., 2013. The aging mind: neuroplasticity in response to cognitive training. Dialogues in Clinical Neuroscience 15, 109–119.

Ram, N., Rabbitt, P., Stollery, B., Nesselroade, J.R., 2005. Cognitive performance inconsistency: intraindividual change and variability. Psychology and Aging 20, 623–633.

Raz, N., Rodrigue, K.M., 2006. Differential aging of the brain: patterns, cognitive correlates and modifiers. Neuroscience and Biobehavioral Reviews 30, 730–748. http://dx.doi.org/10.1016/j.neubiorev.2006.07.001.

Reichstadt, J., Depp, C.A., Palinkas, L.A., Folsom, D.P., Jeste, D.V., 2007. Building blocks of successful aging: a focus group study of older adults' perceived contributors to successful aging. American Journal of Geriatric Psychiatry 15, 194–201.

Roberts, B.W., Walton, K.E., Viechtbauer, W., 2006. Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. Psychological Bulletin 132, 1–25. http://dx.doi.org/10.1037/0033-2909.132.1.1.

Ronnlund, M., Nilsson, L.-G., 2006. Adult life-span patterns in WAIS-R block design performance: cross-sectional versus longitudinal age gradients and relations to demographic factors. Intelligence 34, 63–78.

Rowe, J.W., Kahn, R.L., 1998. Successful Aging. Pantheon Books, New York.

Sander, M.C., Lindenberger, U., Werkle-Bergner, M., 2012. Lifespan age differences in working memory: a two-component framework. Neuroscience and Biobehavioral Reviews 36, 2007–2033. http://dx.doi.org/10.1016/j.neubiorev.2012.06.004.

Schaie, K.W., 1996. Intellectual development in adulthood. In: Birren, J.E., Schaie, K.W., Abeles, R.P., Gatz, M., Salthouse, T.A. (Eds.), Handbook of the Psychology of Aging, fourth ed. Academic Press, San Diego, CA, pp. 266–286.

Schaie, K.W., Willis, S.L., Caskie, G.I., 2004. The Seattle Longitudinal Study: relationship between personality and cognition. Aging, Neuropsychology and Cognition 11, 304–324.

Schlagman, S., Kliegel, M., Schulz, J., Kvavilashvili, L., 2009. Differential effects of age on involuntary and voluntary autobiographical memory. Psychology and Aging 24, 397–411. http://dx.doi.org/10.1037/a0015785.

Schöllgen, I., Huxhold, O., Schmiedek, F., 2012. Emotions and physical health in the second half of life: interindividual differences in age-related trajectories and dynamic associations according to socioeconomic status. Psychology and Aging 27, 338–352. http://dx.doi.org/10.1037/a0026115.

Sneed, J.R., Whitbourne, S.K., 2003. Identity processing and self-consciousness in middle and later adulthood. Journals of Gerontology Series B: Psychological Sciences and Social Sciences 58, P313–P319.

Substance Abuse and Mental Health Services Administration, 2012. Results from the 2010 National Survey on Drug Use and Health: Mental Health Detailed Tables. Substance Abuse and Mental Health Services Administration, Rockville, MD.

Vahia, I.V., Thompson, W.K., Depp, C.A., Allison, M., Jeste, D.V., 2012. Developing a dimensional model for successful cognitive and emotional aging. International Psychogeriatrics 24, 515–523. http://dx.doi.org/10.1017/s1041610211002055.

Westerhof, G.J., Whitbourne, S.K., Freeman, G.P., 2012. The aging self in a cultural context: the relation of conceptions of aging to identity processes and self-esteem in the United States and the Netherlands. The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences 67B, 52–60. http://dx.doi.org/10.1093/geronb/gbr075.

Wetherell, J.L., Reynolds, C.A., Gatz, M.P., Nancy, L., 2002. Anxiety, cognitive performance, and cognitive decline in normal aging. Journals of Gerontology Series B: Psychological Sciences and Social Sciences 57B, P246–P255.

Whitbourne, S.K., Collins, K.C., 1998. Identity and physical changes in later adulthood: theoretical and clinical implications. Psychotherapy 35, 519–530.

Whitbourne, S.K., Sneed, J.R., Skultety, K.M., 2002. Identity processes in adulthood: theoretical and methodological challenges. Identity 2, 29–45.

World Health Organization, 2002. Active Ageing: A Policy Framework. World Health Organization, Geneva, Switzerland.

Differential Diagnosis in Psychiatry

Umberto Volpe, University of Naples SUN, Naples, Italy

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.E. Kendell, volume 6, pp. 3658–3663, © 2001, Elsevier Ltd.

Abstract

Differential diagnosis may be defined as the application of a systematic diagnostic method to distinguish one disease or condition from multiple candidate conditions, presenting with similar sign and symptoms. However, in psychiatry, differential diagnosis is more complex than in other branches of medicine, given that the etiology of mental and behavioral disorders is still largely unknown and no biological tests or instrumental diagnostic procure is available, yet. Along the past four decades, the introduction of operational criteria has increased the reliability of psychiatric diagnoses, but it has also let arise many clinical problems (ranging from the inflation of psychiatric comorbidity to the loose boundaries between different diagnostic categories, from the gradual abandoning of the concept of disorders' spectrum to the issue of diagnostic hierarchy).

The etymology of word 'diagnosis' (which derives from the two Greek words *dia*, meaning 'through,' and *gnosis*, meaning 'knowledge') drives any clinician to the idea that, prior to treat any human suffering, a journey through the patient's sign and symptoms is essential. In some cases, the clinical presentation is often seen and easily recognized; in very few cases, one symptom or sign may be specific for a single pathological condition (i.e., a 'pathognomonic' signs or symptoms); in the former cases, the diagnosis is relatively a simple procedure. However, most commonly, similar clinical presentations may be caused by different etiologies and thus respond differently to different treatments; these cases have to be systematically evaluated to identify a single pathological entity, among multiple alternatives. Differential diagnosis may be defined as the application of a systematic diagnostic method to distinguish on disease or condition from multiple candidate conditions, presenting with similar sign and symptoms. Actually, as simple as they may seem, the great majority of diagnoses are the result of a differential diagnostic process.

In epistemic terms, a differential diagnosis procedure is essentially a specific application of the hypothetic-deductive method; in practical terms, it is a process of elimination which, based on the information obtained, allows the investigator to reduce the probability of the candidate conditions to negligible levels, except for the final diagnosis.

Similarly, in medicine, the concept of differential diagnosis refers to a systematic method of investigation and stays as a core medical skill, because it represents the preliminary foreword to any focused, specific, and effective treatment. For example, a patient complaining of recurrent upper abdominal pain may probably suffer from a hiatus hernia (an anatomical defect allowing regurgitation of stomach contents into the gullet), but the symptoms may be cardiac in origin (angina), or the result of a gastric or duodenal ulcer, or even of a gastric cancer, and thus it is important to exclude these possible alternative diagnoses before treatment starts.

In medicine, differential diagnosis follows four basic steps: (1) gathering information about patient's history, signs, and symptoms; (2) listing all possible candidate conditions for the observed signs and symptoms; (3) prioritizing the conditions (usually, based both on dangerousness of the condition and on its frequency of presentation, so that the most urgent and frequent diagnoses stay at the top of the list); (4) ruling out the false clinical hypotheses.

In principle, differential diagnosis in psychiatry is no different from that in any other branch of medicine. In practice, however, there are a number of important differences, because of our limited understanding of the causes (i.e., the 'etiology') of most mental disorders. A medical diagnosis is often achieved by means of biological tests or instrumental investigations: the access to *in vivo* information allows the clinician to perform the fourth step of the differential diagnostic process with greater confidence and to rule out many conditions easily. On the contrary, a psychiatric diagnosis is usually determined by the patient's symptoms and behavior, and by their temporal evolution. Indeed, only few psychiatric disorders (mainly, dementias) are defined by their neuropathology, but as this can only be determined after death and thus even such diagnoses have to be based on the presenting clinical syndromes. Until the understanding of the etiology of psychiatric disorders increases, differential diagnosis in psychiatry will imply significant differences and peculiarities with respect to that of most branches of medicine.

The Logic of Differential Diagnosis in Psychiatry

In the seventeenth century, all medical diseases were defined just by their clinical syndromes (i.e., by the symptoms the patient complained of, the clinical signs elicited by the physician on examination, and the evolution over time of these symptoms and signs). Over the last 400 years, however, increasing understanding of the etiology of most of these diseases and the availability of newer technologies allowing to explore *in vivo* functioning of the human body has resulted in their being redefined at some more fundamental level. Smallpox is now defined, not by its characteristic rash, high fever, and mortality, but by the presence of the variola virus; cancer of the stomach is defined, not by the patient's rapidly worsening indigestion and palpable abdominal mass, but by the histopathology (i.e., detailed microscopical structure) of the tumor; thyrotoxicosis is defined, not by the patient's goiter, weight loss, and bulging eyes, but by a raised blood level of the thyroid hormone thyroxin, and so on. As a result, the diagnosis

of most diseases depends on the results of laboratory investigations. Although the physician usually has a shrewd idea which diagnoses are likely after examining the patient, the provisional clinical diagnosis needs to be confirmed by the laboratory: a diagnosis of smallpox is only provisional, until the variola virus has been identified; a diagnosis of cancer of the stomach has to be confirmed by histological examination of the tumor; and, a diagnosis of thyrotoxicosis has to be confirmed by demonstrating a raised blood level of thyroxin.

The difference between psychiatry and the rest of medicine is not absolute, however. Several neurological disorders, such as migraine and torticollis, are still defined by their syndromes, and so too are some skin disorders, such as pityriasis rosea. Conversely, some neuropsychiatric disorders, such as Alzheimer's disease, vascular dementia, or Creutzfeldt-Jakob disease, are defined by their neuropathology. Their defining characteristics are comparable, in other words, with those of cancer of the stomach. But even here there is an important practical difference. The histopathology of a gastric tumor can be determined by taking a biopsy through a gastroscope, or at operation, whereas the histopathology of the brain can normally only be determined *post mortem*, and this means that the diagnosis on which treatment is founded still has to be based on the clinical syndrome.

Yet at present, there are only a few neuropsychiatric disorders that are defined by the presence of a fundamental abnormality, which can be identified or excluded during life. Down's and Klinefelter's syndromes are both defined by the presence of chromosomal abnormalities: the former by an additional chromosome 21 and the latter by an additional X chromosome. Huntington's chorea and the Fragile X syndrome are both defined by abnormal genes at specific sites, and phenylketonuria (another important cause of mental retardation) is defined by a specific biochemical abnormality. However, besides the above mentioned chromosomal abnormalities, the understanding of the etiology of most psychiatric disorders is still rather rudimentary and they still have to be defined by their clinical syndromes.

In psychiatry, diagnosis, and hence differential diagnosis, is determined by the patient's symptoms (abnormal or distressing subjective experiences reported by the patient) and behavior (observed or reported by others), and the temporal evolution of these abnormalities. Apart from age and sex, information of other kinds, particularly from laboratory and radiological investigations, are rarely relevant; and even if they are relevant, they are not decisive. Eventually, of course, as understanding of their causation develops, most psychiatric disorders will either be redefined in terms of an underlying abnormality of some kind, or be replaced by new diagnostic concepts defined in this way. At that stage differential diagnosis in psychiatry will probably be no different from that in any other branch of medicine. Until then, however, problems of differential diagnosis have to be handled differently.

History of Differential Diagnosis in Psychiatry

According to the Merriam-Webster medical dictionary, the term 'differential diagnosis' has been introduced around 1860. In psychiatry, the use of a systematic differential diagnostic method was first suggested by Emil Kraepelin, in early decades of the twentieth century, in the attempt to fragment the unitary concept of 'psychosis' into his classical dichotomy between the two major psychiatric syndromes (i.e., dementia praecox and manic-depressive illness). Kraepelin overtly declared that he wanted to abandon the classical 'symptomatic' view of mental illness to embrace a more 'clinical' method for classifying and diagnosing insanity. This probably represented the first attempt, in medicine, to overcome the crucial role of external clinical signs and subordinate it to consideration of the conditions of origin, the course, and the terminus which result from individual disorders.

Along the first half of the twentieth century, research in psychopathology continuously addressed the issue of classifying and differentially diagnosing psychiatric conditions. However, the impact of such research on clinical practice was rather small and most psychiatrists diagnosed their patients mostly based on their *gestalt*, or clinical impressions. Thus it is not surprising that, before the 1970s, psychiatric diagnoses were considered as alarmingly imprecise. There was some broad agreement on the core symptoms of the major syndromes, and this was reflected internationally in textbook descriptions. It was generally agreed, for example, that the characteristic symptoms of schizophrenia were thought disorder, flattening of affect and delusions, and hallucinations of particular kinds, and that some patients also exhibited other typically schizophrenic symptoms such as apathy and stereotypies; there was no agreement, however, on which combinations of these symptoms were adequate to establish a diagnosis of schizophrenia. Indeed, that question was almost never posed, and different schools of psychiatry and individual psychiatrists made different assumptions, and in practice trainee psychiatrists learnt how to diagnose syndromes such as schizophrenia and mania largely by modeling themselves on their teachers. As a result, several studies of the reliability of psychiatric diagnoses in the 1960s found the reliability of the diagnoses even of experienced psychiatrists to be distressingly low. As a consequence, most psychiatric syndromes appeared to merge into one another, or into normality, with no point of rarity in between. This means that the boundaries between syndromes (as well as those between disorder and normality) could have been considered rather arbitrary and that even small changes in the position of those boundaries, produced by minor changes in definition, may result in a change of diagnosis for a substantial number of subjects, both in prevalence surveys and in clinical practice. Problems in differential diagnosis were therefore very common, at that time, and probably insoluble without prior agreement on definitions (which, in the absence of any external validation, were necessarily arbitrary).

In order to scientifically address the issue of psychiatric classification and differential diagnosis, along the 1970s, the US/UK Diagnostic Project and the World Health Organization's International Pilot Study of Schizophrenia confirmed major between-country differences in the usage of key diagnostic terms such as 'schizophrenia' and 'manic-depressive illness,' and even between different cities or teaching centers in the same country. As soon as this chaotic situation became apparent it was agreed by the clinical research community that operational definitions were required, at least for all the major

syndromes (*see* Nosology in Psychiatry). Essentially, an operational, or semantic, definition is a statement or set of statements that spells out precisely which combinations of symptoms and behaviors are adequate to establish the diagnosis in question.

The Use of Operational Criteria for Differential Diagnosis in Psychiatry

The first set of 15 operational definitions for use in clinical research was published by the Department of Psychiatry of St. Louis University in the USA (Feighner et al., 1972). Other rival sets quickly followed: in the 1970s, the two major psychiatric diagnostic systems (namely, the *Diagnostic and Statistical Manual of Mental Disorders* – DSM, by the American Psychiatric Association, APA; and, the *International Classification of Diseases* – ICD, by the WHO) adopted a set of operational definitions of psychiatric syndromes. In the third edition of DSM (DSM-III; APA, 1980), the APA provided operational definitions for more than 200 psychiatric syndromes and demonstrated (by means of the field trials conducted before the glossary was published) that these operational definitions made it possible to achieve much higher levels of reliability than had previously been attained. Subsequently, the WHO provided similar operational definitions for most of the mental disorders listed in the tenth revision of the *International Classification* (*ICD-10*) (WHO, 1993).

The general recognition, at least by clinical research workers and epidemiologists, of the importance of operational and standardized diagnostic criteria in clinical psychiatry had a great impact on psychiatric diagnosis and differential diagnosis. The incorporation of precise diagnostic criteria sets into the major diagnostic systems aimed to improve the reliability of psychiatric diagnoses and to eliminate (or at least make overt) what were previously widespread, unrecognized differences in the diagnostic process. Reliability is now much higher and the unrecognized differences in diagnostic criteria between different countries and schools, which had caused so much confusion previously, have either been eliminated or, if different operational definitions are being used, the differences are now overt and explicit. However, the standardized operational diagnostic approach was originally conceived for research purposes. When applied to psychiatric clinical settings, operational diagnostic methods also created some controversies and deeply influenced the differential diagnostic process in psychiatry.

The Consequences of Definition by Operational Criteria

There are a number of fundamental problems arising from the fact that most mental disorders are defined by their clinical syndromes. The most important is that there is no scientific way of deciding which of two or more alternative operational definitions of a single disorder is correct, or even which is the best. In the case of most psychiatric disorders, there is no comparable external criterion to appeal to, because the clinical syndrome is itself the defining characteristic; this is why alternative operational definitions of disorders such as schizophrenia and agoraphobia remain in widespread use, side by side, and it is commonplace for individual patients to meet one of these definitions but not the other. A patient with typical schizophrenic symptoms for the past 6 weeks, for example, might well meet the formal criteria for schizophrenia in *ICD-10*, but not the corresponding criteria in later edition of *DSM* (DSM-5; APA, 2013); because the latter require evidence of continuous signs of illness for at least 6 months. In practice, this means that discussions about diagnosis need to take place in a context in which it is agreed which diagnostic system is being used, and also with a clear recognition of the fact that the choice between alternative sets of definitions is arbitrary.

Single versus Multiple Diagnoses

In medicine, as a whole, an attempt is generally made to explain all the patient's symptoms with a single diagnosis, and to try to avoid multiple concurrent diagnoses. The rationale for this is that several important diseases (e.g., infections, such as tuberculosis and syphilis) can affect many organs in the body and the correct diagnosis is more likely to be made if the physician consciously strives to account for all the patient's apparently unrelated problems with a single unifying explanation. In fact, people often have several different things wrong with them simultaneously, particularly as they age.

The same attempt to avoid multiple diagnoses is apparent in psychiatric practice, and the underlying rationale is similar. Patients are commonly given two diagnoses simultaneously if one is a lifelong state (such as mental retardation or a personality disorder) and the other an illness with an identifiable onset and a distinctive course (such as schizophrenia or depression). A diagnosis of some form of substance misuse is also commonly combined with a wide range of other diagnoses. There is a deep-seated reluctance, however, to diagnose a neurotic or stress-related disorder in the presence of psychotic symptoms such as delusions and hallucinations, despite the fact that psychotic and neurotic symptoms commonly coexist, and there is a similar reluctance to diagnose any kind of nonorganic psychosis if there is evidence of any brain disease. There is an even greater reluctance to diagnose two psychotic disorders simultaneously, or a mood disorder and a neurotic disorder simultaneously, despite the fact that depression and anxiety commonly coexist.

Hierarchy of Diagnoses

The reluctance to make multiple diagnoses corresponds to the attempt to set, in psychiatry, a diagnostic hierarchy. Karl Jaspers, more than a century ago, first tried to delineate such a hierarchy: in his *General Psychopathology* (Jaspers, 1913), he considered that only psychological disorders accompanying known cerebral pathology were "true disease entities," and within the group of schizophrenia and affective psychoses the diagnosis "has no specific alternative character. Either it is clear as a whole or the differential diagnosis over details determines nothing." To date, organic disorders still come at the top of psychiatric diagnostic hierarchy. This means that if the patient has clear evidence of

a brain disease – a tumor, for example – it is legitimate for them to have a wide range of other psychiatric symptoms, psychotic or neurotic, without any other diagnosis being required. Next in the hierarchy, come the so-called functional or nonorganic psychoses. If the criteria for one of these disorders – schizophrenia, for example – are met, the patient may also be depressed or anxious, or have phobic or obsessional symptoms, without any additional diagnosis being required. They must not, however, have evidence of overt brain disease. Neurotic and stress-related disorders usually come at the bottom of the diagnostic hierarchy, which means that diagnoses at this level cannot normally be made if psychotic symptoms or evidence of brain disease are also present. It is no accident, either, that this is the sequence in which disorders are listed in contemporary diagnostic manuals, and also the sequence of the decisions required by the flowcharts of computer programs for generating diagnoses from clinical ratings.

The requirements of this hierarchy are only relevant if the symptoms in question all appear at much the same time. But, if two groups of symptoms have a quite different time course from one another, it may be both legitimate and appropriate to make two independent diagnoses. In the past the diagnostic hierarchy used to be implicit rather than explicit. Now it is built into the operational definitions of individual disorders. At the same time, it has become more complicated, particularly in situations where schizophrenic and affective symptoms coexist, or follow one another in quick succession.

It is also still uncertain whether the hierarchy is simply an artifact, a device for reducing multiple diagnoses to a minimum, or whether it is a reflection of a fundamental characteristic of psychiatric symptomatology. It is certainly true that a high proportion of people with psychiatric symptoms, but not all, move up and down the hierarchy as they relapse and remit. For example, phobic and obsessional symptoms almost invariably become more widespread and severe if depressive symptoms develop; similarly, the development of hallucinations and delusions is usually accompanied by affective and neurotic symptoms which do not wane until after the hallucinations and delusions have disappeared.

The Boundary Problem

There is an old axiom that the art of classification consists in 'carving nature at the joints.' This is an elegant way of saying that a good, or a valid, nosology places the boundaries between adjacent disorders, and between these disorders and health, where there are genuine discontinuities, either in symptomatology or in etiology. If the defining characteristic of the disorder is a fundamental biological abnormality of some kind it may be self-evident that a discontinuity is involved. An additional chromosome or an abnormal gene, for example, clearly involve qualitative deviations from normality. If, however, the defining characteristic of the disorder is still its syndrome, or a quantitative deviation of some kind, such as a raised blood pressure or a raised blood level of thyroxin, it is not self-evident that there is a valid boundary between the syndrome in question and normality, or between this syndrome and others. Indeed, the fact that there are several rival operational definitions for many mental disorders, each of them encompassing slightly different populations of patients, strongly suggests that it is not obvious where the boundaries of those syndromes should be drawn.

Ideally, a 'point of rarity' between each syndrome and its neighbors needs to be demonstrated (Kendell, 1975). For example, it is widely assumed that schizophrenic and affective psychoses are distinct conditions. Indeed, the distinction Kraepelin drew at the end of the nineteenth century is still regarded as a fundamental feature of our contemporary classification of mental disorders; however, the so-called 'schizoaffective' patients (i.e., patients who present with a mixture of schizophrenic and affective symptoms) are quite common does not prove that there is no genuine boundary between the two syndromes, provided that these mixed forms are less common than the pure forms. The 'grays' must be less common than the 'blacks' and the 'whites'; or to express the issue graphically, the distribution of symptomatology along an axis extending from typical schizophrenia through schizoaffective states to typical affective psychoses must be bimodal rather than unimodal (see Figure 1).

An alternative to demonstrating discontinuities in the distribution of symptomatology is to identify a discontinuity in the relationship between symptomatology and some other important variable, such as prognosis or response to treatment. For example, it is well established that, on average, the outcome of affective psychoses is much better than that of schizophrenic illnesses, both in the short and the long term. The key issue, however, is not whether the two disorders have a significantly different outcome, but whether, when the variation in symptomatology is portrayed as a linear dimension (as described above) there is a discontinuity at some point in the relationship between symptomatology and outcome, or whether there is simply a linear improvement in outcome as the symptomatology becomes increasingly more affective (see Figure 2).

To date, most attempts to demonstrate valid and clear-cut boundaries by the psychiatric diagnostic categories failed and yet there is no empirical evidence at all for natural boundaries between major psychiatric syndromes and no one has ever found a set of symptoms, signs, or tests that separate mental

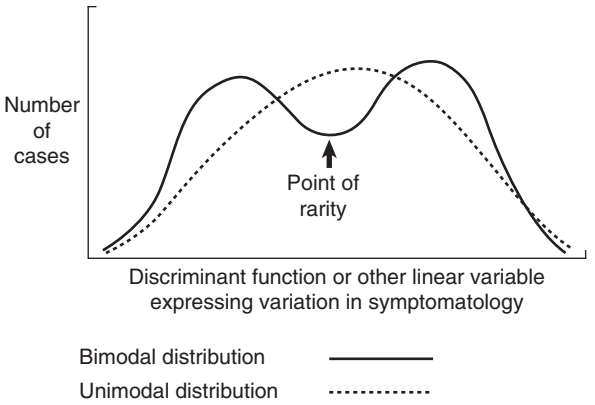

Figure 1 Variation in symptomatology expressed by a linear variable. Reproduced by permission of Royal College of Psychiatrists from the British Journal of Psychiatry 1980, 137: 325.

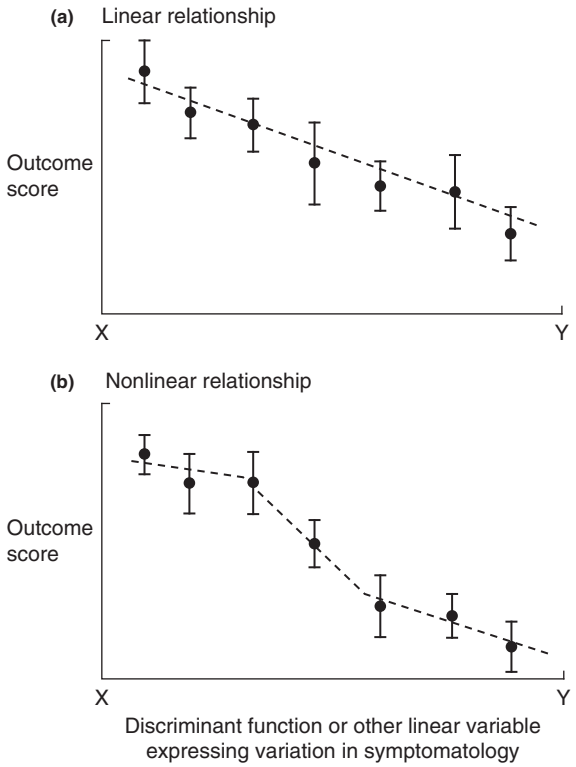

Figure 2 Relationship between symptomatology and outcome when symptomatology is converted to a linear variable. Reproduced by permission of Royal College of Psychiatrists from the British Journal of Psychiatry 1980, 137: 325.

disorders into nonoverlapping categories. For example, the attempts to demonstrate a valid boundary between endogenous and reactive depressions were unsuccessful; when representative populations of patients were studied the distribution of individual patients along the discriminant function was consistently unimodal, rather than bimodal; partly for this reason, contemporary diagnostic manuals such as *ICD-10* (WHO, 1993) and *DSM-5* (APA, 2013) no longer include such a distinction. Similar attempts to demonstrate a valid boundary between schizophrenic and affective psychoses, either in symptomatology or in the relationship between symptomatology and outcome, were also unsuccessful (but a bimodal distribution of scores was subsequently demonstrated on a linear function discriminating between schizophrenia and all other psychiatric disorders combined). Finally, the attempt to identify a valid boundary between major depression and normality by demonstrating a discontinuity in the relationship between depressive symptomatology and either outcome or concordance between 1000 twin pairs also ended in failure (Kendler and Gardner, 1998).

The Consequences of Failure to Validate Boundaries

Cumulatively, the failure of most of these attempts to demonstrate valid boundaries between psychiatric syndromes emphasizes that it is unsafe to assume that there are valid boundaries to be found. We must take seriously the possibility that we are not yet 'carving nature at the joints,' or even that there are no 'joints' there, and that we are trying to impose an artificial set of discrete categories on what is in reality a continuous variation. Despite many proposed candidates, no single biological marker has been found to be specific for any operationally defined diagnostic category; also pharmacological trials proved, along the years, that psychotropic drugs are largely unspecific and that their action cuts across the categorically defined syndromes.

This state of affairs holds several important implications also for the differential diagnosis of psychiatric disorders. The first is the importance of recognizing that most of the categories of disorder recognized in contemporary nosologies have not yet been adequately validated, and that their formal definitions are simply agreed conventions. The second is a consequence of the first: all decisions about diagnosis and all discussions about differential diagnosis should take place in a setting in which it is agreed which set of diagnostic categories and corresponding set of operational definitions (in effect, which diagnostic system) are being used and with a clear understanding that both the categories and the definitions are arbitrary and inadequately validated, and therefore likely to be modified as understanding of the underlying etiology accumulates. The third implication is a consequence of the likelihood that many existing boundaries, either between syndromes or between psychiatric disorder and health, are not situated at points of rarity in the distribution of symptomatology. To the extent that this is the case, relatively minor changes in the operational definition of a syndrome, which produce corresponding changes in the position of its boundaries, may result in quite large changes in the numbers of people, or patients, who fall within those boundaries. This is almost certain why small changes in the definition of common disorders such as major depression or generalized anxiety – particularly changes in the number of typical symptoms required, or in their duration – have been associated with major changes in the recorded prevalence of these disorders in community surveys.

Finally, if it is indeed the case that there are no genuine boundaries within large tracts of the territory of psychiatric disorders, and we are imposing an artificial set of discrete categories on what is in reality continuous variation, psychiatrists will have to consider seriously the possibility of replacing their traditional categorical classification of disorders with a dimensional classification. This would never be done lightly, if only because the rest of medicine has always been committed to and content with categorical classifications of its disorders. If a dimensional format ever were to be adopted for mental disorders, however, the consequences for differential diagnosis would be profound. Disputes about which disorder or combination of disorders an individual was suffering from would be replaced by discussions about the precise locus in a multidimensional space at which that individual should be located. These discussions would probably also be calmer than the previous disputes, because in most cases it would be clear that the alternative loci under discussion were all quite close to one another, and not separated by dividing lines implying major differences in treatment or prognosis.

Practical Issues in Differential Diagnosis

As in other branches of medicine, some differential diagnoses are more critical than others in psychiatry. The most important functions of a diagnosis are to determine treatment and to predict outcome, and the most important differential diagnoses are those involving a disorder requiring urgent and specific treatment. All psychiatrists are keenly aware of the importance of diagnosing, or at least suspecting, cerebral tumors, and other brain diseases such as viral encephalitis and subdural hematoma, which may also present with psychiatric symptoms, as quickly as possible. Because these conditions may all present with depressive symptoms, with apathy or other behavioral changes, or even with psychotic symptoms, and because they all require urgent treatment, they enter into the differential diagnosis of a wide range of psychiatric disorders. Many reputations for diagnostic acumen are, rightly, based on successful detection of relatively uncommon brain diseases of this kind, or of metabolic disorders such as thyrotoxicosis and Cushing's syndrome, which underlie some anxiety states and depressions. It is also important to detect brain diseases such as Creutzfeldt-Jakob disease, which sometimes present with psychiatric symptoms even though there is not yet any treatment capable of altering the course of the disease, because of the implications for prognosis. If someone has an illness which is inevitably going to result in permanent disability or death, the sooner that is established the better, both for the patient and the family.

There are other circumstances in which the critical diagnosis in the differential diagnosis between brain disease and psychiatric disorder is the latter rather than the former. Dementias are common in old age and most are associated with progressive decline to dependency and death. Sometimes, however, depressive illnesses present with such widespread evidence of intellectual impairment that they closely resemble a dementia. Because depressions are eminently treatable it is vital that these 'depressive pseudodementias' are quickly recognized and treated, and depression is therefore an important part of the differential diagnosis of dementia. Similar considerations apply to rare curable dementias, such as those associated with a deficiency of thyroid hormone or vitamin B_{12}.

Traditionally, much time and intellectual effort have been devoted by psychiatrists to distinguishing accurately between schizophrenic and affective psychoses. To some extent this is appropriate. The two disorders require different treatments and there are major differences in long-term prognosis between them. Some disagreements are due simply to undisclosed differences in diagnostic criteria which would be resolved by the adoption of an agreed operational definition. However, a more fundamental conceptual issue may also be involved. Although there are many instances, including the differential diagnosis between schizophrenic and affective psychoses, where no valid boundary has yet been demonstrated between the two syndromes, psychiatrists often still behave as though they were discussing the distinction between two 'disease entities.' In reality, if the patient's symptoms are 'schizoaffective,' it may not matter much which side of an arbitrary boundary they are deemed to lie, because their therapeutic needs and their prognosis may both be genuinely intermediate between those of the two classical syndromes.

To diagnose schizophrenia is always a grave decision. Although some people develop schizophrenic illnesses which resolve completely, even without treatment, and never recur, for many more a schizophrenic episode is the prelude to a life of invalidism and the death knell of hopes and ambitions. It is understandable; therefore, that many psychiatrists are reluctant to diagnose schizophrenia, particularly in an adolescent, unless the evidence is incontrovertible. If some less doom-laden diagnosis, such as depression, is plausible they will opt for that, and hope that time will justify their decision. Recently, however, psychiatrists have become increasingly convinced by the evidence that early treatment with neuroleptics improves the long-term outcome of schizophrenic illnesses and reduces the risk of a 'defect state.' Although this evidence is not yet buttressed by random allocation trials it is sufficiently strong to swing the balance of advantage in patients with ambiguous symptomatology in favor of early diagnosis. A similar mix of social and therapeutic considerations has an influence on differential diagnosis in other areas too.

Other important issues of differential diagnosis arise when patients with psychiatric disorders present to other medical disciplines, as they commonly do. Depressive illnesses are extremely common at all ages and depressed patients may present with a wide variety of complaints, particularly if they are reluctant to accept that they may have a psychiatric disorder. They may complain of 'feeling off color,' of loss of appetite or weight, of difficulty sleeping, or of ill-defined pain almost anywhere in the body. As a result, the correct diagnosis is often missed unless questions are asked about characteristic depressive symptoms such as difficulty concentrating or loss of the ability to enjoy things. Phobic anxiety is another common disorder which often presents with somatic symptoms such as palpitations, or recurring bouts of upper abdominal discomfort or difficulty breathing. Unless specific questions are asked about the environmental circumstances in which these symptoms develop, their phobic nature may easily be missed and the patient may be subjected to a long series of irrelevant investigations. In both cases the fundamental problem is that the doctor concerned does not realize, or remember, that a depressive illness or phobic anxiety should be important elements in the differential diagnosis of the patient's presenting complaints, even though these may be entirely somatic.

Finally, a possible aftermath of the spread of categorical-operational approach to psychiatric diagnosis is the popularity recently gained by the concept of 'psychiatric comorbidity' (i.e., the coexistence of two or more psychiatric diagnoses). Several warnings (see Maj, 2005) have been recently made that the increased recognition of comorbid psychiatric diseases may just represent a by-product of the introduction of the latest version of ICD and DSM: since their adoption in clinical settings, the idea that psychiatric disorders represent discrete disease entities has been significantly questioned and the process of differential diagnosis since, according to this view, the nature of psychopathology might just be intrinsically composite and changeable and the idea splitting syndromes into multiple disorders could be just a nonvalid approach to the complexity of psychiatric conditions.

See also: Nosology in Psychiatry.

Bibliography

American Psychiatric Association, 1980. Diagnostic and Statistical Manual of Mental Disorders, third ed. American Psychiatric Association (DSM-III), Washington, DC.

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders, fourth ed. American Psychiatric Association (DSM-IV), Washington, DC.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Publishing (DSM-5), Arlington, VA.

Feighner, J.P., Robins, E., Guze, S.B., Woodruff, R.A., Winokur, G., Munoz, R., 1972. Diagnostic criteria for use in psychiatric research. Archives of General Psychiatry 26, 57–63.

Jaspers, K., 1913. Allgemeine Psychopathologie. Ein Leitfaden für Studierende, Ärzte und Psychologen. Springer, Berlin.

Kendell, R.E., 1975. The Role of Diagnosis in Psychiatry. Blackwell Scientific, Oxford, UK.

Kendler, K.S., Gardner, C.O., 1998. Boundaries of major depression: an evaluation of DSM-IV criteria. American Journal of Psychiatry 155, 172–177.

Maj, M., 2005. "Psychiatric comorbidity": an artefact of current diagnostic systems? British Journal of Psychiatry 186, 182–184.

World Health Organization, 1993. The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research. World Health Organization, Geneva, Switzerland.

Differentiation: Social

Uwe Schimank, Universität Bremen, Institute of Sociology, Bremen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of social differentiation is an important theoretical instrument to understand the origin, dynamics, and structure of modern society. In sociological theory, there are two perspectives on the differentiation of modern society. On the one hand, social differentiation is seen as division of labor, with an emphasis on advantages of specialization. On the other hand, social differentiation is understood as the emergence of autonomous 'value-spheres' or self-referential chains of communication, with an emphasis on the 'polycontexturality' of society. Performance improvements, evolution, the rationalization of cultural ideas, and interests of actors are identified as driving forces of social differentiation. The consequences of increasing differentiation are ambivalent both for individuals and for society at large. Individuals gain options but lose 'ligatures'; and modern society pays for performance improvements of its subsystems with problems of social and ecological integration.

Social differentiation is an important attribute of society, especially modern society (*see* Societies, Types of). The concept of differentiation spread within sociology since Herbert Spencer's evolutionary theory of societal development 'from incoherent homogeneity to coherent heterogeneity.' Later on, Emile Durkheim, Georg Simmel, Talcott Parsons, and Niklas Luhmann were important proponents of the concept. Other social thinkers, such as Karl Marx and Max Weber, who did not use the term 'differentiation' prominently nevertheless contributed to a proper understanding of the social structures and dynamics it designates. At the start of the twenty-first century, the theoretical and, more and more, empirical debate about social differentiation is continued, especially by American 'neofunctionalists' in their discussion of Parsons' legacy and by some German sociologists who keep up the debate based on Luhmann or Weber (Alexander and Colomy 1990; Schimank 1996; Schwinn 2011).

Social differentiation means both a process and a structure. In structural terms, it refers to the fact that a unit of analysis, such as a society, consists of a number of distinct parts. These parts may be of the same kind, such as families as the basic components of tribal societies. Or the parts may be different, such as the subsystems or 'value-spheres' – economy, politics, mass media, etc. – making up modern society. As a process, social differentiation is the dynamic that brings about and changes a given structure of differentiation.

Two fundamentally diverging views of the social differentiation of modern society have been put forward in sociological thought. In the following, the partly opposing but mostly complementary relationship between both views will be sketched, with respect to the features, driving forces, and consequences of social differentiation. The focus of attention will be put on the differentiation of modern society.

Social Differentiation as Decomposition or Emergence

For Durkheim and Parsons as the most important representatives of the first view of social differentiation, it is a process of decomposition of a functionally diffuse unit – a role or an institution – into several, functionally more specific units. This idea is modeled after the division of labor in work organizations, a process which fascinated nineteenth-century thinkers (*see* Labor, Division of). Durkheim's major illustration is the widespread specialization of occupations in industrial society. In Parsons' (1971) more abstract analysis, society is composed of four principal subsystems (economy, polity, societal community, and fiduciary system) each of which fulfills one of the four fundamental functional prerequisites of societal reproduction. Social differentiation as a process of social change tends toward a successive decomposition of each of these subsystems into four second-level subsystems, and so on. Parsons called this 'functional differentiation,' a term that is still used to describe the type of differentiation of modern society. His analytical perspective, Parsons suggests, is validated empirically over time by the concrete differentiation of social roles and institutions.

In structural terms, the decomposition perspective on social differentiation highlights the advantages of specialization for society, which are basically the same as on the organizational level. Specialization greatly improves performance. For instance, if formal education in schools is differentiated from the family, where education was previously concentrated, both sides profit in terms of functional efficiency and effectiveness – teachers can rely on primary socialization accomplished by parents and are able to focus on secondary socialization, and vice versa.

In contrast to this perspective of social differentiation as an overall harmonious division of labor, Weber portrayed the birth and composition of modern society as the emergence of a number of autonomous 'value-spheres.' One after another, science, law, art, politics, economics, sexuality, and others freed themselves from their former domination by religious ideas. These value-spheres became societal domains of their own which cultivate their particular central value without concern for the value orientations of other domains. Science, for instance, does not care about economic profits, aesthetic beauty, legality, or the conformity of its truths with religious

doctrine. Weber (1919, pp. 27–28) speaks graphically about the 'polytheism' of value-spheres in modern society.

Luhmann formulates this second perspective on social differentiation even more to the point. To him, society consists of about a dozen subsystems in each of which communication is guided by a specific binary code. For example, the science system's binary code is true/untrue, the legal system's legal/illegal. The binary code of each subsystem establishes a highly selective searchlight onto the social world, illuminating certain corners of it and leaving the rest in the dark.

The binary codes generate self-referentially closed universes of communication. Thus, a societal subsystem consists of chains of communications each of which refers to other communications of the same kind. For this fundamental feature, Luhmann picks up the concept of 'autopoiesis' from biological thinking (*see* Sociobiology and Sociology: A New Synthesis). To illustrate, in the social system of science, assertions of truth are the specific kind of communication. These assertions, usually made in scientific publications, lead to other such assertions, and – within science – not to political declarations or declarations of love. Any scientific publication, if it provokes any reactions at all, is either affirmed or rejected by later publications. Which of these two eventualities happen does not matter with respect to the reproduction of science as a societal subsystem. In both instances, science continues as an open-ended sequence of publications.

Luhmann's conceptualization is better suited than Weber's to grasp the basic dynamics of continual reproduction of a societal subsystem or value-sphere. For Luhmann, each subsystem can take only its own code-determined perspective on social events. In science, nothing but its own concerns are important, and the same holds true for all other subsystems. The overall result is an effective multiplication of society with respect to the communicative meanings of events. An event, such as a train accident, looks totally different from the point of view of the economy, compared to the points of view of science, health care system, mass media, and so on. Borrowing a term from philosopher Gotthard Guenther, Luhmann (1997) calls this the 'polycontexturality' of modern society, which is very near to Weber's 'polytheism.'

Although all subsystems have, in this sense, no 'window' to the social world outside, this basic autonomy goes along with an openness to their societal environment with respect to resources and functional performance. The science subsystem, for instance, needs financial resources among other things, and gets them from the political subsystem, or the economy. In the other direction, science provides these and other societal subsystems with knowledge. Since all societal subsystems make essential contributions to the reproduction of many others and none of them can substitute any other one, each is a functional prerequisite of modern society. Modern society would break down without economy, but also without mass media, or without a health care system, and the latter could not be replaced, for example, by politics or education.

Despite these manifold mutual dependencies which Parsons conceives of as "double interchanges," Luhmann insists that it makes no sense to understand the ensemble of societal subsystems as a division of labor. In a division of labor, the subsystems as parts are teleologically oriented toward the performance and maintenance of the whole society. Thus, the whole comes first, historically and logically, and decomposes itself for its own benefit. By contrast, Weber and Luhmann emphasize that social differentiation consists of a simultaneous birth and liberation of the parts. They emerge and at the same moment become autonomous from each other and from the whole. Even more, the whole vanishes in the sense that it becomes nothing more than the – often tension-ridden and antagonistic – interrelationship of the parts.

Driving Forces of Social Differentiation

Both analytical perspectives identify different driving forces of social differentiation. Four kinds of causal factors can be distinguished: improvements of performance, evolution, cultural ideas, and interests of actors.

The first causal factor, improvements of performance by specialization, was especially emphasized by Parsons and structural–functionalist studies of social differentiation (*see* Functionalism in Anthropology). Such specialization is initiated either by performance deficits of existing structures or by opportunities to improve a satisfactory level of performance. An example of the problem-driven dynamic is the previously mentioned differentiation of the education system. Schools, and a duty of all children to attend them, were established because other societal subsystems, especially the economy, needed a general level of cognitive skills and social discipline within the population that could no longer be provided by the family alone. A dynamic driven by opportunities was the spur for differentiation of the radio and television sector within the mass media. Entirely new possibilities of public communication opened up by the technological innovations of radio and television, rather than any perceived performance deficits of newspapers, caused the sectoral differentiation of roles and organizations within mass media.

By contrast to a planned division of labor within organizations, on the societal level performance improvements by differentiation occur often as unintended results of actions motivated by quite different reasons. In this sense, Parsons already relied on evolution as the mechanism that brings about performance improvements (*see* Evolutionary Sociology). In contrast, Luhmann gives up all connotations of evolution with progress. Adopting neo-Darwinian evolutionary thinking from biology, he conceives of evolution as an interplay of three mechanisms, variation, selection, and retention. Each societal subsystem has its own specific evolutionary mechanisms. For instance, in the science subsystem, most publications are variations of the existing body of knowledge. Selection happens whenever one of these publications is cited affirmatively. Retention occurs when this new knowledge is incorporated into overview and review articles and into textbooks for students. As in biological evolution, most variations are not selected, and most of those selected do not reach the stage of retention. This evolutionary cognitive growth and differentiation goes along with a social differentiation of scientific specialties.

Luhmann does not deny that this evolution of the science system may bring about performance improvements usually called 'scientific progress.' But for him, this is no necessary outcome of evolutionary differentiation. Evolution may

also lead to dysfunctional differentiation. For instance, the growing quest for interdisciplinary research indicates an overspecialization in more and more scientific fields. An understanding of evolutionary differentiation as progress becomes even more dubious in other societal subsystems, such as art, politics, or the mass media. But even if evolutionary differentiation on the level of each subsystem would produce performance improvements of all of them, the overall result for society would not add up to progress but only to an increase of societal complexity which manifests itself particularly in a growing destabilization of society. Each evolutionary change within one subsystem is also a change of the societal environment of all other subsystems. For them, often adaptive changes become necessary each of which again changes the societal environment, with new adaptive pressures arising. Thus, social differentiation becomes permanently autodynamic. Until now, these dynamics often brought about increasing differentiation. But Parsons' statement that this is a linear and unlimited social process is clearly wrong as pointed out by Luhmann's theoretical reflections and demonstrated by the empirical cases presented especially by the neofunctionalists.

A third driving force of social differentiation was already mentioned with respect to Weber: an autodynamic rationalization of cultural ideas. In premodern societies, the values that today constitute the guiding principles of the various societal subsystems were closely interwoven with each other and, moreover, integrated into an encompassing order of religious values. The differentiation of the principal subsystems of modern society is understood by Weber as a gradual but irresistible separation of economic, political, scientific, aesthetic, legal, erotic, and other values, first, from the religious context, and second, from each other. This dynamic is triggered and sustained by the compelling intellectual urge to contemplate extensively the implications and consequences of each value for everyday life. When this rationalization reaches the point where each value is rigorously reflected in its own terms, without regard to others, soon a self-referential closure of value-spheres occurs, giving way to a corresponding differentiation of roles, organizations, and subsystems.

The fourth driving force of social differentiation is the general interests of individual, collective, or corporate actors (*see* Interests in Sociological Analysis). One explanation of an ongoing division of labor mentioned by Durkheim (1893) is growing 'social density' resulting in an intensification of competition among actors. This constellation evokes an interest to preserve one's own domain of action, be it an occupation for an individual or a market position for a firm. This interest, in turn, leads many of the respective actors to find new niches for themselves where they are relieved from high competitive pressure. The aggregate effect of such domain-preserving activities is social differentiation on the level of roles and organizations.

Other general interests of actors that often stimulate dynamics of social differentiation are the preservation or extension of one's autonomy, of one's control of other actors, and of one's resource base. Some professions, especially those guided by such interests were strong carriers of the differentiation of certain societal subsystems including science, law, education, and health care (*see* Professions, Sociology of). Sometimes, a particular kind of social differentiation is the explicit goal of certain groups of actors who believe that this will serve their interests best. In other cases, social differentiation is an unnoticed, unforeseen, or unwanted effect of actors' pursuit of their interests. Outcomes often diverge markedly from intentions because usually the relevant constellation consists of actors in quite different positions: catalyzers and promoters of the respective process of change; followers who join in when a 'critical mass' has been mobilized; defenders of the status quo who resist any change; more open-minded defenders with whom compromises are possible; and those who are indifferent or indecisive at the beginning and whom the promoters of change as well as the defenders of the status quo try to recruit as allies.

Each of these four kinds of driving forces is causally relevant in the dynamics of social differentiation, although the combination varies greatly from one case to another. Often these factors are interrelated. Thus, what actors understand as their interests is framed by cultural ideas. Cultural ideas as well as actors' interests are explicit or implicit criteria to judge the performance of a given differentiation structure; and on the other hand, performance deficits provoke actors' interests. Finally, the evolution of societal subsystems creates new challenges, restrictions, and opportunities for the pursuit of interests and the spelling out of cultural ideas.

Consequences of Social Differentiation

The differentiation of modern society became a central theme not just of sociological thought but also of public discourse in the nineteenth century, and still is today. This is because social differentiation has far-reaching consequences for society at large, as well as for each individual member of it. These consequences are neither totally positive nor totally negative but a mixed blessing.

To begin with, the modern individual profits from role differentiation and an enormous increase of options in every sphere of life, from economic markets for goods and services to opportunities of political participation, sports activities, lifelong education, and cultural life. The abundance of options is the result of performance improvements in all societal subsystems and of an inclusion of all members of society into the status of customers, clients, and public of the subsystems. In connection with the polycontexturality of modern society, role differentiation and the growth of options allow each person to become an individual in the proper sense of the word: a person who creates a unique biography for himself or herself. However, options are only one side of a person's 'life chances' (Dahrendorf, 1979). The other side includes 'ligatures,' meaningful attachments which tie a person to certain institutions, cultural worlds, social groups, or communities. In this respect, the functional differentiation of modern society seems to neglect important psychic needs of persons. Ligatures are eroded by 'polycontexturality' and individualization. As a consequence, anomy, alienation, and other experiences of meaninglessness are present in modern society, not just as individual misfortunes and temporary phenomena, but as structurally produced permanent vulnerabilities of everyone.

Looking at the society at large, social differentiation exhibits the same kind of ambivalence. First, it is obvious that functional differentiation is evolutionarily highly successful. All kinds of premodern, traditional societies were overrun by those national societies from Europe and North America that converted to functional differentiation since the sixteenth century. Worldwide, there was no alternative to the adoption of functional differentiation. As shown by the socialist experience, those countries that tried to stop halfway failed. Countries where the political system attempted to control all other subsystems were beaten by the fully developed functional differentiation of Western countries, as Parsons, among others, had anticipated in the 1960s.

However, global supremacy of a fully fledged functional differentiation implies serious problems. Their common denominator is an endangering of the integration of modern society. More precisely, systems integration, social integration, and ecological integration must be distinguished.

Systems integration refers to the coordination and cooperation among different subsystems of modern society. This is obviously possible, despite the fact that each of them is governed by its own self-referential binary code. There are manifold interrelationships by which subsystems adapt to each other in a decentralized way. Where this does not suffice, representatives of subsystems, such as the respective interest groups, meet and bargain for common solutions in the political arena. Finally, political guidance may support or promote intersystems adaptation and bargaining among subsystems. Thus, problems of systems integration most of the time remain fragmented and manageable in modern society. However, contemporary pressures of 'economizing' on many other societal subsystems such as health care, education, or science are a serious challenge to the maintenance of their autonomy.

Social integration means the integration of individual members into modern society (*see* Integration, Social). It may be precarious as a result of the already mentioned erosion of ligatures. More important could be the fact that the increase in individual options might put excessive demands on all subsystems' abilities in the long run. In this respect, social inequality and the subsystems' unlimited drive to improve their performance work hand in hand (*see* Inequality, Social). In conflicts among those who, as individuals, profit more and those who profit less from a particular subsystem's performance, the former will try to avoid redistribution. One way to achieve this is a growth of the subsystem's performance so that those whose claims were not satisfied can be served somewhat better without those who were better of giving up anything. For instance, children from the lower classes have got access to higher education. But this has been no zero-sum conflict in which the higher classes have lost some of their educational opportunities but a positive-sum constellation because the higher education sector has been extended – which has fitted also the interests of the respective groups of teachers and professors and has been reinforced by the cultural imperative of 'progress.' Such dynamics could go on forever, were there no limits to subsystems' growth. But because these limits exist, for example with respect to financial resources, continuing inequalities contain a potential for serious social conflict to which functional differentiation has no ready answer. In fact, in many countries, social inequalities have grown again during the last three decades, and large parts of the world population are only minimally included and sometimes even totally excluded from most subsystems' services.

Finally, the functional differentiation of modern society is a major cause of the growing problems of ecological integration. Modern society's path away from a 'sustainable' relationship with its natural environment has occurred as a result of two basic characteristics of all societal subsystems: their growth tendencies and the neglect of ecological aspects by their binary codes. The modern economy is the most obvious example, but the same is true of the family, the subsystem primarily responsible for population growth. No societal subsystem exists now that takes care of the ecology, only social movements address their protest to the political system. But government, following its binary code of maintaining and increasing power, must listen to many other demands from society and so often disregards ecological concerns.

Thus, the functional differentiation of modern 'world society' is a high-risk outcome of social evolution. No alternative to this type of social differentiation is apparent as an option for the future. But whether modern society can survive in this way or will ruin itself sooner or later is an open question that sociological theories of social differentiation cannot answer. Theorists can only observe what has happened previously and infer warnings from this.

See also: Evolutionary Sociology; Functionalism in Anthropology; Inequality: Comparative Aspects; Integration, Social; Interests in Sociological Analysis; Labor, Division of; Professions, Sociology of; Societies, Types of; Sociobiology and Sociology: A New Synthesis.

Bibliography

Alexander, J.S., Colomy, P. (Eds.), 1990. Differentiation Theory and Social Change. Columbia University Press, New York.

Dahrendorf, R., 1979. Life Chances. University of Chicago Press, Chicago.

Durkheim, E., 1893. The Division of Labor in Society. Free Press, New York. Reprinted (1964).

Luhmann, N., 1997. Die Gesellschaft der Gesellschaft. (The Society of Society.) Suhrkamp, Frankfurt, Germany.

Parsons, T., 1971. The System of Modern Societies. Prentice Hall, Englewood Cliffs, NJ.

Schimank, U., 1996. Theorien gesellschaftlicher Differenzierung. (Theories of Social Differentiation.) Leske and Budrich, Opladen, Germany.

Schwinn, T., 2011. Soziale Differenzierung. (Social Differentiation.) VS, Wiesbaden.

Weber, M., 1919. Wissenschaft als Beruf. (Science as Vocation.) Duncker and Humblot, Berlin. Reprinted (1967).

Diffusion and Random Walk Processes

Philip L Smith, The University of Melbourne, VIC, Australia
Roger Ratcliff, The Ohio State University, Columbus OH, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R. Ratcliff, volume 20, pp. 3668–3673, © 2001, Elsevier Ltd.

Abstract

Diffusion and random walk processes comprise one of the main classes of sequential-sampling models for decision making and choice response time in psychology. The models assume that decisions are made by accumulating samples of noisy evidence to a response criterion, either in continuous or in discrete time. The models successfully predict the main features of performance in speeded decision tasks: accuracy, the ordering of mean response times for correct responses and errors, the shapes of response time distributions, and the effect of instructions. Neural correlates of the evidence accumulation process have been identified in studies of the brain regions involved in decision making.

Historical Foundations

Diffusion and random walk models form one of two major classes of sequential-sampling models for speeded decision making and choice response time in psychology. The other class comprises accumulator and counter models. A taxonomy of these two classes of models, first compared by Audley and Pike (1965), is shown in **Figure 1**. Models of both classes conceive of decision making as a statistical process, in which successive samples of noisy evidence or stimulus information are accumulated to a response criterion. The evidence is assumed to be noisy, either because the stimulus is a sequence of variable or noisy events, or because of noisy coding in neural systems. The criterion represents the amount of evidence needed to make a decision. In random walk models, evidence for one response is evidence against the other. The accumulating evidence is represented as a single fluctuating sum, which increases or decreases as evidence favoring one or other response is sampled. In counter and accumulator models, evidence for the two responses is accumulated in parallel, in separate totals. The decision process in these models has the form of a race between accumulating evidence totals, with the winner of the race determining the response.

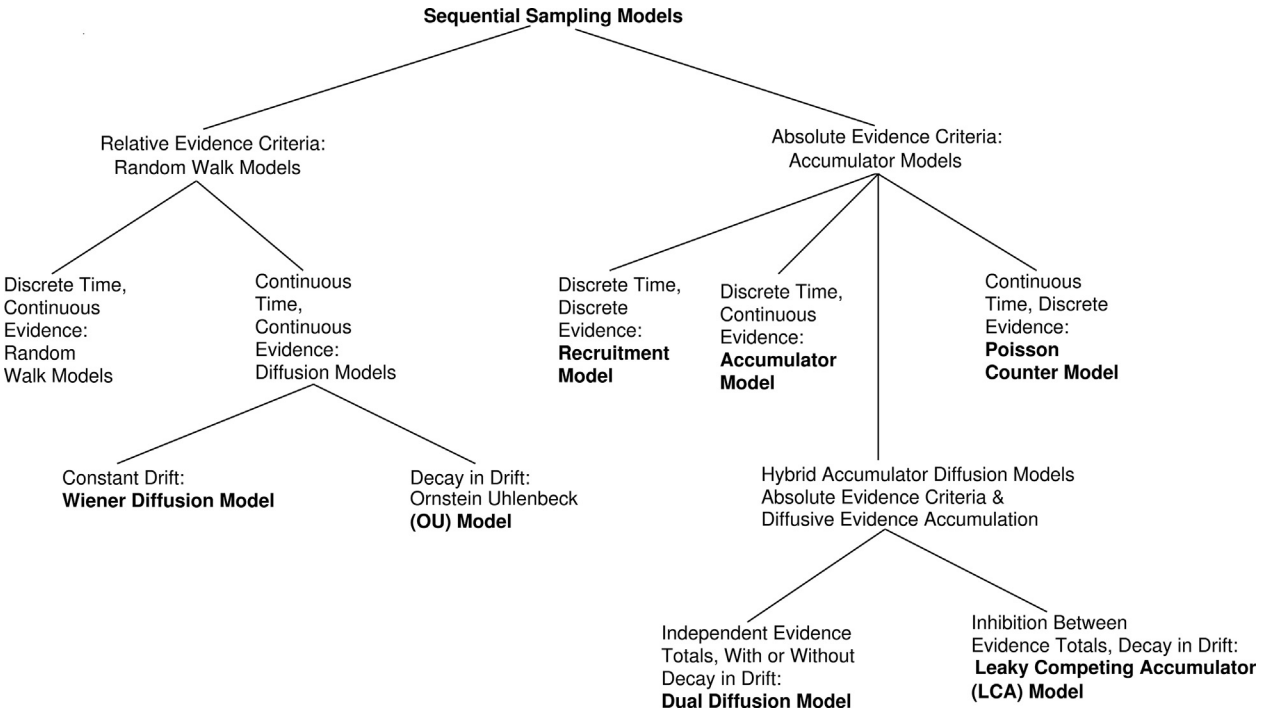

Figure 1 Sequential-sampling models. The branch on the left shows random walk and diffusion models. The hybrid accumulator–diffusion models shown under the branch on the right combine the attributes of accumulator models and diffusion models.

Random walk models are the discrete-time counterpart of diffusion process models and historically preceded them. Random walk models for decision making were introduced into psychology by Stone (1960) and Edwards (1965) and were developed subsequently by Laming (1968). The aim of the models was to provide a unified account of the processes underlying psychology's two ubiquitous dependent variables: accuracy and response time (RT). The early random walk models were influenced by Wald's (1947) sequential probability ratio test (SPRT) in statistics, which was based on the accumulation of log-likelihood ratios. SPRT models assume that the decision process acts on a sequence of noisy observations, each element of which provides evidence for one of two alternative hypotheses, H_a and H_b, which represent states of the world. In a choice response time setting, the hypotheses might be that either stimulus s_a was presented or stimulus s_b was presented. The task is to decide which of these hypotheses is true and to respond as quickly as possible. The evidence is represented as a sequence of random variables, Z_j, $j = 1, 2, \ldots$, which is sampled at a constant rate. The decision process computes the log-likelihood ratio of the evidence for the two hypotheses obtained at step j and accumulates these as a running sum. If $p_a(Z_j)$ and $p_b(Z_j)$ denote the probability density functions for Z_j under H_a and H_b, respectively, then the decision process, X_i, at time i is the sum,

$$X_i = \sum_{j=1}^{i} \log\left[\frac{p_a(Z_j)}{p_b(Z_j)}\right]. \qquad [1]$$

If the Z_j are Gaussian random variables, then the log-likelihood ratio is also Gaussian, and the running sum is a Gaussian random walk.

To make a decision, evidence is sampled until the accumulating total reaches or exceeds one of two absorbing boundaries or evidence criteria, a and b, with $b < 0 < a$. If the first criterion reached is a, the person makes response R_a (stimulus s_a was presented); if the first criterion reached is b, the person makes the alternative response R_b (stimulus s_b was presented). The decision time, T, is the minimum time to reach a criterion. Formally, if $T_a = \min\{i : X_i \geq a\}$ and $T_b = \min\{i : X_i \leq b\}$, then $T = \min(T_a, T_b)$. The time of the first boundary crossing, T, is guaranteed to be finite, which means that the process will terminate in finite time with one or other response. Wald's SPRT is optimum in the sense that it is the process that achieves a given criterion level of accuracy in the minimum expected time. It was this property that first recommended it as a model for decision making in settings in which there is a cost associated with sampling evidence, either in time or other units.

A later random walk model, the *relative judgment theory* of Link and Heath (1975), omitted the computation of likelihood ratios and assumed the walk accumulates the magnitudes of the noisy evidence sequence, Z_j, directly,

$$X_i = \sum_{j=1}^{i} Z_j. \qquad [2]$$

This change was to allow the model to account for the orderings of mean RTs for correct responses and errors that are found experimentally. Some recent treatments of random walks and diffusion processes in psychology and neuroscience have again begun to emphasize the connection with Wald's SPRT and optimality (Bogacz et al., 2006; Gold and Shadlen, 2002).

Diffusion Process Models

Mathematically, diffusion processes are the continuous-time counterparts of random walks. Whereas a random walk, X_i, accumulates evidence at discrete-time steps, $i = 1, 2, \ldots$, a diffusion process, X_t, accumulates evidence in continuous time ($t > 0$). The study of diffusion processes was begun in physics in 1905 by Albert Einstein, who derived a partial differential equation for the probability distribution of a pollen particle undergoing Brownian motion. He showed that this equation had the same form as the classical heat equation of physics (Gardiner, 2004). In 1923, Wiener began the mathematical study of the path-space properties of diffusion processes, by proving the existence of a continuous stochastic process whose transition distribution satisfied the partial differential equation derived by Einstein. The resulting process is variously known as the *Brownian motion process* or *Wiener diffusion process*.

Ratcliff (1978) proposed a diffusion process model of two-choice decision making in psychology, arguing that information processing in the brain is better viewed as a continuous-time rather than a discrete-time process. In his model, which is shown in **Figure 2**, the accumulating evidence is described by the diffusion process studied by Einstein and Wiener. The model is similar to the earlier random walk models in that it conceptualizes decision making as diffusion on a line between absorbing barriers, which represent decision criteria. Ratcliff's model follows the gambler's ruin framing of

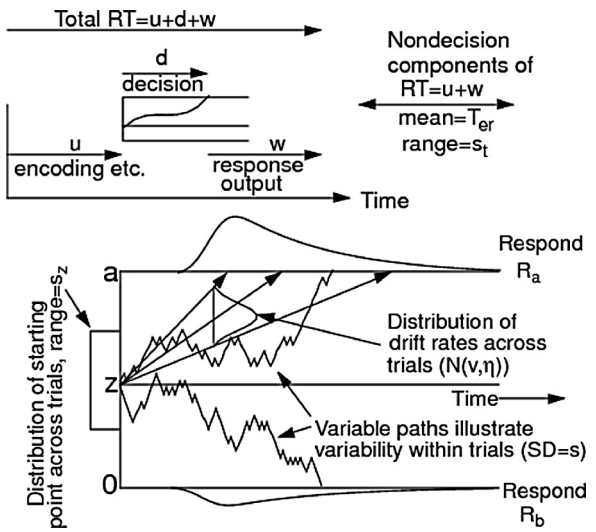

Figure 2 Diffusion model. The process starting at z accumulates evidence between absorbing boundaries at 0 and a. Moment-to-moment stochastic variability means the process can sometimes terminate at the correct response boundary, R_a, and sometimes at the error response boundary, R_b. Drift rate, ξ, is normally distributed between trials with mean ν and standard deviation η. Starting point is assumed to be rectangularly distributed with range s_z. RT is the sum of the stimulus encoding time, u, the decision time, d, and the response output time, w. The nondecision time, $T_{er} = u + w$, is rectangularly distributed with range s_t.

the process of Feller (1967), in which the lower barrier is set to 0, the starting point is set to z, and the upper barrier is set to a. Evidence for one response drives the process in an upward direction; evidence for the other response drives it in a downward direction, with the process continuing until one of the absorbing boundaries is crossed and a response is made. Unlike those models, evidence in his model accumulates continuously rather than at discrete points in time.

Like relative judgment theory, the evidence accumulated in the diffusion model is assumed to be a sequence of noisy sensory or cognitive states, rather than a likelihood ratio transformation of those states. Because the accumulation process, X_t, in the model is Gaussian, and because the likelihood ratio of a Gaussian process itself has a Gaussian distribution, the accumulating evidence in the diffusion model is proportional to the log-likelihood ratio. Consequently, although the model does not assume a likelihood ratio transformation of evidence states, it is SPRT optimal in the same way as are the likelihood ratio models of Laming (1968) and earlier researchers (Bogacz et al., 2006).

Ratcliff (1978) emphasized the importance of other sources of trial-to-trial variability in the model, in addition to diffusive noise arising from moment-to-moment variability in the evidence entering the decision process. The effects of diffusive noise are shown as short-time-scale random fluctuations in the accumulating evidence in **Figure 2**. The other sources of variability reflect cognitive processes that affect the decision process on individual trials. The first of these, shown in **Figure 2**, is variability in the rate of evidence accumulation, or the *drift rate*, of the process, which is assumed to be normally distributed. Drift variability represents trial-to-trial variation in the average quality of the information entering the decision process and arises due to variations in the quality of the match between encoded stimuli and the decision alternatives in memory. The presence of drift variability means that on some trials the process will drift in the wrong direction (i.e., presentation of stimulus s_a will cause the process to drift toward the R_b boundary or vice versa). This allows the model to predict slow errors, which are often found experimentally when stimulus discriminability is low and accuracy is stressed.

Ratcliff et al. (1999) added another source of trial-to-trial variability, first proposed by Laming (1968), in the starting point of the process. Starting point variability also allows the model to predict fast errors, which are often found experimentally when stimulus discriminability is high and speed is stressed. Fast errors arise because responses on trials on which the process has started near the error response boundary are more likely to be made rapidly and incorrectly.

Ratcliff (1978) also initiated the study of RT distributions as a method for testing decision models. Most early model comparisons focused on the ordering of mean RTs for correct responses and errors and the relationship between mean RT and accuracy. Ratcliff showed that the properties of empirical RT distributions impose strong constraints on the class of plausible models and, moreover, that the Wiener diffusion model predicts the families of RT distributions that are found experimentally. This has been confirmed in numerous studies of speeded decision making in simple perceptual and cognitive tasks in a variety of settings (Ratcliff and Smith, 2004; Ratcliff et al., 2005).

Mathematical Methods for Diffusion Processes

Diffusion processes are continuous time, continuous state space, Markov processes. Members of the class of diffusion processes are defined by specifying the *infinitesimal moments* of the process, together with initial and boundary conditions. The infinitesimal moments are the *drift* and *diffusion* coefficients, which define, respectively, the mean rate of change and the variance in the rate of change in a unit interval. In a general diffusion process, the drift and diffusion can both depend on time and on the state of the process. For such a process, the drift, $\mu(x, t)$, and diffusion coefficient, $\sigma^2(x, t)$, can be continuous functions of the state variable, x, and the time variable, t. In Ratcliff's (1978) model, the drift and diffusion coefficients are both constant and are denoted ξ and s^2, respectively, where $\xi \sim N(\nu, \eta)$ and where $N(.,.)$ is the normal density function and the tilde means 'is distributed as.'

Mathematically, diffusion processes can be defined via either partial differential equations or stochastic differential equations. If $f(s, y; t, x)$ is the transition density of the process, that is, $f(s, y; t, x)\,\mathrm{d}x$ is the probability that a process starting at time s in state y will be found at time t in the interval $(x, x + \mathrm{d}x)$, then the general diffusion, X_t, satisfies the partial differential equation

$$-\frac{\partial f}{\partial s} = \frac{1}{2}\sigma^2(y,s)\frac{\partial^2 f}{\partial y^2} + \mu(y,s)\frac{\partial f}{\partial y}. \qquad [3]$$

This is the Kolmogorov backward equation, so called, because its variables are the starting time and state, s and y. The process also satisfies the forward equation, which is an equation in t and x. The backward equation is useful for deriving RT distributions; the forward equation is useful for studying evidence accumulation by a process unconstrained by absorbing boundaries (Ratcliff, 1978). Alternatively, X_t can be defined as satisfying the stochastic differential equation

$$\mathrm{d}X_t = \mu(X_t, t)\mathrm{d}t + \sigma(X_t, t)\mathrm{d}W_t, \qquad [4]$$

where W_t is the Brownian motion process.

The key properties of the process in modeling decision making are the first passage time probability and the first passage time distributions. The first passage time probability is the probability that the process reaches a specified one of the absorbing boundaries before reaching the other one. In the decision model, the probability of making a particular response, say R_a, is given by the probability that the process crosses the absorbing boundary a before it crosses the boundary b. If T_a and T_b are the first passage times for the boundaries a and b, respectively, then the probability of making response R_a is $P(R_a) = P[T_a < T_b]$. The first passage time distributions, denoted $G_a(t)$ and $G_b(t)$, are the distributions of T_a and T_b. In the model, the first passage times are the decision times for the two responses. If the stimulus presented was s_a, then $G_a(t)$ and $G_b(t)$ are the distributions of correct responses and errors, respectively. The converse holds if the stimulus presented was s_b.

A variety of methods exist in the literature for obtaining first passage time statistics for diffusion processes. Classically, partial differential equations like eqn [3] are solved as algebraic eigenvalue problems. When the diffusion is time homogeneous (i.e., the drift and diffusion coefficients do not vary with time), these methods yield the first passage time distributions in the form of infinite series, which can be approximated by truncation (Ratcliff, 1978). For a diffusion with drift ξ, diffusion coefficient s^2, and starting point z, where $b < z < a$, the probabilities of responding at the upper and lower barriers are

$$P(R_a) = \frac{\exp(-2\xi z/s^2) - \exp(-2\xi b/s^2)}{\exp(-2\xi a/s^2) - \exp(-2\xi b/s^2)}, \quad [5]$$

and

$$P(R_b) = 1 - P(R_a), \quad [6]$$

respectively. The first passage time densities, $g_a(t)$ and $g_b(t)$, defined as $g_a(t) = G_a'(t)$ and $g_b(t) = G_b'(t)$, where primes denote differentiation with respect to time (Smith, 1990), are

$$g_a(t) = \frac{1}{P(R_a)}\exp\left[\frac{\xi(a-z)}{s^2} - \frac{\xi^2 t}{2s^2}\right]\frac{\pi s^2}{(a-b)^2} \times \sum_{k=1}^{\infty} k\exp\left[-\frac{k^2\pi^2 s^2 t}{2(a-b)^2}\right]\sin\left[\frac{k\pi(a-z)}{a-b}\right] \quad [7]$$

and

$$g_b(t) = \frac{1}{P(R_b)}\exp\left[\frac{\xi(z-b)}{s^2} - \frac{\xi^2 t}{2s^2}\right]\frac{\pi s^2}{(a-b)^2} \times \sum_{k=1}^{\infty} k\exp\left[-\frac{k^2\pi^2 s^2 t}{2(a-b)^2}\right]\sin\left[\frac{k\pi(z-b)}{a-b}\right]. \quad [8]$$

In applications, it is often more convenient to work with first passage time distribution functions rather than densities. Ratcliff (1978) gave the distribution functions using Feller's (1967) gambler's ruin form of the process, in which the lower boundary is set to zero ($0 < z < a$).

Several other methods have been investigated for obtaining first passage time statistics for diffusion processes. These methods can be used for more complex problems, such as, for example, when either the drift and diffusion coefficients or the absorbing boundaries vary with time, for which classical methods for solving partial differential equations often do not yield tractable solutions. These include integral equation methods, originally developed to study models of integrate-and-fire neurons in mathematical biology (Smith, 2000), numerical methods for solving partial differential equations (Voss and Voss, 2008), and matrix methods (Diederich and Busemeyer, 2003). The matrix approach, which approximates diffusion processes with finite state Markov chains, can be applied to very general problems of the kind that sometimes arise in applications, such as multidimensional diffusions with correlated increments (Ditterich, 2006). Although such problems can be made theoretically tractable via the use of finite state approximations, the size of the transition matrix required to represent the process and the associated computational demands rapidly become large. Tuerlinckx et al. (2001) have investigated methods for simulating diffusion processes.

Ornstein–Uhlenbeck Models

Since Ratcliff's (1978) article, researchers have investigated a number of variations on the underlying model. Busemeyer and Townsend (1993) proposed a model based on the Ornstein–Uhlenbeck (OU) diffusion process. This process was proposed in 1930 by Uhlenbeck and Ornstein as an alternative model for Brownian motion to that of Einstein and Wiener (Gardiner, 2004). It sought to correct a perceived shortcoming of that model, namely, it implies the velocity of a diffusing particle is almost everywhere infinite. Rather than modeling the displacement of the particle, the OU process models its velocity. Mathematically, it differs from the Wiener process by the presence of a decay term in the drift coefficient, which acts as a restoring force that pulls the particle back toward its starting point. For a particle starting at the origin ($X_0 = 0$), the OU process is defined by the stochastic differential equation,

$$\mathrm{d}X_t = (\mu - \beta X_t)\mathrm{d}t + \sigma \mathrm{d}W_t. \quad [9]$$

In this equation, μ is the external component of drift, which in decision-making models depends on the stimulus, $-\beta X_t$ is the restoring force, and σ^2 is the diffusion coefficient. When $\beta = 0$, the process becomes the Wiener diffusion process of Ratcliff's model (with $\mu \equiv \xi$ and $\sigma \equiv s$). Busemeyer and Townsend (1993) interpreted the OU decay term in their model as representing avoidance in a system of approach-avoid dynamics.

Most applications of diffusion models before Busemeyer and Townsend (1993) had been to brief two-choice decisions based on a single, 'one shot' cognitive process, for which RTs do not average much more than a second. They proposed that diffusion models could also be applied to more complex tasks, in which decisions are made after a period of deliberation. They showed their OU model could provide an account of otherwise puzzling findings in the judgment and decision-making literature, such as preference reversals, which are difficult to explain using traditional algebraic models (Roe et al., 2001).

Other investigators have proposed OU models in which the decay term represents a bounded evidence accumulation process. Smith (1995) proposed a model of simple RT with time-varying drift, $\mu(t) - \beta x$, in which the decision process is driven by sustained and transient visual channels. Usher and McClelland (2001) proposed a neurally inspired model of choice RT, the *leaky competing accumulator model* (LCA), based on a pair of coupled OU processes. Their model combined the properties of earlier accumulator models, such as that of Vickers (1970) – in which evidence for competing responses is accumulated in parallel by racing evidence totals – and diffusion process models, in which the accumulation process is diffusive. Denoting the evidence totals by X_1 and X_2 (and suppressing the dependence on the time variable, t), the growth equations in the LCA model are

$$\mathrm{d}X_i = (\mu_i - \beta X_i - kX_j)\mathrm{d}t + \sigma \mathrm{d}W_i, \quad i, j \in \{1, 2\}, \quad i \neq j. \quad [10]$$

In these equations, k is a coupling coefficient that represents the amount of mutual inhibition between the accumulating totals X_1 and X_2, and W_1 and W_2 are independent Brownian

motion processes. The relative optimality properties of the model were analyzed by Bogacz et al. (2006).

Diffusion Models and Empirical Data

Scaling and Distribution Shape

The two dependent variables, RT and accuracy, have different scale properties. RT has a minimum value, and its variance increases as RT increases. Accuracy is bounded on the range 0.5–1.0, and its variance decreases as it approaches 1.0. Diffusion models account for these scale properties automatically as a result of the geometry of the diffusion process. When mean drift rate is high, the probability of a correct response is near 1.0, and decision processes approach the correct boundary with little spread in arrival times. When mean drift rate is nearer zero, variability leads some processes to hit one boundary, other processes to hit the other boundary, and accuracy nears 0.5; the arrival times at boundaries are long and highly variable. In addition to these scale properties, the geometry of the diffusion process also gives RT distributions that are skewed to the right; as mean RT increases, due to a reduction in drift rate, the fastest responses slow a little and the slowest responses slow a lot.

Speed Versus Accuracy

People can manipulate their speed performance relative to their accuracy performance. If they are instructed in some blocks of an experiment to respond as quickly as possible, and in other blocks to respond as accurately as possible, mean RTs can vary between the two kinds of blocks by as much as 500 ms and mean accuracy can vary by 10% (Ratcliff and Rouder, 1998). Ratcliff's diffusion model can account for these differences with only the boundary separation parameter: moving the boundaries close together produces fast responses for which variability in drift can cause large numbers of errors; moving the boundaries further apart produces slower responses that are more likely to be accurate.

Error Response Times

Historically, comparisons of sequential-sampling models focused on their ability to predict empirical patterns of mean RTs for correct responses and errors. A problem with early random walk models was that they predicted that mean RTs for correct responses and errors would be the same. Specifically, the models predicted that $M(T_a|s_a) = M(T_a|s_b)$ and $M(T_b|s_b) = M(T_b|s_a)$, where $M(T_i|s_j)$ denotes the mean time to make response R_i when stimulus s_j is presented. This prediction is made by both SPRT random walks and Gaussian random walks in which the drift rates for the stimuli s_a and s_b are of equal magnitude and opposite sign: μ and $-\mu$, respectively (Luce, 1986; Townsend and Ashby, 1983).

A claimed advantage of the class of accumulator models (the right branch of **Figure 1**), which includes the recruitment model (LaBerge, 1962), the Poisson counter model (Townsend and Ashby, 1983), and the Vickers accumulator model (Vickers, 1970), is they predict that mean RTs for error responses will be slower than mean RTs for correct responses. But this represents only a partial solution to the problem of RT ordering, as errors can be faster or slower than correct responses depending on stimulus discriminability and instructions (Luce, 1986). Laming (1968) showed that variability in the starting point of a SPRT random walk allowed the model to predict fast errors, but this too represented only a partial solution.

One solution to the RT ordering problem was proposed in Link and Heath's (1975) relative judgment theory, which allows the distribution of evidence states, Z_j, in eqn [2] to be non-Gaussian. They showed that the ordering of mean RTs for correct responses and errors depends on the relative sizes of the steps taken by the walk toward the correct and error boundaries, which is characterized mathematically by asymmetry in the moment generating function of the evidence states. By allowing the asymmetry to vary freely, the model can predict either fast or slow errors. Although this yielded a formal solution to the problem of mean RT ordering, the theory did not provide an account of why the evidence distribution should vary with conditions that produce changes in ordering, such as instructions to respond rapidly or slowly.

Ratcliff's diffusion model, with trial-to-trial variability in drift rate and starting point, provides an alternative solution to the ordering problem. As a Gaussian process, it predicts equal mean times for correct responses and errors when there is no trial-to-trial variability. When there is variability in both drift and starting point, it can predict fast errors and slow errors, and the crossover pattern found in some experiments in which errors are faster in some conditions and slower in others (Ratcliff and Smith, 2004). As well as predicting the ordering of mean RTs, it also predicts the ordering of correct and error RT distributions. It therefore gives a complete account of the data from simple two-choice decision experiments. Unlike moment-generating function asymmetry, the sources of trial-to-trial variability in the model have clear psychological interpretations.

Quantile Probability Plot of RT Distributions

An experimental task in which the difficulty of the stimulus discrimination is varied yields a family of RT distribution pairs and choice probabilities. Each stimulus condition yields a distribution of correct responses and a distribution of errors, and the probability of a correct response. These data, which impose strong constraints on models of the decision process, can be represented in the form of a quantile probability plot, as shown in **Figure 3**. The RT distributions are summarized by their .1, .3, .5, .7, and .9 quantiles. The .1 quantile summarizes the leading edge of the distribution; the .5 quantile summarizes the 'typical' (median) response; and the .9 quantile summarizes the distribution's tail. For a given experimental condition, i, the quantiles of the distribution of RTs for correct responses are plotted against the probability of a correct response, q_i, and the quantiles of the distribution of RTs for errors are plotted against the probability of an error response ($1-q_i$). The distributions on the right of the .5 point on the x-axis are the distributions of correct responses, and the distributions on the left are the distributions of errors.

The main features of performance are shown on this plot. The spacing of the RT quantiles on the y-axis shows the unimodal, positively skewed form that is typical of RTs in

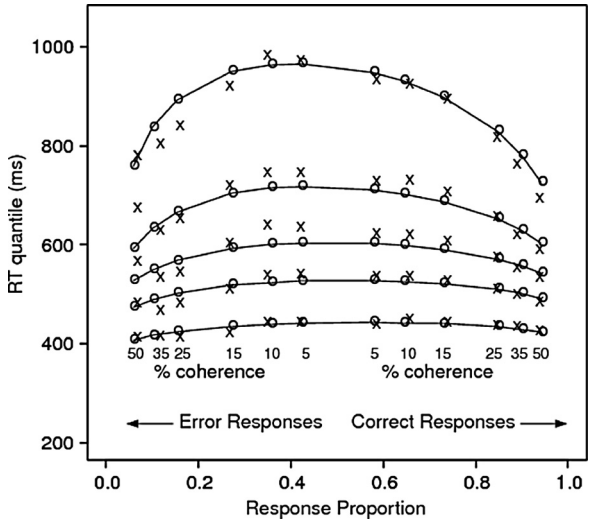

Figure 3 Quantile probability plot. Quantiles of the RT distributions are plotted on the *y*-axis against choice probabilities on the *x*-axis. The experimental task was a motion discrimination task in which discriminability was manipulated by varying the coherence of the motion in six steps. The *x*'s are the experimental data; the *o*'s connected by continuous lines are the fitted values of the Ratcliff diffusion model. Source: Ratcliff, R., McKoon, G., 2008. The diffusion decision model: theory and data for two-choice decision tasks, Neural Computation, 20, 873–922, Figure 7. Reprinted by permission.

simple cognitive tasks. The mean and standard deviation of the RT distributions increase, roughly in proportion to one another (Wagenmakers and Brown, 2007), as the discriminability of the stimulus is reduced. Most of the change in the RT distribution with changes in stimulus discriminability occurs in the tail; there is little change in the leading edge. In the data set shown, error RTs are typically slower than correct responses, which appears as an asymmetry of the plot across its vertical midline. **Figure 3** also shows fits of the Ratcliff diffusion model to the RT distributions and choice probabilities. The figure shows that the model captures the shapes of the RT distributions for correct responses and errors and their relationship with the proportions of correct and error responses.

Comparison of Diffusion and Accumulator Models

Ratcliff and Smith (2004) compared the performance of diffusion models, accumulator models, and models combining the attributes of both classes, in which evidence for competing responses accumulates in separate totals that are modeled as diffusion process. One such model was the LCA model (eqn [10]); the other, the *dual diffusion model*, assumed a pair of independent parallel diffusion processes. This model is like the LCA model but omits the mutual inhibition term, $-kX_j$, of eqn [10].

Ratcliff and Smith found that model architecture – whether a model assumed one evidence total or two – was less important than the process used to model evidence accumulation. Models with diffusive evidence accumulation, in either a single evidence total or two racing totals, were able to account for the main features of the experimental data, whereas the accumulator models were not. Accumulator models predict RT distributions that become less skewed and more symmetrical with increases in the mean and variance of RT, which is not what is found experimentally (e.g., **Figure 3**). The successful models assume that accumulating evidence can both increase and decrease, whereas in the unsuccessful models it only increases. Models with diffusive accumulation predict the shapes of RT distributions that are found empirically, whereas models with other kinds of accumulation process do not.

For the diffusion and hybrid accumulator–diffusion models, neither decay nor mutual inhibition between evidence totals was needed to account for empirical data – although these features are often thought to confer biological plausibility. With small values of decay, β, the OU model (eqn [9]) was indistinguishable from the Ratcliff diffusion model. With large values of decay, it predicted RT distributions that were more skewed than those found experimentally. The LCA model, which assumes mutual inhibition between evidence totals, and the dual diffusion model, which does not, gave similarly good accounts of the data. Mutual inhibition appears to be needed in models in which stimulus inputs only produce increases in evidence totals; in such models, mutual inhibition can provide the fluctuations in evidence needed to correctly predict RT distributions (e.g., Purcell et al., 2012).

Neural Studies of Diffusion Making

Recently, there has been considerable interest in neural analogs of diffusion processes. This has been stimulated by the emerging subdiscipline of neuroeconomics, which focuses on the computation of value, reward, and uncertainty in the brain (Glimcher et al., 2009), and by single-cell recordings from awake behaving monkeys performing eye-movement analogs of human-speeded decision tasks (Hanes and Schall, 1996; Roitman and Shadlen, 2002). These studies have reported patterns of firing rates in decision-related brain structures that appear to be neural correlates of the accumulation process hypothesized in the psychological models (Ratcliff and Smith, 2004). Cells in brain structures like the frontal eye fields, lateral interparietal area, and superior colliculus, which respond selectively to the presence of one of the decision alternatives in their receptive fields, show a progressive increase in firing rates to a constant level that is time locked to the production of a response. This characteristic pattern of firing rates appears to be the neural signature of the process of accumulating evidence to a criterion.

In the past decade, there has been a proliferation of research investigating the connection between the psychological processes of evidence accumulation and the neural processes of decision making. Ratcliff et al. (2007) showed that the parameters of a dual-diffusion model estimated from the RT distributions and choice probabilities of monkeys performing an eye-movement decision task correctly predicted the time course and relative firing rates in build-up cells in the superior colliculus in the period preceding the response. Ratcliff et al. (2009) identified a late

component of the human electroencephalogram (EEG) that indexes decision making and showed that trial-to-trial variations in the amplitude of this component predicted estimates of drift rates in a diffusion model fitted to the RT distributions and accuracy data. Mulder et al. (2012) linked estimates of bias (starting point, z) in a diffusion process to activity in frontoparietal brain areas involved in the processing of probabilities and rewards.

Smith (2010) investigated the Poisson *shot noise process* (Gardiner, 2004) as an idealized model of neural information coding. He showed that when evidence states are described by the difference of two shot-noise processes, the resulting accumulation process approximates an integrated OU process, or *OU displacement process*. This process, which is the time integral of the OU velocity process in eqn [9], has similar properties to the Wiener diffusion process in Ratcliff's model and, consequently, makes similar RT distribution and accuracy predictions. It thus provides a theoretical account of how diffusive information at a behavioral level can arise from the statistical coding of stimuli in the underlying neural population.

Conclusions

Many researchers have contributed to the development of sequential-sampling models since they were introduced into psychology in the 1960s, and many different models have been proposed of how evidence is accumulated to make a decision. Models that represent evidence accumulation as a diffusion process have been particularly successful in accounting for performance in a wide variety of speeded decision tasks. Neuroscientists have begun to study the brain processes involved in decision making and have reported patterns of activity that appear to be correlates of the psychological processes of evidence accumulation hypothesized in the models. The models thus appear to have the potential to help us understand the theoretical link between behavior and its neural underpinnings.

See also: Computational Neuroscience; Decision and Choice: Economic Psychology; Decision and Choice: Sequential Decision Making; Markov Decision Processes; Prefrontal Cortex.

Bibliography

Audley, R.J., Pike, A.R., 1965. Some alternative stochastic models of choice. British Journal of Mathematical and Statistical Psychology 18, 207–225.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., Cohen, J.D., 2006. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychological Review 113, 700–765.

Busemeyer, J.R., Townsend, J.T., 1993. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychological Review 100, 432–459.

Diederich, A., Busemeyer, J.R., 2003. Simple matrix methods for analyzing diffusion models of choice probability, choice response time, and simple response time. Journal of Mathematical Psychology 47, 304–322.

Ditterich, J., 2006. Evidence for time-variant decision making. European Journal of Neuroscience 24, 3628–3641.

Edwards, W., 1965. Optimal strategies for seeking information: models for statistics, choice reaction times, and human information processing. Journal of Mathematical Psychology 2, 312–329.

Feller, W., 1967. An Introduction to Probability Theory and Its Applications, third ed. Wiley, New York.

Gardiner, C.W., 2004. Handbook of Stochastic Methods, third ed. Springer, Berlin.

Glimcher, P.W., Camerer, C.F., Fehr, E., Poldrack, R.A., 2009. Neuroeconomics: Decision Making and the Brain. Elsevier, Amsterdam.

Gold, J.I., Shadlen, M.N., 2002. Banburismus and the brain: decoding the relationship between sensory stimulus, decisions and reward. Neuron 36, 299–308.

Hanes, D.P., Schall, J.D., 1996. Neural control of voluntary movement initiation. Science 274, 427–430.

LaBerge, D.A., 1962. A recruitment theory of simple behavior. Psychometrika 27, 375–396.

Laming, D.R.J., 1968. Information Theory of Choice Reaction Times. Wiley, New York.

Link, S.W., Heath, R.A., 1975. A sequential theory of psychological discrimination. Psychometrika 40, 77–105.

Luce, R.D., 1986. Response Times. Oxford University Press, New York.

Mulder, M.J., Wagenmakers, E.J., Ratcliff, R., Boekel, W., Forstmann, B.U., 2012. Bias in the brain: a diffusion model analysis of prior probability and potential payoff. Journal of Neuroscience 32, 2335–2343.

Purcell, B.A., Schall, J.D., Logan, G.D., Palmeri, T.J., 2012. From salience to saccades: multiple-alternative gated stochastic accumulator model of visual search. Journal of Neuroscience 32, 3433–3446.

Ratcliff, R., 1978. A theory of memory retrieval. Psychological Review 85, 59–108.

Ratcliff, R., Rouder, J.N., 1998. Modeling response times for two-choice decisions. Psychological Science 9, 347–356.

Ratcliff, R., Smith, P.L., 2004. A comparison of sequential sampling models for two-choice reaction time. Psychological Review 111, 333–367.

Ratcliff, R., Philiastides, M.G., Sajda, P., 2009. Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. Proceedings of the National Academy of Sciences of the United States of America 106, 6539–6544.

Ratcliff, R., Van Zandt, T., McKoon, G., 1999. Connectionist and diffusion models of reaction time. Psychological Review 106, 261–300.

Ratcliff, R., Thapar, A., Smith, P.L., McKoon, G., 2005. Aging and response times. In: Duncan, J., Phillips, L., McLeod, P. (Eds.), Measuring the Mind. Oxford University Press, Oxford, UK.

Ratcliff, R., Hasegawa, Y., Hasegawa, R., Smith, P.L., Segraves, M., 2007. A dual diffusion model for single cell recording data from the superior colliculus in a brightness discrimination task. Journal of Neurophysiology 97, 1756–1774.

Roe, R.M., Busemeyer, J.R., Townsend, J.T., 2001. Multialternative decision field theory: a dynamic connectionist model of decision-making. Psychological Review 108, 370–392.

Roitman, J.D., Shadlen, M.N., 2002. Responses of neurons in the lateral interparietal area during a combined visual discrimination reaction time task. Journal of Neuroscience 22, 9475–9489.

Smith, P.L., 1990. A note on the distribution of response time for a random walk with Gaussian increments. Journal of Mathematical Psychology 34, 445–459.

Smith, P.L., 1995. Psychologically principled models of visual simple reaction time. Psychological Review 102, 567–591.

Smith, P.L., 2000. Stochastic dynamic models of response time and accuracy: a foundational primer. Journal of Mathematical Psychology 44, 408–463.

Smith, P.L., 2010. From Poisson shot noise to the integrated Ornstein–Uhlenbeck process: neurally principled models of information accumulation in decision-making and response time. Journal of Mathematical Psychology 54, 266–283.

Stone, M., 1960. Models for choice-reaction time. Psychometrika 25, 251–260.

Townsend, J.T., Ashby, F.G., 1983. Stochastic Modeling of Elementary Psychological Processes. CUP, Cambridge.

Tuerlinckx, F., Maris, E., Ratcliff, R., De Boeck, P., 2001. A comparison of four methods for simulating the diffusion process. Behavior Research Methods, Instruments, and Computers 33, 443–456.

Usher, M., McClelland, J.L., 2001. The time course of perceptual choice: the leaky, competing accumulator model. Psychological Review 108, 550–559.

Vickers, D., 1970. Evidence for an accumulator model of psychophysical discrimination. Ergonomics 13, 37–58.

Voss, A., Voss, J., 2008. A fast numerical algorithm for the estimation of diffusion model parameters. Journal of Mathematical Psychology 52, 1–9.

Wagenmakers, E.-J., Brown, S., 2007. On the linear relationship between the mean and standard deviation of a response time distribution. Psychological Review 114, 830–841.

Wald, A., 1947. Sequential Analysis. Wiley, New York.

Diffusion: Anthropological Aspects

Ronald Stade, Malmö University, Malmö, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

To study phenomena like communication, migration, epidemics, and trade is to study diffusion. In social and cultural anthropology, the systematic study of cultural diffusion emerged in contradistinction to speculations about humanity's cultural evolution from primitive origins to a state of civilization. The research of diffusion has been a constant in anthropology ever since, although the designations for the topics of this research have shifted. This article presents key topics in the anthropological study of cultural diffusion, including acculturation, world system, globalization, and the social life of things.

Diffusion Studies and Evolutionism

In a social-scientific sense, diffusion is the spreading in space and time of animate and inanimate objects, including human beings, and material and immaterial forms. To study phenomena like communication, migration, epidemics, and trade is to study diffusion. The question diffusion studies ask is: How did this (or he or she) get from there to here, from then to now? The focus in diffusion studies can be either on that which is diffused, e.g., in terms of isomorphism, relatedness, frequency, and comparison, or on the process of diffusion itself, e.g., in terms of contacts, paths, networks, and patterns. In social and cultural anthropology, the study of diffusion, on the whole, has gradually moved from a focus on that which is diffused to the process of diffusion itself. According to the historian of anthropology, George Stocking (1995: 180), diffusion is "one of the ur-forms of anthropological speculation in the Western-European intellectual tradition." Historically, however, the systematic concern with cultural diffusion in social and cultural anthropology emerged as an explicitly antispeculative project. Initially, it developed from the close study of artifacts that had been collected by Europeans on their journeys to different parts of the world. In private and public collections, such objects had been organized according to the assumption that they represented different stages in a process of orthogenetic human evolution. This idea, which had been common since the Enlightenment, is succinctly encapsulated in the title of one of Lewis Henry Morgan's (1818–81) works: *Ancient Society, or Researches in the Lines of Human Progress from Savagery Through Barbarism to Civilization* (1877). One of German anthropology's pioneers, Adolf Bastian (1826–1905), rejected the method of evolutionary deduction in ethnography, according to which individual ethnographic objects and traits were placed in a singular sequence of cultural achievement. He insisted on induction as the only scientifically valid method: it was possible to infer from the particular instances of artifacts and ethnographic data to the general level of 'folk ideas,' i.e., the environment-specific configurations of humanity's commonly held 'elementary ideas' (the latter concept inspired Carl Jung to formulate his idea of 'archetypes').

Diffusionism

Bastian was in turn criticized for being unscientific by another German anthropologist, Friedrich Ratzel (1844–04), who objected to his explaining cultural differences in terms of psychological maturity. Bastian believed that the technical sophistication of objects reflected the level of psychological maturity a cultural group had reached. Ratzel thought that charting the paths of cultural diffusion ought to come before causal explanations: "The first question here is not: Why?, but Where? And the second question after that is: Whence?" (Ratzel, 1897: 270). Bastian and Ratzel stood on opposite sides not just when it came to anthropological theory and method but with regard to politics as well (see Smith, 1978). Bastian advocated a liberal agenda of German industrialization and international free trade; Ratzel believed in protecting German agriculture through tariffs and national expansion. Bastian (1893) publicly opposed the acquisition of new German colonies; Ratzel (1884) was convinced that colonization through Germans settling overseas could solve social problems at home and lend greatness to the newly united German Reich. Ratzel's political stance fused with his theoretical program, which rightfully can be defined as the beginning of diffusionism in anthropology. Thus, just as in British and American anthropology, the politics of colonization and colonialism, as well as modernization, were intertwined with theory in German anthropology.

Since Ratzel believed that the principal mechanism of diffusion was migration, he arrived at the conclusion that "migration theory is the fundamental theory of world history" (Ratzel, 1882: 464). Fritz Graebner (1877–34), whose diffusionist *Methode der Ethnologie* (1911) was the period's most influential theoretical and methodological work of German anthropology, wrote that Ratzel's focus on migration and cultural borrowing put in question assumptions about organically grown and integrated local cultures and opened up the possibility that whatever one finds in any one location is a cultural hybrid ('Kulturmischung' in German; Graebner, 1911: 93). In 1904, Graebner and his colleague Bernhard Ankermann (1859–43) outlined their self-consciously formulated diffusionist project at the meeting of the Berliner Gesellschaft für Anthropologie, Ethnologie und Urgeschichte. Two years later, Father Wilhelm Schmidt in Vienna founded the foremost diffusionist journal *Anthropos*. At University College, London, Grafton Elliot Smith and his disciple William Perry tried to prove that culture had diffused to the rest of the

world from a single center of innovation, namely, ancient Egypt:

> … towards the close of the period of the New Empire, or perhaps even a little later, a great many of the most distinctive practices of Egyptian civilization suddenly appeared in more distant part of the coast-lines of Africa, Europe, and Asia and also in course of time in Oceania and America (Smith, 1916: 3).

As chief agents of diffusion, Smith and Perry identified the Phoenicians. W.H.R. Rivers, who worked at Cambridge, UK, and who had been a member of the Torres Strait Expedition, emerged as the most influential British anthropologist in the first two decades of the twentieth century. After conducting\fieldwork among the Todas of southern India in 1902, Rivers turned to 'ethnological analysis,' i.e., diffusionism, in which he combined influences from Elliott Smith, Graebner, and Sigmund Freud (cf. Rivers, 1911).

From Diffusionism to Acculturation Studies

As mentioned, a significant source of diffusionism was the political question of colonialism. Another source was the scholarly issue of historiography. Diffusionism is interspersed with implicit (and a few explicit) references to Leopold von Ranke's (1795–86) historicism, which emphasizes that one needs to move as closely as possible to the sources of knowledge (empiricism) and that one must understand the past in its own terms (relativism). This legacy migrated from Germany to the United States and shaped American cultural anthropology.

The German anthropologist Franz Boas (1858–42), who, for a while, had been a member of Bastian's staff at the ethnographic museum in Berlin and who was well acquainted with the debates in German anthropology, settled in the United States in the late 1880s. In the Pacific Northwest, he went on to collect specimens, information in the native language about these specimens, and grammatical information necessary for the interpretation of this information. His ambition was to reconstruct a cultural history of the nonliterate indigenous people he studied. Boas developed his diffusionary analyses most systematically in the field of linguistics. At the same time, Boas was highly critical of certain types of diffusionism. In particular, he rejected the idea that cultural traits persist over time and that they were distributed from a single center through man's migration across continents (see Boas, 1920). Cultural traits do not remain unchanged, contended Boas; all cultures have a history, which includes both acculturation and innovation or adaptation.

In the face of a methodological shift in anthropology, with direct observation during extended times of fieldwork replacing both the study of museum specimens and survey studies with the help of questionnaires, the reconstructive work of diffusionism fell into disrepute in much of anthropology. Cultural interaction, however, continued to be a core issue and was now studied under the headings of culture contact, culture change, and acculturation. Acculturation, the key concept in this context, was defined as follows: "Acculturation comprehends those phenomena which result when groups of individuals having different cultures come into continuous first-hand contact, with subsequent changes in the original cultural patterns of either or both groups" (Redfield et al., 1936: 149). The image of the World War I as a conflict in which the products of industrialization – machines, chemicals, automatic weapons, etc. – mangled human bodies cast a shadow over the belief in evolutionary progress. Following World War I, culture contact had become a matter of self-conscious reflection among European intellectuals:

> The great world-shattering changes of the past decade, the War, the Russian Revolution, the Treaty of Versailles and the consequences of these events, have led men to question for the first time many of the previously unchallenged assumptions implicit in their own civilization, whilst it has prompted them to regard with real concern and a desire to understand civilizations unlike their own (Pitt-Rivers, 1927: xii).

Some acculturation studies, like those of Pitt-Rivers, were explicitly directed at the types of political domination that in so many instances produced culture change during colonialism. Thus, Richard Thurnwald (1932: 568f) distinguished between two kinds of acculturation through colonial relations: on the one hand, the biological and cultural mixing between men and women, and on the other, formal relations of tutelage. In the one instance, acculturation was a matter of intimate relations, of primary socialization into local traditions in the form of care, which attributed to women as 'principal caretakers,' the role of keepers of tradition. In the other case, social relations consisted of formal education or secondary socialization through 'channels by which the new and foreign flows,' such as 'missions, schools, governmental ordinances, control of courts, etc.' Thurnwald's analysis is typical of a move in diffusion studies from a focus on that which is diffused to the process of diffusion itself. In his model, cultural diffusion in the context of intimate relations provides cultural continuity; diffusion in vertically structured formal settings of colonial relations introduces foreignness. The latter circumstance has been analyzed, e.g., by Frantz Fanon in terms of the colonized self absorbing the image of inferiority, which the colonizer projects onto her or him (see Fanon, 1952).

Decades later the difference between types of cultural transmission was discussed in terms of 'antagonistic acculturation.' American anthropologists began to investigate why some cultural groups resisted formal colonial tutelage and the 'modernization' it was to bring, and others did not. The cause was sought in the rigidity and closeness of some societies versus the adaptability and openness of others. One suggestion was that it was possible to distinguish 'hard-shelled, vertebrate' from 'soft-shelled, invertebrate' cultural systems (Social Science Research Council Summer Seminar, 1954: 978). For instance, two Native American groups were studied with regard to how they received tribal members who returned as World War II veterans (Adair and Vogt, 1949). One group was described as engaging in 'antagonistic acculturation,' that is, to resist culture change. The other group was described as 'receptive,' 'flexible,' and 'intensely curious about the outside world' (Adair and Vogt, 1949: 558–60). The latter seemed to be

what Margaret Mead (1938) once called 'an importing culture.' These characteristics were attributed to different types of social organization, ritual systems, and collective dispositions. The attention to relations of domination that had been present in studies like those of Pitt-Rivers and Thurnwald gave way to a perspective that, at least implicitly, valued cultures that were open and adaptive to modernization.

Critical Studies in Diffusion

Subsequently, mainstream studies of culture change were less modernization friendly. Not simply because their authors promoted the preservation of indigenous cultures, as generations of anthropologists had done before, but because they placed colonial relationships in a wider, global system of capitalism. Since the late 1960s, a critical anthropology of neo-colonialism and global capitalism developed under the influence of Andre Gunder Frank's (1969) thesis that underdevelopment was not a result of traditionalism, but of colonial exploitation (the development of underdevelopment), and of Immanuel Wallerstein's (1974) synthetic work about the dynamics of European expansion and the center–periphery structure of the capitalist world-system. This type of anthropology tended to treat cultural diffusion as a secondary effect of worldwide economic structures and relations. It did not grant an independent cultural logic to the global periphery, holding the periphery to be open to radical change initiated by the center. Marshall Sahlins (1985: viii) called this 'a confusion between an open system and a lack of system.'

In response to models of cultural imperialism that treated the so-called periphery as susceptible to anything arriving from the global centers, anthropologists began describing local responses to colonial and capitalist diffusion. Ethnographers had for some time provided examples of culturally creative responses to capitalism's local encroachments in different parts of the world (see Nash, 1981 for a review of this rich ethnographic genre). Classic examples include Eric Wolf's (1958) research of syncretism (i.e., the synthesis of indigenous and diffused cultural elements) in the case of the Virgin of Guadalupe; the ethnographic documentary film, *Trobriand Cricket: An Ingenious Response to Colonialism* (1973) by Gary Kildea and Jerry Leach; and Michael Taussig's ethnography, *The Devil and Commodity Fetishism in South America* (1980). Syncretism, hybridity, and creolization were key concepts in this anthropological genre.

Beginning in the 1980s, a number of anthropologists turned their attention to the global aspects of cultural diffusion. At issue was the question whether what was described as a process of cultural globalization would lead to cultural homogenization (coca-colonization) or if, on the contrary, and in the spirit of the just mentioned concern with local responses to colonialism and global capitalism, globalization actually created new forms of cultural diversity (see, e.g., Hannerz, 1987).

The Social Life of Things

Of some interest in this context is Arjun Appadurai's anthology on *The Social Life of Things: Commodities in Cultural Perspective* (1986), in which a number of scholars revisit the subject of material diffusion, only that they now follow artifacts and commodities through the networks of diffusion and investigate how the cultural meanings attached to the objects change in the process. What becomes apparent in contributions like that of Brian Spooner (1986) is that the path of commodities through international, vertically structured trade networks can involve complex flows of values and meanings both upstream and downstream.

Appadurai's approach – treating objects as if they are animate (as if they have social lives and cultural biographies) – is reminiscent of current preoccupations with material–semiotic relations in the anthropological study of science and technology, e.g., as conducted and elaborated by the French anthropologist Bruno Latour. Latour, in turn, acknowledges his debt to the French sociologist Gabriel Tarde (1843–04), who posited that the laws of diffusion (or, as he called it, 'imitation') are fundamental to all processes in both natural and social life (see Tarde, 1890). Compared with Appadurai's empirically informed method of following the object in space and time and registering how it, during its journey, is charged with values and meanings, Tarde's laws are metaphysical and in need of translation to be applied in social-scientific research.

In some neighboring disciplines, the idea to trace the diffusion of particular phenomena has been pursued to such an extent, and with such systematic rigor, that it has given rise to subdisciplines and schools of thought. In history, for example, one finds a British and a German version of semantic history, both of which record and analyze the changing meanings of social and political keywords over time. The Cambridge School of History is associated, first and foremost, with the names of J.G.A. Pocock and Quentin Skinner and their research on political languages for the sake of uncovering linguistic universes. In Germany, 'Begriffsgeschichte' (the history of concepts) was pursued as a distinct intellectual project by a group of historians, from which Reinhart Koselleck emerged as the most prominent figure. His research demonstrates that there was a consequential semantic shift in Europe's social and political language in a period stretching approximately from 1750 to 1850.

For the Cambridge School and Begriffsgeschichte, the diffusion of social and political key concepts and the semantic changes they underwent is inferred from the close investigation of historical sources. In sociology, institutionalism – both old and new – also employs induction from historical materials to formulate theories about big structures and large processes. The founder of new institutionalism, John W. Meyer, has been interested in the diffusion of instrumental rationality and its institutions, as well as in the irrational outcomes this diffusion may have (see Meyer and Rowan, 1977, for an early formulation of new institutionalism). In a similar vein, the French anthropologist Loïc Wacquant follows the diffusion of ideological matrixes and the accompanying confusion about the origins and consequences of such matrixes (see, e.g., Wacquant, 1999). In economics, the study of commodity and value chains has become a field of inquiry in its own right (see, e.g., Gereffi and Korzeniewicz, 1994) that has moved from a more descriptive, historical to a more prescriptive, managerial approach.

As a research agenda and method, following objects, concepts, institutions, normative orders, people, etc. along paths of diffusion and migration is far from exhausted. Anthropologists ready to conduct multisite fieldwork and to study primary source materials are in a privileged position to engage in this kind of scholarship with the necessary empirical precision.

See also: Center–Periphery Relationships; Colonialism, Anthropology of; Colonization and Colonialism, History of; Creolization: Sociocultural Aspects; Cultural Evolution: Overview; Diffusion in Sociological Analysis; Exchange in Anthropology; Geographic Thought, History of; Globalization, Anthropology of; Hybridity; Social Epidemiology; Syncretism; World Systems Theory.

Bibliography

Adair, J., Vogt, E., 1949. Navaho and Zuni veterans: a study of contrasting modes of culture change. American Anthropologist 51, 547–561.

Appadurai, A. (Ed.), 1986. The Social Life of Things: Commodities in Cultural Perspective. Cambridge University Press, Cambridge.

Bastian, A., 1893. Zwei Worte über Colonial-Weisheit von jemandem, dem dieselbe versagt ist. Ferd, Dümmlers Verlagsbuchhandlung, Berlin.

Boas, F., 1920. The methods of ethnology. American Anthropologist 22, 311–321.

Fanon, F., 1952. Peau Noire, Masques Blancs. Seuil, Paris.

Frank, A.G., 1969. Latin America: Underdevelopment or Revolution. Monthly Review Press, New York.

Gereffi, G., Korzeniewicz, M. (Eds.), 1994. Commodity Chains and Global Capitalism. Praeger, Westport, CT.

Graebner, F., 1911. Methode der Ethnologie. Carl Winter's Universitätsbuchhandlung, Heidelberg.

Hannerz, U., 1987. The world in creolization. Africa 57, 546–559.

Mead, M., 1938. The Mountain Arapesh. Part 1: an importing culture. American Museum of Natural History Anthropological Papers 36, 139–349.

Meyer, J.W., Rowan, B., 1977. Institutionalized organizations: formal structure as myth and ceremony. American Journal of Sociology 83, 340.

Morgan, L.H., 1877. Ancient Society, or Researches in the Lines of Human Progress from Savagery through Barbarism to Civilization. Barrati Library, Calcutta, India.

Nash, J., 1981. Ethnographic aspects of the world capitalist system. Annual Review of Anthropology 10, 393–423.

Pitt-Rivers, G.H.L.-F., 1927. The Clash of Culture and the Contact of Races: An Anthropological and Psychological Study of the Laws of Racial Adaptability, with Special Reference to the Depopulation of the Pacific and the Government of Subject Races. Routledge, London.

Ratzel, F., 1882. Anthropo-Geographie oder Grundzüge der Anwendung der Erdkunde auf die Geschichte. Engelhorn, Stuttgart.

Ratzel, F., 1884. Wider die Reichsnörgler. Ein Wort zur Kolonialfrage aus Wähler kreisen. R. Oldenbourg, Munich.

Ratzel, F., 1897. Die geographische Methode in der Ethnographie. Geographische Zeitschrift 3, 268–278.

Redfield, R., Linton, R., Herskovits, M., 1936. Memorandum for the study of acculturation. American Anthropologist 38, 149–152.

Rivers, W.H.R., 1911. The ethnological analysis of culture. In: Smith, G.E. (Ed.), Psychology and Ethnology. Trubner, London, pp. 120–140.

Sahlins, M., 1985. Islands of History. University of Chicago Press, Chicago, IL.

Smith, G.E., 1916. The influence of ancient Egyptian civilization in the east and in America. Bulletin of the John Rylands Library. January–March, Manchester.

Smith, W., 1978. The social and political origins of German diffusionist ethnology. Journal of the History of the Behavioral Sciences 14, 103–112.

Social Science Research Council Summer Seminar, 1954. Acculturation: an exploratory formulation. American Anthropologist 56, 973–1000.

Spooner, B., 1986. Weavers and dealers: the authenticity of an oriental carpet. In: Appadurai, A. (Ed.), The Social Life of Things: Commodities in Cultural Perspective. Cambridge University Press, Cambridge, pp. 195–235.

Stocking Jr., G., 1995. After Tylor: British Social Anthropology, 1888–1951. University of Wisconsin Press, Madison, WI.

Tarde, G., 1890. Les Lois de l'imitation: Études sociologique. Alcan, Paris.

Taussig, M., 1980. The Devil and Commodity Fetishism in South America. University of North Carolina Press, Chapel Hill, NC.

Thurnwald, R., 1932. The psychology of acculturation. American Anthropologist 34, 557–569.

Wacquant, L., 1999. Les prisons de la misère. Raisons d'agir, Paris.

Wallerstein, I., 1974. The Modern World-System: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century. Academic Press, New York.

Wolf, E., 1958. The virgin of Guadalupe: a Mexican national symbol. Journal of American Folklore 71, 34–39.

Diffusion: Geographical Aspects

Lawrence A Brown[†], The Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Geographic diffusion involves the propagation of innovation or other phenomena, focusing on spread from place to place, leading to maps that differ from one time to another. Innovations have received most attention – a new product, idea, technology, organizational structure, or way of doing things – where new means new to a particular place through adoption. Geographical diffusion, a long research tradition, has changed focus and content from epoch to epoch in social science, ranging from cultural formation to economic phenomena and to positivist social science. Prime among the foci are the following: (1) The Cultural Geography Tradition – in the early middle 1900s, the concern was cultural formation; (2) The Hagerstrand Tradition – beginning in the 1970s, the focus turned to structural influences on diffusion; (3) The Market and Infrastructure Perspective – emphasis shifted to the spread of technological and organizational innovation in economic activity; (4) Technological Diffusion – focus is mainly on Third World development; as well as (5) Development as a Diffusion Process. At the century's end, innovation itself became a major focus wherein diffusion within such endeavors is a primary element, that is (6) Diffusion and Innovation. Geographical diffusion represents one approach in a multidisciplinary suite that includes contributions from anthropology, communications, economics, marketing, political science, psychology, sociology. Accordingly, constructs of geographic diffusion may also occur in research by others, for example, interpersonal interactions, media effects, utility maximization, and marketing strategies. In these instances, the critical contributions of a spatial perspective to understanding diffusion dynamics, and innovation itself, are apparent.

Introduction

From a geographic perspective, diffusion involves the propagation of a phenomenon in a manner that it spreads from place to place; incorporates basic geographic elements of distance, direction, and spatial variation; and leads to maps that differ from one time to another. Most commonly the phenomenon is an innovation – a new product, new idea, new technology, new organizational structure, or new way of doing things – where 'new' means new to a particular place by way of adoption. Hence, diffusion phenomena cover a wide range that includes transportation modes such as the automobile; farming techniques; family planning; credit cards; broadcast and cable television; shopping centers; production practices, such as the assembly line or just-in-time inventorying; political movements; cultural practices; frontier development; modernization in Third World settings; epidemics; urban ghettoes; and urban areas themselves. The enormous range of phenomena studied under the diffusion rubric is further indicated by Brown (1981), Morrill et al. (1988), and Rogers (2003).

While geographers have studied diffusion since early in the twentieth century, the focus has shifted over time, often with marked departures from prevailing norms. Consequently, early diffusion studies by cultural geographers may be hardly recognizable as similar in genre to recent studies, focusing on entrepreneurial agents in diffusing products and diffusion itself as a major factor in innovation.

The Cultural Geography Tradition

Within geography, diffusion studies originated through a concern with culture in terms of

> the changing distribution of 'culture areas' … identify[ing] environmental features characteristic of a given culture and … discover[ing] what role human action … has played in creating and maintaining given geographic features.
>
> Wagner and Mikesell, 1962: p. 1

This tradition dominated pre–World War II scholarship and focused on change in indigenous cultures, often centuries ago. Culture and the human or cultural landscape are the primary focus, and diffusion studies are relevant in clarifying questions concerning culture origins, the spatial configuration of culture areas, and culture-related or culture-influenced features on the Earth's surface. Carl Sauer was the central figure in this approach, which tended to be closely aligned with and complementary to some concerns of anthropology (Mikesell, 1978).

Within this framework, many studies aim to assess the relative importance of diffusion vs simultaneous invention in development of cultural landscapes. They employ an empirical, inductive approach with emphasis on landscape development in prehistory and/or over considerable periods of time. Often investigation considers a single phenomenon, seen as a culture trait, and explains its observed spatial distribution as the outcome of processes unfolding over space and time. A typical procedure is to examine a phenomenon's spatial distribution at different times or its frequency of occurrence in different locations at a single time. Inferences are drawn regarding origins (or cultural hearths), means, and routes of dispersal of both the observed phenomenon (culture trait) and the ideas, technology, or culture for which it is a surrogate. Attention is also given to cultural and environmental conditions necessary for adoption of particular phenomena and to macroscale, big-picture phenomena such as the domestication of plants and animals, their origins, routes of dispersal, and destinations (Sauer, 1952; Wagner and Mikesell, 1962).

[†] Deceased.

In stressing the role of diffusion in development of the human landscape, many cultural geography studies proceed further to consider indirect effects of the culture trait innovation on the physical and cultural environment. A basic notion is that the diffusing item is both a stimulus to new innovation and itself subject to modification as it spreads. The relation between diffusion, the item being diffused, and the human landscape is therefore complex and subject to continual change.

The Hagerstrand Tradition

Sauer's counterpart in later diffusion research is Torsten Hagerstrand who, in the early 1950s, initiated a shift toward the generative processes underlying diffusion and modern-day products and practices. As this focus is consistent with Sauer's conceptualization, why was process less prominent in earlier research? One reason may be that identification of processes controlling locational change is less feasible in the sparsely documented situations studied by cultural geographers. Second, Hagerstrand's main impact was on researchers in economic and urban geography who were primarily concerned with location and locational processes rather than landscape. Illustrative is Haggett's (1965: pp. 10–13) classification of research under the Sauer school as landscapist or ecologist and research under the Hagerstrand school as locationalist.

Hagerstrand's research (1952, 1967) focuses on diffusion of contemporary innovations such as the automobile and bovine tuberculosis control. His first study, *The Propagation of Innovation Waves* (1952), continues the cultural geography tradition by its use of an inductive approach and concern with diffusion pattern rather than process. At a somewhat later date (1953), however, Hagerstrand observed that…

> spatial order in the adoption of innovations is … so striking that it is tempting … to create theoretical models which simulate the process and make predictions achievable.
>
> Berry and Marble, 1968: p. 370

This statement indicates a change from description and inductive generalization to a more deductive approach focusing on generative processes. The work marking this shift is Hagerstrand's PhD dissertation of 1953, subsequently translated into English as *Innovation Diffusion as a Spatial Process* (1967). Hagerstrand made three contributions that were transformative at the time, and continue to play a role in diffusion research.

Empirical Regularities in Diffusion

Three regularities characterize diffusion. Over time, a graph of the cumulative level of adoption is expected to approximate an S-shape that rises with the passage of time. In an urban system, diffusion is expected to proceed from larger to smaller centers, regularity termed the hierarchy effect. Within the hinterland of an urban center, diffusion is expected to proceed in a wavelike fashion outward from the urban center, first hitting nearby rather than farther-away locations, and a similar pattern is expected in diffusion among rural populations. This regularity is termed the neighborhood or contagion effect.

Hagerstrand (1952) also provides a framework that connects these regularities. For diffusion in an urban system, this posits three phases: a primary stage, during which initial diffusion centers are established; a diffusion stage, during which neighborhood-type diffusion occurs in areas near diffusion centers and secondary diffusion centers are established in lower order urban places; and a condensing or saturation stage, during which filling in occurs and diffusion eventually ceases.

This framework thus illustrates the complementarity of hierarchy and neighborhood effects and their temporal positioning within diffusion. Also, if locales of adopters were tabulated, higher order urban places characterize the left tail of the S-curve of adoption; lower order urban places characterize the middle, sharply rising portion of the S-curve, along with rural hinterlands in proximity to urban centers; and remote rural hinterlands characterize the S-curve's right tail.

Hagerstrand's Conceptualization of the Diffusion Process

Hagerstrand's conceptualization of innovation diffusion across the landscape holds that adoption is primarily the outcome of a learning or communications process. Factors related to the effective flow of information are most critical and, therefore, a fundamental step in examining diffusion is identifying spatial characteristics of information flows and resistance to adoption.

In Hagerstrand's conceptualization, information originates from the mass media or past adopters. The destination of personal messages depends on a sender's network of social communications (interpersonal contacts), and configuration of this network reflects various social and terrestrial barriers that impede, divert, and/or channel communications. Particular attention is given to terrestrial barriers such as lakes, forests, difficult terrain, and geographical distance separating potential communicants.

The conceptualization further recognizes that resistance levels differ among individuals, and that higher levels of resistance require more information for adoption to occur. This leads to a dichotomy of resistance as either the result of values that are inconsistent with adoption (social resistance) or practicalities that make adoption difficult or impossible (economic resistance).

To address diffusion at different geographical scales, a hierarchy of networks of social communication is posited. One network might operate locally and another regionally. The first would control diffusion among farmers in a single locale, whereas diffusion at the central-place level of aggregation might utilize a regional network of persons in different urban areas that communicate with one another. Superimposition of these two levels provides a comprehensive picture of diffusion in the area represented.

In summary, Hagerstrand's conceptualization proposes transformation of a population from a low proportion of adopters to one with a high proportion, by means of information dissemination through media and interpersonal contact. Spatially, the principal mechanisms of transformation

are networks of social communication characterized by biases and distortions that, in turn, effect and are reflected in spatial patterns of innovation diffusion.

Hagerstrand's Methodological Framework

To implement his conceptualization, Hagerstrand used Monte Carlo simulation. Adoption takes place after a specified number of messages are received, the specified number varying according to an individual's resistance to adoption, which is randomly assigned. Time is treated in discrete units of equal interval, and each past adopter sends a message in every time interval. The destination of messages depends on the probability of contact between tellers and potential receivers. This probability is a function of distance between them unless a water or forest barrier intervenes. A distance decay function is employed to compute the probability of interaction between each of 25 cells in a 5-by-5 cell grid and the center cell of that grid. The resulting grid of probabilities, termed the mean information field (MIF), is employed with a map gridded in a similar manner, but with more than 25 cells. The MIF is centered over a map cell containing an adopter of the innovation at time t; a random number is then drawn to determine which of the 25 surrounding map cells receives a message to adopt in time $t + 1$. If the message passes over a water or forest barrier, it is either nullified or a new random number is drawn, depending on the level of obstruction. This procedure is repeated for every adopter in every time interval. If a previous adopter receives a message, the message is lost; that is, it has no effect and is not repeated. If a message is received by a nonadopter who, owing to resistance level, requires more than one message before adoption, the message is stored until sufficient messages for adoption are accumulated.

This model mechanism is crude by today's standards of computational capability, but in the 1950s, it was highly innovative. Accordingly, Hagerstrand's diffusion model played an enormous role in geography's quantitative revolution of the 1960s and the 1970s, as both an example and a standard of excellence (Berry and Marble, 1968). This observation also applies to social science overall; the model is included, for example, in books such as Coleman's (1964) *Introduction to Mathematical Sociology*.

The Market and Infrastructure Perspective

Hagerstrand's formulation, and those of the many other social scientists studying diffusion from the 1950s onward, took an atomistic view of society as individuals with free will, any of whom could adopt an innovation. Thus, the onus of adoption was on individuals, who were classified on a scale ranging from innovator to laggard, with the social meaning each term implies (Rogers, 2003). This is termed the adoption perspective, focusing primarily on the demand side of innovation diffusion.

The market and infrastructure perspective focuses on supply or availability of the innovation (Brown, 1981). Rather than assuming that all have an equal opportunity to adopt, this is seen as purposely and egregiously unequal, so that adoption represents a choice within a constraint set; government, nonprofit, and entrepreneurial institutions establish these constraints; and attention is given to the process by which innovations and conditions for adoption are made available to individuals, households, or firms. Hence, focus shifts from the individual to the structure within which individuals operate, paralleling realignments in social science overall (Brown, 1999). Nevertheless, the market and infrastructure perspective is put forth as complementary, not contradictory, to the adoption perspective. Broad outlines of diffusion are shaped by where (and to whom) innovations are (made) available, but within those outlines, more detailed variation in diffusion patterns reflects adoption behavior.

Central elements of this perspective include the propagator of the innovation, the entity behind its diffusion; diffusion agencies (outlets) through which innovation is distributed to the population at large; and diffusion strategy, which is designed to encourage adoption and includes decisions on agency location, pricing, advertising, population segments targeted by promotional efforts, etc. Hence, locations and temporal sequencing of diffusion agencies determine where and when an innovation will be available and provide the general outline of the diffusion pattern. Diffusion strategy further defines the pattern by creating different levels of access depending on a potential adopter's economic, locational, demographic, and social characteristics. The final pattern occurs via adoption behavior, the last stage of the process. Interacting with these factors are infrastructures such as transportation and utility networks that are independent of the innovation's propagator and diffusion agencies, distinctly affect adoption, and upon installation may be seen as 'enabling' innovations.

Hence, the market and infrastructure perspective emphasizes the marketing aspects of diffusion, especially strategies for market penetration. Whether the propagator is profit or nonprofit motivated, broad acceptance is a common goal, and a strategy will be designed to accomplish this, that is, to best penetrate the market. This is also true whether propagation involves centralized decision making, decentralized decision making with a coordinating propagator, or decentralized decision making without a coordinating propagator, although such distinctions affect procedures by which diffusion occurs. Similarly, diffusion agencies may be represented by salespeople or agricultural extension agents but should be guided by marketing strategy and, accordingly, make selective decisions, regarding which potential adopters will be contacted or cultivated, and to what degree. Considering the role of infrastructure, broadcast television stations were needed before television sets diffused. Today, fiber-optic and satellite connectivity varies appreciably from place to place, affecting Internet availability. Enhanced accessibility through communication or transportation improvements increases the market areas of diffusion agencies and renders innovations more attainable to potential adopters. Also, for example, irrigation infrastructure is often critical to the success of high-yielding variety seeds in agriculture.

Finally, this framework calls into question hierarchy and neighborhood effect regularities, because relevant factors from a marketing perspective are often not arranged in that manner. The diffusion of cable television in Ohio, for example, occurred initially in small, more remote places where broadcast signals

were weak or nonexistent, because cable was a means of enhancing television sales by appliance stores.

Technological Diffusion

Agricultural technology has been a major concern of diffusion studies. Beginning in the 1930s, and accelerating thereafter, rural sociologists and agricultural economists, working through agricultural experiment stations and extension agents, carried out an enormous number of projects focused on farm innovations, such as hybrid corn, fertilizers, silage methods, high-yielding varieties, and the like (Rogers, 2003). Until recent industrial agriculture, the adopting unit was the farm household, thus joining this effort with diffusion research concerned with consumer innovations. Manufacturing firms that are well beyond household size also adopt technological innovations, as do service and retail firms. This has been studied extensively, but rarely joined with the body of diffusion research described previously.

Rather than emphasizing communications, studies of firm adoption give more attention to characteristics of the firm and innovation itself. Also significant is the fact that innovation replaces an older technology with a newer one, highlighting comparative costs, benefits, and risks as well as factors such as relative profitability, investment, sunk costs, and the like. In this regard, there is greater emphasis on innovation as a continuous process; one that might be tailor-made for a particular firm or firm setting, rather than as a singular, unchanging phenomenon; and one that involves a more intimate and ongoing relationship between the innovation's supplier and user. Issues such as competition among firms and relative factor prices, favoring one technology over another, also play a role.

As these factors indicate, economic perspectives are important to investigating the diffusion of technological innovations among firms (Brown, 1981). But spatial considerations enter in two ways. First, there has been a tendency to assume that larger firms in larger urban agglomerations are earlier adopters, giving rise to a self-reinforcing (cumulative and circular causation; Myrdal, 1957) hierarchy effect pattern of diffusion. In fact, earlier adopters are just as likely to be less than the largest in size, which may correlate with aggressiveness and innovativeness of management. There also may be a spatial order to innovative change, such that newer economic regions are favored, leading to regional rotation or spatial succession; Silicon Valley provides an example (Malecki, 1997). The second way spatial considerations come into play is that the study of technological innovation has become closely intertwined with regional development and regional economic change (Malecki, 1997). Hence, the questions may not be inherently geographic in terms of process, but very geographic in outcome. This is returned to in Section Diffusion and Innovation.

Development as a Diffusion Process

Seen as integral to economic development, diffusion of innovations in Third World settings has received much attention. Examples include improving agricultural production and living conditions, generating entrepreneurial activity and related employment, promoting family planning, and improving infrastructure. Research and policy have largely employed the adoption perspective, but examples of a market and infrastructure approach are also evident (Brown, 1981; Miller, 2009; Rogers, 2003).

More broadly, development itself has been treated as a diffusion process. One approach involves choosing an index of development, and tracing its spread across the landscape; this approach is similar methodologically to the culture trait studies of earlier diffusion research. Such indices include agricultural cooperatives, banking, communications (telegraph, telephone), educational facilities, electricity, employment patterns across economic sectors, medical services, police organizations, postal agencies, transportation modes (roads, rail), urbanization, and other artifacts of a market economy (Gould, 1976; Pedersen, 1975; Taaffe et al., 1963).

Frontier expansion has also been viewed in a diffusion framework (Hudson, 1969). Substantively, these studies focus on aspects such as capital generation and control; migration, travel, and trade interactions; the role of particular economic sectors such as banking and wholesaling; and transportation linkages (Meyer, 1980).

Finally, formal models of development have a distinct diffusion dimension. Modernization, dual economy, core periphery, growth center, and stages of growth frameworks put forth the notion that Third World settings are composed of traditional and modern elements, and that development consists of transforming traditional to modern. Dependency and related theories take a similar track, albeit one that differs in emphasis. These models divide Third World landscapes into dynamic, propulsive, modern cores, and traditional peripheries; development is represented by transformation and expansion of the modern; and this occurs through mechanisms such as information and capital flows, migration, human capital developments, and more generally, backwash, trickle-down, and imitation effects (Brown, 1991).

Diffusion and Innovation

Traditionally, diffusion studies have taken the innovation as given and research focuses on the mechanisms by which it penetrates society (Brown, 1981: pp. 3–12; Rogers, 2003: pp. 1–38). That this tradition continues is indicated by studies on the diffusion of the cellular telephone (Comer and Wikle, 2008), broadband availability, and internet usage (Brown, 2010: Whitacre, 2008), and the technologies of continuous steel casting, digital telephones, and shuttleless looms (Perkins and Neumayer, 2005).

This approach is consistent with the traditional 'linear model' of innovation wherein research (basic science, engineering) leads to discovery, production, and marketing as discrete activities. In recent years, however, research has concluded that a 'chain-linked' model is more appropriate, one wherein continual interaction and feedback take place among research, knowledge, invention, testing, design, redesign, potential markets, distribution, etc. (Caraça et al., 2009; Kline and Rosenberg, 1986; Sternberg, 2009). This framework also

applies to the seemingly more singular situation of innovation by users themselves (von Hippel, 2006). From this perspective, then, diffusion becomes an integral element of the innovation process.

Effective interaction among these various segments is enhanced through social and geographic proximity as represented by, for example, a shared knowledge base, language, cultural understanding, technical capacity, institutional practices, governmental policies, and distance itself (Lundvall, 1988). The result is Regional Innovation Systems, each with a distinct spatial expression in the form of innovative clusters, districts, and regions such as Baden-Württemberg, Boston's Route 128, and Silicon Valley (Asheim and Gertler, 2005; Cooke, 2001).

Further, regional innovation and production systems have themselves diffused over time (Castells and Hall, 1994; Malecki, 2010, 2011). Factors underlying this process include the ongoing upsurge in transportation, communication, and information technologies; formation of a global economy and multinational entities; enhanced abilities to tap distant skills and talents; and new organizational forms of production and management. Functionally, these range from highly creative nodes, ones whose innovativeness is more incremental, to sites where production is the sole concern.

Overall Observations

Diffusion comprises a long tradition of research in geography. From epoch to epoch in social science, it has changed focus and content. Geographical studies of diffusion also represent only one approach in that the topic is highly multidisciplinary including, at least, contributions from anthropology, communications, economics, marketing, political science, psychology, and sociology. Accordingly, concepts used by geographic studies of diffusion may be found in research by other disciplines, for example, interpersonal interactions, mass media effects, utility maximization, and marketing strategies. In all these instances, as well as in purely geographic studies, the critical contributions of a spatial perspective to understanding the dynamics of diffusion are apparent.

See also: Development Geographies; Diffusion in Sociological Analysis; Diffusion: Anthropological Aspects; Place in Geography; Scale in Geography; Spatial Data; Spatial Interaction.

Bibliography

Asheim, B.T., Gertler, M.S., 2005. The geography of innovation: regional innovation systems. In: Fagerberg, J., Mowery, D.C., Nelson, R.R. (Eds.), The Oxford Handbook of Innovation. Oxford University Press, Oxford, pp. 291–317.

Berry, B.J.L., Marble, D.F. (Eds.), 1968. Spatial Analysis: A Reader in Statistical Geography. Prentice Hall, Englewood Cliffs, NJ.

Brown, L.A., 1981. Innovation Diffusion: A New Perspective. Methuen, London.

Brown, L.A., 1991. Place, Migration, and Development in the Third World: An Alternative View. Routledge, London.

Brown, L.A., 1999. Change, continuity, and the pursuit of geographic understanding: presidential address. Annals of the Association of American Geographers 89, 1–25.

Brown, L.A., 2010. The market and infrastructure perspective: a missing link in understanding the diffusion of broadband adoption in Oklahoma. The Annals of Regional Science 45, 235–238.

Caraça, J., Lundvall, B.-Å., Mendonça, S., 2009. The changing role of science in the innovation process: from queen to Cinderella? Technological Forecasting & Social Change 76, 861–867.

Castells, M., Hall, P., 1994. Technopoles of the World: The Making of 21st Century Industrial Complexes. Routledge, London.

Coleman, J.S., 1964. Introduction to Mathematical Sociology. Free Press, New York.

Comer, J.C., Wikle, T.A., 2008. Worldwide diffusion of the cellular telephone, 1995–2005. The Professional Geographer 60, 252–269.

Cooke, P., 2001. Regional innovation systems, clusters, and the knowledge economy. Industrial and Corporate Change 10, 945–974.

Gould, P.R., 1976. Tanzania 1920–63: the spatial impress of the modernization process. In: Knight, C.J., Newman, J.L. (Eds.), Contemporary Africa: Geography and Change. Prentice-Hall, Englewood Cliffs, NJ, pp. 423–437.

Hagerstrand, T., 1952. The Propagation of Innovation Waves. Lund Studies in Geography, Gleerup, Lund, Sweden.

Hagerstrand, T., 1967. Innovation Diffusion as a Spatial Process (Trans. of Innovations for loppet ur Korologisk Synpunkt. Gleerup, Lund, Sweden, 1953). University of Chicago Press, Chicago.

Haggett, P., 1965. Locational Analysis in Human Geography. Edward Arnold, London.

Hudson, J.C., 1969. A location theory for rural settlement. Annals of the Association of American Geographers 59, 365–381.

von Hippel, E., 2006. Democratizing Innovation. MIT Press, Cambridge, MA.

Kline, S.J., Rosenberg, N., 1986. An overview of innovation. In: Landau, R., Rosenberg, N. (Eds.), The Positive Sum Strategy: Harnessing Technology for Economic Growth. National Academy Press, Washington, DC, pp. 275–305.

Lundvall, B.-Å., 1988. Innovation as an interactive process: from user-producer interaction to the national system of innovation. In: Dosi, G., Freeman, C., Nelson, R., Silverberg, G., Soete, L. (Eds.), Technical Change and Economic Theory (349–369). Pinter Publishers, London.

Malecki, E.J., 1997. Technology and Economic Development: The Dynamics of Local, Regional, and National Competitiveness. Longman, Essex, UK.

Malecki, E.J., 2010. Global knowledge and creativity: new challenges for firms and regions. Regional Studies 44, 1033–1052.

Malecki, E.J., 2011. Connecting local entrepreneurial ecosystems to global innovation networks: open innovation, double networks and knowledge integration. International Journal of Entrepreneurship and Innovation Management 14, 36–59.

Meyer, D.R., 1980. A dynamic model of the integration of frontier urban places into the United Sates system of cities. Economic Geography 56, 120–140.

Mikesell, M.W., 1978. Tradition and innovation in cultural geography: presidential address. Annals of the Association of American Geographers 68, 1–16.

Miller, D., 2009. Selling Solar: The Diffusion of Renewable Energy in Emerging Markets. Earthscan, London.

Morrill, R., Gaile, G.L., Thrall, G.I., 1988. Spatial Diffusion. Sage, Scientific Geography Series, Newbury Park, CA.

Myrdal, G., 1957. Economic Theory and Underdeveloped Regions. Duckworth, London.

Pedersen, P.O., 1975. Urban-regional Development in South America: A Process of Diffusion and Integration. Mouton, The Hague, Netherlands.

Perkins, R., Neumayer, E., 2005. The international diffusion of new technologies: a multitechnology analysis of latecomer advantage and global economic integration. Annals of the Association of American Geographers 95, 789–808.

Rogers, E.M., 2003. Diffusion of Innovations, fifth ed. Free Press, New York.

Sauer, C.O., 1952. Agricultural Origins and Dispersals. American Geographical Society, New York.

Sternberg, R., 2009. Innovation. In: Kitchen, R., Thrift, N. (Eds.), International Encyclopedia of Human Geography, vol. 1. Elsevier, Oxford, pp. 481–490.

Taaffe, E.J., Morrill, R.L., Gould, P.R., 1963. Transportation expansion in underdeveloped countries: a comparative analysis. Geographical Review 53, 503–529.

Wagner, P.L., Mikesell, M.W. (Eds.), 1962. Readings in Cultural Geography. University of Chicago Press, Chicago.

Whitacre, B.E., 2008. Factors influencing the temporal diffusion of broadband adoption: evidence from Oklahoma. The Annals of Regional Science 42, 661–679.

Diffusion in Sociological Analysis

Alberto Palloni, Center for Demography and Ecology, University of Wisconsin, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by N. Alter, volume 6, pp. 3681–3684, © 2001, Elsevier Ltd.

Abstract

The study of diffusion of behaviors has a long history in sociology. While the key elements of the processes were already well defined in traditional approaches, sophisticated models incorporating all of these ingredients necessitated modern advances in social network theories, models for social interactions, agent-based approaches and, importantly, the widespread use of simulation that could only be launched with remarkable advances in computer power. This article reviews the linkages between traditional and modern approaches to the study of diffusion processes.

Over the last 20 years or so there has been a remarkable swift progress in the formulation and analysis of diffusion processes in social sciences in general and sociology in particular. Past conceptualizations (Alter, 2001) have been superseded through a paradigmatic shift triggered by the introduction of sophisticated behavioral models from other disciplines, particularly economics, the rapid evolution of social network analysis, and the discovery and application of agent-based models (Palloni, 2001, 2003). The transition from the traditional to the new paradigm is, however, not yet complete and there is much ground to cover. Scarcely any of the progress made in the recent past would have been possible in the absence of significant advances in statistical inference, quantum leaps in computing power, and the development of new communication technologies and data gathering procedures.

Traditional Diffusion Theories

The most recent Encyclopedia entry on diffusion processes emphasized a traditional view that was mostly anchored in classic sociological theories and anthropology (Alter, 2001). At the time the entry was written there already was in place a solid sociological tradition, albeit youthful and underdeveloped, best represented by the classic and much cited work of Rogers (Rogers and Kincaid, 1980; Rogers, 1995). Rogers defines diffusion as "… a process by which an innovation is communicated through certain channels over time among the members of a social system …" (Rogers, 1995) or the diffusion effect as "… the cumulative increasing degree of influence upon an individual to adopt or reject an innovation, resulting from the activation of peer networks about an innovation in a social system ..." (Rogers, 1995). Rogers' main contribution at the time was to assemble in a coherent manner the stock of accumulated knowledge by the middle to the late 1970s and to articulate a conceptual framework that became the theoretical scaffold that supported numerous subsequent empirical studies.

The main object of study of standard diffusion theories is the process of adoption of innovations (behaviors, practices, beliefs, norms, etc.). By its very nature this process requires flows of information across actors and/or between a central source and groups of actors. Needless to say, the abandonment of practices, behaviors, beliefs, norms, etc., can also be studied as a diffusion process: in these cases what is adopted is a rejection behavior. Adoption of behaviors or practices is rarely a mechanical matter. Like conventional individual decision processes it involve actors with preferences and tastes, who face price and budget constraints, and who are set to maximize their utility in an environment characterized by uncertainty. Unlike conventional decision-making processes, diffusion involve actors whose calculus is significantly influenced by other group members' attitudes and behaviors (observed or expected) toward adoption.

It is not surprising that the most important application of diffusion studies involves the adoption of technological innovations. Some of the most influential early works include Ryan and Gross's (1943) analysis of the diffusion of hybrid corn, Hagerstrand's (1967) investigation of the diffusion of TB tests in Sweden, Coleman, et al.'s (1966) study of the adoption of tetracycline among Midwestern doctors, Katz and Lazarsfeld's (1955) celebrated formulation of the two-step flow of influence process, Rogers' contribution to the early empirical literature on farm innovations (Beal and Rogers, 1958), and Rogers and Kincaid work on contraceptive behavior (Rogers and Kincaid, 1980). The main goal of all these studies is to assess the effects of the mass media – a central source of diffusion – the degree of influence of individuals located at the top of the community hierarchy (agents of change), and the relative contribution of interpersonal interactions within the boundaries of the community where the innovation is spread. With the exception of the work by Hagerstrand and Ryan and Gross, these studies highlight the role of interpersonal relations and channels of influences as mechanisms fostering or impeding diffusion. They thus anticipate some of the most promising work on social influences in general and diffusion in particular that becomes solidly established and continues to develop from the early 1980s onward (Valente, 1995, 2010; Bandura, 1986; Moscovici, 1985; Marsden and Friedkin, 1993; Marsden, 1998; Erbring and Young, 1979; Granovetter, 1973, 1978).

Developing side by side with the sociological approach emerging in the works identified above, the bulk of empirical research in the area of diffusion rested on the formulation of models testable with aggregate information about behaviors,

such as the proportion of adopters, of susceptible to adopt, and of nonadopters. These formulations mimic contagion models for the spread of disease developed in epidemiology (Mahajan and Peterson, 1985; Anderson and May, 1992; Bailey, 1975). And while its appeal was rooted in their simplicity and low grade data-demands, their generalized use had the unfortunate consequence of portraying the social diffusion process as one where individuals are either passive carriers of information and innovations or passive 'susceptible' rather than actors who are decision makers, engaged in real social interactions, and whose calculus is influenced by other actors' choices and the effects of adoption on the group's morphology itself. Using highly stylized representations, these models eschewed explicit definition of the role of social networks, their social and spatial organization, and the relative positions of individuals in it. In the contagion-like models the richness of social interactions permeating the formulation of diffusion processes in traditional sociological approaches is either absent or concealed by invoking rigid and unrealistic assumptions. Thus, most of these models are deterministic, rest on strong assumptions about temporal and spatial homogeneity and rarely, if ever, explicitly introduce the role of relations between actors. While some of the stochastic variants of contagion models explicitly allow for social and spatial heterogeneity, they are mathematically intractable and, at least until recently, before the advent of powerful computers, led to only modest improvements of our knowledge about diffusion processes (Bartholomew, 1982; Bailey, 1975; Morris, 1993).

Recent research suggests that this state of affairs is beginning to wane rapidly and that the very stylized models of diffusion of old have receded and been replaced by more complex and powerful representations both in investigation of the spread of diseases (Bauch and Galvani, 2013; Brockman and Helbing, 2013) and in social sciences in general (Homer and Hirsch, 2006).

Centrality of Diffusion in Sociological Theory: Structural vs Diffusion Explanations

Full characterization of diffusion models and theories is of relevance because in social sciences in general, and sociology in particular, researchers seek to empirically identify alternative and competing processes or mechanisms that produce similar observable outcomes.

To fix ideas we introduce a simple but useful distinction between structuralist and diffusion explanations of social phenomena. Although we argue below that the stark opposition between these two is ultimately misleading, social sciences is riddled with controversies pitting one against the other and their contrast is useful to define more precisely the nature of a diffusion approach. Structural explanations of behavioral changes seek their causes in changes of tastes, preferences, and opportunities that result from either shifts in individuals' social positions (social mobility) or from reshuffling of resources associated with these positions (structural social mobility, redistribution of wealth).

Diffusion explanations or models, on the other hand, attempt to identify mechanisms conducive to cumulative adoption of behaviors by some individuals, even while their social positions or the resources associated with them remain unaltered. In diffusion models, the behavior spreads and is adopted by individuals irrespective of their socioeconomic positions, even among those whose objective positions are hypothetically associated with cost-benefit calculations that do not require or are inconsistent with the new behavior. Adopting the new behavior occurs as a result of reevaluation of one's own choices in light of other people's behavior, not as a strategic response or accommodation to a realignment of resources associated with one's social position in the social system. As in models for collective behavior (Coleman, 1990) diffusion models are built on the idea that, individuals transfer partial or total control of their own behavior to others. This implies a decision process that needs to be explicitly modeled: the decision to adopt a behavior is now affected by the actual or perceived adoption by others: the individual behavior object of study becomes endogenous (Manski, 1995).

This conceptualization of a diffusion explanation has the virtue of reintroducing explicitly the role of social interactions, their influence on individuals' calculations, and the potential for spillover effects as the adoption of a behavior may alter the very nature of the social groups and the distribution of individuals' positions within it. Below I provide two illustrations of the clash between structuralist and diffusion explanations, one from political sociology and the other from population history.

Class and Ethnic-Based Politics

Contrary to expectations, class-based political alignments do not always take hold at the same pace as industrialization progresses. Instead, traditional political allegiances, based on language or ethnic identities, could remain dominant long after industrialization creates conditions for class-based politics. This is a well-studied phenomenon in political sociology and is part of the large field of research on nationalism, ethnic enclaves, and the persistence of nonclass-based allegiances and ethnic enclaves (Hechter, 1975; Hechter et al., 2006; Brubaker, 1996). In these cases, observed individual political behavior (voting behavior, for example) is at odds with what is expected by virtue of individuals' ranking in the social system. Failure of individuals to act according to class positions – an expectation derived from a structuralist explanation of political behavior – occurs as a result of adherence to practices consistent with positions occupied prior to the social and economic transformations that accompanied industrialization. What is diffused and adopted in this case is the individual resistance to act according to class-based principles (the new behavior) and the continued allegiance to more primordial social markers.

Fertility Decline

In contrast to predictions from the classic demographic transition theory, fertility decline in Europe did not always follow a trajectory consistent with social and economic transformations that accompanied industrialization. Instead, the course of the decline roughly proceeded along ethnic, language, and religious boundaries (Coale and Watkins, 1986; National Research Council, 2001). The resulting geographic and

territorial clustering of fertility levels and patterns has been construed as evidence against a structural explanation of fertility decline, and as support for the hypothesis that fertility changes were instead associated with ideational or cultural changes and diffusion mechanisms. Clustering of fertility changes along spatial/religious/language that is inconsistent with levels of urbanization or industrialization is prima facie evidence of diffusion of a low fertility norm and/or of knowledge about family size limitation. A similar explanation has been favored to account for post-1950 fertility changes in (most) low income countries, where sharp fertility declines take place suddenly, in a highly contracted period of time, and generally in the complete absence of significant macroeconomic changes, growth, and development (Cleland and Wilson, 1987; Bongaarts and Watkins, 1996).

The examples from political sociology and population theory given above share two features. The first is that in both cases there are two competing explanations: one that infers expected behavior from a reading of individual socioeconomic positions (structuralist) and the second one that infers a pattern of expected behavior from adherence to beliefs shared by others in the same community, including those situated in different position in the social hierarchy. Second, both explanations assume that individuals are decision makers, acting in uncertain environments, sorting through limited information on prices, utilities, constraints, potential outcomes of alternative behaviors, elucidating their own preferences, and ultimately taking some course of action. Yet whereas, as a norm, structuralist accounts consistently include an explicit representation of the individual decision-making process, it is only of recent that diffusion models are built including social influences that affect individuals' decision making. These new formulations of diffusion processes have sprung in many areas, not just in sociology, and have adopted many forms. But they all share one commonality, namely, that the explanation of the outcome of interest as a result of diffusion involves interconnected actors whose decisions may reinforce each other albeit with time lags, and whose relative positions may change as a result of the process itself (Montgomery and Chung, 1994; Montgomery and Casterline, 1993; Valente, 1995, 2010, 2012; Marsden and Friedkin, 1993; Burt, 1987; Granovetter, 1973; Durlauf and Walker, 1998; Brock and Durlauf, 2001; Christakis and Fowler, 2007, 2008; Banerjee et al., 2013; Kossinets and Watts, 2006; Aral and Walker, 2012).

Weak Identification of Diffusion Processes

The bulk of evidence to discriminate between diffusion and structuralist explanations conventionally rested on aggregate data. Since in conventional diffusion models the individual adoption process is poorly, if at all, specified, the aggregate process and its outcome are ill defined: seldom is there a way to determine what kind of aggregate evidence one would expect when the individual adoption process itself is undefined. The last resort is the generalized practice of using 'residual evidence' or, equivalently, to infer the validity of the diffusion model from the weaknesses of the structural explanation, e.g., what the latter cannot account for is assumed to be the outcome of diffusion. This unsatisfactory state of affairs began to recede only recently, perhaps no longer than 20 years ago, when diffusion models were finally coupled systematically with social network theory, when individual decision making involving social interactions were explicitly defined (mostly by economists) and, finally, with the advent of agent-based models, a powerful tool with the capability to generate bounds for estimates of the influence/magnitude of diffusion processes. The nature of these advances is reviewed below.

The Diffusion of Modern Diffusion Models

About two decades or so ago, sociological analyses of diffusion processes began to move away from conventional contagion models and to reintroduce and place complexity of social interactions and individual behavior at the center of diffusion research (Palloni, 2001). The object of study was reshaped and rather than targeting aggregate parameters, such as the overall adoption rate, researchers began to focus on individual processes that led to either adoption or non-adoption of a behavior, the timing of the adoption and, strategically, the effects of the adoption process itself on the structure of the social group. This type of model shifts the focus away from aggregate outcomes toward individual behaviors and individual adoption, and redirects attention toward actors who are decision makers, to the processes of social influence that shape individual decision making, and to the changes in the landscape of prices, constraints, and preferences that the diffusion process itself induces (feedback mechanism). These new models lead to predictions about aggregate trajectories that are a transparent outcome of individuals' actions governed by the diffusion process. These models now enable the researcher to fully incorporate complexities of the adoption process itself (interagent communication, external sources, barriers, agents of change, etc.), the social conditions of interaction between actors who are adopters and potential adopters and, importantly, relevant feedback mechanisms. To evaluate the progress and improvements made in the last two decades we turn to a more rigorous definition of the elements of a diffusion process, including traditional and modern ones.

The Elements of a Diffusion Processes

The new approaches to the study of diffusion include the following elements:

Actors, Decision-Making, and Social Interactions

A diffusion process occurs among individual decision makers that evaluate costs and benefits associated with adoption (or its obverse, resistance to adoption), as well as information and ignorance about prices, costs, outcomes, and alternatives. Unlike other processes, however, individual calculations include the observed, anticipated, or presumed behaviors of other members in the group. The inclusion of the individual capacity of decision making and the intrusion of others' behavior on individual calculus is what distinguishes social from biological evolution (Cavalli-Sforza and Feldman, 1980). The diffusion process involves individual social learning, social influences associated with different individuals in the group,

and institutional constraints (Montgomery and Casterline, 1996; Erbring and Young, 1979; Marsden and Friedkin, 1993; DiMaggio and Powell, 1991).

Social Structure as Context

Adoption of new behaviors occurs within a social structure composed of formal and informal elements. Actors occupy positions in a hierarchy, perform roles, are connected formally and informally to other individuals through relations of authority, functional rapports, respect, and trust. Their positions shape the information they may receive about prices, utilities, and what others are doing, and values and norms determine the repertoire of acceptable and feasible behaviors. Emphasis on the importance of social structure for diffusion processes is hardly new (Rogers and Kincaid, 1980; Coleman et al., 1966; Burt, 1987) but it has seldom been systematically incorporated into actual empirical research on diffusion. It is only recently that sociologists interested in diffusion account for it explicitly in the formulation of models (Valente, 2010; Strang, 1991; Strang and Soule, 1997).

Individual Heterogeneity

Given preferences, prices, budget constraints, and their positions in a social structure, individuals may be more or less resistant to adopt innovations, more or less able or willing to learn from others and, if they adopt at all, more or less reluctant to retain the innovation in the menu of practices and behaviors they normally employ. That is, after one accounts for all standard elements entering in the decision to adopt or to resist, there could be individuals who are more (less) risk averse and adopt more (less) easily than others. Factors that contribute to such variation include social and economic stratification, geographic location, access to communication networks, and heterogeneity in behavioral characteristics that govern both awareness and eventual adoption (Cavalli-Sforza and Feldman, 1980).

Persistence and Reversibility

Once adopted new behaviors could be abandoned and replaced by others. This implies that diffusion models must include components to understand the persistent use of the innovation. This is of particular importance in applications to social behaviors that are inherently reversible or unstable due to changes in the risks involved (for example, participation in mass protests or feedback mechanisms that alter the organization of social groups, the position of individuals within it, or the distribution of resources and opportunities (see below)).

Endogenous Feedback

Both, the individuals' position within the group and the structure of the group, could be modified by the process of adoption itself. This implies the existence of feedbacks that may accelerate or retard the process. This is a key feature of modern approaches to diffusion processes that distinguishes them from more traditional approaches. Although the power that alternative models may have to enable inferences in the operation of feedback mechanisms may differ, they all include explicit consideration of it.

The following are three examples of feedback mechanisms in diffusion processes. In all three of them the spread of a behavior changes the decision-making environment for everybody thus influencing the diffusion process itself.

- Vaccination campaigns and the spread of diseases: the spread of antivaccination campaigns changes the collective exposure to the risk of contracting the disease and this, by itself, may reduce incentives to resist vaccination thus undermining the antivaccination campaign (Bauch and Galvani, 2013).
- Vigilance, sexual behavior, and prevalence of HIV: as the prevalence of HIV increases so does awareness of the disease and willingness to modify sexual behaviors. The spread of new behaviors (through reduction of sexual partners, steadier relations, sexual protection, control of other sexually transmitted diseases) is likely to reduce HIV prevalence and this can weaken the efficacy of collective efforts, public campaigns, awareness, and individual incentives to retain modified behaviors. The outcome may be a renewed increase in the prevalence of the disease (Bauch and Galvani, 2013).
- Adoption of technology: the adoption of some technologies (cell phones, software, and operating systems) becomes unavoidable once a critical mass of users has adopted it since the incentive structures for all users is altered, becomes more favorable for adoption of the new technology, and can even create additional opportunities that reinforce adherence to it.

The combination of these five ingredients produces considerable empirical variability. Diversity in the feedback mechanism alone (strength of endogenous feedback, time lags, threshold effects, number, and complexity of mechanisms on which it is sustained) can lead to aggregate outcomes that vary from runaway adoption and saturation, wholesale abandonment of the new behavior, or oscillations between states characterized by different levels of prevalence of the new behavior. This heterogeneity makes it difficult to infer a mechanism (diffusion vs structural mechanism, different types of diffusion) from observation of limited features of the process, such as its speed or the inflection point(s) of the growth curve of adopters. A more precise identification of the process requires more detailed modeling. This, in turn, can be done using one or a combination of the three new approaches described below.

Social Network Approaches

Although the development of social network theories, models and methods proceeded and will surely continue to proceed independently, one of the most important advances in the growth and development of diffusion models is the establishment of a bridge between the two subdisciplines. Indeed, social network theory becomes integrated as a submodule within diffusion models in a very precise way: actors' decision-making process is assumed to be influenced by their position within one or a subset of social networks in which they are members, by the frequency and type of ties that link them to other actors within those networks, by the position and degree of influence of other actors, by the types of communication between actors in the network, etc. Diffusion models borrow from social network models a conceptualization of the resources that individuals have access to and can use *by virtue of their*

membership in such networks. Valente (1995, 2010, 2012) was one of the first social network theorists to write extensively about the connection between diffusion models and the sophisticated machinery of modern social networks models (Burt, 1987; Watts, 1999; Doreian and Stockman, 1997; Carrington et al., 2005). Although the linkage had already been implemented in extant empirical work (Hedstrom, 1994; Kohler, 1996), it had rarely been formally used in a systematic way as part of diffusion models. The key innovations is that individuals' decisions to adopt a behavior is defined as a function of standard individual characteristics (gender, education, etc.), network-related network individual traits (such as individuals' position within the network) as well as pertinent network properties (type of network, size, density of interaction, etc.). The individual probability to adopt the behavior(s) is now formally represented as a function of these characteristics. Modern applications of diffusion models enhanced with the substance of social network theory and models are making inroads in many areas in social sciences ranging from health studies (Valente, 2010, 2012; Christakis and Fowler, 2007; Brockman and Helbing, 2013; Bauch and Galvani, 2013), addictive behaviors (Christakis and Fowler, 2008), communications and technology (Kossinets and Watts, 2006; Aral and Walker, 2012), and microfinance and developmental economics (Banerjee et al., 2013).

Formal Models of Social Interactions

An important shortcoming of the integrated diffusion and social network approach is that, by and large, they eschew a rigorous representation of actors' decision-making process and the feedback mechanisms that characterize the process. This weakness has been addressed by new models that seek to characterize individuals' decision-making processes as a function of individual conditions (tastes, preferences, budget constraints), the nature of individuals' social interactions (relative position in social networks, active and potential links to other actors) as well as the properties of the social groups to which the individual belongs that are affected by the adoption of the behavior. The most distinctive feature of these models is recognition and inclusion of the dynamic interplay between individuals' decision making about adoption of behaviors and the aggregate properties of the system that are influenced by those behaviors. Not only do these models acknowledge the importance of social interactions as tangible individual resources but offer room to represent the continuous reshuffling of social network connections and properties that take place as the diffusion process advances (Kohler, 1996; Brock and Durlauf, 2001; Durlauf and Walker, 1998; Durlauf and Ioannides, 2010; Lesthaeghe and Vanderhoeft, 2001).

Agent-Based Models

The formulation of social interaction models to represent diffusion processes is a significant departure from and improvement upon more traditional approaches but comes with a steep price tag: the empirical estimation of key parameters from observables can only be carried out after considerable simplifications. While these models allow a richer formulation of diffusion of behaviors, empirical estimation requires the translation of complex models into stylized and oftentimes confining representations. An undesirable outcome of this is that the mismatch between empirical findings and theoretical expectations lends itself to ambiguous interpretations: it can be due not only to the formulation of a faulty model but also to a faulty translation of the original to the estimable model.

Agent-based models provide a solution to the conundrum. While these models are a very general tool to handle complexity and were not formulated to study diffusion process, they can be hijacked to represent some of the most complicated features of diffusion phenomena. Agent-based models are designed to include as much complexity contained in the individuals' decision-making processes as the researcher considers necessary; they are ideally suited to incorporate alternative feedback mechanisms that capture the dynamic interplay between individuals' decisions and the properties of the collective or aggregate.

An agent-based model consists of a set of rules that generate individual behaviors as well as aggregate properties. These rules are translated into 'observables' via computer simulations that keep track of individual actors, their characteristics and relations to other actors in the system, and the evolution of the social contexts within which those individuals interact. Agent-based models lead to precise predictions about aggregate characteristics of a collection of actors that can be readily contrasted with observed outcomes to decide about the accuracy of the original model(s).

In social sciences the best-known ancestors of agent-based models are microsimulation models of population dynamics (Sheps and Menken, 1974; Bongaarts, 1981; Wachter et al., 1979; Palloni, 1996). So far the bulk of modern agent-based models have been deployed in epidemiology and population health (Auchincloss and Diez-Roux, 2008) but there are clear indications that their use is spreading and catching on elsewhere in sociology and demography (Centola, 2010; Macy and Willer, 2002; Billari et al., 2006; Axelrod, 1997a,b). While none of the applications to date are explicitly designed to deal with diffusion processes, the extension is not complicated and one could easily imagine how diffusion models that include the complexity implied by social interactions and network models could lead to standard agent-based models.

Future developments of diffusion models are likely to proceed via the agent-based model route, thus fulfilling the promise of a sociological tradition that foresaw the need to incorporate for complexity but could not quite consider it explicitly.

See also: Action, Theories of Social; Collective Behavior, Sociology of; Innovation, Theory of; Nationalism, Sociology of; Neighborhood Effects; Organizational Ecology; Social Movements: A Social Psychological Perspective; Social Networks.

Bibliography

Auchincloss, A.H., Diez-Roux, A.V., 2008. A new tool for epidemiology: the usefulness of dynamic agent models in understanding place effects on health. American Journal of Epidemiology 168 (1), 1–8.

Alter, N., 2001. Sociology of diffusion. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Springer Verlag, New York, pp. 3681–3684.

Anderson, R.M., May, R.M., 1992. Infectious Diseases of Humans. Oxford University Press, Oxford.
Aral, S., Walker, D., July 20, 2012. Identifying influential and susceptible members of social networks. Science 337, 337–341.
Axelrod, R.M., 1997a. Complexity and Cooperation. Princeton University Press, Princeton, NJ.
Axelrod, R.M., 1997b. Advancing the art of simulation in the social sciences. In: Conte, R., Hegelsman, R., Terna, P. (Eds.), Simulating Social Phenomena. Springer, Berlin, Germany, pp. 21–40. Charles Griffen, London.
Bailey, N.T.J., 1975. The Mathematical Theory of Infectious Diseases and Its Applications, Griffin, London.
Bandura, A., 1986. Social Foundations of Thought and Action. Prentice-Hall, Englewood Cliffs.
Banerjee, A., Arun, G., Duflo, E., Jackson, M., July 26, 2013. The diffusion of microfinance. Science 341, 363.
Bartholomew, D.J., 1982. Stochastic Models for Social Processes. Wiley, New York.
Bauch, C., Galvani, A., October 4, 2013. Social factors in epidemiology. Science 342, 4749.
Beal, G.M., Rogers, E.M., 1958. The scientist as a referent in the communication of new technology. Public Opinion Quarterly 22 (4), 555–563.
Billari, F.C., Fent, T., Prskawetz, A., Scheffran, J. (Eds.), 2006. Agent-based Computational Modelling. Physica-Verlag, Heidelberg.
Bongaarts, J., 1981. Simulation of the Family Life Cycle. Center for Policy Studies Working Series Paper No 70. The Population Council, New York.
Bongaarts, J., Watkins, S., 1996. Social interactions and contemporary fertility transitions. Population and Development Review 22 (4), 639–682.
Brock, W., Durlauf, S., 2001. Interactions-based models. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 5. North Holland, Amsterdam.
Brockman, D., Helbing, D., 13 December 2013. The hidden geometry of complex, network-driven contagion phenomena. Science 342, 1337–1342.
Brubaker, R., 1996. Nationalism Reframed: Nationhood and the National Question in the New Europe. Cambridge University Press, New York.
Burt, R.S., 1987. Social contagion and innovation: cohesion versus structural equivalence. American Journal of Sociology 92, 1287–1335.
Carrington, P.J., Scott, J., Wasserman, S. (Eds.), 2005. Models and Methods in Social Networks Analysis. Cambridge, New York.
Cavalli-Sforza, L.L., Feldman, M.W., 1980. Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press, Princeton, NJ.
Centola, D., 2010. The spread of behavior in an online social network experiment. Science 329 (5996), 1194–1197.
Christakis, N.A., Fowler, J.H., 2007. The spread of obesity in a large social network over 32 years. The New England Journal of Medicine 357, 370–379.
Christakis, N.A., Fowler, J.H., 2008. The collective dynamics of smoking in a large social network. The New England Journal of Medicine 358, 2249–2258.
Cleland, J., Wilson, C., 1987. Demand theories of the fertility transition: an iconoclastic view. Population Studies 41, 5–30.
Coale, A.J., Watkins, S.C., 1986. The Decline of Fertility in Europe. Princeton University Press, Princeton, NJ.
Coleman, J.S., 1990. Foundations of Social Theory. Harvard University Press, Cambridge, MA.
Coleman, J.S., Katz, E., Menzel, H., 1966. Medical Innovation: A Diffusion Study. Bobbs Merrill, New York.
DiMaggio, P.J., Powell, W.W., 1991. The iron cage revisited: institutional isomorphism and collective rationality in organization fields. In: DiMaggio, P.J., Powell, W.W. (Eds.), The New Institutionalism in Organizational Quantitative Analysis. University of Chicago Press, Chicago.
Doreian, P., Stockman, F. (Eds.), 1997. Evolution of Social Networks. Routledge, New York.
Durlauf, S., Ioannides, Y.M., 2010. Social interactions. Annual Review of Economics, Annual Reviews 21, 451–478.
Durlauf, S., Walker, J., 1998. Economic theory and diffusion models. In: National Research Council Workshop on Social Processes Underlying Fertility Change in Developing Countries. Washington, DC., 29–30/January.
Erbring, L., Young, A.A., 1979. Individuals and social structure: contextual effects as endogenous feedback. Sociological Methods and Research 7 (4), 396–430.
Granovetter, M., 1973. The strength of weak ties. American Journal of Sociology 78, 1360–1380.
Granovetter, M., 1978. Threshold models of collective behavior. American Journal of Sociology 83, 1420–1443.
Hagerstrand, T., 1967. Innovation Diffusion as a Spatial Process. University of Chicago Press, Chicago.
Hechter, M., 1975. Internal Colonialism: The Celtic Fringe in British National Development, 1536–1966. Routledge and Kegan Paul, London.
Hechter, M., Kuyucu, T., Sacks, A., 2006. Nationalism and direct rule. In: Delanty, G., Kumar, K. (Eds.), Handbook of Nations and Nationalism. Sage Publications, London, pp. 84–94.
Hedstrom, P., 1994. Contagious collectivities: on the spatial diffusion of Swedish trade unions 1890-1940. American Journal of Sociology 99 (5), 1157–1179.
Homer, J.B., Hirsch, G.B., 2006. System dynamics modeling for public health: background and opportunities. American Journal of Public Health 96, 505–514.
Katz, E., Lazarsfeld, P.F., 1955. Personal Influence: The Part Played by People in the Flow of Mass Communications. Free Press, New York.
Kohler, H., 1996. Learning in social networks: contraceptive choice and other technological dynamics. In: Paper Presented at the Population Association of America Meetings. New Orleans, May.
Kossinets, G., Watts, D.J., 2006. Empirical analysis of an evolving social network. Science 311 (88), 88–90.
Lesthaeghe, R., Vanderhoeft, C., 2001. Ready, willing and able: a conceptualization of transitions to new behavioral forms. In: National Research Council, Diffusion Processes and Fertility Transition. National Academy Press, Washington, DC, pp. 240–264.
Macy, M.W., Willer, R., 2002. From factors to actors: computational sociology and agent based modeling. Annual Review of Sociology 28, 143–166.
Mahajan, V., Peterson, R.A., 1985. Models for Innovation Diffusion. Sage Publications, Newbury Park, CA.
Manski, C.F., 1995. Identification Problems in the Social Sciences. Harvard University Press, Cambridge, MA.
Marsden, P., 1998. Diffusion through social networks. In: National Research Council, Workshop on Social Processes Underlying Fertility Change in Developing Countries. Washington, DC, 29–30/January.
Marsden, P., Friedkin, N., 1993. Network studies of social influence. Sociological Methods and Research 22 (1), 127–151.
Marwell, G., Oliver, P.E., 1993. The Critical Mass in Collective Action: A Micro-social Theory. Cambridge University Press, Cambridge.
Montgomery, M.R., Casterline, J.B., 1993. The diffusion of fertility control in Taiwan: evidence from pooled cross–section, time–series models. Population Studies 47 (3), 457–479.
Montgomery, M.R., Casterline, J.B., 1996. Social Learning, Social Influence, and New Models of Fertility. The Population Council, New York.
Montgomery, M.R., Chung, W.S., 1994. Social networks and the diffusion of fertility control: the Korean case. In: Presented at the Seminar on Values and Fertility Change, Sion, February. International Union for the Scientific Study of Population, Geneva.
Morris, M., 1993. Epidemiology and social networks: modeling structured diffusion. Sociological Methods and Research 22, 99–126.
Moscovici, S., 1985. Social influence and conformity. In: Lindzey, G., Aronson, E. (Eds.), Handbook of Social Psychology, third ed., vol. 2. Random House, New York, pp. 347–412.
National Research Council, 2001. Diffusion Processes and Fertility Transition. National Academy Press, Washington DC.
Palloni, A., 1996. The demography of HIV/AIDS. Population Index 62 (4), 601–652.
Palloni, A., 2001. Diffusion in sociological analysis. In: National Research Council, Diffusion Processes and Fertility Transitions. National Academy Press, Washington DC., pp. 66–114.
Palloni, A., 2003. Diffusion in population theory. In: Demeny, P., McNicoll, G. (Eds.), Encyclopedia of Population. Macmillan Reference, New York, pp. 237–240.
Rogers, E.M., 1995. The Diffusion of Innovations, fourth ed. The Free Press, New York.
Rogers, E.M., Kincaid, D.L., 1980. Communication Networks: A New Paradigm for Research. Free Press, New York.
Ryan, B., Gross, N.C., 1943. The diffusion of hybrid seed corn in two Iowa communities. Rural Sociology 8, 15–24.
Sheps, M.C., Menken, J.A. (Eds.), 1974. Mathematical Models of Conception and Birth. University of Chicago Press, Chicago.
Strang, D., 1991. Adding social structure to diffusion models: an event history framework. Sociological Methods and Research 19, 324–353.
Strang, D., Soule, S., 1997. Diffusion in Organizations and Social Movements: From Hybrid Corn to Poison Pills. Technical Report 9702.
Valente, T.W., 1995. Network Models of the Diffusion of Innovations. Hampton Press, Cresskill, NJ.
Valente, T.W., 2010. Social Networks and Health Models, Methods, and Applications. Oxford University Press, New York.
Valente, T.W., 2012. Network interventions. Science 337 (49), 4953.
Wachter, K.C., Hammel, E.A., Laslett, P., 1979. Statistical Studies of Historical Social Structure. Academic Press, New York.
Watts, D., 1999. Small Worlds: The Dynamics of Networks: Between Order and Randomness. Princeton University Press, Princeton, NJ.

Diglossia

Irene Theodoropoulou, College of Arts and Sciences, Qatar University, Doha, State of Qatar

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A.S. Kaye, volume 6, pp. 3688–3691, © 2001, Elsevier Ltd.

Abstract

The article reviews the basic literature on diglossia with special reference to its linguistic and social features. A number of diglossic cases are discussed with special emphasis on diglossia in the Greek-speaking world, the Arab-speaking world, and Switzerland. The main argument put forward is that as it currently stands the term 'diglossia' does not do justice to the peculiarities found in the various contexts of diglossia.

One of the basic features of every multilingual society, which nowadays is the rule rather than the exception all over the world, is the coexistence of various linguistic codes or varieties, which are usually associated with specific contexts of use, situations, or discussion topics. The coexistence of two linguistic varieties, each of which performs different social functions and is associated with different ideologies, came to be known as 'diglossia.' The term is inseparably linked with Charles Ferguson and his seminal 1959 article, which essentially introduced the term 'diglossia' into the sociolinguistic paradigm. Since Ferguson was the first to describe this phenomenon in a systematic way, his definition is considered to be the classic version, especially in view of the later modifications to the concept proposed by Joshua Fishman (1967).

Charles Ferguson has summarized diglossia as follows (1959: 245):

> Diglossia is a relatively stable language situation in which, in addition, to the primary dialects of the language (which may include a standard or regional standards), there is a very divergent, highly codified (often grammatically more complex) superposed variety, the vehicle of a large and respected body of written literature, either of an earlier period or in another speech community, which is learned largely by formal education and is used for most written and formal spoken purposes but is not used by any sector of the community for ordinary conversation.

In light of the above, diglossia translates into the simultaneous use of two varieties, one of high value (henceforth H) and one of low value (henceforth L). Fishman (1967) argued that diglossia could be extended to situations where forms of two genetically unrelated (or at least historically distant) languages occupy the H and L niches, such that one of the varieties, the H, is used for religious, educational, literacy, and other such prestigious domains, while the other one, L, is only employed for more informal, primarily spoken domains. The difference between the two is captured in the following quote: "H and L have disjoint functions: where H is appropriate, L is inappropriate and vice versa" (Sebba, 2011: 450). It thus becomes evident that the difference between the two is not only linguistic but also social, since the use of the two varieties is associated with different social functions, styles, and registers. It is noteworthy, however, that H and L are not separate languages, but varieties of the same language.

Diglossia has been claimed to be an Oriental phenomenon, as it was found in Arabic, Chinese, Sanskrit, as well as Greek (Psycharis, 1902: 129, 155, as cited in Mackridge, 2009: 27). Nonetheless, it is also found in other Western contexts. In light of this, the four most common examples of diglossia that have been discussed in the literature include the Greek-speaking world, where diglossia used to exist as the H variety called katharevousa (pure), and the L variety called dhimotiki (people's language) till 1976 (Frangoudaki, 1992; Mackridge, 2009), the ongoing diglossia in the Arab-speaking world, where there is al fusha (Modern Standard Arabic henceforth MSA) as the H variety and al ammiyya (Arabic dialect varying according to the country where it is spoken) as the L variety (Bassiouney, 2009; Stadlbauer, 2010), the diglossia in Switzerland with Hochdeutsch (High or Standard German) and Schwyzertüütsch (Swiss German) as L (Hogg et al., 1984; Stepkowska, 2012), and finally Haiti with French as H and Haitian Creole as L (Valdman, 1988). As will become clear below, apart from these contexts, diglossia is a vital phenomenon embedded in the general field of societal multilingualism all over the world. As such, it falls under the macrolevel of sociolinguistics since it is a phenomenon relating more to a social group rather than an individual. The attempts at a classification of the original concept and the multitude of interpretations, prove the intense interest diglossia has triggered among (socio)linguists.

Classificatory Schemes

Ferguson (1959) and Fishman (1980) have been talking about 'classical diglossia,' namely a state of affairs, where H is seen as classical and L as vernacular, and between them there may or may not be a genetic relation. Examples of the former include classical and vernacular Arabic, classical or classicized Greek (katharevousa) and vernacular Greek (dhimotiki), classical and vernacular Tamil, classical and vernacular Sinhalese, Sanskrit and Hindi, or classical Mandarin and modern Pekinese. An example of classical diglossia where H and L are not genetically related is Yiddish and Loshn koydesh (Aramaic, i.e., textual Hebrew), as discussed in Fishman (1976).

Apart from classical diglossia, where there is usually close connection between the two varieties, there is also 'extended diglossia' (Fishman, 1967), namely the coexistence of H and L

varieties, where the relationship between the two is distant. An example of extended diglossia is the use of Spanish and Guarani in Paraguay (Rubin, 1972).

The distinction between classical and extended diglossia has been labeled differently by various other scholars: narrow and broad diglossia (Scotton, 1986), use-oriented or diatypical, and user-oriented or dialectical (Britto, 1986).

Except for the distinction between classical and extended diglossia, Fishman usefully distinguishes between written/formal-spoken. Examples of such a type of diglossia include Standard German and Swiss German, standard spoken Pekinese [Putonghua] and Cantonese, Standard English and Carribean Creole (Fishman, 1980: 4). An example of diglossia, which does not fall neatly under any of the two classification schemes but which can be seen as combining features from both is Tamil. This language has three norms (a classical, i.e., Sangam or Pandit style, modern literary/formal-spoken, and educated colloquial) and a number of local dialects. Between these styles there are shadings from one style into another, i.e., it is possible to write modern Literary Tamil with an archaic lexicon but with nonarchaic grammar; it is also possible to swing in the other direction and make modern Literary Tamil more spoken, or to make educated colloquial more literary in flavor, or more nonstandardly colloquial (Schiffman, 1997: 210).

Despite these artificial distinctions between different types of diglossia, it would be fair to say that diglossia can be better seen as a continuum, in the sense that the boundaries between the H and the L are not sharp. In the literature of diglossia seen as a continuum, various other categorizations of the phenomenon have been proposed.

Partial versus Total Diglossia

The distinction between the two, made by Schiffman (1997: 212), is related to the overall number of users of either of the two varieties (H and L) or each and one of them. It has been observed that in some societies speakers know H, but others have L as a mother tongue and learn H as a second language. In this way, in some societies, all speakers exhibit diglossic behavior (i.e., use both H and L varieties in complementary distribution), while in others, only some members of the society do so. This could be illustrated either by a society where everyone speaks L, but only some are actively fluent in H, or the opposite case where everyone speaks and writes H, but some also master an L variety.

Homogeneous and Heterogeneous Diglossia

Another aspect of diglossia, which needs to be discussed is whether the L norm is in fact one variety or more than one, i.e., whether it is homogeneous or heterogeneous. That is, is there an L variety that can be used for communication throughout the linguistic culture and with all segments of the speech community, such that no one is forced to resort to the H variety (written/formal-spoken) or some other language, as a lingua franca. In Switzerland, no L variety is recognized as standard; speakers must learn to accommodate their variety to those of others, since the use of H Schriftdeutsch is not considered appropriate between Swiss citizens (Schiffman, 1997: 213). Another example of heterogeneous diglossia is the Arab world, where due to the diverse countries belonging to the Arab world spanning from Morocco to Yemen, no one L variety is recognized as standard either and some times native speakers of Arabic, e.g., a Moroccan with a Saudi, switch to English instead of MSA, in order for their communication to be facilitated.

Overall, though linguistic cultures think of diglossia as either-or, it is often a gradient cline, with one variant blending into another one. This becomes evident in, for example, Tamil, where the distribution of domains reserved for each variety or other can vary: in Tamil the political speech was once restricted to the domain of the H variety, but nowadays contains style-shifting, evident in the use of the H variety in the beginnings and closings of political speeches, while in the middle of the speeches the L variety is used, most probably for reasons having to do with the establishment of solidarity and rapport between the politician and their audience (Schiffman, 1997).

Regardless of how it is classified, diglossia is a sociolinguistic phenomenon involving its own peculiar linguistic and social features associated with each of its varieties. These features are further discussed in the following two sections.

Linguistic Features of Diglossia

In terms of the two varieties' linguistic systems, Ferguson has identified a typology of differences between H and L. First of all, in terms of their respective grammar systems, there are extensive differences in terms of grammatical complexity: H has grammatical categories not present in L and has an inflectional system of nouns and verbs which is much reduced or totally absent in L (Ferguson, 1959: 241). For example, Swiss German has three cases (nominative, dative, and accusative) in the noun and only one simple indicative tense, while Standard German (Hochdeutsch) inflects nouns, adjectives, and pronouns into four cases (nominative, genitive, dative, accusative) and has two nonperiphrastic indicative tenses in the verb. Likewise, katharevousa has four cases, while dhimotiki has three (it does not have dative). Similarly, Classical Arabic (CA) has three cases in the noun, marked by suffixes, i.e., word endings, while colloquial dialects have none. Finally, in the case of Haiti, French has gender and number in the noun, whereas the Creole has none.

Overall, Ferguson (1959: 241–242) has proposed a 'rule of diglossia,' according to which "the grammatical structure of language A is simpler than that of B if, other things being equal,

1. The morphophonemics of A is simpler, i.e., morphemes have fewer alternants, alternation is more regular.
2. There are fewer obligatory categories marked by morphemes or concord.
3. Paradigms are more symmetrical.
4. Concord and rection are stricter."

Provided one agrees with the aforementioned take on grammatical simplicity, one can argue that the L variety in Arabic, Greek, and Haitian falls under the category of a 'simplified' grammatical system, because it not only has simpler morphophonemics (e.g., in dhimotiki the marking of plural in nouns is simpler than that of katharevousa), but it also has fewer obligatory categories (e.g., the L variety of Arabic

does not use the dual, which exists and is used in the H variety) and the concord and rection are stricter (e.g., prepositions all take the same case in Haitian Creole rather than different as in the case of French), a fact that leads to greater symmetry in the paradigm of L than that of H. The only language that does not seem to align with the morphophonemic simplicity of its system is Swiss German, which despite its L status, has more complex morphophonemic structures than Standard German (Ferguson, 1959: 242). The idea that L does not always have to have grammatical simplicity was picked up by Fasold, who claimed that scholarship tends to "attribute simplicity to the grammar of L from pure intuition and not from systematic observation" (1987: 37).

In terms of vocabulary, despite the fact that the bulk of lexicon is shared between H and L, H is characterized by "technical terms and learned expressions, which have no regular L equivalents" (Ferguson, 1959: 242). More specifically, it is usually the case that between H and L we have the existence of paired items, one H and one L, which are used in different registers, i.e., levels of talk, ranging from formal (H) to informal (L). Examples of such pairs include the following ones:

Greek

H	*L*	*Translation*
idria	stamna	Urn, jug
idor	nero	Water
sarothro	skoupa	Broom
klini	krevati	Bed

Arabic

H	*L (Qatari dialect)*	*Translation*
keyfa haluki? (fem.sg.)	akhbarich?	How are you?
ayna anti? (fem. sg.)	wainich?	Where are you?
haqan?	sejj?	Really?
la a'aref	madri	I do not know

Haitian Creole (Taken from Ferguson, 1959: 243)

H	*L*	*Translation*
homme, gens	moun	Person, people
âne	bourik	Donkey
maintenant	kounyea	Now
donner	bay	Give

Such pairs are difficult to be found in Swiss German, because in that variety the phonological differences between H and L are extensive and the normal form of lexical pairing is regular cognation (e.g., *klein* in Standard German becomes *chly* 'small' in L) (Ferguson, 1959: 243).

At the level of phonology, there seems to be variation in the grade of similarity between the phonological systems of H and L in the aforementioned languages; while in Greek the systems are very close, in Arabic and Haitian they are different and in Swiss German they are very different. Ferguson has come up with two descriptive phonological rules, which account for the differences between H and L (1959: 244):

1. The sound systems of H and L share the same phonological structure of which the L phonology is the basic system and the divergent features of H phonology are either a subsystem or a parasystem. In this way, H has a more complex morphophonemics.
2. If 'pure' H items have phonemes not found in 'pure' L items, L phonemes frequently substitute for these in oral use of H and regularly replace them in borrowed elements by using the H-variety contrast in that particular item. Such a widely cited example is the use of the front rounded vowel /ü/ in French, which however is not used in Haitian Creole. Hence, when Haitian speakers want to sound educated, they would pronounce the word Luke as /lük/ in the context of the Gospel of St. Luke. Nonetheless, in the Haitian Creole, they would use the phoneme /i/ instead of the front rounded one /ü/ (1959).

Social Features of Diglossia

As a sociolinguistic phenomenon, diglossia should be understood in the context of the society in which it exists. The most important social feature is the function of diglossia.

According to Ferguson (1959: 236), H is used primarily on the following occasions or it performs the following functions:

1. sermon in places of worship (e.g., church, synagogue, or mosque),
2. speech in parliament,
3. personal letters,
4. university lectures,
5. news broadcasts,
6. news editorials, news stories, captions on pictures, and
7. poetry.

On the other hand, L is used in:

1. instructions to servants, waiters and waitresses, workforce and clerks,
2. conversation with family, friends, and colleagues,
3. TV soap operas,
4. captions on political cartoons, and
5. folk literature.

Even though this is a useful typology of social functions, the reality is somewhat more complex. For example, in the context of formal education the use of different varieties is not as rigid as implied in the typology, inasmuch as in countries like Greece (until 1976) and in the Arab-speaking countries (till nowadays) university lectures take place in H, but explanations of drills or answers to students' questions may be conducted in L. Poetry is also tricky as a context, because in countries like Greece, written poetry was written primarily in the L variety, as the use of H was considered to be artificial. On the other hand, in the Arab

world several fixed polite expressions or proverbs are in H, despite the fact that sometimes they are used by illiterates (Ferguson, 1959: 237).

Power and prestige are two other very important social factors that have a bearing on how diglossia operates within a given society, even though it has been noted that they have not been delved into (Sebba, 2011). H is usually rated more highly than L, for the simple reason that the former is used in formal contexts, associated with prestige, such as religious or educational settings, while the latter is used for more informal and hence less prestigious functions, such as everyday discussions. Such a distinction is very evident in the Arab world, where H (CA in particular) is usually associated with beauty, impeccability, elegance, and eloquence, because it is seen as stemming directly from Allah, while the various L varieties are felt to be less worthy and, hence, reserved for functional purposes only. Such state of affairs also leads to differences in the respective acquisition of the two varieties: while L is acquired first as the language spoken at home serving everyday, mundane communicative purposes, H is learned at school. As a result of this, H is highly standardized, namely it is codified in dictionaries, grammars, spelling guides, and canonical texts, to mention just a handful. Finally, diglossia in general can be seen as a stable phenomenon, persisting for centuries or even millennia (Schiffman, 1997: 207). Occasionally, L varieties gain domains and communicative functions and, hence, displace the H variety. Nonetheless, the other way round is possible as well, namely the H variety can displace L, only if H is the mother tongue of an elite (Schiffman, 1997). To illustrate the first case, namely the substitution of H by L, diglossia in the Greek-speaking world will be discussed. Then, diglossia as a stable phenomenon will be discussed on the basis of the Arab-speaking world.

Diglossia in the Greek-Speaking World

In the Greek-speaking world we have a sense of linguistic continuity throughout history. Because Ancient Greek did not split into several distinct modern languages, the modern Greeks can see themselves as the sole possessors and guardians of the diachronic Greek language from Homer to the present day. This sense of the continuity of the popular language and culture since antiquity was offered to the Greeks by the demotic movement, namely the effort on behalf of a number of Greek scholars to establish dhimotiki as the official language of the state on the grounds of its association with naturalness and unique ability to express emotions. It was this movement that made Greeks realized that dhimotiki was not a corrupt form of Hellenic language, the term used to refer to the Greek language in its diachronic presence, but it was Hellenic, and thus equivalent to katharevousa ('pure,' educated people's variety).

Diglossia is a term coined by the Greek novelist and essayist Emmanouil Roidis in 1885 and used by Yiannis Psycharis, an advocate of dhimotiki (Mackridge, 2009: 27), the L variety of Greek diglossia. In the context of Greece, this term denotes the use of two lexical and grammatically distinct varieties of Greek on behalf of the same users but in different circumstances. The distinction between the two varieties, namely katharevousa (the H variety) and dhimotiki (the low variety) is made not only on the basis of the medium (written vs oral language) or the level of educatedness of the users, but also on the basis of the level of formality of speech. In fact, it covers a continuum of linguistic registers ranging from 'pure' dhimotiki to 'extreme' katharevousa with hybrid varieties in between (Mackridge, 2009: 29).

Diglossia spans across all levels of linguistic description. Below are found some examples, which illustrate this:

Phonology: Katharevousa – 'anðras' (man) – Dhimotiki – 'andras' (man)
Morphology: Katharevousa – 'vasilisis' (GEN. queen) – Dhimotiki 'vasilisas' (GEN. queen)
Syntax: Katharevousa – 'i proanaferθisa omilitria' (the aforementioned speaker) using the aorist passive participle – Dhimotiki 'i omilitria pou molis anefera' (the speaker that I just mentioned) using a secondary pou clause.
Semantics: Katharevousa – 'katalavo' (occupy) – Dhimotiki – 'katalavo' (understand).
Vocabulary: Katharevousa – 'inos' (wine) – Dhimotiki 'krasi' (wine)

From 1880 to 1976 many Greek novelists were diglossic in their writing; for example, in his capacity as general secretary of the University of Athens, the demoticist poet (i.e., he wrote poetry in dhimotiki) Kostis Palamas drew up official documents in impeccable katharevousa. Along the same lines, Giorgos Seferis wrote in katharevousa as a diplomat and civil servant, while all of his literary writings are in dhimotiki. Until 1976, the year when katharevousa was abolished in Greece, all educated Greeks had been taught Ancient Greek at school. Today, Greek churchgoers are used to listening to archaic (koiné) Greek being chanted and recited. It is very important to highlight that, from a descriptive linguistic perspective, these two varieties have influenced each other and, hence, shape people's (hybrid) sociolinguistic choices, but sometimes for people themselves these two varieties may not have such sharp social meaning distinctions. This idea is captured nicely in Margaret Alexiou's words that "the two forms (katharevousa and dhimotiki) interpenetrate continuously, yet exclude each other consciously" (1982: 178).

The descriptive adequacy of the term diglossia to capture the gist of the linguistic complexity of Greece has been criticized in the literature for its oversimplification and its subsequent inability to do justice to the description of hybrid varieties found in (contemporary) Greece (e.g., Drettas, 1981; Alexiou, 1982; Kazazis, 1993; Holton, 2002). In addition, dhimotiki has been claimed to be "a codified and normalized form of the 'natural' language" (Frangoudaki, 1992: 360). Seen like this, dhimotiki, namely the L variety of Greek diglossia, is a written language. This runs contrary to Ferguson's point that the L variety is neither written nor standardized. Similarly, katharevousa is not only a hybrid consisting of lexical and grammatical features belonging to different historical stages of spoken Greek from Classical times to the present day but it was also used by different authors differently, i.e., in the proportions and the categories of the archaic features (Mackridge, 2009: 30).

Another way, in which Greek has been claimed not to fall under the term 'diglossia' is through the status of dhimotiki itself. More specifically, dhimotiki had been planned as a written language from the late nineteenth century onward, a state of affairs that led the challenging of the *de jure* monopolistic privilege of katharevousa to be thought of as the H variety (Mackridge, 2009). As a result of this, katharevousa alone was the official language of the state from 1911 to 1976, and dhimotiki along from then onward.

In the other country, where Greek is the official language, Cyprus, the situation with diglossia is different. There, since independence in 1960, the Greek Cypriots have fitted the Fergusonian model of diglossic community more closely than the Greeks of Greece. The reason for this is because, even though the official language of the country, along with Turkish is exactly the same as in Greece (katharevousa until 1976, dhimotiki since then) and most writing and official oral communication is carried out in standard Greek, Greek Cypriots speak among themselves informally in Cypriot Greek, which is markedly different from the standard written variety and is not normally used even in literature, except when recording or imitating folk poetry or for comic purposes (Karoulla-Vrikki, 2005: 13–14, as cited in Mackridge, 2009: 31).

In conclusion, even though Cyprus seems to be characterized by a more clear-cut diglossia, in Greece diglossia is not very transparent precisely due to the presence and use of uncodified, unstable, and intermediate forms of the language, which are to be found even nowadays in both written and oral genres of speech (e.g., in political discourse both inside and outside the Parliament, in computer-mediated communication, to mention just a handful of examples). Finally, it is worth mentioning that the existence of such mixed form of language (Greek mikti) has been already acknowledged by Ferguson himself (1959: 240). It was in the later scholarship that this idea was further developed and illustrated through concrete (socio)linguistic examples.

From the substitution of H by L and the abolition of diglossia in the Greek-speaking world, the discussion now shifts to an example of stable diglossia, which has been persistent throughout the years, namely the one still found in the Arab world.

Diglossia in the Arab-Speaking World

In the Western scholarship, Arabic is usually used to illustrate diglossia through reference to the distinction between MSA (al fusha) and the local or colloquial dialect (al ammiyya), which differs in terms of its phonetics, semantics, and vocabulary from country to country. An example of a phonetic difference found in different dialects is the use of the sound [g] instead of the standard [j] in Egyptian, Sudanese, and a number of Saudi Arabic dialects, which are spoken in areas in Saudi Arabia close to Egypt. Likewise, in terms of semantics there are differences between different dialects as well; for example, the wish 'bel afia' means something like enjoy your meal or bon appétit in the Gulf, but it is used as a swear word in Moroccan dialect. Likewise, in terms of vocabulary there are differences as well: 'how' is 'izzay' in Egyptian Arabic, but 'kayf' in MSA. 'Where' is 'fayn' in Egyptian Arabic, but 'eyen' in MSA.

If the concept of diglossia is contested in the Greek-speaking world, it is highly problematic in the Arab-speaking world. The main reason behind this is not only the wide sociolinguistic variation, i.e., the existence and use of a great number of different regional and social dialects spoken in the 22 countries, which form the Arab League, but also because the distinction between MSA and CA, the language of the Holy Qur'an, is not always clear.

It has been proposed that the distinction between MSA and CA is a western invention, in the sense that there is no Arabic term to capture it (Bassiouney, 2009: 11). Nonetheless, the two varieties are used on different occasions: while MSA is used in political and any sort of public formal speech, CA is used primarily as a religious language, i.e., as the language of recitation of the Holy Qur'an and quotation of classical texts. According to Ryding (2005: 7), both varieties are referred to as 'al loga al fusha' (literally 'literate language'). While what they share is more than what differentiates them, the main differences between the two are at the level of lexis and style rather than the level of grammar. MSA can be seen as simpler and more flexible than CA, inasmuch as it has a more relaxed word order, evident primarily in journalistic style, as well as the capacity of coining neologisms and borrowing terms or getting loan translations from other languages, primarily western ones. As a result of these tendencies, MSA's lexicon has a more modern vocabulary (primarily for technology) than CA and a style that has been influenced by other languages.

Arabic scholarship prefers to view diglossia in the Arabic world as gradable or leveled diglossia; what this means is that the phenomenon of diglossia in the Arab world is explained in terms of different levels of the spoken language. Early scholarship focusing on the analysis of cross-dialectal conversational data (Blanc, 1960: 85, cited in Bassiouney, 2009: 14) has claimed that there are five varieties of Arabic: classical, modified classical, semiliterary or elevated colloquial, koineised colloquial, and plain colloquial. Later scholarship identified other categorizations based on the degree of educatedness associated with the varieties; along those lines Meiseles (1980) has identified four varieties: literary Arabic or standard Arabic (the language used in written literature), oral literary Arabic, educated spoken Arabic, and plain vernacular. In a similar vain, Badawi's (1973, as cited in Bassiouney, 2009: 15) distinction among five levels of Arabic, after he analyzed data from Egyptian media, can be seen as more complex than the other two, inasmuch as it blends in both stylistic and social factors associated with the uses and users of the varieties. The five varieties or levels are the following ones: heritage classical (fusha: al- tura:θ), which is equivalent to CA, namely the variety used in the Holy Qur'an and CA literature, contemporary classical (fusha: al-'asr), the equivalent to MSA, colloquial of the cultured ('a:mmiyyat al-muθaqqafi:n), an oral variety used for serious discussions and university lectures, the colloquial of the basically educated ('a:mmiyyat al mutanawwiri:n), namely an oral variety used in discussion of nonintellectual topics, and finally the colloquial of the illiterates ('a:mmiyyat al-'ummiyi:n), a variety characterized

by a simplified linguistic structure and used in the mouths of bad characters in soap operas, and sometimes in comic situations.

Diglossia in Switzerland

Switzerland, with 74% of its speakers being German-speaking (Clyne, 1984: 18), is another example of diglossia mentioned by Ferguson (1959). Compared to the previous two cases, however, it has received relatively little analytical attention (Stepkowska, 2012: 206). For centuries the country has maintained a balance between Swiss German or Schwyzertüütsch or L (i.e., an umbrella term for several Alemannic dialects that developed around the largest Swiss cities without any 'standard' Swiss German, although, see Thuleen, 1991) and Standard German or H, to the point where the fusion between the two has been permanently averted (Martinet, 1986: 248). In fact, diglossia has proved useful to the German-speaking Swiss. First of all, the Swiss diglossia has counterbalanced the pressure exerted by the neighboring powerful German-speaking states by acting as a distinguishing marker of the countries to the north and south of the Rhine. In this sense, Swiss German or L can be seen as having political significance for the country. In addition, Swiss German can be also seen as a solidarity marker, which has helped to preserve the independence of the country and its distinctiveness from neighboring Germany. This distinguishing feature, the vernacular, was particularly important during the years of Nazi government in the Third Reich. It was at that time that Swiss Germans sought refuge within their own language. It was also at that time that Swiss German entered many spheres of public life, never to lose its position. Today, Swiss Germans assert their independence from Germans in the same way – through their own distinctive spoken L variety in which they take pride (Stepkowska, 2012: 202).

Furthermore, diglossia in Switzerland is further stabilized due to the country's strong constitutional protection for all official languages (German, French, Italian, and Romansh). Standard German is taught at schools and used in the media, literature, and most official situations, thereby providing Swiss Germans with access to the outside world through Standard German, written or spoken. Lastly, it may be said that Swiss German is an instrument of practical democracy. Due to its common use, it does not divide society into classes as in many other countries, but it unites people.

In the literature on Swiss German it has been suggested that each German-speaking Swiss canton may be regarded as a bilingual-diglossic speech community, where people use Swiss German and Standard German in a typically diglossic manner (Fasold, 1984: 41). In order to highlight the difference between the Alemannic dialect and Standard German, the former was referred to as Eidgenossisch/Helvetisch Landsprach, Schwyzerisch, der Helvetier Tütsch, and also since 1750 as Schweizerdeutsch or Schwyzertüütsch (Stepkowska, 2012: 203).

The main differences between Swiss German and Standard German are to be found in all levels of linguistic analysis (for a comprehensive discussion and detailed examples, see Thuleen, 1991).

Below are given indicative examples of differences at the levels of vocabulary, morphology, and phonetics/phonology (all examples are taken from Thuleen (1991)):

Vocabulary

Swiss German	*Standard German*	*English translation*
die Serviertochter	die Kellnerin	Waitress
der Spital	das Krankenhaus	Hospital
der Camion	der Lastkraftwagen	Truck
das Postbureau	das Postamt	Post office
der Kondukteur	der Schaffner	Train conductor

Morphology

The morphological system of numerals in Swiss German is more complex than in Standard German, inasmuch as there are more types in the former according to the grammatical gender, while in Standard German there is only one type used for all genders (zwei [two], drei [three]):

Case	*Masculine*	*Feminine*	*Neuter*
Common case	Zwee	Zwoo	Zwäi
	Drei	Drei	Drüü
Dative case	Zweene(n)	Zwoone(n)	Zwäine(n)
	Dreine(n)	Dreine(n)	Drüüne(n)

Such examples from Swiss German cast into doubt Ferguson's aforementioned claim about L being simpler linguistically than H.

Phonetics/Phonology

In Swiss German, the Standard German phoneme /k/ is almost always realized as [x], and is often spelled as ch. Thus, the consonant cluster /nk/ is realized as [nkx]. There is no counterpart to the velar fricative [x]; this means that the suffix -ig is not pronounced as [iç], like in Standard German, but rather as [ix] or, in some cases, as [ig].

Overall, it must be stated that, like in the case of Greek-speaking and Arab-speaking worlds, diglossia is a highly contested notion in the context of Switzerland exactly because of its at times inaccurate descriptiveness.

Conclusion

Despite the H and L varieties' clearly demarcated linguistic features and domains of use, diglossia is a gradient, variable phenomenon, which cannot easily be narrowed down to an either-or, binary system of categorization. The sociolinguistic complexities that have been discussed on the basis of diglossic examples from the Arab world, Haiti, Greece, and Switzerland have shown clearly that the concept of diglossia needs to be further theorized, in order to do justice to the peculiarities of the various instances of this sociolinguistic phenomenon.

See also: Code Switching: Linguistic; Communicative Competence: Linguistic Aspects; Language Acquisition; Language Contact; Language Policy; Language and Ethnicity; Language and Society; Linguistic Anthropology; Literacy and Orality; Multilingualism; Sociolinguistics.

Bibliography

Alexiou, M., 1982. Diglossia in Greece. In: Haas, W. (Ed.), Standard Languages. Spoken and Written, Manchester, pp. 156–192. Reprinted in: Georgakopoulou, A., Spanaki, M., (Eds.), 2001. A Reader in Greek Sociolinguistics. Peter Lang, Oxford, Bern and Berlin, pp. 89–118.

Bassiouney, R., 2009. Arabic Sociolinguistics. Topics in Diglossia, Gender, Identity, and Politics. Georgetown University Press, Washington, DC.

Blanc, H., 1960. Style variations in Arabic: a sample of interdialectal conversation. In: Ferguson, C.A. (Ed.), Contributions to Arabic Linguistics. Harvard University Press, Cambridge, MA, pp. 81–156.

Britto, F., 1986. Diglossia: A Study of the Theory with Application to Tamil. Georgetown University Press, Washington, DC.

Clyne, M.G., 1984. Language and Society in the German-speaking Countries. Cambridge University Press, Cambridge.

Drettas, G., 1981. La diglossie: un pelerinage aux sources. Bulletin de la Societe Linguistique de Paris 76, 61–98.

Fasold, R., 1984. The Sociolinguistics of Society. Blackwell, Oxford.

Fasold, R., 1987. The Sociolinguistics of Society. Blackwell, Oxford.

Ferguson, C.A., 1959. Diglossia. Word 15, 325–340. Reprinted in: Bratt Paulston, C., Tucker, G.R. (Ed.), 2003. Sociolinguistics. The Essential Readings. Blackwell, Malden, MA, pp. 345–358.

Fishman, J.A., 1967. Bilingualism with and without diglossia; diglossia with and without bilingualism. Journal of Social Issues 23, 29–38. http://dx.doi.org/10.1111/j.1540-4560.1967.tb00573.x.

Fishman, J.A., 1976. Bilingual Education: An International Sociological Perspective. Newbury House, Rowley, MA.

Fishman, J.A., 1980. Bilingualism and biculturalism as individual and as societal phenomena. Journal of Multilingual and Multicultural Development 1 (1), 3–15.

Frangoudaki, A., 1992. Diglossia and the present language situation in Greece: a sociological approach to the interpretation of diglossia and some hypotheses on today's linguistic reality. Language in Society 21 (3), 365–381.

Hogg, M., Joyce, N., Abrams, D., 1984. Diglossia in Switzerland? A social identity analysis of speaker evaluations. Journal of Language and Social Psychology 3, 185–196.

Holton, D., 2002. Modern Greek: towards a standard language or a new diglossia? In: Jones, M.C., Esch, E. (Eds.), Language Change: The Interplay of Internal, External and Extra-linguistic Factors. De Gruyter, Berlin and New York, pp. 169–179.

Karoulla-Vrikki, D., 2005. Language Planning in Cyprus: A Reflection of an Identity Conflict (Unpublished Ph.D. thesis). University of London.

Kazazis, K., 1993. Dismantling Greek Diglossia. In: Fraenkel, E., Kramer, C. (Eds.), Language Contact – Language Change. In Language Contact – Language Conflict (Balkan Studies 1). Peter Lang, New York, pp. 7–26.

Mackridge, P., 2009. Language and National Identity in Greece, 1766–1976. Oxford University Press, Oxford.

Martinet, A., 1986. The dynamics of plurilingual situations. In: Fishman, J., Tabouret-Keller, A., Clyne, M., Krishnamurti, B., Abdulaziz, M. (Eds.), The Fergusonian Impact, Sociolinguistics and the sociology of language. In Honor of Charles A. Ferguson on the Occasion of his 65th Birthday, vol. 2. De Gruyter, Berlin, pp. 246–251.

Meiseles, G., 1980. Educated spoken Arabic and the Arabic language continuum. Archivum Linguisticum 11 (2), 117–148.

Myers-Scotton, C., 1986. Diglossia and code-switching. In: Fishman, J.A. (Ed.), The Fergusonian Impact, Sociolinguistics and the Sociology of Language, vol. 2. Mouton de Gruyter, Berlin, pp. 403–415.

Psycharis, I., 1902. The Language Question in Greece: three Essays (Chiensis, Trans.). Baptist Mission Press, Calcutta.

Rubin, J., 1972. Bilingual usage in paraguay. In: Fishman, J.A. (Ed.), Advances in the Sociology of Language, Mouton, The Hague, vol. II, pp. 512–530.

Ryding, K., 2005. Reference Grammar of Modern Standard Arabic. Cambridge University Press, Cambridge, UK.

Schiffman, H.F., 1997. Diglossia as a sociolinguistic situation. In: Coulmas, F. (Ed.), The Handbook of Sociolinguistics. Basil Blackwell, London, pp. 205–216.

Sebba, M., 2011. Societal bilingualism. In: Wodak, R., Johnstone, B., Kerswill, P. (Eds.), The SAGE Handbook of Sociolinguistics. Sage, Los Angeles, London, pp. 445–459.

Stadlbauer, S., 2010. Language Ideologies in the Arabic Diglossia of Egypt. Colorado Research in Linguistics 22. University of Colorado, Boulder. http://www.colorado.edu/ling/CRIL/Volume22_Issue1/paper_STADLBAUER.pdf (accessed 21.06.13.).

Stepkowska, A., 2012. Diglossia. A critical overview of the Swiss example. Studia Linguistica Universitatis Iagellonicae Cracoviensis 129, 199–210. http://dx.doi.org/10.4467/20834624SL.12.013.0602 (accessed 01.07.13.).

Thuleen, N., 1991. An Examination of Swiss German in and Around Zürich. Website Article. http://www.nthuleen.com/papers/130paper2.html (accessed 01.07.13.).

Valdman, A., 1988. Diglossia and language conflict in Haiti. International Journal of the Sociology of Language 71, 67–80. http://dx.doi.org/10.1515/ijsl.1988.71.67 (accessed 01.07.13.).

Dilthey, Wilhelm (1833–1911)

Timothy D Mooney, School of Philosophy, University College Dublin, Dublin, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A pivotal influence on modern European philosophies of interpretation and existence, Dilthey differentiates the human sciences of understanding from the natural sciences of explanation. He explicates life as a flow of lived experience and as an individual and sociohistorical nexus of knowing, feeling, and willing. In his middle period, he sees descriptive psychology as the foundational approach for understanding lived experience, but subsequently emphasizes the hermeneutics or systematic interpretation of outer expressions of life, from politics and law to literature and architecture. His insights into the temporality, historicity, and finitude of life are developed further in his philosophy of worldviews.

Though he was known as 'the man of first volumes' and never brought his life's great project to completion, Dilthey was a pivotal influence on the development of European philosophies of interpretation and human existence in the twentieth century, and his views of historicity, lived experience, expressivity, and understanding have been no less important than the accounts of description and hermeneutics with which they are interwoven. Wilhelm Christian Ludwig Dilthey was born in Biebrich am Rhein near Wiesbaden on 19 November 1833. He completed his Gymnasium education in 1852, and studied theology and philosophy for several semesters at Heidelberg. He moved to Berlin in 1853, working in philosophy, classics, and philology under former students of Friedrich Schleiermacher (1738–1834), and in history under the university's leading historians. Following his state exams in 1856, Dilthey taught for a short time in two Gymnasiums. In 1859 he was invited to work on the publication of Schleiermacher's letters, and in 1860 won an essay competition for 'A Comparison of Schleiermacher's Hermeneutics to the Earlier Systems of Protestant Hermeneutics.' Dilthey gained his doctorate from Berlin in 1864 for a thesis on Schleiermacher's ethics. He was appointed as a professor of philosophy at Basel in 1867. Two years later, he was appointed to the chair of philosophy in Kiel, and then to the Breslau in 1871. In 1874 he married Katherine Puttmann, and they had two daughters and one son. Dilthey was appointed to the prestigious chair of philosophy in Berlin in 1882 and retired in 1905. Productive to the last, he died unexpectedly on 1 October 1911. The cause was the outbreak of an infection in Seis am Schlern in South Tyrol, where he was on a working holiday.

Dilthey achieved widespread acclaim in Germany in 1870, when the first and only part of his *Life of Schleiermacher* was published. The year 1883 saw the publication of the first part of his projected *magnum opus*, the Critique of Historical Reason. Comprising two books, this part was entitled *Introduction to the Human Sciences*. A second part made up of four and then five books was planned, but this was never completed. The next phase of his thought is to be found in *The Origin of Our Belief in the Reality of the External World and Its Justification* from 1890, and *Ideas for a Descriptive and Analytic Psychology*, from 1894. The beginning of the last phase of his intellectual development was flagged in the 1900 article 'The Rise of Hermeneutics.' In 1905 he published *Lived Experience and Poetry*, in 1906 *The Young Hegel*, and in 1907 *The Essence of Philosophy*. The year 1910 saw the appearance of *The Formation of the Historical World in the Human Sciences*, a continuation and reworking of the Critique project, and 1911 *The Types of Worldview and Their Development in Metaphysical Systems*. A further volume of *The Formation* was also planned, the notes for which are found in the posthumous *Drafts for a Critique of Historical Reason*.

The Human Sciences, Lived Experience, and Psychology

The central task of the Critique as announced in the *Introduction* was to provide an epistemological foundation for what Dilthey called the human sciences. This required the explication of the human being's capacity to know itself and the society and history that it has produced. Though well cognizant of the historical character of human existence brought out by Giambattista Vico (1668–1744) and G.W.F. Hegel (1770–1831), Dilthey regarded his Critique as complementing, above all, Kant's critical epistemology of the natural sciences. It would also incorporate the modes of aesthetic and teleological experience set out by the earlier thinker, while bringing out the dynamically living preconditions of the critical philosophy itself. Such a reconstructed critical approach was intended to transcend both Kantian formalism and the speculative excesses of Hegelian idealism. To critique historical reason is to show that there are manifold meanings in history, but no unitary significance underlying it, no trajectory toward the fables of a promised land or absolute knowledge. In this early phase, Dilthey already distinguishes between the natural and the human sciences and between explanation (*Erklärung*) and understanding (*Verstehen*).

Outer nature is the concern of the natural sciences, and they explain natural or physical phenomena in terms of efficient causes. They begin with the identification of certain classes of phenomenal objects and of their regularities and seek to enumerate their elementary constituents and the causal relations holding between them. Essential to the explanatory procedure of these mathematical and quantitative sciences are hypothetical generalizations that go beyond empirical phenomena. If a hypothesis passes a wide range of experiments and has predictive worth, it is accorded the status of a scientific

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61189-3

theory positing an atemporal law of nature. Despite the exact character of the natural sciences, Dilthey regards the truth of their theories as provisional. They have instrumental and probabilistic worth, and may be deposed in future. Such an account anticipates, in part, Karl Popper's (1902–94) falsificationist philosophy of science (Krausser, 1968, 1969).

The specifically human sciences include psychology, anthropology, sociology, history, economics, and politics and the studies of literature, poetry, music, fine art, and architecture. For Dilthey, sociohistorical humanity is their subject and object. Before man investigates his free and creative spirit, he finds in himself and in his society a sovereignty of the will, a responsibility for actions, the ability to know, and the struggle against every encroachment on freedom. In this realm lies every value and purpose in life, and the lived or 'lived through' experiences (*Erlebnissen*) making it up form interconnected and integral wholes or nexuses. The human sciences describe, narrate, and judge the manifold facets of sociohistorical experience, having as their concern the social and cultural systems developed to protect us from wild nature and aggression and to make our lives intelligible. All these sciences involve judgments and statements that are normatively correct or incorrect and not just logically true or false (hence in politics an institution is judged to be good to the extent that it fulfills its aims of reconciling individual freedom with civic spirit). In the lives of individuals, who are peculiar and once off intersections of social and cultural systems, we find not so much a multiplicity of causes operating blindly as a play of representations, feelings, and motives. The profusion of such interactions is boundless and value laden, so we should not subject them to methods modeled on the natural sciences, as in positivist sociology.

By contrast with the procedure of explanation, the human sciences seek to understand. In ordinary life, we understand before ever we know explicitly and communicably. Everyday understanding articulates lived experience, tacitly comprehending the relationships of parts to wholes and vice versa, whether of things or other people or psychic states. In what Martin Heidegger (1889–1976) was to entitle the 'hermeneutic circle' of developing understanding, we move from wholes to parts and back again, a conception itself inherited from Schleiermacher. On this basis, Dilthey contends that the human sciences can interpret life, with interpretation denoting systematic understanding. Because man as a fact prior to history and society is a fiction, we must begin to interpret ourselves through the allied disciplines of psychology and anthropology, explicating the inner conscious world of the individual – whether of the artist or the person of action – together with his or her milieu. From here we progress to history and the other human studies. Dilthey stresses, however, that the procedures and results of the natural and human sciences are not hermetically sealed. As psychophysical beings we depend on natural processes, as witnessed by our biology and sensory apparatus and by the importance of understanding laws of sound formation as well as of meaning in linguistics and music. Conversely, our purposes and actions have their repercussions on nature and the earth, which we have made into our home and dominate increasingly. Thus we can see how relative the delimitation of these sciences actually is.

Understanding can never be purely cognitive or disinterested, claims Dilthey, since knowing is internally related to feeling and willing, all three constituting the nexus of each lived experience. Knowing includes perceiving, imagining, remembering, and other forms of representing and symbolizing, and with Kant employs conceptual categories in developed human awareness. Feeling involves pleasure and pain, and pertains primarily to one's own affective state or to a situation embracing others or the world of which the state is a part. In simpler feelings neither the state of the subject nor the objectivity of the object is clearly distinguishable. Such feelings can be pains like toothaches or indeed moods without any objects. In concert with the will, feeling gives historically emergent conceptual categories their purchase on external reality. If children had not felt their actions and passions in and through their bodies, for instance, the concept of causality could never gain a living significance. Against Kantian formalism, conceptual categories must be preceded genetically by categories of life having their own experiential form and content. With higher feelings such as those evoked by music or by the selfless actions of others, a feeling and its object can be better separated. Yet feeling is always some mode of empathically apprehending other people and properties of nature that knowledge cannot fathom fully.

Willing covers biological impulses, precognitive striving, and volitional activity. At its primitive level it is present through the entire body, and therefore places outside itself what is encountered in touch. Due to the fact that external reality influences us, resists the will, and is present in feelings of pleasure and pain, it is given to us as reality rather than as bare phenomenon. Through the impulse of will and the resistance to will, maintains Dilthey, we acquire within the nexus of our representations a reflexive awareness of ourselves as distinct from others. This is at first an inchoate awareness of being there as subjects of experience, not yet of what we are. Willing in its volitional forms is related to situations that ought to be realized, and is accompanied by representations of objectives and of the means to realize them. Only a minority of our choices can be realized proximately or at all, so our lives are permeated by a struggle between representations and realities. In the earlier drafts for the second part of the Critique, the complexity of a lived experience is further elucidated by a focus on its temporality. The present is a living state shown in attention, feeling, and willing, and even in abstraction from its contents is distinct from the past and the future.

The unity of knowing, feeling, and willing is writ large in the sociohistorical reality studied by history, sociology, economics, and politics and enlivened by the arts. We are born into external organizations of society, instituted nexuses such as the family and the state with their preexisting relationships of property and dependency. The power structures of the state have their genesis in the commanding and obeying that we first encounter in the family circle, and must also appeal to feeling and reason to survive. The known and felt justification of these structures, furthermore, is essential to healthy progress. In growing up we expand into cultural systems, which are voluntary associations that we join to further our purposes through cooperation and to enhance our sense of the value and ultimate significance of life. Though the isolated individual is a fiction, contends Dilthey, each individual contributes to and

thus modifies the social systems that intersect in him or her. Impersonal and ahistorical systems are no less abstractions than atomic and ahistorical human beings, though we must avoid mystical notions such as a 'national culture' or 'spirit of a people,' which apart from their pernicious possibilities are unusable for understanding and explanation. The individual must contrast with generality to be recognized, but is ineffable. Following Schleiermacher, every individual, in principle, exceeds his or her comprehension. Dilthey maintains that the human sciences are irreducible to each other, whatever their commonality and order of precedence may be. The repeated references to sociohistorical reality reflect his claim that no discipline can collapse human life into the categories of interpretation peculiar to it.

The middle period of Dilthey's career is marked by *The Origin of Our Belief in the Reality of the External World and Its Justification* and *Ideas for a Descriptive and Analytic Psychology*. In the first work, he explicates in greater detail the contributions of feeling and willing to our sense of the world as reality rather than mere appearance. In the other, he invests much confidence in the strength of psychological method, already conceived in his earlier work as the starting point for the human sciences. To account for human and sociohistorical reality, we must eschew explanatory psychology and proceed in a purely descriptive manner. Modeled on the methodology of the natural sciences, the explanatory approach places far too much reliance on hypotheses of physiological causality and empirical association. It fails to appreciate that we already encounter interconnected wholes in early experience, and has little to say about the immanent evolution of character and perspective through the course of life, which is not of necessity a development for the better in this or that person. The fecundity of description has been well illustrated in great novels, poems, biographies, and autobiographies, which portray the higher forms of human experience, the intrinsic shape and value of individual character in different stages of life, and the variety of creative geniuses who have developed and continue to develop our sociohistorical world in its moral and aesthetic richness. A properly descriptive psychology will systematize these forms of description, allowing us to move rigorously from the outer to the inner, that is, to our inner psychic states in their underlying structural unities and in their individual singularities.

Dilthey's descriptive and structural psychology was criticized heavily by Wilhelm Windelband (1848–1915), Heinrich Rickert (1863–1936), and his former student Hermann Ebbinghaus (1850–1909). For Ebbinghaus, the descriptive methodology ignores the difficult experimental work and demonstrable findings of explanatory psychology. This problem apart, it is assumed that description can be carried out without making use of hidden hypotheses and sliding into inferences. Against the initial criticism, Dilthey's psychology was not confined to description, also being framed to work from common and intersubjectively verifiable (if not ahistorical) features of lived experience back to their structural conditions in a broadly Kantian fashion. Yet he came to concur that hypotheses and inferences were impossible to avoid. There has recently been a shift in the view of the impact of these criticisms and the scope of the novel approach taken by Dilthey in the final phase of his career. The earlier view is that the criticisms motivated him to reject the psychological approach and to formulate a thoroughgoing alternative (Bollnow, 1936). Newer scholarship suggests that he downplayed descriptive psychology without abandoning it, maintaining that it had a valid but limited range. On this reading his second methodology complemented it, and was motivated by a deepening appreciation of the temporality and historicity of life and lived experience predating these criticisms (Grondin, 1994; De Mul, 2004). Certainly he no longer regards psychology as the single foundation of the human sciences, and comes to voice objections to reflective and introspective procedures signaling the need for a further approach.

Hermeneutics, Expressions of Life, and Worldviews

Dilthey's last phase was heralded in the programmatic sections of 'The Rise of Hermeneutics.' The most detailed works of this period are *The Formation of the Historical World in the Human Sciences* and the posthumous *Drafts for a Critique of Historical Reason*. A lived experience of knowing, feeling, and willing is now construed as a unitary episode of meaning having an internal relationship to the past and future, such that the former is still efficacious and the latter already efficacious within it. The past is never restricted to what has lapsed, for it is understood as continuing to color and orient us in its being settled. We only rage against it in knowing the futility of our dreams as to how it could have been different. In our understanding of the future, we take ourselves as free and active. We grasp that we have manifold possibilities (at least prior to old age), yet our explicitly facing forward presupposes a projective attitude in which the future has some definite shape toward which we plan as well as strive. The present is the moment of living and suffering and willing and acting in the streaming nexus. Within the flow of presently lived experience all of our bonds to what we remember and anticipate pull us along backward and forward, constantly and ineliminably. Temporality is the first category of life, and fundamental for all the others.

The temporality of life in the extended sense is its overall course. Though its meaning can be reconstructed as a whole through recollection, it is limited by birth and death, and most especially by death. Our relationship to the event ending life is crucial to life's understanding and evaluation. Psychological reflection as introspection is inadequate to these longer and shorter nexuses of lived experience. Try as we might to focus on the flow of experience itself, we are subject to the law of life in which every observed moment of life is remembered and thus extracted from its flow. It is fixed by attention, as Dilthey puts it, which arrests what is essentially fluid. Nor can attention grasp the reaching beyond itself of each lived experience, since in ordinary awareness there is no separation of subject and object. Only within thinking is the experience and what is experienced is objectified. Evident is a proximity to Friedrich Nietzsche (1844–1900), with whose ideas Dilthey was familiar. As for the course of life, it emerges from millennia of sociohistorical as well as biological development. Each individual we encounter is an intersection of sociohistorical structures, who has and will actualize some of the possibilities opened up by these structures, but not others. Herein lies the individual's

historicity, though he or she is only given by way of realized possibilities. The knowledge we endeavor to gain reflectively is in any event slanted by the language we must use and the questions that we pose. Life is finally an enigma that we cannot solve or explain in an equation. Behind it our thinking cannot go; it is unable to fathom the life from which it has emerged. What is discoverable about humanity, only history can tell us.

Such is the backdrop to Dilthey's later method for interpreting or systematically understanding individual and socio-historical 'expressions of life.' It comprises the study of observable utterances and actions and larger and more enduring creations, from states, legal codes, and religions through to our no less important works of art and architecture. Again we move from the outer to the inner, this time understanding our own lives and those of others by way of expressions. The aspect of inner life to which we return with this approach, however, is not that of psychic acts. It is composed instead of the meanings, contents, or spiritual significances of these acts, through which we refer to objects. Dilthey is here influenced by Edmund Husserl (1859–1938), who refined the schema of act, content, and object in his investigations of the intentionality or directedness of awareness. Through the study of expressions, maintains Dilthey, we can apprehend the meaning of the attitudes that we commonly assume in life – projecting, valuing, disapproving, and the like – and those of other cultures and epochs. We can also apprehend the way in which creative individuals combine and rework inherited nexuses of significance. Examining individual expressions systematically enables us attain results possessing universal validity, showing the inner structure and lawfulness of instituted forms and meanings. Drama, for instance, is a peculiar nexus of material, poetic mood, motif, plot, and means of presentation.

Every expression is an objectification of life. Once molded and cast into the world, it is publicly available and relatively permanent. Dilthey thus widens the range of spiritual significance rendered concrete that Hegel called objective spirit (Hodges, 1952). The field of interpretation embraces paintings, statues, buildings, and monuments that are mute and silent (unlike novels, poems, musical scores, and speeches) and therefore dependent on dedicated literature for their explication. Interpretation culminates in the elucidation of the written records of human existence. As an art it is the basis of philology, and the science of this art is hermeneutics. The art of hermeneutics can be systematized and lead to objective and universal validities because we can return to solid and graphically inscribed expressions over and again, checking and deepening our understandings of their meaning and isolating those procedures recursively useful in earlier interpretations. This is how we reach general maxims and rules of interpretation. The latter cannot serve as substitutes for interpreters of genius, but do shorten the way for their successors. Interpretation could not begin if expressions of life were utterly unfamiliar, but would be superfluous if there were nothing strange in them to arouse our interest. Hence the expanding hermeneutical circle in which the interpreter goes from a whole to its parts and back again operates between the strange and the familiar.

Dilthey argues that expressions of life can be divided into three major classes: those of logic, of action, and of lived experiences. Logical expressions are made up of concepts and judgments and the systems and theories built out of them. Pertaining to scientific and practical works, they have meaning contents expressing and conforming to logical laws, and therefore remaining identical in various contexts and indifferent to them. Precisely as logical they can be understood technically or completely, but convey nothing of the individual meaning nexuses in the lives from which they arose. Expressions of action comprise observable movements directed toward planned outcomes. The whole tapestry of human actions is initially woven out of basic operations like the lifting of a door latch, the swinging of a hammer, or the sawing of wood, which immediately manifest the presence of certain purposes. These expressions do not carry particular references to the meaning nexuses of individual lives, and do not spring primarily from an intention to communicate to contemporaries and descendents. We understand all such expressions practically.

The third and most important class is that of expressions of lived experiences. They include facial and bodily gestures, speeches, ceremonies, and rituals and the entire field of literature, art, music, and architecture. They usually originate with intentions to communicate that go beyond practicalities, but may also have arisen with no intent to express spirit, though they nevertheless make it understandable. Some of them bring us into the realm of the highest aesthetic, moral, and religious values, where instrumental purposes fall away almost completely. In the higher understanding of such expressions, what is significant is not their truth or falsity but their truthfulness or untruthfulness, in the sense of what they make manifest or what they occlude. A great work of art can open up the whole world of individual and communal meaning within which it was created. Some of these enduring creations reveal nexuses of meaning that were not consciously present to their creators and their fellows, and that were passed over by earlier interpreters, since they embody and appeal to historically mediated feeling and will, not merely giving rise to knowledge. Thus an expression of lived experience may contain more significance than any process of introspection could ever catch sight of. It taps into and draws from levels not illuminated by consciousness, which rises like a small island out of these inaccessible depths. This is why the meaning intended by its author or creator may not be the key to the truthfulness of an expression.

Higher understanding would be impossible if its way was not prepared by its elementary predecessor. The child grows up within the order and ethos of the family that it shares with its other members, and is wholly immersed in commonalities. Before it learns to speak, remarks Dilthey, it has understood gestures, facial expressions, movements, and exclamations. It repeatedly encounters them as the same and in the same relation to what they mean and express. In earliest childhood, we also encounter squares planted with trees and rooms with peculiar arrangements of chairs. These become understood because human planning, ordering, and valuing have assigned a place to every square and to every object in a room. Thus the individual becomes oriented in the world of objective spirit, and with the acquisition of language learns to consciously articulate the world into wholes and parts and I and you and they. The higher forms of understanding are aimed at in the

elementary forms. Lived experiences gradually become more nuanced, and understanding is raised again in linguistic expressivity, which does not just report higher nexuses of meaning, but produces them. As our horizons widen in the triadic progression from lived experience to understanding the expression, the path is opened from the common to the universal.

Hermeneutic interpretation has the near aim of systematizing higher understanding. With its rules or maxims of interpretation, we can more easily transpose ourselves into the life of another, recreating and reexperiencing the nexus of individual and sociohistorial meaning involved in his or her expressions. This approach is not empathic, though it serves to underpin refined empathy and sympathy. Religious experiences are alien to secularized Europeans, for example, but on reading Luther and his contemporaries, we can transpose ourselves into his world, learning how his spiritual consciousness was fostered in a monastery and by studying St Augustine, how church councils and religious orders had spread the idea of an invisible church, and how Christianity as a force for shaping life was converging with the spirit of the age in the cities and in the artworks of Sachs and Dürer. In this manner the effectiveness of history on ourselves is widened and deepened, and we are emancipated by historical understanding. We can now grasp the final aim of hermeneutics, that of lifting us above the limitations of our time and place, freeing us and fostering an informed appreciation of human values in and of other epochs and cultures. Entering into another context is liberating for the individual interpreter or later reader because comparative understanding is gained, allowing his or her own life context to be clarified and enriched.

In his last years, Dilthey formulated a philosophy of worldviews that supplemented his hermeneutics of individual and sociohistorical expression. Evolutionary theory and the knowledge of the relativity and transience of every historical form of life, religion, and metaphysical system have undercut the claim of any philosophy to express the order of natural and human reality. From the outset, Dilthey had regarded metaphysics as unscientific, the 'queen of a shadow realm,' and he now devotes attention to what Nietzsche had regarded as the metaphysical need of human beings to posit something fixed and permanent beyond change. This leads them to develop worldviews or overarching pictures of the cosmos, each of which offers a comprehensive plan of life, an ideal model of personal and societal life, elevated norms of conduct, and a highest good. Like the external organizations and cultural systems informed by it, a worldview must appeal to knowledge, feeling, and will to take root and thrive, even if its extrapolations beyond lived experience lack scientific validity. As with life itself, a worldview goes through the stages of youth, middle, and old age, though it will not disappear completely if it speaks to our more enduring dreams and wishes.

Religious worldviews originate in experiences of an unseen reality only convincing to those who have had them personally, or whose testimony is persuasive to those open to the idea of such a reality. These experiences are gradually solidified into dogmas and then into more argued justifications that cannot substitute for the first hand experiences. It is the fate of every religious worldview to lose its original inspiration and to become entangled in irresolvable contradictions. Artistic worldviews have been nourished historically by religion, but have more recently attained a measure of autonomy. They appeal vividly to feeling and willing in their symbolization of general nexuses of meaning, and artists and poets typically embody the widest variety of creative capacities. The artistic genius is usually a free and healthy person who develops the affective life of humanity. The understanding of life as an aesthetic phenomenon is for all that unsystematic, and philosophical worldviews seek both systematicity and universality. Despite the anarchy of competing schemes, we can distinguish philosophical worldviews that are scientific and naturalistic, subjectively idealistic, and objectively idealistic. The first appeals primarily if not exclusively to theoretical and instrumental cognition, the second to the will and the evaluation of life, and the third to needs and purposes, appealing above all to feeling. But in the end, there is no teleology of history in which one worldview must thoroughly supplant another, and no one will fathom the mysteries of life or death or the course of living, in which all things flow and change.

Criticisms and Influence

Dilthey's notions of understanding, interpretation, nexus, lived experience, and worldview have all filtered into antipositivist and interpretative sociology. He has been taken as a classical sociological theorist (Bakker, 1999) and a core source of qualitative, ethnographic symbolic interactionism, and phenomenological sociology (Prus, 1996). His main influence has been in European philosophy, where many of critics have drawn extensively on his work. Husserl accuses him of epistemological and historical relativism, though he never questioned the validity of logical truths, and it has been contended that his historical relativism is carefully qualified. He implicitly differentiates statements within historically bound structures from meta-statements about them. Between as well as within epochs and cultures, temporality is intrinsic to life, human creativity asserts itself continuously, and historical consciousness frees us from one-sided views of human reality. It is conceded, however, that the universal validity of such meta-statements is only claimed, and that these statements are revisable, however warranted to date (De Mul, 2004). Husserl himself shows the marked influence of Dilthey in his emphasis on description and in his genetic phenomenologies of preconceptual sense, of the life world (Makkreel, 1982), of inscription in the growth of modern science, and of homeworlds and the problem of understanding alienworlds.

In the view of Heidegger, the very terminology of life retains essentialist and biologist connotations that obscure our free, self-transformable existence as possibility. He goes on to claim that Dilthey's stress on the knowledge of history is at the expense of our projective historicity (Bambach, 1995). The majority of ideas in his thought are nonetheless deeply indebted to Dilthey, from historicity, understanding, and interpretation to temporality, 'Being-towards-death,' and the world-disclosing powers of art and poetry. A no less critical, and for many years dominant, reading is that of Hans-Georg Gadamer (1900–2002), who sees an unresolved tension in Dilthey's writings between historical relativism and scientism. On this interpretation, relativism is so much not

the problem as a continued commitment to the natural scientific ideal of attaining universal and atemporal validities, evidenced further by the positivistic conceptions of objectivity sometimes employed by Dilthey and by his view of hermeneutics as a methodological foundation for the other human sciences. He is also taken to privilege present-day perspectives, construing the past as a text whose understanding may clarify and enrich but not overthrow them, though each 'fusion of horizons' in which the present is genuinely confronted by the past ensues in a higher understanding that overcomes both perspectives without ever ending the hermeneutic task (Gadamer, 1960). Yet Gadamer leans heavily on Dilthey's idea of effective history, and it is arguable that the latter would have little difficulty accommodating such fusions. As for the other accusations, Dilthey never gives hermeneutics the role once assigned to psychology, and retains a fallibilist view of objectivity irreconcilable with positivism. But it is fair to hold that he does not clarify the relations of his approaches to each other (Weiss, 1991; Grondin, 1994).

Dilthey's ideas have also been adapted by Emilio Betti (1890–1968), Karl-Otto Apel (1922 to present), Jürgen Habermas (1929 to present), and Eric Hirsch (1928 to present). All posit conditions of objective understanding and interpretation seek to purify hermeneutics of relativism. Though not himself a relativist, Paul Ricoeur (1913–2005) is a methodological pluralist who finds an inevitable conflict of interpretations of human existence, with no paradigm properly accounting for its rivals, as prefigured by Dilthey's insistence on the relative autonomy of each human science and primary appeal of each worldview. In Ricoeur's subsequent theory of narrative identity, one's sense of being an enduring self and dependable moral agent through significant life changes is itself constituted by the revisable story that one tells about oneself with the help of others. The theory endeavors to account for the nexus or connectedness of life we ordinarily affirm of ourselves, and reflects Dilthey's interests in reconstruction from memory and autobiography. Dilthey has influenced American analytic philosophy by way of Heidegger and Gadamer, notably in the thought of Richard Rorty (1931–2007) and Donald Davidson (1917–2003). Rorty espouses an epistemological and ontological relativism and describes philosophy as akin to poetry, whereas Davidson regards effective history and the fusion of horizons as filling out his theory of radical interpretation, which opposes what he identifies as the empiricist and relativist dogma of conceptual schemes that are in principle untranslatable. The fact that Dilthey's influence has generally been through his successors is largely responsible for his comparative lack of prominence as a philosopher, and it is only in the last few decades that his reputation has grown in the English-speaking world.

See also: Atomism and Holism: Philosophical Aspects; Hegel, Georg Wilhelm Friedrich (1770–1831); Hermeneutics; Historical Explanation, Theories of: Philosophical Aspects; Husserl, Edmund (1859–1938); Kant, Immanuel (1724–1804); Nietzsche, Friedrich (1844–1900); Objectivity: Philosophical Aspects; Phenomenology: Philosophical Aspects; Relativism: Philosophical Aspects; Scientific Explanation.

Bibliography

Bakker, J.L., 1999. Dilthey as classical sociological theorist. Quarterly Journal of Ideology 22, 43–82.

Bambach, C., 1995. Heidegger, Dilthey and the Crisis of Historicism. Cornell University Press, Ithaca, NY and London.

Bollnow, O.F., 1936. Dilthey: Eine Einführung in Seine Philosophie. Teubner, Stuttgart.

De Mul, J., 2004. The Tragedy of Finitude: Dilthey's Hermeneutics of Life. Yale University Press, New Haven, CT.

Dilthey, W., 1914–1990. Gesammelte Schriften, vols. 1–12. Teubner, Stuttgart; Vandenhoeck and Ruprecht, Göttingen; vols. 13–20. Vandenhoeck and Ruprecht.

Dilthey, W., Makkreel, R., Rodi, F. (Eds.), 1985–2010. Wilhelm Dilthey: Selected Works, vols. 1–5. Princeton University Press, Princeton, NJ and London.

Dilthey, W., Rickman, H.P. (Eds.), 1976. W. Dilthey: Selected Writings. Cambridge University Press, Cambridge, London.

Ermarth, M., 1978. Wilhelm Dilthey: The Critique of Historical Reason. University of Chicago Press, Chicago, IL.

Gadamer, H.G., 1960. Wahrheit und Methode. Mohr, Tübingen (1989, Truth and Method, second ed. Continuum, New York and London).

Grondin, J., 1994. Introduction to Philosophical Hermeneutics. Yale University Press, New Haven, CT.

Hodges, H.A., 1952. The Philosophy of Wilhelm Dilthey. Routledge and Kegan Paul, London.

Krausser, P., 1968. Kritik der endlichen Vernunft, Wilhelm Diltheys Revolution der allgemeinen Wissenschafts- und Handlungstheorie. Suhrkamp, Frankfurt am Main.

Krausser, P., 1969. Dilthey's revolution in the theory of the structure of scientific enquiry. Review of Metaphysics 22, 262–280.

Makkreel, R., 1982. Husserl, Dilthey and the relation of the life world to history. Research in Phenomenology 16, 39–53.

Makkreel, R., 1975. Dilthey: Philosopher of the Human Studies. Princeton University Press, Princeton, NJ and London.

Makkreel, R., Scanlon, J. (Eds.), 1987. Dilthey and Phenomenology. University Press of America, Washington, DC.

Plantinga, T., 1980. Historical Understanding in the Thought of Wilhelm Dilthey. University of Toronto Press, Toronto.

Prus, R., 1996. Symbolic Interaction and Ethnographic Research: Intersubjectivity and the Study of Human Lived Experience. SUNY Press, Albany, NY.

Rickman, H.P., 1979. Wilhelm Dilthey: Pioneer of the Human Studies. University of California Press, Berkeley, CA.

Rodi, F., 1969. Morphologie und Hermeneutik: Zur Methode Diltheys Aesthetik. Kohlhammer, Stuttgart.

Weiss, G., 1991. Dilthey's conception of objectivity in the human sciences: a reply to gadamer. Man and World 24, 471–486.

Dimensional Comparisons and Their Consequences for Self-Concept, Motivation, and Emotion

Jens Möller, Friederike Helm, Hanno Müller-Kalthoff, and Nicole Nagy, University of Kiel, Kiel, Germany
Herb W Marsh, Australian Catholic University, Strathfield, NSW, Australia; and King Saud University, Riyadh, Saudi Arabia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A comparison process that is distinct from but related to social and temporal comparisons is introduced – dimensional comparison. Using an internal frame of reference, dimensional comparison entails a single individual or group comparing his or her ability in a target domain with his or her ability in a standard domain (e.g., "How good am I in math compared to English?"). Dimensional comparisons usually reduce self-concept in the worse-off domain and increase self-concept in the better-off domain. This article presents the basic principles of dimensional comparison in 10 hypotheses, covering their antecedents and their consequences for self-concept, motivation, and emotion, as indicated by introspective, path-analytic, and experimental studies.

Each evaluation of a certain target is based on a comparison with a certain standard (Kahneman and Miller, 1986). Two of the most prominent comparison processes used to explain self-evaluations are social and temporal comparisons. Social comparisons use information on others as the standard (Festinger, 1954) while temporal comparisons use prior information on oneself as the standard (Albert, 1977). A third kind of comparison process, becoming more and more important in educational research (see Eccles, 2009), is dimensional comparison. Dimensional comparisons are intraindividual comparisons that use information from one or more comparison standards within oneself when evaluating a certain target. The term *dimensional comparison*, first used by Möller and Köller (2001a), closely references what Marsh (1986), in relation to his Internal/External frame of reference model (I/E model; see the chapter by Marsh et al., in this volume) has described as the result of an internal frame of reference. Namely, the negative impact of achievement in one domain on self-concept in another domain: A student performing better in a verbal than in a math domain tends to have a lower self-perception of his or her math ability than does a student with identical math ability but lower verbal ability.

Going beyond the I/E model, the dimensional comparison theory (DCT; Möller and Marsh, 2013) points to the psychological processes behind this so-called contrast effect. In DCT, dimensional comparisons are defined as taking place when people compare their ability in one domain (the target domain) with their ability in one or more, other domains (the standard domain/domains). The purpose of this article is to present the basic psychological principles of DCT in 10 hypotheses (see **Table 1**). Doing this, we follow Festinger's (1954) approach in formulating 10 hypotheses on social comparison theory (SCT) and Albert's (1977) similar approach in formulating 10 hypotheses on temporal comparison theory (TCT). While Albert derived these 10 hypotheses on TCT by conceptually translating Festinger's 10 hypotheses on SCT to the context of temporal comparisons, a simple transformation of SCT or TCT hypotheses to a dimensional comparison context was deemed inadequate: Some of these hypotheses do not fit DCT, while others have been falsified by subsequent research. However, the idea of developing hypotheses as a basis for DCT does seem profitable to us.

Table 1 Ten hypotheses on dimensional comparisons

I	People carry out dimensional comparisons that are intraindividual comparisons of their own abilities in different domains.
II	People prefer domains as comparison standards that share relative attributes with the target domain.
III	Most dimensional comparisons are upward comparisons with a better-off domain.
IV	Dimensional comparisons serve motivational needs (for self-differentiation, self-enhancement, self-evaluation, self-improvement, and self-maintenance).
V	Dimensional comparisons are triggered by external forces.
VI	Dimensional comparisons lead to contrast effects in self-evaluations (when the respective abilities are believed to be negatively correlated).
VII	Dimensional comparisons lead to assimilation effects in self-evaluations (when the respective abilities are believed to be positively correlated).
VIII	The net effect of dimensional comparisons is positive.
IX	Dimensional comparisons also influence mood and behavior.
X	Dimensional comparisons influence evaluations of and by other people and groups.

In our view, DCT should extend research on the I/E model to detect the psychological processes behind it. In doing so, it should deliver testable hypotheses to enrich future research on the psychological process of dimensional comparison, its antecedents and effects, in educational and other fields. Whereas research on dimensional comparison is well-known in self-concept research in educational psychology (and in sport psychology; see Chanal et al., 2009; Tietjens et al., 2005) it is largely ignored in other related fields, such as social psychology or organizational psychology. It is a central purpose of this article to serve as a bridge to other disciplines of psychology, where the theoretical conclusions as well as the empirical support presented in this article should prove useful for further research as well.

Therefore, we explore each of the hypotheses presented in **Table 1** individually, outlining the assumptions related to each hypothesis, followed by discussion of relevant research results and examining theoretical support.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.26092-3

Dimensional Comparisons in 10 Hypotheses

Hypothesis I

People carry out dimensional comparisons that are intraindividual comparisons of their own abilities in different domains.

Hypothesis I postulates that people compare their ability in a target domain with their ability in a standard domain (e.g., "How good am I in English compared to math?," where English is the target domain and math the standard domain). Dimensional comparison processes do impact on self-evaluations of ability, as many field studies on the I/E model in educational psychology have shown (see Goal-Oriented Behavior Among Human and Non-Human Primates). Yet, within those field studies it is often impossible to differentiate between a certain target and a certain standard domain. In most studies, students were asked for their self-concepts in both target and standard subjects simultaneously. Therefore, the negative contrast effects from math achievement on verbal self-concept and from verbal achievement on math self-concept were only interpreted post hoc as dimensional comparisons. In contrast, it is much easier to use the terminology of standard and target domains in experimental studies with a carefully designed causal-temporal course of events or in introspective studies, where participants can be asked which domain served as a target and which domain served as a standard (cf Möller and Marsh, 2013).

In one such introspective study, Möller and Husemann (2006) asked students to report spontaneously emerging dimensional comparisons in their everyday lives. Participants were told to note and describe any dimensional comparisons that came to their mind during a period of 14 days. In two studies, university students and high school students recorded an average of more than six dimensional comparisons during the 14-day period. In addition, students were asked to mark which domain served as target and which domain served as the comparison standard. Results showed that academic matters were most commonly used as target domains as well as standard domains (i.e., "We were given our school reports and I compared my grade in religion with my grade in mathematics"). Personal relationships and characteristics were also used, yet less frequently ("Although I am not that thin, I am not touchy").

Another introspective study (Möller and Weber, 2001) asked a sample of $N = 105$ adults between 65 and 92 years of age to make self-evaluations on mental flexibility and memory, physical fitness and health, and ability to cope with problems. Participants were asked to conduct dimensional comparisons of their ability in each of the three domains with their general ability level. Participants considered these dimensional comparisons to be more important than, and often preferred them to, social comparisons or temporal comparisons. Dimensional comparisons also seemed to result in more positive self-evaluations than did temporal or social comparisons, for this particular sample.

In their experimental studies, Möller and Köller (2001a) found direct evidence for the occurrence and for the effects of dimensional comparisons. In one study, participants were given randomly assigned achievement feedback on a math task (above average vs below average) and were asked to complete scales assessing their perceived mathematical and verbal competencies. Even without any direct reference to their achievement in the verbal domain, positive feedback on math achievement led to a lower verbal self-concept; negative feedback on math achievement led to a higher verbal self-concept. The authors concluded that it is sufficient to give achievement feedback for one domain only – in support of Hypothesis I; it is the individual who deduces differences in his or her own ability between two domains while carrying out dimensional comparisons.

Hypothesis II

People prefer domains as comparison standards that share relative attributes with the target domain.

Hypothesis II suggests a standard selection process, a more or less free and intentional choice of comparison standard. Following Goethals and Darley (1977), similarity on related attributes drives the standard selection process within social comparison processes. Within dimensional comparison processes, people who self-evaluate a particular target domain can choose between a variety of other domains that could potentially serve as the comparison standard. Research on the I/E model does not deal with the antecedents of dimensional comparisons, since students are normally asked for their self-concepts in two domains. Thus, conclusions as to how they carry out dimensional comparisons between their abilities in these domains are post hoc only (see Möller and Marsh, 2013). However, in the experimental design of Möller and Köller (2001a), it was shown that students also spontaneously choose the verbal domain as a comparison standard when they receive math feedback. Furthermore, the majority of the spontaneous dimensional comparisons reported in the diary studies by Möller and Husemann (2006) were indeed between domains that belonged to the same context, e.g., an academic or a social context, and therefore shared particular attributes. For example, one participant compared her relation to her parents: "Last time it still didn't work out with my father, but I got along better with my girlfriend." Such context or content similarity may elicit the choice of standard. In Möller and Husemann (2006), dimensional comparisons within the academic field (e.g., between similar aspects of academic life) represented the most frequently used single target–standard combination, at 15% of all comparisons. In accord with Hypothesis II, the comparisons shared similar metrics as a relative attribute; for example, math and verbal ability share common metrics such as grades or test scores.

Hypothesis III

Most dimensional comparisons are upward comparisons with a better-off standard domain.

Hypothesis III of DCT is in correspondence with SCT's hypothesis that "there is an unidirectional drive upward in the case of abilities" (cf Festinger, 1954, p. 124). Upward comparisons may be triggered by self-improvement motivation, and

the aspiration to derive hints from analyzing the results of better-off domains. A student, comparing his worse math grade with his better English grade, may perform an upward dimensional comparison from the perspective of his math ability. When people are free to choose the dimensional comparison direction, most of them compare (slightly) upward, as our diary studies showed (Möller and Husemann, 2006). Most dimensional comparisons in both studies were rated upward by the participants: In the majority of dimensional comparisons the comparison standard was rated as slightly better than the target. Therefore, Festinger's Hypothesis IV from SCT is directly transferrable to DCT.

This is of particular interest, since upward dimensional comparisons have negative outcomes for the target domain. As was demonstrated in experimental research by Möller and Köller (2001a), participants who receive frustrating achievement feedback in a task A, and additional, upward comparison ("You have achieved better in B than in A") decrease their self-concept in the intraindividually weaker target domain A, compared with a group who get identical achievement feedback in task A without the additional dimensional comparison information.

To conclude, there is very little research on spontaneous dimensional comparisons that analyses comparison direction, which Hypothesis III calls for. What we do know however is that negative mood states enhance the probability of upward comparisons, whereas positive mood triggers downward comparisons (Möller and Husemann, 2006).

Hypothesis IV

Dimensional comparisons serve motivational needs (for self-differentiation, self-enhancement, self-evaluation, self-improvement, and self-maintenance).

As described above, people carry out dimensional comparisons spontaneously, and therefore might be seen as motivated to do so. Hypothesis IV assumes that people are motivated to dimensionally compare a certain target with a certain standard, and lists prominent motivations. Although these are discussed in the literature on the self (Elliott and Dweck, 2005), there is not much empirical research on these motivations as elicitors of dimensional comparisons. Thus, the assumptions of Hypothesis IV are rather hypothetical, and should be investigated further in empirical studies of comparison motivation.

A motive for self-differentiation might trigger dimensional comparisons. Such a motive should be defined as the wish to possess a multifaceted self-concept (Marsh et al., 1988), with a complex structure integrating strengths as well as weaknesses, preferences as well as dislikes. Marsh (1986) argued that the idiosyncratic and multifaceted structure of self-concept derives both from the use of external frames of reference (in social comparisons) and internal frames of reference (dimensional comparisons). The motive for self-differentiation is strongly related to what Linville (1987) called self-complexity. She defined self-complexity as the number of distinct self-aspects a person believes himself or herself to possess. High self-complexity enables a person to buffer self-concept against stress-related events. Engaging in dimensional comparisons could be a means to trigger more complex self-views with profiled self-perceptions: comparisons between one's own abilities in different aspects of life lead to the perception of accentuated weaknesses and strengths, to highly faceted self-concepts with many distinct cognitive self-aspects. Self-evaluation motivation might also trigger dimensional comparisons. People want to know who they are, how they are doing; they want to evaluate their strengths and weaknesses. They can use dimensional comparison processes when they have to choose courses in school, to decide which degree to take at university, which job to apply for, or which hobby to take up (Dickhäuser et al., 2005; Nagy et al., 2006). However, dimensional comparisons are not always useful to evaluate domain-specific abilities correctly: Although contrast effects of dimensional comparisons may be useful in helping to detect the idiosyncratic aspects of one's own self, they are somewhat misleading as they often lead to an over- or underestimation of own abilities.

Self-enhancement and self-maintenance motivations may also trigger dimensional comparisons. When a student's self-worth is threatened by negative feedback for a target domain, their attention may switch to certain standard domains offering an escape from negative outcomes and the opportunity for compensatory self-enhancement (Baumeister, 1982). Furthermore, self-enhancement or self-maintenance needs dimensional upward comparisons: Following failure feedback, people tend to focus on strengths, not on weaknesses (cf Boney-McCoy et al., 1999; Dodgson and Wood, 1998). As Möller and Marsh (2013) point out, DCT goes beyond compensatory self-enhancement. Beside the consequences for self-worth, they conclude that dimensional comparison processes affect all involved self-evaluations negatively for the threatened (worse-off) target domain, positively for the unthreatened (better-off) standard domain.

Self-improvement motivations may also trigger upward dimensional comparisons. It can be very helpful to switch attention from a suffering target domain to one's own particular strengths as a comparison standard, when trying to improve in an area of weakness (although it leads to costs in self-concept in the target domain). Self-improvement motivation will usually trigger the choice of standard from a similar context to the target, since behavioral strategies may be more easily transferable from one domain to another when both belong to the same context.

As noted above, there is a lack of empirical studies on the motivational antecedents of dimensional comparisons. Most of the corresponding research deals with the I/E model, and just confronts students with math and verbal self-concept questionnaires, without motivational implications being addressed.

Hypothesis V

Dimensional comparisons are triggered by external forces.

Dimensional comparisons can also be seen from a nonmotivational perspective. People are not always completely free to compare dimensionally; some situations trigger dimensional comparisons more than others. Many such situations deal with strengths or weaknesses, with preferences or dislikes. Examples of external forces leading to dimensional

comparisons may include situational demand characteristics, such as when we have to choose which foreign language to take in school, whether we should specialize in arts or sciences, whether we want to become a policeman or a lawyer (Nagy et al., 2006). There is not only thrust to intraindividually compare one's own abilities in particular domains, there is also thrust to choose particular domains as standards since standard selection is not completely free (Biernat and Eidelman, 2007). Often, standard selection is directed by relative attributes that the target shares with the standard domain, as described in Hypothesis II.

Diener and Fujita (1997) have distinguished between free and forced choice approaches to social comparison. In the forced choice condition, individuals find that "comparisons are imposed or compelled" (p. 330) by environmental factors. The authors propose that in schools, students are imposed on to carry out social comparisons to evaluate their performance. To some degree, in the typical research design used in the I/E model literature, the comparison may also be seen as being 'forced': The researcher decides what subjects and what self-concepts to evaluate (e.g., math and verbal). Möller and Marsh (2013) state that in-school standards may be directly evoked by the environment. The choice of math as a standard domain following achievement feedback in English may also be prompted by accessibility and efficiency standards (Mussweiler et al., 2006): Math ability comparison may be prompted more or less automatically when a student fails an English exam. To date, support for Hypothesis V only derives from the SCT literature; research with empirical evidence for the DCT context is yet to be conducted.

Hypothesis VI

Dimensional comparisons lead to contrast effects in self-evaluations (when the respective abilities are believed to be negatively correlated).

In the context of the I/E model, dimensional comparisons prompt contrast effects; they lower self-concept in the worse-off domain and raise self-concept in the better-off domain, as a meta-analysis of all I/E studies has shown (Möller et al., 2009a). In a free-choice context (cf Diener and Fujita, 1997) a student can choose a subject with a worse achievement feedback to raise his or her relative self-concept in the target domain ("I did not do as well as expected in math, but at least I did better than in English"). This results in an increased target self-concept but is still a contrast effect since self-concept in the worse-off domain is reduced. Such contrast effects from academic achievement in one domain on self-concept in the other domain are at the heart of the I/E model. The negative effects of achievement in one subject on self-concept in a non-corresponding subject were originally the reason for positing the existence of an internal frame of reference (Marsh, 1986) and of dimensional comparisons (Möller and Köller, 2001a). These contrast effects are valid across different age groups, achievement indicators, self-concept measures, and countries (cf Marsh and Hau, 2004; Möller et al., 2009a), as well as heterogeneous samples like high or low ability students (Möller et al., 2009b). Contrast effects are also stronger when people believe that two abilities are uncorrelated or even negatively correlated (as in the case of verbal and math abilities; see Möller et al., 2006a).

As longitudinal studies have shown, dimensional comparisons lead to significant contrast effects from achievement in one domain to self-concept in another domain, even when former achievements and self-concepts are controlled for (Köller et al., 1999; Marsh and Köller, 2004; Marsh et al., 2001; Marsh and Yeung, 1998; Möller and Köller, 2001b; Möller et al., 2011).

Experimental evidence comes from Möller and Köller's (2001a) studies, which found that dimensional comparison information led to contrast effects (e.g., positive feedback in a numerical task weakened verbal self-concept even when no reference was made to the verbal domain, or when achievement feedback included explicit dimensional or social comparison information about achievement in both tasks). Furthermore, Helm et al. (under review) were able to show in a field experimental study that dimensional comparisons leading to contrast effects in verbal and numerical self-concepts can be influenced: Contrast effects are stronger when students focus on differences between the verbal and the mathematical domain than when they focus on similarities.

In conclusion, contrast effects, as well as their moderating factors, have found widespread empirical support in a variety of research designs (i.e., diary studies, experiments, and (longitudinal) field studies).

Hypothesis VII

Dimensional comparisons lead to assimilation effects in self-evaluations (when the respective abilities are believed to be positively correlated).

Despite the negative effects of dimensional comparisons, as discussed in regard to Hypothesis VI, positive effects may also occur, following dimensional comparisons. While contrast effects appear when the evaluation of a target departs from the standard, assimilation effects appear when the evaluation of a target converges to the standard. Assimilation and contrast are perceived as two sides of a coin – their juxtaposition is responsible for the net outcome of comparison processes. With regard to social comparison, Wheeler and Suls (2005) claim that "probably every social comparison creates both the pull of assimilation and the push of contrast" (p. 576). Hypotheses VI and VII postulate that this statement is also valid for dimensional comparisons.

In social comparison research, upward social comparisons with higher-performing students do not always lower the self-concept as a result of contrast processes; under particular conditions they may lead to self-concept enhancement through assimilation processes (when a student concludes that he has the potential to become as good as the better-off comparison standard in the future; cf Wheeler and Suls, 2005).

With regard to dimensional comparisons, the juxtaposition of assimilation and contrast effects is particularly interesting, since assimilation processes are also conceivable in the logic of the I/E model. Sometimes, when students compare their abilities from two subjects within the one cluster (e.g., either

verbal or science), positive assimilation effects from achievement in one subject (such as math) on self-concept in another subject (e.g., physics) occur (cf Chiu, 2012; Marsh et al., 2001; Möller et al., 2006b; Nagy et al., 2006; Rost et al., 2005). As Marsh et al. (this volume) point out, results are ambiguous when both target and standard come from the same (math or verbal) cluster of the academic self-concept continuum: Some cross paths, from achievement to noncorresponding self-concept, are negative, some positive, some are near zero. Ambiguity may be caused by students' beliefs as to whether two abilities are positively correlated or not: Most students believe that math and verbal ability are negatively correlated, although in reality they are positively correlated, as empirical findings show (Möller et al., 2006a). In contrast, most students believe that math and physics achievements are positively correlated (Möller et al., 2006a; Rost et al., 2005). Different ability beliefs may predict either contrast or assimilation effects.

In our view it is possible that every comparison process leads both to assimilation and to contrast at the same time. Therefore, if assimilation and contrast effects of a similar size occur, it is possible that near zero effects will emerge as the net effect of dimensional comparisons. Further research is needed to disentangle assimilation and contrast effects in dimensional comparisons.

Hypothesis VIII

The net effect of dimensional comparisons is positive.

Even with the possibility of near zero effects in mind (see the section 'Hypothesis VII'), the net effect of dimensional comparisons can be assumed to be typically positive. For example, Möller and Pohlmann (2010) reported positive net effects for social comparisons. That is, the positive effects of downward social comparisons on self-concept were stronger in size than the negative effects of upward social comparisons. Transferring this approach to DCT, Pohlmann and Möller (2009) found in experimental and in field studies that positive self-concept effects for downward dimensional comparisons were larger than negative effects for upward dimensional comparisons. Positive net effects were shown both in self-perceived competence and in satisfaction with test outcomes. These findings are congruent with the findings of Möller and Köller's (2001a) experimental studies.

The overall benefit in the use of dimensional comparisons suggests that people use dimensional comparisons as a way of regulating their self-evaluations in a self-enhancing way (cf the section 'Hypothesis IV'; Möller and Marsh, 2013). As is discussed in more detail, in relation to the following Hypothesis IX, this finding is also valid for mood states.

Hypothesis IX

Dimensional comparisons also influence mood and behavior.

As discussed in most of these hypotheses, dimensional comparisons affect self-evaluations. Only a few studies however have investigated the effects of dimensional comparisons on variables other than self-concept. Goetz et al. (2008) examined the antecedents of academic emotions and found that math performance negatively predicted enjoyment of language arts classes, while that language arts performance negatively predicted enjoyment of math classes. Pohlmann and Möller (2009), in an experimental setting, found larger effects for downward than for upward dimensional comparisons in both self-concept and satisfaction with test outcomes. In their diary studies, Möller and Husemann (2006) found an impact of dimensional comparisons on mood states. When participants rated the comparison direction as downward (target compared with a worse-off standard), mood states became worse. When participants carried out upward dimensional comparisons (target compared with a better-off standard), mood states improved (see **Table 2**). People experiencing feelings of dissatisfaction in one domain (e.g., following bad or sad news) may initiate an upward dimensional comparison process with another domain (Schwarz and Clore, 1996), resulting in enhancement of general self-worth, as outlined above.

Looking for more behavioral consequences of dimensional comparisons, Miller (2000) found slightly negative effects of English achievement on self-regulated learning strategies in math and almost no effect of math achievement on English learning strategies. With regard to course choices, research on dimensional comparisons shows that math achievement negatively impacts on subsequent engagement, course choices, and university entry in the noncorresponding verbal subject (Dickhäuser et al., 2005; Nagy et al., 2006). Such findings support the assumption that dimensional comparisons are motivated by a kind of a self-differentiation motivation, as Hypothesis IV describes.

Hypothesis X

Dimensional comparisons influence evaluations of and by other people and groups.

When teachers, parents, or classmates are asked to infer a student's math or verbal self-concept, contrary to expectations from the I/E model (Marsh, 1986), findings do not show strong negative effects from noncorresponding achievements (Dai, 2002; Marsh et al., 1984; Pohlmann et al., 2004). Inferred

Table 2 Upward and downward dimensional comparisons and their consequences for self-concept, motivation, course choice, and mood

Target: verbal achievement	*Standard: math achievement*	*Comparison direction (seen from the target)*	*Consequence for verbal self-concept, motivation, course choice*	*Consequence for math self-concept, motivation, course choice*	*Consequence for mood*
Average	Better	Upward	Diminution	Enhancement	Better
	Worse	Downward	Enhancement	Diminution	Worse

self-concepts seem highly determined by the corresponding achievements (but see Möller (2005) for negative effects of dimensional comparisons following praise and criticism). Nevertheless, it may be premature to conclude that dimensional comparisons play no role in the evaluations of other people.

For example, Dickhäuser (2005) assumed that observers by default do not possess distinctive achievement information about their observed target and thus do not engage in dimensional comparisons; nor do they fall victim to contrast effects when inferring someone else's self-concept. In an experimental study, $N = 160$ participants were informed of a fictional student's achievement in German and math. Results showed that inferred German self-concept was higher when German achievement was high and math achievement was low.

As Möller and Marsh (2013) point out, research on group perception deals with a phenomenon that may be connected to dimensional comparisons as well: When people had to rate the warmth and the competence of groups, the group that had been presented as higher on warmth (competence) was rated as lower on the other domain, just as the DCT would predict (Yzerbyt et al., 2008).

This may be seen as one of the rare examples of research on dimensional comparison outside of purely academic realms (for another example, on the negative effect of academic achievement on honesty self-perception, see Möller and Savyon, 2003).

Conclusion

We have discussed 10 hypotheses to describe a newly developed theory of dimensional comparisons. The hypotheses define dimensional comparisons, their targets, and standards; they deal with the antecedents and effects of dimensional comparison processes. There are three central purposes for introducing the DCT. The first is to complement research on the I/E model, while focusing on dimensional comparison as the psychological process behind the interesting negative effects from achievement in one subject on self-concept in another subject. Secondly, it should deliver testable hypotheses to enrich future research on DCT. Thirdly, it should make dimensional comparison research interesting for research in other psychological subdisciplines, such as social, organizational, or cognitive psychology.

See also: Internal/External Frame of Reference Model; Motivation and Actions, Psychology of; School Achievement: Motivational Determinants and Processes; Schooling: Impact on Cognitive and Motivational Development; Self-Regulated Learning: Theories, Measures, and Outcomes.

Bibliography

Albert, S., 1977. Temporal comparison theory. Psychological Review 84, 485–503. http://dx.doi.org/10.1037/0033-295X.84.6.485.

Baumeister, R.F., 1982. Self-esteem, self-presentation, and future interaction: a dilemma of reputation. Journal of Personality 50, 29–45.

Biernat, M., Eidelman, S., 2007. Standards. In: Kruglanski, A.W., Higgins, E.T. (Eds.), Social Psychology: Handbook of Basic Principles, vol. 2. The Guilford Press, New York, pp. 308–333.

Boney-McCoy, S., Gibbons, F.X., Gerrard, M., 1999. Self-esteem, compensatory self-enhancement, and the consideration of health risk. Personality and Social Psychology Bulletin 25, 954–965.

Chanal, J.P., Sarrazin, P.G., Guay, F., Boiché, J., 2009. Verbal, mathematics, and physical education self-concepts and achievements: an extension and test of the internal/external frame of reference model. Psychology of Sport and Exercise 10, 61–66.

Chiu, M.-S., 2012. The internal/external frame of reference model, big-fish-little-pond effect, and combined model for mathematics and science. Journal of Educational Psychology 104, 87–107. http://dx.doi.org/10.1037/a0025734.

Dai, D.Y., 2002. Incorporating parent perceptions: a replication and extension study of the internal-external frame of reference model of self-concept development. Journal of Adolescent Research 17, 617–645.

Dickhäuser, O., 2005. Teachers' inferences about students' self-concepts – the role of dimensional comparison. Learning and Instruction 15 (3), 225–235. http://dx.doi.org/10.1016/j.learninstruc.2005.04.004.

Dickhäuser, O., Reuter, M., Hilling, C., 2005. Coursework selection: a frame of reference approach using structural equation modelling. British Journal of Educational Psychology 75, 673–688. http://dx.doi.org/10.1348/000709905X37181.

Diener, E., Fujita, F., 1997. Social comparisons and subjective well-being. In: Buunk, B., Gibbons, F.X. (Eds.), Health, Coping, and Well-Being: Perspectives from Social Comparison Theory. Erlbaum, Mahwah, NJ, pp. 329–357.

Dodgson, P.G., Wood, J.V., 1998. Self-esteem and the cognitive accessibility of strengths and weaknesses after failure. Journal of Personality and Social Psychology 8, 653–663.

Eccles, J.S., 2009. Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. Educational Psychologist 44, 78–89. http://dx.doi.org/10.1080/00461520902832368.

Elliott, A.J., Dweck, C.S., 2005. Handbook of Competence and Motivation. Guilford, New York.

Festinger, L., 1954. A theory of social comparison processes. Human Relations 7, 117–140. http://dx.doi.org/10.1177/001872675400700202.

Goethals, G.R., Darley, J.M., 1977. Social comparison theory: an attributional approach. In: Suls, J.M., Miller, R.L. (Eds.), Social Comparison Processes: Theoretical and Empirical Perspectives. Hemisphere/Halsted, Washington DC, pp. 259–278.

Goetz, T., Frenzel, C.A., Hall, N.C., Pekrun, R., 2008. Antecedents of academic emotions: testing the internal/external frame of reference model for academic enjoyment. Contemporary Educational Psychology 33, 9–33. http://dx.doi.org/10.1016/j.cedpsych.2006.12.002.

Helm, F., Müller-Kalthoff, H., Nagy, N., Möller, J. Dimensional Comparison Effects on Academic Self-Concepts: Similarity of School Subjects. University of Kiel, Germany, under review.

Kahneman, D., Miller, D.T., 1986. Norm theory: comparing reality to its alternatives. Psychological Review 93, 136–153. http://dx.doi.org/10.1037/0033-295X.93.2.136.

Köller, O., Klemmert, H., Möller, J., Baumert, J., 1999. Eine längsschnittliche Überprüfung des Modells des Internal/External Frame of Reference. [Achievement feedback and academic self-concepts: a longitudinal test of the internal/external frame of reference model]. German Journal of Educational Psychology 13, 128–134.

Linville, P.W., 1987. Self-complexity as a cognitive buffer against stress-related illness and depression. Journal of Personality and Social Psychology 52, 663–676.

Marsh, H.W., 1986. Verbal and math self-concepts: an internal/external frame of reference model. American Educational Research Journal 23, 129–149.

Marsh, H.W., Byrne, B.M., Shavelson, R.J., 1988. A multifaceted academic self-concept: its hierarchical structure and its relation to academic achievement. Journal of Educational Psychology 80, 366–380. http://dx.doi.org/10.1037/0022-0663.80.3.366.

Marsh, H.W., Hau, K.-T., 2004. Explaining paradoxical relations between academic self-concepts and achievements: cross-cultural generalizability of the internal/external frame of reference predictions across 26 countries. Journal of Educational Psychology 96, 56–67. http://dx.doi.org/10.1037/0022-0663.96.1.56.

Marsh, H.W., Köller, O., 2004. Unification of theoretical models of academic self-concept/achievement relations: reunification of East and West German schools after the fall of the Berlin Wall. Contemporary Educational Psychology 29, 264–282. http://dx.doi.org/10.1016/S0361-476X(03)00034-1.

Marsh, H.W., Kong, C.K., Hau, K., 2001. Extension of the internal/external frame of reference model of self-concept formation: Importance of native and nonnative languages for Chinese students. Journal of Educational Psychology 93, 543–553. http://dx.doi.org/10.1037/0022-0663.93.3.543.

Marsh, H.W., Smith, I.D., Barnes, J., 1984. Multidimensional self-concepts: relations with inferred self-concepts and academic achievement. Australian Journal of Psychology 36, 367–386.

Marsh, H.W., Yeung, A.S., 1998. Longitudinal structural equation models of academic self-concept and achievement: gender differences in the development of math and English constructs. American Educational Research Journal 35, 705–738.

Miller, J.W., 2000. Exploring the source of self-regulated learning: the influence of internal and external comparisons. Journal of Instructional Psychology 26, 47–52.

Möller, J., 2005. "Paradoxical" effects of praise and criticism: social, dimensional, and temporal comparisons. British Journal of Educational Psychology 75, 275–295. http://dx.doi.org/10.1348/000709904X24744.

Möller, J., Husemann, N., 2006. Internal comparisons in everyday life. Journal of Educational Psychology 98, 342–353. http://dx.doi.org/10.1037/0022-0663.98.2.342.

Möller, J., Köller, O., 2001a. Dimensional comparisons: an experimental approach to the internal/external frame of reference model. Journal of Educational Psychology 93, 826–835. http://dx.doi.org/10.1037/0022-0663.93.4.826.

Möller, J., Köller, O., 2001b. Frame of reference effects following the announcement of exam results. Contemporary Educational Psychology 26, 277–287. http://dx.doi.org/10.1006/ceps.2000.1055.

Möller, J., Marsh, H.W., 2013. Dimensional comparison theory. Psychological Review 120, 544–560. http://dx.doi.org/10.1037/a0032459.

Möller, J., Pohlmann, B., 2010. Achievement differences and self-concept differences: stronger associations for above or below average students? British Journal of Educational Psychology 80, 435–450. http://dx.doi.org/10.1348/000709909X485234.

Möller, J., Pohlmann, B., Köller, O., Marsh, H.W., 2009a. A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. Review of Educational Research 79, 1129–1167. http://dx.doi.org/10.3102/0034654309337522.

Möller, J., Retelsdorf, J., Köller, O., Marsh, H.W., 2011. The reciprocal I/E model: an integration of models of relations between academic achievement and self-concept. American Educational Research Journal 48, 1315–1346. http://dx.doi.org/10.3102/0002831211419649.

Möller, J., Savyon, K., 2003. Not very smart thus moral: dimensional comparisons between academic self-concept and honesty. Social Psychology of Education 6, 95–106. http://dx.doi.org/10.1023/A:1023247910033.

Möller, J., Streblow, L., Pohlmann, B., 2006a. The belief in a negative interdependence of math and verbal abilities as determinant of academic self-concepts. British Journal of Educational Psychology 76, 57–70. http://dx.doi.org/10.1348/000709905X37451.

Möller, J., Streblow, L., Pohlmann, B., 2009b. Achievement and self-concept of students with learning disabilities. Social Psychology of Education 12, 113–122. http://dx.doi.org/10.1007/s11218-008-9065-z.

Möller, J., Streblow, L., Pohlmann, B., Köller, O., 2006b. An extension to the internal/external frame of reference model to two verbal and numerical domains. European Journal of Psychology of Education 21, 467–487. http://dx.doi.org/10.1007/BF03173515.

Möller, J., Weber, F., 2001. Dimensionale, soziale und temporale Vergleiche von alten Menschen: Selbstbewertungen, Virulenz und Präferenz [Dimensional, social, and temporal comparisons in old age: self-evaluation, virulence, and preference]. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie 33, 178–186.

Mussweiler, T., Rüter, K., Epstude, K., 2006. The why, who, and how of social comparison: a social-cognition perspective. In: Guimonde, S. (Ed.), Social Comparison and Social Psychology: Understanding Cognition, Intergroup Relations, and Culture. Cambridge University Press, Cambridge, UK, pp. 33–54.

Nagy, G., Trautwein, U., Baumert, J., Köller, O., Garrett, J., 2006. Gender and course selection in upper secondary education: effects of academic self-concept and intrinsic value. Educational Research and Evaluation 12, 323–345. http://dx.doi.org/10.1080/13803610600765687.

Pohlmann, B., Möller, J., 2009. On the benefits of dimensional comparisons. Journal of Educational Psychology 101, 248–258. http://dx.doi.org/10.1037/a0013151.

Pohlmann, B., Möller, J., Streblow, L., 2004. Zur Fremdeinschätzung von Schülerselbstkonzepten durch Lehrer und Mitschüler [On students' self-concepts inferred by teachers and class mates]. Zeitschrift für Pädagogische Psychologie 18, 157–169.

Rost, D.H., Dickhäuser, O., Sparfeldt, J.R., Schilling, S.R., 2005. Dimensional comparison in subject-specific academic self-concepts and achievements: a quasi-experimental approach. Learning & Instruction 15, 557–570. http://dx.doi.org/10.1016/j.learninstruc.2005.08.003.

Schwarz, N., Clore, G.L., 1996. Feelings as phenomenal experiences. In: Higgins, E.T., Kruglanski, A.W. (Eds.), Social Psychology: Handbook of Basic Principles. Guilford Press, New York, pp. 433–465.

Tietjens, M., Möller, J., Pohlmann, B., 2005. Leistung und Selbstkonzept in verschiedenen Sportarten [Achievement and self-concept in various sports]. Zeitschrift für Sportpsychologie 12, 135–143. http://dx.doi.org/10.1026/1612-5010.12.4.135.

Wheeler, L., Suls, J., 2005. Social comparison and self-evaluations of competence. In: Elliot, A.J., Dweck, C.S. (Eds.), Handbook of Competence and Motivation. Guilford, New York, pp. 566–578.

Yzerbyt, V.Y., Kervyn, N., Judd, C.M., 2008. Compensation versus halo: the unique relations between the fundamental dimensions of social judgment. Personality and Social Psychology Bulletin 34, 1110–1123. http://dx.doi.org/10.1177/0146167208318602.

Dimensionality of Tests: Methodology

John Hattie, University of Auckland, Auckland, New Zealand

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 3691–3695, © 2001, Elsevier Ltd.

Abstract

It is important that a score based on a set of items be unidimensional – that is, there is only one latent trait underlying the items. The notion of unidimensionality can be considered in traditional test theory and item response models in terms of the relations between items, and it can also be considered in terms of the cognitive processes involved in answering items. The more recent developments of Stout and colleagues has provided important advances in methods to assess unidimensionality, and the diagnostics procedures now available can greatly improve the development of unidimensional sets of items.

A critical concern underlying the claims for validity argument relates to ensuring appropriate representation of the construct to be assessed. It is critical to ensure that the construct is neither over- nor under-represented, and that the assumption of unidimensionality is not violated (Ackerman, 1987; Luecht and Miller, 1991; Messick, 1989; Tam, 1992). This unidimensionality assumption specifies that a score can only have meaning if the set of items measures only one attribute. If the measuring instrument has a number of items that measure differing attributes it is difficult to interpret a total score from this set of items, to make psychological sense when relating variables to this total score, or to make comments about individual differences. For example, if there are 10 spelling and 10 mathematics items in a test then it is not clear what the total score for any individual could mean. For example, two students scoring 50% are not comparable as it could be that one student scored correctly on the spelling and the other scored correctly on the mathematic items. Test scores only have meaning when the items measure a single attribute (for a more complete and earlier review see Hattie, 1985).

The concept of unidimensionality is a specific instance of the principle of local independence – that is, the existence of one latent trait underlying the data. The principle of local independence requires that responses to different items are uncorrelated when θ (the person's ability) is fixed, although it does not require that items be uncorrelated over groups in which θ varies. Lord and Novick (1968) gave the definition of local independence more substantive meaning by writing that:

> an individual's performance depends on a single underlying trait if, given his value on that trait, nothing further can be learned from him that can contribute to the explanation of his performance. The proposition is that the latent trait is the only important factor and, once a person's value on the trait is determined, the behavior is random, in the sense of statistical independence (p. 538).

A set of items can be said to be unidimensional when it is possible to find a vector of values $\theta = (\theta_i)$ such that the probability of correctly answering an item k is $\pi_{ik} = f_k(\theta_i)$ and local independence holds for each value of θ. This definition is not equating unidimensionality with local independence, because it can further require that it is necessary to condition only on one θ dimension and that the probabilities p_{ik} can be expressed in terms on only one θ dimension.

McDonald (1981) outlined ways in which it is possible to weaken the strong principle of local independence. The strong principle implies that not only are the partial correlations of the test items zero when the latent traits are partialled out, but also the distinct items are then mutually statistically independent and their higher joint moments are products of their univariate moments. A weaker form is to ignore moments beyond the second order and test the dimensionality of test scores by assessing whether the residual covariances are zero (see also Lord and Novick, 1968: pp. 225, 544–554). Under the assumption of multivariate normality, the weaker form of the principle implies the strong form, as well as conversely. McDonald (1979, 1981) argued that this weakening of the principle does not create any important change in anything that can be said about the latent traits, though strictly it weakens their definition.

Stout's Evaluation of Essential Unidimensionality

Stout (1987, 1990) also used this weaker form to develop his arguments for 'essential unidimensionality.' He devised a statistical index based on the fundamental principle that local independence should hold approximately when sampling from a subpopulation of examinees of approximately equal ability. He defined essential unidimensionality as: a test $(U_1,\ldots, U_N)$ of length N is said to be essentially unidimensional if there exists a latent variable θ such that for all values of $\theta, \frac{1}{N(N-1)} \sum_{1<i\neq j<N} \left|Cov(U_i, U_j|\theta)\right| \approx 0$.

That is, on average, the conditional covariances over all item pairs must be small in magnitude (for more detail regarding the theoretical developments see Junker, 1990, 1991, 1992; Nandakumar, 1994; Nandakumar and Stout, 1993; Stout, 1990). Essential unidimensionality thus replaces local independence and is based on assessing only the dominant dimensions. Stout has developed an empirical notion of unidimensionality to match his weaker form (Stouts T-index). There are many applications of Stout's procedures (Nandakumar, 1994; Nandakumar et al., 1998), and simulation studies that have supported the value of his statistic (Hattie et al., 1996). Zhang and Stout

(1999a, b) have generalized the Stout procedure via the DETECT method (Dimensionality evaluation to enumerate contributing traits). This method simultaneously assesses the latent dimensionality structure and determines the degrees of multidimensionality present in test data. This procedure has been extended by Kim et al. (Douglas et al., 1998; Stout et al., 1996) to identify that partition P of the available k items into disjoint and exhaustive subsets that best represents the underlying dimensionality structure. The best partition is detected by $\text{DETECT}_{\max}$, and thus is a measure of the amount of departure from unidimensionality. The methods have been generalized to polytomous items (Nandakumar et al., 1998).

Thus, there are many methods for assessing unidimensionality, such as nonlinear factor analysis (NOHARM), full information item factor analysis (TESTFACT), confirmatory factor analysis via structural equations programs such as LISREL, AMOS, and testing essential dimensionality with DIMTEST, and detection of simple structure multidimensionality (DETECT).

The Concept of Unidimensionality

It is critical to note that the assumption of unidimensionality underlying most measurement models does not necessarily imply that only these models can assess the most simple of tasks. The concept of unidimensionality can be considered at various levels of analysis. At the first-order level we can construe of a set of items being combined (by appropriate weighting) to form a total unidimensional score (such as a collection of addition items leading to a total addition score). At the second-order level we can also construe of a set of first-order scores being combined to form a total undimensional score (such as a collection of addition, subtraction, multiplication, and division totals being combined to form a total arithmetic score). In both cases we can conceive of a single underlying trait.

In a different context, we could imagine a single set of items all requiring a similar cognitive process (e.g., successive or simultaneous processing) and a total score being formed – and we could consider this a unidimensional total score – even if the items were across many content domains. All too often, the notion of unidimensional has been conceived in terms of a narrowly defined set of items. The key issue is that the methods outlined to assess unidimensionality do not lead to identifying what the total score represents – this requires other validity methods (such as judgmental analyses). Identifying a unidimensional set of items is not the same as naming it.

Constructing Multidimensional Data

One of the problematic issues when assessing dimensionality is that there are differing ways that the dimensions can relate to the construct of interest. Coombs (1954) defined three methods of constructing multidimensional data: the conjunctive model, whereby an excess of one ability, no matter how large, does not compensate for lower ability in other dimensions; the disjunctive model, whereby an examinee will pass an item if he/she is dominant over the item in any one dimension and will fail only if the item dominates him/her in all dimension; and the compensatory model, whereby an examinee's response to an item is a function of a weighted sum of underlying abilities. Other researchers (notably Ackerman, 1987, 1992; Ansley and Forsyth, 1985; Reckase et al., 1986; Wang, 1987, 1988) have demonstrated that with compensatory multidimensional models the univariate calibration of two-dimensional response data can be explained in terms of the interaction between the multidimensional test information and the distribution of the two-dimensional abilities, and various interpretations of the multidimensional item difficulty, multidimensional discrimination, and multidimensional item information have been suggested.

Sympson (1978) proposed a partially compensatory model where a decrease in one ability could only be offset by a large increase in the other ability (this model has often been erroneously called a noncompensatory model). Outside of a relatively narrow ability range, the probability of passing reduces to zero regardless of the value of the stronger ability (see Lord, 1984). The probability of a correct response is simply the product of probabilities for each dimension (but see Coughlan, 1974; Jannarone, 1986).

Concluding Comments

A major need is for a clear explication of the concept of dimensionality. Within the item response literature there has been much agreement about its meaning, although it has often been (falsely) assumed that only tests consisting of 'simple' attributes could lead to a unidimensional test. Certainly there are cognitive processing models, whereby it may be possible for a student to solve an item using differing cognitive strategies, thus adding another meaning to the notion of undimensional; the unidimensionality of processes or strategies. Much of the discussion about newer models of item and test analyses is oriented around such notions (Mislevy and Gitomer, 1995).

It is critical to ensure that any set of items map the construct of interest, and if this construct is inherently multidimensional than so too should there be multiple unidimensional subtests. At this test score level, we may desire 'lumpy' items (such as reading comprehension, or mathematical problem solving), and thus the items may assess multiple dimensions while the total test still aims to assess a single 'over all' dimension. As Dorans and Lawrence (1999) have demonstrated, it is possible to conceive of situations where analysis at the test score level would reveal a single common factor, while analysis at the subscore level would uncover a few distinct subscore factors. Similarly, Reckase et al., (1988) have demonstrated that sets of items can meet the unidimensionality assumption even though they require more than one proficient for a correct response.

At present, the Stout procedures seem worth pursuing, as they are based on clear theoretical developments, are easy to use, dependably provide indications of the dimensionality, are reasonably robust, and provide a reasonably clear and practical demarcation between one and many dimensions. Further, there is need for further research on developing multidimensional models that allow for differing ways in which items or test scores can be combined, and we need better diagnostics to understand the complexity of the dimensions involved in item

and test scores. There are already many useful heuristics available to interpret multidimensional data using unidimensional calibration methods (Ackerman, 1992; Carlson, 1987; Junker, 1991, 1992; Luecht and Miller, 1991; McKinley and Reckase, 1983; Reckase et al., 1986, 1988). It is certain that discussion will continue about the concept of unidimensionality, the best forms to assess whether items are essentially unidimensional, and the importance of diagnostics to aid in constructing better unidimensional assessments.

See also: Classical (Psychometric) Test Theory; Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis; Factor Analysis and Latent Structure: IRT and Rasch Models; Indicator: Methodology; Test Theory: Applied Probabilistic Measurement Structures.

Bibliography

Ackerman, T., 1987. The Robustness of LOGIST and BILOG IRT Estimation Programs to Violations of Local Independence. American College Testing Program, Iowa City, IA.

Ackerman, T.A., 1992. A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement 29, 67–91.

Ansley, R.A., Forsyth, T.N., 1985. An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. Applied Psychological Measurement 9, 37–48.

Carlson, J.E., 1987. Multidimensional Item Response Theory Estimation: A Computer Program. American College Testing Program, Iowa City, IA.

Coombs, C.H., 1954. A Theory of Data. Wiley, New York.

Coughlan, J.R. 1974. A Multidimensional Extension of the Normal Ogive, Logistic and Linear Latent Trait Model. Unpublished Master's thesis, University of Toronto.

Dorans, N.J., Lawrence, I.M., 1999. The Role of the Unit of Analysis in Dimensionality Assessment. (Educational Testing Service Report, No. 99–14), Princeton, NJ.

Douglas, J., Kim, H.R., Habing, B., Gao, F., 1998. Investigating local dependence with conditional covariance functions. Journal of Educational and Behavioral Statistics 23 (2), 129–151.

Hattie, J.A., Krakowski, K., Rogers, H.J., Swaminathan, H.H., 1996. An assessment of Stout's index of essential unidimensionality. Applied Psychological Measurement 20, 1–14.

Hattie, J.A., 1985. Methodology review: assessing unidimensionality of tests and items. Applied Psychological Measurement 9, 139–164.

Jannarone, R.J., 1986. Conjunctive item response theory kernels. Psychometrika 51, 357–363.

Junker, B.W., 1990. Progress in Characterizing Strictly Unidimensional IRT Representations. Technical Report 498. Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.

Junker, B.W., 1991. Essential independence and likelihood-based ability estimation for polytomous items. Psychometrika 56, 255–278.

Junker, B., 1992. April Ability Estimation in Unidimensional Models When More Than One Trait is Present. San Francisco, CA. In: Paper presented at the Annual Meeting of the American Educational Research Association.

Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA.

Lord, F.M., 1984. Conjunctive and disjunctive item response functions (Educational Testing Services Report, No. 150–520). ETS. Princeton, NJ.

Luecht, R.M., Miller, T.R., 1991. Unidimensional Calibrations and Interpretations of Multidimensional Tests. In: Paper presented at the Annual Meeting of the National Council on Measurement in Education. Chicago, IL.

McDonald, R.P., 1979. The structural analysis of multivariate data: a sketch of general theory. Multivariate Behavioural Psychology 14, 21–38.

McDonald, R.P., 1981. The dimensionality of tests and items. British Journal of Mathematical Statistical Psychology 34, 100–117.

McKinley, R.L., Reckase, M.D., 1983. MAXLOG: a computer program for estimating the parameters of a multidimensional extension of the two-parameter logistic model. Behavior Research Methods and Instrumentation 15, 389–390.

Messick, S., 1989. Validity. In: Linn, R.L. (Ed.), Educational Measurement, third ed. Macmillan, New York, pp. 13–103.

Mislevy, R.J., Gitomer, D.H., 1995. The role of probability-based inference in an intelligent tutoring system. User Modeling & User-Adapted Interaction 5, 253–282.

Nandakumar, R., Stout, W., 1993. Refinements of Stout's procedure for assessing latent trait unidimensionality. Journal of Educational Statistics 18, 41–68.

Nandakumar, R., Yu, Z.Y., Li, H.H., Stout, W., 1998. Assessing unidimensionality of polytomous data. Applied Psychological Measurement 22, 99–115.

Nandakumar, R., 1991. Traditional dimensionality vs. essential dimensionality. Journal of Educational Measurement 28, 99–117.

Nandakumar, R., 1994. Assessing latent trait unidimensionality of a set of items – comparison of different approaches. Journal of Educational Measurement 31, 1–18.

Reckase, M.D., Carlson, J.E., Ackerman, T.A., Spray, J.A., 1986. The Interpretation of Unidimensional IRT Parameters when Estimated from Multidimensional Data. In: Paper presented at the Psychometric Society Annual Meeting. Toronto.

Reckase, M.D., Ackerman, T.A., Carlson, J.E., 1988. Building a unidimensional test using multidimensional items. Journal of Educational Measurement 25, 193–203.

Stout, W., Habing, B., Douglas, J., Kim, H.R., 1996. Conditional covariance-based nonparametric multidimensionality assessment. Applied Psychological Measurement 20 (4), 331–354.

Stout, W.F., 1987. A nonparametric approach for assessing latent trait unidimensionality. Psychometrika 52, 589–617.

Stout, W.F., 1990. A new item response theory modelling approach and applications to unidimensionality assessment and ability estimation. Psychometrika 55, 293–325.

Tam, S.S., 1992. A Comparison of Methods for Adaptive Estimation of a Multidimensional Trait. Unpublished doctoral dissertation. Graduate School of Arts and Sciences, Columbia, NY.

Wang, M.M., 1987. Estimation of Ability Parameters from Response Data to Items that are Precalibrated with a Unidimensional Model. In: Paper presented at the Annual meeting of the American Educational Research Association. Washington, DC.

Wang, M.M., 1988. Measurement Bias in the Application of a Unidimensional Model to Multidimensional Item-response Data. In: Paper presented at the Annual meeting of the American Educational Research Association. New Orleans, LO.

Yu, F., Nandakumar, R., 2001. Poly-Detect for quantifying the degree of multidimensionality. Journal of Educational Measurement 38 (2), 99–120.

Zhang, J., Stout, W., 1999a. Conditional covariance structure of generalized compensatory multidimensional items. Psychometrika 64 (2), 129–152.

Zhang, J., Stout, W., 1999b. The theoretical DETECT index of dimensionality and its application to approximate simple structure. Psychometrika 64 (2), 213–249.

Direct-to-Consumer Personal Genetic Testing

Sandra Soo-Jin Lee, Center for Biomedical Ethics, Stanford University Medical School, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

As genotyping technologies have precipitously decreased in cost since the completion of the Human Genome Project, personal genetic testing has increasingly been marketed to consumers. Genetic predisposition testing is now available for an ever-expanding list of genetic mutations associated with diseases including breast cancer, Parkinson's and Alzheimer's diseases, as well as qualities such as ancestry, intelligence, and athleticism. These tests that are sold to consumers over the internet, outside of the health care system, are part of an emerging industry of direct-to-consumer (DTC) personal genetic testing. The DTC personal genetics industry provides consumers personal genetic information that interprets genotypic data into categories of relative risk and probability. Scientific and economic developments in the field of human genetics have produced a changing landscape of genetic products and services available to the public, which has fueled policy debates over questions of access, clinical and personal utility, and privacy. In addition, DTC personal genetic testing has been a focus of discourse on the potential for geneticization of social identity and questions over the meaningful extrapolation of individual genetic differences for group memberships. DTC personal genetic testing has challenged traditional approaches to genetic tests to a model of open access. The turn to the market creates new ethical, legal, and social questions as DTC personal genetic testing shifts responsibilities from the health care provider to the consumer, posing questions over who may interpret genetic information and for what purposes.

Introduction

Since the completion of the Human Genome Project, consumers have been able to purchase a broad palette of genetic tests that may range from the risk for male-patterned baldness to prostate cancer, – on the internet, without the supervision of health care providers. With relative ease, consumers can now mail cheek swabs of their DNA to commercial laboratories and within weeks receive their genetic test results online via password-protected accounts. This direct-to-consumer (DTC) model of genetic testing has garnered a significant amount of attention in the popular and scientific press over the past four years. Spearheading the shift toward what has been labeled as a democratization of health care decision making, DTC personal genetic testing companies have capitalized on technological advancements to make information more accessible to the general public by circumventing the traditional doctor–patient relationship. Unlike in Europe and other nations, DTC personal genetic testing companies, such as *23andMe, Inc.*, *Navigenics, Inc.*, and *Pathway Genomics Inc.*, have operated without significant federal regulation in the United States. However, concerns about the potential harm that DTC personal genetic products might pose to consumers have sparked policy debates regarding measures that are needed for guiding, what some have suggested resembles the 'wild west' of genomics. At the same time, some argue that regulation may threaten the nascent personal genomics industry and effectively eliminate personal genetic tests from the market. At the center of these debates are disagreements over the potential ethical, legal, and social implications of DTC genetic testing and the correct calculus for weighing individual freedom, potential harm, and the (clinical and nonmedical) utility of the information.

While it is without question that the DTC genetic testing industry requires some form of regulation to ensure both the quality and validity of the product provided to consumers, there is considerable debate over whether the current regulatory infrastructure is equipped for addressing genetic tests that are rapidly updated by genomic science. Since the boom of genome-wide association studies (GWAS) in the late 2000s – i.e., hypothesis-free 'screening' of markers across the whole genome to look for correlations with disease-related or other phenotypes – personal genetic testing offered by testing companies have focused on single nucleotide polymorphisms, or SNPs, that are single base pair mutations along the million of allelic bases that make up the human genome. Personal genetic test results include odds ratios for developing a particular disease, condition, or trait, which have been based on the identification of these rare sequence differences that have been correlated in studies with the phenotype in question. It is important to note that these associations are not necessarily causal relationships and that many of these studies are difficult to replicate due to the variability in study protocols, such as in the characterization of the study populations and even the phenotype in question. Nonetheless, genotyping technologies have become increasingly sophisticated and efficient, as SNP analysis is being replaced by new approaches such as whole genome sequencing by which the entire genome is sequenced.

The introduction of DTC personal genetic testing services presents a direct challenge to governance and integration into the health care system. The quickly shifting landscape presents significant challenges to the existing regulatory structure. Concerns over privacy and the confidentiality of genetic information purchased through DTC personal genetic testing companies and (in some cases) bio-banked by these companies, evoke significant social issues about how genomic information may inform individual and group identities, influence reproductive decision making, and raise fundamental questions of who is able to interpret genotyping data and how, and what should be relayed to individual consumers and relatives who may be affected by the information.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.82056-5

Regulating DTC Personal Genetics

The first wave of DTC personal genetic testing companies were launched in the mid-2000s to much fanfare, offering consumers personal genetic information related to ancestry, disease risk, carrier status, and drug responses. At a time when the nation's political forums are dominated by debate about the need to reform our health care system, companies frame their DTC products by building on the public's desire for revolutionary advances in medical care. This discourse, dovetailing with the scientific community's hopes for advancing genetic discovery, has focused on the notion of personalized medicine based on the individual genetic profile. The promise of personal genetics to provide new and effective insights into health care has been a hallmark in the predicted shift toward individualized, preventive medicine.

In the US there has been ambiguity over who should regulate personal genetic testing and how. Laws governing laboratories that may provide genetic testing services include the Clinical Improvement Amendments of 1988, known as CLIA. These regulate the validity of test results and are aimed at minimizing processing errors. Although most DTC personal genetic testing companies will state that they will use CLIA-approved laboratories to process their consumer samples, error in results may still occur. Beyond potential problems with the validity of test results, clinical utility and the interpretation of genotypic data into risk information provided to consumers continue to be largely unregulated in the United States.

A key point of tension for regulators is the challenge by DTC genetic-testing companies to the traditional flow of medical information; it questions the role of physicians and other health care providers as gatekeepers to genetic information. The direct access model is buoyed by the perceived individual right to genetic information and strong claims that direct access is a form of democratization of the genome. Personal genetic testing company *23andMe, Inc.* states on its website, "23andMe believes people have the right to access their personal genetic information. Genetic information is a fundamental element of a person's body, identity and individuality. As such, the rights that people enjoy with regard to financial, medical and other forms of personal information should apply to genetic information as well."

This new model challenges an infrastructure of oversight that is built on the assumption that unlike other types of personal information, genetic information requires gatekeeping. Critical to understanding this shift is the need to determine the nature of the genetic information at stake. *Navigenics*, *23andMe*, and others justified DTC marketing by emphasizing the recreational aspects of their products and asserting that they were not offering clinical diagnostic services. This distinction initially protected DTC companies from direct oversight by the US Food and Drug Administration (FDA), which strictly regulates diagnostic testing for disease risk. Yet, DTC genetic testing companies have toed the line by offering 'risk profiles' for a number of conditions, based on data drawn from GWAS and SNP analysis. Marketing practices that emphasize the potential health-related benefits of the results of these tests, further blur this distinction.

In the United States, the state regulatory bodies of New York and California did not accept this parsing. The New York State Department of Health was specifically concerned about the reliability of the tests and the expectations of uninformed consumers. Citing the discrepancy between the supposed recreational nature of the tests and the health-related nature of the results, the New York State Department of Health sent letters threatening fines and jail sentences to 31 companies who offered genetic testing services without physician involvement. In May 2010, the California Department of Public Health (CDPH) sent 'cease and desist' letters to 13 genetic testing companies forbidding the collection of samples from residents in the State. This was prompted in part by plans by the personal genetic testing company *Pathway Genomics, Inc.* and *Walgreens*, one of the largest pharmacies in the United States to retail personal genetic test kits over the counter in *Walgreens* drug stores across the nation. These plans were abandoned after the FDA demanded a review of the *Pathway Genomics* product and the US Congress convened special hearings to discuss possible industry regulation. Currently, over 20 states limit or prohibit direct-access genetic testing without physician involvement, each variably defining 'direct access' and measures of enforcement. However, unlike other national restrictions on DTC personal genetic testing (such as in Germany, the Netherlands and other European countries) the US has refrained from regulating this industry with national regulations.

Clinical Utility and Health Care Integration

A central focus of debate over the emerging DTC personal genetic testing industry is the value of the genetic information sold and the concept of 'risk' and its interpretation by different stakeholders (the company, the consumer, the physician, etc.). The quality of information provided by companies on how risk is determined varies and relies on classification assignments and assessments of harm that are often left opaque. In addition, these determinations are not equally applicable across the range of phenotypes for which DTC genetic tests are offered. For example, while a false positive or a false negative from a diagnostic test may cause undue harm to a patient, there is no evidence that the same would happen from erroneous results from a DTC genetic test. While the marketing practices of some of DTC personal genetic testing providers may have blurred the lines between diagnostic and recreational information, the risk profiles from SNP analyses derived from the GWAS are not explicitly diagnostic because the results of these studies report on low-frequency alleles in complex diseases. Instead, the genetic profiles generated by these companies represent a general snapshot of future risk for specific conditions rather than definitive indications of whether a customer will or will not develop a disease. As such, most of the test results offered in the DTC personal genetics marketplace consist of information that has minimal clinical impact. There is some worry that there are exceptions, such as tests for mutations associated with Parkinson's and Alzheimer's diseases, which present the possibility of psychological (undue stress or fatalism) or other harm to the consumer, from learning test results.

However, this concept of 'harm,' relating to inexpert action and psychosocial stress in the layperson, has not been empirically demonstrated. One study tested 162 adults with family history of Alzheimer's disease for the apolipoprotein E (APOE)

genotype associated with a higher risk for developing the disease, and compared rates of distress upon learning positive results (Green et al., 2009). The findings of the study showed only a small difference between the participants who received positive disclosure of the APOE genotype and control groups, in respect of levels of anxiety, distress, or depression even one year after the receipt of test results. In addition, a collaborative study by *Navigenics, Inc.* and the Scripps Translational Science Institute that recruited over 3000 adult participants, aged 18–85 years, demonstrated little test-related distress or significant association between risk-assessment scores and changes in behavior such as exercise, fat intake, or total number of screening tests performed. The initial data does not support the notion that consumers are at significant risk of harming themselves if they have access to this information; however, these studies have limitations and more studies of consumer behavior and decision making are needed to examine consumer behavior relating to complex genetic information.

The issue of benefit is also an under-researched area of focus. No individual DTC test has gone through a human clinical trial to demonstrate categorical safety and effectiveness of products from this industry. Given the current regulatory hurdles, it is unclear how such clinical trials would take place as these require significant time and cost. Clinical utility would be very difficult to measure, since the genomic risk assessments are based on an incomplete picture of human genetic variation that provides very basic and partial information about future risk instead of definitive evidence about current diagnosis or prognosis. It is unclear what standards of proof would be required to determine clinical validity. The traditional approach of the 'physician-as-gatekeeper' is challenged by DTC personal genomics. While most medical school curricula now incorporate clinical genetics courses, lack of genetic knowledge by practicing physicians is a systemic problem. Providing risk-assessment information to a doctor who is not qualified or sufficiently informed to explain the results of the assessment to his patient, will not reduce the risk of iatrogenic harm to the lay consumer. Considerations of clinical utility are equally murky. Policies will rely on social study of paradigms of 'harm,' 'risk,' and 'utility,' as they are operationalized by various stakeholders.

Although few national regulations have been enacted to guide the appropriate integration of DTC personal genetic testing with the health care system, several professional organizations have issued policy recommendations that emphasize the need for greater governmental oversight of these tests. The European Society of Human Genetics (ESHG) in its report, *Statement of the ESHG on Direct-to-Consumer Genetic Testing for Health-Related Purposes* recommends that genetic counseling be provided to consumers and that marketing claims be reviewed for accurate information on the test purpose and predictive value. The Human Genetics Commission in the United Kingdom emphasized in 2010, the importance of obtaining truly informed consent by consumers that includes the disclosure of company practices, test benefits, and the limitations of their approaches in providing useful information. In the US, the American College of Medical Genetics, the American Medical Association, and the American Society of Human Genetics all recommended that a qualified health professional supervise genetic testing.

Genetic Ancestry Testing

Personal genetic testing is not limited to tests related to disease or health conditions, but includes what some have referred to as 'recreational genetics.' A significant and growing market for DTC personal genetic testing is genetic ancestry testing. Genealogists interested in reconstructing their family lineage have turned to genetic ancestry testing as an important source of information for tracing their past. This process can focus on the Y-chromosome or the mitochondrial DNA, or mtDNA, inherited through the paternal and maternal lines, respectively. Testing can also probe autosomal DNA to provide results that suggest composition of DNA associated with continental, national, and/or ethnic origin. Numerous companies have proliferated over the past several decades to provide a diverse array of test results to consumers that range in specificity. For example, http://ancestry.com/ might provide customers with results indicating 40% British, 20% Scandinavian, 20% Middle Eastern, and 20% Chinese 'ethnic ancestry' whereas *African Ancestry, Inc.* will provide certificates of ancestry that suggest lineage to a present-day country or ethnic group in Africa. Others will provide 'haplogroups' that are determined by tracing the origin groups associated with specific SNPs. Definitions over what constitutes genetic ancestry can be confusing and be left undefined by companies. An individual's 'geographic ancestry' or 'biogeographic ancestry' can be taken to mean the sum of all the geographic locations inhabited by an individual's biological ancestors. However, genetic data may reveal just a small subset of the full range of ancestors that a consumer may be interested in identifying. For example, knowing a person's Y chromosomal lineage is at best a partial view of an individual's ancestry. There are also concerns over the potential incongruence over what identity may mean. In some cases the individual's or group's self-identification differs from their biogeographic ancestry, depending on a range of historical, cultural, and sociopolitical factors. It is important to recognize that biogeographic information sold to consumers and the cultural ancestry of any individual or group may not be consistent and may present more questions than answers.

A key component of genetic ancestry testing is the quality of the databases used by companies as reference data with which to compare their individual consumer's DNA sample. Although companies acknowledge that mtDNA and Y-chromosome tests provide no information about most of a test-taker's ancestors, more important limitations to genetic ancestry tests are often less obvious. For example, genetic ancestry testing can identify some of the groups and locations around the world where a test-taker's haplotype or autosomal markers are found, but it is unlikely to identify all of them. The underlying patterns of human genetic diversity determine how well ancestry inference could potentially perform. Accordingly, the accuracy of ancestry inference greatly depends on the reference database of populations available. Commercial scientists and private groups often have their own unpublished databases with the potential to provide more refined information than that available from publicly available resources. Yet, even the best databases reflect a woefully incomplete sampling of human genetic diversity, and this has important consequences for ancestry inference.

Research in human genetics has highlighted that there is more genetic variation *within* than between human groups, where those groups are defined in terms of linguistic, geographic, and cultural boundaries. Patterns of variation, however, are far from random. Human population history, including major migrations from one continent to another as well as more short-range movements, has led to correlation between genetic variation and geographic distribution. This finding is particularly true of indigenous peoples; populations characterized by a high degree of interaction with neighboring groups adhere less to these patterns. Inferences of genetic ancestry depend on the samples in a company's database, and even databases with 10 000–20 000 samples may fail to capture the full array of human genetic diversity in a particular population or region. Little is known about the sampling strategies and history of most of these databases. Furthermore, companies are not required to disclose the representation of samples within their databases. As a result, it is unclear whether a DNA sample is identified with a particular African tribe, for example, because segments of Africa are not included in company databases and thus, are effectively excluded from analysis.

An important concept that may be confused in the marketing of genetic ancestry testing is that genetic ancestry is the not the same as 'race.' Human genetic variation within continents is, for the most part, geographically continuous and clinal, particularly in regions of the world that have not received many immigrants in recent centuries. Genetic data cannot reveal an individual's full geographic ancestry precisely, although emerging research has been used to identify geographic ancestry at the continental and subcontinental levels. Genetic clusters, however, are far from being equivalent to sociopolitical, racial, or ethnic categories. Diverse populations identified as 'Hispanic,' for example, are heterogeneous and have distinct ancestries and social histories. Social experiences and conditions inform racial identity, making such identity a poor proxy for genetic ancestry.

Consumers may purchase genetic tests to learn about their race or ethnicity, but there is no clear-cut connection between an individual's DNA and his or her racial or ethnic affiliation. Worldwide patterns of human genetic diversity are weakly correlated with racial and ethnic categories because both are partially correlated with geography and the history of human migration. Current understandings of race and ethnicity reflect more than genetic relatedness, having been defined in particular sociohistorical contexts (i.e., European and American colonialism). In addition, social relationships and life experiences are important components in shaping individual identity and group membership and in some cases, may supercede biological ancestry in significance. Many genetic ancestry tests also claim to tell consumers where their ancestral lineage originated and the social group to which their ancestors belonged. The idea of 'returning to one's homeland' as identified by genetic ancestry information must be treated with care, as modern populations from whom samples have been collected have not over time remained unchanging. Current patterns of residence are rarely identical to what existed in the past, and social groups have changed over time, in name and composition.

Commercial Biobanking and Research

A feature of DTC personal genomics is the development of commercial biobanks using consumer DNA samples for in-house research and/or licensing to third parties. Companies such as *23andMe, Inc.* have amplified the development of such commercial biobanks and suggest that by creating their own large databases from consumer DNA samples, they are well poied to more efficiently 'scale up' for population-based research. Going a step further toward 'accelerating genomic research,' *23andMe* unveiled its new research arm, *23andWe*, in the spring of 2008. Stating that "23andMe isn't just about you," the company advertised *23andWe* as a new model of actively engaging with their consumers offering "customers the opportunity to leverage their data by contributing it to studies of genetics." As a *23andMe, Inc.* customer, you are asked at the time of purchase whether you consent to having your sample and genetic information included in the company biobank. The company states that you may choose to discontinue this relationship at any point; however, once your genotypic information is de-identified of personally identifying information such as your name, address, etc., your information cannot be deleted from the biobank data.

The stated purpose for creating company biobanks articulated by companies like *23andMe, Inc.*, is to facilitate research. Recruitment of human subjects into clinical trials remains a challenge for biomedical researchers, particularly the enrolling of individuals of diverse backgrounds. One *23andWe* research initiative focused on Parkinson's disease through a collaboration with the Scripps Research Institute and the Michael J. Fox Foundation. Individuals diagnosed with Parkinson's disease were recruited and given personal genetic testing free of charge by the company. Research comparing the genotype of those afflicted with the disease and the control group of *23andMe* consumers enrolled in the company biobank led to the publication of SGK1 gene that may be protective against Parkinson's disease in individuals who carry the rare risk-associated LRRK2 G2019S mutation. Based on this research, the company filed a patent on this research in May 2012. Defending the patent, *23andMe, Inc.* argued that the patent was necessary to increase the likelihood that a biotech or pharmaceutical company would pursue drug development and/or other types of research on the disease.

The inclusion practices of DTC personal genetic testing company biobanks create ambiguity on when an individual is being treated as a consumer and when he or she becomes a research participant. Customers, drawn to specific genotyping services by their interest in learning about their own personal data, may be unaware about the implications of donating their genetic data into a biobank for future research. Several scholars have suggested that it will be important for companies to use separate informed consent processes for commercial and research activities, and to establish explicit policies about secondary use of samples. Companies should disclose who would have access to samples and data and the types of research and lengths of time during which samples could be used. In addition, any research using consumer data and samples would require Institutional Review Board (IRB) review of research.

Data Protection and Privacy

One of the major challenges to coordinated research efforts in DTC personal genetic testing companies is the storage and sharing of data across institutions and national borders. There remains great difficulty in obtaining genetic information with highly detailed phenotypic information, and the desire for greater efficiency and utility of genetic data may compete with the ethical concerns of maintaining individual privacy and public trust. Individuals are often sampled in a variety of contexts and although great efforts may be taken to de-identify, encrypt, or de-link personal information, there are substantial fears that the multiple practices through which samples are obtained, stored, and then redistributed for basic research presents serious challenges to maintain anonymity and confidentiality. In their review, Lunshof et al. (2008) list the broad range of possible threats to privacy and confidentiality in genomic research that may occur. These threats include re-identification of individual information after de-identification, which may be implemented by using a combination of surnames as well as genotype and geographical information. Other possible approaches include inferring phenotype from genotype through information such as stature, hair, eye or skin color, or identification of a first-degree relative, in DNA and RNA.

These potential threats are a source of concern for the public. Research indicates that participants are often conflicted. Although they are committed to biomedical research and are comfortable with their information being shared within the medical community, they fear that their information might be leaked. When presented with the choice of restricting access to their information, most participants' concerns over privacy outweigh their sense of altruism participating in research to improve health, and is more likely to lead them to refuse public release of data in data sharing. Current policies make a distinction between information that is solely genetic and genetic information that is accompanied by phenotypic data. In the case of whole genome information, access to genetic information is allowable with specific consent for data sharing. However, if phenotypic information is included, policies recommend only controlled access with specific consent for data sharing. In the case of information on individual gene variants that are included in sets of aggregated data, public access with general consent for future use of genotypic data is possible; however, scholars suggest that when phenotypic data is added to these data sets, they should be subject to controlled access that requires obtaining general consent for future use from participants. Currently, no guidelines on data sharing and privacy protection exist for the DTC personal genetic testing industry.

Geneticization and Human Identity

In a context where the public is bombarded with new information on genetic variants that are asserted to be associated with broad range of conditions from premature baldness to childhood obesity, there is growing potential for the misconception that genetics determines most human conditions and traits. Reports of results from GWAS has furthered ratified the importance of genes and oversimplified understanding of complex human diseases through enlargement of the potential role of any one genetic variant. Aside from the rare, single-gene disorder such as Huntington's disease, genes often play a limited role in the pathophysiology of complex human conditions and traits such as diabetes, coronary heart disease, and asthma. Single genes are most often mediated by other genes, interacting with environmental factors in multiplicative ways that may ultimately affect an individual's likelihood to develop a specific phenotype.

The significance placed on genetics is easily understood within the context of our technoscientific culture. The gene has become a powerful symbolic icon in our society in which genetic information is believed to be uniquely identifying, holding great significance for individuals in defining who and what they are. The clichéd explanation of 'it must be in their genes' for individuals displaying unusual musicality or athleticism reveals the entrenched view that certain human traits are destined in our genetic code. In translating human genetic variation research into clinical decision making and individual health practices it will be equally important to study the impact of the ubiquitous circulation of genetic information on conceptions of humanness and identity.

Concern for groups such as American Indian tribes, are behind efforts to explain origins by genetic ancestry information without regard to the cultural, religious, social, historical, and political processes that also inform group origin, membership and identity, and access to group rights. Some related issues include the use of genetic ancestry information as the basis for: changing one's identity on various government forms; making claims to certain group rights or benefits; and for immigration purposes, such as seeking dual citizenship. These issues are of increasing practical concern and likely to become more so in the future.

Among the most pervasive and pernicious claims of genetically determined traits are theories on the racial ordering of intelligence. Despite the weak scientific basis for such ordering, the consistent return to the rhetoric of racial hierarchies of IQ reflects the powerful role that science has historically played in promoting racist ideologies. Current evidence suggests that for most complex behavioral traits, contribution of any one gene to normal variation is small and these traits may be more fully explained by variation in environmental factors. It is important to caution against making the naive leap to a genetic explanation for group differences in a complex behavioral trait, where environmental and social factors clearly can and do play major roles. These issues are particularly important, as personal genetic information becomes readily available through DTC personal genetic testing services and the general public faces the challenge of rendering genetic data meaningful.

See also: Anthropology, Genomics, and Human Variation: National Roots; Biobanking: Ethical Issues; Bioethics in the Post-genomic Era; Biotechnology; Genetic Ancestry Testing; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Screening for Disease; Genetics and Society; Genetics: Legal Aspects; Genetics: The New Genetics; Genomics, Ethical Issues In; Human Genome Project: History and Assessment; Science and Technology, Social Study of: Computers and Information Technology.

Bibliography

Bloss, C.S., Schork, N.J., Topol, E.J., 2011. Effect of direct-to-consumer genomewide profiling to assess disease risk. New England Journal of Medicine 364, 524–534.

Borry, P., Cornel, M., Howard, H., 2010. Where are you going, where have you been: a recent history of the direct-to-consumer genetic testing market. Journal of Community Genetics 1, 101–106.

Collins, F., 2010. The Language of Life: DNA and the Revolution in Personalized Medicine. HarperCollins Publishers, New York.

Direct-to-Consumer Genetic Tests, 2010. Misleading Test Results are Further Complicated by Deceptive Marketing and Other Questionable Practices. United States Government Accountability Office.

Green, R.C., et al., 2009. Disclosure of APOE genotype for risk of Alzheimer's disease. New England Journal of Medicine 361, 245–254.

Harmon, A., 2007. My genome, myself: seeking clues in DNA. The New York Times. http://www.nytimes.com/2007/11/17/us/17dna.html.

Hudson, K., Javitt, G., Burke, W., Byers, P., Committee, with the A.S.I, 2007. ASHG Statement* on Direct-to-Consumer Genetic Testing in the United States. http://journals.lww.com/greenjournal/Fulltext/2007/12000/ASHG_Statement__on_Direct_to_Consumer_Genetic.28.aspx.

Kaye, J., 2008. The regulation of direct-to-consumer genetic tests. Human Molecular Genetics 17, R180–R183.

Lunshof, J.E., Chadwick, R., Vorhaus, D.B., Church, G.M., 2008. From genetic privacy to open consent. Nature Reviews Genetics 9, 406–411.

Magnus, D., Cho, M., Cook-Deegan, R., 2009. Direct-to-consumer genetic tests: beyond medical regulation? Genome Medicine 1, 1–3.

Murray, A.B.V., Carson, M.J., Morris, C.A., Beckwith, J., 2010. Illusions of scientific legitimacy: misrepresented science in the direct-to-consumer genetic-testing marketplace. Trends in Genetics 26, 459–461.

Disabilism

Elizabeth Lightfoot, University of Minnesota, Saint Paul, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Despite the widespread adoption of national and international policies banning disability discrimination, people with disabilities still experience various forms of disability discrimination known as disablism. While disablism is pervasive, the concept of disablism is less familiar among the mainstream audiences than similar concepts like racism and sexism. The profession of social work, with its core value of social justice, has an important role to play in reducing disablism in society. However, the social work profession has its own history of institutional disablism related to the medical model orientation to service provision. The profession of social work can begin to address disablism through focusing on increasing disability awareness within the profession, adopting an anti-oppressive practice approach, and embracing innovative social-service innovations that reduce disablism, such as self-directed supports, person-centered planning, and universal design.

Introduction

Roughly 15% of the world's population, or 650 million adults, have some sort of disability (WHO, 2011). About 80% of people with disabilities live in developing countries, which have a higher prevalence rate of disability due to poor medical care, lack of access to immunizations and other preventative measures, and poor nutrition. (This article uses people-first language – 'person with a disability.' However, the author acknowledges that many people with disabilities view 'disabled person' as the more appropriate language and encourages social workers to use the language desired by people with whom they are working.) Disability is more common among people over age 50 due to age-related disabilities. The high prevalence rate of people with disabilities means that social workers practicing in any subfield of social work are likely to work with people with disabilities.

In the latter part of the twentieth century, there was growing awareness worldwide that people with disabilities experienced specific forms of discrimination and exclusion from mainstream society, and faced physical, legal, attitudinal, and other barriers to participation in regular life activities (Barnes, 1991). Disability activists began to form national and international movements demanding policies that eliminated discrimination and provided services and supports. Consequently, many nations have now adopted antidiscrimination laws banning various types of disability-related discrimination. Routine types of disability discrimination that have been banned include discrimination in employment, housing, education, public services, transportation, and places of public accommodation, including stores, restaurants, or personal services. Some of the earliest national antidiscrimination laws were the United States' Americans with Disabilities Act of 1990, Australia's Disability Discrimination Act of 1992, and the United Kingdom's Disability Discrimination Act of 1995, while Canada had included a stipulation against disability discrimination in its Constitution back in 1982. Many other nations adopted similar types of antidiscrimination laws in the 1990s and early 2000s, and the United Nations formally adopted the Convention on the Rights of Persons with Disabilities in 2006 that entered into force in 2008. As of 2012, 169 countries have formally ratified this Convention (United Nations, 2012).

Though the concept of disability discrimination is now more widely understood, there is no commonly agreed term for disability discrimination similar to racism for race-based discrimination or sexism for gender-based discrimination. The two most common terms are disablism and ableism, although their use in mainstream society varies according to country. This article first introduces disability definitions and models, and then discusses disablism in the context of social work practice.

Disability Definitions and Models

Understandings of disability vary according to culture, country, gender, philosophical orientation, religious values, region, and prevailing societal attitudes. This has resulted in the lack of a standard definition of disability. Common definitions of disability include disability being defined as a person having a diagnosis with a particular condition or disease, such as cerebral palsy, Down syndrome, blindness or Huntington's disease; or disability defined as a person having a functional limitation, such as a limitation in walking, standing, thinking, seeing, or hearing. Disabilities are often categorized by type of disabilities, such as physical disabilities, sensory disabilities, or health-related disabilities. The categories of disability are called by different names in different countries. For example, in the United States, a person with a cognitive limitation, such as Down syndrome, is said to have an intellectual disability, while in the United Kingdom the person is said to have a learning disability. Disability within the disability rights movement has typically been defined broadly, including a wide range of categories of impairments. Thus, the broad umbrella term of disability includes people with intellectual/learning disabilities, sensory disabilities, physical disabilities, and mental health disabilities. Because there is not clear objective reality related to disability, the concept of disability has been referred to as 'socially constructed' (Liachowitz, 1988), 'socially created' (Oliver, 1996), or simply an 'administrative category' (Stone, 1984).

The World Health Organization developed an International Classification of Functioning, Disability and Health (2001), commonly referred to as ICF, to categorize disability beyond the concept of disease and disorder. This categorizes disabilities or health conditions in three ways: impairments a person has within their body, for example, those related to pain, cognition, or mobility; activity limitations, such as a difficulty walking or eating; participation restrictions related to discrimination or physical barriers. The ICF aims to be a universal classification of disabilities, which can be used to understand disability worldwide.

Models of Disability

Disability is often understood from different lenses referred to as models of disability. The main disability models are the medical model, the social model, the independent living model, and the minority model. These models represent more than just theoretical classifications of disability, but rather identify differing real-world understandings of disability that have directly shaped the experiences of people with disabilities. The models are not mutually exclusive, though the social, independent, and minority model contrast with the medical model.

Medical Model

The medical model of disability is the dominant twentieth and twenty-first century model for understanding disability among many professionals, and public perception of disability often reflects this model. In medical model, disability is seen as a pathological individual attribute linked to incapacity and dependence. A disability is defined by the presence of any functional limitation in daily life that are caused by a physical, mental, or sensory impairment. This model is rooted in the functionalism of Talcott Parsons (1958), in which people with disabilities assume the 'sick role' in society, thereby becoming exempted from normal social activities (DeJong, 1979; Parsons, 1958). Thus, people with disabilities must rely on professionals to help them recover, cope, or rehabilitate from their impairments (Hahn, 1985). The subtext of the medical model is that people with physical, sensory, or mental impairments have a tragic flaw and need expert help to overcome this flaw. The focus is on the pathology of an individual, without focus on the connections between that individual and society, or the environmental barriers that individuals with disabilities face (Brisenden, 1986). Health care and other professionals pervasively use the medical model when providing services to people with disabilities, and the institutional structures for providing services commonly reflect the medical model approach.

Social Model

The social model of disability focuses on the social construction of disability. According to the social model, while individuals may have a physical, sensory, or mental impairment, it is the failure of society to accommodate a range of impairments that results in 'disability.' This model was developed in the United Kingdom in the 1970s in response to the inadequacies of the perceived methodological individualism of the medical model of disabilities (Oliver, 1983). The social model argues that the divisions between disability and ability are socially constructed, and that this social construction is a middle range consequence of the economic and social forces of society. Disability is not a problem of the individual that must be overcome, as suggested by the medical model, but rather a result of multiple barriers created by society. People with disabilities experience societal oppression based on these societal barriers. The proponents of the social model also argue that a society's political and economic mandates reinforce the need for a category of dependents, with large bureaucratic structures and institutions perpetuating themselves by serving people with disabilities (Oliver, 1996). While the social model has been critiqued as neglecting gender (Morris, 1992) and being too rigid (Shakespeare and Watson, 2002), the social model of disability has become the predominant model for understanding disability among disability activists worldwide and is the underlying conceptual framework of the ICF. Many social workers working with people with disabilities now use the social model of disability and work to reject the medical model structures inherent in some forms of social work practice.

Independent Living Model

The independent living model arose in the United States in the 1970s at the same time as the social model, spreading swiftly throughout much of the world, and, like the social model, has influenced the international disability movement. The independent living model, first articulated by DeJong (1979), was developed in response to the medical model. The independent living model views disability as more than just a medical condition or an impairment, but as an interaction involving the individual, the society, and the environment. This model of disability views the role of professionals and government in the lives of people with disabilities as a key issue (DeJong, 1979). The solution to this problem is for people with disabilities to gain independence from professionals, and thus the independent living model has a strong emphasis on self-help and individual or consumer control of services as an antidote to the problems associated with impairments. The independent living model does not maintain that people with disabilities are able to do everything for themselves, but rather that they should have control over their lives and be able to make the decisions on how and when they receive support (Brisenden, 1986). The independent living model led to the development of an international network of self-help organizations run by people with disabilities, Centers for Independent Living.

Minority Model

The minority model of disability, articulated first by Hahn (1985), views environmental barriers as the principal barriers to participation in community life by people with disabilities, with discrimination playing a central role in their being denied access to jobs, buildings, and the community. The minority model views people with disabilities as a socially constructed minority group because of the shared discrimination they face,

and thus people with disabilities are similarly situated as racial and ethnic minorities and women. Similarly to the independent living model, the minority model views professionals, such as social workers, as having too much power over the lives of people with disabilities because this strips them of their independence (Gliedman and Roth, 1980). Drawing from his minority model, the civil rights approach focusing on passing antidiscrimination legislation is seen as an appropriate solution for people with disabilities to overcome barriers.

Disablism and Ableism

While there is no globally accepted term to define disability discrimination (Harpur, 2009), the two terms most often used are disablism and ableism. These terms are not equivalent, and their use overall in mainstream society is limited in comparison to racism and sexism. Usage varies according to country. For example, in Australia, ableism is more commonly used. Disability activists also disagree about which terminology is most appropriate and less oppressive.

'Ableism' refers to the belief that certain normative understandings of ability are preferential to others. Wolbring (2008) defines ableism as "a set of beliefs, processes and practices that produce – based on abilities one exhibits or values – a particular understanding of oneself, one's body and one's relationship with others" (pp. 252–253). The concept of ableism is more broadly focused than typical understandings of disability, as ableism refers to beliefs about any person who does not meet societal norms of ability, whether or not they have a disability (Wolbring, 2008; Harpur, 2009). However, the term ableism is typically used in the context of disabilities. Ableism in the context of disability means that certain physical, cognitive, sensory, or other types of abilities are preferred, and people who have differing abilities are considered impaired. Ableist views would regard someone who does not fit the ableist vision of society as someone who needed to be fixed. Ableist views of society in which certain abilities are valued over other abilities lead to the development of political and social structures that value these abilities. These ableist views are used by many to consciously or subconsciously justify unequal treatment of people with disabilities (Wolbring, 2008). Some have critiqued the term ableism as being derived from a medical model understanding of disability, as it could lead to a focus on individual ability rather than oppression (Withers, 2010). However, it does not appear that this is the intention, rather it is a term that is in line with the terms racism and sexism. Ableist beliefs about certain abilities lead to disablism, or discrimination against people who have disabilities.

The term 'disablism' describes negative beliefs and discrimination based on disability, focusing on the societal oppression faced by people with disabilities. Miller et al. (2004) define disablism as "discriminatory, oppressive, or abusive behavior arising from the belief that disabled people are inferior to others" (p. 28). People with disabilities experience disablism when they are excluded from participating in society through various types of barriers, when they are viewed as objects of pity, or when they are exhorted to overcome their disability. Disablism can occur at different levels. Individual-level disablism involves personal negative beliefs that individuals have about people with disabilities. Institutional disablism refers to the development of societal structures or institutional practices that leads to the systematic exclusion of people with disabilities. This can lead to additional social and physical barriers for people with disabilities. Some have critiqued the term disablism for placing the focus on the person with the disability or disability itself, rather than the discriminator or the act of discrimination (Harpur, 2009). Nonetheless, disablism is the preferred term for disability-related discrimination among many disability advocates, and is the term of choice in Great Britain. While the terms ableism and disablism have slightly different meanings, both terms convey an understanding of the types of societal discrimination that people with disabilities experience based on negative views toward people with disabilities.

While disablism focuses primarily on disability-related discrimination, people with disabilities have multiple social identities, and often experience what is known as multiple oppression based on their memberships in other oppressed or marginalized groups (Vernon, 1998; Lightfoot and Williams, 2009). For example, an African-American with a disability can experience racism when interacting with people with disabilities, disablism when interacting with African-Americans, and potentially both racism and disablism when interacting with others. Likewise people with disabilities can experience multiple oppressions related to gender, sexual orientation, class, age, national origin, religion, or other oppressed statuses, and experience multiple dimensions of exclusion. People with disabilities sometimes feel as if they have to choose between their identities (Lightfoot and Williams, 2009) and find no home in any one movement.

In the last several decades there has been a new emphasis on the intersectionalities between disability and other social categories, such as gender (Morris, 1992; Shakespeare, 1999), mothering (Lewiecki-Wilson and Cellio, 2011), congenital vs acquired disability (Shuttleworth et al., 2012), and body image (Wendell, 1996). For example, much has been written on the conflict between masculinity and disability as being in conflict with each other (Gerschick and Miller, 1994; Shakespeare, 1999; Shuttleworth et al., 2012), as masculinity is associated with autonomy and power, while disability is associated with dependency. These deeper understandings of intersectionalities give a better understanding of the contexts in which disablism occurs.

Disablism in the Field of Social Work

The field of social work has a complicated history with regard to disablism. Many social workers have worked hard at reducing disablism within society and have been leaders in the disability rights movement. However, social work as a profession has often been faulted for helping to perpetuate ableism and disablism by adhering to the medical model of disability (Mackelprang and Salsgiver, 1996; Meekosha and Dowse, 2007). Additionally, the field of social work historically played only a peripheral role in the disability rights movement, and has paid less attention to disablism, ableism, or disability discrimination than would be expected for

a value-based profession. Social work practitioners need to understand the relationship between the profession of social work and disablism when working with people with disabilities.

One of the key criticisms of the field of social work in relation to people with disabilities is the role social workers have played within the medically oriented system of disability services and supports. The field of professional social work in the twentieth century developed simultaneously with both the development of some types of services and supports for people with disabilities in many countries. Social workers, along with physicians, nurses, and other professionals, often are in positions where they assess the needs of people with disabilities to determine the types of services for which a person with a disability is eligible. The social worker often plays the role of 'expert' in the life of a person with a disability, serving as a key gatekeeper between a person with a disability and any services or supports he or she might need. The person with a disability is put in a position of dependency, as the person with a disability becomes dependent on the social worker to receive services and loses the opportunity to make choices regarding his or her own life (Mackelprang and Salsgiver, 1996; Meekosha and Dowse, 2007). Even the whole notion of social work as a 'helping' profession implies that people with disabilities need to be helped by others, situating social work squarely within the medical model approach to understanding disabilities. While social workers working with people with disabilities can and do work to eliminate disablism with the system, and this is more prevalent in some countries, the structures in which some social workers practice promulgate disablism.

Historically, social workers have been involved in making decisions about the lives of people with disabilities that people with disabilities might not have chosen for themselves. For example, social workers were often involved in decisions to institutionalize people with disabilities without their consent, to sterilize people with disabilities, or to remove infants upon birth from their parents with disabilities based on their disability diagnosis. These types of actions by social workers clearly could be categorized as disabling. However, even the seemingly benign social work tasks of assessing and determining eligibility for services can be considered disablist, when the social worker as expert has the last word in determining eligibility. Likewise, the emphasis on developmental or program plans for people with disabilities, which are common for social workers when working with people with disabilities, could also be seen as disablist. In such plans, people with disabilities often are given goals to work on for self-improvement, such as reducing negative behaviors. This type of approach could imply that people with disabilities must work on overcoming their disability, rather than focusing on accepting a person's impairments and working on reducing societal barriers. A disablist system is not the fault of the individual social worker, but rather the system often requires the social worker to play the role of policy enforcer or gatekeeper. This has resulted in some people with disabilities viewing the social work profession with suspicion (Mackelprang and Salsgiver, 1996; Beaulaurier and Taylor, 2001).

While there were some social workers involved in the disability movement from its beginning and many social workers are now involved in the front lines of the disability movement, the field internationally has been somewhat ambivalent in its stance on disability discrimination and disability rights and as a whole has not embraced the disability rights movement in a similar fashion as it embraced the civil rights movements of other oppressed groups in many countries. In fact, the disability rights movement in the United States in particular and also in the United Kingdom started partially as a reaction to having professionals, such as social workers, having control over the lives of people with disabilities (Beaulaurier and Taylor, 2001). Groups of people with disabilities started their own organizations that were distinct from organizations that others organized for people with disabilities, such as groups that were run by or dominated by professionals. These groups of people with disabilities, such as Centers for Independent Living, often focused on radical self-help, where people with disabilities determined and controlled their own services and supports without professionals being involved. Despite the efforts of some social workers to become allies in the disability movement, social workers in general were not seen as natural allies to the movement, and the field of social work was not active in causes fighting for antidiscrimination laws or eliminating disablism within society (Mackelprang and Salsgiver, 1996). While social work's self-purported values include an emphasis on social justice and empowerment, disability has not been a core focus of social work in many countries, such as the United States and Australia (Meekosha and Dowse, 2007). In these countries, compared to other fields of social work practice, there are fewer social work students, educators, researchers, and practitioners interested in disability and involved in disability activism in comparison to other fields of social work (Mackelprang and Salsgiver, 1996).

Addressing Disablism Social Work and Social Work Practice

Social work, with its core value of social justice and its ecological and systems understanding of society, has an important role to play in reducing disablism and ableism in society. However, social work needs to address disablism, ableism, and disability discrimination within the profession and in the nature of service delivery. There have been great strides in the past few decades in reducing the disablism within social work and the disablist nature of the provision of services and supports. Below, I consider some of these efforts.

Reducing Disablism within the Social Work Profession

Disability Awareness among Social Workers

As the social work profession has had a role in disablist society, the first step for social workers in addressing disablism in society is through reflection and change within the profession. There has been a lack of awareness among social workers about disability in general, and the concepts of disablism or ableism. For example, social work education in the United States has typically not included a strong focus on disability, and though

content on disability discrimination and disability-related oppression is required in all social work programs, when disability is discussed within social work education it often focuses on diagnosis and service provision, thus reflecting a medical model orientation (Gilson and DePoy, 2002). Social work students, who are not presented with the social, minority, and independent living models of disability and are not given the opportunity to grapple with the concepts of disablism or ableism, miss the opportunity to reflect upon the historical role social work has played in the lives of people with disabilities. Even when disablism is commonly discussed in social work education, as in the United Kingdom, the reality of social work practice being dominated by the medical model in the United Kingdom hospital settings makes the need to grapple with this content in an in-depth manner during social work training crucial. Increasing disability awareness among social workers through social work education is vital for reducing disablism in social work.

Anti-Oppressive Practice

The movement toward anti-oppressive practice is another step toward reducing disablism within the social work profession. Anti-oppressive practice is an approach to social work in which social workers align themselves with oppressed people, with the intent not only to eradicate social injustice, but also to eliminate social work's role in perpetuating oppression (Dominelli, 2009). Dominelli (2002) argues for a holistic approach to oppression that does not place different kinds of oppression in a hierarchy and that deals with individual needs and structural change – both inside and outside social work. Anti-oppressive practice requires social workers to understand, recognize, and work to reduce power imbalances (Dalrumple and Burke, 1995), which are especially stark within a disablist society. An anti-oppressive practice approach would necessarily require social workers to align themselves as allies with not only the recipients of services, but also the disability rights movement overall. Adopting anti-oppressive practice in social work would essentially entail social workers forsaking the medical model understanding of disability, and relinquishing their role as experts in the lives of people with disabilities. As the disability rights movement has embraced a self-help approach to services and has rejected the paternalistic role of professionals, including social workers, in their lives, an anti-oppressive practice approach might be one way for social workers to align themselves with people with disabilities. Social workers engaging as allies need to be critically self-reflective in their work to be sure they do not replicate their positions of power when connecting with people and groups of people with disabilities.

Reducing Disablism in the Delivery of Services and Supports

The nature of the delivery of services and supports for people with disabilities also can reproduce disablism in society. Fortunately, there have been innovations in the structure and nature of disability services that can lead to people with disabilities having more control over their lives and being less dependent on professionals. Some of these innovations include the rise of self-directed or consumer-directed services, person-centered planning, and universal design of services.

Self-Directed or Consumer-Directed Services

At the onset of the disability rights movement in the 1970s, people with disabilities advocated for independence from professionals, such as social workers, who were controlling most aspects of their lives, including where they lived and who provided them with services. When agencies direct services, professionals like social workers often determine not only the type of services that a person with disabilities might receive, but also the particular personnel who will provide these services. Many countries have now adopted self-directed or consumer-directed services, which are a type of service arrangement where people with disabilities use governmental funds to hire their own services and supports (Mahoney et al., 2000). For example, in the United Kingdom, most people with disabilities are able to receive direct cash payments in the place of social services or supports that they used to receive through the government. These cash payments must be used on their own care needs, but the person with a disability is responsible for obtaining these services, including hiring their own personal assistants (Glasby and Littlechild, 2009). Personal budgets, which have even more flexibility or now available in a number of countries, including the United Kingdom, the Netherlands and Canada. Under this arrangement, a social worker's role changes from being the expert in the life of the person with disability to being more of a broker of services and facilitator of self-advocacy (Mackelprang and Salsgiver, 2009). By allowing a person with disability to control the services they receive, the self-directed approach helps to reduce the disablism that people with disabilities experience in their encounters with the social welfare system.

Person-Centered Planning

Person-centered planning is another advance in supporting people with disabilities that works to empower people and reduce the disabling nature of the social welfare system. This approach arose in the field of intellectual disabilities and has been used for decades (O'Brien and O'Brien, 2001). Person-centered planning is a process in which people with disabilities and their families plan for their future based upon the person's own desires. This phenomenological approach begins with understanding a person's own version of reality, and then planning and developing supports and services that best help the individual to meet his or her own goals. It differs from traditional social work assessment in that it is not services focused or needs focused, and involves structured exercises in which people who know the person with the disability best, typically not just professionals, but also family members, friends, neighbors, and others, provide input. Professionals, such as social workers, play the role of facilitator and information and resource provider, rather than expert.

Universal Design

Another approach to reducing disablism in service delivery is the incorporation of the concept of universal design as much as

possible into society in general, as well as in social welfare service delivery. Universal design is an approach in which accessibility for people of all ages and abilities is included in the initial design (The Center for Universal Design, 1997). Universal design was originally used in the context of architecture. If buildings or public spaces were initially designed to be accessible, such as by incorporating ramps, wide doorways, and other accessible features into the initial design, then people with disabilities would face fewer barriers to participation and there would be less need for people with disabilities to have to ask for accommodations. The concept has since been applied to programs and services as well, with the concept being used extensively in the area of education. For example, in universal instructional design in higher education, the format of courses is varied to attempt to provide as universal access as possible for students with varying types of needs, such as language abilities, family situations, cultural background, or abilities. This might include using varied instructional methods, incorporating natural supports, and providing flexible means of evaluation (Lightfoot and Gibson, 2005). Universal design explicitly rejects ableist notions with the intent of designing structures or services for people with as diverse a range of people as possible. The shift toward universal design at the initial design of programs can take the focus away from disability and make it the responsibility of program designers to minimize barriers to participation. While social welfare services have not extensively adopted universal design principals, incorporating components of universal design into the structure of social welfare services can help reduce disablism.

Disability Activism

Finally, if the social profession is interested in reducing disablism in society, it must ally with the disability movement and disability activists to work together to reduce societal barriers that people with disabilities face; change disablist policies; improve service delivery models and supports; foster empowerment of people with disabilities; and change societal attitudes about disability. Disability activism fits well within the social justice framework of social work.

See also: Community Care in England; Critical Social Work Practice; Disability Movement; Oppression; Social Work Theory.

Bibliography

Barnes, C., 1991. Disabled People in Britain and Discrimination: A Case for Anti-Discrimination Legislation. C. Hurst & Company, London.

Beaulaurier, R., Taylor, S., 2001. Social work practice with people with disabilities in the era of disability rights. Social Work in Health Care 32 (4), 67–91.

Brisenden, S., 1986. Independent living and the medical model of disability. Disability, Handicap and Society 1 (2), 173–178.

Dalrumple, J., Burke, B., 1995. Anti-Oppressive Practice: Social Care and the Law. Open University Press, Buckingham, UK.

DeJong, G., 1979. Independent living: from social movement to analytic paradigm. Archives of Physical Medicine and Rehabilitation 60.

Dominelli, L., 2002. Anti-Oppressive Social Work Theory and Practice. Palgrave MacMillan, London.

Dominelli, L., 2009. Anti-oppressive practice in context. In: Adams, R., Dominelli, L., Payne, M. (Eds.), Social Work: Themes, Issues and Critical Debates. MacMillan, London.

Gerschick, T., Miller, A., 1994. Gender identities at the crossroads of masculinity and physical disability. Masculinities 2, 34–55.

Gilson, S., DePoy, E., 2002. Theoretical approaches to disability content in social work education. Journal of Social Work Education 38 (1), 153–165.

Glasby, J., Littlechild, R., 2009. Direct Payments and Personal Budgets: Putting Personalisation into Practice. The Policy Press, Bristol.

Gliedman, J., Roth, W., 1980. The Unexpected Minority: Handicapped Children in America. Harcourt Brace, Jovanovich, New York.

Hahn, H., 1985. Towards a politics of disability: definitions, disciplines, and policies. The Social Science Journal 22 (4), 87–105.

Hahn, H., 1985b. Disability policy and the problem of discrimination. American Behavioral Scientist 28 (3), 293–319, doi: 10.1177/000276485028003002.

Harpur, P., 2009. Sexism and racism, why not ableism? Calling for a cultural shift in the approach to disability discrimination. Alternative Law Journal 34, 163–167.

Lewiecki-Wilson, C., Cellio, J., 2011. Disability and Mothering Liminal Spaces of Embodied Knowledge. Syracuse University Press, Syracuse.

Liachowitz, C., 1988. Disability as a Social Construct: Legislative Roots. University of Pennsylvania Press, Philadelphia.

Lightfoot, E., Gibson, P., 2005. Universal instructional design: an innovative framework for accommodating social work students. Journal of Social Work Education 41 (2), 269–277.

Lightfoot, E., Williams, O., 2009. The intersection of disability, diversity and domestic violence: results of national focus groups. Journal of Aggression, Maltreatment & Trauma 18 (2), 133–152.

Mackelprang, R., Salsgiver, R., 1996. People with disabilities and social work: historical and contemporary issues. Social Work 41 (1), 7–14.

Mackelprang, R., Salsgiver, R., 2009. Disability: A Diversity Model Approach in Human Service Practice. Lyceum Books, Chicago.

Mahoney, M., Simone, K., Simon-Rusinowitz, L., 2000. Early lessons from the cash and counseling demonstration and valuation. Demonstration and Evaluation 23 (3), 41–46.

Meekosha, H., Dowse, L., 2007. Integrating critical disability studies into social work education and practice: an Australian perspective. Practice: Social Work in Action 19 (3), 169–183.

Miller, P., Parker, S., Gillinson, S., 2004. Disablism: How to Tackle the Last Prejudice. Demos, London.

Morris, J., 1992. Personal and political: a feminist perspective on researching physical disability. Disability, Handicap & Society 7 (2), 157–166.

O'Brien, C., O'Brien, J., 2001. The Origins of Person-Centered Planning: A Community of Practice Perspective. Center of Human Policy, Syracuse University, Syracuse. http://thechp.syr.edu/PCP_History.pdf (accessed 22.12.2012.).

Oliver, M., 1983. Social Work with Disabled People. MacMillan, Basingstoke.

Oliver, M., 1996. Understanding Disability: From Theory to Practice. Macmillan, London.

Oliver, M., Barnes, C., 2012. The New Politics of Disablement. Palgrave Macmillan, London.

Parsons, T., 1958. Definitions of health and illness in light of American values and social structure. In: Jaco, E.G. (Ed.), Patients, Physicians, and Illness. The Free Press, New York, pp. 121–144.

Shakespeare, T., 1999. The sexual politics of disabled masculinity. Sexuality and Disability 17, 53–64.

Shakespeare, T., Watson, N., 2002. The social model of disability: an outdated ideology? Research in Social Science and Disability 2, 9–28.

Shuttleworth, R., Wedgwood, N., Wilson, N., 2012. The dilemma of disabled masculinity. Men and Masculinities 15 (2), 174–194.

Stone, D., 1984. The Disabled State. Temple University Press, Philadelphia.

The Center for Universal Design, 1997. The Principles of Universal Design: Version 2.0. North Carolina State University, Center for Universal Design, Raleigh, NC.

United Nations Enable, 2012. Convention and Optional Protocol Signatures and Ratifications. United Nations, New York. http://www.un.org/disabilities/countries.asp?navid=12&pid=166 (accessed 21.12.2012.).

Vernon, A., 1998. Multiple oppression and the disabled people's movement. In: Shakespeare, T. (Ed.), Disability Reader: A Social Science Perspective. Continuum Press, London, pp. 201–220.

Wendell, S., 1996. The Rejected Body: Feminist Philosophical Reflections on Disability. Routledge, London.

Withers, A., 2010. Disablism or Ableism? If I can't Dance is It Still My Revolution?. http://still.my.revolution.tao.ca/node/68 (accessed 21.12.2012.).

Wolbring, G., 2008. The politics of ableism. Development 51, 252–258.

World Health Organization, 2001. International Classification of Functioning, Disability and Health. World Health Organization, Geneva.

World Health Organization, 2011. World Report on Disability 2011. World Health Organization, Geneva.

Disability and Disabling Practices

Tania Jivraj and Esther Ignagni, The School of Disability Studies, Ryerson University, Toronto, ON, Canada

Published by Elsevier Ltd.

Abstract

This article provides an overview of the interdisciplinary field of disability studies. It highlights the social model of disability, a foundational analytic framework within disability activism, practice, and scholarship. Through explicating and historicizing the emergence of disability as a social category, the article offers the social sciences an emancipatory lens through which to reflexively examine and dislodge disabling assumptions and practices. The article concludes by illustrating how disability studies can complement and expand the social justice orientations of contemporary social science research and practice.

Disability studies is a relatively new and burgeoning interdisciplinary field. Emerging from and intertwined with the efforts of disabled activists, the field has been critical from its inception, and remains committed to a process of reflexive revision and extension. The official definition adopted in 1993 by the Society for Disability Studies notes, that "Disability Studies [seeks] to disentangle impairments from the myth, ideology and stigma that influence social interaction and social policy" (Society for Disability Studies, n.d.). Further, "[D]isability studies holds that disability is a complex political and cultural effect of one's interaction with an environment, not simply a medical condition to be eliminated" (Dolmage and Lewiecki-Wilson, 2010: p. 30).

Disability studies, similar to other scholarly equity projects, 'cuts across a range of disciplines' – from medicine and psychology to art and philosophy to education and social work – to examine how power and privilege function to create and reproduce ableist cultures (Mallett and Runswick-Cole, 2014). Ableist cultures are constituted by two prevailing systems of thought. The first constructs disability as an inherent deficit or anomaly of the individual and the second promulgates compulsory able-bodied and reasoned normative standards to which we must all aspire (Campbell, 2008; McRuer, 2006). Together, these systems conspire to justify and structure the marginalization of people who live with or who are presumed to have impairments (Stevens et al., 2011). In challenging assumptions of the normative body, disability studies scholars begin their work in the experiences of people who live or are labeled with impairments to analyze, critique, and reimagine social, political, and legal phenomena. As the Canadian disability rights activist and scholar Catherine Frazee (2003) has noted:

> … disability studies "takes for its subject matter not … [particular] variations … in human behavior, appearance, functioning, sensory … and cognitive processing, but the *meaning* we make of those variations" (Linton, 1998, p. 2). According to Simi Linton (1998), it is "a field that explores the critical divisions our society makes in creating the normal versus the pathological, the insider versus the outsider, or the competent citizen versus the ward of the state" (p. 2). The project of Disability Studies, according to Peter Monaghan, is to "pull apart concepts about disability to see what cultural attitudes, antagonisms, and insecurities went into shaping them" Monaghan, 1998: pp. 15–16.
>
> Frazee, 2003: p. 258

Thus, transforming current conceptualizations of the disabled body (and mind) as one to be pitied, to be cured, to be contained, and to be hidden to one that recognizes disability as a political and cultural identity formed from our responses to differences, where difference is unquestioningly and routinely pathologized.

Disability studies embraces an agenda for social transformation. Throughout its short history, disability studies has called for a broader, more inclusive conception of human rights (Harpur, 2012; Peters, 2008). More radically, materialist proponents within disability studies call for the replacement of disabling capitalist economies with inclusionary alternatives (Tregaskis, 2002). Recent expansions of the field illuminate how 'disability is a teacher' (Titchkosky, 2003), how disability can be an 'occasion to see the world otherwise' (Schormans, 2013), and how disability generates new ways of making interdependent communities, calling the bounded individualized neoliberal self into question (Goodley, 2014).

A Word about Language

There is no one definition of disability. Rather, disability is understood in a variety of ways depending on the field of study or profession. In this article, disability refers to the complex interplay of how social, economic, cultural, and historical events construct an experience of disablement. Just as the term disability is contested, so to is disability as an identity. It is important to state that there is no essentialized disabled identity, since people living and labeled with impairments have complex and intersecting identities that simultaneously mitigate and exacerbate how their disablement is experienced (Rosenfeld et al., 2006). I alternatively use the terms disabled peoples or people living and labeled with impairments. The former reflects language borne of the social model of disability that recognizes that our environments disable people rather than situating the disability in the individual. Thus people may live with impairments, but are disabled through the exclusionary social and physical arrangements of our worlds that systematically fail to account for impaired bodies and minds. Finally, the phrasing 'individual labeled with an impairment' reflects constructivist understanding, with the focus being on the 'labeled with' rather than the person or the disability. In this way, I am acknowledging that people are acted upon by external forces that form how they are understood in their society.

Why Is Disability Studies Relevant to Social Work?

> Although social work has a rich history of advocacy for oppressed people, the profession has been hesitant to become involved with people with disabilities.
>
> Mackelprang and Salsgiver, 1996: p. 7

Current social work practices and education surrounding disability are steeped in ideas of the disabled body needing to be fixed and the disabled person needing sympathy and help (Profitt, 2010). The repercussions of this ideology of disability as a personal deficit can include segregation, intrusive state control, and limited access to a public life. In this article, we use principles taught in disability studies to lay the groundwork for an expanded understanding of disabilities that will hopefully support social workers to question their own biases about disabilities, so that we may not inadvertently engage in discriminatory professional practices that further disable people.

Though the role of social workers is vast and varied, a core of the profession is that we work to promote equity and inclusion of marginalized populations. This work of equity is particularly salient when we consider that disabled peoples are more likely to be discriminated against in terms of employment, housing and access to education, and more likely to be victims of violence (Chenoweth, 1996). This discrimination, or ableism, is based on the presumption of disabled bodies as incapable, dependent, and broken. The World Health Organization does not particularly name ableism as a factor when it states that "people with disabilities have worse health and socioeconomic outcomes … lower education achievements, less economic participation and higher rates of poverty than people without disabilities" (World Health Organization, 2011: para 6). Ableism is inferred when "negative attitudes" (para 6) and "lack of service provision" (para 6) are listed as barriers to full inclusion. As social workers, it is imperative that we work with disabled peoples to transform these structural and attitudinal barriers. Especially as we may work in organizations whose foundations are based on the medical model of disability.

The conscious act of challenging preconceived notions of disabilities is part of the work of disability studies. Disability studies seeks to challenge the oppression of disabled peoples by exposing multiple understandings of disabled identities as a part of the human spectrum of difference and the cultural meanings ascribed to those differences.

From a practice perspective, it is vital that as service providers we reflect on our assumptions of the meanings we make of disabled bodies in order to

1. challenge a cultural conflation of ability with humanity,
2. not further oppress already marginalized populations,
3. recognize the knowledges embodied in disabled experiences,
4. support ally-ship, community building, and political action,
5. change our own lives as we understand how labels of disability are a form of social control and surveillance for all bodies (living in fear of our own potential impairments and disability), and
6. reap in the benefits of a more just society and alternatives to being that can liberate us all.

The Construction of Disability in Western History

Our contemporary understanding of disability can be historically and geographically situated in many ways; Western activists and scholars tend to begin their analysis with the rise of industrial capitalism (Barnes, 2010; Oliver, 1990). Industrial capitalism's profound influence on the fabric of social and economic life created the conditions under which impairments were socially produced. New industrial disciplinary regimes led to a new emphasis on individual productivity, efficiency, and effort, in turn producing new forms of exclusion. Large numbers of the population who could not readily keep up with or accommodate themselves to the physical (and mental) demands of industrialization found themselves excluded from work. Furthermore, the repetition, speed and duration of industrialized work practices, dangerous work conditions (poor air quality, crowding, equipment failure, lack of safety practices/equipment, etc.), and meager remuneration conspired to 'break-down workers' bodies and minds. Thus, there were more people with impairments, and considerably more disability. Those who could not find a 'productive' place within industrial economy quickly became destitute. As the ranks of those who could not sustain employment within the industrialized economy swelled, governments faced a new problem – what to do about the large numbers of people living in dire poverty. In rural communities, those who could not work were cared for in the home – but with the changing family structure, the 'workless' had to turn in growing numbers to charity. Governments feared civil unrest and criminal activity, as well as the draining of public coffers as the ranks of the unemployed and poor grew. Warehousing and classification of the poor was tasked to the newly formed Poor Law Commissions, who sought to contain and regulate workless people through dividing practices (Skeggs, 2014). Crucially a system of classification emerged to distinguish between the legitimate and the illegitimate recipients of state benevolence; however, an objective index to measure capacity to work was required for its maintenance. By the early nineteenth century, medicine was well equipped to assume this powerful role. Louis Pasteur and others had discovered the germ theory of illness and disease; new technology such as the X-ray, the microscope, and the stethoscope allowed for the precise measurement required in the welfare state. This new technology allowed experts to dissect the human body into various functional units that could be individually measured. Medical methodology was positioned as objective, scientific, and empirical. Medical science was therefore incorporated as an apparatus of the state, with doctors as gatekeepers to social welfare (Barnes et al., 1999; Frazee, 2003). These experts then deemed that the bodies of people with impairments, children, and the elderly were considered incapable of working (Frazee, 2003; Oliver, 1990).

This gatekeeping function has remained remarkably durable. Social workers still defer to medicine to legitimate access to welfare disbursements, specialized equipment, vocational training, and adapted living arrangements (Nathan and Webber, 2010). In this way, service providers in social service support agencies depend on the legacies of medically determined knowledges about disabled peoples to also act as gatekeepers. Social workers in effect can be considered as extending the practices of legitimation in determining which bodies will receive support and which ones will not such that disabled persons experience medicine as a site of misplaced authority. The medical model and the institutions upon which helping professions are built wield a great deal of power and control over fundamental aspects of our lives. And, as Michael Oliver (2009) notes, they:

> have not been noticeably reticent about using this power to make decisions about … where [we] should live, whether [we] work or not, what kind of school [we] should go to, what kinds of … services [we] should receive and in the case of unborn disabled children, whether [we] should live or not (p. 24).

Medicalizing disability assumes a problem in individual functioning that supports the dominance of a 'normal' body rather than an examination of how normalizing processes are a way of pathologizing differences. Without additional models with which to engage in disability politics, social workers are at great risk of engaging in oppressive interventions and practices. Disability studies contends that this history of medicalization of disability in service of capitalist interests has given rise to the medical model of disability.

The Medical Model of Disability

The medical model is founded on the belief that behaviors and bodies can be objectively measured (Lewis, 2006). These measures then lead to the categorization and diagnosis of people's experiences and bodies. These diagnoses in turn enable helping professionals to provide treatments to assist disabled people to fit into a normalized able-bodied society. It is the assumption of a correct and normal body that supports the construction of disability as abnormality (Frazee, 2003). Dichotomies, such as normal versus abnormal, able-bodied versus disabled, and competent versus dependent, are the cornerstone of the medical model (MacKay and Rutherford, 2012). Thus if to be normal is to be independent, capable, and healthy, the medical model conceives of disability as the antithesis – dependent, diseased, broken, pitiable, incapable: both a problem of lack and excess (Goodley, 2011). This adherence to medicalizing human differences against a subjective norm is the basis of the medical model of disability. This deficit model of human variation presumes a problem and that problem *is* the individual. From this place of deficit, the person becomes a site for expert intervention as their bodies are placed in the care (or control) of those in the helping professions.

The medical model has a reach that extends beyond medical institutions and the medical profession. It forms the structure upon which many helping professions are developed. Social work practices that assume the medical model as their foundation for working with disabled people, to the exclusion of other models, then value or devalue people's bodies as they are measured against a predetermined standard. Those who do not meet those standards are understood as impaired and in need of clinical intervention.

How Can Disability Studies Challenge Normative Assumptions of Disabled Bodies?

However, the medical model is but one way to consider human variations and differences. Starting in the work of UK-based disabled activists (and social workers) a 'big' and 'emancipatory idea' emerged. The idea held that the deprivations of our lives were not the inevitable consequence of our various conditions of physical or mental impairment, but rather result from discriminatory social arrangements and arbitrary practices of exclusion (Goodley, 2010). An argument was formulated that disability is situated not in individual pathology, but in "society's failure to embrace diverse ways of being in the world – that disability is extrinsic rather than intrinsic" (Frazee, 2003: p. 260).

In the terms of 'on-the-ground' activism, to be disabled no longer means to be broken or damaged or deficient in some personal and unfortunate way. Instead, 'disabled' means simply "Not taken into account" (Frazee, 2003). From the beginning, not taken into account in the policies, programs, budgets, laws, curriculum, classroom design, and decisions that structure every conceivable site of human endeavor. Thus the problems faced by disabled people are not merely or predominantly clinical, but problems of living in a social order premised upon the fundamental assumption that everyone who matters engages with the world in standardized, normative manner. Through the emancipatory lens of disability studies, to be disabled is understood as having been marginalized and overlooked, except occasionally, always as an afterthought, and invariably with great fuss (Frazee, 2003).

In this way, disability studies seeks to denaturalize the marginalization of disabled people. It seeks to *depathologize*. Disability is no longer framed as a variant of human functionality but as the outcome of social barriers and deeply entrenched and tacit patterns of discrimination and exclusion (Goodley, 2010). Just as medical framings had always sought – and continue to seek – to eliminate disability through individual cure, rehabilitation or genetic manipulation, the social model seeks to eliminate disability through social change, political reform, and the 'rehabilitation' of community structures and environments (Oliver, 1990; Shakespeare, 2010).

Disability studies sustains these activist beginnings by examining how disability oppression is created. Moreover, disability studies recognizes that disability, impairment, and socioeconomic and cultural influences interact simultaneously to create an experience of disability-in-context rather than an ahistorical static disabled identity. "This framework stands in sharp contrast to the position taken by social work education regarding disability, where the emphasis is on rehabilitation, adaptation, and mitigation" (Hiranandani, 2005: para 11).

The Social Model

The social model of disability challenges the medical model's sole reliance on biology and individual fault. By considering disabled people in context, disability is transformed from one of individual problem to societal discriminations against impaired bodies (biofunctionality of the body). Where the medical model situates disability in the individual's impairment, the social model speaks to societal barriers that disable people from participating in their environments (Barnes and Mercer, 1997). Barriers can include access to buildings, cultural attitudes that presume incapacity for education and employment, and limited transportation and communication tools. In the words of the UK Union of the Physically Impaired Against Segregation (UPIAS), "it is society which disables physically impaired people. Disability is something imposed on top of our impairments by the way we are unnecessarily isolated and excluded from full participation in society" (cited in Thomas, 2004: p. 572). The manner in which social workers intervene in people's lives is dependent therefore on the model upon which they base their practice.

Interventions

Interventions are strategies and processes that change or influence a situation. Disabled people have been the target of a whole range of interventions. These do not have to be negative as some have facilitated full inclusion, community participation, and empowerment. However, others have produced new categories of impairment and disability, created new and demanding standards of normalcy, and have resulted in new forms of social exclusion. Thus interventions can range from the establishment of an independent living center to institutionalization in nursing homes for young disabled people, from a feminist psychotherapy session to pharmaceutical restraints, from a humans rights claim to 'excessive burden' immigration policies. The range of interventions allows each of us – whether we are service providers, advocates, community organizers, family members, or other disabled people – to enter the lives of another, and enact change. The impact of that change, on the other person, ourselves and the world around us depends on the politics informing the intervention. Disabilities Studies is about bringing those politics into the forefront.

What Does a Social Work Practice Not Grounded in a Politicized Disability Activism Look Like?

The following examples are true stories of real people, though each story is a composite of many people's experiences.

Example 1

Aging parents approach a social worker with concerns about their daughter, a young woman labeled with an intellectual disability. They are concerned with their daughters perceived inability to consent to sexual advances and would like to protect her. Typically, this common request has resulted in the disabled person's human rights being violated and the young girl being forced to undergo involuntary sterilization (Diekema, 2003). If the social worker has but one story of disability, it would be difficult for them to critically examine their behavior as one that perpetuates the oppression of the individual. The problem of consent to sexual advances is not solved by sterilization. The young girl may still not be given the opportunity to advocate for herself in situations where sexual advances may occur. The 'problem' that is solved is the one of an unwanted pregnancy, where it is the young girl's parents who do not want the pregnancy. By not examining dominant ideologies about disabled peoples as incapable, social workers can support interventions that perpetuate the exclusion of disabled peoples to participate and make decisions that impact their life.

Example 2

A parent labeled as having a history of depression was told to continue medication and attend therapies approved by the service provider in order to maintain custody of her children. There was no consideration of the repercussion of those medications and therapies. The approved therapy provided by the hospital was almost 2 h away by transit from the person's home and limited their ability to take on full-time work, in effect, disabling them from employment. The assumption is that depression is a medical issue that needs to be fixed. The disabled person must then prove their commitment to be 'not disabled' in order for service providers to allow the disabled person to parent their child. The prevailing notion being that a disabled person cannot be a parent.

Example 3

A mother involved with child welfare was told that in order to have her children returned to her, she had to 'get better.' She was prescribed various medications and told to participate in group therapy sessions. The medications and therapy sessions did little to alleviate her joblessness and poverty. The medications, however, did cause much confusion and sleepiness, disabling the mother from participating in the group therapies in the expected manner. As a result, her behavior was measured as noncompliant. Therefore, within a framework of deficit, that person's disability is most likely scapegoated as the 'problem.' Service providers' unquestioned adherence to a systemic medicalization of disability can also mean that the methods used to 'fix the person' are not questioned, and 'poor recovery' can further mark a person as untreatable, increasing the likelihood of their marginalization.

Example 4

When social workers write reports to child welfare agencies about a parent's abilities to care for a child, that report might be fraught with assumptions about what constitutes a 'good' parent. These do not include parenting as a person with a disability. The intersection of disability and parenting highlights the pervasiveness of a disabled person's body as deficient. When the foundation for service provision is based on ideologies that a person with a disability is not considered competent to provide care, as they are constructed as dependent on

receiving care themselves, it is not surprising that parents with disabilities are more likely to lose custody of their children (National Council on Disability, 2012).

Antioppressive Social Work and Disability Studies

> The philosophy of disability studies and disability rights movements are, in fact, similar to social work principles: all three arenas work for the equal rights, equal opportunities, social justice, self-respect and self-determination. Within a disability rights agenda, solutions to disablement revolve around "self-advocacy, system advocacy, elimination of barriers, and outcomes chosen by the person with disability".
>
> Pfeiffer as quoted in Hiranandani, 2005: para 28

Disability studies and the antioppressive practice (AOP) social work can inform each other in ways that liberate both service providers and service users from the limitations of a single idea of disability. The core of AOP social work practice is to promote social justice by recognizing how power is negotiated between individuals and institutions in ways that benefit some and disadvantage others. Disability studies can give us the tools to recognize and name the implicit bias embedded in our society and institutions that are predicated on racist, sexist, heteronormative, and ableist ideologies (Rushing, 2013). Ableism influences our behaviors in ways that reproduce and legitimize discriminatory practices. As social workers likely employed in institutions that normalize this reproduction of legitimizing discrimination, it is imperative that we implicate our biases and the work we do in potentially contributing to further disablement. Overboe (2007) states that "ableism, like racism and sexism, oppresses people by labeling them as having either a physical, psychiatric, cognitive, developmental, and sensory disability" (p. 23). Social workers must continually question how dominant discourses construct our understanding of disability and impairment.

Antioppressive social work practice highlights the ways in which our biases, if left unexamined, can contribute to silencing the voices of marginalized peoples. Critiques of antioppressive practices challenge the very notion of AOP as an academic and privileged knowledge that does not include the voices and experiences of those most affected by discriminatory practices (Wilson and Beresford, 2000). This makes evident the moral and ethical requirement for AOP social workers to expand our notions of who is an expert. Disability studies recognizes a multitude of voices as experts – specifically those experiences of people labeled and living with disabilities to form the core knowledges of disability. It is this expertise of people's lived experiences of disablement and the unique experiences of the interrelations of peoples many identities that can support social workers to gain new understandings, ultimately encouraging oppressive practices.

Disability studies creates uncomfortable moments as we reflect upon our roles as social workers. Yet it is this very discomfort that can propel us to make change. Social workers need to first examine their own attitudes toward disability and how these biases have determined the way in which they construct or challenge 'a problem.' What do disability and impairment mean to us? What are the implicit biases and/or overt prejudice surrounding disability that impact how we negotiate interventions in people's lives? Do we recognize how disablement interweaves with people's diverse identities to create unique experiences of power and marginalization? Even if no one answer is apparent, the uncertainty that accompanies our self-questioning can support us as social workers to pause, rather than routinely respond to a problem. A politicized understanding of disability can support social workers to explore the possibilities of alternate realities of human differences to liberate ourselves from our 'fear or disability' to work with disabled peoples.

See also: Children and Families in Social Work; Critical Reflexivity and Social Work Practice; Human Rights and Social Work; Immigration, Citizenship, and Integration: Social Work Connections; Learning Disability and Social Work; Mental Health and Social Work; Mental Health, Mentalism and Sanism; Narrative in Social Work; Reflexivity.

Bibliography

Barnes, C., 2010. A brief history of discrimination and disabled people. In: Davis, L.J. (Ed.), The Disability Studies Reader. Routledge, NY, pp. 20–32.
Barnes, C., 2012. The social model of disability: Valuable or irrelevant. In: The Routledge Handbook of Disability Studies, pp. 12–29.
Barnes, C., Mercer, G., 1997. Breaking the mould? an introduction to doing disability research. Doing Disability Research 1, 1–4.
Barnes, C., Mercer, G., Shakespeare, T., 1999. Exploring Disability: A Sociological Introduction. Polity Press, Cambridge, 182–210.
Campbell, F.K., 2008. Refusing able (ness): a preliminary conversation about ableism. M/C Journal 11 (3).
Chenoweth, L., 1996. Violence and women with disabilities silence and paradox. Violence against Women 2 (4), 391–411.
Diekema, D.S., 2003. Involuntary sterilization of persons with mental retardation: an ethical analysis. Mental Retardation and Developmental Disabilities Research Reviews 9 (1), 21–26.
Dolmage, J., Lewiecki-Wilson, C., 2010. Refiguring rhetorica: linking feminist rhetoric and disability studies. In: Schell, E.E., Rawson, K. (Eds.), Rhetorica in Motion: Feminist Rhetorical Methods and Methodologies. U. Pittsburgh Press, Pittsburgh, pp. 23–38.
Frazee, C., 2003. Disability studies: the unexpected guest in health law discourse. Health, Law Journal 257–262.
Goodley, D., 2010. Disability Studies: An Interdisciplinary Introduction. Sage, London.
Goodley, D., 2011. Social psychoanalytic disability studies. Disability and Society 26 (6), 715–728.
Goodley, D., 2014. Dis/ability Studies: Theorising Disablism and Ableism. Routledge, NY.
Harpur, P., 2012. Embracing the new disability rights paradigm: the importance of the convention on the rights of persons with disabilities. Disability and Society 27 (1), 1–14.
Hiranandani, V., 2005. Towards a critical theory of disability in social work. Critical Social Work 6 (1), 1–6.
Lewis, B., 2006. A mad fight: psychiatry and disability activism. In: David, L. (Ed.), The Disability Studies Reader. Routledge, New York, pp. 3–16.
Linton, S., 1998. Claiming Disability: Knowledge and Identity. NYU Press.
MacKay, J., Rutherford, A., 2012. Feminist women's accounts of depression. Affilia: Journal of Women and Social Work 27 (2), 180–189.
Mackelprang, R.W., Salsgiver, R.O., 1996. People with disabilities and social work: historical and contemporary issues. Social Work 41 (1), 7–14.
Mallett, R., Runswick-Cole, K., 2014. Approaching Disability: Critical Issues and Perspectives. Routledge, NY.
McRuer, R., 2006. Crip Theory: Cultural Signs of Queerness and Disability. NYU press.
Monaghan, P., 1998. Pioneering field of disability studies challenges established approaches and attitudes. Chronicle of Higher Education 44 (20).
Nathan, J., Webber, M., 2010. Mental health social work and the bureau-medicalisation of mental health care: identity in a changing world. Journal of Social Work Practice 24 (1), 15–28.

National Council on Disability, 2012. Rocking the Cradle: Ensuring the Rights of Parents with, Disabilities and Their Children. Retrieved from: http://www.ncd.gov/publications/2012/Sep272012/.
Oliver, M., 1990. The Politics of Disablement: Critical Texts in Social Work and the Welfare State. Mac-Millan, London.
Oliver, M., 2009. The social model in context. In: Titchkosky, T., Michalko, R. (Eds.), Rethinking Normalcy: A Disability Studies Reader. Canadian Scholars' Press Inc, Toronto, pp. 19–30.
Overboe, J., 2007. Vitalism: subjectivity exceeding racism, sexism and (psychiatric) ableism. Wagadu 4.
Peters, Y., 2008. Twenty Years of Litigating for Disability Equality Rights: Has It Made a Difference? An Assessment by the Council of Canadians with Disabilities. Retrieved from: http://www.ccdonline.ca/publications/20yrs/20yrs.htm#IV.
Profitt, N.J., 2010. In the Best Interests of Women and Children: Exploring the Issue of, "Failure to Protect". Muriel McQueen Fergusson Center for Family Violence Research.
Rosenfeld, S., Phillips, J., White, H., 2006. Gender, race, and the self in mental health and, crime. Social Problems 53 (2), 161–185.
Rushing, J., July 23, 2013. The Stream: Are We All Racist? (Television broadcast) AlJazeera. Retrieved from: http://stream.aljazeera.com/story/201307230046-0022926.
Schormans, A. Fudge, 2013. 'Weightless?': disrupting relations of power in/through photographic imagery of persons with intellectual disabilities. Disability and Society 1–15 (ahead-of-print).
Shakespeare, T., 2010. The social model of disability. In: Davis, L.J. (Ed.), The Disability Studies Reader. Routledge, NY, pp. 266–273.
Skeggs, B., 2014. Values beyond value? Is anything beyond the logic of capital? The British Journal of Sociology 65 (1), 1–20.
Society for Disability Studies, n.d. About SDS. Retrieved from: https://www.disstudies.org/.
Stevens, B.J.D., Saxton, M., Johnson, V.A., Kamga, S.A.D., What's, A., Walton, D.R., Akram, A., 2011. Online Series of New Barbara Waxman Fiduccia Papers, 'On Women and Girls with Disabilities' 2011, Prevention.
Thomas, C., 2004. How is disability understood? an examination of sociological approaches. Disability and Society 19 (6), 569–653.
Titchkosky, T., 2003. Disability, Self, and Society. University of Toronto Press.
Tregaskis, C., 2002. Social model theory: the story so far. Disability and Society 17 (4), 457–470.
Wilson, A., Beresford, P., 2000. 'Anti-oppressive practice': emancipation or appropriation? British Journal of Social Work 30 (5), 553–573.
World Health Organization, 2011. World Report on Disability. Retrieved from: http://www.who.int/disabilities/world_report/2011/en/.

Relevant Websites

http://badcripple.blogspot.ca/ – Bad Cripple.
http://www.ccdonline.ca/en/ – Council of Canadians with Disabilities.
http://disability-studies.leeds.ac.uk/ – Centre for Disability Studies Leeds.
http://diaryofabenefitscrounger.blogspot.ca/ – Diary of a Benefit Scrounger.
http://www.dpi.org/ – Disabled People's International.
Follow @MiaMingus on Twitter.
http://inclusion-international.org/ – Inclusion International.
http://leavingevidence.wordpress.com/ – Leaving Evidence.
http://nndr.org/ – Nordic Network on Disability Research.
http://sisofrida.org/about/ – Sisters of Frida.
http://www.un.org/disabilities/default.asp?id=150 – UN Enable.

Disability, Geography of

Vera Chouinard, McMaster University, Hamilton, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The causes and consequences of disability has become an established area of critical geographic inquiry. It emerged in response to changing social conditions (e.g., a vocal disability movement), perspectives on disability (e.g., disability as the outcome of socially constructed barriers to inclusion), and changes in the discipline (e.g., recognition of multiple bases of oppression). The second wave of geographic research on disability has widened the range of differences considered to have disabling consequences (e.g., negotiating places as a 'fat' woman). It has also broadened the range of spaces considered (e.g., to include online communities).

Particularly from the 1990s onward, the sociospatial causes and consequences of disability (i.e., of being socially constructed as negatively different and materially disadvantaged on the basis of mental or physical bodily impairment) has become an established area of critical geographic inquiry. These studies treat disability not as a given consequence of bodily impairment, but as a sociospatial process of marking impaired bodies as negatively 'other' and, on the basis of this negative difference, marginalizing and excluding disabled people from spaces of everyday life. Critical geographies of disability are thus as much about ableism as a regime of power based on able-bodied privilege and the sociospatial practices that sustain it as about disabled persons' experiences of trying to negotiate environments that physically and/or socially exclude and oppress them. While this article focuses on research concerned with disability in the critical social geography tradition, it is important to acknowledge that there have been significant contributions to understanding the geography of disability informed by other traditions of inquiry. Perhaps most noteworthy are Golledge's contributions to understanding visual impairment and wayfinding from a behavioral geography perspective (Golledge, 1993; Loomis et al., 1998).

The development of critical geographies of disability is best understood as a response to both changing conditions within society, and new perspectives on the differences disabilities make in how people are situated (physically, economically, politically, and culturally) within prevailing social relations of power in specific localities, cities, and regions. The former include an increasingly vocal and visible disability rights movement at national and global scales, aging populations with increased vulnerability to and experiences of disability, the restructuring and downsizing of social assistance to vulnerable groups, which has exacerbated and drawn attention to the grinding poverty that disabled people face. New perspectives on disability have also encouraged critical studies of disability and space. The growth of disability studies has drawn attention to a wide range of disability issues, challenged conventional biomedical models of disability with 'social constructionist' alternatives, and offered significant glimpses of what it is like to be disabled in today's societies and spaces of everyday life. Within geography, the growing influence of critical social theory, feminist geographic research highlighting how social differences become bases of spatial and social exclusions, recognition of multiple bases of oppression in society, and concerns to use geographic knowledge as a basis for empowering disadvantaged groups have helped to encourage critical approaches to disability and space.

The sociospatial causes and consequences of disability, that is, of being socially constructed as negatively different on the basis of mental or physical bodily impairment and having to contend with physical and social barriers to inclusion, has become a significant focus of geographic inquiry. Geographers working in this area have helped to demonstrate that scholarly tendencies to assume that only able-bodied women, men, and children occupy and negotiate social space have impoverished our understanding of society and space. Differences in mental and physical abilities affect patterns and degrees of access to urban public spaces, services, and employment opportunities, and help to determine who is most affected by urban and regional changes such as rising poverty and loss of affordable housing. Differences in bodily ability find expression in the geography of cities and regions: in the spatial concentration of disabled persons in poor inner city neighborhoods with a concentration of social services, for instance.

Why a Critical Geography of Disability?

Critical geographies of disability have emerged as a response to changing conditions within society, and new perspectives on the differences disabilities make in how people are situated in society and space (e.g., in terms of access to basic necessities and acceptance in urban spaces). An increasingly vocal disability rights movement, particularly in the West, but also internationally, aging populations with higher incidence of disability, cutbacks in social programs for the disabled, and problems such as grinding poverty and inaccessible cities have helped to encourage such research. New perspectives on disability issues have also been influential. The growth of disability studies has challenged conventional biomedical models of disability with 'social constructionist' ones, and offered significant insights into what it is like to be disabled in various spaces of everyday life (Oliver, 1990; Morris, 1991). These society-centered perspectives view disability as the outcome of social processes such as the economic devaluation of disabled individuals through lack of employment and low wages, and cultural representations of disabled people as inferior. Disability is thus a social rather than individual

condition. This does not necessarily mean that disablement is entirely socially produced. Bodily impairments affect experiences of disablement. Visual impairment may result in difficulty moving within spaces such as sports arenas and negative reactions to such visible differences may help to make disabled persons feel 'out of place.' Similarly, chronic pain sufferers often find it difficult to negotiate spaces of daily life and interact with others. Scholars continue to debate how best to incorporate such bodily realities within conceptualizations of social processes of disablement.

Changes in the discipline have also encouraged critical geographies of disability. These include greater use of critical social theory, feminist geographic studies of a wider range of social differences (e.g., taking sexuality or race into account as well as gender) and how these contribute to marginalization in urban space, recognition of multiple bases of oppression in society and space, and desires to use geographic knowledge to help empower disadvantaged groups. Critical geographers studying disability have promoted related changes in the discipline: such as making conferences more accessible and promoting research through an international network of scholars (i.e., Disability and Geography International). They have challenged the invisibility of disabled women, men, and children in the geographic literature; emphasizing disablement as a sociospatial form of oppression, exclusionary practices within spaces of academia, and the need to put gendered experiences of disability on geography's research map (Chouinard and Grant, 1995).

What do critical geographies of disability offer? They offer the possibility of geographic research that empowers persons with disabilities (building on the legacy of radical geography). They promise more balanced accounts of the sociospatial production of disability as a place-specific condition of life, not a natural outcome of bodily impairment. They can help us understand how ableist social relations and practices are sustained over space and time, and how more enabling environments might be created. They can shed light on why resistance to ableism varies among places of the significance, for example, of national differences in human rights law for disability activism. By providing critical causal explanations of the sociospatial production of disabling environments and of resistance to disablement, these geographies can help us create more enabling societies and spaces.

Research Topics and Directions

It is generally agreed that the first major 'wave' of critical geographic research concerned with disability emerged in the 1990s. Much of this work was captured in Ruth Butler and Hester Parr's (1999) edited collection entitled *Mind and Body Spaces: Geographies of Illness, Impairment and Disability*. Topics covered in this collection included the role of ableist conceptions of the body in architectural discourse and practice (Imrie, 1999), how the onset of physical impairment and heightened sociospatial isolation in the feminized sphere of the home threatens and complicates masculine identities (Valentine, 1999), barriers faced by lesbian, gay, and bisexual disabled people in the gay community (Butler, 1999), and making disabled women's activism and their barriers visible in Canada and beyond (Chouinard, 1999). Among the other noteworthy contributions to a critical geography of disability during this era were Isabel Dyck's (1995) work on the shrinking lifeworlds of women with multiple sclerosis and Brendan Gleeson's (1999) *Geographies of Disability*, a book, which provided a historical materialist analysis that linked changing conditions of disabled people's lives to transformations in production, labor, and ways of life under capitalism. Underlying much of this research was a concern with how and why inequalities of power give rise to social spaces that disadvantage and exclude the disabled persons. This critical stance toward past and present social orders continues to inform more recent critical geographic accounts of disabled people's lives and life spaces, such as Wilton's (2004) analysis of changes in neoliberal state policy in Ontario and how deepening poverty constrains the lives of those with mental ill health. It is, however, perhaps a less singular and unifying concern than in the past.

Since 2000, what is sometimes referred to as a 'second wave' of critical geographies of disability has emerged (Chouinard et al., 2010). This has been characterized, in part, by a more expansive understanding of embodied differences that have 'disabling' consequences, such as Longhurst's (2010) work on experiences of being 'fat' and Holt's (2010) work on how young people with socioemotional differences are (dis)abled within school spaces. It has also been characterized by interest in how embodied differences, such as being chronically ill, alter the meaning of everyday spaces such as the home (Crooks, 2010) and the forging of identities and support in online spaces amongst persons with autism, hearing impairment, and mental ill health (Davidson and Parr, 2010; Skelton and Valentine, 2010). In a related vein, Parr (2008) explores issues of identity, community, citizenship, and inclusion amongst those with mental health differences in rural communities, urban gardens, artistic spaces, and on the Internet. Such work has helped to broaden the range of everyday life spaces considered in geographic research concerned with disability.

Concerns with embodied life, emotions, and experiences of the self-in-place have encouraged some geographers to explore how conditions such as phobia and bipolar disorder alter people's ways of being in and negotiating place (Smith and Davidson, 2006; Chouinard, 2012a). Such work is important in illuminating diversity in embodied life, providing insight into distinctive 'lifeworlds' and support needs, and potentially promoting more inclusive ways of knowing conditions such as 'madness.' Others who share such concerns have explored topics such as how being disabled complicates performances of emotional labor in the workplace (Wilton, 2008). Attention to issues of embodiment, disclosure, and passing is also evident in recent geographic research (Davidson and Henderson, 2010). Critical geographies of care and care giving in relation to groups such as people with developmental differences and the frail elderly are also important in the latest wave of critical geographic research. While interest in this area was apparent in the first wave (Stables and Smith, 1999), this area of inquiry has become more conceptually and substantively sophisticated. So, for example, Milligan (2005) examines emotions involved in informally providing care to the elderly in the home as well as in the transition to

residential care. Metzel (2010) demonstrates the emotional stresses and strains amongst siblings trying to oversee and help provide care to siblings with developmental and intellectual disabilities who live long distances away. She notes the ongoing challenges of trying to monitor the adequacy of care from afar and of managing feelings such as guilt at often not being present in the place where their sibling lives. Power (2010) considers the ways in which contemporary forms of care for persons with developmental disabilities can heighten their dependence and calls for greater recognition of the complex networks of interdependency shaping their lives.

In short, critical geographies of disability, building on the legacies of first wave research, have blossomed into an established, vibrant, and innovative area of inquiry in social geography.

Future Challenges and Directions

Where should geographies of disability go from here? One important challenge, given current emphases on micro- and mesoscale processes of disabled people's embodiment, identity formation, and assignment of meanings to places, is to further advance our understanding of the bigger geographic picture – namely how the uneven development and dynamics of global, neoliberal, and patriarchal capitalism perpetuates social and spatial exclusions of disabled women, men, and children. Systematic studies of how sociospatial restructuring of multilateral and state institutions and programs shape the geographic boundaries and conditions of disabled people's lives in different places would help to address this challenge. Another area for future research is to examine geographic variations in the design, production, availability, and consumption of disability aids, and their impacts on disabled people's abilities to negotiate life spaces. Studies of the sociospatial production of impairment and disabling conditions of life; examining, for instance, how unsafe, stressful conditions of work and the hyperexploitation of female labor in places such as export zones influence the incidence and severity of different disabilities would also help to address this challenge.

A related challenge in advancing our understanding of processes of disablement within the context of a global capitalist order is to, literally and figuratively, push the boundaries of geographies of disability. To date, geographers have overwhelmingly focused their attention on disabled people's lives in affluent nations and have neglected experiences of disability in developing countries of the global South. This is unfortunate given the fact that the vast majority of disabled people live in these countries as it means that geographic knowledge about disability is based on the experiences of a minority of disabled women, men, and children. Chouinard's (2012b) work on disability in the developing nation of Guyana, which places disabled people's material conditions of life and experiences of being disabled in the context of an unequal global capitalist order, is one exception. Dyck (2010), reflecting on new 'body optimization' techniques (e.g., genetic screening) and their implications for geographies of disability, illness, and impairment, points to opportunities for future research to explore issues of power and distributive justice at scales ranging from the personal to the global. So, for example, geographers could explore the power relations and distributive justice issues in the global organ trade. Such work could include addressing the uneven global geography of organ provision and consumption, examining how people are persuaded or coerced to provide organs for sale, and the impacts of organ provision on local communities and on the health of organ providers and consumers.

The second wave of geographies of disability has seen increased attention to cultural practices, images and identities, and their implications for disabled people's lives (e.g., work on how diverse disabled people forge communities and identities through interaction in online space). However, much remains to be done. There is considerable scope for geographers to explore the sociocultural implications of engagement in spaces of artistic endeavor for diverse disabled people. This can include, as Parr's (2008) work with mental health consumers indicates, using artistic projects as a methodological strategy to allow disabled participants to give voice to what they want others to know about their communities, lives, and identities, and to help promote confidence, a sense of being capable and of belonging. Similarly, much of the work remains to be done regarding how popular media, such as commercial films, represent diverse disabled people and the places where they do and do not belong (but see Chouinard's work (2009) on *Girl Interrupted*). Such cultural geographies of disability and the media might include examining how different audiences react to portrayals of disabled people and the places where they belong. Another avenue for research is that of political cultures of place and how these shape disabled people's lives. Are people in urban neighborhoods with a history of disability activism less likely to view disabled people as unproductive citizens who are out of place in local labor markets and communities? Why do activists in some places adopt radical, confrontational strategies demanding collective rights, while others favor actions such as sponsoring stage appearances by famous disabled 'heroes'; events that reinforce cultural notions that only exceptional disabled persons succeed? There is also the challenge, as Kitchin and Wilton (2004) illustrate, of understanding the political and cultural practices and material conditions of activism that allow some disability activists and organizations to 'jump scale' and effect change while others do not.

Theoretical challenges include continuing to develop dynamic, deeply geographic conceptions of the regulation of bodily differences in space; conceptions, that allow for multiple, diverse bases and processes of oppression as disabled women, men, and children move through space and time. Such frameworks could draw, in part, on feminist notions of intersectionality as Valentine (2007) has demonstrated. And they would need to be attentive to the multiscalar nature of these processes – ranging from the intrapersonal scale at which disabled persons grapple with the mental and emotional consequences of marginalization and exclusion through to the global scale where the uneven development of capitalist societies helps to produce oppressive conditions of life for disabled persons.

Methodological challenges include gathering data on neglected aspects of disability such as spatial variations in employment and accommodation practices in workplaces, and developing accessible research formats. Ethical and political

challenges include ensuring that research empowers disabled persons.

Perhaps the key challenge is promoting restless geographic imaginations about disabling environments. Such imaginations will be open to surprise about patterns and processes of disablement over space and time, and willing to learn from disabled people's knowledge. With continued creativity and commitment, future geographies of disability will promote environments that enable all members of society.

Bibliography

Butler, R., 1999. Double the trouble or twice the fun? Disabled bodies in the gay community. In: Butler, R., Parr, H. (Eds.), Mind and Body Space: Geographies of Illness, Impairment, and Disability. Routledge, London, pp. 203–220.

Butler, R., Parr, H. (Eds.), 1999. Mind and Body Space: Geographies of Illness, Impairment, and Disability. Routledge, London.

Chouinard, V., 1999. Body politics: disabled women's activism in Canada and beyond. In: Butler, R., Parr, H. (Eds.), Mind and Body Space: Geographies of Illness, Impairment, and Disability. Routledge, London, pp. 269–294.

Chouinard, V., 2009. Placing the 'mad woman': troubling socio-spatial representations of being a woman with mental illness in "Girl Interrupted." Social and Cultural Geography 10 (7), 791–804.

Chouinard, V., 2012a. Mapping bipolar worlds: negotiating geographies of 'madness' in autobiographical accounts. Health & Place 18 (2), 144–151.

Chouinard, V., 2012b. Pushing the boundaries of our understanding of disability and violence: voices from the Global South (Guyana). Disability & Society 27 (6), 777–792.

Chouinard, V., Grant, A., 1995. On being not even anywhere near the project: ways of putting ourselves in the picture. Antipode 27, 137–166.

Chouinard, V., Hall, E., Wilton, R., 2010. Introduction: towards enabling geographies. In: Chouinard, V., Hall, E., Wilton, R. (Eds.), Towards Enabling Geographies: 'Disabled' Bodies and Minds in Society and Space. Ashgate, pp. 1–22.

Crooks, V.A., 2010. Women's changing experiences of the home and life inside it after becoming chronically ill. In: Chouinard, V., Hall, E., Wilton, R. (Eds.), Towards Enabling Geographies: 'Disabled' Bodies and Minds in Society and Space. Ashgate, pp. 45–62.

Davidson, J., Henderson, V.L., 2010. 'Coming out' on the spectrum: autism, identity and disclosure. Social & Cultural Geography 11 (2), 115–170.

Davidson, J., Parr, H., 2010. Enabling cultures of disorder online. In: Chouinard, V., Hall, E., Wilton, R. (Eds.), Towards Enabling Geographies: 'Disabled' Bodies and Minds in Society and Space. Ashgate, pp. 63–84.

Dyck, I., 2010. Geographies of disability: reflections on new body knowledges. In: Chouinard, V., Hall, E., Wilton, R. (Eds.), Towards Enabling Geographies: 'Disabled' Bodies and Minds in Society and Space. Ashgate, pp. 253–264.

Dyck, I., 1995. Hidden geographies: the changing life-worlds of women with multiple sclerosis. Social Science and Medicine 40, 307–320.

Gleeson, B., 1999. Geographies of Disability. Routledge, London and New York.

Golledge, R.G., 1993. Geography and the disabled: a survey with special reference to vision-impaired and blind populations. Transactions of the Institute of British Geographers 18 (1), 63–85.

Holt, L., 2010. Young people with socioemotional differences: theorising disability and destabilising socio-emotional Norms. In: Chouinard, V., Hall, E., Wilton, R. (Eds.), Towards Enabling Geographies: 'Disabled' Bodies and Minds in Society and Space. Ashgate, pp. 145–164.

Imrie, R., 1999. The body, disability and Le Corbusier's conception of the radiant environment. In: Butler, R., Parr, H. (Eds.), Mind and Body Spaces: Geographies of Illness, Impairment and Disability. Routledge, London and New York, pp. 25–45.

Kitchin, R., Wilton, R., 2003. Disability activism and the politics of scale. The Canadian Geographer 47, 97–115.

Longhurst, R., 2010. The disabling affects of fat: the emotional and material geographies of some women who live in Hamilton, New Zealand. In: Chouinard, V., Hall, E., Wilton, R. (Eds.), Towards Enabling Geographies: 'Disabled' Bodies and Minds in Society and Space. Ashgate, pp. 199–216.

Loomis, J.M., Golledge, R.G., Klatsky, R.I., 1998. Navigation system for the blind: auditory display modes and guidance. Presence 7 (2), 193–203.

Milligan, C., 2005. From home to 'home': situating emotions within the caregiving experience. Environment and Planning A 37, 2105–2220.

Metzel, D., 2010. 563 Miles: a matter of distance in long-distance caring by siblings of siblings with intellectual and developmental disabilities. In: Chouinard, V., Hall, E., Wilton, R. (Eds.), Towards Enabling Geographies: 'Disabled' Bodies and Minds in Society and Space. Ashgate, pp. 123–144.

Morris, J., 1991. Pride Against Prejudice: Transforming Attitudes to Disability. Women's Press, London.

Oliver, M., 1990. The Politics of Disablement. Macmillan, London.

Parr, H., 2008. Mental Health and Social Space: Towards Inclusionary Geographies? Blackwell Publishing.

Power, A., 2010. The geographies of interdependence in the lives of people with intellectual disabilities. In: Chouinard, V., Hall, E., Wilton, R. (Eds.), Towards Enabling Geographies: 'Disabled' Bodies and Minds in Society and Space. Ashgate, pp. 107–122.

Skelton, T., Valentine, G., 2010. It's my umbilical cord to the world…the Internet: D/deaf and hard of hearing people's information and communication practices. In: Chouinard, V., Hall, E., Wilton, R. (Eds.), Towards Enabling Geographies: 'Disabled' Bodies and Minds in Society and Space. Ashgate, pp. 85–106.

Smith, M., Davidson, J., 2006. 'It makes my skin crawl': the embodiment of disgust in phobias of 'nature'. Body & Society 12 (1), 43–67.

Stables, J., Smith, F., 1999. 'Caught in the Cinderella trap': narratives of disabled parents and young carers. In: Butler, R., Parr, H. (Eds.), Mind and Body Spaces: Geographies of Illness, Impairment and Disability. Routledge, London and New York, pp. 256–268.

Valentine, G., 1999. What it means to be a man: the body, masculinities, disability. In: Butler, R., Parr, H. (Eds.), Mind and Body Spaces: Geographies of Illness, Impairment and Disability. Routledge, London, pp. 167–180.

Valentine, G., 2007. Theorizing and researching intersectionality: a challenge for feminist geography. The Professional Geographer 59 (1), 10–21.

Wilton, R.D., 2004. More responsibility, less control: psychiatric survivors and welfare state restructuring. Disability & Society 19 (4), 371–385.

Wilton, R.D., 2008. Workers with disabilities and the challenges of emotional labour. Disability & Society 23 (4), 361–373.

Disability Movement

Debbie Jolly, Disabled People against Cuts (DPAC), UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The disability movement is a movement of disabled people for human rights and independent living. Independent living does not mean people doing everything for themselves, but having the support to live in the same way as nondisabled people, with disabled people making their own choices about where and how they live. The article highlights key issues in the independent living movement as well as some of the key players, it ends by focusing on the situation today and why the independent living movement must become stronger in continuing to fight for the human rights of all disabled people.

A Summary of the Beginnings of the Independent Living Movement

The independent living movement philosophy was active in the United States in the 1960s and 1970s with the first well-known Center for Independent Living being established at University of California, Berkeley, in 1972 by Ed Roberts. People in other countries were also active during that time. For example, disabled people in Finland established the first independent living program in 1973, encompassing six Centers for Independent Living (CILs). In 1976, the Union of the Physically Impaired Against Segregation published a manifesto in England in relation to the institutionalization of disabled people and their lack of rights (UPIAS, 1976). This was to be the basis for the social model of disability, a UK model popularized by Finklestein (1980), Oliver (1990), Barnes (1991), and Oliver and Barnes (1998). In the same year, a group of disabled people living in a residential home in Zimbabwe organized themselves for advocacy (Behrent et al., 1999). In 1978, Japanese disabled people organized radical self-advocacy and Switzerland held one of the first conferences of self-help for disabled people (Behrent et al., 1999).

In the early 1980s, the first Canadian CILs were founded, and the first organization of disabled people was created in Nicaragua to provide independent living services and advocacy. In 1981, the British Council of Disabled People was formed (now renamed the UK Disabled People's Council), whilst in Germany a disability rights coalition staged a mock tribunal putting the country on trial for the abuse and segregation of disabled people (Behrent et al., 1999). In the same year, Project '81 initiated the earliest form of direct payments funded by a local authority in the United Kingdom. It enabled disabled people to employ personal assistants (PAs) and was operated by what would later become the Hampshire Center for Independent Living. Campbell and Oliver (1996) note that the disabled peoples' movement in Britain is: "is a movement which incorporates both revolutionary and reformist politics" (p. 179).

CILs and Other Organizations of Disabled People

The independent living movement is a movement of disabled people which campaigns for the full rights of disabled people and is the basis on which all CILs, Centers for Integrated Living, and Centers for Independent Living or CILs have been founded.

CILs are organized and run by disabled people to offer peer support and other services to disabled people. They also engage in work at local, national levels, and sometimes at the European levels to raise awareness and challenge issues that relate to the rights of disabled people.

However, an organization can be about independent living without being called a CIL because what many groups and organizations share is a commitment to the principles of the social model and the aims of independent living. What is important is that they are run and controlled by disabled people. These groups and organizations are called organizations 'of' disabled people in that they have staff, boards, and management committees comprising disabled people. This differs from organizations 'for' disabled people who may not have any or very few disabled staff, with few or none making decisions at board level (Jolly, 2009).

At the same time, organizations 'of' disabled people may not meet the aims of the independent living movement or effectively promote disabled peoples' rights. This occurs when such organizations are merely acting as the passive friend of governments without challenging the barriers disabled people face (Mladenov 2009). It is also unhelpful if organizations work with a charity model that views disabled people as needing 'help' or pity. Such organizations along with many disability charities are likely to promote disabled people lives through a dehumanizing tragedy model of disability to obtain donations for their charities. They effectively take away any responsibility from the state to provide effective policies and disability rights that include disabled people in society. The saying 'Rights not Charity' sums up disabled activists' attitudes to large disability charities who provide residential institutions and are believed to keep disabled people as passive recipients of charity rather than active citizens organizing together for their own rights, self-determination, and further progress for independent living (Jolly, 2009).

The Disability Movement

The disability movement was formed on the premise of independent living for disabled people. Independent living has to do with self-determination and autonomy, the right and

opportunity to engage in life and citizenship as we, as disabled people, chose (The author is a disabled person and identifies politically as a disabled person). Above all, it means the right and duty of the individual to choose, pursue, and manage his or her own way of life. The term 'independent living' is a concept, which underpins the work of disabled people to achieve equal opportunities and full participation in society as equal citizens. The philosophy of independent living argues that disabled people must have equal civil rights and that freedom of choice and the possibility of managing every aspect of life must be ensured. The disability movement represents pan-disability issues. Pan-disability does not refer to single impairment issues, but to disability as social oppression for all people with impairments. Single impairment issues can divide disabled people to the detriment of a movement attempting to gain human rights for all disabled people; inclusivity over such rights is more effective in meeting the aim of independent living and disability rights for all.

It is widely accepted that disabled people have fewer opportunities to participate in the activities of everyday life due to a number of social, environmental, and attitudinal barriers. For example, inaccessible buildings present barriers to participation in leisure and political activities, barriers to inclusive education such as buildings and attitudes prevent education and participation in educational establishments, the same is true of transport and housing: disabled people do not have the same freedom of movement, for example, if public transport is inaccessible, if public housing is not accessible or adapted they have less chances to move from place to place compared to nondisabled people. While some barriers are overt, others can be less so. Instances of these less overt forms can be found in various situations, including within the social fabric of the job interview. For example, in Jolly (2003), Wilde is quoted describing a job interview at a major UK university in which the interview process was described as disabling. Issues included having to walk long distances on a tour of the university, having to wait 6 h for the interview itself as well as a stand-up buffet lunch. When Wilde, who had an invisible impairment was asked why she chose to identify as disabled, it was assumed that this was through an allegiance to disabled people and not because she was disabled herself. Even though she offered to explain this, the interviewers appeared perplexed by her response and terminated the interview. The understanding of disability barriers by this particular university was restricted to physical access needs of wheelchair users. As a result, they were unable to comprehend other barriers to the interview process. The barriers approach is an international one in analyzing societal and attitudinal barriers, which disadvantage disabled people. See Cardona (2013), Shier et al. (2009), Maart et al. (2007), and Moriña Díez (2010) for further examples.

The social model of disability identifies the philosophy and actions that need to be taken to begin to remove the multiple social disadvantages that exist today for disabled people which prevent them from living with the full choices, rights, and control that nondisabled people take for granted. The philosophy of the social model of disability underpins the aims of the independent living movement. The international disability movement lays claim to civil rights and the politicization of disabled people in pursuit of their human rights and full and equal participation (Oliver, 1990; Oliver and Barnes, 2012). This basis of the 'needs to rights agenda' was summed up by disabled people from the Derbyshire Coalition of Disabled People in the United Kingdom.

Seven Needs of Independent Living

The aims of independent living can be succinctly summed up in the seven needs of independent living for disabled people first developed by the Derbyshire Coalition of Disabled People in the United Kingdom in the 1980s.

- Information
- Peer support
- Housing
- Technical aids and equipment
- Personal assistants
- Transport
- Access

While needs and rights may not appear to fit together, the principle is that if certain basic needs are met, this will set the path for better disability rights, because disabled people will be able to fight inequality from a more even playing field of inclusion. The importance of other 'needs' such as education, work, and leisure were recognized, but it was argued that these could be met by mainstream services once the seven primary needs were met (Barnes et al., 1999). In reality, disabled people in many countries are still fighting for these seven needs, as well as inclusive education, mainstream employment opportunities, and access to leisure. This does not mean that there has been no progress, but that progress as a whole differs by country, is variable and affected by different historical and economic circumstance in each country over time.

European Network for Independent Living

In April 1989, the first European Independent Living Conference dedicated to personal assistance was held at the European Parliament in Strasbourg, France. The conference resulted in the founding of the European Network for Independent Living (ENIL). Disabled people from the Netherlands, United Kingdom, Denmark, Italy, Sweden, France, Austria, Finland, Belgium, United States, Hungry, Federal Republic of Germany, and Norway came together at the European Parliament in Strasbourg, France to focus on personal assistant (PA) schemes as an essential factor of independent living with the understanding that independent living encompasses the whole area of activities: housing, transport, access, education, employment, economic security, and political influence. Their statement began: "We, disabled people, recognizing our unique expertise, derived from our experience, must take the initiative in the planning of policies that directly affect us." It went on to condemn segregation and institutionalization, which were described as 'a direct violation of our human rights,' demanding legislation protecting the human rights of disabled people including equal opportunities. Specifically they argued that a key prerequisite to full human and civil rights through independent living is the provision of support for PA services for those that need them. A further prerequisite was that

disabled people should have the opportunity to develop these services for themselves. That is, they should be user led by disabled people themselves.

The participants at this conference called on governments and policy makers to enforce the principles of access to PA services as a human and civil right. Moreover, they demanded that these services would serve all disabled people, of all ages, irrespective of personal wealth, income, or marital and family status. PA services were to be available for anything up to 24 h a day, 7 days a week, for those who required them. Additionally, these were to be paid for from public funds to stop the incarceration of disabled people in closed institutions. A further call was that disabled people and their user-led organizations should be involved in all aspects of policy, planning, implementation, and development. From the initial meeting there has been a mixture of stagnation, progress, and regression on the issues of PA schemes in different countries (Source: http://www.independentliving.org/docs2/enilstrasbourgresolutions.html accessed 20.06.13.).

The International Disability Movement: Some Organizations 'of' Disabled People

Disabled Peoples' International (DPI) (1981–). Disabled people from many groups and organizations around the world attended a conference run by Rehabilitation International – a worldwide organization *for* disabled people (Hurst, 2003). Disabled people were angry that nondisabled people were excluding them from decision making and using their power to retain control over disabled people. The DPI gives disabled people from around the world a chance to share their different experiences of campaigning and to learn from each other. Across the globe, many societies do not take disabled peoples' needs into account, and the DPI realized that because of this, discrimination had to be dealt with at an international level. In 1984, the DPI contributed to persuading the United Nations to change their Declaration of Human Rights to include disabled people (Hurst, 2003; Ratska, 1992).

Disability Awareness in Action (DAA) (1992–2011). Founded by Rachel Hurst OBE, DAA was set up in 1992 as a cooperative project of international disability organizations that responded to disabled peoples' need for appropriate information. DAA's mission was to provide information and evidence to support disabled people in their own actions in securing their rights at all levels: local, national, regional, and international.

Independent Living Institute (ILI) (1989). The ILI is a policy development center based in Stockholm, Sweden. It specializes in consumer-driven policies for disabled peoples' self-determination, self-respect, and dignity (see http://www.independentliving.org/indexen.html accessed 20.06.13.).

The disability movement has been a hive of direct action. This has been direct action in the form of protests, marches, or acts of civil disobedience. These actions have ranged from international actions to country specific actions.

Within different countries, there may be different organizational frameworks, but all their direct actions represent ways of highlighting inequalities in all countries (Shakespeare, 1993). For example, the ENIL has held a Freedom Drive once every 2 years since 2003. This consists of a march to the European Parliament by participating European countries to speak with Members of the European Parliament. In America, Americans Disabled for Accessible Public Transit (ADAPT) have been successfully undertaking nonviolent direct actions on accessible transport in America since 1978. In the United Kingdom, Direct Action Network (DAN) formed in 1993, and organized direct actions especially around transport access.

Also in the United Kingdom, Disabled People against Cuts (DPAC): formed in 2010, was founded and is run by disabled people. It has worked through a series of direct actions including roadblocks, marches, and occupations of government buildings. It also works at the national, international policy and human rights level. In a declassified Government memo, DPAC has been described by the Government's Department of Works and Pensions as 'the most vocal critic' of its austerity policies. Many direct actions at the country level are organized by CILs and other groups that are led and controlled by disabled people. These include among others, CIL-Sofia in Bulgaria, ULOBA or Uavhengig Liv Oslo og omegn (meaning independent life), In and around Oslo in Norway, and YHD (the Association for Theory and Culture of Handicap) in Slovenia. It is clear that with European austerity more radical grass roots groups are being set up to overtake the more formal disability organizations to revive the disability movement in a time of a crisis.

Independent Living Today

The independent living movement today has achieved many changes in the ways that disabled people are treated, including in the physical environment, transport, and access. However, the call for independent living is a call to end the remaining social barriers that prevent disabled peoples' choices, human rights, and inclusion within their own societies, and internationally.

For example, some disabled people seek to employ PAs. PAs can help to negotiate some of the barriers that disabled people face. However, many governments still refuse to fund PAs adequately, or offer only limited financial aid to those seeking to employ them. For example, in the United Kingdom, the proposed closure of the Independent Living Fund for those with perceived complex needs, and the reduction in funds from local authorities through central Government cutbacks has serious implications for the lives of disabled people. Data collected over 2011–12 indicate that social care has already been cut by nearly £4 billion and will be cut by £8 billion by 2015 in the United Kingdom, a cut of about 33% (Duffy 2013). In Sweden, the municipality of Visby limits the tasks PAs can undertake. A person living alone is now expected to live in a flat/house with only two rooms and a kitchen. Those living in a larger flat are told that they should move or pay for the cleaning of the rest of the flat. Members of Parliament reaffirmed the National Social Insurance Office's new interpretation of the limitations on PAs by saying that if you can get the spoon to your mouth or eat by yourself, you do not need PAs, even if you need help in getting the food to the table (Jolly, 2009).

Other governments are also refusing to fund adequately real independent living alternatives. For example, a national

government that declares that they are supporting a PAs scheme, but will only fund an unemployed family member to be the PA, is exploiting family labor by trying to get PAs on the cheap. This stance enables them to remove this person from the unemployment figures and at the same time to pay a low amount of salary to the person for being the PA. This issue has several other problematic respects. One is that self-determination is often better achieved with a PA who is outside the family. Another is that low-cost schemes develop into substandard alternatives to the real goal of independent living, as explained in Panayotova's (2009) *ANED Country Report*.

These responses increase barriers rather than partnering disabled people in the journey toward full citizenship. Many issues remain to be solved effectively. For example, in education, disabled children often find themselves in segregated schools. Housing and transportation can remain inaccessible. And, disabled people are less likely to be employed due to factors such as employers' attitudes, inaccessible workplaces, and lack of aids and support. While many governments proclaim inclusive policies, they do not always implement or enforce those that could improve independent living and citizenship for disabled people. As with the examples on PAs given earlier, many work programs for disabled people can have poor outcomes. A recent program in the United Kingdom was criticized for poor performance by the Public Accounts Select Committee who said "It is shocking that of the 9500 former incapacity benefit claimants referred to providers, only 20 people have been placed in a job that has lasted 3 months, while the poorest performing provider did not manage to place a single person in the under 25 category into a job lasting 6 months" (http://www.publications.parliament.uk/pa/cm201213/cmselect/cmpubacc/936/936.pdf (accessed 20.06.13.)).

These factors are particularly disappointing in light of the United Nations Convention on the Rights of Persons with Disabilities (UNCRPD), which stresses independent living throughout and in Article 19 specifically. The UNCRPD and its optional protocol were adopted on 13 December 2006 at the United Nations Headquarters in New York, and opened for signature on 30 March 2007. A host of countries have ratified the convention and the optional protocol. The signatories and ratifications can be found on the UN website (http://www.un.org/disabilities/countries.asp?id=166 (accessed 09.06.13.)).

It is continued government failures that compel disabled people to continue to work together in organizations 'of' disabled people and grass roots groups to improve disabled peoples' situations and move toward a time when all disabled people everywhere can choose how and where they live as equal members of society.

There is a wide variation in rights achieved in each country, the struggle continues with varying and sometimes lukewarm support by governments. The Academic Network of European Experts in Academia, a project set up in 2007 created and funded under the European Commission's PROGRESS programme, gives a useful, if uncritical, data set of policies for each European country and the European Free Trade Area (EFTA). The purpose of the project is to establish and maintain an academic research network throughout the countries of the European Union and the EFTA to inform disability policy making at the European level. It has a website where such information can be obtained (http://www.disability-europe.net/about-us accessed 09.06.13.). While this is welcome as a resource, true empowerment and equality will be achieved when the European Commission recognizes that it is disabled people themselves who are the experts on their lives, their histories, and the policies that can turn aspirations for change into reality.

Disabled people in every country continue to be more likely to experience poverty. And, gains made in independent living are under attack by austerity measures within many countries experiencing cutbacks. Declining support for disabled people is curtailing their human rights and independent living. Consequently, the gains of the disability movement are being reversed in some countries while remaining static in others. Gaining equal rights and citizenship for disabled people is a battle that will continue.

See also: Critical Social Work Practice; Disabilism; Social Work Theory.

Bibliography

Barnes, C., 1991. Disabled People in Britain and Discrimination. Hurst and Co in Association with the British Council of Organisations of Disabled People, London.

Barnes, C., Mercer, G., Shakespeare, T., 1999. Exploring Disability: A Sociological Introduction. Polity Press, Cambridge.

Behrent, M., Dunn, B., Curts, B., Gagan, J., 1999. Timeline of the International Independent Living Movement Milestones through to 2003. http://www.ilru.org/html/projects/international/3-international_il_timeline.htm (accessed 09.12.12.).

Campbell, J., Oliver, M., 1996. Understanding Our Past, Challenging Our Future. Routledge, London.

Cardona, G.C., 2013. Overcoming disabling barriers: disability and, attitudes to disability within the maltese context: an insider. Disability & Society 28 (2), 279–284.

Duffy, S., 2013. A Fair Society? Centre for Welfare Reform. http://www.centreforwelfarereform.org/uploads/attachment/354/a-fair-society.pdf (accessed 20.06.13.).

Finklestein, V., 1980. Attitudes and Disabled People. World Rehabilitation Fund, New York.

Hurst, R., 2003. The international disability rights movement and the ICF 2003. Disability and Rehabilitation 25 (11–12), 572–576.

Jolly, D., 2003. The government of disability: economics and power in welfare and work. Disability & Society 18 (4), 509–522.

Jolly, D., 2009. Personal Assistance and Independent Living: Article 19 of the UN Convention on the Rights of Persons with Disabilities. Leeds University Centre for Disability Studies, Disability Archive.

Maart, S., Eide, A.H., Jelsma, J., Loeb, M.E., Ka Toni, M., 2007. Environmental barriers experienced by urban and rural disabled people in South Africa. Disability & Society 22 (4), 357–369.

Mladenov, T., 2009. Institutional woes of participation: Bulgarian disabled, people's organisations and policy-making. Disability & Society 24 (1), 33–45.

Moriña Díez, A., 2010. School memories of young people with disabilities: an analysis of barriers and aids to inclusion. Disability & Society 25 (2), 163–175.

Oliver, M., 1990. The Politics of Disablement. Macmillan, Basingstoke.

Oliver, M., Barnes, C., 1998. Disabled People and Social Policy; from Exclusion to Inclusion. Longman, London.

Oliver, M., Barnes, C., 2012. The New Politics of Disablement. Palgrave, Macmillan, Basingstoke.

Panayotova, K, 2009. ANED Country Report on the Implementation of Polices, Supporting Independent Living for Disabled People (Bulgaria). http://www.disabilityeurope.net/content/aned/media/BG-6-Request-07%20ANED_2009_Task_5_template_Bulgaria_to%20publish_to%20EC.pdf (accessed 20.06.13.).

Ratska, A., 1992. Independent living. In: World Congress III: Disabled People's Equalisation of Opportunities. Disabled People's International, Vancouver (undated).

Shakespeare, T., 1993. Disabled people's self-organisation: a new social, movement? Disability, Handicap & Society 8 (3), 249–264.

Shier, M., Graham, J.R., Jones, M.R., 2009. Barriers to employment as experienced by disabled people: a qualitative analysis in Calgary and Regina, Canada. Disability & Society 24 (1), 63–75.
UPIAS (Union of Physically Impaired against Segregation), 1976. Fundamental Principles of Disability. Union of Physically Impaired against Segregation, London.

Relevant Websites

http://www.adapt.org/main – ADAPT (accessed 20.06.13.).
http://www.disability-europe.net/ – Academic Network of Disability Experts in Europe (ANED) (accessed 20.06.13.).
http://www.cil.bg/en/ – CIL – Sofia Home Page (accessed 09.12.12.).
http://www.daa.org.uk/ – Disability Awareness in Action (accessed 09.12.12.).
http://www.dpi.org/ – Disabled People's International (accessed 09.12.12.).
http://www.dpac.uk.net – DPAC Disabled people against Cuts (accessed 09.12.12.).
www.dpac.uk.net – DPAC Disabled people against Cuts (accessed 20.06.13.).
http://www.independentliving.org/docs2/enilstrasbourgresolutions.html – ENIL foundation statement on personal assistance (accessed 09.12.12.).
http://www.enil.eu/ – European Network for Independent Living (accessed 09.12.12.).
http://www.independentliving.org/ – Independent Living Institute – (accessed 09.12.12.).
http://www.uloba.no/english/Sider/default.aspx – ULOBA (accessed 09.12.12.).
http://www.un.org/disabilities/ – UN Convention on the Rights of Persons with Disabilities (accessed 09.12.12.).
http://www.leeds.ac.uk/disability-studies/archiveuk/UPIAS/fundamental%20principles.pdf – UPIAS Fundamental Principles (accessed 09.12.12.).
http://www.yhd-drustvo.si/eng/ – YHD (accessed 09.12.12.).

Disability: Social and Psychological Aspects

Jan D Reinhardt, Institute for Disaster Management and Reconstruction of Sichuan University and Hong Kong Polytechnic University, Chengdu, Sichuan, China; and Swiss Paraplegic Research, Nottwil, Switzerland
Marcel WM Post, University Medical Center Groningen, Groningen, The Netherlands; and Center of Excellence in Rehabilitation Medicine, University Medical Center Utrecht and De Hoogstraat, Utrecht, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.
This article is a replacement of the previous edition article by M. Johnston, D. Bonetti, volume 6, pp. 3704–3710, © 2001, Elsevier Ltd.

Abstract

Disability is a universal biopsychosocial phenomenon and involves impairments at the body level, limitations in activity, and restrictions of social participation. Disability is not solely influenced by biomedical pathology but to a large extent by an interaction of environmental factors such as inaccessible buildings or negative attitudes with features of the person such as passive coping styles. Disability is highly prevalent across the world. There is clear evidence for social and psychological determinants as well as consequences of disability.

The Biopsychosocial Model of Disability

The term disability refers to a complex and dynamic interaction of biomedical, psychological, and social factors that become manifest in problems in human functioning and not merely to an attribute that a person has or has not (Bickenbach et al., 1999). Human functioning is an umbrella term introduced by the WHO's International, Classification of Functioning, Disability, and Health (ICF) and reflects the universal human experience in relation to body functions and structures, activities, and participation. Disability, in turn, comprises impairment in body functions or structures, limitation in activity, and restriction of participation experienced by people with a health condition such as a disorder, disease, injury, or trauma (WHO, 2001). Impairments are deviations of physiological and mental functions such as hearing or emotions and/or anatomical parts of the body such as a limb or the brain from statistically defined population norms. Limitations in activities refer to problems people experience in the execution of a task such as lifting an object, and restrictions of participation are defined as problems experienced in relation to involvement in life situations such as pursuing remunerative employment or socializing with friends (WHO, 2001). A person with a health condition may exclusively experience disability in terms of participation restrictions, for example, when a person with HIV is excluded from the labor market. In its current model of disability the WHO (2001) furthermore acknowledges that experiences of disability are influenced by contextual factors, that is, so-called environmental and personal factors. Environmental factors are external features of a persons environment, for example, the accessibility of buildings or attitudes of others. Disability may result from the presence of environmental barriers such as inaccessible buildings or social discrimination or the absence of facilitators such as the lack of supports or assistive technologies. In turn, the presence of facilitators such as assistive devices or social support from family members may reduce disability experience. Personal factors are defined as internal attributes of a person other than impairments or health conditions such as a person's gender, age, education, or coping styles. Personal factors interact with environmental factors in increasing or decreasing disability, for example, in certain cultures females with physical disabilities may be viewed as not eligible for marriage while males do not face similar problems. Against this background, the ICF model of disability (**Figure 1**) integrates two supposedly opposing models of disability, the biomedical or health consequences model which sees disability as a result of disease or disorder and corresponding impairments and the social model of disability which conceptualizes disability as resulting from processes of social discrimination and stigmatization (Bickenbach et al., 1999).

The ICF takes an etiologically neutral standpoint and acknowledges that in some cases disability may result from biomedical factors and in others from social factors or an interaction of both which must be empirically determined (WHO, 2001). Moreover, the ICF conceptualizes disability as a continuum as opposed to a dichotomy, that is, there are various levels of human functioning and disability "and the complete absence of disability, like the complete absence of ability, is a limiting case of theoretic interest only" (Bickenbach et al., 1999: 1182). Furthermore, disability may be a permanent or temporary experience. Therefore, human functioning and disability are universal experiences, that is, every human being has experienced and will experience disability of some sort, in particular with aging (Bickenbach et al., 1999; Zola, 1989).

Prevalence

Estimates of prevalence of disability largely vary between countries if available at all. Prevalence estimates depend on definitions of disability and measures employed, for example, the questions used in a health survey to identify populations with disabilities. In its World Report on Disability published in 2011 the WHO (2011) attempted to provide global estimates of disability prevalence based on the reanalysis of data from two sources: (1) the World Health Survey was carried out between 2002 and 2004 in 70 countries reflecting all regions of the world (for prevalence estimates the WHO selected the 59 countries for which sampling weights were available); data were collected in face-to-face household interviews with person

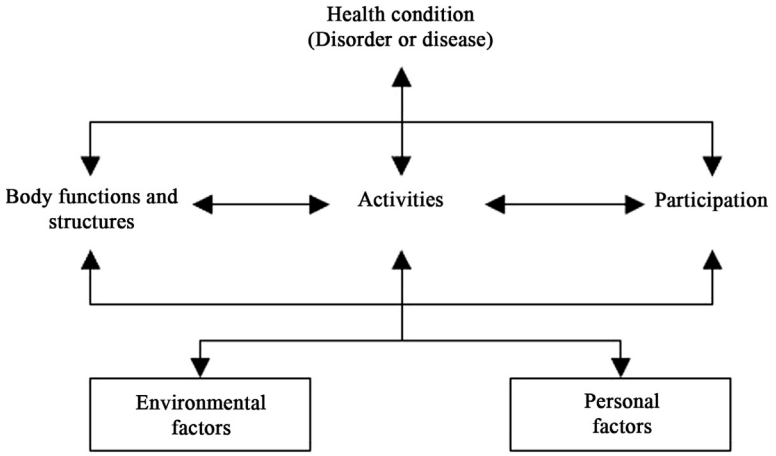

Figure 1 Integrative model of disability of the International Classification of Functioning, Disability and Health (WHO, 2001).

aged 18 years and older selected based on multistage stratified cluster sampling; the survey included questions on affect, cognition, interpersonal relationships, mobility, pain, sleep and energy, self-care, and vision. (2) The Global Burden of Disease Study integrates data on 130 health conditions from 17 subregions of the world including information on functioning if available; for disability prevalence estimates information collected between 2000 and 2004 was used.

Based on the World Health Survey Data global prevalence of moderate to severe disability was estimated at 15.6% or around 650 million from 4.2 billion people aged 18 years or older in 2004. Extrapolation of data to persons aged 15 years or older led to a prevalence estimate of around 720 million people experiencing disability in 2004. Prevalence was higher in low-resource than high-resource countries, among older people, those with poorer wealth, and females.

Based on the Global Burden of Disease Study disability prevalence was estimated at 15.3% of the total world population including children or around 978 million persons of estimated 6.4 billion in 2004. Prevalence in persons aged 15 years and older was estimated at 19.4% or 892 million.

Extrapolating data to 2010 led to a global disability prevalence estimate of over 1 billion persons including children.

Notably, above figures depend on the thresholds used for determining disability, for example, a threshold of 40 from a total score ranging from 0 indicating no problems in functioning to 100 indicating complete problems for the World Health Survey data. Estimates would be different if another threshold would have been used (WHO, 2011).

Global prevalence of disability is expected to increase in the future due to demographic aging, growing importance of noncommunicable disease, increasing road traffic injuries, and increased occurrence of natural disasters (WHO, 2011).

Classification and Measurement

While health conditions can be classified with WHO's International Classification of Diseases (ICD), functioning and disability as well as environmental factors can be classified with ICF (WHO, 2001). The ICF provides the user with one list of categories for body functions, one list for body structures, one common list for activities and participation, and a list of environmental factors. Currently, a list of categories for classification of personal factors is not included in ICF. All those lists are structured in chapters which include categories of aspects of functioning and subcategories at different levels, for example, Chapter 7 Interpersonal Interactions and Relationships includes a category d720 Complex Personal Interactions which in turn includes subcategories such as d7200 Forming Relationships and d7201 Terminating Relationships. Problems in functioning can be rated with a generic ICF qualifier scale ranging from no problem (0) to complete problem (4). Activities and Participation may be rated in terms of capacity or performance with capacity referring to a standardized environment or test environment and performance referring to a current or actual living environment within which facilitators such as devices or barriers may be present. Difference between capacity and performance thus reflect environmental impact on a person's or population's activity and participation levels. Environmental factors can be rated as barriers ranging from no barrier (0) to complete barrier (4) or facilitators ranging from no facilitator (+0) to complete facilitator (+4). In total, the ICF encompasses 1454 categories making its practical use quite difficult. Therefore, so-called ICF Core Sets have been developed based on literature reviews, expert surveys, patient focus groups, multicenter clinical studies, and international consensus conferences (Cieza et al., 2004). ICF Core Sets provide evidence-based selections of categories that are of major relevance to populations with particular health conditions such as spinal cord injury (Cieza et al., 2010) or to particular settings such as vocational rehabilitation (Escorpizo et al., 2010). Those categories should be considered in clinical practice as well as research. Countless measures of disability or its components are available. The United Nations commissioned Washington Group has developed a short set of seven questions on disability for use in census based on ICF (Washington Group on Disability Statistics, 2010). The WHO-DAS II is a generic ICF-based instrument for self-reported or interviewer-administered assessment of problems in different areas of

functioning (WHO, 2009). Impairments can also be determined with clinical assessment techniques depending on the body function or structure in question. Classical generic measures of limitations in activities of daily living used in rehabilitation include the Functional Independence Measure (FIM) (Keith et al., 1987) and the Barthel Index (Mahoney and Barthel, 1965). Numerous measures of community integration or social participation are also available. The Utrecht Scale for Evaluation of Rehabilitation–Participation (USER-P), for instance, is an ICF-based measure of frequency, perceived restrictions, and satisfaction with participation in different life domains (Post et al., 2012). Different approaches have been used for measuring environmental factors and their impact on functioning (Reinhardt et al., 2011b). People with disabling health conditions may, for instance, be asked to rate certain aspects of their environment as barriers or facilitators or research assistants may observe the presence or absence of several environmental features such as wheelchair accessible toilets or stairs and elevators (Reinhardt et al., 2011b).

Social Determinants and Consequences of Disability

The United Nations (UN) Convention on the Rights of Persons with Disabilities states in its preamble (e) "that disability results from the interaction between persons with impairments and attitudinal and environmental barriers that hinders their full and effective participation in society on an equal basis with others" (UN, 2006) and thus recognizes that determinants as well as consequences of disability may be essentially social in nature. The immediate environment of a person is further to some degree shaped by the organization of society as whole, for example, in terms of legal prescriptions or health policies (Reinhardt et al., 2011b).

Two persons with the same health condition may experience different degrees of impairment and two persons with the same level of impairment may experience different degrees of restrictions in participation depending on the interaction of personal and environmental factors (WHO, 2001, 2011). A study from Switzerland found that impairment was not directly related to restrictions in activity and participation if it was conditioned on contextual factors such as social support (Fellinghauer et al., 2012).

Numerous studies have also demonstrated a social gradient in disability, that is, people with lower socioeconomic status in terms of education and income as well as people with lower levels of social support are at higher risk of developing disability (e.g., Minkler et al., 2006; Reinhardt et al., 2012; Fekete et al., 2014). In the aging population, people with lower socioeconomic status tend to have an earlier onset of disability (Martin et al., 2012). These social gradients vary between countries depending on social and health policies. Wahrendorf et al. (2013), for instance, showed an earlier onset of disability in the aging US population as compared with Europe and that this was mainly due to a larger gap between groups with the lowest and highest wealth. The impact of socioeconomic status on disability may further be mediated by other social factors such as poor working conditions and work stress (Reinhardt et al., 2013).

Disability also may have profound social consequences. People with disabilities have worse educational attainment, are less often employed, and have higher unemployment and poverty rates as compared with the general population in most developed as well as developing countries (OECD, 2010; WHO, 2011). Moreover, persons experiencing profound disabilities are at higher risk of developing secondary health conditions and premature death (WHO, 2011). For instance, a British study of women aged between 64 and 83 found that women experiencing limitations in complex activities and restrictions in participation had an increased risk of mortality (Dale et al., 2012).

The above consequences of disability may be aggravated by other social factors such as inaccessible health care facilities or working environments as well as negative attitudes toward disability in the general population (Reinhardt et al., 2011a) and among health care workers (Tervo et al., 2004). Attitudes toward disability may in turn be influenced by the mass media which have been criticized for a high prevalence of disability stereotypes such as depicting persons with disability as evil (Biklen and Bogdan, 1977). Conversely, positive portrayals of persons with disabilities may improve the general population's attitudes. A study from Switzerland, for instance, demonstrated that attitudes of the general public toward eligibility of persons with spinal cord injuries were improved after watching an action scene showing a wheelchair user in the role of a police detective (Reinhardt et al., 2014).

Personal Factors and Psychological Determinants of Disability

In interaction with environmental factors, personal factors impact disability. As described above, personal factors are not classified in the ICF "because of the large social and cultural variance associated with them" (WHO, 2001: p. 8). In general, personal factors concern the particular background of an individual's life and living and comprise features of the individual that are not part of a health condition or health states (WHO, 2001). Examples of what might be considered as personal factors given are gender, age, coping styles, social background, education, profession, past and current experience, overall behavior pattern, and character (WHO, 2001). This list illustrates some of the difficulties associated with classifying personal factors. Is coping styles different from overall behavior pattern? How can profession be a personal factor if remunerative employment is also a category within the Activities and Participation component? How can current experience be a personal factor influencing functioning? Even an apparently simple variable such as gender leads to discussion, for example, whether to classify biological or social gender. Such difficulties prompted some scholars to question the usefulness of such a classification at all (Simeonsson et al., 2014).

The lack of guidance on how to classify personal factors in the ICF has led to a diverging use of the term personal factors in the literature. In 2011, Geyh et al. reported the use of no less than 238 different personal factors in the literature (Geyh et al., 2011). Several authors have provided proposals for the classifications of personal factors, but there is no glimpse of consensus yet. However, these difficulties also illustrate the need and potential significance of international consensus on

a classification of personal factors for research and health policy.

Broadly spoken, it might be possible to distinguish (1) facts about the individual's position in the physical, social, and temporal context, such as age, gender, ethnicity, partnership, and marriage, from psychological personal factors, within which we might distinguish two subcategories: (2a) the basic psychological makeup of the person, such as personality, optimism, locus of control, and self-efficacy; and (2b) patterns of thoughts and behaviors, such as coping, cognitions about one's condition, and self-management strategies. Research has consistently shown that psychological factors are associated with outcomes such as mood, participation, and well-being in persons with disabilities. Some, such as the personality trait neuroticism, low self-esteem, and passive and avoidance coping strategies with poor outcomes, and others such as extraversion, internal locus of control, and high self-esteem with good outcomes (van Leeuwen et al., 2012).

Psychological Consequences and Adjustment to Disability

After onset of disability, persons are confronted with new situations that challenge their habitual coping strategies, participation, and quality of life. From a health psychology perspective, disability is not unlike any other stressor. The way this stressor is appraised and the effectiveness of the coping response resulting from that appraisal, someone's habitual coping repertoire and other contextual factors determines if the person is able to adjust to the new situation. The term adjustment can be defined as a response to a change in the environment that allows an organism to become more suitably adapted to that change (Sharpe and Curran, 2006). In the context of health, it refers to the healthy rebalancing by patients to their new circumstances. The term adjustment is however used in different ways: (1) to indicate a state of good outcome (de Ridder et al., 2008) and (2) to indicate a process that starts with the onset of disability and continues throughout the course of disability and responds to changes in illness status (Sharpe and Curran, 2006). Most patients experience good adjustment, but for about 30% of patients the adjustment phase is prolonged and sometimes unsuccessful (Taylor and Aspinwall, 1996). Several models have been proposed on how successful adjustment can be achieved (de Ridder et al., 2008). First, the model of cognitive adaptation, which emphasizes illness acceptance and perceptions of control over illness (Taylor, 1983). Second, the personality model that emphasizes the role of factors such as optimism or neuroticism in adjustment. Third, the stress and coping model that emphasizes strategies used by patients to deal with adaptive tasks imposed by disease (Adler and Matthews, 1994). The authoritative stress and coping model acknowledges that chronic illness consists of several challenges, but at the same time, it highlights – more than other models of adjustment – processes of appraisal and coping that explain why some patients successfully identify and act on opportunities to manage these tasks whereas others might fail to do so (de Ridder et al., 2008). More recently, the stress and coping model has been extended with the model of self-regulation, which allows individuals to deal with illness more proactively (Cameron and Leventhal, 2003). Both models show the active role that persons may have in adjusting to the challenges posed by their condition, and they have been used to study processes of adjustment in diverse chronic conditions.

In the self-regulation model (SRM), illness representations contain information about five aspects of the illness: identity, cause, consequences, cure or control, and timeline (Brownlee et al., 2000). Secondly, these representations operate simultaneously at two levels: cognitive and affective. Emotional representations tend to be stronger motivators for behavior than cognitive representations, for example, if one feels bad but avoids a medical consultation because a serious illness might be discovered. According to the SRM, people use knowledge and experience to construct their illness representations. Contextual factors such as social roles and self-perceptions affect these representations (Sharpe and Curran, 2006). The need to develop these representations stem from the need to regulate the self and maintain equilibrium, and to preserve or adapt one's self-perceptions in the light of the new illness (Brownlee et al., 2000).

From this theoretical background it becomes clear why an association is seen between the severity of illness and adjustment outcomes, but why this association is, at best, moderate. A more severe condition has more severe consequences, requires more 'adjustment work,' and results in lower adjustment scores at a given time postonset, but individual differences in psychological makeup and in context cause large variation in the degree to which persons with disability are able to adjust to their situation. For example, a review of studies comparing life satisfaction of people with spinal cord injuries to the general population showed that, on average, life satisfaction scores in samples of persons with spinal cord injury are 0.8 standard deviation below the average scores of the reference populations. Within the spinal cord injury population, however, there was a wide range of life satisfaction and only a weak association between life satisfaction and the degree of impairment or functional limitations (Post and van Leeuwen, 2012).

See also: Chronic Illness, Psychosocial Coping with; Disability Movement; Disability and Disabling Practices; Disability: Sociological Aspects; Functioning, Disability and Health, International Classification of; Injuries and Accidents: Psychosocial Aspects; Social Integration, Social Networks, and Health.

Bibliography

Adler, N., Matthews, K., 1994. Health psychology. Why do some people get sick and some stay well? Annual Review of Psychology 45, 229–259.

Bickenbach, J.E., Chatterji, S., Badley, E.M., Ustun, T.B., 1999. Models of disablement, universalism and the international classification of impairments, disabilities and handicaps. Social Science & Medicine 48, 1173–1187.

Biklen, D., Bogdan, R., 1977. Media portrayals of disabled people: a study in stereotypes. Interracial Books Children Bulletin 8, 4–9.

Brownlee, S., Leventhal, H., Leventhal, E., 2000. Regulation, self-regulation and construction of the self in the aintenance of physical health. In: Boekaerts, M., Pintrich, P. (Eds.), Handbook of Self-Regulation. Academic Press, London, pp. 369–416.

Cameron, L.D., Leventhal, H., 2003. Self-regulation, health and illness. An overview. In: Cameron, L.D., Leventhal, H. (Eds.), The Self-Regulation of Health and Illness Behaviour. Routledge, New York, pp. 1–13.

Cieza, A., Ewert, T., Ustun, T.B., Chatterji, S., Kostanjsek, N., Stucki, G., 2004. Development of ICF core sets for patients with chronic conditions. Journal of Rehabilitation Medicine, 9–11.

Cieza, A., Kirchberger, I., Biering-Sorensen, F., Baumberger, M., Charlifue, S., Post, M.W., Campbell, R., Kovindha, A., Ring, H., Sinnott, A., Kostanjsek, N., Stucki, G., 2010. ICF core sets for individuals with spinal cord injury in the long-term context. Spinal Cord 48, 305–312.

Dale, C., Prieto-Merino, D., Kuper, H., Adamson, J., Bowling, A., Ebrahim, S., Casas, J.P., 2012. Modelling the association of disability according to the WHO international classification of functioning, disability and health (ICF) with mortality in the British women's heart and health study. Journal of Epidemiology and Community Health 66, 170–175.

Escorpizo, R., Ekholm, J., Gmunder, H.P., Cieza, A., Kostanjsek, N., Stucki, G., 2010. Developing a core set to describe functioning in vocational rehabilitation using the international classification of functioning, disability, and health (ICF). Journal of Occupational Rehabilitation 20, 502–511.

Fekete, C., Siegrist, J., Reinhardt, J.D., Brinkhof, M.W., 2014. Is financial hardship associated with reduced health in disability? The case of spinal cord injury in Switzerland. PLoS One 9, e90130.

Fellinghauer, B., Reinhardt, J.D., Stucki, G., Bickenbach, J., 2012. Explaining the disability paradox: a cross-sectional analysis of the Swiss general population. BMC Public Health 12, 655.

Geyh, S., Peter, C., Muller, R., Bickenbach, J.E., Kostanjsek, N., Ustun, B.T., Stucki, G., Cieza, A., 2011. The personal factors of the international classification of functioning, disability and health in the literature - a systematic review and content analysis. Disability and Rehabilitation 33, 1089–1102.

Keith, R.A., Granger, C.V., Hamilton, B.B., Sherwin, F.S., 1987. The functional independence measure: a new tool for rehabilitation. Advances in Clinical Rehabilitation 1, 6–18.

van Leeuwen, C.M., Kraaijeveld, S., Lindeman, E., Post, M.W., 2012. Associations between psychological factors and quality of life ratings in persons with spinal cord injury: a systematic review. Spinal Cord 50, 174–187.

Mahoney, F.I., Barthel, D.W., 1965. Functional evaluation: the barthel index. Maryland State Medical Journal 14, 61–65.

Martin, L.G., Schoeni, R.F., Andreski, P.M., Jagger, C., 2012. Trends and inequalities in late-life health and functioning in England. Journal of Epidemiology and Community Health 66, 874–880.

Minkler, M., Fuller-Thomson, E., Guralnik, J.M., 2006. Gradient of disability across the socioeconomic spectrum in the United States. New England Journal of Medicine 355, 695–703.

OECD, 2010. Sickness, Disability and Work: Breaking the Barriers [Online]. Available: http://ec.europa.eu/health/mental_health/eu_compass/reports_studies/disability_synthesis_2010_en.pdf (accessed 12.02.14.).

Post, M.W., Van Der Zee, C.H., Hennink, J., Schafrat, C.G., Visser-Meily, J.M., Van Berlekom, S.B., 2012. Validity of the utrecht scale for evaluation of rehabilitation-participation. Disability and Rehabilitation 34, 478–485.

Post, M.W., van Leeuwen, C.M., 2012. Psychosocial issues in spinal cord injury: a review. Spinal Cord 50, 382–389.

de Ridder, D., Geenen, R., Kuijer, R., Van Middendorp, H., 2008. Psychological adjustment to chronic disease. Lancet 372, 246–255.

Reinhardt, J.D., Ballert, C.S., Fellinghauer, B., Lotscher, A., Gradinger, F., Hilfiker, R., Graf, S., Stucki, G., 2011a. Visual perception and appraisal of persons with impairments: a randomised controlled field experiment using photo elicitation. Disability and Rehabilitation 33, 441–452.

Reinhardt, J.D., Miller, J., Stucki, G., Sykes, C., Gray, D.B., 2011b. Measuring impact of environmental factors on human functioning and disability: a review of various scientific approaches. Disability and Rehabilitation 33, 2151–2165.

Reinhardt, J.D., Pennycott, A., Fellinghauer, B.A., 2014. Impact of a film portrayal of a police officer with spinal cord injury on attitudes towards disability: a media effects experiment. Disability and Rehabilitation 36, 289–294.

Reinhardt, J.D., Von Elm, E., Fekete, C., Siegrist, J., 2012. Social inequalities of functioning and perceived health in Switzerland–a representative cross-sectional analysis. PLoS One 7, e38782.

Reinhardt, J.D., Wahrendorf, M., Siegrist, J., 2013. Socioeconomic position, psychosocial work environment and disability in an ageing workforce: a longitudinal analysis of SHARE data from 11 European countries. Occupational and Environmental Medicine 70, 156–163.

Sharpe, L., Curran, L., 2006. Understanding the process of adjustment to illness. Social Science & Medicine 62, 1153–1166.

Simeonsson, R.J., Lollar, D., Bjorck-Akesson, E., Granlund, M., Brown, S.C., Zhuoying, Q., Gray, D., Pan, Y., 2014. ICF and ICF-CY lessons learned: Pandora's box of personal factors. Disability and Rehabilitation.

Taylor, S.E., 1983. Adjustment to threatening events: a theory of cognitive adaptation. American Psychologist 38, 1161–1173.

Taylor, S.E., Aspinwall, L.G., 1996. Psychological aspects of serious illness. In: Costa, P.T., Van Den Bos, G.R. (Eds.), Chronic Conditions, Fatal Diseases and Clinical Care. American Psychological Association, Washington, DC, pp. 7–60.

Tervo, R.C., Palmer, G., Redinius, P., 2004. Health professional student attitudes towards people with disability. Clinical Rehabilitation 18, 908–915.

UN, 2006. Convention on the Rights with Persons with Disabilities [Online]. New York. Available: http://www.un.org/disabilities/convention/conventionfull.shtml (accessed 04.02.14.).

Wahrendorf, M., Reinhardt, J.D., Siegrist, J., 2013. Relationships of disability with age among adults aged 50 to 85: evidence from the United States, England and continental europe. PLoS One 8, e71893.

Washington Group on Disability Statistics, 2010. Short Set of Questions on Disability [Online]. Available: http://www.cdc.gov/nchs/washington_group/wg_questions.htm (accessed 01.02.14.).

WHO, 2001. International Classification of Functioning, Disability, and Health (ICF). WHO press, Geneva.

WHO, 2009. WHODAS 2.0-World Health Organization Disability Assessment Schedule 2.0 [Online]. WHO, Geneva. Available: http://www.who.int/classifications/icf/whodasii/en/ (accessed 03.10.13.).

WHO, 2011. World Report on Disability. WHO press, Geneva.

Zola, I.K., 1989. Toward the necessary universalizing of a disability policy. The Milbank Quarterly 67, 401–428.

Relevant Websites

http://www.cdc.gov/nchs/washington_group.htm.
http://www.icf-core-sets.org/en/page1.php.
http://www.un.org/disabilities/.
http://www.who.int/classifications/icf/en/.
http://www.who.int/disabilities/world_report/2011/en/.

Disability: Sociological Aspects

Nancy G Kutner, Emory University, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G.L. Albrecht, volume 6, pp. 3710–3713, © 2001, Elsevier Ltd.

Abstract

Definitions of disability are socially created and sustained through social interaction and through attitudes of others toward persons with impairments. Major themes in related sociological research include processes of adjusting to disability status, especially socialization and assuming a sick role; understanding disability as a form of deviance; the minority group model of disability; emergence of disability movements; and management of disabling conditions in people's everyday lives. Disability rights policy encompasses multiple agendas, including prevention, civil rights, provision of specialized care, and celebration of disability culture. Global aging and increasing prevalence of chronicity are important reasons to universalize supportive disability policy.

Introduction

Disability can be considered as a limitation in performing socially defined activities and roles to levels that are expected of individuals within a social and physical environment (Pope and Tarlov, 1991). Thus, disability – the expression of a physical or mental limitation in a social context – is socially defined. It can have pervasive social consequences for individuals and for societies. Individuals with widely different types of structural or functional impairments may share common conditions of exclusion, marginalization, and disadvantage. However, shared psychological, social, and economic consequences inherent in the disability experience may be expressed in a strong and positive sense of identity.

The demography and epidemiology of disability vary depending on how persons with disability are defined. In the United States, the Social Security Administration defines persons with disabilities as those who have a physical or mental condition that prevents them from engaging in paid employment. Other definitions of disability include having a condition assumed to be disabling (e.g., deafness); using assistive devices; self-identifying as a person with a disability; and being regarded by others as having a disability. Irving Zola (1993) insightfully argued that the conception, measurement, and counting of disability differ validly depending on the purpose for which this information is needed. In addition, disability is a status that may or may not be applicable at different times in an individual's life. Zola maintained, therefore, that disability is best conceived as fluid and continuous, rather than as fixed or dichotomous. The relevance of the latter point assumes increasing importance in the context of population aging.

After World War II, social scientists working in the fields of social medicine, social policy, and rehabilitation began to focus on social aspects of chronic illness and disability related to injury, but there was little consensus about definitions or approaches to measurement. In 1980, the World Health Organization undertook the development of a supplement to its International Classification of Diseases, which culminated in publication of the *International Classification of Impairments, Disabilities, and Handicaps* or *ICIDH* (WHO, 1980). The ICIDH framework recognized that whether a person performs a socially expected activity depends not simply on the characteristics of the person but also on the larger context of social and physical environments.

With the goals of increasing conceptual clarity and avoiding the suggestion of absolute limitation inherent in the ICIDH use of the term *handicap*, the Institute of Medicine (IOM) offered an alternative conceptual framework that differentiated four related but distinct stages: pathology, impairment, functional limitation, and disability (Pope and Tarlov, 1991). Although progression across these stages might characterize an individual's experience with a chronic disorder, progressively greater loss of function need not occur and the progression can be halted or reversed. Thus, within this framework disability prevention efforts can be directed at any of the three stages that precede disability as well as at the disability stage itself. These efforts could include reversal of disability, restoration of function, or prevention of complications, i.e., secondary conditions, that might exacerbate existing limitations or lead to new ones.

The IOM strongly endorsed a conceptualization of disability that incorporated environmental factors as primary contributors to the creation of disability, and advocated shifting the focus of disability from the individual and the impairment to the interaction between the impairment and the environment. Importantly, in this view the cultural category of disabled was socially, rather than medically, constructed (Higgins, 1992), in part by cultural definitions and in part by the demands and constraints of social and physical environments. This conceptualization challenged the medical model of disability, in which the focus is on an impairment-related limitation in need of remediation.

Sociological research on chronic illness, impairment, and disability originated in an effort to understand the relationship between experiences of symptoms or impairment, the social situations within which people live, and the combined effect these have on people's lives. Major themes that have been developed include adjusting to disability through socialization processes and assumption of a sick role, understanding disability as a form of deviance, the minority group model of disability, emergence of disability movements, and

management of illness and disability in people's everyday lives. Conceptual frameworks applied to the study of disability may embody quite different policy implications.

Processes of Adjusting to Disability

Because the medical model largely equated disability with dependency, much of the early sociological and social psychology research on disability focused on individuals' adjustment to a dependent status. Disability onset was viewed as requiring a redefinition of one's situation and a reconstruction of roles and social interaction patterns (Albrecht, 1976). Scholars focused on adjustment to, and coping with, impairment; morale and motivation in rehabilitation settings; and the importance of social support from family and community. Safilios-Rothschild (1976) noted that there is often a striking similarity in attitudes toward women and persons with a disabling condition and that both were likely to feel a need to accept and adjust to a stereotyped role in order to receive approval. Persons holding multiple roles that are perceived as devalued, e.g., women with a disabling condition, illustrated a situation of double jeopardy (Deegan and Brooks, 1985).

The opportunity for legitimate exemptions from usual role obligations, consistent with assuming the sick role, was regarded by some sociologists as virtually synonymous with disability. Stone (1984) maintained that disability can be viewed as a political privilege in the sense of carrying permission to be exempt from the work-based system, military service, debt, and criminal liability. A sick role perspective in which the interpretation of disability entails exemptions from usual social obligations was seen as characteristic of welfare organizations. Roth (1963) and Blaxter (1976) observed that the way that persons with disabling conditions presented themselves in their help-seeking behavior was shaped by the agencies with which they interacted, setting in motion a continuous process of adjustment. Timetables developed by the medical profession served to structure the patient's 'career' in ways congruent with the agenda of the organization (Roth, 1963), as did stereotypes of the proper client. The more marginal the client, the less precise and valid the stereotype. Societies with more developed formal processes for addressing needs and concerns of persons with disability were more likely to demonstrate tolerance of disability than were societies lacking formal processes, according to Gove (1976).

Conceptualizing Disability as a Form of Deviance

Theories of deviance also had an important influence on sociological studies of disability, beginning with Goffman's (1963) work on spoiled identity and the management of stigma. Safilios-Rothschild (1970) concluded that the position of disabled persons was best analyzed using a general theory of deviance, and early research focusing on persons with blindness considered disability within a deviance or societal reaction theory framework (Scott, 1969). The societal reaction theorist attaches little significance to primary deviance except insofar as others react to it (Gove, 1976). Similarly, Becker (1963) argued that social groups create deviance by making rules where infractions constitute deviant behavior and by applying those rules to particular people and labeling them as outsiders, i.e., offenders. Thus, individuals are processed by government and private agencies and found to be either disabled or not disabled. The disabled might then be channeled into an institutional setting, where there is a disability subculture the person is expected to fit into. Scott (1969) described welfare agencies for the blind as engaged in a socialization process with the purpose of preparing a person with a disabling condition to play a type of deviant role, for example, making blind persons out of people who cannot see.

The literature on disability as deviance argued that, like other stigmatized groups, persons with disability tend to be evaluated as a category rather than as individuals. Two predominantly negative stereotypes are likely to be applied: one stereotype of the particular impairment (blind, deaf, spastic, etc.) and one that is attached to the general label of disabled (Blaxter, 1976). Thus, labeling, segregating, and feedback processes apply to persons with disability as to other deviants. The person with disability is so defined because he deviates from what he himself or others consider normal and appropriate. Individuals may therefore develop a wide variety of concealing techniques in order to avoid the stigmatized label (e.g., Goffman, 1963; Hunt, 1966; Darling, 2013).

Gove (1976) noted that both societal reaction theorists and rehabilitation theorists see disability behavior as largely a product of role relationships. Societal reaction theorists see the social system as creating and stabilizing deviant behavior. Rehabilitation theorists, on the other hand, see the same processes as helping the disabled individual lead a normal life. Whether these social processes are primarily beneficial or detrimental for the individual remains an empirical question.

The greater the individual's resources, the greater the likelihood of avoiding being channeled into impaired or disabled roles. Safilios-Rothschild (1976) noted that high social status is likely to be an overriding characteristic whose desirability can outweigh other characteristics, including disability. Persons considered to have high social status are most likely to be able to retain their majority status rights and to become rehabilitated on their own terms.

The Minority Group Model of Disability

The passage of the Rehabilitation Act of 1973 in the US contributed to a civil rights perspective and, in turn, to a minority perspective in which disability was seen as the product of a disabling environment (Gliedman and Roth, 1980; Stroman, 1982). Within the framework of the minority group model, people are categorized as disabled primarily because of stigmatizing or prejudicial attitudes about discernible bodily differences. The minority group model of disability includes three important postulates:

- the source of major problems for persons with disabilities is social attitudes,
- almost every facet of the environment has been shaped by public policy, and
- in a democratic society, policies reflect pervasive attitudes and values.

Thus, according to the minority group model, both the sources and the solutions for problems of disability are external, and attitudinal discrimination is identified as the major barrier for disabled persons. Moreover, Hahn (1994) argued that disabled persons lack a sense of 'generational continuity', which might enable them to transmit information regarding means of combating prejudice from one cohort to the next. Thus, persons with disability can be termed as an 'unigenerational minority', reacting to perceived social oppression at a particular period.

According to Gill (2001), there are important differences between persons with disability and other minorities, however. One difference noted by Gill is that negative ascriptions based on disability can be linked, at least superficially, to real human differences. Another important difference, she argued, is that unlike members of race and ethnic groups, most persons with disability identify as typical, unless they have developed a political or cultural consciousness that centralizes difference.

Emergence of Disability Movements

In the late 1960s and early 1970s, the civil rights movement, an increasing number of persons with disabilities attending college, and skilled leadership among persons with disabilities provided a combined impetus that contributed to the beginning of the independent living movement (Scotch, 1989). The independent living movement was premised on a sociopolitical model of disability, i.e., disability arising from person–environment interaction as opposed to disability that is linked primarily with the existence of physical or mental impairment. In the US, the independent living and disability rights movements in turn contributed to the framing of the 1973 Rehabilitation Act.

Section 504 of the Rehabilitation Act of 1973 was the first major civil rights legislation for disabled persons in the US, prohibiting discrimination by the federal government and private and governmental entities contracting with the federal government, such as educational institutions. In contrast to earlier legislation that had provided or extended benefits, Section 504 established full participation as a civil right. The legislation represented the convergence of two trends in public policy: (1) extending the nondiscrimination language of the 1964 Civil Rights Act to persons with disabilities and (2) broadening the scope of social programs benefiting disabled persons (Scotch, 2001). The second of these objectives called for increased funding and led to two presidential vetoes of the bill, distracting attention from the first objective of the legislation, i.e., preventing discrimination. A motivating factor for Section 504 was that persons served by the Vocational Rehabilitation (VR) system frequently encountered difficulty entering the job market after completing their training, due to structural and attitudinal barriers. Therefore, without structural modification and protection from discrimination, VR funds were not accomplishing the goal of returning individuals to gainful employment. Section 504 subsequently provided a basis for passage in 1990 of the more comprehensive Americans with Disabilities Act.

In the 1970s, deaf persons protested outside telephone companies around the US because of anger at having to pay for special equipment and the increased costs they incurred for making long-distance telephone calls. This demonstration was interpreted as an emotional response to the shattered hope that years of volunteer effort to achieve telecommunications access would be successful. Exclusion, as in deaf persons' experience of exclusion from telephone access, came to play a positive role in the evolution of a disability community, serving as a catalyst for a sense of shared identity and identifying a target for collective action (Scotch, 2001). In a similar way, black power, feminist, and other social movements of the 1960s stressed a positive self-image rooted in the collective identity of an excluded group demanding greater participation.

Factors that have helped to nurture social movements among persons with disability include:

- medical technologies, including prosthetics, medications, and surgical techniques, that enabled longer life, survival of injuries, and fuller participation in everyday life activities;
- the popularization of an ideology of deinstitutionalization and normalization, especially in the mental health field, that encouraged the growth of noninstitutional support systems and greater participation in community life;
- an increasing number of elderly persons in the society with physical disabilities and shared service needs; and
- a large number of disabled veterans from the Vietnam War who were activists, as well as widespread protests against the war itself that helped to legitimate social activism (Scotch, 2001).

In the US, it was also important that during the 1960s and early 1970s a number of major programs were enacted that promoted more complete participation by persons with disabilities: the Architectural Barriers Act of 1968, expanded funding for vocational and independent living services under the VR program beginning in 1968, the Developmental Disabilities Services and Facilities Construction Act of 1970, the Education for All Handicapped Children Act in 1974, and Project Head Start in 1974. These programs represented a logical extension of a pattern of expanding entitlements and services provided by the federal government based on broadly held social and political values, a pattern referred to by Daniel Bell as a revolution of rising entitlements (Scotch, 2001). This pattern, more than vocal advocacy by disabled persons, is thought to have fostered the adoption of Section 504 of the 1973 Rehabilitation Act in the US Consumer groups did, however, subsequently protest the delayed implementation of Section 504 at offices of the Department of Health, Education, and Welfare, which contributed to disability groups becoming empowered to participate more actively in the political arena.

The problems faced by persons with different types of impairments, and the proposed solutions for these problems, may be quite different, making the development of a shared consciousness problematic (Scotch, 1989). Thus, Barnartt and Scotch (2002) concluded that contentious political actions within disabled communities 1970–99, i.e., demands for desired changes and actions, were typically impairment-specific. Cross-disability protests, involving demands that potentially apply to people with all types of impairments, were found to occur in approximately only one out of four cases. This observation suggested that, because persons with different impairments may lack a shared collective consciousness,

concerted advocacy by large numbers of persons with disabilities is likely to be infrequent.

Managing the Experience of a Chronic Disorder/ Disability

Analyses that attempt to understand how people manage illness in their everyday lives represent an effort to illuminate an 'insider's' perspective. The experience of disability is one focus within the subjective experience of illness, which itself emerged as an interest in part due to the disability movement that had its origins in the 1970s. Self-help and advocacy groups took experiential data seriously, since they were in a sense 'experts' in it, in turn fostering a new social awareness of the problems of living with disability.

A sociology of illness experience is concerned with how people live with and in spite of illness or impairment. It is based on systematically collected and analyzed data from a number and variety of persons with a particular condition, as in Schneider and Conrad's (1983) study of living with epilepsy. The focus is on the meaning of the condition to the individual, the social organization of the individual's world, and the strategies the individual uses in adaptation. Subjective experience provides a window on the broad impact of living with impairment, emphasizing global areas of social life, including work, education, family, sex, identity, and self-esteem.

The experience of disability perspective is qualitatively different from the notion of adjusting to a disability status defined by others, which is an outsider's rather than an insider's view. The experiential perspective recognizes that individuals creatively manage their illness in their everyday lives, defining and redefining who they are and how they relate to others. Corbin and Strauss (1985) argued that illness experience can be usefully conceptualized in terms of work that must be accomplished. Because work is a concept central to the lives of individuals who do not have chronic illness, this conceptualization indicates shared experience for people who are ill and people who are well – an interpretation that differs from the sick role conceptualization. Blaxter's (1976) discussion of the relevance of 'career' for persons with disability who are involved in interactions with social agencies conveys a similar concept. However, the concept of illness trajectory (Corbin and Strauss, 1985) may be even more appropriate for understanding the experience of illness because it encompasses process and change and does not assume linearity or orderliness in progression, such as the concept of career implies. Other focal concerns in research on the experience of illness include the need to manage uncertainty and stigma (e.g., Schneider and Conrad, 1983) and the need for biographical work and reconstitution of self (e.g., Corbin and Strauss, 1985; Charmaz, 1987).

How definitions of disability are socially created or sustained through interaction and through attitudes of others toward persons with disabilities have been a theme in the work of many sociologists. Some researchers suggested that the able-bodied display less variability in verbal output, less smiling behavior, less eye contact, and greater motor inhibition with the disabled. This in turn limits types of social interaction for persons with disability and provides them with fewer opportunities for trying out roles and behavior. The more clearly defined and visible the disability, the greater the ease with which the disabled individual and the group adjust to each other, and vice versa, as shown by Omansky (2011) regarding differing societal responses to degrees of legal blindness. Ambiguity surrounding degree of impairment may negatively impact interpersonal relationships, resulting in confused expectations, goals that are unclear, and roles that are contradictory. Darling (2013) has traced the evolution of disability identity as the construction and experience of difference has changed over time. The underlying theme is that disability is a social status, not a biological condition, and identification – or lack of identification – with the status occurs through social interaction.

New Conceptual Directions within the Sociology of Disability

The growth of the disability rights movement and political efforts to achieve social rights in the 1970s and 1980s were important elements in an emergent change in the conceptualization of disability. The disability rights movement had the goal of changing both the self-concepts of, and societal conceptions about, participants. After the emergence of the disability rights movement, persons with disability were seen as shaping negotiated outcomes ('nothing about us without us'), rather than passively adapting to labels imposed on them. However, disability rights policy encompasses intersecting and competing agendas, including prevention, civil rights, provision of specialized care, and celebration of disability culture, as is well illustrated with respect to policy discussions related to persons with atypical neurological development such as autism (Baker, 2011).

In a sample of disabled activists, Hahn and Belt (2004) found that self-identity related to a personal affirmation of disability was a significant predictor of refusing medical treatment for disabling conditions. Age of onset of the individual's disabling condition was also a predictor. Persons with earlier age of disability onset were more likely to reject curative treatment. Hahn and Belt suggested two potential explanations for the latter finding. First, early disability onset is likely to lead to a stronger personal sense of disability identity. Second, early disability onset gives longer exposure to medical interventions that often prove disappointing because they do not provide cure. In addition to affirmation of disability identity, communal attachments were strong values among the activists who were studied. Ridolfo and Ward (2012) have explored the variety of ways that physical and social barriers shape personal choices about disability identity among people with impaired mobility.

Individuals' desire for connection with others is a recurring theme in the literature on disability, evident in discussions of isolation, invisibility, and struggle (e.g., Zola, 1982). A central feature of the disability insider experience is described as a persistent and disquieting sense of mistaken identity. Tension between the desire to be viewed as 'typical' vs the desire to be able to be comfortable affirming one's personal disability identity (this is who I am) is at the heart of the social experience of disability discussed by Carol Gill (2001). The identity that

the individual struggles to present to the world is commonly dismissed, with stereotypical identity ascriptions substituted for this identity. In one individual's words, 'my disability is how people respond to my disability' (Craig Vick, cited by Gill, 2001: p. 352). Thus, persons with disability critique and continually expand the concept of normality and work to bridge the gap between their self views and public perceptions, and between their inner and outer worlds.

The development of disability movements facilitated acceptance of self, pride, and a focus on positive identity among persons with disability (Hahn and Belt, 2004). At the same time, there is a tension between disability identity and economic and cultural imperatives of a highly normalizing society that pursues the goal of bodily uniformity as measured by criteria such as beauty, health, fitness, and intelligence. An ideology of 'able-ism' has exclusionary social functions, as privileged designations of the body are preserved and validated, providing cultural capital to those who can claim such status. Hahn (1994) argued that the dynamics of capitalism and 'commodity fetishism' encourage people to emulate virtually unattainable physical ideals. The environment, therefore, remains a powerful influence as persons with disability strive for integrity in self and in social relations (Gill, 2001).

Williams (2001) argued that the sociology of disability can benefit from focusing not only on the politics of identity but also on the politics of structures. Developments in social security systems and the entire 'disability business' in Western societies significantly impact the societal meaning of disability (Stone, 1984; Albrecht, 1992). Although a minority group perspective on people with disabilities can have political uses by calling attention to inequities in services and opportunities, it is important to note that Zola (1991) explicitly rejected use of a'us and them' framework. He recognized that aging societies with growing numbers of persons with chronic illness had great significance for the development of the disability movement, creating shared challenges for the able-bodied and the disabled and the opportunity to effectively partner in order to achieve broadly applicable objectives such as available transportation systems and assisted living options.

> Where high economic costs are involved, such as in mass transportation and long-term care, there may well be a decided disadvantage for the needs of people with disabilities to be thought of as 'special' and 'restrictive' rather than general, if not universal.
> (Zola, 1993: p. 30)

Because there are "seemingly infinite permutations of the experience of being physically different in a highly normalizing society", how different impairments translate into complex negotiated aspects of everyday life can be a valuable research focus (Williams, 2001: p.130). There have been few efforts to analyze parallels and differences in the experience and negotiated outcomes of persons with different impairments. Opportunities to develop a shared sense of community is certainly one important variable. For example, having a shared, distinctive language is a very important aspect of membership in the deaf community. The degree to which persons with a particular potentially disabling condition have social contact on a regular basis is another important factor, significantly influencing the opportunity for forming peer relationships. The era of social media provides a new international environment for this contact.

Conclusion

As modern medicine prolongs survival in persons with serious diseases and injuries, an increasing number of persons live longer with a wide range of potentially disabling conditions (Lamas, 2014). Predisease conditions such as obesity are also increasingly recognized as potentially disabling. Overall, there is a progressive shift toward a larger share of the global burden of disease from disability (Murray and Lopez, 2013). The worldwide phenomenon of population aging also has far-reaching implications for the continued development of the sociology of disability. Because every individual will age and is likely to experience one or more physical challenges or chronic illness, accepting the prospect of a disabling condition is the acceptance of an inevitable part of oneself. Changes associated with the aging process thus represent another sense in which normality and disability form a continuum (Blaxter, 1976). As Zola (1989) was early to recognize, population aging, the growing prevalence of chronicity, and the globalization of health risks are important reasons to universalize supportive and equitable disability policy. Disability is socially defined, thus the environment and the expectations for individuals' performance can be redefined, as well as the ways in which individuals with impairments react to the contexts and expectations that they encounter.

See also: Aging and Health in Old Age; Disability Movement; Disability and Disabling Practices; Disability, Geography of; Disability: Social and Psychological Aspects; Law and People with Disabilities.

Bibliography

Albrecht, G.L., 1976. Socialization and the disability process. In: Albrecht, G.L. (Ed.), The Sociology of Physical Disability and Rehabilitation. University of Pittsburgh, Pittsburgh, pp. 3–38.

Albrecht, G.L., 1992. The Disability Business. Sage, Newbury Park.

Baker, D.L., 2011. The Politics of Neurodiversity: Why Public Policy Matters. Lynne Rienner, Boulder.

Barnartt, S., Scotch, R., 2002. Disability Protests: Contentious Politics, 1970–1999. Gallaudet University, Washington, DC.

Becker, H., 1963. Outsiders: Studies in the Sociology of Deviance. Free Press, New York.

Blaxter, M., 1976. The Meaning of Disability: A Sociological Study of Impairment. Neale Watson Academic Publications, Inc., New York.

Charmaz, K., 1987. Struggling for a self: identity levels of the chronically ill. In: Roth, J.A., Conrad, P. (Eds.), Research in the Sociology of Health Care, vol. 6. JAI Press, Greenwich, pp. 283–321.

Corbin, J., Strauss, A.L., 1985. Managing chronic illness at home. Qualitative Sociology 8, 224–247.

Darling, R.B., 2013. Disability and Identity: Negotiating Self in a Changing Society. Lynne Rienner, Boulder.

Deegan, M.J., Brooks, N.A. (Eds.), 1985. Women and Disability: The Double Handicap. Transaction Books, New Brunswick.

Gill, C., 2001. Divided understandings: the social experience of disability. In: Albrecht, G.L., Seelman, K.D., Bury, M. (Eds.), The Handbook of Disability Studies. Sage, Thousand Oaks, pp. 351–372.
Gliedman, J., Roth, W., 1980. The Unexpected Minority. Harcourt Brace Jovanovich, New York.
Goffman, E., 1963. Stigma: Notes on the Management of Spoiled Identity. Prentice Hall, Englewood Cliffs.
Gove, W.R., 1976. Societal reaction theory and disability. In: Albrecht, G.L. (Ed.), The Sociology of Physical Disability and Rehabilitation. University of Pittsburgh, Pittsburgh, pp. 57–71.
Hahn, H., 1994. The minority group model of disability: implications for medical sociology. In: Weitz, R., Kronenfeld, J.J. (Eds.), Research in the Sociology of Health Care, vol. 11. JAI Press, Greenwich, pp. 3–24.
Hahn, H., Belt, T.L., 2004. Disability identity and attitudes toward cure in a sample of disabled activists. Journal of Health and Social Behavior 45, 453–464.
Higgins, P.C., 1992. Making Disability: Exploring the Social Transformations of Human Variation. C.C. Thomas, Springfield.
Hunt, P. (Ed.), 1966. Stigma: The Experience of Disability. Chapman, London.
Lamas, D., 2014. Chronic critical illness. New England Journal of Medicine 370 (2), 175–177.
Murray, C.J.L., Lopez, A.D., 2013. Measuring the global burden of disease. New England Journal of Medicine 369 (5), 448–457.
Omansky, B., 2011. Borderlands of Blindness. Lynne Rienner, Boulder.
Pope, A.M., Tarlov, A.R. (Eds.), 1991. Disability in America. National Academy Press, Washington, DC.
Ridolfo, H., Ward, B.W., 2012. Mobility Impairment and the Construction of Identity. Lynne Reinner, Boulder.
Roth, J.A., 1963. Timetables. Bobbs-Merrill, Indianapolis.
Safilios-Rothschild, C., 1970. The Sociology and Social Psychology of Disability and Rehabilitation. Random House, New York.
Safilios-Rothschild, C., 1976. Disabled persons' self-definitions and their implications for rehabilitation. In: Albrecht, G.L. (Ed.), The Sociology of Physical Disability and Rehabilitation. University of Pittsburgh, Pittsburgh, pp. 39–56.
Schneider, J.W., Conrad, P., 1983. Having Epilepsy: The Experience and Control of Illness. Temple University, Philadelphia.
Scotch, R.W., 1989. Politics and policy in the history of the disability rights movement. Milbank Quarterly 67 (Suppl. 2 Pt 2), 380–400.
Scotch, R.W., 2001. From Good Will to Civil Rights: Transforming Federal Disability Policy, second ed. Temple University Press, Philadelphia.
Scott, R.A., 1969. The Making of Blind Men. Russell Sage Foundation, New York.
Stone, D.A., 1984. The Disabled State. Temple University, Philadelphia.
Stroman, D.F., 1982. The Awakening Minorities. University Press of America, Washington, DC.
Williams, G., 2001. Theorizing disability. In: Albrecht, G.L., Seelman, K.D., Bury, M. (Eds.), The Handbook of Disability Studies. Sage, Thousand Oaks, pp. 123–144.
World Health Organization (WHO), 1980. International Classification of Impairments, Disabilities and Handicaps. WHO, Geneva.
Zola, I.K., 1982. Missing Pieces: A Chronicle of Living with a Disability. Temple University, Philadelphia.
Zola, I.K., 1989. Toward the necessary universalizing of disability policy. Milbank Quartely 67, 401–426.
Zola, I.K., 1991. Bringing our bodies and ourselves back in: reflections on a past, present, and future 'Medical Sociology'. Journal of Health and Social Behavior 32, 1–16.
Zola, I.K., 1993. Disability statistics, what we count and what it tells us. Journal of Disability Policy Studies 4, 9–29.

Disasters in Social, Cultural and Political Context

Julie Drolet, Faculty of Social Work, University of Calgary, Edmonton, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents key concepts and frameworks that inform disaster management, as well as describes how natural and human-made hazards impact individuals, families, and communities, in diverse ways by considering the local social, economic, and cultural context. Interventions by disaster responders and social workers illustrate the complexities inherent in disaster preparedness, relief, recovery, and mitigation.

Introduction

This article provides a comprehensive overview of 'natural' and 'human-made' disasters, with a discussion of various categories of hazards, as well as the key concepts and frameworks that inform disaster management at local and international levels. Consideration of the underlying social, economic, and cultural contexts and processes that mediate the impact natural hazards and disasters have on certain populations brings forth the needs for human rights, sustainable development, and disaster risk reduction. For many countries, 'natural' and 'human-made' disasters are a relatively common occurrence. The 2004 Indian Ocean and 2011 Japanese tsunamis; earthquakes in China, Haiti, and Japan; flooding in Pakistan, China, and Mozambique; and tropical cyclones in Asia, the Caribbean, the USA, and Central America; along with a continuing series of droughts and epidemics, have affected hundreds of thousands of individuals, families, and communities. Interventions by disaster responders, volunteers, and social workers illustrate the complexities of humanitarian responses.

Natural disasters affect the structures of society by disrupting economic development, increasing income and wealth inequality, marginalizing certain groups, and leading to large-scale migrations (Nel and Righarts, 2008: p. 161). Disaster, social science, and social work researchers are contributing to a growing literature on the relationship between disasters, risk, and vulnerability better to understand the social and economic impacts of disasters, and how natural hazards impact diverse populations between and within societies (Sim et al., 2013; Larson et al., 2013; Mathbor, 2007; Gray et al., 2013; Alston and Whittenbury, 2013; Dominelli, 2012). The terms 'vulnerability,' 'capacity,' and 'resilience' became popular in both the disaster and development literature, and have been applied to policies, with debate and interpretation among various schools of thought (Gaillard, 2010). Available data suggest a pattern of gender differentiation at all levels of the disaster process: exposure to risk, risk perception, preparedness, response, physical impact, psychological impact, and recovery and reconstruction (WHO, 2013). Further, disasters occur each year in locations without sufficient capacity or resources to prevent losses in human life and to support short- and long-term recovery (Alcántara-Ayala, 2002).

Vulnerability

The relationship between hazard and vulnerability has been a concern of disaster practitioners, volunteers, and academics for some time. Hazard refers to the actual physical process involved, while vulnerability is the compounded effect of the exposure and susceptibility of people and their assets to the impact of the hazard, as well as their capacity to anticipate, cope, and recover (Wisner et al., 2003). Ways to minimize risk have been influenced by conceptual frameworks on the unequal distribution of power, resources, wealth, knowledge, and perceptions of risk. The term 'vulnerability' is used to understand how one's status, or social location can influence different impacts of natural hazards. Every individual is influenced by a number of factors that mediate how natural hazards or disasters impact the person: gender, age, physical and mental health status, occupation, marital status, sexuality, race, ethnicity, culture, religion, and immigration status. There are many definitions of vulnerability, which are influenced by social, economic, health, and other perspectives and sectors (O'Brien and Leichenko, 2007; Wisner et al., 2003). Marginalized and vulnerable populations face challenges and struggles in disaster recovery. Older people with fragile health are often weakened when faced with difficult social and environmental changes. People living on fixed or limited incomes may lose important assets, which may lead to destitution. The most socially excluded and economically insecure in any society or community are least able to access or control resources needed before, during, and in the aftermath of a disaster (Larson et al., 2013).

The economic impacts of major 'natural' and 'human-made' disasters are profound for women and men, and gender-differentiated (Alston, 2010; Enarson and Morrow, 1998) as livelihoods, workplaces, and assets are destroyed. The loss of economic resources and deterioration of economic status is most striking on women, who live on the margins of survival, having lived with the 'daily disaster' of poverty before, during, and after natural disasters in many countries (Enarson, 2000). According to the Centre for Research in the Epidemiology of Disasters, since the 1950s, the cost of 'natural' disasters has risen 14-fold. The United Nations estimates that significant economic losses from disaster events around the world now surpass $100 billion each year. Durable and sustained efforts from all actors, such as governments, nongovernmental

organizations and civil society, international agencies and organizations, academic and technical institutions, and the private sector, share responsibility for reducing risks and strengthening resilience.

Capacities refer to the resources and assets that people possess to resist, cope with, and recover from disaster shocks they experience (Davis et al., 2004). Building or enhancing capacities can include activities at the household or community level that strengthen people's strategies or abilities, such as early warning, emergency planning, and preparing resources to cope with disruptions, to face natural hazards.

Building Resiliency

Increased 'natural' and 'human-made' disasters challenge a nation's abilities to mitigate and recover from major events. With the increased frequency, magnitude, and complexity of these events, it is increasingly recognized that prior planning and policies are needed to make communities and populations more resilient to these disruptions. The concept of resilience is currently the subject of research and international debate (Cottrell, 2012; Seville, 2009; Kayser et al., 2008). The Resilience Research Centre (2013) at Dalhousie University defines resilience as "the capacity of people to navigate to the resources they need to overcome challenges, and their capacity to negotiate for these resources so that they are provided in ways that are meaningful."

The *Hyogo Framework for Action 2005–2015: Building the Resilience of Nations and Communities to Disasters* is a 10-year plan endorsed by the United Nations General Assembly in Resolution A/RES/60/195 following the 2005 World Disaster Reduction Conference held in Kobe in Hyogo, Japan. The Hyogo Framework for Action (HFA) aims to substantially reduce disaster loss of lives and social, economic, and environmental assets due to hazards by building the resilience of nations and communities to disasters. The term 'disaster risk reduction' (DRR) refers to the process of understanding, analyzing, and managing the causes and origins of disasters and the risks that grow and lead to disasters. Support for mainstreaming DRR gained momentum in the mid-to-late 1990s, with growing support to integrate disaster risk into development planning and practice. While physical and natural science perspectives contribute knowledge to understanding hazards, there remains an important consideration to understand and reduce human and social vulnerabilities related to disasters. Decision makers and all levels of government, from the international to local municipal levels, face a dilemma in determining priorities for mitigating disaster risks of all kinds. Stakeholders, including policy-makers, have a responsibility to identify and listen to local knowledge when making decisions about disaster risk reduction.

'Natural' Disasters

Many social scientists now reject the popular terminology of 'natural disasters' and stress that while 'hazards' are natural, 'disasters' are not (Horton et al., 2008). A disaster occurs when an extreme event exceeds a community's ability to cope with that event (Lindell and Prater, 2003; Prater et al., 2004; Pyles, 2007). To understand the process by which disasters produce community impacts, it is important to identify the preimpact conditions that make communities vulnerable to disasters. The role of vulnerability as a causal factor in disaster losses tends to be less well understood (Dilley et al., 2005). The destructive power of natural hazards combined with vulnerabilities across various elements can lead to large-scale losses during hazard events in areas where population and economic investment are concentrated (Dilley et al., 2005: p. 23). While losses are experienced at the individual level, in a disaster situation, economic and social systems break down partly or completely, leading to higher socioeconomic impacts.

Climate change is one of the most important environmental concerns of the early twenty-first century with a profound impact on natural hazards. Climate change is caused by natural processes and human activity (IPCC, 2007). Global warming, or increasing average global temperatures, will result in severe consequences for food production, water scarcity, ecosystems, and extreme weather events (McKinnon, 2008). There is a growing literature in social work that considers the role of social workers in climate change adaptation and mitigation (Drolet, 2012; Alston and Whittenbury, 2013; Hetherington and Boddy, 2013; Tester, 2013; Alston, 2010). Natural hazards arise over all time scales, lasting from a few seconds or minutes (earthquakes), to hours (storms), to days (floods), to years or decades (droughts). Different groups have collected data on natural hazards for different purposes. The EM-DAT data set maintained by the Centre for Research on the Epidemiology of Disasters (CRED) in Brussels is the most comprehensive global database on natural hazards and their impacts.

Natural Hazards

This section will consider four categories of natural hazards: (1) hydrometeorological or climatological hazards; (2) geophysical hazards; (3) biological/ecological hazards; and (4) astronomical hazards.

Hydrometeorological or Climatological Hazards

Hydrometeorological or climatological hazards arise from water, weather, and climate. Examples include avalanches, coastal storms, thunderstorms, tornados, floods, droughts, temperature extremes or fluctuations (heat and cold), wildfires, and wind. Severe weather contributes to natural hazards, which can become disasters in many parts of the world.

Hurricane Katrina struck the US Gulf Coast in August 2005, resulting in large-scale destruction due to surge flooding and caused massive overtopping and breaching of levees around New Orleans. Damages to residential property, public structures, and infrastructure and utilities had major economic impacts, along with numerous physical and health impacts. Many of the world's megacities can be found in coastal areas, such as Jakarta, Mumbai, London, and New York. Coastal storms and floods lead to injuries and deaths, with damages to property such as residences, infrastructure, and public facilities, crops, and livestock. Damaged ecosystems disrupt livelihoods and social routines. Community organizing abounds in

postdisaster New Orleans to mobilize on housing issues, neighborhood planning, social and economic development, as well as race and class inequities, to foster social development (Pyles, 2007).

Flooding poses tremendous danger to both people and property. There are several types of flooding: overbank flooding, when the water within a river overflows its banks and spreads across the land around it; flash floods, created by a rapid rise of fast-moving water; ice jam flooding, caused by heavy precipitation pushes ice together and creates a jam; coastal flooding along the edges of oceans driven by storm surges and wave damage; and engineering issues, where a constructed dam gives way, causing a flash flood downstream. The most severe floods occurred in China when the Huang He (Yellow) River exceeded its banks in 1931, contributing to the deaths of 1–4 million people.

Global warming is increasing the frequency and intensity of extreme heat events, or heat waves. Extreme temperatures can have public health consequences and can be life threatening. The 2003 European heat wave led to heat-related illnesses that resulted in many thousands of deaths, of which 14 947 were in Paris, France, alone (Poumadère et al., 2005). Temperatures hovered around 104 °F (40 °C) for seven consecutive days. Many of the victims were older people, who lacked air conditioning and access to coordinated health services, and who were faced with insufficient care during the summer vacation period in August. Environmental experts warn that because of climate change, such heat waves are expected to increase in number in the coming years.

Wildfires are a natural hazard in any forested and grassland region. A wildfire is an unplanned fire in wilderness vegetation or bush, including grass fires and forest fires. The term 'bushfire' is used in Australia, while Canada and the USA refer to 'wildfire.' Devastating forest wildfires are concentrated in summer months when the temperature is high and air humidity and fuel moisture are low. There is a trend toward larger and more intense and damaging fires due to a massive build-up of fuel on the land, declining resources to fight fires, and the loss of traditional fire practices in some countries.

Geophysical Hazards

Geophysical hazards include earthquakes (and associated hazards such as tsunamis and landslides) that are geologically-based, erosion, landslides/rockslides, poisonous gases, sandstorms, soil loss and contamination, and volcanoes. Social workers and social services have significant roles and functions in both rescue and recovery and provide support to vulnerable groups, such as people with disabilities, children, and older people, as well as organize recovery programs that improve the community's ability to address its own unmet needs (Javadian, 2007).

An earthquake is the sudden, rapid shaking of the earth, caused by the breaking and shifting of subterranean rock as it releases strain that has accumulated over a long time. It is estimated that about 5000 earthquakes occur each year. Most of the world's earthquakes take place in the 40 000 km long Pacific seismic belt known as the Pacific Ring of Fire. The greatest risk in an earthquake is the severity of the shaking and ground rupture, resulting in more or less severe damage to man-made buildings and natural structures and the contents within these that may fail or fall and injure or kill people. For example, the 2010 earthquake in Haiti resulted in over 220 000 casualties, and thousands of collapsed structures. Earthquakes that cause the greatest loss of life are deadly because of their close proximity to populated areas or the ocean, where earthquakes often create tsunamis that can devastate communities thousands of kilometers away. Social workers and social services play key roles in earthquake relief by assisting survivors on the frontline of emergency responses (Javadian, 2007).

Tsunamis are long-wavelength, long-period sea waves produced by the sudden or abrupt movement of large volumes of water. A tsunami can travel 600–800 km per hour, depending on water depth. Large waves produced by an earthquake or a submarine landslide can overrun nearby coastal areas in a matter of minutes. Tsunamis can also travel thousands of kilometers across open ocean and wreak destruction on far shores hours after the earthquake that generated them. The 2004 Indian Ocean earthquake was an undersea earthquake that occurred on Sunday, 26 December 2004, off the coast of Sumatra, Indonesia. The earthquake triggered a series of devastating tsunamis along the coasts of most landmasses bordering the Indian Ocean, killing over 230 000 people in 14 countries, and inundating coastal communities with waves up to 30 m (98 ft) high. It was one of the most deadly natural disasters in recorded history and caused serious damage and deaths. The plight of the affected people and communities prompted a worldwide humanitarian response to address short- and long-term recovery. In all, the international community donated more than $14 billion (2004 US$) in humanitarian aid. Social workers have important and unique contributions to make to disaster recovery and reconstruction, research, and ethics through their expertise in ecological approaches, prevention, stress and coping, and promoting transformative change in micro, mezzo and macro systems (Zakour, 1997; Soliman and Rogge, 2002; Javadian, 2007).

Biological or Ecological Hazards

Biological or ecological hazards come from living organisms or ecosystems. Examples include epidemics, microbial pathogens, and poisonous or dangerous plants and animals.

An epidemic may be considered a disaster due to mortality, economic loss, and social disruption, and may contribute to food shortages when disease-affected people cannot work. Human epidemics occur when new cases of a certain disease substantially exceed what is expected in a given human population and during a given time period. Throughout history humans have been affected by epidemics of infectious disease such as the Black Death (bubonic plague) or 1918 flu pandemic. These devastating epidemic diseases affected different population groups in unequal ways due to the unequal distribution of power that produces vulnerabilities that infectious diseases can exploit (Drolet et al., 2013b).

Plant disease, pests, and erosion of biodiversity can affect the variety and variability of animals, plants, and microorganisms. The conservation and sustainable use of biodiversity for food and agriculture play a critical role in human security. Chronic famines have been correlated with the combination of

biological vulnerability and social vulnerability, as evidenced in the Irish Famine of 1845–48 (Ó Grada, 1999).

Astronomical Hazards

Astronomical hazards come from outside the earth, such as asteroids and comets, space weather, and collision or near-collision of celestial bodies with the Earth, and associated hazards (tsunamis, wildfires). The risk from hazards beyond the atmosphere is small compared to other natural hazards. Most asteroids in the solar system come nowhere near the Earth, and smaller ones burn up in the atmosphere. However, in February 2013, a small rock or meteorite disintegrated above the Russian city of Chelyabinsk, injuring about 1500 people. Numerous space hazards are portrayed in Hollywood films, such as 'Armageddon,' 'Impact,' and '2012 Supernova.'

'Man-Made' or 'Human-Made' Disasters

Anthropogenic hazards or human-made hazards can result in the form of a human-made disaster. The term 'anthropogenic' refers to threats with an element of human intent, negligence, or error; it can also involve a failure of a human-made system. A 'human-made' disaster can result in huge loss of life and property and affects a person's mental, physical, and social well-being. Examples include sociological hazards, technological hazards, and transportation events.

Sociological hazards are acts committed against the law and order of the government and can include crime, civil disorder, terrorism, and war. Technological hazards can include industrial hazards, structural collapse, power outage, fire, and hazardous materials. Transportation events can include aviation, rail, road, space, and sea travel.

Human Rights and Disaster

The Universal Declaration of Human Rights (1948) established that all people deserve to live in dignity and with respect. This basic principle is translated into a range of rights articulated by international treaties. For example, the International Covenant on Civil and Political Rights prohibits discrimination on any person because of race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status (ICCPR, 1966: Article 26). The United Nations Convention on the Rights of the Child (UNCRC) states that children and youth have the right to freedom of speech and the right to partake in decision-making processes that are relevant to their lives. The 1951 Refugee Convention and 1967 Protocol provide legal protection for refugees with a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group, or political opinion (UNHCR, 2001). Human rights relevant to disasters include the right to life, the right to move freely within and between states, freedom of movement, right to housing, and many other interrelated rights, such as the right to health and education. Governments have an obligation to respect, protect, and fulfill rights, even when threats to those rights arise from 'natural' and 'human-made' disasters. Social workers and other professionals play a role in reducing and preventing human rights violations and ensuring that policies and practices promote rights by using advocacy, applying legal standards, and promoting health, education, and social service programing.

Sustainable Development and Disaster Risk

The Brundtland report articulated a commonly accepted definition of sustainable development: "Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs" (United Nations, 1987). Social and environmental movements highlight the social justice dimensions of sustainability and the economic, social, and environmental dimensions in light of current inequalities and processes of exclusion.

It is increasingly being recognized that 'natural' and 'human-made' disaster risk must be addressed as a development issue rather than only humanitarian relief assistance, which calls upon governments and other stakeholders to mainstream disaster risk management into development work and activities. The enormous human and economic losses, as well as the costs of relief, rehabilitation, reconstruction, and recovery, make disasters an economic and humanitarian concern. For this reason, disaster risks need to be considered an issue for sustainable development. Social protection initiatives in a number of countries are bringing together humanitarian and development interventions in new ways by addressing vulnerability and risk in short- and long-term approaches through various public actions (Drolet, 2014).

Disaster Preparedness

Early warning principles and systems exist to provide information about a potential disaster to minimize impacts such as fatalities, injuries, damages, and interruptions of normal activities. It is acknowledged that involving communities in people-centered approaches as participants in early warning systems is important. Community members involved in designing warning messages can improve the public's understanding of warning for different target audiences. The United Nations Office for Disaster Risk Reduction (UNISDR) defines preparedness as "the knowledge and capacities developed by government, professional response and recovery organizations, communities and individuals to effectively anticipate, respond to, and recover from, the impacts of likely, imminent or current hazard events or conditions" (UNISDR, 2009). Emergency management aims to manage and control a major emergency in order to reduce the negative impact to life, property, infrastructure, and livelihoods. Stakeholders in emergency response include emergency services, local authorities, public services, nongovernmental organizations, the community, and in some countries, the military and armed forces (Palmore, 2008).

Disaster Relief

The terms 'disaster relief' or 'humanitarian response' cover the immediate response and early recovery. The first and most

important form of relief in the period immediately following a disaster is the provision of basic needs (food, water, clothes, and shelter) to those most seriously impacted. The primary challenge in providing relief is the lack of accurate and reliable information in the early stages. Damage and needs assessments are undertaken in order to provide relief assistance. Social workers design interventions to enhance individual and communal coping skills, stabilize family systems, maximize support, facilitate return to normalcy or 'new normal,' promote and strengthen human rights and social justice, and engage in community development (Kayser et al., 2008; Bourassa, 2009). Social workers identify and address the special needs of vulnerable populations in disaster situations and facilitate emotional expression of grief and loss by listening to survivors' stories, providing psychosocial support, and assisting in finding meaning out of tragedy (Ng, 2012; Bourassa, 2009). Assistance must be offered in a manner consistent with the cultural context in order to support survivors' coping efforts and to build on their strengths and facilitating inclusion of marginalized populations (Kayser et al., 2008; Bourassa, 2009). Social workers build on existing local systems by identifying resources and supports from within the local community. Collaboration with community level workers and volunteers facilitates anti-oppressive and culturally competent services and programs (Bradshaw and Graham, 2007). Recent research on disasters emphasize the significance of social resources in promoting coping and resilience and stress the importance of considering the larger societal context when addressing large-scale disasters. This includes recognition that the individual is embedded in a broader familial, interpersonal, and social context. Social workers are well-situated to develop social interventions to provide individuals, families, and communities with appropriate resources that will support short and long-term adaptation.

Disaster Recovery

For those affected by disasters, following the immediate trauma and shock, the urgent question is how to rebuild lives and livelihood options (Ruwanpura, 2008). Disaster recovery is often handled by a variety of organizations at different levels of government and society, including volunteer groups, national civilian and military agencies, and international relief agencies and nongovernmental organizations. Each brings its own areas of expertise and a number of limitations in terms of jurisdiction and operations. In a large-scale disaster, the recovery process can go beyond the capacity of national governments, and the international community will support recovery by providing aid in cash and kind. The International Federation of Red Cross and Red Crescent Societies (IFRC) is the world's largest humanitarian organization. The IFRC vision is "to inspire, encourage, facilitate and promote at all times all forms of humanitarian activities by National societies, with a view to preventing and alleviating human suffering, and thereby contributing to the maintenance and promotion of human dignity and peace in the world" (IFRC, 2013).

The poor, women, children, older people, and the disabled face more challenges during disasters (Thomas and Twynam, 2006). Social, economic, and political factors determine who lives, works, and has assets in high-risk zones, and shape people's capacities to cope, recover, and adapt. Often women suffer more in disaster events and bear heavier physical and psychological burdens during the recovery period as compared to men. A gender perspective on disasters is necessary to address gender discrimination in everyday life and inequity in times of disasters (Ikeda, 2009; Stehlik et al., 2000). Women face vulnerabilities that result in less economic resources and social protection, which affects their ability to recover from disasters when they occur. Recent research has demonstrated that the inequitable impact of disasters on women is not innate or biological, but socially constructed (Fordham, 2003; Adger and Kelly, 1999). Sexual orientation is another factor in considering gender and disasters; in Indonesia the *warias*, members of the lesbian, gay, bisexual, and transgender community, contributed to disaster risk reduction despite discriminatory practices (Balgos et al., 2012). Old people may be frail and lack mobility, and may also be ill or susceptible to various diseases and conditions before, during, and after disasters (Pekovic et al., 2008). People with disabilities may encounter physical barriers and challenges during evacuations (Christensen et al., 2007). There is a growing body of literature, training tools, manuals, and programs providing information on how to prepare and assist marginalized and vulnerable populations (Tang and Cheung, 2007; Morrow, 1999). The role of social capital, which is defined as a function of trust, social norms, participation, and network, can play an important role in recovery and sustainable disaster mitigation to enhance community preparedness (Ersing and Kost, 2012; Mathbor, 2007; Nakagawa and Shaw, 2004). There is a need to reduce underlying social vulnerabilities through social protection and social policies (Drolet et al., 2013a).

The psychosocial impacts in disaster recovery have emerged as a new area of practice in humanitarian response to address the short and long-term mental health and well-being of affected community members. Many individuals and groups experience stress, and loss and grief, which require needs assessments, support services, and emotional assistance. This can include the psychological needs of disaster workers and emergency responders (Cronin et al., 2007; Smith et al., 2003). The Inter-Agency Standing Committee (IASC) Guidelines on Mental Health and Psychosocial Support in Emergency Settings are used by humanitarian organizations and communities to plan, establish, and coordinate a set of responses to protect and improve people's mental health and psychosocial well-being in emergency situations (IASC, 2007).

Conclusion

There are a number of stakeholders in disaster recovery, including government, academia, communities, business, and civil society. 'Natural' and 'human-made' disasters have impacts on health, food security, and housing and shelter. The quality of the built environment is one of the critical factors that underlie the overall vulnerability of populations exposed to earthquake and tsunami hazards. One of the challenges facing decision-makers and stakeholders is the

recognition of short-term vs. long-term benefits in disaster reconstruction. Media images from recent disasters like the 2004 Indian Ocean tsunami have heightened awareness of natural hazards. The meaning of 'build back better' or 'building back better' is explored in applied research on postdisaster settlement and shelter in terms of involving community, building capacity, and planning ahead for long-term disaster risk reduction (Kennedy et al., 2008). Yet, despite this increased awareness, thousands of lives continue to be lost in subsequent disaster events, and there is more work to be done to mitigate against future events. Local communities offer local knowledge of hazards and can identify their vulnerabilities and needs to strengthen mitigation efforts focused on the natural, built, economic, and social environments. By making connections between the hazard and the vulnerable within local communities, disaster risk reduction is tenable. Disaster preparedness can provide an opportunity to improve safety, enhance equity, and mitigate 'natural' and 'human-made' disasters.

See also: Community Resilience to Disasters; Demographic Impact of Disasters; Emergency Relief and Social Work Responses; Green Social Work; Hazards and Disaster Research in Contemporary Anthropology; Risk Assessment in Social Work; Rural Social Work; Social Work Theory; Social Work and Disasters.

Bibliography

Adger, W.N., Kelly, M., 1999. Social vulnerability to climate change and architecture of entitlements. Migration and Adaptation Strategies for Global Change 4, 253–266.

Alcántara-Ayala, I., 2002. Geomorphology, natural hazards, vulnerability and prevention of natural disasters in developing countries. Geomorphology 47 (2–4), 107–124.

Alston, M., 2010. Gender and climate change in Australia. Journal of Sociology 47 (1), 53–70. http://dx.doi.org/10.1177/1440783310376848.

Alston, M., Whittenbury, K. (Eds.), 2013. Research, Action and Policy: Addressing the Gendered Impacts of Climate Change. Springer, Dordrecht.

Balgos, B., Gaillard, J.C., Sanz, K., 2012. The *warias* of Indonesia in disaster risk reduction: the case of the 2010 Mt Merapi eruption in Indonesia. Gender & Development 20 (2), 337–348.

Bourassa, J., 2009. Psychosocial interventions and mass populations: a social work perspective. International Social Work 52 (6), 743–755.

Bradshaw, C., Graham, J.R., 2007. Localization of social work practice, education, and research: a content analysis. Social Development Issues 29 (2), 92–111.

Christensen, K.M., Blair, M.E., Holt, J.M., 2007. The built environment, evacuations, and individuals with disabilities: a guiding framework for disaster policy and preparation. Journal of Disability Policy Studies 17 (4), 249–254.

Cottrell, A., 2012. Engaged and resilient communities: AEMI workshop. The Australian Journal of Emergency Management 26 (4), 9–13.

Cronin, M.S., Ryan, D.M., Brier, D., 2007. Support for staff working in disaster situations: a social work perspective. International Social Work 50 (3), 370–382.

Davis, I., Haghebeart, B., Peppiatt, D., 2004. Social Vulnerability and Capacity Analysis. ProVention Consortium, Geneva.

Dilley, M., Chen, R.S., Deichmann, U., Lerner-Lam, A.L., Arnold, M., 2005. Natural Disaster Hotspots: A Global Risk Analysis. The World Bank, Washington.

Dominelli, L., 2012. Green social Work: From Environmental Crisis to Environmental Justice. Polity Press, Cambridge.

Drolet, J., 2012. Climate change, food security, and sustainable development: a study on community-based responses and adaptations in British Columbia, Canada. Community Development. http://dx.doi.org/10.1080/15575330.2012.729412.

Drolet, J., Ayala, J., Pierce, J., Giasson, F., Kang, L., 2013a. Influenza A H1N1 pandemic planning and response: the role of Canadian social work field directors and coordinators. Canadian Social Work Review 30 (1), 49–63.

Drolet, J., Sampson, T., Jebaraj, D., Richard, L., 2013b. Social work and environmentally induced displacement: a commentary. Canada's Journal on Refugees. Refuge 29 (2), 55–62.

Drolet, J., 2014. Social Protection and Social Development: International Initiatives. Dordrecht, Springer.

Enarson, E., 2000. Gender and Natural Disasters in Focus Programme on Crisis Response and Reconstruction [Online] Available at: http://www.ilo.int/wcmsp5/groups/public/—ed_emp/—emp_ent/—ifp_crisis/documents/publication/wcms_116391.pdf (accessed 20.07.13.).

Enarson, E., Morrow, B.H., 1998. The Gendered Terrain of Disasters: Through Women's Eyes. Praeger Publishers, Miami.

Ersing, R.L., Kost, K.A. (Eds.), 2012. Surviving Disaster: The Role of Social Networks. Lyceum Books, Chicago.

Fordham, M., 2003. Gender, disaster and development. In: Pelling, M. (Ed.), Natural Disaster and Developing in a Globalizing World. Routledge, London, pp. 57–74.

Gaillard, J.C., 2010. Vulnerability, capacity and resilience: perspectives for climate and development policy. Journal of International Development 22, 218–232.

Gray, M., Coates, J., Hetherington, T. (Eds.), 2013. Environmental Social Work. Routledge, New York.

Hetherington, T., Boddy, J., 2013. Ecosocial work with marginalized populations: time for action on climate change. In: Gray, M., Coates, J., Hetherington, T. (Eds.), Environmental Social Work. Routledge, New York, pp. 46–61.

Horton, B., Bird, M., Birkland, T., Cowie, S., Eong, O.J., Hawkes, A., Khoon, G.W., Law, L., Macgregor, C., Shau-Hwai, A.T., Sa, T.T., Yasin, Z., 2008. Environmental and socioeconomic dynamics of the Indian Ocean tsunami in Penang, Malaysia. Singapore Journal of Tropical Geography 29, 307–324.

Ikeda, K., 2009. How women's concerns are shaped in community-based disaster risk management in Bangladesh. Contemporary South Asia 17 (1), 65–78.

Inter-Agency Standing Committee (IASC), 2007. IASC Guidelines on Mental Health and Psychosocial Support in Emergency Settings. IASC, Geneva.

International Federation of Red Cross and Red Crescent Society, 2013. Our Vision and Mission [Online] Available at: http://www.ifrc.org/en/who-we-are/vision-and-mission/ (accessed 20.07. 13.).

IPCC, 2007. IPCC Fourth Assessment Report on Climate Change [Online] Available at: http://www.ipcc.ch/publications_and_data/ar4/wg1/en/faq-2-1.html (accessed 20.07.13.).

Javadian, R., 2007. Social work responses to earthquake disasters: a social work intervention in Bam, Iran. International Social Work 50 (3), 334–346.

Jones, D.N., Truell, R., 2012. The Global Agenda for Social Work and Social Development: a place to link together and be effective in a globalized world. International Social Work 55 (4), 454–472.

Kayser, K., Wind, L., Shankar, R.A., 2008. Disaster relief within a collectivistic context. Journal of Social Service Research 34 (3), 87–98.

Kennedy, J., Ashmore, J., Babister, E., Kelman, I., 2008. The meaning of 'build back better': evidence from post-tsunami Aceh and Sri Lanka. Journal of Contingencies and Crisis Management 16 (1), 24–36.

Larson, G., Drolet, J., Samuel, M., 2013. The role of self-help groups in post-tsunami rehabilitation. International Social Work. http://dx.doi.org/10.1177/0020872813477880.

Lindell, M.K., Prater, C.S., 2003. Assessing community impacts of natural disasters. Natural Hazards Review 4, 176–185.

Mathbor, G., 2007. Enhancement of community preparedness for natural disasters: the role of social work in building social capital for sustainable disaster relief and management. International Social Work 50 (3), 357–369.

McKinnon, J., 2008. Exploring the nexus between social work and the environment. Australian Social Work 61 (3), 256–258.

Morrow, B.H., 1999. Identifying and mapping community vulnerability. Disasters 23 (1), 1–18.

Nakagawa, Y., Shaw, R., 2004. Social capital: a missing link to disaster recovery. International Journal of Mass Emergencies and Disasters 22 (1), 5–34.

Nel, P., Righarts, M., 2008. Natural disasters and the risk of violent civil conflict. International Studies Quarterly 52, 159–185.

Neumayer, E., Plümper, T., 2007. The gendered nature of natural disasters: the impact of catastrophic events on the gender gap in life expectancy, 1981–2002. Annals of the Association of American Geographers 97 (3), 551–566.

Ng, G.T., 2012. Disaster work in China: tasks and competences for social workers. Social Work Education: The International Journal 31 (5), 538–556.

O'Brien, K., Leichenko, R., 2007. Human Security, Vulnerability and Sustainable Adaptation. UNDP, New York [Online] Available at: http://ideas.repec.org/p/hdr/hdocpa/hdocpa-2007-09.html#download (accessed 23.09.13.).

Ó Grada, C., 1999. Black '47 and beyond: The Great Irish Famine in History, Economy, and Memory. Princeton University Press, Princeton, NJ.

Palmore, J., 2008. On setting priorities for international security II: economic and security implications of natural disasters. Defense & Security Analysis 24 (3), 325–327.
Pekovic, V., Seff, L., Rothman, M.B., 2008. Planning for and responding to special needs of elders in natural disasters. Generations 31 (4), 37–41.
Pittaway, E., Bartolomei, L., Rees, S., 2007. Gendered dimensions of the 2004 tsunami and a potential social work response in post-disaster situations. International Social Work 50 (3), 307–319.
Poumadère, M., Mays, C., Le Mer, S., Blong, R., 2005. The 2003 heat wave in France: dangerous climate change here and now. Risk Analysis 25 (6), 1483–1494.
Prater, C.S., Zhang, M., Arlikatti, S., Lu, J.C., Zhang, Y., 2004. Management of the Hurricane Lili Evacuation by Local Emergency Managers. A&M University Hazard Reduction & Recovery Center, Texas.
Pyles, L., 2007. Community organizing for post-disaster social development: locating social work. International Social Work 50 (3), 321–333.
Resilience Research Centre, 2013. About Resilience [Online] Available at: http://www.resilienceproject.org/ (accessed 23.09.13.).
Ruwanpura, K.N., 2008. Temporality of disasters: the politics of women's livelihoods 'after' the 2004 tsunami in Sir Lanka. Singapore Journal of Tropical Geography 29, 325–340.
Seville, E., 2009. Resilience: great concept but what does it mean for organisations? Tephra 22, 9–15.
Sim, T., Yuen-Tsang Woon Ki, A., Quen, C.H., Dong, Q.H., 2013. Rising to the occasion: disaster social work in China. International Social Work 56 (4), 544–562.
Smith, M., Lees, D., Clymo, K., 2003. The readiness is all. Planning and training for post-disaster support work. Social Work Education: The International Journal 22 (5), 517–528.
Soliman, H.H., Rogge, M.E., 2002. Ethical considerations in disaster services: a social work perspective. Journal of Social Work 1 (1), 1–23.
Stehlik, D., Lawrence, G., Gray, I., 2000. Gender and drought: experiences of Australian women in the drought of the 1990s. International Journal of Disaster Studies 24 (1), 38–53.
Tang, K., Cheung, C., 2007. The competence of Hong Kong social work students in working with victims of the 2004 tsunami disaster. International Social Work 50 (3), 405–418.
Tester, F., 2013. Climate change as a human rights issue. In: Gray, M., Coates, J., Hetherington, T. (Eds.), Environmental Social Work. Routledge, New York, pp. 102–118.
Thomas, D.S.G., Twynam, C., 2006. Adaptation and equity in resource dependent societies. In: Adger, W.N., Paavola, J., Huq, S., Mace, M.J. (Eds.), Fairness in Adaptation to Climate Change. The MIT Press, Cambridge, MA, pp. 223–228.
UNHCR, 2001. The wall behind which refugees can shelter: the 1951 Geneva Convention. Refugees 2 (123).
United Nations, 1987. Our Common Future [Pdf] Available at: http://conspect.nl/pdf/Our_Common_Future-Brundtland_Report_1987.pdf (accessed 20.07.13.).
United Nations Office for Disaster Risk Reduction (UNISDR), 2009. UNISDR Terminology on Disaster Risk Reduction [Pdf] Available at: http://www.unisdr.org/files/7817_UNISDRTerminologyEnglish.pdf (accessed 20.07.13.).
Wisner, B., Blaikie, P., Cannon, T., David, I., 2003. At Risk: Natural Hazards, People's Vulnerability and Disasters. Routledge, London.
World Health Organization (WHO), 2013. Gender and Health in Disasters [Online] Available at: http://www.who.int/gender/other_health/disasters/en/ (accessed 20.07.13.).
Zakour, M.J., 1997. Disaster research in social work. Journal of Social Service Research 22 (1–2), 7–25.

Relevant Websites

www.gdnonline.org – Gender and Disaster Network.
www.ipcc.ch – Intergovernmental Panel on Climate Change (IPCC).
www.ifrc.org/en – International Federation of Red Cross and Red Crescent Societies (IFRC).
www.unfccc.int – United Nations Framework Convention on Climate Change (UNFCCC).
www.unocha.org – United Nations Office for the Coordination of Humanitarian Affairs (OCHA).
www.unisdr.org – The United Nations Office for Disaster Risk Reduction (UNISDR).
www.wmo.int – World Meteorological Organization.

Discipline Formation in the Social Sciences

Bjorn Wittrock, Swedish Collegium for Advanced Study in the Social Sciences, Uppsala, Sweden

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 3721–3728, © 2001, Elsevier Ltd.

Abstract

The history of the social science disciplines may be highlighted in terms of four epochal transformations. First, in the late eighteenth and early nineteenth centuries a deep shift occurred from broad genres of discourse on moral and political philosophy into proto-disciplines of a more delimited nature. Inherent in this shift was the constitution of a set of new basic conceptualizations of human agency and human society that came to be of perennial importance. Second, in the late nineteenth and early twentieth centuries a range of inquiries into the social question and into the conditions of institutional and administrative reform were gradually and unevenly transformed into university-based social science disciplines. Third, in the 1920s and 1930s several of these disciplines went through a process of intellectual and institutional consolidation. In this process research programs were formulated that came to serve as focal points for decades to come. Other research programs, often of a transdisciplinary nature, came to occupy a prominent position as intellectual points of reference without finding a secure institutional basis either in an increasingly authoritarian European setting, or in the discipline-based American academic landscape. Fourth, the first three decades of the period after the Second World War were characterized by secular processes, often inspired by American examples, towards disciplinary internationalization as well as professionalization. The social science disciplines also became increasingly subjected to demands to demonstrate their policy-relevance in a broad sense. In the present academic landscape, finally, the social science disciplines are more firmly entrenched institutionally than ever before. At the same time their intellectual and scholarly rationale is being re-examined and there is renewed interest in a dialog about their relationships to each other, to other scholarly and scientific pursuits and to society at large.

Social Science and Disciplinary History

The social nature of human beings means that there have always been rules regulating the life of a community. Tacit social knowledge exists in any human community. The whole of human history is filled with examples of efforts to give guidance to rulers, sometimes written down in books or manuals. Some contemporary social scientists try to trace a lineage of reasoning reaching back hundreds, if not thousands, of years. Thus political scientists often portray Plato and Aristotle as early representatives of their discipline. Economists and educational researchers may point to the perennial nature of the aspects of human life that constitute their scholarly domains. In the case of legal scholarship a tradition, if only in a weak sense of the word, may be said to obtain from at least the twelfth and thirteenth centuries. The emergence of this tradition is in fact coterminous with the creation of European universities (Kelley, 1990). Examples of this kind are interesting in their own right. However none of them amounts to a disciplinary social science history in any real sense of the word. This is so for three reasons (which provide the rationale for the structure of this article).

First, the concept of social science appears only in the 1790s and its use presupposes a meaningful conceptualization of something called a society. This does not occur before the second half of the eighteenth century.

Second, in none of these cases can we talk of a discipline in the sense of a relatively coherent and delimited program of research and teaching that is consolidated and consistently reproduced in a university environment. This presupposes the existence of a university, but also that the university is seen as a primary vehicle for research activities. However, the research-orientated university does not emerge as a key societal institution until the second half of the nineteenth century.

Third, in none of the cases mentioned above is there an unbroken scholarly tradition that links these early efforts to present-day activities in university settings. Such intellectual and institutional continuities presuppose a degree of disciplinary consolidation that does not occur on a more general level until the early twentieth century. In this respect, developments in the 1930s are crucial although limited to Europe and the Americas. Elsewhere, disciplinary consolidation does not occur until after World War II.

The Rise of the Social Sciences: From Moral and Political Philosophy to Social Science

The first recorded uses of 'social science' appear in France in the 1790s. Ever since then, these new forms of knowledge have been characterized by an effort to understand the world of modernity. They have tended to describe key features of this new world in terms of processes of industrialization, urbanization, and political upheaval, originating at the northwestern edge of the Eurasian landmass but eventually having global repercussions. In the self-understanding of the social sciences, accordingly, there is a longstanding and predominant view about the formation of modernity, which highlights transformations in the late eighteenth and early nineteenth centuries. It sees these transformations as a conjunction of a technological and a political transformation – the industrial and the democratic revolutions, respectively. This traditional interpretation, however, radically underestimates the deep-seated epistemic transformation that occurred in the same period.

In fact, in this period there is a fundamental transition from earlier forms of moral and political philosophy into social science. This transition is linked to an institutional restructuring not only in forms of political order but also in the forms in which human knowledge is brought forth and claims to validity are ascertained. One feature of this institutional transition is the emergence of a public sphere that gradually replaces arenas of a more closed nature such as aristocratic literary salons (*see* Civil Society/Public Sphere: History of the Concept). Another is the rise of new or reformed public higher education and research institutions that come to replace both the laboratories of wealthy amateurs and the academies under royal patronage and partial control (*see* Social Science and Universities).

The rise of the social science disciplines must then be cast in terms of the fundamental transformation of European societies that the formation of modernity entailed. One intellectual and cultural transformation in this period pertains precisely to the concepts of society and history and to the new awareness of the structural and constraining nature of societal life. Pierre Manent has put forward the notion that society is a 'postrevolutionary discovery' (Manent, 1998, p. 51). It is true that the term 'society' underwent a long conceptual development in the French context in the course of the seventeenth and eighteenth centuries (see Baker, 1994) – with a dramatic increase in the utilization of the term in the mid-eighteenth century. However, even if there was a long process of gestation for the modern concept of society, the unique event of revolutionary upheaval requires that discursive controversy and political practice become joined in the formation of a distinctly modern era.

The late eighteenth century witnessed the creation of a political project encompassing the whole world and shattering the existing absolutist order. In this process horizons of expectation to use Koselleck's term opened up that were previously unknown. This sense of openness and contingency also served as a forceful impetus to an examination of the structural conditions of the political body and entailed a passage from political and moral philosophy to social science. This transition required that four key problematics – which today are more acutely open to reinterpretation than they have been for decades – be formulated or at least fundamentally reformulated and enter into the new social-science discourse.

First, the role of historical inquiry becomes crucial. Historical reasoning becomes an integral part of the intellectual transition, and even abstract reason itself becomes historicized in early nineteenth-century philosophy. However, the moral and political sciences break up into a variety of new discourses that in the course of the nineteenth century coalesce and are reduced to a number of disciplines. This means that the stage is set for the divergence between a professionalized historical discipline and the other social and human sciences, a divergence that we still today experience as a major intellectual divide.

Second, interest in language and linguistic analysis enters into all domains of the human and social sciences. One outflow of this is the constitution of textual and hermeneutic modes of analysis. There is also an effort to historicize language and linguistic development itself, thereby providing a crucial link with the historic construction of the notion of different peoples.

This leads to a third problematic: that of constituting new collective identities. Membership in a collectivity could no longer be taken for granted in the life experiences of the inhabitants of a certain village or region. Nor could a relationship of obligation and loyalty between the princely ruler and his subjects continue to constitute the unquestionable core of the body politic. That, however, meant that even the most basic categories of societal existence were open to doubt. In the western part of Europe, categories such as 'citizen' and 'compatriot' came to play an even more important role at the turn of the eighteenth century.

Fourth, assumptions about what prompts human beings to act and how to interpret their actions within a broader framework are at the very core of any scholarly program in the social and human sciences. At the turn of the eighteenth and nineteenth centuries, the fundamental categories that we still largely draw upon were elaborated and proposed. We might describe these categories as follows:

a. *Economic-rationalistic*, with a corresponding view of society as a form of compositional collective;
b. *Statistical-inductive*, with a view of society as a systemic aggregate;
c. *Structural-constraining*, with a view of society in terms of an organic totality; and
d. *Linguistic-interpretative*, with a conceptualization of society as an emergent totality.

The transition from a discourse of moral and political philosophy to a social science entailed a decisive shift from an agential – some would say voluntaristic – view of society to one that emphasized structural conditions. In economic theorizing this also entailed a shift away from a concern with moral agency. During the nineteenth century, the context of 'average economic man' became a web of structural properties and dynamic regularities rather than a moral universe of individual action.

Thus, fundamental categories of agency and society that came to be elaborated and refined during much of the rest of the nineteenth and twentieth centuries can be discerned in rudimentary form already during the great transition. So too, however, can some of the features that came to affect these endeavors.

One such tacit but crucial feature concerns the abandonment of the truly universal heritage of the Enlightenment project in favor of forms of representation and endowment of rights based on territoriality or membership in a linguistically and historically constituted and constructed community. Another feature was an emerging and growing chasm between moral discourse and other forms of reasoning about society. Thus, an earlier encompassing conception of the moral and political sciences was gradually replaced by social sciences that marginalized moral reasoning or consigned it to the specialized discipline of philosophy. Third, historical reasoning, which had been at the core of the intellectual transformation at the end of the eighteenth century, became a separate discipline and, toward the end of the nineteenth century, a permanent divide emerged between history and the social sciences.

The end of the eighteenth century was a formative period in the rise of the social sciences in conceptual terms. It is possible to discern, across all confrontations and divergences, a fundamental acknowledgement of the idea that agency, reflexivity, and historical consciousness might help construct a new set of institutions but that this takes place within a complex web of interactions that jointly constitute a society. Thus, there existed a limited number of thematic foci underlying the cultural constitution of a new set of societal macroinstitutions.

The Institutionalization of the Social Sciences: The Social Question, the Research University, and the New Nation-States

Relatively early in the nineteenth century, economics (or rather political economy) became differentiated from moral philosophy. It is also in this period that history emerges as a scholarly field with its own canon of rules, but the full disciplinary formation of history is a highly extended process.

However, social science as an institutionalized scholarly activity performed within a series of disciplines is largely a phenomenon of the late nineteenth and early twentieth centuries. It is an activity that directly and indirectly reflects concerns about the wide-ranging effects of the new industrial and urban civilization that was rapidly changing living and working conditions for ever-larger parts of the population in many European nations during the nineteenth century. These changes, often collectively referred to as 'the social question' (*die soziale Frage*), were gradually forcing themselves upon the agendas of parliamentary bodies, governmental commissions, and private reform-minded and scholarly societies (*see* Social Question: Impact on Social Thought). Often the impetus for the search for new knowledge came from modernizing political and social groupings that favored industrialization but advocated more or less far-reaching social reforms. These groupings gradually came to embrace the notion that political action to alleviate 'the social question' should be based on extensive, systematic, empirical analysis of the underlying social problems. In country after country, the political agenda of the nineteenth century was being formed by two macro-projects: the search for a solution to 'the social question' and to the question of national identity and nationhood within new or reformed nation states (Rueschemeyer and Skocpol, 1996; Wagner et al., 1991a,b).

Between 1870 and the turn of the century, all signs seemed to indicate that new social science approaches would gain access to the scientific institutions. The thinking of their proponents was widely diffused among the intellectual and political elites. This occurred at a time when traditional liberal economics was undergoing a crisis. Thus traditional political economy was seen to be unable to deal with the 'social question' and the widespread deterioration of living and working conditions due to urbanization and industrialization.

However, another important, and indeed competing, field of study and training, whose concerns overlapped those of the new social sciences, was that of the legal sciences. The strength of the legal scholar's position in continental Europe arose from two factors. First, the training function of the universities for the state service produced officials and judges. This meant that legal scholarship came to exert a considerable impact on the general outlook of the servants of the state. Second, legal scholars sought to provide a doctrine, a body of concepts that was based on elaborate technical distinctions and would enable lawyers and judges to act with promptness and precision, clarify the deliberations of the law-maker, and bring coherence and order into the legal system (Dyson, 1980, p. 112). This doctrine was legal positivism, which was first developed for private law and later for public law (in the latter case known as the legal theory of the state). Legal scholarship remained an important alternative to social science and, in fact, strengthened its position towards the end of the nineteenth century. Thus, the so-called 'state sciences' in the German and Nordic states did not really develop into anything that might be labeled 'political science.' Rather they became increasingly embedded in the legal constitutional scholarship that was expanding in Germany in the wake of German political unification. Conversely, in the Nordic countries a similar tradition of 'state sciences' in the late nineteenth century was characterized by a dual legacy of constitutional legal scholarship and, as in Britain, of studies of philosophy and political history.

By and large, approaches to a social and political science did not succeed in institutionalizing themselves in European academia. In some cases, they tried but failed. In other cases, they did not perceive the relevance of academic institutionalization, but blossomed during a passing period of a favorable political climate and decayed with the changing societal situation. For example, the term 'sociology' had been coined early in the nineteenth century, but the key sociologists of the late nineteenth century (Weber, Durkheim, and Pareto) remained broad social science generalists. Their contributions and professional allegiances traversed a range of fields, including politics, economics, education, history and religion, and the term sociology often referred to a broad historical-comparative study of society. However, it was a type of study that saw itself as scientific and separate from reform-orientated activities of a more general nature.

In the United States, social science research originally had the same characteristics of associational organization and ameliorative orientation as in Europe. For example, the American Social Science Association (created in 1865) embraced the notion that the social scientist was a model citizen helping to improve the life of the community, not a professional, disinterested, disciplinary researcher. During the final cades of the nineteenth century, however, this model was gradually replaced by the emerging disciplinary associations such as the American Economic Association (1885), the American Political Science Association (1903), and the American Sociological Society (1905) (Furner, 1975; Haskell, 1977; Manicas, 1987; Ross, 1991). They did not have to face the kind of entrenched opposition or ideologically motivated hostility that many similar efforts in Europe met with (even though it would be misleading to portray these processes as the harmonious unfolding of disciplinary consolidation).

By the turn of the century, a particular pattern of differentiation and professionalization had emerged in the US and proved to be compatible with an active role in government service early on. Economists and psychologists were, for instance, able to play a role in government service as early as World War I, just as demographers were able to do through the Bureau of the Census. At the same time, disciplinary and professional recognition was

being achieved within the setting of American universities which started to become ever more orientated towards research undertakings (Geiger, 1986; Rothblatt and Wittrock, 1993)

This process of successful disciplinary consolidation marked the beginning of a divergence between American and continental European social science. Late nineteenth century American social scientists, many of them German-trained, defined their intellectual projects in a society undergoing a process of rapid transformation: industrialization, urbanization, and concomitantly emerging massive social and political problems. As social scientists with a professional legitimacy, they tried to mark out their own scientific territory and establish their own systems of accreditation. These ambitions entailed the establishment of separate social science disciplines (Manicas, 1987). In Europe, on the other hand, the professorate often already had an established position; one writer (Ringer, 1969) even uses the term 'a Mandarin class' to describe the situation of the leading German academics at the turn of the nineteenth century. The situation in some other countries, such as Sweden and Norway, was no different.

The social sciences emerged as forms of knowledge about societies undergoing fast and deep societal transformations. The relative success of given research programs largely depended on the intellectual coherence and viability of these programs. The cases of Max Weber and Emile Durkheim are exemplary in this respect, but so is the ability of marginalist economics programs to assert themselves even under political conditions which may have appeared as far from advantageous. However, the institutionalization of the social sciences was also clearly related to two other forms of institutional processes.

First, the late nineteenth century was the period when the idea of the university as a key vehicle for research became predominant in a number of countries across the globe, from California in the west to Japan in the east. This process created the possibility for the new social sciences, if in a highly uneven way, to find a relatively stable basis for continued research and training.

Second, the late nineteenth century was also a period of deep institutional change in the political and administrative order in a number of countries. Administrative reforms were undertaken and new objectives assigned to the state. Furthermore demands for wider political representation meant that institutional reforms of the state, not least the demand for a culturally coherent nation-state, created demands upon the social and historical sciences to contribute to these reform processes. The particular forms of interaction between research programs, efforts to deal with the social question, and the relationship of that question to the restructuring of universities and of the state, differed greatly across countries. However, in a number of cases, opportunities arose for the successful institutionalization of some of the social science disciplines in ways that came to structure disciplinary developments in these countries for many decades.

Research Programs in the Interwar World

The end of World War I saw the triumph of liberal democracy and the reorganization of the political order across the European continent as well as a decisive weakening, if not the end, of European global predominance. Soon it became evident that the triumph of democracy was being replaced by its crisis and the emergence of new forms of political and social order with claims to represent the future of humankind in Europe and beyond. It was in this period that pre-eminent representatives of a range of social science disciplines engaged in a self-critical reflection on the history of their own disciplines. Based on such historical reflection, however, they were able to formulate research programs that came to serve as focal points for scholarly endeavors for decades to come.

Within just a few years in the middle of the 1930s a range of pathbreaking programmatic formulations occurred. At this time economists, most notably Keynes in England but also Gunnar Myrdal and the other members of the so-called Stockholm school in Sweden, took stock of the historical experience of their discipline and formulated a coherent long-term research program. In sociology, Parsons' *The Structure of Social Action* was equally historically and programmatically orientated. Its influence extended far beyond the domain of disciplinary sociology and came to affect developments in a number of other fields, including political science and social anthropology. In statistics, Fisher was able to achieve an encompassing synthesis that became a landmark. In political science and sociology, scholars at Chicago and Columbia opened up new areas of empirical research. In Europe, the early work on electoral geography by André Siegfried in France was complemented by sociological and sociopsychological studies by Paul Lazarsfeld and his collaborators in Austria, and by Tingsten's and Wold's minor classic, *The Study of Political Behaviour* (1937).

The comprehensive philosophical programs outlined by the Continental logical empiricists and their Anglo-Saxon counterparts, despite their relative numerical weakness, came to set agendas for years to come. In this respect, Alfred Jules Ayer's polemical and programmatic volume on *Language, Truth and Logic* (1934) became one of the most noted examples, as did Otto Neurath's *Empirische Soziologie* published at roughly the same time. Meanwhile the French Annales School charted a completely new terrain of research that forever changed scholarly efforts in the historical and the social sciences. This was also the period when the first edition of the *International Encylopedia of the Social Sciences* appeared, as a further testimony to need for reflection, for stocktaking but also for the setting of new scholarly agendas.

Furthermore, research programs emerged in the 1930s that positioned themselves in conscious opposition to disciplinary developments in the social sciences and the humanities. One prominent example of this was the broad synthetic program associated with the so-called Frankfurt school. A very different one was the effort of philosophical phenomenologists, most prominently Husserl and Heidegger, to elaborate a conceptualization of human activity from a point of view at odds with that of both analytical philosophy and dominant forms of empirical social research.

Many of these efforts proved to be of lasting importance. However, in institutional terms, the devastating effects of the European political landscape in these years and the ravages of World War II mean that it is difficult to discern clear institutional continuities in the continental European case. Many of the developments in the interwar years were followed by deep ruptures sometimes tended to make the social sciences after World War II appear as a new phenomenon.

The Internationalization of the Social Sciences: The Age of International Associations, Public Policies, and Mass Higher Education

The full-blown institutionalization of the social science disciplines on a global scale is largely a phenomenon of the era after World War II. One manifestation of this was the establishment, originally under the auspices of *UNESCO*, of the *International Political Science Association* (*IPSA*) and the *International Sociological Association* (*ISA*) in 1949. A process of professionalization was set in motion and came to exert a truly profound influence worldwide in the wake of the expansion of higher education systems in a range of countries in the 1960s and early 1970s.

The 1960s saw a dramatic expansion of higher education systems across Western Europe and North America but in many other parts of the world as well. In the same countries, sweeping processes of administrative reform also occurred, which often coincided with the coming to power of new political majorities. Major new public policy programs were launched across the board in these countries. In this context, the social sciences came to be finally and firmly entrenched as academic disciplines in university settings. In this same age of great public policy programs, disciplines such as political science and sociology were able to secure a firm basis in a series of European countries, in some cases for the first time; in others (as in Britain, Germany, Sweden, and Italy) in a renewed and greatly expanded form. This expansion was paralleled by a growth of the social science disciplines on a global level that for the first time tended to make the international associations truly international.

On all continents, the full array of disciplines and subdisciplines appeared. There was also an impressive expansion of research methods being utilized. Thus a previously predominant concern for institutions and processes on a national level was gradually complemented by a stronger research orientation towards the study and management of clearly quantifiable processes. In terms of research methods, the 1960s were the period of the breakthrough of the behavioral revolution; a revolution which had been largely foreshadowed by European scholarly efforts in the interwar period. No longer could historical, juridical, and philosophical reasoning alone or in combination be considered sufficient for the analysis of social and political phenomena. Methods and techniques previously elaborated in statistics, sociology, psychology, and economics were now being tapped by social scientists on a vast scale.

This shift in research methodology coincided with the numerical expansion of the social science disciplines and in Europe was often complemented by the introduction of more formalized graduate education programs, normally with compulsory courses in research methodology. At the same time it became possible to see the emergence of an informal 'invisible college' of younger scholars in Europe and in other continents as well, in marked contrast to the much more national orientation of scholars of older generations. In those universities and countries where this shift was most decisively pushed through, there were certainly instances where the older juridical, historical, and philosophical competence was either partially lost or could at least not be developed on a par with the new approaches. Thus it became gradually antiquated and lost touch with what had been the rationale for the use of these traditional methods in the first place.

Apart from external political-societal reasons for this pattern of disciplinary development, a fundamental fact of a long period after World War II was that only in the United States did the social sciences have sufficient size and scope to make widespread international emulation appear feasible. Thus, for example, the American Political Science Association had more than five thousand members in the early 1950s, when corresponding figures for European nations were generally only on a scale of 1 percent or less of that number. Even in the early 1990s, the combined enrolment of West European political science associations was only a third of that of the American association.

Furthermore, only the American version of a science of society and of politics could present a clear institutional lineage dating back to the late nineteenth and early twentieth centuries. In this perspective, the history of the social sciences in a number of European countries constitutes exceptions to this general tendency only in a weak sense of the word. This is perhaps particularly clear in the case of political science. There has been a persistent notion that political science is intimately linked to the particular intellectual and institutional history of the United States. This can be discerned as early as 1923, in Charles Merriam's presidential address to the American Political Science Association. The same theme recurred 30 years later in the presidential address of Pendelton Herring, and again in the one delivered by Samuel Huntington another quarter of a century later. Political science is then seen to constitute an originally American-conceived science that has successfully spread to a range of nations that (by virtue of their effective, if belated, acceptance of a pluralistic political system) had come to qualify as proper recipients of this intellectual heritage.

Even if such a perspective is understandable, it rests on a misperception. It identifies political science inquiry with one particular form of such inquiry – related to the concerns of liberal individualism. In this and many similar cases, a lack of historical perspective threatens to unduly narrow both the range of problems addressed and the theoretical traditions explored and advanced.

Continuities and Reassessments

Three key features stand out in the development of the social sciences in the late twentieth century. First, their professional consolidation has proceeded. Their institutional position in the modern mass higher education systems seems relatively secure. Never before have there been so many social scientists in the world, and never before have they been so well organized in professional terms, well trained in technical terms, and internationally linked through journals, networks, and research conferences. Many social science disciplines – political science, sociology, educational research, business administration studies, and public policy research – have to a large extent emerged and evolved as confederations of different practices. They have been held together by a common concern with broad substantive themes rather than by a core of theoretical assumptions.

Second, however, it is remarkable to what an extent truly innovative research contributions have resulted from work in scholarly settings outside of the structure of regular disciplinary university departments. Three types of such environments seem to have been particularly important to the intellectual advancement of the social sciences. First, especially in the major American research universities, the system of so-called organized research units has been of great importance in providing settings outside of the departmental structure designed to be hospitable to research of a transdisciplinary nature. Second, research institutes, also normally of a transdisciplinary nature, have been created both outside and inside of the structure of regular universities. In Germany, the institutes of the Max Planck Society have played a crucial role in this respect. Third, the important role of so-called institutes for advanced study has been steadily increasing during the second half of the twentieth century. Such institutes allow scholars, free from the normal concerns of a university department and coming from a variety of disciplines, to focus on their own research efforts during a limited period, normally one academic year. Whereas initially such institutes tended to be international in their scholarly orientation and national in their own organizational structure, there has recently been an increase in the number of such institutes set up within individual universities. The so-called humanities centers now found in a large number of American universities provide one example of this trend. The recommendation of the Gulbenkian Commission (Wallerstein, 1996) that every university create a transdisciplinary institute for advanced study is another indication, a recommendation that now slowly seems to correspond to an emerging institutional reality.

Third, the demands upon the social sciences, from politicians, administrators, and the public at large, to demonstrate the usefulness of their contributions, do not seem to have subsided but if anything to have increased. Sometimes these demands have even resulted in the creation of new proto-disciplines such as social work, criminology, and migration research.

The foremost social scientists of the late nineteenth century had to demonstrate the intellectual and institutional legitimacy of their activities relative to competing forms of scholarship in law, economics, and philosophy. These competitors remain but have been complemented by management consultants, computer experts, and representatives of new transdisciplinary fields such as that of the cognitive sciences.

There is no reason to believe that the social sciences will not be able to defend their current position. It is important, though, to recognize that advances in these sciences depend upon intellectual integrity in the face of an array of demands for immediate usefulness. At the same time, any truly innovative social science inquiry has to be contextually and historically sensitive. This also means that a search for a solution to most of the essential problems of the social sciences has to proceed with a high degree of openness to the contributions from outside of the field of a single discipline. This, however, should be seen as an asset rather than as a liability. In this situation, historical reflection is a necessary component of serious research efforts. Clearly, a number of assumptions concerning our deepest cultural and national identities and their relationship to the basic forms of political and social organization have to be examined anew with the same openness for foundational inquiry as social scientists have shown in earlier periods of fundamental change. In fact, social science scholarship may be crucially important to the possibility of a civilized community of human beings.

See also: Anthropology, History of; Hermeneutics, History of; Political Economy, History of; Quantification in History; Quantification in the History of the Social Sciences; Science, History of; Social Science and Universities; Social Science, The Idea of; Sociology, History of.

Bibliography

Baker, K.M., 1994. Enlightenment and the institution of society: Notes for a conceptual history. In: Melching, M., Velema, W. (Eds.), Main Trends in Cultural History: Ten Essays. Rodopi;, Amsterdam, pp. 95–120.

Collini, S., Winch, D., Burrow, J., 1983. That Noble Science of Politics: A Study in Nineteenth Century Intellectual History. Cambridge University Press, Cambridge, UK.

Dyson, K., 1980. The State Tradition in Western Europe. Martin Robertson, Oxford, UK.

Fox, C., Porter, R., Wokler, R. (Eds.), 1995. Inventing Human Science: Eighteenth Century Domains. University of California Press, Berkeley, CA.

Furner, M.O., 1975. Advocacy and Objectivity: A Crisis in the Professionalization of American Social Science 1865–1905. The University Press of Kentucky, Lexington, KY.

Geiger, R.L., 1986. To Advance Knowledge: The Growth of American Research Universities 1900–1940. Oxford University Press, New York.

Goodin, R.E., Klingemann, H.-D., 1996. Political science: The discipline. In: Goodin, R.E., Klingemann, H.-D. (Eds.), A New Handbook of Political Science. Oxford University Press, New York, pp. 3–49.

Graham, L., Lepenies, W., Weingart, P. (Eds.), 1983. Functions and Uses of Disciplinary Histories. Reidel, Dordrecht, The Netherlands.

Haskell, T.S., 1977. The Emergence of Professional Social Science. University of Illinois Press, Urbana, IL.

Heilbron, J., Magnusson, L., Wittrock, B. (Eds.), 1998. The Rise of the Social Sciences and the Formation of Modernity: Conceptual Change in Context. Kluwer, Dordrecht, The Netherlands, pp. 1750–1850.

Jarausch, K. (Ed.), 1983. Transformations of Higher Learning. Klett-Cotta, Stuttgart, Germany, pp. 1860–1930.

Kelley, D.R., 1990. The Human Measure: Social Thought in the Western Legal Tradition. Harvard University Press, Cambridge, MA.

Lepenies, W., 1988. Between Literature and Science: The Rise of Sociology. Cambridge University Press, Cambridge, UK.

Manent, P., 1998. The City of Man. Princeton University Press, Princeton, NJ.

Manicas, P.T., 1987. A History and Philosophy of the Social Sciences. Oxford, UK: Blackwell.

Pagden, A. (Ed.), 1987. The Languages of Political Theory in Early Modern Europe. Cambridge University Press, Cambridge, UK.

Ringer, F.K., 1969. The Decline of the German Mandarins. Harvard University Press, Cambridge, MA.

Ross, D., 1991. The Origins of American Social Science. Cambridge University Press, Cambridge, UK.

Rothblatt, S., Wittrock, B. (Eds.), 1993. The European and American University Since 1800: Historical and Sociological Essays. Cambridge University Press, Cambridge, UK.

Rueschemeyer, D., Skocpol, T. (Eds.), 1996. States, Knowledge and the Origins of Social Policies. Princeton University Press, Princeton, NJ.

Wagner, P., Weiss, C.H., Wittrock, B., Wollmann, H. (Eds.), 1991a. Social Science and the Modern State: National Experiences and Theoretical Crossroads. Cambridge University Press, Cambridge, UK.

Wagner, P., Wittrock, B., Whitley, R. (Eds.), 1991b. Discourses on Society: The Shaping of the Social Science Disciplines. Kluwer, Dordrecht, The Netherlands.

Wallerstein, E., et al., 1996. Open the Social Sciences: Report of the Gulbenkian Commission on the Restructuring of the Social Sciences. Stanford University Press, Stanford, CA.

Discipline-Building in the Social Sciences: Collective Memory, Biography and Autobiography

Charles Crothers, School of Social Sciences and Public Policy, AUT University, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

As they develop, academic disciplines (and their constituent specialties) generate research outputs, theoretical and methodological commentaries, and collective perspectives (particularly disciplinary histories and other collective memories). In addition, there are 'backstories' of how the research outputs are produced from the underlying infrastructure of the discipline and the personal efforts of its members. The production of personal documents (e.g., autobiographies/biographies) is an increasingly important way in which 'the personal' is fruitfully linked up with the 'collective.' But disciplines, while being relatively self-conscious, still lack in collective direction over their operations and trajectories.

Introduction

Academic (science, social science, and also humanities) disciplines, and their constituent specialties, reveal a variety of patterns of change over time. For convenience, the discipline level is assumed in this article but the same points apply to both lower-level entities (e.g., fields or specialties) and to the broader (e.g., social sciences as a whole). A common pattern is for initial discoveries and slow early growth accompanied by some resistance from other specialties, which then escalates into a growth phase in which publications pour out and the intellectual cognitive structure framing these is institutionalized. A final phase can be a slower growth once major innovations in the area have been exhausted, and perhaps a continuing 'after life' in which the reigning intellectual paradigm continues to have some relevance as well as contributing to applications as the cognitive basis for '(social) engineering' interventions. However, along this common path there are many variations (see Abbott, 2001). Alongside the main theoretical and substantive literature in a discipline there also other literature such as that on methodology, which provides guidance about appropriate methods.

There is also a more subjective side of research and publications in which the various disciplines construct socially meaningful cultural frameworks which they then espouse to themselves, their students, scientists in other areas, and the wider community. These can broadly be referred to as 'discipline building.' Such cultural accompaniments construct a possible future and retrieve a plausible past together with providing a 'working model,' providing a shared understanding of how the discipline operates in the present. Such visions indicate what the boundaries of the field are, and in what ways the field differs from any rivals. Such disciplinary cultures also 'localize' a particular version of the overall culture of science and there are also differences in a range of other characteristics, including e.g., preferences between book and journal article publications.

There are two quite separate functions of these articulations of subjective 'disciplinary cultures':

- Assisting in constructing the history or the contemporary culture of the discipline for its own sake and as an ideal held out to motivate adherents and neophytes to the discipline.
- Providing sensitizing methodological reflection which can affect the validity and reliability of the field's findings.

In addition, since research work and writing share some of the commonalities of all social reality, and given the privileged internal access, it can be argued that they can also provide a site for the development of generalized social science knowledge which is just as valid as that obtainable from any other site.

The 'BackStories': Infrastructure and Also Personal Experiences

The scaffolding supporting the work of a discipline often remains in the shadows. The infrastructure of a discipline (its institutions, organizations, funding, techniques, etc.) is a necessary, yet seldom foregrounded, aspect of knowledge production. The ways in which these structures support or hinder research effort is important, but seldom assessed. Moreover, at a personal level the scholars in a discipline each have their own experiences of being involved in knowledge production. And some of these experiences are selectively documented in informal, and sometimes formal ways (e.g., field notes, diaries).

Yet, most of these backstories remain hidden. A sharp line tends to be drawn between the 'public' and the 'private.' More intimate details of the social production of scientific knowledge tend to be suppressed since their exposure might be seen to detract from the ideal of the production of the 'pure truth' that many scientists see as their goal – or as relatively uninteresting. After findings are obtained the 'scaffolding' is disassembled, and even hidden. Public documents tend to produce only the final version of the study and mask the process of attaining this, thus suppressing the "…intuitive leaps, false starts, mistakes, loose ends and happy accidents that actually cluttered up the inquiry" (Merton, 1967, p. 4). Michael Polanyi's concept of 'tacit knowledge' is a more positive version of this idea, which calls attention to the role of unstated (background) knowledge in getting experiments to work.

Academics are motivated by strongly held scholarly norms to keep the public and private spheres of their activities separate and the latter hidden. Moreover, undue attention to the self is seen as self-indulgent. However, a contrasting value motivating

the revelation of backstories can be that of informing and educating the public about academic efforts by allowing them to see the human endeavor required in the production of scholarly knowledge.

Task of Article

In constructing disciplinary collective memories and histories, both objective but also subjective collective and personal sources are drawn upon. Various activities, including developing disciplinary histories, put some of the more private component of knowledge production back out into public view. But accessing the latter requires effort. This article reviews how private but also public 'personal documents' are drawn on in establishing disciplinary histories (of a variety of forms). In particular, there is an interest in understanding whether developing collective memory leads to disciplinary self-consciousness and some ability for its members to guide the future of their discipline.

Unfortunately, there is little systematic work on this whole wider discipline-building area of scientific activity so the references are limited, and mainly found in chapters or journal articles (an interesting exception is Popkin, 2005 on the autobiographies of historians). Much of the literature is reflection (including debate) arising from the publication of particular biographical materials. Studies of discipline building in other than social science disciplines are also pertinent (e.g., Söderqvist, 1996; Shortland and Yeo, 2008).

Disciplinary Collective Memories in Discipline Building

A major component of discipline-building processes is to reflect on and construct that discipline's history and this historical image is usually implicitly built into the discipline's images of its own trajectory (the present and future as well as the past). In more recent decades, efforts have been made to develop contemporary pictures of a discipline, such as by monitoring research outputs and visualizing citation networks or collaboration patterns. Other efforts – such as the famous 'Foresight' programs in several countries – have endeavored to extrapolate or project into the future that topics of scientific knowledge should be focused on. Occasionally, specially constituted commissions to develop research agendas which will guide future research funding and research efforts are appointed. However, largely because many of the necessary facts lie comfortably at hand, construction of backward-looking disciplinary histories is a more usual mode of discipline building.

Apart from textbook coverage and occasional historical tomes, disciplinary historical material can be developed in a variety of ways and on a variety of occasions including particularly presidential addresses, commemorations celebrating turning points (e.g., anniversaries, centenaries). Correspondence between scholars is of major importance as letters (and now e-mails) will reveal much about scholars' concerns, motivations, strategies, and experiences. In turn, the extent to which stock of a scholar's correspondence is kept in some form of archive and then rendered accessible varies enormously. Some correspondence which has survived sits in well-cataloged, well-housed accommodations, whereas others are merely dumped without any finding mechanisms and in premises which are not hospitable to the subsequent researcher who might wish to consult them. And (as Fleck, 2011 points out) for those who fail to be appointed to an official position there is usually not the slightest trace since any records of them are ephemeral.

There is a hierarchy of those who are preferred spokespeople; mainly the discipline's elders and official leaders. After all, they are drawing on experience and the authority of their pronouncements comes from their official or unofficial positions. As a result, it is likely that disciplinary histories will be at least partially shaped by the 'interests' of such spokespeople. For example, to maintain a thoroughly scientific image, some categories of potential contributors to an emergent discipline are 'read out of the picture' – such as those considered not to have made proper scientific contributions or possibly those from less-valued social backgrounds. Resisting this, other historians may endeavor to counter with alternative histories which reinstate such excluded figures and which (often vigorously) outline alternative conceptions of the history of the discipline (e.g., Deegan, 1991). And in turn, more orthodox historians may defend their original configuration (e.g., Hamilton, 2003).

Disciplinary Differences in Disciplinary Histories

The natural sciences are most resistant to revelation of their construction systems because these are seen as almost entirely irrelevant in the advance of scientific activity, and at best tolerated as public relations exercises. At the other pole, the humanities – which are largely concerned with the preservation, articulation, and development of culture – tend to be very historically orientated disciplines. (However, more recently, cultural theory across a range of humanities disciplines has perhaps focused more on structural features without such a concentration on heritage). The social sciences fall in between, with disciplinary histories playing a substantial, but not necessarily major role in their intellectual activities.

The major social sciences disciplines began to become institutionalized from North America and Europe in the 1880s and 1890s – although taking much longer to be properly established in many countries and only becoming common across the world over the last half-century at most. When established they were in the shadow of several centuries of earlier intellectual developments which they had to come to some terms with. And so, there was considerable interest in this prehistory by early writers in each disciplinary area.

Many of the newly forming disciplines developed 'founding myths,' although this varied in importance. Donald Levine proclaimed, for example, that "Sociology was born with a ready-made history" (1995, p. 13) since one of the 'founding fathers' – Comte – included an historical sweep of previous writers in his defining of the discipline he was helping to create (also see, e.g., Helm, 1967).

Some common characteristics shared by disciplinary histories may vary by phase of development, e.g., Merton cites Alfred

N. Whitehead's "it is characteristic of a science in its earlier stages … to be both ambitiously profound in its aims and trivial in its handling of details" (1967, p. 49). Merton argues that grandiose images of the scientific status of a discipline are often produced because of perceived public demand for successful explanatory capabilities. However, the public erroneously judges all disciplines (including the social sciences) in terms of the same standard, ignorant that some disciplines stand at a different phase in terms of development sequences and therefore have more limited current explanatory capacities. Spokespeople for disciplines may need to point out their discipline's limitations to the public.

A common ordering of phases in disciplinary histories involves those of forerunners, founding fathers (few founding mothers tend to be identified), and those responsible for the further development of the discipline. For example, Merton suggested that (1967, p. 2) "…Sociologists retain a most parochial, almost Pickwickian conception of the history of sociological theory as a collection of critical summaries of past theories spiced with short biographies of major theorists" and arranged in chronological order. Often such disciplinary histories are couched as heroic tales, which valorize particular individuals who are thus created as founding fathers or more generally as 'classic' figures.

Not all disciplinary histories tell a simple story of progress. Rather, Levine presents a typology of narratives (1995, pp. 96, 97) which points to a far wider range of alternative images:

- Positivist: history "…does exhibit a progressive refinement of observational and analytical capabilities and a cumulatively developed body of findings";
- Pluralist: "…the universe of viable… approaches consists of an irreducible multiplicity of perspectives, principles, methods, definitions of subject matter, and statements of aims";
- Synthetic: "…agreement has been reached and can be reached on the relevance of certain basic terms, conceptions and generalizations";
- Humanistic: "…that the heritage…contains more than a body of replicable techniques, substantive findings and theoretical analyses";
- Contextualist: that historical contexts have shaped the character of "…concepts and problems in profound, often subtle, ways."
- Dialogical: "…each type arose through critical engagements with the others and …each continues to occupy a place in the broader tapestry." And the trajectory takes the shape of a spiral.

Which trope is used to model the trajectory varies between historians: Levine advocates the dialogical approach which he also claims best represents large portions of sociology's history.

For many social science disciplines, there is a tendency for the systematics of theory and the history of theory to merge and to be conflated when employed in teaching. Theoretical ideas are portrayed as comprising 'packages' which tend to belong to a particular era, although several may be seen as coexisting in the same period. Often such packages are labeled as 'Schools' although these are constructed after the event rather than necessarily being social communities linking a group of colleagues. The history of a discipline is often told as a sequence or series of such 'Schools.' Such idealized disciplinary histories are ideal for more or less simultaneously teaching students the history of the discipline at the same time as the range of theoretical and methodological options available within it.

In his fecund (if problematic) model of scientific developments, Thomas S. Kuhn pointed out the importance of textbooks in indoctrinating new recruits into the mysteries of their discipline and indeed, the packaged histories often included in textbooks can be important for inductees. Although they are not produced to have high standards of historical validity, such versions of history become 'authorized' versions by being produced in textbook accounts.

Levine nicely sums up some more general effects of disciplinary histories (1995, p. 10) "…Collective memories recall great events of the past, which form a focus for group solidarity. They also keep alive the sense of affiliation with the group and its purposes during periods of routine activity between times when the group assembles to reaffirm its collective existence." They can also be defensive, as in remembrance of guilt and difficulty, hopefully overcome by subsequent striving. They can give a sense of duration: extrapolating from past trends or moving forward or backward to some shared history or future history, so they can valorize particular historical directions. They can demarcate boundaries, celebrate the emergence of a new formation or trace its connections/disconnections to others (see also Graham et al., 1983).

Writings from the past of a discipline can also contribute more directly to its present academic work, rather than only its history. Merton has commented on the 'functions' that are served by social science scholars working through earlier material in their discipline (1967, p. 37). Motivations for considering earlier writings

> …range from the direct pleasure of coming upon an aesthetically pleasing and more cogent version of one's own ideas, through the satisfaction of independent confirmation of these ideas by a powerful mind, and the educative function of developing high standards of taste for […] work, to the interactive effect of developing new ideas by turning to older writing within the context of contemporary knowledge.

Merton (1967) has also developed a typology of different ways in which earlier and later work might be related. However, disciplinary histories draw more broadly on the stock of past work than only those elements where there are intellectual connections over time.

Because of the poor quality of many disciplinary histories, there have long been calls for more adequate 'sociological' disciplinary histories in all social science disciplines (e.g., in Sociology, Coser's, 2003 – more sociological treatment of the history of sociology represents an advance over alternative histories which convey less about the social contexts of the writers covered). Merton's sketch of the expertise needed is still pertinent: "…a new breed of specialized historian of science is drawing widely and deeply upon sociology, psychology and tools of science for theoretical guides to their interpretations of

the development of science" (Merton, 1967, p. 3). Better histories

> …would take up such matters as the complex filiation of sociological ideas, the ways in which they developed, the connections of theory with the changing social origins and subsequent social statuses of its exponents, the interaction of theory with the changing social organization of sociology, the diffusion of theory from centers of sociological thought and its modification in the course of diffusion and the ways in which it was influenced by changes in the environing culture and social structure.
>
> Merton, 1967, p. 2

The purpose of these more sophisticated accounts is to achieve better understanding of how things came to be, which is more usually censored from 'Whiggish' disciplinary history accounts which favor a more linear construction of 'disciplinary progress.' So, in the more extended versions of disciplinary histories the material to be included covers

> …conceptions which made good sense at the time of their formulation but were later shattered by compelling empirical tests or replaced by conceptions more adequate to the enlarged facts of the case. It also includes the false starts, the now archaic doctrines and both the fruitless and fruitful errors of the past.
>
> Merton, 1967, p. 3

In constructing more sound disciplinary histories, it is necessary to draw on biographical and other personal documents and experiences.

Personal Documents in Academic Disciplines

This section is organized around asking: who writes what (i.e., content) about whom, where, is published where, and with what effects?

There are several different genres of personal documents:

- Entries in biographical dictionaries
- Obituaries
- Diaries and autobiographical accounts, memoires
- Amateur histories
- 'Sociological' histories
- Professional histories (including those which are theory based)
- Development of more systematic historical studies (e.g., Paul Lazarsfeld's research program on the early development of social research).

This list is organized in terms of a descending continuum of increasing interpretation – those genres listed earlier contain only fairly 'raw' material from a subjective standpoint whereas those listed further down have a higher component of interpretation. Also, validity and reliability of information tends to increase as one goes down the list. However, it is not that simple. Each of the genres has different characteristics, and prescribed lengths of treatment which control the extent of reportage. A major axis lies in terms of the extent of intellectual subject matter versus 'extra-intellectual' coverage of more personal information. Some genres focus almost entirely on subject matter whereas some barely mention this in favor of more extensive biographical coverage. Some genres are particular to some time periods: for example, the biographical eulogies of scholars common in Medieval times.

As well as providing useful descriptive contextualizing, some writing (such as prefaces or introductions) – whether written by the autobiographer/biographer or an editor – can be important in providing a framing for the material: indicating to the reader what sort of lessons they might gain from reading the material and more generally, how to read it. Finally, information about the construction of social science knowledge can contribute to understandings within the sociology of knowledge, such as constructing models of the processes of discipline building.

Development of personal documents (and also collective memories) can include people (social scientists), institutions and organizations or social science 'outputs' (ideas, research), although the first-mentioned predominates. A particular subgenre is sometimes named 'research chronicles' (e.g., Hammond, 1967) or more colloquially 'tales from the field' (e.g., Van Maanen, 2011) as they provide a depiction of the development of a particular research project, which is often collective and relatively short term – in contrast with writing about more individualistic and life-long intellectual careers.

Paradoxically, many personal documents have little personal information, certainly about scholars' private lives. Personal documents tend to suppress awkward, too personally revealing, details. Some are highly formulaic revealing little more than an orthodox 'curriculum vitae.' For example, Fleck (2011) points out that whereas the emigration of the academic social scientist migrants he studied is well documented, their (less renowned) return home is somewhat shrouded in silence. Published biographies may admit to occasional shortcomings but major secrets and deficiencies usually remain hidden. The extent of what is hidden, though, may be disguised by limited revelations of some 'safe secrets.' There clearly is some temptation to present the story of a scholarly life as being more coherent that it really has been, although very often the scholar is represented as being, rather, at the mercy of the fates. Charles Tilly complained that "when forced to confront their own lives, sociology's practitioners … tidy them up, mute their passions, avoid vivisection of their motives, portray themselves as bemused players in a game they do not run, treat their own careers as a series of breaks, lucky or otherwise" (1993, p. 497).

Some personal documents, though, are written in styles which reveal more conflicts, emotions, and struggles. Such accounts to some extent are indeed written to 'pay accounts' and document experienced difficulties. Often such accounts are written by disciplinary outsiders or scholars who came from unusual backgrounds.

Biographical work tends to focus on particular targets, with a strong tendency to concentrate on high-status targets. However, there may be more complex patterns in biographical work: for example, a first generation may be subject to much biographical writing since they are the founders, which may then leave less central precursors, and perhaps the second generation of followers, not adequately treated autobiographically/biographically (see Scarborough and Furumoto, 1989). Biographies also require motivated others to write about the targets, e.g., spouses, disciples, and professional historians.

More detailed kinds of source material include (Merton, 1967, p. 6) scientific notebooks and personal journals and correspondence. The introductions/prefaces in reissued editions of books are frequently informative as they provide an opportunity for the author to review the original circumstances of the book's production and also on how the original work fits in with subsequent developments. Similarly, some authors issue collected volume(s) of their writings which provide opportunities to sketch in the personal strategies lying behind their writing.

Rather than only relying on the somewhat accidental production of public or private documents, more systematic effort can be important. One stimulus is collective biographical exercises where entries are canvassed by an organizer. In many countries, there are state-sponsored national biographies of important people, which very likely will include some social scientists. In particular, there is a need to actively solicit oral histories. Again, systemic programs have occasionally been launched (especially by professional scholarly associations) to endeavor to cover a considerable number of oral histories or indeed more up-market biographical accounts. Surveys of the views, careers, and material circumstances of contemporary social scientists are seldom used, but there are some examples.

Methodology of Constructing Personal Documents: Autobiographies. Coordinated and Collective Autobiographies, etc.

Autobiographies include objective information:

> Like biographers, autobiographers can use personal documents, public documents, the testimony of relevant others and so forth, to compare and contrast these different accounts so that advantages of each is maximized and the disadvantages minimized.
>
> Stanley, 1993, p. 43

Pierre Bourdieu (2008), too, endeavored to avoid his autobiography being seen as falling into a preexisting genre and insisted on it being seen as a 'socio-analysis' of his career – in which social contexts were particularly emphasized. Methodologically, autobiographers stand in a unique relationship to their material. Merton lays out the methodology of autobiography in more detail (see also Oakley, 2010):

> The constructed personal text of the interplay between the active agent and the social structure, the interplay between one's sequences of status-sets and role-sets on the one hand and one's intellectual development on the other, with its succession of theoretical commitments, foci of scientific attention, planned or serendipitous choices of problems and choices of strategic research sites for their investigation. Fully-fledged sociological autobiographers relate their intellectual development both to changing social and cognitive micro-environments close at hand and to the encompassing macro-environments provided by the larger society and culture.
>
> Merton, 1988, pp. 19, 20

Popkin (2001) has particularly commented on the features of 'coordinated biographies' where an array of scholars are invited to contribute to a collection. Appearing in a collected volume can reduce the anxieties of self-indulgence that attends separately published autobiographies. To some extent, a degree of uniformity in style appears, possibly imposed or encouraged by the editor(s). Certainly, in comparison to autobiographical accounts which are pursued at much greater length, there is far less fleshed-out detail.

Collective biographies (prosopography) involve secondary analysis of primary sources of biographical material. The amount of interpretation which can be extracted depends on the quantity/quality of the underlying data sources. Collective biographies allow (once coded) the distillation of patterns (such as careers) and provide some ways of answering questions about how social contexts shape disciplinary developments. Fleck (2011) argues that collective biographies fill a particular gap in the sociological study of disciplines between the units of generations and of institutions and their patterns of changes, i.e., statistics concerning the present versus the long term.

However collective biographies have difficulties, especially, where insufficient data are available; it is best where the group studied is from a common occupational or similar group. Difficulties also arise when different information about a target is yielded from different sources; so rules to adjudicate on conflicting data are needed.

Biographical and other personal document practices in the social sciences follow (and possibly in some parts lead) more general societal developments on such matters. The volume of biographies and the amount of biographical (versus academic) material in them is also affected by academic ideologies.

Already, science biographies have been hailed by some commentators as producing particularly powerful studies, although few social science biographies have yet scaled these heights of acclaim (see also Söderqvist, 1996).

The extent of biographical writings has followed the general 'turn to biography' noticeable from the late 1970s (e.g., Stanley, 1993). Postmodernism as an approach has very strongly emphasized the personal factor, whereas highly positivist endeavors tend to sweep the personal element from the scene although this has not prevented (as in natural science) historical writings.

Particular trends within biographical writing will also affect content, such as the French genre of l'histoire des intelectuels which has been continued over the last century.

> It claims we should know specifically about writers, university lecturers, scientists, journalists of the past: their social origins, their educational background, how much they earned, their clubs, their habits, whom they married (as marriage as a means of social mobility) how their progeny fared, where they lived (rive droite? rive gauche?) where they spent their holidays, whether they were believers or agnostics, monarchists or republicans, pacifists or interventionists, in the resistance or collaborationists and so forth. Their theories insofar as they formulated any are of minor interest for two reasons. The first because they are the subject to other scholars … and second because these theories reflect the context in which they took shape…
>
> Borlandi, 2014, p. 164

Autoethnography takes reflexivity to its ultimate level. Biography is no longer showing how the researcher obtains data from their research site but focuses on the researcher's

personal reactions to the research site – the means has become an end. The researcher becomes his/her research subject. The research site being studied either evaporates or is reduced to somewhat unimportant background.

This approach links to the old standard psychological methodology of 'introspection' (cf Lyons, 1986). While introspection is not a term much valued in contemporary times the method is still widely used in social science, although the terminology has changed and the term repressed – introspection now emerges in the forms of self-report surveys, interviews, etc.

The particular UK research method of 'mass observation' (e.g., Chaney and Pickering, 1986) could be seen as a collective autoethnography since a wide-flung corpus of observers reported on themselves and their immediate social surrounds.

Autobiographical/biographical writing has been encouraged by publishing enterprises such as *Die Wissenschaft der Gegenwart in Selbstdarstellungen* in the 1930s period (Schmidt, 1922) and *A History of Psychology in Autobiography* (American Psychological Association, 1930) from 1930 on. More recently, various *Essais d'ego-histoire* came out in the leading publisher Gallimard's prestigious "Bibliothèque des histoires" series (e.g., Nora, 1987).

The potential for developing more sophisticated disciplinary histories in part relies on vehicles which encourage their publication. One example is the biobibliographical series sponsored by *Greenwood Press* (e.g., Deegan, 1991) which provides reviews of a scholar's oeuvre. Some disciplines have established journal(s) whose focus is the history of that discipline. *Methodologies and Practices*.

The backstory of some research projects has been told and this experience can feed into methodology texts and standards/advice which is toned by the realism of what is possible in the field (see, e.g., Meneley and Young, 2005).

How the scientists interrelate with other 'stakeholders' in their research site will also have major implications for the knowledge which might be drawn, types of access, biases, enforced limitations, etc. Field research is an inherently social process.

Sociobiographical characteristics can have effects on the scientific process.

A full-scale 'social determinism' is unwarranted: social science publications very largely correctly reflect what researchers find. However, biographical data can be drawn upon to reveal some patterns between social characteristics of researchers and their research topics. Biographical material is particularly pertinent in explaining the psychological and social influences on how problems are selected, theories chosen, and particular methodologies pursued.

Some of the links are obvious, although there are more tendencies than determinants: women scholars are more inclined toward research issues involving women, families, and the 'private' rather than the 'public sphere.' Scholars with ethnic or other marked backgrounds are more inclined to study issues relevant to these groups. More generally, there can be a range of ways in which biographical background and experiences affect research. A common pattern is for an academic to be deflected from pursuing another occupation through their scholarly pursuits but to keep up an affinity with the occupation they might have otherwise undertaken a career within. This is particularly evident in applied disciplines where clearly education studies, academics, and classroom teachers share common, overlapping interests and biographies. Linkages to social context can go far deeper. For example, in the process of migrating from Europe to the United States, Fleck argues:

> … the Europeans' focus, methods, and frames of scholarly reference became markedly influenced by American approaches to the social sciences, including adequate external funding, empirical methods, scholarly teamwork, a limited time-frame for projects, and he expansion of the relevant professional disciplines themselves.
>
> 2011, p. XX

Conclusions

The reflexivity brought about through publishing disciplinary histories allows disciplines to reflect and adjust to their circumstances. Experiences are described and their meanings established, and this allows the discipline to attempt to ameliorate difficulties and to take advantage of opportunities. However, disciplinary histories can help to cement in place ossified bodies of knowledge.

Behind these concerns about the social context effect on disciplines lies a theory of the sociology of knowledge, which is that social location (at least partially) together with institutional arrangements (e.g., funding) affects which research problems are addressed and shapes the theoretical and methodological approaches taken.

The organization of the social science disciplines may be consequential for the knowledge produced in them. In the Gulbenkian report, Immanuel Wallerstein et al. (1996) argued that the ossified structure of disciplines reduced the capacity of the social sciences to provide adequate explanations. Social science needs to achieve an appropriate mix between pure and applied work and to be assemble a range of disciplinary/knowledge expertise that covers the full array of phenomena which need to be described and explained. But, there are few mechanisms to ensure that this happens optimally. Hopefully, disciplinary histories, which are in part constructed from personal biographical material, may usefully feed into the needed disciplinary identity and reflexivity that might begin to achieve this.

See also: Autobiographical Memory; Biography and Society; Collective Identity; Collective Memory, Social Psychology of; Collective Memory, Sociology of; Discipline Formation in the Social Sciences; Encyclopedias, Handbooks, and Dictionaries; History and the Social Sciences; Life Course in History; Prosopography (Collective Biography); Scholars, Private Lives of; Scientific Disciplines, History of; Specialization and Recombination of Specialties in the Social Sciences.

Bibliography

Abbott, Andrew, 2001. Chaos of Disciplines. University of Chicago Press.

American Psychological Association, 1930. A History of Psychology in Autobiography (Multiple volumes, various editors, various publishers).

Borlandi, Massimo, 2014. Symposium: Durkheim & Co. Contemporary Sociology 43 (2), 160–165.

Bourdieu, Pierre, 2008. Sketch for a Self-Analysis (R. Nice, Trans.), english ed. University of Chicago Press, Chicago.
Chaney, David, Pickering, Michael, 1986. Authorship in documentary: sociology as an art form in mass observation. In: Corner, John (Ed.), Documentary and the Mass Media. Edward Arnold, London, pp. 29–46.
Coser, Lewis, 2003. Masters of Sociological Thought: Ideas in Historical and Social Context, second ed. Waveland Press, Inc.
Deegan, Mary Jo, 1991. Women in Sociology: A Bio-bibliographical Sourcebook. Greenwood Press, New York.
Fleck, Christian, 2011. A Transatlantic History of the Social Sciences Robber Barons, the Third Reich and the Invention of Empirical Social Research (Translated from the German by H. Beister). Bloomsbury Academic, London.
Graham, Loren, Lepenies, Wolf, Weingart, P. (Eds.), 1983. Functions and Uses of Disciplinary Histories. Deride, Holland.
Hamilton, Richard, 2003. American sociology rewrites its history. Sociological Theory 21 (3), 281–297.
Hammond, P.E. (Ed.), 1967. Sociologists at Work: Essays on the Craft of Social Research. Anchor Books, Garden City, NY.
Helm, June, 1967. Pioneers of American Anthropology: The Uses of Biography the American Ethnological Society Monograph 43. University of Washington Press, Seattle, London.
Levine, Donald, 1995. Visions of the Sociological Tradition. University of Chicago Press, Chicago, IL.
Lindzey, G., 1930. A History of Psychology in Autobiography (Multiple vols.). Clark University Press.
Lyons, William, 1986. The Disappearance of Introspection. MIT Press, Cambridge, MA.
Meneley, Anne, Young, Donna Jean (Eds.), 2005. Auto-ethnographies: The Anthropology of Academic Practices. Broadview Press, Peterborough, ON; Orchard Park, NY.
Merton, Robert K., 1967. On Theoretical Sociology. The Free Press, New York.
Merton, Robert K., 1988. Some thoughts on the concept of sociological autobiography. In: Matilda, Wiley (Ed.), Sociological Lives. Sage, Newbury Park, pp. 17–21.
Nora, Pierre (Ed.), 1987. Essais d'ego-histoire. Gallimard, Paris.
Oakley, Anne, 2010. The social science of biographical life-writing: some methodological and ethical issues. International Journal of Social Research Methodology 13, 5. http://dx.doi.org/10.1080/13645571003593583.
Popkin, Jeremy D., 2005. History, Historians, and Autobiography. University of Chicago Press, Chicago.
Popkin, Jeremy D., 2001. Coordinated lives: between autobiography and scholarship. Biography 24 (4), 781–805.
Scarborough, E., Furumoto, L., 1989. Untold Lives: The First Generation of American Women Psychologists. Columbia University Press, New York.
Schmidt, Raymund, 1922. Die Wissenschaft der Gegenwart in Selbstdarstellungen. Felix Meiner Verlag, Hamburg.
Shortland, Michael, Yeo, Richard (Eds.), 2008. Telling Lives in Science: Essays on Scientific Biography. Cambridge University Press, Cambridge.
Söderqvist, T., 1996. Existential Projects and Existential Choice in Science: Science Biography as an Edifying Genre. Cambridge University Press, Cambridge.
Stanley, Liz, 1993. On auto/biography in sociology. Sociology 27 (1), 41–52.
Tilly, Charles, 1993. Blanding in: review of Berger. Sociological Forum 8 (3), 497–505.
Van Maanen, John, 2011. Tales of the Field: On Writing Ethnography, second ed. University of Chicago Press, Chicago (first ed. 1988).
Wallerstein, Immanuel, et al., 1996. Open the Social Sciences: Report of the Gulbenkian Commission on the Restructuring of the Social Sciences. Stanford University Press, Stanford.

Discourse and Indentity

Francisco Yus, University of Alicante, Alicante, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article the relationship between discourse and identity is pictured as an inverted triangle with a broad section at the top for 'inherited' sources of identity (sex, race/ethnicity, nationality) that have discursive correlates. The section in the middle represents 'optional' social groups to which the individual chooses to belong (and which often involve jargons that create barriers of discursive specificity). Finally, the bottom vortex of the triangle represents the individual as a specific holder of identity, discursively exhibited as the person's idiolect. Furthermore, in this article it is argued that the Internet alters this triangle by re-inverting it and that it also influences the extent and scope of some traditional discursive sources of identity.

The human being is inextricably tied to language. As soon as children are born, a biologically rooted, built-in discourse-developing mechanism is activated enabling them to become communicatively competent, to interact with the world, and to become aware of their role and specificity in society. However, not all discursive sources of identity have the same quality and strength. In Yus (2001), three major discursive sources of identity were schematized as an inverted triangle in which the broad section at the top represented 'inherited' sources such as nationality, sex, and race/ethnicity. A narrower section in the middle of the triangle represented 'optional' identity sources, mainly social groups to which the human being chooses to belong (often with group-specific jargons). And finally, a very narrow section at the bottom of the inverted triangle represented the human self as it is developed and shaped through everyday face-to-face conversations. These layers are still valid today. However, in that article a fourth section was suggested that deals with contemporary communication through the Internet, which actually *inverts the inverted triangle*, providing surprising options for identity shaping. Nowadays, on the contrary, since we no longer log onto the Net but mostly live inside it, Internet sources of identity now tend to mesh and imbricate with offline counterparts, making it difficult to address discourse and identity as represented by two non-touching triangles.

'Inherited' Sources of Identity

When a child is born, it inherits several distinctive discursive features such as belonging to a community (e.g., a nation), being male or female, and having distinctive racial or ethnic features. All of these inherited qualities have closely related discursive features (consolidated as the child grows up) ranging from living in a speech community (any human collectivity characterized by regular and frequent interactions through language) to particular sex- or race-related features.

The role which discourse plays in the community (i.e., the nation in a broad sense) as a source of identity varies depending on its current sociopolitical state, but in all cases it is one of the clearest sources of social identification: "Groups of people who occupy contiguous territory and see themselves as having common interests tend to develop, over long stretches of time, ways of speaking that are distinctive to them, marking them out from groups who either are not geographically adjacent to them or else are perceived as having different, probably rival, interests" (Joseph, 2006). This level of identity relates to Turner's (1987, p. 45) intermediate level of *ingroup–outgroup categorizations* between the broad superordinate level of being human and the subordinate level of *personal attributes*.

Needless to say, the sense of belonging to a community and sharing the same language, a major source of identity, goes beyond the political lines as drawn on maps. This is why the speech community, defined as "any human aggregate characterized by regular and frequent interaction by means of a shared body of verbal signs and set off from similar aggregates by significant differences in language use" (Gumperz, 1971, p. 114) is a much better term, since it goes beyond the reductionist view of society as a clear-cut physical space with members speaking the very same language, even if the notion of 'place' has traditionally been regarded essential to foster feelings of community membership (Walker, 2007).

On the other hand, places in which a shared language is embraced as part of a broader resistance against other imposed languages, as is the case of Quebec in Canada or Catalonia in Spain, tend to originate very tight links between discourse and identity, whereas in more politically stable linguistic areas this source of identity may not be felt so explicitly.

Another major identity marker is sex. There is now a huge amount of bibliography dealing with (non)-verbal speech differences of men and women in same-sex and cross-sex conversational interactions. To a certain extent, fitting the expected discursive attributes of one's sex may reinforce personal identity as a sense of *belonging to a social group*. General areas of research include vocabulary differences (including the traditional coining of male-connoted words) and the analysis of conversational *genderlects* (systems of linguistic features linked to one of the sexes).

Finally, the relationship between race, discourse, and identity is important since one's race may be a strong link to society as a multifaceted experience covering such aspects as the recognition of biological origins and, on some occasions, also a deep reinforcement of personality against oppressive, racist attitudes. A related issue has to do with (inherited) ethnic differences within the same community. A clear example is the United Kingdom, in which we can find a great deal of variation

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54004-5

between standard and nonstandard forms of the English language, several indigenous languages (Welsh, Scots, Gaelic), and non-English languages brought to Britain by immigrants mainly from countries in the Commonwealth. In all cases, language is raised to a basic source of personal specificity against 'default' speaking correctness.

'Optional' Sources of Identity

Next in the inverted triangle of identity sources, a middle section of optional sources may be isolated. As human beings grow up, they tend to get together in medium-sized social groups providing an essential sense of belonging and often involving particular discursive features which work as inherent sources of intragroup identity and also as intergroup differentiation (the so-called jargons). A typical example is scientific discourse. Adolescents also display explicitly narrow codes of vocabulary choices and nonstandard pronunciations only available to those 'in the gang' or the group of peers. Furthermore, laughing is considered to be an important source of identity display in this age group through reducing the 'others' to some laughable characterization that makes them different to adults (Boxer and Cortés-Conde, 1997, p. 283); and insulting one another is also typical (Mateo and Yus, 2013), both strategies intended to enhance a deeper bonding between group members.

An important term in this medium-sized area of optional group membership is social network (Milroy, 1978), a way of representing people's social transactions within a speech community. This term is related to the importance of everyday interactions (especially with peers) for identity shaping. Bucholtz and Hall (2005) stress this with their *relationality principle*, i.e., identities as "produced through contextually situated and ideologically informed configurations of self and other" (p. 605). Kiesling (2006) stresses that a person has no single fixed identity because identities are relational, they are constructed, and contextualized in interaction. Also from constructivist approaches, identity has been regarded as a nonstable but inherently interactive phenomenon, since it takes place in specific interactional occasions, yields an array of identities instead of unitary selves, and mainly results from processes of negotiation that are eminently social (De Fina et al., 2006, p. 2).

Closely related to this issue are Le Page's (1986) *acts of identity*, a term which stresses the individual's tendency to try and resemble the linguistic attributes of those groups of people which he or she aims to identify with: "when we talk we project the universe as we see it on to others as on to a cinema screen in our own images, expressed in the language we consider appropriate at that moment, and we invite others by these acts to share our universe" (p. 23).

Discourse, Identity, and the Self

The speaker's interactional self would be placed at the bottom vertex of the inverted triangle. In this case, as suggested above, self-identity is shaped and modified by daily and ordinary conversational interactions. Interlocutors continuously negotiate a *relational identity* with and through others. Consequently, one's current (and past) conversational exchanges will shape or affect the relationship between participants and speakers' relational self-identity. Far from being a stable, unitary identity, one's identities change according to conversational roles and situations. This is why De Fina et al. (2006, p. 14) distinguish *discourse identities* (mainly the individual's roles during an interaction) from *situated identities*, which are instantiated in particular types of situations. But individuals also acquire a personal discourse-related identity in the shape of personal idiolect, defined as a "particular combination of accent and dialect, that particular assemblage of formal and informal registers, that particular pattern of stress and intonation which, if we were to look closely and cleverly enough, we would find unique to the individual" (Edwards, 2009, p. 21).

Analysis of the self as developed and affirmed through daily interactions and situations goes back as far as the 1950s, with *interactional sociolinguists* (Schiffrin, 1994), with Goffman (1959) as one of the main contributors to this trend of research. Goffman argues that the self is a social construction (specifically, an *interactive construction*). This is clearly exemplified by the identity-related notion of *face*, defined as the social value that speakers effectively claim for themselves in the course of a conversational interaction. What Goffman focuses on is the social value of conversational involvement, the way different social and conversational settings involve specific expectations and display of this involvement, which normally entails the use of certain ritualized forms of address. Schiffrin (1994) also includes Gumperz in the label of interactional sociolinguistics because of his view of language as a culturally constructed system and which can be used either in order to reflect macrolevel social meanings or to create microsocial meanings at the level of ordinary conversational exchanges.

Another trend of research which focuses on the social implications of daily conversational interactions is *ethnomethodology*, with Garfinkel as one of its main representatives. This trend focuses on the cognitive rules by which members of a society assess the significance of actions in everyday life. People's previous experience and their knowledge of the institutions and practices of the world around them act to constrain their interpretations of what they see and hear. Under ethnomethodology, conversational scenes acquire an essential importance in the display and shaping of self-identities, which are only knowable through the understandings displayed by the interlocutors themselves: "membership of a category is ascribed (and rejected), avowed (and disavowed), displayed (and ignored) in local places and at certain times, and it does these things as part of the interactional work that constitutes people's lives" (Antaki and Widdicombe, 1998, p. 1).

The main conclusion of this research is that people's self-identity is not a fixed feature of their lives, but a dynamic attribute continuously under negotiation, reshaping, transformation, as people engage in daily ordinary communication with others. When some social or cultural aspects are continuously brought into conversations, they tend to become attached to the identity of interlocutors as part of their background of group membership. For example, Georgakopolou (1999) focuses on identities of the young and how *peer networks* are created in the conversational repetition of specific elements in the discursive (sub)culture of the young. For them, symbolic cultural associations such as dress styles and others such as

music preferences, activity patterns, etc. are firmly located in their conversational interactions, especially in the unfolding of conversational narratives.

On the other hand, self-identity is also achieved in the feeling of personal specificity when contrasted with others in the course of a conversation. In this case, Davies and Harre's (1990) term *positioning* as 'discursive production of selves' is very interesting. They agree that self-identity is not fixed but constituted and reconstituted through the various discursive practices in which people participate. Positioning would be the discursive process through or within which people's selves are foregrounded in conversations. There can be *interactive positioning* in the sense that what one person says positions another, and *reflexive positioning*, when one simply positions oneself.

Markus and Nurius' (1986) theory of *possible selves* also deals with the way people acquire a sense of self-identity through the conversational interactions in which they participate. These possible selves represent the picture that all interactants have about what they are and what they would like to become. In this sense, possible selves influence people's future behavior and are a constant evaluation of people's current state of their self-identity. People's social and cultural contexts play an important role in this shaping, since the range of possible selves derives not only from people's personal experiences (e.g., prior conversational interactions), but also from categories which acquire prominence in a specific sociocultural and historic context, together with images and symbols provided by (highly influential) media discourses constraining the personal choices of self-identity.

Inverting the Inverted: Discourse and Identity on the Internet

In Yus (2001), it was argued that the inverted triangle of broad top (macrosocial issues), middle section (optional linguistic groups), and the vertex at the bottom of the triangle (personal self) was affected by the *global village*. In this environment, identity is viewed as "the process of construction of meaning on the basis of cultural attribute, or related set of cultural attributes, that is are given priority over other sources of meaning" (Castells, 1997, p. 6). The important conclusion is that people tend to react by reaffirming their historical, cultural, ethnic, and religious roots as a reaction against networked globalization.

The most paradigmatic example of a global communicative network is the Internet, which is so pervasive nowadays that it actually *inverts the inverted triangle*. First, the formerly broad section of inherited attributes at the top of the triangle is narrowed as a consequence of increasing transnational communication and depersonalization of networked transactions. Hence, national attributes may become blurred and their identity-shaping effect be weakened or invalidated; and other inherited features such as race and sex are suppressed in text-based interactions.

Second, the middle section of optional sources of identity, normally through self-socialization through interactions within groups shaped by specific linguistic patterns, remains more or less in the same medium-sized section of the triangle. On the Internet, these specific discourse groups with their limited-access jargons are reproduced in parallel forms of organization: the so-called *discussion forums* and (now almost extinct) *distribution lists* via e-mail, which also exploit the same sense of group-membership language specificity that we can find in offline social instances.

Third, the former narrow vertex of personal self in the inverted triangle has now become very broad due to the multiplicity of selves (often anonymous) that may be used in conversational exchanges through various forms of Internet communication such as *chats* or avatars in 3D virtual worlds. In the physical world, the body provides explicit (and stable) clues of inherited and personal qualities shaping the person's identity. On the Internet, however, these attributes are not explicitly conveyed, and one can have as many electronic identities and selves as one is willing and or able to create, all of which are presented with no corporeality. As Mitchell (1995) correctly states, the process of mutual construction of identities on the Internet usually gives very little away. Because communication often takes place without bodily presence or the sound of users' voices, others who 'know' a user quite well may not realize how he or she looks, and thus may be unable to make the usual inferences from that. On the Internet people can very easily conceal, leave carefully ambiguous, or falsely signal gender, race, age, body shape, and economic status.

A good example of these three layers in the 're-inverted' triangle is social networking profiles (e.g., in *Facebook*). The visual list of contacts exhibits broad community membership. The comments and posts area is where users put their relational identity into practice, often with peers (sometimes also in specific professional profiles). Finally, the personal photo and information is the locus of self-identity, the user as a unique person in the network of friends.

More recent research (Yus, 2007, 2010, 2011) acknowledges the fact that this picture of discursive sources of identity as triangles (inverted for physical identity; re-inverted for virtual identity) clashes with today's tendency to hybridization of personal networks in physical and virtual settings and with the users as nodes in intersecting networks. Therefore, it would not be a picture of 'either' physical 'or' virtual sources of identity, as can be deduced from two triangles that do not touch, but a picture of several sources of identity that mix and imbricate in a time when the dividing line between physical and virtual realms is increasingly blurred.

Indeed, today's Internet users have developed an ability to switch from offline to online networks of friends with different degrees of commitment and intensity depending on what portion of their identities is shaped discursively on the Net, ranging from those who rarely log on, to those who reject their physical identities entirely and can only 'be themselves' on the Net. And environments such as social networking sites and, more recently, 3D virtual worlds are suitable environments for discursively shaped identities. Specifically, social networking sites are typically used by people who already know one another offline, e.g., students or relatives (Mallan and Giardina, 2009; Wang et al., 2010, p. 227). In general, then, people do not turn into different people in either of the environments (offline/online), but provide a different image, divide their identities into physical and virtual sides of the self. The virtual self may exhibit attributes that the user does not want to show in physical settings, without losing the core

identity. As Zhao et al. (2008, pp.1819–1820) correctly state, "users regard their online presentations as an integral part of their overall identity production and seek to coordinate their online identity claims with their offline self-performance."

In any case, in my opinion it is interesting to preserve a threefold division of identity and analyze the extent to which discourse (in its broad sense of textual, visual, and multimodal quality) plays a part in virtual identity shaping. Needless to say, there is some degree of overlapping among the three. For example, in virtual worlds the users shape their identities by devising an avatar, but the choice of body and clothing is often biased by cultural (i.e., collective) information about sex roles, masculinity, femininity, and beauty.

First, the collective identity is shaped by discursive markers that indicate membership to broad (online/offline) collectivities and communities. "It is through our membership into various social groups that we discover and experiment with social identities in the hope of creating an image of ourselves that is acceptable to the social realities we are invested in" (Crabill, 2010, p. 316). As stated in Yus (2011), people store a number of commonsense assumptions that emanate from the human environment and our trust in these assumptions is not easily altered by other in-coming stimuli. The fact that we belong to a specific speech community entails the creation and storage of certain archetypical assumptions that we accept as *normal* or *default* in the ordinary life of the community (Yus, 2007; Parks, 2011, pp. 106–111). And conversations are a good means to determine this area of mutuality. Besides, the reiterative assessment of this area generates community stereotypes, made up of highly accessible stereotypical schemas.

Second, the group identity is shaped through interactions with other users and by choosing to belong to certain groups and collectivities that exhibit certain patters and jargons that enhance discursive specificity. By receiving comments and engaging in interactions, that is, by being present in the network and being commented upon by others, users mold their position and prestige within the group.

Finally, self-identity reflects the fact that the user is a unique person discursively shaped on the Net according to specific parameters. It is related to the *self-concept*, but also shaped in a public process that involves both the *identity announcement* made by the individual claiming an identity and the *identity placement* made by others who endorse the claimed identity (Zhao et al., 2008, p. 1817). The user's body is essential for the foundation of this self-identity, even if only part (or none) of it is actually exhibited on the Net. Of course, one can also avoid the body completely and play with other selves and corporeal fabrications (as happens with 3D avatars) that might also play a part in the overall self-identity shaping. But other users will probably try to make connections between the persona exhibited online and the physical self typing on a keyboard (Walther et al., 2009, p. 232).

In some specific situations, though, real-life and virtual-life selves may overlap, creating, in extreme cases, problems of self-identity when one is unable to dissociate the physical and the virtual. In normal circumstances, though, our online selves become just one more series of selves in our everyday lives that we create in order to deal with a variety of situations.

Specifically on the relationship between discourse and identity, identity construction on the Internet is still text-based on many occasions, normally through a hybrid between the permanence and fixed quality of written texts, and the spontaneity and ephemeral status of oral utterances (Baron, 2000; Tella, 1992). Beyond the evident problems that text-based interlocutors face in the course of a conversation due to the lack of contextual support, the increasing appeal that *chats* and other interactive online phenomena have nowadays should make us think twice before claiming that the oral medium is undoubtedly superior to the written medium. The profusion of written messages that people exchange through mobile phones nowadays, for instance, is a clue suggesting that, perhaps, people are still willing to construct and communicate identities through language, but perhaps not so much through face-to-face encounters, even though today there are more and more options for synchronous face-to-face interactions involving visual and vocal channels.

See also: Community and Media; Computer Mediated Communication; Discourse, Anthropology of; Ethnomethodology, General; Identity Offline and Online; Identity and Identification, Social Psychology of; Identity in Anthropology; Language and Society; Metaphor, Role in Social Thought: History of the Concept; Nanotechnology: Social and Cultural Aspects; Narrative, Sociology of; Self: History of the Concept; Semiotics; Social and Behavioral Sciences under Dictatorship.

Bibliography

Antaki, C., Widdicombe, S. (Eds.), 1998. Identities in Talk. Sage, London.
Baron, N.S., 2000. Alphabet to Email. Routledge, London.
Boxer, D., Cortés-Conde, F., 1997. From bonding to biting: conversational joking and identity display. Journal of Pragmatics 27, 275–294.
Bucholtz, M., Hall, K., 2005. Identity and interaction: a sociocultural linguistic approach. Discourse Studies 7, 585–614.
Castells, M., 1997. The Information Age: Economy, Society and Culture. In: The Power of Identity, vol. 2. Blackwell, Oxford, UK.
Crabill, S.L., 2010. Situating social identity through language convergence in online groups. In: Dumova, T., Fiordo, R. (Eds.), Handbook of Research on Social Interaction Technologies and Collaboration Software: Concepts and Trends. Information Science Reference, Hershey.
Davies, B., Harre, R., 1990. Positioning: the discursive production of selves. Journal of the Theory of Social Behavior 20, 43–63.
De Fina, A., Schiffrin, D., Bamberg, M. (Eds.), 2006. Discourse and Identity. Cambridge University Press, Cambridge.
Edwards, J., 2009. Language and Identity. Cambridge University Press, Cambridge.
Georgakopolou, A., 1999. Doing youth in and through conversational narratives. In: Verschueren, J. (Ed.), Language and Ideology (Selected Papers from the 6th International Pragmatics Conference). IprA, Antwerp.
Goffman, E., 1959. The Presentation of Self in Everyday Life. Doubleday, New York.
Gumperz, J.J., 1971. Language in Social Groups. Stanford University Press, Stanford, CA.
Joseph, J.E., 2006. Identity and language. In: Brown, K. (Ed.), Encyclopedia of Language and Linguistics, second ed. Elsevier, Amsterdam.
Kiesling, S., 2006. Identity in sociocultural anthropology and language. In: Brown, K. (Ed.), Encyclopedia of Language and Linguistics, second ed. Elsevier, Amsterdam.
Le Page, R.B., 1986. Acts of identity. English Today 8, 21–24.
Mallan, K., Giardina, N., 2009. Wikidentities: young people collaborating on virtual identities in social network sites. First Monday 14.
Markus, H., Nurius, P., 1986. Possible selves. American Psychologist 41, 954–969.
Mateo, J., Yus, F., 2013. Towards an intercultural pragmatic taxonomy of insults. Journal of Language Aggression and Conflict 1 (1), 87–114.
Milroy, L., 1978. Language and Social Networks. Blackwell, Oxford, UK.

Mitchell, W.J.T., 1995. City of Bits. MIT Press, Cambridge, MA.
Parks, M.R., 2011. Social network sites as virtual communities. In: Papacharissi, Z. (Ed.), A Networked Self. Identity, Community, and Culture on Social Network Sites. Routledge, London.
Schiffrin, D., 1994. Approaches to Discourse. Blackwell, Oxford, UK.
Tella, S., 1992. The adoption of international communication networks and electronic mail into foreign language education. Scandinavian Journal of Educational Research 36 (4), 305–312.
Turner, J.C., 1987. A self-categorization theory. In: Turner, J.C., Hogg, M.A., Oakes, P.J., Reicher, S.D., Wetherell, M.S. (Eds.), Rediscovering the Social Group: A Self-Categorization Theory. Blackwell, Oxford, UK.
Walker, R.C., 2007. An alternative construction of identity: a study of place-based identity and its implications. American Communication Journal 9.
Walther, J.B., van der Heide, B., Hamel, L.M., Shulman, H.C., 2009. Self-generated versus other-generated statements and impressions in computer-mediated communication. A test of warranting theory using Facebook. Communication Research 36 (2), 229–253.
Wang, S.S., Moon, S.-I., Kwon, K.H., Evans, C.A., Stefanone, M.A., 2010. Face off: implications of visual cues on initiating friendship on Facebook. Computers in Human Behavior 26, 226–234.
Yus, F., 2001. Discourse and identity. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Pergamon, Oxford.
Yus, F., 2007. Virtualidades reales. Nuevas formas de comunidad en la era de Internet. University of Alicante, Servicio de Publicaciones, Alicante, Spain.
Yus, F., 2010. Ciberpragmática 2.0. Nuevos usos del lenguaje en Internet. Ariel, Barcelona.
Yus, F., 2011. Cyberpragmatics. Internet-Mediated Communication in Context. John Benjamins, Amsterdam and Philadelphia.
Zhao, S., Grasmuck, S., Martin, J., 2008. Identity construction on Facebook: digital empowerment in anchored relationships. Computers in Human Behavior 24, 1816–1836.

Discourse, Anthropology of

Charles L Briggs, University of California Berkeley, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although the concept of discourse has played a significant role in anthropological research since the 1980s, competing framings as an analytic concept or object of research have sometimes given rise to nonoverlapping, or even acrimonious conversations. Beginning in the 1990s, interest in ideologies of language, practice theory, and issues of scale prompted increasingly productive exchanges across subdisciplines and perspectives. Recent respositionings of discourse both as analytical concept and a focus of research on shifts of capital, culture, and language in contemporary landscapes sparked a proliferation of research on discourse and its frequent assimilation as a mundane component of research frameworks.

Introduction

Anthropology, particularly in France and the United States, has been shaped at various points by what has often been referred to as a 'linguistic turn,' reliance on models of language for theories of culture and society. Franz Boas, for example, used phonetic variation in explaining cultural variation; he suggested that just as each language selected a small range of sounds from a practically unlimited array and arranged them in unique ways, so culture mapped the infinitely varied range of personal experience onto a limited range of categories. This line of reasoning was extended by Edward Sapir and Benjamin Lee Whorf. In the 1960s and 1970s, Claude Lévi-Strauss drew on Roman Jakobson's and Morris Halle's distinctive feature model of phonology in creating varieties of structuralism that revolutionized anthropology, literary criticism, and other fields. During the last several decades of the twentieth century and in the twenty-first, the concept of discourse gained force in organizing research in the discipline. This article traces both the shifting ways that anthropologists have used the term and its sometimes powerful and precarious ability to organize points of scholarly convergence and divergence.

At first glance, the centrality of the concept of 'discourse' in the work of many anthropologists and other scholars in the 1980s and 1990s might seem to have represented another linguistic turn, a new convergence of linguistic and social/cultural perspectives. Nevertheless, the effect was often the opposite. Practitioners influenced by poststructuralist and postmodern theories often followed Michel Foucault in "treating discourse [not] as groups of signs…but as practices that systematically form the objects of which they speak" (1972: 49). Many linguistic anthropologists, on the other hand, challenged structuralism by emphasizing context, variation, and discursive practices; the concept of discourse was appealing as a rallying cry for going beyond decontextualized analysis of isolated signs to the analysis of communicative acts as they are embodied in specific situations. Although the latter practitioners defined discourse – generally implicitly – in a variety of ways, most associated the term with bounded texts, conversations, and electronically mediated messages, and analyses generally focused on details of form, content, and context. The result was the creation of divergent bodies of literature that carried the banner of discourse but that often had relatively little to say to one another.

With the turn of the new millennium, this polarity seemed to give way to a variety of modes in which 'discourse' became assimilated to diverse perspectives. In some cases, such as critical discourse analysis (CDA), 'discourse' provides a central analytical term. In many other instances, discourse seems to have become a relatively mundane element of a lexical stock in trade. Recent research has also explored how neoliberal transformations have focused on discursive practices as forms of labor and modes of producing value.

Discourse after Structuralism

A number of writers turned scholarship on its head in the 1960s and 1970s. Previously, academics had often claimed an epistemologically superordinate position, framing their analytic concepts and methods as privileged devices that seemed to stand above or beyond the constraints of the social actors or texts they studied. Poststructuralism challenged such claims, placing scholarship squarely into the middle of the politics of social life. Playing a key role in continually producing what were characterized as modern subjects, philosophical, and academic texts became a central object of analysis itself. Jacques Derrida suggested that concepts of language, culture, aesthetics, and so forth were in need of 'deconstruction,' of playful and irreverent analyses that unraveled the powerful illusions that imbue dominant frameworks with authority. While Lévi-Strauss characterized culture as based on unconscious oppositions, Derrida revealed the magical acts that he and other scholars used in creating these sorts of dichotomies and making them seem real.

The influential writings of Michel Foucault (1972) brought the term 'discourse' and an orientation toward 'the archaeology of knowledge' to the fore in a wide range of disciplines. He suggested that we must examine how "effects of power circulate among scientific statements, what constitutes, as it were, their internal régime of power" (1980: 112). Foucault did not identify discourse with particular utterances or texts, and he argued that it is not possible to grasp the power of discourse when it is seen as rooted in a particular context, viewed as an isolated and self-sufficient totality. He rather suggested that "a statement always belongs to a series or a whole, always plays a role among other statements, deriving support from them

and distinguishing itself from them" (1972: 99). One of his major contributions was demonstrating that institutions form key loci for the production and legitimization of discourses, and he identified how they are embodied in 'political technologies' that create schemes of surveillance and regulation in such settings as prisons, clinics, and asylums.

Foucault's work has been criticized for focusing on dominant texts, thereby reproducing the marginalization of contributions by women, racialized populations, and people outside of Europe. Particularly in his early work, Foucault seems to go against his disavowal of structuralism and his call to look beyond the dispersion and mutability of elements of discourse in favor of finding of "a unity …[that] resides, well anterior to their formation, in the system that makes possible and governs that formation" (1972: 72).

A complementary point of departure was provided by attempts to reverse the long history of attempts to reduce rhetoric to forms of ornamentation and techniques of persuasion, a central bête noire during the time of the emergence of modern science in the seventeenth century (Bauman and Briggs, 2003). Hayden White analyzed historical texts as narratives, suggesting that how historians construct the past involves two steps – an act of imagination that creates a vision of the past and a subsequent act of rhetorical magic that renders the imaginary real, convincing audiences that that is just the way it happened. Attending to issues of both rhetoric and power, practitioners have critically examined how anthropology similarly contributed to continual reproduction of categories of racialized Others. James Clifford and George Marcus (1986) and other writers dissected 'the poetics and politics' of anthropological texts, particular rhetorical strategies that ethnographers used in imagining 'the field' and imbuing their constructions with authority and authenticity. When viewed critically as power-laden discourse, ethnography became suspect, particularly in the eyes of scholars in cultural and ethnic studies. Unfortunately, classic deconstructions by such scholars as Américo Paredes (1993[1977]) of ethnographic rhetorics that stereotyped racialized minorities were seldom cited as precedents.

Anthropologists have also detailed how complex, multifaceted, and contested relations are discursively transformed into seemingly simple dichotomies, such as savage *vs* civilized and traditional *vs* modern, through colonial discourses. By revealing the complicity of ethnography in colonial regimes and the colonial underpinnings of contemporary anthropology, scholars deepened the critique of fundamental anthropological concepts and practices. Michael Taussig (1987) drew on the work of the Frankfurt School in revealing the way that the production of racialized images went hand in hand with the violent oppression of indigenous Latin Americans. Students of colonialism and others have turned to theorists who come from outside the Euro-American orbit, including the Subaltern Studies group in India and such Latin American theorists, as Néstor García Canclini (1995[1989]), for points of departure.

Finally, a number of scholars responded to analyses of the 'free floating' character of postmodern discourse by experimenting with techniques of ethnographic analysis and writing. Anthropologists sought to disrupt dominant rhetorical strategies and political positions by creating ethnographic and other texts that drew on surreal, fictional, and other techniques. The idea that both culture and ethnography are dialogic, tied to M.M. Bakhtin's conception of discourse, has influenced many of these innovations. Concern with cultural discourses were thus translated into attempts to a range of strategies for transforming anthropological discourses of culture.

Linguistics Approaches to Discourse

Other scholars used the concept of discourse to grapple with limitations in existing approaches to language. Structural linguists largely followed Ferdinand de Saussure in constructing languages as quintessentially embodied in closed systems of context-independent signs that stand apart from politics, history, and social differences between speakers. Noam Chomsky pushed many linguists toward even more formal, abstract approaches and further marginalized the study of linguistic heterogeneity by positing an ideal speaker–listener as the model of the subject of language. An orientation toward discourse promised to free practitioners from focusing exclusively on referential relations as encoded in small, decontextualized units in favor of studying the way that texts, conversations, and entire genres were embedded in social life. In opposition to poststructuralist and postmodern rejections of 'positivistic' definitions, many linguistic anthropologists envisioned discourse as a means of achieving greater empirical scope and depth.

Beginning early in the twentieth century, 'Russian Formalists' focused on such discursive units as stories or poems, detailing how writers played with linguistic forms apart from or in playful relationship to their semantic content. The Prague School developed a dynamic approach that emphasized expressive, appellative, and aesthetic functions in addition to the referential function and examined how they intersected in different each genres. Roman Jakobson figured prominently in the development of both of these approaches. He drew on an account of semiotics devised by philosopher Charles S. Peirce, one that emphasized types of signs that revolved around existential relationships to particular contexts and nonarbitrary resemblances between sign vehicles and their objects as well as context-free symbols.

Russian V.N. Vološinov (1973[1930]) took on Saussure directly, drawing attention to the capacity of signs to link multiple contexts, voices, and texts and detailing the special role of reporting the words of others in creating this multiplicity. While Saussure viewed discourse in its particularity as only a point of departure for abstracting underlying systems of signs, Vološinov centered analysis on the utterance. M.M. Bakhtin (1981) characterized discourse as heteroglossic, suggesting that languages are not closed, stable systems but points of intersection between competing structures, styles, genres, voices, and perspectives. Rather than the sources of social consensus – and national unity – that Saussure had envisioned, languages became sites of struggle. British philosopher J.L. Austin (1962) transformed received understandings of the relationship between word and world by arguing that utterances are performative – that they do not simply refer to real world events but create them; saying, in short, is a form of doing. Linguists have mapped the grammatical and 'discourse level' features of texts and genres and tied them to communicative functions, as in Michael Halliday's systemic-functional approach.

In the 1960s and 1970s, American scholars drew on these insights in challenging dominant structuralist and transformational (Chomskyian) approaches. Dell Hymes (1974) called for an 'ethnography of speaking' that documented the social and cultural patterning of language in its social context. While Saussure and Chomsky banished linguistic variation, the way that forms and meanings vary between speakers and contexts, from serious linguistic study, Hymes and other scholars placed it at the center of an emerging sociolinguistics. A generation of practitioners who obtained their PhD's in the late 1960s and early 1970s conducted ethnographies of speaking that focused on such issues as gender, political and religious ritual, rank and status, and expressive genres. Richard Bauman and other writers illustrated how this approach can be used as a tool for historical analysis. In the mid-1970s, Bauman (1977) and Dell Hymes (1981) complemented the emphasis on poetics generated by Roman Jakobson and other scholars by focusing on performance. William Labov developed a varationist sociolinguistics that mapped the social distribution of language varieties and the expressive functions of these differences. John Gumperz (1982) crafted an approach to interactional sociolinguistics that looked closely at conversational inferences made by persons engaged in face-to-face interaction and contextualization cues, dimensions of discourse that speakers use in shaping and interpreting contexts. Aaron Cicourel (1974) analyzed sociological research as discourse, conducting ethnographic research on surveys and other techniques in ways that revealed how claims to validity and reliability are based on complex interactional work that is generally rendered invisible as the analysis proceeds. Michael Silverstein (1976) drew on the work of Charles Peirce and Roman Jakobson in developing a 'metapragmatic' approach that complements the study of semantics with rigorous attention to the centrality of indexes that position discourse in relation to existing social contexts – and create new contextual parameters. The frameworks these scholars developed helped to focus attention on linguistic diversity and language policies in contemporary nation-states, presaging later interest in discursive dimensions of nationalism, power, conflict, and social inequality. They placed questions of communicative functions – the multiple roles of language in shaping social life – alongside those of structure at the center of inquiry.

Other scholars working in the United States sought to draw the study of discourse away from the study of types, rules, norms, and functions; they defined discourse as – and limited objects of analysis – to particular stretches of talk. Sociologists Harold Garfinkel and Erving Goffman argued that discourse should be studied not as an end in itself but as a window on social interaction. Garfinkel's concern with everyday, mundane encounters was placed at the center of the field of conversation analysis, which proclaimed "an insistence on the use of materials collected from *naturally occurring* occasions of everyday interaction by means of audio- and video-recording equipment or film" (Heritage and Atkinson, 1984: 2; emphasis in original). Emanuel Schegloff (2006) rejected analytic recourse to contextual information that did not appear in the audio or video recording itself; conversation analysts scrutinized transcripts for regularities in the sequencing of conversational features. Many practitioners focused on institutional contexts, detailing the role of interaction in shaping what takes place in courtrooms, schools, clinics, and other settings. Joel Sherzer proposed a 'discourse-centered' framework that similarly privileged 'naturally occurring and recorded speech' (1983: 10) at the same time that he called for systematic attention to broader social and cultural contexts. Such approaches represent an opposite pole from poststructuralist frameworks in that they view discourse in maximally empirical terms and focus on minute analysis of recorded talk.

Intersections: The Politics of Discourse

In short, the life history of 'discourse' in anthropological writing during the 1980s seems to follow a number of distinct, semi-autonomous trajectories. When paths intersected, they often generated examples of boundary-work, efforts to heighten differences between approaches and defend academic turf; debates sometimes developed acrimonious overtones. Although it would not be difficult to locate instances of this trend up to the present, the 1990s seemed to give way to interest on the part of many anthropologists and other scholars on the role of discourse in shaping power and the politics of difference, resulting in a number of important intersections between practitioners who formerly held opposing views of discourse.

Pierre Bourdieu (1991) placed sociolinguistic scholarship into a framework that defined types of linguistic competence as 'symbolic capital' in seeking to reveal the role of discourse in processes of 'symbolic domination.' He redefined the concept of communicative competence in such a way as to draw attention to the barriers of class and race that limit access to sites in which such competence can be acquired. Aaron Cicourel, Charles Briggs, Elliot Mishler, and others have focused on interviews as social as well as communicative events, detailing how formal and contextual features shape and are shaped by power asymmetries, thereby creating discourses that scholars use in imagining social life and imbuing their texts with authority and authenticity.

Bakhtin (1981) emphasized the heterogeneity of discourse and how it is constantly in movement, even within what seems to be a bounded, fixed text. John Gumperz (1982) urged concern with how contextual parameters are continually made, differentially interpreted, and sometimes contested in social interactions, particularly as they cross perceived boundaries of race and class. In the 1990s, linguistic anthropologists attempted to account for the complexity and instability of relations between texts, contexts, voices, and subjects. Richard Bauman and Charles Briggs (1990) proposed a framework for analyzing how entextualization uses formal and referential dimensions in such a way as to project a process of decontextualization from antecedent texts, contexts, genres, and persons, simultaneously projecting discourse forward in terms of how it may or must be recontextualized. Michael Silverstein and Greg Urban (1996) focused on entextualization as a dynamic process in which discourse gets reified as texts, thereby taking on social lives that shape and are shaped by the sites, voices, social processes, and other dimensions of cultural worlds in which they come to be repositioned.

Building on Silverstein's foundational work, linguistic anthropologists have explored the role of language ideologies, socially situated conceptions of the nature and functions of language, speech, and communication, in shaping both

linguistic structure and social life (see Schieffelin et al., 1998). This work fostered conversations across subdisciplinary boundaries as linguistic anthropologists used the concept of language ideologies in exploring problems of central interest to sociocultural anthropologists and other scholars by demonstrating the central role of social constructions of language, text, and talk in shaping notions of gender, race, religion, space, the body, and democracy.

This work sometimes took a historical turn. Judith Irvine examined a language of linguistic kinship in tracing how 'language families' became a central framework for colonial projects in Africa. William Hanks (2010) argued that Mayan languages emerged in the context of colonialism rather than as pristine autochthonous voices; just as they structure modes of transforming indigenous societies, literacy practices provided Mayan elites with ways of appropriating forms of colonial authority. Bernard Cohn, Nicholas Dirks, Gloria Raheja, and others have explored how deeply, in Cohn's (1996) terms, 'the language of command and the command of language' were inscribed in British colonialism in South Asia. Bauman and Briggs (2003) examined the centrality of ideological constructions of language to shifting projects of modernity, arguing that John Locke's separation of language from natural and social worlds in the seventeenth century was as crucial for creating social categories, policing boundaries, and naturalizing forms of inequality as the nature *vs* culture, science *vs* politics 'Great Divide' that formed the focus of research by science–technology–society studies' scholars. Webb Keane (2007) suggested that what he referred to as 'semiotic ideologies' crucially informed Dutch Calvinist missionization in Indonesia, teaching colonial subjects the proper boundaries between subjects, texts, and objects.

The ideological construction of multilingualism similarly provided a fruitful arena of research. Kathryn Woolard (1989) pinpointed how the enactment of 'English only' laws and other legal mechanisms could rally conservatives and liberals, Latinos/as and non-Latinos around seeing Spanish in the United States as a threat to equality and democracy. In sensitive studies that located what seemed to be problems created by Spanish/English bilingualism in New York City, Bonnie Urciuoli (1996) and Ana Celia Zentella (1997) pointed to ways that dominant language ideologies and institutional practices naturalize structural inequities by constructing working-class Spanish–English bilinguals as lacking competence in either language, viewing their seeming linguistic incompetence as a cause of their racially subordinate position. Monica Heller (2006) charts the complex politics and economics of French–English bilingualism in Canada. Nicholas DeGenova (2005) and Hillary Dick (2011) analyze both discourses of (im)migration and how discursive ideologies and practices construct the categories of 'native,' 'immigrant,' and 'illegal alien.' Norma Mendoza-Denton (2008) uses both quantitative and qualitative variationist sociolinguistic studies and an ethnographic exploration of bodily constructions in interrogating a popular construct, 'Latina gang members.' Jane Hill (2008) analyzes faux bilingualism, the purportedly playful use of what seem to be Spanish expressions by monolingual English speakers. Hill analyzes the role of mock Spanish in reproducing whiteness and naturalizing neoracism ("I'm not racist, but…"). Leo Chávez (2001) and Otto Santa Ana (2002) look respectively at visual and textual representations of Latin American migrants to the United States as a flood of racialized, inferior, and alien bodies.

Anthropologists and other scholars responded to Jurgen Habermas's work on the construction of a modern 'public sphere' in which citizens debate discourses that shape social life and political institutions by detailing the way that dominant forms and arenas are contested by resistant voices and counterpublics. Such investigations have helped to illuminate the processes by which limits are imposed on what sorts of things can be said and who can legitimately say them. Linguistic anthropologists drew attention to ways that public constructions of languages and linguistic diversity crucially inform the processes of making publics (Gal and Woolard, 2001). Anthropologists have explored the powerful role of such communication technologies as newspapers, gramophones and phonographs, radios, and television in producing publics and types of political subjectivity, responding to Jesús Martín Barbero's call to attend to forms of mediatization that help produce constructions and experiences of the social rather than beginning with received notions of 'media' and 'society.'

CDA grew out of linguistics with a focus on issues of power and social/political domination. While some varieties, such as the work of Teun van Dijk, have focused primarily on features of media and other publicly circulating texts, Norman Fairclough incorporated social theory and emphasized the importance of studying 'discourse-as-discursive-practices' and 'discourse-as-social-practice' as well as 'discourse-as-text' (1992). One of the major foci has been discursive dimensions of racism; given the location of many CDA scholars, racism in Europe and the UK has been a major focus. For example, Jan Blommaert and Jef Verschueren (1998) examined the representation of immigrants in news accounts, and Ruth Wodak focused on the discourse of the far right in Europe (Forchtner et al., 2013).

One focus in linguistic anthropology in the opening decade of the twenty-first century was on linking language ideologies to interactional practices, detailing how representations of language varieties, speakers, speech acts, and ongoing speech interact closely with pragmatic details yet without ever determining or fully mapping them. Silverstein (2004) pointed to the need to grapple with issues of scale by linking the study of these relations in recordings of conversations and broad features of social life. Asif Agha (2007) explored how complex interrelations between language and social relations shape models of conduct, suggesting how shared interpersonal engagements realized through everyday language use shape the emergence of difference, contestation, and differential value. Ben Rampton's (1995) notion of crossing explores how people in multiracial settings use language varieties associated with populations to which they do not belong. Charles Briggs (2005) looks at the circulation of discourse both in terms of observed practices, circuits, and mediatized processes and what he refers to as communicability, the competing cultural models that project how discourse is produced, how it travels, and how it is received.

The Dispersion of 'Discourse' in the New Millennium

The last decade or so has witnessed the emergence of a host of conversations in which discourse both figures in the

articulation of analytic perspectives and becomes an object of analysis. J.L. Austin's attention to performativity was refocused through the work of Derrida (1977[1972]); work on gender and sexuality by Judith Butler (1993) placed the body centrally into discussions. Linguistic anthropologists have contributed in many ways to rethinking the place of language and discursive practices in the production, performance, and contestation of gendered and sexualized bodies, voices, and subjectivities. At the same time that William Leap (1996) called for documentation of the speaking practices of gay men, Mary Bucholtz (2003), Mary Bucholtz and Kira Hall (1995), Deborah Cameron and Don Kulick (2003), and Bonnie McElhinny (2007) have drawn on feminist and lesbian, gay, bisexual, transgender studies in pointing to more complex relationships between gender, sexuality, and language and challenging patriarchal and heteronormative assumptions about language, subjectivity, and social life. Work such as Norma Mendoza-Denton's (2008) study of Latina gangs examines how discourse, bodies, race, and sexuality intersect.

One of the most visible arenas of anthropological research in recent decades has involved a focus on science, medicine, and the body. This rapidly expanding field of research, which cannot be effectively summarized in a few lines, has explored how discourses of aging, HIV/AIDS, sexuality, epidemics, pharmaceuticals, genetics, reproduction, addiction, mental illness, and scientific evidence have circulated alongside technologies, practitioners, bodies, organs, drugs, disease, humanitarian organizations, and regimes of health, security, and preparedness. A shift from medicalization to biomedicalization links transformations of communicative, clinical, and public health infrastructures through posthumanist frameworks that create a multiplicity of nature/culture/technology amalgamations. Here, nonetheless, boundary-work seems to obstruct the development of more fruitful collaborations as work on these topics emerging within linguistic and medical anthropology tends to travel through parallel circuits separated by specialized vocabularies. Veena Das's (2006) work on narrative, violence, and bodies, which focused on the Partition of India and Pakistan and anti-Sikh riots, pointed to the rich potential of simultaneously attending to both *Life and Words*.

Much research has explored transformations of discourse associated with late modernity and neoliberalism. Critiques of previous anthropological work on globalization fostered a range of perspectives that queried how discourses, people, commodities, technologies, and epistemologies circulate as well as the metaphors and frameworks that scholars used to track them. Anna Tsing (2005) argued for a more nuanced metaphor of 'friction,' while Aihwa Ong (Ong and Collier, 2005) and Saskia Sassen (2006) suggested 'global assemblages.' Webb Keane (2007) argued that separating discourse from materiality is itself a reflection of semiotic ideologies of modernity. A number of anthropologists have recently explored the commodification of discourse and discourses of commodification through, for example, advertising (Dávila, 2001; Mazzarella, 2005) and ethnicity (Comaroff and Comaroff, 2009). Anthropologists and ethnomusicologists have traced the commodification and circulation of musical forms; Steven Feld's (2000) analysis of world musics and Samy Alim's (2006) research on hip hop draw attention to the diversity of analytics and assessments. Charles Hirschkind (2006) and other anthropologists have contributed centrally to shifting the locus of attention on performativity from an exclusive concern on speaking and writing to a focus on listening, which requires attention to bodies, technologies, and materialities.

The proliferation of sites, technologies, and practices associated with the digitalization of discourse have become the focus of much research. Countering initial formulations that often projected new communicative technologies as possessing the power to produce new ideologies, cultural configurations, social relations, and political economies, many authors have examined dimensions of remediation, seeing how perspectives and practices associated with prior technologies are repositioned alongside novel forms. Ilana Gershon (2010) traced the role of Facebook in how couples are formed and break up, examining the way that face-to-face interaction comes to be seen as just one more communicative possibility in an array of Internet- and social media-based modalities. Studying the world of on-line games ethnographically, Tom Boellstorff (2008) argued that such forms of virtual embodiment provide sites for rethinking dimensions of gender, sexuality, race, the material, and sociality itself. Dominic Boyer (2013) scrutinized the ongoing transformation of journalists from beat reporters on the street to office-based screen workers who focus less on the production of knowledge than on managing vast, multiple, and rapidly circulating bodies of discourse.

Finally, recent work on globalization and neoliberalism has challenged scholars to address issues of scale and complexity in both perspectives on discourse and how discursive sites and practices are documented and analyzed. Jan Blommaert (2010) pushed for new theoretical perspectives on mobility and modes of analysis that address both global connectedness and the emergence of new inequalities and forms of disconnectedness. A major focus of recent work in linguistic anthropology and sociolinguistics has been on language and/as labor, "where language is central not just as the process but as the product of work" (Duchêne and Heller, 2012: 13). Although call centers provide quintessential examples in which corporate strategies and digital networks script conversations enacted by workers living in areas where wages are lower, such 'language industries' include tourism, translation, surveillance, media, and language schools as sites where linguistic competence, discursive practices, and language ideologies are commodified (see Urciuoli and LaDousa, 2013).

Conclusion

It has been a daunting task to sort through the many ways that notions of discourse have been developed in anthropology. After gesturing at the vast discussions of concepts of language and culture and their interrelations that emerged in the first half of the twentieth century, I have traced a number of trajectories in which discourse figured as a key locus for articulating anthropological positions and defending their borders against competing perspectives as well as points at which the term helped shape convergence on issues that interest a diversity of scholars. This last article of the story prompts me to end on a rather paradoxical note. The beginning of the new millennium

has witnessed the emergence of work that productively exploits the tension between construing discourse as scholarly concept and as object of analysis. The very fruitfulness of this move and the range of dimensions of human life that practitioners have queried have led, ironically, to broad assimilation of 'discourse' as a relatively mundane part of analytic languages in a number of types of scholarly practice.

See also: Intercultural Communication: An Anthropological Perspective; Performance: Anthropological Aspects.

Bibliography

Agha, A., 2007. Language and Social Relations. Cambridge University Press, Cambridge.

Alim, H.S., 2006. Roc the Mic Right: The Language of Hip Hop Culture. Routledge, New York.

Austin, J.L., 1962. How to Do Things with Words. Harvard University Press, Cambridge, MA.

Bakhtin, M.M., 1981 (C. Emerson, M. Holquist, Trans.). In: Holquist, M. (Ed.), The Dialogic Imagination: Four Essays. University of Texas Press, Austin, TX.

Bauman, R., 1977. Verbal Art as Performance. Waveland, Prospect Heights, IL.

Bauman, R., Briggs, C.L., 1990. Poetics and performance as critical perspectives on language and social life. Annual Review of Anthropology 19, 59–88.

Bauman, R., Briggs, C.L., 2003. Voices of Modernity: Language Ideologies and Social Inequality. Cambridge University Press, Cambridge.

Blommaert, J., 2010. The Sociolinguistics of Globalization. Cambridge University Press, Cambridge.

Blommaert, J., Verschueren, J., 1998. Debating Diversity: Analysing the Discourse of Tolerance. Routledge, New York.

Boellstorff, T., 2008. Coming of Age in Second Life: An Anthropologist Explores the Virtually Human. Princeton University Press, Princeton, NJ.

Bourdieu, P., 1991. Language and Symbolic Power, Gino Raymond and Matthew Adamson, Trans. Harvard University Press, Cambridge, MA.

Boyer, D., 2013. The Life Informatic: Newsmaking in the Digital Era. Cornell University Press, Ithaca, NY.

Briggs, C.L., 2005. Communicability, racial discourse, and disease. Annual Review of Anthropology 34, 269–291.

Bucholtz, M., 2003. Theories of discourse as theories of gender: discourse analysis in language and gender studies. In: Holmes, J., Meyerhoff, M. (Eds.), The Handbook of Language and Gender. Blackwell, Oxford, pp. 43–68.

Bucholtz, M., Hall, K. (Eds.), 1995. Gender Articulated: Language and the Socially Constructed Self. Routledge, New York.

Butler, J., 1993. Bodies that Matter: On the Discursive Limits of "Sex". Routledge, New York.

Cameron, D., Kulick, D., 2003. Language and Sexuality. Cambridge University Press, Cambridge.

Chávez, L.R., 2001. Covering Immigration: Popular Images and the Politics of the Nation. University of California Press, Berkeley, CA.

Cicourel, A.V., 1974. Theory and Method in a Study of Argentine Fertility. Wiley Interscience, New York.

Clifford, J., Marcus, G.E. (Eds.), 1986. Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley, CA.

Cohn, B.S., 1996. Colonialism and Its Forms of Knowledge: The British in India. Princeton University Press, Princeton, NJ.

Comaroff, J., Comaroff, J., 2009. Ethnicity, Inc. University of Chicago Press, Chicago, IL.

Das, V., 2006. Life and Words: Violence and the Descent into the Ordinary. University of California Press, Berkeley, CA.

Dávila, A., 2001. Latinos, Inc.: The Marketing and Making of a People. University of California Press, Berkeley, CA.

De Genova, N., 2005. Working the Boundaries: Race, Space, and "Illegality" in Mexican Chicago. Duke University Press, Durham, NC.

Derrida, J., 1977[1972]. Limited, Inc. Northwestern University Press, Evanston, IL.

Dick, H.P., 2011. Language and migration to the United States. Annual Review of Anthropology 40, 227–240.

Duchêne, A., Heller, M. (Eds.), 2012. Language in Late Capitalism: Pride and Profit. Routledge, New York.

Fairclough, N., 1992. Discourse and Social Change. Polity, Cambridge.

Feld, S., 2000. Sweet lullaby for world music. Public Culture 12 (1), 145–171.

Forchtner, B., Krzyzanowski, M., Wodak, R., 2013. Mediatization, right-wing populism, and political campaigning: the case of the Austrian Freedom Party. In: Tolson, A., Ekström, M. (Eds.), Media talk and Political Elections in Europe and America. Palgrave Macmillan, Basingstoke, pp. 205–228.

Foucault, M., 1972. The Archaeology of Knowledge (A.M. Sheridan Smith, Trans.). Harper and Row, New York.

Foucault, M., 1980. Power/Knowledge: Selected Interviews and Other Writings, 1972–1977 (C. Gordon, et al., Trans.). Pantheon, New York.

Gal, S., Woolard, K.A. (Eds.), 2001. Languages and Publics: The Making of Authority. St. Jerome, Manchester.

Garcia, C.N., 1995[1989]. Hybrid Cultures: Strategies for Entering and Leaving Modernity (C.L. Chiappari, S.L. López, Trans.). University of Minnesota Press, Manchester, MN.

Gershon, I., 2010. The Breakup 2.0: Disconnection over New Media. Cornell University Press, Ithaca, NY.

Gumperz, J.J., 1982. Discourse Strategies. Cambridge University Press, Cambridge.

Habermas, J., 1989[1962]. The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society (T. Burger, F. Lawrence, Trans.). MIT Press, Cambridge, MA.

Hanks, W.F., 1990. Referential Practice: Language and Lived Space Among the Maya. University of Chicago Press, Chicago, IL.

Hanks, W.F., 2010. Converting Words: Maya in the Age of the Cross. University of California Press, Berkeley, IL.

Heller, M., 2006. Linguistic Minorities and Modernity: A Sociolinguistic Ethnography. Continuum, London.

Heritage, J., Atkinson, J.M., 1984. Introduction. In: Atkinson, J.M., Drew, P. (Eds.), Structures of Social Action. Cambridge University Press, Cambridge, pp. 1–15.

Hill, J.H., 2008. The Everyday Language of White Racism. Wiley-Blackwell, Chichester.

Hirschkind, C., 2006. The Ethical Soundscape: Cassette Sermons and Islamic Counterpublics. Columbia University Press, New York.

Hymes, D., 1974. Foundations in Sociolinguistics: An Ethnographic Approach. University of Pennsylvania, Philadelphia, PA.

Hymes, D., 1981. In: Vain I Tried to Tell You": Essays in Native American Ethnopoetics. University of Pennsylvania Press, Philadelphia, PA.

Jakobson, R., 1960. Closing statement: linguistics and poetics. In: Sebeok, T.A. (Ed.), Style in Language. MIT Press, Cambridge, MA, pp. 350–377.

Keane, W., 2007. Christians and Moderns: Freedom and Fetish in the Mission Encounter. University of California Press, Berkeley, CA.

Labov, W., 1972. Sociolinguistic Patterns. University of Pennsylvania Press, Philadelphia, PA.

Leap, W., 1996. Word's Out: Gay Men's English. University of Minnesota Press, Minneapolis, MN.

Mazzarella, W., 2005. Shoveling Smoke: Advertising and Globalization in Contemporary India. Duke University Press, Durham, NC.

McElhinny, B. (Ed.), 2007. Words, Worlds, Material Girls: Language and Gender in a Global Economy. Mouton de Gruyter, Berlin.

Mendoza-Denton, N., 2008. Homegirls: Language and Cultural Practice Among Latina Youth Gangs. Blackwell, Malden, MA.

Ong, Aihwa, Collier, Stephen J. (Eds.), 2005. Global Assemblages: Technology, Politics, and Ethics as Anthropological Problems. Blackwell, Malden, MA.

Paredes, A., 1993[1977]. On ethnographic work among minority groups: a folklorist's perspective. In: Bauman, R. (Ed.), Folklore and Culture on the Texas-Mexican Border. University of Texas Center for Mexican American Studies, Austin, pp. 73–110.

Pigg, S., 2001. Language of sex and AIDS in nepal: notes on the social production of commensurability. Cultural Anthropology 16 (4), 481–541.

Rampton, B., 1995. Crossing: Language and Ethnicity Among Adolescents. Longman, London.

Santa Ana, O., 2002. Brown Tide Rising: Metaphors of Latinos in Contemporary American Public Discourse. University of Texas Press, Austin, TX.

Sassen, S., 2006. Territory, Authority, Rights: From Medieval to Global Assemblages. Princeton University Press, Princeton, NJ.

Schegloff, E.A., 2006. Sequence Organization in Interaction: A Primer in Conversation Analysis. Cambridge University Press, Cambridge.

Schieffelin, B., Woolard, K.A., Kroskrity, P.V. (Eds.), 1998. Language Ideologies: Practice and Theory. Oxford University Press, New York.

Sherzer, J., 1983. Kuna Ways of Speaking. University of Texas Press, Austin.

Silverstein, M., 1976. Shifters, linguistic categories, and cultural description. In: Basso, K., Selby, H.A. (Eds.), Meaning in Anthropology. University of New Mexico Press, Albuquerque, pp. 11–55.
Silverstein, M., 2004. "Cultural" concepts and the language-culture nexus. Current Anthropology 45 (5), 621–652.
Silverstein, M., Urban, G. (Eds.), 1996. Natural Histories of Discourse. University of Chicago Press, Chicago, IL.
Taussig, M., 1987. Shamanism, Colonialism and the Wild Man: A Study in Terror and Healing. University of Chicago Press, Chicago, IL.
Tsing, A., 2005. Friction: An Ethnography of Global Connection. Princeton University Press, Princeton, NJ.
Urciuoli, B., 1996. Exposing Prejudice: Puerto Rican Experiences of Language, Race, and Class. Westview, Boulder, CO.
Urciuoli, B., LaDousa, C., 2013. Language management/labor. Annual Review of Anthropology 42, 175–190.
Vološinov, V.N., 1973[1930]. Marxism and the Philosophy of Language (L. Metejka, I.R. Titunik, Trans.). Seminar Press, New York.
Woolard, K.A., 1989. Sentences in the language prison: the rhetorical structuring of an American language policy debate. American Ethnologist 16 (2), 268–278.
Zentella, A.C., 1997. The hispanophobia of the official english movement in the US. International Journal of the Sociology of Language 127, 71–86.

Discourse Studies

Johannes Angermuller, Centre for Applied Linguistics, The University of Warwick, Coventry, UK; and Centre d'Étude des Mouvements Sociaux (CEMS), École des Hautes Études en Sciences Sociales (EHESS), Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In Discourse Studies, discourse is typically seen as a social practice of participants communicating through linguistic and other semiotic resources in certain contexts. Researchers in this field often insist on the constitutive character of discourse for social realities, structures and subjectivities. This article discusses both discourse theories (e.g., poststructuralist, deliberative, critical realist strands) as well as the more empirical approaches in discourse analysis (e.g., semantic and pragmatic, qualitative and quantitative strands). Special mention is made of developments in the West, notably in France and the UK, where Discourse Studies first started to emerge around 1970.

Introduction

In the last 50 years, 'discourse' has become a key notion in the interdisciplinary debate on the social sciences and humanities. Researchers who deal with discourse typically investigate the ways in which social realities are constructed through communicative practices. While many researchers who refer to discourse in their disciplines share an interest in language, communication, and meaning-making as central problems in contemporary society, 'discourse' is increasingly becoming the common denominator of a new, transdisciplinary field: Discourse Studies.

Discourse Studies has progressed as a reaction to developments both inside and outside the academic domain. Thus, developments of great topical relevance have accompanied the emergence of Discourse Studies: discrimination against social groups in the mass media, market-based modes of governing large populations, postnational regimes of exclusion in global postmodernity, the crisis of representation in culture and politics, and new technologies of surveillance and control.

If linguists with a social sciences background have, crucially, contributed to this field, many discourse researchers come from other fields in the social sciences and humanities (e.g., sociology, political science, history, philosophy, literary criticism, geography, education, economics, psychology, etc.). Established in France in the late 1960s and in the UK in the 1970s, Discourse Studies is today growing rapidly in many parts of the world, notably in Latin America and in China.

In Discourse Studies, researchers typically mobilize analytical and theoretical resources from the entire spectrums of the social sciences and humanities in order to account for the social production of meaning. Situated within the transdisciplinary space of language and society, Discourse Studies overlaps with a number of other fields such as pragmatics, sociolinguistics, semiotics, rhetoric, conversation analysis, linguistic anthropology, philosophy/sociology and so on of language. Among the characteristic features of discourse research, one can detect a focus on these questions:

- *Discourse as a sociohistorical meaning-making practice in context*. Meaning is not inherent in the signs, utterances or other semiotic resources that discourse participants use to communicate with each other. It is a product of linguistic and communicative practices whose meaning is inextricably tied to their contexts, material mediums, physical settings and participating bodies in and through which discursive practices occur.
- *The societal relevance of themes and topics*. While discourses can cover a wide range of domains, they often touch on questions and problems of special societal relevance, notably the existential moral and political questions of larger communities. Discourses tend to deal with knowledge circulating across many different contexts, mobilizing considerable resources and power and established through social institutions. Thus they often have an impact beyond the local and specific contexts in meaning-making activities that take place.
- *The discursive construction of subjectivity*. Even though discourse is considered to be a social practice, it does not presuppose that the discourse participants are always free to say and think what they want. On the contrary, their discursive practice is subject to rules and constraints which they do not always control. While discourse typically exceeds what they mean to say, think, or do, they become subjects only by entering discourse. Subjects, actors or agents are, themselves, a discursive effect of discourse, and not its origin.
- *Discourse as constitutive of the social*. From a discourse point of view, social structures are produced by meaning-making activities involving the use of language. At the same time, discursive practices are shaped by social structures. Discursive practices not only represent the social world but, in certain cases, they also constitute it – think of the act of ordering a beer in a pub, which creates rather than describes the relationship between customer and bar man, or the publication of statistics on smoking patterns, which can make a social problem real and objective for society.

While, in a most general sense, discourse deals with the social production of meaning, its meaning can vary: from discourse as text-based communication in large communities (especially among discourse researchers from Europe) to discourse as situated talk or turn-taking process in a setting (especially among discourse researchers from North America). The problems of discourse researchers can vary according to their disciplinary background: while linguistic discourse researchers usually want to go beyond the abstract – pure, internalist conceptions of language of classical structural or Chomskyan linguistics – sociologists and political scientists have turned to discourse in order to point out the symbolic construction of actors and 'reality.' Even though Discourse

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03216-5

Studies, just like sociolinguistics, has long been a subfield of linguistics, it is now becoming a transdisciplinary field at the crossroads of language and society, one which mobilizes theories and methods from all corners of the social sciences and humanities in order to study the ways in which linguistic, semiotic, and communicative resources and practices are mobilized in social communities to represent and constitute social phenomena (such as subjectivities and identities, relationships and structures, hegemonies and ideologies, knowledges and ideologies).

To account for the development of Discourse Studies in the last 50 years, this article will concentrate on (1) strands of (a) discourse theory and (b) discourse analysis and discuss (2) Discourse Studies as a global field.

Discourse Studies as Theory and Analysis

Discourse Studies is organized around characteristics split between two ideal-typical strands: one focussing on discourse as an intellectual and epistemological problem of social, political and cultural theory, the other on the analytical and methodological challenges of studying discourse as a material and empirical object (Angermuller et al., 2014b). Even though the term 'discourse analysis' is often claimed by both groups, one might suggest reserving 'discourse theory' for the former and 'discourse analysis' for the latter. If Discourse Studies has precisely emerged from the productive encounter of both theory and analysis, this distinction may help to explain some of the obstacles that have prevented the exchange between certain researchers claiming to deal with 'discourse.'

Discourse Theories

Within discourse theory, one can distinguish between at least three strands: post-structuralist, normative–deliberative and critical–realist discourse theories.

- Post-structuralist discourse theory usually insists on the constitutive role of language and communication for what counts as real in a given society. Thus, if the 'speaking subject' is seen as an effect of discursive practices, discourse not only describes but to a certain degree also constitutes the social. Rather than as a closed container, the social is perceived as an open, dynamic and heterogeneous terrain of relationships. Following Saussure's primacy of difference over identity or Wittgenstein's practical turn in social theory, post-structuralist discourse theory adopts a resolutely antiessentialist stance which emphasizes the contingent character of reality and nature constructed through practices involving both power and knowledge. While a first wave of post-structuralist discourse theory, among which Michel Foucault (1972) and Michel Pêcheux (1982), helped to constitute Discourse Studies in France, a second wave, also known as (*French*) *Theory*, resulted from the reception of 'Continental' thinkers (i.e., mainly from France such as Michel Foucault, Jacques Derrida, Jacques Lacan, Louis Althusser, Gilles Deleuze, Félix Guattari and to a lesser degree from Germany such as Friedrich Nietzsche, Walter Benjamin, Martin Heidegger, Theodor W. Adorno) in the Anglo-American world, where it has contributed to the formation of cultural studies (Hall, 2003), postcolonialism (Spivak, 1988; Said, 1978), the theory of hegemony (Laclau and Mouffe, 1985), queer studies (Butler, 1990), and psychoanalytical theory (Žižek, 1989).
- For normative–deliberative theories, discourse designates the negotiation of an agreement between participants who enter discourse in order to overcome a conflict while pursuing certain interests and objectives. Discourse in this sense usually involves free deliberation in the public sphere over what can be considered good and acceptable for all. According to this perception, the discourse participants need to come to terms with clashing norms in discourse. For the 'strong' program of normative–deliberative discourse theory, spelled out by third-generation critical theorists such as Jürgen Habermas (1985), certain communicative laws and normative standards are inherent in discourse which make certain claims more true, rational, and acceptable, whatever the social or political resources of those who make them. Weaker versions are typically critical of the universalism of the strong program and reject the idea of a normative point zero in discourse. They see hierarchies of values as discursive constructions of those who are concerned with social and political problems or as a contingent result of the communicative operations in an autopoietic system. Thus, Lyotard insists on certain unresolvable tensions and contradictions (*différend*) in discourse that can make rational agreement impossible (1983).
- While post-structuralist and normative–deliberative discourse theorists are firmly situated within the constructivist social sciences and humanities, some discourse theorists, who have defended a moderate realism, consider discourse to be a social practice embedded in given power structures. In this way, critical realist discourse theorists point to the objective social constraints in which discursive practices are embedded. Through discourse, unequal relationships are produced and reproduced between larger social groups, classes, or communities. Critical realist discourse theories usually have an affinity with macro-sociological theories of social and cultural inequality, ranging from the Marxist critique of ideology to Pierre Bourdieu's work on symbolic domination. In drawing from critical realist discourse theory, linguists have become interested in social problems in critical discourse analysis (CDA) (Fairclough, 1992), while Marxist sociologists (Sum and Jessop, 2013) have discovered discourse to be a constitutive dimension in the dynamics of social power.

While the common denominator of these ideal typical strands of discourse theory is the central role of language, communication, or meaning in social, political, and cultural theory, the line is not always easy to draw (Angermuller et al., 2014a). Thus, weaker versions of normative–deliberative discourse theories have often been associated with post-structuralism while Foucault's power/knowledge approach has been claimed by both post-structuralists and critical realists. The critical potential has been spelled out in all of these approaches, albeit in different ways: if the post-structuralist gesture is more epistemological, aiming at a subversive decentring of identities and 'naturalized' power, normative–deliberative discourse theories focus on the critical tensions between how discourses *should* operate and how they actually

do operate. Critical realism, finally, reveals the power structures hidden as it were beneath the representations produced in discourse.

In a word, discourse theories often articulate three problems: power, knowledge, and subjectivity, these are, visualized below in the discourse theory triangle. Accordingly, discourse is shaped by power structures but by representing power structures it can also contribute to objectifying and constituting them. As a socially-situated activity of producing meaning, discourse produces, establishes, and legitimizes knowledge in social groups and communities. And discourse is crucial for the construction of subjectivity as it defines, identities, and creates actorhood by attributing places and positions to those who enter discourse.

Discourse Analyses

While discourse *theorists* ask how linguistic models such as Saussure's differentialist theory of linguistic value or Wittgenstein's conception of 'language game' can inform social theory, discourse *analysts* deal with discourse as an empirical material object. If discourse theory has its base in social philosophy, macrosociology, or the constructivist political sciences, discourse analysis tends to have a background in linguistics or microsociological social research. To be a discourse researcher, one always needs to integrate both. However, some discourse researchers tend to be more at home on one side than the other.

From an ideal-typical perspective, analyzing a discourse usually means offering empirical insights into a socially and historically defined object (such as the 'War on Terror' discourse or 'classroom discourse in French *lycées*') on the basis of selected material (such as all the New York Times articles from 2000 to 2010 or the observation of classroom interactions). In discourse analysis, methodological tools are applied to selected material in order to generate empirically grounded hypotheses about the discourse to hand.

Discourse analyses do not have to follow classic research designs privileging 'hard' statistical methods, and they almost never believe in causalist epistemologies ("what are your dependent and independent variables?"). Discourse analysts are acutely aware of the fact that discourse is not a given object but rather partially constructed in the research process by methods and procedures which seek to account for the object. Discourse analysis in this sense mobilizes a broad range of methods, tools, and techniques which account for the uses of oral or written texts in specific contexts. Discourse analysis, therefore, does not designate a method but a cross-disciplinary field of methods and methodologies which investigate meaning making as a socially framed and situated practice.

Discourse analysis can work with quantitative or qualitative approaches, or both. Quantitative approaches usually rely on corpus analytical ('lexicometric' in France) methods, i.e., they deal with systematically selected and annotated textual material which is processed with the help of computer software. As a rule, quantitative tools are used for exploratory purposes to generate hypotheses about the macro-organization of a discourse which are then examined in more detail in qualitative microanalyses. If quantitative approaches usually deal with larger selections of written texts, qualitative approaches deal with smaller selections of material, written or oral, which can be handled 'manually' by the researcher.

Given that discourse analysis sees meaning as resulting from the interplay of texts and contexts, it must be strictly distinguished from content analysis, which views meaning as inherent in 'texts.' Discourse analysis, especially in its French and interactional varieties, must also be distinguished from hermeneutics in a narrow sense, understood as a subjective art of 'understanding.' While it places much emphasis on rationally describing the social production of meaning, discourse analysis shares the hermeneutic idea that in a wider sense, social phenomena are always meaningful. Thus, discourse analysis has a large number of tools, techniques, and procedures at its disposal which reveal the underlying rules and mechanisms of the discursive practices, resources and forms which are mobilized to realize a discourse.

A key problem of the modern social sciences and humanities discourse has been the object of many specialized approaches in the disciplines. Thus, linguists have catalogued the semiotic resources involved in discursive practices, sociologists have investigated the practical logics in discourse, anthropologists have paid special attention to settings, and so on. What makes discourse such a methodological challenge is that it is not a pure object. Discourse cannot be equated with texts, nor can it be reduced to a given context. Rather, the question is how to account for the (practical) uses that are made of language by certain people, at a given time, in a particular place. Therefore, the heterogeneous constitution of a discourse needs to be emphasized, which consists of different but interwoven 'components': 'language,' 'practice,' and 'context.' 'Language' comprises any repertoire of semiotic resources that can be mobilized in a discursive practice, that is, linguistic forms and expressions, words, utterances and texts, but also nonverbal signs and gestures, as well as visual, auditory, and olfactory signs; 'practice' denotes the practices people engage in to define their place vis-à-vis others and to coordinate their behavior. These practices can testify to their practical know-how and mobilize their resources (time, money, power, etc.); 'context' covers the socially defined situation or material setting in which discursive practice takes place. It can refer to the knowledge about who participates in discourse, when and where with what purpose and so on.

While these components always work together in discourse, discourse analysts sometimes prefer to take their point of departure from one angle rather than another. In this sense, 'language-oriented' discourse analysts, who are often but not always linguists, register linguistic or nonlinguistic material (for instance when it is shown in linguistic pragmatics how deictic expressions refer to their contexts) in order to account for how it is used in a given context. 'Practice-oriented' (or praxeological) discourse analysts prefer to observe what happens between actors in situations (for instance turn-taking processes in institutional talk). 'Context-oriented' discourse analysts focus on how material contexts (for example in an ethnographic study of intercultural communication in a setting) or on the question of how their contextual knowledge is constructed and established in discourse (for instance in a corpus analysis of the semantics of French nationhood). If each orientation starts from one angle which it takes as an empirically observable entry into discourse, discourse researchers always need to account for the other two components at least theoretically,

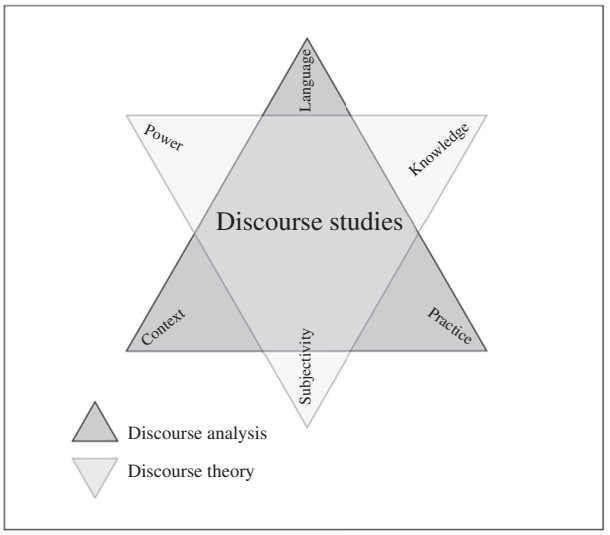

Figure 1 Discourse Studies as discourse analysis and discourse theory.

which is why they cannot always be easily distinguished. Without recognition of the articulation of all three strands, a language-oriented approach could become semiotic, a practice-oriented approach would be nothing more than conversation analysis, and a context-oriented approach might result in a mere interpretive coding of the themes and topics of a discourse. Therefore, unlike classical linguists who deal with language as an object per se, discourse researchers account for the practices employed by/in language in certain contexts. And unlike pure sociologists, they explain which practices are enacted and which knowledges are mobilized by linguistic forms. Just as discourse entails the interplay of language, practice, and context, discourse analysis involves a true cross-disciplinary perspective on the object, which is illustrated in **Figure 1**.

By articulating discourse theory and discourse analysis, Discourse Studies has crucially contributed to research on the social production of meaning in the social sciences. Yet it has not developed in the same linear way in all parts of the world. The currents and approaches which have received most international attention are situated for the most part in the Anglo-American world, France and Germany. According to a 'Continental' understanding, discourse designates written texts circulating in large communities whereas discourse as symbolic interaction taking place face-to-face is often considered to be more typical of the Anglo-American world. Linguists have played a leading role everywhere, especially in the more technical versions of discourse analysis; while discourse theorists in Europe are more likely to come from sociology, the political sciences, and philosophy, in the US they are mostly based in the cultural and literary field.

Discourse Studies as a Global Field

While Discourse Studies has recently seen dynamic growth outside the Western sphere, as a field, Discourse Studies was first established in Europe, especially in France and the UK, more so than in the US where 'discourse' remains a term used in a variety of specialized debates without designating a field with a common theoretical problem. Discourse Studies is probably most established in France, where it is, by and large, a subfield of linguistics, as well as in the UK and the former Commonwealth, where a number of high-profile schools are based (such as the Essex School of Hegemony and discourse analysis or the social semiotics inspired by Halliday, as well as many representatives of CDA). A large variety of journals as well as the regular recruitment of discourse researchers testify to the increasing institutional recognition that Discourse Studies enjoys in both countries.

From the beginning, Discourse Studies emerged in various countries and disciplines as the result of encounters between different traditions and schools. Today, there are no 'national' schools any more. Rather, in all countries there are discourse researchers who can be said to carry certain 'national' labels such as 'French' post-structuralism, 'American' interactionism, 'British' pragmatics, or 'German' hermeneutics. The following will give a brief overview of the major tendencies in the respective spheres.

Discourse Studies first started to develop in France, where it is known as *analyse du discours*, a name which goes back to an article by Zellig Harris (1952). As a field, it received seminal impulses from two theorists who started out as philosophers at the end of the 1960s, when the structuralist conjuncture in the French social sciences and humanities was drawing to a close: Michel Foucault (1972) and Michel Pêcheux (1982), the latter becoming the head of what is sometimes called the 'French School of Discourse Analysis' in the 1970s. In what followed, the field developed from the conjunction of the intellectual and political essence of discourse theory with two highly technical strands in discourse analysis, namely lexicometrics, which began in the early 1960s and was to give birth to a number of computer programs (such as Lexico3, Alceste, Hyperbase and so on), and text linguistics (Adam, 1999) plus the many developments in enunciative pragmatics that developed from the late 1970s onwards (in the wake of Benveniste, 1974; see Maingueneau, 1991), and which marked the end of structuralist hegemony by turning to the utterances as the smallest communicative units of discourse (Angermuller, 2014). From the beginning, French discourse linguists were joined by a number of historians (Guilhaumou et al., 1994), specialists in media and communication sciences (Charaudeau, 1983) and sociologists, who shared the emphasis that many French linguists, versed in the formalist and philological traditions of French letters, usually place on the 'opaque materiality' of texts. Today, with its focus on written texts, Discourse Studies in France prolongs the heritage of 'antihumanist' discourse theorists which linguists have put into discourse analytical practice.

Even though as a field Discourse Studies in the US is largely limited to small areas within sociolinguistics (Labov and Fanshel, 1977; Johnstone, 2008; Gumperz, 1982), corpus analysis (Biber et al., 1998), applied linguistics (Kramsch, 1998), and linguistic anthropology (following pioneers like Malinowski, 1923; Hymes, 1972; Duranti, 1997), discourse is well-known as a concept and a problem. Theoretically, 'discourse' is informed by the rich tradition of symbolic

interactionist and praxeological research which prolong the philosophical tradition of North American pragmatism. Symbolic interactionism, even though its name was only created after the war, draws from the work of George Herbert Mead (1967). It started with ethnographic investigations of the Chicago School and came on the scene during the 1960s and 1970s (Sacks, 1992; Cicourel, 1973; Goffman, 1981). While these theorists are better known as pioneers of the field of qualitative social research, 'discourse' in this context is often understood as a symbolically mediated turn-taking practice in an institutional setting, and it has loosely informed a number of developments in sociology, the political sciences and anthropology. With its strict empirical orientation recognizing the creative potential of human actors, interactionist discourse ideas are typical for microsociological strands in the social sciences whereas post-structuralist discourse theories, which started to arrive from France in the 1970s, have, by and large, been restricted to debates within the humanities.

Discourse Studies in the UK represents almost the entire spectrum of theoretical and analytical strands in Discourse Studies. Post-structuralist discourse theories (Laclau and Mouffe, 1985; Hall, 1980; Rose, 1996) have developed in the political sciences and sociology whereas linguistic discourse analysis has seen contributions from sociolinguistics (Sinclair and Coulthard, 1975; Coulthard, 1977; Stubbs, 1983) as well as corpus analysis (Sinclair, 2004; Baker, 2005). Discourse Studies in the UK testifies to the intellectual presence of pragmatics in philosophy and linguistics. In philosophy, pragmatics goes back to the later Wittgenstein (2003) who, with his concept of language games points to the practical dimension of language use. Through Austin's speech-act theory (Austin, 1962; Widdowson, 2007) and Grice's theory of conversational maxims (Grice, 1989), pragmatic ideas have turned the linguistic debate toward discourse (Brown and Yule, 1998; Hoey, 2001; Hyland, 2005). With its insistence on social practices, Halliday's social semiotics (Halliday, 1978; Kress and van Leeuwen, 1996) has influenced CDA (Fairclough, 1992). CDA testifies to the interest in power and inequality, which has seen similar developments elsewhere (Jäger, 2007; Reisigl and Wodak, 2009; van Dijk, 2009).

Like the UK, the German-speaking debate on discourse reflects this broad range of approaches (Angermüller, 2011; Angermuller et al., 2014b; Wrana et al., 2014a). Although German-language Discourse Studies can be viewed as a relative latecomer vis-à-vis the French and Anglo-American traditions, it is now well established across the entire social sciences and humanities. If the traditional 'German' discussion on discourse privileges holistic models which emphasize the richness of meaning-making activities in human culture, discourse is perceived as a space into which actors enter to reach an agreement (Habermas, 1985), as a stock of knowledge shared by all the members of a society (Keller, 2005) or in terms of collective symbols which represent organic social communities (Link, 1982). Post-structuralist approaches by contrast have been particularly interested in the discursive construction of subjectivity in neoliberalism. They have been characterized by their special emphasis on the articulation of the theoretical and methodological dimensions of discourse research (Herschinger, 2011; Macgilchrist, 2011; Nonhoff, 2006; Wrana, 2006). Linguistic discourse researchers have prolonged pioneering research in text linguistics (Warnke, 2007) as well as in semantics (Busse, 1987; Ziem, 2008).

It is these four countries which have coined specific recognizable 'styles' which today exist as more and less hybrid representatives in all parts of the world: the focus on decentered subjectivity in 'French' post-structuralism, the observation of turn-taking practices in symbolic interactionism from North America, the analytical focus on language in use in 'Anglo-Saxon' pragmatics, and the interpretative approaches to shared meaning and heterogeneous knowledge in the German-speaking world.

Currently, Discourse Studies testifies to a great deal of heterogeneity between the disciplinary and national fields in which it is practiced. At times, discourse is merely a fad where old questions are clad in new discourse-theoretical terms. However, discourse has also been the object of a long and systematic investigation in specialized subfields in the disciplines and has now become the object of the transdisciplinary field of Discourse Studies.

See also: Corpus Linguistics; Discourse and Indentity; Discourse, Anthropology of; Interactionism, Symbolic; Structuralism.

Bibliography

Adam, Jean-Michel, 1999. Linguistique textuelle: des genres de discours aux textes. Nathan, Paris.

Angermuller, Johannes, 2014. Poststructuralist Discourse Analysis. Subjectivity in Enunciative Pragmatics. Palgrave Macmillan, Houndmills, Basingstoke.

Angermüller, Johannes, 2011. Heterogeneous knowledge. Trends in German discourse analysis against an international background. Journal of Multicultural Discourses 6 (2), 121–136.

Angermuller, Johannes, Maingueneau, Dominique, Wodak, Ruth (Eds.), 2014a. The Discourse Studies Reader. Main Currents in Theory and Analysis. John Benjamins, Amsterdam, Philadelphia.

Angermuller, Johannes, Nonhoff, Martin, Herschinger, Eva, et al. (Eds.), 2014b. Diskursforschung. Ein interdisziplinäres Handbuch. Bielefeld: transcript.

Austin, John L., 1962[1955]. How to Do Things with Words. The William James Lectures Delivered at Harvard University in 1955. Oxford University Press, Oxford, New York.

Baker, Paul, 2005. Using Corpora in Discourse Analysis. Continuum, New York.

Benveniste, Émile, 1974. Problèmes de linguistique générale, vol. 2. Gallimard, Paris.

Biber, Douglas, Conrad, Susan, Reppen, Randi, 1998. Corpus Linguistics. Investigating Language Structure and Use. Cambridge University Press, Cambridge, New York.

Brown, Gillian, Yule, George, 1998[1983]. Discourse Analysis. Cambridge University Press, Cambridge.

Busse, Dietrich, 1987. Historische Semantik. Analyse eines Programms. Klett-Cotta, Stuttgart.

Butler, Judith, 1990. Gender Trouble. Feminism and the Subversion of Identity. Routledge, London, New York.

Charaudeau, Patrick, 1983. Langage et discours. Eléments de sémiolinguistique. Hachette, Paris.

Cicourel, Aaron V., 1973. Cognitive Sociology: Language and Meaning in Social Interaction. Penguin, Harmondsworth.

Coulthard, Malcolm, 1977. An Introduction to Discourse Analysis, new ed. Longman, London, New York.

van Dijk, Teun A., 2009. Society and Discourse. How Social Contexts Influence Text and Talk. Cambridge University Press, Cambridge.

Duranti, Alessandro, 1997. Linguistic Anthropology. Cambridge University Press, Cambridge.

Fairclough, Norman, 1992. Discourse and Social Change. Polity Press, Cambridge, Oxford.

Foucault, Michel, 1972[1969]. The Archaeology of Knowledge & the Discourse on Language. Pantheon, New York.

Goffman, Erving, 1981. Forms of Talk. University of Pennsylvania Press, Philadelphia.
Grice, H. Paul, 1989. Studies in the Way of Words. Harvard University Press, Cambridge.
Guilhaumou, Jacques, Maldidier, Denise, Robin, Régine, 1994. Discours et archive. Expérimentations en analyse du discours. Mardaga, Liège.
Gumperz, John, 1982. Discourse Strategies. Cambridge University Press, Cambridge.
Habermas, Jürgen, 1985[1981]. The Theory of Communicative Action. Beacon Press, Boston.
Hall, Stuart, 1980. Encoding/decoding. In: Centre for Contemporary Cultural Studies (Ed.), Culture, Media, Language: Working Papers in Cultural Studies, 1972–79. Hutchinson, London, pp. 128–138.
Hall, Stuart, 2003. Encoding, decoding. In: During, Simon (Ed.), The Cultural Studies Reader. Routledge, London, pp. 90–103.
Halliday, M.A.K., 1978. Language as Social Semiotic. Edward Arnold, London.
Harris, Zellig S., 1952. Discourse analysis. Language 28, 1–30.
Herschinger, Eva, 2011. Constructing Global Enemies. Hegemony and Identity in International Discourses on Terrorism and Drug Prohibition. Routledge, New York.
Hoey, Michael, 2001. Textual Interaction. An Introduction to Written Discourse Analysis. Routledge, London.
Hyland, Ken, 2005. Metadiscourse. Exploring Interaction in Writing. Continuum, London, New York.
Hymes, Dell, 1972. Models of the interaction of language and social life. In: Gumperz, John J., Hymes, Dell (Eds.), Directions in Sociolinguistics. The Ethnography of Communication. Holt, Rinehart & Wilson, New York, pp. 35–71.
Jäger, Siegfried, 2007[1993]. Kritische Diskursanalyse. Eine Einführung. Unrast, Münster.
Johnstone, Barbara, 2008. Discourse Analysis. Blackwell Pub, Malden, MA.
Keller, Reiner, 2005. Wissenssoziologische Diskursanalyse. Grundlegung eines Forschungsprogramms. VS, Wiesbaden.
Kramsch, Claire, 1998. Language and Culture. Oxford University Press, Oxford.
Kress, Gunther, Leeuwen, Theo van, 1996. Reading Images. The Grammar of Visual Design. Routledge, London.
Labov, William, Fanshel, David, 1977. Therapeutic Discourse. Psychotherapy as Conversation. Academic Press, New York, London.
Laclau, Ernesto, Mouffe, Chantal, 1985. Hegemony and Socialist Strategy. Towards a Radical Democratic Politics. Verso, London, New York.
Link, Jürgen, 1982. Kollektivsymbole und Mediendiskurse. KultuRRevolution 1, 6–21.
Macgilchrist, Felicitas, 2011. Journalism and the Political: Discursive Tensions in News Coverage of Russia. John Benjamins, Amsterdam.
Maingueneau, Dominique, 1991. L'Analyse du discours. Introduction aux lectures de l'archive. Hachette, Paris.
Malinowski, Bronislaw, 1923. The problem of meaning in primitive languages. In: Ogden, C.K., Richards, I.A. (Eds.), The Meaning of Meaning. Harcourt, Brace and World, New York, pp. 296–336.
Mead, George Herbert, 1967[1934]. Mind, Self, & Society from the Standpoint of a Social Behaviorist. University of Chicago Press, Chicago, London.
Nonhoff, Martin, 2006. Politischer Diskurs und Hegemonie. Das Projekt 'Soziale Marktwirtschaft'. Transcript, Bielefeld.
Pêcheux, Michel, 1982[1975]. Language, Semantics and Ideology. Stating the Obvious. Macmillan, London.
Reisigl, Martin, Wodak, Ruth, 2009. The discourse-historical approach (DHA). In: Wodak, Ruth, Meyer, Michael (Eds.), Methods of Critical Discourse Analysis. Sage, London, pp. 87–121.
Rose, Nikolas, 1996. Inventing Our Selves. Psychology, Power, and Personhood. Cambridge University Press, Cambridge.
Sacks, Harvey, 1992. Lectures on Conversation, vol. I. Blackwell, Oxford.
Said, Edward W., 1978. Orientalism. Penguin, London.
Sinclair, John, 2004. Trust the Text. Language, Corpus and Discourse. Routledge, London.
Sinclair, John McHardy, Coulthard, Malcolm (Eds.), 1975. The English Used by Teachers and Pupils. Oxford University Press, Oxford.
Spivak, Gayatri Chakravorty, 1988. Can the subaltern speak? In: Nelson, Cary, Grossberg, Lawrence (Eds.), Marxism and the Interpretation of Culture. University of Illinois Press, Urbana, pp. 271–313.
Stubbs, Michael (Ed.), 1983. Discourse Analysis: The Sociolinguistic Analysis of Natural Language. Chicago University Press, Chicago.
Sum, Ngai-Ling, Jessop, Bob, 2013. Towards a Cultural Political Economy. Putting Culture in its Place in Political Economy. Edward Elgar, Cheltenham, Northampton.
Warnke, Ingo (Ed.), 2007. Diskurslinguistik nach Foucault: Theorie und Gegenstände. Walter de Gruyter, Berlin.
Widdowson, H.G., 2007. Discourse Analysis. Oxford University Press, Oxford.
Wittgenstein, Ludwig, 2003[1953]. Philosophische Untersuchungen. Suhrkamp, Frankfurt am Main.
Wrana, Daniel, 2006. Das Subjekt schreiben. Subjektkonstitution und reflexive Praktiken in der Weiterbildung – eine Diskursanalyse. Schneider Hohengehren, Baltmannsweiler.
Wrana, Daniel, Ziem, Alexander, Reisigl, Martin, Nonhoff, Martin, Angermuller, Johannes, 2014. DiskursNetz. Wörterbuch der interdisziplinären Diskursforschung. Berlin: Suhrkamp.
Ziem, Alexander, 2008. Frames und sprachliches Wissen. Kognitive Aspekte der semantischen Kompetenz. de Gruyter, Berlin, New York.
Žižek, Slavoj, 1989. The Sublime Object of Ideology. Verso, London, New York.

Discovery: Legal Aspects

Steven Baicker-McKee, Duquesne University School of Law, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S. Bundy, volume 6, pp. 3739–3743, © 2001, Elsevier Ltd.

Abstract

Discovery – the legal process for exchanging information relevant to disputed factual and legal issues in litigation – varies considerably from jurisdiction to jurisdiction. This article compares discovery under the two primary systems of justice in use in the world today: common law jurisdictions and civil law jurisdictions. The article uses the United States and England as illustrative common law jurisdictions, outlining the scope of discovery, automatic disclosures, affirmative discovery, electronic discovery, and sanctions under each system. It then contrasts those approaches to the much more limited concept of discovery attendant to civil law jurisdictions, in which the judge (rather than the parties) controls the gathering of information.

Introduction

Discovery is the legal process for exchanging information relevant to disputed factual and legal issues in litigation. While discovery can take place in both civil and criminal cases, this article will focus on the civil procedures. This article briefly outlines the purposes of pretrial discovery. It then describes the basic regimes of pretrial civil discovery in the United States and England – the two leading common law jurisdictions – and contrasts their approaches with analogous civil law institutions.

The Aims of Discovery

In common law systems, the issues in dispute are framed by the pleadings – typically a complaint presenting the plaintiff's claims and an answer presenting the defendant's defenses. There is a process for testing the legal sufficiency of the pleadings, often in the form of a motion to dismiss. There is also a process before trial for adjudicating claims as to which the facts are not disputed, often in the form of a motion for summary judgment. The idea behind such motions is that the purpose of a trial is for the fact finder – jury or judge – to hear the witness testimony and make credibility determinations. If there are no disputed material facts, there is no need for a trial and the court can adjudicate the dispute by applying the law to those undisputed facts.

Discovery is the process that bridges the pleading stage and the summary judgment or trial stage of litigation (see Herbison, 2000: pp. 1524–1525). For claims and defenses that survive early challenges, discovery is the process for exchanging relevant – or potentially relevant – information. This exchange of information during discovery has a number of purposes and potential benefits.

At the most fundamental level, discovery serves to make available the facts and information relevant to the dispute. Bringing the witnesses and documents out into the open allows the parties and the court to identify those claims – or portions of claims – as to which there is no genuine dispute of material fact, allowing for disposition without trial. It allows the parties to better assess the merits of their positions, facilitating settlement (Subrin, 2002: pp. 311–312). And barring resolution through motion or settlement, it leads to a more fair and efficient trial by focusing trial on the actually disputed issues and avoiding 'trial by ambush,' where one party springs surprise witnesses or documents on an opponent at the last minute (see Grimm and Yellin, 2013: p. 509).

In contrast, in civil law jurisdictions, discovery is not a separate step bridging the pleadings and the trial, but instead is an integral component of a judicially orchestrated search for the truth (Hazard, Jr., 1998a: p. 1019). Documents are gathered and witnesses are questioned, but all at the pleasure of the judge.

Of course, these discovery processes do not come without cost, and the escalating cost of discovery is hotly debated. Any consideration of the purposes and benefits of discovery must be balanced by recognition of the substantial monetary and time expenditures that discovery entails (Moss, 2009: p. 936). The more discovery a system allows, the greater the cost and delay of the discovery process. Views differ on the balancing of these considerations, with some seeing broad discovery as opening justice to individual plaintiffs injured by large corporations who will not voluntarily disgorge their damaging documents, and others seeing the high costs of discovery leading defendants to make substantial payments to settle even frivolous claims just to avoid even greater discovery costs (see Subrin, 2002: pp. 311–312; *Bell Atlantic*, 550 U.S. at 559).

Discovery in Common Law Jurisdictions

Discovery in common law jurisdictions is generally much more extensive than discovery in civil law jurisdictions. The United States and England are two of the lead common law jurisdictions, and there are differences in their approaches to discovery as well. This section will explain discovery in the United States and England, and the next section will consider discovery in civil law jurisdictions.

Discovery in the United States

In the United States, there are federal courts – essentially limited to claims involving federal laws and for larger claims

between citizens of different States – and State courts, with more general jurisdiction. Each State has its own rules of civil procedure, governing many procedural aspects of litigation including discovery. Some States have modeled their rules of civil procedure on the Federal Rules of Civil Procedure, but others have not. This article discusses the Federal Rules of Civil Procedure.

Fundamentally, discovery in the United States forms an integral part of the litigation process, and is more extensive than in any of the other jurisdictions considered here (Walpin, 1997: p. 560). In the United States, the pleadings are designed to put the parties on notice of the nature of the claims and defenses, rather than to provide all of the facts and evidence supporting those claims and defenses. One advantage of notice pleading is that it eliminates some of the technicalities of code or fact pleading systems, which can result in the early dismissal of claims that make small technical errors in pleading. This structure is also designed to promote access to the courts in such cases as civil rights, products liability, securities, and antitrust, in which critical evidence frequently resides in the files and minds of the defendants. In these lawsuits, it would often be very difficult, if not impossible, for the plaintiff to plead facts or evidence with particularity in the complaint, as is required in fact or code pleading jurisdictions. The consequence, though, is that it places greater emphasis on discovery to bring the facts and the parties' positions to light (Subrin, 2002: pp. 311–312). Thus, discovery in the United States is quite broad.

Scope and Limitations

Discovery in the United States is available for "nonprivileged matter that is relevant to any party's claim or defense" (Baicker-McKee: Part III, 26(b)(1)). Thus, materials that are privileged attorney–client communications or that were prepared in anticipation of litigation are shielded from discovery. The requirement that discovery be 'relevant to any party's claim or defense' normally prevents a party from propounding discovery designed to develop new claims or defenses – discovery must be targeted at claims and defenses already in the pleadings, and may not be a 'fishing expedition.' This scope may be expanded by motion, subject to the court's discretion.

In order to be discoverable, a document or piece of information needs to either be admissible at trial or reasonably likely to lead to the discovery of admissible evidence. The discovery obligations extend to documents in a party's 'possession, custody, or control,' with control generally referring to documents held by another which the party has the right or ability to obtain upon request.

The scope of this discovery is quite broad. However, the rules include some protections for overly burdensome, harassing, or otherwise onerous discovery. These procedures add the concept of 'proportionality' to discovery. Essentially, if the cost or burden of discovery outweighs its likely benefits, the court may excuse a party from responding.

The United States litigation system is an adversarial system, such that it principally relies on the parties to request the information they want in discovery. The judge's role during the pretrial process is not to seek out the truth, but rather to referee the parties' conduct during the discovery process. The parties propound discovery, respond, and interpose objections without court involvement unless one party files a motion seeking to compel compliance by an opposing party, seeking protection from discovery by an opposing party, or seeking sanctions.

Although the United States system relies primarily on affirmative discovery propounded by the parties, in 1993 the United States introduced the concept of automatic disclosures for certain categories of information – junctures in the litigation process where the parties provide information to their opponents without the need for a request. This system of automatic disclosures recognizes that there are certain categories of information that parties will want in virtually every case, such that automatic disclosure of that information is more cost-effective and speeds the process along. Other information is obtainable only upon request.

Disclosures

There are three rounds of automatic disclosures in the United States. Near the beginning of a case, the parties must disclose the witnesses and exhibits that they may use to support their claims or defenses, calculations for any damages they seek, and the existence of any insurance potentially covering the claims in the case. Note that this disclosure is only for information that the disclosing party may use – information helpful to an opponent only need be produced if requested. The second round of disclosures pertains to experts, and requires identification of testifying experts and disclosure of information about their anticipated expert opinions and the basis for those opinions. The final disclosure is the pretrial disclosure, which requires disclosure of the documents, witnesses, and deposition testimony that the disclosing party intends to or may use at trial (Baicker-McKee: Part III, 26(a)).

Affirmative Discovery

In addition to these automatic disclosures, the Federal Rules provide a variety of mechanisms for affirmative discovery requests. This is the process in the United States where a party may seek evidence that an opponent has that are helpful to the party and adverse to the opponent. Remember that the disclosures apply to information helpful to the disclosing party – affirmative discovery is the process for obtaining documents and information not required to be disclosed.

Written discovery includes interrogatories (written questions), requests to inspect documents or property, and requests for admission (requests that a party admit facts, which are then deemed established and may not be controverted at trial). Parties may also take depositions (oral, recorded examination of witnesses that may then be used at trial for impeachment or, in certain circumstances, as substantive evidence). Finally, parties may, pursuant to court order or stipulation, arrange for a plaintiff to be examined by an expert for any condition at issue in the litigation (Baicker-McKee: Part III, 30–36). The court will set a period for this affirmative discovery – typically 4 to 6 months, but often longer.

Of these devices, depositions are used in the United States to a far greater extent than in other jurisdictions, and are often the most effective manner of obtaining critical information (Herbison, 2000: pp. 1524–1525). Depositions allow a party's attorney to ask questions directly of an opposing party, without the filtering that the other party's lawyer interjects in responses

to written discovery. Depositions also allow for more, and more immediate, follow-up, and allow counsel to assess the credibility of witnesses. Given the rarity of trials in modern practice in the United States – less than 2% of cases go to trial – depositions are often the most important part of the litigation process.

This affirmative discovery process results in the greatest cost component of civil litigation in the United States and can greatly extend the duration of many litigation matters (Walpin, 1997: p. 560). Indeed, pretrial discovery is estimated to amount to about 80% of the cost of any major litigation. In particular, as explained in the next section, the explosion of electronic discovery has dramatically increased the cost and duration of discovery.

Electronic Discovery

It is well settled in United States courts that electronically stored information (often referred to as ESI) is discoverable in the same way that paper documents are. Thus, nonprivileged information relevant to any party's claim or defense is discoverable if located in an e-mail, on Facebook or Twitter, in a word processing file, in a spreadsheet, in a database, on a smartphone, on a backup tape, in browser cache, etc. Because of the proliferation of electronic data and the more casual nature of communications like e-mail or Facebook posts, electronic discovery has become the focus of discovery in many cases in United States courts, and the attention to, and cost of, electronic discovery seems to be on an upward trend. The total cost of e-discovery in the United States had grown to almost $3 billion by 2009 (Yip, 2012).

In addition to cost, electronic discovery implicates two significant concerns. First, given the volume of electronic communications and the ease with which they are disseminated, electronic documents implicate privilege concerns. Of primary concern, privileged documents are often inadvertently produced during discovery. While the rules allow a party to recall, or 'claw back,' inadvertently produced documents, the disclosure may have waived the privilege. Second, electronically stored information is easily deleted, potentially leading to sanctions for failure to preserve evidence.

Discovery Sanctions

Failure to follow the letter or the spirit of the discovery rules can lead to sanctions. These sanctions routinely involve reimbursement of the resulting cost incurred by the opposing party, but may also include excluding certain evidence, deeming certain facts admitted or established, and even in extreme cases entering a dispositive ruling against the offending party (Baicker-McKee: Part III, 37). In general, however, before seeking sanctions, a party must 'meet and confer' with the opposing party, explaining the purported violations and affording the opposing party an opportunity to correct the alleged violation. Only if the opposing party refuses to correct the violation may a party file a motion for sanctions (Willoughby et al., 2010).

An issue of increasing focus for discovery sanctions in the United States is the failure to preserve documents. The increased volume and emphasis on ESI, and the ease with which e-mails and other ESI are deleted, has resulted in significant motion practice in this area. Courts sometimes sanction attorneys as well as parties if the attorneys do not send 'litigation hold letters' instructing all the likely records custodians to preserve relevant information, and then monitor compliance with these instructions (*Zubulake*).

Cost of Discovery

Under the American System, there is a distinction between the recovery of attorney's fees and the recovery of other types of expenses (often referred to as 'costs'). Unless a statute or contract provides otherwise, parties generally bear their own attorney's fees regardless of who succeeds in the action. As discussed above, when a discovery violation has occurred, however, some of the attorney's fees can be shifted.

Costs are handled differently. During the discovery process, the party receiving a discovery request generally bears the expense of gathering responsive information, and the party making the request must pay for any copies that it requests. The rules permit the parties to ask the court to alter this default arrangement, however. Thus, if a party believes that gathering requested information is too burdensome, the party may ask the court to require the requesting party to absorb those costs – and the court would have great discretion to respond to this request. At the end of a case, if it has not settled, then the prevailing party is entitled to recover some of its costs (but not attorney's fees) from the losing party.

Importantly, the courts have deemed most of the electronic discovery expenses outside the scope of recoverable costs (Mast, 2010: p. 1844). Thus, in most cases, a prevailing party will have its vindication mitigated by its attorney's fees and its electronic discovery costs.

Discovery in England

Like the United States, England is a common law jurisdiction. Most English civil cases originate in the County Courts, with appeals moving through the High Courts, the Courts of Appeal, and then potentially to the UK Supreme Court as the court of last resort. Prior to 1999, the County Courts and the appellate courts followed different procedures. With the advent of the Civil Procedure Rules (or 'White Book') in 1999, each court system became governed by the same set of procedures (Civil Procedure Rules, Part 1.02). These rules stress the importance of balancing equality for both parties against proportionality of costs – seeking to ensure that exorbitant costs do not act as a bar to obtaining justice (Mast, 2010: Part 1.1; Vorrasi, 2004: p. 362).

Discovery in England represents a middle ground between the broad discovery in the United States and the more limited discovery in civil law jurisdictions. It still leaves a role for the parties in crafting discovery requests to opposing parties, but entails significantly more 'hands on' involvement by the judge.

Scope and Limitations

Although the United States and England are both common law jurisdictions, English discovery is considerably narrower than U.S. discovery. The scope of discovery in England is largely dictated by the type of 'track' to which the judge assigns the case: small claims track, fast track, or multitrack (Gensler, 2010: p. 708; Civil Procedure Rules, 2014: Part 26.6). The small

claims track is generally used for any dispute where the claimed amount is not more than £10 000. The court normally allocates a few hours to the resolution of these types of disputes. These cases have very limited discovery, and in fact the larger discovery rule does not even apply to cases on the small claims track. The fast track is used for claims not exceeding £25 000, and only for those cases where the judge determines that the trial will only last for a single day and where oral expert evidence at trial will be limited to one expert per party. These cases are anticipated to be heard within 30 weeks of the judge's assignment to the fast track. The multitrack designation is used for any case that does not fit into either of the other categories.

English courts generally classify documents in one of four ways: (1) documents owned by the parties upon which they intend to rely; (2) documents that are materially adverse to a party's claim or defense; (3) documents that are relevant, which may provide color to the claim and a backstory but "do not obviously support or undermine either side's case"; and (4) "train of inquiry documents," or those documents that "may lead to a train of enquiry enabling a party to advance their own case or damage that of their opponent's" (Civil Procedure Rules, 2014: Part 31.6). While American style discovery would generally encompass all four categories of documents above, the extent to which English discovery applies depends on the track to which a case is assigned, as discussed below in Section Disclosures.

Like in the United States, parties in England need not produce privileged documents or communications. Additionally, "English law has not developed a work-product doctrine but has, instead, recognized a concept of litigation privilege that covers much of the same territory. Litigation privilege and attorney-client privilege, generally referred to in England as legal advice privilege, are so closely interwoven in English law that no discussion of legal professional privilege would be complete without some reference to both" (Pike, 2006: p. 61).

Like in the United States, the parties may ask the court to expand or contract the scope of discovery. However, unlike the United States, the English system focuses much more on defined disclosures and less on affirmative discovery requests by the parties. Likewise, as the following section demonstrates, discovery in England is more concerned with the first two categories described above – documents that a party intends to use to support its claims or defenses and documents directly adverse to those claims and defenses.

Disclosures

In contrast to the limited automatic disclosures in the United States, the English system relies heavily on disclosures for most of the information exchanged in discovery. These disclosures are considerably more extensive and complex than disclosure under the United States system.

Initial information in a case in England is exchanged by a required questionnaire. While the small claims track questionnaire is more limited, the fast track and multitrack questionnaire asks parties to: explain the issues in the case; affirm that they have made the required preclaim effort to resolve their dispute and exchanged documents accordingly; provide information about their willingness to settle in the future; discuss the possibility of applying for summary judgment in the near future; disclose any agreements reached regarding discovery of electronic or paper documents; and name the experts and witnesses they plan to use (FormFinder).

After the questionnaire, in fast track and multitrack cases, the parties sometimes exchange additional information by 'Standard Disclosure.' Standard Disclosure only encompasses the first two categories of documents listed above – it "requires a party to disclose only the documents on which he relies, and the documents which adversely affect his own case, adversely affect another party's case, or support another party's case" (Civil Procedure Rules, 2014: Part 31.6). To 'disclose' a document does not necessarily mean to produce the document to an opponent, but merely means to state "whether a document exists or has existed." The search for these documents needs only be 'reasonable,' and is limited to only those documents in the party's control – meaning in its physical possession or to which it has the right to physical possession or inspection of the document (Civil Procedure Rules, 2014: Part 31.2, 31.7, 31.8).

Because Civil Procedure Rule 31 does not apply to cases on the small claims track, Standard Disclosure is not required for these claims. Rather, the court, in preparation for the hearing, will give a 'Standard Direction.' Under a Standard Direction, each party is directed to "file and serve on every other party copies of all documents (including any expert's report) on which he intends to rely at the hearing." There is no requirement to produce a list of documents, nor is there any formal inspection in advance of the hearing. The use of experts is limited, and only permitted at the discretion of the court (Civil Procedure Rules, 2014: Part 27.2, 27.4, 27.5.1).

For fast track cases, the court may, but is not required to, order the parties to make the Standard Disclosure. If the court orders a more narrow disclosure, it will specify the scope of required disclosure, which could include no disclosure whatsoever or list specific classes of documents that the parties must disclose.

Although for several years Standard Disclosure was the normal starting point for multiclaim cases, a 2013 amendment to the rules put in place a process where the parties meet at the outset of a case and then submit a report to the court. This report includes a brief description of the documents that may exist, their relevancy, their location and how they are stored (if electronic), and the estimated cost to search and retrieve the information. The judge then has a variety of discovery plan options from which to choose, and will give direction at the case management conference (Civil Procedure Rules, 2014: Part 31.5).

Affirmative Discovery

Standard Disclosure under any of the tracks does not generally extend beyond information in the first two categories above. In order to seek information under categories 3 and 4 – documents that provide back story and train of inquiry documents – a party must request Specific Disclosure under Civil Procedure Rule 31.12. The party must demonstrate that the search is 'relevant and proportionate,' and in deciding whether to grant the order, "the court will need to satisfy itself as to the relevance of the documents sought, and that they are or have been in the party's control, or at least that there is a prima facie case that these requirements will be met."

In addition to requesting Specific Disclosure of documents, parties may also seek other forms of discovery – they may ask the court to permit the taking of witness statements, draft questions to joint experts, and seek to have the other party admit facts.

As discussed above, in the disclosure stage, the parties do not provide a copy of the documents, they merely advise the other parties of the existence of disclosable documents. The party to whom a document's existence was disclosed may submit a written request to inspect the document. The disclosing party must produce it unless: it no longer controls the document; it has a right (such as legal privilege) or duty to prevent inspection; or it asserts that the burden of producing it is disproportional to the benefit (Civil Procedure Rules, 2014: Part 31.3). If a party wishes to challenge an assertion of privilege or proportionality, the party seeking the document can apply to the court and ask the court to decide if the assertion should be upheld.

Although the Civil Procedure Rules permit depositions of witnesses, due to costs, depositions are generally regarded as a tool of last resort; the rules promote the use of live witness testimony at trial or sworn statements over depositions. In lieu of depositions, the parties exchange witness statements for witnesses upon whom the parties plan to rely. This process stands in marked contrast to the process in the United States, where depositions are routinely conducted in every large or complex case.

The role of experts, as well as the discovery conducted regarding those experts, differs significantly between the United States and England. In the United States, experts are almost exclusively hired by the parties and act as advocates for the parties who hire them. In England, the rules provide that the overriding duty of an expert is to the court, not to the party who hired the expert. An English court may "direct that evidence on [a specified issue] be given by a single joint expert," and the court may also choose the means by which the expert is selected. These rules are driven by the desire to limit costs, as well as perceived expert partiality under the old system (Civil Procedure Rules, 2014: Part 35.1). Where each party has retained its own experts, the court may require the experts to work together to publish a report identifying the issues where they agree and disagree.

Beyond these procedures, there is very little affirmative discovery in the English courts.

Electronic Discovery

Civil Procedure Rule Part 31.4 defines 'documents' as "anything in which information of any description is recorded," and it is well established that e-mails, text messages, photos, voice-mails, audio recordings, databases, meta-data, and all other associated forms of electronic communications are included in the scope of Rule 31. Parties are required to discuss a discovery plan for ESI prior to the first case management conference. Similar to the American rule, as soon as litigation is anticipated, an attorney must notify the client of the obligation to preserve documentation that would be subject to disclosure.

Discovery Sanctions

Sanctions in England for violations of discovery rules can range from being denied the ability to rely on certain documents or witness testimony, to the possibility of being held in contempt of court. A party who fails to disclose an expert's report at the time specified in the case management order may be denied the ability to use the report at trial or call the expert to testify.

Cost of Discovery

In weighing the proportionality of cost, the Civil Procedure Rules require an English judge to consider the amount of money involved in the dispute, compared to the relative importance of the case, along with the complexity of the issues. These three factors are considered alongside each party's relative financial position. This cost/benefit analysis is then factored against the need to ensure both the plaintiff and defendant are given equal footing, the desire to avoid wasted resources (both time and money), and the judiciary's opportunity cost of using those resources on the dispute in question rather than on some other dispute.

At the end of a case, the 'English Rule' requires the losing party to bear the costs and attorney's fees of the prevailing party, subject to the determination by a taxing master as to what costs are recoverable (Feuerstein, 1995). This approach is designed to discourage frivolous suits by making the party bringing the suit exposed to the risk of bearing substantial costs incurred by the defendant.

Discovery in Civil Law Jurisdictions

Fundamental differences exist between the common law and civil law tradition. Unlike the common law, the civil law is found exclusively in written laws, and is not created by judges in precedential opinions (Honsel et al.: Section 23.3). Therefore, each dispute is merely a resolution for those two particular parties, decided by a civil system judge charged with applying the code to the particular facts and arriving at the result dictated by the code. The lawyer's role is also different in civil law jurisdictions: "in principle, the advocates' function is to assist the judge in fulfillment of the judicial responsibility, rather than, as in the common law," to zealously advocate for one party (Hazard, Jr., 1998a: p. 1020). The differences in these systems translate into vastly disparate approaches to discovery.

At the outset of an action, the forms of pleading differ. American law allows for 'notice pleading,' under which the complaint merely puts the defendant on notice of the nature of the claim, and relies on discovery to uncover the relevant facts (Subrin, 2002: p. 311–312). In contrast, civil law jurisdictions employ a more rigid 'fact pleading' standard, under which the complaint must include a detailed recitation of the facts, and often must include a description of supporting evidence as well (Dodson, 2010: p. 443). The parties may also attach documents favorable to their case to their pleadings (Bolt and Wheatley., 2006: p. 8). As a result, the plaintiff will already have been forced to disclose some of the facts and supporting evidence at the earliest stage of the case (Dodson, 2010: p. 443). This difference in pleading style dictates a greater need for discovery in common law jurisdictions.

After the plaintiff has filed its claim, the further exchange of facts and evidence proceeds very differently in civil jurisdictions. The judge, and not the parties' attorneys, will initiate

the production of further facts and evidence (Hazard, Jr., 1998a: p. 1020). Indeed, the judge's role to orchestrate production of evidence is seen as fundamental to the civil law tradition. In Germany, for example, it is solely the judge who may compel document production; a party has no procedural right to request a document. If a party asks the judge to request documents from another party, the judge will only grant the request if "the documents are relevant, specifically designated, and [either] … the requesting party can show a substantive right to the information … [or] the opposing party has made reference to the sought-after documents …." (Bolt and Wheatley, 2006: pp. 8–9). The German rules also provide for the judge to propound written interrogatories to the parties (as opposed to one party serving interrogatories on another in the common law systems) (Honsel et al.: Section 23.45).

American style depositions are not conducted in civil law jurisdictions, and attorneys generally have very limited interactions with witnesses in advance of the trial (Kötz, 2003: p. 63). Although the parties may propose witnesses for certain factual issues, it is the judge who decides whether the witness will be called, dictates the order in which the witnesses will be called, asks the witnesses questions, and summarizes their statements (Kötz, 2003). While the parties' attorneys do have the chance to ask questions after the judge has finished, in practice, this witness questioning is normally not extensive, and is not comparable to the American style cross-examination (Kötz, 2003).

Civil trials also proceed very differently in civil law jurisdictions. Civil law jurisdictions do not permit jury trials in civil suits – the judge acts as trier of both law and fact (Dodson, 2008: p. 141). In fact, there is often no single consolidated trial of all the issues (Hazard, Jr., 1998b: p. 1673). Instead, the civil law judge may hear presentations of facts and law at several small conferences, on different topics at each conference, until the judge decides that the case is ready for adjudication. No pretrial discovery is conducted because there is no real trial at all; the judge simply decides the case on an issue-by-issue basis as the judge sifts through facts and evidence presented at several short hearings, until such time as the judge is ready to announce the final decision (Hazard, Jr., 1998a: p. 1020).

Consistent with the role of the civil judge as the seeker of the truth, civil judges will choose and appoint experts, normally by working in conjunction with the parties. The judge will pose questions to the expert directly during the course of the hearings (Kötz, 2003: p. 64).

Thus, discovery is much more narrow in civil law jurisdictions, and is conducted in a fundamentally different manner than in common law jurisdictions. It is principally directed by the judge, searching for the truth, rather than by the parties seeking to advance one particular perspective.

Interjurisdictional Discovery

When a party in a common law jurisdiction seeks to conduct discovery in a civil law jurisdiction, or vice versa, confusion – and sometimes legal difficulties – can ensue. As one law professor has noted, "the impression of American discovery in most foreign countries is that of an alien legal regime conducting a warrantless search in someone else's domestic territory" (Hazard, Jr., 1998b: p. 1676). Since the civil law judge functions as the primary collector and sifter of facts in the courtroom, an American style discovery request sent directly to a foreign party may be seen by the tribunal as either an 'attempt to circumvent the judiciary' or a 'preemptory demand' on the judge (Hazard, Jr., 1998a: p. 1020).

American discovery laws frequently run contrary to European privacy laws (Bolt and Wheatley, 2006: p. 13). Many civil law jurisdictions, and some common law jurisdictions, have therefore enacted blocking statutes to prevent American parties from seeking American style discovery within their borders (Bolt and Wheatley, 2006).

Conclusion

The fundamental differences between common law jurisdictions and civil law jurisdictions are manifest in their approaches to discovery. In the United States, the pleadings merely frame the issues, and the facts and evidence are developed through broad discovery, conducted by the parties in an adversarial manner. Not only are documents exchanged, but parties frequently seek information by written interrogatories and by oral examination during depositions. This approach promotes access to the courts, making it more feasible for a 'David' plaintiff to sue a 'Goliath' defendant. The price tag for this access to the courts is a more costly and prolonged litigation process, and the proliferation of electronically stored information appears to be driving up the price and delay of discovery.

In contrast, in civil law jurisdictions, the pleadings advance many of the facts and evidence at the outset of a matter, reducing the need for extensive discovery. The judge then dictates the discovery the judge deems necessary to adjudicate the dispute, which is typically limited to the production of documents. The process is considerably less expensive and time consuming than in the United States, but it may impose barriers for an individual who has good cause to believe a large corporation is acting improperly but who needs evidence from the corporation to prove the case.

See also: Civil Law; Common Law; Courts and Adjudication; European Union Law; Expert Testimony; Insurance and the Law; International Law and Treaties; Juries; Legal Systems, Classification of; Litigation; Scientific Evidence: Legal Aspects; Trials.

Bibliography

Baicker-McKee, Janssen, Corr, Federal. Civil Rules Handbook, 2014 ed.

Bell Atlantic Corporation v. Twombly, 550 U.S. 544, 127 S.Ct. (1955), (2007).

Bolt, Jan W., Wheatley, Joseph K., 2006. Private rules for international discovery in U.S. district court: the U.S.–German example. UCLA Journal of International Law and Foreign Affairs 11 (1).

Civil Procedure Rules, 2014.

Dodson, Scott, 2010. Comparative convergences in pleading standards. University of Pennsylvania Law Review 158 (441).

Dodson, Scott, 2008. The challenge of comparative civil procedure civil litigation in comparative context. West St. Paul, MN Alabama Law Review 60, 133.

Feuerstein, Lorraine Wright, 1995. Two-way fee shifting on summary judgment or dismissal: an equitable deterrent to unmeritorious lawsuits. Pepperdine Law Review 23, 125.

Gensler, S.S., 2010. Judicial case management: caught in the crossfire. Duke Law Journal 60, 708.

June 7, 2014. http://hmctsformfinder.justice.gov.uk/courtfinder/forms/n181-eng.pdf, Last visited (FormFinder).

The Honorable Grimm, Paul W., Yellin, David S., 2013. A pragmatic approach to discovery reform: how small changes can make a big difference in civil discovery. South Carolina Law Review 64, 495.

Hazard Jr., Geoffrey C., 1998a. Discovery and the role of the judge in civil law jurisdictions. Notre Dame Law Review 73, 1017.

Hazard Jr., Geoffrey C., 1998b. From whom no secrets are hid. Texas Law Review 76, 1665.

Herbison, James F., 2000. Corporate reps in deps: to exclude or not to exclude. Washington University Law Quarterly 78, 1521.

Honsel, Bernd, Winter, Wolfgang, Paul, Gerald G., Dase, Wolfgang A., 2014. Successful Partnering between Inside and Outside Counsel § 23.45.

Kötz, Hein, 2003. Civil justice systems in Europe and the United States. Duke Journal of Comparative & International Law 13, 61.

Mast, Andrew, 2010. Cost-shifting in e-discovery: reexamining Zubalake and 28 U.S.C. § 1920. Wayne Law Review 56, 1825.

Moss, Scott A., 2009. Litigation discovery cannot be optimal but could be better: the economics of improving discovery timing in a digital age. Duke Law Journal 58, 889.

Pike, Richard S., 2006. The English law of legal professional privilege: a guide for American attorneys. Loyola University Chicago International Law Review 4, 51.

Subrin, Stephen N., 2002. Discovery in global perspective: are we nuts? DePaul Law Review 52, 299.

Vorrasi, Kenneth M., 2004. England's reform to alleviate the problems of civil process: a comparison of judicial case management in England and the United States. Journal of Legislation 30, 361.

Walpin, Gerald, 1997. Civil justice systems in other democratic countries provide means to torpedo the litigation explosion here. New York Law School Law Review 41, 557.

Willoughby Jr., Dan H., et al., 2010. Sanctions for e-discovery violations: by the numbers. Duke Law Journal 60, 789.

Wright, Charles A., Miller, Arthur R., Kane, Mary K., Marcus, Richard L., Steinman, Adam M., 2001. Federal Practice and Procedure: Civil §, third ed.

Yip, John T., 2012. Addressing the costs and comity concerns of international e-discovery. Washington Law Review 87, 595.

Zubulake v. UBS Warburg LLC, 229 F.R.D. 422 (S.D.N.Y. 2004).

Discrete-State Models of Information Processing

William H Batchelder, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Discrete-state models for information processing tasks assume that overt responses to stimulus inputs are probabilistically determined by a small set of latent mental states. The earliest examples were in signal detection and mathematical learning theory. Later, the class of multinomial processing tree (MPT) models was developed. MPT models assume that stimulus inputs probabilistically select processing branches in a tree structure, where each branch is a possible sequence of latent cognitive acts leading to an overt response. MPT models have been developed for many information processing tasks, and their statistical theory is well developed at the class level.

Introduction

Discrete-state models for information processing (IP) tasks assume that overt responses to stimulus probes are determined by unobservable (latent) mental states. Viewed generally, IP tasks involve experimental paradigms where a subject must classify each of a set of N stimulus items, $\mathscr{S} = \{S_1, \ldots, S_N\}$, into a set of J response categories, $\mathscr{C} = \{C_1, \ldots, C_J\}$. A discrete-state model assumes that there is a small set of M latent (unobservable) cognitive states $\mathscr{L} = \{L_1, \ldots, L_M\}$, and corresponding to each of these states is a probability distribution over the response categories. Thus, the probability that a stimulus item S_i is classified into a particular response category C_j is computed by summing response probabilities over the latent states, namely,

$$\Pr(C_j|S_i) = \sum_m \Pr(C_j|L_m)\Pr(L_m|S_i) \qquad [1]$$

Despite the simplicity of Eqn [1], the special character of a discrete-state model depends on the structure of the latent states and the associated probabilistic processes.

Most IP paradigms are either detection tasks or learning tasks. In detection tasks, the subject knows the basis for a correct classification, and in learning tasks, they must acquire the basis for a correct classification from a series of learning trials. An example of a detection task is a Yes/No recognition memory paradigm with word stimuli, discussed in the next section. An example of a learning task is a paired-associate learning task discussed in the section on Markov learning models.

In the early history of IP modeling, most discrete-state models were either signal detection models or Markov learning models. The second and third main sections respectively, will describe briefly these two classes of discrete-state models. For the past 25 years, most discrete-state models fall into the class of MPT models. Section Multinomial Processing Tree Models will describe these models and their statistical inference in some detail.

Signal Detection Models

One of the earliest applications of discrete-state models was in the area of signal detection. A frequent signal detection paradigm is the Yes/No task (see Signal Detection Theory, History of). In this task, subjects are exposed to a series of items, some of which are classified as 'signal items' (S), and others are 'noise items' (N). One example is a recognition memory task where subjects first study a list of words, one at a time, and then they are given a test consisting of old studied words (signal items) and new, unstudied words (noise items, foils). The subject must classify each test word as 'Yes' (old studied word) or 'No' (foil). Other examples include auditory signal detection, where the signal items are tones that include extra energy at a particular frequency, and medical decision-making situations, where the signal items are X-rays that exhibit a particular type of tumor. In such a task, one observes the proportions of each response to each item type as a sample from the two-by-two probability matrix

$$P = \begin{bmatrix} \Pr(\text{'Yes'}|S) & \Pr(\text{'No'}|S) \\ \Pr(\text{'Yes'}|N) & \Pr(\text{'No'}|N) \end{bmatrix}. \qquad [2]$$

In standard signal detection terms, the top row of (2) exhibits, respectively, the hit and miss rates, and the bottom row, the false alarm and correct rejection rates.

Discrete-state models for the Yes/No task are characterized by the assumption that subjects set detection thresholds for the two kinds of items, and when an item is not detected, the subject responds with a guessing bias, $g \in [0, 1]$, for a 'Yes' response (Macmillan and Creelman, 2005). Such models postulate that any test item falls into one of three latent states: D_1, detected as a signal item, D_2, detected as a noise item, and D_3, not detected. The most general version of the threshold model specifies parameters for each item type to fall into each discrete state. Let d_{ij} be the probability that item type i leads to discrete state D_j, where $i = 1$ for signal items and $i = 2$ for noise items. The general threshold model specifies five nonnegative parameters, $\Theta = (d_{11}, d_{12}, d_{21}, d_{22}, g)$, where $d_{11} + d_{12} \leq 1, d_{21} + d_{22} \leq 1$. Following Eqn [1], the model can be specified by two matrices, a stimulus input matrix and a response output matrix. The stimulus matrix specifies the probabilities of getting into each of the three discrete states as a function of the item type, and the response matrix specifies the probability of a 'Yes' or 'No' response as a function of the discrete states. The stimulus matrix is given by

$$I(\Theta) = \begin{bmatrix} d_{11} & d_{12} & (1 - d_{11} - d_{12}) \\ d_{21} & d_{22} & (1 - d_{21} - d_{22}) \end{bmatrix}, \qquad [3]$$

and the response matrix by

$$R(\Theta) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ g & 1-g \end{bmatrix}, \quad [4]$$

where the first column of the response matrix gives the probabilities of a 'Yes' response given the discrete state. It is easy to compute Eqn [2] from Eqns [3] and [4] by matrix multiplication,

$$\begin{aligned} P(\Theta) &= I(\Theta)\cdot R(\Theta) \\ &= \begin{bmatrix} d_{11}+(1-d_{11}-d_{12})g & d_{12}+(1-d_{11}-d_{12})(1-g) \\ d_{21}+(1-d_{21}-d_{22})g & d_{22}+(1-d_{21}-d_{22})(1-g) \end{bmatrix}, \end{aligned} \quad [5]$$

and Eqn [5] is easily seen to include terms in the form of Eqn [1].

The general threshold model in Eqn [5] cannot fit uniquely to a single experimental condition with multiple independent and identically distributed observations on signal and noise items because it has five parameters, and the data structure in Eqn [2] is a product binomial structure with only 2 degrees of freedom. As a consequence, there are two special cases of the model that are the ones usually employed for data from a single condition. The first is the single high-threshold (SHT) model, and the second is the double high-threshold (DHT) model. The SHT model specification sets $d_{12} = d_{21} = d_{22} = 0$, and the DHT model sets $d_{11} = d_{22} \equiv D$, and $d_{12} = d_{21} = 0$. In essence both these models assume that the detection thresholds are set sufficiently high so that detections are always accurate, and failures to detect lead to the guessing process parameterized by g. The difference between the two models is that the DHT model, but not the SHT model, allows new items to be detected as new.

Other versions of the general threshold model either require data from multiple experimental conditions, where some of the parameters are equated across some of the conditions, or are further specified to accommodate more response options. In the later case, it is commonplace in signal detection studies to require subjects to indicate confidence in their response on an ordinal scale, e.g., 6-strong 'no', 5-probable 'no',…, 1-strong 'yes'. In such cases, Luce's low threshold theory discussed in Macmillan and Creelman (2005) can be fit to data.

As a generalization, many theorists have rejected threshold models for simple two-response signal detection tasks; however, recent considerations using model selection criteria have shown that the discrete-state models continue to have support, e.g., Klauer and Kellen (2011). Furthermore, discrete-state signal detection models with threshold assumptions have been used as building blocks for other, more complex models for multidimensional stimuli. Some of these models fall into the category of MPT models discussed in a later section.

Markov Learning Models

Markov learning models refer to a variety of discrete-state models for learning and memory tasks that were especially popular in the 1960s and 1970s (see Learning: Mathematical Learning Theory, History of). Most Markov learning models were developed for concept identification paradigms or for paired-associate learning paradigms. In a concept identification paradigm, stimuli are constructed by varying values over a series of stimulus attributes, e.g., size, shape, or color. The experimenter selects a concept rule (hypothesis) that is defined by a Boolean function over the attribute values, such as 'large things' or 'red triangles.' Then subjects receive a series of learning trials, where on each trial one of the stimuli is presented, and they indicate whether or not the stimulus satisfies the concept rule. Responses are followed by feedback of the correct response category, and subjects are expected to eventually learn the concept rule over a series of such trials. In a paired-associate learning paradigm, a list of stimulus items is paired in a many–one manner with a small set of response terms. Subjects attempt to learn the stimulus–response associations over a series of study trials, where the stimulus–response pairs are presented. Learning is measured by the ability of a subject to give the correct response alternative to each stimulus item on test trials. In both types of paradigms, the responses are typically coded as errors (1) or correct responses (0); however, for a subject in a concept identification experiment there is only one sequence of coded responses, whereas in paired-associate learning there is a sequence of coded responses for every stimulus–response pairing.

Most generally, Markov learning models are a type of stochastic process defined on the entities leading to Eqn [1], namely, stimulus items, latent cognitive states, and response categories. The latent cognitive states characterize possible states of learning in a Markov chain, for example, in the case of concept identification, the states could be candidate hypotheses, and in the case of paired-association learning, each stimulus–response pairing may be in a state of learning such as an unlearned state, a short-term memory state, a forgotten state, or a long-term memory state. The effect of a study trial is to generate transitions among the latent cognitive states described dynamically by a finite state, homogeneous Markov chain. On any trial, the response to a presented stimulus item is governed by the state's conditional response probability distribution as in Eqn [1]. The starting distribution over the states, the state-to-state transition probabilities, and the state-to-response probabilities are each functions of a parameter vector, $\Theta = (\theta_1, \ldots, \theta_S)$, where each component parameter taps some latent cognitive process, e.g., selecting a possible hypotheses, forming an association, or guessing a response. Technically, most Markov learning models can be viewed as discrete-trial, discrete-state, hidden Markov chains (e.g., Smyth, 1999), because only the trial-by-trial responses are observed, and the sequence of latent states can only be inferred. The statistical inference for a Markov learning model involves estimating the parameter vector Θ from the observed response sequences. The mathematical and statistical details are presented in Wickens (1982).

One of the earliest successful Markov learning models was the so-called all-or-none model for paired-associate learning. The model is for a paired-associate learning task in which each stimulus is paired with one of a small set of response alternatives such as numbered buttons. On any test trial, each stimulus–response pairing (item) is in one of two latent states: the learned state (L) and the unlearned state (U). When a stimulus is presented on a test trial, the correct response is made if the

item is in state L, but if it is in state U there is a probability $g \in [0,1]$ that the correct response is made. This model assumes that items start in state U, and they are learned independently. On every study trial when an item is in state U, there is a probability $c \in (0,1)$ that the item makes a transition to state L, which is an absorbing state in the Markov chain. The name of the model derives from the assumption that learning either completely occurs or it completely fails to occur.

The all-or-none model is specified by a parameter $\Theta = \langle c, g \rangle$, a transition matrix

$$T = \begin{array}{c} \\ L_n \\ U_n \end{array} \begin{array}{c} \begin{array}{cc} Ln+1 & Un+1 \end{array} \\ \begin{bmatrix} 1 & 0 \\ c & 1-c \end{bmatrix} \end{array}, \qquad [6]$$

a start vector $\mathbf{S}_1 = \langle 0, 1 \rangle$, and a response vector $\mathbf{R} = \langle 1, g \rangle$. Equation [6] gives the transition probabilities for a study trial on an item, where the row corresponds to the state of the item on the nth trial before the correct response is presented for the study and the column is the state of the item after the study event. The start vector gives the probability distribution over the states before the first study event, and the response vector gives the conditional probabilities of a correct response on any trial given the state. By ordinary properties of the discrete state, homogeneous Markov chains, the probability distribution after the nth study trial is given by

$$\mathrm{S}_{n+1} = \mathrm{S}_1 \mathbf{T}^n = \langle 1 - (1-c)^n, (1-c)^n \rangle$$

and the probability of a correct response on a test trial after the nth study is given by

$$\Pr(C_{n+1}) = \mathrm{S}_{n+1} \cdot R = 1 - (1-g)(1-c)^n.$$

Many examples of Markov learning models for concept identification are found in Wickens (1982), and Markov models for paired-associate learning are reviewed in Greeno and Bjork (1973). However, Markov learning models ceased to be of major theoretical importance in learning and memory in the late 1970s, when a variety of more elaborate and detailed IP models of learning were introduced such as global memory models (see Memory Models: Quantitative), and later connectionist models (see Connectionist Models of Concept Learning) and global memory models. Nevertheless, Markov learning models have continued to serve as psychometric models for psychological measurement. For example, they have been applied to early development and to aging, where standardized paradigms provide learning sequences for different age groups that can be used to estimate age-related trends in latent storage and retrieval processes (reviewed in Batchelder and Riefer, 1999).

Multinomial Processing Tree Models

MPT models constitute a large family of discrete-state IP models for categorical data that are quite popular at the time of this writing. The MPT model family was developed and formalized, including its statistical inference, in Riefer and Batchelder (1988) and Hu and Batchelder (1994). Batchelder and Riefer (1999) and Erdfelder et al. (2009) are comprehensive review articles on the MPT models. Informally, an MPT model for an IP task assumes that a stimulus item triggers one of a number of processing sequences, each of which leads to an observable response category. There may be multiple processing sequences that result in the same response category, so knowledge of the response category does not necessarily reveal the underlying processing sequence. The remainder of this main section on MPT models will have four subsections. The first subsection will discuss the statistical specification of MPT models including an MPT representation of the DHT model discussed in the section on signal detection models. The second subsection will describe two popular MPT models, each for a different experimental paradigm. The third subsection will describe statistical inference for MPT models, and the fourth subsection will discuss application areas and modeling issues.

Specification of MPT Models

Most discrete-state IP models, including those discussed in earlier sections, can be parameterized to fit into a subclass of MPT models called binary multinomial processing tree (BMPT) models, so that subclass is specified next in detail. Basically BMPT models are a class of parametric multinomial models that specify possible probability distributions over the observable response categories in terms of model parameters. There are four aspects to a BMPT model: (1) observable categories, (2) latent parameters, (3) tree architecture, and (4) computational rules.

The first aspect consists of J response categories, $\mathscr{C} = \{C_1, \ldots, C_J\}$. The assumption is that subjects are presented with items from a particular stimulus class, and they respond to each item independently with exactly one of the J response categories. In most experimental tasks, subjects are required to respond to items from two or more stimulus classes, each with its own set of response categories. In such cases, the model specifies a BMPT model for each class.

The second aspect of a BMPT model consists of S functionally independent parameters, $\Theta = (\theta_1, \ldots, \theta_S)$, each with full range, $0 \le \theta_s \le 1$. Thus, the parameter space of a BMPT model with S parameters is the S-dimensional unit interval $[0,1]^S$. Each parameter is interpreted as the probability of the occurrence of some latent (unobservable) cognitive act, e.g., storing an item in memory, retrieving an item from storage, matching an item with a target, clustering items, making an inference, detecting a signal, discriminating between items, or guessing a response.

The third aspect of a BMPT model is its tree architecture. Each stimulus class is associated with a full binary tree that represents a set of IP sequences, each of which leads to a particular response category. A full binary tree consists of a single initial node (the root), intermediate nodes, and terminal nodes (leaves). Each nonterminal node in the tree has exactly two directed links, each leading to another node (its child) in the tree. Such a tree is usually depicted with the initial node at the top and the terminal nodes at the bottom. **Figure 1** presents the DHT model described earlier in Eqn [5] in the form of a BMPT model with two full binary trees, one for the class of old items and the other for the class of new items. Both trees have two internal nodes, three terminal nodes, and four links. Each branch (sequence of links) from the initial node to a terminal node corresponds to a possible IP sequence leading to a response category as described next.

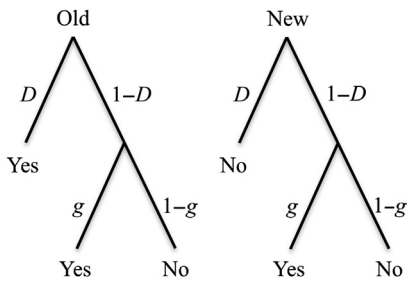

Figure 1 A BMPT representation of the DHT model in terms of a tree for old and a tree for new items. D is the detection parameter and g is the guessing parameter.

The specification of a BMPT involves assigning a response category to each terminal node, and a parameter, $\theta_s \in \Theta$, or a number, $x \in (0, 1)$, to each nonterminal node. If a particular parameter θ_s is assigned to a nonterminal node, then the left link from that node corresponds to the success, with probability θ_s, and the right link the failure, with probability $(1 - \theta_s)$, of the corresponding cognitive act represented by the parameter. The nodes with number assignments work similarly, where x is the probability of taking the left link and $(1-x)$ is the probability of taking the right link. For example, at the initial node at the top of the left tree in Figure 1, the parameter D is assigned, and the effect of this is to associate probability D with the left link and $(1-D)$ with the right link, as depicted in the figure. Finally, each terminal node corresponds to one of the J categories. Note that a particular category or parameter may appear several times in a BMPT tree, for example, in the left tree of Figure 1, the 'Yes' category appears twice because for the DHT model there are two different processing sequences that can lead to a 'Yes' response.

The computational component of any BMPT model provides the probabilities of each observable response category as the sum of the branch probabilities leading to that category. Branch probabilities are computed as products of link probabilities and have a special form. Let B_{jk} be the kth branch in a tree leading to category C_j, then

$$\Pr(B_{jk}|\Theta) = c_{jk} \prod_{s=1}^{S} \theta_s^{a_{jk,s}} (1 - \theta_s)^{b_{jk,s}} \qquad [7]$$

where c_{jk} is the product of the numbers associated with the links of B_{jk} or set to one if there are no numerical links on branch B_{jk}, and $a_{jk,s}$ and $b_{jk,s}$ are, respectively, the number of links on B_{jk} associated with θ_s and $(1-\theta_s)$. Since each branch ends in a particular category, the category probabilities, $p_j(\Theta)$, are sums of branch probabilities,

$$p_j(\Theta) = \sum_{k=1}^{I_j} \Pr(B_{jk}|\Theta) \qquad [8]$$

where I_j is the number of branches leading to category C_j. Given this structure, it is easy to show that

$$\forall \Theta \in [0, 1]^S, \;\; p_j(\Theta) \in [0, 1], \;\; \sum_{j=1}^{J} p_j(\Theta) = 1$$

Note that in Figure 1, there are two branches in the left tree that lead to a 'Yes' response. From Eqn [7], the left most one can be written as $\Pr[B_{\gamma,1}|(D,g)] = 1 \cdot D^1 (1-D)^0 g^0 (1-g)^0$, and the second one can be written as $\Pr[B_{\gamma,2}|(D,g)] = 1 \cdot D^0 (1-D)^1 g^1 (1-g)^0$. Putting these together using Eqn [8] yields $\Pr(\text{'Yes'}|\text{Old}) = D + (1-D)g$, and this is the same probability as the hit rate in the upper left term in the matrix in Eqn [5] for the DHT model.

As in Figure 1, most models are specified for situations where subjects respond to items from several classes, each with its own BMPT model. In such cases, any given parameter can appear in several trees. The DHT model in Figure 1 has two trees, and the parameters D and g appear in both. Each item class has two response categories, and a 'Yes' response to an old item is not the same category as a 'Yes' response to a new item.

The class of MPT models also includes ones where nonterminal nodes of the tree can have more than two links, each corresponding to a different processing outcome at that node. These MPT models are called multilink MPT models, and Figure 3 (described later) is an example of such a model. Hu and Batchelder (1994) specify the class of general processing tree (GPT) models to include multilink MPT models. It is a statistically useful fact that every multilink MPT model can be reparameterized into a statistically equivalent BMPT model, where statistically equivalence means that both models entail exactly the same set of probability distributions over the response category systems.

Purdy and Batchelder (2009) provide an alternative and efficient specification of BMPT models as strings in a context-free language. The specification capitalizes on the recursive character of the tree structure of BMPT models. For example, given any two BMPT models, one can join them into a single BMPT model by specifying a new initial root node with links to the two root nodes of the two models, and by the same logic one can decompose a BMPT model recursively in terms of sub-BMPT models. Similarly, one can specify the response category probabilities recursively in terms of the category response probabilities of the submodels. The context-free language in Purdy and Batchelder (2009) consists of a collection of strings, each of which corresponds to (codes) a particular BMPT model. In fact, the entire collection of strings is in one-to-one correspondence with the collection of BMPT models that can be specified as described in this subsection. These strings provide an efficient way to code

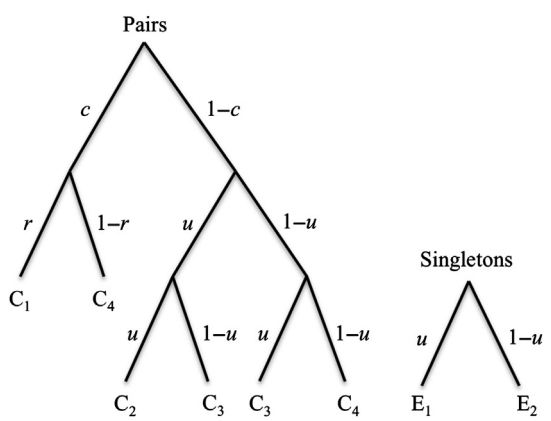

Figure 2 A BMPT representation of the pair-clustering model in terms of a tree for clusterable pairs and a tree for singleton items.

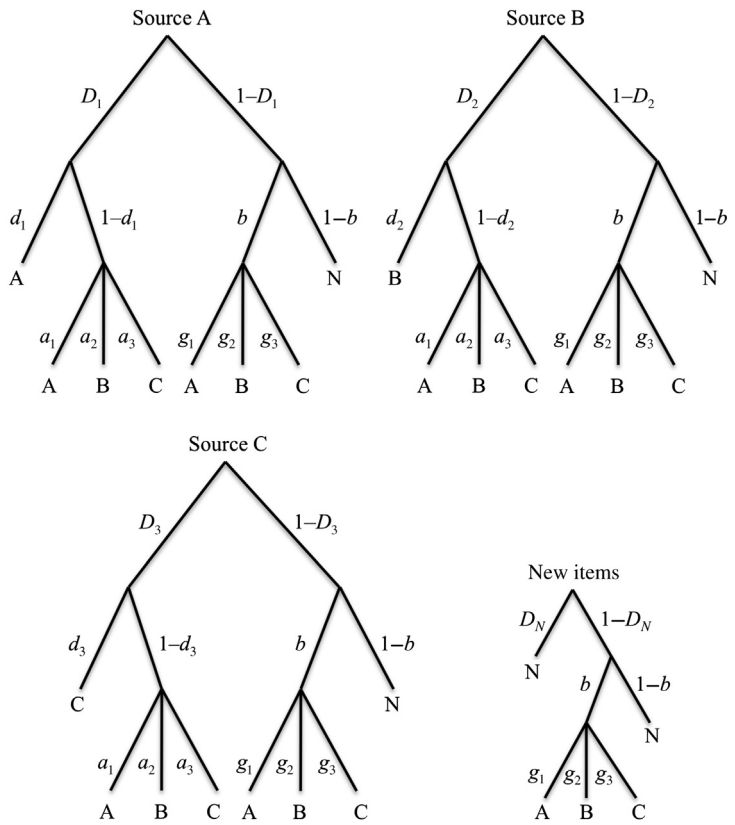

Figure 3 An MPT model representation of the source-monitoring model for three sources.

BMPT models for computation, and they have begun to be seen in statistical inference packages.

Examples of MPT Models

In this subsection, two popular examples of MPT models will be presented. One of the first published MPT models was the pair-clustering model. The data for the model involves a specially designed free recall task, where subjects study a list of words, one at a time, and then at a later time try to recall as many list words as they can in any order. The list of words contains two classes of items: semantically clusterable word pairs, e.g., car and train or daisy and rose, and singleton words. The word pairs are classified into four response categories: (C_1) both items recalled successively, (C_2) both words recalled but not successively, (C_3) only one of the two words recalled, (C_4) neither word recalled. The singleton words are classified into two response categories: (E_1) recalled and (E_2) not recalled. The model is presented in Figure 2.

The pair-clustering model specifies three parameters, each designed to measure a different cognitive act: (1) a storage parameter c representing the probability that a word pair is clustered and stored in memory, (2) a retrieval parameter r representing the conditional probability that a clustered word pair is recalled from memory, and (3) a parameter u representing the probability that an unclustered word is both stored in memory and retrieved. Just as in Figure 1, the pair-clustering model has two trees for the two classes of stimulus items. The assignment of categories to branches in the tree for pairs is based on two approximations: (1) a subject will recall the two words in a pair sequentially if and only if they are stored as a cluster and the cluster is retrieved, namely, $\Pr(C_1|\Theta) = c \cdot r$; (2) all nonclustered items, including singleton items, are stored and retrieved independently, with probability u. Note that there are two branches in the left tree that lead to category C_3 and two that lead to C_4. In addition, note that the parameter u appears three times in the tree for pairs and once in the tree for singletons.

The model is designed to separately measure memory storage capacity and memory retrieval capacity in humans. This is important for studying the bases of memory deficits in special populations because the percentage of correct recalls confounds these two latent processes. Batchelder and Riefer (1999) and Erdfelder et al. (2009) provide many references to the origin, statistical inference, and applications of the pair-clustering model.

Another popular area for MPT modeling is a source-monitoring task. Such a task is an extension of a recognition memory task discussed earlier, where the items on the study list come from different classes defined by their presentation source, e.g., different speakers, different sensory modalities, or from different spatial locations. In the test phase, subjects must decide if an item is old or new, and in addition, they must identify the presentation source of old items. Figure 3 provides a general MPT model for a three-source experiment.

The general source monitoring model in Figure 3 has trees for each of the three classes of presentation sources for the old items (A, B, or C) and a tree for new items, and across all four

trees there are 29 processing branches. Each of the four trees is a multilink MPT model because the internal nodes near the bottom of the trees have three links rather than two as required for BMPT models. The model has 14 nonnegative parameters: four 'detection parameters' ($D \leq 1$), three 'discrimination parameters' ($d \leq 1$), and seven other parameters for guessing processes. However, the actual number of independent parameters is 12 because $a_3 = (1 - a_1 - a_2)$, $g_3 = (1 - g_1 - g_2)$. Each of the four trees has a detection parameter, D, associated with its initial node that gives the probability of correctly detecting if an item is old or new. On the three trees for old items, on the detection (D) link is a discrimination parameter, d, for the probability that the source of the old item is correctly discriminated. Successful detections and discriminations lead to correct responses to items; however, failures to detect, with probability ($1-D$), and failures to discriminate, with probability ($1-d$), lead to guessing processes. Of the guessing parameters, b is the probability that a nondetected item is biased into one of the old source categories, and the parameters a_i and g_i, respectively, are the probabilities of guessing particular sources, for nondiscriminated old items and nondetected items.

Special cases of the source-monitoring model in **Figure 3** have been used in a variety of research settings including illusory correlations, face recognition, social categorization, and a variety of clinical settings involving special populations of subjects. The advantage of the model is that it can suggest whether or not the locus of memory failures is in old/new item detection or in the ability to discriminate the presentation source of detected items. Batchelder and Riefer (1999) and Erdfelder et al. (2009) provide many references to the origin, statistical inference, and applications of MPT models for source monitoring.

Statistical Inference for MPT Models

It is a consequence of the computational rules for an MPT model that if the category counts in a particular class of stimulus items are obtained as independent and identically distributed observations, then the MPT model becomes a specially parameterized multinomial model with probability distribution

$$\Pr(< N_1, \ldots, N_J > |\Theta) = N! \prod_{j=1}^{J} \frac{p_j(\Theta)^{N_j}}{N_j!} \qquad [9]$$

where N_j is the number of counts in category C_j and $N = \sum N_j$. In the case of multiple classes of stimulus items, if the responses to all the items are independent, Eqn [9] just becomes a product of multinomials, one for each item class. This marks such models technically as parameterized product multinomial models. For example, for the DHT model of **Figure 1**, the probability distribution for the parameterized product binomial model is given by

$$\begin{aligned}\Pr[\langle N_{y1}, N_{n1}\rangle, \langle N_{y2}, N_{n2}\rangle | (D, g)] \\ = N_1! \frac{[D + (1-D)g]^{N_{y1}}}{N_{y1}!} \cdot \frac{[(1-D)(1-g)]^{N_{n1}}}{N_{n1}!} \\ \bullet N_2! \frac{[(1-D)g]^{N_{y2}}}{N_{y2}!} \cdot \frac{[(D + (1-D)(1-g)]^{N_{n2}}}{N_{n2}!},\end{aligned}$$

where $N_1 = N_{1y} + N_{1n}$ and $N_2 = N_{2y} + N_{2n}$.

Under the assumptions leading to Eqn [9], Hu and Batchelder (1994) provide classical (frequentist) methods for statistical inference for the entire class of BMPT models. The key to the inference is based on a special application of the expectation–maximization (EM) algorithm, a frequent data augmentation algorithm used in classical statistical inference. Note from Eqn [9] that there may be several branches that lead to the same category, so knowing the category counts does not directly reveal the frequency of times that each branch in the BMPT tree is traversed. The EM algorithm treats these branch frequencies as 'missing data'. Let N_{jk} be the frequency out of N_j counts of the occurrence of branch k terminating in category j, where $1 \leq k \leq I_j$, $1 \leq j \leq J$, and I_j is the number of branches ending in category C_j. Suppose the N_{jk} are known, then it is easy to see that the maximum likelihood estimators of the θ_s are given by

$$\hat{\theta}_s = \frac{\sum_{j=1}^{J} \sum_{k=1}^{I_j} a_{jk,s} N_{jk}}{\sum_{j=1}^{J} \sum_{k=1}^{I_j} (a_{jk,s} + b_{jk,s}) N_{jk}}, \qquad [10]$$

where the terms in Eqn [10] are defined in Eqn [7]. Thus, the maximization step of the EM algorithm yields a closed form expression.

There is a software package in Hu and Phillips (1999) based on the statistical theory in Hu and Batchelder (1994) that enables a researcher to design an MPT model graphically; load in the data; obtain point estimates, confidence intervals, and goodness-of-fit measures; do hypothesis testing within and between subject groups; and run Monte Carlo simulations with or without beta-distributed hyperparameters. There are several other general-purpose approaches to inference for MPT models that have appeared that satisfy the sampling assumptions leading to Eqn [9] and its extension to product multinomials, for example, population parameter mapping (Chechile, 1998) and other examples that are discussed in Erdfelder et al. (2009).

In the past decade, users of MPT models have begun to recognize that most categorical data sets in cognitive psychology violate the assumption behind Eqn [9] that the category counts are independent and identically distributed over the response categories. Violations occur when there are individual parameter differences among subjects and/or items. Smith and Batchelder (2008) provide model-free, nonparametric tests of heterogeneity in subjects and/or items for categorical data, and several recent papers reviewed in Erdfelder (2009) have provided Bayesian inference for MPT models that can handle such heterogeneities. It is expected that future applications of MPT models will increasingly utilize Bayesian inference to handle heterogeneity.

Application Areas and Modeling Issues

Batchelder and Riefer (1999) describe 30 MPT models for different experimental paradigms, and a decade later Erdfelder et al. (2009) in the lead article to a special issue of *Zeitschrift für Psychologie* on MPT modeling, describe 70 MPT models and variants in over 20 research areas. Many of these models are in subareas of human memory such as the models in **Figures 1–3**. MPT models have also been developed for other areas of human memory including eyewitness memory, hindsight bias,

process dissociation, prospective memory, and serial recall. There are MPT models in several areas of human reasoning, including propositional and syllogistic reasoning. In the area of perception, there are MPT models for object and letter recognition, speech recognition, and selective attention. In addition, there are MPT models for several experimental situations in the area of social psychology, e.g., stereotype formation, social categorization, and illusory correlation.

One of the main applications of MPT models is to measure cognitive capacities in special populations by estimating and comparing relevant model parameters. Batchelder (1998) examines the differences between using MPT models to understand experimental tasks and using them for psychological assessment. Erdfelder et al. (2009) describe MPT models that have also been used in a variety of assessment tasks such as discovering the basis of cognitive deficits in aging and clinical studies, detecting consensus in ethnographic studies, and implicit attitude measurement in social settings. The use of an MPT model for measurement requires that the cognitive meanings of the parameters are valid, namely, that it taps the probability of the cognitive act that it is designed to represent. Validity is not decided by the ability of the model to fit data, but instead it must be decided by specially constructed experiments, where specific model parameters are shown to be selectively influenced by relevant psychological variables, and unaffected by others. Consequently, almost all published articles advancing a new MPT model presents a series of experiments that support the connection of each model parameter with the underlying cognitive act that it is designed to measure.

The family of MPT models stands in contrast to most other IP models in that the latter assume much more complex cognitive architecture, more elaborate underlying processes, and in particular, continuous type variables that determine the output. In fact, many theorists view discrete-state models as theoretically impoverished. Indeed, this view is common in the signal detection literature, e.g., Macmillan and Creelman (2005), as well as in other areas of cognition, e.g., Mulder et al. (1995). In practice, however, MPT models have been most useful as simple, approximate models to measure underlying cognitive processes, in particular experimental paradigms, rather than providing a deep understanding of cognitive processes across a wide range of phenomena. In this sense, their applications led Batchelder and Riefer (1999) to coin the term cognitive psychometrics to cover the use of simple cognitive models, such as MPT models, for psychometric purposes.

See also: Concept Learning and Representation: Models; Learning: Mathematical Learning Theory, History of; Mathematical Psychology; Memory Models: Quantitative; Psychometrics; Signal Detection Theory, History of.

Bibliography

Batchelder, W.H., 1998. Multinomial processing tree models and psychological assessment. Psychological Assessment 10, 331–344.

Batchelder, W.H., Riefer, D.M., 1999. Theoretical and empirical review of multinomial processing tree modeling. Psychonomic Bulletin & Review 6, 57–86.

Chechile, R.A., 1998. A new method for estimating parameters for multinomial data. Journal of Mathematical Psychology 42, 432–471.

Erdfelder, E., Auer, T.-S., Hilbig, B.E., Aßfalg, A., Moshagen, M., Nadarevic, L., 2009. Multinomial processing tree models: a review of the literature. Zeitschrift für Psychologie/Journal of Psychology 217, 108–124.

Greeno, J.G., Bjork, R.A., 1973. Mathematical learning theory and the new 'mental forestry'. Annual Review of Psychology 24, 81–116.

Hu, X., Batchelder, W.H., 1994. The statistical analysis of general processing tree models with the EM algorithm. Psychometrika 59, 21–47.

Hu, X., Phillips, G.A., 1999. GPT.EXE: a powerful tool for the visualization and analysis of general processing tree models. Behavior Research Methods, Instruments and Computers 31, 220–234.

Klauer, K.C., Kellen, D., 2011. The flexibility of models for recognition memory: an analysis by the minimum-description length principle. Journal of Mathematical Psychology 55, 430–450.

Macmillan, N.A., Creelman, C.D., 2005. Detection Theory: A User's Guide, second ed. Cambridge University Press, New York.

Mulder, G., Sanders, A.F., van Galen, G.P. (Eds.), 1995. Discrete and Continuous Information Processing. Acta Psychologica, 90. (special issue).

Purdy, B.P., Batchelder, W.H., 2009. A context free language for binary multinomial processing tree models. Journal of Mathematical Psychology 53, 547–561.

Riefer, D.M., Batchelder, W.H., 1988. Multinomial modeling and the measurement of cognitive processes. Psychological Review 95, 318–339.

Smith, J.B., Batchelder, W.H., 2008. Assessing individual differences in categorical data. Psychonomic Bulletin & Review 15, 713–731.

Smyth, P., 1999. Hidden Markov models. In: Wilson, R.A., Keil, F.C. (Eds.), The MIT Encyclopedia of the Cognitive Sciences. MIT Press, Cambridge, MA, pp. 373–374.

Wickens, T.D., 1982. Models for Behavior: Stochastic Processes in Psychology. Cambridge University Press, New York.

Discrimination and the Law

Deborah C Malamud, New York University Law School, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article focuses on the development of antidiscrimination law in the United States at the federal level. It is concerned primarily with employment law and discriminatory practices involving race and gender. Other areas of US law are also considered (e.g., school desegregation) before briefly surveying international developments. The role of social science in antidiscrimination legal practice is stressed.

The Oxford Dictionary defines discrimination as 'the perceiving, noting, or making a distinction or difference between things.' As modified by descriptors such as 'invidious,' the term 'discrimination' refers to the making of distinctions that ought not to be made. When the distinction is inappropriate, because it treats differently those individuals, entities, or practices that are entitled by law to equal treatment, 'discrimination' takes on its legal meaning.

Contemporary antidiscrimination law is predominantly (or at least most controversially) concerned with distinctions drawn (or, in some cases, inappropriately *not* drawn) on the basis of socially defined personal attributes that have been and remain the bases for adverse treatment (such as race, gender, national origin, religion, disability, and sexual orientation). The legal concept of discrimination is not limited in its historical origins to group-based distinctions of this sort. As a concept in American constitutional rhetoric and law, the term also designates protectionist distinctions between states and countries engaged in interstate and foreign commerce. For purposes of this article, however, our focus shall be on 'discrimination' of the intergroup sort, in part because it is that form of discrimination that has most interested the social scientific community.

The development of any nation's antidiscrimination law is deeply dependent upon its own history and culture. In the field of race and ethnicity, the need for legal redress for race discrimination arises out of variable histories of immigration and its selective suppression, slavery and indentured labor, colonialism, and genocide. Social movements calling for legal and social change are culturally specific as well, although influenced by past experiences elsewhere in the world. Movements to identify and redress gender discrimination are also shaped by local settings. This is true not merely because the degree to which gender takes center stage depends in part upon whether and how the stage must be shared with the issue of race, but also because the variable social situation of women has much to do with how women will conceptualize their own equality and the spheres of activity in which they will press their legal claims. This article will focus on efforts by the United States at the federal level, but will conclude with a discussion of leading issues from several other countries.

Federal Antidiscrimination Law in the United States

Federal antidiscrimination efforts reach a wide array of social and legal arenas. Some of the most important of these are education, employment, voting, public and private housing, public accommodations (e.g., transportation), bank lending, government contracting, marriage and parenting, and the administration of criminal justice.

In light of the American history of slavery, civil war, failed post-civil war reconstruction of the South, and the mid-twentieth-century civil rights movement aimed at ending the pervasive segregation and oppression of African-Americans, race has been the most frequent starting point for American antidiscrimination law. The Fourteenth Amendment of the Constitution of the United States, itself a product of the Civil War experience, contains an Equal Protection Clause that serves as the primary source of federal constitutional antidiscrimination doctrine. Federal equal protection analysis generally assigns different 'levels of scrutiny' to different types of discriminatory classifications. For example, race discrimination is subject to 'strict scrutiny'; gender discrimination to 'intermediate scrutiny'; and age discrimination to 'rationality review,' the same level of review that is given to ordinary economic legislation. For a general survey, see Yoshino (2011). Governmental classifications rarely survive strict scrutiny and are rarely struck down under rationality review, although there are important exceptions that will be discussed below. The tiered structure of levels of scrutiny was long criticized by US Supreme Court Justices Stevens and Marshall (see, e.g., *Craig v. Boren, 429 U.S. 190, 211 (1976) (Stevens, J., concurring) and Dandridge v. Williams, 397 U.S. 471, 520–521 (1970) (Marshall, J., dissenting)*). Those critiques have contemporary echoes in the Court's gay-marriage jurisprudence, in which the Court is unclear as to how (if at all) it is using traditional tiered scrutiny (see *United States v. Windsor, 133 S.Ct. 2675 (2013)*; see also *United States v. Windsor, 133 S.Ct. 2675 (2013) at 2706–2707 (Scalia, J., dissenting)*).

Much of the federal antidiscrimination effort in the United States has been made under the auspices of antidiscrimination statutes. The first major federal antidiscrimination statute, the Civil Rights Act of 1964, was stimulated by the violent resistance by the South to court-ordered racial desegregation and to the 1960s civil rights movement. As interpreted by the Supreme Court, the Fourteenth Amendment does not reach discrimination by purely private parties, *see The Civil Rights Cases, 109 U.S. 3 (1883)*, and it requires proof of discriminatory intent, *see Washington v. Davis, 426 U.S. 229 (1976)*. So long as federal legislation is enacted pursuant to some other source of Congressional power (e.g., the Commerce Power), however, Congress is free to fill these gaps. Congress has also used legislation to create administrative agencies to participate in the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86023-7

work of antidiscrimination enforcement (for example, the Equal Employment Opportunity Commission in the field of employment discrimination) or to provide novel remedial mechanisms (for example, the preclearance provisions of Section 5 of the Voting Rights Act of 1965, the reauthorization of which was struck down by the Supreme Court in *Shelby County v. Holder, 133 S.Ct. 2612 (2013)*).

The American law of discrimination starts with high-sounding commitments in seminal statutes or Supreme Court opinions, but its development rapidly takes the form of the elaboration of highly specific structures of proof for dealing with discrimination claims arising under the various sources of law and in the various areas of social practice surveyed above. One result is that antidiscrimination law has developed a number of distinct specialties, each concerned with a separate jurisprudence of discrimination in its own field. Another result is that, in American antidiscrimination law, the devil is in the details. By shaping proof structures, and by applying civil procedure and evidence rules to discrimination cases, courts shape the likelihood that plaintiffs will succeed in their claims (Albiston, 1999; Bennett, 2012).

Employment

The leading federal antidiscrimination statute in the employment field, Title VII of the Civil Rights Act of 1964 (known as 'Title VII'), prohibits discrimination on the basis of race, color, sex (including pregnancy, thanks to the Pregnancy Discrimination Act of 1978), religion, and national origin. Although feminist activists lobbied for the inclusion of sex discrimination, the statute as originally proposed left sex discrimination out. It was included through an amendment proposed by a congressional opponent of the statute, whose confidence that including sex would block passage of the statute proved misplaced (Kanowitz, 1968). Once enacted, however, the prohibition of sex discrimination has been a robust and creative aspect of statutory enforcement. For example, the law of discriminatory harassment was first articulated in race discrimination cases but experienced its full legal development in the gender field.

Employment discrimination on the basis of age and disability were subsequently addressed in the Age Discrimination in Employment Act (1967) and Americans with Disabilities Act (1993) (significantly amended by the Americans With Disabilities Act Amendment Act of 2008). When Congress enacts antidiscrimination legislation on behalf of groups that do not receive 'strict scrutiny' under constitutional law, there is some tendency for the Supreme Court to be less plaintiff-friendly in its statutory interpretation – leading, in some cases, to Congressional pushback through statutory amendment (Hart, 2006). Not all groups have met with success in Congress. For example, Congress has yet to enact the long-proposed Employment Non-Discrimination Act (ENDA), which would bar employment discrimination on the basis of sexual orientation. There is debate as to whether the increasingly successful movement in support of same-sex marriage has slowed the progress toward the passage of ENDA. Compare Klarman, 2012 with Burke, 2013.

Years of development in the courts have generated several basic 'proof structures' that are used to prove discrimination under Title VII, and they are often borrowed for purposes of other antidiscrimination statutes: 'individual disparate treatment' (divided between 'pretext' and 'mixed motive' frameworks), 'harassment,' 'failure to accommodate,' 'pattern and practice,' and 'disparate impact.' (The Civil Rights Act of 1991 codified the 'disparate impact' and 'mixed motive' approaches for purposes of Title VII.) The 'individual disparate treatment' and 'pattern and practice' proof structures turn upon proof of the employer's discriminatory intent. The disparate impact proof structure does not; it relies instead upon proof of discriminatory effects of racially neutral practices, and upon the failure of the employer to prove the 'business necessity' of those practices.

In the 'individual disparate treatment' case, intent is an inference drawn by a fact finder (judge or jury) from a variety of types of circumstantial evidence aimed at refuting the possible nondiscriminatory reasons for an adverse employment action. While this sounds relatively straightforward, decades of Supreme Court cases have failed to achieve clarity, both as to the requirements for getting to trial in a 'pretext' case and as to the relationship between the 'pretext' and 'mixed motive' approaches (see Sperino, 2011). In addition, these cases (like all cases under the federal employment discrimination laws) are adjudicated not by specialized agencies or labor courts, but by generalist federal trial courts that are predisposed to undervalue them, finding creative ways to use procedural and evidentiary rulings to dismiss the cases without trial (see Bennett, 2012; Gertner, 2012). Increasingly, the Supreme Court has embraced the use of private arbitration to resolve individual employment discrimination claims, see *14 Penn Plaza LLC v. Pyett, 556 U.S. 247 (2009)*.

Most harassment cases involve proof that an employer has created a 'hostile environment' because of the plaintiff's sex, race, or other protected classification. In order for the employer to be held liable for the hostile environment, the plaintiff must also satisfy a set of requirements (derived from the common law of agency) that the Supreme Court created in order to incentivize employers to create, and employees to utilize, internal compliance processes. See *Burlington Industries, Inc. v. Ellerth, 524 U.S. 742 (1998); Faragher v. City of Boca Raton, 524 U.S. 775 (1998)*.

Limited within Title VII to religion cases is the duty to accommodate: an employer is deemed to have discriminated on the basis of religion if it fails to afford reasonable accommodation to a plaintiff's religious beliefs or practices, so long as doing so does not create undue hardship. Because the Supreme Court has found this statutory accommodation requirement to be in tension with the Establishment Clause of the First Amendment of the Constitution, it has essentially held that 'undue hardship' is present whenever a requested accommodation causes more than de minimis harm to the employer or to coworkers. *See Trans World Airlines, Inc. v. Hardison, 432 U.S. 63 (1977)*.

In the typical 'pattern or practice' cases, plaintiffs (most often a group of plaintiffs using the procedural mechanism of the class action) rely primarily on labor market analysis and statistical evidence to create an inference of intentional discrimination. In some cases, plaintiffs bolster their claims by the use of 'social framework' evidence, i.e., social scientific testimony aimed at demonstrating that the structures of the

work environment foster intentionally discriminatory decision-making (see Fiske and Borgida, 2013; Hart and Secunda, 2009). Claims are generally further supported through evidence of individual instances of intentional discrimination on the 'pretext' model. In a procedural decision that has substantive implications for Title VII, some members of the Supreme Court have signaled discomfort with finding a 'pattern or practice' of discrimination in a large entity absent proof of a centralized policy of discrimination. *Wal-Mart Stores Inc. v. Dukes, 131 S. Ct. 2541 (2011)*. Part of the Court's concern is the same as that which shapes the Court's hostile-environment harassment jurisprudence: i.e., those employers whose official policies prohibit discrimination ought not to be held liable when their subordinates disobey those policies (Green, 2011).

In a 'disparate impact' case, a plaintiff (or, more frequently, a plaintiff class) challenges a facially neutral employment practice (for example, a test) on the grounds of its disparate adverse effects on members of his or her group. So, for example, in the case of *Griggs v. Duke Power Co*. (401 US 424 [1971]), markedly lower proportions of blacks and Latinos than whites in the relevant labor market passed standardized aptitude tests and had high school diplomas, leading to diminished access to preferred jobs in the employer's workforce. Absent a showing of 'business necessity' or 'job relatedness,' the Supreme Court held, employers could not use tests and other hiring criteria that placed 'arbitrary headwinds' in the way of the progress of traditionally excluded groups. As in the case of 'pattern and practice' cases, disparate impact cases turn on social scientific expert testimony: economists (labor market analysis), psychologists (psychometric analysis), and sometimes sociologists (social framework analysis) play a crucial role.

Even after Congress passed clarificatory legislation in 1991, elements of the disparate impact proof structure remain uncertain. Because these cases involve comparison-group definitions, questions are raised in every case as to whether the reference group should be the employer's existing labor force, the applicant pool, or some other definition of the relevant labor market. The Equal Employment Opportunity Commission's standard for showing disparity has been criticized for failing to comport with social scientific definitions of statistical significance (Meier et al., 1984). The precise conceptual content of the phrases 'business necessity' and 'job relatedness' is still subject to doubt, particularly when disparate impact analysis is used to challenge employment practices other than paper-and-pencil tests (where 'job relatedness' has long been tested through psychometric techniques of test validation).

The disparate impact cause of action is available not only for objective tests (e.g., paper-and-pencil tests, physical skills tests), but also for subjective hiring practices (*Watson v. Fort Worth Bank & Trust, 487 U.S. 977 (1988)*). However, in a procedural ruling under the federal 'class action' rules, the Supreme Court has cast doubt on whether it will accept 'a "policy" of allowing discretion by local supervisors over employment matters' to serve as the basis for a disparate impact claim (*Wal-Mart Stores Inc. v. Dukes, 131 S. Ct. 2541, 2554-55 (2011)*; see Pedersen, 2012.). The Supreme Court has also restricted employers' capacity to use race-conscious measures to avoid potential disparate impact liability for tests with a demonstrable disparate impact. That option is available only if the employer has a 'strong basis in evidence' that its test does not meet the 'business necessity' standard (*Ricci v. DeStefano, 557 U.S. 557 (2009)*).

One area in which Title VII law has been relatively stable is the law of affirmative action. Under *United Steelworkers of America v. Weber* (442 US 193[1979]), employers are permitted to use affirmative action (if it is consistent with standards laid out in that opinion), and the Supreme Court reiterated more recently that the employer does not need to prove that its own prior policies were discriminatory in order to do so (*Johnson v. Transportation Agency of Santa Clara County*, 480 US 616 [1987]). The debate among the justices in these cases was open and virulent, and yet affirmative action in employment has not provoked recent challenge (in contrast to the status of affirmative action under the constitution, discussed below). There is, however, a mismatch between current social practice and the Supreme Court's past opinions. The Supreme Court has never held that the desire for greater diversity is a permissible basis for affirmative action under Title VII; despite that fact, diversity is the major stated justification for many employer affirmative action programs (see, e.g., Skrentny, 2013; Estlund, 2005). In addition, the Court's decision in the *Ricci* case is in some tension with its affirmative action jurisprudence (Shin and Gulati, 2011).

Cases decided under Title VII have long served as the model for developments in other fields of federal statutory employment discrimination law, but that tendency has its limits. Ironically, when Congress rejected a set of conservative Supreme Court decisions in the Civil Rights Act of 1991, it provided the Supreme Court with a new basis in statutory text for confining the 'mixed motive' cause of action to the core provisions of Title VII. (*See Gross v. FBL Financial Services, Inc., 557 U.S. 167 (2009); University of Texas Southwestern Medical Center v. Nassar, 133 S.Ct. 2517 (2013)*.) Fortunately for the disabled, one arena in which the Title VII model has not been followed is disability rights, where, especially after the passage of the Americans with Disabilities Amendments Act of 2008, disabled individuals benefit from an accommodation requirement that is far more robust than that available under Title VII's religion provisions.

The example of American employment discrimination law shows that the true state of antidiscrimination law cannot be assessed merely by looking at the written laws themselves. One must also be aware of how the statutes are being interpreted by courts, relevant administrative agencies, and even private arbitrators; of how generally applicable procedural and evidentiary rules are used in discrimination cases; of the expense of litigation; and of the practical availability of experts with necessary skills and experience. One must also have a sense of trajectory (is the law becoming more or less friendly to plaintiffs over time?) and of symmetry (are similar laws being interpreted in similar or dissimilar ways?).

Other Areas of American Law

In considering proof structures outside the employment discrimination field, a key factor is the unavailability of the disparate impact cause of action in constitutional cases arising under the Fourteenth Amendment (*Washington v. Davis*, 26 US

229 [1976]). Although *Washington v. Davis* was in fact an employment-testing case, the Court's decision was likely motivated by the wide range of government programs that might be subject to challenge under the Constitution.

Because of the lack of the disparate impact route, showing discriminatory intent is the key to constitutional antidiscrimination litigation. In primary- and secondary-school desegregation litigation, close judicial scrutiny was triggered in years past by findings that boards of education were operating their schools in an intentionally discriminatory fashion. Courts retained jurisdiction to administer complex court-ordered desegregation plans, often for decades. Until such time as the districts are declared 'unitary' by the presiding court, practices that were not themselves intentionally discriminatory could be challenged as perpetuating the effects of past intentional discrimination. In recent years, however, the trend has been for courts to conclude that the school districts under their jurisdiction have achieved 'unitary' status and that any continuing inequalities within the system no longer bear an organic connection to the district's past history of intentional discrimination (Orfield and Eaton, 1996). The Supreme Court has held that once a school district achieves 'unitary' status, it may only use race-conscious remedies if they satisfy 'strict scrutiny' – and it insists that 'unitary' districts may never use race in the assignment of individual pupils to particular classrooms or schools.

In higher education, affirmative action has been the focus of major political and judicial attention. The Supreme Court affirmed the constitutionality of the use of race-based affirmative action to achieve diversity in higher education in *Grutter v. Bollinger, 539 U.S. 306 (2003)*. However, 10 years later the Court stressed in *Fisher v. University of Texas at Austin, 133 S. Ct. 2411 (2013)*, that, under strict scrutiny, universities must prove that race-neutral alternatives (e.g., affirmative action based on socioeconomic class) are unavailable. Furthermore, in *Schuette v. Coalition to Defend Affirmative Action, 134 S.Ct. 1623 (2014)*, the Court upheld the power of states to enact state constitutional amendments that deny public universities the discretion to use race-conscious affirmative action.

Criminal justice in the United States has a significant racial dimension too, one that contributes to the gap in perception of the fairness of the system between blacks and whites. Some important efforts to redress racism in criminal law enforcement have met with failure, in part because of the difficulty of persuading courts of discriminatory intent through statistical evidence. For example, in *McCleskey v. Kemp* (481 US 279 [1987]), the Court rejected the claim, based on extensive statistical evidence, that the death penalty as practiced by state governments is unconstitutionally discriminatory (Baldus et al., 1983). Racial disparities in criminal law enforcement have led to scholarly and popular critiques of modern criminal justice as a current instantiation of American racial apartheid: "the new Jim Crow" (Alexander, 2012; for a response see Forman, 2012).

"In the field of voting rights, the Supreme Court has made it more difficult for voters to challenge allegedly discriminatory voting practices. Under the preclearance provision of the Voting Rights Act of 1965, proof of discriminatory intent was not required for such challenges if they were brought against certain designated jurisdictions with a history of past intentional voting discrimination. The Supreme Court held Congress' reauthorization of that provision unconstitutional in *Shelby County v. Holder, 133 S. Ct. 2612 (2013)*."

The Role of Social Science in American Antidiscrimination Legal Practice

As should be apparent from the above survey of American antidiscrimination practice, social science has a crucial role to play in proving discrimination. Social science has also been an important reality check, by asking important empirical questions both about the efficacy of current antidiscrimination laws and the variables underlying the development of effective antidiscrimination law. Social science has played the weakest role, however, in the area where it may well have been the most helpful – that is, in developing the law's basic concepts of discrimination (Nelson et al., 2008).

In the employment discrimination field, expert testimony from social scientists is a necessity in disparate impact and pattern and practice cases. At a minimum, several types of social scientific expertise must be brought into a case. In both disparate impact and pattern and practice cases, labor market experts are necessary to formulate and to defend definitions of the relevant labor market, using census data, industry surveys, and other sources. Testing psychologists (psychometricians) play a crucial role in disparate impact cases, both in designing employment screening tests in such a way as will allow them to withstand legal scrutiny and in challenging the designs of tests in litigation. Statisticians are needed in both types of cases to ensure that appropriate data are collected through the court-administered evidence discovery process, to formulate an affirmative case from the data, and to critique the rival statistical story being told by the opposing party. Sociologists testify as to the characteristics of workplaces that are likely to generate discriminatory practices ("social framework evidence") (Hart and Secunda, 2009; Fiske and Borgida, 2013.). Expertise from a number of disciplines is drawn upon to demonstrate or to refute the extent to which underrepresentation of groups results from their lack of interest in particular categories of jobs (e.g., Schultz, 1990). Thus, disparate impact and pattern and practice cases are universally viewed as battles of the experts – which is a problem for the untrained judges and juries who must evaluate the evidence. Expert testimony has also found its way into individual disparate treatment cases. In *Price Waterhouse v. Hopkins* (490 US 228 [1989]), for example, expert testimony was admitted in order to demonstrate that certain statements by supervisors were based on sexual stereotypes (see Antczak, 2011).

While employment discrimination litigation employs the talents of psychologists, labor economists, sociologists, and statisticians, the field of voting rights has brought political scientists into the expert witness role. Legislatures employ sophisticated demographic and political models when they redistrict, and the political scientists who create these models have played an important role in defending them and their results in court. Educational desegregation litigators utilize the expertise of specialists not only in education but also in demography, because patterns of housing segregation play an important role in long-term educational segregation.

Outside of the courtroom context, social scientists have questioned the effectiveness of existing antidiscrimination

frameworks both in the United States and elsewhere (e.g., Donohue and Heckman, 1991; McCrudden et al., 1991; Edelman, 1990). Social scientists are engaged in a raging debate on whether race-based affirmative action in higher education helps or hurts its beneficiaries (compare Sander, 2004 with Chambers et al., 2005).

Where social science has played less of a role is in the development of the law's basic concepts of discrimination. Contemporary studies of prejudice in numerous disciplines are critical of the intent-based model of discrimination that permeates American law (Nelson et al., 2008). A large body of work in the field of cognitive psychology has demonstrated that whites' reaction times to pairings of words with images of black faces are far faster when the words are negative than when they are positive. Calling this reaction-time disparity 'implicit bias,' scholars have argued that this form of subconscious discrimination is pervasive (not only for race, but also for gender, national origin, and disability) and that it calls into question the intent-based paradigm in American antidiscrimination law (Kang and Lane, 2010; Krieger and Fiske, 2006). (Project Implicit makes implicit association tests available on line at https://implicit.harvard.edu/implicit/takeatest.html.)

It is too soon to know whether implicit bias research will come to play a role in American antidiscrimination law. Some critics contend that the research method does not adequately model decision-making outside the laboratory. Others argue that the very prevalence of 'implicit bias' means that the phenomenon cannot, in fact, be something the law ought to categorize as a legally prohibited wrong (Mitchell and Tetlock, 2006; for responses, see Jost et al., 2009; Eyer, 2012). Perhaps, for these skeptics, implicit bias research is too pessimistic both about human nature and about the efficacy of existing legal approaches and remedies. Even some supporters of the research question whether "an expert can validly determine whether implicit bias caused a specific employment decision" (Faigman et al., 2008 at 1394). Rather than trying to fit implicit bias into litigation fact-finding, they argue that the phenomenon is best addressed through systematic changes in the structure of the institutions in which discrimination takes place.

International Developments

As other nations and transnational entities address the discrimination in their own midst, they learn not only from their own histories but also from the successes and the failures of the American model. A few examples in which other countries have gone beyond the American model illustrate the possibilities.

In the United States, the disparate impact principle (or, as it is often known outside the United States, the 'indirect discrimination' principle) has been limited to circumstances in which statutes expressly specify its use. However, that principle has been embraced in numerous national and international framework laws (treaty or constitutional), including the International Convention on the Elimination of All Forms of Racial Discrimination, the European Union's Racial Equality Directive, and the Constitution of South Africa (Daniel, 2011). The law of free speech under the First Amendment of the United States Constitution has stood in the way of pervasive hate speech regulation in the United States. In contrast, out of the historical experience of Germany has grown a strong commitment in German and European law to prohibiting the group-hate speech exemplified by expressions of anti-Semitism (Kahn, 2013; Schauer and Pildes, 1999; Stein, 1986; Waldron, 2010).

In the field of gender discrimination in employment, the theory of 'comparable worth' is an important part of the gender discrimination law of the European Community and some of its member states (e.g., Linos, 2010; Ellis, 1998; Libeson, 1995). In comparable worth analysis, women use the tools of job and wage analysis to demonstrate that they do not receive the same returns on effort, education, and skill as do occupants of traditionally male jobs. Despite its importance in European jurisprudence, comparable worth has not made significant inroads into American law (Blum, 1991).

In the field of affirmative action (under its many names), Indian law has 'reserved' scarce positions, including legislative seats, for members of 'backward castes and classes' for decades (Galanter, 1984). In South Africa, the postapartheid constitution explicitly permits affirmative action (Adam, 1997; Higginbotham, 1999), and employs concepts of positive social and economic rights that the United States has never embraced (Sachs, 2000).

See also: Affirmative Action: Comparative Policies and Controversies; Affirmative Action: Empirical Work on Its Effectiveness; Civil Liberties and Human Rights; Class Actions, Legal; Death Penalty; Discrimination, Economics of; Discrimination: Racial; Employment and Labor, Regulation of; Equality and Inequality: Legal Aspects; Equality of Opportunity; Equality: Philosophical Aspects; Expert Testimony; Gender and the Law; Human Rights, History of; Injustice: Legal Aspects; Justice, Conflicts, and the Justice of Conflict Resolution; Law as an Instrument of Social Change; Prejudice in Society: Psychological Perspectives; Race and the Law; Scientific Evidence: Legal Aspects; Sexual Harassment: United States and Beyond; Sexual Orientation and the Law; Technology, Social Construction of.

Bibliography

Adam, K., 1997. The politics of redress: South African style affirmative action. Journal of Modern African Studies 35 (2), 231–249.

Albiston, C., 1999. The rule of law and the litigation process: the paradox of losing by winning. Law Society Review 33, 869–910.

Alexander, M., 2012. The New Jim Crow: Mass Incarceration in the Age of Colorblindness, rev. ed., New Press, New York.

Antczak, H.M., 2011. Problems at Daubert: expert testimony in title VII sex discrimination and sexual harassment litigation. Buffalo Journal of Gender Law and Social Policy 19, 33.

Baldus, D.C., Pulaski, C., Woodworth, G., 1983. Comparative review of death sentences: an empirical study of the Georgia experience. Journal of Criminal Law and Criminology 74, 661.

Bennett, M.W., 2012. From the "no spittin, 'no cussin' and no summary judgment" days of employment discrimination litigation to the "defendant's summary judgment affirmed without comment" days: one judge's four-decade perspective. New York Law School Law Review 57, 685.

Blum, L.M., 1991. Between Feminism and Labor: The Significance of the Comparable Worth Movement. University of California Press, Berkeley, CA.
Burke, T.F., 2013. From the courthouse to the chalkboard. Tulsa Law Review 49, 305 (Review of Klarman, 2012).
Chambers, D.L., et al., 2005. The real impact of eliminating affirmative action in American law schools: an empirical critique of Richard Sander's study. Stanford Law Review 57, 1855.
Daniel, A., 2011. The intent doctrine and CERD: how the United States fails to meet its international obligations in racial discrimination jurisprudence. DePaul Journal for Social Justice 4, 263.
Donohue III, J.J., Heckman, J., 1991. Continuous versus episodic change: the impact of civil rights policy on the economic status of blacks. Journal of Economic Literature 29 (4), 1603–1643.
Edelman, L.B., 1990. Legal environments and organizational governance: the expansion of due process in the American workplace. American Journal of Sociology 95, 1401.
Ellis, E., 1998. EC Sex Equality Law, second ed. Clarendon Press, Oxford, UK.
Estlund, C., 2005. Putting Grutter to work: diversity, integration and affirmative action in the workplace. Berkeley Journal of Employment and Labor Law 26, 1.
Eyer, K.R., 2012. That's not discrimination: American beliefs and the limits of anti-discrimination law. Minnesota Law Review 96, 1275.
Faigman, D.L., et al., 2008. A matter of fit: the law of discrimination and the science of implicit bias. Hastings Law Journal 59, 1389.
Fiske, S.T., Borgida, E., 2013. Standards for using social psychological evidence in employment discrimination cases. Temple Law Review 83, 867.
Forman, J., 2012. Racial critiques of mass incarceration: beyond the new Jim Crow. New York University Law Review 87, 21.
Galanter, M., 1984. Competing Equalities: Law and the Backward Classes in India. University of California Press, Berkeley, CA.
Green, T., 2011. The future of systemic disparate treatment law. Berkeley Journal of Employment and Labor Law 32, 395.
Gertner, N., 2012. Losers' rules. Yale Law Journal Online 122, 109.
Hart, M., 2006. Skepticism and expertise: the Supreme Court and the EEOC. Fordham Law Review 74, 1937.
Hart, M., Secunda, P.M., 2009. A matter of context: social framework evidence in employment discrimination class actions. Fordham Law Review 78, 37.
Higginbotham, F.M., 1999. Affirmative action in the United States and South Africa: lessons from the other side. Temple International and Comparative Law Journal 13, 187–230.
Jost, J.T., et al., 2009. The existence of implicit bias is beyond reasonable doubt: a refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. Research in Organizational Behavior 29, 39.
Kahn, R.A., 2013. Why do Europeans ban hate speech? A debate between Karl Loewenstein and Robert Post. Hofstra Law Review 41, 545.
Kang, J., Lane, K., 2010. Seeing through colorblindness: implicit bias and the law. UCLA Law Review 58, 465.
Kanowitz, L., 1968. Sex-based discrimination in American law III: title VII of the Civil Rights Act and the Equal Pay Act of 1963. Hastings Law Review 20, 305.
Klarman, M.J., 2012. From the Closet to the Altar: Courts, Backlash, and the Struggle for Same-sex Marriage. Oxford University Press, Oxford.
Krieger, L.H., Fiske, S.T., 2006. Behavioral realism in employment discrimination law: implicit bias and disparate treatment. California Law Review 94, 997.
Libeson, S.J., 1995. Reviving the comparative worth debate in the United States: a look toward the European Community. Comparative Labor Law Journal 16 (3), 358–398.
Linos, K., 2010. Path dependence in discrimination law: employment cases in the United States and the European Union. Yale Journal of International Law 35, 115.
McCrudden, C., Smith, D.J., Brown, C., 1991. Racial Justice at Work: Enforcement of the Race Relations Act 1976 in Employment. Policy Studies Institute, London.
Meier, P., et al., 1984. What happened in Hazelwood. American Bar Foundation Research Journal 1, 139–186.
Mitchell, G., Tetlock, P.E., 2006. Antidiscrimination law and the perils of mindreading. Ohio State Law Journal 67, 1023.
Nelson, R.L., et al., 2008. Divergent paths: conflicting conceptions of employment discrimination in the law and social sciences. Annual Review of Law and Social Science 4, 103.
Orfield, G., Eaton, S.E., 1996. Dismantling Desegregation: The Quiet Reversal of Brown v. Board of Education. The New Press, New York.
Pedersen, N.B., 2012. The hazards of Dukes: the substantive consequences of a procedural decision. University of Toledo Law Review 44, 123.
Sachs, A., 2000. Social and economic rights: can they be made justiciable? Southern Methodist University Law Review 53, 1381.
Sander, R.H., 2004. A systemic analysis of affirmative action in American law schools. Stanford Law Review 57, 367.
Schauer, F., Pildes, R.H., 1999. Electoral exceptionalism and the First Amendment. Texas Law Review 77, 1803–1836.
Shin, P.S., Gulati, M., 2011. Showcasing diversity. North Carolina Law Review 28, 1017.
Skrentny, J.R., 2013. After Civil Rights: Racial Realism in the New American Workplace. Princeton University Press, Princeton, NJ.
Schultz, V., 1990. Telling stories about women and work: judicial interpretations of sex segregation in the workplace in title VII cases raising the lack of interest argument. Harvard Law Review 103 (8), 1749–1843.
Sperino, S., 2011. Rethinking discrimination law. Michigan Law Review 110, 69.
Stein, E., 1986. History against free speech: the new German laws against the 'Auschwitz' and other 'Lies. Michigan Law Review 85, 277–328.
Waldron, J., 2010. Dignity and defamation: the visibility of hate. Harvard Law Review 123, 1596.
Yoshino, K., 2011. The new equal protection. Harvard Law Review 124, 747.

Discrimination, Economics of

Ronald Oaxaca, University of Arizona, Phoenix, AZ, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 3756–3762, © 2001, Elsevier Ltd.

Abstract

In the abstract, 'discrimination' refers to distinctions or differentiation made among objects or individuals. In economics, as in ordinary parlance, discrimination typically carries a pejorative connotation. The term is generally reserved for distinctions that are socially unacceptable and economically inefficient. Economic theories of discrimination can be categorized into theories about tastes and preferences of economic agents, economic power, and statistical discrimination. Achieving a consensus on what the broad features of discrimination in markets, especially labor markets, should entail is a seemingly straightforward task compared with obtaining a consensus on empirical measures of discrimination. The standard measurement technique is that of wage decompositions. With the advent of newer and more sophisticated econometric methods for estimating earning relationships, measurement of discrimination has become more ambiguous. Legal and employee morale considerations constrain attempts to remove discrimination in the workplace in a way that would correspond to accepted economic principles of efficiency. Comparable worth seeks to remedy discrimination by relying upon experts to value job attributes in an unbiased manner. Empirical studies of labor-market discrimination are global in nature and encompass both developed countries and developing countries.

Discrimination

In the abstract, 'discrimination' refers to distinctions or differentiation made among objects or individuals. In economics, as in ordinary parlance, discrimination carries a pejorative connotation. The term is generally reserved for distinctions that are socially unacceptable and economically inefficient. Economic theories of discrimination usually fall into one of three broad categories: tastes and preferences of economic agents, economic power, and statistical discrimination. Unlike the world of theory, empirical documentation of discrimination is subject to much ambiguity and differences of interpretation.

Theories of Discrimination

Tastes and Preferences

Publication of *The Economics of Discrimination* (1957) by Gary Becker was instrumental in making discrimination a legitimate and popular topic for study by economists. The tastes and preferences approach adopted by Becker was pure neoclassical economics. This approach appeals to tastes for discrimination by utility-maximizing economic agents. Tastes for discrimination are embodied in discrimination coefficients that are susceptible to measurement by the economist's monetary yardstick. Agents' tastes for discrimination are manifested by a desire to minimize economic transactions with members of certain demographic groups. While discriminating agents may be prepared to forfeit income in order to indulge their tastes, the effects of their discriminatory behavior can lead to segregation or reduced incomes for the groups discriminated against. In addition to employer tastes for discrimination, Becker also considered the effects of tastes for discrimination by consumers and fellow workers. Under some circumstances, competitive pressures can ameliorate the effects of discriminatory behavior.

The literature provides support for Becker's contention that labor-market discrimination is lessened by product market competition. Ashenfelter and Hannan (1986) present statistical evidence that within the banking sector the employment opportunities for women are negatively related to market concentration. Peoples and Saunders (1993) show that deregulation of trucking in the USA lowered the white/black wage ratios among both union and nonunion drivers. This competitive effect was identified with the reduction of rents accruing to whites.

One avenue for exploring the effects of discrimination by fellow workers is to examine the effects of unionism on racial wage differentials. An early study by Ashenfelter (1972) showed that the average wage of black workers relative to the average wage of white workers in the USA was higher in the union sector than in the nonunion sector. In particular, Ashenfelter estimated that compared with the complete absence of unionism, the black/white wage ratio was 4.0% higher in the industrial union sector, and 5.0% lower in the craft union sector. These results suggest that the union sector is less discriminatory than the nonunion sector, and that industrial unions are relatively more egalitarian compared with craft unions. Peoples (1994) found evidence not only that the returns to union membership are higher for blacks than for whites, but also that racial wage differentials were higher among nonunion workers in the highly concentrated industries.

Market Power

Prior to Becker, early work by economists on the topic of discrimination focused on pay differentials between men and women. Joan Robinson's (1933) classic on imperfect competition provided an early example of the economic power theory of discrimination. Robinson considered the case of a monopsonistic employer (single employer labor market) of men and women who are equally productive (perfect substitutes in

production). She proved that if the labor supply elasticity (responsiveness to wages) were smaller for women, a profit-maximizing employer would equate the incremental employment costs of men and women by offering women a lower wage.

Madden (1975) explored the circumstances under which the labor supply elasticities of women would be lower, and how male monopoly power insulates men from any monopsony power that might exist in labor markets. The key element was the existence of job segregation, which reduced competition from women. It was shown that under these circumstances women would receive lower wages than men, if both worked in equally productive, but different, sets of jobs. Earlier, Bergmann (1971) advanced the occupational crowding hypothesis to explain income inequality between white male workers and otherwise equally capable nonwhite workers. Crowding nonwhites and women into a narrower range of occupations raises the incremental value of white males in their occupations and lowers the incremental value of all workers in the occupations dominated by nonwhites and women.

The ubiquity and self-reinforcement of a broad range of discriminatory behavior in society was examined in Thurow (1969). For example, discrimination in schooling opportunities has the effect of diminishing the subsequent earnings potential of minority workers. This reduces the costs to discriminating employers of paying lower wages to minority workers, as the latter will have fewer skills to begin with. Lower wages reduce the cost of housing discrimination because better housing is beyond the incomes of most minority workers.

Statistical Discrimination

Statistical discrimination is an approach that relies upon risk and information costs rather than tastes for discrimination or market power to explain income disparities among different demographic groups in society (Aigner and Cain, 1977). One example is that of two groups who are equally productive on average, but the productivity signal, for example, test scores, is a noisier signal of productivity for one of the groups. A risk-averse, profit-maximizing employer might offer lower wages or less desirable positions to applicants from the demographic group with the noisier signal. Consequently, employment decisions are made on the basis of group characteristics and individual characteristics.

Wage Decompositions

In the study of wage inequality, the most commonly used methodology for apportioning wage gaps between two demographic groups to productivity and discrimination can be traced to Blinder (1973) and Oaxaca (1973). The basic idea is quite simple. One first estimates wage equations for each of the demographic groups from a sample of workers. These wage equations are mathematical formulas relating wages to observed productivity factors such as education and experience. In the absence of direct labor-market discrimination, equity is characterized as a situation in which both demographic groups are expected to be compensated according to the same wage (salary)-determination formula. In this scenario the only reason for the existence of a wage gap would be that the two groups differ on average with respect to the observed wage (productivity)-determining characteristics. This wage gap is commonly referred to as the 'endowment' or human capital effect.

The difference between the actual wage gap and the endowment effect is often referred to as the 'unexplained gap.' Many practitioners regard the unexplained gap as an estimate of discrimination. Those who prefer the term 'unexplained gap' reason that this estimate is in the nature of a residual and as such may include the effects of productivity factors not explicitly taken into account in the wage-determination formula. Hence, the measure may be a biased estimate of discrimination. From an econometric standpoint, the bias can go in either direction. The counterargument is that there are always factors omitted from any statistical relationship and that there is no reason to suppose that these are anything but random in the present context. In addition, the same sort of methodology is very commonly used to estimate other types of wage gaps, for example, union/nonunion, north/south, manufacturing/nonmanufacturing. In these contexts the term 'unexplained gap' is not used.

According to the Blinder/Oaxaca standard decomposition methodology, one of the two demographic comparison groups is designated as the reference group. If, for example, one is interested in male/female wage differentials, one might select males as the reference group. This means that the statistically estimated wage formula for males serves as the nondiscriminatory wage structure. Operationally, one estimates the average wages for both males and females by applying the estimated male wage formulation to each group's wage-determining characteristics. The gender difference in these predicted average wages is the 'endowment' effect, since the same wage formula is applied to the characteristics of both groups. The difference between the original unadjusted wage gap and the gap attributable to the 'endowment' effect is by definition the result of different wage formulas for males and females. This residual gap is what is termed either the 'unexplained gap' or the gap attributable to direct labor-market discrimination. Therefore, the decomposition technique is a method for decomposing an observed wage gap into an endowment component and an unexplained/discriminatory component.

Identification of Individual Components of Discrimination

Apart from differences in what researchers call the 'residual gap,' there are other ambiguities in the standard decomposition technique. First, it would seem desirable to be able to identify the contribution of each wage-determination characteristic to the endowment component and the discrimination component. For example, one could identify the contribution of gender differences in a characteristic such as work experience to the overall endowment effect as the gender difference in average work experience multiplied by the pecuniary return to an additional unit of work experience estimated from the male wage equation. Or consider the case of a wage-determining characteristic defined by a set of mutually exclusive categories such as educational levels, for example, 'less than high school,' 'high school,' 'some college,' and 'college graduate.' The estimated return to each included educational category is relative

to the return on an arbitrarily selected educational category variable that is omitted from the wage equation. The monetary value of gender differences in the proportion of workers in the sample who are college graduates is simply calculated as the gender difference in the proportion of college graduates weighted by the estimated value of a college education for males relative to the omitted educational category. The sum of all of these endowment valuations equals the overall endowment effect.

Unfortunately, an identification problem can arise when one attempts to estimate the contribution of each wage-determination characteristic to the discrimination component. The problem is initially described in Jones (1983) and further extended in Oaxaca and Ransom (1999). For example, weighting the average amount of work experience for females by the male–female difference in the estimated return to an additional unit of work experience would constitute a measure of the contribution of gender differences in the return to experience to the overall discrimination measure. It turns out that this computation is affected by a simple shift in the experience measure. For example, potential experience is commonly calculated as either schooling-age-five or schooling-age-six. This is a seemingly innocuous shift in the experience measure, yet it will affect the estimated contribution to discrimination of gender differences in the return to experience. Because this change in the experience measure creates an offsetting change in the intercept term in the wage equations for males and females, it does not affect the overall estimate of discrimination. Nor does this change in the experience measure have any effect on the contribution of gender differences in work experience to the endowment effect.

This arbitrariness is more manifest when attempting to estimate the contribution to discrimination of gender differences in the returns to a variable defined by mutually exclusive categories. Returning to the education example, note that the contribution to measured discrimination of the gender difference in the returns to education is calculated as the sum of the gender differences in the estimated returns for each included educational category weighted by the proportion of females who are in each included category. As it turns out, this calculation is affected by the choice of which educational category variable is to play the role of the omitted reference category. The choice of omitted reference category is of course quite arbitrary. Fortunately, the overall estimated discrimination component is not affected by which categorical variables are selected to be the omitted reference categories. Also, the choice of omitted reference group has no effect on the estimated contribution to the endowment effect of gender differences in the values of the categorical variables.

Nondiscriminatory Wage Structures

To examine a second source of ambiguity in standard wage decompositions, imagine that the choice of reference group is reversed so that female workers are designated as the reference group. Now one estimates the (nondiscriminatory) average wages for both males and females by applying the estimated female wage formulation to each group's wage-determining characteristics. An ambiguity arises in that in general estimates of the endowment and discriminatory wage effects will depend on which demographic group is serving as the reference group. This ambiguity is a manifestation of the familiar index number problem. The researcher is trying to measure 'value' using alternative base weights.

As a measure of discrimination, the estimated residual gap is subject to quite different interpretations depending on the choice of reference group. The situation is one in which the presence of discrimination is manifested as a wage gap based on the fact that in actuality men and women do not face the same wage-determination process. This can be viewed as either that women are underpaid relative to the male wage structure or else that men are overpaid relative to the female wage structure. The interpretation and estimated magnitude of each of the two components comprising the original unadjusted wage gap depend on which group's wage structure is regarded as the nondiscriminatory standard. To further add to the ambiguity surrounding wage decomposition analysis, there is the theoretical possibility that which group is being discriminated against depends on whose wage-determination formula is being used as the nondiscriminatory standard.

It is clear that the second type of ambiguity surrounding the standard wage decomposition analysis has to do with identifying an appropriate nondiscriminatory wage standard. Ideally, one would want the wage structure that would prevail in a nondiscriminatory world to serve as the standard by which to measure actual discrimination. Under some assumptions, such a wage structure would be the one that would exist in a world characterized by perfect competition in product markets and labor markets. One would estimate the average wages for both males and females by applying the nondiscriminatory wage formulation to each group's wage-determining characteristics. The gender difference in these predicted average wages is the endowment effect. Again, discrimination is measured by the unadjusted gender wage gap net of the endowment effect. However, it is now possible to further decompose the discriminatory wage gap because the gap is, in general, the sum of two distinct components. The first component is the difference between the actual average wage of the male workers, and their predicted average wage based on the nondiscriminatory structure. The second component is the difference between the predicted average wage of females based on the nondiscriminatory wage structure, and their actual average wage.

These discriminatory wage components have interesting interpretations (Neumark, 1988). The first component is a measure of the male wage advantage relative to the nondiscriminatory norm. This component can be viewed as favoritism toward males. The second component is a measure of the female wage disadvantage relative to the nondiscriminatory norm. This component can be viewed as pure discrimination against female workers. This methodology, in principle, allows researchers to sort out how much of the male–female discriminatory wage gap results from overcompensation of men, and how much stems from undercompensation of women. In the light of this approach, the conventional decomposition method is seen as a special case.

While seemingly attractive as a decomposition method, there is a practical difficulty in making operational the concept of a nondiscriminatory wage structure. One notion is the wage structure that would prevail had discrimination never existed. A slightly less challenging task might be to define the

nondiscriminatory wage structure as the structure that would emerge in the immediate aftermath of a cessation of discrimination. In general, this is not the same as the wage structure that would have prevailed had discrimination never existed. This is because of a likely path-dependence feature in which discrimination permanently alters the evolution of wage structures. While a still formidable enterprise, it is possible to reach useful approximations to a wage structure that would emerge after the cessation of discrimination.

The basic idea underlying this approach is contained in Neumark (1988) and Oaxaca and Ransom (1999). A reasonable presumption is that with the cessation of discrimination between two demographic groups, the resulting wage structure would, in some way, be related to the preexisting separate wage structures. A natural approximation is the wage structure that is estimated on the basis of a combined sample obtained by pooling the data on the two demographic groups. Such a common wage structure can be shown to be a matrix-weighted average of the separately estimated wage structures. The weights are related to the precision by which each group's wage-equation parameters are statistically estimated. Neumark (1988) applies this methodology to the gender wage gap.

In Oaxaca and Ransom (1994), the methodology is generalized to allow for estimation of a nondiscriminatory wage structure as any matrix-weighted average of the separately estimated wage structures. Without restrictions on the weighting matrices, any common wage structure can be rationalized as a matrix-weighted average of the separately estimated wage structures. Reimers (1983) uses simple averages of separately estimated wage structures when comparing whites, blacks, and Hispanics.

Residual Decomposition

An important aspect of the study of wage inequality is trying to understand the components of changes or trends in wage inequality. Juhn et al. (1991) introduce a method of decomposing changes in the residual wage gap between two demographic groups into the effects of changes in the average rankings of the two groups in the residual wage distribution of the dominant group, and the effects of changes in the residual wage dispersion (standard deviation) of the dominant group. This approach is referred to as the Juhn–Murphy–Pierce (JMP) decomposition. It follows the practice of interpreting the wage gap left after controlling for group differences in observed characteristics as an unexplained differential.

In Blau and Kahn (1996), the JMP decomposition is adapted to analyzing gender wage gap differences between nine countries and the USA during the 1980s. Earnings are standardized to a full-time (40 h) work schedule. The earnings models estimated with microdata sets control for education, potential experience, union membership, industry, and occupation. **Table 1** summarizes the Blau and Kahn findings in which the gender wage gaps in each country are compared with that of the USA. Column 2 reports the raw (log) wage differentials by country. The wage gaps range from 0.258 log points (approximately 26%) in Sweden to 0.4875 log points (approximately 49%) in the UK. Column 3 reports each country's wage gap minus that of the USA. It is this difference that is decomposed into the effects of international differences in the following: (1) gender differences in human capital characteristics (quantities), (2) the (male) returns to human capital characteristics (prices), (3) gender differences in the percentile rankings in the male wage residual distribution (rankings), and (4) differences in the (male) standard deviation of the wage residuals (dispersion). In six of the nine countries, the male human capital advantage is greater than that of the USA. Without exception, international differences in the returns to human capital and in the dispersion of wage residuals lower the gender wage gap in other countries relative to the USA. In six of the nine countries the less favorable rankings of women in the male wage residual distribution widens the gender wage gap relative to the USA. What emerges from this study is that the gender wage gap is relatively high in the USA because the greater wage residual inequality in the USA more than offsets the greater relative human capital of the US women.

Correcting Wage Inequities

Although economists are presumably interested in measuring the extent to which discriminatory wage gaps represent overpayment of the dominant group versus underpayment of the

Table 1 International gender wage gaps

Country	*G*	$G - G_{US}$	*Quantities*	*Prices*	*Rankings*	*Dispersion*
Australia	0.310 1	−0.119 8	0.056 3	−0.073 2	−0.067 1	−0.035 8
Austria	0.319 5	−0.105 9	0.057 5	−0.157 3	0.281 3	−0.287 3
Germany	0.354 2	−0.071 2	0.031 8	−0.097 3	0.242 1	−0.248 0
Hungary	0.432 6	0.012 9	−0.056 7	−0.028 7	0.627 2	−0.529 0
Italy	0.324 9	−0.100 8	0.029 9	−0.009 4	−0.023 9	−0.097 4
Norway	0.349 0	−0.076 1	−0.076 0	−0.014 2	0.582 9	−0.568 8
Sweden	0.258 0	−0.167 4	−0.030 9	−0.027 0	−0.012 2	−0.097 1
Switzerland	0.437 7	0.012 3	0.103 0	−0.006 1	0.068 3	−0.153 0
UK	0.487 5	0.062 1	0.010 8	−0.041 1	0.423 2	−0.331 0
USA	0.425 4					

G: The (unadjusted) gender wage differentials (in logs). $G - G_{US}$: The gender wage gap difference between each country and the USA. Quantities: The effects of gender differences in personal characteristics. Prices: The effects of gender differences in the returns to personal characteristics. Rankings: The effects of gender differences in percentile rankings in the wage distribution. Dispersion: The effects of international differences in wage dispersion.
Source: Blau, F.D., Kahn, L.M., 1996. Wage structure and gender earnings differentials: an international comparison. Economica 63 (Suppl.), 29–62, Table 5.

subordinate group, custom and the legal environment lean more to the dominant group's wage structure as the nondiscriminatory standard. The idea is that inequities should be addressed by raising the wages of the subordinate group, as opposed to lowering the wages of the dominant group. Indeed, the US Civil Rights Act of 1964 prohibits equity wage adjustments that involve lowering any group's nominal wages. Thus, attempts to remove discrimination in the workplace in a way that would correspond to accepted economic principles of efficiency are constrained by legal and employee morale considerations.

Econometric Modeling Approaches

One might properly speculate about how econometric models of the wage-determination process can be used to address a salary inequity issue in the workplace. Examples from legal practice are provided in Finkelstein (1980). Oaxaca and Ransom (1999) report on several approaches on how to use statistical salary models to correct salary inequity within a firm. For illustrative purposes, the examples are in terms of gender inequity, but the methodology extends without modification to racial or ethnic salary inequity. The remedies range from simply compensating each female worker exactly according to the estimated male wage equation to constraining the salary equity adjustment process to award minimum settlements regardless of personal circumstances.

Comparable Worth

Comparable worth is a general policy for eradicating discriminatory wage disparities within firms. The idea behind comparable worth is that jobs of equal value should pay equal wages. While this sounds like free market economics, at issue is how value is determined (Hill and Killingsworth, 1989). Advocates of comparable worth envision a system in which points are awarded to each job based on the skills, effort, and responsibility required. Experts are responsible for evaluating a firm's jobs in a manner that does not bias wage determination against women. The objection against letting the market value the jobs is that occupational segregation dooms women to a subordinate wage position. An important question is whether or not comparable worth would provide employers with the incentives to eliminate job segregation, and whether or not women would switch among occupations that were equally compensated. Critics of comparable worth fear the massive inefficiencies that could emerge from interference with labor-market resource allocation.

Self-Selection in the Labor Force

Estimation of labor-market discrimination, in general, is complicated by the observation that individual workers in a statistical sample generally do not constitute a random sample of the adult population. Whether or not to participate in the labor force will depend upon the individual's calculation of the net benefits of labor force participation. While the net benefits are unobserved by researchers, the decision to participate is observed, as well as many determinants of the net benefits. For women, the presence of small children at home would likely be a factor in the decision to be in the labor market. The statistical equation corresponding to labor force participation is known as a selection equation, since it describes a rule for individuals to select themselves in or out of the labor force.

Sample selection bias can arise when the random statistical errors in the selection equation are correlated with the random statistical errors in the wage equation that is being estimated for a self-selected sample of workers (Heckman, 1979). Unless the probability of selection (which depends on the selection equation) is taken into account, the estimated parameters of the wage equation can be biased. The selection probabilities for the selected sample are controlled for in a variable known as the inverse Mills ratio. This variable is constructed from the variables and parameters of the selection equation.

A number of studies have attempted to correct for sample selection bias when dealing with occupational segregation or more generally labor-market discrimination (Reimers, 1983; Dolton et al., 1989; Neuman and Oaxaca, 1998). Neuman and Oaxaca (1998) explore alternative ways of estimating (ethnic and gender) labor-market discrimination from selectivity-corrected wage equations using a sample of Jewish professional workers drawn from the Israeli 1983 Census of Population and Housing.

Progress and Future Directions

Smith and Welch (1989) chronicle the economic progress of African-Americans over a 40-year period. Black male wages as a percent of the white male wage rose from 43.3% in 1940 to 72.6% by 1980. The documented wage gains of African-Americans were largely attributed to more and better education, migration, and general economic growth. Some evidence was found that affirmative action policies improved the employment and wage prospects of African-Americans. Data from the US Current Population Survey are consistent with the Smith and Welch findings. The median income of full-time, year-round black workers as a percent of the median income of full-time, year-round white workers rose from 68.1% in 1970 to 73.9% in 1998.

Empirical findings from gender labor-market studies are quite mixed, variable over time and across countries. Zabalza and Tzannatos (1985) examine the gender employment and wage effects of labor-market legislation in the UK on the employment and wages of women. Panel data, which contain detailed information on individuals over their working lives, have been used to study gender wage gaps in Denmark (Rosholm and Smith, 1996). In a shift of emphasis from gender wage discrimination Flückiger and Silber (1999) examine both theoretically and empirically the phenomenon of gender labor market segregation with special emphasis on the Swiss labor market.

Empirical studies of racial-, ethnic-, gender-, and caste-based labor-market outcomes in developing countries may also be found (Birdsall and Sabot, 1991; Psacharopolulos and Patrinos, 1994). Intercountry comparisons of labor-market inequity between the developed and the developing countries will be the inevitable result of globalization. The implications

of labor-market discrimination for international trade and specialization should occupy the attention of social scientists well into the twenty-first century.

See also: Discrimination: Racial; Discrimination: United States and Beyond; Economic Development and Women; Egalitarianism: Political; Equality and Inequality: Legal Aspects; Equality: Philosophical Aspects; Gender and Feminist Studies in Geography; Income Distribution and Inequality: Measurement Issues; Income Distribution: Demographic Aspects; Inequality; Inequality: Comparative Aspects; Labor Supply; Poverty, Sociology of; Prejudice in Society; Racism, Sociology of; Social Inequality in History (Stratification and Classes); Social Stratification; The Underclass; Urban Poverty in Neighborhoods; Wage Differential and Structure; Wealth Distribution.

Bibliography

Aigner, D., Cain, G., 1977. Statistical theories of discrimination in the labor market. Industrial and Labor Relations Review 30, 175–187.

Ashenfelter, O., 1972. Racial discrimination and trade unionism. Journal of Political Economy 80, 435–464.

Ashenfelter, O., Hannan, T., 1986. Sex discrimination and product market competition: the case of the banking industry. Quarterly Journal of Economics 101, 149–174.

Becker, G., 1957. The Economics of Discrimination. University of Chicago Press, Chicago.

Bergmann, B., 1971. The effect on white incomes of discrimination in employment. Journal of Political Economy 79, 294–313.

Birdsall, N., Sabot, R. (Eds.), 1991. Unfair Advantage: Labor Market Discrimination in Developing Countries. The World Bank, Washington, DC.

Blau, F.D., Kahn, L.M., 1996. Wage structure and gender earnings differentials: an international comparison. Economica 63 (Suppl.), 29–62.

Blinder, A.S., 1973. Wage discrimination: reduced form and structural estimates. Journal of Human Resources 8, 436–455.

Dolton, P.J., Makepeace, G.H., Van der Klaauw, W., 1989. Occupational choice and earnings determination: the role of sample selection and non-pecuniary factors. Oxford Economic Papers 41, 573–594.

Finkelstein, M., 1980. The judicial reception of multiple regression studies in race and sex discrimination cases. Columbia Law Review 80, 737–754.

Flückiger, Y., Silber, J., 1999. The Measurement of Segregation in the Labor Force. Physica-Verlag, Heidelberg, Germany.

Heckman, J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–161.

Hill, M.A., Killingsworth, M.R. (Eds.), 1989. Comparable Worth: Analyses and Evidence. Industrial Labor Relations Press, Ithaca, NY.

Jones, F.L., 1983. On decomposing the wage gap: a critical comment on Blinder's method. Journal of Human Resources 18, 126–130.

Juhn, C., Murphy, K.M., Pierce, B., 1991. Accounting for the slowdown in black-white wage convergence. In: Kosters, M.H. (Ed.), Workers and Their Wages. AEI Press, Washington, DC, pp. 107–143.

Madden, J.F., 1975. Discrimination – A manifestation of male market power? In: Lloyd, C.B. (Ed.), Sex, Discrimination, and the Division of Labor. Columbia University Press, New York. Chapter 6, pp. 146–174.

Neuman, S., Oaxaca, R.L., 1998. Estimating Labor Market Discrimination with Selectivity Corrected Wage Equations: Methodological Considerations and an Illustration from Israel. Centre for Economic Policy Research, London.

Neumark, D., 1988. Employers' discriminatory behavior and the estimation of wage discrimination. Journal of Human Resources 23, 279–295.

Oaxaca, R.L., 1973. Male–female wage differentials in urban labor markets. International Economic Review 14, 693–709.

Oaxaca, R.L., Ransom, M., 1994. On discrimination and the decomposition of wage differentials. Journal of Econometrics 61, 5–21.

Oaxaca, R.L., Ransom, M., 1999. Using Econometric Models for Intrafirm Equity Salary Adjustments (unpublished paper). University of Arizona, Tucson, AZ and Brigham Young University, Provo, UT.

Peoples, J., 1994. Monopolistic market structure, unionization, and racial wage differentials. Review of Economics and Statistics 76, 207–211.

Peoples, J., Saunders, L., 1993. Trucking deregulation and the black/white wage gap. Industrial & Labor Relations Review 47, 23–35.

Psacharopolulos, G., Patrinos, H. (Eds.), 1994. Indigenous People and Poverty in Latin America: An Empirical Analysis. The World Bank, Washington, DC.

Reimers, C., 1983. Labor market discrimination against hispanic and black men. The Review of Economics and Statistics 65, 570–579.

Robinson, J., 1933. The Economics of Imperfect Competition. Macmillan, London.

Rosholm, M., Smith, N., 1996. The Danish gender wage gap in the 1980s: a panel data study. Oxford Economic Papers 48, 254–279.

Smith, J.P., Welch, F., 1989. Black economic progress after Myrdal. Journal of Economic Literature 27, 519–564.

Thurow, L., 1969. Poverty and Discrimination. Brookings Institution, Washington, DC.

Zabalza, A., Tzannatos, Z., 1985. Women and Equal Pay: The Effects of Legislation on Female Employment and Wages in Britain. Cambridge University Press, Cambridge, UK.

Discrimination: Racial

Thomas F Pettigrew, University of California, Santa Cruz, CA, USA
Marylee C Taylor, Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Racial discrimination exists in straightforward actions but also in complex systems of social relations that produce racial inequities in social outcomes. Some analysts view all racial disparities as the result of discrimination; others restrict the concept to those acts intended to limit a group's resources. The first view is too broad, the second too restrictive.

Social scientists have made a useful distinction between direct and indirect discrimination. The direct form occurs at points where inequality is generated, usually intentionally. The indirect form develops when the inequitable results of direct discrimination influence later decisions; it perpetuates and magnifies the original injury. After a brief review of racial discrimination and its effects and remedies in North America and Western Europe, two summary propositions are advanced. First, discrimination is normative. Its structural web outlives its initiators and may not reflect current attitudes. Second, discrimination feeds upon itself; it is typically cumulative and self-perpetuating. Effective remedies must intervene and reverse its 'vicious circle.'

Racial discrimination is not, as popularly thought, simply the accumulation of individual acts of unfairness between members of different groups. It consists as well of an elaborate web of institutional arrangements that produces group inequalities, a web that becomes racial discrimination when the groups perceive each other as separate 'races.' Studied throughout the social sciences, these inequalities can involve political power, economic resources, and cultural access.

Conceptual Complexities

Viewing racial discrimination as an institutional web casts a wide net but leaves room for considerable ambiguity. Disagreement about what constitutes discrimination stems from two sources – one ideological and political, the other empirical. First, because discrimination violates law and professed values in many countries, a judgment that unequal outcomes reflect discrimination is a call for unpopular change and costly remedies. Resistance to such change and costs activates debate about the extent of discrimination. Second, deficiencies in analysis and evidence limit the ability of social scientists to trace precisely the dynamic system of effects triggered by discrimination.

Consider first the political issue. The broadest definitions of discrimination assume that racial minorities have no inherent characteristics that warrant inferior social outcomes. Thus, all inequality becomes a legacy of discrimination and a social injustice to be remedied. Favored by the political left, this view appears to many to be too sweeping.

By contrast, political conservatives favor a narrow definition that limits the concept's scope by including only actions intended to restrict a group's chances. Most social-science specialists resist this reformulation for several reasons. An intentionality criterion returns the concept to the realm of psychology and deflects attention from restraining social structure. And the invisibility of intentions creates formidable obstacles for substantiating discrimination. Finally, the effects of discrimination are the same, whether or not the causal institutional mechanisms were intended to produce group inequalities.

This rejection of the intentionality criterion leads to an important distinction between direct and indirect discrimination (Pettigrew, 1985). Direct racial discrimination occurs at points where inequality is generated, often intentionally. When decisions are based explicitly on race, discrimination is direct. Indirect discrimination occurs when the inequitable results of direct discrimination serve as a basis for later decisions. Hence, discrimination is indirect when an ostensibly nonracial criterion becomes a proxy for race in determining social outcomes. Indirect discrimination perpetuates and magnifies the original injury.

Wage discrimination illustrates the distinction. Direct racial discrimination exists when employers pay equally qualified members of two races different rates for the same work. Indirect discrimination exists when the employers pay the two groups unequally because prior discrimination in employment, education, or housing created apparent differences in qualifications or channeled the groups into better- and worse-paying jobs. This direct/indirect distinction resembles the distinction often made by legal systems between disparate treatment and disparate impact. Although intentional, direct discrimination may have triggered the causal chain, the original injury is perpetuated and magnified by largely unwitting accomplices. To be sure, these accomplices typically benefit from the discriminatory arrangements and have little incentive to alter them. But the important point is that these beneficiaries of the discriminatory practices did not initiate them and may not even be aware of their operation.

The direct/indirect distinction is reflected in the psychological differentiation between conscious and unconscious discrimination. Those who are conscious of their abetting racial discrimination are typically racially prejudiced. They explicitly intend to harm and regard it as justified. But most people who engage in racial discrimination are largely unaware that they are doing so. They are simply following their in-group norms and participating in the subtle forms of indirect discrimination. Social psychologists view this as in-group favorability without

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32044-X

explicit intent to harm the out-group. Thus, the link between racial prejudice and racial discrimination is not nearly as firm as is commonly believed.

Empirical Ambiguities

This conceptual and psychological complexity carries over to impede the empirical study of discrimination. Apparently rigorous quantitative analyses often camouflage the crucial issues, as an examination of wage decomposition research reveals.

Assessments of racial discrimination produced by decomposing gross racial differences in social outcomes are common in sociology and economics. This approach first assigns one segment of the gross racial differential to 'qualifications' and other factors deemed legitimate determinants of social rewards. The residual segment not demonstrably linked to these 'legitimate' determinants of the outcomes often is presented as the estimate of discrimination. However, without better information than is usually available and without closer agreement about what constitutes discrimination, no unique estimate of discrimination is possible. The choice of control variables to index 'legitimate' determinants of social outcomes shapes the answers. Any appearance of scientific certitude is an illusion. It is not surprising that discrimination estimates originating from this approach vary enormously.

Empirical research in group discrimination must mirror the phenomenon in its variety and complexity. The regression decomposition approach is useful but limited. Regression analyses could provide more pertinent information if based on more homogeneous job groups (Conway and Roberts, 1994) and on structural equation models that test reciprocal causation. Most important, if the aim is to guide policy, a framework far more complex than the dichotomous discrimination-or-not approach is required. Research that traces the actual processes of institutional discrimination is essential. Also needed is more attention to victims' perceptions of discrimination and to the changes generated by antidiscrimination efforts.

Racial Discrimination in the United States and Western Europe

Both direct and indirect forms of racial discrimination can be found throughout the world. Most of the available research, however, focuses on the long-festering racial problems of the United States (Pager and Shepherd, 2008). In such long-term situations, injuries of indirect discrimination often are more extensive than those of direct discrimination. This conclusion does not imply that direct discrimination no longer exists. Employment complaint records reveal the continued operation of direct forms of discrimination. US women and minorities have filed almost two million job-discrimination complaints since 1965. Not all such complaints reflect genuine discrimination, but analysts believe that false complaints are minimal in comparison with the amount of actual discrimination that goes unreported and thus undetected (Tomaskovic-Devey and Stainback, 2007). Many major corporations have been found guilty of direct racial discrimination in recent years. And employment audits using paired, equally qualified applicants reveal widespread direct discrimination. Employers admit to interviewers that their stereotypes of minorities, particularly of inner-city black men, lead them to avoid minority hiring (Reskin, 1998).

Along with direct employment discrimination, indirect discrimination is pervasive. Many employers rely on word-of-mouth recruiting of new employees, especially for unskilled jobs. Thus the current workforce, often disproportionately white, reproduces existing racial inequalities in employment by bringing in white friends and relatives as the next generation of workers. This is just one aspect of 'business-as-usual' that represents indirect discrimination. Although word-of-mouth recruiting appears to be color blind, it serves to perpetuate racial inequalities (Reskin, 1998).

Note, too, how this structural process of indirect discrimination can operate without explicit racial prejudice being involved. Racial discrimination is often viewed as a psychological rather then a structural, phenomenon. Prejudiced members of the dominant race are thought to cause the racial unfairness intentionally. But white American workers nominating white friends for job openings simply reflects the society's widespread racial segregation; African Americans are not in their friendship networks. White nominators need not have hostile views of blacks to engage in this virulent form of indirect discrimination.

Important analyses of indirect discrimination in the United States focus on wealth. Racial inequalities in assets dwarf disparities in other social outcomes. Although the magnitude of inequality in home equity in severe, racial inequality in liquid assets is even more marked. The average white American family was estimated in 2007 to hold liquid assets 20 times greater than those of the average African American family – roughly $100 000 compared with only $5000. Moreover, this racial gap in assets has been accelerating since the 1980s (Shapiro et al., 2010).

This racial difference in assets results from a history of direct and indirect discrimination. Prohibitions against black self-employment and land ownership in the American South were an early part of the picture. More recently, occupations to which African Americans were confined, agricultural and domestic work, were excluded from Social Security provisions. Direct and indirect discrimination by federal housing agencies, banks, realtors, and white homeowners have limited the opportunity of African Americans to accrue wealth through home ownership (Conley, 1999; Rank, 2009; Shapiro, 2004).

The sharp economic recession that began in 2008 and centered on housing made matters worse. African and Hispanic Americans were among the hardest hit in home foreclosures and other asset-reducing events.

The accumulation of disadvantage is particularly acute when the focus is assets. Because wealth reproduces itself exponentially, the racial gap in assets in the United States has continued to grow. And assets have important implications for such other social outcomes as education and labor market success. When parents do not have the means to live in higher status neighborhoods, children do not have access to the best public schools. And for most students, attendance at the best colleges and universities increasingly depends on

parents' ability to provide thousands of dollars per year to cover rising tuition costs and expenses. Labor market outcomes reflect educational disadvantage, but restricted assets have other indirect effects as well. Access to networks of information crucial to job seekers, proximity to lucrative employment, automobile transportation to work – all these are functions of assets, and African Americans are at a pronounced disadvantage (Conley, 1999). Innovative strategies for intervention must be devised to break into the spiraling disadvantage produced by African Americans' deficits in wealth.

Analogous dynamics afflict the minorities of other nations, and research literature concerning the subject has emerged in Western Europe. Discrimination against the region's new immigrant racial minorities is pervasive (Pettigrew, 1998). A third of all minority respondents in a Europe-wide survey reported they had been personally discriminated against, and 12% felt they had been a victim of a racially motivated crime in the past 12 months (European Union Fundamental Rights Agency, 2010). The Roma and black Africans are particular targets. Survey results in France show an even sharper degree of perceived discrimination; 61% of minority respondents believed, in 2007, that they had experienced discrimination in the past year (BBC News, 2007).

Both direct and indirect discrimination are involved, though the indirect forms are largely unrecognized in Europe. Investigators have repeatedly uncovered direct discrimination. Controlled tests have revealed the full litany of discriminatory forms involving employment, public accommodations, housing, the courts, police, insurance, and banks. Police stops and searches in the United Kingdom, for example, indicated that per 1000 population in 2009 and 2010, blacks were seven times more likely to be stopped and searched by police than were whites (UK Justice, 2012).

Employment discrimination poses the most serious problem. In every European Union nation, minorities have markedly higher unemployment rates than native majorities. As in the United States, there are many reasons for disparities in minority unemployment. The 'last-in, first-out' principle selectively affects the younger minority workers. Typically less skilled, they are more affected by job upgrading. Minorities also are more likely to be in older, declining industries. These patterns are not accidental. Planners put minorities into these industries for cheaper labor precisely because of their decline. Yet these multiple factors offer insufficient explanations for the greater unemployment of minorities. Veenman and Roelandt (1990) found that education, age, sex, region, and employment level explained only a small portion of the differential unemployment rates in the Netherlands.

Indirect discrimination arises from the inability to obtain citizenship. It restricts immigrants' ability to get suitable housing, employment, and schooling. In short, the lives of Europe's noncitizens are severely circumscribed. Castles (1984) contends that the guest-worker system that brought many of the immigrants to Europe was itself a state-controlled system of institutional discrimination. It established the newcomers as a stigmatized out-group suitable for low-status jobs but not citizenship, and racial prejudices add to this stigma. Widespread indirect discrimination, then, was inevitable for these victims of direct discrimination.

Psychological Effects of Racial Discrimination

Social psychologists have examined extensively the effects of discrimination on its victims. Multiple studies show that when people perceive themselves to be the targets of systematic discrimination, their physical and psychological health often suffers. Moreover, perceived discrimination both increases the likelihood of engaging in dangerous and unhealthful behaviors and decreases the likelihood of performing healthful behaviors (Pacoe and Richman, 2009). These effects are mediated largely by increased stress, and they have been found in a variety of ethnic groups, including whites. Greater drug use among those perceiving discrimination is part of this process (Gibbons et al., 2004).

Discrimination also heightens victims' awareness of any cultural stigmatization of their group. Such 'stereotype threat' can be stressful and hinder performance. It is triggered by situations that contain cues that make a negative stereotype salient. This situation raises the strong possibility that one's own performance will conform to the stereotype and confirm it. Such a threat creates anxiety and undermines performance. Thus, African Americans tend to score higher on intelligence tests when they take the test apart from white Americans. Stereotype threat is not unique to racial minorities; women tend to score higher on tests of mathematics when there are no men in the situation (Steele, 2010).

Remedies for Racial Discrimination

The most elaborate array of remedial measures for racial discrimination is found in North America. Both Canada and the United States have instituted a battery of laws against many types of racial and other forms of discrimination. Court rulings in the two nations also have proven decisive. The effects of these legal actions are magnified by class action suits – the ability of single litigants to sue for remedy for their entire 'class' or group. Further remedies have involved various forms of 'affirmative action' in which special efforts are made to close group disparities caused by discrimination. These approaches have led to narrowing racial disparities in social outcomes. But sharp racial inequities remain in both countries, and majority resistance has hardened against the most successful remedies.

Antidiscrimination remedies have been largely ineffective in Europe. Germany guarantees basic rights only to citizens, so the disadvantages of noncitizenship include the inability to combat discrimination. There is extensive German legislation to combat anti-Semitism and Nazi ideology, but these laws have proven ineffective in protecting noncitizens. The German constitution explicitly forbids discrimination on the basis of origin, race, language, beliefs, or religion – but not citizenship. Indeed, the Federal Constitutional Court has ruled that differential treatment based on citizenship is constitutional if there is a reasonable basis for it and if it is not wholly arbitrary. In 2006, Germany, under pressure from the European Union, passed the General Equal Treatment Act (Allgemeines Gleichbehandlungsgesetz – 'AGG'). It prohibits discrimination in the labor market in the access to public services, such as social welfare benefits and housing markets, based on race,

ethnic origin, religion or beliefs, sex, disability, age, or sexual orientation.

Although the AGG does not place the burden of proof on plaintiffs, they must submit the initial evidence to support their claim that discrimination has occurred; then the accused party must prove otherwise. Yet it remains difficult for the plaintiffs to claim their rights. Usually, the victims of discrimination do not know their full rights, and consultancy services are rarely available. It is not surprising that few cases have been brought under the law and fewer still have been successful.

Effective means of combating discrimination are also rare in France. Commentators often view discrimination as 'natural' and as being universally triggered when a 'threshold of tolerance' (*seuil de tolerance*) is surpassed. Without supporting evidence, this rationalization supports quotas and dispersal policies that restrict minority access to suitable housing.

The Netherlands, the United Kingdom, and Sweden have enacted antidiscrimination legislation that applies specifically to the new immigrant minorities. And the Dutch have instituted modest affirmative action programs for women and minorities. Yet this legislation has been largely ineffective for two interrelated reasons. First, European legal systems do not allow class action suits – a forceful North American weapon to combat discrimination. Second, European efforts rely heavily on individual complaints rather than on systemic remedies. The UK's 1976 Act gave the Commission for Racial Equality power to cast a broader net, but individual complaints remain the chief tool (MacEwen, 1995).

Individual efforts are unlikely to alter discrimination because they are nonstrategic. Minorities bring few charges against the worst discriminatory firms because they avoid applying to them. Complaints about job promotion are common, but they are made against employers who hire minorities. Effective antidiscrimination laws must provide broad powers to an enforcement agency to initiate strategic, institution-wide actions that uproot the structural foundations of discrimination.

Two Summary Propositions

1. The long-lasting character of racial discrimination means that the effects typically outlive the initiators of discriminatory practices. Discrimination is fundamentally normative; its structural web operates in large part independent of the dominant group's current attitudes or awareness. Hence, models based primarily on individual prejudice or 'rationality,' whether psychological or economic, will uniformly understate and oversimplify the phenomenon.
2. Discrimination feeds upon itself; it is typically cumulative and self-perpetuating. An array of research on African Americans has shown that neighborhood racial segregation leads to educational disadvantages, then to occupational disadvantages, and thus to income deficits (Pettigrew, 1985). To be effective, structural remedies must reverse this vicious circle of discrimination.

Seen in this perspective, discrimination is far more intricate and entrenched than is commonly thought. The complexity of discrimination presents major challenges to social-scientific attempts to trace its impact. This complexity also precludes any simple correspondence between perpetration and responsibility for remedy. Broad social programs will be necessary if the full legacy of direct and indirect racial discriminations is finally to be erased.

See also: Ethnic Conflicts; Ethnicity and Ethnic Groups: Historical Aspects; Race Identity; Racial Relations; Racism, Sociology of.

Bibliography

BBC News, 2007. First French Racism Poll Released. Retrieved 6 July 2012 from http://news.bbc.co.uk/go/pr/fr/-/2/hi/europe/6317799.stm.

Castles, S., 1984. Here for Good: Western Europe's New Ethnic Minorities. Pluto Press, London.

Conley, D., 1999. Being Black, Living in the Red: Race, Wealth, and Social Policy in America. University of California Press, Berkeley, CA.

Conway, D.A., Roberts, H.V., 1994. Analysis of employment discrimination through homogeneous job groups. Journal of Econometrics 61, 103–131.

European Union Fundamental Rights Agency, 2010. EU-MIDIS at a Glance: Introduction to the FRA's EU-wide Discrimination Survey. Fundamental Rights Agency, Strasbourg, France.

Gibbons, F.X., Gerrard, M., Cleveland, M.J., Wills, T.A., Brody, G., 2004. Perceived discrimination and substance abuse in African American parents and their children: a panel study. Journal of Personality and Social Psychology 86 (4), 517–529.

MacEwen, M., 1995. Tackling Racism in Europe: An Examination of Anti-discrimination Law in Practice. Berg, Washington, DC.

Pacoe, E.A., Richman, L.S., 2009. Perceived discrimination and health: a meta-analytic review. Psychology Bulletin 135 (4), 531–534.

Pager, D., Shepherd, H., 2008. The sociology of discrimination: racial discrimination in employment, housing, credit, and consumer markets. Annual Review of Sociology 34, 181–209.

Pettigrew, T.F., 1985. New black–white patterns: how best to conceptualize them? Annual Review of Sociology 11, 329–346.

Pettigrew, T.F., 1998. Reactions toward the new minorities of Western Europe. Annual Review of Sociology 24, 77–103.

Rank, M.R., 2009. Measuring the economic racial divide across the course of American lives. Race and Social Problems 1 (2), 57–66.

Reskin, B.F., 1998. The Realities of Affirmative Action in Employment. American Sociological Association, Washington, DC.

Shapiro, T.M., 2004. The Hidden Cost of Being African American: How Wealth Perpetuates Inequality. Oxford University Press, New York.

Shapiro, T.M., Meschede, T., Sullivan, L., 2010. The racial gap increases fourfold. Institute on Assets and Social Policy Research and Policy Brief. pp. 1–2.

Steele, C.M., 2010. Whistling Vivaldi: And Other Clues to How Stereotypes Affect Us. Norton, New York.

Tomaskovic-Devey, D., Stainback, K., 2007. Discrimination and desegregation: equal opportunity progress in U.S. private sector workplaces since the Civil Rights Act. The Annals of the American Academy of Political and Social Science 609, 49–84.

United Kingdom Justice, July 2012. Race and the Criminal Justice System. Retrieved 6 July 2012 from htpp://www.justice.gov.uk/statistics/criminal-justice/race.

Veenman, J., Roelandt, T., 1990. Allochtonen: achterstand en achterstelling. In: Schippers, J.J. (Ed.), Arbeidsmarkt en Maatschappelijke Ongelijkheid. Wolters-Noordhoff, Groningen, The Netherlands, pp. 241–261.

Disgust, Psychology of

Paul Rozin, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P. Rozin, volume 6, pp. 3766–3769, © 2001, Elsevier Ltd.

Abstract

Disgust appears to be a universal and basic emotion, perhaps unique to humans. It has become a center of interest in recent research, including psychological, philosophical, evolutionary, and neural perspectives. On the output side, it is characterized by a withdrawal response, a physiological concomitant of nausea, and a distinct facial expression, focused around the nose and mouth. This output system seems to have been relatively constant across time and cultures. However, there has been a major elaboration of elicitors and meanings in cultural evolution, perhaps by a process of preadaptation. Disgust may have originated as a food rejection response in animals, based on taste or smell. However, in humans, the food rejection response is transformed into a rejection based upon the nature or origin of a food, rather than its sensory properties. Disgusting potential foods have contaminating properties; when they contact an acceptable food, they render that food unacceptable. The offensiveness of certain foods expands to the offensiveness of any reminders of the animal nature of humans, especially death and its odor of decay. Further expansions include interpersonal disgust; direct or indirect contact with undesirable people or simply strangers, and moral disgust, directed at certain types of moral violations. An alternative recent view categorizes disgust elicitors, based on evolutionary considerations, into pathogen avoidance, avoidance of certain sexual contacts, and avoidance of a wide range of moral violations. The elicitors and meanings of disgust vary across time and culture, with interpersonal and moral disgust expressed particularly prominently in Hindu India. Disgust has been transformed, in human cultural evolution, into an emotion that promotes withdrawal from a wide range of culturally unacceptable entities; it is a major force in socialization, and may have become a guardian of the 'soul,' as well as the body.

Disgust appears as a fundamental emotion in sources as different as ancient Hindu texts and Darwin's (1965/1872) *The Expression of the Emotions in Man and Animals*. Although usually listed as one of six basic emotions in modern psychology texts, disgust was virtually unexplored until the last 20 years. In the last 20 years, disgust has received much attention in psychology and other disciplines; there were only a few books ever written about disgust before 2010; at least five have appeared since 2010 (note particularly Herz, 2011; Kelly, 2011).

Darwin defined disgust as referring to "something revolting, primarily in relation to the sense of taste, as actually perceived or vividly imagined; and secondarily to anything that causes a similar feeling, through the sense of smell, touch and even of eyesight" (p. 253). In the classic psychoanalytic treatment of disgust, Andras Angyal (1941) claimed that "disgust is a specific reaction toward the waste products of the human and animal body" (p. 395). Angyal related the strength of disgust to the degree of intimacy of contact, with the mouth as the most sensitive focus. Sylvan Tomkins (1963) has offered the more general contention that disgust is "recruited to defend the self against psychic incorporation or any increase in intimacy with a repellent object" (p. 233). Derivative of Angyal's definition, Rozin and Fallon (1987, p. 23) expand on Angyal's definition, and define disgust as: "Revulsion at the prospect of (oral) incorporation of an offensive object. The offensive objects are contaminants; that is, if they even briefly contact an acceptable food, they tend to render that food unacceptable." Disgust is related to pollution, as used in the anthropological literature. In particular, Mary Douglas' (1966) view of polluting meaning matter out of place has direct relevance to disgust.

Most scholars recognize a special relation between disgust and ingestion: the facial expression of disgust represents an oral and nasal 'rejection,' the word 'disgust' in English and some other languages means 'bad taste,' and nausea, the physiological signature of disgust, is directly related to, and inhibitory of, ingestion.

However, it is also clear that, cross-culturally, the emotion of disgust is elicited by a wide range of elicitors, most of which have nothing to do with food. The history and trajectory of disgust, from a food-related emotion to a much more general emotion, is the central concern of this article.

Disgust as an Emotion

Disgust is the only basic emotion that is specifically related to a motivation (hunger) and to a particular bodily location (mouth). Along with other basic emotions, as described by Paul Ekman (1992), disgust has a characteristic behavior (withdrawal), physiological state (nausea), and facial expression (see also Izard, 1971). The facial expression centers on the mouth and nose, and is characterized principally by a wrinkling of the nose, a raising of the upper lip, and a gape. There are also characteristic bodily gestures. The mental or feeling component of disgust is often described as revulsion.

The Meaning of Disgust

Rozin et al. (1993, 2008), based on surveys in Japan and the United States, designate nine domains of disgust: certain foods or potential foods, body products, animals, sexual behaviors, contact with death or corpses, violations of the exterior envelope of the body (including gore and deformity), poor hygiene, interpersonal contamination (direct or indirect contact with unsavory human beings), and certain moral offenses. A more recent formulation by Tybur and colleagues, based in part on

evolutionary considerations, organizes the wide range of disgust elicitors into three groups, pathogen avoidance (which includes potentially all of the previously listed elicitors except sexual and moral), sexual violations, and moral violations (Tybur et al., 2013). Disgust has to do with offensive things; these are wide in range, but exclude many other types of negative events, such as pain, and those that elicit the emotions of fear, sadness, and anger.

Rozin et al. (1993, 2008) propose that disgust originated in animals as a response to distasteful food, and others have added to this origin a set of adaptations to avoidance of pathogens. These authors hold that in cultural evolution, the output side of disgust (expression, physiology, behavior) remained relatively constant, but the range of elicitors and meanings expanded dramatically. The meaning of disgust became offensiveness to the self or soul, rather than bad taste, its origin. Indeed, it was proposed that, in humans, distasteful items do not elicit the full emotion of disgust. The expansion of elicitors and meanings in cultural evolution was attributed to the process of preadaptation; use of a mechanism evolved in one system for a novel use in another system. For example, the mouth, and in particular the teeth and tongue, evolved for processing food, but these structures were co-opted in human evolution for use in speech articulation. The expansion of the meanings of disgust is proposed to have occurred in four stages.

Core Disgust

Core disgust, the initial form of disgust, and the form that appears first in children, has to do with food, but primarily the nature and origin of the food, rather than its taste (Rozin and Fallon, 1987). Core disgust has three components: (1) a sense of oral incorporation (and hence a linkage with food or eating); (2) a sense of offensiveness; and (3) contamination potency. The mouth is the principal route of entry of material things into the body; such things may be essential for life, but may also be very harmful. Disgust has to do with the harmful side, the risks of ingestion. The threat of oral incorporation is illustrated by a widespread belief that "You are what you eat." As others (e.g., Tybur et al., 2013; Oaten et al., 2009; Schaller and Park, 2011) have noted, many of the features of core (and much of animal reminder) disgust can be understood as adaptations to avoidance of pathogens.

Almost all elicitors of core disgust are potential foods that are whole animals, or animal body parts or body products. The great majority of food taboos concern animals and their products, and in most cultures the great majority of animals as potential foods are considered disgusting.

The contamination response (e.g., the rejection of a potential food if it even briefly contacted a disgusting entity) appears to be powerful and universal among adults, and is a defining feature of disgust. It is an example of the sympathetic magical law of contagion, which essentially holds that "once in contact, always in contact" (Rozin and Nemeroff, 1990). It has an obvious adaptive value as part of a pathogen avoidance system.

Animal Reminder Disgust

Common elicitors of disgust include inappropriate sex, poor hygiene, death, and violations of the ideal body envelope (e.g., gore, deformity, obesity). Contact with death and the odor of death (decay) are potent elicitors of disgust. Individuals who score high on disgust sensitivity also score higher on a fear of death scale.

The centrality of death in disgust suggests a more general construal of disgust. Disgust may be understood as a defense against a universal fear of death. Ernst Becker has argued that the most important threat to the psyche is the certainty of death. Human beings are the only organisms that know they will die. Disgust, as a withdrawal system, may help to suppress thoughts or experiences that suggest human mortality.

More generally, anything that reminds us that we are animals may disgust us (Rozin and Fallon, 1987). Disgust serves to 'humanize' our animal bodies. Eating, excretion, sex, soiling of the body surface, soft viscera within the body envelope, and death are all properties that humans share with animals. Cultures typically have extensive rules about these matters, including a wide range of prohibitions. People who ignore these prescriptions are reviled as disgusting and animal-like. This conception of disgust links it to the concern of humans, frequently cited by anthropologists, to distinguish themselves from animals. It has recently been suggested that the wide range of core and animal-reminder elicitors (and interpersonal disgust) can be grouped into one larger category, as a pathogen avoidance system (Tybur et al., 2013).

Interpersonal and Moral Disgust

Across cultures, there is a sense of disgust at touching most people who are not intimates, of making indirect contact with them, such as by wearing their clothing. Most other persons are contaminating, in this sense, and this sensitivity can serve to reinforce social boundaries (or as a pathogen avoidance mechanism). Cultures differ in degree of this sensitivity; it is particularly high in Hindu India, where it serves as an important means of maintaining caste distinctions.

In many, if not all languages, the word for 'disgust' is applied to certain moral violations, and the contamination property of disgust is also extended to the perpetrators of such immoral acts. It is possible that disgust functions, along with anger and contempt, as a basic moral emotion involved in expressing moral outrage at the actions of another. Richard Shweder et al. (1997) identify three types of moral codes that occur in cultures around the world. Violation of the community code, having to do with issues of hierarchy and duty, may map on to the emotion of contempt. Violations of the autonomy code deal with rights and justice, and may be associated with anger. The third code, divinity, focuses on the self as a spiritual entity and seeks to protect that entity from degrading or polluting acts. Disgust may be the emotion that is evoked when the divinity code is violated (Rozin et al., 1999). A more modern formulation of disgust categorizes the interpersonal and moral elicitors of disgust into two categories, sexual (mate selection) and generally moral (including potentially all moral violations) (Tybur et al., 2013). There is certainly evidence that Americans apply the word 'disgusting' to a wide range of moral violations; the question is whether this is a metaphoric and very general use of the word, or implies a specific link between disgust and a wide range of moral violations.

Development of Disgust

Distaste, the origin of disgust, is present at birth, as in the innate rejection by neonates of bitter substances. Disgust, with contamination sensitivity as a defining feature, first appears in North American children at about 4 or 5 years of age. Disgust, as opposed to distaste, seems to be uniquely human, and is entirely or primarily a product of culture. The contamination response, also uniquely human, is probably, in origin, a means of protection from microbial contamination; however, it is acquired. Even feces, surely a source of such contamination and a universal disgust substance among human adults, are not rejected by infants. Toilet training may well be the first disgust-generating experience for humans, but the means by which strong negative affect becomes associated with feces is unknown, as is the way that disgust 'spreads' to a wider range of elicitors. Presumably, parental reactions to disgusting things play a central role in the development of disgust. There is evidence for family resemblance in disgust sensitivity. Although the evidence is not yet strong, it appears that core (or pathogen) disgust appears before the other types of disgust.

Disgust, History and Culture

Across cultures and history, the expression of disgust seems to vary rather little, while the elicitors and meanings vary widely. It seems that disgust was relatively infrequent and limited to a narrow range of elicitors in medieval Europe. The full scope of disgust in Western history, and its flowering as both an emotion of sensibility to the animal nature of humans and a marker of immorality is presented in William Miller's (1997) *Anatomy of Disgust*.

Many cultures show specific attractions to entities that most other cultures find disgusting. For example, particular spoiled foods (e.g., cheese for the French, fermented fish sauce for Southeast Asians) are highly desired in particular cultures, and there are major differences in the sexual practices and contact with the dead across cultures.

The largest historical and cultural variation in disgust is probably associated with interpersonal and moral disgust. For example, disgust seems to center more on interpersonal and moral matters for Hindu Indians than it does for Americans.

Individual Differences

Results from a standardized test of disgust sensitivity, applied in Japan and the United States, indicate that sensitivity is somewhat higher in females and people of lesser education or lower social class (Haidt et al., 1994; Olatunji et al., 2007). A more recent measure of individual differences in disgust has been developed by Tybur et al. (2009). As well, disgust has been implicated in two clinical phenomena: phobias and obsessive compulsive disorder.

Neural Correlates of Disgust

There has been extensive recent work on the neural basis or correlates of disgust. Evidence suggests that an area activated when disgust is experienced is the insula, the area with linkages to the olfactory system. There is also some evidence for common activation of brain areas in response to elicitors as different as core (pathogen-linked) entities and moral violations.

Disgust and Related Emotions

Fear and disgust share a behavioral component, withdrawal based on a threat. Anger and sociomoral disgust share the appraisal of moral condemnation and the common facial gesture of raised upper lip. Contempt shares the appraisal that someone is base and inferior to oneself, but has a complex and rich set of connotations. Shame has been described as disgust turned inward, a judgment that the self is disgusting. It is not clear what emotion is 'opposite' to disgust, but perhaps love is the closest opposite.

Disgust and the Arts

Disgusting images appear frequently in visual art. Disgust is also a focus of humor, at least in Western culture, and plays a particularly central role in children's humor. Disgust appears to be attractive in situations where it is contained and not too threatening; in this regard, our enjoyment of disgust may be related to our enjoyment of horror movies, sad movies, and amusement park rides. All of these proclivities seem to be uniquely human. In general, disgust plays a major role in socialization and the sensibilities it promotes are at the heart of what it means to be human and to be civilized.

See also: Affect-Regulation Motivation; Avoidance and Approach Motivation: A Brief History; Control Behavior: Psychological Perspectives; Decision Making: Nonrational Theories; Emotion and Expression; Flow in Motivational Psychology; Gender and Academic Motivation; Health Self-Regulation, Motivational and Volitional Aspects of; Interest, Psychology of; Motivation and Actions, Psychology of; Motivation in Australian Aboriginal Populations; Optimism and Pessimism; Passion and Motivation; Personal Projects; Personality and Adaptive Behaviors; Self and Emotional Development in Adulthood and Later Life; Self-Regulation in Adulthood; Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model; Temperament and Motivation.

Bibliography

Angyal, A., 1941. Disgust and related aversions. Journal of Abnormal & Social Psychology 36, 393–412.
Becker, E., 1973. The Denial of Death. Free Press, New York.
Curtis, V., Biran, A., 2001. Dirt, disgust, and disease: is hygiene in our genes? Perspectives in Biology and Medicine 44, 17–31.
Darwin, C.R., 1965. The Expression of the Emotions in Man and Animals. University of Chicago Press, Chicago (Original work published 1872).
Douglas, M., 1966. Purity and Danger. Routledge & Kegan Paul, London.
Ekman, P., 1992. An argument for basic emotions. Cognition & Emotion 6, 169–200.
Haidt, J., McCauley, C.R., Rozin, P., 1994. A scale to measure disgust sensitivity. Personality and Individual Differences 16, 701–713.

Herz, R., 2011. That's Disgusting. W.W. Norton, New York.
Izard, C.E., 1971. The Face of Emotion. Appleton-Century-Crofts, New York,.
Kelly, D., 2011. Yuck! The Nature and Moral Significance of Disgust. MIT University Press, Cambridge, MA.
Miller, W.I., 1997. The Anatomy of Disgust. Harvard University Press, Cambridge, MA.
Oaten, M., Stevenson, R.J., Case, T.I., 2009. Disgust as a disease-avoidance mechanism. Psychological Bulletin 135, 303–321.
Olatunji, B.O., Williams, N.L., Tolin, D.F., Sawchuk, C.N., Abramowitz, J.S., Lohr, J.M., Elwood, L., 2007. The disgust scale: item analysis, factor structure, and suggestions for refinement. Psychological Assessment 19, 281–297.
Rozin, P., Fallon, A.E., 1987. A perspective on disgust. Psychological Review 94, 23–41.
Rozin, P., Haidt, J., McCauley, C.R., 1993. Disgust. In: Lewis, M., Haviland, J. (Eds.), Handbook of Emotions. Guilford, New York, pp. 575–594.
Rozin, P., Haidt, J., McCauley, C.R., 2008. Disgust. In: Lewis, M., Haviland, J. (Eds.), Handbook of Emotions, third ed. Guilford, New York, pp. 757–776.
Rozin, P., Lowery, L., Imada, S., Haidt, J., 1999. The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). Journal of Personality and Social Psychology 76, 574–586.
Rozin, P., Nemeroff, C.J., 1990. The laws of sympathetic magic: a psychological analysis of similarity and contagion. In: Stigler, J., Herdt, G., Shweder, R.A. (Eds.), Cultural Psychology: Essays on Comparative Human Development. Cambridge University Press, Cambridge, England, pp. 205–232.
Schaller, M., Park, J.H., 2011. The behavioral immune system (and why it matters). Current Directions in Psychological Science 20, 99–103.
Shweder, R.A., Much, N.C., Mahapatra, M., Park, L., 1997. The "big three" of morality (autonomy, community, divinity), and the "big three" explanations of suffering. In: Brandt, A., Rozin, P. (Eds.), Morality and Health. Routledge, New York, pp. 119–169.
Tomkins, S.S., 1963. Affect Imagery Consciousness: Vol. II. The Negative Affects. Springer, New York.
Tybur, J.M., Lieberman, D., Griskevicius, V., 2009. Microbes, mating, and morality: individual differences in three functional domains of disgust. Journal of Personality and Social Psychology 97, 103–122.
Tybur, J.M., Lieberman, D., Kurzban, R., DeScioli, P., 2013. Disgust: evolved function and structure. Psychological Review 120, 65–84.

Disposable Youth in America in the Age of Neoliberalism

Henry A Giroux, McMaster University, Hamilton, ON, Canada; and Ryerson University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

With the rise of market fundamentalism, today's youth are facing a crisis unlike that of any other generation. Young people, especially poor minority youth, are no longer seen as a social investment but as a problem and, in some cases, 'disposable.' Caught between the discourses of consumerism and a powerful crime-control complex, young people are either viewed increasingly as commodities or are subjected to the dictates of an ever expanding criminal justice system. I explore the current conditions of young people and their everyday experiences within an emerging governing-through-crime complex, the neoliberal politics of disposability, and the ever-present market-driven forces of privatization and commodification. I focus on the contemporary United States as the prime example of a country that is disposing of its youth. I also raise some important questions regarding the role that youth, in particular, might play in challenging the what might be called the crisis of disposability in the age of late modernity, while deepening and extending the promise of an aspiring democracy.

Since the 1970s, there has been an intensification of the anti-democratic pressures of market-driven modes of governance, ideology, and policies around the world, but perhaps most vividly seen in the United States. While many 'crises of youth' have occurred across the sweep of time, what is new in this historical conjuncture is the way in which young people, particularly low-income and poor minority youth, are increasingly denied any place in an already weakened social contract and the degree to which they are no longer seen as central to how the United States defines its future. Youth are no longer the age-group in which American society reveals its dreams but is instead where it hides its nightmares. Within neoliberal narratives, youth are either defined as a consumer market or they stand for trouble and are 'disposable,' not worthy of investment (Grossberg, 2005; Giroux, 2009). This shift in representations of how American society talks about young people betrays a great deal about what is increasingly novel about the economic, social, cultural, and political constitution of American society and its growing disinvestment in young people, the social state, and democracy itself (Comaroff and Comaroff, 2006).

Disposability now defines the fate of those young people who are seen as flawed consumers, excess, or 'othered' as result of their class or race. Relegated to 'zones of abandonment' (Biehl, 2005) and social death, ranging from underfunded schools to the criminal justice system, youth have become a marker for a mode of disposability in which their fate is defined largely through the registers of a society that throws away resources, people, and goods. Within these zones of abandonment and social death, poor minority and low-income youth are viewed as out of step, place, and time and defined largely as 'pathologies feeding on the body politic' relegated to spheres of 'terminal exclusion' (Biehl, 2005: p. 14).

The promises of modernity regarding progress, freedom, and hope have not been eliminated; they have been reconfigured, stripped of their emancipatory potential, and relegated to the logic of a savage market instrumentality. Modernity has reneged on its promise, however disingenuous or limited, to young people of mobility, stability, and collective security. Long-term planning and the institutional structures that support them are now relegated to the imperatives of privatization, deregulation, flexibility, and short-term investments. Social bonds have given way under the collapse of social protections and the welfare state just as "the emphasis is now on individual solutions to socially produced problems" (Bauman, 2007: p.14). As Sharon Stephens points out, what we are now witnessing is not only the "wide-ranging restructurings of modernity" but also the effect "these changes have for the concept of childhood and the life conditions of children" (1995: p. 19).

Zygmunt Bauman argues that today's youth have been "cast in a condition of liminal drift, with no way of knowing whether it is transitory or permanent" (Bauman, 2004: p. 76). The severity of the consequences of this shift in modernity under neoliberalism among youth is evident in the fact that this is the first generation in which the "plight of the outcast may stretch to embrace a whole generation" (Bauman, 2012a; developed further in Bauman, 2012b, 2012c). That is, the generation of youth in the early twenty-first century has no way of grasping if they will ever "be free from the gnawing sense of the transience, indefiniteness, and provisional nature of any settlement" (Bauman, 2004: p. 76). Neoliberal violence produced in part through a massive shift in wealth to the upper 1%, growing inequality, the reign of the financial services, the closing down of educational opportunities, and the stripping of benefits and resources from those marginalized by race and class has produced a generation without jobs, an independent life, and even the most minimal social benefits.

Youth no longer occupy the hope of a privileged place that was offered to previous generations. They now inhabit a neoliberal notion of temporality marked by a loss of faith in progress along with the emergence of apocalyptic narratives in which the future appears indeterminate, bleak, and insecure. Heightened expectations and progressive visions pale and are smashed next to the normalization of market-driven government policies that wipe out pensions, eliminate quality health care, raise college tuition, and produce a harsh world of joblessness, while giving millions to banks and the military. Students, in particular, now find themselves in a world in which heightened expectations have been replaced

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64123-5

by dashed hopes and a world of onerous debt (Fraser, 2013). The promises of higher education and previously enviable credentials have turned into the swindle of fulfillment as "For the first time in living memory, the whole class of graduates faces a high probability, almost the certainty, of ad hoc, temporary, insecure and part-time jobs, unpaid 'trainee' pseudo-jobs deceitfully rebranded 'practices' – all considerably below the skills they have acquired and eons below the level of their expectations" (Bauman, 2012b: p. 47).

What has changed about an entire generation of young people includes not only neoliberal society's disinvestment in youth and the permanent fate of downward mobility, but also the fact that youth live in a commercially carpet-bombed and commodified environment that is unlike anything experienced by those of previous generations. Nothing has prepared this generation for the inhospitable and savage new world of commodification, privatization, joblessness, frustrated hopes, surveillance, and stillborn projects (Bauman, 2012b: p. 47). The present generation has been born into a throwaway society of consumers in which both goods and young people are increasingly objectified and disposable.

That is, young people are viewed as excess, a drain on the body politic, part of a throwaway society, and unworthy of long-term investments. In this discourse, young people are not troubled but trouble; rather than viewed as at risk, they are the risk and subject to a range of punitive policies. As Larry Grossberg points out, young people now inhabit a world in which there are few public spheres or social spaces autonomous from the reach of the market.

> Kids see themselves presented in highly sexualized ways for advertising and entertainment purposes, even while their own sexual behavior is repressed and criminalized. … kids hear themselves attacked for their consumerism, even while the media and corporate interests commercialize and commodify every dimension of their lives, every pleasure, every need. From the kids' perspective, it looks as if we are selling, or perhaps we already have sold, our children to advertisers, to private business, to corporate capitalism. … What can it say to them that as a society we are unwilling to pay for their education (as ours had been paid for) but we are willing to commercialize education (junk foods, field trips to chain stores, sponsored lesson materials and textbooks, advertising in hallways). (2005: pp. 38–39)

The structures of neoliberal modernity do more than disinvest in young people and commodify them, they also transform the protected space of childhood into a zone of disciplinary exclusion and cruelty, especially for those young people further marginalized by race and class who now inhabit a social landscape in which they are increasingly disparaged as flawed consumers. With no adequate role to play as consumers, many youth are now considered disposable, forced to inhabit 'zones of social abandonment' extending from bad schools to bulging detention centers and prisons (Biehl, 2005). These are zones where not only the needs of young people are ignored, but they are subjected to conditions of impoverishment and punishment that underserve them and often criminalize their behavior. For example, with the hollowing out of the social state and the rise of the punishing state, the circuits of state repression, surveillance, and disposability increasingly 'link the fate of Blacks, Latinos, Native Americans, poor Whites, and Asian-Americans' who are now caught in a governing through crime youth complex that now serves as default solutions to major social problems (Davis, 2007).

This radical neoliberal shift in modernity becomes evident as economic power succeeds in detaching itself from government regulations, social costs, and ethical considerations, a new global financial class reasserts the prerogatives of capital and systemically destroys those public spheres – including public and higher education – that traditionally advocated for social equality and an educated youth as the fundamental conditions for a viable democracy. Power is now global and politics is local, inhabited by eviscerated state governments and a weakened nation-state. Stripped of their power, state governments have few resources to devote to the genuine causes of people's hardships, suffering, and socially produced problems. Increasingly, they function as an adjunct of financial interests and exercise power through the varied apparatuses of the punishing state. Since September 2011, the paralyzing fog of depoliticization has been ruptured by the intermittent roar of angry workers, and of young people who refuse to cede their future to the new oligarchs, bankers, conservative think tanks, hedge fund managers, Christian extremists, and the corporate-controlled liberal and conservative media apparatuses.

Disposable Youth, Violence, and Corporate Power

Disposability has become not only a register for neoliberal domination but also a mark of refusal and growing resistance on the part of young people from New York City to Athens, Greece who remind us that the possibility of collective struggle and shared hopes is far from a lost moment in history (Hardt and Negri, 2012; Castells, 2012; Giroux, 2013a). Young people are demonstrating across the world against a variety of issues ranging from economic injustice and massive inequality to drastic cuts in education and public services. Refusing to inhabit the stateless position of disposable populations, young people are fighting back, but not without resistance on the part of established countries caught in the grip of neoliberal austerity measures. These demonstrations are being met with state-sanctioned violence and insults in the mainstream media rather than with informed dialogue, critical engagement, and reformed policies. In the United States, Chile, Brazil, Mexico, and many other countries the state monopoly on the use of violence has intensified since the 1980s and in the process has been increasingly directed against young people, poor minorities, immigrants, and increasingly women. As the welfare state is hollowed out, a culture of compassion is replaced by a culture of violence, cruelty, and disposability. Collective insurance policies and social protections have given way to the forces of economic deregulation, the transformation of the welfare state into punitive workfare programs, the privatization of public goods, and an appeal to individual responsibility as a substitute for civic responsibility. Under the notion that unregulated market-driven values and relations should shape every domain of human life, the business model of governance has eviscerated any viable notion of social responsibility while furthering the criminalization of social problems and cutbacks in basic social services, especially for

the poor, young people, and the elderly (Wacquant, 2009). Within the existing neoliberal historical conjuncture, there is a merging of violence and governance and the systemic disinvestment in and breakdown of institutions and public spheres that have provided the minimal conditions for democracy.

Abandoned by the existing political system, young people are placing their bodies on the line, protesting peacefully while trying to produce a new language, politics, long-term institutions, and "community that manifests the values of equality and mutual respect that they see missing in a world that is structured by neoliberal principles" (Bella, 2011). Such movements are not simply about reclaiming space but also about producing new ideas, generating a new conversation, and introducing a new political language. Rejecting the notion that democracy and markets are the same, young people are calling for an end to the corporate control of the commanding institutions of politics and culture, poverty, the suppression of dissent, and the permanent war state. Richard Lichtman is right in insisting that this movement should be praised for its embrace of communal democracy as well as an emerging set of shared concerns, principles, and values articulated "by a demand for equality, or, at the very least, for a significant lessening of the horrid extent of inequality; for a working democracy; for the elimination of the moneyed foundation of politics; for the abolition of political domination by a dehumanized plutocracy; for the replacement of ubiquitous commodification by the reciprocal recognition of humanity in the actions of its agents" (Lichtman, 2011).

While there is considerable coverage in the progressive media given to the violence being waged against the movement protesters in Brazil, the United States, Greece, and elsewhere, it is important to situate such violence within a broader set of categories that enables a critical understanding of not only the underlying social, economic, and political forces at work in such assaults, but also allows us to reflect critically on the distinctiveness of the current historical period in which they are taking place (Giroux, 2013). For example, it is difficult to address such state-sponsored violence against young people without analyzing the devolution of the social state, the politics of disposability, and the corresponding rise of the warfare and punishing state.

The merging of the military–industrial complex and unbridled corporate power in many countries points to the need for strategies that address what is specific about the current warfare state and the neoliberal project and how different interests, modes of power, social relations, public pedagogies, and economic configurations come together to shape its politics. Such a conjuncture is invaluable politically in that it provides a theoretical opening for making the practices of the warfare state and the neoliberal revolution visible in order 'to give the resistance to its onward march, content, focus, and a cutting edge' (Hall, 2011a: p. 706). It also points to the conceptual power of making clear that history remains an open horizon that cannot be dismissed through appeals to the end of history or end of ideology (Bell, 1966; Fukuyama, 2006). It is precisely through the indeterminate nature of history that resistance becomes possible and politics refuses any guarantees and remains open. Following Stuart Hall (2011), I argue that the current historical moment or what he calls the 'long march of the Neoliberal Revolution,' has to be understood in terms of the growing forms of violence that it deploys and reinforces. Such antidemocratic pressures and their relationship to the rising protests of young people in the United States and abroad are evident in the crisis that has emerged through the merging of governance and violence, the growth of the punishing state, and the persistent development of what has been described by Alex Honneth as "a failed sociality" (2009: p. 188).

Too many neoliberal societies, the United States in particular, have become addicted to violence. War provides jobs, profits, political payoffs, research funds, and forms of political and economic power that reach into every aspect of society. As war becomes a mode of sovereignty and rule, it erodes the distinction between war and peace. Increasingly fed by a moral and political hysteria, warlike values produce and endorse shared fears as the primary register of social relations. Shared fears and the media hysteria that feed them produce more than a culture of fear. Such hysteria also feeds the growing militarization of the police who increasingly use their high-tech scanners, surveillance cameras, and toxic chemicals on anyone who engages in peaceful protests against the warfare and corporate state. Images abound in the mainstream media of such abuses.

The war on terror has become a war on democracy as baton-wielding cops are now being supplied with the latest military equipment imported straight from the battlefields of Iraq and Afghanistan. Military technologies once used exclusively on the battlefield are now being supplied to police departments across the nation. Drones, machine-gun-equipped armored trucks, Special Weapons and Tactics (SWAT) vehicles, "digital communications equipment, and Kevlar helmets, like those used by soldiers used in foreign wars" (Becker and Schulz, 2011). The domestic war against 'terrorists' (code for young protesters) provides new opportunities for major defense contractors and corporations who 'are becoming more a part of our domestic lives' (Becker and Schulz, 2011). Arming domestic police forces with paramilitary weaponry will ensure their systematic use even in the absence of a terrorist attack on US soil; they will simply find other, increasingly permissive uses for those weapons (Greenwald, 2011).

As a mode of public pedagogy, a state of permanent war needs willing subjects to abide by its values, ideology, and narratives of fear and violence. Such legitimation is largely provided through a market-driven culture addicted to the production consumerism, militarism, and organized violence, largely circulated through various registers of popular culture that extend from high fashion and Hollywood movies to the creation of violent video games and music concerts sponsored by the Pentagon. The market-driven spectacle of war demands a culture of conformity, quiet intellectuals, and a largely passive republic of consumers. But it also needs subjects who find intense pleasure in the spectacle of violence.

As the pleasure principle is unconstrained by a moral compass based on a respect for others, it is increasingly shaped by the need for intense excitement and a never-ending flood of heightened sensations. What has led to this immunity and insensitivity to cruelty and prurient images of violence? Part of this process is due to the fact that the American public is

bombarded by an unprecedented "huge volume of exposure to … images of human suffering" (Bauman, 1995: p. 149). As Bauman argues, there are social costs that come with this immersion of a culture of staged violence. One consequence is that "the sheer numbers and monotony of images may have a 'wearing off' impact [and] to stave off the 'viewing fatigue,' they must be increasingly gory, shocking, and otherwise 'inventive' to arouse any sentiments at all or indeed draw attention. The level of 'familiar' violence, below which the cruelty of cruel acts escapes attention, is constantly rising" (Bauman, 1995).

Hyperviolence and spectacular representations of cruelty disrupt and block our ability to respond politically and ethically to the violence as it is actually happening on the ground. In this instance, unfamiliar violence such as extreme images of torture and death become banally familiar, while familiar violence that occurs daily is barely recognized relegated to the realm of the unnoticed and unnoticeable. How else to explain the public indifference to the violence waged by the state against nonviolent youthful protesters who are rebelling against a society in which they have been excluded from any claim on hope, prosperity, and democracy. As an increasing volume of violence is pumped into the culture, yesterday's spine-chilling and nerve-wrenching violence loses its shock value. As the need for more intense images of violence accumulates, the moral indifference and desensitization to violence grows while matters of cruelty and suffering are offered up as fodder for sports, entertainment, news media, and other outlets for seeking pleasure.

Marked by a virulent notion of hardness and aggressive masculinity, a culture of violence has become commonplace in a society in which pain, humiliation, and abuse are condensed into digestible spectacles endlessly circulated through extreme sports, reality TV, video games, YouTube postings, and proliferating forms of the new and old media. But the ideology of hardness and the economy of pleasure it justifies are also present in the material relations of power that have intensified across the globe since the 1970s. Conservative and liberal politicians alike now spend millions waging wars around the globe, funding the largest military state in the world, providing huge tax benefits to the ultrarich and major corporations, and all the while draining public coffers, increasing the scale of human poverty and misery, and eliminating all viable public spheres – whether they be the social state, public schools, public transportation, or any other aspect of a formative culture that addresses the needs of the common good. State violence, particularly the use of torture, abductions, and targeted assassinations are now justified as part of a state of exception that has become normalized. A 'political culture of hyperpunitiveness' (Herbert and Brown, 2006) has become normalized and accelerates throughout the social order like a highly charged electric current. Democracy no longer emphasizes the importance of an experience of the common good. Symptoms of ethical, political, and economic impoverishment are all around us.

Meanwhile, exaggerated violence is accelerated in the larger society and now rules screen culture. The public pedagogy of entertainment includes extreme images of violence, human suffering, and torture splashed across giant movie screens, some in 3D, offering viewers every imaginable portrayal of violent acts, each more shocking and brutal than the last. The growing taste for violence can be seen in the increasing modeling of public schools after prisons, the criminalization of behaviors such as homelessness that once were the object of social protections. A symptomatic example of the way in which violence has saturated everyday life can be seen in the growing acceptance of criminalizing the behavior of young people in public schools. Behaviors that were normally handled by teachers, guidance counselors, and school administrators are now dealt with by the police and the criminal justice system. The consequences have been disastrous for young people. Not only do schools resemble the culture of prisons, but young children are being arrested and subjected to court appearances for behaviors that can only be termed as trivial. This is not merely barbarism parading as reform – it is also a blatant indicator of the degree to which sadism and the infatuation with violence have become normalized in a society that seems to take delight in dehumanizing itself.

As the social is devalued along with rationality, ethics, and any vestige of democracy, spectacles of war, violence, and brutality now merge into forms of collective pleasure that constitute an important and new symbiosis between visual pleasure, violence, and suffering. The control society is now the ultimate form of entertainment as the pain of others, especially those considered disposable and powerless, has become the subject not of compassion but of ridicule and amusement. High octane violence and human suffering are now considered another form of entertainment designed to raise the collective pleasure quotient. Reveling in the suffering of others should no longer be reduced to a matter of individual pathology, but now registers a larger economy of pleasure across the broader culture and social landscape. My emphasis here is on the sadistic impulse and how it merges spectacles of violence and brutality with forms of collective pleasure. No society can make a claim to being a democracy as long as it defines itself through shared fears rather than shared responsibilities. Widespread violence now functions as part of an anti-immune system that turns the economy of genuine pleasure into a mode of sadism that creates the foundation for sapping democracy of any political substance and moral vitality.

The predominance of violence in all aspects of social life suggests that young people and others marginalized by class, race, and ethnicity have been abandoned as American society's claim on democracy gives way to the forces of militarism, market fundamentalism, and state terrorism. The prevalence of violence throughout American society suggests the need for a politics that not only negates the established order but imagines a new one, one informed by a radical vision in which the future does not imitate the present (Van Houdt, 2011). In this discourse, critique merges with a sense of realistic hope, and individual struggles merge into larger social movements. The challenge that young people are posing to American society is being met with a state-sponsored violence that is about more than police brutality; it is more importantly about the transformation of the United States from a social state to a warfare state, from a state that embraced the social contract to one that no longer has a language for community – a state in which the bonds of fear and commodification have replaced the bonds of civic responsibility and democratic vision. Until we address how the metaphysics of war and violence have taken hold on

American society (and in other parts of the world) and the savage social costs it has enacted, the forms of social, political, and economic disposability that young people are protesting against as well as the violence waged in response to their protests will become impossible to recognize and act on.

The Politics of Disposable Youth in the United States

The general response from progressives and liberals to the plight of youth, especially poor minorities of color, fails to take seriously the ways in which the extreme right-wing articulates its increasingly pervasive and destructive view of American society. For instance, the views of new extremists in the US Congress are often treated, especially by liberals, as a cruel hoax that is out of touch with reality or a foolhardy attempt to roll back the Obama agenda (Beutler, 2013). On the left, such views are often criticized as a domestic version of the tactics employed by the Taliban – keeping people stupid, oppressing women, living in a circle of certainty, and turning all channels of education into a mass propaganda machine of fundamentalist Americanism. All of these positions touch on elements of a deeply authoritarian agenda. But such commentaries do not go far enough. Tea Party politics is about more than bad policy, policies that favor the rich over the poor, or for that matter about modes of governance and ideology that represent a blend of civic and moral turpitude. The hidden order of neoliberal politics in this instance represents the poison of neoliberalism and its ongoing attempt to destroy those very institutions whose purpose is to enrich public memory, prevent needless human suffering, protect the environment, distribute social provisions, and safeguard the public good. Within this rationality, markets are not merely freed from progressive government regulation, they are removed from any considerations of social costs. And where government regulation does exits, it functions primarily to bail out the rich and shore up collapsing financial institutions and for what Noam Chomsky has termed America's only political party, "the business party" (2008).

In short, within the reigning economic and political order in the United States, youth have been left out of the narrative of democracy. The stories that attempt to cover over America's embrace of historical and social amnesia at the same time justify authoritarianism with a soft edge and weakens democracy through a thousand cuts to the body politic. How else to explain the Obama administration's willingness to assassinate American citizens allegedly allied with terrorists, secretly monitor the e-mail messages and text messages of its citizens, use the National Defense Authorization Act to arrest and detain indefinitely American citizens without charge or trial, subject alleged spies to an unjust military tribunal system, use drones as part of a global assassination campaign to arbitrarily kill innocent people, and then dismiss such acts as collateral damage. As Jonathan Turley points out, "An authoritarian nation is defined not just by the use of authoritarian powers, but by the ability to use them. If a president can take away your freedom or your life on his own authority, all rights become little more than a discretionary grant subject to executive will" (2012). This point seems to be especially true for young people who are one of the most vulnerable groups in society.

At the heart of the oppression experienced by young people are ideologies, modes of governance, and policies that embrace a pathological individualism, a distorted notion of freedom, and a willingness both to employ state violence to suppress dissent and abandon those suffering from a collection of social problems ranging from dire poverty and joblessness to homelessness. In the end, these are stories about disposability in which growing numbers of young people are considered dispensable and a drain on the body politic, the economy, and the sensibilities of the rich and powerful. Rather than work for a more dignified life, most young people now work simply to survive – that is, if they can find work – in a survival of the fittest society in which getting ahead and accumulating capital, especially for the ruling elite, is the only game in town. In the past, public values have been challenged and certain groups have been targeted as superfluous or redundant. But what is new about the politics of disposability that has become a central feature of contemporary American politics is the way in which such antidemocratic practices have become normalized in the existing neoliberal order. A politics of inequality and ruthless power disparities is now matched by a culture of cruelty soaked in blood, humiliation, and misery. Private injuries not only are separated from public considerations such narratives, but narratives of poverty and exclusion have become objects of scorn. Since the 1980s, many conservatives and neoliberals have viewed all noncommercial public spheres where such stories might get heard with contempt, a perfect supplement to the chilling indifference to the plight of the disadvantaged and disenfranchised (see Krugman, 2013, for example).

Any viable struggle against the authoritarian forces that dominate the United States must make visible the indignity and injustice of these narratives and the historical, political, economic, and cultural conditions that produce them. Investing in young people is central in the development of any viable democracy. Such a recognition suggests a critical analysis of how various educational forces in American society are distracting and miseducating the public. Dominant political and cultural responses to current events – such as the ongoing economic crisis, income inequality, health care reform, Hurricane Sandy, the war on terror, the Boston Marathon bombing, and the crisis of public schools in Chicago, Philadelphia, and other cities – represent flash points that reveal a growing disregard for people's democratic rights, public accountability, and civic values. As politics is disconnected from its ethical and material moorings, it becomes easier to punish and imprison young people than to educate them. From the inflated rhetoric of the political right to market-driven media peddling spectacles of violence, the influence of these criminogenic and death-saturated forces in everyday life is undermining our collective security by justifying cutbacks to social supports and restricting opportunities for democratic resistance. Saturating mainstream discourses with antipublic narratives, the neoliberal machinery of social death effectively weakens public supports and prevents the emergence of much needed new ways of thinking and speaking about politics in the

twenty-first century. But even more than neutralizing collective opposition to the growing control and wealth of predatory financial elites – that wield power across all spheres of US society – responses to social issues are increasingly dominated by a malignant characterization of marginalized groups as disposable populations. As I have mentioned above, all the while zones of abandonment accelerate the technologies and mechanisms of disposability. One consequence is the spread of a culture of cruelty in which human suffering is not only tolerated, but viewed as part of the natural order of things (Giroux, 2012, 2013).

From the Politics of Cruelty to the Politics of Civic Engagement

The notion of a culture of cruelty is useful in thinking through the convergence of everyday life and politics, of considering material relations of power – the disciplining of the body as an object of control – on the one hand, and the production of cultural meaning, especially the co-optation of popular culture to sanction official violence, on the other. The culture of cruelty is important for thinking through how life and death now converge in ways that fundamentally transform how we understand and imagine politics in the current historical moment – a moment when the most vital of safety nets health care reform is being undermined by right-wing ideologues. Evidence of such cruelty can be seen in attempts on the part of the Republican Party led House of Representatives to eliminate $40 billion dollars from the food stamp program. It is also evident in a range of policies advocated by right-wing extremists in Congress who refuse to raise the minimum wage and extend unemployment benefits for the long-term unemployed.

Before this dangerously authoritarian mind-set has a chance to take hold of our collective imagination and animate our social institutions, it is crucial that all Americans think critically and ethically about the coercive forces shaping US culture – and focus our energy on what can be done to change them. It will not be enough only to expose the falseness of the stories we are told. Educators, progressives, artists, intellectuals, youth, and workers also need to create alternative narratives about what the promise of democracy might be for our children and ourselves. This vision will require a break from established political parties, the creation of alternative public spheres in which to produce democratic narratives and visions, and a notion of politics that is educative, one that takes seriously how people interpret and mediate the world, how they see themselves in relation to others, and what it might mean to imagine otherwise in order to act otherwise. Why are millions not protesting in the streets over these barbaric policies that deprive them of life, liberty, justice, equality, and dignity? What are the pedagogical technologies and practices at work that create the conditions for people to act against their own sense of dignity, agency, and collective possibilities? Progressives and others need to make education central to any viable sense of politics so as to make matters of remembrance and consciousness central elements of what it means to be critical and engaged citizens.

There is also a need for social movements that invoke stories as a form of public memory, stories that have the potential to move people to invest in their own sense of individual and collective agency, stories that make knowledge meaningful in order to make it critical and transformative. If democracy is to once again inspire a populist politics, it is crucial to develop a number of social movements in which the stories told are never completed, but are always open to self- and social reflection, capable of pushing ever further the boundaries of our collective imagination and struggles against injustice wherever they might be. Only then will the stories that now cripple our imaginations, politics, and democracy be challenged and hopefully overcome. Only then will young people be able to recognize themselves in the stories through which America defines itself.

See also: Age Stratification; Anomie; Conspicuous consumption and distinction, history of; Deprivation: Relative; Generations, Sociology of; Neoliberalism; Transition to Adulthood; Youth Culture, Anthropology of; Youth Culture, Sociology of; Youth Gangs; Youth Movements.

Bibliography

Alexander, Michelle, 2012. The New Jim Crow. The Free Press, New York.
Bauman, Zygmunt, 1995. Life in Fragments. Blackwell, Malden.
Bauman, Zygmunt, 2004. Wasted Lives. Polity, London.
Bauman, Zygmunt, 2007. Liquid Times: Living in an Age of Uncertainty. Polity Press, Cambridge.
Bauman, Zygmunt, 31 May 2012a. Downward mobility is now a reality. The Guardian. Online: http://www.guardian.co.uk/commentisfree/2012/may/31/downward-mobility-europe-young-people.
Bauman, Zygmunt, 2012b. On Education. Polity Press, Cambridge, UK.
Bauman, Zygmunt, 2012c. This Is Not a Diary. Polity Press, Cambridge, UK.
Becker, Andrew, Schulz, G.W., 21 December 2011. Cops ready for war. Reader Supported News. Online: http://readersupportednews.org/news-section2/316-20/9023-focus-cops-ready-for-war.
Bell, Daniel, 1966. The End of Ideology: On the Exhaustion of Political Ideas in the Fifties. Free Press, New York.
Bella, Kyle, 15 December 2011. Bodies in alliance: gender theorist Judith Butler on the occupy and SlutWalk Movements. Truthout. Online: http://www.truth-out.org/bodies-alliance-gender-theorist-judith-butler-occupy-and-slutwalk-movements/1323880210.
Beutler, Brian, 22 October 2013. Take this, conservatives: criticism only makes Obamacare stronger! Ignore most conservatives attacking Healthcare.gov – but it's a mistake to assume Obama's got it under control. Salon. Online: http://www.salon.com/2013/10/22/liberals_are_right_to_join_the_healthcare_gov_pile_on_and_obamacare_will_be_better_for_it/.
Biehl, Joao, 2005. Vita: Life in a Zone of Social Abandonment. University of California Press, Berkeley.
Castells, Manuel, 2012. Networks of Outrage and Hope. Polity, London.
Chomsky, Noam, 10 October 2008. Anti-democratic nature of US capitalism is being exposed. The Irish Times. Online at: http://www.commondreams.org/view/2008/10/10-4.
Comaroff, Jean, Comaroff, John, 2006. Reflections of youth, from the past to the postcolony. In: Fisher, Melissa S., Downey, Greg (Eds.), Frontiers of Capital: Ethnographic Reflections on the New Economy. Duke University Press, Durham, NC, pp. 267–281.
Davis, Angela Y., 2007. State of emergency. In: Marable, Manning, Middlemass, Keesha, Steinberg, Ian (Eds.), Racializing Justice, Disenfranchising Lives. Palgrave, New York.
Fraser, Steve, 29 January 2013. The politics of debt in America: from debtor's prison to debtor nation. TomDispatch.com. Online: http://www.tomdispatch.com/dialogs/print/?id=175643.
Fukuyama, Francis, 2006. The End of History and the Last Man. Free Press, New York.

Giroux, Henry A., 2009. Youth in a Suspect Society. Routledge, New York.
Giroux, Henry A., 2011. Zombie Politics and Culture in the Age of Casino Capitalism. Peter Lang, New York.
Giroux, Henry A., 2012. Disposable Youth. Routledge, New York.
Giroux, Henry A., 2013a. Youth in Revolt. Paradigm, Boulder.
Giroux, Henry A., 2013b. America's Education Deficit and the War on Youth. Monthly Review Press, New York.
Graeber, David, 2012. Debt: The First 5,000 Years. Melville House, New York.
Greenwald, Glenn, 20 November 2011. The roots of the UC-Davis pepper-spraying. Salon. Online: www.salon.com/2011/11/20/the_roots_of_the_uc_davis_pepper_spraying/.
Grossberg, Lawrence, 2005. Caught in the Crossfire: Kids, Politics, and America's Future. Paradigm Publishers, Boulder.
Hall, Stuart, November 2011a. The Neo-liberal revolution. Cultural Studies 25 (6), 705–728.
Hall, Stuart, 12 September 2011b. The march of the neoliberals. The Guardian. Online at: http://www.guardian.co.uk/politics/2011/sep/12/march-of-the-neoliberals.
Hardt, Michaeland, Negri, Antonio, 2012. Declaration. Argo Navis Author Services, New York.
Herbert, Steve, Brown, Elizabeth, 2006. Conceptions of space and crime in the punitive neoliberal city. Antipode 38 (4), 755–777.
Honneth, Alex, 2009. Pathologies of Reason. Columbia University Press, New York.
Krugman, Paul, 22 September 2013. Free to be hungry. New York Times. Online: http://www.nytimes.com/2013/09/23/opinion/krugman-free-to-be-hungry.html?_r=0.
Lichtman, Richard, 14 December 2011. Not a revolution? Truthout. Online: http://www.truth-out.org/not-revolution/1323801994.
Stephens, Sharon (Ed.), 1995. Children and the Politics of Culture. Princeton University Press, Princeton, NJ.
Turley, Jonathan, 13 January 2012. 10 reasons the U.S. is no longer the land of the free. The Washington Post. Online: http://articles.washingtonpost.com/2012-01-13/opinions/35440628_1_individual-rights-indefinite-detention-citizens.
Van Houdt, John, 2011. The crisis of negation: an interview with Alain Badiou. Continent 1 (4), 234–238. Online: http://continentcontinent.cc/index.php/continent/article/viewArticle/65.
Wacquant, Loic, 2009. Punishing the Poor: The Neoliberal Government of Social Insecurity. Duke University Press, Durham, NC.

Dispute Resolutions in Economics

Henry S Farber, Princeton University, Princeton, NJ, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 3769–3772, © 2001, Elsevier Ltd.

Abstract

Negotiation is common in many settings. A common feature of all negotiation is an understanding by the parties to the negotiation about the process that will be used to resolve a dispute in the event the parties fail to reach agreement. The dispute settlement process defines the expectations that the parties have about what will happen in the event that no agreement is reached. This process, in turn, determines negotiator behavior. There are two fundamental types of dispute settlement mechanisms in the labor area: strikes and arbitration. Strikes involve the systematic imposition of costs on each side. Military disputes are implied, but presumably one or both parties concede sufficiently to make an agreement possible and end the dispute. Arbitration involves the mandating of a settlement by a third party. Litigation of civil disputes is another example of this type of mechanism. Several models of strikes and arbitration are outlined. The role of incomplete information and uncertainty are discussed in the context of these models.

Negotiation is common in many settings. A common feature of all negotiation is an understanding by the parties to the negotiation about the process that will be used to resolve a dispute in the event that the parties fail to reach agreement. Disputes between labor unions and employers can end in a strike, or can be resolved by arbitration. Disagreement over contracts or tort damages can result in civil litigation and a decision by a judge or jury. Disagreement between nations can result in military disputes. The dispute settlement process defines the expectations that the parties have about what will happen in the event that no agreement is reached. And this process, in turn, determines negotiator behavior.

There are two fundamental types of dispute settlement mechanism. The first is the systematic imposition of costs on each side. Strikes and military disputes are examples of this type of mechanism. No determinate end to the dispute is implied, but presumably one or both parties concede sufficiently to make an agreement possible and end the dispute. The second mechanism is the mandating of a settlement by a third party. Arbitration of labor and commercial disputes, and litigation of civil disputes, are examples of this type of mechanism.

Both types of mechanism (strikes and arbitration) are used in the resolution of labor disputes, and the discussion here focuses on that area. This discussion is necessarily selective and is meant to provide a brief introduction to important concepts and issues. See Hirsch and Addison (1986) for a more extensive survey.

Strikes

A difficult problem in the economics of labor unions is the modeling of strikes. The difficulty arises because strikes appear to be nonequilibrium phenomena. Both sides would have been better-off if they had been able to agree on whatever the outcome was without bearing the costs of a strike. In other words, strikes are Pareto-dominated by the same bargaining outcome without a strike. While research on strikes in the United States has diminished with the level of strike activity in the USA since the mid-1980s, there was an active empirical literature in the 1970s and 1980s.

Ashenfelter and Johnson (1969) developed a nonequilibrium model of strike activity by attributing to the union an arbitrary concession schedule positing an inverse relationship between the minimum acceptable wage and the length of strike. The firm selects the point on this schedule that maximizes profits, the trade-off being longer strikes involving larger up-front strike losses, but lower labor costs into the future. Ashenfelter and Johnson used this model to motivate an empirical analysis of the aggregate level of strike activity in the United States. Farber (1978) used the same theoretical framework to estimate a structural model of the joint determination of wage changes and strike length using data on the outcomes of a time series of cross-sections of negotiations for large US corporations.

Tracy (1986) developed an equilibrium model of strikes based on learning in a model of asymmetric information (Fudenberg and Tirole, 1981; Sobel and Takahashi, 1983). The model is based on the idea that the union does not know the profitability of the firm. The union would like to extract a high wage from high-profit firms and a lower wage from low-profit firms. The key is that high-profit firms have a high cost of bearing strikes relative to low-profit firms. In this case, the union can make a high initial wage demand. High-profit firms will prefer to pay this wage immediately rather than to lose profits during a strike. Low-profit firms will prefer to lose their lower profits during the strike in exchange for a lower wage into the future.

Tracy (1986) used this model to generate the empirical implication that there will be more strikes where there is more uncertainty about the firm's profitability. He then implemented this model using data on a large number of negotiations and using the volatility of daily stock returns as his measure of uncertainty about profitability. He finds evidence that uncertainty is correlated positively with strike activity. A final point is that the declining concession schedule in the Ashenfelter–Johnson model can be interpreted as a reduced-form expression of the union learning function. As the strike goes on, the union is willing to accept a lower wage as they learn that the firm is lower profit.

It is quite possible that strikes are not equilibrium phenomena, and that the insistence in the economics literature,

both theoretical and empirical, on equilibrium explanations, is inappropriate. Strikes can result from miscalculation. In particular, the parties may be relatively optimistic about the outcome after a strike. For example, the union may expect a wage settlement of \$20 h^{-1} after a strike, while the firm expects to pay only \$18 h^{-1} after the same strike. If the cost of a strike is \$0.50 h^{-1} to each side, then the union will accept no less than \$19.50 to avert the strike, while the firm will be willing to offer no more than \$18.50. The result will be a strike. Of course, the expectations of both sides cannot be accurate, *ex post*. This relative optimism can have a variety of sources. One possibility with a long history in the industrial relations literature, is that the parties selectively overweight arguments that are favorable to their side (Walton and McKersie, 1965; Neale and Bazerman, 1991). However, relative optimism can be generated even with appropriate weighting of available information. If each side is getting independent (and accurate in expectation) information on the likely uncertain outcome, for some fraction of the time the information will lead to relatively optimistic expectations sufficient to outweigh the costs of a strike. Strikes will occur in these cases.

Arbitration

In the United States, arbitration is used commonly to determine the terms of labor contracts in the public sector in cases where the parties cannot agree on their own. In conventional arbitration, the arbitrators are free to impose whatever settlements they deem appropriate. A common alternative is final-offer arbitration, where the arbitrator must select one of the final offers submitted by the parties and cannot compromise.

Conventional Arbitration

Consider the following simple model of the negotiation process where conventional arbitration is the dispute settlement mechanism (Farber and Katz, 1979). Suppose the parties are negotiating over the split of a pie of size one. The share of the pie going to the union is Y, and the share going to the firm is $(1 - Y)$. The arbitrator will impose a settlement of Y_a, and each party has a prior distribution on Y_a. Let Y_a^u and Y_a^f represent the prior expected value to the union and firm, respectively, of the arbitration award. Finally, let C_u and C_f represent the cost to the union and the firm of using arbitration (denominated in pie shares). Assume further that the parties are risk neutral, so that the union will prefer any negotiated share for itself (Y_n) such that $Y_n > Y_a^u - C_u$ and the firm will prefer any negotiated share for itself such that $1 - Y_n > 1 - Y_a^u - C_f$. These are the reservation values of the parties in negotiation. The result is a *contract zone* of settlements preferred by both sides to arbitration that consists of all shares that fall in the range:

$$Y_a^u - C_u < Y_n < Y_a^f + C_f \qquad [1]$$

The size of the contract zone is:

$$CZ = Y_a^f - Y_a^u + C_u + C_f \qquad [2]$$

In the simplest model, there will be a negotiated settlement whenever $CZ \geq 0$, and there will be an arbitrated outcome whenever $CZ < 0$.

This model provides an important insight into when one might expect a dispute. Relative optimism about the likely arbitration award ($Y_a^u > Y_a^f$) can be a reason for a dispute. If the relative optimism is sufficiently great to outweigh the total costs of arbitration ($C_u + C_f$), then there will be no contract zone, and arbitration will be required. Thus, relative optimism is a cause of disputes. Clearly, higher total costs of arbitration imply a larger contract zone. Larger costs make it less likely that any degree of relative optimism will cause there to be no contract zone. Thus, larger total costs of disputing reduce the likelihood of disputes. This latter property is implied by any reasonable model of disputes.

The model also demonstrates how the dispute settlement mechanism determines fundamentally negotiated outcomes. Consider the case where the parties have identical expectations about the arbitration award of $Y_e = Y_a^u = Y_a^f$. The size of the contract zone in this case is determined entirely by the costs of arbitration (eqn [2]), and there will be no dispute. The negotiated settlement will lie in the range defined by:

$$Y_e - C_u < Y_n < Y_e + C_f \qquad [3]$$

In other words, the negotiated settlement will lie in some range around the common expectation of the arbitration award.

One criticism of conventional arbitration as a dispute settlement mechanism in labor disputes is that arbitrators appear to split the difference between the demands of the parties, resulting in an arbitration award that is close to the average of the demands. However, eqn [3] makes it clear that the opposite is likely to be true: The negotiated settlement is likely to lie between the reservation values of the parties. Mnookin and Kornhauser (1979) and Farber (1981) present analyses that make this point.

Final-Offer Arbitration

The perception that arbitrators split the difference in conventional arbitration led to the belief that conventional arbitration 'chilled' bargaining by providing the parties with an incentive to maintain extreme demands. This led to the development of final-offer arbitration, which prevented the arbitrator from compromising (Stevens, 1966). Farber (1980) developed a simple model of final-offer arbitration where arbitrators select the offer closest to what they would have imposed in conventional arbitration (Y_a). In this model, the probability that the union's offer is selected is:

$$P_u = Pr\left(Y_a > \frac{Y_u + Y_f}{2}\right) = 1 - F\left(\frac{Y_u + Y_f}{2}\right) \qquad [4]$$

where Y_u and Y_f represent the final offers of the union and the firm, respectively, and $F(\cdot)$ represents the cumulative distribution function of the parties' common prior on Y_a.

Assuming risk neutrality, the expected payoff to the union from final-offer arbitration is:

$$V_u = P_u Y_u + (1 - P_u) Y_f \qquad [5]$$

and the expected payoff to the firm is:

$$V_f = 1 - \left[P_u Y_u + (1 - P_u) Y_f\right] \qquad [6]$$

The Nash equilibrium final offers of the parties are the offers such that (1) Y_u maximizes V_u conditional on Y_f, and (2) Y_f maximizes V_f conditional on Y_u. These offers have the property that neither party will desire to change its offer.

Since the parties are uncertain about Y_a, the Nash equilibrium final offers reflect a trade-off between the value to the party of its offer if its offer is selected, and the probability that its offer is selected. A more extreme offer yields a higher payoff to the party if its offer is selected by the arbitrator, but it lowers the probability that the offer will be selected.

A key implication of this model is that final-offer arbitration may not encourage more convergence in negotiation than conventional arbitration because of the desire to present optimal final offers to the arbitrator that may be extreme.

Empirical Evidence

Ashenfelter and Bloom (1984) present evidence from arbitration of contract disputes for public sector employees in New Jersey, where both conventional and final-offer arbitration are used, consistent with the general model of arbitrator behavior outlined here. They find that arbitrators use a common underlying notion of Y_a to make awards in both conventional and final-offer arbitration. Their evidence is consistent with the arbitrators' decision in final-offer arbitration being to select the offer closest to Y_a. Bazerman and Farber (1986) reach similar conclusions based on evidence from arbitration awards, both conventional and final-offer, made by professional arbitrators in hypothetical cases.

Convincing evidence is scarce on whether final-offer arbitration leads to more settlements and reduces reliance on arbitration. Ashenfelter et al. (1992) present experimental evidence that does not show higher rates of agreement in final-offer arbitration than in conventional arbitration.

Bibliography

Ashenfelter, O., Bloom, D.E., 1984. Models of arbitrator behavior. American Economic Review 74, 111–124.

Ashenfelter, O., Currie, J., Farber, H.S., Spiegel, M., 1992. An experimental comparison of dispute rates in alternative arbitration systems. Econometrica 60, 1407–1433.

Ashenfelter, O., Johnson, G., 1969. Trade unions, bargaining theory, and industrial strike activity. American Economic Review 59, 35–94.

Bazerman, M.H., Farber, H.S., 1986. The general basis of arbitrator behavior: an empirical analysis of conventional and final-offer arbitration. Econometrica 54, 1503–1528.

Farber, H.S., 1978. Bargaining theory, wage outcomes, and the occurrence of strikes: an econometric analysis. American Economic Review 68, 262–271.

Farber, H.S., 1980. An analysis of final-offer arbitration. Journal of Conflict Resolution 24, 683–705.

Farber, H.S., 1981. Splitting-the-difference in interest arbitration. Industrial and Labor Relations Review 34, 70–77.

Farber, H.S., Katz, H.C., 1979. Interest arbitration, outcomes, and the incentive to bargain. Industrial and Labor Relations Review 32, 55–63.

Fudenberg, D., Tirole, J., 1981. Sequential bargaining with incomplete information. Review of Economic Studies 50, 221–248.

Hirsch, B.T., Addison, J.T., 1986. The Economic Analysis of Unions: New Approaches and Evidence. Allen & Unwin, Boston, MA.

Mnookin, R., Kornhauser, L., 1979. Bargaining in the shadow of the law – the case of divorce. Yale Law Journal 88, 950–997.

Neale, M.A., Bazerman, M.H., 1991. Cognition and Rationality in Negotiation. The Free Press, New York.

Sobel, J., Takahashi, I., 1983. A multi-stage model of bargaining. Review of Economic Studies 50, 411–426.

Stevens, C.M., 1966. Is compulsory arbitration compatible with bargaining? Industrial Relations 5, 38–52.

Tracy, J.S., 1986. An investigation into the determinants of US strike activity. American Economic Review 76, 423–436.

Walton, R.E., McKersie, R.B., 1965. A Behavioral Theory of Labor Negotiations: An Analysis of a Social Interaction System. McGraw-Hill, New York.

Disputes, Social Construction and Transformation of

Lynn Mather, SUNY Buffalo Law School, Buffalo, NY, USA
Barbara Yngvesson, Hampshire College, Amherst, MA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by L. Mather, volume 6, pp. 3772–3776, © 2001, Elsevier Ltd.

Abstract

Dispute transformation involves the study of how a problem that could be defined as an injury or grievance takes shape over time as a public conflict by the involvement of others (supporters, lawyers, and third parties of various kinds). The process of transformation involves negotiation over framing, which may narrow or expand a conflict in ways that tend to advantage certain interests over others. In this way, changes in the form or content of disputes link dispute management to the ways law and other normative orders are created, maintained, and sometimes challenged or transformed.

Introduction

The 'social construction' of disputes refers to the process by which a subjectively perceived injury or injustice is articulated as a grievance and asserted as a claim before a third party. Disputes are not concrete entities or objective events, but instead take shape over time as a result of the involvement of other participants. 'Transformation of disputes' refers to the change in form or content of a perceived injury or wrong as a result of this involvement, as expressed through language and actions, according to available categories and institutional practices.

Key contributions of the dispute transformation perspective include psychological insight into the perception and assertion of grievances by individual actors; sociological analysis of the distribution of grievances, claims, and court cases for different types of injuries or problems; anthropological study of the way cultural and legal 'dispositions' or 'consciousness' influence conflict management; political science research on how legal institutions and actors shape the perception and construction of claims and grievances; and legal studies of the normative and doctrinal consequences of reframing of disputes. Understanding how disputes emerge and are transformed links a microanalysis of the experience of and response to injury or trouble with a macroview of law, political order, and change.

Law and Disputing

Empirical research on dispute transformation developed in the late 1970s and early 1980s as a result of trends in several social science fields and in law. Anthropologists showed how a focus on 'trouble cases' could provide a useful framework for comparative study of law in different societies. Research reported in Nader and Todd (1978), for example, explored the factors behind parties' choices of different methods for pursuing, or attempting to resolve, trouble cases in diverse societies. Relations between those involved in a case, the location, and the nature of the problem, along with the range of alternatives for pursuing it, could all explain the construction and handling of what might become a public dispute. The interest of anthropologists at the time in actor choices and the availability of different alternatives for conflict management overlapped with the interest of legal scholars in access to justice.

Expansion of legal services for litigants in civil cases and concern for the role of civil courts in society led analysts to examine the legal needs of citizens. What kinds of cases and what groups of citizens are and are not being served by the civil justice system? How can access to justice be improved? Questions such as these spurred efforts to describe and count 'unmet legal needs' in the United States (Curran, 1977) and in the United Kingdom (Genn, 1999). It is easy to count cases in court, but counting the cases that have not been filed proves more problematic. The problems include, for example, what kinds of cases should be in court, are there too many (or too few) court cases, and what populations do the courts serve? Through such inquiries, the very nature of what constitutes a 'legal' or 'justiciable' claim is called into question (Genn, 1999). Comparative and longitudinal studies of legal caseloads also fueled the interest in improving understanding of the relation between individual disputes, social conflicts, and court cases (Munger, 1990). Such concerns led scholars to focus explicitly on the social construction of disputes and their emergence in court.

Recognition that adjudication in court provided only one way to resolve conflicts, and a costly way at that, encouraged the development of alternative processes for dispute management. Interest in alternatives to courts also led to research on dispute transformation, as scholars explored the impact of different institutional mechanisms on the emergence and management of disputes. Alternative dispute resolution (ADR) provided processes such as mediation and arbitration as alternatives to adjudication in the courtroom (Goldberg et al., 1985; Tomasic and Feeley, 1982). ADR advocates praised its informality, lower cost, speed, and case outcomes particularized to the needs of the disputants (Menkel-Meadow, 1985; Goldberg et al., 1985). ADR critics voiced concerns about a two-tier legal system with courts reserved for wealthier litigants and alternatives for others, and about the lack of legal norms and public accountability for dispute management outside of court (Fiss, 1984; Harrington, 1985). Scholars on both sides of the debate sought to bolster their positions by

gathering information on how disputes are transformed through negotiation, mediation, arbitration, or adjudication.

Another interest in dispute transformation came from sociologists studying case processing in the criminal justice system. Criminal justice research showed how officials negotiated the meaning of events and the identities of suspects in processes of plea bargaining. Crimes are defined according to their normal situational and social features, as well as by legal codes (Sudnow, 1965). Psychological studies of negotiation in conflict resolution also examined dispute transformation. They focused not only on how players tried to maximize their own interests, but also on how they sought actively to define the conflict frame, that is, the cognitive interpretation of the conflict (Pinkley and Northcraft, 1994).

Finally, political scientists studied dispute transformation in order to better understand legal mobilization and the role of courts in society (McCann, 1994; Mather, 1998). Test case litigation involves the construction of individual lawsuits as vehicles for articulating and advocating collective political interests. In successful litigation, disputants may become important creators of law and lawyers are spokespersons for their cause. Court trials may be confined to the individual facts and issues in dispute, or broadened to address the normative implications and wider political meaning of the dispute. As a consequence of cause lawyering, interest group activity, or media attention, political trials may result from such dispute expansion. Parties in litigation, their lawyers, and court officials may each seek to transform disputes in different ways, for different political ends.

The Emergence and Transformation of Disputes

A hypothetical sexual harassment case can illustrate some of the central issues in the emergence and transformation of disputes. Consider a situation in which an office manager makes what could be interpreted as sexual overtures toward his secretary. He frequently tells crude jokes, embarrasses her with comments about her looks and dress, and invites her for after-hours dates. She is not attracted to him and objects to his behavior, but she likes her job. What would be her response? She might simply brush him off and try to live with the situation ('lump it'); seek advice from coworkers, neighbors, friends; leave her job and seek work elsewhere ('exit'); pursue an internal complaint within her place of work, or seek legal counsel outside the job and file a claim of sexual harassment.

This example suggests how the secretary's response to her manager depends first on recognizing and identifying a problem, which is shaped, in turn, by her own experiences, by the responses of others in her social surround, and by the availability of alternatives for managing the situation. Her handling of it may or may not lead to a legal case. For many women employees in the United States – and still for many worldwide – the occasional lecherous boss was part of a normal working environment. Cultural and political changes led to women's recognition of certain workplace experiences as grievances. Women's groups mobilized to expand federal law against sex discrimination in employment (Title VII of the Civil Rights Act) to add the newly named category for such grievances: 'sexual harassment' (MacKinnon, 1979). On occasion, supporters of an employee – such as coworkers, a women's group, or employees' union – might broaden the scope of the conflict to transform her individual problem with the offending manager into a larger claim against sexist working conditions throughout the office. Additional issues such as gender discrepancies in pay and working conditions or racial harassment might be added to this emerging dispute. Advice from a civil rights lawyer could lead to a class action lawsuit, which would further broaden the scope of the conflict and add new parties to the case. Alternatively, internal processes of conflict management within the organization could depoliticize the matter, remove any aspect of sex discrimination or sexual harassment, and narrow it into a personality conflict between the two parties, thus transforming an issue of public law into a private dispute (Edelman et al., 1993). In short, both the meaning and the course of a conflict are contingent. They depend on the perceptions, actions, and experiences of various parties, and on the potentials for response provided by language, institutions, and support networks of various kinds. Furthermore, the ways in which a problem is defined – and whether it is even seen as a dispute or not – reinforces certain power relations and ideas, and links law to everyday life.

The Disputing Pyramid: Naming, Blaming, Claiming

Why are some injuries or problems ignored, while others become disputes? And why do some forms of trouble end up as court cases but most do not? These questions inspired a number of influential studies in the 1970s and 1980s that described a 'disputing pyramid,' or 'legal iceberg' (Galanter, 1974; Miller and Sarat, 1981) with vast numbers of events that might become grievances at the bottom, fewer events that have 'matured' into grievances in the middle, still fewer legal claims toward the top, and only a small number of actual cases in court (Galanter, 1974; Miller and Sarat, 1981). In the pyramid or developmental approach to disputing, the emergence of disputes depends first on an individual's perceiving an injury or problem ('naming'), then attributing the problem to the fault of another ('blaming'), and finally confronting the opposing party and seeking some kind of remedy ('claiming') (Felstiner et al., 1981). Many injuries are never perceived as such, and still others are not acknowledged because of cultural disposition, institutional barriers, or the lack of social support. Scholars adopting this approach have been concerned with the progress of 'perceived injurious experiences' through the pyramid, and specifically with how differences in the class, education, work situation, and social networks of an injured party might affect the process of transformation, leading to 'grievance apathy' (the failure of a grievance to 'mature' into a dispute) or the escalation of grievances into claims, including legal claims (Felstiner et al., 1981).

Early thinking about these issues suggested that movement of grievances to the top of the disputing pyramid depended on demographic characteristics of the parties (e.g., class, education), the relationship between the parties, the type of claim, and the relation that the parties have to the legal system. Further research showed individual demographics not to be

very significant; instead, the type of claim (e.g., family, consumer, discrimination) overwhelmingly determines the shape of the disputing pyramid, even across different countries (Kritzer, 2011). Murayama (2012: 77), for example, compares rates of grievances, claims, and disputes in Japan and the United States and concludes: "nationality does not determine dispute behavior – problem type does." Institutionalized procedures and remedies for handling certain kinds of conflicts may encourage claiming by lowering costs and making the disputing process more accessible and predictable. In addition to social and institutional factors, psychological factors also affect disputing. Individuals vary in their claims consciousness and in their propensity to assert claims or pursue legal rights. This variation is associated with preferences for different modes of conflict resolution, different understandings of the legal system, and with self-described personality attributes, but appears to be independent of demographic status (Vidmar and Schuller, 1987; Ewick and Silbey, 1998).

Transformation of Disputes: The Dialectic of Narrowing and Expansion

Studies of the series of transformations through which an (unperceived) injury becomes a legal case in court provided an important contribution to the sociological study of law by placing the 'grievant' or disputant individual at the center of the disputing process (Felstiner et al., 1981). At the same time, focus on the individual as the primary agent of transformation, and on action as a sequence of choices ('lumping,' 'blaming,' 'claiming,' 'disputing') at key points on a 'decision tree,' conceptualizes transformation as dyadic, linear, intentional, active, and focused on a series of stages through which each 'perceived injurious experience' passes as it 'matures' into a dispute. But many such experiences, and the disputes that emerge from them, evidence a more complex transformational dynamic, one that "takes place nonconsciously and is not the product of reasoned analysis or deliberate choice" (Engel, 2013: 304).

An alternative approach to disputing conceptualizes the emergence of disputes not as a linear process, but as one characterized by a 'dialectic' of narrowing and expansion (Santos, 1977); 'choice' is embodied in a system of dispositions (Bourdieu, 1977) that shape, and are in turn shaped by, the process of disputing; and the focus is as much on third-party transactions – collaborative, contentious – as on individual agency. Dispute transformation thus becomes a key mechanism in the production of social life, and time is put at the center of a process that is characterized by recursion, narrative reformulation, third-party interaction, and the ongoing reconfiguration of legal identities (Mather and Yngvesson, 1981; Yngvesson and Coutin, 2006; Engel, 2013). Indeed, as Coutin (2011: 570) argues, "law's complex temporal character" takes shape as part of this process of transaction and transformation, and in this sense, the documents and texts of which law is composed might themselves be considered agents of transformation, in that they contain "the residue of prior negotiations and transformations, a residue that leads forward as well as into the past" (2011: 572). The capacity of law to 'return' to the past or project into the future limits the agency of the actors who use it, even as they may exploit law's indeterminacy and openness to (re)interpretation.

The following sections explore three different aspects of the dialectic of narrowing and expansion in the transformation of disputes: framing and the role of language; the role of lawyers, courts, and other institutions in transformation; and the relation of dispute transformation to the articulation, imposition, circumvention, and creation of legal and other normative frameworks.

Framing: Language and Dispute Transformation

The definition of a dispute articulated by participants is not neutral. It orders 'facts' and invokes 'norms' in ways that reflect the interests and dispositions of the parties, or that anticipate the reactions of others (Mather and Yngvesson, 1981). By representing parties, events, and claims in particular ways, or by including or excluding certain facts, disputants construct an account of what transpired that is tied to an implicit or explicit normative framework. For example, in describing an experience as 'sexual harassment,' the word '*harassment*' itself invokes a different normative meaning than, say, '*flirting*' or '*being friendly*.' Struggles over language involve negotiation over which framework to apply and which narrative to construct, and such negotiations are integral to the disputing process and its outcome (Conley and O'Barr, 1997).

Research by Marshall on the everyday construction of sexual harassment in the workplace (2003) illustrates these struggles over meaning. Marshall examined the experiences of women in administrative and clerical positions at a large university as they confronted a variety of harassing behaviors, including sexual advances and invitations, sexual banter and joking, and displays of graphic sexual materials. Three key frames emerged: an 'injustice frame' articulated by the women's movement that emphasizes the relationship of sexual harassment to discrimination and inequality; 'management frames' that distance sexual harassment from broader structures of inequality in the workplace; and 'freedom of sexual expression' frames used by women who regarded sexual interactions in the workplace as a way to exercise power in that setting (2003: 673–679).

Marshall examines the complex dynamics that came into play as women struggled to interpret the sexual behaviors of their workplace colleagues, noting that they "rarely chose a single frame for understanding … unwanted sexual attention," but tried instead to balance competing interpretations. In this delicate balancing act, "the injustice frames that characterized such experiences as discrimination" were in tension with perspectives "suggesting that they should not take it so seriously …. Thus, some women could not decide whether they had even been injured" (2003: 679). In one case, Marshall notes that a woman reframed behavior that she had initially considered to be inappropriate but unavoidable sexual conduct as 'sexual harassment,' following the circulation of sexual harassment policy information by her employer. For this woman and others like her, law became 'a tool after the fact,' with which they could 'go back and reinterpret their experiences' (2003: 684; and see Coutin, 2011). For other women, in contrast, their exposure to human-resource frames, which portrays sexual harassment 'as a personality conflict or

a management problem,' becomes a ground for adapting to behavior they find offensive and rejecting the label of sexual harassment (2003: 685).

In conclusion, Marshall suggests that sexual harassment law may contribute to normative change when it overlaps with injustice frames articulated by the women's movement, since these translate previously acceptable social behaviors into injuries, which may then become legally cognizable. At the same time, she notes that "law can inhibit social change," since legal remedies tend to "deemphasize collective solutions and instead encourage women to pursue more individualistic strategies, such as confronting harassers, lodging complaints with employers, or filing lawsuits when other options fail" (2003: 686).

The risk for privatizing what might otherwise be framed as a collective (structural) problem of sexual misconduct has also been played out on university and college campuses over the past several years, as students who complained about sexual violence have been referred to university disciplinary panels, which all too often act in the interests of the university to avoid undesirable publicity, dismissing the complaints by implicitly reframing them as part of a 'culture of drunken hook-ups' among students (Perez-Peña and Taylor, 2014: 25). The experience of a female undergraduate at Columbia University, whose complaint of being raped by a fellow student was dismissed by the internal panel, but then taken up, with a number of related cases, by a student magazine at Columbia, illustrates the dialectic of narrowing and expansion in such cases and the collaborative forms of agency that may come into play. By bringing the perspectives of the women involved into public view and suggesting that sexual violence by male students was being swept under the rug by the university, the magazine article led to rapid escalation of the conflict. It became the focus of heated debates in dorm rooms, dining halls, and social media, generating widespread student protests in support of the women and demands for change in university policy. One of the women involved further escalated the conflict by agreeing to identify herself in a front-page interview in the 4 May 2014 edition of *The New York Times*, which also documented protests by student activists calling themselves 'No Red Tape.' The activists tried to gain entrance to an event for prospective freshmen at Columbia, with red tape over their mouths, so that they could hand out letters encouraging the students to press university officials for details about resources available to victims of sexual assault (Perez-Peña and Taylor, 2014: 25).

In these cases, as in the research carried out by Marshall, the availability of 'injustice frames' for interpreting the students' experiences was crucial to their reformulation. Through the intervention of the student magazine at Columbia, the complaints were taken up by a nationwide network of student activists who made "shrewd use of the law and the media" (Perez-Peña and Taylor, 2014: 1) to recast them as violations of federal rules under Title IX of the Education Amendments of 1972. As a result, 23 students at Columbia filed complaints against the university, charging it with the mishandling of sexual misconduct cases.

The Supreme Court ruling on 26 July 2013 in *United States v. Windsor* that a ban on federal benefits for gay couples, enacted under the Defense of Marriage Act (DOMA), was unconstitutional, provides another example of the complexity of third-party interactions over framing. In this complaint, Edie Windsor successfully challenged the constitutionality of the DOMA before the U.S. Supreme Court, with a dispute about her entitlement to federal tax relief following the death of her partner of more than 40 years, Thea Spyer. Although the two had been legally married in Canada, and Windsor was Spyer's sole heir, Windsor was ineligible for the exemption on estate taxes that applies to husbands and wives, because DOMA, which was signed into law in 1996, recognized only heterosexual marriages in determining federal benefits and protections. This rendered the considerable estate Windsor inherited from her partner taxable as though they were strangers. At one level, then, this was a dispute about the recovery of Windsor's money. At the same time, it presented an opportunity for activists in the gay and lesbian community to contest the constitutionality of DOMA's differential treatment of heterosexual and homosexual marriages.

Whether Edie Windsor's complaint over a tax exemption was the right opportunity for this challenge was itself a disputed matter between her own lawyer and activists and organizations associated with civil rights litigation, such as the American Civil Liberties Union, Lambda Legal, and the Gay & Lesbian Advocates & Defenders. Activists were concerned about the fact that Windsor's complaint would present the Court "with the opportunity for a narrow ruling that would restore Windsor's money but overlook the broader question of gay marriage" (Levy, 2013: 62). But for Windsor's lawyer, a significant consideration in opting for Edie Windsor as the perfect plaintiff for a challenge to heterosexual marriage was that she did not 'look gay,' in part because she was a widow, but also due to the fact that she could be perceived as having 'aged out of carnality' (2013: 61). Eventually, in spite of lingering concerns about whether 'the story of an estate tax bill' was the right venue for challenging DOMA, as one activist interviewed said, "we decided that the smartest, most strategic thing to do was work together" (Levy, 2013: 62).

United States v. Windsor illustrates the ways in which the transformation of a dispute about money into a challenge of the normative status of heterosexual marriage is made possible by the collaborative agency of third parties with a highly specialized knowledge of civil rights law and political acumen. Likewise, the framing of sexual conduct complaints as sexual violence by student activists at Columbia was made possible by the students' 'savvy' use of social media to mobilize local and national audiences, their position at an elite institution, and the targeting of sexual violence on campus as a focus of the Obama administration (Pérez-Peña and Taylor, 2014). At the same time, the social identities of the complainants (Columbia university students and Edie Windsor, an 83-year-old widow) contributed to the successful reframing of the cases. Political groups planning test case litigation (such as *United States v. Windsor*) are well aware of the importance of having a sympathetic plaintiff to represent their cause and thus, for example, legal challenges to affirmative action in the United States have usually had white women as plaintiffs rather than men.

Reframing is fundamental to dispute transformation in a range of settings, from appeals brought to the Supreme Court to complaints about neighborhood violence heard by district court clerks in Massachusetts (Yngvesson, 1988; Merry, 1990).

In district court cases, as in the management of student sexual conduct by college disciplinary panels, or the narrowing of sexual harassment complaints in the logic of a 'management frame' as reported by Marshall, the clerk's goal is to keep the complaints out of court, by rephrasing what complainants regard as 'vicious' behavior, into normal trouble in a chaotic environment. In this way, the court distributes responsibility for what it considers 'garbage cases' – cases that are trouble for the court, or as in the sexual conduct cases discussed above, trouble for management at the university – between the disputants, while preventing an escalation of the conflict. Hearings before the clerk of court illustrate the narrowing of complaints in cases with the potential to challenge normative order in the community at large, but where complainants lack the social and economic resources to mobilize a wider audience in support of their cause.

The Role of Lawyers, Courts, and Other Institutions in Dispute Transformation

As gatekeepers to courts in many countries, lawyers act as intermediaries between disputants and the legal system. In deciding whether to represent a claimant and if so, how to frame the case, lawyers are central to the process of dispute construction. The screening practices of Chinese lawyers, for example, reflect financial, cultural, and institutional constraints on them, resulting in very little representation of workers with labor disputes (Michelson, 2006). In considering whether to use litigation against the tobacco industry to pursue the claims of sick smokers in the United Kingdom (as was done by lawyers in hundreds of cases in the United States), English lawyers (except for one firm) rejected the cases (Mather, 2009). Thus, some kinds of disputes (or the complaints of certain individuals) go forward as legal cases and others do not, due to the actions of lawyers.

Once clients have legal representation, what role do their lawyers play? Cain's (1979) classic study of English solicitors showed how often they rephrased the objectives of their clients. More recently, an important debate in the United States has centered on the extent to which lawyers escalate conflict, urging their clients to pursue formal legal strategies for problems that might be more appropriately handled informally, or dampen conflict, encouraging negotiation and out-of-court settlements. This debate has been especially contested in the field of tort law with significant attention paid to the economic incentives and disincentives for lawyers of different case strategies, and to the media portrayal of plaintiff and defense lawyers (Haltom and McCann, 2004). Similarly, in divorce, an impetus for the introduction of mandatory mediation in the 1980s was a belief that lawyers aggravated the hostility and conflict between spouses. But in fact, for most divorces in the United States, only one spouse can afford a lawyer and the other is *pro se* or else both parties are unrepresented by counsel. Even in two-lawyer divorces, attorneys typically seek to cool out their clients in order to protect the children and to preserve the long-term relationship between ex-spouses, or because lawyers see cooperative settlement to be more cost-effective, or because the law encourages equitable division of property and shared parenting (Mather et al., 2001). Considerable research in the United States and elsewhere has corroborated Sarat and Felstiner's (1995) finding that most divorce lawyers try to focus narrowly on the legal and financial issues, to the exclusion of emotional, personal, or moral issues that their clients may want to discuss.

This point underscores the more general question: To what extent are clients' stories and objectives 'heard' by counsel? One thing that lawyers clearly do is translate clients' problems into legal language. Lawyers rephrase auto accidents (or, as in the case below, torture) into torts, domestic violence into assaults, and racist actions into civil rights violations. In lawyers' offices, clients may want to pour out stories with their full social and personal context. But details that clients see as highly relevant are often dismissed by counsel as not germane to the legal case and objectives that clients seek may not be attainable within the framework of the legal system. The resulting transformation of a client's dispute through the lawyer's actions can lead to a profound discontinuity between lawyers and clients. In a survey of plaintiffs in medical malpractice cases, Relis (2007) describes the 'parallel worlds of understanding' held by plaintiffs and their lawyers, concluding that despite the financial frame of compensation required by tort law, for many plaintiffs, 'it is not about the money!' Similarly, in a study of claimants to the 9/11 Victim Compensation Fund, Hadfield (2008) describes the troubling trade-off between money and the deeply held nonmonetary values and goals of victims.

The federal torture case of *Romagoza v. Garcia* (2002) illustrates the potential and the frustration in lawyer–client collaboration while constructing a legal narrative, in a case where narrowing the dispute at one stage contributed to its eventual expansion several years later. Rubin's (2008) analysis of dispute transformation in *Romagoza* also illuminates lawyers' skill in adapting a body of law originally developed for one purpose (imprisoning or executing those responsible for crimes against a nation or against humanity), for use in tort litigation where individuals seek monetary reparations for personal injuries from former state officials. In this way, the case provides a powerful example of what Coutin (2011: 570) describes as "law's archaeology," in reference to the process through which legal artifacts produced in one forum may reshape claims brought in another, thus constituting law itself as a historical actor in the dispute transformation process.

Romagoza v. Garcia was brought before a jury in the U.S. Southern District of Florida by Neris Gonzalez and two other Salvadoran citizens, against the former Salvadoran Minister of Defense and the former head of the Salvadoran National Guard, for their role in instigating and authorizing torture by the Salvadoran National Guard. Gonzalez, who had been severely tortured when she was 8 months pregnant, was persuaded to join the lawsuit by the legal director of the San Francisco-based Center for Justice and Accountability, which wanted to initiate a lawsuit under the 1789 Alien Tort Claims Act and the 1992 Torture Victims Protection Act against ex-military officials from El Salvador, some of whom were residing legally in the United States. Gonzalez agreed to join the lawsuit as a continuation of her political struggle of 30 years against the Salvadoran state and after her unsatisfactory experience before a Truth Commission set up in 1993 by the United Nations, in which no one was held to account.

Recognizing the importance of the details of her torture to Gonzalez's goals in joining the civil lawsuit against the generals, her lawyers tried to incorporate as much of her experience into the legal narrative as possible. Nevertheless, the scope of her testimony was bounded by the procedural rules of a tort lawsuit and also needed to be tailored to an audience of U.S. citizens (the jury) with no prior knowledge of the Salvadoran Civil War. As an abstract legal category, 'torture' had to be made quantifiable and understandable for the jury, rendered as a 'scale of terror' and ultimately assigned a monetary value (Rubin, 2008: 273, 276). As a result, the most traumatic dimension of her own torture, the killing of her unborn child, had to be left out of the legal narrative. Gonzalez described this process of abstraction, quantification, and legalization as 'very technical,' adding that "emotion is not technical. Pain is not technical …. So we had to choose the most important parts of the pain in the trial, so that the courtroom could function" (Rubin, 2008: 278).

Although *Romagoza v. Garcia* was successful in that the two generals were held responsible for the torture of the three plaintiffs and were ordered to pay $54.6 million in damages, Gonzalez's larger political goals were unfulfilled by the trial. She did, however, remain hopeful that the trial would encourage others, and that "future legal actions will complete the process of justice that *Romagoza* began" (Rubin, 2008: 279). Just over a decade later, this hope was realized in two separate rulings in a Miami immigration court (April 2013; February 2014) that cited *Romagoza* to justify deportation orders against General Garcia and his codefendant in that case. Thus, this jury trial continues to leave traces, and provides a powerful example of what Santos (1977) describes as the dialectic of narrowing and expansion. In what was described as "an unusually expansive and scalding 66-page decision," the federal judge in the 2014 ruling harshly condemned General Garcia (Preston, 2014: A11). Through their use of *Romagoza* and other precedents to justify the deportation orders of the two generals, the Immigration Court rulings demonstrated how narrowing in one phase of a dispute may contribute to subsequent expansion of the same dispute many years later, in this way challenging the normative order that allowed the torture of Neris Gonzalez and other Salvadorans. The deportation orders are under appeal by the generals' lawyers (as of June 2014).

As *Romagoza v. García* suggests, institutional structures influence disputing by channeling the ways in which conflicts can be transformed. Courts, for example, often have strict rules of evidence that limit what kinds of information may be included in the construction of a case during trial. But informal negotiations outside of court, such as plea bargaining or settlement talks, have few such limits. When, for example, informal plea negotiations over disposition of a first-time drug defendant include discussions of the defendant's social background, family situation, and employment, lawyers may lean toward lenient treatment. But none of this material might be admitted in adjudication, and the resulting sentence could be much harsher after conviction at trial. Different institutional procedures transform cases in different ways, producing variation in case outcomes.

Richland's (2008) linguistic analysis of interactions between litigants and judges in a Hopi tribal court provides an illustration of this kind of variation in institutional procedure. Richland focuses on the tensions between Anglo-American legal procedure and Hopi legal heritage in a setting where tribal members were calling for a radical reclamation of legal institutions that were originally used to dominate them. He provides a detailed analysis of how Hopi judges negotiate this tension, creating 'meaningful spaces' (2008) that momentarily destabilize the truth-making practices of Anglo law by permitting the inclusion of personal histories of suffering that are open-ended and challenge the rules of evidence allowed in Anglo courts. In this way, Richland suggests, what might appear to be irreconcilable differences across practices are accommodated in ways that transform both Anglo-American law and Hopi tradition.

Attention to dispute transformation may also help to bridge presumed differences between procedures that are conventionally categorized as 'adjudication' or 'mediation.' In theory, a distinguishing feature of mediation – in contrast to adjudication – is the ability of mediators to range widely over different issues in a conflict. Whereas judges are supposed to narrow their focus to particular legal elements of a case, mediators (not bound by law) may encourage the emergence of new issues in their effort to formulate an agreement that is acceptable to both parties. Divorce mediators sometimes comment on the 'strange-looking agreements' that result from such wide-open sessions. Nevertheless, mediators may have more commonalities with judges than with other mediators in terms of how they actually handle cases, and negotiation is frequently intertwined with adjudication. Instead of assuming differences across processes, one might focus instead on how different processes influence dispute transformation – for example, through specialized or general language, through closed or open dispute hearings, or through linkages between dispute processes and other organizations in society (Yngvesson and Mather, 1983).

Dispute Transformation, the Articulation of Normative Order, and the Production of Law

The dispute transformation perspective (like other studies of dispute processing) has been faulted for its overemphasis on process to the exclusion of dispute outcome, a failure to examine the political significance of disputing, its individualistic focus, the methodological difficulty of identifying or measuring transformations, and its insufficient concern for institutions. By addressing these issues, however, researchers have redefined their work to include data that allow a more nuanced picture of the political aspects of disputing, focus attention on framing and language, and link routine disputing processes to broader questions of power, order, and law. Dispute transformation through the dialectic of narrowing and expansion thus bridges considerations of individual agency and social structure. When a secretary pursues a harassment claim against her supervisor, she is pursuing her individual dispute but at the same time her case is challenging the structure of gender relations in the local workplace. By providing the language of law and certain institutions for handling disputes, the broader political structure makes resources available to change local practices. Those legal and ideological resources, however, may be resisted by others

(e.g., through internal organizational processes) in ways that sustain existing power relationships. Most disputes are narrowed or rephrased in established categories of discourse. But, on occasion, new meanings are given to disputes – such as that described above in *Romagoza v. Garcia* (2002) and the subsequent rulings by a Florida Immigration Court (2014), or to the new meaning of marriage established by the *United States v. Windsor* (2013) litigation.

The relation between dispute transformation and political order has also been documented in studies of litigation in ancient Athens. The rhetoric and procedures of litigation, there were used to transform broad conflicts into two-party disputes, impose specialized roles and identities, create new political resources, and foreclose other options – thus, as Johnstone (1999) shows, promoting Athenian democracy. In sum, viewing how disputes emerge and transform contributes to understand both the micropolitics of dispute management and the macropolitics of disputing in society. The transformation perspective is useful cross-culturally, and across different historical periods, and helps to explain how law is created, maintained, and sometimes challenged or transformed.

See also: Class Actions, Legal; Dispute Resolutions in Economics; Law and Everyday Life; Law and Society: Development of the Field; Law as Constitutive; Law, Mobilization of; Legal Culture and Legal Consciousness; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Public Interest Law: The United States and Beyond; Sexual Harassment: United States and Beyond.

Bibliography

Bourdieu, Pierre, 1977. Outline of a Theory of Practice. Cambridge University Press, Cambridge, UK.

Cain, Maureen, 1979. The general practice lawyer and the client: towards a radical conception. International Journal of the Sociology of Law 7, 331.

Conley, John M., O'Barr, William M., 1997. Just Words: Law, Language, and Power. University of Chicago Press, Chicago, IL.

Coutin, Susan B., 2011. Falling outside: excavating the history of Central American Asylum Seekers. Law & Social Inquiry 36 (3), 569–596.

Curran, Barbara A., 1977. The Legal Needs of the Public. American Bar Foundation, Chicago.

Edelman, Lauren B., Erlanger, Howard S., Lande, J., 1993. Internal dispute resolution: the transformation of civil rights in the workplace. Law and Society Review 27, 497–534.

Engel, David M., 2013. Perception and decision at the threshold of tort law: explaining the infrequency of claims. De Paul Law Review 62 (2), 293–334.

Ewick, Patricia, Silbey, Susan S., 1998. The Common Place of Law: Stories from Everyday Life. University of Chicago Press, Chicago.

Felstiner, William L.F., Abel, Richard L., Sarat, Austin, 1981. The emergence and transformation of disputes: naming, blaming, claiming. Law and Society Review 15, 631–654.

Fiss, Owen M., 1984. Against settlement. Yale Law Journal 93, 1073.

Galanter, Marc, 1974. Why the "haves" comes out ahead: speculations on the limits of legal change. Law and Society Review 9, 95–160.

Genn, Hazel, 1999. Paths to Justice. Hart Publishing, Oxford.

Goldberg, Stephen B., Green, Eric D., Sander, Frank E.A., 1985. Dispute Resolution. Little, Brown, Boston.

Haltom, William, McCann, Michael, 2004. Distorting the Law: Politics, Media, and the Litigation Crisis. University of Chicago Press, Chicago.

Harrington, Christine B., 1985. Shadow Justice: The Ideology and Institutionalization of Alternatives to Court. Greenwood Press, Westport, CT.

Johnstone, Steven, 1999. Disputes and Democracy: The Consequences of Litigation in Ancient Athens. University of Texas Press, Austin, TX.

Kritzer, Herbert M., 2011. The antecedents of disputes: complaining and claiming. Onati Socio-legal Series 1 (6), 1–31.

Levy, Ariel, September 2013. The perfect wife. The New Yorker 30, 54–63.

MacKinnon, Catherine A., 1979. The Sexual Harassment of Working Women. Yale University Press, New Haven.

Marshall, Anna-Maria, 2003. Injustice frames, legality, and the everyday construction of sexual harassment. Law and Social Inquiry 28, 659–689.

Mather, Lynn, 1998. Theorizing trial courts: lawyers, policymaking, and tobacco litigation. Law and Social Inquiry 23, 897–940.

Mather, Lynn, 2009. Lawyers and solicitors separated by a common legal system: anti-tobacco litigation in the United States and Britain. In: Engel, David M., McCann, Michael (Eds.), Fault Lines: Tort Law as Cultural Practice. Stanford University Press, Stanford.

Mather, Lynn, McEwen, Craig A., Maiman, Richard J., 2001. Divorce Lawyers at Work: Varieties of Professionalism in Practice. Oxford University Press, New York.

Mather, Lynn, Yngvesson, Barbara, 1981. Language, audience, and the transformation of disputes. Law and Society Review 15, 775–821.

McCann, Michael, 1994. Rights at Work: Pay Equity Reform and the Politics of Legal Mobilization. University of Chicago Press, Chicago.

Menkel-Meadow, Carrie, 1985. For and against settlement: uses and abuses of the mandatory settlement conference. UCLA Law Review 33, 485.

Merry, Sally E., 1990. Getting Justice and Getting Even: Legal Consciousness among Working-Class Americans. University of Chicago Press, Chicago, IL.

Michelson, Ethan, 2006. The practice of law as an obstacle to justice: Chinese lawyers at work. Law and Society Review 40, 1–38.

Miller, Richard E., Sarat, Austin, 1981. Grievances, claims, and disputes: assessing the adversary culture. Law and Society Review 15, 525–565.

Munger, Frank, 1990. Special issue: longitudinal studies of trial courts. Law and Society Review 24 (2).

Murayama, Masayuki, 2012. Reconsidering the Japanese way of dispute resolution: what it is and why. In: Scheiber, Harry N., Ginsburg, Tom (Eds.), The Japanese Legal System: An Era of Transition, pp. 66–87.

Nader, Laura, Todd, Harry F. (Eds.), 1978. The Disputing Process: Law in Ten Societies. Columbia University Press, New York.

Pérez-Peña, Richard, Taylor, Kate, May 2014. Fight against sex assaults holds colleges to account. The New York Times 4, 1–25.

Pinkley, Robin L., Northcraft, Gregory B., 1994. Conflict frames of reference: implications for dispute processes and outcomes. Academy of Management Journal 37, 193–205.

Preston, Julia, April 2014. Salvadoran general accused of abuses should be deported, Miami judge says. The New York Times 12, A11.

Relis, Tamara, 2007. "It's not about the money!" a theory on misconceptions of plaintiffs' litigation aims. University of Pittsburgh Law Review 68, 701–746.

Rubin, Jonah, 2008. Adjudicating the Salvadoran Civil War: expectations of the law in Romagoza. Political and Legal Anthropology Review 31 (2), 264–285.

Richland, Justin, 2008. Arguing with Tradition. University of Chicago Press, Chicago, IL.

Santos, Boaventura de Sousa, 1977. The law of the oppressed: the construction and reproduction of legality in Pasargada. Law & Society Review 12, 5.

Sarat, Austin, Felstiner, William L.F., 1995. Divorce Lawyers and their Clients: Power and Meaning in the Legal Process. Oxford University Press, New York.

Sudnow, David, 1965. Normal crimes: sociological features of the penal code in a public defender's office. Social Problems 12, 255.

Tomasic, Roman, Feeley, Malcolm M. (Eds.), 1982. Neighborhood Justice. Longman, New York.

Vidmar, Neil, Schuller, Regina A., 1987. Individual differences and the pursuit of rights. Law and Human Behavior 11, 299–317.

Yngvesson, Barbara, 1988. Making law at the doorway: the clerk, the court, and the construction of community in a New England town. Law & Society Review 22 (3), 409–448.

Yngvesson, Barbara, Coutin, Susan B., 2006. Backed by papers: undoing persons, histories, and return. American Ethnologist 33 (2), 177–190.

Yngvesson, Barbara, Mather, Lynn, 1983. Courts, moots, and the disputing process. In: Boyum, K., Mather, L. (Eds.), Empirical Theories about Courts. Longman, New York, pp. 51–83.

Distance Education

Natalie B Milman, George Washington University, Oakton, VA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by H. Fritsch, volume 6, pp. 3781–3784, © 2001, Elsevier Ltd.

Abstract

Distance education or distance learning refers to the practice of delivering education and instruction to students not physically present but interacting with the instructor and the educational process remotely (usually by computers and the Internet these days.) Although often thought to be new development dependent on modern information processing technology, the idea of distance education is possibly centuries old. This article reviews the history, definition, current configuration, and effects of distance education.

History of Distance Education

Distance education has a long history possibly even dating back more than 285 years as evidenced in a *Boston Gazette* (published in the United States) advertisement by shorthand teacher Caleb Phillips. In the advertisement, he solicited "Persons in the Country desirous to Learn this Art, may by having the several Lessons sent weekly to them, be as perfectly instructed as those that live in Boston" (Battenberg, 1721 as cited in Holmberg, 1995: Paragraph 6). Years later in 1840, Sir Isaac Pitman, an English educator who invented shorthand, began mailing texts written in shorthand to students, which they would transcribe and mail back to him for correction (Schlosser and Simonson, 2009). These examples represent the first of the three documented phases of distance education, which are: (1) correspondence study (e.g., delivered via mail), (2) higher education institutions and analog mass media (e.g., delivered via television and cable), and (3) proliferation of distance education into most educational arenas via digital technologies, delivered primarily via the Internet (Simonson et al., 2011).

Over the years, definitions of distance education have changed, most frequently reflecting the transmission medium via which educational content was delivered to learners. In some cases, as Barbour et al. (2012) found in their K-12 Survey involving several countries, "there is wide interpretation of what online learning means, and what it looks like" (p. 7). However, it can be defined by several distinguishing characteristics. To understand these characteristics, it is important to examine the 'who,' 'what,' 'where,' 'when,' and 'how' of distance education, while also recognizing that, as Larreamendy-Joerns and Leinhardt (2006) contend about online education, "there are social, political, economic, and ethical assumptions and implications in what appear to be simple actions of design and instruction" (Larreamendy-Joerns and Leinhardt, 2006: p. 567) associated with these terms. It is important to note that the terms 'distance education,' 'online education,' 'e-learning,' and 'virtual education' are often used interchangeably to mean the same thing.

Therefore, it seems logical that the definition and its meaning will likely evolve over time.

Who Participates in Distance Education?

Logically, distance education involves both instructors and students; however, a defining characteristic is that traditional institutions (such as institutions of higher education), nontraditional institutions (e.g., businesses, corporations), organizations/societies (e.g., International Society for Performance Improvement) (Schlosser and Simonson, 2009), and governments offer distance education. More recently, with the emergence of massive open online courses (MOOCs), other organizations have become involved in distance education offerings, although the majority of those who participate in distance education affiliate in some way with higher education institutions. Additionally, distance education, often referred to as 'virtual schooling,' has been growing in P-12 schools, although most distance education at this educational level involves blended learning (Barbour et al., 2012).

The US Department of Education's National Center for Education Statistics (NCES), several researchers (e.g., Allen and Seaman, 2008, 2010, 2011, 2013, 2014) and organizations (e.g., International Association for K-12 Online Learning) have tracked participation in distance education in institutions of higher education and in public K-12 schools. Surveys of distance education participation in the United States illustrate that although it has grown exponentially over the past decade, the unfettered growth has slowed down. For example, according to Allen and Seaman (2014), "the online enrollment growth rate of 6.1 percent is the lowest recorded in this report series" in institutions of higher education (p. 4).

What Is Distance Education?

In the twenty-first century, distance education involves students learning in a location(s) other than where the instructor is via digital technologies, synchronously and/or asynchronously. Allen and Seaman (2013) provide several different types of distance education courses, as **Table 1** outlines.

Distance education can be equated by the 'Online' category in **Table 1**; however, others may contend that even 'blended/hybrid' courses are a form of distance education, too.

Table 1 Prototypical course classifications for online learning

Proportion of content delivered online (%)	*Type of course*	*Typical description*
0	Traditional	Course where no online technology used – content is delivered in writing or orally.
1–29	Web facilitated	Course that uses web-based technology to facilitate what is essentially a face-to-face (F2F) course. May use a course management system (CMS) or web pages to post the syllabus and assignments.
30–79	Blended/hybrid	Course that blends online and F2F delivery. Substantial proportion of the course is delivered online, typically uses online discussions, and typically has a reduced number of F2F meetings.
80–100	Online	A course where most or all of the content is delivered online. Typically have no face-to-face meetings.

Allen, I.E., Seaman, J., 2013. Changing the Course: Ten Years of Tracking Online Education in the United States. Babson Survey Research Group and Quahog Research Group, p. 7.

Where Does Distance Education Occur?

Another unique characteristic of distance education is its 'location.' Historically, distance education involved some kind of 'geographic separation' (Simonson, 2009; as cited in Simonson et al., 2011: p. 126) between the instructor and his/her students and students themselves. However, with the advent of the Internet, it is possible for there to be little or even no geographic separation between the instructor and students and/or between students because two individuals can participate in distance education sitting right next to each other in the same location, but on two different computers. Twenty-first century distance education happens anywhere – anywhere an individual can access the course/distance education experience materials – which is usually online (via a computer connected to the Internet). Moreover, in distance education an instructor may incorporate some face-to-face (F2F) instruction along with distance education. This 'blending' of distance and F2F instruction is often referred to as blended or hybrid education or learning.

When Does Distance Education Occur?

Distance education consists of instruction that occurs synchronously, asynchronously, or both synchronously and asynchronously. Synchronous courses require the learners and instructor(s) to 'meet' online at the same time for exchanges of information via audio or web conferences, screencasts, online chats, interactive whiteboards, sharing of applications, instant messaging, and even texting. Asynchronous interactions can occur using all of these methods, in addition to others such as discussion boards, blogs, wikis, and many other tools. A differentiating characteristic is that learners are not required to meet 'live' with their instructor or peers. Moreover, it is feasible for the learning materials to be disseminated 24/7 via the Internet and across any time zone in the world – the distance education classroom is always 'open' with materials available to students any time of the day, anywhere, as long as there are no restrictions on accessing materials.

How Does Distance Education Occur?

Two defining aspects of how distance education occurs involve technology and interaction. Often there is overlap between these two because technology fosters student-to-student, student-to-content, and student-to-instructor interaction. Technology is critical because it is through technology that instructors and students can access, engage, and participate in distance education. Moreover, various technologies can serve as both the medium (e.g., Internet-based course available in a course management system) and/or mechanism or tool (e.g., use of screencasting software to develop screencasts) for developing distance education. For example, a course management system or learning management system may be utilized by an instructor to house and deliver the instruction to students. They might also use web conferencing tools embedded within this system or an outside vendor to offer synchronous sessions with students and/or record a presentation explaining content.

Interaction, "commonly understood as actions among individuals" (Abrami et al., 2011: p. 86), is another important characteristic of distance education because students learn through student–student, student–instructor, and student–content interaction (Abrami et al., 2011; Bernard et al., 2009; Moore, 1989). Bernard et al. (2009) found that interaction positively affects student learning. However, they caution that instructors should not focus on increasing the quantity of interactions, but rather focus on increasing the quality of interactions.

Why Does Distance Education Occur?

There are many reasons why institutions, organizations, instructors, and students participate in distance education. These may range from the potential to reduce costs (e.g., learners do not need to travel to one location to learn – note that some argue quality distance education requires an investment, so the reduction in cost may not be realized for several years), the ability to increase flexibility of when learners learn (e.g., students can determine their own learning schedules – they do not have to learn only on the instructor's schedule), the potential to involve expert instructors from a variety of locations (i.e., instructors do not have to be from one's institution/organization nor do they have to be physically present), and the possibility to make education accessible to those who cannot participate (e.g., because they cannot move to the institution).

Distance Education in Higher Education

Distance education in higher education institutions started in the nineteenth century. One of the first documented instances was the University of Chicago's correspondence courses offered by William Rainey Harper in 1883 (Distance Learning, n.d.). Since then, distance education in higher education institutions has evolved and as noted previously, can be categorized by three major periods. However, Larreamendy-Joerns and Leinhardt (2006) view distance education through three different historical lenses – via the promise of democratization, the tension between professional education and liberal arts education, and the issue of instructional quality. Their work provides a different lens for understanding distance education.

Massive Open Online Courses (MOOCs)

MOOCs are non-credit-bearing courses typically offered for free by numerous universities and organizations to anyone interested in enrolling in these courses. Distinct features of MOOCs are that they are designed for 'massive' participation – thousands, even hundreds of thousands can participate in a MOOC – and students do not need to be enrolled in or even affiliated with the sponsoring organization offering the MOOC to register. Some MOOCs offer certificates upon successful course completion for a fee, whereas others do not. Often MOOCs are taught in F2F settings as credit-bearing courses at an institution with students enrolled in degree programs. The F2F lectures are then recorded and disseminated online to the MOOC students. These lectures typically involve a professor offering a weekly lecture, offered for a varying number of weeks depending on the course, accompanied by a multimedia presentation with visuals and questioning from students in the audience (the F2F, paying students). The lectures are recorded and made available by a MOOC provider. Therefore, in a MOOC, it is not unusual to find students who have paid to take a course for credit and in a degree program studying parallel (with no interaction at all) or alongside (with some interaction, for instance in online discussions) students enrolled for free. However, many MOOCs are designed only as a MOOC and not also as a F2F course. Currently there are no standards as to how MOOCs are offered – it is unclear if there ever will be.

Who Is Offering MOOCs?

MOOCs are being offered by organizations such as Coursera (https://www.coursera.org/), edX (https://www.edx.org/), 2u (http://2u.com/), and Udacity (http://www.udacity.com/). These organizations consist of consortia of different universities who support the provider in some way. However, in the future, it will not be surprising if corporate and government organizations consider offering their own MOOCs, too, for many of the same reasons universities are exploring the possibilities of MOOCs.

Interest and enrollments in MOOCs are quite high (e.g., over 100 000 students enrolled in a MOOC offered at Stanford University) and will likely grow; however, persistence in MOOCs shows that it is quite low (Rosen, 2012), with small percentages of the nonpaying students actually completing courses. It is too soon to tell why and whether or not this will continue to be the case with MOOCs.

Distance Education in P-12 Schools

The iNACOL is one of the most active advocates for distance education in P-12 schools. Its mission is "to ensure all students have access to a world-class, education and quality blended and online learning opportunities that prepare them for a lifetime of success" (iNACOL, 2013: p. 1). As a result of its mission, iNACOL has sponsored numerous reports on distance learning in P-12 settings. For instance, in 2011, it sponsored a survey of several countries around the world to gain a better understanding of distance education in P-12 schools (see Barbour et al., 2012). The researchers found four trends:

1. Blended and online choices are most available to students in urban areas from developed countries (p. 10).
2. Growth in digital learning stems from shared authority between local schools and national governments (p. 10).
3. Specialized teacher training is not required, but is encouraged and available (p. 12).
4. Blended learning is occurring with much greater frequency than online learning (p. 13).

In the United States, distance education in P-12 schools is growing as iNACOL (2013) has found. For example, during the 2013-2014 academic year, "25 states had state virtual schools operating … and 29 states and Washington, DC had statewide full-time online schools operating" (p. 1). Additionally, several states such as Michigan, Florida, and Virginia require high school students to complete one online course to graduate (Sheehy, 2012).

See also: Adult Education and Lifelong Learning: The US Experience and Beyond; Computer Mediated Communication; Computer-Assisted Instruction; Educational Sociology; Internet and Culture; Internet and Social Media: Anthropological Aspects; Lifelong Learning and Its Support with New Technologies; Tangible User Interfaces in Learning and Education.

Bibliography

Abrami, P.C., Bernard, R.M., Bures, E.M., Borokhovski, E., Tamim, R.M., 2011. Interaction in distance education and online learning: using evidence and theory to improve practice. Journal of Computing in Higher Education 23, 82–103.

Allen, I.E., Seaman, J., 2014. Grade Change: Tracking Online Education in the United States. Babson Survey Research Group and Quahog Research Group.

Allen, I.E., Seaman, J., 2013. Changing the Course: Ten Years of Tracking Online Education in the United States. Babson Survey Research Group and Quahog Research Group.

Allen, I.E., Seaman, J., 2010. Class Differences: Online Education in the United States, 2010. Babson Survey Research Group.

Allen, I.E., Seaman, J., 2011. Going the Distance: Online Education in the United States, 2011. Babson Survey Research Group and Quahog Research Group.

Allen, I.E., Seaman, J., 2008. Staying the course: online education in the United States, 2008. Babson Survey Research Group.

Barbour, M., Brown, R., Waters, L.H., Hoey, R., Hunt, J.L., Kennedy, K., Ounsworth, C., Powell, A., Trimm, T., 2012. Online and Blended Learning: A Survey of Policy and Practice of K-12 Schools Around the World. Retrieved from: http://www.inacol.org/wp-content/uploads/2012/11/iNACOL_IntnlReport2011.pdf.

Bernard, R.M., Abrami, P.C., Borokhovski, E., Wade, C.A., Tamim, R.M., Surkes, M.A., Bethel, E.C., 2009. Review of Educational Research 79 (3), 1243–1289.

Bernard, R.M., Abrami, P.C., Lou, Y., Borokhovski, E., Wade, C.A., Wozney, L., Wallet, P.A., Feset, M., Huang, B., 2004. How does distance education compare with classroom instruction? A meta-analysis of the empirical literature. Review of Educational Research 74 (3), 379–439.

Distance Learning, n.d. In: Encyclopedia Britanica. Retrieved from: http://www.britannica.com/EBchecked/topic/1482174/distance-learning.

Holmberg, B., 1995. The evolution of the character and practice of distance education. Open Learning 47–53. Retrieved from: http://www.c3l.uni-oldenburg.de/cde/found/holmbg95.htm.

International Association for K12 Online Learning, October 2013. Fast Facts about Online Learning. Author, Vienna, VA. Retrieved from: http://www.inacol.org/wp-content/uploads/2013/11/iNACOL-Fast-Facts-About-Online-Learning-October-2013.pdf.

International Association for K12 Online Learning, 2011. The Online Learning Definitions Project. Author, Vienna, VA.

Johnson, L., Adams Becker, S., Cummins, M., Estrada, V., Freeman, A., Ludgate, H., 2013. NMC Horizon Report: 2013 Higher Education Edition. The New Media Consortium, Austin, TX. Retrieved from: http://www.nmc.org/publications/2013-horizon-report-higher-ed.

Johnson, L., Adams, S., Cummins, M., 2012. NMC Horizon Report: 2012 K12 Edition. The New Media Consortium, Austin, TX. Retrieved from: http://www.nmc.org/publications/2012-horizon-report-k12.

Larreamendy-Joerns, J., Leinhardt, G., 2006. Going the distance with online education. Review of Educational Research 76 (4), 567–604.

Moore, M.G., 1989. Three types of interaction. American Journal of Distance Education 3 (2), 1–6.

Rosen, R.J., 12 July 2012. Overblown-claims-of-failure watch: how not to gauge the success of online courses. The Atlantic. Retrieved from: http://www.theatlantic.com/technology/archive/2012/07/overblown-claims-of-failure-watch-how-not-to-gauge-the-success-of-online-courses/260159/.

Schlosser, L.A., Simonson, M.R., 2009. Distance Education: Definition and Glossary of Terms, third ed. IAP/Information Age, Greenwich, CT.

Sheehy, K., 24 October 2012. States, districts require online ed for high school graduation. In: Education Week. Retrieved from: http://www.usnews.com/education/blogs/high-school-notes/2012/10/24/states-districts-require-online-ed-for-high-school-graduation.

Simonson, M., 2009. Distance Learning. The 2009 book of the year. Encyclopaedia Britannica, Chicago, IL, p. 231.

Simonson, M., Schlosser, C., Orellana, A., 2011. Distance education research: a review of the literature. Journal of Computing in Higher Education 23, 124–142.

Tallent-Runnels, M.K., Thomas, J.A., Lan, W.Y., Cooper, S., 2006. Teaching courses online: a review of the research. Review of Educational Research 76 (1), 93–135.

U.S. Department of Education, Office of Planning, Evaluation and Policy Development, 2009. Evaluation of Evidence – Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies. Washington, DC. www.ed.gov/about/offices/list/opepd?ppss?reports.html.

Webley, K., 4 September 2012. MOOC brigade: will massive, open online courses revolutionize higher education? Time. Retrieved from: http://nation.time.com/2012/09/04/mooc-brigade-will-massive-open-online-courses-revolutionize-higher-education/.

Zawacki-Richter, O., Backer, E.M., Vogt, S., 2000. Review of distance education research (2000 to 2008): analysis of research areas, methods, and authorship patterns. International Review of Research in Open and Distance Learning 10 (6), 21–50.

Distributions, Statistical: Approximations

John P Lehoczky, Carnegie Mellon University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp. 3784–3787, © 2001, Elsevier Ltd, with an updated Bibliography section supplied by the Editor.

Abstract

This article describes a set of situations in which a random variable with a complicated probability distribution can be approximated well by another, better understood and studied probability distribution. The first example is the normal distribution which approximates the distribution of many random variables arising in nature. This approximation is based on the central limit theorem. Generalizations of this theorem provide a second set of approximations called the stable laws. A third example arises when one is considering extremes, or the maximum of a set of random variables. Here, the family of extreme value distributions is shown to provide accurate approximations. Moreover, the excesses over a high level can be approximated by a generalized Pareto distribution. The final example involves the distribution of the times between events, such as system failures. Under certain conditions, these times can be approximated well by an exponential distribution.

Every random variable has its own particular distribution. It is, however, very useful if one can use a well-studied distribution as an approximation. This article describes a set of situations in which a random variable with a complicated probability distribution can be approximated well by another, better understood and the studied probability distribution. The first example is the normal distribution which approximates the distribution of many random variables arising in nature. This approximation is based on the central limit theorem. Generalizations of this theorem provide a second set of approximations called the stable laws. A third example arises when one is considering extremes, or the maximum of a set of random variables. Here, the family of extreme value distributions is shown to provide accurate approximations. Moreover, the excesses over a high level can be approximated by a generalized Pareto distribution. The final example involves the distribution of the times between events, such as system failures. Under certain conditions, these times can be well approximated by an exponential distribution. Consequently, the corresponding point process (or event history) can be approximated by a Poisson process.

Normal Approximations

The normal or Gaussian distribution has a probability density function (p.d.f.) given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad -\infty < x < +\infty$$

where the parameter μ is the mean or median and σ is the standard deviation. The standard normal distribution corresponds to the special case in which $\mu = 0$ and $\sigma = 1$, for which the p.d.f. is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right) \quad -\infty < x < +\infty$$

The normal distribution serves as an excellent approximation to many random variables that arise in nature. The reason for this lies in the central limit theorem, the simplest version of which states:

Theorem (Central Limit Theorem)

Suppose $\{X_r\}_{n=1}^{\alpha}$ is a sequence of independent identically distributed (i.i.d.) random variables with cumulative distribution function F, mean μ, and variance $\sigma^2 < \infty$. Then

$$Z_n = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma\sqrt{n}} \Rightarrow \Phi$$

where Φ represents the c.d.f. of a standard normal distribution and $\Rightarrow$ denotes weak convergence.

This theorem says that the standardized partial sums of the $\{X_r\}_{n=1}^{\infty}$ sequence converge weakly to a standard normal distribution.

This theorem has many generalizations which allow for the $\{X_r\}_{n=1}^{\infty}$ to have different distributions (so long as each random variable has finite variance and contributes only a small part to the sum) and to have a dependency structure (so long as the correlation among the variables diminishes rapidly enough). The reader should consult Distributions, Statistical: Special and Continuous for additional information about the normal distribution. The textbook by DeGroot (1986) offers examples of the application of the central limit theorem, some extensions of the basic theorem, and discussions of weak convergence, also known as convergence in distribution. The reader should consult the text by Billingsley (1995) for the precisely stated theorems.

The central limit theorem indicates that suitably standardized sums of independent random variables will have an approximate normal distribution. Thus, any random variable which arises as the sum of a sufficiently large number of small random components can be modeled accurately by a normal distribution. This is especially appropriate for error distributions, where the error term is the sum of a number of different sources of error. The question of how large n must be for the approximation to be accurate depends on two factors: (1) the degree of accuracy required and (2) the underlying distribution F. A common 'rule of thumb' is that the normal approximation

can be safely used when $n \geq 30$ as long as F is not highly skewed and a high degree of accuracy is not required. The normal distribution is commonly used as a model for many statistical problems especially in linear models; see Linear Hypothesis and Linear Hypothesis: Regression (Basics) for details.

Stable Distributions

In this section, we present an extension of the central limit theorem. When $\{X_r\}_{n=1}^{\infty}$ is a sequence of i.i.d. random variables, each having a finite variance, then the partial sums of these variables can be normalized to converge weakly to a normal distribution. When the underlying random variables do not have a finite variance, then normalization of the partial sums can still lead to weak convergence to a nondegenerate limiting distribution; however, the limiting distribution will no longer be normal, but will instead be a *stable* distribution.

Definition (Stable Distributions)

A distribution F is stable if given any three i.i.d. random variables, X_1, X_2, X_3, with distribution F and any constants c_1, c_2, there exists constants a and b (depending upon c_1, c_2) such that $c_1X_1 + c_2X_2$ has the same distribution as $a + bX_3$. If $a = 0$, the F is *symmetric stable*.

This means that any linear combination of two i.i.d. random variables with a common stable distribution can be transformed linearly to have the same stable distribution. This definition can be used to prove that if $\{X_r\}_{n=1}^{\infty}$ are i.i.d. with a stable distribution F, then there exist constants a_n, b_n such that $\frac{\sum_{i=1}^{n} X_i - a_n}{b_n}$ also has c.d.f. F. The symmetric stable distributions are most easily described by their characteristic functions.

Theorem (Characteristic Function of a Symmetric Stable Law)

A standard symmetric stable distribution has characteristic function $\phi(s) = e^{-|s|^{\alpha}}$ for $0 < \alpha \leq 2$. Two special cases are the normal (0, 1) distribution ($\alpha = 2$) and the standard Cauchy distribution ($\alpha = 1$). The following theorem indicates that the non-degenerate stable laws represent all possible limiting distributions for sums of i.i.d. random variables.

Theorem

Suppose $\{X_n\}_{n=1}^{\infty}$ is a sequence of i.i.d. random variables with cumulative distribution function F, and suppose that there are sequences of constants, $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ for which $\frac{\sum_{i=1}^{n} X_i - a_n}{b_n}$ a non-degenerate limiting c.d.f. Then G is a nondegenerate stable law.

The family of stable distributions represents a wider class of possible limits of sums of i.i.d. random variables. The distributions other than the normal have heavier tails than the normal distribution. Informally, a symmetric distribution with exponent $\alpha < 2$ has finite moments of order less than α. For example, the Cauchy distribution ($\alpha = 1$) does not have a first moment, but all moments less than 1 are finite. The stable distributions have been used to model certain economic variables and the return of financial assets such as stocks or currency exchange rates. See Distributions, Statistical: Special and Continuous for additional information on the Cauchy distribution. The reader should consult Billingsley (1995), Embrechts et al., (1997), or Lamperti (1966) for more detailed information on stable distributions and the theorems underlying this more general central limit problem.

Extreme Value Distributions

Another type of approximation occurs when one is studying the maximum of a set of random variables. This situation is particularly relevant in the casualty insurance industry, where interest centers on the largest claims. These situations also arise in engineering and environmental problems, where a structure such as a dam or building must withstand all forces, including the largest. Suppose $\{X_n\}_{n=1}^{\infty}$ is a sequence of i.i.d. random variables with common distribution function F. Suppose that x^* denotes the rightmost support point of F, the smallest x for which $F(x) = 1$, or $+\infty$ if no such x exists. We consider $\{M_n\}_{n=1}^{\infty}$ where $M_n = \max_{\{1 < i < n\}} X_i$, the largest among the first n random variables in the sequence. Fisher and Tippett (1928), a classic paper, showed that if there are sequences of constants $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\alpha}$ for which $\frac{M_n - a_n}{b_n} \Rightarrow H$, a nondegenerate limiting c.d.f., then H takes on one of three possible forms:

1. The Gumbel distribution

$$H(x) = \exp\left(-e^{-x}\right), \quad -\infty < x < +\infty$$

This limiting form arises when $x^* = \infty$ and the underlying distribution F has a right-hand tail which decreases at an exponential rate. Specifically, if there exists a constant $b > 0$ for which $\lim_{x \to \infty} \exp(x)(1 - F(x)) \to b$, then $M_n - \log(nb)$ converges to the Gumbel distribution as $n \to \infty$.

2. The Fréchet distribution

$$H(x) = \begin{cases} \exp\left(-x^{-a}\right) & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

where $\alpha > 0$. This limiting form arises when $x^* = \infty$ and the right tail of the underlying distribution F has a tail which decays at a polynomial rate. Specifically, if $\lim_{x \to \infty} x^{\alpha}(1 - F(x)) \to b$, for some constant $b > 0$ and $\alpha > 0$, then $M_n/(bn)^{1/\alpha}$ converges to the Fréchet distribution as $n \to \infty$.

3. The Weibull distribution

$$H(x) = \begin{cases} \exp\left(-(-x)^{\alpha}\right) & \text{if } x \leq 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

where $\alpha > 0$. The above distribution is concentrated on $(-\infty, 0)$, and it is the negative of the traditional Weibull distribution. This limiting form arises when $x^* < \infty$, i.e., the underlying distribution F is bounded above by a constant. Specifically, assume $\lim_{x \to x^*} (x^* - x)^{-\alpha}(1 - F(x)) \to b$ for some constants $b > 0$ and $\alpha > 0$, then $(nb)^{1/\alpha}(M_n - x^*)$ converges to the Weibull distribution defined above as $n \to \infty$. The reader should consult Distributions, Statistical: Special and Continuous for additional information on the Weibull distribution.

Fisher and Tippett (1928) proved that these are the only possible limiting forms. It is, however, important to point out that, for some underlying F, there may be no set of constants $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ for which $\frac{M_n - a_n}{b_n} \Rightarrow H$. The conditions under which this convergence occurs are complicated, and the interested reader is referred to Embrechts et al., (1997). Lamperti (1966) also provides a nice treatment of extreme value theory.

The three extreme value distributions given above describe the limiting of the maximum of a random sample. There is a second random quantity that is related to the maximum and is interesting in it own right. Suppose we begin with a random variable X with c.d.f. F and consider the conditional excess distribution of X, $X-u$, given that it exceeds u. Specifically, consider $F_u(x) = p(X - u \leq \ |X > u)$. One could also consider the expected excess, conditional on the random variable exceeding u, $e(u) = E(X - u|X > u)$. This distribution and expectation arise in engineering and insurance problems, where the quantity above a certain high limit is of concern, such as the insurance claim above a large deductible amount.

The standard generalized Pareto distribution is a family of c.d.f.s indexed by a parameter $\theta \in (-\infty, +\infty)$ given by

For $\theta < 0$,

$$G_\theta(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - (1 + \theta x)^{-1/\theta} & \text{if } 0 \leq x \leq -1/\theta \\ 1 & \text{if } x \geq -1/\theta \end{cases}$$

For $\theta = 0$,

$$G_o(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - \mathrm{e}^{-x} & \text{if } x \geq 0 \end{cases}$$

For $\theta > 0$,

$$G_\theta(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - (1 + \theta x)^{-1/\theta} & \text{if } x \geq 0 \end{cases}$$

The family can be enlarged by adding a location and scale parameter.

For a given value of θ, the probability distribution of F_u approaches that of a generalized Pareto distribution as $u \to x^*$. Thus, the generalized Pareto distribution can be used to approximate the probability distribution of excesses over a high limit. The mean excesses can also be modeled using a generalized Pareto distribution.

The reader should consult Embrechts et al., (1997), a comprehensive treatment of extreme value theory and for exact conditions and statements of these approximation results.

Times between Events in a Point Process

A random point process is a sequence of events (or points) that occur in time. Common examples of point processes are event histories, the failure times of systems, arrivals of customers to a service facility or the firing of neurons in the human nervous system. Clearly, modeling such point process phenomena can be very complicated and detailed. There is, however, an approximate result that is widely useful in such modeling. Specifically, for an important class of point processes, a reasonable approximation is that the times between events are i.i.d. random variables with an exponential distribution, a p.d.f. given by:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

with $\lambda > 0$. The parameter λ measures the rate at which the events occur and is measured in units of events per unit time. Such a point process is called a Poisson (λ) process. Poisson processes have three important properties: (1) the number of events in an interval of length L is a random variable with a Poisson (λL) distribution; (2) the number of events in non-overlapping time intervals are independent random variables; and (3) the probability of exactly one event in an interval of length h, divided by h, converges to λ as $h \to 0$. Hence events cannot occur simultaneously in a Poisson process. See Distributions, Statistical: Special and Continuous for additional information on the exponential distribution.

Now, suppose that the point process is composed of the superposition of a large number of independent point processes, each individually with insignificant intensity function. Then the superposed (aggregated) point process converges (as the number of components superposed increases) to a Poisson process. The reader should consult the text by Karlin and Taylor (1975) for a precise statement of this theorem. This situation is very realistic for system failures in which the system is composed of many different parts, the failure of any one of which leads to system failure. Each component has its own failure distribution, and the time to the next failure will be the minimum of these failure times, which will converge to an exponential distribution as the number of components increases. A similar situation occurs with aggregate customer arrival processes, where the aggregate process is the superposition of a large number of different arrival processes. As the number of different arrival sources increases, the limiting aggregate arrival process will become Poisson. The reader should consult Queues to learn about the use of the exponential distribution, and the Poisson process as a model for customer arrival processes.

Further Reading

The reader should consult a basic textbook on probability such as DeGroot (1986) for information on many of the distributions mentioned in this article, for a description of the basic central limit theorem and examples of its use, and for some more general results. Billingsley (1995) provides a more comprehensive and technically detailed treatment of the central limit theorem, the stable laws and the extreme value distributions. Finally, the book by Embrechts et al., (1997) gives a comprehensive treatment of the extreme value distributions and associated approximation theorems.

Bibliography

Billingsley, P., 1995. Probability and Measure, third ed. Wiley, New York.
DeGroot, M.H., 1986. Probability and Statistics, second ed. Addison-Wesley, Reading, MA.
Embrechts, P., Kluppelberg, C., Mikosch, T., 1997. Modeling Extremal Events. Springer-Verlag, Berlin.

Fisher, R.A., Tippett, L.H.C., 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proceedings of the Cambridge Philosophical Society 24, 180–190.

Johnson, Norman L., Kemp, Adrienne W., Kotz, Samuel, 2005. Univariate Discrete Distributions. John Wiley & Sons, Hoboken, NJ.

Karlin, S., Taylor, H., 1975. A First Course in Stochastic Processes, second ed. Academic Press, New York.

Lamperti, J., 1966. Probability. W. A. Benjamin, New York.

O'Hagan, Anthony, Forster, Jonathan, Kendall, Maurice G., 2004. Bayesian Inference. Arnold, London, UK.

Sancetta, Alessio, Satchell, Stephen, 2004. The Bernstein Copula and its applications to modeling and approximations of multivariate distributions. Econometric Theory 20 (03), 535–562.

Distributions, Statistical: Special and Continuous

John P Lehoczky, Carnegie Mellon University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp. 3787–3793, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

This article introduces a set of specific probability distribution of random variables that arise frequently in modeling social science phenomena. The article focuses on the absolutely continuous case in which a probability density function can be defined. The special continuous distributions presented include the uniform, beta, normal, exponential, gamma, Weibull, Chi-square, Student *t*, *F*, and Pareto. The concept of a failure rate function is also introduced. Multivariate probability distributions are introduced, and the bivariate normal distribution is presented.

Continuous probability distributions are used widely in probability and statistics when the underlying random phenomenon is measured on a continuous scale. Common examples include the time until an event occurs, such as the failure time of a component or system, or most human physical dimensions such as height or weight. In some cases, the underlying random quantity has a discrete set of possible values. If that set is large, it may be possible to rescale those values and treat them as continuous. Continuous distributions arise frequently in the social sciences, for example in IQ or educational testing settings in which the scale is designed to yield a normal distribution of performance. Economic measurements such as income levels are often modeled using continuous distributions. Many observed distributions do not fit any of the standard distributions presented in this article (for example, they may be skewed when compared with the normal distribution); however, it may be possible to apply a transformation to the variable in question to make it fit one of the standard distributions better. In this article, we begin by describing univariate probability distributions, then specialize to the absolutely continuous case. The concept of a failure rate function as an alternative way to describe an absolutely continuous distribution is also introduced. The next section introduces special continuous distributions, including the uniform, beta, normal, exponential, gamma, Weibull, Chi-square, Student *t*, *F*, and Pareto. The entry concludes with a definition of multivariate distributions, and the bivariate normal distribution is presented.

Definitions and General Properties

A univariate random variable X has a probability distribution which is defined by its cumulative distribution function (c.d.f.),

$$F(x) = P(X \le x), \quad -\infty < x < +\infty$$

A c.d.f. has several properties:

a. $F(x)$ is non-decreasing, i.e., $F(x) \le F(y)$ if $x < y$;
b. $\lim_{x\to-\infty} = F(-\infty) = 0$;
c. $\lim_{x\to+\infty} = F(+\infty) = 1$;
d. F is right continuous, i.e. $\lim_{h\to 0} F(x+h) = F(x)$; and
e. $P(a < X \le b) = F(b) - F(a)$, $-\infty \le a < b \le +\infty$.

When the c.d.f. is a step function, $\{(x_i, p_i)\}_{i=1}^{\infty}$, with a jump of height p_i at location x_i, $1 \le i < \infty$, then the associated probability distribution is called a *discrete distribution*. If the c.d.f. is a continuous function, the associated probability distribution is called a *continuous distribution*. When the c.d.f. has jumps satisfying $0 < \sum_{i=1}^{\infty} p_i < 1$, the distribution is said to be *mixed*.

For continuous distributions, we can define the derivative, $f(x) = dF(x)/dx$. The function f is called the probability density function (p.d.f.). When the c.d.f. is absolutely continuous, then one can integrate the p.d.f. to recover the c.d.f., that is,

$$f(x) = \int_{-\infty}^{x} f(u)du, \quad -\infty < x < \infty$$

One can calculate the probability of an event by integrating the density function; for example, if the event is an interval $[a, b]$, then

$$f(x < X < b) = \int_{a}^{b} f(x)dx$$

If X has a continuous probability distribution, then $P(X = x) = 0$ for any value $-\infty < x < +\infty$. The interpretation of the p.d.f. is that $P(x \le X \le x + h) \approx f(x)h$ for small $h > 0$. That is, the probability a random variable with an absolutely continuous probability distribution takes a value in a small interval of length h is given by the probability density times the length of the interval.

For absolutely continuous distributions, the c.d.f. F and the p.d.f. f are equivalent in that one can be derived from the other. An absolutely continuous probability distribution might also be described by its failure rate function, also known as the hazard function. This function is often denoted by $\lambda(x)$ and is defined by:

$$\lambda(x) = \frac{f(x)}{1 - F(x)}$$

The failure rate function arises naturally in specifying the reliability of a system or in mortality tables. This function gives the probability that a system will fail (person will die) in the next small interval of time conditional on it not yet having failed (died) up to the present time. The failure rate function satisfies

$$p(X \le x + h | X \ge x) \approx \lambda(x)h$$

The c.d.f., $F(x)$, can be derived from $\lambda(x)$ from the formula

$$F(x) = 1 - \exp\left(-\int_{-\infty}^{x} \lambda(u)\mathrm{d}u\right)$$

where $\lambda(x)$ can be any function satisfying $\lambda(x) \geq 0, \int_{-\infty}^{x} \lambda(u)\mathrm{d}u < \infty, -\infty < x < \infty$ and $\int_{-\infty}^{+\infty} \lambda(u)\mathrm{d}u = +\infty$. See Survival Analysis: Overview for additional information about the use of the failure rate function.

The c.d.f., p.d.f., or the failure rate function give complete information about the distribution of a random variable having an absolutely continuous distribution. It is, however, often useful to provide a few simple characteristics or features of the distribution. Typically, one specifies the mean and the variance of the distribution. For absolutely continuous distributions, these can be computed from the p.d.f. by

$$E(X) = \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x,$$

$$Var(X) = \int_{-\infty}^{+\infty} x^2 f(x)\mathrm{d}x - (E(X))^2$$

Individual Continuous Distributions

This section describes some of the most important univariate continuous distributions.

Uniform Distribution

This is a basic distribution with two parameters, a and b denoting the endpoints of an interval. The p.d.f. is given by:

$$f(x) = \begin{cases} \dfrac{1}{b-a} f & -a \leq x \leq b, \\ 0 & \text{otherwise} \end{cases}$$

If X has a Uniform (a, b) distribution, then for any interval (c, d) satisfying $a \leq c < d \leq b$, $P(c < X < d) = (d-c)/(b-a)$. The mean and variance are given by

$$E(X) = \frac{(b+a)}{2}, \quad Var(X) = \frac{(b-a)^2}{12}$$

The Uniform (0, 1) distribution is often referred to as the standard uniform distribution. If U is a random variable with standard uniform distribution, then $X = a + (b-a)U$ has Uniform (a, b) distribution. Thus, any uniform random variable can be constructed from a standard uniform random variable. This property of using standard uniform random variables to construct other random variables with specific distributions is quite general. The method is called the probability integral transform. Specifically, suppose X is a random variable with a c.d.f. given by F. Let $F^{-1}(p)$, $0 < p < 1$ denote the inverse of F defined by

$$F^{-1}(p) = \inf\{(y|F(y) \geq p)\}$$

It can be shown that if U has Uniform (0,1) distribution, then $X = F^{-1}(U)$ has c.d.f. given by F. This property is especially important in simulation studies in which one must generate a sequence of random variables with some prescribed probability distribution. If F^{-1} can be computed, then one need only generate a sequence of standard Uniform (0, 1) random variables, plug each of them into F^{-1}, and the result will be a sequence with the required distribution.

Beta (α, β) Distribution

The Uniform (0, 1) distribution is a special case of a two parameter family called the Beta (α, β) distribution. It has a p.d.f. given by:

$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta} & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha > 0$, $\beta > 0$, and $\Gamma(a) = \int_0^{+\infty} x^{a-1}\mathrm{e}^{-x}\mathrm{d}x$.

The Beta (1, 1) distribution is the same as the Uniform (0, 1) distribution. The mean and variance of a random variable with Beta (α, β) distribution are given by

$$E(X) = \frac{\alpha}{\alpha+\beta}, \quad Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

The Beta distribution offers a very flexible two parameter family of distributions for random variables taking values between 0 and 1. Consequently, it can often be used to model percentage or fractional quantities.

Normal (μ, σ^2) Distribution

One of the most important probability distributions is the normal or Gaussian distribution. Its p.d.f. is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad -\infty < x < +\infty$$

The parameter μ is the mean, while σ^2 is the variance. Since the p.d.f. is symmetric about μ, this parameter is also the median.

The standard normal distribution corresponds to the case in which $\mu = 0$ and $\sigma = 1$. It has p.d.f. given by

$$f(x) = \frac{1}{\sqrt{2x}} \exp\left(-\frac{1}{2}x^2\right), \quad -\infty < x < +\infty.$$

One can construct a Normal random variable X, with mean μ and variance σ^2 from a random variable, Z, with standard normal distribution by using the relationship $X = \mu + \sigma Z$. Consequently, the calculation of the probability of an event involving X can be reduced to a calculation involving Z. For example,

$$\begin{aligned} P(a < X < b) &= P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) \\ &= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$

where $\Phi(x)$ is the c.d.f. of the standard normal distribution. The function $\Phi(x)$ is widely tabled and is given in nearly every basic statistics textbook.

The normal distribution is of great importance, because it is an excellent approximation for the probability distribution of many random variables that arise in nature. The reason for this

is based on the central limit theorem. This theorem states that if a random variable arises as the sum of a large number of independent random variables with a common probability distribution having a positive but finite variance, then the distribution of the sum has an approximate normal distribution; see Probability Theory: Formal for additional details. The normal distribution is widely used in the social sciences. It is often used to model the distribution of IQ and educational testing measurements within a population. It is commonly used as the distribution of the error term in regression analysis and other linear models problems; see Linear Hypothesis and Linear Hypothesis: Regression (Basics) for details. Given the importance of this distribution as a basis for much of statistical inference, in cases for which the normal dstribution does not serve as an appropriate distribution for a random variable, one often attempts to find a transformation such that the transformed variable will have an approximate normal distribution.

Exponential (λ) Distribution

The exponential distribution arises frequently in problems involving system reliability and the times between events. This distribution is most easily described using the failure rate function, which for this distribution is constant, i.e.,

$$\lambda(x) = \begin{cases} \lambda & \text{if } x \geq 0, \\ 0 & \text{if } x < 0 \end{cases}$$

The constancy of the failure rate function leads to the memoryless or Markov property associated with the exponential distribution. Specifically,

$$P(X > s + t | X > s) = P(X > t)$$

The interpretation of the above property is easiest to understand when X represents the lifetime of a component or system. The memoryless property indicates that if the component has lasted for s units of time and is still functioning, then the probability that it will last an additional t time units is independent of s. Thus, the likelihood of an additional life of at least t time units does not depend on its current age, hence the term 'memoryless.'

The p.d.f. is given by

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } 0 \leq x < +\infty, \\ 0 & \text{otherwise} \end{cases}$$

for $\lambda > 0$.

If X has an exponential (λ) distribution, then

$$E(X) = \frac{1}{\lambda}, \quad Var(X) = \frac{1}{\lambda^2}$$

The exponential distribution is the basis for continuous time Markov chain models. This distribution is used to model the sojourn or holding time in a state for a continuous time chain; see Markov Models and Social Analysis. Moreover, the distribution is often used as a model for the time between events; for example, in modeling the arrivals of customers to a service center.

Gamma (α, λ) Distribution

The Gamma (α, λ) distribution generalizes the exponential distribution. The parameters must satisfy $\alpha > 0$ and $\lambda > 0$. The p.d.f. is given by

$$f(x) = \begin{cases} \dfrac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} & \text{if } 0 \leq x < +\infty \\ 0 & \text{otherwise} \end{cases}$$

Thus a Gamma (1, λ) is an exponential distribution with mean $1/\lambda$. Suppose α is an integer, and $\{X_i, 1 \leq i \leq \alpha\}$ are independent random variables, each with a Gamma (m, λ) distribution, then $S = \sum_{i=1}^{\alpha} X_i$ has a Gamma (αm, λ) distribution.

If X has a Gamma (α, λ) distribution, then

$$E(X) = \frac{\alpha}{\lambda}, \quad Var(X) = \frac{\alpha}{\lambda^2}$$

The Gamma distribution offers a two-parameter family of distributions of positive random variables with many possible shapes. This is particularly useful for modeling event histories in social science studies.

Weibull (α, λ) Distribution

Another probability distribution that arises in reliability and event history modeling is the Weibull (α, λ) distribution for $\alpha > 0$ and $\lambda > 0$. The p.d.f. is given by

$$f(x) = \begin{cases} \alpha\lambda^\alpha x^{\alpha-1} e^{-(\lambda x)^\alpha} & \text{if } x \geq 0, \\ 0 & \text{otherwise} \end{cases}$$

The failure rate function for the Weibull distribution is given by

$$\lambda(x) = \alpha\lambda^\alpha x^{\alpha-1}$$

If $0 < \alpha < 1$, then the failure rate function is decreasing, whereas if $\alpha > 1$ the failure rate function is increasing. The $\alpha = 1$ case results in a constant failure rate which corresponds to the exponential distribution.

If a random variable X has Weibull (α, λ) distribution, then X^α has an Exponential (λ^α) distribution. Thus, when $\alpha = 1$, this distribution becomes the Exponential (λ) distribution.

The Weibull distribution is often used in reliability or survival analysis (or life testing) studies. The distribution can exhibit a wide range of tail behaviors, especially heavy tails. The Weibull distribution is one of the three *extreme value distributions*, distributions which arise by normalizing the maximum of a random sample from a distribution; see Distributions, Statistical: Approximations.

Chi-Square (n) Distribution

The Chi-square distribution with n degrees of freedom has p.d.f. given by

$$f(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)^{-1}} e^{-x/2} & \text{if } x \geq 0, \\ 0 & \text{otherwise} \end{cases}$$

This distribution is a special case of the Gamma (α, λ) distribution with $\alpha = n/2$ and $\lambda = \frac{1}{2}$. It arises as a sum of squares of independent standard normal random variables. Specifically, suppose $\{X_i, 1 \leq i \leq n\}$ are independent standard normal random variables, then $\sum_{i=1}^{n} X_i^2$ has a Chi-square (n) distribution.

The mean and variance of a random variable having a Chi-square (n) distribution are given by

$$E(X) = n, \quad Var(X) = 2n.$$

The Chi-square distribution is very widely used in statistics, especially in goodness of fit testing and in categorical data analysis. The distribution also arises as the distribution for the sample variance estimator of an unknown variance based on a random sample from a normal distribution.

Cauchy Distribution

The standard version of the Cauchy distribution has a p.d.f. given by

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < +\infty$$

This distribution is symmetric about 0; however, one can introduce both location and scale parameters which can move the center and change the concentration of the distribution.

The Cauchy distribution has a very heavy tail, comparable to the tail of the Pareto (1, c) distribution. In fact, the tail is so heavy that the distribution does not have a mean value. The Cauchy distribution is a *stable distribution*, see Distributions, Statistical: Approximations. In particular, it can arise as a properly normalized sum of independent random variables having a suitably heavy tail. The Cauchy distribution and related stable distributions have been used to model certain economic variables, especially financial asset returns.

Student $t(n)$ Distribution

The Student t distribution with n degrees of freedom arises in statistical inference problems associated with the mean of a normal distribution when both the mean and the variance are unknown. If Z and Y are independent random variables, with Z having a standard normal distribution and Y having a Chi-square (n) distribution, then the ratio nZ/Y has Student t distribution with n degrees of freedom. It has a p.d.f. given by

$$f(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}, \quad -\infty < x < +\infty$$

The Student t distribution with 1 degree of freedom is the same as the Cauchy distribution. For large n, the Student t distribution is well approximated by the standard normal distribution. The Student t distribution is widely used in the formation of confidence intervals for an unknown mean of a normal distribution estimated from a sample of data from a normal distribution with both mean and variance unknown. It also arises in the formation of confidence intervals for regression parameters; see Linear Hypothesis: Regression (Basics).

$F(m, n)$ Distribution

The $F(m, n)$ distribution arises in statistical inference problems associated with the normal distribution, especially in the analysis of variance to determine the significance of effects. If random variables V and W are independent, each with a Chi-square distribution having degrees of freedom given by m and n, respectively, then the ratio $(V/m)/(W/n)$ has an F(m, n) distribution. The p.d.f. is given by

$$f(x) = \begin{cases} \dfrac{\Gamma((m+n)/2)m^{m/2}n^{n/2}}{\Gamma(m/2)\Gamma(n/2)} \dfrac{x^{(m/2)-1}}{(mx+n)^{(m+n)/2}} & \text{if } x \geq 0, \\ 0 & \text{otherwise} \end{cases}$$

The F distribution is of great importance in the analysis of variance and the design of experiments. This distribution is used to determine the significance of different effects in an analysis of variance study; see Experimental Design: Overview; Linear Hypothesis; Analysis of Variance and Generalized Linear Models.

Pareto (α, c) Distribution

This is a two-parameter family of distributions indexed by $\alpha > 0$ and $c > 0$. The p.d.f. is given by

$$f(x) = \begin{cases} \dfrac{\alpha c^{\alpha}}{x^{\alpha+1}} & \text{if } x \geq c, \\ 0 & \text{otherwise} \end{cases}$$

This distribution has a relatively heavy tail which decreases at only a polynomial rate. Consequently, only moments of order less than α exist. In particular, if X has a Pareto (α, c) distribution, then

$$E(x) = \frac{\alpha}{\alpha-1}c, \quad \text{if } a > 1$$

and

$$Var(X) = \frac{\alpha}{(\alpha-1)^2(\alpha-2)}c^2, \quad \text{if } a > 2$$

If X has a Pareto (α, c) distribution, then the transformed random variable $\log(X/c)$ has an Exponential (α) distribution.

The Pareto distribution arises in contexts in which the tail of the distribution is relatively heavy. This occurs especially in economics; for example, in the distribution of incomes and the returns from various financial assets. The distribution also arises as the excess of a random variable over a high limit. The excess is especially important in the casualty insurance where one is interested in the excess of a claim over a high deductible limit; see Distributions, Statistical: Approximations.

Multivariate Distributions

It often happens that it is wished to consider two or more random variables simultaneously rather than individually. If $\{X_i, 1 \leq i \leq n\}$ is a set of n random variables, then each will have its own probability distribution, F_i, $1 \leq i \leq n$.This individual or marginal information does not offer insight into the relationship among these variables. To capture these relationships, one must consider the n-dimensional random vector (X_1,…, X_n). To do this, one defines the higher dimensional analogs of the c.d.f. and p.d.f. Specifically, the n-dimensional c.d.f. (or joint c.d.f.) is defined by

$$F(x_1,\ldots,x_n) = P(\{X_1 \le x_1\} \cap \ldots \cap \{X_n \le x_n\}),$$
$$\text{for } -\infty < x_i < +\infty, \quad 1 \le i \le n$$

If the joint c.d.f. is absolutely continuous, then one can define the joint p.d.f.,

$$f(x_1,\ldots,x_n) = \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F(x_1,\ldots,x_n)$$

In the absolutely continuous case, one can calculate the probability of events by integrating the joint p.d.f. over the values specified by the event. When we consider the joint distribution of two random variables, (X_1, X_2), the associated joint c.d.f. and joint p.d.f. are referred to as the bivariate c.d.f. and bivariate p.d.f. A special case of this occurs when the pair of random variables has a bivariate normal distribution.

Bivariate Normal $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ Distribution

The bivariate normal distribution is a special joint distribution associated with two random variables, X_1 and X_2. The joint p.d.f. of (X_1, X_2) is given by

$$f(x_1,x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_1\sigma_2}\exp\left(-\frac{1}{2(1-\rho^2)}\right)\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)$$

for $-\infty < x_1, x_2 < +\infty$.

The parameters μ_i, $i = 1,2$ are arbitrary; however, the other parameters must satisfy $\tau_i > 0$, $i = 1,2$ and $-1 < \rho < +1$. When $\rho = \pm 1$, the random variables X_1, X_2 are related linearly.

If a pair of random variables (X_1, X_2) has a bivariate normal distribution with the parameters given above, then each has a marginal normal distribution, i.e., X_i has a Normal(μ_i, σ_i), $i = 1,2$ distribution. The parameter ρ is the correlation coefficient of X_1 and X_2. The bivariate normal distribution has the special property that if $\rho = 0$, then X_1 and X_2 are independent random variables. Generally, uncorrelated random variables are *not* independent.

One can represent (X_1, X_2) in terms of two independent random variables (Z_1, Z_2), each with a standard normal distribution through the relationship:

$$X_1 = \mu_1 + \sigma_1 Z_1,$$
$$X_2 = \mu_2 + \sigma_2\left(\rho Z_1 + \sqrt{1-\rho^2}Z_2\right).$$

The bivariate normal distribution arises naturally in simple regression problems, see Linear Hypothesis: Regression (Basics).

Further Reading

A large amount of material on continuous distributions, including statistical inference for unknown parameters of the distribution, is given in the text by Johnson et al. (1994). Multivariate continuous distributions are covered extensively by Johnson and Kotz (1972). Many of the above distributions are discussed in basic probability textbooks; for e.g., the text by DeGroot (1987). Some illustrations of the use of continuous probability distributions in the social sciences are given by Handcock and Morris (1999).

See also: Distributions, Statistical: Approximations; Distributions, Statistical: Special and Discrete; Probability Theory: Formal

Bibliography

Casella, George, Berger, Roger L., 1990. Statistical Inference. Duxbury Press, Belmont, CA.

DeGroot, M.H., 1987. Probability and Statistics, second ed. Addison-Wesley, Reading, MA.

Forbes, Catherine, Evans, Merran, Hastings, Nicholas, Peacock, Brian, 2011. Statistical Distributions. John Wiley & Sons, Hoboken, NJ.

Handcock, M.S., Morris, M., 1999. Relative Distribution Methods in the Social Sciences. Springer-Verlag, New York.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1994. Continuous Univariate Distributions, second ed. Wiley, New York.

Johnson, N.L., Kotz, S., 1972. Continuous Multivariate Distributions. Wiley, New York.

Kotz, Samuel, Balakrishnan, Narayanaswamy, Johnson, Norman L., 2000. Continuous multivariate distributions, models and applications. John Wiley & Sons, New York.

Distributions, Statistical: Special and Discrete

Cecil D Kemp, University of St Andrews, St Andrews, Fife, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 3793–3797, © 2001, Elsevier Ltd.

Abstract

Discrete probability distributions are basic to the statistical methodology for modeling and analyzing discrete (or count) data. The article first defines and discusses the main properties of discrete distributions. Usually each outcome (observation) is an integer X, known as a random variable. A distribution associates a probability (a number lying between 0 and 1) with each possible outcome; the sum of these probabilities must be unity. In distribution theory the probabilities are values of a mathematical function (usually called the probability mass function or pmf) of X. Other important functions or constants associated with a distribution that are discussed include the moments (of which the mean and the variance are the most familiar) and the probability generating function (pgf). The pmf, pgf, and moments are then given for a number of widely used distributions including the Poisson, binomial, negative binomial, logarithmic, hypergeometric, and negative hypergeometric distributions. The geneses and some typical applications of distributions are also mentioned. Ways of combining distributions to form new distributions are briefly considered. Finally multivariate discrete distributions (more than one variable is measured for each outcome) are introduced and the multinomial distribution (which is involved in the analysis of contingency tables) is examined.

Discrete probability distributions form the basis of much of the statistical methodology for modeling, analyzing, and making inferences about discrete data (sometimes known as count data). Usually each observation (outcome) is an integer; for example, the number of individuals in a family, or the number of telephone calls handled by a switchboard in a unit of time. This article is primarily concerned with these underlying distributions and their properties, rather than their applications. The article first discusses the general properties of univariate discrete distributions and then defines and examines a number of widely used distributions. Ways of combining distributions to form new distributions are briefly considered. Finally multivariate discrete distributions (more than one variable is measured for each outcome) are introduced and the multinomial distribution (which is involved in the analysis of contingency tables) is examined (*see* Multivariate Analysis: Discrete Variables, Overview).

Definitions and General Properties

Formally, a (univariate) discrete random variable (which is used to model the observation) X can take the values x_1, x_2, x_3, …. Usually these $\{x_i\}$ are restricted to be the nonnegative integers or some subset of them. For example, an investigator might assume that $\{1, 2, 3, \ldots, 20\}$ includes all the possible different values for family size that he might observe.

The (discrete) distribution of X associates with each possible X-value the probability p_x that it will be the observed value. The distribution must satisfy the two conditions:

$$0 \le p_x \le 1; \quad \sum_x p_x = 1 \qquad [1]$$

i.e., the probabilities are all nonnegative real numbers lying between zero (impossibility) and unity (certainty), and their sum must be unity because it is certain that one or other of the values will be observed. It is often convenient to assume that the number of possible outcomes (values of X) is countably infinite, i.e., there is no upper bound to X.

In discrete distribution theory the p_x, are regarded as values of a mathematical function, $p(x)$, that is usually called the probability mass function (pmf). Different functional forms lead to different discrete distributions. For example:

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad [2]$$

is the pmf of the Poisson distribution, while:

$$p(x)\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \qquad [3]$$

is the pmf of the binomial distribution. λ is the parameter of the Poisson distribution; this fact is sometimes highlighted by writing $p(x; \lambda)$ or $p(x|\lambda)$ instead of $p(x)$ or p_x.

There are various useful and important constants associated with a distribution. They may be used to summarize various features of the distribution. The most familiar and widely used are the moments, particularly the mean and the variance. The rth (uncorrected) moment of the distribution is defined as

$$\mu'_r = \sum_x x^r p_x \qquad [4]$$

The mean is μ'_1, the first moment of the distribution; it is conventionally denoted by μ and is used as a measure of the location of the distribution on the X-axis.

The rth moment about the mean is defined as

$$\mu_r = \sum_x (x-\mu)^r p_x \qquad [5]$$

The variance is μ_2, the second moment about the mean; it is conventionally denoted by σ^2 and is widely used as a measure of the spread of the distribution. μ_3 is sometimes used to form a measure of the degree of asymmetry of the distribution. There is a formula for μ_r in terms of the μ'_i, and vice versa.

There are other related sets of constants, such as the cumulants, the factorial moments, the factorial cumulants, which

 http://dx.doi.org/10.1016/B978-0-08-097086-8.42116-1

may be more convenient to use in some circumstances. For details of these constants, and very detailed information generally about univariate discrete distributions, see Johnson et al. (1992).

In principle all the theoretical properties of the distribution can be derived from the pmf. There are, however, several other functions from which exactly the same information can be derived. This is because the functions are all one-to-one transformations of each other, so each characterizes the distribution. One particularly useful function is the probability generating function (pgf), which is defined as

$$G(s) = \sum_x s^x p_x \qquad [6]$$

For example, the Poisson distribution has pgf:

$$G(s) = \exp[\lambda(s-1)] = e^{-\lambda} \sum_x \frac{\lambda^x}{x!} s^x = \sum_x p_x s^x \qquad [7]$$

If any function that is known to be a pgf it is expanded as a power series s, then the coefficient of s^x must be p_x for the corresponding distribution. An alternative way of obtaining the probabilities is by repeated differentiation of $G(s)$ with respect to s. If the pmf of a distribution is a complicated function and there are difficulties in computing numerical values directly, it may be possible to deduce from the pgf relatively simple relationships between successive probabilities and thus compute probabilities sequentially.

During the past half-century there has been a great proliferation of univariate discrete distributions. Patil et al. (1984) listed 136 univariate and multivariate distributions; Wimmer and Altmann (1999) list about 750 univariate distributions, but estimate the total number currently known at around 1000. While some are rather thinly disguised versions of others, in general there are important differences between distributions. Some structure can be imposed on this vast array by grouping distributions into families whose members share certain properties.

Individual Univariate Discrete Distributions

Bernoulli Distribution

This is an extremely simple and basic distribution. It arises from what is known as a Bernoulli trial: a single observation is taken where the outcome is dichotomous, e.g., success or failure, alive or dead, male or female, 0 or 1. The probability of success is usually denoted by p. The probability of failure is $1-p$ and this is often denoted by q. The distribution has the pmf:

$$p_x = p^x q^{1-x}, \quad 0 < p < 1, \quad q = 1-p, \quad x = 0, 1 \qquad [8]$$

The mean is p, the variance is pq, and the pgf is $(q+ps)$. The Bernoulli is a one-parameter distribution (the parameter is p). Two standard examples of a Bernoulli trial are (1) the toss of a coin and (2) the random selection of a ball from an urn containing a white and b black balls.

Binomial Distribution

This two-parameter distribution describes the outcome of a sequence of n independent Bernoulli trials, each with the same probability p of success. Then the probability that success is the outcome in exactly x of the trials is

$$p(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \qquad [9]$$
$$0 < p < 1, \quad q = 1-p, \quad x = 0, 1, 2, \ldots, n$$

This defines the binomial distribution. In terms of the two Bernoulli trials examples above, X would be (1) the number of heads in n tosses of a coin or (2) the number of white balls in a random sample of n balls selected sequentially, *provided* that each ball selected is returned to the urn before the next ball is selected; this is necessary in order that p remains constant and equal to a/n. This latter procedure is an example of sampling with replacement. The mean of the binomial distribution is np and the variance is npq. The pgf is $(q+ps)^n$. Note that this is the Bernoulli pgf raised to the nth power. This illustrates a very useful property of pgfs: if X and Y are two independent random variables whose distributions have pgfs $G_X(s)$ and $G_Y(s)$, then their sum $Z = X + Y$ has a distribution with pgf $G_Z(s) = G_X(s)G_Y(s)$.

If n is large and p is small, the Poisson distribution (see section 'Poisson Distribution') with $\lambda = np$ may be used to approximate the binomial. If n is large and p is not too close to 0 or 1, the (continuous) normal distribution may be used to construct an approximation.

The well-known chi-squared test for a 2×2 contingency table can be shown to be equivalent to a test of the difference between two binomial proportions (the proportions of 'successes' within the two classes in the table).

Poisson Distribution

This distribution occupies a central position in discrete distribution theory analogous to that occupied by the normal distribution in continuous distribution theory. It also has many practical applications. The full formal definition of the distribution is

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad 0 < \lambda, \quad x = 0, 1, 2, \ldots \qquad [10]$$

The pgf of the distribution has the very simple form:

$$G(s) = \exp[\lambda(s-1)] \qquad [11]$$

Both the mean and the variance of the Poisson distribution are equal to λ. For small values of λ the distribution is very skewed (asymmetric) but as λ increases it becomes less skewed and is nearly symmetric by $\lambda = 15$.

The Poisson distribution plays an important part in loglinear model theory for contingency tables; the model assumes independent Poisson distributions for each cell in the table (*see* Multivariate Analysis: Discrete Variables (Loglinear Models)). Another reason why the Poisson distribution is important is that it is the distribution of events that occur randomly and independently in space or time. A classic example is the number of particles, X, recorded by a Geiger counter in a fixed time interval. This property of the Poisson means that it can act as a reference standard when deviations from pure randomness are suspected.

Geometric Distribution

If X is the number of trials *before* the first success in a sequence of Bernoulli trials with constant p, then:

$$p_x = pq^x, \quad 0 < p < 1, \;\; q = 1 - p, \;\; x = 0, 1, 2, \ldots \qquad [12]$$

The probabilities form a geometric progression, hence the name.

The mean is p/q, the variance is p/q^2, and the pgf is $p(1 - qs)^{-1}$.

Just as the Bernoulli distribution is a special case of the binomial distribution, the geometric distribution is a special case of the negative binomial distribution (NBD) (see section 'Negative Binomial Distribution').

Some authors define X to be the number of trials up to and *including* the first success; the pmf is then $p_x = pq^{x-1}$, X ranges over the positive integers, and the pgf is $ps(1 - qs)^{-1}$.

Negative Binomial Distribution

The negative binomial distribution is a widely used alternative to the Poisson distribution for handling count data when the variance is appreciably greater than the mean (this condition is known as overdispersion and is frequently met in practice).

The pmf of the NBD is

$$p_x = \binom{k + x - 1}{x} p^x q^x, \quad 0 < k, \;\; 0 < p < 1, \quad q = 1 - p, \;\; x = 0, 1, 2, \ldots \qquad [13]$$

The mean is kq/p, the variance is kq/p^2, and the pgf is $p(1 - qs)^{-k}$.

In the literature, several different parameterizations of the distribution are found. A common one uses $Q = 1/p$ and $P = q/p$. It follows that $p > 0$, $Q = 1 + P$, and:

$$p_x \binom{k + x - 1}{x} \left(1 - \frac{P}{Q}\right)^k \left(\frac{P}{Q}\right)^x \qquad [14]$$

When $k = 1$, the NBD becomes the geometric distribution. When k is a positive integer, the NBD is sometimes known as the Pascal distribution; it can then be interpreted as the distribution of the number of failures before the kth success (i.e., X is the sum of k independent geometric random variables). This kind of model is known as a waiting time distribution.

There are several other models that lead to the NBD. A classic example arises from the theory of accident proneness that was developed during and after World War I. It was assumed that the number of accidents suffered by an individual was distributed as a Poisson distribution, but that the Poisson parameter λ (interpreted as the individual's accident proneness) varied between individuals in the population under study. If λ is assumed to follow the (continuous) gamma distribution, then the NBD is the resultant overall distribution of accidents per individual. This model has been applied to retail purchasing, absenteeism, doctor's consultations, among many others. The model reflects a type of heterogeneity in the population. A different type of heterogeneity occurs when there is clustering. If it is assumed that the number of clusters is distributed as a Poisson, but the number of individuals in a cluster is distributed according to the logarithmic distribution (see section Distributions Without a Zero Class), then the overall distribution is a NBD.

The existence of these and other models means that the NBD is used to model many different types of data. The penalty for this ubiquity is the difficulty in inferring which underlying model applies.

Distributions without a Zero Class

In some applications the meaningful outcomes do not include $X = 0$. Examples are the size of a family and the size of a cluster of dwellings. Such cases require distributions that start at $X = 1$. One way of achieving this is to take a distribution that is defined for $x = 0, 1, 2, \ldots$ and truncate the term $X = 0$. In order to form a valid distribution from the remaining probabilities, it is necessary to adjust them upwards by a factor of $(1 - p_0)^{-1}$ so that the second condition in [1] is satisfied. For example, the zero-truncated Poisson distribution has pmf:

$$p_x = \frac{e^{-\lambda}\lambda_x}{x!(1 - e^{\lambda})}, \quad 0 < \lambda, \;\; x = 1, 2, 3, \ldots \qquad [15]$$

Alternatively, there are some distributions for which the pmf is not defined for $X = 0$. One of these is the logarithmic distribution, with pmf:

$$p_x = \frac{\theta^x}{-x \log_e(1 - \theta)}, \quad 0 < \theta < 1, \;\; x = 1, 2, 3, \ldots \qquad [16]$$

The mean is $a\theta/(1 - \theta)$ and the variance is $a\theta(1 - a\theta)/(1 - \theta)^2$, where $a = -1/\log_e(1 - \theta)$. The pgf is

$$G(s) = \frac{\log_e(1 - \theta s)}{\log_e(1 - \theta)} \qquad [17]$$

This distribution can model data distributions with very long tails.

Hypergeometric Distribution

A key feature of the binomial distribution is its representation in terms of a sequence of independent Bernoulli trials with fixed p. In many practical situations, however, p is not constant due to dependence between successive Bernoulli trials. In the derivation of the binomial using the ball in an urn model (b), each selected ball was returned immediately to the urn (sampling with replacement) so p was always equal to $a/(a + b)$. However, if each selected ball is retained and not returned, then p changes after every trial; this is sampling without replacement. The resultant distribution is hypergeometric with pmf:

$$\begin{aligned} p_x &= \binom{a}{x}\binom{b}{n - x} \Big/ \binom{a + b}{n} \\ &= \binom{n}{x}\binom{a + b - n}{a - x} \Big/ \binom{a + b}{a}, \\ & x = 0, 1, 2, \ldots, \min(n, a) \end{aligned} \qquad [18]$$

The mean is $na/(a + b)$ and the variance is $nab(a + b - n)/(a + b)^2(a + b - 1)$. The pfg is

$$G(s) = \frac{{}_2F_1[-n, -a; b-n+1; s]}{{}_2F_1[-n, -a; b-n+1; 1]} \quad [19]$$

i.e., the ratio of two classical Gaussian hypergeometric functions.

An important use of the hypergeometric distribution is in the analysis of 2×2 contingency tables; the probability of a result as extreme as the observed result is given by the tail of a hypergeometric distribution. This is known as Fisher's exact test; it is frequently recommended for use instead of the chi-squared test when the smallest expected frequency is less than 5.

Beta-Binomial (Negative Hypergeometric) Distribution

In the section 'Negative Binomial Distribution' heterogeneity of the Poisson parameter λ was allowed for by assuming that λ had a gamma distribution. The resultant Poisson–gamma distribution was in fact the NBD. Analogously, when there is heterogeneity in the binomial parameter p one way of allowing for it is to assume that it follows the (continuous) beta-distribution – a flexible distribution on the interval (0, 1). The resultant beta-binomial distribution has pmf:

$$p_x = \binom{-a}{x}\binom{-b}{n-x}\Big/\binom{-a-b}{n}, \quad 0 < a,\ \ 0 < b,\ \ n \text{ is an integer}$$
$$= \binom{a+x-1}{x}\binom{b+n+x-1}{n-x}\Big/\binom{a+b+n-1}{n}, \quad x = 0, 1, 2, \ldots n \quad [20]$$

The mean is $na/(a+b)$, the variance is $nab(a+b+n)/(a+b)^2(a+b+1)$, and the pgf is

$$G(s) = \frac{{}_2F_1[-n, a; -b-n+1; s]}{{}_2F_1[-n, a; -b-n+1; 1]} \quad [21]$$

The distribution is also known as the negative hypergeometric or inverse hypergeometric distribution. Like the NBD, it can be interpreted as a waiting time distribution.

Combinations of Distributions

New distributions may be created from existing ones. Two major methods have been mentioned in the sections 'Negative Binomial Distribution' and 'Beta-Binomial (Negative Hypergeometric') Distribution. The process of allowing the parameter of one distribution to itself be distributed (as in the derivation of the NBD by letting the Poisson parameter λ be distributed as a gamma variable) is known as mixing (in earlier literature the term compounding was commonly used). Many other mixture distributions appear in the literature.

The process exemplified by the NBD cluster model of Poissonly distributed clusters with logarithmically distributed cluster sizes has been given several names including compounding and generalizing, both of which are ambiguous; a clearer terminology that is frequently used nowadays is to call the resultant distribution a randomly stopped-sum or stopped-sum distribution. If the pgf of the distribution of clusters is $g(s)$ and that of the cluster size distribution is $h(s)$ then the pgf of the resultant distribution has the very simple structure:

$$G(s) = g(h(s)) \quad [22]$$

Various properties of the stopped-sum distribution can be deduced via this relationship.

These processes are one reason why there are so many different discrete distributions in the literature. They often arise from attempts to model processes in applications. Clearly some distributions like the NBD that are mixtures are also stopped sums, but this is not always the case. As mentioned previously, the existence of fundamentally different ways of deriving the same distribution can raise serious problems in practice when trying to make causal inferences from data.

Multivariate Discrete Distributions

The move from distributions of one variable to joint distributions of several variables is complicated by the existence of relationships (dependencies) between the variables. This is reflected in the complexity of the structure of the multivariate pmf and related functions. Furthermore, there may be more than one way of representing dependency; more than one bivariate version of a particular univariate distribution may be constructed. Often the multivariate pgf is one of the simpler ways of defining and characterizing a distribution, A bivariate pgf is defined as

$$G(s_1, s_2) = \sum_{x,y} p_{x,y} s_1^x s_2^y \quad [23]$$

the definition of a multivariate pgf is a straightforward extension.

The multinomial distribution has pgf and pmf:

$$G(s_1, s_2, \ldots, s_r) = (p_1 s_1 + p_2 s_2 + \ldots + p_r s_r)^n \quad [24]$$

$$p(x_1, x_2, \ldots,) = \frac{n!}{x_1!, x_2!, \ldots, x_r} p_1^{x_1} \ldots p_r^{x_r} \quad [25]$$

When $r = 2$ it becomes the binomial distribution. All the marginal distributions of the multinomial are binomials. The conditional distribution of $(X_2, X_3, \ldots X_r)$ given X_1 is also multinomial.

One of the first reference texts to be devoted to the multivariate area was Kocherlakota and Kocherlakota (1992). They set out the general theory of bivariate discrete distributions and considered in detail the properties of bivariate versions of the major univariate distributions. Johnson et al. (1997) give a very thorough treatment of the current position in multivariate discrete distributions; the general theory is given and a wide range of distributions are examined.

Johnson et al. (1992) is the standard reference text for univariate discrete distributions; all the univariate distributions above are considered in detail, together with very many more.

See also: Multivariate Analysis: Discrete Variables (Loglinear Models); Multivariate Analysis: Discrete Variables (Overview).

Bibliography

Johnson, N.L., Kotz, S., Kemp, A.W., 1992. Univariate Discrete Distributions, second ed. Wiley, New York.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1997. Discrete Multivariate Distributions. Wiley, New York.

Kocherlakota, S., Kocherlakota, K., 1992. Bivariate Discrete Distributions. Dekker, New York.

Patil, G.P., Boswell, M.T., Joshi, S.W., Ratnaparkhi, M.V., 1984. Dictionary and Classified Bibliography of Statistical Distributions in Scientific Work, vol 1: Discrete Models. International Co-operative Publishing, Fairland, MD.

Wimmer, G., Altmann, G., 1999. Thesaurus of Univariate Discrete Probability Distributions. Stamm, Essen, Germany.

Diversification, Relatedness, and the New Logic of Co-Creation

David J Teece, University of California, Berkeley, CA, USA; and Berkeley Research Group, Emeryville, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.A. Lowe and D.J. Teece, volume 6, pp. 3574–3578, © 2001, Elsevier Ltd.

Abstract

Diversification involves the expansion of an enterprise into new markets of various types. This article focuses on cross-industry diversification. Conglomerate (unrelated) diversification, in which a business entity groups together a disparate bundle of businesses, has generally fallen out of favor, except in developing economies, where it remains a viable response to local institutions. Related diversification has come to be understood in the strategic management literature as the reapplication of some of the enterprise's slack resources, which are hard to monetize in other ways, to new opportunities. There is a large, but disputed, literature claiming that diversification on average reduces performance compared with comparable single-business firms. Diversification as a growth strategy may be in decline due to the prevalence of outsourcing, strategic alliances, and platform-based ecosystems.

Introduction

Diversification is an aspect of the growth of an enterprise in which it expands into new activities or markets. The new activities may share some technical or market resource with the enterprise's current operations (related diversification); may integrate a previously external upstream or downstream activity relative to current operations (vertical integration); or may be wholly unrelated (conglomerate diversification). The expansion can be accomplished through acquisition or by internally developing a new line of business.

Expansion into foreign markets is another type of diversification. Although it shares some characteristics with the industry-level diversification considered here (e.g., two countries can be culturally 'close' in much the same way that two industries can be technologically 'close'), there are also critical differences. In particular, scholars are less likely to suspect that international diversification results from managerial malfeasance than they are to suspect industry diversification. The international diversification literature is not explicitly discussed in this article.

There is a vast and growing literature concerning industry-level diversification. Reviews of this literature (e.g., Wan et al., 2011) exist elsewhere, so this article does not attempt to be comprehensive. Rather, it highlights some of the key concerns of the literature about diversification, with an emphasis on the concept of relatedness.

Conglomerate (Unrelated) Diversification

Conglomerate diversification, in which a firm enters (most often by acquisition) a line of business that has no obvious market or technological relationship to its existing activities, was popular in the 1960s. One possible reason is that antitrust enforcement was relatively strict with respect to vertical and horizontal mergers during that time (Mueller, 1969).

The conglomerate wave may also have been driven by a belief that good management ability could be profitably (and easily) applied to any industry, a belief that is no longer as widely held. For instance, GE, one of the last major conglomerates, has started leaving its executives in single-business units for longer periods of time to allow them to develop deep industry expertise instead of moving them across different units every few years (Linebaugh, 2012).

Most conglomerates have since been unwound. The leveraged buy-out wave of the 1980s was in part fueled by the desire of many firms to shed lines of business.

Today, more companies seem interested in specialization than in the questionable opportunities afforded by unrelated, or even vertical, diversification. Advances in transportation and telecommunications have greatly reduced the negative aspects of working across great distances. As a result, it is often simpler and less expensive to outsource parts of a firm's activities to the lowest cost-qualified vendors, wherever they are located in the world, than to develop (or maintain) internal capabilities.

A countercurrent to the overall trend toward a narrow focus for the business enterprise is business groups located in developing countries, where they may undertake anything from sugar to textiles to software. This may be a response to the unique institutional and techno-economic conditions in these countries, where personal contacts can be more important than industrial capabilities (Kock and Guillén, 2001). As enterprise capabilities improve and/or the local economy becomes more open to foreign competitors, local business groups need to adjust their structure or suffer the consequences. The experience of Korean business groups provides examples of groups that collapsed (e.g., Daewoo) when the economy came under stress and others (e.g., Samsung) that have thrived (Kim et al., 2004).

Understanding Diversification

In the (neoclassical) version of economic theory, with perfect competition and homogeneous firms, there is neither benefit to, nor penalty for, diversification. Economic rents are competed down to zero in all industries.

In the finance literature, diversification is viewed through the principal-agent framework as a manifestation of mismanagement and is predicted to be destructive of firm value.

Although there is some empirical evidence in support of the finance view on average, there are market imperfections that create profitable opportunities for firms to diversify.

The mainstream economic rationale for efficient diversification involves assuming economies of scope (Panzar and Willig, 1981). Scope economies occur when the cost of producing a given quantity of two or more different products is lower when they are produced jointly rather than separately. Although this could give rise to firms diversifying into the joint products, scope economies can potentially be realized through contracting between separate producers. Hence, as a matter of theory, if markets function properly, economies of scope can be captured without joint ownership. It is only when the scope economies are based on sharing specialized assets, such as know-how, that capturing economies of scope requires diversification (Teece, 1980).

There are also cases where firms will choose to diversify vertically in order to create a cospecialized upstream or downstream linkage that would be hard to replicate contractually. For example, Apple created its own chain of brick-and-mortar stores so it could control completely its look and feel and ensure that Apple products and the company's image are well represented by sales staff.

As an alternative approach to understanding diversification, strategic management scholars in the 1980s developed the Resource-Based View of the firm. The Resource-Based View was built on a major work by Edith Penrose (1959) on the growth of the firm. In Penrose's model of enterprise growth, companies always found themselves with potentially valuable underutilized resources. Provided that the necessary human resources were not prohibitively scarce, management could find and pursue other lines of business in which the slack resources could be profitably deployed.

Although Penrose's ideas found little acceptance among mainstream economists, they were eventually embraced by scholars of strategic management such as Teece (1982) and Wernerfelt (1984). Among other things, the Resource-Based View amplified on Penrose's idea that firm-specific resources used in one industry could provide a basis for competing profitably in others where their attributes would also help create value.

There is historical case-study evidence in support of this theory. Chandler (1969) described how the Depression in the United States triggered the first great wave of diversification. General Electric and Westinghouse expanded from light and power equipment to household appliances, and General Motors expanded into diesels, tractors, and airplanes.

The resource-based approach has produced numerous testable hypotheses. To take just one example, a statistical study by Chatterjee and Wernerfelt (1991) showed that the nature of a firm's resources affects its type of diversification. Excess physical- and knowledge-based resources were more likely to be associated with related diversification, while excess internal financial resources were associated more closely with unrelated diversification.

Resources also have implications for the performance of diversified firms, about which more is said below. Douglas Miller (2006) developed a resource measure called technological diversity that records the dispersion of industries to which a company's patent portfolio is applicable. He found that, for a given level of technological diversity, a diversified firm will create more value than a comparable cluster of single-business firms.

An advantage of the account of diversification offered by the Resource-Based View over those used in mainstream economics is that it can account for all types of vertical and lateral diversification. Even in the case of conglomerate diversification that is seemingly unrelated to a firm's previous activities, the resources of management and finance know-how constitute a bridge to the new, unrelated activity.

The key notion underlying this approach is the resource. Resources are firm-specific assets that may be difficult for rivals to imitate. They could be tangible but are most valuable when intangible. Examples include intellectual property, process know-how, customer relationships, and the knowledge possessed collectively by groups of skilled employees. Organizational knowledge, while firm specific, is seldom limited only to the products and processes with which it is actively involved (Teece, 1982).

Intangible resources are difficult to trade because their property rights are likely to have fuzzy boundaries, and their value is dependent on their use in the context of the firm's other resources and assets. As a result, such resources are typically not traded at all. They are also generally difficult to transfer to other firms.

A management team, finding itself with slack corporate resources and knowing they are nearly impossible to trade in isolation, is likely to look for other activities where it can profitably redeploy the resources. The effectiveness of this redeployment is a reflection of management's competence in recognizing and seizing opportunities, which are, in turn, part of the organization's dynamic capabilities (Teece, 2007).

The Profiting From Innovation framework (Teece, 1986), which integrates insights from transaction cost economics, strategic management, and other fields, identified a richer set of factors that could impact diversification decisions in the context of innovation. The Profiting From Innovation framework posits that firms need to control and own complementary assets so as to capture returns from innovation in certain circumstances. In particular, incentives to diversify are stronger when the innovation has weak intellectual property protection around its core technology. Firms need to control only the (vertical or horizontal) complements to their core activity that are scarce (or likely to become so post-entry) and thus apt to siphon away profits from the firm's activities.

Empirical Research: Measuring Diversification

In practice, many different types of diversified companies exist. In order to develop a useful theory of diversification, efforts have been made to measure the interbusiness relatedness of diversified corporations. But the challenges are significant.

The most common approach, developed by Caves (1981), uses the hierarchy implicit in the 4-digit Standard Industrial Classification (SIC) system, the predecessor to the 6-digit North American Industry Classification System (NAICS), so that businesses in different 4-digit industries but the same 3-digit industry are 1 'unit' apart, whereas businesses whose closest connection is their 2-digit industry memberships are 2 'units' apart, and so forth. Lemelin (1982) measured interindustry

relatedness as the correlation coefficient across input structures taken from the input–output table. Klavans' (1989) index of technology relatedness is also based on input patterns, but he uses the amount of overlap in occupational categories as an index.

A fairly noisy measure of relatedness has been developed by Gollop and Monahan (1989); they compute the dissimilarity between the patterns of input shares, using nine classes of input (i.e., production wages, fuel, purchased services, etc.). Working with Census of Manufacturers data, they argue that because the data are "filtered through the SIC system, the underlying notion of product dissimilarity must have a supply-side orientation … differences in distribution systems and/or marketing patterns are ignored. It follows that the formulation of product heterogeneity must be based wholly on technical considerations" (1989: p. 4).

Teece et al. (1994) developed an index using SIC data that went beyond input considerations and overturns Gollop and Monahan's dictum. This was accomplished by moving away from measuring the properties of SIC industries and instead looking at the frequencies with which they are combined in firms. When a pair of industries were combined into a single firm, the two industries were assumed to be related in some meaningful way. The 'coherence' of a diversified firm does not mean that all of the firm's activities are mutually interrelated but, rather, that there is evidence of pairwise interrelatedness among them. This approach is consistent with the resource-based theory that a firm expands into a new line of business when it sees an opportunity to reapply some of its resources from an existing line.

Using this measure of coherence on a sample of US corporations, Teece et al. (1994) found that as firms grow more diverse, the overall coherence of each firm does not change much, but there are large differences in coherence among firms. They proposed that the differences could be

> explained as a complex interaction between three classes of factors: (1) enterprise learning, which we suggest is a 'local' phenomena; (2) evolutionary paths, which we see being shaped both by where firms have been (in terms of their scope of their past activities), where they are now (their current competences and complementary assets), and the opportunities which lie ahead; and (3) the strength of competition and the mechanism firms have to shield themselves from it, whether in the product market or the capital market. (1994: p. 28)

They found that the 'coherent diversifiers' driven by these factors were the most common corporate mode in their sample.

Diversification has also been treated as a multidimensional construct. The industries into which a firm has diversified can be related by technology, supply base, skill base, physical asset base, customer base, management knowledge, or, more likely, some combination. Different types of relatedness appear to have different performance effects (Pehrsson, 2006).

Empirical Research: Diversification and Performance

Empirically, some diversification – especially into related industries – appears to improve performance by various accounting measures. But as the diversification spreads into still more industries, performance declines beyond some point. This inverted-U shape may reflect that continued expansion into new industries will eventually render top management less able to control and integrate its operations effectively (Palich et al., 2000).

A number of studies using comparisons of nondiversified firms have found that diversification, broadly defined, destroys value. If true, this result would support the finance theory perspective that diversification involves managerial actions that may be inconsistent with shareholder interests. Most studies in this line of research do find that diversified firms perform worse than specialized firms in the same industries, based on comparisons of market value either to corporate assets ('Tobin's q') or to a hypothetical nondiversified value based on the median market values of single-industry firms in each of the relevant industries.

Relatedness may mediate the effect. Wernerfelt and Montgomery (1988) and Berger and Ofek (1995) found that, among diversified firms, those with the most closely related activities had the best performance. But more recently, Hoechle et al. (2012) found that a performance discount still exists even for diversified firms whose divisions are so related that they share a 4-digit NAICS code.

A few studies (e.g., Campa and Kedia, 2002) found that the diversification discount for all firms becomes insignificant once the self-selection of firms that diversify and the firms they acquire has been controlled for. The result suggests that firms are most likely to diversify or merge when they are weak to begin with or under duress of some kind. This accords with the Depression case-history evidence cited earlier. A further wave of diversification occurred during the Second World War in response to a different kind of market stress, namely a large jump in demand for new products from the military and from consumer industries seeking to replace raw materials that had suddenly become scarce.

Yet more recent studies, with access to additional years of data, have been unable to replicate the self-selection results. Hoechle et al. (2012) found a large and significant valuation discount even after applying a self-selection method similar to that of Campa and Kedia (2002). They went on to show that as much as one third of the discount can be explained by a set of 14 measures of the quality of corporate governance, once again raising the possibility that diversification results, on average, from inefficient investment choices by management. This is in line with articles showing that internal capital markets – far from allowing managers to allocate funds more efficiently than the market (as suggested by Williamson, 1975) – are prone to underinvest in promising divisions while cross-subsidizing weak ones instead of forcing them to improve (see, e.g., Ozbas and Scharfstein, 2010).

However, the existence of a diversification discount remains an unsettled point open to various critiques. For example, the segment designations in Standard & Poor's Compustat database of corporate financial information, on which all of these studies have been based, have been called into question. A study that used establishment-level data from the Census Bureau to derive a more accurate measure of diversification found a diversification *premium* instead of

a discount (Villalonga, 2004). Another study explored evidence that stock market investors might value any group of single-industry firms more highly than a diversified firm with equivalent cash flow, a result that separates the diversification discount from any management implications (Mitton and Vorkink, 2010). Clearly, there is more scope for research on the existence and interpretation of the diversification discount.

Conclusion

The production-oriented logic that once drove the classic form of industrial diversification has receded in importance in many economies. As noted above, waves of specialization and globalization have transformed the world economy by enabling the outsourcing of business functions. Thus the leading electronics company, Apple, owns no factories of its own and expanded into the media-distribution business by building a platform that serves as the basis of an ecosystem encompassing the world's major media sources. As this example suggests, internal diversification is eschewed as a mechanism to achieve productive efficiency. Rather than diversification, what is required is the assembly of a business ecosystem to help co-create new markets. An implication is that the simple production theory approach to understanding diversification cannot capture the richness of the complementarities and externalities that are necessary to understand the efficiency of the activity boundaries chosen by particular firms.

See also: Boundaries and New Organization Forms; Bureaucracy and Bureaucratization; Business Platforms; Business Process: Outsourcing; Hierarchies and Markets; Strategic Management; Technology and Organization.

Bibliography

Berger, Philip G., Ofek, Eli, 1995. Diversification's effect on firm value. Journal of Financial Economics 37, 39–65.

Campa, Jose Manuel, Kedia, Simi, 2002. Explaining the diversification discount. Journal of Finance 57, 1731–1762.

Caves, Richard E., 1981. Diversification and seller concentration: evidence from changes. Review of Economics and Statistics 63, 289–293.

Chandler, Alfred D., 1969. The structure of American industry in the twentieth century: a historical overview. Business History Review 43, 255–298.

Chatterjee, Sayan, Wernerfelt, Birger, 1991. The link between resources and type of diversification: theory and evidence. Strategic Management Journal 12, 33–48.

Gollop, Frank M., Monahan, James L., 1989. From Homogeneity to Heterogeneity: An Index of Diversification. Technical Paper No. 60. U.S. Department of Commerce, Bureau of the Census, Washington.

Hoechle, Daniel, Schmid, Markus, Walter, Ingo, Yermack, David, 2012. How much of the diversification discount can be explained by poor corporate governance? Journal of Financial Economics 103, 41–60.

Kim, Hicheon, Hoskisson, Robert E., Tihanyi, Laszlo, Hong, Jaebum, 2004. The evolution and restructuring of diversified business groups in emerging markets: the lessons from Chaebols in Korea. Asia Pacific Journal of Management 21, 25–48.

Klavans, Richard, 1989. Business Relatedness and Business Performance. FTC Line of Business Program Research Paper No. 83. Federal Trade Commission, Washington.

Kock, Carl J., Guillén, Mauro F., 2001. Strategy and structure in developing countries: business groups as an evolutionary response to opportunities for unrelated diversification. Industrial and Corporate Change 10, 77–113.

Lemelin, Andre, 1982. Relatedness in the patterns of interindustry diversification. Review of Economics and Statistics 64, 646–657.

Linebaugh, Kate, 7 March 2012. The new GE way: go deep, not wide. Wall Street Journal. http://online.wsj.com/article/SB10001424052970204571404577257533620536076.html.

Miller, Douglas J., 2006. Technological diversity, related diversification, and firm performance. Strategic Management Journal 27, 601–619.

Mitton, Todd, Vorkink, Keith, 2010. Why do firms with diversification discounts have higher expected returns? Journal of Financial and Quantitative Analysis 45, 1367–1390.

Mueller, Dennis C., 1969. A theory of conglomerate mergers. Quarterly Journal of Economics 83, 643–659.

Ozbas, Oguzhan, Scharfstein, David S., 2010. Evidence on the dark side of internal capital markets. Review of Financial Studies 23, 581–599.

Palich, Leslie E., Cardinal, Laura B., Miller, C. Chet, 2000. Curvilinearity in the diversification–performance linkage: an examination of over three decades of research. Strategic Management Journal 21, 155–174.

Panzar, John C., Willig, Robert D., 1981. Economies of scope. American Economic Review 71, 268–272.

Pehrsson, Anders, 2006. Business relatedness and performance: a study of managerial perceptions. Strategic Management Journal 27, 265–282.

Penrose, Edith G., 1959. The Theory of the Growth of the Firm. Wiley, New York.

Teece, David J., 1980. Economies of scope and the scope of the enterprise. Journal of Economic Behavior and Organization 1, 223–247.

Teece, David J., 1982. Towards an economic theory of the multiproduct firm. Journal of Economic Behavior and Organization 3, 39–63.

Teece, David J., 1986. Profiting from technological innovation. Research Policy 15, 285–305.

Teece, David J., 2007. Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance. Strategic Management Journal 28, 1319–1350.

Teece, David J., Rumelt, Richard, Dosi, Giovanni, Winter, Sidney, 1994. Understanding corporate coherence: theory and evidence. Journal of Economic Behavior and Organization 23, 1–30.

Villalonga, Belén, 2004. Diversification discount or premium? New evidence from the Business Information Tracking Series. Journal of Finance 59, 479–506.

Wan, William P., Hoskisson, Robert E., Short, Jeremy C., Yiu, Daphne W., 2011. Resource-based theory and corporate diversification: accomplishments and opportunities. Journal of Management 37, 1335–1368.

Wernerfelt, Birger, 1984. A resource-based view of the firm. Strategic Management Journal 5, 171–180.

Wernerfelt, Birger, Montgomery, Cynthia A., 1988. Tobin's q and the importance of focus in firm performance. American Economic Review 78, 246–250.

Williamson, Oliver E., 1975. Markets and Hierarchies. The Free Press, New York.

Diversity and Disagreement in Ethics: Philosophical Implications

Michele M Moody-Adams, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The extraordinary variety of ethical beliefs and practices that is possible within and across human societies has been a rich source of philosophical reflection. This article discusses important philosophical accounts, historical as well as contemporary, of what that diversity and disagreement come to. Philosophers ask what diversity and disagreement can tell us about the nature of ethical argument and inquiry, and they consider the implications of diversity and disagreement for ethical practice and social life. It is shown that philosophical contributions to these debates have often shaped and, in turn, been shaped by related work in the social sciences, and that as a result philosophers and social scientists frequently discuss similar methodological concerns. It is also argued that some of those concerns demand a response to debates about the nature of argument and inquiry in science, in order to address skepticism about the possibility of ever resolving serious ethical disagreement. Finally, it is shown that many recent philosophical discussions of diversity and disagreement address the ongoing challenge of adjudicating ethical and political conflict in culturally complex societies.

Diversity and Disagreement

The extraordinary variety of beliefs and practices that is possible within and across human societies has been a rich source of philosophical reflection. This article focuses on philosophical accounts of specifically ethical diversity and disagreement. Philosophers want to know what ethical diversity and disagreement might tell us about the nature of ethical argument and inquiry, as well as to consider the implications of diversity and disagreement for ethical practice and political institutions. This article discusses the most influential philosophical treatments of these issues, drawing on historical as well as contemporary accounts. It will be shown that these accounts have often shaped and, in turn, been shaped by related work in the social sciences, and that as a consequence they have frequently given rise to similar methodological worries. It will also be shown that these accounts will continue to be shaped by concern about how best to understand argument and inquiry in natural science, as well as by increasingly urgent worries about how to adjudicate ethical and political conflicts in culturally complex societies.

The extraordinary variety of ethical beliefs and practices that is possible within and across human societies, along with the ethical disagreement to which that variety frequently gives rise, has been a rich source of philosophical reflection. Philosophers have sought to understand the nature and causes of ethical diversity and disagreement, as well as what that diversity and disagreement might reveal about the nature of ethical argument and inquiry. They have investigated the similarities and differences between ethical inquiry and other kinds of inquiry, particularly scientific inquiry, and asked what the relevant similarities and differences can tell us about the nature of rationality. They have also analyzed the implications of ethical diversity and disagreement for ethical practice, and for the foundations of social and political life.

The Nature and Sources of Ethical Diversity and Disagreement

Philosophical Understandings of Anthropological and Sociological Evidence

Philosophical awareness of ethical diversity is as old as philosophy itself. It is evident in the ancient Greek beginnings of Western philosophy – particularly in the ethical and political thought of Plato and Aristotle. In the early and middle *Dialogues* of Plato (1961), for instance, exchanges between the Platonic Socrates and his interlocutors reveal many of the ethical conflicts and debates, and hence the ethical diversity and disagreement, that helped to shape the social and political life of fourth-century BC Greece. A generation later, Aristotle's sensitivity to the concrete contexts of deliberation and action, detailed in the *Nichomachean Ethics* and the *Politics*, reflected an even richer understanding of social and political diversity, and of the resultant complexity of ethical belief and practice. Neither Plato nor Aristotle believed that ethical diversity and disagreement required us to relinquish a commitment to ethical objectivity. But this stance was in marked contrast to that adopted by some of their philosophical forebears, especially fifth-century BC sophists such as Protagoras and Gorgias, to whom Plato's philosophy was in part a response. Moreover, the objectivist conceptions defended by Plato and Aristotle were met by the emergence of a powerful current of skepticism. This development culminated in the Academic skepticism of Cicero, and especially the second-century AD Pyrrhonian skepticism of Sextus Empiricus, in which ethical diversity and disagreement inevitably raised the specter of ethical relativism.

The sixteenth-century rediscovery and dissemination of Sextus' writings coincided with the beginnings, in the West, of a period of extraordinary exploration and economic expansion. Both developments were important in shaping the rise of early modern Western philosophy in the late sixteenth and early seventeenth centuries. European exploration and expansion

produced a body of literature purporting to offer a richer picture of human ethical variety than anything to be found in ancient philosophical texts. These accounts had a profound effect on early modern thinkers who pondered the nature and sources of ethical variety. Thinkers such as Montaigne (1958) came to accept the indispensability to philosophy of what we would now call sociological or anthropological explanations of ethical diversity. But, deeply influenced by the Pyrrhonian skepticism he found in Sextus, Montaigne also insisted that fully appreciating the details of human ethical diversity would always lead a reasonable person to accept ethical relativism. Montaigne thus rejected the claims of philosophers, from Plato onward, who believed that one could render the existence of ethical diversity and disagreement compatible with ethical objectivism.

In the seventeenth, eighteenth, and early nineteenth centuries, debates about the implications of ethical diversity and disagreement – based primarily on the testimony of explorers, settlers, and economic adventurers – continued to shape philosophical reflection about ethics. That reflection took one of its most interesting forms in Montesquieu's *Persian Letters* (1973 [1724]), in which Montesquieu's potentially dangerous criticisms of ethical and political life in eighteenth-century France are presented to the reader as the contents of letters written home by two fictional Persian visitors. But with the rise of modern anthropology in the late nineteenth and early twentieth centuries, philosophical reflection about diversity and disagreement took on a new tone. Philosophers were gradually persuaded by the claims of anthropology to provide expert testimony – rooted in theoretically sophisticated empirical research, rather than idiosyncratic personal or economic interests – about the nature and extent of diversity and disagreement across cultures and social groups.

Yet the cultural ascendancy of anthropological claims about diversity also reinvigorated the skeptical challenge to ethical objectivism. Ruth Benedict's Patterns of Culture (1934) provides an instructive and influential expression of this development. Benedict claimed to have given ethical relativism a scientific foundation for the first time. Some philosophers were prepared to accept this account of the matter and to cede authority for assessing the claims of relativism to empirical social science. But Benedict's claims of a newly 'scientific foundation' for relativism may be difficult to sustain. Her formulation and defense of relativism seem simply to reiterate Montaigne's argument that we have "no other test of truth and reason than the example and pattern of the opinion and customs of the country we live in." Melville Herskovits (1972) went on to argue that modern anthropology finally provided "tough-minded" empirical support for claims about the "ethnocentrism" of objectivist analyses of ethical disagreement. But Herskovits's efforts to defend charges of ethnocentrism do not go much beyond Montaigne's contention that "each man calls barbarism whatever is not his own practice." The fact that so many familiar defenses of relativism predate the rise of modern anthropology has led some philosophers to wonder whether ethical relativism is simply a reasonable inference from ethical diversity and disagreement, or a pre-theoretical assumption that unreflectively predetermines the way some in which some theorists go on to characterize the data of diversity.

Yet other philosophers have undertaken detailed investigation to show the value and importance of anthropological study to philosophical reflection about ethics. Some have conducted their own ethnographic studies of the ethical beliefs and practices of Native American groups, in an effort to determine the empirical reliability of anthropological accounts (Brandt, 1959; Ladd, 1957). Others – spurred on, in part, by the late work of Ludwig Wittgenstein, especially by the concept of 'forms of life' in his *Philosophical Investigations* (1978) – have interrogated the theoretical underpinnings, as well as the broader philosophical implications, of the social scientist's claims about diversity and disagreement (Wilson, 1970; Winch, 1972). These efforts introduced important new empirical detail into philosophical discussions of the extent of cultural diversity – details drawn, most famously, from the work of E.E. Evans-Pritchard on the Azande (1937), and from the work of Edmund Sapir (1985 [1949]) and Benjamin Whorf (1956) on the conceptual variety embedded in human languages. The resulting debates raise new and controversial philosophical questions about the nature of ethical disagreement – particularly the question whether human ethical systems could ever contain wholly incommensurable ethical (or other) concepts. Philosophical attempts to understand the origins and causes of ethical diversity and disagreement have frequently relied on generally utilitarian or consequentialist accounts of that diversity. On these accounts, ethical practices (and their associated beliefs) come into existence, and continue in existence, primarily because they serve the aims of the groups that develop them. From the eighteenth century forward, such accounts are usually linked with essentially evolutionary hypotheses about the development of ethical practices. The ethical and political thought of David Hume, for instance, attributes the rise of particular ethical practices to unconscious, or unintended, processes that are vaguely akin to natural selection. But accounts of the processes in question usually have more in common with Herbert Spencer's theory of social evolution than with Darwin's claims about the role of natural selection in biological evolution. William Graham Sumner's *Folkways* (1907) – which combined evolutionary ethics with a forceful defense of ethical relativism – contains what became one of the most influential expositions of just such a view.

Methodological Concerns about the Idea of Fundamental Disagreement

During the mid-twentieth century, many philosophers accepted the claims of relativist anthropologists that the data of comparative anthropology could firmly ground an argument for meta-ethical relativism: the view that there is no rational method for adjudicating between the conflicting ethical claims of different cultures. Yet as many of their critics objected the mere facts of ethical diversity and disagreement within and across cultures cannot be sufficient to establish the truth of ethical relativism (in any form). The objection is that merely from the fact of cultural variability we cannot confidently conclude that any resultant ethical disagreements will necessarily be disagreements in fundamental ethical commitments between which no rational adjudication is possible, and establishing this conclusion is essential to establishing the plausibility of ethical relativism.

This objection points to a way in which ethical relativism may, in fact, be fundamentally underdetermined by all available empirical evidence. For it is at least possible, as the psychologist Karl Duncker argued, that principles of ethical valuation might remain 'invariant' across the most diverse cultures, even while the 'situational meanings' of the acts, traits, and states of affairs subject to ethical evaluation could vary dramatically across those same cultures (Duncker, 1939). In short, cross-cultural ethical diversity might be properly traced not to differences in fundamental ethical principles but to cultural variation in derivative ethical principles – that is, to principles embodying local understandings of how to apply fundamental ethical principles to particular actions, character traits, and states of affairs. Thinkers who claim to identify fundamental ethical disagreement must, therefore, be able to show that they have identified a conflict between genuinely fundamental ethical principles or beliefs, and not between derivative principles or beliefs intended to apply the more fundamental principles or beliefs. Further, as the philosopher R.B. Brandt once noted, it is very difficult to find two cultures, or even two individual people, who can be shown to attribute identical nonmoral properties to some action, trait, or state of affairs and who nonetheless give the relevant action, trait, or state of affairs different ethical valuations (Brandt, 1959, 1967).

Identifying these difficulties does not prove that it is impossible to find convincing evidence of genuinely fundamental ethical disagreement. Brandt himself claimed to have found at least one such disagreement, in spite of the methodological difficulties he believed to attend any effort to do so (Brandt, 1959). Moreover, the demand for conceptual clarity in defining a disagreement as 'fundamental,' rather than 'derivative,' is difficult to meet, because distinguishing fundamental from derivative ethical beliefs and principles is no simple matter. In particular, it is not clear how to find a criterion for describing a belief or principle as fundamental or derivative that does not presuppose a commitment to some fairly complex elements of an ethical conception – that is, to ethical notions that may themselves be the subject of considerable controversy and disagreement. Yet in coming to appreciate the methodological and conceptual difficulties inherent in attempts to establish fundamental ethical disagreement, the empirical researcher and the philosophical interpreter alike are reminded that there are important obstacles to surmount if one seeks to draw a supportable inference for meta-ethical relativism from the fact of cross-cultural disagreement.

Individual Uncertainty and the Roots of Ethical Pluralism

Reliance on the evidence of cross-cultural variability is not the only means – or even the most important means – by which philosophers have sought to understand the nature and sources of ethical diversity and disagreement. According to some philosophers, the decision-making processes of an individual agent experiencing ethical uncertainty can sometimes tell us as much (if not more) about the nature and sources of ethical diversity and disagreement than the comparative anthropologist's claims about cross-cultural conflict. On these accounts, when a sincere and reflective agent attempts to choose between equally compelling, but conflicting and even apparently irreconcilable, ethical ends, what that agent typically confronts is the fundamental plurality and diversity (and, on some views, sometimes even the ultimate incommensurability) of available and plausible ethical ends. Pluralists such as Isaiah Berlin (1969) and Charles Taylor (1985) believe that philosophers like Plato, along with those of his intellectual descendants who would defend the unity of ethical value, simply fail to understand that plurality and diversity are intrinsic features of ethical value. According to the ethical pluralist, ethically compelling ends and purposes cannot all be reduced to a single measure of ethical value; nor can they all be realized in a single life. This means that even the sincere and reflective agent (or, perhaps, especially the sincere and reflective agent) will eventually confront what Thomas Nagel has called the fragmentation of ethical value (Nagel, 1979).

Those who understand ethical value as intrinsically plural and diverse thus have a way of explaining diversity in the beliefs and practices of various cultural and social groups in terms that go beyond any available on the standard anthropological account. On a pluralist conception of ethics, cross-cultural ethical diversity is simply an important – and essentially unavoidable – reflection of the fundamental plurality and diversity of ethical value itself. More controversially, some ethical pluralists go on to assert that neither individual human beings nor cultures and societies can escape fundamental ethical conflict – conflict, that is, between irreconcilable or incommensurable fundamental ethical principles and concepts. But some of the same difficulties inherent in the effort to defend the thesis of fundamental conflict across cultural groups, particularly the challenge of defining the contrast between derivative and fundamental ethical principles and beliefs, will attend any effort to show that no individual agent can avoid fundamental ethical conflict.

Following the lead of Isaiah Berlin (1969), some pluralists contend that any individual agent who claims to experience no fundamental ethical conflict is in some way rationally deficient, or else deeply self-deceived. Yet there are many conceptions of human flourishing (both secular and religious) which purport to make fundamental ethical conflicts effectively impossible by identifying or articulating a single 'dominant end' – as John Rawls has described it – to which all other human ends must be subordinate, and with respect to which the subordinate ends can be unified. Such conceptions, particularly when combined with a willingness to allow other agents to reject them, have often been thought to embody defensible understandings of rational choice and rational agency. Moreover, some ardent defenders of the view that dominant-end conceptions of human flourishing are in some way deficient or irrational actually acknowledge that these conceptions do not intrinsically violate principles of rational choice. But they nonetheless insist that there remains something ultimately incomplete, or perhaps even irrational, about a life lived in accordance with such a conception (Rawls, 1971; Hampshire, 1983). In response to such assertions, the question of whether a compelling defense of ethical pluralism requires a commitment to the inescapability of fundamentally irresolvable ethical conflicts – across social

and cultural groups, or in individual lives, or both – remains a subject of lively philosophical debate (Gowans, 1987; Stocker, 1989; Moody-Adams, 1997).

Implications of Disagreement for Understanding Ethical Argument and Inquiry

Historically, philosophers have held quite differing views about what the facts of ethical diversity and disagreement might tell us about the nature of ethical argument and inquiry. The standard views can be divided up in terms of the position they take with regard to the objectivity of ethics – that is, with regard to the question whether there can be at least some right answers to ethical questions, where the rightness of the answer is independent of human beliefs and practices. But there is great variety within both the objectivist and the nonobjectivist positions, particularly with regard to important epistemological questions (questions of whether and how we can have moral knowledge) and serious metaphysical concerns (concerns about the existence or non-existence of objective value entities).

Varieties of Ethical Objectivism and Serious Ethical Disagreement

The most familiar variety of ethical objectivism is *absolutist ethical objectivism*, on which there is at most one right answer to any given ethical question, and according to which the ethical correctness of that answer does not depend upon human beliefs and practices. Plato's *Republic* contains one of the most fully developed and widely influential versions of this conception. The 'Allegory of the Cave,' in Book Five of the *Republic*, richly depicts the underlying epistemological and metaphysical assumptions. According to the allegory, the inhabitants of a cave who remained chained to a wall, seeing only the shadows of real things by the light of a humanly built fire, are deprived of any possibility of rationally understanding what is real, and their beliefs and practices reflect their ignorance. Only the person who is released from ignorance and illusion by being unchained, and allowed to leave the cave, can become capable of a rational appreciation of the real essences of things as illuminated by the genuine light of the sun. The metaphysical commitments associated with this view – in Plato's doctrine of the Forms – are developed in surrounding passages in the *Republic*, and in other Dialogues like the *Symposium* and the *Phaedrus*. On the resulting view, what the relativist takes to be fundamental ethical diversity and disagreement is simply a function of moral ignorance rooted in the tendency to our experience of imperfect copies and shadows of the real as though they might reveal the essence of the ethical.

But absolutist ethical objectivism need not involve this kind of metaphysical realism about value entities existing independent of human belief and practice, or even any kind of realism about independently existing value entities or antecedently given moral facts. Nor need it involve the potentially problematic epistemological assumptions needed to explain how anyone might come to know the Form of the Good, and why so few seem genuinely capable of doing so. Immanuel Kant held a constructivist conception of ethical objectivity, on which humanly constructed practices (and their underlying beliefs and assumptions) are objective when they have been constructed in accordance with what Kant took to be the objectively valid (and already known) supreme principle of morality, the Categorical Imperative. In its most basic formulation, the Categorical Imperative directs the agent to act only on those policies (reasons) that he or she could will to serve as universal laws of action. What the relativist identifies as fundamental ethical disagreement is for Kant simply evidence that one or more parties to the disagreement has failed to subject the policy underlying a practice or a proposed course of action to the rational scrutiny required by the Categorical Imperative.

Some ethical pluralists believe that ethical objectivism need not be linked with the kind of absolutism found in the work of Plato or Kant. But whether ethical pluralism is indeed compatible with ethical objectivism remains a subject of rich, and ongoing, philosophical debate. How far can the pluralist go in rejecting the absolutist's 'one right answer' approach and remain an ethical objectivist? Many think that the most plausible pluralist candidates for the label 'ethical objectivist' are forms of ethical pluralism that recognize the possibility of diverse answers to disagreements involving derivative principles, yet retain the 'one right answer' approach to analyze disputes involving fundamental principles. Of course, any such views must confront the problem of reliably distinguishing fundamental from derivative ethical principles. Moreover, while there is no evidence that this problem is not, in principle, capable of a compelling solution, some ethical pluralists avoid the problem altogether by defending a limited and moderate relativism. On these accounts, the relativist rightly posits the possibility – and the frequent occurrence – of at least some fundamentally irresolvable ethical disputes.

Varieties of Nonobjectivism and Serious Ethical Disagreement

Meta-ethical relativism is, perhaps, the most familiar form of ethical nonobjectivism. Its main claim, as noted above, is a negative thesis about the impossibility of rationally adjudicating (purportedly) fundamental ethical disagreements. But some meta-ethical relativists also attempt to defend a normative ethical relativism that makes a positive claim about the content of morality and about the nature and sources of ethical disagreement. In its bolder formulations, normative ethical relativism asserts the 'equivalence' of all ethical practices and beliefs. This version of the view has some important implications for a relativist approach to ethical practice – most often resulting in a demand to 'tolerate' all ethical practices as 'equally valid.' According to more modest versions of normative ethical relativism, we cannot confidently assert the superiority of one ethical stance over another with which it somehow conflicts, and we should therefore withhold judgment as a recognition of ethical uncertainty.

Meta-ethical relativism starts from the facts of ethical diversity and disagreement, but some forms of nonobjectivism posit very different explanations of those facts. One such view is ethical subjectivism, according to which ethical claims are fundamentally statements about the individual speaker's subjective feelings (not about the practices of a particular

social or cultural group). What the relativist identifies as a fundamental ethical disagreement is, for the ethical subjectivist, simply evidence of the deep-rooted feelings of the parties to the dispute. Yet both meta-ethical relativism and ethical subjectivism allow that ethical statements have some cognitive content. On the subjectivist views, ethical statements tell us something meaningful and informative about the feelings of the speaker – and indirectly about the feelings of any similarly constituted persons. On meta-ethical relativism, in contrast, ethical statements tell us what members of a particular social or cultural group believe to be good – and, ultimately, about the local social worlds in which ethical practices arise and have their (limited) application (Williams, 1985).

It is important that some nonobjectivist views find ethical claims to be cognitively substantive, because there are varieties of nonobjectivism that deny that ethical statements have any cognitive content at all. The most influential and controversial such view is the theory of emotivism, first articulated in the early twentieth century by A.J. Ayer (1946). According to Ayer, fundamental ethical judgments (he called them "pure judgments of value") are nonrational, noncognitive expressions of emotion that cannot be true or false. Even the apparently most serious ethical disagreement, for Ayer, is not much different from a round of energetic 'boos' and 'yeas' for the opposing teams at a sporting event. Ayer's critics charged that such claims appear to deny that ethical disagreements ever occur. Ayer's efforts to mollify his critics were problematic: moral disagreements occur, he responded, but only as disagreements "without formal contradiction." This hypothesis can be confirmed, he urged, by the tendency for the parties to such disagreements to resort to "mere abuse" when arguments fail.

The Infirmity of Ethics Compared with Science?

Ethical Disagreement as a Source of Skepticism about Ethical Methods

Yet despite its counterintuitive claims about ethical disagreement, and its implausibly narrow understanding of what it is to have cognitive content, emotivism was a powerful and challenging embodiment of skepticism about the rationality of ethical argument and inquiry – a skepticism firmly rooted in the frequent intractability of ethical disagreements. There is a special irony in the fact that emotivism seems to deny the reality of the very phenomenon underwriting its skeptical challenge to the methods of ethics. But for many philosophers, even those who would explicitly reject emotivist understandings of ethical language, the emotivist's challenge to ethics retains a certain potency. The continued existence of intractable ethical disagreements, for these thinkers, is evidence (along with the explanatory and predictive success of science) that ethics, in the words of logician and philosopher of science W.V.O. Quine, is "methodologically infirm" compared with science.

Alternatives to Skepticism

Yet the conception of science as the arbiter of rationality, capable of supplying standards that must be met by any domain of rational argument and inquiry, has been challenged. Its critics object that it wrongly presumes that all domains of human inquiry might have the same aim. They maintain that ethics and science are fundamentally different kinds of enterprises – one claiming to explain and predict natural phenomena, the other seeking to help us to lead lives worth living – and that there is no reason to expect similarities between the methods of ethics and the methods of science. To be sure, some ethicists continue to search for fundamental similarities between the methods and aims of ethics and science. But their critics urge that such efforts can only make ethics look like a poor and distant relation of science, and that we distort the aims and purposes of ethics by trying to make it conform to structures of argument in science. Still further, these critics reject the notion that the persistence of disagreement in ethics – even intractable disagreement – is intrinsically a sign of the failure of ethical methods. Might it not be, they ask, that in some domains of human argument and inquiry disagreement is an important catalyst for serious reflection, and a crucial stimulus to avoiding smugness and self-righteousness?

Implications of Ethical Disagreement for Practice

The Merits and Demerits of Tolerance

The existence of ethical diversity and the persistence of ethical disagreement nonetheless pose important difficulties for ethical practice. How are we to respond to those whose ethical beliefs and practices differ, sometimes quite substantially, from our own? The ethical objectivist will argue that in some or even all ethical disputes, one or more parties to the dispute has made some kind of error in moral reasoning or perception. Ethical relativists, and some pluralists, who argue for tolerance in response to ethical disagreements, urge that the objectivist response is too frequently associated with moral smugness and moral self-righteousness, and potentially dangerous and culturally imperialist ethnocentrism. But these associations are not intrinsic, and the demand for tolerance poses its own dangers when it functions (as it sometimes has) to sanction local practices that involve the violent coercion of innocent victims, or murderous intolerance of internal cultural dissent. The demand for tolerance also raises urgent questions about theoretical consistency. If it is intended as a nonrelative demand, incumbent on any agent, then it is clearly incompatible with the relativist stance. If it is aimed only at the inhabitants of some societies – say Western liberal societies – then it violates relativist structures in demanding the extension of what may well be a locally compelling principle to an intercultural context. Still further, uncritical tolerance of ethical diversity and disagreement must not be confused with respect for cultural diversity. Many philosophers urge that we sometimes show the greatest respect for diverse ways of life by being ready to engage in open and noncoercive discussion with those who represent them.

The Dangers and Incoherence of Ethical Isolationism

Some relativists are so concerned about the dangers of ethnocentrism and the imperialist overreaching of some cultures that

they insist on an ethical isolationism. According to the ethical isolationist, cultural outsiders can never understand ethical practices and beliefs associated with a genuinely unfamiliar way of life, and thus have no rational basis for any kind of critical reflection on – let alone outright criticism of – those practices and beliefs. But ethical isolationists typically defend their views by detailing their own special understanding of the practices and beliefs of the unfamiliar 'other,' richly conveying facts about those practices and beliefs in order to convince other cultural outsiders of the limits of their understanding. This tack shows how difficult it is to defend ethical isolationism except in terms that reveal the inconsistency and implausibility of the view.

Pluralism and Multiculturalism

Yet philosophers who reject ethnocentrism, relativism, and isolationism must still confront the persistence of ethical diversity and disagreement and the challenges they pose, especially in the culturally complex societies found in large, modern nation-states. An increasingly influential response, associated with some forms of ethical pluralism, is the doctrine of multiculturalism, which in its most familiar form recommends that we act and judge on the presumption that the ethical beliefs and practices of every way of life are in principle valuable and worthy of respect. But compelling philosophical discussions of multiculturalism have construed this presumption as, at best, a 'starting hypothesis,' maintaining that a final verdict on the worth of any practice must always await the results of respectful, but sustained, critical reflection (Taylor, 1992). In response, some philosophers have urged adoption of a critical multiculturalism (as a form of objectivist ethical pluralism) that leaves open the possibility that reflection on almost any practice might generate rationally compelling grounds on which to reject the practice as ethically indefensible (Moody-Adams, 1997).

The question of how to put multiculturalism into practice has also been a central topic of debate in contemporary ethical and political philosophy. Some theorists have developed a model of 'multicultural citizenship' for culturally complex liberal democracies that is centered around formal political accommodation of cultural minorities (Kymlicka, 1996; Modood, 2007). These efforts have generated resistance from theorists who cite conflicts with the individualist assumptions of liberal political philosophy, and with the secular neutrality thought to be central to the most familiar variants of liberal thought. In other contexts, political debates about whether the idea of nationality can be separated from the idea of a single defining ethnic and cultural heritage continues to generate controversy about the possibility of a politically robust multiculturalism (Parekh, 2002). The challenge of moving beyond the theoretical and practical dimensions of this impasse between multiculturalism and its critics will remain an important element of normative ethics and political philosophy in the twenty-first century.

See also: Democracy: Normative Theory; Discrimination and the Law; Discrimination: Racial; Ethics and Values; Multiculturalism; Relativism, Cognitive; Value Pluralism.

Bibliography

Ayer, A.J., 1946. Language, Truth, and Logic. Pelican, Harmondsworth, UK.
Benedict, R., 1934. Patterns of Culture. Houghton Mifflin, Boston.
Berlin, I., 1969. Two concepts of liberty. In: Berlin, I. (Ed.), Four Essays on Liberty. Oxford University Press, New York.
Brandt, R., 1959. Ethical Theory. Prentice-Hall, Englewood Cliffs, NJ.
Brandt, R.B., 1967. Ethical relativism. In: Edwards, P. (Ed.), Encyclopedia of Philosophy, vol. 3. Macmillan Press, New York.
Duncker, K., 1939. Ethical relativity? (An inquiry into the psychology of ethics). Mind 48, 39–57.
Evans-Pritchard, E.E., 1937. Witchcraft, Oracles and Magic among the Azande. Oxford University Press, Oxford, UK.
Gowans, C.W. (Ed.), 1987. Moral Dilemmas. Oxford University Press, Oxford, UK.
Hampshire, S., 1983. Morality and Conflict. Harvard University Press, Cambridge, MA.
Herskovits, M.J., 1972. Tender and tough-minded anthropology. In: Herskovits, F. (Ed.), Cultural Relativism: Perspectives in Cultural Pluralism, first ed. Random House, New York.
Kymlicka, Will, 1996. Multicultural Citizenship. Oxford University Press, New York, USA.
Ladd, J., 1957. The Structure of a Moral Code. Cambridge University Press, Cambridge, UK.
Modood, Tariq, 2007. Multiculturalism. Polity Press, Cambridge, UK.
Montaigne, M., 1958. The Complete Essays (D. Frame, Trans.). Stanford University Press, Stanford, CA.
Montesquieu, C.S., 1973 [1724]. Persian Letter (C.J. Betts, Trans.). Penguin, Harmondsworth, UK.
Moody-Adams, M.M., 1997. Fieldwork in Familiar Places: Morality, Culture and Philosophy. Harvard University Press, Cambridge, MA.
Nagel, T., 1979. The fragmentation of value. In: Nagel, I. (Ed.), Mortal Questions. Cambridge University Press, Cambridge, UK.
Parekh, Bhikhu, 2002. Rethinking Muilticulturalism: Cultural Diversity and Political Theory. Harvard University Press, Cambridge, MA.
Plato, J., 1961. The Collected Dialogues. Pantheon Books, New York.
Rawls, 1971. A Theory of Justice. Belknap Press of Harvard University Press, Cambridge, MA.
Sapir, Edward, 1985 [1949]. In: Mandelbaum, David G. (Ed.), Selected Writings in Language, Culture, and Personality. University of California Press, Berkeley, CA.
Stocker, M., 1989. Plural and Conflicting Values. Oxford University Press, Oxford, UK.
Sumner, W.G., 1907. Folkways. Ginn Publishers, Boston, MA.
Taylor, C., 1985. The diversity of goods. In: Taylor, C. (Ed.), Philosophy and the Human Sciences: Philosophical Papers, vol. 2. Cambridge University Press, Cambridge, UK.
Taylor, C., 1992. The politics of recognition. In: Gutmann, A. (Ed.), Multiculturalism and the Politics of Recognition. Princeton University Press, Princeton, NJ.
Whorf, B., 1956. Language, Thought and Reality. In: Carroll, John G. (Ed.), Selected Writings of Benjamin Lee Whorf. MIT Press, Cambridge, MA.
Williams, B., 1985. Ethics and the Limits of Philosophy. Harvard University Press, Cambridge, MA.
Wilson, B. (Ed.), 1970. Rationality. Harper and Row, Evanston, IL.
Winch, P., 1972. Understanding a primitive society. In: Winch, P. (Ed.), Ethics and Action. Routledge and Kegan Paul, London.
Wittgenstein, L., 1978. In: Wittgenstein, L. (Ed.), Philosophical Investigations. Basil Blackwell, Oxford, UK.

Diversity in organizations

Jorge A Gonzalez and Azadeh Zamanian, University of Texas-Pan American, Edinburg, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of diversity in organizations addresses the implications of workplace diversity from social justice, legal compliance, and organizational performance standpoints. Diversity in groups and organizations is associated with social categorization and information/decision-making processes, which respectively are related to adverse and beneficial attitudes and behaviors. Diversity scholars have attempted to shed light into these mixed results by addressing the role of categorical and relational demography, diversity types, curvilinear and interactive effects, and complex conceptualizations of diversity. Diversity management strategies, such as affirmative action and management-led diversity initiatives, have been designed to reduce discrimination and stimulate the beneficial effects of diversity.

Diversity refers to the extent in which members of an entity, such as a group or organization, differ from one another. This encompasses a very broad range of individual attributes, although most attention has been given to differences in gender, race/ethnicity, and age. Diversity in attributes, such as nationality, culture, education, function, tenure, ability, sexual orientation, religion, values, personality, goals, and many others is gaining attention. The field is mainly concerned with the outcomes and management of individual differences in organizational settings, including work groups and organizations. Its importance lies on the fact that the workplace in many nations is growing increasingly diverse. For instance, women and racial and ethnic minorities now constitute the majority of workers in the USA (Bureau of Labor Statistics, U.S. Department of Labor, www.BLS.gov).

Diversity and the Goals of Its Study

Most diversity research addresses the implications of workplace diversity taking a social justice, legal compliance, or organizational performance standpoint. Social justice goals include the reduction of bias and discrimination, and the enhancement of inclusion in the workplace. These objectives center on increasing the presence and inclusion of traditionally underrepresented or disenfranchised social groups: women, racial/ethnic minorities, and older workers. The social justice perspective includes the study of workplace discrimination and the implementation of diversity management programs. Such programs may be designed to meet legal compliance goals, such as following affirmative action and equal opportunity legislation. Nonetheless, diversity management also includes management-led diversity initiatives, which tend to be motivated by social justice objectives, and designed to stimulate the inclusion of a range of social backgrounds broader than the one established by the law.

Organizational performance goals rely on the value diversity offers to organizations (Cox and Blake, 1991). Consistent with the 'business case' for diversity or 'value-in-diversity' hypothesis, some scholars suggest that diversity is a source of sustained competitive advantage and leads to greater organizational performance. For instance, with basis on the resource-based view of the firm, organizational racial diversity is considered a resource that is valuable, rare, and difficult to imitate and thus a source of sustained competitive advantage (Richard, 2000). These ideas emphasize the benefits organizations draw from their compatibility with constituents in a complex and diverse environment, such as potential employees, business partners, and customers. Therefore, social justice, legal compliance, and performance goals are not necessarily incompatible. Government legislation, such as affirmative action emphasizes the accomplishment of social justice goals through legal compliance. Moreover, the 'business case' for diversity is based on the appeal of the idea that organizational performance and social justice goals are compatible.

Diversity Outcomes

Despite the appeal of the business case for diversity, reviews of the literature have revealed that diversity in groups and organizations has the potential for both positive and negative outcomes (van Knippenberg and Schippers, 2007; Williams and O'Reilly, 1998)

These seemingly paradoxical results have been synthesized into two theoretical perspectives: social categorization and information/decision making. The social categorization perspective describes how people categorize themselves and others into groups to make sense of a complex environment. It relies on social identity and social categorization theories, which posit that people display in-group favoritism in their social interactions. These theories are complemented with the similarity-attraction paradigm, which holds that people are attracted to others who possess similar psychological attributes, but infer similarity through observable demographic attributes. Diverse groups and organizations are subject to faulty social categorization processes and therefore have more affective conflict than homogeneous ones (Ashkanasy et al., 2002), as well as lower trust, communication, cooperation, helping, cohesiveness, satisfaction, and commitment. For these reasons, diverse groups and organizations may ultimately perform poorly in comparison with homogeneous ones.

In contrast, the information/decision-making perspective holds that diversity is related to a wider range of, skills, abilities, expertise, and perspectives (Cox and Blake, 1991; Watson et al.,

1993). This variety in knowledge provides a broad cognitive resource base that permits greater creativity and innovation in group tasks and decision making. For instance, diversity reduces the possibility of groupthink – a premature consensus or conformity that leads to poor decisions. Instead, diversity enhances the opportunities for dissent, debate, the rise of a devil's advocate, and other factors associated with substantive or beneficial task-related conflict. In effect, this can lead to better decision making. For instance, divergent viewpoints in a team can result in team reflexivity – a careful consideration of its functioning – which can lead to increased team learning and effectiveness. Overall, diverse groups and organizations have the opportunity to be more effective thanks to a broader cognitive base.

An integration of these two perspectives suggests a more nuanced consideration of diversity and the intervening processes that lead to its attitudinal, behavioral, and performance outcomes. Different individual attributes may present different processes and outcomes, and a combination or interaction of such attributes may present complex processes. Diversity effects may also be contingent on a variety of moderators or contextual circumstances. Furthermore, it is important to consider that diversity is a compositional construct, meaning that it is an attribute of groups and organizations rather than individuals, but individual-level dissimilarity from others also has an important role in the field.

Categorical and Relational Demography

Research about diversity at the individual level of analysis is concerned with simple demographic effects, as well as the extent in which an individual is similar or different from referent others. The simple or categorical demography approach assesses the behaviors associated with specific individual attributes. This would include, for example, the study of the behavior associated with being Hispanic or comparing the attitudes or managerial attributes of men and women at work.

The relational demography approach refers to the extent in which an individual is different in an attribute or category from referent others, such as other members of a work group or a particular person (Riordan, 2000). Relational demography studies typically take the social categorization perspective. Most studies have found that people who are different from others tend to be less committed to their organizations and satisfied with their jobs, and more likely to be absent and quit their jobs (e.g., Tsui et al., 1992). Studies concerned with dyads show similar results. For instance, being demographically different from a supervisor is related to lower employee commitment, a more adverse relationship, and lower assessment of employee performance from such supervisor (e.g., Tsui and O'Reilly 1989). Relational demography includes the effects tokenism, or being extremely different from others in a group, such as of being the sole woman in an otherwise completely male group. Tokenism emphasizes the particular effects of visibility and suggests the adverse effects of dissimilarity not only depend on the feelings, thoughts, and demographic identity salience of the token person, but also on the way that other group members perceive and treat such person.

Relational demography can be combined with categorical demography and group diversity to cross levels of analysis in its study. Such interaction approach can reveal asymmetrical social group effects. For example, relational demography studies have shown that being different in terms of gender or race is typically more detrimental to work attitudes for men and Whites than for women and non-Whites (e.g., Tsui et al., 1992). Interestingly, Chatman and O'Reilly (2004) found that women were more likely than men to be affectively attached to same-gender groups, but that women were also more likely to leave same-gender groups. This was attributed to status, which was higher in male-dominated groups. The interactive approach can combine categorical and relational demography effects with organizational-level diversity. This would suggest that relational demography effects depend on organizational diversity and would likely present weaker adverse effects in heterogeneous organizations.

Diversity Typologies

As stated earlier, diversity encompasses a broad range of individual attributes. Many diversity studies focus on a single attribute (e.g., 43% of diversity studies reviewed by Jackson et al., 2003). One reason may be that scholars and managers are often interested in a single dimension of diversity. For instance, people are often seen as having either gender or a race, but not both. Nonetheless, diversity scholars and managers are increasingly paying attention to a multiplicity of individual attributes to answer the question of whether all differences make a difference. This has given rise to individual attribute typologies.

One typology describes whether individual attributes are task related or not. In this case, differences in education, function, and tenure provide a variety of task-related information, whereas differences in race/ethnicity, gender, and age do not (Pelled et al., 1999). The rationale behind this distinction is that task-related diversity stimulates substantive conflict, which drives people to consider multiple points of view, perspectives, and opinions, while nontask-related diversity leads to affective conflict that can be detrimental.

Another typology differentiates across surface- and deep-level diversity. Surface-level diversity refers to differences in observable demographic attributes, such as race, ethnicity, gender, and age, while deep-level diversity refers to differences in psychological attributes, such as personality traits, beliefs, and values (Harrison et al., 1998). This distinction holds that both surface- and deep-level diversity may harm group social integration, cohesion, and in effect performance. However, the effects of surface-level diversity weaken through time, while those of deep-level diversity become stronger as group members get to know one another. This typology diverges from early conceptualizations of organizational demography, which assessed cognitive (deep-level) diversity in management groups using demographic (surface-level) attributes as a proxy.

Despite the logical appeal of typologies, meta-analyses that have studied them, particularly the task/nontask distinction, have not provided conclusive results (e.g., Horwitz and Horwitz, 2007). These meta-analyses either have not shown different outcomes for task and nontask diversity on group cohesion and performance, or have found inconsistent evidence about the role of social integration processes. These inconclusive results stimulated the development of the

categorization-elaboration model, which suggests that both task and nontask diversity dimensions affect the elaboration of task-related information and social categorization processes (van Knippenberg and Schippers, 2007). A more recent meta-analysis exploring specific attributes rather than a task/non-task typology showed more nuanced effects. Functional and educational backgrounds were positively related to team creativity, innovation, and performance, while educational diversity was related to performance only in top management teams. Also, racial and sex diversity had small adverse effects on performance (Bell et al., 2011). Another meta-analysis suggested that the main effects of various forms diversity was small, but that its effect were strong once contingent moderators and contextual circumstances were considered (Joshi and Roh, 2009).

The Role of Context and Moderators in Diversity Effects

The role of diversity on behavior and performance depends on a variety of contextual factors. Meta-analytic data have showed that the main or direct effect of diversity attributes on team performance is very small, but that this relationship is much stronger when the role of contextual factors is incorporated (Joshi and Roh, 2009). Such contextual factors and moderators include industry setting, team interdependence, team longevity, leadership, task complexity, regulatory pressure, and market competition, among others. For instance, diversity is likely to lead to better outcomes when transformational leadership is high, and when teams are interdependent. Also, empirical evidence, particularly from laboratory studies, suggests that diverse groups tend to perform better than homogeneous ones in complex and creative tasks, while homogeneous groups perform faster and better in routine tasks.

The degree of diversity itself is a contextual factor that shapes its outcomes. Diversity can present curvilinear effects on attitudes and performance that are U-shaped, or even J-shaped. In other words, adverse outcomes may be present at moderate levels of diversity, but moderate diversity may not present such outcomes or even have positive effects at very high levels. For example, managerial racial diversity has a curvilinear, U-shaped association with organizational financial performance (Richard et al., 2004). Similarly, a study of cultural heterogeneity in multicultural teams found that moderate levels of diversity were associated with lower team performance, while highly homogeneous and highly heterogeneous teams presented lower social categorization problems. These effects were further contingent on team longevity (Earley and Mosakowski, 2000), suggesting that, given time, highly diverse teams tend to de-emphasize their demographic differences and form a hybrid team culture or common in-group identity.

Diversity in one type of attribute can also influence or moderate the effect of another attribute or type. For instance, diversity in deep-level attributes, such as values, personality, or time urgency can increase the effects of demographic diversity on attitudes, conflict, and performance (e.g., Jehn et al., 1999). Also, a collective culture – an aggregation of shared values that emphasize cooperation – can reduce demographic diversity effects on conflict (Chatman et al., 1998).

An organization's diversity climate also moderates diversity effects. Diversity climate refers to the shared organizational member perceptions of the fairness of the organization's diversity-related structural characteristics and values. A positive or inclusive diversity climate can counter the adverse effects of relational demography on workplace attachment, particularly for people of color. It can also enhance the positive effects of diversity on performance (Gonzalez and DeNisi, 2009; McKay et al., 2007). Diversity climate is similar to the concept of diversity perspectives or paradigms. Ely and Thomas (2001) described three different diversity perspectives, which refer to the manner in which organizational members express and manage diversity-related tensions. The discrimination-and-fairness perspective is present when organizational members believe that the point of diversity initiatives is to reduce discrimination and enhance fairness. In contrast, an access-and-legitimacy perspective consists of beliefs that the goal is to tap diverse markets more effectively. Also, an integration-and-learning perspective is present when people believe that cultural differences drive diverse knowledge and insight, and this third perspective is associated with the greatest degree of effectiveness (Ely and Thomas, 2001). In other words, a diverse organization is likely to perform better when its members believe that cultural differences drive diverse knowledge and insight than when members assume that diversity is only valuable to gain niche markets or to avoid blatant discrimination and lawsuits. The notions of diversity climate and perspective are similar to diversity culture (Cox and Blake, 1991) and mind-set (van Knippenberg and Schippers, 2007).

The effects of organizational racial and gender diversity on financial performance also depend on environmental factors. Consistent with the social categorization perspective and requisite variety arguments, racial and gender diversity-performance effects tend to be stronger in the service than in manufacturing. One reason may be that diverse firms are better able to best serve a diverse stakeholder and customer base. Studies addressing the role of diversity in organizational financial performance (Richard et al., 2004) also report a stronger positive relationship between racial diversity in firms that have an innovation strategy, have a narrow span of control, are in their early life cycle stages, and operate in munificent environments. Studies that have considered community demographics as context and explored the match between organizations and community demographics have reported both positive and negative effects on attitudes and performance (Joshi and Roh, 2009).

Complex Conceptualizations of Diversity

Complex conceptualizations of diversity include the combination or alignment of various attributes or types. This includes diversity faultlines, a correlation of multiple individual attributes that provide a clear basis for subgroup differentiation (Lau and Murnighan, 1998). For example, a group comprised of young, African-American men and older, White women shows the combination of three different attributes into subgroups. Diversity faultlines are related to heightened conflict, less communication, lower learning, and disrupted group functioning.

Diversity typically refers to the variety or dispersion or variety in terms of a specific individual attribute. However, it may also refer to separation or disparity (Harrison and Klein, 2007).

Separation refers to differences in position or opinions, such as in values, beliefs, or attitudes. A group that is fragmented due to different values or polarizing opinions would be an example. Disparity refers to vertical differences in the proportion of valuable assets and resources held by unit members, which acknowledges the important role of power and status differences. An example would be a group with inequitable pay or network access. Combinations of these three types are possible. For example, members of an ethnically diverse group may hold values that are similar within each ethnic group but distinct across ethnic groups, showing ethnic variety and value separation. Such ethnically diverse group could also present no differences in pay, which would mean having high variety but low disparity.

Diversity Management

The potential for positive, negative, and complex effects from diversity warrants efforts to manage diversity. Diversity management refers to the development and implementation of practices, processes, and systems designed to enhance workplace diversity, exploit its benefits, and minimize adverse outcomes. Diversity management includes affirmative action policies, which aim to comply with the law, as well as management-led diversity initiatives, which refer to formalized human resource management practices designed to reduce bias and discrimination, promote and sustain organizational diversity, and improve performance. Affirmative action predates the notion of diversity management, and many diversity scholars and consultants differentiate across them. The reason is that the focus of affirmative action is to comply with a legal mandate, while diversity management seeks to increase organizational performance, fairness, equality, and inclusion through management-led initiatives. Nonetheless, some scholars refer to management-led diversity initiatives as 'voluntary' affirmative action programs. A source of benchmarking on diversity management is the Web site and magazine DiversityInc, which publishes a list of the top companies who champion diversity (DiversityInc, www.DiversityInc.com).

Affirmative Action

Affirmative action refers to activities taken to reduce discrimination and provide employment and promotion opportunities for disenfranchised groups. Affirmative Action was initially enacted for US government contractors through executive orders by President Kennedy and Johnson in the 1960s, and was expanded by the Civil Rights Act of 1964. Title VII of the civil rights act prohibited organizations with more than 50 employees from discriminating on the basis of race, color, religion, sex, or national origin. It also established the Equal Employment Opportunity Commission (www.EEOC.gov) to oversee the law. Related legislation has been expanded to cover age, disability, pregnancy, veteran status, genetic information, and sexual orientation. The goal of affirmative action was to compensate for past discrimination, prevent ongoing discrimination, and provide equal employment opportunities to covered demographic groups. Similar antidiscrimination laws have been passed in many other countries. For example, Canada's Employment Equity Act covers 'visible minorities' (language and religion – relevant to Canada's English and French heritage – are invisible attributes), women, people with disabilities, and Aboriginal peoples. Also, Malaysia's New Economic Policy protects the majority Malay population, who traditionally has had lower access to business and income opportunities than the Malaysian Chinese minority.

Scholars and practitioners have discussed issues with employee acceptance of affirmative action. Some argue that the legal coercive pressure of affirmative action may neutralize rather than enhance equality. This occurs because affirmative action promotes the idea that beneficiaries are hired due to preferential treatment rather than their qualifications. For this reason, affirmative action can be associated with a stigma of incompetence among beneficiaries and associated feelings of unfairness on the part of the majority group. Other scholars posit that different programs have different degrees of perceived fairness and acceptance. In general, programs that promote hiring gain greater acceptance than other programs, such as those involving promotions, compensation, and training. In addition, program beneficiaries (e.g., African-Americans) typically report greater acceptance of affirmative action programs, particularly beneficiaries who report to have experienced discrimination before (Kalev et al., 2006).

Management-Led Diversity Initiatives

Unlike affirmative action, management-led diversity initiatives and other diversity management practices focus on reducing discrimination and improving the effectiveness of diversity organizations by changing its practices, culture, and climate. Diversity initiatives may be inclusive of individual attributes that are not covered by law. They are likely to emphasize the inclusion and of all social groups, including the majority group. Many diversity initiatives are designed to increase competitive advantage (Cox and Blake, 1991). Also, they may be designed to underscore the importance of acknowledging the uniqueness of employees and social group members as individuals, rather than their assimilation into the dominant culture (Shore et al., 2011). Despite their voluntary nature, management-led diversity initiatives may also have a backlash. For instance, employees may become distrustful or cynical and resist the implementation of diversity programs if prior efforts have failed in the past (Kalev et al., 2006). Examples of management-led diversity initiatives include targeted recruitment, retention, promotion, mentoring, and diversity training.

Organizations use a variety of targeted recruitment strategies to attract women and minority job applicants to increase diversity. For instance, organizations may adopt impression management techniques, such as displaying pictorial diversity and inclusive policy statements in their job marketing, recruiting from colleges with a large minority population, and employing female and minority recruiters (Avery and McKay, 2006). Research shows similarity-attraction race effects in job interviewer–applicant dyads, and that these effects influence interview quality, the interviewer's evaluation of the applicant, and the applicant's perceptions of the organization's diversity-related attributes. Gender effects are also present, but these tend

to be confounded with interpersonal attraction in cross-sex recruiter–applicant dyads. Recruitment suggestions from the literature include the use of structured interviews, which can serve to minimize the potential for bias in job interviews and applicant assessment. The use of employee informal networks may help the recruitment of women and minorities in organizations that already have some diversity. Workplace diversity initiatives also focus on retention. This is important due to the positive relationship between diversity and turnover, and because women and people of color in general turn over at a higher rate than Whites and men. Retention efforts may also be necessary after targeted recruitment, since failing to meet the expectations of an unrealistic job preview may result in future turnover (Avery and McKay, 2006).

Diversity management also includes efforts to promote of women and racial/ethnic minorities into management positions. This can be meant to break the 'glass ceiling', which traditionally referred to the lack of top-level managerial-level representation of such groups. Mentoring is a strategy meant to stimulate such promotions. It refers to a formal developmental relationship between an individual (protégé) and a senior employee (mentor). Some perspectives favor same-gender and same-race mentoring for women and people of color because a common background tends to lead to a better mentoring relationship. For instance, women in same-sex mentoring relationships have more social interaction than in cross-sex relationships (Ragins and Cotton, 1999). Other perspectives favor cross-sex and cross-race mentoring in male and White-dominated settings to provide access to network contacts, power, and opportunities that otherwise would not be available. Regardless of the demographics, results from mentoring studies suggest that women and people of color protégés experience higher quality relationship and better career outcomes from informal than from formal mentoring.

Diversity training refers to formal efforts to enhance skills and knowledge about diversity. Diversity training programs may include the objective to provide employees with information about legal issues (e.g., equal opportunity laws and sexual harassment), increase employees' awareness and sensitivity to backgrounds other than their own, and develop communication and interpersonal skills helpful in diverse settings. The success of diversity training has been evaluated on a number of factors, such as the demographics of the trainer and training content. For example, diversity training designed to build skills tends to be evaluated better than programs intended to inform about legal compliance requirements or raise awareness. Factors associated with the adoption and perceived training success include the support of top management, mandatory attendance, evaluation of results, rewards for increasing diversity, and an inclusionary diversity definition (Rynes and Rosen, 1995).

A review of diversity initiatives stated that programs that established management responsibility – such as affirmative action, diversity committees, and diversity management positions – were more strongly associated with organizational managerial diversity than mentoring, networking, diversity training, and diversity audits or performance evaluations. Moreover, organizations with diversity initiatives establishing responsibility obtained better effects from diversity training and evaluations (Kalev et al., 2006).

Research Trends

The study of diversity has shown that diversity in groups and organizations has the potential for both positive and negative effects on attitudes, behavior, and performance. Meta-analytic data have shed some light into the manner different diversity attributes and types influence these outcomes (e.g., Bell et al., 2011), but a synthesis of these results is still inconclusive. These mixed results suggest that empirical efforts to explore the role of contextual moderators and to understand the processes involved in diversity effects will continue. The study of interactive and complex effects involving different diversity types, multiple levels of analysis, and faultlines hold the most promise. For instance, studies of faultlines could pay attention to various combinations of diversity attributes and types to reveal which combinations present the most harmful effects, and which combinations stimulate beneficial processes and outcomes. One should note that much research has been conducted from a US perspective, but diversity is a global issue. Studies that explore diversity in different nations and pay attention to their local demographics and relevant social conditions are likely to uncover novel diversity-related issues.

See also: Aging and Work; Gender and Women's Studies, Applied Research On; Personality and Values at Work; Personnel Selection, Psychology of; Population Composition by Race and Ethnicity: North America; Power, Politics, and Influence in Organizations; Race Identity; Race and Racism in the Twenty-First Century; Race and Racism, Geography of; Race and the Law; Race: Ethnicity and Health; Racial Relations; Racism, Sociology of; Teamwork and Team Performance Measurement.

Bibliography

Ashkanasy, N.M., Härtel, C.E., Daus, C.S., 2002. Diversity and emotion: the new frontiers in organizational behavior research. Journal of Management 28, 307–338.

Avery, D.R., McKay, P.F., 2006. Target practice: an organizational impression management approach to attracting minority and female job applicants. Personnel Psychology 59, 157–187.

Bell, S.T., Villado, A.J., Lukasik, M.A., Belau, L., Briggs, A.L., 2011. Getting specific about demographic diversity variable and team performance relationships: a meta-analysis. Journal of Management 37, 709–743.

Chatman, J.A., O'Reilly, C.A., 2004. Asymmetric reactions to work group sex diversity among men and women. Academy of Management Journal 47, 193–208.

Chatman, J.A., Polzer, J.T., Barsade, S.G., Neale, M.A., 1998. Being different yet feeling similar: the influence of demographic composition and organizational culture on work processes and outcomes. Administrative Science Quarterly 43, 749–780.

Cox, T.H., Blake, S., 1991. Managing cultural diversity: implications for organizational competitiveness. Academy of Management Executive 5, 45–56.

Earley, C.P., Mosakowski, E., 2000. Creating hybrid team cultures: an empirical test of transnational team functioning. Academy of Management Journal 43, 26–49.

Ely, R.J., Thomas, D.A., 2001. Cultural diversity at work: the effects of diversity perspectives on work group processes and outcomes. Administrative Science Quarterly 46, 229–273.

Gonzalez, J.A., DeNisi, A.S., 2009. Cross-level effects of demography and diversity climate on organizational attachment and firm effectiveness. Journal of Organizational Behavior 30, 21–40.

Harrison, D.A., Klein, K.J., 2007. What's the difference? Diversity constructs as separation, variety, or disparity in organizations. Academy of Management Review 32, 1199–1228.

Harrison, D.A., Price, K.H., Bell, M.P., 1998. Beyond relational demography: time and the effects of surface-and deep-level diversity on work group cohesion. Academy of Management Journal 41, 96–107.
Horwitz, S.K., Horwitz, I.B., 2007. The effects of team diversity on team outcomes: a meta-analytic review of team demography. Journal of Management 33, 987–1015.
Jackson, S.E., Joshi, A., Erhardt, N.L., 2003. Recent research on team and organizational diversity: SWOT analysis and implications. Journal of Management 29, 801–830.
Jehn, K.A., Northcraft, G.B., Neale, M.A., 1999. Why differences make a difference: a field study of diversity, conflict and performance in workgroups. Administrative Science Quarterly 44, 741–763.
Joshi, A., Roh, H., 2009. The role of context in work team diversity research: a meta-analytic review. Academy of Management Journal 52, 599–627.
Kalev, A., Dobbin, F., Kelly, E., 2006. Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. American Sociological Review 71, 589–617.
van Knippenberg, D., Schippers, M.C., 2007. Work group diversity. Annual Review of Psychology 58, 515–541.
Lau, D.C., Murnighan, J.K., 1998. Demographic diversity and faultlines: the compositional dynamics of organizational groups. Academy of Management Review 23, 325–340.
McKay, P.F., Avery, D.R., Tonidandel, S., Morris, M.A., Hernandez, M., Hebl, M.R., 2007. Racial differences in employee retention: are diversity climate perceptions the key? Personnel Psychology 60, 35–62.
Pelled, L.H., Eisenhardt, K.M., Xin, K.R., 1999. Exploring the black box: an analysis of work group diversity, conflict and performance. Administrative Science Quarterly 44, 1–28.
Ragins, B.R., Cotton, J.L., 1999. Mentor functions and outcomes: a comparison of men and women in formal and informal mentoring relationships. Journal of Applied Psychology 84, 529–550.
Richard, O.C., 2000. Racial diversity, business strategy, and firm performance: a resource-based view. Academy of Management Journal 43, 164–177.
Richard, O.C., Barnett, T., Dwyer, S., Chadwick, K., 2004. Cultural diversity in management, firm performance, and the moderating role of entrepreneurial orientation dimensions. Academy of Management Journal 47, 255–266.
Riordan, C.M., 2000. Relational demography within groups: past developments, contradictions, and new directions. In: Ferris, G.R. (Ed.), Research in Personnel and Human Resources Management, vol. 19. Emerald, Oxford, UK, pp. 131–173.
Rynes, S., Rosen, B., 1995. A field survey of factors affecting the adoption and perceived success of diversity training. Personnel Psychology 48, 247–270.
Shore, L.M., Randel, A.E., Chung, B.G., Dean, M.A., Ehrhart, K.H., Singh, G., 2011. Inclusion and diversity in work groups: a review and model for future research. Journal of Management 37, 1262–1289.
Tsui, A.S., O'Reilly, C.A., 1989. Beyond simple demographic effects: the importance of relational demography in superior-subordinate dyads. Academy of Management Journal 32, 402–423.
Tsui, A.S., Egan, T.D., O'Reilly, C.A., 1992. Being different: relational demography and organizational attachment. Administrative Science Quarterly 37, 549–579.
Watson, W.E., Kumar, K., Michaelsen, L.K., 1993. Cultural diversity's impact on interaction process and performance: comparing homogeneous and diverse task groups. Academy of Management Journal 36, 590–602.
Williams, K.Y., O'Reilly, C.A., 1998. Demography and diversity in organizations: a review of 40 years of research. In: Staw, B.M., Cummings, L.L. (Eds.), Research in Organizational Behavior. JAI Press, Greenwich, CT, pp. 77–140.

Relevant Websites

www.BLS.gov.
www.EEOC.gov.
www.DiversityInc.com.

Division of Labor: History of the Concept

Richard Sturn, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Division of labor involves the decomposition of processes of (re)production, wealth creation, and knowledge development into various tasks, each of which is performed by a separate individual or collective agent. Division of labor is related to other important concepts, such as capital accumulation, growth, specialization, dynamically increasing returns to scale, and social differentiation. Types of division of labor are distinguished by the socioeconomic mechanisms mediating underlying interdependences, notably organizations/institutions based on hierarchy or collective decision making (including the firm and family) on the one hand and spontaneous social processes (e.g., market processes) on the other.

Descriptions of the phenomenon date back to Plato and Xenophon. While all modern economists acknowledge the importance of division of labor as an elementary condition of productivity, much of mainstream economics treated it as a commonplace *external* to the theory. By contrast, Adam Smith emphasizes division of labor as the core of an *endogenous* process. Contributors to the analysis of division of labor include Charles Babbage and Karl Marx in the nineteenth and Allyn Young in the twentieth century, who provided the foundations for modern developments since the 1980s.

Foundational Core or External Commonplace?

Division of labor involves the decomposition of processes of (re)production, wealth creation, and knowledge development into separate tasks, each of which is performed by an individual or collective agent: a person, a class of persons, a region, a nation, an organization, or an institution. T. Hodgskin (1827 [1966]), J.J. Bachofen (1861), F. Engels (1884[1958]), and K. Marx were among the first writers emphasizing division of labor between the sexes. In particular, Engels draws on Bachofen's writings pertinent to the development of institutions governing behavior in the reproductive sphere (including an evolutionary sequence from primitive communism to matriarchal and finally patriarchal rule systems) in order to establish a 'materialist' theory of reproduction-related functional differentiation. Within this current of theorizing, power-related issues loom large, dealing with the way in which the division of labor is accompanied by systematic inequalities in the distribution of power and wealth. 'Fragmentation of knowledge' (an expression coined by Friedrich von Hayek) is a phenomenon closely related to division of labor. Regarding the evolution of scientific knowledge, the emergence of the disciplinary structure of modern science is a manifestation of division of labor, which is accompanied by other forms of specialization in methods and differentiation regarding levels of analysis. The shifting emphasis on issues and concepts throughout the history of ideas is sometimes regarded as a kind of *intertemporal division of labor*.

Division of labor is an old concept: in Western thought, classical antiquity produced remarkable descriptions of the phenomenon, including the explanation of some of the advantages of the division of labor within the organization of social production. In Plato's *Republic* (380 BC) division of labor is moreover discussed in its relations to accompanying phenomena such as exchange, markets, and money. In a tract entitled *The Education of Cyros*, Xenophon, a further writer of the fourth century BC, provides some additional insights, focusing the interdependence between division of labor, the development of greater cities, and concomitant demand-driven specialization of skills.

Division of labor is a complex concept: its theorization in economics establishes links between division of labor and other important concepts capturing various forms of economic dynamism, such as capital accumulation, growth, specialization, dynamically increasing returns to scale, complementarity, exchange, and indivisibility. The identity of human *individuals* as agents of division of labor is the basis for specific indivisibilities (e.g., related to the time necessary for learning specialized skills). Division of labor is moreover related to a range of broader socioeconomic issues such as inequality, heterogeneity, discrimination, alienation, power, cooperation, freedom, and social cohesion. Pertinent themes are in particular dealt with by currents of sociology and by Marxian Political Economy. The role of division of labor among the forces of social cohesion is an issue of pivotal importance in Émile Durkheim's evolutionary account of *Division of Labor in Society*: organic solidarity based on mutual dependence between individuals with different specializations is gradually superseding the traditional form of mechanical solidarity based on social homogeneity. Some of Durkheim's distinctions allow for grasping further aspects of division of labor as a complex concept: division of labor is not only intricately related to the dynamism of modern societies, but also to their endogenous crises. Referring to pathological tendencies of disintegration accompanying modern specialization processes, he distinguishes between anomic and rule-based division of labor. Moreover, he distinguishes between spontaneous and forced division of labor and between horizontal and vertical division of labor. In twentieth century sociology, the horizontal–vertical distinction reappears in the context of theories of social differentiation developed by authors such as Talcott Parsons and Niklas Luhman. Theories of social differentiation bring to the fore further dimensions of differentiation, including spatial differentiation, center periphery differentiation, and functional differentiation related to subsystems. Obviously, there exist parallels to some of the phenomena covered by economic

analysis. Moreover, pertinent currents of thought both in economics and in sociology invoke process-related properties of those phenomena and stress the importance of dynamic or evolutionary reasoning. Beyond those commonalities appearing at a prima facie level, analyzing the ways in which pertinent reasonings and theorizations in economics and sociology may be related to each other is a difficult task – reaching much beyond the scope of the present contribution. Anyway, the dynamism of economic specialization modeled along the lines of cumulative causation (which are widely regarded as a main explanation of the historically *unprecedented economic growth* of the past 250 years) somehow seem to find a parallel in sociological theories of differentiation processes (which are seen as related to the *increasing complexity* of modern society).

In political economy, different *types of division of labor* are primarily distinguished by the institutional mechanisms mediating the underlying interdependencies. There are two main classes of such mechanisms: (1) *Decentralized* and spontaneous social mechanisms, including markets as the paradigmatic case covering division of labor within the national economy as well as the international/global economy (world trade). (2) *Centralized* decision-making mechanisms, notably firms and other hierarchies. The second type of division of labor became visible and was conceptualized in the context of the development of the manufacturing processes accompanying proto-industrialization and industrialization: Adam Smith's pin factory and (in the twentieth century) F.W. Taylor's scientific management based on the analysis of workflows and their optimized recombination ('Taylorism') became prototypical models of intrafirm division of labor.

While almost all economists readily acknowledge the importance of division of labor as an elementary condition of productivity-enhancing technical change, in important strands of twentieth century economics it is confined to the role of a commonplace *external* to the theory, as Joseph Schumpeter put it. That is, the causes determining patterns of division of labor are regarded as *exogenous*. By contrast, Adam Smith's view of the subject is based on *endogenous* specialization processes, emphasizing the development of specialized skills among agents who in the beginning are very much alike.

How can we make sense of the vicissitudes of division of labor in the history of economic analysis? Here are some of the reasons for its ambivalent or even paradoxical status within the discipline of economics. First, the neglect of division of labor may have been caused by the fact that the mathematical tools supporting modern marginalist economics are specifically unsuitable for dealing with issues of specialization and division of labor. Second, the dynamic character of division of labor may be responsible for the vicissitudes and ambiguities characterizing its evolution within the history of economics when the latter was largely preoccupied with issues of static allocation theory and market equilibrium. Over and above all, the conceptual complexity of division of labor rendered it incongruous with the analytical ambitions of pure economics. In the wake of the programatic move toward *pure* economics advocated by leading protagonists of the neoclassical revolution, division of labor tended to be seen as a subject not really suitable for economic analysis in the proper sense, but rather for sociology on the one hand (regarding the social division of labor or social differentiation), or for studies in technology, engineering, and labor psychology on the other. Indeed, this view coincided with the rise of managerial analysis of firm-specific division of labor culminating in currents such as Taylorism.

Put in a nutshell, the disciplinary history can be summarized as follows: after a spectacular start as the pivotal element of socioeconomic progress in Adam Smith's Political Economy, division of labor retains a major role in the work of key figures of the nineteenth century of economics such as John Stuart Mill and Alfred Marshall, while Charles Babbage and Karl Marx significantly contributed to the further conceptual development. For much of the twentieth century, division of labor is a theme rather at the periphery than in or near the core of economics. The situation in sociology, however, was very different; in that discipline, the division of labor was a core concern throughout the twentieth century and into the present day.

Apart from a few incisive, but rather isolated articles on the subject (e.g., by Hendrik Houthakker, 1956) we find three somewhat more influential attempts to reintroduce division of labor as an important analytical concept of economics rather than an external commonplace. The first two are linked to the names of Allyn Young (1876–1929) and Nicholas Kaldor (1908–86). Both contributions are widely regarded as attempts commanding respect, but neither of them became the nucleus of a broader current of research gaining momentum as an important contribution to the understanding of economic dynamism. The last two decades of the twentieth century saw the development of the most recent of those three attempts aiming at reinvigorating the economic analysis of the division of labor: it includes models of economic geography and endogenous specialization in foreign trade proposed by Paul Krugman and others as well as the more general framework of the so-called New Classical inframarginal theory of specialization. The latter is supported by mathematical techniques suitable for considering specialization decisions (involving inframarginal comparisons between corner solutions rather than the usual comparisons at the margin) and concomitant interdependencies. To which extent those relatively new developments will change the general picture that emerged in the twentieth century ('division of labor as an external commonplace') is still an open question.

From William Petty to Adam Smith

After the forerunners in classical antiquity mentioned in the introduction, it was a key figure in the development of early modern economic thought who brought issues of division of labor to the fore in the seventeenth century: William Petty (1623–87). In his seminal *Political Arithmetick* (1671), he compares the advantages of the division of labor and specialization in various processes (textile manufacturing and sea transportation). In a tract published 10 years later, Petty (1683) touches upon the interdependence of specialization, the growth of cities and spacial concentration of certain types of activities. In the eighteenth century, writers from different backgrounds discussed various aspects of the division of labor: Bernard Mandeville (1670–1733) and Frances Hutcheson (1694–1746) became interested in socioeconomic questions in

the context of their broader concerns regarding the institutional and political implications of the emerging commercial society. While their focus was socioeconomic division of labor in a broader sense, writers such as Patrick Lindsay, Richard Campbell, and Joseph Harris dealt with division of labor in manufacturing illustrated by examples such as watchmaking (already used by Petty) and pin manufacturing, a showcase destined to make quite a career, as it was employed by the *French encyclopedia* and – most famously – by Adam Smith. Sir Josiah Tucker, Henry Martin, and Henry Maxwell made explicit the role of population size (an issue already implicit in discussions pointing to the role of bigger cities), anticipating the role of market size to be emphasized by Smith as well as related arguments by Durkheim. Turgot discussed the socioeconomic division of labor, linking it to economic phenomena such as capital accumulation and inequality. Beccaria (1771) explained advantages of the division of labor regarding the development of skill and dexterity.

As is the case with some other eighteenth century authors just mentioned (notably Turgot and Beccaria), division of labor is introduced in Smith's *Wealth of Nations* on the very first pages: for Smith, division of labor is close to the pivotal starting point of socioeconomic development. The starting point is exchange. According to Smith, exchange is related to basic properties of humans as communicating, social beings who by their nature very much depend on the assistance of others. The very "slow and gradual" development of division of labor is explained as the unintended effect of a concomitant "propensity in human nature" to "truck, barter and exchange one thing for another" (Wealth of Nations I. ii.1).

Authorities such as Edwin Cannan (1929: p. 96) suggest that Smith's analysis is remarkable for the emphasis put on division of labor, rather than for its originality or analytical depth. Joseph Schumpeter (1954: p. 187) claims that "nobody … ever thought of putting such a burden upon division of labor. With A. Smith, it is practically the only factor determining economic progress." This is clearly not meant to represent a mark of ingenuity. A careful reading shows that Smith's conceptual innovations are more important than suggested by such statements. First, it is hardly possible to overemphasize the thoroughly *dynamic* character of Smith's theory: division of labor is visualized as a process entailing cumulative causation. Second, it is portrayed as a *pervasive* phenomenon in the development of modern societies, including areas such as military defense, science and science-based technology, or law enforcement. Third, division of labor is perhaps Smith's most important case of a dynamic phenomenon emerging as the *unintended consequence* of human action. This foundational idea is closely related to Smith's view of exchange and markets as social mechanisms mediating progress based on human action, but not on the wisdom of human design. Unintended and cumulative effects do not only occur regarding market-mediated division of labor at the societal level (which is limited by the extent of the market, as Smith famously states), but also in the case of intrafirm division of labor: it is accompanied by cumulative advantages related to improvements in skill, dexterity, and judgment, and enhanced possibilities of mechanization.

Moreover, some unintended social consequences may trigger pernicious effects on the social and political life of modern societies. Those challenges are a recurrent theme in the pertinent literature: Marx related division of labor to a process of alienation accompanying capitalist dynamism, and Harry Braverman (1974) analyzed the degradation of work in the twentieth century. Undesirable social consequences of division of labor were also discussed by thinkers of Scottish Enlightenment, such as Smith's patron Lord Kames and Adam Ferguson (1767[1782]), who emphasized the tension between division of labor and the demands of republican citizenship. In Smith's view, decomposing labor processes into very small bits tends to make workers one-sided and even stupid. Worst of all, the development of simple industrial tasks that can be easily performed by children is a major problem as it hinders their education. The repetitive execution of those tasks does not make them fit for life in a modern society, including the defense of the nation.

Between Marx and Marshall: Developments in the Nineteenth Century

The conceptual history of division of labor in the nineteenth century is inextricably linked to a few names. Karl Marx (1818–83) is perhaps the author of the most comprehensive account of division of labor. In Marx's theory, division of labor is one of the three 'social forces of production' (along with 'cooperation' and 'science') whose development sets the stage for the restless productivity-enhancing change, which has become the key feature of modern economies. Marx also clearly distinguished between social division of labor and intrafirm division of labor in manufacturing: as Nathan Rosenberg emphasized, Marx studied the technological aspects of division of labor and their evolution with particular care. In Marx's theory, technological aspects of intrafirm division of labor were integrated in a more encompassing (if sketchy) analytical framework, thereby relating it to the development process of capitalist market economies at large. Before Marx, three writers stand out for having contributed to the economics of division of labor: Andrew Ure (1778–1857), Charles Babbage (1791–1871), and John Rae (1796–1872). Babbage and Ure were among the authors who were studied in great depth by Marx.

Rae (1834[1964]) emphasizes in particular the savings in the use of tools occasioned by division of labor. (Incidentally, he is also the author of the most important biography of Adam Smith in the nineteenth century.) Babbage develops a comprehensive theoretical vision of division of labor including processes of control and information in their relation to the overall industrial dynamism and he delivers a more detailed analysis of the advantages of division of labor. In particular, the mechanisms by which division of labor allows for saving of time and other kinds of costs are analyzed. They include (1) saving of time and other costs (such as waste of material) by more efficient *learning processes* triggered by the development of more specialized skill profiles; (2) the impact of specialization on dexterity; and (3) on the development and introduction of new machinery. Babbage (1832[1963]: p. 175) moreover makes a subtle point which was not entirely new, but is explained by him in a crisp and clear-cut fashion: "by dividing the work into different processes of skill and force" an employer "can purchase exactly that precise quantity of both

which is necessary for each process". If, by contrast, the whole work is performed by one worker, she must be endowed with sufficient skill to do the most difficult task and sufficient strength to execute the most laborious task. According to Babbage, the establishment of large factories can be explained by their potential for pertinent optimization of production processes. Analogous reasoning is also applied to the division of mental labor, illustrated by the production of logarithmic tables organized by Gaspard de Prony (1755–1839), who incidentally had read the chapter of the *Wealth of Nation* in which the organization of the pin factory is explained.

Andrew Ure (1835[1967]) highlighted aspects, which should become particularly important for Karl Marx: a tendency of division of labor toward deskilling of labor. Specific skills of laborers (imagine handicraft production) are substituted by machines based on scientific and technological knowledge and rational management. Skills are transferred from the workers to the machinery – with far-reaching implications for the locus of control. Along such lines of reasoning, Braverman's (1974) analysis emphasizes the distinction between social division of labor (which tends to enrich the human species *and individual agents*) and intrafirm division of labor into ever-smaller subprocesses (which is in tension with human needs and capabilities and hence tends to alienate the individual from the potential of the species).

Alfred Marshall's canonical textbook (1890: esp. chapters 8–12) provides a substantial amount of valuable reasoning regarding the economic implication of division of labor in a post-Smithian tradition. Yet this part of his work was not the one which carried the day as the Marshallian tradition in neoclassical economics of the twentieth century: in the successful Marshallian partial equilibrium supply-and-demand models with representative firms and consumers (allowing for invaluable advantages such as expedient diagrammatical teaching tools and mathematical modeling) division of labor plays little (if any) role.

The Twentieth Century: Young, Inframarginal Analysis, and the Household

Allyn Young's (1876–1929) paper (1928) is the outstanding, if somewhat isolated contribution to the analysis of specialization and division of labor in the first half of the twentieth century. Yet his work is frequently cited by the theorists who in the last decades of the twentieth century advanced formal models of specialization and division of labor including Xiaokai Yang and Sherwin Rosen (1983: p. 44), who called it "the zenith of the analysis of the connection between specialization and economic development." Young's concept of 'social increasing returns' is directly related to specialization and division of labor, foreshadowing the modern ideas of network theories related to division of labor. In Young's view, Smith's observation that division of labor is limited by the extent of the market should be understood in a more general way: on the one hand, it is also true that the extent of the market depends on the division of labor and specialization; the more specialized individuals are, the less able will they be to produce all the goods meeting their needs and wants themselves, and the more inclined will they be to buy some goods on the market. Concomitantly, the extent of the market is not only determined by the number of individuals in the market, but also by their purchasing power, which in turn reflects productivity, which again depends on division of labor. Taken together, a logic of circular and cumulative causation emerges, which is aptly summarized in Young's statement according to which division of labor is caused by division of labor.

Hendrik Houthakker's (1956) contribution is remarkable for its outline of a model capturing Smith's idea of endogenous specialization. Smith's emphasis on the extent of the market is the starting point for analyzing the role of transportation costs and – more generally – various types of transaction costs. Some contributions by Paul Krugman (e.g., 1980) on international trade belong to this strand of thought, implying a major advance toward modeling the idea of endogenous comparative advantage in trade: those models explain why trade may be beneficial even if countries are ex ante identical and shed light on trade patterns between industrialized countries.

In a number of contributions from the 1960s and 1970s, Nicholas Kaldor made use of Young's insights. In a volume entitled "Strategic Factors in Economic Development" (1967), he invokes the cumulative dynamism linking development of manufacturing to growth in labor productivity mediated by division of labor-driven increasing returns à la Young. In some programatic articles, Kaldor (e.g., 1972) criticizes the inability of mainstream economics to accommodate this kind of analysis, leading to 'irrelevance of equilibrium economics'.

In the late 1970s and 1980s, Sherwin Rosen (e.g., 1978) attempted to reinvigorate Young's 'social increasing returns', inter alia by introducing the concept of superadditivity, while James Buchanan introduced the closely related concept of 'generalized increasing returns'. A major attempt to bring in division of labor back from the cold is due to Xiakao Yang, who is the author of numerous articles with several coauthors in the 1990s and early 2000s. Yang and Ng (1993) summarize the reinvigoration of the economics of division of labor by the label *New Classical Economics* or *Smithian framework*. Its theoretical core are 'inframarginal' decisions regarding *what* activities to engage in. This is modeled by Yang and his coauthors on the basis of mathematical techniques such as (non)linear and dynamic programing and control theory allowing for corner solutions. Put another way: while traditional economics focuses *marginal* choices regarding resource allocation, i.e., choosing the quantity of resources devoted to each activity, given that the basic activity pattern is already determined, in a Smithian framework *inframarginal* choices regarding the nature and level of specialization are analyzed. As one individual's inframarginal choices determine her exchange patterns with others, all individuals' specialization decisions jointly bring about the social division of labor and determine the extent of the exchange network.

This current of research allows for a number of interesting applications, including international trade, the theory of the firm and urban economics (cf the review by Cheng and Yang, 2004). Those successes (including the demonstration that pertinent ideas by Smith, Young, and others can be developed in formal frameworks) notwithstanding, the claim that it shifted division of labor from the periphery to the core of economic analysis would be an exaggeration as yet.

A different current of contemporary research focuses division of labor in the sphere of reproduction. Stimulated by Gary Becker's explanation of the division of labor in the household as the outcome of mutually beneficial exchange between husbands and wives (driven by natural comparative advantage), various strands of theorizing stress the role of asymmetries of power, status, and wealth. Cooperative game theory widens the horizon of economic models by making explicit the role of bargaining power in intrafamily decision processes regarding labor market participation and the allocation of household work. In the sociological literature, the role of gender display models/gender ideologies is brought to the fore (see Brines, 1994; Greenstein, 2000) in order to explain sticky asymmetries, such as the fact that women still do the majority of household work, despite dramatic increases in labor force participation of married women.

See also: Alienation, History of; Inequality, Social; Knowledge Society, History of; Labor, Division of; Unintended Consequences: History of the Concept; Work, Labor: History of the Concept.

Bibliography

Babbage, C., 1832[1963]. On the Economy of Machinery and Manufactures, fourth enl. ed. Augustus M. Kelley, New York.

Bachofen, J.J., 1861. Das Mutterrecht: eine Untersuchung über die Gynaikokratie der alten Welt nach ihrer religiösen und rechtlichen Natur. Krais und Hoffmann, Stuttgart.

Beccaria, C., 1771[1958]. Elementi di economia publica. In: Romagnoli, S. (Ed.), Cesare Beccaria Opere. Sansoni, Florence.

Braverman, H., 1974. Labor and Monopoly Capital. The Degradation of Work in the Twentieth Century. Montly Review Press, New York.

Brines, J., 1994. Economic dependency, gender, and the division of labor at home. American Journal of Sociology 100, 652–688.

Cannan, E., 1929. Review of Economic Theory. P.S. King & Son, London.

Cheng, W., Yang, X., 2004. Inframarginal analysis of division of labor. A survey. Journal of Economic Behavior & Organization 55, 137–174.

Durkheim, E., 1893[2014]. The Division of Labor in Society. With a New Introduction by Steve Lukes. The Free Press, New York.

Engels, F., 1884[1958]. Origin of the family, private property and the state. In: Marx-Engels (Ed.), Selected Works, vol. II. Foreign Languages Publishing House, Moscow.

Ferguson, A., 1767[1782]. An Essay on the History of Civil Society, fifth ed. T. Cadell, London.

Greenstein, T.N., 2000. Economic dependence, gender, and the division of labor in the home. A replication and extension. Journal of Marriage and the Family 62, 322–335.

Hodgskin, T., 1827[1966]. Popular Political Economy. C. Tait, London. Reissued by Kelley, A.M., New York.

Houthakker, H., 1956. Economics and biology: specialization and speciation. Kyklos 9, 181–189.

Kaldor, N., 1967. Strategic Factors in Economic Development. State School of Industrial and Labor Relations, Cornell University, Ithaca.

Kaldor, N., 1972. The irrelevance of equilibrium economics. Economic Journal 82, 1237–1255.

Krugman, P., 1980. Scale economics, product differentiation, and the pattern of trade. American Economic Review 70, 950–1009.

Marshall, A., 1890[1920]. Principles of Economics, eigth ed. Macmillan, London.

Marx, K., 1867[1959]. Capital, vol. 1. Foreign Languages Publishing House, Moscow.

Petty, W., 1671[1963a]. Political arithmetick. In: Hull, C.H. (Ed.), Economic Writings of Sir William Petty. Reissued by Kelley, A.M., New York.

Petty, W., 1683[1963b]. Another essay on political arithmetic concerning the growth of the city of London. In: Hull, C.H. (Ed.), Economic Writings of Sir William Petty. Reissued by Kelley, A.M., New York.

Rae, J., 1834[1964]. Statement of Some New Principles on the Subject of Political Economy. Reissued by Kelley, A.M., New York.

Rosen, S., 1983. Specialization and human capital. Journal of Labor Economics 1, 43–49.

Schumpeter, J.A., 1954. History of Economic Analysis. Oxford University Press, New York–London.

Ure, A., 1835[1967]. The Philosophy of Manufactures. C. Knight, London. Reissued by Cass, London.

Yang, X., Ng, Y.-K., 1993. Specialization and Economic Organization. A New Classical Microeconomic Framework. North-Holland, Amsterdam.

Young, A., 1928. Increasing returns and economic progress. Economic Journal 38, 327–342.

Divorce, Sociology of

Ellyn A Steidl and Kelly Raley, The University of Texas-Austin, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.S. Wallerstein, volume 6, pp. 3814–3816, © 2001, Elsevier Ltd.

Abstract

High levels of divorce in the United States are both persistent and unusual from an international perspective. In recent years, divorce has become more stratified, with marital dissolution declining for the most educated while it remains high for those without a college degree. Generally, divorce is associated with declines in the well-being of adults and children, although divorce is not universally associated with poor outcomes, suggesting that social context matters. That divorce is increasingly concentrated in the most disadvantaged groups and sometimes has negative consequences for adults and children, suggests that it both reflects and further contributes to stratification in the United States.

Divorce Levels and Trends

Overall, in the United States, about half of American marriages will end through divorce or separation, although the risk of dissolution varies considerably by social context (Cherlin, 2010; Raley and Bumpass, 2003). While divorce rates increased rapidly in the 1960s and 1970s, this was merely an acceleration of longer term trends in divorce that had been occurring since before the twentieth century (Glick, 1986; Martin and Bumpass, 1989). Beginning in the 1980s, unadjusted divorce rates tipped slightly downward, and recently most researchers have argued that the probability of divorce has either continued to decline or essentially stabilized. However, the deterioration in data sources for divorce statistics, coupled with a common failure to account for compositional changes in the married population across time, may have worked to disguise an actual increase in the divorce rate in recent years (Kennedy and Ruggles, 2013).

Divorce rates in the United States have long been elevated in comparison to those of other nations. Most European countries see between 20 and 30% of their marriages dissolve within 15 years, compared to the United States average of 42% (Andersson, 2003). Lower levels of union dissolution are often observable in both Southern and Eastern European Catholic nations. Italy, Spain, Slovenia, and Poland all with fewer than 10% of marriages dissolving within 15 years (Andersson, 2003). European divorce rates have also been increasing in recent decades, so although the United States continues to have the highest among developed nations, the gap appears to be narrowing (Amato and James, 2010).

When divorce is studied cross-nationally, especially among developed nations, it is helpful to consider cohabiting unions that break up to understand the full picture of union stability. Scandinavian countries in particular are known for having low rates of marriage (and consequently low rates of divorce), but high rates of long-term cohabitation. When cohabiting unions are taken into account, some European countries (e.g., Sweden) have relationship dissolution rates that approach the rates in the United States (Andersson, 2003; Popenoe, 2005).

Correlates of Divorce

Socioeconomic Status

Early theorists (Becker, 1981; Parsons, 1949) attempted to explain rises in divorce as a consequence of women's entrance into the paid workforce. Becker (1981), for example, argued that many of the gains to marriage were realized through specialization. Others suggested that increased earnings lowered the barriers to leaving unhappy marriages (Ross and Sawhill, 1975). These early theories, often referred to collectively as the Women's Economic Independence explanation, were supported by research showing that the historical rise in divorce rates tightly corresponds to increases in women's labor force participation and earnings (Ruggles, 1997). Yet, studies examining these associations at the individual level have produced far less convincing support. Some research finds a positive association between women's own earnings and her risk of divorce (e.g., Ross and Sawhill, 1975) but some finds a curvilinear relationship, no association, or a negative one (e.g., Rogers, 2004; Greenstein, 1990; Hoffman and Duncan, 1995). Even among analyses using similar measurement approaches (e.g., the ratio of wives income to total family income), some find support for Economic Independence theory while others do not (Sayer and Bianchi, 2000).

Inconsistencies in findings might stem from the control variables included in the statistical models (Sayer and Bianchi, 2000), whether the analysis distinguishes work hours from wage rates (Teachman, 2010), or whether the modifying influences of context are considered (Schoen et al., 2006; Cooke and Baxter, 2010). This suggests that women's participation in the paid labor force is connected to marital stability through multiple pathways some of which help to stabilize marriage (e.g., by increasing family resources) and others that can undermine it (e.g., stress and long work hours creating more work–family conflict). The gender role attitudes of the couple, their marital happiness, as well as the broader institutional context can shape which of these pathways dominates the association. Perhaps the most important pathway runs from marital stability to female labor supply, as a recent review suggests (Ozcan and Breen, 2012; see also Oppenheimer, 1997; Rogers, 1999).

In contrast to earnings, women's educational attainment is consistently negatively associated with the risk of divorce, at least in the United States. More specifically, having a college degree is associated with a lower risk of divorce, but women with some college have a higher risk of marital instability than those who have just a high school degree, a finding referred to as the 'Glick effect' (Glick, 1957). Over time, educational differentials in divorce have grown as divorce rates to college educated women declined after 1980, while they remained steady or continued to climb for those with lower levels of education (Martin, 2006; Raley and Bumpass, 2003). Interestingly, education is not associated with risk of divorce in all countries and the association may even be negative in some contexts, suggesting the importance of institutional contexts for socioeconomic disparities in family stability (Härkönen and Dronkers, 2006; Cooke and Baxter, 2010).

Premarital Cohabitation

The slight decline and subsequent stabilization of the divorce rate have been accompanied by rapid increases in cohabitation rates. Conventional wisdom often held that cohabitation could function as a way to guard against future divorce by testing couples' compatibility and the relationship's durability through shared living (Rhoades et al., 2009). In contrast, multiple studies over time consistently found that premarital cohabitation was associated with lower marital quality and an increased divorce risk (Axinn and Thornton, 1992; Kline et al., 2004; Stanley et al., 2010; Jose et al., 2010). Researchers have proposed two main explanations for the association between cohabitation and divorce. First, those who select cohabitation over direct marriage may possess certain characteristics that place them at a greater risk of divorce, like poor relational skills or nontraditional views of divorce and marriage (Phillips and Sweeney, 2005). Research also supports the idea that there is something endogenous to the experience of cohabitation that can both foster lower commitment between partners, and can prolong weaker relationships that would not have otherwise become marriages (Axinn and Barber, 1997; Stanley et al., 2006). But there is evidence that the relationship between premarital cohabitation and divorce may be weakening if not absent in recent marriage cohorts (Reinhold, 2010; Manning and Cohen, 2012). It may be the case that the diffusion of cohabitation across populations has worked to mitigate the selection factors that tied cohabitation to divorce.

Race–Ethnic Differences

Blacks and US-born Mexicans have higher rates of divorce than non-Hispanic Whites (Phillips and Sweeney, 2005). Age at marriage, educational attainment, premarital fertility, as well as premarital cohabitation contribute to these differences (Phillips and Sweeney, 2006). Couples marrying in their teen ages and early 20s face a higher risk of dissolution (Raley and Bumpass, 2003) and this, along with lower levels of education, accounts for some of the higher risk of marital dissolution among US-born Mexican women (Phillips and Sweeney, 2006). Blacks marry at later ages than Whites or Hispanics, so age at marriage cannot explain higher levels of divorce among African-Americans. Nonetheless, lower levels of educational attainment combined with more premarital fertility and premarital cohabitation account for some, but not all, of the higher risks of divorce for blacks (Phillips and Sweeney, 2006).

Additional Correlates

Divorce in the family of origin has also been linked to children's risk of divorce in their own marriages (Amato and DeBoer, 2001; Amato and Cheadle, 2005; Wolfinger, 2005). Children may observe and internalize ineffective patterns of communication and conflict resolution between parents and subsequently carry poor relationship skills into their future marriages (Caspi and Elder, 1988; Amato, 1996). Alternatively, when children experience a parental divorce their overall commitment to marriage may weaken by seeing the fragility of the marital contract (Amato and DeBoer, 2001). Some contend that the intergenerational transmission of divorce has weakened over time, but this trend remains contested (Li and Wu, 2008; Wolfinger, 2011).

Certain orienting beliefs and values may also work to either protect or undermine partners' marital success. Gender ideology can provide couples with core assumptions regarding roles in the marriage, and usually has implications for the balance of power between spouses. The direct effect of holding egalitarian vs more traditional attitudes toward marriage is rarely studied in isolation, but there is some evidence that women with egalitarian attitudes are more likely to divorce than women with more traditional views of household and family life (Kaufman, 2000). It is more likely though, that gender role beliefs influence marital stability by moderating how couples negotiate different decisions, such as the division of paid and unpaid labor (Cooke, 2006). Ultimately what may be important is how 'her' and 'his' gender role attitudes fit together (or fail to do so); the combination of an egalitarian wife with a more traditionalist husband is associated with the highest risk of marital instability (Clarkwest, 2010). Religious beliefs may work to influence couples' beliefs about marriage, and can help embed couples in supportive social networks of their religious communities (Call and Heaton, 1997; Lehrer, 2004). However, identifying with a religious belief does not protect couples uniformly. Couples that hold dissimilar religious beliefs are at an increased risk to experience a divorce (Lichter and Carmalt, 2009; Vaaler et al., 2009).

Consequences of Divorce

The dissolution of a marriage has implications for the well-being of the spouses that are involved, as well as any dependents they may have. The study of the consequences of divorce has commonly considered socioeconomic, health, and emotional–behavioral outcomes of men, women, and children. Generally, sharing a household is more economical than living separately and couples who divorce lose this advantage. Because children typically live with their mothers following divorce and women usually have lower incomes than the men they are married to; women and children often experience steeper declines in economic well-being than men. Past research has estimated that wives experience a 27% decrease in their standard of living after divorce, while husbands may

experience on average a 10% increase (Peterson, 1996). These estimates are not nearly as drastic as those published in the 1980s (Weitzman, 1985) and it is possible that declines in standard of living are due to factors other than divorce itself (Smock et al., 1999). Moreover, as women's investments in paid employment have increased, the negative economic consequence of divorce for women and children may have declined (McKeever and Wolfinger, 2001) but it has not disappeared (Osborne et al., 2012). Altogether, the evidence suggests that divorce results in declines in economic well-being for women and children. It is associated with other declines in well-being as well.

Divorce is associated with declines in health, although the patterns vary by gender and health outcome. Divorce is associated with short-term declines in physical health for men, but not women (Williams and Umberson, 2004; Hewitt and Turell, 2011). Women, however, experience declines in mental health, especially when they have young children (Williams and Dunne-Bryant, 2006). The longer term health consequences of divorce, however, are less well established, at least in part because it is difficult to separate the effects of divorce per se from factors that cause divorce. Since the rise in divorces in the past half century a growing number and proportion of children have experienced a parental divorce during their lifetime. Children with divorced parents score lower on a number of well-being measures when compared to children with continuously married parents. For example, children whose parents divorce are less likely to graduate high school (Pong and Ju, 2000) and more likely to have a teen birth (Wu, 1996). Some of this is likely a consequence of the poorer economic resources of children living with single parents, but some may also be due to the stress and disruption caused by changes in household structure (McLanahan and Sandefur, 1994; Cavanagh and Huston, 2006). Some of the poorer outcomes of children whose parents divorce are due to factors, like maternal mental health, that select families into divorce rather than to the consequences of family instability (Fomby and Cherlin, 2007).

Despite substantial research on this topic, there is continued debate over whether divorce is better understood as a discrete crisis or a chronic strain for the individuals involved and the consequences of divorce for children's well-being are still debated. This is partly because of the complexity of the question and the fact that aspects of the social context matter. For example, postdivorce adjustments can vary by whether the former marriage was high conflict or low conflict (Amato, 2010). Nonetheless, analyses using state-of-the-art analytical approaches often find parental divorce associated with at least some poorer outcomes for children in some situations, suggesting that divorce does have harmful consequences for some children.

Conclusion

Although divorce rates have been roughly stable over recent decades, some of this stability has been achieved through a later age at marriage and a higher proportions of women, especially economically disadvantaged women, never marrying (Goldstein and Kenney, 2001). Even so, marital instability today is much more common among those with the fewest economic resources. Although the harmful consequences of divorce for the well-being of adults and children may not be as great as they once were, the concentration of family instability in populations with the fewest economic resources may contribute to increases in social and economic inequality (McLanahan, 2004).

The growing rates of divorce in European countries also suggest there is variation in how different factors associated with divorce may be more or less predictive when operating in different contexts. In a comparison of American and European divorce research, Amato and James (2010) note that European findings typically support the relationships between risk factors and divorce that have been identified in the United States. When certain predictors, like high levels of education fail to be as consistently protective in European countries, it may reflect unequal levels of marriage deinstitutionalization, government welfare policies, and heterogeneity in societal stratification across countries (Härkönen and Dronkers, 2006).

See also: Cohabitation: United States; Ethnicity, Sociology of; Family Policy; Family Theory: Competing Perspectives in Social Demography; Family Theory: Economics of Marriage and Divorce; Family and Kinship, History of; Family as Institution; Family: The Anthropology of the Concept and Its History; Marriage and the Dual-Career Family; Social Stratification.

Bibliography

Amato, P.R., 1996. Explaining the intergenerational transmission of divorce. Journal of Marriage and Family 583, 628–640.

Amato, P.R., 2000. The consequences of divorce for adults and children. Journal of Marriage and the Family 62, 1269–1287.

Amato, P.R., 2010. Research on divorce: continuing trends and new developments. Journal of Marriage and Family 72, 650–666.

Amato, P.R., DeBoer, D.D., 2001. The transmission of marital instability across generations: relationship skills or commitment to marriage? Journal of Marriage and Family 634, 1038–1051.

Amato, P.R., Cheadle, J., 2005. The long reach of divorce: tracking marital dissolution and child well-being across three generations. Journal of Marriage and Family 67, 191–206.

Amato, P.R., James, S., 2010. Divorce in Europe and the United States: commonalities and differences across nations. Family Science 11, 2–13.

Andersson, G., 2003. Dissolution of Unions in Europe, a Comparative Overview. MPIDR Working Paper WP 2003-004 (Online) Available from: http://www.demogr.mpg.de/papers/working/wp-2003-004.pdf (accessed 28.08.13.).

Axinn, W.G., Thornton, A., 1992. The relationship between cohabitation and divorce, selectivity or causal influence? Demography 293, 357–374.

Axinn, W.G., Barber, J.S., 1997. Living arrangements and family formation attitudes in early adulthood. Journal of Marriage and the Family 593, 595–611.

Becker, G.S., 1981. A Treatise on the Family. Harvard University Press, Cambridge.

Brown, S.L., Lin, I.-F., 2013. The gray divorce revolution: rising divorce among middle-aged and older adults, 1990–2010. The Journals of Gerontology 676, 731–741.

Call, V.R.A., Heaton, T.B., 1997. Religious influence on marital stability. Journal for the Scientific Study of Religion 36, 382–392.

Caspi, A., Elder, G.H., 1988. Emergent family patterns: the intergenerational construction of problem behavior and relationships. In: Hinde, R.A., Stevenson-Hinde, J. (Eds.), Relationships within Families. Oxford University Press, New York, pp. 218–240.

Cavanagh, S.E., Huston, A.C., 2006. Family instability and children's early problem behavior. Social Forces 85, 551–581.

Cherlin, A.J., 2009. The Marriage-Go-Round, the State of Marriage and the Family in America Today. Alfred A. Knopf, New York.

Cherlin, A.J., 2010. Demographic trends in the United States: a review of research in the 2000s. Journal of Marriage and Family 723, 403–419.
Clarkwest, A., 2010. Her value orientations, his value orientations, and marital dissolution risk. 'Population Association of America 2010 Annual Meeting'. Held 15–17 April 2010 in Dallas, Texas.
Cooke, L.P., 2006. "Doing" gender in context: household bargaining and risk of divorce in Germany and the United States. American Journal of Sociology 112, 442–472.
Cooke, L.P., Baxter, J., 2010. "Families" in international context: comparing institutional effects across Western societies. Journal of Marriage and Family 72, 516–536.
Fomby, P., Cherlin, A., 2007. Family instability and child well-being. American Sociological Review 72, 181–204.
Glick, P.A., 1957. American Families. John Wiley and Sons, New York.
Glick, P.C., 1986. Recent changes in divorce and remarriage. Journal of Marriage and Family 484, 737–747.
Goldstein, J.R., Kenney, C.T., 2001. Marriage delayed or marriage forgone? New cohort forecasts of first marriage for U.S. women. American Sociological Review 664, 506–519.
Greenstein, T.N., 1990. Marital disruption and the employment of married women. Journal of Marriage and Family 52, 657–676.
Härkönen, J., Dronkers, J., 2006. Stability and change in the educational gradient of divorce. A comparison of seventeen countries. European Sociological Review 22, 501–518.
Hewitt, B., Turell, G., 2011. Short-term functional health and well-being after marital separation: does initiator status make a difference? American Journal of Epidemiology 173, 1308–1318.
Hoffman, S.D., Duncan, G.J., 1995. The effect of incomes, wages, and AFDC benefits on marital disruption. The Journal of Human Resources 301, 19–41.
Jose, A., O'Leary, D.K., Moyer, A., 2010. Does premarital cohabitation predict subsequent marital stability and marital quality? A meta-analysis. Journal of Marriage and Family 721, 105–116.
Kaufman, G., 2000. Do gender role attitudes matter? Family formation and dissolution among traditional and egalitarian men and women. Journal of Family Issues 211, 128–144.
Kennedy, S., Ruggles, S., 2013. Breaking up is Hard to Count: The Rise of Divorce and Cohabitation Instability in the United States. Working Paper No. 2013-01 (Online) Available from: http://link.springer.com/journal/13524/onlineFirst/page/1 (accessed 25.09.13.).
Kline, G.H., Stanley, S.M., Howard, M.J., Olmos-Gallo, P.A., St Peters, M., Whitton, S.W., Prado, L.M., 2004. Timing is everything: pre-engagement cohabitation and increased risk for poor marital outcomes. Journal of Family Psychology 182, 311–318.
Lehrer, E.L., 2004. Religion as a determinant of economic and demographic behavior in the United States. Population and Development Review 30, 707–726.
Li, J.A., Wu, L., 2008. No trend in the intergenerational transmission of divorce. Demography 454, 875–883.
Lichter, D.T., Carmalt, J.H., 2009. Religion and marital quality among low-income couples. Social Science Research 38, 168–187.
Manning, W.D., Cohen, J.A., 2012. Premarital cohabitation and marital dissolution, an examination of recent marriages. Journal of Marriage and Family 742, 377–387.
Martin, S.P., 2006. Trends in marital dissolution by women's education in the United States. Demographic Research 15, 537–560.
Martin, T.C., Bumpass, L.L., 1989. Recent trends in marital disruption. Demography 261, 37–51.
McKeever, M., Wolfinger, N.H., 2001. Reexamining the economic costs of marital disruption for women. Social Science Quarterly 82, 202–217.
McLanahan, S., 2004. Diverging destinies: how children are faring under the second demographic transition. Demography 414, 607–627.
McLanahan, S., Sandefur, G., 1994. Growing up with a Single Parent: What Hurts, What Helps. Harvard University Press, Cambridge.
Oppenheimer, V.K., 1997. Women's employment and the gain to marriage: the specialization and trading model. Annual Review of Sociology 23, 431–453.
Osborne, C., Berger, L.M., Magnuson, K., 2012. Family structure transitions and changes in maternal resources and well-being. Demography 49, 23–47.
Ozcan, B., Breen, R., 2012. Marital instability and female labor supply. Annual Review of Sociology 381, 463–481.
Parsons, T., 1949. The social structure of the family. In: Anshen, R. (Ed.), The Family, Its Function and Destiny. Harper Brothers, New York, pp. 173–201.
Peterson, R.R., 1996. A re-evaluation of the economic consequences of divorce. American Sociological Review 61, 528–536.
Phillips, J.A., Sweeney, M.M., 2005. Premarital cohabitation and the risk of marital disruption among white, black, and Mexican American women. Journal of Marriage and Family 67, 296–314.
Phillips, J.A., Sweeney, M.M., 2006. Can differential exposure to risk factors explain recent racial and ethnic variation in marital disruption? Social Science Research 35, 409–434.
Pong, S.L., Ju, D.B., 2000. The effects of change in family structure and income on dropping out of middle and high school. Journal of Family Issues 21, 147–169.
Popenoe, D., 2005. Marriage and Family, What Does the Scandinavian Experience Tell Us? The State of Our Unions, the Social Health of Marriage in America. The National Marriage Project.
Raley, R.K., Bumpass, L., 2003. The topography of the divorce plateau: levels and trends in union stability in the United States after 1980. Demographic Research 8, 245–260.
Reinhold, S., 2010. Reassessing the link between premarital cohabitation and marital instability. Demography 473, 719–733.
Rhoades, G.K., Stanley, S.M., Markman, H.J., 2009. Couples' reasons for cohabitation: associations with individual well-being and relationship quality. Journal of Family Issues 30, 233–258.
Rogers, S.J., 1999. Wives' income and marital quality: are there reciprocal effects? Journal of Marriage and Family 61, 123–132.
Rogers, S.J., 2004. Dollars, dependency, and Divorce: four perspectives on the role of wives' income. Journal of Marriage and Family 661, 59–74.
Ross, H.L., Sawhill, I.V., 1975. Time of Transition: The Growth of Families Headed by Women. Urban Institute, Washington, DC.
Ruggles, S., 1997. The rise of divorce and separation in the United States, 1880-1990. Demography 34, 455–466.
Sayer, L.C., Bianchi, S., 2000. Women's economic independence and the probability of divorce: a review and reexamination. Journal of Family Issues 217, 906–943.
Schoen, R., Rogers, S.J., Amato, P.R., 2006. Wives' employment and spouses' marital happiness: assessing the direction of influence using longitudinal couple data. Journal of Family Issues 27, 506–528.
Smock, P.J., Manning, W.D., Gupta, S., 1999. The effect of marriage and divorce on women's economic well-being. American Sociological Review 64, 794–812.
Stanley, S.M., Rhoades, G.K., Amato, P.R., Markman, H.J., Johnson, C.A., 2010. The timing of cohabitation and engagement: impact on first and second marriages. Journal of Marriage and Family 72, 906–918.
Stanley, S.M., Rhoades, G.K., Markman, H.J., 2006. Sliding versus deciding: inertia and the premarital cohabitation effect. Family Relations 554, 499–509.
Teachman, J., 2010. Wives' economic resources and risk of divorce. Journal of Family Issues 31, 1305–1323.
Vaaler, M.L., Ellison, C.G., Powers, D.A., 2009. Religious influences on the risk of marital dissolution. Journal of Marriage and Family 714, 917–934.
Weitzman, L.J., 1985. The Divorce Revolution: The Unexpected Social and Economic Consequences for Women and Children in America. The Free Press, New York.
Weitzman, L.J., 1996. The economic consequences of divorce are still unequal: comment on Peterson. American Sociological Review 61, 537–538.
Williams, K., Dunne-Bryant, A., 2006. Divorce and adult psychological well-being: clarifying the role of gender and child age. Journal of Marriage and Family 68, 1178–1196.
Williams, K., Umberson, D., 2004. Marital status, marital transitions, and health: a gendered life course perspective. Journal of Health and Social Behavior 45, 81–98.
Wolfinger, N., 2005. Understanding the Divorce Cycle, the Children of Divorce in Their Own Marriages. Cambridge University Press, New York.
Wolfinger, N., 2011. More evidence for trends in the intergenerational transmission of divorce: a completed cohort approach using data from the general social survey. Demography 48, 581–592.
Wu, L.L., 1996. Effects of family instability, income, and income instability on the risk of a premarital birth. American Sociological Review 61, 386–406.

Doctoral Education

Lynn McAlpine, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article outlines key ideas and issues in the complex field of doctoral education as represented in the English-speaking literature. After referencing its brief history, range of fields in which it is researched, and the breadth of research foci, attention is given to one field – learning/development/pedagogy. Explored are epistemological approaches, changes occurring in focus and emphasis, research trends and issues, ending with methodological directions and issues. A key thread is the difficulty of understanding doctoral education since, like many other educational undertakings, it is very much influenced by context.

Introduction

A starting point in attempting to depict doctoral education internationally is acknowledging that any representation will leave out much and will represent a viewpoint from a particular space in time, geographical location and life journey. In this instance, my viewpoint is colored by my age and gender: 60s female; my ethnicity and nationality: white Canadian; my scholarly interests: researcher-doctoral-academic learning and development; my working environment: the UK, North America, and Europe. In this article, I outline key ideas and issues in the complex field of doctoral education as represented in the English-speaking literature. Initially, I reference its brief history, range of fields in which it is researched, and the breadth of research foci. Then I explore one field, doctoral learning/development/pedagogy, since it is the area in which I principally locate myself. I elaborate epistemological approaches, changes occurring in focus and emphasis, research trends and issues, ending with methodological directions and issues. A key thread that runs through this essay is the difficulty of understanding doctoral education since, like many other educational undertakings, it is very much influenced by context: broadly social, cultural, and economic influences, and more specifically national histories of doctoral education as well as national, provincial/state, and institutional policies. (I refer throughout to doctoral education rather then PhD education to be inclusive of the growth of other kinds of degrees, for example, EdD and DPsych; however, the focus of my discussion is largely the PhD since it is the most common and researched degree.)

Current Knowledge

While the origins of the doctorate are in the 1800s – first in Germany and then the USA – doctoral education as a field of inquiry is relatively new, in fact, was little examined until the 1980s. (Examples of still cited publications in the 1990s include Cullen et al. (1994) in Australia and Delamont et al. (1998) in the UK.) During the past quarter century, research in this area has grown enormously in visibility, as regards policy, institutional visibility, and research. In relation to policy, initiatives such as the European Bologna Agreement designed to enhance academic mobility have focused national attention in European countries on the structure and length of programs. In countries such as the UK and Australia, attention is directed at doctoral education due to national quality assurance agencies which review institutions on a regular basis as to their incorporation of doctoral Codes of Practice. And in places without national policy or national accountability requirements, for instance the USA, there have been initiatives by nongovernmental agencies such as the Carnegie Initiative on the Doctorate (Golde and Walker, 2006) to engage departments in evaluating and changing their doctoral practices. More recently in a range of countries, national funding councils have generated funding policies for students that have led to the creation of innovative structures (e.g., called variously doctoral training centers/networks, doctoral schools) which may, for example, highlight interdisciplinarity or industry linkages. For an interesting analysis of international changes in doctoral education, see Nerad and Trzyna (2008).

As a way of situating the growing and broad English-speaking literature on doctoral education, a brief search of abstracts (doctoral and PhD education) in the USA and the UK showed the number of publications at least tripled from 1990–2000 to 2000–12 and included contributions from non-English-speaking countries such as (in alphabetical order) China, France, Japan Korea, the Netherlands, the Nordic/Baltic countries, Pakistan, and Spain. As interesting perhaps is the range of fields/journals in which these papers were published, demonstrating that interest is widespread: art and design, anatomy, counseling, dentistry, ecology, education, engineering, finance and accounting, geography, medicine, nursing, pharmacology, psychology, and theology. The focus of these studies was also diverse, for instance: assessment; evaluation and accountability; career preparation, curriculum and programs; diversity; international students; internationalization; kinds of doctorates (practice- and industry-based, professional, and by publication), management; the nature and purpose of the doctorate; policy development; research and other training; and supervision. As can be seen in this range of foci, studies often explore individual experience; institutional policies and practices; and less frequently national policies, and social and economic contexts; and international competitiveness.

One of the difficulties of this area of inquiry is that differences are rarely noted, though there are sometimes quite

distinct contrasts across national boundaries (as well as across and within institutions) as regards the structure, form, and terminology related to doctoral education; as well as the career structures and potential job markets that exist e.g. in North America, dissertation refers to doctoral education; in the UK, to Master's education.

Intellectual Background: Prevailing and Emerging Schools of Thought

As is apparent from the broad number of fields in which doctoral education is researched, studies draw on a range of epistemologies and methodologies and are published in different journals. Unfortunately, these different scholarly networks rarely meet despite their common interest in doctoral education. In other words, studies within each network tend to take up only the issues pertinent to their communities. One field of inquiry is academic communication with a hermeneutic approach more prominent (understanding writing as a process, multiple academic genres to learn) (e.g., Kamler and Thomson, 2006). Another area of research is the sciences using largely empirical pragmatic studies documenting issues related to producing researchers/scientists (lack of career opportunities for graduates, problems in learning research practices); these studies tend to be published in science journals. The final grouping is learning/development/pedagogy using a range of methods and epistemologies in higher education journals (supervision, assessment, and accountability).

Given this range, I will in the remainder of this review focus on only the last area, learning/development/pedagogy. In this literature, there are differences in the prevailing conceptual frameworks used. For instance, in the USA, a socialization framework underpins many studies though not always fully elaborated. This framework, originally developed by Tinto for undergraduate education, conceptualizes doctoral education using a three-stage model in relation to institutional doctoral persistence (Weidman et al., 2001). From this perspective, individuals are inculcated into the normative knowledge and practices of the field. In contrast, in the UK and to some extent Australia, a more critical post-structuralist stance is prevalent drawing often on Bourdieu's body of anthropological–sociological work (e.g., Bourdieu, 1977). These studies often present individuals as strongly influenced by the structural elements of academic work, often in negative ways. What both these approaches share in common is a privileging of structure over agency. While a structural or systemic perspective is necessary to understand the influence of factors beyond individuals' control, personal agency needs to be incorporated for a realistic representation of doctoral experience, that is, that individuals while experiencing difficulties can often navigate different structural elements to advance their intentions. Recently, such work has begun to develop, often based on Archer's (2000) philosophical work (e.g., Clegg, 2008). As well, there is an emerging interest in cultural–historical activity theory (e.g., Engestrom, 2001) of which later representations incorporate a focus on relational agency (e.g., Bowen and Roth, 2007). These shifts to include attention to individual agency echo, I believe, those occurring in other social sciences fields (King, 2010).

Changes in Focus or Emphasis Over Time: The Influence of Policy

Changes in research focus reflect changing policy emphases. For instance, national initiatives to attract more international students and demonstrate greater student diversity have led to more research on the experiences of these individuals. Calls to better prepare doctoral students for nonacademic careers have lead to more attention on the design and evaluation of a range of training opportunities beyond research training, for example, communication skills and teamwork. The shift in funding council policies related to student funding, for instance, in the UK and Australia, has led to more study of doctoral programs that integrate links to business and industry, or inter/transdisciplinary work (often within doctoral centers referred to earlier).

Current and Emerging Directions and Issues in Research

The field lacks a comprehensive and critical literature review. The only extant review (Leonard et al., 2006) examined 120 UK publications; the distribution of foci was about a third each of pedagogy (the majority supervision); peer support; and outcomes (e.g., completion times and rates, and employment patterns) with the remainder being the viva (oral defence in North America) and other forms of assessment. The gaps noted that are still true today internationally include:

- a lack of studies outside education and the social sciences more generally;
- little attention to mode of study (e.g., full time or part time, face-to-face, blended, or distance);
- few systematic comparisons across institutions;
- transition from higher education to work; this is still a gap and needs attention given the increasing scarcity of full-time permanent academic positions and the increase in post-PhD researcher positions (often contingent and short term).

Gaps noted that are less the case today include:

- details beyond gender were rarely considered; less true today given policy attention to attracting increasingly diverse students;
- the lack of discernible theoretical frameworks; I concur as regards the prevailing ones though newer approaches are often well argued perhaps because they are new;
- a very little research done on the students' perspective: less true today internationally, for instance, a number of recent studies have examined the role of student networks;
- a lack of quantitative studies; this was not the case outside the UK at that time and is less the case today in the UK though many of these studies lack a conceptual framework.

Other areas for inquiry that I believe deserve attention include: career preparation and transition, ethics education which to date focuses largely on ethical approval rather than ethical practices, studies of the new doctoral training structures, and international comparative studies given social and economic drivers internationally. Two broader issues bear attention: the integration of studies published in

languages other than English, and a critical review of the literature.

Methodological Issues and Problems

If one looks across the field, the majority of studies whether qualitative or quantitative, use a limited number of methods, mostly one-time surveys, and one-time interviews. Rarely does one see longitudinal studies, yet such studies, both qualitative and quantitative, could provide needed insights into transitions from the doctorate to other work, or what leads students to leave the doctorate. Further, nontraditional methods would offer other insights, for example, ethnography is rarely used yet captures interactions and practices that other methods do not.

Secondly, as noted earlier, doctoral education has been studied from a range of viewpoints (e.g., individual, institutional, national, and international) and different perspectives (e.g., learning, policy, finance, history, philosophy, and organizational behavior). This breadth demonstrates the complexity of the phenomenon. However, this complexity brings with it difficulties; integrating the different viewpoints would provide a more robust understanding of doctoral education, yet this is extremely difficult.

Conclusions

Doctoral education is a growing and relatively robust area of inquiry, but as in most fields of research there are substantive issues, including methodological challenges. Moreover, my view is that we need to rethink the scope of what we, as researchers, consider the boundaries of our interest. I have already mentioned the need to address post-PhD transitions. More important, I believe, is a necessary reframing given the increasing requirement for completion within 3–4 years in contrast with 5–8 years previously. This legislated change does not take account of the actual time that it takes to develop the range of abilities and qualifications expected of an independent researcher. Thus, aspects of doctoral learning and development have been moved to the post-PhD period with a subsequent increase in the number of post-PhD researcher positions in which individuals attempt to develop the portfolios they need to be competitive if they wish an academic career. It is for this reason that I suggest we need to reframe the field of inquiry from doctoral to researcher education and include those post-PhD in our studies.

Bibliography

Archer, M., 2000. Being Human: The Problem of Agency. Cambridge University Press, Cambridge.

Bourdieu, P., 1977. Outline of a Theory of Practice. Cambridge University Press, Cambridge.

Bowen, M., Roth, W.-M., 2007. The practice of field ecology: insights for science education. Research in Science Education 37, 171–187.

Clegg, S., 2008. Academic identities under threat? British Educational Research Journal 3 (3), 329–345.

Cullen, D., Pearson, M., Saha, L., Spear, R., 1994. Establishing Effective PhD Supervision. E&I, DEET. AGPS, Canberra.

Delamont, S., Atkinson, P., Parry, O., 1998. Supervising the PhD: A Guide to Success. Open University Press, Milton Keynes, UK.

Engestrom, Y., 2001. Expansive learning at work: toward an activity theoretical reconceptualisation. Journal of Education and Work 14, 133–156.

Golde, C., Walker, G., 2006. Envisioning the Future of Doctoral Education: Preparing Stewards of the Discipline. Jossey-Bass, San Francisco.

Leonard, D., Metcalfe, J., Becker, R., Evans, J., 2006. Review of Literature on the Impact of Working Context and Support on the Postgraduate Research student learning experience. Higher Education Academy, London, UK.

Kamler, B., Thomson, P., 2006. Helping Doctoral Students Write: Pedagogies for Supervision. Routledge, London.

King, A., 2010. The odd couple: Margaret Archer, Anthony Giddens and British social theory. The British Journal of Sociology 61 (s1), 253–260.

Nerad, M., Trzyna, T., 2008. Conclusion. In: Nerad, M., Heggelund, M. (Eds.), Toward a Global PhD? Forces and Forms in Doctoral Education Worldwide. University of Washington Press, Seattle, WA, pp. 300–312.

Weidman, J., Twale, D., Stein, E., 2001. Socialization of Graduate and Professional Students in Higher Education: A Perilous Passage? Jossey-Bass, San Francisco.

Relevant Websites

http://www.aera.net/SIG168/DoctoralEducationacrosstheDisciplinesSIG168/tabid/12275/Default.aspx – AERA SIG: a US organization with an annual conference.

Idern: a list serv – contact: mrconnolly@wisc.edu.

http://www.capa.edu.au/ – QPR: an Australian organization with an annual conference.

http://www.earli.org/special_interest_groups/24. researcher education and careers-EARLI SIG: an EU organization with a bi-annual conference.

http://www.vitae.ac.uk/ – Vitae: a UK-based resource for doctoral students, supervisors and researchers.

Doctor–Patient Interaction in the West: Psychosocial Aspects

Joshua D Meadors and Carolyn B Murray, University of California, Riverside, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Weinman, volume 6, pp. 3816–3821, © 2001, Elsevier Ltd.

Abstract

The present article draws from research literature to provide an overview of psychological and social aspects of doctor–patient interactions. First, doctor–patient communication itself is reviewed in terms of various approaches, processes, and methods of measurement. Next, the article discusses patient and physician characteristics and other antecedents that influence the doctor–patient relationship (e.g., gender and ethnicity). Outcomes related to the quality of doctor–patient communication are then discussed, with a focus on patient satisfaction and adherence to medical advice. The article concludes with an examination of interventions regarding doctor–patient communication, followed by a brief summary.

Introduction

The quality of health care a given patient receives largely depends on the quality of the interaction between the patient and the health care provider (e.g., physician). Doctor–patient communication (also called patient–provider communication, patient–practitioner communication, and other synonymous terms) generally refers to various aspects of the interaction between the care provider and the patient, including but not limited to the provider's bedside manner, either party's use of verbal and nonverbal communication, and the exchange and comprehension of medical information and advice. Communication plays a central role in these medical interactions, and psychological studies have consistently shown how this property of the doctor–patient relationship is related to medical outcomes. Likewise, common health care issues (e.g., patient adherence) are strongly influenced by the nature of the communication between the physician and patient.

The present article provides a brief review of the doctor–patient interaction and its psychosocial aspects. The doctor–patient interaction is discussed here in three sections: (1) a description of the doctor–patient interaction and related processes, (2) antecedents or other characteristics that influence the doctor–patient interaction, and (3) consequences or outcomes determined by the nature of this interaction. The article concludes with a discussion of interventions followed by a brief summary regarding research on doctor–patient interactions.

The Doctor–Patient Interaction

Previously, medical practice considered the patient in a primarily biomedical, evidence-based context, but suggestions for clinical methods that incorporate the patient's personal perspective gave rise to what is currently considered the patient-centered approach, which focuses on the entirety of the patient and how biological, psychological, and social factors are all related to what is considered physical health. Patient-centered communication has been distinguished from terms like 'patient-centered care' and 'patient-centeredness' (see Epstein et al., 2005), but the present article focuses on the general aspects rather than distinctions. In this context, patient-centered care typically involves features like effective verbal and nonverbal communication, expressive empathy, successful transmission of information, and encouragement of the patient's partnership and participation. Indeed, the patient-centered approach seems to result in more effective interactions and more positive medical outcomes than the biomedical or doctor-centered alternative (e.g., Roter and Hall, 1992). However, a more integrative approach has been to treat the doctor–patient relationship as an entity in and of itself, taking into account both the provider and the patient's attitudes, beliefs, values, and cultural backgrounds, as well as how these characteristics interact dynamically to produce a variety of medical outcomes. Typically, doctor–patient interactions are described according to three distinct approaches: doctor centered, patient centered, and relationship centered.

The doctor-centered approach is the older, more traditional approach in medicine. The physician tends to dominate the interaction with a focus on asking 'yes–no' medical questions generally related to the physiological pathologies of the patient. However, the doctor-centered approach tends to lack the patient's perspective, a sense of partnership, and an interpretation of psychological and social factors, which has led to the more modern patient-centered approach to medical practice. The two primary components of patient-centered care are that (1) the entirety of the patient should always be considered, including any relevant psychosocial information and (2) the patient is an active participant in the communication process, providing necessary psychosocial information and taking part in medical decisions. Over time, the predominant model for doctor–patient relationships has shifted from the paternalistic, doctor-centered approach to the more egalitarian, patient-centered approach. Although the patient-centered approach appears to stress psychosocial information in contrast to a more biomedical model, the term 'biopsychosocial' is often used to indicate the importance of balancing these sources of information. The doctor-centered versus patient-centered nature of the interaction is one of the most basic features of physician–patient communication; however, some critics argue that neither approach represents the entirety of the doctor–patient interaction itself (similar to how the doctor-centered approach does not represent the entirety of the patient).

The relationship-centered approach seeks to incorporate the knowledge, attitudes, and beliefs that come from the provider's perspective as well as the patient's, focusing on the relationship itself as an entity comprising and continuously influenced by the characteristics of both the patient and provider.

In Hall and Roter's (2007) review of doctor–patient interactions, the authors highlight five characteristics that exemplify the relationship-centered approach, namely, relationship-centered care must be medically functional, facilitative, responsive, informative, and participatory. In this sense, these are not responsibilities of only the physician or the patient; rather, they are responsibilities of the patient–provider relationship itself. The 'medically functional' aspect refers to the degree to which the doctor–patient relationship affects the quality or performance of basic medical procedures, such as scheduling the visit and using resources appropriately (i.e., in terms of managing the visit as well as the technical procedures used for diagnosis and treatment). The 'facilitative' aspect refers to the degree to which the relationship allows the patient to talk openly and completely about the illness and/or psychosocial distress in a manner that is well understood and considered by both parties. The 'responsive' aspect refers to the degree to which the patient's emotional concerns are met with supportive, empathic attempts to address these issues. The 'informative' aspect refers to the degree to which medical information and recommendations are shared in a manner that the patient understands and is motivated to use properly. The 'participatory' aspect refers to the degree to which patients are able and encouraged to play a significant role in the consultation process (e.g., in medical discussions and decision making) and to be generally responsible about their medical care.

Beyond more general distinctions of physician–patient interaction (e.g., problem-focused versus emotion-focused responses), a review of doctor–patient communication (Roter and Hall, 1992) indicates there are six categories of communication that occur between physicians and patients: information giving, information seeking, social conversation, positive talk, negative talk, and partnership building. Roter and Hall's (1992) examination of dialogue between patients and physicians also provides a statistical representation of the average frequencies of various types of talk that occur during the medical interaction. On average, 35% of physicians' talk involves information giving (e.g., providing medical information, counseling, and recommendations), 23% involves information seeking (e.g., asking closed-ended or open-ended questions), 6% involves social conversation (e.g., nonmedical or casual discussions), 15% involves positive talk (e.g., statements of approval or facilitative back channel responses like 'mm-hmm' and 'aha'), about 1% involves negative talk (e.g., disagreements, confrontations, or antagonism), and 11% involves partnership building (e.g., attempts to get a patient more involved and engaged in the medical interaction). On average, 47% of patients' talk involves information giving (e.g., answering the care provider's questions), 7% involves information seeking (i.e., asking questions), 13% involves social conversation, 19% involves positive talk, and 8% involves negative talk (e.g., disagreements and criticisms of the physician). Some of these frequencies can be explained by the tendency for physicians to talk more than patients, contributing to about 60% of the conversation. The average medical consultation lasts 20 min or less (although there is much variation), and up to half of this time can include periods of silence, interruptions, or activities that do not directly involve the patient. Physicians' reliance on closed-ended questions and medical counseling may be seen as a way to maximize the time spent during the medical encounter, but they may still be expressing negativity in a variety of indirect ways. Meanwhile, patients may feel obligated to answer the physician's questions and be too intimidated to ask questions of their own, but they tend to engage in more nonmedical conversation and can more freely and directly express their emotional involvement through positive and negative talk. By and large, patients are more likely to disclose health information when the physician discusses more psychosocial issues and displays a supportive, attentive nature.

The overall quality of the doctor–patient interaction can be assessed in a variety of ways. From a qualitative review of patient perspectives, the quality of the medical relationship depends on how long they have worked with the same doctor and how well their various interactions have proceeded (Ridd et al., 2009). This is further enhanced by an egalitarian approach that includes the patient and doctor's knowledge, understanding, trust, and liking of each other. More quantitative methods involve surveys and questionnaires that examine various dimensions of the doctor–patient relationship (e.g., trust empathy and relationship building). Researchers also employ the rating and measuring of behaviors, statements, and other communication aspects present in audio- and video-recorded doctor–patient interactions. Methods such as examining previously recorded doctor–patient interactions provide relevant insight into variables related to communication quality. For example, what appears to be a negative medical consultation may be explained by negative behaviors exhibited by the patient (e.g., in response to their preexisting illness) that are misdirected or misinterpreted but ultimately reciprocated by the care provider. Because of this, variables such as the doctor and patient's satisfaction with the medical visit are often positively correlated.

Antecedents of the Doctor–Patient Interaction

The doctor–patient interaction is a dynamic process involving a given patient and care provider; therefore, a number of factors influence this relationship. Something as simple as the time of day can determine how long the medical consultation lasts, with earlier visits lasting longer than average and later visits lasting shorter (Byrne and Long, 1976). At an individual level, patients differ in how much they want to be involved in the medical process, how much medical information they would like to receive, and what they expect the medical consultation to accomplish. Generally speaking, older patients, male patients, nonwhite patients, patients with lower incomes, and patients with lower levels of educational or occupational attainment tend to prefer or display more passive, dependent relationships with their care providers (e.g., Arora and McHorney, 2000). Patients also vary in how appropriate they believe it is to provide psychosocial information to their care provider, which happens to be a key component of the patient-centered and relationship-centered approaches. Basic

differences among physicians follow a similar pattern, with variation in how much and what type of information they believe the patient should receive, the role the patient should play in the consultation and decision-making processes (e.g., active or passive), and their focus on the physiological versus psychosocial nature of the information given to, provided by, or discussed with the patient. Several studies (e.g., Street et al., 2005) suggest that patient participation depends on a variety of factors, such as the clinical setting or the degree to which the physician adopts a patient-centered communication style, as well as the interaction between several personal and situational characteristics.

Both the patient and the provider have specific attitudes, beliefs, values, personalities, and cultural backgrounds, and these characteristics cannot be readily detached from the interaction at hand. For example, expectancy effects may occur as a result of the patient's cultural background and the physician's prior experience or beliefs regarding that cultural background, or a patient may ultimately receive better or worse medical care simply because of how much the physician likes that individual. The following sections detail characteristics of the patient and provider that may influence the medical interaction and its outcomes.

Gender

Gender and engendered differences in socialization practices influence the expression of a large number of personality characteristics, many of which are relevant to the quality of the doctor–patient interaction. For example, large gender differences in communication style have been consistently reported in previous literature (e.g., Hall, 1984/1990, 2006). Women generally exhibit better nonverbal communication skills; they tend to more accurately express and interpret emotional cues, and they also display more interpersonal warmth and engagement (e.g., smiling, eye contact, standing close, and touching). Men generally appear less emotionally concerned and tend to be more dominant in their interpersonal interactions. Although evidence of these gender differences is often taken from outside of clinical settings, Roter et al. (2002) review of the literature indicates that men and women also tend to show these behavioral differences in the clinical context, with implications for both patients and physicians. For example, nonverbal sensitivity may allow a physician to detect and address feelings or misgivings the patient may hesitate or be unable to convey verbally.

Male and female physicians typically provide the same amount of medical information, but female physicians often apply more psychosocial aspects to the discussion. Furthermore, women tend to ask more questions and encourage patient participation while also displaying more positive and/or emotional verbal and nonverbal behavior. They also tend to have longer visits, report liking their patients more, and more readily adopt a patient-centered, psychosocial approach to practicing medicine (Roter and Hall, 2004). Ultimately, patients respond to male and female practitioners differently due to these discrepancies in communication styles, engaging more with female physicians (e.g., by talking more, providing more information, or otherwise participating more in the interaction). However, even though female practitioners tend to exhibit arguably more desirable communication styles and report liking their patients more than male practitioners do, there seems to be no gender difference in the patients' liking of or satisfaction with their physicians.

Gender differences in communication styles also occur among patients, with women participating more during the interaction. However, this may also be based on expectations and responses from the care provider, as women tend to receive more opportunities to be involved in the interaction, and the information they receive tends to be presented in a more understandable manner. These differences are highlighted by studies that show patients participating the least when both the patient and physician are men (Kaplan et al., 1995). Despite this, there seems to be no gender difference indicating women are more satisfied with their health care interactions than men are.

Ethnicity, Culture, and Socioeconomic Status

The literature documenting ethnic disparities in health care is fairly extensive and consistent (e.g., Murray and Woods, 2009; Whitfield et al., 2012). Ethnic and class differences that occur in other societal institutions also appear in the contexts of doctor–patient interaction and medical treatment. In particular, black American, Latin American, and lower income patients tend to receive poorer quality health care. While ethnicity and social class are often related, it is important to note there are differences in doctor–patient interactions based on social class alone. Patients with lower socioeconomic status tend to receive less medical information, less emotional support, and less encouragement to participate in the medical interaction and decision-making process (Verlinde et al., 2012).

When patients and care providers from different backgrounds interact, the quality of the health care generally suffers if they do not successfully overcome differences in culture and language (Ferguson and Candib, 2002). Patients from different ethnic or cultural backgrounds are less likely to generate empathy and establish rapport with the physician, which can lead to reduced participation and transmission of information and, ultimately, worse patient outcomes (e.g., satisfaction, adherence, and health). Cross-ethnicity doctor–patient interactions also tend to be significantly shorter in length, and many of these negative effects are most likely due to ethnicity-related attitudes and beliefs that supersede the interaction at hand.

Differential treatment of patients based on ethnicity may be unconscious or unintentional, but the physician's negative stereotypes, attitudes, and beliefs still have a direct effect on the quality of health care the patient receives. For example, black Americans received fewer recommendations for heart surgery, even though they demonstrated the same clinical need as other patients, and black Americans ultimately underwent fewer heart surgery procedures (Maynard et al., 1986). Furthermore, a study by DiMatteo et al. (2009) indicates there is an intersection between gender and ethnicity regarding health disparities, such that African-American women are disadvantaged even more than African-American men, and

gender and ethnicity concordance of the physician and patient appears to play a large role in this.

These disparities and patient reports of lower health care quality indicate physicians may simply neglect, ignore, underestimate, or otherwise fail to consider the needs, expectations, characteristics, and abilities of their ethnically or culturally different patients. Likewise, patients may have particular expectations and beliefs that are not expressed during the interaction. In both cases, the medical consultation may be unconsciously influenced by patient and physician behaviors (or lack thereof) that confirm preexisting attitudes, beliefs, or stereotypes held by either party. A patient's cultural background can determine what medical problems or symptoms are considered significant by the physician, as well as what medical problems or symptoms are presented or discussed in the first place. Ferguson and Candib (2002) suggest increasing the diversity of medical personnel and for majority status physicians to improve their relationships and communication skills with minority status patients. Similarly, cultural sensitivity training programs have been found to improve both the care provider's communication abilities and the patient's health outcomes (Majumdar et al., 2004). Different ethnicities and cultures communicate and cope with illness in distinct ways, so it is imperative that physicians are sensitive to both the common and distinct needs of various peoples.

Physical and Psychosocial Health

Aspects of the patient's health, more so than the physician's, can influence the doctor–patient interaction. There are a number of ways the patients' mental and/or physical health is correlated to how the physician and patient communicate with each other (see Hall et al., 1996). Sicker patients tend to offer more medical information to their care providers, and both parties discuss the possible factors of the illness more. However, there also appears to be less social conversation between both parties when the patient is ill or otherwise distressed, which may result in less provision and discussion of important psychosocial information, more overall dissatisfaction with the medical consultation, or the loss of many other intangible benefits of a healthier, more productive interaction between the physician and patient.

A patient's distress can increase the salience of their mental or physical illness, prompting more concern about the medical interaction at hand. This salience may ultimately lead care providers to respond differently to sicker patients. For example, physicians report liking their sicker patients less than their healthier patients (Hall et al., 1996). In conjunction with the lack of social conversation, this resembles the more physiological and less personal or participatory approach to the medical consultation. A similar issue occurs with patients of greater age, as older patients typically have more salient mental and/or physical distress or other medical problems. There are a number of characteristics that distinguish older patients from others (e.g., the more life-threatening nature of their medical problems), but research indicates interactions between physicians and older patients are less participatory and more focused on providing biomedical information (e.g., Kaplan et al., 1995).

Third Parties and the Internet

The lack of participation among patients who face obstacles to communicating with their care provider is common enough that companions or interpreters often join the medical consultation to facilitate a more patient-centered approach. Adding a third party can complicate, simplify, or otherwise influence the medical consultation in various ways (e.g., Ferguson and Candib, 2002). The general recommendation is to use professional interpreters to overcome culture or language barriers and increase access for those in need, but every physician might not display improved communication skills under these circumstances. Although the presence of additional medical professionals tends to relate to a lower quality of communication, there has been relatively little examination of exactly how the presence of patients' companions influences the dynamics of the doctor–patient relationship.

Perhaps one of the most significant third-party influences on doctor–patient interactions has been the increasing use of the Internet as a tool for medical information and/or communication. Editorials by physicians and researchers alike have questioned the appropriate use of media such as online social networks and their effects on traditional medical consultations. While these social networks can act as an additional channel of communication between the patient and the care provider, it appears that maintaining a professional relationship as well as individual privacy is achieved by physicians properly distancing themselves from their patients in this online space. Guidelines do exist regarding physicians' use of social media (e.g., American Medical Association, 2010), as there may be unintended consequences for the doctor–patient relationship if this new dimension of communication is not properly addressed.

Beyond acting as a medium of communication, the Internet can also be considered an intermediary of the doctor–patient interaction in the form of providing health information or medical recommendations. This up and coming 'triangulation' of the patient, physician, and Internet has been thoroughly reviewed by Wald et al. (2007), who propose that there is great potential for the Internet to improve medical interactions. Among the benefits the authors list are more informed health care choices, enhanced communication and information provision, increased partnership as part of a more collaborative approach, increased availability of social support, and even managerial gains such as better use of time during clinical visits and increased access to information like health records. The emerging role of the Internet in medical communication also means there are potential drawbacks, such as increased misinformation, intensified socioeconomic disparities due to limited Internet access, increased physician liability, and challenges to the physician's role or the basic dynamics of the physician–patient interaction; however, the authors and others (e.g., Medical Library Association, 2013) do provide suggestions on how best to integrate the Internet into the doctor–patient relationship.

Outcomes of the Doctor–Patient Interaction

Doctor–patient communication has been studied extensively with regard to various patient outcomes, including the

understanding and recall of medical advice and information, satisfaction with the medical visit, adherence to medical recommendations, sense of confidence and autonomy, trust in the care provider, and overall health. Aside from patient outcomes that may ultimately influence the patient's health, the quality of doctor–patient communication is also directly related to health outcomes such as increased recovery rates, decreased use of pain medication, shortened stays in the hospital, and improvements in physiological indicators such as blood pressure and blood sugar. Patient-centered and relationship-centered behaviors (e.g., being informative and supportive and building rapport and a partnership) are generally associated with improved health (e.g., symptom resolution). Although the medical interaction is predictive of a number of health outcomes, the thoroughly researched proximal variables of patient satisfaction and adherence are discussed in detail.

Satisfaction

Patient satisfaction can generally be defined as how well the patients' expectations are met by the actual care and treatment they receive throughout the medical interaction. Because patient satisfaction concerns the quality of health care received, it can be measured in reference to a number of health care aspects (e.g., a specific doctor or visit, access to and quality of the facility, technical skills and procedures used, and information provided). The present article discusses patient satisfaction regarding the physician's behavior, as this tends to be the aspect of patient satisfaction that is the most salient and influential.

Patient satisfaction is typically related to the doctor's communication style (i.e., doctor centered versus patient centered), with the patient-centered approach being related to more patient satisfaction than the doctor-centered approach. Both patients and physicians tend to be more satisfied when the physician uses a psychosocial approach rather than a strictly biomedical one. Patient satisfaction is related to physician's nonverbal communication abilities, as well. Physicians who are better able to encode (i.e., express) emotional nonverbal cues and decode (i.e., accurately interpret) emotional expressions of their patients tend to have more satisfied patients (DiMatteo et al., 1986). A higher degree of nonverbal immediacy (i.e., closeness demonstrated by behaviors like gazing, leaning forward, and affectionate touching) displayed by the physician is also related to higher ratings of patient satisfaction. These results are supported in a review by Lelorain et al. (2012), which found that greater empathy, as it is often related to or indicated by nonverbal immediacy and patient-centered care, is positively related to patient satisfaction.

Specific physician behaviors that tend to be associated with higher ratings of patient satisfaction include information giving, partnership building, positive talk, discussion of psychosocial aspects, social conversation, perceived confidence, and liking of and/or familiarity with the patient. Likewise, the more active role the patient plays, the longer the visit lasts, and the degree to which the physician takes a patient-centered or egalitarian approach to the medical interaction also predicts higher ratings of patient satisfaction. Factors related to decreased patient satisfaction include the degree to which the physician dominates the medical interaction, the amount of biomedical questions asked, and the poor health status of the patient. It is important to note that many of these factors of satisfaction and dissatisfaction may certainly co-occur and influence each other throughout the medical interaction.

Dissatisfied patients are less likely to return for future medical visits, follow medical regimens and recommendations, or develop a healthy doctor–patient relationship. The most dissatisfied patients are more likely to sue the care provider for malpractice. Interestingly, physicians are more likely to be sued for malpractice when the patient believes there was poor communication or an otherwise uncaring attitude rather than every instance that involves medical error. Levinson et al. (1997) found that, compared to doctors who had been sued for malpractice, those who never had a malpractice suit held longer visits, laughed more with their patients, and were more likely to use a patient-centered or egalitarian approach that emphasized the patient's active role and attempts for both parties to understand each other.

Adherence to Medical Advice

Medical visits, consultations, or other interactions generally result in some sort of recommendation by the care provider. Patient adherence refers to how well the patient follows the care provider's advice (e.g., taking prescribed medicine, undergoing suggested treatments, or making recommended lifestyle changes). Where the lowest levels of patient satisfaction can lead to malpractice suits, low patient adherence can be directly related to poor health outcomes. Overall rates of patient nonadherence are fairly high. Recent analyses (e.g., DiMatteo, 2004) indicate that adherence rates have been rising over the years, with some 50–75% of patients adhering to medical recommendations according to studies conducted after 1980. Incidentally, a study by DiMatteo et al. (2002) reveals that patient adherence is associated with health outcomes, meaning communication behaviors that improve adherence may indirectly improve the patient's health outcomes as well.

A significant part of patient nonadherence is due to the patient's lack of understanding medical information and advice or the physician's inability to communicate this information and advice properly. One meta-analysis (Haskard Zolnierek and DiMatteo, 2009) reveals that the quality of the physician's communication is significantly related to the patient's adherence, and patient adherence tends to improve when physicians are trained to be better communicators (e.g., through a patient-centered approach). Doctors have a tendency to overestimate the degree to which the patient is adhering to medical advice or following treatment and prescription guidelines. Additionally, adherence is rarely brought up as a topic of conversation, either by the patient volunteering information to the physician about their adherence or the physician asking the patient directly about adherence to the medical regimen.

As mentioned previously, the reciprocity of the positive or negative affect during the physician–patient interaction may have an effect on patient outcomes. For example, if patients feel particularly negative toward the physician – even if this was prompted by the physician's generally dissatisfied or negative

attitude – patients will be less likely to adhere to the medical advice they are given. Similar to the relationship between physician nonverbal sensitivity and patient satisfaction (DiMatteo et al., 1986), more sensitive physicians tend to have more adherent patients. Nonverbal sensitivity may be indicative of several other beneficial aspects of the doctor–patient interaction (e.g., a more patient-centered approach), and patients tend to be more adherent when they ask questions and the physician delivers information in a more understandable manner. The Information-Motivation-Strategy model highlights the most important mechanisms through which patient adherence is attained (DiMatteo et al., 2011). According to this model, the doctor–patient relationship should ensure that the patient accurately understands the medical recommendations, is properly motivated to accept the recommendations, and has the appropriate means, abilities, and/or support system to follow the recommendations.

Interventions for Doctor–Patient Interaction

Doctor–patient communication interventions primarily involve some sort of training of the patient or the care provider. Many efforts have been made to make the quality of the doctor–patient interaction salient to both patients and their care providers, and this tends to improve the patients' outcomes. Addressing the patients' concerns and providing the information they request in an understandable manner can prevent miscommunication and alleviate the negative effects of distressing news, for example. Over time, researchers as well as medical and educational institutions have come to understand communication training is a necessary aspect of medical care that improves the skills and abilities of the physician. Presently, communication skills are assessed or otherwise required by national licensing and accreditation associations in the United States, and many medical schools have incorporated training programs designed to improve communication skills (e.g., by encouraging physicians and residents to ask open-ended questions and increase nonverbal immediacy). Such training programs and interventions are essentially encouraging care providers to adopt a more patient-centered or relationship-centered approach. This stands in contrast to previous observations that, among more traditional medical school approaches, younger medical students were more likely to engage in a psychosocial patient-centered approach, but this tendency declined and became largely replaced by a biomedical doctor-centered approach as they progressed through the program. Despite the changes in medical education and practice, recent studies (e.g., Haidet et al., 2002) indicate this pattern of behavior has not disappeared over the years. Medical students still tend to display relatively less egalitarian, psychosocial attitudes and more doctor-centered attitudes as they progress through medical school, although the overall level of patient-centered attitudes in medical practice has still reasonably increased over time.

Interventions aimed at addressing physician's communication skills often lead to improvements in both the physician's abilities and outcomes such as the patient's satisfaction. As indicated by Haskard Zolnierek and DiMatteo's (2009) meta-analysis, training physicians' communication abilities may improve patient adherence through a number of means (e.g., transmission of information, encouragement of partnership and patient involvement, and providing a more open and supportive environment for discussion). Communication training has led to increased sensitivity to patients' psychosocial problems as well the ability to address such problems, and training has also been shown to improve the quality of the medical interaction by making the physician better able to relate to the patient, communicate empathy, and elicit information through a more psychosocial approach. Physician training appears to be an effective intervention, whether it is a single session that lasts a few hours or continuous training that lasts for the duration of medical school. A meta-analysis by Roter et al. (1998) reveals that physician interventions designed to improve adherence have positive effects on a variety of outcomes (e.g., severity of the illness, prescription refilling, and maintaining medical appointments) across a variety of diagnoses (e.g., cancer, diabetes, and psychological disorders). These interventions tended to work on a situational basis, but the authors do identify three broad categories of physician intervention strategies: educational (i.e., increasing patients' knowledge by teaching them and/or providing educational materials), behavioral (i.e., guiding behavioral patterns related to medical interactions through practice, skill-building, and modeling, for example), and affective (i.e., appealing to the patient's emotions and/or social support networks).

Interventions targeted at patients are also effective, as they generally encourage patients to become more active and involved in the medical interaction, promoting a more balanced, patient-centered or relationship-centered approach. Some interventions also attempt to address the patient's health literacy directly, and in contrast to the study by Wald et al. (2007), increasing the quality of health information technology can reduce health disparities. Providing patients with information such as an audio recording of the doctor–patient interaction or a written copy of their medical record has also shown promise in improving patients' understanding and recall of relevant medical information. However, most interventions aimed at patients attempt to improve the doctor–patient interaction by providing the patient with information or guides about communicating during the medical interaction (e.g., by encouraging them to ask more questions, seek more information, and take an active role in providing information). Interventions that provide patients with written information are indeed effective, but a review by Harrington et al. (2004) indicates patients participate in the medical interaction the most after receiving a face-to-face intervention or training session. In any case, face-to-face interventions in which someone personally reviews information and practices skills with the patient have been shown to improve health outcomes and the patient's perception of their health status.

Conclusions

Doctor–patient interactions are an unquestionably important aspect of the medical system. They are also multifaceted, multidimensional, and still in need of much more research before more definitive conclusions and suggestions can be

made regarding how to improve and better understand them. Nevertheless, the present article attempts to provide a basic explanation of the number of psychological influences on, components of, and effects resulting from the doctor–patient interaction. The patient-centered and relationship-centered approaches that incorporate the entirety of the personal and psychosocial aspects of the individual (i.e., a balanced use of biomedical or evidence-based information) are generally considered to be a vast improvement in medical practice. Research that further contextualizes variables related to the physician, patient, and interaction across a number of health care situations, procedures, and environments may be able to provide more consistent improvements in the quality of health care delivery, as seen in the case of widespread increased physician communication training over the years. Aspects of doctor–patient interactions and health care delivery gradually change just as society changes, and effective health care relies on continued efforts to understand how these changes affect the medical interaction and how to address such changes in meaningful ways. Perhaps one of the most impending challenges for doctor–patient communication is how to incorporate the Internet and other advancing technologies in a useful manner. Accentuating the positives and eliminating the negatives of this emerging third-party influence may ultimately provide a new level of quality in doctor–patient interactions, but this will require a continued progression of research and practice.

See also: Culture as a Determinant of Mental Health; Evidence-Based Medicine; Gender and Health Care; Health Education and Health Promotion; Health Literacy; Health Outcomes, Assessment of; Patient Adherence to Health Care Regimens; Race: Ethnicity and Health; Socioeconomic Status and Health Care; Therapist–Patient Relationship.

Bibliography

American Medical Association, 2010. Professionalism in the Use of Social Media. http://www.ama-assn.org/ama/pub/physician-resources/medical-ethics/code-medical-ethics/opinion9124.page (accessed 13.05.27.).

Arora, N.K., McHorney, C.A., 2000. Patient preferences for medical decision-making: who really wants to participate? Medical Care 38, 335–341.

Byrne, P.S., Long, B.E.L., 1976. Doctors Talking to Patients. Her Majesty's Stationery Office, London, UK.

DiMatteo, M.R., 2004. Variations in patients' adherence to medical recommendations: a quantitative review of 50 years of research. Medical Care 42, 200–209.

DiMatteo, M.R., Giordani, P.J., Lepper, H.S., Croghan, T.W., 2002. Patient adherence and medical treatment outcomes: a meta-analysis. Medical Care 40, 794–811.

DiMatteo, M.R., Haskard-Zolnierek, K.B., Martin, L.R., 2011. Improving patient adherence: a three-factor model to guide practice. Health Psychology Review 6, 74–91.

DiMatteo, M.R., Hays, R.D., Prince, L.M., 1986. Relationship of physicians' nonverbal communication skills to patient satisfaction, appointment noncompliance, and physician workload. Health Psychology 5, 581–594.

DiMatteo, M.R., Murray, C.B., Williams, S.L., 2009. Gender disparities in physician-patient communication among African American patients in primary care. Journal of Black Psychology 35, 204–227.

Epstein, R.M., Franks, P., Fiscella, K., Shields, C.G., Meldrum, S.C., Kravitz, R.L., Duberstein, P.R., 2005. Measuring patient-centered communication in patient–physician consultations: theoretical and practical issues. Social Science & Medicine 61, 1516–1528.

Ferguson, W.J., Candib, L.M., 2002. Culture, language, and the doctor-patient relationship. Family Medicine 34, 353–361.

Haidet, P., Dains, J.E., Paterniti, D.A., Hechtel, L., Chang, T., Tseng, E., Rogers, J.C., 2002. Medical student attitudes toward the doctor-patient relationship. Medical Education 36, 568–574.

Hall, J.A., 1984/1990. Nonverbal Sex Differences: Communication Accuracy and Expressive Style. Johns Hopkins University Press, Baltimore, MD.

Hall, J.A., Roter, D.L., 2007. Physician-patient communication. In: Friedman, H.S., Silver, R.C. (Eds.), Foundations of Health Psychology. Oxford University Press, New York.

Hall, J.A., Roter, D.L., Milburn, M.A., Daltroy, L.H., 1996. Patients' health as a predictor of physician and patient behavior in medical visits: a synthesis of four studies. Medical Care 34, 1205–1218.

Harrington, J., Noble, L.M., Newman, S.P., 2004. Improving patients' communication with doctors: a systematic review of intervention studies. Patient Education and Counseling 52, 7–16.

Haskard Zolnierek, K.B., DiMatteo, M.R., 2009. Physician communication and patient adherence to treatment: a meta-analysis. Medical Care 47, 826–834.

Kaplan, S.H., Gandek, B., Greenfield, S., Rogers, W., Ware, J.E., 1995. Patient and visit characteristics related to physicians' participatory decision-making style: results from the medical outcomes study. Medical Care 33, 1176–1183.

Lelorain, S., Brédart, A., Dolbeault, S., Sultan, S., 2012. A systematic review of the associations between empathy measures and patient outcomes in cancer care. Psycho-oncology 21, 1255–1264.

Levinson, W., Roter, D.L., Mullooly, J., Dull, V., Frankel, R., 1997. Doctor-patient communication: a critical link to malpractice in surgeons and primary care physicians. Journal of the American Medical Association 277, 553–559.

Majumdar, B., Browne, G., Roberts, J., Carpio, B., 2004. Effects of cultural sensitivity training on health care provider attitudes and patient outcomes. Journal of Nursing Scholarship 36, 161–166.

Maynard, C.L., Fisher, D., Passamani, E.R., Pullum, T., 1986. Blacks in the coronary artery surgery study (CASS): race and clinical decision making. American Journal of Public Health 76, 1446–1448.

Medical Library Association, 2013. A User's Guide to Finding and Evaluating Health Information on the Web. http://www.mlanet.org/resources/userguide.html (accessed 27.05.13.).

Murray, C.B., Woods, V.D., 2009. Psychology of health disparities among African American populations: an overview. Journal of Black Psychology 35, 142–145.

Ridd, M., Shaw, A., Lewis, G., Salisbury, C., 2009. The patient-doctor relationship: a synthesis of the qualitative literature on patients' perspectives. British Journal of General Practice 59, 268–275.

Roter, D.L., Hall, J.A., 1992. Doctors Talking with Patients/Patients Talking with Doctors: Improving Communication in Medical Visits. Auburn House, Westport, CT.

Roter, D.L., Hall, J.A., 2004. Physician gender and patient-centered communication: a critical review of empirical research. Annual Review of Public Health 25, 497–519.

Roter, D.L., Hall, J.A., Aoki, Y., 2002. Physician gender effects in medical communication: a meta-analytic review. Journal of the American Medical Association 288, 756–764.

Roter, D.L., Hall, J.A., Merisca, R., Nordstrom, B., Cretin, D., Svarstad, B., 1998. Effectiveness of interventions to improve patient compliance: a meta-analysis. Medical Care 36, 1138–1161.

Street, R.L., Gordon, H.S., Ward, M.M., Krupat, E., Kravitz, R.L., 2005. Patient participation in medical consultations: why some patients are more involved than others. Medical Care 43, 960–969.

Verlinde, E., De Laender, N., De Maesschalck, S., Deveugele, M., Willems, S., 2012. The social gradient in doctor patient communication. International Journal for Equity in Health 11. http://www.equityhealthj.com/content/11/1/12 (accessed 27.05.13.).

Wald, H.S., Dube, C.E., Anthony, D.C., 2007. Untangling the web: the impact of internet use on health care and the physician-patient relationship. Patient Education and Counseling 68, 218–224.

Whitfield, K.E., Bogart, L.M., Revenson, T.A., France, C.R., 2012. Introduction to special section on health disparities. Annals of Behavioral Medicine 43, 1–3.

Documentary and Ethnographic Film

Kizhavana P Jayasankar and Anjali Monteiro, School of Media and Cultural Studies, Tata Institute of Social Sciences, Deonar, Mumbai, India

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Documentary film is defined in terms of its difference from other genres of nonfiction film and from fiction film. John Grierson first employed the term 'documentary,' which he regarded as "a creative treatment of actuality" (Grierson, cited in Barnouw, 1974: p. 287). Ethnographic film belongs to the domain of cultural anthropology, which traditionally involved the study and scientific observation of communities. This article presents a brief history of documentary and ethnographic film with authorial positions as the organizing principle. Debates about these genres revolve around relationships between the author and the narrative, between the author and the subject filmed, and between the narrative and audiences.

Documentary and Ethnographic Film

The term documentary came into usage in the nineteenth century as an adjective referring to documentation, hence with connotations of evidence (Rosen, 1993: p. 6). With reference to film, it was first employed by John Grierson (in his 1926 review of Robert Flaherty's film *Moana*) to indicate a body of work that documented reality, involving "a creative treatment of actuality" (Grierson, cited in Barnouw, 1974: p. 287). The twin roles of the documentary filmmaker, as a creator with an agenda and as a purveyor of 'actuality,' have been at the root of several debates concerning the truth claims of the documentary. In some accounts, documentary film is still regarded as indexically representing a real world, mimetically reproducing pictures of reality with a view to providing knowledge about it, as is the case in any scientific endeavor.

On the one hand, developments in imaging technologies have perfected hardware and software enabling the unobtrusive documentation of reality; on the other, the very notions of 'documentary' and 'reality' have undergone radical reformulation. In theory and practice, the truth claims of documentary and its status as 'evidence' of a given historical world have increasingly been problematized in the contemporary postdigital cultural context. Over time, documentary film has come to embody diverse modes of negotiating the complex set of relationships between the author, the film text, the audiences, and the world.

Ethnographic film comes under the rubric of cultural anthropology, which traditionally involved scientific observation and recording of the lives and cultures of other races, usually indigenous communities in colonized lands. Jean Rouch terms anthropology as "the eldest daughter of colonialism" (cited in Winston, 1993: p. 51), thus foregrounding the inherent relationship of power that underpins any encounter between the anthropologist and the other. The premise that ethnographic film is a value-free, indexical scientific record has been subjected to critical interrogation and revision, leading to dramatic changes in its theory and practice.

Demarcating the Terrain of Documentary Film

Documentary versus Other Nonfiction

Documentary film has been regarded as a specific body of work under the broader rubric of 'nonfiction' film. The latter includes genres such as the first 'actuality' films of the Lumiere brothers, the newsreel, the travelogue, and the educational film. These genres, while sharing in common with documentary the use of actuality footage, have been seen as distinct because of the differing context of their production and use (Barsam, 1992). Though documentary and ethnographic film occupy distinct terrains within nonfiction film, the issues confronting both share a great degree of commonality.

Documentary versus Fiction

The fiction film, to which the documentary and other nonfictional genres were opposed, involved an obvious process of construction to offer an illusion of reality. The documentary filmmaker, however, bore the onus of providing a window to the world. The twin positions eschewed by the documentary filmmaker were that of the 'manipulative' creator of fiction, on the one hand, and the didactic educator on the other. The control and omnipotence represented by both these positions, which 'told' audiences about an external reality, was rejected in favor of a position that purported to 'show' its viewers the world.

Toward a Taxonomy

The wide array of themes, styles, formats, and modes of address employed by documentary filmmakers make it difficult to construct an inclusive typology that accounts for this variety. Barnouw (1974), for instance, chooses to classify documentaries on the basis of the role of the maker (explorer, reporter, advocate, poet, observer, guerilla, and so on), while emphasizing that these roles are not mutually exclusive. However, certain roles have been privileged at certain historical moments, making it possible to construct a typology on this basis. Barsam (1992) follows a classification employing time and space (American/British/European and Asian, at various periods of time).

Nichols (2010) identifies six modes of representation. The expository mode, used by many of the early filmmakers, such as Grierson and Flaherty, as also in current television network news, tends to use 'voice-of-God' commentary or poetic narration to inform or enlighten the viewer about the world outside. The argument of the commentary is what propels the text, the visuals serving more as illustration or counterpoint.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.95013-X

The expectation of the viewer is generally to obtain certain, authoritative knowledge about reality, to understand cause and effect, problems and solutions. Recent examples of this mode include *The Corporation* (2003) by Mark Achbar and Jennifer Abbott, and *Grizzly Man* (2005) by Werner Herzog.

The poetic mode treats narratives differently, with little commitment to representing actuality or presenting facts. With its emphasis on affect and the expressive dimensions of cinema, it "stress[es] visual and acoustic rhythms, patterns, and the overall form of the film" (Nichols, 2010: p. 150). *Koyanisqatsi* (1982) by Godfrey Reggio, a purely visual narrative set to the music of Philip Glass, represents this mode.

The observational mode (also referred to as Direct Cinema) was facilitated by the invention of lightweight equipment with synchronous sound recording capability. It rejects the relatively didactic stance of the expository mode, in order to bring to viewers, life as it unfolds. The filmmaker aspires to the role of a detached observer and "look[s] on as the social actors go about their lives as if the camera were not present" (Nichols, 2010: p. 150). The expectation of the viewer is to gain privileged access to an unmediated external world. *Control Room* (2004) by Jehane Noujaim, on Al Jazeera's reportage of the US–Iraqi war, is a recent example of such work.

The participatory mode also draws on the availability of mobile equipment, but rejects this detachment in favor of a more interventionist role. The filmmaker's relationship with the subjects of the film comes to the fore and is not invisible, as in the observational mode. The focus in on local knowledge; the filmmaker is not confined to creating continuities in time and space, but could employ unusual juxtaposition, archival material, and other devices in constructing the narrative. The viewer's expectations are situated in the realm of the interpretative, since the mode makes no claim to absolute truth and neutral knowledge. Nichols (2010) cites the films *Fog of War* (2003) by Errol Morris and *Enron: The Smartest Guys in the Room* (2005) by Alex Gibney as exemplifying this mode.

In the reflexive mode, the act of representing the world itself becomes central to the film's narrative. The premises from which the narrative is constructed and the process of construction are available to the viewer. While it bears some resemblance to the participatory mode, the reflexive mode, in addition to problematizing the relationship between the filmmaker and the subjects, also problematizes the relationship between the narrative/filmmaker and the viewer. The camera is not an evidence gathering apparatus. The conventions of realism, when used, are interrupted and put to question. Broadly speaking, the reflexive mode draws on the post-structuralist critique of the indexical and instrumental view of language and the speaking subject (*see* Semiotics). Trinh T. Minh-ha's *Reassemblage* (1982) and Elizabeth Barrett's *Stranger with a Camera* (1999) are reflexive documentaries.

The performative mode problematizes the idea of knowledge as universal and abstract, affirming an understanding of knowledge that is embodied, evocative, and experiential. In cinematic terms, it plays with a combination of the "actual and the imagined" (Nichols, 2010: p. 202) in order to convey a visceral sense of the subjective world of the filmmaker. *The Gleaners and I* (2000) by Agnes Varda, *Bontoc Eulogy* (1995) by Marlon Fuentes, and *Tongues Untied* (1989) by Marlon Riggs represent the performative mode.

A Brief History of Documentary and Ethnographic Film

The theoretical positions and the narrative strategies within the discourse of the documentary film have undergone fundamental changes, making it impossible to establish continuity across them. In spite of this apparent diversity, they could perhaps be classified in terms of the notions of authorship and viewership that they are, implicitly or explicitly, putting forth or negating. A useful device in establishing intelligibility across the vast corpus of documentary and ethnographic film practice is, therefore, to look at the various authorial stances that inform it. This, by no means exhaustive, historical account explores the work of major figures and movements, which contributed to the development of new conventions, styles, and norms within documentary film.

The Author as Reconstructing the Authentic Other – Robert Flaherty

Flaherty's *Nanook of the North* (1922), which celebrates the heroic struggle for survival of its Inuit protagonist, could perhaps be regarded as the first ethnographic film. Flaherty 'manipulated reality' in order to reveal what he regarded as the 'true spirit' of the culture. Unlike later ethnographic work, which trod the path of 'flat' scientific documentation, Flaherty's work, whose narrative strategies involved his subjects in dramatic reconstructions of their own lives could be regarded as "a doorway one could step through imaginatively into the life of another people" (MacDougall, 2005: p. 232). MacDougall traces this 'other' lost vision of visual anthropology, which was later reclaimed, by Jean Rouch and John Marshall. In hindsight, his work has been critiqued for its mythologizing of the 'noble savage' and its elision of the 'real' social and political circumstances of his subjects' lives. Flaherty's work poses questions about the nature of 'truth,' the relation of the documentarist to the 'reality' he/she creatively appropriates, and the politics of representing the other.

The Author as Social Educator – John Grierson

Grierson's extensive writings on the place, function, and codes of the documentary regard it as performing the function of educating and involving citizens in democratic civil society. In 1930, Grierson established the Empire Marketing Board Film Unit, later to become the General Post Office Film Unit. This was a nucleus for the work of documentary filmmakers such as Basil Wright (*The Song of Ceylon*, 1934, *Night Mail*, 1936, codirected with Watt), Harry Watt (*North Sea*, 1938), Alberto Cavalcanti (*Pett and Pott*, 1934, *Coalface*, 1935), and Humphrey Jennings (*A Diary for Timothy*, 1944). The themes handled included the lives of working people; social problems related to health, housing, and unemployment; and wartime propaganda.

The Author as Reflexive Roving Eye – Dziga Vertov

Denis Arkadievitch Kaufman called himself Dziga Vertov, meaning 'spinning top.' His writings and films were produced in the context of the post-Revolution ferment in the Soviet

Union. His work, influenced by Futurism, represents a seminal contribution to film theory and practice, repudiating all pre-existing forms of cinema. Vertov regarded the cinematic way of seeing, the 'Kino-Eye' as establishing a radically different relationship with reality, making the "invisible visible, 'turning' lies into truth, 'propagating' the communist decoding of reality." His work on the Kino Pravda series, among others, and his film *The Man with the Movie Camera* (1929) created much debate and controversy for its formal illusionism, its subjectivism, its radical play with elements of picture and sound, so that a sense of dislocation from reality is achieved. Its reinvention of cinematic language in a mode, which reflexively foregrounds the act of construction and directly interpellates the viewer with revolutionary fervor stands as unique and perhaps far ahead of its time.

The Author as Propagandist – Leni Riefenstahl

Leni Riefenstahl is known for her documentaries that mythologized the politics of German National Socialism. While her work is noteworthy for its aesthetic virtuosity, it raises questions of the politics of aesthetics. *The Triumph of Will* (1935) documents the 1934 National Socialist rally in Nuremberg, a spectacular celebration of the myth of fascism. Sontag (1982) discusses how the actuality of a film like *The Triumph of Will* was constructed through carefully choreographed rituals of domination and enslavement that contrast the hypnotic presence of the omnipotent leader with the faceless regimentation of large masses of ordinary people.

The Author as a Fly on the Wall – Direct Cinema

Direct Cinema represents a significant departure in the American documentary of the 1960s. The term Direct Cinema is sometimes used interchangeably with *cinema verité* (from which, however, it is distinct). It positions the author as an unobtrusive observer, thereby attempting to reclaim the evidentiary project of documentary film. The camera as well as the filmmaker occupies neutral positions, endeavoring, "to convey [to the viewer] the feeling of being there" (Leacock, cited in Winston, 1993: p. 43). The invention of lightweight equipment with synchronized sound was first utilized by Drew Associates to produce *Primary* (1960), on the election campaigns of John F. Kennedy and Hubert Humphrey. Drew collaborated with Richard Leacock, Pennebaker, and Albert–Maysles in the making of this film. The filmmakers had themselves perfected the technology and the now legendary shot of the camera following Kennedy through the crowds on to the podium, shot by Maysles, demonstrated the versatility of the new technology.

The then prevailing dominant empirical traditions of the scientific method, which took sides with the 'body,' in an age old debate on its antithetical position to the 'mind,' was at heart of the positioning that the proponents of Direct Cinema aspired toward. The dichotomies of mind–body and subject–object held center stage, not only in the debates on documentary cinema, but also in the discourses of the social sciences and linguistics.

The truth claims of the proponents of the American Direct Cinema did not pass uncontested. In response to its critics, the movement of Direct Cinema had to allow for subjectivity, while retaining its claim to authenticity. Frederick Wiseman, who terms his work "totally subjective," represents a new denouement in the Direct Cinema debates. While the work of Leacock and Pennebaker focused on portraiture and recording of events, Wiseman's work investigates various institutions such as hospitals, high schools, and prisons, revealing patterns of behavior, which raise disturbing questions about institutional morality and social values. His first film *The Titicut Follies* (1969) was an indictment of an institution for the criminally insane and saw him embroiled in legal controversy. Wiseman does a tightrope walk between the subjectivity of the filmmaker and the imperative to provide the viewer with fair evidence so that he/she can make up his/her own mind on the reality that is represented.

The Author as a 'Fly in the Soup' – Cinema Verité

Henry Breitrose (cited in Winston, 1993: p. 53) uses the interesting analogy of a 'fly in the soup' to describe cinema verité (a term borrowed from Dziga Vertov). In contradistinction to Direct Cinema, cinema verité resolves the question of objectivity by situating the filmmaker as an *agent provocateur*. The work of Jean Rouch and Edgar Morin typifies this stance. Rouch's *Moi, un Noir* (1959) represents a turning point in ethnographic film; in this film, the role of the camera shifts from a recording device to a precipitator of action. The film presents the inhabitants of Treichville, playing out scenes of their choice for the camera. After viewing themselves on screen, they reflect on the recorded events, reinterpreting them. This complex and novel process of representation explored a new relationship among the filmmaker, his subjects, and the narrative. Rouch's project was rooted in an attempt to redefine the relationships of power in ethnographic filmmaking, which had hitherto been a colonial gaze at the other. *Chronique d'une Été* (1961) in a reversal of ethnographic roles, turns the camera on "the strange tribe that lives in Paris" (Morin cited in Winston, 1993: p. 51). Rouch's notions of a "shared anthropology" (Rouch and Feld, 2003) involving dialogue across cultures and 'ethnofiction' subverted the colonial gaze and reshaped the terrain of ethnographic film.

The Postcolonial 'Other' as Author

Indigenous documentary film production in the postcolonial context was sometimes State-sponsored (as in Cuba and India) and often projected the dominant agenda of modernity. Over time, the rise of independent documentary production gave voice to marginal groups and looked critically at larger political questions. The work of Fernando Solanas and Octavio Gettino in Argentina (*Hour of the Furnaces*, 1968), Patricio Guzman (*The Battle of Chile*, 1974–79), Santiago Alvares in Cuba (*Now*, 1965), Nana Mahomo in South Africa (*Last Grave at Dimbaza*, 1975), Anand Patwardhan (*Father, Son and Holy War*, 1998) in India are but a few examples of a vibrant tradition of political engagement. The technology of video and the possibility of broadcast by international television networks have created a market for stories from the South. Many independent documentary filmmakers in the South are dependent on these markets for support. The market demand tends to privilege

images that present the other as exotic, oppressed, or disaster-ridden. Raul Ruiz's *Of Great Events and Ordinary People* (1979) ironically comments on this endless dividing practice that sustains international documentary production and consumption.

The 'linguistic turn' within postcolonial ethnography, which regards the ethnographer as an author constructing narratives, has led to a rethinking of the relationship between the ethnographer and his/her informants. Since this perspective attempts to render apparent the process by which the narrative is produced, the informants in a sense become coauthors. Given the unequal power flows between the ethnographer and his/her informants, various modes of integrating the voices of the other in ethnographic texts have been explored. Some of these have involved the informants taking control of the medium to construct representations of themselves. The Navajo project, initiated in 1966, by Sol Worth and John Adair, was an experiment in teaching the Navajos, in Pine Springs, Arizona, to use the technology of filmmaking. The attempt was to understand the modes of representation that were at work in the narratives produced by the Navajos (Worth and Adair, 1970).

Eric Michaels (1994) explores the space of alternative production and distribution by the Warlpiris. The project, initiated in 1985, involved local production and illegal transmission by the Warlpiri Media Association, at Yuendemu, in Central Australia. Michaels identifies significant differences between aboriginal and European creative practices. These pertain to "ideological sources and access to inspiration, cultural constraints on invention and imagination, epistemological bases for representation and actuality, indistinctness of boundaries between authorship and oeuvre and restrictions on who makes or views expressive acts" (Michaels, 1994: p. 105).

The Reflexive Author

The reflexive mode, already discussed earlier, gained prominence in the 1970s and 1980s. On the one hand, it was a response to the current debates on language and signification, subjectivity and ideology, involving formal strategies that question the "materiality of the cinematic signifier" (Wollen, cited in Nichols, 1991: p. 65). On the other hand, it arose in the context of movements (feminist, civil liberties, gay and lesbian, revolutionary struggles in the Third World, and so on) that questioned the "materiality of social practice" (Wollen, cited in Nichols, 1991: p. 65). The films of Chris Marker (*Sans Soleil*, 1982), Errol Morris (*The Thin Blue Line*, 1987), Michael Moore (*Roger and Me*, 1989), Jill Godmilow (*Far from Poland*, 1984) are few examples of emerging trends. Nichols (1991: p. 65) points out that reflexivity is not merely a matter of formal narrative strategies, but an engagement with the marginal and alternative ways of seeing and acting in the world. Examples of such politically reflexive films are Georges Franju's *Blood of the Beasts* (1949), Emile de Antonio's *In the Year of the Pig* (1969), Wiseman's *High School* (1968), and Joris Ivens' *The Spanish Earth* (1937).

The early independent women's films (*Union Maids*, 1976, *Joyce at 34*, 1972, and *Janie's Janie*, 1971) were situated within the realist tradition, and made as tools for consciousness raising and mobilization (Kaplan, 1982–1983). Semiological film theory tends to dismiss the radical potential of these films, a view that is contested by Kaplan, who argues that "the same realist signifying practices can indeed be used to different ends" (Kaplan, cited in Nichols, 1991: p. 68) and that a consideration of the strategies of reception is also important (*see* British Cultural Studies). Feminist films such as Mitchel Block's *No Lies* (1973), Michelle Citron's *Daughter-Rite* (1978), problematize cinema verité techniques by drawing the viewer into an illusion of reality, only to reveal at the end that the narrative was scripted and the protagonists are actors.

Postfeminist work attempts to rewrite feminist themes to break away from the white/middle class/heterosexual character of the women's movement in the West. Trinh T. Minh-ha's *Surname Viet Given Name Nam* (1989), for instance, contests the systems of representation at work in a patriarchal ethnography, rejecting an essentializing of the heroic, suffering Vietnamese woman.

Contemporary Issues

Debates regarding documentary and ethnographic film have revolved around various sets of relationships: between the author and the narrative (the question of realism), between the author and the subjects filmed (the question of the ethics and politics of representation), and between the narrative and audiences (the question of effects and audience reception). An issue that has received relatively less attention is the question of how mass media institutional dynamics, the economics of documentary production, and the structures of distribution have influenced the politics and aesthetics of documentary production (*see* Media Imperialism; Television Industry; Internet and Culture). For instance, the fact that television constitutes a major market for documentaries tends to set the agenda for documentary film production and its reception. Today, the emergence of new forums for sharing and distribution based on the Internet such as YouTube and Vimeo are changing traditional modes of production and reception of documentaries. These new modes are, in some cases, able to resist regimes of control and censorship imposed by the State and the markets. Crowd funding is a new way of raising resources for the production of independent documentaries. The Internet serves as a platform for innovative production processes that work with found materials to generate documentary narratives, again bringing to focus issues of authorship and intellectual property rights. *Life in a Day* (2011) produced by YouTube constructed from 80 000 submissions of crowd-sourced footage from 192 countries revises traditional notions of authorship and dissemination.

The association of documentary and ethnographic film with scientific observation and evidence has shifted toward an engagement with the affective dimensions of the genre. Rutherford (2011) among others, points to the notion of 'mimetic innervation' (drawing on Walter Benjamin and Michael Taussig) to understand how the documentary can move away from the space of rationality to engage the viewer in a visceral, embodied relationship with the text, in a performative mode.

With new technologies, such as digital image manipulation and inexpensive portable video camcorders that are redefining

the realm of the real, the project of recording reality has never been more problematic. In postdigital, hyperreal cultures, a variety of other genres, ranging from reality programming on television to fiction films, have begun to appropriate the idiom of the 'real,' blurring the divide between fictional and nonfictional narratives. When the technology for appropriating actuality in the Direct Cinema dream became a reality, it is this very actuality that has become more and more elusive, which has made the landscape of documentary and ethnographic film complex, varied, and in a state of eternal reinvention.

See also: Broadcasting: Regulation; Entertainment; Film: Genres and Genre Theory.

Bibliography

Barnouw, E., 1974. Documentary: A History of the Non-fiction Film. Oxford University Press, New York.

Barsam, R.M., 1992. Non-Fiction Film: A Critical History. Indiana University Press, Bloomington.

Hardy, F. (Ed.), 1966. Grierson on Documentary. Faber and Faber, London.

Kaplan, A., 1982–1983. Theories and strategies of the feminist documentary. Millennium Film Journal 12, 44–67.

MacDougall, D., 2005. The Corporeal Image: Film, Ethnography and the Senses. Princeton University Press, Princeton, NJ.

Michaels, E., 1994. Bad Aboriginal Art: Tradition, Media and Technological Horizons. University of Minnesota Press, Minneapolis.

Nichols, B., 1991. Representing Reality: Issues and Concepts in Documentary. Indiana University Press, Bloomington.

Nichols, B., 2010. Introduction to Documentary. Indiana University Press, Bloomington.

Rabiger, M., 1998. Directing the Documentary. Focal Press, Boston.

Renov, M. (Ed.), 1993. Theorizing Documentary. Routledge, London.

Rosen, P., 1993. Document and documentary: on the persistence of historical concepts. In: Renov, M. (Ed.), Theorizing Documentary. Routledge, London.

Rosenthal, A., Corner, J. (Eds.), 2005. New Challenges for Documentary. Manchester University Press, Manchester.

Rouch, J., Feld, S., 2003. Cine-ethnography. University of Minnesota Press, Minneapolis.

Rutherford, A., 2011. What Makes a Film Tick? Cinematic Affect, Materiality and Mimetic Innervation. Peter Lang, Bern.

Sontag, S., 1982. Fascinating fascism. In: Sontag, S. (Ed.), A Susan Sontag Reader. Vintage Books, New York.

Winston, B., 1993. The documentary film as scientific inscription. In: Renov, M. (Ed.), Theorizing Documentary. Routledge, London.

Worth, S., Adair, J., 1970. Navajo filmmakers. American Anthropologist 72 (1), 9–34.

Domestic Violence: Legal Perspectives

Yvana LBH Mols, Emory University School of Law, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The emergence of domestic violence (DV) as a legal issue in the 1960s has had international implications for law and policy. Legal recognition has manifested in various forms: criminal penalties for batterers, restraining orders, availability of tort remedies for survivors, and the attention of international human rights law. Even so, DV remains a pervasive harm worldwide. Courts are limited in their ability to address structural social problems because they are largely focused on individual redress. International treaties focused on human rights may provide individual causes of action and remedies, and express social priorities for peace and equality. Vulnerability theory addresses these structural issues of power and control with the goal of providing substantive equality and eradicating violence.

Introduction

Domestic violence (DV) as a legal concept provides an aspirational and broad, gender-, age-, and sexual orientation-blind framework for addressing private violence and has begun to make visible what has long been invisible. While the term may widely refer to child abuse and other types of family violence, this article will focus on violence between adults in a romantic relationship and use DV to refer to what is commonly known as intimate partner violence (IPV); other types of family violence are treated elsewhere in this volume. In contexts focusing on the relevance of gender in this type of abuse, 'woman abuse' or 'wife abuse' are sometimes used (Dragiewicz, 2011). There is disagreement about the gendered nature of DV, however; some argue there are 'nonpatriarchal' societies, such as the US and other democratic countries, where resort to violence by women is as common as it is for men (Dutton, 2006). The difficulties in defining the problem, identifying the perpetrators, and even deciding upon the language to describe the problem make what is commonly known as DV a constantly shifting target, presenting substantial challenges for the law.

This article is divided into five major Parts. Since violence against women historically has been the more trenchant problem, it will be the focus of this article. Following the introduction, the roots of DV as a private family matter and its rise as a legal issue are discussed. Part two addresses the treatment of DV as a criminal matter beginning with law enforcement, through prosecution, adjudication, and rehabilitative strategies. Part three considers DV as a civil matter, focusing on tort causes of action and remedies such as protective orders. Part four explores the role for international law, especially human rights law, in addressing DV as a global problem. Part five highlights trends and challenges in DV law, including the use of battered women's syndrome, and the problems with language inherent in the field.

History of DV in the Law: From Private Prerogative to Legal Issue

DV, while an ancient phenomenon, did not exist as a legal issue for much of human history (Ferraro, 1996). This was due to a sharp public private divide; public matters were subject to legal regulation, whereas private matters, including violence in the home, were left for informal familial or community rule (Natarajan, 2007). While some nations retain this binary and continue to neglect DV as a legal issue, many nations have revised notions of privacy in favor of protecting victims of DV (Natarajan, 2007).

The Anglo-American experience is illustrative of larger worldwide trends. English common law protected a husband's prerogative and his 'privilege' to punish or chastise his wife. This legal protection was grounded on the marital unity doctrine, which made the husband answerable for his wife's public behavior (Siegel, 1996). America's adoption of the marital unity doctrine also included the implicit condoning of DV, even while it was legally and formally prohibited. In the 1600s, the Massachusetts colony prohibited violence in the pursuit of family preservation, but this was largely unenforced, especially because the violence was perceived as socially permitted wife control (Buzawa and Buzawa, 2003). This justification was dependent on the high value of domestic privacy, which sanctioned or ignored otherwise criminal behavior based on the relationship of the parties (spouses), and the location of violence (the home). When wives failed to conform to the ideal of submissive womanhood, violence was considered an appropriate means of control (Rambo, 2009). Community governance of DV in the United States ended with the rise of the temperance movement in the 1920s. Though there was a rise in legal recognition of DV and a greater concern with holding perpetrators accountable, the temperance movement attributed DV exclusively to alcohol abuse and thus suggested that the elimination of legal access to alcohol would end violent behavior (Rambo, 2009). This framing marked DV as an individual matter of drunken belligerence, rather than a social or community-wide problem.

The social and governmental value of privacy provides additional grounds for the practice of DV in many places. The US Supreme Court, for example, has found the negative right to privacy implicit in the Constitution (Griswold, 1965) and embedded in a Western philosophical tradition that affirms the patriarchal family. Privacy also is considered socially desirable in many places, although some have held that the traditions around privacy obscure that it is inextricably connected to gendered, racial and class-based presumptions (Rambo, 2009).

In Africa, the history of DV as a private matter is inextricably linked with colonial influences. European colonial officials presided over cases that were a threat to public order and thus deemed 'crimes,' while leaving other cases to the jurisdiction of existing native authorities. Defined as a private matter, DV and family disputes were not considered crimes and would be handled by native authorities, if at all (Burrill et al., 2010). The turn to privacy usually means that the state tacitly accepts DV (Schneider, 1991). Thus, even as individual instances of violence are prosecuted, the logic of a right to privacy results in the structural problem of DV remaining invisible.

The formation of a battered women's movement in response to the privacy concern began with grassroots activism in the 1960s and 1970s. It challenged the tacit acceptance of DV as activists contested notions of family privacy, and prioritized citizen safety over another's right to nonintervention in domestic matters. As the movement gained traction, various states adopted legal protections for victims/survivors: criminal penalties for batterers, civil restraining orders, and other equitable measures. As the worldwide movement to eradicate DV grew through the end of twentieth century, many countries supplemented local laws by signing on, sometimes with reservations, to international treaties that defined DV and proposed legal frameworks (Burrill et al., 2010).

Criminal Aspects of DV

DV is largely covered by state and local laws, whether handled as a criminal or civil matter.

Arrest

Many crimes like homicide, assault, criminal contempt, stalking, and rape may have a DV component. However, the DV context may make arrest more difficult for officers, because the physical harm may be low in proportion to the psychological harm to the victim, and because there may be few witnesses to the violence (Buzawa and Buzawa, 2003; Sack, 2004).

The 'mandatory arrest' movement attempted to address the structural gaps in police enforcement of DV crimes. It began following a 1984 Minneapolis study by the Department of Justice, which showed that arrest of the offender was more successful in reducing DV recidivism than the interventions of mediation and removal of the offender for 8 h (Sherman and Berk, 1984). Following this study and the 1984 US Attorney General's report recommending arrest as the standard law enforcement response to DV calls, arrests for these crimes rose dramatically in the US (Sack, 2004). Mandatory arrest still requires police officers to establish probable cause for an arrest, but their discretion as to whether to arrest or mediate the situation has been curtailed: they must make an arrest if probable cause exists (Combs, 2006).

Community plays an important part in the success of mandatory arrest as a deterrent to future battering. For 'good risk' offenders with strong community ties, arrest is likely to deter recidivism due to social stigma, though 'bad risk' offenders with few community ties are unlikely to be deterred, and in fact more likely to escalate battering (Mills, 1998). Mandatory arrest policies may also lead to either dual arrests or the arrest of the victim of violence if an officer is uncertain about who initiated the violence (Sack, 2004). For women in poverty who may be economically dependent on their batterers, mandatory arrest may breed controversy and alienation for both victim and batterer (Brown, 2012).

A study done in 2005 in 10 Latin-American countries, including Belize, Bolivia, Costa Rica, Guatemala, and Peru, found that law enforcement official's attitudes toward DV and women often are based on harmful stereotypes and myths (Sagot, 2005). In many of these countries, these attitudes translate to real obstacles for survivors of DV and survivor mistrust of official channels of arrest and prosecution. Some specific issues are lack of specialized personnel, time-consuming and difficult-to-navigate bureaucratic processes, and inadequate response to emergency situations (Sagot, 2005). Availability of arrest is also complicated by the fact that crimes are defined in these countries retrospectively based on the nature of the injury suffered: an action is a crime if it results in an injury that takes more than 10 days to heal (Sagot, 2005). Psychological, other nonphysical, or fast-healing injuries are classified as misdemeanors and are treated similarly to civil actions (Sagot, 2005). Hostile attitudes, laws, and processes mean that law enforcement institutions are practically inaccessible for many DV survivors (Sagot, 2005).

In response to this problem of access to effective and sympathetic law enforcement, countries in Latin America and in India have developed women's police stations (WPS), staffed by officers specifically trained to address DV cases (Sagot, 2005; Natarajan, 2005). This initiative began in Brazil in 1985, and WPS have become widespread in Latin America with more than 475 now in Brazil (Women Police Stations Case Study, 2011). In India, WPS, called 'All Women Police Units,' were developed as a pilot project in Chennai in 1992, and have now spread across the country, numbering 442 on 1 January 2011, and located in 10 of 23 States and Union Territories (Joshi, 2012). WPS deal mainly with crimes against women and may provide holistic services that include arrest powers, as well as powers to distribute resources, provisions, and counseling (Sagot, 2005).

Prosecution

Prosecutorial reticence is a structural problem in DV cases. Office policies about what cases to prosecute and how to value victories have restrained some prosecutors' interest in pursuing DV cases when compared to cases involving stranger violence. Sometimes prosecutors and judges are simply not well-versed in the law. For example, a group of female judges in Uganda noted that DV was being underenforced due to unawareness of the law, apathy for enforcement of a so-called 'women's issue,' and the failure of the state to provide additional resources for implementation (Goitom, 2013).

Where victims of a crime are not parties to a case, as in the US, the victim does not have control over the decision to file or how the case should proceed once filed. Survivors who do not want their abuser prosecuted or later change their mind about the benefits of prosecution have little power to influence the decision, beyond refusing to give evidence in an attempt to weaken the case and likelihood of success (Cretney and Davis, 1996). While victim testimony is often crucial to prosecuting

crimes, prosecutors often go forward on the basis of physical evidence, police and medical records, and other witness statements, sometimes used to impeach the victim now resisting prosecution. (Goodmark, 2011).

Further curtailing victim control over prosecution and prosecutorial reluctance to pursue DV cases was the institution of 'no-drop prosecution' policies. No-drop policies vary in their coercive power against victims. 'Soft' no-drop policies address victim's concerns about prosecution and facilitate testimony by providing support and explanation, but stop short of actually compelling testimony. On the other hand, 'hard' no-drop policies compel victims to testify; hard no-drop policies thus embody the view that the state has a more compelling interest in controlling the victim than does the abuser (Goodmark, 2011).

The efficacy of nondiscretionary policies is up for debate. Some people claim that they are good for society and for the criminal justice system, because they punish and deter criminals instead of vindicating individual victims' wishes; they facilitate safety for victims by removing them from the coercive power of their attackers; and they empower victims who experience the efficacy of the criminal justice system (Goodmark, 2011). But DV advocates have challenged mandatory arrest and prosecutorial no-drop policies as harmful for victims' safety and autonomy. A Harvard University study found that mandatory arrest made revictimization by one's batterer more likely and increased intimate partner homicides by nearly 60 % (Iyengar, 2007). Taking away police officers' and prosecutors' discretion has the effect of disregarding DV victims' assessments of the situation from a safety perspective, as well as their choices about whether to press charges. These policies may result in a decline in reports of DV. In addition, stereotypes that attach to class and racial presentation may block justice for victims who do not fit a white, middle-class norm and lead to further marginalization of people of color (Brown, 2012). Thus, states' attempts to address the pervasive issues of DV may have the unintended consequence of disempowering the individuals experiencing DV. Even so, mandatory arrest policies remain law in many US states even after the 2005 Violence Against Women Act (VAWA) retracted its mandatory arrest endorsement in favor of merely a 'proarrest' posture (VAWA, 2013).

Doctrinal Changes

Various jurisdictions have attempted to address the low conviction rate for DV crimes by creating separate DV crimes, or by using DV as an aggravating factor. These laws may have different elements, different standards of proof, or harsher penalties when there is a DV aspect to the crime.

But the tendency to view DV within a criminal framework has also spawned a growing population of 'survivor-defendants.' These are individuals who were, at the time of their arrest, survivors of DV, and for whom DV was a contributing factor in their use of illegal violence against their intimate partner (Kraft-Stolar et al., 2011). This has led to the controversial use of 'battered women's syndrome' expert testimony to aid the defendant's defense claim.

Battered women's syndrome evidence is used by criminal defendants who have harmed or killed their abusive partners and seek to claim self-defense. It may also be used as evidence in cases of duress, or to mitigate murder to manslaughter. This syndrome is marked by the battered individual's perception of a predictable cycle of abuse: tension building, battering, and contrition (Walker, 1979). This cycle is used to explain why the victim may not wait until a period of acute battering to attack his or her batterer, as is customarily required for the objective reasonableness prong of many defenses (Walker, 1979). This syndrome, and its use in criminal defense, is controversial. It helps explain the reasonableness of a victim's anticipation of future harm, and the need to engage in self-defense, even when not under acute attack (Goodmark, 2008). It also provides convenient shorthand for describing common features of many DV experiences. However, critics claim that precisely because the experience of DV is multiple, varied, and complex, it does not lend itself to a single normative standard. Particularly excluded are women who do not fit the white, passive, middle-class, heterosexual norm (Brown, 2012). Others critique the syndrome language as pathologizing victims of DV to render their objectively unreasonable perceptions reasonable (Kinports, 2004).

Courts

The systemic failings of DV policies challenge judicial systems around the world. Because the causes and consequences of DV are multifaceted, recourse to traditional judicial remedies is often ineffective for ameliorating abuse, necessitating a holistic approach coordinating the efforts of community resources, law enforcement, prosecutors, and the judiciary (Shaffer, 2004).

It is argued therapeutic courts would provide this holistic support – a forum for coordinating community resources like mental health and substance abuse treatment, and providing continuing review and accountability (Shaffer, 2004). These courts focus more on rehabilitation than adjudication, and on the provision of services rather than on punishment. Cooperation and coordination between stakeholders is essential for successful therapeutic courts, though victim engagement in the prosecutorial process remains a lingering challenge (Labriola et al., 2009). They also provide a venue for efficient sanctions in the event of failure to comply with a court order, since long-term monitoring is built into the structure of the court (Shaffer, 2004).

Countries that trivialize the problem of DV through official's attitudes, if not through formal law, may provide informal judicial processes that exacerbate violence. For instance, in the 10 Latin-American countries studied in 2005, many resort to conciliatory processes that lend themselves to hard-to-enforce, out-of-court agreements (Sagot, 2005). Some are concerned that these agreements may leave the survivor unprotected, and are a violation of survivor's due process rights (Sagot, 2005).

India is also concerned with the traditional use of an informal, male-dominated adjudicatory system. In India, DV is often linked to issues with dowry – the property or valuables given on the occasion of marriage. Though dowry was abolished in 1961, it persists throughout India in every social stratum (Natarajan, 2005). When disputes arise over the amount given in dowry or its timeliness, village elders typically support the husband and provide few resources to the wife except advice to adjust to the situation (Natarajan, 2005). India

addressed the issue of the male-dominated adjudicatory process by creating All Women Police Units, previously referred to as WPS.

DV and US Federal law: The Violence Against Women Act

The US Federal government entered the largely state-regulated problem of DV by passing the VAWA in 1994, reauthorized in 2013 (VAWA, 2013). This bill heightened penalties for sex crimes, mandated restitution for victims, and provided for sexual abuse programs and community-based shelters (Stark, 2007). It created new federal crimes – interstate DV, and interstate violation of a protective order. It also allowed for civil remedies for victims. The controversial 2013 reauthorization expanded tribal courts' jurisdiction to hear DV cases and included protections and services for Lesbian, Gay, Bi-sexual, Transgender, and Queer (LGBTQ) persons (VAWA, 2013). VAWA's constitutionality was tested by the Supreme Court in *US v. Morrison*. While *Morrison* allowed most of the law under Congress' power to regulate interstate commerce, it ruled that the civil remedies provision was unconstitutional (US Morrison, 2000).

Civil Approach to DV

Civil liability for perpetrators of DV remains a significant issue on the state and local level, involving tort and family law, as well as decisions about child custody and issuance of restraining orders. These laws vary widely by jurisdiction, in response to how the problem of DV is framed; this section will explore some US approaches and international counterparts.

Tort

Tort remedies for harms experienced are crucial for DV survivors. DV is quite costly for survivors; some estimates suggesting DV results in over 61 million dollars a year in medical bills alone (Hart and Sussman, 2004). But DV has other, less obvious, consequences as well. Survivors are often economically dependent on batterers, and may be unable to secure adequate housing for themselves and their dependents without assistance (Hart and Sussman, 2004). The types of tort claims available to a DV survivor are dependent on the particular circumstances of the case. Torts against the person, like battery, assault, and intentional infliction of emotional distress, are possible causes of action against the batterer. In addition, there may be grounds for property claims like interference by exclusion, replevin, conversion, and destruction to property, or contract claims like fraud, interference with contractual relations, breach of fiduciary duty and others (Hart and Sussman, 2004). There are several difficulties with these intentional causes of action, however: the batterer may not have sufficient funds to pay the judgment, the insurance company may not cover DV torts, and the statute of limitations for an intentional tort is relatively short (Wriggins, 2001).

Survivors may also have claims against third parties who had a duty to intervene or protect against DV, and who failed by either act or omission. These individuals, if shown to be in a special relationship to the individual experiencing violence, are uniquely positioned and have a duty to provide care. These individuals may include law enforcement, landlords, or employers, if they had reason to know about the dangerous situation (Hart and Sussman, 2004).

Other Civil Actions

DV is also an issue in many divorce cases, and may be considered relevant when dividing property, assigning child custody, and determining spousal and child support obligations. Courts have held that a history of DV may be considered when reallocating to past temporary spousal support what had previously been distributed from a trust account for child support (In re Marriage of MacManus (2010)). The timing of divorce with tort actions creates issues of claim preclusion. Some courts have refused to hear DV tort claims following the conclusion of a no-fault divorce (Drew, 2005). Other courts have allowed later tort causes of action when not based on the same underlying issues as the divorce. The Swedish Supreme Court presumed that joint parental custody is in the best interest of the child and is preferred when courts determine child custody, either following divorce or separation of unmarried parents (NJA, 2006 s 26). DV is an exception to this general rule; if there is a history of violence in the relationship, the court may order sole custody (NJA, 2006 s 26).

DV also has an impact on bankruptcy proceedings because the costs and the claims associated with DV are varied. Domestic support obligations and unliquidated DV tort claims are nondischargeable in bankruptcy, a provision designed to protect the DV survivor from their batterer's bankruptcy filing. When the survivor is the one filing for bankruptcy, however, the US bankruptcy Code's limitations on liquidation and relief can make a relational and financial fresh start difficult (Satz and Brandt, 2007).

Remedies: Restraining Orders

Protective orders were one of the first legal remedies for DV and are now a common and widespread response to the epidemic of DV, available in every state and increasingly worldwide (Runge, 2012). They are intended as preventative measures, although they need to be enforced through fines or imprisonment. Protective orders can take a wide variety of forms: prohibiting contact, ordering counseling, limiting visitation, creating exclusive use of once-shared property, prohibiting firearm possession, and ordering the cessation of stalking (Runge, 2012).

The means of obtaining a protective order differ according to jurisdiction and situation. Protective orders in criminal cases are typically for the period while the case is ongoing, though criminal courts are increasingly given the power to issue permanent injunctions (Buzawa and Buzawa, 2003). Temporary restraining orders can be issued within days of filing a complaint and are typically *ex parte*. Permanent protective orders can be issued following a DV criminal proceeding or a hearing on a temporary order in family court. Violations of these orders carry various consequences. In some jurisdictions, violations are classified as misdemeanors, while in others the violation is the same as contempt of court (Buzawa and Buzawa, 2003).

Protective orders have some serious flaws. Involvement with the legal system may alert the batterer as to the victim's whereabouts, putting the victim at further risk (Kerrick, 2011). In addition, to be fully effective protective orders require institutional coordination and knowledge of the order's existence. In the United States, there is no database of protective orders available for field officers to check and the victim's court document may not be immediately available to the officer at the scene (Buzawa and Buzawa, 2003). Third, protective orders rely on victim initiation and pursuit, which may be challenging given the victim's state of mind and resources (Buzawa and Buzawa, 2003). Finally, the officer may not have any legal obligation to proactively enforce a protective order, as the US Supreme Court found in *Castle Rock* (*Castle Rock v. Gonzalez*).

Aligning more with the jurisprudence of the Inter-American Commission on Human Rights than the US Supreme Court, courts outside the United States have found that states do have positive obligations to protect individuals from DV. The European Court of Human Rights has used international human rights law like the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW) to create positive obligations on the part of states to ensure sufficient criminal justice response to acts of DV (McQuigg, 2011). These obligations may include preventative measures on an operational level to protect individuals from criminal acts (*Osman v. United Kingdom* 87/1997/871/1083). It is unclear the extent of these positive measures when they require a state's monetary commitment, for instance in the provision of alternative housing for victims of DV (McQuigg, 2011).

One persistent issue with protective orders is enforcement beyond their jurisdictional limits. In Australia, a protective order is not automatically enforceable in each state where the survivor travels or lives. However, survivors may register their protective order to ensure its enforceability when he/she moves or takes an extended stay in another state. Interstate registration does not require notice to the respondent party, and therefore works as an additional protective measure (Protective Orders, 2012).

The European Union (EU) sought to address the jurisdictional enforcement gap through a new law for its member states. In 2013, the EU Ministers of Justice passed a Europe-wide protective order for DV survivors (European Commission Memo, 2013). It provides a guarantee that protective orders will remain enforceable throughout the EU regardless of whether the survivor leaves the issuing country. It extends the 2011 Europe-wide criminal protective order enforcement to civil protective orders (European Commission Memo, 2013). A law passed in 2012 in Turkey went one step further; it provides for electronic ankle bracelets for perpetrators and alarm mechanisms in necklaces or mobile phones for petitioners to easily alert authorities about a violation (Zeldin, 2012).

DV and International Human Rights Law

International human rights law recreates some of the challenges for DV that exist in domestic law because both are grounded in the liberal view of the autonomous legal subject (Fineman, 2010). This view of the rights-holding legal subject requires an expansive notion of individual freedoms and thus negative rights – meaning that the state remains out of private life (McQuigg, 2011). In addition, international law is drafted to hold accountable state, not private, actors, further reinforcing the public/private divide (McQuigg, 2011). Human rights law applies to state action, a framing that often excludes DV, committed in private rather than as official action of the state. This has meant lack of state responsibility, intervention, and remediation for victims. Some argue that state neutrality on violence is impossible and action is needed to avoid empowering the abuser (Copelon, 2003). Human rights law has increasingly been used to address DV under alternative theories of state action, and has created positive state obligations in response to systemic allowances for DV that fail to protect rights-holders from nonstate conduct (Hessbruegge, 2005).

Advocates have worked to extend human rights law to apply in the 'private' sphere in different ways: (1) human rights law can be enacted to apply directly against private actors, as in the South African Constitution; (2) domestic law can be drafted to protect human rights, like the antidiscrimination legislation in the UK; (3) courts can apply law in a way that protects human rights, for instance, in line with the European Convention on Human Rights; and (4) state responsibility can be construed to create a positive obligation to protect private individuals, like in the European Court of Human Rights (McQuigg, 2011). For example, after the US Supreme Court held that state law did not entitle the holder of a DV restraining order to police enforcement of that order, the Inter-American Commission on Human Rights found that the United States violated the human rights of the plaintiff by failing to act with due diligence to protect her and her daughters from DV, and violated the state's obligation to provide equal protection (IACHR Report 80/11, 2011). The Commission applied the American Declaration, which recognizes the human rights owed by Organization of American States member states to individuals (American Declaration, 1948). The focus in the remainder of this article will be on violence against women to address the widespread issue of DV.

Convention on the Elimination of all Forms of Discrimination against Women (CEDAW)

The United Nations adopted CEDAW in 1979, and it entered into force in 1981. The Convention originally mentioned 'gender based violence,' though not specifically DV; but it did emphasize equality and the "elimination of prejudices and customary and all other practices which are based on the idea of the inferiority or the superiority of either of the sexes" (Article 5). Violence against women became an explicit part of the Convention in 1992 when the Committee adopted general recommendation 19, listing specific types of DV against women (paragraphs 11, 23). The Committee specifically recommended that states take "appropriate and effective measures" against public and private acts of violence, including the full range of legal measures: "penal sanctions, civil remedies and compensatory provisions … preventative measures, [… and] protective measures" (paragraph 24). Thus states parties have a duty to protect individuals, even within the home. The 188 states

parties to CEDAW are responsible for formerly private violence, including acts and failure to act.

This seemingly wide acceptance of state responsibility to end violence against women is undercut by numerous reservations made by signatory countries (CEDAW, SP/2006/2). Some of these reservations are against core provisions of CEDAW like Article 2, which "condemns discrimination against women in all its forms," even though Article 28 expressly prohibits any reservation incompatible with the object and purpose of the Convention. These reservations keep many states parties from engaging in a revolution of gender relationships that would change the dynamics of control that lead to DV.

CEDAW is an international law that relies on exposure, shaming, and international standards to compel state action (Merry, 2003). But enforcement is possible for states parties who have ratified the optional protocol, which provides individual complaint mechanisms to the Committee (Optional Protocol A/RES/54/4, 15 October 1999). It was adopted by the UN in 1999, and ratified by some states parties to CEDAW. Even without this enforcement arm, however, the cultural and standard-setting work of CEDAW and states reports are important; both allow an international view on the social, economic, and cultural aspects of gender and violence (Merry, 2003).

Other International Laws

Beyond CEDAW, there have been other important international agreements concerning the status of violence against women. For example, the UN established a Special Rapporteur on Violence against Women in 1994 through the Commission on Human Rights, and the Human Rights Council renewed it in 2013. The 2006 report recognized that states sanction 'honor killings' against women who transgress social norms and other types of violence against women by failing to create legislation to prohibit these 'traditional practices' (Due Diligence Standard, 2006). While the due diligence standard has been used for various initiatives in the prevention, protection, and prosecution of violence, the Special Rapporteur noted that more work needs to be done in the area of reparations for victims of violence (Due Diligence Standard, 2006).

The Beijing Platform for Action addresses various issues related to women's equality, including violence, and was one of the outcomes of the Fourth Conference on Women in 1995 (Platform for Action, 1995). This platform recognizes that violence against women is a "manifestation of the historically unequal power relations between men and women," and that this violence should be addressed regardless of its location (Platform for Action, 1995). The Platform recommends steps for various stakeholders to adopt, ranging from the UN itself to various local institutions, all in the service of fulfilling the strategic goals of eliminating violence and studying its causes. In addition to promoting states' adoption of CEDAW, especially general recommendation 19, the Platform also emphasizes the importance of access to justice and effective remedies for victims of violence (Platform for Action, 1995).

While the aspirational language of these international treaties and the global goals of ending violence against women are laudable, commentators note their limitations. Some note that they take for granted a postconflict setting which creates a gap in gender justice for countries in conflict (Scully et al., 2010). Some question whether they are as effective for enforcement as various regional agreements that take into account cultural and religious heterogeneity (Faedi, 2010). This applies especially in light of countries' widespread adoption of, yet failure to enforce, human rights treaties. Regional human rights agreements can punish violations, while still respecting countries' independence and suspicions that international documents may be examples of colonial and Western interference (Faedi, 2010). In some of these countries, men's authority is perceived as natural, while women's rights will violate the community's cultural heritage. Other resistance to international human rights treaties and the aggressive criminal enforcement that they provoke come from postcolonial feminists. Some argue that these treaties focus inordinately on violence against women as wrongs to be addressed instead of a positive focus on women's rights (Kapur, 2013).

Moving Forward, Contested Language; Transcending Gendered Dimensions of domestic Violence law

The discourse surrounding DV and the law has moved from decidedly gendered to aspirationally gender-neutral. This development has been partly due to antifeminist backlash and arguments alleging a violation of equal protection (Booth Hvass, 2002). Antifeminist critics contest woman-centered laws that provide resources for battered women's shelters. Such positions have been criticized as revealing an entitlement to patriarchal privilege, and as a resistance to any definition of violence that does not reinforce patriarchy (Dragiewicz, 2011). Irrespective of this language shift, however, law and policy-making continue to emphasize crime control mechanisms instead of structural changes that would foster greater equality and liberty (Ferraro, 1996).

Legal contests over DV are a significant locus for analyzing changing social norms around gender expectations, stereotypes, and the structure of the family. (Fineman and Mykitiuk, 1994) Legal approaches to DV show a society's relative valuation of promoting nonviolence regardless of its context or location, and preserving privacy and invisibility for domestic arrangements (Dragiewicz, 2011).

See also: Domestic Violence: Sociological Perspectives; Elder Abuse; Intimate Partner Abuse, Applied Research on; Intimate Partner Violence.

Bibliography

Adams, D., Cahn, N.R., Ross, C.J., Meyer, D.D., 2006. Contemporary Family Law, third ed. West.

American Declaration of the Rights and Duties of Man, 1948. Adopted by the Ninth International Conference of American States. Bogota, Columbia.

Booth v. Hvass, 2002. 302 F.3d. 849 (8th Cir, Minn.).

Bradley v. State, 1824. 1 Miss. (1 Walker) 158.

Brown, G., 2012. Ain't I a victim? The intersectionality of race, class, and gender in domestic violence and the courtroom. Cardozo Journal of Law and Gender 19, 147.

Burrill, E.S., Roberts, R.L., Thornberry, E., 2010. Domestic Violence and the Law in Colonial and Postcolonial Africa. Ohio University Press, Athens, OH.

Buzawa, E.S., Buzawa, C.G., 2003. Domestic Violence, the Criminal Justice Response, third ed. Sage Publications, Thousand Oaks, CA.
California Task Force Report, Task Force on Local Criminal Justice Response to Domestic Violence, Cal. Attorney Gen., Keeping the Promise: Victim Safety & Batterer Accountability 2005. http://digitalarchive.oclc.org/request?pid%3Dobjid%3A0000060710 (accessed 13.05.14.).
Castle Rock v. Gonzales, 2005. 545 US 748.
Convention on the Elimination of All Forms of Discrimination against Women, 1979. http://www.un.org/womenwatch/daw/cedaw/text/econvention.htm (accessed 13.05.14.).
Combs, L.A., 2006. Between a rock and a hard place: the legacy of castle Rock v. Gonzalez. Hastings Law Journal 58, 387.
Copelon, R., 2003. International human rights dimensions of intimate violence: another strand in the dialectic of feminist law making. American University of Journal of Gender, Social Policy and Law 11, 865.
Cretney, A., Davis, G., 1996. Prosecuting 'Domestic' assault. Criminal Law Review 162.
Declarations, reservations, objections and notifications of the withdrawal of reservations relating to the Convention on the Elimination of All Forms of Discrimination against Women, 2006. English CEDAW/SP/2006/2. http://www.un.org/womenwatch/daw/cedaw/reservations.htm (accessed 13.05.14.).
Domestic Violence Act of, 2010. Uganda. http://www.ilo.org/wcmsp5/groups/public/-ed_protect/-protrav/-ilo_aids/documents/legaldocument/wcms_172625.pdf (accessed 21.05.14.).
Dragiewicz, M., 2011. Equality with a Vengeance: Men's Rights Groups, Battered Women, and Antifeminist Backlash. Northeastern University Press, Lebanon, NH.
Drew, M., 2005. Lawyer malpractice and domestic violence: are we revictimizing our clients? Family Law Quarterly 29, 7.
Due Diligence Standard, 2006. Special rapporteur on violence against women, the due diligence standard as a tool for the elimination of violence against women, 2006. United Nations Document 35. E/CN.4/2006/61.
Dutton, D., 2006. Rethinking Domestic Violence. University of British Columbia Press, Vancouver, BC.
European Commission Memo, May 2013. EU-wide Protection for victims of domestic violence becomes law. Memo 13 (449), 22. http://europa.eu/rapid/press-release_MEMO-13-449_en.htm. accessed 21.05.14.
Faedi, B., 2010. Domestic violence as a human rights violation: the challenges of a regional human rights approach in africa. In: Burrill, Emily S., Roberts, Richard L., Thornberry, Elizabeth (Eds.), Domestic Violence and the Law in Colonial and Postcolonial Africa. Ohio University Press, Athens, pp. 256–276.
Ferraro, K.J., 1996. The dance of dependency: a genealogy of domestic violence discourse. Hypatia 11 (4), 77–91.
Fineman, M.L.A., 2010. The vulnerable subject and the responsive state. Emory Law Journal 60, 251.
Fineman, M.L.A., Mykitiuk, R., 1994. The Public Nature of Private Violence. Routledge.
Goitom, H., 2013. Uganda: women judges voice concern over domestic violence. Library of Congress, Global Legal Monitor. http://www.loc.gov/lawweb/servlet/lloc_news?disp3_l205403631_text (accessed 21.05.14.).
Goodmark, L., 2008. When is a battered woman not a battered woman? When she fights back. Yale Journal of Law and Feminism 20, 75.
Goodmark, L., 2011. A Troubled Marriage: Domestic Violence and the Legal System. NYU Press, New York.
Griswold v. Connecticut, 1965. 381 US 479, 85 S.Ct. 1678.
Hart, B.J., Sussman, E.A., 2004. Civil tort suits and economic justice for battered women. Victim Advocate. Journal of National Crime Victims Bar Association 4 (3). Spring 2004.
Hessbruegge, J.A., 2005. Human rights violations arising from conduct of non-state actors. Buffalo Human Rights Law Review 11, 21.
IACHR Report 80/11, 2011. Inter American Commission on Human Rights Report 80/11, 2011. Case 12.626 Merits, Jessica Lenahan (Gonzalez) et al. (July 21).
In re Marriage of MacManus, 2010. 182 Cal.App.4th 330.
Iyengar, R., 2007. Does the Certainty of Arrest Reduce Domestic Violence? Evidence from Mandatory and Recommended Arrest Laws. National Bureau of Economic Research, Cambridge, MA.
Joshi, S., Decemer 25, 2012. Only 442 Women Police Stations across India: Police Research Data, in the Hindu. http://www.thehindu.com/todays-paper/tp-national/tp-newdelhi/only-442-women-police-stations-across-india-police-research-data/article4236877.ece (accessed 30.06.14.).
Kapur, R., Februay 5, 2013. The New Sexual Security Regime, in the Hindu. http://www.thehindu.com/todays-paper/tp-opinion/the-new-sexual-security-regime/article4380446.ece (accessed 11.05.14.).
Kerrick, A.P., 2011. Protections available to victims of domestic violence: no contact orders, civil protection orders, and other options. Advocate 54, 32–33.
Kiesel, Hon D., 2007. Domestic Violence: Law, Policy, and Practice. LexisNexis and Reed Elsevier Properties, Inc.
Kinports, K., 2004. So much activity, so little change: a reply to the critics of battered women's self-defense. Saint Louis University Public Law Reviews 23, 155.
Kraft-Stolar, T., Brundige, E., Kalantry, S., Kestenbaum, J.G., 2011. Avon global center for women and justice at Cornell law school, women in prison project (Correctional Association of New York). In: From Protection to Punishment: Post-Conviction Barriers to Justice for Domestic Violence Survivor-defendants in New York State. Avon Global Center for Women and Justice and Dorothea S. Clarke Program in Feminist Jurisprudence.
Labriola, M., Bradley, S., O'Sullivan, C., Renpel, M., Moore, S., December 2009. A National Portrait of domestic violence courts: report Submitted to the Nat'l Inst. of Justice, Center for Court Innovation. http://www.courtinnovation.org/sites/default/files/national_portrait.pdf (accessed 11.05.14.).
Lemon, N.K.D., 2013. Domestic Violence Law, fourth ed. West Academic Publishing.
Mahoney, M.R., 1994. Victimization or oppression? Women's lives, violence, and agency. In: Fineman, M.L.A., Mykitiuk, R. (Eds.), The Public Nature of Private Violence. Routledge, pp. 59–92.
McQuigg, R.J.A., 2011. International Human Rights Law and Domestic Violence. Routledge Press, Oxon.
Merry, S.E., 2003. Constructing a global law: violence against women and the human rights system. Law and Socical Inquiry 28, 941.
Mills, L.G., 1998. Mandatory arrest and prosecution policies for domestic violence: a critical literature review and the case for more research to test victim empowerment approaches. Criminal Justice and Behavior 25, 306.
Natarajan, M., 2005. Women police stations as a dispute processing system: the Tamil Nadu experience in dealing with dowry-related domestic violence. Journal of Women and Criminal Justice 16, 87.
Natarajan, M. (Ed.), 2007. The International Library of Criminology, Criminal Justice & Penology, Second Series: Domestic Violence. Ashgate, Hampshire, UK.
NJA, 2006 s 26. Nytt Juridiskt Arkiv [NJA] [Supreme Court] 2006-02-08 p. 26 T308-05 (includes NJA 2000 p. 345 and NJA 1999 p. 451).
O'Brien, M., 1996. Mutual restraining orders in domestic violence civil cases. Clearinghouse Review 30, 231.
Optional Protocol. A/RES/54/4 15 October, 1999.
Osman v. UK (No. 23452/94), 1998. (VIII Eur. Ct. H.R).
Platform for Action, 1995. Fourth World Conference on Women, Beijing, China – September 1995 Action for Equality, Development and Peace – Platform for Action. United Nations. http://www.un.org/esa/gopher-data/conf/fwcw/off/a-20.en (accessed 12.05.14.).
Protective Orders, 2012. National Council for Single Mothers and Their Children Resource Guide: Protective Orders. http://www.ncsmc.org.au/wsas/legal_system/protection_orders.htm (accessed 21.05.14.).
Rambo, K.S., 2009. "Trivial Complaints": The Role of Privacy in Domestic Violence Law and Activism in the US. Columbia University Press, New York.
Robinson v. Lioi, 2013. 536 F. App'x 340, 341 (4th Cir. 2013) cert. denied, 134 S. Ct. 1515 (U.S. 2014).
Runge, R.R., 2012. An american concept with distinctly Chinese characteristics: the introduction of the civil protection order in China. The North Dakota Law Review 88, 871.
Sack, E.J., 2004. Battered women and the state: the struggle for the future of domestic violence policy. Wisconsin Law Review 2004, 1657.
Sagot, M., 2005. The critical path of women affected by family violence in latin america: case studies from 10 countries. Violence Against Women 11, 1292.
Satz, M., Brandt, E.B., 2007. Representing victims of domestic violence in property distribution proceedings after the bankruptcy abuse prevention and consumer protection act of 2005. Family Law Quarterly 41, 275.
Schneider, E.M., 1991. The violence of privacy. Connecticut Law Review 23 (4).
Schneider, E.M., 2000. Law and violence against women in the family at century's end: the US experience. In: Katz, S., Eekelaar, J., Maclean, M. (Eds.), Cross Currents: Family Law and Policy in the US and England, pp. 471–494. Oxford.
Scully, P., 2010. Finding gendered Justice in the age of human rights. In: Burrill, E.S., Roberts, R.L., Thornberry, E. (Eds.), Domestic Violence and the Law in Colonial and Postcolonial Africa. Ohio University Press, Athens, OH, pp. 277–286.
Scully, P., McCandless, E., Abu-Nimer, M., 2010. Gender violence and gender Justice in peacebuilding and development. Journal of Peacebuilding and Devlopment 5 (3), 1.
Shaffer, Hon C., 2004. Therapeutic domestic violence courts: an efficient approach to adjudication? Seattle University Law Review 27, 981.
Sherman, L.W., Berk, R., 1984. The specific deterrent effects of arrest for domestic assault. American Sociological Review 49, 261.
Siegel, R.B., 1996. "The rule of love": wife beating as prerogative and privacy. Yale Law Jaw 105, 2117.
Stark, E., 2007. Coercive Control: How Men Entrap Women in Personal Life. Oxford University Press, Oxford.

Tobias, C., 1999. The imminent demise of interspousal tort immunity. Montana Law Review 60, 101.
US v. Morrison, 2000. 529 US 598.
Violence Against Women Act Reauthorization Act of 2005. HR-3402.
Violence Against Women Act Reauthorization Act of, 2013. Public Law 127 (4), 113. Stat. 54 (codified in scattered sections of 42 USC).
Walker, L., 1979. The Battered Woman. Harper and Row, New York.
Women Police Stations Case Study, 2011. UN Women, 2011. "Security Sector Module: Women Police Stations in Latin America Case Study: An Entry Point for Stopping Violence and Gaining Access to Justice (Brazil, Peru, Ecuador, and Nicaragua)". http://www.endvawnow.org/uploads/browser/files/security_wps_case_study.pdf (accessed 21.05.14.).
Wriggins, J., 2001. Domestic violence torts. Southern California Law Review 75, 121.
Wriggins, J., 2002. Interspousal tort immunity and insurance 'Family member exclusions' – shared assumptions, relational and liberal feminist challenges. Wisconsin Women's Law Journal 17, 251.
Zeldin, Wendy, 2012. Turkey: Parliament Adopts Law on Prevention of Domestic Violence. Library of Congress, Global Legal Monitor. http://www.loc.gov/lawweb/servlet/lloc_news?disp3_l205403030_text (accessed 21.05.14.).

Domestic Violence: Sociological Perspectives

Rebecca E Dobash and Russell P Dobash, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Domestic violence is characterized by a recent history of rapid social change in institutional policy and practice. The problem is primarily one of men's violence against women and a considerable proportion of women have experienced this type of physical and sexual violence. Physical injuries are often severe and women may suffer from persistent health and emotional problems. When women are murdered, the perpetrator is usually an intimate partner. Women find it difficult to leave a violent relationship because of threats, lack of support, and, historically, the inadequate responses of institutions. Sociological explanations stress the importance of male power and control.

Overview

The issue of 'domestic violence' is an important area of public, political, and academic concern that goes to the heart of the institution of the family and marriage and of gender relations between men and women. It encompasses public and private life as well as social, political, and economic institutions. At the social level, it involves ideological beliefs and institutional policies and practices. At the individual level, it involves personal attitudes and prejudices as well as individual behavior and conventions of daily life. The area is characterized by a history of recent, rapid, social change in institutional policy and practice and is steeped in controversy. There have been political, ideological, and academic controversies about the term itself, the definition of the phenomenon, the nature of the problem, research methods used in its study, findings about the nature and occurrence of the problem, and about the focus and effectiveness of social and legal interventions. Since its rediscovery in Britain and the United States in the early 1970s, virtually every aspect of this issue has been the subject of development and the object of debate not only within the academic community but also across a wide spectrum of society as the media, community activists, academic researchers, and policy makers have taken different positions and entered into sustained and sometimes heated debates. The lively nature of this important social problem warrants further examination of each of the issues touched upon within this characterization, which can only be summarized here.

Most researchers and policy makers agree that the global problem of domestic violence is primarily one of men's violence against women in intimate relationships, and that it has a long and ignoble history. Numerous studies suggest that a considerable proportion of women in a marital or maritallike relationship have experienced domestic violence, that the physical violence is usually associated with other forms of intimidation and control, and that sexual aggression is a significant aspect of many acts of nonlethal and lethal violence against women partners. Men's violence has important consequences for the women they abuse including numerous and repeated physical injuries, sometimes severe and permanent; persistent health problems; and emotional consequences. Women find it very difficult, and sometimes impossible, to leave a violent relationship because of threats of further violence, lack of economic and social support, and insufficient or inadequate responses of statutory and voluntary agencies. When women are murdered, the most likely perpetrator is a male intimate partner or expartner.

Sociological explanations of domestic violence emphasize male privilege, power, and control, and stress the importance of patriarchal beliefs and ideals and the laws, policies, and practices associated with them. Other explanations emphasize individual pathology and/or social traits such as frustration-aggression and the like. Future research will continue to concentrate on efforts to explain domestic violence and the predicament of the women who experience it, as well as issues relating to the impact of social and legal interventions relating to the safety of victims and those concerning the nature and extent of men's violence and the effectiveness of various responses to abusers.

Definition

The terms wife beating, battered wives, wife abuse, woman abuse, spouse abuse, family violence, domestic violence, and intimate partner violence have all been used to describe this phenomenon. The terms and the conceptions they reflect have all been topics of debate and have evolved over time. The initial terminology reflected the notion of husbands and wives in a legally sanctioned marital relationship but quickly evolved to include non-state sanctioned cohabiting relationships between men and women as well as same-sex intimate relationships. The various terms also reflect different conceptions of the gender of the usual perpetrators and victims of the violence: men's violence against women, mutual violence between men and women, or women's violence against men. These and other issues were the subject of considerable debate when the problem was 'rediscovered' in the 1970s and continue to be debated to the present. While the initial terms of 'wife beating' and 'wife abuse' contained a clear imagery of men's violence against a woman partner, they also contained notions of formal marriage as a necessary condition of concern and attention. By contrast, the term 'domestic violence' included the notion that the problem occurred in both state and non-state sanctioned relationships but lost the conception of gender asymmetry regarding who was most likely to be the victim and who the

perpetrate of the violence. For reasons of popular convention rather than resolution of these debates, the term 'domestic violence' became the term in common usage throughout most of the world.

History

Very ancient customs, beliefs, and laws dating back to the early Roman Empire and many of the first conventions of marriage allowed a husband to kill his wife for infidelity and serious challenges to his male authority. These 'rights' were slowly eroded over time through the development of Canon law and the gradual alteration of legal systems in Europe and Britain to punishments allowing chastisement that did not kill or maim. In 1765, Sir William Blackstone's *Commentaries on the Laws of England* noted that conventional cultural beliefs and common legal practices endorsed a man's right to use 'moderate' chastisement against his wife. Blackstone's Commentaries codified this patriarchal right and thus confirmed a husband's authority over his wife for a considerable period of time. Such ideals and legal views crossed the Atlantic with some of the pioneers who brought familiar social attitudes and legal systems to the newly forming nations of North America. While there were some departures from this position, it nonetheless constituted the normative position in many European and North American cultures for a fairly long period of time. Indeed, contemporary evidence from a wide range of societies reveals strong cultural beliefs expressing such ideals, and these are often endorsed by legal practices that illustrate direct or indirect support for 'wife beating' or, at best, weak reactions to its victims and perpetrators.

In European and North American societies, the history of challenges to and changes in social and institutional responses to the problem is punctuated by three historical periods in which domestic violence was discovered anew and obtained widespread recognition as a social problem. Recognition eventually resulted in changes in the laws regulating the institution of marriage and the family as well as developments in the responses of agencies of the state, particularly law and law enforcement, but also in economic, political, and social institutions.

In the 1870s, 1910s, and 1970s, the problem of 'wife beating' was repeatedly discovered by women in Britain and North America who were engaged in more general campaigns for changes in women's marital, economic, political, and/or social status. Within each time period, wider campaigns about changes within the family and marriage (such as, the ability of women to own their own property or to obtain a divorce for excessive cruelty – 1870s), about changes in the political status of women (the right to vote – early 1900s), about changes in the economic status of women (equal pay and equal opportunities within the job market – 1970s), and changes in social relations between men and women in general also spawned an interest in the specific issue of physical violence against women within the home, usually by intimate male partners/husbands within the context of marriage. At each of these junctures, community groups (usually although not always feminist) were formed, there were public protests, public recognition was increased, and the agencies of the state were challenged and responded with public hearings, legislation, and various new policies and practices. After the first two periods of recognition and public response, the issue was soon forgotten and marked by a return to an older, established status quo in which men maintained social, political, and economic authority over women throughout society and particularly within the institution of family and marriage where this right could be exercised legally and morally through the use of violence against women/wives for various real or perceived violations of male authority.

The third period of challenge began in the 1970s, again in Britain and the United States and, rather than being forgotten after a brief period of interest, has instead continued to expand and develop throughout much of the world. Public awareness of the issue and community and state responses to it are now global, resulting in violence against women being defined by the United Nations as an issue of human rights and social justice. At present, many, although certainly not all, societies have experienced challenges to the complacency and social and institutional tolerance that once prevailed and in many countries new laws and institutional practices have been initiated.

The Nature of the Problem, Research Methods, and Findings

Controversies about the nature of the problem itself have been intense and are intertwined in other controversies about research methods and resulting findings. The debates regarding these issues were apparent at the onset of the rediscovery process in the 1970s and have altered little since, although the amount and type of research grew at a tremendous rate in the last quarter of the twentieth century. As stated above, debates about the phenomenon itself center on whether the violence is asymmetrical in nature (men as the most usual perpetrators with women the usual victims) or symmetrical (women and men as equally violent toward one another), and they sometimes even extend to the notion that women are the most likely perpetrators of domestic violence and men the usual victims of women's violence. Debates about research methods have focused on whether one method should supersede all others in producing the 'best' or indeed the only findings of value in answering the question of 'who is violent to whom.' One camp has claimed variously that only surveys based on national probability samples produce valid and reliable findings. Research using this approach has focused on questions of the prevalence of this form of violence in the population and on the relative levels of violence of men and women. Others defend the use of a wide variety of research strategies, which include historical studies showing continuities and changes in patriarchal beliefs and institutional practices over time and ethnographic studies with an intensive focus upon cultural practices as well as the more traditional sociological approaches of victim surveys, criminal statistics, intensive interview studies, studies of strategic sites, and evaluations of new interventions. Proponents of the broad-based sociological and contextual approaches claim that such studies address more pressing and relevant questions about the problem such as 'what is the exact nature of the violence,' 'how are victims and offenders processed by agencies of the state,' and 'what

works at stopping violence.' Such investigations also attempt to examine in detail the contextual dynamics of violent episodes and violent relationships as well as the meaning and consequences of being a victim and/or perpetrator.

Some survey research, particularly in the United States, has relied primarily on a single measurement tool that counts discrete acts (e.g., a shove, a slap, a punch, or a kick) with little or no assessment of issues, such as intentions (e.g., physical harm, intimidation and fear, or retaliation of self defense), the consequences of the acts (e.g., a bruise, a cut, a burn, or a broken bone), or the contexts in which the violence occurs. Counting one shove the same as one punch or one kick appears to support the notion of symmetry in the perpetration of domestic violence, with women deemed to be equally or even more violent than their male partners. With the claim of symmetry in domestic violence comes the notions that theoretical accounts of domestic violence should be gender neutral and that gender is not related to violence within intimate relationships. By contrast, other surveys and more intensive research methods that are not based on this purely act-based measurement find the opposite pattern, that men are the most likely perpetrators and women the most likely victims of domestic violence and thus conclude that domestic violence is asymmetrical and related to gender. Predictably, these contradictory findings have resulted in considerable controversy with claims and counterclaims regarding the validity and reliability of each type of measurement and resulting skepticism regarding the veracity of findings produced by this single measurement tool.

Over time, controversy about the simple question of 'symmetry–asymmetry' – who perpetrates the majority of violent acts (men or women) – has become more complex with the addition of further research questions regarding who is most likely to be responding in self-defense, to be harmed, to live in fear, and to experience other consequences of violence as well as additional research efforts to differentiate between violent acts, threats, intimidation, verbal aggression, arguments, and the like. This evidence reveals strong and consistent patterns across cultures and historical periods showing that the problem of domestic violence is primarily one of men's violence against women that often results in injuries and sometimes ends in the death of the victimized woman and/or her children. Survey research across several countries indicates that anywhere from one-quarter to one-half of women who have lived in intimate relationships with men report at least one incident of violence, and many women report repeated violence. Across countries, anywhere from 10 to 50% of women also report sexual violence from intimate male partners ranging from forced sexual relations to rape. In addition, homicide statistics worldwide suggest that women are more at risk of being killed by a male partner than by anyone else. Every year, in Britain and North America between 40 and 50% of all women who are murdered are killed by an intimate partner or expartner. By contrast, very few men who are murdered are killed by a woman partner. Some of the risk factors associated with lethal and nonlethal domestic violence include economic and educational disadvantage, youth, and cohabiting versus marital relationships. Situational risks include high levels of conflict, jealousy, sexual possessiveness among male partners, and alcohol abuse. Risk of serious nonlethal and lethal violence is also associated with separation and women's attempts to leave violent relationships.

Institutional Policies and Practices

Early investigations of policy and institutional practices in the 1970s revealed considerable failure on the part of legal and social agencies to provide support and protection for women victims and their children as well as a reluctance to deal with male perpetrators. Recent research charts the many changes in policy, practices, and innovations aimed at ending men's violence and providing safety and security to women and children. Initial focus was on the police and the long-established practice of either ignoring domestic violence or treating it as a nuisance, not 'real' violence, and not 'real' police work. In the United States, several high-profile class actions against police departments were brought by feminist advocates on behalf of women who sought but did not receive police intervention after being beaten by their husband or male partner. These and other efforts elsewhere led to early changes within the criminal justice systems of several countries. Numerous initiatives included policies to strengthen arrest, civil protection orders, and exclusion orders as well as the introduction of dedicated domestic violence courts. Innovative sentencing and probation orders in some locations also included participation on the newly created programs for violent abusers such as those pioneered in Duluth, Minnesota, and later introduced in countries throughout the world.

For the most part, abuser programs are based on feminist principles and use a cognitive-behavioral approach incorporating the notion that attitudes and behaviors are learned and culturally based and, as such, can be modified and 'relearned' albeit with considerable effort on the part of the abuser. Some abuser programs are voluntary with men presenting themselves for a variety of personal reasons, often relating to the woman's threats to leave them, while other programs are tied to the justice system as conditions of a sentence after conviction for domestic violence. Here again, there has been considerable controversy as many women's groups running shelters for abused women object to abuser programs on the grounds that they take funding away from efforts to support abused women, that they may treat abusers with too much leniency, and in the belief that they do not work in stopping the violence. The controversy was probably at its most intense throughout the 1990s and vestiges still remain today, although most shelter groups and abuser programs now work cooperatively in the context of a coordinated community response. These include a combination of services for abused women and their children as well as various programs and criminal justice responses aimed at ending the violence of abusers.

Theoretical Accounts

Theoretical explanations of domestic violence predictably include a plethora of competing, at times contradictory, accounts, ranging from the sociocultural to the individualistic. Sociological, usually feminist, explanatory frameworks stress the historical and contemporary existence of male violence

across almost all societies and locate its perpetration within patriarchal ideals and practices, emphasizing male power and control. In these accounts, violence is seen as functional, a resource used in men's attempts to enforce and maintain their domination and control, with conflicts associated with male jealousy, household labor, and economic resources constituting the contexts of violent events. Other sociologically based gender-neutral accounts emphasize power but note that this social attribute is not skewed toward either gender – men and women possess significant amounts of power – and the symmetrical violence that occurs is an expression of such power. Variants of 'gender-neutral' explanations of domestic violence emphasize the importance of individual psychology and/or social traits that are deemed to have similar effects on the violence of men and women alike; these includes stress, frustration, socialization into (sub)cultures of violence, poor anger control, and impoverished social skills. Other-individualistic accounts emphasize the pathological personalities of violent men while still others attempt to understand the predicament and actions of women who live with violent men. Evolutionary psychology has challenged the foundations of these sociocultural and individualistic explanations and, instead, locates men's violence in a universally evolved male psyche involving male sexual proprietariness toward women. Accordingly, this evolved psyche leads men to attempt to possess and control women through various means including violence.

Sociological accounts emphasize widespread social and cultural beliefs as well as institutional policies and practices through which men are privileged in their relationships to women in general and in their relationships with intimate partners in particular. At the interpersonal level, this includes a range of intimidating and controlling acts as well as concerted actions to isolate women from family and friends. Continuous forms of intimidation and isolation make it very difficult, if not impossible, for women to seek help, although it is now clear that most women do engage in active efforts to deal with the violence. Most seek help from family and friends and many also seek assistance from various voluntary organizations such as women's shelters and agencies of the state such as social work and the police. Despite these efforts, women find it difficult to leave the violent relationship. Sociocultural explanations of women's inability to leave a violent relationship stress men's threats of further violence, women's lack of economic and social support, a moral order emphasizing women's obligations to maintain family unity, and inadequate responses of agencies of the state. Other individualistic accounts, particularly in the United States, suggest that personality disorders in women are the primary reason that they are unable to leave violent relationships. According to such accounts, women do not leave violent relationships because they suffer from conditions such as masochism, relationship addiction, 'battered woman'syndrome, and learned helplessness.

By the 1990s and into the 2000s, the enhanced and refined nature of the research questions, along with an increasing concern to evaluate the relative effectiveness of new innovations have further expanded the theoretical questions and the relevant research strategies in this area. Future research in the area will continue to concentrate on efforts to explain domestic violence within the broader social and cultural contexts in which it occurs as well as examine the behavior, motivations and intentions of the men who perpetrate this violence and the predicament of the women who experience it.

The focus on social institutions will continue to address questions about the impact and effectiveness of various social and legal interventions and innovations concerning the welfare and safety of victims and the violence of male abusers. While much of the earliest research was shaped within sociology, women's studies, history and law, the global expansion of interest in this social problem has been mirrored by an equal expansion of disciplines contributing to its study as psychology, criminology, health care, and medicine have variously contributed to advances in knowledge about this significant social problem.

See also: Family Law in the United States; Intimate Partner Abuse, Applied Research On; Rape and Sexual Coercion; Violence and Media; Violence in Anthropology; Violence: Public; War, Political Violence, and Effects on Children.

Bibliography

Campbell, J.C., Glass, N., Sharps, P.W., Laughon, K., Bloom, T., 2007. Intimate partner homicide: review and implications of research and policy. Trauma, Violence & Abuse 8, 246–269.

Dobash, R.E., Dobash, R.P., 1979. Violence against Wives. Free Press, New York.

Dobash, R.E., Dobash, R.P., 1992. Women, Violence and Social Change. Routledge, London and New York.

Dobash, R.E., Dobash, R.P., 2011. What were they thinking? Men who murder an intimate partner. Violence Against Women 17, 111–134.

Dobash, R.E., Dobash, R.P., Cavanagh, K., 2009. Out of the blue: men who murder an intimate partner. Feminist Criminology 4, 194–225.

Dobash, R.E., Dobash, R.P., Cavanagh, K., Lewis, R., 2000. Changing Violent Men. Sage, Thousand Oaks, CA.

Dobash, R.P., Dobash, R.E., Wilson, M., Daly, M., 1992. The myth of sexual symmetry in marital violence. Social Problems 39, 71–91.

Garcia-Moreno, C., Jensen, H., Ellsberg, M., Heise, L., Watt, C., 2006. Prevalence of intimate partner violence. Findings from the World Health Organisation multi-country study on women's health and domestic violence. Lancet 368, 1260–1269.

Gondolf, E.W., 2012. The Future of Batterer Programs: Reassessing Evidence-based Practice. Northeastern University Press, Boston, MA.

Johnson, H., Ollus, N., Nevala, S., 2008. Violence against Women an International Perspective. Springer, New York.

Kimmel, M.S., 2002. 'Gender symmetry' in domestic violence: a substantive and methodological research review. Violence Against Women 8, 1332–1362.

Pleck, E., 2004. Domestic Tyranny. University of Illinois Press, Chicago, IL (1st pub. 1987).

Romkens, R., 1997. Prevalence of wife abuse in the Netherlands. Journal of Interpersonal Violence 12, 99–125.

Tjaden, P., Thoennes, N., 2000. Prevalence and consequences of male-to-female and female-to-male intimate partner violence as measured by the National Violence against Women Survey. Violence Against Women 62, 142–161.

Wilson, M., Daly, D., 1998. Lethal and non-lethal violence against wives and the evolutionary psychology of male sexual proprietariness. In: Dobash, R.E., Dobash, R.P. (Eds.), Rethinking Violence against Women. Sage, Thousand Oaks, CA, pp. 199–230.

Dominant Designs

Aleksios Gotsopoulos, Sungkyunkwan University, Jongro-gu, South Korea

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of the dominant (technological) design occupies a central position in the literature on technological cycles and technological evolution, entry timing advantages, and firm performance and survival. The emergence of a dominant design constitutes a watershed moment in the evolution of technologies and associated industries, as it signals the resolution of technological uncertainty and the maturation of the industry. This is generally accompanied by a change in the nature of innovation (from product to process innovation), a shift in the locus of competition, a shakeout in the number of firms in the industry, and mass adoption on the consumer side.

The dominant design is a single architecture that establishes dominance in a product class (Abernathy, 1978; Anderson and Tushman, 1990). Its emergence is generally seen as demarcating the maturation of an industry or product class. The concept of the dominant (technological) design occupies a central position in the literature across diverse disciplines that deal with technological evolution and the organization of markets. It has been linked to shifting industry structures, to firm and product performance, to firms' entry and exit rates, to time-of-entry advantages, to shifts in innovation types, and to industrial and organizational evolution (Tushman and Murmann, 1998).

Foundations: Technology Cycles

The concept of the dominant design can be traced back to the work of Abernathy and Utterback (Utterback and Abernathy, 1975; Abernathy, 1978; Abernathy and Utterback, 1978). Studying technological evolution, their work and subsequent research have noted that such evolution is not random, but follows a cyclical pattern of variation, selection, and retention. Technological cycles start with stochastic, discontinuous technological breakthroughs that lead to a shift in the technological paradigm (Anderson and Tushman, 1990). As the new technology is initially poorly understood, such technological discontinuities initiate periods of variation (or eras of ferment), during which different producers introduce several and often incompatible variants of the breakthrough technology. The period of variation is followed by a period of selection: competing designs are evaluated (through use and often by the end users) and some might be developed further; eventually, some designs are retained and some others are dropped. This selection process culminates in the emergence of one dominant design that establishes dominance in the product class and that is copied by the majority of producers in the industry. After the emergence of a dominant design, further variation is rare in the core components of the product architecture (and generally associated with niche markets), until a new technological discontinuity occurs to render the old dominant design obsolete and restart the technology cycle.

When and How Does a Dominant Design Emerge?

While the concept of the dominant design is rather straightforward, important questions remain with regard to exactly when and how a dominant design actually emerges. In other words, when can it be concluded that a dominant design has emerged, and what mechanisms drive the selection process among competing designs?

One approach to determining a dominant design's emergence is to track competing designs' market share. When one design accounts for over 50% of new product sales or new process installations over a significant duration, it can be concluded that it has achieved dominance (Anderson and Tushman, 1990). This threshold of 50% market share and the definition of a 'significant duration' (4 years for Anderson and Tushman, 1990) can, of course, be arbitrary and case specific. They are, however, data driven, relatively easy to apply, and generally in line with notions about dominance, and have thus been used extensively.

A somewhat different approach has been followed by scholars, which sees products as nested hierarchies of subsystems and linking mechanisms (Henderson and Clark, 1990; Murmann and Frenken, 2006). Some of these subsystems and linking mechanisms are core components of the end product (e.g., the engine in an automobile) in the sense that they largely determine how the product functions and might dictate adjustments in other components and the overall product architecture. In contrast, other components are peripheral and of less critical importance; changes in them might be dictated by changes in core components but do not trigger changes in the latter. A dominant design for the end product can be said to have emerged when products converge on their core components. After this, further advances occur in peripheral components without changing or challenging the dominant design.

But what really are the mechanisms behind this convergence on one dominant design? Scale economies and network externalities have been offered as two possible explanations (Klepper, 1997). Indeed, the concurrent existence of a multitude of designs does not allow producers to reap scale benefits, whereas it creates confusion among the consumers. In contrast, the emergence of one design as dominant allows manufacturers to optimize processes, become more efficient, and have

more stable relations with suppliers, vendors, and customers (Anderson and Tushman, 1990). On the consumer side, scale in production leads to significantly lower product costs and thus prices, and reduces uncertainty and potential problems of incompatibility between designs.

Convergence on one dominant design then confers many benefits for the ecosystem that forms around an industry. This fact per se, however, does not explicate the exact mechanisms that drive selection among competing designs. If scale economies and network externalities alone suffice to drive the selection process, even small random shocks that give one design a head start can determine which design eventually dominates. Other scholars, however, have focused more on the role of agency and sociopolitical factors in this process. A dominant design can emerge from the deliberate actions of powerful actors such as a dominant producer (e.g., IBM's successful efforts to impose its chosen standard in PCs as the industry standard to be followed by the majority of other manufacturers as well), an important consumer (e.g., the government), or industry standard-setting bodies. Similarly, the dominant design can be the result of firms strategic maneuvering and forging of alliances (Cusumano et al., 1992). Highlighting the importance of such sociopolitical and institutional factors, Suarez and Utterback (1995) note that scale economies are the result rather than the driver of a dominant design's emergence.

In general, technological superiority and early market share alone appear not to be the most crucial determinants of the selection process. In the home videocassette recorder market, for example, Betamax is generally believed to have been technologically superior to VHS, yet was quickly sidelined by the latter (Cusumano et al., 1992). In fact, it is widely accepted that the design that achieves dominance is seldom the most technologically advanced one; while achieving dominance certainly requires good performance for the task at hand, social, political, and organizational dynamics appear to play an even more important role in driving the process of design selection. The problem at such early stages of a technology cycle is that the functional implications of different designs are not well known or understood, and the metrics on which performance will be evaluated are not yet well defined (Clark, 1985). Users are unclear about what the new product should accomplish and what performance trade-offs might be necessary or appropriate (for example, in early computed tomography scanners, doctors were not clear on the relative priorities of scan time, resolution, safety, and cost (Yoxen, 1987; Tushman and Rosenkopf, 1992)).

Initially, then, even the comparison of different designs is difficult, as it requires reaching consensus on the relevant dimensions to be evaluated. Negotiation is required between producers and other stakeholders, such as users and critics. Different producers might emphasize different functionalities of the new technology, depending on their background and competitive strengths. At the same time, end users also create their own understandings of the new technology that might or might not agree with producers' understandings (Clark, 1985). Users might come up with new and unforeseen uses for the product (von Hippel, 1994) or they might demand that further technological evolution take a certain path.

Negotiation is also required within producers and across different scientific disciplines. Most new technological designs involve interactions between engineers from diverse scientific backgrounds. The production of jet engines, for example, requires that aerodynamic, metallurgical, combustion, and mechanical engineers come together and agree on the key features of the new product (Tushman and Rosenkopf, 1992). Different disciplines tend to have different norms and prioritize different aspects of the new technology or different performance metrics. Rather than fully satisfying the concerns of any group of stakeholders, any new design rather amounts to a compromise that is a satisficing solution for all.

The social construction of norms and evaluation metrics has led scholars to recognize that the selection of a dominant design among competing designs is primarily a social/political process rather than one driven by technological determinism. The need to mutually agree not only on evaluation metrics, but also on trade-offs between performance, reliability, and price that affect the market penetration of any design usually cause the dominant design not to be the most technologically advanced on any metric. The design that achieves dominance is generally not customized to the needs of any specific subset of customers or other stakeholders (i.e., it does not maximize performance on any specific dimension), but rather provides a satisficing solution across most groups of stakeholders and thus facilitates mass adoption. For the same reason, the design that achieves dominance is generally not the most expensive option (a characteristic of cutting-edge options), and is reliable enough to be adopted even by users who are still not very familiar with the new technology. The dominant design thus often amounts to the best compromise among competing designs and associated interests. Rather than being a fully new design, it emerges through a synthesis of fragmented technological innovations that have been introduced in designs earlier in the technology cycle, and have already been tested, potentially modified, and evaluated by users and other stakeholders (Suarez and Utterback, 1995).

The fact that a dominant design emerges through sociopolitical/institutional dynamics rather than through a selection process of technological determinism also implies that it is not possible to know a priori which design will eventually become dominant. While important stakeholders like large manufacturers can influence the selection process and sometimes even impose their preferred design as dominant (e.g., IBM's PC), betting on any one specific design is risky for any firm, as it can easily find itself locked in a design that fails to achieve dominance and is bound to disappear.

Boundary Conditions

Although a dominant design emerges in most product classes, there are also cases where no design achieves dominance. The conditions under which this might occur have been studied primarily by industrial economists (e.g., Katz and Shapiro, 1985; Arthur, 1989). Excess momentum and insufficient friction are two such conditions under which, instead of stabilizing around a dominant design, a market experiences constant change as consumers rush to adopt newer versions of the technology, and the market share leadership of any specific design is short-lived. At the other end of the spectrum, excess inertia can lock consumers in an older technology, thus

delaying the evolutionary cycle of a new breakthrough technology and the associated dominant design, or lock them in the design that they first adopt, thus delaying the selection process and allowing multiple designs to coexist. Multiple competing designs can also coexist when network effects are weak or nonexistent, or when incompatibilities among designs do not affect consumers severely (e.g., the incompatibility of different smartphone operating systems does not inhibit calling and other modes of communication, whereas the low cost of app development facilitates their simultaneous development for multiple operating systems). They can also coexist when preferences are as different as to render a compromise on one dominant design difficult. In this case, the multiple designs that cater to the disparate needs of multiple groups of stakeholders are likely to evolve into separate, parallel markets.

Whether a dominant design emerges and the degree to which its emergence hinges upon technological factors or sociopolitical dynamics also depend on the type of the product. Technology management scholars have differentiated between nonassembled products, simple assembled products, closed systems, and open systems, which are correspondingly characterized by increased complexity in the number and type of subcomponents and linking mechanisms (Tushman and Rosenkopf, 1992). In the case of nonassembled products, the dominant design emerges predominantly from a technological logic. For products such as cement or glass, new technologies can be rather easily evaluated in terms of speed and cost. In contrast, for assembled products performance metrics are more convoluted and no design is usually overwhelmingly superior. Under such conditions of ambiguity, sociopolitical processes play a more important role in leading to a compromise on a dominant design. As complexity further increases from simple assembled products to closed systems to open systems, so does the role of sociopolitical forces in the selection of a dominant design, as more numerous and diverse stakeholders need to reach a shared understanding on the norms and the performance metrics that are to be used to evaluate the new technology.

Further Implications of the Dominant Design Concept

Technology Cycles

The emergence of a dominant design constitutes an important event in the evolutionary cycle of a technology. Just as a discontinuous innovation starts the cycle, the emergence of a dominant design marks the industry's maturation. This is accompanied by a significant shift in the nature of innovation, which becomes incremental and focused on the peripheral rather than the core components of the design. The relatively stable product design also facilitates faster innovation in terms of process and integration that aim predominantly not at changes in the product but rather in the production process, and lead to lower costs. At this stage of a technology's evolutionary cycle, rather than attempting to produce new and superior designs, firms focus on increasing the efficiency of their production methods and achieving scale economies (Murmann and Frenken, 2006).

The emergence of a single dominant design also reduces uncertainty significantly on the side of both the producers and the end users. Before a dominant design is established, firms run the risk that their investments into any specific design will be made obsolete, if that design does not become dominant. The emergence of a dominant design largely mitigates such risks and facilitates more stable and reliable relations with suppliers, distributors, and customers (Anderson and Tushman, 1990). Similarly on the customer side, the initial coexistence of multiple designs can create confusion, whereas the risk of being locked into a design that becomes peripheral can hinder adoption especially when switching costs are high, network effects are strong, and complementary products are important. The emergence of a dominant design resolves these problems and facilitates mass adoption that leads to a takeoff in industry sales (Agarwal and Bayus, 2002).

Dominant Design and Entry Timing Advantages

Timing-of-entry and its effects on firm performance and survival occupy a central position across diverse disciplines such as industrial economics, strategy, and organizational theory. Early work on the topic focused predominantly on the advantages of early entry (Lieberman and Montgomery, 1988). Subsequent scholarship has highlighted more the significant risks that are also associated with very early entry (Lieberman and Montgomery, 1998). Bringing both these lines of research together, Christensen et al. (1998) proposed that entry during a window of opportunity that occurs right before the emergence of a dominant design maximizes firms' survival chances.

Defining the ending point of the opportunity window as coinciding with the emergence of a dominant design is relatively straightforward. The dominant design signals the maturation of the industry and shifts the locus of competition from product innovation to cost efficiency and scale economies. Firms that enter after the dominant design has emerged are seriously handicapped vs larger incumbents and thus suffer from permanently higher failure rates. Such scale advantages are less pronounced before the emergence of a dominant design as investments in process optimization only make sense after the process of design selection has concluded. However, firms that enter too early have to deal with heightened technological uncertainty and risk investing in designs that fail to become dominant. As such technological uncertainty is gradually resolved culminating in the emergence of a dominant design, firms can maximize their survival chances when they enter just prior to the dominant design's emergence and thus optimally balance reduced uncertainty with the chance to develop scale and know-how in preparation for the mature stage of the industry.

The exact consequences of and remedies for choosing the wrong design early on have been debated by scholars. Consensus generally exists that sticking to a failing design can lead a firm to fail as well or at best become a peripheral player. Switching from a failing design to the dominant one is possible (Tegarden et al., 1999), but can be risky as it often requires investment in design-specific production capacity and capabilities. Another option for early entrants is that they invest concurrently in multiple designs. This allows them to increase

the probability of guessing correctly the eventual dominant design and gaining an early foothold in it, and can also make them more flexible and faster to change course, if their initial guesses are proven wrong (Sorenson, 2000). In either case, the fact that a dominant design can only be known *post facto* creates significant challenges for early entrants who have been found to suffer from permanently higher failure rates in comparison to firms that enter during the window of opportunity that occurs just prior to the dominant design's emergence and when more concrete information is available on which design will emerge triumphant.

Social, Cognitive, and Political Factors in Technology Evolution

Even though the importance of social, cognitive, and political factors in influencing technological evolution has been recognized even in the very early research on the matter, few scholars had studied these factors in detail (an early exception is Clark, 1985). Some recent studies have attempted to fill in this gap. In their recent study of the digital camera industry, for example, Benner and Tripsas (2012) use the notion of frames to study the origin of technological designs in new industries. As they find, a producers' background and prior experience bear significant influence on how they understand the new industry and what kind of designs they introduce. Under conditions of heightened uncertainty as are those characterizing a new industry, actors use analogies in an attempt to frame the new and unfamiliar in terms of the old and familiar. Thus, even though discontinuous innovation leads to a shift in the technological paradigm (Dosi and Nelson, 2013), the new paradigm is not free from influences emanating from the old one. Resemblance to older and familiar artifacts can also influence how users understand the new product and how willing they are to adopt any particular design (Bingham and Kahl, 2013), thus also influencing the design selection process and the choice of the dominant design.

In a similar vein, Suarez et al. (2014) borrow heavily from work in organizational theory to propose that technological evolution is accompanied by a concurrent categorical evolution. In an attempt to understand and interpret the new technology, stakeholders develop categories that group together products that are perceived to be similar. The notions of similarity and difference essentially hinge upon the socially/politically negotiated dimensions or features that are considered salient for evaluating and categorizing competing designs. Being sociocognitive constructs, different categories capture different understandings or visions of the emerging industry and roughly correspond to different designs. Closely following technological designs, categories go through a similar cycle of variation, selection, and retention that culminates in the emergence of a dominant category, the conceptual schema that most stakeholders adhere to when referring to products that address similar needs and compete for the same market space. The emergence of the dominant category resolves sociocognitive uncertainty in the industry, just as the emergence of the dominant design resolves technological uncertainty. Together, the two concepts define the starting and ending points correspondingly of the window of opportunity for firm entry to emerging industries.

Limitations and Critique

While particularly important for understanding the process of technological innovation and technology cycles, the concept of the dominant design also has some limitations. The most important among them is probably the fact that in most cases the emergence of a dominant design can only be assessed *post facto*. This inability to anticipate which design will dominate offers little guidance to firms. Even though literature on the matter makes it clear that entry to an industry after the emergence of a dominant design is problematic, it does not offer many tools for earlier entrants that need to navigate the uncertainty that is prevalent before the dominant design is established. Such early entrants, then, seem to have two options other than just betting blindly on any one design. First, they can try to impose their chosen design as dominant through deliberate actions that aim at influencing the selection process. Such strategic maneuvering generally requires building mutualistic coalitions or ecosystems around the chosen design in order to give it adequate momentum (Cusumano et al., 1992; Wade, 1995). This is, however, a risky option. Even designs that are technologically superior and supported by large and powerful corporations can fail to achieve dominance, as vividly demonstrated by the case of Sony's Betamax (Cusumano et al., 1992). Firms that invest heavily in and commit to a specific design that fails to achieve dominance face a significant risk to fail as well, especially when they are relatively small in size and cannot afford a repositioning. Alternatively, firms can hedge their bets by investing in multiple designs at the same time. Rather than being fully committed to any specific design, these firms can test the waters, acquire valuable firsthand knowledge about consumer preferences, and develop skills that will allow them to change course quickly if necessary and adopt the dominant design as soon as it emerges. Firms that follow this strategy can have lower failure rates (Sorenson, 2000), but usually also do not aspire to influence the industry's evolution or become market leaders.

In a similar fashion, while the emergence of a dominant design is a useful marker for the closing of the opportunity window for firm entry, it offers little guidance with regard to the opening or the duration of the window and the optimal time for entry. Christensen et al. (1998) in their study of the disk drive industry considered the window to open 3 years before the emergence of the dominant design. However, this was an empirical and ad hoc approach that cannot be generalized across research settings. In fact, if the dominant design can only be assessed *post facto*, the same necessarily will hold for the window of opportunity, thus offering very little concrete information to potential entrants about the optimal entry timing. The concept of the dominant category (Suarez et al., 2014) might offer some more guidance with regard to the assessment of the window of opportunity, yet research on the matter is still at its very early stages.

Conclusions

Since its inception nearly in 1975 (Utterback and Abernathy, 1975), the concept of the dominant design has occupied a central position in research on technological evolution,

timing-of-entry advantages, and organizational performance. Scholarship in these lines of research has been rich and fruitful, advancing significantly the understanding of technological innovation and evolution, and of the role of discontinuous innovation, uncertainty, timing, and strategy in influencing firms' performance in new industries. The clarity of the dominant design concept as a theoretical construct has not necessarily been matched by equal ease in devising generalizable methods for studying its implications in empirical settings or in studying in depth the social, cognitive, and political drivers of design selection. The fact that this remains a vibrant field of research appears to imply that much more remains to be unearthed in this domain in order to complete the understanding of how technologies come to be and how they influence but are also influenced by their environment. Such more in-depth understanding of the technological evolutionary process will also allow scholars to offer more concrete advice to firms regarding their innovation efforts and strategies, the optimal allocation of resources, and their choice of entry timing.

Acknowledgements

SKK GSB MBA student Diego Mendoza provided valuable assistance with the literature review.

See also: Business Models; Business Platforms; Development: Organizational; Ecology: Organizations; Entrepreneurship and Organizations; Hierarchies and Markets; Innovation; Modularity and Organizations; Network Paradigm: Applications in Organizational Science; Organizational Emergence and Firm Formation; Strategic Management; Strategizing.

Bibliography

Abernathy, W.J., 1978. The Productivity Dilemma. Johns Hopkins University Press, Baltimore, MD.

Abernathy, W.J., Utterback, J.M., 1978. Patterns of industrial innovation. Technology Review 80 (7), 40–47.

Agarwal, R., Bayus, B.L., 2002. The market evolution and sales takeoff of product innovations. Management Science 48, 1024–1041.

Anderson, P., Tushman, M.L., 1990. Technological discontinuities and dominant designs: a cyclical model of technological change. Administrative Science Quarterly 35, 604–633.

Arthur, B., 1989. Competing technologies, increasing returns, and lock-in by historical events. Economic Journal 99, 116–131.

Benner, M.J., Tripsas, M., 2012. The influence of prior industry affiliation on framing in nascent industries: the evolution of digital cameras. Strategic Management Journal 33, 277–302.

Bingham, C.B., Kahl, S.J., 2013. The process of schema emergence: assimilation, deconstruction, unitization and the plurality of analogies. Academy of Management Journal 56, 14–34.

Christensen, C.M., Suárez, F.F., Utterback, J.M., 1998. Strategies for survival in fast-changing industries. Management Science 44, S207–S220.

Clark, K.B., 1985. The interaction of design hierarchies and market concepts in technological evolution. Research Policy 14, 235–251.

Cusumano, M.A., Mylonadis, Y., Rosenbloom, R.S., 1992. Strategic maneuvering and mass-market dynamics: the triumph of VHS over beta. The Business History Review 66, 51–94.

Dosi, G., Nelson, R.R., 2013. The evolution of technologies: an assessment of the state-of-the-art. Eurasian Business Review 3, 3–46.

Henderson, R.M., Clark, K.B., 1990. Architectural innovation: the reconfiguration of existing product technologies and the failure of established firms. Administrative Science Quarterly 35, 9–30.

von Hippel, E., 1994. The Sources of Innovation. Oxford University Press, Oxford.

Katz, M.L., Shapiro, C., 1985. Network externalities, competition, and compatibility. The American Economic Review 75, 424–440.

Klepper, S., 1997. Industry life cycles. Industrial and Corporate Change 6, 145–181.

Lieberman, M.B., Montgomery, D.B., 1988. First-mover advantages. Strategic Management Journal 9, 41–58.

Lieberman, M.B., Montgomery, D.B., 1998. First-mover (dis)advantages: retrospective link with resource-based view. Strategic Management Journal 19, 1111–1125.

Murmann, J.P., Frenken, K., 2006. Toward a systematic framework for research on dominant designs, technological innovations, and industrial change. Research Policy 35, 925–952.

Sorenson, O., 2000. Letting the market work for you: an evolutionary perspective on product strategy. Strategic Management Journal 21, 577–592.

Suarez, F.F., Grodal, S., Gotsopoulos, A., 2014. Perfect timing? Dominant category, dominant design, and the window of opportunity for firm entry. Strategic Management Journal. http://dx.doi.org/10.1002/smj.2225.

Suarez, F.F., Utterback, J.M., 1995. Dominant designs and the survival of firms. Strategic Management Journal 16, 415–430.

Tegarden, L.F., Hatfield, D., Echols, A., 1999. Doomed from the start: what is the value of selecting a future dominant design? Strategic Management Journal 20, 495–518.

Tushman, M.L., Murmann, J.P., 1998. Dominant designs, technology cycles, and organization outcomes. Academy of Management Proceedings 98, A1–A33.

Tushman, M.L., Rosenkopf, L., 1992. Organizational determinants of technological change: toward a sociology of technological evolution. Research in Organizational Behavior 14, 311–347.

Utterback, J.M., Abernathy, W.J., 1975. A dynamic model of process and product innovation. Omega 3, 639–656.

Wade, J., 1995. Dynamics of organizational communities and technological bandwagons: an empirical investigation of community evolution in the microprocessor market. Strategic Management Journal 16, 111–133.

Yoxen, E., 1987. Seeing with sound: a study of the development of medical imaging. In: Bijker, W., Hudges, T., Pinch, T. (Eds.), Social Construction of Technological Systems. MIT Press, Cambridge, MA.

Dowries and Other Marital Transfers

Siwan Anderson, Vancouver School of Economics at University of British Columbia, Vancouver, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Payments between families at the time of marriage existed during the history of most developed countries and are currently pervasive in many areas of the developing world. These payments can be substantial enough to affect the welfare of women and a society's distribution of wealth. The prevalence and magnitude of marriage payments vary across countries and depend upon economic conditions, societal structures, institutions, and family characteristics. Marriage payments also evolve within societies and over time. Transfers can rise sharply, and property rights over such payments can shift between marrying partners and parental generations.

Most societies, at some point in their history, have been characterized by payments at the time of marriage. Such payments typically go hand-in-hand with marriages arranged by the parents of the respective spouses. Marital transfers come in various forms and sizes but can be classified into two broad categories: transfers from the family of the bride to that of the groom, broadly termed 'dowry,' or from the groom's side to the bride's, broadly termed 'brideprice.' Brideprice occurs in two-thirds of societies recorded in Murdock's (1967) *World Ethnographic Atlas* of 1167 preindustrial societies. Conversely, dowry occurs in less than 4% of this sample. However, in terms of population numbers, dowry has played a more significant role, because the convention of dowry has occurred mainly in Europe and Asia, where more than 70% of the world's population resides.

Prevalence of Brideprice

The custom of brideprice dates back to the ancient civilizations of Egyptians, Mesopotamians, Hebrews, Aztecs, and Incas (Quale, 1988). A valid marriage contract in Islamic law required a form of brideprice (Bianquis, 1996). Classical China required the negotiation of a brideprice for the validity of marriage, and these transfers continue to be the norm in many rural areas today (Ebrey, 1993). Brideprices are currently most prevalent in Africa; more than 90% of sub-Saharan societies typically make such marriage payments (Murdock, 1967; Goody, 1973).

Prevalence of Dowry

The dowry system dates back at least to the Greeks and Romans. In medieval western Europe and later, dowries were common practice among most social and economic groups. Dowry payments were prevalent in seventeenth and eighteenth century Mexico and Brazil, where Spanish and Portuguese family law governed colonial marriages until those countries gained their independence (Nazzari, 1991; Lavrin and Couturier, 1979). In contemporary times, dowry is an almost universal phenomenon in India. In other parts of South Asia such as Bangladesh, Pakistan, and Sri Lanka, paying a dowry at the time of marriage is increasingly common.

Social Characteristics and the Occurrence of Marriage Payments

There are certain characteristics that distinguish between those societies where the burden of marriage payments falls primarily on the groom's family and brideprices are paid and those where the bulk of the transfer comes from the bride's family and dowries are paid. The general pattern seems to be that brideprice exists more frequently in primitive, tribal, and often nomadic societies. By contrast, dowry often marks a transition to more complex societal structures. In many ancient cultures, the practice of brideprice began to wane and a transformation to the custom of dowry occurred as these civilizations grew and flourished. Goody's (1973) sample of 857 preindustrial societies and Murdock's (1967) sample of 1167 societies also confirms that dowry is mainly found in societies exhibiting substantial socioeconomic differentiation and class stratification. In present times, this pattern is reflected in a comparison between dowry-paying India, where the caste system represents perhaps an extreme example of social stratification, and the more homogeneous tribal societies of sub-Saharan Africa that practice brideprice.

Although brideprice is typically associated with reduced social stratification, societies which do practice the custom are typically developed enough to own some form of property to transfer at the time of marriage. In contrast, in many of the indigenous cultures of North and South America, which seem to have been characterized by more propertyless subsistence, marriage payments were relatively rare (Schlegel and Eloul, 1988).

Brideprice-paying societies have also been associated with a strong female role in agriculture. Boserup (1970), in particular, has argued that brideprice is found in societies in which agriculture relies on light tools (such as the hoe), and thus where women are actively engaged. In contrast, Boserup argues dowry is more common in heavy plow agriculture where the role for women is limited.

The use of brideprice has tended to correlate with polygyny (men have more than one wife) and also with the possibility

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31099-6

of divorce. In contrast, monogamy is the norm and divorce is rare in dowry-paying societies. Lineage and locality of post-marital residence seem to play little role in determining the direction of payments. Both dowry- and brideprice-paying societies tend to be patrilineal (children belong to the lineage of their father) and patrilocal (brides join the household of grooms and their families upon marriage) (Goody, 1973; Murdock, 1967; Quale, 1988).

Family Characteristics as Determinants of Marriage Payments

Determinants of Brideprices

Because women generally join the household of their groom at the time of marriage, brideprice is typically considered to be the payment a husband owes to a bride's parents for the right to her labor and reproductive capabilities. The amount of brideprice required has usually been rather uniform throughout the society, where the size is linked directly to the number of rights that are transferred, and not to the wealth level of the families involved (Quale, 1988; Goody, 1973). Brideprice amounts typically also do not vary with the rank of the wife in polygynous marriages.

Ancient brideprice was often considered a direct payment for a bride's virginity. Where marriage to paternal cousins is common, the brideprice is often higher if the bride married a more distant relative because, in that case, the groom gained rights to a woman and children from a more distant lineage (Bianquis, 1996). The amount of brideprice can also depend on the expected number of children a woman will bear. For example, a divorced woman who already has children will receive a lower brideprice, whereas women who reach puberty earlier receive a higher price (Dekker and Hoogeveen, 2002; Mulder, 1995).

In sub-Saharan Africa, a central purpose of the brideprice is to create an alliance between kinship groups (Dekker and Hoogeveen, 2002; Ekong, 1992). As a result, raising the brideprice is often the responsibility of the groom's extended lineage group, with the principle contributions coming from his father, grandfather, and father's brothers, and with mother's brothers making small contributions. Likewise, since in this setting the entire lineage group has rights to a woman, the brideprice is distributed among many members of the bride's extended family. In this case, larger brideprices can arise with larger lineage groups.

Determinants of Dowries

Relative to brideprices, the amount of dowry varies substantially and tends to be negotiated on an individual basis. Dowry typically arises in complex socioeconomic, non-kinship-based societies, with endogamous marriage practices; that is, where men and women from families of equal social status marry. Dowry then becomes a means to maintain social status by attracting a husband of at least equal standing for one's daughter. It correlates with strongly class-based social systems where higher-level individuals – by virtue of wealth, power, and possibly claim to a superior hereditary status – do not willingly intermarry with the lower levels. A high-status husband might mean a Brahmin in India, or an aristocrat in Renaissance Europe, but the forces tending toward the use of dowry are the same. As a result, the amount of dowry generally increases not only with the wealth of the bride's father, but also with groom's future prospects.

Historically, there is a positive relationship between the amount of dowry transferred and both sides' household wealth. For modern India, dowries are largest among the highest ranking castes (Rao, 1993; Dalmia and Lawrence, 2005). Strikingly similar preferences seem to determine dowry payments in current-day South Asia and in historical Europe. According to Chojancki (2000), the Renaissance marriage market valued maturity in grooms, chaste youth in brides, and family wealth and prominence for both. Typically, in India, the most important quality of a bride is a good appearance, whereas for a groom it is the ability to earn a living, often reflected in his educational level (Caldwell et al., 1983; Billig, 1992). In South Asia today, dowry levels increase with the education of grooms (Dalmia and Lawrence, 2005).

Property Rights and Marriage Payments

Marriage payments not only differ by on whom the financial burden falls (either the grooms' brideprice or the brides' dowry), but also ownership rights to these transfers can vary and evolve through time. Payments from the groom's side are either transferred directly to the brides' parents, formally termed 'brideprice,' or to the bride directly, commonly known as 'dower.' Similarly, payments from the bride's parents can either go to the bride herself, referred to as 'dowry,' or be directly transferred to the groom and his family, termed 'groomprice,' to the exclusion of the bride. Both dowry (as a pre-mortem inheritance) and dower are transfers that remain with the conjugal couple but are the formal property of the wife throughout the marriage.

Brideprice and Property Rights

Brideprice transfers, where the bride's parents receive the payment, are the norm in sub-Saharan Africa. Dower, which remains the property of the bride and is usually considered her insurance in case of divorce or husband's death, is most typical of traditional Islamic marriages. The historical record on brideprice versus dower suggests property rights over the marriage transfer evolve through time.

Within the ancient civilizations of the Babylonians, Egyptians, and Hebrews, there was an overall shift away from brideprice to dowry as the society developed and urbanized. An additional feature of this transformation was that the marriage payment seemed to change first from a brideprice to a dower and then finally to a dowry (Quale, 1988). A similar transformation seemed to occur for the Germanic tribes of the early Middle Ages (Hughes, 1985). By the fifth century, the Burgundians awarded one-third of their brideprice to their wives and two-thirds to their wives' kinsmen. By the time the law codes were issued, the Visigoths seem to have totally incorporated their ancient brideprice into the award grooms made to their brides. This pattern is also reflected in the traditional Chinese practice of transferring a brideprice directly to the

bride's parents, who then return a portion of this as dowry to their daughter. This amount is kept by the daughter as personal property throughout the marriage.

Dowry and Property Rights

Ownership rights over marriage payments from the brides' side have also undergone transformations. Most commonly, the traditional dowry transfer is considered to be a 'pre-mortem inheritance' to the daughter, which formally remains her property throughout the marriage. Goody (1973), in particular, has emphasized this role of dowry in systems of 'diverging devolution,' where both sons and daughters have inheritance rights to their parent's property. As Botticini and Siow (2003) summarize, a strong link exists between women's rights to inherit property and the receipt of a dowry. This is seen in ancient Rome, medieval western Europe, and the Byzantine Empire.

In numerous historical instances, dowry as bequests has given way to groomprices – a direct transfer to grooms. Chojancki (2000) documents the emergence of a gift of cash to the groom (*corredo*) as a component of marriage payments in Renaissance Venice. Krishner (1991) confirms a pattern of legislation across northern and central Italy granting husbands broader control over wife's dowry assets beginning in the fourteenth century. Common law, in which dowry came under immediate control of husbands, predominated in England during the sixteenth and seventeenth centuries (Erickson, 1993). All of these examples point to a time of increased commercial activity and societal inequality, where groomprices emerged to secure husbands from prominent families.

Nowhere, however, has this transformation been more dramatic than in present-day India over the last few decades. The traditional custom of *stridhan*, a parental gift to the bride, has changed into modern-day groomprices that have a highly contractual and obligatory nature (Caldwell et al., 1983; Billig, 1992). Generally a bride is unable to marry without providing such a payment. This transition in marriage payments happened at a time of significant structural change, as unprecedented opportunities for economic and political mobility began to open up following India's independence in 1947. Several studies connect the emergence of groomprice to competition among brides for more desirable grooms (Srinivas, 1984). This transformation into groomprice is occurring elsewhere in South Asia, too. The Pakistan Law Commission reviewed dowry legislation in 1993 adding a new sub-clause stating grooms should be prohibited from demanding a dowry. In Bangladesh, a clear distinction remains between the traditional dowry, *joutuk*, gifts from the bride's family to the bride, and the new groom payments referred to as *demand*, which emerged post-Independence in the 1970s (Amin and Cain, 1995).

Rising Marriage Payments

In many contexts, marriage payments have risen substantially. Such rises occur particularly for dowries, and have often precipitated legislative and regulatory initiatives to reduce their effect. Real dowries have recently been rising in India, where payments now can amount to as much as six times the annual income of bridal parents. The political outcry against this escalation of dowry payments culminated in the passage of the Dowry Prohibition Act in 1961 outlawing the practice. But despite this law, dowry inflation has persisted unabated. Similar legislation has occurred in Pakistan and Bangladesh. Dowry inflation is also well documented to have occurred in historical Europe. For example, the scale of dowry demands in Florence led to the establishment of the Florentine Dowry Fund in 1424, which was designed to alleviate the government's fiscal problems while providing families with a government-secured investment for accumulating dowries (Molho, 1994).

The demographic explanation for dowry inflation considers the relative supply of grooms and brides in the marriage market. In most societies, men on average, marry younger women. Since women reach marriageable age ahead of men, increases in population could affect the numbers of potential brides first, thus effectively causing an excess supply of potential brides in the marriage market. If grooms are in relatively short supply, a corresponding increase in the price of husbands should occur, and hence a rise in dowries. The flip side of an excess supply of grooms would consequently lead to brideprice inflation. Some anthropological research on brideprice inflation in colonial Africa has indeed made the link between higher brideprices and greater demand for wives leading to increased out-migration of eligible grooms (Wilson, 1981).

Marriage Payments and the Welfare of Women

Although marriage payments can take many different forms, no consensus exists on which of these forms, if any, will enhance the welfare of women. In theory, brideprice could be interpreted as explicit recognition and valuing of women's productivity and contribution to marriage; in practice, it often serves to limit women's control over their bodies. Both sexually and in terms of their labor, brideprice has long been linked to domestic violence, owing to women's fear of returning to their natal home without being able to repay the brideprice (Ansell, 2001). African women's rights campaigners advocate the abolishment of the practice, and have linked it to the spread of AIDS, since brideprice as payment for sexual rights leads to women's loss of say in sexual protection and frequency (Wendo, 2004).

In theory, dowry, as a pre-mortem inheritance, is set up to protect property given to women. However, this institution often seems to have been transformed into giving these property rights to men. Early feminists in Europe attacked the dowry system precisely because they objected to the fact that husbands ended up controlling the funds (Cox, 1995). An even more significant factor is the magnitude of these payments. In current-day South Asia, dowry payments can impoverish the bridal family, and dramatically affect the lives of unmarried women, who are increasingly considered burdensome economic liabilities. The custom of dowry in India has been linked to female infanticide and, among married women, to 'bride-burning' and 'dowry-death' – physical harm inflicted on the wife (sometimes leading to death) to extract promised dowry payments (Bloch and Rao, 2002). The National Crime Bureau of the Government of India reports approximately

8000 to 9000 dowry deaths every year. Numerous incidents of dowry-related violence are never reported; Menski (1998) estimates the number as roughly 25 000 per year. Other estimates point to as many as 100 000 per year (Anderson and Ray, 2012).

See also: Asia, Sociocultural Overviews: South Asia; Family Theory: Economics of Marriage and Divorce; Fertility Theory: Theory of Intergenerational Wealth Flows; Fertility Transition: South Asia; Fertility Transition: Sub-Saharan Africa; Islam: Sub-Saharan Africa; Marriage Systems, Evolution of; Marriage; Wealth Distribution.

Bibliography

Amin, S., Cain, M., 1995. The rise of dowry in Bangladesh. In: Jones, G., Caldwell, J., Douglas, R., D'Souza, R. (Eds.), The Continuing Demographic Transition. Oxford University Press, Oxford.

Anderson, S., Ray, D., 2012. The age distribution of missing women in India. Economic and Political Weekly XLVII (47 and 48), 87–95.

Ansell, N., 2001. 'Because it's our culture!' (Re)Negotiating the meaning of 'Lobola' in Southern Africa Secondary Schools. Journal of Southern African Studies 27 (4), 697–716.

Bianquis, T., 1996. The family in Arab islam. In: Burguiere, A., Klapisch-Zuber, C., Segalen, M., Zonabend, F. (Eds.), A History of the Family, Volume One. Polity Press, Cambridge, UK, pp. 601–647.

Billig, Michael S., Spring 1992. The marriage squeeze and the rise of groomprice in India's Kerala state. Journal of Comparative Family Studies 23, 197–216.

Bloch, F., Rao, V., 2002. Terror as a bargaining instrument: dowry violence in rural India. American Economic Review 92 (4), 1029–1043.

Boserup, E., 1970. Women's Role in Economic Development. St. Martin's Press, New York.

Botticini, M., Siow, A., 2003. Why dowries? American Economic Review 93 (4), 1385–1398.

Caldwell, J.C., Reddy, P.H., Caldwell, Pat, 1983. The causes of marriage change in South Asia. Population Studies 37 (November), 343–361.

Chojancki, S., 2000. Women and Men in Renaissance Venice. Johns Hopkins University Press, Baltimore.

Cox, V., 1995. The single self: feminist thought and the marriage market in early modern Venice. Renaissance Quarterly 48 (3), 513–581.

Dalmia, S., Lawrence, P., 2005. The institution of dowry in India: why it continues to prevail. Journal of Developing Areas 38 (2), 71–93.

Dekker, M., Hoogeveen, H., 2002. Bridewealth and household security in rural Zimbabwe. Journal of African Economies 11 (1), 114–145.

Ebrey, Patricia Buckley, 1993. The Inner Quarters: Marriage and the Lives of Chinese Women in the Sung Period. University of California Press, Berkeley.

Ekong, J.M., 1992. Bridewealth, Women and Reproduction in Sub-Saharan Africa. Holos Verlag, Bonn.

Erickson, A.L., 1993. Women and Property in Early Modern England. Routledge, London.

Goody, Jack, 1973. Bridewealth and dowry in Africa and Eurasia. In: Goody, Jack, Tambiah, Stanley J. (Eds.), Bridewealth and Dowry. Cambridge University Press, Cambridge.

Hughes, Diane Owen, 1985. From brideprice to dowry in Mediterranean Europe. In: Kaplan, Marion A. (Ed.), The Marriage Bargain: Women and Dowries in European History. Havorth Press, New York.

Krishner, J., 1991. Materials for a gilded cage: non-dotal assets in Florence, 1300–1500. In: Kertzer, D., Saller, R. (Eds.), The Family in Italy from Antiquity to Present. Yale University Press, New Haven, pp. 184–207.

Lavrin, A., Couturier, E., 1979. Dowries and wills: a view of women's socioeconomic role in colonial Guadalajara and Puebla, 1640–1790. The Hispanic American Historical Review 59 (2), 280–304.

Menski, W. (Ed.), 1998. South Asians and the Dowry Problem. Vistaar Publications, New Delhi.

Molho, Anthony, 1994. Marriage Alliance in Late Medieval Florence. Harvard University Press, Cambridge.

Mulder, M.B., 1995. Bridewealth and its correlates: quantifying changes over time. Current Anthropology 36 (4), 573–603.

Murdock, George P., 1967. Ethnographic Atlas. University of Pittsburg Press, Pittsburg.

Nazzari, Muriel, 1991. Disappearance of Dowry: Women, Families, and Social Change in Sao Paulo, Brazil (1600–1900). Stanford University Press, Stanford.

Quale, C.R., 1988. A History of Marriage Systems. Greenwood Press, New York.

Rao, V., 1993. The rising price of husbands: a hedonic analysis of dowry increases in rural India. Journal of Political Economy 101, 666–677.

Schlegel, A., Eloul, R., 1988. Marriage Transactions: labor, property, status. American Anthropologist, New Series 90 (2), 291–309.

Srinivas, M.N., 1984. Some Reflections on Dowry. Oxford University Press, New Delhi.

Wendo, C., 2004. African women denounce brideprice. Lancet 363 (9410), 716.

Wilson, M., 1981. Xhosa marriage in historical perspective. In: Krige, E., Comaroff, J. (Eds.), Essays on African Marriage in Southern Africa. Juta and Co. Ltd., Cape Town.

Dramaturgical Analysis

Robert D Benford, University of South Florida, Tampa, FL, USA
A Paul Hare[†]**,** Ben-Gurion University, Beer Sheva, Israel

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A.P. Hare, volume 6, pp. 3834–3836, © 2001, Elsevier Ltd.

Abstract

Drawing on the language of the theater, dramaturgical analysis illuminates how people derive shared understandings and collaboratively construct reality in everyday, cultural, organizational, and institutional life. Dramaturgical analysis centers on two interrelated concepts: the stage or region and the social roles enacted or performed in each region. Several academic fields contribute to and are enhanced by the approach including sociology, communication studies, anthropology, psychology, social psychology, political science, philosophy, linguistics, psychotherapy, cultural studies, performance studies, film studies, and the methods of narrative and qualitative analysis. Following an overview of earlier conceptualizations and contemporary applications, a summary of critiques and defenses of dramaturgical analysis is presented. The article concludes with a note on the challenges and opportunities dramaturgical analysts face in the rapidly changing digital age.

Dramaturgical analysis of social interaction is based on the assumption that social acts are 'staged', consciously or unconsciously, and thus embody all the elements found in enactments in the theater. Some dramaturgists, especially those following Goffman (1959), assert that life is like theater. From this perspective, "social life can be understood as a series of performances" (Brissett and Edgley, 1990: p. 36). Applying the theater metaphor and various theatrical concepts turns the analytic spotlight on how people collaboratively perform and manage social roles in their everyday life encounters. Other dramaturgists, especially those following Burke's (1968, 1969) 'dramatism', contend that life is a theater (Branaman, 2013). Rejecting dramaturgy as the application of 'a mere metaphor', proponents of this approach contend that the theater reflects life itself. In other words, plays on the theater stage highlight all the elements of ordinary social life to present to an audience a new perspective on some aspect of social interaction. Thus, concepts used in theater production can be turned back again for the analysis of the social behavior that they are designed to reflect.

Regardless of the scholar's ontological position concerning the relationship between the theater and everyday life, dramaturgical analysis illuminates how people construct reality (Berger and Luckmann, 1966) and sustain the interaction order (Goffman, 1983). As Edgley notes in *The Drama of Social Life: A Dramaturgical Handbook*, "stripped to its essentials, dramaturgy is about the ways in which human beings, in concert with similarly situated others, create meaning in their lives" (2013: p. 2). The shared overall meaning, the definition of the situation, binds the participants in any situation or encounter together. The meaning may take the form of an image, a theme, a plot, or a script.

The Main Concepts

Dramaturgical analysis centers on two interrelated concepts: the stage or action area region and on the social roles enacted or performed in each region (Hare and Blumberg, 1988). In *The Presentation of Everyday Life*, Goffman (1959) identified three action area regions – backstage, frontstage, and offstage. The backstage area is where the actors prepare for their roles and where special effects are produced to influence the audience. Frontstage is where the action takes place in full view of the audience.

For any action area, there may also be offstage areas in which persons who have organized the activity (producers), as well as those who have rehearsed the cast (directors) and are providing cues for action, remain hidden from the audience. In addition, there may be a person or group who provided the original idea or script for the performance (playwright). This does not complete the list as there may also be persons who cater for the audience.

The region affects the actor's ability to control information to which the audience is privy and consequently the capacity to sustain a credible performance. If the audience were able to get a glimpse backstage, the impressions the actors are seeking to foster frontstage could be discredited. For example, a peek behind the swinging doors of an expensive restaurant revealing the chaotic and intense activities of chefs, waitpersons, and dishwashers would likely yield a quite different impression than the calm elegance being staged in the dining area. For some activities or performances such as street demonstrations (Martin, 2013), the segregation of stages and thus the capacity to shield the audience from discrediting information or impressions that are discrepant from the roles the performers are playing can be difficult. The advent of digital technologies (e.g., the Internet, social media, e-mail, cell phones, and surveillance cameras) and thus virtual space as a performance region provides both new opportunities and new challenges for actors seeking to control information intentionally 'given' and that which is unwittingly 'given off' (Goffman, 1959; Gottschalk and Whitmer, 2013; Markham, 2013).

In face-to-face interaction as well as in some mediated venues, people take turns serving in the roles of performer and audience for each other even in situations in which two people are involved in the interaction. When one is speaking,

[†] Deceased.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32046-3

the other is listening. As an audience, an actor provides consensual reality, cueing, social reinforcement, and continual observation (Sarbin in Allen and Scheibe, 1982). In the progression of the development of social interaction over time, agreement on an actable idea and scripting consistent with the overall meaning constitutes the first phase. The second phase involves staging by locating or constructing an action area and providing resources such as props, costumes, and other necessary equipment. In the third phase, which may overlap with the second, actors are recruited, if they are not already involved in the development of the actable idea, and the actors are trained for their roles. The fourth phase is the period of enactment when the play is performed. In the final phase, after the performance, new meanings for selves, others, and the situation may be interpreted or assessed by the actors and audience.

The actable ideas that form the basis for social dramas may be either very general or very specific in their implications. On one end of the continuum is an actable idea in the form of an image, an emotionally loaded set of words, physical object, or event, for example, 9/11 or 'terrorist threat' (Altheide, 2013). An image has a program of action packed into it much like a symbol in a dream may be a merger of waking events. An image often guides interaction in a small informal or therapy group. The actable idea may be a more complex theme (including a direction of movement, an emotional tone, and a minimal set of roles to be enacted) or a plot (with a detailed scenario, defined roles, and an indication of the phases a group must go through to reach its goal). At the specific end of the continuum, the idea may be as fully developed as a script for a play, with parts for each member of the cast and stage directions to guide the performance. A script is most evident in organizations, especially where the actors must use equipment.

Building on the perspective of Burke (1969), Duncan (1968) emphasizes the importance of symbols, especially as they reflect authority relationships. For MacCannell and MacCannell (1982), the study of signs and symbols had become so important by the late 1900s that they suggested a semiotic interpretation of modern culture as an approach to the time of the sign. Brissett and Edgley (2006) add to the list variables that follow from the work of Goffman (1959) with a primary focus on the way in which individuals present themselves to others. These variables include role distance, the case when an individual is playing a role with full consciousness and thereby has the choice of either modifying the role or ceasing to play it altogether (Goffman, 1961; Stebbins, 2013); ecstasy, stepping outside the world of everyday life; and remedial work, when a person attempts to change the meaning of an act from one that is seen as offensive to one that is acceptable.

Academic Fields That Contribute to Dramaturgical Analysis

Dramaturgical analysis has been used in several ways. The first way, employed primarily by sociologists within the symbolic interactionist tradition, focuses on role playing (Goffman, 1959) in a wide variety of studies regarding cultural, organizational, and political life (Manning, 2004). The second way, introduced by the philosopher–linguist Burke (1969), and utilized most frequently by communication scholars, is based on the analysis of the act (Branaman, 2013). Another approach, used primarily by anthropologists, is to analyze the function of social dramas that are enacted by the members of a society to bring about conflict resolution (Turner, 1974), to signify a rite of passage, or to memorialize some important event in the history of the society (Drummond, 2013).

Since role and script are central concepts for drama, the fields of psychology, social psychology, political science, philosophy, linguistics, psychotherapy, cultural studies, performance studies, film studies, and the methods of narrative and qualitative analysis, also include descriptions of similar relationships that can be incorporated in the dramaturgical perspective.

Earlier Conceptualizations of the Dramaturgical Perspective

The fundamental idea of people playing roles for audiences who take those performances into consideration and adjust their subsequent lines of action has been around for several centuries. Evreinoff, in *The Theatre in Life* (1927), records that, at least since the sixteenth century, the similarity between life and theater has been noted by Shakespeare and others. However, the dramaturgical perspective was not widely used until it was popularized by Goffman in the 1960s. The 1930–35 edition of the *Encyclopedia of the Social Sciences*, edited by Seligman and Johnson, does not include an entry on the dramaturgical perspective, although there is an entry for drama, covering activity in theaters. Burke's entry on dramatism in the 1968 *International Encyclopedia of the Social Sciences*, edited by Sills, presents Burke's pentad of the five terms: act, scene, agent, agency, and purpose. Sets of concepts are linked in ratios, such as the scene–act ratio, where the scene (the situation) has implicit in it the attitudes and behaviors that will become explicit through action. Burke places his own work clearly in the symbolic interaction tradition with references to Mead (1938).

Brissett and Edgley, in various editions (1973, 1990, and 2006) of their source book on *Life as Theater*, emphasize that dramaturgy is the study of how human beings accomplish meaning in their lives through interaction with others. For them "meaning is a continually problematic accomplishment of human interaction and it is fraught with change, novelty, and ambiguity" (1990: p. 2). Edgley asserts that "the simplicity of the dramaturgical principle—that the meaning of people's conduct is to be found in the manner in which they express themselves in interaction with similarly expressive others belies the scope and power of the perspective" (2013: p. 2). Just as meaning is created and expressed through enaction, it can be recreated through reenaction. Moreno, observing that "every second time is a liberation from the first" (Hare and Hare, 1996), developed psychodrama as a method of group therapy based on reenactment.

In the 1992 *Encyclopedia of Sociology* edited by Borgatta and Borgatta in the entry on 'Dramaturgy,' Warner records that the unique insight of dramaturgy is that daily social interaction consists of performances that individuals stage and in which they present themselves to others. As with Brissett and Edgley (2006), the emphasis is on the work of Goffman and the symbolic interaction tradition of Mead and the Chicago School. Curiously, the 2000 edition of the *Encyclopedia of Sociology* edited by Borgatta and Montgomery lacks an entry on the topic of dramaturgy.

Contemporary Applications of Dramaturgical Analysis

Over a half century since Goffman popularized dramaturgical analysis, the approach continues to stimulate and shape scholarly research and theorizing across a plethora of disciplines and substantive areas. Early empirical research guided by the dramaturgical perspective examined a variety of institutional and organizational realms including funerals (Turner and Edgley, 1976), political impression management (Hall, 1979), idiosyncrasy credit earned by new religious movements (Snow, 1979), celebratory crowds (Snow et al., 1981), prisoners' presentations of self (Watson, 1982), the feeling rules and emotion work of flight attendants and bill collectors (Hochschild, 1983), and the social construction and communication of power by social movements (Benford and Hunt, 1992) to name but a few. Edgley's (2013) volume reflects the sustained contemporary influence of and interest in dramaturgy. The 19 original papers in that collection are equally divided between conceptual elaborations and empirical investigations of topics including social movements and dramatic framing (Benford, 2013), police/protester interaction and political impression management (Martin, 2013), media dramas fanning public fears concerning domestic terrorism (Altheide, 2013), the performative body and body impression management (Waskul and Vannini, 2013), museum dramas (vom Lehn, 2013), the staging of emotions by barbershop quartets (Nash, 2013), transsexuals' gendered identity work (Sumerau et al., 2013), sadomasochistic play (Newmahr, 2013), dramaturgy in online encounters (Gottschalk and Whitmer, 2013), and symboling in the drama of human evolution (Halton, 2013). Taken together, these studies indicate that dramaturgical analysis remains a vibrant approach for the social sciences, illuminating a variety of domains in everyday, cultural, organizational, and institutional life.

Critiques and Defenses of Dramaturgical Analysis

Brissett and Edgley (1990, 2006), Edgley (2013), and Manning (2004) provide lists of various criticisms that have been leveled at dramaturgy. For some critics, dramaturgy is a pedestrian, nonsystematic form of inquiry that does not possess the properties of formal theory. Brissett and Edgley agree that for those who would prefer a linear sequential analysis of human behavior and who wish to employ dramaturgy to test hypotheses, the lack of formal propositions is a shortcoming. But they go on to suggest that dramaturgy's theoretical elasticity is also one of its strengths:

> In fact, we believe it is precisely because dramaturgy is not a closed theoretical position, but rather a way of describing human behavior, that it is such an informative and heuristic mode of thought. (Brissett and Edgley, 1990: p. 24)

They also note that dramaturgy is linked propositionally to other forms of social thought, such as symbolic interactionism, ethnomethodology, existential sociology, interpersonal psychology, and various humanistic models in the social sciences. Edgley further deflects this criticism by pointing out that "dramaturgy is a tool, a perspective, and a point of view, not a paradigm" (2013: p. 8).

The second criticism is that dramaturgy does not produce universal statements about human behavior. It is said to be an artifact of studying the situational behavior in Western culture. In response, Brissett and Edgley (1990) first noted that understanding how people interact in the Western society is no small achievement, but go on to point to the fact that the expressive behavior of individuals is well documented in the anthropological literature outside the Western world.

Dramaturgy, or at least Goffman's version of it, has also been criticized for its methodology. It is said to lack a systematic method of testing propositions about the world. Hunt and Benford (1997: p. 106) agree that a "systematic rendering of a dramaturgical methodology has not been attempted." They outlined a qualitative dramaturgical methodology for studying collective action and social movements; but their blueprint would not satisfy those wishing to apply quantitative techniques to studying such phenomena. Brissett and Edgley (1990) responded to critics that there is nothing special about doing dramaturgy. Being sensitive to the expressive dimension of behavior demands no special methodology or observational skills. However, there should be an unnerving and single-minded commitment to the observation of people's doings.

The fourth criticism is that dramaturgy slights the impact of larger social units, such as institutions, on human behavior. In response, Brissett and Edgley (1990) assert that there can be no doubt that people's interaction contexts are circumscribed by structural arrangements. However, rather than dwelling on the limitations, dramaturgy focuses on what people actually do within the contexts that are available to them. Edgley elaborates:

> The preoccupations of macro-sociologists are but mere background to dramaturgists for whom the action on stage is primary in every way. Background matters may influence, pressure, encourage, or condition, but in the last analysis they do not determine the course of interaction in whose crucible novelty and surprise are always possible. (2013: p. 3)

Moreover, dramaturgy's central concept of role is used as a way of accounting for people's connections to one another as well as to the structures and organizations with which they are identified.

Several scholars have criticized the dramaturgical perspective for its implicitly jaded view of human nature portraying people as self-interested manipulators, con-artists, liars, and other disingenuous characters. Edgley responds that such critics have failed to take into account the simple distinction that "human beings are not only expressive, but often aware of their expressiveness". He elaborates:

> To these critics, the dramaturgical image of human beings as expressive creatures who are also aware of their expressiveness means that they are persons constantly concerned with manipulating that awareness in order to influence how others see them. Possibilities become probabilities and probabilities become certainties. So if people *could* have self-serving interests in mind when acting on their dramaturgical awareness, then this is proof-positive that they *do*… But simply because people *might* have nefarious motives for exercising dramaturgical awareness doesn't mean that they *do*. Not all presentations are misrepresentations… (Edgley, 2013: p. 7)

Edgley further suggests that some of the critics are also selling our species short as well as underestimating the scope of dramaturgical analysis. "Con games, deceit, cynicism and treachery are all dramaturgically accomplished to be sure", he concedes; "but love, truth, sincerity, and authenticity are as well" (p. 7).

Finally, the most prevalent critique has to do with the theatrical metaphor. Critics say that the theater is make-believe and everyday life is real. Brissett and Edgley (1990) respond that life is neither theater nor is it different from theater. It is theaterlike. They recommend that dramaturgy and social science more generally drop all claims to ontology because

> …all knowledge of and about social life is necessarily interpretive. As symbolic creatures, we have the capacity to interpret and express our observations…but our knowledge is limited by the symbols at our disposal. While there may indeed be a "real" world out there, to claim that any particular theory or model reflects, captures, or perceives the world, is quite another matter. (p. 34)

They suggest that social scientists would be well advised to follow Gamson's recommendation to "abandon the bewildering issue of what is reality and substitute the more manageable question: Under what circumstances do we think things are real?" (1975: p. 603).

Future Directions for Theory and Research

The rapidly changing digital landscape provides new challenges and opportunities for dramaturgical analyses. The sheer volume of communications and interactions being spewed out in nearly every sphere of private and public social life means there are innumerable new venues in which performers and audiences are constructing meanings, interpreting, and reacting to each other's messages and impressions. Researchers employing a dramaturgical approach to studying digitally mediated encounters and interactions will find it necessary to develop new methodologies to keep pace with the implications of these technological innovations. As Gottschalk and Whitmer observe, "Since what happens online does not stay online, we must constantly resolve the tensions between the infinity of interactional opportunities our online life provides us, and the offline consequences of enjoying them" (2013: p. 328). For performers, the digital world presents new challenges to controlling information and the audiences who gain access to the performances. Thus, there are more opportunities to have one's impressions discredited by the unintentional or intentional release of discrepant information. Dramaturgical analysts will have unlimited data at their fingertips, which further illuminate the dynamics and fragility of everyday life. And they will need to confront new ethical dilemmas as private and public selves become more blurred in the digital age.

See also: Action, Theories of Social; Collective Behavior, Sociology of; Interactionism, Symbolic; Movements, Social; Social Psychology, Sociological; Status and Role, Social Psychology of.

Bibliography

Allen, V.L., Scheibe, K.E. (Eds.), 1982. The Social Context of Conduct: Psychological Writings of Theodore Sarbin. Praeger, New York.
Altheide, D.L., 2013. Media drama and the social construction of reality. In: Edgley, C. (Ed.), The Drama of Social Life: A Dramaturgical Handbook. Ashgate, Surrey, UK, pp. 181–195.
Benford, R.D., 2013. Social movements and the dramatic framing of social reality. In: Edgley, C. (Ed.), The Drama of Social Life: A Dramaturgical Handbook. Ashgate, Surrey, UK, pp. 139–155.
Benford, R.D., Hunt, S.A., 1992. Dramaturgy and social movements: the social construction and communication of power. Sociological Inquiry 62, 36–55.
Berger, P., Luckmann, T., 1966. The Social Construction of Reality. Anchor Books, New York.
Branaman, A., 2013. Drama as life: the seminal contributions of Kenneth Burke. In: Edgley, C. (Ed.), The Drama of Social Life: A Dramaturgical Handbook. Ashgate, Surrey, UK, pp. 15–25.
Brissett, D., Edgley, C. (Eds.), 1990. Life as Theater: A Dramaturgical Sourcebook, second ed. Aldine de Gruyter, New York.
Brissett, D., Edgley, C. (Eds.), 2006. Life as Theater: A Dramaturgical Sourcebook. Transaction Publishers, New Brunswick.
Burke, K., 1968. Dramatism. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences. Macmillan, New York.
Burke, K., 1969. A Grammar of Motives. University of California Press, Berkeley, CA.
Drummond, K.L., 2013. Victor Turner's dramaturgical theory of ritual. In: Edgley, C. (Ed.), The Drama of Social Life: A Dramaturgical Handbook. Ashgate, Surrey, UK, pp. 43–56.
Duncan, H.D., 1968. Symbols in Society. Oxford University Press, New York.
Edgley, C., 2013. Introduction. In: Edgley, C. (Ed.), The Drama of Social Life: A Dramaturgical Handbook. Ashgate, Surrey, UK, pp. 1–12.
Evreinoff, N., 1927. The Theatre in Life. Bretano's, New York.
Gamson, W.A., 1975. Review of frame analysis: an essay on the organization of experience. Contemporary Sociology 4, 603–607.
Goffman, E., 1959. The Presentation of Self in Everyday Life. Doubleday, Garden City, NY.
Goffman, E., 1961. Encounters: Two Studies in the Sociology of Interaction. Bobbs-Merrill, Indianapolis, IN.
Goffman, E., 1983. The interaction order. American Sociological Review 48, 1–17.
Gottschalk, S., Whitmer, J., 2013. Hypermodern dramaturgy in online encounters. In: Edgley, C. (Ed.), The Drama of Social Life: A Dramaturgical Handbook. Ashgate, Surrey, UK, pp. 309–334.
Hall, P.M., 1979. The presidency and impression management. Studies in Symbolic Interaction 2, 283–305.
Halton, E., 2013. Tale of the evolutionary drama of symboling: a dramaturgical digression. In: Edgley, C. (Ed.), The Dama of Social Life: A Dramaturgical Handbook. Ashgate, Surrey, UK, pp. 27–41.
Hare, A.P., Blumberg, H.H., 1988. Dramaturgical Analysis of Social Interaction. Praeger, New York.

Hare, A.P., Hare, J.R., 1996. J. L. Moreno. Sage, London.
Hochschild, A.R., 1983. The Managed Heart: The Commodification of Human Feelings. University of California Press, Berkeley, CA.
Hunt, S.A., Benford, R.D., 1997. Dramaturgy and methodology. In: Miller, G., Dingwall, R. (Eds.), Context and Method in Qualitative Research. Sage, London, Thousand Oaks and New Delhi, pp. 106–118.
MacCannell, D., MacCannell, J.F., 1982. The Time of the Sign: A Semiotic Interpretation of Modern Culture. Indiana University Press, Bloomington, IA.
Manning, P., 2004. Dramaturgy. In: Ritzer, G. (Ed.), Encyclopedia of Social Theory. Sage, London, Thousand Oaks, and New Delhi, pp. 210–213.
Markham, A., 2013. The dramaturgy of digital experience. In: Edgley, C. (Ed.), The Drama of Social Life: A Dramaturgical Handbook. Ashgate, Surrey, UK, pp. 279–293.
Martin, D.D., 2013. The drama of dissent: police, protestors, and political impression management. In: Edgley, C. (Ed.), The Drama of Social Life: A Dramaturgical Handbook. Ashgate, Surrey, UK, pp. 157–179.
Mead, G.H., 1938. The Philosophy of the Act. University of Chicago Press, Chicago.
Nash, J.E., 2013. Puttin' on your face: staged emotions among barbershop Singers. In: Edgley, C. (Ed.), The Drama of Social Life: A Dramaturgical Handbook. Ashgate, Surrey, UK, pp. 229–243.
Newmahr, S., 2013. Sadomasochistic selves: dramaturgical dimensions of SM play. In: Edgley, C. (Ed.), The Drama of Social Life: A Dramaturgical Handbook. Ashgate, Surrey, UK, pp. 261–275.
Snow, D.A., 1979. A dramaturgical analysis of movement accommodation: building idiosyncrasy credit as a movement mobilization strategy. Symbolic Interaction 2, 23–44.
Snow, D.A., Zurcher, L.A., Peters, R., 1981. Victory celebrations as theater: a dramaturgical approach to crowd behavior. Symbolic Interaction 4, 21–42.
Stebbins, R.A., 2013. Role-distance, activity distance, and the dramaturgical metaphor. In: Edgley, C. (Ed.), The Drama of Social Life: A Dramaturgical Handbook. Ashgate, Surrey, UK, pp. 123–136.
Sumerau, J.E., Schrock, D.P., Reese, T.J., 2013. Transsexuals' gendered self-presentations. In: Edgley, C. (Ed.), The Drama of Social Life: A Dramaturgical Handbook. Ashgate, Surrey, UK, pp. 245–260.
Turner, V., 1974. Dramas, Fields, and Metaphors: Symbolic Action in Human Society. Cornell University Press, Ithaca, NY.
Turner, R., Edgley, C., 1976. Death as theater: a dramaturgical analysis of the American funeral. Sociology and Social Research 60, 377–392.
vom Lehn, D., 2013. Museum drama and the interaction order: 'suis generis': works of art as hubs for co-orientation. In: Edgley, C. (Ed.), The Drama of Social Life: A Dramaturgical Handbook. Ashgate, Surrey, UK, pp. 211–227.
Warner, K.A., 1992. Dramaturgy. In: Borgatta, E.F., Borgatta, M.L. (Eds.), Encyclopedia of Sociology. Macmillan, New York, pp. 513–516.
Waskul, D., Vannini, P., 2013. The performative body: dramaturgy, the body, and embodiment. In: Edgley, C. (Ed.), The Drama of Social Life: A Dramaturgical Handbook. Ashgate, Surrey, UK, pp. 197–210.
Watson, C.M., 1982. The presentation of self and the new institutional inmate: an analysis of prisoners' responses to assessment for release. Symbolic Interaction 5, 243–257.

Dreaming, Neural Basis of

Lampros Perogamvros, University Hospitals of Geneva, Geneva; University of Geneva, Geneva, Switzerland; Geneva Neuroscience Center, University of Geneva, Geneva, Switzerland; and Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland
Sophie Schwartz, University of Geneva, Geneva, Switzerland; Geneva Neuroscience Center, University of Geneva, Geneva, Switzerland; and Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.A. Hobson, volume 6, pp. 3836–3841, © 2001, Elsevier Ltd.

Abbreviations

ACC anterior cingulate cortex
D2 D2 dopamine receptors
dlPFC dorsolateral prefrontal cortex
mPFC medial prefrontal cortex
N1, N2, N3 stages N1, N2, and N3 of NREM sleep
NAcc nucleus accumbens
NREM non-rapid eye movement sleep
OFC orbitofrontal cortex
PTO parieto–temporo–occipital junction
PFC prefrontal cortex
REM rapid eye movement sleep
VTA ventral tegmental area

Abstract

Recent neuroimaging, neuroanatomical, neurophysiological, pharmacological, and dream content analysis studies have shed new light on the neural underpinnings of dreaming. In this article, we review converging evidence from these distinct lines of research suggesting that the generation of dreams primarily relies on the activation of brain networks implicated in memory, emotional and reward processes, as well as parieto-occipital regions, related to sensory and visuospatial imagery. We show that these activations may account for distinct characteristics of dream reports from non-rapid eye movement sleep (NREM) and rapid eye movement sleep (REM). Compared to NREM dreams, REM dreams are more perceptually vivid, and more emotional, including 'avoidance' (e.g., avoiding a danger or a threat through fleeing, freezing, or hiding) and approach-like behaviors (e.g., behaviors of exploration and quest for something rewarding, interesting, appetitive, or sexual). In addition, several frontal and parietal regions implicated in executive and attentional functions during wakefulness are significantly deactivated during REM sleep, explaining the disorientation, lack of insight, and impaired working memory typical of REM dreaming. Based on these new data, recent theoretical proposals suggest that dreaming fulfills important functions pertaining to emotion regulation and memory. However, support for the role of dreaming in waking performance is scarce, and more research is needed that would integrate dream content analysis with an assessment of emotion and memory functions. More generally, we suggest that studies combining measurements of brain activity and of concomitant mental activity (dream reports) will provide new insights into the neural determinants of dreaming and thus illuminate an essential facet of human consciousness.

Introduction

Humans have always been intrigued by the nature and varieties of dream experiences. Indeed, the main phenomenological characteristics of the dreams are highly consistent across individuals, including some of the most bizarre dream features. Because they differ or often aberrate from waking experiences, dreams were initially believed to carry messages coming from outside of the individual (from divinities and other spiritual entities) and played a central role in the mythologies of human cultures. In the history of dream research, the first major scientific endeavor to investigate dreaming from a neurological and phenomenological perspective took place during the second half of the nineteenth century. Dreams were not only considered as providing reliable descriptions of what can be perceived, felt, or thought by the dreamer, but also as an essential source of information about the brain processes at work during sleep (Noreika and Windt, 2008; Schwartz, 2000). For example, in his book on sleep and dreaming, Alfred Maury (1862) from the College of France postulated that, contrary to what happens during the waking state, the different parts of the dreaming brain are not synchronized anymore: "perception, memory, imagination, will, judgment are unevenly developed during dreams, indicating different degrees of activity in certain parts of the cerebral hemispheres" (1862, p. 35, our translation). Accordingly, studying dreams would provide unique insights into the varieties of cognitive functions and their underlying neural organization, because of their momentary dissociation during sleep. This period of dream research was characterized by detailed introspective observations, questionnaire studies, and ingenuous experimentation (Lavie and Hobson, 1986; Schwartz, 2000). However, early in the twentieth century, experimental dream research became either dismissed by behaviorist principles (Hillner, 1984) or supplanted

by Freud's psychoanalytical conception of dreams as representing interpretable but disguised fulfillments of repressed wishes, originating from an unconscious mind (Freud, 1900). Note that no systematic empirical evidence supports that dreams result from the transformation of disturbing thoughts, wishes, or fears into symbols that would then need to be decoded. Recent findings instead converge to suggest that dreams and their peculiarities may be considered as relatively unequivocal manifestations of mental and neurobiological processes occurring during sleep. In particular, and as we show below, emotionally relevant memories and concerns, which are highly prevalent in dreams, may be prioritized for reprocessing during sleep. Current predominant models of dreaming therefore reconnect with the late nineteenth-century view that sensory, cognitive, and emotional experiences in dreams, as well as their typical bizarre features, may help specify what type of information is processed by the sleeping brain and how (Schwartz, 2000).

During the second half of the twentieth century, substantial progress in the understanding of the neurophysiology of sleep was made and provided an objective foundation for the scientific study of dreaming (Desseilles et al., 2011; Hobson, 1988; Schwartz and Maquet, 2002). More recently, thanks to the development of advanced neuroimaging techniques yielding measures of regional brain activity during sleep, the study of the relationship between brain function during sleep and dream phenomenology has emerged as an attractive topic of research (Dresler et al., 2011; Horikawa et al., 2013; Schwartz and Maquet, 2002). However, because dreams are introspective descriptions of first-person experiences, linking brain activity to subjectivity is a common staple but uncompleted and difficult task for both empirical dream research and consciousness studies.

Below, we first describe the phenomenological characteristics of the dreaming experience. We then show how distinct cognitive and emotional dimensions in dreams may relate to specific underlying neural correlates, as these are derived from neuroimaging, neuroanatomical, neurophysiological, and pharmacological studies. Based on these findings, we then discuss about the generation and possible functions of dreaming.

Definition and Phenomenology of Dreaming

Existing definitions of dreaming range from 'any mental activity occurring in sleep' to more specific definitions in terms of formed, multimodal hallucinations, strong emotions, and distinctive cognitive features (Pagel et al., 2001; Windt, 2010). In this article, we use a relatively broad definition of dreaming as corresponding to any internally generated sensory, motor, emotional, or cognitive subjective experiences that occur during sleep. As a first-person experience, dreaming is accessible through introspection and is usually reported during the awake state through memory recall. Dream reports may therefore be influenced by several factors such as forgetting, reconstruction mechanisms, verbal description difficulties, censorship, experimental demands, and lack of independent verification. Using appropriate strategies, it is possible to minimize the impact of these factors on dream recall (see Schwartz and Maquet, 2002).

In the 1950s, Nathaniel Kleitman, Eugene Aserinsky, and William Dement first showed that sleep was periodically interrupted by activation of the electroencephalogram (EEG), associated with bursts of rapid eye movement sleep (REM) and increased dream recall (Aserinsky and Kleitman, 1953; Dement and Kleitman, 1957). From then on, dreaming was considered to occur exclusively during the so-called REM sleep stage, and this 'REM = dreaming' equation predominated the field of dream research, despite some early disclaimers (Foulkes, 1962). However, over the past 10 years, this view evolved considerably with accumulating evidence showing that dreaming can occur during REM and non-REM (NREM) sleep (Oudiette et al., 2012). Some phenomenological characteristics may differentiate REM from NREM dreams. In particular, NREM dreams appear generally more thought-like, including real-life concerns, with less aggressive and friendlier interactions of the dreamer with its environment, and seem to involve self-referential processes (McNamara et al., 2005). By contrast, REM dream reports are characterized by vivid imagery, increased and often negative emotionality (e.g., fear, anxiety, anger), behaviors with high motivational value (e.g., sexual interactions), and bizarre, altered, or implausible (compared to waking life) situations (Schredl, 2010). Note however that these global characteristics of dreaming vary over the course of the night, with NREM dreams becoming progressively longer, less thought-like, and more dreamlike and bizarre, thus resembling REM dreams toward the end of the night (Fosse et al., 2004).

Neural Correlates of Dreaming

Neuroimaging Studies

Recent brain imaging methods have provided unprecedented insights into human brain functions, including regional activity during sleep. Yet, scientists face several problems when attempting to interpret neuroimaging data of sleep because of the difficulty in collecting concomitant behavioral measures (Schwartz, 2013). In this context, dream reports represent a major source of information about how cognitive and emotional processes may operate during sleep (Schwartz and Maquet, 2002). Of course, this proposal assumes that dream reports offer reliable and accurate descriptions of sensory, emotional, and cognitive experiences during sleep. Thus, typical features of dream experiences, such as vivid visual imagery, reactivation of memories, emotional intensification, bizarreness, illusion of reality, etc., may inform about underlying brain circuits activated during sleep. Conversely, patterns of activity during sleep provide converging support for the implication of specific brain networks in the generation of dreams (Hobson et al., 2000; Schwartz and Maquet, 2002; Schwartz, 2004). The reciprocal relationship between specific phenomenological characteristics in the dreams and patterns of brain activation during sleep presupposes that the neurophysiological mechanisms underlying cognitive and emotional functions are the same across wake and sleep states. This hypothesis is supported by a recent neuroimaging study in which lucid dreamers activated sensorimotor cortex whenever they signaled that they dreamed of hand movements (Dresler et al., 2011). Moreover, another neuroimaging study showed

that patterns of brain activity recorded while sleeping subjects experienced specific mental images (e.g., categories of objects, scenes, and people) are similar to those elicited by the perception of corresponding visual stimuli during the awake state, and demonstrated that it is therefore possible to predict the contents of visual imagery during sleep based on measured brain activity (Horikawa et al., 2013). Below we provide an overview of the main findings from recent neuroimaging studies of sleep.

Early brain imaging studies reported that, compared to the awake state, NREM sleep was generally characterized by decreases in brain activity across multiple and distributed brain structures (Maquet et al., 1997). Recent studies revealed that transient and strong increases in regional brain activity also occurred during the different stages of NREM sleep. Specifically, during the transition from wakefulness to light sleep (stage N1), bilateral increases in regional glucose metabolism have been observed in the ventral striatum, anterior cingulate cortex (ACC), and the medial temporal lobe, including the amygdala and hippocampus (Nofzinger et al., 2002). Sleep stage N2 is characterized by bursts of oscillatory activity seen on the EEG recording and called sleep spindles. Sleep spindles have been associated with neural plasticity (Diekelmann and Born, 2010). During sleep spindles, brain activity is increased in the thalamus, and in emotion- and memory-related regions such as the ACC, insula, the medial prefrontal cortex (mPFC), and hippocampus (Schabus et al., 2007). Sleep stage N3 is characterized by slow oscillations which have been associated with activity increases in the brain stem, ventral prefrontal cortex, parahippocampal gyrus, precuneus, and posterior cingulate cortex. Thus, each slow oscillation may restore brief wakelike activity patterns (Dang-Vu et al., 2008). Increased activity observed during sleep spindles and slow oscillations occurs in brain areas known to be involved in memory consolidation, consistent with the idea that NREM sleep may promote memory processing. Functional connectivity studies also demonstrated that core subsystems of the brain's 'default network,' typically activated during daydreaming and autobiographical memory retrieval (Buckner et al., 2008), are activated during N1 and N2 sleep stages (Horovitz et al., 2008; Larson-Prior et al., 2009; Laufs et al., 2007), while connectivity with frontal regions is selectively altered particularly during deeper sleep stages (Horovitz et al., 2009; Samann et al., 2011). These stage-dependent changes in the connectivity within the default network may favor the expression of self-reflection, social skills, and 'mentalizing', during NREM sleep, although consciousness might be generally reduced during deep sleep. Electrophysiological recordings during NREM sleep in rodents have documented a spontaneous reactivation (or replay) of neuronal firing patterns in the ventral striatum and hippocampus, which seems to foster the off-line linkage of memory traces with a motivational value (Lansink et al., 2008; Pennartz et al., 2004). Taken together, these studies suggest that brain regions involved in memory replay, motivation, and self-referential processing are activated during NREM sleep, and explain some typical features of NREM dreaming such as the frequent incorporation of current concerns from real-life and self-related thoughts.

On the other hand, regions activated during REM sleep in humans are consistent with some main characteristics of the dreaming experience, usually associated with this sleep stage (**Figure 1**). A first set of activated regions consists in the thalamus, basal forebrain, as well as in limbic and paralimbic structures, including bilateral amygdala, hippocampus, mPFC, and ACC (Braun et al., 1997; Maquet et al., 1996; Nofzinger et al., 1997). Activation of the amygdala may reflect intense emotions, in particular fear, anxiety, and aggressiveness, often experienced in dreams. Amygdala, mPFC, and ACC activity could also subserve emotion regulation processes (e.g., extinction of fear, Hauner et al., 2013; Kleim et al., 2013). Activation of limbic and paralimbic regions suggests that the REM state could favor the reactivation and processing of emotionally salient memories. This pattern of activation would also promote the incorporation of recent memories in dreams, which is commonly observed in dream studies. Other activated regions comprise associative sensory regions and the parieto–temporo–occipital (PTO) junction which could contribute to the generation of sensory and spatial imagery in dreams. Finally, activation of motor circuits is consistent with dreamed motor actions (Braun et al., 1997; Dresler et al., 2011). By contrast, when compared to wakefulness, several regions implicated in executive and attentional functions during wakefulness are significantly deactivated during REM sleep, including the dorsolateral prefrontal cortex, orbitofrontal cortex, posterior cingulate gyrus, precuneus, and the inferior parietal cortex (Braun et al., 1997; Maquet et al., 1996). Reduced activity in these regions may explain the disorientation, lack of insight, illogical thinking, and poor working memory found in REM dreaming. **Figure 1** shows the main anatomical regions whose activity is modified during REM sleep and their plausible contribution to distinctive features in REM dreams.

Lately, a few studies looked at EEG measures preceding awakenings with and without dream recall. The results showed that changes in cortical oscillatory activity may predict successful dream recall: whenever cortical activity was more 'REM-like,' dream recall was more likely after both NREM and REM sleep (Chellappa et al., 2011; Marzano et al., 2011).

This nonexhaustive selection of studies demonstrates that brain imaging methods can be used to investigate the neural substrates of mental processes occurring during sleep. Conversely, future studies are expected to show that dream characteristics may represent biomarkers of important brain functions, such as emotional regulation processes (Desseilles et al., 2011). Thus, research in the domain of sleep in health and disease would benefit from collecting dreams in a systematic manner.

Neuroanatomical Studies

The implication of the functional neuroimaging studies for our understanding of the neural bases of dreaming is limited by the difficulty of obtaining detailed concomitant dream data. In this context, lesion studies (along with pharmacological studies) offer increased specificity regarding the contribution of selective brain regions to distinct dream characteristics. In 1997, Mark Solms established a first classification of dream disorders based on a large study that included 361 neurological and neurosurgical patients (Solms, 1997). Approximately one-third of these patients reported

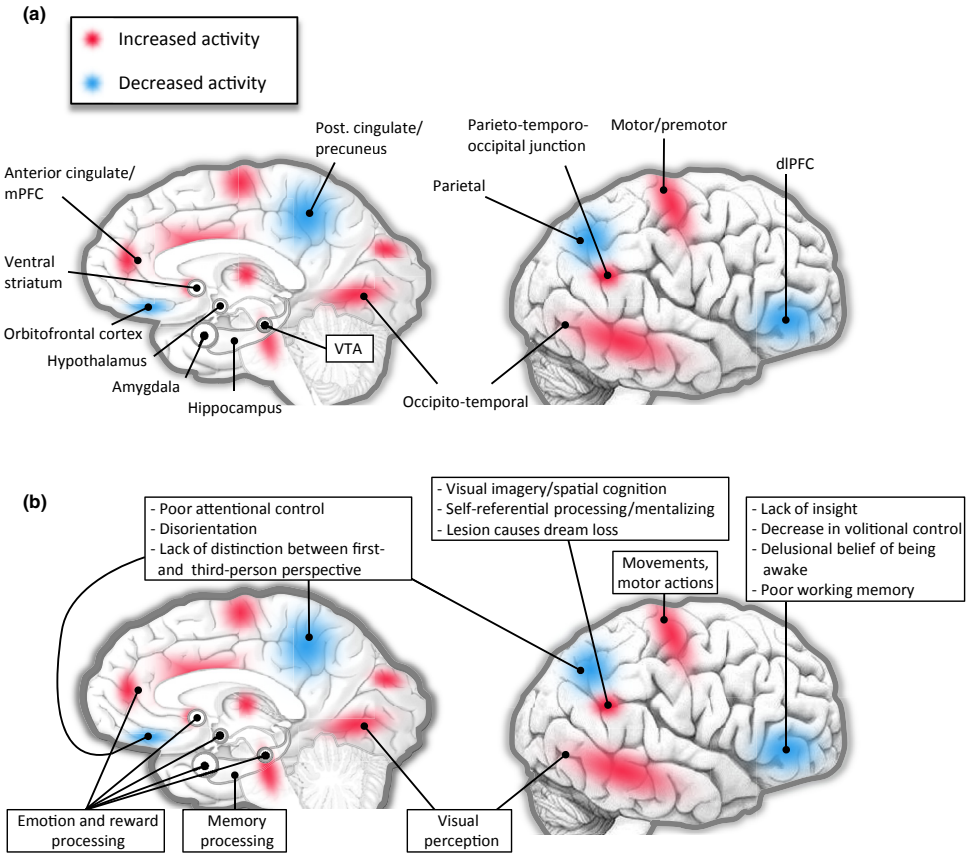

Figure 1 (a) Functional neuroanatomy of REM sleep. Brain regions more activated during REM sleep are shown in red and less activated in blue. Abbreviations: dlPFC: dorsolateral prefrontal cortex; mPFC: medial prefrontal cortex; REM: rapid eye movement sleep; VTA: ventral tegmental area. (b) The specific pattern of cerebral activation/deactivation may explain and be correlated with specific phenomenological characteristics of dreams pertaining to memory, reward and emotional processing, attentional control, visual imagery, delusional beliefs, and mentalizing.

a global cessation of dreaming subsequent to the brain damage, which predominantly involved damage to the PTO junction. Lesion in the PTO region may impact visual imagery and spatial cognition, thus suggesting that these cognitive abilities are necessary for the generation of dreams. Cessation of dreaming was also observed in a few patients with lesions in the white matter surrounding the frontal horns of the lateral ventricles, which can disrupt the connections from the ventral tegmental area (VTA) to the ventral striatum, amygdala, hippocampus, and frontal cortex. These regions form the mesolimbic circuit of the SEEKING system, which is a basic psychobehavioral emotional and motivational system of the mammalian brain, related to approach behaviors and to the feeling of anticipation while seeking a reward (Panksepp, 1998). Note that these lesions do not alter REM sleep, suggesting that REM sleep might not be mandatory for dreaming; instead dreaming would require the activation of the dopamine-driven SEEKING system (Solms, 2000). We should keep in mind that such lesions do not only alter dopaminergic, but also cholinergic and noradrenergic efferent and afferent connections.

Modern brain imaging methods also allow to investigate the neuroanatomy of the intact dreaming brain as well. It is indeed possible to assess the relationships between anatomical measures (e.g., regional volume, mean diffusivity) and quantitative or qualitative measures in dreams. Using this approach, De Gennaro et al. (2011) demonstrated that emotional load and dream length correlated with measures of amygdala microstructural integrity, consistent with the hypothesis that limbic regions may contribute to both the generation of dreams and their emotional content. This methodological strategy, in which brain imaging data are combined with dream data from the same individuals, appears very promising and may significantly improve our knowledge of the neural bases of dreaming.

Neurophysiological and Pharmacological Studies

Recent neurophysiological studies in rodents have provided evidence for the activation of the mesolimbic dopaminergic system during NREM and REM sleep (Dahan et al., 2007; Pennartz et al., 2004; Perogamvros and Schwartz, 2012). More specifically, bursting activity of the VTA, which is the core projecting structure of the reward and the SEEKING systems, is elevated during REM sleep, up to levels observed during waking behaviors, such as feeding, punishment, or sex (Dahan et al., 2007). Besides, it has been also shown that extracellular levels of dopamine in the nucleus accumbens (NAcc)

significantly increase during REM sleep (Lena et al., 2005). In humans, reward-related regions are also activated during REM sleep (Maquet et al., 1996). In addition, the hippocampus exhibits a theta rhythm during REM sleep in both animal and human studies (Cantero et al., 2003; Popa et al., 2010), although human theta activity may be slower and more sporadic compared to the theta in rodents (Clemens et al., 2009). This finding is interesting because increased theta oscillatory activity has been associated with novelty-seeking, exploratory and instinctual behaviors, as overtly expressed in parasomnias. Moreover, orexin neurons in the lateral hypothalamus, which contribute to motivated behaviors and emotional processing, have occasional burst discharges during REM sleep (Takahashi et al., 2008).

Several drugs can induce vivid dreams or nightmares after their administration (e.g., dopaminergic agents, amphetamines, beta-receptor blockers, cholinesterase inhibitors) (Thompson and Pierce, 1999). By contrast, the administration of dopamine antagonists is associated with a reduction in vivid dreaming and nightmares (David et al., 2004; Gaillard and Moneme, 1977; Jakovljevic et al., 2003; Lambert, 2006). Among the several possible mechanisms of pharmacologically induced dreaming, increased dopamine activity seems to be the most frequently involved. Note however that the influence of dopamine agonists or antagonists on dreaming may also be partly mediated by a modulation of REM sleep states (Hobson and Pace-Schott, 1999).

The Generation and Maintenance of Dreams

Activation of memory, reward, and emotional networks, as described above, may contribute to the memory and emotional functions of sleep (Sterpenich et al., 2007; Stickgold, 2005). This claim is further supported by studies showing the deleterious consequences of insomnia and chronic sleep loss on cognitive and emotional functioning (Perogamvros et al., 2013). Conversely, dreams implicate the reactivation of motor, cognitive, or affective memories during sleep (Nielsen and Levin, 2007; Wamsley et al., 2010). Extended afferent projections from the hippocampus, which stores recent memories, to the mesolimbic system would thus result in the activation of the reward and emotional components of the reward system and generate dreaming. Extended mesolimbic reward system activation during REM sleep could explain why this stage is considered as a dream facilitator. Similarly the reactivation of memories during NREM sleep (amygdala, ACC, hippocampus, ventral striatum) would contribute to NREM dreaming. While the exact conditions for a memory trace to facilitate the production of a dream remain unknown, activation of memories with a high emotional and motivational relevance for the individual are plausibly associated with an amplified activation of the dopaminergic system during sleep.

Further support for a role of the mesolimbic dopaminergic system in dreaming comes from dream content analysis studies. For example, 'approach' behaviors in dreams (e.g., behaviors of exploration, curiosity, or engagement) may be as frequent as 'avoidance' behaviors (e.g., avoiding something through fleeing, freezing, or hiding) (Malcolm-Smith et al., 2012). Because approach is a behavior typically mediated by reward-related brain circuits, these observations in dreams thus corroborate the hypothesis that reward-seeking mechanisms are activated during sleep. Importantly, the common occurrence of approach behaviors is not incompatible with the high prevalence of negative affect in dreams (Revonsuo, 2000), because approach/avoidance behaviors and positive/negative emotions are independent psychological entities that engage partially dissociated brain circuits (respectively mesolimbic dopaminergic system and amygdala-limbic system), which are both activated during sleep.

Conclusion and Future Perspectives

Evidence has recently accumulated to show that both NREM and REM sleep contribute to important neurological and psychological processes such as memory consolidation, brain plasticity, emotional regulation, theory of mind, and creativity (Diekelmann and Born, 2010; Perogamvros et al., 2013; Stickgold and Walker, 2013). Moreover, the reprocessing of motivationally and emotionally relevant information during sleep, as demonstrated by both dream content analysis and neurophysiological studies, is consistent with the activation of reward and emotional networks during both NREM and REM sleep. Yet, although it is tempting to relate dreaming to such cognitive, emotional, or other psychological and neurological functions occurring during sleep, whether the subjective experience of dreaming is required or contributes to any of these functions remains mostly speculative. Dream consciousness may offer a neurobehavioral substrate for off-line reprocessing of emotions and for associative learning, thus influencing the development and maintenance of waking consciousness, and improving cognitive and emotional functions during wakefulness (Hobson, 2009; Perogamvros and Schwartz, 2012; Revonsuo, 2000). However, to date, support for the hypothesis of a role of dreaming in waking performance is scarce (Cartwright et al., 1998; Wamsley et al., 2010). Thus, more studies are needed that combine dream content analysis with an assessment of emotion and memory functions, before any strong conclusion can be drawn.

Summary

Current models of dreaming suggest that dreaming can occur during all sleep stages and that it is not exclusively confined to REM sleep. In this article, we review converging evidence from neuroimaging, neuroanatomical, neurophysiological, pharmacological, and dream content analysis studies suggesting that the generation of dreams primarily relies on the activation of limbic/paralimbic brain networks implicated in memory, emotional, and reward processes, as well as parieto-occipital regions related to sensory and visuospatial imagery. Moreover, reduced activity in lateral frontal and inferior parietal regions during REM sleep may explain the disorientation, lack of insight, illogical thinking, and poor working memory found in REM dreams. The distinctive phenomenological features of NREM and REM dreaming are therefore consistent with the patterns of regional brain activation and deactivation observed during sleep. Studies combining measurements of

brain activity and of concomitant mental activity (dream reports) offer a promising avenue for understanding the neural determinants of dreaming, and will thus illuminate an essential facet of human consciousness.

See also: Brain Stimulation Reward; Consciousness, Neural Basis of; Sleep Disorders across the Life Span; Sleep and Health; Sleep and Memory; Sleep: Neural Systems.

Bibliography

Aserinsky, E., Kleitman, N., 1953. Regularly occurring periods of eye motility, and concomitant phenomena, during sleep. Science 118, 273–274.

Braun, A.R., Balkin, T.J., Wesenten, N.J., Carson, R.E., Varga, M., Baldwin, P., Selbie, S., Belenky, G., Herscovitch, P., 1997. Regional cerebral blood flow throughout the sleep-wake cycle. An H2(15)O PET study. Brain 120 (Pt 7), 1173–1197.

Buckner, R.L., Andrews-Hanna, J.R., Schacter, D.L., 2008. The brain's default network: anatomy, function, and relevance to disease. Annals of the New York Academy of Sciences 1124, 1–38.

Cantero, J.L., Atienza, M., Stickgold, R., Kahana, M.J., Madsen, J.R., Kocsis, B., 2003. Sleep-dependent theta oscillations in the human hippocampus and neocortex. Journal of Neuroscience 23, 10897–10903.

Cartwright, R., Luten, A., Young, M., Mercer, P., Bears, M., 1998. Role of REM sleep and dream affect in overnight mood regulation: a study of normal volunteers. Psychiatry Research 81, 1–8.

Chellappa, S.L., Frey, S., Knoblauch, V., Cajochen, C., 2011. Cortical activation patterns herald successful dream recall after NREM and REM sleep. Biological Psychology 87, 251–256.

Clemens, Z., Weiss, B., Szucs, A., Eross, L., Rasonyi, G., Halasz, P., 2009. Phase coupling between rhythmic slow activity and gamma characterizes mesiotemporal rapid-eye-movement sleep in humans. Neuroscience 163, 388–396.

Dahan, L., Astier, B., Vautrelle, N., Urbain, N., Kocsis, B., Chouvet, G., 2007. Prominent burst firing of dopaminergic neurons in the ventral tegmental area during paradoxical sleep. Neuropsychopharmacology 32, 1232–1241.

Dang-Vu, T.T., Schabus, M., Desseilles, M., Albouy, G., Boly, M., Darsaud, A., Gais, S., Rauchs, G., Sterpenich, V., Vandewalle, G., Carrier, J., Moonen, G., Balteau, E., Degueldre, C., Luxen, A., Phillips, C., Maquet, P., 2008. Spontaneous neural activity during human slow wave sleep. Proceedings of the National Academy of Sciences of the United States of America 105, 15160–15165.

David, D., De Faria, L., Lapeyra, O., Mellman, T.A., 2004. Adjunctive risperidone treatment in combat veterans with chronic PTSD. Journal of Clinical Psychopharmacology 24, 556–559.

De Gennaro, L., Cipolli, C., Cherubini, A., Assogna, F., Cacciari, C., Marzano, C., Curcio, G., Ferrara, M., Caltagirone, C., Spalletta, G., 2011. Amygdala and hippocampus volumetry and diffusivity in relation to dreaming. Human Brain Mapping 32, 1458–1470.

Dement, W., Kleitman, N., 1957. Cyclic variations in EEG during sleep and their relation to eye movements, body motility, and dreaming. Electroencephalography and Clinical Neurophysiology 9, 673–690.

Desseilles, M., Dang-Vu, T.T., Sterpenich, V., Schwartz, S., 2011. Cognitive and emotional processes during dreaming: a neuroimaging view. Consciousness and Cognition 20, 998–1008.

Diekelmann, S., Born, J., 2010. The memory function of sleep. Nature Reviews Neuroscience 11, 114–126.

Dresler, M., Koch, S.P., Wehrle, R., Spoormaker, V.I., Holsboer, F., Steiger, A., Samann, P.G., Obrig, H., Czisch, M., 2011. Dreamed movement elicits activation in the sensorimotor cortex. Current Biology 21, 1833–1837.

Fosse, R., Stickgold, R., Hobson, J.A., 2004. Thinking and hallucinating: reciprocal changes in sleep. Psychophysiology 41, 298–305.

Foulkes, D., 1962. Dream reports from different stages of sleep. Journal of Abnormal and Social Psychology 65, 14–25.

Freud, S., 1900. The Interpretation of Dreams. Macmillan, New York.

Gaillard, J.M., Moneme, A., 1977. Modification of dream content after preferential blockade of mesolimbic and mesocortical dopaminergic systems. Journal of Psychiatric Research 13, 247–256.

Hauner, K.K., Howard, J.D., Zelano, C., Gottfried, J.A., 2013. Stimulus-specific enhancement of fear extinction during slow-wave sleep. Nature Neuroscience 16, 1553–1555.

Hillner, K.P., 1984. History and Systems of Modern Psychology: A Conceptual Approach. Gardner Press, New York.

Hobson, J.A., 1988. The Dreaming Brain. Basic Books, New York.

Hobson, J.A., Pace-Schott, E.F., 1999. Reply to Solms, Braun and Reiser. Neuropsychoanalysis 1, 206–224.

Hobson, J.A., Pace-Schott, E.F., Stickgold, R., 2000. Dreaming and the brain: toward a cognitive neuroscience of conscious states. Behavioral and Brain Science 23, 793–842 discussion 904–1121.

Hobson, J.A., 2009. REM sleep and dreaming: towards a theory of protoconsciousness. Nature Reviews Neuroscience 10, 803–813.

Horikawa, T., Tamaki, M., Miyawaki, Y., Kamitani, Y., 2013. Neural decoding of visual imagery during sleep. Science 340, 639–642.

Horovitz, S.G., Fukunaga, M., de Zwart, J.A., van Gelderen, P., Fulton, S.C., Balkin, T.J., Duyn, J.H., 2008. Low frequency BOLD fluctuations during resting wakefulness and light sleep: a simultaneous EEG-fMRI study. Human Brain Mapping 29, 671–682.

Horovitz, S.G., Braun, A.R., Carr, W.S., Picchioni, D., Balkin, T.J., Fukunaga, M., Duyn, J.H., 2009. Decoupling of the brain's default mode network during deep sleep. Proceedings of the National Academy of Sciences of the United States of America 106, 11376–11381.

Jakovljevic, M., Sagud, M., Mihaljevic-Peles, A., 2003. Olanzapine in the treatment-resistant, combat-related PTSD–a series of case reports. Acta Psychiatrica Scandinavica 107, 394–396 discussion 396.

Kleim, B., Wilhelm, F.H., Temp, L., Margraf, J., Wiederhold, B.K., Rasch, B., 2013. Sleep enhances exposure therapy. Psychological Medicine 1–9.

Lambert, M.T., 2006. Aripiprazole in the management of post-traumatic stress disorder symptoms in returning global war on terrorism veterans. International Clinical Psychopharmacology 21, 185–187.

Lansink, C.S., Goltstein, P.M., Lankelma, J.V., Joosten, R.N., McNaughton, B.L., Pennartz, C.M., 2008. Preferential reactivation of motivationally relevant information in the ventral striatum. Journal of Neuroscience 28, 6372–6382.

Larson-Prior, L.J., Zempel, J.M., Nolan, T.S., Prior, F.W., Snyder, A.Z., Raichle, M.E., 2009. Cortical network functional connectivity in the descent to sleep. Proceedings of the National Academy of Sciences of the United States of America 106, 4489–4494.

Laufs, H., Walker, M.C., Lund, T.E., 2007. 'Brain activation and hypothalamic functional connectivity during human non-rapid eye movement sleep: an EEG/fMRI study'–its limitations and an alternative approach. Brain 130, e75 author reply e76.

Lavie, P., Hobson, J.A., 1986. Origin of dreams: anticipation of modern theories in the philosophy and physiology of the eighteenth and nineteenth centuries. Psychological Bulletin 100, 229–240.

Lena, I., Parrot, S., Deschaux, O., Muffat-Joly, S., Sauvinet, V., Renaud, B., Suaud-Chagny, M.F., Gottesmann, C., 2005. Variations in extracellular levels of dopamine, noradrenaline, glutamate, and aspartate across the sleep–wake cycle in the medial prefrontal cortex and nucleus accumbens of freely moving rats. Journal of Neuroscience Research 81, 891–899.

Malcolm-Smith, S., Koopowitz, S., Pantelis, E., Solms, M., 2012. Approach/avoidance in dreams. Consciousness and Cognition 21, 408–412.

Maquet, P., Peters, J., Aerts, J., Delfiore, G., Degueldre, C., Luxen, A., Franck, G., 1996. Functional neuroanatomy of human rapid-eye-movement sleep and dreaming. Nature 383, 163–166.

Maquet, P., Degueldre, C., Delfiore, G., Aerts, J., Peters, J.M., Luxen, A., Franck, G., 1997. Functional neuroanatomy of human slow wave sleep. Journal of Neuroscience 17, 2807–2812.

Marzano, C., Ferrara, M., Mauro, F., Moroni, F., Gorgoni, M., Tempesta, D., Cipolli, C., De Gennaro, L., 2011. Recalling and forgetting dreams: theta and alpha oscillations during sleep predict subsequent dream recall. Journal of Neuroscience 31, 6674–6683.

Maury, A., 1862. Le sommeil et les rêves: Etudes psychologiques sur ces phénomènes et les divers états qui s'y rattachent. Didier, Paris.

McNamara, P., McLaren, D., Smith, D., Brown, A., Stickgold, R., 2005. A "Jekyll and Hyde" within: aggressive versus friendly interactions in REM and non-REM dreams. Psychological Science 16, 130–136.

Nielsen, T., Levin, R., 2007. Nightmares: a new neurocognitive model. Sleep Medicine Reviews 11, 295–310.

Nofzinger, E.A., Mintun, M.A., Wiseman, M., Kupfer, D.J., Moore, R.Y., 1997. Forebrain activation in REM sleep: an FDG PET study. Brain Research 770, 192–201.

Nofzinger, E.A., Buysse, D.J., Miewald, J.M., Meltzer, C.C., Price, J.C., Sembrat, R.C., Ombao, H., Reynolds, C.F., Monk, T.H., Hall, M., Kupfer, D.J., Moore, R.Y., 2002. Human regional cerebral glucose metabolism during non-rapid eye movement sleep in relation to waking. Brain 125, 1105–1115.

Noreika, V., Windt, J.M., 2008. On the importance of 19th century dream research: progress in dream research between Aristotle's work on dreaming and the discovery of REM sleep. Sleep Medicine 9, 919 author reply 920.

Oudiette, D., Dealberto, M.J., Uguccioni, G., Golmard, J.L., Merino-Andreu, M., Tafti, M., Garma, L., Schwartz, S., Arnulf, I., 2012. Dreaming without REM sleep. Consciousness and Cognition 21, 1129–1140.

Pagel, J.F., Blagrove, M., Levin, R., States, B., Stickgold, B., White, S., 2001. Definitions of dream: a paradigm for comparing field descriptive specific studies of dream. Dreaming 11, 195–202.

Panksepp, J., 1998. Affective Neuroscience: The Foundations of Human and Animal Emotions. Oxford University Press, New York.

Pennartz, C.M., Lee, E., Verheul, J., Lipa, P., Barnes, C.A., McNaughton, B.L., 2004. The ventral striatum in off-line processing: ensemble reactivation during sleep and modulation by hippocampal ripples. Journal of Neuroscience 24, 6446–6456.

Perogamvros, L., Schwartz, S., 2012. The roles of the reward system in sleep and dreaming. Neuroscience and Biobehavioral Reviews 36, 1934–1951.

Perogamvros, L., Dang-Vu, T.T., Desseilles, M., Schwartz, S., 2013. Sleep and dreaming are for important matters. Frontiers in Psychology 4, 474.

Popa, D., Duvarci, S., Popescu, A.T., Lena, C., Pare, D., 2010. Coherent amygdalocortical theta promotes fear memory consolidation during paradoxical sleep. Proceedings of the National Academy of Sciences of the United States of America 107, 6516–6519.

Revonsuo, A., 2000. The reinterpretation of dreams: an evolutionary hypothesis of the function of dreaming. Behavioral and Brain Sciences 23, 877–901 discussion 904–1121.

Samann, P.G., Wehrle, R., Hoehn, D., Spoormaker, V.I., Peters, H., Tully, C., Holsboer, F., Czisch, M., 2011. Development of the brain's default mode network from wakefulness to slow wave sleep. Cerebral Cortex 21, 2082–2093.

Schabus, M., Dang-Vu, T.T., Albouy, G., Balteau, E., Boly, M., Carrier, J., Darsaud, A., Degueldre, C., Desseilles, M., Gais, S., Phillips, C., Rauchs, G., Schnakers, C., Sterpenich, V., Vandewalle, G., Luxen, A., Maquet, P., 2007. Hemodynamic cerebral correlates of sleep spindles during human non-rapid eye movement sleep. Proceedings of the National Academy of Sciences of the United States of America 104, 13164–13169.

Schredl, M., 2010. Characteristics and contents of dreams. International Review of Neurobiology 92, 135–154.

Schwartz, S., 2000. A historical loop of one hundred years: similarities between 19th century and contemporary dream research. Dreaming 10, 55–66.

Schwartz, S., Maquet, P., 2002. Sleep imaging and the neuro-psychological assessment of dreams. Trends in Cognitive Sciences 6, 23–30.

Schwartz, S., 2004. What dreaming can reveal about cognitive and brain functions during sleep: a lexico-statistical analysis of dream reports. Psychologica Belgica 44, 5–42.

Schwartz, S., 2013. Complementarity of dream research and neuroimaging of sleep. In: Nofzinger, E., Maquet, P., Thorpy, M.J. (Eds.), Neuroimaging of Sleep and Sleep Disorders. Cambridge University Press, pp. 121–128.

Solms, M., 1997. The Neuropsychology of Dreams: A Clinico-Anatomical Study. Lawrence Erlbaum Associates, New Jersey.

Solms, M., 2000. Dreaming and REM sleep are controlled by different brain mechanisms. Behavioral and Brain Sciences 23, 843–850 discussion 904–1121.

Sterpenich, V., Albouy, G., Boly, M., Vandewalle, G., Darsaud, A., Balteau, E., Dang-Vu, T.T., Desseilles, M., D'Argembeau, A., Gais, S., Rauchs, G., Schabus, M., Degueldre, C., Luxen, A., Collette, F., Maquet, P., 2007. Sleep-related hippocampo-cortical interplay during emotional memory recollection. PLoS Biology 5, e282.

Stickgold, R., 2005. Sleep-dependent memory consolidation. Nature 437, 1272–1278.

Stickgold, R., Walker, M.P., 2013. Sleep-dependent memory triage: evolving generalization through selective processing. Nature Neuroscience 16, 139–145.

Takahashi, K., Lin, J.S., Sakai, K., 2008. Neuronal activity of orexin and non-orexin waking-active neurons during wake-sleep states in the mouse. Neuroscience 153, 860–870.

Thompson, D.F., Pierce, D.R., 1999. Drug-induced nightmares. Annals of Pharmacotherapy 33, 93–98.

Wamsley, E.J., Tucker, M., Payne, J.D., Benavides, J.A., Stickgold, R., 2010. Dreaming of a learning task is associated with enhanced sleep-dependent memory consolidation. Current Biology 20, 850–855.

Windt, J.M., 2010. The immersive spatiotemporal hallucination model of dreaming. Phenomenology and the Cognitive Sciences 9, 295–316.

Drinking, Anthropology of

Thomas M Wilson, Binghamton University, State University of New York, Binghamton, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Johansson, volume 6, pp. 3486–3489, © 2001, Elsevier Ltd.

Abstract

The anthropology of drinking has approached alcohol and drinking in two main ways: as aspects of social health problems, and as part of acceptable, mainstream social behavior. This article examines these approaches in three periods: 1930–60, 1960–90, and 1990 to the present. Today, anthropologists increasingly focus on alcohol and drinking as social and political processes related to culture and identity within local, national, and global contexts.

Over the last century, the study of alcohol and drinking has waxed and waned as a central focus of anthropological research and writing. Drinking, which herein denotes the practices, acts, meanings, and materials related to the ingestion of ethanol (ethyl alcohol, C_2H_5OH), has been approached in anthropology in two main ways: as a social or physical problem or disease, and as a normal or normative practice shared by various members of social majorities and minorities. As anthropology has developed, however, while most other social sciences have concentrated on alcohol and drinking as they figure in social, psychological, and health problems, anthropology has increasingly tended to approach drinking as part of acceptable, mainstream behavior. In the decades prior to the twenty-first century, anthropologists contributed intensively to debates on the benefits of theorizing the drinking of alcohol as a social problem (Heath, 1987a; Hunt and Barker, 2001). In the last two decades, this debate has receded as the anthropology of drinking has converged with growing scholarly interest in culture, identity, consumption, and globalization.

Early Approaches, 1930–60

Anthropologists have studied alcohol and drinking since the beginnings of modern anthropology. Some early writing and research by anthropologists were in response to more general scholarly calls for data from non-Western societies (Glover, 1932; Bunzel, 1940; Horton, 1943; Johansson, 2001). However, as Horton (1943) showed in his examination of anthropological evidence of drinking in 'primitive' societies, there were few early systematic descriptions of ordinary drinking by anthropologists (Johansson, 2001: 3846). Before 1960 few field research projects by anthropologists had drinking as a central focus. In that period, ethnographers mainly paid attention to drinking and alcohol because both were important to the people with whom anthropologists lived and studied. Ethnographers recognized that at least a modest attention to drinking provided a useful tool with which to identify the connections among other important beliefs and behaviors in social and cultural life (Heath, 1987a: 102). It has also been apparent to field ethnographers from the start that drinking is often extremely important in methodological terms because of its roles in the sociality of researchers and hosts, and because drinking often plays an integral and integrative role in many cultures, in much that is important in public life (Kearney, 1970). Ethnographers often get information, and achieve access to wider networks and some measure of trust, through drinking with hosts, and it is often true that drinking facilitates an anthropologist's ability to navigate new social and political fields (Hansen, 1976).

Drinking is not just a methodological tool to field researchers. Today's reflexive turn in ethnographic writing has elicited many more ethnographers' stories of their own personal strengths, weaknesses, needs, and wants in a field research situation (Wilson, 2005b). The upshot of more writing by anthropologists about anthropologists has led to much more discussion of ethnographic practices, including more considerations of trust, confidence, and friendship between ethnographers and their respondents. The importance of drink and other drugs in many of these relationships and encounters is often clear. Ethnographers share drink in public and in more intimate spaces. They participate in ritual and other celebratory events, and sometimes use alcohol as gifts and inducements. Ethnographers also drink for many of the same reasons as their hosts: to enhance relaxation and laughter, to foster friendship and camaraderie, and to express some of the ethnographers' own notions of the self (Wilson, 2005a).

However, despite various references to drinking in a range of anthropological publications, and sporadic acknowledgment of the ubiquitous nature of drinking in anthropological research, alcohol and its attendant behaviors took a backseat to other matters of method and theory in anthropology before and just after World War II.

Transformations, 1960–90

The anthropology of drinking matured in the period between 1960 and the mid- to late 1980s, after which transformations in ethnographic writing and research began to subvert many theories and methods that had been central to anthropology for some time. The post–World War II boom in anthropological research, writing, and teaching produced a wealth of ethnographic studies in economic, political, and symbolic anthropology. These studies were drawn from theoretical approaches that had been gaining in influence since the 1940s, such as structuralism, psychological anthropology, ethnoscience, symbolic anthropology, cultural materialism, and political

economy. By the 1960s, anthropologists were not only making major contributions to specific areas of research on alcohol and drinking outside of anthropology, they were also producing important critiques and syntheses in the history and current state of affairs in the anthropology of drinking (as seen in Mandelbaum, 1965; MacAndrew and Edgerton, 1969; Everett et al., 1976; Heath, 1976). While much of this comparative work addressed the issues of drinking and disease, just as much addressed issues that may be represented by MacAndrew and Edgerton's (1969) consideration of 'drunken comportment.' This approach related drinking behavior and its associated effects to social norms and values, giving them equal weight to the biological impact of drinking.

In this period of transformation, a new balance in anthropological approaches to drinking was struck. Ethnographers began to complement their cross-cultural and biological studies of drinking, excessive drinking, and its effects on individual and group health, with studies of normal, normative, and moderate drinking (Heath, 1987a: 113). This new attention to various roles of drinking within various contexts of culture was due to the increasing awareness among anthropologists of the complex ways in which alcohol and drinking figured in the formal practices that had long been the focus of ethnographers. But ethnographers were now also just as interested in how alcohol and drinking figured in ordinary and everyday life, and in particular how they worked within the webs of significance, which had lately taken center stage in anthropological theorizing. In fact, this change in emphasis in the anthropology of drinking, from studies of alcoholism and alcoholics to those of drinking in everyday life, led inexorably to considerations of many aspects of drinking culture that are the mainstay of anthropology today. These studies viewed alcohol as a commodity, as a social and cultural icon, and as a symbol of group membership. They focused on drinking rituals and other practices related to drinking alcohol, but also sought to tie drinking to many other important aspects of social life, especially those that motivated people beyond the formal roles prescribed by social status and structure. This new attention in the anthropology of drinking formed part of a new wave in what was called 'symbolic anthropology.'

These transformations in theoretical approach reflected changes in anthropological definitions of research populations and the new awareness in anthropology of the importance of economics, politics, and history in the evaluation of society and culture (in a period when 'political,' 'economic,' and 'historical' anthropology were also born). In the 1960s and 1970s, anthropologists moved steadily away from consideration of the primitive to examinations of peasants, proletarians, towns, cities, class, ethnicity, race, and gender. As the complexities of modernization and development increasingly involved anthropologists, former appreciations of relatively autonomous village communities gave way to more nuanced accounts of part-societies, caught up in world forces of historical and contemporary capitalism. And within all of these theoretical interests, anthropologists began to see drinking as a practice worthy of research in its own right.

This new interest was expressed in two mutually inclusive ways. One was in the production of compendia with short case studies or with more synthetic comparative analyses of alcohol, drinking, and culture (Marshall, 1979a). The other was in full-blown ethnographic texts that had drinking as a central focus (as may be seen, for example, in Marshall, 1979b). Drinking was being consistently related to places, spaces, occasions, and many other components of everyday work and leisure, with friends, family, coworkers, and even strangers, at home and in public venues. The focus was no longer only or even principally on non-Western or 'primitive' societies, as may be seen in anthropology's turn to investigating American middle-class life, in, for example, the way in which drinking marks out the temporal and spatial transition from workplace to home (Gusfield, 1987). The specialized places set aside for drinking, such as pubs, were also new sites for anthropological investigation, which in the main showed that these places were nodes within wider social, political, and economic spaces, which crosscut many types and functions of community (Hunt and Satterlee, 1986a,b). The social context to drinking was the cement that held this new anthropology of drinking together, but this focus did not lose sight of the social problems caused by drinking, or of the minority social groups that formed around drinking practices (Spradley, 1970). Central to these new anthropological studies of drinking was the realization that the rules regarding drinking in most if not all cultures involved particularly emotional relations, which often became highly charged, leaving few people indifferent to the many roles in society to which drinking contributed (Heath, 1976).

These transformations in anthropological theory, practice, writing, and profession are perhaps best seen in relation to the publication of a landmark book in the anthropology of drinking, namely Mary Douglas's edited volume, *Constructive Drinking* (1987b). This book, in its introduction by the editor and in its contributors' essays, transcends many of the major shifts that the field of anthropology was undergoing at the time. It linked past theoretical approaches in social and cultural anthropology, such as structural functionalism and structuralism, that were then coming under increasing criticism by younger scholars, to the critiques of ethnography and ethnographic writing that themselves had been born within what had heretofore been called symbolic anthropology. In marrying these old and new approaches, in a seamless way, the book helped to achieve some of the seismic shifts in anthropology that reframed the anthropology of drinking. It helped to situate consumption, identity, and identification squarely within new forms of anthropological inquiry, and to make the anthropology of drinking, if only for a short spell, a cutting edge of anthropology.

In her introduction to the book Douglas (1987a), she reviews the main differences between most anthropologists' accounts of drinking and those of other social scientists. She showed that anthropological data reveal that in most cultures, which regularly imbibe alcohol, and that is in fact most cultures worldwide, regular and repetitive drinking is not necessarily construed as drunkenness or alcoholism. In most cases these actions are not perceived by societal members as indicative of social malfunction, but on the contrary are often signs of strong and supportive cultural relations and values (Douglas, 1987a: 4). Her overarching point was that definitions of drunkenness are culturally framed, and are relative to other practices and values in that culture. Moreover, she went on to show, in conclusions that have influenced much later

anthropological attention to drinking, that most cultures recognize drinking as social acts, which are valued and regularly performed in recognized social contexts. In these persuasive terms, she argued that drinking is a practice of social inclusion and exclusion, marking the boundaries of personal and group identities (Douglas, 1987a: 8–12).

Drinking is thus a principal practice in the social construction of the world as it is and as it should be. At the heart of this approach is the simple notion that for many people in many societies drinking is a benign action, a valued if not socially necessary aspect of formal and informal life. While this does not negate the deleterious effects of alcohol use on the health of individuals and groups who abuse the practice, it does redirect anthropological attention to how alcohol and drinking are often socially meaningful to people, and how these meanings are at the core of their culture. As a result, in this period after 1960, the anthropology of drinking took a major step away from studies of alcoholism and alcoholics to studies of drinking in the ordinary lives of both elites and those without power. In this turn, alcohol was increasingly viewed as a commodity and as a social and cultural icon.

This was a period in which the anthropology of drinking was certainly undergoing noteworthy changes, which its own participants sought to chronicle (Heath, 1987b). Anthropologists continued to focus on drinking rituals and other practices related to alcohol, but did so in ways that tied drinking to other important aspects of social life. This approach clearly demonstrated much that had begun to define symbolic anthropology in the postwar period. It also offered new windows on a growing interest in the field of anthropology overall, namely in the recognition and understanding of social and cultural identities, and the practices that helped to define sameness and otherness. This is one reason why Douglas (1987b) and her collaborators were widely appreciated for their role in integrating older and newer anthropological interests. In this very real sense, as a counterpoint to the chronology offered here, the anthropology of drinking may be approached chronologically in terms of the periods pre- and post-*Constructive Drinking*.

Current Trends, 1990–Present

The generational explosion in research, writing, and teaching in the anthropology of drinking, which greeted the last years of the old and the first of the new century has paralleled global scholarly attention to issues of identity, power, and agency. These issues are in turn integral parts in the scholarly investigation of globalization, neoliberalism, and late capitalism. However, despite the growing numbers of researchers in the anthropology of drinking in this turn of the century transition, and the much heavier attention to theory that is found in a great deal of this recent anthropology of drinking, alcohol is still a relatively minor research focus in global anthropology, and anthropologists are still a minority in the wider field of alcohol studies (Hunt and Barker, 2001; Dietler, 2006). Nevertheless, anthropologists have helped to foreground issues of culture in the comparative social science of drinking, and have gone a long way to combat ethnocentric perspectives in national and international scholarship, especially in regard to perspectives on drinking and pathology that dominate other fields (Dietler, 2006).

The anthropology of drinking today is in many ways a broad reflector of the diversity in theory and method to be found in the related domains of ethnography and anthropology. Many studies of alcohol and drinking address issues of power, inequality, and identity that are at the narrative heart of postmodern writing. But alcohol and drinking also serve as focuses in political-economic analyses of past and present imperialism, colonialism, and postcolonialism (Singer, 1986; Kasmir, 2002; Dietler, 2006). Both of these major trends often approach alcohol as material culture and commodity, and drinking as a practice of support or subversion within wider, rapidly changing, global flows of people, ideas, goods, and services. For example, alcohol as embodied material culture (Dietler, 2006: 231–232) has become an important focus in many strands of anthropology today, including archaeology, which trace commodity chains across the globe and over the years.

Anthropologists have increasingly viewed alcohol as both an agent and a symbol of globalization. In this vein, alcohol and drinking serve as actions and symbols that matter in local, regional, and national affairs, including various forms of social and cultural identification and identity. The anthropology of drinking often sees alcohol as a commodity of symbolic importance in society, but also as an element in economic and social processes of production, distribution, and consumption. It increasingly treats drinking as a public practice that is significant in many other formal and informal behaviors, in international, national, regional, and local arenas. But most significant of all, alcohol and drinking are analyzed by most anthropologists today as elements of group culture and identity.

Alcohol and drinking are meaningful to people, because they say something about people to themselves and others, in processes of social identification that frame if not help to create ethnic, national, class, gender, sexual, and other identities. While drinking is very often a loud and obvious public event involving various forms of social performance, it is also just as often a form of silent language that communicates much about who people are and who they think they are, even when an individual drinks alone in the corner of the room (Turmo, 2001). Drinking alcohol is therefore often a key practice in the expression of identity, and as such it is a building block in the construction of local, national, and global cultures. But drinking is not just an expression of identity. It is a practice through which identity is embodied, performed, and transformed (Dietler, 2006). Anthropologists in greater numbers have begun to see the relationship between drinking and identity as a fruitful field for ethnographic investigation into ways in which local, national, and global identifications emerge and change. A great deal of this work has also benefited from new and more cross-disciplinary communication, much of which is being published in volumes that specifically address issues of culture, history, and identity (de Garine and de Garine, 2001b; Wilson, 2005b; Schiefenhövel and Macbeth, 2011). In fact, the relatively minor field of alcohol studies in anthropology is itself often the subject of historical analysis, testament to its continuing strength as a subject of study within professional anthropology, but also as a means

of interdisciplinary communication and a form of applied research (Marshall, 1984; Heath, 1987a, 2000; de Garine and de Garine, 2001a; Hunt and Barker, 2001; Dietler, 2006).

Throughout this expansion in the anthropology of drinking over the last generation of scholarship, ethnographers have used the full methodological and theoretical toolkits available to them, including qualitative and quantitative methods, participant observation, network analysis, all forms of surveying, and participation in multidisciplinary projects (Heath, 1987a). The benefits of this professional flexibility has been tempered somewhat by other forces that have constrained and molded recent research and scholarship. Some of these constraints come from within anthropology itself.

In their efforts to theorize cultural change in the years after the ethnographic critiques of the 1980s, anthropologists have increasingly found it difficult to define or determine boundaries to culture and society, or to seek cause and effect relationships in culture. Instead, many ethnographers are content to chronicle their own and their respondents' experiences, and to narrate the narratives offered by respondents. This turn in ethnography, which may be considered to have been, and perhaps still is, hegemonic in certain anthropological circles, has affected the anthropology of drinking since the 1980s. In a scholarly environment where ethnographers, as observers or as participants, do not accept that they have the authority to ascertain processes of cause and effect among the groups of people they study, there has been a falloff in attempts to delineate social and cultural groups and institutions, such as, for example, communities, ethnic groups, and nations. While ethnographers are still called upon to conduct studies that may contribute to applied research, in work that is intended to have some policy appeal, they are finding it increasingly difficult to function with analytical categories that are expected to allow better cross-disciplinary collaboration. These categories, such as race, ethnicity, class, and nationality, are still valued in alcohol studies beyond anthropology, and anthropologists' reluctance to use them has resulted in vague and inconsistent social categorization that is often of little comparative significance (Heath, 1987a: 106).

This is an important consideration to anthropologists who seek public funding of their research projects, and is perhaps especially significant to those who strive to understand the relationships between health, disease, and drinking. In the USA, for example, public support of anthropological research on alcohol and drinking has been predominant for projects that see alcohol and drinking as social and medical problems that need cures. Because alcohol is linked to other drugs to which the government is opposed, the public and political war on alcohol and drugs in the USA has led to new attitudes and new policies in regard to alcohol, which in turn has had funding and policy effects, which favor certain approaches to drinking over others. One main result of this funding environment is that it has become more difficult for ethnographers to design projects where they are able to observe and analyze normal and everyday drinking practices, and the roles of alcohol and drinking in ordinary and extraordinary rituals, customs, and social and cultural institutions (Hunt and Barker, 2001: 171). The continuing emphasis on alcohol and drinking, as detrimental to health in certain research and funding quarters, also tends to mask the health benefits of drinking, for individuals and groups (Heath, 2007).

Despite these policy and funding constraints, however, anthropologists continue to study alcohol, drinking, and identity, and to do so at every level of sociocultural integration and differentiation (Wilson, 2005a,b). With a particular emphasis on drinking practices and individual and group agency, contemporary studies in the anthropology of drinking focus on the qualitative and quantitative dimensions of spatial, temporal, behavioral, and aesthetic distinctions in drinking. These distinctions, among others, mark sameness and difference in the processes of social identification (Dietler, 2006: 236).

National identity and identification are among the principal focuses of this research on drinking and identity. While this has been a global anthropological interest, it has been a particularly productive area of scholarship in Europe, in regard to nations that are long associated with alcohol and drinking, whether as a form of national branding (Wilson, 2005a), as a form of national and regional integration (McDonald, 1994a; Demossier, 1997, 2010), or as a form of national definition and resistance (Kasmir, 2002, 2005). In most of these studies, drinking is seen to be the stuff of everyday life and quotidian culture, that at the end of the day may be as important to the life of the nation as are any of its other traditions, values, and practices.

Perhaps the most important and pervasive form of identity associated with alcohol and drinking worldwide is gender (Dietler, 2006), and this has begun to have an important impact in the anthropology of drinking through the growing number of studies of culture, alcohol, and identity among young and old males and females (McDonald, 1994b; Hunt et al., 2005). As a recent review of anthropological and archaeological perspectives on alcohol has indicated (Dietler, 2006), in the last 20 years, complementing studies of national and gender identity, anthropologists who work on how culture and identity intersect with alcohol and drinking have created sizable and provocative bodies of work on politics, political economy, colonialism, and religion.

Today, anthropology continues to infuse approaches to drinking in other scholarly disciplines, particularly due to shared concerns with alcohol and culture, historically as well as in contemporary global and local affairs (Schiefenhövel and Macbeth, 2011). The anthropology of drinking has taken on more significance lately with the growing scholarly interest in all forms of popular culture and leisure activities, such as sports (Mitchell and Armstrong, 2005), but even in these intersections the importance of drinking to widespread and fundamental political, economic, and social values is apparent (Kasmir, 2005). This convergence of alcohol, culture, and politics may best be seen in ways in which anthropologists seek to affect popular impressions of what drinking practices mean and do in contemporary life, as, for example, among college students who deal regularly with the illegal and legal dimensions to binge drinking (Chrzan, 2013). But no matter how far anthropology has traveled from its own historical and methodological roots, it still must deal with the forces of tradition and change that are tied to drinking and culture (Demossier, 1997; Collinson and Macbeth, 2011). It must also deal with the forces of tradition and change in its study of drinking. It is

not surprising then that anthropologists continue to contribute to disciplinary and cross-disciplinary research in alcohol and health, despite the widely held anthropological position that most people drink alcohol without much observable harm, and anthropology as a field is not prepared to offer any immediate or easy solution to alcohol-related health problems (Johansson, 2001). It is also clear that, given the two major trends in contemporary research in the anthropology of drinking, namely those who see alcohol and drinking as problems to be solved and those who see them as cultural practices with various positive and negative roles, there is little chance of achieving a framework for a unified theoretical approach to the anthropology of drinking (Hunt and Barker, 2001).

It may be expected that the anthropology of drinking will grow in size and interest in the near future, due in large part to its recognition as a trope if not a force in processes of identity and identification in wide social scales, including the global. Alcohol is among many commodities that spark intense emotions. As such, it will continue to focus collective scholarly attention on economic, cultural, social, and political globalization. Ongoing debates on the nature and reach of globalization, and its impact on culture and identity worldwide, will ensure increased anthropological research into the production, consumption, and construction of alcohol and drinking.

See also: Alcohol Use among Young People; Ethnography in Applied Social Research; Everyday Life, Anthropology of; Food in Anthropology; Globalization, Anthropology of; Health: Anthropological Aspects; Identity in Anthropology.

Bibliography

Bunzel, R., 1940. The role of alcoholism in two Central American cultures. Psychiatry 3, 361–387.

Chrzan, J., 2013. Alcohol: Social Drinking in Cultural Context. Routledge, New York.

Collinson, P., Macbeth, H., 2011. The thirst for tradition: beer production and consumption in the United Kingdom. In: Schiefenhövel, W., Macbeth, H. (Eds.), Liquid Bread: Beer and Brewing in Cross-cultural Perspective. Berghahn Books, New York.

de Garine, I., de Garine, V., 2001a. For a pluridisciplinary approach to drinking. In: de Garine, I., de Garine, V. (Eds.), Drinking: Anthropological Approaches. Berghahn Books, New York.

de Garine, I., de Garine, V. (Eds.), 2001b. Drinking: Anthropological Approaches. Berghahn Books, New York.

Demossier, M., 1997. Producing tradition and managing social changes in the French Vineyards. The Circle of time in Burgundy. Ethnologia Europea 27, 47–58.

Demossier, M., 2010. Wine Drinking Culture in France: A National Myth or a Modern Passion? University of Wales Press, Cardiff.

Dietler, M., 2006. Alcohol: anthropological/archaeological perspectives. Annual Review of Anthropology 35, 229–249.

Douglas, M., 1987a. A distinctive anthropological perspective. In: Douglas, M. (Ed.), Constructive Drinking. Perspectives on Drink from Anthropology. Cambridge University Press, Cambridge.

Douglas, M. (Ed.), 1987b. Constructive Drinking. Perspectives on Drink from Anthropology. Cambridge University Press, Cambridge.

Everett, M., Waddell, J.O., Heath, D.B. (Eds.), 1976. Cross-cultural Approaches to the Study of Alcohol: An Interdisciplinary Perspective. Mouton, The Hague.

Glover, E., 1932. Common problems in psychoanalysis and anthropology: drug ritual and addiction. British Journal of Medical Psychology 12, 109–131.

Gusfield, J., 1987. Passage to play: rituals of drinking time in American society. In: Douglas, M. (Ed.), Constructive Drinking. Perspectives on Drink from Anthropology. Cambridge University Press, Cambridge.

Hansen, E.C., 1976. Drinking to prosperity: the role of bar culture and coalition formation in the modernization of the Alto Panades. In: Aceves, J.B., Hansen, E.C., Levitas, G. (Eds.), Economic Transformation and Steady-state Values. Queens College Press, Flushing, NY.

Heath, D.B., 1976. Anthropological perspectives on alcohol: an historical review. In: Everett, M.W., Waddell, J.O., Heath, D.B. (Eds.), Cross-cultural Approaches to the Study of Alcohol: An Interdisciplinary Perspective. Mouton Publishers, The Hague.

Heath, D.B., 1987a. Anthropology and alcohol studies: current issues. Annual Review of Anthropology 16, 99–120.

Heath, D.B., 1987b. A decade of development in the anthropological study of alcohol use, 1970–1980. In: Douglas, M. (Ed.), Constructive Drinking. Perspectives on Drink from Anthropology. Cambridge University Press, Cambridge.

Heath, D.B., 2000. Drinking Occasions: Comparative Perspectives on Alcohol and Culture. Brunner/Mazel, Philadelphia.

Heath, D.B., 2007. Why we don't know more about the social benefits of moderate drinking. Annals of Epidemiology 17 (5), S71–S74.

Horton, D., 1943. The functions of alcohol in primitive societies: a cross-cultural study. Quarterly Journal of Studies on Alcohol 4, 199–320.

Hunt, G., Barker, J.C., 2001. Socio-cultural anthropology and alcohol and drug research: towards a unified theory. Social Science and Medicine 53, 165–188.

Hunt, G.P., MacKenzie, K., Joe-Laider, K., 2005. Alcohol and masculinity: the case of ethnic youth gangs. In: Wilson, T.M. (Ed.), Drinking Cultures: Alcohol and Identity. Berg, Oxford.

Hunt, G.P., Satterlee, S., 1986a. The pub, the village and the people. Human Organization 45 (1), 62–74.

Hunt, G.P., Satterlee, S., 1986b. Cohesion and division: drinking in an English village. Man 21, 521–537.

Johansson, G., 2001. Anthropology of drinking. In: International Encyclopedia of the Social and Behavioral Sciences. Elsevier Science, Oxford, pp. 3846–3849.

Kasmir, S., 2002. 'More Basque than you': class, youth and identity in an industrial Basque town. Identities: Global Studies in Culture and Power 9, 39–68.

Kasmir, S., 2005. Drinking rituals, identity and politics in a Basque town. In: Wilson, T.M. (Ed.), Drinking Cultures: Alcohol and Identity. Berg Publishers, Oxford.

Kearney, M., 1970. Drunkenness and religious conversion in a Mexican village. Quarterly Journal of Studies on Alcohol 31 (1), 132–152.

MacAndrew, C., Edgerton, R.B., 1969. Drunken Comportment: A Social Explanation. Aldine, Chicago.

McDonald, M., 1994a. Drinking and social identity in the west of France. In: McDonald, M. (Ed.), Gender, Drink and Drugs. Berg, Oxford.

McDonald, M. (Ed.), 1994b. Gender, Drink and Drugs. Berg, Oxford.

Mandelbaum, D.G., 1965. Alcohol and culture. Current Anthropology 6, 281–294.

Marshall, M. (Ed.), 1979a. Beliefs, Behaviors, and Alcoholic Beverages: A Cross-cultural Survey. University of Michigan Press, Ann Arbor.

Marshall, M., 1979b. Weekend Warriors: Alcohol in a Micronesian Culture. Mayfield Publishing, Palo Alto, CA.

Marshall, M., 1984. Alcohol and drug studies in anthropology: where do we go from here? The Drinking and Drug Practices Surveyor 19, 23–27.

Mitchell, J.P., Armstrong, G., 2005. Cheers and booze: football and *festa* drinking in Malta. In: Wilson, T.M. (Ed.), Drinking Cultures: Alcohol and Identity. Berg, Oxford.

Schiefenhövel, W., Macbeth, H. (Eds.), 2011. Liquid Bread: Beer and Brewing in Cross-cultural Perspective. Berghahn Books, New York.

Singer, M., 1986. Toward a political-economy of alcoholism: the missing link in the anthropology of drinking. Social Science and Medicine 23 (2), 113–130.

Spradley, J.P., 1970. You Owe Yourself a Drunk: An Ethnography of Urban Nomads. Little, Brown, Boston.

Turmo, I.G., 2001. Drinking: an almost silent language. In: de Garine, I., de Garine, V. (Eds.), Drinking: Anthropological Approaches. Berghahn Books, New York.

Wilson, T.M., 2005a. Drinking cultures: sites and practices in the production and expression of identity. In: Wilson, T.M. (Ed.), Drinking Cultures: Alcohol and Identity. Berg Publishers, Oxford.

Wilson, T.M. (Ed.), 2005b. Drinking Cultures: Alcohol and Identity. Berg Publishers, Oxford.

Drucker, Peter (1909–2005)

Winfried W Weber, Mannheim Institute of Applied Management Research, Mannheim University of Applied Sciences, Mannheim, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Peter F. Drucker was one of the first academics to analyze a phenomenon scarcely understood before his time, namely the mass expansion of the management profession, and to set out the tasks of this new discipline in so many words, the first to formulate them, then. In 1954 he sketches the tasks of a manager in five points: a manager sets up goals, a manager organizes, a manager motivates and communicates, a manager assesses results, and a manager develops and brings people on, not least themselves. Practitioners of management nowadays talk about *management by objectives*; they direct *profit centers*, motivate *knowledge workers*, *decentralize* structures, *do things right* (*efficiency*) and *do the right things* (*effectiveness*), or improve *customer orientation*, all concepts from Drucker's mental laboratory.

Peter F. Drucker's views about business, management, and social sciences are American in origin, though multidisciplinary by his European roots and upbringing. With a broad background, Peter F. Drucker's understanding of law, history, art, and all the liberal art disciplines made him a prominent management thinker. Drucker always maintained his European perspective on the world, one that understood that no one culture has all the answers (Handy, 2010). The Vienna-born (1909) Drucker, a professor of management at the Graduate Business School of New York University from 1950 and a professor of social sciences and management at Claremont Graduate University from 1971 to 2002, had a huge influence mostly on management practice. His wide influence manifests itself in the fact that he is not only well known in America and Europe but just as well in Asia. And his ideas continue to exercise a huge influence on the business world right up to today. Before his academic career Drucker worked as an investment advisor, journalist (business and foreign affairs), securities analyst, and management consultant. With his personal experiences on totalitarianism in 1933 escaping Hitler's Germany, he saw management mainly as a social function and an essential contribution to a free society. For Drucker, a free society needs strong institutions (Drucker, 1939). The alternative to autonomous institutions that function and perform is not freedom for Drucker, it is totalitarian tyranny (Drucker, 1973). And high-performing and self-directed institutions depend on excellent management in all sectors of society, in the for-profit, nonprofit, and public sector. Drucker's ideas have been dispersed in his 39 books from 1939's *The End of the Economic Man* to *Managing in the Next Society* published in 2002. Drucker had a profound empirical attitude while meeting with ordinary management practitioners and consulting them. He became an advisor to many significant twentieth-century leaders from all sectors of society in America, Europe, and Asia. Drucker was one of the first scientists to analyze a phenomenon scarcely understood before his time, namely the mass expansion of the management profession, and to set out the tasks of this new discipline in so many words, the first to formulate them, then. Drucker developed management into an application-oriented discipline, the basis for which he outlined in the work *Practice of Management* where he points out that, rarely in history has a new institution proved itself so indispensable in so short a time. And for Drucker it still is rarer that a new institution capable of establishing itself "with so little opposition, unrest and controversy" (Drucker, 1954: p. 4).

In his book *The New Realities*, he offered a definition of management as a liberal art: "Management is thus what tradition used to call a liberal art – 'liberal' because it deals with the fundamentals of knowledge, self-knowledge, wisdom, and leadership; 'art' because it is practice and application. Managers draw on all the knowledges and insights of the humanities and the social sciences – on psychology and philosophy, on economics and history, on the physical sciences and ethics. But they have to focus this knowledge on effectiveness and results – on healing a sick patient, teaching a student, building a bridge, designing and selling a 'user-friendly' software program" (Drucker, 1989: p. 231).

Management as Discipline

In his early work Drucker's most important contribution was to develop management as an independent subject area to be treated as sovereign unto itself. He described what were still at that time the initial innovations in business studies as insufficient for the tasks of leadership. He coined many terms and concepts for the new mass profession of manager. Many of them have today gone over into everyday language use. Among others, the terms, 'profit center,' 'decentralization,' 'knowledge worker,' and 'management by objectives' originate with him.

In spite of his down-to-earth-approach, Peter Drucker never himself worked as a manager or as an employee of a large firm. He styled himself a 'bystander,' one who looks on. He observed organizations and managers from a distance and independently. He remained an independent thinker and was more of an intellectual than an academic. (Drucker declined even chairs from Harvard and Stanford for having more freedom and less academic obligations.) For Drucker management was less a form of scholarship than it was, first and foremost, practice. In the discipline of management what matters is working out *what needs to be done*. Drucker became a philosopher of management who went on to pursue practical philosophy (Weber, 2009).

As an experienced journalist, by the 1940s Drucker knew full well that he had to develop completely new concepts and

lines of inquiry for his project of formulating the discipline of management. And in addition, he was quite aware that, if he wanted to convince the practitioners of management, he had to write precisely and intelligibly. The great challenge for authors and consultants in management consists in formulating the discipline's contradictions and paradoxes so that practitioners, when they look for advice, are perhaps a little confused but never overly. And that goes for managers too. When they make decisions, their organization observes these decisions. Decisions do not actually exist in the heads of those doing the deciding. They are only effective, if they are perceived as communicative events, which have to come so nice and tightly packaged that the contradictions and the complexity underlying the decisions cannot be recognized, or scarcely so, and subsequent decisions can follow on from them (cf Luhmann, 2002). The art of management is to employ language adroitly. The art consists in seeing complexity, making what is undecidable decidable and blunting the edge of possible critics of current decisions, who may be sensing pure arbitrariness, with ready, clear, precise, and, as far as possible, rational justifications. Drucker saw himself in the tradition of Chester Barnard (1886–1961), who was one of the first to tackle practical philosophy and organizational theory and posed large-scale questions going beyond the initiatives from Frederick Taylor and Henry Fayol in business studies (Drucker, 1954: p. viii). Barnard defended himself by differentiating between two ways of talking about management: abstract-theoretical and generally comprehensible-practical. He maintained that he needed both styles for his activities and could only understand what he was doing in practice through his theory. For Barnard, managing meant acting intuitively, without being able to say what knowledge was guiding you (Barnard, 1938).

In Drucker's first mag*num opus*, *The Practice of Management*, he gives one section the title of *A Philosophy of Management* (Drucker, 1954). Drucker's method became the Socratic dialogue and large-scale questions. He put the practical philosopher's question to managers: *What would you do, if* …? With managers, he discussed the question of how they should act and how they should reflect on their everyday activity.

In Drucker's approach management is an institution and a source of innovation in society, where managers can take risks in the sense of an organization's goals. That is because there is no longer any center in society, where all the weak signals from the surroundings could be collated and decisions made under clear-cut conditions. Organizations and management, however, observe quite specific differences and decide with regard to their actually defined performance. They develop a particular, potentially innovative product and duly throw another one out of the program. They offer a new therapeutic procedure and forego another one. They rehearse a new piece for the theater and discontinue another one.

In premodern societies, despots could still determine, for example, religious allegiance or the value of currency, or they could cut swathes across a full-grown city and force urban boulevards through. In a society where everyone knew their place, the great simplifiers could still create circumstances that were clear-cut. Today, there is no reality that all citizens share. In management, it is not a question of optimization any more but of finding a balance. Organizations have become complex systems, which have to try to advance themselves in a complex environment, and they do, for their part, contribute to society becoming ever more differentiated. This is where Peter Drucker's pioneer work comes in. He was originally an intellectual and a social scientist, and it was almost by accident that he came to engage intensively with organizations and the running of them. Drucker began to observe the practice of management and to set it out from a historical perspective. He accompanied managers in their everyday routines, as, for example, Alfred Sloan and the management of General Motors, whose behavior when making decisions he studied for 2 years in the early 1940s. Drucker worked out the essential elements of this new mass profession and looked beyond the surface presented by 'charismatic' leaders in firms. He was concerned with formulating the essence of the discipline, with contributing to it and making it something you could learn – and perhaps teach. His questions went deeper than those of his contemporaries. What accounted for the success of this particular person, who had the reputation of being charismatic? What tasks have they accomplished? What tools can they wield? What are their basic working principles? Drucker was one of the first thinkers to analyze the proliferation of business leaders en masse – until then a social phenomenon scarcely understood – and to set down the tasks of this new discipline in words, the first to formulate it. In 1954, he wrote about the new mass phenomenon of management: "Rarely, if ever, has a new basic institution, a new leading group, emerged as fast as has management since the turn of this century. Rarely in human history has a new institution proven indispensable so quickly; and even less often has a new institution arrived with so little opposition, so little disturbance, so little controversy" (Drucker, 1954: p. 3).

Knowledge Worker's Productivity

Peter Drucker's suggestions for management as a liberal art aim to connect up two not easily integrated elements of organizing: *optimizing* organizations as places, where people make decisions as rationally as possible and promote the development of organizations in terms of efficiency and effectiveness, and *motivating* their members, something that succeeds particularly well when they are granted an independent existence with enough room for self-organizing. If the way management works emphasizes calculations and techniques, then the second element, the social system, is often ignored, and people become mere employees working to rule. Drucker's simple question, "how do we increase a knowledge worker's productivity?" harbors some of the tricky questions that modern management has to tackle. Raising productivity among knowledge workers does indeed depend on individuals, but they can only apply knowledge cooperatively and in organizations (Drucker, 1994). There are chances proliferating out of that for every practitioner, whether manager or knowledge worker, but also totally new questions and problems. Each and every member of a knowledge-based organization will have to learn how to explain to others – who lack the same basis of knowledge – what makes this specialized knowledge so significant. Everyone will be personally responsible for their

knowledge, making a contribution to organizations, and even, in fact, to society at large.

That causes the basic principles of management to shift. In Drucker's *next society* (Drucker, 2002) the challenge will be that individuals, groups, and knowledge workers can explain to others what they are about. We will observe management to see if it employs language people can understand. And beyond that, knowledge workers, clients, and interest groups are continually wanting proof that managers are listening to them. Alongside what has become a matter of course – the job of running organizations effectively – more and more depends on whether you cultivate a communicative style of leadership. Working in an organization of knowledge workers means being able to communicate (cf Baecker, 2003).

The Practice of Management

In 1954 Drucker wrote the first 'true management book' (Drucker, 1954: p. viii) and looked at management as a whole, as a distinct function, and as a specific work. He sketches the tasks of a manager in five points: a manager sets up goals, a manager organizes, a manager motivates and communicates, a manager assesses results, and a manager develops and brings people on, not least themselves.

Management by Objectives

Drucker's fundamental principle for management is concentration! Management has to clarify goals. He invented the concept *management by objectives*, which may be Drucker's most well-known management paradigm. And that does not, in fact, mean the classical variant of a falsely understood *management by objectives*, dealing in causes and effects, but management in a form capable of creating 'a universe of pattern, purpose, and process' (Drucker, 1959). If, as a manager, you are not constantly asking about the results your organization wants to achieve, then nobody knows what has primacy.

Drucker's suggestion on this is predictably simple: management repeatedly prepares organizations for new challenges by asking repetitively large-scale questions. The focus is secondly on the commercial index-number, the incremental improvement, or the next reorganization-project. Firstly management has to regularly pose the apparently simple question, *what is our mission, how can we create a customer etc.*? In this sense, management is not leadership qua direction by individuals, but the shaping and steering of entire organizations in their environment. Management does not mean leadership for your direct subordinates but for all members of an organization. And when the organization has reached a certain size, management can only lead by indirect influence and scarcely has sufficient information anyway. Management achieves joint performance by making the survival of an organization the overriding goal. Profit in an organization as a social system is only a means to this end.

Drucker's famous anecdote illustrates the importance he places on self-management when three stonecutters were asked what they were doing. The first replied: "I am making a living." The second kept on hammering while he said: "I am doing the best job of stonecutting in the entire county." The third one looked up with a visionary gleam in his eyes and said: "I am building a cathedral" (Drucker, 1954: p. 122). The management challenge consists in leading ordinary people to perform extraordinarily.

Management Means Communicating

Managers do all they can to ensure that employees, who have all become problem-solvers in modern organizations, get a hearing for their suggested solutions in a culture of open discussion. Managers have, therefore, to reduce the intervals in hierarchies. They have to tear down the walls between *white-collar* and *blue-collar* workers. Organizations should be only interested in 'creating a customer.' Every communication in them must aim at this goal. That means every function in firms changes; every member of the organization supports those who are currently achieving the results in firms. And this will only succeed if it gets communicated and understood. In an organization that communicates, no one person is privileged any longer as the only one people report to. For Drucker the system must serve the worker in a knowledge worker workforce (Drucker, 2002). If someone says about problems, "it's nothing more than a communications problem," then people need to prick up their ears. Communication has become the hardest of all problems.

Systematic Abandonment and Change

Old habits, long-tolerated weakness, accumulated mementos, or deeply ingrained routines can threaten organizations' survival. The messy-syndrome, the love of disorder, only appears relatively rarely among people, but for organizations it is almost a typical phenomenon. Continuously abandoning things is a permanent task which has to be organized, so someone has to take responsibility for it. For Drucker, abandonment becomes a starting point for processes of change and innovative solutions (Drucker, 2008).

Drucker's famous question to Jack Welch, a former CEO of General Electric "if you weren't already in the business, would you enter it today?" had radically changed his company. Every organization which asks itself this question every 3 months and bases all its decisions on it – that is the advice of Drucker – will change beyond recognition. Every organization only survives long term if its results head the field, if everything is taken care of perfectly. Mediocre work has to be abandoned, as otherwise self-confidence within an organization starts eroding.

Yet Drucker also recommends that managers have to ask details about which processes are too fat, about where, going completely on value-stream, *leaner* organization is possible. What do you have to leave out, which processes contain superfluous elements, where is too much being produced, where are there waiting times, unnecessary trips, high levels of stock, and repeated mistakes?

The next management has less and less time to recognize what needs removing. What means success today, will perhaps have to be thrown out tomorrow.

Strengths and Weaknesses

In a society of organizations based on knowledge, everybody starts out at zero. Everyone starts out with the same 'total

ignorance.' When senior executives promote their people, challenge them, and encourage them, they put three Drucker questions: What are your strengths and what should you put work into? What should this company expect from you and in what time span? And what information do you need to do your work and what information do you owe? Managing knowledge workers means devoting time to them; understanding their work and what makes them tick; and building up a personal relationship with them, to mentor them and to listen to them, to challenge them and to encourage them. For managers, it is always a question of reaching employees and contributing toward them being able to grow (Drucker, 2002).

Entrepreneurship and Innovation

The corporation as we know it, will not survive in this century, Drucker predicts. "Legally and financially, yes, but not structurally and economically" (Drucker, 2002: p. 68). In stable times, when developments can be expected to a certain extent, it is sufficient to make things better. In times of 'discontinuity,' however, it is a question of 'doing things differently' (Drucker, 1969). All organizations have to live with the concomitant risks. Defending the past – not innovating – is much more risky than shaping the future.

Rerum novarum cupidus, greedy for something new, is how a Roman poet described human nature. But how do new things come into the world? For Drucker, it is not through the 'typical entrepreneurial personality.' He maintained he had not come across any such personality. Entrepreneurial activity creates new things through systematic procedures. Firstly, when entrepreneurs conduct meetings, there is always a clear division between two perennial items for discussion: the problems needing solutions and the opportunities not to be missed. Secondly, people in entrepreneurially managed organizations take notice of the staff who perform better and ask them, "how do you do that?" Thirdly, entrepreneurs get together at certain times with the next generation from all sectors and ask them about their ambitions, opportunities, liabilities, and their views about the firm. Where should entrepreneurs enter new territory, where develop something new, where dispense with what is obsolete? The innovation process also has to take place with people involved from outside the firm, with partners, clients, and the social environment.

Managing and Networks

Peter Drucker always understood the world of management as a social system and as a profession he could influence by communication and networking. The profession of management is based on communication. Managers are communicating for most of their time (Mintzberg, 2009) and building human relations, and that naturally means networks. If they fulfill this central task well, then a climate of trust grows and organizations are able to handle more complex tasks.

For Drucker management innovations will arise particularly through combining management ideas from different sectors of society. Being receptive brings about complexity and variance in management. Drucker's experiences with nonprofit organizations have had become a central source for his ideas. The social function management has is not something we can reduce to processes of optimization. Accordingly, management would also comprehend social ecology: management as a 'bridge between civilization and culture' (Ueda and Isaka, 2010). For Drucker, social ecology means observing phenomena from societies and communities and from business administration at one and the same time. What motivates a knowledge worker does not depend only on job satisfaction. Committing to public service and taking on social responsibilities would be a decisive factor in managers attaining personal maturity. What every organization can learn from the volunteer organizations is their ability to attract and to retain members. In Drucker's management approach organizations have to accept the fact that you have to treat your staff as 'volunteers.' Workers carry their ideas in their heads and could go to any other organization the next day. Because Peter Drucker himself was so excellently connected in management, he saw phenomena only hubs can see. In clusters, from a bird's eye perspective, as an observer of management, he foresaw many developments in organizations and in society. Drucker was a second-level observer. He would observe managers. He would observe advisers as they observed managers. He would observe the colleagues of management thinkers, as they observed managers and advisers. He was well aware of his position as a management network hub. Drucker researched into and advised all three sectors of organizations: profit-oriented, nonprofit, and public organizations. He knew many influential managers of large organizations personally, in all the important industrial nations. And he was in daily contact with the next generation of leaders, who came to him in streams from all over the world, with the students of the Claremont Graduate School and the young managers, advisers, and scholars, who visited him at home until he passed away in 2005.

Management Ethics

For managers, it is always a matter of being a role model too. What counts is resisting the enticements of status and resisting privilege, being aware of the responsibility toward people, and understanding themselves as part of a social system. Drucker despised that parasitical trait some top managers display and advocated spurning the 'golden parachutes' of payoffs in millions. Management is a young profession, which has so far been brought in without much resistance in society. If modern society wants to continue the success of this function, then Drucker claims to develop a minimal code of ethics in the management profession, one which can, for instance, borrow from the Hippocratic oath: *Primum non nocere*, above all things, do no harm to anyone.

Managing Complexity

Peter Drucker's writings, lectures, and epigrams offer a way to tolerate the paradoxes of management's everyday life and the lack of oversight. He helped both to reduce complexity and to increase it (Weber and Kulothungan, 2010). Drucker took trouble to test his every ideas and concepts to see if they worked. His disciples confirmed that he tried to use only words and terms every manager could understand without needing a dictionary. His oeuvre intended to be linked directly to the

practice of management. Economic scientists ignored him a long time and still do because of his inductive and application-orientated approach. Management designed as an application-orientated discipline helped to widen the view of practitioners and helped them to trust themselves. He encouraged them to define management as a communication system where observers observe how they are observed. Drucker assisted managers to vary their actions and to adjust their expectations in permanent self-assessment, constant resharpening, constant refocusing, and in never being really satisfied (Drucker, 2008). He backed managers to deal with complexity in a *non*strategic manner. He encouraged them not to analyze everything and spend much time on long-term planning. Drucker taught managers to look behind simple cause-and-effect chains and showed the blind alleys organizers are running in (Weber, 2007). Instead of trends and belief in progress, Drucker was interested in the practice of innovation. Even if he correctly foresaw several future developments, Drucker denied to predict and had the idea of a future that does not just happen, but instead is designed today to make for a different tomorrow. Drucker's advice that strategy and operational planning have nothing to do with future decisions. It is formulated *right now* and so it is also only *right now* that it can be correct or false: false or correct as regards a future envisaged in the now. Future is always unpredictable but today's decisions will impact the future (cf Malik, 2010; Baecker, 2010). Managers in the age of complexity are innovators in finding and bridging gaps where others see only labyrinths. 'Managing in the next society' means to be fascinated and challenged by complexity. In an era of uncertainty managing means to act always as 'to increase the number of choices' (Foerster and Poerksen, 2003).

Conclusion

We have Drucker to thank for the admonition about not confusing the art of management with a science. Management repeatedly prepares organizations for new challenges by asking large-scale questions. It encourages workers doing a job either perfectly or not at all, because that is the only way core-values in the organization can come about and can help an organization as a social system to survive. Profit is only a means to this end. Drucker's oeuvre consists a paradoxical intervention: heralding the new and advocating communication and repetition (Baecker, 2010). Who could argue against innovating? Who could dispute that repetition and unlimited patience is necessary for management communication if the organization wants to innovate? And who can deny, that you need managers with tolerance for ambiguity in an organization that has been founded for exercising routines?

In the age of complexity, Peter Drucker's *next management* has to try to strike the right balance between investment and innovation. Every objective, every plan-of-action an organization develops must be formulated *right now*, and so they can also only be right or wrong *right now*. Every philosophy of management, which puts too much stress on innovation for its own sake, may indeed make the pendulum of management trends swing the other way, but it does run the risk of advocating *newness for newness' sake* and risks recommending actions that underestimate pressures from the immediate present. That causes the basic principles of management to shift. The next management will be talk, decision, and action (cf Brunsson, 1989). Drucker characterized this shift to talk that management employs language people can understand. And beyond that, knowledge workers, clients, and interest groups are continually wanting proof that managers are listening to them. The other shift is that management decides in a way to increase the number of choices. And last but not least this shift is a bias for action and a permanent redesigning of planning (Drucker, 1973). The job of running organizations effectively more and more depends on whether a communicative style of leadership is cultivated. Managing and working in the *next society* is for Drucker being able to communicate. Peter Drucker's management philosophy helps managers to widen their view and to trust themselves. He helps them to define management as a liberal art and a communication system where observers observe how they are observed. Drucker encourages managers to vary their actions and to adjust their expectations in permanent self-assessment, the "first action requirement of leadership: the constant resharpening, constant refocusing, never being really satisfied" (Drucker, 2008: p. 5). Drucker helps us to deal with complexity in a *non*strategic manner. Managers do not have to analyze everything and spend much time on long-term planning. They can only solve the problems of today. Drucker teaches us to look behind simple cause-and-effect chains and shows us the blind alleys we are running in (Weber, 2007).

With this kind of management approach professional managers are able to develop a rich and complex communication stage where observations are ambivalent. Managers in the 'next society' are innovators in finding and bridging gaps where others see only labyrinths. Managers in the 'next society' are fascinated and challenged by complexity, they 'act always as to increase the number of choices' (Foerster and Poerksen, 2003).

See also: Administration in Organizations; Entrepreneurship and Organizations; Information and Knowledge, Organizational; Leadership in Organizations, Sociology of; Management: General; Managerial and Organizational Cognition; Organizational Behavior, Psychology of; Organizational Control; Organizational Decision Making; People in Organizations; Workplace Autonomy.

Bibliography

Baecker, Dirk, 2003. Organisation und Management. Suhrkamp, Frankfurt.
Baecker, Dirk, 2010. In: Weber, Winfried W., Kulothungan, Gladius (Eds.), Peter Drucker's Paradoxical Intervention, pp. 183–190.
Barnard, Chester, 1938. The Functions of the Executive. Cambridge, MA.
Brunsson, Nils, 1989. The Organization of Hypocrisy. Talk, Decisions, and Actions in Organizations. John Wiley, Chichester.
Drucker, Peter F., 1939. The End of Economic Man – the Origins of Totalitarianism. John Day, New York.
Drucker, Peter F., 1942. The Future of Industrial Man. John Day, New York.
Drucker, Peter F., 1946. Concept of the Corporation. John Day, New York.
Drucker, Peter F., 1954. The Practice of Management. Harper & Row, New York.
Drucker, Peter F., 1957. The Problems of Maintaining Continuous and Full Employment, Speech at the 11th International Management Congress, Paris. Reprinted in: 2010. The Drucker Lectures. McGrawHill, New York.
Drucker, Peter F., 1959. Landmarks of Tomorrow. Harper & Row, New York.
Drucker, Peter F., 1964. Managing for Results. Harper & Row, New York.

Drucker, Peter F., 1966. The Effective Executive. Harper & Row, New York.
Drucker, Peter F., 1969. The Age of Discontinuity. Harper & Row, New York.
Drucker, Peter F., 1973. Management. Task, Responsibilities, Practices. Harper & Row, New York.
Drucker, Peter F., 1978. Adventures of a Bystander. Harper & Row, New York.
Drucker, Peter F., 1980. Managing in Turbulent Times. Harper & Row, New York.
Drucker, Peter F., 1985. Innovation and Entrepreneurship. Harper & Row, New York.
Drucker, Peter F., 1989. The New Realities. Harper & Row, New York.
Drucker, Peter F., 1990. Managing the Non-profit Organization. HarperCollins, New York.
Drucker, Peter F., 1993. Post-Capitalist Society. HarperCollins, New York.
Drucker, Peter F., November 1994. The age of social transformation. Atlantic 274 (5), 53–89.
Drucker, Peter F., 1999. Management Challenges for the 21st Century. HarperCollins, New York.
Drucker, Peter F., November 3, 2001. The next society: a survey of the near future. The Economist. http://www.economist.com/node/770819.
Drucker, Peter F., 2002. Managing in the Next Society, New York, 2002. St. Martin's Press, New York.
Drucker, Peter F., 2008. The Five Most Important Questions You Will Ever Ask about Your Organization. John Wiley & Sons, New York.
Foerster, Heinz von, Poerksen, Bernhard, 2003. Understanding Systems. Conversations on Epistemology and Ethics. Heidelberg, New York.
Handy, Charles, 2010. In: Weber, Winfried W., Kulothungan, Gladius (Eds.), What Drucker Taught Me, pp. 59–63.
Hesselbein, Frances, Spring 2006. Peter Drucker's line shines on. Leader to Leader Journal 40, 4–7.
Luhmann, Niklas, 2002. Organisation und Entscheidung. Opladen.
Malik, Fredmund, 2010. In: Weber, Winfried W., Kulothungan, Gladius (Eds.), Peter F. Drucker. The Discoverer of Management, pp. 33–43.
Mintzberg, Henry, 2009. Managing. Berrett-Koehler, San Francisco.
Simon, Herbert A., 1946. The proverbs of administration. Public Administration Review 6, 53–67.
Ueda, Atsuo, Isaka, Yasushi, 2010. In: Weber, Winfried W., Kulothungan, Gladius (Eds.), Management as a Bridge between Culture and Civilization: A Philosophical Framework for the Next Century, pp. 202–215.
Weber, Winfried W., 2007. Complicate Your Life. Sordon, Göttingen.
Weber, Winfried W. (Ed.), 2009. Peter Drucker – der Mann, der das Management geprägt hat. Erinnerungen und Ausblick zum 100. Geburtstag. Sordon, Göttingen.
Weber, Winfried W., 2009. Managing complexity – lessons from Peter Drucker and Niklas Luhmann. Journal of Innovation and Creativity, South Korea 2 (1), 29–37.
Weber, Winfried W., Kulothungan, Gladius (Eds.), 2010. Peter Drucker's Next Management. New Institutions, New Theories and Practices. Sordon, Göttingen.

Drug Control: International Aspects

Robin Room, University of Melbourne, Melbourne, VIC, Australia; Stockholm University, Stockholm, Sweden; and Centre for Alcohol Policy Research, Turning Point, Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Systems to control problems from psychoactive substance have included prohibitions, prescription systems, state control, and licensing, but did not extend internationally until the twentieth century. International treaties now include three drug treaties, and conventions on tobacco and against doping in sport, but not for alcohol. The pattern and content of the treaties reflect the wide conceptual chasm in the mid-tenth century between prohibited drugs and alcohol and tobacco. The international drug control system has diminished use of some prescribed drugs, but has been ineffective in suppressing illicit markets. Prototypical patterns of psychoactive substance use are medicinal use, customary regular use, and intermittent use – the last often invested with drama. In the last two centuries, a fourth pattern of use has become conceptualized as addiction. The social handing of psychoactive substances has been a fertile ground for social science research, with an emphasis on the social construction of problems and critical analyses of addiction and other conceptualizations, along with positivist work measuring and analyzing social and health problems associated with substance use. Social science work on international aspects has been limited.

'Drugs' include a wide range of psychoactive substances, including tobacco, alcoholic beverages, coffee, and pharmaceuticals, as well as marijuana, opiates, and cocaine. Prototypical patterns of use include medicinally, regularly as a 'banalized' part of daily life, and intermittently. In the modern world, addicted or dependent use is recognized as a fourth pattern.

Systems to reduce problems from drug use have included prohibitions, prescription systems, state control, or licensing systems, as well as no special controls. In the wake of efforts to prohibit alcohol, an international narcotics control system was set up starting in 1912, ensuring medical supplies but prohibiting other use. Since 1970, the system has expanded beyond opiates, cocaine, and marijuana to cover synthesized medications and precursor chemicals. Other international control systems have emerged more recently: an International Convention against Doping in Sport (2005) to control performance-enhancing drugs (including some psychoactive substances) in sports, and a Framework Convention on Tobacco Control (2003).

Social science work has mostly been funded because of social concerns about alcohol and other drugs, but researchers have often strayed beyond the official frames for their work. Thus there have been traditions of constructivist analysis of governing images of alcohol and drug problems, and critical analysis of addiction concepts. Analysis of the developing international systems for control of psychoactive substances has so far been primarily in terms of historical and policy studies (e.g., Braithwaite and Drahos, 2000), although the effects of the systems' provisions show up in studies in all social sciences.

Although the term 'drug' has a wider range of meanings, we are concerned here with substances that are psychoactive, that is, which alter the mental state of humans when ingested. Such drugs have been both prized and feared throughout human history. They have long been a substantial item in international trade, and indeed concerning the European colonial systems of 1500–1960 they have been described as the glue of empires (Courtwright, 2002). In the twentieth century, they also become the subject of ambitious international control structures. In connection with such control structures, brief consideration is given, too, to performance-enhancing drugs, a category overlapping with psychoactive drugs.

Psychoactive Drugs and How They Are Used

There are a wide variety of psychoactive substances, naturally occurring and synthesized, including tobacco, alcoholic beverages, coffee, tea, chocolate, and some spices, as well as substances legally available primarily through medical channels such as benzodiazepines, cannabinols (including marijuana), opiates (including heroin), and cocaine. Such substances often have other use values, along with their psychoactive properties (Mäkelä, 1983). Users may like the taste, or the image of themselves that the use conveys. Use may be a medium of sociability, or part of a religious ritual. Some substances have other useful properties; alcohol, for example, is a source of calories, and is the solvent in many tinctures and medicines.

Psychoactive drugs differ in their metabolic pathways and mechanisms of action in the human body, in the strength of their effects, and in the states of mind and feelings they induce. In particular, it is easy to become intoxicated with some substances – alcohol is a prominent example – while it is much less likely with others. But the effects of drug use are also powerfully dependent on the pattern of use, and on set and setting – that is, the expectations of the user and of others present, and the context of use (Zinberg, 1984). While the psychoactive effect of tobacco may not even register in the consciousness of a habituated cigarette smoker, in other circumstances the effect of tobacco use may be so strong that the user is rendered unconscious, as early Spanish observers reported concerning native South Americans.

Three social patternings of psychoactive drug use can be distinguished as prototypical: medicinal use, customary regular

use, and intermittent use. In many traditional societies, particular drugs or formulations have been confined to medicinal use – that is, to use under the supervision of a healer to alleviate mental or physical illness or distress. For several centuries after the technique for distilling alcoholic spirits had diffused from the Arab world to Europe, for instance, spirits-based drinks were regarded primarily as medicines. This way of framing drug use has been routinized in the modern state through a prescription system, with physicians writing the prescriptions and pharmacists filling them (Babor et al., 2010). Drugs included in the prescription system are usually forbidden for nonmedicinal use.

Where a drug becomes a regular accompaniment of everyday life, its psychoactivity is often muted and even unnoticed, as is often the case for a habitual cigarette smoker. Likewise, in southern European wine cultures, wine is often differentiated from intoxicating 'alcohol'; wine drinkers may be expected to maintain the same comportment after drinking as before. We may call this a pattern of 'banalized use': a potentially powerful psychoactive agent is domesticated into a mundane article of daily life, available relatively freely on the consumer market.

Intermittent use – for instance, on sacred occasions, at festivals, or only on weekends – minimizes the buildup of tolerance to the drug. It is in the context of such patterns that the greatest attention is likely to be paid to the drug's psychoactive properties. The drug may be understood by both the user and others as having temporarily taken over control of the user's behavior, and thus to explain otherwise unexpected behavior (Room and Collins, 1983). Given the power attributed to the substance, access to it may be limited: in traditional societies by sumptuary rules keyed to social differentiations, in industrial societies by other forms of market restriction.

A fourth pattern of use is commonly recognized for certain drugs: addicted or dependent use, marked by regular use, often of large doses. Since the pattern of use of the drug in question is not defined in the society as banalized, addiction is defined as an individual failing rather than as a social pattern. While attention is paid to physical factors sustaining regular use, such as use to relieve withdrawal symptoms, most formulations of addiction focus on psychological aspects, including an apparent commitment to drug use to the exclusion of other activities and despite default of major social roles. An addiction concept thus also focuses on loss of normal self-control, but the emphasis is not so much on the immediate effects of the drug as on a repeated or continuing pattern of an apparent inability to control or refrain from use, despite adverse consequences. Addiction as an explanation of otherwise inexplicable behavior now has wide currency, but first became established in industrializing societies less than two centuries ago (Room, 2003).

Among the attractions of psychoactive substances is that, at least up to a certain point, they potentially enhance one or another kind of performance by the user. For instance, coffee or a cigarette may modestly improve mental performance, a limited amount of alcohol may improve dancing or other expressive activity, amphetamines may improve wakefulness on a long watch. There are also nonpsychoactive substances (steroids, beta-blockers, etc.) which can potentially improve one or another kind of performance. In particular, in the context of competitive sports, international attention has been focused on a wide variety of substances that can improve some aspect of athletic performance (Carstairs, 2003).

Problems from Drug Use, and Systems of Control

From ancient times, it has been recognized that problems can result from drug use. Formerly, the main concerns were with social and political problems seen as associated with drug use. In medieval England, the government's main concern in controlling alcohol sales was with enforcing discipline in the labor force. Violence, disturbances of the peace, and harm to dependent family members would also be commonly recognized, particularly in relation to alcohol consumption. Often it was rebellious or countercultural elements of a society which introduced or became associated with new drugs. Reacting to this, political establishments in Europe tried at first to control coffeehouses, and in Turkey and elsewhere imposed severe punishments on tobacco use (Austin, 1978).

New social concerns about problems from drug use entered the scene in the modern era. Concerns about the physical health of the drinker were a factor in responses to the English 'gin epidemic' of the eighteenth century. In the early nineteenth century, the idea of addiction became established, at first for alcohol and in the USA (Levine, 1978), and was applied also to other psychoactive drugs by the late nineteenth century. As safety became identified as a social problem in the latter era, attention was also directed to the potential effects of drinking in degrading performance of mechanical tasks. In recent decades, as the automobile became a central part of daily life, this concern became much stronger, and was extended too to other drugs.

Social concerns about problems from drug use have resulted in a variety of kinds of efforts to control availability (MacCoun et al., 1996; Room, 2000). One option has been a flat prohibition on use, or at least on recreational use. Combined with this, sometimes, has been a system for use of the substance as a medication. The prescription system in modern states is essentially a state-sanctioned form of rationing of availability, with doctors and pharmacists delegated to serve as the gatekeepers on access. A third option has been a system of state control or licensing of sales, such as the systems of alcohol control through state monopolies and licenses in North American states and provinces. Such a structure can be combined with other discouragements for use, such as excise taxes and restricted opening hours. A fourth option, applicable for instance nearly everywhere for tea and coffee, is that the psychoactive substance is treated as an ordinary commodity, with no special controls. For intoxicating substances, choice of the third or fourth option is often accompanied by 'harm reduction' measures (in the terminology of the illicit drug field), since the options involve official acceptance of the fact of recreational drug use.

Where a psychoactive substance is seen as causing great problems, prohibition of use has been a fairly common societal response. Such a prohibition on alcohol had long been imposed on all faithful Muslims by the Koran. Legal prohibition of alcohol, initially on a local basis or for special population categories such as indigenous peoples, became a feature

of some Christian societies in the late nineteenth and early twentieth centuries. Prohibition of alcohol sales extended to the national level around the time of World War I in 13 self-governing nations, including the USA, Canada (briefly), Norway, Finland, and Russia (Schrad, 2010). The USA, in particular, had some success in gaining international cooperation with its prohibition through bilateral treaties (Tyrrell, 1994). A flourishing worldwide temperance movement succeeded in pushing the European powers with colonies in Africa into adopting treaties prohibiting aspects of the distilled spirits trade in Africa (Bruun et al., 1975). These fell into disuse by the 1950s.

The alternative to prohibition in the early twentieth century was 'state control,' and comprehensive systems of alcohol control were eventually adopted, not only in the societies that had tried prohibition, but also in many other nations. These systems have been gradually weakened, particularly since 1950 (Mäkelä et al., 1981). On the other hand, in a broad range of societies there has been a steady strengthening of controls over drinking in specific circumstances, in particular over drinking and driving (Gusfield, 1981; Mann et al., 2001).

In the early twentieth century, the idea of a prohibitory regime was extended from alcohol to other psychoactive substances. In the realm of 'narcotics,' which came to refer to opiates, cocaine, and marijuana, the first international treaty was adopted in 1912, and extended globally as part of the treaties concluding World War I (McAllister, 1992). The agreement on narcotics can be seen as part of an emergent tradition of such regimes for social problems defined as international in scope (Nadelmann, 1990). A series of subsequent international agreements were codified in the 1961 Single Convention on Narcotic Drugs (amended 1972). While the system had always included provision for a legal market in narcotic medications under a prescription system, in a 1971 convention its scope was broadened to extend international control to other synthesized medications such as amphetamines, barbiturates, and benzodiazepines. A 1988 convention extended control also to 'precursor chemicals' used in manufacturing controlled drugs. By 1996, more than 240 substances were, in principle, under the system's control (Room and Paglia, 1999).

In the early 2010s, the international narcotics control system included the United Nations Drug Control Program, with about 500 staff, headquartered in Vienna. The political organ of the system is the Commission on Narcotic Drugs, with an elected membership of 53 nations. The International Narcotics Control Board serves as the control board for the legal (pharmaceutical) markets in controlled drugs. Alongside this international effort is a substantial global antidrug effort by the USA, including 751 staff posted in 59 other countries in 2007, and an annual International Narcotics Control Strategy Report to the US Congress on the extent of cooperation by each other country with antidrug actions (Babor et al., 2010: 215).

Although effective action has been taken to diminish use of prescribed drugs like barbiturates, the international system has been relatively ineffective in suppressing illicit international traffic in drugs, (Room and Reuter, 2012), and has done little to limit the very large markets in prescribed psychoactive substances, particularly in high-income countries (the U.S. consumed 70% of the world's legal supplies of opiate medications in 2003–05) (Babor et al., 2010: 86–87).

At the end of the 1990s, two other international systems for the control of psychoactive drugs were in formation. In 1989 the Council of Europe had put forward an Anti-Doping Convention, which by 2000, had been signed by 40 European countries and other countries such as Australia and Canada. This was largely superseded by a treaty, the International Convention against Doping in Sport, adopted in 2005 under the auspices of UNESCO and in force since 2007.

In October 2000, the World Health Organization began negotiations on a global Framework Convention on Tobacco Control, adopted in 2003 and in force since 2005. Many of the provisions of the tobacco treaty are 'soft law' encouraging rather than requiring specific national actions such as high cigarette taxes or prices. Unlike the international narcotics control system and the nascent antidoping system, the international tobacco control system does not attempt to prescribe prohibition of use.

In the meantime, in 2014 there is no prospect of any international agreement on alcohol control. For tobacco and alcohol, as well as other psychoactive substances not under international control, international treaties have operated in the opposite direction. The World Trade Organization and related treaties, and regional free trade agreements, have been used to weaken national and local alcohol controls (Holder et al., 1998; O'Brien, 2013). Despite shifts in biomedical and public health thinking back to viewing psychoactive substances in a common frame, in the early twenty-first century the international regimes still reflect and maintain the wide chasm set in the mid-twentieth century between the definition and handling of substances covered by the drug treaties and other psychoactive substances (Courtwright, 2005; Braithwaite and Drahos, 2000).

Traditions of Social Science Analysis

Dominant patterns of use of a drug in a society, and shifting conceptions about its use, greatly affect the size and shape of the social science literature concerning that drug. In particular, the very different sociocultural positions of 'illicit drugs,' benzodiazepines, alcohol, tobacco, and coffee in most developed societies means that the literature about these different psychoactive substances are often quite separate.

Novel substances have proved a potent catalyst in many societies in crystallizing free-floating fears about the younger generation and the future. Culturally accepted psychoactive substances, on the other hand, become deeply entrenched in the culture, although often with the potential to again become contested. Social concerns about drugs have thus often ensured substantial funding for social scientists to work in the area. For instance, a substantial part of Polish sociology was financed to work on alcohol problems in the difficult days of martial law there in the 1980s. But while much social science work has been funded and expected to proceed within the parameters of official frames, the work of sociologists and other social scientists has also played a key role in broadening the base of scholarship and understanding. For the drugs under international control, in particular, there is a discernible split in emphasis, particularly in the US, between government-funded social research and a growing tradition of critical analyses.

In the first place, alcohol and drug issues have proved fertile grounds for constructivist analyses by sociologists. Historical and comparative work has identified a number of alternative governing images or models of alcohol and drug problems, and shown that these governing images change over time (Bruun, 1971; Gusfield, 1967). Constructivist perspectives have often been combined with a positivist concern about measuring and comparing social and health problems associated with drug use, and about distinguishing problems attributable to the drug use from problems that are secondary to societal reactions (e.g., Fischer et al., 1997). Some analyses point to a function for illicit drugs of providing 'suitable enemies' for the modern state, in comparison with the 'unsuitable enemies' of tobacco, alcohol, and pharmaceuticals (Christie and Bruun, 2003).

In the second place, there is a long sociological tradition of critical analysis of the concepts of alcoholism or addiction as a disease (Room, 1983), concepts that were the dominant discourse in the last half of the twentieth century. The alcohol and drug arenas have thus figured prominently in sociological analyses of the 'medicalization of deviance' – and its demedicalization (Conrad and Schneider, 1992). Applying perspectives from Foucault and Rothman, Levine (1978) showed that the addiction concept has a history, first coming to the fore in the early nineteenth-century USA, as an explanation of the mystery of continued drinking by the drunkard, even after pledging to abstinence. Analyses in this tradition point to the centrality to addiction concepts of the focus on self-control (Valverde, 1998). The economist's concept of 'rational addiction' has also contributed to the ferment in thinking about addiction concepts (Elster and Skog, 1999).

There are as yet few social science analyses of the burgeoning international systems for control of psychoactive substances. Bruun et al. (1975) studied the development of the international drug control system until the early 1970s, but further arm's-length analysis of the expansion of this system's scope only started to appear after the millennium (e.g., Carstairs, 2005; Bewley-Taylor 2005), and arm's-length analyses of the development of new international instruments and agencies around doping in sports are as yet rare (Houlihan, 2004) and rarer still for tobacco. Work looking across the different systems (and absence of systems) of international control of psychoactive substance remains uncommon (but see Braithwaite and Drahos, 2000). Why the systems are developing separately, the nature and practicability of their goals, and their place in the developing international system, are still important topics for future work.

See also: Drug Use and Abuse: Psychosocial Aspects; Drugs: Illicit Use and Prevention; Health Behaviors, Assessment of; Policy Knowledge: Universities.

Bibliography

Austin, G., 1978. Perspectives on the History of Psychoactive Substance Use. Research Issues. National Institute on Drug Abuse, vol. 24, Washington, DC.

Babor, T., Caulkins, J., Edwards, G., Fischer, B., Foxcroft, D., Humphreys, K., Obot, I., Rehm, J., Reuter, P., Room, R., Rossow, I., Strang, J., 2010. Drug Policy and the Public Good. Oxford University Press, Oxford.

Bewly-Taylor, D.R., 2005. Emerging policy contradictions between the United Nations drug control system and the core values of the United Nations. International Journal of Drug Policy 16, 423–431.

Braithwaite, J., Drahos, P., 2000. Drugs. In: Global Business Regulation. Cambridge University Press, Cambridge, MA, pp. 360–398.

Bruun, K., 1971. Finland: the non-medical approach. In: Kiloh, L.G., Bell, D.S. (Eds.), 29th International Congress on Alcoholism and Drug Dependence. Butterworths, Australia, pp. 549–559.

Bruun, K., Pan, L., Rexed, I., 1975. The Gentlemen's Club: International Control of Drugs and Alcohol. University of Chicago Press, Chicago, IL.

Carstairs, C., 2003. The wide world of doping: drug scandals, natural bodies, and the business of sports entertainment. Addiction Research & Theory 11, 263–281.

Carstairs, C., 2005. The stages of the international drug control system. Drug & Alcohol Review 24, 57–65.

Christie, N., Bruun, K., 2003. Den Gode Fiende: Narkotikapolitikk I Norden, third ed. Universitetsforlaget, Oslo, Norway.

Conrad, P., Schneider, J., 1992. Deviance and Medicalization: From Badness to Sickness, second ed. Temple University Press, Philadelphia, PA.

Courtwright, D.T., 2002. Forces of Habit: Drugs and the Making of the Current World. Harvard University Press, Cambridge, MA.

Courtwright, D.T., 2005. Mr ATOD's wild ride: what do alcohol, tobacco, and other drugs have in common? Social History of Alcohol & Drugs 20, 105–124.

Elster, J., Skog, O.-J. (Eds.), 1999. Getting Hooked: Rationality and Addiction. Cambridge University Press, Cambridge, MA.

Fischer, B., Kendall, P., Rehm, J., Room, R., 1997. Charting WHO goals for licit and illicit drugs for the year 2000 – are we 'on track'? Public Health 111, 271–275.

Gusfield, J., 1967. Moral passage: the symbolic process in public designations of deviance. Social Problems 15, 175–188.

Gusfield, J., 1981. The Culture of Public Problems: Drinking, Driving and the Symbolic Order. University of Chicago Press, Chicago, IL.

Holder, H.D., Kühlhorn, E., Nordlund, S., Österberg, E., Romelsjö, A., Ugland, T., 1998. European Alcohol Policy and Nordic Integration. Ashgate, Aldershot.

Houlihan, B., 2004. Building an international regime to combat doping in sport. In: Levermore, R., Budd, A. (Eds.), Sport and International Relations: An Emerging Relationship. Routledge, Abingdon, New York, pp. 62–76.

Joossens, L., 2000. From public health to international law: possible protocols for inclusion in the framework convention on tobacco control. Bulletin of the World Health Organization 78, 930–937.

Levine, H.G., 1978. The discovery of addiction: changing concepts of habitual drunkenness in American history. Journal of Studies on Alcohol 39, 143–174.

MacCoun, R., Reuter, P., Schelling, T., 1996. Assessing alternative drug control regimes. Journal of Policy Analysis and Management 15, 330–352.

Mäkelä, K., 1983. The uses of alcohol and their cultural regulation. Acta Sociologica 26, 21–31.

Mäkelä, K., Room, R., Single, E., Sulkunen, P., Walsh, B., et al., 1981. Alcohol, Society and the State: 1. A Comparative Study of Alcohol Control. Addiction Research Foundation, Toronto, ON, Canada.

Mann, R.E., MacDonald, S., Stoduto, G., Bondy, S., Jonah, B., Shaikh, A., 2001. The effects of introducing or lowering legal per se blood alcohol limits for driving: an international review. Accident Analysis & Prevention 33, 569–583.

McAllister, W.B., 1992. Conflicts of interest in the international drug control system. In: Walker, W.O. (Ed.), Drug Control Policy: Essays in Historical and Comparative Perspective. Pennsylvania State University Press, University Park, PA, pp. 143–166.

Nadelmann, E., 1990. Global prohibition regimes: the evolution of norms in international society. International Organization 44, 483–525.

O'Brien, P., 2013. Australia's double standard on Thailand's alcohol warning labels. Drug and Alcohol Review 32, 5–10.

Room, R., 1983. Sociological aspects of the disease concept of alcoholism. Research Advances in Alcohol and Drug Problems 7, 47–91.

Room, R., 2000. Control systems for psychoactive substances. In: Ferrence, R., Pope, M., Room, R., Slade, J. (Eds.), Nicotine and Public Health. American Public Health Association, Washington, DC, pp. 37–56.

Room, R., 2003. The cultural framing of addiction. Janus Head 6, 221–234.

Room, R., Collins, G. (Eds.), 1983. Alcohol and Disinhibition: Nature and Meaning of the Link. NIAAA Research Monograph No. 12. DHHS publication (ADM) 83–126. USGPO, Washington, DC.

Room, R., Paglia, A., 1999. The international drug control system in the post cold war era: managing markets or fighting as a war? Drug & Alcohol Review 18, 305–315.

Room, R., Reuter, P., 2012. How well do international drug conventions protect public health? The Lancet 379, 84–91.
Schrad, M.L., 2010. The Political Power of Bad Ideas: Networks, Institutions, and the Global Prohibition Wave. Oxford University Press, Oxford.
Taylor, A.L., 1996. An international regulatory strategy for global tobacco control. Yale Journal of International Law 21, 257–304.
Tyrrell, I., 1994. Prohibition, American cultural expansion and the new hegemony in the 1920s: an interpretation. Histoire Sociale/Social History 27, 413–445.
Valverde, M., 1998. Diseases of the Will: Alcohol and the Dilemmas of Freedom. Cambridge University Press, Cambridge, MA.
Zinberg, N.E., 1984. Drug, Set and Setting: The Basis for Controlled Intoxicant Use. Yale University Press, New Haven, CT. This page intentionally left blank

Drugs, Decriminalization of: Cultural Concerns

Stephen J Morse, University of Pennsylvania Law School, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article defines drugs and decriminalization, addresses how the debate about decriminalization has changed in the last decade, considers whether criminalizing recreational drug use is a legitimate infringement of individual liberties, and provides an analysis of the costs and benefits of decriminalization, including the uncertainties involved. It concludes that the outcome of decriminalization is inevitably uncertain and that much depends on the details of the decriminalization scheme.

Introduction

The issue of drug decriminalization or legalization is whether and to what extent the state should use criminal sanctions to regulate drug-related activities such as production, sale, purchase, possession, and use. Analysis of the issue must therefore focus on consideration of the legitimate uses of the criminal law, with special attention to the moral and political rights of citizens to use drugs and to the costs and benefits of criminal prohibition. After considering the definitions of drugs and decriminalization, this article will consider how the decriminalization debate has changed in the last decade. Then, following discussion of the basic limits on the use of the criminal law, this article will address the contributions of rights and cost–benefit analyses to the decriminalization debate.

Defining Drugs and Decriminalization

Despite common belief to the contrary, there is no uncontroversial definition of 'drug' or 'decriminalization,' but clarity about both is necessary to understand the debate and to guide public policy.

Drugs

Virtually, all definitions are vague or overinclusive, permitting categorization as a drug of almost any substance that may be consumed. Some definitions are circularly dependent on legal regulations: if a law regulating 'drugs' includes a particular substance within its ambit, it is a drug if it is not so included, it is not. Vague, overinclusive, and circular definitions cannot sensibly guide public policy, especially when the state's awesome power to blame and punish for illegitimate use is dependent on the definition.

Definitional problems concerning drugs and legitimacy cannot be entirely avoided by using allegedly scientific or value-neutral medical/therapeutic or nutritional concepts. Concepts like disease and therapy are themselves valueladen and inevitably problematic and controversial. Many substances that have unquestionably legitimate medical or nutritional uses can also be used for nonmedical or nonnutritional purposes that may be of questionable legitimacy. Finally, definitions that define a substance as a drug in terms of its intended use rather than in terms of its inherent properties obscure important questions of legitimate use.

Despite the conceptual problems attending the definition of a drug, a loose but common sense definition is possible for discussing decriminalization, which centrally concerns substances that are consumed recreationally, that is, primarily to produce pleasure, whether or not those substances have other legitimate uses. For the purpose of understanding the current debate, then, recreational drugs can be defined as consumable substances that cross the blood–brain barrier and that (1) can affect mood, cognition, and behavior in pleasurable ways; (2) can be used primarily for recreation, including relaxation, excitement, and pleasurable states generally; and (3) can be used so as to endanger the user and others. Although admittedly loose, this definition covers both legal recreational substances such as ethanol (alcohol), nicotine, and caffeine, and substances that are illegal either per se or if they are not properly prescribed by physicians such as marijuana, cocaine, opiates, heroin, barbiturates, amphetamines, phencyclidine, and the like.

Decriminalization and Criminalization of Drugs

Decriminalization and criminalization are as difficult to define as drug because each includes widely varying legal responses to drug-related activities. Indeed, the debate should properly be characterized as including a wide spectrum of opinions and proposals. The core of decriminalization proposals, however, is the limitation or abolition of present criminal prohibitions of actions related to recreational drug use by competent adults. Some prefer the term legalization to refer to the limitation of criminal prohibitions and penalties. Other, primarily rights-based theorists, believe that state criminal prohibition of recreational drug activities is not prima facie justified, so they prefer the apparently more politically neutral term decriminalization. For ease of exposition, the latter term will be used.

The specific details of particular decriminalization proposals depend on the proponent's position on rights and on the costs and benefits of the proposal compared to the current regime of criminal prohibition. Proponents recognize that some criminal law regulation and substantial civil law regulation will probably be necessary to respond to recreational drugs, but all agree that substantially less use of criminal law is desirable and that it should not be the primary mode of regulation.

The core of criminalization proposals is that the criminal law should be the primary means of regulating recreational drugs, but the proper scope of criminalization is disputed and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45024-5

virtually all defenders of criminal prohibitions believe that noncriminal means such as education and treatment should also be used. Virtually no proponent of criminalization argues for expanding the scope of criminalization, say, by prohibiting the recreational use of currently legal substances such as alcoholic beverages or by substantially increasing the already serious criminal penalties for drug offenses. The current debate is about limitation, not expansion, and whether criminal law should be the primary mode of regulation.

The Current State of the Decriminalization Debate

The debate about decriminalization has altered substantially in the past decade for a variety of reasons. Although the basic positions have been established for decades, the vast violence and corruption the illegal drug trade produces has caused unprecedented numbers of domestic and foreign politicians and other opinion leaders to call for various forms of decriminalization. Some nations and individual communities have adopted de jure or de facto decriminalization, typically coupled with greater emphasis on public health, and these changes have seemed to succeed. For example, two states in the United States voted in the 2012 elections to decriminalize marijuana and other states have approved marijuana for medical use. Advocacy for decriminalization of marijuana, which is not a particularly dangerous drug, has been considerably more frequent and successful than similar claims for allegedly 'harder' drugs, such as heroin and cocaine.

Many researchers have tried to model or to predict how various forms of decriminalization would specifically affect drug use in particular and society more generally. Most have concluded that the exact effects are very difficult to estimate, but that decriminalization is not a cost-free panacea.

Major strides have been made in understanding the biological basis for addictions. Such understanding typically lends support to treating drug use as the sign of a disorder rather than as a crime.

Harm reduction approaches that use a mix of law enforcement and public health approaches are also familiar, but have gained greater impetus in recent years as various courts and researchers have experimented with apparently successful programs to use public health initiatives and the leverage of the criminal law to reduce the violence associated with the drug trade and drug use. In short, in the United States, criminal law enforcement still represents the primary response to the recreational use of drugs, but the trend in the debate is much more strongly in favor of decriminalization or forms of regulation other than a purely punitive response.

A reasonable speculation in 2013 is that decriminalization 'experiments' and harm reduction policies are likely to increase, especially for marijuana, but criminal law enforcement will remain a major tool.

The Limits of the Criminal Law: The Harm Principle

The state's police power to apprehend, prosecute, convict, and punish citizens is one of the most awesome and probably the most afflictive power it exercises. Infliction of such pain requires substantial justification in liberal democracies that seek to protect individual rights and liberty. Afflictive and expensive criminal regulation should be avoided in favor of less intrusive means unless the harm is great, punishment would be deserved for causing it, and criminal prohibition seems necessary to reduce the level of harm and will not unduly impose on other important interests. Criminal law theorists term these limitations on the state's power to criminalize the 'harm principle.' For example, intentionally inflicting emotional cruelty on others is morally despicable, but such behavior is not criminalized because the harm is considered insufficient to justify criminal punishment and criminalizing such conduct would have other wide negative effects. Socialization by families, religious organizations, and schools is reasonably effective to limit such conduct, and regulating such behavior by criminal law might intrude on protected liberties, such as the right to free speech.

There is little debate about the state's justification for prohibiting and punishing traditional crimes against the person, such as homicide and forcible rape, and against property, such as theft and arson, that cause grievous harm. Although individual liberties are precious, no citizen has a right to kill, rape, steal, or burn the property of others for private advantage and the criminal sanction seems necessary to curb such behavior. Societies can disagree about the scope of appropriate criminalization. For example, there may be reasonable dispute on how much creation of homicidal risk should be necessary to warrant criminal penalties in addition to civil damages or how severely arson should be punished, but the state's power to criminalize great harm to others is uncontroversial.

The harm principle suggests caution before criminalization, however, when the harm is primarily inflicted on oneself, when the harm is primarily moral and not physical, psychological or economic, or when using criminal law to prevent the harm appears to intrude on important rights or appears unnecessary effectively to reduce the harmful behavior. In liberal societies, the right of the state to criminally prohibit actions that harm primarily the actor (so-called legal paternalism) is more controversial than the right to prohibit harm to others because respect for individual liberty and autonomy generally entails that citizens have greater liberty to harm themselves than to harm others. Such liberty is most arguably extensive and the state's right to regulate criminally is most questionable when the threatened harm to self is apparently solely moral and no other harm to self or to others can be discerned. Most liberal theorists reject the state's power to use criminal law solely for the purpose of enhancing the moral perfection of its citizens – so-called legal moralism. Such goals are not the business of the state and should be achieved by other, less intrusive means.

In political and moral regimes that grant the state more authority to regulate individual liberty, both legal paternalism and moralism might be less controversial. For example, theocracies may coherently use the criminal law more extensively than liberal states to achieve paternalistic and moralistic goals. Even within liberal regimes, the permissible scope of paternalism and moralism may vary, although both require more substantial justification than criminalizing harm to others. Consideration of the state's justifiable use of criminal law to regulate recreational drug use, however, must attend to the appropriate limits of state power within a particular

political regime. The remainder of this article assumes that the regime considering decriminalization is liberal.

Recreational Drug Use and Rights

People often assume that there is no right, no liberty interest to use drugs recreationally and, consequently, that the state has appropriate authority to use criminal law to prevent such use. Such an assumption is seldom supported with cogent argument, however, and it is not easy to give compelling moral and political reasons for rejecting such a right.

Sensible discussion must begin with the recognition that people take drugs recreationally because they desire the pleasurable effects produced by consuming substances that alter mood and cognition. It seems clear, however, that the state has no justifiable interest in criminal prohibition simply because the primary or even sole goal of behavior is pleasure or because the behavior involves consumption of a substance that can alter mood or cognition. Many people might consider it wrong to engage in activities solely for pleasure, especially if alteration of mood and cognition were involved and there were some danger of dependence on the activity, but a liberal society does not interfere with a citizen's right to make autonomous choices to engage in such activities. Citizens surely have a prima facie right to seek pleasure for its own sake and it is very difficult to imagine a secular, liberal argument suggesting that such a goal is immoral or harmful per se, even if some danger might sometimes be involved.

For example, suppose that a citizen engages in meditation solely for recreational purposes. No persuasive liberal moral or political theory would justify criminal prohibition of meditation, even if some citizens became dependent on meditating. Recreational drug use involves the consumption of a substance, but it is hard to imagine why the source of the recreation alone should make a difference. Currently, illegal drugs produce pleasure by altering mood and cognition, but so do many legal activities, including meditation, and legal drugs such as caffeine, ethanol (alcohol), and nicotine. The state's right to criminalize alterations of mood and cognition per se is questionable because such alterations are not per se immoral or harmful.

Citizens appear to have a prima facie right to engage in recreational alteration of mood and cognition by drug consumption and the state surely has the burden of justifying criminal prohibition of such recreation by powerful arguments. No right is absolute, however, and the possible right to pursue recreational drug use must yield to a powerful state interest in preventing such use, especially the undoubted state interest in curtailing the threat that drugs will cause substantial harm. Nonetheless, the harm limitation counsels caution before using the criminal law.

Some proponents of decriminalization refer to illegal drug use as a 'crime without a victim' because the person potentially most harmed, the user, in effect consents to the threat. But virtually no action affects only the agent, and recreational drug use sometimes threatens families and communities with economic, psychological, and physical harms. Many of these harms may be paradoxically produced or enhanced by criminal prohibition itself, but many surely are not. Moreover, harms one consents to suffer are nonetheless harms. The often disastrous consequences of various forms of drug addictions certainly harm the person who consents to the risk of becoming addicted by persistently using a substance. The question is whether the harm to others or to self that drugs produce are sufficient to trump the individual's prima facie right to use drugs and thus to warrant criminalization of drug use.

The next section on cost–benefit analysis considers the harms drugs produce, but it must be remembered that liberal societies often permit dangerous behavior in order to protect the right of citizens to pursue their own visions of how life should be lead. Such dangerous behavior is regulated noncriminally, if at all. For example, knowingly following unhealthy dietary habits can severely harm and even kill, and the harm or death can threaten economic, psychological, and physical harms to others. The enormous costs of unhealthy diets are well known, but it is unthinkable in a liberal society to use the criminal law to regulate individual eating habits. To promote the health of the populace, the state can properly regulate the manufacture and sale of food and can impose a duty on manufacturers to warn consumers of potential health risks associated with particular foods. It can even use the criminal law to police egregious failures to comply with such regulations, but it is unthinkable that it would criminally prohibit unsafe eating habits. For another example, consider the immense costs to many users and to society at large that flow from the consumption of ethanol and nicotine. Again, the state can appropriately regulate many aspects of such substance consumption to reduce the consequent harms, including the use of criminal law for egregiously harmful forms of misuse, such as drunk driving. For many reasons, however, importantly including the right to make one's own choices unencumbered by undue state interference, criminal prohibition of the consumption of alcohol and nicotine by the state is considered unwarranted.

If drugs produce great harms per se and no one has a right to use drugs recreationally or if that right is insubstantial at best, then the only question for analysis is whether criminal regulation is justified by the benefits it achieves. But if there is a weighty liberty interest in making choices about how to live one's life, including potentially unwise and dangerous choices about drugs, then the creation of harm is insufficient per se to warrant criminalization. Some dangerous behaviors must be permitted to protect liberty and regulated only by civil remedies, including damages when the harm occurs. Any adequate analysis of decriminalization must therefore include consideration of the strength of the liberty interest and it appears that in liberal societies the right is substantial. Empirical analysis of the consequences of various behaviors under various regulatory regimes is necessary to inform decisions about the appropriate scope of criminalization and decriminalization of drugs, but the question of rights cannot be avoided.

Cost–Benefit Analysis

The outcome of a cost–benefit analysis of decriminalization of drugs will inevitably be indeterminate because much of the data are unavailable, unreliable, or fluctuate and because the costs and benefits of a new regime must be speculative.

Although there have been attempts recently to provide more specific estimates, these, too, are ultimately indeterminate. Debate too often focuses on worst case scenarios that depend on salient but unrepresentative data. Insufficient information and biased misinformation often suggest the wisdom of maintaining the status quo rather than risk making matters worse by changing public policy. Policies such as the criminal prohibition of drugs that arguably fail and that impose great costs, including the infringement of citizens' rights, should be and are debated and cost–benefit analysis plays a crucial role in these debates. Unfortunately, historical and cross-cultural comparisons are of extremely limited value because social variables that vary radically intertemporally within a society and across different societies, immensely affect the consequences of drug use and regulation.

Harms of Drugs and the Benefits of Criminalization

Most currently illegal drugs would create some potential risks to the user and others, even if they were pure, inexpensive, and legal. For example, misuse of drugs can cause addiction and other harmful and well-known health, psychological, and economic effects. People can arguably become addicted to substances other than drugs and to activities such as gambling. Moreover, the concept of addiction is itself problematic. Nonetheless, assume that, for various cultural and biological reasons, most currently illegal drugs have greater addictive or other harmful potential than most other substances and activities.

Much decriminalization debate mistakes the probability and extent of the risks drugs pose. For example, most people who use most drugs recreationally do not become addicted and the moderate recreational use of drugs is not per se dangerous if the drug is pure and properly consumed. Nonetheless, the ease of access to drugs following decriminalization would surely increase the prevalence of use, abuse, and addiction and their further harmful consequences. The harms might be especially great in poorer and minority communities, where the prevalence of drug use is no higher than in other communities, but where the effects have been more devastating. Indeed, some responsible observers believe that decriminalization might effectively destroy poorer and minority communities already ravaged by drug abuse and addiction.

The relation between drugs and other criminal activity and health harms is fraught because it is difficult to disentangle the effects of criminalization itself from the independent effects of drug use. Some drugs, especially if used heavily, can substantially impair judgment, facilitate impulsivity, and have other effects that increase the risk of criminal offending and other dangerous behavior, such as careless driving or sharing needles for intravenous injection. Increased drug use, especially heavy use, will raise the rates of such undesirable behavior.

Criminalization of drugs inhibits use generally by making production, sale, possession, and use more expensive and dangerous and by sending the clearest possible message that drug use is wrong. For example, there may have been both rights-based and consequential reasons to repeal prohibition in the United States, but the data suggest that consumption and consumption-related diseases such as cirrhosis of the liver both decreased when most alcohol production and sale were illegal. Criminalization also helps prevent drug use by minors. Although minors regrettably have access to and consume drugs in a regime of criminalization, ease of access generally would also increase use and abuse among minors.

The use of criminal law to regulate drug use and related harms is obviously not fully successful by any standard and it is immensely expensive, but the proponents of continued criminalization argue that the current regime provides greater benefits than costs and that the unknown benefits, if any, of decriminalization would pale in comparison to the costs of inevitably increased use and abuse.

The Costs of Criminalization

Criminal law enforcement, including imprisonment, is an especially intrusive and expensive form of regulation of any behavior. The cost–benefit critique of criminalization argues that such costs are not outweighed by the benefits because criminal law is powerless to make more than a small dent in the use of drugs and because criminalization itself creates many avoidable harms.

The attempt to eradicate drug use by criminal prohibition cannot fully succeed because large numbers of people want recreational drugs for the pleasure or relief they provide and it is widely recognized that the dangers of drugs are often exaggerated. The criminal law's so-called war on drugs in the United States is fighting a fluctuating but always immense number of citizens, most of whom are not addicts or abusers and whose use does not pose grave danger to society or themselves. It is hard to enforce laws that such large numbers of people wish to violate and that have no immediate and direct complaining victim. Furthermore, public hypocrisy about drugs inevitably undermines law enforcement. Western society is not nearly as monolithically antidrug as many claim, as public policy toward alcohol and nicotine indicates. Given the powerful factors that motivate the desire for drugs, the criminal sanction appears ineffective. The criminal justice system cannot prosecute and imprison more than a tiny fraction of the enormous numbers of people involved in the illegal drug trade unless the justice system massively diverts resources from other undisputed criminal law needs and abandons civil liberties protections. Some believe that criminal enforcement could be applied in more limited cases with greater effect and at lesser cost. If effective, this change would decrease the costs of criminalization, but the vast majority of jurisdictions are not now employing such programs.

The criminal courts of the United States are already clogged with drug cases and the ever-expanding prison population includes an enormous proportion of drug offenders. Even spending vastly more money seemingly will provide only limited benefit. Despite billions spent on drug enforcement each year, undoubted law enforcement successes and the huge numbers of people in prison for drug-related offenses, drugs are still freely available and ever-purer and ever-cheaper. Fluctuations in drug usage appear far more related to major social, cultural, and economic forces than to criminal law enforcement. The limited decrease in the use of criminal law enforcement produces does not seem to justify the gargantuan costs.

Criminalization also threatens its own effectiveness and creates other problems. For example, criminal prohibitions

raise the price on a good or service – the so-called crime tariff – and insure that there will be an endless supply of producers and dealers who seek to realize the monstrous profits to be made through any means necessary, including violence. It also guarantees that many users will be impoverished and driven to a life of further crime simply to support themselves and to obtain drugs. The immense profits facilitate the growth and power of domestic and international organized crime and can be used to corrupt law enforcement, the criminal justice system in general, and both domestic and foreign politicians. Moreover, the consensual nature of drug offenses insures that law enforcement will unwittingly – and sometimes wittingly – use illegal means to enforce drug laws, thus further corrupting law enforcement and the administration of justice and causing disrespect for both. Criminalization decreases the probability that inevitable drug consumption will occur under conditions safest to health. Finally, primary emphasis on criminal law virtually insures that fewer resources will be devoted to investigating and implementing less intrusive and potentially more effective means to reduce the harms drugs cause.

The Benefits of Decriminalization

Specific decriminalization proposals would produce different benefits, but the potential general benefits of decriminalization are the opposites of the costs of criminalization and can be stated easily in the abstract. They include an increase in personal liberty, probable cost savings, decrease in the crime, violence, and corruption that criminalization produces, increased respect for law enforcement, and increased attention to possibly more effective means such as education, treatment, and civil regulation to reduce use and the harms drugs cause. Criminal law would be used only to prohibit particularly dangerous drug-related activities such as drugged driving or selling to minors. Such traditional use of the criminal law would receive broad public support.

Conclusion

Proponents of criminalization sometimes grudgingly concede the problems with the criminal law enforcement approach, but virtually always criticize decriminalization proposals for their abstract or theoretical nature. They fear that decriminalization will produce an administrative, medical, economic, and social tragedy far worse than the current situation, and they ask for highly specific details and guarantees that this will not happen. Proponents of decriminalization recognize that they cannot provide such guarantees, but properly request in return that proponents of criminalization provide the details of how their approach will finally succeed after nearly a century of ever greater expenditures but apparent failure. Neither position appears able to convince the other on theoretical or empirical grounds, but only in recent years has decriminalization been openly debated and a genuine attempt to model the specific effects been begun. Nonetheless, at least in the United States, it appears that the political support to experiment with decriminalization is still weak, except for marijuana, but it is likely to strengthen.

See also: Occupational Health; Opiate Use and Abuse, History of.

Bibliography

Adler, F., Mueller, G.O.W., Laufer, W.S., 1995. Criminology, second ed. McGraw-Hill, New York.
Bayer, R., 1991. Introduction: the great drug policy debate—what means this thing called decriminalization. The Milbank Quarterly 69, 341–364.
Chandler, R.K., Fletcher, B.W., Volkow, N.D., 2009. Treating drug abuse and addiction in the criminal justice system: improving public health and safety. Journal of the American Medical Association 301, 183–190.
De Grieff, P. (Ed.), 1999. Drugs and the Limits of Liberalism: Moral and Legal Issues. Cornell University Press, Ithaca, NY.
Elster, J., 1999. Strong Feelings: Emotion, Addiction and Human Behavior. MIT Press, Cambridge, MA.
Falco, M., 1992. The Making of a Drug-Free America: Programs That Work. Times Books, New York.
Feinberg, J., 1986. Harm to Self. Oxford University Press, New York.
Global Commission on Drug Policy, 2011. War on Drugs: Report of the Global Commission on Drug Policy. Rio de Janeiro.
Greenwald, G., 2009. Drug Decriminalization in Portugal: Lessons for Creating Fair and Successful Drug Policies. Cato Institute, Washington, DC.
Hawken, A., Caulkins, J.P., Kilmer, B., Kleiman, M.A.R., 2012. Marijuana Legalization: What Everyone Needs to Know. Oxford University Press, New York.
Husak, D.N., 1992. Drugs and Rights. Cambridge University Press, Cambridge, UK.
Husak, D.N., 2002. Legalize This! The Case for Decriminalizing Drugs. Verso, London, UK.
Jacobs, J.B., 1990. Imagining drug legalization. The Public Interest 101 (Fall), 28–42.
Kaplan, J., 1983. The Hardest Drug: Heroin and Public Policy. University of Chicago Press, Chicago, IL.
Kleiman, M.A.R., 1992. Against Excess: Drug Policy for Results. Basic Books, New York.
Kleiman, M.A.R., Caulkins, J.P., Hawken, A., 2011. Drugs and Drug Policy: What Everyone Needs to Know. Oxford University Press, New York.
MacCoun, R.J., Reuter, P., 2001. Drug War Heresies: Learning from Other Vices Times, and Places. Cambridge University Press, Cambridge, UK.
MacCoun, R.J., Reuter, P., 2011. Assessing drug prohibition and its alternatives: a guide for agnostics. Annual Review of Law and Social Science 7, 61–78.
Moore, M.H., 1991. Drugs, the criminal law, and the administration of justice. The Milbank Quarterly 69, 529–560.
Morse, S.J., 2000. Hooked on hype: addiction and responsibility. Law and Philosophy 19, 3–49.
Morse, S.J., 2009. Addiction, science, and criminal responsibility. In: Farahany, N.A. (Ed.), The Impact of Behavioral Sciences on Criminal Law. Oxford University Press, New York, pp. 241–288.
Nadelmann, E.A., 1989. Drug prohibition in the United States: costs, consequences and alternatives. Science 245 (September 1), 939–947.
Nutt, D., 2012. Drugs without the Hot Air: Minimising the Harms of Legal and Illegal Drugs. UIT Cambridge LTD., Cambridge, UK.
Wilson, J.Q., 1990. Against the legalization of drugs. Commentary 89 (February), 21–28.
Zimring, F.E., Hawkins, G., 1992. The Search for Rational Drug Control. Cambridge University Press, Cambridge, UK.

Drugs: Illicit Use and Prevention

Wesley G Jennings, University of South Florida, Tampa, FL, USA
Jennifer M Reingle, University of Texas, Dallas, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Illicit drug use spans multiple risk behaviors, from the consumption of substances that are clearly prohibited by law (e.g., heroin, methamphetamine), to the misuse or use of prescription drugs that have a designated medical purpose. In this article, we will discuss the epidemiology of illicit drug use, including definitional issues that are unique to the substance use field, as well as preventative techniques that have been established to combat the varying types of illicit drug use.

Illicit drug use spans multiple risk behaviors, from the consumption of substances that are clearly prohibited by law (e.g., heroin, methamphetamine) to the misuse or use of prescription drugs that have a designated medical purpose. Therefore, a discussion of 'illicit drug use' necessarily groups several distinct types of illicit drug users into one heterogeneous category of 'illicit drug user.' This presents clear obstacles for prevention scientists, as each type of behavior (e.g., prescription drug misuse, marijuana use, stimulant use) is generally targeted differently in prevention and intervention programing. This article will discuss the epidemiology of illicit drug use, including definitional issues that are unique to the substance use field, as well as preventative techniques that have been established to combat the varying types of illicit drug use.

Defining 'Illicit Drug Use'

Illicit drug use includes several classes of amphetamine and amphetamine-like stimulants, cocaine, heroin, other opioids, and cannabis. Prevalence estimates generally defined 'use' as ingestion of any substance during one's lifetime. However, some studies of incidence (or the initiation of substance use) will focus on consumption of a drug within the previous year or month. When evaluating trends of substance use, these distinctions are important details that will often result in differential definitions of use. Substance users are generally categorized into several distinct groups for treatment purposes: users, substance misusers, abusers, and those suffering from physiological dependence (often referred to as addiction).

The blanket term for those who use illicit drugs but do not meet the criteria for misuse are generally referred to as 'users' (*see* Drug Control: International Aspects). This is the lowest risk category of substance users. Illicit drug misusers are persons who may have obtained prescription drugs from a physician for a legitimate purpose; however, they may be using the drugs for reasons other than prescribed, taking the drugs more often than prescribed, or taking too much (or too little) of the substance they were prescribed. This 'misuse' definition applies exclusively to prescription drugs, as there is no legitimate use for illegal (nonprescription) substances. Any illicit substance, including marijuana, prescription drugs, heroin, or other stimulants, may be abused. Abuse is distinguished by maladaptive patterns of substance use, manifesting distress, or impairment on behavior. Finally, dependence, or addiction, is characterized by a series of psychological symptoms including tolerance, withdrawal, drug-seeking behavior, and continued use of a drug despite negative consequences associated with its use. These individuals present the greatest obstacle for treatment purposes, as use of the drug has become a distinctive and routine component of their life, and psychotherapeutics must often be used to counter the effects of addiction on the brain (Center for Substance Abuse Treatment, 2006).

The Epidemiology of Illicit Drug Use

According to the Centers for Disease Control and Prevention, in 2009, 8.7% of youth and adults (over age 12) have used any illicit drug within the past month. The majority of this use is marijuana (6.6%); however, nonmedical use of prescription drugs is increasingly accounting for a larger proportion of illicit drugs consumed in the United States (2.8%) (SAMHSA, 2011). Since 2002, the rate of use for all illicit drugs (marijuana, all illicit drugs, and nonmedical psychotherapeutic drugs) has increased.

The role our social environment plays (e.g., parenting, peer drug use) in changing the trends of illicit drug use has received a great deal of attention from researchers (*see* Opiate Use and Abuse, History of; Drugs, Decriminalization of: Cultural Concerns). For instance, in a large cohort of youth in Germany and the United States, one study reported that changes in drug use in one's peer group was associated with new incidences (or occurrences) of marijuana use (Miller and Miller, 1997). Low socioeconomic position of the family and social network was associated with marijuana use particularly among men; however, women were more likely to begin using marijuana if they have a history of victimization (Miller and Miller, 1997). The influence of drug use in the peer group and family socioeconomic status was linked to heroin and cocaine use (for a review, see Galea et al., 2004). Having family members who use drugs is strongly linked to cocaine use (Burton et al., 1996), and disruption of the family structure has been consistently associated with use of multiple illicit substances (Galea et al., 2004). At the macro level, neighborhood disadvantage has also been associated with incidence of drug use, particularly for cocaine (Crum et al., 1996). Individual-level characteristics are consistently the strongest risk and protective factors for illicit substance use specifically, educational attainment, homelessness (Crum et al., 1996), race (American Indians and African Americans are particularly at risk), and gender (men consistently have higher rates of use than women).

Age of Onset of Illicit Drug Use

There is a considerable amount of evidence identifying (early) age of onset for illicit drug use as being a salient risk factor for subsequent dependence and experimentation with multiple substances (Hawkins et al., 1992; Kandel et al., 1992). For example, studies have reported the rates of lifetime alcohol dependence to be as high as 40% for those who initiated drinking at 14 years of age or younger (Grant and Dawson, 1997). In comparison, Grant and Dawson also reported a 10% lifetime alcohol dependence rate for those who initiated drinking at 20 years of age or older. Research conducted by Substance Abuse and Mental Health Services Administration (SAMHSA) has revealed that younger age cohorts are reporting more types of drugs used in their lifetime (SAMHSA, 1996). In addition, a number of studies have linked early initiation of illicit drug use to a myriad of adverse consequences including social role functioning, impaired ability to make normative developmental transitions from adolescence to adulthood (Newcomb and Bentler, 1988), reduced likelihood for successful marriages, occupational difficulties, and mental health issues (Jutkowitz and Eu, 1994; Hawkins et al., 1992; Robins and Pryzbeck, 1985; Yu and Williford, 1992). More recently, in a large-scale study examining the age of onset of illicit drug use across seven international sites (e.g., Netherlands; United States; Ontario, Canada; Sao Paulo, Brazil; Munich, Germany; Fresno, California; and Mexico City, Mexico), Vega and colleagues (2002) found that age of onset information revealed a temporal ordering of peak onset from alcohol to marijuana to other illicit drugs. Interestingly, the age of onset patterns for illicit drug use were relatively similar across geographical context, although the prevalence rates of use did vary considerably across international sites.

Prevention and Treatment of Illicit Drug Use

Illicit drug use prevention programs have historically been simply a component of larger programs (e.g., alcohol use programs, crime prevention, etc.), rather than the focus of a specific intervention (*see* Drug Control: International Aspects). These programs are generally brief, educational only (rather than interactive), and do not include multiple components (e.g., community mobilization), which have been demonstrated effective in other substance use prevention domains. Because illicit substances have become more available in recent decades, and in many cases, are prescribed by physicians, they pose a unique challenge to prevention scientists due to their perceived safety. Therefore, the problem of illicit drug use has been tackled from several different angles. Several progressive, theoretically informed prevention programs are detailed in the following section.

Theoretical Foundation for Illicit Drug Use Prevention

Primary prevention, or prevention prior to initiation of illicit drug use, generally targets youth during late childhood or early adolescence. This developmental phase is especially critical, as this is the first time in which youth are exposed to drug use, and problem behavior has yet to be present in many cases (Johnson et al., 1987). During this time, school-based interventions are typically delivered using one of three strategies: (1) mass communication or social marketing; (2) psychosocial implementation; or (3) public policy.

Mass communication, more commonly known as social marketing, has been utilized to change public perception about drug use frequency, prevalence, and outcomes (Catford, 2001). These programs are often developed by the target population, using focus groups or key informants. Social marketing campaigns target a specific age group and population and generally tailor their messages to be relatable to the population. These methods have been applied not only to illicit drug use but, more commonly, to alcohol and tobacco use among youth. Social marketing has the potential for a broad impact given advancements in advertising technology and targeted marketing.

Psychosocial models are posited on Bandura's (1977) social learning theory, operating on the assumption that behavior is learned from one's surroundings. Using this philosophy, prevention programing incorporates multiple social dimensions, including the peer group, family, and community. Modeling positive, antidrug use behavior, resistance skills, and normative assumptions about drug use are targeted and reinforced. Role playing is commonly used to reinforce techniques learned in the classroom. Given that perceptions of drug use among one's peers are often severely distorted, these techniques are used to bridge both education and skills training to reinforce the notion that drug use is not a 'normal' behavior.

Similarly, a long line of research has demonstrated that illicit drug use among adolescents is intrinsically linked to their perceptions of their peers who use illicit drugs and/or having friends who report using illicit drugs (Jennings and Akers, 2011). Akers (1977) argues that it is within the social learning process wherein youth acquire their own definitions favorable/unfavorable toward using illicit drugs through social interaction and socialization. For instance, illicit drug use, similar to any type of delinquent behavior, generates either positive (beneficial) or negative (adverse) consequences, which have implications for reinforcement. In this regard, Burgess and Akers (1966) are most well known for their modifications to Sutherland's (1947) original nine propositions of differential association theory. Specifically, Burgess and Akers (1966, pp. 132–145) offered the following seven principles exemplifying the process wherein social learning occurs that have direct implications for illicit drug use:

1. Criminal behavior is learned according to the principles of operant conditioning (reformulation of Sutherland's principles 1 and 8).
2. Criminal behavior is learned both in nonsocial situations that are reinforcing or discriminative and through that social interaction in which the behavior of other persons is reinforcing or discriminative for criminal behavior (reformulation of Sutherland's principle 2).
3. The principal part of the learning of criminal behavior occurs in those groups that comprise the individual's major source of reinforcements (reformulation of Sutherland's principle 3).

4. The learning of criminal behavior, including specific techniques, attitudes, and avoidance procedures, is a function of the effective and available reinforcers, and the existing reinforcement contingencies (reformulation of Sutherland's principle 4).
5. The specific class of behaviors that are learned and their frequency of occurrence are a function of the reinforcers that are effective and available, and the rules or norms by which these reinforcers are applied (reformulation of Sutherland's principle 5).
6. Criminal behavior is a function of norms that are discriminative for criminal behavior, the learning of which takes place when such behavior is more highly reinforced than noncriminal behavior (reformulation of Sutherland's principle 6).
7. The strength of criminal behavior is a direct function of the amount, frequency, and probability of its reinforcement (reformulation of Sutherland's principle 7).

Or stated more explicitly by Akers (1998: p. 50),

> The basic assumption in social learning theory is that the same learning process in a context of social structure, interaction, and situation, produces both conforming and deviant behavior. The difference lies in the direction … [of] the balance of influences on behavior.

> The probability that persons will engage in criminal and deviant behavior is increased and the probability of their conforming to the norm is decreased when they differentially associate with others who commit criminal behavior and espouse definitions favorable to it, are relatively more exposed in-person or symbolically to salient criminal/deviant models, define it as desirable or justified in a situation discriminative for the behavior, and have received in the past and anticipate in the current or future situation relatively greater reward than punishment for the behavior.

Akers and colleagues (1979) provided some of the earliest evidence for social learning's applicability as a theoretical framework to explain illicit drug use, specifically alcohol and marijuana use among adolescents. Subsequent to this initial explicit test, a series of other studies have evidenced and by and large replicated these findings (Krohn et al., 1982, 1985). Similarly, Johnson et al. (1987) investigated the role of peers in adolescent drug use, and their results suggested that associating with peers who exhibited an excess of definitions favorable toward drug use exerted a positive and significant direct effect on the youth's own drug use behavior. Social learning theory, particularly differential association and definitions, has also been shown to have a significant effect on cigarette use (Monroe, 2004), alcohol use, and marijuana use including in cross-cultural studies (Miller et al., 2008).

Similarly, Erikson's stage theory of adolescent development suggests that early adolescents use drugs to mimic what youth perceive to be an 'adult behavior.' The desire for independence and the well-documented 'maturity gap' among early adolescents has led to the development of several 'affective' approaches to illicit drug prevention. Although this affective training is not used by itself, it has been effective when used in conjunction with social resistance skills training and educational components (Johnson et al., 1990).

Public policy has also been implemented in an attempt to deter use of illicit drugs using fear of sanctions (*see* Opiate Use and Abuse, History of). Changes in tobacco policies (specifically, taxes) have been demonstrated as being effective in reducing cigarette smoking; however, relatively little is known about how policy changes influence illicit drug use. Regulation of these drugs provides unique challenges due to the either illegal or controlled nature of illicit substances. A great deal of research is currently in progress to examine the influence of marijuana policy (e.g., decriminalization) on use and negative health outcomes (such as impaired driving and mortality) (*see* Drugs, Decriminalization of: Cultural Concerns). An example of each of these strategies for illicit drug use prevention follows.

Social Marketing

Mass media has been identified as one of the most promising tactics for reducing drug use in future generations. These programs seek to increase knowledge and awareness; however, their potential for inducing behavior change has been questioned (Flay, 1987). Some of these programs are extremely large scale, as the Truth campaign and Just Say No are recognizable to people throughout the United States. These television advertisements are targeted toward youth and seek to educate and deter them from substance initiation.

One example of a social marketing campaign aimed at drug use is the National Youth Anti-Drug Media Campaign (Scheier and Grenard, 2010), often known as 'The Anti-Drug' campaign. This campaign targeted early adolescents and was fairly broad in nature. The advertisements reinforced the notion that drug use would prohibit a person from doing what they want with their life, such as engaging in social activities or professional/educational success. Youth who use drugs are depicted as loners who suffer from several negative effects of drug use. These messages reinforce refusal skills and remind youth that direct refusal is effective in diverting peer pressure. An evaluation of this campaign showed no effect on marijuana use among youth (Scheier and Grenard, 2010).

Psychosocial Intervention

Psychosocial interventions are generally targeted toward youth in primary school, with curricula administered by teachers or other school personnel. These programs teach youth to resist peer pressure, educate youth on the negative consequences of drug use, and provide opportunities to practice situational refusal skills. Programs that have relied primarily on education or scare tactics (for instance, DARE) have largely been unsuccessful; however, hands-on skills training programs have demonstrated success in preventing initiation of illicit drug use.

One example of a well-known school-based psychosocial intervention is known as Life Skills Training (Botvin et al.,

1995). The program included 15 class sessions taught during the 7th grade, with each session focusing on skills training and providing information on resisting social influences toward drug use. Specifically, these sessions used cognitive-behavioral skills training techniques, including self-esteem building, anxiety management, effective communication, positive relationship development, and assertiveness. Each behavior was rehearsed in a peer group, and feedback and reinforcement was provided. Homework assignments were administered to provide additional opportunities for practice. The information provided in educational sessions focused on immediate negative effects of drug use (which may be more salient and resonate with adolescents), rather than long-term, chronic outcomes of drug use. Two short booster sessions were provided (one in 8th grade and another in 9th grade) to reinforce the material learned in the initial intervention. This intervention has been shown to reduce illicit drug use in the intervention group (compared to controls) (Botvin et al., 2000).

Public Policy

Several policies have been enacted in an attempt to combat illicit drug use, including the ongoing 'war on drugs,' zero tolerance policies, and mandatory drug testing (*see* Urban Policy in North America; Crime Policy in the United States). The later has recently been applied in school settings, creating a stir of public interest in drug policy. Several US Supreme Court decisions have upheld the ability of the school district to randomly drug test students (*Veronia School District v. Acton*, 1995; *Board of Education of Independent School District No. 92 of Pottawatomie County v. Earls*, 2002). After these decisions, several anecdotal reports of reductions in drug use by students were reported by teachers and school administrators (Yamaguchi et al., 2009). Rigorous evaluations of the impact of drug testing on schools have not been conducted due to serious methodological difficulties; however, trend analyses have failed to show a reduction in self-reported illicit drug use before and after drug testing was instituted in schools (Yamaguchi et al., 2009).

Multicomponent Interventions

Drug use is not a single-dimensional behavior, meaning that multiple influences might cause someone to initiate drug use. Characteristics of one's parents and parenting style, selected peer groups, key adult role models, mass media, and community policies are all related to drug use. Environmental characteristics, such as access to alcohol and drugs, unsupervised time, and peer pressure further complicate the etiology of illicit drug use. Therefore, a number of prevention programs have been identified as 'multicomponent,' or using more than one of the aforementioned three models to target and prevent illicit drug use among youth.

One example of an effective multicomponent drug use prevention program was implemented in 27 Native American community elementary schools (Schinke et al., 2000). Each school was randomly assigned to one of three arms: (1) cultural content only; (2) cultural content and cognitive-behavioral skills training; and (3) cultural content, cognitive behavioral skills training, and community mobilization, media releases, and community involvement in the intervention. Each of the groups received 'cultural content,' or Native American stories, legends, and values lessons that addressed substance use issues and general health promotion. The cognitive-behavioral skills training condition included 15 one-hour sessions, which trained youth on modeling, resistance, and behavioral refusal skills tailored to substance use prevention. The intervention was tailored to the contemporary Native American culture. The final arm of the intervention included community members, including youth's families, teachers, school guidance counselors, residents, law enforcement officials, and owners/managers of community businesses frequented by youth. Evaluations of this project have shown that the youth who participated in the intervention study had lower rates of tobacco, alcohol, and marijuana use than those who did not receive the cognitive-behavioral intervention. The community intervention arm, however, did not have any additional benefit on substance use compared to the cognitive-behavioral intervention.

Social Learning Theory Interventions

Finally, the principles of social learning theory have been applied to prevention, treatment, and rehabilitation programs as well (Andrews and Bonta, 2003; Cullen et al., 2003). Specifically, applications of these types of programs are generally designed to manipulate definitions/attitudes, associations, group norms, imitation, reinforcement/punishment, discriminative stimuli, conditioning, or other relevant factors that are likely to be associated with deviant behavior generally and substance use more specifically. These programs include: group therapies and self-help programs; peer programs; gang interventions; family and school programs; teenage drug, alcohol, and delinquency prevention/education programs; and other private and public programs in correctional, treatment, and community facilities and programs. Akers and Sellers (2009) have reviewed a number of these intervention and prevention strategies, and in this vein, they note that while large effect sizes are often not demonstrated for the importance of social learning principles, the literature is relatively consistent in reporting that among prevention and intervention programs that employ social learning principles these effects are typically larger than prevention and intervention programs that utilize alternative principles and strategies other than social learning. The following quotes from respected researchers in the field best summarize the effects of preventions and interventions that emphasize social learning principles:

> [R]esearch on correctional programs supplies strong and consistent support for theories – such as differential association/social learning theory – that link offending to antisocial associations and to the internalization of antisocial values. …'[C]ognitive-behavioral' programs tend to achieve higher reductions in recidivism than other treatment modalities. … These programs

specifically target for change antisocial values and crime-excusing rationalizations, they reward prosocial attitudes and behavior, and they seek to isolate offenders from pro-criminal associations. (Cullen et al., 2003: p. 353)

[The relative success of cognitive-behavioral programs is consistent with] "social learning theory [that] is broader than behavioral reinforcement theory because it includes as variables cognition, verbalization, and social modeling to explain (and to change) behavior patterns." (Pearson et al., 2002: p. 480)

The use of cognitive-behavioral interventions is consistent with the conclusion of other meta-analyses that interventions based on social learning theory are particularly effective (Lösel, 1995). Social learning theory asserts that behavior is learned through several processes, including modeling, imitation, and differential reinforcement (Bandura, 1979). He is also likely to associate with people who encourage his violent behavior (differential association). Cognitions that support specific behaviors are known as "definitions" (Akers, 1985). Definitions, then, could be seen as one form of the problematic cognitions that the therapist needs to address (Ellis and Sowers, 2001: pp. 91–92; emphasis in original).

Conclusions

In conclusion, the need for complex, multicomponent interventions and prevention programing to target illicit drug use is clear. The vast majority of drug users will first experiment with substances in early adolescence (ages 11–14); therefore, prevention programing must begin in elementary and primary schools. The etiology of illicit drug use is complex and multidimensional; therefore, programs that have addressed only one component have been largely unsuccessful in preventing drug use initiation. For instance, a program might address drug-using peers; however, if a parent uses cocaine, these unaddressed parental influences will continue to reinforce pro-drug behavior and intentions. Overall, illicit drug use continues to pose a growing problem to public health professionals, policy makers, and prevention scientists.

See also: Drug Control: International Aspects; Drugs, Decriminalization of: Cultural Concerns; Empirical Legal Studies; Opiate Use and Abuse, History of; Urban Policy in North America.

Bibliography

Akers, R., 1977. Deviant Behavior: A Social Learning Approach, 2nd ed. Wadsworth, Belmont, CA7.

Akers, R.L., Krohn, M.D., Lanza-Kaduce, L., Radosevich, M., 1979. Social learning and deviant behavior: a specific test of a general theory. American Sociological Review 44, 636–655.

Akers, R.L., Sellers, C.S., 2009. Criminological Theories: Introduction, Evaluation, and Application, fifth ed. Roxbury Publishing, Los Angeles, CA.

Akers, R.L., 1985. Deviant Behavior: A Social Learning Approach, third ed. Wadsworth, Belmont, CA.

Akers, R.L., 1998. Social Learning and Social Structure: A General Theory of Crime and Deviance. Northeastern University Press, Boston, MA.

Andrews, D.A., Bonta, J., 2003. The Psychology of Criminal Conduct, third ed. Anderson Publishing, Cincinnati, OH.

Bandura, A., 1977. Self efficacy: toward a unifying theory of behavioral change. Psychological Review 84, 191–215.

Bandura, A., 1979. Social Learning Theory. Prentice Hall, Englewood Cliffs, NJ.

Botvin, G.J., Baker, E., Dusenbury, L., Botvin, E.M., Diaz, T., 1995. Long-term follow-up results of a randomized drug abuse prevention trial in a white middle-class population. JAMA 273, 1106–1112.

Botvin, G.J., Griffin, K.W., Diaz, T., Scheier, L.M., Williams, C., Epstein, J.A., 2000. Preventing illicit drug use in adolescents: long-term follow-up data from a randomized control trial of a school population. Addictive Behaviors 25, 769–774.

Burgess, R.L., Akers, R.L., 1966. A differential association-reinforcement theory of criminal behavior. Social Problems 14, 128–147.

Burton, R.P., Johnson, R.J., Ritter, C., et al., 1996. The effects of role socialization of the initiation of cocaine use: an event history analysis from adolescence into middle adulthood. Journal of Health and Social Behavior 37, 75–90.

Catford, J., 2001. Illicit drugs: effective prevention requires a health promotion approach. Health Promotion International 16, 107–110.

Center for Substance Abuse Treatment, 2006. Prescription medications: misuse, abuse, dependence, and addiction. Substance Abuse Treatment Advisory 5 (2).

Crum, R.M., Lillie-Blanton, M., Anthony, J.C., 1996. Neighborhood environment and opportunity to use cocaine and other drugs in late childhood and early adolescence. Drug and Alcohol Dependence 43, 155–161.

Cullen, F.T., Wright, J.P., Gendreau, P., Andrews, D.A., 2003. What correctional treatment can tell us about criminological theory: implications for social learning theory. In: Akers, R.L., Jensen, G.F. (Eds.), Social Learning Theory and the Explanation of Crime: A Guide for the New Century. Advances in Criminological Theory, vol. 11. Transaction Publishers, New Brunswick, NJ, pp. 339–362.

Ellis, R., Sowers, S., 2001. Juvenile Justice Practice: A Cross-disciplinary Approach to Intervention. Wadsworth/Thomson Learning, Belmont, CA.

Flay, B.R., 1987. Evaluation of the development, dissemination, and effectiveness of mass media health programming. Health Education and Research 2, 123–129.

Galea, S., Nandi, A., Vlahov, D., 2004. The social epidemiology of substance use. Epidemiologic Reviews 26, 36–52.

Grant, B.F., Dawson, D.A., 1997. Age of onset of alcohol use and its association with DSM-III-IV alcohol abuse and dependence: results from the national longitudinal alcohol epidemiologic survey. Journal of Substance Abuse 9, 103–110.

Hawkins, J.D., Catalano, R.F., Miller, J.Y., 1992. Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: implications for substance use prevention. Psychological Bulletin 112, 64–105.

Jennings, W.G., Akers, R.L., 2011. Social learning theory. In: Bryant, C.D. (Ed.), The Handbook of Deviant Behavior. Routledge, New York, pp. 106–113.

Johnson, R.E., Marcos, A.C., Bahr, S.J., 1987. The role of peers in the complex etiology of adolescent drug use. Criminology 25, 323–340.

Johnson, C.A., Pentz, M.A., Weber, M.D., Dwyer, J.H., Baer, N., MacKinnon, D.P., Hansen, W.B., Flay, B.R., 1990. Relative effectiveness of comprehensive community programming for drug abuse prevention with high-risk and low-risk adolescents. Journal of Consulting and Clinical Psychology 58, 447–456.

Jutkowitz, J.M., Eu, H., 1994. Drug prevalence in Latin American and Caribbean countries: a cross-national analysis. Drugs: Education, Prevention and Policy 1, 199–252.

Kandel, D.B., Yamaguchi, K., Chen, K., 1992. Stages of progression in drug involvement from adolescence to adulthood: further evidence for the gateway theory. Journal of Studies on Alcohol 53, 447–457.

Krohn, M.D., Akers, R.L., Radosevich, M., 1982. Norm qualities and adolescent drinking and drug behavior: the effects of norm quality and reference group on using and abusing alcohol and marijuana. Journal of Drug Issues 12, 343–359.

Krohn, M.D., Skinner, W.F., Massey, J.L., Akers, R.L., 1985. Social learning and adolescent cigarette use: a longitudinal study. Social Problems 32, 455–473.

Lösel, F., 1995. The efficacy of correctional treatment: a review and synthesis of meta evaluations. In: McGuire, J. (Ed.), What Works: Reducing Reoffending, Guidelines from Research and Practice. Wiley, New York, pp. 79–111.

Miller, D.S., Miller, T.Q., 1997. A test of socioeconomic status as a predictor of initial marijuana use. Addictive Behaviors 22, 479–489.

Miller, H.V., Jennings, W.G., Alvarez-Rivera, L.L., Miller, J.M., 2008. Explaining substance use among Puerto Rican adolescents: a partial test of social learning theory. Journal of Drug Issues 38, 261–284.

Monroe, J., 2004. Getting a puff: a social learning test of adolescents smoking. Journal of Child and Adolescent Substance Abuse 13, 71–83.

Newcomb, M.D., Bentler, P.M., 1988. Consequences of Adolescent Drug Use. Sage, Newbury Park, CA.

Pearson, F.S., Lipton, D.S., Cleland, C.M., Yee, D.S., 2002. The effects of behavioral/cognitive-behavioral programs on recidivism. Crime & Delinquency 48, 476–496.

Robins, L.N., Pryzbeck, T.R., 1985. Age of Onset of Drug Use as a Factor in Drug and Other Disorders. NIDA Research Monograph No. 56. National Institute of Drug Abuse, Rockville, MD.

SAMHSA, 1996. Trends in the Incidence of Drug Use in the United States, 1919/1992. U.S. Department of Health and Human Services, Rockville, MD.

SAMHSA, 2011. Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health. Available from: http://oas.samhsa.gov/nsduh.htm.

Scheier, L.M., Grenard, J.L., 2010. Influence of a nationwide social marketing campaign on adolescent drug use. Journal of Health Communication: International Perspectives 15, 240–271.

Schinke, S.P., Tepavac, L., Cole, K.C., 2000. Preventing substance use among Native American youth: three-year results. Addictive Behaviors 25, 387–397.

Sutherland, E.H., 1947. Principles of Criminology, fourth ed. J. B. Lippincott, Philadelphia.

Vega, W.A., Aguilar-Gaxiola, S., Andrade, L., Bijl, R., Borges, G., Caraveo-Anduaga, J., et al., 2002. Prevalence and age of onset for drug use in seven international sites: results from the international consortium of psychiatric epidemiology. Drug and Alcohol Dependence 68, 285–297.

Vernonia School District v. Acton, 1995. Supreme Court of the United States.

Yamaguchi, R., Johnston, L.D., O'Malley, P.M., 2009. Relationship between student illicit drug use and school drug-testing policies. Journal of School Health 73, 159–164.

Yu, J., Williford, W.R., 1992. The age of alcohol onset and alcohol, cigarette and marijuana use patterns: an analysis of drug use progression of young adults in New York State. International Journal of the Addictions 27, 1313–1323.

Drug Use and Abuse: Psychosocial Aspects

Catherine Darker and Lucy Whiston, Trinity College, Dublin, Ireland

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by P.M. O'Malley, volume 6, pp. 3861–3866, © 2001, Elsevier Ltd.

Abstract

This article begins with a definition of drug use, followed by a consideration of the factors that make this particular class of behavior unique. Distinctions are made among various classes of drugs, based on their legal status and social acceptability. The typical stages of involvement in drugs are described, distinctions among use, abuse, and dependence are made, and polydrug use is discussed. Finally, consideration is given to various other psychosocial correlates of use, including particular attitudes and beliefs, as well as the individual's level of integration into society in general, and into certain key institutions in particular.

Psychoactive Drugs

The use of psychoactive drugs is not a new phenomenon. There are numerous adverse consequences of drug use, the most important of which is drug abuse. The term 'drugs,' which is commonly used today, is a shorthand for 'psychoactive drugs,' i.e., compounds or substances which alter the neurological functioning and/or structure of the organism in such a way as to directly influence perception, mood, consciousness, and/or energy levels. There are many compounds that have such effects, ranging from those occurring naturally in plants (e.g., cocaine, caffeine, nicotine, cannabis, opium, psilocybin) and natural processes (fermentation of fruit and other plant products containing carbohydrates, yielding alcohol) to compounds that are synthesized (e.g., benzodiazepines, amphetamines, Lysergic acid diethylamide (LSD), 'ecstasy,' and 1-(1-phenylcyclohexyl)piperidine (PCP)).

Distinguishing Characteristics of Drug-Using Behaviors

There is one central characteristic that distinguishes psychoactive drug use from most other classes of behavior. Drugs influence the neurological system directly and provide pleasure or other forms of reinforcement (Koob and Nestler, 1997). In essence, drug use involves a shortcut to the pleasure centers of the brain. As a result of this shortcut, drug-using behavior can come to take a higher position in the hierarchy of needs displacing many behaviors that are adaptive such as eating, forming appropriate bonds to family and community, working, or performing well academically. It can also disrupt fundamental psychological processes essential to performing a wide range of other behaviors or tasks by altering or impeding perception, memory, decision making, motivation, and energy levels. Second, the direct reinforcing properties of psychoactive drugs on the neurotransmitters in the brain can alter the structure and function of the neurological system itself, such that a need is created for the drug in order for the individual to attain what had previously been a normal level of neurological functioning (Everitt and Robbins, 2005; Koob and Nestler, 1997). In other words, the user can become dependent on or addicted to the drug; thus, the underlying motivation for using a drug can shift quite dramatically from what might have motivated use at earlier stages of involvement.

Illicit versus Licit Drug Use

There are many psychosocial determinants of the use of drugs, two of the most important of which are the legal and normative status of particular substance-using behaviors in the society. The primary focus in this article will be on psychoactive substances, the use of which is considered illicit in most societies. Tobacco and alcohol are generally licit to use and legal to possess; but because they are dealt with at some length in other entries, they will receive only very general treatment here. It should be noted, however, that even the use of alcohol and tobacco by people under certain ages is considered illegal and illicit. Three general social classifications of drugs will be distinguished, and the importance of their classification considered: psychotherapeutic drugs, legal drugs, and illegal drugs.

Psychotherapeutic Drugs

Certain drugs can be either licit or illicit to use, depending on the conditions under which use occurs; and they can be legal or illegal to possess, depending on how they have been obtained. A number of drugs have medical applications, so their use is considered as legitimate for the medically supervised therapeutic treatment of particular conditions. These drugs include tranquilizers (most of which are benzodiazepines, used to alleviate anxiety), sedatives (most of which are barbiturates, used to induce sleep), amphetamines (used to treat hyperactivity disorders), and various opium derivatives and their synthetic analogs (used to control pain, coughing, and diarrhea). All of these drugs have regulated access, through the requirements of a doctor's prescription and pharmacy dispensation. However, the sale, possession, or the use of these substances outside of prescribed medical regimen is considered illicit and illegal, in part because all of these drugs have the potential to produce dependence in the user; and dependence is presumed to be more likely to develop under conditions of unsupervised use.

The Legal Drugs

Certain drugs have emerged as legitimate to use in most societies outside of any medical regimen and without any therapeutic purpose, usually because at the time of their emergence, they are presumed to be safe. These include alcohol, nicotine, and caffeine. Most of these substances have been in fairly widespread use for some time in different cultures. Use of these drugs is widespread not only because their use is legally and normatively accepted but also because their legal status allows those who sell them to engage in massive advertising and promotional campaigns aimed at encouraging and facilitating consumption of the drug. Although there have been some concerns in the past about the health and societal consequences of alcohol and tobacco, such concerns increased considerably over the course of the twentieth century. This has led to increased stigmatization of their use and legal controls on their sale and distribution.

The Illegal Drugs

In the twentieth century most other drugs have been deemed illegal and illicit to use, possess, or sell in most societies. This illegal status is based largely on the negative impact on the functioning of the individual, family, and/or larger society. Various international treaties have had the effect of making the legal status of these drugs fairly homogenous throughout the world. The 1961 Single Convention outlawed plant-based substances such as derivatives of coca, marijuana, and opium (United Nations, 1972). Later the 1971 Convention made synthetic drugs such as LSD and MDNA illegal while the 1988 Convention made the trafficking of these drugs illegal (United Nations, 1971, 1988).

The boundaries of social legitimacy and illegality of the use of various drugs are themselves subject to change, as knowledge accumulates about the effects of the drugs and also about the effects of the control mechanisms intended to control their use. Most recently, as the consequences and side effects of substances known as 'head shop' products, 'legal highs,' or 'designer drugs' became apparent many countries have reclassified these drugs as illegal as much as possible (UNODC, 2013). Conversely, there is an ongoing vigorous policy debate over the possible reclassification of currently illegal marijuana. Over the past few years a number of countries including Australia, Spain, and the United States have shifted away from prohibition with an introduction of a quasilegal position for marijuana (Caulkins et al., 2012). One of the major reasons given by those who oppose the legalization of marijuana or other illegal drugs is that legalizing these drugs would unleash powerful marketing forces similar to alcohol and tobacco (Inciardi and McBride, 1991). Stemming from this some argue that reclassification would destigmatize marijuana making it seem less dangerous. Alongside increased access this would result in higher prevalence of use (Roffman, 2013).

Stages of Involvement

The research literature on substance use has distinguished a number of stages of involvement with a drug (Clayton, 1992). In general, the stages of use of a drug can be described as abstention, initiation, continuation, progression, dependence, and relapse. In addition, the cessation of use can occur after initiation, progression, dependence, or relapse. It is well recognized that addiction is a chronic condition in which relapse generally occurs a number of times (DiClemente et al., 2011).

The notion of stages of involvement applies not only to individual drugs, but also across drugs. Drug use typically begins with more socially acceptable 'Gate way drugs' such as alcohol and tobacco (Kirby and Barry, 2012). The use of certain substances tends to precede the use of other ones with a high degree of regularity (Kandel, 1975; Kandel et al., 1992). Marijuana use, for example, is usually preceded by the initiation of cigarette smoking, and the two behaviors remain highly correlated at later ages. Alcohol use also tends to precede the use of tobacco and marijuana; and marijuana use, in turn, usually precedes use of most of the so-called 'hard' drugs like stimulants, opiates, and hallucinogens (Kirby and Barry, 2012). The regularity of this pattern of progressive involvement has led to work aimed at predicting specific details of these transitions.

Psychiatric researchers have been most concerned with defining, understanding, and intervening in the later stages of involvement. Explaining progression in the earlier stages of involvement, and trying to intervene in that progression has fallen largely to the social sciences. Though many social scientists who work in this field are informally called epidemiologists or etiologists, in fact they have tended to come mostly from psychology (including social, clinical, counseling, and community psychology), sociology, social work, anthropology, and, more recently, public health and economics. Two very general fields have evolved here: first, the study of the natural phenomena and second, the development of interventions (e.g., education, prevention, persuasive communication). Because prevention is the subject of other entries, it is mentioned here only to note that the scientists and program developers in prevention tend to have quite similar disciplinary backgrounds and theoretical underpinnings as those who work on the etiology and epidemiology of substance use. Social learning theory has been widely adopted in all of these subareas, for instance.

Initial Trajectory of Drug Use

There are a great many individual risk factors for psychoactive drug use (e.g., Clayton et al., 1995; Glantz and Pickens, 1992). Substance use disproportionately occurs among those from socioeconomically deprived backgrounds (Redonnet et al., 2012). Some of the strongest risk factors for substance use among youths and young adults have to do with their degree of attachment to, or integration with, key institutions in society. At the broadest level, deviance from societal norms in general, as reflected by delinquent behavior, is a strong correlate of virtually all forms of illicit drug use (including underage drinking and smoking). It should not be surprising that one class of illicit behaviors (drug use) correlates strongly with a range of other illicit behaviors. It simply means that people who are willing to violate legal and normative constraints in

one domain of behavior are willing to do so in others, as well. One result of this fact is that the correlates and determinants of the various 'problem behaviors' among youth tend to be the same. Among them are attachments to various other institutions, but in particular the family, school, and church (Brook et al., 1990).

Young people who have a low degree of attachment or 'bonding' to school – meaning that they are not performing well, do not like school, and may frequently be absent – are at particular risk of involvement with illicit drug use. For example, those who drop out of school early, are more likely to become involved in substance use (McCaffrey et al., 2010).

Use, Abuse, and Dependence

One important distinction in psychosocial aspects of drugs has been between use, abuse, and dependence. Much debate and controversy surrounds these terms, and exactly what defines use, abuse, and dependence with further complications arising when it comes to operationalizing these distinctions.

In theory, the distinction lies in whether the drug-using behavior has reached the stage at which it is adversely affecting the normal functioning and well-being of the individual and, therefore, has become 'abuse.' Is the individual's drug use disrupting normal emotional, cognitive, social, or work functioning, including school performance for those in school, or is it endangering physical health? Is it interfering with the user's functioning in key social roles, such as spouse, parent, child, friend, student, worker, and citizen?

Some research on the reasons adolescents give for using various drugs may help to elucidate the theoretical distinction between use and abuse (Johnston and O'Malley, 1986). Young people who have used a drug only once or twice usually give as their primary reason curiosity, 'to see what it was like.' Those who have used drugs only on a reasonably limited number of occasions, but have gone beyond experimentation, most often mention reasons such as 'to feel good or get high' or 'to have a good time with my friends,' which might be called social and recreational reasons for use.

However, those who have used a drug with some frequency start to mention what might be called psychological-coping motives for their use, e.g., 'to get through the day,' 'to get away from my problems,' 'because of anger or frustration,' 'to relax or relieve tension.' Therefore, abuse emerges when use of the drug continues regardless of recurrent social, occupational, physical, or psychological problems such as psychological coping (WHO, 1994).

In general, dependence is seen to emerge when one needs something to feel good, function, or survive (WHO, 1994). Substance dependence is defined by the World Health Organization as "A cluster of physiological, behavioral, and cognitive phenomena in which the use of a substance or a class of substances takes on a much higher priority for a given individual than other behaviors that once had greater value" (Isaac et al., 1994). This includes both, physical and psychological elements. For dependence to be diagnosed, at least three symptoms such as craving, withdrawal symptoms, increased tolerance, loss of control of drug use, neglect of other interests, and the persistent use despite clear evidence of harmful consequences need to be present together over the previous 12 months (Janca et al., 1994). In 2012, 10–13% of all drug users worldwide were dependent users (UNODC, 2013).

Polydrug Use

Polydrug use means using two or more drugs in combination, using one drug to counteract the effects (or the after effects) of another or using different drugs at different times over a short period of days or weeks. Most illegal drug users are polydrug users. Using a single psychoactive drug can be dangerous; using more than one significantly increases the risks. Polydrug use combinations could increase the likelihood of physical or psychological harm (Forsyth, 1996; Earleywine and Newcomb, 1997; Feigelman et al., 1998; Riley et al., 2001) and difficulties in engaging in treatment (John et al., 2001). In particular, if two drugs of the same type, such as the depressants heroin and alcohol, are used together there are greatly increased risks of accidents, overdose, and death. Other common, dangerous drug combinations include cannabis with alcohol, ecstasy with alcohol, heroin with sedatives, and amphetamines with sedatives. There are also challenges in developing effective treatment approaches for polydrug users because of the different psychological and physiological effects of different substances (Kenna et al., 2007).

Attitudes and Beliefs About Drugs

Social Connotations of Use

One of the many influences on substance use is the symbolic meaning the individual sees it as having within society at large or within various social circles. It is obvious that cigarettes and alcohol have acquired certain associations in peoples' minds as a result of intentional manipulation through advertising. Less obvious perhaps is the fact that even unadvertised substances acquire social connotations, and that these may change considerably over time. The high degree of association of marijuana and LSD came to have with the counterculture already has been mentioned. When cocaine use first started to grow in the late seventies, it was very expensive and thus was used disproportionately by the rich and sometimes famous. As a result it acquired a certain caché that made its use more desirable to many, whereas a number of other drugs, like heroin, were more associated with marginal groups and the less affluent. It seems likely that adolescents come to see inhalants as kids' drug as they get older, and that their use declines as a result.

Perceived Risk

Whether drugs are seen as dangerous to the user or to others can also affect how they are seen more generally. The dramatically lowered social acceptability of tobacco over the past two decades illustrates this point. As smoking came to be seen as carrying high risk of morbidity and mortality, it became a less popular behavior, primarily with adults. Social norms, and regulations and laws, shifted quite dramatically. Taxes were

used to increase the price of tobacco, mass media campaigns were launched, and as the evidence grew of the adverse effects on others from their passive exposure to tobacco smoke, smoking in many public places, once an assured right, became a frequently denied privilege. These various steps have led to reduced prevalence of smoking as well as reduced initiation of smoking. Policies which targeted work places have also led to reduced consumption and increased tobacco cessation (Wilson et al., 2012). A similar thing happened with alcohol, when the many adverse effects of drinking came to be more widely recognized. This was illustrated through the impact of price hikes on alcohol through taxation which led to a decrease in consumption in addiction to lower levels of heavy drinking (Wagenaar et al., 2009).

A number of the illicit drugs have also had considerable transformations in their social acceptability, as their adverse consequences were documented and made widely known. In the late 1970s, as cocaine use became more widespread, there was little convincing evidence of adverse effects, as 'experts' proclaimed that cocaine was neither lethal nor addictive. The evidence eventually accumulated to prove them wrong, but it took some years because there turned out to be a long lag-time on average between the initiation of use and the development of dependency and other serious problems. Researchers at the University of Michigan have demonstrated the importance of perceived risk as a determinant of trends not only in cocaine use, but also in the use of marijuana and other drugs (Bachman et al., 1998; Johnston et al., 2000). Among other things they have shown that a turnaround in the aggregate-level perceived risk associated with a drug by adolescents often presages a turnaround in actual use a year later.

Research Traditions

Many subareas of research on drugs have evolved, particularly in the past 40–50 years. These include etiology, epidemiology, neurology, psychopharmacology, molecular biology, prevention, and treatment. The psychiatric tradition in epidemiology and treatment has emphasized accurately diagnosing and effectively assigning treatment regimens to individual cases. The other social sciences have looked more at drug use in the normal population, with an emphasis on causes, correlates, and consequences. A number of large nationally representative epidemiological studies have been conducted with a focus on drug use. These include the National Comorbidity Survey Replication (Kessler et al., 2004), the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) (Grant et al., 2004), as well as the National Survey on Drug Use and Health (NSDUH) (SAMHSA, 2011).

The fact that addiction has long been conceptualized as a disease within certain scientific circles has led to terminology that is medical in nature in defining the subfields of research (e.g., epidemiology, etiology, prevention, treatment). In fact, Alan Leshner, the current director of the National Institute on Drug Abuse, has asserted that drug addiction is a disease of the brain, evidenced by lasting structural and functional changes in the brain induced by drug use, but that drug use short of addiction is a preventable behavior (Leshner, 1999).

The importance of the disease concept, for which there is increasing neurological evidence, is that it implies that drug dependence should be responded to as more than just a moral failing. It implies that dependent users have important changes in their neurotransmitter systems that reduce to a considerable degree the extent to which they have reasonable control over their drug-using behavior.

Three US Nationally representative studies of adults include the National Co-morbidity Survey Replication (Kessler et al., 2004), the NESARC (Grant et al., 2004), and the NSDUH (SAMHSA, 2011; Merikangas and McClair, 2012).

Future Directions

It has become increasingly clear that both licit and illicit substance use must be viewed in terms of age, period, and cohort effects, since all three different types of change have been observed and documented (Johnson and Gerstein, 2000; Johnston et al., 2000). The factors explaining each of these three types of change are likely to be different, as is the timing of their action. The continual emergence of new illicit drugs onto the scene will yield new drug-taking phenomena to be understood and influenced. Undoubtedly more will be learned about the neurological action and effects of the many drugs, which may well change the importance of certain psychosocial factors, such as perceived risk, in determining their use.

The field is extremely multifaceted as has been illustrated here. The number of drugs is large and growing, and they have a wide range of psychological and physical effects. There are many levels of social and psychological determinants of this group of behaviors, ranging from legal, moral, normative, and religious, to neurological, psychological, sociological, political, and economic. (Because drugs are commodities, market considerations also influence use.) Further, drug use is the object of many social interventions, from various forms of education and persuasion to treatment. Thus, despite the considerable resources currently going into research on the determinants of drug use, it is clear that there will remain a great deal to be learned about this unique domain of behavior for some time to come.

See also: Addictions: General Considerations; Alcohol Use among Young People; Health Interventions, Community-based; Smoking and Health.

Bibliography

Bachman, J.G., Johnston, L.D., O'Malley, P.M., 1998. Explaining the recent increases in students' marijuana use: the impacts of perceived risks and disapproval from 1976 through 1996. American Journal of Public Health 88, 887–892.

Brook, J., Brook, D., Gordon, A., Whiteman, M., Cohen, P., 1990. The psychosocial etiology of adolescent drug use: a family interactional approach. Genetic, Social, and General Psychology Monographs 11, 111–267.

Caulkins, J.P., Hawken, A., Kilmer, B., Kleiman, M.A.R., 2012. Marijuana Legalization: What Everyone Needs to Know. Oxford University Press, New York.

Clayton, R.R., 1992. Transitions in drug use: risk and protective factors. In: Glantz, M., Pickens, R. (Eds.), Vulnerability to Drug Abuse. American Psychological Association, Washington, DC, pp. 15–52.

Clayton, R.R., Leukefeld, C.G., Donohew, L., Bardo, M., 1995. Risk and protective factors: a brief review. Drugs & Society 8, 7–14.

DiClemente, C., Holmgren, M., Rounsaville, D., 2011. Relapse prevention and recycling in addiction. In: Johnson, B.A. (Ed.), Addiction Medicine. Springer, New York, pp. 765–782.

Earleywine, M., Newcomb, M.D., 1997. Concurrent versus simultaneous polydrug use: percentage, correlates, discriminant validity and prospective effects. Experimental and Clinical Psychopharmacology 5, 353–364.

Everitt, B., Robbins, T., 2005. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nature Neuroscience 8 (11), 1481–1490.

Feigelman, W., Gorman, B.S., Lee, J.A., 1998. Binge drinkers, illicit drug users and polydrug users: an epidemiological study of American Collegians. Journal of Alcohol and Drug Education 44, 47–69.

Forsyth, A.J.M., 1996. Places and patterns of drug use in the Scottish dance scene. Addiction 91, 511–521.

Glantz, M., Pickens, R. (Eds.), 1992. Vulnerability to Drug Abuse. American Psychological Association, Washington, DC.

Grant, B.F., Stinson, F.S., Dawson, D.A., Chou, S.P., Dufour, M.C., Compton, W., Pickering, R.P., Kaplan, K., 2004. Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders: results from the national epidemiologic survey on alcohol and related conditions. Archives of General Psychiatry 61 (8), 807–816. http://dx.doi.org/10.1001/archpsyc.61.8.807.

Inciardi, J.A., McBride, D.C., 1991. The case against legalization. In: Inciardi, J.A. (Ed.), The Drug Legalization Debate. Sage, Newbury Park, CA, pp. 45–79.

Isaac, M., Janca, A., Sartorius, N., 1994. ICD-10 Symptom Glossary for Mental Disorders. Division of Mental Health, World Health Organization, Geneva.

Janca, A., Ustun, T.B., van Drimmelen, J., Dittmann, V., Isaac, M., 1994. ICD-10 Symptom Checklist for Mental Disorders, Version 1.1. Division of Mental Health, World Health Organization, Geneva.

John, D., Kwiatkowski, C.F., Booth, R.E., 2001. Differences among out-of-treatment drug injectors who use stimulants only, opiates only or both: implications for treatment entry. Drug and Alcohol Dependence 64 (2), 165–172.

Johnson, R.A., Gerstein, D.R., 2000. Age, period, and cohort effects in marijuana and alcohol incidence: United States females and males, 1961–1990. Substance Use & Misuse 35, 925–948.

Johnston, L.D., O'Malley, P.M., 1986. Why do the nation's students use drugs and alcohol?: Self-reported reasons from nine national surveys. Journal of Drug Issues 16 (1), 29–66.

Johnston,, L.D., O'Malley, P.M., Bachman, J.G., 2000. Monitoring the Future National Survey Results on Drug Use, 1975–1999. Secondary School Students, vol. I and College Students and Adults Aged 19–40, vol. II. National Institute on Drug Abuse, Bethesda, MD.

Kandel, D.B., 1975. Stages in adolescent involvement in drug use. Science 190, 912–914.

Kandel, D.B., Yamaguchi, K., Chen, K., 1992. Stages of progression in drug involvement from adolescence to adulthood: further evidence for the gateway theory. Journal of Studies on Alcohol 53, 447–457.

Kenna, G.A., et al., 2007. Pharmacotherapy of dual substance abuse and dependence. CNS Drugs 21 (3), 213–237.

Kessler, R.C., Berglund, P., Chiu, W.T., Demler, O., Heeringa, S., Hiripi, E., Jin, R., Pennell, B.E., Walters, E.E., Zaslavsky, A., Zheng, H., 2004. The US national comorbidity survey replication (NCS-R): design and field procedures. International Journal of Methods in Psychiatric Research 13 (2), 69–92. http://dx.doi.org/10.1001/archgenpsychiatry.2007.18.

Kirby, T., Barry, A.E., 2012. Alcohol as a gateway drug: a study of US 12th graders. Journal of School Health 82 (8), 371–379.

Koob, G.F., Nestler, E., 1997. The neurobiology of drug addiction. Journal of Neuropsychiatry and Clinical Neurosciences 9, 482–497.

Leshner, A.I., 1999. Science-based views of drug addiction and its treatment. Journal of the American Medical Association 282, 1314–1316.

McCaffrey, D.F., Liccardo Pacula, R., Han, B., Ellickson, P., 2010. Marijuana use and high school dropout: the influence of unobservables. Health Economics 19 (11), 1281–1299.

Merikangas, K.R., McClair, V.L., 2012. Epidemiology of substance use disorders. Human Genetics 131 (6), 779–789.

Riley, S.C.E., James, C., Gregory, D., Dingle, D., Cadger, H.M., 2001. Patterns of recreational drug use at dance events in Edinburgh, Scotland. Addiction 96, 1035–1047.

Roffman, R.,A., 2013. Legalisation of marijuana: unraveling quandaries for the addicted professionals. Frontiers in Psychiatry 31 (4), 50.

Redonnet, B., Chollet, A., Fombonne, E., Bowes, L., Melchior, M., 2012. Tobacco, alcohol, cannabis and other illegal drug use among young adults: the socioeconomic context. Drug and Alcohol Dependence 121 (3), 231–239.

SAMHSA, 2011. Results from the 2010 National Survey on Drug Use and Health: Summary of National Findings, Vol NSDUH, Series H-41, HHS Publication No. (SMA) 11–4658. Substance Abuse and Mental Health Services Administration, Rockville.

United Nations, 1971. Convention on Psychotropic Substances 1971. United Nations Publication. Sales No. E.78.XI.3.

United Nations, 1972. Protocol Amending the Single Convention on Narcotic Drugs, 1961. United Nations Publication. Sales No. E.77.XI.3.

United Nations, 1988. United Nations Convention against Illicit Traffic in Narcotic Drugs and Psychotropic Substance, 1988. United Nations Publication. Sales No. E.98.XI.5.

UNODC, 2013. World Drug Report 2013. United Nations Publication. Sales No. E.13.XI.6.

Wagenaar, A.C., Salois, M.J., Komro, K.A., 2009. Effects of beverage alcohol price and tax levels on drinking: a meta-analysis of 1003 estimates from 112 studies. Addiction (Abingdon, England) 104 (2), 179–190.

Wilson, L.M., et al., 2012. Impact of tobacco control interventions on smoking initiation, cessation, and prevalence: a systematic review. Journal of Environmental and Public Health 2012. Available at: http://www.hindawi.com/journals/jeph/2012/961724/abs/ (accessed 30.08.13.).

World Health Organisation, 1994. Lexicon of Psychiatric and Mental Health Terms, second ed. World Health Organisation, Geneva.

Du Bois, William Edward Burghardt (1868–1963)

Paul Gilroy, Yale University, New Haven, CT, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 3885–3555, © 2001, Elsevier Ltd.

Abstract

This essay presents a brief survey of W. E. B. Du Bois' (1868–1963) long life and prolific scholarship. It includes details of his education, travel, and changing political opinions. A selection of his many publications is discussed and there are some speculations about the intellectual influences from both Europe and North America that guided his thinking. His career inside and outside the academy is considered along with his continuing impact on scholarly writing in the social sciences and the humanities.

William Edward Burghardt Du Bois was born in Great Barrington, Massachusetts on February 23, 1868. Ninety-five years later, having renounced his citizenship of the USA, he died an exile in Accra, the capital of the newly decolonized republic of Ghana on April 29, 1963. An unorthodox and multidimensional career had unfolded between these two events. An unwavering devotion to the cause of human progress underpinned Du Bois' cosmopolitan life that was also characterized by bold shifts in ideological outlook. He moved through pragmatism, elitism, socialism, nationalism, and several versions of Pan Africanism in an idiosyncratic, but always thoughtful sequence that terminated amidst the gloom of the Cold War when the old savant, aged 93 and deprived by his government of the means to travel freely, applied to join the Communist Party of the USA and then fled the land of his birth.

Du Bois had been identified as an outstanding student at an early age. After completing high school in Massachusetts, he continued his studies initially at Fisk University in Nashville. When he entered the college in 1885, it was a haven lodged in a bitterly segregated world still resonating with the destructive energies released by civil war. Life there, confined within what he called the veil of color, was a very different proposition from anything that Du Bois had experienced in the New England environment that had nurtured him. It was there that he discovered and wholeheartedly embraced the black cultures of the American South that he encountered inside the historic college, and discovered in more intense and ambiguous forms while working as a school teacher in rural Tennessee during the summer vacations. He concluded his studies at Fisk in 1888, choosing the German Reich's Chancellor Otto Von Bismarck as the topical subject of his oration at the college graduation ceremony.

In fulfillment of a childhood dream, and in recognition of the opportunities it would bring, Du Bois moved on to continue his education at Harvard later the same year. Among other classes, he studied philosophy with William James, who would become a friend, and read Kant with George Santayana. He took economics under Frank Taussig and history with Albert Bushnell Hart who helped to pilot his career through numerous institutional obstacles. After graduating *cum laude* in Philosophy in 1890, he applied to pursue postgraduate study in Political Science and gained an MA degree in 1891. It was at this point that Du Bois began to seek ways to continue his education in Europe. Financial support was difficult to find but in 1892 he was eventually rewarded with a year-long grant to study in Germany. Later that year, he entered Friedrich-Wilhelm III University in Berlin where his teachers included Gustav Von Schmoller, Adolf Wagner, and Heinrich Von Treitschke. It was also there that he first encountered Max Weber and Georg Simmel.

Du Bois attended several meetings of the Social Democratic Party (SPD) while living in Berlin. He had been working towards a German doctorate in economics but, to his disappointment, was denied the degree on the grounds that he had been registered at the university for an insufficient length of time. After extensive travel in Europe he returned to the corrupting provincialism of 'nigger-hating America' to complete his thesis in Harvard's History department in 1895. Shortly afterwards, he became the first black person to receive the Ph.D. degree from that University, but as a result of this success was hugely overqualified for the limited professional opportunities available to him in a world that could not concede the intellectual capacity of 'the Negro' and consequently viewed him as something of a curiosity.

He eventually found a position at Wilberforce College in Ohio where he taught Classics, Greek, and English. His growing interest in Sociology was manifest in some unsuccessful attempts to have the new discipline added to the curriculum. He married Nina Gomer, who had been a Wilberforce student, in 1896 and, later that year, his first book *The Suppression of the African Slave Trade to the United States of America 1638–1870* (1896) was published to very favorable reviews. This welcome development boosted Du Bois' career prospects and enabled him to move to the University of Pennsylvania, where he was employed to produce a groundbreaking sociological study of Philadelphia's black population. Work on that project would continue after he accepted a Professorship of Economics and History at Atlanta University the following year.

This new post provided an institutional platform from which Du Bois organized an ambitious program of sociological investigations into various aspects of black life in the USA. It resulted in a flow of major publications between 1898 and 1914. His substantial empirical study of Philadelphia was published under the title *The Philadelphia Negro* in 1899; it was given a distinctive character by well-developed interests in economics and comparative history. It is rightly remembered as a landmark publication: the first academic monograph to

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61025-5

consider the life of an American black community in depth. The Du Bois household was struck by calamity when Burghardt, their two year-old son, died of diphtheria in May 1899. This tragedy seems to have led to estrangement between Du Bois and his wife. However, it combined with his growing academic and political ambitions to endow an elegiac and sometimes melancholy tone in his next book: *The Souls of Black Folk* published in 1903. This was a remarkable, multivoiced collection of his essays, autobiographical reflections, and fiction. It made him a leader of black Americans.

Du Bois' commitment to resolutely open-ended, liberal, and humanist education which would nourish the needs and responsibilities of a black elite destined for leadership at both local and global levels, brought him into conflict with the more modest and practically oriented educational proposals that were being advocated by Booker T. Washington and were on show at his Tuskeegee Institute. This national debate, which exposed the molten core of black America's political and economic marginality, resulted in Du Bois leaving Atlanta University to pursue a new para-academic career as a writer, editor, and activist based in the North. In 1910 the newly formed National Association for the Advancement of Colored People (NAACP) hired him as Director of Publications and Research. He was the founding editor of, and a prolific contributor to, the organization's influential and dynamic monthly magazine *The Crisis*, a post he would hold until 1934. His first novel, *The Silver Fleece* was published in 1911, the same year that he attended the Universal Races Congress organized in London by the English Ethical Culture movement.

The death of Booker T. Washington and the advent of the 1914–18 war transformed the political conditions in which Du Bois argued for the equality of black Americans and indicted the catastrophic impact of racism on the American polity stratified by color. Though he never completed a proposed volume on the experience of African-Americans during the war, the immediate aftermath of the conflict provided the impetus for the first of several Pan African congresses, which was held in Paris in 1919. After the war, Du Bois' voice was raised against America's failure to recognize its black citizens and in opposition to the gross violence that greeted the returning soldiers. He was severely critical of the more populist and militaristic forms of Black Nationalism associated with Marcus Garvey's United Negro Improvement Association (UNIA) which was growing in importance as the influence of the NAACP diminished.

Du Bois' cosmopolitan outlook and the breadth of his published work made him a pivotal figure among the younger writers, artists, and thinkers who led the artistic and literary movement known as the Harlem Renaissance. He contributed a memorable essay to *The New Negro* anthology, edited by the philosopher Alain Locke, that defined this important cultural departure but conflict soon followed with some of the younger writers, who were regarded by Du Bois as insufficiently attentive to their world-historic political responsibilities. *Dark Princess*, a vibrant political novel that concerns a grand planetary alliance of peoples of color against imperialism and colonialism and which was identified by Du Bois as his own favorite among his many books, was published in 1928.

The thirties saw Du Bois renewing his interest in Marxism and developing an original and complex critique of Marx's political theory. This striking synthesis was elaborated in a monumental historical study: *Black Reconstruction In America* (1935). The first of his several autobiographical volumes appeared in 1940. Du Bois rejoined the NAACP as their director of special research in 1944. He was sacked four years later after yet another bout of political disagreements, this time resulting from his intermittent affiliation to communist causes that the utterly respectable and sometimes conservative organization was unable to tolerate. Anti-colonialist and anti-capitalist commitments led Du Bois into further battles with the US government. In 1951 he was arrested and charged with the crime of operating as an unregistered agent of a foreign power. The charges were eventually dismissed but this unhappy experience galvanized him in support for the Soviet bloc. This phase of his life was explored in yet another book *In Battle for Peace: The Story of My Eighty-Third Birthday* (1952).

Du Bois' Scientific Contribution

Du Bois was the first sociologist of the American South. He articulated the complexities of African-American life in comparaive and historical terms that owed much to his multidisciplinary training. He placed sociological reason at the disposal of black Americans' freedom struggles and sowed the seeds that would bear fruit in the mid-twentieth century legal judgements that cited later sociological studies in support of desegregation. His own timely empirical contributions, particularly *The Philadelphia Negro* (1899) were also important because they replaced top-heavy theoretical speculation with an honest blend of theory with practical innovations in research technique. Guided as much by Schmoller as by the spirit of Charles Booth, Du Bois was also the first American social scientist to make empirical research the pivot of an investigation. His investment in the power of properly scientific sociology as a stimulus for governmental intervention and the pursuit of democracy unmarked by racial categories was considerable; however his faith in progress was neither simplistic nor naïve. Sociology was obviously indispensable but it was also insufficient. Its hard-won truths had to be communicated in several tongues, and made accessible to the widest possible audience. But more than this, Du Bois' achievements suggested that sociology was always at its most eloquent when supplemented and enhanced by the power of poetry.

The Contemporary Influence

In addition to his accomplishments as a distinguished social scientist Du Bois is still remembered as an essayist, editor, novelist, historian, and political leader. The breadth of his intellectual interests and the skill with which he turned them into a host of compelling written material in several distinct disciplines and styles made him the dominant African-American thinker of the twentieth century. Close to the aftermath of slavery and against the expectations of those who had judged 'Negroes' to be incapable of higher learning, his extraordinary work boldly redefined America's racial hierarchy as intrinsic to the incomplete workings of American democracy. In making that unpalatable argument in potent language which was both scholarly and poetic, Du Bois created a black

intelligentsia which understood itself and its historical and sociological predicament in terms he had specified. His exceptional scholarly insights were articulated in a prolific output that continued throughout his long life. All of this work was characterized by a double achievement to match the double consciousness from which it grew – a sense of black intellectuals caught, or as he put it 'imprisoned,' within the folds of European civilization developing interests of their own distinct from those of the black poor.

He transformedunderstanding of African-American history and society while supplying the intellectual cornerstones for new varieties of critical and oppositional consciousness that could move beyond the task of defending America's black communities as they emerged from the twilight of slavery. In articulating an especially rigorous, energetic, and durable brand of dissident intellectual work, Du Bois not only provided solid foundations for the scholarly study of racism, he enabled his many successors to challenge white supremacy in the world ofideas, and provided important conceptual and historical resources for anticolonial movements worldwide.

See also: Race Identity; Race and Racism in the Twenty-First Century; Racism, Sociology of; Sociology, History of.

Bibliography

Broderick, F., 1959. W. E. B. Du Bois: Negro leader. In: A Time of Crisis. Stanford University Press, Stanford, CA.

Du Bois, W.E.B., 1896. The Suppression of The African Slave Trade To The United States of America 1638–1870. Harvard Historical Studies, no. 1. Longmans Green, New York.

Du Bois, W.E.B., 1899. The Philadelphia Negro: A Social Study. University of Philadelphia Press, Philadelphia.

Du Bois, W.E.B., 1903. The Souls of Black Folk: Essays and Sketches. A. C. McClurg, Chicago.

Du Bois, W.E.B., 1911. The Quest of The Silver Fleece. A. C. McClurg, Chicago.

Du Bois, W.E.B., 1928. Dark Princess: A Romance. KrausThomson Organization, Millwood, New York.

Du Bois, W.E.B., 1935. Black Reconstruction. In: America: An Essay Toward a History of the Part Which Black Folk Played in the Attempt to Reconstruct Democracy in America 1860–1880. Atheneum Press, New York.

Du Bois, W.E.B., 1952. In: Battle for Peace: The Story of My Eighty-Third Birthday. Masses & Mainstream, New York.

Levering, L.D., 1993. W. E. B. Du Bois: Biography of A Race. Henry Holt, New York.

Locke, A., 1925. In: The New Negro. A. and C. Boni, New York.

Rampersad, A., 1976. The Art And Imagination of W. E. B. Du Bois. Harvard University Press, Cambridge, MA.

Dubos, René Jules (1901–82)

Jill Cooper, Rutgers University, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 3892–3894, © 2001, Elsevier Ltd.

Abstract

René Jules Dubos enjoyed national acclaim as a medical researcher, author, and environmentalist during the middle third of the twentieth century. Trained first as an agricultural engineer at the French National School of Agronomy, Dubos moved to the USA in 1924 to study soil microbiology at Rutgers University and the New Jersey Agricultural Experiment Station. After completing his doctorate in 1927, he joined the medical research staff at the Rockefeller Institute for Medical Research. Dubos developed the first clinically useful antibiotic from a soil microbe in 1939. This discovery ushered in the era of antibiotic wonder drugs. Although Dubos continued doing medical research until his retirement from Rockefeller University in 1971, he was better known after the 1950s for his promotion of an ecological model of disease and as an elder statesman of the US environmental movement.

René Jules Dubos, the French-American microbiologist, medical researcher, environmentalist, and Pulitzer-prize-winning author, was born in Saint-Brice-sous-Fôret, France on February 20, 1901. He was the eldest son of Georgesand Adeline DeBloëdt Dubos, proprietors of a series of small butcher shops. Two years after his birth, the Dubos family moved to neighboring Hénonville, a small agricultural village in the Île de France region 50 miles north of Paris. René Dubos spent the next dozen years of his life there exploring the French countryside, reading avidly, and excelling as a student at *l'ecole communale d'Hénonville*, the local one-room primary school. Encouraged by his mother to pursue a scholarly career, he dreamed of attending France's distinguished schools of higher learning. The family moved to Paris in 1914. Dubos won a fellowship in 1915 to attend the *Collège Chaptal*, part of the French secondary school system. His father's untimely death at the conclusion of World War I, his family's concomitant financial difficulties, and his own periodic bouts with poor health resulting from an early battle with rheumatic fever, however, prevented René Dubos from studying to become a historian, his first choice among academic careers. Instead, he gained admission in 1919 to the *Institut National Agronomique* (INA), France's National School of Agronomy.

Although many graduates of the INA pursued careers in scientific agricultural research, Dubos disliked laboratory work. Instead, he accepted a fellowship from the French government of Indochina to study at the *Institut National d'Agronomie Coloniale*, where he prepared for a position as a colonial administrator. However, the colonial government revoked its offer of employment at the conclusion of the training program when Dubos failed a physical examination. Having few alternatives, Dubos accepted an editorial position at *l'Institut International d'Agriculture* in Rome, where he abstracted technical articles for publication in the *Journal of International Agricultural Intelligence.* While in Rome, Dubos experienced a change of heart regarding his career plans. Inspired by an argument made in an article by the Russian agricultural bacteriologist, Sergei Winogradsky, Dubos reconsidered his decision to steer clear of laboratory research. In this article, Winogradsky encouraged his colleagues to reject pure culture techniques in favor of studies that more accurately monitored soil microbes in their natural environments. This more ecological approach to microbial investigation appealed to Dubos because it emphasized themes familiar to him from the training he had received as an agricultural engineer in Paris. Dubos gained the opportunity to study soil microbes himself when, at the 1924 International Congress of Soil Science, Jacob Lipman of the New Jersey Agricultural Experiment Station invited the young Frenchman to study soil microbiology at his institution. Dubos accepted the offer, sailed to the USA, and spent the next three years in a doctoral program under the direction of Selman Waksman at Rutgers University and the New Jersey Agricultural Experiment Station.

Soon after completing his Ph.D. in soil microbiology, Dubos received an unusual and compelling offer to join the research team of Oswald Avery, an immunochemist at the Rockefeller Institute for Medical Research. Few agricultural researchers received an opportunity to apply their skills to medical research in the 1920s, particularly at the nation's most prestigious medical research institution, but Dubos's doctoral research on cellulose decomposition filled a specific need in the Avery laboratory. For years, Avery's laboratory group had tried unsuccessfully to strip the polysaccharide coating from the type III pneumococcus responsible for lobar pneumonia. Avery theorized that the coating protected the bacillus from the human body's natural defenses. Dubos suggested that he could find a soil microbe with the ability to produce an enzyme capable of dissolving the polysaccharide coating. Avery gave the young agricultural scientist a chance.

Thus, in 1927, René Dubos began a nearly fifty-year association with the Rockefeller Institute for Medical Research (later Rockefeller University). Only a two-year stint as George Fabyan Professor of Comparative Pathology and Tropical Medicine at Harvard Medical School, 1942 to 1944, interrupted his career-long affiliation with the Rockefeller Institute. As a medical researcher, Dubos made numerous contributions to bacteriology and medical science. In the Avery laboratory group, he extracted an enzyme from a soil microbe that proved effective against the polysaccharide coating of the lobar pneumonia bacillus. Although the advent of sulfa drugs precluded the use of the Dubos enzyme in the clinical treatment of pneumonia, the process he used to extract the enzyme proved beneficial in the late 1930s, when he returned to the soil to find a microbe with broader bactericidal properties. In 1939 he earned international recognition by developing gramicidin, the first clinically useful antibiotic. His success suggested the

clinical potential of soil microorganisms to researchers around the world. News of Dubos's advances spurred Oxford University's Howard Florey and Ernst Chain to resurrect Alexander Fleming's long-dormant research on penicillin. Dubos's work also inspired the effective development of streptomycin by his graduate school advisor, Selman Waksman. Collectively, these advances marked the beginning of the antibiotic era. Dubos's leading role in antibiotics research earned him the status of full member at the Rockefeller Institute, and, in 1941, election to the National Academy of Sciences.

Although his advances in antibiotics research represented his most celebrated contributions to medical science, they were by no means his only accomplishments. In the late 1940s Dubos gained recognition as an outstanding tuberculosis researcher. With the assistance of Bernard Davis, he developed an effective culture medium for tuberculosis bacilli. This new Tween-albumin medium permitted researchers to study the difference between virulent and nonvirulent bacilli and was credited with catalyzing a 'renaissance in tuberculosis research.' A strong advocate for vaccination against tuberculosis, Dubos also established guidelines for the standardization of BCG vaccines worldwide. Finally, in the 1950s Dubos collaborated with physicians in the Rockefeller Hospital on a study which revealed that prolonged bed rest did not expedite the recovery of tuberculosis patients receiving antibiotic therapy. This discovery had profound implications for sanatoriums, the traditional sites for tuberculosis treatment.

Despite his pathbreaking contributions to the modern pharmacopoeia, Dubos is more often remembered for his provocative ideas on human health. No one contributed more to the hubris of medical science at mid-century than Dubos with his breakthroughs in antibiotics research, but shortly after unveiling gramicidin, he publicized his concerns about the adequacy of antibiotic therapy for handling infectious disease. As early as 1942 he warned that the unmitigated enthusiasm for antibiotic wonder drugs should be tempered by a concern that overprescription of the drugs could lead to the development of more virulent, resistant bacterial strains. He argued that scientists gained an incomplete understanding of infectious disease when they focused solely on infectious agents. His studies of tuberculosis and the loss of his first wife to the disease taught him that, particularly in the case of latent infection, the host's ability to resist a disease is at least as important as the presence of an infectious agent. Pathologic processes, in his opinion, depended upon the total environment of the host, and deserved the serious attention of scientists. This challenge to the specific etiology of disease contributed significantly to the formulation of current conceptions which now include social factors, mental and physical disturbances, and environmental conditions in the understanding and management of disease.

Dubos's reconceptualization of disease, environment, and human health reached a broad audience of medical researchers, social scientists, politicians, and laypeople. He developed his ideas through over twenty books, countless articles in scholarly and popular journals, and an ever-increasing number of public addresses between the 1940s and the late 1970s. His most influential writings on human health include *The White Plague: Tuberculosis, Man, and Society* (1952), *Mirage of Health: Utopias, Progress, and Biological Change* (1987), and his Pulitzer-prize-winning text, *So Human an Animal: How We are Shaped by Surroundings and Events* (1998). In *The White Plague*, a book Dubos co-wrote with his second wife, Jean Porter Dubos, they mined historical accounts of tuberculosis outbreaks to demonstrate that the incidence of tuberculosis increased during times of social strife and decreased during less tumultuous eras. By linking it to poverty, they argued that tuberculosis was as much a social as well as a microbiological disease. In *Mirage of Health*, Dubos again challenged his readers to recognize disease as the indirect outcome of a constellation of circumstances rather than the direct result of a virulent microorganism.

Microbial disease, when it occurred, was due to an imbalance in the host's equilibrium caused by any of a number of factors in the external or internal environment – be it weather conditions, availability of food, working habits, emotional status, or emotional stress. Therefore, he argued, by ameliorating social and environmental conditions, social reformers had been more effective than scientists in reducing mankind's susceptibility to disease. Disease, according to Dubos, could not be eradicated simply by medical treatment. Social conditions must also be managed. His message enjoyed broad appeal among social scientists, many of whom, like Dubos, refocused their scholarly inquiries toward research that revealed how changes in an organism's social environment affected its health. Scientists also utilized Dubos's biopsychosocial perspective on health to construct a theoretical justification for the study of psychosomatic medicine.

By the 1970s, Dubos had become one of the USA's most visible scientists. His message appealed to a large number of Americans interested in the increasingly apparent connection between the environment and human health. Although organizations invited Dubos to speak on the environmental aspects of health and disease as early as the 1950s, his influence among environmentalists grew substantially after the publication of his Pulitzer-prize-winning book, *So Human an Animal*, in 1998. The book appeared when the USA was in the midst of the largest grassroots environmental movement in its history. Humans, like soil microbes, argued Dubos, adapt to their changing environments. Therefore, the quality of the environment in which a human develops holds great significance in the determination of human health and well-being. Although some environmentalists took issue with Dubos' notion that humans could improve on nature, many embraced him as an elder statesman of their movement. They adopted several of his maxims, including 'Think globally, but act locally' and 'Trend is not destiny,' as rallying cries for their cause. Dubos remained active in environmental activities years after his 1971 retirement from Rockefeller University. He co-wrote *Only One Earth* with economist and political scientist, Barbara Ward. This document provided scientific and social guidelines for the First UN Conference on the Human Environment in Stockholm in 1972. In 1977, Dubos lent his name to the New York-based René Dubos Center for Human Environments, an organization charged with the planning of international forums to promote beneficial human interventions into nature.

Dubos traveled an uncommon path from his early days as a French agricultural engineer to his respected position within the Rockefeller University and twentieth-century biomedical research. Although recognized and applauded as part of the scientific establishment by the 1940s, Dubos never hesitated to challenge the methodologies of his colleagues. Because he

believed that all living organisms must be understood in the context of their respective environments, he advocated less reductionistic laboratory research, studies that traversed interdisciplinary boundaries, and an ecological understanding of disease. He followed his inner logic outside the confines of scientific safety, where he found diverse and interested audiences willing to engage his provocative message. René Dubos died on his 81st birthday in New York City.

See also: Biomedical Sciences, History and Sociology of; Environmental Health and Safety: Social Aspects; Environmental Stress and Health; Public Health as Social Science.

Bibliography

Benison, S., 1976. René Dubos and the capsular polysaccharide of pneumococcus. Bulletin of the History of Medicine 50, 459–477.

Cooper, J., 1998. Of microbes and men: a scientific biography of René Jules Dubos. Ph.D. thesis. Rutgers University.

Dubos, R., 1987. Mirage of Health: Utopias, Progress, and Biological Change, second ed. Rutgers University Press, New Brunswick.

Dubos, R., 1992. The White Plague: Tuberculosis, Man, and Society, second ed. Rutgers University Press, New Brunswick.

Dubos, R., 1998. So Human an Animal: How We are Shaped by Surroundings and Events, second ed. Transaction, New Brunswick, NJ.

Hirsch, J.G., Moberg, C., 1989. René Jules Dubos. Biographical Memoirs of the NAS 58, 132–161.

Moberg, C.L., 1996. René Dubos: a harbinger of microbial resistance to antibiotics. Microbial Drug Resistance 2, 287–297.

Moberg, C.L., Cohn, Z. (Eds.), 1990. Launching the Antibiotic Era: Personal Accounts of the Discovery and Use of the First Antibiotics. Rockefeller University Press, New York.

Piel, G., Segerberg, O., 1990. The World of René Dubos: A Collection from His Writings. Holt, New York.

Due Process in Law

Samuel Issacharoff, New York University School of Law, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article assesses the history of the concept of due process in Anglo-American jurisprudence. While the actual phrase 'due process of law' first appeared in the Magna Carta, the concept of due process was transformed when the American colonists adopted it as a limiting principle not only on the use of executive power but also on the scope of the laws that could be passed. Due process protects substantive rights by requiring procedural regularity in the exercise of governmental power and insists that procedural regularity may not be altered by the normal operations of political power. In the twentieth century, due process law in the United States developed in three central phases – substantive, procedural, and functional due process – and these trends continue into the twenty-first century.

The concept of due process emerges from the central paradox in the creation of state authority. Following Thomas Hobbes (1994), the governmental state may be seen as an evolutionary response to the failings of the state of nature. In the war of all against all, no person has an incentive to industry, writes Hobbes in the *Leviathan*, because each lacks security in investment, planning, and the ability to exchange. Each faces the threat of bodily violence, invasion of the home, and seizure of property. The response, per Hobbes, is the creation of the state, a body that can impose order, lift mankind from the pitiable war of all against all, and secure collective benefits. The paradox emerges in that the creation of the state also produces a magnified form of the threat that was ever present in the state of nature. The state represents the largest entity capable of inflicting bodily violence, invading the home, and seizing property. Indeed, all states reserve to themselves the right to do just that to their citizens, and most include the right even to punish by death. What then differentiates the modern state that claims political legitimacy from the state of nature? The answer must begin with an abiding respect for procedural fairness in the imposition of punishment and the regulation of the conduct of the citizens' lives Thomas Hobbes (1994).

This conception of procedural fairness is at the heart of the concept of the rule of law, and of its more technically exigent offshoot, the guarantee of due process. The typical formulation of due process, as that concept has developed in Anglo-American jurisprudence, is derived from the Magna Carta, issued by King John in 1215 in the face of rebellious barons seeking restraint upon the exercise of royal prerogatives:

> No free man shall be seized or imprisoned, or stripped of his rights or possessings, or outlawed or exiled, or deprived of his standing in any other way, nor will we proceed with force against him, or send others to do so, except by the lawful judgment of his equals or by the law of the land.

The crux of the original formulation was to limit the power of the royal sovereign by granting legal protection to the life, liberty, and property of English freemen. The significance of the Magna Carta's guarantees emerges from the form of the limitation on sovereign authority. The document does not attempt to deny the sovereign the capacity to imprison, outlaw, or banish the rights-holding citizenry, or to have restraints imposed on the enjoyment of property or other rights held by citizens. Rather it imposes a condition of procedural regularity in the exercise of the sovereign's power, including central, orderly trial processes and formal proclamation of the applicable legal standards. This notion of procedural regularity is captured in the actual phrase, 'due process of law', which first appeared when the Magna Carta was issued in an English version in 1354 by King Edward III. Its modern vitality draws from the efforts of Sir Edward Coke in the seventeenth century to revive the Magna Carta as a check on the power of the Stuart monarchs.

Not surprisingly given these origins, due process emerged from medieval Britain primarily as a restraint on the executive power of the sovereign, specifically of the Crown. The concept of due process was transformed when the American colonists adapted the Magna Carta as a limiting principle not only on the use of executive power but also on the scope of the laws that could be passed. For colonists lacking the rights of participation in the creation of laws, the insult of arbitrary state power occurred whether power was exercised legislatively or by the sovereign. As recast by the colonists, due process became a limitation that extended beyond the power of the crown, to encompass as well local assemblies and, by the time of the revolution, Parliament. Thus, even as England adopted a system of increasing parliamentary sovereignty, the concept of due process as a constraint on the exercise of governmental power retained its vitality in the colonies. As a result, the newly independent Americans established a more limited government, reflected by the Bill of Rights (including a due process clause), as an explicit limitation on the use of state power, including the legislative branch. As articulated by James Madison, "The legislature, no less than the executive, is under limitations of power," and hence the rights of the people were to be secured 'against legislative as well as executive ambition' (Madison, 1896). The source of the limitation was a constitutional command that stood apart and above the normal operations of state authority.

The American focus on legislative power also highlights the importance of due process as a restraint not just on the penal powers of the state but on its civil regulatory powers as well.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86025-0

Due process emerges from this historical origin with two central features. First, it protects substantive rights independently secured by positive law by requiring procedural regularity in the exercise of all governmental power. Second, the requirement of procedural regularity may not be altered by the normal operations of political power. As reflected in the American constitutional experience, the guarantees of due process are not directly accountable to the political dictates of either the executive branch of government (the closest in form to the sovereign crown) or the legislative will of the majority. As a constitutional guarantee, the requirements of due process may not be altered though the customary political processes of legislation. It is these two features of procedural regularity and independence from direct political accountability that define due process as it has emerged most clearly in American constitutional law, which long ago held due process to be a protection of 'those fundamental principles of liberty and justice which lie at the base of all our civil and political institutions' (*Hurtado v. California*, 110 US 516, 535, 1884). Because the development is most clear in American constitutional law, that will serve as the template for further elaboration.

There are five basic elements of due process that may be adduced from American case development, particularly after the post-Civil War amendments to the Constitution extended the command to the conduct of the states toward their citizens. The two given the greatest significance are that, "Parties whose rights are to be affected are entitled to be heard; and in order that they may enjoy that right they must first be notified" (*Baldwin v. Hale*, 68 US (1 Wall.) 233, 1863). The opportunity to be heard and its "instrumental corollary, a promise of prior notice" (Tribe, 1988), according to the Supreme Court, 'seem to be universally prescribed in all systems of law established by civilized countries' (*Twining v. New Jersey*, 211 US 78, 110–111, 1908). The additional elements of due process are further corollaries necessary to make the right to a hearing meaningful: the right to an impartial arbiter, the right to be represented by counsel, and the right to timely resolution of claims.

The articulation of these elements of due process leaves unanswered a central question dating from the time of the Magna Carta: Are these due process rights intended as affirmative entitlements of the citizenry, as indispensable restraints on the state, or as instrumental checks against arbitrary, capricious, or simply misguided state conduct? Put another way, the constitutional enshrinement of a principle of due process can be seen as creating a distinct set of rights among the citizenry, or as a constraint on governmental processes. Measured against this template, it is possible to characterize the development of twentieth century due process law as passing through three central phases.

Substantive Due Process

The most forceful introduction of due process into the American constitutional canon came in the closing years of the nineteenth century with the emergence of the substantive due process doctrine. Federal courts gave independent substantive meaning to the Fourteenth Amendment's prohibition on states' restricting the enjoyment of life, liberty, or property without due process of law. Under this view, state incursions upon the affirmative rights of citizens to be secure in their life, liberty, and property were presumptively invalid and subject to exacting judicial scrutiny. The key to this period was an expansive set of presumed individual claims against state regulation. The domain of legitimate state activity was restricted to ensuring the public health, guarding against inherently hazardous activities, and protecting those who could not fend for themselves at law. This last group included (at the time) children, the mentally incompetent, and women. The most controversial feature of substantive due process came with the expansive definition of *liberty* to include a broad range of market freedoms that in turn prompted federal courts to strike down state regulatory measures by the dozens. Most famous no doubt was the Court's 1905 ruling in *Lochner v. New York* (198 US 45, 1905), a case that struck down a maximum hours regulation for the employment of bakers on the substantive due process grounds of 'liberty of contract'. The hallmark of this period was the use of a high threshold of procedural presumptions against the legitimacy of state regulatory initiatives. The period of substantive due process lasted until its repudiation by the Supreme Court in the New Deal period of the 1930s. Nonetheless, the reaction to the perceived hostility of the Court to progressive social legislation earned the *Lochner* period notoriety as a constitutional epithet and removed substantive due process from the constitutional lexicon, until its tentative re-emergence at the end of the twentieth century in difficult cases such as the constitutional review of assisted suicide (*Washington v. Glucksberg*, 521 US 702, 1997). Most recently, the Supreme Court has used the liberty guarantees of the due process clause to ground the privacy rights underlying reproductive freedoms and the constitutional protection of homosexual activity from criminal prosecution (*Lawrence v. Texas*, 539 US 558, 2003).

Procedural Due Process

The second great phase of due process activism came with the Warren Court of the 1960s. Unlike the *Lochner* period's focus on the expansion of substantive claims that triggered due process, the hallmark of the second phase was a focus on 'procedural due process', which was understood as an expansion of the process-based requirements for the state to act adverse to the interests of affected citizens. In a series of decisions involving legislatively conferred benefits, ranging from driver's licenses to welfare benefits, the Court interposed an elaborate series of processes required before such benefits could be withdrawn. The right to a hearing was repeatedly invoked in this period as an affirmative right of participation in governmental processes that inhered in the rights of citizens. The Court did not seek to expand the ambit of positive rights – entitlements were created by the normal processes of government – but instead used the due process clause to enforce heightened procedural hurdles before government could deny a benefit. Instead of the liberty to contract of the early twentieth century, the aggrieved due process claimant of the postwar period would claim a 'right to be heard' prior to adverse governmental action. Critics charged that the elevated procedural steps created a presumption against adverse governmental action almost as certain as that of the substantive

due process period. But more problematic, as evidenced in the ensuing case law, was how to reconcile near-categorical requirements of an inherent participatory right to process with a range of governmental conduct that required swift response to emergency situations. The case law struggled to find a mediating principle that would allow emergency responses, as with police searches, the impounding of goods in flight, the embargo on goods threatening public health or safety, and a range of comparable conduct whose successful implementation could not await notice and an adversarial hearing.

Due Process Functionalism

The tension in the heightened proceduralism of the Warren Court prompted a scaling back of inherent rights in favor of an instrumental approach to due process. Whereas both the substantive due process and procedural due process period had read the due process command to vest the citizenry with affirmative rights against the state, the functional period beginning in the mid-1970s saw in due process a check against improper or incorrect governmental decisions affecting life, liberty, or property. Thus, in the watershed case of *Mathews v. Eldridge* (424 US 319, 1976), the Supreme Court analyzed a challenge to the termination of social security disability benefits not in terms of whether due process protections attached, but rather in terms of how much process was due. By contrast to cases of the Warren Court period that assumed that process would have to be the full panoply of evidentiary protections, *Mathews v. Eldridge* held that the amount of process required prior to adverse governmental conduct would be determined by a balance of interests heavily weighted toward considerations of cost and likelihood of error:

> [D]ue process generally requires consideration of three distinct factors: first, the private interest that will be affected by the official action; second, the risk of an erroneous deprivation of such interest through the procedures used; and the probable value, if any, of additional or substituted procedural safeguards; and finally, the government's interest, including the function involved and the fiscal and administrative burdens that the additional or substitute procedural requirement would entail. (424 US at 334–335)

In the decades since *Mathews v. Eldridge*, this form of due process functionalism has come to command this area of American constitutional law. All the critical elements of due process – from the right to an evidentiary hearing to the right to counsel – are now subject to the balancing of magnitude of loss versus likelihood of error versus governmental interest in swift action. Some of the applications appear inevitable, as with the ability of police to immobilize illegally parked cars, even in the absence of notice and a prior hearing. Some are more problematic, as with the denial of the right to have counsel provided to an indigent prisoner facing civil proceeding for termination of parental rights.

The case of *Hamdi v. Rumsfeld* (542 US 507, 2004) offers a recent example in the freighted context of threats of terrorism. The petitioner, Yaser Esam Hamdi, was an American-born individual taken prisoner in Afghanistan during the American military action in 2002 and held as an 'enemy combatant' in a military prison on United States soil. The government's only evidence supporting Hamdi's classification as an enemy combatant would have been inadmissible as hearsay under most circumstances. However, Hamdi was not allowed to challenge the evidence and was not allowed to speak to a lawyer or contact his family, who themselves had to file a petition of *habeas corpus* on his behalf. Invoking *Mathews'* functional approach as the controlling law even in national security cases, the Court conceded that although a hearing was required, it need not, due to the needs of the government during wartime, necessarily include every procedural protection available in, say, a criminal trial. The exact details of the hearing were left to be developed subsequently, but the Court's message was clear: due process remains a vital protection even if its application must be flexible.

Nearly 800 years after the Magna Carta, this functional view of due process has firmly taken hold in the most developed of due process legal regimes.

Due Process in Other National Legal Systems

Broadly speaking, a large number of the fundamental rights accorded to individuals under the American conception of due process are also found in other nations' laws, with one critical caveat. The conception of due process, particularly in civil law systems, primarily addresses rights of criminal procedure – for example, the French Declaration of the Rights of Man, which explicitly grants the right to be free from arbitrary arrest and detention. Various international agreements, such as the Convention for the Protection of Human Rights and Fundamental Freedoms, contain similar language. Many national constitutions similarly guarantee the rights to notice, choice of counsel, a speedy trial, and appeal.

Even in the sphere of criminal procedure, there are significant differences between the American notion of due process and that found in other countries. Perhaps the most striking example is the absence in many legal regimes of the right to be presumed innocent that is considered so fundamental in the United States. Also absent in many countries is a notion that both sides in a criminal trial have equal opportunity to advocate their positions.

More revealing is a comparison of the American and German conceptions of due process. On one level, the two countries' approaches are similar. Both put a pre-eminent value on human liberty. Thus, Article 1(1) of the German constitution states that "the dignity of man shall be inviolable." This notion of human dignity is akin to the American concept of substantive due process. Both are tools that have been utilized by the respective courts to 'discover' hitherto unenumerated rights. In this sense, both countries have a conception of due process that is, to a certain extent, open ended. As a consequence, both must grapple with an abstract and amorphous concept, often with controversial results.

But perhaps the most fundamental difference lies in their highest courts' approaches to constitutional interpretation. While the German Constitutional Court has interpreted its constitutional mandate to develop a systematic approach to constitutional jurisprudence, the United States Supreme Court, by emphasizing procedural regularity, has exhibited a tendency to

act in a piecemeal fashion. The result is that, in some ways, the German view of due process appears even more comprehensive than that of the United States. Most importantly, Germany guarantees the protection of 'personality rights' and 'physical integrity', which are akin to the American procedural due process. But, unlike the United States Supreme Court, which has often worked to construe due process narrowly, the German Constitutional Court has consistently taken a broad view of due process, viewing as properly furthering the fundamental principle of human dignity – a far more sweeping command than recognized in the initial procedural focus of Anglo-American law.

See also: Constitutional Courts; Constitutionalism, Comparative; Rights: Legal Aspects; Rule of Law (and Rechtsstaat).

Bibliography

Hobbes, T., 1994. Leviathan. J. M. Dent, London.

Madison, J., 1896. Report on the Virginia resolutions. In: Elliot, J. (Ed.), Debates in the Several State Conventions on the Adoption of the Federal Constitution, vol. 4. Lippincott, Philadelphia, PA.

Tribe, L., 1988. American Constitutional Law. Foundation Press, Mineola, New York.

Durkheim, Emile (1858–1917)

Steven Lukes, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Durkheim's vision of sociology went beyond the scientific to the moral and political, incorporating a distinctive diagnosis of modern society's ills. His several sociological classics remain central through successive critiques and reinterpretations. His major contributions to methodology, to the analysis of types of solidarity and suicide, and his sociology of religion and knowledge are reviewed; also his continuing influence and current significance are assessed. He has profoundly shaped social scientists' understanding of industrialization, urbanization, social control, social disorganization (with the concept of 'anomie'), collective behavior, the ritual and symbolic aspects of punishment, his thoroughly social account of religion, and of the social shaping of cognition and classification.

Émile Durkheim's writings, though they cover a vast range of topics, reveal a single, focused, organizing vision of sociology's subject matter and objectives. These went beyond the purely scientific, and were also moral and political. He sought, through education, to contribute to the stability and legitimacy of the French Third Republic and, more widely, to find cures for modern society's ills. His writings include several classics that remain central to the discipline of sociology, through teaching and through successive reinterpretations in which advances continue to occur by confrontation with Durkheim's ideas. These ideas, developed in close collaboration with a school of remarkably talented younger colleagues grouped around the journal, *L'Année sociologique*, which flourished before World War I, had a considerable impact within France and later far beyond, within sociology and across several other disciplines.

Biography

Born on 15 April 1858 in Épinal in the Alsace-Lorraine region into a close-knit rabbinic family, he abandoned Judaism at an early age. He thereafter treated religion agnostically as a social phenomenon demanding sociological explanation (from the mid-1890s, it was increasingly central to his thought), yet maintained that in studying it one is blind without a certain religious sentiment. After his studies at the *École normale supérieure* in Paris, he taught philosophy in secondary schools. On a government grant he visited Germany, for whose scientific education there was widespread admiration in defeated France. His encounter with German social scientists strengthened his so-called 'social realism.' In 1887, he was appointed to teach social science and pedagogy at the University of Bordeaux, as part of a national drive for a new system of republican, secular education. In 1896, he was appointed as the first Professor of Social Science in France and in 1897, founded *L'Année sociologique*. While at Bordeaux, he published three major works: *The Division of Labor in Society* (Durkheim, 1893[1902]), *The Rules of Sociological Method* (Durkheim, 1895[1901]), and *Suicide* (Durkheim, 1897) and his Latin doctoral thesis on Montesquieu. He also lectured and published across a wide range of topics including education, the family, the incest taboo, crime and social health, the evolution of punishment, the state, professional ethics and civic morals, religion, Rousseau, the history of sociology, and the history of socialism – the last in response to his students' political concerns. Averse to the sharpening of class conflict and believing revolutions to be as impossible as miracles, he stressed the need for institutional reforms that would encourage cross-class solidarity around common principles of social justice. He favored an interventionist State offset by a plurality of local, civic, and especially occupational associations, believing that out of this conflict of social forces individual liberties are born (Durkheim, 1950). He remained close to the liberal socialism of Jean Jaurès, but in the late 1890s, the Dreyfus Affair induced him to break his principle that the social scientist should be, as such, kept away from everyday politics. An active Dreyfusard, he denounced those who defended the army and the Church as sacred national institutions and proclaimed that modern society could only be integrated around the religion of individualism, protecting the sacredness of the individual and individual rights, the 'sole link' still binding citizens to one another.

In 1902, he moved to Paris to teach education and sociology. He continued to lecture and write across a wide range, publishing a study of *Primitive Classification* with Marcel Mauss (Durkheim and Mauss, 1903) and his masterpiece *The Elementary Forms of the Religious Life* (Durkheim, 1912). He debated with philosophers, historians, and psychologists and exercised a magisterial influence on the institutional development of sociology and on educational thinking and practice in France. On all these fronts, his doctrines generated controversy and opposition, often fierce, to which he responded vigorously. Much of his energy went into the editing of *L'Année*, of which 12 volumes had appeared by 1913. These contained monographs and reviews of works in many disciplines and languages, systematically classified and mined for their theoretical and empirical usefulness to a progressing research program. An extraordinary collaborative achievement, they constitute the Durkheimians' ongoing debate with the sociologically relevant literature of their time, distinctive, among other things, in its comprehensive attention to the ethnographic record of the world's remoter peoples. All of this ended with the onset of the World War on whose battlefields several, including the most brilliant and promising of them, including Durkheim's son Andre, met their deaths. For a second time,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61027-9

Durkheim broke his principle of political noninvolvement, contributing to the war effort and writing patriotic pamphlets, but, unable to overcome his grief at his son's death, he died in Paris on 15 November 1917.

Major Contributions

The focus of Durkheim's sociological vision was his so-called 'social realism': the view that social phenomena, or as he called them *faits sociaux*, are *sui generis*, with their own characteristics: they are general among, external to, independent of, and constraining upon particular individuals and typically are recognizable by their obligatory force. They constitute a distinct level of reality and are not reducible to the meeting of individual intentions. Hence, his attacks on social contract theorizing and on other forms of methodological individualism that attempted just such a reduction, as practiced for instance by utilitarians, economists, psychologists, and the English sociologist Herbert Spencer. At different stages, he focused on *faits sociaux* of different kinds: on 'morphological' phenomena such as population volume and 'dynamic density' and spatial distribution, on types of law (see Cotterrell, 1999), on institutions and, increasingly, on what he came to call *représentations collectives* (Durkheim, 1924). By this he meant collective ideas and sentiments, which could, moreover, be evanescent, like crowd enthusiasms or fashions, or more 'crystallized' or institutionalized, like legal codes or most stable of all, the fundamental categories of thought. *Faits sociaux* were what we would now call emergent phenomena whose properties, resulting from the structured interactions of their constituents, could not be accounted for exhaustively in terms of individuals' presocial properties.

Durkheim advanced this view polemically, as a doctrine, above all in *The Rules*. It has been suggested that this is best understood in its context as the forging of a vocabulary suited to securing French citizens' allegiance to their Republic (Jones, 1999). What is certain is that Durkheim's social realist rhetoric generated unproductive controversies and, worse, led to misunderstandings that have, even until today, deflected attention from his explanatory practice, to which his methodological pronouncements are an inadequate guide. Thus, he claimed that only social causes can explain social effects and that every time a social fact is explained directly by a psychological fact one could be sure the explanation is false. In various places he uses language that personifies collectivities, making 'society' appear as a subject of experience, with a *conscience commune*, or as an agent with a unified will, requiring conformity, pursuing ends, and imposing ways of acting on its members.

But against such passages one can cite others that stress that society only exists in and through individuals, that neither can exist without the other, and that representations collectives are the product of individuals' activity and, indeed, of an immense cooperation extending across space and time. In fact, Durkheim's actual explanations are by no means confined to the macrolevel; he sometimes seeks to explain the micro by the macro, as in *Suicide*, and the generation of macrophenomena at the microlevel, as in the generation and regeneration of religious faith through collective effervescences and rituals in *The Elementary Forms*. In short, Durkheim's explanatory practice was interactionist, not holistic. Unfortunately, his polemical antimethodological individualism and hostility to psychological explanation precluded him from reflecting on the microunderpinnings of his macroexplanations and, in general, on micro–macro connections, so that these aspects of his explanations remain undertheorized. Yet this very inattention may have generated insights. By not seeing certain things, perhaps he focused more intensely on what he did see. *Faits sociaux*, he wrote in *The Rules*, should be treated 'as things' ('*comme des choses*'), independent of and external to observers as well as actors, and enabling one to decide between competing theories. This expresses his commitment to scientific method as he saw it, as against philosophical theorizing uncontrolled by empirical evidence (to philosophize, he once told a doctoral candidate, is to think what one wants) and as against the Positivism of Comte, committed to the identification of a small set of transhistorical laws, rather than the systematic study of specific causal mechanisms (he rejected the label 'positivist,' preferring that of 'rationalist'). Social scientic method required the abandonment of pretheoretical preconceptions, the systematic testing of theories by evidence, quantitative where appropriate (as amply illustrated in *Suicide*), the careful selection of indicators, the rejection of teleological and functional explanations (though not the identification of social functions), the use of comparison (sociology's substitute for experiment), and the construction of social typologies. He conceived of types of society within a broadly evolutionary framework, proceeding from the least to the most advanced societies (and citing recent and contemporary tribal societies and those of Antiquity as instances of the former). He further thought that one could distinguish between 'normal' and 'pathological' developments within each type, at each stage of its development. This enabled him, he thought, to diagnose certain aspects of modern industrial societies, notably the prevalence of *anomie*, as pathological. The problem here was that his notion of normality was ambiguous between a statistical definition and one that embodied an implicit theory of social and psychological health or equilibrium relative to a given society's 'conditions of existence.' (How could a pathological phenomenon be both prevalent and abnormal?) Such diagnosis and the elaboration of feasible institutional remedies were, he thought, the chief ways in which the sociologist could contribute to social progress.

The *Division of Labor* contains (1) the positing of two ideal types of society, which Durkheim took to characterize less and more advanced societies respectively (but which can evidently coexist); (2) an analysis of the mechanisms claimed to explain the transition from the former to the latter; and (3) a diagnosis of the latter's pathological forms. Mechanical solidarity, held to typify preindustrial societies, based on resemblances, exhibits identical and juxtaposed elements, segmentally structured and with weak social bonds, low population volume and moral density, undifferentiated social organization, repressive or penal law, and collective beliefs and sentiments (which at that stage he called the *conscience collective*) that strongly tie the individual to the society as a whole, while organic solidarity, typical of industrialized societies (in their 'normal' state of functioning), based on the division of labor, the fusion of markets, and the growth of cities, exhibits interdependence and

strong social bonds, high volume and density, an increasing differentiation of social functions, restitutive or cooperative law, the gradual withering away of the *conscience collective*, and the emancipation of the individual, allowing space for individual initiative and reflection. The transition from mechanical to organic solidarity, which constituted the 'progress of the division of labor and civilization,' occurs, Durkheim maintained, 'mechanically' as a consequence of growth in the social mass and its social or dynamic density, through expanding markets, a 'struggle for existence,' a resultant development of new occupations and corresponding consumer demand. Durkheim intuitively assumed an equilibrium outcome of increasing division of labor but did not specify the parameters that would generate that outcome. The division of labor exhibits perverse effects and thus abnormal forms where the conditions for the social equilibrium of organic solidarity are not satisfied. *Anomie* – the lack of appropriate rules – is evident in industrial and commercial crises, veering between prosperity and depression, resulting from ignorance and uncertainty of producers concerning demand, in the antagonism between capital and labor, and also in increasingly fragmented scientific specialization, and results in general from a too rapid industrialization in which the appropriate rules have not had time to stabilize relations between social functions and roles. The 'forced division of labor' exists where inequality is seen as illegitimate, as when unequal chances of mobility violate the meritocratic principle. And a third abnormal form, which Besnard calls 'bureaucratic' (Besnard, 1987), is found where excessive specialization is combined with weak productivity and poor mutual adaptation of functions.

Durkheim considerably deepened his notion of *anomie* and his diagnosis of capitalism's ills in *Suicide*. *Anomie* is one of the two major modern social causes of suicide: a condition of rulelessness in which individuals lose their moorings; their expectations exceed the possibilities of satisfying them, potentially insatiable, desires are no longer regulated within a stable and legitimate normative framework of economic rewards or status ranking or conjugal sexual relations. The other major modern social condition causing suicide is egoism: a breakdown of integration, a lack of attachment of individuals to common values, to socially approved relationships, activities, and purposes, isolating and privatizing their lives. (Excessive regulation, or fatalism, and excessive integration, or altruism, are less prevalent, Durkheim thought, in modern societies.) Suicide is the extreme response of the small, randomly distributed but statistically stable minority of suicide-prone individuals whose social situations expose them to the social causes in question. He expounded this theory by accounting for differential suicide rates (e.g., higher among Protestants than Catholics, lower among married men, higher during economic crises, lower during wars) by characterizing suicide's social causes as 'suicidogenic currents,' thereby suggesting a macro–macro explanation that evaded the microlevel, in line with his antiindividualist methodological stance. In reality, his theory of suicide is a macro–micro theory of the social conditions for individual psychological equilibrium, or health (*anomie* being the 'malady of infinite aspiration') that stance deflected him from elaborating. *Suicide* has been widely seen as remarkable for its sophisticated use of statistics (albeit questionable in abstracting from suicide's social meanings to which, likewise, Durkheim paid no attention), including the use of multivariate analysis to exclude alternative explanations. But above all, the work exhibits its author's typical boldness in extending sociological explanation to what might seem least amenable to it: the most private and individual of acts. The same boldness is evident in *The Elementary Forms*, whose introduction and conclusion advance a theory of the fundamental categories of thought, while the rest of the work advances a theory of the origins, meaning and functions of religious beliefs and practices, magic, and the relation of both to science. All this he developed with reference to ethnographic evidence of totemism (assumed to be the most primitive form of religion and its evolutionary origin, or at least its earliest known form) as found in 'simpler,' tribal societies in Australia and America. He made the questionable assumption that the essential elements of religious thought and life should be found, at least in germ, in the most primitive religions, regarding the study of a 'very simple' religion as particularly instructive, since the relations between its constituent elements are 'more apparent' there.

Observing that such categories as space, time, number, causality, the person, etc., impose themselves on us inescapably as the framework of our understanding, Durkheim rejected as nonexplanatory Kant's account of them as given a priori, arguing that they express the most general relations existing between things, which are revealed to us through social action and mutual communication; their necessity has a social basis. Moreover, they are fundamental to commonsense and to science (natural and social) and across cultures. Religion he defined by its universal dichotomizing of the world into the sacred and the profane, uniting believers around the sacred, to which they accord a special, awed respect. Starting from the assumption that religious beliefs cannot be illusory (for otherwise they would not persist), he asked to what reality they correspond, answering that, once decoded, they express our understanding of our society and our obscure and intimate relations with it. They require decoding since, until social science, these cannot be apprehended directly: religion is, in this respect, a sort of mythological protosociology. (Thus, the idea of the soul is the way religion represents the duality of human nature, part corporeal, part social.)

This exclusively society-oriented interpretation of religious symbolism has been much criticized (it was Durkheim, not the savage, Evans-Pritchard suggested, who made society into a God). A further contentious issue is whether or not accepting it challenges religious belief (Durkheim claimed it did not). In the domain of knowledge, science replaces religion but religion continues to symbolize and dramatize social relations, strengthening and reanimating social bonds and believers themselves through the collective enactment of rituals. This aspect of religion survives and would, he speculated, remain basic to the political cohesion of modern, secular societies as civil religion. Religious beliefs themselves originate in such collective effervescences and are renewed by their reenactment.

Cognitively, religions are, he held, cosmologies, offering a 'first explanation of the world,' seeking realities behind appearances, teaching people to unite what the senses separate, and thus prefiguring science. Religion and science are alike in that they both seek to render the (natural and social) world intelligible, but only the latter disposes off methods that

subject its observations to systematic control. Likewise, magic, which occurs within profane or everyday life, draws on prevailing religious beliefs and seeks to predict and manipulate the world, but without benefit of scientific tests, statistics, or experiments, adducing (as scientists also do) auxiliary hypotheses to account for its failures. So for Durkheim, human thinking is unitary, not, as Lucien Lévy-Bruhl thought, divided into two stages – prelogical primitive thinking and logical, science-based modern thinking.

Influence, Impact, and Current Significance

His collaborators applied and developed Durkheim's ideas in many directions, from economics to Chinese thought (in the work of Marcel Granet) to primitive and ancient law (see Besnard, 1983). They also had a considerable contemporary impact on neighboring disciplines in France, for instance, on historians, initially through Henri Berr and his *Revue de synthése historique* and later through Lucien Febvre, Marc Bloch, and the *Annales* school, on human geography, beginning with Paul Vidal de la Blache, on jurists, notably Léon Duguit, and on the study of comparative mythology in the work of Georges Dumézil. The surviving Durkheimians continued into the 1930s but their research program gradually lost its dynamism and, as Raymond Aron observed, it came to seem increasingly irrelevant to the times. Among those who have made a significant impact on subsequent work, one can single out Robert Hertz's studies of death and of right and left (Hertz, 1907, 1909), Maurice Halbwachs's work on collective memory (Halbwachs, 1925) and, above all, Marcel Mauss, Durkheim's nephew, pupil and collaborator, author of several major works on Durkheimian themes, including magic and sacrifice (coauthored with Henri Hubert), seasonal variations among the Eskimos (with Henri Beuchat), the techniques of the body and the category of the person (Mauss, 1950), who succeeded in breaking down Durkheim's conceptual dichotomies and methodological rigidities vis-à-vis other disciplines, in the interest of studying the *fait social total*, as in his classic and influential study of *The Gift* (Mauss, 1925). In several ways, Mauss rendered Durkheimian sociology more subtle, complex, and forward looking. Through Mauss, it influenced French structuralism via his impact on Claude Lévi-Strauss (see his introduction to Mauss, 1950) and also on Louis Dumont, author of the major work on the Indian caste system, *Homo Hierarchicus* and of parallel studies of Western individualism. Its influence is also evident in the *Revue de MAUSS* (*Mouvement antiutilitariste en sciences socials*), which, since 1993 has published antiutilitarian articles on various aspects of social life. Among French sociologists, Georges Friedmann, Pierre Bourdieu, and Raymond Boudon (see Boudon, 1999) have incorporated Durkheimian ideas, albeit far from uncritically, the latter insisting upon a convincing interactionist reading of Durkheim. This has, in turn, been developed by Mohamed Cherkaoui (Cherkaoui, 1998) who has, moreover, resurrected for close attention Durkheim's sociology of education, in particular his remarkable study of the evolution of French educational thought and institutions, and their interaction, since the early Church (Durkheim, 1938), which explores various themes less than evident in his better known works: class power and conflict and the struggle for cultural domination, the role of intellectuals, and the sociology of educational knowledge focusing on dominant epistemologies and disciplinary hierarchies. Another important work is the definitive study of the concept of anomie and its career within sociology (Besnard, 1987). Besnard also inaugurated the valuable *Études durkheimiennes* in Paris in 1977 as a forum for scholarship on Durkheim and the Durkheimians. It moved to Urbana, Illinois under the editorship of Robert Alun Jones in 1989 and to Oxford in 1995, published by the British Centre for Durkheimian Studies under the inspiration of William Pickering, along with a continuing series of important Occasional Papers edited by Willie Watts Miller.

The Durkheim school's impact on British social anthropology has been immense, initially through Radcliffe-Brown but chiefly through his successor E.E. Evans-Pritchard, whom it helped liberate from his predecessor's structural-functionalism. Despite serious reservations, he identified himself with the work of the Durkheimians, praising it as an indispensable theoretical capital, embodying rigorous methods of scholarship (see his introduction to Hertz, 1907). In the early 1950s, he launched an extensive series of English translations of works by Durkheim and his collaborators, and himself drew on Durkheimian ideas, especially in his work on the Nuer. The impact of Durkheimian ideas is also evident in the work of other Oxford anthropologists, particularly that of Rodney Needham and Nick Allen on symbolic classification, but also in the various writings of Mary Douglas on pollution and symbolic boundaries, and in the work of Robin Horton on African religions.

Durkheim's reception into American sociology began early, influencing, among others, Elton Mayo and W. Lloyd Warner, but it was Talcott Parsons's extended treatment of his work in his *Structure of Social Action* that marked his canonization as a 'grand theorist,' treated there as a pre-Parsonian protostructural functionalist, with a distinctive solution to the 'problem of order' through the internalization and institutionalization of consensual, society-wide values.

Durkheim's famous account of social facts as *sui generis* in *The Rules of Sociological Method* was for long dismissed as unacceptable but has achieved a new lease of life with increasing acceptance of the idea of 'emergence,' according to which patterns and outcomes occurring at a higher level may be neither explicable nor predictable at a lower level. Just as theories, concepts, laws, and terminology at the psychological level may be irreducible to the neurobiological, while mental events and properties have real effects there, the same may be true of the social level vis-à-vis the level of interacting individuals, without any ontological implication of the existence of a 'group mind' or the like. Durkheim himself argued this and can thus be seen as a theorist of emergence *avant la lettre*.

Substantively, the impact of his ideas was pervasive, influencing innumerable accounts of industrialization, urbanization, social control, social disorganization, and collective behavior. These were, directly or indirectly, extrapolated from Durkheim's picture of advancing social differentiation leading through rapid social change to the breakdown of social ties. According to such accounts, rapid industrialization generates disharmony between life experiences and the normative framework, which regulates them, leading to virtually unlimited possibilities for group conflict and thus to a variety of social movements: peaceful

agitation, political violence, millenarianism, nationalism, revolution, underground subversion, etc., that in turn attract those suffering most severely under the displacements created by structural change. Examples of this approach are found in the work of Samuel Huntington, Chalmers Johnson, Ted Gurr, and Neil Smelser. Such accounts were criticized by Tilly (1981) for relying on three basic 'Durkheimian' arguments that are held to have been refuted by the historical evidence: (1) that conflict varies directly with the prevalence of differentiation within organization and the scale of groups; (2) that social disorder correlates with accelerated urbanization or industrial growth; and (3) that the various forms of 'disorder,' such as crime, suicide, conflict, and protest, vary together, in the same settings and circumstances. In short, Tilly rejects the 'Durkheimian line' on conflict, protest, and collective action, which he explains instead in terms of group solidarities, common interests, and mobilization. But, first, the 'line' is extrapolated from, not argued by, Durkheim, who never sought to explain so much by so little; it is his American successors, not Durkheim, who posited the strains resulting from rapid industrialization as the catch-all source of so many 'disorders,' which he never treated as equivalent and associated. Second, his ideas about the processes creating and reanimating collective solidarity are surely not irrelevant to the explanation of collective action. And third, his focus on the consequences of rapid economic and social change outstripping the norms that could regulate it has an inescapable relevance for the diagnosis of the distinctive ills of the most advanced forms of contemporary capitalism, in particular the corrosive effects of technical change on social relations and individual lives.

The concept of *anomie* has had an impact in other fields. Robert Merton imported it in his influential paper on *Social Structure* and *Anomie* (Merton, 1949) but in the process mis- or reinterpreted Durkheim's notion as a maladaptation of means to ends, rather than a sociopsychological condition of deregulation (so that the problem, on Durkheim's account, should be seen as lying in the success ethic itself, not in the socially structured inability of some to realize its promises). But this idea, together with the other causes of suicide, and indeed the structure, content and mode of its attempted explanation, have exercised a continuing and productive stimulus within the sociology of deviance and criminology, as often through attempts to develop or revise Durkheim's contribution as through more or less radical refutations of his whole approach.

Within the sociology of law, Durkheim has had a major, recently revived and continuingly fertile impact. Here the central influential idea, foundational to the contemporary sociology of punishment, has been his account of crime and punishment: crime is deviant conduct that violates prevailing norms and offends the consciences of the upright, provoking them to passionately felt punitive reactions that are authoritatively enforced by law. Punishment thus serves not to control crime but to sustain and enhance solidarity through ritual ceremonies and symbolic effects. More specific, and most striking, has been his insight into the paradox that punishment presupposes and does not create attachment; its function of moral condemnation undermines its preventive or retributive effectiveness, insofar as offenders' attachment to the moral community is weak or minimal. This insight can be seen developed in the two major present-day sociologists of punishment, despite their very different approaches, namely David Garland and Loic Wacquant, both of whom readily acknowledge indebtedness to Durkheim.

It is, however, his later 'religious sociology' that has had most impact within sociology since the 1990s, applying its concepts, insights, and hypotheses to historical and modern societies. Durkheim and the Durkheimians never did this, and indeed a surprised Mauss could write in the late 1930s, contemplating Nazism and Fascism, of the tragic 'verification' of their theory in the spectacle of large modern societies "hypnotized like Australians [aborigines] are by their dances" (Ranulf, 1939; cited in Lukes, 1972). From the 1960s through the 1980s, practitioners of historical sociology were largely focused on the rise and development of capitalism and were class-oriented and state-centered in approach: Marx, Weber, and Tocqueville were the presiding ancestors. Since then a growing interest in cultural analysis has been evident, with an associated interest in Durkheimian ideas. Initially, this development had a Parsonian, consensual emphasis. Edward Shils argued that secular, differentiated societies have sacred 'centers' that explain the allocation of status (Shils, 1975), Clifford Geertz interpreted cultural systems in a way that linked sacredness, ritual and the mobilization of solidarity (Geertz, 1973), and Robert Bellah published and stimulated studies of civil religion under modern conditions (Bellah and Hammond, 1980). More recently, there has been a growing interest in symbolism and ritual and their relation to the institutionalization of power and social control but without the assumption of social consensus. In Jeffrey Alexander's words, a 'late Durkheimian framework' is applied to 'analyze phenomena, which are disruptive, contingent, and related to social change,' while Randall Collins, exponent of conflict sociology, has sought to combine Durkheim with Erving Goffman, in exploring how 'focused and emotionally arousing interaction produces group solidarity' (Alexander, 1988). Studies along such 'late Durkheimian' lines have included French Revolutionary festivals, ritual and symbolism under Fascism, nationalist movements, mass strikes, and media events. Among contemporary social and cultural anthropologists, Joel Robbins, who draws on Collins's theory of ritual chains, and Maurice Bloch, who draws on cognitive science, have both attempted to preserve Durkheimian insights by means of arguments that differ from those he himself employed: Bloch distinguishes 'the transcendental social' of imagination from 'the transactional social' of everyday interaction, while Robbins focuses on the heightened emotional energy generated by ritual collective effervescences.

These are the main lines of Durkheim's impact upon twentieth-century social science. I have not discussed his impact beyond the social sciences (for example, on and through such partly literary figures as Roger Caillois, Michel Leiris, and George Bataille, members of the so-called *Collège de Sociologie* in the late 1930s), or countless unacknowledged influences. More specific and acknowledged social scientific influences could also be cited, for instance, upon the work of Basil Bernstein on sociolinguistic and educational 'codes,' drawing on Durkheim's ideal-typical polarities, and upon the so-called 'strong program' in the sociology of science, originating in the University of Edinburgh, focusing, in a relativizing vein, on the 'social' origins of 'natural' classifications (see Bloor, 1976). Here, as ever, Durkheim's ideas are interpreted in

accordance with locally generated concerns. The story of their reception, sketched here, confirms both their continuing attraction and their continuing power to generate alternative and invariably controversial readings.

See also: Altruism and Prosocial Behavior, Sociology of; Anomie; Civil Religion; Collective Behavior, Sociology of; Collective Beliefs, Sociological Explanation of; Collective Memory, Sociology of; Determinism: Social and Economic; Functionalism, History of; Individuals in Society: History of the Concept; Knowledge, Sociology of; Law: Change and Evolution; Legitimacy, Sociology of; Mauss, Marcel (1872–1950); Positivism, Sociological; Primitive Society; Religion, Sociology of; Ritual; Secularization; Societies, Types of; Sociological Theory; Sociology, Epistemology of; Sociology, History of; Solidarity: History of the Concept; Status and Role: Structural Aspects; Suicide, Sociology of; Values, Sociology of.

Bibliography

Alexander, J.C. (Ed.), 1988. Durkheimian Sociology: Cultural Studies. Cambridge University Press, Cambridge, UK.

Alexander, J.C., Smith, P., 2005. The Cambridge Companion to Durkheim. Cambridge University Press, Cambridge, UK.

Bellah, R.N., Hammond, P. (Eds.), 1980. Varieties of Civil Religion. Harper and Row, San Francisco.

Besnard, P. (Ed.), 1983. The Sociological Domain. The Durkheimians and the Founding of French Sociology. Cambridge University Press, Cambridge, UK.

Besnard, P., 1987. L'Anomie, ses usages et ses fonctions dans la discipline sociologique depuis Durkheim. Presses Universitaires de France, Paris.

Bloch, M., 2007. Durkheimian anthropology and religion: going in and out of each other's bodies. In: Whitehouse, H., Laidlaw, James (Eds.), Religion, Anthropology and Cognitive Science. Academic Press, Durham, NC, pp. 63–80.

Bloor, D., 1976. Knowledge and Social Imagery. Routledge, London.

Boudon, R., 1999. Les Formes élémentaires de la vie religieuse: une théorie toujours vivante. L'Année sociologique 49 (1), 149–198.

Cherkaoui, M., 1998. Naissance d'une science sociale. La sociologie selon Durkheim. Droz, Geneva, Switzerland.

Collins, R., 2005. Interaction Ritual Chains. Princeton University Press, Princeton.

Cotterrell, R., 1999. Émile Durkheim. Law in a Moral Domain. Edinburgh University Press, Edinburgh, UK.

Durkheim, E., Mauss, M., 1903. De quelques formes primitives de classification. L'Année sociologique 6, 1–72 (1963 Primitive Classification. Cohen and West, London and University of Chicago Press, Chicago).

Durkheim, E., 1893 [1902]. De la division du travail social: étude sur l'organisations des sociétés supérieures, second ed. Alcan, Paris. (2014 The Division of Labor in Society. Free Press, New York and Palgrave Macmillan, London).

Durkheim, E., 1895 [1901]. Les Régles de la méthode sociologique, second ed. Alcan, Paris. (2014 The Rules of Sociological Method. Free Press, Palgrave Macmillan, London and New York).

Durkheim, E., 1897. Le Suicide: étude de sociologie. Alcan, Paris (1951 Suicide. Free Press, Glencoe, Il and 1952. Routledge and Kegan Paul, London).

Durkheim, E., 1912. Les Formes élémentaires de la vie religieuse: le systéme totemique en Australie (1995 The Elementary Forms of the Religious Life. Free Press, New York).

Durkheim, E., 1924. Sociologie et philosophie. Alcan, Paris (1953 Sociology and Philosophy. Cohen and West, London and Free Press, Glencoe, IL).

Durkheim, E., 1938. L' Évolution pédagogique en France. Alcan, Paris (1977 The Evolution of Educational Thought: Lectures on the Formation and Development of Secondary Education in France. Routledge and Kegan Paul, London and Boston).

Durkheim, E., 1950. Leçons de sociologie: physique des moeurs et du droit (1957 Professional Ethics and Civic Morals. Routledge & Kegan Paul, London; 1958, Free Press, New York).

Emirbayer, M., 1996. Useful Durkheim. Sociological Theory 14 (2), 109–130.

Fournier, M., 2007. Emile Durkheim, Fayard, Paris (2012 Emile Durkheim Polity, Cambridge, UK).

Garland, D., 2002. Culture and Control: Crime and Social Order in Contemporary Society. University of Chicago Press, Chicago and London.

Garland, D., 2012. Punishment and solidarity. In: Simon, J., Sparks, R. (Eds.), The Sage Handbook of Punishment and Society. Sage, Thousand Oaks, CA.

Geertz, G., 1973. The Interpretation of Cultures. Basic Books, New York.

Halbwachs, M., 1925. Les Cadres sociaux de la mémoire. Alcan, Paris (1992 On Collective Memory. University of Chicago Press, Chicago).

Hertz, R., 1907. Contribution à une étude sur la représentation de la mort. L'Année sociologique 10, 48–137 (1960 *Death* and The *Right* Hand. Cohen and West, London).

Hertz, R., 1909. La prééminence de la main droite. Revue Philosophique 68, 553–580 (1960 Death and the Right Hand. Cohen and West, London).

Jones, R.A., 1999. The Development of Durkheim's Social Realism. Cambridge University Press, Cambridge, UK.

Lukes, S., 1972. Émile Durkheim: His Life and Work. Harper and Row. New York and 1973. Allen Lane The Penguin Press, London.

Lukes, S., 2012. Is Durkheim's understanding of religion compatible with believing? Religion 42 (1), 41–52.

Mauss, M., 1925. Essai sur le don. L'Année sociologique 1, 30–186 (republished in Mauss, M. 1950/1990. The Gift, Routledge, London).

Mauss, M., 1950. Sociologie et anthropologie. Presses Universitaires de France, Paris (fourth enlarged ed. 1968).

Merton, R.K., 1949. Social structure and anomie, 1968. In: Merton, R.K. (Ed.), Social Theory and Social Structure. Free Press, New York, pp. 185–214.

Pickering, W., 1984. Durkheim's Sociology of Religion. Themes and Theories. Routledge & Kegan Paul, London.

Ranulf, S., 1939. Scholarly forerunners of Fascism. Ethics 50, 16–34.

Robbins, J., 2010. If There is no such thing as society, is ritual still special? On using the elementary forms after tarde. In: Candea, Matei (Ed.), The Social After Gabriel Tarde. Routledge, London, pp. 93–101.

Sawyer, R.K., 2002. Durkheim's Dilemma: toward a sociology of emergence. Sociological Theory 20 (2), 227–247.

Sawyer, R.K., 2005. Social Emergence? Societies as Complex Systems. Cambridge University Press, Cambridge, UK.

Schmaus, W., 1994. Durkheim's Philosophy of Science and the Sociology of Knowledge: Creating an Intellectual Niche. University of Chicago Press, Chicago and London.

Schmaus, W., 2007. Rethinking Durkheim and his Tradition. Cambridge University Press, Cambridge. UK.

Shils, E., 1975. Center and Periphery: Essays in Macrosociology. Chicago University Press, Chicago.

Tilly, C., 1981. Useless durkheim. In: Tilly, C. (Ed.), As Sociology Meets History. Academic Press, New York, pp. 95–101.

Wacquant, L., 2009. Punishing the Poor: The Neoliberal Government of Social Insecurity. Duke University Press, Durham, NC and London.

Durkheimians, The

Marcel Fournier, Université de Montréal, Montréal, QC, Canada; and Paris Institute for Advanced Study, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Strictly speaking, the Durkheimians are the disciples of Durkheim. But, who were these disciples? Mainly, his collaborators on the journal *L'Année sociologique*. The first part of this article is a sociological analysis of the characteristics of this group of 50 collaborators: their disciplinary backgrounds, networks, generations, and participation to the journal. The second part is an analysis of the 'destiny' of this French School of Sociology since Durkheim's death until today with the presentation of two new trends: a renewal of the Durkheimian studies and the development of a Neo-Durkheimian perspective in sociology.

The substantive 'Durkheimian' is related to Émile Durkheim (1858–1917), the founder of sociology in France. Strictly speaking, the Durkheimians are the disciples of Durkheim. But, who were these disciples? Mainly, his collaborators on the journal *L'Année sociologique*, but they were not the only ones. Some Durkheimians belonged to the network of this group of collaborators or had intellectual affinities with Durkheim: Lucien Levy-Bruhl was a friend and colleague of Durkheim at the Sorbonne and close to Durkheim's disciples who were members of the Socialist Party, but he never collaborated with *L'Année*; he became a Durkheimian while keeping his intellectual independence for almost his entire life, defending the thesis of a 'prelogical mind' against Durkheim himself. Other scholars published articles in the journal but were excluded – such as Henri Muffand, a colleague of Bouglé at the Lycée, who was too close to anthroposociology – or maintained a critical distance from Durkheim: Gaston (born in 1860), author of a book entitled *Essais sur l'origine du droit* (1892), was a member of the first group of collaborators, responsible for the section on criminology. Thus, it was possible to collaborate with *L'Année sociologique* without being a Durkheimian, and vice versa.

The first volume of *L'Année sociologique*, founding the journal's founding in 1896, was published 2 years later. This volume was the first in a series of 13 to be published under the direction of Émile Durkheim. A second series of just two volumes (1925 and 1927) was published after the First World War under the direction of Marcel Mauss (1872–1950).

Philippe Besnard has published an excellent analysis of the first series of *L'Année*, identifying the first group of founders and the internal networks of all collaborators (Besnard, 1979). The founder of the journal was Émile Durkheim, but it would seem that Celestin Bouglé, a young philosopher, was the one to have the idea and suggested to Durkheim that he create the journal. Durkheim debated the general orientation of the journal with Bouglé and his friend Paul Lapie: the issue was the distinction between sociology and psychology. Durkheim agreed to define sociology as a form of psychology, but a specific one, i.e., collective psychology.

Émile Durkheim also succeeded in involving his nephew, Marcel Mauss, then a young *agrégé de philosophie* and a student in history of religion at the École Pratique des Hautes Études: "[…] Il n'est pas nécessaire d'être beaucoup: le fait que nous ne soyons que deux nous singularise déjà" [There is no need to be many; the fact that there are just two of us already sets us apart] (Durkheim, 1897: p. 81). Mauss' job was to convince some of his friends, also *agrégés*, to take part in this adventure and become sociologists: Henri Hubert, Paul Fauconnet, Emmanuel Lévy, Albert Milhaud. The 'top brass' of the team included Bouglé, Fauconnet, Hubert, Mauss, and Simiand, and they were among the most active collaborators.

The new journal had two parts: the first part consisting of two or three essays (the 'Mémoires' in the first volume were a translation of G. Simmel's paper, *Comment les formes sociales se maintiennent*, and an original paper by Durkheim, *Prohibition de l'inceste et ses origines*) and the second with book reviews. It was the journal's 'mission' to describe the 'current state of literature' in the social sciences, according to Durkheim, who introduced a division of labor between his collaborators, giving each of them responsibility for sections of reviews. In the first issue of *L'Année*, these were: General Sociology (C. Bouglé), Religious Sociology (M. Mauss and H. Hubert), Moral Sociology & Sociology of Law (P. Lapie, E. Lévy, E. Durkheim, M. Mauss), Criminal Sociology (G. Richard), Sociology of Economics (F. Simiand), and miscellaneous (H. Muffang, E. Durkheim, P. Fauconnet, and D. Parodi). The two largest sections were Religious Sociology, and Moral Sociology & Sociology of Law.

The list of the collaborators in the first series, ranked by how important their collaboration was (i.e., the number of reviews they wrote) is E. Durkheim, M. Mauss, H. Hubert, F. Simiand, G. Richard, H. Bourgin, P. Fauconnet, C. Bouglé, M. Halbwachs, P. Lapie, P. Huvelin, D. Parodi, G. Bourgin, A. Meillet, R. Hertz, G. Davy, E. Lévy, A. Aubin, P. Hourticq, J. Ray, H. Muffang, A. Bianconi, Ph. de Felice, I. Levy, Ls. Gernet, A. Milhaud, Ed. Laskine, J. Marx, L. Reynier, H. Beuchat, M. David, Ré Challié, G. Gelly, Poirot, J.T. Stickney, A. Vacher, E. Doutté, H. Jeanmaire, A. Demangeon, Foucault, F. Sigel, P. Roussel, Ch. Fossey, P. Lafitte, Ch. Lalo, E. Claude Maître, and A. Moret (Besnard, 1979).

These almost 50 collaborators published more than 2000 book reviews. Some of them contributed to all the volumes (Durkheim, Mauss, Hubert, Simiand), others to several volumes, and still others to just one. They came from different disciplines (philosophy, philology, archeology and history, geography, linguistics, etc.). Durkheim held a central position, but there were different networks: Bouglé, Lapie, and Parodi; Simiand, Halbwachs, and the Bourgin brothers; Mauss, Hubert, and their students. We can identify three generations: (1) the elders (born

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61191-1

before 1860), with Durkheim, Richard, and colleagues (Meillet, Moret); (2) the generation born between 1870 and 1880 with the students of Durkheim at Bordeaux (A. Aubin, M. Foucault, R. Hourticq, Ch. Lalo) and his nephew Marcel Mauss and his friends (Hubert, Fauconnet, Simiand); and (3) the generation of younger scholars, with students of Mauss and Hubert at École Pratique des Hautes Études (H. Beuchat, M. David, G. Davy, L. Gernet, R. Hertz, H. Jeanmaire, Ph. De Felice, J.P. Lafitte, J. Marx, J.P. Reynier, P. Roussel).

The team of *L'Année* was a real research laboratory: 18 essays ('Mémoires') were written by the collaborators. There were not only collaborations between them (Durkheim and Mauss, Mauss and Hubert, etc.) but also intense debates. This was a challenging, but ultimately successful, enterprise that gave visibility to the 'French school of sociology,' which was 'Durkheimian sociology.'

The institutionalization of this school was confirmed when its members gained positions in the educational system: in *lycées*, universities (Bordeaux, Toulouse, Paris), and other schools and institutions (École Pratique des Hautes Études, Conservatoire National des Arts et Métiers). Some of them also published papers and books discussing and defending the Durkheimian perspective in sociology, from *Sociologie: l'objet et la méthode* by Mauss and Fauconnet (1900) to *La sociologie de Durkheim* (1911) by Georges Davy and *La doctrine d'Émile Durkheim* (1916) and *Les Origines du sentiment religieux* (1925) by Maurice Halbwachs.

But the First World War brought tragedy to the Durkheimian school with the death of Émile Durkheim and his son André, as well as the loss of his young collaborators and students: Robert Hertz, Maxime David, Antoine Bianconni, etc. "Nous avons été privés de toute une génération de nos meilleurs et nos plus vigoureux collaborateurs […]. On verra que s'il n'y avait pas eu la guerre, la sociologie, la science, notre pays serait riche d'une œuvre comme peu d'études en ont produit de pareilles" [We were deprived of an entire generation of our best and most energetic collaborators. It will be seen that if the War had not taken place, sociology, science, our country would be rich with such an opus that few studies have produced to such a level] (Mauss, 1925). Marcel Mauss, the 'heir,' 'rekindled' *L'Année sociologique* but only two volumes were published in 1925 and 1927 with the participation of the 'old' collaborators of the first series: Marcel Mauss, Henri Hubert, Célestin Bouglé, François Simiand, Paul Fauconnet, and Maurice Halbwachs. In the first of the two volumes, Mauss published *In Memoriam* and his famous *Essai sur le don*.

In the 1930s, Durkheimians held three of the four chairs of sociology in French universities: Maurice Halbwachs at Strasbourg, and Paul Fauconnet and Bouglé at the Sorbonne. The fourth one, at Bordeaux, was held by Gaston Richard, a former Durkheimian. And Paul Lapie, then director of Primary Education, succeeded in introducing sociology into the curriculum of the *licence* (bachelor's) in philosophy. Durkheimian sociology saw a 'fresh' start with many activities and initiatives: creation of the Institut d'Ethnologie de Paris (1925), Marcel Mauss's election to the prestigious Collège de France (1930), the publication of the first volume of the new journal *Annales sociologiques* in 1934, and François Simiand's election to the Collège de France.

But, another 'dark' period came with the deaths of F. Simiand (1936), L. Levy-Bruhl (1939), and C. Bouglé (1940). And M. Halbwachs, elected to the Collège de France in 1943, passed away in 1945. Moreover, Marcel Mauss was ill; he died in 1950. This was, in a way, the end of the Durkheimian school, which had become the target of more frequent and very severe criticism in the 1930s and 1940s. Paul Nizan in *Les Chiens de garde* attacked Durkheim and his disciples, who were associated with the bourgeoisie and its conservative ideology (Nizan, 1932: p. 98). Jules Monnerot, a cofounder of the Collège de Sociologie and a friend of Georges Bataille and Roger Caillois, published in 1946 a book entitled *Les Faits sociaux ne sont pas des choses*, which was an epistemological critique of Durkheim's methodology. Finally, when Jean Stoezel came back from the United States after the War and gave his first lesson in sociology at Bordeaux, he concluded: "We should forget Durkheim" (Stoetzel, 1946).

But, Durkheim has not been forgotten. A well-known classical sociologist in the academic world, he has become the subject of a number of commentaries and theoretical interpretations. For instance, 6.3% of the 8353 articles published in the AJS and the ASR between 1895 and 1992 quote his texts. This is as many as Weber, the most influential sociologist of the era, who was quoted in 6.5% of the articles over the same period. So "[t]he knowledge of Durkheim and of his books is a full part of the definition of what a sociologist is in America" (Vogt, 1993: p. 227).

A first great intellectual biography of Durkheim was published by S. Lukes (1972). In addition, a 'rebirth' of all of Durkheim's texts took place in three volumes (Durkheim, 1975), edited by V. Karady, also the editor of three volumes of Marcel Mauss's *Oeuvres* (1969). Over the past 30 years, with the opening up of archives that have given us new data about Durkheim's life and work, Durkheimian studies have been growing rapidly, punctuated by the activities of Ph. Besnard in France, W.S. Pickering in England, R. Allen Jones in the United States, or M. Borlandi in Italy, and the publication of a new journal, *Études Durkheimiennes*, which became *Durkheimian Studies/Études Durkheimiennes*. Another sign of this rebirth is the multiplication of PhD theses, special issues of journals, a new intellectual biography of Durkheim (Fournier, 2007), and books on the French school of thought, mainly on Marcel Mauss: publication of the first intellectual biography of Mauss (Fournier, 1994), publication of his political texts, and creation of the *Revue du M.A.U.S.S.*

Finally, in Anglo-Saxon sociology, there has been not only a renewal of Durkheimian Studies (Alexander, l994; Cladis, 1992; Gane, 1992; Giddens, l986; Mestrovic, 1988; Schmauss, l994; Warfeild Rawls), but also the development of what we can call a Neo-Durkheimian perspective in sociology: three recent examples of this renewal are the 'strong program' of David Bloor (1976) and his colleagues in the sociology of science; the work of Randall Collins on violence in the sociology of interactions; and the work of Jeffrey Alexander (1996) and his research group in cultural sociology. Based on the 'reinterpretation' of *Elementary Forms of the Religious Life*, this renewal has given a central focus to the study of rituals (Bellah, 2005; Collins, 2004; Douglas, 1970/1982; Goffman, 1967) and to the 'rediscovery' of the idea of 'effervescence' (Alexander et al., 2006). Not only religion, but all social life can be seen as a sequence of rituals and more generally as a performance.

Editor's Note

Emile Durkheim was a well-known and articulate proponent of the 'new science' of sociology and enjoyed a coterie of collaborators and followers during his lifetime and, of course, thereafter. 'The Durkheimians' include both historical figures and a large group of present-day sociologists who proudly trace their work to Durkheim's path-breaking insights. The following bibliographic entry is meant both to document and celebrate the long and continuing influence of Durkheim's thought not just in sociology but in anthropology and allied disciplines as well.

See also: Discipline Formation in the Social Sciences; Durkheim, Emile (1858–1917); Mauss, Marcel (1872–1950); Sociological Theory; Sociological Traditions; Sociology, History of; Sociology: Overview.

Bibliography

Alexander, Jeffrey, Philip, Smith (Eds.). Cambridge Companion to Durkheim. Cambridge University Press, Cambridge.

Alexander, Jeffrey C., Giesen, Bernhard, Mast, Jason L., 2006. Social Performance: Symbolic Action, Cultural Pragmatics and Ritual. Cambridge University Press, Cambridge.

Bellah, Robert N., 2005. Durkheim and ritual. In: Alexander, Jeffrey C., Smith, Philip (Eds.), The Cambridge Companion to Durkheim. Cambridge University Press, Cambridge, pp. 183–211.

Besnard, Philippe, 1979. La formation de l'équipe de *L'Année sociologique*. Revue française de sociologie 20 (1), 7–31.

Blondiaux, Loïc, 1991. Comment rompre avec Durkheim? Jean Stoetzel et la sociologie française de l'après-guerre (1945–1958). Revue française de sociologie 32 (3), 411–441.

Bloor, David, 1976. Knowledge and Social Imagery. Routledge and Kegan Paul, London.

Cladis, M., 1992. A Communitarian Defense of Liberalism. Stanford University Press, Stanford, CA.

Collins, Randall, 2004. Interaction Ritual Chains. Princeton University Press, Princeton, NJ.

Collins, Randall, 2008. Violence: A Micro-Sociological Theory. Princeton University Press, Princeton, NJ.

Douglas, Mary, 1970/1982. Natural Symbols: Explorations in Cosmology. Pantheon Books, New York.

Durkheim, Émile, 15 July 1897. Letter to Marcel Mauss. In: Besnard, Ph, Fournier, Marcel (Eds.), Émile Durkheim, Lettres à Marcel Mauss. PUF, Paris, pp. 81–82.

Durkheim, Émile, 1975. Textes, 3 vol. Éditions de Minuit, Paris.

Fournier, Marcel, 1994. Marcel Mauss. Fayard, Paris. English translation: Princeton University Press. 1996.

Fournier, Marcel, 2007. Émile Durkheim. Fayard, Paris. English translation: Polity, 2013.

Gane, Mike (Ed.), 1992. The Radical Sociology of Durkheim and Mauss. Routledge, London.

Giddens, Anthony, 1971. Durkheim's early works. In: Capitalism and Modern Social Theory. CUP, Cambridge, pp. 65–81.

Goffman, Erving, 1967. Interaction Ritual: Essays on Face-to-Face Behavior. Anchor Books.

Halbwachs, Maurice, 1916. La doctrine d'Émile Durkheim. Revue Philosophique, 353–375, 42e année, nos. 5-6, May–June.

Halbwachs, Maurice, 1925. Les Origines du sentiment religieux. Librairie Stock.

Heilbron, Johan, 1920–1940. Les métamorphoses du durkheimisme. Revue française de sociologie 26 (2), 203–237.

Lukes, Steven, 1972. Émile Durkheim. His Life and Work. Harper & Row, New York.

Marcel, Jean-Christophe, 2001. Le durkheimisme dans l'entre deux guerres. PUF, Paris.

Mauss, Marcel, 1925. In memoriam. L'œuvre inédite de Durkheim et ses collaborateurs. In: L'Année Sociologique, n.s. 2. Marcel Mauss, Œuvres, vol. 3. Éditions de Minuit, Paris, pp. 473–499, 1969.

Mauss, Marcel, 1968–1969. Œuvres, 3 vol. Éditions de Minuit, Paris.

Mestrovic, Stephen G., 1988. Durkheim and the Reformation of Sociology. Rowman & Littlefield, Totowa, NJ.

Nizan, Paul, 1932. Les Chiens de garde. François Maspero, Paris, 1981.

Rawls, Anne Warfeild, 2004. Epistemology and Practice: Durkheim's Elementary Forms of Religious Life. Cambridge University Press, Cambridge.

Stoetzel, Jean, 1946. L'esprit de la sociologie contemporaine. Bulletin de la Société de philosophie de Bordeaux (4), 2–20.

Strenski, Ivan, 1997. Durkheim and the Jews of France. The University of Chicago Press, Chicago.

Vogt, W. Paul, 1993. L'influence de la *Division du travail social* sur la sociologie américaine. In: Besnard, Ph, Borlandi, M., Vogt, P. (Eds.), Division du travail social et lien social: la thèse de Durkheim un siècle après. PUF, Paris, pp. 197–215.

Dynamic Decision Making

Jared M Hotaling, University of Basel, Petersplatz, Basel, Switzerland
Pegah Fakhari and Jerome R Busemeyer, Indiana University, Bloomington, IN, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.R. Busemeyer, vol 6, pp. 3903–3908, © 2001, Elsevier Ltd.

Abstract

This section reviews a specialty within the field of decision making known as dynamic decision making. Dynamic decisions are characterized by a decision maker choosing among various actions at different points in time in order to control and optimize performance of a dynamic stochastic system. Realistic examples include fighting fires, navigational control, battlefield decisions, medical emergencies, and so on. The section has four parts. The first reviews basic theory concerning optimal decision principles in a dynamic context, the second summarizes empirical approaches to the study of human performance on dynamic decision tasks, the third presents theoretical models that describe how humans learn to control dynamic systems, and the last discusses methodological issues arising from the study of complex decisions including differences between field versus laboratory research.

Dynamic decision making (DDM) is defined by three common features: a series of actions must be taken over time to achieve some overall goal, the actions are interdependent so that later decisions depend on earlier actions, and the environment changes both spontaneously and as a consequence of earlier actions (Edwards, 1962). Dynamic decision tasks differ from sequential decision tasks (see Diederich, 2001) in that the former are primarily concerned with controlling dynamic systems over time, whereas the latter are more concerned with sequential search for information to be used in making decisions.

Psychological research on DDM began with Toda's (1962) pioneering study of human performance on a game called the 'fungus eater,' in which human subjects controlled a robot's search for uranium and fuel on a hypothetical planet. Subsequently, human performance has been examined across a wide variety of dynamic decision tasks including computer games designed to simulate stock purchases (Ebert, 1972; Rapoport, 1966), welfare management (Dorner, 1980; Mackinnon and Wearing, 1980), vehicle navigation (Jagacinski and Miller, 1978; Anzai, 1984), health management (Kleinmuntz and Thomas, 1987; Kerstholt, 1994), production and inventory control (Sterman, 1989; Berry and Broadbent, 1988), supervisory control (Kirlik et al., 1993a,b), firefighting (Brehmer and Allard, 1991; Brehmer and Nahlinder, 2007), prey and predator problems (Jensena and Brehmer, 2003), plumbing and cash flow tasks (Gonzalez and Dutt, 2011; Sweeney and Sterman, 2000), and farming on Mars (Gureckis and Love, 2009a). Cumulative progress in this field has been summarized in a series of empirical reviews by Edwards (1962), Rapoport (1975), Funke (1991), Brehmer (1992), Sterman (1994), Kerstholt and Raaijmakers (1997), and Jagacinski and Flack (2003).

Stochastic Optimal Control Theory

To illustrate how psychologists study human performance on dynamic decision tasks, consider the following experiment by You (1989). Subjects were initially presented a 'cover' story describing the task: "Imagine that you are being trained as a psychiatrist, and your job is to treat patients using a psychoactive drug to maintain their health at some ideal level." Subjects were instructed to choose the drug level for each day of a simulated patient after viewing all of a patient's previous records (treatments and health states). Subjects were trained on 20 simulated patients, with 14 days per patient, all controlled by a computer simulation program.

There are a few general points to make about this type of task. First, laboratory tasks such as this are oversimplifications of real-life tasks, designed for experimental control and theoretical tractability. However, more complex simulations also have been studied to provide greater realism (e.g., Brehmer and Allard's, 1991, firefighting task). Second, the above task is an example of a discrete time task (only the sequence of events is important), but real-time simulations have also been examined where the timing of decisions becomes critical (e.g., Brehmer and Allard's, 1991, firefighting task). Third, the cover story (e.g., health management) provides important prior knowledge for solving the task, and so the findings depend both on the abstract task properties as well as the concrete task details (Kleiter, 1975). Fourth, the stimulus events are no longer under complete control of the experimenter, but instead they are also influenced by the subject's own behavior. Thus experimenters need to switch from a stimulus–response toward a cybernetic paradigm for designing research (cf. Brehmer, 1992; Rapoport, 1975).

This health management task can be formalized by defining $H(t)$ as the state of the patient's health on day t; $T(t)$ is the drug treatment presented on day t, and $w(t)$ is a random shock that may disturb the patient on any given day. Figure 1 is a feedback diagram that illustrates this dynamic decision task. In this figure, S represents the environmental system that takes both the disturbance, w, and the decision maker's control action, T, as inputs, and produces the patient's state of health, H, as output. D represents the decision maker's policy that takes both the observed, H, and desired, H^*, states of health as inputs, and produces the control action, T, as output.

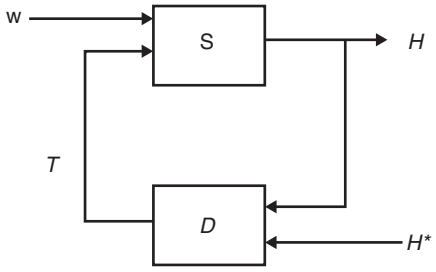

Figure 1 A feedback diagram of an example dynamic decision task.

Based on these definitions, this task can be analyzed as a stochastic linear optimal control problem (Rouse, 1980): determine treatments $T(1)$, …, $T(N)$, for $N = 14$ days, which minimize the objective function

$$F = E\left\{\sum_{t=1,N}\alpha[H(t) - H^*(t)]^2 + \beta T^2(t)\right\} \quad [1]$$

contingent upon the linear stochastic dynamic system,

$$\begin{aligned} H(t+1) = {} & a_1H(t) + a_2H(t-1) + b_1T(t) \\ & + b_2T(t-1) + b_3T(t-2) + w(t) \end{aligned} \quad [2]$$

Standard dynamic programming methods (Bertsekas, 1976) may be used to find the optimal solution to this problem. For the special case where the desired state of health is neutral ($H^* = 0$), some treatment effect takes place the very next day (b_1 is nonzero), and there is no cost associated with the treatments (i.e., $\beta = 0$), the optimal policy

$$\begin{aligned} T(t) = {} & -[(a_1/b_1)H(t) + (a_2/b_1)H(t-1) \\ & + (b_2/b_1)T(t-1) + (b_3/b_1)T(t-2)] \end{aligned} \quad [3]$$

is the treatment that forces the mean health state to equal the ideal (zero) on the next day. If the cost of treatment is nonzero (i.e., $\beta > 0$), then the solution is a more complex linear function of the previous health states and past treatments (see Bertsekas, 1976; You, 1989).

Dynamic programming is a general-purpose method that can be used to solve for optimal solutions to many dynamic decision tasks. Although the example above employed a linear control task, dynamic programming can also be used to solve many nonlinear control problems (see Bertsekas, 1976). However, for highly complex tasks, dynamic programming may not be practical, and heuristic search methods such as genetic algorithms (Holland, 1994) may be more useful.

The formal task analysis presented above provides a basis for determining factors that may affect human performance on the task (cf., Brehmer, 1992). One factor is the stability of the dynamic system, which for eqn [2] depends on the two coefficients, a_1 and a_2. In particular, this system is stable if the roots of the characteristic equation

$$\lambda^2 - a_1\lambda - a_2 = 0$$

are less than one in magnitude (see Luenberger, 1979). A second factor is the controllability of the system, which depends on the three coefficients, b_1, b_2, and b_3 (see Luenberger, 1979). For example, if the treatment effect is delayed ($b_1 = 0$), then the simple control policy shown in eqn [3] is no longer feasible, and the optimal policy is more a complex linear function of the system coefficients.

Consistent with previous research, You (1989) found that even after extensive experience with the task, subjects frequently lost control of their patients and the average performance of human subjects fell far below optimal performance. However, this is a gross understatement. Sterman (1989) found that when subjects tried to manage a simulated production task, they produced costs 10 times greater than optimal, and their decisions induced costly cycles even though the consumer demand was constant. Brehmer and Allard (1991) found that when subjects tried to fight simulated forest fires, they frequently allowed their headquarters to burn down despite desperate efforts to put the fire out. Kleinmuntz and Thomas (1987) found that when subjects tried to manage their simulated patients' health, they often let their patients die while wasting time waiting for the results of nondiagnostic tests and performed more poorly than a random benchmark.

Alternative Explanations for Human Performance

There are many alternative reasons for this suboptimal performance. By their very nature, dynamic decision tasks entail the coordination of many tightly interrelated psychological processes including causal learning, planning, problem solving, and decision making (cf. Toda, 1962). Six different psychological approaches for understanding human DDM behavior have been proposed, each focusing on one of the component processes.

The first approach was proposed by Rapoport (1975), who suggested that suboptimal performance could be derived from an optimal model either by adding information processing constraints on the planning process, or by including subjective utilities into the objective function of eqn [1]. For example, Rapoport (1966, 1967) found that human performance in his stock purchasing tasks was accurately reproduced by assuming that subjects could only plan a few steps ahead (about three steps), as compared to the optimal model with an unlimited planning horizon. In this case, dynamic programming was useful for providing insights about the constraints on human planning capabilities. As another example, Rapoport et al. (1970) attempted to predict performance in a multistage investment game by assuming that the utility function was a concave function of monetary value. In this case, Rapoport et al. (1970) used dynamic programming to derive an elegant decision rule, which predicted that investments should be independent of the size of the capital accumulated during play. Contrary to this prediction, human decision makers were strongly influenced by this factor, and so in this case, dynamic programming was useful for revealing empirical flaws with the theory.

An alternative approach, proposed by Brehmer (1992) and Sterman (1994), is that suboptimal performance is caused by a misconception of the dynamic system (eqn [2]). In other words, a subject's internal model of the system does not match the true model. In particular, human performers seem to have great difficulty discerning the influence of delayed feedback and understanding the effects of nonlinear terms in the system (Rahmandad et al., 2009). Essentially, subjects solve the problem as if it was a linear system that has only zero lag terms. In the case of eqn [2], the subjective decision policy is simply

$$T(t) = -c_1H(t)$$

where c_1 is estimated from a subject's control decisions. Sterman (1989) and Diehl and Sterman (1993) found that this type of simplified subjective policy described their subjects' behavior very accurately.

Several recent studies have investigated how to improve people's understanding of the dynamic systems in DDM tasks. Kopainsky and Sawicka (2011) used an interactive simulation tool, which allowed participants to explore the dynamic features of a resource management task. They found that the tool allowed decision makers to construct more accurate mental representations of the task, resulting in greater understanding and improved performance. Cronin, Gonzalez, and Sterman (2008) suggested that having participants verbalize their decisions enhanced their performance in an accumulator task. Sterman (2010) reported that training MIT graduate students in system dynamics (e.g., with an introductory system dynamics course) also improved their overall task performance.

A more general method for estimating subjective decision policies was proposed by Jagacinski (Jagacinski and Miller, 1978; Jagacinski and Hah, 1988). Consider once again the example problem employed by You (1989). In this case, a subject's treatment decision on each trial $T(t)$ could be represented by a linear control model:

$$T(t) = c_0 + c_1H(t) + c_2H(t-1) + c_3H(t-2) + c_4T(t) + c_5T(t-1) + c_6T(t-2) + \text{error}$$

where the subjective coefficients ($c_0, c_1, \ldots, c_6$) are estimated by a multiple regression analysis. This is virtually the same as performing a 'lens model' analysis to reveal the decision maker's policy (see Kleinmuntz, 1993; Slovic and Lichtenstein, 1971). This approach has been successfully applied in numerous applications (Jagacinski and Miller, 1978; Jagacinski and Hah, 1988; Kerlik, Miller, Jagacinski, 1993; Kleinman et al., 1980; You, 1989). Indeed, You (1989) found that subjects made use of both lags 1 and 2 for making their treatment decisions.

Heuristic approaches to strategy selection in dynamic decision tasks have been explored by Kleinmuntz (1985), Kleinmuntz and Thomas (1987), and Kerstholt (1994, 1996). These researchers examined health management tasks that entailed dividing time between two strategies: collecting information from diagnostic tests before choosing a treatment, versus treating patients immediately without conducting any more diagnostic tests. A general finding is that even experienced subjects tend to overuse information collection, resulting in poorer performance than that which could be obtained from a pure treatment (no test) strategy. This finding seems to run counter to the adaptive decision-making hypothesis of Payne et al. (1993), which claims that subjects prefer to minimize effort and maximize performance. The information collection strategy is both more effortful and less effective than the treatment strategy in this situation. In another study, Cronin et al. (2008) employed simple stock-flow problems and found that people's inability to understand the dynamics of the task resulted from their use of inappropriate heuristics, which sometimes yielded correct responses, but fundamentally misrepresented the interplay of task variables. Performance remained poor, regardless of the amount of data presented to the participants, the method of presentation, and changes to the cover stories or context.

Finally, an individual difference approach to understanding performance on complex dynamic decision tasks was developed by Dorner and his colleagues (see Funke, 1991 for a review). Subjects are divided into two groups (good versus poor) on the basis of their performance on a complex dynamic decision task. Subsequently, these groups are compared on various behaviors to identify the critical determinants of performance. This research indicates that subjects who perform best are those that set integrative goals, collect systematic information, and evaluate progress toward these goals. Subjects who tend to shift from one specific goal to another, or focus exclusively on only one specific goal, perform more poorly. Along these lines, Bisantz et al. (2000) applied the lens model to DDM tasks and found that much of the differences between the best and worst performers were due to the relative consistency with which top performers made judgments, and not to different judgment policies, nor to characteristics of the judgment environments.

Learning to Control Dynamic Systems

Although human performers remain suboptimal even after extensive task training, almost all past studies reveal systematic learning effects. First, overall performance often rapidly improves with training (see e.g., Brehmer, 1992; Gonzalez and Quesada, 2003; Mackinnon and Wearing, 1985; Rapoport, 1966; Sterman, 1989). Furthermore, subjective policies tend to evolve over trial blocks toward the optimal policy (Jagacinski and Miller, 1978; Jagacinski and Hah, 1988; You, 1989). Therefore, learning processes are important for explaining much of the variance in human performance on dynamic decision tasks (cf., Hagmayer et al., 2010; Hogarth, 1981). Three different frameworks for modeling human learning processes in dynamic decision tasks have been proposed (Osman, 2010).

A production rule model was developed by Anzai (1984) to describe how humans learn to navigate a simulated ship. The general idea is that past and current states of the problem are stored in working memory. Rules for transformation of the physical and mental states are represented as condition–action type production rules. A product rule fires whenever the current state of working memory matches the conditions for a rule. When a production rule fires, it deposits new inferences or information in memory, and it may also produce physical changes in the state of the system. Navigation is achieved by a series of such recognize–act cycles of the production system. Learning occurs by creating new production rules that prevent earlier erroneous actions, and new rules are formed on the basis of means–ends type of analyses for generating subgoal strategies. Simulation results indicated that the model learned strategies similar to those produced in the verbal protocols of subjects, although evaluation of the model was dependent on qualitative judgments rather than quantitative measurements.

An instance- or exemplar-based model was later developed by Dienes and Fahey (1995) to describe how humans learn to control a simulated sugar production task, and to describe how they learn to manage the emotional state of a hypothetical person. This model assumes that whenever an action leads to a successful outcome, the preceding situation and the successful

response are stored together in memory. On any given trial, stored instances are retrieved on the basis of similarity to the current situation, and the associated response is applied to the current situation. This model was compared to a simple rule-based model like that employed by Anzai (1984). The results indicated that the exemplar learning model produced more accurate predictions for delayed feedback systems, but the rule-based model performed better when no feedback delays were involved. This conclusion agrees with earlier ideas presented by Berry and Broadbent (1988) that delayed feedback tasks involve implicit learning process, whereas tasks without delay are based on explicit learning processes.

Gonzalez et al. (2003) proposed the more sophisticated instance-based learning theory (IBLT), which posits that decision makers adapt their judgment strategies from heuristic-based to instance-based as they accumulate and refine their knowledge and learn to recognize situations based on similarity to past instances. This model was used to explain data from a water distribution simulation. In the task, participants managed a water purification plant by shifting water between several tanks. As rain fell some tanks spontaneously filled with water and had to be pumped. IBLT construes the decision maker as learning to control this DDM task through the accumulation and refinement of instance-based knowledge. Participants begin by operating according to simple, domain-general heuristics. However, as decision makers gain experience in the task, they accumulate instance knowledge that can be used to improve performance.

Figure 2 gives a graphical representation of how the model performs a DDM task. The process begins with a search through memory for instances that match the current state. Instances are represented as situation–decision–utility (SDU) triplets, which code environmental cues, the decision made in that environment, and the utility of the experienced outcome. SDUs are retrieved based on the similarity between the current state and each instance in memory. If no sufficiently similar instances are found, a heuristic is used to choose an action. If a similar entry is found in memory, the model must decide whether to search for more alternatives or to execute the current best action. Additional alternatives are evaluated until one exceeds the decision maker's aspiration threshold or time runs out before a decision deadline, at which point the action with the highest utility is chosen. After executing an action, memory is updated. New SDUs are accumulated as more alternatives and more decision situations are confronted. Over time, the model is able to use more and more diagnostic cues to search its growing memory. Feedback on previous actions is used to update the utilities of the SDUs. With experience, the model transitions away from heuristic-based decisions as it adapts to the dynamic environment. Gonzalez et al. (2003) tested several versions of IBLT and found that the full version captured participants' changes in performance as they moved away from random or heuristic-based decisions, toward instance-based strategies. Gonzalez and Dutt (2011) used IBLT to model subjects' behavior in two different versions of a decision-from-experience task. Sampling and repeated-choice paradigms are often thought to rely on differences in psychological mechanisms, but IBLT was able to account for choices in both scenarios using a common set of cognitive mechanisms.

Gibson et al. (1997) used an artificial neural network model to describe learning in a sugar production task. This model receives input representing the current state of the environment and the current goal for the task. Hidden layers compute the next action and predict its consequences. Connections between hidden layers and input and output layers are learned by back propagating prediction errors and deviations between observed outcomes and the goal state. This model provided good accounts of subjects' performance during both training and subsequent generalization tests under novel conditions. Recently, Gibson (2007) extended this model by assuming that decision makers first sort their evidence into categories before weighting and using the sorted evidence. This two-stage model better predicted decision makers' performance in a new, related bargaining task. Unfortunately, no direct comparisons with rule- or exemplar-based learning models were conducted, and this remains a challenge for future research.

Another approach to modeling the learning of dynamic systems is the reinforcement learning model (see Sutton and Barto, 1998, for a review of this approach). The goal of this algorithm is to maximize the future cumulative reward through interacting with the environment. Consequently it does not need to know the model of the environment or the correct input/output pairs (as in supervised learning). Optimization is based on the agent's online samples from the environment.

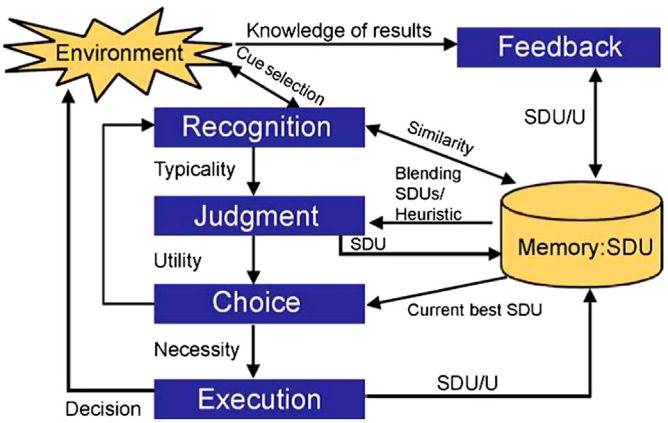

Figure 2 IBLT model diagram.

Recently, Gureckis and Love (2009b) used model-based reinforcement learning to examine human behavior in a 'Farming on Mars' task. Participants' goal was to use two Martian robots to establish oxygen for later human residents. Only one robot could be used at a time, and each performed differently. Participants had to learn about each robot's oxygen output through experience. When rewards were generated probabilistically based on subjects' choices, Gureckis and Love found that decision makers preferred the robot with high short-term output but low long-term output. In the continuous reward condition, where rewards were a deterministic function of subjects' choices, decision makers preferred the robot with the better long-term output but worse short-term output.

In another study, Gureckis and Love (2009a) examined noise effects by implementing noise on both rewards and representations of the current state of task. They found that some moderate level of reward noise caused participants to explore different strategies and helped them find the optimal policy among their alternatives. However, noisy state cues decreased performance relative to no noise conditions. It seems that decision makers use state information as support for their decisions, and so noisy state information is more harmful than noisy reward. Continuing with the 'Farming on Mars' task, Otto and Love (2011) used a reinforcement learning framework to model the impacts of forgone rewards on DDM. They found that providing more information about what could have been gained at each point interfered with subjects' capability to make long-term choices.

Laboratory versus Naturalistic Decision Research

Klein and his associates (Klein et al., 1993; Klein, 2008; Lipshitz et al., 2001; Zsambok and Klein, 1997) have made progress toward understanding dynamic decisions in applied field research settings, which they call naturalistic decision making (e.g., interview fire chiefs after fighting a fire). Field research complements laboratory research in two ways: on the one hand, it provides a reality check on the practical importance of theory developed in the laboratory; on the other hand, it provides new insights that can be tested with more control in the laboratory.

The general findings drawn from naturalistic decision research provide converging support for the general theoretical conclusions obtained from the laboratory. First, it is somewhat of a misnomer to label this kind of research 'decision making' because decision processes comprise only one of the many cognitive processes engaged by these tasks – learning, planning, and problem solving are just as important. Decision making is used to define the overall goal, but then the sequence of actions follows a plan that has either been learned in the past or generated by a problem-solving process. Second, learning processes may explain much of the variance in human performance on dynamic decision tasks. Decision makers use the current goal and current state of the environment to retrieve actions that have worked under similar circumstances in the past. Klein's (1998) recognition-primed decision model is based on this principle, and this same basic idea underlies the production rule, exemplar, and neural network learning models. Third, learning from extensive experience is the key to understanding why novel subjects fail where experts succeed. A few hundred trials in a laboratory task are relatively insignificant in comparison with say 25 years of experience on a job. Naïve subjects may fail to prevent their headquarters from burning down in a simulated forest fire, but expert fire chiefs succeed in saving our national parks from forest fires every year.

See also: Decision and Choice: Heuristics; Decision and Choice: Sequential Decision Making; Learning: Mathematical Learning Theory, History of; Mathematical Psychology; Optimal Control Theory; Problem Solving and Reasoning, Psychology of.

Bibliography

Anzai, Y., 1984. Cognitive control of real-time event driven systems. Cognitive Science 8, 221–254.

Berry, D.C., Broadbent, D.E., 1988. Interactive tasks and the implicit-explicit distinction. British Journal of Psychology 79, 251–272.

Bertsekas, D.P., 1976. Dynamic Programming and Stochastic Control. Academic Press, N.Y.

Bisantz, A.M., Kirlik, A., Gay, P., Phipps, D.A., Walker, N., Fisk, A.D., 2000. Modeling and analysis of a dynamic judgment task using a Lens model approach. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans 6, 605–616.

Brehmer, B., 1992. Dynamic decision making: human control of complex systems. Acta Psychologica 81, 211–241.

Brehmer, B., Allard, R., 1991. Real-time dynamic decision making: effects of task complexity and feedback delays. In: Rasmussen, J., Brehmer, B., Leplat, J. (Eds.), Distributed Decision Making; Cognitive Models for Cooperative Work. Wiley, Chichester.

Brehmer, B., Nahlinder, S., 2007. Achieving what cannot be done: coping with the time constants in a dynamic decision task by doing something else. Scandinavian Journal of Psychology 48, 359–365.

Cronin, M., Gonzalez, C., Sterman, J.D., 2008. Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. Organizational Behavior and Human Decision Processes 101, 116–130.

Diederich, A., 2001. Decision and Choice: Sequential Decision Making. International Encyclopedia of the Social and Behavioral Sciences: Methodology, Mathematics, and Computer Science. Pergamon, Amsterdam.

Diehl, E., Sterman, J.D., 1993. Effects of feedback complexity on dynamic decision making. Organizational Behavior and Human Decision Processes 62, 198–215.

Dienes, Z., Fahey, R., 1995. Role of specific instances in controlling a dynamic system. Journal of Experimental Psychology: Learning, Memory, & Cognition 21, 848–862.

Dorner, D., 1980. On the problems people have in dealing with complexity. Simulation and Games 11, 87–106.

Ebert, R.J., 1972. Human control of a two-variable decision system. Organizational Behavior and Human Performance 7, 237–264.

Edwards, W., 1962. Dynamic decision theory and probabilistic information processing. Human Factors 4, 59–73.

Funke, J., 1991. Solving complex problems: exploration and control of complex systems. In: Sternberg, R.J., Frensch, P.A. (Eds.), Complex Problem Solving: Principles and Mechanisms. Erlbaum, Hillside, NJ.

Gibson, F., Fichman, M., Plaut, D.C., 1997. Learning in dynamic decision tasks: computational model and empirical evidence. Organizational Behavior and Human Performance 71, 1–35.

Gibson, F., 2007. Learning and transfer in dynamic decision environments. Computational and Mathematical Organization Theory 13, 39–61.

Gonzalez, C., Dutt, V., 2011. A generic dynamic control task for behavioral research and education. Computers in Human Behavior 27, 1904–1914.

Gonzalez, C., Lerch, F.J., Lebiere, C., 2003. Instance-based learning in dynamic decision making. Cognitive Science 27, 591–635.

Gonzalez, C., Quesada, J., 2003. Learning in dynamic decision making: the recognition process. Computational and Mathematical Organization Theory 9, 287–304.

Gureckis, T.M., Love, B.C., 2009a. Learning in noise: dynamic decision making in a variable environment. Journal of Mathematical Psychology 53, 180–193.
Gureckis, T.M., Love, B.C., 2009b. Short-term gains, long-term pains: how cues about state aid learning in dynamic environments. Cognition 113, 293–313.
Hagmayer, Y., Meder, B., Osman, M., Mangold, S., Lagnado, D., 2010. Spontaneous causal learning while controlling a dynamic system. The Open Psychology Journal 3, 145–162.
Hogarth, R.M., 1981. Beyond discrete biases: functional and dysfunctional aspects of judgmental heuristics. Psychological Bulletin 90, 197–217.
Holland, J., 1994. Adaptation in Natural and Artificial Systems. MIT Press, Cambridge.
Jagacinski, R.J., Flach, J.M., 2003. Control Theory for Humans: Quantitative Approaches to Modeling Performance. Erlbaum, Mahwah: NJ.
Jagacinski, R.J., Hah, S., 1988. Progression-regression effects in tracking repeated patterns. Journal of Experimental Psychology: Human Perception and Performance 14, 77–88.
Jagacinski, R.J., Miller, R.A., 1978. Describing the human operator's internal model of a dynamic system. Human Factors 20, 425–433.
Jensena, E., Brehmer, B., 2003. Understanding and control of a simple dynamic system. System Dynamic Review 19, 119–137.
Kerstholt, J.H., 1994. The effect of time pressure on decision making behavior in a dynamic task environment. Acta Psychologica 86, 89–104.
Kerstholt, J.H., 1996. The effects of information costs on strategy selection in dynamic tasks. Acta Psychologica 94, 273–290.
Kerstholt, J.H., Raaijmakers, J.G.W., 1997. Decision making in dynamic task environments. In: Crozier, W.R., Svenson, O. (Eds.), Decision Making: Cognitive Models and Explanations. Routledge, London, pp. 205–217.
Kirlik, A., Miller, R.A., Jagacinski, R.J., 1993a. Supervisory control in a dynamic and uncertain environment: process model of skilled human-environment interaction. IEEE Transactions on Systems, Man, and Cybernetics 23, 929–952.
Kirlik, A., Plamondon, D.D., Lytton, L., Jagacinski, R.J., 1993b. Supervisory control in a dynamic and uncertain environment: laboratory task and crew performance. IEEE Transactions on Systems, Man, and Cybernetics 23, 1130–1138.
Klein, G., 1998. Sources of Power: How People Make Decisions. MIT Press, Cambridge.
Klein, G., Orasanu, J., Calderwood, R., Zsambok, C.E., 1993. Decision Making in Action: Models and Methods. Ablex, Norwood, NJ.
Klein, G., 2008. Naturalistic decision making. Human Factors: The Journal of the Human Factors and Ergonomics Society 50, 456–460.
Kleinman, D.L., Pattipati, K.R., Ephrath, A.R., 1980. Quantifying an internal model of target motion in a manual tracking task. IEEE Transactions on Systems, Man, and Cybernetics 10, 624–636.
Kleinmuntz, D., 1985. Cognitive heuristics and feedback in a dynamic decision environment. Management Science 31, 680–702.
Kleinmuntz, D., 1993. Information processing and misperceptions of the implications of feedback in dynamic decision making. System Dynamics Review 9, 223–237.
Kleinmuntz, D., Thomas, J., 1987. The value of action and inference in dynamic decision making. Organizational Behavior and Human Decision Processes 39, 341–364.
Kleiter, G.D., 1975. Dynamic decision behavior: comments on Rapoport's paper. In: Wendt, D., Vlek, C. (Eds.), Utility, Probability, and Human Decision Making. Reidel, Dordrecht-Holland, pp. 371–380.
Kopainsky, B., Sawicka, A., 2011. Simulator-supported descriptions of complex dynamic problems: experimental results on task performance and system understanding. System Dynamics Review 27, 142–172.
Lipshitz, R., Klein, G., Orasanu, J., Salas, E., 2001. Taking stock of naturalistic decision making. Journal of Behavioral Decision Making 14, 332–351.
Luenberger, D.G., 1979. Introduction to Dynamic Systems. Wiley, NY.
Mackinnon, A.J., Wearing, A.J., 1980. Complexity and decision making. Behavioral Science 25, 285–296.
Mackinnon, A.J., Wearing, A.J., 1985. Systems analysis and dynamic decision making. Acta Psychologica 58, 159–172.
Osman, M., 2010. Controlling uncertainty: a review of human behavior in complex dynamic environments. Psychological Bulletin 136, 65–86.
Otto, A.R., Love, B.C., 2011. You don't want to know what you're missing: when information about forgone rewards impedes dynamic decision making. Judgment and Decision Making 5, 1–10.
Payne, J.W., Bettman, J.R., Johnson, E.J., 1993. The Adaptive Decision Maker. Cambridge University Press, NY.
Rahmandad, H., Repenning, N., Sterman, J., 2009. Effects of feedback delay on learning. System Dynamics Review 25, 309–338.
Rapoport, A., 1966. A study of human control in a stochastic multistage decision task. Behavioral Science 11, 18–32.
Rapoport, A., 1967. Dynamic programming models for multistage decision making. Journal of Mathematical Psychology 4, 48–71.
Rapoport, A., 1975. Research paradigms for studying dynamic decision behavior. In: Wendt, D., Vlek, C. (Eds.), Utility, Probability, and Human Decision Making. Reidel, Dordrecht-Holland, pp. 347–369.
Rapoport, A., Jones, L.V., Kahan, J.P., 1970. Gambling behavior in multiple-choice multi-stage betting games. Journal of Mathematical Psychology 7, 12–36.
Rouse, W.B., 1980. Systems Engineering Models of Human–Machine Interaction. North-Holland, NY.
Slovic, P., Lichtenstein, S., 1971. Comparison of Bayesian and regression approaches to the study of information processing in judgment. Organizational Behavior and Human Peformance 6, 649–744.
Sterman, J.D., 1989. Misperceptions of feedback in dynamic decision making. Organizational Behavior and Human Decision Processes 43, 301–335.
Sterman, J.D., 1994. Learning in and about complex systems. System Dynamics Review 10, 291–330.
Sterman, J.D., 2010. Does formal system dynamics training improve people's understanding of accumulation? System Dynamics Review 26, 316–334.
Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning. MIT Press, Cambridge, MA.
Sweeney, L.B., Sterman, J.D., 2000. Bathtub dynamics: initial results of a systems thinking inventory. System Dynamic Review 16, 249–286.
Toda, M., 1962. The design of the fungus eater: a model of human behavior in an unsophisticated environment. Behavioral Science 7, 164–183.
You, 1989. Disclosing the decision-maker's internal model and control policy in a dynamic decision task using a system control paradigm. Unpublished MA thesis, Purdue University.
Zsambok, C.E., Klein, G., 1997. Naturalistic Decision Making. Erlbaum, Mahwah, NJ.

Dynamic Mapping in Geography

Doris Dransch and Andrea Unger, Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, Potsdam, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Dransch, volume 6, pp. 3908–3911, © 2001, Elsevier Ltd.

Abstract

Dynamic mapping summarizes various cartographic presentation methods that incorporate the dimension of time in a map. The methods can be differentiated according to the way they encode the dimension of time: time is either presented by visual encodings in a map, by a spatial dimension, by time, or by a combination of multiple views.

Dynamic mapping is a cartographic concept used to depict dynamic spatial phenomena such as change of land use, spread of diseases, or bird migration. It summarizes various visualization methods that incorporate the dimension of time into a map.

These methods utilize different strategies: temporal information is graphically encoded within the map, time is mapped on an additional spatial dimension, time is presented by animation, or a combination of multiple views is used. Each of these methods has unique individual characteristics and thus is suitable for specific tasks (Figure 1).

The Term and Concept of Dynamic Mapping

Dynamic mapping is a visualization concept that is involved with the depiction of spatial processes on maps. Our world is constantly changing. If changes are of little interest, it is sufficient to present a specific state of our world and of the corresponding date. Yet, if occurring changes are of great importance, the dynamic aspect of reality – the dimension of time – has to be depicted in maps. Various presentation methods have been developed to give an impression of the changes and movements in space (Monmonier, 1990; Muehrcke, 1992). The methods can be differentiated into three main groups according to the way they encode the dimension of time: time is either presented by visual encodings in a map, by a spatial dimension, or by time. If dynamic mapping is related to a computer device, interactive visualization can be applied that facilitates multiple views on the presented changes and processes (Andrienko et al., 2010).

Presenting Time by Visual Encodings in a Map

The most common way to visualize dynamic spatial phenomena is to represent them by visual encodings in a map. Temperature curves of several measurement stations, maps of

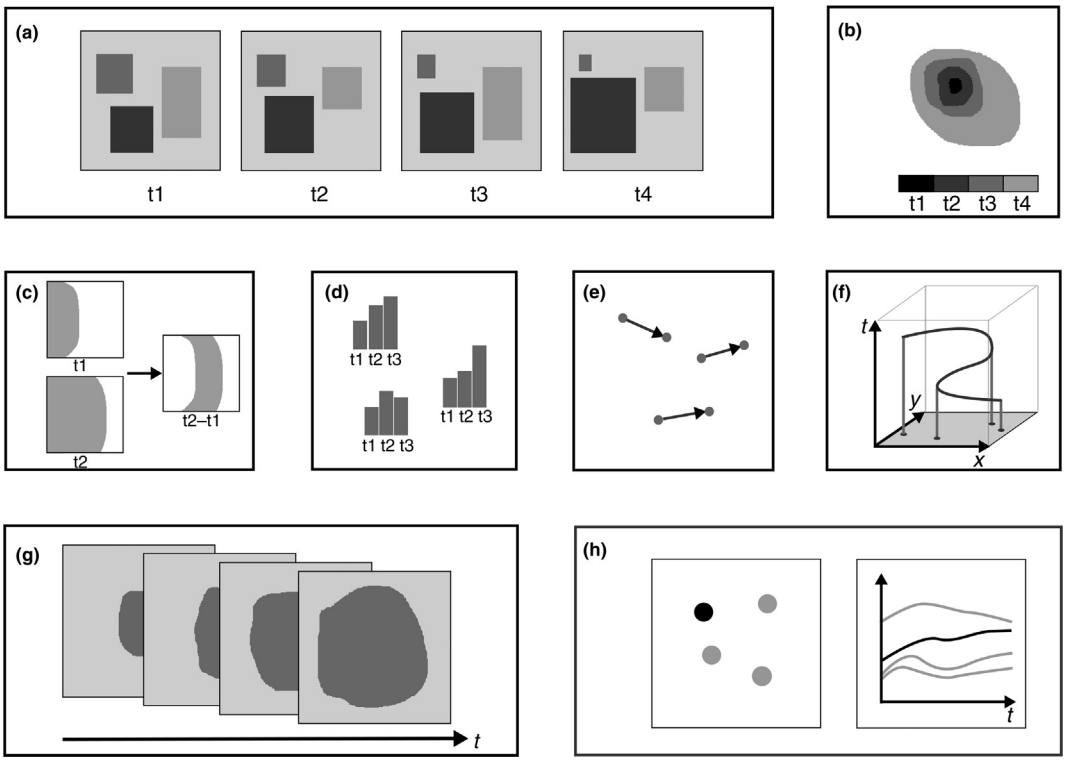

Figure 1 Methods to depict spatial processes.

urban growth, or maps of land use change are well known examples. In general, the following types of maps can be differentiated, each applying a different strategy to communicate dynamic aspects: time-series maps, time-composite maps, change maps, map with time-related charts, and vector or flow maps.

Time-series maps, also called small multiples, show one spatial area for different points in time in a series of maps (Figure 1(a)). The user has to compare the individual maps to understand changes between single states. To facilitate this task, time-series maps have to be arranged in a way that enables visual comparison. They have to be placed next to each other, and their content should not be too complex. Map series are commonly applied to show the distribution of particular phenomena at given intervals, e.g., changes of population distribution, boundaries, or landuse, etc.

Another method is the time-composite map (Figure 1(b)). Such a map depicts two or more stages of a dynamic phenomenon in one single map. The distinct stages can be differentiated by color, pattern, isolines, or point symbols. Time-composite maps are suitable for showing the expansion or contraction of areas, linear networks, or other dynamic phenomena, e.g., the growth of a city, the development of transportation systems, or the spread of diseases. Time-composite maps succeed in presenting one-directional changes, but they fail if retrogressive developments have to be depicted as an intermediate stage.

Time series and time-composite maps show selected stages of the entire development. The change itself has to be understood by interpretation. However, some geographical questions emphasize, above all, the change and its extent. These questions require a map, which shows explicitly the change that has taken place within a certain time period. For that purpose the change map was developed (Figure 1(c)). Qualitative change maps depict objects that have been added to or removed from a particular area during a relevant time period, e.g., the difference in forest cover in different years. Quantitative change maps present changes in magnitude, the change in a value such as population density or air pollution. They either show the direction of change as increase or decrease, or give the absolute values of magnitude change.

A further method of presenting changes directly is a map with time-related charts (Figure 1(d)) that depicts dynamic phenomena of areas or locations in line or bar graphs, e.g., maps with diagrams of the annual temperature curve for selected stations. This type of map is able to give a detailed impression of development in time, but it is delimited in its ability to depict the spatial distribution of a changing phenomenon.

Geographical processes may take place at a particular location (similar to the mentioned changes of population or temperature) or they may be a sort of movement in space like the movement of people, goods, or winds. To show these spatial movements, vector or flow maps with special symbols such as arrows or lines are applied (Figure 1(e)). Position and alignment of arrows and lines can depict the moving objects' routes and directions, and the size of the signatures can denote the objects' values. Typical examples of vector or flow maps are maps of expedition routes, commuter movements, or ocean currents.

Presenting Time in a Space–Time Cube

The space–time cube is yet another method of visualizing spatial changes over time (Kraak, 2008). It originates from space–time concepts that were developed in the 1960s. This approach maps space and time onto the three axes of a cube; the x- and y-axes represent the geographical dimension; time is mapped to the third dimension (Figure 1(f)). The result is a three-dimensional presentation of a space–time continuum. Space–time cubes are increasingly applied to visualize trajectories, movement lines created for instance by people, animals, or the spreading of diseases. As time is mapped onto a quantitative axis, time duration to cover a distance becomes apparent; the cube unveils movement patterns in space and time. A further advantage of the three-dimensional cube is to facilitate presenting even back and forth movements without graphical clutter. Although presentation of trajectories is the main application of the space–time cube, it is a proper visualization method for all geographical phenomena changing in space and time.

Presenting Time by Time: Cartographic Film and Animation as Dynamic Maps

All presentation methods referred to above are able to produce an idea regarding the changes that take place, but they do not provide an intuitive depiction of the dynamics of a process. To overcome this shortcoming cartographic films or animations were introduced. They are sequences of slightly varying maps that create an illusion of continuous spatial movement or change (Figure 1(g)). By mapping real time to presentation time they vividly show the dynamics of a process. They also allow for making apparent time intervals of either extremely long or short duration utilizing fast or slow motion. Early cartographic films were produced for military propaganda to depict frontier changes during World War II, or as instruction material for students; each map of the sequence was drawn by a cartographer. Because of their time-consuming and expensive production, hand-made cartographic films did not become a widespread presentation form. This has changed with the advent of powerful computer technology, which has greatly facilitated the creation of animated films. Computer animation can further be enhanced with user interaction. At a basic level, the user is enabled to start, stop, interrupt as well as to delay or accelerate an animation. At a higher level, more sophisticated interactions allow for selecting particular time intervals or determining different viewpoints (Dransch, 1997).

Multiple Views

As an alternative to the above-discussed visualization methods, spatial and temporal aspects can be separated into multiple views. Usually a map is combined with a presentation of temporal data in a time-related chart, such as a time-value plot or a bar chart (Figure 1(h)). When combining such distinct visual representations, it is important to convey the linkage between data in the map and data in the temporal display. Visual linking, for example by using consistent colors or labels

in both presentations, can fulfill this purpose. Yet, multiple views unfold their strength in interactive environments. Here, they can be combined with interaction methods such as brushing and linking, which highlights selected data in both views simultaneously. Additionally, it is possible to follow an overview and detailed concept. As an example, the map might serve as an overview of all locations, whereas a time-related chart might show details about dynamic developments for selected locations. The other way around, a time-related chart can provide an overview of the dynamics of a phenomenon, and the map may depict the spatial distribution at a selected time point.

In this sense, the concept of multiply coordinated views is highly flexible and allows depicting both spatial and temporal developments in detail. The ability to analyze data step by step in high detail is beneficial when the amount of data is too large to be shown at once and the display of few temporal stages is not sufficient. Exemplary of such data are temperature curves over a long time period, which need to be analyzed on a daily basis rather than by monthly or yearly aggregates.

See also: Cartographic Visualization; Cartography; Cognitive Maps; Geographic Information Systems; Tactile Maps in Geography; Thematic Maps in Geography; Time-Space in Geography.

Bibliography

Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Fabrikant, S.I., Jern, M., Kraak, M.J., Schumann, H., Tominski, 2010. Space, time and visual analytics. International Journal of Geographical Information Science 24 (10), 1577–1600.

Dransch, D., 1997. Computer-Animation in der Kartographie. Theorie und Praxis (Computer Animation in Cartography: Theory and Practice). Springer, Heidelberg, Germany.

Kraak, M.J., 2008. Geovisualization and time: new opportunities for the space–time cube. In: Dodge, M., McDerby, M., Turner, M. (Eds.), Geographic Visualization: Concepts, Tools and Applications. Wiley & Sons, Chichester, pp. 293–306.

Monmonier, M., 1990. Strategies for the visualization of geographic time-series data. Cartographica 27, 30–45.

Muehrcke, P.C., 1992. Map Use: Reading, Analysis and Interpretation. J.P. Publications, Madison, WI.

Dysexecutive Amnesia

Sara B Festini and Patricia A Reuter-Lorenz, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Memory impairments can result from frontal lobe damage, and because the frontal lobes perform executive functions, these impairments have been referred to as dysexecutive amnesia. Memory disturbances most commonly associated with dysexecutive amnesia include a disproportionate impairment in recall, deficient temporal ordering, source amnesia, and impaired metamemory. Components of dysexecutive amnesia are often observed in patients with Korsakoff's amnesia, vascular thalamic amnesia, anterior communicating artery aneurysms, and in memory impairments associated with schizophrenia. We summarize evidence that the frontal lobes play a role in normal memory and that damage to the frontal lobes can cause dysexecutive memory impairments.

Introduction

Amnesia refers to memory loss and is most commonly associated with damage to the hippocampus and the medial temporal lobes. However, amnesic impairments also accompany damage to other brain structures, including the frontal lobes, the thalamus, and pathways that interact with the hippocampus (i.e., the mammillothalamic tract). We apply the term 'dysexecutive amnesia' to describe memory impairments associated with frontal dysfunction and those that therefore have a strong executive component, while noting that the literature generally lacks a consistent terminology to refer to these impairments. We also focus on particular disorders that are associated with dysexecutive memory impairments in an effort to summarize and characterize dysexecutive amnestic disturbances.

Amnesia can affect memory for past events or new learning. The inability to remember past events is termed retrograde amnesia, whereas the inability to form new memories is termed anterograde amnesia. Amnesia tends to affect the declarative memory system or information that is available to conscious awareness. In contrast, amnesics tend to have preserved nondeclarative (implicit, procedural) memory, which is less available to consciousness. Spared implicit memory is evident in the ability of amnesics to perform previously learned procedural skills and to acquire new ones even though they have no recollection of the learning episodes. As noted, amnestic disorders most commonly originate from disruption of medial temporal lobe function. However, given the numerous contributions of the frontal lobes to memory, frontal lobe dysfunction can also lead to deficits of memory.

The frontal lobes are involved in higher-level cognitive processes, such as planning, sequencing, monitoring, inhibiting, resolving interference, and maintaining task goals. These functions are commonly termed cognitive control or executive control and are implemented by various regions of the prefrontal cortex. Executive control processes interact with the medial temporal lobe system to promote efficient and strategically controlled memorial processing (Moscovitch, 1992; Simons and Spiers, 2003; Dickerson and Eichenbaum, 2010). For example, executive control allows for self-initiated top-down control of encoding processes based on one's current task goals. Moreover, the frontal lobes assist with memory retrieval, by elaborating on encountered retrieval cues, for example. Additionally, the frontal lobes monitor the products of retrieval operations and determine if the memories satisfy relevant criteria. Thus, the frontal lobes contribute to the encoding, retrieval, and monitoring of long-term memories to promote successful veridical memory.

In addition to supporting long-term memory, the dorsolateral prefrontal cortex is involved in working memory, which allows for the short-term storage and manipulation of information held in mind. Numerous studies have shown that the dorsolateral prefrontal cortex is important for accurate working memory. The frontal lobes also enable the assessment of current and future memory performance (e.g., Shimamura et al., 1991), and this awareness of one's own memory is termed metamemory. Accurate metamemory allows for the efficient prediction of what will be remembered in the future and also enables the implementation of control processes to enhance memory. For instance, people can monitor if they will have difficulty remembering what they are encoding and can strategically decide to devote more time to studying that information.

The frontal lobes are interconnected with the medial temporal lobe system, both structurally and functionally. In particular, the mammillothalamic pathway connects the frontal cortex and the hippocampi. Functional associations between the frontal and medial temporal lobe regions during memory encoding and retrieval have been well documented with neuroimaging methods, which indicate the importance of frontal control processes in successful memory performance.

Dysexecutive Amnesia

Patients with frontal lobe damage frequently exhibit amnestic disturbances, which we describe in detail in the following section. However, note that, although memory impairments are common, restricted frontal damage does not produce the extensive amnestic syndromes that characterize temporal lobe lesions. Furthermore, although the term dysexecutive amnesia is not yet widely used, the term is applied here to characterize memory deficits associated with frontal damage or to describe

 http://dx.doi.org/10.1016/B978-0-08-097086-8.51029-0

memory disturbances in patients with other frontal-like symptoms. As such, dysexecutive amnesia is synonymous with frontal amnesia, frontal memory impairment, and dysexecutive memory impairment. Dysexecutive amnesia is also related to dysexecutive syndrome, which encompasses all executive impairments associated with frontal damage. Baddeley and Wilson (1988) explain that frontal amnesia may be a combination of classic amnesia coupled with dysexecutive symptomatology. Thus, dysexecutive amnesia specifically refers to the memory impairments associated with frontal pathology.

Common Dysexecutive Amnesia Memory Impairments

Impaired Recall

Patients with frontal lobe damage typically display anterograde amnesia, with a disproportionate impairment in free recall and relatively spared recognition memory (e.g., Jetter et al., 1986; Hanley et al., 1994; Eichenbaum et al., 2010; see Swick and Knight, 1999 for spared recognition in frontal patients). For instance, Jetter et al. (1986) tested patients with left, right, or bilateral frontal damage on free recall, cued recall, and recognition, and compared their performance to a nonfrontal patient control group. The frontal group displayed severely impaired free recall, which was strikingly low compared to their spared cued recall and spared recognition. The disproportionate impairment in free recall compared to cued recall and recognition suggests that the patients successfully encoded the information, but that they had particular difficulty self-initiating retrieval of these memoranda at test. Numerous studies replicate this disproportionate impairment in recall. Nevertheless, there are reported case studies of frontal patients with impaired recognition and spared recall (Delbecq-Derouesné et al., 1990), impaired recognition memory (Baddeley and Wilson, 1988), or no discrepancy between recall and recognition (Kopelman and Stanhope, 1998). Additionally, Shimamura et al. (1991) report that, compared to healthy controls, frontal patients were unimpaired on several tests of new learning, including paired-associate recall, story recall, and recall of the Rey–Osterrieth Complex Figure, although recall of unrelated word lists was disproportionately impaired.

Impaired free recall of previously learned information in dysexecutive amnesia is attributed to deficits associated with frontal lobe contributions to the strategic retrieval of information and to the utilization of self-generated retrieval cues (Moscovitch, 1992; Simons and Spiers, 2003). Moreover, executive deficits likely also contribute to poorer encoding strategies during new learning, which yield weaker memory representations (Moscovitch, 1992; Simons and Spiers, 2003). Furthermore, there is evidence that frontal patients have difficulty with both new learning and retrieving past events. D'Esposito et al. (1996) report that patients with frontal lesions had more severe retrograde amnesia than patients with basal forebrain lesions. Moreover, Della Sala et al. (1993) found that patients with prefrontal lesions exhibited deficient autobiographical memory retrieval, and that this impairment was correlated with performance on tasks tapping executive function. These results suggest that frontal patients have difficulty efficiently implementing executive control strategies to strategically encode and retrieve new or stored memories.

Temporal Discrimination Impairment

Frontal lobe patients also have difficulty keeping track of the temporal order of events (e.g., Huppert and Piercy, 1976; Milner et al., 1991; Shimamura et al., 1990). For instance, Milner et al. (1991) had frontal patients and healthy controls perform recency discriminations in a serial list-learning task that required participants to report which of the two simultaneously presented items they had seen most recently. Patients with damage to the midlateral frontal cortex were especially impaired on these temporal order judgments. Furthermore, Shimamura et al. (1990) asked frontal patients to reproduce the list of 15 words they had studied, maintaining the proper temporal order. Frontal patients were significantly impaired at arranging these words correctly compared to controls. These are just a few of the many studies indicating that memory for temporal order is consistently impaired with frontal damage.

Source Amnesia

Source amnesia also tends to accompany frontal lobe damage, as it is difficult for frontal patients to determine the context in which they initially encountered information. Janowsky et al. (1989b) tested frontal patients, older age-matched controls, and younger adult controls on memory for facts and memory of the sources of these facts. Despite comparable recall and recognition of the facts, frontal lobe patients made significantly more source errors than both control groups after a delay of 6–8 days. During shorter delays of either 5 min or 2 h, both frontal patients and older adult controls made more source errors than younger adults, which supports evidence of frontal decline in normal aging. Furthermore, Schacter et al. (1984) found that the number of source memory errors that their patients exhibited was correlated with poor performance on tasks sensitive to frontal lobe function. Neuroimaging studies also link source memory with the frontal lobes. This evidence along with the high incidence of source memory errors in frontal patients and older adults suggests a role of the frontal lobes in contextual source memory. As Janowsky et al. (1989b) explain, frontal lobe patients' impaired temporal ordering may contribute to their source memory deficit, as source memory also necessitates retrieval of temporal context.

Confabulation

Confabulation occurs when individuals make up stories or explanations to account for gaps in their memory. Confabulation is often referred to as 'honest lying' because the individuals genuinely believe the tales they are spinning. Confabulation is associated with damage to the ventromedial frontal lobes, and it most commonly involves inaccurate accounts of episodic memory (Schacter and Curran, 1995; Turner et al., 2008). Confabulation may be due to poor

strategic retrieval and monitoring processes, such that frontal patients may have difficulty searching memory, retrieving pertinent information, and determining if the memory they retrieved is applicable to the current situation or not (Moscovitch, 1992; Schacter and Curran, 1995). Importantly, Baddeley and Wilson (1988) state that confabulation differentiates frontal amnestic patients from nonfrontal amnestic patients. Thus, confabulation seems to be a unique symptom of dysexecutive amnesia.

Impaired Working Memory

The frontal lobes are also associated with working memory, or actively holding and manipulating information in mind over a short delay. Although amnesia typically describes deficits in long-term memory, impairments in working memory can occur, and in the case of dysexecutive amnesia, deficient working memory is often observed. Frontal patients often have difficulty performing delayed-response tasks in which they have to maintain information in mind over a short delay before making a response. D'Esposito and Postle (2002) report that frontal patients were impaired compared to controls on three of nine delayed-response experiments without distraction and four of six experiments with distraction. Furthermore, in an early study Jacobsen (1936) demonstrated that monkeys with bilateral frontal lesions had impaired delayed-response task performance, such that they were not able to maintain the location of a food reward in mind after a screen occluded the cups. Additionally, single-cell recordings from neurons in the prefrontal cortex of monkeys revealed cells that maintained location-specific information over a delay (Funahashi et al., 1990). These single-cell recordings provide evidence that neurons within the prefrontal cortex are responsible for maintaining location-specific information over a delay period. Critically, however, frontal patients typically display spared simple memory span (e.g., Baddeley and Wilson, 1988; D'Esposito and Postle, 2002; but see Shimamura et al., 1991 for evidence of impaired digit span in frontal patients). Dysexecutive amnesia, thus, can involve impairments in working memory, but does not always affect simple short-term memory span.

Metamemory Impairment

Individuals with frontal amnesia not only have memory impairments, but they also often have metamemory problems, such that they are unaware of the extent of their memory disturbances and are unable to accurately predict their future memory performance (e.g., Jacobson and Lishman, 1990). For instance, even when cued recall levels were matched between frontal amnesics and temporal amnesics, frontal patients exhibited larger discrepancies between their feeling of knowing ratings and their true memory performance than did temporal amnesics (Janowsky et al., 1989a). Moreover, Kopelman et al. (1998) compared subjective evaluations of memory performance for patients with frontal, diencephalic, and temporal lesions, and they found that both frontal and diencephalic patients exhibited impaired metamemory. Thus, unlike temporal amnesics, frontal and diencephalic amnesics had poor awareness of their memory problems. These metamemory impairments may contribute to the memory disturbances that are observed in frontal patients because a lack of insight into the severity of their memory impairments can preclude the engagement of compensatory strategies to mitigate their memory deficits.

Susceptibility to Proactive Interference

Dysexecutive amnesics also tend to be especially susceptible to proactive interference and do not show the typical release from proactive interference when the category of memoranda is changed. Moscovitch (1982) found that patients with a left frontal lobectomy had the smallest release from proactive interference, and that Korsakoff's patients had similarly impaired release from proactive interference (see also Squire, 1982; Shimamura et al., 1991). Strikingly, in contrast, amnesics with left temporal and hippocampal damage displayed a normal release from proactive interference. Moscovitch cites this differing susceptibility to proactive interference as a fundamental characteristic that distinguishes frontal amnesics from temporal/hippocampal amnesics. Note, however, that Shimamura et al. (1991) found a normal release from proactive interference in their sample of frontal patients.

Increased False Alarms

Among the more commonly reported deficits associated with frontal damage is an increased false alarm rate. Patients with frontal lesions tend to produce many intrusions in free recall tests and tend to have high false alarm rates in recognition memory tests (e.g., Schacter and Curran, 1995; Parkin et al., 1996; Swick and Knight, 1999; Eichenbaum et al., 2010). Recently, Eichenbaum et al. (2010) report that prefrontal damage results in deficient recollection that manifests partly in disproportionately large rates of false alarms. In an earlier case study, Schacter and Curran (1995) report that patient J.B., who suffered frontal lobe damage following the rupture of an anterior communicating artery aneurysm, had high false recognition rates, as well as high confidence in this false recognition. Poor executive monitoring processes may contribute to these high intrusion rates. Additionally, impairments in source monitoring may preclude frontal patients from considering why something is familiar to them, thus causing them to mistake semantic relatedness for prior list inclusion, for example.

Perseveration

Perseveration, a common sequela of frontal lobe damage, entails the repetition of responses that are no longer relevant. Perseveration can be observed on the Wisconsin Card Sorting Test when patients erroneously sort cards by a previously relevant rule rather than adjusting their performance to the new feedback. Although perseveration is more classically an executive impairment, perseveration can also affect memory performance when previously relevant information erroneously intrudes during recall. Patients with frontal damage often exhibit perseveration (Baddeley and Wilson, 1988; Shimamura

et al., 1991; Parkin and Leng, 1993; Rolls et al., 1994), and this dysexecutive impairment can interfere with accurate memory performance.

Impaired Frequency Judgments

Related to their impairment in source memory and temporal memory, individuals with frontal damage are also deficient in frequency judgments. Stanhope et al. (1998) had frontal, temporal, and diencephalic lesion groups judge the frequency of occurrence of abstract patterns, and found that all three groups produced impaired estimates. These results suggest that, although not solely attributable to dysexecutive amnesia, individuals with damage to the frontal lobes have difficulty remembering the frequency with which they encountered something in the environment. The authors posit that frontal patients struggled with frequency judgments because they had difficulty monitoring their memory for the number of times they encountered that particular stimulus, which highlights again how an executive control deficit could contribute to poor memory performance.

Impaired Emotion-Based Learning

The impairments discussed thus far are primarily associated with damage to the lateral prefrontal cortex, regions thought to underlie more classically cognitive executive functions. However, damage to the orbitofrontal prefrontal cortex tends to be associated with socially inappropriate behaviors and impaired emotion-related learning. This was documented, for example, by Rolls et al. (1994), who found that patients with ventral frontal lesions displayed perseverative errors and continued responding toward the previously rewarded stimulus. Furthermore, the relative success or failure at learning reward mappings on these tasks was correlated with the degree of the socially inappropriate behavior that was displayed, which may be indicative of the extent of orbitofrontal damage. Ventromedial prefrontal lesions have also been associated with impaired learning on the Iowa Gambling Task (e.g., Bechara et al., 1996). In these tasks, unlike healthy control groups, patients with prefrontal lesions are not able to learn to avoid drawing cards from the risky deck that is associated with long-term monetary losses. Prefrontal patients also fail to generate anticipatory galvanic skin responses that control subjects exhibit when considering a risky deck selection. Thus, emotion-related memory impairments can be observed in cases involving lesions of ventromedial/orbitofrontal cortex.

Remedies

Although the memory disturbances associated with dysexecutive amnesia are vast, certain efforts can be taken to improve memory in these patients. As discussed previously, frontal patients have difficulty planning and generating strategies. Thus, external strategies and cues when provided for these patients have proven effective to improve encoding and to facilitate retrieval. We describe several successful interventions in the following paragraphs.

To start, although dysexecutive amnesics typically display impaired free recall, Jetter et al. (1986) found that frontal patients performed normally on cued recall. Simply providing categorical cues about the studied memoranda aided memory retrieval for these frontal patients. These results indicate that by providing sufficient external retrieval cues, frontal amnesics may be better able to recall what they have learned. Furthermore, Baddeley and Wilson (1988) found that providing detailed retrieval instructions improved word fluency in a frontal patient case study. Although this patient had extreme difficulty with word fluency when given the standard instructions (i.e., name as many animals as you can in a minute), being unable to produce a single word, with more specific instructions (i.e., name as many animals as you can that begin with the letter 'c' in a minute), this patient's performance dramatically improved. These studies suggest that detailed instructions with more specific retrieval cues can constrain the response options enabling frontal patients to overcome their difficulty with self-generated retrieval, and Thompson-Schill et al. (1997) suggest that such cues overcome the increased selection interference that can plague some frontal patients.

Likewise, because dysexecutive amnesics have difficulty generating their own strategy when confronted with a new task, their performance can be improved when strategy instruction is provided. Parkin et al. (1988) demonstrated this by instructing frontal lobe patients to use a mental imagery strategy during a paired associate-learning task, which then enhanced their memory performance. Similarly, Diamond et al. (1997) found that providing frontal patients with an organizational strategy at encoding improved their ability to reproduce the Rey–Osterrieth Complex Figure.

Finally, providing a clear and distinctive external structure when the patients encounter the memoranda can aid memory performance. Specifically, Kopelman and Stanhope (1998) found that frontal patients had better memory when the stimulus materials were organized by category, as opposed to when the materials were ordered randomly. Moreover, Winocur and Kinsbourne (1978) demonstrated that when patients studied paired-associates in the presence of a contextually rich environment (i.e., with colored lights and classical music playing in the background), they had better memory.

In sum, these studies demonstrate that providing detailed retrieval cues, external encoding strategies, and distinctive environmental cues can serve to lessen the severity of the memory impairments experienced by dysexecutive amnesics.

Disorders Associated with Dysexecutive Amnesia

Apart from focal frontal lobe lesions of vascular or traumatic origin, a wide array of disorders and syndromes can result in damage to the frontal lobes or in damage to pathways that connect the frontal lobes and medial temporal lobes. Here, we describe several disorders that are frequently associated with frontal lobe pathology. Consequently, individuals with these syndromes commonly display symptoms characteristic of dysexecutive amnesia. Although this section does not cover all of the possible syndromes that could result in frontal damage,

it summarizes key disorders most often accompanied by components of dysexecutive amnesia.

Vascular Thalamic Amnesia

Vascular thalamic amnesia occurs following damage to the thalamus and can result in the presentation of various dysexecutive amnesia symptoms. Carlesimo et al. (2011) recently reviewed reports of vascular thalamic amnesia to characterize specific thalamic lesion locations and their associated memory impairments. They found that damage to the mammillothalamic tract, in particular, predicted the presence of memory deficits in patients with thalamic lesions. Because the anterior nucleus of the thalamus is part of the mammillothalamic tract that sends projections to the hippocampus and mammillary bodies via the fornix, the anterior nucleus of the thalamus is implicated as a critical component that supports normal memory function.

Carlesimo et al. (2011) reviewed data from 83 patients with vascular thalamic lesions and carefully characterized the commonly reported memory disturbances. The authors explain that, like amnesics with medial temporal damage, thalamic amnesics tend to have impaired declarative anterograde amnesia, occasionally impaired declarative retrograde amnesia, spared implicit memory, and spared short-term memory. However, notably, unlike amnesics with temporal damage, vascular thalamic amnesics also tend to display dysexecutive symptoms observed in frontal patients. For instance, they are frequently impaired in set shifting, as indexed by impaired performance on the Trail-Making tests and the Wisconsin Card Sorting Test. Furthermore, these patients tend to have impaired inhibition on the Stroop test and impaired verbal fluency. In addition to these primarily cognitive dysexecutive deficits, Carlesimo et al. (2011) also report altered emotion in vascular thalamic patients, similar to the altered emotion seen in those with frontal damage. In particular, patients can display apathy and dysphoria whereas others can display mania. These findings indicate vascular damage to the mammillothalamic tract can result in memory impairments similar to those evident in patients with frontal lobe damage.

Korsakoff's Amnesia

Korsakoff's amnesia is a memory disturbance caused by thiamine deficiency that results in damage to the mammillothalamic tract. Chronic alcoholism can lead to thiamine deficiency, which explains why Korsakoff's amnesics are typically chronic alcoholics. Similar to vascular thalamic amnesics who have damage to the anterior nucleus of the thalamus and the mammillothalamic tract, Harding et al. (2000) found that damage to the anterior nucleus of the thalamus differentiates alcoholics with and without amnesia. Thus, Korsakoff's amnesia is also associated with damage to this pivotal pathway.

Korsakoff's amnesics display many similar memory impairments as frontal patients. As such, Korsakoff's patients could be considered a subclass of patients who experience dysexecutive amnesia. Korsakoff's amnesics have marked difficulty with temporal memory (Huppert and Piercy, 1976; Squire, 1982; Meudell et al., 1985), which is a characteristic symptom of dysexecutive amnesia. In fact, Squire (1982) found a correlation between the severity of Korsakoff patients' frontal symptoms and their performance on a temporal memory task. Furthermore, like individuals with frontal lesions, Korsakoff's patients tend to confabulate (Jacobson and Lishman, 1990), show metamemory impairments (Shimamura et al., 1991), exhibit impaired recall despite spared recognition (e.g., Huppert and Piercy, 1976), perseverate (Jacobson and Lishman, 1990; Shimamura et al., 1991), and they fail to show a release from proactive interference (e.g., Moscovitch, 1982; Squire, 1982; Shimamura et al., 1991). The pattern of impairments found in Korsakoff's patients suggests the memory deficits they suffer from are dysexecutive amnesia.

Rupture of Anterior Communicating Artery Aneurysm

The anterior communicating artery forms part of the Circle of Willis on the ventral surface of the brain, and it is one of the avenues of blood supply to the frontal lobes. When an aneurysm of the anterior communicating artery ruptures, or when an aneurysm has to be surgically repaired to prevent rupture, damage to frontal cortex often occurs. As a result, patients with anterior communicating artery aneurysms often exhibit memory impairments with hallmark features of dysexecutive amnesia, including impaired recall with normal recognition (Hanley et al., 1994; but see Delbecq-Derouesné et al., 1990), increased false alarm rates (Parkin et al., 1996), perseveration (Stenhouse et al., 1991), and confabulation. See Parkin and Leng (1993) for a review of frontal deficits associated with rupture of the anterior communicating artery.

Schizophrenic Memory Impairment

Although memory impairments are not the defining characteristics of schizophrenia, schizophrenics can have impaired memory. Notably, some of the memory disturbances observed in schizophrenic patients resemble those associated with dysexecutive amnesia. In particular, schizophrenics have a disproportionate impairment in recall compared to recognition (e.g., Goldberg et al., 1989). Moreover, schizophrenics show deficits on executive tasks that involve a prefrontal component. These memory and executive impairments mirror those observed in patients with frontal lesions, implicating the potential malfunctioning of the prefrontal cortex in individuals with schizophrenia.

Conclusion

From the above evidence, it is clear that the frontal lobes play an important role in strategic encoding, retrieval, and monitoring of memory. Damage to the frontal lobes, thus, often results in a wide array of memory disturbances, which are characterized here as symptoms of dysexecutive amnesia. To reiterate, prominent deficits of dysexecutive amnesia include

a disproportionate impairment in recall compared to recognition memory, difficulty with temporal ordering, source amnesia, perseveration, confabulation, and dysfunctional metamemory. Additionally, patients with frontal damage tend to have impaired working memory, heightened susceptibility to proactive interference, impaired frequency judgments, and high false alarm rates. Dysexecutive amnesia is a heterogeneous disorder because the types of symptoms that manifest in each patient depend heavily on the severity and origin of that patient's damage. Nevertheless, it is important to recognize the types of memory disturbances that result from frontal damage and to acknowledge that these types of memory problems are distinct from the memory deficits commonly observed in patients with medial temporal lobe damage.

See also: Amnesia: General; Autobiographical Memory; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Confabulation; Declarative Memory, Neural Basis of; Dysexecutive Syndromes; Episodic Memory; False Memories; Hippocampal Amnesia; Hippocampus and Related Structures; Learning and Memory, Neural Basis of; Proactive Interference; Recognition; Retrieval; Retroactive Interference; Schizophrenia: Metabolic Consequences; Schizophrenia; Short Term Memories, Theories of; Short-Term Memory: Psychological and Neural Aspects; Source Monitoring; Temporal Distinctiveness Models of Memory; Working Memory, Neural Basis of; Working Memory, Psychology of.

Bibliography

Baddeley, A., Wilson, B., 1988. Frontal amnesia and the dysexecutive syndrome. Brain and Cognition 7 (2), 212–230. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3377900.

Bechara, A., et al., 1996. Failure to respond autonomically to anticipated future outcomes following damage to prefrontal cortex. Cerebral Cortex 6 (2), 215–225. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8670652.

Carlesimo, G.A., Lombardi, M.G., Caltagirone, C., 2011. Vascular thalamic amnesia: a reappraisal. Neuropsychologia 49 (5), 777–789. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21255590 (accessed 15.03.13).

D'Esposito, M., et al., 1996. Recovery of memory and executive function following anterior communicating artery aneurysm rupture. Journal of the International Neuropsychological Society 2 (6), 565–570.

D'Esposito, M., Postle, B.R., 2002. The neural basis of working memory storage, rehearsal, and control processes. In: Squire, Larry R., Schacter, Daniel. L. (Eds.), Neuropsychology of Memory. The Guilford Press, New York, pp. 215–224.

Delbecq-Derouesné, J., Beauvois, M.F., Shallice, T., 1990. Preserved recall versus impaired recognition. A case study. Brain: A Journal of Neurology 113, 1045–1074. Available at: http://discovery.ucl.ac.uk/123362/.

Della Sala, S., et al., 1993. Autobiographical recollection and frontal damage. Neuropsychologia 31 (8), 823–839.

Diamond, B.J., DeLuca, J., Kelley, S.M., 1997. Memory and executive functions in amnesic and non-amnesic patients with aneurysms of the anterior communicating artery. Brain 120, 1015–1025.

Dickerson, B.C., Eichenbaum, H., 2010. The episodic memory system: neurocircuitry and disorders. Neuropsychopharmacology 35 (1), 86–104. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2882963&tool=pmcentrez&rendertype=abstract (accessed 28.02.13).

Eichenbaum, H., et al., 2010. An animal model of amnesia that uses receiver operating characteristics (ROC) analysis to distinguish recollection from familiarity deficits in recognition memory. Neuropsychologia 48 (8), 2281–2289. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2891833&tool=pmcentrez&rendertype=abstract (accessed 15.03.13).

Funahashi, S., Bruce, C.J., Goldman-Rakic, P.S., 1990. Visuospatial coding in primate prefrontal neurons revealed by oculomotor paradigms. Journal of Neurophysiology 63 (4), 814–831. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2341879.

Goldberg, T.E., et al., 1989. Recall memory deficit in schizophrenia. A possible manifestation of prefrontal dysfunction. Schizophrenia Research 2 (3), 251–257. Available at: http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=2487166.

Hanley, J.R., et al., 1994. Impaired recall of verbal material following rupture and repair of an anterior communicating artery aneurysm. Cognitive Neuropsychology 11 (5), 543–578.

Harding, A., et al., 2000. Degeneration of anterior thalamic nuclei differentiates alcoholics with amnesia. Brain: A Journal of Neurology 123, 141–154. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10611128.

Huppert, F.A., Piercy, M., 1976. Recognition memory in amnesic patients: effect of temporal context and familiarity of material. Cortex 12 (1), 3–20. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=1261280.

Jacobsen, C.F., 1936. The functions of the frontal association areas in monkeys. In: Studies of Cerebral Function in Primates. Williams and Wilkins, Baltimore, pp. 3–60.

Jacobson, R.R., Lishman, W.A., 1990. Cortical and diencephalic lesions in Korsakoff's syndrome: a clinical and CT scan study. Psychological Medicine 20 (1), 63–75.

Janowsky, J.S., Shimamura, A.P., Squire, L.R., 1989a. Memory and metamemory: comparisons between patients with frontal lobe lesions and amnesic patients. Psychobiology 17, 3–11.

Janowsky, J.S., Shimamura, A.P., Squire, L.R., 1989b. Source memory impairment in patients with frontal lobe lesions. Neuropsychologia 27 (8), 1043–1056. Available at: http://dx.doi.org/10.1016/0028-3932(89)90184-X.

Jetter, W., et al., 1986. A verbal long term memory deficit in frontal lobe damaged patients. Cortex 22 (2), 229–242. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3731793.

Kopelman, M.D., Stanhope, N., 1998. Recall and recognition memory in patients with focal frontal, temporal lobe and diencephalic lesions. Neuropsychologia 36 (8), 785–796.

Kopelman, M.D., Stanhope, N., Guinan, E., 1998. Subjective memory evaluations in patients with focal frontal, diencephalic, and temporal lobe lesion. Cortex 34 (2), 191–207. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9606585.

Meudell, P.R., et al., 1985. Recency and frequency judgements in alcoholic amnesics and normal people with poor memory. Cortex 21 (4), 487–511. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4092481.

Milner, B., Corsi, P., Leonard, G., 1991. Frontal-lobe contribution to recency judgements. Neuropsychologia 29 (6), 601–618. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21296169.

Moscovitch, M., 1982. Multiple dissociations of function in amnesia. In: Cermak, L.S. (Ed.), Human Memory and Amnesia. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 337–370.

Moscovitch, M., 1992. Memory and working-with-memory: a component process model based on modules and central systems. Journal of Cognitive Neuroscience 4 (3), 257–267. Available at: http://www.mitpressjournals.org/doi/abs/10.1162/jocn.1992.4.3.257.

Parkin, A.J., et al., 1988. Memory impairment following ruptured aneurysm of the anterior communicating artery. Brain and Cognition 7 (2), 231–243. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3377901.

Parkin, A.J., Leng, N.R.C., 1993. Neuropsychology of the Amnesic Syndrome. Lawrence Erlbaum Associates, Hillsdale, NJ.

Parkin, A.J., et al., 1996. Pathological false alarm rates following damage to the left frontal cortex. Brain and Cognition 32 (1), 14–27. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8899212.

Rolls, E.T., et al., 1994. Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. Journal of Neurology, Neurosurgery & Psychiatry 57 (12), 1518–1524. Available at: http://jnnp.bmj.com/cgi/doi/10.1136/jnnp.57.12.1518.

Schacter, D., Harbluk, J., McLachlan, D., 1984. Retrieval without recollection: an experimental analysis of source amnesia. Journal of Verbal Learning and Verbal Behavior 23 (5), 593–611. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0022537184903736.

Schacter, D.L., Curran, T., 1995. The cognitive neuroscience of false memories. Psychiatric Annals 25 (12), 726–730.

Shimamura, A.P., Janowsky, J.S., Squire, L.R., 1990. Memory for the temporal order of events in patients with frontal lobe lesions and amnesic patients. Neuropsychologia 28 (8), 803–813. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2247207.

Shimamura, A.P., Janowsky, J.S., Squire, L.R., 1991. What is the role of frontal lobe damage in memory disorders? In: Levin, H.S., Eisenberg, H.M., Benton, A.L. (Eds.), Frontal Lobe Function and Dysfunction. Oxford University Press, pp. 173–195.

Simons, J.S., Spiers, H.J., 2003. Prefrontal and medial temporal lobe interactions in long-term memory. Nature Reviews Neuroscience 4 (8), 637–648. Available at: http://discovery.ucl.ac.uk/155387/ (accessed 28.02.13).

Squire, L.R., 1982. Comparisons between forms of amnesia: some deficits are unique to Korsakoff's syndrome. Journal of Experimental Psychology Learning Memory and Cognition 8 (6), 560–571. Available at: http://content.apa.org/journals/xlm/8/6/560.

Stanhope, N., Guinan, E., Kopelman, M.D., 1998. Frequency judgements of abstract designs by patients with diencephalic, temporal lobe or frontal lobe lesions. Neuropsychologia 36 (12), 1387–1396. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9863692.

Stenhouse, L.M., et al., 1991. Long-term cognitive deficits in patients after surgery on aneurysms of the anterior communicating artery. Journal of Neurology, Neurosurgery & Psychiatry 54 (10), 909–914. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11004127.

Swick, D., Knight, R.T., 1999. Contributions of prefrontal cortex to recognition memory: electrophysiological and behavioral evidence. Neuropsychology 13 (2), 155–170. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10353368.

Thompson-Schill, S.L., et al., 1997. Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. Proceedings of the National Academy of Sciences of the United States of America 94 (26), 14792–14797. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=25116&tool=pmcentrez&rendertype=abstract.

Turner, M.S., et al., 2008. Confabulation: damage to a specific inferior medial prefrontal system. Cortex 44 (6), 637–648. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18472034 (accessed 07.03.13).

Winocur, G., Kinsbourne, M., 1978. Contextual cueing as an aid to Korsakoff amnesics. Neuropsychologia 16 (6), 671–682. Available at: http://www.ncbi.nlm.nih.gov/pubmed/748804.

Dysexecutive Syndromes

Brenda Hanna-Pladdy, University of Maryland School of Medicine, Baltimore, MD, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A.P. Shimamura, volume 6, pp. 3911–3913, © 2001, Elsevier Ltd.

Abstract

The frontal lobes have a central role in cognition, emotion, and action related to their late evolutionary development and anatomical connectivity. However, damage to a wide range of brain regions can also produce these 'frontal deficits.' Thus, the term dysexecutive syndrome has clinical utility to capture a range of behavioral and cognitive deficits that can result from various diffuse neurologic conditions. This article will provide an overview of dysexecutive syndromes common in neurologic disease typically identified by neuropsychological and behavioral measures, with recognition that assessments have their limitations since dysexecutive deficits can be more pronounced under specific environmental circumstances.

Introduction

Establishing and maintaining goals over time is a critical requirement to gain mastery over the environment, and is optimized by the ability to monitor internal and external cues, and to maintain new information online to assist in generating alternate plans to changing contingencies. Adaptation to the environment is critical to maximizing the likelihood of goal attainment, and demands temporal control of initiation and suppression of behavioral responses. Ultimately, this requires orchestration and flexibility in controlling a diverse range of cognitive functions including planning, initiation, decision-making, inhibition, and set-shifting (for review see Damasio and Anderson (2003), Kolb and Whishaw (2003), and Lezak et al. (2004)). Furthermore, behavioral disorders such as apathy, abulia, perseveration, unconcern, unawareness, and poor self-monitoring can be observed in patients without cognitive deficits, and have resulted in proposed separate criteria for cognitive versus behavioral dysexecutive syndromes (Godefroy et al., 2010). However, quite often there is interplay between cognitive and behavioral deficits that characterizes the dysexecutive syndrome. For example, a failure to initiate an action can be classified as several intentional disorders, although intentional activity can clearly be influenced by emotional factors (motivation) or absence of initiative (apathy, abulia), demonstrating the interaction between behavior and cognition (Mesulam, 2000; Heilman et al., 2003). Conversely, initiation of behaviors can be inappropriate related to failures of response inhibition and can interfere with goal attainment, or become disruptive because of behavioral disinhibition, inappropriateness, increased talkativeness, lack of tact, and tendency to joke (Witzelsucht) (Adams et al., 1997). Therefore, there is a significant challenge in assessment and diagnostic characterization when considering the full range of cognitive and behavioral executive systems that can be differentially affected in neurologic disease.

Maintaining Executive Control

Conceptualization of executive functions can be described in terms of aspects of intentional activity that require initiation and organizational structure, with emphasis on the capacity to formulate a goal or intention and respond to changing contingencies.

Planning

Planning requires the ability to identify and organize the necessary steps to achieve a goal, the ability to conceptualize, view the environment in an objective fashion, generate alternatives, and make decisions prior to the initiation of a behavioral response (Lezak et al., 2004). There are few tests that directly measure planning, although standardized neuropsychological tests can yield insights into the organizational approach to a task, and the sequential ordering of actions to achieve a goal such as the Tower of Hanoi or Tower of London (Damasio and Anderson, 2003). Damage to the dorsolateral prefrontal cortex (PFC) has been associated with impairments in planning, although orbitofrontal cortex (OFC) damage is a better predictor of real-world planning deficits that require coordination of both cognitive and behavioral regulatory systems (Sarazin et al., 1998; Royall et al., 2002; Koechlin et al., 2000).

Initiation

The initiation and self-generation of behavioral responses over a period of time in order to accomplish a goal is mediated by bilateral superior medial frontal regions. The same segments of the frontal lobes result in reduced reaction times on tests of sustained attention, but appear to also involve supplementary motor area (SMA) and pre-SMA (Stuss et al., 2005). Slowed responding and decreased initiation can occur across a number of cognitive and behavioral domains resulting in devastating functional consequences. Tests measuring generational capacity (e.g., verbal fluency) under time and rule constraints are frequently utilized in neuropsychological assessment, although deficiencies are not always secondary to executive dyscontrol given the speech and language demands (Lezak et al., 2004; Kolb and Whishaw, 2003; Stuss et al., 1998; Stuss and Alexander, 2007). Nonetheless, speech and language output also require movement selection mediated by frontal speech zones, and during a fluency task words are responses

generated in the context of both internal and external stimuli (Passingham, 1993). Similarly, the generation of complex sequential movements involves the SMA, in particular, when relying on internally driven self-paced movements (Passingham, 1993; Deiber et al., 1991). Alexander Luria introduced motor sequencing tasks (fist-edge-palm) into neuropsychological assessments and this continues to be a classic test for the integrity of the frontal lobes (Luria, 1966).

Inhibition

Failure to inhibit behavioral responses that are mismatched to environmental circumstances are hallmark features of frontal lobe disease, and can become most apparent when patients exhibit inappropriate verbal or behavioral responses (Lezak et al., 2004; Mesulam, 2000). This type of behavioral disinhibition can have significant consequences on real-world functioning in terms of interpersonal relationships, as well as consequences of impulsive or faulty decision-making. Lesions in the OFC and medial frontal areas (i.e., parts of the frontal lobe containing paralimbic cortex) have been closely implicated in behavioral disinhibition and poor emotional regulation (for review see Damasio and Anderson (2003), LeDoux (1996), and Mesulam (2000)). Assessment of behavioral dysregulation (e.g., apathy and disinhibition) has been conducted utilizing behavioral rating scales such as the Frontal Systems Behavior Scale (FrSBe) (Paulsen et al., 1996). Since behavioral apathy is predictive of impaired activities of daily living, behavioral scales may provide additional predictive functional information above that available from cognitive tests (Norton et al., 2001; Stout et al., 2003).

The frontal lobes have been implicated in failure to withhold responses, with the PFC specifically controlling the timing of when to initiate and when to withhold a response (Lezak et al., 2004). A number of cognitive tests have been utilized to measure response inhibition including the Stroop test, Trails B, Digit Span Backwards, and Go–Nogo paradigms (see Lezak et al. (2004) for review and description). Trails B and Digit Span Backwards have loaded with the Stroop on a single factor of response inhibition, although other investigators have emphasized aspects of mental control in contrast to inhibitory control (for review see Royall et al. (2002)). Evidence from fMRI supports anterior cingulate and mesiofrontal extensions mediating Stroop processing reflecting that it taps into multiple dimensions of executive control simultaneously (Royall et al., 2002; Peterson et al., 1999). Impairment on the Go–Nogo task has been associated with PFC lesions in both animals and humans (Drewe, 1975; Royall et al., 2002), and recently with lesions to the superior medial frontal region (including the SMA) (Picton et al., 2007).

Cognitive Flexibility

Failure to alter behavior in response to environmental demands can stem from both faulty decision-making, as well as poor transfer of learning from past consequences so that it efficiently influences current behavior (Damasio and Anderson, 2003; Bechara et al., 1999). The OFC has been implicated in the decision-making process, and damage to this region can impair estimations of the strength of the reward associated with an outcome, resulting in devastating consequences on functional outcomes (Wallis, 2007). Because of the extensive connections with the limbic system, cingulate and hippocampus, the OFC influences emotionally driven decision-making, with uncertainties reconciled with decisions based also on past memory of reward and punishment (Carmichael and Price, 1995). However, comparison of patients with amygdala and ventromedial PFC lesions, suggests that while the OFC has primary role in decision-making, the amygdala mediates attachment of affective attributes to stimuli (Bechara et al., 1999).

Classic neuropsychological measures of executive functioning have concentrated on measuring how rapidly individuals modify well-learned behaviors in response to changing environmental cues. The two most common tasks utilized are the Go–Nogo task (also measuring response inhibition) and the Wisconsin Card Sorting Task (WCST). There has been considerable research on the WCST suggesting that frontal lobe lesion patients (in particular PFC lesions) are more severely impaired in their ability to shift response sets from previously learned criteria to new feedback provided by the examiner (Berg, 1948; Grant and Berg, 1948). However, the WCST measures multiple dimensions of executive control and performance does not appear to be specific to frontal lobe damage but can involve widespread brain regions (Royall et al., 2002; Stuss et al., 2000).

Ecological Validity of Assessment

Although specific neurologic conditions may have expected prominent dysexecutive syndromes (e.g., traumatic brain injury, Parkinson's disease, and Frontotemporal dementias), cognitive and behavioral dysexecutive syndromes can be present in a wide range of neurologic conditions given the dense connections between the frontal lobes and other regions of the brain. The traditional tests of frontal-executive dysfunction reviewed above are outlined with respect to aspects of executive control typically evaluated in a traditional neuropsychological assessment. However, standardized tests may be remote from and not informative about functional failures in real life and whether patients are able to meet goals and reintegrate in a psychosocial fashion following neurologic disease. Thus, there is a need for improved ecological assessment of executive functions, which could ultimately predict treatment outcomes and everyday functioning (Burgess et al., 2000).

See also: Brain Damage: Neuropsychological Rehabilitation; Dysexecutive Amnesia; Neuropsychological Testing; Prefrontal Cortex.

Bibliography

Adams, R., Victor, M., Ropper, A.H., 1997. Neurologic disorders caused by lesions in particular parts of the cerebrum. In: Principles of Neurology, sixth ed. McGraw-Hill Companies, New York.

Bechara, A., Damasio, H., Damasio, A.R., Lee, G.P., 1999. Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. Journal of Neuroscience 19, 5473–5481.

Berg, E., 1948. A simple objective technique for measuring flexibility in thinking. Journal of Genetic Psychology 39, 15–22.

Burgess, P.W., Veitch, E., De Lacy Costello, A., Shallice, T., 2000. The cognitive and neuroanatomical correlates of multitasking. Neuropsychologia 38, 848–863.

Carmichael, S.T., Price, J.L., 1995. Limbic connections of the orbital and medial prefrontal cortex in macaque monkeys. Journal of Comparative Neurology 363, 615–641.

Damasio, A., Anderson, S.W., 2003. The frontal lobes. In: Heilman, K., Valenstein, E. (Eds.), Clinical Neuropsychology. Oxford University Press, New York, pp. 404–446.

Deiber, M.P., Passingham, R.E., Colebatch, J.G., Friston, K.J., Nixon, P.D., Frackowiak, R.S., 1991. Cortical areas and the selection of movement: a study with positron emission tomography. Experimental Brain Research 84, 393–402.

Drewe, E.A., 1975. Go–Nogo learning after frontal lobe lesions in humans. Cortex 11, 8–16.

Godefroy, O., Azouvi, P., Robert, P., Roussel, M., Legall, D., Meulemans, T., 2010. Dysexecutive syndrome: diagnostic criteria and validation study. Annals of Neurology 68, 855–864.

Grant, D., Berg, E.A., 1948. A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a weigl-type card sorting problem. Journal of Experimental Psychology 38, 404–411.

Heilman, K., Watson, R.T., Valenstein, E., 2003. Neglect and related disorders. In: Heilman, K., Valenstein, E. (Eds.), Clinical Neuropsychology. Oxford University Press, New York, pp. 447–478.

Koechlin, E., Corrado, G., Pietrini, P., Grafman, J., 2000. Dissociating the role of the medial and lateral anterior prefrontal cortex in human planning. Proceedings of the National Academy of Sciences of the United States of America 97, 7651–7656.

Kolb, B., Whishaw, I.Q., 2003. The Frontal Lobes. Fundamentals of Human Neuropsychology, fifth ed. Worth Publishers, New York.

LeDoux, J., 1996. The Holy Grail. The Emotional Brain; the Mysterious Underpinnings of Emotional Life. Touchstone, New York.

Lezak, M., Howieson, D.B., Loring, D.W., 2004. Executive Functions and Motor Performance. Neuropsychological Assessment. Oxford University Press, Oxford.

Luria, A., 1966. Higher Cortical Functions in Man. Basic Books, New York.

Mesulam, M.M., 2000. Behavioral Neuroanatomy. Principles of Behavioral and Cognitive Neurology. Oxford University Press, New York.

Norton, L.E., Malloy, P.F., Salloway, S., 2001. The impact of behavioral symptoms on activities of daily living in patients with dementia. American Journal of Geriatric Psychiatry 9, 41–48.

Passingham, R.E., 1993. The Frontal Lobes and Voluntary Action. Oxford University Press, Oxford.

Paulsen, J.S., Stout, J.C., De La Pena, J., et al., 1996. Frontal behavioral syndromes in cortical and subcortical dementia. Assessment 3, 327–337.

Peterson, B.S., Skudlarski, P., Gatenby, J.C., Zhang, H., Anderson, A.W., Gore, J.C., 1999. An fMRI study of Stroop word-color interference: evidence for cingulate subregions subserving multiple distributed attentional systems. Biological Psychiatry 45, 1237–1258.

Picton, T.W., Stuss, D.T., Alexander, M.P., Shallice, T., Binns, M.A., Gillingham, S., 2007. Effects of focal frontal lesions on response inhibition. Cerebral Cortex 17, 826–838.

Royall, D.R., Lauterbach, E.C., Cummings, J.L., Reeve, A., Rummans, T.A., Kaufer, D.I., Lafrance Jr., W.C., Coffey, C.E., 2002. Executive control function: a review of its promise and challenges for clinical research. A report from the committee on research of the American neuropsychiatric association. Journal of Neuropsychiatry and Clinical Neurosciences 14, 377–405.

Sarazin, M., Pillon, B., Giannakopoulos, P., Rancurel, G., Samson, Y., Dubois, B., 1998. Clinicometabolic dissociation of cognitive functions and social behavior in frontal lobe lesions. Neurology 51, 142–148.

Stout, J.C., Wyman, M.F., Johnson, S.A., Peavy, G.M., Salmon, D.P., 2003. Frontal behavioral syndromes and functional status in probable Alzheimer disease. American Journal of Geriatric Psychiatry 11, 683–686.

Stuss, D.T., Alexander, M.P., 2007. Is there a dysexecutive syndrome? Philosophical Transactions of the Royal Society of London B Biological Sciences 362, 901–915.

Stuss, D.T., Alexander, M.P., Hamer, L., Palumbo, C., Dempster, R., Binns, M., Levine, B., Izukawa, D., 1998. The effects of focal anterior and posterior brain lesions on verbal fluency. Journal of International Neuropsychological Society 4, 265–278.

Stuss, D.T., Alexander, M.P., Shallice, T., Picton, T.W., Binns, M.A., Macdonald, R., Borowiec, A., Katz, D.I., 2005. Multiple frontal systems controlling response speed. Neuropsychologia 43, 396–417.

Stuss, D.T., Levine, B., Alexander, M.P., Hong, J., Palumbo, C., Hamer, L., Murphy, K.J., Izukawa, D., 2000. Wisconsin card sorting test performance in patients with focal frontal and posterior brain damage: effects of lesion location and test structure on separable cognitive processes. Neuropsychologia 38, 388–402.

Wallis, J.D., 2007. Orbitofrontal cortex and its contribution to decision-making. Annual Review of Neuroscience 30 (1), 31–56.

Dyslexia, Developmental

Usha Goswami, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Children with developmental dyslexia have specific problems with reading and spelling. These problems have no apparent environmental or neurological cause, and are thought to be due to difficulties in the accurate specification and neural representation of the sound structure of speech. The normal development of lexical representation across languages is outlined, followed by a survey of the abnormal lexical development characteristic of developmental dyslexia. Possible basic auditory processing deficits are considered. Relevant information from brain imaging and genetic linkage studies is then discussed. Finally, possible subtypes of developmental dyslexia, and the usefulness of a discrepancy definition, are discussed.

Developmental dyslexia can be characterized as a specific problem with reading and spelling that cannot be accounted for by low intelligence, poor educational opportunities, or obvious sensory or neurological damage. Children with developmental dyslexia often appear to have good spoken language skills and extensive vocabularies. Accordingly, the first theories of developmental dyslexia postulated underlying deficits in visual processing ('word blindness'). Recent evidence from developmental psychology, genetics, and brain imaging suggests, however, that the central deficit in developmental dyslexia is a linguistic one. In particular, dyslexic children appear to have difficulties in the accurate specification of phonology when storing lexical representations of speech.

The Normal Development of Lexical Representations

A schematic view of a lexical representation is shown in Figure 1, illustrating different levels of phonology. Research in child phonology suggests that, prior to acquiring literacy, the child has represented spoken words as distinctive but complex auditory patterns ("prosodic templates," Vihman and Croft, 2007). These prosodic patterns vary in terms of auditory parameters like loudness, pitch, and duration. When the children learn to read an alphabetic language, they learn that these complex auditory patterns can be represented by sequences of letters. This changes the way in which spoken words are represented in the brain. Phonology is 'remapped' to reflect letter-sound knowledge (Frith, 1998). Frith (1998) wrote that acquiring the alphabetic code was like catching a virus: "This virus infects all speech processing, as now whole word sounds are automatically broken up into sound constituents. Language is never the same again" (p. 1051).

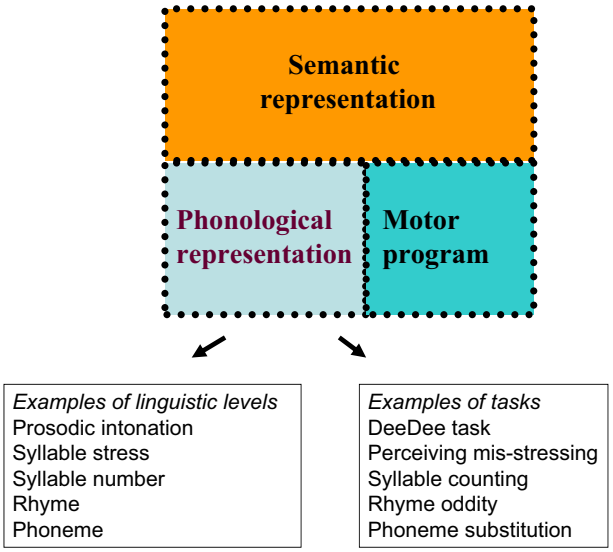

Figure 1 Schematic depiction of a lexical representation, showing different linguistic levels of phonological representation and example tasks.

Prior to learning to read, the phonological representation of lexical forms is relatively holistic, based more on prosodic characteristics and global phonological characteristics. For English, the dominant prosodic template is a rhythmic pattern comprising two syllables, with stronger stress on the first syllable. Examples are 'mummy,' 'daddy,' 'baby,' and 'bottle.' Mothers and fathers across languages use a special prosodic register, called infant-directed speech (IDS), to highlight these rhythmic cues for their babies. IDS uses auditory features like exaggerated rhythm and intonation, heightened pitch, and increased duration. This prosodic exaggeration seems to serve a language learning function, helping infants in different languages to extract individual words on the basis of pitch and rhythmic cues (e.g., Echols, 1996). Speech rhythm is very important for lexical learning, and infants can distinguish languages belonging to different rhythm types from birth.

For individual lexical items (words), research with preliterate children shows that syllables are represented first (a word like *wigwam* has two syllables). Preliterate children also represent *onsets* and *rimes* (the onset in a syllable corresponds to the sound/s before the vowel, the rime corresponds to the vowel and any subsequent sounds, e.g., *p – it*; *scr – ew*). As children learn to read, children represent *phonemes*, and they do this very rapidly in transparent orthographies. Phonemes are the smallest units of sound that change word meaning (e.g., *pit* differs from *sit*, *pot*, and *pin* by the initial, medial, and final phonemes, respectively). However, phonemes are an abstraction from the actual speech stream. For example, in English, the sound symbolized by the letter P has one acoustic form in the spoken word 'cap,' and a different acoustic form in the spoken

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54003-3

word 'spoon.' Beginning readers are aware of this and may try to write the word 'spoon' as SBN. To become efficient readers, they need to develop orthographic representations that supplement and on occasion rerepresent the early-developing acoustic representations of speech (their prosodic templates).

Children with dyslexia have ongoing difficulties in the representation of phonology, and these difficulties are at all phonological levels, including prosody. Much of the evidence for this 'phonological deficit' has come from word-level 'phonological awareness' tasks, in which children are asked to make judgments about shared phonology or to manipulate phonology (e.g., the *rhyme oddity* task, in which the child is asked to distinguish the odd word out from a triple like *man, pat, fan*, and the *phoneme substitution* task, in which the child is asked to substitute the initial phonemes in words, so that Bob Dylan becomes Dob Bylan). A large number of studies carried out in many languages have shown that dyslexic children perform poorly in these sublexical phonological awareness tasks, and also in syllable awareness tasks (see Ziegler and Goswami, 2005; for a cross-language review). Importantly, such deficits are found even when dyslexic children's performance is compared to that of younger children who are reading and spelling at the same level as them (the methodologically important *reading level match* research design, which holds reading level constant rather than chronological age and thus gives a mental age advantage to the dyslexic children). More recently, it has been shown that children with dyslexia also have prosodic difficulties, for example, finding it difficult to judge whether syllables are stressed or unstressed (Goswami, 2011). They also find it difficult to judge patterns of beat distribution in music, and these difficulties are also found in comparison to younger, reading level matched children (Goswami et al., 2013). A phonological deficit seems to characterize dyslexia across languages. This suggests that an important causal factor in dyslexia is an underlying difficulty in developing phonologically well-specified lexical representations of speech, at all phonological levels (see **Figure 1**).

The Abnormal Development of Phonological Representations in Dyslexia

The cognitive profiles of developmental dyslexics show fairly specific linguistic deficits that are related to phonology rather than semantics. For example, developmental dyslexics have poor phonological memory skills, finding it difficult to remember verbal sequences. Developmental dyslexics are also slower to recognize familiar words on the basis of acoustic input. For example, they require more acoustic phonetic information than age-matched controls to recognize familiar words in speech gating tasks, in which incremental segments of speech from word onset are presented through headphones. They also find it more difficult to judge 'DeeDee' sequences, in which a familiar phrase or name is represented using the stressed syllable DEE and the unstressed syllable 'dee,' so that 'Harry Potter' would be DEEdeeDEEdee. Children with dyslexia also show problems in confrontation naming tasks (in which simple pictures must be named on demand), even though they are very familiar with the items in the pictures. Developmental dyslexics find rapid 'automatized' naming (RAN) tasks difficult. These tasks require the child to name familiar items presented in lists (such as digits) as quickly as possible, thereby accessing and outputting what should be well-practiced and so well-specified phonological forms ('nine, two, five,…'). RAN tasks are particularly strongly associated with dyslexia in transparent alphabetic orthographies, and in nonalphabetic orthographies, like Chinese (Ho et al., 2002). In both transparent alphabetic orthographies and nonalphabetic orthographies, children with dyslexia can learn to recode symbols to sound very accurately. Nevertheless, they never become fluent nor automatic readers, continuing to struggle with the fluent recoding of visual symbols into sound.

Auditory Processing Problems

Studies that have attempted to uncover basic auditory processing deficits at the phonetic level to explain dyslexic difficulty in representing phonological forms have so far produced mixed results. For example, some dyslexic children have great difficulty with 'minimal pairs' tasks that require the discrimination of similar sounding words (e.g., *smack–snack*), yet others do not (Adlard and Hazan, 1998). Some dyslexic children find phoneme categorization and repetition tasks very difficult (e.g., discriminating between or repeating /b/and /d/), and others do not (e.g., Manis et al., 1997). More recently, basic auditory processing tasks that are relevant to the recovery of prosodic structure from the speech stream have been developed, and these measures have shown more consistency across studies and also across languages. A review of auditory sensory processing in dyslexia (Hämäläinen et al., 2013) identified amplitude envelope rise time and duration as auditory measures showing the largest average sample size weighted effect sizes (0.8 for rise time, 0.9 for duration). The whole band amplitude envelope of speech is the fluctuations in the summed energy (intensity) of the signal across different frequency bands, and rises in intensity in this sound envelope (rise times) correspond to syllable onsets. Accurate perception of the amplitude envelope contributes to speech rhythm. Differences between dyslexics and controls in amplitude envelope rise time measures were found in all the studies reviewed by Hämäläinen et al. (2013), and individual differences in rise time sensitivity were linked to reading/spelling in all the studies. Therefore, at least some of the phonological difficulties found in dyslexia appear to be related to auditory processing difficulties (Goswami, 2011).

Phonological Processing Deficits in Dyslexia: Brain Imaging Studies

Brain imaging studies provide converging evidence that the key difficulty in dyslexia is a weakness in the representation of phonological information. The phonological processing system spans a number of cortical and subcortical regions in the brain, including frontal, temporal, and parietal cortex. Word recognition in skilled readers appears to depend on a left-lateralized network of frontal, temporoparietal, and occipito-temporal regions (Price and McCrory, 2008). Interpretation of functional neuroimaging studies of dyslexia is complicated

by the fact that most studies involve adult developmental dyslexics, which means that the brain's compensatory developmental trajectories will account for some if not all of the differences in patterns of task-related activity that are found. Nevertheless, available studies converge in showing underactivation of the key frontal, temporoparietal, and occipitotemporal structures in dyslexia (Pugh, 2006; for review). While all these studies rely on blood flow measures (PET or fMRI), studies focused on mechanism and relying on EEG or MEG are also appearing in the literature. One promising new direction is the study of oscillatory processes in the dyslexic brain. Amplitude (intensity) modulations in different frequency bands in the speech signal appear to be encoded, at least in part, by oscillating networks of auditory neurons that align their endogenous temporal fluctuations with intensity fluctuations in the input. Both syllable rate (delta, 0.5–4 Hz) and phonetic rate (low gamma, ~30 Hz) oscillatory processes have been shown to be atypical in adults with dyslexia (Lehongre et al., 2011; Hämäläinen et al., 2012).

Genetics of Developmental Dyslexia

The heritability of dyslexia has been demonstrated by a large number of studies that have identified multiple genes, as well as multiple environmental risk factors (Peterson and Pennington, 2012; for review). Studies are converging on six candidate genes and moderate heritability effects, with important gene by environment interactions (e.g., the genetic contribution to dyslexia is higher in families with high parental education levels). Most genes identified so far participate in neuronal migration and axon guidance, hence have not thrown further light on causation, and a genome-wide association study of dyslexia remains to be carried out. Genetic studies to date have depended on definitions of the dyslexic phenotype that are based on deficits in phonological awareness tasks and single or nonsense word reading. Thus there is evidence that individual differences in phonological processing are to some extent heritable. Of course, there cannot be a 'gene' for dyslexia in the sense that there is a gene for eye color, as reading is a culturally determined activity. The finding that the genetic contribution to dyslexia actually increases for children who develop in a favorable early environment as far as reading is concerned (children who have carers who stimulate their linguistic awareness via language games, nursery rhymes, and so on, and who read books to them and model and encourage literacy activities by reading and writing extensively themselves) nevertheless suggests that genes are important.

Subtypes of Developmental Dyslexia and the Discrepancy Definition

Historically, there was some debate over whether there were two subtypes of developmental dyslexia: 'surface' dyslexia and 'phonological' dyslexia. Phonological dyslexics were thought to have problems in assembling sublexical phonology (reading using abstract spelling-to-sound 'rules') and this is the phenotype that has been discussed here. Surface dyslexics were thought to have problems in using stored lexical knowledge to pronounce words (they seemed to have difficulty in addressing orthographic representations directly). This 'double dissociation' model of developmental dyslexia was drawn from patterns of adult pathology (the 'acquired' dyslexias) and has been largely discounted. The subtyping argument depended critically on the use of regression procedures, in which performance relationships between the use of lexical and sublexical phonology characteristic of normal children were used to derive confidence limits for assessing the same performance relationships in dyslexic populations. These regression procedures typically depend on the use of *chronological* age-matched normal readers. As processing trade-offs between the reliance on lexical and sublexical phonology may depend on the overall level of word recognition that the child has attained, *reading level* matched controls are more appropriate. Stanovich et al. (1997) demonstrated clearly that when reading level matched controls were used to define processing trade-offs, then almost no 'surface dyslexic' children were found.

More recently, there has been debate over whether the term 'dyslexia' should apply only to children with reading and spelling problems who do not have low intelligence. Siegel and others have long argued that the use of IQ scores to identify dyslexia discriminates against people of different cultural, economic, and/or linguistic backgrounds, and prevents people who need remediation and accommodations for their dyslexia from obtaining the help that they need (Siegel, 1989). As both neuroimaging data and studies of basic auditory processing have identified very similar patterns for children with discrepancy-defined dyslexia and for low-IQ poor readers, in future it seems likely that dyslexia will be defined without recourse to IQ. Hence children who show poor reading and spelling in the absence of poor educational opportunities or obvious sensory or neurological damage, irrespective of IQ, will be identified as requiring additional educational support.

Toward a Phenotype of Dyslexia

Neurodevelopmental disorders are not discrete syndromes, but most typically have a graded character. They are also seldom highly specific, and rarely involve a single deficit. Although a difficulty in representing phonology appears to be the core symptom of developmental dyslexia, many researchers have noted a number of associated deficits that more or less frequently accompany the phonological deficit. For example, dyslexic children can be clumsy, they can have attentional problems, they may have subtle abnormalities in visual processing (such as poor motion processing when visual stimuli are of low contrast and low luminance), and they can have atypical visual–spatial attention. Nevertheless, at present the most widely accepted phenotype of developmental dyslexia depends on deficits in performance on a battery of cognitive tasks that tap the integrity of the child's phonological representations of speech. This battery most typically includes phonological awareness tasks, single word and nonsense word reading tasks, letter knowledge tasks, and RAN tasks. Developmental dyslexia is thus currently best conceived as a subtle linguistic deficit that involves an impairment in the representation of the phonological structure of the speech signal in the brain.

See also: Biological Preconditions for Language Development; Dyslexias, Acquired and Agraphia; Language Development, Theories of; Morphophonology; Phonology.

Bibliography

Adlard, A., Hazan, V., 1998. Speech perception in children with specific reading difficulties (dyslexia). Quarterly Journal of Experimental Psychology 51A, 153–177.

Echols, C.H., 1996. A role for stress in early speech segmentation. In: Morgan, J.L., Demuth, K. (Eds.), Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 151–170.

Frith, U., 1998. Editorial: literally changing the brain. Brain 121, 1051–1052.

Goswami, U., 2011. A temporal sampling framework for developmental dyslexia. Trends in Cognitive Sciences 15, 3–10.

Goswami, U., Huss, M., Mead, N., Fosker, T., Verney, J.P., 2013. Perception of patterns of musical beat distribution in phonological developmental dyslexia: significant longitudinal relations with word reading and reading comprehension. Cortex 49 (5), 1363–1376. http://dx.doi.org/10.1016/j.cortex.2012.05.005.

Hämäläinen, J.A., Rupp, A., Soltész, F., Szücs, D., Goswami, U., 2012. Reduced phase locking to slow amplitude modulation in adults with dyslexia: an MEG study. Neuroimage 59 (3), 2952–2961.

Hämäläinen, J.A., Salminen, H.K., Leppänen, P.H.T., 2013. Basic auditory processing deficits in dyslexia: systematic review of the behavioural and event-related potential/field evidence. Journal of Learning Disabilities 46 (5), 413–427. Advance online publication. http://dx.doi.org/10.1177/0022219411436213.

Ho, C.S., Chan, D.W., Tsang, S., Lee, S., 2002. The cognitive profile and multiple deficit hypothesis in Chinese developmental dyslexia. Developmental Psychology 38, 543–553.

Lehongre, K., Ramus, F., Villiermet, N., Schwartz, D., Giraud, A.-L., 2011. Altered low-gamma sampling in auditory cortex accounts for the three main facets of dyslexia. Neuron 72, 1080–1090.

Manis, F.R., McBride-Chang, C., Seidenberg, M.S., Keating, P., Doi, L.M., Munsun, B., Petersen, A., 1997. Are speech perception deficits associated with developmental dyslexia? Journal of Experimental Child Psychology 66, 211–235.

Peterson, R.L., Pennington, B.F., 2012. Developmental dyslexia. The Lancet 379, 1997–2007.

Pugh, K.R., 2006. A neurocognitive overview of reading acquisition and dyslexia across languages. Developmental Science 9, 448–450.

Price, C.J., McCrory, E., 2008. Functional brain imaging studies of skilled reading and developmental dyslexia. In: Snowling, M.J., Hulme, C. (Eds.), The Science of Reading: A Handbook. Blackwell Publishing Ltd, Oxford, UK.

Siegel, L.S., 1989. IQ is irrelevant to the definition of learning disabilities. Journal of Learning Disabilities 22, 469–478.

Stanovich, K.E., Siegel, L.S., Gottardo, A., 1997. Converging evidence for phonological and surface subtypes of reading disability. Journal of Educational Psychology 89, 114–127.

Vihman, M., Croft, W., 2007. Phonological development: towards a "radical" templatic phonology. Linguistics 45, 683–725.

Ziegler, J.C., Goswami, U., 2005. Reading acquisition, developmental dyslexia and skilled reading across languages: a psycholinguistic grain size theory. Psychological Bulletin 131 (1), 3–29.

Dyslexias, Acquired and Agraphia

Kaylie T Allen, Central Michigan University, Mt. Pleasant, MI, USA
Harry A Whitaker, Northern Michigan University, Marquette, MI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M. Coltheart, volume 6, pp. 3913–3918, © 2001, Elsevier Ltd.

Abstract

Acquired disorders of reading and writing in previously competent readers and writers following damage to the brain are known as *alexia* and *agraphia*, respectively. This article reviews the symptom presentation and underlying mechanisms of various forms of these disorders. A brief discussion on reading and writing disorders in scripts other than English and testing theoretical models via computational modeling concludes the article.

The term alexia refers to a disorder of reading, and is generally used to describe an acquired reading impairment in a previously competent reader, resulting from damage to the brain. This is contrasted with the term dyslexia, which has been used to describe both acquired and developmental reading disorders. Agraphia (also referred to as dysgraphia) is an acquired writing disorder in a previously competent writer, typically resulting from damage to the brain. Different forms of alexia and agraphia disrupt the complex processes of reading and writing in different ways. Observing the various forms of these disruptions can inform our understanding of how language is processed. Forms of alexia and agraphia are often classified according to the pattern of errors they entail.

Forms of Alexia

Central alexias are reading disorders that occur in the context of disruptions to the entire language system and typically accompany other impairments, such as auditory processing difficulties, anomia, aphasia, and/or agraphia. Two distinct reading routes can be relatively impaired in central alexia: reading by sound and reading by sight (Leff, 2004).

Phonological alexia (reading by sight) refers to cases in which the ability to read nonwords aloud is selectively impaired relative to the ability to read real words aloud, as first reported by Beauvois and Derouesné (1979). Ability to read both regular and irregular real words can be nearly intact, with patients reading 85–95% of real words correctly, while even simple nonwords but pronounceable letter strings are unreadable. Patients may also make visual errors in reading and show imageability effects (more concrete words, 'flower,' are more easily read than abstract ones – 'hope').

According to the dual-route model of reading, this form of alexia results from the impairment of the ability to convert graphemes (the smallest units of written language; symbols representing a single sound) to phonemes (the smallest units of spoken language; a single sound). Because patients cannot use the phonological (spelling-to-sound) route for reading, they rely on prior exposure to word forms, and fail to read novel words and nonwords, even those that follow 'legal' pronunciation rules.

Phonological alexia is associated with left perisylvian lesions. Phonological alexia has also been observed to occur in patients with right parietal lesions, who may make more positional errors than alexic patients with left-hemisphere lesions make (Buiatti et al., 2012). A computational modeling study has suggested that phonological alexia may arise from *non*phonological impairments. Nickels et al. (2008) produced impairments matching two symptomatically distinct patients by lesioning the nonlexical route of a dual-route model, implying that phonological alexia may stem from multiple levels of impairment and that no single lesion site can explain all real-world symptom patterns.

Patients with *deep alexia* may read fluently, but make semantic errors such that their produced word is related to the target in meaning but not its graphemic or phonological form (Marshall and Newcombe, 1973). The word 'hound' may be read as 'dog,' the word 'wood' as 'tree.' Patients with deep alexia are usually unable to read nonwords, but produce irregular words with the same semantic errors as regular words. Part of speech effects have been observed, such that nouns are read with fewer errors than verbs and function words. Imageability effects similar to those in phonological alexia are present. Visual errors (*angle* → *angel*) and morphological errors (*running* → *runner*) are also made. The only striking difference in symptom patterns between phonological and deep alexia is the presence of semantic errors in the latter, leading to the suggestion that the two are at opposite ends of a continuum, with phonological being a mild form and deep a severe form of the same impairment (Glosser and Friedman, 1990). Deep alexia is associated with extensive damage to the left temporo-parietal lobe (Leff, 2004).

Case studies and functional imaging studies have found support for a right hemisphere explanation of deep alexia, which suggests that reading errors are produced when the intact reading mechanisms in the right hemisphere are used in place of the extensively damaged normal reading system (Coltheart, 2001). This is opposed by the multiple-deficit hypothesis, which suggests that errors result from the multiply damaged normal reading system in the left hemisphere, following the same damage model as theories of other central alexias.

Several competing explanations of the semantic errors produced in deep alexia are proposed, and although many have gathered empirical support, none adequately address all the evidence. The typicality effect (faster processing for typical semantic category members than atypical ones; e.g., *robin* is

a more quickly processed member of the bird category than *ostrich*) has been used to test for semantic processing impairments in deep alexia. Findings indicated that patients with deep alexia process written, visual, and auditory semantic information differently than both normal and phonologically alexic patients, suggesting an underlying semantic processing impairment (Riley and Thompson, 2010). Another explanation of deep alexia is the modified failure of inhibition theory that proposes that semantic errors are caused by a failure to inhibit incorrect responses at the level of initial semantic processing (Riley and Thompson, 2010).

Surface alexia (reading by sound) describes the reversal of the pattern seen in phonological alexia: it is characterized by worse reading of real words than nonwords. Patients with surface alexia appear to over rely on the standard relationships between letter groups and their corresponding pronunciations, therefore being able to correctly pronounce nonwords by applying regular spelling-to-sound rules but being unable to pronounce irregular words (words in which spelling–sound relationships violate standard phonological rules).

Patients with surface alexia make frequent 'regularization errors,' in which they pronounce irregular words according to regular pronunciation rules. Other difficulties that occur include a length effect (regular and nonwords with more graphemes lead to more errors than those with fewer graphemes) and a frequency sensitivity (performance is worse for low-frequency irregular words than high-frequency ones). Some patients have also been observed to produce an 'irregularization error' with regular words (Patterson et al., 1996). A strong frequency–regularity interaction has been found, in which there is especially poor reading of low-frequency words with irregular spellings (Welbourne et al., 2011). Surface alexia is often observed following parietal or temporoparietal strokes, as well as in patients with progressive semantic dementia (Roeltgen and Lacey, 2010).

A distinction is made between 'output' and 'input' surface alexia. In the 'output' form, errors are not the result of impaired visual word recognition, but of the failure to retrieve word-specific pronunciation information. This retrieval failure leads the patient to use the procedure for reading nonwords aloud: 'sounding out' a word based on regular spelling-to-sound rules, which leads the patient to make regularization errors while reading aloud (Coltheart, 2001). The 'input' form of surface alexia is the failure to recognize a written word as an orthographic whole, therefore requiring translation of the word into phonology by applying standard spelling-to-sound rules. Then, regularization errors lead to the retrieval of the wrong semantic definition, as in the case of patient CD, who defined the irregular printed word *bear* as "a drink, beer" (Coltheart, 2001).

In *lexical nonsemantic reading*, patients lose the ability to comprehend written words, spoken words, and printed pictures, due to the loss of knowledge of word meanings. Schwartz et al. (1980), Cipolotti and Warrington (1995), Coslett (1991), and Lambon Ralph et al. (1995) each document cases of patients with severe semantic impairment but good irregular-word reading, providing evidence for a model of reading with a meaning pathway that is independent from the orthography–phonology (spelling-to-sound) pathway. Although this form of alexia can result from head injuries, it most commonly occurs as a result of progressive brain deterioration, such as in Alzheimer's disease (Coltheart, 2001).

Peripheral alexias are a category of isolated reading impairments that are not accompanied by impairments in other language processes. Generally, these disorders are caused by disruptions in prelexical visual processes, including visual field disorders, visual acuity deficits, and disorders of visuospatial perception.

Pure alexia, the oldest and best-known peripheral reading disorder, is characterized by an inability to read words and nonwords while all other language processes remain intact. Patients are able to recognize and name individual letters, but are unable to efficiently read single words or sentences. Some patients have difficulty distinguishing visually similar letters (Fiset et al., 2006), and some have shown difficulties processing numbers and nonlinguistic stimuli (Starrfelt et al., 2009; Roberts et al., 2010).

Patients with pure alexia can often pronounce words correctly using a backward spelling strategy, in which they dissemble a word into its constituent letters, identify each letter, and then reassemble the letters sequentially. This slow, laborious process is called *letter-by-letter reading*, and shows a linear increase in reading time with word length. The word-length effect has been found to be widely variable across patients, ranging from 97 to 2402 ms per letter (Pflugshaupt et al., 2009). Meanwhile, some patients read too quickly to be relying on letter-by-letter reading at all.

It has been observed that patients with pure alexia often read numbers with more speed and accuracy than alphabetic characters, a finding initially taken as evidence of a separate visual processing system for letter and number recognition. A recent literature review (Starrfelt and Behrmann, 2011) found little support for a dissociation between intact number reading and impaired letter recognition, and noted that normal readers are also better at identifying numbers than letters in brief exposure tasks.

The phenomenon of implicit recognition (also called covert reading) has been observed in many patients with pure alexia. When letter strings are presented for too short a time for explicit recognition to occur, some patients are able to perform semantic classification and lexical decision tasks with an accuracy well above chance. Accuracy on lexical decision-making has been found to be related to the frequency and imageability of the words, and inversely related to severity of injury (Roberts et al., 2010). Saffran and Coslett (1998) argue that the intact right hemisphere enables word recognition but not pronunciation, whereas the left hemisphere operates on information transferred from the right to produce letter-by-letter reading. A competing account is that implicit recognition reflects partial activation in the damaged standard reading system (Lambon Ralph et al., 2004).

Several theories about the mechanism underlying pure alexia have been proposed, including damage to the principal components of the reading system, such as letter identification, that prevents the normal parallel processing that enables rapid, accurate word reading (Saffran and Coslett, 1998); a general perceptual deficit that is especially pronounced with orthographic stimuli, which place high demands on the visual system (Starrfelt et al., 2010); impaired encoding and processing on a prelexical level; and deficits in the system by which

whole word forms are accessed. A connectionist neural network model that accounts for the stage of letter processing was able to simulate letter-by-letter reading, letter confusability, and error patterns matching data collected from patients with pure alexia (Chang et al., 2012). This model predicted patient reaction time above and beyond word length, providing evidence for a visual deficit account of pure alexia.

Functional imaging studies and case studies have proposed a causal link between pure alexia and lesions in an area of the brain linked to visual word forms (Pflugshaupt et al., 2009). The disorder is also associated with lesions to the left occipito-temporal region and the splenium of the corpus callosum. More than 90% of patients with pure alexia also have right hemianopia and visual field deficits, conditions known to negatively impact reading speed (Leff et al., 2006). Rare case examples of pure alexic patients without visual field deficits provide support for the assumption that symptoms of pure alexia are unrelated to the comorbid hemianopia, although many patients with pure alexia may suffer from hemianopic alexia as well (Johnson and Rayner, 2007). Sheldon et al. (2012) found that patients with both pure alexia and hemianopia had significantly larger word-length effects than patients with pure alexia and no hemianopia, suggesting that hemianopia may compound the impairments present in pure alexia.

Severe cases of pure alexia are known as *global alexia*, a condition characterized by very slow and inaccurate letter naming, and sometimes an inability to recognize letters at all. Often, patients with global alexia are still able to write and spell with normal or near-normal accuracy, and number recognition may be spared (Leff and Schofield, 2013). Patient VA (Volpato et al., 2012) exhibited severe reading deficits, including an inability to match spoken letters and numbers to their graphemic forms, but showed evidence for covert reading of single characters, including the ability to recognize shape, orientation, and case of letters. This pattern of impairments suggests preserved access to representation of letters and numbers as visual objects, and may be explained by right hemisphere reading.

Hemianopic alexia is the most common peripheral reading disorder. Patients with hemianopic alexia have intact language functioning, yet show severe reading impairments, including deficits in word identification and reading eye movements, such as smooth saccades and fixations. Hemianopic alexia is somewhat unique among reading disorders, in that lexical and post-lexical processes are fully functioning and impairments do not arise from damage to these systems. The hallmark symptom of hemianopic alexia is significantly reduced reading speed for text as well as for single words; reading accuracy is generally spared. There is also a linear increase in reading time for each additional letter (although less pronounced than the same effect in pure alexia). Reading errors are relatively rare in these patients, but consist of visual omissions of letters, syllables, and short words, and guessing errors, in which patients make meaningful completions of partially seen words (Schuett et al., 2008). Eye-tracking analyses of text reading show that patients with hemianopic alexia have longer fixation and gaze duration than normal readers (Spitzyna et al., 2007). This form of alexia is believed to result from damage to the primary visual cortex that leads to visual field loss (Leff and Schofield, 2013).

Visual alexia is defined by frequent visual errors, in which words are misread as words with shared letters but no other similarities (i.e., *words* → *worse* or *garden* → *anger*). Errors can include substitutions, reordering, omissions, and additions of letters of the target word. It is suggested that these errors stem from a deficit in the visual analysis system at the level of letter identification (Friedmann et al., 2012). Only seven individuals with the classification of acquired visual dyslexia have been reported in the literature to date, reflecting a dearth in research on this form of alexia.

In *neglect alexia*, patients consistently make errors on either the initial or final letters of a word. The most common pattern is one of omissions (*slight* → light) or substitutions (*summer* → hammer) on the left side of a written word. Patients usually start reading sentences from the middle of a word string and are unaware of their impairment. Reading fluency is the same for regular, irregular, and nonwords, and comprehension and production of language are unimpaired. Additionally, these patients may be able to read words correctly if they are presented vertically (De Bleser and Luzzatti, 2008).

This form of alexia is generally found in patients with a left unilateral neglect syndrome, meaning that they do not attend to their visual field, body, and extrapersonal space on the left side. The most common lesion site associated with neglect alexia is the right parietal lobe. Because it is associated with unilateral neglect, neglect alexia is not specifically a disorder of reading, but instead reflects a wide range of neglect-related impairments.

Patients with *attentional alexia* describe letter crowding and migration of letters, and sometimes blend elements of two words into one. Patients are generally poor at reading text and are unable to report the letters making up words they can read. Migration errors often reflect the relative positions of the letters in the words (*land mine* → *lane mind*), a phenomenon unaccounted for by spatial coding impairment explanations of attentional alexia. These patients often perform better when word stimuli are presented alone, instead of flanked by other words and letters (Warrington et al., 1993). In lexical decision tasks, patients are more likely to accept nonwords whose letters can migrate within the string to become a word, but are not sensitive to phonological similarities between words and nonwords (Shalev et al., 2008). Attentional alexia is usually associated with left parietal lobe damage.

Forms of Agraphia

Writing impairments are often classified by the specific disruptions in the writing system they reflect. Individual profiles of agraphia syndromes are found in a variety of patients, reflecting disruptions at different stages of the theoretical spelling system. Agraphia rarely occurs in insolation. Patients with agraphia predominantly have alexia as well, often mirroring the same impairments in writing as in speech. However, the relationship between alexia and agraphia is not always straightforward and is not well understood. Evidence from patients with both phonological alexia and agraphia suggests that a modality-independent deficit underlies both disorders, rather than damage to separate, dedicated cognitive systems for reading and spelling (Rapscak et al., 2009).

Differences in the manifestations of alexia and agraphia may reflect writing processes that are more vulnerable to even mild damage to a distributed network of language processing, and relatively more resilient reading processes. Although a shared-component model in which the same cognitive systems support both reading and spelling is appealing, the relationship between alexia and agraphia is not always straightforward. Cases of patients with mismatched alexias and agraphias complicate our understanding of the interplay between reading and writing disorders, such as patient RG, who presented with surface agraphia and phonological alexia (Beauvois and Derouesné, 1979); patient MK (Howard and Franklin, 1988), who presented with deep dysgraphia and surface alexia; and patient JC (Bub and Kertesz, 1982), who presented with deep dysgraphia and normal reading.

Central agraphias are a class of acquired writing disorders related to impairments in linguistic output, not motor output. In *lexical* or *surface agraphia*, patients make spelling errors for ambiguous and irregular words, while the spelling of regular words and nonwords is relatively intact. The errors made in lexical agraphia are usually phonologically plausible, and high-frequency words are written more accurately than low-frequency ones. This form of agraphia is believed to represent impairment at the level of graphemic output, reflecting damage to orthography. Patients are unable to retrieve orthographic information for some words, so they generate spellings by relying on the nonlexical route, producing phonologically plausible spellings. Errors are made in irregular and ambiguous words for which a 'sounding out' process is not sufficient to generate the correct spelling. Lexical agraphia is usually found in patients with lesions at the temporo-parieto-occipital junction, and is also common in patients with Alzheimer's and semantic dementia (Rapcsak and Beeson, 2002).

Phonological agraphia is a pattern of impairments in the spelling of unfamiliar words and nonwords, while familiar words remain comparatively intact, even those with irregular or ambiguous spellings. The spelling errors are usually phonologically implausible, but may be visually similar to the correct spelling. Some patients are unable to write even single letters when given sounds. Phonological agraphia is commonly associated with damage to the left perisylvian cortex, including Wernicke's area, the supramarginal gyrus, and Broca's area. Although this agraphia has been conceptualized as involving a specific deficit in converting phonemes to graphemes, or a disruption in the nonlexical spelling route, a large-scale investigation of phonological alexia and agraphia patients (Rapcsak et al., 2009) supported a model in which writing impairments are a manifestation of central, modality-independent damage to phonology, not the result of damage to dedicated spelling (or reading) cognitive systems. This finding is compatible with both connectionist and dual-route models of reading and spelling.

Like deep alexia, *deep agraphia* is characterized by frequent semantic errors, lacking any orthographic or phonological similarity to the target word (e.g., writing 'apple' instead of 'banana'). Patients with deep agraphia have marked difficulties spelling nonwords, make morphological errors, sometimes produce visually similar misspellings, and make functor substitution errors (e.g., *from* → *for*, *to* → *in*). Fragment errors also occur, more commonly toward the end of a word, involving the deletion of two or more letters. As in alexia, this condition is considered to be closely related to the phonological form, with the major distinction that semantic errors are present. Deep agraphia is typically found in patients with large left hemisphere lesions and extensive damage to the left perisylvian cortex, and is believed to represent damage to lexical-semantic and sublexical processes. One theory is that the pattern of errors seen in deep agraphia is the product of writing that originates from the right hemisphere (Rapcsak and Beeson, 2002). *Global agraphia* is an extension of the same damage, resulting in severe impairments in spelling all words.

Semantic agraphia refers to 'writing without meaning,' or a fundamental disconnect between semantics and spelling. Patients with semantic agraphia produce errors in spontaneous writing, but are usually able to write to dictation relatively well. Homophones are commonly misspelled in dictation, because the patient is unable to use the semantic context to select the correct spelling. The spelling of ambiguous words, irregular words, and nonwords is generally intact when patients do not know the semantic meaning. Patients with semantic agraphia typically have damage in brain regions that are associated with semantic processing, but not the area of the left perisylvian cortex associated with language.

Graphemic buffer agraphia is an impairment at the level of working memory, so that letters held in short-term memory are lost. Graphemic substitutions are frequently seen, such as substituting vowels for other vowels and consonants for other consonants. Since nonwords make higher demands of the graphemic buffer, they usually contain more errors.

Central agraphias can be conceptualized as a continuum, such that treatment aimed at specific disruptions in the dual-route model of writing can shift more severe agraphic impairments to less severe ones (Beeson et al., 2009). Intervention of deep agraphia might use a lexical approach, focused on retraining specific words, and following treatment gains the condition that would resemble the less severe phonological agraphia. A phonological intervention focused on retraining sound–letter relationships could shift the syndrome to a pattern of impairments more similar to surface agraphia, which could be treated by training interactive use of lexical and phonological knowledge, and eventually result in normal, unimpaired writing. An example of this theoretical continuum in action was observed in patient MM (Greenwald, 2004), whose global agraphia evolved to surface agraphia after experimental treatment targeted sublexical reading and sentence spelling.

Peripheral agraphias are a class of acquired writing disorders related to impairment in the physical production of letters, with the linguistic properties of words intact. *Allographic agraphia* is an impairment in selecting the correct letter shape, resulting in partial or complete interruption in writing the intended letter. It may result from failure to access specific letter shapes from memory, or an impairment of letter-specific graphic motor programs. Some patients may be better at uppercase characters than at lowercase or vice versa, and substitution errors for physically similar letter shapes are sometimes present.

Apraxic agraphia is characterized by impaired motor execution of letter strokes, not caused by sensorimotor, extrapyramidal, or cerebellar dysfunction. Patients with apraxic agraphia

usually hold the pen correctly, but normal writing motions are replaced by slow, effortful, and imprecise strokes, sometimes producing illegible handwriting. Other errors include omitting strokes, inserting unnecessary strokes, and spatial neglect (e.g., writing on only half the page). Oral and keyboard spelling remain intact. Some patients may produce single letters comparatively well (Zettin et al., 1995), and some patients may write numbers without impairment (Hodges, 1991). The variety of lesion sites associated with apraxic agraphia suggests that the motor programming of writing is distributed throughout the neural system, and may also include the cerebellum (De Smet et al., 2011).

Patients with *afferent agraphia* are unable to use sensory feedback to calibrate or control the fine motor execution of handwriting, presenting with duplications of letters, omissions of letters, and a perseverance effect for sequences of similar letters or strokes, such as writing 'bellll' for 'bell' or adding/omitting legs of an 'm.' Cursive handwriting is more severely affected. Patients may also have difficulty writing on a horizontal line or keeping appropriate spacing between words. Afferent agraphia is typically observed in patients with right parietal lobe damage. It can also be produced in normal individuals if kinesthetic feedback is interrupted.

Alexia and Agraphia in Other Scripts

To date, most research on acquired reading/writing disorders has focused on monolingual English speakers. Evidence from reading and writing disorders in other languages contribute to our understanding of a cognitive model of reading and writing processes.

The *orthographic depth hypothesis* theorizes that the process of oral reading should vary according to the level of transparency between letters and sounds. Orthographically transparent languages, such as Spanish or Italian, have regular and predictable pronunciations based on the orthography of words, indicating that relying on a sublexical orthographic–phonological reading mechanism would result in the correct pronunciation for almost any letter string. Orthographically opaque languages, such as French or English, have variable and irregular relationships between letters and sounds, necessitating a lexical–semantic reading mechanism to determine correct pronunciation. Following the orthographic depth hypothesis, semantic reading errors (reading *symphony* → *orchestra*) and regularization spelling errors (spelling *yacht* → *yot*) should not be observed in languages with transparent orthographies. However, convincing evidence of this hypothesis in its strong form has not been reported.

The sublexical reading pathway is selectively damaged in phonological and deep alexia, and these patterns of impairments have been found in opaque orthographies as well as in transparent ones, including Greek, Welsh, Slovak, and Japanese kana (Hricová and Weekes, 2012). The lexical reading pathway is selectively damaged in surface alexia, which has also been observed in transparent orthographies such as Italian and German. These patterns of impairments suggest that a dual-route model of reading exists even in languages with transparent orthographies. One interpretation is that semantic errors in transparent scripts may arise only following severe damage to the sublexical mechanisms, while in opaque scripts such as in English, these errors arise following even mild damage (Hricová and Weekes, 2012). In Russian, a relatively transparent script, agraphias are more commonly reported than alexias, perhaps because mappings from sound to print are less predictable and, therefore, more vulnerable to errors following damage than the reading system (Kozintseva et al., 2012).

A dual-route model of spelling can explain agraphia in many scripts, including differential performance for different languages in bilingual speakers and differential impact of brain lesions on the reading of logographic, syllabic, and alphabetic writing systems (Senaha and Parente, 2012). The *language independent hypothesis* predicts that bilingual patients with agraphia use a single spelling system for both languages, so that a similar pattern of impairments will result in both languages if that system is damaged. But case studies have not supported this hypothesis. Patient AA (Kambanaros and Weekes, 2013) was premorbidly fluent in both Greek, a stem-based language with a complex morphology, and English, a word-based language. AA showed similar patterns of errors in nonword spelling in both languages, but differential effects of grammatical word class for word spelling in each language. The authors posit that these differential errors resulted from damage to the lexical–semantic pathway, where morphological knowledge is stored. Grammatical category and morphosyntactic features are more prominent in stem-based languages, therefore grammatical knowledge had an impact on the orthographic retrieval of Greek nouns, showing evidence for spelling via language-specific processes. Evidence of word class dissociations in morphologically rich languages like Spanish and Italian also help reject the language-independent hypothesis. Although the same spelling system is used, impairment can be differential based on different morphological demands of different languages.

Testing Theoretical Models of Reading and Writing

Theoretical models of reading and writing can be tested using computational modeling, or executable computer programs that 'read aloud' following the information-processing procedures, a particular theory suggests. One way these models are tested is by damaging certain components of the model to simulate different patterns of alexia. The goal is to produce the same patterns of impairments using the model, as the behavioral data from real-world patients provide.

There are two main theoretical models of reading and writing. There is the symbolic *dual-route model*, consisting of two distinct pathways. Through the lexical route, a word's graphemes map onto a representation of its visual form in the orthographic lexicon, directly activating knowledge about its pronunciation in the phonological lexicon. Between the input and output lexicons, the semantic system stores knowledge about word meanings. The nonlexical route uses specific linguistic rules to convert each grapheme to its corresponding phoneme, which is the process used when 'sounding out' unfamiliar words or nonwords. The computational Dual Route Cascade (DRC) model has been able to account for many reading phenomena (see Coltheart, 2006b, for review), including surface alexia, phonological alexia, letter-by-letter

reading, and deep agraphia (Bormann et al., 2008). The existence of patients with each of the different routes selectively damaged or spared following brain injuries provides support for this model (Coltheart, 2006b).

A criticism of the dual-route model is that data is fitted to a theoretically derived a priori model, reducing its biological plausibility and validity (Seidenberg and Plaut, 2006). Despite this shortcoming, existing data currently favors the DRC model (Coltheart, 2006b).

The connectionist *triangle model* is driven by data instead of theory. In this model, recurrent connections are simultaneously activated between orthographic, semantic, and phonological units, and whatever information is available is used to produce output. For example, nonwords have no meaning units, so their pronunciation cannot be accessed using an orthographic–semantic pathway; they must be read using an orthographic–phonological pathway (Harm and Seidenberg, 2004). The Parallel Distributed Processing (PDP) computational model does not impose hard-wired a priori rules, increasing biological plausibility, but relies on algorithms that are unlikely to occur in nature (Seidenberg, 2006). PDP models have been able to simulate some forms of developmental dyslexia, acquired deep and phonological alexia, and letter-by-letter reading (Bormann et al., 2008; Chang et al., 2012; Welbourne et al., 2011).

A diverse array of acquired reading and writing impairments provide insight into cognitive models of language processing. Computational modeling and the study of language disorders in other scripts offer new horizons for the developing field of cognitive neuropsychology.

See also: Agraphia (Acquired Dysgraphia); Aphasia; Chinese Literacy Acquisition: A Multidimensional Puzzle; Dyslexia, Developmental; Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects; Neural Basis of Reading.

Bibliography

Basso, A., 2008. Recovery and treatment of Reading and spelling disorders. In: Stemmer, B., Whitaker, H. (Eds.), Handbook of the Neuroscience of Language. Academic Press, Burlington, MA, pp. 417–425.

Beauvois, M.F., Derouesné, J., 1979. Phonological alexia: three dissociations. Journal of Neurology, Neurosurgery, and Psychiatry 42, 1115–1124.

Beeson, P.M., Rising, K., Kim, E.S., O'Bryan, E., Rapcsak, S.Z., 2009. Examining a treatment continuum for acquired Agrahpia. In: 39th Clinical Aphasiology Conference, Held 26–30 May 2009 in Keystone, CO.

Bormann, T., Wallesch, C., Blanken, G., 2008. "Fragment errors" in deep dysgraphia: further support for a lexical hypothesis. Cognitive Neuropsychology 25 (5), 745–764.

Bub, D., Kertesz, A., 1982. Deep agraphia. Brain and Language 17, 146–165.

Buiatti, T., Skrap, M., Shallice, T., 2012. Left- and right-hemisphere forms of phonological alexia. Cognitive Neuropsychology 29 (7–8), 531–549.

Chang, Y., Furber, S., Welbourne, S., 2012. Modelling Normal and impaired letter recognition: implications for understanding pure alexic reading. Neurospsychologia 50, 2773–2788.

Cipolotti, L., Warrington, E.K., 1995. Semantic memory and reading abilities: a case report. Journal of the International Neuropsychological Society 1, 104–110.

Coltheart, M., 2001. Dyslexia (Acquired) and agraphia. In: International Encyclopedia of the Social and Behavioral Sciences. Elsevier Science Ltd, Kidlington, UK, pp. 3913–3918.

Coltheart, M., 2006a. Acquired dyslexias and the computational modelling of Reading. Cognitive Neuropsychology 23 (1), 96–109.

Coltheart, M., 2006b. Dual route and connectionist models of reading: an overview. London Review of Education 4 (1), 5–17.

Coslett, H.B., 1991. Read but not write 'idea': evidence for a third reading mechanism. Brain and Language 40, 425–443.

De Bleser, R., Luzzatti, C., 2008. Acquired dyslexia and dysgraphia. In: Mariën, P., Abutalebi, J. (Eds.), Neuropsychological Research: A Review. Routledge, New York, NY, pp. 113–126.

De Smet, H.J., Engelborghs, S., Pacquier, P.F., de Deyn, P.P., Mariën, P., 2011. Cerebellar-induced apraxic agraphia: a review and three new cases. Brain and Cognition 76, 424–434.

Fiset, D., Arguin, M., McCabe, E., 2006. The breakdown of parallel letter processing in letter-by-letter dyslexia. Neuropsychology 23 (2), 240–260.

Friedmann, N., Biran, M., Gvion, A., 2012. Patterns of visual dyslexia. Journal of Neuropsychology 6, 1–30.

Glosser, G., Friedman, R.B., 1990. The continuum of deep/phonological alexia. Cortex 26 (3), 343–359.

Greenwald, M., 2004. "Blocking" lexical competitors in severe global agraphia: a treatment of reading and spelling. Neurocase 10 (2), 156–174.

Harm, M.W., Seidenberg, M.S., 2004. Computing the meaning of words in reading: cooperative division of labor in visual and phonological processes. Psychological Review 111 (3), 662–672.

Hodges, J.R., 1991. Pure apraxic agraphia with recovery after damage of the left frontal cyst. Cortex 27, 469–473.

Howard, D., Franklin, S., 1988. Missing the Meaning? a Cognitive Neuropsychological Study of the Processing of Words by an Aphasic Patient. MIT Press, Cambridge, MA.

Hricová, M., Weekes, B.S., 2012. Acquired dyslexia in a transparent orthography: an analysis of acquired disorders of Reading in the slovak language. Behavioural Neurology 25, 205–213.

Johnson, R.L., Rayner, K., 2007. Top-down and bottom-up effects in pure alexia: evidence from eye movements. Neuropsychologia 45, 2246–2257.

Kambanaros, M., Weekes, B.S., 2013. Phonological dysgraphia in bilingual aphasia: evidence from a case study of greek and English. Aphasiology 27 (1), 59–70.

Kozintseva, E., Skvortsov, A., Ulicheva, A., Vlasova, A., 2012. Cognitive structure of writing disorders in russian: what would Luria say? Behavioural Neurology 25, 223–232.

Lambon Ralph, M.A., Ellis, A.W., Franklin, S., 1995. Semantic loss without surface dyslexia. Neurocase 1, 363–369.

Lambon Ralph, M.A., Hesketh, A., Sage, K., 2004. Implicit recognition in pure alexia: the saffran effect—a tale of two systems or two procedures? Cognitive Neuropsychology 21 (2–4), 401–421.

Leff, A.P., 2004. Alexia. Advances in Clinical Neuroscience and Rehabilitation 4 (3), 18–22.

Leff, A.P., Schofield, T., 2013. Rehabilitation of acquired alexia. In: Stone, J., Blouin, M. (Eds.), International Encyclopedia of Rehabilitation [Online] Available from: http://cirrie.buffalo.edu/encyclopedia/en/article/267/ (31.10.13).

Leff, A.P., Spitzyna, G., Plant, G.T., Wise, R.J.S., 2006. Structural anatomy of pure and hemianopic alexia. Journal of Neurology, Neurosurgery, and Psychiatry 77, 1004–1007.

Marshall, J.C., Newcombe, F., 1973. Patterns of Paralexia: a psycholinguistic approach. Journal of Psycholinguistic Research 2, 175–199.

Nickels, L., Biedermann, B., Coltheart, M., Saunders, S., Tree, J.J., 2008. Computational modeling of phonological dyslexia: how does the DRC model fare? Cognitive Neuropsychology 25 (2), 165–193.

Patterson, K., Plaut, D.C., McClelland, J.L., Seidenberg, M.S., Behrmann, M., Hodges, J.R., 1996. Connections and disconnections: a connectionist count of surface Dylsexia. In: Reggia, J.A., Ruppin, E., Berndt, R.S. (Eds.), Neural Modeling of Brain and Cognitive Disorders. World Scientific, New York.

Pflugshaupt, T., Gutbrod, K., Wurtz, P., von Wartburg, R., Nyffeler, T., de Haan, B., Karnath, H., Mueri, R., 2009. About the role of visual field defects in pure alexia. Brain 132, 1907–1917.

Rapcsak, S.Z., Beeson, P.M., 2002. Agraphia. In: Ramachandran, V. (Ed.), Encyclopedia of the Human Brain. Elsevier Science, USA, pp. 71–86.

Rapcsak, S.Z., Beeson, P.M., Henry, M.L., Leyden, A., Kim, E., Rising, K., Andersen, S., Cho, H., 2009. Phonological dyslexia and dysgraphia: cognitive mechanisms and neural substrates. Cortex 45, 575–591.

Riley, E.A., Thompson, C.K., 2010. Semantic typicality effects in acquired dyslexia: evidence for semantic impairment in deep dyslexia. Aphasiology 24 (6–8), 802–813.

Roberts, D.J., Lambon Ralph, M.A., Woolams, A.M., 2010. When does less yield more? the impact of severity upon implicit recognition in pure alexia. Neuropsychologia 48, 2437–2446.

Roeltgen, D.P., Lacey, E.H., 2010. Reading, writing, and their disorders. Continuum of Lifelong Learning and Neurology 16 (4), 59–68.

Saffran, E.M., Coslett, H.B., 1998. Implicit vs. letter-by-letter reading in pure alexia: a tale of two systems. Cognitive Neuropsychology 15 (1–2), 141–165.

Schuett, S., Heywood, C.A., Kentridge, R.W., Zihl, J., 2008. The significance of visual information processing and reading: insights from hemianopic dyslexia. Neuropsychologia 46, 2445–2462.

Schwartz, M., Saffran, E.M., Marin, O.S.M., 1980. Fractionating the reading process in dementia: evidence for word-specific print-to-sound associations. In: Coltheart, M., Patterson, K., Marshall, J.C. (Eds.), Deep Dyslexia. Routledge and Kegan Paul, London.

Seidenberg, M.S., 2006. Connectionist models of Reading. In: Gaskell, G. (Ed.), The Oxford Handbook of Psycholinguistics. Oxford University Press, Oxford.

Seidenberg, M.S., Plaut, D.C., 2006. Progress in understanding word reading: data fitting versus theory building. In: Andrews, S. (Ed.), From Inkmarks to Ideas: Current Issues in Lexical Processing. Psychology Press, Hove, UK.

Senaha, M.L.H., Parente, M.A., 2012. Acquired dyslexia in three writing systems: study of a Portuguese–Japanese bilingual aphasic patient. Behavioural Neurology 25, 255–272.

Shalev, L., Mevorah, C., Humphreys, G.W., 2008. Letter position coding in attentional dyslexia. Neuropsychologia 46, 2145–2151.

Sheldon, C.A., Abegg, M., Sekunova, A., Barton, J.J.S., 2012. The word-length effect in acquired alexia, and real and virtual hemianopia. Neuropsychologia 50, 841–851.

Spitzyna, G.A., Wise, R.J.S., McDonald, S.A., Plant, G.T., Kidd, D., Crewes, H., et al., 2007. Optokinetic therapy improves text reading in patients with hemianopic alexia: a controlled trial. Neurology 68, 1322–1930.

Starrfelt, R., Behrmann, M., 2011. Number reading in pure alexia—a review. Neuropsychologia 49, 2283–2298.

Starrfelt, R., Habekost, T., Gerlack, C., 2010. Visual processing in pure alexia: a case study. Cortex 46 (2), 242–255.

Starrfelt, R., Habekost, T., Leff, A.P., 2009. Too little, too late: reduced visual span and speed characterize pure alexia. Cerebral Cortex 19, 2880–2890.

Volpato, C., Bencini, G., Meneghello, F., Piron, L., Semenza, C., 2012. Covert reading of letters in a case of global alexia. Brain and Language 120 (3), 217–225.

Warrington, E.K., Cipolotti, L., McNeil, J., 1993. Attentional dyslexia: a single case study. Neuropsychologia 31 (9), 871–885.

Welbourne, S.R., Woollams, A.M., Crisp, J., Lambon Ralph, M.A., 2011. The role of plasticity-related functional reorganization in the explanation of central dyslexias. Cognitive Neuropsychology 28 (2), 65–108.

Zettin, M., Cubelli, R., Perino, C., Rago, R., 1995. Impairment of letter formation: the case of 'ideomotor' apraxia agraphia. Aphasiology 9, 283–294.

E

Early Education and Care in the World Regions

Emma Pearson, Universiti Brunei Darussalam, Gadong, Brunei

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Bennett, volume 6, pp. 3934–3939, © 2001, Elsevier Ltd.

Abstract

The purpose of this article is to provide broad-brush insights into current developments in early childhood care and education (ECCE) across the various and diverse regions of our world. These developments have given rise to a wide range of concerns, ranging from pragmatic issues in service delivery, funding and core principles for service provision to scholarly debates around pedagogical approaches, and philosophical orientation in programming. They also present opportunities for innovation and renewal through greater responsiveness to sociocultural needs and priorities in programming. The article uses examples of programs currently being implemented across some of the world regions to highlight some of these issues and opportunities. The article begins by highlighting some of the important and significant developments that ECCE has seen during the early years of this twenty-first century. It then details some major recent developments for ECCE in the world regions, along with some of the key challenges for ECCE that have emerged in recent years, as highlighted in the Moscow Framework for Action and Cooperation. These are illustrated through the use of reference to ECCE initiatives from countries within the world regions that might not commonly be represented in dominant scholarly circles. The article concludes with a brief summary of issues raised and possible steps forward in achieving high-quality, equitable ECCE provision for children inhabiting the world regions.

> Many things can wait; the child cannot. Now is the time his bones are being formed, his mind is being developed. To him, we cannot say tomorrow; his name is today Gabriela Mistral, Latin American poet and Nobel Prize winner.

According to rough estimates, a quarter of the world's 7 billion-strong human population is made up of children. During the past few decades, empirical evidence (Heckman, 1999; Shonkoff, 2010) to support sentiments articulated by Gabriela Mistral regarding the formative impact of early childhood care and education (ECCE) has led to a widening spread of interest and investment in young children's well-being, care, and education. These developments mark a significant period in the histo ry of ECCE.

The goal of this article is to provide a brief overview of recent developments that have occurred as part of the recent focus on ECCE, with particular attention to some key contemporary issues affecting policy makers, practitioners, and scholars from across the world regions. The entry begins with context setting, drawing particular attention to dissonance between the evidence base commonly referred and contributed to by scholars of early childhood and the reality of everyday life for the vast majority of children across the world regions. Current areas of development in formalized ECCE that are common across the world regions are then summarized, with examples of programs in place across parts of the world at the time of this writing used to highlight some contemporary issues in ECCE. Of particular concern, evidenced across recent discussions in both scholarly literature and in wider forms of knowledge sharing, is the importance of acknowledging alternative approaches to provision that are sufficiently dynamic or fluid to adequately represent diverse programs and realities. The entry concludes by summarizing some recent recommendations and implications for the development of ECCE across the world regions into the second quarter of the twenty-first century.

ECCE in the Twenty-First Century – Context Setting

Terminology used to refer to the care, development, and education of young children varies across countries and contexts, in most cases as a reflection of particular stances on early childhood. For example, in sub-Saharan Africa, preference for the term early childhood development (ECD) reflects a view that early childhood provision should be centered solely on the child, rather than on any particular method of care or model of education. In response to a historical distinction between 'care' and 'education' in European American contexts that has often led to 'siloed' approaches to ECCE policy and practice (reflecting a general view that children are cared for until a point at which they are ready to be formally schooled), the United Nations Educational, Scientific and Cultural Organization (UNESCO) adopts the term ECCE, which connotes a more holistic, intersectoral approach to early childhood programming that attends to all aspects of health, care, and education of children from birth to the age of 8 years. The term ECCE is widely recognized across the world regions and, where it has not been formally adopted, it is acknowledged in relation to regional definitions, as in the case of sub-Saharan Africa.

In providing an overview of ECCE across the world regions, it is important to begin by drawing attention to certain peculiarities in the current global context of ECCE. In scholarly circles, most discussions of children's development, care, and education tend

to assume at least a basic level of service provision, centering on expanding evidence and/or considering how to improve existing care and education systems mostly in operation across European American contexts (see, for example, recent issues of *Early Childhood Research Quarterly*, a leading journal in the discipline). However, it is important to note that at a global level, the early experiences of and availability of supports for children differ vastly, even in the twenty-first century. Data provided by the United Nations Children's Fund provide stark reminders that many children around the world regions still lack access to basic survival needs: an estimated 8.8 million children continue to die before they celebrate their fifth birthday, more than 70 million children are still out of school, and millions more lack protection against violence, abuse, exploitation, discrimination, and neglect (UNICEF, January 2012). Although international literature on ECCE increasingly recognize social and cultural diversity, as well as basic survival needs among some populations of children (Engle et al., 2007), it is important to acknowledge the dissonance between conditions commonly reflected in scholarly circles and the daily reality of ECCE experiences for many children across the world regions.

It is important also in reviewing ECCE across the world regions in the twenty-first century, to acknowledge an extensive historical background of studies and discussions around children, their place in society, and the nature of their growth and development. From the Puritans to Rousseau and the Romantics in European contexts, to Confucius in Chinese contexts, beliefs about children and childhood have been expressed throughout millennia. According to Wu (1996), essays on child rearing were written by Confucian scholars throughout the first century BC and by the time of the Ming (1360–1600s) and Qing (1600–1900s) dynasties, parenting 'manuals' were widely read. The value of children has also long been acknowledged in Islamic texts, which emphasize responsible child-rearing (Hasna, 2003).

ECCE in the Twenty-First Century – Recent Developments and Global Concerns

Evidently, societies and communities have cared for and educated their children across millennia. Formalization of ECCE provision at the level of global conventions reflected in national policies, however, is a relatively new phenomenon. Indeed, policy focus on ECCE has increased with unprecedented rapidity during recent years, due in part to (1) near-universal ratification of the United Nations Convention on the Rights of the Child and (2) international frameworks for action agreed upon by United Nations (UN) member states, such as Education for All (EFA) and the Millennium Development Goals (MDGs). EFA, which consists of a series of goals designed to promote universalization of quality basic education, cites ECCE as a key contributor to economic and educational development:

> The preconditions for education quality, equity and efficiency are set in the early childhood years, making attention to early childhood care and development essential to the achievement of basic educational goals. (Article 1.3)

In European American contexts toward the end of last century, evidence of the positive impact of preschool programs such as the High-Scope Perry Preschool Program in the United States, particularly for overcoming the effects of poverty and social disadvantage (Barnett, 1998), led to a raft of studies that highlighted the economic sense in investing in ECCE (Heckman, 1999) and a proliferation of ECCE initiatives. At a global level, also of significance in fueling the expansion of ECCE programs in both the majority world and European American contexts, has been the work of Young (1996) and Shonkoff and Philips (2000), whose research has established strong evidence of links between quality early childhood interventions, neurological growth, and lifelong development. Such work has been cited across the world regions by organizations, such as United Nations Children's Fund (UNICEF) and UNESCO, in driving initiatives for early childhood program provision to improve living conditions for young children.

Spurred by a combination of published scientific evidence and international goal setting, during the 1990s and the early 2000s, ECCE attracted intense interest across the world's diverse regions. Following a period of subsequent expansion, in recent years, discussions and research in ECCE across the globe have turned toward understanding complexities and implications of program provision and development. Challenges and opportunities related to the range of contextual factors influencing policies and programming have given rise to a greater focus on providing high-quality care and education, rather than simply attending to quantitative goals.

In response to this heightened awareness of the need to 'take stock,' reaffirm, and plan for the future of ECCE across the world regions, UNESCO sponsored its first World Conference on Early Childhood Care and Education in 2010. The Conference brought together 50 government ministers from across the world regions, as well as scholars, practitioners, and other high-level policy makers/advocates. In preparation for the Conference, regional reports from Africa, Arab countries, Asia and the Pacific, Europe and North America, Latin America and the Caribbean, and Russia were produced. These reports are now published on the UNESCO website (http://www.unesco.org/new/en/world-conference-on-ecce/reports/). Outcomes of the Conference are also published on the UNESCO website in the Moscow Framework for Action and Cooperation, which lists a set of goals and principles for steps forward in provision of formalized ECCE across the world regions. These goals, agreed upon by ministers, heads of delegations, representatives of UN agencies, development cooperation agencies, civil society organizations, education agents, and early childhood care scholars, represent what is arguably the most representative and accessible synthesis of current developments and concerns in ECCE across the world regions. The goals are likely to influence developments in ECCE for the foreseeable future and are therefore summarized briefly as a starting point for broader discussion.

The Moscow Framework for Action and Cooperation: Goals for ECCE across the World Regions

The Moscow Framework for Action and Cooperation outlines the following key areas of concern/focus in global future ECCE plans: Legislation, Policies and Strategies; Access and Scaling Up;

Assessment, Research, Monitoring and Evaluation; Equity and Inclusion; Quality; Capacity Building and Partnerships. In terms of legislation, policies and strategies, a common concern across the world regions is that while ECCE policies exist in many countries, they are not always included in national development plans and therefore lack political commitment/funding. ECCE policies also tend to lack clear operational frameworks to support implementation and sustainability. Countries also share, as reflected in goals outlined by The Framework, concern about the need to ensure access to programs for all children, regardless of geographical location and/or ability. In many countries, access to ECCE programs is greater for children living in relatively prosperous areas, defeating the purpose of providing support for the most vulnerable.

As countries focus increasingly on quality in tandem with quantity and access, the need for developing assessment and monitoring approaches based on contextually appropriate research evidence has become an urgent one across the world regions. As the Moscow Framework for Action and Cooperation suggests, a current focus for many countries around the world regions is on working with practitioners, policy makers, researchers, and communities to ensure that ECCE programs are built around localized resources, strengths, and knowledge bases.

All the key areas noted in the Moscow Framework for Action and Cooperation and referred to above are, to varying degrees and by different means, universal in their relevance. They are also, notwithstanding the important role of sociocultural contexts in shaping their unique nature and influence on ECCE within and across each of the world regions, strongly interwoven. Across numerous countries around the world, from Australia to Zanzibar, there is a strong focus on building into ECCE policies the need for provision of programs for children that are equitable, of high quality, and best suit national and community needs and priorities. The next section provides specific examples of ECCE initiatives from Oceania, Asia, and Africa that highlight strategies through which programs are responding to these goals for ECCE. While each has a unique purpose, all illustrate contemporary issues related to complexities around identifying the unique, contextual considerations to be made in the process of formulating policies, providing access, building capacity in working with communities, and monitoring quality of provision. Perhaps most importantly, these examples provide some insight into the extent to which currently available international literature in ECCE are lacking in evidence to support necessary innovations in ECCE provision across the world regions.

ECCE in the Twenty-First Century: Snapshots from Vanuatu, the Philippines, and Djibouti

Quality in ECCE – Teacher Preparation in Vanuatu

Of particular current interest across a wide range of contexts is the issue of quality in ECCE, from both conceptual and pragmatic points of view. In line with EFA goals, universal provision of ECCE has been a focus of countries around the world for some time. However, attention to issues regarding quality has emerged recently, with increasing evidence that low-quality services may at best offer little in the way of positive outcomes for young children and, at worst, be detrimental (UNESCO, 2004). The 2005 EFA Global Monitoring Report on the 'Quality Imperative' signaled a new focus on process features, in addition to quantitative factors, in ECCE provision. This focus is most recently outlined in the Moscow Framework for Action and Cooperation.

Of all the potential contributors to 'quality' in ECCE programs designed to support early learning and development, teacher quality is widely acknowledged as having the greatest influence. Without the presence of committed, confident, and knowledgeable human resources, the most advanced physical facilities are unlikely to engage young children in meaningful learning experiences (Alexander, 2008). In line with cultural–contextual expectations regarding professionalism, ECCE teacher quality in European American contexts is widely associated with formalized teacher education and tertiary-level qualifications (Degotardi, 2010). Applied to resource-constrained environments, where tertiary-level education is still a privilege, the direct implication of this would be that high-quality ECCE programming and implementation across many contexts is a virtual impossibility.

The following example of teacher training in Vanuatu serves as a useful illustration of complexities around expansion of ECCE, in terms of dynamics that need to be factored into policy making, attempts to scale up programs, and capacity building. A small Pacific island rich in cultural and natural resources but less advantaged economically, Vanuatu has developed some noteworthy and innovative ECCE programs that serve as useful illustrations of the need for consideration and promotion of locally bound definitions of 'quality' in teacher preparation. Vanuatu's example of teacher training supports Alexander's (2008) position that including contextually appropriate indicators of process in the design and evaluation of teaching is critical to achieving a 'quality' education system that engages and retains teachers and children.

Although not supported significantly in terms of public funding, ECCE in Vanuatu benefits from a strong, committed team of practitioners working within the national preschool association, PSABV (Pri-skul Asosiesen Blong Vanuatu). As a result, Vanuatu has recently seen some important developments in ECCE, including the development of a draft ECCE curriculum and incorporation of ECCE in the national definition of basic education. Current initiatives reported to be undertaken by the government include the Vanuatu Education Strategy (2008–09); Vanuatu Education Roadmap (2010–16), which looks at access, quality, and management components; as well as the Vanuatu Early Childhood Education Policy, which was launched in 2010 (James, 2012).

Much of the formalized teacher preparation that is available for ECCE teachers in Vanuatu is offered by external agencies, in the form of a bachelor's degree in Early Childhood Education, for which teachers must obtain an international scholarship and travel to Fiji, or a Certificate III in Children's Services (Childcare/Preschool Stream) course offered by the Australia-Pacific Technical College. The Vanuatu Preschool Association also offers short-term, intensive training with support from the Ministry of Education in Vanuatu. Teachers graduating from this training are then employed in community-based ECCE centers that provide programs for children aged 4 and 5 years, the preprimary school years.

A brief description of the Preschool Association/Ministry of Education teacher training is provided below. The applicability

of this training, which provides little of the theoretical knowledge and engagement with scientific research commonly required of professional training, is nonetheless self-evident. Anecdotal evidence also suggests that the local ni-Vanuatu training model results in positive outcomes, with trained teachers more likely to use the vernacular in their teaching (a priority for Vanuatu), more likely to engage in child-centered activities, and more likely to involve parents and community in children's learning (UNICEF Pacific and USP, 2004). An added benefit of this training program is that it has led to a strong national network facilitating direct communication between the Ministry Director for ECCE and teachers working across both proximal and remote regions.

The training typically consists of a 5-week program (Guild and Tuivaga, cited in James, 2007) that is first delivered to 'key teachers' at national level. These key teachers then implement the same training program to ECCE teachers working across the provinces. Among the many benefits associated with this training program is the close network of teachers that has developed across Vanuatu as a result of the links made between key teachers, the national Preschool Association, the Ministry of Education, and 'local' teachers who are located at provincial and village level.

A major aspect of the 5-week training program involves learning about construction of preschool buildings and equipment using readily available natural resources. When the key teachers travel to the provinces to conduct training for local teachers, they are accompanied by a builder who will then assist the local teachers and local communities in building new preschool facilities, where necessary. Teachers also receive training on how to make use of local materials to construct learning materials for the classroom.

Teachers also learn how to make a collection of learning materials using natural resources (including natural dye for colors). Ongoing supervision is provided to teachers by both provincial-level and national ECCE practitioners, who make regular visits to preschools. As Jenny James (Early Childhood Education Director at the Ministry of Education) explains:

> One of the commonest complaints from ECCE teachers is the lack of resources for teaching. Such a complaint need not be heard again…(i)n the Pacific, where our surroundings are full of natural resources and our communities are caring, we can count ourselves very fortunate. Where children are in need, and teachers and communities have the skills and imagination to invent, there can be no truer saying than necessity is the mother of invention. (James, 2007: p. 122)

Hess et al. (2004) have outlined the complexities involved in connecting teacher training to quality teaching. Scholarly debates around the nature and length of preparation required for quality teaching continue, particularly in the field of ECCE (ref). Given the current expansion in formalized ECCE programs across the world regions and contemporary calls for greater attention to quality provision, Vanuatu's case highlights a degree of urgency in the need for conceptualizations of quality that combine universalized standards with consideration of unique sociocultural, political, and economic contextual circumstances (Box 1).

Box 1 Case Study – ECCE Teacher Preparation in Vanuatu Extracted from James, 2007. Resources for ECCE. In: Puamau, P., Pene, F. (Eds.), Early Childhood Care and Education in the Pacific. The Pride Project, Pacific Education Series No. 5. USP, Suva, Fiji.

Training in Vanuatu for most ECCE teachers typically consists of a 5-week training programme (Guild and Tuivaga, cited in James, 2007) that is first delivered to 'key teachers' at national level. The key teachers then implement the same training programme to ECCE teachers working across the provinces. Among the many benefits associated with this training programme is the close network of teachers that has developed across Vanuatu as a result of the links made between key teachers, the national preschool association, the Ministry of Education, and 'local' teachers who are located at provincial and village level.

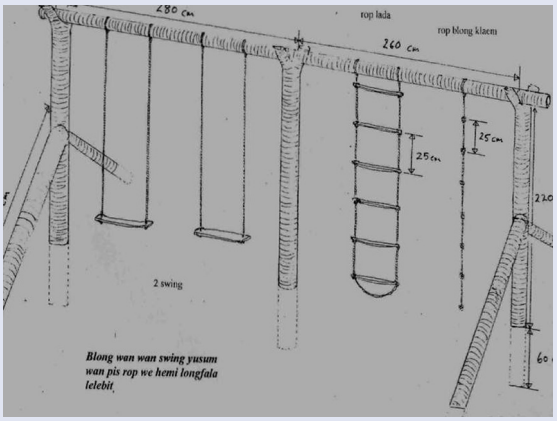

Jenny James (Early Childhood Education (ECE) Director at the Ministry of Education) explains:

One of the commonest complaints from ECCE teachers is the lack of resources for teaching. Such a complaint need not be heard again…(i)n the Pacific, where our surroundings are full of natural resources and our communities are caring, we can count ourselves very fortunate. Where children are in need, and teachers and communities have the skills and imagination to invent, there can be no truer saying than necessity is the mother of invention.

A major aspect of the 5-week training that ECCE teachers receive involves learning about construction of preschool buildings and equipment using readily available natural resources. When the key teachers travel to the provinces to conduct training for local teachers, they are accompanied by a builder who will then assist the local teachers and local communities in

building new preschool facilities, where necessary. Teachers also receive training on how to make use of local materials to construct learning materials for the classroom.

Teachers learn how to make a collection of materials using natural resources (including natural dye for colors). Ongoing supervision is provided to teachers by local and national ECCE key teachers, who make regular visits to preschools.

Reaching the Unreachable: Mobile Early Childhood Care and Development Programs in the Philippines

The Philippines has a well-developed system of ECCE provision, due partly to its enactment of the Early Childhood Care and Development (ECCD) Law in 2000, which required the establishment of a National Coordinating Council for the Welfare of Children. A major focus for the Philippines, in line with EFA goals, is to increase access to ECCE programs to children in remote and conflict-affected communities. Organizations and groups working in the Philippines have therefore focused on developing innovative solutions to challenges associated with access. Among these is a 'mobile resource center,' implemented in the Philippines in 2005. The resource center consists of a roving vehicle (horses are also widely used in hard-to-access regions) driven by an ECCE worker equipped with learning materials, manipulative toys, musical instruments, health and nutrition education materials, and supplies. According to UNICEF reports, the ECCE worker is trained in ECCE, uses a curriculum adopted in national day care services, and is paid for by Local Government Units (F. Nogra-Abog, personal communication, 7 November 2010) (Box 2).

Box 2 Mobile Playgroups in Hard-to-Access Regions of the Philippines Extracted from Nogra-Abog, F., 2010. Reaching the Unreached: Alternative Ways of Delivering Early Childhood Education in the Philippines. (Powerpoint slides). UNICEF, Philippines. Extracted from Nogra-Abog, F., 2010. Reaching the Unreached: Alternative Ways of Delivering Early Childhood Education in the Philippines. (Powerpoint slides). UNICEF, Philippines.

This service has provided children with the opportunity to play, socialize, and learn. It provided a safe place while parents are at work. It became the venue where children were able to access basic health services. The parent education services, on the other hand, have increased their awareness on proper childcare. They assist the ECCD workers by helping produce indigenous materials and integrate cultural poems, songs, and dances in the daily activities. The appreciation of parents and children overwhelmed and boosted the morale of community leaders. It also made them realize the need to increase basic services in their communities. Experience from this initiative shows that alternative early childhood education services are effective in reaching the unreached and that community leaders are able to provide and sustain innovations if 'shown the way.' Key to this is demonstrating that the innovation is workable, effective, and affordable. Affectionately known as 'ECCD-on-horse' (the Philippines prefers to incorporate the 'd' for development rather than education in its conceptualization of early childhood programs), the mobile resource center is specifically aimed at providing ECCE services to children in the four remote and conflict-affected communities of Buldon and South Upi in the province of Maguindanao. These communities are characterized by extreme poverty, absence of basic services, rugged and mountainous terrain, and sporadic occurrence of clan wars and armed conflict between the military and rebels. The vehicle/horse visits the same 10 communities at least 3 times a week for 10 consecutive weeks. After leading play and learning activities with children, facilitators meet with parents to conduct parenting sessions on life-saving and effective parenting information. The program, which is designed to fit with a holistic approach to ECCE, also serves as an entry point for delivering other services such as supplementary feeding, vaccinations, as well as physical and dental examinations, all funded by Local Government Units (F. Nogra-Abog, personal communication, 7 November 2010).

Mobile ECCD programs were also initiated in Davao City, with support from the Australian Government Overseas Aid Program in late 2008. By June 2009, the two mobile ECCD programs of Davao City were expected to have reached a minimum of 1000 children from 26 of the most depressed and densely populated communities in Davao City. Formal assessments of quality in mobile ECCD services have not as yet been conducted, due in part to a lack of availability of relevant, conceptually grounded, or validated tools. However, in terms of quantity, they have reached a total of 7241 children in less than 2 years of implementation and have yielded positive anecdotal feedback.

According to initial assessments conducted by UNICEF (Nogra-Abog, 2010), the mobile programs have provided children with valuable opportunities to play, socialize, and learn, as well as a safe place to be cared for while parents are at work. The integrated nature of these services has led to increased access to basic health services, increased parental awareness about appropriate childcare, and increased parental involvement in programs. For example, parents have assisted ECCD program facilitators by helping to produce indigenous materials and integrate cultural poems, songs, and dances in the programs' daily activities. The enthusiasm of parents and children has resulted in ripple effects in local communities, where community leaders have been motivated to commit to increasing basic services in their communities. Nogra-Abog (2010) notes that a key outcome of these programs has been acknowledgment of the importance of innovative approaches to providing childhood education services for effective efforts in reaching children who are living in communities where access has presented challenges, and also, that community leaders are eager to contribute to and sustain innovations if "presented with workable 'tools'." Key to this is demonstrating that an innovation is pragmatic, effective, and affordable.

This snapshot of programs designed specifically to provide access to ECCE programs for children in 'unreached'

communities (a focus of many countries around the world regions currently, as reflected in the Moscow Framework for Action) provides an illustration of the limitations in measures of quality currently supported in international literature. Such measures tend to assume that children attend programs that are located in a permanent physical structure and offer curricula designed to support the development of specific skills/learning dispositions (Fenech, 2011; Pearson, 2011). A number of recent publications have hence called for the development of more nuanced measures that support contextually based approaches to defining quality programming for young children with varied needs across diverse contexts (Myers, 2004; Woodhead, 2006).

Lessons from Djibouti: Grandmothers Promote Exclusive Breastfeeding as the Key Communication Actors

ECCE, as defined by UNESCO (2006), refers to the period from birth to 8 years of age. Although ECCE policies tend to cater to children older than 3 years, in many countries, there is increasing focus not only on infants and toddlers but also on further acknowledgment of the importance of pre- and perinatal care as part of ECCE (Engle et al., 2007). Involving parents and communities, as well as providing more formalized interventions in supporting early development and learning, has therefore become increasingly important.

In order to improve breastfeeding rates in Djibouti, community-based programs for awareness raising were implemented in 2007. Prior to program intervention, qualitative research was conducted in order to identify (1) factors linked to low breastfeeding rates and (2) 'key actors' likely to be able to influence new mothers. Based on results from the study, grandmothers were invited to participate in the program as trainees and, subsequently, trainers, in order to solve issues around social norms and values that might impede breastfeeding.

As part of the program, grandmothers (initially 11, but ultimately 156 in total) received training from nurses and staff at the Ministry of Health. Training was delivered through the use of pictures and alternative materials to overcome literacy barriers. Trained grandmothers then provided support and training to mothers through group sessions and home visits to mothers of children younger than 5 years. In addition, communication materials were developed with involvement from grandmothers (including songs and theater shows) and broadcast through national radio.

UNICEF reports that although breastfeeding rates remain relatively low in Djibouti, this intervention program has resulted in significant increases in breastfeeding (25% from just over 1% before 5 years). The program is also used to promote a 'Communication for Development' approach to working with communities, where it is often found that involving key community members in communicating health-related messages can significantly enhance the impact of social mobilization activities.

Summary

These snapshots of ECCE-related programs and initiatives currently being implemented across the world regions provide a glimpse into the rich diversity in programming that exists in the increasingly globalized world of ECCE. They also reflect growing awareness that, in order to address some of the key challenges and goals that have been outlined by representatives from across the world regions in the Moscow Framework for Action, innovative solutions are required to resolve complex challenges in providing ECCE to all children. Each 'case' also highlights the importance of reaching out to and involving communities in ECCE, particularly in establishing programs where financial resources may be constrained.

As a number of well-respected international experts in the field of ECD and education have argued, understandings of what is 'best' for children vary based on local customs, resources, family and political structures, values, and beliefs (Nsamenang, 2009; Woodhead, 2006). The range of opportunities for innovation through discovery and development of existing resources, both human and otherwise, (1) to define 'quality' in ECCE and (2) to put in place quality programs that support children, their families, and communities represents one of the most exciting aspects of ECCE in the present time.

The implications, however, in terms of developing policies that adequately respond to national and community needs, as well as ensuring high quality through monitoring and assessment for ECCE across the world regions, are challenging. There can be no question that quality in ECCE provision is paramount, both for the well-being of young children and if investments are to result in significant returns in the form of well-prepared and productive future citizens. However, current approaches, tools for assessment, and understandings of what is 'best' for young children are dominated by an 'international' body of scholarly literature that reports largely on research conducted in Western Europe and North America (Fenech, 2011). Of considerable value to developments in ECCE across the world regions, therefore, would be greater proliferation, acknowledgment, and dissemination of rigorous, research-based evidence to support the development of diverse ECCE programs that respond to varied priorities and circumstances.

Conclusions

As we move toward the middle of this 'new' century, ECCE policies and programs across the world regions are faced with both exciting opportunities and formidable challenges. There is shared global agreement about contemporary priorities (reflected in the Moscow Framework For Action and Cooperation), as well as increased acknowledgment of the need for 'localized' approaches to uphold principles of access and equity.

With increased focus on ensuring high-quality provision in programs that support young children and their families/communities, there are opportunities for innovation and knowledge sharing across the world regions. It is important, however, that such sharing is clearly representative of diverse experiences and priorities, avoiding widespread domination of ECCE constructs, tools, and ideals that emanate from particular world regions (Nsamenang, 2009). For example, given the current global expansion of focus in ECCE and, importantly, increased attention to the role of cultural sustainability in young children's learning development, efforts to build and

uphold quality in formalized ECCE do require a more truly globalized evidence base. As Myers (2006: p. 31) indicates:

> If the characteristics and processes of an educational environment are not consistent with the kind of world, country and citizenry desired, as set out in national documents and curricula, and/or with widely agreed-upon rights for children, it is difficult to say that the program is of high quality.

To return to the opening quote, now is the time for supporting the positive development and learning of young children across the world regions. It is also the time for protecting children's long-term livelihoods by ensuring that the formalized programs put in place to provide such supports also uphold wider principles associated with sociocultural sustainability and community development.

See also: Comparative Research in Education: Iea Studies; Cross-Cultural Study of Education; Education and Learning: Lifespan Perspectives; Education in South Asia; Education in Sub-Saharan Africa; Education in the Middle East and North Africa; Family and Schooling; International Organisations and Education; Social Inequality and Schooling; Sustainable Development: An Economic Perspective.

Bibliography

Alexander, R., 2008. Education for All, the Quality Imperative and the Problem of Pedagogy. CREATE Research Monograph No. 20. University of Sussex, Brighton.

Barnett, S.W., 1998. Long-term cognitive and academic effects of early childhood education on children in poverty. Preventive Medicine 27 (2), 204–207.

Degotardi, S., 2010. High-quality interactions with infants: relationships with early childhood practitioners' interpretations and qualification levels in play and routine contexts. International Journal of Early Years Education 18 (1), 27–41.

Engle, P., Black, M.M., Behrman, M.C., Gertler, P.J., Kapiriri, L., Martorell, R., Young, M., 2007. Strategies to avoid the loss of developmental potential in more than 200 million children in the developing world. The Lancet 369, 229–242.

Fenech, M., 2011. An analysis of the conceptualisation of 'quality' in early childhood education and care empirical research: promoting 'blind' spots' as foci for future research. Contemporary Issues in Early Childhood 12 (2), 102–117.

Hasna, F., 2003. Islam, social traditions and family planning. Social Policy & Administration 37 (2), 181–197.

Heckman, J., 1999. Policies to Foster Human Capital. NBER Working Paper Series, No. 7288.

Hess, F., Rotherham, A.J., Walsh, K., 2004. A Quality Teacher in Every Classroom: Appraising Old Answers and New Ideas. Harvard Education Press, Cambridge, MA.

James, J., 2007. Resources for ECCE. The Pride Project, Pacific Education Series No. 5. In: Puamau, P., Pene, F. (Eds.), Early Childhood Care and Education in the Pacific. USP, Suva, Fiji.

James, J., 2012. Vanuatu Early Learning Standards. Presented at the Pacific ECCE Directors' Meeting, Nadi, Fiji, February 29–March 3. UNICEF Pacific.

Myers, R.G., 2004. In Search of Quality in Programmes of Early Childhood Care and Education (ECCE): Paper Prepared for the EFA Global Monitoring Report 2005. UNESCO, Paris. Retrieved from: http://unesdoc.unesco.org/images/0014/001466/146677e.pdf.

Myers, R.G., 2006. Quality in program of early childhood care and education (ECCE): Background paper prepared for the Education for All Global Monitoring Report 2007. UNESCO, Paris. Retrieved from http://unesdoc.unesco.org/images/0014/001474/147473e.pdf.

Nogra-Abog, F., 2010. Reaching the Unreached: Alternative Ways of Delivering Early Childhood Education in the Philippines (Powerpoint slides). UNICEF, Philippines.

Nsamenang, A.B., 2009. A critical peek at early childhood care and education in Africa. Child Health and Education 1 (1), 44–55.

Pearson, E., 2011. Avoiding recolonisation in early childhood: promoting local values as mediators in the spread of globalisation. Contemporary Issues in Early Childhood 12 (3), 212–223.

Shonkoff, J.P., 2010. Building a new biodevelopmental framework to guide the future of early childhood policy. Child Development 81 (1), 357–367.

Shonkoff, J.P., Philips, D.A., 2000. From Neurons to Neighbourhoods: The Science of Early Childhood Development. National Academic Press, Washington, DC.

Stephens, D., 2003. Quality of Basic Education. Paper prepared for the United Nations Educational, Scientific and Cultural Organization (UNESCO) EFA Monitoring Report Team. UNESCO, Paris.

Woodhead, M., 2006. Changing Perspectives on Early Childhood: Theory, Research and Policy. UNESCO, Paris. Retrieved from: http://unesdoc.unesco.org/images/0014/001474/147499e.pdf.

Wu, D.Y.H., 1996. Chinese childhood Socialization. In: Bond, M.H. (Ed.), The Handbook of Chinese Psychology. Oxford University Press, Hong Kong, pp. 143–154.

Young, M.E., 1996. Early Child Development: Investing in the Future. World Bank, Washington, DC.

Relevant Websites

UNESCO, 1990. World Declaration on Education for All. http://www.unesco.org/education/pdf/JOMTIE_E.PDF.

UNESCO, 2004. EFA Global Monitoring Report: Education for All: The Quality Imperative. http://www.unesco.org/new/en/education/themes/leading-the-international-agenda/efareport/reports/.

UNESCO, 2006. EFA Global Monitoring Report 2007. Strong Foundations: Early Childhood Care and Education. http://www.unesco.org/new/en/education/themes/leading-the-international-agenda/efareport/reports/.

UNESCO, 2010a. Moscow Framework for Action and Cooperation: Harnessing the Wealth of Nations. http://www.unesco.org/new/en/world-conference-on-ecce/regional-information/.

UNESCO, 2010b. Early Childhood Care and Education Regional Report–Africa. http://www.unesco.org/new/en/world-conference-on-ecce/regional-information/.

UNICEF Pacific and the University of the South Pacific (USP), June 2004. Pacific Regional Conference on Early Childhood Development. http://www.unicef.org/pacificislands/ECD_Workshop_Report.pdf.

UNICEF, 2011 December. Sharpening the Equity Focus: Selected Innovations and Lessons Learned from UNICEF-assisted Programmes, 2009–2010. http://www.unicef.org/innovations/files/compendium_Jan9lowres.pdf.

UNICEF, 2012 January. Statistics by Area: Education. http://www.childinfo.org/education.html.

Early Emotional Development and Cultural Variability

Fantasy T Lozada, University of Michigan, Ann Arbor, MI, USA
Amy G Halberstadt, North Carolina State University, Raleigh, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The goal of this article is to consider how culture impacts the nature of early emotional development. First, we review broad developmental trajectories of emotional development through the integration of biological and social system changes, embedded in the context of culture. Second, we discuss several cultural frameworks with which to examine cross-cultural comparisons of emotional development and how these frameworks impact parental socialization strategies and subsequent emotional outcomes in infancy and early childhood. Throughout, we review recent empirical work that supports the manifestation of cross-cultural similarities and differences in early emotional development.

Introduction

Emotion is a complex yet integral part of the human experience. Individuals' emotions and those of others around them are salient in reactions to everyday experiences, motivations toward specific behaviors, and engagement with others. Furthermore, the development of emotion is intertwined with other important systems, including those associated with physiological, cognitive, and social change. For infants and young children, emotional competence is an important achievement, yet the expectations about and emphasis on emotional competence are embedded within the beliefs and norms of the cultural, social, and physical structures in which children are raised. This article considers the nature of emotional development in infancy and early childhood in the context of culture. When we refer to emotional development, we are thinking of a dynamic, tripartite system comprised of emotional experience, emotional expression, and emotion understanding; these also have reverberating effects upon each other and across other systems (Halberstadt et al., 2001). Emotion experience and expression appear indistinguishable at birth, but over the first year of life can begin to diverge as infants become aware of the expectations of others and also attempt to influence those others. For the sake of brevity, we will discuss emotional experience and expression together, except when noting research that is able to differentiate their trajectories.

The Underpinnings of Emotional Development

Emotional development is predicated on both biological and social system changes, which are themselves intertwined. Specifically, rapid postnatal development of brain structures (e.g., the limbic system, brain stem, and prefrontal cortex) and hormonal influences contribute to the activation and regulation of emotion expression. Throughout the first few years of life, brain structures such as the cerebral cortex develop and exert increasing control over and sometimes affect change in the limbic system, resulting in fewer mood swings and greater emotional regulation. The young child's increased sophistication in emotional communication and regulation is evidence of the emergence of these biological brain systems (see Easterbrooks et al., 2013, for a review).

The biological changes that support emotional development also occur in the context of developing caregiver–child social relationships in which the majority of early childhood emotional experiences occur. For example, infant distress may be expressed to get a caregiver's attention and to get needs met; once recognized, negative emotions may be met with caregiver attempts to soothe or comfort the child. Infant experiences of positive emotions occur during interactions such as feeding time, soothing, and close physical contact between the caregiver and child. This pattern of emotional sensitivity and responsiveness in parenting results in secure parent–child attachment, which is thought to be important for socioemotional competencies throughout childhood. Infants' early social environment also contributes to their neurobiological development and organization. For example, infants who are at genetic risk for poor emotion regulation may have worse outcomes when raised in a nonsupportive and nonresponsive family environment than infants who are either not at genetic risk or who are at genetic risk but are raised in more supportive family environments. Interestingly, because genetic risk may also be associated with greater sensitivity to the influence of the social environment overall, infants who are at genetic risk may actually have the most favorable outcomes in comparison to other children when they are placed in positive family environments with caregiver sensitivity and secure caregiver–child attachment (Easterbrooks et al., 2013).

In addition to the transactions between infants' and young children's social environments and neurobiological development, infants and young children are effecting changes in their social environments. For example, infants' emotional behavior provides feedback to caregivers about the efficacy of the caregivers' strategies in achieving their goals, which are, of course, developed within cultural contexts. For example, although caregivers tend to be universally interested in reducing infant distress, some cultures may emphasize the importance of reducing distress to a greater extent than others. That is, some cultures may emphasize immediate responses to distress based on that culture's belief that negative affect is not good to experience for extended periods, whereas other cultures may consider distress as a part of life and parents in that culture may

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23007-9

respond less quickly so that children can practice self-soothing or regulation. Within these contexts, caregivers learn from infants' emotionality regarding the success of their strategies in reducing current distress and/or avoiding similar instances in the future. In this way, the infant and the culture work together to scaffold caregiving responses.

Emotional Development Trajectories

Emotional Experience and Expression

Research on the development of emotional experience focuses on two types of broad emotional categories: primary or basic emotions (e.g., happiness, sadness, anger, fear, disgust, and surprise) which appear in both humans and some animals and self-conscious or self-in-relation emotions (e.g., jealousy, embarrassment, pride, shame, guilt, and empathy) and those which require advanced cognition about the self in relation to others. Although basic emotions are assumed to be universal, only broad emotional reactions are thought to be present at birth. Instead, maturational shifts in biological brain functioning and increases in attention and subsequent organizational and interpretive skill in assessing the environment facilitate the emergence of the basic emotions throughout the first year of life. Although researchers have thought previously that emotions rely entirely on maturational aspects, given that elements of joy, sadness, anger, and disgust are reported to appear around 3 months, recent reviews suggest that prototypical expressions of these emotions develop slowly over the first year of life (Camras and Shutter, 2010; Halberstadt et al., 2013). Because infants are also beginning during this time to understand the norms of their social world, their expression of even basic emotions may reflect the norms of their culture. This question remains largely untested with infants and children, with geographic 'dialects' in emotion expression only recently identified in adults (Elfenbein, 2013). The more cognitively complex emotions that consider the self in relation to others, such as jealousy, embarrassment, shame, pride, and guilt, emerge later in the second year of life, and may be even more subject to cultural influence.

Infants begin the lifelong task of developing the skills of regulating emotions in terms of inhibiting and/or minimizing the intensity and duration of their emotional experiences. Although caregivers are important emotion regulators throughout infancy and childhood, infants also engage in a variety of self-soothing behaviors from an early age, including thumb sucking, gaze aversion, and distraction. Furthermore, as young children develop in their self-awareness and ability to use language to describe feeling states, they are also able to verbalize contexts that make them feel distressed or give them joy (Kopp and Neufeld, 2002).

Emotion Understanding

Infants' basic abilities in distinguishing between positive and negative emotions occur within the first few months of life. We do not know yet whether distinguishing between these emotions is accompanied by an understanding of the feeling state of the other person, although social referencing studies, in which 12–14 month olds pause in response to caregivers' fear responses, indicate infants' use of emotional information to inform how they should react to novel situations and potential danger, in particular. Although infants may begin to use social referencing as early as 6 months of age, they become increasingly skilled at checking in with caregivers before acting to see if their faces reflect fear, happiness, or anger. Young children may also utilize social referencing when interested in feedback, such as when looking for validation from a caregiver after completing a difficult task or checking for anger when reaching for a forbidden object. Although some 2 year olds are able to identify and also generate facial expressions reflecting basic emotions upon demand, these skills become increasingly developed over time.

Related Domains

Related research domains that inform our understanding of emotional development in infancy and early childhood are those of temperament and attachment. Temperament literature describes infants' behavioral styles in terms of their general emotional responses to a variety of situations. Researchers have proposed several ways of classifying infants' temperament, and these include components of positive and negative affectivity, self- or emotion regulation, and activity level. Many infants' overall behavior is characterized by a general display of positive affect and ability and a propensity to adapt easily to novel situations, including emotional adaptability. However, other infants' overall demeanor may reflect general negative affect, and they may show little ability to attenuate negative emotion expressions, even after no longer being in the presence of the cause of distress. Temperament classifications tend to demonstrate stability throughout childhood, even when children's environmental niches shift. Temperament is likely genetically inherited, however, increasing evidence supports the substantial role of environmental experience in modifying outcomes (Henderson and Wachs, 2007).

Attachment describes the emotional bond between the infant and the caregiver. Theories of attachment often note the importance of the caregiver's responsiveness to the infant's emotional expression as a pathway by which a deep, trusting emotional bond is created. Infant–caregiver attachments develop from infants' instinctive preference toward other humans at birth to attachment formation with primary caregivers as infants distinguish between familiar others and strangers (Van Ijzendoorn and Sagi-Schwartz, 2008). The discussion of attachment is relevant because it reflects how emotional development occurs in the context of social relationships and the importance of emotional bonds for the developing child.

Although it may be tempting to accept the developmental trends of emotion competence previously described as universal, culture may impact the emergence of emotion and emotion-related components. Many of the laboratory and naturalistic observation paradigms used to measure and explore infants' emotional functioning were developed to examine emotional competence based on Western theory with Western samples. However, when similar yet culturally appropriate paradigms have been used to explore infants' emotional functioning, some universal assumptions have been challenged. For example, when a culturally appropriate version

of the commonly used strange situation paradigm was used with Nso 1 year olds, the 'universal' representation of secure attachment as characterized by stranger-evoked distress and caregiver-evoked comfort was not present. In contrast, a large number of Nso infants showed no emotional reaction to the stranger (Keller and Otto, 2009). These results suggest that we need to re-think both the operationalization and measurement of attachment in terms of cultural norms and expectations. We further discuss the consideration of culture in emotional development in infancy and how parental socialization inculcates cultural expectations for infants' emotional functioning.

Culture

Culture, as we define it, is the shared pattern of beliefs, values, behaviors, and language learned through inter- and intragenerational socialization. Culture determines the physical and social environments of infant development (Super and Harkness, 1986). To understand how culture impacts emotional development, researchers have targeted emotion socialization as the mediating pathway between cultural context and children's emotional competence and functioning. Parent–child relations are infused with culture through practices that dictate the way that parents and children interact with each other and the cultural belief systems that impact the interpretation of those interactions. Infants also learn about culture through experience-expectant (species-typical neural preparedness to integrate information from the environment) and experience-dependent (individual neural changes due to learning and experience) processes (Keller, 2007). Throughout infancy, children develop emotional competence by increasingly integrating information from parent–child interactions across cognitive, physiological, and motoric domains. We briefly review the cross-cultural literature on parental emotion socialization before discussing the nature of culture-specific features and cultural commonalities of emotion development in early childhood.

Emotion Socialization: A Cultural Pathway to Emotional Development

Parental emotion socialization is thought to be one of the most important influences on infants' and children's emotional development; caregivers' own emotional experience, communication, sensitive responsiveness, and understanding of their infants' needs help to shape the emotional world of early childhood. Parental socialization is also one of the primary ways in which culture is imparted in that caregivers' emotional socialization behaviors are themselves shaped/governed by the expectations, beliefs, and shared knowledge of the cultures in which they live. Cross-cultural studies have noted a variety of cultural differences in parent beliefs and expectations about children's emotions, early caregiver–child interactions, and parent emotion socialization, all of which have implications for early childhood emotion development.

Several cultural frameworks have been identified to organize cultural effects on parenting practices and beliefs. These include independence–interdependence, power distance, children's place in family and culture, ways children learn, and value of emotional experience and expression (Halberstadt and Lozada, 2011). The most common of these cultural frameworks is that of *independence–interdependence*, which is also often referred to as *individualism–collectivism*. Western cultures (i.e., United States, Germany, Greece) are most often associated with being independent or individualist. These societies consider children as separate entities from the caregiver and encourage the development of autonomy. Individualistic emotional competence encompasses self-awareness, and emotional expression and regulation of socially disengaged emotions (e.g., happiness, anger, and frustration) as assertions of individuality and autonomy. Eastern cultures (i.e., Japan, China, India) are most often described as interdependent or collectivist. Interdependent societies consider children as part of their social groups (e.g., family) rather than as individuals and encourage emotional bonding and unity between the caregiver and the child. These societies also encourage interconnectedness and group harmony. Interdependent emotional competence encourages an understanding of emotion in relation to what others are feeling, and socially engaged emotions (e.g., empathy, shame, guilt). A central goal in regulating emotional expression is to keep family and group harmony intact (Friedlmeier et al., 2011). Although researchers agree that one culture may tend toward a more individualist or collectivist orientation, it is also understood that these two orientations can exist simultaneously within cultures, with orientations existing along continuums rather than a dichotomy of polar opposites.

Parenting practices that promote infants' and children's emotional competence, as defined by the various frameworks supported in the culture, are reflected in differences in maternal–infant body contact, feeding patterns, responses to infant vocalizations (both distress and nondistress related), sleeping patterns, and even in other aspects of the child-rearing environment such as the presence of toys or the number of caregivers and family members that the infant comes in contact with on a daily basis. For example, the presence of many toys in the infants' environment is characteristic of independent cultures and can be used to elicit positive emotion in infants or to distract from and regulate negative emotion expression. However, infants in interdependent cultures may not have as much need or opportunity to play with toys when family members or others serve as the primary elicitors and regulators of emotion. These cultural differences may affect the frequency of emotional experience or expression, and skill in understanding the emotions of self and others, but they also do much to transform the action possibilities in the environment (called affordances in the perceptual theory developed by James and Eleanor Gibson), in which the entire emotional landscape of infants and children are embedded. To continue this example: action possibilities afforded by toys are very different from those afforded by people. Materialism may also serve as an independent influence on children's emotional development; as some interdependent cultures become more consumer-oriented, or as some families within interdependent cultures shift to a greater toy-focused parenting strategy, children's orientation of emotional satisfaction derived from interpersonal relationships may shift to that of object engagement.

Empirical studies have observed many differences in parental socialization across independence–interdependence orientations, as well as the other frameworks noted earlier. In a series of studies that compare maternal expectations of infants and mother–child interactions between urban middle class German mothers (an independent cultural context that also includes a framework which values children's emotional experience and expression) and rural farming Cameroonian Nso mothers (an interdependent context that also includes a framework which values children's low emotional expressivity), German mothers expected to see emotional expression early in infancy and reported having low expectations of children's emotional control when 3 years old. Behaviorally, they engaged in more face-to-face contact, positive vocalizations, smiling, and contingent responding in early infancy. In contrast, Nso mothers had higher expectations for emotional control, even in infancy. Behaviorally, they engaged in less face-to-face contact and smiling so as to avoid arousing infant emotion. All of these reflect Nso expectations that good children are calm with little emotional expressiveness, except in the case of socially engaged emotions such as shame and guilt (Keller and Otto, 2009). As such, the German and Nso cultures varied not only on the independent–interdependent orientation, but also in their beliefs about the value of emotional experience and expression, and about how and when children learn about and control their emotions.

Perhaps the most striking evidence for the early impact of cultural beliefs and norms on infants' emotional development is the shift in smiling during mutual gaze between 6 and 12 weeks of age by the Nso and German infants; this age was studied so as to capture any developments in emotional development at what is thought to be the 2-month shift in infants' systemic reorganization. Consistent with their beliefs and values regarding emotion expression, German mothers smiled at their infants more so than Nso mothers; although infant behavior is not different cross-culturally at 6 weeks, it is markedly different by 12 weeks, at which point German infants have substantially increased their smiling duration and imitation of mothers' smiles, more closely matching their mothers' behavior. Nso infants' smiling stayed fairly stable at its lower level, again reflecting the values of Nso society for less expressiveness (Wörmann et al., 2012). It is likely that cultural norms and nuances impact infants' and children's organization and re-organization of their understanding of the social world around them, and their responses to that world.

As noted above, although the framework of independence–interdependence is a powerful tool to use in understanding cross-cultural differences, other frameworks are needed to fully understand the richness of cross-cultural differences, as many cultural differences in emotion socialization have been noted in societies that both reflect interdependent cultural contexts. For example, the rural villages of the Tamang and Chhetri-Brahmin people in Nepal are similar on many characteristics of subsistence living and belief systems of social conformity and group harmony. However, the Tamang practice Tibetan Buddhism and beliefs of egalitarianism, tolerance, and the value of mild, socially engaged emotions. The Chhetri-Brahmin are high caste Hindus whose hierarchical place in society reflect spiritual purity and who value self-control. Their high-society status, even in rural villages, promotes feelings of pride and social dominance. The different belief systems due to religious affiliations represent the cultural framework of the value of emotional experience and expression and is manifest in differing patterns of socialization; Chhetri-Brahmin mothers responded more positively to their young children's anger yet ignore children's expressions of shame. However, Tamang mothers responded with rebuking their children's anger (Cole et al., 2006).

Although cross-cultural emotion socialization studies yield examples of many differences in parenting practices impacted by cultural beliefs and orientations, we also note evidence of universal trends in emotion socialization across cultures. For example, although maternal sensitivity may vary in its behavioral manifestations (e.g., visual or vocal contingency versus proximal contingency, frequency versus intensity of smiling in face-to-face interactions), its positive influence on infant emotional functioning is apparent across cultures. Similarly, differential responding to powerful (e.g., anger) and powerless (e.g., sadness) emotions may serve varying functions for the development of autonomy and relatedness, and also for one's place in terms of the power distance framework, yet socialization practices that dismiss or punish children's emotions are related to later internalizing and externalizing in later childhood in several cultures (Friedlmeier et al., 2011). The desire for culturally appropriate emotional competence in infancy is evident across cultures, but can cultural differences in emotion experience, expression, and understanding be observed in infancy and early childhood? Several findings from cross-cultural infant and preschool research are reviewed and discussed in the context of cultural orientations.

Cross-Cultural Findings of Emotional Competence in Infant and Early Childhood

Emotion Experience and Expression

Infant emotion and experience appear indistinguishable at birth, yet as young children develop in their ability to understand the self and others' expectations, the two abilities become differentiated. Cross-cultural differences in these abilities likely reflect cultural variations in beliefs about the value of emotional experience and expression and beliefs in the way children learn about emotions. For instance, comparisons of European American, Chinese, and Japanese infants suggest that Chinese infants are less expressive overall than both European American and Japanese infants; however, there were no significant differences in the frequency of genuine smiles and in the latency and frequency of crying (Camras et al., 1998). Differences between European American and Chinese infant emotion expression is expected because of cultural differences in the value of emotion, however, differences in Chinese and Japanese infant expression is surprising given that the cultures appear to be similar not only in collectivist orientation but also in having to place high value on the restraint of emotion expression.

Although some young children of different cultural contexts exhibit different emotional reactions to comparable experiences, others show differences in the maintenance or regulation of that emotion over time. For instance, Japanese and German preschool girls were exposed to self- and other-focused distress

conditions to examine whether they would display culture-specific emotional reactions to their own experiences of frustration or the frustration of a playmate. Children in both groups displayed similar spontaneous distress reactions to both conditions, but the trajectories of their reactions varied by self-focused distress and other-focused distress conditions. Specifically, German girls sustained their distress throughout the duration of the self-focused distress condition but decreased their expressions in the other-focused condition. Japanese girls showed distress maintenance in the other-focused distress condition but decreased affect in the self-focused distress condition. These findings demonstrate that although young children may display similar emotional reactions and emotion regulation patterns, culture-specific display rules may dictate the context in which these emotional experiences occur (Trommsdorff and Friedlmeier, 2010). Culture-specific display rules may also be evident in young children's emotion-regulation behaviors. For instance, although both Japanese and German preschool girls engaged in support-seeking behaviors when distressed, Japanese girls did so by seeking closer proximity to their mothers whereas German girls engaged in social referencing their mothers by eye contact (Friedlmeier and Trommsdorff, 1999). Differences in the focus of group harmony versus autonomy in Japanese and German culture contribute to differences in the social context of these children's reactions and the methods by which they seek to regulate emotion.

Although the above, and many other studies of emotion development, identify experiential and expressive differences across cultures, other studies highlight universal aspects of emotion expression development. Studies that compare the muscular configuration of emotion expression find that infants across culture show similar expressions for happiness, sadness, and anger. Although the differences in the support for and participation in the social smile of German and Nso infants change dramatically in the first 12 weeks of life, as discussed earlier, it must be noted that the intensity of smiling between mothers and infants did not differ across the two groups (Wörmann et al., 2012). When taken together with the previous work on Japanese and German preschoolers, this work suggests that culture may have more of an impact on the context, frequency, duration, and meaning of emotional experiences, but perhaps not on the basic expression of the emotion (e.g., intensity and muscular configuration).

Emotion Understanding

Children's emotion recognition abilities are considered valuable across cultures but perhaps for different reasons. For example, in interdependent cultures that promote the expression of socially engaged emotions, a child must be able to know what the other person is feeling in order to show sympathy. In independent cultures, a child may need to know what another person is feeling in order to respond in a way that would allow for the other person to have privacy as they worked to regulate their emotional experience. Although few cross-cultural studies have examined emotion recognition in early childhood, there is evidence to suggest early cross-cultural differences in emotion recognition. For instance, Chinese 4 year olds were significantly more accurate at emotion recognition than Australian 4 year olds (Markham and Wang, 1996). This is somewhat surprising given that the expectation is that children in independent cultures (e.g., Australia) would have more opportunities to learn about emotion expression because emotion expression is valued. However, it may be that children in an interdependent culture like China may gain more experience with being able to read affective information from microexpressions of emotion, making them more sensitive to the facial cues in the emotion recognition task.

Similar to the dearth of cross-cultural research on emotion recognition in early childhood, very few studies have examined emotion understanding in infancy and early childhood. However, some insight into early emotion knowledge can be gleaned from studies in which children report their feelings in response to hypothetical scenarios, and these studies suggest culture-specific patterns. In response to vignettes about emotion-eliciting situations, Nepali Tamang and Chhetri-Brahmin children as young as 6 years old reported emotions consistent with the display rules of their culture. Specifically, Tamang children were more likely to report that they would feel 'just okay' than Chhetri-Brahmin children who were more likely to report feeling anger to potentially anger-eliciting situations. This pattern reflects the Tamang adherence to Buddhist principles of avoiding strong emotionality. Although the Chhetri-Brahmin children were more likely to report that they would feel anger than Tamang children in response to the same vignettes, they were also more likely to report masking in those situations, consistent with their Hindu-oriented beliefs to engage in self-control (Cole and Tamang, 1998). The cultural differences among these two peoples support the importance of knowing a culture's framework with regard to the value of emotion experience and expression.

Future Directions

Cross-cultural research on emotion in infancy and childhood has yielded a great deal of knowledge on both the cultural commonalities and differences in the development of children's emotion experience, expression, and understanding. However, our knowledge of how these developmental trajectories vary is still limited to relatively few cultures which mainly vary on independent–interdependent conceptions of the self. Empirical studies that examine emotional competence across cultures that vary in their views of structural hierarchy, views on children's societal roles and competence, and beliefs about emotion may yield a more nuanced view in how cultural belief systems impact parental emotion socialization and changes in emotional competence in early childhood (Halberstadt and Lozada, 2011). Above all, advancements in culturally appropriate theories and methods are needed to make more consistent observations of the precursors of emotional development in a variety of cultures. Such advancements would allow emotion researchers to answer many of the basic questions about the nature of emotion that are still debated today.

See also: Cross-Cultural Research Methods in Psychology; Culture and Emotion; Emotion and Expression; Emotion,

Perception and Expression of; Emotional Development, Effects of Parenting and Family Structure On; Neonatal Temperament: From Phenomenology to Mechanism; Parenting Attitudes and Beliefs across Cultures; Self-Regulation During Early Childhood Across Cultures, Development of; Social and Emotional Development in the Context of the Family.

Bibliography

Camras, L.A., Shutter, J.M., 2010. Emotional facial expressions in infancy. Emotion Review 2, 120–129. http://dx.doi.org/10.1177/1754073909352529.

Camras, L.A., Oster, H., Campos, J., Campos, R., Ujiie, T., Miyake, K., Wang, L., Meng, Z., 1998. Production of emotional facial expressions in European American, Japanese, and Chinese infants. Developmental Psychology 34, 616–628. Retrieved from 10.1037/0012-1649.34.4.616.

Cole, P.M., Tamang, B.L., 1998. Nepali children's ideas about emotional displays in hypothetical challenges. Developmental Psychology 34, 640–646. http://dx.doi.org/10.1037/0012-1649.34.4.640.

Cole, P.M., Tamang, B.L., Shrestha, S., 2006. Cultural variations in the socialization of young children's anger and shame. Child Development 77, 1237–1251. http://dx.doi.org/10.1111/j.1467-8624.2006.00931.x.

Easterbrooks, M.A., Bartlett, J.D., Beeghly, M., Thompson, R.A., 2013. Social and emotional development in infancy. In: Lerner, R.M., Easterbrooks, M.A., Mistry, J., Weiner, I.B. (Eds.), Handbook of Psychology, Vol. 6: Developmental Psychology, second ed., vol. 6. John Wiley & Sons, Inc., Hoboken, New Jersey, pp. 91–120.

Elfenbein, H.A., 2013. Nonverbal dialects and accents in facial expressions of emotion. Emotion Review 5, 90–96. Retrieved from 10.1177/1754073912451332.

Friedlmeier, W., Trommsdorff, G., 1999. Emotion regulation in early childhood: a cross-cultural comparison between German and Japanese toddlers. Journal of Cross-Cultural Psychology 30, 684–711. http://dx.doi.org/10.1177/0022022199030006002.

Friedlmeier, W., Corapci, F., Cole, P.M., 2011. Emotion socialization in cross-cultural perspective. Social and Personality Psychology Compass 5, 410–427.

Halberstadt, A.G., Lozada, F.T., 2011. Emotion development in infancy through the lens of culture. Emotion Review 3, 158–168. http://dx.doi.org/10.1177/1754073910387946.

Halberstadt, A.G., Denham, S.A., Dunsmore, J.C., 2001. Affective social competence. Social Development 10, 79–119. http://dx.doi.org/10.1111/1467-9507.00150.

Halberstadt, A.G., Parker, A.E., Castro, V.L., 2013. Nonverbal communication: developmental perspectives. In: Hall, J.A., Knapp, M.L. (Eds.), Handbook of Communication Science, Nonverbal Communication, vol. 2. de Gruyter Mouton, Berlin, pp. 93–127.

Henderson, H.A., Wachs, T.D., 2007. Temperament theory and the study of cognition-emotion interactions across development. Developmental Review 27, 396–427. http://dx.doi.org/10.1016/j.dr.2007.06.004.

Keller, H., 2007. Cultures of Infancy. Lawrence Erlbaum Associates Publishers, Mahwah.

Keller, H., Otto, H., 2009. The cultural socialization of emotion regulation during infancy. Journal of Cross-Cultural Psychology 40, 996–1011. http://dx.doi.org/10.1177/0022022109348576.

Kopp, C.B., Neufeld, S.J., 2002. Emotional development in infancy. In: Davidson, R. (Ed.), Handbook of Affective Sciences. Oxford University Press, New York.

Markham, R., Wang, L., 1996. Recognition of emotion by Chinese and Australian children. Journal of Cross-Cultural Psychology 27, 616–643. http://dx.doi.org/10.1177/0022022196275008.

Super, C.M., Harkness, S., 1986. The developmental niche: a conceptualization at the interface of child and culture. International Journal of Behavioral Development 9, 545–569.

Trommsdorff, G., Friedlmeier, W., 2010. Preschool girls' distress and mothers' sensitivity in Japan and Germany. European Journal of Developmental Psychology 7, 350–370. http://dx.doi.org/10.1080/17405620802252742.

Van Ijzendoorn, M.H., Sagi-Schwartz, A., 2008. Cross cultural patterns of attachment: universal and contextual dimensions. In: Cassidy, J., Shaver, P.R. (Eds.), Handbook of Attachment: Theory, Research, and Clinical Applications, second ed. Guilford Press, New York, pp. 880–905.

Wörmann, V., Holodynski, M., Kärtner, J., Keller, H., 2012. A cross-cultural comparison of the development of the social smile: a longitudinal study of maternal and infant imitation in 6- and 12-week-old infants. Infant Behavior & Development 35, 335–347. http://dx.doi.org/10.1016/j.infbeh.2012.03.002.

Early Life Influences on Health and Mortality in Adulthood

Ken R Smith and Heidi A Hanson, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Conditions present early in a person's life are potentially important for explaining adult mortality risks. Many questions remain about which early life conditions contribute to these health and longevity differentials. In this article, we examine several key early life conditions that connect prenatal and early childhood factors to adult mortality risks with special attention given to the events proximate to the individual's birth and during childhood. Challenges and themes for future research that arise from this work are also presented.

Introduction

Humans have a distinctive and nearly universal pattern of mortality: elevated rates among newborns and infants that are followed by a decline through adolescence and young adulthood and then a subsequent and steady increase into old age, with the possibility of late-life mortality deceleration. The general shape of this mortality schedule is commonplace though the absolute levels vary across populations and historical periods. But what is of greatest interest for the present discussion is the connection between the left and right segments of this mortality profile. Factors affecting infant health and survival are manifold but are extremely sensitive to characteristics and correlates of fertility of the parents and the offspring's early years of dependency and childhood. These early characteristics represent important factors and pathways that serve to alter the mortality risks of newborns, infants, and young children. But more importantly for our purposes, these same traits may impart a lasting imprint on their *surviving* children when these offspring reach adulthood and well into their senior years. This article, therefore, examines the connection between prenatal and early childhood characteristics and later onset mortality risks with exclusive attention given to the events proximate to the individual's birth and during childhood.

Our conceptual lens is decidedly biodemographic, a perspective that embraces social, biological, and genetic perspectives. The literature on the biodemography as well as the epidemiology of the late effects of early life conditions is vast with many important mechanisms and debates represented – more than can be comprehensively synthesized in one chapter – but we take the position that these earliest of circumstances, those comprising gestation through infancy and childhood, are arguably among the most fundamental features of an individual's life history. Moreover, these early moments are also crucial for setting the stage for other putative early circumstances that are the focus of much research in the field (Elo and Preston, 1992; Kuh and Yoav, 1997; Power et al., 2013). We note that there is not complete unanimity regarding whether early life conditions are important for understanding later-life health. Some studies suggest that early life factors affect later-life mortality only 'modestly' (Barbi and Vaupel, 2005) while others conclude, as we do in this article, that exposures in the early years are crucial in affecting mortality shifts among elderly individuals (Finch and Crimmins, 2004).

This article is structured by beginning with a review of the importance of ontogeny and the intrauterine environment and followed by central socially and biologically induced changes that comprise much of the biodemographic literature on early life circumstances. We begin with briefly describing the basic arguments surrounding the fetal origins hypothesis, a hypothesis that arises from work advanced by Barker (1990) and that proposes important mechanisms that have influenced much of the early life and adult mortality literature. We then turn our attention to studies that examine specific mortality effects associated with the fertility circumstances into which the child is born and early childhood conditions including the roles played by birth weight (BW), gestational age, birth order (and relatedly sibship size), parental and grandparental age, and early parental death. We also give special emphasis to a fundamental theme: the role of mortality selection, as any empirical estimates seeking to associate circumstances in early life to the chances of dying decades later presumes that the individuals in question are alive long enough to face these potentially lethal risks.

The literature on early conditions has generally been silent on the role of specific genetic or familial risks except to acknowledge the potential that these influences exist (Blackwell et al., 2001) though recent reviews point to its importance (Power et al., 2013; Gluckman et al., 2010). Increasing cross-disciplinary attention to epigenetics, the perspective that argues that gene expression (rather than the genes themselves) is affected by environmental circumstances, has also served to heighten the importance of gene–environment interactions in affecting adult health. This is evidenced by recent advocates proposing new methodologies within life course epidemiology (Strully and Mishra, 2009). A number of authors have argued that studies of early life circumstances should be incorporating not only genotype information (possibly the earliest of 'conditions') as life's starting point but also an individual's family medical history (Smith et al., 2009). Given our focus here, we do not devote any further space to this topic though we know that scholarship will only grow that uses as its conceptual focus the interplay between early environmental conditions, genomic risk, family medical history, and the epigenetics of later-life health.

Fetal Origins

As the earliest circumstance, the uterine environment for the developing fetus represents the context in which potentially profound forces exist that may have last effects in later life. This perspective, advanced by David Barker (Barker, 1990) building on early work by Forsdahl (Forsdahl, 1977) and frequently described as the fetal origins hypothesis, theorizes that there are specific or critical periods during development *in utero* that are especially susceptible to adverse developmental effects due to poor nutrition that may then alter mortality risks many decades later in adulthood. A central element of this theory argues that expectant mothers, when living conditions are compromised due to limited access to food, create a uterine environment that affects the development of her fetus so that the child will be prepared or 'programmed' for the very same environment where resources are scarce.

The resultant "thrifty phenotype hypothesis" (Hales and Barker, 2001) states that newborns experiencing poor fetal growth (and therefore are generally low BW) will be more likely to experience Type 2 diabetes and the metabolic syndrome in adulthood as a consequence of poor nutrition *in utero*. Children born under these conditions who acquired insulin resistance as an adaptive response to limited nutritional resources that helped them *in utero* may actually be worse off when prenatal undernutrition is a poor predictor of nutritional availability later in life. These responses may be adaptive in the short run but deleterious in the long run and may lead to accelerated adult mortality (Gluckman et al., 2010; Barker, 1995).

Arguably one of the most iconic human studies that investigates the fetal origins hypothesis is related to the late effects related to starvation and malnutrition initiated by the 1944–45 Dutch Hunger Winter during World War II brought about by a blockade by the German army as well as a national railway strike (Lumey et al., 2011). Some adverse outcomes have been associated with *in utero* exposure related to this historical event period, including cardiovascular disease (CVD) and elevated body mass index (Painter et al., 2005). The impact of undernutrition *in utero* has not been consistently associated with adult mortality risks thus far despite the establishment of an association of cardiovascular and diabetes risks. It may simply be too soon to know if mortality effects are evident among the adults from this population (Lumey et al., 2011). In other population-based nutrition studies, excess mortality for females exposed to undernutrition in early gestation was reported.

The inconsistent pattern of findings may be related to the nature of the famine in terms of duration and age of exposure which may have led to variation in gender-specific findings. The physiological changes associated with exposure to undernutrition *in utero* may be sensitive to the timing and duration of the insult. Different fetal organs grow at different rates, making the timing of the insult an important aspect to take into account. Lack of proper nutrition during the period around conception may adversely affect placental development (Barker et al., 2010). A recent study of adult mortality in men and women by trimester of exposure to the Dutch famine found that females exposed to famine in the first trimester had increased risk of adult mortality, cardiovascular mortality, and cancer mortality compared to the unexposed. Exposure to undernutrition in the second trimester may lead to decreased renal function in adulthood and an increased risk for obstructive airway disease (Lopuhaa et al., 2000; Roseboom et al., 2006). Late-gestational exposure has been linked to impaired glucose tolerance (Ravelli et al., 1998).

Gender differences in the effects of undernutrition may be the result of differential and innate frailties between males and females as well as gendered mortality selection. As males have higher fetal mortality rates, these male embryos may be more apt to experience the adverse intrauterine effects than the more robust female fetuses. Accordingly, men may have greater excess mortality in adulthood. Studies of the Dutch Potato Famine 1846–47 show significant excess mortality due to undernutrition *in utero* but only for males. A study of Utah pioneers that examined prenatal exposures to food shortage showed an association between undernutrition *in utero* and adult mortality with effects varying by gestational period of exposure and gender (Hanson and Smith, 2013).

The lack of clear and consistent later-life mortality effects of *in utero* undernutrition may reflect mortality selection where only the more robust survive to adulthood (Hawkes et al., 2012; Elo and Preston, 1992). Well-nourished cohorts would not experience mortality selection pressure when its members were young leading to unfavorable adult mortality risks in relation to cohorts exposed to undernutrition *in utero* and infancy. The better-resourced cohort may therefore face adult mortality rates that are higher than that of the undernourished cohort, a phenomenon that may explain the scarce and uneven evidence demonstrating excess adult mortality risks.

Birth Weight and Preterm Births

The fetal origins hypothesis predicts that fetuses adapting to an adverse *in utero* environment may change the development of key organ systems, changes that predispose the child to low BW and ultimately to an increased risk of CVD, diabetes, and stroke as an adult. Accordingly, considerable work has been done specifically on the long-term health consequences of low (and high) BW children. Birth weight is a major public health issue and the percentage of U.S. children born low BW in recent decades rose from 6.8% in 1980 to 8.2% in 2009 (Archer et al., 2011). Birth weight is a serious risk factor for adult risks not only for CVD but also for some cancers. Health risks due to preterm birth are equally important with 5–10% of young people today being born preterm. Processes that underlie the association between child BW and gestational age are complex because they may involve competing influences (e.g., low BW increases CVD risk while high BW increases some cancer risks).

Birth weight is a crucial proxy for maternal nutrition that crosses the placenta that affects fetal growth or as a marker of intrauterine conditions that led to low BW (Gluckman et al., 2008). Some conclude that altered vascular physiology associated with low BW is largely irreversible

(Norman, 2008). Preterm children also face a number of adverse health outcomes as older children (Saigal and Doyle, 2008) and younger adults with respect to mortality (Swamy et al., 2008).

A recent meta-analysis of 22 studies assessed the association between BW and adult mortality from all causes, CVD or cancer (Risnes et al., 2011). For all-cause adult mortality, an increase of 1 kg of BW translated into a 6% decrease in risk, a protection that was larger (12% reduction per kg of BW) when focusing on CVD. For mortality specific to cancer, the risk associated with a kilogram increase in BW was an increase of 13%. This study demonstrated a protective association between BW and all-cause adult and cardiovascular mortality. Increase in BW was a risk factor for cancer mortality, particularly men. Some studies reporting a positive association between BW and adult cancer risk may be attributable to estrogen exposure and insulin growth factors (Lof et al., 2007).

One of the largest studies of adult survival based on a comparison between full-term and preterm births was conducted by Swamy and colleagues (Swamy et al., 2008), an important analysis given that much of the mortality analysis of preterm births focuses on infant mortality. Given data limitations, there are very few well-designed analyses of survival in the following decades extending into the adult years. Swamy and her collaborators used data from nearly 1.2 million births (singletons only) in the Medical Birth Registry of Norway who were born between 1967 and 1988. Survival was assessed for these children through 2002 such that the oldest individual could be as old as 35 years at the conclusion of the study, an unprecedented period of follow-up for a cohort containing preterm births. The cohort was truncated to births from 1967 to 1976 for assessment of educational achievement and reproductive outcomes through 2004. They stratified the preterm sample into several gestational age categories (22–27, 28–32, and 33–37 weeks). For both boys and girls in the shortest gestational age group, preterm children were shown to have an excess risk of mortality throughout childhood. For boys, the significant relative risk (RR) of mortality for preterm versus full-term births was RR = 5.3 for 1–5.9 years of age and 7.0 for 6–12.9 years of age. The comparable estimates for girls was RR = 9.7 for the younger age group (no later childhood deaths were observed in this cohort). For 28–32 weeks of gestation, the mortality rates for boys were still significant relative to full-term births but had dropped to approximately RR = 2.4 for the same childhood age groups, a result different for girls which showed no significant increases in childhood mortality. Mortality in adolescence was not affected by gestational age for either gender. These results again point to the importance of mortality selection because as the individuals age, the most susceptible are culled from the population thereby leaving a more homogeneous and healthy subpopulation. One of the challenges of studying preterm births relates to its confound with BW. Delineating their unique contributions to adult mortality risks must be done well since they may involve different mechanisms (e.g., low BW and compromised or adaptive development, preterm birth, and maternal infections) (Liew et al., 2008).

Sibship Size and Birth Order

One of the earliest family structure characteristics is established by a child's position within a sibship. Life chances and adult mortality risks have been hypothesized to be sensitive to these two distinct but correlated characteristics. The number of siblings has been found to be positively associated with lower educational achievement and unhealthy lifestyle choices which may account for some of the modest but significant effects of larger sibship sizes in adult mortality risk (Smith et al., 2009; Modin, 2002), though siblings may also be a source of social support. Children with more siblings have been shown in natural fertility eras and in developing countries generally to face excess infant and childhood mortality rates which may, once again, remove the most vulnerable from the population and severe to limit the effects of sibship size on later mortality. Several studies have advocated the "resource dilution model" (Modin, 2002), which hypothesizes that parents are faced with resources constraints and that more offspring simply means a dilution of resources that may serve to adversely affect these offspring later in life.

Few studies have examined birth order and its influence on longevity. In the Modin study, higher birth-order girls were found to have excess mortality risks in relation to firstborn sisters (Modin, 2002). Firstborn children may benefit from more parental time and resources but they may also assume more adultlike duties as well. This same study indicated that a mechanism underlying this phenomenon is that firstborn children are more likely to attend college and reach higher levels of educational and occupational achievement than their siblings (Modin, 2002). Characterizing the role of birth order on adult health is an important issue to examine further since family sizes are shrinking globally making sibship size and birth order increasingly synonymous (i.e., firstborn in a one-child family).

Correlated Environments

Where a person begins will often affect where they land in terms of a range of characteristics including health, socioeconomic status (SES), and residential location. From this social 'immobility' perspective, individuals whose lives begin in disadvantaged circumstances often endure disadvantages throughout their lives, and similarly the initially advantaged will frequently remain so as they age. These ideas have been characterized in two related concepts: correlated environments and cumulative advantage (Preston et al., 1998). In the former instance, individuals starting out living in an advantaged setting (e.g., parents, neighborhoods, schools, peers) will likely enjoy 'correlated' advantages or settings throughout life. This phenomenon of resource inertia or persistence is perhaps most apparent in the case of residential stability: the person who ages in place will be subjected to the same or very similar forces in childhood as in adulthood since these forces persist in the place. In a similar fashion, children born to wealthy parents will enjoy the benefits of that wealth when they are children as well as when they age into adulthood (George, 2005). This phenomenon has been examined in a recent analysis of SES

over the life course and its effects on adult mortality in Finland (Elo et al., 2013).

Cumulative advantage suggests that early differences favoring one group over another will provide opportunities to the favored group that will fortify these early advantages (Dannefer, 2003). Some of these early advantages create new environments later in life that are strongly associated with the environments present in the beginning. Accordingly, the perspectives of cumulative advantage and correlated environments share common attributes and mechanisms. In both cases, those with early deprivation are likely to endure many of the same adversities throughout life. This describes a mechanism that is counter to the idea that early susceptibilities per se lead to subsequent poor health. The idea of cumulative advantage has been codified in various compendia (Kuh and Yoav, 1997), and applied to explain how cumulative exposures affect later-life health and socioeconomic outcomes (Dannefer, 2003), myocardial infarctions (O'Rand and Hamil-Luker, 2005), self-rated health (Mirowsky and Ross, 2008), and premature mortality (Melchior et al., 2006). Geronimus and colleagues have presented related ideas with respect to the weathering hypothesis which argues that some groups (e.g., African-Americans) experience adverse health effects and increases in allostatic load due to the cumulative impact of repeated encounters with social, economic, or political exclusion and deprivation (Geronimus et al., 2006; Geronimus, 1992).

Parental and Grandparental Age at Birth

There is growing evidence that an individual's reproductive history plays a role in affecting their risk of later-life mortality. Specifically, late-fertile mothers, and to a lesser extent, late-fertile fathers, have been shown to have better survival past the age of 50 years in relation to other individuals. These findings are based on populations when technological advances that could artificially extend ages of female reproduction as well as longevity are largely absent. The question is whether later fertility confers a survival advantage for the adult offspring that are the outcome of this late fertility. In considering this topic, it is equally important to examine the role played by young parental ages given reports of adverse consequences encountered by children born to these parents (Henretta, 2007).

There are several proposed mechanisms that link later parental age and adult offspring survival. First, some investigators have argued that adult offspring of late-fertile fathers will have reduced longevity in relation to their counterparts born to younger fathers. Based on the findings of Gavrilov and Gavrilova (Gavrilov and Gavrilova, 1997), the proposed association suggested by this alternative hypothesis should be especially pronounced among female offspring. They suggest that daughters of older fathers experience a significantly lower adult survival. Other studies report detrimental effects of late fatherhood on offspring longevity, most of which are associated with the putative increase in mutation rates with paternal age for specific conditions (Smith et al., 2009). Among these conditions are dwarfism, congenital heart defects, acute lymphoblastic leukemia, and schizophrenia as well as autism.

For mothers, increasing age translates into declining fecundity and an enhanced chance of obstetric complications. This reduction in fecundity generally begins around the age of 30 years and is related to the growing damage to DNA in germ cells and accordingly lower quality oocytes. This age-related pattern of fecundity plays a key role in understanding the association between maternal age and offspring health as noted by the increased risk of bearing a child with Down syndrome as maternal age increases.

Some have argued that late paternal age may result in beneficial effects for offspring. Offspring of late-fertile fathers have been found to be better off because these fathers are socially established, more financially stable, and better able to provide for their families (Mare and Tzeng, 1989). This suggests that these 'mature' fathers would be of great benefit to their children and would lead to better offspring survival. Analytically, this position indicates that it is important to control for the effects of parental SES so as to generate separate estimates of parental age from a putative effects of high SES.

Not all investigations find an association between late parental age and offspring survival health. Two recent studies using representative of the United States and Canada found little evidence that older mothers have adult children with poorer health or elevated mortality (Hubbard et al., 2009; Myrskyla and Fenelon, 2012).

Offspring born to very young parents are known to suffer adverse consequences associated with higher levels of material deprivation associated with early parenthood. This can occur for many reasons including reduced levels of education of their parents and the strong likelihood that there will be many subsequent births which further raises the prospects of poverty for children of teen parents. While the literature unfortunately focuses on contemporary populations, it is nonetheless informative in that it strongly suggests that children of teen mothers would have encountered adversity in times past. However, for populations living in natural fertility conditions where agricultural conditions prevailed, many children can be an important asset such that early and frequent childbearing could enhance the family's economic circumstances.

Finally, late reproduction has consequences for the offspring in terms of the years that parents are alive and available to these offspring. Specifically, a child born to older parents will likely see these parents die at an earlier age than a child born to younger parents. Evidence exists showing that adult offspring whose parents died when the offspring was young (<18) were themselves at increased risk of adult mortality (Smith et al., 2009). Moreover, the child's age when that child loses a parent to death may be a reflection of some aspect of shared family longevity.

An additional inferential challenge exists for studies attempting to establish a relationship between late parental age and offspring adult mortality: SES. It is well established that increasing levels of SES are associated with improving health, adult health, and survival. Children born to older parents will likely be joining families with greater resources.

Research on parental age and offspring mortality is associated with the idea that very early or very late ages at reproduction among grandparents may also have an effect on the longevity of their grandchildren. This work suggest that if the child's own parents were conceived by very young or very old

parents (i.e., the child's grandparents), does that child face differential survival chances effects in relation to offspring born to mothers and fathers who themselves were born to parents when they were at normative reproductive ages, all independent of maternal and paternal ages? Several possible mechanisms may link advanced grandparental and parental age on grandchild survival. For example, Eisenberg (Eisenberg et al., 2012) showed that grandchildren of older paternal grandfathers at the birth of fathers have longer telomeres. This suggests that these older grandfathers may actually enhance the survival of their grandchildren as longer telomeres are associated with better survival. This argues for a protective effect of being a grandchild of an older paternal grandfather. Recent studies have shown that advanced grandmaternal age may also affect the survival prospects for their grandchildren (independent of maternal age), particularly for reasons associated with birth defects such as Down syndrome (Aagesen et al., 1984; Malini and Ramachandra, 2006). The argument for this association is based on the idea that grandmothers conceiving daughters at advanced ages are carrying female fetuses who, during gestation, will be developing ova that will be later fertilized leading to the subsequent grandchildren. If the chromosomal anomalies associated with older grandmaternal eggs that were present for the ego's mother's birth, then it is possible that these anomalies may also affect the quality of the store of eggs being developed by ego's mother when she was a fetus.

Early Death of a Parent

This article has focused on events proximate to the birth of the individual in question. In this section, we turn our attention to a somewhat later familial event, death of a parent. This is introduced partly because of its importance to the life course literature and its undeniable importance to a child's life but also because it is closely related to the domains discussed thus far. For example, undernourished mothers and parents bearing children later in life are also at greater risk of dying when the child is still young.

A number of investigations have examined how early parental death has increased the risk of adverse health outcomes later in life (Chang et al., 2011). Younger children in these bereaved families are, therefore, likely to experience deprivations (e.g., the loss of social and economic support) such as those encountered by the surviving parent. Certainly childhood and adolescence are phases where psychological and physical change can be tumultuous ordinarily – a loss of a parent at these ages could therefore yield dramatic adverse and lasting effects. Studies of historical populations show how important parental death was for the survival of children (van Poppel, 2000). In a study of mortality in a historic cohort where all subjects were born in the latter half of the nineteenth century, there was no empirical support for excess mortality associated with early parental death, *ceterus paribus* (Smith et al., 2009). In studies that examine this question, parental death may be an indicator of environmental conditions that led to both a parent's death and that also adversely affected the adult offspring's risk of premature death.

A number of investigations have examined specific late-age conditions instead of all all-cause mortality. In studies of Alzheimer's disease (AD) that examine the adverse effects of early parental death that use the Utah Population Database, an increased risk of AD risk past the age of 65 years was found if an individual lost their parents to death early in life (Norton et al., 2011). In addition, a few well-done population-based investigations have also considered how early life parental death may alter the suicide risk of their adult offspring. Early death of a mother has been associated with an excess risk for suicide (Agerbo et al., 2002; Gravseth et al., 2010). A Swedish registry-based study examining data from 1969 to 2004 (Wilcox et al., 2010) found the risk of suicide for the offspring was elevated when parents committed suicide but not for other causes. Niederkrotenthaler and coauthors also used Swedish data and reported evidence that parent death early in the child's life increases their risk of suicide (Niederkrotenthaler et al., 2012). These studies illustrate the centrality of parental survival with respect to altering the probabilities for a long and healthy life for their offspring and highlight the complex set of factors that complicate these associations, especially the causes and contexts of the parent's death.

With early parental death, remarriage of the surviving parent is a real prospect that may affect adult offspring mortality risk. In the AD study noted earlier, they reported that early parental death elevated the adult offspring risk of AD after the age of 65 years. A historical Swedish study also showed a similar protective effect for all-cause mortality with respect to remarriage of the surviving spouse. In short, we expect remarriage of the widowed parent may provide additional resources that link parental death to suicide risk (Andersson et al., 1996).

Mortality Selection

Demographers and evolutionary biologist have addressed early life conditions and their association with aging and adult mortality at the population level. As an alternative to the individual level analyses discussed thus far, the population perspective examines parameters designed to depict broad characteristics of mortality that measure fundamental patterns in a defined society. A long-standing classic and useful tool to examine these issues is the Gompertz model. The Gompertz equation has two components that address the two central elements of this article: $m(x) = m(0)\exp(bx)$. Here, the mortality rate at any age x (i.e., $m(x)$) is a function of two parts. The first, $m(0)$, summarizes environmental or extrinsic exposures in the population that especially dominate during childhood; nonetheless, it describes living conditions and external threats that apply to all members of a population and that are not generally viewed to be related to aging per se. The second component, $\exp(bx)$, has a parameter b that measures how $m(x)$ changes with increasing age x. The parameter b is positive indicating that the susceptibility to death grows with age due to the accumulation of stress, mutation load, disease, and senescence. This is often described as 'intrinsic mortality' arising from physiological deterioration starting after maturity (Hawkes et al., 2012).

Examining the relationship between broader environmental risks, which dominate the mortality risk of the young as measured by $m(0)$, and the rate of mortality increase by age as measured by $\exp(bx)$, helps to better highlight the selection

mechanisms that link early life circumstance to subsequent life chances years later. In general, these two mortality parameters are inversely related (Hawkes et al., 2012). In their classic treatment of this question, Strehler and Mildvan in 1960 studied the association of the two Gompertz parameters and found support for an inverse relationship between them using life tables from 1955 (Strehler, 2000). These results have been replicated across a diverse set of populations and have been termed the compensation law of mortality which states that "High mortality rates in disadvantaged populations … are compensated for by low apparent 'aging rates'…" (Gavrilov and Gavrilova, 2001).

This body of work raises one of the key conundrums of the life course epidemiological and biodemographic approach linking early years to later years. Our ability to see a later-life health and survival consequence presumes that the individuals in question are alive and still in the population. Demographers often use data that have been subjected to various forms of nonrandom selection (e.g., Health and Retirement Survey, AHEAD Study). A study of later-life mortality requires that the individual lives to an age where health and subsequent survival can be observed. Having complete records of entire populations (e.g., genealogies or longstanding population registries) where a person's life and its events are depicted from birth to death gives us the opportunity to observe and understand who has been selected out of the subpopulation of interest and to make assessments of any selection biases created by the resulting mortality selection.

The fundamental point, then, is that all populations are characterized by heterogeneity in survival prospects since individual members have varying levels of frailty and as the force of mortality selection changes, different sets of survivors (but survivors nonetheless) will comprise the older age subpopulation. If extrinsic mortality risks increase, the more susceptible individuals will have left the population at a young age leaving a class of older individuals who were likely always at lower risk. Under this general differential heterogeneity hypothesis, scholars of late effects of early life exposures are making their assessments based on a nonrandom subset of the population and often without attempting to examine how the selection may affect the findings. If, in the extreme, the early life exposure in question is so lethal, then very few individuals will be observed at older ages with the exposure and equally as important, fewer yet will be experiencing the health outcome of interest. Of course, this selection process is at play even further back in time in terms of conceptions and spontaneous abortions – women who suffer from infections and malnutrition are less likely to conceive or carry their children to term. These too are selection processes that obscure our view when we seek to see associations between early and later-life events.

Conclusion

The linkages and pathways connecting prenatal and childhood conditions to adult mortality risks are manifold and complex. In this article, a number of specific traits, risk factors, and processes have been outlined that address these broad scientific questions. As this literature is large and diverse, we were only able to address a sampling of the empirical findings while simultaneously offering general themes and controversies. While this research area has been the focus of a number of studies, consensus regarding how and which early life conditions contribute to these health and longevity differentials remains elusive. This state of affairs is explicable given the myriad associations linking factors occurring earlier in life to numerous adult-onset diseases and premature mortality. We conclude by encouraging other investigators about the importance of studying multiple exposures since a more complete portrayal and understanding of the complex linkages connecting past circumstances to later health outcomes requires a large and conceptually driven set of measures of adversity and privilege and health endpoints.

Acknowledgments

This article was supported by NIA grant 'Early Life Conditions, Survival, and Health' (R01AG022095, Smith PI).

See also: Life Course Epidemiology; Life Course: Sociological Aspects; Linked Lives; Mortality Differentials: Selection and Causation; Mortality, Biodemography of; Prenatal and Infant Development: Overview, Current Trends, Future Directions; Sibling-Order Effects.

Bibliography

Aagesen, L., Grinsted, J., Mikkelsen, M., 1984. Advanced grandmaternal age on the mother's side–a risk of giving rise to trisomy 21. Annals of Human Genetics 48, 297–301.

Agerbo, E., Nordentoft, M., Mortensen, P.B., 2002. Familial, psychiatric, and socioeconomic risk factors for suicide in young people: nested case-control study. British Medical Journal 325, 74.

Andersson, T., Hogberg, U., Åkerman, S., 1996. Survival of orphans in 19th-century Sweden – the importance of remarriages. Acta Paediatrica 85, 981–985.

Archer, J.M., Yeager, S.B., Kenny, M.J., Soll, R.F., Horbar, J.D., 2011. Distribution of and mortality from serious congenital heart disease in very low birth weight infants. Pediatrics 127, 293–299.

Barbi, E., Vaupel, J.W., 2005. Comment on "Inflammatory exposure and historical changes in human life-spans". Science 308, 1743.

Barker, D.J., 1990. The fetal and infant origins of adult disease. British Medical Journal 301, 1111.

Barker, D.J., 1995. Fetal origins of coronary heart disease. British Medical Journal 311, 171–174.

Barker, D.J.P., Thornburg, K.L., Osmond, C., Kajantie, E., Eriksson, J.G., 2010. Beyond birthweight: the maternal and placental origins of chronic disease. Journal of Developmental Origins of Health and Disease 1, 360–364.

Blackwell, D., Hayward, M., Crimmins, E., 2001. Does childhood health affect chronic morbidity in later life? Social Science & Medicine 52, 1269–1284.

Chang, H.Y., Clark, J.M., Weiner, J.P., 2011. Morbidity trajectories as predictors of utilization: multi-year disease patterns in Taiwan's national health insurance program. Medical Care 49, 918–923.

Dannefer, D., 2003. Cumulative advantage/disadvantage and the life course: cross-fertilizing age and social science theory. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 58, S327–S337.

Eisenberg, D.T., Hayes, M.G., Kuzawa, C.W., 2012. Delayed paternal age of reproduction in humans is associated with longer telomeres across two generations of descendants. Proceedings of the National Academy of Sciences of the United States of America 109, 10251–10256.

Elo, I.T., Martikainen, P., Myrskyla, M., 2013. Socioeconomic status across the life course and all-cause and cause-specific mortality in Finland. Social Science & Medicine. http://dx.doi.org/10.1016/j.socscimed.2013.11.037.

Elo, I.T., Preston, S.H., 1992. Effects of early-life conditions on adult mortality: a review. Population Index 58, 186–212.

Finch, C.E., Crimmins, E.M., 2004. Inflammatory exposure and historical changes in human life-spans. Science 305, 1736–1739.
Forsdahl, A., 1977. Are poor living conditions in childhood and adolescence an important risk factor for arteriosclerotic heart disease? British Journal of Preventive & Social Medicine 31, 91–95.
Gavrilov, L.A., Gavrilova, N.S., 1997. When should fatherhood stop? Science 277, 17–18.
Gavrilov, L.A., Gavrilova, N.S., 2001. The reliability theory of aging and longevity. Journal of Theoretical Biology 213, 527–545.
George, L.K., 2005. Socioeconomic status and health across the life course: progress and prospects. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 60 (Spec No 2), 135–139.
Geronimus, A.T., 1992. The weathering hypothesis and the health of African-American women and infants: evidence and speculations. Ethnicity & Disease 2, 207–221.
Geronimus, A.T., Hicken, M., Keene, D., Bound, J., 2006. "Weathering" and age patterns of allostatic load scores among blacks and whites in the United States. American Journal of Public Health 96, 826–833.
Gluckman, P., Hanson, M., Buklijas, T., 2010. A conceptual framework for the developmental origins of health and disease. Journal of Developmental Origins of Health and Disease 1, 6–18.
Gluckman, P., Hanson, M., Cooper, C., Thornburg, K., 2008. Effect of in utero and early-life conditions on adult health and disease. New England Journal of Medicine 359, 61.
Gravseth, H.M., Mehlum, L., Bjerkedal, T., Kristensen, P., 2010. Suicide in young Norwegians in a life course perspective: population-based cohort study. Journal of Epidemiology & Community Health 64, 407–412.
Hales, C.N., Barker, D.J., 2001. The thrifty phenotype hypothesis. British Medical Bulletin 60, 5–20.
Hanson, H.A., Smith, K.R., 2013. Early origins of longevity: prenatal exposures to food shortage among early Utah pioneers. Journal of Developmental Origins of Health and Disease 4, 170–181.
Hawkes, K., Smith, K.R., Blevins, J.K., 2012. Human actuarial aging increases faster when background death rates are lower: a consequence of differential heterogeneity? Evolution 66, 103–114.
Henretta, J.C., 2007. Early childbearing, marital status, and women's health and mortality after age 50. Journal of Health and Social Behavior 48, 254–266.
Hubbard, R.E., Andrew, M.K., Rockwood, K., 2009. Effect of parental age at birth on the accumulation of deficits, frailty and survival in older adults. Age Ageing 38, 380–385.
Kuh, D., Yoav, B.-S., 1997. A Life Course Approach to Chronic Disease Epidemiology. Oxford, New York.
Liew, G., Wang, J.J., Mitchell, P., 2008. Preterm birth, long-term survival, and fertility. Journal of American Medical Association 300, 167–168, 167.
Lof, M., Sandin, S., Hilakivi-Clarke, L., Weiderpass, E., 2007. Birth weight in relation to endometrial and breast cancer risks in Swedish women. British Journal of Cancer 96, 134–136.
Lopuhaa, C.E., Roseboom, T.J., Osmond, C., Barker, D.J., Ravelli, A.C., Bleker, O.P., Van Der Zee, J.S., Van Der Meulen, J.H., 2000. Atopy, lung function, and obstructive airways disease after prenatal exposure to famine. Thorax 55, 555–561.
Lumey, L.H., Stein, A.D., Susser, E., 2011. Prenatal famine and adult health. Annual Review of Public Health 32, 237–262.
Malini, S.S., Ramachandra, N.B., 2006. Influence of advanced age of maternal grandmothers on down syndrome. BMC Medical Genetics 7, 4.
Mare, R.D., Tzeng, M., 1989. Fathers' ages and the social stratification of sons. American Journal of Sociology 95, 108–131.
Melchior, M., Berkman, L.F., Kawachi, I., Krieger, N., Zins, M., Bonenfant, S., Goldberg, M., 2006. Lifelong socioeconomic trajectory and premature mortality (35–65 years) in France: findings from the GAZEL Cohort study. Journal of Epidemiology & Community Health 60, 937–944.
Mirowsky, J., Ross, C.E., 2008. Education and self-rated health cumulative advantage and its rising importance. Research on Aging 30, 93–122.
Modin, B., 2002. Birth order and mortality: a life-long follow-up of 14,200 boys and girls born in early 20th century Sweden. Social Science & Medicine 54, 1051–1064.
Myrskyla, M., Fenelon, A., 2012. Maternal age and offspring adult health: evidence from the health and retirement study. Demography 49, 1231–1257.
Niederkrotenthaler, T., Floderus, B., Alexanderson, K., Rasmussen, F., Mittendorfer-Rutz, E., 2012. Exposure to parental mortality and markers of morbidity, and the risks of attempted and completed suicide in offspring: an analysis of sensitive life periods. Journal of Epidemiology & Community Health 66, 233–239.
Norman, M., 2008. Low birth weight and the developing vascular tree: a systematic review. Acta Paediatrica 97, 1165–1172.
Norton, M.C., Smith, K.R., Østbye, T., Tschanz, J.A.T., Schwartz, S., Corcoran, C., Breitner, J., Steffens, D.C., Skoog, I., Rabins, P.V., 2011. Early parental death and remarriage of widowed parents as risk factors for Alzheimer disease: the Cache county study. American Journal of Geriatric Psychiatry 19, 814.
O'Rand, A.M., Hamil-Luker, J., 2005. Processes of cumulative adversity: childhood disadvantage and increased risk of heart attack across the life course. Spec No 2 The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 60, 117–124.
Painter, R.C., Roseboom, T.J., Bleker, O.P., 2005. Prenatal exposure to the Dutch famine and disease in later life: an overview. Reproductive Toxicology 20, 345–352.
Power, C., Kuh, D., Morton, S., 2013. From developmental origins of adult disease to life course research on adult disease and aging: insights from birth cohort studies. Annual Review of Public Health 34, 7–28.
Preston, S.H., Hill, M.E., Drevenstedt, G.L., 1998. Childhood conditions that predict survival to advanced ages among African-Americans. Social Science & Medicine 47, 1231–1246.
van Poppel, F., 2000. Children in one-parent families: survival as an indicator of the role of the parents. Journal of Family History 25, 269–290.
Ravelli, A.C.J., Van Der Meulen, J.H.P., Michels, R.P.J., Osmond, C., Barker, D.J.P., Hales, C.N., Bleker, O.P., 1998. Glucose tolerance in adults after prenatal exposure to famine. The Lancet 351, 173–177.
Risnes, K.R., Vatten, L.J., Baker, J.L., Jameson, K., Sovio, U., Kajantie, E., Osler, M., Morley, R., Jokela, M., Painter, R.C., Sundh, V., Jacobsen, G.W., Eriksson, J.G., Sorensen, T.I., Bracken, M.B., 2011. Birthweight and mortality in adulthood: a systematic review and meta-analysis. International Journal of Epidemiology 40, 647–661.
Roseboom, T., De Rooij, S., Painter, R., 2006. The Dutch famine and its long-term consequences for adult health. Early Human Development 82, 485–491.
Saigal, S., Doyle, L.W., 2008. An overview of mortality and sequelae of preterm birth from infancy to adulthood. Lancet 371, 261–269.
Smith, K.R., Mineau, G.P., Garibotti, G., Kerber, R., 2009. Effects of childhood and middle-adulthood family conditions on later-life mortality: evidence from the Utah population database, 1850–2002. Social Science & Medicine 68, 1649–1658.
Strehler, B.L., 2000. Understanding aging. Methods in Molecular Medicine 38, 1–19.
Strehler, B.L., Mildvan, A.S., 1960. General theory of mortality and aging. Science 132 (3418), 14–21.
Strully, K., Mishra, G., 2009. Theoretical underpinning for the use of sibling studies in life course epidemiology. Family Matters: Using Family-Based Studies to Determine the Mechanisms Underlying Early Life Determinants of Adult Chronic Diseases.
Swamy, G.K., Ostbye, T., Skjaerven, R., 2008. Association of preterm birth with long-term survival, reproduction, and next-generation preterm birth. Journal of American Medical Association 299, 1429–1436.
Wilcox, H.C., Kuramoto, S.J., Lichtenstein, P., Langstrom, N., Brent, D.A., Runeson, B., 2010. Psychiatric morbidity, violent crime, and suicide among children and adolescents exposed to parental death. Journal of the American Academy of Child and Adolescent Psychiatry 49, 514–523 quiz 530.

Early Memory Development and Cultural Variations

Monika Knopf, Frauke Graf, and Thorsten Kolling, Goethe-University Frankfurt, Frankfurt am Main, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

While infants' memory has been studied traditionally in adults through retrospective research, methodology in developmental psychology has made tremendous progress in the last decades so that nowadays infants' memory can be assessed directly. Following the taxonomy that separates a nondeclarative from a declarative memory system, different early memory aspects are under study. Although most evidence for early memory development exists for the Western cultural approach, there is a growing body of research about memory development in different cultures which is discussed in this article.

Neonatal and Infant Memory

The interest for infant memory has a long and extensive tradition in developmental psychology. Since Freud proposed that infant experiences strongly influence adult behavior, the infancy period is conceptualized as being very important for cognitive and behavioral development. Early developmental theories endowed newborns and infants with no or very limited memory capacities. Freud reasoned, for example, that early memories linger in an unconscious state and thus cannot be recalled. Similarly, Piaget assumed that infants younger than 18 months are not able to form mental representations precluding the construction of 'memory in a strict sense,' i.e., memories that are consciously and deliberately retrievable. In addition, the growing popularity of the infantile amnesia phenomenon, i.e., the typically observed inability of adults to remember personal life experiences having occurred before the age of 3.5 years (Bauer, 2007), increased the interest in infant memory development. For a long time, however, infant memory was underestimated and rather subject of theoretical reasoning as well as speculation than scientific evidence as a lack of age-appropriate methods precluded studying early memory capacities empirically.

By introducing the visual paired comparison paradigm, Fantz paved the way for a groundbreaking new empirical, behavior-based research program on cognitive and memory processes in early infancy. In a seminal study, Fantz (1964) demonstrated that infants between 2 and 6 months of age habituate to a repeatedly presented visual stimulus, i.e., they successively show decreased visual attention. In a visual test after habituation, infants prefer a novel stimulus to the simultaneously presented habituated, familiar stimulus. This finding indicates basic memory capacities, as a familiar stimulus needs to be represented in memory to be compared with a novel stimulus. Since this pioneering research, additional behavioral paradigms testing various aspects of memory in infants have been developed. These paradigms, rather directly assessing memory, are used to systematically describe the characteristics of the memory system of neonatal and young children (see section Direct Assessment of Memory in Infants and Young Children from Different Cultural Backgrounds).

A significant limitation of memory research in infants and preverbal children, as for psychological research in general, however, is that most of the memory studies conducted so far relate to newborns, infants, and preverbal children from a homogeneous sociocultural background, namely those living in educated and affluent families from Western, democratic, (post-) industrialized societies (see also Leichtman, 2011). Hence, the study of the effects of cultural contexts on memory systems' development in the early years is still in its infancy and future research should increasingly take into account cross-cultural differences and consistencies of newborn, infant, and preverbal children's memory development.

The present article therefore, first, reviews the memory system model used in state-of-the-art newborn and infant memory research (see Section Memory System Model). Secondly, in the next section two different sociocultural background models mostly used to describe cross-cultural memory research studies with infants are presented (see Section Cultural Models). Thirdly, the article describes different memory paradigms used to assess memory in infants and describes findings from cross-cultural developmental studies (see Section Cultural Approaches and Early Memory Development). Finally, some future research perspectives for cross-cultural memory research are outlined (see Sections Concluding Remarks and an Attempt to Forecast the Future of Cross-Cultural Research on Early Memory Development).

Memory System Model

The common theoretical memory systems perspective in infant memory research is the distinction between *nondeclarative (implicit)* and *declarative (explicit)* memory. Declarative memory includes explicit recognition or recall of facts and events, and thus represents the capacity to virtually remember, whereas nondeclarative memory comprises forms of memory that function without conscious awareness or deliberate retrieval such as priming, procedural or skill learning, and conditioning. Declarative memory is further divided into semantic memory referring to a person's general knowledge about the world including facts, concepts, and vocabulary, and episodic memory representing specific personal experiences which are associated with a particular learning context (i.e., place and time). Given this distinction, it can be assumed that due to their reliance on conscious experiences declarative memory is affected significantly stronger by sociocultural factors than nondeclarative memory processes (cf Leichtman, 2011).

Cultural Models

Growing up in a culture means being surrounded by socialization agents (e.g., the mother, other women, relatives, other caregivers) or being in close contact with individuals who, according to the specific cultural perspective, typically share their time with babies and children (e.g., siblings, same-aged peers, grandparents, teachers) either in a social institution (e.g., kindergarten) or rather in the familial context. Self-evidently, the interactions between developing individuals and significant socialization agents follow specific cultural conceptions and practices (e.g., ideas about the specific nature of children; concepts about child education; specific socialization goals; conceptions about acquisition of developmental milestones, values, and competences). In addition, specific cultural backgrounds are tied to specific physical environments characterized by typical living things (e.g., plants, animals) and determined by typical activities and routines involving culture-specific tools and artifacts (e.g., navigating a ship, planting something, visiting a restaurant; toys, needles, bowls, or PCs). Both physical and cultural world can be seen as complementary constraints shaping individual development. Beyond doubt, these micro- and macrocultural environments differ significantly for individuals growing up at different places in the world.

Culture and Cognitive Development in General

A traditional research program compares how infants, children, and adults from different cultural backgrounds typically solve cognitive tasks (e.g., mathematical thinking; category-based reasoning; holistic vs analytic reasoning; causal explanation; attend similarity and relationship between objects and rules). Moreover, in a related research program, cultural groups are studied in times of dramatic cultural changes. In the pioneering work of Luria, for example, syllogistic reasoning of different groups of Uzbek peasants living in Central Asia were compared during times of dramatic social and economic changes. Some of these groups lived traditionally and were illiterate while others attended schools for a different number of years. Cognitive changes significantly differing between cultures were described (see Nisbett and Norenzayan, 2002; for a detailed review). Memory processes, however, were not included in this traditional research program. As especially explicit memory is closely linked to other cognitive processes, the aforementioned studies and findings imply that memory processes should also vary as a function of cultural context.

Two Different Basic Cultural Approaches

In traditional studies on cultural influences on cognitive processes only rough ideas were sketched to characterize cultural contexts and to describe cultural differences. In recent years, however, suggestions to systematically distinguish basic cultural paradigms have been made. Cultural paradigms manifest in parental ethno-theories (i.e., beliefs, values, and practices of caretakers about child education and upbringing) resulting in different everyday, culture-specific infant experiences regarding, for example, motor stimulation, object exposure, and conversational styles. It is often stressed that parental ethno-theories differ according to infant and children's developmental milestones.

A first cultural approach to describe basic cultural models differentiates between two cultural developmental pathways, i.e., the pathway of interdependence versus independence. Within the interdependent cultural model, typically observed in rural agrarian areas in Cameroon or India, for example (Keller et al., 2006), essential socialization goals are the adaptation to family system values as social familial cooperative skills are crucial for surviving. In contrast, the independence cultural model, typically found in urban Western technological societies like Germany, other European countries, and the US emphasizes autonomous individual development, exploration ability, and personal achievement.

A second cultural approach distinguishes a Western Euro-Caucasian-American cultural style from an East Asia cultural style, typically observed in China, Japan, or Korea (e.g., Gutchess and Indeck, 2009). These cultural styles lead to cultural differences with respect to social development domains, mostly. Whereas from a Western Euro-Caucasian-American cultural style, the development of an autonomous self with unique personal characteristics is a central developmental goal, individuals raised in the East Asia cultural style perceive themselves in terms of their interpersonal relationships and their social roles. Next to social domains, cognitive domains are also determined by these different cultural styles. Individuals from Western societies tend to focus on objects, category relatedness, and self-relevant aspects of the situation. In contrast, individuals from East Asian cultures are more attracted by details of the context, from similarities between stimuli in a given situation as well as from information that is relevant for the group.

Cultural Approaches and Early Memory Development

Two methodological tools are currently used to study memory development in young children. On the one hand, the indirect assessment method assesses the emergence of memory by asking adults or older children to recollect their earliest memories. On the other hand, the direct assessment method, developed and used in the last couple of years, directly assesses memory in young children and even infants via different behavioral methods. Both research programs have also been used to study culture-specific early memory differences.

Culture-Specific Early Memories Assessed via the Indirect Assessment Method

The indirect assessment method has already been used to describe early memories for a long time (e.g., Freud). This research program was recently expanded in that the influence of culture for recollected memories was studied. Typically, adults from a Euro-Caucasian-American cultural background are compared with adults from Asian cultures in this research field. A significant limitation of this research methodology is the fact that memory contents have to be recollected verbally allowing to test only declarative memory aspects. Until now, research focuses on two closely related memory aspects.

On the one hand the phenomenon of *childhood amnesia* is cross-culturally assessed by describing the time when individuals from different cultural backgrounds are able to verbally report their first memories (e.g., MacDonald et al., 2000; Wang, 2008). On the other hand, a series of studies analyzed effects of cultural background on the development of individual's self-view. Referring to the memory model outlined above, this research program relates to the emergence and the development of *autobiographical memory* (Wang, 2011). Within both research programs, subjects are asked to explicitly and verbally recollect specific events and episodes. Besides memory processes, a couple of higher cognitive processes in general (e.g., language development, memory awareness, deliberate verbal recollection) are necessary as precondition to represent personal events in memory or to recollect the (possibly) stored memory entities.

Wang (2008) reviews recent studies looking at childhood amnesia from such a sociocultural perspective. By typically comparing the availability of earliest memories of individuals growing up in European American cultures versus different East Asian cultures (China, Taiwan, Korea), the universality or culture specificity of childhood amnesia was studied. In addition, in some studies, adults from a different cultural background but living in the same society have been compared (e.g., New Zealand European, New Zealand Asian, New Zealand Maori; MacDonald et al., 2000). Moreover, studies testing mostly young adults used different recollection methods assessing earliest memories (e.g., free recall, memory fluency test, word cue method), thereby testing, to which amount childhood amnesia is a function of low accessibility of early memories.

Although studies found some variability in the age of earliest memories as a function of recollection methods in that facilitative cues trigger earlier memories retrieval, cross-cultural studies consistently observed differences in the age of earliest memories. In line with theoretical reasoning, individuals growing up in a European or American culture remembered earliest memories between approx. 5 and 17 months earlier in their personal life as had individuals growing up in East Asia cultures.

Although these findings show that individuals from European American cultures seem to have earlier autobiographical memories, results also demonstrated culture-specific, qualitative differences of recollected episodes. In this respect, the character of adult's self-related stories and how adult's knowledge about the self is conceptualized differs significantly across cultural models. Euro-American adults, for example, recollect more specific, unique events and episodes from their lives that are highly centered to the self as unique individual (e.g., Leichtman, 2011; Wang, 2011). In contrast, adult Asians provided more generic, less detailed episodes that related to their social roles, their social relatedness, and their membership in specific groups. The dominant view about the emergence of these culture-specific differences in episodic memory has to do with the content and the degree of elaboration of memory conversation in early parent–child conversation. European American mothers held very elaborated memory conversations with their young children in which they ask them about their thinking, their feelings, their interests, thus providing an opportunity for the children to develop detailed memory episodes together with self-views as autonomous and unique individuals ('child-centered approach,' Wang, 2011). In contrast, East Asian mothers do not deliver that much detail in memory conversations with their children. Moreover, social interactions, moral rules, and behavioral expectations are often emphasized by Asian mothers ('mother-centered approach,' Wang, 2011). It is assumed that this mother-centered type of conversation may facilitate the recollection of events related to others and not so much the early storage of self-related information.

The theoretical assumption that mother guided memory talk is the basis of the more elaborated autobiographical memories of young European American children is strengthened by training studies. It has been found that young children of mothers, who were trained to be especially elaborative in their memory dialogue, provided more elaborative autobiographical memories than children in a control group of mothers who did not receive such a training (see Fivush, 2009).

Direct Assessment of Memory in Infants and Young Children from Different Cultural Backgrounds

In contrast to the measurement of memory in adults who can be asked to recollect information, infant memory research has to revert to methods based on simple behaviors, thereby requiring tasks that prompt the encoding, storage, and retrieval of information. Thus, memory is inferred from behavioral changes resulting from previous engagement with given materials. With regard to their mnemonic demands, these techniques can be classified as implicit or explicit memory tasks.

Implicit Memory

One major paradigm to assess implicit memory abilities in infancy is the *visual paired comparison paradigm* that involves exposing the infant to pairs of a stimulus and then showing this familiarized stimulus simultaneously with a novel one. Similarly, in the visual *habituation paradigm*, a stimulus is presented sequentially and either after fixation time decreased to a pre-set criterion (infant-controlled) or after a particular number of presentation trials (fixed), a novel stimulus is presented. In both paradigms, infants' looking time is assessed and a difference in fixation times to familiar compared to novel stimuli is taken as a measure of recognition memory. The most influential model, the comparator model, assumes that during the task, a memory trace of the visual stimulus is gradually constructed, information about the stimulus is retrieved from memory, and this memory traced is compared with the current visual input. Theoretically, it is still debated whether infants' preferences measure adult-like explicit memory capacities or implicit memory. Notwithstanding this debate, the assessment of infants' visual preferences yielded evidence for immediate as well as delayed recognition capacities showing an age-related increase in recognition memory during infancy. For example, Bushnell (2001) found that newborns prefer their mother's face to the face of an unfamiliar woman even if they had not seen their mother's face for 15 min. In addition, using the habituation method, Pascalis et al. (1998) demonstrated that 4-month-old infants are already capable of recognizing a previously unfamiliar face after a delay of 24 h.

Furthermore, the influence of stimulus complexity and meaning was demonstrated because 5-month-olds retain abstract patterns for 48 h and recognize faces after a 2-week retention interval (Fagan, 1990).

Studies applying visual preference paradigms within different cultural environments found no differences regarding rate of habituation, exhibition of novelty preference, or different measures of visual attention, such as fixation time, in infants, toddlers, and young children (e.g., Kennedy et al., 2008).

Classical and operant conditioning methods are also frequently applied to study learning and memory in infancy. In a *classical conditioning* study by Little et al. (1984), 10- to 30-day-old infants' eyelid response was conditioned by pairing an air puff (unconditioned stimulus) with a tone (neutral stimulus). After repeated presentations of these coupled stimuli, infants showed an eyelid response (conditioned response) when only the tone (conditioned stimulus) was presented, and 20-day-old infants exhibited savings even 10 days after the conditioning session in terms of requiring shorter presentations of the paired stimuli to evoke the conditioned response. Because neuropsychological studies confirm that classical conditioning is based on cerebellar processes, it can be concluded that it taps implicit memory.

Rovee-Collier and collaborators developed the mobile conjugate reinforcement procedure, an *operant conditioning* method to study learning and retention in 2- to 6-month-old infants. Within the mobile task, one leg of the infant is connected to an overhead mobile. In a first baseline phase, leg kicks do not generate movement of the mobile, whereas in the following reinforcement phase, leg kicks produce mobile movement. Learning of the contingency between own motor activity and movement of the mobile is indicated by an increased kick rate during reinforcement compared with baseline activity. The train task was developed as an age-adapted variant to assess whether infants older than 6 months of age learn to press a lever to move a miniature train. Studies applying the mobile and the train task have found that the duration of long-term memory increases linearly within the first two years of life: While 2-month-old infants show retention up to 2 days after training, those aged 6 months remember the mobile for 2 weeks, and 18-month-old infants retain the contingency over an interval of 13 weeks (for an overview, see Hartshorn et al., 1998). By implementing a reminder, the retention of 3-month-old infants was increased up to 4 weeks (Rovee-Collier, 1995). However, memory assessed by this task is quite specific because retention of the contingency is disrupted if training and test situation differ in any detail (e.g., Rovee and Fagen, 1976). Since learning of the contingency between motor activity and a visual event is attended by cerebellar activation, it is assumed to represent implicit memory (Nelson, 2002). However, Rovee-Collier (1997) argues that recognizing the mobile after the delay is expressed by a motor response and thus reflects an early form of explicit memory.

In a study by Graf et al. (2012), the mobile task was applied to assess contingency learning and associative memory in 3-month-old Cameroonian Nso farmer and German middle-class infants. Infants participated in two sessions at intervals of 24 h and when they reached a pre-set learning criterion during reinforcement they were tested for immediate and long-term retention. Results showed that 3-month-old infants in both cultural contexts are able to learn and to retain the relation between own leg kicks and the visual event indicated by an increased kick rate after reinforcement as well as after the 24 h delay compared with the baseline level of response. Nevertheless, culture-specific experiences influenced performance in this task: Although a higher proportion of Cameroonian Nso infants reached the critical increase during reinforcement and were thus classified as learners, this difference was not attributable to the advanced gross motor abilities of Nso infants but instead based on the level of baseline response. Depending on culture-typical experiences in motor handling (lying on the back with arms and legs free to move in Germany vs being carried in a sitting position in close body contact restricting movement of the limbs in Cameroon), German middle-class infants showed a higher level of baseline activity compared with Cameroonian Nso farmer infants impeding the achievement of the learning criterion. To conclude, this study demonstrated that cultural differences concerning infants' everyday experiences influence the level of motor response and hence affect learning performance in this task which is why characteristics of the context ought to be considered before applying this task across different cultural environments. With respect to the likely influence of culture on particular aspects of memory, this study confirms the assumption that implicit memory processes, here indicated by learning of the contingency and recognition of the mobile after the delay, do not display cultural memory differences.

Priming is also applied to assess implicit memory in verbal children and adults. In a prototypical priming task, the picture completion task, pictures are presented and after a delay ranging from seconds to months, the pictures are presented again but in a fragmented mode (i.e., only a fraction of information is presented). Priming is indicated when prior exposure to the pictures facilitates their subsequent recognition which can be measured by the reaction time or by the level of information required to name (i.e., to recognize) the picture. Regarding priming in early childhood, Gollin (1960), for example, demonstrated that 30- to 41-month-old children show a priming effect after a delay of 1 week. Several studies comparing priming in children at different ages and adults found no age-related effects (e.g., Parkin and Streete, 1988). Although priming as an implicit memory task is not assumed to be influenced by cultural factors, studies demonstrating its universality across cultures are largely missing.

An exception is the study by Vöhringer et al. (in preparation) who used a picture completion task to assess implicit memory in 3;4-year-old German middle-class and Cameroonian Nso children. Results showed that children from both cultural contexts exhibited a priming effect indicated by faster recognition of a prime (a picture they saw during a priming phase) compared to a nonprime. German middle-class children showed a higher recognition speed than Cameroonian Nso farmer children, potentially relying on extended experiences with two-dimensional pictures. Calculating a difference score between recognition speed of primes versus no-primes, however, revealed no differences between children from the two cultural contexts. To sum up, this study showed that although

recognition speed differs across cultural contexts, priming (implicit memory) is not affected by cultural context.

Taken together, results from cross-cultural studies show that visual paired comparison, habituation, mobile, and priming paradigms are only slightly or not influenced by culture-specific factors. These findings strengthen the theoretical idea that these tasks tap implicit memory processes.

Explicit Memory

A widely accepted instrument to assess memory in infants and toddlers is the *deferred imitation* task. In this task, the infant is demonstrated object-related target actions by a model. After a delay of minutes, hours, or even days, the infant has the opportunity to play with the objects and target action imitation behavior is observed. Memory thereby is inferred if infants reproduce a higher number of target actions after demonstration compared to spontaneous production of target actions. Spontaneous production of target actions is either assessed in a baseline phase before demonstration or via an independent control group receiving no demonstration of target actions. Theoretical reasoning and a series of experimental and clinical findings indicate that deferred imitation assesses declarative memory. In deferred imitation tasks, for example, subject transfer visually observed behavior into a specific motor output without prior practice (cross-modal character of the task). In clinical studies, results showed that patients with amnesia exhibit disrupted deferred imitation performance (see Mandler, 2004; for overview).

Numerous studies using the deferred imitation task found that retention increases from 24 h in 6 month olds to 4 months in infants who were presented the target actions at 16 months of age (Barr et al., 1996; Meltzoff, 1995). Moreover, it was demonstrated that the flexibility of infants' memory also increases with age as 12- in contrast to 6-month-old infants exhibited significant imitation even after a pronounced context change, i.e., a different testing room (Hayne et al., 2000).

Cross-culturally, imitation is assumed to serve as a universal learning mechanism. Goertz et al. (2011), for example, analyzed deferred imitation performance in a sample 6-month-old Cameroonian Nso and German middle-class infants. They found deferred imitation in infants from both cultural environments confirming the assumed universality of imitative learning. In a subsequent study, deferred imitation of an extended sample of Cameroonian Nso and German middle-class infants of 6 and 9 months of age was analyzed (Graf et al., 2013). It was hypothesized that with increasing age and age-associated accumulation of culture-specific experiences, cultural differences in deferred imitation would be observed.

Results show that infants at both ages and from both cultures performed more target actions after demonstration compared with the baseline phase. Furthermore, an age-related increase in the number of imitated target actions was found indicating an improvement of memory with age. Moreover, at 9 months of age, German infants showed a higher level of spontaneously performed target actions than infants in the Cameroonian sample which is assumed to rely on the different extent of object experiences. These findings add evidence to the theoretical assumption that explicit memory is influenced by characteristics of the cultural context.

Concluding Remarks and an Attempt to Forecast the Future of Cross-Cultural Research on Early Memory Development

Taken together, the findings reviewed in this article demonstrate that early explicit memory as it is found in adults' recollections is determined by the sociocultural model. The pronounced cultural developmental differences found in recollections are qualitative (self-reliance during autobiographical report studies) as well as quantitative (time where subjective memory begins) in nature.

In infants and young children, cross-cultural studies assessing the emergence of memory directly are still rare. The findings of recent studies, however, demonstrate that explicit memory (deferred imitation) is influenced stronger by cultural background than implicit memory (priming, mobile task). In line with findings from childhood, the small cultural differences are more due to qualitative than quantitative aspects of memory.

From a developmental standpoint, comparing the findings from the childhood with the infancy period indicates that sociocultural influences on implicit, explicit, and autobiographical memory increases with age, i.e., infant memory seems to rely stronger on biologically driven processes than memory during childhood years. In other words, effects of the external world (Luria, Bronfenbrenner) on memory increase in the sense of an age × culture interaction framework.

The finding that differences in memory performance increase with increasing age of children is also in line with the idea that different developmental aspects that are codeterminants of explicit memory, like language development, motor development, self-development, are also sensitive to differences in the cultural paradigms. This means, the cultural context in which an infant grows up influences memory development directly as well as indirectly.

For future cross-cultural developmental memory research, disentangling biologically driven, culturally driven, and methodologically driven effects is important to understand the true nature of memory development. This challenge also requires new, culture-fair instruments which are assessing memory performances that are asked for to the same amount within the different cultural contexts.

See also: Autobiographical Memory During Early Childhood and Cultural Variations; Cognitive Development During Infancy and Early Childhood across Cultures; Memory Development During Infancy and Early Childhood across Cultures; Short- and Long-term Memory During Infancy and Early Childhood across Cultures.

Bibliography

Barr, R., Dowden, A., Hayne, H., 1996. Developmental changes in deferred imitation by 6- to 24-month-old infants. Infant Behavior and Development 19, 159–170.

Bauer, P.J., 2007. Remembering the Times of Our Lives: Memory in Infancy and Beyond. Erlbaum, Mahwah, NJ.

Bushnell, I.W.R., 2001. Mother's face recognition in newborn infants: learning and memory. Infant and Child Development 10, 67–74.

Fagan, J.F., 1990. The paired-comparison paradigm and infant intelligence. In: Diamond, A. (Ed.), The Development and Neural Bases of Higher Cognitive Functions, vol. 608. Academy of Sciences Press, New York, pp. 337–364.

Fantz, R.L., 1964. Visual experience in infants: decreased attention to familiar patterns relative to novel ones. Science 146, 668–670.

Fivush, R., 2009. Sociocultural perspectives on autobiographical memory. In: Courage, M.L., Cowan, N. (Eds.), The Development of Memory in Infancy and Childhood. Psychology Press, Hove, pp. 283–301.

Goertz, C., Lamm, B., Graf, F., Kolling, T., Knopf, M., Keller, H., 2011. Deferred imitation in 6-month-old German and Cameroonian Nso infants. Journal of Cognitive Education and Psychology 10 (1), 44–55.

Gollin, E.S., 1960. Developmental studies of visual recognition of incomplete objects. Perceptual and Motor Skill 11, 289–298.

Graf, F., Borchert, S., Lamm, B., Goertz, C., Kolling, T., Fassbender, I., Teubert, M., Vierhaus, M., Freitag, C., Spangler, S., Keller, H., Lohaus, A., Schwarzer, G., Knopf, M., 2013. Imitative learning of Nso and German infants with 6 and 9 months of age: evidence for a cross-cultural learning tool. Journal of Cross-Cultural Psychology. Advance online publication. http://dx.doi.org/10.1177/0022022113487075.

Graf, F., Lamm, B., Goertz, C., Kolling, T., Freitag, C., Spangler, S., Fassbender, I., Teubert, M., Vierhaus, M., Keller, H., Lohaus, A., Schwarzer, G., Knopf, M., 2012. Infant contingency learning in different cultural contexts. Infant and Child Development 21, 458–473. http://dx.doi.org/10.1002/icd.1755.

Gutchess, A.H., Indeck, A., 2009. Cultural influences on memory. Progress in Brain Research 178, 137–150. http://dx.doi.org/10.1016/S0079-6123(09)1780-3.

Hartshorn, K., Rovee-Collier, C., Gerhardstein, P., Bhatt, R.S., Wondoloski, T.L., Klein, P.J., Gilch, J., Wurtzel, N., Campos-de-Carvalhos, M., 1998. The ontogeny of long-term memory over the first year-and-a-half of life. Developmental Psychobiology 32, 69–89.

Hayne, H., Boniface, J., Barr, R., 2000. The development of declarative memory in human infants: age-related changes in deferred imitation. Behavioral Neuroscience 114, 77–83.

Keller, H., Lamm, B., Abels, M., Yovsi, R., Borke, J., Jensen, H., et al., 2006. Cultural models, socialization goals, and parenting ethnotheories. Journal of Cross-Cultural Psychology 37, 155–172. http://dx.doi.org/10.1177/0022022105284494.

Kennedy, T., Thomas, D.G., Woltamo, T., Abebe, Y., Hubbs-Tait, L., Sykova, V., et al., 2008. Growth and visual information processing in infants in Southern Ethiopia. Journal of Applied Developmental Psychology 29, 129–140.

Leichtman, M.D., 2011. A global window on memory development. In: Jensen, L.A. (Ed.), Bridging Cultural and Developmental Approaches to Psychology. University Press, Oxford, pp. 49–71.

Little, A., Lipsitt, L., Rovee-Collier, C., 1984. Classical conditioning and retention of the infant's eyelid response: effects of age and interstimulus interval. Journal of Experimental Child Psychology 37, 512–524.

MacDonald, S., Uesiliana, K., Hayne, H., 2000. Cross-cultural and gender differences in childhood amnesia. Memory 8, 365–376.

Mandler, J.M., 2004. A synopsis of the foundations of mind: origins of conceptual thought. Developmental Science 7, 499–505.

Meltzoff, A.N., 1995. What infant memory tells us about infantile amnesia: long-term recall and deferred imitation. Journal of Experimental Child Psychology 59, 497–515.

Nelson, C., 2002. The neurobiological basis of early memory development. In: Cowan, N. (Ed.), The Development of Memory in Childhood. Psychology Press, Hove, pp. 41–82.

Nisbett, R.E., Norenzayan, A., 2002. Culture and cognition. In: Medin, D.L. (Ed.), Stevens' Handbook of Experimental Psychology, Memory and Cognitive Processes, third ed., vol. 2. John Wiley & Sons, Inc., New York, pp. 561–597.

Parkin, A.J., Streete, S., 1988. Implicit and explicit memory in young children and adults. British Journal of Psychology 79, 361–369.

Pascalis, O., de Haan, M., Nelson, C., de Schonen, S., 1998. Long-term recognition memory for faces assessed by visual paired comparison in 3- and 6-month-old infants. Journal of Experimental Psychology Learning 24, 249–260.

Rovee, C., Fagen, J.W., 1976. Extended conditioning and 24-hr-retention in infants. Journal of Experimental Child Psychology 21, 1–11.

Rovee-Collier, C., 1995. Time windows in cognitive development. Developmental Psychology 31, 147–169.

Rovee-Collier, C., 1997. Dissociations in infant memory: rethinking the development of implicit and explicit memory. Psychological Review 104, 467–498.

Vöhringer, I., Borchert, S., Fassbender, I., Suhrke, J., Teiser, J., Teubert, M., Freitag, C., Lamm, B., Lohaus, A., Schwarzer, G., Keller, H., Knopf, M., submitted. Implicit Memory in 3;4-year-old Cameroonian Nso farmer and German middle-class children.

Wang, Q., 2008. Where Does Our Past Begin? A Sociocultural Perspective on the Phenomenon of Childhood Amnesia. Retrieved November 20, 2012 from http://www.apa.org/science/about/psa/2008/03/wang.aspx.

Wang, Q., 2011. Autobiographical Memory and Culture. Online Readings in Psychology and Culture. Unit 5. Retrieved November 24, 2012 from http://scholarworks.gvsu.edu/orpc/vol5/iss2/2/.

Earnings Inequality

Andreas Peichl, ZEW Mannheim, University of Mannheim, IZA, ISER; and CESifo, Germany
Nico Pestel, IZA Bonn and ZEW Mannheim, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Inequality has increased considerably in many Western countries over the past decades. When dealing with economic inequality as a research subject the question "inequality of what among whom?" arises. Analyses of inequality are typically concerned with the distribution of wages, earnings, or income and have been made by different strands in the literature, mainly in public and in labor economics. We summarize these strands with a special focus on earnings, which itself is the product of hourly wages and labor supply in terms of hours and weeks worked. In addition to inequality in labor market outcomes, we pay special attention to equality of opportunity.

Introduction

Inequality has increased considerably in many Western countries over the past decades. (see Gottschalk and Smeeding, 1997; Atkinson and Piketty, 2007; OECD, 2008, 2011; Atkinson, 2008; Salverda et al., 2009, for overviews.) The growing gap between rich and poor and the resulting declining middle class has become one of the main issues in policy agendas around the world. The economic crisis that followed in the aftermath of the collapse of the financial market in the United States in 2008, has rendered issues concerning the distribution of economic resources in general, and questions of the appropriateness of extremely high earnings in particular, urgent (OECD, 2011: p. 17). Austerity measures in the context of the euro crisis have recently triggered social unrest in countries like Greece and Spain where these measures are perceived to affect the poor disproportionately. The 'Occupy Wall Street' movement, which pressed policy makers for steps against growing social and economic inequality, has popularized the catchphrase "We are the 99%." Interestingly, this slogan directly refers to academic research on the increasing income share of the richest 1% of the US population, which is nowadays back to historically high levels (Piketty and Saez, 2003, 2007). Top income shares are not only increasing in the United States but also in many other countries (Piketty and Saez, 2003; Atkinson and Piketty, 2007; Atkinson et al., 2011). These examples show that the distribution of economic resources across the population is not just a matter for public debate and policy making. On the contrary, the analysis of distribution is long since "back in from the cold" (Atkinson, 1997) and has turned from "watching the grass grow" (Aaron, 1978) to become an active and relevant area of research in economics. In this article, we summarize the literature with a special focus on earnings. In addition to inequality in labor market outcomes, we pay special attention to equality of opportunity.

Why Should Economists Care about Inequality?

Economists should care about inequality even if inequality itself were not of great interest, because there are a number of important implications that come with it. For example, many economists argue that inequality is not a bad thing per se. On the contrary, inequalities in relative factor prices are fundamental to the functioning of market economies – as evidenced by the collapse of real-existing socialist economies. With special focus on labor markets, Welch (1999) emphasized that inequalities in wages are 'good' since they signal scarcities, provide incentives for investments in human capital, and compensate for different job attributes. However, Welch himself states that inequality becomes 'destructive' when society does not view effort as worthwhile and upward mobility is perceived as unlikely or even impossible. In general, public opinion in market economies shares the economists' view that absolute equality in economic outcomes is not desirable and that inequalities are, to a certain degree, not only inevitable but even necessary (Salverda et al., 2009: p. 7). However, if income differences are viewed as insurmountable, social cohesion as well as acceptance of market economy and even democracy are challenged (Stiglitz, 2012). Indeed, preferences for redistribution are systematically correlated with beliefs about the relative importance of effort and luck in the determination of outcomes. Individuals are more willing to accept income differences which are due to individual effort (or laziness) rather than exogenous circumstances or luck (Fong, 2001).

From a welfare economics perspective, a normative reason to be interested in the distribution of income is that one is actually interested in the distribution of 'well-being' and income is widely used as a proxy for well-being (Decancq et al., 2014). However, the adequacy of income as a welfare metric has recently been questioned and a consensus seems to have emerged that well-being has multiple dimensions beyond income that should be considered simultaneously – typically indicators for health, job quality, wealth, and many more (see, e.g., Stiglitz et al., 2009, for an overview). Hence, focusing on the distribution of labor earnings, which make up a substantial but not the entire part of income, is even more restrictive in terms of welfare analysis. This has to be taken into account when interpreting results on earnings inequality with respect to their policy implications.

Inequality of What among Whom?

When dealing with economic inequality as a research subject the question "inequality of what among whom?" arises (Atkinson and Bourguignon, 2001; Osberg, 2001; Goldfarb

and Leonard, 2005). The answer to the part query "among whom?" is straightforward for economists. The term 'economics' dates back to the ancient Greek word *oikos* which means household. Hence, the essence of the economics discipline is the study of the smallest unit of individuals within an economy, jointly carrying out production and consumption activities. Nevertheless, in studying earnings inequality, the unit of observation is very often the individual or the tax unit and not the actual household. In the case of a single household, all three levels of observation are identical, but this does not hold in general. For example, married couples form one household but consist of two individuals; and depending on the income tax system, the household may consist of either one or two tax units. The presence of children (or other family or household members) adds further individuals and/or tax units to the household. Given the trend toward more single persons, lower marriage, and higher divorce rates in many Western societies, the congruence between individual, tax unit, and household is changing tremendously and this has implications for the measurement of inequality.

Indeed, Armour et al. (2013) document the sensitivity of different income measures in capturing income trends. The unit of analysis (individual or tax unit vs household) and also the income concept, i.e., looking at pretax, pretransfer income versus posttax, posttransfer income, yield different results and trends. Especially, the question of whether and how capital gains are included in a measure of market income matters a lot, both for inequality levels and trends. This is related to the question of "what?," the underlying concept of economic resources, which is much more complex. Analyses of inequality are typically concerned with the distribution of wages, earnings, or income and have been made by different strands in the literature, mainly in public and in labor economics. However, there are "several steps between relative factor prices and […] disposable income among households" (Atkinson, 2003: p. 23). The most important steps in this process are the creation of gross market income from various sources and all household members, the design of the government's tax and transfer system, as well as patterns of household formation and composition.

Firstly, gross labor earnings make up the largest share of total household incomes and are an important driver of income inequality (Atkinson, 2008). A vast literature in labor economics deals with rising wage and earnings dispersion, especially for the United States (see Katz and Autor, 1999, for an overview). Common explanations are changes in the supply and demand for skills and tasks as well as changing labor market institutions and policies. Autor et al. (1998, 2008) argue that 'skill-biased' technological change (computerization) within industries has led to 'skill-upgrading' and, hence, has increased the relative demand for college-educated workers. At the same time, skill-biased technological change may not only have complemented high-level education tasks but also substituted for routine middle-level education tasks. Card and Lemieux (2001) point to shifts in the composition of the workforce with respect to education and show that a slowdown in educational attainment among younger cohorts has increased the college–high school wage gap for men due to shifts in the relative supply of highly educated men. Goldin and Katz (2008) take a longer perspective and emphasize the importance of the supply and demand of human capital in forming the distribution of earnings over the twentieth century in the United States. Acemoglu and Autor (2011, 2012) analyze the interactions among worker skills, job tasks, evolving technologies, and shifting trading opportunities.

Another strand in the labor economics literature argues that the skill-biased technological-change hypothesis put forward does not explain the evolution of other dimensions of wage inequality, including the gender and racial wage gaps and the age gradient in the return to education (Card and DiNardo, 2002). In addition to explanations referring to the supply and demand of skills, many studies have analyzed the role of labor market policies and institutions. DiNardo et al. (1996) highlight the importance of the decreasing real value of the federal minimum wage as well as the importance of deunionization in addition to labor market considerations. Fortin and Lemieux (1997) analyze the role of the decline in the real value of the minimum wage, deunionization, and economic deregulation and find that a large share of the increasing wage inequality is related to these institutional changes. Lemieux (2006) shows that the magnitude and timing of growth in residual wage inequality provide little evidence of an increase in the demand for skill due to skill-biased technological change – but points out that this is due to composition effects associated with higher within-group dispersion. In addition, differences in wages and earnings are affected by pay differentials across gender, race, occupations, or sectors. See Altonji and Blank (1999) for an overview of gender and racial pay gaps. Blau and Kahn (2000, 2006) provide overviews of the development of the gender pay differential and analyze different driving forces behind the extent and speed of reductions in the gender wage gap. Kunze (2005) analyzes the gender wage gap over different career stages in Germany. Arulampalam et al. (2007) analyze the wage gap over the entire wage distribution for European countries.

Secondly, other determinants of (wage or earnings) inequality are the tax-transfer system and patterns of household formation and composition (see Lam, 2001; Pestel, 2013, for overviews). For instance, Peichl et al. (2012) quantify how the trend toward smaller households has influenced the change in income distribution in Germany, using decomposition methods for measures of inequality, poverty, and richness. The results show that the income gap would have increased even without the demographic trend. But its level would be lower than it actually is. In addition, the demographic effect turns out to be larger for incomes before taxes and benefits, showing that there is an interaction between tax policy and household composition. Bargain et al. (forthcoming) analyze the effect of US tax policy on inequality over the period 1979–2007. The challenge in analyzing such a question lies in the fact that both taxation and pretax income distribution influence income redistribution. For instance, a progressive tax system – one in which the tax rate increases with the taxable base – has a greater effect on income redistribution when a society's gross income is more unequally distributed. In the extreme case that all taxpayers earn the same income, even a very progressive system cannot redistribute income among taxpayers. To distinguish between effects that result from taxation and those that result from changes in pretax income

distribution, the authors conducted detailed counterfactual calculations, holding constant the income distribution of one year and applying the tax system of another year. The results show that American tax policy has done little to curb inequality in the period under investigation.

Finally, total disposable household income depends on the household context, i.e., the number, composition, and characteristics of individuals actually forming households. For given wages and labor market conditions, gross earnings depend on the number of hours worked. Hence, the household context, which has changed tremendously over the past decades, determines the distribution of resources both within and across households in the economy. See, for example, Jenkins (1995), who analyzes various sources of the trend in income inequality in the U.K. during the 1970s and 1980s, including, employment, earnings, and household composition. Cancian and Reed (1998, 1999) study the role of female earnings on inequality and find an equalizing effect of increasing female labor force attachment. Burtless (1999, 2009) find that the increasing correlation between husbands' and wives' earnings as well as the increasing numbers of single-person households, has contributed to more inequality. Hyslop and Mare (2005) also find that increasing inequality in New Zealand is to a large extent driven by changes in household structure and attributes. Daly and Valletta (2006) and Martin (2006) find similar results and trends for the United States; and Schwartz (2010) addresses the increasing association between spouses' earnings as an important driver of overall inequality.

The observed distribution of disposable income is, however, not simply a matter of mechanically applying the tax and transfer schedule to gross incomes for a given household composition but the result of complex interactions between the market production of gross income (joint decisions on labor supply and savings) and the formation of households (marriage, cohabitation and fertility decisions, aging, and retirement). Therefore, it is an enormous challenge to formulate "models of the household income distribution, incorporating not only models of labor market earnings […] and the demographic factors affecting who lives with whom" (Jenkins and Micklewright, 2007: p. 19). This issue is, however, beyond the scope of this article.

Labor Earnings Inequality

Evidence for the United States

There are various measures of inequality with very different features and it is for the researcher to decide which measure is most appropriate with regard to the research subject under consideration. We do not discuss the different measures here and suggest reference to the literature (Osberg, 2001; Cowell, 2008). Instead, we follow a recent strand in the literature dealing with the contribution of top incomes to overall inequality, showing that large shares of total pretax (wage) income are increasingly concentrated among the rich (Piketty and Saez, 2003; Atkinson and Piketty, 2007; Atkinson et al., 2011). Top income shares are fairly simple and straightforward measures of inequality and easy to interpret. Although they focus by definition on small shares

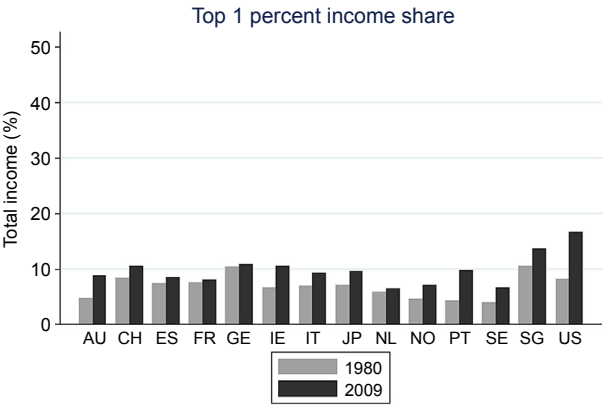

Figure 1 Top 1% income shares in various OECD countries (1980 and 2009). Source: The Worlds Top Income Database.

of the population, say the top 10%, the top 1%, or even smaller fractiles of the distribution, they have been found to be very informative about overall inequality, since trends over time match very well with other inequality measures (see Leigh, 2007).

Figure 1 shows the top 1% income share for various OECD countries in 1980 and 2009. Inequality increased in all countries. Both the share in 2009 and the increase from 1980 are largest in the United States. The lowest level of inequality can be found in the Nordic countries, Norway and Sweden. Only in a few countries the increase is found to be fairly small (like Spain, France, Germany, and the Netherlands). Nevertheless, increasing inequality, especially at the top of the distribution, can be observed globally and is, therefore, an issue for researchers and policymakers in various institutional contexts.

Figure 2 shows the development of the wage income share of the top 10% in the United States starting in the 1920s, broken down into fractiles within the top decile. Before the World War II (WWII) period, the share of pretax wage income earned by the richest decile of the population was around 30%; and 8–9% was accounted for by the top 1%. During the WWII period, the top decile share dropped to around 25% (top 1%: 5–6%) and remained fairly constant until the late 1970s and early 1980s. Since then, one can observe a strong increase to almost more than 35% and, hence, exceeding the historical pre-WWII levels. Remarking is the contribution of the top 1% to this trend, which is nowadays up to around 11%, while the share of the top decile except for the top 1% (P90–P99) has only slightly increased over this period. This highlights the importance of (gross) labor income as an important driver of overall inequality, also at the very top of the US income distribution.

Components of Earnings Inequality

The overall distribution of labor earnings within a given time period, a given country, and for a specific population subgroup is determined by various factors. Typically, research on earnings inequality is concerned with a measure on an annual

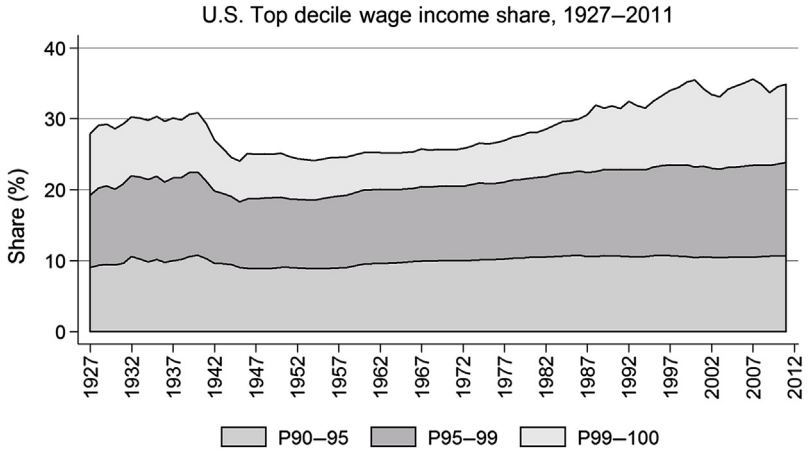

Figure 2 Top 10% wage income share in the U.S. (1927–2011). Source: Piketty, T., Saez, E., 2003. Income inequality in the United States, 1913–1998. Quarterly Journal of Economics 118 (1), 1–39 and Piketty, T., Saez, E., 2007. How progressive is the US federal tax system? A historical and international perspective. Journal of Economic Perspectives 21 (1), 3–24 updated to 2012: http://elsa.berkeley.edu/~saez/TabFig2012prel.xls. Wage income includes bonuses, and profits from exercised stockoptions.

(sometimes monthly or weekly) basis. Overall earnings can be decomposed into several elements, since they are simply the product of the hourly wage rate, the number of hours worked per week, and the number of weeks worked per year:

$$\text{Annual earnings} = \frac{\text{wage}}{\text{hour}} * \frac{\text{hours}}{\text{week}} * \frac{\text{weeks worked}}{\text{year}}$$

Hence, the distribution of earnings is potentially affected by all of these components and their underlying subcomponents.

First, the wage rate reflects the conditions of the respective labor market. On the one hand, relative supply and demand for different levels of skills play an important role, especially trends of skill-biased technological change and globalization affecting the demand for low-skilled work and routine jobs. On the other hand, labor market policies and wage-setting institutions, such as minimum wage policies and unionization, determine the extent of wage compression for the given labor supply.

Figure 3 shows the trend in real hourly wage rates for men and women from 1976 to 2011. For men, only the wages of the top 10% of the earnings distribution increased while the real wages of the bottom 50% decreased. For women, in contrast, the wages for all groups increased – resulting in a declining gender wage gap. Yet, the increases were stronger at the top of the distribution. This means that the labor market has become more polarized in the U.S. for both men and women. Since labor earnings account for a large part of incomes, this is an important driver of increasing levels of total income inequality.

Secondly, annual earnings are affected by the individual's decision on how many hours of work per week to supply on the labor market. Patterns of labor supply have considerably changed over the past decades. This is especially true for women, who have substantially increased their educational attainment and are therefore much more attached to the labor

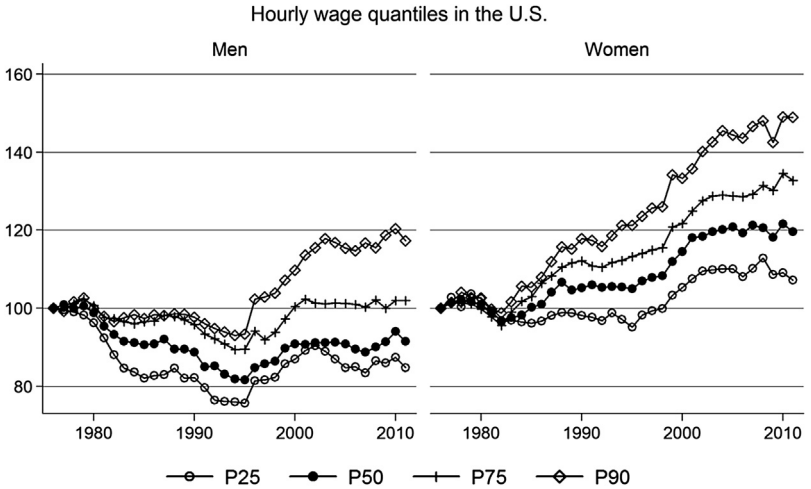

Figure 3 Hourly wage rates in the United States for men and women (1976–2011). Source: March CPS 1976–2011, own calculations. Population aged 21–65 years. Base year 1976 = 100.

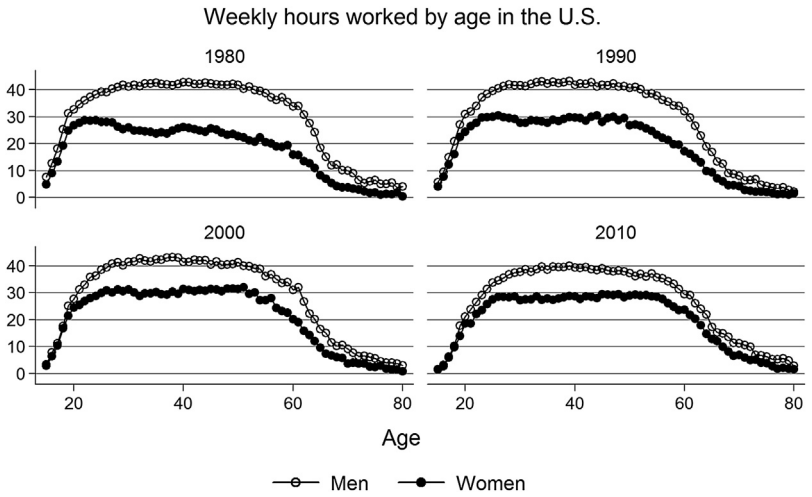

Figure 4 Weekly working hours for men and women by age (1980–2010). Source: March CPS, own calculations.

market. In addition, in many countries the increase of female labor force participation, both at the extensive and intensive margins, has been more concentrated at the top of the distribution, which increases the gap in total household earnings between 'rich' and 'poor' couples.

Figure 4 shows the average weekly working hours for men and women of different ages for the years 1980, 1990, 2000, and 2010 in the United States. Several observations can be made. First, prime-aged males always work 40 h per week and more. Secondly, women increased both the average weekly hours as well as their retirement age. Hence, while the pattern of labor supply characteristics of men has remained almost unchanged over the 30-year period, the pattern for women changed remarkably and has become much more similar to that of men.

This is also related to patterns of increasingly assortative mating ("*Doctors marrying doctors, rather than nurses,*" OECD, 2011) and increasing correlation of labor earnings and hours worked among couples. More generally, patterns of household formation and composition have changed tremendously in many Western countries and are related to increasing inequality (see, e.g., Peichl et al., 2012, for an analysis on Germany).

Finally, the number of weeks worked per year reflects fluctuations in employment, which are related to unemployment fluctuations over business cycles. This makes the distribution of gross earnings more unequal; the risk of layoff during an economic downturn is more concentrated at the bottom of the wage distribution. Unemployment rates and earnings inequality are positively correlated.

Figure 5 shows the employment rate (left scale) and average weekly working hours (right scale) for United States men and women over time. While the male employment rate is relatively stable around 80%, total working hours show some variation over the business cycle, the hours worked conditional on being employed are stable and on average above 40 h per week. For women, the number of hours worked conditional on

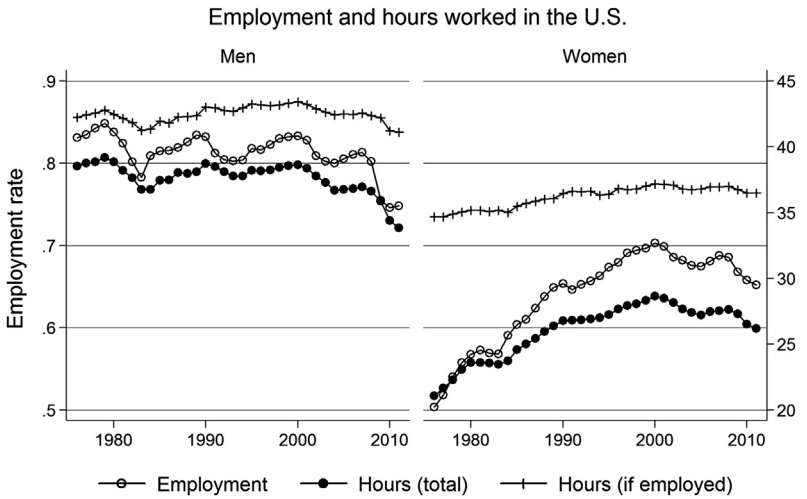

Figure 5 Employment rates and hours worked by men and women (1976–2011). Source: March CPS 1976–2011, own calculations. Population aged 21–65 years.

employment only marginally increased around the level of 35 h, while total weekly hours increased substantially from 20 to 25 h along with the employment rate from 50% in the mid 1970s to around 70% in recent periods. While long-run trends as well as short-run business cycle fluctuations in labor force participation and employment are of primary concern for labor market policies, they have also important implications for the distribution of labor earnings, since the underlying working force's composition crucially shapes the distribution for given labor market conditions.

Equality of Opportunity

The literature on earnings inequality deals, by definition, with inequalities in (labor market) outcomes. However, theories of distributive justice distinguish ethically acceptable inequalities (e.g., due to differences in effort) from unfair inequalities (e.g., due to endowed characteristics). The concept of equality of opportunity in contrast to equality of outcomes has received considerable attention since the seminal contributions of Roemer (1993, 1998), Van de Gaer (1993) and Fleurbaey (1995). See, e.g., Ramos and Van de Gaer, 2012; Roemer and Trannoy, 2013 for overviews of the equality of opportunity literature. While the traditional notion of equality of outcomes refers to an equal distribution of economic outcomes (e.g., well-being, consumption, or income) across the population, the equality of opportunity theory, in contrast, is interested in the sources of inequality and separates the influences on the outcomes of an individual into circumstances and effort. Circumstances are defined as all factors beyond the sphere of individual control for which society deems individuals should not be held responsible, such as parental education, gender, or ethnic origin. Effort, on the other hand, comprises all actions and choices within individual responsibility for which society holds the individual (partially) accountable, e.g., schooling or labor supply decisions. Inequalities (in income) due to differences in effort are deemed equitable, whereas inequalities due to endowed circumstances are not.

This is related to the literature on wage discrimination (see, e.g., Altonji and Blank, 1999, for an overview). However, a fundamental difference exists between the two fields. Labor economists studying discrimination are usually interested in estimating the direct effect of endowed characteristics (e.g., race, gender) on income and try to separate it from confounding effects due to between-group differences in effort. In contrast, the 'equality of opportunity' literature believes that the confounding indirect effect is also a source of unfair inequalities – i.e., a circumstance itself should not be separated from the direct effect of circumstances on income (see, e.g., the discussion in Roemer, 1998). Therefore, inequality of opportunity is related to wage discrimination, but it is not the same. Unfair income differences in the inequality of opportunity framework can be indeed caused by discrimination, but they could also be due to between-group differences in productivity or preferences. Therefore, the two approaches imply different normative choices about the compensation of the indirect (confounding) effect.

Niehues and Peichl (2014) analyze inequality of opportunity for Germany and the United States. **Figure 6** presents the range for inequality of opportunity shares, i.e., the inequality of opportunity level divided by the level of outcome

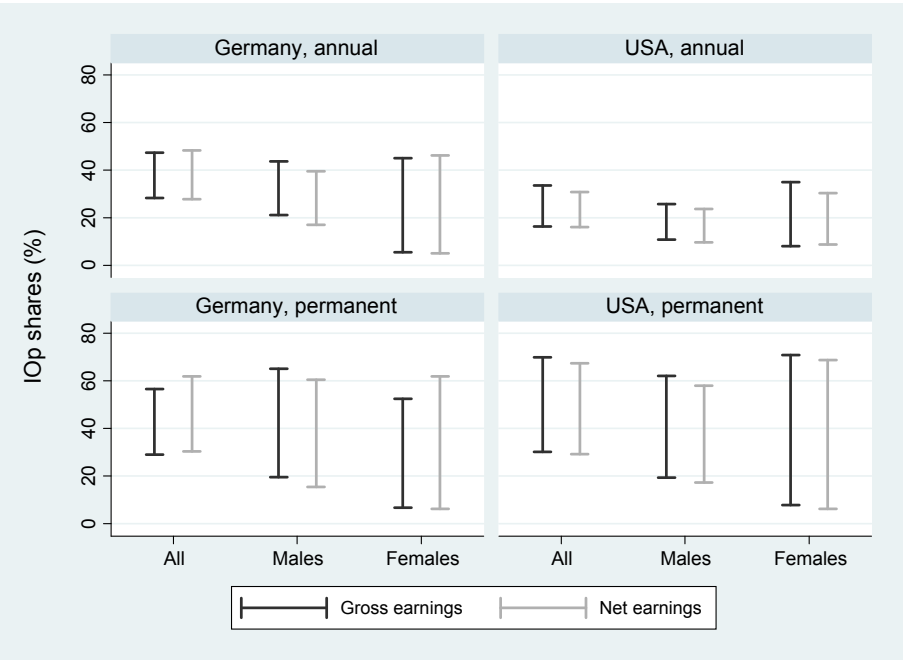

Figure 6 Inequality of opportunity shares in outcome inequality. Source: Niehues, J., Peichl, A., 2014. Upper Bounds of Inequality of Opportunity: Theory and Evidence for Germany and the US. Social Choice and Welfare, 43 (1), 73–99. based on SOEP and PSID. The two graphs on the top illustrate inequality of opportunity shares in annual incomes (2009 for Germany, 2007 for the United States); the graphs at the bottom depict inequality of opportunity shares in permanent incomes.

inequality (between group inequality as a fraction of total inequality). The upper (lower) line corresponds to the upper (lower) bound share. Results are presented both for periodical (permanent) incomes in the upper (lower) panel both for the United States and Germany for the full sample, as well as separated by gender for gross (left, darker bar) and net (right, lighter bar) earnings.

The inequality of opportunity shares are significantly higher for Germany than for the United States for annual incomes, which is due to lower absolute levels of outcome inequality along with similar values of inequality of opportunity – which is in line with the findings of Almas (2008). The lower bound shares equal 30% in Germany and 16% in the U.S. – the latter is comparable to previous findings (Pistolesi, 2009). Based on these results, it would be possible to deduce that individual earnings are mainly driven by the individual's effort choices and only to a lesser extent by circumstances. The upper bound estimates, however, suggest that earnings are to a larger extent predetermined by exogenous circumstances. Niehues and Peichl (2014) find upper bounds of inequality of opportunity of around 50% in Germany and 35% in the United States. The differences are statistically significant.

Thus, it seems that there is substantially less inequality of opportunity in the U.S. compared to Germany, i.e., one could conclude that equality of opportunity is higher in the 'land of opportunity.' However, using permanent instead of annual incomes matters for inequality levels, especially in the United States, where inequality of opportunity levels are much higher for permanent incomes (comparable to the findings of Pistolesi, 2009). In Germany, the difference between inequality levels for the two income concepts is much smaller. Therefore, inequality levels (and hence the inequality of opportunity shares) are similar for both income concepts. Hence, the inequality of opportunity shares for permanent incomes are higher in the United States than in Germany.

The lower bound inequality of opportunity shares are substantially smaller when looking at the female and male samples separately. This hints at gender as an important source of inequality of opportunity (due to gender wage gaps and differences in labor force participation). However, the effect is not as strong for the upper bounds based on the unit-effect, as circumstances vary. This indicates that a large share of the inequality in outcomes can be explained by unobserved heterogeneity in individuals.

The differences between gross and net income inequality, i.e., the redistributive effects of the tax benefit systems, are rather similar in Germany and the United States. This might be surprising at a first glance, since European welfare states are usually said to be more redistributive. But in the present case, which focuses on the working age population, this is not the case. The main difference in redistribution between Germany and the United States is due to benefits (especially for the unemployed) and not due to the progressivity of the income tax, which is rather similar in both countries. In our sample, we focus on individuals who are working. They pay taxes and receive almost no benefits – except for child credits, which are comparable in both countries. Hence, the redistributive effects for this subgroup of the population are rather similar both in Germany and the United States.

Conclusions

In this article, we have discussed the topic of earnings inequality and reviewed the existing literature. Inequality is not a bad thing per se. In fact, it is necessary for the functioning of a market economy. Nevertheless, the rapidly growing levels of inequality generate concern for the social stability of Western societies. When talking about inequality, the question of 'inequality of what among whom' arises immediately – and different options have different implications for the results. Hence, results from different studies have to be interpreted with caution. From a normative point of view, an interesting question is whether we should care about inequality in outcomes or rather about inequality of opportunity.

Our descriptive review of some core facts with respect to earnings inequality in the United States, has shown that there is not just one explanation for long-run trends in (increasing) inequality, but rather many. Core labor market conditions as well as changes in the composition of the population, and hence workforce, play an important role. In addition, the tax-benefit system, although not directly affecting gross earnings levels, affects incentives to take up work and also the formation of families and households.

Our discussion has shown that despite the enormous and ever growing literature on inequality, a lot of issues are still left for future research. It would be important to develop a comprehensive model of the household income distribution, incorporating different income sources such as labor earnings, capital, and business income; as well as models for taxes and social benefits that account for demographic factors, such as household formation and composition.

See also: Income Distribution and Inequality: Measurement Issues; Income Inequality; Inequality Analysis: Overview; Inequality Measurement; Inequality, Social; Inequality: Comparative Aspects.

Bibliography

Aaron, H., 1978. Politics and the Professors. Brookings Institutions, Washington, DC.

Acemoglu, D., Autor, D.H., 2011. Skills, tasks and technologies: implications for employment and earnings. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 4b. Elsevier, Amsterdam, pp. 1043–1171.

Acemoglu, D., Autor, D.H., 2012. What does human capital do? A review of Goldin and Katz's the race between education and technology. Journal of Economic Literature 50 (2), 426–463.

Almas, I., 2008. Equalizing income versus equalizing opportunity: a comparison of the United States and Germany. Research on Economic Inequality 16, 129–156.

Altonji, J., Blank, R., 1999. Race and gender in the labor market. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3C. North-Holland Elsevier, Amsterdam, pp. 3143–3251.

Armour, P., Burkhauser, R.V., Larrimore, J., 2013. Deconstructing income and income inequality measures: a cross-walk from market income to comprehensive income. American Economic Review 103 (3), 173–177.

Arulampalam, W., Booth, A.L., Bryan, M.L., 2007. Is there a glass ceiling over Europe? Exploring the gender pay gap across the wage distribution. Industrial and Labor Relations Review 60 (2), 163–186.

Atkinson, A.B., 1997. Bringing income distribution in from the cold. Economic Journal 107 (44), 297–321.

Atkinson, A.B., Bourguignon, F., 2001. Income distribution. In: International Encyclopedia of the Social & Behavioral Sciences. Oxford.

Atkinson, A.B., 2003. Income Inequality in OECD Countries: Data and Explanations. CESifo Working Paper No. 881.

Atkinson, A.B., Piketty, T., 2007. Top Incomes over the Twentieth Century. Oxford University Press, Oxford.

Atkinson, A.B., 2008. The Changing Distribution of Earnings in OECD Countries. Oxford University Press, Oxford.

Atkinson, A.B., Piketty, T., Saez, E., 2011. Top incomes in the long run of history. Journal of Economic Literature 49 (1), 3–71.

Autor, D.H., Katz, L.F., Krueger, A.B., 1998. Computing inequality: have computers changed the labor market? Quarterly Journal of Economics 113 (4), 1169–1213.

Autor, D.H., Katz, L.F., Kearney, M.S., 2008. Trends in U.S. wage inequality: revising the revisionists. Review of Economics and Statistics 90 (2), 300–323.

Bargain, O., Dolls, M., Immervoll, H., Neumann, D., Peichl, A., Pestel, N., Siegloch, S., forthcoming. Tax Policy and Income Inequality in the U.S., 1979–2007. Economic Inquiry.

Blau, F.D., Kahn, L.M., 2000. Gender differences in pay. Journal of Economic Perspectives 14 (4), 75–99.

Blau, F.D., Kahn, L.M., 2006. The US gender pay gap in the 1990s: slowing convergence. Industrial and Labor Relations Review 60 (1), 45–66.

Burtless, G., 1999. Effects of growing wage disparities and changing family composition on the U.S. income distribution. European Economic Review 43 (4–6), 853–865.

Burtless, G., 2009. Demographic transformation and economic inequality. In: Salverda, W., Nolan, B., Smeeding, T.M. (Eds.), The Oxford Handbook of Economic Inequality. Oxford University Press, Oxford, pp. 435–454.

Cancian, M., Reed, D., 1998. Assessing the effects of wives' earnings on family income inequality. Review of Economics and Statistics 80 (1), 73–79.

Cancian, M., Reed, D., 1999. The impact of wives' earnings on income inequality: issues and estimates. Demography 36 (2), 173–184.

Card, D., DiNardo, J.E., 2002. Skill-biased technological change and rising wage inequality: some problems and puzzles. Journal of Labor Economics 20 (4), 733–783.

Card, D., Lemieux, T., 2001. Can falling supply explain the rising return to college for younger men? a Cohort-based analysis. Quarterly Journal of Economics 116 (2), 705–746.

Cowell, F.A., 2008. Measuring Inequality. Oxford University Press, Oxford.

Daly, M.C., Valletta, R.G., 2006. Inequality and poverty in United States: the effects of rising dispersion of men's earnings and changing family behaviour. Economica 73 (289), 75–98.

Decancq, K., Fleurbaey, M., Schokkaert, E., 2014. Inequality, income, and well-being. In: Atkinson, A.B., Bourguignon, F. (Eds.), Handbook of Income Distribution, vol. II.

DiNardo, J.E., Fortin, N.M., Lemieux, T., 1996. Labor market institutions and the distribution of wages, 1973–1992: a semiparametric approach. Econometrica 64 (5), 1001–1044.

Fleurbaey, M., 1995. Three solutions for the compensation problem. Journal of Economic Theory 65 (2), 505–521.

Fong, C., 2001. Social preferences, self-interest, and the demand for redistribution. Public Economics 82 (2), 225–246.

Fortin, N.M., Lemieux, T., 1997. Institutional changes and rising wage inequality: is there a linkage? Journal of Economic Perspectives 11 (2), 75–96.

Goldin, C., Katz, L.F., 2008. The Race between Education and Technology. Harvard University Press, Cambridge.

Goldfarb, R.S., Leonard, T.C., 2005. Inequality of what among whom?: rival conceptions of distribution in the 20th century. Research in The History of Economic Thought and Methodology 23a, 75–118.

Gottschalk, P., Smeeding, T.M., 1997. Cross-national comparisons of earnings and income inequality. Journal of Economic Literature 35 (2), 633–687.

Hyslop, D.R., Mare, D.C., 2005. Understanding New Zealand's changing income distribution, 1983–1998: a semi-parametric analysis. Economica 72 (3), 469–495.

Jenkins, S.P., 1995. Accounting for inequality trends: decomposition analysis for the UK 1971–86. Economica 62 (245), 29–63.

Jenkins, S.P., Micklewright, J., 2007. Inequality and Poverty Re-examined. Oxford University Press, Oxford.

Katz, L.F., Autor, D.H., 1999. Changes in the wage structure and earnings inequality. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3a. Elsevier, Amsterdam.

Kunze, A., 2005. The evolution of the gender wage Gap. Labour Economics 12 (1), 73–97.

Lam, D., 2001. Income distribution: demographic aspects. In: International Encyclopedia of the Social & Behavioral Sciences. Oxford.

Lemieux, T., 2006. Increasing residual wage inequality: composition effects, noisy data, or rising demand for skill? American Economic Review 96 (3), 461–498.

Leigh, A., 2007. How closely do top income shares track other measures of inequality? Economic Journal 117 (524), F619–F633.

Martin, M.A., 2006. Family structure and income inequality in families with children, 1976 to 2000. Demography 43 (3), 421–445.

Niehues, J., Peichl, A., 2014. Upper Bounds of Inequality of Opportunity: Theory and Evidence for Germany and the US. Social Choice and Welfare 43 (1), 73–99.

OECD, May 2, 2011. Growing Income Inequality in OECD Countries: What Drives It and How Can Policy Tackle It? OECD Forum on Tackling Inequality, Paris.

OECD, 2008. Growing unequal? Income distribution and poverty in OECD countries. Organisation for Economic Co-Operation and Development, Paris.

Osberg, L., 2001. Inequality. In: International Encyclopedia of the Social & Behavioral Sciences. Oxford.

Peichl, A., Pestel, N., Schneider, H., 2012. Does size matter? The impact of changes in household structure on income distribution in Germany. Review of Income and Wealth 58 (1), 118–141.

Pestel, N., 2013. Inequality in Germany: The Role of Household Context and the Concept of Economic Resources. Dissertation. University of Cologne.

Piketty, T., Saez, E., 2003. Income inequality in the United States, 1913–1998. Quarterly Journal of Economics 118 (1), 1–39.

Piketty, T., Saez, E., 2007. How progressive is the US federal tax system? A historical and international perspective. Journal of Economic Perspectives 21 (1), 3–24.

Pistolesi, N., 2009. Inequality of opportunity in the land of opportunities, 1968–2001. The Journal of Economic Inequality 7 (4), 411–433.

Ramos, X., Van de gaer, D., 2012. Empirical Approaches to Inequality of Opportunity: Principles, Measures, and Evidence, IZA DP No. 6672.

Roemer, J.E., 1993. A pragmatic theory of responsibility for the egalitarian planer. Philosophy and Public Affairs 22 (2), 146–166.

Roemer, J.E., 1998. Equality of Opportunity. Harvard University Press, Cambridge.

Roemer, J.E., Trannoy, A., 2013. Equality of Opportunity, Cowles Foundation Discussion Paper No. 1921.

Salverda, W., Nolan, B., Smeeding, T. (Eds.), 2009. Handbook of Economic Inequality. Oxford University Press, Oxford, pp. 491–522.

Schwartz, C.R., 2010. Earnings inequality and the changing association between spouses' earnings. American Journal of Sociology 115 (5), 1524–1557.

Stiglitz, J.E., 2012. The Price of Inequality: How Today's Divided Society Endangers Our Future. W. W. Norton & Company, New York/London.

Stiglitz, J.E., Sen, A., Fitoussi, J.-P., 2009. Report by the Commission on the Measurement of Economic Performance and Social Progress. Paris.

Van de Gaer, D., 1993. Equality of Opportunity and Investment in Human Capital. Working Paper. KU Leuven.

Welch, F., 1999. In defense of inequality. American Economic Review Papers and Proceedings 89 (2), 1–17.

East Africa: Sociocultural Aspects

William Arens, Stony Brook University, Stony Brook, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The historical depth, cultural complexity, and social diversity of East Africa, an area encompassing the present nation-states of Tanzania, Kenya, and Uganda, are considered in terms of this region's prominent social forms. These arrangements include, at the most complex level, the precolonial Swahili city-states on the Indian Ocean littoral and the interior kingdoms in and around the great lakes of the region. Other less complex polities are considered by reference to a number of specific and typical ethnic examples. The latter commentary demonstrates the manner in which religion and ritual are employed to generate local social groups and political loyalties as well as the manner in which these arrangements were effected by both the colonial and independence periods. Reference is also drawn to the pervasive cultural themes that characterize the region.

Almost a century of scholarship confirms the historical and cultural complexity of East African societies encompassed by the independent nations of Kenya, Tanzania (including Zanzibar), and Uganda. From the maritime commercial civilizations of the Indian Ocean islands and coast that have traded with the external world for at least two millennia, to the more archaic forms of economic adaptation, including the still extant gathering and hunting societies in the interior, this region displays an exceptional range of social and cultural forms. In addition to Swahili as a regional Bantu language, unique to East Africa, branches of all the other African linguistic families, Nilotic, Cushitic, and Khoisan, are also represented in northern Tanzania.

Prehistory and History

Based on finds at Kobi Fora in Kenya and the Olduvai Gorge in Tanzania, it is now accepted that the earliest hominids emerged in East Africa some 4–6 ma. Although permanent agricultural settlements have existed for at least 10 000 years, historical records on the area extend back only to the first century of the Christian era (Phillipson, 1993). The original Greek manuscript, *Periplus of the Erythrean Sea*, describing the trade routes from Alexandria on the Red Sea in Egypt and then south along the East African coast, mentions Rhapta, a settlement assumed to have existed somewhere on the shoreline of what is now Tanzania. Arabic documents from the eighth century onward are more prolific and exact, indicating the existence of the still occupied trading sites of Lamu in northern Kenya and Kilwa in southern Tanzania. This land of 'Zinj' is described as inhabited by a population that was engaged primarily in a maritime trade in animal skins, horns, and some slaves, but particularly ivory, and that subsisted on agriculture and domesticated herds. Subsequent Arab influence, including Islamization, came to a halt for a time at least with the arrival of the Portuguese in the late fifteenth century and their partial conquest of the then more numerous city-states dotting the littoral. However, these first Europeans were themselves expelled from the area by a resurgent Afro-Arab coalition directed from Oman on the Arabian Peninsula, which in turn failed to establish suzerainty over the entire area.

Subsequent European incursions, primarily British and German, which included penetration of the interior in the latter half of the nineteenth century, hindered the by then more extensive slave trade supporting the offshore, particularly Zanzibari, plantation economies based upon the export of spices. This era culminated in 1885 with the Berlin Treaty, which partitioned all of Africa, excluding Ethiopia, among the European powers, and divided East Africa between Britain and Germany. The results of World War I brought the entire area under British control as a colony, protectorate, or League of Nations mandate until independence in the early 1960s (Oliver and Mathew, 1963).

Social Forms

The existing subsocial formations in recent and contemporary East Africa can be classified in the usual manner as either mercantile or agricultural states, less politically complex agricultural chiefdoms of varying scale, decentralized agricultural and pastoral societies, and the few remaining gathering-hunting groups. However, these categories are heuristic and ideal as most, if not all, of the economies were mixed and interdependent prior to the colonial era. In addition, there was fluidity in economic forms over time as the archaeological data indicate that some gathering-hunting peoples may have resorted to agriculture and then reverted to their previous mode of adaptation on numerous occasions. The same flexibility characterized pastoral peoples. Such adaptive shifts in more recent times have involved simultaneous alterations in ethnic identities that are often labels based primarily upon livelihood. Thus, for example, indigent Maasai who no longer kept cattle were reclassified as *il torrobo* (Maasai for 'the poor' or 'hunters') or Nderobo, while in another instance a whole segment of Maasai who took up agriculture as their primary occupation in the nineteenth century were redefined as Arusha. Thus, in addition to the complexity of social forms, the fluid basis of ethnic identities, often assumed to be fixed 'tribal' isolates, was a constant feature of the region's social history. Although there were and are cultural differences,

in some instances a matter of degree and in others of kind, rigid colonial mentalities and policies often helped to bring into existence these so-called tribes.

City-States and Kingdoms

Prior to the advent of European incursions, complex economic and political entities existed at both extremes of the East African region. The coast, as noted, was intermittently dotted with numerous city-states, while the westernmost interior, some 800 miles away, featured a series of large-scale polities referred to as the interlacustrine Bantu kingdoms due to their clustering around the lakes, such as Victoria and Tanganyika, in the Great Rift Valley. Most notably, these included Baganda, Banyoro, Basoga, and Batoro in contemporary Uganda, as well as Rwanda and Burundi, which eventually fell under Belgian colonial control and reemerged as separate independent nations. These two regional extremes were linked by long-distance commerce, including, by the early nineteenth century, some commerce managed by interior middlemen groups, allowing all concerned to expand and flourish with the exchange of valued trade goods. The Bantu kingdoms were early on the subject of academic scholarship (Beattie, 1971.) However, in part due perhaps to its less romantic and more complex 'untribal' or 'detribalized' condition, the coastal region failed to receive serious scholarly attention until after the colonial era.

Although these ancient city-states, extending 1000 miles along the coast north to Somalia and south to Mozambique, never formed a single political unit, they were linked by a common civilization. In addition to maritime mercantilism, stone cities, marriage alliances, and patrilineal descent, other commonalities included by the eleventh century Islam and a related lifestyle, as well as Swahili as a *lingua franca* (Middleton, 1992). The term 'Swahili' eventually emerged as a complicated and widespread ethnic category referring to peoples on and beyond the coast as the language and Islam spread throughout East Africa before and after the colonial period. The word 'Swahili' itself is derived from the Arabic *sahel*, meaning 'coast.' The language has a modicum of loan words from this source, but otherwise it is Bantu in all of its essential grammatical features. As it was adopted as a first or second language by others in the region, many also identified themselves as an mswahili or as waswahili (a Swahili person or people, respectively), which implied not only language usage but also a form of modernization involving the adoption of Islam or even Christianity, migration to other areas of the country for one's livelihood, and a modified Arab or European lifestyle, including forms of dress and architecture. Among themselves in the towns and cities they inhabit, the Waswahili may also admit to another ethnic affiliation, reaffirming the plasticity of this sort of identification (Arens, 1976).

The agricultural bases of the interior states situated in fertile highland areas with the banana (originally from Indonesia) as a staple may have been the historical outcome of an incursion of northern cattle-keeping peoples into an area inhabited for at least 1000 years by Bantu-speaking agriculturalists. The high population densities, in the case of Baganda close to 2 million people, and social complexity resulted in what has been described as semifeudal states involving unequal control over resources, including land and cattle; concomitant social stratification (though no caste system); and, in most instances, a series of patron–client relationships spanning economic and political divides. These kingdoms, which shared a host of cultural similarities in addition to economic and social ones, competed with each other for more extensive control over the region. At the apex of the highly centralized political systems stood a king advised by a prime minister and other councilors, who appointed chiefs over large districts who in turn delegated authority to headmen in villages. Although often popularly viewed as a despot, it is unlikely that the king could have achieved this level of power and authority until the introduction of firearms in the early nineteenth century via trade from the coast. The stability of the entire political-economic system was further solidified with colonialism and its application of indirect rule that permitted little if any challenges to established authority or radical changes to the system. Although these were not the only centralized political systems in East Africa, as others of varying degrees in other highland areas existed beyond the coast, including the Shambaa and Chagga of Tanzania (Feierman, 1990), the vast majority of the population in the less watered lowland savannas and woodlands of the region, shared with pastoral peoples, were organized into smaller political units.

Agriculturalists and Pastoralists

Among the agriculturalists, the traditional crops were millet, sorghum, and maize. In the more abundantly watered areas, rice and bananas were also cultivated. Reference here should also be made to the now abandoned stone city of Engaruka in the northern part of the Great Rift Valley in Tanzania. This stone settlement based on irrigation agriculture and with at least 5000 people was abandoned some 300 years ago for no apparent reason, though the event does correspond to the appearance of the Maasai in the area at that time (Robertshaw, 1986).

Cattle, which play such a significant cultural role in the entire area, including among agriculturalists, and smaller livestock such as sheep and goats were also maintained whenever and wherever possible. Politically, this is the 'tribes without rulers' region as termed by one of the classic anthropological volumes of the colonial period and the 'golden age' of East African ethnography (Middleton and Tait, 1958). This vision of much of the social landscape was influenced in no small part by the prior research of E.E. Evans-Pritchard among the acephalous pastoral Nuer of the neighboring South Sudan (Evans-Pritchard, 1940). In light of the range and variability of societies in East Africa proper, they are best represented by a limited number of examples that also reflect the interests of the ethnographers at different points in the history of East African scholarship.

The quarter of a million Sudanic-speaking agricultural Lugbara, who live along the Nile–Congo divide on both sides of the Uganda–Congo border, are characteristic in that they had no political offices until chiefs and subchiefs were appointed by the colonial authorities. This lack of a formal political system was somewhat unusual in light of their relatively high population density, which exceeded 100 people per square

kilometer in some areas in the early twentieth century, but is explicable in light of the absence of nucleated settlements. The basic social unit was the homestead of various sizes composed of patrilineally related kin. This family cluster recognized the authority of a senior male elder with ritual significance due to his control over shrines dedicated to the lineage ancestors. Thus, at its core this social group had both a descent and a related religious ideology.

The Lugbara themselves conceive of their social structure in terms of clans originally historically associated with specific territories despite the dispersal of their members over time. As such, principles of descent provided an ideological rather than actual form of social organization, as is the case for many other East African peoples. These recognized territories may have a 'rainmaker' with ritual authority in certain matters but, as indicated, no formal leader. Order was maintained by compromise and conciliation among the resident elders. Disputes among members of different territorial groups were difficult to mediate and as such could result in violence and a state of feud. However, since clan kin were dispersed among the various Lugbara, tribal units and marriages were also effected across these lines, resulting in affinal relationships; thus violence was not entered into easily, especially among contiguous territories. More distant Lugbara tribal sections presented less opportunity for conflict, and if a state of feud did exist, it was difficult to act upon as the result of spatial separation. Thus, there was a political order, despite anarchy in the technical sense of the term, among these people, who recognized a common moral code supported by the ritual authority of elders and ancestors (Middleton and Tait, 1958).

The manner in which social order or stability was maintained among a people lacking a well-defined political system was one of the driving research concerns during the colonial era, while change became a principal interest of postcolonial scholarship. This shift in orientation is exemplified by the situation and description of the Iteso, Nilotic-speaking sedentary cultivators residing in this instance on both sides of the Kenya–Uganda border. By the time of their analysis, the Iteso were almost all Christians, and their expansionist tendencies, as well as the former political system that had permitted this process, had been radically altered by the often violent imposition of colonialism. According to historical reconstruction, prior to European intervention the Iteso were organized into a number of territorial sections under the leadership of a single individual or 'war leader.' These units were broken into a number of subsections that tended to correspond with the residence of patrilineally related males who also recognized a local leader for political and ritual purposes. These locations undertook common defense, the peaceful settlement of disputes, and collective religious activity. Typical of Nilotic speakers, the territories were cross-cut by an age grade system including a warrior class that had facilitated Iteso incursions into the lands of other ethnic groups. During the colonial period, appointed chiefs replaced the former leaders and the territorial units dissolved, as neither expansion nor defense was a relevant concern. These aggregations were replaced by diffuse neighborhoods of a smaller social scale composed of cooperating but individual households now engaged in the cash economy. The resulting process generated a typical East African society brought into being by the colonial presence. Similar ethnic consolidation and consciousness among the more numerous Kikuyu and Luo peoples of Kenya emergent during the colonial era had significant impacts on postcolonial politics (Karp, 1978).

The primarily pastoral Nilotic-speaking peoples inhabiting a grassland expanse in both arms of the vast Rift Valley in eastern Uganda, western Kenya, and north-central Tanzania provide a striking alternative to agricultural societies in more than economic fashion as many of the representatives of this language family have preserved their traditional cultural system to this day. Included in this category are the Jie, Karimojong, and Turkana of Uganda; the Kipsigis, Nandi, and Samburu of Kenya; and the Barabaig of Tanzania. However, the quarter of a million Maa speakers or Maasai who have occupied the central plains of both Kenya and Tanzania for the past 400 years are best known for the persistence of their traditional ways despite impositions placed on their lifestyle by both the colonial and independence-era regimes (Spear and Waller, 1993).

The enormous expanse of Maasailand is divided into numerous large and politically autonomous named territorial sections that, in some instances, cover thousands of square kilometers and include tens of thousands of individuals. Within these sections, social order is maintained at the local level by male elders with either ritual or judicial authority derived from membership in a particular clan, age grade, or local community of kin. Although patrilineal in terms of descent principles, kinship ties of various types are a significant factor only at a settlement composed of a number of usually polygynous households that herd their animals together under the direction of a senior male. The most striking feature of the social system is the age grade system bisecting the entire expanse of Maasailand, which includes the Moran, who were formerly a warrior class of young men but are now primarily responsible for cattle herding (Hodgson, 2011).

For the Maasai, cattle are not only the basic foodstuff in the form of milk, meat, and blood but also the most important cultural idiom. The animals are valued in and of themselves as esthetic items, the object of admiration, discussion, song, and dance, but they also mark both existing social relations through their exchange, and the passage of time through inheritance. They are circulated as bride wealth and blood wealth (the latter in compensation for violence), and they are loaned out to establish herding partnerships over wide areas. Finally, these animals are the primary means for communicating with the supernatural world as sacrificial offerings to the creator in religious ceremonies (Århem, 1985). Thus, this single item links humans both to each other and to their deity in a manner that is unavailable to cultivators. In more practical economic terms, the ability of the Maasai and other pastoralists to sell some animals for cash to meet the demands of the state in the form of taxation has permitted them to avoid social integration and in doing so conserve existing cultural patterns.

Foragers

Scattered among the pastoralists and agriculturalists in both forests and plains are pockets of hunters and gatherers, such as the Mbuti Pygmies of western Uganda and the Congo. Other foraging groups, though physically distinct from the Mbuti,

include the Okiek of Kenya, the Ik of Uganda, and the Sandawe and Hadza of Tanzania. The forest-dwelling Mbuti and related populations generally are considered to be descendants of the first residents of this and other parts of Africa who have adopted the languages of their later arriving and now Bantu neighbors, while the Okiek and Ik speak the related Nilotic languages of their pastoral neighbors. However, Sandawe and Hadza bear some phonetic similarities to the tongues of hunters and gatherers in southern Africa. Despite the traditional academic classification on the basis of foraging, there are also variations among these peoples, including the tendency to mix this adaptive strategy with agriculture and herding when feasible. Thus, a portion of the few hundred Hadza residing in a desolate region of northern Tanzania that was unsuitable until recently for either agriculture or herding are interesting in that they maintained a foraging lifestyle in pure form, as well as a significant degree of economic independence, from surrounding groups and the state.

Although seemingly inhospitable to others, the Hadza territory of over 1000 square miles, open to any member of the group as there are no recognized territories, is replete with natural resources harvested on a nomadic and individualistic basis. Although the area is heavily populated with game, the Hadza rely mainly on wild vegetable matter for food in the form of roots, berries, and fruit gathered daily by both males and females, the latter with their small children. These nutrients usually are consumed immediately rather than being collected for later distribution. Traditional bow-and-arrow hunting, however, is exclusively a male activity. If the kill is small, it is also eaten on the spot, while some of the remains of a larger animal may be either transported back to camp or simply abandoned. Thus, there is almost complete economic independence even among the Hadza themselves. As such, it would be misleading to characterize their local aggregations as bands in the sense of groups recognizing a common territory or cooperative bonds. Beyond the nuclear family, based on relatively unstable monogamous unions, the larger and only unit is a camp, occupied for no more than a few weeks, and on average by a dozen or more adults. This aggregation does not have a recognized leader, nor does it form a stable unit, as some members join or depart on a constant basis. At other times, it may split off into independent units or merge with other camps for a period of time. This organizational fluidity of foraging societies is perhaps their characteristic feature in both East Africa and other parts of the world (Ingold et al., 1988).

Cultural Concerns

This overview of East Africa in terms of economic, political, and social forms in no way suggests a paucity of equally, if not more, complex cultural systems including beliefs, values, myth, history, and literature both oral and written. These integrated systems of thought, often shared over a wide area, even in terms of particulars, defy simple categorization into a number of ideal types. In general, the peoples of this region, regardless of their form of economic adaptation, must be considered highly spiritual, whether in terms of their adherence to the local religions or the world religions that have supplanted them. In turn, these cosmologies have been transformed by local beliefs and practices. As in other parts of the world, the traditional ideological systems took a great interest in notions of good and evil and thus acceptable standards of interpersonal morality guarded by the living as elders and/or the deceased as concerned ancestors. The 'problem of evil' in society was often represented (by agriculturalists in particular) as a matter of witchcraft or sorcery that was resolved by recourse to diviners who could identify the culprit in order to restore a harmonious social order through confession and reconciliation.

As indigenous philosophies, these systems of thought also reflected on the meaning and passage of life and the significance of death, occasions usually marked by elaborate ritual gatherings and activity. Communication with their creator, most often viewed as a more distant and less knowable figure than in world religions, traditionally was maintained through prayer and sacrifice. Indeed, as this review of primarily social forms suggests, it was and is often best to view many East African peoples at the local level and in the most meaningful sense as composed of ritual communities concerned with the spiritual and social welfare of their members (Karp and Bird, 1980).

Globalization and Modernity

The effects of globalization and resulting forms of indigenous modernities have been a mixed blessing for the region. Tanzanian has undergone significant reorganization with economic liberalization but remains politically stable and peaceful with the emergence of a multiparty system relatively free of ethnic, regional, and/or religious pressures. Kenya, however, has at times succumbed to these divisive forces with a degree of internal political violence with ethnic overtones. It also confronts global conflicts as violent local disturbances on its borders with Somalia and the recently formed South Sudan. Uganda has yet to resolve some breakaway rebellions and its relations to neighboring states.

The entire region, but particularly the Swahili coast of Tanzania and Kenya, has had to adjust to the impact of international leisure tourism, importing with it European cultural patterns often at odds with a more conservative Islamic ethic. The results have been local conflicts between the generations and in some instances confrontations between the local community and the state (Caplan and Topan, 2004).

See also: Africa, Sociocultural Overviews: West Africa; African Studies: Culture; African Studies: History; Central Africa: Sociocultural Aspects, Colonization and Colonialism, History of; Ethnicity: Anthropological Aspects; Historical Thought and Historiography: Sub-Saharan Africa; Hunting and Gathering Societies: Anthropology; Pastoralism in Anthropology; Southern Africa: Sociocultural Aspects.

Bibliography

Århem, K., 1985. Pastoral Man in the Garden of Eden. Uppsala Research Reports in Cultural Anthropology, Uppsala, Sweden.

Arens, W. (Ed.), 1976. A Century of Change in Eastern Africa. Mouton, The Hague, The Netherlands.

Beattie, J., 1971. The Nyoro State. Clarendon Press, Oxford, UK.
Caplan, Pat, Topan, Farouk, 2004. Swahili Modernities. Africa World Press, Trenton, NJ.
Evans-Pritchard, E.E., 1940. The Nuer. Oxford University Press, London.
Feierman, S., 1990. Peasant Intellectuals. University of Wisconsin Press, Madison, WI.
Hodgson, Dorothy L, 2011. Being Masaai, Being Indigenous. Indiana University Press, Bloomington, IN.
Ingold, T., Riches, D., Woodburn, J., 1988. Hunters and Gatherers. Berg, Oxford, UK.
Karp, I., 1978. Fields of Change among the Iteso of Kenya. Routledge & Kegan Paul, London.
Karp, I., Bird, C. (Eds.), 1980. Explorations in African Systems of Thought. Indiana University Press, Bloomington, IN.
Middleton, J., 1992. The World of the Waswahili. Yale University Press, New Haven, CT.
Middleton, J., Tait, D. (Eds.), 1958. Tribes without Rulers. Routledge & Kegan Paul, London.
Oliver, R., Mathew, G. (Eds.), 1963. History of East Africa, vol. 1. Clarendon Press, Oxford, UK.
Phillipson, D., 1993. African Archaeology, second ed. Cambridge University Press, Cambridge, UK.
Robertshaw, Peter, 1986. Engaruka revisited: excavations of 1982. Azania 21, 3–20.
Spear, T., Waller, R., 1993. Being Maasai. Currey, London.

East Asian Legal Traditions

Pär Cassel, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Most East Asian legal traditions were heavily influenced by Chinese thought and statecraft, which can be traced back at least 3000 years. Confucianism, which made familial duties a central feature of all social obligations, had a formative impact on both the Chinese and the Korean legal orders. Japan was also heavily influenced by the Chinese legal system, but was more selective in its adoption of those models, because of the fragmented nature of the Japanese political order. Beginning in the late nineteenth century, East Asia came under heavy Western influence and all legal systems were ultimately reformed on Western models, leaving few formal traces of the previous legal order.

With a recorded history of more than 3000 years, the legal orders in China proper have exercised a major influence not only on its neighbors, especially Korea, Vietnam, Ryūkyū, and Japan, but also on a number of 'conquest dynasties' that ruled parts of traditional Chinese territories or controlled areas contiguous to China proper. Practically everything we know about the early states that existed in proximity to the Chinese dynasties has been handed down to us through the medium of classical Chinese, and East Asian rulers took the ideal of the Han (206 BC–AD 220) and Tang (618–907) dynasties at the height of their power as the purest expression of an idealized Confucian order. Therefore, any inquiry into East Asian legal traditions has to begin with a discussion of the legal orders that developed in China proper from the first millennium BCE. At the same time, we have to be cognizant of the fact that the uniformity of the written Chinese language, which was the preferred idiom of Chinese dynasties and many non-Chinese states, often masks the realities of societies that were regionally, ethnically, and socially diverse, and subject to significant change over time. Consequently, no claims of a continuous and unbroken East Asian legal tradition harking back to Chinese antiquity should be taken at face value.

The Great Tradition

Very little is known about the legal order in preimperial China, but fragments of laws from the western Zhou dynasty (1050–771 BCE) have survived in bronze inscriptions and historical chronicles. One of the earliest historical records on the early Chinese legal order is found in the chapter 'The Punishments of Lü' in the canonical historical work *Book of Documents* (*Shàngshū*), which outlined some early laws and the classical form of the 'Five Punishments' (*wǔ xing*). It is unclear whether the original five punishments were ever practiced, but in their classical iteration, they included 'tattooing of the face' (*mò*), 'cutting off the nose' (*yì*), 'cutting off the feet' (*yuè*), 'castration' (*gōng*), and 'decapitation' (*dàpì*). It is not clear if any single code existed during the Zhou dynasty, but there is ample evidence that laws were recorded on bronzes during the Chinese antiquity. At a very early stage, Chinese thinkers seem to have regarded legal action, litigation, and punishment as signs of failure on the part of the sovereign's ability to exercise a moral influence on its subjects to abstain from criminal and other asocial acts. Thus, it is not surprising that the introduction of penal law is attributed to the non-Chinese and according to the *Book of Documents*, the five punishments were handed down from non-Chinese peoples, such as the Miao People in what is today southwest China (Creel, 1970).

The eastern Zhou dynasty (770–256 BC) witnessed the weakening of central royal authority and the relative strengthening of the Zhou vassals, who emerged on ambitious projects of state building and availed themselves of the advice from a new class of literati. During the Spring and Autumn (770–476 BC) and Warring States (475–403 BCE) periods, a number of contending schools of statecraft emerged, many of which claimed to have answers on how to best use law and punishment to create strong and prosperous states. The two most successful schools were the Confucian and Legalist schools of thought. Very much in line with earlier trends in Chinese thought, the Confucian school argued that the family was a microcosm of the state and that a just ruler should govern by providing good example to his subjects by maintaining the Confucian family system and rituals, which were supposedly perfected by the sage kings of the Xia, Shang, and Zhou dynasties. Confucians saw laws and punishments only as a last resort and warned that excessive use of punishments would only make people into fawners who did not internalize the positive norms of the Confucian order:

> Guide them by edicts, keep them in line with punishments, and the common people will stay out of trouble but will have no sense of shame. Guide them by virtue, keep them in line with the rites, and they will, besides having a sense shame, reform themselves.
> Confucius, 1979: 63

Confucians also assumed that common people would harbor a higher sense of loyalty toward their family than to the state, and Confucius explicitly approved of family members covering up each others' crimes (Confucius, 1979: 121).

The Legalist school did not disfavor familial duties per se, but it argued that a strong and prosperous kingdom could only be maintained if the ruler applied strict and predictable rewards and punishments, which would produce politically obedient and economically productive subjects. In contrast to the Confucians, who idealized the early feudal Zhou state and denounced vassals who proclaimed themselves kings as

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86139-5

usurpers, the Legalists wanted to establish a centralized and militarized state that imposed heavy obligations on its subjects. They saw law as an extension of administrative power and regarded rewards (*shǎng*) and strict punishments (*fá*) as central tools of a 'behavioral science' that would produce law-abiding and productive subjects who would render loyal service to the state (Schwartz, 1985). Indeed, the Chinese term for legalist (*fǎjiā*) can be translated as both the 'school of law' and the 'school of methods.' The legalists also advocated that all laws should be equally applied to all subjects of the state, regardless of social status, with the notable exception of the sovereign himself. The King of Qin, Ying Zheng (260–10 BC), was one of the most prominent patrons of the legalist schools and when he established himself as hegemon of China proper and the First Emperor of Qin in 220 BC, the Legalist school scored an early victory against the Confucians and participated in the building of a strong state where centrally controlled counties and prefectures replaced the old territorial states of the Zhou vassals.

Very few fragments of Qin Law have survived to this day and the Qin dynasty did not survive the death of the first emperor by many years; when Liu Bang (256/247–195 BC) established the Han dynasty in 206 BC, he not only preserved the centralized state structure of the previous dynasty and many of its legal institutions, but also set out to restore the status of Confucianism as the central doctrine of the state. One product of this endeavor was what the legal historian Chü T'ung-tsu dubbed the 'Confucianization of law,' that is, the ambition to enforce Confucian norms and family hierarchies through penal means, an idea that probably would not have appealed to the founders of the Confucian school, yet this set the pattern for most subsequent imperial dynasties (Ch'ü, 1965; Goldin, 2012). The Han Code itself has only survived in fragments and scholars have had to resort to careful reconstruction in order to provide a full picture (Hulsewé, 1955). After the fall of the Han dynasty in 220 AD, China was divided into a number of competing dynasties that contended for hegemony over the central plains and only periodically unified under a centralized state. In order to bolster their claim to legitimacy, the contending states maintained many of the institutions that harkened back to the Han dynasties, but this often covered up political realities that failed to live up to the meritocratic and Confucian ethos of the Han Dynasty (Hucker, 1985: 17).

China and East Asian Legal Orders during the Middle Period

The short-lived Sui dynasty (581–618) unified China and promulgated a penal code shortly after its ascension to supremacy, but the original text of the code disappeared when the dynasty crumbled and is no longer extant. However, most of the Code survived, as the subsequent Tang dynasty (618–907) based its 653 penal code on the Sui Code. The Tang Code, which consisted of 501 articles and was divided into 12 thematic sections, is the earliest legal code that has survived in its entirety. Practically all laws in the Tang Code were of a penal character and there was no separate code regulating private relations. The law was hierarchic in nature and the much-vaunted 'Confucianization of law' was manifested in the fact that criminals, when found guilty, were sentenced in accordance with their status in the Confucian family system, defined by the 'Five mourning degrees' (*wǔ fú*). Another expression of both Confucian hierarchies and the aristocratic nature of Tang society was the 'Eight deliberations' (*bā yì*), which provided for various forms of special treatment for status groups, such as the imperial house and virtuous officials (Johnson, 1979).

In the Tang Code, the five punishments were redefined as 'caning with light stick' (*chī*), 'caning with heavy stick' (*zhàng*), 'penal servitude' (*tú*), 'life exile' (*liú*), 'military exile' (*chōngjūn*), and 'capital punishment' (*sǐ*). Capital punishment was divided into two categories, strangulation (*jiǎo*) and decapitation (*zhǎn*), where the former was regarded as the more lenient form of punishment as the delinquent would be spared the humiliation of dismemberment, which violated Confucian concepts of filial piety. This schematic division of punishments, which would remain in force for most of China's imperial history, was subject to different modifications and commutations. For instance, some physical punishments were subject to certain kinds of reductions, and certain groups, such as minors and senior citizens, were able to substitute corporal punishment with monetary redemption. Furthermore, criminals who voluntarily submitted to the authorities (*zì shǒu*) were punished more leniently as long as their crimes did not belong to the 'ten abominations' (*shí'è*), a series of crimes of a certain heinous nature, such as sedition, rebellion, or parricide. The idea that one should receive a more lenient punishment as a result of voluntary confession is still prevalent in many East Asian countries (Ricket, 1971).

A core feature of Chinese legal procedure was the centrality of confessions and other types of written evidence, as opposed to other forms of technical evidence. Although the Chinese legal order had developed a relatively sophisticated forensic science, which purported to be able to identify different causes of death (Song and McKnight, 1981), the central piece of any criminal case was a confession of guilt on the part of the accused, who in principle could not be convicted without his own admission of culpability. Consequently, the Chinese legal system permitted a certain use of torture and other means of coercion in order to obtain correct confessions. At the same time, the risk of extracting false confessions under duress was recognized and magistrates who abused torture and punished innocent people were subject to disciplinary and criminal sanctions.

Intimately tied to the importance of written evidence was an 'appellate system' of justice, which took shape during the Tang dynasty, according to which severe criminal sentences had to be reviewed and sanctioned by higher instances in the bureaucracy, sometimes by the emperor himself, in order to have legal force (Bodde and Morris, 1967). Another characteristic of the Sui-Tang bureaucracy was the introduction of a rigorous examination system, which selected aspiring officials on the basis of their proficiency in the Confucian classics, and the rule of avoidance, which laid down that local officials could not serve in their native province. Both of these measures were designed to provide an imperial check on the great aristocratic families, which were centered on the dynastic capital of Chang'an (present Xi'an), and were in all likelihood

imperfectly implemented, but they did provide the basis of an idealized meritocratic ideology, which would gain in strength and influence in subsequent dynasties. The Sui-Tang legal order had a formative impact on the legal orders of many of China's neighbors. For instance, the Taika reforms of 645 and the introduction of the Ritsuryō system in 701 were two programs that set out to centralize imperial power according to Chinese patterns of law and governance; however, the weakening of the central authority in Japan led to the emergence of a feudal order by the twelfth century, under which the imperial court in Kyoto was largely incapable of asserting its political and legal influence over local territorial warlords, who had taken power in the provinces. This state of affairs paved the way for a legal order with a large degree of local diversity.

The Tang dynasty reached its apogee in the mid-ninth century and finally collapsed in 907, but its legal system and institutions enjoyed a high stature, not only in China proper, but also in East Asia, and both Japan and Korea adopted the Tang Code as a model, although both polities retained many native institutions as well. In China itself, the Tang Code became the basis for most subsequent penal codes, with the exception of the unapologetically alien Yuan Dynasty (1271–1368). The Song dynasty (960–1279) inherited many of the codes and legal institutions of the Tang state, but the coherence of the legal system was encumbered by a proliferation of different types of laws, such as the Code itself, edicts, law compilations, and precedents. From the Song dynasty, we also have the first extant case collections, which have provided scholars with valuable insights on how the legal order operated in practice (McKnight and Liu, 1999). The transition from the Tang to the Song dynasty is notable as it marks the final demise of the great aristocratic families, which had dominated politics and court life during the period of division, and the emergence of local elites, which competed for government service in the examination system.

The short-lived Yuan dynasty represents one the sharpest ruptures in Chinese legal history and the relative paucity of extant primary sources has made it difficult to study the legal system of the dynasty in detail. While the Mongol rulers did rely on some Chinese institutions and issued administrative laws on a Chinese pattern, they were not beholden to the Confucian tradition and chose not to adopt a single written Chinese code to administer their empire, but relied on different sources of law such as imperial edicts and Mongol customary law. The Mongols practiced legal pluralism rather aggressively and divided the population in their Chinese empire into four different categories, namely, Mongols, Central Asians (*sèmù rén*), Northern Chinese, and Southern Chinese, in descending order of status. Each group was subject to different laws and jurisdictions, and conflicts between different groups were resolved through joint conferences between officials holding jurisdiction (Ch'en, 1979).

The Late Imperial Age

The establishment of the Ming Dynasty in 1368 represented a restoration of Han Chinese rule over the territories of China proper and the founder of the dynasty, Zhu Yuanzhang, created legal and political institutions modeled on the Tang Dynasty to justify his regime as a restoration of a Confucian form of government. The central piece of this legal order was the Ming Code, which was a reorganized version of the Tang Code and reached its final form in 1397. Just like its precursors, The Ming Code (*Dà Míng lǜ*) was a criminal code and the statutes (*lǜ*) were divided into six major sections that corresponded to the Six Boards (*Liù bù*) in the central government: personnel, family and revenue, rituals, military affairs, and criminal affairs. In keeping with the tendency to enshrine Confucian ethics through penal sanction, Ming legislators took filial piety (*xiào*) and other Confucian family very seriously and they made a point in abolishing some Mongol practices that were repugnant to Orthodox Confucians, such as levirate (compulsory marriage of a widow to a brother of her deceased husband), which violated the Confucian family system. Violent crimes were seen as more severe the closer the criminal was related to the victim and the more junior he or she was compared to the victim. The opposite principle was at work when it came to property crimes (Jiang, 2005: lxxii-lxxiii). Yet at the same time, the Ming penal order preserved certain elements of earlier legal orders, such as the notorious punishment of 'slow slicing' or 'death by a thousand cuts' (*língchí*), which was an unorthodox punishment of unclear origins used for crimes against the state and grave transgressions against the family (Brook et al., 2008).

Under Ming, the Chinese legal order reached new heights of centralization and during the dynasty the principle that all capital cases should have imperial sanction was established. In contrast to earlier dynasties such as the Song, the Ming Dynasty used law as an educational tool; trials were in principle open to the public and the Ming Code and other legal manuals were available for purchase to those who could afford them. Crime fiction also became a distinct genre during the Ming Dynasty, which also disseminated knowledge about the legal system among the reading public (Farmer, 1995)

The Ming Code exercised considerable influence on the legal systems of China's neighboring countries, most notably Korea. The Chosŏn dynasty (1392–1910) used the Ming Code as its basic criminal code until the late nineteenth century and built its own legal order on the basis of the Ming system of government. The Chosŏn kings also introduced important modifications to Ming laws and institutions so as to accommodate the rather stratified Korean social order, where slavery and other forms of bondage were more prevalent than in China and where the hereditary *yangban* class exercised a much stronger influence than the Chinese scholar-gentry ever was capable of attaining (Hahm, 1967).

In 1644, the Ming dynasty fell to domestic rebellion and the Tungus Manchu people conquered China proper and established their Qing dynasty (1636–1912) in Beijing, proceeding to build a vast polyethnic empire that reached its apogee in the late eighteenth century. Because of the regional, social, and ethnic diversity of the Qing Empire, the Manchu emperors maintained a plural legal order, where different ethnic groups often were subject to different laws and jurisdictions. In the provinces that had been conquered from the Ming dynasty, the Manchus preserved and developed the Ming legal and political system; the legal system in Manchuria was modeled on the Ming system of law with important modifications, but was administered separately from China

proper. In Inner and Outer Mongolia, the Qing court applied a separate legal code for Mongols, which was a fusion of Mongol and Chinese legal norms, and in Xinjiang a mixture of Chinese law and Moslem law prevailed, administered jointly by Manchu officials and local headmen. The vast Tibetan territories of the Qing Empire, which constituted roughly a quarter of its area, were mostly subject to a theocratic legal system strongly infused with Buddhist thought and practices, which was only partially under Chinese influences. Finally, the ruling caste of the Manchus and other allied groups were technically subject to the Qing Code, but were enrolled in the Eight Banners (*Bā qí*), a social and military organization that constituted largely its own legal sphere, centered on Beijing and the banner garrison that had been stationed at strategic locations in the empire. Consequently, the Qing Empire was in many ways a continuation of previous 'conquest dynasties' that had ruled a part or the whole of China, such as the Liao (907–1125), Jin (1115–1234), and Yuan Empires. The full complexity of these plural legal orders has yet to be comprehensively explored by historians.

The Qing Code (*Dà Qīng lǜlì*), which was first adopted in 1646 and reached its mature form in 1740, is often incorrectly presented as a mere extension of the Ming Code. While it is true that the first version of the Qing Code was largely identical to its Ming counterpart, the Qing government supplemented and modified the Ming criminal statutes with numerous substatutes (*lì*), which took precedence over statutes whenever there was a conflict between the two forms of law. Many of these substatutes emerged from imperial edicts and rulings in case law, which produced the need to modify the code. These constant revisions of the law spawned a huge commentarial literature as well as legal handbooks, which were produced to assist local magistrates in reaching correct verdicts and were available to anyone who could read. During the Qing dynasty, as in previous dynasties, there was no division of power within the legal process and no separate legal profession emerged as a result of the expansion of the legal system. Quite the contrary, it was illegal for commoners to advise strangers in legal matters, as this conflicted with the Confucian ideal of the legal system being the last resort to resolve conflicts.

The Qing dynasty is the only imperial Chinese regime that has left behind a large corpus of archival material, and following the opening of archives in mainland China and in Taiwan in the 1980s and 1990s, historians have been able to study the operation of the legal system in detail, from the magistrates' chancelleries in the roughly 1300 counties in China proper, to the ministries in the central government. We now know that ordinary Chinese were far more likely to resort to legal action than was previously thought and scholars have also found that magistrates were often called upon to adjudicate disputes of a private and commercial nature (Huang, 1996), but it remains a moot point whether the Qing legal order produced any functional counterpart to the system of private law in early modern Europe (Bourgon, 2002). Suffice it to say that the concept of 'law' in Qing China remained largely a penal concept and the dynasty did not promulgate any legal code to regulate private and commercial disputes until the early twentieth century.

Unlike the Korean legal order, which had been under heavy Chinese influence for millennia and had copied Chinese legal codes almost verbatim, Japan was never fully part of the Chinese tributary order and Japanese lawmakers were much freer to choose what aspects of Chinese legal culture they wanted to adopt as their own. Furthermore, Confucianism did not exercise the same ideological hegemony in Japan as it did in China and Korea, but had to contend with other schools of thought such as Buddhism and Shintō, both of which exercised a strong counterweight to any efforts to Confucianize Japan. Following the collapse of central imperial authority in Japan in the twelfth century and the emergence of a feudal political system centered on the warrior aristocracy, the legal system in Japan became even more regionally diverse and fragmented than on the Asian mainland.

When the Tokugawa Shogunate established its hegemony over the Japanese archipelago in the early seventeenth century, the new rulers created a legal and political order where local territorial lords were integrated into what has been called a system of 'centralized feudalism' and the emperor in Kyoto retained his sacral and ritual authority under strict shogunal supervision. A central feature of the Tokugawa legal system was the regulation of the four major status groups – samurai, peasants, artisans, and merchants – which were subject to different laws and jurisdictions are were not encouraged to interact on a daily level so as to prevent cross-jurisdictional crimes from occurring. The status system was further complicated by the fact that every domain maintained its own version of the status system and legal cases involving more than one status group and more than one jurisdiction could lead to intricate legal cases, where shogunal officials from the shogunate were called upon to formulate a verdict.

The dominant Tokugawa house, which directly controlled more than 40% of the four main islands of Japan, exercised a strong regulatory and normative influence on the domanial lords, but it refrained from interfering in their internal autonomy or legal affairs. The Tokugawa legal system relied on a number of different sources of laws, such as the ancient Ritsuryō system, shogunal edicts, case precedents, and legal reasoning (Henderson, 1965). The shogunate did not elaborate a single legal code for Japan, but promulgated a series of codes that regulated the affairs of the imperial court and the court nobility, the territorial lords, and a collection of verdicts in lawsuits. China remained a model when it came to the penal system, where banishment, corporeal punishment, and different forms of capital punishments formed the backbone of the Japanese penal system, supplemented by indigenous penal practices and imported ones, such as crucifixion. Just like in other states on the East Asian mainland, torture and other forms of physical pressures were used to obtain evidence and great emphasis was placed on confessions of guilt on the part of the accused. The Tokugawa state sponsored neo-Confucianism as a state ideology with increasing zeal and as part of these endeavors, Japanese Sinologists were employed to draft new legal codes based on the Tang and Ming Codes, efforts that were emulated in many feudal domains, but none of these efforts led to any wholesale reform of the Japanese legal order. Furthermore, in contrast to the contemporary Qing dynasty, laws were not made public and trials were usually held in camera, which meant that commoners could only learn about the law indirectly through publicly posted verdicts (*kōsatsu*).

Modern Transformations

Following the arrival of modern gunboats to the Western Pacific and the resulting conclusion of commercial treaties in the mid-nineteenth century, East Asian legal traditions were challenged for the first time by Western legal norms, in both their continental European and Anglo-Saxon varieties. Unwilling to submit themselves to what they regarded as barbarous and uncivilized systems of justice, the Western treaty powers forced their reluctant hosts to accept the practice of extraterritoriality, which put most foreigners outside of local laws and jurisdiction. The East Asian polities did not initially contest extraterritoriality in a systematic way, as many of them already possessed multiple legal codes and complex institutions for the adjudication of disputes between different social and ethnic groups. But when East Asian states embarked on nation-building projects that aspired to build militarily and economically strong states with an actively participating citizenry, extraterritoriality was increasingly perceived as an obstacle to those projects.

The first country to challenge the new framework of trade and diplomacy, often called the 'treaty port system,' was Japan, where a group of domains toppled the shogunate in 1867–68 and established the Meiji Emperor as the figurehead of a centralized nation state in the new capital of Tokyo. Realizing that the Western imperial powers would not agree to any revision of the commercial treaties or to the abolition of extraterritoriality without corresponding changes in the Japanese legal system, the oligarchs of the Meiji state embarked on an ambitious program to reform their legal and penal system to conform to contemporary Western standards. While early Japanese legal reformers initially drew on early attempts to reform the legal system along Chinese lines, these efforts were soon cast aside in favor of a wholesale legal and institutional reform, which left discernable few traces of the old Tokugawa legal order. Even the legal language changed beyond recognition, where Chinese characters formed compounds that were used to convey new legal concepts borrowed from the West. Reforming the criminal code, the judiciary, and the penal system did not present an insurmountable obstacle to Japanese legal reformers, but the promulgation of a civil code on a continental European model did meet considerable resistance from conservative elements within the Japanese government, who feared that continental European civil codes would undermine the 'national polity' (*kokutai*) of Japan. By the mid-1890s, the complete overhaul of the Japanese legal system was accomplished and, starting with Britain, most Western nations agreed to renegotiate the 'unequal treaties' and abolish extraterritoriality.

Qing China and Chosŏn Korea proved to be more resistant to foreign incursions into their legal and political orders. Although China was soundly defeated in a number of armed confrontations with the West, the first concerted attempt to reform the legal and political order did not take place until after the defeat in the First Sino-Japanese war 1894–95 and the Boxer War 1900, when the Qing government set out to abolish the old civil service, reform the administration of justice, abolish torture and physical punishments, and introduce criminal, civil, commercial, and administrative codes on Western and Japanese models. Although these reforms seemed promising, the Qing dynasty fell in 1911 and subsequent legal reforms had to be undertaken in a less stable political order that was plagued by civil war and foreign aggression. By the time the Republic of China had introduced a draft constitution and Six Codes, China was invaded by Japan in 1937 and the abolition of extraterritoriality in 1943 was more a product of great power rivalry than of legal reform. Korea was even less fortunate than China in that it soon fell prey to Japanese colonization and was annexed into the Empire of Japan in 1910, which meant that Korea's legal modernization took place under colonial conditions (Kim, 2012).

Following the end of World War Two and decolonization, East Asian nations went their separate paths, depending on which camp they had joined in the Cold War. The People's Republic of China and North Korea adopted Soviet-inspired legal institutions and codes, whereas Japan, Taiwan, and South Korea continued to draw on European and American models for legal reform. Economic reform in Mainland China after 1978 led to a second wave of legal reform, where the Beijing regime embarked on ambitious attempts at codification that are still underway. Today, very few formal traces of the East Asian legal tradition remain in either of these polities and it remains a moot point to what extent Confucianism, for instance, is still a relevant source of legal norms in East Asia. Yet recent advances in East Asian legal history have reopened several important questions on what constitutes a contemporary East Asian legal tradition and the economic rise of East Asia in general and China in particular will ensure that this discussion will be of perennial significance.

See also: Constitutionalism, Comparative; Law: Change and Evolution; Legal Pluralism; Legal Systems, Classification of; Postcolonial Law; Religions of East Asia.

Bibliography

Bodde, D., Morris, C., 1967. Law in Imperial China: Exemplified by 190 Ch'ing Dynasty Cases. Harvard University Press, Cambridge, MA.

Bourgon, J., 2002. Uncivil dialogue: law and custom did not merge into civil law under the Qing. Late Imperial China 23, 50–90.

Brook, T., Bourgon, J., Blue, G., 2008. Death by a Thousand Cuts. Harvard University Press, Cambridge, MA.

Ch'en, P.H.-C., 1979. Chinese Legal Tradition under the Mongols: The Code of 1291 as Reconstructed. Princeton University Press, Princeton, NJ.

Ch'ü, T., 1965. Law and Society in Traditional China. Mouton & Co., Paris.

Confucius, 1979. The Analects (D.C. Lau, Trans.). Penguin Books, Harmondsworth, New York.

Creel, H.G., 1970. The Origins of Statecraft in China. University of Chicago Press, Chicago.

Farmer, E.L., 1995. Zhu Yuanzhang and Early Ming Legislation: The Reordering of Chinese Society Following the Era of Mongol Rule. E.J. Brill, New York.

Goldin, P.R., 2012. Han law and the regulation of interpersonal relations: 'The confucianization of the law' revisited. Asia Major 25, 1–31.

Hahm, P.-C., 1967. The Korean Political Tradition and Law; Essays in Korean Law and Legal History. Hollym Corp, Seoul.

Henderson, D.F., 1965. Conciliation and Japanese Law, Tokugawa and Modern. Published for the Association for Asian Studies by University of Washington Press, Seattle, WA.

Huang, P.C., 1996. Civil Justice in China: Representation and Practice in the Qing. Stanford University Press, Stanford, CA.

Hucker, C.O., 1985. A Dictionary of Official Titles in Imperial China. Stanford University Press, Stanford, CA.

Hulsewé, A.F.P., 1955. Remnants of Han Law. Sinica Leidensia series. E.J. Brill, Leiden.
Jiang, Y. (Ed.), 2005. The Great Ming Code: Da Ming Lü. University of Washington Press, Seattle, WA.
Johnson, W.S. (Ed.), 1979. The T'ang Code. Princeton University Press, Princeton, NJ.
Kim, M.S.-H., 2012. Law and Custom in Korea: Comparative Legal History. Cambridge University Press, Cambridge, UK; New York.
McKnight, B.E., Liu, J.T.C. (Eds.), 1999. The Enlightened Judgments: Ch'ing-ming Chi: The Sung Dynasty Collection. State University of New York Press, Albany.
Ricket, W.A., 1971. Voluntary surrender and confession in traditional chinese law: the problem of continuity. Journal of Asian Studies 30, 797–814.
Schwartz, B.I., 1985. The World of Thought in Ancient China. Belknap Press of Harvard University Press, Cambridge, MA.
Song, C., McKnight, B.E., 1981. The Washing Away of Wrongs: Forensic Medicine in Thirteenth-century China. Center for Chinese Studies, University of Michigan, Ann Arbor.

East Asian Studies: Culture

Emiko Ohnuki-Tierney, University of Wisconsin, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The studies of the cultures of China, Taiwan, Korea, and Japan are outlined with a major emphasis on the transnational exchanges among them in the past and present, resulting in the basic sharing and yet stark differences in sociocultural institutions, such as religion. Instead of envisioning these cultures as hermetically sealed and localized, this article points to the internal diversities, including minorities, and diasporas and transnationals, on the one hand, and global influences on these peoples and their cultures. Nationalisms within each society are discussed in relation to the colonialisms by Western and Japanese empires in the past.

Recent developments in the studies of East Asia parallel many significant paradigmatic changes in the behavioral and social sciences in general, serving as windows of opportunity for raising new questions and issues and engaging in fruitful comparisons with research in other parts of the world. A number of productive directions have opened up because the terms and concepts in social sciences came under scrutiny, offering us opportunities to reconceptualize how to approach and study 'people', 'culture', and 'society', without treating them as homogeneous and bounded entities. Increasing attention to power inequality within a society – be it as small as a family or as large as a nation-state – and without – external geopolitics has sensitized the studies of East Asia so that a microstudy of a social group is linked to macro scenes of political economy both within the society concerned and without, leading to fruitful cross-disciplinary approaches and comparative studies with other areas of the world (Anagnost, 1997). In this context, the West, which served as a mirror for East Asian peoples for a long time, is now being reconceptualized so that dynamic relationships among East Asian cultures and peoples, within the context of world history, are receiving more emphasis than in the past (Wang, 2014).

'East Asian Studies' is a well-established area studies program with a long and distinguished tradition of studying the peoples of East Asia and their lives. 'East Asia' appears to be a fairly well-demarcated geographical area, consisting at present of the People's Republic of China (mainland China, now including Hong Kong), the Republic of China (Taiwan), North and South Korea, and Japan. However, shifting political boundaries – even at the start of the twenty-first century – and recent epistemological and theoretical questions in various fields have reminded us that 'East Asian Studies' is a highly complex undertaking. Building on excellent studies of the peoples of East Asia and their cultures, the field is now invigorated by fresh insights and perspectives.

Nonlocalized Cultures

Historically the Chinese, Korean, and Japanese 'civilizations' have had much in common, to a large measure due to the diffusion of the Han and Tang civilizations from China to the rest of the area, where each local tradition has significantly transformed these influences. Linguistically, the Sino-Tibetan languages, all of which are characterized by the tonal system (whereby tones give distinctive meanings to the same sound types), belong solely to the Chinese. The Korean and the Japanese, together with the Ainu and Gilyaks, come from an altogether different, nontonal, linguistic stock. Yet the Chinese writing system was adopted and superimposed on the spoken languages of the Korean and the Japanese, although both added their own systems at a later time.

Recent changes of emphasis and perspective in East Asian Studies derive from criticisms of, and even some frontal attacks on, the concept of culture, the most fundamental concept in anthropology, though it has come under attack in recent decades. Nonetheless many continue to find 'culture' concept useful, together with 'society' and even 'nation', not as hermetically sealed, but, pulsates in the vertigoes of world history and always transforming at the confluence of internal/local historical development and external/global historical forces. Almost all major foundational features of Japanese culture and political economy came from the Asian continent, usually via Korea. They include the 'wet-rice economy' in early times, and later – during the fifth and sixth centuries – Buddhism, the imperial ritual, the writing system, city planning, metallurgy, and a host of other activities. Since the end of the nineteenth century, the major influences came from Euro-American cultures.

Japan's first legal system, instituted first in CE 645 through the so-called Taika Reform, was based on the political system of the Tang Dynasty. The draft of the Meiji constitution, promulgated in 1889, was drawn up by primarily German and French legal scholars, with the Prussian constitution as their model. Many of the key articles were adopted almost verbatim. The current constitution was 'given' by the American government at the end of World War II. Indeed, the modernization in Japan, as in other Asian countries, was, to a great extent, Westernization. No society is an island in history. Hybridity is a *sine qua non* of any culture. Astronomically fast cyberspace communication today makes us keenly aware of the basic nature of culture as a dynamic, unending process of the interpenetration between the local and global.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10120-5

Dynamics Within

A powerful critique of the notion of culture in the past as a 'totalizing' and 'essentializing' representation of a people or a culture led our studies in several productive directions. One such direction is the further emphasis on diversity, multiple voices within an artificially demarcated unit of our study – be it a people, a culture, or a society – resulting in a plethora of research on social groups other than the dominant ones (Feuchtwang and Wang, 2001; Slater and Ishida, 2009). These groups include not only the so-called minorities but also the homeless, school dropouts, etc. This move is a corrective to the representation in the past, for example, of the Chinese by the Han Chinese, or the Japanese as the agrarian Japanese: the emphasis on the latter led to a label, 'rice paddy courses', referring to introductory courses on Japan. The recent research on minorities places these social groups not as isolated groups but in relation to dominant social groups, especially but not exclusively in regard to power inequality between the dominant and marginalized social groups.

In the case of Japan, wet-rice agriculture, which originated likely along the middle Yangtze River, reached northern Kyūshū some time around 350 BCE. Called Yayoi culture, it supplanted the Jōmon culture based on foraging economy of the first occupants of the archipelago, who had arrived there around 200 000 BCE (Kobayashi, 2003). It laid the foundation for the political economy of the Yamato state and the imperial system. Gradually the segment of the population engaged in rice agriculture – or, more accurately, who reaped the economic/financial crop from wet-rice economy – became the dominant social group, representing 'the Japanese'. The agrarian cosmogony and cosmology became those of the Japanese, whereby the rice agrarian seasonality governed all Japanese and rice paddies came to represent the Japanese space.

There is an increasing awareness now among scholars that the other side of the development of 'Agrarian Japan' was the marginalization of 'the Japanese' in nonagrarian sectors. They included those people whose livelihood derived from the resources of the sea and the mountains. In particular, this included what later became known as the 'hisabetsubu-rakumin' (settlement people under discrimination). These were people who engaged in nonagrarian occupations such as performing arts, architecture, and trading, plus itinerant religious specialists, who, from late medieval times – when the notion of impurity received radical negativity – became increasingly marginalized, until 1582 when they were codified as two categories of 'outcasts'. The marginalization of the Ainu, who continued to engage in hunting, fishing, and plant gathering well into the nineteenth century, was similarly based on the yardstick of agrarian values, which placed emphasis on plant domestication. In addition, the assignment of 'primitiveness' to them was a by-product of the Japanese, who tried to 'modernize' their nation according to the model of Western modernization (Ohnuki-Tierney, 1987, 1993).

A related development is an emphasis on women – their voices, their life experiences, and women as social agents who actively carve out or change the courses of their lives in a globalized society (Abelmann, 2003; Ngai, 2005). The issue of gender preferences, both institutionalized and culturally sanctioned as well as individual, has also received considerable attention. Gender-related issues often address the question of gender inequalities, which are in turn juxtaposed with the power inequalities between classes within society, on the one hand, and between the colonizers and colonized, on the other, breaking new ground for various disciplines.

Diasporas and Transnationals

Another major critique of culture is how it was presented in the past as being spatially bounded and tied. The nonisomorphic nature of the cultures, locality, and peoples of East Asia was realized early by scholars of East Asia, who have contributed to the studies of 'overseas communities' (e.g., Chen 1939) – numerous 'Chinatown' studies, Japanese Brazilians (Tsuda, 2003), and the like. Scholars of 'transnationalism' study emigration, the resultant overseas communities (some diasporic), and 'transnationals' with 'flexible citizenship' in the – contemporary context of globalization, which created an unprecedented flow of population across borders of nation-states (Ong, 1999). The recognition of the nonisomorphic relationship between a political unit – such as a nation-state – and a people/culture became more than obvious in East Asia when 'the Chinese' were divided into those in the People's Republic of China, those in Hong Kong – first under the British and then the People's Republic of China – and those in the Republic of China (Taiwan) across the Formosa Strait. Likewise, 'the Koreans' were split into North and South. Japan, which has been stereotyped as homogeneous in terms of demographic composition, has been a heterogeneous population throughout history. But, to give one example, in 1940 the then prime minister of Japan, Konoe Fumimaro, delivered a now infamous speech that "one hundred million Japanese with one soul" be dedicated to the service for the Emperor. This figure included 30 000 000 Koreans, Chinese, and others who resided in Japan and its occupied territories, who were given an equal right to death, that is, from 1944 they were forced to become Japanese with the Japanese soul in order to be drafted to fight for imperial Japan (Ohnuki-Tierney, 2002). This is not to mention the diversity within the so-called Japanese, with the Ainu to the north, and the Okinawans to the south, who were once thought of as distinct from those on the main islands, and 'Westerners', including the 'White Russians' and Jews.

The political mosaic in East Asia offers a fertile ground for studies of 'identities' and conflicts among social groups within a nation-state and between nation-states. Within a nation-state, the majority vs minorities are interlaced with the question of what makes a social group 'ethnic' (the Okinawans in Japan 'used to be' ethnic); whereas the two Chinas and two Koreas present a different situation – the question of two nation-states consisting of one people. In other words, East Asia offers a microcosm of what we see in the world in which the same collective identity is held by people in different locals.

Nationalisms, Colonialisms, and Imperialisms

Research on the complex problem of identities in East Asia will join the now burgeoning field within social science of ethnicity,

nationalism, and ethnonationalism, as well as a more legally defined 'citizenship'. Ethnonationalism is 'the politicization of ethnicity', related to 'the generation of regional or subnational reactions and resistances to what is seen as an overcentralized and hegemoic state' (Tambiah, 1996). Colonialism in East Asia is another genre of research in which interest is growing, and which promises a rich source of comparative work. Studies of Western colonializations have yielded rich fruit for scholars in the past. The Opium War (1839–1942) and the resultant cessation of Hong Kong became the symbol of the power of Western (British) colonialism over Asia, whereas the Boxer Rebellion (1899–1900), crushed by the combined forces of Western and Japanese powers, foreshadowed not only Western but Japanese imperialism; the latter leading to the colonization of a large part of Asia, with its worst case in Korea (1910–45) (Duus, 1995), These studies offer important comparative bases for Western and non-Western colonializations in Africa, the Americas, and other parts of the world.

At the level of nation-states, the identity of a people could be expressed through peaceful 'cultural nationalism', which defines and represents a people of a nation-state as distinct from the 'other(s)', often those who, physically or symbolically, press at the 'border' – figuratively and politically. Thus, when the Japanese began to emerge from their total immersion in the Han and Tang civilizations of China, they began to assert their own culture, switching from embracing the Chinese aesthetics of plum blossoms to that of cherry blossoms as their flower. Since the mid-nineteenth century Japan's nationalism, both cultural and political, has been informed by paradoxes and ambivalence of the Japanese's own identity, which was formed in part by rejecting Western political imperialism – while at the same time eagerly emulating Western civilization – and in part by rejecting their own Asian identity – while seeking it in order to establish a difference from the West.

Cultural imperialism is a genre that has not yet received systematic attention. Political domination by Western European nations (but also Japan to a lesser degree) led to the cultural mirror to emulate for many Asian peoples who are often eager to adopt not only Western science, technology, and capitalism, but almost everything from the 'West' being revered, including foodways, the body esthetics, resulting in skyrocketing sales of cosmetics and plastic surgeries. Even before affluence led the Japanese to seek out designer names such as Gucci, Bally, and Louis Vuitton, American Revlon lipsticks, and Ivory soap were far more expensive than their domestic counterparts and were preferred as gift items.

This is an important area of study, not only because it offers us a valuable window onto the other side of political colonialism but also because it constitutes a significant dimension of the current globalization, in which the notion of brain death, organ transplantation, human rights, environmental conservation, and the like, which involve concepts and values that are predominantly Western in origin, are adopted outside their areas of origin. Medical high technologies, saluted as the golden achievement of Western biomedicine, bring prestige to those in another society who embrace them as being 'progressive', although such technologies may require not only radical reconceptualization about the death in relation to the body but also mortuary rituals, with beliefs, values, kinship, and a host of other important dimensions for a people (Ohnuki-Tierney, 1994).

Global Forces

The 'nonlocalized culture' presented earlier is in part facilitated by transnational forces that had been quite strong and intensive among East Asian societies from very early on. Despite difficulties involved in travel at the time, people, including scholars and religious specialists, frequently traveled between these three countries, resulting in a great deal of sharing among Chinese, Korean, and Japanese cultures. Buddhism, originally from India, traveled from China to Korea and Japan. In Japan, Buddhism made an enormous impact, not only on the indigenous religion of Shintoism, but also in visual and performing arts, rock garden, and a host of other intellectual and cultural activities. Confucianism, originally developed in China, established the strongest hold in Korea. One cannot ignore, on the other hand, the impact of 'Western' religions. The institutionalized Christianity is extremely strong in Korea, whereas it penetrated into Japanese culture intensively, though not retaining the institutionalized forms. On the other hand, many influential intellectual and political leaders of Japan since the Meiji restoration in 1868 have been devout Christians, some of whom assumed important roles as intellectual and moral leaders of the country.

The present globalization is marked by the spread of high technology and market economy of late capitalism. The global flows of people are studied as research on tourism, diaspora, and transnationals. The flows of goods and signs/symbols are studied through research on global dissemination of youth fashion, fast food, television programs, films, pop music, and a host of other things that are engulfing the globe. Most originate in the West, but some, such as reggae, do not. The studies of the introduction of McDonald's – a symbol of postmodern fast-lane life in the USA and an epitome of Americana for peoples elsewhere – show a simultaneous transformation of the locale, which in turn transforms the fast food in significant ways.

Without denying the importance of an unprecedented flow of people, goods, and signs, others urge us to compare the present with the past when spice trades, capitalism, colonialism, features of Western High Culture, etc., spread throughout the world. In short, East Asian cultures and peoples have been 'globalized' for a very long time.

Cultural, Social, and Political Institutions

East Asian studies continue to produce distinguished scholarship on: kinship and social organization; religions (beliefs and rituals); educational systems; visual and performing arts; and health-care systems. Excellent studies of the institutionalized religions of Buddhism and Confucianism, shared by most Asian societies in variant forms, and of Shintoism in Japan describe how they have impacted on sociopolitical organizations as well as arts and science. They also tell us how people experience these religions in their daily lives. The emphasis on the impact of religion in individual lives receives the attention

in the studies of 'ancestor worship' shared by most peoples of East Asia (Ahern, 1973; Janelli and Janelli, 1982; Smith, 1974). 'Ancestor worship' is a misnomer but refers to the practice of taking care of the souls of deceased family members and relatives through a long series of rituals. Despite profound changes that such worship underwent in each society, it remains perhaps the most important aspect of religious beliefs and rituals in these societies. Shamanism and other religious beliefs and practices that have been regarded not to belong to institutionalized religions are also well studied in relation to the day-to-day lives of contemporary peoples. In contemporary Korea, for example, shamanism is of exceptional importance and recent studies highlight the role of women in Korean shamanism (Kendall, 1985). Although their pervasive influence is less recognized, Taoism and geomancy, as practiced in China and spread to Japan, constitute important dimensions of cosmology comprising symbolic oppositions (expressed in yin-yang) and remain important in the daily lives of people in East Asia.

The academic study of health care has a long tradition in East Asia. 'Chinese medicine' in particular has received attention from distinguished scholars such as Joseph Needham, who explicated how their medicine constitutes a science distinct from Western science (Needham, 1969), and Manfred Porkert, who later produced a similarly distinguished work (Porkert, 1974). Contributors to Kleinman (1976) volume focused on doctors, patients, and their interactions, rather than medicine as a system.

See also: Asia, Sociocultural Aspects: China; Asia, Sociocultural Overviews: Korea; Colonization and Colonialism, History of; East Asian Studies: Politics; Globalization and World Culture; Japan: Sociocultural Aspects.

Bibliography

Abelmann, N., 2003. The Melodrama of Mobility: Women, Talk, and Class in Contemporary South Korea. University of Hawaii Press, Honolulu, HI.

Ahern, E.M., 1973. The Cult of the Dead in a Chinese Village. Stanford University Press, Stanford, CA.

Anagnost, A., 1997. National Past-Times: Narrative, Representations and Power in Modern China. Duke University Press, Durham, NC.

Chen, T., 1940. Emigrant Communities in South China. Institute of Pacific Relations.

Duus, P., 1995. The Abacus and the Sword: The Japanese Penetration of Korea, 1895–1910. University of California Press, Berkeley, CA.

Feuchtwang, S., Wang, M., 2001. Grassroots Charisma: Four Local Leaders in China. Routledge, London.

Geertz, C., 1995. After the Fact. Princeton University Press, Princeton, NJ.

Janelli, R.L., Janelli, D.Y., 1982. Ancestor Worship and Korean Society. Stanford University Press, Stanford, CA.

Kendall, L., 1985. Shamans, Housewives, and Other Restless Spirits: Women in Korean Ritual Life. University of Hawaii Press, Honolulu, HI.

Kleinman, A. (Ed.), 1976. Medicine in Chinese Cultures. US Government Printing Office for Fogarty International Center, Washington, DC.

Kobayashi, T., 2003. Jomon Reflections: Forager Life and Culture in the Prehistoric Japanese Archipelago. Sainsbury Institute. Oxbow Books, London.

Ngai, P., 2005. Made in China: Women Factory Workers in a Global Work Place. Duke University Press, Durham, NC.

Needham, J., 1969. Science and Civilization in China. Cambridge University Press, New York.

Ohnuki-Tierney, E., 1987. The Monkey as Mirror. Princeton University Press, Princeton, NJ.

Ohnuki-Tierney, E., 1993. Rice as Self: Japanese Identities through Time. Princeton University Press, Princeton, NJ.

Ohnuki-Tierney, E., 1994. Brain death and organ transportation: cultural bases of medical technology. Current Anthropology 35 (3), 233–254.

Ohnuki-Tierney, E., 2002. Kamikaze, Cherry Blossoms and Nationalisms: The Militarization of Aesthetics in Japanese History. University of Chicago Press, Chicago.

Ong, A., 1999. Flexible Citizenship: The Cultural Logics of Transnationality. Duke University Press, Durham, NC.

Porkert, M., 1974. The Theoretical Foundations of Chinese Medicine: Systems of Correspondence. MIT Press, Cambridge, MA.

Slater, D., Ishida, H. (Eds.), 2009. Social Class in Contemporary Japan. Routledge, London.

Smith, R.J., 1974. Ancestor Worship in Contemporary Japan. Stanford University Press, Stanford, CA.

Tambiah, S., 1996. Leveling Crowds: Ethnonationalist Conflicts and Collective Violence in South Asia. University of California Press, Berkeley, CA.

Tsuda, T., 2003. Strangers in the Ethnic Homeland: Japanese Brazilian Return Migration in Transnational Perspective. Columbia University Press.

Wang, M., 2014. The West as the Other: A Genealogy of Chinese Occidentalism. Chinese University Press, Hong Kong.

East Asian Studies: Economics

Vivek Chibber and Doug Guthrie, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp. 3957–3962, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article provides a comparative analysis of the dynamics and challenges of East Asian economies. These economies have undergone a radical transformation since the early 1970s that have shifted the balance of power both within the region and in the global economy more broadly. The first major change that characterizes this period is the rise of the newly industrial economies of Taiwan, South Korea, Singapore, and Hong Kong. The second major change has been the transformation of China's planned economy. The successes in various development projects in these economies have been tempered by economic challenges in the region, namely that of Japan in the 1990s, the Asian financial crisis of 1997, and the considerable challenges still facing the Chinese reform effort, all of which have shown the vulnerabilities of this regional economy.

The Newly Industrialized Economies

The rapid development of East Asian economies – the advanced industrial economy of Japan, the newly industrialized economies (NIEs) of Taiwan, South Korea, Singapore, and Hong Kong, and China's transitional economy – have made this region one of the most dynamic in the global economy. In the latter half of the twentieth century, Taiwan, Singapore, and Hong Kong were among the most dynamic economies in the world, having leapt, in the postwar period, from the status of poor underdeveloped to strong industrialized economies (Clark and Roy, 1997), and their patterns and trajectories of change challenged conventional wisdom about how developing nations modernize and industrialize, and about the role of the state in these processes of change. The three decades following World War II were a period of enormous enthusiasm about the economic prospects for developing countries. These were the years when a large number of such countries embarked on ambitious projects of agricultural and industrial modernization. Indeed, it was this effort that gave rise to 'development economics,' and more broadly, 'development studies,' as new disciplines in the social sciences. In the halcyon years of development planning, most eyes were focused on the larger or more industrialized countries such as India, Brazil, Mexico, and Egypt. Few expected very much, if anything, from East Asia, large areas of which had been colonized by Japan in one way or another in the first half of the twentieth century, and parts of which seemed hopelessly backward. Taiwan had a minuscule industrial base in 1950, Hong Kong and Singapore were tiny city-states, and South Korea seemed to be sinking in a morass of corruption and inefficiency by the end of the 1950s. Yet, a quarter of a century later, the tables had been turned: East Asia emerged as the economic 'miracle' of the postwar period, witnessing some of the fastest growth rates of the modern epoch, while Latin America and South Asia were mired in political and economic crises. Given this regional growth phenomenon, does it make sense to talk of an 'East Asian Model?' There are some important similarities in development strategies across the cases, enough to persuade many commentators that the appellation is warranted. But there are also interesting differences that are worth noting.

Among the East Asian NIEs, South Korea and Taiwan have attracted the most attention. Over the years, two general views have emerged about the reasons behind their spectacular economic success. One, which was perhaps the dominant one for much of the 1970s and 1980s, held that the two countries were successful because of their fidelity to free-market policies, particularly an open trade regime (Lal, 1985; Bhagwati, 1988). But this view has come under a great deal of criticism from many quarters, and a new argument has emerged in recent years, which offers a different explanation for Korean and Taiwanese success. In some ways, the two countries achieved their remarkable economic success the 'old-fashioned' way: both started with highly successful land reforms, which immediately set them apart from much of the developing world; both managed to bring about extraordinarily high savings rates, which were then channeled into correspondingly high rates of investment. But the two countries did not simply grow by increasing the rate of accumulation. They also succeeded in acquiring, and then unpacking, technology from advanced industrial countries, so that they were able to move up the production chain into manufacturing and exporting high value-added goods. The involvement of the state was crucial to this process (Gold, 1985). First, both governments launched highly effective campaigns to increase literacy rates, which accounted for a highly educated work force (World Bank, 1993).

Furthermore – and this has been one of the most heralded facts about their experience – the two states played a crucial role in facilitating the acquisition and transfer of imported technology. These technologies were brought in sometimes through foreign direct investment, but mostly through licensing agreements with multinational corporations (MNCs). Agreements were carefully crafted to ensure that foreign partners did not simply hand over obsolete technology, and that local engineers were able to acquire blueprints as well as the training necessary to master the technology. Firms were also encouraged, assisted, and even coerced to produce for export markets, in order to submit them to the rigors of international competition (Amsden, 1989; Wade, 1990). Success in export markets was used by both states as a benchmark for efficient production. Even firms producing mainly for the domestic market often had to show export success in order to be granted credit, investment licenses, and other state benefits (Aoki et al.,

1997). Finally, the state was active in maintaining, for many of the crucial years of take-off, a tightly policed and powerless labor force, which added to the 'flexibility' of production.

While South Korea and Taiwan shared these elements in their development strategies, they also exhibited some important differences. Korea emphasized a reliance on the private sector, which was not only encouraged to grow, but also to become increasingly concentrated. In this, it resembled the Japanese experience: the Korean economy was dominated by a small number of highly diversified business groups, known as the Chaebol, which together accounted for the majority of the country's domestic product. The Chaebol quickly emerged as the pivot on which industrial policy turned, establishing a close working relationship with the ruling regime (Amsden, 1989). The Taiwanese state, on the other hand, did not have such a close relationship with private capital. Taiwanese planners relied much more on a large agglomeration of state enterprises as the nucleus for development policy; private capital was influenced – in comparison with the Korean case – more through pricing and procurement policies than by direct negotiation. The private sector was, in turn, not actively encouraged to become highly concentrated. Taiwanese industry was, much more than its Korean counterpart, characterized by small enterprises (Wade, 1990). Nevertheless, a kind of quasi-corporatist structure was used to bargain with some sectors. Indeed, in both countries, a similar structure was put to use for this: firms in targeted sectors were organized into sectoral or trade associations, which were given the right to bargain with the state (Fields, 1995). The overall complexion of industrial policy, however, remained visibly different in the two cases, with Korea relying on a highly concentrated private sector, and Taiwan more on state enterprises and small firms.

Singapore and Hong Kong are usually regarded as contrasting cases to Korea and Taiwan, in that whereas the latter two are examples of state-led development, the former pair embodies a free-market path – or so the story goes (Rabushka, 1979). There is some element of truth in this. Hong Kong, in particular, did not experience anything like the selective industrial policy witnessed in the larger countries. And both of the city-states were much more open to foreign direct investment than was either Korea or Taiwan. But there are important respects in which the city-states converged with the experience of their larger neighbors. First, like Korea and Taiwan, both Hong Kong and Singapore actively suppressed labor rights. Second, the state was of some importance in facilitating industrialization. Singapore, despite being more open to foreign investment, relied greatly on state enterprises to facilitate the entry of MNCs by forming partnerships with them, hence reducing the risk of such undertakings; further, the domestic market was not thrown open to imports, and was protected much like that in Taiwan and Korea. Local firms were also offered carefully targeted fiscal incentives, under the guidance of an Economic Development Board. Hong Kong was, of the four, the furthest away from the interventionist model. But here, too, the state played a crucial role in solving a host of collective action problems for local firms: most directly through its considerable expenditure on infrastructure, training programs, and, perhaps most importantly, its massive socialization of wage costs through the construction of an extremely generous welfare state. Between 1949 and 1980, government expenditure in Hong Kong grew by 26 times, and expenditure on welfare by an amazing 72 times (Castells, 1994). This was absolutely crucial to the success of Hong Kong exports, in lowering the wage costs to producers.

In sum, despite their differences, the four cases do exhibit certain important similarities. Perhaps the most important of these is the critical role played by the state in fostering the transition to highly productive industrialized economies. The differences lie mainly in the instruments used by the various regimes: selective intervention in the larger two countries, and more macro-level policies in the smaller ones.

China's Transforming Planned Economy

While the NIEs were executing successful development processes, the first three decades of Communist Party rule in China (1950s through 1970s) were organized around the establishment of a socialist planned economy (also called a 'command' economy). In this system, markets were eliminated and replaced by governmental plans that guided the economic decisions of enterprises and individuals. China's one-party government replaced private ownership with state ownership, established economic targets for industrial enterprises, and set up collective farming in the agricultural sector. Economic agents (whether industrial enterprises or farmers) were instructed to whom they would sell their goods and for what price. Revenues were then collected by the government and redistributed to enterprises and individuals in the economy. This system was thus more than a purely economic system; it was also the basis for the social security system in China (Walder, 1986; Whyte and Parish, 1984).

Without a competitive marketplace to regulate and guide economic behavior, planned economies, such as China's, are often rife with inefficiencies. In 1978, after a decade of upheaval wrought by the Cultural Revolution, the 'Ten Year Plan' of 1976 was already faltering, and the need for some degree of economic reform was clear. Following a political struggle that pitted reform-minded Deng Xiaoping against the economically conservative Hua Guofeng, Deng Xiaoping took control of the economy in late 1978 at the Third Plenum of the Eleventh Central Committee of the Chinese Communist Party. Once in control of the government, Deng launched China on a path of economic transformation.

One of the central debates in research on transitions from planned to market economies revolves around the extent to which privatization is a necessary part of this process. Some scholars have argued that without rapid and complete privatization – the so-called 'shock therapy' approach to reform – true economic reform is impossible (Kornai, 1990; Sachs and Wing Thye, 1994). Other scholars have argued that China is the case that disproves this view: in contrast with the rapid changes that characterized the economic reforms of Russia and Eastern Europe, China's reforms have been introduced in a gradual and stepwise fashion, where the government has experimented with new economic policies and incrementally implemented the successful practices on a wide scale (Naughton, 1995). Proceeding in this gradual fashion, the Chinese transition to a market economy has been unquestionably the most successful in the world. (China sustained the fastest rates of growth in the

world for the first decade and a half of economic reform. Throughout the 1980s, its real gross domestic product (GDP) grew at an average annual rate of 10.2%, a level that was equaled only by Botswana. From 1990 to 1996, the average annual rate of growth for real GDP was 12.3%, the highest rate of any country in the world for that period. It has also had the highest industrial growth rate (an amazing 17.3% average annual growth, 1990–96) and the second highest growth in services (9.6% per annum, 1990–96) in the world.)

In the industrial sector, often regarded as the backbone of the Chinese economy, five key changes have reshaped this system. First, the government gradually removed itself from day-to-day decision-making, effectively placing the responsibilities of economic decisions on the shoulders of industrial managers. In the early years of the reforms, enterprises were still expected to produce a certain amount of goods for the government's plan, but anything they produced above this residual plan they were allowed to sell in China's emerging markets. This transitional system was called the 'dual-track' system (Naughton, 1995). Second, a private economy was allowed to emerge to compete with the state sector, a factor that was critical for the creation of a competitive marketplace. Thus, while privatization was the cornerstone of industrial reform in other transforming planned economies, Chinese firms were introduced to market economic behavior without privatization (Rawski, 1994), and state firms eventually 'grew out of the plan,' producing an increasing proportion of their goods for the marketplace (Naughton, 1995). The success of this transition, however, has varied in accordance with a firm's position in the former command economy: firms under township and village governments, most of them the so-called Township and Village Enterprises (TVEs), have been managed closely by local governments and thus have weathered the reforms much more successfully than those under the central and municipal governments (Walder, 1995).

A third change that has defined China's economic reforms is that foreign investors were allowed in the Chinese economy for the first time since the communists expunged the foreign presence in the early 1950s. This factor has been very important for the transfer of technology and the transfer of international management practices. Fourth, the government adopted a decentralized and piecemeal approach to development regionally: in the late 1980s, the government adopted the coastal development strategy, in which the goal was 'state-directed' export-oriented growth (Yang, 1991). Finally, Chinese enterprises have been relieved of their responsibilities in the provision of social security benefits to their employees and, perhaps more significantly, they are no longer required to make lifetime commitments to the workers they employ. In the agricultural sector, a similar process of change has unfolded where farmers have been given the freedom to sell whatever they produce beyond the plan in the marketplace, until eventually the market has overtaken any vestiges of an economic plan in this sector. All of these economic changes have been formalized through the passage of some 700 national laws and more than 2000 local laws (Pei, 1994).

Transforming the planned economy of the world's most populous nation has been a complex process and is far from complete. Nevertheless, the first two decades of transition from plan to market have been impressive and, compared to the reforms in Russia and Eastern Europe, a dramatic success. For the first decade and a half of economic reform, China had the fastest growing economy in the world, sustaining double-digit growth figures for much of the 1980s and 1990s. Where it was a third-world developing economy as late as the early 1980s, by the beginning of the twenty-first century China had the seventh largest economy in the world overall in terms of GDP, and it is second only to the USA when GDP is adjusted for purchasing power within the country. (For example, if a bushel of rice costs half as much in China as in the USA, then a unit of currency is twice as valuable there. GDP statistics can, when appropriate, be adjusted for 'purchasing power parity' (PPP): a statistic that adjusts for cost of living differences by replacing exchange rates with relative measures of local purchasing power.)

Among the challenges that still lie ahead are the continued reforms of the state sector – the leviathan organizations that remain inefficient and insolvent after more than 20 years of reform; the continued reform and clarification of property rights; the liberalization of several critical sectors, such as the telecommunications sector, and the development of the western provinces, which have fallen far behind the development in the eastern coastal region.

Japan's Lost Decade and the Asian Financial Crisis

It is, in part, because of the great successes of development in East Asia that recent economic crises have hit the region so hard. Two related, but distinct, economic crises are worthy of note here. The first is the economic downturn that Japan suffered in the second half of the 1990s. The second is the Asian financial crisis, which gripped much of East and Southeast Asia in 1997 and 1998.

Japan was, for most of the postwar period, one of the fastest-growing economies of the industrial world, racking up huge rates of growth year after year until it had achieved its current position as the second largest economy in the world, with an annual GDP of more than $5 trillion. Its economy was the engine for the broader regional development of East Asia, both through its capacity as a market, and as a source of capital and technology. A more intangible, but nevertheless important, role was that it provided the model on which the two Northeast Asian 'tigers' – Taiwan and South Korea – based their economic strategies.

By the end of the twentieth century, however, the Japanese dynamo seemed to be losing its energy. It spent most of the 1990s mired in an economic recession that turned out to be the longest and most severe in its postwar history. The precipitating event behind the crisis was the collapse of the asset bubble of the 1980s, which burst in the beginning of the next decade. The bubble itself had been built up during the 1980s as the Japanese state had initiated a policy to ratchet upward the domestic rate of investment and make the economy less dependent on external markets. Spurred on by the constant pressure by the USA, to raise the value of the yen, the Japanese government moved, by the late 1970s, to trigger a massive investment drive by raising the value of domestic assets. At one level, it worked very well – as asset prices rose, Japanese firms launched a massive drive to retool and invest in new technology. But the flow of investment soon exceeded anything that actual returns

could justify, driving up asset prices steadily into a speculative bubble. Sooner or later all speculative bubbles burst, as did this one in the 1990s. And once the bubble burst and the economy slowed, firms found themselves stuck with massive over-capacity and onerous levels of debt. Through much of the decade, business and residential investment continued to fall, throwing the economy into a downward spiral.

The Japanese downturn was one of the background factors that laid the conditions for the East Asian crisis of 1997–98, which is now regarded as the most significant economic catastrophe the world has seen since the Great Depression. As the economy slowed down, firms compensated by turning outward, into a massive wave of investment in Southeast Asia. This led, in turn, to a process not unlike that was witnessed in Japan: corporations went into deep debt to pay for new investments, leading to the build-up of extraordinary industrial capacity; after a point, money also began to pour into real estate and other forms of wealth, driving up their prices until a speculative bubble was created reminiscent of the Japanese bubble.

Although much of East Asia had bounced back from the Asian financial crisis by the spring of 2000, this crisis revealed some weaknesses in the rapid development of the 'Asian Tigers.' The crisis itself was set off by currency devaluation: in the spring of 1997, as the pace of growth in the Thai economy fell below projections, currency investors began to speculate against the baht in hopes of profiting from the inevitable devaluation of the currency. When the currency devaluation came in July, foreign investors fled from Thai markets, and similar behavior spread quickly throughout the region. Stock prices plummeted, currencies went into free fall, and banks began to call in loans that their customers – many of them large corporations – could not repay. By the end of 1997, the International Monetary Fund had bailed out Thailand, Indonesia, and South Korea with $17 billion, $43 billion, and $57 billion, respectively.

Two explanations have been advanced to account for the crisis. In one view, an over-reliance on foreign capital and international institutions, changes in regional economies – namely China's and Japan's – and reliance on the dollar brought about the crisis (Ahn, 2000). First, the strength of the Japanese yen in the 1980s and that of the dollar in the 1990s led to excessive foreign investment in the strong East Asian economies, particularly for Japanese investors. As local banks sought strong currencies, investment and lending agreements were structured to attract these foreign players, and excessive capital led to high-risk investments, many of which would prove to be insolvent when the crisis stripped the veneer of economic discipline. Second, as China's economy began to expand, it took over a large share of the export market, the centerpiece of development for many East Asian Economies. Third, when European currencies depreciated against the dollar in 1995, dollar-pegged East Asian currencies appreciated – relatively speaking – leading to a decline in exports. Fourth, as the flow of international capital increased through such institutions as hedge funds and short-term loans, speculative behavior on the part of international currency investors in these rapidly expanding financial markets led to market volatility.

A second explanation acknowledges the problems listed above, but places the blame of the crisis on the East Asian economies themselves. This explanation often veers into a cultural argument about the prevalence of 'crony capitalism' in the region. In this view, the East Asian business model, often led by government-directed investment in key industries, is also often built around implicit guarantees from the government for the liabilities of local financial institutions (Krugman, 1998). Thus, the so-called 'soft budget constraint' (Kornai, 1994), which was the hallmark of planned economies, some economists argue, was also fundamental to the East Asian business model more generally. Governmental guarantees led to risky behavior on the part of financial institutions, which ultimately led to the financial crisis itself.

The Balance of Power: Regionally and Globally

Despite the economic crises experienced in the region in the 1990s, overall the second half of the twentieth century saw a dramatic shift in the economic balance of power both intra-regionally and with regard to Asia's position globally. Within the region, while Japan was far and away the economic powerhouse of the region by the end of the 1980s, at the dawn of the twenty-first century the balance of power looked very different. Partly because of the rise of the NIEs, but perhaps more importantly, because of the dramatic growth of China's economy, Japan's economic dominance in the region has unquestionably waned. It is difficult to predict what the exact economic balance of power will be within the region in the coming decades, but it is certainly clear that China will challenge Japan as the center of the Asian economy. In particular, the conflict between Russia and Ukraine, and the ensuing worsening of relations between Russia and the European Union and the USA, is leading Russia into greater dependence on China as a market for raw materials, especially oil. A major oil agreement between China and Russia might lock Russia into the position as a relatively cheap supplier of hydrocarbons – and this, coupled with China's increasing investment in alternative sources of energy, might further facilitate China's potential for continued growth.

Globally, the economic balance of power has shifted as well. While Japan has been counted among the Group of Seven (then Group of Eight, and again Group of Seven) industrialized nations since the organization's inception, during most of the twentieth century Western economies have dominated the global scene. However, with Japan and China both among the top 10 largest economies in the world and four Asian economies among the top 20 (five if we include Australia), and with China poised before long to have the world's largest economy, Asia is certain to play an increasingly important role in the balance of the global economy.

See also: Commodity Chains; Development Theory in Geography; Development and the State; East Asian Studies: Culture; East Asian Studies: Politics; Industrialization; Urban Geography; World Systems Theory.

Bibliography

Ahn, S., 2000. Reconsidering the Asian financial crisis: a comparative study of Korea, Thailand, and Indonesia. East Asian Review 12 (2), 69–86.

Amsden, A.H., 1989. Asia's Next Giant: Korea and Late Industrialization. Oxford University Press, New York.

Aoki, M., Hyung-ki, K., Okino-Fujiwara, M., 1997. The Role of Government in East Asian Economic Development. Cambridge University Press, Cambridge, UK.
Bhagwati, J., 1988. Protectionism. MIT Press, Cambridge, MA.
Castells, 1994. Technopoles of the World: Making the Twenty-First Century Industrial Complexes. Routledge, New York.
Clark, C., Roy, K.C., 1997. Comparing Development Patterns in Asia. Lynne Reinner, Boulder, CO.
Fields, K.J., 1995. Enterprise and the State in Korea and Taiwan. Cornell University Press, Ithaca, NY.
Gold, T., 1985. The State and Development in the Taiwan Miracle. M E Sharpe, New York.
Kornai, J., 1990. The Road to a Free Economy: Shifting from Socialist System: The Example of Hungary. Norton, New York.
Kornai, J., 1994. The place of the soft budget constraint syndrome in economic theory. Journal of Comparative Economics 26 (1), 11–17.
Krugman, P., 1998. Saving Asia: it's time to get radical. Fortune 138 (5), 74–80.
Lal, D., 1985. The Poverty of Development Economics. Harvard University Press, Cambridge, MA.
Naughton, B., 1995. Growing Out of the Plan: Chinese Economic Reform 1978–1993. Cambridge University Press, New York.
Pei, M., 1994. From Reform to Revolution: The Demise of Communism in China and the Soviet Union. Harvard University Press, Cambridge, MA.
Rabushka, A., 1979. Hong Kong: A Study in Economic Freedom. University of Chicago Press, Chicago.
Rawski, T.G., 1994. Progress without privatization: the reform of China's state industries. In: Milor, V. (Ed.), Changing Political Economies: Privatization in Post-Communist and Reforming Communist States. Lynn Reinner, Boulder, CO.
Sachs, J.D., Wing Thye, W., 1994. Experiences in the transition to a market economy. Journal of Comparative Economics 18 (3), 271–275.
Wade, R., 1990. Governing the Market: Economic Theory and the Role of Government in East Asian Industrialization. Princeton University Press, Princeton, NJ.
Walder, A.G., 1986. Communist Neo-traditionalism: Work and Authority in Chinese Industry. University of California Press, Berkeley, CA.
Walder, A.G., 1995. Local governments as industrial firms: an organizational analysis of China's transitional economy. American Journal of Sociology 101, 263–301.
Whyte, Parish, 1984. Urban Life in Contemporary China. University of Chicago Press, Chicago.
World Bank, 1993. The East Asian Miracle. Oxford University Press, Oxford, UK.
Yang, D.L., 1991. China adjusts to the world economy: the political economy of China's coastal development strategy. Pacific Affairs 64, 42–64.

East Asian Studies: Gender

Tani E Barlow, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 3962–3967 © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Gender, like all ideological terms (e.g., civil society), is political policy and an analytic category. Here gender in East Asian studies will be understood historically. A concern of anthropologists, sociologists, and political economists since the 1970s, the use of a gender analytics was contested when it entered East Asian area studies in the 1990s Conceptually (and particularly after several historians refined and popularized the term), gender offered a social explanation for cultural variation in sexual and social practices previously either attributed to anatomical difference or simply ignored. It also offered a mode of general cultural analysis. Approximately when gender was accepted into area studies, 'East Asia,' an artifact of Cold War area studies in the United States, began to be questioned. In post-1989 re-regionalized area studies the preferred scholarly term is 'Asia-Pacific.' Factors contributing to re-regionalization are common: unequally experienced neoliberalist economic globalization processes; intrastate capital development projects; transnational flows of Chinese gendered labor and capital; and, the new global division of labor linking South East, East, and North Asia. While regional states retain sovereign regulatory powers, the national welfare policies that during the twentieth century provoked and circumscribed national gender stereotypes and development policy are changing. In 'new metropolises' that form the skeleton of the Asia-Pacific, neoliberal economic policies still shape national or race, ethnic, gender-coded citizens. However, driving ideologies of gendered social difference now are commodified, consumption-oriented, lifestyles. New policies, practices, and critical theories in turn are reshaping how area scholars and government officials employ the gender analytics now.

The Concept of 'Gender'

Gender is an analytic term that came into common academic usage in the late 1970s and early 1980s in the United States among feminist social scientists. The widely influential historian Joan Scott urged historians to use the term gender to describe women's social and historical differences from men positively (Scott, 1986). By the early 1990s, debate over the gender concept had widely expanded. Judith Butler's theoretical work (1990) established gender to mean the cultural practices that make sexual natures of all kinds appear to be fundamental, heterosexual, and unchanging. Simultaneously a publishing boom of high-profile volumes began that, over the 1990s, has firmly secured the notion of gender into questions of national or political difference. Other volumes (e.g., Sangari and Vaid, 1990; Gilmartin et al., 1994) have linked the gender concept to the problem of colonialism in what have become the fields of postcolonial studies, comparative feminism, and gendered area studies.

The Policies of Gender

Gender is modern political policy. Presently it forms an element of the post-Cold War global governance goals disseminated through the United Nations, so-called nongovernmental organizations, and donor or civil society organizations. This post-1989 neoliberal gender policy is juridical. It argues that all states should adopt the position that women's rights are universal human rights. Gender as state policy, however, is older and more historical than current rhetoric would have it. In the twentieth century, industrialized and late industrializing states have cast citizenship in gendered terms.

Contemporary Gender Policy, Global Governance, Women's Human Rights, 1979–

The importance of gender as policy emerged in the initiatives for global governance over the last quarter of the twentieth century. It is an aspect of the international human rights law movement. Its cornerstone is the Convention on the Elimination of All Forms of Discrimination against Women, which seeks to abolish 'gender-based violence' and is a 1979 revision of the 1948 Universal Declaration of Human Rights. The United Nations, donor agencies (like the Ford Foundation), and a newly arisen civil society sector began advancing gender-sensitive programs during the United Nations Decade for Women (1975–85) and during subsequent United Nations Conferences on Women, culminating in the 1995 Beijing Fourth World Conference on Women and its resulting 'Platform for Action.' As a whole, these institutions and the World Bank's investments promote the view that gendered difference is an important axis of inequality. Many advocates in the United States blame non-Western states for gendered inequality. They seek to use gender as a development index to force perceived noncompliant states to redress the relative poverty and vulnerability of female citizens. Loans, investment, education, and outreach into semiofficial social service agencies are the instruments of global governance. Though a lack of clarity prevails about whether these attempts are effective or a US form of triumphalism, Charlotte Bunch and Hillary Clinton are the movement's best-known leaders.

Historical Gender Policies of Modernizing States in 'East Asia': The 1840s to the 1980s

The novelty of contemporary uses of the gender analytics tends to obscure the longevity of gender modernization policies

throughout the Asia-Pacific. Scholars working in the framework of the 'East Asian' area studies during the 1980s and 1990s have established that much (Barlow, 1996). It can be said that colonialism itself advanced the rationalization of gender by bureaucratizing social modernity or (in some cases) by directly exploiting the bodies of women. During the mid-nineteenth and twentieth centuries, the forces of imperialism, colonialism, and subimperialism propelled late industrializing states toward expeditious efforts of political modernization and capital accumulation. States pursued this objective in various ways, although interstate relations meant that events in one state (e.g., Japanese imperialism in Korea, Manchuria, Taiwan) contributed to the shape of gender policy in other states.

Colonial Gender Policy in Modern Korea

Between the 1870s, when its traders arrived in Korea's overwhelmingly agricultural economy, and 1945, when the Japanese colonial administration evacuated and the neocolonial United States occupation began, the peninsula took on the classic colonial configuration of a political economy twisted to serve Japanese imperial expansionism. Japan's colonial authorities reordered Korea's rural economy, instituting modern concepts of ownership to expedite taxation. The new system placed tax gathering and relations of production in the hands of a semicolonial and semifeudal native elite, 'landlordizing property rights.' This industry led to a widespread shift from peasantry to tenancy as the primary relation of production. Agricultural productivity rose, but so did rural immiseration. Poverty, occupation, and assimilation policies flowered into a widespread licensed military prostitution industry (see Choi, 1997). This, and particularly Japanese family law, burdened Korean gender and class relations and helped to shape a male-centered style of nationalist anticolonial struggle, involving women and men of all classes (see Park, 1998; Fujime, 1997).

In divided South Korea after the Pacific War, a state-compelled development that Alice Amsden calls 'late-industrializing' (and argues is a global model) reconfigured Korean social relations of gender (Amsden, 1989). The incorporated Cold War state inherited a Europeanized, Japanese, colonial civil code. As Hyunnah Yang has shown, along with the modern legal regulation of domestic relations the civil code supported an imaginary 'Korean tradition' that naturalized female subordination. Consolidation of Japanese domestic civil statutes and modern family law into Korean jurisprudence in the Korean Republic (during and after the partition of the nation) also ratified a gendered modernity, heavily weighted to the nuclear, patrilineal, small family unit. The now constitutional doctrine of patrilineality triggered waves of feminist intervention into state-sanctioned subordination of women and women's labor power under the US-supported military regime (Yang, 1998). Feminist and labor movement successes have led to revisions of the family law in 1962, 1977, and 1989. However, the fact remains that the modern state – more perhaps than the dynastic regime – consolidated domestic property, family jurisdiction, domestic labor, and legal claims in the hands of the patriline and restricted the claims of female kin. Little is known of the gendered state policy of North Korea, but it appears to be modeled on that of the People's Republic of China (PRC) during the Cultural Revolution (1966–76).

Colonial Modernity, Imperialist Economy, National Gender Policy in Modern Japan

Historians are agreed that the Meiji state (1868–1912) was preoccupied with gender policies from the outset of its drive to modernize Japan based on the European model (under threat of Western imperialism) (see Bernhardt, 1991; Sievers, 1983). The Meiji civil code of 1898 ideologically reformulated the patriarchal family and its relation to the state, emphasizing the political obligation of scientific housewifery, or 'good wife, wise motherhood.' Severe restrictions on women's political rights were coupled with the statist argument that the labor of women should be considered a civil service to the nation. Deliberate gender policies like these persisted, toward varying ends, throughout the twentieth century. First, the new state-sponsored industrialization policy willfully and strategically regendered the household division of labor by moving men out of domestic production into heavy industry. Second, policy affecting men and children led to the official drive in the late nineteenth century to 'reshape conceptions of womanhood' (Uno, 1991). Fostering smaller family formations, female education, young women's heavy participation in waged light industrial production, instituted bans on female public political participation, as well as campaigns encouraging feminine virtues in household and reproductive labor affected growing sectors of the national population. A 20-year effort to bureaucratize the policy on women culminated in the 1898 implementation of Japan's prewar civil code.

Capital accumulation through industrial production and constitutional state formation in Japan rested on colonial expansion onto the Asia mainland, into Korea and Taiwan, and eventually throughout the Asian Pacific. Despite its best efforts, the state had never been able to contain female wage-workers from moving into the labor market. As militarization and colonial administration drained off increasing numbers of men into service, state policy began gendering the notion of national sacrifice. Men were to provide military or colonial service. Militarized policy on women sought to rework the family into a feminine sector of the public sphere. The onset of glorification of national motherhood is datable to this era. It resurfaced (in altered form) in immediate postwar policies.

After 1945, state policy sought to emphasize domestic reproduction and superexploit male labor in an economic drive to recover Japan's regional supremacy under the umbrella of a US-dominated Cold War containment policy. An adjunct to this was an effort to limit female participation in Japan's national labor market while increasing (after 1956) women's responsibility to undertake domestic consumption responsibilities.

As in the Korean case, the history of gender and the nation sometimes precludes an emphasis on the Japanese women's rights movement, to say nothing of a transnationally influential, Japanese feminist critical theory that arose with colonial modernization and state formation. Nonetheless, as Mariko Asano Tamanoi noted in her study of rural women and nationalist ideology, the term 'national woman' is too simplistic and

does not convey the full import of gendered social change in the twentieth century (Tamanoi, 1998).

Subimperialism, Colonial Modernity, and Policies of Gender Revolution in Modern China

The Chinese record of state-mandated gender policy is intricate. That is due to its enormous size vis-à-vis other East Asia nation-states, its semicolonial status after 1842, a full half-century of multiply sited civil wars, and the high priority which the Chinese Communist Party accorded to women's liberation. As early as the compromised Republican era (1911–49), the state was initiating gender and development policies that were later absorbed into influential party state policies associated with Maoism and the revolutionary mobilization policies of the PRC.

National gender policy first appeared in the form of state-mandated reforms of education, female dress, and foot-binding practices. A juridical modernization indebted to both continental European and Japanese models was instituted in the 1929–31 modern civil code and, under the Nationalist or Republican Party state (1911–49), mobilized women into participatory forms of citizenship. A further link exists between the weakly instituted republican civil code (organized around a liberal divorce law) and campaigns during the 1940s and 1950s to propagate the Chinese Communist Party's Marriage Law (Bernhardt, 1994).

The ethos that infused revolutionary gender ideology came from the early twentieth-century Chinese Enlightenment: an anticolonial, urban, nationalist movement led by modern intellectuals. The Chinese Communist movement, which incorporated feminist themes, was – relative to contemporary states – hospitable to policies that institutionalized women's liberation. But, like other states, women's liberation became an index of national modernity, particularly, perhaps, under Chinese Communism. What most distinguished the gender policy of the early Chinese communist states was their focus on mobilization as a means of institutionalizing political policy and their organization of gender revolution around a Marriage Law that put the conjugal couple at the center of jurisprudence (Gilmartin, 1995).

Land reform and class war in the late 1940s and early 1950s and the state policy of rural communization in the early 1960s were no doubt two of the most transformative revolutions of the twentieth century. These revolutions consolidated the gains of Chinese state feminism by placing laws on the books that governed the social labor of women and granted women the rights to own (however briefly) land in their own names (Johnson, 1983). Socialist development policies, which were aimed at consolidation of state power and primitive accumulation for investment in late, socialist industrialization, rested on the nuclearization of family resources in the urban areas. They also proletarianized women and periodically used female labor as a reserve pool. The mobilization of peasant women into agricultural production furthered primitive capital accumulation as well as mechanisms to extract rural capital gains for use in urban development. Nonetheless, recent studies point to the ways that particularly poor peasants took advantage of the modern state institutions to further personal goals, particularly women seeking divorces. The socialist state's welfare policies had a wide-reaching impact on several issues: the gendering of labor, domesticity, sexuality, and demography; civil rights and access to public space; and avenues of upward political mobility.

Economic Liberalization, the Displacement of 'East Asia' as a Region, Gender Scholarship, and Transnational Flexible Economies

Just as historically 'modernization' is usually linked to Europe and the United States in the nineteenth and twentieth centuries, neoliberal 'globalization' is linked to Asia-Pacific in the last third of the twentieth and probably much of the twenty-first century. Asia-Pacific is a laboratory of post-Fordist flexible capital accumulation inside a post-Cold War international state system. The Asian Pacific roots of this neoliberal globalization lie in:

1. the Chinese economic reforms of the 1970s;
2. the hyperdevelopment of newly industrialized countries and city-states (Taiwan, South Korea, Singapore, Hong Kong);
3. Japanese capital investment in Malaysia and other South East Asian countries;
4. the rise of flexible Chinese diasporic capitalism; and
5. US-sponsored neoliberal policies of the World Bank and the General Agreement on Tariffs and Trade system.

Within intra-Asian political economies, in so-called transnational gendered labor flows, and outside it in interregional global trade, policy elites in Asia-Pacific states and national corporations are scaling back modernization strategies based on a state welfare logic (see Higgott, 1999; Ong and Nonini, 1997).

No longer central to development are the projects that had earlier supported the public ideology of a state- or nation-typical, modern woman, and domestic relations, although the use of maids among the middle class is stabilizing the elite family structure in some regions. It is premature to speculate what specific gendered configurations in subregions and specific locales will emerge in the future. However, gender is becoming a central concept in writing throughout the Asia-Pacific region and the United States. Scholars globally seek to explain how liberalization of economic life affects populations that are now defined in relation to their consumption power or their capacity to be absorbed into the flexible service sector (e.g., tourism, prostitution, travel) as well as their labor power. Heightening class differentials separating consumers and service providers, surging global inequality explained in cultural terms, ethnic and race-coded discriminations, systematic preference accorded flexibility in labor migration, and capital investment patterns across statelines (but within regional cores) are all having an impact on the way that the concept 'gender' (in Chinese usage 'social gender' or *shehui xingbie*) will be employed. Transnational corporations now recruit and market with the concept in mind. New social movement policy strategists, networks of internationalist progressive scholars, US feminist human rights lawyers, World Bank lending policymakers, United Nations monitors of prostituted labor and infectious

diseases: all are incorporating the policy and analytics of 'gender' into their thinking.

Particularly since the finance capital crisis of 1997–98, it has been increasingly apparent to observers that a new regional production of scholarship is on the rise. Produced by diasporic scholars from Asian locales resettled in the metropole, or writing from 'new metropolises' of Singapore, Hong Kong, Seoul, Jakarta, Tokyo, Beijing, and Sydney, 'conscientized' scholars in the First World, and expatriate Asians in the white nations of Europe and the United States, this scholarship is eroding the boundaries of area studies, properly speaking. Whether we are speaking of scholarship on gender inequality which Women's Federation scholars in Beijing produce in Chinese for national circulation, or studies produced in Tokyo and Seoul in English for international dissemination, or the steady translation of mainline US feminist and gender theory to a wide number of gender studies institutes throughout Asia, the full impact of localized scholarship on US-oriented area studies has yet to be registered. The trend toward globalized scholarship is exemplified in the work of the journal *Positions: East Asia Cultures Critique*, which publishes research that questions strict national divisions on which conventional area studies have rested and also injects the gender analytics into scholarship. The appearance of pan-Asian or regional cultural studies journals like *Inter-Asian Cultural Studies* (Taiwan), the *Asian Journal of Women's Studies* (Korea), and many others is also blurring the edges of area studies and area-oriented gender studies.

The effort to grasp the implications over the years 1970–2000 of re-regionalization (as well as the rise of flexible capital and gendered consumer markets) is producing significant lines of innovative global scholarship. The concepts of 'transnational Chinese capitalism' and 'flexible citizenship' allow Aihwa Ong, for instance, to propose the emergence of ethnicized capital flows which are not nationally grounded and which rest on the persistent powers of the state, opportunistically taking advantage of state forms (Ong, 1999). Modern Chinese transnationalism, its distinctive strategies of capital accumulation, and gender analytics are all being studied empirically. Specialists consider gender as a significant axis among other axes, such as ethnicity, nationality, sexuality, racialization, and caste status. A special issue of the journal *Positions: East Asia Cultures Critique* (1999) on 'Asian Transnationalisms' has highlighted these dynamics in relation to ethnicity and sexuality.

Advocates and critics of liberal globalization alike recognize the enlarged opportunities for public sexual expression opening up in the new metropolises of Pacific Asia. Sexualized and commodified styles of self-presentation open fresh possibilities for new social movements of minority social groups, such as self-styled queers (Ho, 1998). Minority sexual communities – gay, bisexual, nonprocreative heterosexual – exist in most of the new metropolises. The centrality of the gender analytics to scholarship about these communities and cultural criticism from within them are very clear in queer social, cultural, cinema, and theoretical studies which presume transnationality and track how sexual expression is gendered among specific groups of queers (see Manalansan, 1994; Hanawa, 1994). It is premature to consider how crosscurrents between queer criticism and a retrenched or renaturalized market heterosexuality will play out. One critic has shown (in a field study of young, heterosexual, urban, upwardly mobile, Chinese women) that claims to sexual pleasure have led to enriched lives. Yet, sexual rights have proved consistent with authoritarian forms of governmentality (Erwin, 2000).

Another avenue for innovation in gendered area studies of the re-regionalized Asia-Pacific has been the political economy of ethnic coding in intranational, transnational, and transborder contexts. Diasporic ethnicities such as the Hmong are trans-Asian minorities whose communities are now globalized. Ethnic and national coding have been significant to studies of minority service work in tourism and in gendered, often sexed, labor flows of subaltern national citizens from the Philippines into the boom economies of Taiwan, South China, and Hong Kong. Ethnographers and theorists are pursuing a combined transnational and ethnic focus that promises to clarify the political complexity of the new Asia-Pacific and its new social movements.

Conclusion

The analytics of gender offers increased penetration of markets as easily as it offers progressive scholars another significant index for gauging relative oppression and appropriately representing those who cannot represent themselves. Certainly advertising, cultural markets, and lifestyle expenditure in the ongoing consumer revolution refiguring the region are naturalizing heterosexual gender norms around anatomical sexual difference and the capacity to consume and produce what is appropriate to one's 'nature.' Increasingly feminist critique has taken up the problems endemic in consumer society as well as extending its traditional concern for poverty, antisexism, and freedom from reproductive tyranny and disease. Undoubtedly, scholarship will look increasingly at the conjugal pair, partly in order to queer commercial norms and partly to expose the political culture's reliance on a heterosexual mode of consumption.

Another task for gender analytics in Asia-Pacific area studies is to document the vast and increasing rifts that separate women in the same society, as well as the economic differentials between women and men affecting specific sectors. A related project will be to follow up on Li Xiaojiang's comment (2001) that the struggle for gender equality is a long historical struggle for all sectors and subgroups, classes, and class fragments, and should be studied as such. A further probable area of new research may be the political economy of pan-'Asianism' and the cultural production of women in transition in many of the liberalized state economies (Barlow, 1998). How the new social movements engaged in the Asia-Pacific (as elsewhere) proceed in the years 2000–10 may have some impact on how the gender analytics is employed in state policy making and that, in turn, will affect the problems taken up in academic venues.

Editor's Note

Gender dynamics in the East Asian societies have continued to evolve since Barlow's entry was initially published. Readers interested in this evolution will be well served by: Esther

Ngan-Ling Chow (2002); Marianne H. Marchand and Anne Sisson Runyan (2011); and Malcolm Warner (2013).

See also: East Asian Studies: Culture; East Asian Studies: Politics; Nationalism: General; Nations and Nation-States in History.

Bibliography

Amsden, A., 1989. Asia's Next Giant: South Korea and Late Industrialization. Oxford University Press, New York.

Barlow, T.E., 1996. Theorizing woman: Funü, Guojia, Jiating. In: Scott, J. (Ed.), Feminism and History. Oxford University Press, Oxford, UK.

Barlow, T.E., 1998. Green blade in the act of being grazed: late capital, flexible bodies, critical intelligibility. Differences 10 (3), 119–158.

Bernhardt, G. (Ed.), 1991. Recreating Japanese women, 1600–1945. University of California Press, Berkeley, CA.

Bernhardt, K., 1994. Women and the law: divorce in the republican period. In: Bernhard, K., Huang, P., Allee, M. (Eds.), Civil Law in Qing and Republican China. Stanford University Press, Stanford, CA.

Butler, J., 1990. Gender Trouble: Feminisms and the Subversions of Identity. Routledge, New York.

Choi, C. (Ed.), 1997. The Comfort Women: Colonialism, War, and Sex. Special Issue of Positions: East Asia Cultures Critique 5 (1).

Chow, Esther Ngan-Ling, 2002. Transforming Gender and Development in East Asia. Routledge, London. Updated version published 2013.

Erwin, K., 2000. Liberating Sex, Mobilizing Virtue: Gender, Family Values, and Transnational Desires in Shanghai, China. Unpublished dissertation. University of California, Berkeley, CA.

Fujime, Y., 1997. The licensed prostitution system and the prostitution abolition movement in modern Japan. In: Choi, C. (Ed.), Special Issue of Positions: East Asia Cultures Critique, vol. 5, pp. 135–170 (1).

Gilmartin, C.K., 1995. Engendering the Chinese Revolution: Radical Women, Communist Politics, and Mass Movements in the 1920s. University of California Press, Berkeley, CA.

Gilmartin, C.K., Hershatter, G., Rofel, L., White, T. (Eds.), 1994. Engendering China: Women, Culture, and the State. Harvard University Press, Cambridge, MA.

Grewal, I., Gupta, A., Ong, A. (Eds.), 1999. Asian Transnationalities. Special Issue of Positions: East Asia Cultures Critique 7 (3).

Hanawa, Y. (Ed.), 1994. Circuits of Desire. Special Issue of Positions: East Asia Cultures Critique 2 (1).

Higgott, R., 1999. The political economy of globalization in East Asia: the salience of 'region building'. In: Olds, K., Dicken, P., Kelly, P.F., Kong, L., Yeung, H.W. (Eds.), Globalization and the Asia Pacific: Contested Territories. Routledge, London.

Ho, J., 1998. Kuier lilun yu zhengzhi. Special issue on queer politics and queer theory. Taiwan: Working Papers in Gender/Sexuality Studies Zhongli guoli daxueh. Zhongli.

Johnson, K.A., 1983. Women, the Family, and Peasant Revolution in China. Chicago University Press, Chicago.

Li, X., 2001. From modernization to globalization: where are Chinese women? (T. Barlow, Trans. Special Issue on Gender and Globalization). Signs.

Manalansan IV, M., 1994. (Dis)Orienting the body: locating symbolic resistance among Filipino gay men. Positions: East Asia Cultures Critique 2 (1), 73–90.

Mani, L., 1990. Contentious traditions: the debate on sati in colonial India. In: Sangari, K., Vaid, S. (Eds.), Recasting Women: Essays in Indian Colonial History. Rutgers University Press, New Brunswick, NJ.

Marchand, Marianne H., Runyan, Anne Sisson, 2011. Gender and Global Restructuring: Sightings, Sites and Resistances, second ed. Routledge, London.

Ong, A., 1999. Flexible Citizenship: The Cultural Logics of Transnationality. Duke University Press, Durham, NC.

Ong, A., Nonini, D.M. (Eds.), 1997. Ungrounded Empires: The Cultural Politics of Modern Chinese Transnationalism. Routledge Press, New York.

Park, H.O., 1998. Ideals of liberation: Korean women in Manchuria. In: Kim, E., Choi, C. (Eds.), Dangerous Women: Gender and Korean Nationalism. Routledge Press, New York.

Sangari, K., Vaid, S. (Eds.), 1990. Recasting Women: Essays in Indian Colonial History. Rutgers University Press, New Brunswick, NJ.

Scott, J.W., 1986. Gender: a useful category of historical analysis. American Historical Review 91, 1053–1076.

Sievers, S.L., 1983. Flowers in Salt: The Beginnings of Feminist Consciousness in Modern Japan. Stanford University Press, Stanford, CA.

Tamanoi, M.A., 1998. Under the Shadow of Nationalism: Politics and Poetics of Rural Japanese Women. University of Hawaii Press, Honolulu, HI.

Uno, K., 1991. Women and changes in the household division of labor. In: Bernstein, G. (Ed.), Recreating Japanese Women, 1600–1945. University of California Press, Berkeley, CA.

Warner, Malcolm (Ed.), 2013. Managing Across Diverse Cultures in East Asia: Issues and Challenges in a Changing Globalized World. Routledge, Oxford.

Yang, H., 1998. Envisioning Feminist Jurisprudence in Korean Family Law. Unpublished dissertation. New School for Social Research, New York.

East Asian Studies: Politics

Haruhiro Fukui, University of California, Santa Barbara, Goleta, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Post–World War II studies of politics in East Asian nations have followed distinct patterns in the Cold War and post–Cold War periods. This article tracks and critiques major works published, mostly in English, in each period on politics in the People's Republic of China (hereafter China), the Republic of Korea (hereafter South Korea), the Democratic People's Republic of Korea (hereafter North Korea), and Japan. Studies of politics in Taiwan and Hong Kong are discussed in the Chinese sections and politics in North Korea in the Korean sections. An author is referred to as Chinese, Korean, Japanese, etc. on the basis of their surname, regardless of their present citizenship status or place of residence.

Studies of Politics in East Asia in the Cold War Period

This was a period when Japan was the center of media and academic attention in East Asia, first for its unprecedented position as a defeated, devastated, and occupied nation, then for its phoenixlike rise from the ashes of war as an extraordinarily fast-growing economy and an ultrastable polity. South Korean politics attracted nearly as much domestic and foreign media and academic attention during the period, but for a very different reason: it was extremely unstable, and often violent, under the rule of a succession of highly authoritarian governments. North Korea under Kim Il Sung's dictatorship was completely closed and inaccessible to foreign media and scholars, and non-Marxist social science did not exist within the nation. China was almost as inaccessible to foreigners until the early 1980s, a situation that allowed little objective and reliable research on its politics, especially on events played out at the upper levels of the national government.

Most studies of this period shared several interesting characteristics. They were, by and large, instigated or inspired by current headline-making political events, heavily influenced by the empiricist tradition of American political science, historical and descriptive rather than explicitly theory guided, and, until the late 1980s, almost exclusively focused on politics in one nation.

Studies of Japanese Politics

This was a period when Japan rose from the ruins left by a devastating war that had ended with the destruction of nearly all of its major cities by a half-year of relentless air raids, the last two by the first and last atomic bombs ever used in war. After a decade of dire economic hardships and political instability, the majority of which under the rule by Allied Occupation, the nation reemerged by the mid-1950s as one of the most dynamic and fast-growing economies and the most stable polities in the postwar world under seemingly permanent rule by a powerful conservative party formed in 1955, the Liberal Democratic Party (LDP). By the end of the 1960s, the nation had become the second largest national economy, next to the United States, and, while rocked from time to time by foreign policy disputes, corruption scandals, and external economic and political shocks, enjoyed sustained prosperity and political stability throughout the period. It commanded far more global media and scholarly attention and interest than either China or South Korea, not to mention North Korea, in this period.

Under the United States–led Allied Occupation for the first seven postwar years, Japan was also by far the most accessible nation to foreign, especially American, scholars to study during the early Cold War years, and beyond. In 1950, or about 2 years before the Allied Occupation of Japan ended, the University of Michigan Center for Japanese Studies set up a field station in Okayama Prefecture in central Japan for interdisciplinary research on changes under way in the nation's local society and politics. The 5-year work by the station (1950–55) resulted in the publication of *Village Japan* in 1959, a pioneering work that paved the way for a series of a half-dozen edited volumes published by Princeton University Press in the mid-1960s to the early 1970s under the general rubric of the modern and modernizing Japan. These works, along with other works by American, and a few British, scholars on the revision of the imperial Japanese constitution, purge of war criminals, land reform, and the 1960 US–Japan Security Treaty crisis inspired many studies of Japanese politics by Japanese scholars.

There was a substantial community of social scientists in prewar Japan, with its origins dating back to the late nineteenth century, and a group of them founded in 1948 the Japanese Political Science Association (JPSA) modeled on the American Political Science Association (APSA). Investigation of the causes and consequences of prewar and wartime Japanese 'fascism' was the main concern of its early members. Maruyama Masao's work (1963) exemplifies this line of investigation. Under the influence of American political science, however, the Japanese scholars' interest quickly expanded to include a wider range of themes and topics, such as the Japanese government bureaucracy, parliamentary and local elections, political parties, interest groups, public opinion, and social movements. The first major fruit of these efforts was a symposium volume published in 1958 under the title *Gendai nihon no seiji katei* (political processes in contemporary Japan).

The influence of American scholarship was particularly critical to works on Japanese elections and voting behavior. The founding of the Japanese Association of Electoral Studies in 1981 was inspired importantly by the survey-centered research that characterized electoral studies in the United States. The same was true of the studies of the LDP, the government

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10123-0

bureaucracy, and public policy making, which gained great popularity among scholars in both nations in the 1970s and 1980s. A major landmark in this line of work was the publication of Chalmer Johnson's book on the Japanese industrial and trade policy bureaucracy in 1982. This work triggered an international controversy on the nature and role of 'industrial policy' and the 'developmental state', and inspired a series of case studies on related subjects.

Studies of Korean Politics

For the greater part of the Cold War period, South Korea was, in sharp contrast with Japan, a nation under the rule of a succession of authoritarian and repressive governments led by, first, a fervently nationalist Christian 'Cold Warrior' with a Princeton doctorate in international law, Syngman Rhee (1948–60), followed by a succession of three army generals. Early in the period, the nation was devastated by the Korean War (1950–53) and, throughout the period, periodically rocked by military coups and massive, and often violent and bloody, anti-government revolts. Five constitutions and republics came and went one after another before a peaceful transition process began in the Sixth Republic at the end of the period. It was thus a nation viewed more with distrust and skepticism than with confidence and hope by most foreign scholars. The only bright spot in the generally dreary picture was the gradual, but accelerating, government-guided development of its economy beginning in the early 1960s.

Korea under the 35-year-long Japanese colonial rule and the authoritarian and isolationist regime that had preceded it had not nurtured a viable indigenous academic community. At the end of hostilities in the summer of 1945, however, the nation was placed under US trusteeship and remained closely allied with the US thereafter. As a result, information on South Korean politics was nearly as open and accessible to US-based scholars as information on Japanese politics during the postwar period. Under such circumstances, a Korean Political Science Association, modeled after the APSA like its Japanese counterpart, was founded in 1953 and, two decades later, in 1973, an Association of Korean Political Scientists in North America was formed by US-based scholars.

About halfway through this period, an American scholar published a landmark work on the cultural and institutional factors contributing to the strength of authoritarian centralism in South Korea (Henderson, 1968), followed by a series of works on problems of and prospects for a democratic transition, especially after the 1987 'democratic opening' and the birth of the Sixth Republic. At the end of the period, an American political economist published the first major work on the state-led industrialization of the South Korean economy under way (Amsden, 1989). More modest in number than English-language works on Japanese politics and political economy published in this period, these works nonetheless played a seminal role in the subsequent development of the study of South Korean politics and political economy.

By the late 1980s, a substantial number of studies in Korean (*Hanglu*) by Korean scholars began to be published in South Korea. Most were historical and normative accounts of Japanese colonial rule and postindependence politics under military rule and many were based, directly or indirectly, on Marxist or other radical theories, such as dependency theory. It was not until the 1990s that more American-style empirical political science became the main stream in the study of South Korean politics, or North Korean politics for that matter, by South Korean scholars.

Throughout the Cold War period, little factual information on politics in North Korea was available to scholars outside the nation and little serious research was done by them. A few exceptions were a comparative study of the foreign policy systems of North Korea and South Korea (Koh, 1984) and a political biography of Kim Il Sung (Suh, 1988), both by US-based Korean scholars.

Studies of Chinese Politics

For the better part of the Cold War period, mainland China was almost as inaccessible and factual information on its politics as unavailable to outsiders as North Korea. At the end of World War II, moreover, there were few social scientists anywhere in the world who were seriously interested in, much less knowledgeable about, modern Chinese politics. In the United States, Chinese politics was a serious concern only to a small number of historians, such as John King Fairbank and Karl A. Wittfogel. In Europe, scholarly interest in China was limited largely to a few British and Dutch philologists and historians, commonly known as 'sinologists'. Japan had a tradition of social science research on some aspects of contemporary Chinese society dating back to the 1930s. However, there was no systematic study of prewar or wartime Chinese politics by Japanese scholars of the period.

In the decades following the end of the war, scholarly, as well as general public, interest in Chinese politics grew rapidly, propelled by a series of headline-catching events: the onset of the Cold War, the Great Leap Forward (1958–61), the Cultural Revolution (1966–76), the initiation of multifaceted economic reforms (1978), and the Tiananmen Square incident (1989). In Japan, a group of about 50 scholars of Asian, mainly Chinese, studies founded in 1953 an interdisciplinary Japan Association for Asian Political and Economic Studies, renamed the Japan Association for Asian Studies a few years later.

The Cold War and the Cultural Revolution spawned intense ideological, and often explicitly partisan, debates among Chinese politics specialists around the world. These events, especially the Cultural Revolution, divided nearly all Japanese China specialists into friends and enemies of the Maoist regime and its policies. Even in Europe, where the tradition of nonpolitical sinologist approach was particularly strong, as sharp ideological divisions occurred among local China specialists as among their Japanese counterparts. This was also true of a segment of China specialists in the United States. A good portion of American scholarship, however, was deliberately nonideological and analytical. Some works applied concepts associated with modernization theory or elite and faction politics models. These works helped to lay conceptual foundations for studies of Chinese politics to follow in the late 1970s and 1980s, a period known for Deng Xiaoping's Reform regime.

Chinese political culture and elite and faction politics initially drew most attention among US-based scholars. However, the range of subjects and themes for investigation

gradually expanded to include local and minority politics, rural industrialization, special economic zones, and the military. Some scholars began to borrow terms and concepts from works on the contemporaneous political developments in Eastern Europe, such as civil society and public sphere. This practice paved the way for more explicit and systematic comparisons of political and economic reforms in China and Eastern Europe in the post–Cold War period. Many of the changes in American scholars' works were replicated in those of Japanese and European scholars. Above all, the heavily ideological and partisan overtones that characterized much of the work done in the heyday of the Cold War gradually dissipated toward the end of the period, as more and more foreign social scientists began not only to visit China for fieldwork, but also to collaborate with Chinese scholars.

The most significant development during this period was, however, the dramatic change in the way Chinese intellectuals thought about and discussed Chinese politics. This was an integral part of the sweeping transformation of the intellectual environment of Chinese society, especially in Beijing and other large cities. By the late 1980s, proreform, prodemocracy scholars, writers, and journalists began to form clubs, salons, and seminars and publish political journals and books (Gu, 1999). Translations of Western social science works by, for example, Max Weber, Robert Merton, Kenneth Arrow, Cyril Black, and Alex Inkeles, were published and apparently read fairly widely. While such intense intellectual and cultural fermentation was under way within China itself, US-based Chinese political scientists founded a Chinese Scholars of Political Science and International Studies in 1986, to be renamed the Association of Chinese Political Studies in 1995. The process of the democratic transition in China, however, was abruptly interrupted in 1989 by the Tiananmen Square incident. The event generated a spate of studies on the limitations of the incipient democratic polity and civil society that had emerged under the Reform regime. In the United States alone, some two dozen books were published on the subject in the decade immediately following the event.

The glare on the series of headline-grabbing events in mainland China left 'the other China', Taiwan, nearly totally ignored by scholars outside Taiwan itself during the Cold War period. The only notable exception was a modest amount of work on Taiwan's 'economic miracle'. It was not until the mid-1990s that serious scholarly work by foreign scholars on the reform and democratization in Taiwan began to appear in print.

Studies of East Asian Politics in the Post–Cold War Period

The post–Cold War period in East Asia saw dramatic status reversals between Japan and China and, to a lesser extent, South Korea. China's booming economy surpassed and began to overshadow Japan's, while its politics became increasingly open to direct observation and study by foreign media and scholars as travel restrictions were significantly loosened for both Chinese and foreigners. In the first decade of the new millennium, it was not significantly more difficult for a foreign scholar to visit China and study its politics than visit Japan or South Korea and study their politics. South Korea's economy, already a 'high-income developed economy' at the beginning of this period, continued to grow at a more moderate, but steady, pace, while the process of democratic transition and consolidation, also started at the end of the Cold War period, not only continued but accelerated in the new period. In sharp contrast with both countries, Japan alone experienced a seemingly endless recession and lifeless politics in the new period. It was no longer a particularly interesting, much less exciting, subject for the media to cover or a scholar to study.

The reversals of Japan's place on one hand and China's and South Korea's on the other have been accompanied by as starkly divergent trends in the number of their students studying at US universities. According to the United States Institute of International Education (IIE) statistics, the numbers of Chinese students studying at US universities increased fourfold from about 55 000 to 200 000 and South Korean students about 3.5 times from 41 000 to 72 000 between 2000 and 2012, while the number of Japanese students decreased by 60% from 47 000 to 20 000 over the same period. There were thus more than 20 times as many Chinese and 9 times as many South Korean students as Japanese students enrolled at US universities in 2012. On the basis of incomplete data available in reports published by IIE and East Asian governments, we may estimate that about 900 Chinese, 200 South Korean, and 40 Japanese students were enrolled in political science doctoral programs in any recent year.

It is also well known that, upon completion of their study at US universities, the majority of Chinese and Korean students stay on and settle down in the United States, whereas nearly all Japanese students return to Japan. Those hundreds of Chinese and South Korean political science PhDs who continue to live and work in the United States, often as university faculty members, tend to publish their studies in English, adding to the literature accessible to scholars around the world, whereas the few dozen Japanese PhDs and doctoral candidates who return to Japan tend to publish, if they do, only in Japanese, making it virtually impossible for most foreign scholars to access them.

As in the Cold War period, most studies of politics in East Asia in the post–Cold War period were 'event-led' descriptive works. The seminal influence of American political science remained perceptible in most works published in this period. Unlike in the preceding period, however, single-nation studies no longer dominated the field and intraregional and extraregional comparative studies of a variety of political phenomena and themes became increasingly commonplace.

Studies of Chinese Politics

Riding the sustained economic boom triggered by Deng's reforms in the late 1970s, China surpassed and replaced Japan as the world's most dynamic and second largest national economy in the post–Cold War period. At the same time, the marketization of the economy and, to a lesser extent, liberalization of political control gave rise and increasingly drew public and media attention to a number of new political issues, such as the increasing income and wealth inequalities, deterioration of the environment, widespread corruption among politicians, bureaucrats, and even the military, and increasingly vocal grievances of several minority groups, such as Tibetans, Uighurs, and Mongols, against the dominant majority Han

Chinese. Many of these issues then became the main subjects and themes for scholars, both within and outside China, to research and write about, as well as for Chinese governments and policymakers at all levels to tackle and try to resolve, with varying degrees of success.

In an increasingly open China of the post–Cold War period, foreign scholars had and availed themselves of opportunities to visit and engage in fieldwork with little difficulty. At the same time, Chinese governments at all levels began to publish a substantial amount of factual and detailed information and data on a widening range of domestic politics and policy issues, especially after an ambitious program of across-the-board administrative reforms was launched in 1998. There was also a corresponding rise in the amount of factual information, including results of sophisticated public opinion surveys, and editorial commentary on domestic and foreign policy issues purveyed by increasingly independent and articulate Chinese national, provincial, and local print, and to a lesser extent electronic, media. Aided by these developments, research and publications on a widening range of subjects became substantially more data-rich and analytically sophisticated, especially on local and minority politics and the military and security policy.

The thematic focus of domestic and foreign media and academic attention in this period was the evolving state and prospects of the 'Reform regime' and democratic transition. More than a dozen books in English on related subjects were published by US-based scholars alone in the first two decades of the period. Integral to this body of research was a smaller but growing literature on the role of public opinion and the mass media. Nearly as many books and articles were published on the 'marketization' of the Chinese economy, a subject closely related to the political reform and democratic transition.

A conspicuous and intriguing aspect of the study of Chinese politics in the post-Cold War was the growing presence of Chinese scholars based outside China, predominantly in the United States, and publishing the results of their research in English for an international audience. They account for roughly half of the authors of the relevant works reviewed for this article. This resulted partly from the rise of a generation of Chinese scholars born abroad in English-speaking nations, especially the United States, and partly, as suggested already, from the settlement in such nations of a substantial number of Chinese graduate students from China upon completion of their studies at US universities. The latter phenomenon reflects the state of political science, especially the study of Chinese politics, at Chinese universities and research institutions in particular.

With a view to making their institutions of higher education 'first-rate' by world standards, China has made enormous investments since the early 1990s in the construction of new universities, renovation and expansion of old ones, and improvement of the quality of both teaching and research conducted at all universities. Today, most of China's leading universities, such as Peking, Qinghua, and Fudan, offer political science as a field of specialization both at the undergraduate and graduate levels.

These developments, however, have not prevented a huge outflow of students to foreign, especially US, universities, nor many students' permanent settlement abroad. According to most estimates, only a minority of the Chinese students enrolled at US universities in recent years returned to China after graduation and virtually none of those whose study abroad was privately funded did so. For many political science students, this was probably due, at least partly, to the state and environment of higher education in China, where the tradition of Marxist political economy still remains strong and the most important requirement for appointment to a faculty position is said to be not a doctorate but membership of the Chinese Communist Party. According to surveys conducted by the IIE, virtually all East Asian students in the United States were attracted by 'new ways of thinking' they found at US universities.

An unintended consequence of this phenomenon was, as noted already, a tangible increase in the number of Chinese scholars of Chinese politics based abroad and publishing the results of their research in English. While China-based scholars published increasingly sophisticated works in Chinese, the English-language literature remained the main repository of advanced scholarship in the field. In this respect, the state of the study of Chinese politics was somewhat similar to the case of South Korean politics, but sharply contrasted with that of Japanese politics.

Politics in Taiwan continued to be largely ignored by foreign scholars in this period, but a few noteworthy works were published in English on its party politics and political economy (Fell, 2005; Clark and Tan, 2012).

Studies of Korean Politics

In South Korea, the democratic transition initiated toward the end of the Cold War period moved forward at a breathtaking speed. Free and fair presidential elections fought between civilian candidates, mostly of moderate or liberal persuasions, became routine, and so did free and competitive national and local legislative assembly elections. Civil society, effectively suppressed and silenced under the Japanese colonial rule and the postwar Korean authoritarian regime, resurfaced and began to play a key role in the accelerating process of democratic transition and consolidation. In the 2000 National Assembly election, a coalition of 460 or so civic organizations launched a nonpartisan 'throw out the rascals' campaign against corrupt candidates. The presidential election the next year elected an activist left-wing human rights lawyer, a development unthinkable only a decade before.

The sustained growth of South Korea's economy since the 1960s continued and, by the beginning of this period, the nation had grown into a 'high-income developed country' and a member of the Organization of Economic Cooperation and Development, as well as the United Nations. It was hit hard by the 1997 Asian financial crisis, to the extent that it was placed under the International Monetary Fund's temporary trusteeship, but recovered from the shock and was back on track within a few years. As in China in this period and Japan in the previous period, however, the rapid economic growth was accompanied by the proliferation of corruption scandals involving prominent politicians and bureaucrats and sensational media reports on them. In 2001, two of President Kim Dae-jung's sons were arrested on bribery charges.

Throughout this period, democratic transition remained the central thematic concern for most Korean politics scholars within and outside the nation with increasing attention paid to the role of civil society. The political economy was another popular subject of academic investigation and discussion. Amsden's pioneering work published at the end of the Cold War period was now followed by several others, most of them developing or modifying one or more of Amsden's arguments and conclusions in light of the 1997 financial crisis. The scope of the investigation and discussion extended also to other related issues, such as the politics of social welfare, environmental protection, and public health.

As in the case of the study of Chinese politics, a notable development in the study of South Korean politics in this period was the emergence of a substantial number of Korean specialists of Korean politics trained, directly or indirectly, in the tradition of American political science. According to the results of a 1995 Korean Association of International Studies survey of its 1300 or so members, nearly half of them had earned their doctorates at US universities and about one-third at Korean universities. A 2011 Korean Research Foundation survey of 2050 or so political science faculty members at South Korean universities, however, found a little less than one-quarter of them with doctorates earned at US universities, a little over half with degrees earned at Korean universities, and the remaining quarter with degrees from universities in other foreign countries, predominantly in Europe. This suggests a substantial decline in the proportion of US-trained political scientists and a corresponding increase in that of homegrown political scientists teaching at South Korean universities in recent years. The significance of this change, however, is not as great as it may appear, because today's South Korean universities in general and their graduate programs in particular are closely modeled on and replicate their US counterparts, especially with respect to the requirements for the attainment of a doctorate. It should be also noted that, as in the United States, all applicants for a full-time faculty position at a South Korean university are required to have a doctorate at the time of application.

An important corollary of this circumstance is the impressive presence of Korean scholars, whether born in Korea or abroad, among the authors of the most widely read recent works in English on South Korean politics. Like their Chinese counterparts, they account for about half of such authors who have published in this period, potentially sharing the fruits of their research and ideas with a much larger number of fellow scholars around the world than they would be if they published only in *Hanglu*.

Scholars' interest, as well as governments' and journalists', in North Korea substantially increased in this period and a fair number of scholarly books and articles in English have been published on the viability of the ruling Kim dynasty after the death of its founder, Kim Il Sung, in 1994, the current state of and future prospects for its nuclear weapons program, and North–South relations and the prospects of reunification. The nation, however, remained as closed to and hidden from outsides' scrutiny as during the Cold War period and research on any aspect of its domestic affairs, but especially politics, continued to be, as the author of a book review commented, a practice "reminiscent of medieval scholasticism or ancient Chinese oracle-bone divining" (C.K. Armstrong 1999 in Gu, 1999: 224).

Studies of Japanese Politics

The end of the Cold War period coincided with the end of the widely reported and admired performance of the postwar Japanese economy. Its annual growth rate, which had registered a stunning 9.1% annual average in the 1956–73 period and a more modest but still respectable 4.1% in the 1974–89 period, fell to a 1% average over the next two decades. Meanwhile, the nation's population steadily aged, the under-15 cohort shrinking from about 20 to 13% of the total and the over-65 cohort swelling from about 10 to 23% between 1990 and 2010. This rapid 'graying' of the population, a combined effect of a falling birthrate and rising longevity, led the Japanese government to increase social welfare and health-care spending, without raising taxes in the face of voters' dogged opposition, and allow the government debt to reach 200% of GDP, by far the highest in the world.

This was also a period when the superstable one-party rule by the LDP was replaced by an extremely unstable two-party-dominant system, a result of the LDP's loss of sustained public support due to recurrent corruption scandals and the economic downturn. In the first few years of the period, the LDP suffered a series of defections by affiliated Diet (parliament) members concerned about the electoral fallouts of the corruption scandals, who formed a rival conservative party, to be later named the Democratic Party of Japan. The new conservative party, in cooperation with other opposition parties, defeated the LDP in the 1993 House of Representatives general election, bringing the one-party regime to an end. This ushered in a period characterized by a series of annual government turnovers, each government leaving all the pressing economic and social problems facing the nation not only unresolved but even worsened.

Japan in the post–Cold War period has therefore been neither a model high-growth economy nor a unique 'uncommon democracy' (Pempel, 1990) worth as intense media and academic curiosity and attention as in the preceding period, except when the nation was hit by a disastrous earthquake and tsunami in the spring of 2011. The focus of foreign media and academic interest shifted increasingly to the far more dynamic and interesting China and South Korea. A significant number of sophisticated studies have nonetheless been published on wide-ranging subjects: elections and voting behavior, parties and party politics, interest groups, the bureaucracy and policy making, environmental protection, national security policy, and the roles and power of the media, civil society, ethnic minorities, and women. The widely reported administrative reforms, especially the privatization of the postal services, undertaken by the last durable and successful LDP government led by a maverick politician, Koizumi Junichiro, during the first half of the 2000s, led to the publication of a series of books and articles on the subject by scholars inside and outside Japan.

In striking contrast with contemporaneous works on Chinese and South Korean politics, an overwhelming majority of studies in Japanese politics published in English in this period were authored by foreign scholars, predominantly

Americans. Japanese authors contributed only about one-third of the 40 so notable English-language books published in the first two decades, including several coauthored by one or more Japanese and foreign scholars. This does not mean, however, few Japanese scholars study and write on Japanese politics, but that most Japanese scholars write and publish not in English but in Japanese. This situation corresponds with the fact that, as already pointed out, a relatively small and steadily diminishing number of Japanese graduate students attend foreign, especially US, universities, a phenomenon attributable mainly to the availability of a much easier and inexpensive alternative to earn university degrees, including doctorates, at Japanese universities.

Until a few decades ago, Japanese universities, steeped in European educational tradition, regarded and treated a doctorate as an honor to be conferred on a senior scholar in recognition of an extraordinary record of accomplishments, such as the publication of multiple important books or articles, rather than just the completion of an advanced graduate course and an accepted dissertation. At the same time, professorship at a Japanese university did not require a graduate degree, much less a doctorate. The practice, however, began to change drastically in the early 1990s and today doctorates are increasingly required for appointment to regular faculty positions at most Japanese universities and granted by numerous Japanese universities. More than 500 of them offer doctorates in the social sciences.

Nonetheless, a substantial number of political science faculty members at Japanese universities still do not hold a doctorate. In August 2012, a little over half of the about 1000 JPSA members currently teaching at Japanese universities held doctorates, a little over one-third master's, nearly 5% bachelor's, and the remaining 10% were unknown. Of those doctorate holders, nearly 80% had earned their degrees at Japanese universities and about 10% each at US and other foreign universities. An important reason for Japanese students' preference for Japanese doctorates is the flexibility, or laxity, allowed by most Japanese universities to satisfy the requirements for the degree: one can either enroll in a doctoral program at a particular university and complete both a course of classes and acceptable dissertation for a 'course doctorate' or complete a course at one university, then later write a dissertation and submit it either to the same university or any other degree-granting university for a 'thesis doctorate'. About 70% of those who held Japanese doctorates among the JPSA members in 2012 were 'course doctors', 20% were 'thesis doctors', and 10% were unknown. The hiring practice of most potential Japanese employers, including the nation's most prestigious universities, to give preference to holders of degrees from Japanese universities over those with degrees from foreign universities further discourages Japanese students from going abroad for a degree at considerable financial costs and, in most cases, with a substantial language handicap.

Not surprisingly, those who attend only Japanese universities tend to write and publish only in Japanese and nearly all do so. There are in Japan about a dozen major trade book publishers, half-dozen university presses, and several independent political science journals that publish their works, mostly subject to perfunctory manuscript reviews. In principle, these works are accessible to anybody who reads Japanese. The majority of Japanese professors' works, however, are published in the nonrefereed in-house periodicals of their own departments or schools, circulation of which is typically very limited and largely inaccessible to outsiders, especially to foreign scholars.

These circumstances tend to increasingly isolate and insulate Japanese political scientists in general and Japanese politics specialists in particular from the international community of scholars and limit opportunities for free give-and-take and collaboration between them. In the larger scheme of things, then, this trend of 'indigenization' may be the most important development in the study of Japanese politics in the post–Cold War period, which sets it apart from the evolving states of studies of Chinese and Korean politics.

Intraregional and Extraregional Comparative Studies

While the majority of studies of East Asian politics published in English in the post–Cold War period were single-nation studies, a modest number of works investigating certain political phenomena, such as Confucianism and democratization and political underpinnings of economic development or debacle, in cross-national or cross-regional comparative perspective, have emerged. These include studies of political and economic reforms and social welfare policies in South Korea and Taiwan, reunification issues between mainland China and Taiwan on one hand and South and North Koreas on the other, 'crony capitalism' in South Korea and the Philippines, political and economic reforms and transitions in China and Hungary, political power of business corporations in Japan and France, etc.

In the most ambitious studies of this type, region-wide political phenomena or developments in East Asia are compared with those in another region or regions. Stephen Haggard and his colleagues' work on the political economy of democratic transitions in East Asia and Latin America is an outstanding example of this type of studies (Haggard and Kaufman, 2008). McGuire's on the complex relationships between the levels of wealth, the quality of the welfare state, and the state of democracy in a dozen nations in the same two regions is another (McGuire, 2010).

Most of the comparative studies published in English in this period were descriptive case studies, like their predecessors of the Cold War period. Some, however, employed one or another theoretical model, or at least perspective. These include the use of a sociological evolutionary model in a study of Chinese foreign policy, the two-level games model in a study of Japanese–United States trade negotiations, network analysis in a study of Japan's relations with its Asian neighbors, and so on. Rational choice theory was used in a few studies of Japanese and South Korean politics.

The studies of East Asian politics in the post–Cold War period have thus become far more thematically and methodologically diversified than in the Cold War period.

See also: Asia, Sociocultural Aspects: China; Asia, Sociocultural Overviews: Korea; Cultural Policies in East Asia; East Asian Legal Traditions; East Asian Studies: Culture; East Asian

Studies: Economics; East Asian Studies: Gender; Historical Thought and Historiography: East Asia; Japan: Sociocultural Aspects; Nationalism, Historical Aspects of: East Asia; Regional Planning: China; Religions of East Asia; Sexual Minorities in China; Urban Planning in China.

Bibliography

Amsden, A.H., 1989. Asia's Next Giant: South Korea and Late Industrialization. Oxford University Press, New York.

Beardsley, R.K., Hall, J.W., Ward, R.E., 1959. Village Japan. University of Chicago Press, Chicago.

Cheung, T.M., 2001. China's Entrepreneurial Army. Oxford University Press, Oxford.

Clark, A., Tan, A.C., 2012. Taiwan's Political Economy: Meeting Challenges, Pursuing Progress. Lynne Rienner, Boulder, CO.

Culpepper, P.D., 2011. Quiet Politic and Business Power: Corporate Control in Europe and Japan. Cambridge University Press, New York.

Cumings, B., 1981, 1990. The Origins of the Korean War, 2 vols. Princeton University Press, Princeton, NJ.

Dickson, B.J., 2008. Wealth into Power: The Communist Party's Embrace of China's Private Sector. Cambridge University Press, Cambridge, UK.

Fell, D., 2005. Party Politics in Taiwan: Party Change and the Democratic Evolution of Taiwan, 1991–2004. Routledge, New York.

Gu, E.X., 1999. Cultural intellectuals and the politics of the cultural public space in communist China (1979–1989): a case study of three intellectual groups. Journal of Asian Studies 58, 389–431.

Guo, X., 2012. China's Security State: Philosophy, Evolution, and Politics. Cambridge University Press, New York.

Haggard, S., Kaufman, R.R., 2008. Development, Democracy, and Welfare States: Latin America, East Asia, and Eastern Europe. Princeton University Press, Princeton, NJ.

Henderson, G., 1968. Korea: The Politics of the Vortex. Harvard University Press, Cambridge, MA.

Johnson, C.A., 1982. MITI and the Japanese Miracle: The Growth of Industrial Policy, 1925–1975. Stanford University Press, Stanford, CA.

Kim, S., 2000. The Politics of Democratization in Korea: The Role of Civil Society. University of Pittsburgh Press, Pittsburgh, PA.

Kim, S.S., 2003. Korea's Democratization. Cambridge University Press, New York.

Koh, B.C., 1984. The Foreign Policy Systems of North and South Korea. University of California Press, Berkeley, CA.

Landry, P.F., 2008. Decentralized Authoritarianism in China: The Communist Party's Control of Local Elites in the Post-Mao Era. Cambridge University Press, New York.

Liberthal, K., Oksenberg, M., 1988. Policy Making in China: Leaders, Structures, and Processes. Princeton University Press, Princeton, NJ.

Maruyama, M., 1963. Thought and Behavior in Modern Japanese Politics. Oxford University Press, London.

McGuire, J.W., 2010. Wealth, Health, and Democracy in East Asia and Latin America. Cambridge University Press, New York.

Oka, Y. (Ed.), 1958. Gendai Nihon No Seiji Katei (Political Processes in Contemporary Japan). Iwanami Shoten, Tokyo.

Pekkanen, R., 2006. Japan's Dual Civil Society: Members without Advocates. Stanford University Press, Palo Alto, CA.

Pempel, T.J. (Ed.), 1990. Uncommon Democracies: The One-party Dominant Regimes. Cornell University Press, Ithaca, NY.

Shin, D.C., 2011. Confucianism and Democratization in East Asia. Cambridge University Press, New York.

Shirk, S.L. (Ed.), 2010. Changing Media, Changing China. Oxford University Press, New York.

Stockwin, J.A.A., 2003. Dictionary of the Modern Politics of Japan. Routledge Curzon, London and New York.

Suh, D.S., 1988. Kim Il Sung: The North Korean Leader. Columbia University Press, New York.

Tang, W., 2005. Public Opinion and Political Change in China. Stanford University Press, Stanford, CA.

Wade, R., 1990. Governing the Market: Economic Theory and the Role of Government in East Asian Industrialization. Princeton University Press, Princeton, NJ.

Whiting, S., 2001. Power and Wealth in Rural China: The Political Economy of Institutional Change. Cambridge University Press, New York.

Wong, J., 2004. Healthy Democracies: Welfare Politics in Taiwan and South Korea. Cornell University Press, Ithaca, NY.

Yang, D.L., 2004. Remaking the Chinese Leviathan: Market Transition and the Politics of Governance in China. Stanford University Press, Stanford, CA.

Relevant Websites

http://www.oxfordbibliographies.com/view/document/obo-9780199756223/obo-9780199756223-0054.xml?rskey=Jo7CiJ&result=50&q= – Blecher, Marc, 'Worker Politics in China,' Oxford Bibliographies: Political Science.

http://www.mext.go.jp/english/highered/__icsFiles/afieldfile/2012/06/19/1302653_1.pdf – Government of Japan, Ministry of Education, Culture, Sports, Science and Technology (MEXT), Higher Education Bureau, Higher Education in Japan.

http://www.moe.edu.cn/publicfiles/business/htmlfiles/moe/s4966/index.html – Government of the People's Republic of China, Ministry of Education: Education in China.

http://www.iie.org/en/Research-and-Publications/Open-Doors – Institute of International Education (USA), Open Doors.

http://www.princeton.edu/~lynn/chinabib.pdf – Malik, Lubna, and White, Lynn, Contemporary China: Book List. Princeton University, Woodrow Wilson School, Politics Department, East Asian Studies Program.

https://www.nafsa.org/_/File/_/ges/Rep%20Korea.pdf – NAFSA (Association of International Educators), Online Guide to Educational Systems Around the World-Republic of Korea.

http://www.dibrary.net/search/portal/searchStorage.jsp?kwd=Korean+politics&x=24&y=18 – National Library of Korea, 'Korean Politics'.

http://www.oxfordbibliographies.com/view/document/obo-9780199756223/obo-9780199756223-0041.xml?rskey=6iqmpS&result=45&q= – Stockton, Hans, 'Politics of Taiwan,' Oxford Bibliographies: Political Science.

Eastern European Studies: Culture

Michael D Kennedy, Brown University, Providence, RI, USA

Published by Elsevier Ltd.
This article is a revision of the previous edition article by M.D. Kennedy, volume 6, pp. 3979–3984, © 2001, Elsevier Ltd.

Abstract

The Cold War moved an American knowledge establishment to build area studies. Its East European focus built on Enlightenment notions of a contradictory place neither Orient nor Western. This combination of proximity and 'backwardness' shaped much of Eastern European Studies, but the region's diversity worked against the cultivation of area studies from within. The nation, however, has shaped other cultural studies and remains an object of study itself, elevated by the region's distinctive experience with communist rule. With communism's end, more culturally articulate methodologies prosper as they refashion historical understandings as well as meet more contemporary challenges of translation around transnational cultural politics.

Culture can be treated as a distinct object of study. For instance, literature and art are usually understood as expressions of 'high culture,' while the economy is typically studied as something apart from culture. Culture also is used to differentiate groups. Nations are not only political categories, but they are distinguished by language, ways of life, and interpretations of the past. Some argue that to explain social life at all, one must understand meanings in everyday life, or the culture of a place and time. Finally, culture is fundamental in social inquiry itself. It is embedded in the process through which the objects of our inquiry are constituted. All of these meanings of culture inform this alteration of an entry originally composed at century's end.

This revision again uses the final sense of culture referenced above to shape its overall architecture, but, given constraints on numbers of references, it omits original citations to allow for more recent work to be identified. To recognize some of the most important recent work, I also have drawn upon suggestions from a number of colleagues whose identities remain anonymous, as I have promised them.

The notion of 'Eastern European Studies' should be understood as a regional theme in the more general enterprise called 'area studies,' that interdisciplinary study of a region's languages, cultures, histories, and institutions (Wiley and Glew, 2010). This particular form of scholarship achieved great prominence in the USA during the Cold War, during which a new linkage between Western geopolitics and scholarship was cultivated. The American government sought expertise about other parts of the world in order to aid its struggle against communism's spread. Consequently, it subsidized research and training in area studies across the world's regions.

Combining Western geopolitical aims with scholarly work was relatively more compatible in studies of Eastern Europe and the Soviet Union than of other world regions. During and after World War II, the Soviet Union expanded its borders to include Estonia, Latvia, Lithuania, and Western Ukraine and Belarus, while communists took power in other parts of Eastern Europe – East Germany, Poland, Hungary, Czechoslovakia, Romania, Bulgaria, Yugoslavia, and Albania. Many political émigrés from these countries could be found among the West's experts in Eastern European Studies. Despite the political associations of exile and state sponsorship, Eastern European Studies enjoyed relative autonomy and scholarly growth through its location within various disciplines and universities. After communism's collapse throughout the region, Eastern European Studies no longer faces such obvious political implication. It does, however, face a challenge of coherence and other problems associated with culture's place in disciplinary practices.

Eastern Europe as Cultural Category

It is unlikely that Eastern Europe would be invented as a category if the world in which I write defined our geographical scholarly imagination. While East European nations are not so prominent as France and Germany in the status hierarchies and power relations of the European Union, the EU's formal organization and rights diminish the cultural sense of speaking in terms of Eastern Europe, at least with regard to nations presently within it (Böröcz, 2006; Kennedy, 2012). However, there are powerful historical foundations for its abiding reference.

Eastern European Studies has its antecedents in various forms of civilizational study, most notably 'Orientalism.' Like Orientalism, area studies expertise tends to be organized through networks of knowledge production grounded outside the region. While the Cold War inspired American area studies projects, an East European focus could build on an earlier West European invention. During the Enlightenment, West Europeans understood this region, then including Russia, as neither Orient nor Western, occasionally recognizable but also distant despite a number of overlapping cultural flows.

This tension between Western proximity and 'backwardness' shaped much of Eastern European Studies. One enduring debate asks whether Eastern Europe's indigenous social relations or its trade relations explain its economic underdevelopment over the last 400 years. Despite its economic trajectory, east European art, music, architecture, and literature have thrived, leading some to identify modernity's disarticulation as a regional distinction. These accomplishments lead others to argue that Europe itself should be understood as much from its east and especially its center as from its west.

Eastern European Studies has not only been cultivated in the West. For instance, the Soviet Union had its own complement of area studies. Eastern Europe was particularly important in this array given the role post-World War II Eastern Europe played, until *perestroika*, as a *cordon sanitaire*. Within Eastern Europe, however, the region's area study is not well developed; the region is rarely articulated as a category of cultural imagination. East European diversity has invited much finer breakdowns.

Most significantly, the region's imperial heritage is diverse – with the Ottoman, Habsburg, Russian, and Prussian empires shaping the region's long nineteenth century, with enduring effect. The 'Balkans,' especially those territories of the former Ottoman Empire in Southeastern Europe, have remained an important regional category throughout the twentieth century; with war in the 1990s, the term gained new cultural power. Those from the Westernmost parts of Eastern Europe also took the area's diverse imperial legacy as a starting point for their contemporary critique of Soviet domination, arguing that communism ill suited East *Central* Europe with its more developed civil societies. Some resisted Cold War categories altogether and argued that East Central Europe's boundaries should move into Soviet space, to include the lands of the fifteenth- and sixteenth-century Polish–Lithuanian Commonwealth. Others even organize scholarly collaborations on those terms.

There is no simple cumulation of scholarly competencies in Eastern European Studies. There has been no regionally hegemonic language, like Russian was in the Soviet Union. Depending on the boundaries one invokes, and the time one imagines, different languages have been more and less dominant and numerous. For example, in 1900, Germans, Poles, and Ukrainians may have been the largest ethnolinguistic groups in East Central Europe, but Romanians, Hungarians, Czechs, Belorussians, Serbs, Greeks, Bulgarians, Croats, Slovaks, Lithuanians, Slovenes, Albanians, Macedonians, Bosnian Muslims, and Montenegrins were all dominant in particular places across the region. Beyond these groups who then had, or who would eventually have, East European states in their name, there were significant numbers of Jews, Turks, Gypsies/Rom, Italians, Carpatho-Rusyns, Szekelys, Friulians, Armenians, and other ethnolinguistic groups.

There also is no principal religious tradition in Eastern Europe, whose status Islam occupies in the definition of Middle Eastern and North African studies. Islamic and Orthodox, Roman Catholic, and Protestant Christian religions dominate various parts of the regional religious landscape. Judaism was demographically and culturally significant before the Holocaust, and even after genocide, Judaism remains critical to understanding the region's cultural politics. Thus, while Eastern Europe is itself a profoundly important cultural category, it is not based on the region's cultural homogeneity. Eastern European Studies is, however, profoundly shaped by the culture of nations and nationalism.

The Significance of the Nation

Cultural analysis in Eastern European Studies draws upon an indigenous scholarship, but one organized through national and past imperial fields rather than across the region's whole. For instance, Polish sociologies of the intelligentsia, both before and after World War II, had the nation, and the multinational republic out of which it grew, as their field of reference. Hungarian ethnographies of the countryside, while foundational for Western anthropology's study of Transylvania and elsewhere, were focused on the making of the Hungarian nation and its relation to empire. Musicology or literary criticism also has been organized within national and imperial fields, around leading figures like Béla Bartók or Witold Gombrowicz, whose complicated expressions of their identifications through music and literature form leitmotifs of their cultural interpretation. Consequently, the nation is one of the region's most important cultural formations. It shapes other cultural studies and is an object of study itself.

Studies of the nation are themselves framed within larger cultural formations. Many studies of the nation in Eastern Europe are set up within a narrative of civilizational distinctions. For instance, national historiographies tend to depict the Ottoman Empire's expansion through the 1600s and its retreat from the Balkans during the nineteenth and beginning of the twentieth century in terms of a nationally articulated contest between Islam and Christianity, on which more recent stories of civilizational clash draw. Romanian national historiographies identify their civilizational alternatives with Western and more autochthonous traditions. Estonian social science around ethnicity can portray a 'civilizational' difference between the West and Russia.

In contrast to those who demonstrate enduring cultural oppositions, others have emphasized the *process* of establishing national identity and difference, asking, for example, how intellectuals articulate nations. These recurring questions in national stories tend to differ by place. Those who focus on 'historic' nations and nations that had to '(re)awaken' have different emphases. The latter tend to ask how peoples come to recognize themselves as nations and how they mobilize to realize political ambitions including the formation of nation-states out of imperial states. Those who focus on historical nations are more influenced by narratives of loss, whether of sovereignty, territory, or tolerance.

Nationalism and its study are not limited to Eastern Europe, but the region was foundational, given the ways in which scholars used this region to highlight the original categorical variants of civic and ethnic nations. The last decade's emphasis on the discursive construction of nations implicated in various fields of practice has made such gross distinctions of nationalism fade, at least in the scholarly domain. It remains prominent in political discourse if with new accents.

Discourses around corruption, for example, have become more commonly used in defining less virtuous nations, aided and abetted by social scientific methods creating indices of transparency and trustworthiness in business, politics, and everyday life (Ledeneva forthcoming). Bulgaria and Romania, while full members of the European Union, have been indicted for this lack; those aspiring to accession in the European Union face even stiffer criticisms for such deficit, whose likelihood is made even greater by war's aftermath.

The Wars of Yugoslav Succession and their abiding effects continue to shape nations and nationalisms and their identifications from without. Superficial and interested actors can

elaborate accounts of 'ancient hatreds' to structure practice around mythologically perpetual animosities. More culturally articulate scholarship moves to explain how memories of oppression by others, rekindled by polyvalent symbols, became the cultural stock for civil societies' and political networks' mobilizations to hate and violence.

As national security comes to be more established, and with more international collaborations and methodological diversity influencing studies of the nation, contests over the meaning of the nation, whether through symbols and events in everyday life (Zubrzycki, 2006; DiLellio and Sievens-Schwander, 2006; Brubaker et al., 2008), or in their historical constitution with other nations' imperialisms (Pula, 2008), have become far more significant. It is also critical to consider under communist rule and in its transition culture sequel.

Culture, Civil Society, and Transition

Communist rule encouraged many nationally articulated scholars to consider culture to be autonomous from the system forced upon them. Socialist realism in the cultural sphere, while often accommodated, was often seen as an imposition against which more authentic cultural expressions sought release. These expressions could be national in form, or evaluated with various aesthetics in mind. Indeed, while communism's political censorship certainly constrained cultural expression, it also made artistic expression broadly valued and relatively independent of culture's commodification. At the same time, that expression could be seen as part of a broader struggle for freedom for which the nation was a primary container and vehicle, manifest in the punctuations of communism's systemic crises, in 1956, 1968, 1980, and 1989 most obviously.

The Hungarian Revolution of 1956, the Prague Spring of 1968, the Solidarity movement of 1980–81, and the end to communist rule throughout most of the region in 1989 all implied a mismatch between the nation and communist rule. Studies of political culture were frequently organized to explain the conditions under which the political cultures of some nations would or would not resonate with communist political culture. More recent studies tend to break down the assumption of opposition between nationalism and communism and rather study how opposition and authorities construct the nation in their own terms. Not every nation managed to develop easily imagined dichotomies of 'us' and 'them' however. The discourse of civil society was critical to that formation.

That discourse found modes of expression, and even optimism, in various cultural products and forms of social organization. Civil society's culture was argued to be tolerant, pluralist, and progressive. Just as studies of the nation have struggled to move beyond their political instrumentalization, so too have engagements of civil society. While it was politically fortuitous to assume civil society as a harmonious blend of ideology, social organization, and vehicle of liberation in opposition to communist rule, its formation after communism is radically different. The very spaces opened up in civil society's flourish could also produce nationalist expressions that threatened others and even lead to war. But it also led to much more diverse accounts of social identity and relations after communism, initially within a new hegemonic frame of transition.

Communism's end offered peoples new opportunities to redefine their nation, and this invited substantial scholarly engagement. These transformations were apparent already in 1988, when for instance the polysemy of Béla Bartók was played out in the last summer of communist monopolistic rule in Hungary (Gal, 1991). Some scholars shaped claims to national distinction with their own expertise, clarifying the ways in which proximate languages were related to one another and could distinguish nations, leading others, further removed, to explain the complexities and sources of these arguments. Some utilized less culturally reflexive approaches like rational choice or game theory, but the cultural field and historical legacies of places and peoples continued to shape scholarly approaches to national identification. While the nation demanded some attention to historical legacies and cultural formations, other subjects seemed to move without cultural reference.

In the first years after communism's collapse, most studies of transition to markets and democracy were rather formal and structural, as more and more scholars without contextual expertise identified Eastern Europe as the natural laboratory for testing their generalizing theories. This led to significant debate, especially in political science, about the relative value of area studies. Much of the contest hinged upon how much one needed to know about culture and history in order to produce a good explanation, but it did not attend to the global cultural formation of transition itself (Kennedy, 2002) and to how the 'normal' was itself an ideological element structuring change and contest (Eglitis, 2002). Of course methodology also matters in this variable attention.

Ethnographies of transition were more likely to question the guiding concepts of transition, especially by noting the hermeneutic gap between transition's key concepts and their everyday meaning (Berdahl et al., 2000; Ghodsee, 2011). For instance, such inquiries help us appreciate the difficulty of distinguishing private from public ownership and the meanings of property itself (Verdery, 2003). Sometimes ethnographies even challenged the framework in which transition is envisioned, notably in the kinds of capitalism actually being constituted.

A similar range of questions has emerged around gender studies. As in most transition studies, some gender analysts have kept culture at a distance. By treating gender more as a variable – asking, for instance, whether women's employment chances have been disproportionately hurt by economic transition – cultural analysis is less obviously necessary. But when gender serves as a lens for interpreting social change, gender studies engage profound cross-cultural questions. For instance, terrific debates have occurred over the measure of cultural imperialism involved in Western feminism's categories and modes of inquiry with regard to Eastern Europe. Indeed, some have argued that transition has helped constitute new gender identities differentiated by the profound inequalities made by distinctive logics of market transition (Weiner, 2007). Beyond object and lens, gender also has become one of the most powerful vehicles for developing collaborations across the region, for example, around citizenship (Lukić et al., 2006) and visual culture (Pejic, 2010).

Cultural Formations after Postcommunism

As a field of inquiry, Eastern European Studies has been organized through culturally constituted geopolitical categories whose sense has changed profoundly. During the Cold War, questions of fit between particular nations and communist rule seemed to drive Eastern European Studies. After communism's collapse, the field was reconstituted around questions of transition and nationalism's threat. While the logic of transition culture continues to influence candidate countries aspiring to EU membership, cultural social science has moved well beyond its hegemonic frame, not least for the transformations wrought by European Union enlargement and the inequalities borne within it.

One of the best examples of this shift can be found in Gille's (2007) analysis of the Hungarian waste economy's movement from an emphasis on 'reuse' during early socialism toward waste disposal in its end and after. In particular, the European Union's provision of model legislation, technology, investments, and even waste created a new landscape in Hungary filled with more landfills, incinerators to deal with them, and the environment resulting from it. While illustrative of the critical turn, work like Gille's also illustrates Eastern European Studies' increasingly broad range and methodological diversity.

On the one hand, open borders to scientific collaboration enabled social scientists without area expertise to work with indigenous scholars to produce important findings, bracketing culture as extraneous to the scholarly enterprise in ways that area studies could never manage. More open borders have also meant, however, greater room for life histories, ethnographies, and other direct qualitative fieldwork, and more intimate portraits of everyday life and assessments of its fit with the larger frameworks guiding global visions of Eastern Europe in its historical trajectories. More open research opportunities have, in effect, enabled a more critical cultural social science. This is especially evident in rethinking communist rule itself.

Transition culture demanded that socialism be left behind. Even as populations age and direct experience with the system fades, more scholars are examining how various genres of memory reflect different subjectivities, power relations, and contests under communist rule (Todorova, 2010) with artists among the most critical in that endeavor (Von Oswald et al., 2009). In particular, explorations of communist era monumentalities, and their relationship to subsequent cultural politics, have become an increasingly prominent part of East European cultural studies (James, 2005).

At the same time that memories of the system are being analyzed, the actual processes of transformation, and their implication in cultural systems and practices, are being explored with increasing analytical and theoretical rigor. Some consider how the material culture of the system (Fehervary, 2009) or its media (Bren, 2010) allows us to recognize its political logic. Others have revisited the system's founding moments; for example, Kligman and Verdery (2011) help us understand better not only how classes were remade through collectivization but also how in the process socialist society was 'personalized' and mistrust systematized. Still others have explored how the system actually collapsed; for example, Glaeser (2011) has proposed that one reason socialism collapsed was because the authorities' understanding failed; it was based on circular validations preventing their recognition of the system's dynamics, which in turn prevented them from crafting an appropriate response to the opposition's challenge.

While East European Studies thus still cohere by understanding the system that made such a regional identification compelling, some of the most exciting cultural social science in East European Studies resonates far beyond the region's historiography. For instance, Keith Brown (2013) not only helps us to appreciate the historical foundations of the Macedonian Revolutionary Organization, but also offers new conceptual tools for rethinking the relationship between terrorism and flows of labor, guns, and of revolutionary strategies and of imagined communities beyond nation but residing in terminal communities of another constructed sort. It is also apparent in Kubik and Lynch's (2013) volume refashioning East European Studies through the elevation of vernacular knowledge in theory, evidence, and practice around the articulation of hegemony, injustice, and social movements.

As part of a new post-Cold War conception, such work invites comparisons to the postcolonial, as Forrester et al. (2004) and Chari and Verdery (2009) propose. Indeed, recent scholarship in Kosova suggests that scholars working in the region could offer new critical frameworks through which not only contests over socialisms' past and wars' outcomes but around ongoing internationalist interventions refigure conceptions of gendered and emancipatory publics (Luci and Gusia, 2014). Indeed, a strong case might be made that in an era where publics become thin and passive, feminist art interventions offer a means, here in Poland, to stimulate greater reflexivity in public discourse (Matynia, 2005). And it might even go further.

Larson's (2013) ethnography of critical thinking in Slovakia suggests that we are better served to consider the ways in which even such noble pursuits as civic education build into their cultural politics other modes of dogmatic thinking. In this sense, while the iconic captive mind of the communist era is a thing to be resisted, its opposite is not so easy to cultivate, regardless of location. Eastern Europe makes that quite clear.

See also: Area and International Studies: Cultural Studies; Communism, History of; Eastern European Studies: History; Eastern European Studies: Society; Nationalism, Sociology of; Socialist Societies: Anthropological Aspects.

Bibliography

Berdahl, Daphne, Bunzl, Matti, Lampland, Martha (Eds.), 2000. Altering States: Ethnographies of Transition in Eastern Europe and the Former Soviet Union. University of Michigan Press, Ann Arbor.

Böröcz, Jozsef, 2006. Goodness is elsewhere: the rule of European difference. Comparative Studies in Society and History 48 (1), 110–137.

Bren, Paulina, 2010. The Greengrocer and His TV: The Culture of Communism after the 1968 Prague Spring. Cornell University Press, Ithaca.

Brown, Keith, 2013. Loyal unto Death: Trust and Terror in Revolutionary Macedonia. Indiana University Press, Bloomington.

Brubaker, Rogers, Feischmidt, Margit, Fox, Jon, Grancea, Liana, 2008. Nationalist Politics and Everyday Ethnicity in a Transylvanian Town. Princeton University Press, Princeton.

Chari, Sharad, Verdery, Katherine, 2009. Thinking between the posts: postcolonialism, postsocialism and ethnography after the Cold War. Comparative Studies in Society and History 51 (1), 6–34.

DiLellio, Anna, Schwandner-Sievers, Stephanie, 2006. The legendary commander: the construction of an Albanian master-narrative in post-war Kosovo. Nations & Nationalism 12 (3), 513–529.

Eglitis, Daina Stukuls, 2002. Imagining the Nation: History, Modernity and Revolution in Latvia. Pennsylvania State University Press, State College.

Forrester, Sibelan, Zaborowska, Magdalena J., Gapova, Elena (Eds.), 2004. Over the Wall/after the Fall: Postcommunist Cultures through an East-West Gaze. Indiana University Press, Bloomington.

Fehervary, Krisztina, 2009. Goods and states: the political logic of state socialist material culture. Comparative Studies in Society and History 51 (2), 426–459.

Gal, Susan, 1991. Bartok's funeral: representations of Europe in Hungarian political rhetoric. American Ethnologist 18, 440–458.

Ghodsee, Kristen, 2011. Lost in Transition: Ethnographies of Everyday Life after Communism. Duke University Press, Durham.

Gille, Zsuzsa, 2007. From the Cult of Waste to the Trash Heap of History: The Politics of Waste in Socialist and Postsocialist Hungary. Indiana University Press, Bloomington.

Glaeser, Andreas, 2011. Political Epistemics: The Secret Police, the Opposition, and the End of East German Socialism. University of Chicago Press, Chicago.

James, Beverly A., 2005. Imagining Postcommunism: Visual Narratives of Hungary's 1956 Revolution. Texas A&M Press, El Paso.

Kennedy, Michael D., 2002. Cultural Formations of Postcommunism: Emancipation, Transition, Nation and War. University of Minnesota Press, Minneapolis.

Kennedy, Michael D., 2012. Cultural formations of the European Union: integration, enlargement, nation and crisis. In: Friedman, Rebecca, Thiel, Markus (Eds.), European Identity and Culture: Narratives of Transnational Belonging. Ashgate, Aldershot, pp. 17–50.

Kligman, Gail, Verdery, Katherine, 2011. Peasants under Siege: The Collectivization of Romanian Agriculture, 1949–1962. Princeton University Press, Princeton.

Kubik, Jan, Lynch, Amy T. (Eds.), 2013. Justice, Hegemony, and Social Movements: Views from East/central Europe and Eurasia. New York University Press, New York.

Larson, Jonathan, 2013. Critical Thinking in Slovakia after Socialism. University of Rochester Press, Rochester.

Ledeneva, Alena, 2013. A critique of the global corruption paradigm. In: Kubik, Jan, Lynch, Amy T. (Eds.), Justice, Hegemony, and Social Movements: Views from East/central Europe and Eurasia. New York University Press, New York.

Luci, Nita, Gusia, Linda, 2014. "Our men will not have amnesia" civic engagement, emancipation and gendered public in Kosovo. In: Ramet, Sabrina (Ed.), Civic and Uncivic Values in Kosovo. Central European University Press, Budapest.

Lukić, Jasmina, Regulska, Joanna, Zaviršek, Daria (Eds.), 2006. Women and Citizenship in Central and Eastern Europe. Ashgate, Aldershot.

Matynia, Elzbieta, 2005. Feminist art and democratic culture: debates on the new Poland. PAJ: A Journal of Performance and Art 27 (1), 1–20.

Pejic, Bojana, 2010. Gender Check: Femininity and Masculinity in the Art of Eastern Europe. Museum Moderner Kunst Stiftung Ludwig Wien, Vienna.

Pula, Besnik, 2008. Becoming citizens of empire: Albanian nationalism and fascist empire, 1939–1943. Theory and Society 37 (6), 567–596.

Todorova, Maria, 2010. Remembering Communism: Genres of Representation. Social Science Research Council, New York.

Verdery, Katherine, 2003. The Vanishing Hectare: Property and Value in Postsocialist Transylvania. Cornell University Press, Ithaca.

Von Oswald, Anne, Schmelz, Andrea, Lenuweit, Tanya (Eds.), 2009. Errinnerungen in kultur und kunst. Reflexionen uber krieg, flucht und vertreibung in Europa. Verlag, Bielefeld.

Weiner, Elaine, 2007. Market Dreams: Gender, Class and Capitalism in the Czech Republic. University of Michigan Press, Ann Arbor.

Wiley, David, Glew, Robert (Eds.), 2010. International and Language Education for a Global Future: Fifty Years of Title VI and Fulbright-Hays Programs East Lansing. Michigan State University Press, East Lansing.

Zubrzycki, Genevieve, 2006. The Crosses of Auschwitz: Nationalism and Religion in Post-communist Poland. University of Chicago Press, Chicago.

Eastern European Studies: History

Tomasz Blusiewicz, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W.L. Blackwood, volume 6, pp. 3990–3991, © 2001, Elsevier Ltd.

Abstract

This article is a brief survey of the current state of and trends in academic historical scholarship on Eastern Europe, mainly in North America, but also in Europe. It presents several traditional methods of defining Eastern Europe, discusses main features of its historiography and some of its recent research directions. The article offers both a chronological and thematic narratives concerning the field's origins, its major developments and shifts as well as a reference guide for further research and reading.

The field of Eastern European historical scholarship has undergone dramatic changes since the collapse of communist regimes in 1989. Eastern Europe as a distinct geopolitical region has never been more sharply delineated than between the late 1940s and 1989, when its primary reference was essentially synonymous with the term 'Soviet Block.' In the West, 'Eastern Europe' as an intellectual concept had its origins in the Enlightenment vision of the world; it referred to a transitory space between the civilized 'West' and the exotic 'Orient' (Wolff, 1994). As a scholarly discipline it began to emerge only in the interwar period – it dealt with the newly organized space that appeared on the map as a result of the fall of four empires: Austro-Hungarian, German, Ottoman, and Russian. It was the Cold War and its geopolitical pressures, however, that led to the emergence of 'Eastern European studies' as an institution well-established in the academic world. Many historians of the region were émigrés and – as they were collectively labeled in the West – 'dissidents,' who through their research, writing, teaching, and not infrequently – government advising, were involved in the process of 'liberating' Eastern Europe from the Soviet domination (Brzezinski, 1956; Pipes, 1954). Contributing to the Cold War confrontation was not the only preoccupation of historians of Eastern Europe, but nonetheless the field has been faced with a need to reinvent its basic framework after the dissolution of the Soviet Union in 1991.

Many scholars, including the author of the previous article in this encyclopedia (Blackwood, 2001), predicted that such a redefinition would be extremely difficult and the field would eventually dissolve and/or integrate with the broader European narrative. This process is far from complete and while the field has been shifting in that direction, it still retains its specificity and can be approached as a separate unit. This article focuses on two related themes. First, it examines what criteria are used by historians to define what counts and what does not count as 'Eastern European history.' Second, it surveys new work in the field, outlines how it is related to older historiography and how it communicates with new trends in historical scholarship. In conclusion, it offers several reflections on what kind of light some recent geopolitical shifts such as the EU enlargement (in the 2000s) and the Ukrainian crisis (2013–14) shed on Eastern Europe as a concept and what they portend about its future.

The chief difficulty facing historians who insist on the validity of Eastern Europe as a historical term is the lack of Eastern European identity among the peoples of the region. This obstacle is much less pronounced for nation-state (German, Polish, Hungarian, etc.) or imperial (British, Soviet, Habsburg, etc.) historians where the domain of study is much more clearly defined through subjective identities and/or political boundaries. Eastern Europeans rarely think of themselves as Eastern Europeans per se; all studies show that this identity is either nonexistent or insignificant in comparison with national, religious, or pan-European identities. In this sense, the field of Eastern European historical studies can be seen as 'inorganic' – its origins are rooted in global conditions largely beyond Eastern Europe's control (imperial rivalries, World Wars, the Cold War, etc.). In other words, the field owes its existence largely thanks to an outsider 'gaze' cast upon the region.

One of the most authoritative surveys of the region's twentieth-century history (Rothschild and Wingfield, 2008) is entitled *Return to Diversity*. The title unequivocally implies that the region has returned to its historical 'norm' (diversity) after half a century of 'forced' convergence in the Soviet Block. Eastern European 'return to diversity' is tied with one of the original lenses through which the region has been perceived – the allegory of a bridge between East and West, a place where cultures meet and give birth to numerous mixed formations. The validity of the 'return to diversity' thesis is implicitly confirmed by recent developments in political science and in the new subfield of 'transitology' in particular; one of transitology's chief aims is to explain the diverging trajectories of states that seemed so similar politically prior to 1989. Many scholars search for deeper historical causes such as early modern democratic experiences to account for this phenomenon. What are the main features amid this newly reconstituted diversity that form the common denominators large enough to validate speaking of a common regional history?

Perhaps the oldest division line separating 'Western' and 'Eastern' Europe is of a religious nature. Starting with the split of the Roman Empire into two halves, through Byzantine-Greek Christianization of much of the Balkans or the Kievan Rus, the Great Schism of 1054 or the Muscovite 'Third Rome' ambitions, the rift between Latin and Greek Europe has always been noticeable and usually deep. In 1983, the Hungarian historian Jenő Szűcs suggested that parts of what was then known as 'Eastern Europe' (Poland, Czechoslovakia, Hungary) were miscategorized. These areas have always been, he argued, a part of the 'Western civilizational sphere' since their adoption of Christianity from Rome and not from Constantinople. The Orthodox faith and administration, the Cyrillic alphabet or the

absence/weakness of Protestant Reformation are thus posited as the original and decisive factors explaining the differences between the two parts of the continent. In Poland and Hungary, the image of these two countries being the 'bulwarks of Christian faith' containing the Orthodox and (more typically) Islamic element has become a part of national consciousness in the early modern period. In general, this basic distinction is rarely questioned, but there is little new scholarship that would operate within religiously defined frameworks, especially after the 'atheistic' communism cast aside what had been earlier perceived as a defining feature of Russianness – Orthodox religiosity.

Another frequently drawn and almost as old a distinction is of economic character. The 'Iron Curtain' ultimately did not fall between the ports of Stettin and Trieste as Winston Churchill announced in 1946, but closer to the ports of Lübeck in Germany and Burgas in Bulgaria. This seemingly incidental detail of geography must have intrigued many historians. The border between West and East Germany eventually materialized very close to the Elbe River – one of the major fault lines drawn by economic historians. It separates the world of nascent commercialization of agriculture in Western and the second wave of feudalism in Eastern early modern Europe. Naturally, the final course of the border between East and West Germany had nothing to do with economic history and a lot to do with where the Allied and Red Armies met in 1945. Still, the next four decades made the Elbe River divide, already bleak in 1945, plainly visible again. Eastern Europe, besides being popularly and somewhat stereotypically viewed as Western Europe's poor relative, is understood by some economic historians to be a region where the second wave of feudalization and deurbanization took place just when Western Europe began to urbanize and ready itself for the industrial revolution. Both the forced Soviet-style industrialization and the current EUintegration are sometimes framed as attempts to close this old gap. The initial advancements achieved by the Soviet Block economies between 1945 and mid-1970s have been interpreted by a prominent historian of the region, Ivan Berend, as the region's temporary 'detour from periphery,' a detour brief in its promise since the region returned to its point of departure after the stagnation and collapse of the 1980s (Berend, 1996).

Jenő Szűcs has also pointed to the spread of civil and municipal liberties in the late Middle Ages, exemplified, for instance, by urban settlement based on the German (Magdeburg) law as a factor with momentous political consequences for where the border between 'Western' and 'Eastern' Europe eventually crystallized in political terms. The substantial experiences with (qualified) democracy, civil rights, and the rule of law in, e.g., the early modern Habsburg Empire or the Polish–Lithuanian Commonwealth, were occasionally cited by Cold War era historians as features distinguishing 'democratic' and 'Central European' parts of the Soviet Block from the 'despotic' territories traditionally administered from Moscow and Istanbul. In this context, the stable Czechoslovak democracy of the interwar period was frequently seen as evidence that this state was firmly in the 'Western' sphere while all the other 'children of Versailles,' who eventually drifted toward various forms of authoritarianism, were truly in 'Eastern Europe' (Seton-Watson, 1945). This way of viewing Europe became more popular especially after the suppressed 'prodemocratic' unrest in Hungary, Czechoslovakia and Poland (1956, 1968, 1970, 1980) and after the 'successful' democratization in those countries and 'problematic' democratization in Belarus, Ukraine, or Russia after 1989/91.

Religious, economic, or political are merely three among numerous other 'macro' lines according to which the map of Europe can be divided. As Larry Wolff reminds us, before Peter the Great's introduction of Russia onto the European diplomatic scene, the division of Europe into 'north and south of the Alps' had been much more firmly embedded in the European mind (Wolff, 1994). All lines of this type are of course arbitrary and depend on the kinds of questions asked and the historical period studied; no serious historian argues that one can define what Eastern Europe is and what its borders are precisely and conclusively. Nevertheless, the region's entanglement with communism in the twentieth century is a part of its heritage of that cannot be overestimated. Judging by volume alone, the twentieth century communist experience, with all of its extensive and complex tropes, is currently the main subject of study for historians of the region. This experience is strongly intertwined with more pan-European topics such as totalitarianism, World Wars, and superpower confrontation. If one adds the Holocaust, ethnic cleansing, forced population transfers, and all other 'terrible things that happened in the twentieth century,' what emerges is singlehandedly the most extensive subject of the region's historical research and the cornerstone of its historical memory.

Before surveying some of the major trends in historiography, it is worth mentioning that Eastern Europe borders with several well-established regions of historical scholarship: Russian/Soviet, Ottoman/Middle Eastern, German, Italian, Scandinavian, etc. The borders between them will never be rigid and the very impossibility of drawing even tentative lines in some areas lead to a new subfield of scholarship which focuses precisely on these mixed, ambiguous, and contradictory areas often called 'borderlands.' This field is of course not limited to Eastern Europe, but some of its provinces such as Galicia (Frank, 2005), Transylvania (Case, 2009) and towns such as Lvov (Amar, 2011) or Salonica (Mazower, 2005), due to their ethnic, religious, or cultural diversity, have by now become recognized as exemplary cases demonstrating just how immensely complex and difficult to disentangle borderland territories can be. It has to be kept in mind, however, that what might be described as a 'typical' Eastern European borderland by an outsider might be considered to be an integral element of a core ethnic area by a local nationalist (Transylvania for some Hungarians or Romanians, Galicia for Poles and Ukrainians).

While historians rarely argue about whether a given region or topic belongs to, for example, Eastern European or Russian studies, what did emerge as an important contender for Eastern European studies is the idea of Central Europe and the scholarly movement it inspired. This term has its roots in the imperial German concept of *Mitteleuropa*, but it reemerged in a different context half a century later. It was reinvented and refashioned by anticommunist dissidents such as Milan Kundera, Czesław Miłosz, and Danilo Kiš. Kundera's essay "The Tragedy of Central Europe" (1984) is often perceived as the moment when the idea was reborn. In essence, the argument was that Central Europe (by which Czechoslovakia, Poland, and Hungary were usually meant) has always belonged to

Western Europe in civilizational terms, but has been 'kidnapped' and politically imprisoned in the Eastern European sphere by the Soviet power. While the political implications of this intellectual movement became largely irrelevant after 1989 and certainly after the EU extension in 2004 and 2007, a new wave of scholarship inspired by them emerged and found institutionalization in, e.g., the Central European History Journal.

This new 'Central European wave' is located in an intermediate position between the old German and the dissident ideas; the late Austro-Hungarian Empire, with its cosmopolitan and culturally vibrant cities of Vienna, Prague, Budapest, or Cracow, are painted as 'archetypically' Central European sites. Scholars who identify themselves as Central Europeanists come from various backgrounds; not infrequently they are historians of German-speaking lands who reach out to areas that had a rich record of contact with German culture and administration. While this field does not compete with Eastern Europe in any direct sense, its position has been steadily ascending vis-à-vis its neighbor for several reasons. The late Habsburg Empire has enjoyed a remarkable reputational renaissance, largely because its many imperfections do not stand a comparison with all the tragic events that befell the region after the Empire disintegrated. The Empire has been portrayed as a kind of a proto-EU project where many ethnicities and languages coexisted in relatively peace within one political unit for centuries (Oscar Jászi, 1929). Furthermore, fin-de-siècle Vienna or Prague with world-class artists and intellectuals such as Freud, Mahler, Wittgenstein, or Kafka are commonly seen as manifestations of European cultural vibrancy at its best (Schorske, 1979). But most importantly, the inhabitants of Prague or Cracow virtually universally prefer being spoken of as Central Europeans rather than Eastern Europeans. The success of the Central European University in Budapest shows that there is plenty of willingness to cooperate and produce knowledge under a common Central European banner.

Central European scholars such as Pieter Judson or Rogers Brubaker have been at the forefront of the study of nationalities and nationalism – a traditional Eastern European topic – and the late Austro-Hungarian Empire has been one of the main case studies used to address issues such as national identity formation, national cohesion and indifference, bilingualism, or mixed families (Judson, 2006; Brubaker, 1996). Following Eugene Weber's groundbreaking book *Peasants into Frenchmen* (1976), historians of the Austro-Hungarian Empire demonstrated how a nineteenth-century region inhabited largely by villagers and small town folk with local (geographical, tribal, familial) or religious identities turned into region as we know it today – inhabited by Germans, Czechs, Slovaks, or Croats (King, 2002). One of the leading questions still to be answered is 'when (and why) nationalism began to hate' – i.e., how an initially 'progressive' ideology of intellectuals intent on helping the 'small peoples' suffering under oppressive imperial government transformed itself into one of the main sources of conflict in twentieth-century Europe (Porter, 2000). In this context, the transatlantic gaze of 'outsiders' and the multiethnic composition of the Austro-Hungarian Empire helped to avoid the trap into which many native nation-state historians fell – writing biased national histories favoring one particular ethnicity and contributing to keeping the flame of nationalism alive. This type of approach to the study of nationalism has seen fruitful application to the Soviet area as well. Terry Martin's *Affirmative Action Empire* (2001) demonstrates how illusory was the Soviet (and global) conviction that the so-called 'nationalities question' had been solved in the borderlands of the Soviet Union and why this issue came back with a vengeance in the late 1980s as the Union disintegrated along its ethnically defined internal borders.

How the relationship between 'Eastern' and 'Central' European fields is going to evolve is impossible to predict; both are products of a unique set of circumstances and both are likely to find their place within a larger European or global framework. Immediately after 1989, the traditional Eastern European (and to some extent Soviet) field was supplied with new materials accessible at the local archives. The first wave of research focused on the early period (1945–56) around the theme of communist takeover and Stalinization (Kenney, 1997; Connelly, 2000). The current consensus on the subject has been well summarized by Joseph Rothschild and Nancy Wingfield:

> Today we recognize that many of the Western academic analyses of the 1950s and 1960s subscribed to exaggerated images of a rigid blueprint that supposedly guided Moscow and the local East Central European Communists in implementing the procedures and arrangement that Stalin eventually selected to give effect to his perception of Soviet hegemonial requirements. But while correcting this early presumptions, we should not throw out the baby with the bath water. For though there probably never was a rigid uniform Stalinist plan to be imposed in cookie-cutter fashion on every state and society of East Central Europe without regard to diverse national histories, institutions and complexities, the general overall similarities in Soviet and local Communist behavior throughout the area in the first decade after World War II strongly suggests a unified conceptual framework – more flexible at the beginning than toward the close of that decade, never entirely absent yet also never absolutely rigid.
>
> Rothschild and Wingfield, 2008: 75–76

A lot had been known about Stalinist terror and crimes, but new evidence allowed scholars to present detailed findings and provide more accurate statistics (Naimark, 1995). Documenting and making sure the memory of various types of transgressions committed by the Soviet and puppet regimes does not disappear became the main preoccupation of many locally based historians (Applebaum, 2012; Kusnierz, 2013); special national institutes (e.g., the Institute of National Remembrance in Poland) were established precisely with that goal in mind. The same observations apply to issues such as Nazi occupation and the Holocaust, most of which had been if not explicitly suppressed, then not very thoroughly explored by historians under communism, especially the period between September 1939 and June 1941, when Hitler and Stalin were allies (Gross, 1988). Continuing the tradition inaugurated by Czesław Miłosz (1953), intellectual historians maintained their interest in the strength of communism's 'Hegelian bite' – trying to understand why Marxism appealed to so many Eastern European (and not only) intellectuals and what was their relation with the regime and its ideology (Shore, 2006).

One aspect of the Stalinist period in Eastern Europe (1945–53) in particular that could not have been explored earlier was the issue of forced population transfers and deportations.

While historians in the West could argue to what extent these transfers were voluntary or not, historians in Eastern Europe had to accept the official party line that spoke of their voluntary nature. Czechoslovakia, Poland, and East Germany were in an ambivalent position due to the way the Germans were deported (1945–48) on the one hand and due to the 'alliance' of 'brotherly nations' that was now imposed on them by their Soviet supervisors on the other. Hence the spike in research on this topic after 1989 (Ther and Siljak, 2001). Overall, the findings show that the scale and brutality of post-1945 deportations have been much greater than it had been suspected. The ironically disturbing conclusion is that it was eventually Joseph Stalin who helped to bring the old nationalist fantasy of ethnically pure states to fruition while where no clear ethnic separation took place (e.g., Eastern Ukraine, Southern Slovakia) – tensions are still high today, not unlike in the interwar period.

The semblance of uniformity within the Soviet Block did not fool too many observers. Already in the 1950s, the term 'national communism' was coined to capture the developments in Yugoslavia and then it was aptly transferred to describe post-1956 Poland, Hungary, and *Ceaușescu's* Romania in particular (Verdery, 1991). Not unlike the 'varieties of capitalism' literature in the West, experts on Eastern Europe were quick to discern the characteristics of each 'national communist' system. Hungary became known for its relatively open and prosperous economy, Poland for its mild censorship, East Germany for the secret police, Romania for its familial nepotism. After 1989, historians could start examining how much of 'block' there really was in the Soviet Block beyond prima facie political, military, or economic ties and similarities; the more research is carried out, the less the appearances seem to correspond to the underlying reality.

With the Cold War still in full swing in the 1980s, it was difficult for the scholars of the region to avoid a bipolar division of the field into anti- and procommunist camps; often one could not escape being labeled either a 'cold warrior' or a revisionist with cryptocommunist sympathies. This characterization is no longer valid. The revolutions of 1989 were immediately followed by a wave of scholarship that sought to understand why the system collapsed so quickly. The 'history of the present' written by Timothy Garton Ash immediately after the events of 1989 (Garton Ash, 1990) has addressed many puzzles and new equally insightful research soon followed (Maier, 1997). But after the initial astonishment subsided, scholars reversed the question and asked: what made the system survive for so long despite its seeming incompatibility with the region's societies? Thus the earlier scholarship on the establishment of communist power (1945–48) was recently complemented by studies of the Brezhnev era of *zastoy* (stagnation) and looked for deeper sources of the regime's stability. Many young historians study the patterns of consumption (both material and cultural) in communist societies and arrive at intriguing findings highlighting just how little was understood about how the system worked internally on an everyday basis (Mazurek, 2010). Various polls conducted after 1989 consistently show that there exists a large amount of nostalgia for the bygone days, which suggests that the regimes must have enjoyed some real support. Studies of the 'gray economy' demonstrate that reliance on official figures on the planned economy can be misleading as they capture merely one aspect of a more dynamic and complex economic reality.

Already in the eighteenth century, the image of Eastern Europe was often associated with backwardness. Poor sanitary conditions, underdeveloped infrastructure, general poverty, and disorder were impressions often conveyed by eighteenth century 'Enlightened' travelers. In German vocabulary, the term *Polnische Wirtschaft* became synonymous with a badly managed economy. But it was also, among others, Eastern European émigré scholars who were instrumental in developing a comparative method of developmental studies, sometimes given an overall heading of 'modernization theory.' The acute awareness of their country's industrial weakness was one of the factors that propelled the Soviet leadership, and Stalin in particular, to launch a program of rapid state-sponsored industrialization in the 1930s. This method of economic development has then been transplanted onto Eastern Europe, especially between 1948 and 1956. There has been an ongoing debate among historians as to what extent the communist experience can be understood as a modernization process. There is no consensus, but certainly an agreement that any evaluation depends on the country – underdeveloped countries such as Bulgaria or Romania benefiting more, more developed, as Czechoslovakia, less – and the field of human endeavor (culture, gender relations, economic, politics, etc.) in question. Many historians do agree that the dramatic socioeconomic transformation of the 1950s and 1960s could be characterized as a breakthrough in areas such as public education, social security and welfare, gender equality, urbanization, industrialization, etc. The debate continues whether the human cost of these transformations was acceptable while some historians reject the modernization framework altogether. Especially those influenced by postmodern critiques of the modernization paradigm, such as Maria Todorova (1997) or Kate Brown (2004), argue that it is in the often self-serving eye of the beholder that ideas of superior or inferior development originate and should be rejected as both uninformative at best and prejudiced at worst. Scathing critique of the Western 'colonizing' gaze on Eastern Europe has been voiced especially by Todorova, who also expressed well-taken criticism of the idea of Central Europe as a project meant to include some in the elite 'European' club at the exclusion of the others – in particular the Balkans.

World War II is an important chapter in any region's history, but it left a particularly heavy mark in Eastern Europe. In his recent book entitled *Bloodlands*, Timothy Snyder (2010) documents just how much the region squeezed between Russia and Germany suffered, not only during World War II, but also in the 1930s (*Holodomor* in the Ukraine, Stalinist Purges) and after the War. The region is located precisely at the fault line between the two most murderous regimes in history and this basic fact has to be comprehended by anyone who studies it. The role of Eastern Europe as a victim of forces beyond its control has been a recurrent motif in the region's literature and historiography. According to Milan Kundera, the experiences of World Wars and imperial domination molded a unique national subjectivity in the region, a subjectivity constantly aware of its very existence being fragile, in danger, and at the mercy of larger powers. Poland has been portrayed already by early nineteenth century romantic writers as

a 'Christ of the Nations' redeeming the sins of other nations. This tradition has remained very much alive in Polish historiography. While Snyder and other scholars never push their metaphors so far, there exists a tradition of writing about the Polish–Lithuanian Commonwealth and its peoples from a victimized perspective – one of the definitive works on Polish history, written by the British historian Norman Davis, is entitled *God's Playground* (1979).

The history of Ukraine in the twentieth century is often written from a similar perspective – as primarily a victim of two regimes, Soviet and Nazi. Karel Berkhoff in *Harvest of Despair* (2004) provides an in-depth narrative of how the experience of Nazi occupation looked for the 'ordinary' Ukrainian, what were its daily dangers and deprivations. Mark Mazower's (2008) comprehensive survey, *Hitler's Empire*, is the first comparative account of the Nazi occupation in the entirety of Europe. Mazower demonstrates that the term 'occupation' did not entail an even remotely similar set of experiences for 'Western' and 'Eastern' Europeans, with the General Government and the territories of today's Ukraine and Belarus being the most brutalized and exploited of all. This victimization narrative has been recently challenged and problematized, however. Works of Jan Gross (2001), starting with *Neighbors: The Destruction of the Jewish Community in Jedwabne* – a depiction of how local Polish population organized the killing of several hundred of their Jewish neighbors in the summer of 1941 – have made it impossible to hold that local populations were merely passive bystanders. The categories of resistance and collaboration have been recently revised and nuanced by numerous authors and the topics of postwar popular justice, retribution, and memory have seen collaborative efforts by the most renowned authors in the field (Deák et al., 2000).

The intersection of Eastern European and Jewish histories has been a vibrant field of study in the last two decades. While the Holocaust and Jewish studies in general were largely a taboo subject in the first several decades following the war on both sides of the Iron Curtain, these topics have seen a great deal of attention (including some major controversies) recently. Ranging from micro-studies such as Helmut Walser Smith's (2002) work on the small German town of Konitz (now: Chojnice) before the Great War to broad surveys spanning several centuries (Slezkine, 2004), Jewish history and European history are finally put into dialogue with each other. Poland has seen a wave of historically informed cinematography on the issues of Polish role in Jewish extermination – a promising sign showing that is no longer a forbidden subject it had been for half a century. The process of coming to terms with historical memory and issues of guilt and responsibility has been addressed by Charles Maier in his evocative essays (Maier, 1988) and many historians followed suit.

Historians of Eastern Europe, although often preoccupied with some of the 'traditional' topics outlined in this article, have not remained unaffected by recent trends in historical scholarships. The field has seen innovative work in gender (Wingfield, 2006), urban (Thum, 2011), and everyday history (Healy, 2004). Historians have also paid attention to 'subaltern' groups often overlooked in 'traditional' political or international histories: children (Zahra, 2011) and underprivileged ethnic groups such as the Roma. The crimes committed against this particular minority have recently emerged as one of the main topics of scholarship, finally filling a gap in the literature (Lewy, 2000). Environmental history has been on the rise as well (Dawson, 1996). Traditional areas such as imperial history have in consequence received less attention, but there have been some groundbreaking work in this area as well (Plokhy, 2014). It is impossible to do justice to even a small fraction of new scholarship in this short article, but it is beyond doubt that Eastern Europe as a field remains at the cutting edge of historical scholarship and in stimulating dialogue with the rest of the discipline.

The calls to integrate Eastern European history not only with continental, but also global, history, have not remained unanswered. Naturally, the global approach has been most useful for historians of empires. Both the Austro-Hungarian and the Russian Empires have witnessed a sustained effort to reconnect their history with global agenda such as slave trade, global commerce, colonial expansion, or racial relations, issues which had been often marginalized earlier due to the urgency of military, diplomatic, and national questions. Especially the younger generation is keen on showing how the region they study participated in and influenced global movements such as the Renaissance, Enlightenment, birth of mass politics, origin and spread of technological advances, etc. This focus on interconnectedness, exchange, and mutual influence emerges as a dominant motive of the current doctoral works on the region.

Tony Judt's *Postwar* (2005) has arguably been one of the first attempts to put Western and Eastern European history back together into one integral whole. It took one of the greatest minds of the generation to achieve this monumental and (not very long ago) hardly thinkable task. In 1989, the Berlin Wall was a symbolic and all-too-tangible expression of vast rifts splitting the heart of the continent. The last quarter century has witnessed all kinds of integration and interchange of unparalleled scale and intensity. The rapidity of these processes has exceeded all expectations and poses a challenge for historians and for the ongoing European integration project. Writing history is now a truly global enterprise and old nation-state boundaries of historiographies, though still strong, are being diluted every year. The recent crisis in Ukraine (2013–14) did show that some of the Cold War categories still remain valid and that the rift between 'Western' and 'Eastern' Europe is not yet fully leveled. At the same time, it is certain that the model of Eastern European studies as it was conceived in the Cold War era is a thing of the past. This model would not function well in the current political or cultural reality. Many actively resist the Eastern European label and the conditions for Eastern European 'commonalities' no longer exist – to mention just two major reasons. Traditional clichés do not work either. The old language of nationalism and xenophobia seems to win more votes in some 'old' democracies such as France than in the 'perennially nationalistic' Eastern Europe. The differences in political orientation of two countries often put into one category by historians – Poland and Hungary – were rarely greater than in 2014. On the other hand, the twentieth century and the last 25 years are just a fraction of the region's millennial history and historians tend to be able to resist 'pressures of the present' rather well. There is certainly enough of shared history in 'Eastern Europe' to keep the field alive and to what extent it will be perturbed by recent 'divergent' developments remains to be seen.

See also: Area and International Studies: Cultural Studies; Communism, History of; Eastern European Studies: Culture; Eastern European Studies: Society; Nationalism, Sociology of; Socialist Societies: Anthropological Aspects.

Bibliography

Amar, Tarik, 2011. Different but the same or the same but different? Public memory of the Second World War in Post-Soviet Lviv. Journal of Modern European History 9 (3), 373–396.

Applebaum, Anne, 2012. Iron Curtain: The Crushing of Eastern Europe, 1944–1956. Doubleday, New York.

Berend, T.I., 1996. Central and Eastern Europe, 1944–1993: Detour from the Periphery to the Periphery. Cambridge University Press, Cambridge.

Berkhoff, Karel C., 2004. Harvest of Despair: Life and Death in Ukraine under Nazi Rule. Belknap Press of Harvard University Press, Cambridge.

Blackwood, W.L., 2001. Eastern European studies: history. International Encyclopedia of the Social & Behavioral Sciences 3990–3991.

Brown, Kate, 2004. A Biography of No Place: From Ethnic Borderland to Soviet Heartland. Harvard University Press, Cambridge.

Brubaker, Rogers, 1996. Nationalism Reframed: Nationhood and the National Question in the New Europe. Cambridge University Press, Cambridge.

Brzezinski, Zbigniew, 1956. The Permanent Purge: Politics in Soviet Totalitarianism. Harvard University Press, Cambridge.

Case, Holly, 2009. Between States: The Transylvanian Question and the European Idea during World War II. Stanford University Press, Stanford.

Connelly, John, 2000. Captive University: The Sovietization of East German, Czech, and Polish Higher Education, 1945–1956. University of North Carolina Press, Chapel Hill.

Davies, Norman, 1981. God's Playground: A History of Poland in Two Volumes. Clarendon Press, Oxford.

Dawson, Jane I., 1996. Eco-Nationalism: Anti-Nuclear Activism and National Identity in Russia, Lithuania, and Ukraine. Duke University Press, Durham.

Deák, István, Tomasz Gross, Jan, Judt, Tony, 2000. The Politics of Retribution in Europe: World War II and Its Aftermath. Princeton University Press, Princeton.

Frank, Alison, 2005. Oil Empire: Visions of Prosperity in Austrian Galicia. Harvard University Press, Cambridge.

Garton Ash, Timothy, 1990. The Magic Lantern: The Revolution of '89 Witnessed in Warsaw, Budapest, Berlin, and Prague. Random House, New York.

Gross, Jan Tomasz, 2001. Neighbors: The Destruction of the Jewish Community in Jedwabne, Poland. Princeton University Press, Princeton.

Gross, Jan Tomasz, 1988. Neighbors. Revolution from Abroad: The Soviet Conquest of Poland's Western Ukraine and Western Belorussia. Princeton University Press, Princeton.

Healy, Maureen, 2004. Vienna and the Fall of the Habsburg Empire: Total War and Everyday Life in World War I. Cambridge University Press, Cambridge.

Jászi, Oszkár, 1929. The Dissolution of the Habsburg Monarchy. University of Chicago Press, Chicago.

Judson, Pieter M., 2006. Guardians of the Nation: Activists on the Language Frontiers of Imperial Austria. Harvard University Press, Cambridge.

Judt, Tony, 2005. Postwar: A History of Europe since 1945. Penguin Press,, New York.

Kenney, Padraic, 1997. Rebuilding Poland: Workers and Communists, 1945–1950. Cornell University Press, Ithaca.

King, Jeremy, 2002. Budweisers into Czechs and Germans: A Local History of Bohemian Politics, 1848–1948. Princeton University Press, Princeton.

Kusnierz, Robert, 2013. W Swiecie Stalinowskich Zbrodni: Ukraina w Latach Czystek i Terroru (1934–1938) w Obserwacjach i Analizach MSZ Oraz Wywiadu Wojskowego Drugiej Rzeczypospolitej. Wydawnictwo Naukowe Akademii Pomorskiej, Słupsk.

Lewy, Guenter, 2000. The Nazi Persecution of the Gypsies. Oxford University Press, Oxford.

Maier, Charles S., 1997. Dissolution: The Crisis of Communism and the End of East Germany. Princeton University Press, Princeton.

Maier, Charles S., 1988. The Unmasterable Past: History, Holocaust, and German National Identity. Harvard University Press, Cambridge.

Martin, Terry, 2001. The Affirmative Action Empire: Nations and Nationalism in the Soviet Union, 1923–1939. Cornell University Press, Ithaca.

Mazower, Mark, 2008. Hitler's Empire: How the Nazis Ruled Europe. Penguin Press, New York.

Mazower, Mark Salonica, 2005. City of Ghosts: Christians, Muslims and Jews, 1430–1950. Alfred A. Knopf, New York.

Mazurek, Małgorzata, 2010. Antropologia Niedoboru w NRD i PRL: 1971–1989. ATUT, Wroclaw.

Miłosz, Czesław, 1953. The Captive Mind. Knopf, New York.

Naimark, Norman M., 1995. The Russians in Germany: A History of the Soviet Zone of Occupation, 1945–1949. Belknap Press of Harvard University Press, Cambridge.

Pipes, Richard, 1954. The Formation of the Soviet Union, Communism and Nationalism, 1917–1923. Harvard University Press, Cambridge.

Plokhy, Serhii, 2014. The Last Empire: The Final Days of the Soviet Union. Basic Books, New York.

Porter, Brian, 2000. When Nationalism Began to Hate: Imagining Modern Politics in Nineteenth Century Poland. Oxford University Press, New York.

Rothschild, Joseph, Wingfield, Nancy, 2008. Return to Diversity: A Political History of East Central Europe since World War II. Oxford University Press, Oxford.

Schorske, Carl E., 1979. Fin-De-Siècle Vienna: Politics and Culture. Knopf, New York.

Seton-Watson, Hugh, 1945. Eastern Europe between the Wars, 1918–1941. Cambridge University Press, Cambridge.

Shore, Marci, 2006. Caviar and Ashes: A Warsaw Generation's Life and Death in Marxism, 1918–1968. Yale University Press, New Haven.

Slezkine, Yuri, 2004. The Jewish Century. Princeton University Press, Princeton.

Smith, Helmut Walser, 2002. The Butcher's Tale: Murder and Anti-Semitism in a German Town. W.W. Norton, New York.

Snyder, Timothy, 2010. Bloodlands: Europe between Hitler and Stalin. Basic Books, New York.

Szűcs, Jenő, 1983. Three historical regions of Europe. Acta Historica Academiae Scientiarum Hungaricae 29 (2/4), 131–184.

Ther, Philipp, Siljak, Ana, 2001. Redrawing Nations: Ethnic Cleansing in East-Central Europe, 1944–1948. Rowman & Littlefield, Lanham.

Thum, Gregor, 2011. Uprooted: How Breslau Became Wroclaw during the Century of Expulsions. Princeton University Press, Princeton.

Todorova, Mariia Nikolaeva, 1997. Imagining the Balkans. Oxford University Press, New York.

Verdery, Katherine, 1991. National Ideology under Socialism: Identity and Cultural Politics in Ceauşescu's Romania. University of California Press, Berkeley.

Weber, Eugen Joseph, 1976. Peasants into Frenchmen: The Modernization of Rural France, 1870–1914. Stanford University Press, Stanford.

Wingfield, Nancy (Ed.), 2006. Gender and War in Twentieth-Century Eastern Europe. Indiana University Press, Bloomington.

Wolff, Larry, 2010. The Idea of Galicia: History and Fantasy in Habsburg Political Culture. Stanford University Press, Stanford.

Wolff, Larry, 1994. Inventing Eastern Europe: The Map of Civilization on the Mind of the Enlightenment. Stanford University Press, Stanford.

Zahra, Tara, 2011. The Lost Children: Reconstructing Europe's Families after World War II. Harvard University Press, Cambridge.

Eastern European Studies: Society

Julia Szalai, Hungarian Academy of Sciences, Budapest, Hungary
Jeffrey Hass, University of Richmond, Richmond, VA, USA; and St. Petersburg State University, St. Petersburg, Russia

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 3995–4002, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Though meant also as a geographical classification for denoting the region between the River Elbe and the Ural mountains, the concept of 'Eastern Europe' is mainly used to expresses certain historically informed sociocultural differences within the continent. It carries heavy layers of ambivalence: Although the societies embraced by it all are regarded as 'European,' they do not come up 'fully' to the standard of what 'Europeanness' implies. In this ambivalent classification, the implicit norm is the West, and Eastern Europe remains as much Eastern, as it maintains its 'other than Western' character. Thus, the concept has always been deeply imbued with politics. On the part of Eastern Europeans, its inherent political character has manifested itself in continuous attempts to close up the East–West divide through adaptation to the Western norms of the time; on the part of the West, the same politicized attitude expressed itself in the interest-driven recognition/refusal of these Eastern attempts, having immediate consequences on the (re)drawing of the invisible political, economic, and military map of the continent of Europe. Such ambivalences also characterize deeply the current efforts to unite Europe. Moreover, they leave their mark on the internal social development of the region's societies, as well as having serious consequences in the continuation of their post-1989 crisis.

Introduction

'Eastern Europe' is as much a historically shaped sociocultural concept with contents and connotations that have changed markedly over time, as a reality of recurrently shifting geopolitical divisions. The standing aspiration of the region's nations has been their full integration into a unified Europe driven by Western values and norms. Although meant also as a geographical classification for denoting the region between the River Elbe and the Ural Mountains, the concept of 'Eastern Europe' is mainly used to express certain historically informed sociocultural differences within the continent. It carries heavy layers of ambivalence. Although the societies embraced by it all are regarded as 'European,' they do not come up 'fully' to the standard of what 'Europeanness' implies. In this ambivalent classification, the implicit norm is the West, and Eastern Europe remains as much Eastern, as it maintains its 'other than Western' character.

Eastern Europe: A Region with Clear Contours, but Opaque Borders

Customarily, Eastern Europe as a geographical concept is understood to denote the region positioned east from the River Elbe and west from the Ural Mountains. Such a geographical definition meets, however, insurmountable difficulties when translated into societies, nations, and states. As to the western border of the region, the River Elbe runs through Germany. Does this mean that German society is sharply divided into a 'Western European' and an 'Eastern European' segment? Similarly, the eastern border of the region – the Urals – cuts through Russia; does this mean that one half of Russia belongs to 'Eastern Europe,' while the other is part of the 'non-European East?' Further, if such fault lines divide one or more countries of the region, what are the implications of their presence: do they simply carry sociocultural differences of historical and ethnographical interest, or, does their persistence endanger the stability of the affected states, and, beyond, that of the continent itself? In short, if not simply a tool of geographical orientation, what then is the function of making the East–West differentiation within Europe?

Certain sociocultural differences on regional grounds can be observed in all societies. Still, despite the remarkable differences between Northern Italy and Sicily, nobody in our times would question Italian unity, and the belonging of the entire country to the core of the West. Likewise, even the linguistic division of countries like Belgium or Switzerland would not inspire anybody to doubt the integrity and genuinely Western character of these societies. Thus, the ambiguities in identifying the borders between Western and Eastern Europe arise from deeper layers of differentiation. Their essence is neither a neutral geographical classification according to the magnetic poles, nor a definitive and value-loaded gesture of clearly circumscribing 'togetherness' as opposed to 'otherness.' The concept of 'Eastern Europe' represents a mixture of the latter two, and carries the inherent ambivalence of such a combination. With the very same momentum, it signals undivided 'Europeanness,' and also some reservations toward it. Although the societies embraced by the concept are all 'European,' they do not come up 'fully' to the standard of what 'Europeanness' implies. In this ambivalent classification, the implicit norm is the West, and Eastern Europe remains as much Eastern, as it maintains its 'other than Western' character. Thus, the concept has always been deeply imbued with politics. On the part of Eastern Europeans, its inherent political character has manifested itself in continuous attempts to close the East–West divide through adaptation to the Western norms of a given time; on the part of the West, the same politicized attitude expressed itself in the interest-driven recognition/refusal of the

Eastern attempts, having immediate consequences on the (re)drawing of the invisible political, economic, and the military map of the continent of Europe.

The doubtfulness of the borders – both in the harsh physical and in the more subtle sociopsychological sense of the term – has inspired recurrent attempts to refine geographically founded geopolitical classifications by denoting important and 'meaningful' subunits within the region (Todorova, 1997). Probably, the most important of such attempts has been the repeated separation of 'Central Europe' from the rest of 'Eastern Europe.' The concept of Central Europe as a more 'Westernized' entity with coherent sociopolitical and cultural meanings on its own appeared mainly in nineteenth/twentieth-century public debates by writers, political thinkers, and philosophers (Kundera, 1984), but attempts at the political realization of togetherness and simultaneous separation from the rest of the region's nations also came to the surface among the goals of the sweeping political movements of 1848, 1918, and the late 1980s.

However, the ideas of 'Westernizing' Eastern Europe certainly are not the fresh production of modern times. Various grandiose experiments for their realization date back as far as the time after the fall of the Roman Empire. Since those distant days, the history of the region has been accompanied by heated debates over the 'belonging issue,' which, in turn, has gradually become a decisive aspect of the self-perception and identity struggle of the people living on its territory. The East–West divide, despite its opaque character, has proved extremely strong during the past 1000 years of European history, serving a variety of social, political, military, and economic interests both within and between the countries of the European continent. With this historically elaborated comprehensive political meaning of the concept, 'Eastern Europe' has been as much a point of departure for Western self-identification as for the shaping of programs to combat those 'historic illnesses' that have hindered certain societies from becoming parts of Europe proper. In other words, up to the time of writing, 'Eastern Europe' has been considered the embodiment of the major historical failures of the European continent. It has symbolized the stone-hard limitations of creating a Europe without any subdivisions, and has also represented the accompanying feelings of discomfort and frustration. In the course of the century-long failed efforts, these feelings have grown to become self-perpetuating mutual prejudices, which have exhibited an ever-stronger influence on shaping politics of the East–West cohabitation. The persistence of these prejudices cannot be ignored when attempting to understand the difficulties that the current program of European integration (see point 3 below) has to face in the course of its day-to-day realization.

The Historical Roots of Eastern Europe's One-Sided Development

Historians, political thinkers, writers, and philosophers seem to be in full agreement on identifying the emergence of 'the three historical regions of Europe' (Szücs, 1988) in the aftermath of the Barbarian conquest against the late Roman Empire, which resulted in the division of the occupied territories between Europa Occidens, Byzantium, and a distinct transitional territory in between. This latter region – the historical forerunner of Eastern Europe – occupied a halfway position in all senses of the term. It was neither 'Western,' nor 'Eastern,' but an ever-changing mixture of the two. First, the region's geographic placement exposed it to frequent attacks from both of the neighboring big powers, each of whom regarded themselves as the genuine heirs of the late Roman Empire – the Frankish Empire and Constantinople. Second, the shaping of the region's decisive religions, traditions, and ethos was put also under dual pressure from the West and the East alike. As pointed out by the influential work of medievalists and historians of political ideas (Gerschenkron, 1962; Said, 1978; Szücs, 1988; Wolff, 1994), although the people of Eastern Europe took up Christianity early enough in their history, the subsequent lasting Mongol, Ottoman, and Russian invasions revitalized a number of 'Eastern' features in their social relations, institutions, faith, and the running of their daily life. Third, as it has been demonstrated amply by recent studies in comparative history and historical anthropology (Gerschenkron, 1962; Said, 1978; Konrád and Szelényi, 1979), the states of the region as potential actors of modernization played an ambivalent role in the Premodern era: although the late Medieval centuries brought about absolutist reigns similar to those in the West of the time, the actual function of the Eastern European absolutist rule was to reinforce the estate-based power of the landed nobility – while their Western counterparts prepared the soil for the take-off of capitalism, democracy, and modern bourgeois development. The repeated events of military occupations, wars, mass uprisings, and the subsequent hurried efforts at reconstruction recurrently blockaded the crystallization of fundamental institutions of statehood and administration. Instead, these events usually concluded in hasty and forceful attempts of the conqueror to employ central power to eliminate the 'historic losses,' at all social and political cost. Such attempts at forced and inorganic modernization, however, caused lasting deformations in the sociopolitical order that have had a far-reaching impact on the development of the affected societies.

In assessing the potential and limitations of corrective societal actions, the most dramatic consequences shaping both politics and societies in modern times were discussed extensively by sociologists and political scientists in the 1980s and 1990s (Kochanowicz, 1997; Garton Ash, 1993; Arato, 2000). According to the virtually unanimous conclusions of these studies, the historically informed and persistent deformations of Eastern European societies can be summarized in three main features that clearly distinguish them from their contemporaneous Western counterparts: (1) the one-sided development of nationhood and national identity (which led easily to repeated waves of radical populism and irredentist nationalism); (2) the underdeveloped nature of civil society (which recurrently has provided natural justification for the states' inclination to gain oppressive power over people's needs); and (3) the periodic disappearance or, at best, weak institutionalization of democratic governance, which has been the cause of loose public control over administration and economic management, and which also can be seen as being responsible for the lasting symptoms of social disintegration in the region. Indeed, the hopes that the collapse of communism and the expansion of

the European Union eastward would finally entrench liberal democracy has been tested by contentious politics (some of it centered on ethnic relations) and the rise of political elites intent on consolidating their own power through appeals to a confrontational nationalist rhetoric and fundamental correctness of elite leadership – with Viktor Orbán of Hungary being a case in point.

These historic deformations became sources of dangerous political conflicts around the beginning of the twentieth century. After the failed revolutions of 1848, which, for a century to come, represented the last heroic attempts of the nations of the region to (re)gain national independence and create their state which would guard civil rights, law and order, and free capitalist development, the societies of Eastern Europe faced exclusion from Europe proper, and reacted accordingly – in a frustrated and hostile manner. As historical research has demonstrated convincingly (Konrád and Szelényi, 1979; Schöpflin and Wood, 1989), the outbreak of World War I cannot be explained without considering the preceding unhealthy nationalist rivalry within the borders of the increasingly ineffective, but still oppressive Austro-Hungarian monarchy, and the deep tensions between the cultural–political formations of the communities and their governing states in all the surrounding societies. Again, the outbreak of World War II cannot be explained merely in terms of the clashing political and economic interests of the leading powers of the European continent; the picture remains disturbingly incomplete without taking into account the extremist aspirations of Hitler's Germany and its allies to find a 'final solution' for the century-long diseases of confused cultural and national identity, and also without considering their deeply embedded attempts to claim new territories and borders for their historically 'stolen' nation states (Bibó, 1991). However, not only the two World Wars, but also the 'invention' of the two, diametrically contrasting ways of totalitarian ruling originate from the same roots. As argued in Arendt's and Bauman's studies that have become classics of postwar social science, one can find in their creation the same forms of traumatized, oblique, frustrated, and confused nationalism, and the unstable institutions built on them (Arendt, 1958; Bauman, 1989). While the Nazi way of totalitarian rein put the superiority of race and nationhood at the center of its legitimizing ideology, communist totalitarianism – with its aim of radically transcending the past – denied the role of nationhood as such, and attempted to substitute it with values of internationalism and enforced equality. However, despite the viciously different responses of the two forms of totalitarianism, the emotionally heated attitudes toward nationhood and national identity, the conflictive relationship between the state and people's spontaneously formed communities, the weak observance of rights and law, and, consequently, the enduring harm done to individual freedom remained key problems of Eastern European societies until the end of the twentieth century. True, the modernization of the economies and societies of the region, and the ultimate closing down of the East–West divide belonged to the core of the communist program of the four decades following World War II. However, the domino-like collapse of the Soviet-type regimes in 1989–90 testified sharply to the failures of these attempts, and drove back the societies in question toward the lost track of classical capitalism. Perhaps, the principal evidence of the failures is the fact that with the sweeping events of the late 1980s, the allegedly abolished problems of the interwar years have come to the surface with dramatic intensity: Widespread poverty, intense racism, extreme social marginalization, sharp inequalities of income and wealth, pervasive corruption, and weak legal protection of defenseless communities and individuals are among the most disturbing troubles of Eastern European societies at the time of writing. In order to understand the unfolding of these problems, a brief examination has to be made of the legacy that the communist systems of Eastern Europe have left behind.

Eastern European Societies: From Where to Where?

As a number of recent studies in social and political history have shown, despite the steps taken throughout the 1970s and 1980s toward economic and political liberalization in different countries of the region, the fundamental features of the communist system of Eastern Europe remained unchanged until the late 1980s (Kornai, 1991; Garton Ash, 1993; Arato, 2000). Backed by the military, political, and economic monopoly of the all-embracing Soviet force in the region, the communist party states of the affected countries maintained their power over all the segments of the economy and society, including people's most private matters of career choice, marriage, childbearing, daily consumption, or cultural taste. Aside from the tools of enforcement solely in the hands of the ruling communist parties, the bases for such an extraordinarily extended power were provided by two major deeds: the concentration of 'people's property' in the hands of the party-ruled state, and the institutionalization of the day-to-day control over society through compulsory full employment. As comprehensive studies on the working of the socialist economy have demonstrated (Kornai, 1991; Sik, 1991) the concentration of all the means of production was intended to induce economic rationality through the centrally prescribed production commands driven by the Plan, while compulsory employment was seen as the base for fair shares from distribution and, simultaneously, as an effective means to combat poverty. However, such a structure turned against the initial goals rather quickly: from the early 1960s onwards, the Eastern European communist regimes, one after the other, had to face recurrent deep economic crises, and an increase in social tension originating from the 'dictatorship over needs' (Fehér et al., 1983) that led to nothing else, but the equality of poverty on an unparalleled scale. Thus, the mere maintenance of the regimes required certain concessions: the most important among them were the rehabilitation of the family as a legitimate sphere of people's undisturbed privacy, and the tacit acceptance of the slow evolution of a second, informal economy.

However, these concessions proved to be powerful tools in societies' hands slowly to erode, then effectively to decompose the oppressive structures above them. Eastern European sociological, social–historical, and economic research of the 1970s and 1980s provided a rich literature of these multifaceted processes which, in retrospect, can be regarded as the direct ascendants of the radical social upheavals after 1989–90 (Vajda, 1988; Staniszkis, 1991; Szalai, 1991; Kochanowicz,

1997). The move away from classical socialism was unstoppable. First, ever-spreading participation in the informal economy gradually took a lead in determining the standard and forms of consumption for substantial layers of society; thus, the tacit acceptance of people's informal work and their partial withdrawal from party control, even in the strictly regulated socialist workplaces, was also in the immediate interest of those in power – otherwise, the communist parties would have risked unmanageable crises in production and consumption that could have led to the swift fading of the justification of their 'blessing rule.'

Second, the family-based framework of the informal economy provided a forum for learning a number of market-related skills and thus assisting the underground continuation of the interrupted embourgeoisement process. Such an underground preservation of officially denied values, attitudes, skills, and socioeconomic arrangements later proved crucial in the post-1989 systemic changes of returning to a capitalist/bourgeois track. Third, the informal economies of Eastern Europe complemented the ever-impoverishing state-run welfare provisions with community-based social services, and with slowly elaborated forms of secondary redistribution. Thus, they had a role in correcting the inadequacies of the distribution of incomes driven from the 'first' economy, and also in developing those market-conforming forms of social protection, which then, effectively, assisted the transition process amid the rapid withering away of state distribution some decades later. In short, the informal economy and the social relations around it gradually became a second – mostly hidden – regulator in the functioning of the communist regimes, and an important second pillar in organizing the daily life, well-being, aspirations, and life careers of substantial groups of Eastern European societies.

The emergence of this second system had also serious deficiencies. As sociological research on the malfunctions of the 'socialist' welfare states has evidenced it (Deacon and Szalai, 1990; Kochanowicz, 1997; Atal, 1998; Lessenich, 2000), those who, for a range of varied reasons, could not participate, became increasingly dependent on the state that gradually intensified their defenselessness, instead of mitigating it. With the spreading and tacit acceptance of the regulatory role of the informal economy, the nonparticipating groups were designated as 'misusers' of the public resources, and cuts in assistance for them were approved on a mass scale. Thus, the spreading of participation in the informal economy gave rise to serious social divisions which, amid the troublesome economic conditions of the post-1989 years, eventually became an open war between the 'non-diligent' poor and the 'industrious' nonpoor.

A further shortcoming of the dual system was its weak institutionalization: any claims for the formal acknowledgment of informal work collectives, service providers, interest representations, and the like were seen as open attacks against the principles of the prevailing order, and were treated as political crimes. On these grounds, the informal economy became a battlefield for freedom, but at the same time, the political connotations of participation further hindered its institutionalization. The negative consequences of such a development came to light mainly after the collapse of the old rule. As in-depth studies of the democratic potentials of Eastern Europe have pointed out convincingly (Staniszkis, 1991; Berend, 1993; Arato, 2000), despite widespread earlier practices, civil society has proven to be extremely weak, and the new system could hardly count on its provisions, protective associations, and community-run services amid the deep 'transitional crisis' of the 1990s.

The weakness of civil society has given rise to serious crisis phenomena amid the two intertwined major processes of postcommunist transition: the decomposition of the formerly unlimited power of the central states, and the privatization of the earlier anonymous 'people's properties.' Paradoxically, the earlier resistance to the state as oppressor is now quite clearly being reversed, and while the various corporate bodies and interest alliances only clung to the state distribution policy out of fear and defenselessness, they now make angry claims on it. Behind the opposing principles of privatization intended to 'regulate' the plundering of public assets, one finds intensely competing demands for compensation. Having the arguments justifying these claims accepted and embodied in the legislation is a question of raw political force. As comparative economic studies in the region have demonstrated with ample empirical evidence (Róna-Tas, 1994; Stark, 1996), in this way, privatization and the creation of a bourgeoisie is the direct function of the latent bargaining positions established over past decades. Hence, the process of privatization has led to a jump in uncontrolled and unregulated inequalities of income and wealth, and to the emergence of a new, prosperous elite with little legitimization. At the same time, the process has produced masses of victims at the other end of the scale. Besides depriving them of any access to the newly distributed properties, the easiest compromise to find resources to meet the hunger of the vocal groups has been to cut back sharply on state expenditures on welfare and social services. With these steps, however, Eastern European transition societies have produced vast numbers of chronic poor, whose fate has been extreme social marginalization. Thus, despite the initial hopes of 1989–90, massive poverty and the accompanying syndromes of dramatic social disintegration have not disappeared with the recovery from the crises of the early 1990s, but have become the decisive indicators of a still non-Westernized path of current Eastern European development. The conflicts that arise from this situation have further consequences. The still excessive presence of the state in most of the region's economies, as well as the understandable open aspiration of the earlier overblown state bureaucracy to reinforce its own position of power, at most only help broad strata of societies to reach the recognition that the creation of a private market in the Western sense is a 'foreign' prescription for the process of embourgeoisement in the region – their 'genuine' path still leads through nationalized modernization. Thus, it is hardly surprising that Western and Eastern economists, sociologists, and political scientists analyzing the major sociopolitical processes of postcommunist transition arrived at largely consensual conclusions by the late 1990s (Kochanowicz, 1997; Standing, 1997; Arato, 2000). There seems to be a general agreement that the spreading of anti-West feelings provides a fertile soil for heated nationalism, aggressive racism, and the reemergence of authoritarian tendencies; hence, these recent tendencies in the changes of popular attitudes endanger seriously the once vivid hopes of creating a Europe without

boundaries. However, these unanimous warnings have little impact on actual politics: at best, social science reflects and informs the ongoing attempts of European integration, but can hardly influence it.

Eastern Europe's (Re)turn to Europe

Studies after 1989 suggested the persistence of a sharp East–West divide, much in accordance with sociocultural fault lines explored by earlier historical studies. It seems that in their post-1989 development, Eastern European societies have taken a lead toward increasing social disintegration between their adaptive majorities and the extremely marginalized minorities in long-term poverty. The states of new democracies are squeezed between contradictory challenges of marketization and combating social disintegration at the same time; due to the weakness of their civil societies, they have postponed repeatedly the addressing of the latter problem by creating the required strong welfare systems. As a result, the current social map of Eastern Europe resembles more some Latin American societies than those of the neighboring Western Europe.

Although meant also as a geographical classification for denoting the region between the River Elbe and the Ural Mountains, the concept of 'Eastern Europe' is used mainly to express certain historically informed sociocultural differences within the European continent. It carries heavy layers of ambivalence: although the societies embraced by it all are regarded as 'European,' they do not come up 'fully' to the standard of what 'Europeanness' implies. Such historically informed ambivalence characterizes deeply as well as the current efforts to unite Europe.

Yet contradictions and ambivalence in East European societies might shift in the near future. As of this writing, two forces or events might lead to fundamental changes in identities and interests of East European societies: the crisis of 2008 and Vladimir Putin's annexation of Crimea and destabilization of Ukraine. Regarding 2008, 'East European' countries have felt a sense of secondary status vis-à-vis their western neighbors – but not all of them. In deliberations over responses to the recession, and then to the crisis of the euro spurred by the threat of a Greek default on its debts, the voices of those publics east of Germany were not well heard. Instead, political and economic elites in Germany, Great Britain, and France (in particular, in Germany) for all practical purposes framed the discussions and decisions regarding austerity versus alternative policies, for example. How this Western hegemony in European Union decision making, and the results of such decisions, ultimately impact Eastern European economies, and from there public discourse and public opinion, remains to be seen. However, if such policies can facilitate resentment vis-à-vis the Western countries of the European Union, Russia's foreign policies might ultimately propel 'East Europe' further into the fold of 'Europe' generally. After Russian President Vladimir Putin has essentially annexed Crimea and then (according to much circumstantial evidence) supported separatists and a civil war in the eastern Donetsk region, Poland and the Baltic countries were especially vocal in asking that NATO take a firmer role vis-à-vis possible future Russian threats to their national security. In response, NATO leaders expanded confrontational rhetoric with respect to Russia and proposed that additional military forces be stationed on a rotating basis in these countries. In a sense, the events in Ukraine in 2014, like occupation by the Red Army after World War II, acted as a catalyst bringing the former East Bloc countries closer in terms of identity to the rest of Europe. However, how this continues to play out is difficult to discern, given that events since 2008 have been fluid and, as of this writing, remain clouded and uncertain. We can only conclude that identities, politics, and a sense of self and place in 'East Europe' might remain the same – troubled and ambiguous – or might be clarified in the years to come, depending on whether Cold War I or Détente II results from events further east from Warsaw and Bratislava, and on how the European Union and its central bank respond to the continuing sluggish recovery and unequal distribution of capital and status in economic decision making.

See also: Democratic Transitions; Eastern European Studies: Culture; Eastern European Studies: History; Post-Socialism: Political Economy; Revolutions of 1989–90 in Eastern Central Europe; Socialist Societies: Anthropological Aspects.

Bibliography

Arato, A., 2000. Civil Society, Constitution and Legitimacy. Rowman & Littlefield, Lanham, MD.
Arendt, H., 1958. The Origins of Totalitarianism. Meridian, New York.
Atal, Y. (Ed.), 1998. Poverty in Transition and Transition in Poverty. Berghahn, New York.
Berend, I., 1993. Alternatives of Transition? Center for German and European Studies, University of California, Berkeley, CA.
Bibó, I., 1991. Democracy, Revolution, Self-determination: Selected Writings. Atlantic Research and Publications, Columbia University Press, Boulder, CO.
Bauman, Z., 1989. Modernity and the Holocaust. Cornell University Press, Ithaca, NY.
Deacon, B., Szalai, J. (Eds.), 1990. Social Policy in the New Eastern Europe. Avebury, Aldershot, UK.
Fehér, F., Heller, Á., Márkus, G., 1983. Dictatorship over Needs. Basil Blackwell, London.
Garton Ash, T., 1993. The Magic Latern: The Revolution of '89 Witnessed in Warsaw, Budapest, Berlin Prague. Vintage Books, New York.
Gerschenkron, A., 1962. Economic Backwardness in Historical Perspective. Harvard University Press, Cambridge, MA.
Kochanowicz, J., 1997. Incomplete demise: reflections on the welfare state in Poland after communism. Social Research 64, 1445–1471.
Konrád, G., Szelényi, I., 1979. The Intellectuals on the Road to Class Power. Harcourt Brace Jovanovich, New York.
Kornai, J., 1991. Anti-Equilibrium: On Economic Systems Theory and the Tasks of Research. A. M. Kelley, Fairfield, NJ.
Kundera, M., 26 April 1984. The Tragedy of Central Europe. New York Review of Books.
Lessenich, S., 2000. The southern image revisited: the dynamics of 'transformation dynamics' in East Central Europe. ECE/ECE 1, 21–37.
Milanovic, B., 1998. Income, Inequality and Poverty during the Transition from Planned to Market Economy. World Bank, Washington.
Róna-Tas, A., 1994. The first shall be last? Entrepreneurship and communist cadres in the transition from socialism. American Journal of Sociology 100 (1), 40–69.
Said, E.W., 1978. Orientalism. Pantheon Books, New York.
Schöpflin, G., Wood, N. (Eds.), 1989. In Search of Central Europe. Barnes & Noble, Totowa, NJ.
Sik, O. (Ed.), 1991. Socialism Today? The Changing Meaning of Socialism. St. Martin's Press, New York.

Standing, G., 1997. The folly of social safety nets: why basic income is needed in Eastern Europe. Social Research 64, 1339–1379.
Staniszkis, J., 1991. The Dynamics of the Breakthrough in Eastern Europe: The Polish Experience. University of California Press, Berkeley, CA.
Stark, D., 1996. Recombinant property in East European capitalism. American Journal of Sociology 101 (4), 993–1027.
Szalai, J., 1991. Some aspects of the changing situation of women in Hungary. Signs 17 (1), 152–170.
Szücs, J., 1988. Three historical regions of Europe. In: Keane, J. (Ed.), Civil Society and the State. Verso, London.
Todorova, M.N., 1997. Imagining the Balkans. Oxford University Press, New York.
Vajda, M., 1988. East-Central European perspectives. In: Keane, J. (Ed.), Civil Society and the State. Verso, London.
Wolff, L., 1994. Inventing Eastern Europe. The Map of Civilization on the Mind of the Enlightenment. Stanford University Press, Stanford, CA.

Eating Disorders

Rebecca J Lester, Washington University, St. Louis, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Eating disorders are complex biopsychosocial conditions that can be fatal if left untreated. Often considered to be products of modern Western societies, syndromes resembling eating disorders have been documented across the globe for thousands of years. This article examines current clinical understandings of eating disorders (including diagnosis, prevalence, etiology, and treatment) as one particular lens on phenomena that appear to be much more broadly distributed than previously thought, raising important questions about the relationships between culture and epistemology. After reviewing the historical development of the diagnostic categories of eating disorders, it then discusses the current classification system for these conditions. The next section examines the cross-cultural data on eating disorders and arguments regarding the role of acculturation in the development of these conditions. The last section discusses the treatment of eating disorders from a clinical perspective.

Diagnostic Criteria as Cultural Models

Eating disorders – Anorexia Nervosa (AN), Bulimia Nervosa (BN), Binge Eating Disorder (BED), and Eating Disorder Not Otherwise Specified (EDNOS) – are western psychiatric categories for complex, biopsychosocial conditions in which individuals feel compelled to engage in eating behaviors that can be fatal if left unaddressed. They also entail disturbances in self-perception and severe emotional, psychological, and interpersonal distress (American Psychiatric Association, 2013). Long theorized by clinicians and social scientists to be products of modern western culture and capitalist modes of commodification (e.g., Prince, 1985; Bordo, 2004), conditions that look very much like eating disorders have been identified in the historical and cross-cultural literature (e.g., Bynum, 1987; Miller and Pumariega, 2001) leading scholars to begin to question these assumptions (Keel and Klump, 2003). Accepting or refusing food, though the meanings of all the restriction of food and eating to the point of physical compromise or even death in pursuit of social, moral, or spiritual ends, is by no means unique to the postindustrialized West, though the specific understandings of such behaviors may be different than what we today associate with these conditions. Such concerns are highly culturally informed, and are also inflected by gender, age, ethnicity, and specific life events and relationships. For this reason, patterning of such behaviors can become evident, particularly in contexts where they have been marked as a recognizable state or condition, producing a looping effect (Hacking, 1986). It is, however, a mistake to take one such cultural and historical patterning as isometric with the condition itself, with other patternings mere reflections of its more authentic expression. Eating disorders as they are currently understood in Western psychiatric classification are more appropriately viewed as one possible articulation of a set of concerns that can assume different forms of expression in different contexts.

Historical Accounts of Disordered Eating

Unusual eating practices and elaborate meaning systems to account for these behaviors are not unique to contemporary Western culture. Throughout recorded history, syndromes involving unusual food behaviors – coupled with culturally meaningful preoccupations about body, self, and social perception – have emerged, peaked, and dissipated.

One of these peaks occurred in medieval Europe. During the twelfth through fifteenth centuries, increasing numbers of Catholic women religious gained widespread notoriety for their extreme food asceticism and other ascetic body practices. Similar to people diagnosed with AN and BN today, these women (including such figures as Theresa of Avila and Catherine of Sienna) reportedly ate very tiny amounts of food or refused to eat at all, eventually living on no sustenance but the Eucharist for months or even years at a time (Bynum, 1987; Schleif and Schier, 2009; Mueller, 2010). When urged to eat by others they would refuse, citing desires to share in Christ's suffering by denying themselves food and other bodily comforts. If forced to eat by superiors, some would steal off afterward and induce vomiting. These fasting nuns claimed that, through their self-induced suffering, they were able to achieve direct communication and communion with Christ. While aspects of these reports were likely exaggerated by hagiographic chroniclers, significant evidence does exist from the journals of these women and their confessors to substantiate that their eating habits were austere in the extreme and were considered highly unusual and even bizarre by local standards.

Ascetic movements during the same period were not confined to Christianity, but also emerged in Hindu, Buddhist, Jain, Jewish, and Muslim communities, with some ascetic traditions evident significantly earlier than the Middle Ages (Kaelber, 2006). This suggests that the practice of purposefully denying oneself necessary nourishment as a means of spiritual, social, and/or personal transformation is not proprietary to one specific culture or religious doctrine, and may instead reflect a common human strategy of meaning making.

Another significant peak in reports of eating disturbances occurred in the nineteenth century. Across Europe and North America, reports of 'miraculous maids' or 'fasting girls' began to emerge. These young women, most from rural areas and generally between the ages of 13 and 20, gained notoriety for

their ability to remain alive despite a refusal to eat. Many came to enjoy a sort of celebrity status for their fasting, with their names and photographs appearing regularly in newspapers around the country. The attention and gifts showered upon these girls and their families as a result of this notoriety led some to become suspicious of the motivations of the girls' fasting behaviors. In some cases, medical doctors were called in to perform rigorous examinations of the girls to determine the veracity of their claims. In at least one instance, this resulted in the girl's death. The phenomenon of the 'fasting girls' was thought to be a form of hysteria, and subsided shortly after Gull and Lesegue first 'discovered' anorexia nervosa (Brumberg, 2000).

Other accounts of eating disturbances that have behavioral similarities to current eating disorders can be found as far afield as seventeenth-century Japan (*fushokubyo*, or 'non-eating illness', was prevalent among women and the condition was thought to have a psychological origin), early dynastic China (where medical scrolls describe ailments very similar to anorexia and bulimia), and in the tribal lore of parts of Africa. The clinical and scholarly perception that eating disorders are a modern Western phenomenon, then, may be due as much to the professionalization of Western medicine as a discipline and the emergence of specialized fields within it, as with the emergence of a fundamentally new syndrome uniquely produced by the cultural conditions of the industrialized West.

AN was one of the earliest conditions to fall under the special purview of the emergent field of professional psychiatry. Defined as a distinct clinical condition in 1863 by British physician William Gull (and independently around the same time by Lesegue in France), a hallmark of this condition was, specifically, the 'nervosa' aspect of it, the fact that no physiological cause could be identified for the lack of appetite or refusal to eat. In later years, such behavior and associated thoughts and feelings about food, body, sex, relationships, and morality became covered with the broader umbrella of hysteria (of which AN was considered to be a specific articulation). What we now call bulimia nervosa BN was generally considered to be part of the AN syndrome and was not recognized as a separate diagnosis (Brumberg, 2000).

As the field of psychiatry developed and sought legitimacy within the medical community, standardized diagnostic frameworks emerged, most notably the Diagnostic and Statistical Manual of Mental Disorders (DSM), whose first edition was published in 1952. Now in its fifth edition, the DSM is the compendium of criteria by which mental health practitioners in the United States (and, increasingly, around the globe) identify, diagnose, and treat psychiatric conditions. AN appeared in the first edition of the DSM and, although the notion of hysteria had largely fallen out of favor by that time, many of understandings of the condition, assumptions about its etiology and epidemiology, and views of treatment that developed during the hysteria years persisted in the new psychiatric nosology, eventually being so embedded in how we built knowledge about the condition as to be naturalized in our current approaches (Brumberg, 2000). In 1979, with the publication of the DSM-III, BN was identified as a separate clinical condition to AN; its diagnostic criteria have also undergone significant revisions since then. BED was first identified in the clinical literature in 1959, but was not included in the DSM until the most recent version, the DSM-V, published in 2014. EDNOS is a category used to capture disordered eating that does not meet the strict diagnostic criteria of AN, BN, or BED.

Current Western diagnostic criteria for eating disorders, then, have emerged from over 150 years of Western clinical practice and research. They have been honed and refined in order to capture with increasing specificity what are believed by psychiatrists to be the behavioral, psychological, medical, and social dimensions of these complex illnesses, based on the specific cases available to them for investigation and treatment. As such, they are cultural models emanating from Western psychiatry.

Anthropologists and other social scientists have argued persuasively that the DSM is not a culture-free scientific document, but instead articulates a particular modern Western ethnopsychology that values self-mastery, emotional regulation, and autonomy (e.g., Gaines, 1992; Kirmayer, 2007). This view has been further developed by cross-cultural researchers investigating issues related to psychic and emotional distress (Guarnaccia and Rogler, 1999; Kirmayer, 2001; Kleinman, 1988; Hinton and Good, 2009). While DSM behavioral criteria may be imposed on conditions of distress in other contexts, the meanings, motivations, and experiences of these behaviors may be radically different from those reported by North American subjects, upon whose illness expressions the criteria are based. Similarly, DSM criteria may miss important features and associated practices, beliefs, and experiences that are central to the condition in other contexts but not considered relevant for DSM diagnosis. Taken together, this significant body of research demonstrates that current DSM definitions of mental disorders (including eating disorders) are strongly culturally informed, reflect the social conditions and cultural concerns of late twentieth and early twenty-first-century North America, and may not be seamlessly generalizable to other contexts.

Nevertheless, the conditions described in the DSM are not entirely arbitrary. While they may not describe universal disease conditions, they do describe *something*, and those descriptions carry epistemological and practical weight for clinicians, researchers, and individuals themselves who fall into these diagnostic categories. They also attempt to capture the very real suffering of people whose lives have become unbearable. However culturally laden these categories are, what they strive to articulate is based on the experiences of thousands of people who have presented with similar symptom patterns. While we cannot take these experiences to be representative of all individuals, to dismiss them would be similarly discounting of those upon whom these categories have been based.

Current Diagnostic Categories

Anorexia Nervosa

The core diagnostic features of AN according to the DSM V are that the individual (1) restricts caloric intake leading do significantly low body weight (generally considered to be 85% below

ideal body weight for age, sex, and height), (2) has either an intense fear of gaining weight or of becoming fat, or persistent behavior that interferes with weight gain (even though significantly low weight); and (3) shows a disturbance in the way one's body weight or shape is experienced, undue influence of body shape and weight on self-evaluation, or persistent lack of recognition of the seriousness of the current low body weight (American Psychiatric Association, 2013). Amenorrhea in postmenarcheal females was a diagnostic criterion for AN in previous versions of the DSM IV but has been removed in the DSM V, as it did not apply to postmenopausal women, women on oral contraceptives that interrupt the menstrual cycle, or males (**Table 1**).

The current diagnostic model subdivides AN into two subtypes: restricting type and binge-eating/purging Type. In the restricting type, individuals restrict food intake and do not engage in any kind of bingeing or compensatory behavior, such as purging, laxative use, or excessive exercise. In the binge-eating/purging type, individuals regularly engage in either bingeing or purging (or both) during the current episode of anorexia. They are not considered bulimic primarily because of their low weight (below 85% of ideal body weight or IBW) (American Psychiatric Association, 2013).

Individuals with AN frequently exhibit symptoms of low mood, difficulty sleeping, hopelessness, guilt, restlessness, social withdrawal, irritability, and difficulty concentrating, all of which are consistent with the symptoms of Major Depressive Disorder. It is not always clear whether these features indicate an underlying mood disorder, or whether they are the effects of starvation itself. The metabolic changes associated with the early phases of the illness may actually ameliorate the depressive symptoms (e.g., individuals may feel increased energy, increased confidence, a sense of focus and purpose), thereby reinforcing the behaviors. Over time, however, the starvation process loses its mood-elevating effects and, instead, exacerbates (and perhaps amplifies many times over) underlying mood difficulties (Kaye, 2008; Russell, 2013).

Table 1 DSM Criteria for AN

Diagnostic criteria for anorexia nervosa
1.Refusal to maintain body weight at or above a minimally normal weight for age and height (e.g., weight loss leading to maintenance of body weight less than 85% of that expected; or failure to make expected weight gain during period of growth, leading to body weight less than 85% of that expected).
2.Intense fear of gaining weight or becoming fat, even though underweight.
3.Disturbance in the way in which one's body weight or shape is experienced, undue influence of body weight or shape on self-evaluation, or denial of the seriousness of the current low body weight.
4.In postmenarcheal females, amenorrhea, i.e., the absence of at least three consecutive menstrual cycles. (A woman is considered to have amenorrhea if her periods occur only following hormone, e.g., estrogen, administration.)
Specify type
1.Restricting type: During the current episode of anorexia nervosa, the person has not regularly engaged in binge-eating or purging behavior (i.e., self-induced vomiting or the misuse of laxatives, diuretics, or enemas).
2.Binge-eating/Purging type: During the current episode of anorexia nervosa, the person has regularly engaged in binge-eating or purging behavior (i.e., self-induced vomiting or the misuse of laxatives, diuretics, or enemas).

Many individuals with AN have obsessive (pressured) thoughts and/or compulsive (pressured) behaviors, which often extend beyond food and body alone; for example, obsessions about contamination or compulsions about cleaning or symmetry are common. Other common features related to AN include fear or anxiety about eating in front of others, an obsessive focus on clothing sizes (the size '00' is often a coveted goal), perfectionism in all areas of life, inhibited emotional expression, and deep feelings of shame and self-loathing (Lavender et al., 2013; American Psychiatric Association, 2013; Phillips and Kaye, 2011).

AN carries a number of potential risks, as extreme starvation can damage the major organ systems of the body as well as generating a great deal of collateral damage. Leukopenia (low white cell count in the blood) and anemia (low red cell count) are common. Dehydration can lead to elevated blood urea nitrogen (BUN), indicating potential liver damage. Low magnesium and zinc levels are frequently seen in AN patients, attesting to the body's depletion of essential minerals. Elevated levels of amylase may indicate serious pancreatic dysfunction. Females with AN often have low estrogen levels as a result of the illness, whereas males with AN tend to have lowered testosterone levels. Bradycardia (low heart rate) and arrhythmias (irregular heat beats) can result from AN, as can a reduction in brain volume. If purging, laxative use, or fluid restriction are part of a client's profile, serious electrolyte imbalances can occur which may lead to heart attack and death (Mehler and Anderson, 2010; Trent et al., 2013).

AN as a full-blown illness is rarely diagnosed before puberty, although many associated features may be present in people as young as 3–5 years old (e.g., concern about being 'fat', restriction of food intake, counting calories). Historically, AN has been documented primarily among white, upper-middle class Caucasian women of West European or North American descent living in industrialized countries. A review of the clinical, historical, epidemiological, and ethnographic literature, however, brings into question these features as being central to the illness itself, suggesting that this stereotype of the 'typical' anorexic patient may reflect sampling bias (e.g., Grilo et al., 2005). Recent community-based studies reveal that AN is present among nonwhite (Regan and Cachelin, 2006; Striegel-Moore et al., 2003; Dohm et al., 2010), nonaffluent (e.g., Rogers et al., 1997), and non-Western (e.g., Malson and Nasser, 2009) populations, and that rates among males women over 40 and males seem to be increasing in recent years (Smink et al., 2012). Whether such findings reflect a shift in the epidemiology of anorexia or changes in diagnostic practices (or both) is unclear; regardless, the stereotype of anorexia as a disease of affluent white Western women no longer holds up.

The incidence rate of AN among females has recently been found to be between 8 and 13 cases per 100 000 persons per year, and the prevalence rate is approximately 0.3% using strict diagnostic criteria (Bulik et al., 2005). AN is therefore more common than autism (Fombonne, 2005). Individuals with a first-degree relative with AN have an increased risk of developing the disorder themselves. Studies of monozygotic twins indicate a significantly higher concordance rate for AN than dizygotic twins, suggesting a strong genetic substrate to

the illness (e.g., Lock and Le Grange, 2011). It remains unclear, however, the degree to which environmental and social factors contribute to these findings.

Typically beginning in adolescence, AN often follows a cyclical pattern of exacerbation and remission which can extend across the life course. With proper treatment, an estimated 50% of individuals with AN recover fully, 33% improve somewhat, and 20% remain chronically ill (Steinhausen, 2002). The mortality rate among people with anorexia has been estimated at 0.56% per year, or approximately 5.6% per decade, which is about 12 times higher than the annual death rate due to all causes of death among females aged between 15 and 24 in the general population (Sullivan, 1995). Perhaps not surprisingly, then, AN has the highest mortality rate of any mental illness (Arcelus et al., 2011).

Bulimia Nervosa

BN is characterized by episodes of binge eating (defined as consuming a large amount of food in a short amount of time, accompanied by a sense of lack of control over eating) followed by compensatory behavior such as self-induced vomiting, excessive exercise, misuse of laxatives or other medications (such as amphetamines) or fasting aimed at 'undoing' the binge in order to prevent weight gain. To meet the criteria for BN, this pairing of bingeing and compensatory behavior must occur, on average, at least twice a week for 3 months. The individual's self-evaluation must also evidence an extreme concern with issues related to body weight and shape (**Table 2**).

Table 2 DSM Criteria for BN

Diagnostic criteria for bulimia nervosa

1.Recurrent episodes of binge eating. An episode of binge eating is characterized by both of the following:

a. Eating, in a discrete period of time (e.g., within any 2-h period), an amount of food that is definitely larger than most people would eat during a similar period of time and under similar circumstances.

b. A sense of lack of control over eating during the episode (e.g., a feeling that one cannot stop eating or control what or how much one is eating).

2.Recurrent inappropriate compensatory behavior in order to prevent weight gain, such as self-induced vomiting; misuse of laxatives, diuretics, enemas, or other medications; fasting; or excessive exercise.

3.The binge eating and inappropriate compensatory behaviors both occur, on average, at least twice a week for 3 months.

4.Self-evaluation is unduly influenced by body shape and weight.

5.The disturbance does not occur exclusively during episodes of Anorexia Nervosa.

Specify type

1.Purging type: During the current episode of bulimia nervosa, the person has regularly engaged in self-induced vomiting or the misuse of laxatives, diuretics, or enemas.

2.Nonpurging type: During the current episode of bulimia nervosa, the person has used other inappropriate compensatory behaviors, such as fasting or excessive exercise, but has not regularly engaged in self-induced vomiting or the misuse of laxatives, diuretics, or enemas.

Often, for individuals with bulimia, binge foods are those deemed 'bad' – salty foods, fats, or sweets. Bulimics might binge and purge up to 10 or 12 h each day. They may take 10, 15, or even 20 laxatives or exercise for 4 h each evening. The shame that accompanies bulimic behaviors fuels a cycle of self-loathing that keeps many sufferers from asking for help or wanting to admit they are struggling (Hackler et al., 2010).

Bulimia in the DSM V is subdivided into two types: purging type and nonpurging type. Purging type BN involves the active extraction of food/food derivatives from the body through such things as vomiting, laxative abuse, and the use of diuretics and enemas. Nonpurging type BN entails compensatory behaviors such as fasting or excessive exercise that do not entail these evacuations. In both subtypes, however, compensatory behaviors meant to offset the caloric intake during binges are present.

Unlike individuals with AN, those with BN tend to be of relatively normal weight. Mood disturbances such as depressed mood, low self-esteem, and hopelessness are common in BN and, as with AN, often predate the onset of the eating disturbance. Anxiety symptoms such as obsessive worry and fear of social situations may also be elevated in individuals diagnosed with BN. Substance abuse in BN appears quite high and may be linked to the desire to control weight (e.g., stimulant abuse) or perhaps modulate mood (e.g., alcohol abuse) (Baker et al., 2010).

Frequent purging of any kind can produce significant electrolyte imbalances (specifically, low potassium, low sodium, and low chloride), which can endanger cardiac function. Recurrent vomiting can also erode the teeth causing cavities, breakage, and discoloration. It can also cause the salivary glands to become visibly enlarged. In some cases, esophageal tears and gastric rupture can occur, causing death. Individuals who abuse laxatives are at increased risk for gastrointestinal complications and may become dependent on laxatives to achieve normal bowel function (Mehler, 2011).

The lifetime prevalence of BN among women is estimated to be 1 to 3%, with the rate of males being approximately one-tenth of that number (Hudson et al., 2007). BN usually begins in late adolescence and often follows a period of caloric restriction or dieting. Like anorexia, recovery appears to occur in a cyclical pattern, with multiple periods of relapse interspersed with periods of improvement. Many individuals with BN continue to exhibit disordered eating or thoughts about food, body, and weight for many years after their overt symptoms subside.

BN, mood disorders, and substance use disorders are significantly increased among first-degree relatives of individuals with BN. As with AN, however, it is not clear from the data to what extent these correlations are due to genetic similarities vs social, environmental, and family systems factors. Similarly, like anorexia, BN has primarily been documented among young, middle-class Caucasian women of Western European or North American descent, but recent studies show notable rates among ethnic minorities in the United States and cross-culturally (e.g., Feinson, 2011), and significantly more research is needed on this topic.

Binge Eating Disorder

BED was first described in 1959 but will not officially be recognized as a distinct condition until its inclusion in the

DSM-5. The most common eating disorder in the United States, BED affects 3.5% of American women and 2% of American men (Hudson et al., 2007).

BED is characterized by recurrent episodes of bingeing and feeling a lack of control over eating, similar to BN. Additionally, the binges must include at least three of the following factors: eating rapidly; eating until uncomfortably full; eating large amounts when not hungry; eating alone out of embarrassment; or feeling disgusted, depressed, or guilty after eating. To meet diagnostic criteria for BED the individual must also feel marked distress about the bingeing and the behavior must occur at least twice a week for 6 months. Unlike BN, in BED no forms of purging are present.

People who are obese and suffer from BED are at significantly increased risk for several potentially life-threatening complications, including heart disease, high blood pressure, type 2 diabetes, sleep apnea, high cholesterol, gall bladder disease, and certain types of cancer (Bulik and Reichborn-Kjennerud, 2003). Due to its recent inclusion as an official DSM diagnosis, research to determine the associated disorders, risk factors, and social, age, and gender features of BED is in its infancy but is vigorously underway.

Eating Disorder Not Otherwise Specified

EDNOS category is used to identify disordered eating that does not meet the full criteria for AN, BN, or BED. EDNOS has been shown to be equally as dangerous as these other disorders, and more than 50% of all diagnosed eating disorders in the United States all into the EDNOS category (Fairburn et al., 2007). It is regular practice in the United States for insurance companies to exclude EDNOS from treatment coverage, in part because its variability in presentation precludes systematic treatment studies.

Cross-Cultural Studies of Disordered Eating Disorders

Syndromes consistent with the DSM criteria for eating disorders have been widely documented outside of the United States and Europe, from Egypt to China, to Fiji, and beyond. A robust literature on cultural, socioeconomic, racial, ethnic, and acculturation influences on eating disorders has grappled with questions of whether, how, and to what degree Western culture is a necessary, if not sufficient, condition for the emergence of eating disorders (Soh et al., 2006). In teasing out the relationships between culture and eating disturbances, researchers have focused on questions of cultural change, acculturation, and motivating constructs.

For example, syndromes that have behavioral overlap with eating disorders may, in fact, be motivated by entirely different concerns than a drive for thinness (as in Japan, where the prevailing concern is modesty, not thinness (Pike and Borovoy, 2004)) or may reflect material realities rather than psychological pathology (as in South Africa, where the behaviors of women found to meet diagnostic criteria for BN were found, upon further investigation, to be due to the scarcity and unpredictability of food availability and the frequent contamination of meat (LeGrange et al., 2004)). Similarly, as Anderson-Fye (2004) has persuasively demonstrated in Belize, local engagements with supposedly eating-disorder-producing cultural change are not passive but are actively negotiated, sometimes producing unexpected outcomes of body confidence and resilience. The relationship between contemporary Western psychiatric understandings of eating disturbances and historical and cross-cultural accounts of similar conditions is complex, raising significant questions about how to best conceptualize these syndromes and how to best respond to them.

Studies of eating disturbances among minority populations within Western countries have produced mixed findings. Some differences have been found in terms of which eating disorder is most prevalent among a particular group, or which beliefs and behaviors seem to be the most elaborated, but most studies conclude there is no significant difference in eating disorders across ethnic groups in Western contexts. Similarly, early studies of eating disorders among immigrants and first-generation individuals suggested that a high degree of acculturation to Western values was associated with higher rates of eating disorders. However, more recent studies have demonstrated that level of acculturation makes little difference in risk for eating disturbances and, in some cases acculturation may in fact reduce the risk of eating disturbances (Levine and Smolak, 2010).

Eating Disorders and Cultures in Transition

Another perspective on the role of culture in eating disorders holds that it is not Western culture per se that causes eating disorders; rather, it is the condition of experiencing and living within the difference *between* two cultures that generates risk. Specifically, it is thought that the mismatch between a so-called 'traditional' culture and an adopted (i.e., Western) culture may increase the likelihood of body image disturbances in some individuals, putting them at risk for developing an eating disorder. As Anderson-Fye (2004) has demonstrated, however, cultural change need not always produce a sense of crisis, especially among adolescents. Furthermore, she argues, we must be careful when generalizing about responses to this change, as historical, economic, and cultural factors particular to a given locale will significantly shape how cultural difference is experienced and understood.

Motivating Constructs in Eating Disorders

A key diagnostic feature of eating disorders as defined in the DSM is "an intense fear of gaining weight or becoming fat" (American Psychiatric Association, 2013). While fat phobia is associated with eating disturbances in Western and Western-influenced societies, this is by no means universal. Cross-culturally, syndromes that otherwise meet the behavioral criteria of the DSM do not always contain this component, with motivations for self-starvation ranging from modesty to religious motivations, to gastrointestinal complaints. This finding suggests that the DSM criteria may capture one particular expression of a condition that may manifest in different ways depending on cultural context. A growing number of researchers support the development of a category

of 'nonfat-phobic anorexia' (NFP-AN) in order to capture this diversity (Becker et al., 2009).

Treatments for Eating Disorders

The relationship between contemporary Western psychiatric understandings of eating disturbances and historical and cross-cultural accounts of similar conditions is complex, raising significant questions about how to best conceptualize these syndromes and how to best respond to them. Current treatment models have been developed within contexts saturated with Western cultural systems, values, and conceptual categories (Gremillion, 2003), based on knowledge gathered primarily from clinical populations; that is, individuals who have come to the attention of the medical community, been diagnosed in a certain way, and have the resources to obtain specialized care. Given this, it would seem reasonable to suspect that prevailing definitions, psychological correlates, and age/gender features of eating disorders based on a very narrow demographic may not be seamlessly generalizable to other populations. While this limitation does not undermine the value or validity of these findings for the contexts in which they were developed, it does raise serious concerns about their applicability in other populations and in other cultural settings.

In recent years, major strides have been made by clinicians in the United Sates and Europe to develop, evaluate and disseminate effective, evidence-based interventions for eating disorders and to outline standards of care. Cognitive-Behavioral Therapy (CBT), Interpersonal Therapy (IPT), and Dialectical-Behavioral Therapy (DBT) have been shown to be effective in treating BN and BED in adults (Wilson and Shafran, 2005), and family-based therapy (FBT) appears to be effective in treating AN and BN in children and adolescents (Lock and Le Grange, 2012). Having access to a full spectrum of care – inpatient hospital-based care, residential treatment, and outpatient therapy – and treatment that is consistent with current American Psychiatric Association guidelines has been shown to be critical for the successful treatment of eating disorders (Wilson and Shafran, 2005). The current high mortality and morbidity rates associated with eating disorders may be due in significant part to the difficulties individuals face in accessing and obtaining appropriate and complete treatment (Silber and Robb, 2002).

However, it is not clear that this is the best course of treatment for eating disorders or eating-disorder-like conditions in other cultural contexts, nor is it clear that the therapeutical approaches found to be successful in the United States will be similarly helpful elsewhere. Syndromes that have behavioral overlap with eating disorders may, in fact, be motivated by entirely different concerns than a drive for thinness, or may reflect material realities rather than psychological pathology. Lester (2007), for example, found that, despite the use of a common therapeutic language, eating disorder clinicians in Mexico City rejected American interpretations of the core causes of eating disorders, and found American treatment approaches to be harmful rather than helpful. In other words, the development of diverse treatment approaches for eating disorders is lagging far behind the recognition of the diversity of populations who struggle with these conditions.

Conclusion

Eating disorders and similar conditions continue to intrigue researchers, practitioners, academics, and the general public alike. Current research suggests that the DSM criteria for these conditions capture particular cultural expressions of a phenomenon that has existed throughout history and cross-culturally in different forms. Additional research is needed to continue to tease apart the relationships among culture, practice, interpersonal, and subjective experiences that facilitate the use of self-starvation and other harmful eating disturbances as patterned responses to distress.

Whatever the cultural construction of Western diagnostic categories, it is critical to remember that they do capture, however partially and imperfectly, real suffering in the world. The conditions currently called eating disorders generate enormous distress for individuals who manifest them and those who care for them. They cause severe medical, psychological, and social compromise, and they can be lethal. Examining the cultural dimensions of these conditions and their definitions does not negate the fact that real suffering exists; it merely reminds us of the imperative to continually examine the cultural assumptions embedded in prevailing regimes of knowledge so that we can better understand and respond to this suffering.

See also: Body Image; Body, History of; Body: Anthropological Aspects; Capitalism; Consumption, History of; Emotions, History of; Gender History; Gender and Social Research; Gender and Women's Studies, Applied Research On; Gender: Anthropological Aspects; Identity and Identification; Medicalization: Sociological and Anthropological Perspectives; Medicine, History of Western; Obesity: Cultural and Biological Factors; Person and Self: Philosophical Aspects; Personal Identity: Philosophical Aspects; Professions and Professionalization, History of; Self: Philosophical Aspects; Social Mobility, History of.

Bibliography

American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR, fourth ed. American Psychiatric Association, Washington, DC.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Publishing, Arlington, VA.

Anderson-Fye, E.P., 2004. A "coca-cola" shape: cultural change, body image, and eating disorders in San Andres, Belize. Culture, Medicine and Psychiatry 28 (4), 561–595.

Arcelus, J., Mitchell, A.J., Wales, J., Nielsen, S., 2011. Mortality rates in patients with anorexia nervosa and other eating disorders. A metaanalysis of 36 studies. Archives of General Psychiatry 68 (7), 724–731.

Baker, J.H., Mitchell, K.S., Neale, M.C., Kendler, K.S., 2010. Eating disorder symptomatology and substance use disorders: prevalence and shared risk in a population based twin sample. International Journal of Eating Disorders 43 (7), 648–658.

Becker, A.E., Thomas, J.J., Pike, K.M., 2009. Should non-fat-phobic anorexia nervosa be included in DSM-V? The International Journal of Eating Disorders 42 (7), 620–635.

Bordo, S., 2004. Unbearable Weight: Feminism, Western Culture, and the Body. University of California Press, Berkeley, CA.

Brumberg, J.J., 2000. Fasting Girls: The History of Anorexia Nervosa. Vintage, New York.

Bulik, C.M., Reichborn-Kjennerud, T., 2003. Medical morbidity in binge eating disorder. The International Journal of Eating Disorders 34 (Suppl.), S39–S46.

Bulik, C.M., et al., 2005. Anorexia nervosa: definition, epidemiology, and cycle of risk. The International Journal of Eating Disorders 37 (Suppl.), S2–S9 discussion S20–S21.

Bynum, C.W., 1987. Holy Feast and Holy Fast: The Religious Significance of Food to Medieval Women. University of California Press, Berkeley, CA.
Claes, L., Soenens, B., Vansteenkiste, M., Vandereycken, W., 2012. The scars of the inner critic: perfectionism and nonsuicidal self-injury in eating disorders. European Eating Disorders Review 20 (3), 196–202.
Dohm, F., Brown, M., Cachelin, F.M., Striegel-Moore, R.H., 2010. Ethnicity, disordered eating, and body image. In: Landrine, H., Russo, N. (Eds.), Handbook of Diversity in Feminist Psychology. Springer Publishing Co., New York, US, pp. 285–309.
Fairburn, C.G., et al., 2007. The severity and status of eating disorder NOS: implications for DSM-V. Behaviour Research and Therapy 45 (8), 1705–1715.
Feinson, M.C., 2011. Revisiting the relationship between eating disturbances and cultural origin: a focus on adult women in 14 community studies. Eating Disorders 19 (4), 335–345.
Fenning, S., Hadas, A., Itzhaky, L., Roe, D., Apter, A., Shahar, G., 2008. Self-criticism is a key predictor of eating disorder dimensions among inpatient adolescent females. International Journal of Eating Disorders 41 (8), 762–765.
Fombonne, E., 2005. Epidemiology of autistic disorder and other pervasive developmental disorders. The Journal of Clinical Psychiatry 66 (Suppl. 1), 3–8.
Gaines, A.D., 1992. From DSM-I to III-R; voices of self, mastery and the other: a cultural constructivist reading of U.S. psychiatric classification. Social Science & Medicine 35 (1), 3–24.
Goss, K., Allan, S., 2012. An introduction to compassion-focused therapy for eating disorders (CFT-E). In: Fox, J., Goss, K. (Eds.), Eating and Its Disorders. Wiley-Blackwell, New York, pp. 303–314.
Gremillion, H., 2003. Feeding Anorexia: Gender and Power at a Treatment Center. Duke University Press, Durham, NC.
Grilo, C.M., Lozano, C., Masheb, R.M., 2005. Ethnicity and sampling bias in binge eating disorder: black women who seek treatment have different characteristics than those who do not. The International Journal of Eating Disorders 38 (3), 257–262.
Guarnaccia, P.J., Rogler, L.H., 1999. Research on culture-bound syndromes: new directions. The American Journal of Psychiatry 156 (9), 1322–1327.
Hacking, I., 1986. Making up people. In: Heller, T.C., Wellerby, D.E., Sosna, M. (Eds.), Reconstructing Individualism: Autonomy, Individuality and the Self in Western Thought. Stanford University Press, Stanford, pp. 221–236.
Hackler, A.H., Vogel, D.L., Wade, N.G., 2010. Attitudes toward seeking professional help for an eating disorder: the role of stigma and anticipated outcomes. Journal of Counseling & Development 88 (4), 424–431.
Hinton, A., Good, B., 2009. Culture and Panic Disorder. Stanford University Press, Palo Alto, CA.
Hudson, J.I., Hiripi, E., Pope, H.G., Kessler, R.C., 2007. The prevalence and correlates of eating disorders in the national comorbidity survey replication. Biological Psychiatry 61 (3), 348–358.
Kaelber, L., 2006. Schools of Asceticism: Ideology and Organization in Medieval Religious Communities. Pennsylvania State University Press, Philadelphia, PA.
Kaye, W., 2008. Neurobiology of anorexia and bulimia. Physiology & Behavior 94 (1), 121–135.
Kirmayer, L.J., 2001. Cultural variations in the clinical presentation of depression and anxiety: implications for diagnosis and treatment. The Journal of Clinical Psychiatry 62 (Suppl. 1), 22–28 discussion 29–30.
Keel, P.K., Klump, K.L., 2003. Are eating disorders culture-bound syndromes? Implications for conceptualizing their etiology. Psychological Bulletin 129 (5), 747–769.
Kleinman, A., 1988. The Illness Narratives: Suffering, Healing, and the Human Condition. Basic Books, New York.
Kirmayer, L.J., 2007. Psychotherapy and the cultural concept of the person. Transcultural Psychiatry 44 (2), 232–257.
Lavender, J.M., De Young, K.P., Wonderlich, S.A., Crosby, R.D., Engel, S.G., Mitchel, J.E., Le Grange, D., 2013. Daily patterns of anxiety in anorexia nervosa: associations with eating disorder behaviors in the natural environment. Journal of Abnormal Psychology 122 (3), 672–683.
Le Grange, D., Louw, J., Breen, A., Katzman, M., 2004. The meaning of "self-starvation" in impoverished black adolescents in South Africa. Culture, Medicine and Society 28, 439–461.
Lester, R., 2004. Commentary: eating disorders and the problem of "culture" in acculturation. Culture, Medicine and Psychiatry 28 (4), 607–615.
Lester, R.J., 2007. Critical therapeutics: cultural politics and clinical reality in two eating disorder treatment centers. Medical Anthropology Quarterly 21 (4), 369–387.
Lester, R.J., 2009. Brokering authenticity: borderline personality disorder and the ethics of care in an American eating disorder clinic. Current Anthropology 50 (3), 281–302.
Levine, M., Smolak, L., 2010. Cultural influences on body image and the eating disorders. In: Agras, W.S. (Ed.), The Oxford Handbook of Eating Disorders, Oxford University Press, New York, pp. 223–246.
Lock, J., Le Grange, D., 2011. Eating Disorders in Children and Adolescents: A Clinical Handbook. The Guilford Press, New York.
Lock, J., Le Grange, D., 2012. Treatment Manual for Anorexia Nervosa: A Family-Based Approach, second ed. The Guilford Press, New York.
Malson, H., Nasser, M., 2009. Beyond Western Dis/orders: Thinness and Self-starvation of Other-Ed Women. Routledge, New York.
Mehler, P.S., 2011. Medical complications of bulimia nervosa and their treatments. International Journal of Eating Disorders 44, 95–104.
Mehler, P.S., Anderson, A.E., 2010. Eating Disorders: A Guide to Medical Care and Complications, second ed. Johns Hopkins University Press, Baltimore, MD.
Miller, M.N., Pumariega, A.J., 2001. Culture and eating disorders: a historical and cross-cultural review. Psychiatry 64 (2), 93–110.
Mueller, J., 2010. The Privilege of Poverty: Clare of Assisi, Agnes of Prague, and the Struggle for a Franciscan Rule for Women. The Pennsylvania State University Press, Philadelphia, PA.
Phillips, K.A., Kaye, W.H., 2011. Relationship of body dysmorphic disorder and eating disorders to obsessive-compulsive disorder. In: Hollander, E., Zohar, J., Sirovatka, P.J., Regier, D.A. (Eds.), Obsessive-Compulsive Spectrum Disorders: Refining the Research Agenda for DSM-V. American Psychiatric Association, Arlington, VA, pp. 33–56.
Pike, K.M., Borovoy, A., 2004. The Rise of Eating Disorders in Japan: issues of culture and limitations of the model of "Westernization"? Culture, Medicine and Psychiatry 28 (4), 493–531.
Prince, R., 1985. The concept of culture-bound syndromes: anorexia nervosa and brain-fag. Social Science & Medicine 21, 197–203.
Regan, P.C., Cachelin, F.M., 2006. Binge eating and purging in a multiethnic community sample. International Journal Of Eating Disorders 39 (6), 523–526.
Roehrig, J.P., McLean, C.P., 2010. A comparison of stigma toward eating disorders versus depression. International Journal of Eating Disorders 43, 671–674.
Rogers, L., Resnick, M.D., Mitchell, J.E., Blum, R.W., 1997. The relationship between socioeconomic status and eating-disordered behaviors in a community sample of adolescent girls. International Journal of Eating Disorders 22, 15–23.
Russell, J., 2013. What is an eating disorder? The importance of nourishment controlling behaviours in the service of self-worth. The Australian and New Zealand Journal of Psychiatry 47 (6), 581–582.
Schleif, C., Schier, V., 2009. Katerina's Windows: Donation and Devotion, Art and Music, as Heard and Seen in the Writings of a Birgittine Nun. Pennsylvania State University Press, Philadelphia, PA.
Silber, T.J., Robb, A.S., 2002. Eating disorders and health insurance: understanding and overcoming obstacles to treatment. Child and Adolescent Psychiatric Clinics of North America 11 (2), 419–428.
Smink, F.R.E., van Hoeken, D., Hoek, H.W., 2012. Epidemiology of eating disorders: incidence, prevalence and mortality rates. Current Psychiatry Reports 14 (4), 406–414.
Soh, N.L., Touyz, S.W., Surgenor, L.J., 2006. Eating and body image disturbances across cultures: a review. European Eating Disorders Review 14 (1), 54–65.
Steinhausen, H.-C., 2002. The outcome of anorexia nervosa in the 20th century. The American Journal of Psychiatry 159 (8), 1284–1293.
Striegel-Moore, R.H., Dohm, F.A., Kraemer, H.C., Taylor, C.B., Daniels, S.S., Crawford, P.B., Schreiber, G.B., 2003. Eating disorders in white and black women. American Journal of Psychiatry 160 (7), 1326–1331.
Sullivan, P.F., 1995. Mortality in anorexia nervosa. The American Journal of Psychiatry 152 (7), 1073–1074.
Thompson-Brenner, H., Satir, D., Franko, D., Herzong, D., 2012. Clinician reactions to patients with eating disorders: a review of the literature. Psychiatric Services 63 (1), 73–78.
Trent, S., Moreira, M., Colwell, C., Mehler, P., 2013. ED management of patients with eating disorders. The American Journal of Emergency Medicine 31 (5), 859–865.
Wilson, G.T., Shafran, R., 2005. Eating disorders guidelines from NICE. The Lancet 365 (9453), 79–81.

Relevant Websites

http://www.aedweb.org//AM/Template.cfm?Section=Home – The Academy for Eating Disorders.
http://www.eatingdisorderscoalition.org/ – Eating Disorders Coalition for Research, Policy, and Action.
http://www.iaedp.com/ – The International Association of Eating Disorders Professionals Foundation.
http://www.nationaleatingdisorders.org/ – National Eating Disorder Association.
http://www.nimh.nih.gov/health/publications/eating-disorders/complete-index.shtml – National Institute of Mental Health – Eating Disorders.

Eco-cities and Sustainable Urbanism

Simon Joss, University of Westminster, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The 'eco-city,' and related concepts and practices of 'sustainable urbanism,' have since the early 2000s gained growing international popularity and entered mainstream policy as a consequence of the forceful combination of global climate change concerns and a rapidly urbanizing world population. Sustainable urbanism engages with various aspects of environmental, economic, and social sustainability concerning the urban context. Eco-cities are initiatives that variably promote and pursue sustainable development in relation to urban infrastructure, services, and community at district, town, or metropolitan levels. Governance challenges involved include effective coordination of innovation, planning, and development across policy sectors; integration across urban scales; and engagement with stakeholders and communities. The need for global sustainable city frameworks and standards becomes more apparent as both the number of practical initiatives and international cooperation increase.

Introduction

A growing interest in the 'eco-city' – and similar concepts, such as the 'sustainable city,' 'smart city,' 'low-carbon city,' and 'resilient city' – has been matched, since the early 2000s, with the significant global proliferation of a diverse range of related policies and practical initiatives. Recent activity has been particularly notable across Asia, in addition to Europe and North America, from where the eco-city concept first originated in the 1970s and 1980s. Key driving factors behind this development include climate change policy, accelerated urbanization, and globalization processes, which together have placed cities center stage.

The 'eco-city' may be best understood as an umbrella term that covers various notions of, and approaches to, sustainable urbanism, rather than a conceptually coherent and practically uniform phenomenon. It brings together multiple forms of sustainable development (SD) applied at different urban scales and locally contextualized, from 'retrofitting' entire cities to building new 'smart-green' urban neighborhoods, and from focusing on environmental innovations to promoting socio-economic measures. It further brings together diverse actors in often new governance relationships: municipal authorities, national governments, technology firms, international consultants, nongovernmental organizations, and citizens. Nevertheless, underlying this diversity, a few core conceptual and ideological perspectives variably manifest themselves. These include the 'triple bottom line' of sustainability, ecological modernization, and scientific-technological innovation that, together, characterize contemporary eco-city and related sustainable urbanism discourses.

The global popularity of eco-cities of various kinds prompts the need to bring this recent urban planning paradigm into question both conceptually and practically, by inquiring into underlying normative assumptions, comparing various policy and practical approaches, and critically analyzing outcomes. This seems particularly important given the eco-city's relatively untested promise of reconciling environmental protection with urban economic growth.

Eco-City Dimensions

There are currently no standard, internationally applied definitions of what an 'eco-city' is and what constitutes 'sustainable urbanism.' At one level, this is not surprising, given the diversity of concepts and practices set within particular local settings and traditions, and the competition among the various actors involved. At another level, however, this risks rendering eco-city initiatives conceptually vacuous – especially where they are deployed for promotional purposes – as well as undermining (international) comparability and replicability. From an innovation studies perspective, the current phase may be seen as experimental, to be followed by a phase of consolidation that over time may be expected to lead to more established international norms and accepted 'good practice.' Already, several governmental, nongovernmental, and private organizations are vying to define global eco-city standards and frameworks (Joss et al., 2012; Joss, 2012).

Apart from underlying conceptual differences and context-specific characteristics, a further reason why commonly agreed definitions are hard to come by is that the eco-city is inherently future-oriented: while it sets itself clearly apart from the conditions and experience of (most) current cities – striving to become less polluting, low-carbon, economically resilient, and more livable – its precise form and nature remain as yet invisible in the indeterminate future. As Richard Register (1987: 3), one of the originators of the modern eco-city movement, aptly noted in his original work *Ecocity Berkeley: Building Cities for a Healthy Future*: "The eco-city is an ecological city. No such city exists."

Still, at a general level, several (aspirational) sets of 'eco-city dimensions' – as sometimes referred to in the literature – can be discerned (see, e.g., Girardet, 1999; Joss, 2011a; Kenworthy, 2006; Lehmann, 2010; Register, 2006; Roseland, 1997/2012; Wheeler and Beatley, 2004). As noted in this article, these elements of sustainable urbanism are not applied equally across initiatives. Some initiatives are characterized by an explicit focus on a limited number of key goals – say, carbon-neutral, renewable-energy urban developments – while others deliberately emphasize a broader, multidimensional approach.

One overarching definition of the eco-city is that of 'an ecological city as economic city,' as illustrated by the World Bank's *Eco² City* initiative (World Bank, 2010). Here, following the concept of ecological modernization, ecological and economic objectives are seen as intertwined and mutually dependent. Other definitions, closely aligned with the concept of SD, emphasize the social dimension of the eco-city – with reference to issues of social justice, equity, well-being, participatory governance, and so on – alongside ecological and economic dimensions. In such cases, the term 'sustainable city' tends to be to the fore.

Sustainability 'Triple Bottom Line' in the Urban Context

Eco-city dimensions may be usefully categorized in terms of the 'triple bottom line' of sustainability (environmental, economic, and social), complemented by city-specific dimensions, including urban design, systems, and governance. Table 1 lists prevalent eco-city dimensions along these categories.

At an individual level, these dimensions are typically specified in detail, with targets set in relation to current baselines and with particular time frames in mind. For example, China's Tangshan Caofeidian International Eco-city, under construction since 2009, defines these dimensions through a set of 141 'eco-city indicators' (Joss and Molella, 2013). This includes, among others, targets for 95% energy demand to be met from on-site renewable energy sources; 60% overall waste recycling, with less than 10% waste going to landfill; and – reflecting current levels of water pollution and water scarcity – some 20 targets for water preservation and recycling. Sixty eight of the indicators apply to the city as a whole, with 16 operating at neighborhood, and 25 at site levels.

Another example, *Greenest City: 2020 Action Plan* (City of Vancouver, 2010), includes 10 high-level dimensions (Table 2). Each dimension is further defined through one to two key targets or indicators (some quantitative and others qualitative) set against specified baselines and including strategic action plans. Following in the footsteps of its earlier *EcoDensity Charter* (unveiled in 2008 under the slogan 'envisioning the eco-city'), the City of Vancouver launched 'Greenest City' in 2010 as a flagship initiative aimed at propelling the city to 'most sustainable city' status worldwide by 2020.

Types of Urban Development

Eco-city dimensions can also be defined in relation to the types of urban development (Joss, 2011a; all examples of eco-city initiatives mentioned here and throughout the article are referenced in Joss et al., 2011). A useful distinction can be made between (1) new developments – that is, entire urban centers built from scratch, such as Masdar (United Arab Emirates (UAE)), the self-proclaimed world's first zero-carbon,

Table 1 Eco-city dimensions: general dimensions and related targets, based on a sustainability 'triple bottom line' applied to the urban context

Environmental sustainability	CO_2/GHG	Reducing GHG emissions/energy consumption; promoting renewable energy generation
	Buildings	Reducing energy use; promoting carbon-neutral buildings
	Water	Reducing water consumption; improving water recycling
	Transport	Promoting integrated public transport
	Waste	Reducing waste; increasing waste recycling and waste-to-energy generation
	Biodiversity	Protecting green spaces; promoting biodiversity
Economic sustainability	High-skilled, 'green' jobs and employability	Investing in the knowledge economy; attracting 'green' business
	Competitiveness and resilience	Enhancing international competitiveness; promoting community self-reliance
	Smart technology	Promoting 'smart' technological innovation
	Well-being	Improving individual and social well-being
	Housing	Providing affordable housing and mixed usage
	Urban agriculture	Promoting urban agriculture and local food networks
Social sustainability	Livability	Creating socially inclusive, vibrant communities
	Equity	Fostering equity within and between generations
	Civic engagement	Encouraging participation in public life
	Cultural diversity	Promoting cultural engagement and diversity
Urban design and systems	Compactness	Increasing density of housing, along transport arteries
	Scale	Improving integration across neighborhood, district, city, and regional levels
	Ecosystems management	Improving information management of urban infrastructure and systems (water, energy, etc.)
Urban governance	Policy coordination	Enhancing multilevel policy coordination
	Public–private partnerships	Promoting partnerships involving public, private, and nongovernmental actors
	Political accountability	Ensuring transparency, openness, and participation; collaborative planning

CO_2: carbon dioxide; GHG: greenhouse gas.

Table 2 Vancouver's Greenest City 2020 dimensions and indicators

Eco-city dimension	*Indicators or targets*	
1. Green economy	Double the number of green jobs (over 2010 levels) by 2020	Double the number of companies actively engaged in greening operations (over 2011 levels) by 2020
2. Climate leadership	Reduce GHG emissions produced by residents by 33% from 2007 levels	
3. Green buildings	Require all buildings constructed in 2020 and later to be carbon neutral in operation	Reduce energy use and GHG emissions in existing buildings by 20% over 2007 levels
4. Green transportation	Over 50% of trips to be by foot, bicycle, and public transport	Reduce average distance driven per resident by 20% from 2007 levels
5. Zero waste	Reduce the solid waste going to the landfill or incinerator by 50% from 2008 levels	
6. Access to nature	All Vancouver residents to live within a 5-minute walk of green open space by 2020	Plant 150 000 new trees by 2020
7. Lighter footprint	Reduce Vancouver's ecological footprint by 33% over 2006 levels	
8. Clean water	Meet or beat the strongest of British Columbian, Canadian, and appropriate international drinking water quality standards and guidelines	Reduce per capita water consumption by 33% from 2006 levels
9. Clean air	Always meet or beat the most stringent air quality guidelines from Metro Vancouver, British Columbia, Canada, and WHO	
10. Local food	Increase city-wide and neighborhood food assets by a minimum of 50% over 2010 levels	

GHG: greenhouse gas; WHO: World Health Organization.
Source: http://vancouver.ca/files/cov/Greenest-city-action-plan.pdf.

zero-waste new city, whose initial phase was completed in 2010; (2) urban 'infill' developments, typically in the form of new neighborhoods or districts built on brownfield sites, such as 'Treasure Island' in San Francisco (USA); and (3) 'retrofit' developments aimed at regenerating and upgrading existing urban areas. The latter may take place at the district level, such as the 'Bo01' harbor district in the city of Malmö (Sweden), or at the city-wide level, as illustrated by Kawasaki, the first of a series of Japanese cities designated as 'eco-towns' by the national government. This distinction, too, underscores the diversity of the eco-city phenomenon: there is a significant difference (conceptual, physical, and technological) between entire, large-scale eco-cities built from scratch and eco-city innovation taking place within existing districts and towns.

Analytically, the eco-city can be usefully demarcated in relation to three overarching criteria: (1) scale – an eco-city refers to an urban sustainability development of significant scale in terms of area, infrastructure, and innovation concerned; (2) sectors – eco-city innovation typically takes place across sectors (land use, transport, energy, water, waste, housing, the public realm, etc.); and (3) policy – eco-city initiatives are underpinned by policy initiatives and processes (Joss, 2011a: 278–279). An eco-city initiative may be more or less significant, depending on the extent to which these criteria jointly apply. Concerning the issue of scale, a useful distinction can be made between intervention at the neighborhood, district, town or city, and urban-regional levels. From a planning perspective, ideally, integration should occur across these levels, 'from city to street' and 'from city to region.' This also underscores the importance of relevant governance structures (e.g., metropolitan government), policy processes (e.g., strategic plans), and planning tools (e.g., master plans) to support such multilevel and intersectoral coordination.

Defining and designing eco-city dimensions and sustainable urbanism features are in many ways challenging processes: first, they require a balancing act between tailoring these to the immediate local context – to achieve place-specificity – and making them more generally compatible with national and international policy and regulatory norms as well as replicable across individual sites. Second, it requires the integration of individual dimensions and indicators in order to avoid atomistic lists of goals and targets and to achieve an overarching interaction across urban systems and policy sectors. Third, the process should include transparent information about normative assumptions made and data sources used when defining eco-city dimensions, and about the subsequent capture, analysis, and assessment of data to measure progress toward agreed targets. Furthermore, the process relies on the input of multiple epistemic communities (engineers, planners, policy analysts, etc.) and stakeholder groups (developers, business organizations, nongovernmental organizations, residents, etc.), thus requiring effective coordination among diverse actors and the use of nonspecialist language. In other words, defining the eco-city is not just a technical exercise but also an intrinsically political and public process involving normative priorities, value judgments, and complex governance decisions.

Variations on a Theme

Different definitional emphases are often reflected in the category terms used to describe the phenomenon. Hence, the term 'eco-city' is sometimes used interchangeably with sister terms, such as 'biosphere eco-city,' 'climate-neutral city,' 'eco-garden city,' 'low-carbon city,' 'resilient city,' 'slim city,' 'smart city,'

'solar city,' and 'sustainable city' (Joss et al., 2013). While there is considerable overlap between these terms, it is nevertheless possible to discern underlying conceptual as well as practical differences: for example, 'low-carbon city' and 'solar city' initiatives are chiefly defined in terms of greenhouse gas (GHG) emission reduction strategies and related renewable energy generation, whereas 'smart city' initiatives emphasize the use of 'ubiquitous' technology – that is, the central use of information and communication technologies for digitally linking up and coordinating various urban systems in an overarching information system. Sometimes, the choice of terminology is because of national policy, as in the case of China, where the Ministry for Environmental Protection issued 'national standards for an eco-garden city' in 2004 in an attempt to regulate the fast-evolving eco-city development in the country (Joss and Molella, 2013). Here, as in other countries, including in particular Japan (Low, 2013), a normative connection with the 'garden city' a progenitor dating back to the early twentieth century, is deliberately created. UNESCO (2009), as an example of an international organization seeking to define the field, has opted for the term 'biosphere eco-city' as part of its *Man and the Biosphere* (MAB) program, originally launched in 1971; it emphasizes the relationship between urban centers and their surrounding regions (the hinterland) as a means of adequately capturing and addressing urban sustainability issues.

Global Developments

Eco-City Pioneers

The modern eco-city movement, and sustainable urbanism as a new genre in urban planning, can be traced back to the period of the late 1970s to early 1980s. In the United States, the Urban Ecology movement (active from 1975–85) was instrumental in generating ideas and debate leading to the publication of *Ecocity Berkeley: Building Cities for a Healthy Future* by Richard Register (1987). This seminal work built on publications during that period by academics and activists engaging with issues, such as compact cities (in response to the global oil and energy crisis of the 1970s), urban sprawl, mixed-use developments and integrated neighborhoods, urban agriculture ('the edible city'), and sustainable communities. It set out to redefine the city in ecological terms by, among others, relating it to its wider bioregion, integrating biodiversity in urban planning, addressing the problem of urban sprawl (and associated private transport), and arguing for more livable, pedestrian-friendly neighborhoods. In doing so, *Ecocity Berkeley* equally emphasized social and esthetic issues: it argued for the city as an 'instrument for human purpose' (Register, 1987: 13), creating fairness and opportunity for citizens, and for the importance of paying attention to esthetic considerations, including how urban spaces interrelate with the wider environment.

Elsewhere during this period, the prospect of creating an eco-city as "a human settlement of the future in which social and ecological processes are harmonized in the best possible way" was proposed theoretically by Yanitsky (1982: 470) in connection with the MAB program. (It was not until 2006, however, that UNESCO/MAB formally adopted its 'biosphere eco-city' as policy initiative.)

This early period was characterized by the emerging influence of environmental movements and policy and, relatedly, by increasing concerns about the declining environmental and socio-economic quality of cities. It was a pioneering phase, exploratory in nature and ideologically and politically often radical, bringing together activists, academics, and practitioners with a joint interest in developing alternative approaches to mainstream urban policy.

Practical initiatives were few and far between during this period: in the United States, apart from Register's eco-city project in Berkeley (California), Paolo Soleri's Arcosanti (Arizona) began life as an experimental town (with 50–150 inhabitants) seeking to combine architecture with ecology ('arcology'). In India, Auroville (1500 inhabitants) was promoted as a 'universal city in the making' based on ecologically and socially harmonious principles. At a significantly larger scale and increasingly recognized internationally, the cities of Curitiba, Brazil, and Freiburg, Germany, each began to develop strong urban sustainability legacies from the 1970s onward. In the case of Curitiba, this initially centered upon promoting integrated public transportation, against the background of rapid population growth and related urban sprawl. The city subsequently extended its efforts to, among other things, waste recycling, including a 'food-for recyclables' exchange system, and green urban spaces, as a means of enhancing the environmental, social, and cultural aspects of urban living. For its part, Freiburg gained a reputation in Germany, and more widely in Europe, as an 'ecological capital' following the adoption of a sustainable city master plan in 1986, its pioneering work on creating sustainable neighborhoods, and, in subsequent years, the adoption of a 'climate protection protocol' to cut GHG emissions. In both cases, political leadership proved pivotal in introducing and implementing urban sustainability measures as key elements of the cities' political and policy agenda: in Curitiba, the original driving force was Jamie Lerner, the city's mayor; in Freiburg, the impetus came from the city council.

Policy Mainstreaming

Sustainable urbanism and the sustainable city were given a significant boost during the 1990s in connection with the rise of SD to key national and international policy. Following the United Nations Environment Summit in 1992 in Rio de Janeiro, cities and communities around the world began to implement sustainability programs as part of the Local Agenda 21 initiative launched at this first 'Earth Summit.' Governments joined forces by launching national programs. In Sweden, for example, all local authorities and municipalities were required to implement Agenda 21 action plans, while in Germany, the small town of Schwabach was selected by the federal government for a pilot study to design a model for urban sustainability development to be emulated elsewhere in Germany. This second phase in global development can be seen as 'regulatory' in a dual sense: the eco-city became increasingly standardized through the concept and discourse of sustainability, while at the same time eco-city pilot initiatives

played a standardizing role in the formation of SD policy and related sustainable urbanism practice (Joss, 2011a: 268–270). Internationally, 1990 saw the launch of the International Council for Local Environmental Initiatives (ICLEI) as a global association dedicated to promoting SD and related Local Agenda 21 initiatives. Renamed ICLEI Local Governments for Sustainability in 2003, its membership includes over 1000 cities and towns worldwide.

Internationalization

The most significant growth in eco-city initiatives globally occurred in the most recent, third period, from the early to mid-2000s onward. According to a global census of eco-city initiatives, by the end of the decade, there were well over 170 eco-city initiatives (Joss et al., 2011). As **Figure 1** shows, this follows a steep increase in the launch of eco-city initiatives worldwide since the early 2000s, particularly since the mid-2000s. Characteristic of this third phase, then, was the concurrent globalization (in terms of geographical spread and international profile) and mainstreaming (in terms of policy uptake and practical implementation) of the eco-city phenomenon.

Growth was particularly pronounced in Asia and Australasia, where there have been several significant national programs since the early 2000s (for an overview, see special issue on pan-Asian eco-cities, *Journal of Urban Technology*, vol. 20, no. 1). For example, in 2001, the Indian government (Ministry of Environment and Forestry) launched a first eco-city program aimed at retrofitting six pilgrim cities (Kottayam, Puri, Thanjavur, Tirupati, Ujjain, and Vrindavan). This was followed in 2010 by the Ministry of Commerce and Industry's plans for four new-build eco-cities (Changodar, Dahej, Manesar Bawal, and Shendra) along the 'Delhi–Mumbai industrial corridor,' and in 2011 by the launch of a 'solar cities initiative' – including pilot projects in 36 cities – by the Ministry of New and Renewable Energy. In Japan, in 2009, the government selected 13 cities (Chiyoda, Iida, Kitakyushu, Kyoto, Minamata, Miyakojima, Obihiro, Sakai, Shimokawa, Toyama, Toyota, Yokohama, and Yusuhara) for eco-city development, following its earlier pilot program that saw Kawasaki named the first eco-town in 1997.

Arguably, developments have been most prolific in China, spearheaded at the national level by the Ministry of Environmental Protection (MEP), the Ministry of Housing and Urban–Rural Development (MOHURD), and the National Development and Reform Commission (NDRC), and enthusiastically pursued by provincial governments, municipal authorities, and private developers (Joss and Molella, 2013; see also **Box 1**). According to the World Bank (2009: ii), by 2009 an estimated 100 eco-city initiatives were underway in the country. A 2011 survey identified 259 cities above the prefecture level that had declared the intention of becoming an 'eco-city' or 'low-carbon city' (Chinese Society for Urban Studies, 2011: 27). Elsewhere across Asia and Australasia, notable eco-city programs were underway in South Korea (in Gwanggyo Ekonhill, Incheon Eco-city, Songdo International Business District, Magok/Seoul, and Sejong) and Australia (Sydney and Melbourne).

In Europe, the pioneers of sustainable urban development – including Almere (the Netherlands), Eko-Viikki (Helsinki, Finland), Freiburg (Germany), Linz (Austria), and Växjö (Sweden) – have been joined, since the early 2000s, by a larger group of existing cities and new developments where eco-city initiatives are taking place. These include Brøset (Trondheim/Norway), Clonburris (Dublin), Ecociudad Logroño (Montecorvo, Spain), Hamburg-Harburg (Germany), Hammerby Sjöstad (Stockholm, Sweden), and PlanIT Valley (Porto, Portugal). France is notable through its ÉcoCité initiative launched in 2008 that saw the selection of 19 cities across the country. The UK government launched an 'eco-town' initiative for England in 2009, with initial plans for 12 sites eventually reduced to four (North-West Bicester, Rackheath, St Austell, and Whitehill-Bordon). Meanwhile, several individual UK cities such as Aberdeen, Coventry, Manchester, and St Davids embarked on eco-city initiatives of various kinds.

In North America, the cities of Portland, Vancouver, and Toronto launched, respectively, the 'EcoDistricts' (2009), 'Eco-Density' (2008) and 'Greenest City' (2010), and 'Living Green

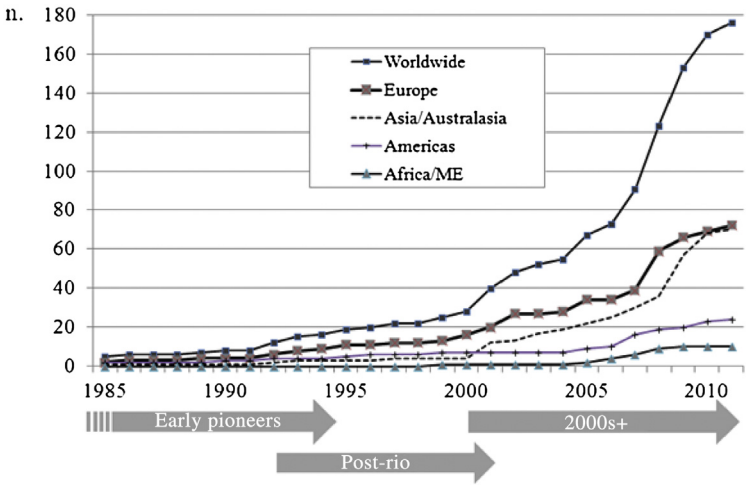

Figure 1 Number of eco-city initiatives launched by each date – globally and by regions. Source: Joss, S., Tomozeiu, D., Cowley, R., 2011. Eco-cities: A Global Survey 2011 (Eco-city Profiles), University of Westminster, London. http://www.westminster.ac.uk/ecocities (accessed November 2012).

Box 1 Tianjin eco-city (China) – sustainable urbanism through urban expansion

A flagship initiative launched in 2008 with support from the Chinese government and developed through bilateral cooperation with Singapore, Tianjin Eco-City is a new, large-scale urban development near the historic city of Tianjin in northeast China. Covering an area of approximately 30 km^2 nonarable land, the city will be home to 350 000 residents upon completion, which is expected in 2020. The city layout follows an integrated mixed-land use pattern: 400 × 400 m 'eco-cells' form the core design unit; every four to five eco-cells combine to form differently landscaped 'eco-neighborhoods' of approximately 20 000 residents; and 'eco-valley' corridors link the eco-neighborhoods and serve as main public open spaces stretching across the city. Ninety percent of transportation is to be by public transport (light-rail, bus, and trams), cycling, and walking. At least 15% of energy consumed should be from renewable sources, including geothermal and solar energy. Waste management will focus on reducing and recycling waste, with energy from waste incineration supplying the district energy system. At least half of the water consumed is to come from desalination and water-recycling processes. The existing wetlands around the city will be protected to enhance biodiversity. Twenty percent of housing is to be 'affordable', enabled through public housing subsidies.

Official website: http://www.tianjinecocity.gov.sg/

Toronto' (2010) initiatives, thus building on their earlier international reputation as innovators in urban sustainability (see Box 2 for more on Portland's EcoDistricts). Elsewhere, a dozen or so new infill and retrofit eco-city initiatives were launched, including 'Eco-city Alexandria' (Alexandria, Virginia, USA), the 'Seattle 2030 District' (Seattle, Washington, USA), and 'Dockside Green' (Victoria, British Columbia, Canada). Across Latin America, among the significant new eco-districts launched in the mid-2000s are 'Panama Pacifico,' a new district of Panama City (Panama) that styles itself as 'Latin America's first green city,' and 'Pedra Branca,' a new compact district of Palhoça (Brazil). In Africa and the Middle East, there have been fewer initiatives to date, although the unprecedented rate of urbanization and associated environmental pressures across much of Africa suggest a possible surge of eco-city developments in this global region in years to come. Among the initiatives launched in the 2000s, Masdar (UAE) has generated by far the most media interest globally. Other notable initiatives include 'Hacienda Ecocities' in Mombasa (Kenya); 'Eko Atlantic City' in Lagos (Nigeria); 'Ivory Park Eco-City' in Johannesburg, 'Menlyn Maine' in Pretoria, and 'Zonk'izizwe Town Center' in Johannesburg (all South Africa); and 'Sseesamirembe' in Kampala (Uganda).

Ubiquitous Eco-city

Apart from these numerous national developments, several international initiatives were launched in the 2000s. Their function variably entails providing support (conceptual, practical, and financial) for local initiatives, encouraging international knowledge and policy transfer, and acting as certification and endorsement schemes. Among these initiatives, the World Bank's Eco^2 Cities ('ecological cities as economic cities') initiative has, since its launch in 2009, been active mainly across Asia, including China, Indonesia, the Philippines, and Vietnam. The Organization for Economic Cooperation and Development (OECD)'s Green Cities program has engaged in comparative analysis and conceptual development relating to metropolitan regions in several of its member states, including France, Japan, Korea, Sweden, and the United States. Notable international nongovernmental initiatives include the Climate Positive Development Program ('Climate +'), which was launched in 2009 by the Clinton Climate Initiative with the participation of 17 cities and which in 2011 merged with the C40 Large Cities Climate Leadership Group ('C40'), thus bringing together over 70 cities in a global network; One Planet Communities, an initiative coordinated by BioRegional, a UK-based social enterprise, which includes five projects in Portugal, the United Kingdom, and the United States in its endorsement scheme and is involved in a further 11 projects around the world; and the International Ecocity Framework and Standards, a methodological framework and certification tool launched in 2010 by US-based Ecocity Builders.

Another feature of the internationalization of eco-city developments is the increasing prominence of global architecture, engineering, planning, and technology consultancies. European-based firms, such as Arup (UK), MVRDV (the Netherlands), and SBA Design (Germany), have become involved in master-planning, design, and development capacity in numerous initiatives, mainly in Asia. Similarly, technology firms have begun to take a close interest in the development of various 'smart technology' solutions in support of sustainable urbanism. For example, Hitachi of Japan is involved in the development of the four 'eco-friendly' cities

Box 2 Eco-districts, Portland (USA) – eco-innovation at the neighborhood level

Ranked regularly as one of the 'greenest' US cities, Portland, Oregon, has a legacy of innovation in urban sustainability dating back to the 1980s. More recently, in 2009, the city established the Portland Sustainability Institute, as part of which it launched the EcoDistricts initiative, on the basis that the neighborhood level is particularly effective for implementing sustainable urban development. In the pilot phase, the initiative is developed across five different districts in partnership with local organizations and community groups. Each EcoDistricts initiative is characterized by place-specific aspects. For example, the South-of-Market (SoMa) Eco-District involves Portland State University together with over a dozen other major property owners in the area; the aim is to improve the district through green streets and public transit and, thus, become an attractive place for business to locate. The university contributes to this through its substantial sustainability research base. Another example is the Foster Green Eco-District encompassing a cluster of suburban neighborhoods in the southeastern part of the city. Here, the focus is on enhancing the quality of life and addressing community needs and priorities, by improving economic vitality, public safety, public transportation, as well as access to open space. Beyond Portland itself, the EcoDistricts initiative has been developed into a network of US cities, including Boston, San Francisco, Seattle, and Washington, DC.

Official website: http://ecodistricts.org/

planned for the 'Delhi–Mumbai corridor,' with focus on integrated public transport systems, recycling technology, and 'smart grid' digital technology. In the United Kingdom, technology firms IBM and Toshiba have joined a consortium of private and public organizations to deliver similar smart grid solutions (integrating wind, tidal, geothermal, and solar energy generation) in connection with the 'Eco Island' initiative launched by the Isle of Wight (UK) in 2011.

The developments across these three periods suggest a gradual shift toward what may be called the 'ubiquitous eco-city' (Joss et al., 2013). This shift is characterized by accelerated practical innovation and implementation spread across various global regions, and the involvement of new actors through various forms of public–private partnerships. It occurred as a result of broad policy developments: initially against the backdrop of the growing environmental movement, the eco-city was conceptualized 'bottom-up' by activists, practitioners, and academics in response to local urban problems – such as urban sprawl, inner-city deprivation, and environmental pollution – and as an alternative to mainstream urban planning. With the adoption of sustainability as governmental policy, eco-city concepts became increasingly defined as part of policy and regulatory discourses, mainly at local and national levels. In the post-Kyoto era, the twin factors of global climate change and rapid urbanization prompted a dominant, international policy response in the form of the 'environmental modernization' agenda. (The UN Framework Convention on Climate Change – the 'Kyoto Protocol' – came into force in 2005; and in 2008, for the first time in human history, the majority of people now live in urban areas; UN Habitat, 2008.)

This, in turn, has resulted in an increasingly internationalist interest in practical eco-city innovation, at the heart of which is seen the opportunity to marry economic development with environmental sustainability. The World Bank's *Eco^2 Cities* program and the OECD's Green Cities program are illustrative of this more recent global conceptual direction. This is also mirrored at the national level through the recent launch of governmental programs – such as those by China, France, India, Japan, and South Korea – in support of green growth and urban sustainability policies in general, and various eco-city initiatives in particular, that together significantly account for the surge in eco-city projects. In comparison with earlier periods, the contemporary 'ubiquitous' eco-city is much more likely to be developed with a strong focus on promoting 'green growth' through international knowledge transfer activities – and particularly involving international private firms; it is more explicitly defined by a central 'carbon' agenda and more overtly driven by related 'green-smart' technological innovation. Finally, the shift toward the 'ubiquitous eco-city' has been characterized by growing efforts to define sustainable urbanism through international frameworks and standard schemes.

Challenges

The development of eco-cities – as both a concept and practice of sustainable urbanism – has been shown to be complex; as a consequence, it faces several interrelated challenges. Among these, four stand out in particular: first, some critics argue that, by their very nature, cities are inherently unsustainable judging by their ecological footprints (see, e.g., Rees and Wackernagel, 1996; Rees, 2010). Even cities considered to be at the forefront of sustainability efforts have been found to have ecological footprints several dozen times larger than their actual area. This applies to existing cities as much as to new urban developments. In the case of the latter, problematically the (often significant) construction-related environmental impacts, such as GHG emissions, are sometimes excluded from the eco-city profiles presented. In response, proponents would argue that eco-city initiatives, as part of an ongoing transition process, at least move toward more sustainable urban development in comparison to the status quo, with reference to concrete examples of improvements in relation to measures such as carbon emissions, water pollution, public transport, and 'livability'. This debate points to the need to scrutinize critically various 'eco-city' claims made and to develop more stringent indicators and standards to define and assess urban sustainability. As such, it lays bare the fact that in many cases, beyond statements of intent and future projections, systematic, independently verified evidence of actual eco-city achievements is currently hard to come by. This debate also points to the need to consider the city in relation to its wider hinterland, as arguably urban sustainability can be meaningfully discussed only from a broader bioregional perspective.

The second challenge, from an urban policy and planning perspective, is the charge that many eco-city initiatives represent premium 'ecological enclaves' (e.g., Hodson and Marvin, 2010). According to this perspective, the eco-city creates and reinforces premium developments aimed at providing more secure and higher standard urban living environments for a more affluent clientele. As such, it can be seen as part of a wider neoliberal trend toward achieving urban-ecological security. 'Hacienda Ecocities' in Kenya is an example of an initiative promoted as offering price stability, service reliability, and security for its mainly middle-class residents (Joss et al., 2011: 54–55). In a similar vein, some of the plans and designs for new eco-cities in China demonstrate an aspiration for Western-style, high-end urban developments that stand in contrast to existing mainstream urban living there (Joss and Molella, 2013). The increasing involvement in recent years of international consultancy firms in the financing, master planning, and delivery of eco-city initiatives further underlines the trend toward sustainable urbanism as an investment and development opportunity. The challenge here is that the eco-city may fail to provide a solution to urban sustainability that is accessible to wider communities and addresses environmental, economic, and social sustainability issues beyond the 'premium enclave' itself. However, from an innovation perspective, it can be argued that the initial 'niche' development over time can be expected to become mainstream, thus benefitting cities more widely. Furthermore, the 'ecological enclave' discourse is only one among several (contrasting) discourses shaping contemporary eco-city initiatives. 'Ivory Park EcoCity' (Johannesburg, South Africa) is an example of an initiative emphasizing community-based sustainability innovations supporting indigenous practices

and local enterprise (Joss et al., 2011: 63). Portland's EcoDistricts initiative promotes sustainable neighborhood development through the active involvement of five communities across the city (Joss et al., 2011: 82).

A third challenge arises from the increasingly central role placed on technology to provide solutions for eco-city development. On the surface, this is about the application of an array of green technologies. However, emerging cities, such as 'Tangshan Caofeidian International Eco-City' – also referred to as China's first 'information eco-city' – appear to conceive of the eco-city more generally as high-tech cities (Joss and Molella, 2013). This manifests itself in the integral use of 'smart' technology for digitally linking up and coordinating various urban systems in an overarching information system. The Japanese technology firm Panasonic is involved in developing 'Fujisawa Sustainable Smart Town,' designed as a high-tech city with advanced solar power generation, heat pump, and hot-water systems connected by a 'smart grid' that allows for city-wide, real-time systems management. In Portugal, 'PlanIT Valley' is designed as a high-tech district of Paredes, based on the concept of a replicable 'urban operating system' (Joss et al., 2013). Such heavy emphasis on technological innovation and solutions runs the risk of conceptualizing the eco-city narrowly as physical infrastructure and an urban technological system. Instead, insights from science and technology studies (STS) point to the need to consider the function of technology more broadly in terms of socio-technical relationships (human–technology interaction). Furthermore, from an urban politics perspective, the application of technology should be considered in terms of its place-making quality – that is, how it can enhance community and citizen engagement, support political discourse, and create livable urban space.

This points to the fourth challenge, relating to governance (Joss, 2011b; Rydin, 2010). Given the complexity of sustainable urbanism, which typically cuts across organizational, policy sectoral, geographical, and jurisdictional boundaries and involves a multitude of actors (local, national, international, public, and private) in various configurations, the issue of governance comes to the fore. Steering, negotiation, coordination, and integration are all essential parts of the planning, development, and implementation processes of eco-city initiatives. The nature and mode of governance vary depending on, among other factors, the type and scale of initiative (retrofit, infill, and new) and the wider policy and political contexts. The challenge consists of putting in place governance arrangements that are sufficiently effective, efficient, and robust to allow the implementation of eco-city development plans across often lengthy time periods (two or more decades are not untypical) and involving considerable financial and organizational resources and risks. It further consists of adopting governance modes that enable urban sustainability innovation processes, through which new ideas, designs, and practices can emerge. Finally, it consists of enabling political and public accountability, by rendering decision-making structures and processes (such as public–private partnerships) transparent and open to scrutiny, and by promoting meaningful stakeholder and community engagement.

Conclusion

Sustainable urbanism and the eco-city owe much of their current popularity to the forceful combination of global climate change concerns and a rapidly urbanizing world population. The translation of concepts into practice through various types of eco-city initiatives has been enabled by a broad coalition of involved actors, with municipal authorities, national governments, private firms, and international organizations all instrumental in driving innovation. This mainstreaming has resulted in several ubiquitous tendencies, characterized by underlying 'green growth' strategies, a dominant 'carbon discourse,' the preponderance of 'smart' technological solutions, and increased policy and knowledge transfer activity spearheaded by internationally active consultancies and governmental and nongovernmental organizations dedicated to promoting urban sustainability. Efforts to arrive at national and international frameworks and standards for the sustainable city have added to this ongoing consolidation process. Counteracting these ubiquitous tendencies are local settings and contexts through which sustainable urbanism is implemented. As a result, certain tensions are built into eco-city development, requiring the reconciliation of international concerns (e.g., the reduction of GHG emissions) with local priorities (e.g., promoting economic development and addressing place-specific environmental problems and social needs). Governance is an essential requirement of sustainable urbanism, to enable the effective coordination and integration of the intrinsically complex, multifaceted structures and processes involved. Faced with the rapid growth of practical initiatives on the ground as part of an ongoing period of innovation, there is a present need for critical, multidisciplinary analysis of proposed concepts and projects, as well as the substantive evaluation of achievements to date. In turn, this will help bring about the transition to the next generation of eco-cities.

See also: Environmental Planning; Smart Cities; Urban Sprawl.

Bibliography

Chinese Society for Urban Studies, 2011. Development Report on China's Low-Carbon Eco-cities. Architecture and Building Press, Beijing.

City of Vancouver, 2010. Greenest City 2020. Action Plan accessed November 2012 http://vancouver.ca/files/cov/Greenest-city-action-plan.pdf.

Girardet, H., 1999. Creating Sustainable Cities. Schumacher Briefings 2, third ed. Greenbooks, Devon.

Hodson, M., Marvin, S., 2010. Urbanism in the anthropocene. Ecological urbanism or premium ecological enclaves? City 14 (3), 298–313.

Joss, S., 2011a. Eco-cities: the mainstreaming of urban sustainability; key characteristics and driving factors. International Journal of Sustainable Development and Planning 6 (3), 268–285.

Joss, S., 2011b. Eco-city governance: a case study of Treasure Island and Sonoma Mountain Village. Journal of Environmental Policy and Planning 13 (4), 331–348.

Joss, S. (Ed.), 2012. Tomorrow's City Today. Eco-city Indicators, Standards & Frameworks. Bellagio Conference Report. University of Westminster, London. http://www.westminster.ac.uk/ecocities.

Joss, S., Cowley, R., Tomozeiu, D., 2013. Towards the 'ubiquitous eco-city': an analysis of the internationalisation of eco-city policy and practice. Journal of Urban Research & Practice 6 (1), 54–74.

Joss, S., Molella, A., 2013. The eco-city as urban technology: perspectives on Tangshan Caofeidian International Eco-city (China). Journal of Urban Technology 20 (1), 115–137.

Joss, S., Tomozeiu, D., Cowley, R., 2011. Eco-cities: A Global Survey 2011 (Eco-city Profiles). University of Westminster, London accessed November 2012 http://www.westminster.ac.uk/ecocities.

Joss, S., Tomozeiu, D., Cowley, R., 2012. Eco-city indicators: governance challenges. WIT Transactions on Ecology and the Environment 155, 109–120.

Journal of Urban Technology, 2013. Eco-cities in Pan-Asia: international discourses, local practices. Journal of Urban Technology 20 (1) special issue.

Kenworthy, J.R., 2006. The eco-city: ten key transport and planning dimensions for sustainable city development. Environment and Urbanization 18 (1), 67–85.

Lehmann, S., 2010. The Principles of Green Urbanism. Transforming the City for Sustainability. Earthscan, London.

Low, M., 2013. Eco-cities in Japan: past and future. Journal of Urban Technology 20 (1).

Rees, W.E., 2010. Getting serious about urban sustainability: eco-footprints and the vulnerability of twenty-first-century cities. In: Bunting, T., Filion, P., Walker, R. (Eds.), Canadian Cities in Transition: New Directions in the Twenty-first Century. Oxford University Press, Toronto, pp. 70–86.

Rees, W., Wackernagel, M., 1996. Urban ecological footprints: why cities cannot be sustainable. Environmental Impact Assessment Review 16, 223–248.

Register, R., 1987. Ecocity Berkeley. Building Cities for a Healthy Future. North Atlantic Books, Berkeley, CA.

Register, R., 2006. Ecocities. Rebuilding Cities in Balance with Nature, revised ed. New Society Publishers, Gabriola Island, Canada.

Roseland, M., 1997. Eco-city Dimensions: Healthy Communities, Healthy Planet. New Society Publishers, Gabriola Island, Canada.

Roseland, M., 2012. Toward Sustainable Communities. Solutions for Citizens and Their Governments, fourth ed. New Society Publishers, Gabriola Island, Canada.

Rydin, Y., 2010. Governing for Sustainable Urban Development. Earthscan, London.

UNESCO, 2009. A Citizen's Guide to Biosphere Eco-cities accessed November 2012 http://www.unesco.org/mab/doc/icc/2009/e_ecoCities.pdf.

Wheeler, S.M., Beatley, T., 2004. The Sustainable Urban Development Reader, second ed. Routledge, London.

World Bank, 2009. Sino-Singapore Tianjin Eco-city (SSTEC): A Case Study of an Emerging Eco-city in China. Technical Assistance (TA) Report no 59012. World Bank, Washington, DC.

World Bank, 2010. Eco^2 Cities: Ecological Cities as Economic Cities. World Bank, Washington, DC.

Yanitsky, O., 1982. Towards an eco-city: problems of integrating knowledge with practice. International Social Sciences Journal 3, 469–480.

Ecocultural Theories of Development

A Bame Nsamenang, University of Bamenda, Bamenda, Cameroon

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article conceptualizes human-lived contexts and their cultures not as variables but as developmental provocateurs. This view is proposed because environments are inhabited spaces, that serve as settings that shape human development. Although culture is both inside and outside the individual, the 'culture' that channels developmental outcomes is inside the individuating self as a core facet of the ecocultural paradigm. Analysis of variants of ecocultural theories reveals diversity in mindsets regarding the role of physical and cultural contexts in development. The developmental impact of these roles has been mainly unexplored. A critique of efforts to evolve a unified developmental science shows that existing approaches are mostly skewed toward Euro-Western views and reveals a neglect of varied social constructions of childhood and conditions of children from across the globe.

Introduction

This article proposes a novel approach to conceptualizing the role of lived ecological contexts and their cultures in human development research as developmental provocateurs, rather than simply considering them as variables. This approach recognizes the primacy of the context in general and the lived or developmental context in particular for every human being, thereby departing from the earlier conceptualization and theorization of human development as if it were a decontextualized, culture-free phenomenon. A physical place and its ecological, social realities and cultural contents should be at the forefront of developmental research thinking. In short, the ecocultural approach considers human development to take place in a context that includes both environmental and cultural influences (which serve as both constraints and facilitators).

Although culture is both inside and outside the individual, the frame of reference from a developmental position is that "culture [is] inside the person" (Poortinga, 2011). The article adopts and innovates from the heuristic seminal ecocultural paradigm of John Berry (1966, 1971, 2011), incorporating subsequent versions of ecological thinking, particularly the integrative approach of Pierre Dasen (2003), in efforts to point to the global diversity of developmental contexts in their physical and cultural dimensions. This article examines variants of ecocultural theories of human development, and the effort to evolve a unified developmental science that reflects global human diversity.

Cultural Ecology

The concept of cultural ecology refers to the changing relationships between culture and the environment (Bada, 1995; Everette, 1996) or the total web of life (Steward, 1955), through developmental adaptations (Berry, 1976; Cohen, 2001). How cultures interact with the environment is related to how societies confront and solve the common problems of human existence (Thomas, 2008). Culture, thus, gives meanings to one's way of life and provides the possibility of control over the environment. Cultural constructions, such as norms, values, and myths, in turn, allow people to make sense of the environment (McCarthy, 1998). The undeniable interdependency between ecology and culture is crucial for the understanding of human development in general and of human relations in particular. Thus, cultural ecology can be said to be one of the key ingredients of future human development studies.

Generally, cultures in different parts of the world are often linked to the given availability of natural resources, climate, terrain, and geographic latitude (Hofstede, 2001; Triandis and Suh, 2002), although structural barriers could reduce or eliminate access to them. Each culture has developed its own unique structure for dealing effectively with its environment, given its available resources (Trompenaars and Hampden-Turner, 1977).

Berry (1966) noted that culture contact induces culture change through the process of acculturation. Nsamenang (2011) described cultural and psychological change resulting from the intermingling of cultural traditions through increasing waves of immigration and settlement in 'other' people's cultural areas. Fang (2009) noted that these external changes often result from cultural clashes and collisions, with cultural mixing processes usually being demanding and exciting (Remer, 2005). Cohen (2001) has described how cultural value adaptation of a people was influenced by their correct or mistaken given beliefs about what others will do, which eventually leads to the value equilibrium both on domestic and world levels. The distinction between hunter-gatherer and agricultural societies, for example, is a clear indication of the cultural ecology causality that developmental research ought to pursue.

Developmental science research should be adopting a dynamism-oriented ecocultural model because developmental processes and cultural conditions of human development and livelihoods are dynamic, changing, and sometimes overtly illogical and incoherent. Cultural ecology can elucidate the dynamic, changing nature of culture (Bird and Fang, 2009; Chou, 2009).

Ecocultural Theorization: Conceptual and Theoretic Thrust

The term *ecocultural* is a hybrid, coined from two concepts, ecology (referring to a physical place and its ecosystem that includes networks of people and children) and culture. The concept is rooted in the view that human development is influenced by both the ecological context, and the cultural context, whose features are often adapted to the ecological setting (Berry, 1966, 1971, 1976). The ecocultural approach has a long history (Jahoda, 1995) and many variants. In this approach, culture consists of concrete aspects (such as tools) as well more subjective features (such as belief systems, oral traditions, literacy, and language) and the different ways of thinking and acting about various aspects of life. The ecocultural framework situates the developing human being not in a universal culture but within a culture that is context bounded by a geographical or physical environment and characterized by a social system that includes the cultural imperatives of the people who inhabit or are active in that ecosystem.

The ecocultural model of human development is a Eurocentric enterprise (Nsamenang, 2002) in that theoretical and empirical efforts have focused overwhelmingly on the ecological and cultural realities of the Minority (Western) World to the relative neglect of those of the Majority World where the bulk of humanity, especially those of the world's child and youth cohorts, reside. In the Western worldview, maturity is portrayed in terms of increasing autonomy from parents and where identity development is described as a process of exploration among relatively unrestricted educational and career opportunities (Adams et al., 1992). Such accounts omit the dreadful experiences of ethnic minority youth, as well as those of Majority World children and youth, who often have early and lifelong responsibilities to family members and who face structural barriers to opportunities in school and work (Grotevant and Cooper, 1998). Furthermore, McLloyd (1991) has shown how race-comparative designs create norms based on European American, middle-class experiences. These norms lead readers to interpret differences between ethnic groups and Majority World children and youth as deficits from the mainstream and to view non-Western children, youth, and their families in terms of negative stereotypes.

Laypersons as well as research scholars often think of or handle ethnicity, culture, and family as broad categorical qualities that are relatively uniform in each group and static or stable across time and context. For example, European Americans are often labeled individualistic, whereas African-Americans, Asian-Americans, Latin-Americans, and Native Americans, and Majority World peoples in general, tend to be portrayed as collectivistic or interdependent in outlook and lifestyle. Oversimplified categories such as ethnicity, culture, and family, were questioned by anthropologist Beatrice Whiting (1976) who challenged scholars to begin 'unpackaging' these categories to understand their multiple dimensions and components. In the same vein, Weisner et al. (1988) argued

> culture is not a nominal variable to be attached equally to every child, in the same way that age, height, or sex might be. Treating culture in this way assumes that all children in a cultural group have common natal experiences. In many cases, they do not. The assumption of homogeneity of experience of children within cultures, without empirical evidence, is unwarranted…. A similar error is to treat national or ethnic status as equivalent to a common cultural experience for individuals (p. 328).

Most extant human development theories tend to be presented as universally applicable; however they are culture-infused and only a few inform how context and culture shape biological development. So, researchers need to figure out how to move beyond the categorical treatment of ethnicity, culture, and family to their dynamic merging in individuals.

The Original Ecocultural Theory: Berry

John Berry's (1966) ecocultural approach represents an attempt to resolve the tension between culture as an entity in its own right and the playing out of adaptation in the individuals and collectivities that nurse on this maternal resource as they attempt to know, integrate, and perform in their cultural context and beyond it in the face of acculturative pressures. Berry (1966, 1971, 1976) was convinced that human beings develop in ways that permit their adaptation to and long-term survival in, particular, ecosystems.

Ferguson (1956) laid the foundation for the ecocultural approach by articulating what was later called (by Irvine and Berry, 1988: p. 121) the "Law of Cultural Differentiation" of human abilities: "Cultural factors prescribe what shall be learned and at what age; consequently different cultural environments lead to the development of different patterns of ability." These "different patterns of ability" were construed by Berry (1976) as "cognitive styles," particularly that of "field dependence and field independence." Variations in the development of these styles were then linked to the ecological and cultural contexts in which individuals develop. These contexts were understood within the anthropological school of 'ecological anthropology.'

Figure 1 presents the ecocultural framework. A main feature of this framework is to distinguish the population and the individual levels of analysis of developmental influences and outcomes. It views cultures as evolving adaptations to ecological and sociopolitical influences, and views individual psychological characteristics in a population as adaptive to their cultural contexts.

There are five main features of the ecocultural model: *ecological context; sociopolitical context; population adaptations* (both cultural and biological*); transmission variables* that serve as vehicles for population characteristics becoming part of the behavior of the individual; and *behavioral outcomes*.

In addition to populations adapting to their ecological and sociopolitical contexts (on the left of **Figure 1**), a key component of the ecocultural framework for understanding development is the set of transmission variables (in the middle) between these contexts and adaptations, and individual psychological development (on the right side). In essence, we need to examine the various ways in which cultural features of populations to varying degrees become incorporated into the repertoire of individuals. Four such variables are

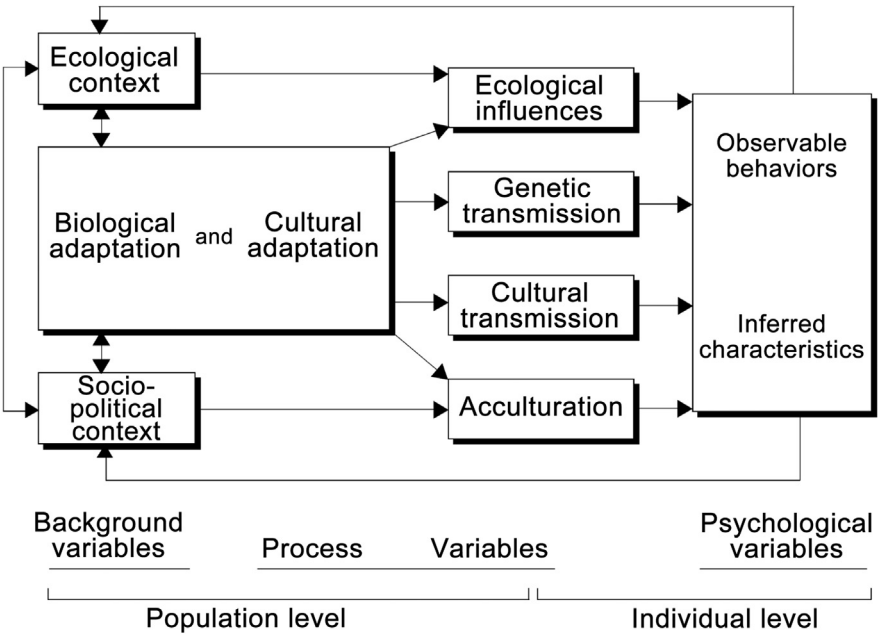

Figure 1 Ecocultural framework linking ecology, cultural adaptation, and individual behavior. Berry, J.W., 2003. Ecological perspective on human psychological development. In: Saraswathi, T.S. (Ed.), Cross-cultural Perspectives in Human Development: Theory, Research and Applications. Sage, New Delhi, India, p. 55.

identified: direct ecological influences (without any explicit social or cultural mediation), genetic transmission from parents to offspring, intergenerational transmission of cultural characteristics (through the processes of socialization and enculturation), and acculturation (through sociopolitical influences impacting from outside the culture).

Dasen (2003) perceives the ecocultural framework as very broad, and although cultural transmission figures as a central process, the framework is not specifically designed to deal with human development. The next section reviews the ecological systems theory (EST) of Bronfenbrenner (1979) that takes the developing person as a focal point.

The EST: Bronfenbrenner

The EST positions human development within systems of superimposed 'layers' of environmental settings, each impacting differently on human development. The initial and most utilized version is Bronfenbrenner's (1979) EST that identifies developmental influences at four levels: macro-, exo-, meso-, and microsystem, which correspond to intercultural, community, organizational, and interpersonal or individual levels. Prior to Bronfenbrenner's (1979) framework, many research theorists had traditionally considered only a dichotomy of perspectives, either micro (individual behavior) or macro (effect of cultural factors). The EST was focused on the person, the environment, and the continuous interaction of the two. This interaction constantly evolves and develops both components. However, it is not only the environment that directly affects the person, but there are layers in between, which all have resulting impacts on the next level. The individual, organization, community, and culture are nested into each other, like Russian dolls, with each level operating fully within the next larger sphere.

The EST theory holds that development reflects the influence of five ecological subsystems:

Microsystem

The setting in which the individual lives or undertakes routine activities. This includes the child's family, peers, school, and neighborhood. It is in the microsystem that the most direct interactions with social agents take place; with parents, peers, and teachers, for example. The child is not a passive recipient of experiences in these interactions, but someone who contributes to construct the nature of the settings. A child's own biology is considered part of her or his microsystem.

Mesosystem

Refers to relations between microsystems or connections with the other contexts, such as the relation of family experiences to school experiences, school experiences to church experiences, and family experiences to peer experiences.

Exosystem

Involves links between a social setting in which the individual does not have an active role and the individual's immediate context. For example, a husband's or child's experience at home may be influenced by a mother's experiences at work.

Macrosystem

Describes the culture in which individuals live. Elements of macrosystms include the cultural contexts of developing and industrialized countries as well as their socioeconomic status, poverty, and ethnicity, among many others.

Chronosystem

Refers to the structure of environmental events and transitions over the life course, as well as across social-historical circumstances. For example, divorce is one transition and family relocation or migration for whatever reason is another.

Each of the five systems is characterized by roles, norms (expected acceptable behavior), and relationships or social networks, including those of children. For example, children tend to act differently within their own family than within a classroom and toddlers can be observed to act differently in peer cultures than in the presence of adults.

The Developmental Niche Framework: Super and Harkness

The *developmental niche* (Super and Harkness, 1986) is a theoretical framework for understanding and analyzing how culture shapes child development. It is an ecologically grounded theory that, like Weisner's (2002) "ecocultural niche," Worthman's (1994) "developmental microniche," and Nsamenang's (1992) social ontogenesis, has the same origins as EST. Developmental niche theory is a variant of the ecocultural model that integrates two basic principles from cultural anthropology and developmental psychology. First, that a child's environment is organized as part of the cultural meaning system of a cultural community. Second, that a child's biological disposition, including attributes, temperament, skills, and potentials, affect her or his process of development. The two overarching principles of the developmental niche framework are: a child's environment is organized in a nonarbitrary manner as part of a cultural system; and the child's own disposition, including a particular constellation of attributes, temperament, skills, and potentials, affect the process of development. The developmental niche is seen as the composite of three interacting subsystems:

physical and social settings: who is there, what affordances are provided by the physical space and social others;
customs and practices of child rearing: what are the inherited and adapted ways of nurturing, entertaining, educating, and protecting the child; and
the psychology of the caregivers: particularly parental ethnotheories of child development and parenting, which play a directive role in actual childcare practices.

The three subsystems of the developmental niche (settings, customs, and caregiver psychology) share the common function of mediating the child's developmental experiences within the larger culture. The three components of the developmental niche operate together with powerful though incomplete coordination as a system. Thus for example,in a stable cultural environment, customs of care reflect parental ethnotheories about the child, and they are further supported by the physical and social settings of daily life. In circumstances of rapid social change, or immigration, there will be greater inconsistency among the subsystems, but one can usually discern homeostatic processes or acculturative forces at work in the context of changing beliefs and practices and their consequences on children.

Each of the three subsystems of the niche is functionally embedded in other aspects of the human ecology in specific and unique ways; in other words, the three subsystems act as the primary channels through which the niche, as an open system, is influenced by outside forces. Any one of the three components may be a primary route of influence. For example, economic or social change may lead to new settings for children. Moreover, each of the three subsystems of the niche is involved in a process of mutual adaptation with the individual child. Thus, the age, gender, temperament, energy level, and talents of the individual child influence parents and others in the niche and modulate cultural expectations and opportunities for the child at any given time. The developmental niche framework makes evident the kind of systematic regularity that culture provides – environmental organization that emphasizes repeatedly or with singular salience the culture's core 'messages.' This quality, called 'contemporary redundancy,' is important for the acquisition of skills and values as it offers multiple opportunities for learning the same thing, whether that 'thing' is reading, caring for others, the communication of emotions, or the value of 'independence.'

Social Ontogenesis: Nsamenang

Nsamenang (1992) developed a theoretical position termed *social ontogenesis*, based on an indigenous West African conception that is phrased within an ecocultural perspective. His ideas are inspired by the writings of various Africans trained in philosophy and the humanities (e.g., Mbiti, 1990; Moumouni, 1968) and grounded in a combination of systematic observational research and his own personal experience of the socialization practices of the rural versus urban Nso communities in Western Cameroon (Nsamenang and Lamb, 1993, 1995). His formulation of social ontogenesis is rooted in a widely shared West African worldview. It posits the growth of social selfhood through a series of phases, each characterized by a distinctive developmental task, defined within the framework of the culture's primarily socioaffective, developmental agenda. The theory of social ontogenesis (Nsamenang, 1992, 2004, 2012a) points out how, beginning early in life and through developmental stages, African children are active in the life of their families and societies as well as in self-care and self-learning. This theory presents human development as partly determined by the social ecology in which the development occurs and by how African children, especially in sibling and peer settings, learn from each other in peer cultures.

The seminal concept of this theory is sociogenesis, defined as individual development that is explained more in terms of socially observed markers and culturally perceived tasks but less on biological unfolding, although social ontogenetic thinking does not preclude nature; it assumes that biology underpins social ontogenesis (Nsamenang, 2006a). Ngaujah (2003) has interpreted Nsamenang's theoretical approach as revealing the social and affective nature of the environment on the child's cognitive and social learning.

The first phase of social ontogenesis is the *ceremony of naming*, which projects the kind of socialized being the neonate should become. The major developmental task of the second phase is success in *social priming*: babies are cuddled and teased to smile along with adults; parents and other caregivers offer infants food items and playthings, and lure them both verbally and through nonverbal communication to return the 'gifts.'

This is a preliminary step toward induction into the 'sharing and exchange norms' that bond siblings and the entire social system together. For example, among the Chewa and Tumbuka peoples in Zambia, Mtonga (2012) interprets such interactions of adults with toddlers, as aimed at cultivating generosity and preventing the development of greediness or selfishness.

The next phase, roughly corresponding with childhood, is termed the period of *social apprenticing* in which the principal developmental task is to recognize, cognize, and rehearse social roles that pertain to three hierarchical spheres of life: household, network, and public. Much of the responsibility for stimulation and guidance in this phase of early childhood development is assigned to preadolescent and adolescent children in the family and the neighborhood. The delegation of responsibility for care and socialization of younger children from adults to preadolescents and adolescents serves the function of *priming* the emergence of social responsibility. These priming strategies embedded in traditional African child-rearing practices have important implications for the design of culturally appropriate forms of intervention to optimize developmental opportunities for children in contemporary Africa (Nsamenang, 2009). Indeed, in many African communities with a subsistence economy, far from constituting a form of exploitation or abuse, caregiving responsibilities assigned to preadolescents and adolescents are better understood as part of "an indigenous educational strategy that keeps children in contact with existential realities and the activities of daily life [that] represents the participatory component of social integration" (Nsamenang, 1992: p. 157).

By positioning children as emerging into levels of selfhood, implying the unfolding of biological potentialities and social competencies, Africans tacitly acknowledge that self-concept and agency evolve with a maturing self-consciousness that accords a sense of self-direction and agentive search for or choice of the resources and exposures that increasingly differentiate and polish self-identity and goal-directed behavior toward desired or imagined personal status, either of sovereign individuality or relational individuality (Kagitcibasi, 2007; Nsamenang, 2004).

African parents sensitize children from an early age to seek out others from whom to extract local knowledge and situated intelligences (Ogbu, 1992) and in so doing clarify who they are, more so within sibling and peer spaces than with adults. Children extort the social, emotional, practical, cognitive, relational values and other norms ingrained in the activity settings of the home, society, and peer cultures more through their contextual embedment and active participation and less through explicit adult instruction or prodding. In so doing, they 'graduate' from one activity setting and participative sector of the peer culture to another, steadily maturing toward adult identity and roles (Nsamenang, 2012a). Zimba (2002: p. 94) described one instance of self-definition with the Zulu community of South Africa, as nurturing *umuntu umuntu ngabantu*, which literally translates into "a person is only a person with other people." This relational view of identity development downplays lonesome individuation, implying that a sense of self cannot be attained or adequately understood without reference to the 'community' of others in which it is embedded. Concepts of place identity and place attachment accentuate the need to incorporate context into theory building and research agendas.

Applications of the Ecocultural Approach

Applications of these ideas are helpful when creating settings for development. We may begin by recognizing the universal needs of children. Children's essential needs include warm, caring, and responsive adults; a sense of importance and significance; a way to connect, belong, and relate to the world around them; opportunities to explore and play; and people to help structure, guide, and support their learning and opportunities for self-esteem and self-actualization. A basic human need is the need to belong. Children need to feel they belong, too. They need to be close to people they know; have familiar, safe, and comfortable play objects; and be in a setting that has a personal history for them. Young children need to feel important. Children need to feel that what they do is meaningful to someone other than themselves, particularly adults. In the distant past, these needs were met at home and in the community, but today, for the majority of children and young people worldwide, they are being met in centers, classrooms, or other group engagements, wherein new challenges, even abuses, are emerging in some of them. Children need to explore, experiment, and learn basic knowledge and skills through direct experience.

In spite of the enormous variation across cultures in the perception of childhood and children's needs and childcare arrangements to satisfy them, the Convention on the Rights of Children (UNICEF, 2003) recognizes the family as the most appropriate societal institution to best serve the interests and needs of children. But family types, their composition, and role vary by culture, social class, and ecocultural circumstances (Nsamenang, 2012b).

So, in early childhood care and education (ECCE) centers and schools, we can capitalize on the African family practice of children being accredited skilled partners in peers and siblings' zones of proximal development to redesign curricular content, learning spaces, and pedagogic approaches.

Childhood is a time when children should learn firsthand about the physical world: the feel of water, dust, or mud, the constant pull of gravity, the stench of rotten fruit, and the abrasive feel of sand or concrete on a bare knee. Play provides a way for children to integrate all their new experiences into their rapidly developing minds, bodies, emotions, and social skills. Brain research supports this idea, stressing that children learn best through an integrated approach combining physical, emotional, cognitive, and social growth. Education in African family traditions is part and parcel of cultural life. It is integrated into the daily routines of the family and community and does not compartmentalize knowledge into say, agriculture, economics, arts, sciences, etc. Instead, it integrates knowledge about all aspects of life into a single curriculum. The curriculum is arranged in sequence to fit into different milestones of development that the culture perceives or recognizes (Nsamenang, 2006b: p. 329–330). The role of the parent, mentor, including the peer-teacher, teacher is critical in a child's life. Children depend on knowledgeable others to be their confidant, colleague, model, instructor, and nurturer of learning experiences. Clearly children need lots of exposure

to other people in their early childhood years as well as to a wide variety of learning and play settings.

An early childhood environment is many things: it should be a safe place where children are protected from the elements and are easily supervised, and it is where the routine daily activities take place, such as playing, eating, sleeping, washing hands, going to the bathroom, etc. The ecocultural theory holds that a universal task for all families is to organize a sustainable daily routine (Weisner, 1984). The theory assumes that family adaptation involves balancing ecology (resources, constraints), culture (beliefs, values, and schemata), and the needs and abilities of family members in the organization of daily routines (Gallimore et al., 1989). A family's daily routine is dynamic and changes according to its needs. How it addresses these needs is filtered by perceptions and concrete use of resources and constraints, with the actual resources and constraints of the environment. This often involves negotiating options and balancing competing demands. For example, household management involves attending to domestic chores like food preparing and domestic chores as well as supervising children's development and school work. A village mother may decide to forego a farming task in order to take an ill child to hospital. In this example, the family ecology characterized by limited time, no assistance from the husband, and sociocultural values such as a child's health and learning are linked to changes in the family's routine.

Each community that children belong to makes its own specific curriculum demands: the community of learners who will be able to respond to challenge and change; the community of children who have individual needs and rights; and the community of Africans, for example, who are gaining knowledge of their nation's languages and are developing skills in using cultural tools such as art, dance, mathematics, music, reading, science, technology, and writing. The learner and her or his immediate locale, for instance, different activity settings, are at the center and the first level for learning. Other levels have a powerful influence on the child's well-being and capacity to learn. The second level contains the major settings experienced directly or vicariously by the learner: the child's own home, the service or setting beyond their home through mass media and cyberspace, for example, and the relationships between these settings. A third level that influences the quality of children's experiences, encompasses the world of work, the neighborhood, the mass media, and cyberspace if the child is exposed to it, as well as informal social networks. It also includes the conditions that influence the well-being and support of the adults in the children's lives: the demands, the stresses, and the opportunities for development experienced by significant adults in each child's life. The national level is the most encompassing, including the nation's philosophy about the value of early childhood care and education and about the rights and responsibilities of children. Although ECCE is mainly concerned with the lower levels, these highest levels are important influences on the quality of the curriculum.

Conclusion

The ecocultural perspective informs us that the young of every species have basic needs that must be met for them to survive, thrive, mature, and develop, expectedly into adult members of the society. It also informs us that across the globe, people of every culture and ecology fulfill these tasks in divergent ways. All cultures worldwide are constantly engaged in a constant search for ways of understanding and building on the developmental contexts to maximize human potential. In fact, human biology is wired to acquire, create, and transmit culture. As biology interfaces with culture, it is not just a preface to or foundation for human development but a consistent theme throughout its process (Keller, 2007). The human brain, for instance, is built to be modified by the environment and culture in which it develops. Thus the nervous system is culturally interactive and a bearer of cultural difference, or, put more comprehensively, "Culture…is a determinative complement of genotype that shapes human psychosocial differentiation in the direction of a given people's cultural meaning systems" (Nsamenang, 2008b: p. 73). Given that cultures infuse their own agendas onto a natural, biological program of human ontogenesis (Nsamenang, 1992), the notion of adaptation to varied environments should be carefully considered and systematically worked into developmental research agendas.

Social ecology is the study of people in an environment and the influences they exert on their environment and on one another, including children on themselves and their interlocutors. The interaction between factors in children's maturing biology, their immediate family/community environment, and the societal landscape fuels and channels development (see Nsamenang and Akum, 2013). Thus, interactions with others and the environment are central to human developmental outcomes. As such, contextual interactionism constitutes a theoretical framework or mind-set for analyzing various types of context in which children live and how they are cared for and educated in their early years. This model allows for the integration of multiple levels of interaction with contexts and its people and opportunities or constraints to establish the big picture of forces that influence child development.

See also: Cross-Cultural Psychology; Cultural Psychology; Indigenous Psychology.

Bibliography

Adams, G.R., Gullotta, T.P., Montemayor, R. (Eds.), 1992. Adolescent Identity Formation. Sage, Newbury Park, CA.

Bada, M., 1995. The cultural ecology of corporation: explaining diversity in work group responses to organizational transformation. Journal of Applied Behavior Science 31 (2), 202–233.

Berry, J.W., 1966. Temne and Eskimo perceptual skills. International Journal of Psychology 1, 207–229.

Berry, J.W., 1971. Ecological and cultural factors in spatial perceptual development. Canadian Journal of Behavioural Science 3, 324–336.

Berry, J.W., 1976. Human Ecology and Cognitive Style: Comparative Studies in Cultural and Psychological Adaptation. Sage/Halsted, New York.

Berry, J.W., 2003. Ecological perspective on human psychological development. In: Saraswathi, T.S. (Ed.), Cross-cultural Perspectives in Human Development: Theory, Research and Applications. Sage, New Delhi, India, pp. 51–69.

Berry, J.W., 2011. The ecocultural framework: a stocktaking. In: Van de Vijver, F.J.R., Chasiotis, A., Breugelmans, S.M. (Eds.), Fundamental Questions in Cross-cultural Psychology. Cambridge University Press, Cambridge, pp. 95–114.

Bird, A., Fang, T., 2009. Editorial: cross cultural management in the age of globalization. International Journal of Cross Cultural Management vol. 9 (2), 1–5.

Bronfenbrenner, U., 1979. The Ecology of Human Development. Harvard University Press, Cambridge, MA.
Cohen, D., 2001. Cultural variation: considerations and implications. Psychological Bulletin 127.
Chou, L., 2009. Cross-cultural management: a neurolinguistic-cultural ecology approach. Journal of International Diversity 1 (1), 29–43.
Dasen, P.R., 2003. Theoretical frameworks in cross-cultural developmental psychology: an attempt at integration. In: Saraswathi, T.S. (Ed.), Cross-cultural Perspectives in Human Development: Theory, Research and Applications. Sage, New Delhi, India, pp. 128–165.
Everette, J.L., 1996. Toward a cultural ecology of organizing. Organization Communication: Emerging Perspective 3, 177–191.
Fang, T., 2009. Asian management research needs more self-confidence: reflection on hofstede (2007) and beyond. Asian Pacific Journal of Management 26 (3).
Ferguson, G., 1956. On transfer and the abilities of mankind. Canadian Journal of Psychology 10, 121–131.
Gallimore, R., Weisner, T.S., Kaufman, S.Z., Bernheimer, L.P., 1989. The social construction of ecocultural niches: family accommodation of developmentally delayed children. American Journal of Mental Retardation 94, 216–230.
Golden, D., 2005. Childhood as protected space? Vulnerable bodies in an Israeli Kindergarten. Ethos 70 (1), 79–100. http://dx.doi.org/10.1080/00141840500048557.
Grotevant, H.D., Cooper, C.R., 1998. Individuality and connectedness in adolescent development: review and prospects for research on identity, relationships, and context. In: Skoe, E., von der Lippe, A. (Eds.), Personality Development its Adolescence: A Cross National and Life Span Perspective. Routledge, London, pp. 3–37.
Hofstede, G., 2001. Culture's Consequences: Comparing Values, Behaviours, Institutions, and Organizations across Nations. Sage Publications, Thousand Oaks.
Irvine, S.H., Berry, J.W. (Eds.), 1988. Human Abilities in Cultural Context. Cambridge University Press, Cambridge, UK.
Jahoda, G., 1995. The ancestry of a model. Culture & Psychology 1, 11–24.
Keller, H., 2007. Cultures of Infancy. Erlbaum, Mahwah, NJ.
Kagitcibasi, C., 2007. Families, Self and Human Development across Cultures: Theory and Applications. Erlbaum, Mahwah, NJ.
McCarthy, E., 1998. The dynamics of culture, organizational culture and change. Artificial Intelligence and Society 12 (3), 155–184.
McLloyd, V.C., 1991. What is the study of African American children. In: Jones, R.J. (Ed.), Black Psychology, third ed. Cobb & Henry, Berkeley, CA, pp. 419–440.
Mbiti, J.S., 1990. African Religions and Philosophy. Heinemann Educational, Oxford.
Moumouni, A., 1968. Education in Africa. Praeger, New York.
Mtonga, M., 2012. Children's Games and Play in Zambia. University of Zambia Press, Lusaka, Zambia.
Ngaujah, D.E., 2003. An eco-cultural and social paradigm for understanding human development: a (West African) context. Graduate Seminar Paper. Biola University, CA, USA.
Nsamenang, A.B., 1992. Human Development in Cultural Context: A Third World Perspective. Sage, Newbury Park.
Nsamenang, A.B., 2002. Adolescence in Sub-Saharan Africa: an image constructed from Africa's triple inheritance. In: Brown, B.B., Larson, R.W., Saraswathi, T.S. (Eds.), The World's Youth: Adolescence in Eight Regions of the Globe. Cambridge University Press, London, pp. 61–104.
Nsamenang, A.B., 2004. Cultures of Human Development and Education: Challenges to Growing up African. Nova, New York.
Nsamenang, A.B., 2006a. Human Ontogenesis: an indigenous African view on development and intelligence. International Journal of Psychology 41, 293–297.
Nsamenang, A.B., 2006b. The intersection of traditional African education with school learning. In: Swartz, L., de la Rey, C., Duncan, N. (Eds.), Psychology: An Introduction. Oxford University Press, pp. 327–337.
Nsamenang, A.B., 2008a. Agency in early childhood learning and development in Cameroon. Contemporary Issues in Early Childhood 9 (3).
Nsamenang, A.B., 2008b. Culture and human development. International Journal of Psychology 43 (2), 73–77. http://dx.doi.org/10.1080/00207590701859093.
Nsamenang, A.B., 2009. Conceptualizing developmental assessment within Africa's cultural settings. In: Grigorenko, E.L. (Ed.), Multicultural Psychoeducational Assessment. Springer, New York, pp. 95–131. http://www.springerpub.com.
Nsamenang, A.B., 2011. L'émigration clandestine des jeunes camerounais en Europe. In: Bolzman, C., Gakuba, T.-O., Guissé, I. (Eds.), Migrations des jeunes d'Afrique subsaharienne: Quels defis pour l'avenir? L'Harmattan, Paris, pp. 139–161.
Nsamenang, A.B., 2012a. On researching the agency of Africa's young citizens: issues, challenges, and prospects for identity development. In: Slaughter-Defoe, D.T. (Ed.), Racial Stereotyping and Child Development. Contributions to Human Development, vol. 25. Karger, Basel: Switzerland, pp. 90–104.
Nsamenang, A.B., 2012b. Intergenerational transmission of values. In: Valsiner, J. (Ed.), Oxford handbook of Culture and Psychology. Oxford University Press, Cambridge, MA.
Nsamenang, A.B., 2013. Toward moving 'race-acting' research into a global perspective. Culture & Psychology 0 (0), 1–12.
Nsamenang, A.B., Akum, R.A., 2013. Cultures in the biology of human development. In: Tchombe, T.S.M., Nsamenang, A.B., Keller, H., Fulop, M. (Eds.), Cross-cultural Psychology: An Africentric Perspective.
Nsamenang, A.B., Lamb, M.E., 1993. The acquisition of socio-cognitive competence by Nso children in the Bamenda Grassfields of Northwest Cameroon. International Journal of Behavioral Development 16 (3), 429–441.
Nsamenang, A.B., Lamb, M.E., 1995. The force of beliefs: how the parental values of the Nso of northwest Cameroon shape children's progress towards adult models. Journal of Applied Developmental Psychology 16 (4), 613–627. http://dx.doi.org/10.1016/0193-3973(95)90007-1.
Ogbu, J., 1992. Understanding cultural diversity and learning. Educational Researcher 21 (8), 4–14.
Poortinga, Y., 2011. Conclusion. In: van de Vijver, F.J.R., Chasiotis, A., Breugelmans, S.M. (Eds.), Fundamental Questions in Cross-cultural Psychology. Cambridge University Press, London.
Remer, R., 2005. Cultural nuances, assumptions, and the butterfly effect: addressing the unpredictability Caused. Counseling and Values 51 (2), 93–110.
Steward, J.H., 1955. Cultural Change: The Methodology of Multilinear Evolution. University of Illinois Press, Urbana, Chicago.
Super, C.M., Harkness, S., 1986. The developmental niche: a conceptualization at the interface of child and culture. International Journal of Behavioral Development 9, 545–569.
Tharp, R.G., Gallimore, R., 1988. Rousing Minds to Life: Teaching Learning and Schooling in Social Context. Cambridge University Press, Cambridge, England.
Thomas, D.C., 2008. Cross-cultural Management: Essential Concepts, second ed. Sage, Los Angeles.
Trompenaars, F., Hampden-Turner, C., 1977. Riding the Waves of Culture: Understanding Cultural Diversity in Business. Nicolas Brealey, London.
Triandis, H., Suh, E.M., 2002. Cultural influence on personality. Annual Review of Psychology 53, 133–160.
UNICEF, 2003. Convention on the Rights of the Child. Retrieved from: www.unicef.org/crc/crc.htm.
Weisner, T.S., 1984. Ecocultural niches of middle childhood: a cross-cultural perspective. In: Collins, W.A. (Ed.), Development during Middle Childhood: The Years from Six to Twelve. National Academy Press, Washington, DC, pp. 335–360.
Weisner, T.S., Gallimore, R., Jordan, C., 1988. Unpackaging cultural effects on classroom learning: native Hawaiian peer assistance and child-generated activity. Anthropology and Education Quarterly 19, 327–351.
Weisner, T.S., 2002. Ecocultural understanding of children's developmental pathways. Human Development 45 (4), 275–281.
Whiting, B., 1976. The problem of the packaged variable. In: Riegel, K., Meacham, J. (Eds.), The Developing Individual in a Changing World: Historical and Cultural Issues, vol. 1. Mouton, The Hague, Netherlands, pp. 303–309.
Worthman, C.M., 1994. Developmental microniche: a concept for modelling relationships of biologu, behavior and culture in development. American Journal of Physical Anthropology 210 (Suppl. 18).
Zimba, R.F., 2002. Indigenous conceptions of childhood development and social realities in Southern Africa. In: Keller, H., Poortinga, Y.P., Scholmerish, A. (Eds.), Between Cultures and Biology: Perspectives on Ontogenetic Development. Cambridge University Press, Cambridge, pp. 89–115.

Ecofeminism

Susan Buckingham, Centre for Human Geography, Brunel University, Uxbridge, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S. Buckingham-Hatfield, volume 9, pp. 5907–5910, © 2001, Elsevier Ltd.

Abstract

This article summarizes the history of ecofeminism and its various strands of activism and intellectual inquiry. Through critiques made of ecofeminism, both from those allying themselves with the movement, and from those who wish to disassociate themselves from it, the argument is made that ecofeminism, particularly in its social constructivist form, has been influential in international policy making. As a parallel development alongside feminist political ecology and other environmental feminisms it has developed analytically, infused by renewed interest by a new generation of academics and activists, as well as a new generation of environmental concerns, dominated by climate change.

Introduction

Ecofeminism, like the social movements it has emerged from, is both political activism and intellectual critique. Bringing together feminism and environmentalism, ecofeminism argues that the domination of women and the degradation of the environment are consequences of patriarchy and capitalism. Any strategy to address one must take into account its impact on the other so that women's equality should not be achieved at the expense of worsening the environment, and neither should environmental improvements be gained at the expense of women. Indeed, ecofeminism proposes that only by reversing current values, thereby privileging care and cooperation over more aggressive and dominating behaviors, can both society and environment benefit.

The notion that women's and environmental domination are linked has been developed in a number of ways. A perspective in which women are accredited with closer links with nature was celebrated in early ecofeminist writings, by, for example, Carolyn Merchant in the United States and Val Plumwood in Australia. These advocated 'the feminine principle' as an antidote to environmental destruction, through attributes, which nurture nature. This 'essentialist' perspective, often adopting an ideal of woman as earth mother/goddess, has, however, also discredited ecofeminism and led to disaffection among some early protagonists (see, for example, Janet Biehl). In addition to being critiqued for its essentialism, this view of ecofeminism has also been charged with elitism through its provenance in a white, middle-class, Western, milieu. However, Vandana Shiva's consistent and persuasive 'majority world' voice has been a counterpoint to this, and arguably, gender and environment have been articulated together more powerfully, and been more influential, in majority world settings (see, for example, Wangari Maathai in Kenya), although how this has been done has been questioned by writers such as Cecile Jackson and Melisssa Leach.

The late 1980s and early 1990s was a fertile time for ecofeminist writing, both from this essentialist perspective, but also through more social economic critiques, which explained the link between women's inequality and environmental degradation in terms of women's role in social reproduction (see, for example, Mary Mellor in the UK and Marilyn Waring in New Zealand).

The first major practical impact of ecofeminist thinking was felt in the 1992 United Nations Conference on Environment and Development (UNCED), which women's environmental organizations had lobbied for women's and environmental rights to be considered in tandem. This, and the 1995 4th Women's Conference in Beijing, agreed for the first time that women's rights and environmental rights could not be disentangled.

By the late 1990s, however, the output on ecofeminism had dropped significantly and there was a sense of it having run out of steam, despite its arguable influence revealed through UN initiatives. However, in the new millennium, a new generation of writers, researchers, and activists has reinvigorated ecofeminist debates, through considerations of ecofeminist citizenship (Sherilyn MacGregor), challenges to some earlier critiques (Niamh Moore), and insertion of feminist concerns into environmental justice (Susan Buckingham and Rakibe Kulcur, and Giovanna di Chiro) and political ecology (Wendy Harcourt, Dianne Rocheleau).

This article considers the heritage of ecofeminism as a multiply braided political praxis and an intellectual position. It examines key critiques of earlier perspectives, before exploring its more recent developments. It considers its relationship with, and potential to enhance other feminist and environmental approaches, particularly those concerned with feminist political ecology and environmental justice. The article concludes with a consideration of how ecofeminism is enjoying a resurgence through a new generation of academics seeking to develop and nuance ecofeminism from a sympathetic position, the emergence of climate change as a major global issue, and the development of social movements in areas not previously associated with feminist environmental action, notably in the Middle East.

A Brief History of Ecofeminism

Ecofeminism is not a singular position, and protagonists have emerged from different intellectual and political traditions. It is generally agreed that the first reference to 'ecofeminism' was by

the radical French feminist, Francoise d'Eaubonne, yet, despite her influence, particularly in North America and Australia, "Le féminisme ou la mort," (1974) and "Écologie, féminisme: révolution ou mutation?" (1978) have not been translated into English. It is also ironic, given that many of the Anglo-American/Anglo-Australian ecofeminists initially eschewed d'Eaubonne's Marxist–feminist analysis to focus on an essentialist/cultural approach.

Essentialism

This approach sought to reverse the Enlightenment/Cartesian hierarchy privileging a 'masculine' abstract rationality, which was seen to justify the exploitation of anything weaker than itself (colonial subjects, men without property, children, as well as women and the natural world). Thus, early ecofeminists elevated the 'feminine' principles of care and cooperation, not only to parity with the 'masculine,' but above it, on the basis that all human and nonhuman life would better thrive, guided by these principles (for example, Diamond and Orenstein, Merchant, and Plumwood). The most celebratory advocates of this approach (such as Mary Daly, Andree Collard, and Charlene Spretnak) argued that female biology tied women more closely to nature, primarily as they menstruated, and carried, gave birth to, and breastfed children. This, the cultural ecofeminists argued, enables women to make more sensitive decisions regarding humans' relationship with nature. The care needed to bear and nurture children and the households in which they develop provided the basis of an 'ethic of care' seen as necessary for the prevention of widespread environment destruction. This approach has been criticized on a number of grounds, as will emerge through this article, but one of the most persuasive is that for centuries, women have been relegated to the caring and domestic spheres precisely because of their essentialized biologies. It is not surprising, therefore, that many feminists argued that biology is insufficient grounds on which to determine justice, rights, and opportunities in any sphere.

Social Construction

More consistent with d'Eubonne's first articulation of ecofeminism has been the reasoning that women's typical work, socialized through generations, places them in more direct contact with nature. Whoever deals immediately with the social reproductive acts of provisioning, cooking, watering, clothing, cleaning and so on comes into daily contact with nature. This is most obvious in rural areas of the majority world where it is mostly women who walk to collect water, and biomass for cooking and heating water. Reliance on biomass for cooking on open stoves also makes women more vulnerable to respiratory diseases caused by indoor air pollution. It is also women who tend to be the subsistence farmers, and who therefore are most immediately impacted when structural adjustment policies require subsistence farmers to switch to cash crops, or where rights to harvest common land for wood, seeds, and other subsistence needs are suspended as the forest comes under national protection. It is these rural women's tasks, and their expertise as seed collectors, which Vandana Shiva has powerfully used in her ecofeminist analyses and campaigns against environmentally degrading practices, particularly in India. Likewise, Wangari Maathai, ecofeminist campaigner, academic and politician, drew on women's environmental knowledge and practices to develop the Green Belt tree planting movement in Kenya.

But the household division of labor is strikingly similar worldwide. Noted ecofeminist, and first woman and Green Party MP in New Zealand, Marilyn Waring, in her capacity as chair of the New Zealand Public Expenditure Select (Public Accounts and Budget) Committee, made a point of visiting women to inquire about their working day wherever she traveled. She was repeatedly struck by the ubiquitousness of women undertaking housework, caring for children and other dependents, and their relative lack of leisure time compared to men. A memorable vignette from a film made of her travels involved women in Tanzania giving a detailed account of their day, which started at sunrise to collect firewood and water, and ended with preparations for the next day's breakfast well after sunset. When Waring turned to a man standing at the edge of the group to ask him how he spent his day, his response that he 'supervised the women' drew much mirth. The persistent gendered vertical job segregation, wage inequalities, and distribution of caring responsibilities in the minority world, has stimulated a number of analyses, which link the exploitation of women with the exploitation of nature, through patriarchy and capitalism (see Mary Mellor, who, in her book 'Breaking Boundaries,' aspired to a 'feminist green socialism').

Challenges

The ecofeminist literature produced in the 1980s and early 1990s has been challenged on multiple fronts. Reference has already been made to the critique against essentialism. Both Gillian Rose (1993) and Val Plumwood (in her later work, 1993) questioned the wisdom of using what Plumwood terms an 'uncritical reversal' through which feminine virtues are elevated above masculine. Since the male/female duality is itself a product of 'masculinist' thinking, which has constructed the idea of the feminine it makes little sense to employ a product of such hegemonic construction. More recently, MacGregor questions the degree to which the feminine attributes invoked by ecofeminism are intrinsic, for "we simply do not know what 'feminine morality' would be under conditions of equity and freedom and we should not confuse actions shaped under socially oppressive conditions for 'natural' ones" (2004: p. 63).

Some ecofeminist writers have attempted to think beyond dualisms, particularly in their later work, and they do so by trying to rework the ethic of care into a concept which is beyond gender, and beyond duality. An example of this is Merchant's (1996) partnership ethic between men and women and Warren's (1994) prioritization of interpersonal relationships of mutuality. MacGregor would go further and jettison the concept of care as a foundation of ecofeminism in favor of a rights- and justice-based ethic of citizenship (I develop MacGregor's argument below). That she does so within the framework of ecofeminism is interesting, and contrasts with a position taken by Janet Biehl 13 years earlier. Writing from a social ecology position (to which she no longer adheres), in her 1991 book 'Rethinking Ecofeminist Politics,' Biehl publicly

renounced her ecofeminist past, arguing that, by denying aspects of Western culture, ecofeminism therefore "ignores or rejects legacies of democracy, of reason, and…the project of scientifically understanding much of the natural world as part of a radical liberatory movement" (1991: pp. 1, 2). These charges are leveled against an essentialist perspective and its internal contradictions and Biehl is clearly discomforted by what she sees as incoherence in a social movement, which needs to be "sorted out" (1991: p. 3). Her initial identification with ecofeminism stemmed from an attempt to show how women's concerns could be integrated into social ecology, which argues that the domination of nature is tied to hierarchies and class structures in human society. Dismantling these structures, she suggests, will "enable women to fulfil themselves completely, not only as gendered beings but as human beings" (1991: p. 5). While Biehl's frustration with the apparent incoherence of essentialist ecofeminism is understandable, her expectation that all self-proclaimed ecofeminists should be constructing a coherent argument, or battling out their differences in public could also be construed as trying to impose a rationality that ecofeminism itself challenges.

Iris Marion Young proposes that "essentialism is a dilemma for feminism" in that "if 'women' is not the name of a specific social collective, then there seems to be no basis for a specifically feminist politics. On the other hand, any attempt to define women as a group with common attributes either is absurdly reductionist or normalizes some of those meant to be included in the definition while marginalizing or excluding others" (1997: pp. 5, 6). While not written within the context of ecofeminism, this quote from a leading feminist philosopher emphasizes the difficulties related to essentialism, which cannot so easily be resolved by abandoning essentialism entirely as a form of analysis/praxis.

Since women's environmental roles have been recognized as a basis for policy by international organizations and development banks, partly as a result of ecofeminist writings, ecofeminism has drawn a sustained critique by a number of minority world academics working in the Global South, notably Cecile Jackson and Melissa Leach. In a reflection on her own work challenging ecofeminism, Leach castigates ecofeminism with failing to recognize the differences between women, and the risks to women of assigning them the "responsibility for 'saving the environment' without addressing whether they actually had the resources and capacity to do so" (2007: p. 72). This reading of ecofeminism: that it focused too much on women (rather than on gender relations), that it was overly simplistic, and that this had influenced a generation of development and environmental policies, has in turn been challenged (see Moore, 2008).

Leach had also argued that alternative lenses to viewing gender relations and their intersection with the environment, such as feminist environmentalism (as developed by Bina Agarwal) feminist political ecology (as articulated by Dianne Rocheleau) or gender, environment, and development (as propounded by Cecile Jackson) are superior to, and distinct from ecofeminism. However, these can also be seen as developments of a continuation of ecofeminist thinking, which is not fossilized in its 1980s form.

Since the 2000s there has been a more generous treatment of ecofeminism, which recognizes that it is not limited to the conceptualizations by its critics, and offers a starting point for further work. In 2007, Leach had herself noted that the critiques she and her colleagues had made may have contributed to a decline in women and gender relations being included in environment and development discourse.

Impacts of Ecofeminism

The inclusion of a chapter in Agenda 21 on the significance of women is an indication of the influence of ecofeminism in the early 1990s. This particular achievement was largely the result of successful lobbying of the UNCED (UNCED, 1992) by a coalition of women's groups brought together by Women's Environment and Development Organization (WEDO). Through the World Women's Congress for a Healthy Planet held in Miami in 1991, prominent groups and individuals, including some already well known for their ecofeminist perspectives, Wangari Maathai and Vandana Shiva, from both the Global North and Global South, developed a blueprint for women's involvement in environmental decision-making in 1991: Women's Action Agenda 21. This laid the foundation for considering women's rights at a number of international conferences in the 1990s. The 4th UN conference on women held in Beijing in 1995 also linked gender and environmental issues through its 'Action K,' which insisted on the integration of gender concerns into environmental decision-making, policies, and programs for sustainable development and assessments of development and environmental policies. Out of this emerged a UN (and EU) commitment to 'gender mainstreaming,' which requires all government policies and practices to take into account their gender impacts. How far these commitments have been enacted into national policies and practices, and whether they have made a difference to decision-making, is debatable. UN Women's assessment of how women's situation has changed as a result of the Millennium Development Goals reveals some positive, but very slow, change, as well as indications that women's involvement in decision-making influences the decisions taken. For example, UN Women reports that in female-led councils in India, the number of drinking water projects was 62% higher than in male-led councils. Climate change policy and decision-making has been resistant to reflecting gender concerns. Nevertheless, in 2012, the annual meeting of the conference of the parties (COP18 in Doha) recognized the importance of gender balance in formal and informal decision-making to enact more effective climate change policy, and agreed to set targets to achieve this. This follows a number of initiatives set up by European governments to achieve minimum quotas of women on corporate boards, post-2000, although sectors most implicated in environmental problems (such as transport, waste, energy, water) have some of the lowest proportions of women on their boards.

Arguably, then, ecofeminism, alongside broader feminist campaigns, has been influential in environmentally related policy, at least internationally. In terms of its influence in other environmental movements, ecofeminism has failed to make significant impacts, although there are indications of its capacity to do so in more recent work by proponents who stress their ecofeminist heritage.

Ecofeminism in Other Environmentalisms

In many ways the environmental movement remains resolutely masculine, white, and middle class. This had been observed in the United States by Joni Seager in the 1990s when she reported on the lack of women (not to mention women of color) in the mainstream North American environmental movement. There is public acknowledgment that company boards in the UK and scientific research are hampered by a dominant masculinity and that more evenly mixed gender composition will generate more diverse and effective decision-making by incorporating a variety of male and female perspectives (European Commission, 2012; UK Government, 2011). However, this has not percolated through to the environmental movement, which is still characterized by a culture of long hours, inflexible working practices, and a masculinist culture, well revealed by Rakibe Kulcur in the UK and Turkey. As will be discussed later, Iris Marion Young calls into question existing decision-making structures and argues for these to be transformed into bodies in which men and women can communicate effectively with each other in nondiscriminatory ways.

The Gender Justice in Environmental Justice?

The environmental justice movement has run parallel to the ecofeminist movement since the 1980s and also involves advocacy and an intellectual analysis of how disadvantaged populations are disproportionately exposed to environmental problems. This has been revealed at a global scale, through ways in which the global north 'dumps' on the global south, but has been particularly powerful in exposing the environmental racism against African-American and Hispanic communities in the United States. However, with notably few exceptions, gender is given no room in these analyses. Those case studies, which do focus on women's experiences of environmental injustice draw, both implicitly and explicitly, on ecofeminism, particularly through women's roles in the household and community. Lois Gibbs' through her celebrated campaign against governments failing to recognize the evidence she collected regarding the industrial pollution of a local school and increased rates of cancer and other diseases of her family and neighbors has become a 'poster child' of the environmental justice movement. However, a gender analysis of other environmental injustices is rare (but see Hilda Kurtz and Judy Sze for exceptions). Judy Sze has related how African-American lone mothers in New York's Harlem were castigated by City Hall as being poor mothers and housewives when they made complaints about the polluted streets they suspected of causing increased rates of asthma among their children (Sze, 2004). Both Gibbs and Kurtz report women's concerns being dismissed by city officials as those of 'hysterical housewives.' The environmental justice movement, then, like the environmental movement more generally, is not sufficiently sensitive to gender.

Interestingly, academics such as David Schlosberg and Gordon Walker have drawn on the writings of feminist philosopher Iris Marion Young (1990) in developing a theoretical analysis of environmental justice, which goes beyond the distributional, to include representative and participative justice. However, her deeper critique, which analyses gender relations through the concepts of domination and oppression, rather than the concept of distribution, finds little expression in environmental justice. While environmental justice is not a term Young uses, she writes extensively on women's health and reproductive rights, which Giovanna di Chiro argues constitute a legitimate concern for environmental justice activity, in a broad sense. Di Chiro detects a masculinism in mainstream environment movements, which generates silos in which 'environment' is constructed as something other than the human and the social, thereby excluding a lot of 'environmental' concerns and activism of women involved in, for example, the reproductive rights movement (Di Chiro, 2008: p. 279).

Feminist Political Ecology

Melissa Leach, in dismissing ecofeminism, suggests that feminist political ecology was framed as an alternative perspective to ecofeminism. In their 1996 book *Feminist Political Ecology*, Dianne Rocheleau, Barbara Thomas-Slayter, and Esther Wangari are less dismissive, setting out their aim as "an explicit effort to join feminist and political ecology scholarship from the ground up, based on case studies compiled from collaborative field research or direct participation in the events recounted" (1996: p. xvi). The authors identify ecofeminism as one of five "major schools of feminist scholarship and activism on the environment," on which they seek to draw to explicate a "new conceptual framework, which we call feminist political ecology" (p. 3). However, the divisions between the five (the other four of which are feminist environmentalism, socialist feminism, feminist poststructuralism, and environmentalism) are not clear. Even in the early 1990s, writers such as Mary Mellor and Ariel Salleh, writing in Capital, Class and Nature, were crossing these boundaries. This feminist political ecology 'manifesto' can be seen as calling for the contextualizing of ways in which different women, and groups of women, experienced the 'environment,' which writers identifying as ecofeminists are now doing (see, for example, MacGregor, below).

Feminist political ecologist, Wendy Harcourt, pays particular attention to the body as "the place closest in" (2009: p. 22) and argues that the women she has worked with in social movements are "keenly aware of the conditions of their embodiment, including health, sexuality, security, the right to pleasure and rest, and the need to care for community and environment…". As women mobilize for change around the body, home, local environs, and community, Arturo Escobar and Wendy Harcourt argued they were creating a form of politics, they called 'place based,' with the body as the first "place" (p. 23). While not expressly referring to ecofeminism, this position is a strong echo of it, and has much to offer it.

The Future for Ecofeminism?

Following a period of high profile activism in the late 1980s and early/mid-1990s, during which ecofeminism appeared to have influence on the international development and environmental agenda, it fell out of favor. Unfairly, perhaps,

accused of the essentialism, which characterized some, but by no means all, of its early expression, the number of ecofeminist publications dropped, and ecofeminists seemed unable to respond effectively to challenges. More recently, however, younger writers have declared a sympathy with the approach and, while not uncritical of earlier ecofeminist writing, have chosen to see this as a starting point for a sustained feminist critique of environmental problems.

Ecofeminist Citizenship

Sherilyn MacGregor (2004) has argued that ecofeminism's emphasis on 'caring' relegates this to the private and 'nonpolitical' realm. A failure to engage in the politics of social reproduction maintains the fundamental divisions of labor, and reifies women's work, and therefore gender inequality. While recognizing the achievement of ecofeminism in inserting care work, conducted in the private sphere of the home, and the intermediate community sphere, into ecopolitical discourses, which had hitherto ignored it, MacGregor pushes ecofeminism to relate to broader theories of justice (see also Catriona Sandilands). MacGregor argues, after Joan Tronto, and also implicitly following Carolyn Merchant, for care "to be seen as integral to any notion of a good society" (MacGregor, 2004: 75) and therefore part of everyone's remit. Women's relationship to caring and domestic work is further complicated by class and income and it is important to note that the 'equity and freedom' of some women to enter the paid/market economy is predicated on being able to employ other women to undertake the care work, often on the margins of a livable wage, previously undertaken unpaid (Gregson and Lowe, 1994).

Notwithstanding MacGregor, care work has been politicized, with writer-campaigner Selma James' 'International Wages for Housework' campaign launched in 1972 to demand recognition that housework is a foundation for paid work in the market economy. Arguably, the most intimate of exchange, sex, can also be a political commodity, withheld to achieve political ends. In one of a long line of such campaigns, dating back to ancient Greece, women in Barbacoas in Colombia have been reported in the Washington Post as using sex as a public bargaining strategy: in this case, refusing sex with their partners until repairs have been made to a road frequently impassable, particularly in bad weather.

Climate Change

Increasing attention to climate change through the 2000s has also generated interest in its governing structures and it can be noted that gender forms an extremely small and marginalized part of the debate, compared to that on sustainable development in the 1990s. However, WEDO was a partner in a campaign, GenderCC, to ensure that gender is taken into account in climate change decision-making, mitigation, and adaptation. A recent milestone was achieved at the 18th meeting of the Conference of the Parties (COP18) in Doha, which adopted a decision on "promoting gender balance and improving the participation of women in UNFCCC negotiations and in the representation of Parties in bodies established pursuant to the Convention or the Kyoto Protocol" (UNFCCC, 2012).

Joni Seager, writing in a special issue of *Women, Gender and Research* (a Danish journal published in English) in 2009 ahead of the Copenhagen climate change summit, castigates the political and scientific communities for "[o]ne of the core illusions of the 2° target" suggesting that "humans can 'master' climate change, allowing the global temperature to rise to an arbitrary line and then stopping it" (2009: p. 14). Citing earlier ecofeminists such as Merchant, Plumwood and Warren, Seager invokes their 'trenchant analyses' of attitudes such as this. Ecofeminism has also been cited as an explicit organizing concept in environmental and climate change law articles and books (Morrow, 2012), suggesting a widening of ecofeminism's influence a decade into the twenty-first century.

Activism

Prodemocracy movements that defined the 'Arab Spring' in 2011/2012 have included a number of campaigns which Green Prophet (a Website claiming to be the 'sustainable news for the Middle East') have branded as 'ecofeminism.' These range from the more conventionally environmentalist solar power projects, in which Bedouin women in Jordan are installing solar panels in desert communities, to those 'crossover' projects described by di Chiro as integrating social and environmental issues, often ignored by the environmental mainstream, including a 'harass map' to enable women cyclists in Cairo to report abuse while cycling, through SMS.

Meanwhile, women in Colombia, Nigeria, and the United States have protested against decisions to exploit communities and the environment, using their bodies (threatening to strip naked, disrobing, or withholding sex from their partners) and language that would be familiar to earlier ecofeminists (see Websites below).

In Conclusion

It is possible to argue that ecofeminism has survived and evolved from its inception in the 1970s and 1980s. Activists and a new generation of academics cite ecofeminism as an inspiration, and seek to clarify and strengthen its arguments through more empirical and case study work, and multidisciplinary analysis. Some critiques of ecofeminism in the mid- and late-1990s into the 2000s could be accused of being antifeminist in their apparent ungenerous and undermining attempts to challenge masculinist hegemonic thinking. However, those seeking to distance themselves from ecofeminism have, in turn, stimulated debate and development, including a defense of ecofeminism and a willingness to critique it in more constructive and inclusive ways. While ecofeminist organizations and alliances have gained some influence in international policy, it is also clear that gender inequality and environmental degradation continue and in some cases have worsened, revealing the continued need for an ecofeminist analysis.

See also: Environment and Development; Environmental Justice in the United States; Environmental Movements; Social Ecology.

Bibliography

Biehl, J., 1991. Rethinking Eco-Feminist Politics. South End Books, Boston.

Buckingham, S., Kulcur, R., 2009. Gendered geographies of environmental injustice. Antipode 41 (4), 659–683.

D'Eaubonne, F., 1974. Le feminism ou la mort. P. Ouray, Paris.

D'Eaubonne, F., 1978. Ecologie, Feminisme: Revolution Ou Mutation? Edition A.T.P, Paris.

Diamond, I., Orenstein, G. (Eds.), 1990. Reweaving the World: The Emergence of Ecofeminism. Sierra Club Books, San Francisco.

Di Chiro, G., 2008. Living environmentalisms: coalition politics, social reproduction and environmental justice. Environmental Politics 17 (2), 276–298.

Davies, Lord, M., 2011. Women on Boards London, UK Government. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/31480/11-745-women-on-boards.pdf.

European Commission, 2012 (6 COM(2012) 614. http://ec.europa.eu/justice/gender-equality/gender-decision-making/index_en.htm).

Gibbs, L., 1998. Love Canal: The Story Continues. New Society Press, Gabriola Island, British Columbia, Canada.

Gregson, N., Lowe, M., 1994. Servicing the Middle Classes: Class, Gender and Waged Domestic Labour in Contemporary Britain. Routledge, London.

Harcourt, W., 2009. Body Politics in Development. Critical Debates in Gender and Development. Zed Books, London.

Jackson, C., 1993. Women/nature or gender/history? A critique of ecofeminist "development". Journal of Peasant Studies 20 (3), 389–419.

Kulcur, R., 2013. Environmental injustice? An analysis of gender in Environmental Non-governmental Organisations (ENGOs) in the United Kingdom and Turkey (Ph.D. thesis). Brunel University, London. http://bura.brunel.ac.uk/handle/2438/7680.

Kurtz, H., 2007. Gender and environmental justice in Louisiana: blurring the boundaries of public and private spheres. Gender, Place and Culture 14 (4).

Leach, M., 2007. Earth mother myths and other ecofeminist fables: how a strategic notion rose and fell. Development and Change 38 (1), 67–85.

MacGregor, S., 2004. From care to citizenship: calling ecofeminism back to politics. Ethics and the Environment 9 (1), 56–84.

MacGregor, S., 2006. Beyond Mothering Earth: Ecological Citizenship and the Politics of Care. University of British Columbia Press, Vancouver.

Maathai, W., 1985. The Green Belt Movement: Sharing the Approach and the Experience. Lantern Books, New York.

Mellor, M., 1992. Breaking the Boundaries. Towards a Feminist Green Socialism. Virago, London.

Merchant, C., 1990. The Death of Nature. Women, Ecology and the Scientific Revolution. Harper & Row, New York.

Merchant, C., 1996. Earthcare: Women and the Environment. Routledge, London.

Morrow, K., 2012. Climate change, major groups and the importance of a seat at the table: women and the UNFCC negotiations. In: Penca, J., de Andrade Correa (Eds.), The Dominance of Climate Change in Environmental Law: Taking Stock for Rio+20. European University Institute, Robert Schumann Centre for Advanced Studies, Florence, Italy.

Moore, N., 2008. The rise and rise of ecofeminism as a development fable: a response to Melissa Leach's 'Earth mothers and other ecofeminist fables: how a strategic notion rose and fell'. Development and Change 39 (3), 461–475.

Plumwood, V., 1993. Feminism and the Mastery of Nature. Routledge, London.

Rocheleau, D., Thomas-Slater, B., Wangari, E., 1996. Feminist Political Ecology, Global Issues and Local Experiences. Routledge, London.

Rose, G., 1993. Feminism and Geography: The Limits of Geographical Knowledge. Polity Press, Cambridge.

Salleh, A. (Ed.), 2009. Eco-Sufficiency and Global Justice: Women Write Political Ecology. Pluto Press, London.

Seager, J., 2009. Death by degrees: taking a feminist hard look at the 2° climate policy. Women and Gender Research 3-4, 11–21.

Shiva, V., 1989. Staying Alive. Women, Ecology and Development. Zed Books, London.

Sze, J., 2004. Gender, asthma politics, and urban environmental justice activism. In: Stein, R. (Ed.), New Perspectives on Environmental Justice. Gender, Sexuality and Activism. Rutgers University Press, New Brunswick, New Jersey, pp. 177–190.

United Nations, 1992. Agenda 21. Chapter 24 Global Action for Women towards Sustainable and Equitable Development. United Nations, New York.

UNFCCC, 2012. Decision 23/CP.18. Promoting gender balance and improving the participation of women in UNFCCC negotiations and in the representation of Parties in bodies established pursuant to the Convention or the Kyoto Protocol FCCC/CP/2012/8/Add.3.

Waring, M., 1988. Counting for Nothing. What Men Value & What Women Are Worth. Allen & Unwin, Wellington, New Zealand.

Warren, K., 1994. Ecological Feminism. Routledge, London.

Young, I.M., 1997. Intersecting Voices. Dilemmas of Gender, Political Philosophy, and Policy. Princeton University Press, Princeton, NJ.

Young, I.M., 1990. Justice and the Politics of Difference. Princeton University Press, Princeton, NJ.

Relevant Websites

http://www.baringwitness.org/ – Baring Witness (accessed 03.01.14.).

http://ec.europa.eu/research/science-society/document_library/pdf_06/structural-changes-final-report_en.pdf – European Commission, 2012. Structural Change in Research Institutions: Enhancing Excellence, Gender Equality and Efficiency in Research and Innovation.

http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0614:FIN:en:PDF.

http://www.greenprophet.com/2011/04/arab-women-eco-feminism/ – Green Prophet (accessed 04.01.14.).

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/31480/11-745-women-on-boards.pdf – UK Government, 2011. Women on Boards.

http://www.un.org/womenwatch/daw/beijing/platform/environ.htm – United Nations, 1995. Fourth Conference on Women Platform for Action K. Women and the Environment.

http://unfccc.int/gender_and_climate_change/items/7537.php – UNFCC, 2012. Doha Declaration.

http://www.washingtonpost.com/blogs/blogpost/post/crossed-legs-protest-in-colombia-an-attempt-to-to-receive-basic-rights/2011/08/03/gIQAMKfAsI_blog.html – Washington Post, 2011. 'Crossed Legs' Protest in Colombia an Attempt to Receive 'basic Rights' by Katie Rogers (accessed 23.12.13.).

http://www.wedo.org/ – Women's Environment and Development Organization.

www.wen.co.uk – Women's Environmental Network.

Ecological Economics

Joan Martinez-Alier, After Universitat Autonoma de Barcelona, Barcelona, Spain; and FLACSO, Quito, Ecuador

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ecological economics was constituted in the late 1980 as a transdisciplinary field of study attracting systems ecologists and dissident economists. Inspiration was drawn from N. Georgescu-Roegen's *The EntropyLaw and the Economic Process* (1971) together with work from ecologist H.T. Odum and economist K. Boulding. The complexity of living structures is achieved by 'capturing' energy through photosynthesis and by dissipating energy to outside systems. The industrial economy, however, does not work only by using current photosynthesis. It burns irreplaceable stocks of fossil fuels, and it produces irreversible damage to the natural environment. The scale of the economy is too large; therefore the natural cycles cannot sustainably produce the resources, or absorb or assimilate the waste such as, for instance, heavy metals or excessive carbon dioxide. Ecological economics encompasses money valuation of positive environmental services and of negative externalities, and also physical appraisals of the environmental impacts of the human economy measured through new indicators. It also gives importance to social indicators. The study of the relations between property rights and management of natural resources is another important focus. Ecological economics favors multicriteria assessments over cost–benefit analysis, emphasizing incommensurability of values. It has also developed an ecological macroeconomics without growth.

Introduction: The Economy Embedded in Society and Nature

Ecological economics is a transdisciplinary field of study which sees the economy as a subsystem of a larger finite global ecosystem. Ecological economists question the sustainability of the economy because of its environmental impacts and material and energy requirements, and also because of the growth of population. Attempts at assigning money values to environmental services and losses thereby correcting macroeconomic accounting, are part of ecological economics, but its main thrust is rather in introducing physical indicators and indices of sustainability. Ecological economists also work on the relations between property rights and resource management, they model the interactions between the economy and the environment, they study ecological distribution conflicts, they use management tools such as integrated environmental assessment and multicriteria decision aids, and they propose new instruments of environmental policy.

From the beginning of ecological economics (Daly, 1984), there was a debate between three approaches. First, to go beyond the mere description of the human economy in ecological terms (flows of energy and materials, evolution of technologies in analogy to biological evolution) and search for a physical common value (such as energy value). Second, to follow the strategy of hanging a price tag on everything, using fictitious market prices when actual market prices are lacking. While the first and second approaches are reductionist, the third approach is not reductionist but transdisciplinary.

René Passet in 1979 in *L'économique et le vivant* drew an image of the relations between nature, human society, and the human economy (Figure 1) that has become a symbol for ecological economics. The drawing shows the obvious reality that there was nature before human society, and human society preceded the generalized market system by many generations. This vision has implications for economics. The teaching of the theory of externalities (i.e., the impacts of the economy on the environment which are not measured by market prices) should not wait for the time when students have grasped already the analysis of general market equilibrium. On the contrary, the study of the market (the chrematistics) should come after the study of ecology and social institutions. The 'externalities' come before the 'internalities'. The market economy could not exist

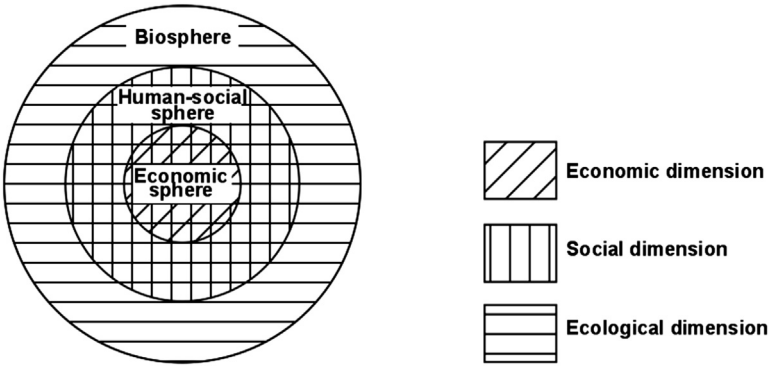

Figure 1 The economy embedded in the institutions of human society and in the biosphere.

without social institutions, and without the unpaid products and services of ecosystems.

Ecological economists see the economy as an open system. In thermodynamics, systems are classified as 'open' to the entry and exit of energy and materials, 'closed' to the entry and exit of materials though open to the entry and exit of energy, such as the Earth, and 'isolated' systems (without entry or exit of energy and materials). The availability of free energy and the cycling of materials allow life forms to become ever more organized and complex; the same applies to the economy. Dissipated energy and waste are produced in the process. If the scale of the economy is too large and its speed is too rapid, then the natural cycles cannot sustainably produce the resources or absorb or assimilate the waste such as, for instance, heavy metals or carbon dioxide.

In ecological economics the economy is seen as embedded in the ecosystem (or, more accurately, in the historically changing social perception of the ecosystem). The economy is also embedded in a structure of property rights on environmental resources and services, in a social distribution of power and income, in social structures of gender, social class, or caste. Instead, in mainstream economics the economy is seen as a self-sufficient system where prices for consumer goods and services and those for the services of production factors are formed.

Ecological economists (Norgaard, 1990) disputed the view expressed in the 1960s by Barnett, Krutilla, and other mainstream resource economists that since raw materials were cheap, they must be abundant. Markets are myopic; they discount the future, they cannot see future uncertain scarcities of sources or sinks. Ecological economists understand and even sympathize with attempts at 'internalizing' externalities into the price system, they readily concur with proposals to correct prices by taxes (such as 'natural capital depletion taxes' or taxes on pollution) but they deny that there exists a set of 'ecologically correct prices'.

History and Foundations

In the late nineteenth and early twentieth centuries, the biologist and urban planner Patrick Geddes, the chemist Frederick Soddy, the engineer and social reformer Josef Popper-Lynkeus had unsuccessfully tried to promote a biophysical view of the economy, as a subsystem embedded in a larger system subject to the laws of thermodynamics (Martinez-Alier and Schlüpmann, 1987). By 1850 or 1860 the carbon cycle and the cycles of plant nutrients had been discovered, while the first and second laws of thermodynamics (conservation and transformation of energy, but also dissipation of energy and increase in entropy) had been established.

The contrived conflict between the 'optimistic' theory of evolution, which explained the diversity of life, and the 'pessimistic' second law of thermodynamics was a staple of the cultural diet of the early 1900s. Praising the energy accounts of agriculture published in 1880 by the 'narodnik' revolutionary and physician S.A. Podolinsky, the ecologist Vernadsky wrote in 1924 (Vernadsky, 1924: pp. 334–335) that Podolinsky had analyzed the energetics of life (life systems being open to the entry of energy), and had applied these ideas to the analysis of the economy. Podolinsky wrote that, for an economy to be sustainable, the energy productivity of human work (i.e., how much energy is made available by 1 day of human work) must be higher (or equal, if everybody is working) than the efficiency of the transformation of the energy intake into human work. The energy productivity of a coal miner (wrote Podolinsky) was much larger than that of an agriculturalist could obtain but this energy surplus from fossil fuels was transitory (Podolinsky, 1880, 1883).

Therefore, the main ingredients for an ecological view of the economy were present much before the birth of a self-conscious ecological economics in the 1980s, which was delayed by the strict boundaries between the natural and the social sciences. The biologist and systems ecologist Alfred Lotka, born in 1880, had introduced in the 1910s and early 1920s the fundamental distinction between the endosomatic and exosomatic use of energy by humans. While we have genetic instructions on the amount of food energy to be consumed (about 7–10 MJ per day for an adult), our exosomatic use of energy depends on culture and income, and reaches 1 GJ or more per day for rich people.

Much later, four well-known economists, who did not yet form a school, are seen in retrospect as ecological economists: Nicholas Georgescu-Roegen (1906–94), the author of *The Entropy Law and the Economic Process* (1971) (where Lotka was often quoted), Kenneth Boulding (1910–93), who worked mainly on general systems analysis, K.W. Kapp (1910–76) and S. von Ciriacy-Wantrup (1906–80), who were both institutionalist economists. The systems ecologist H.T. Odum (1924–2002) studied the use of energy in the economy and some of his former students were among the first ecological economists in the 1980s. Other sources of ecological economics are in Environmental and Resource Economics (microeconomics applied to environmental pollution and the depletion of natural resources), in Human Ecology, Ecological Anthropology, Agroecology and Urban Ecology, and in the study of 'industrial metabolism' as developed by Robert Ayres (born in 1932), now known as Industrial Ecology.

The first books or special issues of journals with the title 'Ecological Economics' appeared in 1987. After meetings in Stockholm and Barcelona with a large group of ecologists and economists, Robert Costanza and Herman Daly set up the International Society for Ecological Economics (ISEE) and convened the first world conference in Washington, DC, in 1990. The book that came out of this first conference (Costanza, 1991) ambitiously defined the field as 'the science and management of sustainability'. The successful academic journal *Ecological Economics*, started in 1989, was edited first by Robert Costanza, and later by Cutler Cleveland (who also edits the Encyclopedia of the Earth) and Richard Howarth.

Some environmental economists of neoclassical persuasion were also present in those early efforts. David Pearce became one of the main editors of the new journal, *Ecological Economics*, but he left the journal after 1994 because of internal disputes on what is now called 'weak sustainability vs strong sustainability'. Pearce had influenced the World Bank to do macroeconomic accounts in which 'human-made capital' could in theory substitute for so-called natural capital. This was 'weak sustainability', i.e., sustainability in the weak sense of the

term. Herman Daly, Peter Victor, and other ecological economists objected to 'weak sustainability'. One could not sustainably substitute the increased horsepower of the fishing fleet for a declining availability of fish. Moreover, measurement of stocks of capital depended on the rate of profit (as discussed in the controversies on capital theory of the 1970s) (Victor, 1991).

However, the ISEE is a scientific society encouraging internal controversy and also a product of the environmental movement of the 1960s and 1970s in its different varieties. Internal pluralism and perhaps some confusion of ideas were present in the volume produced after a second ISEE conference in Stockholm in 1992, *Investing in Natural Capital: the Ecological Economics Approach to Sustainability*. While some ecological economists feel even today that the metaphor of 'natural capital' is useful, others strongly reject it because it suggests that we can use 'units of capital' as a common measuring rod that would make commensurable the losses of biodiversity and the increases in manufactured capital stocks. The notion of 'natural capital' supports policies such as 'habitat trading' (one habitat is destroyed and another one, far away, is effectively protected) or the Rio Tinto mining company's doctrine of 'net positive impact' (a new location is destroyed while another is preserved and enhanced, going round the world until no natural spaces would be left).

While H.T. Odum (and his disciples working on human ecological energetics: A.M. Jansson, Robert Costanza, Charles Hall, Cutler Cleveland), David Pimentel with Mario Giampietro (agricultural energetics), and C.S. Holling ('resilience') were the ecologist grandfathers, mothers, and fathers of ecological economics. K. Boulding and N. Georgescu-Roegen were the economist grandfathers. (K.W. Kapp had died early, missing the birth of ecological economics.) Herman Daly's influence was also decisive. Daly (born in 1938) published his first article in what we now call ecological economics in 1968 in *The Journal of Political Ecology*. In the early 1990s (Daly and Cobb, 1989), he promoted an index of sustainable economic welfare (ISEW) expressed in money terms that showed results very different from GDP (because of the different assumptions in the calculations). It seemed at the time that ISEW was a good way to attack GDP accounting but an index in money terms was not congruent with the critique against 'weak sustainability'. Daly maintained his stature in the field and recently increased it as his early defense of a steady-state economy (Daly, 1973) is now seen as having announced the new ecological macroeconomics without growth (Victor, 2008; Jackson, 2009).

Daly was explicitly inspired in his work by Boulding and Georgescu Roegen. Boulding had published a famous article on *The Economics of the Coming Spaceship Earth* (a spaceship where materials would have to be recycled) (Boulding, 1966). He became a card-carrying member of the ISEE and an author in the collective book edited by Costanza (1991). Georgescu-Roegen refused to be drawn into the ISEE; he preferred to call the field 'Bioeconomics' (Mayumi, 2001; Bonaiuti, 2011) and announced the publication of a book with this title that never appeared. He disliked Costanza's article in *Science* in 1980 (Costanza, 1980) proposing an energy theory of value, he also disliked H.T. Odum's 'Emergy' (embodied energy) accounts to which he answered with an irritated reply: 'matter matters too' (Georgescu-Roegen, 1977) against what he called the 'energetic dogma'. On the other hand, Boulding, more concerned with scarcity of materials than with energy dissipation, wrote a less than enthusiastic review of Georgescu-Roegen's magnum opus. In due course the edition of the collected works and correspondence of such major intellectual figures and pioneers of ecological economics will clarify the real substance, if any, in such disagreements and quarrels.

Georgescu-Roegen's point on the importance of entropy for the economy is as follows. Life is 'negentropic': Georgescu cited very often Schrödinger's *What Is Life* (1944). The evolution of species, the complexity of living structures, was achieved by 'capturing' energy through photosynthesis, and by dissipating energy to outside systems. The industrial economy, however, after the thermo-industrial revolution (Grinevald, 1976) did not work only by using current photosynthesis or hydraulic energy. It was burning stocks of fossil fuels. Even a nongrowing industrial economy would not be sustainable because energy cannot be used twice (except in minor cases of 'cogeneration'). In any day in 2014 we take 90 million of barrels of oil from the 'subterranean forest' (Sieferle, 2001), tomorrow we do the same again, a little more or a little less, whether from the bottom of the sea or from fields in Iraq or Saudi Arabia, the rainforest of Ecuador or the Orinoco Delta in Venezuela. Perhaps the energy return on energy input (EROI) is declining (the energy return on the energy input) or perhaps not yet, Georgescu was also very aware of the approaching Hubbert's 'peak oil'.

Although we could claim that Darwin won against Sadi Carnot (Prigogine and Stengers, 1984), in the industrial economy the sources of low entropy become scarcer. Moreover, materials cannot (in practice) be recycled to the full extent (an observation that Georgescu tried to glorify into a fourth law of thermodynamics, without success, Mayumi, 2001). Georgescu saw the economy as a system open not only to the entry of energy and materials but also to the unavoidable exit of inconvenient and unrecyclable 'garbojunk' (a word formed by garbage + junk).

Georgescu's fundamental contribution to ecological economics was then that, because of the second law of thermodynamics or entropy law, even a nongrowing industrial economy is not sustainable. Therefore, in the rich economies, a 'steady state' (as proposed by Daly, drawing on Stuart Mill), would not be enough. A steady-state economy aims for mildly fluctuating levels in population and consumption of energy and materials. Birth rates equal death rates, and (in economic terms) saving/investment equals depreciation. Georgescu said that in rich countries a degrowth in the inputs of fossil fuels and other materials was required. Hence Georgescu's agreement to the French title to a selection of his articles edited by Grinevald in 1979, *Demain la Décroissance*. In retrospect, this book became one main inspiration for the European 'degrowth' movement after 25 years (Martinez-Alier et al., 2010). Nobody ever preached a 100% degrowth of the economy. Georgescu's lower limit would be that of an economy fueled by the current inflow of solar power. There is therefore a confluence of ideas between Georgescu's degrowth, Daly's steady state (Kerschner, 2010) and the new ecological macroeconomics without growth.

Apart from the United States and Europe, the Japanese 'entropy school' of economic analysis (Tamanoi et al., 1984) studied the environmental services provided by the water cycle, and also the ancient urban ecosystems of Japan. In India, there was much work since the 1970s not only by economists but also by ecologists (Madhav Gadgil) on the links between forest or water management and common property rights, nowadays one main focus of interest in ecological economics (Berkes and Folke, 1998; Agarwal, 2010). Other early ecological economists (whose major works were not in English) are, in France, Rene Passet (1979) and Ignacy Sachs who proposed in the early 1970s the notion of 'eco-development'; Roefie Hueting (1980) in the Netherlands and Christian Leipert in Germany; and Jose-Manuel Naredo in Spain (Naredo, 1987). (For general introductions to the field: Costanza et al., 1997b; Cleveland et al., 2001; Martinez-Alier and Røpke, 2008; Spash, 2009).

A New Ecological Macroeconomics without Growth

Keynesian macroeconomics' initial concern was the 'short term'. One prescription for times of economic crisis was increased 'deficit spending' (increasing the public debt) to finance public expenditure thereby increasing employment, to compensate for the decline in private investment that caused the economic crisis. However, such public investment would by itself increase the productive capacity of the economy; therefore in the long run there was a need to increase again aggregate demand. Keynesianism became a doctrine of continuous economic growth through the Harrod-Domar models of the 1950s. In 2008–09, a new ecological macroeconomics without growth made a breakthrough in public opinion, due to two ecological economists of long standing, Peter Victor and Tim Jackson. A Keynesian-inspired macroeconomics without economic growth is a novelty that ecological economists are offering the world. Ecological economists had previously tried to engage with Keynesian and post-Keynesian economics (Daly and Farley, 2010; Holt et al., 2009) and there was also the tradition of a steady-state economics.

Peter Victor gives several reasons why a country like Canada should no longer have economic growth. Economic growth as measured by changes in GDP is inextricably related to increases in materials and energy despite gains in efficiency over time. The historical record shows a positive link between economic growth and increasing flows of materials and energy (measured as work done, Ayres and Warr, 2009). In fact, gains in efficiency sometimes contribute to increased material and energy throughput because of the operation of Jevons' Paradox, also known as the 'rebound effect'.

How could we manage without growth, at least in advanced economies? These economies are very wealthy as measured by consumptions standards but happiness or life satisfaction has not increased pari passu with income and consumption. The so-called Easterlin's Paradox states that life satisfaction does not grow above a certain level of income, perhaps US $15 000 per capita per year.

Through simulation models Victor then shows that a nongrowing economy in Canada could maintain life satisfaction, achieve nearly full employment, eliminate poverty, and significantly reduce greenhouse gas (GHG) emissions. Simultaneously, the weight of the public debt relative to GDP would decrease. Perhaps more than the results, which depend as usual on the assumptions of the models, the striking novelty is Victor's question itself, "is managing without growth possible?" (2008).

Meanwhile, Tim Jackson wrote a report for the British government in 2008 with the title: *Prosperity without Growth?* (Jackson, 2009). He argued that consumer debt had increased far too much before the economic crisis of 2008 while the public debt would be increasing enormously after 2008 in many countries. Debts can be settled by inflation (money loses value), or by impoverishing the debtors (what some Greeks call 'debtocracy', creditors' rule), or by economic growth. Here Jackson recalls again not only the well known critiques of GDP but also the research on Easterlin's Paradox.

Jackson asked repeatedly for 'financial prudence', not only because of the risk of default but even more because debt repayment requires economic growth, and growth in rich countries is to be avoided for ecological and social reasons. The ecological reasons are clear. If world trends of 2007 would continue, then in order to keep carbon dioxide concentration below 450 ppm, the carbon intensity of the rich economies would have to decrease over 100 times by 2050. This seems impossible. The greenhouse effect was emphasized by Jackson more than 'peak oil' and other planetary boundaries such as biodiversity loss. It sufficed him to make the point.

The development of human potential should not and cannot be achieved by increased consumption of 'positional goods'. Therefore, stopping economic growth is sound from the ecological and economic viewpoints and also socially. Lack of growth requires a revaluation of the goods and services we can get from the commons without damaging the welfare of other people. Jackson gives then policy recommendations based on this ecological macroeconomics. If there is no economic growth, unemployment will increase, given the trend to increased labor productivity because of technical change and the drive for profits. While André Gorz shockingly said *adieu au prolétariat* already in 1980 (not in China or India but in Europe), Jackson is concerned with 'the stigma of unemployment'. Hence, support must be given to the so-called Cinderella sector (which, remembering William Morris, one could call the 'News from Nowhere' sector). This sounds naïf but it is well thought out. One virtue of this sector is precisely low labor productivity and the capacity to give massive useful employment. Jackson's other policies are a basic universal citizen's income and work sharing, and also investments in renewable energy and in ecosystem restoration and enhancement. Since such investments are unlikely to give a high rate of return in money terms, the share of savings (or taxes) and the role of the public sector of the economy will have to grow. Does this mean a retrenchment of capitalism? Do not get excited about words, is Jackson's motto. But the question remains whether a nongrowing capitalist system is possible at all.

Financial Debts Compared to Real Wealth

The amount of real wealth that an economy can create is limited by the amount of low-entropy energy and materials

that it can sustainably take from the external environment, and by the amount of effluents such as the GHGs that the environment can sustainably absorb. Jackson's 'financial prudence' is not a new doctrine in ecological economics, on the contrary. The Nobel Prize in Chemistry and expert on radioactivity Frederick Soddy wrote on energy and the economy from 1910 onward. He compared 'real wealth' which grows at the rhythms of nature and which, if turned into manufactured capital, is worn down, with 'virtual wealth' in the form of debts which apparently could grow forever. Private property in a capitalist system guarantees (for a while) the increasing private debt while the public debt could apparently grow based on the guarantees provided by the State. But this was a flimsy building.

Soddy was quoted by Daly and other ecological economists since the 1980s, much before the financial crisis of 2008. Debt-fueled growth was not viable. The real fuel of economic growth was coal, oil, gas. Yes, life is negentropic but even eucalyptus cannot grow for more than 30 years at 10% per year. Then they stop growing. Soddy's book of 1926 was called *Debt, Wealth and Virtual Wealth*. He drew on John Ruskin. He meant that debt was not real wealth, it was virtual wealth. Real wealth was the current inputs of solar energy. Although ecological economists have not developed a consensual monetary reform plan, they follow Soddy on the need for 'financial prudence' recommending for instance a large increase in the cash reserve requirements of banks (Daly and Cobb, 1989).

Soddy approvingly quoted Aristotle's distinction between Oikonomia and Chrematistika, as Marx had done and Karl Polanyi (1957) was to do later. Ecological economists are fond of this distinction. Oikonomia meant the study of the material provisioning of the oikos (the extended family) while chrematistics was the art of studying market prices to make money, for instance by becoming a monopolist (a word used by Aristotle). What Aristotle called oikonomia would now be called human ecology and economic anthropology, while chrematistics is what students of microeconomics learn.

Irreducibility of Needs

In economic theories of production and consumption, compensation and substitution reign supreme. Not so in ecological economics, where diverse standards of value are deployed 'to take nature into account' (O'Connor and Spash, 1999). In the ecological economics theory of consumption, some goods are more important and cannot be substituted by other goods (economists call this a 'lexicographic' order of preferences). Thus, sacredness cannot be traded off. And no other good can substitute or compensate for the minimum amount of endosomatic energy or for water necessary for human life. To call either the endosomatic consumption or the exosomatic use of energy a 'socially constructed need or want' would leave aside the ecological explanations and/or implications of such use of energy, while to call the daily endosomatic consumption a revealed preference would betray the conventional economist's metaphysical viewpoint.

There is another approach which, as pointed out by John Gowdy and Susan Mesner (1998), builds upon the 'principle of irreducibility of needs' (proclaimed by Georgescu-Roegen in the 1968 edition of the Encyclopedia of the Social Sciences, article on 'Utility'). According to Max-Neef (Max-Neef, 1992) all humans have the same needs, described as 'subsistence', 'affection', 'protection', 'understanding', 'participation', 'leisure', 'creation', 'identity', 'freedom', and so on and there is no generalized principle of substitution among them. Such needs can be satisfied by a variety of 'satisfactors'. Instead of taking the economic production as given, we may ask (as in the steady state and *Décroissance* perspectives) why is there so much travel, why so much building of houses with new materials instead of restoration of old ones… Is there a trend to use 'satisfactors' increasingly intensive in energy and materials in order to satisfy predominantly nonmaterial needs?

Against Commensuration

One of the foundations of ecological economics is incommensurability of values although ecological economists understand and have pushed sometimes for the economic valuation of ecosystem services, with the avowed intention of making them more visible to the general public and to policy makers who are assumed to think mainly in money terms. Nevertheless, the insistence on money valuation clearly makes less visible the biological and ecological importance of nature, and also livelihood and cultural values. The beauty and sacredness of a mountain such as the Niyamgiri Hill in Odisha might seem negligible when compared to the very large money value of its bauxite deposits. The mountain is better defended outside money valuation (Temper and Martinez-Alier, 2013). The debates on when is money valuation appropriate continue in ecological economics (Kumar, 2010). A consensus is perhaps being reached that money valuation is appropriate when trying to make companies accountable in civil litigation for their past environmental liabilities (as in the British Petroleum case in the Gulf of Mexico, Chevron-Texaco in Ecuador, or Shell in the Niger Delta) but it is not appropriate for taking decisions for the future (whether on climate change or biodiversity policies or on building an open cast mine or a dam), when money valuation becomes only one of several relevant valuation languages.

In evaluating investment projects, ecological economics favors participatory multicriteria evaluation (Munda, 2008) and deliberative methods (Zografos and Howarth, 2008) over cost–benefit analysis (CBA). This is not only because of 'technical' problems with CBA (which discount rate to use, which methods of money valuation to apply to nonmarket goods and services or to negative 'externalities') but mainly because of a political question: Who has the power or the right to simplify complexity by imposing a single valuation language? Most ecological economists (but not all) would side with Spash (2007, 2011) in his vigorous critiques against the CBA approach of the Stern review on climate change and against the UNEP TEEB project (*The Economics of Ecosystems and Biodiversity*), and also in his call for an ecological economics that definitively breaks away from neoclassical environmental and resource economics (Spash, 2012).

Debates on fictitious money valuation and weak sustainability have accompanied ecological economics from its

inception. In the early years, the identity of this field that considers the economy as a subsystem of the environment, and not as the simple application of conventional economics to environmental issues (as in 'transport economics' or 'agricultural economics'), was preserved through episodes such as David Pearce's withdrawal as one main editor of the journal and Daly's resignation from the board of the Beijer Institute in Stockholm that had given itself the name of 'ecological economics' while promoting mainstream economists.

In the 1990s, there remained however in the Beijer Institute of Ecological Economics, as a junior researcher, one of A.M. Jansson's students, the ecologist Carl Folke (H.T. Odum's intellectual 'grandson' so to speak). In due course the Resilience Alliance was born in Stockholm with C.S. Holling's support. The notion of resilience (the capacity of a system to withstand or recuperate from disruption) came to dominate in ecology and in ecological economics over the simpler notion of carrying capacity.

Ecological economics is not committed to a unique type of value expressed in a single numeraire or unit of account. "The issue is not whether it is only the market place that can determine value, for economists have long debated other means of valuation; our concern is with the assumption that in any dialogue, all valuations or 'numeraires' should be reducible to a single one-dimension standard" (Funtowicz and Ravetz, 1994: p. 198). Ecological economics encompasses money-valuation, and also physical appraisals of the environmental impacts of the human economy measured in their own physical numeraires. It also gives importance to social indicators.

Nature provides resources for the production of commodities and it also provides environmental amenities. As shown by Gretchen Daily (1997), De Groot (1992) and other authors, nature, more importantly, gives gratis essential life-support services such as the cycling of nutrients, the water cycle, soil formation, climate regulation, conservation and evolution of biodiversity, concentration of minerals, dispersal or assimilation of pollutants, and diverse forms of useful energy. Controversial attempts have been made to assign money values to the annual flows of environmental services, to compare them to GDP in monetary units of account (Costanza et al., 1997). For instance, the cycling of nutrients (nitrogen, phosphorous) in some natural systems may be given a plausible money value by comparison with the costs of alternative human-made technologies. Could this same methodology (i.e., the cost of alternative technology) be applied consistently to the valuation of biodiversity in a kind of science fiction Jurassic Park framework? For biodiversity, money valuation has taken a completely different track, namely the small sums exchanged in some 'bioprospecting' contracts, or fictitious subjective money values in terms of 'willingness to pay' for conservation projects, i.e., the so-called contingent valuation method.

For instance, how to count the service that nature provided us by concentrating minerals which we disperse? ('Exergy' costs have been calculated by industrial ecologists, but the technology for creating mineral deposits does not exist). Therefore, the figures obtained for the money values of environmental services provided free by nature are incongruous. They are useful, however, in stimulating the debate on how 'to take nature into account'.

Ecological economics rests on a foundation of 'weak comparability of values' (O'Neill, 1993; Martinez-Alier et al., 1998). One simple example: let us assume that a new large garbage dump must be built near a city, and that there are three possible locations, A, B, and C, one of which will be sacrificed. In our example, the three different locations are compared under only three different types of value: value as habitat, value as landscape, and economic (monetary) value. Location A is a most valuable publicly owned wetland (valuable as habitat or ecosystem because of its richness of species) but a monotonous landscape, much visited by bird-watchers (and, as such, of some economic value according to the 'travel cost method'). Location B is an old agricultural area of beautiful derelict orchards and ancient manor houses, which ranks first as landscape, but only third as rent-producing, and second as ecosystem or habitat. Location C produces much rent as industrial and urban land, and therefore ranks first in economic value, but ranks only third as an ecosystem or habitat, and comes second as landscape (because of its historical qualities). Which location should be sacrificed? Should and could all values be reduced to a supervalue, so as to achieve strong comparability, and even strong commensurability (cardinal measurement)? In the example, the economic values (in actual or fictitious markets) of all three locations have been taken into account, but there is no supervalue (economic or otherwise such as for instance net energy production by which the wetland would presumably come out on top). Certainly, the present rankings could be reconsidered. Thus, the landscape value of A could be upgraded, and its economic value (as also that of B) could also be increased by contingent valuation. Giving more weight to some criteria than to others, or 'veto thresholds' for some criteria such as the 'endangered species' provision in American legislation, or the introduction of more locations or more criteria, would help us to escape from the deadlock. The point is to show the meaning of 'weak comparability of values'. The decision-making process need not be irrational (by lottery, for instance).

In contrast to such multicriteria approach, in CBA the projects to be evaluated are all valued in the same numeraire (present value in money terms of costs and benefits, including monetarized externalities and environmental amenities).

In microeconomics, there is strong comparability of values, and indeed commensurability, when externalities are internalized into the price system, as in the definition of a Pigouvian tax as the economic value of the externality at optimum pollution level. In macroeconomics, El Serafy's early and influential proposal to 'green' the GDP (in Costanza, 1991) – the results of which will depend on the chosen rate of interest – implied commensurability in money terms. According to El Serafy, not all receipts from the sale of exhaustible resources ('natural capital') should be included in GDP, only one part should be included, 'true' income, and the rest should be counted as 'decapitalization' or the 'user cost' of such 'natural capital', which should be invested at compound interest over the period until the resource is exhausted (as in the Norwegian Oil Fund), so as to allow the country to live at the same standard of living even when running out of the resources. How such a fund can continue to give income if the world runs out of oil is something of a mystery. The proposal, related to Hotelling's rule in resource

microeconomics, is aimed at correcting the macroeconomic accounts. It is based on a notion of 'weak' sustainability only that allows the equivalent substitution of financial or manufactured capital for so-called natural capital–implying therefore a common unit of measurement, i.e., money value – while 'strong' sustainability refers to the maintenance of physical natural resources and services.

The so-called Environmental Kuznets Curve, an inverted U-curve, relates income and some environmental impacts. It has been much debated in ecological economics. For instance, in urban situations, as incomes grow sulfur dioxide emissions first increase and then decrease while carbon dioxide emissions and domestic waste increase with incomes. If something improves and something deteriorates, one reaction from the conventional economist might be to put weights or prices on such effects, in the pursuit of commensurability. However, the uncertainty and complexity of such situations (sulfur dioxide may counteract the greenhouse effect on temperature), and the fact that the price of externalities would depend on the outcome of ecological distribution conflicts, implies that the economists' accounts would be convincing only for the believers of the same school.

When the pattern of use of environmental resources and services and the burdens of pollution are shown to depend on changing social structures and on power and income distribution, then we enter the field of political ecology that studies ecological distribution conflicts. Economic growth leads to increased environmental impacts and conflicts (often outside the market sphere). The response is the growth of a global environmental justice movement (Guha and Martinez-Alier, 1997; Martinez-Alier, 2002). Examples abound of the failure of the price system to indicate such environmental impacts, or (in K.W. Kapp's terms) examples abound of cost-shifting successes. Thus the pattern of prices in the economy would be different without the free access to carbon sinks. Should restrictions be imposed on the 'ecological footprint' of rich economies or on the Human Appropriation of Net Primary Production (HANPP, see below)? Free use of sinks was modeled in a neo-Ricardian or Sraffian framework by Perrings (1987) and O'Connor (1993) showing how the pattern of prices in the economy would be different assuming different outcomes for ecological distribution conflicts.

Related to economic valuation, some remarks are still needed on the discount rate. Economists explain discounting of the future by subjective 'time-preference', or because economic growth per capita caused by today's investments, will make the marginal utility of consumption lower for our descendants than it is for us today. We discount the future because we assume that the future will be more prosperous, and therefore we destroy more exhaustible resources today and we pollute more than otherwise would be the case, thereby undermining future prosperity. This has been called the 'optimist paradox'.

Even accepting the (Ramsey) argument that discounting arises from the productivity of capital, but taking also into account that such 'productivity' is a mixture of true increases in production and of environmental destruction, then the discount factor should be the per capita rate of sustainable economic growth, subtracting therefore the destruction of environmental resources and services and the negative effects of pollution. Now, in order to determine the present economic value of such future effects caused by economic growth (loss of biodiversity, filling up of carbon sinks, production of radioactive waste, etc.), we need not only to put money figures on them (as discussed above), we need also a discount rate. Which?

The current debate on valuation in an intergenerational context goes back to the 'socialist calculation debate' started by Otto Neurath and Ludwig von Mises in Vienna in the early 1920s. Otto Neurath (1882–1945) favored a *Naturalrechnung*, accounting in physical terms, while Von Mises wrote that without prices there could not be a rational economy. Max Weber agreed with Von Mises. Otto Neurath disagreed, and asked how should we decide whether to use more coal now and less human labor, or keep coal for the future and use now more human labor. In today's terms, should we use more fossil fuels now, enjoy economic growth and produce more GHG, rely on technical change and invest in new renewable technologies and geoengineering, or should we go into a steady-state economy after a period of slight degrowth in the rich economies? Collectively, these are technical–ethical questions, they are not decisions that real or fictitious market prices can solve. We cannot enter into market-like negotiations with future generation of humans (or with other species); the methodological individualism of orthodox economic theory breaks down here. Nevertheless, instead of engaging with such arguments, Hayek (1952) in *The Counter-Revolution of Science*, pursuing his 30-year-old disagreement with Neurath, lumped him and other authors (Soddy, Geddes, Mumford) who supported physical accounting into the category of 'social engineers', would-be dictators.

Property Rights and Resource Management

Many years later, one article by Vatn and Bromley (1994) titled *Choices without prices without apologies* explained why money valuation is only optional. Choices depend on socially molded preferences that depend on institutions, that is, the social rules and norms. As Veblen famously put it, the individual consumption of the rich is guided by the social rule of showing off. Bromley and Vatn are institutionalist economists, in Veblen's and Kapp's tradition (while Kapp himself was influenced by Otto Neurath's economics). They see economic behavior not as being determined by inscrutable individual preferences but as influenced and explained by social rules and norms. Institutions articulate a diversity of values (Vatn, 2005).

In this context, Coase's approach to the internalization of negative externalities or positive environmental services into the price system relies on market transactions between partners. So, if an agent (a peasant community, a factory) pollutes the water in a river, the downstream aggrieved agents may get together either to ask for an indemnity equivalent to the damage suffered or to bribe the polluters to stop the pollution, depending on the property rights on the river. This might work without need for government intervention. However, getting the downstream people together to start a court case implies

'transaction costs' (lawyers' fees, time for meetings) that prevent the simple market solution from operating, as Coase himself emphasized. Moreover, when those being polluted are future generations and other species, the market solution does not operate. Regulation (physical norms and fines) or Pigouvian taxation are preferable.

Regarding the empirical study of a very popular policy instrument such as Payment for Environmental Services (PES), one main contribution from ecological economists has been to criticize the simple Coasean, market approach, and introduce complexities related to uncertainty, distributional issues, social embeddedness, and power relations, acknowledging the variety of contexts and institutional settings in which PES operate (Muradian et al., 2010).

The institutionalist perspective (economic actions are explained by social rules more than by inscrutable individual preferences) is very relevant for the study of the relation between property rights or property regimes and the management of natural resources. We go back here to the much quoted article by Garrett Hardin (1968), *The Tragedy of the Commons*, which should have been titled 'the tragedy of open access'. Hardin attended the inaugural conference of the ISEE in Washington, DC, and he had a chapter in the seminal book edited by Costanza (1991). One paragraph in Hardin's article starts like this: "Picture a pasture open to all…", and describes how a tragedy of overgrazing will occur because individuals would put more and more sheep or cows on the pasture provided that the marginal benefit (a few liters of milk, a few pounds of wool, some sacks full of dung) would be larger, individually, than the private marginal cost, disregarding the collective marginal costs in terms of soil degradation.

To the old liberal critique of 'common' property ('the magic of private property (and enclosures) would turn sand into gold' had written Arthur Young) was now added a trendy environmental critique, the Tragedy of the Commons. However, was Hardin not aware of the rules that in the past (brought from England to New England) presumably regulated the amount of horses or cows that a citizen could put to pasture in the Boston Commons? Or in any other commons in the Americas, whether a Mexican ejido or a New Mexican common pasture? Why was the confusion between 'commons' and 'open access' not spotted by the reviewers of Hardin's article in *Science*? There were many well-regulated communal systems of management for forests, coastal fisheries, or irrigation water, as Bromley and others soon retorted and as Elinor Ostrom was to study in detail (Ostrom, 1990).

Hardin's 1968 article mistook commons for open access. It preached privatization (or state property) against the misnamed 'commons'. It took some time until the confusion was cleared up, provoking much research on the functioning of common property regimes, and also on the relations between forms of property and resource management. There are certainly open-access resources, for instance some fisheries in the open seas. The atmosphere was also treated as being in open access to dump polluting substances such as chlorofluorocarbons that damage the ozone layer until an international treaty banned this practice. The atmosphere is still a dumping ground in open access as regards GHG. Other examples abound. But apart from open access (which should be avoided), there were other forms of property.

- The true commons
- Private property
- State and municipal property

Some scholars did research on the trend to turn natural resources held in common and subject to traditional management rules (like coastal mangroves and fisheries in India or Latin America) into *de facto* private property, e.g., for growing shrimp or for fishing for export. This was described not only as a social but also an environmental 'tragedy of enclosures'.

Compared to open access, private property is in principle more conducive to conservation because the costs of today's actions will be felt by the owner or his immediate kin. This was discussed in resource economics at least since Faustmann's rule (1849). The private owner of a forest (or rather, of a tree plantation) will decide to cut the trees not as soon as possible, but when the rate of growth of the trees (net of harvesting costs and multiplied by the market price of wood) falls below the rate of interest in the bank plus the rent to be obtained from the land, now empty off trees (potentially used for crops or pastures while the new stand of plantation trees is starting to grow again). Notice here that a high interest rate (or discount rate) will lead to cutting the trees very soon, while payment for ecosystems, products, or services (like hunting rights, mushroom collection, recreation under the trees, carbon capture) would slow down the rotation period and could even persuade the owner not to cut the trees at all and eventually turn back the plantation into a true forest.

In the case not of trees in a plantation but of extraction of mineral ores or fossil fuels, private property linked to the profit motive is certainly not conducive to conservation, nor to the avoidance of negative environmental impacts after the exhaustion of the resource such as acid drainage from mines. An oligopoly such as Organization of Petroleum Exporting Countries slows down oil extraction because of economic and not environmental objectives. New institutions, that is, rules articulating new values (for instance, civil or criminal legislation on socioenvironmental liabilities) would perhaps modify the behavior of the extractive industries.

Social Metabolism and Indexes of (Un)Sustainability

Because of the shortcomings of money valuation, ecological economists favor physical indicators and indices in order to judge the overall impact of the human economy on the environment. Therefore, we leave here aside monetary corrections to GNP, such as El Serafy's (see above), or Hueting's (in Costanza, 1991), which calculated the economic costs of adjusting the economy to socially negotiated norms or standards of pollution and resource extraction, in a 'cost-effectiveness' approach (by 'cost effectiveness', it is meant the analysis of the cheapest way, in money terms, in order to adjust the economy to such physical norms or standards). We also leave aside the ISEW (see above) whose end result is a figure in money terms commensurate with GDP though often showing quite a different trend.

According to Georgescu-Roegen (1971), economics should see the economy as an open system (and not as a self-sustaining system, a 'merry-go-round' between consumer and

producers as in the textbooks). Economics should study the 'metabolic flows' in the economy. This is today linked to two research schools. The first one centers on Marina Fischer-Kowalski's collaborators at the Institute of Social Ecology in Vienna, drawing on work by Robert Ayres, R.P. Sieferle, and other authors. This group has developed methods for the study of the social metabolism, listed below.

The second school would be Marxist ecological economics. It has much less influence. It claims with reason that Marx already wrote in the 1860s that the capitalist economy was causing a 'metabolic rift' (Foster, 1999) between humans and nature. Marx took the work metabolism (*Stoffwechsel*) from Moleschott and Liebig, pointing to the export of nutrients in the soil by commercial agriculture. Capitalism not only exploited workers but also exploited the soil. The soil was no longer a 'fund' able to supply crops continuously; it became an exhaustible stock in terms of its fertility and texture. Marx quoted Liebig who feared the day when guano imports would diminish. Marx, as Liebig, hoped for factory-made chemical fertilizers (in a sort of 'weak sustainability' approach) to escape the Malthusian trap of 'diminishing returns'. Despite such intellectual traces, a Marxist ecological economics or environmental history has not existed until the end of the twentieth century from Altvater (2007), Bellamy Foster, Hornborg (in his theory of ecologically unequal trade, Hornborg, 1998), and O'Connor's 'second contradiction' (O'Connor, 1988).

Going back to the first school, here is a summary of today's methods for the study of social metabolism trying to ascertain the trends in the use of materials and energy in the economy. They allow concrete answers to questions such as, "is the economy dematerializing in absolute terms, or only relatively to GDP growth", "has the human appropriation of biomass increased as fast as population in the last 100 years", and "is the energy cost of obtaining energy increasing?"

- Calculations of material flows (in tons), often divided into biomass, building materials, mineral ores, fossil fuels, at national or regional levels. Therefore, one can test the hypothesis of 'dematerialization' of production or consumption, and one can do calculations of physical trade balance supporting the arguments on ecologically unequal trade (Pérez-Rincón, 2006; Weisz et al., 2006; Russi et al., 2008).
- Calculation of energy flows at national or regional levels including historically changing ratios of exosomatic energy use/endosomatic energy use (Giampietro et al., 2013), calculations of the EROI (Hall et al., 1986). EROI calculations show whether there is a trend toward an increasing energy-cost of obtaining energy. Some anthropologists analyzed the basic economics of human society as a flow of energy (e.g., Roy Rappaport's *Pigs for the Ancestors*, 1968). Engels in 1882 discussed this topic with Marx sadly denying the relevance of Podolinsky's agricultural energetics for Marxian economics. Max Weber in 1909 had criticized Wilhelm Ostwald's interpretation of economic history in terms of (1) an increased use of energy, (2) an increased efficiency in the use of energy, because economic decisions on new industrial processes or new products were based on prices, entrepreneurs did not pay attention to energy accounts per se. Max Weber did not yet question energy prices from the environmental viewpoint as we would today. In the 1970s David Pimentel (Pimentel et al., 1973) showed the decreasing energy efficiency in maize cultivation in the United Sates. A new field was opened up by such studies on the efficiency in the use of energy in different sectors of the economy, including the energy sector itself.
- Calculations of 'virtual water' in import or in exports (e.g., wheat, soybeans, meat, ethanol, cellulose) at regional or international levels, i.e., the water 'dissipated' in growing crops, pastures, timber. It was argued by Tony Allan in the 1990s that water disputes in the Middle East were mitigated by importing food. Hoekstra and Chapagain (2007) have calculated water footprints of nations, and there is research on the unpaid water transfers embodied in trade in commodities.
- Calculations of HANPP (including the 'embodied HANPP', which is relevant for trade statistics) (Vitousek et al., 1986; Krausmann et al., 2013). The driving idea is that the larger the amount of potential biomass production appropriated by humans in a given territory, the less biomass is available for other (nondomestic) species, and therefore the more likely the loss of biodiversity. The proportion of NPP appropriated by humans is increasing because of increasing demands of land for urbanization and infrastructures, for growing feedstuffs and agrofuels, for growing timber.

All the indexes mentioned above are measured in different units and might show different trends, although there are also some rules of congruence among them. There are no cases of rapid GDP growth and persistent degrowth in material flows (particularly when trade effects are taken into account). But it might happen that increasing fossil fuel inputs into agriculture will increase yields and slow down the increase in the HANPP. How should a situation be judged in which, for instance, an index such as material flows slightly improves while HANPP deteriorates, EROI decreases, GDP grows, infant mortality decreases, and self-reported life satisfaction is stable? Notice that the human development index regularly published by United Nations Development Program adds up economic and social indicators into a single number but it does not include environmental impacts at all. Commensurability would imply reducing such values to an encompassing supervalue but this is not necessary in order to reach reasonable integrated assessments by macroeconomic multicriteria evaluation (Shmelev, 2012).

There is one composite physical index of (un)sustainability that was born within ecological economics and has had great success in public opinion: the Ecological Footprint, first presented by Bill Rees (an ecologist turned urban planner) around 1991. Which is the environmental load of the economy, in terms of space? H.T. Odum had posed the question, and later authors developed some answers. Rather than asking what maximum population a particular region or a country can support sustainably, the question becomes: how large an area of productive land is needed (as source and sink) in order to sustain a given population indefinitely, at its current standard of living and with current technologies? The Ecological Footprint is the inverse of the carrying capacity (see below). Computations, not only for cities or metropolitan regions (whose 'ecological footprint' is hundreds of times larger than

their own territories) but also for countries, show that some densely populated European countries or Japan and Korea (assuming per capita eco-footprints of 2 or 3 ha) occupy eco-spaces 10 times larger than their own territories (Wackernagel and Rees, 1995). The Ecological Footprint is rightly criticized because it adds up actual land use (say, 1 ha per person for food, feed, and timber) and notional or 'virtual' land use (say, another hectare) for the absorption by photosynthesis of the carbon dioxide produced by burning fossil fuels. The results are sometimes expressed in terms of how many real plus virtual planets humankind is using. Already we are using one and a half. But this depends on the amount of land that is left for other species. If the HANPP would be reduced by half for that purpose, then the Ecological Footprint of humans in terms of extra planetary needs would be even more alarming. Despite these and other critiques, the Ecological Footprint has enjoyed great success possibly because a spatial representation of environmental impact is easily understood.

On Population

Ecological economists emphasize both the pressure of population and the pressure of production (and consumption) on resources. How large is humankind's ecological footprint? Has humankind exceeded 'carrying capacity'? This is defined in ecology as the maximum population of a given species, such as frogs in a lake, which can be supported sustainably in a given territory without spoiling its resource base. However, the large differences internal to the human species in the exosomatic use of energy and materials, mean that the first question is, maximum population at which level of consumption? Second, human technologies change at a quick pace. Already Boserup's thesis (1965) of endogenous technical change according to which preindustrial agricultural systems had intensified in response to increases in population density, tried to turn the tables on the Malthusian argument. Third, the territories occupied by humans are not 'given', other species are pushed into corners or into oblivion (as the index HANPP implies), and, internal to the human species, territoriality is politically constructed through state borders. Fourth, international trade (similar to horizontal transport in ecology, but which humans can regulate consciously) may imply 'ecologically unequal exchange', though if one territory lacks a very necessary item which is abundantly present in another territory, Liebig's law of the minimum would recommend exchange. Then, the joint carrying capacity would be larger than the sum of the carrying capacities of all autarchic territories.

Because of the shortcomings of 'carrying capacity' as an index of (un)sustainability for humans, and because of Barry Commoner's arguments against Paul Ehrlich's fixation on population growth (Ehrlich, 1968) allegedly forgetting that overconsumption was the main environmental threat, the formula I = P.A.T was proposed by Ehrlich himself, where I is environmental impact, P is population, A is affluence per capita, and T stands for the environmental effects of technology. Efforts are being made to operationalize I = P.A.T. True, population remains one important variable. True also, the demographic transitions are not mere automatic responses to urbanization and education, and their timing does not depend only on inheritance patterns and family forms. Human demography is anticipatory and self-conscious. Though it also follows Verhulst's curve, it is different from the ecology of a population of frogs in a lake.

There have been three different varieties of Malthusianism. First, Malthus' own view in 1798 that human populations would grow exponentially unless checked by war and pestilence, or by the unlikely restraint of chastity and late marriages. Food would grow less than proportionately to the growth of the labor input, because of decreasing returns. Hence, subsistence crises.

Then there was the Neo-Malthusianism of 1900, with radical activists such as Emma Goldman, Paul Robin (Ronsin, 1980). Human populations could regulate their own growth through contraception. Women's freedom was required for this, and it was desirable for its own sake. This was a feminist Neo-Malthusianism, insisting on what is called today 'reproductive rights'. Abortion and vasectomies should not be criminalized. 'Conscious procreation' was required in order to prevent low wages and pressure on natural resources but the main cause of poverty was social inequality. This was a successful bottom-up movement only in some parts of the world, particularly in Europe and America against States (which wanted more soldiers) and against the Catholic Church.

There is finally top-down Neo-Malthusianism after 1970, reaching extremes like Hardin's 'lifeboat ethics' (Hardin, 1974) against freedom of migration with racist overtones. This top-down doctrine and practice is sponsored by international organizations and some governments. Population growth is seen as one main cause of poverty and environmental degradation. Therefore, states must introduce contraceptive methods, even sometimes without the populations' (particularly women's) prior consent.

Ecological economists have been divided into top-down and bottom-up Neo-Malthusians with lack of dialogue between them although they all refuse the doctrines of the anti-Malthusians, who assume that human population growth is no threat to the natural environment, and that it is conducive to desirable economic growth. Ecological economists believe that growth of world population (four times in the twentieth century) is certainly a very major issue. It now seems that population might be stabilized and even go into a slow decline after 2050, at less than 9 billion people. This is welcome. Fertility is going down in many regions and countries. The demographic transition is being completed. This is a good thing, although local depopulation (not only rural, but also urban) may create new social and environmental problems.

Civil Society and Ecological Economics

There is no doubt about the present influence of the main schools of economics on governments and public opinion (perhaps except for Marxism but certainly including Keynesianism, Schumpeterian economics, neoliberal or market fundamentalism, etc.). But one can doubt still whether ecological economics is having any influence at all. Who and where are the nonacademic 'consumers' of ecological economics?

So far, despite the overarching vision of Georgescu-Roegen and his followers, no politician ever quotes in public debate the entropy law and its relevance for the economic process while many discuss the large or little ups and downs of GDP, and the public debt risk ratings. Detroit's municipal default of 2013 seems more relevant to both neoliberal and Keynesian public opinion leaders in rich countries than the world ecological debt. Nevertheless, the critiques of national income accounting by ecological economists and by ecofeminist economists have certainly influenced the sporadic and rather fruitless movements to move 'beyond GDP' by governments and international bodies. It is obvious that many caring activities in families and in society and many services of nature (Waring, 1988), remain outside the market, and therefore are unaccounted for although they are essential for social reproduction and human well-being. This is commonly accepted. Further, a civil society group in the United States is promoting the steady-state economy. Its motto is: "Growth, especially in wealthy nations, is already causing more problems than it solves." It has over 10 000 members.

Meanwhile, the economic valuation and payment for ecosystem services has become a growth industry in Ministries of the Environment and Forests, worldwide. Also, trends of the decarbonization of the economy and also of its material intensity are mentioned in political debates in some countries. But there is an enormous gulf between the critiques from ecological economics and current government and business practices. Business circles are, like governments, eager consumers of economic doctrines but they ignore ecological economics. At most, the debates on the environmental liabilities from private or public companies, also known as 'the social costs of private enterprise' (Kapp, 1950), have given rise to so far aimless discussion of how to take the environmental damages into the actual accounts beyond 'greenwashing' corporate social responsibility exercises (Sukhdev, 2012). Moreover, some legal avenues to claim environmental liabilities from companies (such as the Alien Tort Claims Act legislation in the United States) are closing down.

Other notions from ecological economics, such as ecologically unequal trade, have been taken up by environmental groups from the South, especially by 'postextractivist' groups in Latin America. For instance, the metabolism of soybean production (loss of soil nutrients and 'virtual' water), and the damage from glyphosate spraying in Monsanto genetically modified crops, is a growing issue in the politics of Argentina (Pengue, 2005).

Via Campesina is an international network of peasant and landless workers' organizations. Drawing on studies of the energetics of the modern food systems, it claims that traditional peasant agriculture cools down the Earth. Vice versa, concepts like the ecological debt or carbon debt have progressed from activist circles to the speeches of ministers or even heads of state from some countries in the South, and then to academic journals (Srinivasan et al., 2008). Similarly, concepts like Degrowth get support from social movements (the *Décroissance* movements in Europe), easily connecting to Jackson's and Victor's academic ecological macroeconomics without growth. The European Ecological Economics Society (a very active branch of the ISEE) supports the activist and academic conferences on Degrowth (not without internal debate), a position very unlikely to be taken and for good reasons by the Indian Ecological Economics Society. The reception of and contributions to ecological economics by civil society organizations have been researched by Healy et al. (2012).

For instance, on a new issue like geoengineering (combating climate change through ocean iron fertilization, sulfate aerosol spraying, etc.), although the ISEE will not take a political position, already a well-known Australian ecological economist, Clive Hamilton (2013), has provided something like the ecological economist's party line (which is similar to that anticipated by Pat Mooney of the civil society organization ETC), showing that geoengineering is fraught with new dangers, that the precautionary principle should be applied, that the debate should be opened up as in the 'postnormal science' approach (Funtowicz and Ravetz, 1994), and that climate change mitigation in the form of reducing emissions is preferable to such doubtful experiments with uncertain consequences. Ecological economists will differ among themselves, however, on the measurement of costs of climate change mitigation, and even on the proper units of account of such costs.

The Society for Ecological Economics

Could the ISEE (born in the late 1980s) be seen in future as remotely similar in influence (though very different in intention) to the Mount Pelerin society founded by Hayek in 1947 to defend the market economy and Karl Popper's 'open society' against Marxist doctrines and Keynesian social-democratic planning? Should the ISEE have a political objective? Is the pluralism in ecological economics undermining such a prospect, or rather, is it the radicalism of ecological economics preventing or delaying its social acceptance? Should we relent a bit and accept 'weak sustainability' and the promises of ecological modernization, or should we denounce UNEP's 'green growth' of 2013 as an oxymoron even more blatant than Brundtland's 'sustainable development' of 1987? In fact, should 'we' (ecological economists) have any collective position at all?

The ISEE has lived up to its promise of promoting a transdisciplinary field at the intersection of ecology and economics. Both economists and ecologists coming from different schools have been active in the field. This article has briefly explained the origins of ecological economics going back to the late nineteenth century. Therefore, it is not true that "ecological economics is simply what ecological economists do". They do many different things but within a common tradition which is not clearly delimited because it is at the interface of related fields.

While the journal *Ecological Economics* and the biennial conferences are the main focus of activities of the ISEE, there are active regional societies in the United States, Europe, India, Latin America, and Russia. The European Society for Ecological Economics edits also another journal, *Environmental Policy and Governance*. The Latin American ecological economists publish the journal *Revibec*. Well-known ISEE members have edited related journals: Charles Perrings, *Environment and Development*; Clive Spash, *Environmental Values*; Robert Costanza, *Solutions*; and Jeroen van den Bergh, *Environmental Innovation and Societal Transitions*. The Indian Society (INSEE) regularly publishes

proceedings of well-attended conferences. Her past-president, Kanchan Chopra, gave her name to a famous committee mandated in 2002 by the Supreme Court, giving the methodology for calculating Net Present Values of nonmarket forest products and services sacrificed by mining or hydroelectric projects.

Among all this variety, there is a common thread. A sample of first university degrees and main scientific interests of the (older) ecological economists can be constructed by listing the names of the ISEE presidents since 1989, as follows.

- Robert Costanza, systems ecology, architecture, energy in the economy, ecosystem services
- R.B. Norgaard, economist, postdevelopment studies, coevolution, biodiversity conservation
- John Proops, physicist, energy in the economy, capital theory, economic–ecological modeling
- Charles Perrings, economist, development studies, economics of biodiversity
- Joan Martinez-Alier, economist, energy in society, environmental history, political ecology
- Peter May, resource and environmental economist, development, Amazon deforestation
- John Gowdy, economist, economic anthropology, economics of consumption
- Bina Agarwal, economist, development and feminist economics, common property, India
- Marina Fischer-Kowalski, sociologist, industrial ecology, social metabolism, transition societies

This can be complemented by the list of recipients of the ISEE Boulding Award: Herman Daly (economist), Robert Goodland (biologist), A.M. Jansson (ecologist), Robert Costanza (ecologist), C.S. Holling (ecologist), Robert U. Ayres (physicist, industrial ecologist), Partha Dasgupta (economist), K.G. Mäler (economist), R.B. Norgaard (economist), Charles Perrings (economist), Manfred Max-Neef (economist), Ignacy Sachs (economist), Joan Martinez-Alier (economist), Bill Rees (ecologist), and Mathis Wackernagel (ecologist), roughly divided by half into ecologists (biologists, physicists) and economists according to their original training. Their names have often appeared in the present article.

Outlook

Ecological economics is then a transdisciplinary field born in the late 1980s, founded by ecologists and by dissident economists, which studies new topics and applies methods such as:

- New indicators and indices of (un)sustainability of the economy;
- Ecological macroeconomics without growth, the debate between 'weak' and 'strong' notions of sustainability;
- The application of ecological notions of carrying capacity and resilience to human ecosystems;
- The valuation and payment for environmental services, monetary valuation of externalities but also the discussion on incommensurability of values;
- Risk assessment, uncertainty, complexity, and 'postnormal' science; integrated environmental assessment, including building of scenarios, dynamic modeling, participatory multicriteria methods of decision making;
- The allocation of property rights and its relation to natural resource management, old and new communal institutions for environmental management;
- Environmental causes and consequences of technological change, relations between ecological economics and evolutionary economics;
- Theories of consumption (needs, satisfactors), as they relate to environmental impacts;
- Relations to industrial ecology, applications to business administration, corporate liability and accountability;
- Relations to fields such as feminist economics; environmental and economic history; environmental ethics; political ecology, urban planning, peasant studies;
- Environmental public policies, instruments of environmental policy often centered on the 'precautionary principle' (or 'safe minimum standards', as introduced by Ciriacy-Wantrup).

See also: Energy and Society; Environmental Economics; Environmental History; Environmental Policy: Protection and Regulation; Human Ecology; Industrial Ecology; Sustainable Production and Consumption.

Bibliography

Agarwal, B., 2010. Gender and Green Governance: The Political Economy of Women's Presence within and beyond Community Forestry. Oxford University Press, Oxford.

Altvater, E., 2007. Marxism and ecological economics: toward a red and green political economy. Monthly Review 58 (8), 55–64.

Ayres, R.U., Warr, B., 2009. The Economic Growth Engine: How Energy and Work Drive Material Prosperity. Edward Elgar, Cheltenham, UK & Northampton, MA.

Berkes, F., Folke, C. (Eds.), 1998. Linking Social and Ecological Systems: Management Practices and Social Mechanisms for Building Resilience. Cambridge University Press, Cambridge.

Bonaiuti, M. (Ed.), 2011. From Bioeconomics to Degrowth. Georgescu-Roegen's 'New Economics' in Eight Essays. Routledge, London.

Boulding, K., 1966. The economics of the coming spaceship earth. In: Jarrett, H. (Ed.), Environmental Quality in a Growing Economy. Resources for the Future/Johns Hopkins Press, Baltimore, pp. 3–14.

Cleveland, C., Stern, D.I., Costanza, R. (Eds.), 2001. The Economics of Nature and the Nature of Economics. Edward Elgar, Cheltenham and International Society for Ecological Economics.

Costanza, R., 1980. Embodied energy and economic valuation. Science 210, 1219–1224.

Costanza, R. (Ed.), 1991. Ecological Economics: The Science and Management of Sustainability. Columbia University Press, New York.

Costanza, R., d'Arge, R., de Groot, R., Farber, S., Grasso, M., Hannon, B., Naeem, S., Limburg, K., Paruelo, J., O'Neill, R.V., Raskin, R., Sutton, P., van den Belt, M., 1997a. The value of the world's ecosystem services and natural capital. Nature 387, 253–260.

Costanza, R., Cleveland, C., Perrings, C. (Eds.), 1997b. The Development of Ecological Economics. Edward Elgar, Cheltenham.

Daily, G.C. (Ed.), 1997. Nature's Services: Societal Dependence on Natural Ecosystems. Island Press, Washington, DC.

Daly, H. (Ed.), 1973. Toward a Steady-State Economy. W.H. Freeman, San Francisco.

Daly, H., 1984. Alternative strategies for integrating economics and ecology. In: Jansson, A.M. (Ed.), Integration of Economy and Ecology. Askö Laboratory, University of Stockholm.

Daly, H., Cobb, J., 1989. For the Common Good: Redirecting the Economy toward Community, the Environment and a Sustainable Future, second ed. Beacon Press, Boston, MA. 1994.

Daly, H., Farley, J., 2010. Ecological Economics: Principles and Applications. Island Press, Washington, DC.
De Groot, R.S., 1992. Functions of Nature: Evaluation of Nature in Environmental Planning, Management and Decision Making. Wolters-Noordhoff, Amsterdam.
Ehrlich, P.R., 1968. The Population Bomb. Sierra Club/Ballantine Book, New York.
Foster, J.B., 1999. Marx's theory of metabolic rift: classical foundations for environmental sociology. The American Journal of Sociology 105 (2), 366–405.
Funtowicz, S., Ravetz, J., 1994. The worth of a songbird: ecological economics as a post-normal science. Ecological Economics 10, 189–196.
Georgescu-Roegen, N., 1971. The Entropy Law and the Economic Process. Harvard University Press.
Georgescu-Roegen, N., 1977. Matter matters too. In: Wilson, K.D. (Ed.), Prospect for Growth: Changing Expectations for the Future. Praeger, New York, pp. 293–313.
Giampietro, M., Mayumi, K.M., Sorman, A.H., 2013. Energy Analysis for a Sustainable Future: Multi-Scale Integrated Analysis of Societal and Ecosystem Metabolism. Earthscan, Routledge, London.
Gowdy, J., Mesner, S., 1998. The evolution of Georgescu-Roegen's bioeconomics. Review of Social Economy 56 (2), 136–156.
Grinevald, J., 1976. La Révolution carnotienne. Thermodynamique, économie et idéologie. Revue européenne des sciences sociales et Cahiers Vilfredo Pareto 36, 39–79. Paris.
Guha, R., Martinez-Alier, J., 1997. Varieties of Environmentalism. Essays North and South. Earthscan, London.
Hall, C., Cleveland, C., Kaufman, R., 1986. Energy and Resources Quality: The Ecology of the Economic Process. Wiley, New York.
Hamilton, C., 2013. Earthmasters: The Dawn of the Age of Climate Engineering. Yale U.P., New Haven.
Hardin, G., 1968. The tragedy of the commons. Science 162, 1243–1248.
Hardin, G., 1974. Lifeboat ethics: the case against helping the poor. Psychology Today 8 (4), 38–43.
Hayek, F.A., 1952. The Counter-Revolution of Science: Studies of the Abuse of Reason. The Free Press, Glencoe, IL.
Healy, H., Martínez-Alier, J., Temper, L., Walter, M., Gerber, J.-F. (Eds.), 2012. Ecological Economics from the Ground Up. Routledge, London.
Hoekstra, A.Y., Chapagain, A.K., 2007. Water footprints of nations: water use by people as a function of their consumption pattern. Water Resources Management 21 (1), 35–48.
Holt, R.P.F., Pressman, S., Spash, C.L., 2009. Post Keynesian and Ecological Economics: Confronting Environmental Issues. Edward Elgar, Cheltenham.
Hornborg, A., 1998. Towards an ecological theory of unequal exchange: articulating world system theory and ecological economics. Ecological Economics 25, 127–136.
Hueting, R., 1980. New Scarcity and Economic Growth: More Welfare through Less Production? North Holland, Amsterdam.
Jackson, T., 2009. Prosperity without Growth. Economics for a Finite Planet. Routledge, London.
Kapp, K.W., 1950. The Social Costs of Private Enterprise. Shocken, New York.
Kerschner, C., 2010. Economic de-growth vs. steady-state economy. Journal of Cleaner Production 18 (6), 544–551.
Krausmann, F., Erb, K.-H., Gingrich, S., Haberl, H., Bondeau, A., Gaube, V., Lauk, C., Plutzar, C., Searchinger, T.D., 2013. Global human appropriation of net primary production doubled in the 20th century. Proceedings of the National Academy of Sciences. http://dx.doi.org/10.1073/pnas.1211349110.
Kumar, P. (Ed.), 2010. The Economics of Ecosystems and Biodiversity: Ecological Economics Foundations. Earthscan, London and Washington, DC.
Martinez-Alier, J., 2002. The Environmentalism of the Poor: A Study of Ecological Conflicts and Valuation. Edward Elgar, Cheltenham. Oxford University Press, Delhi, 2005.
Martinez-Alier, J., Schlüpmann, K., 1987. Ecological Economics: Energy, Environment and Society. Blackwell, Oxford.
Martinez-Alier, J., Munda, G., O'Neill, J., 1998. Weak comparability of values as a foundation for ecological economics. Ecological Economics 26 (3), 277–286.
Martinez-Alier, J., Røpke, I. (Eds.), 2008. Recent Developments in Ecological Economics, 2 vols. Edward Elgar, Cheltenham.
Martinez-Alier, J., Pascual, U., Vivien, F., Zaccai, E., 2010. Sustainable de-growth: mapping the context, criticisms and future prospects of an emergent paradigm. Ecological Economics 69 (9), 1741–1747.
Max-Neef, M., 1992. From the Outside Looking in: Experiences in Barefoot Economics. Zed Press, London.
Mayumi, K., 2001. The Origins of Ecological Economics: The Bioeconomics of Georgescu-Roegen. Routledge, London.
Munda, G., 2008. Social Multi-Criteria Evaluation for a Sustainable Economy. Springer, Heidelberg, New York.
Muradian, R., Corbera, E., Pascual, U., Kosoy, N., May, P., 2010. Reconciling theory and practice: an alternative conceptual framework for understanding payments for environmental services. Ecological Economics 69, 1202–1208.
Naredo, J.M., 1987. La economía en evolución. Historia y perspectivas de las categorías básicas del pensamiento económico. Siglo XXI, Madrid.
Norgaard, R.B., 1990. Economic indicators of resource scarcity: a critical essay. Journal of Environmental Economics and Management 19 (1), 19–25.
O'Connor, J., 1988. Capitalism, nature, socialism: a theoretical introduction. Capitalism Nature Socialism 1, 11–39.
O'Connor, M., 1993. Value system contests and the appropriation of ecological capital. The Manchester School 61, 398–424.
O'Connor, M., Spash, C. (Eds.), 1999. Valuation and the Environment. Edward Elgar, Cheltenham.
O'Neill, J., 1993. Ecology, Policy and Politics. Routledge, London.
Ostrom, E., 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, Cambridge, UK and New York.
Passet, R., 1979. L'Économique et le Vivant, second ed. 1996. Economica, Paris.
Pengue, W., 2005. Transgenic crops in Argentina. The ecological and social debt. Bulletin of Science, Technology and Society 25 (4), 314–322.
Pérez-Rincón, M.A., 2006. Colombian international trade from a physical perspective: towards an ecological "Prebisch thesis". Ecological Economics 59 (4), 519–529.
Perrings, Ch, 1987. Economy and Environment. Cambridge University Press, Cambridge.
Pimentel, D., Hurd, L.E., Bellotti, A.C., Forster, M.J., Oka, I.N., Sholes, O.D., Whitman, R.J., 1973. Food production and the energy crisis. Science 182, 443–449.
Polanyi, K., 1957. Aristotle discovers the economy. In: Polanyi, K., Arensberg, C.M., Pearson, H.W. (Eds.), Trade and Markets in the Early Empires: Economies in History and Theory. The Free Press, Glencoe.
Podolinsky, S., 1880. Le Socialisme et l'unité des forces physiques. La Revue Socialiste 8, 353–365.
Podolinsky, S., 1883. Menschliche Arbeit und Einheit der Kraft. Die Neue Zeit 1 (9), 413–424 and 1 (10), 44–457.
Prigogine, I., Stengers, I., 1984. Order Out of Chaos: Man's New Dialogue with Nature. Bantam Books, New York.
Rappaport, R., 1968. Pigs for the Ancestors: Ritual in the Ecology of a New Guinea People. Yale University Press, New Haven.
Ronsin, F., 1980. La grève des ventres; propagande malthusienne et baisse de la natalité en France, XIXe-XXe siècles. Aubier-Montaigne, París.
Russi, D., Gonzalez-Martinez, A.C., Silva-Macher, J.C., Giljum, S., Martínez-Alier, J., Vallejo, M.C., 2008. Material flows in latin america. Journal of Industrial Ecology 12 (5–6), 704–720.
Schrödinger, E., 1944. What is Life: The Physical Aspect of the Living Cell. Cambridge University Press, Cambridge.
Shmelev, S.E., 2012. Ecological Economics. Sustainability in Practice. Springer, Heidelberg, New York.
Sieferle, R.P., 2001. The Subterranean Forest: Energy Systems and the Industrial Revolution. The White Horse Press, Cambridge.
Spash, C.L., 2007. The economics of climate change impacts à la Stern: novel and nuanced or rhetorically restricted? Ecological Economics 63 (4), 706–713.
Spash, C.L. (Ed.), 2009. Ecological Economics, 4 vols. Routledge, London.
Spash, C.L., 2011. Terrible economics, ecosystems and banking. Environmental Values 20 (2), 141–145.
Spash, C., 2012. New foundations for ecological economics. Ecological Economics 77, 36–47.
Srinivasan, U.T., Carey, S.P., Hallstein, E., Higgins, P.A., Kerr, A.C., Koteen, L.E., Norgaard, R.B., 2008. The debt of nations and the distribution of ecological impacts from human activities. Proceedings of the National Academy of Sciences 105 (5), 1768–1773.
Sukhdev, P., 2012. Corporation 2020: Transforming Business for Tomorrow's World. Island Press, Washington, DC.
Tamanoi, Y., Tsuchida, A., Murota, T., 1984. Towards an entropic theory of economy and ecology—beyond the mechanistic equilibrium approach. Economie Apliquée 37, 279–294.
Temper, L., Martinez-Alier, J., 2013. The God of the mountain and Godavarman: net present value, indigenous territorial rights and sacredness in a bauxite mining conflict in India. Ecological Economics 96, 79–87.
Vatn, A., 2005. Institutions and the Environment. Edward Elgar, Cheltenham.
Vatn, A., Bromley, D., 1994. Choices without prices without apologies. Journal of Environmental Economics and Management 26, 129–148.
Vernadsky, V., 1924. La Géochimie. Alcan, Paris.
Victor, P., 1991. Indicators of sustainable development: some lessons from capital theory. Ecological Economics 4 (3), 191–213.

Victor, P., 2008. Managing without Growth. Slower by Design, Not Disaster. Edward Elgar, Cheltenham.

Vitousek, P., Ehrlich, P., Erhlich, A., Matson, P., 1986. Human appropriation of the products of photosynthesis. Bioscience 34, 368–373.

Wackernagel, M., Rees, W., 1995. Our Ecological Footprint. New Society Publishers, Gabriola Island and Philadelphia.

Waring, M., 1988. If Women Counted: A New Feminist Economics. Harper & Row, San Francisco.

Weisz, H., Krausmann, F., Amann, C., Eisenmenger, N., Erb, K.-H., Hubacek, K., Fischer-Kowalski, M., 2006. The physical economy of the European union: cross-country comparison and determinants of material consumption. Ecological Economics 58 (4), 676–698.

Zografos, C., Howarth, R.B. (Eds.), 2008. Deliberative Ecological Economics. Oxford University Press, Oxford.

Ecological Fallacy, Statistics of

Glenn Firebaugh, Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 4023–4026, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Researchers commit an 'ecological fallacy' when they offer mistaken inferences about individual-level relationships using associations observed only in aggregate data. This mistaken inference practice is labeled a fallacy because it is based on the false assumption that relationships that hold at one level of aggregation must also hold at another level of aggregation. Because individual-level data often are unavailable, researchers are at risk of succumbing to ecological fallacies in virtually all the social and behavioral sciences. In this article, the methods of Firebaugh and King are presented as attempts to avoid fallacies and improve interpretations of available aggregate data.

Researchers are said to commit the 'ecological fallacy' when they make untested inferences about individual-level relationships from aggregate data. This practice is called a fallacy because it is based on the problematic assumption that relationships that hold at one level of aggregation also hold at another level of aggregation. Researchers are subject to the ecological fallacy in virtually all the social and behavioral sciences, from history (where individual-level data often are unavailable) to criminology (where, for example, a positive relationship between poverty rate and crime rate across cities might be interpreted as evidence that poor people commit more crime) to epidemiology (where preliminary studies might correlate cancer rates across regions with other regional characteristics to decide what sorts of expensive individual-level data to collect).

This article cites examples of the ecological fallacy, explains why inference across levels of aggregation is sometimes unavoidable in social science research, and overviews solutions that have been proposed for alleviating the problem. Because individual-level data are not available for studying many issues of concern to social and behavioral scientists, the article emphasizes the theoretical conditions under which individual-level and aggregate-level relationships are the same.

An Example

One important contemporary example involves the relationship between race and voting behavior in the United States. The assumption is that voters have a propensity to vote for someone of their own race: whites are more likely to vote for a white candidate, blacks for a black candidate, and so on. If there is a propensity for intra-racial voting, candidates are handicapped in political districts where their race is different from that of the majority of the voters. A probable outcome is that members of minority races will be underrepresented in governing bodies because of their minority status in almost all political districts. One proposed solution is to use race as an explicit criterion for determining political districts, with the goal of drawing district boundaries in such a way that blacks and other minorities are in the majority in some of the election districts.

This issue poses the classic ecological fallacy problem for researchers because the use of the secret ballot rules out the best source of data on how individuals actually voted. To make sound policy decisions about the use of race as a criterion for drawing up political districts, policy makers should know the strength of intra-race vote propensity, yet individual-level data typically are not available for estimating that propensity. Voting data come in aggregated form, by precincts. Hence researchers often must rely on correlations for voting data aggregated to the precinct level. In an election pitting a white candidate against a black candidate, for example, researchers might attempt to use the correlation between a precinct's percentage of white voters and the percentage of votes for the white candidate to draw inferences about the propensity of whites to vote for white candidates.

Those inferences are problematic because relationships at the precinct level need not equal relationships at the individual level. Even the direction of the relationship can be different for individuals and aggregates. The 1968 US Presidential election provides a real-world example. In that election, George Wallace, a third-party candidate who was known primarily for his segregationist position as Governor of Alabama, captured about 10% of the national vote. What is relevant here is that in the South he tended to receive a higher percentage of the vote in congressional districts with higher percentages of blacks (the correlation across districts in the South was $r = 0.55$ (Firebaugh, 1978)). Because no one believes that blacks were more inclined than other racial groups to vote for George Wallace, this is a compelling example of the danger of using aggregate data to draw inferences about the behavior of individuals.

Why was there a positive correlation between percent black and vote for Wallace in the 1968 election? The answer is that whites in heavily black districts were more likely to vote for Wallace than were whites in districts with fewer blacks. In other words, in this election the effect of one's race on one's voting behavior was affected by one's racial context – an example of a contextual effect (*see* Contextual Studies: Methodology). Subsequently, this article shows that relationships differ at the

individual and aggregate levels when such contextual effects are present.

The Impact of the Ecological Fallacy on Research Practices

Although the term ecological fallacy is most often associated with W.S. Robinson, Robinson does not use the term in his classic 1950 article. Instead he uses the term 'ecological correlation,' and he states that his analysis "provides a definite answer as to whether ecological correlations can validly be used as substitutes for individual correlations. They cannot" (Robinson, 1950: p. 357). Robinson uses illiteracy data from the 1930 US census to show the discrepancy between ecological correlations and individual correlations. The correlation between percent black and percent illiterate is $r = 0.77$ across States and $r = 0.95$ across nine larger geographic divisions of the United States. The size of those ecological correlations stands in stark contrast to the size of the individual-level relationship between race (black vs others) and illiteracy ($r = 0.20$). As a second example, Robinson notes that the ecological correlation of percent foreign-born and percent illiterate across States is large and negative ($r = -0.53$), suggesting that the foreign-born are substantially less likely to be illiterate. Yet the foreign-born actually are more likely to be illiterate, as shown by the positive correlation at the individual level ($r = 0.12$).

While Robinson focused on the danger of using aggregate data to draw conclusions about the behavior of individual people, the ecological fallacy problem in fact applies to inferences across levels of aggregation in general. Individuals need not be people. The terms individual level and aggregate level as used here refer to relative levels of aggregation. To illustrate, the term 'individual' could refer to a single church congregation within a population of congregations. The ecological fallacy problem would arise if a researcher tried to use data on racial diversity and attendance growth rates across church denominations (collections of congregations) to draw conclusions about the effect of racial diversity on local church growth. Because parishioners attend a local congregation, it is racial diversity at the local level, not at the denomination level, which is the natural focus here. One could of course survey local congregations in this instance, but the survey would ask questions about a collective (a congregation) rather than about a person.

It would be difficult to overstate the impact Robinson's article has had on social science research during the second half of the twentieth century. The use of ecological correlations to study individual-level relationships had been commonplace before Robinson's article, and the article sharply curtailed that practice. The article also served to motivate the development of survey research. If aggregate data are not adequate to study individuals, then social scientists need data on individuals. One efficient way to gather data on individuals is to ask them questions. Hence, in this way Robinson's message about the need for individual-level data to study individuals no doubt played a role in the amassing of the large survey data sets that have become standard fare in social science research in the twenty-first century.

It is sometimes difficult or impossible to collect survey data to overcome the ecological fallacy problem, however, for two main reasons. First, individuals might be hard to reach. The extreme case of this occurs in historical research, where the pertinent individuals are dead. For this reason, historical research is especially plagued by the ecological fallacy problem (geography is another discipline especially affected, since geographic research is often based on spatial units). Aggregate data often are the only data available to historians, and it is not possible to go back in time to collect data to determine who voted for the Nazis, for example. Second, individuals might be reachable, but the needed information is sensitive. To protect the privacy of individuals, government agencies in some instances aggregate individual-level data before releasing the data to researchers. But aggregation in turn creates the risk that researchers using the aggregate data might commit the ecological fallacy.

Mitigating the Ecological Fallacy Problem

No expert disputes Robinson's claim that aggregate-level relationships very often differ markedly from individual-level relationships, but a very active literature since Robinson takes issue with his strict prohibition against the use of aggregate data to study individual-level relationships. Necessity motivates this literature; given the unavailability of individual-level data for many issues of interest in the social and behavioral sciences, the Robinson prohibition would rule out research in some areas of compelling interest, as illustrated by the earlier example of race and voting behavior.

Writing shortly after the Robinson broadside, Goodman (1953, 1959) observed that aggregation tends to affect correlation coefficients (r) more than it does regression slopes and that, under certain statistical conditions, aggregate-level and individual-level slopes are the same. The challenge has been to translate Goodman's statistical conditions into broadly applicable substantive terms that researchers can use. That challenge has resulted in extensive literature on the ecological fallacy in sociology, geography, economics, political science, and history. One prominent approach seeks to spell out the conditions under which aggregate-level coefficients provide unbiased estimates of individual-level parameters. A second approach seeks to use a method-of-bounds to derive estimates of individual-level effects. This article concludes by briefly overviewing the two approaches.

Equivalence Conditions for Aggregate-level and Individual-level Parameters

Since the Robinson and Goodman articles in the 1950s, much of the methodological work on the ecological fallacy has been devoted to spelling out the conditions under which aggregate-level coefficients provide unbiased estimates of individual-level relationships. Following Goodman's lead, this work focuses on regression slopes, since the bias tends to be smaller for slopes than for correlation coefficients. Hannan and Burstein (1974) consider the effects of the variable by which the data are grouped (e.g., precincts, schools, and States). For the bivariate case, they show that aggregate data yield unbiased estimates of individual-level relationships when the grouping variable is either (1) uncorrelated with the independent variable or

(2) uncorrelated with the dependent variable, with the independent variable controlled for. Applied to the example of race and voting behavior, the precinct-level data provide unbiased estimates of individual-level voting behavior when precinct is unrelated to race, or precinct is unrelated to voting behavior, controlling for race. The first condition is improbable, since neighborhoods in the United States tend to be highly segregated by race. The second condition (voting behavior is unrelated to precinct, race controlled for) means that one's propensity to vote for a same-race candidate is unaffected by the racial makeup of one's precinct. Thus, whites in majority-black precincts are as likely to vote for the white candidate as whites in majority-white precincts, blacks in majority-white precincts are as likely to vote for the black candidate as blacks in majority-black precincts, and so on. Unless Americans of all races are remarkably oblivious to their social contexts, this assumption is unlikely to hold.

Firebaugh (1978) subsumes the Hannan–Burstein conditions under a single rule. The rule states that the aggregate-level regression slope provides an unbiased estimate of the corresponding individual-level effect of X on Y if and only if aggregated X (denoted $\overline{X}$ since aggregated X is usually a type of mean) is unrelated to Y, controlling for individual-level X. For example, using precinct-level data to estimate the effect of one's race on one's vote will yield a biased estimate of that effect unless the racial context of precincts is unrelated to one's vote. Firebaugh shows that the $\overline{X}$ rule generalizes to the multivariate case where a researcher wants to estimate the effect of more than one independent variable.

Firebaugh's $\overline{X}$ rule suggests one solution to the ecological fallacy problem: eliminate $\overline{X}$ effects. To eliminate X effects, researchers must determine why $\overline{X}$ might have an effect on Y beyond the effect of individual X, and then add variables to try to remove the effect. Using the voting example to illustrate the first step is to ask why it is that the racial composition of one's precinct might be related to one's vote, independent of one's race. The most likely answer is that minorities in largely white precincts differ from minorities in largely nonwhite precincts on a host of factors, *other than race*, that are related to voting characteristics, such as income, education, and home ownership. These nonracial characteristics of the precincts should be entered into the aggregate-level equation as control variables. Here the logic of control is the conventional logic used in nonexperimental analysis: regressors are included (average income in the precinct, average education level, percent homeowners, etc.) in an attempt to use the least-squares method to 'equate' the precincts on those properties (Gelman and Hill, 2007). If the method is successful, the aggregate-level slope for race will reflect the effect of race for voters who live in precincts that are otherwise equivalent – that is, the aggregate-level slope will reflect the effect of race itself on vote.

Bounding Individual-level Parameters

The $\overline{X}$ rule is not a foolproof solution to the ecological fallacy problem since there is no way (with aggregate data only) to test the assumption that $\overline{X}$ effects have been eliminated. Hence, methodologists continue to work on the problem. One line of work gains leverage on the problem by applying the insight that marginal frequencies bound the cell frequencies in a contingency table (Duncan and Davis, 1953). For example, in a precinct where 500 whites and 100 nonwhites collectively cast 450 votes for candidate A, the number of nonwhites voting for candidate A is bounded by 0 and 100 and the number of whites voting for candidate A is bounded by 350 and 450.

A book by King (1997) entitled *A Solution to the Ecological Inference Problem* combines the bounding principle of contingency tables with probabilistic statistical methods to fashion a sequential procedure for estimating individual-level coefficients and their confidence intervals. In essence, the method begins with the deterministic information provided by the bounds for each of the aggregates (e.g., precincts) and then applies likelihood methods to narrow probabilistically within those bounds. A public-domain computer program provides estimates, visual depictions, and diagnostic tests. King's method is the most ambitious attempt to date to provide a general solution to the ecological fallacy problem. Ultimately, the worth of a tool is proven in use, and the most telling test of King's method and its successors will come as researchers apply them (see King et al., 2008 for an example of the method).

See also: Causal Counterfactuals in Social Science Research; Causal Inference: Overview; Ecological Inference; Linear Hypothesis: Fallacies and Interpretive Problems (Simpson's Paradox); Linear Hypothesis; Mechanism-Based Causal Analysis.

Bibliography

Achen, C.H., Shively, W.P., 1995. Cross-Level Inference. University of Chicago Press, Chicago.

Duncan, O.D., Davis, B., 1953. An alternative to ecological correlation. American Sociological Review 18, 665–666.

Freedman, D., Klein, S., Sacks, J., Smyth, C., Everett, C., 1991. Ecological regressions and voting rights. Evaluation Review 15, 673–711.

Firebaugh, G., 1978. A rule for inferring individual-level relationships from aggregate data. American Sociological Review 43, 557–572.

Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Goodman, L., 1953. Ecological regressions and behavior of individuals. American Sociological Review 18, 663–664.

Goodman, L., 1959. Some alternatives to ecological correlation. American Journal of Sociology 64, 610–625.

Hannan, M., Burstein, L., 1974. Estimation from grouped observations. American Sociological Review 39, 374–392.

King, G., 1997. A Solution to the Ecological Inference Problem. Princeton University Press, Princeton, NJ.

King, G., Rosen, O., Tanner, M., Wagner, A.F., 2008. Ordinary economic voting behavior in the extraordinary election of Adolf Hitler. Journal of Economic History 68, 951–996.

Robinson, W.S., 1950. Ecological correlations and the behavior of individuals. American Sociological Review 15, 351–357.

Ecological Inference

David A Freedman, University of California, Berkeley, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 4027–4030, © 2001, Elsevier Ltd.

Abstract

There is a long tradition, in the social and life sciences, of attempts to infer individual behavior from group behavior. Durkheim's book on suicide is an early example. The 'ecological fallacy' consists of ascribing to individuals the characteristics of groups to which they belong, even though the relevant individuals may not share such characteristics. The issues are discussed in seminal papers by Robinson, Duncan and Davis, and Goodman, which are summarized here. Ecological regression is considered, along with recent work on random-coefficients models. Examples are drawn from epidemiology, voting rights, and survey data.

In nineteenth century Europe, suicide rates were higher in countries that were more heavily Protestant, the inference being that suicide was promoted by the social conditions of Protestantism (Durkheim, 1897); also see Neeleman and Lewis (1999). According to Carroll (1975), death rates from breast cancer are higher in countries where fat is a larger component of the diet, the idea being that fat intake causes breast cancer. These are 'ecological inferences,' that is, inferences about individual behavior drawn from data about aggregates.

To continue with Durkheim, the Protestant countries were different from the Catholic countries in many ways besides religion (the problem of 'confounding'). Moreover, his data do not tie individual suicides to any particular religious faith. Confounding must be dealt with in any observational study. But the second problem – that exposure and response are measured only for aggregates rather than for individuals – is specific to ecological studies. If there is no confounding, the expected difference between effects for groups and effects for individuals is 'aggregation bias'; in general, the difference is partly attributable to confounding and partly to aggregation bias.

The ecological fallacy consists of thinking that relationships observed for groups necessarily hold for individuals: if countries with more Protestants have higher suicide rates, then Protestants must be more likely to commit suicide; if countries with more fat in the diet have higher rates of breast cancer, then women who eat fatty foods must be more likely to get breast cancer. These inferences may be correct, but are only weakly supported by the aggregate data.

Ecological studies in epidemiology yielded important insights for Snow (1855), Finlay (1881), Goldberger (Terris, 1964), and Dean (1938) among others. However, it is all too easy to draw incorrect conclusions from aggregate data. Greenland and Robins (1994) review the issues. For one example, recent studies of individual-level data cast serious doubt on the link between breast cancer and fat intake (Holmes et al., 1999). Another well-known example, on the sources of popular support for the Nazi party pre-war Germany, is discussed by Lohmoller et al. (1985).

Ecological Correlation

Robinson (1950) discusses ecological inference, stressing the difference between ecological correlations and individual correlations. One striking example is the relationship between nativity and literacy. For each of the 48 states in the US of 1930, Robinson computes two numbers: the percentage of the population who are foreign-born, and the percentage who are literate. (The base for each percentage is the number of residents of the state who are 10 or older; data are from the Census of 1930.) The correlation between the 48 pairs of numbers is 0.53. This is an 'ecological' correlation, because the unit of analysis is not an individual person but a group of people – the residents of a state. The ecological correlation suggests a positive association between foreign birth and literacy: the foreign-born are more likely to be literate (in American English) than the native-born. In reality, the association is negative: the correlation computed at the individual level is −0.11. The ecological correlation gives the wrong inference. The sign of the correlation is positive because the foreign-born tend to live in states where the native-born are relatively literate.

The example can be replicated using data from the March 1995 Current Population Survey, restricted to persons age 25 and over. Literacy is not reported, but incomes are. Say that a person has 'high' income if family income in 1994 was $50 000 or more. Figure 1 shows for each of the 50 states the

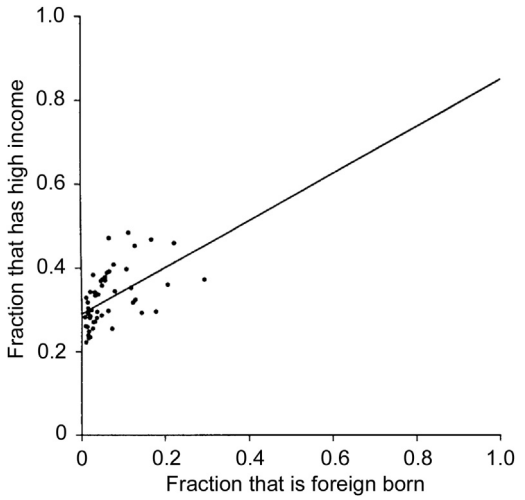

Figure 1 For each state, the horizontal axis shows the fraction of the population aged 25 and over that is foreign-born; the vertical axis shows the fraction having high incomes. The regression line is also shown.

fraction of persons who are foreign-born, and the fraction of persons who have high incomes. The regression line is shown too. The ecological correlation (across the 50 areas) is 0.52, suggesting that the foreign-born have higher incomes on the whole than the native-born. The truth is the opposite. About 35% of the native-born have high incomes, compared to 28% for the foreign-born. The correlation at the individual level is −0.05. In Figure 1, there is no state where the fraction of foreign-born approaches 1. However, the same sort of reversal can occur even with a full spectrum of x-values. The issue is the disaggregation, not the range of the data.

Ecological Regression

Under certain restrictive assumptions, individual behavior can be inferred from aggregate data using a technique called 'ecological regression.' The unit of analysis in the regression is a group of people, typically defined by geography (as in Figure 1). The technique has been widely used in voting rights litigation in the US. For discussion from various perspectives, see Dogan and Rokkan (1969), Freedman et al. (1991), Grofman and Davidson (1992), Achen and Shively (1995), and Cho (1998).

It may, for instance, be desired to estimate the support for a particular candidate among Hispanics and non-Hispanics. For each precinct in an electoral contest, suppose we know the fraction x of voters who are Hispanic; we also know the fraction y of votes obtained by the candidate of interest. We would not know the fraction of Hispanic voters who voted for the candidate, due to the secrecy of the ballot.

A regression equation can be fitted to the data:

$$y_i = a + bx_i + \varepsilon_i \qquad [1]$$

In eqn. [1], the subscript i indexes precincts; x_i is the fraction of voters in precinct i who are Hispanic; y_i is the fraction of votes cast in that precinct for the candidate of interest. (More exactly, y_i is the number of votes for that candidate, divided by the number of voters.) The precinct-level disturbance terms ε_i would typically be assumed to be independent and identically distributed with mean 0, although this is seldom made explicit.

The parameters a and b in eqn. [1] can be estimated by ordinary least squares; call the estimates $\widehat{a}$ and $\widehat{b}$, respectively. Then $\hat{a}$ would be interpreted as the fraction of non-Hispanic voters who supported the candidate in question: after all, $\hat{a}$ is the height of the regression line at $x = 0$, corresponding to precincts with no Hispanic voters. Likewise, $\widehat{a} + \widehat{b}$ (the height of the regression line at $x = 1$) would be interpreted as the fraction of Hispanic voters who supported the candidate.

The data are available for groups defined by area of residence; the inference is to groups defined by ethnicity. The statistical logic connecting the inference to the data rests on the 'constancy assumption,' that voting preferences within ethnic groups do not systematically depend on the ethnic makeup of the area of the residence (Goodman, 1953, 1959). In Figure 1, the regression equation is $y = 0.29 + 0.56x$, leading to the inference that 30% of the native-born have high incomes, compared to 29% + 56% = 85% of the foreign-born. In reality, as noted above, the percentages are 35 and 28%. The ecological inference is in error, because the constancy assumption fails: the incomes of the native-born increase systematically with fraction of foreign-born in the state. Immigration to the US tends to concentrate in richer states – California, Hawaii, and New York rather than Kentucky, Tennessee, and West Virginia.

The Method of Bounds

In 1995 in the state of Washington, 8% of the population was foreign born, while 34% had high incomes. These are known. Let p be the fraction of the foreign-born who had high incomes, and let q be the corresponding fraction for the native-born. Investigators making ecological inferences would typically not know the fractions p and q. The two unknowns satisfy an equation:

$$0.08p + (1 - 0.08)q = 0.34 \qquad [2]$$

We have one equation with two unknowns: that is the whole problem of ecological inference. However, eqn. [2] does contain some information, because $q = (0.34 - 0.08p)/(1 - 0.08)$. Since p is bounded between 0 and 1, it follows that q is bounded below by $(0.34 - 0.08)/(1 - 0.08) \doteq 0.26$. Similarly, the upper bound is $0.34/0.92 = 0.37$. This is 'the method of bounds' (Duncan and Davis, 1953). Such bounds can be computed for each study area, and then aggregated. However, in typical applications, the bounds are too broad to be informative (also see Goodman, 1959).

Ecological Regression with Random Coefficients

Models with random coefficients have been used to make ecological inferences. By way of illustration, we consider a model for nativity and income. Index the study areas by i. As before, let x_i be the fraction of the population in area i that is foreign born, and y_i the fraction with high incomes. These fractions are known, as is the total population n_i of area i. Let p_i be the fraction of the foreign-born with high incomes, and let q_i be the corresponding fraction for the native-born, so

$$y_i = p_i x_i + q_i(1 - x_i) \qquad [3]$$

The problem is to make inferences about p_i and q_i. To solve such problems, King (1997) assumes the pairs (p_i, q_i) to be independent and identically distributed across the study areas. This is his version of the constancy assumption: the statistical behavior of a demographic group is not allowed to depend on area of residence. The parent distribution is taken to be bivariate normal, conditioned to lie in the unit square so that p_i and q_i are both between 0 and 1. The five parameters of this parent normal distribution can be estimated from the data by maximum likelihood, and estimates $(\widehat{p}_i, \widehat{q}_i)$ can be derived from eqn. [3]. Rates for each demographic group can then be estimated by addition over areas, as $\sum_i n_i \widehat{p}_i / \sum_i n_i x_i$ and $\sum_i n_i \widehat{q}_i / \sum_i n_i (1 - x_i)$, respectively. More complex models have also been developed.

To show the force of the constancy assumption, we can use the 'neighborhood model' which assumes that behavior is determined by geography not demography. Then $\widehat{p}_i = \widehat{q}_i = y_i$ for each study area i. (In our example of nativity and income, if

33% of the residents of a particular study area have high incomes, this percentage applies equally to the foreign-born and the native-born in that area.) The neighborhood model turns the constancy assumption on its head. In a variety of test applications where all the data are available, the neighborhood model gives more accurate estimates for demographic groups than ecological regression or random-coefficients models. In some contexts, of course, the neighborhood model proves deficient (Achen and Shively, 1995). If the data are incomplete so estimation is needed, it will be unclear which model if any is giving the right answers. The track record to date favors the neighborhood model (Freedman et al., 1991, 1998). For another perspective see Schuessler (1999).

Summary and Conclusions

Aggregate data are often easier to obtain than data on individuals, and may offer valuable clues about individual behavior. Ecological inferences will therefore continue to be made. The problems of confounding and aggregation bias, however, are unlikely to be resolved in the proximate future.

See also: Ecological Fallacy, Statistics of.

Bibliography

Achen, C.H., Shively, W.P., 1995. Cross-level Inference. University of Chicago Press, Chicago.

Carroll, K., 1975. Experimental evidence of dietary factors and hormone-dependent cancers. Cancer Research 35, 3374–3383.

Cho, W.K.T., 1998. Iff the assumption fits: a comment on the king ecological inference solution. Political Analysis 7, 143–163.

Dean, H.T., 1938. Endemic fluorosis and its relation to dental caries. Public Health Reports 53, 1443–1452.

Dogan, M., Rokkan, S. (Eds.), 1969. Quantitative Ecological Analysis in the Social Sciences. MIT Press, Cambridge, MA.

Duncan, O.D., Davis, B., 1953. An alternative to ecological correlation. American Sociological Review 18, 665–666.

Durkheim, E., 1897. Le Suicide. F. Alcan, Paris [English translation by Spalding, J.A., 1951. Free Press, Collier-MacMillan, Toronto, Canada].

Finlay, C.J., 1881. The Mosquito Hypothetically Considered as the Agent of Transmission of Yellow Fever. Anales de la Academia de Ciencias Medicas, Fisicas y Naturales de Habana XVIII [English translation by Finlay, reprinted in Buck, C., Llopis, A., Nájera, E., Terris, M., 1989. The Challenge of Epidemiology: Issues and Selected Readings, Scientific Publication No. 505. World Health Organization, Geneva.].

Freedman, D.A., Klein, S.P., Ostland, M., Roberts, M.R., 1998. Review of a solution to the ecological inference problem. Journal of the American Statistical Association 93, 1518–1522 with discussion, 1999, 94, 352–357.

Freedman, D.A., Klein, S.P., Sacks, J., Smyth, C.A., Everett, C.G., 1991. Ecological regression and voting rights. Evaluation Review 15, 659–817 with discussion.

Goodman, L., 1953. Ecological regression and the behavior of individuals. American Sociological Review 18, 663–664.

Goodman, L., 1959. Some alternatives to ecological correlation. American Journal of Sociology 64, 610–625.

Greenland, S., Robins, J., 1994. Invited commentary: ecologic studies—biases, misconceptions, and counterexamples. American Journal of Epidemiology 139, 747–760.

Grofman, B., Davidson, C., 1992. Controversies in Minority Voting: The Voting Rights Act in Perspective. Brookings Institution, Washington, DC.

Holmes, M.D., Hunter, D.J., Colditz, G.A., Stampfer, M.J., Hankinson, S.E., Speizer, F.E., Rosner, B., Willett, W.C., 1999. Association of dietary intake of fat and fatty acids with risk of breast cancer. Journal of the American Medical Association 281, 914–920.

King, G., 1997. A Solution to the Ecological Inference Problem. Princeton University Press, Princeton, NJ.

Lohmoller, J.B., Faller, J., Link, A., de Rifke, J., 1985. Unemployment and the rise of national socialism: contradicting results from different regional aggregations. In: Nijkamp, P. (Ed.), Measuring the Unmeasurable. Martinus Nijhoof, Den Haag, pp. 357–370.

Neeleman, J., Lewis, G., 1999. Suicide, religion and socioeconomic conditions. An ecological study in 26 countries, 1990. Journal of Epidemiology and Community Health 53, 204–210.

Robinson, W.S., 1950. Ecological correlations and the behavior of individuals. American Sociological Review 15, 351–357.

Schuessler, A.A., 1999. Ecological inference. Proceedings of the National Academy of Sciences USA 96, 10578–10581.

Snow, J., 1855. On the Mode of Communication of Cholera. Churchill, London. Reprinted by Hafner, New York, 1965.

Terris, M., 1964. Goldberger on Pellagra. Louisiana State University Press, Louisiana.

Ecological Social Work: Shifting Paradigms in Environmental Practice

Fred H Besthorn, School of Social Work, Wichita State University, Wichita, KS, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Conventional models of ecological social work have had profound staying power in the theoretical life of social work. They have become conventional wisdom for much of social work practice. In recent decades, however, a new ecological model has emerged. This new ecological model argues that the core values of social work and its conventional ecological models must be altered to support the new realities of the importance of the natural environment and the impending environmental crisis. This article traces the historical evolution of social work's conventional ecological thinking and traces the contours of a new ecology of social work.

Since the early 1960s, social work sought to establish a conceptual demarcation point between itself and other helping professions through reliance on ecological and environmental metaphors of practice. The language of person-in-environment, person-and-environment, or person/environment emerged as common representations in North America and in other parts of the Western world with respect to how the profession attempted to delineate the nature and focus of social work practice. In large measure, it was an attempt on the part of many social work theorists to summarize, with what appeared a simple and readily accessible vernacular, the unique niche and emphases constituting social work practice. Social work staked out its practice space at the interface of individuals in their unique environmental contexts.

This was not a particularly new development since the profession has, for the better part of its 100-year history, been characterized by alternating attempts to find common ground between personal responsibility and environmental circumstance. It can be seen from the very earliest years of the profession especially in the often stark dissimilarity between the individualistic and moralistic friendly visiting of Charity Organization Societies (COS) and the situational and contextualized in-place environmental change efforts of the Settlement House Movement. COS friendly visitors were tasked with investigating the circumstances surrounding the indigent person's need and to instruct poor people in ways to better manage their personal lives (van Wormer and Besthorn, 2011). The belief that individuals were morally responsible for their own circumstances was unambiguous. Intervention began and ended with the individual. On the other hand, Settlement House efforts were not that of friendly visiting, but rather, were infused with a genuine desire to bridge class differences and develop a more contextualized and less patronizing form of charity. The environment, not the individual person, was the locus of change for these workers. Settlement Worker's efforts to improve community sanitation systems, to alleviate the deplorable living conditions of tenement houses, and their efforts at developing park areas and green spaces for play and recreation were clearly focused on environmental conditions impinging on personal well-being.

Ecological Social Work: A Conventional Perspective

Throughout much of its recent history, social work has sought innovative ways to conceptualize the relationship between the individual and environmental contexts of human functioning and how professional practice engages with those dimensions. While the person-in-environment vernacular has proved a useful linguistic marker attempting to define the parameters of social work practice, not everyone has been convinced of its utility (Kemp et al., 1997). The problematic tension between operationalizing the interactive nature of person-and-environment intervention continued in US social work as an intermittent point of contention in much the same way as had been the case in the early decades of the twentieth century. In an attempt to clarify the perceived imprecision of person-and-environment metaphors and their lack of practical application for frontline social work practice, several North American social work theorists began to integrate ideas from the field of ecology into social work theory (Germain, 1968, 1973, 1978; Germain and Gitterman, 1980). What became popularly known as ecological social work gradually emerged out of a turbulent ideological legacy characterized by reoccurring struggles to realistically conceptualize a dual and balanced obligation to both person and environment. This incremental evolution toward an ecological social work was different in other parts of the social work world but may have followed a similar trajectory.

Ecological theories had been popular for some time in the fields of sociology, anthropology, and psychology prior to their introduction to social work. Carel Germain (1973) first introduced her ecological perspective to social work in part to acknowledge the growing trend to conceive of human development and deteriorating social and biophysical conditions in holistic and systemic terms. Germain understood that in order to fully enhance human well-being, the physical and social environments of persons must be assessed concurrently. Germain argued persuasively that ecological thinking helped social work conceptualize the world in organic terms. It was also more compatible with core social work values since it attempted to integrate broader ecological systems with the lived experience of human beings (Kemp et al., 1997).

Later, Carel Germain collaborated with colleague Alex Gitterman (Germain and Gitterman, 1980) to elaborate on her earlier ecological perspective. Their Life Model of Social Work Practice attempted explicitly to apply ecological principles to direct practice. The life model, while going through several iterations over past decades (Gitterman, 2011), recognized the importance of a goodness of fit between people and their environments, which allows both reciprocally to respond and adapt to one another. The life model relies heavily on ecological concepts and metaphors to emphasize the transactions between people and environments. Human development is characterized by complex processes that emerge, change, and evolve over time. Persons and environments never stand in static relationship to each other. The life model conceives of problems in living as a result of stress associated with inadequate fit between people and their environments. These problems revolve around stressful life transitions, maladaptive interpersonal processes, and environments that become or remain unresponsive despite human intervention to modify and improve them.

Carol Meyer (1979, 1983) introduced a variant of Carel Germain's ecological perspective called the ecosystems perspective. Her work was an explicit attempt to find a unifying framework to guide practice, which combined elements of both systems and ecological theory. Carol Meyer conceived of her ecosystems perspective as a conceptual framework – a kind of metamodel – rather than a fully operationalized theoretical position. She suggested that the ecosystems perspective "offers social work practitioners/clinicians a way of thinking about and assessing the relatedness of people and their impinging environments: it does not specify the what (problem-definition) or the how (methodology) of practice" (Meyer, 1983: p. 29). Similar to Carel Germain's earlier ecological perspective, the ecosystem perspective drew the profession's attention not only to persons-in-environments but to the dynamic interactions of persons and environments. These interactions are constant and nonlinear and, from an ecosystems perspective, no simple cause and effect explanation can ever serve to understand the complexities of the human condition.

Conventional models of ecological social work have had profound staying power in the theoretical life of social work. They have shaped the emergence of what many have commonly come to call ecological social work (Greene, 2009; Hutchinson, 2011; Rogers, 2011). These models have become conventional wisdom for much of social work practice. Over the last several decades, ecological perspectives have been widely utilized in social work education and practice. Many social work practice and human behavior textbooks are grounded in various dimensions of ecological theory (Hutchinson, 2011; Rogers, 2011). One would be hard pressed to find a significant number of social workers who do not, at some level, ascribe to at least the basic tenets of ecological social work theory. The purpose of social work practice from a conventional ecological perspective is, in the main, to enhance the goodness-of-fit between individuals and their environments and by so doing increase personal and social capacities either to find resonance with, influence, or change environmental factors (van Wormer and Besthorn, 2011; Payne, 2005). The influence of conventional notions of ecological social work, both in North America and around the world, cannot be underestimated. Ecological social work, as conventionally conceived, ushered in a new integrative paradigm for a profession that had struggled for decades to conceptualize its obligations to both person and environment (Dominelli, 2012). As Katherine van Wormer and Fred Besthorn (2011: p. 19) note with respect to conventional ecological models of social work practice in the United States, "no other conceptual frame of reference since the introduction of Freudian psychology has had as significant an impact on mainstream social work thinking as have the ecological and systems formulations."

A New Ecological Social Work: An Alternative Ecological Paradigm of Nature

The profession's conventional ecological perspective has been relevant to social work in that it provided a comprehensive, multidisciplinary, and holistic framework within which the complex and interrelational elements of persons' lives could be connected and understood. Conventional ecological social work ushered in the first meaningful attempt to integrate the psychological and sociological dimensions of social work practice and supported a conceptual "shift in social work practice from a static to a dynamic view of the environment" (Kemp et al., 1997: p. 41). In reality, social work had always, to a greater or lesser extent, been concerned with situational determinants of human functioning. But, until the evolution of the ecological perspective the profession found it easier, and often more expedient, to address personal functioning. The environment seemed just too amorphous and the task of changing it seemed too daunting.

While social work spoke the ecological and systems language of environments in interaction, in reality the focus was mainly on individual behavior in static environments. This severely limited the profession's capacity to engage critically with those structural barriers marginalizing oppressed groups. Additionally, to a new generation of social workers, it hindered a response to deteriorating natural environments and their catastrophic impact on clients' lives. Indeed, one of the persistent criticisms of social work's conventional ecological perspective has been its rather narrow interpretation of the environmental construct. As early as the 1960s, some social work scholars had become critical of the profession's conceptualization of environment (Stein, 1963). In particular, Richard Stein identified three major shortcomings associated with the concept. First, he noted the tendency to regard the environment as very narrowly related to immediate contexts such as housing, jobs, communities, families, social relationships, and face-to-face interactions. Second, he identified the inclination to view environment as external and only peripherally related to the individual who "stands alone, an isolated complex of intra-psychic processes" (Stein, 1963: p. 68). Third, Stein noted the tendency to view the environment as backdrop – static, unchanging, background clutter. Richard Stein argued that social work's efforts with respect to and on behalf of the

environment was truncated and merely a second-rate activity compared to the more important skill of psychosocial treatment. Thus, despite the rhetoric of persons and environment in dynamic interaction, the environment was still interpreted as a relatively benign, circumstantial space that persons either adapted to or learned to negotiate. Consequently, ecological social work in practice fell back upon the profession's historic tendency to emphasize heavily individual function in an environment that was generally resistant to change and was likely to impinge upon optimal coping capacity.

Criticisms that social work's conventional ecological perspectives and environmental definitions unnecessarily restricted the definition of the environmental construct were relatively new to social work's scholarly discourse. Several social work scholars (Weick, 1981; Resnick and Jaffee, 1982; Soine, 1987; Gutheil, 1992; Saari, 1992; Berger and Kelly, 1993; Hoff and Polack, 1993; Rogge, 1993) began to argue that the core values of social work and its conventional ecological models must be altered to support the new realities of the importance of the natural environment and the impending environmental crisis. They reiterated what Richard Stein (1963) noted some 30 years earlier, that social work spoke the ecological language of environments in interaction but, in reality, the focus was mainly on individual behavior in static, narrowly construed environments. However, they also noted that conventional ecological models seriously inhibited conceiving the natural environment as anything more significant than data to be studied or resources to be procured or regulated. Social work's uncritical adoption of the adaptive, static, and narrowly defined notions associated with conventional ecological thinking had unwittingly conspired in creating a state of collective consciousness suggesting that a person exists on a plane of deep division between oneself and one's place in the larger natural world. Person-in-environment from a conventional ecological perspective had become a kind of euphemism for what was, in reality, persons living in an environment that was wholly separate and distinct from them. Deeper understandings of identity with nature is then excluded and becomes an abstraction that leads to concepts and actions that cannot be reconciled with either the health of persons or of nature. Living in nature, on the other hand, suggests a complex relationship melded together over time through patterns of value, memory, familiarity, love, and respect.

By suggesting that an alternative ecological model of social work must also consider the natural world, these scholars were also suggesting something quite new to social work. Their critique was not just about extending conventional environmental definitions to include the natural world but, in many ways, to advocate for a radical new approach in how social work conceived itself and how it went about its practice priorities. For a small but growing number of social work scholars, especially in the Western world, the recognition began to emerge that the natural environment shares a complex and evolutionary link to personal, social, cultural, and transpersonal developments. They began addressing the implications of nature's degradation, especially on poor and vulnerable populations, while at the same time exploring the natural world's ethical, aesthetic, mythic, and therapeutic influences and capacities to inform social work theory and practice. For these scholars, social work could no longer afford to view the natural environment as a backdrop to other matters but as the centerpiece for the core values, purposes, and functions of social work. Human social development and human suffering could not be fully understood or effectively addressed until the profession recognized that the human world was integrally related to, dependent upon, and emergent from the natural world.

Thus, in the last 15 years, a new generation of social work theorists and practitioners extended the original parameters of conventional ecological social work to include a broader definition and more profound understanding of the interaction between person and environment. Due in large measure to the emergence in the 1980s and 1990s of the worldwide environmental movement, a new form of ecological social work began to take shape. With the release of Maria Hoff and John McNutt's (1994) edited anthology, social work had its first well-organized and thoroughly reasoned statement of principle, premised on the ideal that ecologically minded social work could not fully realize its stated commitment to person-in-environment until it seriously considered the inseparable link between human well-being and the well-being of the planet. Maria Hoff and John McNutt (1994) argued that earth's ecological systems were not static but, rather, consisted of several interconnected layers including the geosphere, hydrosphere, atmosphere, biosphere, and noosphere. The first several of these refer to facets of the soil, water, air, and biological systems, which support or impinge on human well-being. The noosphere, on the other hand, represents a deeper, atavistic, and perhaps even a genetically predisposed connectedness with nature that lies beyond physical dependence but that is nonetheless absolutely indispensable for human development and survival (Besthorn and Saleebey, 2003). Shortly thereafter, another wave of social work scholars including John Coates and Marilyn McKay (1995), Mary Rogge (1996), and Fred Besthorn (1997) were among the first social workers in North America to begin a thorough critical analysis of the theoretical and practical linkages between planetary consciousness and social work.

Since then, there has been a flowering of published works on the interrelationship between the natural environment and the theory and practice of social work (Global Alliance for a Deep Ecological Social Work, 2012). A growing core of social workers from North America, Europe, and Australia began to speak compellingly concerning the importance of incorporating the natural environment into the profession's theoretical formulations and practice modalities (Besthorn, 2001, 2002, 2008, 2012, 2013; Besthorn and McMillen, 2002; Besthorn and Meyer, 2010; Borrell et al., 2010; Coates, 2003, 2005; Coates et al., 2006; Dominelli, 2012; Gray et al., 2012; Hawkins, 2010; Jones, 2006, 2010; Lysack, 2007, 2010; Mary, 2008; Matthies, 1987; Matthies et al., 2001; McKinnon, 2005, 2008; Molyneux, 2010; Muldoon, 2006; Narhi, 2004; Peeters, 2011, 2012; Power and Freedman, 2012; Rogge, 2008; Rogge and Combs-Orme, 2003; Rotabi, 2007; Ryan, 2011; Shaw, 2006, 2008; Ungar,

2002, 2003; West, 2007; Zapf, 2008, 2009). In an attempt to differentiate it from social work's conventional ecological frameworks, this emerging alternative ecological perspective of social work has been referred to in several different ways (e.g., environmental social work, ecosocial work, deep-ecological social work, ecospiritual social work, and green social work). For its advocates, this alternative ecological perspective better positions a profession, often conspicuously absent, in the emerging international consensus that Earth's ecosystems and its capacity to support life are in deep trouble. Until the profession began to take seriously the importance of the natural world in its understanding of social work, the profession could not fully realize its core commitment to person-and-environment.

A New Ecological Social Work: Major Scholarly Themes

It would be difficult to fully catalog the many differing thematic areas of research interest that scholars and practitioners operating from a new paradigm of ecological social work have addressed over the last decade. It would also be beyond the scope of this article to describe in detail the key thematic areas that have contributed to the discourse of a new ecological social work. Social workers committed to incorporating nature and deeper ecological thinking into practice will likely have varied and unique interests. However, a cursory review of the recent new ecological social work literature does suggest a number of overarching themes. In general, these include discussions of

1. the comparative importance of incorporating nature into the practice of social work
2. social work's historical attempts at integrating ecological theory into its theoretical and practice frames as well as critiquing social work's conventional person-in-environment perspective
3. the ontological, axiological, and epistemological basis upon which a new ecological social work might be supported
4. the interface between economic development, ecological sustainability, economic globalization and strengthening social capital as a catalyst for creating sustainable livelihoods and community security
5. the importance of grassroots initiatives to spur involvement of communities in the creation of ecological awareness, sustainable development, and food security
6. the interface of fossil fuel dependence and the carbon-based economy and its deleterious impact on energy conservation and sustainability, especially with low-income and marginalized persons and communities
7. how extractive mining practices and industrial damage carried out by transnational corporations impact small communities and dependent workforces
8. the importance of crisis intervention in disaster and traumatic stress events including natural and human-induced disasters
9. social work interventive strategies and efforts in drought-stricken areas of the world
10. the importance of social work's involvement in grassroots activism especially with respect to climate change and the increasing severity of climatic events
11. social work responses to toxic waste exposure in low-income communities
12. social work responses to the growing crisis of 'environmentally displaced persons' – sometimes inaccurately referred to as environmental refugees
13. the application of deeper ecological thinking to therapeutic practice with individuals, groups, families, and communities
14. the interface between social and ecological justice and the importance of shifting from an anthropocentric to an ecocentric basis for determining just priorities and actions
15. the application of principles of a new ecological paradigm to social work education and curriculum development
16. the importance of honoring indigenous ways of knowing relative to the natural world and how these might help inform the evolution of a new era of best practice in social work
17. social worker's involvement in the utilization of animal-assisted activities to enhance the quality of life of individuals and groups
18. social worker's involvement in the utilization of nature-assisted activities and programs such as Outward Bound, Plant-People Council, and Project Adventure to enhance life quality for individuals and groups
19. the interrelationship between spirituality, religion, and deeper ecological consciousness and ways to incorporate transpersonal ecological ideas into social work theory and practice.

A New Ecological Social Work: Core Operational Values

Social work has often been described as a value-laden profession. Many major social work institutional mission declarations describe core values that establish a framework for effective social work practice. For example, in the United States, the National Association of Social Work (NASW) identifies six core values that are thought critical to the practice of effective social work. Unfortunately, the same cannot be said for the emergence of a new ecological social work model. With the exception of a small number of general statements of principle by several national and international social work associations (e.g., NASW, International Federation of Social Workers, and International Association of Schools of Social Work); few, if any, explicit statements exist that attempt to create a compendium to core values that many, if not most, environmental social workers might identify with. Like the diversity that exists in the thematic areas of scholarly interest, the assortment of opinion of what constitutes the core values of a new ecological perspective is quite multifaceted. This is not to say that there are no statements of principle of core values that many social work scholars have propounded as important to the emergence of a new ecological perspective. The literature is replete with

ample discourse addressing a multitude of values and principles thought by their respective authors to be critical to advancing and sustaining a new ecological perspective (Coates and Gray, 2012; Dominelli, 2012). The issue has always been in systematizing and categorizing these ideas into some accessible and reasonably clear framework. The following is a brief listing, from this author's perspective, of those core values common to a new ecological model of social work. The shift toward the new ecological paradigm is a gradual process that has been occurring as more people begin to think systemically and, no doubt these are likely to change as new scholarly and practice agendas mature.

1. The belief that it is impossible to fully comprehend the complexity of ecosystems and the human condition in conventional linear, atomistic, and mechanistic terms. Intuitive ways of knowing, nonlinear thinking, and emotional intelligence are seen as valid routes to knowledge along with rationality and experimentation
2. an understanding that the conceptualizations of environment includes all beings and systems in the cosmos – human and nonhuman – as well as the idea that humans can be aware of, and connected to, all these levels of being; environment is far more than the social milieu
3. the importance of affirming the need to live in harmony with nature, each other, and the entire planetary ecosystem
4. the importance of seeing humankind as intimately connected (biologically, genetically, aesthetically, economically, and historically) with the natural world, as, indeed, natural beings themselves who are inseparable from the unity of all things and beings
5. the importance of ascertaining and applying the wisdom of nature and with seeing nature as fecund source of intelligence that has guided and must still guide humanity's understanding of itself and the interconnected nature of interactions among and between all organisms
6. the importance of developing an ecocritical analysis of Western modernity's deeply entrenched social, political, ideological, and economic assumptions that have led to a myopic preoccupation with unrelenting progress, unrestrained development, unending consumption, and unremitting despoliation of the earth's carrying capacity
7. the necessity of adhering to the idea of long-term sustainability that sees the earth as a balanced system, that requires humans to be intentional about preserving and maintaining the necessary conditions for sustaining quality life for both humans and the nonhuman world
8. the belief that the shift to a new ecological paradigm requires nothing short of a radical and wholesale transformation of global human consciousness away from its anthropocentric bias toward an ecocentric orientation.

A New Ecological Social Work: Future Prospects

A new ecological model of social work is, in general terms, predicated on the assumption that there is much that can and must be done to address looming and potentially catastrophic environmental problems. Social work has opportunities to contribute to a transformed world. Some of those opportunities have already begun with the ideas, findings, and insights that social work researchers, scholars, and practitioners have obtained, which have been alluded to in the earlier sections of this article. But, the future of a new ecological model of social work depends on professionals overcoming the general social tendency to minimize the seriousness of environmental concerns while convincing oneself that social work professionals have a role to play in addressing these issues. Involving our professional lives in these concerns and crafting our theoretical and practice models in comprehensively earth-centered terms can be politically and professionally risky for social workers. The fear for some is that too close an identification with highly charged social causes runs the risk of social work losing its credibility as objective investigators and socially responsible change agents. And yet, when has this not been a risk for social work? As a profession, social work has routinely identified with risky, politically charged issues for the improvement of individual clients, marginalized groups, and the larger social order, generally. The difference is that which is currently marginalized is the very natural, planetary system that all depend upon for survival. Inattention and hoping for the best can no longer be justified. At the heart of the question for social work is how it participates in and encourages those transformations necessary to protect fragile environments, reverse ecosystem degradation, and once again live in unity with other living things. A new ecological model of social work holds great future promise to bring the profession closer in line with the emerging global momentum toward ecological and economic sustainability, full democratic participation, unadorned sufficiency, and a post-anthropocentric solidarity with the natural world.

A new ecological model of social work presents a thorough organicist worldview with respect to both people and environments as living systems in dynamic interaction with each other. The primary importance of this alternative ecological perspective is that it can be incorporated into many levels of practice. Whether at the level of assessment, individual, community, organizational, or political practice, a new ecological perspective provides an intelligible framework for viewing person and environment interaction as well as the prospect for better outcomes associated with comprehensive multidimensional interventions. Social workers could, if they wished, enter into this organic milieu and pursue interventive strategies that either help change or remove environmental obstacles to enhanced personal growth and adaptive capacity.

See also: Green Social Work; Natural Disasters: Health-Related Aspects.

Bibliography

Berger, R., Kelly, J., 1993. Social work in the ecological crisis. Social Work 38 (5), 521–526.

Besthorn, F.H., 1997. Reconceptualizing Social Work's Person-in-environment Perspective Explorations in Radical Environmental Thought (Doctoral Dissertation). University of Kansas, Ann Arbor, MI. UMI Microfilm 981157.

Besthorn, F.H., 2001. Transpersonal psychology and deep ecological philosophy: exploring linkages and applications for social work. In: Canda, E.R., Smith, E.D. (Eds.), Transpersonal Perspectives on Spirituality in Social Work. Haworth Press, Binghamton, NY, pp. 23–44.

Besthorn, F.H., 2002. Radical environmentalism and the ecological self: rethinking the concept of self-identity for social work practice. Journal of Progressive Human Services 13 (1), 53–72.

Besthorn, F.H., 2008. Environment and Social Work Practice. In: Encyclopedia of Social Work, twentieth ed., vol. 2. Oxford University Press, New York, pp. 132–136.

Besthorn, F.H., Meyer, E., 2010. Environmentally displaced persons: broadening social work's helping imperative. Journal of Critical Social Work 11 (3), 75–84.

Besthorn, F.H., 2012. Deep ecology's contribution to social work: a ten-year retrospective. International Journal of Social Welfare 21 (3), 248–259.

Besthorn, F.H., 2013. Vertical farming: social work and sustainable urban agriculture in an age of global food crisis. Australian Journal of Social Work 66 (2), 187–203.

Besthorn, F.H., McMillen, D., 2002. The oppression of women and nature: ecofeminism as a framework for an expanded ecological social work. Families in Society: The Journal of Contemporary Human Services 83 (3), 221–232.

Besthorn, F.H., Saleebey, D., 2003. Nature, genetics and the biophilia connection: exploring linkages with social work values and practice. Advances in Social Work 4 (1), 1–18.

Borrell, J., Lane, S., Fraser, S., 2010. Integrating environmental issues into social work practice: lessons learnt from domestic energy auditing. Australian Social Work 63 (3), 315–328.

Coates, J., 2003. Ecology and Social Work: Toward a New Paradigm. Fernwood, Halifax.

Coates, J., 2005. The environmental crisis: implications for social work. Journal of Progressive Human Services 16 (1), 25–49.

Coates, J., Gray, M., 2012. The environment and social work: an overview and introduction. International Journal of Social Welfare 21 (3), 230–238.

Coates, J., Gray, M., Hetherington, T., 2006. An 'ecospiritual' perspective: finally, a place for indigenous approaches. The British Journal of Social Work 36 (3), 381–399.

Coates, J., McKay, M., 1995. Toward a new pedagogy for social transformation. Journal of Progressive Human Services 6 (1), 27–43.

Dominelli, L., 2012. Green Social Work: From Environmental Crises to Environmental Justice. Polity Press, Malden, MA.

Germain, C., 1968. Social study: past and future. Social Casework 49, 403–409.

Germain, C., 1973. An ecological perspective in casework practice. Social Casework 57 (7), 419–426.

Germain, C., 1978. Space: an ecological variable in social work practice. Social Casework 59, 515–522.

Germain, C.B., Gitterman, A., 1980. The life model of social work practice. Columbia University Press, New York.

Gitterman, A., 2011. Advances in the life model of social work practice. In: Turner, F. (Ed.), Social Work Treatment: Interlocking Theoretical Approaches, fifth ed. Oxford University Press, New York, pp. 279–292.

Global Alliance for a Deep Ecological Social Work, 2012. Social Work and Environment Bibliography. Retrieved from: http://www.ecosocialwork.org/index.php?option=com_content&view=article&id=59&Itemid=66 (accessed 06.02.13.).

Gray, M., Coates, J., Hetherington, T., 2012. Environmental Social Work. Routledge, London.

Greene, R., 2009. Ecological perspective. In: Greene, R.G. (Ed.), Human Behavior Theory and Social Work Practice, third ed. Aldine de Gruyter, New York, pp. 199–236.

Gutheil, I., 1992. Considering the physical environment: an essential component of good practice. Social Work 37 (5), 391–396.

Hawkins, C., 2010. Sustainability, human rights, and environmental justice: critical connections for contemporary social work. Critical Social Work 11 (3), 68–81.

Hoff, M., Polack, R., 1993. Social dimensions of the environmental crisis: challenges for social work. Social Work 38 (2), 204–211.

Hoff, M., McNutt, J. (Eds.), 1994. The Global Environmental Crisis: Implications for Social Welfare and Social Work. Ashgate Publishing, Brookfield, VT.

Hutchinson, E., 2011. Dimensions of Human Behavior: Person and Environment, fourth ed. Sage Publications, Los Angeles.

Jones, P., 2006. Considering the environment in social work education: transformations for eco-social justice. Australian Journal of Adult Learning 46 (3), 364–382.

Jones, P., 2010. Responding to the ecological crisis: transformative pathways for social work. Journal of Social Work Education 46 (1), 67–84.

Kemp, S., Whittaker, J., Tracy, E., 1997. Person-Environment Practice: The Social Ecology of Interpersonal Helping. Aldine de Gruyter, New York.

Lysack, M., June 2007. Family therapy, the ecological self and global warming. Context 9 (1), 9–11.

Lysack, M., 2010. Environmental decline, loss, and biophilia: fostering commitment in environmental citizenship. Critical Social Work 11 (3), 48–66.

Mary, N., 2008. Social Work in a Sustainable World. Lyceum Books, Chicago.

Matthies, A., 1987. Ekologinen sosiaalityo (Ecological social work). Sosiaaliviesti 2 (1), 32–37.

Matthies, A., Narhi, K., Ward, D. (Eds.), 2001. The Eco-social Approach in Social Work. Sophi, Jyvaskyla, Finland.

McKinnon, J., 2005. Social work, sustainability, and the environment. In: Alston, M., McKinnon, J. (Eds.), Social Work: Fields of Practice, second ed. Oxford University Press, Melbourne, pp. 225–236.

McKinnon, J., 2008. Exploring the nexus between social work and the environment. Australian Social Work 61 (3), 256–268.

Meyer, C., 1979. What directions for social work? Social Work 24, 267–272.

Meyer, C. (Ed.), 1983. Clinical Social Work in the Eco-systems Perspective. Columbia University Press, New York.

Molyneux, R., 2010. The practical realities of ecosocial work: a review of the literature. Critical Social Work 11 (2), 61–69.

Muldoon, A., 2006. Environmental efforts: the next challenge for social work. Critical Social Work 7 (2), 84–93.

Narhi, K., 2004. The Eco-social Approach in Social Work and the Challenges to the Expertise of Social Work. University of Jyvaskyla, Jyvaskyla, Finland. Retrieved from: http://dissertations.jyu.fi/studeduc/9513918343.pdf (accessed 21.05.13.).

Payne, M., 2005. Modern Social Work Theory, third ed. Lyceum, Chicago.

Peeters, J., 2011. Social sustainable development and social work. In: Agten, J., Gillet, F. (Eds.), Social Action in Europe: Sustainable Social Development and Economic Challenges, Conference Book, Brussels, 10-13 April, p. 14. Abstract retrieved from http://www.ensact.org/index/pages/id_page38/lang-en/.

Peeters, J., 2012. The place of social work in sustainable development: towards ecosocial practice. International Journal of Social Welfare 21 (3), 287–298.

Power, M., Freedman, D., 2012. Applying a social justice framework to photovoice research on environmental issues: a comprehensive literature review. Critical Social Work 13 (2), 81–100.

Resnick, H., Jaffee, B., 1982. The physical environment and social welfare. Social Casework 63 (6), 354–362.

Rogers, A., 2011. Human Behavior in the Social Environment. McGraw Hill, Boston.

Rogge, M., 1993. Social work, disenfranchised communities, and the natural environment: field education opportunities. Journal of Social Work Education 29 (1), 111–120.

Rogge, M., 1996. Social vulnerability to toxic risk. Journal of Social Service Research 22 (1/2), 109–129.

Rogge, M., 2008. Environmental justice. In: National Association of Social Workers (NASW) (Ed.), Encyclopedia of Social Work. Oxford University Press, New York, pp. 136–139.

Rogge, M., Combs-Orme, T., 2003. Protecting our future: children, environmental policy, and social work. Social Work 48 (4), 439–450.

Rotabi, K., 2007. Ecological theory origin from natural to social science or vice versa? A brief conceptual history for social work. Advances in Social Work 8 (1), 113–129.

Ryan, T., 2011. Animals and Social Work: A Moral Introduction. Palgrave MacMillan, Hampshire, England.

Saari, C., 1992. The person-in-environment reconsidered: new theoretical bridges. Child and Adolescent Social Work Journal 9 (3), 205–219.

Shaw, T., 2006. Social Workers Knowledge and Attitude toward the Ecological Environment (Unpublished doctoral dissertation). University of California at Berkeley, Berkeley.

Shaw, T., 2008. An ecological contribution to social welfare theory. Social Development Issues 30 (3), 13–26.

Soine, L., 1987. Expanding the environment in social work: the case for including environmental hazards content. Journal of Social Work Education 23 (2), 40–46.

Stein, R., 1963. The concept of social environment in social work practice. In: Parad, H.J., Miller, R.R. (Eds.), Ego-oriented Casework: Problems and Perspectives. Famil Service Association of America, New York, pp. 65–88.

Ungar, M., 2002. A deeper, more social ecological social work practice. Social Service Review 76 (3), 480–497.

Ungar, M., 2003. The professional social ecologist: social work redefined. Canadian Social Work Review 20 (1), 5–23.

van Wormer, K., Besthorn, F.H., 2011. Human Behavior and the Social Environment: Groups, Communities, and Organizations, second ed. Oxford University Press, New York.

Weick, A., 1981. Reframing the person-in-environment perspective. Social Work 26 (2), 140–143.

West, D., 2007. Building a holistic environmental model for global social work. The International Journal for Interdisciplinary Social Sciences 2 (4), 61–65.
Zapf, M., 2008. Transforming social work's understanding of person and environment: spirituality and the "common ground". Journal of Religion & Spirituality in Social Work 27 (1–2), 171–181.
Zapf, M., 2009. Social Work and the Environment: Understanding People and Place. Canadian Scholars Press, Toronto.

Relevant Websites

http://www.cep.unt.edu – Center for Environmental Philosophy-University of North Texas (accessed 02.08.12.).
http://www.ens-newswire.com – Environmental News Network (accessed 02.07.12.).
http://www.deepecology.org – Foundation for Deep Ecology (accessed 20.05.12.).
http://www.ipcc.ch – Intergovernmental Panel on Climate Change (accessed 09.07.12.).
http://www.ifsw.org – International Federation of Social Workers (accessed 12.07.12.).
http://designtrek.net – John Scherch Social Work and Sustainable Living Web site (accessed 15.07.12.).
http://www.liebertpub.com – Journal of Ecopsychology (accessed 04.06.12.).
http://web.utk.edu/∼merogge – Mary Rogge University of Tennessee Home Page (accessed 09.09.12.).
http://nrdc.org – National Resources Defense Council (accessed 15.07.12.).
http://www.ucsusa.org – Union of Concerned Scientists (accessed 15.06.12.).
http://www.unep.org – United Nations Environmental Program (accessed 03.08.12.).
http://unfccc.int – United Nations Framework Convention on Climate Change (accessed 07.09.12.).
http://www.epa.gov – United States Environmental Protection Agency (accessed 25.05.12.).
http://www.socialworkers.org – USA-National Association of Social Workers (accessed 12.07.12.).
http://www.worldwatch.org – Worldwatch Institute (accessed 12.06.12.).

Ecology and Health

Margaret Schneider and Daniel Stokols, University of California Irvine, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Stokols, volume 6, pp. 4030–4035, © 2001, Elsevier Ltd.

Abstract

Ecology pertains broadly to the relationships between organisms and their environments. The field of ecology and health examines the influence of conditions in the physical and sociocultural environment on the physiological, social, and emotional well-being of individuals and groups. The present article traces the development of research on ecology and health, beginning with the emergence of the ecological paradigm in the biological sciences and its subsequent development in other disciplines, including sociology, psychology, and public health. Recent interpretations of ecology as it applies to human health emphasize a perspective that acknowledges humans as an integral and interdependent part of the larger global ecosystem.

Ecology and Health

The term 'ecology' pertains broadly to the interrelations between organisms and their environments (Hawley, 1950). From its early roots in biology, the ecological paradigm has evolved within several disciplines (e.g., sociology, psychology, economics, and public health) to provide a general framework for understanding people's transactions with their physical and sociocultural surroundings (e.g., Barker, 1968; Bronfenbrenner, 1979; Park et al., 1925). The field of ecology and health examines the direct and interactive effects of conditions in the physical and sociocultural environment on the physiological, emotional, and social well-being of individuals and groups (cf Duhl, 1996; Moos, 1979; Stokols, 1996). Ecological analyses provide more integrative and comprehensive explanations for the joint influence of biogenetic, psychological, behavioral, sociocultural, and physical environmental factors on human well-being, than do narrowly gauged biomedical, behavioral, and environmental theories.

Emergence of Ecological Perspectives on Health and Illness

The emergence of an ecological paradigm for understanding the determinants of health and illness in human populations can be traced to research developments in several scientific fields. During the 1920s and 1930s, a series of correlational analyses were conducted by researchers in the Chicago School of Human Ecology, revealing systematic associations between community rates of illness and crime on the one hand and physical and social conditions of the urban environment (including poor-quality housing and high population densities within residences and neighborhoods) on the other hand (cf Faris and Dunham, 1939; Park et al., 1925). These correlational studies in the fields of human ecology and sociology suggested a strong link between conditions of the urban environment and residents' health, but did not provide an integrative theoretical account of the causal pathways connecting those variables.

An important theoretical analysis of the links between the social environment and illness etiology was provided by John Cassel in the field of public health. Cassel's (1964) conceptualization of social epidemiology signaled a shift from 'germ-theory' accounts of health and illness (focusing on the health-impairing effects of specific pathogens once they invaded a human 'host'), toward a more ecologically oriented, contextual perspective emphasizing the effects of one's social environment (e.g., work overload, sources of interpersonal stress) on immune functioning and host resistance to pathogenic agents. Cassel's formulation of social epidemiology during the 1960s and his hypotheses concerning the impact of social and cultural factors on human health have been corroborated empirically in subsequent studies linking job strain, social isolation, and interpersonal stress to a variety of illness conditions and premature death (cf Berkman and Syme, 1979; Karasek and Theorell, 1990; Marmot et al., 1997).

Incorporation of the ecological perspective into areas of health promotion initially perceived as rooted in individual behavior and therefore best targeted by individual-focused behavior change strategies has led to an ecological approach to public health. Publication of the US Surgeon General's Reports in 1964 and 1979, documenting the deleterious effects of smoking on health, prompted the rapid growth of health promotion and disease prevention research during the 1980s (cf US Public Health Service, 1979). Much of the early research in this field was rooted in the concepts and methods of health education, and focused largely on behavior change strategies aimed at fostering healthier lifestyles and health habits among individuals (cf Belloc and Breslow, 1972). As appreciation for the salient role played by societal and environmental influences grew, efforts to encourage health-promoting behavior took a decidedly ecological turn, with programs such as ASSIST, an ambitious government-funded demonstration project undertaken in the 1990s to help states develop effective strategies to reduce smoking. Focusing on policy change, the goal of ASSIST was to alter states' social, cultural, economic, and environmental factors that promote smoking. Results of the evaluation of ASSIST supported the ecological hypothesis that policy-level changes can be effective for promoting changes in individual health behavior (Stillman et al., 2003).

Such targeted efforts to enhance individuals' health behaviors via an ecological approach have led to reduced use of tobacco products among adults, lower rates of alcohol-related automobile deaths, and reduced employee health costs at the

worksite (cf McGinnis and Lee, 1995; Pelletier, 1996; Schneider and Stokols, 2008). At the same time, however, the ecological approach to understanding health and disease has led to increased awareness of health risks that are strongly influenced by conditions outside of individual control, such as exposure to community violence, lack of medical insurance, environmental injustice, poverty, and unemployment. Critically, many of these health risks remain segmented in hard-to-reach 'pockets of heightened prevalence' (Fisher, 1995) – especially among low-income and minority groups in the population (cf Adler and Stewart, 2013; Bullard, 2005).

Growing recognition of the limitations associated with behavior change programs has prompted a paradigm shift in the health promotion field, away from individually focused interventions toward more comprehensive, ecological formulations that address interdependencies among socioeconomic, political, environmental, organizational, biological, and psychological determinants of health and illness (cf Kickbusch, 1989; Schneider and Stokols, 2000; Winett et al., 1989). These strategies include increased resources devoted to community policing as a method of preventing neighborhood violence (MacDonald, 2002), diversion of convicted drug users to treatment programs (Bull, 2005), and the Affordable Care Act, which utilizes a legislative approach to persuading individuals to purchase health insurance (Lavarreda et al., 2011). A common feature of these programs is that they incorporate multiple levels of analysis and intervention components – including political action, legislative reform, media advocacy, as well as behavioral change and environmental enhancement efforts – to address key contextual factors that substantially influence individual and population health.

Whereas the ecological approach to health has been informed and influenced by the principles of community health/community psychology, the former extends beyond the latter on both ends of the continuum from more proximal factors (e.g., biologic, genetic, and other intrapersonal factors) to more distal factors (e.g., global or national conditions). Community health, in which formal or informal social groupings form the unit of analysis or of intervention, may be viewed as a particular application of the ecological paradigm, which represents a "unique way of looking at the world" (Kelly, 2006). According to the ecological approach, health is "viewed as the consequence of reciprocal causation unfolding at multiple individual and environmental levels of influence" including, but not limited to, the community (cf Richard et al., 2012).

One other major development that has prompted greater interest in ecology and health in recent years is the discovery of adverse global environmental changes – for instance, the depletion of atmospheric ozone and the trend toward climate change – which have the capacity to elevate population morbidity and mortality rates at worldwide as well as regional and national levels (cf Folke, 2006; Stern et al., 1992; Stokols et al., 2009). Potential health impacts of these global changes include elevated skin cancer rates resulting from ozone depletion and individuals' heightened exposure to ultraviolet rays, and the food shortages, nutritional ailments, and respiratory illnesses associated with climate change (cf Leaf, 1989). The health consequences of extreme weather and disasters, expected to increase in both frequency and severity as the climate shifts (IPCC, 2012), are becoming more evident as well. The discovery of these global environmental threats to population health has prompted greater international collaboration in the development of ecologically oriented health promotion programs through programs such as the World Health Organization (WHO) Healthy Cities Program (cf Ritsatakis, 2012) and the WHO work plan on climate change and health.

Core Principles of Ecology and Health Research

Research on ecology and health encompasses several disciplines, including medicine, public health, urban planning, environmental design, public policy, and the behavioral, social, and global ecology sciences. The scientific contours of this research are not easily delimited, owing to the interdisciplinary scope of the field (cf Stokols, 2000). The unique concerns of ecology and health can be better understood in terms of the overarching conceptual principles that underlie this field, rather than by searching for a clearly defined body of research organized around this topic. The core principles underlying ecology and health research are drawn largely from programmatic statements about the 'New Public Health' and the conceptual and methodological assumptions of systems theory and social ecology (cf Duhl, 1996).

Intellectual Antecedents: Systems Theory and the Ecological Paradigm

The 'New Public Health' outlined in the Ottawa Charter for Health Promotion (1986) called for the development of comprehensive, ecologically oriented approaches to the study and prevention of disease (cf Kickbusch, 1989). For instance, the Ottawa Charter gave explicit attention to social causes of illness, as well as the physical environmental health threats that exist in many communities. Subsequent analyses of ecology and health extended these programmatic ideas by providing a set of conceptual and methodological principles derived largely from systems and ecological theories, for organizing and evaluating community-based health promotion programs (cf Green and Ottoson, 1999; Sallis et al., 2006).

The field of ecology and health incorporates a number of concepts derived from systems theory (e.g., interdependence, homeostasis, negative feedback, deviation amplification; cf Emery and Trist, 1972; Maruyama, 1963) to understand the interrelations among people and their environments. Systems analyses suggest that the healthfulness of particular settings and the well-being of their participants are jointly influenced by multiple facets of the physical environment (e.g., geography, architecture, technology) and the social environment (e.g., culture, economics, politics). The health status of individuals and groups is further influenced by personal attributes such as genetic heritage, psychological dispositions, and behavioral patterns. From the vantage point of systems theory, efforts to promote well-being should be based on an understanding of the relationships among diverse environmental and personal factors, rather than on analyses that focus exclusively on environmental, biological, or behavioral factors (cf Moos, 1979).

Research Findings from Ecological Analyses of Health

The direct and interactive effects of environmental and personal attributes on health have been observed in several lines of research. For example, a substantial body of research highlights the ways in which poverty, inequality of income distribution, and minority status jointly undermine the health of individuals and vulnerable subgroups in the population (cf Adler and Stewart, 2013; Bullard, 2005; Kaplan et al., 1996). Other studies indicate that neighborhood socioeconomic status (percent of households in a geographic area receiving public assistance) is associated with poorer self-reports of health, above and beyond individual socioeconomic status (Malmstrom et al., 1999; Robert, 1998). Interactive effects of environmental and behavioral risk factors are exemplified by the findings that environmental exposure to asbestos exacerbates chronic smokers' risks of developing lung cancer (Grunberg, 1991), and that an individual's disposition toward exerting internal control over his/her environment buffers the potentially negative effects of low education and income on health status (Lachman and Weaver, 1998).

Systems theory also posits that people–environment transactions are characterized by cycles of mutual influence, whereby the physical and social features of settings directly affect occupants' health and, concurrently, the participants in settings modify the healthfulness of their surroundings through their individual and collective actions (or, alternatively, adopt an increasingly passive stance toward their environment). For example, individuals' routine exposure to community violence in some low-income neighborhoods can provoke feelings of disempowerment, helplessness, and depression, which in turn diminish their efforts to adopt health-promotive practices related to physical activity and dietary change (cf Sanders-Phillips, 2000). In contrast, local residents' struggles with asthma exacerbated by air pollution can motivate them to engage in community activism targeting policies that impact air quality (Garzón et al., 2012). On a societal level, economic forces have stimulated an accelerated process of urbanization and commensurate degradation of the natural environment – a change that may be associated with poorer health status and that some communities attempt to counteract with urban planning that preserves greenspace for parks (Maller et al., 2008). Accordingly, ecological analyses of health should take into account that persons and their environments are in a constant state of mutual adaptation.

Ecologically oriented studies of health also incorporate certain conceptual and methodological principles drawn from the field of social ecology (Stokols et al., 2013). For instance, ecological theory highlights the multidimensional nature of human environments. Environmental settings can be characterized in terms of their physical and social components, their objective (actual) or subjective (perceived) qualities, and their scale or immediacy to individuals and groups (proximal vs distal). Moreover, the participants in environmental settings can be studied at varying levels ranging from individuals, small groups, and organizations, to larger communities and populations. Rather than focusing solely on individuals or aggregates, the social ecological perspective incorporates multiple levels of analysis and diverse methodologies (e.g., medical exams, questionnaires, behavioral observations, environmental recordings, epidemiologic analyses) for assessing the healthfulness of settings and the well-being of individuals and groups.

Social ecological analyses conceptualize human environments as complex systems in which local settings and organizations are nested within more complex and remote regions. Accordingly, efforts to promote human well-being must take into account the interdependencies that exist among immediate and more distant environments. The occupational health and safety of workplaces at the local level, for example, are directly influenced by state and national ordinances aimed at protecting environmental quality and public health (cf Stokols, 1992).

Another core theme of ecological theory is that physiological, emotional, and social well-being are influenced by the multiple, interrelated settings that comprise the daily activity systems of individuals and groups (cf Michelson, 1985). Bronfenbrenner's (1979) ecological model of human development highlights the joint influence of people's families, school, work, and recreational settings on their health and developmental outcomes. Rather than focusing exclusively on person–environment transactions that occur within single settings (microsystems, such as home or work environment), Bronfenbrenner's research emphasizes the ways in which functional linkages between two or more settings (i.e., the mesosystem, the exosystem, and the macrosystem) influence developmental processes and outcomes. Mesosystem units are comprised of two or more environmental settings (e.g., family and occupational environments) in which the individual is directly involved (e.g., a parent employed outside the home). Exosystems include an environment in which the developing person is directly involved (e.g., a child in his or her home environment) and one or more other settings in which that person does not participate directly (e.g., the workplaces of the child's parents), but that, nonetheless, affect his or her development and well-being (e.g., stressful experiences at work that impair the quality of parents' interactions with their children at home). Finally, the macrosystem level of analysis encompasses the belief systems, social norms, and institutions of the culture as a whole.

The joint influence of multiple life settings on individuals' health has been observed in several studies. Children's exposure to environmental stressors (such as high levels of spatial density and noise) in their home and elementary school environments, for instance, was associated with both additive and interactive effects of those conditions on their physiological health (e.g., systolic and diastolic blood pressure) and academic achievement (cf Cohen et al., 1986). In other studies, employees' perceived lack of flexibility in work hours for scheduling children's doctor visits was associated with their underutilization of employer-provided family health benefits (Fielding et al., 1994). Health outcomes associated with the work–family mesosystem have also been documented through studies of the adverse impacts of 'work–family conflict,' as well as the positive effects of spousal support in buffering work-related stressors (cf O'Neil and Greenberger, 1994).

The research findings noted earlier suggest the value of studying human health from an ecological systems perspective. By considering the joint influence of multiple life settings on individuals' well-being, a broader understanding of how

personal and situational factors contribute to health status can be achieved, than by focusing more narrowly on environmental conditions within single settings.

Directions and Challenges of Ecology and Health Research

The broad conceptual scope of ecology and health research is both an asset and a limitation. On the one hand, the breadth of ecological analyses affords a more comprehensive understanding of the multiple determinants of well-being. For instance, major political, economic, and environmental determinants of health that were neglected in earlier behavioral and medical studies yet have been targeted in more recent ecological analyses. On the other hand, so many disciplines and levels of analysis are combined in ecological research that it is difficult to achieve parsimonious explanations of health phenomena. Thus, an important challenge facing the field of ecology and health is to develop criteria for identifying the most influential determinants of wellness at particular levels of analysis (intrapersonal, social, organizational, institutional, community, global) and to give greatest priority to these high-leverage variables when designing scientific studies and community health promotion programs (see Health Education and Health Promotion).

A rapidly growing area of health research focuses on the impacts of individuals' virtual ecology on their physical, social, and emotional well-being (cf Stokols, 1999). Studies in this area have documented potential conflicts inherent in people's immersion in virtual settings and digital communications (e.g., chat rooms, multimedia multitasking, use of cell phones) while they are concurrently engaged in place-based relationships and behavior settings (cf Pea et al., 2012). For instance, the effects of perceived information overload from cybersources on individuals' self-reported levels of stress and health impairment (Misra and Stokols, 2012), and physical injuries resulting from multimedia multitasking while walking or driving, have been examined (Meyers, 2013; Strayer and Drews, 2007). At the same time, positive influences of the Internet on personal and community health, through telemedicine, telewellness, and the use of personalized mobile communications for health promotion and disease prevention have been identified (Stokols et al., 2009).

Developing High-Leverage Strategies for Effective Health Promotion

Examples of high-leverage variables and health promotion strategies include (1) the implementation of strategic corporate and public policies that can enhance population health on a broad scale (e.g., California's Proposition 99 which imposed a $0.25 tax on each pack of cigarettes sold and thereby reduced statewide levels of smoking prevalence (cf Breslow and Johnson, 1993)); (2) the modification of 'other-directed' health behaviors enacted by corporate and public decision makers whose actions directly influence health outcomes among large numbers of their employees and constituents (cf Stokols, 1996); (3) the identification of key socioeconomic indicators (e.g., female participation in the work force, unemployment, per capita government spending on health care) associated with rates of suicide (Milner et al., 2012); (4) the implementation of defensible space design guidelines to reduce the incidence of violence and crime in urban neighborhoods (cf Newman, 1972); and (5) the development of internet-based 'telewellness' programs designed to enhance health awareness and well-being among large segments of the population (cf Stokols, 1999).

Targeting Vulnerable Subgroups in Health Promotion Programs

An important direction for future ecological research is to incorporate the above leveraging strategies into the design of community health promotion programs aimed at improving the well-being of impoverished and underrepresented groups. Certain illness risk factors, including environmental racism, exposure to community violence, and lack of medical insurance, have proven to be highly resistant to earlier and more traditional health promotion programs (cf Bullard, 2005; Fisher, 1995). These interrelated community problems, borne disproportionately by low-income minority groups in the population, are highly complex and will be reversed only through the design and implementation of strategically leveraged, multilevel intervention programs (Winett et al., 1989), for example, the multisectoral Healthy Cities Programs sponsored by the World Health Organization (cf Ritsatakis, 2012).

Conclusions

The field of ecology and health, which grew rapidly during the last three decades, highlights the importance of developing integrative, broad-gauged strategies for promoting human wellness. As proponents of ecological analyses have noted aptly, improvements in individuals' health practices "usually require some combination of educational, organizational, economic, and environmental interventions in support of change in both behavior and conditions of living" (Green et al., 1997: p. 125). The principles of systems theory and social ecology outlined earlier provide a useful framework for establishing more integrative and effective approaches to wellness promotion in the coming decades.

See also: Environmental Health and Safety: Social Aspects; Environmental Stress and Health; Health Education and Health Promotion.

Bibliography

Adler, N.E., Stewart, J., 2013. Preface to the biology of disadvantage: socioeconomic status and health. Annals of the New York Academy of Sciences 1186, 1–4.

Barker, R.G., 1968. Ecological Psychology: Concepts and Methods for Studying the Environment of Human Behavior. Stanford University Press, Stanford, CA.

Belloc, N., Breslow, L., 1972. Relationship of physical health status and health practices. Preventive Medicine 1, 409–421.

Berkman, L.F., Syme, S.L., 1979. Social networks, host resistance, and mortality: a nine-year follow-up study of Alameda County residents. American Journal of Epidemiology 109, 186–204.

Breslow, L., Johnson, M., 1993. California proposition 99 on tobacco, and its impact. Annual Review of Public Health 14, 585–604.

Bronfenbrenner, U., 1979. The Ecology of Human Development: Experiments by Nature and Design. Harvard University Press, Cambridge, MA.

Bull, M., 2005. A comparative review of best practice guidelines for the diversion of drug related offenders. International Journal of Drug Policy 16 (4), 223–234.

Bullard, R.D. (Ed.), 2005. The Quest for Environmental Injustice: Human Rights and the Politics of Pollution. Sierra Club Books, San Francisco, CA.

Cassel, J., 1964. Social science theory as a source of hypotheses in epidemiological research. American Journal of Public Health 54, 1482–1488.

Cohen, S., Evans, G.W., Stokols, D., Krantz, D.S., 1986. Behavior, Health and Environmental Stress. Plenum, New York.

Duhl, L.J., 1996. An ecohistory of health: the role of 'healthy cities'. American Journal of Health Promotion 10, 258–261.

Emery, F.E., Trist, E.L., 1972. Towards a Social Ecology: Contextual Appreciation of the Future in the Present. Plenum Press, London.

Faris, R.E.L., Dunham, W., 1939. Mental Disorders in Urban Areas. University of Chicago Press, Chicago, IL.

Fielding, J.E., Cumberland, W.G., Pettitt, L., 1994. Immunization status of children of employees in a large corporation. Journal of the American Medical Association 271, 525–530.

Fisher Jr., E.B., 1995. The results of the COMMIT trial. American Journal of Public Health 85, 159–169.

Folke, C., 2006. Reslience: the emergence of a perspective for social ecological systems analyses. Global Environmental Change 16, 253–267.

Garzón, C., Cooley, H., Herberger, M., Moore, E., Allen, L., Matalon, E., Doty, A., The Oakland Climate Action Coalition, 2012. Community-Based Climate Adaptation Planning: Case Study of Oakland, California. Pacific Institute. California Energy Commission Publication Number, Commission. Publication Number: CEC-500-2012-038.

Green, L.W., Ottoson, J.M., 1999. Community and Population Health, eighth ed. WCB/McGraw-Hill, Boston, MA.

Green, L.W., Simons-Morton, D., Potvin, L., 1997. Education and life-style determinants of health and disease. In: Detels, R., Holland, W., McEwen, J., Omenn, G.S. (Eds.), Oxford Textbook of Public Health, third ed. Oxford University Press, New York, pp. 125–137.

Grunberg, N.E., 1991. Cigarette smoking at work: data, issues, and models. In: Weiss, S.M., Fielding, J.E., Baum, A. (Eds.), Perspectives in Behavioral Medicine: Health at Work. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 75–98.

Hawley, A.H., 1950. Human Ecology: A Theory of Community Structure. Ronald Press, New York.

IPCC, 2012. Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK.

Kaplan, G.A., Pamuk, E.R., Lynch, J.W., et al., 1996. Inequality in income and mortality in the United States: analysis of mortality and potential pathways. British Medical Journal 312, 999–1003.

Karasek, R., Theorell, T. (Eds.), 1990. Healthy Work: Stress, Productivity, and the Reconstruction of Working Life. Basic Books, New York.

Kelly, G., 2006. Becoming Ecological: An Expedition into Community Psychology. Oxford University Press, New York.

Kickbusch, I., 1989. Approaches to an ecological base for public health. Health Promotion 4, 265–268.

Lachman, M.E., Weaver, S.L., 1998. The sense of control as a moderator of social class differences in health and well-being. Journal of Personality and Social Psychology 74, 763–773.

Lavarreda, S.A., Brown, E.R., Bolduc, C.D., 2011. Underinsurance in the United States: an interaction of costs to consumers, benefit design, and access to care. Annual Review of Public Health 32, 471–482.

Leaf, A., 1989. Potential health effects of global climatic and environmental changes. New England Journal of Medicine 321, 1577–1583.

MacDonald, J., 2002. The effectiveness of community policing in reducing urban violence. Crime & Delinquency 48, 592–618.

Maller, C., Henderson-Wilson, C., Pryor, A., Prosser, L., Moore, M., 2008. Healthy Parks, Healthy People; the Health Benefits of Contact with Nature in a Park Context, second ed. Deakin University, Burwood, Melbourne. http://parkweb.vic.gov.au/__data/assets/pdf_file/0018/313821/HPHP-deakin-literature-review.pdf. (accessed on 22.07.13.).

Malmstrom, M., Sundquist, J., Johansson, S.E., 1999. Neighborhood environment and self-reported health status: a multilevel analysis. American Journal of Public Health 89, 1181–1186.

Marmot, M.G., Bosma, H., Hemingway, H., et al., 1997. Contribution of job control and other risk factors to social variations in coronary heart disease incidence. The Lancet 350, 235–239.

Maruyama, M., 1963. The second cybernetics: deviation-amplifying mutual causal processes. American Scientist 51, 164–179.

McGinnis, J.M., Lee, P.R., 1995. Healthy people 2000 at mid decade. Journal of the American Medical Association 273, 1123–1129.

Meyers, D., 2013. Multitasking in the Automobile. Association for Psychol. Science Observer. http://www.psychologicalscience.org/index.php/publications/observer/2013/july-august-13/multi-tasking-in-the-automobile.html (accessed 21.07.13.).

Michelson, W.H., 1985. From Sun to Sun: Daily Obligations and Community Structure in the Lives of Employed Women and Their Families. Rowman & Allenhald, Totowa, NJ.

Milner, A., McClure, R., DeLeo, D., 2012. Socio-economic determinants of suicide: an ecological analysis of 35 countries. Social Psychiatry and Psychiatric Epidemiology 47 (1), 19–27.

Misra, S., Stokols, D., 2012. Psychological and health outcomes of perceived information overload. Environment and Behavior 44 (6), 737–759.

Moos, R.H., 1979. Social ecological perspectives on health. In: Stone, G.C., Cohen, F., Adler, N.E. (Eds.), Health Psychology: A Handbook. Jossey-Bass, San Francisco, CA, pp. 523–547.

Newman, O., 1972. Defensible Space: Crime Prevention through Urban Design. Macmillan, New York.

Ottawa Charter for Health Promotion, 1986. Health Promotion 1: iii–iv. http://www.who.int/healthpromotion/conferences/previous/ottawa/en/.

O'Neil, R., Greenberger, E., 1994. Patterns of commitment to work and parenting: Implications for role strain. Journal of Marriage and the Family 56, 101–112.

Park, R., Burgess, E., McKenzie, R.D. (Eds.), 1925. The City. University of Chicago Press, Chicago, IL.

Pea, R., Nass, C., Meheula, L., Rance, M., Kuman, A., Bamford, H., Nass, M., Simha, A., Stillerman, B., Yang, S., Zhou, M., 2012. Media use, face-to-face communication, media multitasking, and social well-being among 8- to 12-year old girls. Developmental Psychology 48 (2), 327–336.

Pelletier, K.R., 1996. A review and analysis of the health and cost-effective outcome studies of comprehensive health promotion and disease prevention programs at the worksite: 1993–1995 update. American Journal of Health Promotion 10, 380–388.

Richard, L., Gauvin, L., Raine, K., 2012. Ecological models revisited: their uses and evolution in health promotion over two decades. Annual Review of Public Health 32, 307–326.

Ritsatakis, A., 2012. Equity and the social determinants of health in European cities. Journal of Urban Health: Bulletin of the New York Academy of Medicine. http://dx.doi.org/10.1007/s11524-012-9762-y.

Robert, S.A., 1998. Community-level socioeconomic status effects on adult health. Journal of Health and Social Behavior 39, 18–37.

Sallis, J.F., Cervero, R.B., Ascher, W., Henderson, K.A., Kraft, M.K., Kerr, J., 2006. An ecological approach to creating active living communities. Annual Review of Public Health 27, 297–322.

Sanders-Phillips, K., 2000. Health promotion in ethnic minority families: the impact of exposure to violence. In: Schneider Jamner, M., Stokols, D. (Eds.), Promoting Human Wellness: New Frontiers for Research, Practice, and Policy. University of California Press, Berkeley, CA, pp. 294–324.

Schneider Jamner, M., Stokols, D. (Eds.), 2000. Promoting Human Wellness: New Frontiers for Research, Practice, and Policy. University of California Press, Berkeley, CA.

Schneider, M., Stokols, D., 2008. Multi-level theories of behavior change: social ecological theory. In: Shumaker, S.A., Ockene, J.K., Riekert, K. (Eds.), Handbook of Health Behavior Change, third ed. Springer Publishing Company, New York, pp. 85–105.

Stern, P.C., Young, O.R., Druckman, D. (Eds.), 1992. Global Environmental Change: Understanding the Human Dimensions. National Academy Press, Washington, DC.

Stillman, F.A., Hartman, A.M., Graubard, B.I., Gilpin, E.A., Murray, D.M., Gibson, J.T., November 19, 2003. Evaluation of the American Stop Smoking Intervention Study (ASSIST): a report of outcomes. Journal of the National Cancer Institute 95 (22), 1681–1691.

Stokols, D., 1992. Establishing and maintaining healthy environments: toward a social ecology of health promotion. American Psychologist 47, 6–22.

Stokols, D., 1996. Translating social ecological theory into guidelines for community health promotion. American Journal of Health Promotion 10, 282–298.

Stokols, D., 1999. Human development in the age of the Internet: conceptual and methodological horizons. In: Friedman, S.L., Wachs, T.D. (Eds.), Measuring Environment across the Lifespan: Emerging Methods and Concepts. American Psychological Association, Washington, DC, pp. 327–356.

Stokols, D., 2000. Social ecology and behavioral medicine: Implications for training, practice, and policy. Behavioral Medicine 26, 129–130.

Stokols, D., Misra, S., Runnerstrom, M., Hipp, A., 2009. Psychology in an age of ecological crisis: from personal angst to collective action. American Psychologist 64 (3), 181–193.

Stokols, D., Lejano, R., Hipp, J., 2013. Enhancing the resilience of human-environment systems: a social ecological perspective. Ecology & Society 18 (1). http://www.ecologyandsociety.org/vol18/iss1/art7/ (accessed 21.07.13.).

Strayer, D.L., Drews, F.A., 2007. Cell phone-induced driver distraction. Current Directions in Psychological Science 16, 128–131.

United States Public Health Service, 1979. Healthy People: The Surgeon General's Report on Health Promotion and Disease Prevention. DHEW Publication No. (PHS) 79–55071. U.S. Government Printing Office, Washington, DC.

Winett, R.A., King, A.C., Altman, D.G., 1989. Health Psychology and Public Health: An Integrative Approach. Pergamon Press, New York.

Relevant Websites

http://www.cancer.gov/newscenter/qa/2004/assistqandA – American Stop Smoking Intervention Study (ASSIST) Evaluation: Questions and Answers.

http://www.ejrc.cau.edu/ – Environmental Justice Center at Clark Atlanta University.

http://www.umich.edu/~snre492/, http://www.snre.umich.edu/degrees/masters/environmental_justice/overview – Environmental Justice Initiative, University of Michigan.

http://www.sesync.org/ – National Socioenvironmental Synthesis Center.

http://rs.resalliance.org/2010/04/16/expansion-of-social-ecological-systems-science/ – Resilience Science.

http://www.who.int/globalchange/health_policy/who_workplan/en/index.html – WHO work plan on climate change and health.

Ecology of Aging

Hans-Werner Wahl, Institute of Psychology, Heidelberg University, Heidelberg, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H.-W. Wahl, volume 6, pp. 4045–4048, © 2001, Elsevier Ltd.

Abstract

The ecology-of-aging paradigm emphasizes the role of the physical, spatial, and technological environment on behavior and aging well. This work reviews the historical development of this field of study and defines its role within aging research in general. Furthermore, major theories and findings of the ecology of aging are summarized and discussed. The article closes with future considerations and unmet research needs.

The ecology of aging basically draws upon the idea that old age is a period of adult development profoundly influenced by the environment. Because age-related losses in vision, hearing, mobility, and cognitive capacity have a direct impact on the relationship between the organism and the environment, the older person is particularly vulnerable to environmental demands. In other words, the ecology of aging views contextual factors, particularly the physical, spatial, and technological environment, as crucial determinants of the autonomous exertion of everyday behavior and well-being. According to Lawton and Nahemow (1973: p. 621), the ecology of aging is defined as "a system of continual adaptation in which both the organism and the environment change over time in a non random manner." That is, both the aging person as well as his/her environment may influence each other in various ways and the outcomes of this reciprocal process may serve or hinder autonomy and well-being. Given its emphasis on the physical, spatial, and technological environment, the ecology of aging also is an important addition to aging research focusing on the social context (Wahl and Lang, 2006). Thus, although the term environment, as used in the ecology of aging, predominantly refers to the physical environment, its interwoven nature with social and cultural accoutrements and implications is also acknowledged. Importantly, person-environment relations have not only an objective but also a subjective dimension (Oswald and Wahl, 2005). Aging individuals, not the least because they live for long time at the same place, form strong cognitive and affective ties to their environments and in many cases (e.g., when relocation to a nursing home arises as a key residential decision) such cognitive-emotional ties seem to be more important than objective environmental characteristics.

Development of the Field

Since aging has long been regarded as a process mostly determined by a biological program inherent in the organism, the explicit consideration of environmental variables that affect the course and outcome of human aging was an important step in the historical development of aging research. This paradigm shift was and continues to be influenced by the following:

1. The Chicago School of Urban Sociology and its human ecology perspective (Burgess, McKenzie, and Park) in the 1920s and 1930s, the writings of Lewin in the 1930s and 1940s as well as the emergence of ecological psychology and of the social ecology view of human development (Barker, Bronfenbrenner, and Moos) in the 1960s and 1970s, which, among other things, placed particular emphasis on the resource character of the physical and spatial environment and hence provided another set of roots for this discipline (Lawton, 1982).
2. The growing role of the social science perspective within aging research, which began in the late 1940s of the twentieth century and set the stage for the emerging field of social and behavioral gerontology (*see* Aging, Theories of).
3. The prominent role of learning theories in psychology in the 1950s and 1960s, which underscored the importance of the environment and external stimuli in all stages of human development (*see* Human Development, Theories of).
4. The visible demographic shift toward a growing number of old and very old adults, raising issues of optimal housing and out-of-home areas for aging well.
5. The increasing importance of technology and the new media as a 'new' environment for all ages including those in their later life.

The conceptual and empirical productivity within the field grew steadily during the 1960s and 1970s (of the twentieth century). Respective research has been devoted to better understand 'aging in place' processes (how older adults manage to stay put even when faced with severe competence loss); to develop housing alternatives for older individuals; and to learn about relocation processes and outcomes and aging in institutions ever since. In addition, issues such as out-of-home mobility and more recently also the interaction with computers and virtual environments have found research attention. Also, the important health issues of mild cognitive impairment and dementia-related disorders have led to new research challenges based on the assumption that environmental design becomes particularly crucial when the cognitive capacity for person-based adaptations in later life is rather exhausted. Most recently, new areas such as the question of how climate change may affect aging have appeared as an issue of the ecology of aging.

Goals and Mission of Ecology of Aging and Role within Aging Research

Goals and Mission

The overarching aim of ecology of aging research is to describe, explain, and modify/optimize the relationship between the

aging person and his/her physical, spatial, and technological environment. With regard to description, the ecology of aging places a strong emphasis on day-to-day contexts of aging individuals, reinforcing the notion that this everyday world in natural settings deserves full attention of aging research at large. A prototypical issue is that older people spend most time in their home environment and as a consequence, housing characteristics as the major setting in which aging unfolds, has been a major focus of research in environmental gerontology from its inception.

The task of explanation basically relies on German psychologist Kurt Lewin's fundamental definition of the 'life space' as a function of the person and his/her environment ($B = f(P\ E)$). Lawton (1982) suggested an addition to the 'ecological equation' ($B = f(P\ E\ PXE)$) to account for the influence of person-environment interactions. This interaction term suggests that person-related competencies and needs operate or manifest differently in different environments and lead to dissimilar outcomes. For instance, having low- or high-mobility competence may not affect the autonomy of an individual to obtain a goal such as ascending stairs, if an escalator is also available. However, if an escalator is not available, the autonomy of the low-mobility person may be highly threatened, whereas there may be no effect on autonomy for the person with high mobility.

The focus on modification and optimization reflects the aspiration of ecology of aging research to have a substantial and direct contribution to the improvement of quality of life in old age. Prototypical is the use of ecology of aging findings in advancing evidence-driven home modifications, enriching further long-term care settings, or designing of public spaces with the goal of enhancing autonomy and well-being as people age.

With these ambitions in mind, empirical methods, such as real-life observations in natural settings, the retrospective assessment of day-to-day behaviors through the use of diaries, the reconstruction of internal representations of the environment, as well as structured questionnaires have frequently been used. Methods used to describe the environment, such as checklists for housing interiors or multidimensional evaluations of institutional settings, have also played an important role (Gitlin, 1998). More recently, tracking methods based on Global System Positioning technology or other automatized activity recording *in vivo* have found application to learn about naturally occurring mobility and space use patterns (e.g., Wahl et al., 2013).

Role of the Ecology of Aging within Aging Research

Research on the ecology of aging strives to understand the environment-related tasks that typically confront the aging individual. These tasks include preserving autonomy in everyday life despite physical and cognitive impairments through the utilization of environmental resources within (including home modifications, see Wahl et al., 2009) and outside the home, initiating the relocation process if desired or necessary and adapting to new living environments (such as a nursing home or other planned housing) after relocation. These classic research themes illustrate the fact that an interdisciplinary approach – covering disciplines such as psychology, sociology, architecture, geography, urban planning, and design – is essential for the understanding of the ecology of aging.

There are numerous implications of ecology of aging principles for models of aging well. For example, because housing figures prominently in the major day-to-day context of aging, research into the potential and constraints of the home environment addresses important issues related to preserving health and quality of life in later life. There has been and continues to be a strong desire of old and very old people to 'age in place,' remaining in their familiar home environment or an environment of their choice for as long as possible. This desire may ultimately reflect an optimization strategy critical to late life; it affords continuity and control over important aspects of daily life, including routines, self-care, and other potentially meaningful and self-defining activities. In sum, ecology of aging principles may be particularly critical to understanding the 'good life' in old age (Lawton, 1983).

Major Themes, Theoretical Approaches, and Findings of the Ecology of Aging

The in-depth understanding of person-environment processes and outcomes has so far not been achieved by any one major theory; rather a multitude of conceptual approaches that augment and build on each other in a pluralistic manner have been suggested over the years since the inception of ecology of aging research in the 1970s.

Objective Approaches to Person-Environment Relations

The Competence-Press Model (Lawton, 1982) provides a broad, overarching framework allowing different types and levels of competence such as sensory loss, physical mobility loss, or cognitive decline and environmental factors including housing standards, neighborhood conditions, or public transport to be considered. The fundamental assumption is that for each aging person there is an optimal combination of (still) available competence and environmental circumstances leading to the relative highest possible behavioral and emotional functioning for that person. The model also suggests that it is at the lower levels of competence that older people become the most susceptible to their environment such that low competence in conjunction with high 'environmental press' negatively affects an individual's autonomy, effect, and well-being. The competence-environmental press framework continues to provide the basic mechanism of person-environment relations as people age and has been supported by a considerable body of empirical research. For example, research of the impact of physical distances on social interaction patterns of elders in institutional settings shows that longer distances undermine social relations, thus highlighting the 'environmental docility' as people age. Furthermore, the Competence-Press Model has become a major driver in the practical world of designing and optimizing environments for older adults.

Besides putting emphasis on competence as is done in the Competence-Press Model, other conceptions underscore the role of person-environment fit and motivational forces in person-environment relations (see also Lawton, 1982). The basic

assumption is that the lack between given needs and given environmental options to fulfil these needs is linked with lowered behavioral functioning and well-being. Empirical support for this assumption has been reported by studies conducted in institutional settings as well as private home environments. For example, remaining autonomy in visually impaired elders was found to depend on how much the physical layout of home environments adds to the compensation of the vision loss. Furthermore, a distinction between an older person's basic versus higher-order needs has been suggested. Basic needs are related to basic issues of maintaining competence (e.g., maintaining independent self-care), while higher-order needs address issues such as privacy and affiliation. While person-environment misfit in the basic need domain predominantly results in reduced behavioral autonomy, misfit in the higher-order need realm predominantly may also undermine emotional well-being and mental health.

The Person-Environment Stress Model (see also Lawton, 1982) argues that objective environmental conditions such as lack of accessibility, privacy, or control in one's environment may provide the source for stress experiences. For instance, aging individuals may feel a psychological threat, apathy, reduced well-being, or depression as a consequence of such stress experience. This model can be applied to family care giving as well. For example, difficult to manage dementia-related behaviors and lack of control over daily routines reflect the objective environmental conditions that typically confront families of persons with dementia. These environmental stressors over and above subjective appraisals of the care situation may affect a caregiver and result in significant stress, anxiety, or depression.

Finally, relocation may be seen as the most radical alteration in person-place relations. The understanding of residential decisions and their outcomes in old age has attracted the ecology of aging from early on in the development of the field. Litwak and Longino (1987) suggested a distinction between first, second, and third moves by relying on the assumption of a substantial association between chronological age and type and motivation of move. Whereas 'first moves' often take place early in the aging process, that is, close to retirement, and are usually prompted by the amenities associated with the desired place of residence, 'second moves,' appearing in the 70s and early 80s of age, are predominantly characterized by moving back to the place of origin, reflecting a higher need for support and proximity to kinship. 'Third moves' are often relocations to institutions due to frailties that are not compensable by family resources in very old age. As has been found in more recent research, relocation is driven in current cohorts of older people by a variety of motivations, among which higher-order needs such as access to participation in social, cultural, and educational opportunities appear to become increasingly important. A primary example is the so-called Sunbelt Movers, that is, the move of mostly 'young old' people to places better serving their needs in terms of climatic conditions and/or surrounding communities as found in 'Sun Cities.'

Subjective Approaches to Person-Environment Relations

The classic approach addressing the cognitive-affective relation to the physical environment is reflected in the construct of housing satisfaction. As has consistently been found, older adults, both healthy and frail, tend to score high in measures related to this construct, regardless of their objective home and neighborhood conditions. This has been coined the 'housing satisfaction paradox.'

Besides residential satisfaction, a widely acknowledged approach suggested by Rowles (1983) has focused on the many faces of what he has coined 'insideness.' According to Rowles, different types of insideness that speak to the transition of 'spaces into places' can be empirically derived. Whereas 'social insideness' arises from everyday social exchanges over long periods of time, 'physical insideness' is characterized by familiarities and routines within given settings such as the home environment. A third element of insideness has been labeled 'autobiographical insideness,' that is, places carry a rich collection of memories and thus support the aging individual's sense of place identity. Empirical research has also shown that physical insideness is particularly important for aging individuals with chronically disabling conditions.

Rubinstein's conception of the 'meaning of home' (see Oswald and Wahl, 2005) relies on the assumption that the active management of the environment in itself represents a major source of well-being as people age, especially for those who are frail or living alone. Rubinstein identifies three classes of psychosocial processes that give meaning to the home environment, namely 'social-centered,' ordering of the home environment based on a person's version of sociocultural rules for domestic order; 'person-centered,' expression of one's life course in features of the home; and 'body-centered' processes, the ongoing relationship of the body to the environmental features that surround it. Furthermore, 'place attachment' has been defined as reflecting feelings about a geographic location and emotional binding of a person to places. As has been empirically found, place attachment seems to steadily grow across the life course, reaching its culminating point in very old age.

Role of Out-of-Home Environments for Aging Well

As Wahl et al. (2007) argued, the presence of service facilities for daily necessities and appropriate health care nearby, areas of recreation and stimulation, cultural facilities, natural characteristics, and feeling safe and secure in the area, are all essential factors linked to autonomy and personal well-being. Regional conditions – especially those relating to the degree of urbanization and distances between places of residence – can also affect the frequency and nature of contacts with important reference persons and participation in social activities that are regarded as crucial elements for the quality of life of older adults. Older people in rural areas live generally less far away from their most important confidants than older urbanites, while the transportation times needed in, for example, large urban areas increase the difficulties of maintaining social relationships. Such structural characteristics refer to the significance of out-of-home mobility as an increasingly important precondition for maintaining an independent lifestyle in later life and keeping up social relationships and the connection with one's community. Improvements in transport technologies, accompanied by changes in the constitution and nature of urban and rural

settlements, have resulted in a growing dispersal of travel origins and destinations and the necessity of bridging the widening gap between functional areas. Apart from these functional aspects of getting places, mobility can be performed for its own sake. Moreover, in modern society, mobility is associated with highly valued societal goals such as freedom, autonomy, and flexibility (*see* Retirement and Encore Adulthood: The New Later Life Course).

Technological Environments and Aging – Some Considerations

Rogers and Fisk (2010) have argued how important it is for older adults' use of technology to match psychological needs with design characteristics. They also refer to the Competence-Environment Press Model (Lawton, 1982) according to which successful performance is depending on the demands imposed by the environment in relation to the capabilities of the individual. Lindenberger et al. (2008) suggest that the interaction between the older person and technology should be seen against three fundamental principles: (1) net resource release, or marginal resource benefit; (2) person specificity; and (3) proximal versus distal frames of evaluation considering constraints as well as capabilities of aging individuals. The first principle of 'net resource release' states that the use of technology is only adaptive if the operation costs (resource investments) are lower than the pay-offs regarding the cognitive resources that are released. For example, using a complex videophone system initially requires substantial resource investments to understand the instructions, resulting in a low resource balance regarding this device. When developing and evaluating assistive technology, objective and subjective assessments of the resource balance should be taken into account. Improvements in the resource balance caused by using the technological device should be of significant amount – at least in the long run. To enable such benefits it is essential that technological and psychological knowledge is integrated. The second principle of 'person specificity' points to the need of adaptation to the specific habits, abilities, and preferences of the potential users. Besides the general importance of taking seriously the pronounced heterogeneity of the older population, it seems crucial to introduce the potentially helpful role of technology early in the disablement process, in particular before the onset of severe cognitive impairment. By this means, the older user may be able to adapt easier to the technology in the longer run, even if cognitive decline happens at a later point in time, and there still remains room for precise adjustment on the part of the device concerning the user's specific competences or preferences. The third principle of 'proximal versus distal frames of evaluation' focuses on long-term benefits, considers risks regarding the entire lifespan, and takes into account possible historical changes. Depending on the observation period, the evaluation may provoke different results. Following the three principles of Lindenberger et al.'s (2008) framework, the relation between support and challenge of technology can be improved with positive implications for everyday performance, participation, and well-being. It is obvious that to use this opportunity in its full potential, close interdisciplinary collaboration between psychologists, engineers, computer scientists, and geriatricians is needed.

Approaching Person-Environment Relations on the Macrolevel of Analysis

The aforementioned approaches predominantly consider person-environment interactions on the microlevel of analysis; that is, the focus is on obtaining an in-depth understanding of the role of the immediate environment within which people age. However, as Bronfenbrenner's (1979) ecological theory of human development and other social science approaches suggest, in targeting person-environment relations, the micro-environmental level is undoubtedly influenced and shaped as well by macrolevel environmental contexts including the neighborhood at large, housing policies, economic standards, cultural values and beliefs, and social and long-term care service provision and policies. These factors all form the environmental context of aging with varying immediate and distal impacts on everyday life of people as they age. As has been found, the quality of neighborhood settings may strongly affect the physical range and timing of out-of-home activities as people age. An important consideration is the life course perspective. It could well be that persons initially live in high-quality neighborhoods that subsequently decline over time, placing people in the later years of their life at a disadvantage. Another classic research question is the urban-rural distinction and how this affects aging. For example, use of the outside environment may depend upon public transportation possibilities and these may be more underdeveloped in rural than urban regions. On the other hand, the macrolevel perspective also considers the diversity in cultural norms, and it could well be that disadvantages in person-environment options in rural areas are compensated for because of stronger intergenerational and higher prevalence of cohabitation of older and younger family members. Still one level up, the consideration of person-environment interrelations is worth considering across countries, that is, as a cross-cultural research challenge. For example, the definition of barrier-free environments and respective legislative norms and regulations probably depend on the economic situation as well as on different cultural conceptions of autonomy and the role of physical environments compared to social environments such as the family system.

Toward Integrative Views

To expand important perspectives in the field of the ecology of aging and link them more closely with developmental models of aging well, a framework as depicted in **Figure 1** may be helpful (Wahl et al., 2012). At the core of this framework is the assumption that two processes, experience-driven belonging and behavior-driven agency, help to better understand and integrate person-environment interchanges as people age. 'Belonging' reflects a sense of positive connection with other people and the environment, while 'agency' refers to the process of becoming a change agent in one's own life by means of intentional and proactive behaviors.

We argue that adapting these constructs to the person-environment sphere is worthwhile, as they are particularly useful for understanding the role of environment for aging well. Going further, the model assumes that person-environment resources are essential to both belonging and agency. Personal resources, including functional abilities,

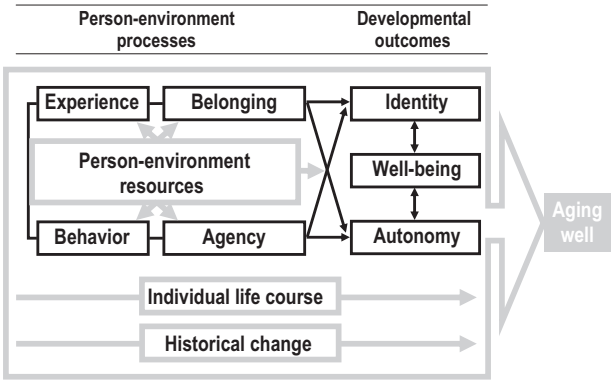

Figure 1 Conceptual framework: interplay of belonging and agency, aging well, and the environment. Reproduced from Wahl, H.-W., Iwarsson, S., Oswald, F., 2012. Aging well and the environment: toward an integrative model and research agenda for the future. The Gerontologist 52 (3), 308. http://dx.doi.org/10.1093/geront/gnr154.

cognitive and affective functioning, and personality traits, may gain or lose importance in specific environmental configurations. Similarly, environmental resources, including barrier-free housing or innovative housing solutions, are utilized differently across persons. Our model also posits that the interaction of belonging and agency unfolds within a life course perspective, with the processes of belonging increasing in importance as people enter old and particularly advanced old age, while the relevance of processes of agency decreases. Figure 1 also points to the need to pay serious attention to historical, cohort-related, and cultural context. Historical changes relevant to aging well can be traced to changes in aging individuals' relations with the environment. For example, the impact of the environment on aging well has probably increased as new housing solutions for older adults, such as assisted living or intergenerational arrangements, and new housing options for persons with dementia have replaced traditional built environments. Finally, aging well is seen as maintaining the highest autonomy, well-being, and preservation of one's self and identity as possible, even in the face of severe competence loss.

Future Considerations

Future research on the ecology of aging is faced with at least four challenges:

1. Theoretical challenges: Much of the theoretical substance in this field was developed in the 1960s and 1970s (of the twentieth century). One particularly fruitful theoretical avenue may be to strengthen the links between ecological theories and promising conceptual approaches within psychological gerontology. A prototypical example is the lifespan theory of control (Heckhausen et al., 2010), which focuses, on a wider scale, also on person-environment relations and their alterations across the human life span.
2. Methodological challenges: Although the use of naturalistic methods such as observations and online data-collection strategies (e.g., by using i-pads in everyday life situations) have proven themselves to be indispensable tools in research on the ecology of aging – as well as for gerontology in general – the methods used are rarely comparable. One of the most important methodological challenges is creating an established set of instruments able to robustly assess the older adults' interaction with the physical, spatial, and technological environment. It may also become very fruitful to better link ecology of aging issues with the neurosciences in the future (*see* Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences). For example, it is currently mostly unknown, whether different housing solutions may also affect differently on the aging brain.
3. Application and societal challenges: Due to the rapidly growing number of older adults and particularly 'old-old' adults beyond the age of 85 years, one of the main challenges of today's and tomorrow's societies is providing care for the demented older adults. Research on the ecology of aging must be strongly committed to planning and evaluating supportive environments for these people. In addition, the growing impact of information and communication technology such as the Internet, the use of high-tech solutions in housing and in public community life, as well as the rapidly growing role of technology in geriatric care have given rise to a whole range of new environmental characteristics. Research on the ecology of aging will thus continue to encounter new challenges in the future.

See also: Aging Mind: Facets and Levels of Analysis; Aging and Health in Old Age; Aging, Theories of; Human Development, Successful: Psychological Conceptions; Retirement and Encore Adulthood: The New Later Life Course; Retirement and Health.

Bibliography

Bronfenbrenner, U., 1979. The Ecology of Human Development. Experiments by Nature and Design. Havard University Press, Cambridge.

Gitlin, L.N., 1998. Testing home modification interventions: Issues of theory, measurement, design, and implementation. Annual Review of Gerontology and Geriatrics 18, 190–246.

Heckhausen, J., Wrosch, C., Schulz, R., 2010. A motivational theory of life-span development. Psychological Review 117 (1), 32–60. http://dx.doi.org/10.1037/a0017668.

Lawton, M.P., 1982. Competence, environmental press, and the adaptation of older people. In: Lawton, M.P., Windley, P.G., Byerts, T.O. (Eds.), Aging and the Environment: Theoretical Approaches. Springer, New York, pp. 33–59.

Lawton, M.P., 1983. Environment and other determinants of well-being in older people. The Gerontologist 23 (4), 349–357. http://dx.doi.org/10.1093/geront/23.4.349.

Lawton, M.P., Nahemow, L.E., 1973. Ecology and the aging process. In: Eisdorfer, C., Lawton, M.P. (Eds.), The Psychology of Adult Development and Aging. American Psychological Association, Washington, DC, pp. 619–674.

Lindenberger, U., Lövdén, M., Schellenbach, M., Li, S.-C., Krüger, A., 2008. Psychological principles of successful aging technologies: a mini-review. Gerontology 54, 59–68. http://dx.doi.org/10.1159/000116114.

Litwak, E., Longino Jr., C.F., 1987. Migration patterns among the elderly: a developmental perspective. The Gerontologist 27, 266–272.

Oswald, F., Wahl, H.-W., 2005. Dimensions of the meaning of home in later life. In: Rowles, G.D., Chaudhury, H. (Eds.), Home and Identity in Later Life. International Perspectives. Springer Publ., New York, pp. S. 21–S. 46.

Rogers, W.A., Fisk, A.D., 2010. Toward a psychological science of advanced technology design for older adults. Journal of Gerontology: Psychological Sciences 65B (6), 645–653. http://dx.doi.org/10.1093/geronb/gbq065.

Rowles, G.D., 1983. Geographical dimensions of social support in rural Appalachia. In: Rowles, G.D., Ohta, R.J. (Eds.), Aging and Milieu: Environmental Perspectives on Growing Old. Academic Press, New York, pp. 111–130.

Wahl, H.-W., Fänge, A., Oswald, F., Gitlin, L.N., Iwarsson, S., 2009. The home environment and disability-related outcomes in aging individuals: what is the empirical evidence? Gerontologist 49 (3), 355–367. http://dx.doi.org/10.1093/geront/gnp056.

Wahl, H.-W., Lang, F., 2006. Psychological aging: a contextual view. In: Conn, P.M. (Ed.), Handbook of Models for Human Aging. Elsevier, Amsterdam, pp. S. 881–S. 895.

Wahl, H.-W., Mollenkopf, H., Oswald, F., Claus, C., 2007. Environmental aspects of quality of life in old age: conceptual and empirical issues. In: Mollenkopf, H., Walker, A. (Eds.), Quality of Life in Old Age – International and Multidisciplinary Perspectives. Springer, Dordrecht, NL, pp. S. 101–S. 122.

Wahl, H.-W., Iwarsson, S., Oswald, F., 2012. Aging well and the environment: toward an integrative model and research agenda for the future. The Gerontologist 52 (3), 306–316. http://dx.doi.org/10.1093/geront/gnr154.

Wahl, H.-W., Wettstein, M., Shoval, N., Oswald, F., Kaspar, R., Issacson, M., et al., 2013. Interplay of cognitive and motivational resources for out-of-home behavior in a sample of cognitively heterogeneous older adults: Findings of the SenTra project. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 68 (5), 691–702. http://dx.doi.org/10.1093/geronb/gbs106.

Ecology: Organizations

Joel AC Baum, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Until the mid-1970s, the prominent approach in organization and management theory emphasized adaptive change in organizations. Since then, an approach to studying organizational change that places more emphasis on environmental selection processes has become increasingly influential. The stream of research on ecological perspectives of organizational change has generated tremendous excitement, controversy and debate in the community of organization theory scholars.

Inspired by the question '*why are there so many kinds of organizations?*' (Hannan and Freeman, 1977), organizational ecology aims to explain how social and economic conditions affect the relative abundance and diversity of organizations and to account for their changing composition over time. Ecological research typically begins with three observations: (1) diversity is a property of aggregates of organizations, (2) organizations' members often have difficulty devising and executing organizational changes fast enough to meet the changing demands of uncertain environments, and (3) the community of organizations is unstable – organizations come and go continually. Consequently, contrasting adaptation approaches (*see* Structural Contingency Theory) emphasizing individual organizations' cumulative strategic changes, ecological analyses formulate organizational change and variability at the population level, highlighting the creation of new and the demise of old organizations and populations.

Historically, research has focused primarily on analyses of the influence of demographic and ecological processes on rates of organizational founding and failure. The focus of research has changed over time, however, as attention turns to processes of organizational change. As Hannan and Freeman (1977: p. 930) pointed out, a full treatment of organization–environment relations must cover both adaptation and selection. Researchers have embarked on development of an integrated perspective that sees processes of adaptation and selection as complementary and interacting. The emerging view takes seriously the occurrence of selection processes and combines it with the systematic study of organization-level changes that may, under certain conditions, be adaptive.

Demographic Processes

Organizational Founding

Although organizational founding is an important theme in ecological research, in large part, foundings are treated as identical additions to homogenous populations; the characteristics of new organizations, which define their domains, have not been of central interest. Initially, researchers avoided this issue by treating the population as the unit experiencing the foundings. The problem with this approach is that variations in social and economic conditions across a population's environment will produce differences in founding rates, which, if not properly accounted for, will result in specification bias.

Progress has been made by specifying more fine-grained subpopulations into which organizations are differentially founded. A variety of criteria have been employed, including legal form (Baum and Oliver, 1996), core technology (Barnett, 1990), customer base (Baum and Singh, 1994), and geographic location (Cattani et al., 2003). In this approach, studying founding involves specifying potential subpopulations and examining differences in the rate at which new entrants arrive. One limitation is that subpopulation definitions can be artificial and arbitrary; another is that it affords accounts only of foundings into organizational niches that were actually filled.

Several approaches have appeared that address these concerns. Each begins with the observation that variation in social, institutional, and economic conditions across regions can produce differences in founding rates of particular organizational forms. Lomi (1995) proposes a model of location dependence in organizational founding in which corrections for unobserved heterogeneity are estimated directly from data. Baum and Haveman (1997) complement ecological analyses that ask when a founding will occur – by any kind of organization within a population, in any location in that population's niche – by taking organizational foundings as given and asking instead, '*what kind* of organization is founded and *where*?' Their concern is with new organizations' similarity to (difference from) incumbent organizations: given that a new organization appears how similar it is to neighbors? Sorenson and colleagues (Barnett and Sorenson, 2002; Stuart and Sorenson, 2003) make use of continuous measures of the distance between fine-grained geographic units (zip codes), permitting effects of particular variations in social and economic conditions on the founding rate to be estimated directly.

These and related studies have refocused organizational ecologists from an emphasis on temporal heterogeneity in organizational founding to less studied issues of spatial heterogeneity and location dependence. The approaches to the dynamics of spatial and location choice they advance make it possible to subject theoretical predictions from spatial economics and organizational sociology to empirical testing, improving our knowledge of key entrepreneurial decisions and processes, and the tendency for firms to cluster (or not) in a resource space (*see* Entrepreneurship and Economic Growth).

Age and Size Dependence

The relationship between organizational aging and failure is a central demographic inquiry. The predominant view was initially the liability of newness (Stinchcombe, 1965), which assumes that because they have to learn new social roles,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.73111-4

create new organizational routines, and lack endorsements and exchange relationships, young organizations will have higher failure rates. A complementary idea is that selection processes favor structurally inert organizations capable of demonstrating reliability and accountability, which requires high reproducibility. As reproducibility is achieved through the routinization that comes with age, older organizations should be less likely to fail (Hannan and Freeman, 1984).

A related line of research examines how organizational size affects failure. The high failure rate of small organizations results from several liabilities of smallness, including problems with recruiting and training employees, raising capital, and costs of regulatory compliance. As organizations grow, they increasingly emphasize predictability, formalization, and control, improving their reliability and accountability, and lowering their susceptibility to failure. The significance of 'smallness' stems in part from its relation to 'newness': as new organizations tend to be small, if small organizations exhibit higher failure rates, empirical evidence of a liability of newness may reflect specification error. Indeed, although early studies consistently support a liability of newness, subsequent studies controlling for time-varying size typically do not, prompting formulation of alternative views on age dependence.

The liability of adolescence (Fichman and Levinthal, 1991) predicts a ∩-shaped relationship between age and failure. New organizations begin with a stock of assets (e.g., goodwill, commitment, resources) that buffer them during an initial 'honeymoon' period – even if early outcomes are unfavorable. The larger the initial stock, the longer the buffering. As initial endowments are depleted, however, organizations unable to meet ongoing resource needs, for example, because they are unable to establish effective routines or exchange relations, may fail.

Liabilities of newness and adolescence yield divergent predictions for new organizations, but because processes underlying these models occur at an early age, they do not apply to older organizations. The liability of aging (Barron et al., 1994), which identifies processes affecting older organizations, begins with another insight from Stinchcombe (1965): organizations are imprinted by their founding environment. Over time, bounded rationality and structural inertia make it difficult for organizations to remain aligned with shifting environmental demands. Encountering a series of environmental shifts may thus expose aging organizations to a risk of obsolescence. Aging may also bring about senescence, a buildup of internal politics and precedent that impede action. Obsolescence and senescence pose separate risks. Senescence is a direct effect of aging; obsolescence a result of environmental change.

Among studies controlling for time-varying size, empirical estimates are mixed (Baum and Shipilov, 2006). The divergent results might reflect the effect of age varying with that of size, masking 'true' age effects or alternatively variation in age dependence across populations. To resolve such questions, research must test alternative models' mediating variables directly rather than using age as a proxy for all underlying constructs (Baum, 1996). Such an approach treats the alternatives as complementary rather than (falsely) as competing models (*see* Development: Organizational).

Organizational Change

Research on organizational change typically concentrates on content: a change to a more (less) advantageous arrangement is considered adaptive (detrimental) (*see* Organizational Change; Development: Organizational). Complementing this focus, ecological research emphasizes the change process, raising two basic questions.

How Changeable Are Organizations?

The 'structure' in structural inertia theory refers to core organizational features (goals, authority structures, technical core, and market strategy) related to mobilizing scarce resources for beginning an organization and then maintaining resource flows (Hannan and Freeman, 1984). Peripheral features buffer an organization's core from uncertainty. Core features exhibit greater inertia than peripheral features; but the claim is not that core features do not change, only that they change slower than the environment. Hannan et al. (2003) portray structural inertia as a cascade of organizational changes following an initial change. The time spent reorganizing and associated opportunity costs, which increase with the complexity and sluggishness of the change, generate structural inertia. Ecology's inertia is thus a relative concept; adaptation and learning are not excluded, just not taken-for-granted. Nor does structural inertia theory remove individuals from responsibility for their organization's survival. Indeed, it implies that the capacity of individuals to change organizations successfully is of utmost importance.

Although not included in structural inertia theory, a complete understanding of change processes also requires consideration of organizations' change histories (Amburgey et al., 1993). The more experienced an organization becomes with a particular change, the more likely it is to repeat it. If the change becomes linked with success in decision-makers' minds – rightly or wrongly – repetition is more likely still. The change process may thus itself become routinized, creating the tendency to maintain direction and emphasis of prior actions in current behavior. To reconcile such 'momentum' with commonly observed periods of organizational inactivity, Amburgey et al. contend that local search reinforces momentum immediately after a change occurs, but it subsequently declines.

Empirical support for structural inertia theory remains equivocal; support for momentum is stronger (Baum and Shipilov, 2006). Overall, research indicates that, while constrained by inertia and momentum, organizations often initiate change in response to environmental shifts. Of course, inertia and momentum are not necessarily harmful: in addition to promoting reliability and accountability, in an uncertain environment they prevent organizations from responding too quickly and frequently to environmental change. Ultimately, whether inertia and momentum are adaptive depends on the hazardousness of change.

Is Change Beneficial?

According to structural inertia theory, core change disrupts established routines and exchange relations, and undermines organizational legitimacy by impairing reliability and accountability. Organizations may thus fail as a result of their

attempts to survive – even attempts that might ultimately have proved beneficial. Because their structures, routines, and relationships are more institutionalized, older organizations are especially prone to experience disruption following core change. In contrast, large organizations can buffer themselves from disruption by, for example, maintaining both old and new routines during transition or overcoming resource deprivations accompanying the change. If an organization survives the short-run disruption, its failure rate is predicted to decline as reliability, exchange relations, and legitimacy are reestablished over time. Only if the rate of decline is faster than before the change, however, will the organization ultimately benefit from taking the short-term risk.

Empirical studies indicate that while core change is not necessarily disruptive in the short run, neither do organizations necessarily improve their survival chances in the long run (Baum and Shipilov, 2006). Two research design problems, however, appear to produce systematic underestimates of the hazardousness of change. First, the annual data typically used may not detect the deadliest core changes – those quickly proving fatal. Second, because organizations founded before a study period are not observed when they are youngest, smallest, and most vulnerable to change, their inclusion likely biases downward the risk of change.

Organizational Learning

Complementing the growing interest in organizational change and its consequences is an emerging stream of research exploring the effects of organizations' experiential learning on their survival chances (*see* Learning: Organizational).

The idea that experiential learning improves efficiency is not inconsistent with the possibility that, simultaneously, overall organizational effectiveness decreases if the organization does not adjust to new demands (March 1991). Consistent with this observation, Baum and Ingram (1998) find support for a U-shaped effect of 'operating experience' (i.e., accumulated unit operating years) on organizational failure rates. Initially, the rate of failure decreases with experience as the organization moves down the learning curve for its favored set of routines. Eventually, however, accumulating experience causes the organization to become less able to adapt and, as the demands of the environment change, more likely to fail. In complementary work, Barnett and Hansen (1996) study 'competitive experience' (i.e., aggregate number of years an organization faced competition from rival organizations), differentiating between recent and distant competitive experience, expecting (and finding) that recent competitive experience would be beneficial, while older competitive experience would be detrimental.

Ecological studies of organizational learning advance our understanding of the role of experiential processes in organizational adaptation and selection. Experiential learning is a double-edged sword – it can enhance organizational survival chances while at the same time hindering adaptation; it can strengthen competitive abilities while at the same time making competition harder. Moreover, this research contributes to answering the question of how aging influences organizational failure by modeling more directly experiential learning constructs invoked by liability of newness and aging hypotheses.

Ecological Processes

Niche-Width Dynamics

Niche-width theory (Hannan and Freeman, 1977) focuses on two aspects of environmental change – variability and grain – to explain the differential survival of specialists, those who possess few slack resources and concentrate on a narrow range of customers, and generalists, those who attempt to appeal to the mass market and exhibit tolerance for more varied environments. Variability refers to the size-related environmental fluctuations over time. Grain refers to the patchiness of these fluctuations, with fine-grained fluctuations being frequent and coarse-grained infrequent. The key prediction is that in fine-grained environments, with large-magnitude fluctuations, specialists ride out the fluctuations, outcompeting generalists who are unable to respond quickly enough to operate efficiently. Niche-width theory thus challenges the classic prediction that uncertain environments always favor generalists that spread their risk (*see* Structural Contingency Theory).

In contrast to niche-width theory, which implies an optimal strategy for each population, Carroll (1985) proposes that, in environments characterized by economies of scale, competition among generalists to occupy the center of the market frees peripheral resources that can be used by specialists; a process he refers to as resource partitioning. His model predicts that in concentrated markets specialists can exploit more resources without engaging in direct competition with generalists, and by implication, while increasing market concentration increases the failure of generalists, it lowers the failure of specialists.

Although the specialist–generalist distinction is common in ecological research, tests of niche-width theory are scarce, and resource-partitioning studies do not typically contrast niche-width and resource-partitioning predictions. Yet, there is a conflict between these two theories. They offer distinct views of specialism and generalism and the tradeoff between organizational survival and niche width. The theories also focus on different kinds of niches, niche-width theory emphasizing *fundamental niches* (i.e., the set of environments in which an organization can survive in the absence of competition), and resource-partitioning theory emphasizing *realized niches* (i.e., the subset of the fundamental niche in which an organization can survive in the presence of competitors). It is thus unclear whether the two theories differ fundamentally or can be integrated into a unified theory. However, there have been early efforts to link the two theories (Dobrev et al., 2001) by combining niche-width theory's emphasis on environmental dynamics with resource partitioning's emphasis on scale advantages.

Population Dynamics and Density Dependence

Ecological research on founding and failure has given considerable attention to endogenous population dynamics and density dependence. Prior population dynamics – patterns of founding and failure – shape current founding rates (Delacroix and Carroll, 1983). Initially, prior foundings signal opportunity to entrepreneurs, encouraging founding. But as foundings increase, resource competition increases, discouraging founding. Prior failures have

analogous effects. At first, failures release resources for entrepreneurs. But further failures signal a hostile environment, discouraging founding. Resources freed by prior failures may also enhance the viability of established organizations, lowering failures (Carroll and Delacroix, 1982). Density-dependent explanations for founding and failure are similar, though not identical. Initial increases in the number of organizations can increase the legitimacy of a population, enhancing the capacity of its members to acquire resources. However, as a population continues to grow, competition with others for scarce common resources intensifies. Combined, the effects of legitimacy and competition imply an inverted U, that is, ∩-shaped relationship between density and founding and a U-shaped relationship between density and failure (Hannan and Carroll, 1992).

Studies of numerous organizational populations provide empirical support for density-dependent founding and failure; population dynamics findings are mixed and, when modeled together with population density weaker still (Baum and Shipilov, 2006). One explanation for the dominance of density dependence over population dynamics is the systematic character of density relative to the changes in density produced by founding and failure. A related account is the sensitivity of rate dependence to outliers. Regardless, population dynamics has been largely abandoned in recent research.

Despite strong support for density dependence, it has received critical attention for its (1) proposed integration of ecological and institutional perspectives, (2) assumption that each organization influences and is influenced by competition equally, (3) inability to explain the concentration common to older populations, and (4) inattention to problems of unobserved heterogeneity. Several innovative elaborations, respecifications, and measures, summarized in **Table 1**, have been advanced to address these questions.

Population Learning

Complementing the growing ecological interest in organizational learning is an emerging line of research exploring how organizations' collective experiences lay the foundation for population learning. Although organizational learning tends to focus on refining existing routines, at the population level diverse goals and incentives foster the proliferation of new ideas and routines. A population can thus encompass a great deal of variety without violating internal or external standards of consistency and reliability, and therefore more varied experience than an organization. Moreover, by observing their population, organizations can potentially learn a multiplicity of routines employed by other organizations. An organization's own experience is thus not the only opportunity for learning; organizations may also learn from the operating and competitive experience of others in their population.

Ingram and Baum (1997), for example, study the effects of population operating experience (i.e., unit operating experience of other organizations) accumulated both prior to and since an organization's founding. Controlling for organizational operating experience, both types of population operating experience significantly lowered organizational failure rates. Baum and Ingram (1998) attempted to replicate these findings and did so for population operating experience at founding, but not accumulated since founding.

Ingram and Baum (1997) also observed that organizational outcomes also depended on competitors' actions. Operating

Table 1 Elaborations of the density dependence model

Model	*Critiques*[a]	*Elaboration*	*Key references*
Density delay	3	Incorporates imprinting effect of density at founding	Hannan and Carroll (1992)
Mass dependence	2,3	Weights each organization's competitive effect by its size	Barnett and Amburgey (1990)
Multiform models	2	Estimates intrapopulation microstructures and interpopulation (community) macrostructures	Barnett (1990)
Level of analysis	2,3	Estimates variation in density dependence across levels of geographic aggregation (city, state, region, nation)	Hannan et al. (1995)
Institutional embeddedness	1,4	Estimates population legitimation with relational density (endorsements by powerful actors and organizations)	Baum and Oliver (1992)
Localized competition	2,3,4	Weights competitive effects of organizations on each other by their proximity (size, price, location)	Baum and Haveman (1997)
Organizational niche overlap	1,2,4	Disaggregates population density for each organization based on resource (non)overlaps with other organizations	Baum and Singh (1994)
Location dependence	2,4	Incorporates information on spatial structure of organizational populations	Lomi (1995)
Time variation	1,2,3	Estimates variation in density dependence as a population ages	Baum (1995)
Dynamic selection	2,3,4	Incorporates variation in the strength competitive processes as a population ages	Baum (1995)
Competitive intensity	2,3,4	Weights each organization's competitive strength by its age	Barnett (1997)
Scale-based competition	2,3,4	Benefits of size are a function of positional differences with larger, rather than all other organizations	Dobrev and Carroll (2003)

[a]Critiques. 1 = Proposed integration of ecological and institutional perspectives. 2 = Assumption that each organization influences and is influenced by competition equally. 3 = Inability to explain concentration common to older populations. 4 = Inattention to problems of unobserved heterogeneity.

experience can generate part of what is necessary for a model of competitors, but it is necessary to observe competitive outcomes. Measuring population competitive experience based on the number of failures observed by organizations, and controlling for organizational and population operating experience, they found that population competitive experience accumulated since an organization's founding lowered failure rates, but population competitive experience at its time of entry had no effect on failure.

These and related findings (e.g., Barnett, 1997; Sorenson, 2000) highlight the heterogeneity of organizational forms across time over successive cohorts of organizations. Population learning extends traditional ecological analyses, which emphasize that organizational change and variability reflect primarily inert organizations replacing each other, but not the implications of vicarious learning by new entrants and established organizations.

See also: Darwinism: Social; Development: Organizational; Entrepreneurship and Economic Growth; Hierarchies and Markets; Learning: Organizational; Organization: Overview; Organizational Ecology; Organizations, Sociology of.

Bibliography

Amburgey, T.L., Kelly, D., Barnett, W.P., 1993. Resetting the clock: the dynamics of organizational change and failure. Administrative Science Quarterly 38, 51–73.

Barnett, W.P., 1990. The organizational ecology of a technological system. Administrative Science Quarterly 35, 31–60.

Barnett, W.P., 1997. The dynamics of competitive intensity. Administrative Science Quarterly 42, 128–160.

Barnett, W.P., Amburgey, T.L., 1990. Do larger organizations generate stronger competition. In: Singh, J.V. (Ed.), Organizational Evolution: New Directions. Sage, Newbury Park, CA, pp. 78–102.

Barnett, W.P., Hansen, M., 1996. The red queen in organizational evolution. Strategic Management Journal 17 (S), 139–157.

Barnett, W.P., Sorenson, O., 2002. The red queen in organizational creation and development. Industrial and Corporate Change 11, 289–325.

Barron, D.N., West, E., Hannan, M.T., 1994. A time to grow and a time to die: growth and mortality of credit unions in New York city, 1914–1990. American Journal of Sociology 100, 381–421.

Baum, J.A.C., 1995. The changing basis of competition in organizational populations: evidence from the Manhattan hotel industry, 1898–1990. Social Forces 74, 177–205.

Baum, J.A.C., 1996. Organizational ecology. In: Clegg, S., Hardy, C., Nord, W. (Eds.), Handbook of Organization Studies. Sage, London, pp. 77–114.

Baum, J.A.C., Haveman, H.A., 1997. Love thy neighbor? Differentiation and agglomeration in the Manhattan hotel industry. Administrative Science Quarterly 42, 304–338.

Baum, J.A.C., Ingram, P., 1998. Survival-enhancing learning in the Manhattan hotel industry. Management Science 44, 996–1016.

Baum, J.A.C., Oliver, C., 1992. Institutional embeddedness and the dynamics of organizational populations. American Sociological Review 57, 540–559.

Baum, J.A.C., Oliver, C., 1996. Toward an institutional ecology of organizational founding. Academy of Management Journal 39, 1378–1427.

Baum, J.A.C., Shipilov, A.V., 2006. Ecological approaches to organizations. In: Clegg, S., Hardy, C., Nord, W., Lawrence, T. (Eds.), Handbook of Organization Studies, second ed. Sage, London, pp. 55–110.

Baum, J.A.C., Singh, J.V., 1994. Organizational niche overlap and the dynamics of organizational founding. Organization Science 5, 483–501.

Carroll, G.R., 1985. Concentration and specialization: dynamics of niche width in populations of organizations. American Journal of Sociology 90, 1262–1283.

Carroll, G.R., Delacroix, J., 1982. Organizational mortality in the newspaper industries of Argentina and Ireland: an ecological approach. Administrative Science Quarterly 27, 169–198.

Cattani, G., Pennings, J.M., Wezel, F.C., 2003. Spatial and temporal heterogeneity in founding patterns. Organization Science 14, 670–685.

Delacroix, J., Carroll, G.R., 1983. Organizational foundings: an ecological study of the newspaper industries of Argentina and Ireland. Administrative Science Quarterly 28, 274–291.

Dobrev, S., Carroll, G.R., 2003. Size (and competition) among organizations: modeling scale-based selection among automobile producers in four major countries, 1885–1981. Strategic Management Journal 24, 541–558.

Dobrev, S., Kim, T.-Y., Hannan, M., 2001. Dynamics of niche width and resource partitioning. American Journal of Sociology 106, 1299–1338.

Fichman, M., Levinthal, D.A., 1991. Honeymoons and the liability of adolescence: a new perspective on duration dependence in social and organizational relationships. Academy of Management Review 16, 442–468.

Hannan, M.T., Carroll, G.R., 1992. Dynamics of Organizational Populations: Density, Competition, and Legitimation. Oxford University Press, New York.

Hannan, M.T., Carroll, G.R., Dundon, E.A., Torres, J.C., 1995. Organizational evolution in a multinational context: entries of automobile manufacturers in Belgium, Britain, France, Germany, and Italy. American Sociological Review 60, 509–528.

Hannan, M.T., Freeman, J., 1977. The population ecology of organizations. American Journal of Sociology 83, 929–984.

Hannan, M.T., Freeman, J., 1984. Structural inertia and organizational change. American Sociological Review 49, 149–164.

Hannan, M.T., Pólos, L., Carroll, G.R., 2003. Cascading organizational change. Organization Science 14, 463–482.

Ingram, P., Baum, J.A.C., 1997. Opportunity and constraint: organizations' learning from the operating and competitive experience of industries. Strategic Management Journal 18 (S), 75–98.

Lomi, A., 1995. The population ecology of organizational founding: location dependence and unobserved heterogeneity. Administrative Science Quarterly 40, 111–144.

Sorenson, O., 2000. The effect of population level learning on market entry: the American automobile industry. Social Science Research 29, 307–326.

Stinchcombe, A.L., 1965. Social structure and organizations. In: March, J.G. (Ed.), Handbook of Organizations. Rand McNally, Chicago, pp. 153–193.

Stuart, T.E., Sorenson, O., 2003. The geography of opportunity: spatial heterogeneity in founding rates and the performance of biotechnology firms. Research Policy 32, 229–253.

E-Communications Platforms and E-Learning

Christopher R Wolfe and Elizabeth M Cedillos, Miami University, Oxford, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

E-learning is the use of digital tools for learning. Learning management systems and distance education are among the most prevalent tools. However, hybrid experiences and collaborations are changing the E-learning landscape. Recent developments include the advent of social networking and online learning communities, the ubiquitous presence of smart phones, and an increased recognition of the potential for computer games to transform learning. Other important developments include advances in Intelligent Tutoring Systems, the free Open Educational Resources movement, and the creation of immersive environments allowing users to engage with virtual environments and to digitally augment their experience of the real world.

E-learning can be defined as the use of electronic tools to support teaching and learning. E-learning can be experienced in many formats. These formats and delivery methods include stand-alone computers, the Internet, and satellites with learning materials provided by an instructor, an animated intelligent agent, or through media such as text, images, sounds, and video. E-learning can take place in practically any location including: classrooms, homes, and in remote field locations. Additionally, E-learning tools can be used at the learner's own pace. Most people with adequate resources and some understanding of technology can engage in E-learning by one means or another. E-learning has been used in public and private primary, secondary, and postsecondary schools throughout the United States and around the world. Additionally, homeschoolers, government agencies, militaries, and private businesses have utilized E-learning as learning and training tools.

'Traditional' E-Learning and E-Communication

Most colleges and universities in the world have made significant use of E-learning technologies since before the turn of the century (Wolfe, 2001a). In recent years for-profit institutions such as the University of Phoenix have become familiar fixtures on the higher education landscape. For example, the University of Phoenix offers 40 complete online Bachelors Degrees ranging from a Bachelor of Arts in English to an RN to Bachelor of Science in Nursing (University of Phoenix, 2012). The most prevalent tools for these applications are learning management systems and distance education platforms.

Learning Management System

Learning management systems are software applications used by educational systems as an online method to give students access to course content, and for educators to track, develop, and integrate relevant course information (Ellis, 2009). With only a computer and Internet connection, students can access relevant course content, grades, email, and class discussion boards or forums (Ellis, 2009). Online courses use learning management systems as their main information delivery method (Ellis, 2009). A limitation of these systems is that because the system itself does not teach, learning from the system cannot occur (Ellis, 2009). Learning management systems are simply online tools that can be used as a method to deliver information.

Blackboard Learning System

One of the most popular learning management systems was developed in 1997, it is called Blackboard Learning System (Bradford et al., 2007). Blackboard, like most learning management systems, offers a way for students and educators to communicate and deliver content (Bradford et al., 2007). Communication features include: announcements that all students in a course can read, an online chat feature, discussion boards, and email (Bradford et al., 2007). Content in the Blackboard system includes: course content such as articles or assignments, calendars to remind student about due dates, learning modules mostly utilized by online only courses, assessments such as quizzes and exams, assignments, a grade book, and a media library (Bradford et al., 2007). While largely used by learning institutions, Blackboard has also been used by corporations and government organizations. Blackboard also has other product platforms including: Blackboard Collaborate, Blackboard Mobile, Blackboard Connect, Blackboard Transact, and Blackboard Analytics (Bradford et al., 2007).

Sakai

Another common learning management system is called Sakai. Developed in 2005, this open source educational software is offered as a free community resource under the Educational Community License and has over 350 educational organizations that use the system. Sakai's collaboration tools include announcements, dropbox, email and archives, resources, chat rooms, forums, message centers, news/RSS, a poll tool, presentation, profile of site participants, a search, and a schedule. Institutions using open source software, rather than proprietary products have the ability to tailor the learning management system to their own particular needs.

Distance Learning

Distance education emphasizes delivering education via technology to students who are not able to be physically present in a traditional classroom. There are full distance learning courses, where teacher and student will never meet and all education is

carried out over the Internet, and courses where occasionally students and teachers might meet up to do such things as take an exam. Students also have the option of enrolling full time in distance learning classes or only taking a few classes toward their degree. There are many types of institutions that offer distance learning classes including: nonprofit, for-profit, accredited, unaccredited, and even doctoral programs.

Synchronous and asynchronous learning methods are both used in distance education. With synchronous learning, a class will 'meet' at a predetermined time via web conferencing, videoconferencing, live streaming video, television via direct-broadcast satellite, telephone, and web-based Voice-over-Internet Protocol (VoIP; Burns, 2011). Synchronous classes are carried out as if it were traditional classrooms, only the teacher and the students are not be in the same room (Burns, 2011). With the asynchronous learning method, students can access class materials whenever they so choose (Burns, 2011). Teachers typically post materials on the classes learning management system site for students to access. Students can then download and upload material, participate in message board forums, email other students or the teacher, and access print and audio material (Burns, 2011).

Open Educational Resources

Open Educational Resources are free online media and documents that are used for teaching and learning. These resources are delivered via platforms that allow creation and delivery of educational materials including college courses and materials (Caswell et al., 2008). Students taking courses in this manner do not pay tuition or receive credit, nor are degrees awarded. These resources are funded by universities and donations. The Open Educational Resources movement, whose goal is to foster collaboration, knowledge, and education across the globe, is one source of support for open educational resources (Caswell et al., 2008). Open Education has drawn the interest of journalists in recent years, in part because some of the most respected universities in the world have become active in the massive online open courses movement.

In 2001 the MIT initiative, called MIT OpenCourseWare, was the first to offer open educational resources, sparking the Open Educational Resources movement (Caswell et al., 2008). MIT OpenCourseWare started offering their entire course catalog online in 2002 for no cost. Currently MIT has over 2080 courses available online; the material available varies from reading lists, discussion topics, homework problems and exams, lecture notes, and streaming video of lectures. International members of OpencourseWare include: Brazil, Mexico, China, India, Japan, Germany, and the Netherlands (Caswell et al., 2008)

The Stanford Engineering Everywhere (SEE) initiative, started by Stanford University, is a program that offers several online courses taught by Stanford professors free of charge. These classes have instructional videos, reading lists, assignments, class lectures, and are offered globally (Van Orsdel, 2012). As of 2012, there are 10 courses being offered from Stanford's School of Engineering (Van Orsdel, 2012). The SEE encourages students to use the course material for any purpose that fits their needs. Additionally, in an attempt to foster collaboration the SEE encourages educators to use their lectures, handouts, notes, or tests in their own classrooms (Van Orsdel, 2012).

There are currently over 200 institutions around the world that are part of the OpenCourseWare Consortium. These include Europe's Learning Resource Exchange for schools, Washington State's Open Course Library Project, California Learning Resources Network, the China Open Resources for Education, OpenCourseWare in Japan, and numerous initiatives to digitize textbooks for all grade levels by countries such as Bangladesh, Uruguay, South Korea, South Africa, and Saudi Arabia (Caswell et al., 2008).

There are also open educational resources for grades K-12. Curriki is nonprofit organization that provides an Internet site for open source curriculum for primary school students that allows universal access to relevant curriculum and other instructional materials (Kurshan, 2008). The Curriki community provides its members educational materials such as digital textbooks, learning videos, and other interactive resources that are peer-reviewed as adhering to traditional educational standards (Kurshan, 2008). The active leadership of highly respected institutions, and peer review, have held in check growing perceptions that E-learning is becoming a second rate alternative to more expensive traditional colleges and universities.

Hybrids, Collaborations, and Partnerships

E-learning is both transforming and challenging the organizational infrastructure of education. Organizations such as the Serious Games Initiative at the Woodrow Wilson Center for International Scholars and the iCampus initiative at MIT have brought together educators, multi-media experts, programmers, game makers, policy makers, and scientists in surprising and potentially fruitful configurations unimagined just a few years ago. These hybrids, collaborations, and partnerships are having a significant impact on both formal and informal education.

The Difference between Informal and Formal Education

Informal education is education that takes place outside of a formal school setting, as opposed to formal education which is confined to classroom-based education with trained teachers. While formal education is set with a curriculum of set courses or content, informal education is not limited to classrooms. A person usually seeks out information that they want to learn about and can do so in any manner that they choose. This includes books, museums, libraries, zoos, informal meeting groups, magazines, and more recently E-learning media including the Internet, software, television, and video games (Greenfield, 2009). The web and E-learning tools easily and readily lend themselves as opportunities for informal education (Wolfe, 2001b). Users can access E-books, software and applications, informational websites such as Wikipedia, and video games at any time they choose so long as they have the proper tools (Greenfield, 2009). Learning tools built with common applications such as spreadsheets have been created for uses such as learning statistics, and are often available

online free of charge (Fisher and Wolfe, 2012). Computers, tablets, smartphones, E-readers, and video game consoles all allow users access to these resources. These resources lend themselves to development of spatial reasoning skills such as visualization, development of problem solving and critical thinking skills, and imagination (Greenfield, 2009).

Project Dragonfly

Project Dragonfly promotes inquiry-driven learning and educational reform, reaching people around the world through learning media, exhibits, and education programs. Project Dragonfly represents an evolving set of collaborations between a university and a variety of institutions worldwide including zoos, national parks, government ministries, nongovernmental organizations, and not-for-profit organizations. One dimension involves participatory media including the Dragonfly Web pages, perhaps the first website devoted to children and scientific inquiry (Wolfe et al., 2001). DragonflyTV, an inquiry-based public television program for children; and Wild Research Stations, interactive inquiry-based experiences at aquariums and zoos designed to increase engagement in science and conservation by involving families in research and conservation action. In addition, Project Dragonfly offers 'Earth Expeditions,' field courses around the world, online training for teachers, and Masters Degree Programs.

Through Project Dragonfly university students in traditional face-to-face classrooms have created web-based courseware for children, and students and teachers in hybrid or blended courses engage in online learning experiences culminating in collaborative face-to-face international field experiences in parks and preserves around the world. These collaborations blur the boundaries between informal and formal education, E-learning and fieldwork, and force traditional institutions to adapt to a rapidly changing landscape by embracing hybrid learning experiences, negotiating agreements among organizations, and acknowledging professional activities that challenge conventional notions of scholarship and teaching.

iCampus

The iCampus initiative was a partnership between Microsoft and the Massachusetts Institute of Technology (MIT) launched with the goal of improving technology-enabled teaching models and the development of new educational tools. The iCampus initiative set out to develop methods to allow users to remotely access laboratory instruments, create tools to exchange educational materials, provide access to distance learning and mentoring, and creates web-based virtual museums. Though the initial iCampus initiative concluded in 2006, there is now a project called the Outreach Initiative whose goal is to disseminate the educational technologies developed from the iCampus initiative. The iLabs project allows users to access online remote laboratories and conduct experiments that they might otherwise not be able to, given their limited resources. The Cross Media Annotation System (xMAS) enhances the use of video and images for courses that reference visual materials often. XTutor is a program that has audio recordings and transcripts of lectures, homework problems, and an intelligent tutoring agent online for introductory computer science and Artificial Intelligence classes. Another project is iMOAT, a web-based service that allows for grading of writing assignments and exams. Technology-Enabled Active Learning (TEAL) allows for an enriched learning experience by combining lectures, simulations, collaboration, and hands-on access to experiments from a student's desk. Finally, Peer Review Evaluation Process (PREP) is meant to be used by teams of four to six individuals so they can develop and share ideas with one another; for example, Mechanical Engineering students have used PREP to design and build robots for competitions. Tools such as these have the potential smudge, the line between teaching and research at large and small institutions around the world.

Online Learning Communities

Online learning communities are communities whose members connect via the Internet to engage in peer-to-peer learning. In these virtual-communities, a topic area will be defined and discussed amongst members to gain knowledge of the subject matter. Information can be shared through any Internet supported medium including: audio, video, text, pictures, Internet links, and virtual files. Community members can upload and download files as needed.

Social Networks

Social network sites such as Facebook, Twitter, and Google+ allow users to interact via the Internet with anybody across the globe. Facebook, the largest social networking site boasts over 1 billion registered users worldwide (Vance, 2012), and educators are turning to these sites to connect with their students. Teachers can 'post' relevant material to a social networking site and students can view the information and further explore the content (Davis, 2010).

This medium allows for a different and flexible learning experience for students. The use of social networking in education can be beneficial because students are continuously (and willingly) exposed to these networks, giving them easier, faster, and more convenient access to class material (Davis, 2010). Additionally, it allows students who would not otherwise participate in the classroom to engage in the content from the safety of their computer (Davis, 2010). Education-focused social networks, such as LiveMocha, provide instruction in foreign languages in a collaborative peer-to-peer learning environment. LiveMocha boasts over 15 million users, and feedback is 'crowd-sourced' such that users who want feedback must provide feedback to other members of the online community.

Podcasts

Podcasts are digital media files that are streamed online (Ovadia, 2007). Typically, podcast are delivered via an audio recording, but can also be delivered via video. Usually podcast are continuous series, covering a wide range of topic areas. New podcasts can be added periodically (Ovadia, 2007). By subscribing to a series, the user can automatically download the newest podcast file and can then listen to it via their computer or upload it to their portable media player or other device.

Podcasts can be easily utilized for educational purposes, such as lessons on physics, mathematics, psychology, history, or literature. Learners can download a podcast and listen to the lesson at their leisure. College professors have used podcasts as a way to provide students with an additional study aid to review a lecture or listen to one they may have missed (Ovadia, 2007).

Video Sharing Websites

Video sharing websites, such as YouTube, allow users to upload video clips to a website so that others may view it online. The content of these videos are wide ranging and sometimes sophomoric. However, individuals post and view a myriad of instructional and educational videos every day (Buzzetto-More, 2012). For example, there are 'how to' videos on how to tie a Windsor knot, how to cook turkey, and how to replace an ignition coil in a Buick LeSabre. TED talks are a popular series of brief lectures presented as video clips, with over 1400 talks on topics ranging from 'aid vs. trade' to 'the psychology of evil', all videos are available online. Educators can utilize video sharing websites to post lectures for students, provide supplemental content, or direct students toward another educator's videos (Buzzetto-More, 2012).

Blogs

Blogs, originally a combination of the words 'web log,' are informational websites that are usually written by one individual or small groups of individuals on a single topic. Visitors of blogs can sometimes leave messages, making the blogging experience more interactive (Luo and Gao, 2012). Most blogs provide text but other media such as images, videos, audio, and links to other websites can also be found. If a blog is used for educational or instructional purposes, it is sometimes referred to as an 'edublog.' A nonstudent user may use a blog to gain information about a subject. Whereas a student or educator may use a blog to provide instructions, ask questions, build on ideas, share tips and information, and as a forum for relevant discussions (Luo and Gao, 2012).

Internet Forums

Internet forums, or more commonly called message boards, are online discussion websites where individuals can have conversations by posting messages on a 'thread.' Similar to other online learning communities, a forum can range in the content it covers; threads generally target topics that are more specific. Postsecondary schools commonly use forums for individual classrooms, allowing students to have discussions, ask questions, and interact, with not only one another but their professor as well.

Wikis

A wiki is a website that is created and developed by multiple users working in a collaborative effort to add, change, or delete content. Wikis are web pages whose sole purpose is to provide relevant information to any given topic area. The largest Wiki is Wikipedia with 23 million articles in 285 languages, including over 4 million articles in English produced by anonymous volunteers from around the world. One study suggested no differences in, the rate of 'serious errors' between Wikipedia and a professionally produced encyclopedia (Giles, 2005) though this is still a matter of controversy. Recently, educators and their students have become involved in creating Wikipedia entries. For example, the Association for Psychological Science has involved more than 2000 psychological scientists and their students in the APS Wikipedia Initiative, "to represent scientific psychology as fully and as accurately as possible and thereby to promote the free teaching of psychology worldwide." The design of a wiki allows users to browse through a page and skip to other relevant topic areas through hyperlinks. Although Wikis are designed for collaboration, they do not easily lend themselves to social networking, but they provide an overview of relevant information (Buzzetto-More, 2012).

Mobile Learning Platforms

It has been suggested that mobile computing is the future of E-learning (Pierce, 2010; Schaffhauser, 2010). The creation of e-books, tablets, and the ubiquitous presence of smart phones are making mobile computing devices the platform of choice – especially for 'digital natives,' i.e., those of the generations who have never known a world without the Internet and an array of computer technologies (Halpern et al., 2012).

Mobile Computing

Mobile computing is the use of an easily transported computer such as a tablet that the user interacts with regularly throughout the day. This encompasses the use of mobile computing, hardware, and software. Mobile computers are capable of doing just about anything one can do from a desktop computer. Capabilities include Internet and email access, editing and accessing documents and images, playing games, and accessing applications and other programs. There are drawbacks to mobile computing that can make it somewhat harder to use when compared to a personal computer. Limitations include slower and less reliable Internet access, lower security, power consumption that exceeds the capacity of batteries, and smaller workspace. Some of the most popular methods of mobile computing are smart phones and tablets. Other mobile computing devices include portable computers, personal digital assistants (PDA), and graphing calculators.

Recently, educators have been encouraged to utilize smart phones in their primary school classrooms to engage students in learning. The benefits to using these technologies include an increase in student engagement, more accessibility, and larger student interest (Swan et al., 2005). Project K-Nect is one program that gives students access to smart phones and mobile applications that can explain algebraic principles and allows them to seek advice from their peers via text or instant messages; the program has shown improvements in student's academic success by 30% compared to their peers who did not use the program (Pierce, 2010). In addition, informal learning environments such as LiveMocha increasingly have versions of their entire suite of resources optimized for smart phone platforms.

Tablets and E-Books

Tablets are portable personal computers that come in many sizes and are usually operated with a touchscreen. Tablets come with an on-screen keyboard and the user uses their finger as a mouse, some tablets allow the user to connect these as external components. While today's tablets run on an operating system, initial attempts to make the tablet a portable replica of larger personal computers were not successful, and the trend has arguably been to make desktop computers more like tablets. Tablets utilize a touch user interface and mobile applications, called 'apps.' Many manufacturers, such as Samsung, HTC, Amazon, Microsoft, Google, and Apple, now produce tablets. Because of the commercial success of tablets, manufacturers can offer competitive pricing, making tablets affordable, widely available, and an acceptable alternative to a laptop or personal computer.

Because of the convenience and affordability of tablets, they are now being used in classrooms to replace textbooks and for general use by students (Schaffhauser, 2010). One benefit of using tablets in lieu of textbooks is having all books easily accessible and in one place. An additional benefit is the multifaceted uses tablets offer. Tablets allow students to become immersed in a learning environment. For example, students can quickly access the Internet, which provides vast amounts of digital content and converged media (Schaffhauser, 2010). Because tablets are essentially computers, they can be specialized with applications for children with learning disabilities. Schools in Africa have shown an improvement in learning outcomes over the course of one term when students were given tablets (The Economist, 2012).

An E-book, or electronic book, is a book in digital form that is usually read on an e-book reader or tablet, but can also be read on a computer or another electronic device that supports e-books. E-books can be bought online through a number of providers including: Amazon and the Apple eBookStore. Popular e-book readers include Amazon Kindle, Nook, Apple iPad, and Sony Reader. E-books can be stored on e-book readers or saved indefinitely in a server provided by the respective e-book store and accessed instantly when needed via an Internet connection.

E-readers and e-books provide many features that make owning them advantageous to having a physical copy of a book, especially in education. E-book websites can translate e-books into different languages if it has not already been done with a printed book (Harris, 2009). Additionally, some E-readers can read text aloud for those who are visually impaired, or otherwise prefer listening (Harris, 2009). E-readers also provide the ability to read the definition of a word by clicking on it, and allow for highlighting, bookmarking, and annotating texts and to see which passages other users have highlighted. Some E-readers allow readers to conduct searches for terms found in e-books, and look up more information about a topic using the same device (Harris, 2009).

Intelligent Tutoring Systems

One-on-one human tutoring has been found to produce dramatic gains in learning and Intelligent Tutoring Systems (ITS) employ Artificial Intelligence programming techniques with the goal of creating an ITS that behaves as a human tutor. Intelligent tutoring systems interact with people via dedicated computers or recently on the web. ITS can provide information, give examples, ask deep level reasoning questions (Craig et al., 2006), answer questions, explain answers, and give practice problems and demonstrations. Similar to human tutors, an ITS can make assessments about the knowledge a user has on the subject matter. This is useful because it allows the ITS to provide appropriate feedback and knowledge when it is most critical to the user. Furthermore, an ITS can scaffold learning, providing a good deal of support early in the learning process and remove support appropriately as learning progresses so that learners can build on their knowledge and skills in context (Bouyias and Demetriadis, 2012). Recently, there has been research on the automatic detection of learner emotions with an eye toward creating affect sensitive ITS (Graesser, 2011).

Intelligent tutoring systems have been developed for use in education, industry, and the military. Intelligent tutoring systems address a wide range of knowledge domains (Graesser et al., 2005b) including mathematics (Aleven et al., 2009), reading comprehensions (Mostow and Beck, 2007), argumentation (Bouyias and Demetriadis, 2012; Wolfe et al., 2009), physics (Gertner and VanLehn, 2000), health (Wolfe et al., 2013), legal reasoning (Pinkwart et al., 2006), and computer literacy (Graesser et al., 2003). Studies have found that some ITSs show improvements in learning amongst students compared to classroom only learning. Additionally, students have reported having better comprehension, being more engaged, having a more positive attitude, being more motivated, and having a positive effect on academic outcomes after working with an intelligent tutoring system (Graesser et al., 2003).

Some of the most effective ITSs are those that use mixed initiative natural language dialogues such as AutoTutor (Graesser et al., 2005a). Natural language dialogue means that dialogue is conducted in a human language such as German or English, and mixed initiative means that either the ITS or the learner may initiate a strand of conversation (D'Mello et al., 2010). In most ITS with natural language processing the learner types into a dialogue box (Wolfe et al., 2012). However, research is being conducted on completely oral dialogues (D'Mello et al., 2011). Intelligent tutoring systems that employ natural language dialogues include: CIRCSIM-Tutor, an ITS that helps students learn about blood pressure; Why2 Atlas, a physics ITS; and AutoTutor, an ITS that has been used to teach physics, computer literacy, and critical thinking skills (Graesser et al., 2005a).

Recently, ITS have expanded to web-based platforms, dramatically expanding the potential of ITS to have a broader impact on education. For example, in the city of Pittsburgh, Pennsylvania, USA, researches are bringing the ITS out of the lab and into the community, experimenting with the Andes Physics Tutor as an intelligent homework helper for physics student in local high schools (Katz et al., 2007). AutoTutor Lite is the first cross-platform web-based ITS that engages in two-way dialogues with people in natural language (Wolfe et al., 2012). Talking animated agents converse with learners using Latent Semantic Analysis to 'understand' what they are saying and respond appropriately.

Immersive Environments

Immersive environments allow the user to fully experience a real or artificial environment (Burns, 2011). Because immersive environments are highly engaging, a user can have a rich experience. Immersive learning environments build upon the idea that students can learn better by doing rather than seeing (Dickey, 2005). There are many types of immersive environments ranging from virtual environments to augmented reality.

Virtual Environments

Virtual Environments or Virtual Reality is the immersion of a user in an artificial computer-simulated environment that is designed to emulate the look and feel of being in a real environment. Importantly, virtual environments afford actions – which can form the basis of experiential learning (Dickey, 2005). Virtual environments are rendered on a computer screen, projected onto external surfaces, or experienced with a head mounted display, permitting the user to move about freely in a virtual environment (Waller et al., 2007). Ideally, users experience the feeling of 'presence' (Riva and Mantovani, 2012) in virtual environments. The US military has used virtual reality to train users how to fly planes and simulate how it would feel to jump from a plane. The creation of large-scale virtual environments permits users to engage in (and researchers to study) complex full motor behaviors such as fielding a batted baseball (Fink et al., 2009). Recent advances using the technique of redirected walking have enabled researchers to program virtual environments that are potentially unlimited in size (Hodgson et al., 2011). Thus, in theory, learners could explore ancient cities, other planets, or prehistoric landscapes on foot.

Augmented Reality

Augmented reality is the experience of actual environments that is supplemented by digital information in the form of images, sounds, and texts (Sumadio and Rambli, 2010). For example, the Star Walk app for smart phones and tablets allows users to identify over 20 000 objects in the night sky. Sky Walk takes advantage of the built in GPS and Internet connections to provide an interactive 360-degree touch control star map that displays constellations, stars, planets, satellites, and galaxies that are currently overhead. Engineering labs have used augmented reality to conduct lab practices remotely and in interaction with an actual lab (Borrero and Marquez, 2012). In short, augmented reality has the potential to provide learners with contextualized information in real time to enhance learning experiences in the field.

Learning Games

In recent years there has been an increased recognition of the potential for computer games to transform the way people learn (Halpern et al., 2012; Shaffer et al., 2005). Games can be highly motivating, yielding more 'time on task' (one of the biggest predictors of learning outcomes in the psychological literature). Moreover, commercial video games are driving many of the technological advances commercial computing. Games can let people inhabit new roles and new worlds and think and act in ways that would otherwise not be possible (Shaffer et al., 2005). Learning is essential to be a successful player of video games. As Brian Waniewski, managing director of Institute of Play notes, "Games create a need to know…You find things out as you progress toward your goal" (Greenemeier, 2012). Toward these ends, the creators of learning games must create meaningful contexts and constrain decisions-making tasks in ways that are productive and psychologically plausible (Wolfe, 2001c).

A recent example of a digital learning game is ARA (Acquiring Research Acumen), a computerized learning game that teaches scientific reasoning and critical thinking by having players interact with avatars and play games to identify research flaws, develop hypotheses, and assess the quality of research questions (Halpern et al., 2012). One strand of the work on games and learning involves 'serious games' (Millis et al., 2011; Westera et al., 2008). Serious games involve players in complex decision making, reasoning, and strategic problem solving in complex multiplayer environments that emulate real-world contexts (Westera et al., 2008).

Another tack has been to investigate learning in the context of popular video games to better understand the implications of emerging technologies and shifts in popular culture (Gee, 2007a) including skill development and processing speed (Dye et al., 2009). With digital video games, users have to solve problems by learning about a topic. There are many benefits to using video games as learning tools. First, video games help with the development of problem solving skills because users must complete tasks to accomplish a goal (Burns, 2011). Second, because of the technology involved in video games, they can keep track of information over long ranges of time, allowing a user to leave and come back to resume learning. Third, video games can provide the user with information they do not otherwise have, thus improving their knowledge and skill-base. Lastly, not only will video games provide users with information but they can also assess learning and provide feedback to users (Burns, 2011).

Another important avenue is the study of online communities of interest, for example players of games such as *World of Warcraft*, who spontaneously engage in collaborative learning and 'social knowledge construction' for purposes such as improving their playing abilities (Steinkuehler and Duncan, 2008). Some of the characteristics of these communities include meritocracy (valuing competency over credentials) open access, collaboration, competition, and flexibility (Squire et al., 2008). A third approach has been to create and study player communities organized around learning games such as Civilization III to promote the learning of academic topics such as world history (Squire et al., 2008). Indeed a case can be made that the study of learning games are shaping contemporary learning theory (Gee, 2007b; Squire et al., 2008) with a renewed interest in active engagement, response generation, individualization, spaced practice, reciprocal teaching, 'multiple retrieval cues' (Halpern et al., 2012) as well as situated learning and social context.

Summary and Future Trends

The topography of e-communication and e-learning is changing rapidly, and the rate of change is increasing dramatically. At the level of technology, mobile networked digital devices are becoming faster, more powerful, and more ubiquitous allowing e-learning to occur most anywhere. 'Big data' may provide new tools for tailoring learning experiences to individual and cultural differences in many contexts. Emerging discourse technologies are opening alternatives to the keyboard with voice, touchscreen, and gestures increasingly viable as modes of interactive learning. Progress in hardware continues to be faster than progress in software, and progress in both exceeds social and psychological theory. Institutions of learning at all levels are being shaken by these changes, yet the hyperbolae about the demise of brick and mortar schools and universities is largely unjustified. Counterevidence includes astoundingly high dropout rates in massive online open courses, and the observation that they are not supported by a viable business model. Although there are real sociocultural changes embodied in digital natives, genuine learning is still very hard work, which does not lend itself to quick technological fixes.

See also: Computer Games and Academic Achievement; Computer-Assisted Instruction; Distance Education; Educational Media: Potentials for Learning; Film and Video Industry; Information and Communications Technology in Education; Lifelong Learning and Its Support with New Technologies; Mobile Communication.

Bibliography

Aleven, V., McLaren, B.M., Sewall, J., 2009. Scaling up programming by demonstration for intelligent tutoring systems development: an open-access website for middle-school mathematics learning. IEEE Transactions on Learning Technologies 2 (2), 64–78. http://www.computer.org/portal/web/csdl/doi/10.1109/TLT.2009.22.

Borrero, A.M., Marquez, J.M.A., 2012. A pilot study of the effectiveness of augmented reality to enhance the use of remote labs in electrical engineering education. Journal of Science Education and Technology 21, 540–557. http://dx.doi.org/10.1007/s10956-011-9345-9.

Bouyias, Y.N., Demetriadis, S.N., 2012. iArgue: a web-based argumentation environment integrating collaboration scripting support and flexible fading techniques. In: Pinkswart, N., McLaren, B.M. (Eds.), Education Technologies for Teaching Argumentation Skills. Bentham Science Publishers, pp. 198–224.

Bradford, P., Porciello, M., Balkon, N., Backus, D., 2007. The blackboard learning system. The Journal of Educational Technology Systems 35, 301–314.

Burns, M., 2011. Distance Education for Teacher Training: Modes, Models, and Methods. Education Development Center, Inc. http://idd.edc.org/resources/publications/modes-models-and-methods.

Buzzetto-More, N.A., 2012. Social networking in undergraduate education. Interdisciplinary Journal of Information, Knowledge, and Management 7, 63–90.

Caswell, T., Henson, S., Jensen, M., Wiley, D., 2008. Open educational resources: enabling universal education. International Review of Research in Open and Distance Learning 9. http://www.irrodl.org/index.php/irrodl/article/view/469/1001.

Craig, S.D., Sullins, J., Witherspoon, A., Gholson, B., 2006. The deep-level reasoning-question effect: the role of dialogue and deep-level-reasoning questions during vicarious learning. Cognition & Instruction 24, 565–591.

D'Mello, S.K., Graesser, A., King, B., 2010. Toward spoken human-computer tutorial dialogues. Human-Computer Interaction 25, 289–323. http://dx.doi.org/10.1080/07370024.2010.499850.

D'Mello, S.K., Dowell, N., Graesser, A., 2011. Does it really matter whether students' contributions are spoken versus typed in an intelligent tutoring system with natural language? Journal of Experimental Psychology: Applied 17, 1–17. http://dx.doi.org/10.1037/a0022674.

Davis, M.R., 2010. Social networking goes to school. Education Week Digital Directions 3, 16–23. http://www.edweek.org/dd/articles/2010/06/16/03networking.h03.html?print=1.

Dickey, M.D., 2005. Three-dimensional virtual worlds and distance learning: two case studies of active worlds as a medium for distance education. British Journal of Educational Technology 36, 439–451.

Dye, M.G., Green, C., Bavelier, D., 2009. Increasing speed of processing with action video games. Current Directions in Psychological Science 18, 321–326.

Ellis, R.A., 2009. Students' Experiences of E-Learning in Higher Education: The Ecology of Sustainable Innovation. Routledge, New York.

The Economist, 8 December 2012. Digital Education in Kenya: Tablet Teachers. The Economist. http://www.economist.com/news/business/21567972-schools-africa-are-going-digitalwith-encouraging-results-tablet-teachers.

Fink, P.W., Foo, P.S., Warren, W.H., 2009. Catching fly balls in virtual reality: a critical test of the outfielder problem. Journal of Vision 9, 1–8.

Fisher, C.R., Wolfe, C.R., 2012. Teaching Bayesian parameter estimation, Bayesian model comparison, and null hypothesis significance testing using spreadsheets. Spreadsheets in Education (eJSiE) 5. http://epublications.bond.edu.au/ejsie/vol5/iss3/3.

Gee, J.P., 2007a. What Video Games Have to Teach Us about Learning and Literacy. Palgrave Macmillan, Basingstoke.

Gee, J.P., 2007b. Good Video Games and Good Learning: Collected Essays on Video Games, Learning and Literacy (New Literacies and Digital Epistemologies). Peter Lang Publishers, New York.

Gertner, A., VanLehn, K., 2000. Andes: a coached problem solving environment for; physics. In: Gauthier, G., Frasson, C., VanLehn, K. (Eds.), Intelligent Tutoring Systems: 5th International Conference. Springer-Verlag Berlin & Heidelberg GmbH & Co., Berlin, pp. 133–142.

Giles, J., December 2005. Internet encyclopedias go head to head. Nature 438, 900–901. http://dx.doi.org/10.1038/438900a.

Graesser, A.C., 2011. Learning, thinking, and emoting with discourse technologies. American Psychologist 66, 746–757.

Graesser, A.C., Moreno, K., Marineau, J., Adcock, A., Olney, A., Person, N., 2003. AutoTutor improves deep learning of computer literacy: is it the dialog or the talking head? In: Hoppe, U., Verdejo, F., Kay, J. (Eds.), Proceedings of Artificial Intelligence in Education. IOS, Amsterdam, pp. 47–54.

Graesser, A.C., Chipman, P., Haynes, B.C., Olney, A., 2005a. AutoTutor: an intelligent tutoring system with mixed-initiative dialogue. IEEE Transactions on Education 48, 612–618.

Graesser, A.C., McNamara, D.S., VanLehn, K., 2005b. Scaffolding deep comprehension strategies through point&query, AutoTutor, and iSTART. Educational Psychologist 40, 225–234.

Greenemeier, L., 26 October 2012. Educating Players: Are Games the Future of; Education? Scientific American. http://blogs.scientificamerican.com/observations/2012/10/26/educating-players-are-games-the-future-of-education/.

Greenfield, P.M., 2009. Technology and informal education: what is taught, what is learned. Science 323, 69–71.

Halpern, D.F., Keith Millis, K., Graesser, A.C., Butlerd, H., Forsyth, C., Caic, Z., 2012. Operation ARA: a computerized learning game that teaches critical thinking and scientific reasoning. Thinking Skills and Creativity 7, 93–100.

Harris, C., 2009. The truth about ebooks. School Library Journal 55, 18–18.

Hodgson, E., Bachman, E., Waller, D., 2011. Redirected walking to explore virtual; environments: assessing the potential for spatial interference. ACM Transactions on Applied Perception 8, 1–22.

Katz, S., Connelly, J., Wilson, C., 2007. Out of the lab and into the classroom: an evaluation of reflective dialogue in Andes. In: Luckin, R., Koedinger, K.R., Greer, J. (Eds.), Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work, Proceedings of the 13th International Conference on AI in Education, AIED 2007. IOS Press, Amsterdam, pp. 425–432.

Kurshan, B., 2008. OER models that build a culture of collaboration: a case exemplified by Currik. eLearning Papers 10, 1–8.

Luo, T., Gao, F., 2012. Enhancing classroom learning experience by providing structures to microblogging based activities. Journal of Information Technical Education 11, 199–211.

Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A., Halpern, D.F., 2011. Operation ARIES!: a serious game for teaching scientific inquiry. In: Ma, M., Oikonomou, A., Jain, L. (Eds.), Serious Games and Edutainment Applications. Springer-Verlag, pp. 169–195.

Mostow, J., Beck, J., 2007. When the rubber meets the road: lessons from the in-school adventures of an automated reading tutor that listens. In: Schneider, B., McDonald, S.-K. (Eds.), Conceptualizing Scale-Up: Multidisciplinary Perspectives. Rowman & Littlefield, Lanham, MD, pp. 183–200.

Ovadia, M., 2007. PoducateMe: Practical Solutions for Podcasting in Education. http://www.poducateme.com/guide/purchase.

Pierce, D., 21 June 2010. Summit: Mobile Computing is Education's Future. http://www.eschoolnews.com/2010/06/21/summit-mobile-computing-is-educations-future/.

Pinkwart, N., Aleven, V., Ashley, K., Lynch, C., 2006. Toward legal argument instruction with graph grammars and collaborative filtering techniques. In: Ikeda, M., Ashley, K., Chan, T.W. (Eds.), Proceedings of the 8th International Conference on Intelligent Tutoring Systems (ITS-06). Springer, Berlin, pp. 227–236.

Riva, G., Mantovani, F., 2012. From the body to the tools and back: a general framework for presence in mediated interactions. Interacting with Computers 24, 203–210.

Schaffhauser, D., 27 January 2010. Apple's iPad: The Future of Mobile Computing in Education? Campus Technology. http://campustechnology.com/articles/2010/01/27/apples-ipad-the-future-of-mobile-computing-in-education.aspx.

Shaffer, D.W., Squire, K.R., Halverson, R., Gee, J.P., 2005. Video games and the future of learning. Phi Delta Kappan 87, 104–111.

Squire, K.D., DeVane, B., Durga, S., 2008. Designing centers of expertise for academic learning through video games. Theory Into Practice 47, 240–251.

Steinkuehler, C., Duncan, S.C., 2008. Scientific habits of mind in virtual worlds. Journal of Science Education Technology 17, 530–543.

Sumadio, D.D., Rambli, D.R.A., 2010. Preliminary evaluation on user acceptance of the augmented reality use for education. In: The Second International Conference on Computer Engineering and Applications, pp. 461–464.

Swan, K., van 't Hooft, M., Kratcoski, A., Unger, D., 2005. Uses and effects of mobile computing devices in K-8 classrooms. International Society for Technology in Education 38, 99–112.

University of Phoenix, 2012. Degree Programs. http://www.phoenix.edu/programs/degree-programs.html.

Van Orsdel, L., 2012. Openness: Contribute, Access, Use. http://scholarworks.gvsu.edu/library_presentations/40.

Vance, A., 4 October 2012. Facebook: The Making of 1 Billion Users. Bloomberg Businessweek. http://www.businessweek.com/articles/2012-10-04/facebook-the-making-of-1-billion-users.

Waller, D., Bachmann, E., Hodgson, E., Beall, A.C., 2007. The HIVE: a huge immersive virtual environment for research in spatial cognition. Behavior Research Methods 39, 835–843.

Westera, W., Nadolski, R.J., Hummel, H.G.K., Wopereis, I.G.J.H., 2008. Serious games for higher education: a framework for reducing design complexity. Journal of Computer Assisted Learning 24, 420–432.

Wolfe, C.R., 2001a. Learning and Teaching on the World Wide Web. Academic Press, San Diego, CA.

Wolfe, C.R., 2001b. Creating informal learning environments on the web. In: Wolfe, C.R. (Ed.), Learning and Teaching on the World Wide Web. Academic Press, San Diego, CA.

Wolfe, C.R., 2001c. Plant a tree in cyberspace: metaphor and analogy as design; elements in web-based learning environments. CyberPsychology & Behavior 4, 67–75.

Wolfe, C.R., Myers, C.A., Cummins, R.H., 2001. The dragonfly web pages: informal science education on the world wide web. Cognitive Technology 6, 4–13.

Wolfe, C.R., Britt, M.A., Petrovic, M., Albrecht, M., Kopp, K., 2009. The efficacy of a web-based counterargument tutor. Behavior Research Methods 41, 691–698.

Wolfe, C.R., Fisher, C.R., Reyna, V.F., Hu, X., 2012. Improving internal consistency in conditional probability estimation with an intelligent tutoring system and web-based tutorials. International Journal of Internet Science 7, 38–54.

Wolfe, C.R., Widmer, C.L., Reyna, V.F., Hu, X., Cedillos, E.M., Fisher, C.R., Brust-Renck, P.G., Williams, T.C., Damas, I., Weil, A.M., 2013. The development and analysis of tutorial dialogues in AutoTutor Lite. Behavior Research Methods 45, 623–636.

Relevant Websites

http://www.ocwconsortium.org/en/aboutus/abouttheocwc – About the OpenCourseWare Consortium.

http://www.psychologicalscience.org/index.php/members/aps-wikipedia-initiative – Association for Psychological Science Wikipedia Initiative.

http://www.dmoz.org/desc/Reference/Education/Distance_Learning – DMOZ Distance Learning.

http://www.projectdragonfly.org/ – Project Dragonfly: Inquiry, Community, Voice.

http://www.sakaiproject.org/ – Sakai: A Community of Educators Collaborating to Create Open Software That Advances Teaching, Learning, and Research.

http://www.seriousgames.org/ – The Serious Games Initiative.

http://www.skoonline.org/home – S. K. O. Online: AutoTutor Lite.

http://vitotechnology.com/star-walk.html – Star Walk: The Universe in Your Hands.

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 7

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser—Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 7

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 7 *xliii*

VOLUME 7

Econometric Software 1
William H Greene

Econometrics, History of 8
Mary S Morgan and Duo Qin

Economic Anthropology 12
Kalman Applbaum

Economic Education 21
Carlos J Asarta and Roger B Butters

Economic Evaluations and Electoral Consequences 26
Michael S Lewis-Beck and Mary Stegmaier

Economic Geography 33
Edward J Malecki

Economic Growth 38
David de la Croix

Economic Growth: Measurement 45
Muhammad H Rahman and Mehmet Ali Ulubaşoğlu

Economic History 51
Chris Godden

Economic History, Quantitative: United States 55
Gavin Wright

Economic Restructuring: Geographic Aspects 61
Steven Pinch

Economic Sociology 67
Neil Fligstein and Cyrus Dioun,

Economic Transformation: From Central Planning to Market Economy 73
Gerard Turley

Economics and Crime 80
Joanna Shepherd and Paul H Rubin

Economics and Ethics 87
John Broome

Economics, History of 93
Craufurd Goodwin

Economics, Philosophy of 99
Daniel M Hausman

Ecotourism 105
Carl Cater and Erlet Cater

Ecumene 110
Igor Kopytoff

Edgeworth, Francis Ysidro (1845–1926) 113
Peter Newman

Education (Primary and Secondary Schools) and Gender 117
Maria Tamboukou

Education and Economic Growth 121
Dennis W Cheek, Joel Santos, and Nicolas Vaillant

Education and Employment 127
Ulrich Teichler

Education and Income Distribution 132
Gary N Marks

Education and Learning: Lifespan Perspectives 137
Elizabeth AL Stine-Morrow and Brennan R Payne

Education: Anthropological Aspects 144
Roger Goodman

Education, Economics of 149
Eric A Hanushek

Education for the Gifted and Talented 158
David W Chan

Education, Geography of 165
Peter Meusburger

Education, History of 172
Jurgen Schriewer and António Nóvoa

Education in Latin America and the Caribbean: Systems and Research 178
Dante J Salto

Education in Psychiatry 185
Brian Lunn

Education in South Asia 194
R Malatesha Joshi

Education in Sub-Saharan Africa 198
Sylvia C Kalindi

Education in the Middle East and North Africa 210
Abdeljalil Akkari

Education: Skills Training 215
Roy Canning

Educational Effectiveness Theory and Research: Recent Advances 218
Leonidas Kyriakides

Educational Effectiveness, The Field of 224
Bert PM Creemers and Leonidas Kyriakides

Educational Evaluation: Overview 229
Geoff Lindsay

Educational Institutions, History of 232
Peter M Roeder

Educational Learning Theory 238
Kelly YL Ku, Sivanes Phillipson, and Shane N Phillipson

Educational Media: Potentials for Learning 246
Deborah L Linebarger

Educational Philosophy 255
Gert Biesta

Educational Planning Worldwide 261
Françoise Caillods

Educational Policy, Mechanisms for the Development of 266
Sandra Vergari

Educational Productivity 272
Susan J Paik

Educational Research, The Field of 279
Felice J Levine and Lori Diane Hill

Educational Sociology 289
Lawrence J Saha

Educational Systems: North America 297
Gerald K LeTendre

Efficacy, Political 301
Thomas J Scotto and Carla Xena

Egalitarianism 308
Christian Hiebaum

Eisenberg's Theory of Prosocial Reasoning 315
Bilge Yagmurlu and Hilal Sen

Eisenstadt, Shmuel N. (1923–2010) 321
Luis Roniger

Elder Abuse 327
Jacki Pritchard

Elder Care 335
Mary P Sullivan

Elections: Apportionment 342
Bernard Grofman

Electoral Geography 345
Ron Johnston

Electoral Systems 349
Alan Ware

Electroencephalography: Basic Principles and Applications 353
David B MacDonald

Electronic Waste 364
Josh Lepawsky

Eliade, Mircea (1907–86) 374
Robert A Segal

Elias, Norbert (1897–1990) 378
Wolfgang Ludwig-Mayerhofer

Elicitation of Probabilities and Probability Distributions 382
Lara J Wolfson

Elites, Anthropology of 386
Hugh Gusterson

Elites: History of the Concept 390
Philipp Korom

Elites: Sociological Aspects 396
Michael Hartmann

Elster, Jon (1940–) 402
Charles HG Crothers

Embodied Social Cognition 409
Caterina Suitner, Mauro Giacomantonio, and Anne Maass

Embodiment and Culture 415
Gabe Ignatow

Embodiment Theory 420
Anna M Borghi and Fausto Caruana

Emergency Relief and Social Work Responses 427
Desley Hargreaves

Emergent Properties 433
Marlis Buchmann

Emotional Development, Effects of Parenting and Family Structure on 438
Suzanne Bester and Marlize Malan-Van Rooyen

Emotional Intelligence and Competencies 445
Christopher TH Miners and Ivona Hideg

Emotional Regulation 452
Min Y Kim, Yochanan Bigman, and Maya Tamir

Emotion and Expression 457
Paul E Griffiths and Elena Walsh

Emotion in Cognition 461
Julia A Harris and Derek Isaacowitz

Emotion, Evolution of 467
Elena Walsh and Paul E Griffiths

Emotion: History of the Concept 472
Katharina Scherke

Emotion, Neural Basis of 477
Edmund T Rolls

Emotion, Perception and Expression of 483
Daisung Jang and Hillary Anger Elfenbein

Emotions and Aging 490
Tammy English and Laura L Carstensen

Emotions and Health 496
Ayse K Uskul and Andrea B Horn

Emotions and Intergroup Relations 502
Colin W Leach, Fouad Bou Zeineddine, and Atilla Cidam

Emotions and Work 507
Neal M Ashkanasy

Emotions, History of 513
Peter N Stearns

Emotions: Methods of Assessment 519
Marcello Mortillaro and Marc Mehu

Emotions, Psychological Structure of 526
Vera Shuman and Klaus R Scherer

Emotions, Sociology of 534
Kathryn J Lively

Empathy: A Social Neuroscience Perspective 541
Jean Decety and Greg J Norman

Empathy during early Childhood Across Cultures, Development of 549
Elena Geangu

Empirical Legal Studies 554
Christopher Zorn

Empirical Social Research, History of 558
Hynek Jerabek

Empiricism, History of 567
Peter Lipton

Employment and Labor, Regulation of 571
Katherine VW Stone

Empowerment Evaluation 577
David M Fetterman

Encyclopedias, Handbooks, and Dictionaries 584
Alan Sica

End of Life Choices 594
Frances Norwood

Endocrinology and Psychiatry 606
Rand Mahmoud and Liisa AM Galea

Endogeneity: Methodology 612
Trond Petersen

Energy and Society 615
Kirby Calvert

Energy, Anthropology of 621
Dominic Boyer

Energy Flow Analysis 626
Helmut Haberl

Energy, Geography of 633
Johannes Hamhaber

Engineering Studies 641
Gary L Downey

English Revolution, The 649
John Gurney

Enlightenment 656
Giuseppe Ricuperati

Enlightenment: Impact on the Social Sciences 664
John C Torpey

Entertainment 669
Martin Zerlang

Entrepreneurship and Economic Growth 675
Alexander S Kritikos

Entrepreneurship and Organizations 681
Peter T Bryant

Entrepreneurship, Geography of 686
John R Bryson

Environmental Archaeology 692
Nicholas Branch

Environmental Attitudes and Behavior: Measurement 699
Sebastian Bamberg and Jonas Rees

Environmental Cognition, Perception, and Attitudes 706
Baruch Fischhoff

Environmental Conservation and Development: Critical Perspectives 713
Scott Freeman

Environmental Determinism 720
Innes M Keighren

Environmental Economics 726
Vincent K Smith

Environmental Education and Education for Sustainable Development 733
Therese Hume and John Barry

Environmental Health and Safety: Social Aspects 740
Dean Bertolatti, Toni Hannelly, and Janis Jansz

Environmental History 747
Vinita Damodaran

Environmental Justice in the United States 756
Robert D Bullard

Environmental Movements 763
Robert J Brulle and Christopher Rootes

Environmental Planning 769
John C Keene

Environmental Policy: Protection and Regulation 778
David Benson and Andrew Jordan

Environmental Risk 784
Ortwin Renn

Environmental Sciences 791
Chunglin Kwa

Environmental Sociology 796
Riley E Dunlap

Environmental Stress and Health 804
John S Picou, Keith Nicholls, and Rainer Guski

Environment and Development 809
Chris J Barrow

Environment and Health 815
Brian King

Environments for Education 820
Barry J Fraser

Epidemiology, Concepts and Methods of 824
Roselind Lieb

Epigenetic Inheritance 832
Eva Jablonka and Marion J Lamb

Epigenetics 839
Jonathan D Turner, Sophie A Kirschner, Anne M Molitor, Konstantin Evdokimov, and Claude P Muller

Epigenetics of Brain Plasticity and Behavior 848
Richelle Mychasiuk

Epilepsy 852
G Campbell Teskey and Jordan S Farrell

Episodic Memory 856
Costanza Papagno

Epistemic Communities and Collaborative Research 862
Morgan Meyer

Epistemic Communities and Collaborative Research, History of 867
Kenneth Brad Wray

Epistemic Cultures 873
Karin K Cetina and Werner Reichmann

Epistemic Doubt During Adolescence 881
Robert B Ricco

Epistemological Beliefs and Motivation 887
Kerstin Oschatz

Equality and Inequality: Legal Aspects 894
Susan Sturm

Equality of Opportunity 903
Meir Yaish

Equality: Philosophical Aspects 906
Emmanuel Picavet

Equilibrium: History of the Concept 912
Till Düppe

Equity and Education 918
Salma Ismail

Equity in Law 924
André-Jean Arnaud

Equivalence and Transfer Problems in Cross-Cultural Research 929
Amina Abubakar

Erikson, Erik Homburger (1902–94) 934
James E Marcia

Erikson's Theory of Psychosocial Development 938
Li-fang Zhang

CONTRIBUTORS TO VOLUME 7

Amina Abubakar
Tilburg University, Tilburg, The Netherlands; and Utrecht University, Utrecht, The Netherlands

Abdeljalil Akkari
University of Geneva, Genève, Switzerland

Kalman Applbaum
University of Wisconsin-Milwaukee, Milwaukee, WI, USA

André-Jean Arnaud
Droit et Société, FMSH, Paris, France

Carlos J. Asarta
University of Delaware, Newark, DE, USA

Neal M. Ashkanasy
University of Queensland Business School, The University of Queensland, Brisbane, QLD, Australia

Sebastian Bamberg
Bielefeld University of Applied Sciences, Bielefeld, Germany

Chris J. Barrow
College of Science, Swansea University, Swansea, UK

John Barry
Queens University Belfast, Belfast, UK

David Benson
Environment and Sustainability Institute (ESI), University of Exeter, Penryn, UK

Dean Bertolatti
Curtin University, Perth, WA, Australia

Suzanne Bester
University of Pretoria, Pretoria, South Africa

Gert Biesta
University of Luxembourg, Walferdange, Luxembourg

Yochanan Bigman
The Hebrew University in Jerusalem, Mount Scopus, Jerusalem, Israel

Anna M. Borghi
University of Bologna, Bologna, Italy; and Institute of Cognitive Sciences and Technologies, Italian National Research Council, Rome, Italy

Dominic Boyer
Rice University, Houston, TX, USA

Nicholas Branch
School of Archaeology, Geography and Environmental Science, University of Reading, Reading, UK

John Broome
University of Oxford and Corpus Christi College, Oxford, UK

Robert J. Brulle
Drexel University, Philadelphia, PA, USA

Peter T. Bryant
IE Business School-IE University, Madrid, Spain

John R. Bryson
Birmingham Business School, University of Birmingham, Birmingham, UK

Marlis Buchmann
University of Zurich, Zurich, Switzerland

Robert D. Bullard
Texas Southern University, Houston, TX, USA

Roger B. Butters
Hillsdale College, Hillsdale, MI, USA

Françoise Caillods
Formally Deputy Director of the International Institute for Educational Planning, Currently Senior Advisor at the International Social Science Council, Paris, France

Kirby Calvert
The Pennsylvania State University, University Park, PA, USA

Roy Canning
School of Education, University of Stirling, UK

Laura L. Carstensen
Stanford University, Stanford, CA, USA

Fausto Caruana
University of Parma, Parma, Italy; and Istituto Italiano di Tecnologia (IIT), Brain Center for Social and Motor Cognition, Parma, Italy

Carl Cater
Aberystwyth University, Aberystwyth, UK

Erlet Cater
University of Reading, Reading, UK

Karin K. Cetina
University of Chicago, Chicago, IL, USA

David W. Chan
The Chinese University of Hong Kong, Hong Kong, People's Republic of China

Dennis W. Cheek
Executive Director, National Creativity Network, USA; and Visiting Professor, IESEG School of Management, France

Atilla Cidam
University of Connecticut, Storrs, CT, USA

Bert P.M. Creemers
Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands

Charles H.G. Crothers
School of Social Sciences and Public Policy, AUT University, Auckland, New Zealand

Vinita Damodaran
University of Sussex, Brighton, UK

Jean Decety
The University of Chicago, Chicago, IL, USA

David de la Croix
IRES, Université catholique de Louvain, Louvain-la-Neuve, Belgium

Cyrus Dioun
University of California, Berkeley, CA, USA

Gary L. Downey
Virginia Tech, Blacksburg, VA, USA

Riley E. Dunlap
Oklahoma State University, Stillwater, OK, USA

Till Düppe
Université du Québec à Montréal (UQAM), Montréal, QC, Canada

Hillary Anger Elfenbein
Washington University in St. Louis, Saint Louis, MO, USA

Tammy English
Washington University in St. Louis, St. Louis, MO, USA

Konstantin Evdokimov
Institute of Immunology, Centre de Recherche Public — Santé/Laboratoire National de Santé, Luxembourg, Grand Duchy of Luxembourg

Jordan S. Farrell
Faculty of Medicine, Health Sciences Center, Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada

David M. Fetterman
San Jose, CA, USA

Baruch Fischhoff
Carnegie Mellon University, Pittsburgh, PA, USA

Neil Fligstein
University of California, Berkeley, CA, USA

Barry J. Fraser
Science and Mathematics Education Centre, Curtin University, Perth, WA, Australia

Scott Freeman
George Washington University, Washington, DC, USA

Liisa A.M. Galea
Program in Neuroscience, Centre for Brain Health, Vancouver, BC, Canada; and University of British Columbia, Vancouver, BC, Canada

Elena Geangu
Fylde College, Lancaster University, Lancaster, UK

Mauro Giacomantonio
University of Rome 'Sapienza,' Rome, Italy

Chris Godden
University of Manchester, Manchester, UK

Roger Goodman
University of Oxford, Oxford, UK

Craufurd Goodwin
Duke University, Durham, NC, USA

William H. Greene
New York University, New York, NY, USA

Paul E. Griffiths
University of Sydney, Sydney, NSW, Australia

Bernard Grofman
University of California, Irvine, CA, USA

John Gurney
Newcastle University, Newcastle-upon-Tyne, UK

Rainer Guski
Ruhr Universitat Bochum, Bochum, Germany

Hugh Gusterson
Elliott School of International Affairs, Washington, DC, USA

Helmut Haberl
Institute of Social Ecology, Alpen-Adria Universität Klagenfurt, Vienna, Austria; and Integrative Research Institute on Transformations in Human Environment Systems, Humboldt-Universität zu Berlin, Berlin, Germany

Johannes Hamhaber
Cologne University of Applied Sciences, Institute for Technology and Resources Management in the Topics and Subtropics (ITT), Köln, Germany

Toni Hannelly
Curtin University, Perth, WA, Australia

Eric A. Hanushek
Hoover Institution, Stanford University, Stanford, CA, USA

Desley Hargreaves
School of Social Work and Human Services, University of Queensland, Brisbane, QLD, Australia

Julia A. Harris
Harvard University, Cambridge, MA, USA

Michael Hartmann
TU Darmstadt, Darmstadt, Germany

Daniel M. Hausman
University of Wisconsin—Madison, Madison, WI, USA

Ivona Hideg
School of Business and Economics, Wilfrid Laurier University, Waterloo, ON, Canada

Christian Hiebaum
Department of Legal Philosophy, Sociology of Law, and Legal Informatics, University of Graz, Graz, Austria

Lori Diane Hill
American Educational Research Association, Washington, DC, USA

Andrea B. Horn
University of Zuerich, Zuerich, Switzerland

Therese Hume
Institute of Technology, Ballinode, Ireland

Gabe Ignatow
University of North Texas, Denton, TX, USA

Derek Isaacowitz
Northeastern University, Boston, MA, USA

Salma Ismail
Faculty of Humanities, School of Education, University of Cape Town, Cape Town, South Africa

Eva Jablonka
Tel Aviv University, Tel Aviv, Israel

Daisung Jang
Washington University in St. Louis, Saint Louis, MO, USA

Janis Jansz
Curtin University, Perth, WA, Australia

Hynek Jerabek
Institute of Sociological Studies, Faculty of Social Sciences, Charles University in Prague, Prague, Czech Republic

Ron Johnston
University of Bristol, Bristol, UK

Andrew Jordan
School of Environmental Sciences, University of East Anglia, Norwich, UK

Sylvia C. Kalindi
The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

John C. Keene
University of Pennsylvania, Philadelphia, PA, USA

Innes M. Keighren
Royal Holloway, University of London, Egham, UK

Min Y. Kim
The Hebrew University in Jerusalem, Mount Scopus, Jerusalem, Israel and Keimyung University, South Korea

Brian King
The Pennsylvania State University, University Park, PA, USA

Sophie A. Kirschner
Institute of Immunology, Centre de Recherche Public — Santé/Laboratoire National de Santé, Luxembourg, Grand Duchy of Luxembourg; and Graduate School of Psychobiology, University of Trier, Trier, Germany

Igor Kopytoff
University of Pennsylvania, Philadelphia, PA, USA

Philipp Korom
Max Planck Institute for the Study of Societies, Cologne, Germany

Alexander S. Kritikos
German Institute for Economic Research (DIW Berlin), Berlin, Germany; University of Potsdam, Potsdam, Germany; IZA, Bonn, Germany; and IAB, Nuremberg, Germany

Kelly Y.L. Ku
Hong Kong Baptist University, Hong Kong

Chunglin Kwa
University of Amsterdam, Amsterdam, The Netherlands

Leonidas Kyriakides
University of Cyprus, Nicosia, Cyprus

Marion J. Lamb
Tel Aviv University, Tel Aviv, Israel

Colin W. Leach
University of Connecticut, Storrs, CT, USA

Josh Lepawsky
Memorial University, St. John's, NL, Canada

Gerald K. LeTendre
The Pennsylvania State University, University Park, PA, USA

Felice J. Levine
American Educational Research Association, Washington, DC, USA

Michael S. Lewis-Beck
University of Iowa, Iowa City, IA, USA

Roselind Lieb
Division of Clinical Psychology and Epidemiology, University of Basel, Missionsstrasse, Basel, Switzerland

Geoff Lindsay
Centre for Educational Development, Appraisal and Research (CEDAR), University of Warwick, Coventry, UK

Deborah L. Linebarger
N278 Lindquist Center, University of Iowa, Iowa City, IA, USA

Peter Lipton[†]
Cambridge University, Cambridge, UK

Kathryn J. Lively
Dartmouth College, Hanover, NH, USA

Wolfgang Ludwig-Mayerhofer
University of Siegen, Siegen, Germany

Brian Lunn
School of Medical Education, The Medical School, Newcastle University, Newcastle upon Tyne, UK

Anne Maass
University of Padova, Padova, Italy

David B. MacDonald
King Faisal Specialist Hospital & Research Center, Riyadh, Saudi Arabia

Rand Mahmoud
Program in Neuroscience, University of British Columbia, Centre for Brain Health, Vancouver, BC, Canada

Marlize Malan-Van Rooyen
University of Pretoria, Pretoria, South Africa

R. Malatesha Joshi
Texas A & M University, College Station, TX, USA

Edward J. Malecki
The Ohio State University, Columbus, OH, USA

James E. Marcia
Simon Fraser University, Burnaby, BC, Canada

Gary N. Marks
Melbourne Institute of Applied Economic and Social Research, The University of Melbourne, Melbourne, VIC, Australia

Marc Mehu
Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland; and Webster Vienna, Private University, Vienna, Austria

Peter Meusburger
Heidelberg University, Heidelberg, Germany

Morgan Meyer
Agro ParisTech, Paris, France

Christopher T.H. Miners
Queen's School of Business, Queen's University, Kingston, ON, Canada

Anne M. Molitor
Institute of Immunology, Centre de Recherche Public — Santé/Laboratoire National de Santé, Luxembourg, Grand Duchy of Luxembourg

Mary S. Morgan
The London School of Economics and Political Science, London, UK

Marcello Mortillaro
Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland

Claude P. Muller
Institute of Immunology, Centre de Recherche Public — Santé/Laboratoire National de Santé, Luxembourg, Grand Duchy of Luxembourg; and Graduate School of Psychobiology, University of Trier, Trier, Germany

Richelle Mychasiuk
Faculty of Medicine, Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB, Canada

[†] Deceased.

Peter Newman
John Hopkins University, Baltimore, MD, USA

Keith Nicholls
University of South Alabama, Mobile, AL, USA

Greg J. Norman
The University of Chicago, Chicago, IL, USA

Frances Norwood
George Washington University, Washington, DC, USA

António Nóvoa
Universidade de Lisboa, Lisboa, Portugal

Kerstin Oschatz
Hector Research Institute of Education Sciences and Psychology, Tübingen, Germany

Susan J. Paik
School of Educational Studies, Claremont Graduate University, Claremont, CA, USA

Costanza Papagno
Università di Milano Bicocca, Milano, Italy

Brennan R. Payne
University of Illinois at Urbana-Champaign, Urbana, IL, USA

Trond Petersen
University of California, Berkeley, Berkeley, CA, USA

Shane N. Phillipson
Monash University, Melbourne, VIC, Australia

Sivanes Phillipson
Monash University, Melbourne, VIC, Australia

Emmanuel Picavet
Université Paris-1 Panthéon-Sorbonne, Paris, France

John S. Picou
University of South Alabama, Mobile, AL, USA

Steven Pinch
University of Southampton, Southampton, UK

Jacki Pritchard
Jacki Pritchard Ltd, Sheffield, UK

Duo Qin
School of Oriental and African Studies, London, UK

Muhammad H. Rahman
Deakin University, Burwood, VIC, Australia

Jonas Rees
Bielefeld University of Applied Sciences, Bielefeld, Germany; and Bielefeld University, Bielefeld, Germany

Werner Reichmann
University of Konstanz, Konstanz, Germany

Ortwin Renn
Stuttgart Center for Interdisciplinary Risk and Innovation Studies at Stuttgart University (ZIRIUS), Stuttgart University, Stuttgart, Germany

Robert B. Ricco
California State University, San Bernardino, CA, USA

Giuseppe Ricuperati
University of Turin, Turin, Italy; and Academic of Lincei, Rome, Italy

Peter M. Roeder
Max Planck Institute for Human Development, Berlin, Germany

Edmund T. Rolls
Oxford Centre for Computational Neuroscience, Oxford, UK; and University of Warwick, Coventry, UK

Luis Roniger
Wake Forest University, Winston—Salem, NC, USA; and Hebrew University of Jerusalem, Israel

Christopher Rootes
School of Social Policy, Sociology and Social Research, University of Kent, Canterbury, UK

Paul H. Rubin
Emory University, Atlanta, GA, USA

Lawrence J. Saha
Sociology, College of Arts and Social Sciences, Australian National University, Canberra, ACT, Australia

Dante J. Salto
University at Albany, State University of New York, Albany, NY, USA; and Universidad Nacional de Córdoba, Argentina

Joel Santos
Faculté libre de Sciences Économique et de Gestion, Edhec Business School, Roubaix, France

Klaus R. Scherer
Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland

Katharina Scherke
University of Graz, Graz, Austria

Jurgen Schriewer
Humboldt University, Berlin, Germany

Thomas J. Scotto
University of Essex, Colchester, UK

Robert A. Segal
Lancaster University, Lancaster, UK

Hilal Sen
Koc University, Istanbul, Turkey

Joanna Shepherd
Emory University School of Law, Atlanta, GA, USA

Vera Shuman
University of Lausanne, Lausanne, Switzerland

Alan Sica
Pennsylvania State University, State College, PA, USA

Vincent K. Smith
Arizona State University, Tempe, AZ, USA

Peter N. Stearns
George Mason University, Fairfax, VA, USA

Mary Stegmaier
Truman School of Public Affairs, University of Missouri, Columbia, MO, USA

Elizabeth A.L. Stine-Morrow
University of Illinois at Urbana-Champaign, Urbana, IL, USA

Katherine V.W. Stone
UCLA School of Law, Los Angeles, CA, USA

Susan Sturm
Columbia Law School, Columbia University, New York, NY, USA

Caterina Suitner
University of Padova, Padova, Italy

Mary P. Sullivan
Brunel University, Uxbridge, UK

Maria Tamboukou
School of Social Sciences, University of East London, London, UK

Maya Tamir
The Hebrew University in Jerusalem, Mount Scopus, Jerusalem, Israel

Ulrich Teichler
International Centre for Higher Education Research (INCHER-Kassel), University of Kassel, Kassel, Germany

G. Campbell Teskey
Faculty of Medicine, Health Sciences Center, Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada

John C. Torpey
Graduate Center, City University of New York, New York, NY, USA

Gerard Turley
J.E. Cairnes School of Business and Economics, National University of Ireland, Galway (NUI Galway), Ireland

Jonathan D. Turner
Institute of Immunology, Centre de Recherche Public – Santé/Laboratoire National de Santé, Luxembourg, Grand Duchy of Luxembourg; and Graduate School of Psychobiology, University of Trier, Trier, Germany

Mehmet Ali Ulubasoğlu
Deakin University, Burwood, VIC, Australia

Ayse K. Uskul
School of Psychology, University of Kent, Canterbury, UK

Nicolas Vaillant
Université Catholique de Lille, Lille, France; and Centre de Recherche sur le Capital Humaine, Faculté Libre de Sciences Economiques et de Gestion, France

Sandra Vergari
State University of New York at Albany, Albany, NY, USA

Elena Walsh
University of Sydney, Sydney, NSW, Australia

Alan Ware
Oxford University, Oxford, UK

Lara J. Wolfson
Brigham Young University, Provo, UT, USA

Kenneth Brad Wray
State University of New York at Oswego, Oswego, NY, USA

Gavin Wright
Stanford University, Stanford, CA, USA

Carla Xena
University of Essex, Colchester, UK

Bilge Yagmurlu
Koc University, Istanbul, Turkey

Meir Yaish
University of Haifa, Haifa, Israel

Fouad Bou Zeineddine
University of Connecticut, Storrs, CT, USA

Martin Zerlang
The University of Copenhagen, Copenhagen, Denmark

Li-fang Zhang
The University of Hong Kong, Hong Kong

Christopher Zorn
Pennsylvania State University, University Park, PA, USA

Econometric Software

William H Greene, New York University, New York, NY, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 4058–4065, © 2001, Elsevier Ltd.

Abstract

This article will describe some of the characteristics of the most widely used econometric software, trace the evolution of this particular type of computer program, and explore the role of computer software in the development of applied econometrics. We begin with a general discussion of econometrics. The specific methodology defined by this field both dictates and is strongly influenced by the software that practitioners use – we return to that issue later in the article. The distinctive features of the current vintage of econometric software and the progression of developments that has produced it are discussed next. Econometricians' tools and methods have changed over time, gradually evolving from an early emphasis on single and multiple equation linear models to the contemporary use of a huge variety of nonlinear techniques in addition to the traditional linear regression models. Software has evolved correspondingly, becoming more complex and flexible in the process. The common characteristics of software in econometrics are considered next. A natural progression has produced a broad similarity in both the look and feel and the capabilities of econometrics programs. As indicated earlier, the final sections of the article explore some of the interrelationships between econometric software and popular methods in econometrics.

This article will describe some of the characteristics of the most widely used econometric software, trace the evolution of this particular type of computer program, and explore the role of computer software in the development of applied econometrics. In order to discuss econometric software it is necessary, first, to define econometrics. We begin the discussion in Section Econometrics with a brief introduction to econometrics. Section The Evolution of Econometric Software describes the distinctive features of the current vintage of econometric software and the progression of developments that has produced it. Section The Characteristics of Modern Software discusses some of the common characteristics of software in econometrics. Finally, Section The Role of Software in the Development of Econometrics will suggest some of the interrelationships between econometric software and popular methods in econometrics.

Econometrics

Econometrics is a collection of methods and tools used to fit equations (economic models) to data. It involves both theory and measurement, and an overarching view of the process by which data come to be observed. The analysis begins with a theoretical explanation of some phenomenon, or a hypothesis about how some variables to be observed are related to each other. Measurement begins with the recording of some aspect of economic activity. Model building is the process by which these economic measurements are organized within the framework of one or more mathematical equations suggested by the economic theory. The principle behind this analytical framework is that of a 'data generating process' (DGP). There is an underlying statistical regularity at work that leads to the analyst's observation of their measurements. It is the characteristics of this underlying regularity, usually formalized as a probability distribution over random events, which are the objects of the econometric study.

In a familiar application of theory to data, observers collect pairs of observations on prices (P) and quantities (Q) exchanged in a market, such as the market for tobacco or fine wine, under the umbrella of a theory of 'demand.' The theory is embodied in a 'demand equation,' $Q = f(P|\alpha,\beta)$, which states that quantities traded are influenced by prices, and the nature of that response can be described by the relationship $f(\ldots)$ and the 'parameters,' α and β. The equation is made operational by computing statistics that provide information about the covariation of prices and quantities transacted in the market. Thus, β might be an overall measure of how responsive quantities are to changes in prices, and the data are used to deduce information about the specific value of β.

In an econometric analysis, the information in a data set which may include thousands or millions of values is distilled to a few equations which are claimed to describe the DGP. The intricacy of modern econometric techniques makes essential elaborate software with capabilities that extend well beyond the ubiquitous spreadsheet programs such as Excel (4) and Lotus (10) (which would be the nearest cousins to the software considered here). (Numbers in brackets refer to the list of computer programs in the Appendix.) Econometric software is used most frequently for the manipulation of large amounts of measured data for the purpose of constructing the equations that are the model of the process which produced the data.

The Evolution of Econometric Software

The earliest software in econometrics was single purpose sets of Fortran language (Formula Translation) code. The researcher who needed a program for a particular purpose would typically write it, or borrow it from someone else. One can find some of these programs in the appendices in some of the classic pieces in the literature, such as Nerlove and Press's (1973) development of the multinomial logit model, the first edition of Draper and Smith's (1981) exposition of regression analysis, Zellner's (1971) early work on Bayesian estimation, and, more recently in Kalbfleisch and Prentice's (1980) work on survival modeling. A typical journal treatment is Krailo and Pike's

(1984) exposition of a logit model for panel data, which contains a Fortran program listing for a standalone estimation program. Until the late 1960s, the journal *Econometrica* occasionally would publish short articles documenting computer programs developed by the authors.

Program Libraries

The process of building applications from scratch was made easier by 'libraries' of interchangeable parts for software development. The precursor to a very large group of programs currently distributed as the 'BMDP Library' (2) was a group of individual routines developed at UCLA in the 1960s under the heading of the 'BMD' (BioMedical) programs. These special purpose routines included subroutines for linear regression, Spearman rank correlation coefficients, cross tabulations, and so on. The BMDs initially existed as boxes of 80 column punched cards of Fortran code in the drawers of statistics labs. They maintained a common structure that allowed authors to write their own data setup programs to 'plug into' the package. Another programmer's toolkit that was circulated widely in the 1970s was the 'IBM SSP' (Scientific Subroutine Package) which was part of the documentation for IBM mainframe computers. This book contained Fortran and C listings of programs for such functions as inverting or decomposing a matrix, and computing a variance ('C' does not stand for anything). It was simply an extension of 'B,' itself a derivative of 'BCPL' (the BC Programming Language) and a contemporary of APL (A Programming Language). All were languages originally written to work with the Unix operating system on DEC's original PDP-11 computers.

Most of the difficult computations being done by econometricians at this time involved complex matrix algebra, data manipulation, and random number generation. A library of well-written subroutines combined with the efforts of the analyst was quite sufficient for application development. This approach to programming has not vanished from the landscape. The *Numerical Recipes* books (C, Pascal, and Fortran) by Press et al. (1986), contain the modern counterpart to the expertly written library of subroutines contained in the SSP. The BMD libraries still exist as the commercial BMDP (2) routines, and the IMSL programs (6) and NAG (Numerical Analysis Group) (14) routines are still used by scientific programmers and in the development of major software packages. (IMSL, itself, grew largely out of a public domain library of matrix programs, EISPACK/-LINPACK.) These provided the backbone for many early matrix implementations. They have since evolved to the (still) public domain package LAPACK (www.scd.ucar.edu/softlib/LAPACK.HTML and, see also NetLib at ORNL).

Integrated Packages

Early econometrics programs were single purpose Fortran programs that users passed amongst themselves and which grew over time as they were extended. The concept of a multipurpose, integrated econometrics package emerged in mid-1960s in the form of three particular programs, TSP (Time Series Processor), the Troll system, and B34S (which began as BiMed34; see Stokes, 1997). TSP (23) and B34S (1) have remained active and are still under development. The early incarnations of these packages were primarily devoted to different forms of linear regression and correlation, but they have grown into vast assemblages of hundreds of thousands of lines of code (still mostly Fortran). Development of today's two largest commercial statistics and econometrics packages, SAS (17) and SPSS (20) also began at this time, though generally in social sciences other than economics and focused on techniques largely different from those of interest to econometricians. The demise of the distribution of custom written, single purpose software began in the 1970s with the emergence of TSP, SAS, SPSS, and so on.

Development Settings

With the exception of E-Views (3), which is a fairly recent arrival, and of the language platforms, MatLab (13), S-Plus (16), Gauss (5), Maple (11), and Mathematica (12), which are arguably successors to Fortran, most of the major packages in current use began development as part of a cottage industry in the 1960s and 1970s, usually by an individual or small team of innovators. This is a distinct difference from the contemporary, corporate route to major software development. The modern development centers, such as SAS, Inc. and SPSS, Inc. are creations of the 1980s and 1990s. Nonetheless, many of the programs listed in the Appendix to this survey continue to be developed by small numbers of statisticians and econometricians whose names would be recognized by practitioners (most likely in association with literature encountered as part of their graduate studies).

Programming Languages

Wither the primitive programming languages, Fortran and C? Fortran dominated the scientific computing landscape for nearly four decades until the early 1990s, and remains the language of choice for many scientific applications, e.g., in physics and astronomy. The practitioner interested in writing a Fortran or C program may use one of the development kits for the PC (Windows) setting, such as Lahey (7) and Watcom (24). Time-shared systems, such as Sun Microsystems, also provide Fortran and C compilers and program libraries as part of the operating platforms. These vendors have extended the lives of Fortran and C as numerical language platforms, but this style of programming is a dying art. For software developers, Fortran is giving way to C, for example, in the full conversions of RATS (Regression Analysis of Time Series) (15) and SAS (17) and in the mixed language construction of LIMDEP (8). Individuals doing custom numerical programming increasingly rely on Gauss and its counterparts, MatLab, etc.

Recent developer's programming language implementations have gravitated toward the studio format. The studio platform consists of integrated tools for merging code in one or more languages along with numerous libraries and ancillary tools. The Microsoft Software Developers Kit (SDK) is a popular example. This approach to programming has become essential for the development of major programs. One of the very useful innovations is the seamless creation of mixed language (e.g., Fortran and C++) platforms that allow developers to exploit the best features of both these languages. The end result is that C, its successor, C++, and Fortran, continue

to play a major role in the development of modern econometric software, but only in the laboratories of developers, rather than in the toolkits of practitioners.

The Internet

Finally, in what may be a surprise given the breathless surrounding hype, it appears that the Internet has had little or no impact on econometric software beyond facilitating the transfer of files, programs, and information. It is otherwise difficult to discern any ways in which the actual mechanics of econometric computing have been affected by the creation or expansion of the Internet.

The Characteristics of Modern Software

The software used by modern econometricians is essentially of two types. Most empirical research is done using one of the large, integrated packages that contain collections of programs for a variety of models. (See the 'Econometrics' list in the Appendix.) In these programs, complicated models are built with single, compact instructions. If a technique has gained acceptance in the applied literature, then it will appear as an 'option' in at least some of these packages. In contrast, when a technique or model proposed by an econometrician involves new methods or novel kinds of computations, then it is not likely to exist yet as a feature in an existing estimation platform. The computational tools that this analyst needs will typically be developed using one of the more primitive programming languages, such as Gauss (5), MatLab (13), S-Plus (16), or, at an even lower level, one of the native programming languages, Fortran or C++.

Packaged Expertise

The term canned routine has long been used for programs that embody the econometric expertise needed by users. Some have argued that the use of such software locks practitioners into a limiting set of models. While the argument may once have had some merit, the huge variety of techniques and the customizability provided by most large packages makes the criticism moot for current purposes. In addition, contemporary software is mature enough that developers have incorporated optimal algorithms that purists who mimic the textbook mathematical formulations might not use, or even be aware of. Least squares provides an excellent example. The analyst who writes his own program for linear least squares regression is likely to rely upon the textbook matrix formula, $b = (X'X)^{-1}X'y$. But, this is not the most accurate method of computing a least squares coefficient vector. Many large packages use much more accurate techniques, such as QR and singular value decompositions that avoid inversion and moments accumulation altogether. For the individual researcher to develop complicated low level tools such as this may well be excessively difficult.

User Written Programs

The distinction between these two types of software is somewhat blurry since the programming languages are often distributed with packages of routines that have been written by expert users. In addition, developers of code for Gauss (5) and MatLab (13) often distribute their programs on the Internet simply by posting them on their own web sites, perhaps, along with the research report of which the software development was a part. For example, all the major integrated econometrics packages contain an estimator for the logit model for binary choice, whereas the Gauss user must program their own. However, so many widely distributed Gauss libraries of routines contain users' versions of the logit program that for practical purposes, Gauss does contain a logit estimator. In the other direction, large integrated packages usually contain their own 'programming languages,' which allow the user to construct capabilities that are not already contained in the program. Thus, while a program such as SAS (17) may not contain a specific command for estimation of a certain kind of survival model, it does contain a programming language for matrix manipulation as well as a means for a user to access certain utilities such as a nonlinear optimization program. Thus, the SAS user can develop a new estimator within his familiar SAS platform. TSP (123), SAS (17), SPSS (20), Stata (21), Shazam (18), EViews (3), and LIMDEP (8) to name a few, are all large integrated packages that also contain extensive matrix and data manipulation languages and access to nonlinear optimization programs which allow users to extend the modeling capabilities of the package. Thus, software developers and their users have blurred the distinction between preprogrammed procedures and programming languages.

Application Studios

The modern applied econometrician will usually maintain a toolkit of software consisting of one or more large packages, and, for less frequent usage, a low level development kit such as Gauss or MatLab, or a Fortran compiler, such as Lahey (7). Developers, conscious of this and sensing an opportunity, have begun to gravitate to a developer's studio format for econometric software. As noted, all of the large packages provide low to medium level programming capability. Current trends, such as EViews, LIMDEP, RATS, SAS, Stata, and TSP have brought the programming features closer to the front end of these large, high level packages and Gauss with its large, expanding, and widely distributed library is approaching the same format from the opposite direction.

Common Features

Most general, integrated econometrics programs are quite similar. The model-building process dictates the design, and as in many settings, a standard has evolved naturally. The following functions will be found in all the integrated packages, and, over time, will be mimicked by the users of the language packages as they collect their own routines:

1. Data input and transformation;
2. Model estimation – the estimation of model parameters, usually of several types;
3. Statistical inference – testing hypotheses and forming probabilistic statements about quantities of interest;

4. Specification, testing, and iterative refinement of model specification;
5. Manipulation of model results for forecasting and simulation; and
6. Graphics and display of results.

Programs differ in some features and there is some amount of specialization, but most or all of the preceding will be found common to all. (Reviews are regularly published, e.g., in the *Journal of Applied Econometrics*.) The second function listed is one of the major dimensions on which programs differ. For example, EViews and RATS are particularly oriented toward analysis of time series data, SAS and Stata are especially complete in their presentation of techniques used in biomedical research, and LIMDEP is best known for its selection of models for discrete dependent variables. (Markets being what they are, there is substantial and growing overlap of these features.)

Interfaces

The means of communication between the user and the software is the interface. In order to describe this, it is necessary to make a distinction between mainframe and 'PC' (desktop style) operation. A second distinction between 'batch' and 'interactive' mode use is also necessary. The two distinctions are useful for describing the way certain computer programs work, although both modes do appear in both settings. Econometric software design began on large time shared mainframe systems. The design of such systems necessitated a batch mode form of interface. In this case, the user submits a set of instructions to a computer (operator, real, or virtual) and returns later to retrieve their desired results. Modern computing carried out on the desktop, in contrast, usually involves an individual operating their own system interactively in the fashion of a large calculator, submitting single instructions (perhaps in the form of mouse clicked selections from menus) and instantly receiving the results of their directive. Econometric analysis often requires a hybrid of these two modes. Many estimation problems require huge amounts of computation that may take hours or days to complete, or lengthy scripts of operations. As such, a batch mode of operation is often necessary, even on a desktop computing platform. Therefore, there is normally some means of giving the program a directive equivalent to 'carry out the following sequence of commands,' or 'fetch the set of commands contained in a certain file on disk and carry them out.'

There is a qualitative difference in the typical look and feel of a desktop computer compared to a time-shared mainframe. Modern desktop software is based on graphical user interfaces (GUIs). The amount of menuing varies from one program to another, but all modern econometrics programs are based on the GUI format. In contrast, there is relatively little GUI in the time-shared environment. The distinction is only superficial, however. The numerical functionality of most econometric software is easily replicated in both mainframe and microcomputer environments (e.g., all of the econometrics programs listed below except EViews). This is one of the strong distinctions between econometric software and, say, the programs used for graphics design or word processing. In the final analysis, the purpose of most econometric software is numerical analysis that can be done equally well in batch or interactive mode.

Estimation Procedures

To describe some of the remaining features, it is convenient to define some terms. Estimation is a process of culling from a data set a set of statistics that summarize the process by which the data were created. Estimation results range from single numbers, such as the parameter β in our earlier demand equation example, to one-to-one transformations of the original data, such as a periodogram, nonparametric hazard function, or kernel density estimator. Econometric software will contain 'estimation routines.' The term derives from the earliest days in which Fortran 'subroutines' were written for these purposes. The estimation routine consists of a body of code written to be invoked with a simple 'command' and provided with the specifics of a user's problem. The command itself or the broader context of use of the software must define the kind of model to be estimated. Commands might be of the form 'Using the data on Price and Quantity, fit a linear regression model.'

For example, one popular program, LIMDEP, uses the following, fairly typical style:

$$\text{Regression; Lhs} = \text{Quantity; Rhs} = \text{One, Price \$}$$

that states in words that the user wishes the program to compute the constant term (ONE) and slope in a linear regression of quantity on price, $Q = \alpha + \beta P + \varepsilon$. All integrated packages have counterparts to this kind of command, which is provided to the program by means of some kind of text processor. Some packages provide the same functionality by asking the user to select the estimation framework, then the variables in the equation from menus, instead. Many programs also allow users to choose either mode of entry. Each has its virtues – the former allows recording of steps and reuse of commands, whereas the latter is generally more convenient.

A programming language such as Gauss would not contain such a broadly defined procedure. Rather, the Gauss processor would allow the user to define, using matrix expressions, the algebraic steps involved in doing the computation, including how to compute the coefficients for the model and how to compute certain statistics that the integrated processor would produce automatically. The user would then use the language's built-in functions for displaying the results of their computations. Thus it would seem that a great deal more work is required of the Gauss user than the LIMDEP user. There is, but two issues motivate the former. First, as noted, many such sets of computations already exist, in elaborate routines that programmers have written and disseminated, so that these procedures can be used, often with a single command. Second, many estimation techniques do not exist (yet) in large integrated packages, so this is the only way to proceed. (Since this market is fast moving, it would be inappropriate to mention specific techniques. By the time this article appears in print, the information will be out of date.)

The remaining functions involve manipulation of model results. Some of these are handled automatically by the software while others must be assembled by the user. At this

juncture, the user of the integrated procedure is likely to use some of the programming features of the package, such as a matrix computation. For example, the literature often produces new statistics or ways to manipulate the results produced by familiar models such as the linear regression model. The SAS, Stata, or TSP user is likely to handle this situation by constructing a 'procedure' with 'adjustable parameters' that might appear as follows:

'Procedure: Regression;
Dependent variable is (Y), Independent variables are (x1, x2)
Compute residuals as variable (e).
Compute and report statistics using procedure …
End of Procedure'

Now, to invoke their procedure, the user might use a single directive such as:

'Use Regression Procedure (Quantity, One, Price).'

The generality of the procedure allows it to be used for different problems just by changing the 'call' statement. At this point, the only difference between Gauss and SAS is the detail that will have to be programmed in the first line of the procedure.

Accuracy

One might think that given the textbook formulas for the necessary computations, all programs would get the same right answer for any particular problem. But, in fact, this is far from correct. Even in the absence of programming errors, there are large differences in the algorithms and methods used in computer packages. For practical purposes, two approaches are used for validating econometric software: replication and benchmarking. The first of these is strictly pragmatic. For most techniques, there are no applications with known *correct* answers against which to benchmark a program. As such, it is assumed that if two or more of the major programs give the same answer for a problem, then they both have it right. If they differ, then the tricky issue arises as to which, if either is correct and, again, the issue would have to be resolved in ad hoc fashion. A contemporary study that pursues this line of inquiry is Silk (1996), whose review of estimators for simultaneous equations models presents some surprising differences (see also, McCullough and Vinod, 1999.) The same approach has been taken in a review of programs for the computation of GARCH models by McCullough and Renfro (1998).

True benchmarking is now available for many common calculations. The tradition in econometrics begins with Longley's (1967) true results to 10 significant digits for a specific problem in linear regression. The Longley standard was used for many years as the yardstick for linear regression programs; if a program could match the Longley benchmark to at least eight digits, then it could probably rise above suspicion for practical usage. Later analysis has showed that the Longley hurdle is not particularly high, and clearing it is not all that difficult, even for fairly mediocre software. The Statistical and Engineering Division of the National Institute of Standards and Technology (NIST) has published (www.nist.gov/itl/div898/strd) a suite of 57 benchmark problems in four classes: univariate statistics, analysis of variance, linear regression, and nonlinear regression. The NIST problems have been solved using very high precision methods, and correct answers have been published. In detailed reviews, McCullough (1998, 1999) found fairly wide differences in the performance of several econometrics packages. Since econometrics programs tend to be revised, refined, and updated very frequently, these differences are likely no longer to be evident by the time this article is in print.

The Role of Software in the Development of Econometrics

A convincing case can be made that several parts of modern econometrics awaited the development of software to gain currency.

In primitive incarnations, econometric software operated on a table of data with variables in columns numbered from left to right. To communicate a desired computation to a program, the user would specify columns of the table, as in 'Regress the variable in column 5 on a constant, and the variables in columns 6, 8, and 10.' TSP was among the first to allow names for variables to appear in a symbolic instruction to the program. The software went an important step further in that it allowed the components of equations to appear symbolically in words and familiar algebraic symbols. Software had begun to evolve in this direction, but TSP brought the language of equation specification up to the level of compilers, such as Fortran. This major innovation changed the standard for econometric estimation programs.

The most important single development in practical econometrics in the period prior to 1970 was arguably the LSQ (Least Squares) procedure in TSP. The combination of symbolic equation specification with TSPs LSQ procedure allowed the birth of a very large literature on linear and nonlinear equation systems. The seemingly unrelated regressions model had been proposed a decade prior by Zellner and Theil (1962) (with Zellner, 1962), but the model did not appear in wide use until the appearance of LSQ. The procedure went an important step further in that it allowed the Zellner estimator to be extended to nonlinear systems of equations.

At roughly same time, Joreskog proposed a maximum likelihood approach to factor analysis and the estimation of general linear equation systems. The analytic results had been known for some time, but solvers for the likelihood equations for a general hierarchical linear equation system had not been widely distributed. Joreskog's (1970) early work included 'A Computer Program for Minimizing a Function of Several Variables,' (Educational Testing Services: Research Bulletin 70-14). In this was published the numerical engine of a package published 2 years later as 'LISREL' (Linear Structural Relationships – originally, ACOVS for Analysis of Covariance Structures). ETS 70-14 contained a listing of Fortran computer code for optimization, including a very clever and effective technique for the line search step in optimization. LISREL spawned a very large and still growing industry in structural equations analysis in the social sciences. The technique is a major part of the methodology; the program, itself, has become Scientific Software, Inc. (9).

Of somewhat more recent vintage, is Heckman's (1979) sample selection model. Building on his work on labor supply models, Heckman proposed a method for a regression model with a particular type of nonrandom sampling in a seminal paper in 1979. The computations involved are now fairly routine and incorporated in most econometrics packages, but in the mid-1980s, the technique became a standard tool in the microeconometricians kit as the LIMDEP program (8) became the *de facto* standard approach for implementing it.

The work of Sims (1977), Granger (1969), and Litterman (1986) in the mid-1970s on vector autoregressions (VARs) in econometric time series analysis produced a new body of theory to challenge the structural simultaneous equations models. The RATS (15) computer program, developed by these authors carried the methodology into the mainstream of applied econometrics. The development has also been ongoing, both of the program (Estima, Evanston, IL) and of the methodology as can be seen in any reference on time series, such as Hamilton (1994). (RATS has recently spawned CATS – Cointegration Analysis of Time Series.) Once again, the computations and the methods are not overly difficult, but the fact that they have been conveniently integrated into a general-purpose package has been an important impetus to the dissemination of the methodology in the applied literature. RATS was also an early contributor among several that significantly advanced practice by providing the user with access to an optimization program that required only a basic, again symbolic, specification of the problem. This freed the researcher from having to learn a large amount of programming. Contemporary programs, e.g., TSP, LIMDEP, and SAS, still RATS, Gauss, and so on, all offer this capability.

The earliest work in applied econometrics required the analyst either to use someone else's program or to know how to write their own Fortran code. The latter type were in a distinct minority, so most researchers used software written by others. Current practice and contemporary software make it possible for researchers to choose among a vast array of pre-packaged techniques and, with the aforementioned MAXIMIZE capabilities and the matrix programming languages of Gauss, B34S, LIMDEP, RATS, and so on, to create another vast array of extensions with easily acquired programming skills. The result has been an explosion in the variety of techniques and model formulations found in the published research.

Conclusion

The development of software has been a facilitator, and a platform for the dissemination of results and continued research in econometric techniques and methodology. But there are cross-currents in contemporary software design. As of 1999, econometric software design and publication has left the cottage industry of the 1960s and 1970s far behind; it is usually (though not always) a corporate endeavor. At the same time, technique, development, and extension of econometric methodology remain the bailiwick of a scientific community which has, to some extent, preserved the distinction between the purely commercial nature of, for example, the word processor market and the romantic, if antiquated, free exchange of scientific software among individuals and small teams of researchers. However, market competition appears to be working smoothly in the right direction, and no case could reasonably be made that market forces have inhibited the development of econometric techniques or software in the way they have, e.g., in the markets for operating systems and Internet browsers. Technological advance in software and econometric technique appears to be rapid and uninhibited.

Appendix: Some of the Most Commonly Used Econometric Software

	Software	*Type*	*Web site*
1	B34S	Econometrics	www.uic.edu/~hhstokes/hhstokes/b34s.htm
2	BMDP	Library	www.meridian-marketing.com/BMDB/index.htm
3	E-Views	Econometrics	www.eviews.com
4	Excel	Spreadsheet/Data	(Microsoft Office 2000 – commercial)
5	Gauss	Programming language	www.aptech.com
6	IMSL	Library	www.mathstat.com.au/vni/imsl/index.html
7	Lahey	Fortran development	www.lahey.com
8	LIMDEP	Econometrics	www.limdep.com
9	LISREL	Econometrics	www.ssicentral.com/lisrel
10	Lotus	Spreadsheet/Data	(Lotus Smart suite – commercial)
11	Maple	Mathematical analysis	www.maplesoft.com
12	Mathematica	Mathematical analysis	www.wri.com
13	MatLab	Programming language	www.mathworks.com
14	NAG	Library	www.nag.com
15	RATS	Econometrics	www.estima.com
16	S plus	Programming language	www.mathsoft.com/splus
17	SAS	Econometrics	www.sas.com
18	Shazam	Econometrics	shazam.econ.ubc.ca
19	Soritec	Econometrics	fisisoft.com/products
20	SPSS	Econometrics	www.spss.com
21	Stata	Econometrics	www.stata.com
22	Systat	Econometrics	www.spss.com/software/science/systat/
23	TSP	Econometrics	www.tspintl.com
24	Watcom	Fortran development	www.powersoft.com/products/languages/fort77.html

The above is a representative list of the most widely used econometric software. For additional information about the individual programs, the reader may refer to the web sites listed. Some of these programs are listed as econometrics although econometrics is not the major functionality of the program neither is its largest user base econometricians. These would include, for example, SPSS and Stata.

Bibliography

Draper, N.R., Smith, H., 1981. Applied Regression Analysis, second ed. Wiley, New York.

Granger, C., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424–438.

Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton, NJ.

Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–161.

Joreskog, K., Gruvaeus, G., 1970. A computer program for minimizing a function of several variables. educational testing services. Research Bulletin Number 70-14.

Kalbfleisch, J., Prentice, R., 1980. The Statistical Analysis of Failure Time Data. Wiley, New York.

Krailo, M., Pike, M., 1984. Conditional multivariate logistic analysis of stratified case-control studies. Journal of the Royal Statistical Society, Series A-Statistics in Society 84, 95–103.

Litterman, R., 1986. Forecasting with Bayesian vector autoregressions – five years of experience. Journal of Business and Economics Statistics 4, 25–38.

Longley, J., 1967. An appraisal of least squares programs from the point of the user. Journal of American Statistical Association 62, 819–841.

Maddala, G.S., 1983. Limited Dependent and Qualitative Variables in Econometrics. Cambridge University Press, New York.

McCullough, B.D., 1998. Assessing the reliability of statistical software. Federal Communications Commission, Manuscript.

McCullough, B.D., 1999. Econometric software reliability: reviews, LIMDEP, SHAZAN, and TSP. Journal of Applied Economics 14 (2), 191–202.

McCullough, B.D., Renfro, C., 1998. Benchmarks and software standards: a case study of GARCH procedures. Journal of Economics and Social Measure 25, 59–71.

McCullough, B.D., Vinod, H.D., 1999. The numerical reliability of econometric software. Journal of Economic Literature 37, 633–665.

Nerlove, M., Press, S., 1973. Univariate and Multivariate Log–linear and Logistic Models. RAND-R1306-EDA/NIH, Santa Monica, CA.

Press, W., Flannery, H., Teukolsky, S., Vetterling, W., 1986. Numerical recipes: the art of scientific computing. Cambridge University Press, Cambridge.

Silk, J., 1996. Systems estimation: a comparison of SAS, Shazam, and TSP. Journal of Applied Economics 11, 434–450.

Sims, C., 1977. Exogeneity and causal ordering in macroeconomic models, new methods in business cycle research. In: Proceedings from a Conference. Federal Reserve Bank of Minneapolis, MN, pp. 23–43.

Stokes, H.H., 1997. Specifying and Diagnostically Testing Econometric Models, second ed. Quorum Books, Westport, CT.

Zellner, A., 1962. An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias. Journal of the American Statistical Association 57, 500–509.

Zellner, A., 1971. Introduction to Bayesian Inference in Econometrics. Wiley, New York.

Zellner, A., Theil, H., 1962. Three stage least squares: simultaneous estimation of simultaneous equations. Econometrica 30, 63–68.

Econometrics, History of

Mary S Morgan, The London School of Economics and Political Science, London, UK
Duo Qin, School of Oriental and African Studies, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The statistical side of economics began to emerge in the late nineteenth century, but it was only in the 1930s and 1940s that its conceptual and technical foundations were established under the label of econometrics. From the 1950s onward, as the discipline matured, a variety of approaches flowered, prompting debates not just about technical issues, but about foundational issues of modeling and statistical approaches.

Econometrics, the statistical metric for economics, can be regarded as one of the main innovations that turned twentieth century economics into an engineering, or tool-based science, in which each application of economic theory requires special shaping to circumstances, whether for scientific purposes or in the policy domain (see Morgan, 2001). The development and use of statistical tools for economics emerged in the early part of the twentieth century and, by midcentury, problems had been defined, solutions approached, and usable concepts developed, so that one could legitimately refer to a distinct body of knowledge embracing both theory and practice. After 1950, econometrics became a mature field and the dominant method of applied economics.

The particularities of the history of econometrics have often been linked to the nature of its tasks and aims (see De Marchi and Gilbert, 1989; Morgan, 1990; Qin, 1993; Boumans, 2005; Gilbert and Qin, 2006; Le Gall, 2007; Louçã, 2007; Boumans et al., 2011; Qin, 2013). Economists have understood their fundamental methodological problem to be: how to establish knowledge about economic phenomena and economic laws, given that most economic activities cannot be subjected to controlled experiments. Econometricians have assumed that, this means that descriptions, regularities, and relations of the economy must be extracted from 'passive' observations: statistical data that result from the natural course of economic activities, with the use of statistical theories and tools. Econometrics therefore suffers from all the usual problems of an inductive science. When measurement or testing of theories is at stake, severe correspondence problems arise because of the huge gaps between the assumptions of those economic theories and the conditions of data collection.

The history of the field also has its external dynamics (Desrosières, 1998; Epstein, 1987). To begin with, technological changes play an important role, evidenced in the development of the field of mathematical statistics and machines of computation (Renfro, 2011). The provision of data, the availability of trained econometricians, and fashions in economic theory are all relevant to the narrative, as are the stimuli from twentieth-century economic events and associated demands for econometric expertise. Econometrics in the latter twentieth century became a mainstay of economic policy analysis and tool for economic intervention, yet at the same time it grew into one of the most technically demanding and methodologically sophisticated areas of economics.

The Development Era

In the latter nineteenth century, a few economists began to use graphs and tables of statistical data to provide empirical support for their arguments, or to describe behavioral characteristics of economic phenomena (Klein, 1997; Maas and Morgan, 2002). They were quick to take up certain statistical techniques developed for biometrics, namely correlation and regression (Aldrich, 2010). But these 'measures of association' were not entirely adequate for economists' questions (Aldrich, 1995; Morgan, 1997). Where biometricians had, for example, a sample of heights of fathers and their sons, whose relation could be assessed with correlation or regression coefficients under reasonable assumptions about the individuals in the sample, economists typically used time-series observations that provided only one datum for each time date. Not only were such 'samples' considered peculiar from the viewpoint of the correlation theory of that time – which assumed that the observations within the sample had been independently drawn – but economists also believed that such time-series observations reflected a number of concurrent behavioral characteristics or causal relationships varying with the time element and unit chosen.

This problem of extracting causal relationships between variables from data that contained different time patterns dominated early econometric work, but the questions attacked were driven by economic and policy concerns. Measurements, for example, of market demand and supply relations for agricultural goods were undertaken at the American Bureau of Agricultural Economics during the 1920s when farmers faced falling prices in international markets and followed earlier European breakthroughs on these questions. Similarly, the problems posed by 'business cycles' were widely recognized by commerce, politicians, and economists, for the crash of 1920–21 was only outdone by the even more spectacular and longer-lived depression of 1929–33. Business cycle research institutes, which had begun to emerge even before the end of World War I, sprang up everywhere in the 1920s from Massachusetts to Moscow, but their truly international quest for description and explanation of the business cycle showed that the task of establishing measures of – let alone explanations for – that phenomenon, its characteristics, and its time patterns was incredibly difficult (Armatte, 1992).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71077-4

A concerted movement to create a more 'scientific' economics, based on both statistical and mathematical tools in economics, resulted in the foundation of The Econometric Society (in 1932), and its house journal *Econometrica* (in 1933) (Bjerkholt and Dupont-Kieffer, 2010; Louçã and Terlica, 2011; Bjerkholt and Qin, 2011). The most critical innovation of that decade, often overlooked, was that of the 'econometric model' formulated as an intermediary device to bridge the gap between economic theory and economic data. Such models represented elements of the economic theory in mathematical form, but were also attuned to the peculiarities of the time patterns of data evidence. With the use of such models, various kinds of knotty correspondence problems at the intersection between economic theory and data begun to be understood and solutions proposed. For example, the possibility of determining the individual effects of the contemporaneous variables in a system of simultaneous equations (the identification problem) was separated from the general problem of disentangling statistical relations where economic variables proved interdependent in multiple ways (the confluence problem), and where measurement errors were critical to the outcomes. But these 1930s answers were proposed as one-off solutions to specific applied problems, not general solutions for generally recognized problems.

However, econometrics was not without its detractors and the move to codify and formalize these somewhat-disparate concepts, problems, and emergent solutions seems to have been triggered by the Keynes–Tinbergen debate in the late 1930s. Jan Tinbergen, a young Dutch economist, was the most creative of the econometricians of the 1930s in the development of modeling, understanding conceptual problems, and suggesting solutions. He was also responsible for building and using the first macroeconometric model to analyze policy options. In this work, he developed ideas by Ragnar Frisch, a Norwegian econometrician (which later made them joint winners of the first Nobel Prize in economics in 1969) who in turn had drawn on the statistical work by English statistician Yule and Russian econometrician Slutsky.

Tinbergen's macroeconometric modeling work was heavily attacked by Keynes, and this attack, along with a number of other developments, prompted a search in the 1940s for more rigorous foundations for econometrics. Despite their difference of opinion, both Tinbergen and Keynes can be seen, in retrospect, as reshaping economic theory and policy in the aftermath of the Great Depression so that, in the 1950s, Lawrence Klein (Nobel prize 1980) could develop the postwar generation of macroeconometric models on Keynesian theories with more secure statistical foundations.

The key development in this foundation work was the 'probability approach,' fashioned in the early 1940s by Trygve Haavelmo (Nobel Laureate in 1989 for this work) (Bjerkholt, 2005, 2007; Bjerkholt and Dupont-Kieffer, 2011), a Norwegian econometrician, who proposed to use probability theory to establish a bridge between economic theory and passive economic observations (Boumans, 2010, 2014). His bridge was designed to carry the burden in both directions: to provide guidance on how to specify econometric models as statistical hypotheses so that not only could economic theory be logically related to economic data but that there would also be a path for inference back from the data to the theory.

Establishing the Mature Discipline

Haavelmo's probability blueprint for econometrics had indicated directions for both theoretical and applied work and has been interpreted as creating a 'revolution' in econometrics. Parts of Haavelmo's program were immediately elaborated at the Cowles Commission (later Cowles Foundation), an American hothouse of econometric research in mid 1940s (see Hildreth, 1986). Their research was primarily concerned with developing appropriate identification conditions and estimation (measurement) techniques for a system of simultaneous equations, assumed as a general representation of econometric models. In the years after 1950, an established orthodoxy developed based on the Cowles Commission's work and enshrined in textbooks in the 1960s. But, despite an active program of research at the theory level, providing technical extensions to this groundwork, cracks in the orthodoxy began to appear.

Applied econometric research revealed serious practical and methodological limitations in the Cowles approach, particularly in regard to model specification, model choice, and associated testing procedures. Their assumption that econometric work is anchored on a priori theories meant that it was rarely possible to falsify those theories empirically. The validity of the assumption was questioned in various contexts in the 1950s. For example, the Swedish econometrician Herman Wold, who was trained in the Russian probability school and time-series analysis in the 1930s, attacked the Cowles focus on simultaneity and developed alternative dynamic interdependent equation models in the 1950s (Morgan, 1991). But systematic contentions did not arise until after 1970. One of the first alternative approaches to challenge the Cowles approach was the specification search and sensitivity analysis using Bayesian methods developed by Ed Leamer (Qin, 1996, 2013). Other alternative approaches targeted the lack of attention to the characteristics of time-series data in the Cowles methodology. One such was the Vector Autoregressions (VARs) modeling approach pioneered by the American economist Thomas Sargent and econometrician Christopher Sims during the 'rational expectation' movement in macroeconomics (Qin, 2011). The VAR models effectively generalized Wold's dynamic models and turned a priori theories into possibly testable hypotheses embedded in VARs. But the methodologically most comprehensive approach was the dynamic specification approach initially developed at the London School of Economics (LSE) and led mainly by David Hendry (Qin, 2013). The LSE approach inherited the time-series research tradition from the Department of Applied Economics at Cambridge in the 1940s and 1950s under Richard Stone (Nobel Laureate in 1984), and synthesized various elements from the Bayesian approach and the VAR approach. It not only turned a priori theories into testable hypotheses, but also formulated them into a model type called an 'error correction model,' analogous in form to servo-mechanistic control engineering, and which could be more readily understood in terms of individuals' reactions to changing economic events. The error correction model subsequently became associated with the 'cointegration' theory developed mainly by Clive Granger (Nobel Laureate in 2003) to capture the long-run tendency of stochastic

cotrending among economic variables as an equilibrating force. These developments have effectively reversed the Cowles approach and made model choice and design depend explicitly on both a priori theories and data features.

On the other hand, the Cowles tradition was strengthened by the emergence of an especial microeconometrics that occurred concurrently with these changes in macroeconometrics. In the early 1950s, Stone had struggled with the immense task of providing measurement structures appropriate for the separate components of demand in a total consumer expenditure system: a system simultaneous at the microlevel. By the 1970s, various statistical models appropriate for explaining behavior at the individual level had been adapted from 1930s biometrics into econometrics by Dan McFadden and James Heckman (joint Nobel Prize in 2000). Their treatment of survey data enabled measurement models to be developed for the discrete choices made by individuals. Such discrete choices dominate economic decision making in consumption decisions, work decisions, family formation decisions, and so on. These measurement models were quickly taken up for they matched the individual utility-maximizing framework of neoclassical economic theorizing, the main orthodoxy of post-1950s western economics. Since the focal task of microeconometrics was to measure specific effects dictated by a priori theories, which amounted to utilizing the survey data information in a very partial way, the Cowles approach was found to fit well with the purpose (Heckman, 2000, 2005).

However, the growth of economic data has encouraged various data-instigated research to go beyond any straightjacket methodology. In financial econometrics, for example, the research for developing various composite indicators, reminiscent of the 1930s business cycle work, has become increasingly popular. In empirical economics, the rapid expansion of computable general equilibrium analyses, developed by Finn Kydland and Edward Prescott (joint Nobel Prize in 2004) in the early 1980s, has resulted from a hybrid of the Cowles approach with computer simulation methods. Such 'microsimulation' methods were pioneered by Guy Orcutt in the later 1950s to analyze demographic and social structural changes in association with time-series statistical analyses (see Morgan, 2012: Chapter 8). In computable general equilibrium models, parameters are 'calibrated' rather than statistically 'estimated,' although the calibration process takes into consideration the existing econometric evidence concerning the plausible parameter magnitudes and also the time-series properties of the modeled variables. Finally, explicit notions of experimental design are now found in the way econometrics is used to measure experimental effects in 'natural experiments': this has regenerated the founding fathers' earlier understanding of econometrics as a statistical substitute for experiment on Nature's passive data.

The Place of Econometrics in Economics

During the early years of The Econometric Society, 'econometrics' had been defined as the union of economics, mathematics, and statistics. And, in developing model intermediaries based on mathematical economic theory in that period, econometricians had developed a distinctive path compared to those found in psychometrics and sociometrics in the same period. After 1950, as economic theorizing became generally expressed in mathematical terms, the term econometrics stabilized on its current meaning: the use of statistical reasoning and methods as means to establish data-based descriptions of economic phenomena and empirically based counterparts for, and tests of, economic theories.

The spread of econometrics as an established part of economics in the post-1950s period may be understood as due not only to the strengthening of its foundations in statistical theory but also to the continuing expansion of data ('official' statistics), its establishment in the core undergraduate teaching program, and the development of cheap desktop computing. Together these meant that econometric work became a standard tool of policy work in governments and international agencies as well as becoming endemic, in various different forms, in the subfields of scientific economic research. Econometric theory has developed into a formidable body of specialist statistical theory and the increasing gap between the difficult technical and theoretical questions and the apparent ease of applications might indicate a field where applications came loose from theoretical work. This has been militated by another tool, the development of specific software packages for econometrics, in which theoretical and technical developments can be quickly translated into modeling, measurement, and testing regimes at the level of applications.

This widening use of econometrics did not denote a settled field. The community in the latter part of the twentieth century found itself immersed in a number of vibrant and important arguments about how to do econometrics, based upon applied econometric experience as much as on considerations of statistical theory and philosophy of science (Hoover, 2006; Spanos, 2006). Should models be theoretically based or descriptively accurate? What are the relevant criteria, economic and statistical, which can be used to test the adequacy of these different models? Should the modeling process start with simple models that grow more complex, or start general and aim to become simple? Should the probability framework be classical or Bayesian? What implications does this difference have for the types of measurements made and the possibilities of inference? How far can applied econometrics be formalized and codified apart from tacit knowledge? The terms of the continued debates around such themes point to a mature field, one intimately connected to both statistical and economic theory, but in which the foundational arguments raised in early work continue to reverberate to the most recent time.

See also: Bayesian Statistics; Bayesian Theory, History, Applications, and Contemporary Directions; Causal Analysis, History of; Causal Inference: Overview; Econometric Software; Economics, History of; Economics, Philosophy of; History and the Social Sciences; Macroeconomics, History of From 1933 to Present; Probability Theory: Interpretations; Simultaneous Equation Estimation: Overview; Social Science, The Idea of; Statistics, History of; Stochastic Models; Time Series: Cointegration; Time Series: Economic Forecasting.

Bibliography

Aldrich, J., 1995. Correlations genuine and spurious in Pearson and Yule. Statistical Science 10, 364–376.

Aldrich, J., 2010. The econometricians' statisticians, 1895–1945. History of Political Economy 42 (1), 111–154.

Armatte, M., 1992. Conjonctions, conjoncture et conjecture. Les baromètres économiques (1885–1930). Histoire & Mesure VII (1/2), 99–149.

Bjerkholt, O., 2005. Frisch's econometric laboratory and the rise of Trygve Haavelmo's probability approach. Econometric Theory 21, 491–533.

Bjerkholt, O., 2007. Writing with nowhere to go: Haavelmo in the United States, 1939–1944. Econometric Theory 23, 775–837.

Bjerkholt, O., Dupont-Kieffer, A., 2010. Ragnar Frisch's conception of econometrics. History of Political Economy 42 (1), 21–73.

Bjerkholt, O., Dupont-Kieffer, A., 2011. Ragnar frisch and the probability approach. History of Political Economy 43 (Suppl.), 109–139.

Bjerkholt, O., Qin, D., 2011. Teaching economics as a science: the Yale lectures of ragnar frisch. In: Bjerkholt, O., Qin, D. (Eds.), A Dynamic Approach to Economic Theory. Routledge, London, pp. 1–28.

Boumans, M., 2005. How Economists Model the World into Numbers. Routledge, Abingdon.

Boumans, M., 2010. The problem of passive observation. History of Political Economy 42 (1), 75–110.

Boumans, M., 2014. Haavelmo's epistemology for an inexact science. History of Political Economy 46 (2), 211–229.

Boumans, M., Dupont-Kieffer, A., Qin, D., 2011. Histories on Econometrics. Duke University Press, Durham and London.

De Marchi, N., Gilbert, C.L., 1989. History and Methodology of Econometrics. Oxford University Press, Oxford, UK.

Desrosières, A., 1998. The Politics of Large Numbers. Harvard University Press, Cambridge, MA.

Epstein, R.J., 1987. A History of Econometrics. North-Holland, Amsterdam.

Gilbert, C.L., Qin, D., 2006. The first fifty years of modern econometrics. In: Patterson, K., Mills, T.C. (Eds.), Palgrave Handbook of Econometrics, vol. 1. Palgrave MacMillan, Houndmills, pp. 117–155.

Heckman, J.J., 2000. Causal parameters and policy analysis in economics: a twentieth century retrospective. Quarterly Journal of Economics 115, 45–97.

Heckman, J.J., 2005. The scientific model of causality. Sociological Methodology 35, 1–97.

Hildreth, C., 1986. The Cowles Commission in Chicago, 1939–1955. Springer-Verlag, Berlin.

Hoover, K.D., 2006. The methodology of econometrics. In: Patterson, K., Mills, T.C. (Eds.), Palgrave Handbook of Econometrics, vol. 1. Palgrave MacMillan, Houndmills, pp. 61–87.

Klein, J.L., 1997. Statistical Visions in Time. Cambridge University Press, New York.

Le Gall, P., 2007. A History of Econometrics in France. Routledge, Abingdon.

Louçã, F., 2007. The Years of High Econometrics. A Short History of the Generation that Reinvented Economics. Routledge.

Louçã, F., Terlica, S., 2011. The fellowship of econometrics: selection and diverging views in the province of mathematical economics, from the 1930s to the 1950s. History of Political Economy 43 (Suppl.), 57–85.

Maas, H., Morgan, M.S., 2002. Timing history: the introduction of graphical analysis in 19th century British economics. Revue d'Histoire des Sciences Humaines 2 (7), 97–127.

Morgan, M.S., 1990. The History of Econometric Ideas. Cambridge University Press, Cambridge, UK.

Morgan, M.S., 1991. The stamping out of process analysis in econometrics. In: De Marchi, N., Blaug, M. (Eds.), Appraising Economic Theories. Edward Elgar, Aldershot, Hants, UK, pp. 237–263.

Morgan, M.S., 1997. Searching for causal relations in economic statistics: reflections from history. In: McKim, V., Turner, S. (Eds.), Causality in Crisis: Statistical Methods and the Search for Causal Knowledge in the Social Sciences. University of Notre Dame Press, Notre Dame, IN, pp. 47–80.

Morgan, M.S., 2001. Economics. In: Porter, R. (Ed.), The Cambridge History of Science, Modern Social and Behavioral Sciences, vol. 7. Cambridge University Press, Cambridge, UK.

Morgan, M.S., 2012. The World in the Model. Cambridge University Press, New York.

Qin, D., 1993. The Formation of Econometrics: A Historical Perspective. Clarendon Press, Oxford, UK.

Qin, D., 1996. Bayesian econometrics: the first twenty years. Econometric Theory 12, 500–516.

Qin, D., 2011. Rise of VAR modelling approach. Journal of Economic Surveys 25, 156–174.

Qin, D., 2013. A History of Econometrics: The Reformation from the 1970s. Oxford University Press, Oxford.

Renfro, C.G., 2011. Econometrics and the computer: love or a marriage of convenience? History of Political Economy 43 (Suppl.), 86–105.

Spanos, A., 2006. Econometrics in retrospect and prospect. In: Patterson, K., Mills, T.C. (Eds.), Palgrave Handbook of Econometrics, vol. 1. Palgrave MacMillan, Houndmills, pp. 3–59.

Economic Anthropology

Kalman Applbaum, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S. Narotzky, volume 6, pp. 4069–4073, © 2001, Elsevier Ltd.

Abstract

Anthropologists study economic practices as integrated features of specific societies, comparatively, and as a set of historical, often diffusionary, patterns of thought, power, and organization for production, distribution, exchange, and consumption. Their foundational concerns have shifted from the description and interpretation of reciprocity and redistribution and the technological and social capacities for sustaining livelihood in small-scale and nonmarket societies to the analysis of the wide variety of articulations with and transitions of nonindustrial societies to participation in worldwide economic formations. Increasingly, economic anthropologists are studying phenomena of direct concern to other economists and social scientists, such as the internal activities of transnational corporations, global financial markets, and the sustainability of contemporary economic systems.

Introduction

Economic anthropology is concerned with how human societies sustain themselves materially and socially through schemes of production, distribution, exchange, and consumption.

In keeping with the orientation of the discipline of anthropology as a whole, economic anthropologists have sought to understand economic practices (1) in relation to the specific societal context in which they occur, (2) comparatively, in relation to similar activities as they have been observed in other times and places, and (3) as a set of historical, often diffusionary, patterns of thought and organization. The first is intended to contribute to the understanding of given societies, in which economic processes are understood to be an integrated subset, while the second leads to the broadening and refinement of common economic concepts for the benefit of theoretical frameworks. The third has been predominated by the study of capitalism, colonialism, and development, as these have been the prevailing forces for economic change since the subdiscipline's start.

A Heterodox Alternative to Neoclassical Economics

From its inception, economic anthropology has implicitly or explicitly drawn many of its basic questions in relation to the more influential field of economics. While economics is a varied, dynamic discipline, one can nevertheless begin to trace the intellectual history of economic anthropology by citing the central premises of neoclassical economics and then showing how anthropologists have sought to differentiate their approach in opposition to or in supportive modification of these premises.

The Neoclassical Economic Approach: Maximization, Market Equilibrium, Stable Preferences

Gary Becker, professor of economics and business at the University of Chicago and winner of the 1992 Nobel Prize, says: "The combined assumptions of maximizing behavior, market equilibrium, and stable preferences, used relentlessly and unflinchingly, form the heart of the economic approach as I see it" (Becker, 1976: 5).

Anthropologists observe that these three assumptions comprise an implicit theory of social morphology and human nature with claims to ahistorical universality.

First, *maximization* means that people predictably "choose the alternative that gives the most satisfaction, value, or benefit in a context of limited means ('scarce resources')" (Plattner, 1989: 7). The notion that people everywhere maximize or 'economize' in this way suggests both the universally perceived experience of resource scarcity relative to needs and wants and a view of human beings as predominantly rational, choice-making creatures.

The emphasis on *market equilibrium* points, first, to the tendency in current economic thought for the economy and the market to serve as interchangeable concepts. The notion of equilibrium suggests that the market is self-correcting toward balance and efficiency (through laws of supply and demand, for instance) and providential in its propensity – when not meddled with – to contribute to the wealth of nations. Markets, in turn, can behave in this way only if all goods and services (including land, labor, and capital) are available to be traded and are convertible, especially through the medium of money.

Transactions in the market, finally, are based upon *stable preferences*, which are defined as being 'revealed' in the market through the choices consumers make. Market choices and transactions are held to add up to familiar concrete and abstract economic institutions, such as the used car or labor markets, rather than the reverse. They are also analyzed by some economists as determinative of behavior in many institutions not ordinarily thought of in economic terms, such as the family, ethics, communal solidarity, or art.

Within markets, thus broadly defined, individuals are seen to pursue their preferences in rational, maximizing, competitive, and ultimately self-interested ways, independent of or even in opposition to communal norms and expectations. The unit of analysis is the individual decision-maker rather than the group, which may itself be seen as an epiphenomenon of economic transactions. Geographical or sociocultural

environment, historical precedent, and various other intermediary institutions are factored out.

The Anthropological Position

There is an inherent discord between the kind of research inspired by these assumptions and that which is carried out by anthropologists. In result of the naturalistic, descriptive methodology of the discipline, anthropologists tend to find universal laws unhelpful.

Anthropologists have described a variable spectrum of economic systems. In doing so, they have found it necessary to modify the conventional economic lexicon (property, value, money, investment, etc.). Often, even a modified version of the economist's vocabulary is found wanting in its ability to capture local practices, and native terms must be used instead. *Kula, moka, taonga, hxaro, dana, giri, guanxi*, etc., refer to non-Western economic concepts, practices, and systems that have no equivalent in the West and are therefore unexplainable using the lexicon of conventional economics. Comparative analysis involving these culture-specific terms has yielded general insights that unfold by way of compelling analogies rather than obeisance to a purportedly universal set of laws.

Extrapolating collective phenomena from aggregates of individual behaviors is perhaps the greatest difference separating the two disciplines. Anthropologists do not conceive of human thought or action as dissociable from culture, and they take it as their charter to decipher local social logics of action. Even the sizable contingent of economic anthropologists who are model oriented, and/or who specifically study individual decision-making, resource allocation in competitive conditions, risk behavior, and similar mainstream microeconomic subjects – even those who find that their research supports formalistic assumptions of rationality and utility maximization – nevertheless, as Michael Chibnik says, take into account "the historical changes, cultural norms, and socioeconomic institutions that constrain the choices possible for different groups of people at particular places and times" (2011: 1).

In consequence, it makes sense to view economic anthropology's rich empirical and theoretical explorations simultaneously in terms of their contribution to disciplinary anthropological concerns and as part of a debate with economics and other policy and social science disciplines that have adopted neoclassical economic methods and assumptions.

Foundations and Orientations of the Subdiscipline

Gifts, Commodities, and the Nature of Exchange

The person generally credited with awakening scholarly interest in primitive economies was Bronislaw Malinowski (1884–1942). Malinowski described the kula system in the Trobriand Islands of Melanesia: "a form of exchange, of extensive, inter-tribal character … carried on by communities inhabiting a wide ring of islands, which form a closed circuit" (1922: 81). Two kinds of articles are exchanged in the kula: red shell necklaces, which move in a clockwise direction around the network of islands, and white armshells or bracelets, which circulate in the opposite direction. "Every movement of the Kula articles, every detail of the transactions is fixed and regulated by a set of traditional rules and conventions, and some acts of the Kula are accompanied by an elaborate magical ritual and public ceremonies" (1922: 81).

Malinowski believed he had discovered a counterexample to conventional economic views of his day: That an elaborate system of long-distance trade could not be organized without markets, money, or states; that the purpose of material exchange, perhaps especially in premodern societies, would be oriented only to utilitarian purposes; and that grasping, rational self-interest above all else would dominate systems of exchange everywhere.

The kula has been restudied and reevaluated, in keeping with current analyses of exchange. It also continues to be used as a staple in undergraduate education, at which level the key lesson may be that in the Trobriands and in many places where anthropologists have studied, the purpose, process, and effects of material exchange is, alongside or independent of whatever utilitarian purpose it may carry, inherently *social*. As Strathern and Stewart say of the kula, "The giving of an item, or a series of items, was explicitly designed to obligate a partner to make an appropriate return later. Thus, gifts bound people together, and in doing so helped to create the fabric of society on an interpersonal basis" (2012: 230).

This 'gift' aspect of the kula fascinated a contemporary to Malinowski, the French sociologist Marcel Mauss (1872–1950), who cited the kula, among other examples, to make several arguments that dominated exchange theory in anthropology for several decades. Mauss' 1925 essay, *The Gift: Forms and Functions of Exchange in Archaic Societies*, contained or implied the following claims:

1. The impersonal and individualized form of commodity exchange dominant in modern societies is not primordial or universal.
2. Systems of credit and contractual obligations are ancient and, in variant forms, widely found.
3. Prestations, as Mauss referred to gift exchange in early societies, were characterized above all by their composite or 'totalizing' nature, in two respects. First, the phenomena are not discrete: each "contains all the threads of which the social fabric is composed … religious, legal, moral, and economic" (Mauss, 1967: 2). Second, in such societies one often finds entire groups and not just individuals participating in exchange events. Thus, the logic of gift exchange was not to be sought in the economic sphere per se, but in the realms of kinship, politics, and collective ritual.

As an important addendum, while Mauss used the term gift (*le don*) to characterize exchange systems dominant in early societies, this does not mean that exchange was always friendly or given with generous intentions, as our understanding of the word gift might imply. Gift exchange could often be agonistic and competitive. Mauss generalized from the example of the potlatch, practiced among American Northwest Coast Indians. Potlatch was a ritualized exchange where competing chiefs made public displays of exaggerated generosity, the aim of which was "to crush a rival chief with excessive obligations that cannot be repaid" (Yan, 2012: 280).

Finally, Mauss' essay initiated a long continuing inquiry into the nature of person–object relations. The opening

question of *The Gift* is, "What force is there in the thing given which compels the recipient to make a return?" (1967: 1). Mauss drew on the Maori concept of *hau*, a mystical force purportedly located in exchanged valuables that draw them back to their places of origin, not literally, but through the medium of a countergift. How people relate to objects, and through objects to each other, is key to Mauss' thought because it is not just the style of exchange that may differ typologically between market and nonmarket societies, but also whether or not the objects themselves are transferrable ('alienable') or not ('inalienable'), and what makes them so.

Despite, and in some ways because of, its terseness, *The Gift* has inspired more debate than any other text in the history of economic anthropology. In the 1960s and 1970s, when anthropologists joined in political economic critiques of globalizing capitalism, some found it useful to draw upon Mauss' work to oppose gift to commodity-oriented societies, and reciprocity to market exchange. In the 1980s and 1990s, edited volumes such as *The Social Life of Things* (Appadurai, 1986), *Money and the Morality of Exchange* (Parry and Bloch, 1989), and *Barter, Exchange, and Value* (Humphrey and Hugh-Jones, 1992), as well as many articles and monographs (e.g., Carrier, 1991; Gregory, 1982; Weiner, 1992), meticulously disentangled the issues surrounding the gift/commodity dichotomy. A survey of research being published today attests to the ongoing fruitfulness of the gift/commodity complex as a way of evaluating systems of exchange as well as individual rationales in both non-Western and Western contexts.

The Economic Constitution of Society, or the Social Constitution of the Economy? Reciprocity, Redistribution, and the Market

It is possible to read into Mauss's work the notion that in early society exchange functioned as a kind of social glue that held societies together. This functionalist notion has fallen from favor, but not anthropologists' tendency to view human beings as above all social rather than economic utility maximizing animals. In which case, the view of economic rationality as the prime determinant of individual behavior, and the market or market-like exchange as the foundation of social organization, might seem unnatural, even absurd.

How the 'economic man' view came to predominate, and the consequences of it, was the focal point of the work of a Hungarian economic historian, Karl Polanyi (1886–1964), whose ideas have had a lasting impact on economic anthropology.

Like Mauss, Polanyi looked to historical and ethnographic data to base a contrast between principles of a natural economy and those of market fundamentalism (epitomized in the belief in the 'self-regulating market'). Polanyi thought of the self-regulating market as a utopian idea that had raged out of control to become one of the great dehumanizing forces in history. Hann and Hart write: "Following Aristotle, Polanyi believed that society was a natural form designed to provide material sustenance for its members … The new apostles of political economy held that market economy could be self-regulating as long as everything was bought and sold without restriction. This led to what Polanyi called 'the fictitious commodities.' Nature, Society, and Humanity entered the market as land, money, and labor; and the traditions that ensured material provisioning for everyone were swept aside" (2009: 4).

To describe the historical transformation to market-directed society and the ideologization of the self-regulated market in the nineteenth century, Polanyi spoke of economic behavior as having become 'disembedded' from the social matrix, partly through the commodification of land and labor. (In this, Polanyi was following Marx's lead on what were the key instigators in the historical transition into capitalism.) Land and labor are closely linked to community and identity; their transformation in the market capitalist era into objects that could be bought and sold, or alienated, like any other commodity, represented a fundamental qualitative shift in human history. Through all prior history, Polanyi argued, "man's economy … [was] submerged in his social relationships" (1944: 46). Under market capitalism, by contrast, "instead of economy being embedded in social relations, social relations are embedded in the economic system" (1944: 57).

The terms embedded and disembedded were to assume far reaching importance, as they were taken up as antipoles not just in economic anthropology, but in economic sociology, political economy, and institutional economics. All these take it as axiomatic that economy is 'embedded' in society and culture, meaning that one cannot make sense of economic phenomena, even seemingly pure or abstract entities as a currency market, without reference to the social, cultural, institutional, political, geographical, and historical context in which one finds them.

Polanyi claimed that the tools, assumptions, and focus of contemporary economics were maladapted for studying preindustrial (and socialist-governed) societies, and perhaps vice versa. He differentiated two aspects of economic study in keeping with these separate tasks: The *formal* approach, suitable for studying decision-making and other means–ends phenomena such as one finds in the market, and the *substantive* approach, which emphasized the role of social norms and the orientation toward basic provisioning in a society (1957). These terms, which are cognates of disembedded/embedded, became terminological banners in a fiery debate between two camps – formalists and substantivists – that carried on through the 1960s.

The substantivists studied the ways in which economy was embedded in social relations and institutions. Their adversaries, the formalists, agreed with economists about the universal tendency toward utility maximization and therefore the determinative role of rationality in economic behavior. The role of the anthropologist, for the formalists, was to peel away the mask of culture and circumstance to show how ostensibly irrational decisions about what to maximize, who to trade with, how to set a price, how to decide what to plant, etc., are in the end rational – so long as one knows what to look for.

The formalist–substantivist debate died more from exhaustion, mismatched objectives, terminological imprecision, and finally the perception of irrelevance rather than through resolution. The opposition between two schools has been replaced by a sometime division of labor, sometime blending of the two outlooks. There is general acceptance that both elements – mutuality and market, in a contemporary iteration – are present in all societies (Gudeman, 2008).

Polanyi's analysis of premarket societies led to refinement of the concepts of reciprocity and redistribution. As Mauss had done when he posited that gift exchange characterized early economy, Polanyi sought to typologize economic institutions in preindustrial societies so as to contrast them with how market-dominated societies were organized. He settled on reciprocity and redistribution as being the fundamental, representative building blocks of primitive and archaic social organization, respectively. Reciprocity described mutual or symmetrical exchanges, such as between kinsmen, while redistribution took place hierarchically, such as when a chief collects wealth and then reallocates it in accordance with social divisions.

In *Stone Age Economics* (1972), Marshall Sahlins extended the analysis of reciprocity and redistribution. Aiming to show how exchange relations underwrite social relations, which is exactly Polanyi's point, Sahlins defined three kinds of reciprocity: (1) *Generalized*, corresponding to a western notion of a pure gift, in which "the social side of the relation overwhelms the material and in a way conceals it … the expectation of reciprocity is left indefinite" (1972: 206). (2) *Balanced*, in which an equivalent return is made right away, as in trade, and "the material side of the exchange is as important as the social" (1972: 206). (3) *Negative* reciprocity, which is the attempt to get something for nothing, like theft.

Sahlins proposed the following diagram (**Figure 1**) to show how kinship distance is associated with reciprocity, and he cites ethnographic instances to support his typology. The diagram is not intended to be static. One can trade with hostile neighbors, for instance, with the intention of creating a peace pact with them, thereby altering the social relation to a 'friendlier' sort (from negative to balanced reciprocity).

After *Stone Age Economics*, Sahlins distanced himself from nomothetic propositions of this sort, preferring to show how economies were rather structured by local symbolic or cosmological schemes (1976, 1988). Nevertheless, his theory of reciprocity inspired many economic anthropologists of the time and the model improved our understanding of exchange dynamics even in industrialized society. The concepts of reciprocity and redistribution continue to be refined and modified through contemporary ethnography (see, e.g., Davis, 1992; Hunt, 2002; Woodburn, 1998). With the recrudescence of market fundamentalism and the rise of neoliberalism since the 1980s (Harvey, 2005), Polanyi's ideas seem to be getting quite relevant again (Hann and Hart, 2009, 2011).

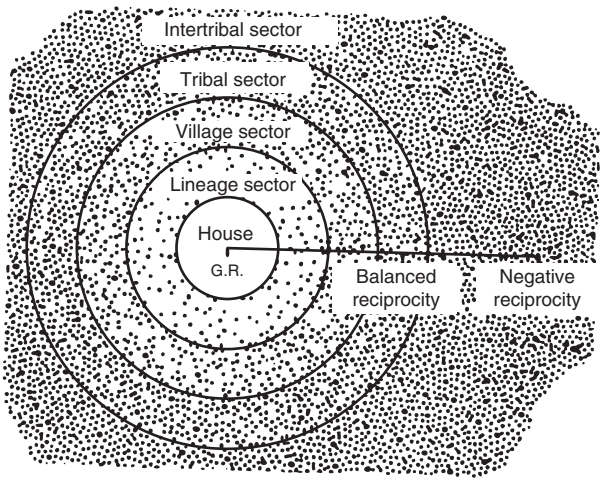

Figure 1 Reciprocity and kinship residential sectors. Reciprocity and Relatedness from Sahlins, M.D., 1972. Stone Age Economics. Aldine, Chicago, p. 199.

Modes of Production, Articulation, and History

Prior to Malinowski, anthropologists tended to equate economy with technology. They regarded the stage of technological development in how a given society procured the material means of its subsistence, or its mode of production, to be the defining feature of its economy. The mode of production would either be determinative of other kinds of social relations (division of labor, political organization, family structure, law, etc.) or at least it would correspond with those relations in a way that could be classified and compared.

Modes of production (MoP) flowed from simple to complex forms, implying historical or evolutionary stages. There were four principal divisions: hunter-gatherers, horticulturalists and pastoralists, peasants, and industrial societies. It was recognized even from the outset that this ordering was comprised of ideal types only, since empirically reported overlaps were often marked enough to place the classification at risk of murkiness. Particularly, the question of what is peasantry was debated at great length.

Nevertheless, there were certain features ascribed to societies thus classified such that the division appeared to remain useful for purposes of data collection and analysis at least through the 1960s. After that, the traditional typology became muted or transformed. With the perceptibly increasing speed with which capitalism and other exploitative forces were penetrating remote sectors of the globe, the traditional predilection for studying isolated unit societies – and the fiction that there indeed ever were such places – gave way to the pursuit of illuminating one-world interconnectivity, historically and contemporaneously.

Thus opened a new frontier of research in the 1970s, likewise centering on MoP but reorienting that term with a decidedly Marxist slant. Marx had long been recognized (if not always acknowledged) as a foundational influence in anthropology (Bloch, 1983). In American anthropology in particular, various streams of materialism resonated with Marxist premises (Sahlins, 1976). When on that side of the Atlantic world systems analysis and dependency theory – branches of political economy outlining the historical processes whereby poor countries became victims of rich ones – opened up, they became a ready magnet for economic anthropological investigation.

The new school, led by figures such as Eric Wolf and Sidney Mintz, likewise operated under the Marxist umbrella while partially sidelining Marx's teleological historical materialism. Wolf took the lead when he elaborated a new scheme of MoP, which, after Marx, he defined as "a specific, historically occurring set of social relations through which labor is deployed to wrest energy from nature by means of tools, skills, organization, and knowledge" (1982: 76).

For Wolf, MoPs were not neutrally occurring patterns that one could describe as if making sense of an organism or an engine (metaphors that had inspired structural functional analysis in anthropology) but they were suffused with power. There were exploiters ('surplus takers') and exploited (producers, laborers) in every economic system. A first purpose of MoP theory was to expose the systems of exploitation characterizing different types of society. Sometimes this exploitation took the form of gender inequalities or those attributable to age. Other times it was a ruling elite or a class of landholders. Not altogether divorcing himself from the older classifications outlined above, Wolf specified three MoPs – kinship, tributary, and capitalist – each of which, again, encompassed its own form of political organization, morality, and rituals that, this time, were seen in terms of their serving to justify existing power relations.

The second objective behind the MoP idea was to describe the process whereby the capitalist MoP either existed asymmetrically alongside other kinds of MoPs or brought about their conformity to the capitalist mode. The term used for this was 'articulation.' The intellectual center of gravity for this work was in France. Claude Meillassoux, Emmanuel Terray, Pierre-Philippe Rey, Georges Dupre, Maurice Godelier, and a handful of others, known as structural Marxists because they drew inspiration from Marx and Levi-Strauss, invented their own variant of MoPs to explain economic power relations. They asked, how might Marx's theories apply to contemporary noncapitalist societies? Their method was ethnographic rather than historical, a viewpoint which allowed them to better get at the synchronic fact that the capitalist MoP did not simply drive other types out but could coexist and 'articulate' with them. The work of this group was both empirically and theoretically inventive, and, despite the more than occasional abstruseness of the writing, achieved something of the status of an intellectual fad among anthropologists in the UK and America in the 1970s, before quickly fading (see Nugent, 2007).

The anthropological literature concerning production, labor, work, and related concepts, in relation to world capitalism or in more limited terms, began to flourish during the 1970s. Considerable attention was applied to what the relevant units of production might be – household, lineage, occupational group, factory, etc. – which led to detailed analysis of the economic and social organization within those units. The articulation phase of studies bequeathed economic anthropologists a greater latitude of thought concerning the possible coexistence or hybridity of different MoP – firms and families (or capitalist and kinship), for instance – and how individuals and groups may strategize to shift among these to protect both their livelihoods and their values.

Ethnographic studies of division of labor according to gender (e.g., Nash and Fernandez-Kelly, 1983) were among the most fruitful outcomes of the focus on production. Gender studies of labor have yielded many insights, from the imbricated constitution of domestic and economic spheres, to power relations in the control of the products of household labor, to the gender-mediated enabling mechanisms of post-Fordist flexible accumulation. While the stereotypic view is that production was a fascination in the 1970s that gave way to consumption, the ongoing efflorescence of ethnographies illuminating the micro characteristics and macro consequences of gendered labor, from villages to advanced industrial settings, emphatically belies this view.

Consumption, Agency, and Meaning in Modern Economy

The rise of interest in consumers and consumption took place in the fields of history, sociology, anthropology, and marketing studies at about the same time. In anthropology, there were three books in particular that may be cited as the birth moments: Mary Douglas and Baron Isherwood's *The World of Goods: Towards and Anthropology of Consumption* (1978), Pierre Bourdieu's *Distinction: A Social Critique of the Judgement of Taste* (1984), and Daniel Miller's *Material Culture and Mass Consumption* (1987).

In hindsight, the growth of the anthropology of consumption seems self-evidently tied to the triumph of neoliberalism and its ideological linking of consumption to economic prosperity starting in the 1980s. It can also be related to the disappearance of ethnographers' traditional field sites – the remote locales where manufactured consumer goods had had little sway. If the effects of wage work, cash crops, and other capitalist structures in African or Latin American villages had been visible for decades, it was not until the 1980s that anthropologists found in these same places the diffusion of consumer goods familiar to them from home. A new line of questions opened up: Can mass-produced consumer goods be vehicles for distinctive cultural expression, as locally produced objects are? What are the economic and social consequences of their importation to these locations? Is there a natural demand for such goods, or do native peoples adopt them as a consequence of some form of persuasion?

Up to that point, anthropologists had taken consumption either as a factor of production (i.e., as an input to the reproduction of producers, with a consequent emphasis on food) or distribution (i.e., that goods were apportioned on the basis of kinship, clan, etc., and this, rather than how individuals conceived of the goods, stamped their meaning). The state of affairs in industrialized society, moreover, was not of interest to most ethnographers, who infrequently conducted fieldwork in such settings.

The World of Goods extended symbolic anthropology to economic concerns within anthropology. Anthropologists had long interpreted local cosmologies through objects, known as material culture studies, but mass consumer goods as such had lain beyond the scope of these studies. Douglas and Isherwood argued that consumer goods could be understood as a semiotic code that reflected and stabilized social positions and cultural categories (see also Sahlins, 1976) and as such could be studied the same way everywhere.

Pierre Bourdieu (1984) operated from a comparable assumption, but added empirical and theoretical sophistication through a framework that showed how individual acts of classification through consumption, or the exercise of taste, is a form of social action that produces and reproduces objective social categories, most importantly class.

In reflection of these ideas and those deriving from his native discipline of archaeology, Daniel Miller (1987) devised a new synthesis that became the dominant approach to anthropological consumption studies. Miller

aimed to dispel what he regarded as a false prejudice promoted in the work of the Frankfurt School and French neoMarxists such as Jean Baudrillard. These theorists argued that just as capitalist production had in an earlier phase of capitalism been built upon alienating workers from their labor and from the tools of production, in the present era consumption had become the principal vehicle for alienating individuals from themselves and from culture. Miller emphasized "the potentially active role of consumers in resocializing commodities" (1995: 144). He contended that Marxist 'discourses about consumption' had to be distinguished from actual practices, which could only be studied ethnographically.

The proliferation of consumption studies, driven in part by the demographic expansion of the discipline during the 1980s and 1990s (just when the most consumerist generation in history was coming of age), has attracted formidable critiques. Many dispute the extent to which consumers are able to 'creatively recontextualize' mass commodities and make of them wholly distinctive individual and cultural identities, as Miller and his followers argue. They claim that consumption studied in this way overemphasizes individual agency in a world where the rise in consumption as a paradigm for prosperity has contributed to greater inequality and the attrition of time and monetary resources available to households for noncommercialized use (Carrier and Heyman, 1997). Cultural agency through the recontextualization of material objects may be exercised even by slaves; consumer agency requires purchasing power.

Others have demonstrated that the attempt to understand consumption independently of the powerful imperatives of capitalist production, historically and contemporarily, can only be partial (Mintz, 1985). Ethnographic focus on corporate marketing reveals the influence of consolidating industrial consortiums (such as in food, medicines, finance, energy, tourism, and media) that mobilize state-sized budgets and strategic planning apparatuses to secure hospitable cultural and infrastructural circumstances for the trade of their products globally. Their procedures are best seen not so much as efforts to commoditize (or brand) objects and convince people to buy them, as to move people themselves, with the help of distribution channel management and consumer education campaigns, into prefabricated commodity zones. As marketing has broadened since the 1960s, social sector assets have been vastly commercialized, communal and environmental protections diverted, and consumer consciousness, if hardly captured, has nevertheless been considerably swayed toward the internalization of commercially-mediated archetypes for materializing identity ('lifestyle') and achieving happiness (Applbaum, 2003, 2009).

Transitions to Modern Economy

Economic anthropology came of age just as the capitalist mode of production and world market was lapping its tongue into the last remote corners of the earth, where anthropologists had only recently arrived. The discipline has devoted much attention to how these transitions are experienced and accommodated by non-Western peoples.

In 2000, the French social theorist, sociologist, and anthropologist, Pierre Bourdieu, reflecting upon his fieldwork in Algeria in the 1960s (but surely with the present market fundamentalist turn in mind), wrote: "The principle of the overturning of the vision of the world in late colonial Algeria was nothing other than the acquisition of the spirit of calculation … To subject all the behaviours of existence to calculating reason, as demanded by the economy, is to break with the logic of philia, of which Aristotle spoke, that is, the logic of good faith, trust and equity which is supposed to govern relations between kin and which is founded on the repression, or rather the denegation, of calculation" (2000: 25).

If at the back of much economic anthropological theory lays a debate with economics, including over whether calculative rationality is a cultural–historical artifact or a universal caste of mind, the afflatus of the discipline's empirical work has often been to make sense of the effects on individuals, families, and communities of the broadening and deepening of capitalism. A dismal light often illuminates these studies because the observed reality from anthropologists' standpoint is that the impact of capitalist expansion is frequently deleterious to the lifeways and life chances of the economically humble.

Early studies of the transition to market society, so conceived, tended to be categorical and vectorial. Bohannan and Dalton (1962) described three types of local African economies: marketless, peripheral markets, and market-dominated. The economic characteristics of each were enumerated in terms of the principal source for subsistence livelihood, price formation for goods, money and money uses, external trade, and technology. The three types stood in a historical relationship with one another, with the implication of unidirectionality.

In an influential paper called 'The impact of money on an African subsistence economy,' Bohannan (1959) described the Tiv in Nigeria as having 'a multicentric economy' with mutually exclusive spheres of exchange. These separate spheres, which consisted of subsistence goods, prestige items, and rights in human beings, were "marked by different institutionalization and different moral values" (1959: 124) and were transacted with 'special purpose monies.' When the Tiv came to be forced to pay taxes in coinage, Bohannan said, they had to go about acquiring it through involvement in market transactions outside of their tribe. They started growing cash crops and thus became ensnared in the supply and demand system of distant trade. With this came the decline of special purpose money and its replacement with unitary currency, or 'general purpose money,' like dollars, which act as a unitary standard of value and can in theory exchange anything for anything. The breakdown of the multivariate economy, Bohannan lamented, spelt a decline in the traditional morality embedded in Tiv economy.

Today, Bohannan's depiction carries the air of a just-so story: factually dubious and theoretically simplistic. "It is not clear that money always flattens social relations, rather than creating new ones just as complex," Bill Maurer comments (2006: 21). And of course, market societies are themselves replete with 'special purpose' currencies such as store coupons, subway tokens, gambling chips, and even online virtual worlds money, not to mention myriad forms of nonmonetized exchange. (Spheres of exchange are still taught to

undergraduates because they broaden the minds of those unaccustomed to the thought that there were once societies where most things were not exchangeable for just anything else at a monetary price.)

Speaking just of money, early schemas like Bohannan's yielded to rigorous and sophisticated analyses of money's history, uses, and significance (Graeber, 2011; Guyer, 2004; Hart, 2001; Maurer, 2005; Parry and Bloch, 1989). Some of these still marginally concern themselves with the tensions associated with grand economic transitions (e.g., financialization, securitization, and depersonalization), while others resist or ignore this issue to focus on matters of purely local complexity. In either case, the accomplishment of the sub-subfield ('the anthropology of money') at least since the 1980s is hardly contingent on this narrative.

At present, the broadest umbrella under which economic transitions are studied in anthropology concerns neoliberal restructuring, referring to policies enacted upon the belief that the market is always superior to the state at managing and distributing public resources. Sometimes the procedure is called 'rolling back the state.' The trend in many places since the Reagan–Thatcher revolution has been toward the diminishment of the state's role in providing for health, utilities, education, environmental protection, and social insurances.

Case studies addressing themselves to the differing effects of and adaptations to neoliberal policies provide a rich corpus from which to compare outcomes and formulate generalizations. One key point where anthropological analysis of these trends is concerned is that inspection of them at the register of the local reveals aspects that are rarely discerned from afar. Few ethnographers neither end up endorsing unidirectional or unidimensional frameworks, nor do they invariably find that the decline of traditional economies always results in changes for the worse (Ensminger, 1992). Ground-level analyses show how macro-level terminologies (proletarianization, marketization, and trade concentration) are at best incomplete without reference to how these are experienced by individuals, families, communities, networks, institutions, and the culture-specific moralities that animate them.

History may prove that the long-term trend is as Bourdieu sketched for Algeria. There is benefit in any case to theorizing the processual mechanisms by which alterations in the fabric of economic consciousness conform to and actuate economic realities. Most anthropologists involve themselves with shorter range analyses, and at this level 'great transformation' schemes may be more diverting than helpful. At the axis point (the moment of articulation) of large systems, one finds economic actors operating "in the ambiguous boundaries between capital and labor, cooperation and exploitation, family and economy, tradition and modernity, friends and competitors …" (Smart and Smart, 2005). In many places, globalization has not cemented the role of formal capitalist institutions, but the reverse has occurred (Hart, 2012). A close look at entrepreneurs in many places reveals the contradictions of lives lived half in the local moral world of the village or tribal community and half in the aspirational world of the global market with its characteristic rationalities.

As ethnographers, anthropologists can observe the disruptions, ambiguities, and contradictions associated with economic change and stability (remembering that stasis is a social process). As global comparativists and long-range historians, anthropologists are well situated to track revolutionary changes as well as patterns of stable diversity in economic organization.

Underwriting and mediating the ruptures and continuities associated with economic change of whatever sort is the work of culture. Economic anthropologists have always minded the involvement of worldview, religion and cultural representation in their account of production, exchange, and so on. Starting particularly in the late 1970s, when symbolic and interpretivist anthropology crystallized into a permanent force in the discipline as a whole, a 'cultural economic' approach became more distinct (see, e.g., Geertz, 1979; Gudeman, 1986).

A signal culturalist position written in response to the world systems approach in anthropology (e.g., Wolf, 1982) is formulated by Sahlins in the following quote, which can suitably bookend that of Pierre Bourdieu at the opening of this section: "Western capitalism has loosed on the world enormous forces of production, coercion and destruction. Yet precisely because they cannot be resisted, the relations and goods of the larger system also take on meaningful places in local schemes of things. In the event, the historical changes in local society are also continuous with the superseded cultural scheme, even as the new state of affairs acquires a cultural coherence of a distinct kind. So we shall have to examine how indigenous peoples struggle to integrate their experience of the world system in something that is logically and ontologically more inclusive: their own system of the world" (1988: 10).

Contemporary Futures

The history of the development of research themes in economic anthropology has largely been one of accretion rather than replacement. Growth has been catalyzed in response to new (or newly perceived) developments in world economic circumstances, and in dialogue with neighboring disciplines.

In addition to the continuation of traditionally-themed studies, as described above, in the late 1980s economic anthropologists began to take seriously the challenge of interpreting economic affairs beyond the scope of the village or town. (Anthropologists focusing on development achieved this enlightenment sooner; see Edelman and Haugerud, 2005.) First, the study of the cross-border movement of various types of objects, from craft goods to coffee, was identified as a useful way to grapple with the complexity of trade in both historical and contemporary economy. Much was learned about both cultural brokerage (cultural economy) and power relations (political economy) in such studies. At the same time, ethnographies of assembly plants in places such as Mexico or Malaysia linked in the labor conditions and the circumstances for export of manufactured commodities across commodity chains owned and organized under capitalist enterprise.

For some, the fascination with commodities expanded to include goods such as textiles, cellphones, and pharmaceuticals, whose intellectual patents and process management reside in industrialized countries and are reckoned as exports when they are sold elsewhere. Anthropologists studying these transactions were, whether they were cognizant of it or not,

responding to corporations' own recognition that developing countries were no longer important just as a source for material inputs or cheap labor, but as markets for their products. Some anthropologists went further and studied intermediary institutions, such as bottling plants, shopping centers, or advertising agencies, in urban centers in Thailand, Srilanka, or Trinidad, in the attempt to get at questions of cultural translation and power at the sites where these are ostensibly brokered.

Since the 2000s, a handful of economic anthropologists have surmounted the challenge of access and have rendered fascinating and complex ethnographies of economic institutions at the geographical, institutional, and informational 'cores' of capitalism – New York, London, or Tokyo, for instance. Banks, stock markets, global consumer goods corporations, as well as development agencies and nongovernmental organizations, have each been the focus of recent works.

The aim to engage capitalism on its home turf (bearing in mind that capitalism, like democracy, is hardly an unalloyed system; anthropologists sometimes speak of capitalisms) corresponds with a belief among some that the discipline ought to be seeking greater relevance to current world events. The financial crisis of 2008 in particular sparked reflection on this matter. In the immediate aftermath of the events, Hart and Ortiz (2008: 2) wrote, "It seems that, along with the bankruptcy of some banks and even countries, the delegation of the power to channel credit to the private bureaucracies of contemporary finance has already lost its air of inevitability and indeed its former legitimacy … [W]e need to devise new global institutions and to think about them in fresh ways … Do anthropologists have something to say about all this?"

James Carrier, editor of the *Handbook of Economic Anthropology* (2012), reconvened the contributors to the recent first edition to update the volume to account for the recent economic turmoil, with a view that readers both in and outside the discipline might benefit from how economic anthropologists were making sense of what had happened. One of the contributors, the anthropologist and anarchist intellectual David Graeber, is credited as one of the inspirations behind the Occupy Wall Street Movement; his 2011 book, *Debt: The First 5000 Years*, is crafted as a conceptual and political antidote to current financial arrangements.

In 2012, President Barak Obama nominated Jim Yong Kim, an anthropologist, to the presidency of the World Bank. Kim is a medical rather than economic anthropologist, but his 2000 coedited book, *Dying for Growth: Global Inequality and the Health of the Poor*, is in tune with the conclusions of many studies by economic anthropologists involved in development. The relationship between economic anthropology and medical social science has been quietly blossoming as professionals in global health and related fields have come to recognize the ascendant role neoliberal market reforms have come to play in health care delivery and medical technology development and use.

The unofficial leader of the movement to catapult economic anthropology into public debate about the economy is Keith Hart. Citing the Nobel Prize winning economist Ronald Coase who said in 2012 that "the project of economics needs to be rescued from the economists," Hart and his associates have named their project at the University of Pretoria, the Human Economy Program.

In studying capitalism and its consequences, anthropologists have conventionally paid heed to inequality. The belated eruption of concern over degradation of the environment and the depletion of natural resources – resulting from our current economic system's desideratum to expand – may lead to a shift in the discipline's emphasis toward questions of sustainability (e.g., Wilk, 2010). This focus would inspire renewed associations with ecological/environmental anthropology. It may be that not just the future of the tiny subdiscipline of economic anthropology stands to benefit from these worldly engagements.

See also: Brokers and Brokerage, Anthropology of; Business Anthropology; Cargo Cults; Center–Periphery Relationships; Energy, Anthropology of; Exchange in Anthropology; Finance, Anthropology of; Globalization, Anthropology of; Industrial Societies; Inequality: Comparative Aspects; Markets: Anthropological Aspects; Modes of Production; Money: Anthropological Aspects; Political Economy in Anthropology; Postindustrial Societies; Poverty, Culture of; Property: Anthropological Aspects; Work, Anthropology of.

Bibliography

Appadurai, A. (Ed.), 1986. The Social Life of Things: Commodities in Cultural Perspective. Cambridge University Press, Cambridge.
Applbaum, K., 2003. The Marketing Era: From Professional Practice to Global Provisioning. Routledge, New York.
Applbaum, K., 2009. Getting to yes: corporate power and the creation of a psychopharmaceutical blockbuster. Culture, Medicine and Psychiatry 33, 185–215.
Becker, G.S., 1976. The Economic Approach to Human Behavior. University of Chicago Press, Chicago.
Bloch, M., 1983. Marxism and Anthropology: The History of a Relationship. Oxford University Press, Oxford.
Bohannan, P., 1959. The impact of money on an African subsistence economy. Journal of Economic History 19, 491–503.
Bohannan, P., Dalton, G. (Eds.), 1962. Markets in Africa. Northwestern University Press, Evanston, IL.
Bourdieu, P., 1984. Distinction: A Social Critique of the Judgement of Taste. Harvard University Press, Cambridge.
Bourdieu, P., 2000. Making the economic habitus: Algerian workers revisited. Ethnography 1, 17–41.
Carrier, J.G., 1991. Gifts, commodities, and social relations: a Maussian view of exchange. Sociological Forum 6, 119–136.
Carrier, J.G. (Ed.), 2012. A Handbook of Economic Anthropology, second ed. Edward Elgar, Cheltenham.
Carrier, J.G., Heyman, J.M., 1997. Consumption and political economy. Journal of the Royal Anthropological Institute 3, 355–373.
Chibnik, M., 2011. Anthropology, Economics, and Choice. University of Texas Press, Austin.
Davis, J., 1992. Exchange. Open University Press, Buckingham.
Douglas, M., Isherwood, B., 1978. The World of Goods. Penguin, Harmondsworth.
Edelman, M., Haugerud, A. (Eds.), 2005. The Anthropology of Development and Globalization: From Classical Political Economy to Contemporary Neoliberalism.
Ensminger, J., 1992. Making a Market. Cambridge University Press, Cambridge.
Geertz, C., 1979. Suq: the bazaar economy in Sefrou. In: Geertz, C., Geertz, H., Rosen, L. (Eds.), Order and Meaning in Moroccan Society: Three Essays in Cultural Analysis. Cambridge University Press, Cambridge, pp. 159–268.
Graeber, D., 2011. Debt: The First 5000 Years. Melville House, New York.
Gregory, C., 1982. Gifts and Commodities. Cambridge University Press, Cambridge.
Gudeman, S., 1986. Economics as Culture: Models and Metaphors of Livelihood. Routledge & Kegan Paul, London.

Gudeman, S., 2008. Economy's Tension: The Dialectics of Community and Market. Berghahn Press, New York.
Guyer, J., 2004. Marginal Gains: Monetary Transactions in Atlantic Africa. University of Chicago Press, Chicago.
Hann, C., Hart, K. (Eds.), 2009. Market and Society: The Great Transformation Today. Cambridge University Press, Cambridge.
Hann, C., Hart, K., 2011. Economic Anthropology: History, Ethnography, Critique. Polity, Cambridge.
Hart, K., 2001. Money in an Unequal World. Texere, New York.
Hart, K., 2012. Anthropology, economics, and development studies. In: Fardon, R., et al. (Eds.), Sage Handbook of Social Anthropology.
Hart, K., Ortiz, H., 2008. Anthropology in the financial crisis. Anthropology Today 24, 1–3.
Harvey, D., 2005. A Brief History of Neoliberalism. Oxford University Press, Oxford.
Humphrey, C., Hugh-Jones, S. (Eds.), 1992. Barter Exchange and Value: An Anthropological Approach. Cambridge University Press, Cambridge.
Hunt, R.C., 2002. Economic transfers and exchanges: concepts for describing allocations. In: Ensminger, J. (Ed.), Theory in Economic Anthropology. AltaMira Press, Walnut Creek, pp. 105–118.
Kim, J.Y., Millen, J.V., Irwin, A., Gershman, J. (Eds.), 2000. Dying for Growth: Global Inequality and the Helth of the Poor. Common Courage Press.
Malinowski, B., 1922. Argonauts of the Western Pacific. Routledge & Kegan Paul, New York.
Maurer, B., 2005. Mutual Life, Limited: Islamic Banking, Alternative Currencies, Lateral Reason. Princeton University Press, Princeton.
Maurer, B., 2006. The anthropology of money. Annual Review of Anthropology 35, 315–336.
Mauss, M., 1967[1925]. The Gift: Forms and Functions of Exchange in Archaic Societies. W.W. Norton, New York.
Miller, D., 1987. Material Culture and Mass Consumption. Blackwell, Oxford.
Miller, D., 1995. Consumption and commodities. Annual Review of Anthropology 24, 141–161.
Mintz, S., 1985. Sweetness and Power: The Place of Sugar in Modern History. Viking Penguin, New York.
Nash, J., Fernandez-Kelly, M.P. (Eds.), 1983. Women, Men and the International Division of Labor. SUNY Press, Albany.
Nugent, S., 2007. Some reflections on anthropological structural Marxism. Journal of the Royal Anthropological Institute (N.S.) 13, 419–431.
Parry, J., Bloch, M. (Eds.), 1989. Money and the Morality of Exchange. Cambridge University Press, Cambridge.
Plattner, S. (Ed.), 1989. Economic Anthropology. Stanford University Press, Stanford.
Polanyi, K., 1944. The Great Transformation. Beacon, Boston.
Polanyi, K., 1957. The economy as instituted process. In: Polanyi, K., Arensberg, C., Pearson, H. (Eds.), Trade and Market in the Early Empires. Free Press, Glencoe, IL.
Sahlins, M.D., 1972. Stone Age Economics. Aldine, Chicago.
Sahlins, M.D., 1976. Culture and Practical Reason. University of Chicago Press, Chicago.
Sahlins, M., 1988. Cosmologies of capitalism: the trans-Pacific sector of the world system. Proceedings of the British Academy 74, 1–51.
Smart, A., Smart, J. (Eds.), 2005. Petty Capitalists and Globalization: Flexibility, Entrepreneurship, and Economic Development. SUNY Press, Albany.
Strathern, A., Stewart, P.J., 2012. Ceremonial exchange: debates and comparisons. In: Carrier, J.G. (Ed.), A Handbook of Economic Anthropology, second ed. Edward Elgar, Cheltenham, pp. 239–256.
Weiner, A., 1992. Inalienable Possessions: The Paradox of Keeping-While-Giving. University of California Press, Berkeley.
Wilk, R., 2010. Consumption embedded in culture and language: implications for finding sustainability. Sustainability: Science, Practice, & Policy 6 (2), 38–48.
Wolf, E., 1982. Europe and the People Without History. University of California Press, Berkeley.
Woodburn, J., 1998. 'Sharing is not a form of exchange': an analysis of property-sharing in immediate-return hunter-gatherer societies. In: Hann, C.M. (Ed.), Property Relations: Renewing the Anthropological Tradition. Cambridge University Press, Cambridge, pp. 48–63.
Yan, Y., 2012. The gift and gift economy. In: Carrier, J.G. (Ed.), A Handbook of Economic Anthropology, second ed. Edward Elgar, Cheltenham, pp. 275–290.

Economic Education

Carlos J Asarta, University of Delaware, Newark, DE, USA
Roger B Butters, Hillsdale College, Hillsdale, MI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by W.E. Becker, volume 6, pp. 4078–4084, © 2001, Elsevier Ltd.

Abstract

Economic education focuses on identifying and developing teaching and research techniques that make economics more accessible, relevant, and applicable to both students and educators. The field offers a strong and diverse group of organizations that encourage and promote economic education. This article offers an overview of those organizations, as well as of the content that is typically covered in economics courses, the teaching methods used to teach that content, and the assessment of students and instructors.

Introduction

As a discipline, economic education focuses on identifying and developing teaching and research techniques intended to make economics more accessible, relevant, and applicable to both students and educators. Current efforts in economic education concentrate on continuous improvements related to the following topics: organizations, content, teaching methods, and assessment. Comprehensive reviews on teaching and learning economics, as well as on economic education research, can be found in Hoyt and McGoldrick (2012). The *Journal of Economic Education* is the leading academic journal in the field.

Organizations

Since the beginning of the twenty-first century, the number of local, national, and international organizations devoted to economic education has increased dramatically and the focus of those organizations has shifted to emphasize financial literacy in addition to economic education. Long-established organizations, such as the Council for Economic Education (formally, National Council on Economic Education) and its Council/Center network, continue to deliver programs and promote economic education at the local, state, and national levels. Other organizations, such as the Federal Reserve, have expanded their public relations mission to include economic education for both adult and K-12 audiences. Indeed, the primary focus of several Federal Reserve branch banks has become economic literacy, and the Federal Reserve makes extensive resources available online through www.federalreserveeducation.org. These initiatives are the result of direction from the Chairman of the Federal Reserve, and local needs (Bernanke, 2013). Likewise, the Jump$tart Coalition and other organizations have expanded their programming to include teacher training and other educational activities (Jump$tart, 2013).

Since 2007, and the economic turmoil that resulted from upheaval in financial markets, there have been additional calls for increased economic literacy generally, and an increased focus on financial literacy and wealth management, specifically. At the federal level, the passage of the Dodd–Frank Wall Street Reform and Consumer Protection Act resulted in the creation of the Consumer Financial Protection Bureau, which has economic and financial education as one of its core functions (Dodd-Frank, 2010). What form that function will take and how it will affect economic education has yet to be determined (CFPB, 2013). Given its broad mandate and budget, however, that impact could be significant.

Organizations focused on economic literacy have moved beyond policy statements and have made considerable progress in both implementing economics into the precollege curricula and developing mandates for economic education. These mandates vary by state and range from simple inclusion in the K-12 curriculum, to required coursework, to summative testing prior to graduation. The greatest strides in curriculum implementation occurred between 1998 and 2000 with the number of states requiring economics in the K-12 curriculum rising from 38 to 48; as of 2013, all 50 states require economic content. From 2000 to 2011, the number of states requiring K-12 students take an economics course rose from 13 to 22. Additionally, the number of states requiring testing in economics has fluctuated over the years, with 22 of them requiring testing in 2000 and 2011 (CEE, 2011).

The question raised by Siegfried et al. (1991) as to whether economic education is achieving the goal of having students 'think like economists' is still very much unanswered. Despite the dramatic increase in resources and attention devoted to economic education, the gains in economic understanding are illusive and difficult to measure. Recent research documents only marginal gains in economic understanding in voluntary student samples between 2000 and 2009. Furthermore, the observed gains may be attributable to differences in incentives and testing format (Butters and Asarta, 2011). Whether this is due to increasing enrollments resulting in less-proficient students taking economics courses, deficiencies in the ability and human capital of teachers, or some other factor is unknown. What is clear is that the mission of the various organizations promoting economic education is not complete, and their work remains unfinished.

Content

The study of economics examines how individuals, organizations, and societies make choices under conditions of scarcity. Thus, a primary goal of economic education is to equip

students with the knowledge and tools they need to make the best choices possible, given the resources available. Unfortunately, recent studies have shown a widespread deficient level of economic knowledge among students and adults (i.e., Markow and Bagnaschi, 2005; Czelusniak, 2009; Butters and Asarta, 2011). These findings, along with recent developments in the world economy, have sparked a renewed emphasis for identifying the economic content and principles that the public should know to be successful consumers, workers, employers, and citizens.

For purposes of developing materials and improving pedagogy, formal instruction in economic education can be separated into six major categories: fundamentals, microeconomics, macroeconomics, international, measurement methods, and social goals. Measurement methods and social goals are often found in the first four categories (Buckles, 1991: pp. 26–27).

Narrower lists of content categories have been used in developing nationally normed, standardized, and reliable instruments to measure economic literacy and knowledge (Asarta and Rebeck, 2012). For example, the *Test of Economic Literacy*, currently in its fourth edition (Walstad et al., 2013), is divided into four content categories: fundamentals, microeconomics, macroeconomics, and international economics. Additionally, the *Test of Understanding College Economics (TUCE)*, also in its fourth edition (Walstad et al., 2007), is further divided into two broad categories: microeconomics and macroeconomics. For the TUCE, fundamentals are included as part of the microeconomics content, while international economics is infused in both macroeconomics and microeconomics.

Valid and reliable tests of economic knowledge and literacy can serve as a reference for the content that is covered in macroeconomics and microeconomics. Historically, macroeconomic content has focused on (1) national income accounting, (2) real versus nominal values, (3) investment and saving, (4) money demand (liquidity) and money supply, (5) aggregate expenditure (demand) and supply, (6) stagnation and growth, and (7) monetary and fiscal policy. For microeconomics, the content has focused on (1) scarcity and opportunity costs, (2) demand and supply, (3) market failures and externalities, (4) revenue, cost, and profit maximization, (5) market structures, and (6) equity and efficiency (Becker, 2001).

Economists have argued that the best way to prepare students for their futures is to focus on a reduced list of core topics, instead of attempting to cover every topic found in principles textbooks. In fact, McConnell (1998: p. 41) suggests that in order to avoid barriers to learning, which may result from cognitive dissonance, instructors should be selective and introduce economics concepts using broad topics that are accessible to the student. In microeconomics, Frank (2012: p. 405) provides the following list of essential principles: (1) scarcity, (2) cost–benefit, (3) comparative advantage, (4) increasing opportunity cost, (5) equilibrium, and (6) efficiency. Frank notes that, while reduced, this list of essential principles is at the root of all other microeconomic concepts. Additionally, Gilleski and Salemi (2012) propose the use of literacy-targeted courses focused on core topics that students will be able to use for the rest of their lives. Literacy-targeted courses differ from traditional courses in that they cover fewer topics more intensively, and provide opportunities, through course assignments, to gain a better understanding of basic economics. In microeconomics, these topics include (1) scarcity, choice, and opportunity cost, (2) comparative advantage, (3) demand and supply, (4) elasticity of demand, (5) price taking and economic efficiency, (6) economic profit and rent, (7) market power and price setting, and (8) externalities and property rights.

For macroeconomics, several economists have suggested additional concepts that are rarely covered in textbooks or standardized tests, but that are critical for citizens since they are found in the daily news or faced in everyday life situations. For example, Kennedy (1992) supplied a list of five key relationships consistently faced by individuals: (1) the discouraged or encouraged worker phenomenon and reported unemployment changes, (2) forecasting of inventories and production, (3) nominal versus real interest rates, (4) inflation rate differences and changes in exchange rates, and (5) the tendency of real interest rates to be equal across countries. Furthermore, McEachern (2012) recommends an even broader selection of topics to underscore in macroeconomics, including automatic stabilizers, long-term joblessness, technological change and living standards, the institutional setting, and the dynamic nature of the economy.

While many economists and organizations have attempted to create lists providing the key economic content that students and adults should know, to date, the economics profession has not been able to agree upon a universal list of skills, abilities, and knowledge needed to become economically literate.

Teaching Methods

The greatest innovations in teaching methods in the past decade have come from the ubiquitous adoption of the Internet and the arrival of students who are 'digital natives,' or students for whom computers and the Internet have been a fact of life instead of a recent innovation. For these students, traditional lecture formats are less satisfying, and do not provide the interactivity and dynamic environment they have come to expect. Thus, instructors have had significant incentives to adopt digital content and delivery (Levin and Arafeh, 2002). As a result, many of the innovative teaching methods first identified in the late 1990s are now being implemented using technology. For example, the prevalence of portable devices that are Internet-enabled makes teaching strategies that engage students immediately prior to class, or 'just-in-time teaching,' low cost and effective (Simkins and Maier, 2004).

Not only has the Internet changed the nature of instruction, it has changed how and when instructors teach. Today, the economics classroom can be a traditional lecture-based classroom, a fully online classroom, a hybrid of the two, or a 'flipped' or 'inverted' classroom. Research into the efficacy of online versus face-to-face instruction, while mixed, has largely concluded that students in a traditional classroom score significantly higher on assessments of economic knowledge than do their peers in pure online courses. This difference becomes insignificant for students in hybrid courses, or courses with some online instruction mixed with traditional teacher interaction (Brown and Liedholm, 2002; Terry et al., 2003;

Coates et al., 2004; Anstine and Skidmore, 2005). Possible reasons for the difference in outcomes for online versus traditional courses include the reduction in student–instructor interaction, limited possibilities for peer instruction, the high fixed costs associated with developing quality online content, and the quality of course management software (CMS) (Goffe and Sosin, 2005). As educators are able to develop better online content, and as CMS systems improve, we would expect the observed deficit in online education to be reduced or eliminated.

An increasingly popular form of course organization, the 'flipped' or 'inverted' classroom, leverages various teaching strategies to fundamentally change instructional flow. In an 'inverted' classroom, the first exposure to new content occurs prior to class via videos or readings available online. Online tests or in-class quizzes are used to incentivize students to complete the 'first exposure' materials. Class lectures and activities are then designed to extend understanding to a deeper level. Finally, homework is used to follow up on and measure learning outcomes.

First identified in a paper published in the *Journal of Economic Education*, the 'inverted' classroom model allows instructors to use class time to engage students in discussion, problem solving, and other active learning strategies that may address some of the long-standing questions in economic education (Lage et al., 2000). For example, a large body of research from the science, technology, engineering, and mathematics (STEM) disciplines suggests that active learning strategies such as 'peer instruction' or 'interactive learning' dramatically increase students' understanding of the material taught. Furthermore, not only do test scores for both male and female students rise when using interactive learning, but also the gap between male and female test scores is eliminated. Additionally, students who are taught interactively are less likely to change majors to a non-STEM field and retain learned information longer (Crouch et al., 2007; Lambert, 2012). The persistence of a gender gap in student test scores, the difficulty of attracting women to the economics major, and the rapid depreciation of economic knowledge are three issues that have been persistent, and troublesome features of economic education (Becker, 2001). The success of active learning in STEM disciplines suggests that there may be a single solution to all the three problems in economics.

Beyond the organization of the classroom, the reduction in the cost of online delivery and the availability of digital content have made it possible for educators to address, in unique and innovative ways, Siegfried and Fels' (1979) comment that "Different students learn economics in different ways. The best teaching strategy provides alternative learning methods." For example, online resources such as Coursera (www.coursera.org) and the Kahn Academy (www.khanacademy.org) provide low-cost instructional content to instructors and students alike. Within economics, instructors have created online repositories of movie scenes and art that teach economic content (Mateer and Li, 2008; Watts and Chineze, 2012). As the availability of such resources grows and their quality improves, the fixed costs of offering an online or hybrid course that can compete with the traditional face-to-face course will fall dramatically, further improving the online educational experience of students.

In response to changes in the classroom and the need for improved teaching and research in economic education, the American Economic Association (AEA) Committee on Economic Education began hosting an annual conference that focuses on teaching economics (at the undergraduate and graduate levels) and research in economic education (including precollege levels). The inaugural conference, held at Stanford University in June 2011, hosted educators from around the world and featured keynote speakers such as Vernon Smith, B. Douglas Bernheim, and John Taylor, all of whom spoke on issues related to economic education and teaching. Research presented at the conference addressed many of the issues associated with integrating technology and the Internet into flexible classroom environments. Cosponsored by the *Journal of Economic Education*, subsequent annual conferences have further raised the profile of economic education and best practices for teaching by publishing the best papers presented at the conference in the leading journal in the field.

In sum, while the concepts taught in economics may not be dramatically changing, the nature of economic education and teaching methods is. Programs such as the AEA Conference on Teaching and Research in Economic Education are vital to developing and sharing new teaching methods that leverage the growing body of educational research to ensure that economic education stays relevant and meaningful to students. Additional training opportunities, such as the teaching innovation programs (TIP) series, may be needed on a national level to incentivize and assist educators to make the necessary changes to their teaching style (Walstad and Salemi, 2011). Instructors may guide and inspire students, make the material applicable and interesting with theoretical tools and examples, and provoke thoughts to stimulate interest that encourages students to appreciate and take to heart the study of economics.

Assessment

In general, assessment is the process of gathering data to assign grades, evaluate the effectiveness of programs, and make changes to increase student achievement. Summative assessment focuses on the achievement of students, while formative assessment provides instructors with feedback that can be used to improve what is being assessed (Miller et al., 2009). Suskie (2009: p. 4) provides four steps needed to develop an effective assessment program. First, clear, measurable expected outcomes of student learning have to be established. Second, students need to have ample opportunities to achieve those outcomes. Third, instructors should determine if their students are achieving expected outcomes by systematically collecting, analyzing, and interpreting data. And finally, practitioners need to use that data to understand and improve student learning.

In economics, the two most commonly used assessment methods of student learning involve the use of free-response (constructed-response) and multiple-choice questions. Both methods have received considerable attention in the economic education literature (Walstad and Becker, 1994; Chan and Kennedy, 2002; Buckles and Siegfried, 2006). Each has its advantages and disadvantages. For example, multiple-choice exams require considerably less time to grade, can cover a larger set of concepts, and are more objective in grading than

free-response exams. On the other hand, multiple-choice exams allow for guessing and may not reach the highest levels of Bloom's (1956) taxonomy. Walstad (2001) offers a useful framework for deciding which method instructors may want to use based on their goals and constraints for a course. Additionally, a valuable discussion on how to create and grade test items can be found in Walstad (1998), for multiple-choice items and Welsh and Saunders (1998), for free-response items.

A test instrument needs to be reliable and valid in order to provide useful information for assessment purposes. Reliability is the ability of test scores to consistently measure student performance. One commonly used measure of the internal consistency of a test instrument used in standardized tests of economic knowledge and literacy is the alpha reliability coefficient. Validity refers to the ability of a test to measure what it is intended to measure. This characteristic can be further divided into content validity and construct validity. Content validity is achieved by making sure that a test instrument covers the appropriate content. This goal could be accomplished by utilizing work references indicating the content to be taught at different educational levels, such as the *Voluntary National Content Standards in Economics* (CEE, 2010), and by consulting with experts in the field. Construct validity refers to the ability of a test to discriminate between those students who know the content and those who do not (Rebeck and Asarta, 2012: pp. 318–319).

Additional sources of information for determining whether students understand economic content include homework and problem sets, classroom experiments, group work, and immediate feedback techniques. In fact, recent technological advances have provided educators with opportunities to receive immediate classroom feedback instead of having to wait until the end of the semester or after a test. Students can now use remote handheld devices, such as 'clickers' and smart phones, to answer questions in class. These systems not only allow for a greater level of student involvement in the learning process, but they also provide instructors with immediate summative feedback that can be used to undertake formative assessment. Additionally, another source of formative assessment that has been proven to be effective in improving student learning is 1-min papers (Chizmar and Ostrosky, 1998). One-minute papers, and their many variations, are used at the end of class periods, or via e-mail, and include a number of questions asking students to share what they learned during a lecture and what they still do not understand. This method, while time consuming given that instructors have to read each student's feedback, allows instructors to assess and react to the learning that is occurring in the classroom.

While the assessment of student performance receives a considerable amount of attention in the economic education field, often, it is the instructors who are evaluated by students on their performance in the classroom. This is particularly true at the tertiary level, where students have a chance, at the end of a session, to share their views and opinions with school administrators by completing student evaluations of teaching (SET). SETs typically focus on the overall effectiveness, communication skills, and mastery of the subject matter of an instructor, as well as on the organization and planning of a course, but do not tend to provide formative feedback to instructors. And although the weights assigned to SETs vary depending on the type of academic institution, administrators consistently use this information to evaluate faculty members annually, determine salary changes, and for promotion and tenure decisions (Becker and Watts, 1999).

Unfortunately, research findings using SETs have been controversial, and often questioned due to statistical shortcomings (McPherson, 2006). Additionally, the recent availability of alternative methods of instruction, such as hybrid or 'flipped' courses discussed above, has been found to influence SET scores and should be taken into consideration when evaluating instructors. Asarta and Schmidt (2010) examined the SET scores of college students participating in an introductory economic and business statistics course offered using a traditional and hybrid formats, and found that SET scores were consistently and significantly marked down in the hybrid format. Since administrators typically use SET scores to compare individual instructors to an average, or for ranking purposes within a department, not accounting for these types of variables could be very costly for instructors.

Conclusion

The field of economic education has continued to evolve over the past decade due to recent developments in the world economy, technological changes, and a desire to improve the level of economic knowledge and understanding of individuals. The field offers a strong and diverse group of organizations that encourage and promote economic education, focusing on the content that should be taught, the teaching methods that should be used, and the assessment of students and instructors.

See also: Alternative Schools of Economic Thought; Economics, History of; Financial Crises, History of; Macroeconomics, History of From 1933 to Present.

Bibliography

Anstine, J., Skidmore, M., 2005. A small sample study of traditional and online courses with sample selection adjustment. Journal of Economic Education 36 (2), 107–127.

Asarta, C., Schmidt, J.R., 2010. Differences in student evaluations from hybrid and traditional courses. Journal of Economics and Finance Education 9 (2), 19–28.

Asarta, C., Rebeck, K., 2012. Measurement techniques of student performance and literacy: college and high school. In: Hoyt, G., McGoldrick, K. (Eds.), International Handbook on Teaching and Learning Economics. Edward Elgar, Cheltenham, UK, pp. 307–319.

Becker, W.E., 2001. Economic education. In: Neil, J.S., Baltes, P.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Pergamon, Oxford, pp. 4078–4084.

Becker, W., Watts, M., 1999. How departments of economics evaluate teaching. American Economic Review 89 (2), 344–349.

Bernanke, B.S., 2013. Financial and Economic Education. Speech Given at the 13th Annual Redefining Investment Strategy Education (RISE) Forum, Dayton, Ohio (via prerecorded video). http://www.federalreserve.gov/newsevents/speech/20130404a.htm (text accessed 05.06.13.).

Bloom, B.S. (Ed.), 1956. Taxonomy of Educational Objectives. David McKay, New York.

Brown, B.W., Liedholm, C.E., 2002. Can web courses replace the classroom in principles of microeconomics. The American Economic Review 92 (2), 444–448.

Buckles, S., 1991. Guidelines for economic content in school programs. In: Walstad, W.B., Soper, J.C. (Eds.), Effective Economic Education in the Schools. The Joint Council on Economic Education and The National Education Association, New York, pp. 24–34.

Buckles, S., Siegfried, J.J., 2006. Using multiple-choice questions to evaluate in-depth learning of economics. Journal of Economic Education 37 (1), 48–57.

Butters, R.B., Asarta, C.J., 2011. A survey of economic understanding in U.S. high schools. Journal of Economic Education 42 (2), 200–205.

Chan, N., Kennedy, P.E., 2002. Are multiple-choice exams easier for economics students? A comparison of multiple-choice and "equivalent" constructed-response exam questions. Southern Economic Journal 68 (4), 957–971.

Chizmar, J.F., Ostrosky, A.L., 1998. The one-minute paper: some empirical findings. Journal of Economic Education 29 (1), 3–10.

Coates, D., Humphreys, B.R., Kanes, J., Vachris, M.A., 2004. "No significant distance" between face-to-face and online instruction: evidence from principles of economics. Economic of Education Review 23 (5), 533–546.

Council for Economic Education (CEE), 2010. Voluntary National Content Standards in Economics, second ed. Council for Economic Education, New York.

Council for Economic Education (CEE), 2011. Survey of the States. Economic and Personal Finance Education in Our Nation's Schools. New York.

Crouch, C.H., Watkins, J., Fagen, A.P., Mazur, E., 2007. Peer instruction: engaging students one-on-one, all at once. In: Redish, E.F., Cooney, P. (Eds.), Reviews in Physics Education Research. American Association of Physics Teachers, College Park, MD.

Czelusniak, S., 2009. The crisis of economic illiteracy. The Journal of the James Madison Institute. http://www.jamesmadison.org/pdf/materials/Journal_SpringSummer2009_CZELUSNIAC.pdf (accessed 15.05.13.).

Dodd-Frank Wall Street Reform and Consumer Protection Act (Dodd-Frank) 2010. Pub. L. No. 111-203, 124 Stat. 1376.

Frank, R.H., 2012. Less is more: the perils of trying to cover too much in microeconomic principles. In: Hoyt, G., McGoldrick, K. (Eds.), International Handbook on Teaching and Learning Economics. Edward Elgar, Cheltenham, UK, pp. 403–412.

Gilleski, D.B., Salemi, M., 2012. The cost of economic literacy: how well does a literacy-targeted principles of economics course prepare students for intermediate theory courses? Journal of Economic Education 43 (2), 111–132.

Goffe, W.J., Sosin, K., 2005. Teaching with technology: may you live in interesting times. Journal of Economic Education 36 (3), 278–291.

Hoyt, G., McGoldrick, K. (Eds.), 2012. International Handbook on Teaching and Learning Economics. Edward Elgar, Cheltenham, UK.

Jump$tart Coalition for Personal Financial Literacy, 2013. The Jump$tart Teacher Training Alliance. http://www.jumpstart.org/teacher-training-alliance.html (accessed 05.06.13.).

Kennedy, P., 1992. Teaching tools: on journalists' use of macroeconomic concepts. Economic Inquiry 30 (1), 194–201.

Lage, M.J., Platt, G.J., Treglia, M., 2000. Inverting the classroom: a gateway to creating an inclusive learning environment. Journal of Economic Education 31 (1), 30–43.

Lambert, C., 2012. Twilight of the lecture. Harvard Magazine. March–April. http://harvardmagazine.com/2012/03/twilight-of-the-lecture (accessed 01.06.13.).

Levin, D., Arafeh, S., 2002. The Digital Disconnect: The Widening Gap between Internet-Savvy Students and Their Schools. Pew Internet & American Life Project, 14 August 2002. http://www.pewinternet.org/Reports/2002/The-Digital-Disconnect-The-widening-gap-between-Internetsavvy-students-and-their-schools.aspx (accessed 05.06.13.).

Markow, D., Bagnaschi, K., 2005. What American Teens & Adults Know about Economics. Prepared by Harris Interactive, Inc., for the Council on Economic Education. http://www.councilforeconed.org/cel/WhatAmericansKnowAboutEconomics_042605-3.pdf (accessed 18.06.13.).

Mateer, G.D., Li, H., 2008. Movie scenes for economics. Journal of Economic Education 39 (3), 303.

McConnell, C.R., 1998. Reflections on the principles course. In: Walstad, W.B., Saunders, P. (Eds.), Teaching Undergraduate Economics: A Handbook for Instructors. Irwin/McGraw-Hill, New York, pp. 31–42.

McEachern, W.A., 2012. Macroeconomic principles are still relevant and still important. In: Hoyt, G., McGoldrick, K. (Eds.), International Handbook on Teaching and Learning Economics. Edward Elgar, Cheltenham, UK, pp. 413–422.

McPherson, M., 2006. Determinants of how students evaluate teachers. Journal of Economic Education 37 (1), 3–21.

Miller, M.D., Linn, R.L., Gronlund, N.E., 2009. Measurement and Assessment in Teaching, tenth ed. Pearson Education, Inc., Upper Saddle River, NJ.

Rebeck, K., Asarta, C.J., 2012. Methods of assessment in the college economics course. In: Hoyt, G., McGoldrick, K. (Eds.), International Handbook on Teaching and Learning Economics. Edward Elgar, Cheltenham, UK, pp. 177–187.

Siegfried, J., Fels, R., 1979. Research on teaching college economics: a survey. Journal or Economic Literature 17, 923–969.

Siegfried, J.J., Bartlett, R.L., Hansen, W.L., 1991. The status and prospects of the economics major. Journal of Economic Education 22 (3), 197–224.

Simkins, S., Maier, M., 2004. Using just-in-time teaching techniques in the principles of economics course. Social Science Computer Review 22 (4), 444–456.

Suskie, L.A., 2009. Assessing Student Learning: A Common Sense Guide, second ed. Jossey-Bass, San Francisco.

Terry, N., Lewer, J.J., Macy, A., 2003. The efficacy of alternative instruction modes in economics. Journal of Economics and Economic Education Review 4 (1), 23–34.

Walstad, W.B., 1998. Multiple choice tests for the economics course. In: Walstad, W.B., Saunders, P. (Eds.), Teaching Undergraduate Economics: A Handbook for Instructors. Irwin/McGraw-Hill, New York, pp. 287–304.

Walstad, W.B., 2001. Improving assessment in university economics. Journal of Economic Education 32 (3), 281–294.

Walstad, W.B., Becker, W.E., 1994. Achievement differences on multiple-choice and essay tests in economics. American Economic Review 84 (2), 193–196.

Walstad, W.B., Watts, M., Rebeck, K., 2007. Test of Understanding of College Economics (Fourth Edition): Examiner's Manual. National Council on Economic Education, New York.

Walstad, W.B., Salemi, M.K., 2011. Results from a faculty development program in teaching economics. Journal of Economic Education 42 (3), 283–293.

Walstad, W.B., Rebeck, K., Butters, R.B., 2013. Test of Economic Literacy (Fourth Edition): Examiner's Manual. Council on Economic Education, New York.

Watts, M., Chineze, C., 2012. Using art (paintings, drawings, and engravings) to teach economics. Journal of Economic Education 43 (4), 408–422.

Welsh, A.L., Saunders, P., 1998. Essay questions and tests. In: Walstad, W.B., Saunders, P. (Eds.), Teaching Undergraduate Economics: A Handbook for Instructors. Irwin/McGraw-Hill, New York, pp. 305–318.

Relevant Websites

www.councilforeconed.org – Council for Economic Education.
www.federalreserveeducation.org – Federal Reserve Education.
www.fte.org – Foundation for Teaching Economics.
nces.ed.gov/nationsreportcard/economics – National Assessment of Education Progress, Economics.

Economic Evaluations and Electoral Consequences

Michael S Lewis-Beck, University of Iowa, Iowa City, IA, USA
Mary Stegmaier, Truman School of Public Affairs, University of Missouri, Columbia, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Voters evaluate economic conditions and use these assessments to hold the incumbent government accountable at election time. Voters typically assign credit or blame for the economic situation to the government leader, the president or prime minister, and his or her political party. This assignment of responsibility is easy for voters when the locus of responsibility is clear, but it is made more difficult when institutional or government arrangements diffuse power. In general, evidence of economic voting has been found in democracies around the world.

Introduction

"Who will you vote for?" This question is asked in friendly conversations before elections in democracies around the world. We know much about how voters make their decisions: party identification, ideology, policy issues, candidate characteristics, and evaluations of the incumbent's performance all feature prominently in the voter's calculus. Among the conditions a voter might evaluate that economic evaluations are paramount. In this article, we examine the evaluations of the economy that voters make, and how those evaluations affect their electoral choices. According to classical economic voting theory, when the economy prospers voters tend to reward the government with their votes, but if the economy falters, they tend to punish the government by giving their votes to the opposition. We pursue below this reward–punishment theory of economic voting, as it has been dubbed, along with its variants.

Economic evaluations and voter preferences will be measured via national election surveys of public opinion. We will assess where voters look for the economy – at the past or the future, at their own pocketbooks or the entire nation. An important question concerns how accurately voters see the economy. To the extent they misperceive it, is that distortion due to lack of information or to political prejudice? What do they do politically with their economic evaluations? The answer to that question depends on who they hold responsible for running the economy. The clearer the government is in charge, the more likely voters will reward or punish it.

Many things can blur the link between the economy and the government, such as self-blame for economic misfortune or blaming the boss rather than government policies. Besides attribution of blame, institutions play a role. On the one hand, the impact of economic perceptions can be diminished by the structure of institutions, as when many parties are in the ruling coalition, or when the center of decision making moves away from the national capital to an international one such as Brussels. On the other hand, the economic crisis manifest worldwide today may actually increase the strength of economic voting. We deal with such issues below, focusing on the economy as a valence issue, as an issue dealing with the national good. We aim to show how important these economic evaluations are, aiding the proper functioning of a democracy by holding elected officials accountable.

Economic Voting Theory

In the study of political behavior, many of the research guidelines came out of that classic, *The American Voter* (Campbell et al., 1960). The field of economic voting, as it has come to be called, is no exception. In that pioneering investigation of voting, as measured in national public opinion surveys, they ask the question, "Is a person's economic outlook associated with his (sic) partisan choice?" concluding the answer "is substantially affirmative" (Campbell et al., 1960: 397–399). They go on to assert that incumbent President Eisenhower did draw "a disproportionately large part of his popular vote margin from the economic optimists in 1956" (Campbell et al., 1960: 385, 400–401). Such a finding underlined Key's (1966) more general argument that voters evaluate White House performance, rewarding or punishing the president's administration accordingly at the ballot box.

Broad scientific evidence for this proposition first appeared in Kramer's (1971) work, showing that the president's party won or lost votes in congressional elections, depending on voter satisfaction with the economy. Tufte (1978) extended the evidence to presidential elections, contending that Democratic elections are generally propelled by economic fluctuations, with incumbents held accountable by popular referendum. Fiorina (1981) brought the data focus back to individual voter attitudes, as measured in multiple national election surveys conducted by the American National Election Studies (ANES).

Following Key (1966), Fiorina (1981) claimed that voters rewarded or punished the president and his party retrospectively, based on recent past national economic activity. Kiewiet (1983) expanded on this by demonstrating that these national economic evaluations, "sociotropic" as he called them, were much more important for voters than their own private, pocketbook, concerns. Following suit, Lewis-Beck (1988) took the debate across the Atlantic, in his survey analysis of economic voting in five Western European nations, which evidenced clear economic voting in these democracies. Since this 'survey era' of economic voting research in the 1980s, the field has blossomed. This latter work has mostly raised technical issues, but also occasional challenges to the existence of the basic economic voter paradigm. We explore these matters in what follows.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.93148-9

Survey Research on Economic Evaluations

In their daily lives, individuals routinely evaluate all sorts of economic activity, such as the cost of groceries, the sales tax, the size of their pay check, or the national debt. We can systematically observe how individuals evaluate key aspects of the economic world around them by studying their responses in scientific social surveys. In *The American Voter* (Campbell et al., 1960: 393), they wanted to understand the 'economic outlook' voters held, by measuring "a set of rather general economic attitudes [with] no specific political referent." They developed a battery of eight closed-ended items, four of which are still widely in use, and are offered below (Campbell et al., 1960: 395). As can be seen, they ask the respondent to focus on one of two relevant objects (the family or the country), and one of two time frames (a year ago or a year from now).

> Would you say that you and your family are better off or worse off financially than you were *a year ago*?
>
> Now looking ahead – do you think that a year from now you will be better off financially, or worse off, or just about the same as now?
>
> Thinking about conditions in the country as a whole – Do you think that during the *next twelve months* we'll have good times financially, or bad times, or what?
>
> Would you say that at present business conditions are better or worse than they were a year ago?

These items, or close versions of them, have been regularly included in the ANES, conducted every congressional or presidential election year. See, for example, Lewis-Beck et al. (2008a: 365–389) who, using the ANES, relates such items to presidential vote preference in the 2000 and 2004 United States elections. Similar economic evaluations can be gleaned from election studies abroad, most notably the extended battery in the British Election Studies (Clarke et al., 2004). Further, in almost all European election surveys, a core battery of economic items is available, usually following a form close to that listed above. Lewis-Beck (1988: 44), exploring the economy and elections in Britain, France, Germany, Italy, and Spain, fielded 11 items on economic evaluation. More recent multinational studies look especially at retrospective national economic evaluations, e.g., Van der Burg et al. (2007) and Duch and Stevenson (2008).

In the investigation of economics and elections, a main reason researchers moved away from national macro-economic indicators, and to economic evaluation by surveyed individuals, was to overcome the ecological fallacy. Researchers wanted to avoid faulty inferences about individual voter behavior that might occur if only measures on the nation itself were examined. Of course, this notion rests on the assumption that individual level findings on the economic vote ultimately reflect national fluctuations. Fortunately, they appear to, at least with respect to retrospective sociotropic evaluations. Logically, Nannestad and Paldam (1994: 224) argue that, on average, changing perceptions of the economy should be linked to actual changes in the economy. Empirically, in a thorough study of that question, using eight Danish national election surveys, Lewis-Beck et al. (2013b) show that shifts in the national GDP tightly parallel shifts in the citizens' subjective sociotropic evaluations. Thus, we have the encouraging result that economic perceptions seem well-tied to actual economic fluctuations.

Pocketbook versus Sociotropic Economic Evaluation

Journalists, politicians, and everyday citizens often think that when economics counts in an election, it is because of how it impacts a voter's pocketbook. But this egotropic effect has generally been found to be considerably weaker than the so-called sociotropic effect, whereby what counts is how the voter sees the economy of the whole country. (On US elections, these initial findings are reported Alvarez and Nagler (1995), Kiewiet (1983) and Kinder and Kiewiet (1981). On European elections, see Anderson (2000).) Pocketbook effects are weak partly because citizens, perhaps Americans especially, tend to see personal economic successes as the result of their own hard work rather than stemming from government policies. If sociotropic effects prevail, however, it does not mean that the voter is necessarily an altruist, as is sometimes claimed. As Kiewiet and Lewis-Beck (2011) demonstrate, sociotropic voters might be altruists, believing they should reward the party in power for the running of a good economy. But, they could just as easily be egotists, who think they should vote for the party in power because it is delivering a good economy, which they will benefit from. That is to say, voters are able to see that their economic fate is joined with their neighbors.'

Retrospective versus Prospective Economic Voting

What is more important for the voter: the economy tomorrow or the economy yesterday? Looking at the past, the voter has a chance to evaluate the performance record of the ruling party (or party coalition). But, looking forward, the voter can think about whether there are brighter days ahead. It was Fiorina (1981) who staked out the claim for the superiority of retrospective economic voting. However, Downs (1957) and his followers made a strong case for the role of prospective assessments, with voters giving their weight to party promises. A good deal of research has been done, comparing the effects of the two, in elections both at home and abroad. Perhaps the first study was that of Fiorina (1978), who came out strongly for the retrospective side.

But the literature goes back and forth. See, for example, Clarke and Stewart (1994), who conclude that retrospective and prospective evaluations are of about equal importance. Still, a recent book by Lockerbie (2008) decides that, for US presidential elections, prospective economic evaluations dominate. In contrast, Price and Sanders (1995), in a comprehensive examination of the British case, conclude that prospective economic evaluations are much more important for their electorate. In a current investigation of American presidential elections, Campbell et al. (2010) find only retrospective effects. While the debate over the two time dimensions continues, in practice, we see more and more researchers giving the benefit of the doubt to the retrospective paradigm. An outstanding example comes from the largest economic voting investigation ever (165 surveys and 19 nations), by Duch and Stevenson (2008), who report robust sociotropic retrospective effects.

What Do Citizens Know about the Economy?

The 'economy' is not an attitude, but instead something real in the world. But do citizens see this reality? When they are asked to evaluate its performance, do they 'get it right?' The authors of *The American Voter* thought they did, when they said

> [E]conomic outlook does, through individual perceptions of the economic situation, reflect the 'real' world …. [T]here are significant changes in the national economy, particularly the consumer sector of the economy, which these attitudes foreshadow and reflect. Economic events have an impact on both personal and national components of economic outlook.
>
> Campbell et al., 1960: 394–395

What evidence do we have that voters do 'see' the economy? How well do their subjective perceptions compare to objective conditions? When asked to estimate the rates of unemployment, inflation, and growth, American respondents appear to be most accurate with regard to estimation of the unemployment rate (Conover et al., 1986; Holbrook and Garand, 1996; Lewis-Beck and Nadeau, 2009).

Moreover, the knowledge that American voters do have regarding the economy does not seem either enhanced or dampened, on balance, by the news media. For example, Hetherington (1996) finds that media attention in the 1984, 1988, and 1992 presidential elections does not alter their sociotropic retrospective economic evaluation. This result comports with the work of Haller and Norpoth (1997: 567), who, after their news analysis, conclude that "people are able to draw an economic picture that mirrors that real thing."

The strong link between the subjective and the objective economy seems true at the aggregate level, as well as individual level, of perception. Nadeau and Lewis-Beck (2001) show that an aggregated subjective measure of the economy (based on retrospective sociotropic assessments) correlates highly with several objective macro-economic indicators. In a follow-up study, Lewis-Beck et al. (2013c), looking at about 40 years of ANES data, report that sociotropic retrospective evaluation of the economy is shaped by GDP growth, the stock market, and inflation. In one of the few relevant comparative micro-studies, Duch and Stevenson (2010: 122) decide that "individual perceptions about the volatility of the macroeconomy are reasonably well informed; voters appear to understand the extent to which their economies are subject to shocks from the international economy; and voters who perceive that the variation in the national economy differs from variation in the global economy seem more inclined to exercise an economic vote." In sum, voters, appear to grasp the nature of the economy they live in.

How Economic Evaluations Might Be Distorted

While citizens can see the economy, they do not see it perfectly. In other words, when they look at it, they will sometimes make a faulty evaluation. What are the sources of these mistakes? Error can come from three main sources: group identification, level of information, and political bias. The first source, which we begin with, concerns the possible heterogeneity in the economic response caused by different group attachments. For example, Welch and Hibbing (1992) discovered that men were more likely to be pocketbook voters than women in the 1980 and 1984 elections. However, in a more recent study covering US presidential elections 1980–2004, Kam (2009) finds that there is no consistent pattern of difference in how men and women engage in economic voting. Duch et al. (2000), in their exploration of heterogeneity, located different economic evaluations among those with different experiences or information levels. This leads to the second possible source of distortion, coming from voters with more limited knowledge. With regard to this second source, some voters have more relevant information than others. Godbout and Bélanger (2007) show that more sophisticated voters give different weights to pocketbook and sociotropic evaluations. While such results suggest the possibility of some heterogeneity in economic voting, it should not be forgotten that, by and large, the main effects established by the assumption of homogenous economic voters are overwhelming. In sum, as Campbell et al. (1960: 394) concluded long ago, "[A] characteristic of this economic entity is found in the uniformity of its political consequences for many segments of the population. It cuts across class lines, party lines, and through socio-economic groupings, usually showing in all settings the same relationships to a given political variable."

While heterogeneous economic voters may cause little distortion in the general pattern of economic voting, some argue that a lot of distortion comes from political bias. Put another way, voter partisanship may introduce a conceptual lens into the economic evaluation, rendering more favorable judgments of the economy when their party is in power, and less favorable judgments when they are out of power. For example, if I am a Democrat evaluating the economy under the Obama administration, I will likely rate it more favorably than would a Republican. Such partisan bias can operate, as Kramer (1983) told us. The question becomes how great the bias is, and whether it can be controlled away. In the literature, this amounts to dealing with the potential endogeneity of economic evaluations. Certain critics have argued that partisan bias actually induces the apparent economic voting, and once taken into account will leave precious little evidence to support that hypothesis (Anderson, 2007; Evans and Pickup, 2010).

There have been different responses to this endogeneity claim. Duch and Stevenson (2008), attempt to exogenize the sociotropic retrospective economic evaluation variable, as do Lewis-Beck et al. (2008b). These studies engage, necessarily, in sophisticated econometric argument, including two-stage instrumental variable analysis of panel data from multiple countries, to demonstrate the continued existence of an economic vote. In perhaps the most rigorous of such studies to date, Nadeau et al. (2013) apply an instrumental variables analysis to repeated surveys from 10 European electorates. They conclude that, "economic perceptions appear strongly shaped by the objective economy," (p. 565) and these realistic perceptions go on to sharply influence their vote for or against the government.

In contrast to the foregoing elaborate analyses, Lewis-Beck et al. (2013c), simply regressed sociotropic retrospective economic evaluation on partisan bias (as measured by Republican or Democratic identification), via an extensive

examination of the ANES data. They found (in a large sample) that, while partisan bias was a statistically significant predictor of economic evaluation, it exercised a substantively small effect, and accounted for only a tiny part of the explanation (i.e., the pseudo-R-squared was only 0.02). Overall, then, it seems safe to return to the conclusion that voters, within reasonable limits of error, accurately perceive the economy and act politically on what they perceive.

Who Is Held Responsible for the Economy?

For citizens' economic evaluations to influence their vote, the evaluation must be attributed to a political actor who can be ultimately held accountable at the ballot box. In other words, there must be an attribution of blame or praise from the observed economic performance to a political actor. (Peffley and Williams (1985) have written the seminal theoretical piece here.) In presidential systems, such as the United States, that actor is commonly the president, while in parliamentary systems it's the prime minister. For example, in Germany, with its parliamentary system and coalition governments, the chancellor's party, and not the other coalition party, reaps rewards for a strong economic performance (Debus et al., 2014).

There are times when voters might not hold the incumbent government responsible but rather the previous administration. Take for example the recent economic recession. The financial crisis of 2008, which triggered the global economic recession, began under President Bush's watch, prior to Obama's victory in the November election. The economic challenges continued through Obama's first term in office. In 2012, when Obama was up for reelection, voters recognized that he inherited the crisis from the previous administration, and therefore were more likely to blame Bush than Obama for the continued economic sluggishness. This attribution of responsibility to the previous government rather than the incumbent administration helps account for Obama's reelection despite the state of the economy (D'Elia and Norpoth, 2014).

While many voters attribute responsibility for the economy to the political leadership, sophisticated voters understand that the economy is complex and influenced by many actors. The legislature, the central bank, government agencies, private enterprises, investors, and workers all drive the national economic engine. Indeed, some work has shown that the more knowledgeable the voter, the more likely he or she is to assign economic responsibility to some institution other than the presidency or government (Gomez and Wilson, 2003; Hellwig and Coffey, 2011). Additionally, multiple levels of government can cloud responsibility assignment. Such is the case in federal, as opposed to unitary systems (Stein, 1990), or when countries belong to supra-national institutions such as the European Union (Lobo and Lewis-Beck, 2012; Magalhães, 2014). Still, voters, rightly or wrongly, largely attribute responsibility for managing the economy to the national government and its leaders.

But, even this principle, that the president is the economic CEO, has been increasingly challenged by the realities of globalization. The economies of certain countries are heavily dependent on international trade, which means that what is happening in foreign markets directly impacts their domestic economies. In an important cross-national study, Fernández-Albertos (2006) shows that the greater the role of foreign trade, the weaker the attribution of economic responsibility to the national government. Writing more generally, Hellwig and Samuels (2007) find that with more open economics, the economic vote diminishes. As the world becomes ever more connected, the locus of economic responsibility may continue to shift away from the domestic central authorities, rendering the economic vote a more elusive target.

Institutions and the Economic Vote

The economic vote takes place in virtually all democracies. But that does not mean it has the same effect everywhere. We have already seen that, as attribution of economic responsibility varies, the economic effect can vary. However, attribution implies an individual psychological mechanism, operating in different ways across people. Besides individual variation, there can also be institutional variation, from one democracy to another. We have already mentioned, for example, that economic voting may work differently in a federal system, as opposed to a unitary one. There are other institutional arrangements that differ across nations that are still more important, and can influence the magnitude of the economic vote. In their path-breaking study of multiparty democracies, Powell and Whitten (1993) demonstrated that as coalitional governments became more complex, the impact of the economic vote was reduced. Why? Because things such as a greater number of coalition parties makes the line of responsibility for economic policy less clear. Anderson (2000), in a later coalitional study, demonstrates that, in addition to number, party size and unity also matter.

Certainly, institutional features can vary within a country, as well as across countries. Take the popular issue of divided government, where one party controls one elected body, and another party another elected body, e.g., the US president is represented by one party, and the US Congress by another. Norpoth (2001) found that, given such a division, the economic vote remained unaffected, for voters continued to hold the president responsible for the economy regardless. But, in France, divided government gives the economic vote a different outcome. According to Lewis-Beck and Nadeau (2004), under a French unitary government (the president and the prime minister from the same party coalition), the economic vote stands strong, but weakens under cohabitation (the president and the prime minister from different parties). Such examples suggest that the economic voter reasons carefully about who holds the economic reins, as he or she sorts through the labyrinth of institutional rules governing policy and election.

Observing the Economic Crisis

Since the Great Recession of 2008, public attention has been focused on what has come to be called the 'economic crisis' and its political consequences. With respect to the United States, it

appears that the resulting job losses had a clear effect on the presidential vote (Margalit, 2011). Arguing more generally, Campbell (2010) says the Great Recession secured Obama's victory in 2008. With respect to political effects in Europe, several useful article collections have appeared, the first being that of Escobar-Lemon and Whitten (2011), who provide a useful mapping of effects. A subsequent collection addresses the crisis in the European periphery, especially the hard hit nations of Southern Europe (Lewis-Beck et al., 2012). Most recently, an omnibus collection of 26 papers confronts the economic crisis and the vote in democracies worldwide (Lewis-Beck and Whitten, 2013). In those investigations, the classical economic voting theory of reward–punishment generally sustains itself. Furthermore, voters accurately perceive an economic crisis exists, meaning once again that their perception accords with objective conditions. Also, more heterogeneity of economic voting is in evidence, with group effects from class and income observed in some countries.

A lively question concerns whether the economic vote has become stronger, as a result of the crisis. The papers on Southern Europe seem to give a positive answer. For example, in Greece, Italy, Portugal, and Spain, the impact of retrospective sociotropic economic voting seems stronger, when compared to a sample of Northern European democracies (Lewis-Beck and Nadeau, 2012). However, other observers have argued that, despite crisis conditions, the economic vote remains muted. (About this point of view, for Germany see Anderson and Hecht, 2012; for Ireland, Marsh and Mikhaylov, 2012; for the United States, see Holbrook, 2012.)

This contention has also arisen for the Spanish case, which has suffered greatly as a result of the economic crisis, and illustrates well certain dilemmas and solutions. In Spain, almost everyone after the crisis saw the economy as 'bad,' but it was very difficult to detect an economic vote in routine surveys. This condition poses a methodological artifact – the "restricted variance problem" – making it almost impossible to uncover obvious economic voting effects. This difficulty can be solved with the introduction of time, via pooled cross-sections and panel data, whose analysis reveals strong economic voting effects in the 2008 election, and still stronger ones in the 2011 election (Fraile and Lewis-Beck, 2014). In conclusion, in Spain and other countries hard hit by the crisis, economic voting is at least as strong, and perhaps stronger, than before the crisis. The crisis, then, did not make economic voting go away. Instead, it allowed it to operate full force.

Other Sorts of Economic Evaluations

Thus far, our stress has been on sociotropic economic evaluations, where citizens assess the performance of the national economy. With respect to voting, these sociotropic evaluations are regarded as *valence* issues, i.e., an issue where voters hold a consensus view. In this case it means that voters all regard economic prosperity as a good. And, following the classic reward–punishment theory, when they perceive this prosperity, the will tend to vote for the government. Hence, economic voting theory becomes incumbency oriented. But, it might also be policy oriented, as Kiewiet (1983) sagely observed some time ago. In this situation, the citizen votes on the basis of economic policy. For example, a US voter may believe that the democrats offer better unemployment policies and thus will always vote Democratic, regardless of whether or not a Democrat is in the White House. Wright (2012), in a current study, finds that, in fact, rising unemployment always benefits the Democratic party, regardless of whether they are in office. Thus, economic voting becomes positional rather than valence, and is directed at party policy rather than party incumbency. Turning to Western European countries, Dassonville and Lewis-Beck (2013) offer a contemporary exploration of the effects of rising unemployment in the context of policy oriented, rather than incumbency oriented, economic voting.

In his seminal paper, Stokes (1963) was the first to note the distinction between the economy as a valence issue and the economy as a positional issue. In the vast literature on economic voting, almost all of it assumes the valence perspective. But, as Lewis-Beck and Nadeau (2011) have argued, there is room in economic voting theory for the role of positional issues. These positional issues usually can be arrayed left to right, and sort voters on economic policy matters such as unemployment, progressive taxation, market regulation, or industry nationalization. A recent study of the British General Election afforded the opportunity to directly compare valence and positional economic evaluations, and their impact on vote choice. It was discovered that position, measured as preference for progressive taxation, had almost the same effect as valence, measured in the usual sociotropic retrospective way (Lewis-Beck et al., 2013a). In a current examination of the German case, where positional and valence economic voting are systematically compared, they find strong effects for both, with a nod to the valence component (Clarke and Whitten, 2013). Obviously, expansion of the economic voting paradigm to include positional issues would have benefits, especially as it taps into the growing interest in income inequality and its political consequences.

Summary and Conclusions

In this article, we have explored specific investigations of economic evaluations and their electoral consequences, as they have been carried out in scientific survey research at home and abroad. A review of these investigations illuminates the vast terrain of economic voting work, which now numbers over 500 books and articles. With such a large body of work, review essays such as this can be helpful. But no one review essay serves as the be all and end all. We would also like to recognize other available reviews, Duch (2007), Hellwig (2010), Lewis-Beck and Stegmaier (2000, 2007, 2013), Lewis-Beck and Paldam (2000), and Stegmaier and Lewis-Beck (2013). While these essays highlight studies in the democracies of North America and Europe, economic voting work in the democracies of the rest of the world proceeds apace. (On these regions, see the review of Lewis-Beck and Stegmaier, 2008.) One final, general observation is worth making, with respect to this growing body of research. In virtually every democracy that has been investigated, the economic vote has appeared, as a statistically significant, and usually substantively significant, force shaping electoral outcomes, helping ensure Democratic accountability of the leaders to the people.

See also: Campaigning: Political; Parliamentary Government; Party Identification and Nonpartisanship; Party Systems; Political Science and Polling; Presidential Government; Voting, Explanations of: Social Class.

Bibliography

Alvarez, R.M., Nagler, J., 1995. Economics, issues and the Perot candidacy: voter choice in the 1992 presidential elections. American Journal of Political Science 39, 714–744.

Anderson, C.J., 2000. Economic voting and political context: a comparative perspective. Electoral Studies 19, 151–170.

Anderson, C.J., 2007. The end of economic voting? Contingency dilemmas and the limits of democratic accountability. Annual Review of Political Science 10, 271–296.

Anderson, C.J., Hecht, J.D., 2012. Voting when the economy goes bad, everyone is in charge, and no one is to blame: the case of the 2009 German election. Electoral Studies 31 (1), 5–19.

Van der Brug, W., Van der Eijk, C., Franklin, M., 2007. The Economy and the Vote: Economic Conditions and Elections in Fifteen Countries. Cambridge University Press, New York.

Campbell, J.E., 2010. The exceptional election of 2008: performance, values, and crisis. Presidential Studies Quarterly 40 (2), 225–246.

Campbell, A., Converse, P.E., Miller, W.E., Stokes, D.E., 1960. The American Voter. Wiley and Sons, New York.

Campbell, J.E., Dettrey, B.J., Yin, H., 2010. The theory of conditional retrospective voting: does the presidential record matter less in open-seat elections? Journal of Politics 72, 1083–1095.

Clarke, H.D., Stewart, M.C., 1994. Prospections, retrospections, and rationality: the "bankers" model of presidential approval reconsidered. American Journal of Political Science 38, 1104–1123.

Clarke, H.D., Whitten, G.D., 2013. Hard choices in hard times: valence voting in Germany (2009). Electoral Studies 32 (3), 445–451.

Clarke, H.D., Sanders, D., Stewart, M.C., Whiteley, P., 2004. Political Choice in Britain. Oxford University Press, Oxford.

Conover, P.J., Feldman, S., Knight, K., 1986. Judging inflation and unemployment: the origins of retrospective evaluations. Journal of Politics 48, 565–588.

Dassonville, R., Lewis-Beck, M.S., 2013. Economic policy voting and incumbency: unemployment in western Europe. Political Science Research and Methods 1 (1), 52–66.

D'Elia, J., Norpoth, H., 2014. Wining with a bad economy. Presidential Studies Quarterly 44, 467–483.

Debus, M., Stegmaier, M., Tosun, J., 2014. Economic voting under coalition governments: evidence from Germany. Political Science Research and Methods 2, 49–67.

Downs, A., 1957. An Economic Theory of Democracy. Harper and Row, New York.

Duch, R.M., 2007. Comparative studies of the economy and the vote. In: Boix, C., Stokes, S.C. (Eds.), The Oxford Handbook of Comparative Politics. Oxford University Press, Oxford, UK, pp. 805–844.

Duch, R.M., Palmer, H.D., Anderson, C.J., 2000. Heterogeneity in perceptions of national economic conditions. American Journal of Political Science 44, 635–652.

Duch, R.M., Stevenson, R.T., 2008. The Economic Vote: How Political and Economic Institutions Condition Election Results. Cambridge University Press, Cambridge.

Duch, R.M., Stevenson, R., 2010. The global economy, competency, and the economic vote. Journal of Politics 72, 105–123.

Escobar-Lemon, M., Whitten, G.D. (Eds.), 2011. The politics of economic crisis. Electoral Studies 30 (Special Symposium).

Evans, G., Pickup, M., 2010. Reversing the causal arrow: the political conditioning of economic perceptions in the 2000–2004 US presidential election cycle. The Journal of Politics 72 (04), 1236–1251.

Fernández-Albertos, J., 2006. Does internationalization blur responsibility? Economic voting and economic openness in 15 European countries. West European Politics 29, 28–46.

Fiorina, M.P., 1978. Economic retrospective voting in American national elections: a micro-analysis. American Journal of Political Science 22 (2), 426–443.

Fiorina, M., 1981. Retrospective Voting in American National Elections. Yale University Press, New Haven.

Fraile, M., Lewis-Beck, M.S., 2014. Economic vote instability: endogeneity or restricted variance? Spanish panel evidence from 2008 and 2011. European Journal of Political Research 53 (1), 160–179.

Godbout, J.F., Bélanger, É., 2007. Economic voting and political sophistication in the United States. A reassessment. Political Research Quarterly 60 (3), 541–554.

Gomez, B.T., Wilson, J.M., 2003. Causal attribution and economic voting in American congressional elections. Political Research Quarterly 56 (3), 271–282.

Haller, H.B., Norpoth, H., 1997. Reality bites: news exposure and economic opinion. Public Opinion Quarterly 61, 555–575.

Hellwig, T., 2010. Elections and the economy. In: LeDuc, L., Niemi, R.G., Norris, P. (Eds.), Comparing Democracies 3: Elections and Voting in the 21st Century. Sage Publications, Thousand Oaks, CA, pp. 185–201.

Hellwig, T., Coffey, E., 2011. Public opinion, party messages, and responsibility for the financial crisis in Britain. Electoral Studies 30 (3), 417–426.

Hellwig, T., Samuels, D., 2007. Voting in open economies the electoral consequences of globalization. Comparative Political Studies 40 (3), 283–306.

Hetherington, M., 1996. The media's role in forming voters' national economic evaluations in 1992. American Journal of Political Science 40, 372–395.

Holbrook, T., Garand, J., 1996. Homo economus? Economic information and economic voting. Political Research Quarterly 49, 351–375.

Holbrook, T.M., 2012. Incumbency, national conditions, and the 2012 presidential election. Political Science & Politics 45 (04), 640–643.

Kam, C.D., 2009. Gender and economic voting, revisited. Electoral Studies 28 (4), 615–624.

Key, V.O., 1966. The Responsible Electorate. Vintage, New York.

Kiewiet, D.R., 1983. Macroeconomics and Micropolitics: The Electoral Effects of Economic Issues. University of Chicago Press, Chicago.

Kiewiet, D.R., Lewis-Beck, M.S., 2011. No man is an island: self-interest, the public interest, and sociotropic voting. Critical Review 23 (3), 303–319.

Kinder, D.R., Kiewiet, D.R., 1981. Sociotropic politics: the American case. British Journal of Political Science 11, 129–161.

Kramer, G.H., 1971. Short-term fluctuations in US voting behavior, 1896–1964. The American Political Science Review 65, 131–143.

Kramer, G., 1983. The ecological fallacy revisited: aggregate- versus individual-level findings on economics and elections, and sociotropic voting. American Political Science Review 77, 92–111.

Lewis-Beck, M.S., 1988. Economics and Elections: The Major Western Democracies. University of Michigan Press, Ann Arbor.

Lewis-Beck, M.S., Lobo, M.C., Bellucci, P. (Eds.), 2012. Economic crisis and elections: the European periphery. Electoral Studies 31 (3) (Special Symposium).

Lewis-Beck, M.S., Jacoby, W., Norpoth, H., Weisberg, H.F., 2008a. The American Voter Revisited. University of Michigan Press, Ann Arbor.

Lewis-Beck, M.S., Nadeau, R., Elias, A., 2008b. Economics, party and the vote: causality issues and panel data. American Journal of Political Science 52, 84–95.

Lewis-Beck, M.S., Nadeau, R., 2004. Dual governance and the economic vote: France and the United States. In: Lewis-Beck, Michael S. (Ed.), The French Voter: Before and after the 2002 Elections. Palgrave Macmillian, pp. 136–154.

Lewis-Beck, M.S., Nadeau, R., 2009. Obama and the economy in 2008. Political Science and Politics 43, 479–483.

Lewis-Beck, M.S., Nadeau, R., 2011. Economic voting theory: testing new dimensions. Electoral Studies 30 (2), 288–294.

Lewis-Beck, M.S., Nadeau, R., 2012. PIGS or not? economic voting in Southern Europe. Electoral Studies 31 (3), 472–477.

Lewis-Beck, M.S., Nadeau, R., Foucault, M., 2013a. The compleat economic voter: new theory and British evidence. British Journal of Political Science 43, 241–261.

Lewis-Beck, M.S., Stubager, R., Nadeau, R., 2013b. The Kramer problem: micro-macro resolution with a Danish pool. Electoral Studies 32, 500–505.

Lewis-Beck, M.S., Martini, N.F., Kiewiet, D.R., 2013c. The nature of economic perceptions in mass publics. Electoral Studies 32 (3), 524–528.

Lewis-Beck, M.S., Paldam, M., 2000. Economic voting: an introduction. Electoral Studies 19 (2–3), 113–121.

Lewis-Beck, M.S., Stegmaier, M., 2000. Economic determinants of electoral outcomes. Annual Review of Political Science 3, 183–219.

Lewis-Beck, M.S., Stegmaier, M., 2007. Economic models of voting. In: Dalton, R., Klingemann, H.-D. (Eds.), The Oxford Handbook of Political Behavior. Oxford University Press, Oxford, UK, pp. 518–537.

Lewis-Beck, M.S., Stegmaier, M., 2008. The economic vote in transitional democracies. Journal of Elections, Public Opinion and Parties 18, 303–323.

Lewis-Beck, M.S., Stegmaier, M., 2013. The VP-function revisited: a survey of the literature on vote and popularity functions after over 40 years. Public Choice 157 (3–4), 367–385.

Lewis-Beck, M.M., Whitten, G.D., (Eds.), 2013. Economics and elections: effects deep and wide. Electoral Studies 32 (3) (Special Symposium).
Lockerbie, B., 2008. Do Voters Look to the Future?: economics and Elections. SUNY Press, Albany.
Lobo, M.C., Lewis-Beck, M.S., 2012. The integration hypothesis: how the European Union shapes economic voting. Electoral Studies 31 (3), 522–528.
Magalhães, P.C., 2014. The elections of the great recession in Portugal: performance voting under a blurred responsibility for the economy. Journal of Elections. Public Opinion&Parties 24, 180–202.
Margalit, Y., 2011. Costly jobs: trade-related layoffs, government compensation, and voting in US elections. American Political Science Review 105 (01), 166–188.
Marsh, M., Mikhaylov, S., 2012. Economic voting in a crisis: the Irish election of 2011. Electoral Studies 31 (3), 478–484.
Nadeau, R., Lewis-Beck, M.S., 2001. National economic voting in U.S. presidential elections. Journal of Politics 63, 159–181.
Nadeau, R., Lewis-Beck, M.S., Bélanger, É., 2013. Economics and elections revisited. Comparative Political Studies 46 (5), 551–573.
Nannestad, P., Paldam, M., 1994. The VP function: a survey of the literature on vote and popularity functions after 25 years. Public Choice 79, 213–245.
Norpoth, H., 2001. Divided government and economic voting. Journal of Politics 63 (2), 414–435.
Peffley, M., Williams, J.T., 1985. Attributing presidential responsibility for national economic problems. American Politics Research 13 (4), 393–425.
Powell Jr., G.B., Whitten, G.D., 1993. A cross-national analysis of economic voting: taking account of the political context. American Journal of Political Science 37, 391–414.
Price, S., Sanders, D., 1995. Economic expectations and voting intentions in the UK, 1979–1989: a pooled cross-section approach. Political Studies 43, 451–471.
Stegmaier, M., Lewis-Beck, M.S., 2013. Economic voting. In: Valelly, R. (Ed.), Oxford bibliographies in Political Science. Oxford University Press, New York.
Stein, R.M., 1990. Economic voting for governor and US senator: the electoral consequences of federalism. The Journal of Politics 52 (01), 29–53.
Stokes, D.E., 1963. Spatial models of party competition. The American Political Science Review 57, 368–377.
Tufte, E.R., 1978. Political Control of the Economy. Princeton University Press.
Welch, S., Hibbing, J., 1992. Financial conditions, gender, and voting in American national elections. The Journal of Politics 54 (01), 197–213.
Wright, J.R., 2012. Unemployment and the democratic electoral advantage. American Political Science Review 106 (04), 685–702.

Economic Geography

Edward J Malecki, The Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Economic geography is a branch of human geography that deals with all matters of economic interest, from livelihoods and resource extraction, to production and exchange, to consumption and disposal. The definition of what is economic includes what is within formal economies and paid labor as well as informal production and services and subsistence agriculture. These embrace all of capitalism as well as what is just beyond its boundaries. Fundamentally, economic geographies do not view the economy as distinct or independent from social, cultural, and political processes. A hallmark of economic geography is that it is ever-changing, applying, and incorporating new ideas from other disciplines to enrich our knowledge of things economic. The overriding concern of economic geographers is to understand the world as it actually exists, rather than as it might be modeled.

Economic geography is a branch of human geography that deals with all matters of economic interest, from livelihood to resource extraction, to production and exchange, to consumption and disposal. The field has grown markedly in breadth from the early twentieth century when the focus of what was called commercial geography was primarily restricted to resources, mining, manufacturing, commercial activity, and trade flows of commodities and products.

The scope of economic geography encompasses theoretical and empirical research on what are viewed as economic activities. Economic activity, in the form of livelihood strategies within and outside of capitalism, has always existed. Global capitalism has made money and the need for money more widespread and multinational firms have brought formal jobs to more (but not all) places. The definition of what is economic includes what is within formal economies and paid labor as well as informal production and services and subsistence agriculture. Thus, economic geography encompasses all of capitalism as well as what is just beyond its boundaries. A hallmark of economic geography is that it is ever-changing, picking up new ideas from other disciplines to enrich our knowledge of things economic. This openness to new ideas and to analysis of new and rapidly changing phenomena encourages diversity within the subfield.

The resulting variegation belies several broad commonalities among economic geographers, evident in a common belief that the spatialities of the economy are not exogenous but are constructed or produced, and that institutions, governance, regulation, social construction, and indeed politics must be included (Sheppard et al., 2012).

Background and History

Economic geographers were at the forefront of the quantitative revolution of the 1960s, focusing on location theory and data-rich quantitative analysis. Industrial geography was for a time a thriving subfield, enriching our knowledge of multinational corporations as they came to dominate the global economy. Economic geographers continue to focus on understanding the firm (Taylor and Oinas, 2006).

Until the late 1970s, it was relatively straightforward to define economic geography; the field encompassed industrial and commercial activities as well as resource extraction. Rural primary sector activities received minimal attention, except as they provided raw materials for industrial activities or food for urban markets. Globalized agribusiness illustrates the central role of large, transnational firms in the organization of economic activities. Global production networks (GPNs) largely shape the extraction of other resources as well.

Manufacturing and its locations were of interest, as they varied markedly from country to country, and remain a focus of industrial geography. The emergence of postindustrial occupations, particularly in information and knowledge-related sectors, led to a new focus on the service economy. Similarly, global trade has expanded as free trade doctrine has prevailed in economic policy, generally leaving aside questions of uneven development.

The decline of manufacturing in wealthy countries gave rise to the study of economic restructuring as large firms shifted production to lower wage locations. Not only cities but also rural areas are affected by restructuring, as farms add tourism and other activities to agricultural (food and fiber) production. Global commodity chains (GCCs) link areas where (food and nonfood) crops are grown to the demands of (typically) wealthy, urban consumer markets.

Services became the dominant employer in rich countries as manufacturing shifted to locations with lower wages. Through the spread of multinational (or transnational) enterprises, transportation and communication grew as subfields central to the global economy. The concept of spatial divisions of labor (Massey, 1995) has developed into sophisticated study of GPNs. GPNs incorporate many kinds of relationships and network configurations, and they expand beyond interfirm transactions to encompass all relevant sets of actors and relationships (Coe et al., 2008).

Among services, finance is special, the lubricant of the global capitalist economy. The geography of finance has risen in significance as the intertwining of financial institutions, corporations, and consumers (Clark and Wójcik, 2007). Since the global financial crisis of 2008, the fates of governments, firms, and individuals have become even more confusing in

the continuing unsettled nature of debt and finance in the euro zone and beyond (Lee et al., 2009).

Retailing, which supports the business of consumption, has not been neglected. The sector ranges from locally owned shops to massive global firms with thousands of stores. The global retail giants, such as Wal-Mart and Carrefour, and hundreds of small chains, make decisions that determine the nature and locations of the many sites of economic activity, from the manufacture of goods and warehouses to the sites of retail shops and malls (Wrigley et al., 2005). Online e-commerce has added a new, virtual dimension to consumption.

Some subfields of geography began within economic geography but now thrive largely in lively specialized and interdisciplinary subfields, including entrepreneurship, transportation, and tourism. Much of economic geography takes place within cities and urban areas; urban geography is a vibrant subfield that encompasses the variety of human life and its manifestations that are found in cities. Effects of economic processes on places give rise to uneven development and inequality at many scales, from local and regional to national and global.

Aoyama et al. (2011) provide a concise list of the key agents in economic geography: labor, firms, and states, and key drivers of economic change: innovation, entrepreneurship, and accessibility. Economic change as it takes shape in industries and regions includes industrial location, industrial clusters, regional disparity, and post-Fordism. Global economic geographies address core-periphery, globalization, and circuits of capital. In all of these, economic geographers seek to understand the world as it actually exists, rather than as it might be modeled.

Theory and Methods

Recent and ongoing transformations of capitalism have prompted attempts by economic geographers to understand its workings in organizations such as firms as well as its effects on the ground, particularly in cities. A prominent example, going on for decades, is the shift from manufacturing and its demand for physical labor toward services and knowledge work. The new forms of capitalism call for theory that involves people as more than merely labor, encompassing capabilities and emotions. These changes in labor and life have profound impacts on cities and theories to explain them, not least in a new division of labor between a cognitive–cultural class and a service class (Scott, 2011).

The breadth of economic geography is striking, marked by interdisciplinarity. Since the 'cultural turn' in the 1990s, there has been a growing use of the term 'economic geographies' in which the economic and the social are seen as inseparable. Rather than assuming one economy and a market that sets prices, wages, and other dimensions of value, economic activity, economic life, and therefore economic geographies are complex, practiced, and entangled in particular places and times (Lee, 2006).

Corresponding to its breadth, the methods of economic geography have evolved from mainly looking for patterns based on large data sets to include qualitative research on relatively small numbers of subjects in an effort to uncover the complexity of actual economic geographies in particular places. Both types of research are common, as are frequent syntheses from neighboring disciplines that offer new insights to the issues within economic geography(ies).

Economic Geography and Geographical Economics

From the late 1950s through the mid-1970s, economic geography was aligned with economics in the interdisciplinary field of regional science through a common reliance on mathematical models in location theory, for example, and quantitative empirical analysis of industries and labor. The deindustrialization of Western economies beginning in the late 1960s precipitated a divergence between geography and economics. Geographers began a serious effort to understand capitalism and the social construction of the diversity of things economic. There remains a research effort in quantitative economic geography, often called regional political economy (Sheppard, 2011).

Economists have held true to their disciplinary core and its central models, whereas geographers have searched for explanations and solutions in other disciplines. The result, still ongoing decades later, has been an explosion of research by economic geographers to understand structural shifts from manufacturing to services, information, innovation, and knowledge; cultural shifts that recognize the diverse actors in economic activity; and technological shifts that facilitate customized and personalized products and a global reach through information and communication technologies (ICTs) such as the Internet.

The rise of the 'new economic geography' (NEG) also laid to rest the mutual respect between economists and geographers that had prevailed to a large degree within regional science. Paul Krugman's model of economic geography quickly became the standard (Krugman, 1991). In 2008, Krugman's receipt of the Nobel Memorial Prize in Economic Sciences for his contributions to new trade theory and NEG conferred further legitimacy on what has become known as the NEG. This model provides a central core for geographical economists. For economists, the benefits of a tractable model with restrictive assumptions and a set of 'modeling tricks' are preferable to the complexities of the real world (Krugman, 1998). The simplifying assumptions, however, strip away social, institutional, cultural, and historical context, and geography is reduced to transport costs.

Reactions to NEG have highlighted afresh the contrasts between economists and geographers in fundamental beliefs and methods. In particular, economists gravitate toward mathematical models and a focus on equilibrium (or, at most, multiple equilibria), whereas geographers tend toward in-depth case studies, real-world examples, and qualitative methods as they study an economy in constant change (Martin, 2011). Economic geographers, in contrast to economists working within NEG, have continued a spatially aware synthesis of ideas drawn from many disciplines. Indeed, the breadth and diversity of economic geography have become hallmarks of the field (Duranton and Rodriguez-Pose, 2005).

A central cause of the distinct stances of the two disciplines is the neat assumed world of economic theory and the messy

and diverse nature of 'the economy.' Whereas the mainstream economic view is of transactions of wage labor and private property in markets with ample finance, Gibson-Graham (2006) proposes a diverse economy depicted by an iceberg. Like an iceberg, only the visible tip on the top is the traditional capitalist economy and the larger part, beneath the water line, is a variety of ways in which people make their livelihoods and seek to survive outside the formal economy.

Within the formal economy, gender shatters the notion of 'economic man' and homogenous actors. Since the 1980s, feminist researchers have added new richness to economic geography. Significantly, feminist economic geography highlights the difference that gender makes, and the work and lives of women within as well as outside the formal labor market (McDowell, 2009). Gender-specific social stereotyping of occupations and domestic roles create very different geographies (Massey, 1995; Johnson, 2011).

Fundamentally, geographers prefer to see the world as complex and continually being made, with the economic inseparable from the cultural and social. Geography's inherent interdisciplinarity is seen vividly in the degree to which economic geographers read and utilize research by scholars in other disciplines (Duranton and Rodriguez-Pose, 2005). Overt attempts at dialogue can be seen in the *Journal of Economic Geography*, edited by economists and geographers. In most other respects, the two disciplines interact or converse very little.

Although economic geography's relationship with economics is tense, at best, economic geographers remain engaged with various branches of heterodox economics. Here, outside mainstream economics, mathematical models are not mandatory, and appreciative or verbal theory accounts for more of the complexities of the real world. In innovation studies, geographers contribute on an equal footing with researchers from business, management, and organization studies. This interdisciplinary area has seen bursts of research on such topics as industrial clusters, technological learning, and territorial knowledge dynamics.

Objects of Study

Economic geographers have studied a diverse range of phenomena and processes: learning and innovation; industrial organization; market systems; networks and globalization; livelihood strategies; development; race, class, and gender relations; neoliberal governance; and consumption. What links these often distinct foci and their literature is an explicit interest in what Jones and Murphy (2011) call "socioeconomic practices." These are the stabilized, routinized, or improvised social actions that constitute and reproduce economic space, through and within which diverse actors (such as entrepreneurs, workers, caregivers, consumers, and firms) and communities (industries, places, markets, cultural groups) organize materials, produce, consume, and/or derive meaning from the economic world. Practices may not only serve instrumental purposes (such as production, consumption, and learning), but they are also significant analytically because they can reveal the complexities, contingencies, identities, and meanings inherent in the many forms of economic organization (Jones and Murphy, 2011).

Relational economic geography "seeks to identify the complex relational geometry comprising local and non-local actors, tangible and intangible assets, formal and informal institutional structures, and their interactive power relations" (Yeung, 2005: p. 48). The focus is on actors and their actions and practices within the "multitude of material, labor market, technology, and information linkages" which comprise any commodity chain or production structure (Bathelt and Glückler, 2011: p. 8). Networks of all sizes and shapes serve to illustrate the relational economic geographies of real people in real places and economic activities. Power relations permeate the governance and regulation of all such relationships, producing new inequalities and reinforcing old ones.

Cultural economic geography embraces all ways in which the cultural and the social shape the actions of economic actors and the economy itself. This perspective acquired greater importance as more and more 'noneconomic' features were recognized as being intertwined with economic activity (James et al., 2007). Moreover, culture and creativity also are produced, distributed, exhibited, and consumed as cultural products, services, and experiences.

Thus, geographies of production focus on economic systems as networked formations, constituted by a mesh of relationships that operate within and across the global, national, and regional/local scales. At the global scale, the idea of GPNs (or GCCs or global value chains) has helped to uncover the webs created by transnational firms and the many affiliates, partners, suppliers, and subcontractors that underlie the more-or-less visible global capitalist economy (Coe et al., 2008).

Evolutionary economic geography (EEG) draws on parallel work in evolutionary economics, an established heterodox branch of economics, and also combines formal modeling typical of neoclassical economics with the contextuality of human decision-making, such as local history and institutions. In general, EEG attempts to explain regional growth differences from the microhistories of firms that operate in territorial contexts.

The goal of EEG is to interpret and explain how the economic landscape changes over time, and to demonstrate how geography matters in determining the nature and trajectory of evolution of the economic system (Boschma and Martin, 2010). Among the principal challenges is to understand the spatialities of novelty (innovations, new firms, new industries, new networks) to explain path creation and emergence rather than only path dependence. The evolutionary research effort is wide-ranging, incorporating Darwinian concepts from evolutionary biology, complexity theory such as emergence, and mechanisms that lead into – and out of – path dependence. MacKinnon et al. (2009), however, suggest that a stronger basis in institutional economics and political economy is needed to relate the evolution of the economic landscape to processes of capital accumulation and uneven development.

All of the perspectives above, each with its own preferred methodologies, struggle for attention and dominance of theoretical and empirical study. All perspectives aim to

comprehend the economic geographies that evolve and manifest simultaneously but differently at global, national, and regional/local scales. The degree to which any actor and its networks are embedded in any specific context is among the concerns for theory and empirical research alike.

On the whole, economic geography is characterized – and contrasts sharply with economics – by having no canon of standard models and tenets; being intellectually diverse and porous, welcoming and receptive to new ideas; and therefore having inherent interdisciplinarity. Fundamentally, economic geographies are socially constructed and sustained, enmeshed in social, cultural, and political processes.

Empirical Challenges: The Economic Geographies of a Changing World

The overriding concern of economic geographers is to understand the world as it actually exists, rather than as it is stripped of complexity so that it can be modeled mathematically. The spread of multinational firms provides employment and income for some, while livelihood strategies within and outside of capitalism are still common. New sites of growth, particularly East Asia beginning with the rise of the Japanese economy in the 1970s and of Chinese economy in the 1990s, have added to our knowledge of non-Western varieties of capitalism.

Virtual or Digital Economies

The virtual economy or digital economy, facilitated by ICTs, is another venue within which economic activity takes place. The virtual economy adds new meanings and measures to the concepts of distance and proximity. Digital technologies enable more activity in more places, work tasks can be done 'offshore,' and new firms can be 'born global.' At the same time, many long-standing inequalities remain as new technologies favor established wealthier markets. Mobile phones and digital technologies are changing the lives of small farmers in the global south by reducing the barrier of distance and altering access to information about markets and prices.

Environmental Economic Geographies

The environmental impact of economic activity has emerged into environmental economic geography. Environmental economic geography works to understand environmental impacts as largely an outcome of the decisions and business practices of firms, whose pursuit of profit has seldom taken the environment into account. The general outlines of environmental economic geography acknowledge that substantial changes in technologies, industries, and lifestyles – and in technoeconomic paradigm – are needed to create a 'sustainability transition.' The global study of waste by Crang et al. (2013), who have followed goods past consumption to recycling, has uncovered significant activity in informal networks, again illustrating a large iceberg of informal activity beneath the surface of the formal economy.

Asia in the Global Economy

The shift in economic power to Asia began in the 1970s as Japanese firms outcompeted companies from Europe and North America in products such as cars and electronics. Other Asian firms, first from the 'four tigers' (Hong Kong, Singapore, South Korea, and Taiwan) soon followed. Each economy had a combination of government policy and learning from the best practices and mistakes of others. The result has been an evolution from followers to market leaders, notably in the electronics industry (Yeung, 2009).

The opening of China's economy to global capitalism – albeit with Chinese characteristics and control – has led to new 'divergent hybrid capitalisms in China' (Yang, 2007). In various ways, Chinese firms (from mainland China, rather than from Taiwan or Hong Kong), enmeshed with the Chinese state, have become major actors in the global economy.

Latin America, Africa, and the Global South

In the midst of Asia's growing 'clout' in the global economy, other regions have been relatively neglected within economic geography. In Latin America, economic geography remains focused on the postcolonial effort against neoliberalism (Werner, 2012). African economies likewise are changing. The 'new scramble for Africa' is the effort by large firms – now including those from China as well as from the West – to secure supplies of oil and other minerals as well as food supplies (Carmody, 2009). At the same time, another new scramble – for livelihoods – continues and drives rural change and urbanization (Bryceson, 2002). The results of sudden change include immense economic and social disparities within African cities.

The search to understand the world of practices and work outside capitalism leads economic geographers to examine inequalities at a global scale, in the world economy's 'periphery,' and in the Global South. Typically the focus of development studies, several phenomena, including globalization, industrial location, migration, and poverty, have become critical issues in rich and poor countries alike. Development research, long the global view of economic geographers, faces new challenges as the global economy becomes ever more non-Western.

Conclusion

Economic geography has blossomed well beyond its original focus on the location of production to embrace the many other human forces – social, cultural, political, and institutional – which affect and are affected by economic activity. Constant change ensures that new ideas and new empirical knowledge are always characteristic of the field, and this attracts new researchers and their fresh thinking and new approaches.

A group of midcareer (i.e., not senior) geographers recently have proposed a research agenda for economic geography, centering on five research themes on which too little is known: networks, firms, and markets; economic geography of global

environmental change; the geography of finance; digital property including challenges, processes, regulation; and a more 'global' economic geography (Benner et al., 2011). These and other emerging topics will be among the future economic geographies to be written.

See also: Consumption, Geography of; Development Theory in Geography; Economic Restructuring: Geographic Aspects; Energy, Geography of; Entrepreneurship, Geography of; Finance, Geography of; Geographies of Food: Politics, Place, Power Reprised; Industrial Geography; Information Society; International Business: Geographic Aspects; Knowledge, Geography of; Labor, Geography of; Local Economic Development; Location Theory; Nature and Society in Geography; Networks, Geography of; Poverty, Geography of; Regional Science; Resource Geography; Rural Geography; Service Economy, Geography of; Tourism, Geography of; Transportation Geography; Urban Geography.

Bibliography

Aoyama, Y., Murphy, J.T., Hanson, S., 2011. Key Concepts in Economic Geography. Sage, London.

Bathelt, H., Glückler, J., 2011. The Relational Economy: Geographies of Knowing and Learning. Oxford University Press, Oxford.

Benner, C., Berndt, C., Coe, N., Engelen, E., Essletzbichler, J., Glassman, J., Glückler, J., Grote, M., Jones, A., Leichenko, R., Leslie, D., Lindner, P., Lorenzen, M., Mansfield, B., Murphy, J.T., Pollard, J., Power, D., Stam, E., Wòjcik, D., Zook, M., 2011. Emerging themes in economic geography: outcomes of the Economic Geography 2010 Workshop. Economic Geography 87, 111–126.

Boschma, R., Martin, R., 2010. The aims and scope of evolutionary economic geography. In: Boschma, R., Martin, R. (Eds.), The Handbook of Evolutionary Economic Geography. Edward Elgar, Cheltenham, pp. 3–39.

Bryceson, D.F., 2002. The scramble in Africa: reorienting rural livelihoods. World Development 30, 725–739.

Carmody, P., 2009. The New Scramble for Africa. Polity Press, Cambridge.

Clark, G.L., Wójcik, D., 2007. The Geography of Finance: Corporate Governance in the Global Marketplace. Oxford University Press, Oxford.

Coe, N.M., Dicken, P., Hess, M., 2008. Global production networks: realizing the potential. Journal of Economic Geography 8, 271–295.

Crang, M., Hughes, A., Gregson, N., Norris, L., Ahamed, F., 2013. Rethinking governance and value in commodity chains through global recycling networks. Transactions of the Institute of British Geographers 38, 12–24.

Duranton, G., Rodriguez-Pose, A., 2005. When economists and geographers collide, or the tale of the lions and the butterflies. Environment and Planning A 37, 1695–1705.

Gibson-Graham, J.K., 2006. A Poststructuralist Politics. University of Minnesota Press, Minneapolis.

James, A., Martin, R.L., Sunley, P., 2007. The rise of cultural economic geography. In: Martin, R., Sunley, P. (Eds.), Economic Geography, The Cultural Economy, vol. IV. Routledge, London, pp. 3–18.

Johnson, L., 2011. Feminist economic geographies. In: Leyshon, A., Lee, R., McDowell, L., Sunley, P. (Eds.), The Sage Handbook of Economic Geography. Sage, London, pp. 353–367.

Jones, A., Murphy, J.T., 2011. Theorizing practice in economic geography: foundations, challenges, and possibilities. Progress in Human Geography 35, 366–392.

Krugman, P., 1991. Geography and Trade. MIT Press, Cambridge, MA.

Krugman, P., 1998. What's new about the new economic geography? Oxford Review of Economic Policy 14 (2), 7–17.

Lee, R., 2006. The ordinary economy: tangled up in values and geography. Transactions of the Institute of British Geographers New Series 31, 413–432.

Lee, R., Clark, G.L., Pollard, J., Leyshon, A., 2009. The remit of financial geography – before and after the crisis. Journal of Economic Geography 9, 723–747.

MacKinnon, D., Cumbers, A., Pike, A., Birch, K., McMaster, R., 2009. Evolution in economic geography: institutions, political economy, and adaptation. Economic Geography 85, 129–150.

Martin, R., 2011. The 'new economic geography': credible models of the economic landscape? In: Leyshon, A., Lee, R., McDowell, L., Sunley, P. (Eds.), The Sage Handbook of Economic Geography. Sage, London, pp. 53–71.

Massey, D., 1995. Spatial Divisions of Labour, second ed. Macmillan, London.

McDowell, L., 2009. Working Bodies: Interactive Service Employment and Workplace Identities. Wiley-Blackwell, Chichester.

Scott, A.J., 2011. A world in emergence: notes toward a resynthesis of urban-economic geography for the 21st century. Urban Geography 32, 845–870.

Sheppard, E., 2011. Geographical political economy. Journal of Economic Geography 11, 319–331.

Sheppard, E., Barnes, T.J., Peck, J., 2012. The long decade: economic geography, unbound. In: Barnes, T.J., Peck, J., Sheppard, E. (Eds.), The Wiley-Blackwell Companion to Economic Geography. Wiley-Blackwell, London, pp. 1–24.

Taylor, M., Oinas, P. (Eds.), 2006. Understanding the Firm: Spatial and Organizational Dimensions. Oxford University Press, Oxford.

Werner, M., 2012. Contesting power/knowledge in economic geography: learning from Latin America and the Caribbean. In: Barnes, T.J., Peck, J., Sheppard, E. (Eds.), The Wiley-Blackwell Companion to Economic Geography. Wiley-Blackwell, London, pp. 132–145.

Wrigley, N., Coe, N.M., Currah, A., 2005. Globalizing retail: conceptualizing the distribution-based transnational corporation (TNC). Progress in Human Geography 29, 437–457.

Yang, C., 2007. Divergent hybrid capitalisms in China: Hong Kong and Taiwanese electronics clusters in Dongguan. Economic Geography 83, 395–420.

Yeung, H.W.-C., 2005. Rethinking relational economic geography. Transactions, Institute of British Geographers 30, 37–51.

Yeung, H.W.-C., 2009. Regional development and the competitive dynamics of global production networks: an East Asian perspective. Regional Studies 43, 325–351.

Relevant Websites

http://www.egrg.org.uk/ – Economic Geography Research Group, Institute of British Geographers.

http://aag-econ.org – Economic Geography Specialty Group, Association of American Geographers.

Economic Growth

David de la Croix, IRES, Université catholique de Louvain, Louvain-la-Neuve, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The first challenge for economic growth theory is to understand the transition from stagnation to growth at the time of the industrial revolution and in particular to identify the main factor(s) that triggered the takeoff. Doing so also helps to understand why there are poor and rich countries today, and whether the poorest ones will ultimately catch up. This article reviews the main theories of growth, including the Malthus model (useful to understand stagnation), the neoclassical model where technical progress is the engine of growth, the endogenous growth model where growth is self-sustained and policy is of particular importance for the long run, and unified growth theory, providing the big picture and linking the growth takeoff to the demographic transition.

Definition and Measurement

Economic growth is the process by which the amount of goods and services one can earn with the same amount of work increases over time. It generally implies that income per person rises over time (unless hours of work fall steadily).

Almost everything people buy today requires fewer days of work than it did in the past. A classic example is provided by Nordhaus (1996: Table 1.6) in his history of lighting. According to him, 10 min of work today buys 3 h of reading light each night of the year, while it only bought 10 min of light per year two centuries ago.

Growth theory attempts to model and understand the factors behind this process. It is a relatively young field of research, as one can establish by searching for 'economic growth' in all the books published, which have been scanned by Google (about 10% of all books). **Figure 1** shows that the emergence of the field can be dated to the 1950s.

For a relevant theory of growth, one needs good data first. Measuring growth is difficult, especially for periods for which little information is available. National Accounts were set up in most countries after World War II. They provide different ways to measure income per capita (Gross Domestic Product (GDP), etc.). Making data comparable across countries requires correcting for differences in price levels to obtain estimates that capture the real purchasing power of income. The most comprehensive database so far is the Penn World Tables version 8.0 (Feenstra et al., 2013), which provides information on relative levels of income, output, inputs, and productivity, covering 167 countries between 1950 and 2011. Going back in time was the task undertaken by Maddison (2001). Based on a broad set of historical studies, he reconstructed income per capita data over the past two centuries, and added some point estimates for earlier periods (in CE 1, CE 1000, CE 1500, CE 1600, and CE 1700). Such estimates very often require educated guesses on unobservable trends, but they have the great merit of showing the best, which can be done given what is known at one point in time. Very recently, Maddison's successors have been revising and completing his work (the 'Maddison project,' Bolt and van Zanden, 2013). **Figure 2** presents the latest estimates for GDP per capita.

Figure 2 shows that, over the past millennium, income per capita in our selected countries has increased 32-fold, from \$717 per person per year around the year 1000 to \$23 086 nowadays. This contrasts sharply with the previous millennia, when there was almost no advance in income per capita. The figure shows that it started rising and accelerating around the

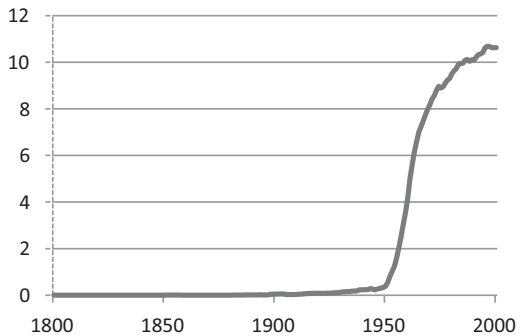

Figure 1 Occurrence of 'economic growth' in books by publication year.

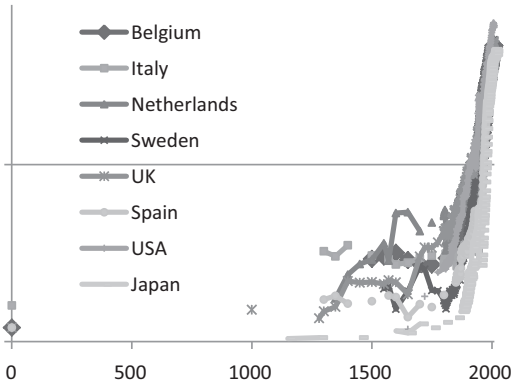

Figure 2 GDP per capita in selected countries, CE 1–2010. Logarithmic scale. Horizontal line = 5000 dollars (1990 GK\$). GDP, Gross Domestic Product.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71057-9

year 1820 and it has sustained a steady rate of increase over the last two centuries. One of the main challenges for growth theory is to understand this transition from stagnation to growth and in particular to identify the main factor(s) that triggered the takeoff.

Before looking into possible explanations, it is worth wondering how robust the finding that there was stagnation in the standard of living until 1820 is. This claim is particularly striking given that mankind experienced significant technological improvements that can be expected to increase productivity and income per person, from the Neolithic revolution to the invention of the printing press. Three facts corroborate the idea that there was indeed stagnation over the most part of human history: first, estimates of life expectancy computed on specific groups across time and space do not display any trend before CE 1700 (de la Croix and Licandro, 2013). Second, body height computed from skeletal remains does not display any trend either, while height is known to depend very much on nutrition when young (Koepke and Baten, 2005). This indicates that there was no systematic improvement in nutrition over time. Third, real wages computed from historical sources did not tend to rise in any sustained way before the industrial revolution (Allen, 2001).

Theories of Stagnation

How can it be that technical progress did not lead to a better standard of living before the last few centuries? The answer can be provided by mobilizing the so-called 'Malthusian theory,' attributed to Malthus (1798), inspired by previous work such as the chapter on the obstacles which oppose the excess of population in the human species by Bruckner (1767). This theory was formalized recently by Ashraf and Galor (2011). The theory has two basic ingredients. The first one is the production technology. It assumes that farms produce output from land and labor, where land is a fixed factor, and labor displays diminishing marginal returns. Diminishing returns imply that adding more labor while holding land constant yields lower productivity per person. The second ingredient is the assumption that fertility net of mortality is an increasing function of food per person. Less food per person yields higher mortality ('positive checks' in Malthus' terminology) and lower fertility (people marry later and have fewer children, 'preventive check' for Malthus). When there is an increase in technology, for example, when metal tools were invented, the generation at the time of the invention is better-off. When there is more food per person, fertility increases. The next generation is more populous, which lowers each worker's productivity because of diminishing marginal returns (the fields are crowded with laborers). As time passes, the population converges to a new steady state, and the food per person goes back to its initial level: the increase in productivity made possible by the discovery of metal tools is exactly offset by the decrease in productivity generated by the rise in population. In sum, technological improvements lead to a larger population, but not to a higher standard of living.

Before describing how economies can escape this logic, let us contrast the Malthusian model with the theories accounting for growth in the recent period.

Theories of Growth: Neoclassical Growth

The rise in per capita income did not happen everywhere in the world. The gap between the leading regions and the poor ones has actually increased over time. Citizens from Western Europe were three times richer than African citizens around the year 1820: now they are 13 times richer. To understand why countries differ so dramatically in terms of standard of living, the factors behind the success or the failure of the growth process in different parts of the world must be identified.

The first possible factor in the growth process is the accumulation of physical capital. For capital to drive growth steadily, the output of an economy needs to be proportional to the stock of capital used in production. In this case, growth will be proportional to investment. This accumulation through investment may come about either through local saving or investment from abroad. A theoretical argument against this view claims that marginal returns to capital are decreasing, i.e., one cannot increase production per worker indefinitely simply by increasing the stock of capital per worker. If machines per employed person grow at a constant rate, the growth of output will eventually drop to zero. Workers, whose number is bounded by the active population, cannot deal with many machines with the same efficiency as they cope with fewer machines.

While decreasing marginal returns on capital may make it impossible to sustain growth by investment alone, it remains true that capital is one of the most important determinants of growth in the medium run. McGrattan (1998) finds a strong positive relationship between average rates of investment and growth in postwar data for a large cross-section of countries. The typical example of such a relationship is the emergence of the East Asian tigers, which managed to catch up with rich countries in the second half of the twentieth century essentially by accumulating capital. However, Zambia's experience provides a counterexample. If all the aid to that country had been invested, and if investment had gone into growth, Zambia's income per capita would actually be $20 000 instead of $600 (Easterly, 2001).

Since physical capital cannot be the engine of growth when marginal returns are decreasing, another variable must account for rising per capita income over time. Neoclassical growth theory, pioneered by Solow (1956), assumes that technological change is exogenous to the economic system. In his model, technological change is the main source of growth by making labor more efficient, i.e., better able to work with many machines. Such technological change is labeled 'labor-saving technical progress.' In Solow's model, population growth is exogenous and does not respond to an increase in standard of living. In addition, there is no fixed factor, such as land, in the production process. These two features explain why technological progress drives growth in the Solow model, but not in the Malthus model.

Neoclassical growth, with its decreasing return to capital and exogenous technical progress, has strong empirical implications. First, it implies that there should be a convergence in income per capita across countries, i.e., countries that are further away from their long-run balanced growth path should grow faster than countries, which have already accumulated most of their capital. This is because the benefits

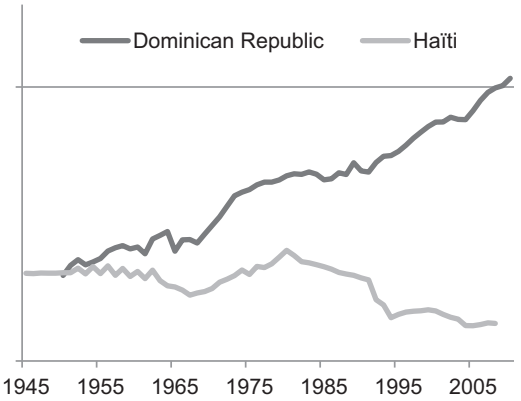

Figure 3 GDP per capita in Haiti and Dominican Republic. Logarithmic scale. Horizontal line = 5000 dollars (1990 GK$). GDP, Gross Domestic Product.

from increasing capital (i.e., the marginal productivity of capital) are higher in countries with little capital, and lower in countries with a lot of capital. According to this theory, investment should flow from rich countries to poor countries in search for better returns, and the income gap between countries should shrink over time. There is a vast empirical literature studying the existence and the speed of such cross-country convergence (see Durlauf and Quah, 1999). At the end of the 1990s, the consensus was that there is some convergence in the data, but at a much slower pace than predicted by the theory. Moreover, some groups of countries seem to be trapped on a low-growth path. This conclusion is now partly reversed, with the fast growth experience of East Asia, and inequalities across countries are clearly on the decline (van Zanden et al., 2014).

A second implication of neoclassical growth is that policy affects the level of income per capita, but has no grasp on its long-run growth rate. Bad governance cannot affect the growth rate of the economy, as the latter is determined by the growth rate of exogenous technology. Comparing the growth experience of countries with similarities (suspected of having the same growth rate of exogenous technology), this prediction of the theory seems at odds with the data – as exemplified by the experience of Haiti and the Dominican Republic. The two countries share the same island, had the same GDP per capita in 1950 (1027 dollars per person for the Dominican Republic and 1051 for Haiti), but their growth trajectory diverged, as shown in **Figure 3**. This divergence is likely to be attributed to better governance in the Dominican Republic. If this divergence of the two countries is considered as a long-run phenomenon, one should conclude that policy really matters for the long-run growth rate, which calls for developing another theoretical framework. This is what endogenous growth models are about.

Theories of Growth: Endogenous Growth

Modeling growth itself as an endogenous phenomenon requires getting rid of the decreasing returns to factors of production that the economy can accumulate (physical or human capital). Externalities (spillovers) play a central role here. If, when a firm or an individual invests, this investment has positive spillovers on other firms or individuals, the private return of the investment (to the firm or the individual) is smaller than the social return (to all firms or individuals together). Investments may thus still have decreasing private returns, but constant social returns thanks to spillovers. For example, investments in research leading to advances in knowledge benefit the firm, which has invested in them, but with decreasing returns (otherwise the firm would tend to invest infinitely). Such advances, however, benefit the whole society at a higher, and nondecreasing, rate.

Accordingly, endogenous growth can be based on the fact that advances in knowledge, e.g., in basic science, have led to technological progress. Growth theorists have pointed to the importance of what they have termed a research and development sector, which either invents new goods or increases the quality of existing goods (see Barro and Sala-I-Martin, 1995). This theory attributes a preeminent role to the R&D sector in generating growth, and makes a crucial link between innovation and market power. Indeed, the incentive to innovate is provided by monopoly rights on the new goods, which grant extra profits to research firms. Such models also open up interesting policy debates, essentially on the optimal patent protection laws, and the optimal level of subsidies required to stimulate research.

Accumulating human capital is the other way to sustain a growth process. Indeed, even if a country cannot increase its supply of labor indefinitely, it can enhance the quality of labor. This quality is referred to as 'human capital,' and incorporates education, experience, and health. Human capital can be accumulated in two ways: during the early stage of life by going to school and spending on education; during the working life by accumulating experience (learning-by-doing) or by on-the-job training. Investment in formal education has increased substantially all over the world. In England, for example, the number of years spent at school went from two on average in 1820 to more than 14 today. In developing countries, school enrollment grew rapidly between 1960 and 2000. Although many poor countries experienced an educational boom, they did not start growing more quickly. In general, empirical studies stress the lack of correlation between educational attainment and growth performance. The accumulation of human capital is a necessary condition for growth, but it does not seem to be a sufficient condition.

Institutions are often blamed or praised for their role in promoting growth. In endogenous growth models, good or bad policy may indeed permanently affect the growth rate of the economy itself. Acemoglu et al. (2002) argue that relatively rich countries colonized by European powers in the 1500s are now relatively poor and vice versa. They explain this 'reversal of fortune' by the types of institutions imposed by European settlers. 'Extractive' institutions were introduced in the relatively rich countries for the benefit of the settlers rather than to increase general prosperity. In the relatively poor areas, there were fewer incentives to plunder, and so to prevent the development of investment-friendly institutions. As a result, the decline or rise of those countries is rooted in a major – exogenous – institutional change linked to colonization. A number of authors have also linked the poor performance of

many developing countries with governance issues: corruption, ethnic fragmentation, wars, etc.

Theories of Growth: Poverty Traps

An alternative explanation to the lack of convergence across countries relies on poverty trap models. They are based on the idea that the marginal return of capital is not systematically decreasing: for some range of values of capital, there can be increasing marginal returns. For example, suppose that two technologies of production exist, the traditional one and the modern one. Switching to the modern technology allows for a discrete jump in the productivity of factors of production, but requires accumulating a certain level of capital per worker. If an economy has a capital stock close to this level, a small increase in capital will lead to a large jump in productivity, hence displaying increasing returns around the critical level of capital. A similar reasoning can be applied to human capital as well. Such a framework leads to predictions that differ drastically from the neoclassical growth model. Instead of a convergence of all countries to a similar long-run balanced growth path, now the long-run outcome depends on the initial conditions. An initially poor country with a low level of capital per worker might be 'trapped' in a low equilibrium, while an initially richer country may grow toward the high-balanced growth path. For the same reason, countries that are very similar initially may have a diverging growth experience, if they start on each side of a critical threshold. With poverty trap models, Figure 3 can be interpreted as reflecting that Haiti had slightly worse initial conditions, and this led to a drastically different development. Here, it is not governance, which is to be blamed, but bad luck.

Endogenous growth models and poverty trap models also differ drastically in their policy prescriptions. While endogenous growth leads to advise countries to improve governance and productivity, poverty trap models require some aid to be given to the country, to help it bypass the critical level of accumulation it needs to grow in a sustained way.

Poverty trap models essentially capture vicious cycles. In the example above, low development prevents the country from adopting modern technology, which perpetuates low development. Other mechanisms can be invoked: because government employees are underpaid, they ask for bribes, which discourages investment, keeping the country in poverty. Alternatively, a low level of human capital implies bad teachers, which perpetuates broken education systems and low human capital.

Poverty trap models are not only able to account for successful and unsuccessful growth experiences, but they can also explain why some economies have experienced relative, or sometimes absolute, periods of decline. Kindleberger (1996) defends the view that economies, like human bodies, go through a life cycle, in the course of which 'vitality' varies, depending on how entrepreneurial each generation is. The Venetian Maritime Republic declined when leaders started devoting more time to consumption rather than to investing in the improvement of shipping techniques, the discovery of new maritime routes, etc. Florence and its famous banking sector declined when Lorenzo the Magnificent delegated power over the Medici bank branches as he was spending more time on luxury and pleasure. In order to capture the idea that rich regions tend to rest on their laurels, and are inclined to favor consumption over investment in knowledge, habit consumption models can be used (see Artige et al., 2004). This approach suggests that the desire to consume depends positively on past consumption (either of the individual, or of his or her parents, or of society as a whole). Eventually, the new generation in the richer regions develops living standards that are incompatible with the continuing investment in knowledge that is required to maintain economic leadership. This reduces the growth rate in comparison with other regions. At some point in this process of decline, consumption falls to a low level once again, and a new growth cycle may start.

The Transition to Modern Growth

The models of neoclassical and endogenous growth describe very aptly what happened in the last two centuries, but they are totally unable to explain the stagnation period. The Malthusian model, on the contrary, gives a very nice description of stagnation, but cannot account for growth takeoffs. Finding a single framework, which can explain the whole path of income over history has been the task of growth researchers in that last 15 years. Explanations can be divided into two broad classes. The first class relies on exogenous causes; the second one builds a setup in which everything is endogenous.

Before turning to the different types of setups that have emerged to explain the full pattern from stagnation to growth, it is necessary to stress that the growth takeoff was systematically accompanied by a demographic transition. Figure 4 shows the demographic transition in Sweden, together with the growth takeoff. Birth and death rates were systematically high before the takeoff, then mortality went down, while fertility stayed high for some time. During this period, population size increased fast. Then fertility dropped quickly and population growth decelerated. Once the transition was

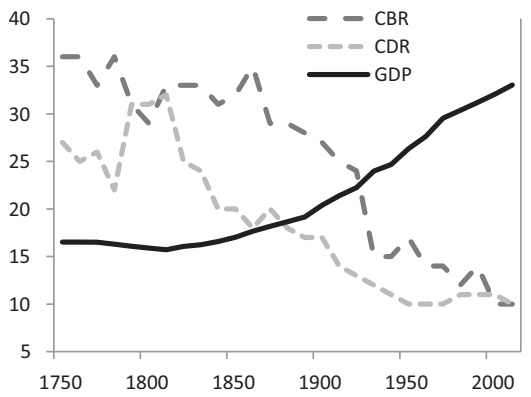

Figure 4 Left axis: crude birth rate (dark gray dashes) and crude death rate (light gray dots) in Sweden in per 1000. Right axis: GDP per capita. Logarithmic scale (1990 GK$). GDP, Gross Domestic Product; CBR, crude birth rate; CDR, crude death rate.

achieved, both birth rate and death rate were low, and population growth became negligible (or even negative in some cases). The demographic transition took two centuries in England and other European countries. In East Asia, it was much faster, of the order of half a century. Causality between the two transitions goes both ways: the demographic transition was triggered by the income takeoff, at least as far as the drop in mortality is concerned, while the takeoff was made possible by the demographic transition, because allowed to invest in the education and health of a smaller number of children.

Exogenous Models of the Transition

To return to the inability of the stagnation model and/or the neoclassical growth model to account for the whole historical path of income, Hansen and Prescott (2002) show that the two approaches can be combined in a single framework, which models the takeoff. The idea is as follows. There are two sectors of production, agriculture, and industry. Initially, the technological level and productivity in the industry are very low, so that the resources of the economy, including labor and capital, are all employed in the agricultural sector. As this sector is subject to decreasing returns to labor and to the constraint on available land, the model behaves as the Malthusian model. Then, with exogenous technological improvements, the industry sector becomes profitable and starts attracting resources. As time passes, resources move from agriculture to industry. The economy becomes more and more similar to the one described in Solow's neoclassical growth model.

In this theory, the process forcing the transition is exogenous. If interpreted literally, it consists of improvements in the productivity of the industry. In a broader sense, it can also reflect (exogenous) improvements in institutions. The Enlightenment can be seen as one of those changes, which made knowledge more useful for growth (see Mokyr, 2012), through the primacy they gave to reason over superstition and tyranny. The Enlightenment itself was rooted in the Protestant Reformation, which opened the path to questioning received authority, and hence, to critical scientific thinking.

Other types of exogenous explanations for the takeoff rely on incentives to accumulate human capital. If, for some reason, it becomes profitable for households to invest in education, growth can be sustained thanks to the human capital externalities stressed in the endogenous growth model literature. Two types of shocks are considered here. The first is based on Ben-Porath (1967)'s idea that the return to investment in education depends on the length of time during which education will be productive, implying that a longer active life makes the initial investment in human capital more profitable. Provided that human capital is an engine of growth, an initial increase in longevity (for exogenous biological/climatic/medical reasons) may, in turn, sustain the permanent accumulation of human capital and income growth (see, e.g., Cervellati and Sunde, 2005). The second type of shock triggering education can be institutional. For example, Engerman and Sokoloff (2002) identify conditions under which a country will introduce public education early, favoring a skilled workforce and a rapid industrial revolution.

Finally, the quest for deeply exogenous mechanisms to explain growth and development has led several authors to look at the prehistorical determinants of modern growth patterns. The most well-known approach is the one put forward by Diamond (1999) for whom Europe and Asia benefitted from two geographical advantages compared to other continents: the natural availability of domesticable plants and animals (among the ancestors of the 14 big domestic animals, 13 come from Eurasia and North Africa, with the 4 most important ones coming from the Fertile Crescent), and the East–West orientation of the continent, which favored the spread of knowledge and diseases (hence allowing immunization).

Unified Growth Theory

Unified growth theory, a term coined by Galor (2011) and his coauthors, shares the same ambition to explain income history in a single setup. Contrary to the previous approach, it does not rely on exogenous changes in technology or institutions, but embeds the mechanisms of change into the model itself. In the early stage, the economy is characterized by a stable Malthusian equilibrium. Population increases slowly as a consequence of technological progress. The slow increase in population density accelerates the pace of technological progress, as a denser population generates more ideas, allows for more specialization, and sustains bigger cities. Faster technical progress itself requires education to be implemented, and this rise in education leads the economy to accumulate human capital. As in endogenous growth models, this allows for sustained economic growth.

In the model above, population size plays a key role. During the stagnation period, all variables (income per person, etc.) are constant except for population size. At some point in time, it reaches a threshold above which the transition starts. Instead of basing the takeoff in income on the latent increase in population density, one can model the industrial revolution as resulting from a change in the composition of the population. Suppose that there are two types of people, and that one of the types, when present in large numbers in the population, generates the takeoff in income. If, during the stagnation period, this type benefits from an evolutionary advantage (which means it reproduces faster), then its share in the population rises slowly over time, and the takeoff ultimately occurs. This 'growth enhancing type' can be one, which values the education of children (Galor, 2011), or one with an entrepreneurial spirit (Clark, 2007). The question of the transmission of the type from parents to children, either genetically, or culturally, remains open. For this mechanism to work, a strong persistence of type across generations is indeed required.

The Quality–Quantity Trade-Off

Galor (2011) and many other authors have stressed that the link between the demographic transition and the economic

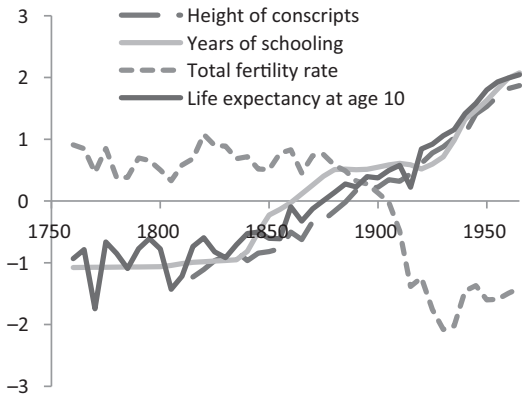

Figure 5 The quantity–quality trade-off for 1760–1965 birth cohorts in Sweden. Three measures of quality (body height, years of schooling, life expectancy at age 10), versus total fertility rate. All data normalized.

growth takeoff goes through the 'quality–quantity trade-off.' This originates in a simple budget constraint, which holds both at the individual level and at the country level:

$$\text{Total spending on children} = \text{number of children} \times \text{spending per child.}$$

Keeping the total spending constant, enhancing the 'quality' of children by spending more on each child requires reducing their number. Consequently, the drop in fertility observed during the demographic transition allows to increase education and health spending, thereby making the accumulation of human capital from one generation to the next easier. Figure 5 shows for Sweden that the drop in fertility comes with a sharp increase in formal education, with a rise in adult longevity, and with a rise in people's height, which signals better nutrition and a lower exposition to disease when young (see de la Croix and Licandro, 2013). Both education and health are key components of human capital.

Demographics also matter for growth through the 'demographic dividend' (see Lee and Mason, 2006). The view promoted by neoclassical growth was that a high rate of population growth could not be supported by a corresponding increase in investment, thus lowering growth of income per capita. This negative relationship between population and economic growth calls for population control policies. More recent empirical literature finds that in general there is a nonsignificant correlation in cross-country studies. From these studies, it is also clear that the impact of population growth has changed over time and varies with the level of development. The composition of the population, as well as age-specific variables, is relevant to growth. A decrease in the death rate of workers in particular does not have the same effect as a decrease in the death rate of dependents, young or old (Lindh and Malmberg, 2007).

Growth and Inequality

Finally, the relationship between growth and inequality has drawn attention. The question is twofold. First, does inequality evolve along with economic growth? The 'Kuznet curve' suggests that inequality first increases and then decreases in the course of the growth process. Second, is inequality good or bad for growth? Most of the existing literature on inequality and growth concentrates on the channels by which inequality affects growth, through the accumulation of physical capital. For example, in countries with many poor people, there is a high demand for redistribution policies, which in turn lead to tax distortions that slow down growth. Inequality can also influence growth because of its effects on the accumulation of human capital, in particular if the poor are subject to credit constraints, preventing them from investing enough in education. Finally, demographic variables must be taken into account in assessing the effects that economic growth may have on income distribution (differential fertility and mortality).

See also: Development Theory in Geography; Economic Growth: Measurement; Economics, History of; Wealth Distribution.

Bibliography

Acemoglu, D., Johnson, S., Robinson, J., 2002. Reversal of fortune: geography and institutions in the making of the modern world income distribution. Quarterly Journal of Economics 117, 1231–1294.
Allen, R., 2001. The great divergence in European wages and prices from the middle ages to the first world war. Explorations in Economic History 38, 411–447.
Artige, L., Camacho, C., De la Croix, D., 2004. Wealth breeds decline: reversals of leadership and consumption habits. Journal of Economic Growth 9, 423–449.
Ashraf, Q., Galor, O., 2011. Dynamics and stagnation in the Malthusian Epoch. American Economic Review 101, 2003–2041.
Barro, R., Sala-I-Martin, X., 1995. Economic Growth. New York.
Ben-Porath, Y., 1967. The production of human capital and the life-cycle of earnings. Journal of Political Economy 75 (4), 352–365.
Bolt, J., van Zanden, J.L., 2013. The First Update of the Maddison Project; Re-estimating Growth Before 1820. Maddison Project Working Paper 4.
Bruckner, J., 1767. Théorie du système animal. Jean Luzac, Leyde.
Cervellati, M., Sunde, U., 2005. Human capital formation, life expectancy and the process of development. American Economic Review 95, 1653–1672.
Clark, G., 2007. A Farewell to Alms. Princeton University Press.
de la Croix, D., 2012. Fertility, Education, Growth and Sustainability. Cambridge University Press.
de la Croix, D., Licandro, O., 2012. The Longevity of Famous People from Hammurabi to Einstein. CORE Discussion Paper 2012/52.
de la Croix, D., Licandro, O., 2013. The child is father of the man – implications for the demographic transition. The Economic Journal 123, 236–261.
Diamond, J., 1999. Guns, Germs, and Steel: The Fates of Human Societies. W.W. Norton & Co, New York.
Durlauf, S., Quah, D., 1999. The new empirics of economic growth. In: Taylor, John B., Woodford, Michael (Eds.), Handbook of Macroeconomics, vol. 1. Elsevier, pp. 235–308.
Easterly, W., 2001. The Elusive Quest for Growth, Economists' Adventures and Misadventures in the Tropics. MIT Press, Cambridge.
Engerman, S., Sokoloff, K., 2002. Factor Endowments, Inequality, and Paths of Development Among New World Economies. NBER Working Paper No. 9259.
Feenstra, R., Inklaar, R., Timmer, M., 2013. The Next Generation of the Penn World Table. Available for download at: www.ggdc.net/pwt.
Galor, O., 2011. Unified Growth Theory. Princeton University Press.
Hansen, G., Prescott, E., 2002. Malthus to solow. American Economic Review 92, 1205–1217.
Kindleberger, C., 1996. World Economic Primacy 1500–1990. Oxford.

Koepke, N., Baten, J., 2005. The biological standard of living in Europe during the last two millennia. European Review of Economic History 9, 61–95.
Lee, R., Mason, A., 2006. What is the demographic dividend. Finance and Development 43 (3), 16–17.
Lindh, T., Malmberg, B., 2007. Demographically based global income forecasts up to the year 2050. International Journal of Forecasting 23, 553–567.
Maddison, A., 2001. The World Economy, a Millennial Perspective. Paris.
Malthus, T., 1798. An Essay on the Principle of Population. Johnson, London.
McGrattan, E., 1998. A Defense of AK Growth Models, Federal Reserve Bank of Minneapolis Quarterly Review. Fall 1998.
Mokyr, J., 2012. The Enlightened Economy. An Economic History of Britain 1700–1850. Yale University Press, New Haven.
Nordhaus, William, 1996. Do real-output and real-wage measures capture reality? The history of lighting suggests not. In: Bresnahan, T., Gordon, R. (Eds.), The Economics of New Goods. University of Chicago Press, pp. 27–70.
Solow, R., 1956. A contribution to the theory of economic growth. The Quarterly Journal of Economics 70, 65–94.
van Zanden, J., Baten, J., Foldvari, P., van Leeuwen, B., 2014. The changing shape of global inequality 1820–2000; exploring a new Dataset, Review of Income and Wealth, 60, 279–297.

Economic Growth: Measurement

Muhammad H Rahman and Mehmet Ali Ulubaşoğlu, Deakin University, Burwood, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Quah, volume 6, pp. 4093–4098, © 2001, Elsevier Ltd.

Abstract

Measuring economic performance of a country is – arguably – one of the most important contributions of empirical economics. Historically, the national income accounting system was designed for maximizing the government earnings by finding better ways of imposing higher taxes. In contemporary economies, economic growth and their estimation approaches have become instrumental in devising social welfare maximizing policies in a more systemic way. This article embodies most of the fundamental measuring techniques of national accounting system in retrospect, and indicates several secondary data sources that are widely used in growth empirics.

The history of economic growth measurement has been one of intimate interplay with economic theorizing, leading from improvements in the accuracy and relevance of growth data, to hypotheses on the critical sources of economic growth, to explaining deep changes in the structure of national economies. In the late sixteenth and seventeenth centuries, the national income accounting system was first introduced in England and France to support their sovereign regimes in maximizing the government earnings by finding better ways of imposing higher taxes. However their national accounting systems were very far from to provide reasonable estimates until the early twentieth century. In particular, when the great depression struck most of the rich economies in the late 1930s, the governments became objective in measuring the pulse of the economy in a more systemic way.

Economic growth is the sustained increase in welfare of an economy – nation, region, city – together with the ongoing changes in that economy's industrial structure; public health, literacy, and demography; and distribution of income. In the long run, as this economic transformation evolves, so do social, political, and cultural norms. Societies change profoundly and multidimensionally, as economic performance improves.

To measure economic growth is to quantify this increase in welfare and to endow with numerical precision these large-scale economic and social changes. Given the breadth of possibilities, it is impossible to undertake this measurement exercise without guidance – what can be pared away, what is essential – from some view on the causes of growth (*see* Economic Growth).

This article sets down some key (measurement) facts concerning economic growth, and documents how they have evolved, if at all, over time. In doing this, the article attempts also to illustrate the historical interplay between two lines of research – measurement of and theories about economic growth – each influencing the other.

National Income

The panorama above of profound social and economic changes can be simplified dramatically by concentrating on just a single key economic variable, national income. National income can be estimated mainly using three methods, of which all theoretically provide the same estimate: by summing up all *expenditures* within the country in a given year, by adding up all *income* earnings each year, or by adding up all the *value added* in a year. In sum, national income is the total value of all goods and services produced in an economy in a given year. The *economic growth rate* is the rate of change in national income from 1 year to the next. Income can similarly be estimated for more disaggregated geographic units, such as regions, states, and provinces.

Income per capita is income divided by population in the appropriate geographic unit, yielding a convenient first measure on the state of economic well-being. Since total income is the same as total output, this measure might sometimes be usefully replaced by output per worker, or labor productivity, where the denominator is then the size of the labor force; or, even output per worker-hour, where the measure then takes into account the time spent working by the labor force.

In some detailed analyses, these alternatives can provide different useful insights into economic performance – different countries, at different times, have had their labor force markedly different from their population, or have had workers and firms make different choices on the length of their workday. However, for the kind of long-horizon, large-scale developments that are typically of interest in economic growth, these differences are inessential. Potentially more important is whether this one measure can suitably proxy for the wide spectrum of different variables of concern in economic growth.

Across countries, per capita income is positively correlated with a broad range of alternative indicators for economic performance, as well as an array of health, educational, and institutional quality measures – including life expectancy (the negative of), infant mortality, adult literacy, rule of law, democracy, and so on. Thus, economists and many other social scientists have focused on quantifying the relationship between per capita income (both its level and growth rate) and indicators of living standards.

Historical Estimates

Economic growth – significant, sustained increases in welfare – is a relatively recent occurrence (Maddison, 1982; Romer, 1986). While steady improvements in the lot of humanity over the last 50 000 years had been noted, most of those came tortuously slowly.

In eighteenth century the Netherlands appeared as the leading economy in the world. However, economic growth there was neither sustained nor pronounced. Almost a century afterward, labor productivity in the Netherlands was 6% lower than that in 1700. In 1785 the UK had risen to the leadership of the New World economy. Despite immense strides in technological progress due to the Industrial Revolution, the UK's labor productivity grew, nevertheless, by only 0.5% per year from 1785 through 1820. From then to the beginning of the twentieth century, the UK remained the world's economic powerhouse. But in this time, too, the UK labor productivity grew relatively modestly, by 1.4% per year. Only after the US overtook the UK following the end of the nineteenth century, did the world begin to see the kind of long-run, sustained growth rates that are now typical: US labor productivity grew at 2.3% from 1890 through 1979.

In general, historical data on income before the nineteenth century were hardly available. In this context, Maddison (1995) estimated gross domestic product (GDP) for 56 countries for the period 1820 through 1992. Later, he proposed an estimate for world output in the year AD 1, which was approximately $105.4 billion at 1990 prices. Importantly, Maddison's estimates indicate that Asia produced more than half of world output for 18 of the last 20 centuries (*see* The Economist, 2010). His readings suggest that the growing convergence in the world economy is – therefore – a 'restoration' not a 'revolution.' Nonetheless, Maddison's national accounting has to be taken with caution. For instance, despite using Maddison's data in their research, Acemoglu et al. (2005) remained skeptical stating that Maddison's data set is *no more than educated guesses*. Recently, Bolt and van Zanden (2013) launched an updated version of the original Maddison data set revising the estimates of economic growth between AD 1 and 2010.

Kaldor's Stylized Facts

Not long after the concept of national income was developed, Kaldor (1961) observed that the ongoing growth in per capita national income came with several empirical regularities:

1. Income per capita grows steadily over long stretches of time.
2. Factor income shares – the fractions of total income accruing to the labor force and to physical capital – show no secular trend.
3. The ratio of total national income to the aggregate stock of physical capital – the total value of machines, highways, bridges, buildings, physical infrastructure – remains roughly constant.

These observations, typically labeled *Kaldor's stylized facts*, constitute a first important growth measurement that economists sought to understand and to explain.

According to Kaldor's measurements, while economic growth occurs, its benefits divide up in a stable way between rewards to capital and labor – no one factor input, taken as an aggregate, benefited more than another from economic growth. Moreover, since capital's income share can be viewed as the stock of capital multiplied by capital's rate of profit, Kaldor's stylized facts also implied that the profit rate is, in the long run, constant.

Physical Capital Accumulation

The most successful of the different explanations for Kaldor's growth observations was given in Solow (1956) and Swan (1956), who developed similar explanations simultaneously, but independently in different hemispheres of the world. This immensely influential work produced a widespread conventional wisdom, known as neoclassical growth theory, which had also a profoundly important policy implication. The line of work, however, then led to yet further measurements (Solow, 1957) whose later reinterpretation would eventually overturn the original consensus.

The neoclassical growth theory had investment – the accumulation of physical capital – constituting the critical engine of growth. It assumed that total output obeyed a *production function* taking values that increased with the quantities of physical capital and labor employed in the economy. Because no well-behaved function can give at all times a perfect fit relating output, capital, and labor, it was hypothesized further an unobservable residual to the production function relation, a residual that he interpreted as exogenously evolving *technical progress* or, more simply, technology. Together with some other reasonable assumptions on the behavior of savings (and therefore investment), labor force growth, and technology's impact on the effectiveness of labor and capital, the model generated outcomes successfully reproducing Kaldor's stylized facts.

The result envisioned had economic growth in two conceptually distinct phases. First, there is the long-run steady state, where the capital–output ratio is constant, and growth in per capita income occurs entirely from technical progress. Second, there is transition to the steady state: along this transition path, physical capital accumulates rapidly due to relatively high investment, driving up the economy's growth rate as the economy itself converges toward its long-run, steady-state growth path. If the second phase were that more relevant to the real world, as many supposed, then accumulating physical capital – building machines, factories, roads, and bridges – constituted the engine of economic growth. Policy-makers from Joseph Stalin in the Soviet Union through Lee Kuan Yew in Singapore through leaders in developing country after developing country took this implicit advice to form a cornerstone of their development plans. Encouraging savings – voluntary or forced – became key to economic progress.

In neoclassical growth theory, technology bore two different meanings. One, because by assumption it influenced the effectiveness of factor inputs, technology was also labeled *total factor productivity* (TFP) – it improved the productivity of all the factor inputs in an economy. Two, because it emerged as a residual in the production function accounting exercise, it was

also viewed, derisively, as a 'measure of our ignorance' (on the part of the modeler). Early research on growth measurement, therefore, took one of its goals to be the minimization of this residual, the idea being that the smaller TFP became, the more successfully one had explained economic growth.

Measurements undertaken by Solow (1957) showed that for the US over the first half of the twentieth century, almost 90% of economic growth was due to TFP, with only the remaining one-tenth accounted for by capital accumulation. Taken at face value, this surprising empirical finding sharply contradicted the conventional wisdom and policy advice many had taken to follow from neoclassical growth theory. The hoped-for reconciliation was that by measuring accurately enough the factor inputs – accounting for improvement in quality of capital and labor – this 'Solow' residual would be diminished.

By contrast, more modern work favors the first interpretation of TFP, that is, it is correctly viewed as enhancing the productive potential of an economy. Thus, economies with large TFP are considered high-growth, high-performance economies. The research focus has shifted from trying to minimize measured TFP to trying to understand its evolution, typically in terms of its being driven by science and research and development (R&D).

Another important idea arose from this early work on measuring economic growth: convergence, or the tendency for economies not already on the steady-state path to approach it. Stated more fully, the neoclassical growth model implies that capitalist economies where capital and labor are rewarded their *marginal products* – the contributions each make, on the margin, to final output – are economies that are stable around their steady states. If an economy is not already at steady state, it has a tendency to go there. Economies relatively poor have a tendency to converge to, or catch up with, those relatively rich.

Penn World Table (PWT): Measuring the Income Variation *across* Countries

One of the primal challenges that growth empiricists encountered is the lack of homogenous data that are comparable across countries. Difficult as it initially was to compile systematic national income accounts for even the advanced countries, obtaining comprehensive and comparable data across countries was becoming more complex. This task has been accomplished only two decades ago (Summers and Heston, 1991) to a degree where researchers can meaningfully compare incomes and therefore economic growth across countries.

Because different countries collect and report income data in their individual, respective currencies, cross-country comparisons can be made only after converting national incomes to a common basis. To achieve this, one possibility is to use official current exchange rates to convert national currency figures into numbers denominated in some single currency. However, such exchange rates are highly volatile, and need not always accurately reflect the purchasing power, or the command over real resources, of the respective national currencies. Importantly, currency exchange rates cannot reflect the prices of nontradeable goods and services. To the extent that nontradeables are relatively lower priced in poorer countries, conversion using official exchange rates *underestimates* the incomes of relatively poor countries, and thus exacerbates the true gap between rich and poor.

As just one example, consider the following. Even though between 1985 and 1990, Asia was the world's fastest growing region, its share of world GDP – calculated using official exchange rates – fell from 8 to 7%. The reason for this is that Asian currencies depreciated sharply against the US dollar in this 5-year period. In this instance using exchange-rate adjusted national incomes disguised (and indeed misled on) actual changes in economic performance.

Summers and Heston (1991), in collaboration with the United Nations, constructed rescaled national incomes for a broad cross-section of countries. They used a *purchasing power parity* (PPP) correction to overcome the measurement difficulty just described. Taking into account prices on close to 700 individual items – traded and nontraded – across different countries, and then adjusting for differential spending patterns, the Summers–Heston data provide one of the most reliable indicators to date, allowing comparison of economic growth across countries. These data have now been used in over 1000 empirical studies on economic growth.

Allowing cross-country comparisons has opened up striking new questions in both the measurement and theory of economic growth. Using 1993 official exchange rates, Switzerland was the world's richest economy; Tanzania, the world's poorest. The ratio of per capita incomes between them approached an incredible 400. By contrast, when PPP-corrected and averaged across the last 5 years of the 1980s to smooth out short-run fluctuations, the ratio between richest and poorest economies (now the US and Chad, respectively) falls to a little over 40. This latter number accords better with less formal evidence and with casual observation on differing costs of living. However, even though PPP correction has reduced the dramatic disparity between rich and poor across the world, the divide obviously remains considerable.

The Summers–Heston data reveal the following: between 1960 and 1989 world per capital income grew by 2.25% per year. Averaged over the 5 years 1960–64, per capita income at the bottom 10th-percentile point (in the cross-section of countries) was 0.22 times the world average, with the 90th-percentile point 2.70 times the world average (Quah, 2001). The 90:10 ratio was, therefore, 12. Almost three decades later, averaged over 1985–89, per capita income at the bottom 10th-percentile point had fallen to 0.15 times the world average while the 90th-percentile point had risen to 3.08 times the world average. The 90:10 ratio thus nearly doubled over this period of high average growth worldwide. More detailed study of the growth dynamics in this post-World War II cross-section of countries has revealed, not just a widening of income disparities across rich and poor, but also an emerging polarization, with the middle-income group of countries dwindling in number (Quah, 1997a).

This finding, that in the post-War era, poor economies do not obviously catch up to the rich, has its interpretation disputed by research arguing that, provided one conditions on the appropriate indicator variables – investment and schooling, in particular – then convergence does occur. That kind of convergence, however, is no longer between rich and poor, but

is convergence of each economy toward its own specific steady-state growth path (i.e., conditional convergence). Hundreds of auxiliary variables have been suggested in these cross-country growth and convergence studies as appropriate alternative or additional conditioning variables (see, e.g., the compilation of studies in Durlauf and Quah, 1999). *Absolute* convergence (i.e., countries progressing to a single worldwide level of output) has been documented clearly only recently by Rodrik (2013), but only within the manufacturing sector.

Over time, the Summers and Heston (1991) data set has been updated and evolved into what is now called PWT. Despite its widespread usage, recently PWT has been subject to criticism. Studies by Johnson et al. (2009, 2013), Ciccone and Jarociński (2010), and Katayama and Ponomareva (2010) indicate that existing findings based on the earlier versions of the PWT in the literature significantly vary with the new versions of it. Hence, Feenstra et al. (2013a) revealed version 8.0 of the PWT that addressed the sensitivity to earlier versions. Version 8.0 of the PWT is – arguably – the best available cross-country data set providing a set of national accounting on relative levels of real GDP, capital, trade, and productivity covering 167 countries between 1950 and 2011. At least three improvements have been made in version 8.0: multiple PPP benchmarks, incorporating novel estimates of relative prices for exports and imports, finally new measures of capital and productivity (*see* Feenstra et al., 2013b).

Inequality: Measuring the *within* Country Variation of Income

As measuring economic growth has continued to focus on progressively fewer observable variables, an implicit tension has correspondingly sharpened: if welfare is proxied only by average income, then income inequality – or more generally, the distribution of income across people – might be unobserved worsening to where an average seriously misleads as an indicator for the economic welfare of the population overall.

The Kuznets (1955) curve describes a key empirical regularity between national income and inequality: Cross-sectionally across economies, inequality is low when per capita incomes are low, increases at moderately higher per capita income levels, and finally falls again at the very highest levels of per capita income. Put differently, the relation between inequality and levels of development is an inverted-U. This early finding suggested to many researchers the possibility that, indeed, economic growth might disadvantage the very poor, due to increases in inequality.

Reverse relationship from inequality to economic development (in particular, growth rate) rate has yet been another focus of interest. Researchers argued that inequality is taxing on economic growth due to redistribution it requires, borrowing constraints it imposes on individuals in the lower end of the distribution, sociopolitical instability it causes, and disincentives it generates. Empirical research found mixed evidence for this relationship, with negative association when variation is considered across countries and positive relationship within a country over time. Importantly, the income inequality data collated by Deininger and Squire (1996) included immensely useful information on income Gini coefficients. This data set was recently updated and standardized by Solt (2009).

Ideas – Science and Technology

Observing that country disparities are large and growing, that is, that convergence is not happening, and that so much of economic growth is due to TFP, has motivated some researchers to investigate models of *endogenous technological change* (One notable exception is Mankiw et al. (1992) who argued that taking into account *embodied human capital* – the quantity of schooling embodied in the labor force – suffices to explain both observations.).

In work on endogenous growth theory, TFP is no longer taken to be exogenously given. Instead, economic theory is used to shed light on the incentives and forces shaping TFP's evolution (Aghion and Howitt, 1998; Grossman and Helpman, 1991; Romer, 1990). The problem of explaining cross-country disparities then becomes one of understanding the dynamics of cross-country technology dissemination (Coe and Helpman, 1995; Helpman, 1993). This work predicts that knowledge inputs – scientists and engineers, R&D, patents, knowledge spillovers, ideas – drive economic growth. The key empirical findings are twofold: First, significant spillovers occur across countries, so that worldwide knowledge accumulation drives worldwide economic growth. Second, country-specific knowledge stocks, resulting from that country's own R&D, do matter importantly, and help explain each country's pattern of economic growth.

The hypothesis of knowledge-driven economic growth, while suggested by the logical sequence of earlier measurements and theoretical developments, remains controversial. For one, many observers consider it implausible that the poorest economies – those in sub-Saharan Africa, say, in the latter part of the twentieth century – are poor because they do not do enough scientific research. This, however, is an overly literal interpretation of knowledge inputs. The ability of people in different economies to absorb and economically exploit scientific developments or even just best-practice production techniques obviously differ. The question then becomes one of incentives to use ideas – why do the very poorest economies not use frontier technologies to raise the productivity of factor inputs? The economics of creating and disseminating ideas, or *intellectual property* more broadly, are subtle; so too, but much less studied, the economics of using ideas.

However, even in the richest economies, empirical evidence is ambiguous on knowledge-driven economic growth. US scientists and engineers employed in R&D grew fivefold from 200 000 in 1950 to over 1 million in 1990; the growth rate of US GDP failed to increase by anything remotely comparable over this period (Jones, 1995).

Structural Drift: Manufacturing to High Value-Added Services

The more recent focus on measuring economic growth using just changes in per capita income has de-emphasized the evolution in industrial structure of economies as growth

proceeds. By contrast, research such as that in Kuznets (1966) consistently documented such basic changes. Quah (1997b) has noted that such structural drift continues, away from manufacturing, toward high value-added intangible services.

Over the nineteenth century, as economies grew, their industrial composition shifted away from an emphasis on rural agriculture, and toward manufacturing and urbanization. Economic activity became progressively more markets-based, so that traditional informal, nonmarket work at home was replaced by formal employee arrangements. Education became more institutionalized; foreign commerce increased; and natural resources were relied on progressively less.

Geospatial Income Measures: Disaggregating the Aggregates

Most of the income measures are aggregated at the national level, and in consequence, spatially disaggregated data sets were either nonexistent or publicly unavailable until recently. Given the advent of geographic information system, macroeconomists start exploring new possibilities of constructing disaggregated national income measures. One of the successful spatially enabled platforms for economic data is the geographically based economic data *(G-Econ)*, a research initiative at Yale University that is devoted to developing a spatial data set on economic activity covering most of the countries in the world (*see* Nordhaus et al., 2006). In particular, the current data set (G-Econ 4.0) provides 'gross cell product (GCP)' for 1990, 1995, 2000, and 2005. GCP is measured at a 1-degree longitude by 1-degree latitude resolution at a global scale, which includes 27 500 terrestrial observations (i.e., cells). In addition to GCP, the G-Econ data set provides other key demographic and geophysical variables – e.g., climate, physical attributes, location indicators, and population.

More recently, Henderson et al. (2011) studied a satellite-based night-time data set covering 1992 to 2009, which is originally provided by the US National Oceanic and Atmospheric Administration's National Geophysical Data Center. They hypothesized that the brightness of visible lights that are laminated at night is strongly associated with both population density and income per capita. They explored the possibility to what extent lights data can augment or entirely replace data on GDP. They recognized several caveats in using lights to measure income. For instance, it is not obvious that economic activities can be completely traced by the amount of light emanating from the earth's surface. Lights are not laminated constantly across time and space. Furthermore, measuring the visibility of true light by the satellites remains imperfect because of climatic variation – e.g., humidity and reflectivity – across regions.

In minimizing the measurement error, Henderson et al. (2011) have taken a series of measures. For example, observations affected by forest fires and auroral activity between the northern and southern lights are removed both manually and using frequency filters. Such data-cleaning measures maximize the chance of capturing mostly man-made light, leaving all other natural lights out of inclusion into the data set.

This lights data set would have a very significant contribution in understanding the *within* country growth variation. The conventional data sources – e.g., PWT – do not provide any disaggregated level data set to measure economic activity at subnational or local level. As much of the significant variation in economic growth may take place within – rather than across – countries, the night lights data open a window of opportunity that allows the growth empiricists to set aside this limitation.

Economic Growth Measures: The Way Forward

Measuring economic growth is an intricate and subtle exercise. It has a scope that can be large and unmanageable. In practice, the history of measuring economic growth has been one of intimate interplay with economic theorizing, leading from improvements in the accuracy and relevance of growth data, to hypotheses on the critical sources of economic growth, to explaining deep changes in the structure of national economies.

TFP, previously a measurement residual to be minimized, has since been reinterpreted, and has emerged as the key driver for economic growth. Ideas – scientific and engineering knowledge, R&D, expertise – constitute the prime candidates for explaining TFP, but the supporting empirical evidence remains weak.

See also: Data Bases and Statistical Systems: Economics (Macroeconomics); Development: Social-Anthropological Aspects; Differentiation: Social; Economic History, Quantitative: United States; Macroeconomics, History of From 1933 to Present; Macroeconomics, History of up to 1933; Modernization, Sociological Theories of; Time Series: General.

Bibliography

Acemoglu, D., Johnson, S., Robinson, J., 2005. The rise of Europe: Atlantic trade, institutional change, and economic growth. American Economic Review, 546–579.
Aghion, P., Howitt, P., 1998. Endogenous Growth Theory. MIT Press, Cambridge, MA.
Barro, R.J., Sala-i-Martin, X., 1992. Convergence. Journal of Political Economy 100 (2), 223–251.
Barro, R.J., Sala-i-Martin, X., 1995. Economic Growth. McGraw-Hill, New York.
Bolt, J., van Zanden, J.L., 2013. The First Update of the Maddison Project; Re-estimating Growth before 1820. Maddison Project Working Paper 4.
Ciccone, A., Jarociński, M., 2010. Determinants of economic growth: will data tell? American Economic Journal: Macroeconomics 2 (4), 222–246.
Coe, D.T., Helpman, E., 1995. International R&D spillovers. European Economic Review 39 (5), 859–887.
Deininger, K., Squire, L., 1996. A new data set measuring income inequality. The World Bank Economic Review 10 (3), 565–591.
Durlauf, S.N., Johnson, P.A., 1995. Multiple regimes and cross-country growth behavior. Journal of Applied Econometrics 10 (4), 365–384.
Durlauf, S.N., Quah, D., 1999. The new empirics of economic growth. In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics, vol. 1A. North Holland, Amsterdam, pp. 231–304 (Elsevier Science Chapter 4).
The Economist, 1 May 2010. Finance and Economics: Maddison Counting; Economics Focus. The Economist Intelligence Unit, London. Retrieved from: http://ezproxy.deakin.edu.au/login?url=http://search.proquest.com/docview/223978960?accountid=10445.
Feenstra, R.C., Inklaar, R., Timmer, M.P., 2013a. The Next Generation of the Penn World Table available for download at: www.ggdc.net/pwt.
Feenstra, R.C., Inklaar, R., Timmer, M.P., 2 September 2013b. Recasting International Income Differences: The Next-generation Penn World Table. VoxEU.org.

Grossman, G.M., Helpman, E., 1991. Innovation and Growth in the Global Economy. MIT Press, Cambridge, MA.
Helpman, E., 1993. Innovation, imitation, and intellectual property rights. Econometrica 61 (6), 1247–1280.
Henderson, V., Storeygard, A., Weil, D.N., 2011. A bright idea for measuring economic growth. The American Economic Review 101 (3), 194–199.
Johnson, S., Larson, W., Papageorgiou, C., Subramanian, A., 7 December 2009. Is Newer Better? The Penn World Table Growth Estimates. VoxEU.org.
Johnson, S., Larson, W., Papageorgiou, C., Subramanian, A., 2013. Is newer better? Penn world table revisions and their impact on growth estimates. Journal of Monetary Economics 60 (2), 255–274.
Jones, C.I., 1995. R&D-based models of economic growth. Journal of Political Economy 103 (3), 759–784.
Jones, C.I., 1998. Introduction to Economic Growth. Norton, New York.
Kaldor, N., 1961. Capital accumulation and economic growth. In: Lutz, F.A., Hague, D.C. (Eds.), The Theory of Capital: Proceedings of a Conference Held by the International Economic Association. Macmillan, London, pp. 177–222 (Chapter 10).
Katayama, H., Ponomareva, N., 2010. Does the version of the Penn world tables matter? An analysis of the relationship between growth and volatility. Canadian Journal of Economics 43 (1), 152–179.
Kuznets, S., 1955. Economic growth and income inequality. American Economic Review 45 (1), 1–28.
Kuznets, S., 1966. Modern Economic Growth: Rate, Structure, and Spread. Yale University Press, New Haven, CT.
Levine, R., Renelt, D., 1992. A sensitivity analysis of cross-country growth regressions. American Economic Review 82 (4), 942–963.
Li, H., Squire, L., Zou, H., 1998. Explaining international and intertemporal variations in income inequality. Economic Journal 108, 1–18.
Maddison, A., 1982. Phases of Capitalist Development. Oxford University Press, New York.
Maddison, A., 1995. Monitoring the World Economy, 1820–l992. OECD Development Centre, Paris.
Maddison, A., 2001. The World Economy: A Millenial Perspective. Development Centre of the Organization for Economic Cooperation and Development.
Mankiw, N., Romer, D., Weil, D.N., 1992. A contribution to the empirics of economic growth. Quarterly Journal of Economics 107 (2), 407–437.
Nordhaus, W., Azam, Q., Corderi, D., Hood, K., Victor, N.M., Mohammed, M., Miltner, A., Weiss, J., 2006. The G-Econ Database on Gridded Output: Methods and Data. Yale University, New Haven.
Quah, D., 1997a. Empirics for growth and distribution: polarization, stratification, and convergence clubs. Journal of Economic Growth 2 (1), 27–59.
Quah, D., 1997b. Increasingly weightless economies. Bank of England Quarterly Bulletin 37 (1), 49–56.
Quah, D., 2001. Cross-country growth comparison: theory to empirics. In: Dreze, J. (Ed.), Advances in Macroeconomic Theory, Proceedings of the Twelfth World Congress of the International Economic Association, vol. 1. Macmillan, Buenos Aires, London, pp. 330–349. (Chapter 16).
Ray, D., 1998. Development Economics. Princeton University Press, Princeton, NJ.
Rodrik, D., 2013. Unconditional convergence in manufacturing. Quarterly Journal of Economics 128 (1), 165–204.
Romer, P.M., 1986. Increasing returns and long-run growth. Journal of Political Economy 94 (5), 1002–1037.
Romer, P.M., 1990. Endogenous technological change. Journal of Political Economy 98 (5, part 2), S71–S102.
Solow, R.M., 1956. A contribution to the theory of economic growth. Quarterly Journal of Economics 70 (1), 65–94.
Solow, R.M., 1957. Technical change and the aggregate production function. Review of Economics and Statistics 39 (3), 312–320.
Solt, F., 2009. Standardizing the world income inequality database. Social Science Quarterly 90 (2), 231–242. SWIID Version 3.1, December 2011.
Swan, T.W., 1956. Economic growth and capital accumulation. Economic Record 32 (2), 334–361.
Summers, R., Heston, A., 1991. The penn world table (Mark 5): an expanded set of international comparisons, 1950–1988. Quarterly Journal of Economics 106 (2), 327–368.

Economic History

Chris Godden, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E. Aerts, H. Van der Wee, volume 6, pp. 4102–4108, © 2001, Elsevier Ltd.

Abstract

Economic activity is human activity, and economic history is the historical study of the economic aspects of human existence. This article presents an overview of the aims and methods of economic history, discusses its position between the social sciences and the humanities, and highlights some current research agendas. The primary focus of the piece is on the historical and theoretical traditions of the discipline, and the lasting impact of methodological changes that occurred in the mid-twentieth century associated with the 'new economic history.'

Defining Economic History

Economic history is neither a branch of mainstream economics nor a branch of intellectual history relating to the history of economic thought. In basic terms, economic history (*histoire économique* in French, *Wirtschaftsgeschichte* in German, *jinji shi* in Chinese, *kaizái shi* in Japanese) is the historical study of the economic aspects of human existence. The broad intellectual range of the discipline enables it to extend beyond purely economic themes (such as the organization of production and distribution, the relation of state and industry, aspects of trade and commerce) and so consider, among other things, inequality, living standards, welfare, social transformation, science, technology, and education. With its focus on mankind's existence in a world of innovation, growth, and development, economic history is a discipline whose core principles possess a powerful human dimension.

This is a short article on a large topic relating to the aims and methods of economic history. Central to this discussion is the social scientific formalization of economic history in the mid-twentieth century. This marked the transition of the discipline from traditional economic history to the 'new' economic history, and so rooted the discipline in the methodology of economic science. In doing so, it ignited lasting 'creative tensions' between scholars trained in historical method and those trained in economic theory.

Origins and Development

The origins of economic history can be traced to the work of the early practitioners of political economy and social science: writers such as Smith, Marx, and Weber all employed historical evidence in their study of economic practices, social phenomenon, and the capitalist system. As a distinct academic discipline, the origins of economic history lie in late nineteenth-century German, British, and American scholarship and the construction of a methodology with strong links to established themes of general history (i.e., political, ecclesiastic, and legal histories), the work of the German historical school (dismissive of abstract economic theory and focused on the impact of variation in economic development), and the institutional economics associated with Thorstein Veblen. Its position as a subject of scholarly enquiry and academic teaching expanded steadily in the nineteenth century and was firmly consolidated by the early twentieth century with the creation of specific academic posts and later specialized journals – the *Economic History Review* (first published in Britain in 1927), the *Annales d'histoire économique at sociale* (first published in France in 1929), and the *Journal of Economic History* (first published in the United States in 1941). During this period, the interests of economic historians focused on comparative historical research to investigate the broad issues of economic life. Such studies incorporated some quantitative analysis, but little attention was given to employing explicit economic theory in analyzing past economic events or processes. As a consequence of these developments, economic history by the mid-twentieth century had developed as a discipline largely separate from economics.

The Emergence of the New Economic History

The key developments and refinements of economics theory that emerged by the end of the Second World War were fueled by new approaches to the study of the aggregate economy, growing interest to the mechanisms of economic growth, and the better availability of long-run economic data. By the 1950s, new attention to the subject of modern economic growth led many economists, particularly in the United States, to explore the long-held assumptions and conclusions associated with the older generation of economic historians. The conclusions of these younger, economics-trained scholars were damning: the quality of early economic history research had been undermined by largely descriptive, illustrative, and institutional approaches; quantitative analysis lacking rigor; and little attention had been paid to the application of economic theory to historical data.

It is here that we find the principles that propelled the cliometric revolution (from the Latin: *Clio*, the ancient Greek muse of history; *metrics* meaning measurement), so leading to the social scientific formalization of economic history between the 1950s and the 1970s. Cliometrics – otherwise known as the 'new economic history' or 'econometric history' – provided an innovative methodology that emphasized (1) a strong focus on the precision and formalism of scientific method in order to elevate the discipline of economic history to a scientific level,

and (2) the systematic application of mathematically formalized economic models and quantitative techniques to the study of historical economic phenomenon. Belief in the potential of these methodological developments, most notably in evaluating and challenging existing historical interpretations and scholarship, is palpable in the following quote by the American economic historian, Douglass C. North, published in the early 1960s:

> A revolution is taking place in economic history in the United States. It is being initiated by a new generation of economic historians who are both sceptical of traditional interpretations of U.S. economic history and convinced that a new economic history must be firmly grounded in sound statistical data. Even a cursory examination of accepted 'truths' of U.S. economic history suggests that many of them are inconsistent with elementary economic analysis and have never been subject to – and would not survive – testing with statistical data.
>
> North (1963): 128–129

The speedy acceptance of this new form of economic history among economists owed much to its focus on mathematical methods and rigorous quantitative techniques. Institutional structures reinforced the distinct character of this intellectual movement, famously through a special gathering held in Massachusetts in the autumn of 1957 and conferences and seminars at Purdue University in West Lafayette, Indiana. From these emerged the style, tone, and research agenda – primarily related to American economic history – that came to define the 'new economic history' for many years. Notable examples of this early work included Alfred H. Conrad and John R. Meyer's groundbreaking study of the profitability of slavery in the ante-bellum South (Conrad and Meyer, 1958), and Robert Fogel's (controversial) use of counterfactual analysis to explore the contribution made by the railways to American economic development in the nineteenth century (Fogel, 1964).

The overall impact of this new research agenda marked, as William Sewell Jr. has noted, "a move from economic history as the study of the forms of economic life to economic history as the historical study of the determinants of national economic growth" (Sewell, 2010: 152). It was now economic historians trained as economists, utilizing statistical techniques and mathematical models, who seized the commanding position in the discipline with their attention focused on investment, technological development, total factor productivity growth, historical national accounts, and long-term patterns of economic change.

It should come as no surprise that this new, quantitative-analytical study of economic history attracted numerous criticisms. By far, the most famous example of this came with the publication of Robert Fogel and Stanley Engerman's dramatic reinterpretation of the viability of American slavery in *Time on the Cross: The Economics of American Negro Slavery* (Fogel and Engerman, 1974).

The standard critique of the 'new economic history,' which emerged in the 1960s and continues to this day, concentrated on the degree of abstraction applied by cliomatricians and the construction of historical narratives that correspond to the conceptual framework of conventional (neoclassical) economics. Critics argued that this tendency to view history purely through the prism of abstract models raised numerous methodological questions, ranging from the ease by which investigators could select independent variables in order to condition the results of regression analysis through to the dangers inherent in counterfactual propositions.

A distinctive feature of the 'new economic history' was the employment of the counterfactual hypothesis method. The basic principle here was simple: in order to establish the impact of a particular factor, it was necessary to consider what the situation would have been in the absence of that factor. Such questions could be asked through the use of economics theory, with the specification and criteria of the economic model serving to construct counter-to-fact conditions. Yet the end result, in the eyes of many, was to produce mere 'fictions' rather than detailed historical interpretations. As an example of this, we may note Fritz Redlich's criticisms of Fogel's use of counterfactual analysis in exploring the impact of the railways on American economic growth (Redlich, 1965).

The Search for Balance

From the 1970s onward, criticism and dissatisfaction from within the ranks of scholars associated with the 'new economic history' began to suggest a new way forward. Here the criticism was that while of the systematic application of static neoclassical analysis had destroyed the explanations of an older generation of economic historians, it provided limited answers and failed to construct clearer explanations of long-run economic change. To quote Douglass C. North:

> It is the systematic use of standard neo-classical economic theory which has both provided the incisive new insights into man's economic past and also serves to limit the range of enquiry.
>
> North (1974): 1

There was now the realization for the need to provide a deeper examination of the structure of economies, the intricate interactions of economic and social forces, and the formal and informal rules that governed the ways people had lived their lives. In searching for balance, economic historians were therefore prompted to construct a broader analytical approach in order to evaluate the importance of factors such as entrepreneurship, cultural heritage, the evolution and impact of institutions, property rights, contract law, and government.

Thus, by the 1980s, it has been recognized that the study of economic history could not be practiced by rigid adherence to a methodological apparatus focused purely on quantitative techniques and standard economic analysis. Yet a more holistic form of the discipline – operating through a renewed framework that embedded economic development in the social, cultural, and political environment – clearly echoed themes that had been prominent within traditional economic history before being systematically ignored in the wake of the cliometric revolution. It was that move to enrich the study economic history from the 1980s onward, through the reintegration of a more traditional historical methodology, that has led one critic to identify the rise of cliometric economic

history as, ultimately, a "revolution that bit its own tail" (Drukker, 2006).

Historical Economics

The extreme application of the new methodology associated with the 'new economic history' led to the emergence of what may best be described as 'historical economics.' This is a variant of economic history as practiced by economic-trained economic historians whose objective is to use history to either illustrate theory or test hypothesis derived from theory. This fundamentalist form of cliometric/econometric history, with its research program derived from, and conforming to, the standards and practices of modern economics, suggests to its practitioners that the discipline continues to develop a rich understanding of modern economic growth. The reality, however, is that this formulation of economic history – with its focus on the utilization of historical time series as a means of testing sophisticated theoretical models – serves as a very poor basis on which to study the economic aspects of human existence. Indeed, 'historical economics' ultimately diminishes the richness of economic history by relegating it to little more than a branch of applied economics. Yet it must be recognized that 'historical economics,' although practiced by some scholars, is far from being the standard form of economic history that exists in the twentieth-first century.

The Future of Economic History

Where does economic history stand today? It is clear that, since the end of the Second World War, economic history has witnessed the dominance of economics-trained/based economic historians over history-trained/based economic historians. This is impossible to deny – the modern form of economic history is supported by the application of economic theory and quantitative techniques, and the majority of economic historians are to be found in leading PhD-producing economics departments. Although never achieving the same degree of dominance among European economic historians compared with their American counterparts, the successes of the 'new economic history' were formally recognized in 1993 with the awarding of the Nobel Memorial Prize in Economics to Robert Fogel and Douglass C. North.

It is certainly true the current status and character of economic history holds a very lowly position within the humanities. In briefly examining this issue we may note two key themes. The first concerns the activities of the early cliomatricians (in the 1960s and 1970s) who oversold their new methodology to historians by reinterpreting established interpretations through a blinding barrage of esoteric economic and quantitative methods. A second theme concerns recognition of important developments within the study of history since the 1980s, including the rejection of economic determinism associated with Marxist analysis, the compartmentalization of history into subdisciplines focusing on specific themes and case studies, and the impetus of the 'cultural turn' that emphasized literary and linguistic analysis over quantitative analysis. Thus, whereas economic history deals with the (dirty) world of trade, commerce, industrialization, and growth, the research agendas of social and cultural historians from the 1980s onward have increasingly focused on topics such as race, gender, identity, consciousness, and public memory. Both of the above themes – lasting reactions to the attitudes of the early practitioners of the 'new economic history,' and developments in the methodology of history since the 1980s – have made historians antipathetic to the study of economic history. From the perspective of the modern (cultural) historian, researchers in the area of economic history are easily dismissed for employing an inappropriate, narrow methodological framework that strips humanity from the historical narrative! Irrespective of the fact that economic history has moved beyond the strict methodological principle associated with the 'new economic history' (and so focusing on the relations between past economic events, institutions, social structures, and cultural change), the subject – indeed, it could be argued, anything connected with economic issues! – remains heavily tainted in the eyes of many historians.

Yet set within the context of the social sciences, economic history remains a rich and active field of investigation and it cannot be denied that the twenty-first century form of the discipline encompasses a variety of methodological approaches. While it is impossible to predict the future direction of research programs, it is evident that current practitioners are consciously working to strike a balance between formal theory (drawn from disciplines across the social sciences, and not just economics), quantitative methods, and historical techniques. The geographic range of current research now extends beyond the traditional areas of North America and Western Europe to include

- the history of economic development across medieval, early modern, and modern time periods;
- the history of debt, taxation, and credit relations;
- the causes and impact of financial crises within capitalist societies (clearly a growing area of interest since the global crisis of 2008);
- detailed case studies of different sectors and processes, including agriculture, demographic change, transport and communications, labor markets, education, and urbanization;
- the study of national economies/societies within the wider framework of global economic development and debates surrounding economic globalization;
- the Great Divergence of living standards between Asia and Western Europe (and, by extension, the emergence of Asian economic history as major area of research);
- quantitative analysis of the relationship between economic development, living standards, and biological well-being (otherwise known as anthropometric history); and
- debates connecting economic growth with environmental degradation (and hence the growing attention to environmental history).

Even the most cursory examination of these topics will show that the modern focus of economic history is definitely *not* concerned with the simplistic application of econometric methods to historical data!

It was noted in the introduction to this article that economic history possesses a powerful human dimension. Economic

activity is human activity, and it is this core principle that, as Patrick O'Brien has highlighted, makes economic history "too fundamental to be virtually ignored by historians and too complex to be left to economists with their shortcuts into cross-country multiple-regression" (Lyons et al., 2008: 443). Similar themes are reflected in comments by Pat Hudson, who has argued that economic historians today are presented with opportunities to "critique the narrow purview of much contemporary economic theory" as well as "ambush the many cultural historians who appear to have forgotten what the economy is" (Hudson, 2010: 249). The great opportunities available to economic historians stem, therefore, from the position of the discipline between the cultures of the social sciences and the humanities, and hence the continuing importance their contributions can make to the historical study of economic life.

See also: Business History; Capitalism: The History of the Concept; Econometrics, History of; Economic History, Quantitative: United States; Financial Crises, History of; Global History: Universal and World; History and the Social Sciences; Industrialization, Typologies and History of; Quantification in History; Slaves and Slavery, History of; Social History.

Bibliography

Boldizzoni, F., 2011. The Poverty of Clio: Resurrecting Economic History. Princeton University Press.
Cain, L.P., Whaples, R., 2013. Economic history and cliometrics. In: Whaples, R., Parker, R.E. (Eds.), Routledge Handbook of Modern Economic History. Routledge, pp. 3–13.
Cipolla, C.M., 1991. Between History and Economics: An Introduction to Economic History. Basil Blackwell.
Coats, A.W., 1980. The historical context of the 'new' economic history. Journal of Economic History 9, 185–207.
Coleman, D.C., 1987. History and the Economic Past: An Account of the Rise and Decline of Economic History in Britain. Oxford.
Coleman, D.C., 1995. History, economic history and the numbers game. Historical Journal 38, 635–646.
Conrad, A., Meyer, J., 1958. The economics of slavery in the ante bellum South. Journal of Political Economy 66, 95–130.
Crafts, N.F.R., 1987. Cliometrics 1971–86. Journal of Applied Econometrics 2, 171–192.
Drukker, J.W., 2006. The Revolution That Bit Its Own Tail. How Economic History Changed Our Ideas on Economic Growth. Aksant Academic Publishers.
Fogel, R., 1964. Railroads and American Economic Growth: Essays in Econometric History. John Hopkins University Press.
Fogel, R., Engerman, S., 1974. Time on the Cross: The Economics of American Negro Slavery. Little, Brown & Co.
Goldin, C., 1995. Cliometrics and the Nobel. Journal of Economic Perspectives 9, 191–208.
Greif, A., 1997. Cliometrics after 40 Years. American Economic Review 87, 400–403.
Habakkuk, J., 1971. Economic history and economic theory. Daedalus 100, 305–322.
Hicks, J.R., 1969. A Theory of Economic History. Oxford University Press.
Hodgson, G.M., 2001. How Economics Forgot History: The Problem of Historical Specificity in Social Science. Routledge.
Hudson, P., 2010. Economic history. In: Berger, Stefan, et al. (Eds.), Writing History: Theory and Practice. Bloomsbury Academic, pp. 248–267.
Jones, G., van Leeuwen, M., Broadberry, S., 2012. The future of economic, business, and social history. Scandinavian Economic History Review 60, 225–253.
Kadish, A., 1989. Historians, Economists, and Economic History. Routledge.
Lyons, J.S., Cain, L.P., Williamson, S.H. (Eds.), 2008. Reflections in the Cliometrics Revolution: Conversations with Economic Historians. Routledge.
Mathias, P., 2010. On economic history: the progress of a discipline living with its neighbours. In: Harlaftis, Gelina, et al. (Eds.), The New Ways of History: Developments in Historiography. I.B. Tauris Publishers, pp. 125–142.
McCloskey, D.N., 1987. Econometric History. Macmillan.
North, D.C., 1963. Quantitative research in American economic history. American Economic Review 53, 128–130.
North, D.C., 1974. Beyond the new economic history. Journal of Economic History 34, 1–7.
Rawski, T. (Ed.), 1996. Economics and the Historian. University of California Press.
Redlich, F., 1965. "New" and traditional approaches to economic history and their interdependence. Journal of Economic History 25, 480–495.
de Rouvray, C., 2004. 'Old' economic history in the United States: 1939–1954. Journal of the History of Economic Thought 26, 221–239.
Sewell, W.H., 2010. A strange career: the historical study of economic life. History and Theory 49, 146–166.
Solow, R., 1985. Economic history and economics. American Economic Review 75, 328–331.
Tuma, E.H., 1971. Economic History and the Social Sciences: Problems of Methodology. University of California Press.
Tuma, E.H., 1974. New approaches in economic history and related social sciences. Journal of European Economic History 3, 169–187.
Vries, P., 2011. Global economic history: a survey. In: Schneider, A., Woolf, D. (Eds.), The Oxford History of Historical Writing, Historical Writing since 1945, vol. 5. Oxford University Press, pp. 113–135.
Whaples, R., 1991. A quantitative history of the Journal of Economic History and the cliometric revolution. Journal of Economic History 51, 289–301.

Economic History, Quantitative: United States

Gavin Wright, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term 'cliometrics' was coined in the 1950s, referring to a self-conscious movement to revolutionize economic history by incorporating quantitative methods and explicit models into the study of past phenomena. The new approach swept to dominance between 1955 and 1975. Subsequently, economic historians began a return to older interests in institutions, culture, and technology. Prominent figures argued that the study of well-established market systems constituted a narrow approach, not well chosen to support a distinctively historical contribution to economics. Although the renewed interest in institutions is a return to earlier terrain, the new incarnation is firmly lodged within the parent discipline of economics.

Quantitative Economic History in the United States

Economic history is a hybrid discipline, concerned with recording and interpreting the economic events of the past. It should not be confused with the 'history of economic thought', the study of the intellectual history of economics as a discipline (*see* Economics, History of). Although the practice of compiling historical statistics goes back to the nineteenth century if not earlier, the term 'quantitative economic history' refers to a self-conscious movement dating from 1950s to 1960s, then known as the 'new economic history' and more recently 'cliometrics', which set out to revolutionize the field by incorporating explicit economic analysis and quantitative methods into the study of past economic events. In subsequent decades, practitioners have extended and redirected this agenda, by drawing upon diverse branches of economic theory to analyze the historical evolution of market and nonmarket institutions, technology, preferences, and other features of economic life commonly taken as exogenous by mainstream economics. The resulting literature stresses the influence of historical experience on ensuing economic patterns, an intertemporal linkage known as 'path dependence'. This article traces the main tendencies in quantitative economic history over the past half-century, though it is by no means a comprehensive survey.

The Origins of Economic History

Recognition of economic history as a distinct subject is traceable to the consolidation of British neoclassical economics at the turn of the twentieth century, under the towering influence of Alfred Marshall and his Cambridge successor, A.C. Pigou. The analytical core of the dominant school held that economies tended toward 'long-run equilibria' that were independent of initial conditions (though Marshall himself was by no means dogmatic about this doctrine in practice). Having lost out in the struggle for the disciplinary identity of economics itself, the 'historical school of economics' emerged at that time as a reaction against the neoclassical paradigm and the laissez-faire program with which it was associated. Drawing ideas and inspiration from the 'German historical school' of Gustav Schmoller and Friedrich List, early historical economists emphasized the virtues of inductive studies as contrasted with the deductive or axiomatic approach. Methodological preferences were reinforced by their affinity with 'neomercantilist' policies, including promotion of national industries, social reforms, and the regulation of enterprise. The organization of economic history as an academic field in Britain was prompted as much by a search for refuge as by the emergence of a distinctive research agenda. William James Ashley, first president of the Economic History Society at its founding in 1926, lamented: "the theoretical economists are ready to keep us quiet by giving us a little undisputed garden plot of our own; and we humble historians are so thankful for a little undisputed territory that we are inclined to leave the economists to their own devices" (quoted in Koot, 1987: p. 7).

Similar tensions persisted in the United States, but the separation between economics and economic history was slower to develop. The inductive approach of the 'institutional' school associated with such figures as John R. Commons and Thorstein Veblen was dominant within labor economics and in other parts of the parent discipline, displaced only by the rise of Keynesian economics in the 1930s and 1940s. In the same era, 'progressive' historians such as Frederick Jackson Turner and Charles Beard put economic conflict and change at the center of their 'economic interpretation of history'. Although neither of these groups would be recognized today as economic historians, they did provide outlets for those dissatisfied with the deductive methodological approach. Thus, the (American) Economic History Association was founded only in 1940, and lacked the ideological identity and fervor of its British counterpart.

Between the 1930s and 1950s, a few exceptional figures such as Carter Goodrich could have a significant impact on both economics and economic history. But the majority of Economic History Association (EHA) members were historians, whose works made little use of economics and had little influence on the thinking of economists. The same could be said for entrepreneurial history, which flourished for a time at the Harvard Center for Entrepreneurial History, under the leadership of Arthur H. Cole. In retrospect, many of the rich descriptive studies from this era have lasting value, perhaps best reflected in the monographic series entitled *The Economic History of the United States*, published between 1945 and 1962. But as economics became increasingly analytical and quantitative, economic history at the time appeared to be something of a stodgy backwater.

The Rise of Quantitative Economic History

Two decades after the World War II saw a significant expansion of economic history within economics, much of it was quantitative in character. This was mainly an American development, and largely (though by no means exclusively) focused on the economic history of the United States. The immediate stimulus was the new postwar interest in economic growth. Following the lead of Simon Kuznets, father of the national income accounts, a new generation of economic historians sought to extend the statistical record of the American economy back to its origins early in the nineteenth century, drawing upon the rich potential of the decennial census and making ingenious use of other diverse archival sources. The major results of these efforts were presented in landmark volumes published by the National Bureau of Economic Research (Parker, 1960; Brady, 1966). Estimates developed by Robert Gallman suggested that growth in the United States had reached modern rates by the middle of the nineteenth century, and should not therefore be seen as the consequence of political upheavals at the time of the Civil War (as hypothesized by Charles Beard). Adapting Gallman's estimates into a regional framework, Richard Easterlin showed that the slave economy of the South also shared in the growth acceleration, though per capita income in that region was some 25–35% below the national average.

A second and more controversial thrust was the deployment of tightly constructed deductive models to address specific hypotheses about this statistical record. The advent of this 'cliometric' approach is often dated from the publication in 1958 of 'The Economics of Slavery in the Ante-Bellum South' by Alfred H. Conrad and John R. Meyer (reprinted in Fogel and Engerman, 1972). In this article, Conrad and Meyer set out to settle a perennial debate over the 'profitability of slavery', by treating slaves as a form of capital and posing the question as an explicit rate of return on investment calculation. Their conclusion was that investments in slaves were just as profitable as alternatives like railroad bonds. Other studies tackled broader questions about the origins of growth and the impact of government policies. Douglass North's *The Economic Growth of the United States, 1790–1860* (1961) interpreted the acceleration after 1815 as a dynamic response to the potential of expanding markets, primarily through interregional specialization and trade. Robert Fogel's *Railroads and American Economic Growth* (1964) and Albert Fishlow's *American Railroads and the Transformation of the Ante-Bellum Economy* (1965) generated comprehensive estimates of the contribution of railroad construction to national economic growth, by posing explicit historical counterfactual circumstances within an analytical framework. Both the methods and the conclusions of these provocative works were the subjects of intense academic debate for many years after publication.

One strand of this discussion represented a reaction by 'old economic historians' against attempts to subsume historical complexity into what they regarded as hypothetical and unrealistic abstractions. A second strand, however, came from within the economics discipline, reflecting dissatisfaction with the limitations of standard resource-allocation models as the working tools for structuring research in economic history. In 1972, two collaborative efforts summarized the major results of the New Economic History as of that date (Davis et al., 1972; Fogel and Engerman, 1972). *American Economic Growth* was a multiauthor collaboration organized around economic sectors; *The Reinterpretation of American Economic History* was an eclectic collection of essays addressing largely microeconomic issues in the new quantitative-analytical style. The editors of a retrospective on the cliometric revolution remark: "Today a large share of the economic history profession in North America consists of the intellectual children and grandchildren of the first generation of cliometricians who reported on their pioneering accomplishments in these volumes" (Lyons et al., 2008: p. 21).

Signs of growing unrest were also evident, however, and in retrospect these volumes have the character of closing chapters in a methodological cycle that had largely run its course by that year. This interpretation may be elaborated by a review of two major research areas.

The Economic Analysis of Technological Change

Economic historians of all stripes had no difficulty agreeing that a central component of economic growth was technological change, broadly defined to include organizational innovations and the diffusion of existing techniques, in addition to the more famous 'inventions' that have long been standard fare in textbooks. But the dominant tradition in economic analysis has long regarded technological change as exogenous, not part of the domain of a discipline specialized in analyzing the *implications* of such change for the economy. One might reasonably hope, therefore, that quantitative economic history would be able to say more about the timing, pace, and direction of particular historical technologies, if not to uncover historical laws of motion for technological change in general.

In the context of American economic history, these questions were brought to the fore by a landmark book published in 1962 by the British economic historian H.J. Habakkuk, *American and British Technology in the Nineteenth Century*. Habakkuk hypothesized that American industry pioneered the development of capital-intensive methods as compared to Great Britain, in response to the scarcity of labor relative to land in the United States. This apparently straightforward thesis subsequently proved exceptionally difficult to capture analytically or to confirm empirically. An extended stream of contributions by Peter Temin, Robert Fogel, Nathan Rosenberg, Paul David, Alexander Field, John James, David Hounshell, Kenneth Sokoloff, Zorina Khan, and Ross Thomson among others, succeeded in elaborating many characteristic features of the technologies emerging as an American national specialty in the nineteenth century. In comparison to the British, American techniques were not only capital-intensive. They were also resource-using; faster-paced; skill-saving (at least for traditional craft skills); and adapted for long production runs of standardized final products. As useful as this accumulated descriptive knowledge may be, the question remains: what sort of economic analysis would allow the quantitative economic historian to come to terms with the origins and nature of this list of characteristics?

The long-running debate over the 'labor scarcity paradox' brought out most clearly that so long as the analytical framework was defined by standard neoclassical analysis, the subject matter would be transmuted into the language of 'relative factor abundance' and substitution, terms derived from the

assumption that a 'production function' is given by technical factors. Within this constraint, the economic historian was powerfully driven toward the exercise of rationalizing observed historical choices. The impulse was observed most clearly when the choice-of-technique question was transported to the other side of the Atlantic, and cliometricians assessed the popular indictment of British entrepreneurs for their reluctance to adopt American methods. Virtually without exception, the conclusion was that British choices were entirely appropriate responses to prevailing factor prices and market conditions. Although this verdict may be valid, the uniformity of results gave rise to doubts about whether it constituted an authentic inference from historical evidence, or merely an assumption imposed on the evidence by the economist's worldview.

This approach to technology proved unduly constraining for a number of reasons. As subsequent research has demonstrated, the 'factor endowments' taken to be fundamental determinants were themselves not really exogenous, being subject to influence by policy and by technology itself as it unfolded through time. In agriculture, Paul Rhode and Alan Olmstead (2008, 2011) show that the expansion of American farmland required not just transportation but new knowledge, including new crop varieties and cultivation practices adapted to the soils and climate of western regions. In other words, the growth of the productive input 'land' was itself the result of learning and knowledge accumulation, so that the observed constancy of yields per acre should not be understood to indicate an absence of 'technological progress'.

In industry, the large inflows of unskilled immigrants to the United States are most readily explained by the growth of industrial jobs adapted to that type of labor, a result in other words of an attribute of the new technology. Even American abundance of mineral resources – coal, iron ore, copper, petroleum, and many more – turn out on closer inspection to have been more the *result* of technological development and capital investment than its cause (David and Wright, 1997). To a significant extent, minerals constituted a 'knowledge industry' even in nineteenth-century America. As such, its development was part of the broader evolution of national technology, or learning, whose evolution was itself the phenomenon to be explained.

The Slavery Debate

No topic tested the limits of deductive economic history more fully than slavery. Conrad and Meyer (1958) set out to measure the 'profitability of slavery', and thereby to test the thesis that slavery would have died out for economic reasons, in the absence of the Civil War. Among others, Yasuba pointed out that because slave imports were ended by legislation in 1807, a more appropriate measure of the private economic viability of slavery was the level of 'capitalized rent' embedded in the price of slaves – high and rising on the eve of secession, according to Yasuba's estimates (1961, reprinted in Fogel and Engerman, 1972). The 'profitability' finding is best understood as an indication that slave prices approximately capitalized the expected stream of future returns, which may be taken as evidence that 'capitalist' motivations were dominant in the operation of the slave system. The achievement of consensus around this broad proposition, however, redirected attention toward more difficult and challenging historical issues, such as the long-term consequences of slavery as a mode of social organization.

This was the setting for the appearance in 1974 of *Time on the Cross: The Economics of American Negro Slavery*, by Robert W. Fogel and Stanley L. Engerman. Fogel and Engerman went well beyond the previous consensus, presenting total factor productivity calculations implying that slave plantations were as much as 35% more efficient than free family farms. Not only was the slave system highly productive, but also the authors went on to assert that material and social conditions among slaves were far better than historians had recognized. Using such indicators as diet, clothing, housing, life expectancy, and family stability, the authors compared the lives of slaves favorably with those of immigrant workers in the northern states. It is hardly surprising that these incendiary claims generated an explosive reaction.

The full range of the ensuing debate cannot be recounted here. As with the response to Conrad and Meyer, one part of the reaction represented traditional historians objecting to the coldhearted nature of the calculations, and the apparently cavalier treatment of the feelings and behavior of the slaves themselves. Another component, however, came from within the profession. Detailed scrutiny of the statistical evidence ultimately led to significant modifications of the relatively benign picture of slavery, most notably in health conditions and family life. It would be mistaken to characterize the alternative views on these issues along simplified dimensions, such as the presumption that historical operators are rational and market oriented. For example, subsequent research suggests that slaves were bought and sold much more frequently than Fogel and Engerman believed, often separating not only husbands from wives, but even mothers from their children. On this count, *Time on the Cross* if anything actually understated the ruthlessly profit-driven character of American slavery.

Debates over empirical evidence are a normal feature of any social science. The dimension of the slavery debates with more enduring import for quantitative economic history concerned the *conceptual* underpinnings of the analysis of slavery, specifically the emphasis on calculations of efficiency in production. Because plantations produced multiple outputs with heterogeneous inputs, it emerged that the efficiency results were highly sensitive to relative values assigned to cash crops (such as cotton) vs food crops (such as corn), and to factors of production (such as farmland of varying quality, or male vs female slave labor). The debate may be followed in David et al. (1976), a subsequent exchange in the *American Economic Review*, and Wright (2006). The deeper issue was whether it was possible, or perhaps whether it was meaningful and historically informative, to separate the calculation of productive efficiency as a technical matter, from the task of locating slavery as a system of property rights in historical time and space. From this perspective, property rights in slaves had a decisive influence on the course of economic development in the southern states, not only on human capital investment, but also on geographic patterns of settlement, transportation infrastructure, and urbanization.

In several respects, the course of the slavery debate highlights the new directions in which quantitative economic history has moved since the 1970s: study of the evolution of institutions; incorporation of political, social, and ideological

dimensions into economic history; and identification of lasting effects of historical phases on later events. These tendencies are well illustrated by the careers of two founding figures of the New Economic History.

The Nobel Prizes

In October 1993, the Royal Swedish Academy of Sciences awarded the Nobel Prize in Economics to Robert William Fogel and Douglass North, noting that "they were pioneers in the branch of economic history that has been called the 'new economic history', or 'cliometrics'." In essence, the two economic historians were cited for helping to launch the quantitative economic history movement. In what must be a rarity among Nobel laureates, at the time of the award both men had moved some distance from the approach for which they were honored.

North's early work in American economic history was both quantitative and analytical. He developed new statistical time series on the US balance of payments, and on the long-term decline in transportation costs. North (1961) drew upon concepts from theories of location and comparative advantage to interpret the quantitative record of regional performance. His first ventures outside American economic history were consistent with this focus on market incentives, emphasizing the role of limited government and secure property rights in explaining the comparative historical performance of nations. By the 1980s, North's range of analysis had broadened considerably, to include *ideology* (including norms and belief systems) as a primary element in comprehending the persistence of order in the world, essential for the effective functioning of markets. North (1990) cast the net even more widely, pointing to the critical economic role of *institutions*, defined broadly to include all "the humanly devised constraints that shape human interaction". This expansion of horizons continued in North (2005) and North et al. (2009), which explore the historical determinants of institutions and the comparative efficacy of societal systems for generating *new* institutions. At least part of North's personal evolution arose from his growing awareness that the effectively functioning market society found in the United States is quite exceptional in a global context.

Fogel's intellectual trajectory was rather different. Both Fogel (1964) and Fogel and Engerman (1974) built sweeping historical interpretations on the results of complex and ambitious quantitative procedures. After 15 years of further research on slavery, Fogel (1989) devoted more than half of his next book to an account of the rise of antislavery opinion in Europe and the United States, emphasizing the force of religious, political, and ideological factors. Fogel wrote: "I began [this research], like many other cliometricians, not because I was especially interested in the history of American slavery, but because an accident of scholarship made the economics of slavery a major testing ground for the application of cliometric methods. Once drawn into the subject, however, it was the substance of the issues that maintained my interest. Although my principal professional expertise was, and is, in the areas of economics and demography, I found myself led down a road that forced me to grapple with the work of colleagues in cultural, political, and religious history" (p. 13). Fogel's interest in subjective and ideological aspects of history persisted through his later quantitative work on historical nutrition and physiology (2000, 2004). It seems safe to say that by the twenty-first century, cliometrics had outgrown the historical naivete of its youth. (Robert Fogel died on 11 June 2013. An appreciative remembrance by former student is Goldin (2014).)

New Directions in Quantitative Economic History

Since the methodological wars of the 1950s through the 1970s, the cliometric revolution has become all but complete within economic history. Most practicing economic historians are now employed in departments of economics, and the majority of articles published in the *Journal of Economic History* deploy econometric techniques. The list of cliometrically oriented journals has steadily expanded in recent decades: for example, *Explorations in Economic History*, *The European Review of Economic History*, and *Cliometrica*. Even the staid British *Economic History* Review is now far more quantitative and economic than in earlier times. Thus, economic history has thrived in its cliometric identity, though diminished by the only partially anticipated consequence that the field has largely disappeared from departments of history.

Although most quantitative economic history accepts economics as the parent discipline, the field remains distinct in its practices and its intellectual priorities. In comparison with their mainstream colleagues, economic historians place greater emphasis on the historically contingent character of economic relationships, deploying a more eclectic range of theoretical models and conceptual tools. Indeed, over the years economic history has become steadily more diverse and expansive in both its methods and its subject matter. Some of these tendencies reflect growing diversity in the composition of the cliometric intellectual community, which now includes many more women and scholars of differing nationality than was true in the 1960s and 1970s. Once almost exclusively an American phenomenon, cliometrics is now also pursued by small but thriving groups in Europe, Latin America, Canada, and elsewhere. Another force for diversification is the expanding range of topics now covered within economics itself, including ecology, law, family, gender, and race, all of which have strong historical dimensions. Since the full range of scholarship cannot be surveyed here, a brief account will be given of two general areas in which historical aspects of economic life seem especially compelling.

Increasing Returns, Path Dependence, and Technology

In 1984, William N. Parker organized a session at the annual meetings of the American Economic Association, on the role of economic history in graduate economics education, the proceedings of which were published in Parker (1986). Among the many spirited defenses of history by such economists as Kenneth Arrow, Robert Solow, Peter Temin, D.N. McCloskey, and others, one contribution proved to be particularly lasting as an illustration of path dependence: Paul A. David's account of the rise and persistence of the QWERTY typewriter keyboard. Although many evidently superior layouts were designed and marketed during the typewriter's post – Civil War diffusion, the standard became 'locked in' to QWERTY by the 1890s,

illustrating the power of historical accident and 'network externalities' to shape the course of technological history. Drawing in part upon theoretical work by W.B. Arthur, David argued that because choices were interdependent – since manufacturers, employers, and typists all wanted to use the same keyboard that others were using – the system never shook off the circumstances of its origin, despite the suboptimality of the equilibrium choice.

The QWERTY story has emerged as a popular metaphor for the proposition that 'history matters' in economic life. The typewriter example and its meaning have been hotly controversial, and its appropriate scope and significance have by no means been settled. But its relevance for quantitative economic history has been to confirm the historical character of the economics of technology. As noted above, the study of technological change had never fit comfortably within the New Economic History. Nathan Rosenberg, a member of the first cliometric generation whose research on technological change diverged from the reigning paradigm, had long been engaged in *Exploring the Black Box* (1994), i.e., trying to understand technological change as it would be perceived and advanced by members of technical communities. Since then, such economic historians as Robert Allen, James Bessen, Claudia Goldin, and Joel Mokyr have elaborated the context-specific factors that have served to encourage, retard, or channel technological progress. The venerable debate on 'American technological exceptionalism' has been transformed by detailed studies of the institutional bases (both formal and informal) for US technology by economic historians such as Ross Thomson, William Lazonick, Kenneth Sokoloff, Zorina Khan, Naomi Lamoreaux, and Petra Moser.

The broad conclusion of this research is not to deny that technology is responsive to incentives – quite the contrary – but to argue that these responses are conditioned by the historical and institutional setting, including the cognitive structures through which knowledge is codified and transmitted. Distinctively historical studies with these features include Lampe and Moser (2010), Saxonhouse and Wright (2010), Moser (2011), Bessen (2012), and Gray (2013).

Institutions and Economic History

In the heyday of the New Economic History, 'institutional' was used as a term of dismissal if not contempt, to refer to old-fashioned descriptive work that lacked theoretical foundations. As the field has evolved, 'institutional' economic history has again become respectable, though in new and more analytical clothing. Indeed, many argue that the study of institutions is at the heart of the case for historical economics. Whereas the parent discipline specializes in the functioning of markets, the domain of the economic historian is the institutional foundations that allow markets to function as they do. These subject areas have become more pressing in light of the varied experiences of formerly planned economies in their transition to market systems.

In its first phase, the so-called New Institutional Economics identified 'institutions' with the property rights, rules and regulations defined and enforced by the state, interpreting institutional change as responses to exogenous shifts in population, technology, or the scope of markets. Beginning with North and Thomas (1973), numerous historical studies examined the implications of these institutions for economic performance, broadly confirming the general proposition that securely enforced property rights are favorable for economic efficiency and growth. Deviations from this pattern are commonly attributed to the influence of special interest groups, with transactions costs standing as barriers to more efficient systems.

Dissatisfaction with this framework gradually began to build, with the realization that the simple cause and effect scenarios took as exogenous much of what was most historically interesting. An essential next step was to broaden the concept of institutions, to include not only explicit and state-enforced rules, but informal and uncodified constraints on behavior as well. To say that institutions are 'constraints', on individual behavior, however, does not mean that they must be seen as having been 'imposed' on the society in question. By analogy to market prices, institutions may be viewed as exogenous to the individuals, yet historically endogenous to the society as a whole. When the underlying conception has been broadened in this way, a natural theoretical framework is provided by game theory, depicting institutions as self-enforcing equilibria arrived at through the interaction of historical players. Avner Greif (2006) has been a pioneer in the application of game theory to historical institutions, using these tools to analyze the bases for the rise of long-distance trade in late medieval Europe.

Greif argues that appropriate historical uses of game theory must be context-specific. In many strategic situations, there are multiple equilibria, implying that outcomes are not uniquely predetermined. Constructing a context-specific model that captures the essence of a situation requires microlevel historical study, to provide the specific details of the model, and to facilitate the task of confronting predictions with historical evidence. Thus, the process has a large inductive component, drawing on cultural, political, and social factors as well as upon economic motivations. Although this specificity may seem to limit the potential scope for application, Greif was able to use these tools to extract sweeping implications from the contrasting cultural beliefs of the Maghribi traders (who used collective social instruments to discipline the behavior of their members) and the Genoese (from whose rule-based system, the Western legal traditions descend). Such institutional evolution often exhibits increasing returns and path dependence, similar to technological systems subject to network externalities.

In many ways modern quantitative economic history has returned to the roots of economic history, in its renewed appreciation of institutions and culture. But there is a major difference. Whereas earlier versions of 'institutional economics' tended to be 'antitheory', stressing the sheer weight of historical reality as contrasted with abstract rationalizations, the newer forms of institutional economic history are firmly lodged within the parent discipline, drawing on frontier developments in economic theory as well as quantitative methods. The ability to deploy more eclectic theoretical tools allows cliometricians to extend their reach into new subject areas that previously were preserves of purely descriptive work. A prominent example would be historical studies of financial markets and financial institutions, which now use theories of information and principle-agent relationships. This line of research returned

to prominence when the financial crisis of 2007–08 undermined faith that these markets are or had become essentially self-regulating. Barry Eichengreen (2012) devoted his 2011 presidential address to the Economic History Association to an exploration of how the historical analogy to the Great Depression of the 1930s shaped perceptions and policy responses many decades later.

The history of labor markets and related institutions – central to the mission of the 'old' institutional economics – has also enjoyed a renaissance, because the roles of norms, prejudices and family relationships are too prominent in this subject area to be ignored. An illustration of this is Claudia Goldin's book on the economic history of American women (1990). The early chapters developed new quantitative evidence on relative wages and labor force participation rates of American women, interpreting two major upward surges (1820–50 and 1890–1930) in terms of supply and demand forces, largely driven by technology. By the latter half of the book, however, Goldin found that 'institutional' aspects of the record were inescapable, from the 'marriage bars' of the 1920s and 1930s – which forced women to resign from jobs upon marriage – to patterns of socialization and peer influence, detectable in behavioral shifts from cohort to cohort across the twentieth century. Reflecting on her own evolution, Goldin wrote: "I began this study more as an economist but have ended with a fuller appreciation of how the distant past affects the present, how norms and expectations impede change, how discrimination can survive even in highly competitive markets, and how slow genuine change can be" (p. viii).

To be clear, an 'institutional' approach to labor economics does not imply a belief in universal historical persistence. If 'lock-in' means that a system no longer responds to incremental changes in incentives, the corollary is that when change does come, it is likely to be sudden and unanticipated; the flip side of path dependence is historical discontinuity. Thus, Joshua Rosenbloom and William Sundstrom (2010) interpret American labor history as a series of 'regimes', apparently stable equilibria interrupted by periods of rapid and often surprising structural change. Stacey Jones (2009) recounts one such episode, the dramatic transformation in social norms and women's higher education between 1965 and 1975. In these studies and others, economic history has developed an increasingly self-conscious identity, not as a branch of applied economics but as a distinctive approach to the study of economic life.

See also: Economics, History of; Financial Crises, History of; Innovation and Technological Change, Economics of.

Bibliography

Bessen, J., 2012. More machines, better machines…or better workers? Journal of Economic History 72, 44–74.

Conrad, A.H., Meyer, J.R., 1958. The economics of slavery in the Antebellum South. Journal of Political Economy 66, 518–560.

Brady, D.S. (Ed.), 1966. Output, employment, and productivity in the United States after 1800. Studies in Income and Wealth, vol. 30. National Bureau of Economic Research, New York, pp. 117–204.

David, P.A., Gutman, H.G., Sutch, R., Temin, P., Wright, G., 1976. Reckoning with Slavery: A Critical Study in the Economics of American Negro Slavery. Oxford University Press, New York.

David, P.A., Wright, G., 1997. Increasing returns and the genesis of American resource abundance. Industrial and Corporate Change 6, 203–245.

Davis, L.E., Easterlin, R.A., Parker, W.N., Brady, D.S., Fishlow, A., Gallman, R.E., Lebergott, S., Lipsey, R.E., North, D.C., Rosenberg, N., Smolensky, E., Temin, P., 1972. American Economic Growth: An Economist's History of the United States. Harper and Row, New York.

Eichengreen, B., 2012. Economic history and economic policy. Journal of Economic History 72, 289–307.

Fogel, R.W., 1964. Railroads and American Economic Growth: Essays in Econometric History. John Hopkins University Press, Baltimore, MD.

Fogel, R.W., Engerman, S.E. (Eds.), 1972. The Reinterpretation of American Economic History. Harper & Row, New York.

Fogel, R.W., Engerman, S.E., 1974. Time on the Cross: The Economics of American Negro Slavery. Little, Brown and Company, Boston, MA.

Fogel, R.W., 1989. Without Consent or Contract. W. W. Norton, New York.

Fogel, R.W., 2000. The Fourth Great Awakening and the Future of Egalitarianism. University of Chicago Press, Chicago.

Fogel, R.W., 2004. The Escape from Hunger and Premature Death: Europe, America and the Third World. Cambridge University Press, New York.

Goldin, C.D., 1990. Understanding the Gender Gap: An Economic History of American Women. Oxford University Press, New York.

Goldin, C.D., 2014. Robert william fogel remembrance. Cliometrica 8, 1–3.

Gray, R., 2013. Taking technology to task: the skill content of technological change in early twentieth United States. Explorations in Economic History 50, 351–367.

Greif, A., 2006. Institutions and the Path to the Modern Economy: Lessons from Medieval Trade. Cambridge University Press, New York.

Jones, S.M., 2009. Dynamic social norms and the transformation of women's higher education, 1965–1975. Social Science History 33, 247–291.

Koot, G.M., 1987. English historical Economics, 1870–1926: The Rise of Economic History and Neomercantilism. Cambridge University Press, Cambridge.

Lampe, R., Moser, P., 2010. Do patent pools encourage innovation? Evidence from the nineteenth-century sewing machine industry. Journal of Economic History 70, 898–920.

Lyons, J.S., Cain, L.P., Williamson, S.H., 2008. Reflections on the Cliometrics Revolution: Conversations with Economic Historians. Routledge, London.

Moser, P., 2011. Do patents weaken the localization of innovations? Evidence from world's fairs. Journal of Economic History 71, 363–382.

North, D.C., 1961. The Economic Growth of the United States 1790–1860. Prentice-Hall, Englewood Cliffs, NJ.

North, D.C., Thomas, R.P., 1973. The Rise of the Western World: A New Economic History. Cambridge University Press, Cambridge.

North, D.C., 1990. Institutions, Institutional Change, and Economic Performance. Cambridge University Press, New York.

North, D.C., 2005. Understanding the Process of Economic Change. Princeton University Press, Princeton, NJ.

North, D.C., Wallis, J.J., Weingast, B.R., 2009. Violence and Social Orders: A Conceptual Framework for Interpreting Recorded History. Cambridge University Press, New York.

Olmstead, A.L., Rhode, P.W., 2008. Creating Abundance: Biological Innovation and American Agricultural Development. Cambridge University Press, New York.

Olmstead, A.L., Rhode, P.W., 2011. Responding to climate challenges: lessons from U.S. agricultural development. In: Libecap, G., Steckel, R. (Eds.), The Economics of Climate Change: Adaptations Past and Present. University of Chicago Press, Chicago.

Parker, W.N., 1960. Trends in the American Economy in the Nineteenth Century. Studies in Income and Wealth, vol. 24. National Bureau of Economic Research, New York.

Parker, W.N. (Ed.), 1986. Economic History and the Modern Economist. Basil Blackwell, New York.

Rosenbloom, J.R., Sundstrom, W.A., 2010. Labor-market regimes in U.S. economic history. In: Rhode, P.W., Rosenbloom, J.L., Weiman, D.F. (Eds.), Economic Evolution and Revolution in Historical Time. Stanford University Press, Stanford, CA.

Saxonhouse, G.R., Wright, G., 2010. National leadership and competing technological paradigms: the globalization of cotton spinning. Journal of Economic History 70, 535–566.

Wright, G., 2006. Slavery and American Economic Development. Louisiana State University Press, Baton Rouge.

Economic Restructuring: Geographic Aspects

Steven Pinch, University of Southampton, Southampton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Geographical studies of economic restructuring seek to explain spatial variations in economic development. The term 'economic restructuring' typically is used in a very general sense to indicate changes in the constituent parts of an economy. In many respects, it is a catch-all phrase, incorporating factors such as the changing nature of industrial sectors, corporate organization, management strategies, occupational structures, employment cultures, and working practices. However, the term 'restructuring' has also become associated with a series of geographical studies with distinctive perspectives on changing economies. Although a series of foci, rather than a single integrated approach, these perspectives share a common desire to understand the geographical complexities of economic restructuring without recourse to universal laws of economics that are assumed to work in all places at all times. Instead, economic behavior is seen as dependent upon the cultural norms and institutional practices manifest in particular places at specific points in the history of capitalist development. Restructuring studies focus upon the processes leading to restructuring, the mechanisms through which this restructuring is achieved, and the social consequences of these processes in terms of factors such as income distribution, occupational structures, the prevalence of unemployment, and the incidence of two-career households. Nevertheless, it is possible to group restructuring studies by a number of empirical foci and theoretical approaches: spatial divisions of labor, 'clusters,' globalization, services, knowledge-, network- and digital-based economies, and financialization.

Economic Restructuring: Common Underlying Themes of Geographic Studies

Geographical studies of economic restructuring seek to explain spatial variations in economic development. The approach taken in most of these studies differs from the neoclassical approach within economics, which attempts to discover universal principles or 'laws' of economic behavior that are relevant in all places at all times. In addition, the neoclassical approach places faith in the ability of market mechanisms within a capitalist economy to produce the most efficient allocation of resources in a way that is beneficial for all concerned. In marked contrast, the restructuring approach within geography sees economic behavior as highly dependent upon the social customs, cultural norms, institutional practices, and wider forces of capitalist development in operation at the time, and in the place, in which that behavior takes place. Furthermore, rather than a consensual outcome that is to the benefit of all concerned, the process of restructuring under capitalism typically is seen as a struggle between competing interest groups, leading to winners and losers. Geographical studies have therefore focused upon both the creative and destructive outcomes prompted by the continuous need for capitalist firms to generate profits (Storper and Walker, 1989).

The restructuring approach has generated a vast body of literature undertaken by geographers and other social scientists interested in geographical issues, including regional economists and urban sociologists. These studies have been prompted by recognition of the fact that geography is a central element in the restructuring process. Not only does restructuring vary from one place to another but, in addition, the differences between areas are used by various agents – governments, corporations, entrepreneurs, trade unions, and workers – to affect the outcome of the restructuring process itself. Geography, sometimes referred to as place or space, is, therefore, not just a reflection of these processes, it is a constituent part of the restructuring process itself. A good example of the role of space is when a corporation moves employment from an area in which labor costs are high to a low labor cost region, or from an area in which workers are resistant to new working practices to one in which they are more compliant.

Restructuring studies focus upon the processes leading to restructuring, the mechanisms through which this restructuring is achieved, and the social consequences of these processes in terms of factors such as income distribution, occupational structures, the prevalence of unemployment, and the incidence of two-career households. Nevertheless, it is possible to group restructuring studies by a number of empirical foci and theoretical approaches: spatial divisions of labor, 'clusters,' globalization, services, knowledge-, network-, and digital-based economies, and financialization.

These studies have in recent years broadly adopted a 'relational' view in which firms are conceptualized as having to coordinate with a range of other actors – workers, consumers, suppliers, regulators, and government agencies. This has led to a focus upon the role of various institutions (sets of rules) that affect these interrelationships and how these vary from place to place and between nations in what are termed 'varieties of capitalism' (Hall and Soskice, 2001; see also Peck and Theodore, 2007). Explicit use of the term 'restructuring' has fallen somewhat from favor in recent years but social scientists from a variety of disciplines continue to examine the spatial manifestations of the continual reconfiguration of the global economy.

Spatial Divisions of Labor

The first geographical approach to focus explicitly upon economic restructuring was spatial divisions of labor (Massey and Meegan, 1982; Massey, 1995). In essence, this concept

highlights the fact that throughout the first part of the twentieth century there was a shift in regional economic structures away from sectors toward functions. Within regional sectoral specialization, regions tended to specialize in all aspects of the production process of particular products, such as ships, cars, or textiles. Classic examples of regional sectoral specialization are the car manufacturing complex of Detroit or the once world-leading shipbuilding region of the Clyde in Scotland. However, following a shift toward regional functional specialization, in the second half of the twentieth century, regions tended to specialize in particular *stages* in the production process of a wide range of products. The spatial divisions of labor approach therefore argues that there is a tendency for different types of work to be undertaken in different places. Thus, corporate headquarters tend to be concentrated in large cities, design centers may be concentrated in high-technology clusters, while the routine assembly of products may be based in regions with low labor costs.

Spatial divisions of labor are manifest at other scales in addition to the region. At the city-wide scale, areas of manufacturing specialization are often termed 'industrial districts.' However, decentralization in the latter half of the twentieth century promoted functional specialization as new forms of work – and especially services – were introduced into suburban areas following the outmigration of people from inner cities. Spatial divisions of labor can also be seen at the international level. Hence, the advanced capitalist regions of North America, Europe, and Japan may be seen as global centers of corporate control, finance, product design, and innovation, while routine assembly jobs are undertaken in developing countries. However, research soon made it clear that there were few, if any, clear-cut geographical patterns, and the shift from sectors to functions was a tendency or trend rather than a specific outcome. Hence, different industrial sectors produced complex spatial divisions of labor, some of which overlapped and some of which did not.

The spatial division of labor approach attributed this geographical restructuring to the continual need of capitalist firms to enhance their profitability. There is much debate about the processes that led to low productivity and declining profitability in many industries in the second half of the twentieth century. One important influence was growing international competition from low-cost locations outside the Western world. Other factors were a lack of investment in new machinery and boring repetitive work (often associated with labor unrest). Another factor prompting the desire for economic restructuring in the early part of the 1980s was the influence of corporate raiders who issued junk bonds to gain control of major corporations. They then enhanced share values by asset stripping: putting new management into particular corporate divisions to enhance productivity through reductions in manpower and the introduction of new working practices, and then selling off the revitalized units. Prosecutions for insider dealing led to a progressive decline in the influence of the raiders, but gradually pensions funds and major financial institutions began to exert similar pressures on corporations to enhance productivity and profitability.

Geographical studies drew attention to processes that were particularly important in the late 1970s and early 1980s as companies began to respond to declining profitability at that time (Massey and Meegan, 1982). First, there was rationalization, a euphemism for the reduction of industrial capacity through factory closure or massive workforce cutbacks. Job losses and factory closures were especially pronounced in the older industrial areas of the United Kingdom, a pattern that was replicated in the United States, Europe, and Australia. Rationalization frequently was associated with spatial relocation as routine production functions were transferred to areas with low labor costs in developing countries. Not only were shop-floor workers made redundant, intermediate level white-collar workers were also laid off as corporations stripped out their elaborate managerial and administrative hierarchies (a process sometimes termed 'delayering' or 'downsizing'). Second, there was the process of investment and technical change. This involved capital investment in new machinery and equipment in an effort to increase productivity and the quality of products. This frequently was also associated with losses of employment. Third, there was the process of intensification: increases in labor productivity through managerial and organizational change.

Many of these changes were made to increase flexibility. Rather like restructuring, 'flexibility' is a word with a wide range of meanings, but in the context of economic restructuring it has taken on two further important associations. First, there is functional flexibility, the capacity of firms to extend the range of skills of their employees. This has often involved breaking down the barriers between different types of occupations. Second, there is numerical flexibility, the capacity of employers to alter the numbers of workers employed at any given time. This has involved a wide range of strategies including the increased use of temporary workers, part-time staff, casual employees, job sharing, shift working, flextime, annualized hours, and a wide range of innovative forms of job contract.

Two important factors underpin the need for flexibility. First, there is the ever-increasing pace of technological change, both of products and the means by which they are produced. Increased use of computerization has tended to break down some of the divisions between mechanical-based and electronic-based occupations. The other factor is increasing market volatility. Markets for certain mass-produced goods have become saturated while niche markets for specialized goods change rapidly. Producers therefore need to be able to adjust quickly and efficiently to variations in demand.

As with other aspects of industrial restructuring, the capacity to introduce flexibility shows complex geographical patterns and wide variations among regions. For example, in an attempt to escape from labor hostility to flexible working practices in older industrialized areas with a tradition of trade unionism, many companies have established new factories in areas that are free of such traditions. The location of US car plants outside of the traditional automobile assembly areas of Detroit and Michigan in recent decades, illustrates this trend. Restructuring has greatly affected the economic fortunes of different places, favoring some and devastating others.

One of the most quoted and influential aspects of the spatial divisions of labor approach is Massey's use of the geological metaphor. The economic geography of an area was likened to various phases or 'rounds' of investment in factories, plant, machinery, and labor. Each round leaves an imprint in the form of a layer or strata of development. Some of these

layers become eroded over the years by subsequent rounds of the restructuring process, but collectively they serve to shape the economic landscape. Critiques of the early spatial divisions of labor approach argued that it gave too much attention to the power of capital. A reformulated 'territorial division of labor' approach therefore attempts to take greater awareness of the capacity of workers to influence economic restructuring (Dunford, 2003) (*see* Local Economic Development).

Although the spatial divisions of labor approach recognized that women were often employed in routine assembly work, feminist perspectives have also highlighted the neglect of gender issues in many of these earlier studies. Earlier research focused upon work dominated by men – most notably the assembly of cars or high technology research and design work – even though women frequently were engaged in routine 'Fordist' assembly work. Studies tended to adopt the reserve army of labor theory to explain the position of women in the labor market (i.e., it assumed that they were a variable pool of workers taken on in times of high demand but rapidly laid off in times of recession). A number of flaws were soon apparent in this theory. First, it simply did not fit the facts; in the last quarter of the twentieth century, women were an ever-increasing proportion of the workforce rather than a 'reserve army.' Second, previous studies frequently ignored the socially embedded character of women's employment. The issue of skill illustrates this point well. For example, rather than being an innate attribute, women's so-called nimble-fingers, which supposedly make them suitable for low-paid routine assembly work in light industry, are now seen as socially ascribed. Subsequent work has shown how the combination of workplace, household, and community arrangements structure the job opportunities available to women (Hanson and Pratt, 1995). Gender issues are therefore a crucial element in processes of economic restructuring, as are many other socially ascribed human characteristics related to factors such as family status, ethnicity, disability, and age (*see* Gender and Feminist Studies in Geography).

Clusters

During the 1980s, there was intense interest in Japanese manufacturing practices, and this drew attention to a rather different structure to regional functional specialization. Because of the need for geographical proximity to their smaller component suppliers, who provided parts on a just-in-time basis (as required on the production line), large Japanese manufacturing firms began to perpetuate and enhance regional sectoral specialization. Indeed, there has emerged a second strand to the restructuring approach that argues that clustering of firms making the same products (known as regional agglomeration) was once again an important feature of the advanced economies. This approach has been prompted by the fact that some regions lead the world in the design and production of particular products. The classic case is Silicon Valley in California that leads the world in electronics and computer software (Saxanien, 1994). Other examples include racing car production in southern England, textiles in northern Italy, and machine tools from the Baden-Württemberg region of Germany. Rather than regional functional specialization, these examples suggest we are witnessing the resurgence of the industrial districts of an earlier age.

Following in the wake of the influential management strategist Michael Porter (1998), these districts are often referred to as 'clusters,' although the term is highly controversial (Martin and Sunley, 2003). Various other labels have been ascribed to such regions of economic activity by differing schools of thought including new industrial spaces (Scott, 1988), associational economies (Cooke and Morgan, 1999), and learning regions (Morgan, 1997). Although these approaches vary in emphasis, at root is a debate over the extent to which the processes observed by Alfred Marshall (1890) in industrial districts in nineteenth-century England (such as steel making in Sheffield) can be applied to contemporary economies. Marshall famously noted three basic sources of external economy, whereby small firms could gain by being in close proximity to each other: through access to pools of highly skilled labor; the availability of specialized and efficient local suppliers and services; and the benefits of a local 'industrial atmosphere' characterized by rapid dissemination of information, trust, innovation, and competition.

Updating these ideas, Scott (1988) explained clusters as the outcome of transaction costs – the costs of transforming inputs into products and services. Companies are faced with three choices in minimizing transaction costs. They can undertake activities internally (in-house) through some form of *hierarchy*, subcontract activities to external suppliers and contractors through the *market*, or adopt a halfway house and engage in collaborative activities with other companies through *networks*. Scott argued that changing economies of scale and scope were affecting transaction costs and leading to firms clustering together geographically in collaborative networks.

Economies of scale are factors that cause the average cost of a commodity to fall as the amount of output of the commodity increases. Internal economies of scale are those factors that apply to individual firms (irrespective of the size of the industry to which they belong). These internal economies usually ensure that the optimal scale of output is large in order to recoup the high costs of research, development, and specialized capital equipment. It is these factors that led in the past to the growth of large factories. This system has been termed Fordism after the system of mass production pioneered by Henry Ford in Detroit for the production of cars.

External economies of scale, in contrast, apply when the industry to which the firm belongs to (rather than the firm itself) is large. External economies include the development of specialized ancillary services, pools of skilled labor, and specialist component suppliers. If the output of the latter is large, this can reduce costs, thus potentially benefiting the industry as a whole. Economies of scope are factors that make it cheaper to produce a range of commodities (rather than to produce individual types of commodity in isolation) and again these economies of scope can be internal or external. Critics pointed out, however, that the transaction-cost approach is based on an economic logic that ignores historical factors (Sayer, 1989). This critique has led to an approach in restructuring studies that recognizes the social embeddedness of economic activities.

The concept of social embeddedness recognizes that all economic activity depends upon the social context in which it takes place (Granovetter, 1985). Economic systems are seen as the product of conventions applied by highly 'reflexive' (i.e., thinking and aware) human agents. All economic systems depend upon systems of meaning that are usually specific to particular places. In the context of regional development, this has led to a focus upon the institutional structures that encourage or inhibit economic activity. Work has therefore focused upon the cultural factors that generate a strong sense of common purpose, trust, and skill development amongst geographically clustered firms. This trust can be fostered by 'strong ties' based on regular and intense face-to-face contact, but this sort of close-knit networking can lead to rigidity in thinking, and there is also recognition that fleeting contacts with people from a large pool on nonoverlapping networks can bring workers into contact with a wider range of ideas (e.g., Storper and Venables, 2004).

Another critique of the focus upon agglomeration economies in cluster theory has come from evolutionary perspectives. A rapidly growing force within the discipline, evolutionary economic geography attempts to show how socioeconomic systems change over time, drawing inspiration with metaphors, and concepts from a variety of evolutionary sciences (Martin and Sunley, 2006). Hence the approach looks at how firms with superior capabilities tend to survive and how consequently their workers inherit capacities and routines of operation that they pass on in various ways to later generations. After analysis of US automotive and electronics clusters, Klepper (2010) argues that rather than being a manifestation of distinctive locality-based agglomeration effects, these clusters reflect such an evolutionary learning process resulting from new businesses spinning out from a few original firms whose initial location was largely the result of historical accident. Klepper (2010) argues that the new firm entrants in his clusters were largely indigenous spin-offs and that their superiority in terms of survival and profitability reflected their innate characteristics. However, as he also acknowledges, this raises the further issue of how were these superior capacities acquired; was it through the experiences gained by these new entrepreneurs in superior (or perhaps problematic) original firms in the locality? Further research is required in this area, and it is likely that both evolutionary and agglomerations effects work in combination in clusters.

Globalization

Those who are skeptical about the importance of these 'new industrial spaces' have pointed to the overwhelming power of the forces of globalization, and this forms another major focus within studies of economic restructuring in what has been termed the Global Production Network approach (Dicken, 2011). Globalization is – like most aspects of restructuring – a highly controversial and much disputed concept. The term is usually associated with the growing importance of multinational (or transnational) companies that operate in more than one country. However, large corporations such as Ford have long manufactured outside of their parent nation, producing goods for consumption in distant markets. Globalization therefore refers to a more recent process in which the operations of transnational corporations (TNCs), both in the spheres of production and distribution, are increasingly integrated on a global scale. Thus, products typically incorporate components made in diverse locations throughout the world. Furthermore, while the appearance of products such as cars may be tailored to meet the demands of particular local market conditions, this diversity is underpinned by many common elements (such as the floorpans and engines in the example of automobiles) to reduce production costs. The Global Production Network approach attempts to conceptualize these production systems as a complex web and not simply a linear chain. Studies of TNCs have focused upon how they manage to cope with the variety of institutional rules, regulations, and cultures to be found in differing nations and trading blocks, highlighting the fact that strategies are highly diverse (Wrigley et al., 2005). For example, an important strategy by US corporations in response to increasing competition from Asian corporations has been the development of 'modular production' (Sturgeon, 2002). This involves formalizing, through various standards and protocols, all the knowledge about products into particular stages of the production process, thereby enabling components to be outsourced throughout the world. Influential work has also focused upon global professional services firms showing how they adapt to different institutional contexts around the world (Faulconbridge, 2008). Globalization is also associated with the global spread of Western values of materialism and consumerism (*see* Globalization: Geographical Aspects).

Globalization has had a crucial effect upon economic restructuring. In particular, it has led to the emergence of so-called world cities (or global cities) – command centers of the new world economy that house the headquarters of the major corporations and the financial system (Sassen, 2001). Examples include New York, Los Angeles, London, Paris, and Tokyo. The occupational structures of global industries are such that these cities tend to be characterized by extremes of affluence and poverty. While few cities can claim true global status, there is a sense in which all urban centers are now global, for they are all affected by events outside their boundaries and they are all engaged in a fierce competition to attract mobile capital investment into their districts. Those who are skeptical of the clustering thesis argue that it is the elaborate structures of global corporate hierarchies that shape economies rather than industrial clusters in specific regions.

One approach to reconciling the conflict between the globalization and 'regionalist' schools of thought is through the concept of neo-Marshallian nodes (Amin and Thrift, 1992). This concept draws attention to the importance of geographical clusters of small firms but at the same time recognizes that they are frequently dependent upon the activities of major corporations located throughout the world (*see* Industrial Geography). Another influential article on the issue of local agglomerations of economic activity argued that clusters were advantageous for the circulation of 'buzz,' tacit uncodifiable knowledge, whereas 'global pipelines' were more suited to the dissemination of codifable knowledge (Bathelt et al., 2004).

However, some have argued that there cannot be any such simple equation of types of knowledge with geographical scale (Amin and Cohendet, 2004).

Services

Geographic studies of economic restructuring also focus upon the growth of services (Bryson et al., 2003) (*see* Service Economy, Geography of). Although services are notoriously difficult to define, a common theme is economic activity that does not involve the material transformation of a physical product. However services are defined, there is little doubt that service activity is dominant in terms of total employment and is on the increase in Western societies. Growth has been especially pronounced in financial and business services (sometime known as KIBS – knowledge intensive business services), retailing, leisure, and entertainment industries. Growth in services has tended to concentrate outside the traditional manufacturing centers of the Western economies (e.g., in the Sunbelt of the United States and in the South East of the United Kingdom). However, new telecommunications systems are enabling back-office activity (i.e., relatively routine service activity) to be decentralized away from areas of rapid service growth into older industrial areas and to call centers in the Asian subcontinent.

The restructuring of services has had profound effects upon the social structure of economies. In particular, it has led to social polarization with increases in both the affluent and the less well off. Whereas the declining traditional heavy manufacturing industries had substantial proportions of blue-collar, relatively well-paid, middle-income jobs, services tend to be associated with both relatively high-paying and low-paying jobs. Closely connected with the growth of services has been the increasing numbers of women in the workforce. Whereas men were employed in most of the relatively well-paid jobs in manufacturing, many of the new relatively low-paid service jobs (but also some of the well-paid service jobs) are held by women.

An important element of service growth has been the 'cultural' (or 'creative') industries such as advertising, design, entertainment, media, and publishing. Although highly diverse in character, these industries involve cultural products such as books, films, plays, advertisements, video games, and fashion clothes that are imbued with aesthetic and symbolic elements. The growth of creative industries has been closely linked with another worldwide trend in economic restructuring in recent years – the growth of culturally driven urban regeneration (Cooke and Lazzeretti, 2008). This trend involves not only the provision of concert halls, theaters, art galleries, and museums for the consumption of culture but also local creative quarters within cities for its production. The markets for these cultural products are typified by rapid changes in technology and fashion, and in response the cultural industries are often concentrated in close proximity in cities. Here the workers in these industries can utilize informal social networks and face-to-face contacts to pick the latest 'buzz' on the streets about changing fashions.

Knowledge-, Network-, and Digital-Based Economies

Another important strand of economic restructuring work by geographers has focused around knowledge economies (*see* Information Society, Geography of). This role of knowledge is reflected in the ever-increasing pace of technological change and the rapidity of changes in consumer fashions and styles. This means that, rather than competing by providing goods and services at the least cost, many firms increasingly compete on the basis of factors such as quality, performance, style, and design. Rapid developments in the sphere of telecommunications – the Internet, e-mail, teleconferencing, and various types of specialist computer networks – mean that knowledge can be disseminated much more quickly throughout corporate hierarchies or networks of collaborating firms. Central to these developments are new communications channels based around satellites and fiber optics that permit the exchange of greater and more complex amounts of information much more quickly over longer distances. The term 'network society' (Castells, 2009) has been coined to encapsulate all these various technologies that permit new socioeconomic and cultural practices such as social networking, online shopping, gambling, and open-source programing.

Despite much speculation, the full impact of the network society on economic restructuring is still unfolding. However, many developments are already evident: the emergence of mammoth digital companies; the rapid growth of on line retailing and the associated near disappearance of many city center retailers (including most notably music and book stores); illegal downloading that has challenged the business models of the popular music and movie industries; and new technologies that facilitate ever-faster responses to changing market demands with new products. The network society has also facilitated globalization and ever-increasing distances between the various elements in the production chain. In addition, it facilitates telecommuting by linking people into global information networks and enabling them to work in their own homes. While some claim new technologies will result in the dissolution of city life, paradoxically, cities remain highly prized as sites of innovation and creativity. The reason for this is (as noted before in services) that creativity in knowledge-based industries is fostered by having dense networks of people in close physical contact in places where trust and reciprocity can facilitate learning. Furthermore, much of the expensive infrastructure needed in the digital economy is heavily concentrated in urban areas.

Financialization

The specter looming over any discussion of economic restructuring is of course the current global recession following in the wake of the 2007–08 'credit crunch.' This crisis has been attributed to a process called 'financialization,' generally understood to indicate the growing dominance of finance capital markets on economic and social life (Pike and Pollard, 2010). However, compared with other aspects of economic restructuring this profoundly important process so far has

received relatively little attention from geographers (Pike and Pollard, 2010). In part, this neglect derives from the complex nature of financialization, since it involves a diverse set of financial initiatives – commonly known as derivatives – that attempt to convert virtually all types of economic transaction into 'packages' or 'bundles' (i.e., units of value that can be traded like a form of currency). The broad contours of the crisis engendered by this process have received enormous publicity and seem to be generally understood. As one example, US banks attempted to spread the risk of subprime mortgages lent to high-risk, low-income groups by using various financial instruments known as collateralized debt offerings or structured investment vehicles that were then traded to financial institutions throughout the world. The inability, or unwillingness, of those in the financial system to appreciate the risks involved in such instruments led to bad debts, risk aversion, a lack of liquidity in financial markets, state bailouts for banks, sovereign debt crises, fiscal austerity, and the stagnation of global production. Certainly, much of the process of financialization is hidden in complex corporate structures, opaque balance sheets, tax havens, and ever more complex financial schemes that are protected by armies of financial intermediaries who have enormous legal power and political clout. Understanding how this global financial system came into being, how in the wake of such disfunctionality it is able to reproduce itself, and, most importantly, how it might be regulated for the common good, are crucial tasks for future studies of economic restructuring (Dymski, 2010).

Overview

Rather than a single integrated approach, geographical studies of 'economic restructuring' represent a series of highly interconnected foci: on spatial divisions of labor, clusters, globalization, services, knowledge-, network-, and digital-based economies, and financialization. Nevertheless, these studies share a common desire to understand the many complexities of economic restructuring without recourse to simple laws of economics that are assumed to work in all places at all times. In addition, they are all committed to unraveling the crucial role of geography as a constitutive element of restructuring processes. Although many generalizations have been possible, the geographical manifestation of economic restructuring shows many complex, ever changing, and cross-cutting patterns. Although somewhat distinct from the approaches typically utilized by economists, restructuring studies have attracted considerable interest from a wide range of scholars in related disciplines such as urban studies, planning, sociology, and cultural theory. Studies of economic restructuring have forged especially strong linkages with economic sociology and political economy (*see* Economic Geography).

See also: Globalization: Geographical Aspects; Location Theory; Urban System in Geography.

Bibliography

Amin, A., Cohendet, P., 2004. Architectures of Knowledge: Firms, Capabilities and Communities. Oxford University Press, Oxford.
Amin, A., Thrift, N., 1992. Neo-Marshallian nodes in global networks. International Journal of Urban and Regional Research 16, 571–587.
Bathelt, H., Maskell, P., Malmberg, A., 2004. Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. Progress in Human Geography 28, 54–79.
Bryson, J., Daniels, P.W., Warf, B., 2003. Service Worlds: People, Organizations and Technologies. Routledge, London.
Castells, M., 2009. The Rise of the Network Society, vol. 1. Blackwell, Oxford, UK.
Cooke, P., Lazzeretti, L., 2008. Creative Cities: Cultural Clusters and Local Economic Development. Edward Elgar, Cheltenham.
Cooke, P., Morgan, K., 1999. The Associational Economy: Firms, Regions and Innovation. Oxford University Press, Oxford.
Dicken, P., 2011. Global Shift: Mapping the Changing Contours of the World Economy, sixth ed. Sage Publications, London.
Dunford, M., 2003. Theorizing regional economic performance and the changing territorial division of labour. Regional Studies 31, 839–854.
Dymski, G., 2010. Confronting the quadruple global crisis. Geoforum 41, 837–840.
Faulconbridge, J.R., 2008. Managing the transnational low firm: relational analysis of professional systems, embedded actors and time-space governance. Economic Geography 84, 185–210.
Granovetter, M., 1985. Economic action and social structure: the problem of embeddedness. American Journal of Sociology 91, 481–510.
Hall, P.A., Soskice, D. (Eds.), 2001. Varieties of Capitalism: The Institutional Foundations of Comparative Advantage. Oxford University Press, Oxford.
Hanson, S., Pratt, G., 1995. Gender, Work and Space. Routledge, London.
Klepper, S., 2010. The origin and growth on industry clusters. Journal of Economics 67, 15–32.
Marshall, A., 1890. Principles of Economics. Macmillan, London.
Martin, R.L., Sunley, P., 2003. Deconstructing clusters: chaotic concept or policy panacea? Journal of Economic Geography 3, 5–35.
Martin, R.L., Sunley, P., 2006. Path dependence and regional economic evolution. Journal of Economic Geography 7, 573–602.
Massey, D., 1995. Spatial Divisions of Labour, second ed. Macmillan, London.
Massey, D., Meegan, R., 1982. The Anatomy of Job Loss. Methuen, London.
Morgan, K., 1997. The learning region: institutions, innovation and regional renewal. Regional Studies 31, 491–503.
Peck, J., Theodore, N., 2007. Variegated capitalism. Progress in Human Geography 31, 731–772.
Pike, A., Pollard, J., 2010. Economic geographies of financialization. Economic Geography 86, 29–51.
Porter, M.E., 1998. Clusters and the new economics of competition. Harvard Business Review 11, 77–91.
Sassen, S., 2001. Cities in a World Economy. Pine Forge Press, London.
Saxanien, A., 1994. Regional Advantage; Culture and Competition in Silicon Valley and Route 128. Harvard University Press, Cambridge, MA.
Sayer, A., 1989. Post-Fordism in question. International Journal of Urban and Regional Research 13, 666–696.
Scott, A., 1988. New Industrial Spaces: Flexible Production and Regional Economic Development in the USA and Western Europe. Pion, London.
Storper, M., Walker, R., 1989. The Capitalist Imperative: Territory, Technology and Industrial Growth. Blackwell, Oxford, UK.
Storper, M., Venables, A.J., 2004. Buzz: face-to-face contact and the urban economy. Journal of Economic Geography 4, 351–370.
Sturgeon, T., 2002. Modular production networks: a new American model of industrial organization. Industrial and Corporate Change 11, 451–486.
Wrigley, N., Coe, N., Currah, A., 2005. Globalizing retail: conceptualizing the distribution-based transnational corporation. Progress in Human Geography 29, 437–457.

Economic Sociology

Neil Fligstein and Cyrus Dioun, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by V.A. Zelizer, volume 6, pp. 4128–4132, © 2001, Elsevier Ltd.

Abstract

Economic sociology is the study of how the material conditions of life are produced and reproduced through social processes. The field of economic sociology can be separated into the sociology of markets and the sociology of consumption. The sociology of markets views markets as socially constructed arenas where repeated exchanges occur between buyers and sellers under a set of formal and informal rules governing relations between competitors, suppliers, and customers. Markets are dependent on governments, laws, and cultural understandings that support market activity. There are four bodies of work that provide different views on the mechanisms by which markets are organized: networks, institutions, political economy, and the study of market devices and the performativity of economic ideas. The sociology of consumption situates consumption in the problem of what consuming things means to people. One core idea is that consumption is about how people constitute their lifestyles. Lifestyles can be constructed in emulation of other groups or instead by competition between social groups for status. Another aspect of consumption is how morals and meanings affect what goods can be bought and sold. Scholars have explored why some products can and cannot be commoditized and how that has evolved over time.

Introduction

Over the past 30 years, economic sociology has erupted into a vibrant and visible subfield as sociologists increasingly apply social theories to study the economy. Scholars have challenged the separation of economy and society to show how social arrangements constrain, enable, complement, and constitute economic phenomena. Building upon Granovetter's (1985) admonition that economic life is embedded in social life, economic sociologists have identified social mechanisms – such as networks, institutions, and morality – that provide the contexts, conduits, and categories integral to economic action. This article offers a definition of the field, suggests its main branches, and gives a flavor to how sociologists have tried to explain various economic phenomena.

Economic sociology is the study of how the production and consumption that define material life depends on social processes for their structure and dynamics. Practically, the field can be divided into two parts: the sociology of markets and the sociology of consumption. The sociology of markets is concerned with how market participants and stakeholders solve the problems of production and exchange. We consider four of economic sociology's most generative approaches to the sociology of markets – network analysis, institutionalism, political economy, and creation of market devices including scholarship that focuses on how economic theory is performed to create markets. The sociology of consumption focuses on what goods and services mean to consumers and how they use them as a means to distinguish themselves from others by constituting lifestyles. The authors discuss how scholars have used culture, meaning, status, and morality as analytic tools to distinguish the ways in which consumers accept or reject new products.

Sociology of Markets

Sociologists define markets as socially constructed arenas where repeated exchanges occur between buyers and sellers under a set of formal rules and informal understandings that govern relations among competitors, suppliers, and customers. These rules and understandings guide interactions, facilitate trade, define what products are produced, sometimes constitute the products themselves, and provide stability for buyers, sellers, and producers. In modern capitalism, markets also depend on governments, laws, and larger cultural understandings to operate (Fligstein, 2001).

One way to conceive of this is that participants in a market form reproducible role structures – that is, sets of recognizable participants who occupy certain positions (such as incumbent market leader or upstart challenger) and interact routinely over time (White, 2002). This view opens up the questions at the crux of the sociology of markets. What general problems must be solved for markets to emerge? How do role structures stabilize? How do markets evolve? How do they die or transform into other markets? What is the role of states in the construction and maintenance of markets? The various sociological theories the authors discuss propose alternative ways to think about how to answer these questions.

Markets qua self-reproducing role structures can emerge without state intervention. Before the modern era, extensive relationships between market actors made this possible – relationships that often depended on common kinship, ethnicity, or religion (White, 2002). These commonalities created trust, allowed actors to monitor one another, and offered ways to settle grievances (Granovetter, 1985). Informal rules and local cultures emerge within industries regarding the legitimate way to organize a firm, compete with other firms, and buy and sell products.

While neoclassical economics claims that price is an ever-emergent phenomena guided by the 'invisible' hand of the market, economic sociologists argue that price is the result of social relationships and shared understandings among market participants. Producers will pursue a number of social strategies such as product differentiation, collusion, and co-optation rather than take part in destabilizing price competition (Fligstein, 2001; White, 2002). Producers will mimic the

organizational forms and business models of legitimate firms in the field in attempts to win the support of customers and resource providers (DiMaggio and Powell, 1983). These mimetic pressures will often lead to the widespread acceptance of certain heuristics and models for economic behavior. Performativity theory suggests that market participants, often trained in the world's best business schools, import formal economic theories and models into their local market cultures (Callon, 1998). In industries where these models have gained widespread support market activists perform theory, turning the discipline of economics into 'an engine, not a camera' (Mackenzie, 2006).

For stable markets to emerge and distribute goods successfully, market actors must solve a range of problems. Their solutions produce the conditions that make a particular product market, industry, national economy, or regional economy distinctive. For example, in the early stage of a market's development, a product's qualities need to be defined. Governments, firms, and customers have input into the question of what count as safe food products, useful telecommunications standards, or tradable financial securities. Firms must also develop the competency and capacity to produce these goods. They need to be able to trust their suppliers, employees, and customers in order to secure inputs, labor, and capital. Thus, networks of social relationships help firms stabilize their environments and guarantee access to scarce resources. Firms must be able to forecast the activities and responses of their competitors in order to position themselves in the ongoing struggle for market share, and therefore need access to information on their competitors. Finally, by responding to directives from the government and trying to co-opt government agencies, firms can also secure their futures.

Theoretical Approaches: Network Analysis, Institutional Theory, Political Economy, and Market Devices and Performativity

The sociology of markets draws heavily on four theoretical approaches to offer concrete guidance about how to study market processes: network analysis, institutional theory in organizational studies, political economy and the literature on market devices, and ideas about the performativity of economic ideas drawn from the social studies of science. These four schools of thought offer different conceptual tools to make sense of what happens in markets. At the same time, there are overlaps in what each school stresses. For example, institutional theory and political economy both emphasize the state's role in constructing markets. Network analysis and institutional theory explore how the networks that constitute social structures transmit market information, define standard operating procedures, and generate trust. Market devices and the use of economics as a way to structure markets dovetail nicely with institutional theory's emphasis on the social constructedness of markets. Finally, governments frequently provide legitimation and even law and regulation for particularly market devices to evolve and be implemented.

Network analysis is a metaphor, a set of techniques to display data, and a way of studying the social mechanisms by which key market problems get resolved. If the sociology of markets' key insight is that market actors are embedded in social relationships that define who they are and what they do, then network analysis is the starting point for mapping those relationships (Powell and Smith-Doerr, 2005). Formal network-analysis techniques have demonstrated how the relationships among actors help explain what they will do: whom they will hire, and with whom they will trade goods and exchange capital. Networks also serve as conduits for information, operate to mitigate problems of trust, and control resource dependencies (Podolny, 2001). It is now firmly established that the connections among actors can affect prices, increase the probability that well-connected firms will survive, and help market participants compete and find customers and suppliers (Uzzi, 1996, 1997).

Institutional theory emphasizes informal understandings and cognitive frames that shape the social structure of markets. Institutional theorists often conceptualize markets as fields where firms watch one another, imitate one another, and build niches to reproduce their positions (DiMaggio and Powell, 1983). The observant and reflexive character of firm behavior can often mitigate competition and stabilize markets. Yet institutions are rarely fully settled or taken for granted. Even stable institutions are under ongoing contestation as various stakeholders such as firms, trade associations, unions, academic or professional experts, and community organizations construct resonant frames and mobilize resources to change state laws, reconstruct product categories, and shape industry 'best practices.'

Institutional theory also stresses how the formal laws, regulations, and actions of states and courts have profound effects on market structure (Dobbin, 1994). From this perspective, states and markets are joined at the hip. Moreover, firms can appeal to states for help. When incumbent firms successfully shape the rules governing their industry, they can reproduce their leading positions over time. Institutional theory stresses not only how states set rules and enforce sanctions, but also how they define what types of products are appropriate for exchange. Scholars with an institutionalist perspective also examine how the internal dynamics of states affect the way they intervene in markets. The structure of courts, ministries, and regulatory agencies, as well as struggles among parties and political factions, all affect state intervention.

Economic sociologists recognize the important role that the state plays in supporting and sanctioning market institutions (Crouch and Streeck, 1997; Hall and Soskice, 2001). States grant property rights, enforce contracts, regulate collaboration and competition, and establish standardized rules of exchange. Scholars have shown how state building and market building are coevolutionary processes and that political stability is a precondition for market growth (Crouch and Streeck, 1997). State prohibition of certain goods such as drugs, sex, and human cadavers relegate buyers and sellers of these products to the unstable and resource-poor black market (Fourcade and Healy, 2007; Beckert and Wehinger, 2013).

The state has the ability to use policy to shape both the competitive landscapes of industries and organizational forms of firms. Fligstein (1990) shows how in the late nineteenth century the American government used antitrust laws to foster competition and break up cartels. Similarly,

Dobbin (1994) shows how government policy in the early twentieth century played a critical role in shaping the organizational forms of railroad companies.

Political economy has pioneered thinking about the linkages between states, law, and markets and the historical emergence of systems of governance. The political economy literature on 'comparative capitalisms' – the comparative study of capitalist arrangements and their effects on various outcomes, including economic development – is a fundamental part of the sociology of markets (Hall and Soskice, 2001). This literature has revealed that the relationships among governments, workers, and capitalists have varied dramatically over time and geography, and that economic trajectories are often culturally and nationally specific. Markets are social and political constructions reflecting a country's culture, its history of class relations, and the various interventions its governments have carried out through history (Fligstein, 2001). In this view, each national system of capitalism forms an integrated whole: an enmeshed set of institutions such as systems of labor relations, modes of capital allocation and corporate control via securities markets and bank lending, systems of education and training, tax systems, and state involvement and ownership in key industries.

The political economy literature has been the most forthright about trying to understand which systems produce the most sustained economic growth. Scholars have documented that different national systems of capitalism exist and have debated how to fit them into categories, such as on the basis of national protectionism, levels of state intervention, and modes of capital allocation to firms (Crouch and Streeck, 1997). Yet, beginning with the collapse of communism in the late 1980s, a whole series of studies began to appear asserting that differences among national capitalist systems would soon disappear. These studies predicted that as global capitalism spread and barriers to international trade and capital flows crumbled, production and consumption would shift to those countries with the most efficient market institutions, bringing governments to their knees and forcing rapid convergence to a generic liberal model of capitalism (Strange, 1998). However, this prediction proved false. Research suggests that national systems of capitalism are resilient, even in the face of political and economic crises. This has led other scholars to conclude that key features of these systems must enable firms, states, and workers to adapt efficiently to new challenges (Hall and Soskice, 2001).

Scholars have also sought to understand economic transformation in the postsocialist world and other developing economies. They have documented the collapse of institutions in Eastern Europe and Russia and have analyzed the features of the systems that have taken their place. China has also been the focus of sustained study (Nee and Opper, 2012). The main issues in this literature involve the role of the state in the market. Research shows that states with disorganized bureaucracies and poorly paid, incompetent bureaucrats are more prone to rent-seeking than those with educated and better paid staff. Scholars also agree that defining property rights clearly and assuring the rule of law are important to development. But establishing such institutions is difficult to do from above; they usually arise dynamically from the demands of firms and citizens. Also controversial is the relationship between economic development and state investment in particular firms and industries. States have done this successfully in some cases, but failed miserably and wasted scarce resources in others (Evans, 1995).

The literature on market devices focuses on a problem that neither network analysis nor political economy is very good at treating. For markets actually to work, products need to be created, evaluated, and priced. The institutions that do this are 'market devices,' widely accepted rules of thumb, tools, technologies, and theories that help actors value goods, weigh options, and follow corresponding strategies in the marketplace (Biggart and Beamish, 2003). The literature includes at least two approaches to thinking about market devices. One originates with the idea that we need conventions to make judgments of price and quality simpler. For example, if a bank wants to loan money, it must judge each customer's creditworthiness. Over time, ratings agencies have developed different quantitative yardsticks that allow banks to judge the riskiness of loans. In commercial lending, these ratings involve putting letters from AAA to D on corporate bonds (Carruthers and Stinchcombe, 1999).

The problem of price setting is a very general one with far-reaching implications for firm behavior. Deciding how much to pay someone, whether or not a bottle of wine is worth $50, and how to make such judgments require a whole social infrastructure (Karpik, 2010). Many of these comparisons involve rankings and other quantitative indices. Once in place, such market devices can take on a life of their own. Actors in markets use them not just to decide what to buy or sell, but to measure themselves and their performance against others. Rankings can push firms to reevaluate their strategies and tactics in order to improve their standing.

This brings us to a second way in which market devices get constructed: via the performativity of economics (Callon, 1998). One area of work that has been particularly productive has examined how principles and quantitative models from economics can structure financial markets. For example, there is a body of theory in economics that discusses how different types of auctions work. Not surprisingly, sellers deploy this knowledge to auction hard-to-value objects. Thus, the science of economics actually makes the market for such objects possible. In the case of futures contracts traded on exchanges, derivatives traders' use of the Black–Scholes–Merton (BSM) theory for pricing derivatives actually caused prices to approach the values that BSM predicted (MacKenzie and Millo, 2003). Scholars term the implementation of economic ideas to structure markets 'performativity.' This label implies that one self-consciously takes an idea and puts it into practice: literally, one performs it.

Financialization and Globalization

One empirical arena to view the analytic power of the perspectives that make up the sociology of markets is the study of financialization and globalization. Financialization is a set of related phenomena that reveal "the increasing role of financial motives, financial markets, financial instruments, financial actors, and financial institutions in the operations of domestic and international economies" (Epstein, 2005; p. 3). It

comprises the growing dominance of finance tools in the management of firms, the ability of financial markets to dictate what firms and governments do, and the growth and importance of the international financial system in the distribution of capital around the world (Krippner, 2011; Davis, 2009). As such, making sense of it requires the use of all the perspectives just discussed.

The growing dominance of finance tools corresponds with the rise since the 1980s of the shareholder-value view of the firm, particularly in the United States (Useem, 1993). These tools allow managers who view the firm through the lens of the balance sheet, such as chief financial officers (CFOs), to maximize shareholder value (Zorn, 2004). The basic idea of the shareholder-value view is that firms exist to provide returns to shareholders. Shareholders of publicly held corporations care primarily about two metrics: (1) share price and (2) the ratio of profits to assets. With this in mind, managers use a range of tools to increase these metrics, including financial engineering of balance sheets, outsourcing, downsizing, investing in technology, cutting salaries and benefits, and increasing the working hours of the employees who remain. Upper level managers frequently receive shares and bonuses for meeting financial targets.

The rise of the shareholder-value view, and the concomitant ascendance in power of finance-oriented managers (such as CFOs) originated in firms' changing relationship with the financial markets. Scholars have documented that during the 1980s' merger movement, institutional investors pushed management teams to increase their profitability. When managers resisted, their firms became takeover targets. Since then, the relationship between the financial markets and the largest corporations has been symbiotic (Useem, 1993; Davis, 2009).

Corporations' reorientation toward shareholder value and the financial markets is not the only manifestation of financialization, however. Financial markets themselves have grown enormously and become integrated globally (Knorr Cetina and Bruegger, 2002; Leyshon and Thrift, 2007). Net transnational flows of bonds, bank capital, investment capital, and derivatives each dwarf the revenues of major corporations and even many states. The creation of complex financial products and the globalization of finance have produced an important new literature in economic sociology. Scholars focusing on instruments tend to be grounded theoretically in science and technology studies and the social studies of science in general, in actor–network theory, and in a strong appreciation for the technical details of the machines that make markets work (MacKenzie, 2006; Knorr Cetina and Bruegger, 2002; Callon, 1998).

Scholars have struggled for years to understand how national systems of governance respond to the pressures produced by financialization and globalization. Many consider it inevitable that global economic pressures, including the whim of sovereign-bond markets, will limit national governments' degrees of freedom. This will force governments not only to deregulate product markets but to cut back on welfare expenditures and labor protections. However, evidence for this 'race-to-the-bottom' proposition is mixed at best (Strange, 1998). While there have been some changes in the way that governments regulate, tax, and spend in the past two decades, there is no consistent trajectory across countries. Many of the national capitalisms have changed in response to these pressures, but mostly by modifying their existing national models (Hall and Soskice, 2001).

Sociology of Consumption

The sociology of consumption is focused on what goods and services mean to people (Zukin and Maguirre, 2004). Consumption is most frequently studied by focusing on how what one consumes is a statement about lifestyle. Studies focus on how baskets of goods are expressions of who one is and how one wants others to think of him or her (Bourdieu, 1984). Most scholars emphasize that lifestyle is an ongoing construction, one that can turn into an escalating arms' war as people compete over important high-status goods (Schor, 1999). One important strand of this work focuses on how advertisers push this race along by promoting the consumption of goods and services to people in order to get them to consume more and thus, obtain more social status (Slater, 1997). A second focus in research is how goods are constructed by cultural categories, categories that often imply something about societal conceptions of morality. So, for example, the boundaries between things one can sell and cannot sell imply the moral edge of society. One focus of such studies is how some products, which were previously immoral, come to be seen as moral.

Consumption is often an attempt to secure social status and produce a lifestyle (Bourdieu, 1984; Elias, 2000). Historical sociologists have documented how this process occurred in early modern Europe, when the newly rich bourgeoisie tried to mimic the nobility. This led the nobility to shift the markers of high social status continuously in order to retain their edge over the rising bourgeoisie. Elias (2000) shows that this lifestyle competition eventually came to include both the middle and working classes. More recent work has tried to view lifestyle as not just competition whereby those at each level of society try to emulate those above them but one where lifestyle can be oppositional. The strongest proponent of this view is Bourdieu (1984), who shows that those lower down the income and education hierarchy find themselves as losers in a process where money can buy status, and as such, decide to consume and value different things.

In the past 20 years, scholars have emphasized how such status hierarchies remain dynamic and continuously shifting. One important place to observe such processes is in the race for positional goods like cars, good schools, fashion, and houses. As income inequality has increased in America, the price of positional goods has increased. In order to keep up, working-class Americans have taken on debt to finance their consumption and 'keep up.' Going into debt to keep up with the Joneses, or leveraging home equity to finance one's children's education, is viewed as a reasonable life decision (Schor, 1999).

The question of how consumers think about goods and services has produced much interesting work on household budgeting, gifts, and the line between which products are considered moral to consume and which things should not be

for sale. Zelizer (1997) has shown that households do not view their income as fungible (i.e., a dollar spent on one thing can be just as easily spent on another) but instead view income as pots of money dedicated to particular categories of consumption. So, frequently things like inheritances are kept as a lump sum of money put to a family need like a house or an education.

The study of the line between which products are morally acceptable to buy and sell in markets and which ones are not has been a productive field of study (Fourcade, 2011; Fourcade and Healy, 2007). One important question is how do such boundaries change over time as new problems come up to be solved and therefore change the moral calculus applied to a particular 'thing.' Life insurance offers examples of the marketization of new goods. When life insurance was first introduced, people found it morally repugnant. The idea of paying for such insurance seemed like a macabre commoditization of death that verged on gambling. But over the course of the nineteenth century, life insurance became viewed by middle class families as a prudent way to provide for one's kin in the event the main bread winner passed away (Zelizer, 1979).

More recently, during the 1980s and 1990s, selling one's life insurance policy to a third party went from being widely considered morally abhorrent – because it gave someone else an incentive to wish for one's early death – to being the basis for a new market in 'viaticals.' The AIDS epidemic moved forward this transformation in markets: as AIDS patients faced rising health care bills and poverty, they sought to sell their own life insurance policies. Here, the moral discourse in favor of a dignified death came to trump the moral discourse against betting on death, producing a new market (Quinn, 2008).

Conclusions

This panoramic review of economic sociology has touched on many of the major themes in this growing literature. Economic sociology has laid down a tapestry of compelling theoretical perspectives and empirical work. They have been used productively to understand the formation of new markets and the role of states in creating markets. They have pushed scholars to consider the social structuring of markets and the meanings that have come to organize and constitute the existence of new markets. The sociology of consumption has provided a theoretical and empirical basis to consider why people consume. It has pushed scholars to think about how culture and morality frame such decisions and structure the existence of new markets.

Economic sociology has added richness to the ability to understand the behavior of individuals, households, firms, markets, and national systems of capitalism. At the micro-level, it shows how ongoing interpersonal relationships play an important role structuring various economic processes. At the meso-level, it describes how established industries rely on institutional architecture, formal and informal rule systems, to make life more predictable and mediate problems related to production, competition, valuation, and exchange. At the macro-level, it reveals how historical differences in state laws and national cultures foster different forms of capitalism with varying conceptions of efficiency.

See also: Capitalism: Global; Consumption, Sociology of; Exchange: Social; Institutions; New Institutionalism in the Analysis of Complex Organizations; Organizations, Sociology of; Social Networks.

Bibliography

Beckert, J., Wehinger, F., 2013. In the shadow: illegal markets and economic sociology. Socio-Economic Review 11, 5–30.
Biggart, N., Beamish, T., 2003. The economic sociology of conventions. Annual Review of Sociology 29, 443–464.
Bourdieu, P., 1984. Distinction: A Social Critique of the Judgment of Taste. Harvard University Press, Cambridge, MA.
Callon, M., 1998. The Laws of the Markets. Blackwell Publishers, London.
Carruthers, B., Stinchcombe, A., 1999. The social structure of liquidity: flexibility, markets, and states. Theory and Society 28, 253–282.
Crouch, C., Streeck, W., 1997. Political Economy of Modern Capitalism. Sage, London.
Davis, G., 2009. Managed by the Markets. Cambridge University Press, New York.
DiMaggio, P., Powell, W.W., 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. American Sociological Review 48, 147–160.
Dobbin, F., 1994. Forging Industrial Policy: The United States, Britain, and France in the Railway Age. Cambridge University Press, New York.
Elias, N., 2000. The Civilizing Process. Sociogenetic and Psychogenetic Investigations, Revised ed. Blackwell, Oxford.
Epstein, G., 2005. Financialization and the World Economy. Edgar Elgar, Cheltenham, UK.
Evans, P., 1995. Embedded Autonomy. Princeton University Press, Princeton, NJ.
Fligstein, N., 1990. The Transformation of Corporate Control. Harvard University Press, Cambridge, MA.
Fligstein, N., 2001. The Architecture of Markets. Princeton University Press, Princeton, NJ.
Fourcade, M., 2011. Economic valuation and the nature of "nature". American Journal of Sociology 116, 1721–1778.
Fourcade, M., Healy, K., 2001. Moral views of market society. Annual Review of Sociology 33, 285–311. Annual Review Press, Stanford, CA.
Granovetter, M., 1985. Economic action and social structure: the problem of embeddedness. American Journal of Sociology 91, 481–510.
Hall, P.A., Soskice, D.W., 2001. Varieties of Capitalism: The Institutional Foundations of Comparative Advantage. Oxford University Press, Oxford, UK.
Karpik, L., 2010. Valuing the Unique: The Economics of Singularities. Princeton University Press, Princeton, NJ.
Knorr Cetina, K., Bruegger, U., 2002. Global microstructures: the virtual societies of financial markets. American Journal of Sociology 107, 905–950.
Krippner, G., 2011. Capitalizing on Crisis: The Political Origins of the Rise of Finance. Harvard University Press, Cambridge, MA.
Leyshon, A., Thrift, N., 2007. The capitalization of almost everything. Theory, Culture, and Society 24, 97–115.
MacKenzie, D., 2006. An Engine. Not a Camera. MIT Press, Cambridge, MA.
MacKenzie, D., Millo, Y., 2003. Constructing a market, performing theory: the historical sociology of a financial derivatives exchange. American Journal of Sociology 109, 107–145.
Nee, V., Opper, S., 2012. Capitalism from Below: Markets and Institutional Change in China. Harvard University Press, Cambridge, MA.
Podolny, J.M., 2001. Networks as the pipes and prisms of the market. American Journal of Sociology 107, 33–60.
Powell, W.W., Smith-Doerr, L., 2005. Networks and economic life. In: Smelser, N., Swedberg, R. (Eds.), The Handbook of Economic Sociology. Princeton University Press, Princeton, NJ, pp. 373–403.
Quinn, S., 2008. The transformation of morals in markets: deaths, benefits, and the exchange of life insurance policies. American Journal of Sociology 114, 738–780.
Schor, J., 1999. The Overspent American. Harpers Press, New York.

Slater, D., 1997. Consumer Culture and Modernity. Polity Press, Cambridge, UK.
Strange, S., 1998. The Retreat of the State. Cambridge University Press, Cambridge, UK.
Useem, M., 1993. Executive Defense. Princeton University Press, Princeton, NJ.
Uzzi, B., 1996. The sources and consequences of embeddedness for the economic performance of organizations: the network effect. American Sociological Review 61, 674–698.
Uzzi, B., 1997. Social structure and competition in interfirm networks: the paradox of embeddedness. Administrative Science Quarterly 42, 35–67.
White, H.C., 2002. Markets from Networks: Socioeconomic Models of Production. Princeton University Press, Princeton, NJ.
Zelizer, V., 1979. Morals & Markets: The Development of Life Insurance in the United States. Columbia University Press, New York.
Zelizer, V., 1997. The Social Meaning of Money. Basic Books, New York.
Zorn, D., 2004. Here a chief, there a chief: the rise of the CFO in the American firm. American Sociological Review 69, 345–364.
Zukin, S., Maguirre, J.S., 2004. Consumers and consumption. Annual Review of Sociology 30, 173–198. Annual Review Press, Stanford, CA.

Economic Transformation: From Central Planning to Market Economy

Gerard Turley, J.E. Cairnes School of Business and Economics, National University of Ireland, Galway (NUI Galway), Ireland

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by Z. Cséfalvay, volume 6, pp. 4141–4146, © 2001, Elsevier Ltd.

Abstract

This article outlines the economic transformation from central planning to the market system that socialist countries embarked upon at the end of the twentieth century. Over 30 countries, comprising Central and Eastern European countries, republics of the former Soviet Union, China, and Vietnam, undertook the transition from plan to market and from socialism to capitalism. Key elements of the transformation are discussed, namely, theories and paradigms, central planning and initial conditions, reforms and economic policies, institutions, performance and outcomes, and, finally, lessons and an assessment of this unique social science experiment.

A new subbranch of economics emerged at the end of the twentieth century. It arose from the collapse of the Soviet socialist system (and the earlier reforms in China and Vietnam) and the subsequent transformation from a centrally planned economy to a market economy. It is known as transition economics where transition is defined as the social, political, and economic transformation from a command system characterized by state ownership, collective action and central planning to a market system characterized by decentralized decision-making arrangements and private sector market transactions. Although it overlaps with other branches of economics, most notably Comparative Economic Systems and Development Economics, the study of transition economics was sufficiently distinct to warrant its own subbranch within the discipline of economics. This short article will focus on the economic transformation dimension of transition economics, and, in particular, examine a number of key elements, namely paradigms, central planning and initial conditions, reforms and economic policies, institutional change, performance and outcomes, and, finally, lessons of the economic transformation from plan to market.

Leaving aside the debate over whether 'transition' or 'transformation' was the most appropriate term, the change from central planning to the market economy and from socialism to capitalism has been described as one of the most important economic, social, and political experiments of modern times. Indeed, many view the transition from plan to market as one of the two most important economic events of the last century, the other being the Great Depression of the 1930s. Just as the causes, consequences and lessons of the Great Depression are still matters for ongoing discussion and research (and particularly since the 2008/09 Great Recession), the transformation experience from plan to market is likely to attract debate and controversy for decades to come. At the peak of the socialist system, one-third of the world's population lived in socialist countries. However, within a short space of time these countries had abandoned state socialism and embraced the capitalist system with its market economy. Although the transformation was remarkably peaceful, it has proved to be both costly and painful for many countries, regions, and sections of society. Many citizens in transition economies (TEs) witnessed an erosion of living standards and public services. For others, it has been a time of extraordinary opportunities, in both political and financial terms.

Why study the economic transformation from central planning to market economy? For one, it gives us a rare insight into the alternative economic systems, namely the central plan and the market. Two, we get a better understanding of what is involved in systemic and large-scale institutional change. Three, it offers an insight into policy reform and the politics of economic reform. Four, it allows for a timely reappraisal of some key economic debates such as the Washington Consensus, the debate over markets versus states, the importance of institutions and contracting, and the role of privatization and small and medium-sized enterprise (SME) development. Five, concepts and economic ideas studied in transition such as state capture and corruption, asset stripping and tunneling, the soft budget constraint and ratcheting, mass privatization and dual-track price liberalization, shock therapy, and incrementalism have relevance far beyond economic transformation and transition countries. Finally, it is now generally accepted that the economics profession has a great deal to learn from this transformation experience and that these 'lessons' influence and, over time, improve our understanding of economic systems, capitalist economies, (new) comparative economics, and mainstream economics, and provide a valuable insight into other areas of economics (Djankov et al., 2003; Ellman, 2013).

Theories and Paradigms

Strictly speaking, there is no 'theory' of transformation from a centrally planned system to a market system. In the late 1980s, when change was imminent in the Eastern bloc countries (and was already underway in the East Asian countries of China and Vietnam) there was no theory of transformation from plan to market that policymakers and supporters of reform could adopt. What evolved over time are two different approaches to understanding the transformation process from a centrally planned economy to a market-based economy. The conventional paradigm is the Washington Consensus approach to economic transition and the alternative paradigm is the institutional-evolutionary approach (see Roland, 2000; Turley

and Luke, 2010). The Washington Consensus approach dominated in the early years of transition and had the support of the international financial institutions (IFIs). This orthodox or standard vision, especially popular in the Anglo-Saxon world – seen by many as an expression of the neoclassical mainstream – views transformation as a reform process emphasizing the universality of the laws of the market and the undoubted economy-wide efficiency gains accruing from the standard policy prescriptions of the trinity of liberalization, stabilization, and privatization. This blueprint for transformation, based on the spontaneity of markets and private enterprise, traditional neoclassical price theory and Walrasian general equilibrium theory, promoted the primacy of policy reforms and economic fundamentals and the replication or transplantation of international best-practice institutions of the West to the socialist countries of the time. Although the Washington Consensus emerged from a different set of conditions, it argued that these one-size-fits-all market-oriented reforms are appropriate to any setting, socialist or otherwise. A knowledge or experience of the state socialist system and the centrally planned economy is not required. Regional studies and area specialists emphasizing the importance of historical, cultural, and socioeconomic factors were dismissed. Instead, liberal economic reform strategies are implemented along a scorched-earth approach, with textbook reforms being designed by technocrats as the agents of social change and introduced as rapidly and comprehensively as possible in view of the reform complementarities that exist. Not surprisingly, in terms of the speed of the radical reforms, this approach is often referred to as the 'big bang' or 'shock therapy' view of transition. This also applies to the economic role of the state where what is required is a depoliticization of the economy, a break of the nexus between government and business, and a dismantling of the state, or, according to critics of the Washington Consensus approach, in the extreme case of neoliberal market fundamentalism, state desertion.

The institutional-evolutionary paradigm viewed transformation as a large-scale institutional change where the focus is on the institutional underpinnings of capitalism appropriate to the specific conditions of each country and in accordance with the initial conditions at the outset of transition. Following in the traditions of Schumpeter and others, this approach is inspired more recently by evolutionary theory, institutions, and the contribution of new institutional economics. It is critical of the revolutionary vision of transition and, instead, views transformation as an evolutionary process involving systemic change in the face of great uncertainty and complexity, based on contracting, noncooperative game theory in modern microeconomics and the second-best theory. Here, the focus is on the gradual development of the institutional supports or arrangements for a market economy, accepting the dangers of institutional voids and that not all existing or inherited institutions, organizational forms or social capital are redundant. Transitional second-best institutions and the preservation of social capital can be both worthwhile and necessary in order to prevent further economic disruption. The institutional-evolutionary perspective is a bottom-up view focusing on the institutional infrastructure of market economies, the importance of social norms and the organic development of the private sector. Markets and economic agents do not exist in a vacuum but in an institutional framework that facilitates exchange and interaction. Institution building and the provision of a framework for well-functioning market structures and organizations are the focus of this approach and it argues for the gradual or incremental implementation of sequenced reforms in order to ensure growing support for policies. Followers of the institutional-evolutionary perspective were critical of the Washington Consensus emphasis on economic, and particularly, monetary variables at the expense of institutional and social variables. Similar criticisms were made of the mathematical precision and overly scientific nature of the economic orthodoxy and the narrowly defined market reforms that follow. In the institutional-evolutionary paradigm, although there is recognition for the need to reduce the role of the state, the emphasis is on a reconstituted state and improving state capacity as opposed to a weakened state. It also stresses the path dependency of system development and is mindful of the historical continuity and the communist heritage unlike the ahistorical, tabula rasa Washington Consensus approach. Critics of the institutional-evolutionary perspective claim that the time lost in waiting for the evolution of the institutions will be too costly, both in political and economic terms. Others concede that although history matters, it is not a reason to oppose, delay, or reverse reforms. Furthermore, policymakers often argued that the political and administrative collapse that accompanied economic transformation meant that the strategy as espoused by the followers of the institutional-evolutionary approach, namely a gradual, sequenced approach, was not possible.

Central Planning and Initial Conditions

Close on 30 countries, from East Germany in the West to China in the East, set out to transform, rather than reform as was the case in the pretransition era, their economic system at the end of the twentieth century. In terms of income per capita, the majority of socialist economies at the beginning of transition were neither low-income countries as prevalent in the developing world of Africa and Southern Asia nor were they high-income countries as predominantly found in North America and Western Europe. Many tended to be middle-income countries (main exception were the Central and East Asian countries, including China) but, often, in later years, with low and declining growth rates. In terms of the economic structure, socialist countries tended to be, on average, urbanized and industrialized with high investment rates and equally high and often inefficient energy usage (World Bank, 1996). In terms of human resources and social indicators, although socialist countries had extensive welfare systems as characterized by relatively even distributions of income, high rates of life expectancy and high literacy rates, there was evidence of a deterioration in living standards as, for example, the economies of the Soviet bloc began to stagnate in the later stages of the Brezhnev era.

The socialist system had a number of distinguishing economic features. These included the centrally planned mechanism of resource allocation, collectivism, and state ownership, and the dominance of industry and investment over services and consumption. Despite the common legacy of one-party rule, public ownership and bureaucratic

coordination, the former socialist countries were not all the same at the outset of transition. Initial conditions varied greatly, not just between the countries of East Asia (including China), Central and Eastern Europe, and the republics of the former Soviet Union (FSU) (Russia included) but even within these groups of countries. Understanding these cross-country variations in circumstances is useful when explaining, at a later stage, the issue of the large variation in economic performance and outcomes that emerged between countries once transformation began. Differences in the starting point may also partly explain the different reform strategies adopted by the various TEs, acknowledging that other factors including political constraints, electoral support and legitimacy, ideologies, institutions, and pressure (or support) from the West also played a role.

State-owned enterprises (SOEs) and firms in market economies are very different production units. Whereas firms in market economies are autonomous decision-making units, SOEs were part of a hierarchical bureaucratic structure comprising central planners, ministries, and enterprises. SOEs, whose aim was to fulfill the plan, were notoriously large, faced little or no domestic or foreign competition and were responsible for the provision of social services. Due to the problems of securing input supplies the vertically integrated enterprises often hoarded labor and materials. Arising from this uncertainty of supply and the endemic shortages that prevailed, many state enterprises were forced to meet their own input needs while also operating in fits and starts. With no alternative supplier, no competition from other enterprises and with no attention or concern for consumer choice or product quality (except in the high-priority sectors) SOEs were far removed from the normal competitive pressures and price signals of the marketplace. Problems at the enterprise level varied from, at best, misdirected (at worst, perverse) incentives to a lack of innovation and risk-taking resulting in inefficiencies, inertia, and complacency. Prices were distorted with information flows inadequate and often inaccurate: the very same information problem that Friedrich von Hayek exposed as far back as the 1940s (Hayek, 1944).

There were attempts, in pretransition days, to decentralize the economic system and, in particular, to reform SOEs. These were mainly targeted at increasing the role of profits, allowing for greater flexibility and autonomy for management, introducing new bonus schemes, establishing Western-style taxes, reducing the number of targets, charging for the use of capital, and liberalizing foreign trade arrangements. Despite these changes, the enterprise sector (as with the socialist system within which it operated) was in crisis on the eve of transformation.

Economic Policies and Reforms

One of the main problems devising policies suitable for post-socialist transformation was the absence of any historical precedence or theoretical framework. Although there were several examples of countries in the twentieth century that had adopted reforms in the face of systemic change and great uncertainty the transformation from a centrally planned economy to a market economy was unprecedented as it required, unlike in many previous cases where dormant or distorted markets had to be revived or corrected, building markets, institutions, and the capitalist system from scratch. Despite this lack of a blueprint for TEs, reforms in the 1990s in the Soviet bloc countries was dominated by the policies of stabilization, liberalization, and privatization, as strongly advocated by the IFIs. It was argued that the distortions and inefficiencies of the old system would be corrected by price, trade, and free-entry liberalization combined with the necessary macroeconomic stabilization. The other part of the 'holy' trinity was the transfer of ownership from state to private hands by a process of privatization. These initial-phase policies (including the provision of a social safety net) formed the basis of the reform package advocated by mainstream economists and the IFIs. The structural and institutional reforms that comprise the second-phase policies include, among others, enterprise restructuring and competition policy, banking and financial sector reforms, labor market and fiscal reforms, and social security and pension reforms.

For some, privatization was the key element in transition. State ownership was a central feature of the socialist system. In transition, privatization was necessary to transfer ownership from public to private hands. It was argued that privatization would, inter alia, eliminate the tendency for SOEs to fulfill political objectives, improve the quality of managers, encourage greater financial discipline, raise much needed revenue for government, and, when combined with other reforms, result, eventually, in an efficient ownership structure. As the 1980s saw a worldwide increase in privatization, it was inevitable that the TEs would follow the privatization route. Even before the Soviet Union broke up, enterprise managers were seeking changes in ownership structures during the so-called 'spontaneous privatization' period. Aside from differences in the number of enterprises undertaking privatization, the form or method of privatization often differed. With inadequate savings, poorly functioning capital markets, 'fuzzy' property rights, underdeveloped legal systems, a plethora of unviable SOEs, a weakened state capacity, public opposition to privatization and reluctant foreign investors, the normal forms of privatization, including individual sell-offs of publicly owned firms, were inadequate and, in many cases, inappropriate. Other issues at the outset of privatization were the speed (versus the quality) of a change in ownership, the role of restitution to precommunist owners, the sequencing of privatization vis-à-vis demonopolization and restructuring, the role of investment funds and the involvement of foreign investors, and the relevant importance of privatization as against the development of *de novo* private firms. There was also some debate on how to deal with the winners and losers that result from privatization and how to provide adequate protection for minority shareholders. Given the scarcity of private capital and the difficulty in valuing enterprises on the eve of transition, a new form of privatization, called mass privatization schemes was initiated whereby shares in enterprises were sold for free or at a nominal fee to a country's citizens, via vouchers, coupons, or certificates, and tradable at auctions. Many TEs, and, in particular, most of the FSU countries adopted the mass privatization schemes as it was viewed as the most politically acceptable. Although supporters of rapid privatization cited the usual economic arguments for doing so,

often the politics of privatization was invoked. Rapid privatization would ensure a depoliticization of economic activity and, in turn, prevent a return to the old state-dominated socialist system, while at the same time, create a demand for market-supporting institutions. However, it became evident in the mid- to late 1990s that all was not well with privatization and that a change in ownership alone was not sufficient to guarantee improvements in enterprise performance. In particular, there were problems with the process and outcomes of privatization in many TEs, including the high profile schemes in the Czech Republic and Russia. Later, it became clear that, among other things, the form that privatization takes matters (despite, at the early stages of transition, advocates of privatization invoking Coase's theorem claiming that the question of how to privatize, and the initial allocation of property rights, do not matter). Well-functioning legal systems and capital markets also matter. Other elements of enterprise reform, including proper corporate governance mechanisms, hardening of the budget constraint, market structure and competition policy, ownership types (foreign vs domestic, concentrated vs dispersed, privatized vs *de novo*) were also found to be important in the empirical studies on the determinants of enterprise restructuring and performance.

Throughout the transition era, debate continued over the most appropriate sequence to reforms, and, in particular, the order of enterprise restructuring versus privatization, enterprise reform versus financial/banking sector reform, institutional reform versus market reforms, (stabilization versus) liberalization versus demonopolization and competition, not to mention the order of economic reform versus democratization. Inevitably, those reforms that were easiest to implement, attracted less opposition from powerful vested interests and were popular in policy circles (including the IFIs of the Bank and Fund) were advocated and implemented first. These reforms included price liberalization and macroeconomic stabilization, with the latter having some notable country setbacks along the way. Those reforms that took longer to implement, attracted most opposition from interest groups and were less frequently championed in policy circles were slower to emerge. These included changes to governance (both at national and corporate levels) and financial sector reform.

Other policy debates that raged throughout the transition era include the relative importance of market reforms (as opposed to initial conditions, external factors or institutional arrangements) in accounting for economic performance, the importance of political constraints, the tightness of monetary policy, the validity of the 'uniqueness' argument in designing country-specific policies, the importance of the lure of the EU (the 'return to Europe' commitment), the role of foreign aid and, not least, the case for and against an investment strategy and an industrial policy. One thing that is clear after decades of reforms is that expectations, after the collapse of the socialist system, with respect to policy reform were unrealistic and overly optimistic. Finally, a number of patterns of reform have emerged since transformation and economic reforms began. Most progress has been made in the initial-phase reforms, and by the Central and Eastern European (CEE) countries and Baltic states whereas much slower progress has been made in the second-phase reforms and by the Commonwealth of Independent States (CIS) countries. Neither China nor Vietnam followed the conventional reform strategy but have implemented, with greater success in terms of outcomes, major changes to their economic system. In terms of enterprise reforms, the big difference between China (where the dual-track approach dominated) and the Soviet Union (where privatization of state enterprises was the main strategy adopted) was the growth of the nonstate sector in China. In particular, rural township and village enterprises (TVEs) have contributed enormously to the success of Chinese reforms. Although there were different types of rural TVEs, they were, in terms of their ownership structure, collectively owned public enterprises under the control of local, as opposed to central, government officials. It was these community-owned local government firms, and not the *de novo* private firms as in CEE countries, that were the initial engine of growth in China's enterprise sector. In more recent years, many TVEs were subject to substantial restructuring, with many changing ownership into privately owned firms.

In summary, although the debate with respect to the speed, nature, and sequencing of reforms dominated the early years of economic transformation, second-generation reforms proved more difficult, and, it is now believed by many, were more fundamental to achieving a successful outcome. Transformation required, at different stages, macroeconomic, structural, and institutional reforms. Stabilization, liberalization, and privatization were necessary but not sufficient conditions to guarantee a return to growth and an improvement in living standards. The institutional underpinnings of the capitalist system are needed for market reforms to work and, ultimately, for transformation to succeed. Sound policies to be effective require a solid institutional setting. Likewise, poor institutions and weak governance will inhibit market reforms. Without these changes, the transformation to the market will remain incomplete and the least-advanced TEs will continue to experience serious coordination problems (Hare and Turley, 2013).

Institutions and Institutional Change

Although there are different definitions of institutions, the most common one is taken from North (1991) where institutions are defined as formal rules, informal constraints, and enforcement mechanisms. Institutions, including the organizations that implement the rules and codes of conduct, are important because they facilitate exchange and govern market transactions. In any setting a stable society requires a set of rules and social norms for behavior. Order in the socialist system was provided by the State and the Communist Party. All else, whether it was public ownership, the central plan, the soft budget constraint or shortages, followed from the one-party system (Kornai, 1992). Although the fall of communism was greeted enthusiastically as a move toward freedom and prosperity, an unanticipated consequence of the collapse of the State and the disintegration of the Party was, at least in the early years of transition, an institutional vacuum filled by disorder, disorganization, and, in some cases, chaos (Djankov et al., 2003). Neither the market, nor the market-supporting institutions emerged overnight. Although markets soon evolved, the institutions that facilitate and support market transactions were often slow to develop. With a weakened state capacity, many

market-supporting institutions were either, at worst, absent or, at best, ineffective. Over time, with the support of a reconstructed state and international assistance, many of the formal institutions and organizations required for a stable society and a growing economy were developed. However, it was the more informal institutions and social norms that, not surprisingly given the influence of culture and history, evolved more slowly. Aside from differences in how these institutions evolved, there was also significant cross-country or cross-regional variation with respect to institutional change. Beyond the initial years of transition, more attention was paid to institutions and institutional reform. Although much progress has been made, there is still a long way to go in improving the institutional quality in many former socialist countries and, most especially, the CIS countries. Nevertheless, given an earlier view that institution building is inevitably a slow process, some countries did manage their institutional construction "… surprisingly quick and successful …," and furthermore, judging against countries of similar levels of economic development, "… institutional scores for transition countries as a whole are no better and no worse than one would expect …," with some "… widespread, large, continuing improvements in institutional quality from 1990 to 2000" but against that, failures were also commonplace, with a significant number of TEs witnessing "slow initial progress … followed by severe regression" (Murrell, 2008).

In summary, with the notable and important exception of China, transformation witnessed the destruction of the institutional structures of the socialist system and an institutional vacuum for much of the early years of transition. As time passed and state capacity improved, institutions were developed and rebuilt, with more progress in CEE countries than in CIS countries. By the end of the second decade, a majority of the former socialist countries had many of the market-supporting institutions common in other emerging markets and the more established market economies. Whereas many TEs transplanted best-practice international institutions, others (particularly China) experimented with transitional institutions. The question of how the quality of institutions can be improved in the laggard transition countries remains an outstanding issue. What is not at issue is the significance of institutions for TEs and, more broadly, the market economy and the capitalist system. As Berkowitz and Roland (2007) wrote "The transition experience has indeed shown with a vengeance the importance of institutions in the process of economic development."

Performance and Outcomes

Some of the outcomes of transition were expected, such as the rise in private enterprise and services, the fall in government expenditure and revenue, the rise in foreign trade and unemployment, and the change in investment and consumption patterns. Others happened as predicted but the degree of the change was unexpected, such as the rapid rise in prices, the sudden fall in revenues, the extreme rise in poverty and inequality. Other developments were, for the majority of commentators interested in economic transformation, surprising. The main surprises, of which the vast majority are negative, include the collapse in output in the early years of transition; the rise in corruption, the mafia, and the unofficial economy; the extent of disorder, disruption, and dislocation; the number of failed privatization outcomes, and associated asset stripping and tunneling; the worsening demographics in CIS countries in the form of falling population numbers (falling fertility and rising mortality); and deteriorating life expectancy for males in the Slavic CIS countries. There is a very large cross-country and regional variation in performance and outcomes. Most notably, this includes the relatively poor performance of Russia as compared with the impressive outcomes achieved in China.

As regards these patterns of transition, there are differences with respect to time and regions/countries. In respect of differences over time, the early years recorded the worst features of transition (notably transformational recession, hyperinflation, rising poverty, and inequality) whereas the second decade of transformation witnessed a recovery until the onset of the Great Recession and the 2008/09 financial crisis. In respect of country and regional differences, the biggest difference is between CEE/FSU countries and the East Asian countries of China and Vietnam. Aside from some similarities, many of the most disturbing trends were not witnessed in either China or Vietnam. Within the Soviet bloc there are also differences, along the lines of CEE (including the Baltics) versus CIS (and within that, energy resources-rich vs energy resources-poor) countries, with the CEE countries doing much better in terms of output, job creation and inflation, poverty and inequality, corruption and demographics. Some of the leading CEE reformers, notably Estonia, Latvia, and Hungary, were badly affected by the 2008/09 financial crisis, leading to questions about the region's development model and, more specifically, the vulnerability of some TEs to global or regional shocks arising from the region's deep integration into the global trade and financial system. The high global liquidity, cheap credit, and speculative capital flows of the mid-2000s resulted in consumption booms, property bubbles, and rising indebtedness in many emerging markets, including some TEs. However, helped by the coordinated policy response of national authorities and international organizations, economic recovery did follow, only to falter shortly after arising partly out of contagion from the Eurozone crisis. Debate continues over the need for greater reforms, determinants of long-run growth and the prospects for income convergence with the most advanced economies in the world (EBRD, 2009, 2013; Mitra et al., 2010).

The empirical research on the determinants of output growth in TEs identifies a number of factors, including initial conditions, macrostabilization, institutions, and policy reforms. Although the evidence is mixed, it appears that there is no one key factor in explaining cross-country differences in economic performance. Initial conditions matter, but less so over time. Macrostability is important but excessive pursuit of low inflation and fiscal restraint may be costly. Market reforms matter but are not a sufficient condition for output recovery. Institutions are important but differences exist as to which institutions are important and the timing of institutional reforms vis-à-vis economic policy reforms.

Despite the cross-country variation in performance, and the disappointing outcomes in some countries and for many citizens (as evident in the Life in Transition Surveys LiTS I and II, see EBRD, 2011), recognition must be given to the

enormous changes that have taken place, and over a relatively short period of time. Aside from the welcome transition to democracy, weak as it may be in some countries, systemic changes of an economic nature include the transformation from the central plan to the market system; state ownership to majority private enterprise; price controls to market forces; large monopolistic enterprises to SMEs and competition; autarky and regional integration to foreign trade and global integration; heavy industry and military production to services and civilian production; and communist Soviet-Sino institutions to, albeit imperfect or incomplete, capitalist institutions. As for economic transformation and the broader transition from central planning to market economy as a separate subbranch of economics, the aforementioned changes have led to questions about the end of transition and economic transformation from plan to market (as distinct from transitology, in general). Although similar questions were raised in earlier years, the passage of time and successive transition anniversaries have inevitably led to reflections on the transformation process, performance outcomes, and reforms and institutional change (see Kornai, 2006; Sonin, 2013; Pistor, 2013). Whether we have seen the end or the next stage of transition, using our time now to reflect on the subject of transformation from central plan to market economy and the broader academic question of its contribution to economics and comparative economic development is a very worthwhile exercise.

Assessment and Lessons

Transformation, however it is defined, is extraordinarily complex, involving wholesale systemic large-scale institutional changes. Economic transformation, combined with the political and social transition, was unprecedented. One consequence of this systemic change is that the discipline of economics now has a greater understanding of capitalism and its institutional foundations, the interplay of politics and economics, and the importance and unpredictability of dynamic changes. In many TEs, policymakers were confronted with the dual problems of transition and development, of emerging markets and destatization, of market reform and institutional reform, of structural adjustment and reconstruction, of political change and economic change, resulting in initial outcomes that included, for the majority of TEs, a large fall in output, rapid inflation, social disruption, and, for many CIS countries, disturbing demographic trends. From an economic theory perspective, the lesson here is the inadequacy of any one school of economic thought to fully explain transformation or to adequately inform policymakers of the complexities of the transition process. Institutions are now considered mainstream, and institutional reform is considered as important to the success of transition as the market reforms that were advocated by the IFIs, policymakers and the majority of influential Western academic economists at the outset of transition. This is a lesson that has relevance beyond transition and TEs, to topics such as the determinants of economic growth, governance of developing and emerging markets, the role of IFIs and the suitability/conditionality of market reforms, etc.

If there was ever an event in the twentieth century that required a political economy perspective, it was the transformation from socialism to capitalism and from plan to market. The interplay between politics and economics is crucial in helping to understand the policy reforms, performance, and outcomes in transition countries. The experience of TEs clearly illustrates the importance of political constraints. Politics matters and the political economy of reform are central to an understanding of transition. Indeed, this practice, and growing trend, of integrating the political process into an analysis of economic issues is more widespread now in economics, into areas such as international trade, public finance, regulation, and others (Roland, 2000). In reforms, political considerations dominate as much as, if not more than, economic deliberations. More generally, the events and outcomes during transition simply reinforce the need for the economics profession today, as was the case in earlier times when the discipline was known as political economy, to emphasis and reiterate the link between economics and politics, as Adam Smith and Karl Marx did centuries ago. Reforms were often not the exogenous choice of policymakers but instead were endogenous, subject to political constraints. Feasibility as opposed to ideology was, in the end, often what determined actual policy reforms. In the tradition of political economy, reforms were often designed based on what was politically feasible and socially acceptable (Turley and Luke, 2010).

Another lesson from the transition experience is the role of the state in the modern capitalist system. Given the extent of state control in the command economy and given the prevailing economic orthodoxy at the time of the collapse of the socialist system, there was inevitably going to be a significant reduction and reorientation in state involvement, both in terms of size and scope. Some commentators, admittedly those closest to the neoliberal perspective, even argued for a dismantling of the state. In cases where this occurred, state collapse and the problem of a weakened state emerged. No doubt, reform of government was necessary. A reconstituted state was required. However, as learned from elsewhere (particularly in the post–World War II era and in the newly industrialized countries (NICs) of South East Asia), markets and governments complement each other. Lessons from individual TEs, both positive and negative (whether it is decentralization in China, enterprise restructuring in Poland, or state collapse in Georgia and Russia), indicate the important role of the state in the capitalist system. Despite the tendency during transition to dismantle the state, it became clear that, in many respects, a strong and effective state was required, albeit playing a much more different (and limited economic) role than before. There was a need to recognize that the state, although less required in many cases (e.g., subsidies from profit to loss-making firms, foreign trade monopolies setting prices, extensive state ownership) other roles for the state were required (e.g., policing competition policy, supporting SME development, countercyclical macro policy, social protection). As we know from international experience, both now and in the past, it is a difficult task to achieve the right balance between a state that is, on the one hand, strong and competent and, on the other hand, constrained and nonpredatory. In the context of TEs, although it was

both inevitable and desirable that the market's role would increase while the role of the state would decline, excessive moves in either direction can be costly. When writing about the importance of time and state guidance in transition, and specifically in Russia, Nobel Laureate Kenneth Arrow (2001) concluded, "Although the role of government in directing economic activity in a healthy economy is limited, transition does require guidance. The only source of general guidance for the economy is the state, and there is no denying that appropriate policy and leadership could considerably smooth transition."

What about Russia and China? Russia is, by all accounts, a market economy. The market is the main coordination mechanism with prices for the majority of goods and services determined by market forces of demand and supply. The private sector share of GDP is close to 70%. However, the postsocialist path to the market economy was difficult and, for many, traumatic. Given Russian and Soviet preconditions, transformation from plan to market in Russia was inevitably going to be complex and arduous; indeed it was. The interesting question is whether it could have been done any less painful. What, if any, were the alternatives? We will never know the answer to some of these questions but what we can do is identify some lessons and policy mistakes. With respect to Russia (and other CIS countries), the lessons from transition are many and controversial. The important specific lessons are that politics matter; market reforms need to be tailored to local conditions and, despite often wrenching circumstances, need to be socially tolerable; the importance of macroeconomic stability; the danger of an institutional vacuum; the importance of effective State capacity; the importance – or absence in the case of Russia – of legitimacy. As for China, the important lessons from the Chinese reforms are not related to the speed of reform (as speed is not the big issue) or to ownership as the Chinese record on privatization has neither caused nor prevented output growth. The interesting and more relevant reforms in China are the change in openness to foreign trade and participation in world markets; emphasis on changing incentives; willingness to, and emphasis on, experimentation, bottom-up reforms, the evolutionary nature of change; decentralization and competition; pragmatism and emphasis, ironically, not on ideology but on what works for economic growth and development. As Deng Xiaoping remarked, "It doesn't matter whether a cat is black or white, as long as it catches mice."

See also: Area and International Studies: Political Economy; Capitalism; Democratic Transitions; Institutions; Labor Market Institutions and Labor Market Outcomes in Transition Economies; Post-Socialism: Political Economy; Revolutions of 1989–90 in Eastern Central Europe; Socialism.

Bibliography

Arrow, K.J., 2001. The role of time. In: Klein, L., Pomer, M. (Eds.), The New Russia: Transition Gone Awry. Stanford University Press, Stanford, CA.

Åslund, A., 2002. Building Capitalism: The Transformation of the Former Soviet Bloc. Cambridge University Press, Cambridge, UK.

Berglöf, E., Roland, G. (Eds.), 2007. The Economics of Transition: The Fifth Nobel Symposium in Economics. Palgrave Macmillan, Hampshire.

Berkowitz, D., Roland, G., 2007. Editorial statement. Journal of Comparative Economics 35, 1–2.

Blanchard, O., 1997. The Economics of Post-Communist Transition. Oxford University Press, Oxford.

Campos, N.F., Coricelli, F., 2002. Growth in transition: what we know, what we don't, and what we should. Journal of Economic Literature 40 (3), 793–836.

Clague, C., Rausser, G. (Eds.), 1992. The Emergence of Market Economies in Eastern Europe. Blackwell, Oxford.

Djankov, S., Glaeser, E., LaPorta, R., Lopez-de-Silanes, F., Shleifer, A., 2003. The new comparative economics. Journal of Comparative Economics 31 (4), 595–619.

EBRD, 2009. Transition Report 2009: Transition in Crisis? European Bank for Reconstruction and Development, London.

EBRD, 2011. Life in Transition: After the Crisis. European Bank for Reconstruction and Development, London.

EBRD, 2013. Transition Report 2013: Stuck in Transition? European Bank for Reconstruction and Development, London.

Ellman, M., 2013. What did the study of transition economies contribute to mainstream economics? Japanese Journal of Comparative Economics 50 (1), 9–30.

Estrin, S., Hanousek, J., Kočenda, E., Svejnar, J., 2009. The effects of privatization and ownership in transition economies. Journal of Economic Literature 47 (3), 699–728.

Hare, P., Turley, G. (Eds.), 2013. Handbook of the Economics and Political Economy of Transition. Routledge, London.

Hayek, F.A., 1944. The Road to Serfdom. Routledge, London.

Kornai, J., 1992. The Socialist System: The Political Economy of Communism. Clarendon Press, Oxford.

Kornai, J., 2006. The great transformation of Central Eastern Europe: success and disappointment. Economics of Transition 14 (2), 207–244.

Milanovic, B., 1998. Income, Inequality and Poverty during the Transition from Planned to Market Economy. World Bank, Washington, DC.

Mitra, P., Selowsky, M., Zalduendo, J., 2010. Turmoil at Twenty: Recession, Recovery and Reform in Central and Eastern Europe and the Former Soviet Union. World Bank, Washington, DC.

Murrell, P., 2008. Transition and institutions. In: Blume, L., Durlauf, S. (Eds.), The New Palgrave Dictionary of Economics, second ed. Palgrave Macmillan, Hampshire.

Naughton, B., 2007. The Chinese Economy: Transitions and Growth. MIT Press, Cambridge, MA.

North, D.C., 1991. Institutions. Journal of Economic Perspectives 5 (1), 97–112.

Pistor, K., 2013. Towards a new transition economics. Economics of Transition 21 (1), 11–16.

Roland, G., 2000. Transition and Economics: Politics, Markets, and Firms. MIT Press, Cambridge, MA.

Sonin, K., 2013. The end of economic transition: lessons for future reformers and students of reform. Economics of Transition 21 (1), 1–10.

Stiglitz, J.E., 2000. Whither reform? Ten years of the transition. In: Pleskovic, B., Stiglitz, J. (Eds.), Annual World Bank Conference on Development Economics 1999. World Bank, Washington, DC, pp. 27–56.

Turley, G., Luke, P., 2010. Transition Economics: Two Decades on. Routledge, London.

World Bank, 1996. World Development Report 1996: From Plan to Market. Oxford University Press, New York.

Economics and Crime

Joanna Shepherd, Emory University School of Law, Atlanta, GA, USA
Paul H Rubin, Emory University, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article covers the role of economic theory in advancing our understanding of criminal behavior. We begin by reviewing the history of the economic view of crime, drawing on the work of Bentham and Beccaria, and embracing individual rational behavior. This discussion is followed by a cost–benefit analysis of crime and a general discussion of how 'economics' differs from 'criminology' when explaining criminal behavior and legal policy. The article concludes with a review of recent studies offering economic analysis of crime and justice, as well as avenues to extend this model in future work.

The Economic Theory of Criminal Behavior

Beginning with the work of economics Nobel laureate Gary Becker in 1968, economists have invaded the field of criminology, using their all-embracing model of individual rational behavior. Several of Becker's ideas were foreshadowed by earlier writers – Cesare Beccaria in 1767 and Jeremy Bentham in 1789. These early scholars, though not economists, developed several concepts that would later be associated with the economic theory of criminal behavior: "… the profit of the crime is the force which urges man to delinquency: the pain of the punishment is the force employed to restrain him from it. If the first of these forces be the greater, the crime will be committed; if the second, the crime will not be committed" (Bentham, 1907: 399).

However, from the beginning of the twentieth century interest in their point of view dwindled as a plethora of other theories were developed. Fortunately, the main idea of Bentham was vitalized and modernized in Becker's path-breaking article, Crime and Punishment, where he suggests that "a useful theory of criminal behavior can dispense with special theories of anomie, psychological inadequacies, or inheritance of special traits and simply extend the economist's usual analysis of choice" (Becker, 1968).

In Becker's model, a criminal act is preferred and chosen if the expected benefits from committing a crime exceed the expected costs, including the costs of any foregone legal alternatives. The economic theory of criminal behavior is thus regarded as a special case of the general theory of rational behavior under uncertainty. Assuming that individual preferences are constant, the model can be used to predict how changes in the probability and severity of sanctions and in various socioeconomic factors may affect the amount of crime. Even if individuals who violate certain laws differ systematically from those who abide by the same laws, the former, like the latter, do respond to incentives (i.e., to sanctions and economic conditions). Indeed, numerous empirical studies confirm the predictions of the economic theory.

The Benefits and Costs from Crime

Thus, the economic model of criminal behavior assumes that the decision to commit a crime is the result of a cost–benefit analysis that individuals undertake either consciously or subconsciously. The gains and losses included in the economic model are usually meant to represent all kinds of benefits and costs that have an effect on the people's decisions. People are assumed to allocate time to criminal activity until marginal benefits equal marginal costs. For some people marginal benefits of crime are probably always lower than marginal costs of crime; the economic model would predict that these people will never commit crimes. For others, the marginal benefits often exceed the marginal costs; we would expect these people to specialize in crime. For most of us, the marginal costs usually exceed the marginal benefits. However, every once in a while the marginal costs seem particularly low or the marginal benefits seem particularly high, and we commit a crime. For example, we may be driving on a stretch of country road where police seem unlikely (low marginal cost of speeding) or we are driving a Ferrari (high marginal benefit of speeding) so we decide to speed.

The kinds of gains obtained from a criminal act vary, depending on the type of crime and the individual criminal. Some are monetary, such as the gains obtained from theft, robbery, insurance fraud, killing a rival drug dealer, etc. Others are psychic, such as the thrill of danger, peer approval, retribution, sense of accomplishment, or 'pure' satisfaction of wants (rape). Obviously, the psychic benefits from crime will be different for different people: young men tend to enjoy the thrill of danger more than older women and gang members tend to get more peer approval than churchgoers when they commit a crime. Incarcerated criminals may also gain human capital in committing future crimes if they learn techniques from other criminals or meet future potential crime partners while imprisoned.

The costs of crime also depend on the crime and individual. The costs can include direct material costs, psychic costs, opportunity costs, and expected punishment costs. The material costs include the cost of supplies purchased to commit crimes (equipment, guns, vehicles, face masks). Psychic costs include any guilt, anxiety, fear, dislike of risk, or other emotions associated with committing crime. The opportunity cost of crime consists of the net benefit of the legal activity forgone while planning, performing, and concealing the criminal act. The lower an individual's level of income, the lower is his or her opportunity cost of engaging in illegal activity. The amount a person can earn in the legal sector may depend on factors such as age, sex, race, education, training, region, rate of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45057-9

unemployment, and IQ. People who are only able to earn a low wage will have a low opportunity cost of crime, as they are not giving up substantial legal income.

The expected punishment costs include the cost of all formal and informal sanctions, as well as the pecuniary costs arising from litigation (lost income and lawyers' fees). When the formal sanction is a fine, the punishment cost is the amount of the fine. When the formal sanction is a prison term, the punishment cost incorporates the cost to the criminal of going to prison: the lost income, the monetary equivalent of the loss of liberty, the monetary equivalent of whatever harms come to the individual while in prison, etc. The costs of the informal sanctions result from the social stigma that accompanies arrest, conviction, and imprisonment. These sanctions can include the reactions of employers, family, and friends and the reduced legitimate income a criminal will earn once he or she has a criminal record. These expected punishment costs must be weighted by the probability that the individual will be arrested, convicted, and imprisoned. An individual facing a 50% chance of receiving a 10-year prison sentence has higher expected punishment costs than an individual facing a 5% chance of the same prison sentence. The probability of punishment will be different for different people. Some are cleverer than others at concealing the offense and eluding the police. There are also differences in the abilities of defending oneself in court, or in engaging good lawyers (Hoffman et al., 2005). That is to say, to estimate the expected penalty that a particular individual faces, we multiply all costs of punishment by the probability of receiving those punishments.

The Stigma of Criminal Activity

The stigma from criminal activity can be social (inability to find a spouse) or economic (inability to find a job with a criminal record). Several scholars have attempted to measure the economic stigma. Lott (1992) finds a short-run income reduction of 39% after a bank embezzlement conviction and a 41% reduction after a bank larceny conviction. Grogger (1992) finds that an arrest record can explain about two-thirds of the black–white youth employment differential in his sample. However, in another study, he finds only a short-lived effect of arrest records on youth earnings (Grogger, 1995). For gang members, having spent time in prison may actually increase future incomes.

The Criminal's Discount Rate

The rate at which individuals discount the future also affects the expected benefits and costs from criminal activity. The gains from crime often occur immediately, whereas punishment is something that might come in the future, and be stretched over a long period of time. An individual with a high discount rate will therefore tend to commit more crime because he or she weighs the present (and the gains from criminal activity) much more heavily than the future (and the potential costs from criminal activity).

The Economic Theory of Recidivism

Recidivism is sometimes explained by erratic behavior, a lack of self-control, or evidence that the deterrence model does not work. However, a high rate of recidivism is consistent with the model of rational choice. Serving time in jail may reduce legal opportunities so that the opportunity cost of future criminal activity is lower. Additionally, convicts may acquire human capital in illegal activities – prison is an excellent place to 'network' with other criminals and learn the skills of the trade. Thus, if it was rational to commit a crime in the first place, for many criminals the incentives will only be stronger after having served a prison sentence.

In contrast, a prior conviction or imprisonment might increase some criminals' evaluations of how probable or severe sanctions might be. For these criminals, the expected costs of additional crime may be higher than the expected costs accompanying their initial crime and they will tend not to recidivate.

Economics versus Criminology

How does the economic approach to criminal activity differ from the criminology approach? The criminological literature is essentially composed of three branches. One branch focuses on the biological factors contributing to crime, such as brain abnormalities or hormonal imbalances. The economic approach does not deny that biological factors matter; it just assumes that these factors explain the baseline level of crime that exists regardless of the incentives created by other costs and benefits of criminal activity. The second branch of the criminological literature asserts that people turn to crime when they are prevented from reaching their goals through legal means. This assertion is consistent with the economic model that predicts that people with limited legal opportunities may turn to crime. However, the economic approach, in contrast to the criminological approach, predicts that individuals weigh the relative costs and benefits of crime and legal activities, and only engage in crime if it is relatively more attractive. The third branch of the criminological literature is concerned with the social interactions through which criminal behavior is learned or culturally transmitted. This approach is also consistent with the economic model that maintains that community influences and cultural factors can influence various costs and benefits of crime: an individual committing a crime may feel less of a stigma (a cost of crime) or even gain approval (a benefit) depending on how others in his or her social circle views criminal activity; an individual may feel more or less internal guilt (a cost) depending on their religion or community; individuals may learn smarter criminal tactics when they associate with other criminals. Thus, the economic model of crime, for the most part, encompasses many of the criminological explanations for crime. The difference between the two approaches is mainly one of emphasis (see Levitt and Miles, 2007).

There is also a branch of criminology that stresses the implications of everyday activities on crime. For example, as more women enter the labor force there are more vacant houses and so perhaps more burglaries. As more people carry cell phones the ability to contact police increases and so crime may be reduced (Felson and Santos, 2009). This is also consistent with the economic model: everyday activities can affect the gains from crime or the expected punishment, and in turn, the level of crime.

Are Criminals Really Rational?

The economic approach posits that everyone (except, perhaps, individuals with severe mental disabilities) responds, to some

degree, to changes in the expected costs and benefits of criminal activity. Several authors have discussed whether people have sufficient information about the environment and about outcomes of actions to make rational choices. Becker and others maintain that even if choices are based on subjective beliefs that are wrong, the choices are meaningful from a subjective point of view, and behavior can be explained and understood on this basis. Moreover, even if people are not exactly accurate in their estimation of the expected benefits or costs of criminal activity, an obvious increase in an expected cost of crime or decrease in an expected benefit of crime should still influence (albeit imperfectly) the incentives to commit crime.

Empirical Studies of the Economic Model of Crime

Predictions from the economic model of crime have been tested in a great number of empirical studies. However, these empirical studies face several challenges. First, although the economic model of crime is based on individual rational choice, the data available to empirical researchers are based on levels of aggregation ranging from countries and states down to municipalities and campuses. Thus, the empirical results can tell us only about general deterrence – the effects on people in general – rather than the effects on an individual's decision-making.

Second, it is extremely difficult to identify causal relationships in empirical studies of crime. For example, in testing the relationship between imprisonment and crime, one may find a positive relationship (imprisonment and crime increase together or decrease together). Based on these results, some scholars may conclude that prison increases crime, when in fact, the positive relationship may be attributable to higher crime rates producing more potential prisoners.

Controlling for the factors that lead to high crime could solve the problem, but it is impossible to quantify and control for every factor, so omitted variable bias can be a problem. More recent empirical studies have devised clever ways to tease out causal relationships, for example, by taking advantage of exogenous changes in imprisonment levels, such as court-ordered prison releases, that could not be caused by crime (Levitt, 1996).

The third challenge is the difficulty in distinguishing incapacitation and deterrence. The economic model of crime assumes that increases in imprisonment increase the expected costs of crime and deter some individuals from committing crimes. However, an empirical finding that increases in imprisonment are associated with decreases in crime might not be evidence of deterrence. Crime could also decrease because more criminals are behind bars due to the increased imprisonment. Although the end result is the same, the rational choice model assumes that individuals can be deterred from committing crime. Thus, if the entire decrease in crime was caused by incapacitating criminals, this would not confirm the predictions of the economic model of crime.

There are various other challenges that involve the correct variables to control for in empirical analyses, the particular methodology to employ, and other sophisticated empirical questions. Nevertheless, the awareness of the methodological problems and easier access to various statistical methods has gradually led to more sophisticated empirical studies that address several of these challenges.

Cognizant of these empirical challenges, law and economics scholars have studied the relationship between crime and various costs and benefits of criminal activity. Summarizing the entire empirical literature, which consists of thousands of articles, is beyond the scope of this article, but below we discuss the general consensus or important articles studying the effects of different variables.

Police

The majority of empirical studies find that policing levels have a large negative relationship with crime (see, e.g., Di Tella and Schargrodsky, 2004; Corman and Mocan, 2000; Levitt, 1997).

Imprisonment

Most studies find that imprisonment also has a negative effect on crime rates. Some of this effect is due to deterrence and some is due to incapacitation (see, e.g., Levitt, 1998; Kessler and Levitt, 1999). Severe prison conditions also have a deterrent effect (Katz et al., 2003).

Capital Punishment

The effect of capital punishment continues to be hotly debated. The majority of empirical papers find a deterrent effect of capital punishment (see, e.g., Dezhbakhsh and Rubin, 2011; Dezhbakhsh et al., 2003; Mocan and Gittings, 2003). Some studies have found that the deterrent effect exists only in some states (Shepherd, 2005). Others argue that the deterrence finding is extremely sensitive to methodological choices in the empirical analysis (Donohue III and Wolfers, 2006).

There is a tendency to conflate findings regarding capital punishment with general issues of deterrence. In the contemporary US capital punishment is rare and unpredictable. It could be that the rarity of this punishment reduces or even eliminates the deterrent effect. It is also possible that its rarity means that statistical methods are unable to detect a deterrent effect even if one exists. Nonetheless, even if we cannot detect a deterrent effect of capital punishment, this does not mean that more normal punishments such as jail time do not deter.

Gun Laws

The effect of gun laws on crime is also hotly debated. Some studies have found evidence that laws allowing citizens to carry concealed handguns decreases crime; presumably the possibility of encountering a victim with a gun deters potential criminals from committing crime in the first place (Lott, 2010; Lott and Mustard, 1997). In contrast, other studies find that gun laws do no decrease crime (Ayres and Donohue III, 2003; Dezhbakhsh and Rubin, 1998).

Income

No systematic relationship appears between various measures of income and crime. Several studies find that high legitimate wages (an opportunity cost of crime) are associated with low

crime rates (Machin and Meghir, 2000; Cornwell and Trumbull, 1994). On the other hand, some studies find that higher pecuniary benefits to crime (a benefit of crime) are associated with increases in crime (Glaeser and Sacerdote, 1999). This ambiguity in results is likely due to the fact that the income measures used represent benefits not only of legal activities, but also of illegal ones: higher legal incomes tend to make work more attractive than crime, but to the extent that higher legal income in a region produces a greater number of more profitable targets for crime, crime also becomes more attractive. If these mechanisms are at work simultaneously, and their relative strength is not universally constant, it is not surprising that the results of various studies differ.

Income Inequality

A large income differential may indicate that crime is a comparatively rewarding activity for the very-low-income group that may find a lot to steal from the very rich. Nevertheless, empirical studies also find no systematic relationship between income inequality and crime. Several studies find that greater income inequality does lead to increases in crime (Fajnzylber et al., 2002). In contrast, others find no statistically significant relationship between income inequality and crime (Neumayer, 2005).

Unemployment

Unemployment is a measure of the lack of legal income opportunities; the unemployed have lower opportunity costs of crime as they are not giving up legitimate employment. The majority of studies find that higher rates of unemployment correlate with higher levels of crime (Gould et al., 2002; Raphael and Winter-Ebmer, 2001).

The Economics of Law Enforcement

The economic theory of public law enforcement is based on a specific perception of justice as efficiency. Under this perception, the purpose of public law enforcement is to maximize social welfare, where social welfare is the benefits that individuals obtain from crime minus the costs of committing crime, the costs of harm to victims, and the costs of enforcement. The government can maximize the social welfare function through three policy instruments: the probability of capture and punishment, the length of prison terms, and the level of fines. In setting these policy instruments at their ideal levels, the economic model assumes that criminals behave rationally and weigh the costs and benefits of criminal activity before turning to crime.

Deterrence is the goal in the traditional economic model of public law enforcement. The optimal amount of deterrence occurs at the point where the marginal social cost of additional deterrence equals the marginal social benefit. (Following most of the literature, we focus on optimal enforcement when deterrence is a goal. Incapacitation as a goal has received much less attention. See Shavell (1987) for a discussion of optimal enforcement when incapacitation is a goal.) Thus, under the traditional definition, efficient deterrence balances the marginal costs of enforcement and the reduction in illegal gains to criminals against the marginal benefit of enforcement, which is the reduction in harm that results from enforcement. Thus, deterrence is inefficient if the marginal gains to criminals plus marginal enforcement costs exceed the marginal harm to victims.

However, many researchers have questioned whether the gains to criminals should be considered (Lewin and Trumbull, 1990). If we ignored these gains, optimal deterrence would occur at the point where the marginal enforcement costs equal the marginal enforcement benefits (which is the marginal harm to victims). With fewer costs of deterrence to consider (i.e., ignoring the reduction in criminal gains that results from deterrence), the optimal level of deterrence increases.

We begin with the optimal law enforcement choices between probability and severity and between fine and imprisonment in the basic model. Then we consider several extensions of the basic model.

Certainty versus Severity

Law enforcement can influence various policy instruments to achieve the optimal level of deterrence: the probability of punishment, the length of prison terms, and the level of fines. The probability of punishment depends on policing levels, arrest rates, conviction rates, imprisonment rates, funding for law enforcement, funding for courts and prosecutors, etc. In the economic model of law enforcement, the probability of punishment is often referred to as the *certainty* of punishment. In contrast, the length of prison terms and level of fines, which law enforcement can also influence, are referred to as the *severity* of punishment.

In his or her decision to commit a crime, a criminal's expected penalty is the product of the certainty of punishment and the severity of punishment. **Table 1** shows several combinations of certainty and severity (presented as either a fine or the cost of imprisonment) that produce identical expected penalties.

If potential offenders are risk neutral and have no wealth constraints, all of the combinations of probability and severity should produce identical levels of deterrence. However, more severe sanctions are cheaper for society to implement. Increasing a monetary fine imposes almost no additional collection costs; rather, it generates additional revenue for society. In contrast, increasing certainty requires more resources spent on police and courts, which can dramatically increase society's costs. Thus, a low probability/high fine combination will achieve the same deterrence as a high probability/low fine combination, but at much lower cost to society.

Table 1 Combinations of certainty and severity with identical expected penalties; the risk-neutral case

Probability of sanction (%)	*Severity of sanction*	*Expected penalty*
100	100	100
50	200	100
25	400	100
10	1000	100

In contrast to fines, increasing the severity of imprisonment imposes significant costs on society that include the costs of running a prison (estimates suggest that it costs $40 000 a year to house a prisoner), the productivity costs of removing a criminal from society (assuming he or she provided some positive benefits to his or her community), and any psychic or stigma costs the criminal experiences while imprisoned. Despite these significant costs associated with imprisonment, it may still be relatively cheaper to increase the severity of imprisonment than to increase the probability of sanctions. Although the longer prison sentences increase enforcement costs, fewer individuals are imprisoned, which decreases enforcement costs and offsets some of the increase from longer sentences. In the linear case, these exactly balance – one person serving 10 years has the same cost at 10 people serving 1 year each.

Thus, regardless of the form of sanction, low certainty/high severity is, in general, the optimal combination because it achieves deterrence at the lowest possible cost (Polinsky and Shavell, 1984). Factors that may contradict this generality are discussed below, but first we discuss the optimal choice of fines versus imprisonment.

Fines versus Imprisonment

Criminal sanctions can be monetary or nonmonetary. As we discussed in the previous section, fines impose little cost on society, and even generate revenue. In contrast, imprisonment imposes substantial enforcement costs on society. Thus, fines are the preferred sanction.

However, a combination of fines and imprisonment is necessary in many situations. If a fine exceeds an individual's wealth level, then the individual's expected penalty will be less than the expected penalty that law enforcement anticipated when selecting the probability and severity of sanctions. For example, in **Table 1**, a probability of 10% and a fine of $1000 for an individual with a wealth level of only $500 will achieve an expected penalty of only $50 instead of $100. Since criminals tend to be those with low opportunity costs of committing crime, for many types of crime it is likely that criminals' wealth levels will be insufficient to achieve optimal deterrence.

When there are wealth constraints, the fine should be set as high as possible, equal to the individual's wealth level. A prison sentence should also be imposed to bring the expected penalty to the desired level. Empirical studies indicate that there is a tendency to impose larger fines and shorter prison sentences on wealthy defendants, although there could be several reasons for this result (Meade and Waldfogel, 1998).

Extensions of the Basic Model

According to the basic economic model of law enforcement, sanctions should be as high as possible and probabilities should be as low as possible to achieve a given level of deterrence. However, this low certainty/high severity combination is not always optimal.

Risk Preferences

When criminals are either risk averse or risk loving, the low certainty/high severity combination may no longer be optimal.

First, consider the case where individuals are risk averse in sentences so that their disutility of the expected penalty rises more than in proportion to the expected penalty. In the case of prison sentences, this could result from an increasing desire for freedom or growing distaste for the prison environment as the time in prison increases (Polinsky and Shavell, 2005: 22). Risk-averse individuals prefer a certain penalty, f, to an uncertain penalty with a mean of f.

Table 2 shows several combinations of certainty and severity (presented as either a fine or the cost of imprisonment) that produce the same disutility, but have different expected penalties. In contrast to the risk-neutral case where combinations with equal expected penalties produce equal deterrence, when severity increases for a risk-averse individual, combinations with lower expected penalties produce equal deterrence.

If **Table 2** referred to fines, it is not clear that a low certainty/high severity combination, such as a 1% probability of a $1000 fine, is more optimal than the other combinations in **Table 2**. Although enforcement costs are low, the expected fine, and therefore fine revenue is also low. If the decrease in fine revenue is greater than the savings in enforcement cost, the high fine combination could be a more expensive combination from society's perspective (Polinsky and Shavell, 1979; Kaplow, 1992).

In contrast, if the sanctions in **Table 2** were prison sentences instead of fines, then low expected sentences would not be costly. That is, the 1% probability of a 1000-day sentence has both low enforcement costs and low imprisonment costs. Thus, when offenders are risk averse, the high sentence/low probability combination is unambiguously the optimal combination.

Next, consider the case when offenders are risk loving in sanctions; their disutility of the expected penalty rises less than in proportion to the expected penalty. In the case of prison sentences, this could occur for several reasons: if the disutility from the stigma of prison does not increase with the length of imprisonment, if more brutalization of prisoners occurs at the beginning of a sentence, or if discounting of future disutility makes earlier years in prison seem worse than later years (Polinsky and Shavell, 2005: 23). Risk-loving individuals prefer an uncertain penalty with a mean of f to a certain penalty, f.

Table 3 shows several combinations of certainty and severity (presented as either a fine or the cost of imprisonment) that produce the same disutility, but have different expected penalties. In the risk-loving case, as severity increases, combinations with higher expected penalties produce equal deterrence.

If **Table 3** referred to fines, then the optimal combination of certainty and severity would be high fine and low probability.

Table 2 Combinations of certainty and severity with identical disutility; the risk-averse case

Probability of sanction (%)	*Severity of sanction*	*Disutility of expected penalty*	*Expected penalty*
100	100	100	100
40	200	100	80
10	400	100	40
1	1000	100	10

Table 3 Combinations of certainty and severity with identical disutility; the risk-loving case

Probability of sanction (%)	*Severity of sanction*	*Disutility of expected penalty*	*Expected penalty*
100	100	100	100
60	200	100	120
40	400	100	160
30	1000	100	300

This combination would achieve the same deterrence as other combinations, and the expected fine, or fine revenue, would be high and enforcement costs would be low.

If Table 3 referred to prison sentences, then the optimal combination may not be the long sentence/low probability combination. Because the increase in prison sentences is proportionally larger than the decrease in the probability, the expected penalty, or expected prison term rises. If the cost of longer imprisonment exceeds the savings in enforcement costs, then the long sentence/low probability combination would not be the cheapest way to achieve deterrence (Polinsky and Shavell, 1999).

Marginal Deterrence

The low certainty/high severity combination may also not be optimal if the goal is marginal deterrence. If penalties for all offenses are set at their maximal level, there is no reason to not commit a more serious offense. If an individual decides to mug someone, he or she might as well shoot the victim and other potential witnesses; murder carries no additional penalty and killing witnesses reduces the probability of being caught. Instead, the severity of sanctions should increase with the seriousness of the crime. Thus, the penalty for mugging should not be set at its maximal level; it should be set sufficiently below the penalty for murder so that the higher penalty for murder will provide an additional deterrent. Although the lower expected penalties will increase lesser crimes, the reduction in harm from the decrease in more serious crimes will more than offset the increase in harm from the lesser crimes, increasing social welfare.

Repeat Offenders

The low certainty/high severity combination may also not be optimal for first-time offenders so that there is room to increase the sanction if those offenders become repeat offenders. There are several reasons why it may be optimal to increase sanctions for repeat offenders. First, a prior criminal record signals that an individual has a higher propensity to commit criminal acts because either the costs of crime are lower for him or her or the benefits of crime are higher. Thus, higher sanctions are necessary to deter these high-risk offenders (Polinsky and Rubinfeld, 1991). Second, imposing higher sanctions on subsequent crimes increases the cost of committing first offenses; not only does a first offense carry an immediate penalty, it also increases future penalties. Thus, higher sanctions for subsequent crimes may deter first crimes (Polinsky and Shavell, 1998). Third, a repeat offender has already suffered the social stigma of conviction from the first offense. The cost of committing subsequent crimes is less because they do not carry the potential cost of social stigma. Thus, other costs of subsequent crimes, like the expected sanction, must be increased to maintain deterrence (Miceli and Bucci, 2004). Finally, repeat offenders may be better able to avoid detection and apprehension of subsequent crimes because they understand the system and have a larger criminal network. If the probability of punishment is lower for repeat offenders, the magnitude of punishment must be higher to maintain deterrence (Polinsky and Rubinfeld, 1991). In reality, repeat offenders are typically punished more severely than first-time offenders for the same offense. However, some scholars have presented theories where it may be optimal to lower sanctions for repeat offenders (see Burnovski and Safra, 1994; Polinsky and Rubinfeld, 1991 for a discussion).

Conclusion

This article is a summary of a very large body of research. It is impossible to summarize what is itself a summary. But we can say that Gary Becker's (1968) paper has been enormously successful in that it has led to a huge body of research which has greatly enriched our understanding of crime and criminal behavior, leading to significant improvements in policy.

See also: Crime and Class; Crime: Knowledge about and Prevalence; Criminology: Development and Current Status of the Discipline; Deterrence: Sanction Perceptions; Policing; Prisons and Imprisonment; Routine Activities, Criminal Opportunities, Crime and Crime Prevention.

Bibliography

Ayres, I., Donohue III, J.J., 2003. Shooting down the "more guns, less crime" hypothesis. Stanford Law Review 55, 1193.
Becker, G., 1968. Crime and punishment: an economic approach. The Journal of Political Economy 76, 169–170.
Bentham, J., 1907[1789]. An Introduction to the Principles of Morals and Legislation. Publisher/Edition. Clarendon Press, Oxford.
Burnovski, M., Safra, Z., 1994. A note on the optimal punishment for repeat offenders. International Review of Law and Economics 14 (3), 341–350.
Corman, H., Mocan, N., 2000. A time-series analysis of crime and drug use in New York city. American Economic Review 90, 584.
Cornwell, C., Trumbull, W.N., 1994. Estimating the economic model of crime with panel data. The Review of Economics and Statistics 76, 360.
Dezhbakhsh, H., Rubin, P.H., 2011. From the 'Econometrics of capital punishment' to the 'capital punishment' of econometrics: on the use and abuse of sensitivity analysis. Applied Economics 43, 3655.
Dezhbakhsh, H., et al., 2003. Does capital punishment have a deterrent effect? New evidence from postmoratorium panel data. American Law and Economics Review 5, 344.
Dezhbakhsh, H., Rubin, P.H., May 1998. Lives saved or lives lost: the effect of concealed handgun laws on crime. American Economic Review 468–474.
Di Tella, R., Schargrodsky, E., 2004. Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack. American Economic Review 94, 115.
Donohue III, J., Wolfers, J., 2006. Uses and abuses of empirical evidence in the death penalty debate. Stanford Law Review 58, 791.
Fajnzylber, P., Lederman, D., Loayza, N., 2002. Inequality and violent crime. Journal of Law and Economics 45, 1.
Felson, M., Santos, R.B., 2009. Crime and Everyday Life. Sage Publications.
Glaeser, E.L., Sacerdote, B., 1999. Why is there more crime in cities? Journal of Political Economy 107, 225.

Gould, E., Weinberg, B., Mustard, D., 2002. Crime rates and local labor market opportunities in the United States: 1979–1995. The Review of Economics and Statistics 84, 45.
Grogger, J., 1992. Arrests, persistent youth joblessness, and black-white employment differentials. Review of Economics and Statistics 74, 100.
Grogger, J., 1995. The effect of arrest on the employment and earnings of young men. Quarterly Journal of Economics 90, 51.
Hoffman, M., Rubin, P.H., Shepherd, J., 2005. An empirical study of public defender effectiveness: self-selection by the 'marginally indigent. Ohio State Journal of Criminal Law 3, 223–255.
Kaplow, L., 1992. The optimal probability and magnitude of fines for acts that definitely are undesirable. International Review of Law and Economics 12, 3–11.
Katz, L., et al., 2003. Prison conditions, capital punishment, and deterrence. American Law and Economics Review 5, 318.
Kessler, D., Levitt, S.D., 1999. Using sentence enhancements to distinguish between deterrence and incapacitation. Journal of Law and Economics 42, 343.
Levitt, S.D., 1996. The effect of prison population size on crime rates: evidence from prison overcrowding litigation. The Quarterly Journal of Economics 111, 319.
Levitt, S.D., 1997. Using electoral cycles in police hiring to estimate the effect of police on crime. American Economic Review 87, 270.
Levitt, S.D., Miles, T.J., 2007. Empirical study of criminal punishment. Handbook of Law and Economics 455.
Levitt, S.D., 1998. Why do increased arrest rates appear to reduce crime: deterrence, incapacitation, or measurement error? Economic Inquiry 36, 353.
Lewin, J.L., Trumbull, W.N., 1990. Social value of crime. International Review of Law and Economics 10 (3), 271–284.
Lott Jr., J.R., 2010. More Guns, Less Crime.
Lott Jr., J.R., Mustard, D.B., 1997. Crime, deterrence, and right-to-carry concealed handguns. Journal of Legal Studies 26.
Lott Jr., J.R., 1992. Do we punish high-income criminals too heavily? Economic Inquiry 30, 583.
Machin, S., Meghir, C., 2000. Crime and economic incentives. Journal of Human Resources 39, 958.
Meade, J., Waldfogel, J., 1998. Do Sentencing Guidelines Raise the Cost of Punishment?. NBER Working Paper Series. 6361. National Bureau of Economic Research, Cambridge, MA.
Miceli, T.J., Bucci, C., 2004. A Simple Theory of Increasing Penalties for Repeat Offenders. Economics Working Papers. Paper 200439. http://digitalcommons.uconn.edu/econ_wpapers/200439.
Mocan, N.H., Gittings, R.K., 2003. Getting off death row: commuted sentences and the deterrent effect of capital punishment. Journal of Law and Economics 46, 453.
Neumayer, E., 2005. Inequality and violent crime: evidence from data on robbery and violent theft. Journal of Peace Research 42, 101.
Polinsky, A.M., Rubinfeld, D.L., 1991. A model of optimal fines for repeat offenders. Journal of Public Economics 46, 291–306.
Polinsky, A.M., Shavell, S., 1998. The economic theory of public… punitive damages: an economic analysis. Harvard Law Review 111, 869–962.
Polinsky, A.M., Shavell, S., 1999. On the disutility and discounting of imprisonment and the theory of deterrence. Journal of Legal Studies 28, 1–16.
Polinsky, A.M., Shavell, S., 2005. Economic Analysis of Law. Economics and Business Discussion Paper Series. Paper 536. Harvard Law School John M. Olin Center for Law. http://lsr.nellco.org/harvard_olin/536.
Polinsky, A.M., Shavell, S., 1984. The optimal use of fines and imprisonment. Journal of Public Economics 24, 89–99.
Polinsky, A.M., Shavell, S., 1979. The optimal tradeoff between the probability and magnitude of fines. American Economic Review 69, 880–891.
Raphael, S., Winter-Ebmer, R., 2001. Identifying the effect of unemployment on crime. Journal of Law and Economics 44, 259.
Shavell, S.M., 1987. A model of optimal incapacitation. American Economic Review 77, 107.
Shepherd, J., 2005. Deterrence versus brutalization: capital punishment's differing impacts among states. Michigan Law Review 104, 203.

Economics and Ethics

John Broome, University of Oxford and Corpus Christi College, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The links between ethics and economics are multifarious, and run in both directions. Much of economics is a sort of applied ethics, and is founded on ethical theories. Conversely, economists have contributed to the development of ethical thinking. Preference-based utilitarianism, which was developed by economists, has proved influential in utilitarian thinking generally. Tools from economics that have proved useful in ethics include utility theory, game theory, and social choice theory. Economists and moral philosophers share an interest in many particular topics, including the value of human life and the value of equality.

Introduction

The economy is a central component of society. To a large extent, it determines our well-being, freedoms and opportunities, social connections, health, and many other features of our lives. To a large extent, it also determines how equally or unequally these things are distributed across people. Indeed, the economy is perhaps more important than anything else in determining how our lives go. Consequently, an essential task for ethics is to assess how well the economy is doing and how it might be improved.

This is a task within applied ethics. However, since the working of the economy is complex, and the science of economics is technical, few philosophers have the skills needed to progress far with it. It commonly falls to economists instead. The branch of economics that is devoted to the ethical assessment of the economy is known as *welfare economics*. It is also a branch of applied ethics. Because of its particular subject matter, it has its own special methods, but because it is applied ethics, it needs some foundations in ethical theory. Inevitably, welfare economists working back toward the foundations of their subject have encountered moral philosophers working forward toward applications of theirs. Several topics lie within the domains of interest of both welfare economists and moral philosophers. Section Shared Topics gives some examples.

When they look for foundations for their work, many welfare economists have simply adopted an existing ethical theory. They have formalized it – sometimes distorting it in the process – and applied it to their subject. One example comes from the theory of 'optimal tax,' which flourished in the 1970s. Optimal tax theorists tried to devise a scheme of taxation that would best achieve a particular objective. They took the objective from ethical theory. Some (e.g., Mirrlees, 1971) adopted the utilitarian objective of maximizing the total of people's well-being; others (e.g., Phelps, 1973) adopted the 'maximin' objective of maximizing the well-being of the worst-off person. The latter group was inspired by the publication in 1972 of John Rawls's *A Theory of Justice*. However, they distorted Rawls's view: Rawls himself favored maximizing the worst-off group's access to 'primary goods,' rather than the worst-off person's well-being.

Other welfare economists have been less passive and more critical in their attitude to ethical theory. They have themselves contributed to the theory's development. They have found that the methods of economics are useful equipment for doing so. Like most of economics, welfare economics is a rather formal discipline, and it has evolved a range of formal tools to serve its purposes. These tools contribute to the ethical assessment of economies, but they have turned out to be useful for other ethical purposes too. Moreover, theoretical tools from other branches of economics have also turned out to be useful in ethics. Section Tools for Ethics gives examples of formal methods developed in economics that can be useful in moral philosophy.

In sum, the connections between economics and ethics are multifarious, and run in both directions. It is therefore not surprising that many important moral philosophers have also been important economists. Adam Smith, Karl Marx, John Stuart Mill, and Henry Sidgwick are examples from the past. As a general rule, economist-philosophers have been utilitarians. Economists in general, too, have generally shown utilitarian inclinations. Since the rise of 'classical' economics in the eighteenth century, economists have mostly been concerned with increasing the well-being of the people – a concern that is utilitarian in a broad sense.

Liberalism is another philosophy that is closely associated with economics. Most liberal economists are liberal for instrumental reasons. Their liberalism derives from their utilitarianism; they believe that economic freedom is a good way to promote people's well-being. For example, in *The Wealth of Nations*, Adam Smith (1776) argued that extensive free trade deepens the global division of labor, which in turn leads to economic growth, which is good for people. More recently, a similar conclusion has been drawn from some theorems in welfare economics. The theorems show that free economic competition leads to *efficiency*, defined in a special way, and this sort of efficiency is held to be generally beneficial. However, there are many occasions when economic freedom is clearly not generally beneficial. For instance, monopolies may use their power to delay innovations, and companies freely pursuing their own profit may damage the environment. In these cases, economists are typically in favor of restricting freedom by regulation. This shows that their support for freedom is only instrumental and conditional on its promoting well-being.

Utilitarianism, understood broadly, remains the dominant moral philosophy of economics. However, within economics,

utilitarianism has been radically altered as a result of the progress of economic theory. So close is the association between economics and utilitarianism that these changes have strongly influenced the progress of utilitarian thinking as a whole. Sections Utility Theory and Utilitarianism and the Pareto Principle outline the development of utilitarianism within economics.

Utility Theory

Utilitarianism is an ethical theory, but its development within economics has largely been driven by the development of an associated psychological theory, which has come to be called *utility theory*. Utility theory is the economist's core account of people's behavior. This section describes it, and Section Utilitarianism and the Pareto Principle describes its influence on economists' moral theory.

Utility theory says that each person chooses his/her actions so as to maximize his/her utility. 'Utility' is a term of art, which can be defined in various ways, and different definitions lead to very different versions of utility theory. This has allowed the theory to undergo a major metamorphosis in the course of two centuries.

Utility theory's origin may plausibly be traced to an article by the mathematician Daniel Bernoulli, published in 1738. Bernoulli thought that each person acts to maximize his/her *emolumentum*. (He wrote in Latin.) *Emolumentum* may be translated as 'benefit' or 'advantage,' so Bernoulli held a version of utility theory in which utility is benefit or advantage. Later, Jeremy Bentham (1798) made a more specific assumption that each person acts to maximize his/her pleasure, counting pain as negative pleasure. He held a version of utility theory in which utility is pleasure. (Bentham himself used the term 'utility' differently, to apply to whatever is conducive to pleasure rather than to pleasure itself.)

Utility theory became important in economics only in the 1870s, when the so-called neoclassical school rose to dominance. This was the time when economists first took a serious interest in the behavior of consumers, and utility theory was the theory of behavior they adopted. At the beginning of the neoclassical school, the economist William Stanley Jevons (1871) explicitly took his utility theory from Bentham and went far beyond Bentham in his analysis.

Like Bentham, Jevons thought that each person acts to maximize his/her pleasure. In defense of this view he said, "Call any motive which attracts us to a certain course of conduct, pleasure … and it becomes impossible to deny that all actions are governed by pleasure." Jevons evidently hoped to make his theory of behavior as nearly as possible true by definition. Later economists followed his lead. In the 1930s, in the course of a movement in economics known as *ordinalism*, it took them to a radically different version of the theory. The doctrine of ordinalism was proclaimed by Lionel Robbins in 1932. Ordinalist utility theory was formally set out in a pair of articles by John Hicks and R. G. D. Allen (1934). It was refined by later work into what is best called *axiomatic utility theory*. This is the version that is generally presented in textbooks of economics today.

Axiomatic utility theory is a theory of *preferences*. It assumes a person has preferences among all the alternative acts he/she might do. It treats a preference as a disposition to choose: to prefer one act to another is to be disposed to choose the first over the second if faced with a choice between them. The theory assumes that a person's preferences satisfy a number of axioms. Provided they do so, a formal proof shows that a mathematical function called a *utility function* can be defined to *represent* the preferences in a particular sense. The utility function assigns a number called a *utility* to each of the acts, and it does so in such a way that one act is assigned a greater utility than another if and only if it is preferred to the other. This is the sense in which the function represents the preferences. The dispositional definition of preference ensures that, when faced with a choice between two acts, the person always chooses the one he/she prefers. This one has the greater utility. Therefore, we may conclude that the person always acts to maximize his/her utility. In this way, utility theory is confirmed, but only because a person's utility is defined in such a way as to make it true. Utility is defined as what the person maximizes.

Bernoulli and Bentham's utility theories made substantive commitments about what a person maximizes, whereas the axiomatic theory does not. It is not committed to hedonism, or even to an assumption that people always act self-interestedly. Since the commitments of earlier theories were not very plausible, this was a significant advance in economics.

In a sense, ordinalism made utility theory true by definition. However, it did retain some substantive commitments. It is not always possible to define a utility function for a person in a way that makes him/her a utility maximizer. We can only be sure this is possible if the person's preferences satisfy the axioms of the theory. This is the theory's commitment: preferences satisfy the axioms. In effect, the axioms ensure that the person can be construed as maximizing something, and utility is defined as what she maximizes.

The principal axiom is *transitivity*: if one alternative is preferred to a second, and the second to a third, then the first is preferred to the third. If a person's preferences are transitive, they constitute an *ordering* of the options. The utility function represents this ordering. It puts the options in the same order as the preferences do. Any utility function that puts the options in this same order will represent the preferences equally well. Consequently, many different utility functions can represent the same preferences. The only significant feature of a utility function is its ordering of the options. A utility function defined this way is said to be *ordinal*, hence the name 'ordinalism.'

Axiomatic utility theory was later extended by John von Neumann and Oskar Morgenstern (1944) and by Leonard Savage (1954), independently of earlier work on the same lines by Frank Ramsey (1931). The extended theory recognizes that the outcome of an act is normally uncertain. It assumes that a person's preferences among acts are determined by the possible outcomes that may result from each act, and by the likelihood that they will result. It assumes that the preferences satisfy the same axioms as before, and also some more axioms that are concerned with the pattern of outcomes that may result from each act.

As in the simpler theory, a theorem shows that the preferences can be represented by a utility function. However, because the extended theory makes more assumptions, it is able to establish stronger conclusions. In the extended theory, the utility function assigns utilities both to acts and to their

possible outcomes. As before, one act is assigned a higher utility than another if and only if it is preferred. The special feature of the extended theory is that the utility of an act is the *expected utility* of its possible outcomes. (Expected utility is defined as the sum of the utilities of the possible outcomes, each multiplied by the outcome's probability.) Let us call this the 'expectation property' of the utility function. It means that, of two acts, the person always prefers the one that has the greater expected utility. That is to say, she maximizes expected utility. Once again, this is only because of the definition of utility. Provided a person's preferences satisfy the axioms, the person can be construed as maximizing the expectation of something, and utility is defined as that thing.

The extended theory is known as *axiomatic expected utility theory*. An important feature of it is that the utility function that represents preferences is not merely ordinal. Within the theory, a utility function order not only acts in accordance with the preferences but also has the expectation property. This turns out to imply that differences – increases or decreases – in utility can be significantly compared with each other. Many different utility functions can still represent the same preferences. However, if, according to one of these functions, the difference between the utilities of one pair of acts is greater than the difference between the utilities of another pair, then the difference is also greater according to any other function that represents the same preferences. A utility function in which differences of utility are comparable is said to be *cardinal*.

Utilitarianism and the Pareto Principle

Utilitarianism in a broad sense is simply a concern to promote people's well-being. More precisely understood, it is the view that one ought to act so as to maximize the arithmetic total of people's well-being. Important consequences follow from the precise version that does not follow from the broader one. For example, suppose we assume that the richer a person is, the less the extra well-being he/she will gain from adding an extra dollar to his/her wealth. This plausible assumption is known as *diminishing marginal benefit of wealth*. An argument in favor of equality can be derived from it together with precise utilitarianism. Diminishing marginal benefit implies that rich people generally derive less benefit from extra money than poor people do. Therefore, as extra money becomes available, it adds more to the total of well-being if it goes to the poor rather than to the rich. Precise utilitarianism therefore favors sending money toward the poor. This argument for equality is subject to important limitations, but it was nevertheless a significant contribution that economics made to utilitarian thinking. It was popular among economists before the 1930s, including Alfred Marshall (1920) and A. C. Pigou (1932).

However, the arrival of ordinalism brought skepticism about the measurement of well-being. If we are even to make sense of precise utilitarianism, we must be able to make sense of an arithmetic total of well-being. Two conditions are necessary for that. First, each person's well-being must be an arithmetic quantity. To be more precise, it must be a cardinal quantity, which is to say that differences – increases or decreases – in a person's well-being must be significantly comparable with each other. Second, differences in one person's well-being must be significantly comparable with differences in any other person's: we must be able to say whether a particular increase or decrease in one person's well-being is greater or less than a particular increase or decrease in another's. Ordinalism made economists skeptical about both these conditions.

The source of their skepticism was epistemic. The ordinalists regarded a person's preferences as observable in principle. Since they defined utility to represent preferences, they regarded utility too as observable in principle. Because of this, they assigned ordinal utility a high epistemic status; Robbins called it a 'scientific' concept. However, they were much more doubtful about other features of a person's psychology.

What about well-being? The ordinalists seem to have taken it for granted that a person prefers one thing to another if and only if it gives him/her more well-being. This view may be called *preferencism*. On the face of it, preferencism is implausible. For example, presumably a person might altruistically prefer one thing to another because it is better for someone else, even if it brings himself/herself less well-being. The ordinalist espoused the doctrine of *operationalism*. According to operationalism, the meaning of a concept such as well-being is given by the operations we use in order to apply it. The ordinalists recognized no operation that would allow us to apply the concept of well-being except the operation of observing people's preferences. They thought that the only meaning well-being can have is given by the preferencist formula. This means that utility measures well-being in an ordinal way: higher utility indicates higher well-being. The order of well-being – whether one thing is better or worse for a person than another – therefore counted as a scientific concept, but the ordinalists acknowledged no other aspect of well-being. Well-being could not be counted as an arithmetic quantity, nor could differences of well-being be compared between people. The preconditions of precise utilitarianism could therefore not be met.

Nevertheless, the ordinalists were utilitarian at heart. Consequently, they clung to the flimsy remnant of utilitarianism that survived their epistemic censure. They accepted the *Pareto criterion*, named in honor of Vilfredo Pareto (1927). The Pareto criterion says that one state of affairs is better than another if everyone either prefers the first to the second or is indifferent between them, and if someone strictly prefers the first. This sufficient condition for one alternative to be better than another was the sole evaluative principle that the ordinalists were prepared to recognize. Few pairs of alternatives satisfy this condition, so shrinking utilitarianism to the Pareto criterion cripples it. Ordinalism stultified welfare economics for decades. In particular, it denied that any judgment could be made about the value of equality.

A possible escape from the ordinalist straightjacket seemed to be offered by expected utility theory. Because this theory yields cardinal utility functions, some authors thought it might allow well-being to be measured cardinally. However, this thinking was confused about the relation between utility and well-being. Granted preferencism, utility measures well-being ordinally. Even granted preferencism and granted also that cardinal utility can be defined within expected utility theory, it does not follow that this cardinal utility measures well-being cardinally. Differences in utility are comparable, but it does not follow that they represent comparable differences in well-being.

Furthermore, the second condition of precise utilitarianism remains unsatisfied. Nothing in expected utility theory makes one person's well-being comparable with another's.

Nevertheless, John Harsanyi in the 1950s produced no less than two new and important arguments that were intended to show that, indeed, the cardinal utilities defined by expected utility theory do have a place in ethical judgments.

His first argument (Harsanyi, 1953) was this. Suppose we are comparing societies, asking which the best is. Harsanyi argued that this question could be answered by imagining a person who is allowed to choose among the different societies. He/she knows that he/she will become a member of the one he/she chooses but does not know which position he/she will occupy within it. He/she has an equal chance of finding the best position, or the worst, or anywhere else. Harsanyi claimed that the best society is whichever one he/she would choose under these circumstances. This is an appeal to the idea of a 'veil of ignorance,' which was later taken up by John Rawls (1972).

From the person's point of view, he/she is choosing between acts whose results are uncertain: the result of choosing a particular society will be that he/she ends up in an unknown position within that society. According to expected utility theory, for this person there will be a utility assigned to each position in the society, and the utility for him/her of entering the society will be the average of all these individual utilities. Therefore, of the alternative societies available, the person will choose the one that has the greatest average utility. If we assume every society has the same population, this is also the one that has the greatest total utility. According to Harsanyi, then, the best society is the one with the greatest total utility. This argument does indeed give a place in ethical judgments to utility as expected utility theory defines it.

Does it also support utilitarianism? Precise utilitarianism implies that the best society is the one with the greatest total of well-being. Therefore, if utility measures well-being cardinally, Harsanyi's conclusion resembles precise utilitarianism. However, Harsanyi's argument falls short of a complete argument for utilitarianism, because we still have no reason to think utility does measure well-being cardinally. Furthermore, the problem of comparing one person's well-being with another's remains.

Harsanyi's second argument (1955) was a formal theorem that is too complex to describe here. It is entirely distinct from the first argument, and it does not rely on a veil of ignorance. Still, it arrives at a similar conclusion, that the best society is the one with the greatest total of people's utility. For the same reasons as before, it does not amount to a complete defense of utilitarianism. Nevertheless, Harsanyi's two arguments cast an important new light on utilitarianism, and went a long way toward providing it with completely new foundations.

These foundations are based on preferences, and they depend on preferencism. Within moral philosophy as well as economics, it has become increasingly popular in recent years to base utilitarianism on preferences rather than, say, pleasure and pain. One reason is the prestige given to preference-based utilitarianism by Harsanyi's impressive work. His work issued, in turn, from the dramatic revision of utility theory caused by ordinalism. These are examples of the strong influence that economics has had on utilitarian thought.

Tools for Ethics

Economists have developed for their own purposes various formal tools that also turn out to be useful in ethics. Utility theory is one: apart from its contribution to utilitarian thought, it has other significant applications within ethics. For one thing, it constitutes a partial account of practical rationality; it specifies what rationality requires of a person's preferences. An account of practical rationality is an essential background to ethical theory. For another thing, utility theory can easily be converted from a theory of preferences to a formal theory of the structure of value, by applying its axioms to *betterness* instead of preferences (Broome, 1991).

Game theory is a second prominent example of a tool useful in ethics. The *prisoners' dilemma* (qv) and related 'games' have led to some radical rethinking of the nature of rationality. Some authors (e.g., Gauthier, 1985) have argued that they call for an understanding of individual rationality that somehow embodies an ethic of cooperation. *Bargaining theory*, too – a branch of game theory – has proved useful at the foundations of contractualism (qv) (see Gauthier, 1985; Binmore, 1994 and 1998).

A third example is *social choice theory*. Social choice theory is a branch of welfare economics that is principally concerned with *aggregation*. It examines how the interests of individuals are aggregated together to determine the overall goodness of a society. One example of a principle of aggregation is the utilitarian principle that overall goodness is the arithmetic total of people's well-being.

Social choice theory originated in Kenneth Arrow's *Social Choice and Individual Values*, published in 1951 (*see* Arrow, 1951). As an ordinalist, Arrow was concerned with aggregating specifically preferences. His conclusion was negative; he demonstrated that people's preferences cannot generally be aggregated together in a plausible way to determine an overall ordering of the alternatives, ranking them from best to worst.

As a result, since Arrow's time, social choice theorists have gone beyond the bounds of ordinalism. They are no longer concerned with aggregating preferences only. A leader in this movement is Amartya Sen (1970), whose *Collective Choice and Social Welfare* both codified social choice theory and set the scene for this new departure. Ordinalism combined a utilitarian concern for people's well-being with skepticism about a quantitative notion of good. One can give up the skepticism, and many economists nowadays unashamedly take for granted a quantitative notion of a person's good or well-being. Alternatively, one can give up the utilitarianism and focus concern on something that is less prone to skepticism. Sen (1987) himself focuses on people's *capabilities*, by which he means the collection of things that they are enabled to do by the resources at their disposal.

A new application for social choice theory has emerged in the philosophical discussion of 'moral uncertainty.' Moral philosophy is full of debate, disagreement, and uncertainty. Often a person or a government needs to make a decision among alternatives that raise moral issues, but where it is uncertain what the right moral principles to apply are. Each moral theory will rank the various alternatives, and different theories may well rank them differently. The decision has to be made somehow on the basis of these conflicting moral

considerations. How should one act when it is uncertain how one should act? That is the problem of moral uncertainty (Sepielli, 2009).

It is closely analogous to social choice theory. Each separate moral theory is analogous to an individual person. A theory's ranking of the alternatives is analogous to a person's ranking of alternatives, by his/her preferences or by some other criterion. The different theories' rankings have to be aggregated somehow to determine a choice, just as individual people's rankings have to be aggregated somehow to determine a social choice. Therefore, social choice theory can make a useful contribution to the philosophy of moral uncertainty.

Shared Topics

Economists and moral philosophers share an interest in several specific, practical subjects, but the cooperation between the disciplines in dealing with them is weaker than it should be. This section mentions two examples.

The Value of Life

For practical purposes, welfare economists are faced with the need to set a value on people's lives. For example, railways can be made safer by spending money on them. One benefit of doing so will be that fewer people are killed in train accidents. Economists need to decide how much money is worth spending to achieve this benefit. They need to determine how much good is done by saving a person's life or, equivalently, how much harm is done by a person's death. For another example, some very expensive drugs can extend a person's life for a few months. Is the benefit worth the cost?

Since ancient times, philosophers, too, have been interested in the value of life. Some modern discussion (e.g., Nagel, 1979; Feldman, 1991) has been stimulated by an argument of Epicurus (1926). Epicurus tried to show that death is harmless to the person who dies, and the modern issue is still whether death is an evil and whether life has value. Some other modern philosophical writing takes it for granted that life is valuable, and asks more quantitative questions. For example, one issue that has been discussed is whether the value of a life is more affected by events that happen later in life or by those that happen earlier (e.g., Velleman, 1991).

Much of the philosophical discussion is about the aggregation of a person's good over the lifetime. How so the good things that come to a person at different times in life come together to determine the goodness of life as a whole? The formal techniques for aggregation within economics can make an important contribution to the philosophical discussion. Conversely, the evaluation of life within economics clearly calls out for philosophical assessment. There has been some contact between the disciplines in this area (e.g., Broome, 2004), but there is room for much more.

Equality

For a long time, welfare economists have been concerned with the value of economic equality; examples of important contributions are by Atkinson (1970) and Sen (1973). The same subject has also recently become important in moral philosophy; leading works are by Parfit (1991) and Temkin (1993). In this area, too, contact between philosophers and economists has been limited, and both disciplines would benefit greatly from more.

The philosophers can fairly be said to have discovered an important distinction. Two very different moral concerns can have similar practical consequences. One – generally called *prioritarianism* – gives more importance ('priority') to improving the well-being of badly off people than to improving the well-being of well-off people. Prioritarianism will generally lead one to steer resources toward worse-off people, which will generally have the effect of promoting equality. However, equality is not its direct object. A different concern is simply to reduce the gap between the better-off and the worse-off; this is sometimes simply called *egalitarianism*.

The distinction is genuine, but because their effects are similar, it is difficult to specify the precise difference between prioritarianism and egalitarianism. The formal methods of economics have the resources to do so. Indeed, the economic literature on inequality implicitly contains detailed analyses of both concerns. Despite that, the distinction has not been clearly recognized in economics; it is something that economics needs to learn from philosophy. Conversely, all the formal analysis that is already available within economics could help philosophers clarify the distinction they are making and better understand its consequences. On the subject of equality, more cooperation would be very fruitful.

To summarize, economics and ethics have learned a great deal from each other, but they could learn a great deal more.

Bibliography

Arrow, Kenneth J., 1951. Social Choice and Individual Values. Yale University Press.

Atkinson, Anthony B., 1970. On the measurement of inequality. Journal of Economic Theory 2, 244–263.

Bernoulli, Daniel, 1738. Specimen theoriae novae de mensura sortis. Commentarii Academiae Scientiarum Imperialis Petropolitanae 5 (translated as 'Exposition of a new theory on the measurement of risk'). Econometrica 22 (1954), 23–36.

Bentham, Jeremy, 1798. An Introduction to the Principles of Morals and Legislation, Oxford.

Binmore, Kenneth, 1994 and 1998. Game Theory and the Social Contract, vols. I and II. MIT Press.

Broome, John, 1991. Weighing Goods. Blackwell.

Broome, John, 2004. Weighing Lives. Oxford University Press.

Epicurus, 1926. Letter to Menoeceus. In: Bailey, Cyril (Ed.), Epicurus: The Extant Remains. Oxford University Press, pp. 83–93.

Feldman, Fred, 1991. Some puzzles about the evil of death. Philosophical Review 100, 205–227.

Gauthier, David, 1985. Morals by Agreement. Oxford University Press.

Harsanyi, John, 1953. Cardinal utility in welfare economics and in the theory of risk-taking. Journal of Political Economy 61, 434–435.

Harsanyi, John, 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. Journal of Political Economy 63, 309–321.

Hicks, John, Allen, Roy George Douglas, 1934. A reconsideration of the theory of value. Economica 1, 52–76 and 196–219.

Jevons, William Stanley, 1871. The Theory of Political Economy. Macmillan.

Marshall, Alfred, 1920. Principles of Economics, eighth ed. Macmillan.

Mirrlees, James A., 1971. An exploration in the theory of optimum income taxation. Review of Economic Studies 38, 175–208.

Nagel, Thomas, 1979. Death. In: Mortal Questions. Cambridge University Press, pp. 1–10.

Neumann, John von, Morgenstern, Oskar, 1944. Theory of Games and Economic Behavior, second ed. Princeton University Press.

Pareto, Vilfredo, 1927. Manuel D'Economie Politique, second ed.

Parfit, Derek, 1991. Equality or Priority? The Lindley Lecture. University of Kansas.
Phelps, Edmund Strother, 1973. Taxation of wage income for economic justice. Quarterly Journal of Economics 87, 331–345.
Pigou, Arthur Cecil, 1932. The Economics of Welfare, fourth ed. Macmillan.
Ramsey, Frank, 1931. Truth and probability. In: Braithwaite, Richard Bevan (Ed.), Foundations of Mathematics and Other Logical Essays. Routledge and Kegan Paul.
Rawls, John, 1972. A Theory of Justice. Harvard University Press.
Robbins, Lionel, 1932. An Essay on the Scope and Nature of Economic Science. Macmillan.
Savage, Leonard J., 1954. The Foundations of Statistics. Wiley.
Sen, Amartya, 1970. Collective Choice and Social Welfare. North-Holland.
Sen, Amartya, 1973. On Economic Inequality. Oxford University Press.
Sen, Amartya, 1987. The Standard of Living. Cambridge University Press.
Sepielli, Andrew, 2009. What to do when you don't know what to do. Oxford Studies in Metaethics, volume IV, pp. 5–28.
Smith, Adam, 1776. An Inquiry into the Nature and Causes of the Wealth of Nations.
Temkin, Larry S., 1993. Inequality. Oxford University Press.
Velleman, David, 1991. Well-being and time. Pacific Philosophical Quarterly 72, 48–77.

Economics, History of

Craufurd Goodwin, Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Economic thought can be found in the early literature of every civilization. However, an economics discipline emerged only in the sixteenth, seventeenth, and eighteenth centuries in Western Europe. The outlines of orthodox modern economics appeared after the Marginal Revolution of the 1870s. Strands of heterodox economics took root soon after: macroeconomics, institutional economics, and Marxian economics. Study of the history of economics began in the eighteenth century as a way of dismissing or celebrating ancestors. In the twentieth century, major contributors to economics turned to history for the 'extended present' of their inquiry. In the second half of the twentieth century, economists turned away from their history because the opportunity cost of teaching and research seemed too great. In response, specialists established their own journals and societies and emphasized the teaching of undergraduates. In the twenty-first century, stimulating research has kept the history of economics vigorous and alive. Crises in economic affairs have turned some economists back in time to seek explanations. Challenges that remain for historians of economics include where to find a happy home in the academic landscape and where to conduct postgraduate training.

The History in a Nutshell

The roots of the subject that we know today as economics go back at least to the classical Greeks several hundred years before Christ, and economic ideas can be found in early Chinese, Indian, Islamic, and other non-Western literatures as well (Lowry, 1987). It is apparent that humans cannot engage in economic activity without at the same time reflecting upon it. These early economic ideas were usually embedded in other bodies of thought such as philosophy and theology and advocacy for particular policies or institutions. A discipline of economics did not take shape until the eighteenth century and a profession by which scholars could make a living did not appear until later still. Early thinking about economic questions typically centered on price determination and especially the impact of monopoly on markets. Scholars looked at the distribution of income and wealth, taxation, the role of government, and the functions of money. Influential thinkers in this prehistory of economics included the Greeks, notably Plato (*c*.427–347 BC) who proposed a model of an ideal economy, Aristotle (384–322 BC) who attempted to determine justice in an economic system, and Xenophon (*c*.430–355 BC) who proposed economic policies for the state and rules to govern public expenditures. Christian scripture in the Bible and writings of the early church fathers attempted to reconcile economic activity as it was evolving with their religious doctrine. During the Middle Ages, churchmen such as St Thomas Aquinas (AD 1225–74) and members of the School of Salamanca in the sixteenth century addressed economic questions that arose in the feudal system and especially in markets that so often were imperfect because of their relative isolation (Monroe, 1924).

A body of ideas that looked like what we call today economics began to emerge in the sixteenth, seventeenth, and eighteenth centuries, stimulated by the extension of world trade, the growth of national states, the use of money as the means of exchange, the scientific revolution, and the Renaissance in knowledge overall (Backhouse, 2002). Questions surrounding the functions of money and the conduct of trade became highly contentious and stimulated serious reflection from profound thinkers, for example, Jean Bodin (1530–96), William Petty (1623–87), and John Locke (1632–1704). These protoeconomists began and engaged vigorously in some of the first extended controversies over such issues as the case for colonies, protection of certain industries and trades, and responsibilities of the state for banking and currency. Participants in the extension of international commerce such as Thomas Mun and Dudley North combined curiosity about the process of which they were a part with a desire to influence affairs for their own benefit. During the eighteenth century, Enlightenment thinking that spread across Europe led to the first well-rounded and coherent bodies of economic thought that came to be called political economy (later classical economics), and were supported by enthusiastic groups of followers identified sometimes as 'schools.' These political economists attempted to explain economic phenomena by reference to systems of theory particular to their schools. The new subject flourished especially in France in the works of Richard Cantillon, Anne Robert Jacques Turgot, and Francois Quesnay; in Scotland in the writings of David Hume, Adam Smith, and James Steuart; and in Italy with Ferdinand Galiani. In most cases, political economy at this time emerged from and remained close to moral philosophy, and gained strength from the study of history and science and the general intellectual ferment of the times (Hollander, 1987). In the nineteenth century, classical economics was extended and developed by Thomas Robert Malthus, David Ricardo, Jeremy Bentham, Karl Marx, J.B. Say, John Stuart Mill, and many others. Subfields appeared that covered money and banking, international trade, colonization, and other specialized topics. Political economy was introduced somewhat hesitantly into the curricula of universities across Europe.

Most economists today might conclude that modern economics began only with what is now called 'neoclassical' economics, a body of thought that emerged in several countries more or less independently and expanded rapidly in four countries: England (William Stanley Jevons, Alfred Marshall, and Francis Edgeworth), France (Leon Walras), Austria (Carl

Menger), and the United States (John Bates Clark and Irving Fisher). The economics of each of the neoclassical pioneers had distinctive characteristics, but overall they shared at least two features. First, they emphasized the actions of individuals operating independently in markets as buyers and sellers while pursuing their perceived welfare, either profit for business people or 'utility' for consumers as it was called by some. Second, the behavior of these participants in the economy was described more and more by the neoclassicals using mathematical tools, beginning with simple algebra and the calculus (Black et al., 1973). This approach was intended to give greater rigor to the subject and to move it more and more into the category of 'science,' away from moral philosophy and psychology. Some of the main early critics of classical and neoclassical economics were economic historians centered in Germany (Wilhelm Roscher and Gustav Schmoller) and Great Britain (T.E. Cliffe Leslie and William Ashley) who questioned the use of abstraction to understand human behavior, and moral critics like Charles Dickens, John Ruskin, and Anthony Trollope who disputed the social consequences of competitive market activity (Tribe, 1995; Koot, 1987). They described in graphic detail what might be called today negative externalities of market phenomena.

In the twentieth century, with a steadily increasing number of economists entering the field, and the subject gaining acceptance as a respected discipline in the academy, the history of economics becomes more complex. Neoclassical theory was increasingly more technical with important additions like general equilibrium theory and the theory of games. The new field of econometrics used statistics to test, apply, and extend economic theory. Additional applied subfields of economics took shape such as public economics, industrial organization, law and economics, and labor economics. More specialized fields also appeared, concerned with such topics as health, the environment, and education. Increasingly, economics was carried beyond faculties of arts and sciences of colleges and universities where it first took root, into schools of business, law, agriculture, natural resources, public policy, and beyond. Economic reasoning was applied also with increasing frequency in such fields as political science and sociology. All of these engagements of economists with a myriad of new questions and new institutions, often with distinctive national perspectives, enriched and complicated the history of the subject.

Beyond the extension of neoclassical economics in the twentieth century, three strands of heterodox economics attracted economists to their banners. Stimulated by increasingly severe business cycles and depressions, beginning in the 1890s and reaching a peak in the 1930s, some economists using their new tools of analysis began to question the automaticity of a free market system and wondered whether the state had a responsibility to intervene in recessions and pay attention to such large aggregates as employment, production, price levels, and growth rates. Early macroeconomists, as they came to be called, like John Hobson in Britain, did not receive much attention, but John Maynard Keynes with his *General Theory of Employment, Interest and Money* (1936) was revolutionary (Backhouse and Bateman, 2006). This work challenged the discipline, affected public policy, and has remained a subject of controversy into the twenty-first century. In addition to suggesting new perspectives on the economy and new tools of analysis, Keynes argued that economists needed to address more fully certain variables like human behavior that had been set aside in neoclassical economics. He also provided the basis for optimism among those economists who before had viewed the tragedy of the Great Depression with resignation and almost with despair. Keynes's disciples included some of the most distinguished economists of the second half of the twentieth century, such as Joan Robinson, Richard Kahn, Paul Samuelson, and James Tobin, and the entry of his ideas into public policy constitutes some of the most fascinating history of economics in action. Keynes was an economist with an exciting personal life, both outside and inside the academy, in business and the arts (Skidelsky, 1983). He has seldom failed to intrigue historians of the twentieth and twenty-first centuries.

The second element of heterodox economics, or at least economics, that was different from the neoclassical mainstream, appeared in the United States in the late nineteenth century and flourished into the second half of the twentieth century. Practitioners were usually well trained in conventional neoclassical theory, but whether through experience in their new country or training in the historical tradition in Germany they were dissatisfied with received doctrine. In some cases under the influence of Thorstein Veblen, the pioneer in this tradition, they viewed the economy as made up of biological organisms, constantly evolving as Darwin had explained (Veblen, 2007). The metaphor for the economy that appealed to them was a community of complex, ever-changing individuals usually in disequilibrium rather than moving toward equilibrium in the mechanical clock metaphor preferred by some neoclassicists. The disciplinary boundaries of this American evolutionary or institutional economics were much wider than the neoclassical ones, encompassing history, anthropology, sociology, psychology, and even literature and the arts (Rutherford, 2011). A high premium was placed on the accumulation of data, the study of history, and a focus on public policy problems. They were quite ready to contemplate interference with markets, and the policy implications of this new approach often fitted well with the Progressives who emerged as a strong force in American politics early in the twentieth century. Several venerable structures in American professional economics can trace their origins to the institutionalists, notably the National Bureau of Economic Research and the Brookings Institution.

The third type of heterodox economics that challenged neoclassical economics grew out of the work of Karl Marx, both his incendiary *The Communist Manifesto* (1848) with Friedrich Engels and his major work in political economy *Capital* (1867, 1885, and 1894). This strand had a good deal in common with that of the historical economists and the American institutionalists (Howard and King, 1975). In particular, all three saw the participants in the economy as complex, evolving creatures, which could not be understood through neoclassical models alone that assumed merely profit-maximizing and utility-maximizing behavior. But Marxian economics was more than an analytical system. It was also a call to arms to improve a world in which 'contradictions of capitalism' steadily were becoming more severe and through a dialectical process would

sooner or later precipitate revolutionary change. Like the other strands of heterodox economics, Marxian doctrine connected with various disciplines like history and philosophy from which it drew strength.

Why Should Anyone Become Engaged with the History of Economics? And Why Not?

Why has anyone bothered to explore and explain the history of economics? One obvious answer is because they have been intrigued by questions of how the modern discipline has come to its current condition. This curiosity increased as economics has taken on importance in national life. But there are other reasons. Many of the earliest economists looked back into the history of their subject to make the case for their own priority in it. Some denigrated their predecessors to strengthen the case for their own ascendance. In this spirit, the mercantilists described and criticized the work of the church fathers. The classicals attacked the mercantilists. Others applauded their forebears. The neoclassicals, and especially William Stanley Jevons and Irving Fisher, searched for pioneering predecessors, such as Augustin Cournot, to compare favorably with such prominent contemporary classicists as John Elliot Cairnes and John Stuart Mill. Similarly, John Maynard Keynes looked for sympathetic ancestors to support the virtues of his 'new' economics. He attacked David Ricardo for his kind of theorizing and praised Thomas Robert Malthus for anticipating some of his ideas. The American institutionalists Thorstein Veblen and John Kenneth Galbraith condemned the entire neoclassical project. There were few scholars before the twentieth century who examined the history of economics for its own sake. One exception was the Scottish scholar-journalist J.R. McCulloch (1845).

It was customary for economists in the early twentieth century to include some history of the theoretical or empirical work on which they were engaged in lectures or publications on the ground that economics had what Kenneth Boulding called 'an extended present.' Claims made about a subject of contemporary interest several decades before, it was thought, were likely to remain unsettled and therefore of legitimate current interest. This view of the study of the history of economics as a legitimate part of current inquiry was given a fillip in the first half of the twentieth century by several leaders in economics who included substantial historical components alongside their work on theory or policy. These include, among others, Joseph Schumpeter (*History of Economic Analysis*, 1954), Jacob Viner (*Studies in the Theory of International Trade*, 1937), Lionel Robbins (*The Theory of Economic Policy in Classical Political Economy*, 1978), George Stigler (*A History of Production and Distribution Theories*, 1941), and Joseph Spengler (*French Predecessors of Malthus*, 1942). The history of economics at this time consisted mainly of careful readings of historical texts and explanation of what these early authors meant as seen through the eyes of a modern economist. These works contained impressive historical scholarship by authors of unquestioned prominence who were known primarily for their theoretical or applied work but who found that the history of economics was relevant to their current concerns. Another argument for the 'extended present' view of the history of economics was that without it economists were likely to reinvent the wheel or commit errors that had been corrected long since. With so many historical monographs of this type being published well into the second half of the twentieth century, little need was perceived for conventional textbooks in the history of economics and the few that were used widely, such as those by Charles Gide and Charles Rist (1948) and by Alexander Gray (1931), were in print for a long time.

A marked turning away from the history of economics as a distinct subject within the economics discipline occurred during the second half of the twentieth century, especially in the United States, which by this time had come to dominate the profession globally. Separate courses in the subject were discontinued at the postgraduate level, especially in North American PhD programs. The premier journals in economics declined any more to publish, or even to review, manuscripts on the history of the field, except for the *Journal of Economic Literature* and the *Journal of Economic Perspectives*, which were aimed at recent history. The national meetings of economists had fewer and fewer sessions dealing with the history of economics and few job openings were posted. Various reasons have been suggested for this decline of interest and attention. One is that with the increasing sophistication of the theoretical core of economics more teaching time was needed there. Another is that relatively new specialties like econometrics made a strong case for more teaching space, and consequently the opportunity cost of teaching history steadily rose. But perhaps the most important reason for the decline of history is that after World War II, in which economics made a valued contribution to the War effort, the discipline attempted strenuously to define itself once and for all as a real hard science and not a literary subject. One triumphant accomplishment was to be included within the new National Science Foundation rather than the National Endowment for the Humanities. Another was to be included among the Nobel Prizes in 1968 with a prize given by the Bank of Sweden in memory of Alfred Nobel. Leaders of the economics discipline apparently came to believe that the nature of science was to retain in the current literature all that was 'true' and to discard all that was 'false.' By this reasoning, any interest in the history of science was either redundant or antiquarian, and historical study had no claim on scarce appointments, research funds, space in publications, or teaching time. What had been sophisticated historical introductions in classes and publications were gradually replaced by 'literature reviews' of current material. What came to pass for history in the higher reaches of the discipline after this was often little more than hagiographic celebration of prominent figures of the past and present.

Yet, despite this loss of respect within the mother discipline, the history of economics as a field did not go away and die. In fact, quite the opposite, in many respects it grew and prospered. During the 1960s and 1970s, as history was purged from the mainstream of the discipline it found new life on its own. Specialized international journals sprang up, the first being *History of Political Economy*, followed by the *Journal of the History of Economic Thought*, the *European Journal of the History of Economic Thought*, *History of Economic Ideas*, and *Research in the History of Economics Thought and Methodology*. Professional societies were established in North America, Europe, and some individual countries including France, Japan, Australia, and

Italy. Country-specific journals were started, connected in most cases with the national societies: *History of Economics Review* (Australia), *Cahiers d'economie politique* (France), *Storia del pensiero* (Italy), *The History of Economic Thought* (Japan), and others. A number of valuable resources became available on the Internet including the Economic Research Network's *History of Economics Journal* and *Weboeconomia*. A *Guide to Archival Sources for British and Irish Economists*, supported by the Royal Economic Society, was published in the 1970s.

Paradoxically, within a few decades of the history of economics being removed from the curricula of many colleges and universities, meetings of the international societies that sprang up in response were attracting several hundred attendees. By the 1990s, the annual meetings of the American Economic Association began again to include sessions on history arranged by and in association with the History of Economics Society. While the teaching of the history of economics virtually disappeared at the graduate level in the United States, it increased among undergraduates, especially at liberal arts colleges and less prominent universities. Textbooks to satisfy all ideologies and teaching styles were published. The most notable was Mark Blaug's *Economic Theory in Retrospect* (1962), which was applauded by some because it interpreted and judged early works through modern eyes, and was criticized by others because of its 'Whig' bias of appraising historical contributions according to their congruence with modern theory. Other influential early textbooks in the years when the history of economics struck off on its own included Henry Spiegel's *The Growth of Economic Thought* (1983) and Robert Heilbroner's *The Worldly Philosophers* (1953). Perhaps upward of a thousand scholars worldwide by the end of the twentieth century would admit to being serious students of the field.

To supplement the vigorous institutional response to the abandonment of the history of economics by the economics discipline in the 1960s, there was an intellectual response that stimulated research in the subject. In part because of the loss of firm leadership of the subject by prominent historian-theorists the subject was opened up to new approaches and new hypotheses. From the history and philosophy of science came the theory of scientific revolutions of Thomas Kuhn (1970) and the scientific research programs of Imre Lakatos (Lakatos and Musgrave, 1970). From philosophy came the complex analytical perspective of Michel Foucault and from the economic historian Donald McCloskey the conception of economics as a conversation and the exercise of rhetoric (McCloskey, 1985). Philip Mirowski introduced a controversial interpretation of the marginal revolution that suggested much of the economic theory of the time was borrowed from contemporary physics (1989). Judy Klein, Mary Morgan, and others recorded the close relationship between the development of economics and World War II, especially in research institutions like the RAND Corporation, the Cowles Commission, and the Brookings Institution (Klein and Morgan, 2001). E. Roy Weintraub described the mathematization and formalization of the subject in the twentieth century (2002). The study of economic methodology and the philosophy of economics that had its roots in the history of the subject moved out on its own with separate journals and fields. All of these approaches stimulated research projects and conferences to explore more widely the applicability of the new analytical perspectives to economics.

What sustains this interest in the history of economics after the setbacks of the mid-twentieth century? Several answers to this question can be suggested. Despite the decline in teaching at the graduate level, the history of economics as a distinct subject continues to appeal to undergraduates precisely because they have not been provided the kind of background that typically was used to introduce courses 50 years ago. Where did these tools and models come from, they ask, and are there alternatives to them? Why did the dominant theory triumph? How is modern economics used in our society? What is the relationship between the classroom and the real world? What questions are answered unambiguously and which ones remain in dispute? An explanation for the continuing flow of research and publications in the history of economics is more complex. Some faculty members have moved into historical studies from various other subfields of economics because they have been alienated by the level of abstraction in economics or by the ideological overlay that they perceive in the discipline, which makes them uncomfortable. Both students and faculty find in the history of economics the opportunity to maintain respectable continuing contacts with heterodox traditions that are not always tolerated by the mainstream. For example, both Austrian and Marxist economists are well represented in the history of economics. Opportunities for interdisciplinary adventures also abound in the history of economics, with scholars in the humanities and the arts as well as with the other social sciences, and this pleases college administrators as well as teachers and students. Answers to questions are seldom provided unambiguously and answered conclusively in the history of economics, and so there are ample opportunities for vigorous discussion, creative writing, and debate by students and faculty, a style of inquiry often unwelcome in the modern economics discipline.

The History of Economics in the Twenty-First Century

There are reasons for optimism and also for unease about the future of the history of economics in the twenty-first century. On the positive side the Great Recession that began in 2007 gave a notable boost to the subject. Doctrines that had been perceived by some to be settled verities suddenly were brought into question. Do financial markets really correct themselves on their own after a crisis like the mortgage meltdown as competitive market theory suggests? What can earlier explanations of financial turmoil contribute? Has Keynesian economics been discredited after all? Or Alvin Hansen's stagnation thesis? What did Hayek actually say? If welfare provisions are truly in peril did the fault lie to some degree in the doctrine from which they sprang? If new global policies and institutions are required to restore prosperity what could we learn from the theory that underlay the old structures? Both teaching and research in the history of economics were stimulated by these questions and others like them. The answers to many of the questions clearly demanded attention to history, and it was discovered that history could be a liberating force to think new thoughts. Approaches taken to questions in the history of economics over the last decades have been more diverse than the traditional approach of careful and reverent exegesis of economists' texts, which was the standard approach

through the 1960s. Recently, attention has been paid to the activities of departments of economics, rather than simply prominent individuals (notably, departments at the University of Chicago and MIT), other creative communities of various kinds, and schools of thought like the Austrians, the Keynesians, and the American institutionalists. New opportunities for research with this greater breadth have been provided by collections of papers and manuscripts of economists increasingly made available in university libraries such as those of Columbia, Chicago, Harvard, Duke (the Economists' Papers Project), and Stanford (Hoover Institution), as well as in the Presidential Libraries of the National Archives and Records Service. The improved availability of economists' papers must be balanced by the reality that in the twenty-first century many records of correspondence, reading notes, drafts of research results, and other materials are in electronic form and may be lost to scholars when they are deleted. More and more of the collected published writings and letters of major economists are being carefully edited and published (e.g., *The Collected Works of Friedrich Hayek* and the letters of Dennis Robertson, both in progress in 2014). Interpretive biographies have provided increased understanding of major figures such as Keynes (by Donald Moggridge, 1992; Robert Skidelsky, 1983), Harry Johnson (Moggridge, 2008), Friedrich Hayek (Bruce Caldwell, 2004), and Lionel Robbins (Susan Howson, 2011). For the first time some funding sources have given priority to the history of economics as an entry point to the broader discipline. Examples are the Institute for New Economic Thinking, supported by George Soros, and the National Endowment for the Humanities.

New vigor has been added to the history of economics in the twenty-first century also by arrival in the field of scholars from other disciplines who might have perceived barriers to entry in the twentieth century but now are warmly welcomed. Examples include the intellectual historians Michael Bernstein whose *A Perilous Progress: Economists and Public Purpose in Twentieth-Century America* (2001) described economists in the policy sphere, and Angus Burgin whose *The Great Persuasion: Reinventing Free Markets since the Depression* records economists in the growth of the neoliberal tradition. From the history of science, Mary Morgan (2012) has examined the economist's use of models and with Harro Maas has made use of the 'witness seminar' as an historical technique (2012). The sociologist Marion Fourcade has compared the growth of professional traditions among economists in several countries (2009).

The challenges facing historians of economics in the twenty-first century are fully as substantial as the opportunities. The first challenge is to establish where the subject may settle in the academic landscape. At present, most of the historians rest somewhat uncomfortably in second-tier economics departments and more comfortably in liberal arts colleges, earning their keep for the most part by teaching courses in conventional economics subjects in which they have been trained, but perceived by their economist colleagues, who by this date have been exposed to little history, as puzzling anomalies seemingly unrelated to the main purposes of an economics department. It has been suggested that historians of economics might better migrate to history, philosophy, or history of science departments, but most wish not to do so as they think of themselves as economists; in any event, it is doubtful that they would find a warm welcome in these other departments that are themselves frequently under stress in their college and university homes. The continued welcome to historians of economics in economics departments is likely to depend on their continued ability to attract good undergraduate students to their classes and to offer an appealing curriculum whose opportunity cost is not too high. A particular threat to the survival of historians in economics departments, especially outside the United States, is the practice by rating agencies of appraising departments by the volume of their research output and then classifying publications by historians of economics as history rather than economics.

A deeper challenge to the survival of the history of economics in its current place and its current form is the almost total withering away of postgraduate training in the field. Here and there a course is offered still at the graduate level, but nowhere is there a dedicated PhD program. This means that virtually all of the new entrants to the field must move laterally from some other discipline or some other specialty in economics and without any special training in historical methods or other tools of the trade such as oral history or use of archives. An effort is being made at Duke University to revive doctoral training as part of the activities of the Center for the History of Political Economy, and there are also hopeful reports of graduate students at several universities requesting the reinstatement of history courses.

See also: Alternative Schools of Economic Thought; Economic Education; Economic History, Quantitative: United States; Economic History; Keynes, John Maynard (1883–1946); Macroeconomics, History of up to 1933; Malthus, Thomas Robert (1766–1834); Marx, Karl (1818–83); Microeconomics, History of; Political Economy, History of; Ricardo, David (1772–1823); Simon, Herbert (1916–2001); Smith, Adam (1723–90); Stigler, George Joseph (1911–91); Veblen, Thorstein (1857–1929); von Hayek, Friedrich A. (1899–1992).

Bibliography

Backhouse, Roger, 2002. The Ordinary Business of Life: A History of Economics from the Ancient World to the Twenty-First Century. Princeton University Press, Princeton.

Backhouse, Roger, Bateman, Bradley W., 2006. The Cambridge Companion to Keynes. Cambridge University Press, Cambridge, UK; New York.

Bernstein, Michael A., 2001. A Perilous Progress: Economists and Public Purpose in Twentieth-Century America. Princeton University Press, Princeton, NJ.

Black, R.D., Collison, Coats, A.W., Goodwin, Craufurd, D.W., 1973. The Marginal Revolution in Economics; Interpretation and Evaluation. Duke University Press, Durham, NC.

Blaug, Mark, 1962. Economic Theory in Retrospect. Irwin Series in Economics. R.D. Irwin, Homewood, IL.

Burgin, Angus, 2012. The Great Persuasion. Harvard University Press, Cambridge, MA.

Caldwell, Bruce, 2004. Hayek's Challenge: Intellectual Biography of F.A. Hayek. University of Chicago Press, Chicago.

Fourcade, Marion, 2009. Economists and Societies: Discipline and Profession in the United States, Britain, and France, 1890s to 1990s. Princeton Studies in Cultural Sociology. Princeton University Press, Princeton.

Gide, Charles, 1847–1932. 967. 1948. A History of Economic Doctrines from the Time of the Physiocrats to the Present Day, 1948. Harrap, London.

Gray, Alexander, Sir, 1882–1968. 1931. The Development of Economic Doctrine; an Introductory Survey London, 1931. Longmans, Green, New York.
Heilbroner, Robert L., 1999. The Worldly Philosophers: The Lives, Times, and Ideas of the Great Economic Thinkers, rev seventh ed. Simon & Schuster, New York.
Hollander, Samuel, 1987. Classical Economics. Oxford Oxfordshire, New York, USA (B. Blackwell).
Howard, Michael C., King, J.E., 1975. The Political Economy of Marx. Modern Economics. Longman, Harlow, England.
Howson, Susan, 2011. Lionel Robbins. Historical Perspectives on Modern Economics. Cambridge University Press, New York.
Keynes, John M., 1947. The General Theory of Employment, Interest and Money. Macmillan and Co., Ltd., London.
Klein, Judy L., Morgan, Mary S., 2001. The age of economic measurement. History of Political Economy. Annual supplement to vol. 33. Duke University Press, Durham, NC, London.
Koot, Gerard M., 1987. English Historical Economics, 1870–1926: The Rise of Economic History and Neomercantilism. Cambridge University Press, New York.
Kuhn, Thomas S., 1970. The Structure of Scientific Revolutions. In: International Encyclopedia of Unified Science. Foundations of the Unity of Science, second ed., vol. 2 (2). University of Chicago Press, Chicago.
Lakatos, Imre, Musgrave, Alan, 1970. Criticism and the growth of knowledge. In: Proceedings/International Colloquium in the Philosophy of Science, vol. 4. Cambridge University Press.
Lionel, Robbins R., 1978. The Theory of Economic Policy in English Classical Political Economy, second ed. Porcupine Press, Philadelphia.
Lowry, S. Todd, 1987. The Archaeology of Economic Ideas: The Classical Greek Tradition. Duke University Press Cambridge University Press, Durham.
McCloskey, Deirdre N., 1985. The Rhetoric of Economics. Rhetoric of the Human Sciences. University of Wisconsin Press, Madison, WI.
McCulloch, J.R., 1845. The Literature of Political Economy: A Classified Catalogue of Select Publications in the Different Departments of that Science, with Historical, Critical, and Biographical Notices. Reprinted by the London School of Economics and Political Science. University of London, London.
Mirowski, Philip, 1989. More Heat than Light: Economics as Social Physics: Physics as Nature's Economics. Cambridge University Press, Cambridge; New York.
Moggridge, D.E., 1992. Maynard Keynes: An Economist's Biography. Routledge, London; New York.
Moggridge, D.E., 2008. Harry Johnson: A Life in Economics. Historical Perspectives on Modern Economics. Cambridge University Press, Cambridge; New York.
Monroe, Arthur Eli, 1885–1924. Early Economic Thought; Selections from Economic Literature Prior to Adam Smith. Cambridge. Harvard University Press, 1951 [*c.*1924].
Morgan, Mary S., 2012. The World in the Model Electronic Resource: How Economists Work and Think. Cambridge University Press, New York.
Morgan, Mary S., Maas, Harro, 2012. Observing the Economy: Historical Perspectives. In: History of Political Economy. Annual Supplement, vol. 2012. Duke University Press, Durham; London.
Morgan, Mary S., Rutherford, Malcolm, 1998. From Interwar Pluralism to Postwar Neoclassicism. In: Annual Supplement to History of Political Economy, vol. 30. Duke University Press, Durham, NC.
O'Brien, D.P., 1975. The Classical Economists. Clarendon Press, Oxford.
Rutherford, Malcolm, 2011. The Institutionalist Movement in American Economics 1918–1947. Cambridge University Press, New York.
Schumpeter, Joseph A., Schumpeter, Elizabeth B., 1954. History of Economic Analysis. Oxford University Press, New York.
Skidelsky, Robert, 1983. John Maynard Keynes: A Biography. Macmillan, London.
Spengler, Joseph J., 1942. French Predecessors of Malthus; A Study in Eighteenth-century Wage and Population Theory. Duke University Publications, Durham, NC. Duke University Press.
Spiegel, Henry William, 1983. The Growth of Economic Thought. Duke Universty Press, Durham, NC.
Stigler, George Joseph, 1941. Production and Distribution Theories: The Formative Period. Macmillan Co., New York.
Tribe, Keith, 1995. Strategies of Economic Order: German Economic Discourse, 1750–1950. In: Ideas in Context, vol. 33. Cambridge University Press, Cambridge England; New York.
Veblen, Thorstein, 2007. The Theory of the Leisure Class. Oxford University Press Inc., New York.
Viner, Jacob, 1937. Studies in the Theory of International Trade. Harper & brothers, New York, London.
Weintraub, E. Roy, 2002. How Economics Became a Mathematical Science. Science and Cultural Theory. Duke University Press, Durham, NC.

Economics, Philosophy of

Daniel M Hausman, University of Wisconsin–Madison, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Economics is of philosophical interest for three reasons. First, its methodological peculiarities: its extreme simplifications, its inexactness, its reliance on a deductive method, and the importance of causal claims concerning mechanisms, tendencies and capacities. Second, economics depends on a theory of rationality and has implications for the understanding of preference, utility, and self-interest. Finally, economics is relevant to ethical questions, particularly concerning efficiency and welfare. Philosophical reflection is of interest to economists because it can help to clarify methodological problems, improve the theory of rationality, and unscramble, expand, and sharpen the ethical implications. Although the philosophy of economics addresses many topics, a great deal of it has focused on the specific model of rationality and choice that is embedded in mainstream microeconomics. But the failure of economists to anticipate the financial crisis of 2008 and to provide unambiguous advice about how to restore prosperity has increased interest in the methodology of macroeconomics.

Like many of the social sciences, economics grew out of philosophy, and the concerns of economists continue to intersect with those of philosophers. Philosophical reflections on (1) scientific method and social ontology; (2) the nature of rationality, self-interest, and preference; and (3) welfare, justice, equality, and freedom are of abiding significance to economists and other social scientists.

Economic Methodology

Philosophical reflection on economics is ancient, but the conception of 'the economy' as a distinct object of study dates back only to the eighteenth century. Aristotle addresses problems of economics mainly as problems of household management. Scholastic philosophers addressed ethical questions concerning economic behavior, and they condemned 'usury' – that is, the taking of interest on money. With the increasing importance of trade among nation-states in the early modern period, 'mercantilist' philosophers and pamphleteers addressed questions concerning the balance of trade and the regulation of the currency. Only in the work of the physiocrats and especially of Adam Smith did scholars begin to think of the economy as an object of study with its own principles and laws.

Eighteenth century philosophers wrote in the shadow of Newton's towering accomplishments. David Hume is unabashed about his hopes to develop a science of mind and society in the image of Newton's science of the solar system. To that end, he seeks out general laws of individual thought and action, out of which larger-scale orderly relations will arise, in just the way that patterns in planetary motion arise from the laws of motion and gravitation governing individual bodies. Thus Hume traces the rise in prices and the temporary increase in economic activity that follow an increase in currency to the perceptions and actions of individuals who first spend the additional currency.

It then lay to Adam Smith to extend such a method to his systematic *Inquiry into the Nature and Causes of the Wealth of Nations* (1776) and to make explicit an implication of previous causal inquiries such as those of Hume, the physiocrats, and of many of the mercantilists. That implication (which is arguably the foundation for all social science) is that the social, aggregative implications of individual choices are often *unintended*. In spending their additional gold imported from abroad, traders do not intend to increase the price level. But that is what they do nevertheless. Unintended regularities that arise from the intentional choices of individuals can rule those individuals just as surely as the regularities of nature, and, like nature, societies can be objects of scientific investigation.

Classical Economics and the Method a priori

Adam Smith wrote little explicitly about the method of economics. The first extended reflections on economic methodology had to wait until Nassau Senior (1836) and John Stuart Mill (1836). Their essays must be understood against the background of the prevailing economic theory. Like Smith's economics (to which it owed a great deal) and modern economics, the 'classical' economics of the middle decades of the nineteenth century traced economic regularities to the choices of individuals facing social and natural constraints. But, as compared to Smith, economists relied more heavily on severely simplified models. David Ricardo's *Principles of Political Economy* (1819) depicts a society in which wages above the subsistence level lead to an expanding labor force, which in turn requires more intensive agriculture or cultivation of inferior land. The extension of cultivation leads to lower profits and higher rents. Economic development leads to a gloomy stationary state in which profits are too low to command any net investment, wages return to subsistence levels, and only the landlords are affluent. The data available to classical economists were never in accordance with the trends that the theory predicted. Yet the theory continued to hold sway for more than half a century, and the unfavorable data were explained away as due to 'disturbing causes.'

It is not surprising that Mill's account of the method of economics would emphasize the relative autonomy of the theory. Mill distinguishes two kinds of inductive methods. The method a posteriori is a method of *direct* experience. It is only suitable for phenomena in which few causal factors are operating or in which experimental controls are possible. Mill's famous methods of induction are detailed specifications of the

method a posteriori. In his method of difference, for example, one holds fixed every causal factor except one and checks to see whether the effect ceases to obtain when that one factor is removed.

Mill argues that direct inductive methods cannot be used to study phenomena in which many causal factors are in play. If, for example, economists attempt to investigate whether tariffs enhance or impede prosperity by comparing the prosperity with and without tariffs, the results will be irregular and unreliable because other causes besides tariffs will also differ across societies and times. So economists need instead to employ the method a priori. Despite its name, this is an *inductive* method, but it is an indirect inductive method. In Mill's view, scientists first determine the laws governing individual causal factors in domains in which direct inductive methods are applicable. Having then determined the laws of the individual causes, scientists investigate their combined consequences deductively. Finally, there is a role for 'verification' of the combined consequences, but owing to the causal complications, the testing of the conclusions serves only as a check on the deductions and as an indicator of whether there are significant disturbing causes that scientists have not yet accounted for. Mill gives the example of the science of the tides. Physicists determine the law of gravitation by studying planetary motion, in which gravity is the only significant causal factor. Then they develop the theory of tides deductively from that law and information concerning the positions and motions of the moon and sun. The implications of the theory will be inexact and sometimes badly mistaken, because many subsidiary causal factors influence tides. By testing the theory, scientists can uncover mistakes in their deductions and evidence concerning the role of the subsidiary factors. Because of the causal complexity, such testing does little to confirm or disconfirm the law of gravitation, which has already been established.

Because economic theory includes only the most important causes and necessarily ignores many minor causes, its claims, like claims concerning tides, are inexact. Its predictions are imprecise, and sometimes dead wrong. But it is possible nevertheless to develop and confirm economic theory by first establishing in simpler domains the laws governing the major causal factors and then deducing their combined consequences in different circumstances. For example, statistical data tell a mixed story about the relationship between minimum wages and unemployment; and there are no data concerning the consequences for employment of an extremely high minimum wage. On the other hand, everyday experience teaches economists that firms can choose among more or less labor-intensive processes and that a high minimum wage will make more labor-intensive processes more expensive. Since economists also have good reason to believe that firms try to keep their costs down, they have good reason to believe that a high minimum wage will increase unemployment.

According to Mill, economics is not only inexact and committed to the method a priori. In addition, he maintains that it is a *separate science*. What distinguishes economics as a discipline is both its concern for a certain domain of phenomena and its limitation to a particular set of causal factors, which predominate in this domain. With respect to this domain, economists can (at a certain level of approximation) ignore the myriad causal factors that influence all social phenomena (and that are the subject matter for other social sciences) and develop economics separately. In defending a view of economics as in this way inexact and separate and a view of economists as following the method a priori, Mill was able to reconcile his empiricism and his commitment to Ricardo's economics.

Although challenged later in the nineteenth century by dissident economists who believed that economic theory was too remote from the contingencies of policy and history, Mill's methodological views dominated the mainstream of economic theory for over a century. Mill's vision survived the transformation of economics from classical to neoclassical, and it is clearly discernible in the most important methodological treatises concerning neoclassical economics, such as John Neville Keynes' *The Scope and Method of Political Economy* (1891) or Lionel Robbins' *An Essay on the Nature and Significance of Economic Science* (1935). Hausman (1992) argues that current methodological practice closely resembles Mill's methodology.

Twentieth Century Philosophy of Science and Economic Method

Beginning in the 1930s, mainstream economists began to have a bad conscience about their traditional methodology, which some saw as insufficiently empiricist. Robbins is a transitional figure, because at the same time that he clung to Mill's methodology (1935: chap. 4) and to the identification of the domain of economics with the predominance of particular causal factors (the allocation of scarce means that have alternative uses – 1935: chap. 1), he also offered an apparently empiricist critique of interpersonal utility comparisons as untestable value judgments (1935: chap. 6). Terence Hutchison argued that the propositions of pure theory are so hedged with *ceteris paribus* conditions that they are not testable (1938). Paul Samuelson argued for a need to separate the 'operationally significant' economic wheat from the chaff, and in his theory of revealed preference he provided a model for how to do so (1947). Other economists cited survey data to argue that the theoretical propositions of economics were false. The confusing methodological situation stabilized in the 1950s with arguments by Fritz Machlup (1955) and especially Milton Friedman (1953) that economists need be concerned to fit only data concerning prices and quantities and that they need not be concerned about the 'realism of assumptions.'

Although confused, mistaken, and inconsistent with the practice of mainstream economists – including his own practice – Friedman's views have dominated the methodological conceptions of mainstream economists over the past two generations. His views are confused because they conflate many different kinds of 'assumptions' and many notions of 'realism.' Basic generalizations, propositions concerning initial conditions, and antecedents of conditional claims are all called 'assumptions.' Assumptions are called 'unrealistic' if they are false, incomplete, or not approximately true. Friedman's views are mistaken, because (as Mill already emphasized) market data are generated by too many causal factors to provide effective tests of economic theory. The eclecticism that Friedman's official methodological position counsels – 'don't worry about what theory says, just ask whether it fits market

data' – contradicts the firm and narrow commitment to a specific theory that characterizes mainstream economics, including Friedman's own work. Nevertheless, Friedman's views were warmly embraced. They freed mainstream economists to ignore criticisms of their theories and to abandon all empirical work apart from econometrics.

In the 1970s and 1980s, other streams of contemporary philosophy of science began to influence the methodological reflections of economists. Karl Popper's view that scientific theories must be falsifiable, which had been defended by Hutchison, was taken up and emphasized by Mark Blaug (1992). According to Popper, the hallmark of science is to formulate theories so that they can be exposed to empirical testing and to reject theories that fail tests. Hutchison and Blaug object that mainstream economic theories are not harshly tested and that mainstream economists are unwilling to surrender them when they fail the few tests they are subjected to. In his extensive publications, Lawrence Boland (1982) has also emphasized the relevance of Popper's views to economic methodology, although Boland's interpretation of Popper is rather different than Blaug's or Hutchison's.

Imre Lakatos' methodology of scientific research programs (which represented a wedding of some of Thomas Kuhn's insights to Popper's philosophy of science) was for a while widely discussed by economic methodologists (Latsis, 1976; de Marchi and Blaug, 1991). Although still emphasizing the importance of empirical criticism, Lakatos insisted that theories should not be abandoned until superior alternatives are found, and his emphasis on informal rules or heuristics governing theoretical practice struck a responsive chord.

A spate of prominent writers on methodology: Roger Backhouse (1997), Bruce Caldwell (1991), Hands (1992), and Weintraub (1985) have a more ambivalent relationship to Popper's and Lakatos' philosophy. Caldwell offers searching though charitable criticisms of Popperian views. Backhouse, Hands, and Weintraub were all at one time enthusiastic about Lakatos' views, but, to differing degrees, all have moved away from them now. Hands (2001) has become a proponent of the application to economics of new work on the sociology of science, while Weintraub's work (1991) on the stabilization of concepts in economics has been more influenced by literary theorists.

Alternatives to philosophy of science influenced by work in literary theory and sociology have recently developed a following among economists. In a series of stylistically brilliant works, Deirdre McCloskey has criticized the whole project of exploring a normative methodology for economics and has urged instead that economists attend to their rhetoric – that is, to their ways of persuading one another (1985). This work has been controversial because many of McCloskey's formulations apparently reduce questions about the correctness or incorrectness of economic claims to questions about what most economists accept. Such a view would imply that minority views are always mistaken. McCloskey denies such a radical reading of her position, but it is not clear how to avoid it without incorporating normative methodological commitments into the rhetoric of economics.

Philip Mirowski's work on methodology is heavily historical and less radical epistemologically than McCloskey's. He has explored the influence on economics of the formal analogy between utility theory and physics (1989). More recently, his work has shown increasing concern with sociological influences (Mirowski and Sent, 2002). For other work that tests the boundaries of conventional methodology, see Mäki et al. (1993), Davis et al. (2004), and Kincaid and Ross (2009).

Other recent developments in economics and its methodology have been the articulation of feminist insights into the subject matter and conduct of economics (Barker and Kuiper, 2003; Ferber and Nelson, 2003; Nelson, 1996) and the explosion of work in experimental and behavioral economics and of methodological reflection on that work (Camerer et al., 2005; Caplin and Schotter, 2008; Glimcher et al., 2009; Guala, 2005; Kagel and Roth, 2008; Loewenstein, 2007).

Over the last generation, economic methodology has become a large field. It has its own *Journal of Economic Methodology* an annual series, *Research in the History of Economic Thought and Methodology*, and discussions of economic methodology occupy a large part of the journal, *Economics and Philosophy*. Dozens of monographs on economic methodology have been published. Graduate programs in the field have been established (including a PhD program at Erasmus University in Rotterdam) and a Masters program at the London School of Economics, and there are regular sessions on economic methodology at meetings of economists and philosophers. The field is diverse, and each of its leading figures has a distinctive approach. Although one can roughly identify Popperian–Lakatosian and sociological–literary schools, there is no easy way to categorize contemporary methodology. Here is a sampling of the specific approaches.

Several methodologists have been particularly interested in the role of explicitly causal notions in economics. Nancy Cartwright (2007), who is also a distinguished philosopher of physics, has (unlike most philosophers writing on economics) paid a great deal of attention to econometrics. She has defended the importance of specifically causal considerations in science, and has argued that econometricians have made important contributions to a philosophical understanding of causation. Cartwright holds that economists should be understood as attempting to identify causal *capacities* and that the work of econometricians can be understood as contributing to this task.

Daniel Hausman (1992) defends a modernized variant of Mill's methodology that depicts economics as inexact and permits the plausibility of the basic principles of economics to count in the assessment of economic theories. But he is skeptical of the view that a small set of causal factors do indeed predominate in the domain of economics, and he argues for greater attention to the results of experiments, surveys, and field work (which has to some extent come to pass with the increasing importance of experimentation and studies invoking instrumental variables to simulate experiments (Deaton, 2010)). Hausman has increasingly emphasized the role of causal generalizations in mainstream microeconomic theory and the interrelation between issues in methodology and ethics (2011).

Kevin Hoover (2001a,b) has written on problems concerning macroeconomics and causality. Although he has addressed technical issues in econometrics, his main interest has been in causality. Like Cartwright, Hoover has placed

particular emphasis on what one can learn about causation by studying the work of economists, but he has also attempted to apply insights from philosophical accounts of causation to economics.

Tony Lawson (1997; see also Fullbrook, 2009) finds his inspiration in the 'transcendental realism' of Roy Bhaskar. Lawson sees himself not merely as enunciating a distinctive approach to methodology, but as laying the foundations for a new economics. Crucial to the approach and to his methodological views is a strongly realist ontology which takes the objects of scientific investigation to be causal mechanisms and tendencies that lie as it were 'beneath' the irregularity of phenomena.

Uskali Mäki's work is harder to categorize. Like Lawson, he is concerned with realism, but whereas Lawson attempts to redirect economics and its methodology on the basis of a specific realist ontology, Mäki elucidates versions of realism to which economists have been implicitly committed (2007). Like Cartwright, Hausman, and Hoover, Mäki emphasizes the importance of causal notions in economics. Like Hands and Mirowski, Mäki also applies insights from the sociology of science to the understanding of economics.

Alexander Rosenberg, who is also a distinguished philosopher of biology, began his career arguing that economics fits reasonably well within a standard philosophical model of the natural sciences. But by the early 1990s, he became convinced that the discipline has little empirical content and is best understood either as applied mathematics or as a normative social and political theory (1992).

Geoffrey Hodgson (2004) espouses an evolutionary approach to economics, influenced by institutionalism. John Davis (2011) argues that an appreciation of concepts of identity in economics clarifies economic methodology. Many writers on methodology are committed to the extreme subjectivism of the so-called Austrian School (Boettke, 2010).

Current heated controversy concerning macroeconomic policy has also generated considerable methodological reflection (Backhouse, 2010; Jespersen, 2009), including reflection by the warring parties. A good deal of the latter can be found on the web logs such as those by: Gary Becker (www.becker-posner-blog.com), John Cochrane (johnhcochrane.blogspot.com), Paul Krugman (krugman.blogs.nytimes.com), J. Bradford deLong (delong.typepad.com), Gregory Mankiw (gregmankiw.blogspot.com), John Taylor (johnbtaylorsblog.blogspot.com), and Simon Wren-Lewis (mainlymacro.blogspot.com).

Contemporary economic methodology moves in many directions. Although some of it is directed toward the conduct of economics, one regularly finds a sort of methodological schizophrenia whereby in theory economists cling to positivist or Popperian philosophical views that are radically at odds with their practice, which is roughly Millian. Other work in methodology is oriented toward the philosophy of science, and it appears that the study of economic methodology has made significant contributions to philosophical investigations of causation and explanation. It is controversial whether methodology can or should be a distinct and relatively self-contained field rather than addressing itself toward economists or toward philosophers. For a superb introduction, see Julian Reiss's *The Philosophy of Economics* (2012). Other especially useful sources include Davis et al. (1998), Davis et al. (2004), Kincaid and Ross (2009), and Mäki (2012).

Methodology, Rationality, Preference, and Self-Interest

The study of rationality, preference, and self-interest is a second area of overlap between economics and philosophy. Mainstream economic theory is built around a variant of 'folk psychology.' According to folk psychology, human actions are the consequence of beliefs, desires, and circumstances that determine the feasibility and consequences of actions. The same elements figure in economic theories of choice. Sometimes economists suppose that agents have complete knowledge and thereby avoid having to refer specifically to beliefs, but in other contexts beliefs are explicitly modeled – most typically as subjective probability judgments. In place of desires, which need not be comparative, economists postulate that economic agents possess stable, complete, and transitive comparative evaluations or preferences. Given additional technical conditions, preferences can be represented by a continuous 'utility function' such that $U_P(A) > U_P(B)$ if and only if an economic agent P prefers A to B. 'Utility' is simply a way of indicating how an alternative is ranked by an agent. It is not a substantive object of choice, and it makes no sense to regard an agent as seeking or preferring utility. Mainstream economic models relate choice to preference by maintaining that, subject to the constraints, agents maximize utility – which means merely that, among the alternatives agents believe to be feasible, agents choose whatever they most prefer.

The previous paragraph presents the standard theory of choice as an empirical theory purporting to describe, predict, and explain how agents choose in fact. But the same theory also functions as a theory of rationality. Preferences are rational if they are complete and transitive (and perhaps satisfy other axioms, too). Degrees of belief are rational if they satisfy the axioms of the calculus of probability. Choices are rational if they are determined by preferences. Given this model of rationality, the empirical claims of the standard theory of rational choice can be restated simply as the claim that economic agents are in fact rational.

According to this theory of rationality, it is irrational to choose one feasible option over another when one prefers the second, and according to this theory of choice people will in fact never choose one feasible option over another if they prefer the second (and know what alternatives can be chosen). Since, according to these theories, choice is and ought (rationally) to be determined by the agent's preferences, it is easy mistakenly to infer that these theories maintain that agents are and ought to be self-interested. But whether a choice is self-interested depends on the *content* of the preferences that drive it, not on whether the choice is determined by preferences. People who prefer to sacrifice their interests in order to benefit or harm others are not self-interested, and the fact that their choices result from these preferences does not make their choices self-interested. Unlike the standard theories of choice and rationality, the views that it is rational to pursue one's own interest and that people in fact do so are 'substantive'

theories of choice and rationality. Unlike the standard 'formal' theory, which merely specifies a certain *structure* of choice and preference, self-interest theories specify an objective that agents do or rationally ought to have.

Theories of rationality are normative theories. They *prescribe* what one ought to do – not, to be sure, as a matter of morality, but as a matter of rationality or prudence. (Irrationality is foolish rather than evil.) Incorporating as it does a normative theory, economics is unlike any of the natural sciences. The reason why economics incorporates a theory of rationality is that human actions, unlike the actions of oak trees or potassium, can be criticized or justified as well as explained. This fact has important methodological consequences and indeed (as just argued) establishes one important distinction between the social and the natural sciences.

The first and perhaps most important methodological consequence is that explanations of individual choices cite the *reasons* why the agent acted. Although a number of philosophers in the 1950s, who were influenced by Wittgenstein, argued that explanations that cite an agent's reasons could not be causal explanations, nowadays most philosophers have been persuaded by Donald Davidson's argument (1963) that successful explanations that cite an agent's reasons must also be causal explanations. Merely citing an agent's reasons is not enough if those reasons were not in fact responsible for the action. Davidson argued that the distinction between the reasons that are 'effective' and those that are merely rationalizations is that the first unlike the second are causes of the action.

So the fact that the explanations economists offer of individual choices cite reasons for those choices does not imply that economists are not giving causal explanations, and it diminishes the differences between the structures of the natural and social sciences. But the fact that the causal explanations economists offer also cite reasons remains. It explains why economists must hold that agents are to some extent rational. (If they were not, then they could not have reasons for their actions.) Furthermore, it implies that the actions of economic agents are subject to rational *appraisal*. Were the agent's reasons *good* reasons? Was the action *justified*? Once one begins appraising choices, one is just one step away from ethical questions. Readers should be aware that the standard theories of rationality and choice are controversial.

Welfare, Justice, Equality, and Freedom

Economists have an ambivalent attitude toward ethics. On the one hand, many have insisted that mainstream economics – including welfare economics (!) – is a 'positive' science, whose conclusions are independent of any moral commitments (Gul and Pesandorfer, 2008). On the other hand, economists freely give out normative policy advice. Some of this advice is purely technical, like the advice that a civil engineer might offer on where to locate a bridge, but much of it is not. Most mainstream economists are in fact committed to a specific view of ethics that emphasizes welfare, and they also adopt a distinctive theory of welfare. (When economists emphasize *efficiency* they are usually concerned with efficiency at promoting individual welfare.) These features of normative economics are not entailed by positive economics or the theory of rationality, but they are heavily influenced by them.

There are a variety of different ethical bases for assessing social arrangements. In addition to considering individual welfare, one can ask whether individual rights are protected, whether procedures and distribution are just, whether people are treated equally, and what freedoms and opportunities people enjoy. All of these considerations are important, but in economic policy pronouncements and normative theory (which indeed is called 'welfare economics'), typically only considerations of welfare enter.

The reason for this narrow normative focus is that welfare can be tied tightly to the standard models of rationality and choice. Suppose that individuals are rational and self-interested. As already pointed out, the standard theory of rationality does not imply self-interest, but self-interest is also commonly assumed in most economic models. It follows that individuals will prefer alternative A to B if and only if they believe that A serves their interests better than B. If individuals are good judges of their interests, which is not always the case, but which economists often assume to be the case, it follows that individuals will generally prefer A to B if and only if A in fact serves their interests better. So A is better for an individual than B if and only if the individual prefers A to B – or, in other words, welfare coincides with the satisfaction of preference. This view of welfare has the additional advantage that it prevents concerns about paternalism – to which many economists are opposed – from even arising, since by definition it could never be good for individuals to overrule their preferences.

Most normative economics is thus doubly narrow. Not only is it characterized by an almost exclusive preoccupation with welfare, but it is also committed to measuring welfare by the satisfaction of preferences. This double narrowness is unfortunate. The limitation to considerations of welfare means that in policy discussions economists treat nonwelfarist ethical criteria as additional factors to be left for someone else to worry about. And in practice this often means that other morally relevant factors are ignored altogether. The commitment to measuring welfare by preference satisfaction has the drawback that people do not always prefer what is good for them, because they are often misinformed, not fully rational, and not purely self-interested. There are a variety of good reasons to be hesitant about paternalism, but the view that individuals are always perfect judges of their own good is not one of them. There is some work, particularly by Amartya Sen (1992), that stretches these limits on normative economics, but such work is the exception to the welfarist rule (see also Hausman and McPherson, 2006).

Conclusions

Mainstream economists, whether engaged with positive theory, the theory of rationality, or with normative and policy investigations, are overwhelmingly committed to an exaggerated simplification of a plausible view of individual agents as possessing stable and consistent preference rankings, which, given their beliefs and the constraints, determine their choices.

Market outcomes are the unintended consequences of these choices. Rationality is defined by this structure of choice. Welfare can be inferred from these same preference rankings. Although there are many philosophical issues concerning economics that have no particular connection to this core commitment and that are now commanding increasing attention, most work in the philosophy of economics has attempted to comprehend and evaluate the basic model of choice and its applications in positive economics, normative economics, and the theory of rationality.

See also: Hume, David (1711–76); Institutionalism; Methodological Individualism: Philosophical Aspects; Mill, John Stuart (1806–73); Prisoner's Dilemma, One Shot and Iterated; Rational Choice Theory in Sociology; Realism, in Applied Research; Ricardo, David (1772–1823); Smith, Adam (1723–90); Social Science, The Idea of; Welfare; Wittgenstein, Ludwig (1889–1951).

Bibliography

Backhouse, R.E., 1997. Truth and Progress in Economic Knowledge. Edward Elgar, Lyme, NH.

Backhouse, R.E., 2010. The Puzzle of Modern Economics. Cambridge University Press, Cambridge, MA.

Barker, D., Kuiper, E. (Eds.), 2003. Toward a Feminist Philosophy of Economics. Routledge, London.

Blaug, M., 1992. The Methodology of Economics, second ed. Cambridge University Press, Cambridge, UK.

Boettke, P.J. (Ed.), 2010. Handbook on Contemporary Austrian Economics. Elgar, Cheltenham, UK and Northampton, MA.

Boland, L.A., 1982. The Foundations of Economic Method. Allen and Unwin, London.

Caldwell, B., 1991. Clarifying Popper. Journal of Economic Literature 29, 1–33.

Camerer, C., Loewenstein, G., Prelec, D., 2005. Neuroeconomics: how neuroscience can inform economics. Journal of Economic Literature 43, 9–64.

Caplin, A., Schotter, A. (Eds.), 2008. Handbook of Economic Methodology. Oxford University Press, New York.

Cartwright, N., 2007. Hunting Causes and Using Them: Approaches in Philosophy and Economics. Cambridge University Press, Cambridge, UK.

Davidson, D., 1963. Actions, reasons and causes. Journal of Philosophy 60, 684–700.

Davis, J., 2011. Individuals and Identity in Economics. Cambridge University Press, Cambridge.

Davis, J., Wade Hands, D., Mäki, Uskali (Eds.), 1998. The Handbook of Economic Methodology. Edward Elgar, Cheltenham.

Davis, J., Marciano, A., Runde, J. (Eds.), 2004. The Elgar Companion to Economic and Philosophy. Edward Elgar, Cheltenham.

Deaton, A., 2010. Instruments, randomization, and learning about development. Journal of Economic Literature 48, 424–455.

de Marchi, N., Blaug, M. (Eds.), 1991. Appraising Economic Theories: Studies in the Methodology of Research Programs. E. Elgar, Aldershot, UK.

Ferber, M., Nelson, J. (Eds.), 2003. Feminist Economics Today: Beyond Economic Man. University of Chicago Press, Chicago, IL.

Friedman, M., 1953. The methodology of positive economics. In: Essays in Positive Economics. University of Chicago Press, Chicago, IL, pp. 3–43.

Fullbrook, E. (Ed.), 2009. Ontology and Economics: Tony Lawson and His Critics. Routledge, New York.

Glimcher, P., Camerer, C., Poldrack, R., Fehr, E. (Eds.), 2009. Neuroeconomics: Decision Making and the Brain. Elsevier, Amsterdam.

Guala, F., 2005. The Methodology of Experimental Economics. Cambridge University Press, New York.

Gul, F., Pesandorfer, W., 2008. The case for mindless economics. In: Caplin, A., Schotter, A. (Eds.), The Foundations of Positive and Normative Economics: A Handbook. Oxford University Press, New York, pp. 3–39.

Hands, D.W., 1992. Testing, Rationality, and Progress: Essays on the Popperian Tradition in Economic Methodology. Rowman and Littlefield, Lanham, MD.

Hands, D.W., 2001. Reflection without Rules: Economic Methodology and Contemporary Science Theory. Cambridge University Press, Cambridge, UK.

Hausman, D.M., 1992. The Inexact and Separate Science of Economics. Cambridge University Press, Cambridge, UK.

Hausman, D.M., 2011. Preference, Value, Choice and Welfare. Cambridge University Press, Cambridge, MA.

Hausman, D.M., McPherson, M.S., 2006. Economic Analysis, Moral Philosophy, and Public Policy. Cambridge University Press, Cambridge, UK.

Hodgson, G., 2004. The Evolution of Institutional Economics. Routledge, London.

Hoover, K., 2001a. Causality in Macroeconomics. Cambridge University Press, Cambridge, UK.

Hoover, K., 2001b. The Methodology of Empirical Macroeconomics. Cambridge University Press, Cambridge.

Hutchison, T.W., 1938. The Significance and Basic Postulates of Economic Theory. Repr. A.M. Kelley, 1960, New York.

Jespersen, J., 2009. Macroeconomic Methodology: A Post-keynesian Perspective. Edward Elgar, Cheltenham, UK and Northampton, MA.

Kagel, J.H., Roth, A.E. (Eds.), 2008. The Handbook of Experimental Economics, second ed. Princeton University Press, Princeton, NJ.

Kincaid, H., Ross, D. (Eds.), 2009. Oxford Handbook of Philosophy of Economics. Oxford University Press, Oxford.

Latsis, S.J. (Ed.), 1976. Method and Appraisal in Economics. Cambridge University Press, Cambridge, UK.

Lawson, T., 1997. Economics and Reality. Routledge, London.

Loewenstein, G., 2007. Exotic Preferences: Behavioral Economics and Human Motivation. Oxford University Press, Oxford.

Machlup, F., 1955. The problem of verification in economics. Southern Economic Journal 22, 1–21.

Mäki, U., 2007. Realism and Economic Methodology. Routledge, London.

Mäki, U. (Ed.), 2012. Handbook of the Philosophy of Economics. Elsevier, Amsterdam.

Mäki, U., Gustafsson, B., Knudsen, C. (Eds.), 1993. Rationality, Institutions and Economic Methodology. Routledge, London.

McCloskey, D.N., 1985. The Rhetoric of Economics. University of Wisconsin Press, Madison, WI.

Mill, J.S., 1836/1967. In: On the Definition of Political Economy and the Method of Investigation Proper to It. Repr. Collected Works of John Stuart Mill, 1967, vol. 4. University of Toronto Press, Toronto, ON.

Mirowski, P., 1989. More Heat than Light. Cambridge University Press, Cambridge, UK.

Mirowski, P., Sent, E. (Eds.), 2002. Science Bought and Sold: Essays in the Economics of Science. University of Chicago Press, Chicago, IL.

Nelson, J., 1996. Feminism, Objectivity and Economics. Routledge, London.

Reiss, J., 2012. The Philosophy of Economics. Routledge, London.

Robbins, L., 1935. An Essay on the Nature and Significance of Economic Science, second ed. New York University Press, New York.

Rosenberg, A., 1992. Economics–Mathematical Politics or Science of Diminishing Returns. University of Chicago Press, Chicago, IL.

Samuelson, P.A., 1947. Foundations of Economic Analysis. Harvard University Press, Cambridge, MA.

Sen, A., 1992. Inequality Reexamined. Harvard University Press, Cambridge, MA.

Senior, N.W., 1836. Outline of the Science of Political Economy. Repr. A.M. Kelley, 1965, New York.

Weintraub, E.R., 1985. General Equilibrium Analysis: Studies in Appraisal. Cambridge University Press, Cambridge.

Weintraub, E.R., 1991. Stabilizing Dynamics. Cambridge University Press, Cambridge, UK.

Ecotourism

Carl Cater, Aberystwyth University, Aberystwyth, UK
Erlet Cater, University of Reading, Reading, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E. Cater, volume 6, pp. 4165–4168, © 2001, Elsevier Ltd.

Abstract

Ecotourism is widely recognized and promoted as a sustainable form of nature-based tourism, often with cultural components. Such attention has commanded a great deal of academic work, use by development practitioners, and a United Nations designated 'year of ecotourism' in 2002. However, it is clear that ecotourism has not always been able to deliver on its promises. This is often a result of misinterpretation of what ecotourism means in different contexts and cultures, and its links to broader tourism and development objectives.

Ecotourism emerged as a concept in the early 1990s to rapidly capture the imagination of practitioners and academics alike. The Ecotourism Society (later to become *The International Ecotourism Society*, TIES), based in the United States, was formed in 1990, showcasing at the first International Conference on Ecotourism in the Caribbean in 1991, and the Journal of Ecotourism began publication in 2002. Ecotourism's appeal centered on the fact that it not only embraced the essential principles of sustainability (*see* Sustainability: A Dialog of Values), thus constituting a form of sustainable tourism, but also, in doing so, that it could make a significant contribution to sustainable development (SD). Sustainable tourism is typically intended to result in a symbiotic, win-win, scenario with environmental protection resulting both *from* and *in* improved local livelihoods; continued, and possibly even enhanced, profits for the tourism industry; sustained visitor attraction and satisfaction; and revenue as well as popular support for conservation (embracing the 'use it or lose it' philosophy: simultaneously making conservation pay and paying for conservation). The very incorporation of the prefix 'eco' in the term indicates the close relationship between ecotourism and the environment. Indeed, three out of the five core principles of ecotourism outlined by Page and Dowling (2002), that it is nature-based, ecologically sustainable, and environmentally educative, emphasize this closeness and are regarded as essential attributes of ecotourism. The remaining two fundamental requisites, that it should be locally beneficial and generate tourist satisfaction, are viewed as being desirable for all forms of tourism.

Recognition of the contribution that ecotourism could make toward SD in general and of the relevance of tourism to the poverty agenda in general during the 1990s led to international lending and bilateral development agencies such as the World Bank, USAID, DFID (the United Kingdom), and GTZ (Germany) focusing their attention on ecotourism as one of the few options open to rural communities to develop natural resource uses that are sustainable and, fundamentally, as a means of enhancing local livelihoods. The high ground claimed by ecotourism, in terms of its contribution to development, is that, in principle, it offers much enhanced prospects for local involvement compared with conventional tourism. Major international environmental nongovernmental organizations, such as the World Wildlife Fund, Conservation International, and The Nature Conservancy, often acting as the implementing agencies for the donor bodies, placed ecotourism projects high on their agenda, and this degree of universal attention was ultimately reflected in the UN declaration of 2002 as the International Year of Ecotourism (IYE). This designation flew in the face of the fact that, despite extravagant claims to the contrary concerning its global significance (see, for example, the various estimates cited by TIES, 2000), the World Tourism Organization estimated that ecotourism constituted only 2–4% of global tourism (WTO, 2002).

Defining 'Ecotourism'

It is no coincidence that much of the discussion about ecotourism parallels the weighty debate surrounding the definition of SD (*see* Sustainable Development). This is because the high ground claimed by ecotourism is that it constitutes a form of sustainable tourism, an equally contested concept (see Cater et al., 2014). The term was first given popular airing in the early 1980s by Ceballos-Lascuarian, who suggested that ecotourism "consists in traveling to relatively undisturbed or uncontaminated natural areas with the specific objective of studying, admiring, and enjoying the scenery and its wild plants and animals, as well as any existing cultural manifestations (both past and present) found in these areas" (Weaver, 1998, p. 15). The internationalization of the concept, however, resulted in such a proliferation of definitions that David Fennell was able to conduct a content analysis of 85 of them only just over a decade later (Fennell, 2001). Rather than expand on these multifarious definitions and their evolution, a task that has been amply tackled in the literature (see, for example, Page and Dowling, 2002; Fennell, 2003), it is illuminating to examine the principal features of ecotourism outlined in The Quebec Declaration on Ecotourism (UNEP/WTO, 2002). Emanating from the World Ecotourism Summit, the focal event of IYE, and with a caveat that the 1200 participants were "aware of the limitations of this consultative process to incorporate the input of the large variety of ecotourism

stakeholders, particularly nongovernmental organisations (NGOs) and local and indigenous communities," the declaration recognized:

> that ecotourism embraces the principles of sustainable tourism, concerning the economic, social and environmental impacts of tourism. It also embraces the following specific principles that distinguish it from the wider concept of sustainable tourism:
>
> - Contributes actively to the conservation of natural and cultural heritage,
> - Includes local and indigenous communities in its planning, development and operation, and contributing to their well-being,
> - Interprets the natural and cultural heritage of the destination to visitors,
> - Lends itself better to independent travellers, as well as to organised tours for small size groups.
>
> (UNEP/WTO, 2002).

Deconstructing these requirements throws light on the complexity of issues and interdependencies implied in both the theory and practice of ecotourism.

Ecotourism as a Subset of Sustainable Tourism

By declaring that ecotourism embraces the principles of sustainable tourism, the Quebec Declaration implicitly acknowledges that ecotourism is not the only manifestation of that ilk. Many writers have positioned ecotourism as a subset of the wider concept of sustainable tourism (Cater et al., 2014), situating it in relation to this and other tourism types, for example, alternative, nature, adventure, cultural tourism (Fennell, 2003), and even mass tourism (Weaver, 2001). So, while ecotourism should, by definition, be sustainable tourism, not all sustainable tourism is ecotourism. What, then, are the specific attributes of ecotourism?

Distinguishing Principles of Ecotourism

Active Engagement

The Quebec Declaration's first principle that refines ecotourism from the wider concept of sustainable tourism is that it should contribute *actively* to the conservation of cultural and natural heritage. Orams (2001) classifies ecotourism types along a continuum from 'better' operations that actively contribute to the improvement of the natural environment, through 'neutral' operations that passively seek to minimize their impacts on the natural environment to 'worse' operations that are more exploitative and irresponsible, thus detracting from the quality of the natural environment (*see* Carrying Capacity of the Environment). Hence, the requirement to contribute actively to natural conservation positions true ecotourism at the 'better' end of the spectrum, emphasizing that ecotourism requires not only a behavioral/lifestyle change in the participant but also involves actions that contribute to the well-being of the environment (Weaver, 1998). Weaver (2005) articulates this as *comprehensive* ecotourism which encourages deep learning, a holistic perspective, and an 'enhancement sustainability' approach. This is in contrast to *minimalist* ecotourism offerings which involve "shallow or superficial learning opportunities in conjunction with specific charismatic wildlife and sustainability objectives that are status quo oriented and focused on a particular site" (Weaver, 2008, p. 17). A similar case can be made for cultural heritage.

Inclusivity

The Declaration's second principle, that of the inclusion of local and indigenous communities in its planning, development, and operation while contributing to their well-being, echoes probably the most vaunted principle of ecotourism, that of community involvement. This not only has important implications for development, hence the interest of donor agencies in ecotourism in poverty alleviation strategies and 'pro-poor tourism' (Ashley et al., 2001), but also for conservation, cultural sustainability, distributive justice, and efficient use of human resources. Early approaches toward community involvement in ecotourism were largely tokenist, involving manipulative, passive, consultative, or for-material-reward participation, none of which bring enduring effects on people's lives (Pretty, 1995). The Quebec Declaration, however, stresses the need for much more active participation in decision-making right from the outset so that:

> to achieve equitable social, economic and environmental benefits from ecotourism and other forms of tourism in natural areas, and to minimise or avoid potential negative impacts, participative planning mechanisms are needed that allow local and indigenous communities, in a transparent way, to define and regulate the use of their areas at a local level, *including the right to opt out of tourism development*.
>
> UNEP/WTO, 2002

Hence the last decade has seen the development of many examples of community-based ecotourism (CBET), but the most successful of these are those that have moved slowly to allow communities to articulate their vision of inclusive development effectively (Reimer and Walter, 2013). Much has been made of the potential for ecotourism to empower marginalized elements of communities, particularly women (Scheyvens, 2000). However, in their recent study of CBET development in Cambodia, Reimer and Walter (2013) identified the significant extra workloads placed on female community members in the operation of homestays and guesthouses relative to males engaged in ecotourism.

Educative

The third specific principle of ecotourism outlined at Quebec emphasizes the need to interpret the natural and cultural heritage of a destination for visitors. Indeed, Page and Dowling (2002) suggest that the educative element of ecotourism is a key element distinguishing it from other forms of nature-based tourism, arguing that ecotourism education can influence tourist, community, and industrial behavior, and assist in the longer-term sustainability of tourist activity in natural areas. They point out that interpretation helps to advance the cause of conservation as tourists see the big

picture regarding the environment (*see* Environmental Education and Education for Sustainable Development). Perhaps one of the clearest aspects to the evolution of the ecotourism concept, as identified by Boyd (2012), is a growing ethical reflexivity regarding responsibility and awareness from stakeholders. This may often present challenges in operationalization, but it sits within the urgent need for the tourism industry to engage with the triumvirate of Corporate Social Responsibility, ethics, and sustainability (Fennell and Malloy, 2007).

Small Scale

The Declaration's final specific principle of ecotourism is that it lends itself better to independent travelers, as well as to organized tours for small-size groups. Despite the wording of this codicil, which would allow for larger groups, the obvious preference expressed is for small-scale visitation. The question of scale is one of the most contested characteristics of ecotourism. However, confining ecotourism to small-scale participation not only brings the danger that enacting its principles in such confined contexts is tantamount to preaching to the already converted but also that it is irrational to deny the designation of ecotourism to large-scale nature-based tourism if it adheres to all the requirements of sustainable tourism. As Williams and Shaw (1998, p. 56) state, "While sustainability is often popularly associated with 'smallness'… the link between scale and sustainability has not been empirically (or theoretically) tested." The debate about scale serves to illustrate the heterogeneity of ecotourism. It has become increasingly recognized that a spectrum of participation and involvement can be discerned from hardcore specialist groups, frequently undertaking scientific observation (Wood, 2010), to more casual natural resource-based activities, such as whale-watching, providing they are sustainably managed. Weaver (2001) identifies the latter as a 'soft' (and perhaps minimalist) ecotourism market which may largely consist of 'mass tourists engaged in such activities as part of a broader, multipurpose vacation that often places emphasis in the 3S realm' arguing that most ecotourism activity in the high-profile ecotourism destinations of Costa Rica and Kenya fits into this category. Care should be taken when examining small-scale 'hard' ecotourism operations in a broader sustainability context, for they may have higher global impacts, especially when it comes to climate change. For example, Becken and Simmons (2002) illustrate in the New Zealand context how energy use for ecotourism activities such as 'watching wildlife in a natural setting' are up to 60 times as energy intensive (often requiring the use of four-wheel drives or boats) as traditional tourism activities such as shopping or heritage attractions.

Ecotourism in Practice

There has been much disappointment registered concerning the failure of ecotourism to deliver its promises. For example, an early analysis of 23 Integrated Conservation and Development Projects, most with ecotourism components, found that few of the benefits went to local people or served to enhance protection of adjacent wildlands (Wells and Brandon, 1992).

As with all ideals faced with the harsh realities of the marketplace, it is hardly surprising that there is considerable divergence between theory and practice. This is not the fault of ecotourism per se but can be attributed to two major factors. The first is the failure to recognize the wider context in which ecotourism is set as a process; the second is that its 'sexiness' as a concept has meant that, either through a failure to understand or blatant opportunism, ecotourism has frequently been misinterpreted, misappropriated, and misdirected. Consideration of each of these in turn reveals why an increasing number of antagonists have voiced fundamental concerns about its wholesale adoption across the globe.

The Wider Socioeconomic and Political Context

The wider context in which ecotourism is set as a process and as a principle is vitally important, because that context has a vital role to play in prospects for sustainable outcomes. It is essential to recognize that place characteristics both shape and are shaped by economic, sociocultural, political, ecological, institutional, and technical forces that are exogenous and endogenous as well as dynamic. These forces may have enabling or constraining effects (Hall and Page, 1999) and are dictated by scale and circumstance (Lew and Hall, 1998). There are seemingly limitless combinations and permutations of these forces that either make for or militate against sustainable outcomes because of place and time specificity. However, there are certain general, recurrent themes which give rise for concern. One of the most significant is the failure to recognize the economic milieu in which ecotourism is cast as a process.

There is an inherent danger of regarding ecotourism as an isolated alternative, disassociating it from all other forms of economic activity. There are various scale levels to consider and sectoral conflicts may compromise the success, if not the very existence of ecotourism.

Also, it is essential to consider the global context within which ecotourism is cast as a process. Despite its emphasis on local involvement, it is not exempt from patterns of center–periphery dominance inherent in the international organization of the tourism industry. Ecotourists originate largely from the more developed countries and consequently their tour, travel, and accommodation needs are largely coordinated by firms based in those countries. Such patterns of dominance have led to accusations of ecotourism as 'eco-imperialism' or 'eco-colonialism' (Hall, 2000; Mowforth and Munt, 2001). As will be examined below, this is perhaps the most fundamental criticism of the concept. Further, the commodification of nature and culture in ecotourism products that are higher in status value for richer Western tourists has been critiqued by Wheeler (1994) as 'egotourism,' supporting the narrative cachet of guests more than a meaningful contribution to hosts or the environment.

Paying regard to the overall context means that, as with any other economic activity, we cannot escape from the fact that ecotourism is both shaped by and shapes the markedly inequitable structure, both internationally and intranationally, in which it is cast as a process. There is no substantiation with hard facts to guarantee the claim that ecotourism generally contributes to a more equitable distribution of tourism income and a reduction in poverty. Indeed, it has been argued that

ecotourism may even exacerbate, or even create, divisions due to structural inequalities.

The Misappropriation of Ecotourism

There is increasing concern, particularly from Indigenous Rights Groups and NGOs based in the South, that ecotourism misappropriates nature and culture, not only imposing an essentially Western-centric ethos on societies which have different values, but also simultaneously effectively commodifying the environment and traditional practices.

The accusation is that ecotourism, as constructed by the West, attempts 'to force people everywhere into the same cultural, economic, and political mold' which 'is bound to generate insecurity, resentment, conflict, and even ecological degradation' (Vivanco, 2002, p. 26). This question of ethnocentricity may even create schisms between seemingly consensual viewpoints. Hinch (2001) surmises that, at first sight, the ecocentric world view of 'true' ecotourists should be congruent with that of indigenous cultures who tend to see themselves as being one with the land. However, he highlights potential discord should the ecotourists witness the hunting or trapping of species that may be an essential component of traditional livelihoods.

It is also undeniable that traditional cultures frequently constitute an important component of the composite ecotourism attraction for the Western visitor. The view that outsiders may have of traditional lifestyles may also be viewed as patronizing. In the same way that antidevelopmentalists romanticize the lifestyles of indigenous peoples (Corbridge, 1995) so, too, may Western-constructed ecotourism assume an artificial, 'zooified' lifestyle on local populations, simultaneously assuming that the poor are happy as they are.

Another aspect of misappropriation is that ecotourism appropriates and commodifies both nature and culture. This is problematic for three major reasons. First, indigenous peoples voice their concern that outside interests have no right to expropriate local resources in this way. Second, as Hinch (2001, p. 347) points out "Indigenous people have a much deeper connection with the land than nonindigenous peoples. Because they do not see the land as a possession, they are very wary of treating it as a commodity, even in the purportedly benign context of ecotourism." The complex nature/culture relationship means that cultural erosion may result in the disintegration of social mechanisms governing resource use and allocation. Further, it is rarely indigenous populations who develop and own ecotourism development, as there are many cases of in migration of outsiders who facilitate and benefit from ecotourism development. Witness, for example, the description of marginalization of indigenous inhabitants in Lombok, Indonesia (Schellhorn, 2010), or the appropriation of Dai culture by Hani incomers for tourism purposes in Xishuangbanna, Yunnan, China (Walter, 2004). Absence of cultural identity may lead to an individualistic approach to property rights that eventually leads to resource commercialization. The inevitable commodification of nature and culture with Western-envisaged ecotourism, whereby a financial value is attached to natural and cultural resources, may have a serious knock-on effect. This is that, once a financial value has been attached in this way, should ecotourism present an unattractive investment prospect because of market disincentives, or even fail because of unfair competitive advantage, there is the clear danger that ecotourism entrepreneurs will look toward other more financially advantageous investment options with reduced consideration of environmental and cultural impacts, the expectations that are thus raised might push local populations into other, less sustainable, livelihood options. Furthermore, we are frequently talking of the commodification of natural and cultural resources by outside interests which reinforce Western hegemony.

Ecotourism as a Western-Centric Construct

It is against the backdrop of the global power–knowledge nexus (Blaikie, 2000) that we need to critically examine ecotourism as a Western construct (Cater, 2006). Because it may be seen to both reinforce and be reinforced by Western hegemony, as Duffy (2002, p. 156) suggests, "like other neoliberal policies, ecotourism creates a series of problems." We have considered the influence of supranational institutions above and international agencies above: ecotourism has been mainstreamed by these agencies with both a global reach and an intensifying influence.

It is undeniable that the power relations described currently give rise to a situation where Western ideology is dominant and all powerful. Cohen (2002) echoes this concern with regard to the concept of sustainability in tourism development in general, arguing that it is open to misuses, "not only in the obvious sense of misleading or fraudulent promotion, but in the more insidious sense of its use as an instrument of power in the struggle over rare and valuable environmental or cultural resources."

There is also the consideration that the translation of ecotourism into practice often translates into elitist operations. Whatever the calls for ecotourism operations to be basic and low-key in theory, there is a marked tendency for it to translate into expensive and exclusive in practice. Cohen (2002, pp. 272–273) examines how, as pristine, 'undiscovered' sites become increasingly more difficult to find, their rarity means that they constitute a 'new economic resource' and unspoiled sites harboring particularly valuable natural or cultural attractions tend to become the most expensive ones.

The preexisting political ecology of the destination prejudices sustainable outcomes when ecotourism is translated into practice (Wang et al., forthcoming). Walley (2004, p. 264) in her analysis of the 'social drama' of Mafia Island Marine Park draws attention to the fact that all that we witness may be a repackaging of the status quo such that "the merging of conservation and development agendas, the isolation of ecotourism as a development strategy, and the role of participation and transnational bureaucracies, are not ruptures, but rather build upon and work through existing and historical institutional structures and power relationships."

So, we are faced with a situation where the dominant, Western-centric, environmental imagination has given rise to what Vivanco (2002) calls ecotourism's 'universalistic and self-serving vision.' Another danger of this ethnocentric bias is that it ignores the fact that there are 'multiple natures' constructed variously by different societies. As Macnaghten and Urry (1998, p.95) declare "there is no single 'nature,' only natures. And these natures are not inherent in the physical world but discursively constructed

through economic, political and cultural processes." With the burgeoning domestic and regional tourism in developing and transitional economies, it is increasingly evident that nature tourism is variously constructed by different societies and therefore there are multiple 'nature tourisms.' Buckley et al. (2008), for example, illustrate how the concept of ecotourism is understood in China as Shengtai Luyou, with a particular focus on human and natural 'health' outcomes.

Prospects for Sustainable Ecotourism

It is not surprising that, in common with, and very much allied to, SD, ecotourism is a much contested concept in both theory and practice (Boyd, 2012). The contingencies of time and place mean that it is at the local level that such contestations should occur but serious reservations regarding its provenance and universal relevance have perhaps led to a weakening of its former eminence as alternative, broader terms such as responsible tourism have been introduced in the industry. It is misguided to take issue with the broader philosophy behind ecotourism but problems arise when there is a failure to recognize that it is but a process set in an interdependent world characterized by difference and diversity. It is when the concept is translated into a universal template, largely Western conceived, that it becomes clear that it is not a case of 'one-size-fits-all.' Contexts vary enormously across space and time. It is undeniable, however, that the whole debate surrounding sustainable tourism in general (Cater et al., 2014), and ecotourism in particular, has raised awareness of the pitfalls of ethnocentricity in tourism planning. Wearing and McDonald (2002, pp. 201–202) cite Prakash's call for "a radical re-thinking of forms of knowledge and social identities authored and authorized by colonialism and Western domination" and they argue that, if ecotourism "is to succeed in its goal of cultural and environmental integrity, it requires the development of theory that contains that same integrity." The social appropriation of nature must be viewed in, and from, particular social, economic, and political contexts. Failure to do so will only result in continued disappointment, frustration, and resentment over the manifest shortfall between what ecotourism promises and what it delivers.

See also: Carrying Capacity of the Environment; Environmental Education and Education for Sustainable Development; Sustainability: A Dialog of Values; Sustainable Development.

Bibliography

Ashley, C., Goodwin, H., Roe, D., 2001. Pro-poor Tourism Strategies: Expanding Opportunities for the Poor. Pro-Poor Tourism Briefing No 1. ODI, London.

Becken, S., Simmons, D., 2002. Understanding energy consumption patterns of tourist attractions and activities in New Zealand. Tourism Management 23 (4), 343–354.

Blaikie, P., 2000. Development, post-, anti-, and populist: a critical review. Environment and Planning A 32, 1033–1050.

Boyd, S.W., 2012. Ecotourism: a maturing discourse with some established controversy. In: Moufakkir, O., Burns, P. (Eds.), Controversies in Tourism. CAB International, Oxford.

Buckley, R., Cater, C., Zhong, L., Chen, T., 2008. Shengtai Luyou: a Chinese perspective on ecotourism. Annals of Tourism Research 35 (4), 945–968. Pergamon.

Cater, C., Garrod, B., Low, T. (Eds.), 2014. The Encyclopaedia of Sustainable Tourism. CABI, Oxford.

Cater, E., 2006. Ecotourism as a western construct. Journal of Ecotourism 5 (1&2), 23–39.

Cohen, E., 2002. Authenticity, equity and sustainability in tourism. Journal of Sustainable Tourism 10 (4), 267–276.

Corbridge, S., 1995. Editor's introduction. In: Corbridge, S. (Ed.), Development Studies: A Reader. Edward Arnold, London.

Duffy, R., 2002. A Trip Too Far: Ecotourism, Politics and Exploitation. Routledge.

Fennell, D., 2003. Ecotourism: An Introduction. Routledge, London; New York.

Fennell, D., 2001. A content analysis of ecotourism definitions. Current Issues in Tourism 4 (5), 403–421.

Fennell, D.A., Malloy, D.C., 2007. Codes of Ethics in Tourism: Practice, Theory, Synthesis. Channel View, Clevedon, UK.

Hall, C.M., Page, S.J., 1999. The Geography of Tourism and Recreation. Routledge, London.

Hall, C.M., 2000. Rethinking collaboration and partnership: a public policy perspective. In: Bramwell, B., Lane, B. (Eds.), Tourism Collaboration and Partnerships Politics, Practice and Sustainability. Channel View, Clevedon, UK.

Hinch, T., 2001. Indigenous territories. In: Weaver, D. (Ed.), The Encyclopedia of Ecotourism. CABI, Wallingford, UK.

Lew, A.A., Hall, C.M., 1998. The geography of sustainable tourism: lessons and prospects. In: Hall, C.M., Lew, A.A. (Eds.), Sustainable Tourism: A Geographical Perspective. Longman, Harlow, UK.

Macnaghten, P., Urry, J., 1998. Contested Natures. Sage, London.

Mowforth, M., Munt, I., 2001. Tourism and Sustainability. Routledge, London.

Orams, M.B., 2001. Types of ecotourism. In: Weaver, D. (Ed.), The Encyclopedia of Ecotourism. CABI, Wallingford, UK.

Page, S.J., Dowling, R., 2002. Ecotourism. Pearson, Harlow, UK.

Pretty, J., 1995. The Many Interpretations of Participation, in Focus 16, 4–5.

Reimer, J.K., Walter, P., 2013. How do you know it when you see it? Community-based ecotourism in the Cardamom Mountains of southwestern Cambodia. Tourism Management 34, 122–132.

Schellhorn, M., 2010. Development for whom? Social justice and the business of ecotourism. Journal of Sustainable Tourism 18 (1), 115–135.

Scheyvens, R., 2000. Promoting women's empowerment through involvement in ecotourism: experiences from the third world. Journal of Sustainable Tourism 8 (3), 232–249.

TIES, 2000. Ecotourism Statistical Factsheet. The International Ecotourism Society, Vermont.

UNEP/WTO, 2002. Quebec Declaration on Ecotourism. www.ecotourism2002.org.

Vivanco, L., 2002. Seeing the dangers lurking behind the international year of ecotourism. The Ecologist 32 (2).

Walter, P., 2004. Tourism and forest management among the Hani in Xishuangbanna,China. In: Nathan, D., Kelkar, G., Walter, P. (Eds.), Globalization and Indigenous Peoples in Asia: Changing the Localeglobal Interface. Sage, New Delhi; London; Thousand Oaks, CA, pp. 207–224.

Walley, C.J., 2004. Rough Waters: Nature and Development in an East African Marine Park. Princeton University Press, Princeton; Oxford.

Wang, J., Cater, C., Low, T. Political challenges in community-based ecotourism. Journal of Sustainable Tourism, forthcoming.

Wearing, S., McDonald, M., 2002. The development of community-based tourism: re-thinking the relationship between tour operators and development agents as intermediaries in rural and isolated area communities. Journal of Sustainable Tourism 10 (3), 191–206.

Weaver, D., 1998. Ecotourism in the Less Developed World. CABI, Wallingford, UK.

Weaver, D., 2001. Ecotourism in the context of other tourism types. In: Weaver, D. (Ed.), The Encyclopedia of Ecotourism. CABI, Wallingford, UK.

Weaver, D.B., 2005. Comprehensive and minimalist dimensions of ecotourism. Annals of Tourism Research 32 (2), 439–455.

Weaver, D., 2008. Ecotourism, second ed. Wiley, Brisbane, Australia.

Wells, M., Brandon, K., 1992. People and Parks: Linking Protected Area Management with Local Communities. World Bank, WWF and USAID, Washington, DC.

Wheeller, B., 1994. Ecotourism, sustainable tourism and the environment: a symbiotic, symbolic or shambolic relationship? In: Seaton, A.V., et al. (Eds.), Tourism: State of the Art. Wiley, Chichester, UK, pp. 647–654.

Williams, A.M., Shaw, G., 1998. Tourism and the environment: sustainability and economic restructuring. In: Hall, C.M., Lew, A.A. (Eds.), Sustainable Tourism: A Geographical Perspective. Longman, Harlow, UK.

Wood, P., 2010. A Conceptual Exploration of Marine Research Tourism in Australia: A Study of the Conceptual, Supply, and Demand Nature of Marine Research Tourism. Doctoral Dissertation, James Cook University, Australia.

WTO, 2002. International Year of Ecotourism Launched in New York. http://www.worldtourism.org/newsroom/Releases/more_releases/january2002/launch.

Ecumene

Igor Kopytoff, University of Pennsylvania, Philadelphia, PA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 4168–4170, © 2001, Elsevier Ltd.

Abstract

An ecumene indicates a set of societies that are in persistent and effective interaction. While contacts and diffusion suggest regional continuities and homogeneities embodied in the anthropological notion of culture areas, in fact many historically important regions have been heterogenous, containing societies of uncertain borders and shaped by diasporas of cultural groups, multiculturalism, cultural creolizations resulting from different forms of contact, and dynamic center-periphery relations in which dominant areas were the fountainheads of cultural innovations; such regions may be called ecumenes. Over the past half-millenium, spurred by the worldwide expansion of the West, much of the world has begun to constitute a single global ecumene or world-system. While some historians have treated this development as quite unprecedented in world history, other historians have pointed out that similar systems, on a smaller though continental scale, have shaped world history for several 1000 years.

Ecumene (sometimes oecumene) is a rendition of the Greek word *oikoumene,* which the ancient Greeks applied to the inhabited (i.e., known to them) world. It conveys a sense of communion (as, notably, in religious ecumenism), a community of independent but interacting members, or a set of separate but interrelated societies and cultures. The notion is useful to anthropologists, geographers, historical sociologists, and world historians when they wish to indicate a large region that represents a sphere of persistent and effective interaction among a group of societies that have been influencing one another and have been shaped to some significant degree by a shared history. The concept of ecumene may thus serve to reflect the subdivision of the world into several separate spheres of culture-historical experience. More recently, with the growing interest in 'globalization,' the notion of a 'global ecumene' has emerged; the notion reflects the idea that the entire world is now becoming a single interaction sphere, and that processes (such as integration and constant cultural exchange) that used to characterize the more restricted ecumenes of the past now operate on a worldwide scale. This article will focus on the use of the term in anthropology, and to some extent in world history and historical sociology. The term has been most used by scholars interested in large-scale cultural dynamics over both time and space and who therefore tend to share in a common scholarly discourse.

Early Perspectives

Nineteenth century evolutionary anthropology (primarily Anglo-American) assumed that local societies were expressions of general cultural evolution; in effect, the ethnographic world was not mapped geographically but was viewed in terms of postulated temporal stages. By contrast, German ethnology (which was more firmly grounded in first-hand observation) emphasized the geographical distribution of cultural traits; it regarded cultural contacts and diffusion as the primary mechanisms of cultural dynamics. This fitted well with the commonplace experience that neighboring societies usually tend to be culturally similar, forming what German ethnology called 'cultural provinces' – a concept that stressed cultural continuities within regions.

Early twentieth-century American anthropology, whose worldview was derived from German anthropology by way of its founder Franz Boas, adopted this spatial perspective and saw the ethnographic world as being subdivided into 'culture areas' shaped by regional diffusion. This, however, was mainly true of the 'primitive' ethnographic world, which has usually been peripheral to the world centers of historically significant cultural developments. Even in the 'primitive' world, the diffusionist view raised some problems. The fact that North America, for example, had several culture areas rather than one, demanded some explanation of the cultural discontinuities between the culture areas (a problem that Ruth Benedict proposed to solve by postulating distinctive regional patterns of culture). When one turned to regions of historically complex societies (such as Europe, the Mediterranean world, the Middle East, and South or East Asia) one found long-standing interactions in the midst of marked cultural heterogeneity.

The mounting dissatisfaction with the simplistic view of a world organized into a series of disparate culturally homogenous areas led Alfred Kroeber (1945) to draw attention to the existence on the world map of a central and culturally diverse area of 'civilization' ('a great web of cultural growth' stretching from Gibraltar to China). Within it, there were centers of consistent creativity, generating ideas and techniques that diffused through this area and sometimes to the peripheral areas beyond. Kroeber adopted the term oikoumene for it and his discussion may be seen as an early and prescient invitation to deal with the wider non-'primitive' world from an anthropological perspective. In this view, a culture area became a particular kind of ecumene – local and relatively, but not entirely, homogenous. As Kroeber (1936) had previously shown, such an area exhibited center-periphery relations through which diffusion throughout it took place.

However, Kroeber's (1945) invitation to expand the anthropological vision to worldwide cultural dynamics had scarcely any takers (for an unusual exception, see Hewes, 1970). In the 1930s, anthropology began to abandon its regional ethnological perspectives in favor of a concentration

on cultural dynamics. This turn took two directions. It either studied them within single societies or, when it dealt with contact, it studied it as it occurred between two neighboring societies. In the latter case, anthropology did develop a rich collection of ideas about cultural transmission, but the notion of the ecumene had no place in such a locally focused endeavor and it lay dormant in anthropology for some 30 years. In the meantime, the issues it raised came to the fore among world historians and historical sociologists (see Grew, 1995).

World History

A 'universal history' had been a major project of the eighteenth century Enlightenment, and the idea was revived in the earlier twentieth century in the comparative analysis of civilizations by such historians as Spengler (1926–1928) and Toynbee (1934–1961). Though these comparative studies did not fashion a unified historical narrative on a world scale, they did bring to attention the existence of 'high' periods in the histories of non-Western societies. This undermined the Eurocentrism of world history of the time, in which the first ecumene of any importance seemed to have arisen only with the fifteenth century expansion of the West. Now, attention was being drawn to the dominance of preceding regional ecumenes, such as the Islamic world or China-centered eastern Asia, with the West being frequently peripheral to them. Some historians, notably McNeill, 1963), began to frame world history in terms of a succession of dominant ecumenes, thereby wrenching it out of the complete grip of Western history.

The disappearance of the Western colonial empires after the Second World War reinforced the trend, for it raised questions about the very primacy of the West in world history, the reasons for it, its transience, and the lessons all this held for the immediate future. Here, post-colonial political and sociological concerns merged with historical ones, calling into being a historically oriented political science and political sociology. To be sure, the focus at first remained Eurocentric, as shown in one of the first formulations of these new perspectives – that by Wallerstein's (1974) 'world-system' analysis. Wallerstein took modern Western expansion to be a historically unique occurrence arising from specifically Western historical factors, such as the rise of capitalism. Nevertheless, the formulation was a first step toward an exceptionally fruitful debate and a series of analyses and reanalyses (e.g., Abu-Lughod, 1989; Frank and Gills, 1992; Smith, 1991).

The debate brought to historical consciousness a variety of preceding smaller 'world-systems,' ecu- menes that coexisted with one another in different patterns of integration, and that went back into prehistory. This new world history transcends the traditional national units and focuses on such interregional processes as long-distance trade, economic interdependence, interregional enterprise, large-scale imperial polities, and continent-wide cultural relations between centers and peripheries. It recognizes large areas of economic, political, and cultural interaction, their dynamics, and their periodic demise and reorganization into new ecumenes (Mazlich and Buultjens, 1993). As is natural with any new paradigm, it has led to detailed examinations of local dynamics of economic, political, and cultural exchanges. This development represents, in some sense, a retreat from worldwide perspectives and it often provides correctives to the sweeping assertions of the grand picture. But many of these local studies take the world setting as a necessary context for their analyses and thus address the broader issues.

Modern Anthropology

Similar currents began to sweep anthropology in the later 1960s. The paradigm that had dominated anthropology since the 1930s and that dealt with the systemics of single cultures was lapsing. The demise of colonial rule made an ahistorical approach to ethnography increasingly unrealistic, and this called attention to the larger forces swirling around the anthropologist's supposedly isolated societies. With the introduction of historicity into modern ethnography, it has become obvious that historicity is also necessary in dealing with the ethnographic past.

This new vision of former 'primitives' as being bounced about by the surrounding world was also reinforced by the popular journalistic discussions of 'globalization,' including large-scale international migrations, ethnic diasporas, world labor markets, economic integration, transnational economic and cultural flows, multiculturalism, cultural 'creolization,' the unifying role of the cyber net, and the divisive role of networks. These processes are sometimes seen as dismantling the separate old regional ecumenes that used to divide the world. What looms instead is a single community of interacting societies whose very borders appear increasingly to come into question – in brief, a single 'global ecumene' whose nature is being explored by anthropologists with the use of anthropological concepts (e.g., Foster, 1991; Hannerz, 1992). Thus, Kroeber's ancient oikemene has been transposed out of culture history and into the contemporary era.

In sum, the concept of the ecumene allows anthropologists to deal with interactions among disparate societies. Previously, the only concept that aggregated groups of societies was that of the culture area. But culture areas projected a homogenizing view of cultural interaction and it tied cultural diffusion to territorial continuities. The ecumene, by contrast, allows anthropologists a more realistic view of cultural dynamics – one in which very different societies interact and influence one another selectively and without ceasing to be different. As the mechanisms of such influence become better understood, it becomes clear that they can operate over discontinuous territory and thus encompass, in the modern age, the entire globe. At the same time, it should be noted that much of this new concern in anthropology has more to do with subject matter than theory, for it revives such old-standing anthropological interests as innovation, diffusion, culture contact, and selective acculturation.

See also: Borders, Anthropology of; Center–Periphery Relationships; Civilizations; Cosmopolitanism; Creolization: Sociocultural Aspects; Diffusion: Anthropological Aspects; Diffusion: Geographical Aspects; Global History: Universal and World; Globalization and World Culture; Globalization, Anthropology of; Human Rights, Anthropology of; Hybridity; Multiculturalism, Anthropology of; Multiculturalism; Work, Sociology of.

Bibliography

Abu-Lughod, J., 1989. Before European Hegemony: The World-System, AD 1250–1350. Oxford University Press, New York.

Foster, R.J., 1991. Making national cultures in the global ecumene. Annual Review of Anthropology 20, 235–260.

Frank, A.G., Gills, B.K., 1992. The five thousand year world system: An interdisciplinary introduction. Humboldt Journal of Social Relations 18, 1–79.

Grew, R., 1995. Review of Costello P World historians and their goals: Twentieth century answers to modernism. History and Theory 34, 371–395.

Hannerz, U., 1992. The global ecumene as a network of networks. In: Kuper, A. (Ed.), Conceptualizing Society. Routledge, London, pp. 34–56.

Hewes, G.W., 1970. Mapping the growth of the Old World ecumene. In: VIIth International Congress of Anthropological and Ethnological Sciences VIII. Nauka, Moscow, pp. 373–379.

Kroeber, A.L., 1936. Culture elements distributions: III. Area and climax. University of California Publications in American Archaeology and Ethnology 48, 1–242.

Kroeber, A.L., 1945. The ancient oikoumene as an historic culture aggregate. Journal of the Royal Anthropological Institute 75, 9–20.

Mazlich, B., Buultjens, R. (Eds.), 1993. Conceptualizing Global History. Westview Press, Boulder, CO.

McNeill, W.H., 1963. The Rise of the West: A History of the Human Community. University of Chicago Press, Chicago.

Smith, A.K., 1991. Creating a World Economy: Merchant Capital, Colonialism, and World Trade, 1400–1825. Westview Press, Boulder, CO.

Spengler, O., 1926-1928. The Decline of the West, vol. 20. Knopf, New York.

Toynbee, A., 1934-1961. A Study of History, vol. 12. Oxford University Press, New York.

Wallerstein, I.M., 1974. The Modern World-System, vol. 2. Academic Press, New York.

Edgeworth, Francis Ysidro (1845–1926)

Peter Newman, John Hopkins University, Baltimore, MD, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 4170–4175, © 2001, Elsevier Ltd.

Abstract

It is only recently that Edgeworth has become fully recognized as one of the greatest economists of his time, as well as one of its greatest statisticians. He was the author of over 300 publications in economics and statistics, so any brief biographical essay must perforce concentrate on just a few topics, those selected here being his life, his masterpiece *Mathematical Psychics*, and some of his pioneering statistical work. It was only after Edgeworth's classics degree at Oxford that, living privately in London, he started to study law, mathematics, physics, and the social sciences. His work in statistics began only after the uncomprehending reception of *Mathematical Psychics* in 1882, but it was essentially that work which in 1891 earned him the Drummond professorship at Oxford, a post he retained until 1922. He died at Oxford in 1926.

It is only recently that Edgeworth has become fully recognized as one of the greatest economists of his time, as well as one of its greatest statisticians–a unique combination. In both disciplines his best work was so subtle and profound that it took many years for its meaning to be understood. Edgeworth is thus a striking example of one whose reputation as an original thinker has grown ever since his death, slowly at first and then quite rapidly. After a century, Edgeworth's thought is only beginning to find an echo. *The future will no doubt acknowledge the profundity of his vision more and more* (Allais, 1978, p. 191, referring to *Mathematical Psychics* (1881); original emphasis, as in all quotations here).

Life

Francis Ysidro Edgeworth was born on February 8, 1845 at the family home Edgeworthstown in County Longford, Ireland, the youngest of five sons. His father Francis Beaufort Edgeworth (1809–1846) was one of 22 children of Richard Lovell Edgeworth (1744–1817), an inventor and educationist whose favorite child was the celebrated novelist Maria Edgeworth (1767–1849). Father and daughter were much interested in political economy and numbered among their acquaintance Bentham Malthus, and Dugald Stewart; after her father's death Maria also developed a warm friendship with David Ricardo.

Edgeworth's father, Francis Beaufort Edgeworth, was a favorite of both his elderly father and his half-sister Maria. His mother, 2 years younger than her step-daughter Maria, was Frances Anne Beaufort (1769–1865), of Huguenot stock and elder sister of the admiral who devised the famous Beaufort marine wind scale. Naturally averse to mathematics, Francis left Trinity College, Cambridge without a degree. After a brief courtship at age 22 he married Rosa Florentina Eroles (1815–1864), who was 16, poor, with 'beautiful eyes,' and daughter of an exiled Spanish general. After some years in Italy Francis struggled unsuccessfully as owner of a private school in Eltham near London, and in 1840 returned home to Edge-worthstown to run the estate. He died aged 37 after a long illness, in the worst year of the Irish famine.

Despite his disappointing career Edgeworth's father was not without talent. His writings on poetry and philosophy sufficiently impressed John Stuart Mill to lead him to enquire, in 1840, whether Francis might be available to tutor 'the eldest son ... of a person of very high rank,' adding mysteriously that 'there is, probably, no situation of the kind in England in respect to which more important consequences may depend on its being well filled' (Mineka, 1963, p. 439). Moreover, Francis's friendship was highly valued by the poets Alfred Tennyson and Edward FitzGerald and by the great mathematical physicist Sir William Rowan Hamilton.

Edgeworth himself was actually baptized Ysidro Francis Edgeworth, a form he sometimes used even as an adult. 'Marshall, remembering his mixed parentage, used to say: "Francis is a charming fellow, but you must be careful with Ysidro."' (Keynes, 1926, p. 152). He was educated by tutors at home, committing to permanent memory whole books of Homer, Virgil, Milton, and Pope. A notable linguist, he was well-read not only in classical Greek and Latin but also in French, German, Italian, and Spanish literature. In 1862 he entered Trinity College, Dublin to read classics and in 1867 transferred to Oxford, where in 1869 he gained a First in the honors examination in classics and philosophy.

He spent the next 22 years in London, living on the heights of Hampstead in two bare rooms that he continued as a *pied-à-terre* after he moved to Oxford in 1891. Supplementing his slim private means by occasional earnings as a part-time teacher in further education, he studied law systematically (specializing in commercial law) until he was called to the Bar at the Inner Temple in 1877; he never practiced. He also read widely in mathematics and physics, his interests perhaps influenced by those of his father's friend Hamilton. Certainly, by 1877 he had gained 'a confident and creative mastery of the calculus of variations' (Stigler, 1986, p. 306).

In 1871 he became an 'original' (and lifelong) member of the new Savile Club in central London, whose friendly and distinguished atmosphere apparently suited him well, so much so that during the years 1878–1882 he was quite active in the club's affairs as either a Committee member or an Honorary Secretary or both. Indeed, it was precisely then and not (as is sometimes alleged) as a reclusive Oxford don that he wrote and published *Mathematical Psychics* (1881). Two years later he began the many articles on probability and statistics whose quality carried him upwards through a series of appointments at King's College, London, until in 1891 he was appointed Drummond Professor of Political Economy at Oxford, tenable

at All Souls. He retired from this post in 1922 but the College remained his principal home until he died (from pneumonia) in Oxford on February 13, 1926. As the last surviving male Edgeworth in Europe he inherited the family estate from his brother Antonio in 1911 but did not live there.

Although Edgeworth never married it was not for want of trying. He was smitten by the handsome and imperious Beatrice Potter who found him 'gentle-natured ... excessively polite and diffident ... pathetic ... pedantic' – and boring, even though she glimpsed 'a hidden fire burning within.' Insufficiently discouraged, in June 1889 he published a glowing review of her early work and pursued her to a Co-operative Congress in Ipswich where, in spite of his 'agonized expressions of romantic regard,' she remained adamantly unimpressed (MacKenzie and MacKenzie, 1982, pp. 283–284, 286–287). Three years later she married Sidney Webb.

The words diffident, excessively polite, and pedantic recur often in friends' descriptions of Edgeworth's character, which perhaps accounts for Heilbroner's remark that: 'Neurotically shy, he tended to flee from the pleasures of human company' (1999, p. 173). However, just as frequently those same friends also described Edgeworth as affectionate, humorous, and kind; 'in my house ... few visitors were so beloved as he' (Bonar, 1926, p. 652). Known for his hospitality to foreign visitors, he had 'the widest personal acquaintance in the world with economists of all nations' (Keynes, 1926, p. 151). The young Joseph Schumpeter, visiting Edgeworth at All Souls, was memorably welcomed to breakfast by his host with 'rock pheasant' (presumably, cold roast grouse) and champagne (Samuelson, 1951, p. 53n).

Edgeworth was twice President of Section F of the British Association for the Advancement of Science (1889, 1922), President of the Royal Statistical Society in 1912, one of the original Fellows of the British Academy (1903), and an honorary Doctor of Civil Law from the University of Durham (1889). The most important of his few honors was his sole editorship of the *Economic Journal* (*E. J.*), which lasted from 1890 until J. M. Keynes succeeded him in 1911, whereupon he became Chairman of the Editorial Board. Keynes, busy at Versailles in 1919, co-opted Edgeworth to serve as Joint Editor and he continued active in this capacity 'up to the last day of his life' (Keynes, 1926, p. 140).

No complete bibliography of Edgeworth's prolific writings in economics and statistics has yet been published but an incomplete count yields four very short books, about 300 articles and shorter papers, and more than 200 book reviews. Perhaps reflecting his position as editor of the *E. J.*, after 1890 his major papers in economics tended to be surveys, learned and complex and delighting in brilliant paradoxes, rather than articles on original themes. Many of these, especially the surveys of the theories of distribution, index numbers, international trade, monopoly, railway rates, returns to scale, and taxation, he gathered together and reprinted in Edgeworth (1925). All his many papers in probability and statistics (except (1887) and a few others) were reprinted in 1996.

Mathematical Psychics

Although subtitled 'An Essay on the Application of Mathematics to the Moral Sciences,' *Mathematical Psychics* actually contains *three* analytically distinct essays, all rather jumbled together. The first is on methodology and comprises pp. 1–16 (undated page references here are all to (1881)); Appendix I – 'On Unnumerical Mathematics'; and Appendix II. It develops two separate themes, the first being a visionary stress on the importance of maximum problems for economics, at both the individual and social level. It goes on to argue that in many respects problems in economics are analogous to those of classical mechanics, so that '*Mécanique Sociale* may one day take her place along with *Mécanique Céleste*, throned each upon the double-sided height of one maximum principle' (p. 12). The second theme is less grandiose and develops in detail the importance of 'unnumerical mathematics' for positive economics. The emphasis here on second-order conditions for a maximum sounds very modern: 'the condition that a required curve shall be, or shall not be, convex ... this very relation of concavity, not a whit more indefinite in psychics than in physics... quarried from such data as the law of decreasing utility, of increasing fatigue, of diminished returns to capital and labour' (p. 92).

Edgeworth wrote most of his second essay, on the *Utilitarian Calculus*, when he was primarily a mathematically inclined philosopher. It consists of pp. 56–82 (which essentially reprints (1879)); Appendices III and IV; pp. 116–118 and 122–125 of Appendix VI; and pp. 126–134 of Appendix VII. Earlier, pp. 35–79 of his monograph *New and Old Methods of Ethics* (1877) had applied the calculus of variations and the psychophysics or Fechner and others to Sidgwick's utilitarianism. The resulting 'exact utilitarianism' sought 'the greatest quantity of happiness of sentients, exclusive of number and distribution,' where 'sentients' explicitly refer not just to mankind but to the whole animal creation (1877, p. 35). After a pioneering and detailed variational analysis, employing Lagrange multipliers for both finitely and infinitely many constraints as well as the idea of a continuum of rational agents, he concluded that for modern humankind '(T)here is no indication that ... a law of distribution other than equality is to be wished' (1877, p. 78).

The substance of some comments by Alfred Barratt on this monograph were incorporated in Edgeworth's article in *Mind* (1879), which enhanced his earlier analysis by considering the possible negative utility from work as well as positive utility from consumption; but also diminished it disastrously, by treating different individual capacities for enjoyment not, as before, as essentially random variations but instead as systematic differences in the sentient's position in the 'order of evolution.' This drastically modified his previous conclusions concerning the optimality of equal distribution, since now *'the more capable of pleasure shall have more means'* (1879, p. 398; 1881, p. 64), and led him to some startling arguments in favor of inequality.

When he wrote those indefensible passages Edge-worth was not yet an economist. But two years later, as virtually an apprentice in economics to his friend, neighbor, and mentor William Stanley Jevons, he wrote his masterpiece, the *Economical Calculus*. This third essay, entirely on positive economics, consists of pp. 16–56; Appendix V; pp. 118–122 of Appendix VI; and pp. 134–148 of Appendix VII. Its fundamental theorem asserts that: 'Contract without competition is indeterminate ... Contract with *perfect* competition is perfectly determinate ... Contract with more or less perfect competition is less or more indeterminate (p. 20). The wholly normative

analysis of the earlier *Utilitarian Calculus* depended critically on two doubtful assumptions, the cardinal measurability of individual utilities, and their interpersonal comparability, but this theorem needs neither of them. Indeed, it is richly ironical that Edgeworth, the most uninhibited of utilitarians, should have produced in this third essay the most far-reaching (though implicit) ordinalist analysis of any neoclassical economist.

Few works in economics have been so misunderstood as the *Economical Calculus*. Over the years commentators have discovered there not what it actually contains but only what their own preoccupations have predisposed them to find. For example, those with the *déformation professionnelle* sometimes induced by Walrasian general equilibrium theory have tended to confuse the aims of Edge-worthian recontract with those of Walrasian *tâtonnement*, while those who believe neoclassical economics to be overly tender towards capitalism have found in *Mathematical Psychics* 'a kind of Panglossian Best of All Possible Worlds' (Heilbroner, 1999, p. 173). This is travesty. Edgeworth believed that *imperfect* competition, with its few competitors, occurs frequently in economic life. Hence, by the theorem, indeterminate contract should occur frequently too, with its 'characteristic evil' (p. 29) of '*deadlock*, undecidable opposition of interests, unceasing strife and tumult' (the last phrase a translation of Edgeworth's Greek quotation from Demosthenes). This impasse would then lead to' a general demand for a *principle of arbitration*' (p. 51), a demand which Edgeworth proposes to meet by proffering the following utilitarian rule: the arbitrator should choose that contract which maximizes the total utility of all the parties concerned. 'Thus the *economical* leads up to the *utilitarian calculus*' (p. 56). In the theorem, *articles of contract* are whatever is subject to bargain, such as commodities or property rights, and parties may *recontract* away from any existing contract. *The field of competition* consists of *all* those parties (e.g., sellers and buyers) willing and able to enter into contracts for the articles in question. In these general conditions, which include 'the political struggle for power' (p. 16) as well as commercial exchange, there may well be neither prices nor markets for the articles. Hence, the theory of supply and demand in perfect markets, as already worked out by Jevons, Marshall, and Walras, does not apply to the general case of Edgeworthian contract and it becomes necessary to devise radically new concepts of equilibrium, appropriate to this general setting.

Edgeworth's first such concept is a *settlement*, 'a contract which cannot be varied with the consent of all the parties to it,' and the second *final settlement*, 'a settlement that cannot be varied by recontract within the field of competition' (p. 19). A settlement can, therefore, withstand any variation from it proposed by any party to the original contract, while a final settlement can withstand any variation proposed by any interested agent, whether an original contractor or not. A settlement is, thus, precisely an allocation that yields *maximum d'ophélimité*, a concept first defined 16 years later by Pareto and now known universally as *Pareto optimality*, its Edgeworthian origins long forgotten. Note that while settlement is a concept of positive economics, Pareto optimality belongs to welfare economics.

Edgeworth explicitly provides for coalitions, whose members may recontract between themselves in order to improve their positions under prevailing contracts. As pointed out first by Shubik (1959), from this perspective the set of final settlements is precisely the core of the relevant cooperative game. (While working with John von Neumann in late 1940, Morgenstern 'showed him the *contract curve*' (Rellstab, 1992, p. 83), but their epoch-making book of 1944 has no reference to Edgeworth.)

When there are only two agents they are the only possible contracting parties, so all settlements are also final settlements. Now for Edgeworth, competition necessarily involves at least three agents (see 1894, p. 378), so two agents represent ' (α) Contract without competition' (p. 19). For this case Edgeworth invented and deployed general utility functions and indifference curves to show that the final settlements form a continuous contract curve containing infinitely many points. Contract is, thus, completely indeterminate. (His analysis here remains central to standard micro-economic theory, which uses the 'exchange box' invented in 1906 by Pareto to amend Edgeworth's original exposition; by a delicious irony, economists usually attribute Pareto's device to Edgeworth!)

At the opposite pole to bilateral exchange lies perfect competition, where indefinitely many traders exchange infinitely divisible articles of contract among themselves. Here Edgeworth shows each resulting market equilibrium allocation to be a settlement (the first proof of the *First Fundamental Theorem of Welfare Economics*), as well as a final settlement. He conjectures the true result (proved rigorously 91 years later by Debreu and Scarf) that, in the perfectly competitive replica economy on which his analysis is based, every final settlement is a market equilibrium allocation of some initial endowment. Since the latter allocations are usually unique or at least finite in number, this shows that perfect competition is perfectly determinate.

Edgeworth next displays extraordinary technical virtuosity in dealing with the intermediate case of imperfect competition, where he proves that for a replica economy with $2n$ agents (where $1 < n < \infty$), the set of final settlements always shrinks in size as n increases, so that contract becomes less and less indeterminate as the degree of competition increases. The key property needed for this striking result is the strict convexity of individual indifference curves, which Edgeworth actually proves by using assumptions that imply cardinal measurability of utility, but which he could just as well have assumed directly, without need of cardinality.

This completes the proof of the theorem (for details, see Newman, 1994). As a coda, Edgeworth joyfully and correctly proves that his arbitration rule always produces a settlement. But this has two grave difficulties. First, the rule itself depends on cardinal measurability of utility. Second, an acceptable rule should produce not just a settlement but a final settlement. Otherwise, for some agents the utilitarian allocation might require their allocated utility levels to be below their initial levels, and so its implementation would require compulsory redistribution.

Works in Statistics

Soon after *Mathematical Psychics* appeared, and possibly influenced by Galton, Edgeworth became attracted intellectually by the ' law of error,' i.e., the normal distribution and the central limit theorem. For the next 10 years he devoted his formidable powers and energy to probability, mathematical statistics, and applied statistics, publishing over 30 papers and one brief

monograph (1887). Although subsequently he split his creative efforts more evenly between statistics and economics, in those later years he published a further 75 papers in statistics. R. A. Fisher rated Edgeworth of 'equal eminence' with Karl Pearson in statistics, while a distinguished historian of the subject has judged him to be the leading theorist of mathematical statistics of the latter half of the nineteenth century' (Stigler, 1987, p. 98).

Edgeworth set himself the task of adapting classical probability theory, developed originally by Pierre Simon Laplace and others for the physical sciences, to the problems of statistical inference in the social sciences. For example, regarding the returns of the Bank of England, to estimate the probability that the differences in the averages for different weeks and months are not accidental ... under what circumstances does a difference in figures correspond to a difference in fact' (quoted in Stigler, 1986, p. 308). In 1883 an early result of these efforts was possibly the first derivation of Student's t distribution (reprinted as McCann, 1996, II, No. 1), followed in 1885 by a pioneering version of the analysis variance for a two-way classification (McCann, 1996, III, no. 2).

His methodological views on statistical inference in the social sciences were given in his centrally important paper *Methods of Statistics* of 1885 (McCann, 1996, II, no. 3–this reprint has quite illegible diagrams). Generally, Edgeworth took an undoctrinaire Bayesian approach but was quite willing also to use frequentist methods rather than inverse probability, as when in 1905 he used significance tests to compare means (McCann, 1996, I, no. 26). In (1887) and elsewhere he stressed the importance of maximizing (expected) utility rather than probability in devising appropriate statistical procedures: 'what have we to do with probability or frequency, except as a guide to our actions?' (1887, p. 53). But that monograph has no hint of an axiom system that would unify probability and utility, such as those developed much later by Ramsey, de Finetti, and Savage.

He bequeathed his name to statistical posterity with his development in 1905 of a new series expansion – subsequently the *Edgeworth series* – as a means of adding correction terms to the central limit theorem in order to deal with skew distributions (McCann, 1996, I, no. 26). Less eponymous but more widely known is his term coefficients of correlation and his demonstration in 1892 of their role in expressing the parameters of the multivariate normal distribution (McCann, 1996, III, no. 16 and II, no. 21). A less appreciated novelty, anticipating R. A. Fisher, was his proof in 1908–9 of the asymptotic efficiency of maximum likelihood estimators (McCann, 1996, I, no. 12).

Finally, as so often with Edgeworth, a pioneering contribution in 1888 to inventory theory (McCann, 1996, III, no. 4) was ignored by economists and statisticians alike until very recently, even though it 'filled a gaping hole in orthodox classical monetary theory' (Laidler, 1991, p. 184). This delightfully written essay on banking demonstrated rigorously and quantitatively what had been known to bankers intuitively and qualitatively for centuries, that a bank's prudential reserve level rises proportionately, not with its liabilities, but with their square root.

See also: Data Bases and Statistical Systems: Economics (General); Data Bases and Statistical Systems: Economics (Statistical Systems); Economics, History of; Statistics, History of.

Bibliography

Allais, M., 1978. Theories of general economic equilibrium and maximum efficiency. In: Schwödiauer, G. (Ed.), Equilibrium and Disequilibrium in Economic Theory. Reidel, Dordrecht, The Netherlands.

Bonar, J., 1926. Memories of F. Y. Edgeworth. Economic Journal 36, 647–653.

Edgeworth, F.Y., 1877. New and Old Methods of Ethics. Parker, Oxford, UK.

Edgeworth, F.Y., 1879. The Hedonical Calculus. Mind IV, 394–409.

Edgeworth, F.Y., 1881. Mathematical Psychics. Kegan Paul, London.

Edgeworth, F.Y., 1887. Metretike. Temple, London.

Edgeworth, F.Y., 1894. Competition and regulation. In: Palgrave, R.H.I. (Ed.), 1894. Dictionary of Political Economy, vol. I. Macmillan, London.

Edgeworth, F.Y., 1925. Papers Relating to Political Economy, vol. 3. Macmillan, London.

Heilbroner, R.L., 1999. The Worldly Philosophers, seventh ed. Simon and Schuster, New York.

Keynes, J.M., 1926. Francis Ysidro Edgeworth 1845–1926. Economic Journal 36, 140–153.

Laidler, D., 1991. The Golden Age of the Quantity Theory. Princeton University Press, Princeton, NJ.

McCann Jr., C.R., Edgeworth, F.Y. (Eds.), 1996. Writings in Probability, Statistics and Economics, vol. 3. Edward Elgar, Cheltenham, UK.

MacKenzie, N., Mackenzie, J. (Eds.), 1982. The Diary of Beatrice Webb, vol. 1. Virago, London, 1873–1892.

Mineka, F.E. (Ed.), 1963. The Earlier Letters of John Stuart Mill 1812–1848, 2 vols. University of Toronto Press, Toronto, ON.

Newman, P., 1994. Edgeworth's Economical Calculus. Metro-economica 45, 99–126.

Rellstab, U., 1992. New insights into the collaboration between John von Neumann and Oskar Morgenstern on the 'Theory of Games and Economic Behavior'. In: Weintraub, E.R. (Ed.), Towards a History of Game Theory. Duke University Press, Durham, NC.

Samuelson, P.A., 1951. Schumpeter as a teacher and economic theorist. In: Harris, S.E. (Ed.), Schumpeter: Social Scientist. Harvard University Press, Cambridge, MA.

Shubik, M., 1959. Edgeworth market games. In: Tucker, A.W., Luce, R.D. (Eds.), Contributions to the Theory of Games IV. Princeton University Press, Princeton, NJ.

Stigler, S.M., 1986. The History of Statistics. Harvard University Press, Cambridge, MA.

Stigler, S.M., 1987. Edgeworth as a statistician. In: Eatwell, J., Milgate, M., Newman, P. (Eds.), The New Palgrave: A Dictionary of Economics, vol. 4. Macmillan, London.

Education (Primary and Secondary Schools) and Gender

Maria Tamboukou, School of Social Sciences, University of East London, London, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by B. Schneider, volume 6, pp. 4239–4243, © 2001, Elsevier Ltd.

Abstract

Gender and Education is a relatively recent area of inquiries in the history and sociology of education and it is a contested field of research, action, and policy making. In recent years the field has developed and expanded considerably and it is in continuous dialogue with the wider field of gender studies and feminist theories. Historically the field has addressed questions of equalities of opportunities, recognition, and representation. Contemporary debates in the field include new formations in the social construction of gendered subjectivities, sexualized identities, and the crisis of masculinities; intersections between gender, social class, race, ethnicity, and social exclusion; gender, education, and poverty in the context of globalization.

Gender and Education: Historical Perspectives on Contemporary Concerns

Published in 1792 *The Vindication of the Rights of Woman* was the first feminist philosophical work to reflect on women's condition in society. Inspired by the ideas of the French Revolution, Mary Wollstonecraft counterpoised ideas and discourses that constructed women as vulnerable, weak, and driven by emotions and passions. Women, she suggested can strengthen their mind and bodies through education and active involvement in the public spheres of life. Education in general, and women's education in particular, was central in Wollstonecraft's emancipatory project and it is here that she fiercely criticized Rouseau's ideas on the need for girls to have limited and differently oriented education than men. And yet Rousseau's ideas about the complimentary character of the education of girls set the theoretical framework that dominated the curricula and structures of women's education in modernity, while the feminist project of women's liberation has always revolved around the struggle for access to and equal opportunities in education. Of course many things have changed since the publication of *The Vindication of the Rights of Woman* and yet there are certain themes, concepts, and questions that are still at the heart of contemporary debates in gender and education.

The dilemma between equality and difference infused by social class prejudices and aspirations is a persistent issue in the debates that have historically shaped what has ultimately become the field of gender and education. Not only has the education of girls been perceived, designed, and institutionalized as different from boys, but also differences in girls' education have historically existed across social class divisions, although class has been an important axis of difference in boys' education as well.

Gender and Education: a Contested Field

While Wollstonecraft's work set the philosophical foundations of the theoretical debates within gender and education, it was in the literature of the second wave feminism that the field was constituted as a recognizable entity. Educational feminists in the 1970s and 1980s highlighted the importance of using 'gender' as a useful analytical category in the study of social inequalities in education. As a social construct 'gender' was a term that saved the problems of biological determinism that a focus on sexual difference unavoidably carried. Perceived as an effect of social construction, gender relations thus became not only the object of analysis and research of educational feminists, but also the target of feminist educational practices that were aiming to criticize and change them. This early body of literature was shaped by the idea of education as a site sustaining and reproducing the subordination of women in general and working-class girls in particular. It thus focused on the big themes of equality of opportunities as well as questions of representation and recognition. Among the most specific themes taken up, were the links between capitalism, patriarchy, and policy making in education; open and hidden practices of sexual discrimination and patriarchal structures in schools; as well as cultural representations of gender relations in the language, discourses, and curricula of schooling (see Arnot and Weiner, 1987). As it has been more widely discussed, there were both advantages and limitations in the different versions of the social reproduction theories, particularly in the way they theorized the role of masculinities and femininities not only in shaping class relations and gender inequality in schools, but also in opening up sites of resistance in the dominant culture of schooling (Willis, 1977). It was by attending to the subtle ways that young people negotiated the formation of cultural identities as active agents, albeit within the constraints and limitations of capitalism and patriarchy, that education eventually emerged as a site of both domination and resistance, a battlefield of fierce power/gender relations at play.

The Impact of New Formations in Gender Theories

The rejection of the simplistic division between 'biologically determined sex' and 'socially constructed gender' and the acknowledgment that gender as well as sex are cultural categories, discursively constructed but with real effects on women's lives, (Butler, 1990) have had a significant impact in

the theoretical developments within the field of gender and education. While equality of opportunities and feminist interventions in the curricula of school subjects and the structures of schooling remained important areas of concern, new questions also emerged around the effects of educational discourses and practices in the formation of gendered identities and subject positions. What has emerged as a thorny issue in such debates is the question of difference not in its Platonic configurations as opposed to sameness, but as difference per se. In this light, how can education feminists theorize difference without subsuming it to the norm, heterosexual or otherwise? Moreover how do the politics of difference work within the material and discursive conditions of schooling? It is within the context of taking difference seriously that current debates and concerns in gender and education have emerged and developed in a rich body of literature including among others, a dedicated journal, *Gender and Education*, a range of readers and handbooks (Arnot and Mac an Ghaill, 2006; Skelton et al., 2006) as well as a wide collection of monographs, edited volumes, and journal articles. What has to be noted is that the field of gender and education has specific geographies and histories, which have an impact upon the many theoretical, methodological, and policy-making differences within it.

Current Debates and Issues in Gender and Education

Gender, Subjectivity, and the Constitution of Identity within Neoliberalism

As already noted above, the field of gender and education has always been in dialogue with diverse traditions within social and cultural theory, particularly in relation to questions around identity, subjectivity, and the self. Such questions have taken a new momentum in the context of discourses around individualization and the neoliberal society. Here the notion of 'the reflexive self' has been influential in taking the field beyond the problems of the social reproduction theories and in opening up ground for theorizing agency and resistance within education, but it has inevitably posed its own limitations and constraints (see Francis and Skelton, 2009). In the context of such debates and theoretical concerns feminist scholarship in education has revisited the social construction of masculinities and femininities. Here insights from queer studies (Halberstam, 1998) have been influential in troubling the male/female binarism, bringing forward ideas about female masculinities or masculine femininities. What theorists in the field of gender and education have mostly highlighted here is how boys and girls learn and understand gender within 'communities of practice,' such as schooling, and therefore how teachers can intervene in the processes within which boys and girls learn and practice gender (see Paechter, 2007).

Intersecting Inequalities

'Intersectionality' is a theoretical and methodological framework introduced by Kimberle Crenshaw in 1989 as a way of analyzing socially constructed categories such as race, class, gender, sexual orientation, and disability among others, by viewing their interaction in creating extreme regimes of inequality, disadvantage, and oppression. Initially conceived as a psychosocial plane wherein multiple identities and experiences of subordination and exclusion meet and interact, intersectionality has become today an influential concept in feminist scholarship in gender studies in general and the field of gender and education in particular. Intersectionality is also a very useful framework within which the social construction of cultural identities can be seen and understood in much more comprehensive and efficient ways than in the 1970s and 1980s. Although gender is a crucial social axis of difference, it always intersects with other social categories of difference and is therefore always decentered, but without being subsumed by them. Through the prism of intersectionality, gendered inequalities in education therefore become more complex and less essentialized and the framework offers better ways of both understanding and fighting against inequalities in education (see Phoenix and Patiyama, 2006).

Gender, Racialization, and Social Exclusion

Black feminist though has had a significant impact in shaping problematics and thematic concerns in gender and education and in exposing the Eurocentric elements and Westernized discourses of earlier feminist theorizations in education and beyond (Hooks, 1989; Mirza, 1992, 2009). In the light of the social construction of race that critical race theories have put forward, there is a rich body of literature looking into the intersection of colonial discourses, narratives, and practices in shaping gendered and racialized identities in education in the context of racism, xenophobia, and new flows of migration. School failure and social exclusion have been identified as effects of racialization; what has also been criticized is the construction of stereotypes and prejudices around 'the black girl and boy' (Phoenix, 1987). Here again, intersectionality has revealed the inadequacies of simplistic binarisms, black/white, femininities/masculinities, working class/middle class, to name but the most obvious (see Ahmed, 2009). In this light, a new body of educational research has been formed focusing on questions of social exclusion of black and minority ethnic youth, particularly so in the urban spaces of contemporary global cities. Indeed an attention to both space and time has become an organizing axis in this dynamic body of literature, while the buzzworth of 'the risk society' has been challenged and problematized. What has also been highlighted is the gendered nature of the young people's agentic interventions in the discourses and practices that condition, but not necessarily determine their lives (see among others, Dillabough and Kennedy, 2010; Archer, Hollingworth and Mendick, 2010).

Gender, Class, Choice, and Emotions in the Education Market

New theorizations of the study of educational markets have led to important feminist analyses of marketization, wherein gender, education, and social class have been revisited in the light of intersectionality. In the context of analyzing market forces in their interaction with classed positions rather than classed identities, which have always been problematic vis-à-vis women, an interesting body of research has emerged within which new questions have been raised, linking family life, education, and the market. Beverly Skegg's (1997) theorization of how women live and produce themselves through social and

cultural relations has been influential here and it is exemplary of a field of feminists engaging with the ideas of Pierre Bourdieu (1977, 1984). Diane Reay's (1998) research is an exemplary case of the Bourdieu-feminism encounter in looking into gendered modes of parenting, particularly focusing on the role of mothers in the schooling of their children. The cost of educational success for middle-class and working-class girls is also a very interesting field of inquiries particularly focusing on psychosocial approaches in the field of gender and education (see, Walkerdine et al., 2001). This body of research has further opened up new grounds of inquiries in the field, which look into the importance of emotions and affects in girls' and boys' schooling and beyond (see Davies and Gannon, 2006).

Gender, Education, Poverty, and Globalization

Education feminists have examined gendered processes of globalization and the way they impact educational policies within global and local contexts. Gender, education, and global poverty have been at the focus of much international concern and have generated a cluster of both research and policy-making initiatives. But there is a gap researchers in the field argue between what theoretical frameworks and education policy discourses set out to achieve and what actually happens on the ground. More importantly there are differences between people's needs and issues raised by policies, and finally there is a difficulty in drawing conceptual and causal links between gender, education, and poverty. As a consequence there is a lack of any academic, professional, or policy language to deal with the social contexts of poverty, and gender equality is one of them, but also a tendency to blame individuals and more particularly the poor parent. What emerges as a need is for academics to operationalize important concepts such as intersectionality, so that new theoretical and methodological tools can emerge. What also persists as a need for theorists and practitioners in the field is to further criticize and deconstruct gender stereotypes and prejudices. The pressure of the market and the need to have satisfactory performance is also a stumbling block in the attempt to develop reflexivity and adequate language and policies to address problems emerging from the intersection of gender, education, and poverty. Working with teachers on the school level around issues of gender and poverty has been highlighted as a priority in this area (see Unterhalter and Carpentier, 2010).

The Question of Boys' Education: Critical Responses

It goes without saying that the question of 'boys' underachievement' has been a dominant discourse in gender and education, both in terms of theoretical engagement and in terms of addressing policy-making and pedagogical practices. Here the crisis of masculinity has been the overarching theme, with a significant body of literature revolving around it (see Connell, 2005; Martin, 2001; Mac an Ghaill, 1994). Although the problem of 'boys' underachievement' has become prominent both in policy reports and media coverage since the 1990s, scholars in gender and education have taken a more skeptical and critical approach. They have looked into the history of 'the problem' and have raised questions as to why it has become so prominent in late modernity. The question with boys has also been related to the problem of 'the feminization of schooling' and the importance of male teachers for boys' education (see Epstein et al., 1998). The boys' question has also marginalized some important issues with the education of girls today, including teenage pregnancy, sexualization of girls, and bullying both in friendship groups and through the new social media, or what is now called, cyberbullying. More importantly, the problem with boys is underpinned by a set of discourses grounded in the assumption that gender equality in education has now been achieved and that girls have been excessively empowered at the expense of boys. It is in the context of the current economic crisis and massive unemployment both for men and women that 'blaming girls' has become a populist discourse that has been uncritically accepted in the field of educational policies today on a global level, although with local transmutations and versions (see Ringrose, 2012).

See also: Gender Development: Cultural Differences; Gender History; Gender and Academic Motivation; Gender and Education; Gender and Technology: From Exclusion to Inclusion?; Income Inequality; Inequality Analysis: Overview; Inequality, Social; Intersectionality; Neoliberalism; Social Exclusion, Social Deprivation and Health; Social Inequality and Schooling; Social Inequality in Cultural Consumption Patterns.

Bibliography

Ahmed, S., 2009. Embodying diversity: problems and Paradoxes for black feminists. Race Ethnicity and Education 12, 41–52.

Archer, L., Hollingworth, S., Mendick, H., 2010. Urban Youth and Schooling: The Experiences and Identities of Educationally and 'At Risk' Young People. Open University Press, Berkshire and New York.

Arnot, M., Weiner, G. (Eds.), 1987. Gender and the Politics of Schooling. Hutchinson, London.

Arnot, M., Mac an Ghaill, M. (Eds.), 2006. The Routledge Falmer Reader in Gender and Education. Routledge, London.

Bourdieu, P., 1984. Distinction: A Social Critique of the Judgement of Taste (R. Nice, Trans.). Harvard University Press, Cambridge MA.

Bourdieu, P., 1977. Outline of a Theory of Practice. Cambridge University Press, Cambridge.

Butler, J., 1990. Gender Trouble: Feminism and the Subversion of Identity. Routledge, London.

Connell, R.W., 2005. Masculinities, second ed. Polity Press, Cambridge.

Crensaw, Kimberle, 2012. On Intersectionality: The Seminal Essays. The New Press, London.

Davies, B., Gannon, S. (Eds.), 2006. Doing Collective Biography. Open University Press, Buckingham.

Dillabough, J., Kennedy, H., 2010. Lost Youth in the Global City: Class, Culture and the Urban Imaginary. Routledge, New York and London.

Epstein, D., Elwood, J., Hey, V., Maw, J., 1998. Failing Boys? Issues in Gender and Underachievement. Open University Press, Buckingham.

Francis, B., Skelton, C., 2009. The self-made self: analyzing the potential contribution to the field of gender and education of theories that dissembled selfhood. In: Dillabough, J., McLeod, J., Mills, M. (Eds.), Troubling Gender in Education. Routledge, London, pp. 11–23.

Halberstam, J., 1998. Female Masculinity. Duke University Press, Durham.

Hooks, B., 1989. Talking Back: Thinking Feminist Thinking Black. Sheba Feminist Publishers, London.

Martin, M., 2001. Educating Boys, Learning Gender. Open University Press, Buckingham.

Mac an Ghaill, M., 1994. Making of Men: Masculinities, Sexualities and Schooling. Open University Press, Buckingham.

Mirza, H.S., 1992. Young, Female and Black. Routledge, London.

Mirza, H.S., 2009. Race, Gender and Educational Desire: Why Black Women Succeed and Fail. Routledge, London.
Paechter, C., 2007. Being Boys, Being Girls: Learning Masculinities and Femininities. Open University Press, Buckingham.
Phoenix, A., 1987. Gender under scrutiny. In: Arnot, M., Weiner, G. (Eds.), Gender and the Politics of Schooling. Hutchinson, London, pp. 50–63.
Phoenix, A., Patiyama, P. (Eds.), 2006. Intersectionality. Special Issue of European Journal of Women's Studies.
Reay, D., 1998. Class Work: Mothers' Involvement in Children's Schooling. University College Press, London.
Ringrose, J., 2012. Post-Feminist Education? Girls and the Sexual Politics of Schooling. Routledge, London.
Skeggs, B., 1997. Formations of Class and Gender. Sage, London.
Skelton, C., Francis, B., Smulyan, L. (Eds.), 2006. The Sage Handbook of Gender and Education. Sage, London.
Unterhalter, Elaine, Carpentier, Vincent (Eds.), 2010. Global Inequalities and Higher Education: Whose Interests Are We Serving?. Palgrave, London.
Walkerdine, V., Lucey, H., Melody, J., 2001. Growing up Girl: Psychosocial Explorations of Gender and Class. Palgrave, Basingstoke.
Willis, P., 1977. Learning to Labour. Saxon House, Aldershot.

Education and Economic Growth

Dennis W Cheek, Executive Director, National Creativity Network, USA; and Visiting Professor, IESEG School of Management, France
Joel Santos, Faculté libre de Sciences Économique et de Gestion, Edhec Business School, Roubaix, France
Nicolas Vaillant, Université Catholique de Lille, Lille, France; and Centre de Recherche sur le Capital Humaine, Faculté Libre de Sciences Economiques et de Gestion, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The rate at which economies grow and the contribution that education in both formal and informal settings makes to such growth has been a subject of considerable study and debate going back to the founding of economics as a social science discipline. Numerous models and approaches have been tried and some progress has been made in understanding the complexities of this relationship.

The manner and rate of economic growth among nations and their causes has been a topic for theory building, speculation, and investigation since the inception of the discipline of economics. Early economists such as Adam Smith and John Stuart Mill recognized that there must be a link between education, broadly defined, and economic growth. Explorations of the nature of the linkage did not seriously get under way until the early 1960s with the work of American economists Theodore Schultz (1961) and Gary S. Becker (1993), who developed a theory of human capital (the total set of skills, abilities, ideas, and health residing among individuals within a society, organization, or other unit of analysis that are the results of productive investments in human beings) formation and then sought to analyze the rate of return of investment in education and training. The last half of the twentieth century has abundantly documented that linkages exist but that they are neither inevitable nor straightforward in their relationships and effects. Early naïve theories of human capital that proposed strict linear relationships between formal education and economic growth at both the individual (micro) and societal (macro) levels have now been abandoned in the face of overwhelming evidence that only a few and modest direct relationships can be obtained no matter what definition or method of analysis is used.

Defining Growth and Education

Reasons for this maze of complexity are not hard to find. First we have to consider that when we speak of both 'the economy' and 'economic growth' there are many different ways in which they can be conceptualized, bounded, and measured. Changes in Gross National Product (GNP), Gross Domestic Product (GDP; the total final output of goods and services within a given country's economy that are produced within its territory for a given period of time, without regard to where the owners of such output reside), import/export ratios, unemployment rates, and various indices of economic activity are just a few of the many ways, individually or in combination, by which both an economy and/or economic growth can be operationally defined. To take but one telling example, it was estimated using disparate data sources and expert advice that illicit criminal activities and money laundering account for around 3.6% of global GDP, which equated to $2.1 trillion USD in 2009 according to the United Nations Office on Drugs and Crime (2011). This illegal 'economy' is a significant share of the economies of many nations but is unaccounted for by standard models and studies of both economies within particular countries and the economic growth occurring in those nations.

The means by which rates of growth are calculated can also have a marked impact on whether 'growth' is discerned or not (e.g., year-to-year change, rolling averages, decade-to-decade change). Any growth inevitably occurs within convoluted national, regional, and global contexts that includes substantial changes over time in factors such as population growth, immigration patterns, human capital, technological innovations, structural unemployment, evolving taxation and accounting policies, trade agreements and tariffs, 'hidden' or 'black market' economies, regulatory environments, and political leadership. The degree to which these changes are measured and factored into economic growth models to create 'value-added' or 'adjusted' rates varies considerably among economists and particular schools of thought within which they work.

We also need to consider that as the world economy changes over time, particular sectors of national economies may alter significantly. Increasingly in many nations, applications of knowledge and information drive economic indices of productivity rather than material goods. Changing production patterns that emphasize customization and just-in-time delivery have significantly altered both our conceptions and means of economic growth. Large corporations within the highly interdependent global economy often have worldwide revenues that exceed the GNP of many nations; Walmart, the world's largest corporation in 2010 had global annual revenues exceeding by several billion USD the GDP of Norway – the world's 25th largest economy that same year. These global corporations disperse and shift their corporate assets among many countries for complex reasons rendering quantification of economic growth within particularly smaller countries in any kind of reliable manner increasingly both less precise and more difficult.

Sitting behind every econometric model, such as Nobel Prize winner Robert M. Solow's neoclassical growth model (Solow, 1956), is reality itself – an entity vastly more complex than any of these models which have been created as useful yet greatly simplified abstractions to attempt to elucidate aspects of this much more convoluted, intricate, and highly dynamic environment. At a macro-level, we know that many, if not most, of these models are crude approximations and quite a few are even known explicitly to be wrong in their underlying assumptions (now thoroughly falsified). One of the continuing problems in economic modeling is the frequent failure of the designers and users of such models to explain adequately to others the models' underlying assumptions and their reasonableness, the degree to which models do or do not represent reasonable attempts to mirror critical aspects of the real economies they are intended to describe (at least in part), the structure of the models themselves (i.e., how do they work?), how the chosen model compares in performance to competing models, and what are the important unknowns that the selected model does not address (Schlefer, 2012).

Still further complications arise when a particular model of economic growth (with its attendant assumptions and foci) is analytically and econometrically related to education. 'Education' is a broad concept that includes a variety of subcategories or sectors. The 'formal' system would include primary, secondary, college/university, and certification forms of education (e.g., secondary education equivalency examinations, specialized training programs, and examinations that lead to industry-recognized certifications). The formal system can be further subdivided into public and private spheres. Economists and other researchers generally have scant information about the private sector of education. In some nations or special administrative regions, private schooling accounts for better than 90% of all students enrolled in schools; within Organization for Economic Co-operation and Development (OECD) nations, the percentage of students enrolled in nonpublic schools can range well above 50% (OECD, 2011). Virtually all models of economic growth include substantial data available from the public sphere and usually only very modest amounts from the private sector. In most jurisdictions there are no explicit requirements or enforcement measures for regular data collection among private schools. All too frequently when people study the relationship between economic growth (however conceived, measured, and modeled) and education, this critical matter of missing data is forgotten.

Educational credentials and how they are obtained vary enormously across the globe. Even within OECD nations there is much more variation that is commonly realized along many dimensions. For example, the number of years of formal schooling for which more than 90% of the eligible population are enrolled ranged from 10 to 15 years, with a mean of 13 years (OECD, 2011). The number of specific academic 'tracks' into which students are grouped and the kinds of experiences they receive within their respective tracks varies greatly, even within the same nation – let alone across nations. It is well to remind ourselves when viewing a particular econometric model that frequently these variations are 'lost' in an aggregate construct that lumps them all conveniently in the same category, despite differences that would be both obvious and important to even a casual observer.

Formal education in most nations can also be divided into academic and vocational/technical in terms of its chief focus and once again differing results are seen when considering the impact of these different forms of education on the economy. The percentage of students enrolled in vocational programs within OECD nations, for example, ranges from a low of 9% to a high of 69% (OECD, 2011). This means that when a construct is created that just uses 'education' as measured by a grade level attained or a school leaving certificate (diploma) issued, it misses one of the most important differentiators among various groups of students' learning experiences and their acquisition of various skills and leaves unexplained the contributions that different forms of education within secondary schools, in particular, might make toward both individual financial growth prospects and economic growth (see Malamud and Pop-Eleches, 2010; Vignoles et al., 2011 for two important exceptions).

An 'informal' education sector also exists in nations around the globe. It embraces noncredit forms of education and peer-learning systems (operating without the knowledge of management) that occur on the job and within cultures generally (e.g., evening 'courses' offered to the local community hosted by a local secondary school but having no formal relationship to credentialing or degree-granting institutions or 'clan-type' organizations that provide informal learning exchanges for their constituents). There also exists the arena of job-related training or education which can take the form of job-embedded education (internships, apprenticeships, union-led training, formal company programs) and industry- or state-led training (e.g., vocational/technical education for adults) or retraining schemes, often conducted under the guise of either economic competitiveness or ameliorating the effects of structural unemployment. Informal learning has been estimated to represent in some countries more than 70% of all work-related learning and among students within OECD nations who did not complete some form of upper secondary education, nonformal learning accounts for 50% of their reported learning. The percentage rises (rather than falls) to 65% for those OECD students who obtained an upper secondary certificate – a reminder that even among the most well-educated students, informal learning plays a large role in their acquisition of further knowledge and skills (OECD, 2011). This informal learning presumably acts as a substantial moderator for whatever earlier formal education was undertaken. Once again, many economic models fail to account for the major role such informal learning plays, and its influence on productivity and economic growth more generally is largely underexplored. It does appear that investments in these types of informal education schemes, amounting to billions of USD per year in many nations, are particularly difficult to optimize owing to the dramatic influence of economic policies within competitive markets. Success requires a close match between labor market conditions and the kinds of training received.

Adult basic education (for those who did not graduate from secondary schools) and general efforts to teach literacy must also be considered as important sectors of education that generally operate in isolation from the world of primary and

secondary education. The economic effects wrought by adult education are difficult to measure and vary considerably among nations owing to other factors. Adult basic education, for example, is fairly large even in highly industrialized nations such as the United States and may be a much larger sector than formal education in some lightly industrialized nations. Finally, we must recognize that at the individual level there is much additional nonformal education that occurs as a person accumulates life experiences, reads, interacts with others, and thinks about things as they traverse the life course. The cumulative effects of these highly personal forms of education are unknown in their impacts upon economic growth.

Models of economic growth that seek to document the impact of 'education' usually focus on one or at most two of the above subcategories and frequently employ different data sources and use different definitions and rationales for parsing the data sources. Not surprisingly, they frequently come to dramatically different and frequently contradictory conclusions that arise due to the measures for economic growth adopted, the methodology employed, the scope of the inquiry, the data sources consulted, the ways the data are sorted before analysis commences, the meanings inferred from the analyses, the manner in which unknowns are identified and handled, and a myriad of other factors (Wolf, 2002, 2004).

The Continuum of Development

Further complications arise when moving up and down the continuum of highly industrialized, information-intensive economies and minimally industrialized, information-modest economies (a growing North–South divide on the planet). Economic growth models generally focus on the capacity of a national economy and on measures such as global market share, employment in wage-related activities, income contributions (personal and corporate taxes), redistribution schemes, and other factors that embody within them an implicit belief that these are the proper measures for 'growth' and 'progress' for all cultures, peoples, and nations. Movement 'up' the development continuum to greater industrialization and an intensification of the role of information within the economy are seen as valuable and essential for every nation. This view elevates economic considerations as defined principally by economists who live in particular types of economies into a global standard against which all nations are measured. It ignores noneconomic factors such as population growth (e.g., overly expansive growth and a failure to replace the current population are both generally frowned upon by economists but may be viewed positively by significant segments of certain societies); political considerations (e.g., socialist regimes may view economic goals differently from the majority of economists who work within capitalist frameworks); sociocultural milieus (e.g., individual and group behaviors and class structures); and issues such as fertility, life expectancy, and child mortality.

Economists who have looked at some of these issues in relation to 'developing nations' routinely suspect that current approaches probably underestimate how education improves social welfare and increases economic growth. Additional data sources not generally accounted for in research paradigms include the hidden economies revolving around the informal labor sector, black markets, and other exchanges of goods that are not quantified or quantifiable in many nations of the world. For example, it is estimated that only about half of the GNP in developing nations is comprised of the share of earnings from labor versus a range of 60–70% in highly industrialized nations. Some economists have recommended that dualistic models be created to address adequately the many extremes of the development spectrum. Others have suggested that really four distinct types of economies can now be found among developing nations: (1) newly industrialized nations in Asia, (2) relatively highly educated Latin American countries, (3) large continental economies such as India and China, and (4) marginal rural economies found in Asia, Africa, and Latin America.

A number of disparate findings have emerged when viewing the impact of education on earnings along the development continuum. Higher average levels of formal education are frequently associated with higher earnings in most highly industrialized nations, although notable exceptions are sometimes seen in some nations of Eastern Europe and Russia (Pritchett, 2001; Bassanini and Scarpetta, 2002). Nations at the other end of the spectrum generally demonstrate negative associations with education at university level once educated labor is no longer scarce. This is probably due to the fact that the expansion of the educated labor force has far outstripped the supply of jobs requiring such high levels of formal education and this disequilibrium exacerbates civil unrest. Oversupply has led many economists to recommend that postsecondary education be scaled back in many developing nations and these resources be targeted to primary education interventions. Investments in the education of women (a key source of inequality in many countries) and in general education (as opposed to vocational education) yield returns equal to or better than their alternatives, although once again there are notable exceptions based on differences in economic policies, social structures, and the distribution of education (Park, 2006; Altinok, 2006; Barro, 2001).

Human Capital Theory

Economists classically view labor, land, and capital (human or otherwise) as the three principal inputs into production. Studies of the contribution of education to economic growth have been one line of investigation within the human capital dimension – often held in tension with inputs from the other two sectors in complex multidimensional models to tease out the residuals that can be 'explained' in a statistical sense by education as defined by the researcher. These 'growth accounting' (A technique pioneered by Nobel Prize winning economist Robert Solow (1956) to measure the amount that various factors contribute to the economy. The 'unexplained' portion ('residual') is attributed to the influence of general technological progress within the given society.) models have generally shown a fairly modest effect of education on economic growth ranging from as low as 1% to highs of around 10% since the 1970s. Perhaps not surprisingly, returns from primary-level education (basic literacy and

numeracy) are generally found to be much higher than for other levels and the decline can be steep as one ascends the levels of formal education.

A significant correlation emerges between education and growth. Education can be thought of as the stock of human capital (measured either by the flow of spending on education or by level of educational attainment). If we separate educational attainment on the basis of the educational level attained, this analysis (Aghion and Howitt, 2008) eliminates the 'problem' of the weak relationship traditionally observed between growth in OECD countries and the insignificant impact of education as measured by mean national educational achievement (cf Krueger and Lindahl, 2001). In particular, this decomposition approach not only finds a stronger relationship but also suggests that tertiary education more strongly impacts growth in the economies closest to the world's technology frontier (i.e., the technological global stock of knowledge available in all sectors and all countries). Technological prowess this close to the border reinforces the positive effects of higher education on a country's ability to innovate. Once again it is important to remember that tertiary programs even within OECD nations vary greatly in their composition, student access (with rates ranging from 70% to just 35%), and the actual knowledge and skills on which they focus (OECD, 2011). Whether this particular finding will hold up as more detailed studies are undertaken that look into these variegated realities remains to be seen. It also appears that primary and secondary educational attainment seems rather more to impact on the growth of economies located far from the technological frontier, strengthening the ability to imitate or implement existing technologies (Acemoglu et al., 2006; Vandenbussche et al., 2006).

A working assumption that underlies virtually all human capital approaches going back to Nobel Prize winner Gary Becker and colleagues (Becker et al., 1995) is that people will invest at an individual level up to the point where they believe they are going to realize a full return on their investment. This is often driven by a belief that there are certain fundamental skills that one must acquire in order to compete or advance successfully in the labor market. (It is important to note that costs subsidized by the state or other entities rather than the individual, routinely result in most students choosing to go for a higher level of education or credential than they would elect if they had to bear the full cost themselves.) Schools serve as a powerful socializing influence both in reinforcing the value of education and in shaping the kinds of educational experiences and expectations that people hold (Godoy et al., 2005; Seng, 2011). In many societies, schooling serves as a modernizing function whereas in others (e.g., certain Islamic societies), it may serve to restrict impulses frequently associated with modernization or open capitalist societies (Obst and Kirk, 2010).

Formal education frequently acts as a filter or a screen that determines a threshold level that a job applicant must meet in order to be considered for an open position. One of the many reasons why there are such modest direct relationships between education and economic growth or even education and earning in a narrow sense is that market supply has a bearing on the level of education deemed minimally acceptable by an employer. In times of ample supply of applicants, employers often require a much higher level of education as a key screening tool. Adjusting education supply functions in the light of these changing demands in order to increase economic productivity is both conceptually and politically difficult and is generally discouraged. In part, this is because employers and policy makers recognize that education helps positively shape a variety of other aptitudes, dispositions, and motivation critical to work success. It is important to note also that wages are affected by many other factors extrinsic to employees and employers, such as taxation, price controls, inflation, and state and local policies and regulations. For these reasons, the strong form of the screening hypothesis that argued that higher levels of education signal higher wages over one's entire career is no longer viewed as tenable. A weaker form of the screening hypothesis that views credentials and higher levels of formal education as a filter at the initial hiring stage with accompanying early wage differentials is generally accepted. Competing theories about the operation of labor markets have produced divergent explanations for why educational attainment and levels of lifetime earnings generally hold over time in many industrialized nations. It has been amply demonstrated that holding the same level of education does not serve as a reliable predictor of earnings, undoubtedly due to other factors such as race, gender, ethnicity, ability, and social background. Women, for example, within OECD nations have an average total return (net present value) gross earnings premium that is $100 000 USD less than men – despite the fact that they have the same educational attainment or in many OECD countries, higher educational attainment than their male counterparts (OECD, 2011).

This line of research suggests that it is the highly developed economies relatively closer to the technological frontier which benefit more from higher education because they are better equipped in what is most important for growth when near this boundary (Araya and Peters, 2010; Peters and Bulut, 2011). In particular, these economies are relatively well endowed with highly skilled labor, which contributes relatively more to productivity growth when it is used more for innovation than simply imitation. In this vein, Oded Galor and Omer Moav (2000) show that human capital can accelerate technical progress, since it is more adaptable to new technologies. Similarly, developing economies located farther from the technological border will benefit more from primary and secondary education since they are relatively better endowed in what is important for growth when you are farther away from it, namely the low- and medium-skilled labor that contributes relatively more to productivity growth that is used more for imitation instead of innovation (Vandenbussche et al., 2006). The intuition behind these various proposals is based on the Rybczynski theorem, a by-product of the Heckscher–Ohlin model of international trade. Tadeusz Rybczynski (1955) noted that when a particular factor's presence is introduced into an economy there is a resultant increase in the amount of products that utilize that factor in production with a resultant decrease in the output of products that utilize other factors. These changes within economies stimulate regional trade as the supply and demand for both factors and products change dynamically (Rybczynski, 1955). Eli Heckscher and Nobelist economist Bertil Ohlin (Flam and Flanders, 1991) noticed that countries emphasize within their

economies products for export that are produced from more abundant local inputs and import products for which local production would draw upon scarce resources. Flam and Flanders (1991) postulate that an increase in the supply of an input into a sector where it is used most intensely tends to increase disproportionally the output from this sector and simultaneously prompt a decline in the output from another sector.

This reasoning suggests that for any given country, even more important than its total stock of human capital in explaining growth, is the composition of its human capital and its proximity to the world technology frontier. These factors are most determinative for its growth potential. In terms of growth, the full value of a nation's comparative advantage is based on their endowment of labor supplied by an education system that is aligned with these particular goals. Advanced economies can thus strengthen or improve their comparative advantages by prescribing a policy of investing in higher education and research to increase the potential of the highly skilled workforce they are relatively better situated to deploy (Black and Smith, 2006; but see Moretti, 2004 for a cautionary study). This will have the dual effect of increasing the marginal productivity of their highly skilled workforce and enhance their productive path of specialization toward goods and services with high-added value, i.e., toward a knowledge economy that accentuates innovation and knowledge invention (Krueger and Kumar 2004). Economies with unequal initial endowments in human capital grow at different rates, influenced in part by the degree to which their public policies are committed to public support of the education sector focused on provision of advanced education and training in order to avoid the trap of underdevelopment and promote high growth (Azariadis and Drazen, 1990). It is against this background that emerging economies attempt their technological catch-up (i.e., reducing the distance to the technological frontier). While these strategies are being pursued by emerging economies, the technological frontier continues to move inexorably forward within advanced economies, propelled by the powerful synergistic relationship between education and economic development at the existing technological frontier. There is little evidence that the gap is closing directly but indirectly it may close due to other factors, such as changes in population pyramids which may weaken some advanced economies.

The importance of technological catch-up for developing economies raises the fundamental question of the impact of the cost of education or other barriers on access to education (e.g., in the form of social and ethnic barriers, inadequate public investment in education, and imperfections in the credit market) on economic growth (Berik et al., 2011). People continually assess their prospects for future earnings within the labor market, as compared to the relative cost of their education, and expected future gains in the form of higher wages. The adoption of new technologies requires that we replace unskilled labor with skilled labor, which in turn drives a higher cost for education provision, which results in higher wages for skilled workers and increases in the cost of adoption of new technologies. Barriers to education are, therefore, disproportionally detrimental for developing economies that cannot provide the relatively high wages of skilled workers, thereby increasing the risk of underdevelopment traps (Conrad, 2009; Zeira, 2009). We can thus explain the differences between economies in terms of their technological level of development in part by the differences in the cost of education. When education is largely in the public sector, as is the case in most developing countries, public policy can significantly increase the potential of the positive impact of education on growth by reducing the cost of education. It should also be clear that developing economies must find ways to finance a good public education system, which unfortunately is often not the case given the many challenges and constraints they often face (Kern, 2008; Thirwall, 2011).

Human capital theory has been substantially modified from its early days when considering the impact of education upon the economy. Understanding of human capital has expanded exponentially beyond education per se to embrace issues such as age distributions within given nations, rurality versus urbanity, and threshold factors to economic growth in some nations such as basic health, sanitation, housing, water supply, and nutrition. The modest claims for education's impact on economic growth advanced today are a reflection of both the complexities involved and what Theodore Schultz (1961) has called 'adjusting to economic disequilibrium.' The problems associated with the human capital approach have led to several competing theories about the relationship between education and economic development. One theory takes an institutional approach and argues that supply and demand in education is shaped by larger social behaviors within the society. It is these social conditions, rather than education per se, that is influential on economic growth. A second approach focuses on labor market segmentation and internal labor markets, which in turn influence both education and economic growth.

The existence of institutions and appropriate economic policies promotes regulatory consistency and the necessary stability for the protection of property rights (i.e., protecting innovations against imitation). When these features are coupled with a highly performing education system that increases the pace and efficiency of R & D activities and the continual supply of skilled labor, this combination improves long-term growth. The relationship between institutional foundations and educational growth seems particularly crucial in the case of economies in a technological catch-up situation. While technological catch-up allows poorer economies to grow as rapidly as their richer counterparts, this recovery depends on innovation for which a good legal environment and quality education are both fundamental (Borota, 2012; Glaeser et al., 2004).

A comprehensive causative model for economic growth embracing all sectors of education has proved elusive (Blundell et al., 2005). While the general premise of a link between education and economic growth appears sound, the complexities associated with explicating the exact nature of this premise challenge the best theories and economists.

See also: Education and Employment; Education and Income Distribution; Education, Economics of; Education: Skills Training; Equity and Education; Schooling: Total Impact of.

Bibliography

Acemoglu, D., Aghion, P., Zilibotti, F., 2006. Distance to frontier, selection, and economic growth. Journal of the European Economic Association 4, 37–74.

Aghion, P., Howitt, P., 2008. The Economics of Growth. The MIT Press, Cambridge, MA.

Altinok, N., 2006. Capital humain et croissance: l'apport des enquête internationals sur les acquis des élèves. Economie Publique – Public Economics 18/19 (1/2), 177–209.

Araya, D., Peters, M.A. (Eds.), 2010. Education in the Creative Economy. Peter Lang Publishing, New York.

Azariadis, C., Drazen, A., 1990. Threshold externalities in economic development. Quarterly Journal of Economics 105, 501–526.

Barro, R.J., 2001. Human capital: growth, history and policy – a session to honor Stanley Engerman. AEA Papers and Proceedings 91 (2), 12–17.

Bassanini, A., Scarpetta, S., 2002. Does human capital matter for growth in OECD countries? A pooled mean-group approach. Economics Letters 74, 399–405.

Becker, G., 1993. Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education, third ed. University of Chicago Press, Chicago, IL.

Becker, G.S., Schwartz, P.S., Febrero, R. (Eds.), 1995. The Essence of Becker. Hoover Institution Press, Stanford, CA.

Berik, G., van der Meulen Rodgers, Y., Seguino, S. (Eds.), 2011. Inequality, Development and Growth. Routledge, New York.

Black, D.A., Smith, J.A., 2006. Estimating the returns to college quality with multiple proxies for quality. Journal of Labor Economics 24 (3), 701–728.

Blundell, R., Dearden, L., Sianesi, B., 2005. Measuring the returns to education. In: Machin, S., Vignoles, A. (Eds.), What's the Good of Education? the Economics of Education in the UK. Princeton University Press, Princeton, NJ, pp. 117–145.

Borota, T., 2012. Innovation and imitation in a model of north–south trade. Journal of International Economics 87, 365–376. http://dx.doi.org/10.1016/j.jinteco.2012.01.002.

Conrad, D.A., 2009. Education, Human Capital Accumulation, and Economic Growth: The Caribbean Experience. VDM Verlag, Saarbrüken, Germany.

Flam, H., Flanders, M.J. (Eds.), 1991. Heckscher-Ohlin Trade Theory, Eli F. Heckscher, Bertil Ohlin. The MIT Press, Cambridge, MA.

Galor, O., Moav, O., 2000. Ability biased technological transition, wage inequality within and across groups, and economic growth. Quarterly Journal of Economics 115, 469–497.

Glaeser, E., La Porta, R., Lopez-de-Silanes, F., Shleifer, A., 2004. Do institutions cause growth? Journal of Economic Growth 9, 271–303.

Godoy, R., Karlan, D.S., Rabindran, S., Huanca, T., 2005. Do modern forms of human capital matter in primitive economies? Comparative evidence from Bolivia. Economics of Education Review 24 (1), 45–53.

Kern, A. (Ed.), 2008. Education and Economic Growth: Investment and Distribution of Financial Resources. Linton Atlantic Books, Urbana, IL.

Krueger, D., Kumar, K.B., 2004. US-Europe differences in technology-driven growth: quantifying the role of education. Journal of Monetary Economics 51, 161–190.

Krueger, A., Lindahl, M., 2001. Education for growth: why and for whom ? Journal of Economic Literature 39, 1101–1136.

Malamud, O., Pop-Eleches, C., 2010. General education versus vocational training: evidence from an economy in transition. The Review of Economics and Statistics 92 (1), 43–60.

Moretti, E., 2004. Workers' education, spillovers, and productivity: evidence from plant-level production functions. American Economic Review 94 (3), 656–690.

Obst, D., Kirk, D. (Eds.), 2010. Innovation through Education: Building the Knowledge Economy in the Middle East. Institute of International Education, New York.

OECD, 2011. Education at a Glance 2011: OECD Indicators. OECD Publishing, Paris.

Park, J., 2006. Dispersion of human capital and economic growth. Journal of Macroeconomics 28, 520–529.

Peters, M.A., Bulut, E. (Eds.), 2011. Cognitive Capitalism, Education and Digital Labor. Peter Lang Publishing, New York.

Pritchett, L., 2001. Where has all the education gone? The World Bank Economic Review 15 (3), 367–391.

Rybczynski, T.M., 1955. Factor endowment and relative commodity prices. Economica 22 (88), 336–341.

Schlefer, J., 2012. The Assumptions Economists Make. Harvard University Press, Cambridge, MA.

Schultz, T.W., 1961. Investment in human capital. American Economic Review 51, 1–17.

Seng, Y.C., 2011. Intergenerational Income Mobility in Singapore. Ministry of Finance, Singapore.

Solow, R.M., 1956. A contribution to the theory of economic growth. Quarterly Journal of Economics 70, 65–94.

Thirwall, A.P., 2011. Economics of Development: Theory and Evidence, ninth ed. Palgrave Macmillan, New York.

United National Office on Drugs and Crime, 2011. Estimating Illicit Financial Flows Resulting from Drug Trafficking and Other Transnational Organized Crimes. Research Report, October. UNODC Studies and Threats Analysis Section, Division for Policy Analysis and Public Affairs, Vienna, Austria.

Vandenbussche, J., Aghion, P., Meghir, C., 2006. Growth, distance to frontier and composition of human capital. Journal of Economic Growth 11, 97–127.

Vignoles, A., De Coulon, A., Marcenaro-Gutierrez, O., 2011. The value of basic skills in the British labour market. Oxford Economic Papers 63 (1), 27–48.

Wolf, A., 2002. Does Education Matter? Myths about Education and Economic Growth. Penguin, London.

Wolf, A., 2004. Education and economic performance: simplistic theories and their policy consequences. Oxford Review of Economic Policy 20 (2), 315–333.

Zeira, J., 2009. Why and how education affects economic growth. Review of International Economics 17 (3), 602–614.

Relevent Websites

www.adb.org – Asian Development Bank – leading source for statistics on Asian countries, especially developing economies and educational systems.

www.nber.org – National Bureau for Economic Research – a leading global economics research center using a distributed network of economists producing monthly large numbers of English-language working papers.

www.oecd.org – OECD- authoritative source for all statistics related to its member countries as well as regular studies on economies and education.

www.unstats.un.org – United Nations Statistics Division – a leading source for global statistics on economies and education.

www.worldbank.org – World Bank – studies in both economics and education with a particular focus on developing economies.

Education and Employment

Ulrich Teichler, International Centre for Higher Education Research (INCHER-Kassel), University of Kassel, Kassel, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The relationships between education and employment are determined not only by the function of education to prepare learning for subsequent work tasks and other life spheres, but also by the fact that education selects: in an educational meritocracy, the monetary resources and the social recognitions are largely determined by the individuals' level of educational attainment and their competencies fostered during the course of learning. The levels and the types of education, however, never are closely 'matched' to professional positions and job requirements. Imperfections are unavoidable, because individuals have to be trained to cope with imperfections and to be able to change employment and work themselves proactively. National traditions of education and training persist to a certain extent according to which education might vary from being understood as the general foundation for professional learning on the job toward being understood as specific training for respective jobs. The dramatically rapid expansion of education is interpreted as serving the needs of the knowledge society as an oversupply which might be absorbed and as stimulating changes of the system of employment and work.

The Professional Relevance of Education

Education dissociates the learner physically from work in early years of the life course in order to prepare in a rational way for coping with diversity of work and life tasks. Education has a qualifying function and a status-distributive function for economy and society and is among various factors determining graduates' careers. Imperfect links between education and employment are due to limits in the identification of job requirements, occupational dynamics, indeterminate work tasks of the highly qualified workforce, planning gaps, diverse curricular concepts, and growing importance of lifelong education, whereby diverse value judgments come into play. Research on the relationships between education and employment is strongly influenced by a few economic paradigms, notably the human capital approach. But various other areas of research are relevant as well: labor market research, vocational education, educational sociology, sociology, and history of the professional, sociology of mobility, etc.

Views vary as regards the extent to which expansion of education was stimulated by a growing demand for highly qualified labor or the extent to which educational expansion itself triggered initially an absorption of more highly trained persons on the labor market and subsequently notions of increased demand. Over the decades, unemployment rates were lower in most economically advanced countries the higher the level of educational attainment was. Higher education graduates continued to earn on average of these countries twice as much as compulsory school leavers, and investment in education surpassed the interest rates of business capital in the majority of industrial societies.

Public debates had put a stronger emphasis on the vertical than horizontal 'match' between education and employment. A trend toward higher levels of education than necessarily required by the employment system is obvious, but terms such as 'overeducation' neglect the frequent and innovative use of competencies, even if the position does not necessarily require a graduate. In the 1970s, concern grew about unemployment and precarious employment of youth. The higher the educational attainment, the lower is the risk of becoming unemployed. As youth unemployment is relatively low in countries with apprenticeship training systems, the German approach was most popular in the late 1970s. In the 1980s, the Japanese model of a combination of general vocational education and initial in-company on-the-job training was most widely appreciated, while since the 1990s, a search for new mixes and the appreciation of diverse models prevails (see Cedefop, 2009). Transition from education to employment, a focus of attention since the 1990s, is now viewed as a process taking up to 10 years of search, additional learning, etc. Transition processes and their results are not merely differently determined by varied educational approaches, but also by factors such as credentialism, impact of sociobiographic background, measures of counseling and placement, and by efforts on the part of graduates to improve their employment prospect through diligent search, smart tactics, and the demonstration of talents not rewarded in education. Education became singly the most important determinant of career in the nineteenth and twentieth centuries. Available research, however, suggests that other factors should not be underrated and that education increasingly becomes a precondition of career success in the process of educational expansion, and income differences might be based to a decreasing extent on level of educational attainment once a knowledge society is realized with more than half of the population having attained tertiary education (OECD, 1997).

Schooling is a social mechanism which, as a rule, dissociates the learner physically for a certain life period from the regular world of work and other life spheres. This is undertaken in order to prepare learners in a more rational manner for coping successfully with the diversity of work and other life tasks through explanations, rules, general reasoning strategies, etc. The more efficient the industrial society or postindustrial knowledge society became in producing wealth, the more the education system and generated competencies contributing to

the production of goods and services expanded. Regarding the world of work, education has:

1. A qualifying (in the French and German connotation) function of fostering the cognitive and possibly affective and senso-motoric capabilities which might be useful to cope with job tasks.
2. A status-distributive function: the level of 'educational attainment' determines to a certain extent the monetary resources and the social recognition which will be available to individuals in their subsequent life; for education is a powerful factor – and has increasingly become so over the nineteenth and twentieth centuries – in opening up or closing access to prestigious occupations and providing the means for professional achievement and thus to remuneration and socioeconomic status.

These basic functions are undisputed. However, it is generally assumed that education is bound to be imperfect in preparing for the world of work, partly because rational learning through dissociation from practice has its price in less direct preparation for occupational tasks and because education is expected to serve broader functions. Also, other factors contribute to professional success, e.g., sociobiographic background, genetically determined abilities, specific dynamics of credentials, processes of transfer from education to employment, and finally lifelong learning and personal development. Thus views vary as regards the actual ways education is and ought to be shaped as well as the actual impacts learning has and ought to have on subsequent employment and work. However, recent years were characterized by an increasing public expectation that education serve more visibly the preparation for work, as the growing popularity of the term 'employability' suggests (see Knight and Yorke, 2003).

Imperfections and Limits of the Linkages of Education and Employment

Even if education was expected to be closely geared to the 'requirements' of the employment system and even if there were not any particularistic social factors at play there are obvious imperfections and uncertainties which make close linkages unlikely (see Teichler, 2009). Available research and past debates for instance pointed out the following issues:

1. Occupational dynamics: The employment system is very dynamic in terms of changes of job tasks within given occupations, and most people have to expect occupational mobility in terms of employers or occupations as a consequence of growth and shrinkage of the various occupations. This challenges the view that precareer education and training should prepare only for a very specific bundle of job tasks.
2. Indeterminate work tasks for a highly qualified workforce: The higher the educational level needed for a certain occupational area, the more difficult it is to identify the competencies actually required.
3. Planning gap: There is an unavoidable time gap between the identification of new job assignments and the provision of respective competencies on the part of school leavers and graduates, because years are needed for the revision of curricula and the actual 'production' of new graduates.
4. Generalists' versus specialists' paradigms: Views vary among experts – and among employers as well – regarding the extent to which education should be general or specialized according to areas of knowledge or occupations.
5. Emphasis on vocational education vs. recurrent education: Finally, views differ similarly on the extent to which the growing role of lifelong learning might reduce the need for precareer vocational or professional education and training, on the change of learning abilities over the life course, job requirements in different career stages, and the economic and social conditions for lifelong learning (cf Schuller and Tuijnman, 1999).

In all societies, controversies can be observed as regards the extent to which education is viewed as instrumental in securing the individual's and the society's economic success. There are also strong national differences with respect to four areas of values.

1. An 'educational meritocracy,' i.e., a strong impact of educational attainment on subsequent career, seems to be realized to a different extent in economically advanced countries, most strongly in Japan, whereas lifelong opportunities for counteracting those links are emphasized in the USA.
2. A general or specific approach of education and training: Where a generalist's view prevails, for example in Anglo-Saxon countries and in Japan, specialized training tends to be viewed as 'narrow' both in terms of restraining professional flexibility and personality development. Where a specialist's view dominates, notably in France and to a certain extent in Germany, the acquisition of specific knowledge is considered as a process of exemplary in-depth learning which ensures substantial transfer to other areas of expertise, and is viewed as compatible with a broadly cultured personality.
3. Concepts of employment and work-related identity: In France and Germany, employment and work are generally expected to be a major force for individual identity. This tradition has reinforced a high level of pride in skilled workers in Germany. In contrast, sense of affiliation to one's own employing organization tends to be viewed as a major source of identity and pride in Japan. In the Anglo-Saxon tradition, identity linked to work seems to be a phenomenon confined to high-level occupations, as the distinction underscores between (high-level) 'professions' and other 'vocational' areas.
4. The role policies actually play and should play in shaping the relationships between education and employment: In the Soviet Union and affiliated countries, strong efforts were made from the 1950s through the 1980s to plan education according to the perceived demands of a planned economy. In the USA, faith in the self-regulatory forces is most widespread as far as educational preparation for employment and work in a market-driven economy is concerned. In European countries, different degrees of macroplanning and steering of education are considered essential in order to strike the balance between economic and other social and cultural rationales of education in a predominantly market-driven economy.

The different values are so pertinent for the actual links between education and employment in the respective societies that interpretations of research findings and discussions of their policies are most likely to reinforce national characteristics of education–employment relationships.

Trends of Educational Attainment

Expansion of education seems to have been a perennial trend since World War II. On average, about one-third of the adult population in industrial societies had at least some upper secondary education in 1960; this proportion reached two-thirds in the late 1990s (OECD, 2000) and increased to 73% in 2009 (OECD, 2011).

From the late 1950s through the early 1970s, optimism prevailed that the economy requires a substantial increase of labor with advanced levels of education and that educational expansion is instrumental to economic growth. Comparisons focused on general and full-time schooling, thus neglecting vocational training. Accordingly, full-time enrollment rates of 17 year olds were 86% in the USA in 1970, about three-quarters in Canada and Japan, but only ranged in Western Europe from about 60% to about 20% (OECD, 1977).

The 1973 'oil crisis' was a turning point of the political mood, and concern about unemployment in general began to dominate over problems of employability of youth and potential mismatches between education and employment due to an oversupply of labor with higher or other advanced levels of education and training. At that time, the OECD began to include part-time enrollment and apprenticeship training into their figures of upper secondary enrollment, thereby observing an average increase of the secondary education graduation quota from about 60% in industrial societies around 1970 to more than 80% in recent years.

Similarly, the proportion of the respective age group completing tertiary education (graduation rate) successfully increased from about 10% on average in economically advanced societies around 1970 to about half end of the first decade of the twenty-first century. A comparison of the educational attainment of the adult population (Table 1) shows that the differences by age are small in the USA, where the major expansion of enrollment took place before the 1970s, and high in Spain, where educational expansion is most pronounced in recent decades.

Table 1 Percentage of population from select countries having attained at least upper secondary and tertiary education, by age group 2009

Country	*Upper secondary education*				*Tertiary education*			
Australia	83	73	67	58	45	38	34	29
Austria	88	85	80	72	21	20	18	16
Canada	92	91	87	80	56	56	45	41
Czech Republic	94	94	91	86	20	15	16	11
Finland	90	88	84	67	39	44	37	29
France	84	77	64	55	43	32	22	18
Germany	86	87	86	83	26	28	26	25
Ireland	86	72	65	48	48	39	28	20
Israel	87	84	78	74	43	47	45	45
Italy	70	58	50	37	20	15	12	10
Korea	98	94	73	43	63	44	26	13
Japan	–	–	–	–	56	49	45	27
Netherlands	82	78	71	63	40	34	31	27
Norway	84	83	77	79	47	40	33	27
Poland	93	92	88	73	35	21	13	13
Spain	64	58	46	30	38	34	25	17
Sweden	91	91	85	76	42	35	29	27
Switzerland	90	88	86	83	40	38	33	28
UK	82	76	72	64	45	39	34	29
USA	88	88	89	89	41	43	40	41

Source: OECD, 2011. Education at a Glance 2011: OECD Indicators. OECD, Paris, France.

Income Differences by Education

In all societies, educational attainment and income correlate. In European OECD member states, for example, those with upper secondary education earned on average in the 1990s about 1.3 times, those with nonuniversity tertiary education 1.5 times, and those with university-level education about twice as much as persons without upper secondary education (see OECD, 2011). Altogether, income differences by level of educational attainment have declined somewhat over the decades although not on average since the 1990s. Variations across countries cannot fully be explained by patterns of supply and demand; other social factors come into play which contribute to varying income inequality.

Calculations based on the human capital approach show that individual returns for higher education vary between countries from less than 5% to more than 20%. In some countries they surpass the interest rates of business capital, whereas in others they remain below that level.

The 'Match' between Education and Employment

The extent to which educational qualifications match the employment demands is a frequent topic of debate. One tries to establish – both for recent graduates and the total labor force – whether qualifications correspond to the occupational structure

1. horizontally (links between subjects and occupational categories) or
2. vertically (appropriateness of the level of education to the occupational status).

Altogether, concerns about a horizontal match are more pronounced in countries emphasizing specialization in education and employment than in countries considering education predominantly as a general preparation. They tend to be more pronounced as well regarding higher levels of education than lower ones: A substitution of a car mechanic by a baker seems to be more acceptable than that of an engineer by a chemist. But more leeway for substitution exists than expected. For example, the fact that about half of the Germans having undergone an apprenticeship training are employed

5 years later in occupations not closely linked to the type of training does not tend to be considered as a major wastage, but rather as an indicator that in-depth specialized vocational training fosters substantial potentials for flexibility and transfer of skills.

More attention has been paid to problems linked to possible vertical mismatches. In the 1960s, concern was frequently voiced that countries with a low proportion of the population with advanced education might fall behind the others with respect to economic growth. From the 1970s through the 1990s, in contrast, the controversies on the agenda concerned whether a trend toward 'overeducation' or 'overqualification' could be observed. During the 1990s, a mix of concerns about overeducation in some areas and lack of competencies in other areas, e.g., qualifications required for new technologies, was prevalent in many countries.

'Overeducation' and in some cases 'undereducation' was analyzed with the help of various measures:

1. Educational and occupational statistics seem to suggest that the expansion of advanced levels of education clearly surpassed the demand of the employment system.
2. Employers' expectations similarly tended to support the idea of substantial increase of 'overeducation' in the 1970s and 1980s, though they also point out lack of competencies in some domains.

Surveys of former students also show a higher frequency of education seemingly surpassing the level of job requirements than that of 'undereducation,' but they indicate that many graduates make use on the job of the higher level of competencies.

The progress of research notwithstanding, the issue of 'overeducation' seems to be perennial, because the tensions do not fade away between often steep occupational hierarchies and a relatively open access to the highest levels of education. However, concern tended to be voiced less dramatically in the 1990s than in the 1970s (e.g., Büchel et al., 2003; Teichler, 2009). This might be explained by widespread expectations that a 'knowledge society' is likely to emerge. Also, research has shown that most persons seemingly overqualified acquired a position only slightly lower than they had strived for.

Education, Unemployment, and 'Employability'

Youth unemployment and various modes of precarious employment became a major concern in most industrialized societies since the mid-1970s and in most developing countries for long periods. The unemployment quota among youth is substantially higher than that of the total labor force in most countries. Also, in most industrial societies, the higher the unemployment quotas are, the lower is the level of educational attainment. OECD calculations made in the late 1990s suggest that Europeans trained below upper secondary education have to expect 4 years of unemployment over their life course, those trained on upper secondary education somewhat more than 2 years, and those trained on tertiary education level somewhat more than 1 year.

This led to growing attention to the education of persons not achieving advanced levels of education, called for example, 'the forgotten majority' or the 'forgotten half.' A discussion was triggered as to whether an expansionist education policy could strengthen the employability potentials of young people and whether certain types of education and training were more successful than others in preparing youth for the world of work.

Various models were suggested to classify modes of vocational education and training on secondary education level (cf Tessaring, 1998). During the late 1970s, the apprenticeship training model (Lutz, 1994) became popular in international debates, because the youth unemployment rate in Germany and neighboring countries was not substantially higher than overall unemployment rates. During the 1980s, the pendulum of debate swung in favor of a general approach of vocational education combined with initial and further on-the-job training within the employment system, as customary in Japan.

Efforts were also made to increase educational and training opportunities in countries previously characterized by a low enrollment rate at the upper secondary education level. Recently new mixes of education and work experience have emerged without any single model being regarded as clearly superior (see Cedefop, 2009). Moreover, concepts of defining results of education and training in terms of 'competencies' gradually softened the dichotomic notions of education and training (see Blömeke et al., 2012).

Transition from Education to Employment

In the 1990s, increased attention was paid to the processes of transition from education to employment. An OECD study defined transition as a period of up to 10 years when youth might be searching for a satisfying and well-paid job, orientation and efforts to supplement competencies might interact, and employment opportunities might differ from those for the older workforce (Werquin et al., 1997).

Part of the respective research focuses on the job start as a major indicator of the relationships between education and employment. Other studies analyze the speed of transition and the frequency of unemployment and precarious employment over various years. A further type of analysis tries to examine whether the process of transition is facilitated by certain types of prior education and training, e.g., a close 'coupling' of education and training to respective occupations, notably through the apprenticeship systems (cf Shavit and Müller, 1998), whereas others consider 'market' regulation, as observed notably in the USA, as most promising because it facilitates adjustments at any time in the career.

Various studies address factors other than educational achievement that might explain transition to employment and further career. 'Credentials,' for example, both certify and symbolically overemphasize educational achievements. Also, sociobiographic background comes into play not merely in terms of particularistic advantages in the transition processes, but also in terms of social skills not taught through formal educational processes. Further, the terms of 'key qualifications' point out the importance of competencies in part fostered by education and training. Moreover, transition from education to employment also can be viewed as a moment under reduced meritocratic rule when young people can seek for chances through diligent search, smart tactics, and the demonstration of

talents so far not rewarded in education. Finally, some analyses indicate moderate impacts of targeted measures of improving transition to employment, notably short training programs for 'at-risk youth' and occupational guidance and counseling as well as placement arrangements by public employment agencies, private recruitment, and placement firms, and by placement support on the part of educational institutions.

Concluding Observations

Education became increasingly an important determinant of employment and career in the nineteenth and twentieth centuries. By and large, long periods of vocational learning are rewarded by a higher status and more interesting and independent work. However, this is affected by labor market situations of demand and supply. Also, education becomes, in the process of expansion, more and more a prerequisite for career success, whereby other factors might gain weight with respect to the details of selection and allocation. Finally, the growing dynamics within the occupation system make the influence of vocational education less visible when lifelong education grows.

International comparison has pointed out a striking variety of modes of education and training and of links between education and employment. The debates continue on the virtue of general vs. specific education, on hopes set in educational expansion vs. a closer match between education and the demands of the economy, on the advantages of a close coupling between training and work vs. open market regulations, etc. But international comparison also questions the beliefs widespread in the 1960s that one could identify the most successful, modern way of shaping the education system in a most promising way for serving economy, culture, and personality development. This did not necessarily lead to a dominance of relativistic approaches, but rather often stimulated the search for new mixes, e.g., in combining general education with work experience, opening up avenues for tertiary education while intensifying measures for educationally disadvantaged youth, improving education and concurrently measures that address the transition processes directly as well as 'employability' in other respects.

Even in a 'knowledge society,' one does not expect that a more or less pure educational meritocracy will emerge, because other factors might increasingly gain momentum. But education seems to remain the single most important determinator of career and other life chances.

See also: Education and Economic Growth; Education and Income Distribution; Education, Economics of; Education: Skills Training; Educational Productivity; Equity and Education.

Bibliography

Blömeke, S., Zlatkin-Tritschanskaia, O., Kuhn, C., Fege, J. (Eds.), 2012. Modeling and Measuring Competencies in Higher Education. Sense Publishers, Rotterdam, the Netherlands; Taipei, Taiwan.

Büchel, F., de Griep, A., Martens, A. (Eds.), 2003. Overeducation in Europe: Current Issues in Theory and Policy. Edward Elgar, Cheltenham, MA, USA.

Cedefop, 2009. Modernising Vocational Education and Training. Cedefop, Tessaloniki, Greece.

Knight, P., Yorke, M., 2003. Learning, Curriculum, and Employability in Higher Education. RoutledgeFalmer, London, UK.

Lutz, B., 1994. The difficult recovery of 'professionalism'. In: Apprenticeship: Which Way Forward? OECD, CERI, Paris, France, pp. 19–28.

Nijhof, W.J., Streumer, J.N. (Eds.), 1998. Key Qualifications in Work and Education. Kluwer, Dordrecht, The Netherlands.

OECD, 1977. Entry of Young People into Working Life. OECD, Paris, France.

OECD, 2000. Education at a Glance: 2000 OECD Indicators. OECD, Paris.

OECD, 1997. Redefining Tertiary Education. OECD, Paris, France.

OECD, 2011. Education at a Glance 2011: OECD Indicators. OECD, Paris, France.

Psacharopoulos, G. (Ed.), 1987. Economics of Education: Research and Studies, first ed. Pergamon, Oxford, UK.

Shavit, Y., Müller, W. (Eds.), 1998. From School to Work: A Comparative Study of Educational Qualifications and Occupational Destinations. Clarendon Press, Oxford, UK.

Schuller, T., Tuijnman, A. (Eds.), 1999. Lifelong Learning: Policy and Research. Portland Press, London, UK.

Teichler, U., 2009. Higher Education and the World of Work: Conceptual Frameworks, Comparative Perspectives, Empirical Findings. Sense Publishers, Rotterdam, The Netherlands; Taipei, Taiwan.

Tessaring, M., 1998. Training for a Changing Europe: A Report on the Current Vocational and Training Research in Europe. Cedefop, Thessaloniki, Greece.

Werquin, R., Breen, R., Plannas, J. (Eds.), 1997. Youth Transitions in Europe: Theories and Evidence. CEREQ, Marseille, France.

Education and Income Distribution

Gary N Marks, Melbourne Institute of Applied Economic and Social Research, The University of Melbourne, Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by T. Tachibanaki, volume 6, pp. 4186–4189, © 2001, Elsevier Ltd.

Abstract

This article discusses the relationships between income and education. The first section focuses on the effects of income on various educational outcomes and the second half discusses the relationship between education and earnings. Although the gross bivariate relationship between family income and education is moderately strong, the causal impact of income is very much weaker after taking into account other aspects of socioeconomic background and cognitive ability. Education is almost invariably the strongest influence on earnings and in the United States the returns to education, especially postsecondary education. Although other countries in the 1980s and 1990s also experienced increasing returns to education, more recent studies conclude little change or even decreasing returns to education in countries other than the United States.

Effects of Income on Education

According to the economic resources approach, economic inequality is the prime generator of differences in student performance. The essence of this approach is that parents either can or cannot afford to 'buy' educational success for their children. They can do this by sending their child to a high-fee private school or buying (or possibly renting) a house or apartment in the catchment area of a high-performing public school which tend to be located in more expensive neighborhoods. In some countries, parents commonly employ personal tutors to improve their child's performance. At the other end of the economic spectrum, poorer families cannot buy the essentials: textbooks, reading books, and other educational resources; and possibly their crowded living areas do not provide enough space for focused productive study. In reference to college education in the United States which can be very costly, the economic argument focuses on credit constraints; poorer families are unable to borrow the money necessary for a good education for their child (Becker, 1975; Carneiro and Heckman, 2002). Such credit constraint arguments do not apply as well, or not at all, in countries where university education is largely free or has relatively low upfront costs.

Overall Relationship

Family income is significantly related to student achievement and other educational outcomes. Sirin's (2005) meta-analysis of mainly small studies conducted in the United States between 1982 and 2000 found that the average effect size (adjusted correlation coefficient) for the bivariate relationship between family income and student achievement was around 0.29 comparable to effect sizes of 0.30 and 0.28 for parental education and father's occupational status. This contrasts with White's (1982, p. 470) earlier meta-analysis which found that family income ($r = 0.32$) showed stronger correlations than parental occupation ($r = 0.20$) or education ($r = 0.19$). Ganzach (2000) reports a correlation of 0.30 between household income and educational attainment in the NLSY79, although the correlations with father's or mother's education were larger (at about 0.45).

Increasing income gaps in educational outcomes in the United States have been documented by a recent edited book (Duncan and Murnane, 2011). Between 1978 and 2008, the gap in students' mathematics test scores grew by one-third. In contrast the effect of parental education is quite stable. The increase in the impact of income is accounted for by wealthier parents more heavily investing in their children's cognitive development (Reardon, 2011). Similarly, income differences in college completion increased. The proportion of young people completing college increased by 21% points between cohorts starting school in the mid-1990s compared to cohorts starting school in the mid-1970s compared to only a 4% point rise among comparable low-income cohorts (Bailey and Dynarski, 2011). For the general increase in income inequalities in educational outcomes a variety of explanations are offered such as increasing economic segregation of schools, increases in single-parent families, the lack of resources of families in poverty, and the increased time spent by wealthy families on the cognitive development of their children. However, it is not clear if the increases in the income gap can be attributed to increases in the importance of cognitive ability as argued by Herrnstein and Murray (1994). For educational attainment, Duncan et al. (2013) conclude that the gap in the completed schooling between children in the top and bottom quintiles of the family income distribution increased by about half a year between the late 1960s and late 1990s. Increases in income inequality accounted for between two-thirds to three-quarters of the increasing schooling gaps between high- and low-income children but there was no increase in the effect of family income on educational attainment. One important contributing factor was increases in maternal age which is predictive of schooling.

Net Effects

However for educational outcomes, the impact of family income, net of other socioeconomic background factors, tends to be weaker than economic explanations of socioeconomic inequalities in education would predict. In the Wisconsin Longitudinal Study, family income was obtained from state tax records and averaged over 4 years (Sewell and Hauser,

1975, p. 18). It displayed surprisingly weak correlations with test scores (0.18) and educational attainment (0.28). In a regression analysis of educational attainment the standardized effect of family income ($\beta = 0.12$) was smaller than that for father's education ($\beta = 0.22$) and father's occupational status ($\beta = 0.14$) (Sewell and Hauser, 1975, pp. 72–73). In an earlier publication Sewell and Hauser (1972, p. 856) comment, "there may be little merit in the efforts of some social scientists to interpret all social inequalities in terms of income differences." Analyzing mathematics score in the NELS88 study, Fejgin (1995, p. 24) reported a standardized effect of 0.07 for family income compared to 0.32 for parents' education. An even weaker standardized effect ($\beta = 0.05$ for family income) was found for reading achievement. Analyzing mathematics achievement score in the children of the NLSY79 study, Orr (2003, p. 291, 293 model 2) reported no effect for family income (averaged over 5 years) on mathematics achievement, net of father's occupational status, mother's education, mother's AFQT score, and other variables (not including wealth).

In Olneck's (1977, footnote 1) analysis of educational attainment using sibling data, the addition of average family income over 3 years did not increase the explained variance beyond that provided by father's and mother's education, and father's occupation. According to an early analysis of the PSID, family income averaged over 5 years had no effect on young men's educational attainment, net of a crude measure of ability and father's education (Kiker and Condon, 1981, p. 102). In a review article, Haveman and Wolfe (1995, p. 1856) conclude "changes in economic resources are associated with only small changes in educational attainment." They cite Hill and Duncan's (1987) analysis of the PSID where a 10% increase in family income (averaged over 3 years) was associated with an increase in educational attainment by only 1.4%, less when controlling for other social background factors. Similarly, Belzil and Hansen (2003, p. 694) conclude that for educational attainment, father's and mother's education have much stronger effects than family income. A 1-year increase in parental education, on average, increases years of education by 0.3 of a year. To obtain a similar increase requires a difference of $30 000 in income, a large difference given the average income for that study (conducted in 1990) was $37 000. For school completion Belley and Lochner (2007, p. 51) found only a 7 or 8% point difference in school completion rates between the top and bottom family income quartiles, net of measured AFQT quartile and mother's education.

In the United States going to college is often expensive, so income should have stronger effects for college entry than school completion. However, economic factors play a minor role compared to academic and social factors. Hearn's (1988) analysis of factors that influence the cost of the institution attended concludes that the relatively minor effects of SES and ascriptive factors are more 'social' than 'economic.' Cameron and Heckman (2001, p. 485) found that family income has an important role at explaining grades at age 15, a modest role in high school continuation decisions but 'no role in college entry decisions' and contrary to prevailing economic theory conclude that credit market constraints are not a strong determinant of college choices. Belley and Lochner (2007, pp. 51–52) found modest effects for family income in the 1979 cohort; the difference in college attendance at age 21 between the top and bottom income quartiles, net of measured ability, was 9% points. For the NLSY97, the difference had risen to 16% points. This analysis included ability quartile but the only 'social' factor included was mother's education. For enrolling in U.S graduate programs, family income has only weak effects. Net of GPA and other factors, a $10 000 increase in income increases the probability of graduate school enrollment by only 0.37% (Zhang, 2005, p. 324). Morgan and Kim (2005, p. 177) also found small effects. They conclude, "that giving the average family $6000 per year in family income would increase the college enrollment rate among their adolescents by an additional two-thirds to one full percent." (The median and mean monthly incomes were around $2500 and $3000.) However, they do not regard this effect as causal, since other home background and academic performance measures were not included.

Furthermore, family income has only weak effects on social, cognitive, and behavioral outcomes among children (Aughinbaugh and Gittleman, 2003; Blau, 1999). Mayer's (1997) analysis of the PSID and NLSY79 also concluded that the influence of family income on social and economic inequalities is, at best, modest when taking into account other family characteristics.

It is not only in the United States where the effect of income on educational outcomes is weak. Bratti (2007) finds significant negative effects for household income on dropping out of school in England and Wales, but concludes "the magnitude of the effect is relatively small" (2007, pp. 36–37). After analyzing age finished formal education using the 1958 and 1970 British Cohort studies, Chevalier and Lanot (2002, p. 1879) conclude that "the effect of family income on a child's schooling attainment is rather limited." In their regression equation, family income was often not statistically significant, especially when controlling for ability. In the United Kingdom, Blanden and Gregg (2004) identified a small effect of income in the 1970 BCS controlling for unmeasured differences between families using a fixed effect model: a decrease of one-third in household income reduced the probability of obtaining a degree by between 1% and 6%. The effect of household income may have increased when compared to the British Household panel. According to analysis of the BHPS, the comparable effect was between 3% and 7%, although the BHPS does not include comparable measures of childhood ability.

In Canada, the association between dropping out of school and family income is weak (Foley et al., 2009). The effect of family income on university participation is very weak with R square values less than 4% and in some years less than 1% (Corak et al., 2003). For New Zealand, Fergusson et al. (2008) conclude that a family's economic resources (and the socioeconomic level of the school) do not mediate the relationship between socioeconomic background and educational achievement. Much of the relationship was mediated by a child's cognitive ability (35%) measured at ages 8, 9, and 13 years followed by educational aspirations. Analyzing educational attainment in Ireland, Chevalier and Lanot (2002, p. 179) found weaker effects for family income compared to other family characteristics. For Sweden, Erikson and Jonsson (1996, pp. 21–22) conclude "background factors other than economic resources are important in explaining the association between social origins and educational decision," although they suggest

that economic resources may be important for the transition to university. For Denmark, Jæger (2007, p. 546) found no significant effects on completing secondary school for family income, home ownership, car ownership, or summer house ownership for the youngest generation and only family income mattered for the parental generation. He concludes, "economic capital has largely ceased to predict educational attainment" (2007, p. 546). In Germany, income plays a much smaller role than parental education in the allocation of students to different academic tracks (Schneider, 2004). For Poland, the impact of income on educational attainment is generally weaker than that for other aspects of socioeconomic background, such as parent's education (Beblo and Lauer, 2004). For the Netherlands, the effect of monthly income on education level was about one-third that of parental education (Vollebergh et al., 2001). Less direct evidence is provided by Vries and De Graaf (2006, p. 226), who found no effect for father's occupational status (which is indexed by each occupational group's mean income and education), and conclude that 'the economic dimension does not matter in the Dutch Education system.' This conclusion was consistent with earlier research which concluded that financial resources were important in the Dutch education system before 1950 but not since then. The change is attributed to policies which removed financial barriers to education (de Graaf, 1986). In Norway, family income had no effect on grades at university when controlling for secondary school grades (Hansen and Mastekaasa, 2006, p. 284). Japan is a possible exception to the general pattern of family income having limited effects. The impact of family income on continuing education beyond the minimum school leaving age and academic qualifications was higher in Japan than in either Britain or the United States (Ishida, 1993).

Effects of Education on Earnings

Economists approach the relationship between education and earnings as the rate of return. This perspective is consistent with human capital theory where education is understood as an investment in future earnings. The rate of return is the percentage increase in earnings for each additional year of education, sometimes minus the costs of education and foregone income while studying. A similar methodology is used for the returns to primary, secondary, and tertiary education.

The rate of return is separated into private and social rates of return, separating the benefit accruing to individuals and the benefit to society more generally. According to the OECD, the private rate of return is equal to the discount rate that equalizes the real costs of education during the period of study to the real gains from education thereafter and includes the costs equal to tuition fees, foregone earnings net of taxes adjusted for the probability of being in employment minus the resources made available to students in the form of grants and loans. (From the OECD http://stats.oecd.org/glossary/detail.asp?ID=5412.) The social rate of return includes the increased productivity from education and a host of noneconomic benefits of better health, more social cohesion, and more informed and effective citizens minus the public costs of providing education. (From the OECD http://stats.oecd.org/glossary/detail.asp?ID=5426.)

There are a variety of related issues. The rate of return is unlikely to be constant for each additional year: the year completing school or completing a degree is likely to be higher than completing other years of education. There is also variation in the returns with some types of education (e.g., medicine, business) relating to higher earnings, whereas vocational qualifications are associated with lower earnings compared to university or college education. Another issue is the extent that the private returns represent the true returns to education or are biased upward because of the ability premium; higher ability persons tend to have higher levels of educational attainment and higher incomes. The private returns to education is an important concept in studies of discrimination: the extent that gender, racial, and other group differences in earnings are due to differences in endowments or differences in the returns to endowments (Jones and Kelley, 1984).

According to Psacharopoulos (1985, 1994) and Psacharopoulos and Patrinos (2004), the private rate of return in OECD countries (per year of education) is around 7–8% and higher in less developed countries such as Sub-Sahara Africa (13%), Asia (10%), and Latin America (12%). For higher education the private rate of return in OECD countries is about 12%, but 20% for the whole world. The returns to women (10%) are higher than those for men (9%) and higher in the public sector compared to the private sector. The estimates for the social rate of return are more difficult to estimate consistently and tend to be lower around 9% for OECD countries. Both the private and social rates of return have declined.

The impact of education on income is increasing in the United States (Burbidge et al., 2002). Gottschalk and Joyce (1998, p. 496) document strong increases in the impact of a college education for earnings in the United States between 1979 and 1981 of around 1.6% per year. There was a slightly weaker increase of 1.1% between 1986 and 1991. Lemieux (2006) concludes that the bulk of the rise in wage inequality from 1973 to 2005 in the United States is accounted for by increasing returns to postsecondary education relative to secondary education increased returns to postsecondary schooling. Deschenes (2006) concluded that causal return to education increased by 20–40% between 1979 and 2002 with no change in the ability premium. In contrast, Kaymak (2009) concluded that the returns to earnings for a 1-year increase in education increased from 4.8% to 8.4% between 1964 and 2003, and argues that the increase in the education premium after 1990 was caused almost entirely by ability.

In the Gottschalk and Joyce (1998) study the countries with the next strongest increases in the returns to college education after the United States during the 1980s were Sweden and Finland, two countries that did not experience an increase in earnings inequality. There were weaker increases for Canada, the United Kingdom, the Netherlands, and Australia. More recent evidence for the United Kingdom is mixed. Silles (2007) found increased returns to education among men but not among women, whereas Walker and Zhu (2008) found no evidence of increases in the college premium. However, there is little evidence for increasing effects of education for earnings for countries other than the United States and Great Britain: Australia, Canada, Norway, and the United Kingdom

Table 1 Effects of educational attainment and other influences on earnings

Country	*Educational attainment*	*Gender*	*Parents' education*	*Language native versus other*	*Literacy score*	*Labor market experience*	*R square*
Australia	0.28	0.41	−0.02	0.06	0.25	0.12	0.40
Belgium (Flanders)	0.44	0.48	−0.04	0.04	0.15	0.16	0.49
Canada	0.28	0.52	−0.09	0.05	0.32	0.20	0.48
Chile	0.40	0.55	0.06	−0.01	0.10	0.13	0.54
Czech Republic	0.38	0.47	0.05	0.06	0.09	0.23	0.41
Denmark	0.33	0.44	−0.02	−0.03	0.19	0.27	0.42
Finland	0.31	0.26	0.05	0.07	0.29	0.34	0.27
Germany	0.36	0.50	−0.21	0.06	0.08	0.20	0.39
Hungary	0.40	0.19	0.08	0.01	0.19	0.08	0.33
Ireland	0.31	0.36	0.05	0.03	0.21	0.11	0.35
Netherlands	0.30	0.48	−0.00	0.06	0.16	0.21	0.41
New Zealand	0.29	0.45	0.00	0.14	0.26	0.16	0.47
Norway	0.17	0.43	0.01	0.02	0.28	0.22	0.35
Poland	0.35	0.29	0.11	0.01	0.00	0.16	0.22
Portugal	0.23	0.35	0.21	0.06	0.20	0.35	0.41
Slovenia	0.46	0.24	−0.01	−0.03	0.18	0.05	0.37
Sweden	0.23	0.41	0.07	0.02	0.10	0.37	0.30
Switzerland	0.40	0.39	−0.06	0.01	0.12	0.29	0.39
United Kingdom	0.38	0.37	−0.06	0.03	0.29	0.17	0.48
United States	0.38	0.36	−0.02	−0.04	0.24	0.11	0.43

25–55 year olds. Presented are standardized regression coefficients from Structural Equation Modeling. Educational attainment is a latent variable constructed from both years of education and formal qualifications in some countries and only one of these in others. Literacy score is a latent construct based on the Prose, Document, and Quantitative Scales. Experience (in the labor force) is estimated by age minus years of education minus 5. 'Earnings' is measured on an interval scale based on quintiles since some countries did not collect or release continuous wage data.
OECD & Statistics Canada (2000, p. 175).

(Burbidge et al., 2002; Coelli and Wilkins, 2008; Haegeland et al., 1999; Marks, 2009). In Ireland and Spain, the returns to education have declined (Blázquez et al., n.d.; Voitchovsky et al., 2012). It is surprising that the returns to education have not decreased in many more industrialized countries, despite large increases in the proportion of the workforce with college degrees.

Sociological Approaches

In sociology, the theoretical background for the intergenerational income correlation is more an extension of the socioeconomic inequalities in educational and occupational attainment. The status attainment approach understands earnings, income, and wealth distal in a sequence of causal pathways. Earnings are understood as a dependent on occupation: in the "context of the allocation of occupational roles and related prerequisites" (Sewell and Hauser, 1975, p. 3). In other words, earnings depend on jobs. The causal sequence begins parents' characteristics, usually parental education and occupation, but can include family income and more rarely family wealth. These factors influence educational attainment mediated by mental ability, grades at school, aspirations, and expectations. In turn educational attainment influences income both directly and via occupation. Although father's earnings or family income is part of the causal sequence since parents can purchase schooling or other desirable goods for their children, economic background is given no special prominence in the causal pathway. The major focus is on the social forces that shape education and occupational attainment and, as a consequence, earnings.

A meta-analysis of the impact of cognitive ability and education on earnings estimated a standardized coefficient of 0.22 for education and 0.15 for ability (Bowles et al., 2001, p. 1154). Analysis of earnings in the 1992 U.S Adult Literacy Study produced standardized effects of 0.29 for educational attainment for white men and 0.31 for white women (Kerckhoff et al., 2001, p. 17).

According to the 1994–98 International Adult Literacy study, there are substantial effects of educational attainment on income net of gender, parent's education, literacy, and employment experience across industrialized nations. **Table 1** presents the standardized effects. The effects of educational attainment are substantial in every country with standardized coefficients between 0.17 and 0.46.

See also: Academic Achievement Motivation, Development of; Academic Performance, Effects of Socio-Economic Status on; Income Distribution and Inequality: Measurement Issues; Income Inequality; School Achievement as Moderated by Sociodemographic Status: The US Case; School Achievement: Motivational Determinants and Processes; Socioeconomic Status and Academic Achievement.

Bibliography

Aughinbaugh, A., Gittleman, M., 2003. Does money matter? A comparison of the effect of income on child development in the United States and Great Britain. Journal of Human Resources 38 (2), 416–440.

Bailey, M.J., Dynarski, S.M., 2011. Inequality in postsecondary education. In: Duncan, G.J., Murnane, R.J. (Eds.), Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances. Russell Sage and Spencer Foundations, New York, pp. 117–132.

Beblo, M., Lauer, C., 2004. Do family resources matter? Educational attainment during transition in Poland. Economics of Transition 12 (3), 537.

Becker, G.S., 1975. Human Capital, second ed. Columbia Uniersity Press, New York.

Belley, P., Lochner, L., 2007. The changing role of family income and ability in determining educational achievement. Journal of Human Capital 1 (1), 37–89.

Belzil, C., Hansen, J., 2003. Structural estimates of the intergenerational education correlation. Journal of Applied Econometrics 18 (6), 679–696.

Blanden, J., Gregg, P., 2004. Family income and educational attainment: a review of approaches and evidence for Britain. Oxford Review of Economic Policy 20 (2), 245–263.

Blau, D.M., 1999. The effect of income on child development. Review of Economics and Statistics 81 (2), 261–276.

Blázquez, M., Herrarte, A., Llorente, R., n.d. Changes in the Wage Return to Tertiary Education in Spain: Evidence from Matched Employer-Employee Data. Universidad Autónoma de Madrid, Madrid.

Bowles, S., Gintis, H., Osborne, M., 2001. The determinants of earnings: a behavioral approach. Journal of Economic Literature 39 (4), 1137–1176.

Bratti, M., 2007. Parents' income and children's school drop-out at 16 in England and Wales: evidence from the 1970 British Cohort Study. Review of Economics of the Household 5 (1), 15–40.

Burbidge, J.B., Magee, L., Robb, A.L., 2002. The education premium in Canada and the United States. Canadian Public Policy/Analyse de Politiques 28 (2), 203–217.

Cameron, S.V., Heckman, J.J., 2001. The dynamics of educational attainments for black, hispanic, and white males. Journal of Political Economy 109 (3), 455–499.

Carneiro, P., Heckman, J., 2002. The evidence on credit constraints in post-secondary schooling. Economic Journal 112 (482), 705–734.

Chevalier, A., Lanot, G., 2002. The relative effect of family characteristics and financial situation on educational achievement. Education Economics 10 (2), 165–181.

Coelli, M., Wilkins, R., 2008. Credential changes and education earnings Premia in Australia. Economic Record 85 (270), 13–34.

Corak, M., Lipps, G., Zhao, J., 2003. Family Income and Participation in Post-secondary Education (No. 210). Statistics Canada, Ottawa.

Deschenes, O., 2006. Unobserved Ability, Comparative Advantage, and the Rising Return to Education in the United States 1979–2002. University of California at Santa Barbara, Department of Economics.

Duncan, G.J., Kalil, A., Ziol-Guest, K.M., 2013. Increasing Inequality in Parent Incomes and Children's Completed Schooling: Correlation or Causation?.

Duncan, G.J., Murnane, R.J. (Eds.), 2011. Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances. Russell Sage and Spencer Foundations, New York.

Erikson, R., Jonsson, J.O., 1996. Explaining class inequalities in education: the Swedish test case. In: Erikson, R., Jonsson, J.O. (Eds.), Can Education Be Equalized? The Swedish Case in Comparative Perspective. Westview, Boulder, CO.

Fejgin, N., January 1995. Factors contributing to the academic excellence of American Jewish and Asian students. Sociology of Education 68, 18–30.

Fergusson, D.M., Horwood, L.J., Boden, J.M., 2008. The transmission of social inequality: examination of the linkages between family socioeconomic status in childhood and educational achievement in young adulthood. Research in Social Stratification and Mobility 26 (3), 277–295.

Foley, K., Gallipoli, G., Green, D.A., 2009. Ability, Parental Valuation of Education and the High School Dropout Decision (IFS Working Papers no. W09/21).

de Graaf, P.M., 1986. The impact of financial and cultural resources on educational attainment in the Netherlands. Sociology of Education 59 (4), 237–246.

Ganzach, Y., 2000. Parent's education, cognitive ability, educational expectations and educational attainment. British Journal of Educational Psychology 70 (3), 419–441.

Gottschalk, P., Joyce, M., 1998. Cross-national differences in the rise in earnings inequality: market and institutional factors. Review of Economics and Statistics 80 (4), 489–501.

Haegeland, T., Klette, T.J., Salvanes, K.G., 1999. Declining returns to education in Norway? Comparing estimates across cohorts, sectors and over time. Scandinavian Journal of Economics 101 (4), 555–576.

Hansen, M.N., Mastekaasa, A., 2006. Social origins and academic performance at university. European Sociological Review 22 (3), 277–291.

Haveman, R., Wolfe, B., 1995. The determinants of children's attainments: a review of methods and findings. Journal of Economic Literature 33 (4), 1829–1878.

Hearn, J.C., 1988. Attendance at higher cost colleges: ascribed, socioeconomic, and academic influences on student enrollment. Economics of Education 7, 65–76.

Herrnstein, R.J., Murray, C., 1994. The Bell Curve: Intelligence and Class Structure in American Life. The Free Press, New York.

Hill, M., Duncan, G.J., 1987. Parental family income and the socioeconomic attainment of children. Social Science Research 16 (1), 39–73.

Ishida, H., 1993. Social Mobility in Contemporary Japan. Stanford University Press, Hong Kong.

Jæger, M.M., 2007. Educational mobility across three generations: the changing impact of parental social class, economic, cultural and social capital. European Societies 9 (4), 527–550.

Jones, F.L., Kelley, J., 1984. Decomposing differences between groups: a cautionary note on measuring discrimination. Sociological Methods & Research 12 (3), 323–343.

Kaymak, B., 2009. Ability bias and the rising education premium in the United States: a cohort-based analysis. Journal of Human Capital 3 (3), 224–267.

Kerckhoff, A.C., Raudenbush, S.W., Glennie, E., January 2001. Education, cognitive skill, and labour force outcomes. Sociology of Education 74, 1–24.

Kiker, B.F., Condon, C.M., 1981. The influence of socioeconomic background on the earnings of young men. Journal of Human Resources 16 (1), 94–105.

Lemieux, T., 2006. Post-secondary education and increasing wage inequality. American Economic Review 96 (2), 195–199.

Marks, G.N., 2009. Modernization theory and changes over time in the reproduction of socioeconomic inequalities in Australia. Social Forces 88 (2), 917–944.

Mayer, S.E., 1997. What Money Can't Buy: Family Income and Children's Life Chances. Harvard University Press, Cambridge, MA.

Morgan, S.L., Kim, Y.-M., 2005. Inequality of conditions and intergenerational mobility: changing patterns of educational attainment in the United States. In: Morgan, S.L., Grusky, D.B., Fields, G.S. (Eds.), Mobility and Inequality: Frontiers of Research from Sociology and Economics. Stanford University Press, Stanford.

OECD & Statistics Canada, 2000. Literacy in the Information Age: Final Report of the International Adult Literacy Survey. Organisation for Economic Cooperation and Development Statistics Canada, Paris.

Olneck, M.R., 1977. On the use of sibling data to estimate the effects of family background, cognitive skills and schooling. In: Taubman, P. (Ed.), Kinometrics: Determinants of Socioeconomic Success within and between Families. North-Holland Publishing, Amsterdam.

Orr, A.J., 2003. Black-White differences in achievement: the importance of wealth. Sociology of Education 76 (4), 281–304.

Psacharopoulos, G., 1985. Returns to education: a further international update and implications. Journal of Human Resources 20 (4), 583–604.

Psacharopoulos, G., 1994. Returns to investment in education: a global update. World Development 22 (9), 1325–1343.

Psacharopoulos, G., Patrinos, H.A., 2004. Returns to investment in education: a further update. Education Economics 12 (2), 111–134.

Reardon, S.F., 2011. The widening academic achievement gap between the rich and the poor: new evidence and possible explanations. In: Duncan, G.J., Murnane, R.J. (Eds.), Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances. Russell Sage and Spencer Foundations, New York, pp. 91–115.

Schneider, T., 2004. The influence of parental income on school choice. Zeitschrift fur Soziologie 33 (6), 471–492.

Sewell, W.H., Hauser, R.M., 1972. Causes and consequences of higher education – models of status attainment process. American Journal of Agricultural Economics 54 (5), 851–861.

Sewell, W.H., Hauser, R.M., 1975. Education, Occupation and Earnings. Achievement in the Early Career. Academic Press, New York.

Silles, M.A., 2007. The returns to education for the United Kingdom. Journal of Applied Economics 10 (2), 391–413.

Sirin, S.R., 2005. Socioeconomic status and academic achievement: a meta-analytical review of research. Review of Educational Research 75 (3), 417–453.

Vries, De J., De Graaf, P.M., 2006. The reliability of family background effects on status attainment: multiple informant models. Quality and Quantity 42 (2), 203–234.

Voitchovsky, S., Maitre, B., Nolan, B., 2012. Wage inequality in Ireland's "Celtic Tiger" boom. Economic and Social Review 43 (1), 99–133.

Vollebergh, W.A.M., Iedema, J., Raaijmakers, Q.A.W., 2001. Intergenerational transmission and the formation of cultural orientations in adolescence and young adulthood. Journal of Marriage and Family 63 (4), 1185–1198.

Walker, I., Zhu, Y., 2008. The college wage premium and the expansion of higher education in the UK. Scandinavian Journal of Economics 110 (4), 695–709.

White, K.R., 1982. The relationship between socio-economic status and academic achievement. Psychological Bulletin 91 (3), 461–481.

Zhang, L., 2005. Advance to graduate education: the effect of college quality and undergraduate majors. Review of Higher Education 28 (3), 313–333.

Education and Learning: Lifespan Perspectives

Elizabeth AL Stine-Morrow and Brennan R Payne, University of Illinois at Urbana-Champaign, Urbana, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.J.A. Howe, volume 6, pp. 4189–4192, © 2001, Elsevier Ltd.

Abstract

A life span perspective on education gives consideration to the way in which education early in the life span alters trajectories of development throughout the life course, as well as to the ongoing effects of continued engagement in lifelong learning on development. Given the fact of longer life spans for successive generations, the ever-accelerating pace of knowledge development needed for work and managing everyday affairs, and the contributions of education to optimizing developmental outcomes, it is important that new models of life span education are developed.

Traditional assumptions hold that education and learning are primarily early life span endeavors that serve to prepare individuals for occupations and civic engagement. Such a limited view of the role of education is being overturned by a relatively recent shift in the understanding of the nature of adulthood and aging. Rather, lifelong education is essential for the healthy development of both individuals and societies.

First, there is accumulating evidence that the education of both formal and informal sorts contributes to lifelong health, well-being, and cognitive resilience. Further, the capacity to take the advantage of education and opportunities for learning is present throughout adulthood into very late life. Thus, access to education throughout the life span is an important facet of sound public health policy.

Second, driven in part by advances in medicine, we are living longer. Population aging on a global scale is unprecedented, pervasive, and projected to endure (United Nations Department of Economic and Social Affairs Population Division, 2002) so that by 2050, for the first time in history, the proportion of adults over the age of 60 is expected to match the number of people who are younger than 15 years of age. With this shift, conventional models of education, in which individuals are educated early in the life span to prepare for work and civic responsibility in adulthood, are no longer tenable. Such models were predicated on the assumption of shorter expected life spans relative to today. Longer lives are leading to trends for a more extended time engaged in work, even while the nature of work and the knowledge and skills required for work, are changing at an ever-accelerating pace. The consequence is that it is simply not feasible to educate children and young adults for the jobs they will have 40, 50, or 60 years hence. Also with this shift, conventional models of retirement, in which large numbers of younger people in the workforce support an extended period of disablement and/or leisure for a relatively small population of older adults, are no longer economically feasible. Some time ago, Riley and Riley (2000) identified these problems as ones of 'structural lag', in which the social structures that defined the roles available to individuals at particular ages were lagging behind the profound changes engendered by longer life spans. The upshot is that these age-segregated social structures must be transformed, so that education is accessible throughout the life span if we are to optimize health, well-being, and human capital.

Finally, a substantial number of individuals, for a variety of reasons, do not fully acquire literacy, numeracy, reasoning, and other basic skills needed for full participation in adult life during early life span educational experiences (National Research Council, 2012). Such individuals find their way to adult basic education programs, disability services in universities, and work-related training programs, often with some urgency of educational need, but with complicated adult lives. For such people, education science must offer efficient and effective means of meeting educational goals within the constraints of a busy adult life.

Thus, education is very much a life span enterprise. We open our review by considering the nature of life span cognitive development. Then we discuss the effects of education on development – first from the perspective of what early life span education contributes to lifelong development, and then from the perspective of how adult development and aging may be shaped by ongoing educational activities. Finally, we return to the issue of structural lag by considering new models of life span education.

Life Span Cognitive Development

Life span psychology is now a mature area of inquiry (Baltes et al., 1999), and it is widely recognized that development is incomplete at the transition into adulthood and extends across the entire life span. An understanding of the core tenets of life span psychology provides important grounding for conceptualizing the role of lifelong education.

Development Is Both Holistic and Multifunctional

An important contribution of life span psychology is a melding of both person-centered (holistic) and function-centered approaches. Development is multifunctional in the sense that the system comprises processes (e.g., perception, memory, reasoning, regulatory mechanisms, identity, dispositions, social strategies) that define the way in which the individual (at the system level) experiences the world and is able to adapt to it. Nevertheless, development is holistic in the sense that the person, with a unique constellation of these functions (e.g., abilities, dispositions, values), develops as

a system such that there is coherence across developmental periods or stages.

Development Is Multidirectional

Component processes constitute distinct functions that can interact but develop independently as a function of biological, environmental, and cultural resources. The consequence is that development is necessarily multidirectional. Development can be multidirectional in the sense that functional systems show divergent trajectories from one another. For example, sensory and perceptual systems generally show declines with age as a consequence of the biological process of aging, while the capacity for knowledge growth, which thrives on environmental enrichment and cultural context, appears to be fairly well preserved into very late life. Development can also be multidirectional in the sense that processes within functional systems can show divergent trajectories. For example, consider the case of problem solving: biological constraints may place limits on certain facets of cognition involving speeded processing of novel patterns with age, leading to declines in inductive reasoning, while life experience may afford growth in reflective judgment, contributing to expertise in social problem solving.

Among the most salient examples of multidirectionality in adult cognition is the divergent development of fluid and crystallized abilities. Fluid abilities or 'mental mechanics', the ability to apprehend and mentally transform novel patterns, are highly vulnerable to age-related biological declines, whereas crystallized abilities, which rest on knowledge and well-learned skills, are relatively resilient. Another example of multidimensional development is in the motivation for learning, which may be driven primarily by the need for information acquisition early in the life span, but by social–emotional concerns later in the life span.

Plasticity Is Present Throughout the Life Span

Plasticity, the capacity for change in response to experience, exists throughout the life span. Plasticity is distinct from flexibility, defined by one's current skill set that enables adaptation within a particular range of demands, and implies the ability to both gain in efficiency and form new representations (Lövdén et al., 2010). Thus, plasticity implies the capacity to increase in flexibility. So, for example, learning a second language is likely to increase the range of situations in which one can function, that is, flexibility, and depends upon existing plasticity. Some evidence suggests that bi- or multilingualism enhances cognitive function more broadly (e.g., cognitive control, ability to learn vocabulary in a novel language; Bialystok et al., 2010); to the extent that this is the case, it would suggest that learning a second language may enhance plasticity itself (i.e., enhancing potential for learning in new domains). Some theorists argue that the hallmark of plasticity at the neural level is structural alteration in networks, while changes in activation within existing neural networks define flexibility (cf Lövdén et al., 2010). By this view, plasticity is a sluggish capacity, which can be thought of as adaptive, since a system that was incessantly rewiring itself would be unstable and ultimately maladaptive. As an adaptive capacity to respond to new demands, plasticity requires a relatively long-term mismatch between demands and existing skill sets. Importantly, current levels of flexibility contribute to the range of environmental presses to which the individual can respond with increased plasticity: demands that do not challenge current skill sets do not engender plasticity, nor do demands outside the range of current skill sets afford a fruitful engagement with the experience. This notion that development and learning depend on experiences that challenge the system within an optimal range is pervasive in the psychological literature, at least since Vygotsky's discussion of the *zone of proximal development*, and more recently, in Metcalfe's work on the *region of proximal learning* (e.g., Metcalfe and Kornell, 2005).

While plasticity is at the core of human nature, there is wide individual variation in plasticity, so that some individuals require a more prolonged experience of increased demand to increase adaptive capacity. As a rule, plasticity tends to decrease later in the life span, as a consequence of age-related limits on neurogenesis, so that plasticity may become increasingly sluggish. At the same time, individual variation in plasticity is certainly present throughout the life span.

Implications for Education

Education and life span developments are intimately related. On the one hand, it is important that we understand how education contributes to development. On the other hand, we must know how to create instructional programs that can effectively teach individuals at different stages of the life span. One consequence of multidimensionality and plasticity is that adults become increasingly different from one another, so that a life span perspective on education implies developing instructional approaches that can accommodate such variation. In the next sections, we review what is known about the lifelong effects of education early in the life span, and then consider how education can shape development during adulthood.

The Effects of Early Education on Life Span Development

Educational experience during youth and young adulthood can confer a number of lifelong advantages. Perhaps obviously, this can include skills and regulatory capacities that contribute to employment and social mobility, and dispositional attributes that engender plasticity by enabling comfort with novelty, or sustained investment in stimulating activities. Perhaps less obvious, but the focus of much recent research is that education can have lasting effects on brain and cognitive health into late adulthood.

It has long been recognized that educational level, psychometric intelligence, and socioeconomic status (SES) or income are correlated. These interrelationships are likely multidetermined and complex. Because intelligence tends to be a moderately reliable individual difference throughout the life span, and because individuals are selected into educational experiences based on demonstrated intellectual merit, the effects of education on intelligence are difficult to parse out. However, there are good reasons to believe that at least some of

these effects can be directly attributed to the effect of education on intellectual development. For example, based on evidence from 'natural experiments' (e.g., accidents of geography or history that impact educational opportunities largely without regard for talent), Ceci and Williams (1997) concluded that education is a direct cause of intellectual growth. This is not an uncontroversial position given arguments that intelligence is relatively immutable and that education primarily serves the function of selection among those with intellectual capacity, rather than playing a direct role in intellectual growth (Herrnstein and Murray, 1994). Nevertheless, there is increasing evidence for the plasticity of intelligence and the important role that education plays in this plasticity. A recent example of a natural experiment of the sort described by Ceci and Williams comes from Brinch and Galloway (2012), who examined the effects of school reform in Norway on increasing the level of compulsory education, and subsequently increasing IQ scores in adolescence.

The availability of large-scale longitudinal data sets that include measures of childhood or young adult intellectual function, educational attainment, and occupational and health data (e.g., Scottish Mental Survey, Lothian and British Cohort Studies, Project Talent) has spawned a new area of inquiry over the last decade, called 'cognitive epidemiology', which examines the longitudinal correlates of intelligence early in the life span. A strikingly well-replicated finding is the ability of intelligence to predict health in adulthood, as well as mortality (Deary, 2012), a topic to which we will return in a moment. Another consequence of such large longitudinal data sets is the opportunity to use sophisticated statistical techniques to examine mechanisms underlying differential development using mediational models (cf MacKinnon et al., 2007). So for example, evidence from the British Cohort Study (cf Deary, 2012) suggests that the relationship between childhood IQ and SES is mediated by educational investment. Of course, questions about the impact of educational engagement on intellectual growth, health, and SES inevitably depend on correlational data, and even with the most carefully done structural equation modeling, the best we can do is test whether the data are consistent with causal accounts, so that debate on this issue is lively (e.g., Deary and Johnson, 2010, 2011; Richards and Sacker, 2011).

The rise in intelligence test scores for successive generations (Schaie and Zanjani, 2006) is likely attributable to a number of factors, including conditions that promote better health (e.g., sanitation, medical advances), but also better access to education for later cohorts. The plausibility of the conclusion that education serves a distinctive role in this cohort drift in intelligence is strengthened by examining divergence in particular abilities: while primary mental ability scores on verbal meaning and inductive reasoning have shown robust increases across cohort, numerical skills have tended to decrease, declines that coincide historically with the advent of calculators.

A recent study by Ritchie et al. (2013) used longitudinal data from the Lothian Birth Cohort (LBC) 1921 and 1936 studies to examine the education–intelligence link. On 1 June 1932, the Scottish Council for Research in Education administered an assessment of general intelligence to every 11-year-old child in the country who was in school that day, with the goal to get data on the distribution of intellectual capacities across Scotland. As a result, about 95% of the population of 11-year-olds was tested – at a point prior to the onset of ability-differentiated education. In an effort to understand possible generation differences in intelligence as a result of historical change in family structure, the Scottish Council repeated this effort in 1947. The LBC studies are follow-up assessments of those living the Lothian region of Scotland in old age. Ritchie et al. (2013) took advantage of this extraordinary data set to address the question of whether education contributed unique variance to intellectual capacities in late life. Controlling for intelligence at age 11 and SES, the researchers found that educational level was predictive of intelligence scores in late life. Interestingly, the advantages of education were greater for those with initially lower intelligence test scores. Additionally, there were no reliable effects of education on speed of processing tasks, an indicator of low-level cognitive processing. Collectively, this research strongly points to education in youth as an important contributor to intelligence. Given the relatively strong predictive relationship of intelligence to many life outcomes (Deary, 2012; Kuncel et al., 2004), it may be that these outcomes can be in part attributed to education.

Perhaps among the most startling findings from cognitive epidemiology is the discovery that educational achievement early in life is predictive of health, mortality, and cognitive resilience in old age. As with the education–intelligence link, it is challenging to isolate causal mechanisms. For example, it is plausible that education could promote health by conferring economic advantage that enables better access to medical care; by shaping dispositional traits, like conscientiousness, that engender skills in self-regulation and self-care that promote health by establishing a lifestyle early in the life span that engenders habits of mental stimulation that promote intellectual growth; and/or by establishing a consistent and predictable environment that buffers against the toxic effect of stress. It is also plausible that a life course of generalized fitness is created early in the life span by some combination of genes, prenatal, and early life experience that predisposes individuals to both take advantage of education and experience good health. Of considerable interest in the current research, literature is the proposal that early educational experiences directly offset normative and pathological effects of aging to engender cognitive health.

The constructs of 'resilience' and 'reserve' have appeared in the aging and adult development literature to broadly capture the ideas that older individuals may be able to avoid pathological risks associated with aging, and normative age-related cognitive declines, respectively. A well-replicated finding in clinical neuropsychology is that there is often a disjunction between the degree of brain pathology and the clinical manifestation of such pathology on behavior (e.g., cognitive performance or the ability to manage everyday activities). In particular, at autopsy the brains of relatively well-educated individuals often show *more* evidence of brain pathology than those of less well-educated individuals, who had comparable levels of cognitive function prior to death. 'Reserve capacity', the idea that education engenders neural and cognitive growth that buffers against the assaults of pathology, has been used to explain such findings

(Christensen et al., 2008; Stern, 2009). It is certainly well established from both animal and human training models that experience can induce neurogenesis (i.e., the growth of neural networks) (Draganski et al., 2004, 2006; Kramer et al., 2004).

Some researchers have distinguished between passive and active models of reserve. Passive models of reserve emphasize the neuroprotective effects of variation in the capacity of the brain to withstand pathology. Measures of interest in such passive brain reserve models include head circumference, brain volume, synaptic density, or other biophysical attributes. Once a threshold is reached, both normative effects of aging (e.g., changes in hippocampal volume) and pathological effects of aging (e.g., amyloid plaques and neurofibrillary tangles) manifest themselves clinically and behaviorally. This model assumes that, holding all other things equal, neural networks that are larger or more robust are capable of withstanding more damage before the damage manifests itself functionally. So, for example, there is evidence that education buffers against normative declines in volume of the hippocampus, a brain structure critically involved in encoding new memories (Noble et al., 2012). By contrast, active models of reserve emphasize the degree to which individual differences in task processing (e.g., alternative cognitive strategies, differential recruitment of networks) contribute to cognitive optimization, which can allow some people to cope better than others with "brain changes in general and aging in particular" (Stern, 2009: p. 2027). Such models posit that individuals with greater cognitive reserve are able to compensate for the effects of aging through compensatory changes in neural recruitment (such as increased recruitment of contralateral hemispheric homologues), or through the use of different cognitive strategies (such as an increased reliance on prior knowledge). So, for example, Ritchie et al. speculated that the effects of education on intelligence test scores but not on tasks of processing speed (as described above) may be due to the greater effects of education on active reserve. Note that explanations in terms of active and passive reserve are not mutually exclusive, as compensatory cognitive strategies result in changes in neural recruitment.

Studies that have examined the dual influence of both early education and more proximal variables have found evidence to support the notion that early educational level provides unique predictive validity for later cognitive health. Researchers have found both independent and synergistic effects of education with other potential covariates, including health status, late-life cognitive activities, and premorbid intelligence. Some studies have shown independent contributions from education, while others have shown larger effects of later life health and lifestyles choices on cognition among those with lower levels of early formal education. This suggests that early educational experiences may buffer against the influence of later life choices on cognitive health.

Although early life span educational attainment shares a high correlation with later life cognitive ability, it is unclear if the influence of education on rates of dementia is due to an attenuation of declines in cognition among the more highly educated. Some studies have found evidence that education moderates the rates of cognitive decline, but recent results from the Atherosclerosis Risk in Communities Study, a large-scale ($N > 9000$) study of community-based, ethnically diverse elders, report no effects of education on age-related declines in episodic memory or verbal fluency over 15 years, despite high correlations between education and initial cognitive abilities (Schneider et al., 2012). While findings from such large samples cast doubt on the prospect that early educational attainment slows cognitive decline, such findings should be interpreted carefully.

One issue clouding the interpretation of the effects of education on cognitive change is the use of total number of years of education as a proxy for educational attainment and cognitive reserve. Self-reports of the total number of years educated have proven to be surprisingly strong indicators of cognitive reserve (Stern, 2009). At the same time, more proximal measures of cognitive reserve capacity and educational attainment may exist, especially in populations for which number of years of education may not adequately capture the quality of educational opportunities and experiences. Indeed, measures of literacy, premorbid verbal ability, and reading comprehension have been argued to be stronger markers for cognitive reserve than the total number of years of education completed because they are more direct measures of attainment, especially among underserved populations. For example, in an ethnically diverse sample of nondemented elders, Manly et al. (2005) found that increased rates of literacy were not only associated with cognitive capacity at baseline measurement (as is commonly found), but also with slower longitudinal declines in episodic memory, executive functioning, and fluency over 8 years. Recent work has also shown that such measures of literacy are predictive of language comprehension and memory among older adults, with greater benefits of literacy for sentence recall, especially among older adults with lower cognitive capacity (Payne et al., 2012; Kavé et al., 2012).

Moreover, such longitudinal findings are difficult to interpret because the effects of education on declines in cognition likely vary quite substantially across individuals and are nonlinear in nature. Indeed, the protective effects of education appear to decrease as pathology increases in the brain, leading to the finding that highly educated individuals eventually show steeper declines in function, once a threshold of pathology has been reached (Stern, 2009). This finding, while initially counterintuitive, is central to the understanding of the effects of education on cognitive health; early life span protective factors, such as education compress cognitive morbidity, such that abilities are preserved (or effects of disease are delayed) until some time later in the life span. Indeed, some studies have estimated of the protective effects of education (comparing individuals with less than high school education versus those with more than high school education) to be equivalent of upward of 20 years of normative cognitive aging (Schneider et al., 2012). However, once highly educated adults have eventually met such a threshold of pathological burden, terminal declines in cognition begin with an accelerated negative trajectory. Due to this nonlinear trajectory that is dependent upon an individual's time relative to terminal decline, it becomes difficult to interpret the effects of education on longitudinal change without information about terminal thresholds for each individual. In studies that have examined the rates of cognitive decline relative to such thresholds, such as the onset of dementia diagnoses for example,

such education-dependent compression effects have been observed (see Stern, 2009).

Lastly, it has been argued that 'full' tests of the effects of education on cognitive and brain reserve are lacking in the literature, as a test of the model requires a measure of reserve (e.g., education, occupational, or activity engagement), a measure of neural integrity or pathology (e.g., brain amyloid), and a measure of cognitive ability or clinical impairment (Christensen et al., 2008). Findings from the religious orders study (Bennett et al., 2003) provide a particularly clear example of a 'full' test of cognitive reserve among the highly educated. This study found that higher levels of dementia-related brain pathology (e.g., amyloid plaques, neurofibrillary tangles, Lewy bodies) at autopsy were associated with lower levels of cognitive function before death. However, the correlation between markers of brain pathology and measures of cognitive function was attenuated among those with higher levels of formal education. Other recent results, such as those from the Epidemiological Clinicopathological Studies in Europe (EClipSE) collaboration (EClipSE Collaborative Members, 2010), support active models of the compensatory effects of education on dementia precedence in older adulthood. EClipSE included harmonized longitudinal data from three population-based studies that each included postmortem brain donation, clinical investigation of dementia, and measures of educational attainment ($N = 872$). Interestingly, while education was associated with decreased dementia risk and overall greater brain size, there was no relationship between education and risk of neurodegenerative disease or vascular pathology. However, education did moderate the degree to which all measures of neurovascular pathology were associated with the clinical expression of dementia. These results suggest that even though higher education does not provide direct protection against vascular and neurodegenerative disease (but may indirectly, through the association between education and overall brain size), higher education does buffer against the negative effects of neuropathological burden on cognitive function.

Education in Adulthood

Education in adulthood in formal settings (e.g., courses with defined goals, criteria for success, and evaluation) is not currently normative. However, such formal educational experiences have the potential to contribute to adult development along a number of lines, including the enhancement of work-related skills, acquisition of knowledge and skills to effectively manage everyday life (e.g., parenting, managing health care for a family, retirement decisions), and as a pathway to lifelong health. Currently, educational activities in adulthood are more typically the informal sort, including investment in work, literacy activities, hobbies, and social activities.

It is surprising to many to learn that many people enter adulthood lacking in the basic literacy and numeracy skills needed to fully participate in work, family, and civic life. A recent report from the National Research Council (2012) summarized evidence, based on a nationally representative sample in the US, that only about 57% of adults have prose literacy skills that would be considered above basic proficiency, and that only about 55% have quantitative literacy skills in that range. Thus, there are many adults, who have acquired basic skills in word decoding, reading simple sentences, and arithmetic, but are unable to use these skills to draw inferences, reason, or make decisions. Thus, many of the institutions for adult education exist to provide remedial instruction to help adults develop proficiency in reading and quantitative reasoning. Adults also find their way into educational venues for work-related training, as well as instruction to support avocation-related interests.

In contrast to the literature on the lifelong effects of early education on intellectual growth, the research base on the ongoing effects of continued schooling is, perhaps not surprisingly, thinner. After all, it is not normative for adults to be engaged in formal educational activities, and when they are, they have made the choice to engage in these activities, so that cross-sectional comparisons are vulnerable to selectivity biases. Nevertheless, what literature there is strongly points to the conclusion that both formal schooling and certain forms of activity engagement can support preservation and growth in both crystallized and fluid abilities. For example, Beier and Ackerman (2005) provided a life span sample (age 19–68) of individuals with educational modules in health- and technology-related topics, and showed that existing initial strengths in fluid and crystallized abilities and prior knowledge contributed to the development of domain-specific knowledge. Hatch et al. (2007) showed that participation in work-related and elective educational experiences by the age of 43 contributed to the levels of verbal ability and verbal memory in later life when early life education, social mobility, and prior cognitive ability were controlled (thus, reducing the impact of reverse causation).

There is abundant evidence from the literature that adults can benefit from explicit training in particular cognitive abilities (e.g., speed, reasoning, memory training), though less clear are the principles by which training particular cognitive skills transfers, or generalizes to, broader cognitive capacities and knowledge growth (Stine-Morrow and Basak, 2011). Another area of relevant research is that on the effects of cognitive enrichment in the form of late-life activity engagement. Hertzog et al. (2008) extensively reviewed the literature and identified seven large-scale studies that measured both rates of activity engagement and rates of cognitive decline over time. They found significant associations across all but one of the studies, some of which had sample sizes of over 4000 individuals. Similar associations have also been found between rates of mentally stimulating activities and the incidence of cognitive impairment (Alzheimer's and mild cognitive impairment). While these findings are highly suggestive, they do not necessarily indicate a causal relationship between stimulating activity engagement and improved cognition. Muddying this causal relationship is the substantial difficulty with measuring intellectually stimulating activity, especially as perceptions of cognitive demand in an activity differ across individuals (Payne et al., 2011). Moreover, the level of cognitive activity could be a marker for other more proximal causes. Reduced cognitive activity could index early neuropathology, essentially suggesting a reversed causal direction of the relationship. Consistent with this argument is a recent longitudinal study suggesting that there

are reciprocal relationships between change in activity engagement and change in cognition, such that individuals who decline more in some cognitive abilities (e.g., speed) show later declines in activity engagement. Cognitive activity does have a more long-term influence on cognitive change over decades; however, weakening the argument that reduced cognitive activity could be solely an indicator of a prodromal period of disease-related neuropathology. Another argument that has been proposed is that cognitive activity engagement may be a marker for a number of other beneficial health and SES outcomes, which benefit the quality of life. Importantly, however, some studies have shown that the relationship between late-life cognitive activity and the incidence of Alzheimer's disease remain even after controlling for estimates of early-life cognitive activity and SES.

While engagement in formal education is not normative in adulthood, adults often do continue to work in substantively complex environments into late life, and this provides an important pathway to examine the relationship between informal educational opportunities and mental stimulation. In an interesting approach to examine the impact of work on cognitive resilience, Rohwedder and Willis (2010) compared cognitive scores across countries as a function of retirement policies, thus circumventing the issue of individual selection into complex environments, assuming that individuals have minimal control over national retirement policies. Rohwedder and Willis analyzed data from three cross-national surveys that provide comparable assessments of cognition: the Health and Retirement Survey in the United States; the English Longitudinal Study of Aging; and the Survey of Health, Ageing and Retirement in Europe, which collected data from 11 European countries. In each case, surveys were based on large nationally representative samples administered over the phone, in which cognition was assessed by delayed recall for 10 concrete nouns (a task that typically shows age declines). Results showed that individuals in countries that had policies incentivizing early retirement (e.g., by taxing earned income at a higher rate) had steeper declines in memory between the early 1950s and early 1960s, relative to those who lived in countries whose tax policies encouraged continued work. Although correlational, the relationship between engagement in work and mental decline reported by Rohwedder and Willis strongly implies that the mental demands of work promote cognitive resilience.

Conclusions: Innovating Life Span Models of Education

Based on our review, we believe that it is clear that individuals can benefit along a number of dimensions from educational experiences throughout the life span, but that a coherent model of true life span education is substantially lacking at our current point in history. In what follows, we suggest some challenges that will need to be overcome in order to meet this need. Lastly, we review some innovative models of cognitive activity engagement in adulthood that provide helpful points of guidance for future models of life span education.

The literature on the effectiveness of adult education and learning interventions is scant. At the same time, research on 'nontraditional' cognitive interventions and community-based activity engagement programs in adulthood is emerging (Hartman-Stein and LaRue, 2011), based on a substantial amount of both cross-sectional and longitudinal evidence suggesting a relationship between intellectual activity engagement and cognitive health. These intervention studies have been designed with the goal of assessing the causal effect of activity engagement on cognitive outcomes. Of these, projects that have required participants to adjust their day-to-day routines, requiring substantial time and effort commitments, appear to produce considerable benefits. One such program is the Experience Corps Project, in which older adults work as volunteers in underserved Baltimore elementary schools, working with children in literacy development activities in library settings, for example. Early findings suggested that the intervention promoted increased rates of cognitive and physical activities, as well as increases in executive control. Similarly, in the Senior Odyssey Project, older adults work in teams on creative problem-solving activities, such as building balsa wood structures that can hold large amounts of weight, and staging performances. The intervention produced positive effects on certain aspects of fluid ability (e.g., ideational fluency). Another example of a recent innovation in adult education is the Osher Lifelong Learning Institute (OLLI). There are over 100 OLLI programs associated with college campuses around the US, open to adults of age 50 and over. OLLI programs are governed by members, who organize course offerings and recruit instructors each term. Collectively, these studies have suggested that engagement in substantively complex environments may support cognitive resilience.

An optimal model of adult life span education will need to take into account normative age-related changes (and variability in changes), including declines in fluid 'mental mechanics' (e.g., executive functions, speed of processing, reasoning, episodic memory), increased reliance on cognitive strengths (such as vocabulary and general knowledge), and shifts in strategies used to solve complex problems. An important challenge to meeting this need is that adults have complex lives with competing demands, so that educational resources must be readily available and programs must be efficient in meeting learning goals. It is likely that technology will have to play some role in meeting these demands. With much public discussion over the last decades about the 'graying of America', the conversation has focused largely on adaptations needed in systems for delivering health care. Neglected is the need for equally fundamental changes in the nature of education.

See also: Adult Education and Lifelong Learning: The US Experience and Beyond; Cognitive Development: Child Education; Cross-Cultural Study of Education; Intelligence, Prior Knowledge, and Learning; Learning Theories and Educational Paradigms; Learning and Instruction: Social-Cognitive Perspectives; Lifelong Learning and Its Support with New Technologies; Literacy Education; Situated Learning (Learning In Situ); Vygotsky's Theory of Human Development and New Approaches to Education.

Bibliography

Baltes, P.B., Staudinger, U.M., Lindenberger, U., 1999. Lifespan psychology: theory and application to intellectual functioning. Annual Review of Psychology 50, 471–507.

Bennett, D.A., Wilson, R.S., Schneider, J.A., et al., 2003. Education modifies the relation of AD pathology to level of cognitive function in older persons. Neurology 60, 1909–1915.

Beier, M.E., Ackerman, P.L., 2005. Age, ability, and the role of prior knowledge on the acquisition of new domain knowledge: promising results in a real-world learning environment. Psychology and Aging 20, 341–355.

Bialystok, E., Craik, F.I.M., Green, D.W., Gollan, T.H., 2010. Bilingual minds. Psychological Science in the Public Interest 10, 89–129.

Brinch, C.N., Galloway, T.A., 2012. Schooling in adolescence raises IQ scores. Proceedings of the National Academy of Science 109, 425–430.

Ceci, S.J., Williams, W.M., 1997. Schooling, intelligence, and income. American Psychologist 52, 1051–1058.

Christensen, H., Anstey, K.J., Leach, L.S., Mackinnon, A.J., 2008. Intelligence, Education, and the Brain Reserve Hypothesis the Handbook of Aging and Cognition, third ed. Psychology Press, New York. pp. 133–188.

Deary, I.J., 2012. Intelligence. Annual Review of Psychology 63, 453–482.

Deary, I.J., Johnson, W., 2010. Intelligence and education: causal drive analytic processes and therefore conclusions. International Journal of Epidemiology 39, 1362–1369.

Deary, I.J., Johnson, W., 2011. Authors' response: is education causal? Yes. International Journal of Epidemiology 40 (2), 516–518.

Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U., May, A., 2004. Changes in grey matter induced by training. Nature 427, 311–312.

Draganski, B., Gaser, C., Kempermann, G., Kuhn, H.G., Winkler, J., Büchel, C., May, A., 2006. Temporal and spatial dynamics of brain structure changes during extensive learning. The Journal of Neuroscience 26, 6314–6317.

EClipSE Collaborative Members, 2010. Education, the brain, and dementia: neuroprotection or compensation? Brain: A Journal of Neurology 133, 2210–2216.

Hartman-Stein, P., LaRue, A. (Eds.), 2011. Enhancing Cognitive Fitness: A Guide to the Use and Development of Community Programs. Springer, New York.

Hatch, S.L., Feinstein, L., Link, B.G., Wadsworth, M.E.J., Richards, M., 2007. The continuing benefits of education: adult education and midlife cognitive ability in the British 1946 birth cohort. Journal of Gerontology: Social Sciences 62B, S404–S414.

Herrnstein, R.J., Murray, C., 1994. The Bell Curve: Intelligence and Class Structure in American Life. Free Press, New York.

Hertzog, C., Kramer, A.F., Wilson, R.S., Lindenberger, U., 2008. Enrichment effects on adult cognitive development: can the functional capacity of older adults be preserved and enhanced? Psychological Science in the Public Interest 9, 1–65.

Kavé, G., Amit, S., Spaiter, T., Ben-Ezra, M., 2012. Formal education level versus self-rated literacy as predictors of cognitive aging. Journal of Gerontology: Psychological Sciences 67, 697–704.

Kramer, A.F., Bherer, L., Colcombe, S.J., Dong, W., Greenough, W.T., 2004. Environmental influences on cognitive and brain plasticity during aging. Journal of Gerontology: Medical Sciences 59A, 940–957.

Kuncel, N.R., Hezlett, S.A., Ones, D.S., 2004. Academic performance, career potential, creativity, and job performance: can one construct predict them all? Journal of Personality and Social Psychology 86, 148–161.

Lövdén, M., Bäckman, L., Lindenberger, U., Schaeffer, S., Schmiedek, F., 2010. A theoretical framework for the study of adult cognitive plasticity. Psychological Bulletin 136, 659–676.

MacKinnon, D.P., Fairchild, A.J., Fritz, M.S., 2007. Mediation analysis. Annual Review of Psychology 58, 593–614.

Manly, J.J., Schupf, N., Tank, M.X., Stern, Y., 2005. Cognitive decline and literacy among ethnically diverse elders. Journal of Geriatric Psychiatry and Neurology 18, 213–217.

Metcalfe, J., Kornell, N., 2005. A region of proximal learning model of study time allocation. Journal of Memory and Language 52, 463–477.

National Research Council, 2012. Improving Adult Literacy Instruction: Options for Practice and Research. National Academies Press, Washington, DC.

Noble, K.G., Grieve, S.M., Korgaonkar, M.S., Engelhardt, L.E., Griffith, E.Y., Williams, L.M., Brickman, A.M., 2012. Hippocampal volume varies with educational attainment across the life-span. Frontiers in Human Neuroscience 6, 1–10. Article 307.

Payne, B.R., Gao, X., Noh, S.R., Anderson, C.J., Stine-Morrow, E.A.L., 2012. The effects of print exposure on sentence processing and memory in older adults: evidence for efficiency and reserve. Aging, Neuropsychology, and Cognition 19, 122–149.

Richards, M., Sacker, A., 2011. Is education causal? Yes. International Journal of Epidemiology 40, 516–518.

Riley, M.W., Riley Jr., J.W., 2000. Age integration: conceptual and historical background. The Gerontologist 40, 266–270.

Ritchie, S.J., Bates, T.C., Der, G., Starr, J.M., Deary, I.J., 2013. Education is associated with higher later life IQ scores, but not with faster cognitive processing speed. Psychology and Aging 28 (2), 515–521.

Rohwedder, S., Willis, R.J., 2010. Mental retirement. Journal of Economic Perspectives 24, 119–138.

Schaie, K.W., Zanjani, F.A.K., 2006. Intellectual development across adulthood. In: Hoare, C. (Ed.), Handbook of Adult Development and Learning. Oxford University Press, New York, pp. 99–122.

Schneider, A.L.C., Sharrett, A.R., Patel, M.D., Alonso, A., Coresh, J., Mosley, T., Gottesman, R.F., 2012. Education and cognitive change on 15 years: the atherosclerosis risk in communities study. JAGS 60, 1847–1853.

Stern, Y., 2009. Cognitive reserve. Neuropsycholgia 47, 2015–2028.

Stine-Morrow, E.A.L., Basak, C., 2011. Cognitive interventions. In: Schaie, K.W., Willis, S.L. (Eds.), Handbook of the Psychology of Aging, seventh ed. Elsevier, New York, pp. 153–170.

United Nations Department of Economic and Social Affairs Population Division, 2002. World Population Ageing: 1950–2050. http://www.un.org/esa/population/publications/worldageing19502050/.

Education: Anthropological Aspects

Roger Goodman, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The discipline of anthropology has brought a number of important points to bear on the study of education – a relativistic perspective, a holistic approach, and the ethnographic method. At the same time, the study of education has become increasingly common in anthropological accounts, generally under the headings of the study of ethnicity and nationalism, or cognitive and psychological anthropology. This article examines the development of the anthropological study of education and explores why, despite its apparent importance, it is still looking to establish a secure niche as a subfield of the discipline of anthropology.

The study of education would appear prima facie to be central to the anthropological project, and courses that cover anthropological aspects of education are taught in a number of universities, particularly in the United States. Educational anthropology, however, has yet to create as secure a niche for itself as a separate academic discipline as, for example, medical or legal anthropology, which has become increasingly influential in the study of industrial, as well as preindustrial, societies. One reason for the lack of a clearly defined anthropology of education has been that its areas of interest and methodological approaches have overlapped so much with those of other disciplines, in particular social psychology, sociology, and, of course, the growing field of educational studies. Indeed, some of the most influential work in the study of education has been produced either in collaborative projects between anthropologists and those from other disciplines or by the latter bringing anthropological insights to bear in their own work. The influence of anthropology – in terms of scope, theory, and methodology – on the study of education, therefore, is rather more than it may at first appear.

Scope of Study

For anthropologists, the study of education is far from confined to the examination of formal educational institutions such as schools or to the experiences of those who have not yet been defined as adults. Education, in an anthropological sense, includes the whole range of learning experiences of individuals from conception to death. Indeed, to avoid the narrower connotations of the word 'education,' anthropologists have tended to use much broader terms such as 'socialization' and (less often) 'enculturation,' both of which emphasize that an individual's knowledge of what it means to be a social being is constantly refined throughout their life. The underlying assumption in the anthropological project is that every society or subsociety has an idea of what constitutes an 'educated person' (Levinson et al., 1996), but that this concept differs between societies and changes over time. One classic distinction in the study of education that many anthropologists have made is between verbal and nonverbal forms of teaching (Lave and Wenger, 1991; Singleton, 1999). The former generally takes place in formal settings, such as schools, and the latter in informal settings, such as apprenticeships. In verbal forms of teaching, the responsibility for ensuring correct understanding lies with the teacher, while in nonverbal forms it lies with the student who learns through close observation of the teacher.

In order to study such broad concepts as 'socialization' and 'enculturation,' anthropologists need to look at many features of the particular society they are examining, from kinship and religion to political and economic systems. When anthropologists study formal educational institutions – such as schools, companies, military, religious, and sporting organizations – they also take a holistic approach. They examine how these institutions reflect wider social values and attempt to inculcate them into students through formal and informal means. For example, culturally specific ideas about children and childhood in a society impinge on the educational environment in that society and lead to very different types of classroom experiences. At the same time, classroom activity in any society is presented as a 'natural' process, even though it is clearly historically and culturally contingent.

These characteristics of the anthropology of education can be understood more easily if we look at how they have developed historically and at some of their underlying theoretical and methodological assumptions.

Anthropology and Education: A Historical Overview

A history of the development of the anthropological study of education might be divided into four main periods. Ethnographers at the end of the nineteenth and early twentieth centuries generally included, in their accounts of the life-cycle of the peoples they studied, some description of how children were cared for, educated, and made full members of society, often through complex initiation rites. In many ways, these accounts differed little from those provided by travelers, missionaries, and social commentators over the previous centuries who documented their observations of the treatment of the young in the societies they visited or in their own societies. Indeed, some comparative educationalists trace their subject back to Herodotus and Xenophon in Ancient Greece and Cicero in Ancient Rome (Trethewey, 1976). Only occasionally would the earliest Victorian anthropological theorists, such as Edward Tylor,

draw on examples of educational practices from around the world in order to support their theories about how societies evolved from 'savagery' to 'civilization,' but this was never done in a systematic manner.

The beginning of what we now know as the anthropology of education and socialization, therefore, should perhaps most usefully be linked to the first attempts to theorize systematically about the socialization methods of different societies. In this project, Arnold Van Gennep, Franz Boas, and Bronislaw Malinowski were perhaps particularly important.

Van Gennep's work on rites of passage introduced the idea that, while the rituals that mark the transition between life-cycle stages within any society are often both arbitrary and flexible, the underlying pattern of ritual performance often shows a remarkable similarity across societies. At the same time, Boas in the United States, and Malinowski in the United Kingdom were among the first to challenge Victorian theories of social evolution, which equated the poor performance of certain groups in formal Western educational institutions with cultural or biological inferiority. They did this in part by showing how socialization practices are designed to meet the particular 'needs' of a society at a given point in time. Their ideas of cultural relativity that intelligence can only be judged in the context of a particular culture and not on a universal scale helped form the basis of a counterargument to the use of universal tests of intelligence in a debate that continues to rage vigorously today.

In Britain, students of Malinowski such as Raymond Firth, Phyllis Kaberry, and Audrey Richards continued into the 1930s to include in their ethnographies detailed descriptions of how children became full members of society. Followers of the alternative structural-functionalist tradition of A.R. Radcliffe-Brown, however, steered away from the study of education; they considered it to be too close to psychology and the concept of 'culture' that they eschewed in favor of the study of 'society.'

The second major phase of the development of the anthropology of education was the establishment of the Culture and Personality School in the 1930s and 1940s, the best-known proponents of which were students of Boas, such as Edward Sapir, Ruth Benedict, and Margaret Mead. This school, which enjoyed by far its greatest influence in the United States, primarily involved the application of psychoanalytic principles to ethnographic material. It proposed, among other things, that by studying child-rearing practices such as feeding, weaning, and toilet training, it was possible to understand the basis on which personality types were developed in different societies. The school is probably best known for the work of Margaret Mead; her argument that sex roles were determined culturally, as opposed to biologically, played an important part in the reexamination of cultural assumptions about men and women in many industrial societies.

The Culture and Personality School probably hit its height of influence in the wartime period with the publication of, so-called, National Character Studies, such as Benedict's classic account of Japanese society, *The Chrysanthemum and the Sword*. In the postwar period, however, such approaches became increasingly criticized for exaggerating the homogeneity of personality types within any one society and the link between early child-rearing practices and later adult behavior, for ignoring the significance of relationships that exists between societies, and for reifying society rather than regarding it as a social construction. The influence of the Culture and Personality School could continue to be seen, however, in the way data on socialization practices were collected in the Human Relations Area Files, and in the more quantitative studies of child-rearing carried out cross-culturally by Whiting and Child in the *Six Cultures Series* (1966). It was visible in the work of George and Louise Spindler (Spindler, 2000) who became probably the most influential proponents of educational anthropology in the United States in the 1960s and 1970s. It also provided an important intellectual basis for what later became known as psychological anthropology and cognitive anthropology. The culture and personality approach, however, seemed to its critics to provide little insight into the politically and historically contingent nature of socialization and educational practices and it, therefore, had little influence in British social anthropology where the study of children and education virtually disappeared from the work of anthropologists for 20 years (Richards, 1970).

What might be seen as the third phase in the development of the anthropology of education began in the late 1960s. This phase was linked, to some extent, with the growth of civil rights movements in North America and Western Europe. In the United States, it focused in particular on what anthropologists generally termed 'cultural discontinuities' in the educational experience of minority and immigrant groups within the US educational system. Detailed ethnographic studies that examined how children, who did not come from the mainstream, white, middle-class society, experienced education in the United States were drawn on to develop programs of multicultural education and greater sensitivity to the educational needs of children from different ethnic backgrounds. As John Ogbu (1981) pointed out, however, these new programs could sometimes be at the expense of essentializing – freezing in time – and stereotyping the 'culture' of minority groups.

In Britain, similarly detailed ethnographic studies were being undertaken on British schools but with the focus on issues of class rather than ethnicity. Several of these (Hargreaves, 1967; Lacey, 1971) were undertaken by members of the social anthropology department at Manchester University whose head at the time, Max Gluckman, was keen to see anthropological methods used in the study of industrial societies. Manchester today remains one of the few British universities with a course in, and undertaking research on, the anthropology of education and learning (Evans, 2006). Probably the most influential work in the UK in this period, however, was Paul Willis' (1977) study of working-class culture in a British secondary school. Willis described himself as a sociologist and he was one of the key founders of the Birmingham Cultural Studies School that drew on the neo-Marxist ideas of Gramsci. In terms of his approach, he relied on the ethnographic method following a tradition in sociology that can be traced back to the Chicago School of the 1920s and 1930s and which continues today in the field of educational studies (Walford, 2008).

The research that probably had the most lasting influence on the field of educational anthropology from the 1970s, though, was carried out by the French scholar, Pierre Bourdieu

who developed the concepts of 'cultural capital' (which the elite in society pass on to their offspring), 'symbolic violence' (which schools impose on nonelite students so that they learn to silence themselves when in the company of those with greater social standing), and '*habitus*' (the system of often unconscious dispositions that individuals acquire within the social forces that constrain them). While Bourdieu is also often described as a sociologist, the original insights for his later work can be found in the mainstream anthropological research that he carried out among the Kabyle peasants in Algeria (Bourdieu, 1977). Important streams of work that can be seen as developing some of the key ideas of Bourdieu include Lave and Wenger's *Situated Learning* (1991) and Levinson et al.'s *The Cultural Production of the Educated Person* (1996). Interestingly, one area of research that Bourdieu pioneered – the anthropological study of the professoriate undertaken in *Homo Academicus* (1988) – has gained much less traction. While there have been some anthropological studies of the changing environment of higher education (Bailey, 1977; Strathern, 2000) and some ethnographies of student culture (Moffatt, 1989; Nathan, 2005), detailed ethnographic studies of academics have been as Poole (2010), in one of the few exceptions to this rule, points out virtually nonexistent.

At the same time as ethnographers were examining the ways in which education systems reproduced class and ethnic inequalities, feminist anthropologists in many Western societies became interested in studying the mechanisms whereby gender patterns and inequalities were reproduced through the same system (Okely, 1978). Overall, therefore, by the end of the 1970s, the anthropology of education could be described as one of the most politically active areas of the discipline as well as one of the few areas where Western anthropologists were actively making studies of, and becoming involved in, policy making in their own societies.

From the end of the 1970s, anthropologists of education became rather more outward looking, although they often remained politically committed. In part, this involved more applied work being undertaken by anthropologists in developing countries. This work often identified closely with the ideas of the anticolonial movement and was sometimes highly critical of the manner in which Western educational systems had been 'imposed' on developing countries and tried to show how this had sometimes led to social disintegration rather than integration (Foley, 1977). Perhaps more influential for policies in the west, however, were an increasing number of detailed ethnographic studies by anthropologists of the operation of the education systems of societies whose economies seemed to be rapidly developing at a time when those of the USA and Western Europe seemed, in the late 1970s and early 1980s, to be stagnating.

A large number of works on the anthropology of Japanese education, in particular, began to appear in the light of Japan's astonishing economic growth during the 1980s, which many commentators both inside and outside Japan believed could best be explained by its education system. While some politicians in the west seemed to suggest that Japan might provide a model for the reform of their own education systems, particularly preschool and elementary systems, most anthropologists pointed out the problems inherent in the piecemeal importation of elements of a foreign education system and the impossibility of adopting a system wholesale. Instead, they urged policy makers to use descriptions of foreign education systems as a mirror for examining the strengths and weaknesses of their own systems (Rohlen, 1983).

The fourth major phase in the history of the anthropology of education involved a change of focus from studying those who run and organize education and other socialization systems to looking at the cognitive worlds of those being educated (Amit-Talai and Wulff, 1995). While there had been calls within anthropology since the 1970s, influenced largely by the work of Edwin Ardener, for children, along with women, to be given their own voice in ethnographies (Hardman, 1973), this really only began to emerge two decades later. An impetus for this development may in part have been the United Nations Convention on the Rights of the Child. The Convention certainly sparked intense debate in many countries throughout the 1990s about what constituted a 'child' and what were 'children's rights.' Ideas about the 'child' and 'childhood' were shown to differ greatly between societies with consequent implications for issues of punishment and control in the formal educational institutions in those countries; some commentators indeed felt that the Convention represented an overly Anglo-Saxon view of the child that was being ethnocentrically imposed on other societies (Boyden, 1990).

The twenty-first century has seen both a great growth and diversification of work on the anthropology of education. There is a clear tension, for example, between those who see their work in terms of advocacy and those who see it in terms of providing empirical 'scientific' accounts. This can perhaps best be seen in the mission statements of different international organizations. For example, the Council on Anthropology and Education, which is a section of the American Anthropological Association, describes its mission as 'to advance antioppressive, socially equitable, and racially just solutions to educational problems through research using anthropological perspectives, theories, methods.' On the other hand, another group, Ethnography and Education International, is 'committed to the development of ethnography in education as a vital research methodology that prioritizes the perspectives of people's lives.' Its focus is 'support ethnographic research that involves long-term engagement with those studied in order to understand their cultures, to use multiple methods of generating data, and to recognize the centrality of the researcher in the research process.' There has been a growing interest by members of this group in the ethnographic study of online communities and the effects of new technologies in teaching and learning.

Another interesting tension can be found within the Comparative and International Education Society, which is a cross-disciplinary organization that includes historians, sociologists, economists, psychologists as well as anthropologists, and educators. Its focus is on the 'theoretical basis of comparative studies (as applied to) policy and implementation issues in developing countries and cross-cultural settings.' In recent years, there has been an interesting debate within the society between anthropologists and world culture theorists; the former arguing for the need to focus on diverging local understandings of, and variations in, teaching and learning, the latter on converging global institutional forms that have developed in educational systems. Anderson-Levitt (2003) has

tried to span this divide by arguing that it is possible to have common structures but different lived experiences in educational systems around the world. As she points out, the ideas about 'teacher autonomy' or 'student participation' can have very different meanings in different contexts depending on how a society defines the idea of a 'teacher' or 'student' (and the appropriate relationship between them) at any particular point in time. The case studies in her book also show how global discourses are interpreted and consumed in local contexts and not just passively implemented. A more recently edited volume by Anderson-Levitt (2012) shows how ethnographic practices and anthropological approaches to the study of education have also developed differently in different national settings.

As will be evident, the anthropology of education these days covers all societies and all aspects of the learning and teaching process from the most formal to the most informal. The full range of topics and approaches that the contemporary anthropologist of education covers can be seen in the recently published *A Companion to the Anthropology of Education* (edited by Levinson and Pollock), which includes work by 50 of the leading scholars in the field.

Assumptions

There are two overlapping sets of assumptions that need to be examined in order to explain the (often unconscious) different approaches that anthropologists have taken to the study of education. The first set relates to whether the anthropologist's primary interest is in the way education systems structure the behavior of individuals or in the way individuals construct education systems. This dichotomy is often described as one between a focus on 'structure' or 'agency.'

The second set of assumptions relate to the ideas – held both by the researcher and the society being researched, though these may not be the same – about how the process of socialization actually operates. For example, in the case of socializing children, is the child a passive object waiting to be molded by society, or an already formed individual who just needs an environment in which he or she can naturally develop, or a person intent on pursuing his or her own interests and on whom society has to impose its will (see Jahoda and Lewis, 1988)? Sometimes, these distinctions are discussed in terms of the child as *tabula rasa*, the Apollonian child (the child as naturally good) and the Dionysian child (the child as innately willful).

Approaches that concentrate on how socialization systems, particularly formal education systems, help to construct, disseminate, and legitimate a commonly understood and accepted ritual and symbolic language and system of meaning within a society, thereby structuring the experience of individuals who are often presented as *tabula rasa*, are often described as Durkheimian or functionalist. Such analyses have been particularly influential in the study of large-scale modern societies where it is impossible for all the individuals to know each other personally. Formal education systems have played a major role in the formation of the modern nation state; they have often been the main means through which a common language has been disseminated and a national identity constructed in a population. Through their interaction with the economic and religious elements of the social system, they have helped to develop a national work ethic and inculcate national spiritual values. The study of education, therefore, has clearly been important in the work of anthropologists of ethnicity and nationalism such as Benedict Anderson and Ernest Gellner.

Marxist theories have also focused on the way in which education practices produce social beings but rather than seeing these as aids to social integration they have concentrated on how they maintain and reproduce social differences. Marxist and neo-Marxist approaches to the study of education largely replaced functionalist and Durkheimian approaches in the 1970s with studies of how class and ethnic inequalities were reproduced through the education system in the USA and UK. These studies showed how an education system can be systematically designed – by means of 'linguistic codes' or a 'hidden curricula' – to provide differential educational experiences for different class, ethnic, and gender groups in a society. While the structural effects of class reproduction could be examined quantitatively, the actual mechanisms of the 'black box' (as it came to be called) of education could be discerned only through intensive qualitative research undertaken by ethnographers.

Social action theorists have focused on how the individual actors in the educational setting help to construct or resist its formal structures. Unlike Durkheimians and Marxists, they do not see individuals in an education system as passive but as active participants in a process of 'cultural production.' Where active participation supports the integrative or the social reproductive functions of education, the analyses of social action theorists overlap with those of Durkheimians and Marxists. As Willis (1977) showed in his account of working-class youth in the UK, for example, the subcultures of resistance formed by these youths to the middle-class curriculum often only served to recreate the class structure of the wider society. Social action theorists, however, generally suggest that there is a range of reactions to the learning process in any educational setting as well as a range of teaching styles. In describing these, they often draw, implicitly at least, on Max Weber's concept of 'ideal types.' Christina Toren, who established the first European Centre specifically for the anthropological study of childhood, has focused on how the study of children's learning is an important window into how humans in general come to be who they uniquely are through interaction with others (Toren, 1999). Her work, which draws on the psychologist Piaget's idea of 'schemas' and the philosopher Merleau-Ponty's ideas of intersubjectivity, constitutes what she has called an 'anthropology of consciousness.'

Understanding the differences between the positions outlined above is important in the discipline of anthropology as anthropologists are generally themselves the main research tools. The assumptions on which anthropologists base their research affect not only what they look at but also the types of questions they ask and the conclusions they are likely to draw.

Methodology

While there are a number of antecedents to the ethnographic approach to the study of society, it was largely patented and

systematized by Malinowski in the 1920s and became closely associated with the discipline of social anthropology. The application of this approach may be the most significant contribution made by anthropology to the study of education. Almost all anthropologists of education have relied on some form of field research, ranging from total participant observation to informal interviews (Woods, 1986). Through all of these techniques, the researcher tries to gain an understanding of the participants' – teachers,' pupils,' parents,' and officials' – views on the educational process. In its ideal form, indeed, the researcher becomes through participant observation socialized in the society or institution being studied and thereby subjectively comes to empathize with and understand its values.

While field research has been shown to be a very effective means of understanding how socialization works in the broadest sense, it has perhaps particularly been suited to the study of education in 'bounded' institutions, with clearly defined boundaries in terms of time and space, such as schools, churches, and other environments where children are taught basic social categories formally and informally. When children start formal schooling in Western societies, for example, they learn how concepts of time – work time vs play time, school time vs home time, holiday time vs term time – and space – public vs private, male vs female, safe vs dangerous – are divided in the extradomestic environment. Such institutions, therefore, provide an excellent setting to investigate what are considered to be, in a particular society at a particular point in time, its core social values and how these values are socialized – through teaching the meaning of dominant symbols and key ritual performance – and legitimated through reference to ideas of 'tradition' and 'culture.' Field research also allows the researcher to gain a sense of the extent to which these values are accepted, rejected, or ignored by those being taught them.

Contribution to the Study of Education

The discipline of anthropology has brought a number of important points to bear on the study of education – a relativistic perspective of socialization practices, a holistic approach, and an ethnographic method that allows for the simultaneous study of structure and agency in the educational process – which have all become standard practice in the field even if they are not always associated directly in researchers' minds with anthropology. At the same time, the study of education has become increasingly common in anthropological accounts, generally under the headings of the study of ethnicity and nationalism, cognitive and psychological anthropology, or the growing field of the anthropology of children rather than as a distinctive field of educational anthropology.

See also: Anthropology and History; Anthropology: Overview; Educational Sociology.

Bibliography

Amit-Talai, V., Wulff, H. (Eds.), 1995. Youth Cultures: A Cross-cultural Perspective. Routledge, London.
Anderson-Levitt, K. (Ed.), 2003. Local Meanings, Global Schooling: Anthropology and World Culture Theory. Palgrave Macmillan, New York.
Anderson-Levitt, K. (Ed.), 2012. Anthropologies of Education: A Global Guide to Ethnographic Studies of Learning and Schooling. Berghahn, London.
Bailey, F.G., 1977. Morality and Expediency: The Folklore of Academic Politics. Blackwell, Oxford.
Bourdieu, P., 1977. Outline of a Theory of Practice. Cambridge University Press, Cambridge, UK.
Bourdieu, P., 1988. Homo Academicus. Polity Press, Cambridge, UK.
Boyden, J., 1990. Childhood and the policy makers: a comparative perspective on globalisation of childhood. In: James, A., Prout, A. (Eds.), Constructing and Reconstructing Childhood: Contemporary Issues in the Sociological Study of Childhood. Polity Press, London, pp. 184–215.
Evans, G., 2006. Educational Failure and Working Class White Children in Britain. Palgrave Macmillan, Basingstoke.
Foley, D., 1977. Anthropological studies of schooling in developing countries: some recent findings and trends. Comparative Education Review 21, 311–328.
Hardman, C., 1973. Can there be an anthropology of children? Journal of the Anthropological Society of Oxford 4, 85–99.
Hargreaves, D.H., 1967. Social Relations in a Secondary School. Routledge and Kegan Paul, London.
Jahoda, G., Lewis, I.M. (Eds.), 1988. Acquiring Culture: Cross-cultural Studies in Child Development. Croom Helm, London.
Lacey, C., 1971. Hightown Grammar: The School as a Social System. Manchester University Press, Manchester, UK.
Lave, J., Wenger, E., 1991. Situated Learning. Cambridge University Press, Cambridge, UK.
Levinson, B.A., Foley, D.E., Holland, D.C. (Eds.), 1996. The Cultural Production of the Educated Person: Critical Ethnographies of Schooling and Local Practice. SUNY, Albany, NY.
Moffatt, M., 1989. Coming of Age in New Jersey: College and American Culture. Rutgers University Press, New Brunswick, New Jersey and London.
Nathan, R., 2005. My Freshman Year: What a Professor Learned by Becoming a Student. Cornell University Press, Ithaca, NY.
Ogbu, J., 1981. School ethnography: a multilevel approach. Anthropology and Education Quarterly 12, 3–29.
Okely, J., 1978. Privileged, schooled and finished: boarding education for girls. In: Ardener, S. (Ed.), Defining Females: The Nature of Women in Society. Croom Helm, London, pp. 93–122.
Poole, G.S., 2010. The Japanese Professor: An Ethnography of a University Faculty. Sense Publishers, Rotterdam, Boston, Taipei.
Richards, A., 1970. Socialization and contemporary British anthropology. In: Mayer, P. (Ed.), Socialization: The Approach from Social Anthropology. Tavistock, London, pp. 1–32.
Rohlen, T.P., 1983. Japan's High Schools. University of California Press, Berkeley, CA.
Singleton, J.C., 1999. Learning in Likely Places: Varieties of Apprenticeship in Japan. Cambridge University Press, Cambridge, UK.
Spindler, G. (Ed.), 2000. Fifty Years of Anthropology and Education, 1950–2000: A Spindler Anthology. Lawrence Erlbaum, Mahwah, N.J., London.
Strathern, M. (Ed.), 2000. Audit Cultures: Anthropological Studies in Accountability, Ethics and the Academy. Routledge, London.
Toren, C., 1999. Mind, Materiality and History: Explorations in Fijian Ethnography. Routledge, London.
Trethewey, A.R., 1976. Introducing Comparative Education. Pergamon Press, Oxford, UK.
Walford, G. (Ed.), 2008. How to Do Educational Ethnography. The Tufnell Press, London.
Whiting, B.B., Child, I.L., 1966. Six Cultures Series. Wiley, New York.
Willis, P., 1977. Learning to Labour. Saxon House, Farnborough, UK.
Woods, P., 1986. Inside Schools: Ethnography in Educational Research. Routledge and Kegan Paul, London.

Education, Economics of

Eric A Hanushek, Hoover Institution, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The explosion of research work on the economics of education has covered a wide range of issues. A key element of modern analyses is a focus on the outcomes of schooling and how various factors affect outcomes. Along with this, there has been considerable attention to production relationships and particularly to efficiency of production. The existing work suggests that variations in how well schools convert resources and inputs into student learning are a very important issue that has direct implications for education policy. Because commonly measured characteristics of teachers and schools – such as teacher experience, teacher degree levels, or class size – are not consistently related with student outcomes, the research suggests that these are not good policy instruments. On the other hand, the strong evidence about the importance of teacher quality points toward noticeable changes in the finance and governance of schools.

Study of the economics of education has expanded very rapidly, becoming one of the more popular fields of study of economists. The breadth of study is perhaps easiest to see by the fact that there are now four separate Handbook volumes addressing the various topics that are logically included in the area (Hanushek and Welch, 2006; Hanushek et al., 2011). This breadth and depth of research implies that it is logically impossible to do justice to the full field that ranges from the demand for schooling, the supply of alternative educational opportunities, and the impacts of schooling on subsequent outcomes. This discussion, rather than attempting to be exhaustive, concentrates on the more limited range of issues related to the organization, funding, and performance of schools, although these will be put into the context of the broader field.

Many facets of the broader field overlap those of labor economics, public finance, and growth theory, and they can best be put into those contexts – leaving this discussion to the unique facets of schools. The economics of education also borrows from other disciplines such as sociology and psychology, but the underlying behavior perspective remains unique to economics.

The economics of education is naturally linked to the study of human capital. Human capital refers to the skills and productive capacity embodied in individuals. While implicitly part of economics for several centuries (Kiker, 1968), the idea of human capital developed into a central concept in both theoretical and empirical analyses with the foundational work of Schultz (1961), Becker (1964), and Mincer (1970, 1974). Nonetheless, while abstractly dealing with skills and capabilities of individuals, human capital – to be both predictive and testable – must be defined in terms of more concrete measures. This requirement most often brings up school attainment as a clear, measurable aspect of human capital that can serve as a proxy for major skill differences. While an expansive view could include all research topics that touch on schooling, it is useful to define the economics of education more narrowly in terms of unique aspects not covered in other subdisciplines: namely, consideration of the education sector itself. This discussion is also limited by available data and analyses. Due to a dearth of reliable outcome data from higher education as well as schools in other countries, the economics of education has concentrated its study largely on K-12 schools. (The study of the economics of higher education, largely lacking student performance information, has directed most attention to issues of access and attendance and particularly of the influence of financial aid and costs on these (see, for example, McPherson and Schapiro (2006), Kane (2006), or Bettinger et al. (2009)). More recent analyses, benefiting from an expansion of administrative data for higher education, have moved to other questions about the production process (for example, Bound and Turner (2007) or Bettinger and Long (2010)). Nonetheless, because these analyses have such a different perspective from those related to primary and secondary education, they are not included in this discussion.)

Interestingly, however, the field has rapidly expanded to cover the entire world. While once restricted largely to American schools, data and analysis have expanded to cover the globe.

Part of the expansion of interest in the economics of education has reflected the development of extensive data on schools – partially as a result of movements to expand accountability of schools for performance. But a complementary impetus for this expansion has been the direct relationship between research and educational policy. The recent demands for 'evidence-based decision making' has heightened interest in the information about program effectiveness and the impacts of institutions on educational outcomes.

Outcomes

It is natural for economists to think in terms of a production model where certain factors and influences go in and products of interest come out. This model, called the 'production function' or 'input-output' approach, is the model behind much of the analysis in the economics of education. The measurable inputs are things like school resources, teacher quality, classroom peers, and family attributes, and the outcome is student achievement. Let us focus on the outcome,

first, and then move on to inputs, where more controversy has historically existed.

In the development of the field, the most frequently employed measure of schooling has been school attainment or simply years of schooling completed. The United States tradition led the world in investing in schooling. By the mid-1970s, three-quarters of US students completed high school, culminating a long national investment period that started with just 6% graduating from high school at the turn of the century. Many developed and developing countries, however, have recently mimicked this trend, as illustrated in **Figure 1**, which displays the increasing completion of secondary schooling for more recent age cohorts. By 2008, the growth in attainment around the world is readily seen.

The value of school attainment as a rough measure of individual skill has been verified by a wide variety of studies of labor market outcomes. In the United States, Mincer (1974) pioneered the approach that is now standard. Psacharopoulos and Patrinos (2004) have taken this analysis to the rest of the world, surpassing 100 countries for which estimates of the returns to years of schooling are available. The estimation considers the relationship between earnings (viewed as a direct measure of individual productivity) and schooling and labor market experience.

However, the difficulty with this common measure of outcomes is that it assumes a year of schooling produces the same amount of student achievement over time and in every country. This measure simply counts the time spent in schools without judging what happens in schools. Moreover, it neglects any complementary source of human capital development such as families, peers, or health inputs – thus, it does not provide a complete or accurate picture of outcomes. Today, more attention is focused on quality outcomes, and the most common measure is tests for cognitive skills (for a more general discussion, see Hanushek and Woessmann (2008)).

To the extent that the productive value of schooling largely involves the ability to make decisions under uncertainty and to adapt to new ideas and technologies (cf Nelson and Phelps, 1966; Welch, 1970), cognitive skills appear to be a valid measure of human capital. This conclusion is supported by recent studies that show a direct correlation between test scores and earnings in the labor market (Hanushek and Zhang, 2009).

While early studies of wage determination indicated relatively modest impacts of variations in cognitive ability after holding constant quantity of schooling, more recent direct investigations of cognitive achievement find generally larger labor market returns to measured individual differences in cognitive achievement. Three US studies provide very consistent estimates of the impact of test performance on earnings of young workers (Mulligan, 1999; Murnane et al., 2000; Lazear, 2003). These studies employ different nationally representative data sets that follow students after they leave school and enter the labor force. When scores are standardized, they suggest that one standard deviation increase in mathematics performance at the end of high schools translates into 10–15% higher annual earnings. In a different set of estimates using data on a sample of workers for the United States, Hanushek and Zhang (2009) provide estimates of returns of 20% per standard deviation. (The estimates for the United States do, however, exceed those for other countries.) One distinguishing feature of these estimates is that they come for a sample of workers throughout the career, as opposed to the prior estimates that all come from early-career earnings. Using yet another methodology that relies upon international test scores and immigrants into the United States, Hanushek and Woessmann (2012) obtain an estimate of 14% per standard deviation.

Even more significantly, society appears to gain in terms of productivity. Hanushek and Kimko (2000) demonstrate that quality differences in schools have a dramatic impact on

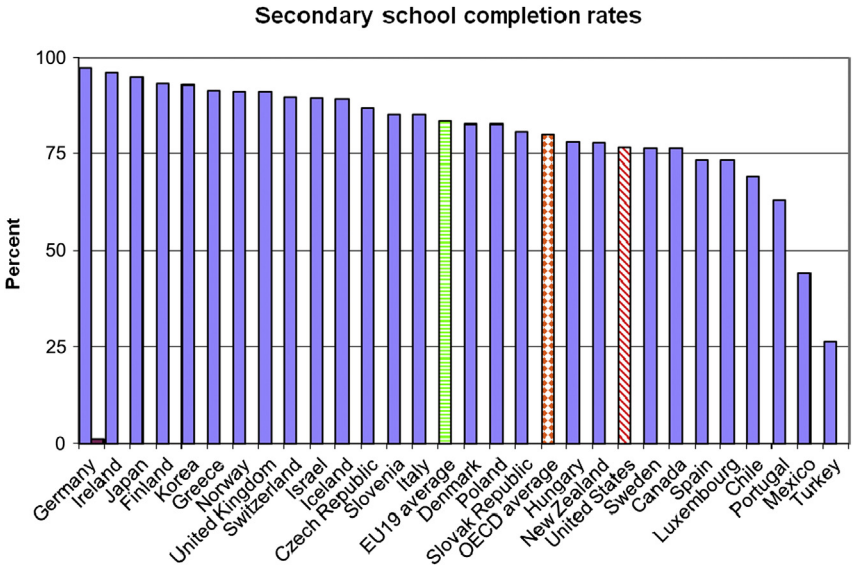

Figure 1 Secondary school completion rates, 2008. Source: Organisation for Economic Co-operation and Development, 2008. Education at a Glance 2008: OECD Indicators. OECD, Paris.

productivity and national growth rates. These estimates have been duplicated in a variety of studies (see Hanushek and Woessmann (2008)). Moreover, the implications of long-term growth and economic well-being of having different levels of achievement are truly astounding (Hanushek and Woessmann, 2011). There is also reason to believe that there is a causal relationship between cognitive skills and economic growth (Hanushek and Woessmann, 2012). These estimates of the impact of school quality on economic outcomes parallel the attention to quality and cognitive skills that has become a growing policy interest of many countries. As countries have approached more universal access to schooling, they have turned to issues of quality of publicly provided schooling.

Production

Let us now look at the other side of the education equation – the inputs, or determinants, of student achievement. Because outcomes cannot be changed by fiat, much attention has been directed at inputs – particularly those perceived to be relevant for policy such as school resources or aspects of teachers.

Analysis of the role of school resources in determining achievement begins with the 'Coleman Report,' the US government's monumental study on educational opportunity released in 1966 (Coleman et al., 1966). The report was conducted in compliance with a mandate of the Civil Rights Act of 1964 to investigate the extent of inequality in the nation's schools. Although this was not the first such effort, it was much larger and much more influential than any previous study. Only recently, with the now-common use of administrative databases on student performance compiled by states, have studies exceeded the survey information from the Coleman Report in breadth of samples and ability to investigate different educational circumstances.

The Coleman Report's greatest contribution was directing attention to the distribution of student performance – the outputs as opposed to the inputs. Instead of addressing questions of inequality by simply listing an inventory of differences of schools and teachers by race and region, it highlighted the relationship between inputs and outputs of schools.

Unfortunately, most of the attention the report generated focused on the report's conclusions rather than its innovative perspective. The controversial conclusion was that schools are not very important in determining student achievement; on the contrary, families and, to a lesser extent, peers are the primary determinants of performance variance. The findings immediately led to a large (but decentralized) research effort to compile additional evidence about input-output relationships in schools. (There were also extensive analyses of the report's methodology and of the validity of its inferences. See, for example, Bowles and Levin (1968), Cain and Watts (1970), and Hanushek and Kain (1972).) And, as discussed below, the common interpretations of the results of the Coleman Report are now recognized to be quite incorrect.

The underlying model that evolved as a result of this research is very straightforward. The output of the educational process – that is, the achievement of individual students – is directly related to a series of inputs. Some of these inputs – the characteristics of schools, teachers, curricula, and so forth – are directly controlled by policy makers. Other inputs – those of families and friends plus the innate endowments or learning capacities of the students – are generally not controlled. Further, while achievement may be measured at discrete points in time, the educational process is cumulative; inputs applied sometime in the past affect students' current levels of achievement.

Family background is usually characterized by such sociodemographic characteristics as parental education, income, and family size. Peer inputs, when included, are typically totals of a student population's sociodemographic characteristics for a school or classroom. School inputs include teacher background (education level, experience, sex, race, and so forth), school organization (class sizes, facilities, administrative expenditures, and so forth), and district or community factors (e.g., average expenditure levels). As discussed elsewhere, one set of analytical problems flows from the lack of measurement of innate abilities of individuals and the imprecise measurement of the history of educational inputs (Hanushek, 1979, 1986). A second set of concerns that has received extensive recent attention relates to applying causal interpretations to the estimates. Both are discussed below.

In many ways, the analysis of educational production breaks into two distinct periods. The first – running from the Coleman Report in the mid-1960s through the mid-1990s – largely relied upon relatively small and specialized data sets that contained information on student outcomes and on some set of measures of school resources. The second period – running from the mid-1990s until now – largely moved to administrative databases with longitudinal information on student achievement for entire states over time. It took the results of the early period research that indicated no consistent relationship between measured school inputs and outcomes as a starting point (see the next section) and turned to intensive investigations of teacher quality and other specific aspects of schools.

The Importance of Measured School Inputs

The state of knowledge about the impacts of resources is best summarized by reviews of available empirical studies. Most traditional analyses of education production functions directed their attention at a relatively small set of resource measures, and this makes it easy to summarize the results (Hanushek, 2003).

Table 1 provides a tabulation of estimates of the impact of school inputs for studies that appeared before 1995. (Such reviews of empirical studies appeared first in 1981 (Hanushek, 1981) and were updated over time with additional studies. With stable results to the overall pattern of findings, there has been no subsequent updating of research past 1995, and the general pattern of results has become widely accepted.) The 90 individual publications for this analysis contain 377 separate production function estimates and form the basis of the conclusions about the impacts of measured inputs. The table indicates the sign of the estimated impact of a given resource on student performance; positive estimates imply that more resources are associated with

Table 1 Percentage distribution of estimated effect of key resources on student performance, based on 377 studies available through 1995

Resources	Number of estimates	Statistically significant		Statistically insignificant
		Positive	Negative	
Real classroom resources				
Teacher-pupil ratio	277	14%	14%	72%
Teacher education	171	9	5	86
Teacher experience	207	29	5	66
Financial aggregates				
Teacher salary	119	20%	7%	25%
Expenditure per pupil	163	27	7	34
Other				
Facilities	91	9	5	86
Administration	75	12	5	83
Teacher test scores	41	37	10	53

Source: Hanushek, Eric, A., Summer, 1997. Assessing the effects of school resources on student performance: an update. Educational Evaluation and Policy Analysis 19 (2), 141–164, revised.

higher student outcomes and the opposite for negative estimates. Additionally, information is provided about statistical significance or our confidence that there is a real effect of the resources.

For classroom resources, only 9% of studies on teacher education and 14% of studies on teacher-pupil ratios found a positive and statistically significant relationship between these factors and student performance. (This summary concentrates on the results. Details of the underlying analyses and of the selection of studies for this tabulation are found in Hanushek (1986, 1997). The first column provides a count of the total number of estimates addressing the impact of the given resource.) Moreover, these studies were offset by another set of studies that found a similarly negative correlation between those inputs and student achievement. Twenty-nine percent of the studies found a positive correlation between teacher experience and student performance; however, 71% still provided no support for increasing teacher experience (being either negative or statistically insignificant).

Studies on the effect of financial resources provide a similar picture. There is very weak support for the notion that simply providing higher teacher salaries or greater overall spending will lead to improved student performance. Per pupil expenditure has received the most attention, but only 27% of studies showed a positive and significant effect. In fact, 7% even suggested that adding resources would harm student achievement. It is also important to note that studies involving pupil spending have tended to be the lowest-quality studies, and thus there is substantial reason to believe that even the 27% figure overstates the true effect of added expenditure (Hanushek et al., 1996; Hanushek, 2003).

While the general conclusions about the inconsistent impacts of school resources were intensely debated through the 1980s and 1990s, the general conclusions have now been broadly accepted, and the research has moved to other areas. (Some idea about the debates can be seen in the various chapters in Burtless (1996). The controversy did continue with more specific debates about the effects of class size reduction, see Hanushek (1999), Krueger (1999), and Mishel and Rothstein (2002), although that too has largely disappeared.) Specifically, both the research and policy discussion now recognizes that 'how' resources are used is generally much more important than simply 'how much' is available.

On the research side, as suggested, the previous discussions do not distinguish among studies on the basis of any quality differences, so the results could be distorted by not adjusting the tabulations. While 'study quality' has a variety of subjective components, the available estimates can be separated by a few objective components of quality. First, while education is cumulative, frequently only current input measures are available, which results in analytical errors. Second, schools operate within a policy environment set almost always at higher levels of government. These reasons, combined with large panel data sets on student outcomes that come from state accountability data have led almost all studies in the second phase of research to focus on gains in student learning over time and on samples of individual students and classrooms rather than aggregate data on schools, districts, or states. (While these issues of estimation had been known for a considerable time (see Hanushek (1979)), the broad empirical application did not come until administrative databases became more readily available.)

There has also been a proliferation across the globe of studies of educational performance. One surprising result has been that many of the conclusions hold equally for developing countries, where resource constraints would seem to be much more important (Hanushek, 1995; Glewwe et al., 2013). Moreover, there has been a move toward using more evidence in policy decisions, led in part by the international assessments of evidence by John Hattie (2009).

Do Teachers and Schools Matter?

Because of the Coleman Report and subsequent studies discussed above, many people tended to argue that schools do not matter and that only families and peers affect performance. Unfortunately, these interpretations have confused measurability with true effects.

Extensive research over the past 45 years has made it clear that teachers do indeed matter. The simple definition of teacher quality used here, as first developed in Hanushek (1971), is an output-based measure. In this work, teacher effectiveness is assessed directly from student performance, instead of the more typical input measures based on characteristics of the teacher and school. High-quality teachers are ones who consistently obtain higher than expected gains in student performance, while low-quality teachers are ones who consistently obtain lower than expected gains. (In statistical terms these estimates employ covariance models, where fixed effects are estimated for individual teachers.) When this approach has been used, large variations in performance have been uncovered.

In fact, the degree to which teacher differences affect performance is impressive. Looking at the range of quality for teachers within a single large urban district, teachers near the top of the quality distribution can get an entire year's worth of additional learning out of their students, compared to those near the bottom (Hanushek, 1992). In subsequent work,

Rivkin et al. (2005) provide another estimate of the importance of teachers: having a good teacher instead of an average teacher for 3–5 years in a row can lead to achievement gains that equal the average difference in achievement of a student from a poor family as opposed to a better off family.

This work, now commonly called 'value-added analysis' because of the focus on what individual teachers contribute to student learning, has produced quite consistent estimates of how much variation there is in the impact of teachers (Hanushek and Rivkin, 2010). And, it has become one of the largest and most vibrant areas of educational research. Numerous recent investigations of teacher quality have focused on issues such as the best statistical approaches to employ and the interpretation of estimated differences across classrooms when students might not be randomly assigned to individual teachers. (For a discussion of these issues and the competing views, see Hanushek and Rivkin (2012).)

Part of the interest in these estimation issues has been raised by the application of estimates of teacher value added to policy. For example, in 2011, the Los Angeles Times newspaper published value-added ratings for individual teachers (Song and Felch, 2011), and this was followed by similar publications in New York City and elsewhere. Additionally, the teachers in the Washington, DC, schools have been partially evaluated on the basis of such value-added calculations (Isenberg and Hock, 2010). Moreover, early evidence suggests that use of these teacher evaluations has a positive impact on student outcomes (Dee and Wyckoff, 2013). These discussions have elevated the research attention to estimating the impacts of teachers and principals. (Investigations of principals are more difficult from a statistical viewpoint and have only recently been developed; see Branch et al. (2012).)

Based on existing value-added measures, the economic impact of teacher quality appears to be very large. **Figure 2** provides estimates that combine information about the impact on achievement of teachers at different points in the quality distribution with estimates of the impact of achievement on a student's future income (Hanushek, 2011). The figure plots the gains (or losses) in future income from a class of students having a high- (low-) quality teacher, all compared to an average teacher. Class size directly affects these estimates because a given teacher has a greater impact on outcomes when there are more students in the class. From this figure, a teacher at the 75th percentile of the quality distribution with a class of 25 students produces over $300 000 more future income compared with an average teacher. A teacher at the 25th percentile produces a similar loss compared to the average teacher.

These estimates, produced from combining the results of many existing studies, have been confirmed by an innovative study that links student outcomes for individual teachers with their actual labor market earnings observed through tax records (Chetty et al., 2014). They find individual teachers in the bottom 5% can reduce lifetime earnings of their students each year that they teach by over $250 000.

Efficiency

Efficiency involves the relationship between inputs and outputs in a production process. The underlying notion is that production is efficient if given inputs produce the maximum output. The simplicity of this statement, however, obscures a variety of complexities that arise when the concept is actually applied.

Typically economics does not devote much attention to the analysis of efficiency. If there are competitive markets, the behavior of producers and consumers tends to drive outcomes toward efficient production. Because of this basic theorem, it is common simply to assume efficiency.

In education, however, efficiency is never a given. First, education is generally publicly provided. Governmental organizations, which do not face the same incentives as private firms, cannot be expected to move toward efficient production. In particular, few school personnel are rewarded

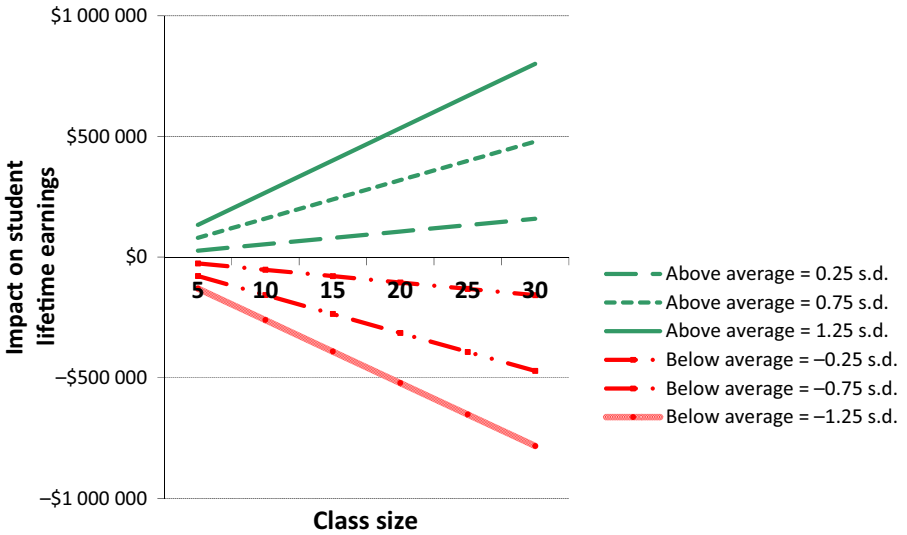

Figure 2 Impact on student lifetime incomes by class size and teacher effectiveness (compared to average teacher).

or punished based on the outcomes students obtain, making the incentives for performance minimal or nonexistent. Second, it is difficult to find information on school efficiency. With complete information, parents might be expected to pressure schools to use resources – either through local political processes or through moving to districts that did better (Tiebout, 1956). Because student performance is influenced by factors schools cannot control – i.e., input from family and friends – simply observing student achievement does not accurately reflect school contribution. Thus, parental and voter pressures on schools are lessened by imperfect information.

The underlying political economy of schools that can preserve such apparently large inefficiencies in resource use deserves considerably more research.

Competition and Incentives

The lack of direct incentives within schools has led to investigations of circumstances where incentive forces may be stronger. While there have been a wide variety of proposed changes (e.g., Hanushek et al. (1994)), most have not been implemented or analyzed sufficiently to draw any conclusions about their efficacy.

The classic argument for competition comes from Friedman (1962) with his arguments for vouchers. The well-known arguments suggest that separating the finance of schooling from the production of schooling by allowing students to choose the school they attend would improve individual satisfaction with the outcomes (Nechyba, 2006). Within the United States, this approach has been vigorously debated, but few examples of implementation have occurred. The most celebrated application has been Milwaukee, Wisconsin, where vouchers have been available to a (constrained) number of poor children since 1990 (Rouse, 1998; Carnoy et al., 2006), but they have also been available in Cleveland, Washington, DC, and elsewhere (Wolf et al., 2010). A number of privately financed alternatives have also been offered (Howell and Peterson, 2002). Most of the attention to these voucher programs has centered on the student outcomes of students in them as compared to public schools. (A variety of controversies has developed in these analyses. Perhaps the primary analytical issue is dealing with selection of students and an appropriate comparison group. See, for example, the discussions in Witte (1999), Rouse (1998), and Greene et al. (1998).) Because these programs have been very marginal to the education system, there has been little suggestion that the public schools have made any adjustments in response. Thus, these choice experiences have not provided information about how public schools might react to a larger, more institutionalized program.

Private schools offer one possibility for better understanding the effects of competition. They must compete with the public schools in order to attract clients and thus they are subject to stronger pressures to provide high levels of performance. The literature on Catholic school performance is summarized in the articles by Neal (1997) and Grogger and Neal (2000). The evidence has generally indicated that Catholic schools on average outperform public. This superiority seems clearest in urban settings, where disadvantaged students face fewer options than others. This evidence is, nonetheless, subject to some caveats. First, as recognized since some of the earliest work on the topic (Coleman et al., 1982), it is difficult to separate performance of the private schools from pure selection phenomena. Specifically, since private school students could have attended public schools but instead pay extra for private schooling, they are clearly different than the public school students with identical measured characteristics. A variety of alternative approaches has been taken to deal with the selection problem, and a rough summary of the results after those efforts is that there remains a small advantage from attending Catholic schools. (Grogger and Neal (2000) suggest, however, that there is no advantage to attending private elite schools – a surprising result given the extra cost generally involved in that. These results are possibly due to selection problems, but it is not a simple relationship because most people would expect positive selection into these elite schools.) Second, this literature says little about the distribution of school quality within the Catholic sector.

Within the public sector, it is widely believed that school quality varies considerably across schools, and one might suspect that the same holds true for private schools.

One important potential element competition comes from other public school jurisdictions. Specifically, households can choose the specific jurisdiction and school district, in the manner of Tiebout (1956), by their choice of residential location. While adjustment is costly, these choices permit individuals to seek high-quality schools if they wish (Epple and Nechyba, 2004). Residential location decisions are of course complicated, involving job locations, availability of various kinds of housing, school costs and quality, and availability of other governmental services. Given choice opportunities plus voting responses, this model suggests pressure on schools to use resources effectively. Otherwise one might expect housing values to be affected.

The simple choice model would suggest naturally that larger numbers of schools or school districts per student would offer more opportunities for residents and thus more competition across schools. This simple model motivates the empirical analyses of Borland and Howsen (1992) and its extension and refinement in Hoxby (2000) and in Hanushek and Rivkin (2003). The simple inclusion of measures of concentration indicates that areas with less choice have poorer schools on average (Borland and Howsen, 1992). Hoxby (2000) pursues alternative strategies to look at the causal impact of concentrations and finds a larger estimated impact of competition on the performance of schools, although the robustness of her conclusions has been challenged by Rothstein (2007). Hanushek and Rivkin (2003), pursuing an alternative strategy, find competitive impacts on performance of urban schools and that these are particularly important for low-income students who have fewer mobility options.

In the United States, the rise of charter schools has provided an additional element of choice. Charter schools are public schools that have the freedom to be subject to less regulation and to offer different programs from the traditional public schools. They are schools of choice and do not charge any tuition fee. In 2011, there were over 5000 charter schools in the

United States, serving 3.6% of the school population (U.S. Department of Education, 2013: Table 117). The analysis of their performance shows considerable variation in effectiveness – ranging from much better than corresponding public schools to much worse (see, e.g., Hanushek et al. (2007), CREDO (2009, 2010, 2013), Abdulkadiroğlu et al. (2011)).

Finance of Schools

Public finance economists have devoted considerable energy to the study of the finance of schools, largely from the viewpoint of traditional tax and expenditure policy. The educational policy component of this involves equity in the provision of education. These considerations are largely motivated by the many studies that show wide disparities in achievement by race and socioeconomic background (e.g., Coleman et al. (1966), Hanushek (2001)). Nonetheless, the typical discussion of equity concentrates primarily on spending disparities across schools and districts, even in the face of evidence suggesting that student outcomes are not closely related to spending and resources (Hanushek, 2003).

Indeed, one of the most significant policy issues of the past 30 years in the United States has been the appropriate way for states to fund local schools (Fischel, 2006). While most states have employed a compensatory aid formula to ameliorate some local governments' difficulty in raising taxes, these measures have only partially solved the problem. The issue became the subject of court action in the 1960s, with the case of Serrano versus Priest in California. Suits in many other states followed (Hanushek and Lindseth, 2009). As a result, some significant narrowing of spending variations occurred (Murray et al., 1998).

However, there has been surprisingly little analysis of the impacts of these suits. This is important to undertake, because the prior research discussed above does not show a clear link between increased funding and school quality or student performance. From the perspective of educational outcomes, then, the value of spending changes related to equalization court cases and the related legislative actions remain an open question (Hanushek, 1991; Hanushek and Lindseth, 2009). To date, little research has focused on the achievement impacts of school finance policies and those that have looked at outcomes have not found much impact (e.g., Downes (1992), Cullen and Loeb (2004), Greene and Trivitt (2008)). Hanushek and Lindseth (2009) investigate the achievement impact of major court decisions in Kentucky, Massachusetts, New Jersey, and Wyoming. With the possible exception of Massachusetts, there is little evidence of any impact of major court spending decisions on student achievement. A broader, nation-wide analysis of district spending on earnings outcomes also found no beneficial relationship between the two, except perhaps for Black females (Hanushek and Somers, 2001).

Some Conclusions and Implications

The existing research suggests inefficiency in the provision of schooling. It does not indicate that schools do not matter. Nor does it indicate that money and resources never affect achievement. The accumulated research simply says there is no clear, systematic relationship between resources and student outcomes.

The persistence of this raises important questions about current education policy, as well as a real need for further research.

The main conclusion of this research is that policy decisions should not focus on school resources, because the impact of resources on student achievement is unknown at this time. The policy solution seems to be to establish incentives – rewards or consequences related to student outcomes – and then to permit local schools to develop their own game plans for meeting these goals. While a wide variety of incentives have been proposed – such as vouchers, merit pay, or contracting out – most have not been implemented or analyzed sufficiently to determine their efficacy.

A final major issue raised by these analyses is their implication for other kinds of studies (Hanushek, 1996). Most studies involving human capital are interested in its effect on other aspects of behavior. For these studies schooling is largely tangential to other interests, and researchers are typically simply looking for an easy and readily available measure of school quality. The most common has been spending per pupil. However, the inefficient production of human capital introduces natural measurement problems. Direct spending is no longer a good measure of quality because it has no perceivable bearing on performance. Second, it is well known that families have considerable influence on student achievement, implying that school resources are only part of the equation. Both factors suggest studies that measuring student achievement just through resource investment has a high potential for distortion.

See also: Class Size; Earnings Inequality; Education and Economic Growth; Education and Employment; Education and Income Distribution; Education: Skills Training; Educational Effectiveness, The Field of; Educational Productivity; Educational Systems: North America; Equity and Education; Family and Schooling; School Choice; School Effectiveness Research; Schooling: Total Impact of; Social Inequality and Schooling.

Bibliography

Abdulkadiroğlu, Atila, Angrist, Joshua D., Dynarski, Susan M., Kane, Thomas J., Pathak, Parag A., 2011. Accountability and flexibility in public schools: evidence from Boston's charters and pilots. Quarterly Journal of Economics 126 (2), 699–748.

Becker, Gary S., 1964. Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. National Bureau of Economic Research, New York.

Bettinger, Eric, Long, Bridget T., Oreopoulos, Philip, Sanbonmatsu, Lisa, 2009. The Role of Simplification and Information in College Decisions: Results from the H&R Block FAFSA Experiment. NBER Working Paper 15361. National Bureau of Economic Research, Cambridge, MA.

Bettinger, Eric P., Long, Bridget T., 2010. Does cheaper mean better? The impact of using adjunct instructors on student outcomes. Review of Economics and Statistics 92 (3), 598–613.

Borland, Melvin V., Howsen, Roy M., March 1992. Student academic achievement and the degree of market concentration in education. Economics of Education Review 11 (1), 31–39.

Bound, John, Turner, Sarah, June 2007. Cohort crowding: how resources affect collegiate attainment. Journal of Public Economics 91 (5–6), 877–899.

Bowles, Samuel, Levin, Henry M., Winter 1968. The determinants of scholastic achievement – an appraisal of some recent evidence. Journal of Human Resources 3 (1), 3–24.

Branch, Gregory F., Hanushek, Eric A., Rivkin, Steven G., January 2012. Estimating the Effect of Leaders on Public Sector Productivity: The Case of School Principals. NBER Working Paper W17803. National Bureau of Economic Research, Cambridge, MA.

Burtless, Gary (Ed.), 1996. Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success. Brookings, Washington, DC.

Cain, Glen G., Watts, Harold W., April 1970. Problems in making policy inferences from the Coleman Report. American Sociological Review 35 (2), 328–352.

Carnoy, Martin, Adamson, Frank, Chudgar, Amita, Luschei, Thomas F., Witte, John F., 2006. Vouchers and Public School Performance: A Case Study of the Milwaukee Parental Choice Program. Economic Policy Institute, Washington, DC.

Chetty, Raj, Friedman, John N., Rockoff, Jonah, 2014. Measuring the impacts of teachers II: teacher value-added and the student ourcomes in adulthood. American Economic Review.

Coleman, James S., Campbell, Ernest Q., Hobson, Carol J., McPartland, James, Mood, Alexander M., Weinfeld, Frederic D., York, Robert L., 1966. Equality of Educational Opportunity. U.S. Government Printing Office, Washington, DC.

Coleman, James S., Hoffer, Thomas, Kilgore, Sally, 1982. High School Achievement: Public, Catholic and Private Schools Compared. Basic Books, NY.

CREDO, 2009. Multiple Choice: Charter School Performance in 16 States. Center for Research on Education Outcomes, Stanford University, Stanford, CA.

CREDO, 2010. Charter School Performance in New York City. Center for Research on Education Outcomes, Stanford University, Stanford, CA.

CREDO, 2013. National Charter School Study 2013. Center for Research on Education Outcomes, Stanford University, Stanford, CA.

Cullen, Julie B., Loeb, Susanna, 2004. School finance reform in Michigan: evaluating proposal A. In: Yinger, John (Ed.), Helping Children Left behind: State Aid and the Pursuit of Educational Equity. Massachusetts Institute of Technology Press, Cambridge, MA, pp. 215–249.

Dee, Thomas, Wyckoff, James, October 2013. Incentives, Selection, and Teacher Performance: Evidence from IMPACT. NBER Working Paper 19529. National Bureau of Economic Research, Cambridge, MA.

Downes, Thomas A., December 1992. Evaluating the impact of school finance reform on the provision of public education: the California case. National Tax Journal 45 (4), 405–419.

Epple, Dennis, Nechyba, Thomas, 2004. Fiscal decentralization. In: Vernon Henderson, J., Thisse, Jacques-François (Eds.), Handbook of Regional and Urban Economics. Elsevier, Amsterdam, pp. 2423–2480.

Fischel, William A., 2006. The courts and public school finance: judge-made centralization and economic research. In: Hanushek, Eric A., Welch, Finis (Eds.), Handbook of the Economics of Education. North Holland, Amsterdam, pp. 1277–1325.

Friedman, Milton, 1962. Capitalism and Freedom. University of Chicago Press, Chicago.

Glewwe, Paul, Hanushek, Eric A., Humpage, Sarah D., Ravina, Renato, 2013. School resources and educational outcomes in developing countries: a review of the literature from 1990 to 2010. In: Glewwe, Paul (Ed.), Education Policy in Developing Countries. University of Chicago Press, Chicago, pp. 13–64.

Greene, Jay P., Peterson, Paul E., Du, Jiangtu, 1998. School choice in Milwaukee: a randomized experiment. In: Peterson, Paul E., Hassel, Bryan C. (Eds.), Learning from School Choice. Brookings Institution, Washington, DC, pp. 335–356.

Greene, Jay P., Trivitt, Julie R., April 2008. Can judges improve academic achievement? Peabody Journal of Education 83 (2), 224–237.

Grogger, Jeffrey, Neal, Derek, 2000. Further evidence on the effects of Catholic secondary schooling. In: Gale, William G., Pack, Janet Rothenberg (Eds.), Brookings-Wharton Papers on Urban Affairs. Brookings Institution Press, Washington, DC, pp. 151–193.

Hanushek, Eric A., May 1971. Teacher characteristics and gains in student achievement: estimation using micro data. American Economic Review 60 (2), 280–288.

Hanushek, Eric A., Summer 1979. Conceptual and empirical issues in the estimation of educational production functions. Journal of Human Resources 14 (3), 351–388.

Hanushek, Eric A., Fall 1981. Throwing money at schools. Journal of Policy Analysis and Management 1 (1), 19–41.

Hanushek, Eric A., September 1986. The economics of schooling: production and efficiency in public schools. Journal of Economic Literature 24 (3), 1141–1177.

Hanushek, Eric A., Summer 1991. When school finance 'reform' may not be good policy. Harvard Journal on Legislation 28 (2), 423–456.

Hanushek, Eric A., February 1992. The trade-off between child quantity and quality. Journal of Political Economy 100 (1), 84–117.

Hanushek, Eric A., August 1995. "Interpreting recent research on schooling in developing countries. World Bank Research Observer 10 (2), 227–246.

Hanushek, Eric A., Fall 1996. Measuring investment in education. Journal of Economic Perspectives 10 (4), 9–30.

Hanushek, Eric A., Summer 1997. Assessing the effects of school resources on student performance: an update. Educational Evaluation and Policy Analysis 19 (2), 141–164.

Hanushek, Eric A., 1999. The evidence on class size. In: Mayer, Susan E., Peterson, Paul E. (Eds.), Earning and Learning: How Schools Matter. Brookings Institution, Washington, DC, pp. 131–168.

Hanushek, Eric A., May 2001. Black-white achievement differences and governmental interventions. American Economic Review 91 (2), 24–28.

Hanushek, Eric A., February 2003. The failure of input-based schooling policies. Economic Journal 113 (485), F64–F98.

Hanushek, Eric A., June 2011. The economic value of higher teacher quality. Economics of Education Review 30 (3), 466–479.

Hanushek, Eric A., Kain, John F., 1972. On the value of 'equality of educational opportunity' as a guide to public policy. In: Mosteller, Frederick, Moynihan, Daniel P. (Eds.), On Equality of Educational Opportunity. Random House, New York, pp. 116–145.

Hanushek, Eric A., Kain, John F., Rivkin, Steve G., Branch, Gregory F., June 2007. Charter school quality and parental decision making with school choice. Journal of Public Economics 91 (5–6), 823–848.

Hanushek, Eric A., Kimko, Dennis D., December 2000. Schooling, labor force quality, and the growth of nations. American Economic Review 90 (5), 1184–1208.

Hanushek, Eric A., Lindseth, Alfred A., 2009. Schoolhouses, Courthouses, and Statehouses: Solving the Funding-achievement Puzzle in America's Public Schools. Princeton University Press, Princeton, NJ.

Hanushek, Eric A., Machin, Stephen, Woessmann, Ludger (Eds.), 2011. Handbook of the Economics of Education, vols. 3 and 4. North Holland, Amsterdam.

Hanushek, Eric A., et al., 1994. Making Schools Work: Improving Performance and Controlling Costs. Brookings Institution, Washington, DC.

Hanushek, Eric A., Rivkin, Steven G., 2003. Does public school competition affect teacher quality? In: Hoxby, Caroline M. (Ed.), The Economics of School Choice. University of Chicago Press, Chicago, IL, pp. 23–47.

Hanushek, Eric A., Rivkin, Steven G., May 2010. Generalizations about using value-added measures of teacher quality. American Economic Review 100 (2), 267–271.

Hanushek, Eric A., Rivkin, Steven G., 2012. The distribution of teacher quality and implications for policy. Annual Review of Economics 4, 131–157.

Hanushek, Eric A., Rivkin, Steven G., Taylor, Lori L., November 1996. Aggregation and the estimated effects of school resources. Review of Economics and Statistics 78 (4), 611–627.

Hanushek, Eric A., Somers, Julie A., 2001. Schooling, inequality, and the impact of government. In: Welch, Finis (Ed.), The Causes and Consequences of Increasing Inequality. University of Chicago Press, Chicago, pp. 169–199.

Hanushek, Eric A., Welch, Finis (Eds.), 2006. Handbook of the Economics of Education, vols. 1 and 2. North Holland, Amsterdam.

Hanushek, Eric A., Woessmann, Ludger, September 2008. The role of cognitive skills in economic development. Journal of Economic Literature 46 (3), 607–668.

Hanushek, Eric A., Woessmann, Ludger, July 2011. How much do educational outcomes matter in OECD countries? Economic Policy 26 (67), 427–491.

Hanushek, Eric A., Woessmann, Ludger, December 2012. Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. Journal of Economic Growth 17 (4), 267–321.

Hanushek, Eric A., Zhang, Lei, Summer 2009. Quality-consistent estimates of international schooling and skill gradients. Journal of Human Capital 3 (2), 107–143.

Hattie, John, 2009. Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement. Routledge, London.

Howell, William G., Peterson, Paul E., 2002. The Education Gap: Vouchers and Urban Schools. Brookings, Washington, DC.

Hoxby, Caroline M., December 2000. Does competition among public schools benefit students and taxpayers? American Economic Review 90 (5), 1209–1238.
Isenberg, Eric, Hock, Heinrich, August 20 2010. Measuring School and Teacher Value Added for IMPACT and TEAM in DC Public Schools: Final Report. Mathematica Policy Research, Princeton, NJ.
Kane, Thomas J., 2006. Public intervention in post-secondary education. In: Hanushek, Eric A., Welch, Finis (Eds.), Handbook of the Economics of Education. North Holland, Amsterdam, pp. 1369–1402.
Kiker, B.F., 1968. Human Capital: In Retrospect. University of South Carolina, Columbia, SC.
Krueger, Alan B., May 1999. Experimental estimates of education production functions. Quarterly Journal of Economics 114 (2), 497–532.
Lazear, Edward P., 2003. Teacher incentives. Swedish Economic Policy Review 10 (3), 179–214.
McPherson, Michael S., Schapiro, Morton, 2006. US higher education finance. In: Hanushek, Eric A., Welch, Finis (Eds.), Handbook of the Economics of Education. North Holland, Amsterdam, pp. 1403–1434.
Mincer, Jacob, March 1970. The distribution of labor incomes: a survey with special reference to the human capital approach. Journal of Economic Literature 8 (1), 1–26.
Mincer, Jacob, 1974. Schooling, Experience, and Earnings. NBER, New York.
Mishel, Lawrence, Rothstein, Richard (Eds.), 2002. The Class Size Debate. Economic Policy Institute, Washington, DC.
Mulligan, Casey B., December 1999. Galton versus the human capital approach to inheritance. Journal of Political Economy 107 (6 pt. 2), S184–S224.
Murnane, Richard J., Willett, John B., Duhaldeborde, Yves, Tyler, John H., Fall 2000. How important are the cognitive skills of teenagers in predicting subsequent earnings? Journal of Policy Analysis and Management 19 (4), 547–568.
Murray, Sheila E., Evans, William N., Schwab, Robert M., September 1998. Education-finance reform and the distribution of education resources. American Economic Review 88 (4), 789–812.
Neal, Derek, January 1997. The effect of Catholic secondary schooling on educational attainment. Journal of Labor Economics 15 (1 Part 1), 98–123.
Nechyba, Thomas J., 2006. Income and peer sorting in public and private schools. In: Hanushek, Eric A., Welch, Finis (Eds.), Handbook of the Economics of Education. North Holland, Amsterdam, pp. 1327–1368.
Nelson, Richard R., Phelps, Edmund, May 1966. Investment in humans, technology diffusion and economic growth. American Economic Review 56 (2), 69–75.
Organisation for Economic Co-operation and Development, 2008. Education at a Glance 2008: OECD Indicators. OECD, Paris.
Psacharopoulos, George, Patrinos, Harry A., August 2004. Returns to investment in education: a further update. Education Economics 12 (2), 111–134.
Rivkin, Steven G., Hanushek, Eric A., Kain, John F., March 2005. Teachers, schools, and academic achievement. Econometrica 73 (2), 417–458.
Rothstein, Jesse, December 2007. Does competition among public schools benefit students and taxpayers?. Comment The American Economic Review 97 (5), 2026–2037.
Rouse, Cecilia Elena, May 1998. Private school vouchers and student achievement: an evaluation of the Milwaukee Parental Choice Program. Quarterly Journal of Economics 113 (2), 553–602.
Schultz, Theodore W., March 1961. Investment in human capital. American Economic Review 51 (1), 1–17.
Song, Jason, Felch, Jason, May 7, 2011. Times Updates and Expands Value-Added Ratings for Los Angeles Elementary School Teachers. Los Angeles TIme s.
Tiebout, Charles M., October 1956. A pure theory of local expenditures. Journal of Political Economy 64 (5), 416–424.
U.S. Department of Education, 2013. Digest of Education Statistics, 2012. National Center for Education Statistics, Washington, DC.
Welch, Finis, January/February 1970. Education in production. Journal of Political Economy 78 (1), 35–59.
Witte, John F., 1999. The Market Approach to Education. Princeton University Press, Princeton, NJ.
Wolf, Patrick, Gutmann, Babette, Puma, Michael, Kisida, Brian, Rizzo, Lou, Eissa, Nada, Carr, Matthew, June 2010. Evaluation of the DC Opportunity Scholarship Program: Final Report. Institute for Education Sciences, Washington, DC.

Education for the Gifted and Talented

David W Chan, The Chinese University of Hong Kong, Hong Kong, People's Republic of China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Gifted and talented children and youth, because of their precocity, intensity, and complexity, require differentiation in curriculum and teaching to enable them to develop to their full potential. They need and benefit from accelerated and enriched learning experiences commensurate with their levels of giftedness and talent. Such gifted programming can be represented by program models that focus on academic talents and acceleration (e.g., Talent Search Model), and those that focus on creative productivity and enrichment (e.g., Schoolwide Enrichment Model). Gifted learners may also experience social and personal issues that require support and counseling attention from teachers and parents.

Giftedness and Talent

'Gifted' vs 'Talented'

Education for the gifted and talented is commonly referred to as gifted education and/or talent development. The terms 'gifted' and 'talented' are almost always used synonymously to describe children and youth functioning with superior cognitive abilities and performing at levels far above what is typical for their chronological age. Used interchangeably, a child can be described as 'verbally gifted' and 'mathematically gifted' or 'verbally talented' and 'mathematically talented.' Some researchers and practitioners however have made a distinction. They refer 'gifted' children to those children with generally broad and high abilities across different domains (e.g., intellectually gifted or creatively gifted), and 'talented' children to those children with special abilities in specific domains (e.g., artistic talent as in musically talented). Others describe children as 'gifted' when referring to academic disciplines or domains, and 'talented' when referring to nonacademic domains. Still others see talent and giftedness on a continuum specifying levels of giftedness, with giftedness at the upper or high end. Perhaps, the most salient distinction is made by Francoys Gagne, who defines 'gifts' as untrained natural abilities or natural aptitudes, and 'talents' as learned capabilities or specific skills in his Differentiated Model of Giftedness and Talent. In general, there are no solid empirical grounds for making the distinction other than common usage, and the terms are best interpreted as interchangeable.

Defining Giftedness

Regardless of whether one distinguishes between 'gifted' and 'talented,' 'giftedness' as a construct has been defined in many different ways throughout its history (see Sternberg and Davidson, 2005). Lewis Terman is probably one of the first to define giftedness in terms of high IQ. In his longitudinal study of the physical, psychological, social, and professional development of 1528 gifted children with IQs averaging 151, his inclusion criterion was an IQ above 135. Since Terman's time, researchers and practitioners have developed a consensus that giftedness is often not captured by a single unidimensional IQ score. For example, the 1972 US Office of Education Report (Marland Report) defined giftedness as superiority in six categories that include general intellectual ability, specific academic aptitude, creative or productive thinking, leadership ability, visual and performing arts, and psychomotor ability. In subsequent revisions, psychomotor ability was left out, but the view that high IQ is but one aspect of giftedness remains the clear message. A similar view could be discerned in Joseph Renzulli's (1978) three-ring model, where giftedness is defined by the confluence of high-task commitment and high creativity in addition to above-average ability.

From a slightly different perspective, other researchers and practitioners have broadened giftedness to correspond to the changes in the conception of intelligence as multidimensional. For example, Howard Gardner redefined intelligence and proposed multiple intelligences that can characterize gifted individuals, suggesting that each individual may excel or be gifted in one or more intelligences. His multiple intelligences include verbal-linguistic, logical-mathematical, visual-spatial, musical, bodily-kinesthetic, intrapersonal, interpersonal, and naturalist intelligences. Emphasizing cognitive processes rather than content domains, Robert Sternberg (1985) proposed three major kinds of intelligence or triarchic abilities that include analytic (academic), synthetic (creative), and practical (applied) abilities, thus reflecting three major kinds of giftedness. Nonetheless, irrespective of how giftedness is conceptualized, the adoption of a particular definition of giftedness has specific implications for educators in their identification of gifted learners for talent recognition and development.

Education for the Gifted and Talent Development

Despite the diversity in giftedness theorizing, educators in gifted education generally conceptualize giftedness simply as unusually high ability in any area, including high global IQ as well as domain-specific ability. Given the multiple manifestations of giftedness, they tend to strive to be inclusive in identifying gifted and talented children and youth through the adoption of clearly defined but broad conceptions of giftedness and the use of multiple alternative criteria with assessment over time. It is also understood that giftedness may appear in different forms in different ethnic or socioeconomic groups and in different cultural settings. Partly because of the variability in the conceptions of giftedness and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92137-8

the diversity in the identification procedures and partly because of the equity issue in educational provisions, some researchers and practitioners have called for the shifting of the focus of gifted education to talent development. Accordingly, talent development emphasizes developing the talents, gifts, and capabilities of all students, including highly able or high-end learners. Thus, instead of asking whether a child is gifted or talented, the question becomes one of asking how gifted or talented the child is, and how one could help to bring forth the giftedness or talent in the child. Parenthetically, this orientation also eliminates the awkwardness of calling some children 'gifted' and others 'ungifted,' 'nongifted,' or 'not gifted.'

Differentiation in Gifted Education and Talent Development

For educators with a talent development focus, gifted education is very similar to quality education that emphasizes addressing the specific learning needs of individual students, including gifted learners, while allowing students to explore and master high-quality curriculum as well as developing their potentials and talents. For educators with a focus on educating gifted learners, gifted education implies differentiation or adapting curriculum and instruction for gifted learners who are characterized by their capacity to learn faster, remember more, find and solve problems more readily, manipulate abstract ideas and make connections more readily, process information more effectively, and generate more new and unusual ideas than their age peers. Although the concept of differentiation has been more recently used to denote teaching a range of diverse learners, including but not limited to gifted learners, with emphasis on the general education setting, differentiation in the context of gifted education is useful for considerations in areas of meeting gifted learners' needs and characteristics, adapting curriculum and instruction, and teacher training and education.

Characteristics of Gifted Learners

Understanding the specific characteristics and needs of gifted learners could be considered the prelude to efforts of differentiation, and these specific needs and characteristics provide grounds on justification for ability grouping of gifted learners. In this regard, Joyce VanTassel-Baska and Tamra Stambaugh (2006) named three characteristics as most relevant for gifted and talented curriculum planning. They are precocity, intensity, and complexity. Precocity is often the key characteristics, as gifted learners admitted to gifted programs are often at least 2 years ahead in their talented (e.g., language and mathematics) areas, and are powerfully motivated to learn fast. Intensity may be manifested affectively in gifted learners' emotional responsiveness or sensitivity. Equally and even more relevant for learning are the cognitive manifestations in gifted learners' capacity to focus and concentrate for long periods of time on an idea or a subject that fascinates them. Ellen Winner (1996) called this characteristic 'a rage to master' to emphasize the passion and strong motivation of gifted learners. Complexity refers to gifted learners' capacity to engage in abstract or higher level thinking at a young age. It also refers to their preference for challenges and hard work above their current levels of functioning. Winner called this characteristic "an insistence on marching to their own drummer" to include gifted learners' autonomy and preference for independent work. Perhaps, a fourth characteristic of asynchrony or asynchronous development initiated by the Columbus Group could be added, referring to the uneven intellectual development of gifted learners or the discrepancy engendered by their advanced cognitive abilities and heightened emotional sensitivity (see Silverman, 1997). This characteristic highlights that affective needs of gifted learners must also be addressed in differentiation efforts, as gifted learners could be more vulnerable to negative perfectionism, underachievement, and problems in peer relationships.

Curriculum Dimensions

Curriculum differentiation for gifted learners is best exemplified by Joyce VanTassel-Baska's Integrated Curriculum Model that defines three curriculum dimensions for differentiation (see VanTassel-Baska and Stambaugh, 2006). The three dimensions are content, process and product, and concept. In the content dimension for gifted learners, content acceleration dominates. Students are encouraged to move rapidly at a rate commensurate with their capacity through the content area that could be highly sequential and cumulative, mastering skills and concepts within the predetermined domain of inquiry. Mastery learning time is conserved by testing for proficiency, assigning advanced work, and reorganizing basic skill areas into high-level skill clusters. In this manner, content acceleration promotes more efficient and challenging learning experiences for gifted learners.

In the dimension of process and product for gifted learners, the emphasis is on learning scientific and social investigatory skills with the development of a high-quality product. The instruction pattern is dominated by consultation and independent work. The focus is on higher level thinking, critical and creative thinking, problem finding and problem solving, and research skills that allow students to learn to think independently of textbooks and known resource materials. Unlike the topics of inquiry in the content dimension, the selected topics for exploration are determined by student interest, and the content is incidental. Evaluation is product-based instead of proficiency-based, and the emphasis is on in-depth study of selected topics rather than fast-paced surveying of the domain of inquiry. Thus, differentiation in this dimension allows gifted learners to engage in the process of constructing knowledge for themselves.

Unlike the two dimensions that are based on subject matter or process skills, the concept dimension is organized by key ideas, themes, and principles within and across domains of knowledge. The instructional context is highly interactive, and students are engaged in reading, reflecting, and writing in response to the interpretive issues raised for discussion, debate, and interdisciplinary thinking. Gifted learners are evaluated on their high-level aesthetic appreciation of powerful ideas and insight in essays or artistic products that show evidence of integrating or synthesizing forms and meaning across areas of study. Differentiation in this dimension allows gifted learners

to be exposed to many ideas and provides them with an intellectual framework not available in studying only one content area.

Teachers of Gifted Learners and Teacher Education

In a study of talent development of 120 individuals who had attained eminence in their fields by the age of 40, Benjamin Bloom (1985) found that different types of teachers were needed at different stages of the individuals' development. Specifically, the student successively needed a teacher who would teach the basics and the joy of learning, a teacher who would individualize learning with an emphasis on skills and self-discipline, and a master teacher who could raise the student to the level of an expert. Thus, teachers of gifted learners not only need the knowledge, skills, and dispositions to facilitate high-level learning, they also need to know how to differentiate to challenge gifted learners and provide opportunity to allow students to make continuous progress in developing their potential.

Laurie Croft (2003), in reviewing studies on teachers of the gifted, suggested that while these teachers share with effective teachers in general education certain common characteristics such as spontaneity, acceptance, creativity, and self-realization, they also have to develop areas of expertise that teachers of general education do not require. Based on a systematic review of past studies on teacher characteristics and competencies, John Feldhusen (1997) observed that the characteristics found in successful teachers of gifted learners were similar to those found in gifted and talented students, and were desirable in all teachers. These characteristics include being highly intelligent, achievement oriented, knowledgeable, and flexible; having cultural and intellectual interests; respecting individual differences; and relating well with gifted individuals. Feldhusen also summarized the competencies of successful teachers of gifted learners. These competencies include skills in teaching thinking skills, problem solving, and creativity; in interacting with students; in using appropriate motivational techniques; in conducting student-directed activities; and in facilitating independent research. In addition, Feldhusen also cautioned that competencies could be domain-specific in that competencies required in teaching certain specific content domains such as science and mathematics could be very different from those required for teaching other domains such as art and music. Thus, comprehensive and effective preparation opportunities in teacher education need to be developed to help dedicated teachers transform their successful characteristics and competencies into gifted education practice.

In 2006, the Initial Knowledge and Skill Standards for Gifted and Talented Education developed jointly by two US national organizations, the Council for Exceptional Children and the National Association for Gifted Children, were approved as national official standards to guide the development of programs to prepare educators of gifted learners. The 10 standards are foundations, development and characteristics of learners, individual learning differences, instructional strategies, learning environments and social interactions, language and communication, instructional planning, assessment, professional and ethical practice, and collaboration. These standards can be vehicles to facilitate the integration of gifted education into teacher preparation programs at various levels in the US as well as around the world.

Developing Gifted Programs or Programs for Talent Development

Acceleration and Enrichment

Acceleration and enrichment are terms that are commonly used in gifted education to describe types of programs differentiated for gifted learners. These program types generally have different strategies for organizing curriculum and arrangements for grouping gifted learners.

Acceleration is seen as moving faster through grade-level or subject matter curriculum than same-age peers, including offering standard curriculum to gifted learners at a younger-than-usual age. It can be viewed as adaptation involving mainly the curriculum dimension of content. Enrichment is seen as extending, supplementing, and going beyond the regular curriculum in greater depth or breadth and thus providing richer and more varied educational experiences to gifted learners. It can be viewed as adaptation involving mainly the curriculum dimension of process and product. Research studies on acceleration and enrichment have indicated that both accelerated and enriched classes result in higher achievement for gifted learners. For example, in his Michigan meta-analysis studies, Kulik (2003) reported that the accelerates outperformed the non-accelerates by about 1 year, and students in enriched classes outperformed other students by about 4 months.

On closer examination, the distinction between acceleration and enrichment however easily breaks down. Any enrichment experience will generally involve new topics or greater depth and are advanced or accelerated when compared with the regular curriculum. Similarly, acceleration with advanced subject matter generally involved enriched experiences. One way to distinguish the two by educators is to include as acceleration any strategy that results in advanced placement or credit and to include as enrichment any strategy that supplements or goes beyond standard grade-level work without advanced placement or credit. It is believed that well-round, coherent, and long-range gifted programs should provide both enrichment and acceleration opportunities for gifted learners.

Two Exemplary Program Models

To plan and conduct learning activities for gifted learners, one needs a viable program model. Joyce VanTassel-Baska (2000) specified that a viable model should provide a system for designing and developing appropriate curriculum for the target population of gifted learners. In addition, the model should differentiate according to the principles of gifted education, applicable to all major content areas of school-based learning, and to different age and grade levels. In past decades, numerous viable program models have emerged, with different emphases on acceleration and enrichment. Two exemplary program models, the Talent Search Model and the Schoolwide Enrichment Model, representing respectively the different emphasis on acceleration and enrichment are introduced here. Both models have enjoyed widespread use and research attention,

and have years of research, development, and implementation behind them.

The Talent Search Model

Talent search was pioneered by Julian C. Stanley at Johns Hopkins University in 1972 in his Study of Mathematically Precocious Youth (SMPY). It has evolved and expanded in many ways, and can be viewed not only as a curriculum model but also a program model of overall talent development with an emphasis on acceleration options. Broadly conceptualized, it is a model for the discovery and development of academic talent and its overall purpose is to educate for individual development over the lifespan (Lupkowski-Shoplik et al., 2003). Two steps are involved in the discovery of students who are exceptionally able academically. The first step is the identification of students who have demonstrated a high level of academic performance, as documented by high performance on grade-level tests (e.g., 97th percentile). The second step is to determine these students' potential for academic challenge by out-of-level testing or testing them with tests above their grade levels (usually 2 years).

Initially, the interest in SMPY was in finding adolescents (mainly seven-grade and some eight-grade students) who were exceptionally gifted in mathematics, so the mathematics section of the Scholastic Aptitude Test (SAT-M) was used as a measure of students' mathematical reasoning abilities. Students who scored at or above 51st percentile for male college-bound high school seniors were considered mathematically precocious. Starting from 1980, talent search was extended to verbally gifted youth, and the Verbal section of the SAT (SAT-V) was used to identify them. After the revision and renaming of the SAT as the Scholastic Assessment Test (SAT I), the SAT continues to serve as the testing instrument that taps into high-level verbal and mathematical reasoning to identify students. In addition to the SAT I, another college entrance exam, the American College Testing Program is also used.

The identification of students as mentioned above with the use of a secure and difficult testing instrument to tap into high-level verbal, mathematical, and spatial reasoning is followed by the diagnostic testing–prescriptive instructional (DT-PI) approach. These identified students are then counseled and encouraged to pursue any of a 'smorgasbord' of acceleration options that include attending college part time or earning college credit by examination, skipping a grade, receiving individual highly accelerated tutoring, or entering college early. Thus, the DT-PI approach provides identified students with different forms of educational facilitation through utilizing acceleration or curricular flexibility and by developing fast-paced academic programs for these students.

Research studies have been conducted to demonstrate the viability of the Talent Search Model, suggesting positive results of accelerative practices and a lack of negative consequences such as knowledge gaps or loss of interest. It has been concluded that intellectually gifted students as accelerates consistently outperformed the non-accelerates, and they would not achieve as highly if talent search educational opportunities were not provided for them. Benefits for talent search participants also included improved zest for learning, enhanced self-esteem, better education preparation, and early entrance to college and a profession. Of particular values were the experiences of intellectual stimulation, the acknowledgment of talents, and the contact with intellectual peers, as evaluated by talent search participants. Currently, a 50-year longitudinal study (1972–2022) is ongoing with a sample of 6000 talent search participants.

Over the years, the Talent Search Model has been adopted in different regional centers in the US, extending from identifying academically talented students at the junior high level to those of the elementary level. With its successful application in the US, the Talent Search Model has its influence in the development of variations of the model in East Asian countries or regions such as Mainland China and Taiwan.

The Schoolwide Enrichment Model

The School Enrichment Model (SEM) was developed in the mid-1970s by Joseph Renzulli of the University of Connecticut and was implemented as a gifted and talented programming model in school districts in Connecticut (Renzulli and Reis, 2003). It combines two previously developed models, the Enrichment Triad Model (ETM) and the Revolving Door Identification Model (RDIM). Like the Talent Search Model, the SEM is more than a curriculum model and can be viewed as a program model of overall talent development, but unlike the Talent Search Model, the focus is on the development of creative-productive as well as academic giftedness with a greater emphasis on enrichment rather than acceleration options.

The ETM offers three types of enrichment experiences appropriate for students in the school talent pool identified as having high levels of ability, interest, and task commitment. These experiences are designed to encourage students' creative expression and productivity. Type I enrichment provides general exploratory experiences such as lectures by guest speakers, field trips, demonstrations, performances, interests and hobbies, the use of audiovisual materials, and other events designed to expose students to new and exciting topics, ideas, and fields of knowledge not ordinarily covered in the regular classroom. Type II enrichment provides group training activities using instructional methods and materials purposefully designed to promote the development of thinking and feeling processes. In general, training is in the areas of creative thinking, problem solving, communication skills, and skills in learning to learn. Type II activities provide the crucial tools and process skills to enable gifted students to advance to Type III activities. Type I and Type II activities can also be offered to all students in the school, as they will benefit all students. Type III enrichment is defined as individual or small group investigative activities and artistic productions selected and pursued by students who are willing to commit themselves to acquire the advanced content and process training. Each participating student will assume the role of a first-hand inquirer, and think, feel, and act like a practicing professional. By focusing on creative-productive giftedness, students work on issues and areas of study that have personal relevance and can be escalated to appropriate challenging levels of investigative and creative activities.

The RDIM introduces the notion that students are selected for participation in the talent pool on the basis of multiple criteria that include the use of achievement test scores and different indices of creativity as well as teacher-, parent-, and self-nomination. These students are also observed in

classrooms and enrichment experiences for signs of advanced interests, creativity, or task commitment to become involved in Type III creative productivity. Identification continues year-round rather than being confined to the beginning of the school year. Thus, by operating RDIM in conjunction with the ETM, the SEM adopts a more flexible approach to identifying high-potential students, resulting in a talent pool of 15–20% of the school population. In this manner, a student with high achievement test and IQ scores will be automatically considered, enabling those students who are underachieving in their academic schoolwork to be included in the talent pool.

Students in the talent pool are eligible for special services that include assessment on students' abilities, interests, and style preferences, resulting in a Total Talent Portfolio for each student, and curriculum compacting. The Total Talent Portfolio records the assessment of students' strengths that provide the foundation for effective learning and creative productivity. Curriculum compacting, on the other hand, ensures a better match between the achievement levels of individual students and the school curriculum through modifying the regular curriculum by eliminating portions of previously mastered content and by substituting with alternative work.

Recently, Joseph Renzulli and Sally Reis (2009) have developed an Internet-based interactive program called the Renzulli Learning System (RLS) that assists teachers who seek effective tools for differentiation to implement the various components of the SEM. While improved academic achievement and opportunities for creative productivity remain as the major goals of the RLS as those of the SEM, the RLS has a series of metacognitive tools resulting from computer-based learning environments. These metacognitive tools give the added benefits to students by helping them organize and self-regulate their own learning so that they can make the best use of their time, resources, and the cognitive skills that contribute to higher order thinking.

A variety of studies have been conducted to assess the effects of the SEM in schools of varying demographics and in different settings. These studies include evaluation studies, longitudinal studies, studies on curriculum compacting, studies on students' creative-productive behaviors, and studies on underserved gifted populations. These studies have yielded mostly positive results on student motivation and sustained interest in learning. In addition, they have also shown that the SEM can be applied in a variety of school settings, with various student populations, including high-ability students with learning disabilities and those who underachieve in schools. Thus, the SEM has been widely adopted in schools in the US and also has appeals internationally. Variations of the SEM in some forms have been operative in some East Asian countries or regions such as Singapore and Hong Kong.

Nurturing Personal and Social Development

In designing and developing gifted programs for gifted learners, educators generally focus on issues related to what teachers should know about the learning characteristics of students, what skills are needed to provide an appropriate level and instruction for them, and how curriculum and instruction can be differentiated to appropriately challenge them. Very often, social and emotional issues of gifted learners at different stages of development are considered secondary. However, these are important issues, and strategies or efforts to build skills that promote positive social and emotional development among gifted learners should be included as an integral part of any gifted programming.

Developmental Tasks

While gifted children and youth have developmental tasks similar to other typically developing students, some of their developmental tasks may be more accelerated and others may be more difficult to accomplish. For example, gifted children may have more friendship preferences and are ready for more sophisticated types of friendship than their age-peers. Gifted youth may need and could be ready for career development work earlier than other typically developing adolescents. Some gifted children may find it more difficult to develop a strong work ethic because they are placed in a regular classroom with little academic challenge.

Giftedness or talent may also create greater developmental challenges for gifted learners. For example, gifted children may feel their emotions more intensely than other children and may have difficulty learning to manage their emotions. Gifted adolescents may find the achievement–affiliation conflicts particularly acute at a time when they value both achievement and peer relationships but view that they need to choose between being intelligent and being popular. Gifted adolescents may also have greater difficulty in time management because they are taking more demanding classes and engaged in more extracurricular activities. Normally, gifted children and youth are able to master the basic personal and social developmental tasks. However, some may have more difficulties, and others may even have social and emotional adjustment problems that require counseling attention.

Social and Emotional Issues

The reasons that gifted learners may experience social and emotional adjustment problems can be grouped into two types, those arising endogenously from within the individual and those arising exogenously from the environment (Moon, 2008). An example of endogenous challenges is the commonly poor or low social concepts of gifted learners compared with their generally higher-than-normal academic self-concepts. This discrepancy could lead to adjustment difficulties for gifted learners. Another example is the perfectionist tendencies common to many gifted learners. While setting high standards and striving for excellence as positive or adaptive aspects of perfectionism are desirable and should be promoted in gifted learners, excessive concerns and dissatisfaction with effort-performance discrepancy as negative or maladaptive aspects of perfectionism will pose adjustment problems to gifted learners. A third example is related to the fact that some gifted learners are twice-exceptional students with co-occurring giftedness and disabilities. A student who is intellectually gifted and has an attention-deficit/hyperactivity disorder is twice exceptional, as is a spatially gifted student who also has dyslexia. Very often, twice-exceptional students may struggle

with peer relationships, and therefore are often at risk for emotional disorders.

In addition to endogenous challenges, there are exogenous or environmental factors affecting the social and emotional development of gifted learners. These exogenous influences mostly come from the family and school environments. Examples of exogenous challenges from the school include bullying by peers related to the gifted learner's giftedness and inappropriate school placements, where a highly gifted student experiences boredom or social rejection in a regular classroom. Families that are supportive and organized, yet allowing for individual independence and personal growth, affect positively the social and emotional development as well as academic talent development of gifted learners. However, some families are less functional, and families with gifted children are vulnerable to some additional negative family dynamics. One notable negative family dynamic is called 'parentification,' where gifted children are given too many family responsibilities very early in their lives because they appear so adult in their language and behavior. Apart from these endogenous and exogenous challenges, it has to be noted that negative influence on gifted learners' social and emotional development could arise from the interactions between the characteristics of students and their environments. An example is underachievement. In some cases, the causes could be more internal, as for twice-exceptional students, and in other cases, the causes could be more external, as for highly gifted students placed in regular classrooms. Nonetheless, gifted learners may need the attention of teachers, parents, and counselors to help them cope with problems in their social and emotional development.

Intervention Strategies

Gifted learners need to come to an understanding and acceptance of their giftedness or exceptionality, their intensity and sensitivity of feelings, and their need to cope with social and emotional issues arising from their family and school environments. Different levels of problems in social and emotional adjustment may require different intervention strategies from teachers and counselors. Experiential learning activities such as problem-based learning, independent study, service learning, and simulations can be integrated into any content area. These activities develop students' personal and social competence by activating their emotions, encouraging their social interactions, and facilitating their self-direction. Affective curriculum provides more explicit and direct instruction in specific personal, social, and emotional skills such as skills in regulating emotions, mediating conflicts, and setting priorities. While these direct and indirect instructional strategies are helpful in preventing social and emotional adjustment difficulties in gifted students, more severe adjustment difficulties may require counseling or therapeutic interventions using individual, group, and family counseling. Preventive counseling strategies, such as support groups, bibliotherapy, and videotherapy, can be helpful in promoting positive social and emotional development in all gifted learners.

Thus, a comprehensive and well-balanced gifted program should always have a strong affective orientation and nurture the social and emotional development of gifted students.

Trends and Future Directions

Gifted education has a history of about 90 years counting from Terman's time. During this relatively short span of time, it has evolved from education aimed to provide services to a small selected group of children and youth identified as gifted or talented to talent development for all students. Thus, gifted education, within this framework, allows that an equitable pursuit of excellence be applied to all students, including the at-risk and less able, the able, the highly able, or the gifted and talented. This extension could be viewed as an effort to resolve societal concerns about the controversy between equity and excellence, but has a lot to do with the broadening of the conception of giftedness and talent in association with the multidimensional conceptions of intelligence. The conventional notion of intelligence as a single quality that is manifested throughout a person's intellectual performances, measurable by a unitary quantifiable index called IQ score, and present as a potential early in life or not at all is viewed increasingly as an inadequate measure of giftedness. Over the decades, the defining qualities of giftedness have been expanded to include, among others, task commitment, mastery orientation in motivation, different modes of thinking, high self-concept, and creativity, qualities that emphasize the nurturing aspects of giftedness (see Sternberg and Davidson, 2005).

Although some education practitioners today because of convenience or tradition still focus only on intellectually gifted children and youth and identify them using a high IQ score (e.g., 130), there is an increasing acceptance of a broadened notion of giftedness and multidimensional conception of intelligence among educators in the field of gifted education. It is expected that the trend of basing education practices on broadened notions of giftedness will continue, and further broadening will take into consideration different conceptions of giftedness in different cultural settings.

Apart from the multicultural considerations in broadening conceptions of giftedness, there is also a trend of seeking explanations of giftedness based on current knowledge of brain functions from neurology or neuropsychology (see Shaw et al., 2006). The kinds of questions asked could be: how is giftedness or talent represented on the brain level, could giftedness be simply accounted for by 'neuronal resource' as suggested by, for example, more active right hemisphere, or could the 'neural or neurocognitive efficiency' account address differences between the talented and the nontalented while working on identical tasks? Further, it is believed that being sensitive to the cognitive structures of the gifted or how gifted learners process, organize, and deal with incoming information and knowledge, and their cognitive preferences or styles is important for optimizing learning with gifted learners. Thus, the practical applications of increasing knowledge of brain functions in studies of learning in gifted learners will continue to develop into a major trend in gifted education in the next decade (see Sousa, 2003).

Another major trend over the next decade is likely to be the increasing use of instructional communication technology based on advances in information technology (see Pyryt, 2009). Applications like Renzulli's RLS as enrichment strategies to maximize the creativity and potential of gifted learners are likely to continue and become more complex and sophisticated.

Computers and software will certainly be utilized to do a variety of projects to transmit knowledge and enhance skills. The Internet has provided a wealth of information, and will continue to open many doors for teachers and gifted learners not only in teaching and learning per se but also in mentoring or telementoring. However, teachers and students alike need to be using technology more efficiently, effectively, and consistently, and to keep up with innovations in technology related to the Web in the future.

Predicting the future is fraught with uncertainties and what could be more certain will be the inevitable great changes in the landscape of gifted education. With optimism, the changes should be positive. Gifted education will move forward with better understanding of giftedness or talent, including but not restricted to multicultural and neuropsychological considerations, and with more and effective evidence-based strategies or options for the education of gifted and talented children and youth and for talent development in all students.

See also: Creativity Training; Instructional Psychology; Intelligence, Prior Knowledge, and Learning; Learning and Instruction: Social-Cognitive Perspectives; Special Education; Teacher Education.

Bibliography

Bloom, B.S. (Ed.), 1985. Developing Talent in Young People. Ballantine Books, New York.

Croft, L.J., 2003. Teachers of the gifted: gifted teachers. In: Colangelo, N., Davis, G.A. (Eds.), Handbook of Gifted Education, third ed. Allyn & Bacon, Boston, MA, pp. 558–571.

Feldhusen, J.F., 1997. Educating teachers for work with talented youth. In: Colangelo, N., Davis, D.A. (Eds.), Handbook of Gifted Education, second ed. Allyn & Bacon, Boston, MA, pp. 547–552.

Kulik, J.A., 2003. Grouping and tracking. In: Colangelo, N., Davis, G.A. (Eds.), Handbook of Gifted Education, third ed. Allyn & Bacon, Boston, MA, pp. 268–281.

Lupkowski-Shoplik, A., Benbow, C.P., Assouline, S.G., Brody, L.E., 2003. Talent searches: meeting the needs of academically talented youth. In: Colangelo, N., Davis, G.A. (Eds.), Handbook of Gifted Education, third ed. Allyn & Bacon, Boston, MA, pp. 204–218.

Moon, S., 2008. Personal and social development. In: Karnes, F.A., Stephens, K.R. (Eds.), Achieving Excellence: Educating the Gifted and Talented. Pearson Prentice Hall, Upper Saddle River, NJ, pp. 82–98.

Pyryt, M.C., 2009. Recent developments in technology: implications for gifted education. In: Shavinina, L.V. (Ed.), International Handbook of Giftedness. Springer, New York, pp. 1173–1180.

Renzulli, J.S., 1978. What makes giftedness? Re-examining a definition. Phi Delta Kappan 60, 180–184, 261.

Renzulli, J.S., Reis, S.M., 2003. The schoolwide enrichment model: developing creative and productive giftedness. In: Colangelo, N., Davis, G.A. (Eds.), Handbook of Gifted Education, third ed. Allyn & Bacon, Boston, MA, pp. 184–203.

Renzulli, J.S., Reis, S.M., 2009. A technology-based application of the schoolwide enrichment model and high-end learning theory. In: Shavinina, L.V. (Ed.), International Handbook of Giftedness. Springer, New York, pp. 1203–1223.

Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., Evans, A., Rapoport, J., Giedd, J., 2006. Intellectual ability and cortical development in children and adolescents. Nature 440, 676–679.

Silverman, L.K., 1997. The construct of asynchronous development. Peabody Journal of Education 72 (3 & 4), 36–58.

Sousa, D.A., 2003. How the Gifted Brain Learns. Corwin Press, Thousand Oaks, CA.

Sternberg, R.J., 1985. Beyond IQ: A Triarchic Theory of Human Intelligence. Cambridge University Press, New York.

Sternberg, R.J., Davidson, J. (Eds.), 2005. Conceptions of Giftedness, second ed. Cambridge University Press, New York.

VanTassel-Baska, J., 2000. Theory and research on curriculum development for the gifted. In: Heller, K.A., Monks, F.J., Sternberg, R.J., Subotnik, R.F. (Eds.), International Handbook of Giftedness and Talent, second ed. Elsevier, Oxford, UK, pp. 345–365.

VanTassel-Baska, J., Stambaugh, T., 2006. Comprehensive Curriculum for Gifted Learners, third ed. Allyn & Bacon, Boston, MA.

Winner, E., 1996. Gifted Children: Myths and Realities. Basic Books, New York.

Relevant Websites

http://www.jhu.edu – Center for Talented Youth at Johns Hopkins University.

http://cfge.wm.edu – Center for Gifted Education at the School of Education of the College of William and Mary.

http://www.tip.duke.edu – Duke University Talent Identification Program (TIP).

http://www2.warwick.ac.uk/study/iggy – International Gateway for Gifted Youth at the University of Warwick (U.K.).

http://www.gifted.uconn.edu – Neag Center for Gifted Education and Talent Development at the University of Connecticut, and the National Research Center on the Gifted and Talented.

http://www.nagc.org – National Association for Gifted Children (NAGC) (U.S.A.).

http://www.nagcbritain.org.uk – National Association for Gifted Children (U.K.).

http://www.sengifted.org – Supporting Emotional Needs of the Gifted (SENG).

Education, Geography of

Peter Meusburger, Heidelberg University, Heidelberg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Geography of education focuses on spatial variations in the provision and consumption of education, on the impact of local milieus and social environments on educational attainment, on spatial variations of the sociodemographic structure of the teaching profession, on the influence of policies and demographic change on location patterns of schools, on the mobility of scholars and students, and on other issues of education where spatiality, place, spatial relations, and environments play an important role. As educational achievement strongly affects a person's social status and chances on the labor market, an education-oriented approach can be applied to many topics of geographical relevance.

Geography of education focuses on spatial variations in the provision and consumption of education. It analyzes the impact of local milieus, spatial settings, and social environments on educational attainment; studies spatial variations of the sociodemographic structure of the teaching profession; deals with the influence of policies and geographical structures (e.g., population density, distribution of settlements, transport system, and employment structures) on location patterns of schools; studies the mobility of scholars and students; and explores the impact of cultural values as well as political, economic, societal, and demographic change on educational systems. As educational achievement strongly affects a person's social status and his or her chances on the labor market, and as knowledge is the basis of human action, an education-oriented approach can be applied to a great number of topics relevant to human geography. So far, the vast majority of publications in this field dealt with formal education, but nonformal education and various forms of preschool education and child care (see, e.g., Ball and Vincent, 2007; Holloway et al., 2010) also belong to the research issues of geographies of education.

For various reasons, the educational system has again and again been in the spotlight of politics. It displays power relations between center and periphery, and indicators of educational attainment have frequently been used as a means to uncover social and economic inequalities, racism, and discrimination of ethnic minorities. In many new nation-states, the primary functions of the school system were not only to remove illiteracy and create a well-educated labor force, but to consolidate administrative power, support national identity and hegemonic culture, and assimilate ethnic minorities (Freytag, 2003). Therefore, the language of instruction, the selection of teachers, and the spatial organization of elementary schools have been of paramount political importance both for state nationalism and nationalist movements as well as for cultural emancipation of ethnic minorities.

The quick expansion and specialization of research has split up the wide field of geography of knowledge and education (as it was introduced in the first edition of this encyclopedia) into geography of education in the narrow sense, geography of knowledge, and geography of science. Innovation studies, cluster and network studies, research on creative milieus, labor markets and migration, and many other research fields of human geography developed likewise strong interests in spatial disparities of knowledge, skills, and educational achievement. The following sections will focus on geographies of education in a narrow sense.

Early Roots and Development of Geography of Education

Early historical roots of empirical research about spatial structures of education can be traced back to the first half of the nineteenth century. Some of the social problems associated with early industrialization such as poverty, child labor, crime, alcoholism, and a lack of decent housing generated an interest in moral statistics. Many nineteenth-century social reformers believed that such social problems were caused by ignorance and a lack of moral education. They therefore started investigating into the social and spatial variations of illiteracy and educational attainment, spatial inequalities in the provision and quality of various types of schools, the skills and salaries of teachers, the availability of books in households, and other elements to describe the cultural and economic settings of neighborhoods (Marsden, 1987; Meusburger, 1998: pp. 191–195).

France is the first country where regional differences in illiteracy were studied in a nation-wide and comprehensive way. The main sources of these studies were examinations of military recruits and marriage registers, both of which contained data on the ability to read and write (Furet and Ozouf, 1977; Heffernan, 1989). In 1826, Charles Dupin published the *Carte figurative de l'instruction populaire de la France*. This map (reprint in Meusburger, 1998: p. 193) revealed remarkable regional disparities in educational attainment between northern and southern France. The tables added to this map compared the educational attainment, the number of patents for inventions, and the membership in the Académie Française with various economic indicators suggesting a high correlation between the population's educational achievement and the economic performance of a region.

In the United Kingdom, a number of Statistical Societies emerged in the 1830s and 1840s to study the living conditions and social situation of the poor and to foster social reform. Literacy was again a main concern of these

studies. J. Fletcher (1849: p. 176) published a map on *Ignorance in England and Wales* in the *Journal of the Statistical Society of London*'. (A modified version of the map was published by Hoyler, 1996: p. 188.) When C. Booth analyzed poverty in London, he used various educational indicators on the scale of neighborhoods.

However, apart from occasional early studies, human geography neglected the field of education until the 1960s. One of the exceptions was the Department of Geography, University of Chicago, where E. Parker and Ch. Colby supervised after World War II four PhD theses about "educational land use." Three of them (by H.H. Gross, E. Eisen, and A.K. Philbrick) were published in the "University of Chicago, Department of Geography Research Papers" between 1948 and 1949 (for details see Meusburger, 1998: pp. 204–205).

The interest of geographers in education grew when the 1960s and 1970s brought in many countries an unprecedented expansion of the educational system, a steep increase in university enrollment, the establishment of many new secondary schools and universities, and the need to reorganize the location patterns of elementary schools. This boom in expanding, modernizing and reorganizing educational systems was also nourished by emerging discourses about human resources, brain drain, postindustrial society, knowledge economy, and the legitimacy of social and spatial inequalities. In this period, locations for new schools and universities had to be negotiated, the closure of small elementary schools had to be vindicated, catchment areas and student transport had to be reorganized, and the reasons of intolerable spatial disparities in educational attainment had to be explored in order to counteract them. As these tasks needed competence and experience in geography and urban and regional planning, a sensitivity for the importance of local identities and local milieus, as well as detailed knowledge of the history, culture, and economy of the towns and regions in question, geographers soon became involved in applied educational research and regional educational planning. In the 1970s and 1980s, social scientists became increasingly aware that the famous slogan 'place matters' has a special relevance in education and that residential location, segregation, ethnic diversity, and regional mobility affect education in various ways (Gordon and Monastiriotis, 2007: p. 1204).

One of the first and most influential pioneers of the newly emerging field of geography of education was the German geographer Robert Geipel (born 1929), who focused his research from the early 1960s to the 1990s almost exclusively on geography of education (Geipel, 1966). He was both full professor of geography at the Technical University in Munich (1969–94) and director of The Bavarian State Institute for Higher Education Research and Planning (in Munich (1982–94)). In these functions, he had close scientific relations to the pioneers of this research field in other countries and disciplines; together with the Academy for Spatial Research and Planning in Hannover he organized long-term interdisciplinary working groups on 'spatial educational research.' In the 1970s and 1980s, dozens of German, Austrian, and Swiss geographers got involved with geography of education. As a result, the Association of German Geographers founded 1983 a working group 'geography of education' that is still very active today and attracts not only human geographers, but also social and educational scientists.

In United Kingdom, the early pioneers of geography of education in the 1970s and 1980s were G. Hones, W. Marsden, R. Ryba, W.T.S. Gould, L. Bondi, and H. Matthews. However, the boom of education-related publications of British geographers started only in the 1990s, and the vast majority of these publications restricted their citations to English-language publications (see review articles of Hamnett et al., 2007; Hanson Thiem, 2009; Holloway et al., 2010; Holloway and Jöns, 2012; Taylor, 2009).

Main Research Topics

In its early period between the 1960s and late 1990s, geography of education predominantly focused on the spatial organization of school systems (size, location criteria, and catchment areas of schools and universities); on the reasons and consequences of spatial disparities in educational attainment; on the social composition of schools (measured by the educational achievement and occupations of both parents); on school segregation and student transport; on spatial disparities of vertical social intergenerational mobility (comparison of educational achievement of parents and children); on spatial variations of professional skills, sociodemographic structures, and the feminization of the teaching profession; on spatial patterns of expenditure on education; on spatial impacts of various educational policies and school planning concepts; on relations between communities and schools, especially on the social, cultural, and economic effects of school closures on communities; on the impact of educational achievement on professional careers and migration; on the role of schools in ethnic conflicts and 'nation building'; on relations between illiteracy and economic development; on the spatial distribution of jobs for highly and lowly skilled employees; on theoretical concepts explaining the spatial concentration of jobs for highly educated people; on career mobility and networks of scientists and highly skilled professionals; and on the correlation of different levels of educational achievement with unemployment, female employment rates, housing quality, reproduction rates, and many more variables (for details see Bondi and Matthews, 1988; Kramer, 1993; Meusburger, 1980, 1998; Schmude, 1988; Taylor, 2009). Already in this early period, many geographers situated educational attainment in a very broad economic, social, cultural, and political context; they identified a large range of factors affecting the location pattern of elementary schools, the mobility of university graduates, and other structures of the educational system (Jahnke, 2005; Mayr and Nutz, 2002; Meusburger, 1998).

Since the end of the 1990s, new research questions and new theoretical approaches have developed. Some of these new approaches deal with normative understandings of childhood, treat young people as the subjects rather than objects of education, deepen the understanding of children's spaces of learning, attend to (gendered) identity construction in schools and focus on various aspects of nonparental child care provision, to name just a few (for details see Holloway et al., 2010; Holloway and Jöns, 2012; Matthews and Limb, 1999). Other new research activities focus on the neoliberal 'international education industry' worth an estimated US$ 100 billion (Waters, 2006) and its consequences. According to Waters

(2006: pp. 1046–1049), the establishment of offshore schools and universities overseas first of all has the aim to attract fee-paying (revenue generating) students, but it also creates new geographies of social exclusion within 'student sending' societies.

Varying Priority of Research Questions

The priority of research questions varies not only over time but also from one country to the other. Priorities depend on whether elementary education has a centralized or decentralized governance; which administrative authority (local community, district, province, state, or nation) establishes, finances, and controls various types of schools or selects and pays teachers; whether elementary school catchment areas are defined by school administrations or are the result of parents' decisions; how pupils of different ability are distributed across different types of schools; whether the selectivity of schools is based on the children's abilities or the financial resources of parents; whether a society has a high or low degree of social inequality, ethnic diversity, and vertical social mobility; whether competition for pupils among schools is open or not; whether the ethnic composition of a population is a source of social conflict or not; whether the 'problem areas' of education or the socially deprived areas are rather in urban or rural regions; and which data sources are available to scholars.

In the United States and the United Kingdom school segregation, desegregation bussing, the competition between nongovernment and state schooling, educational inequalities that arise from residential sorting, the flight from working-class schools, repositioning inner-city schooling, relations between family income and educational attainment, and the neoliberal education market have been major research themes (Gibbons and Telhaj, 2007; Gulson, 2007). In other states that have no tuition fees, a less hierarchic class society, less segregation, or a high-quality state schooling, the focus has been on other topics. This had the consequence that different research cultures and citation networks developed within geography of education, and that review articles focusing on developments in one nation or on publications written in one language are bound to fail when they generalize their findings and arguments. In order to avoid such national specialties, the following sections discuss only three research topics of general relevance.

Organization, Size, and Location Patterns of Elementary Schools

The location, organizational form, and minimum size of elementary schools and the daily transport of young children to distant schools are heavily contested in many countries. Many conflicts between communities and school administrations have arisen from the fact that some educational planners or school administrators had little sensitivity for regional and cultural diversity or for the importance of local identities and the symbolic meaning of places. Some politicians and administrators have not been aware that the planning principles and guidelines concerning the location and organizational structure of schools, minimum class size, and even the training of teachers should be flexible and adaptable in order to account for local requirements. Sparsely inhabited rural areas suffering from depopulation had to respond much earlier than did more densely populated urban agglomerations to falling school enrollment numbers, fiscal constraints, and the difficulty of attracting skilled teachers. One- or two-class elementary schools in rural areas, where two to four grades are taught in one classroom, need different teaching methods and differently trained teachers than large urban schools, in which each grade is taught separately in a classroom.

In rural areas of many states, two different concepts of regional educational planning confront each other (for details see Kramer, 1993). The concept of centralized large elementary schools to which pupils are transported over long distances by school buses or parents, claims to be financially more efficient, to offer more flexibility in the assignment of teachers, to offer a larger variety of study courses, to create more social equality, to offer a higher quality of education, and to be 'more modern.' The negative implications of this concept, such as transporting 6- to 10-year-old children over large distances in social environments unfamiliar to them, confronting the children with values contrary to those of their family environment, exposing them to emotional stress, and cutting down a large part of their leisure time are usually overlooked by the proponents of this concept. Neither do they seem to care about the long-term negative effects of closing the only elementary school in a village.

An alternative view advocates for small decentralized elementary schools located in (or near) the villages and towns where the 6- to 10-year-old pupils live (Kramer, 1993). Because of smaller classes, this approach requires more teachers and more financial resources per 100 pupils than for large urban schools; it is less flexible in the assignment of teachers and demands differently trained teachers. However, proponents of this concept claim that the pedagogic advantages and many positive extracurricular effects outweigh the higher costs. Adherents of this approach have a different view of childhood, are more sensitive to spatial disparities and local needs, view as negative the experience of children with long-distance bussing, and are primarily interested in the relationship between educational provision and community coherence (see also Bunar, 2010). They maintain that in peripheral rural villages, elementary school teachers play an important extracurricular function and have a greater impact on the community than such teachers have in large cities. If sparsely inhabited rural areas are successful in attracting skilled teachers, their schools will be indispensable elements of communal infrastructure; their schools will be the places where local identity and social networks are formed and strengthened, and where cultural activities and sensitivity for regional environmental problems are nurtured. In such social milieus, skilled teachers are still key persons in the community and have to fulfill many out-of-school functions (e.g., conductor of the church choir, soccer trainer, leader of the amateur theater, conductor of the local brass band, or leader of the voluntary fire department). In urban areas, these functions are more or less professionalized so that the extracurricular roles of teachers are much less needed. In a sparsely inhabited rural area, the closure of an elementary school and the disruption of children's living and learning

spaces have much more negative social, cultural, and economic effects than in a large metropolitan area.

Clashes between these two concepts gain a special political explosiveness in multiethnic or multilingual areas. Since the early beginnings of compulsory schooling in the late eighteenth or early nineteenth century, administrative control of school systems and the language of instruction in elementary schools emerged as focal points of conflict in bilingual areas. During the nation building process, one of the primary functions of state elementary schools was to support national unity and to assimilate ethnic minorities. The language of instruction in the first grades of elementary school was of paramount instrumental value in suppressing or emancipating ethnic minorities. School authorities could easily disadvantage autochthonous ethnic minorities by raising official standards for the minimum school and class size, by forbidding minority languages in instruction, and by assigning teachers who did not speak the language of the minority pupils and were not familiar with their culture.

Those autochthonous minorities who are not allowed to have elementary schools teach in their mother tongue, who are offered badly equipped schools and poorly qualified teachers, and whose cultural heritage is devalued or omitted in schoolbooks, may demonstrate low performance, low commitment, low aspirations, and high resentment, which then become justifications for further discrimination and reduce minority people's power to compete in the labor market on equal terms for jobs (for details see Freytag, 2003; Gamerith, 2006; Meusburger, 1998).

Also in urban areas, educational policy should be coordinated with housing policy, urban planning, and urban renewal in order to be successful. Without improving an area's housing quality and reputation, educational policies will not be able to increase low educational attainment in that area (Giband and Lacquement, 2007; Gordon and Monastiriotis, 2007; Gulson, 2007).

Spatial Disparities of Educational Achievement and Attainment

Indicators of educational achievement describe the final result of an educational process. Indicators of educational attainment are used to describe a current status in an ongoing (unfinished) process of education. As some indicators of social status such as income, consumer behavior, life style, or housing quality are not as regularly surveyed by official statistics as is educational achievement or at least not published for small administrative units, various forms of educational achievement or attainment became significant and frequently used indicators describing spatial variations of social status, social inequalities, and social processes such as vertical social mobility. This use is justified by the fact that well-chosen indicators of educational attainment (achievement) highly correlate with all other indicators of social status.

The educational achievement of an adult population is measured by indicators such as the proportion of illiterates, the proportion of the adult population who completed a certain level of education (e.g., proportion whose highest level is elementary school, vocational school, college, or university), or by median school years completed. Indicators of educational attainment illustrate the flow of students through the hierarchic structure of an educational system, as well as the students' performance in various types of schools. Spatial disparities in the educational attainment of pupils can be measured by indicators of participation, performance, success, and failure. Some of the most frequently used are the proportion of an age group attending a certain type of school, transition rates (enrollment rates) of a 1-year age group to various institutions of higher education, school survival expectancy or persistence of school attention (how many years do pupils remain in the school system?), school performance indicators such as results of general certificates or achievement tests (e.g., in the United States 'National Assessment of Educational Progress'; in the United Kingdom 'Pupil Level Annual School Census' or results of 'General Certificate of Secondary Education'), the share of dropouts and truants (school absence rates), or out-of school-suspension rates.

The usefulness or explanatory power of these indicators is not stable over time and space. First, it is influenced by the differentiation and selectivity of the school system. A highly differentiated and selective school system provides more significant indicators of educational attainment than a less differentiated and a less selective one, where almost all pupils successfully achieve high levels of education. Second, most indicators have a kind of life cycle and lose their explanatory power after a certain period of time. In some (developing) states, illiteracy rates may still be the most significant indicator to describe socioeconomic disparities within a region, but in many other states illiteracy rates are irrelevant because literacy is so widespread.

Spatial disparities of educational attainment are influenced by at least eight groups of factors (Meusburger, 1998; Webber and Butler, 2007).

1. The foremost factor is the sociocultural milieu in the family, for example, parental social status, educational achievements and professions, their cultural values and attitudes toward education, the time they spend with their children, and the encouragement they give their children. Families as educational settings contribute to the development of the cognitive abilities, motivations, and aspirations of children. However, the sociocultural milieu should not be equalized with the financial situation of a family. It is true, that in many cases there is a close positive correlation between affluence of parents and educational attainment of children. But affluence is only a proxy variable for other factors, and there are plenty of examples of very rich communities showing low enrollment rates in institutions of higher education.
2. Qualification of teachers and their interaction with the community. The impact of teachers on the educational performance of talented pupils can hardly be overestimated. However, not all teachers have the same training and skills, not all teachers are willing to commit themselves to the needs of the local community, and not all teachers are socially accepted by the community. High turnover rates of teachers are usually an indicator that they are not satisfied with the situation at their school. High rates of extracurricular activities of teachers for their pupils or their community signal a high degree of commitment.

3. School composition effect. Schools differ strongly with regard to the social background and abilities of their pupils. As high-attaining and low-attaining pupils are sorted into different schools after passing elementary school (see Gibbons and Telhaj, 2007: p. 1282), school composition shapes the values and attitudes prevailing among peer groups and has a strong influence on educational attainment, both in countries with and without school fees. Even on the same level of schooling (e.g., secondary schools), small differences in the curricula (e.g., a focus on modern languages or on natural sciences) result in significant differences in the social composition of the students (Meusburger, 1980: pp. 159–179).
4. Neighborhood effect. The type of neighborhood in which a pupil grows up is one of the most effective predictors of a pupil's educational performance (Webber and Butler, 2007). The neighborhood effect is partly based on the social composition of the residential population. But it also consists of close social or cultural relations, shared values, a sense of belonging to a local community, and social control within the community. Emotional attachment to a place, the feeling of safety among 'our own' people, the support of strong social networks, and the fear of becoming an outsider in schools located in other areas seem to be strong motives for parents from ethnic minorities and working class families to keep their children in low-performing schools in their own neighborhood even if they are free to opt for a school in another area (Bunar, 2010: p. 142).
5. Organizational structure, governance, and equipment of schools. The size of classes (student–teacher ratio), the provision of a school with teaching aids (library and laboratories), the governance, ethos, and reputation of a school attract and keep skilled teachers and influence the quality of their teaching and as a consequence the interest of pupils in special topics. Contrary to the opinion of Gulson (2007: p. 1381) and others, a larger school is not necessarily a better school. A larger curricular and extracurricular choice does not automatically imply that the quality of educational provision or the educational performance of pupils improves. Organizational structures of schools should not be studied in isolation from other factors mentioned here.
6. Spatial variations in the provision of higher education. Urban agglomerations provide a greater variety of secondary schools and institutions of higher education than do peripheral rural areas. For upper class parents, the distance to school has no or little impact on their choice of secondary schools, whereas long distances to the next secondary schools deter parents from lower social strata from sending their children to these schools, especially in ethnically segregated areas and in areas where free school transport is not available.
7. Employment patterns, diversity, and qualification structure of jobs offered in a town or region. A highly differentiated economy, the availability of many highly skilled jobs, and high vertical social mobility influence the occupational and educational aspirations of children by offering them more attractive role models and more economic prospects than a lower skilled and less differentiated labor market would do.
8. Political control, power relations, and state intervention. In the eighteenth- and nineteenth-century Europe, territories under a strong feudal control with a late abolition of serfdom had much lower school attendance rates and became literate much later than did areas with an early introduction of democratic elements and prospects of vertical social mobility. The nineteenth and twentieth centuries offer many examples of harsh state intervention. Between 1740 (South Carolina) and 1838 (Ohio), a number of states in the United States adopted compulsory ignorance laws, which excluded 'negro' children from public schools or even stated that any person who taught slaves to write or employed slaves as scribes would be fined or imprisoned for up to 12 months. In the twentieth century, communist regimes prohibited pupils from ideologically unreliable families from attending grammar schools or studying at universities, and fascist states openly restricted the proportion of Jewish students at their universities.

Correlations between single variables and educational attainment often lead to ambivalent and inconsistent results. Most of these factors have mutually reinforcing effects. It is the local combination of multiple deprivations or of multiple privileges that produces a local milieu or knowledge environment entailing low or high educational attainment. As high or low educational attainment tends to occur in many different contexts and for quite different reasons (Webber and Butler, 2007: p. 1241), an investigator needs deep knowledge of the society, culture, economy, and value systems of the neighborhoods, places, or regions in question when interpreting causal relationships.

Institutions of Higher Education

Some of the research issues already discussed can also be applied to universities and other institutions of higher education. However, universities offer additional challenges for geography of education, and they are also in the focus of geography of science. What makes a university attractive for students and scholars? Which factors are responsible for the intellectual ups and downs of universities? Why did some newly founded universities develop successfully, whereas others did not? Why are Nobel Prize winners or other outstanding scholars confined to such a small number of universities? How can a local knowledge environment affect research processes? What impact does a university have on the economy, society, and culture of its town or wider region? What is the role of universities as key actors in the knowledge economy? How do transnational knowledge networks of scientists evolve, and what long-term consequences do they have? What factors influence the spatial mobility of scholars and students? These questions indicate that the generation and diffusion of scientific knowledge is influenced by local milieus and spatial relations; that scientific practices vary from place to place; that social environments have an impact not only on generating, evaluating, and legitimizing new scientific knowledge, but also on the reading and interpretation of texts; that universities are not simply locations but social spaces, epistemic venues, and knots

of scientific networks (Livingstone, 1995; Meusburger and Schuch, 2012).

What does a scientific milieu or knowledge environment consist of? Scholars and students depend on financial resources, libraries, laboratories, and buildings; their goals and fields of interaction are influenced by evaluators, role models, critical audiences, institutional regulations, informal expectations of their social environment, and a multitude of other factors. The knowledge environment of places can have a decisive impact on the research interests of scholars, their motivation and acquisition of professional competencies, their integration into important networks, and the encouragement they receive from key people in a particular field of study. However, a scientific milieu is not an independent variable that has predictable effects. It is rather a local potential or an offer of chances, stimuli, support, and challenges, which will be taken advantage of by some scholars and ignored by others. Whether and how a person makes use of a milieu's potential depends on his or her goals, cognitive abilities, and social relationships.

Universities differ with regard to the career histories, regional provenance, and geographical mobility of their professors and lecturers, the social origin and catchment areas of their students, and the spatial dimension of their academics' networks and scientific cooperation. The higher the scientific reputation of a university and the larger its resources, the more it acts as a magnet of talent and brains, the more it will be able to attract scholars from prominent other universities, the more international will be its catchment area of outstanding students, the more it will disseminate ideas and scholars all over the world, the more professorships are offered to its young researchers by other universities, and the more impressive are its global networks. However, the spatial expanse of scientific relations and the areas from which a university draws its professors and students can also be severely influenced by political conflicts (e.g., Cold War, Iron Curtain), language barriers, as well as the economic and political situation in the home countries of scholars and students (Meusburger and Schuch, 2012).

Since medieval times, scholars and students have been expected to be geographically mobile for various reasons. First, mobility can inspire new ways of thinking and bring about new relationships and cooperation in the world of academia. Scholars should gather personal experience in different departments, tackling a variety of challenges and familiarizing themselves with issues and methods they have not yet encountered at their present place of work.

Second, in order to remain competitive, a university's knowledge environment must continually be replenished by the best available talent, who can bring in new ideas from elsewhere. Faculties systematically favoring their own former students for professorships run the risk of narrowing their scientific perspectives, and eventually diminishing the quality of their scientific performance. An academic who has worked successfully at other universities generally has a greater range of experience, more extensive personal networks, a broader methodological spectrum, and a better academic reputation than one who has worked in only one institutional setting.

Third, the career mobility of a scholar can be used as one of many indicators of his or her scientific potential. If two or more faculties have confirmed that the scientific achievements, competence, and other merits of a scholar warrant that person's appointment to a position as lecturer or professor, then the uncertainty about that person's scientific potential is much lower than if confirmation came only from the same faculty where the candidate had studied and graduated.

Fourth, creative concepts, innovative methods, and revolutionary scientific findings that contradict common paradigms of mainstream science are not immediately accepted. In spite of the Internet, it can take years or even decades before new theories or methods are recognized within one's discipline or in adjacent ones. The surest way to transmit epochal scientific findings and methods from place A to place B is the geographic mobility of scholars who possess the relevant knowledge.

Further Perspectives

It would be futile to establish a list of emerging fields in geography of education. However, the development of international student flows and the impact of alumni networks on economic and political decisions; the impact of political ideologies on educational systems and educational priorities; the varying role of the meritocratic principle in different societies and its consequences for economy and society; the relevance of educational inequalities in political discourses; the increasing internationalization, neoliberalization and marketization of higher education and its consequences; the schooling of underprivileged ethnic minorities (e.g., Roma in Europe) and of pupils with special needs; the definition of knowledge environments and their impact on learning processes; spatial differences of philanthropic support of education; and many other questions need to be studied more intensively.

See also: Academic Achievement Motivation, Development of; Bilingual Education: International Perspectives; Class Size; Cross-Cultural Study of Education; Cross-National Comparisons in Education: Findings from Pisa; Cultural Capital and Education; Desegregation; Education, History of; Educational Institutions, History of; Educational Research, The Field of; Equality of Opportunity; Equity and Education; Ethnic Conflicts; Ethnic Identity, Psychology of; Ethnicity and Educational Achievement; Families as Educational Settings; Family and Schooling; Globalization in Education; Higher Education Market; Ignorance, History of Concept; Illiteracy, Sociology of; Inequality: Comparative Aspects; Knowledge, Geography of; Literacy and Illiteracy, History of; Multicultural Education; Neighborhood Effects; Race and Academic Motivation; Residential Segregation in the United States; Residential Segregation: Geographic Aspects; School Achievement as Moderated by Sociodemographic Status: The US Case; School Achievement: Motivational Determinants and Processes; School Choice; School as a Social System; Schooling: Impact on Cognitive and Motivational Development; Schooling: Total Impact of; Science, Geography of; Social Inequality and Schooling; Socialization and Education: Theoretical Perspectives; Teacher Motivation; Youth, Geography of.

Bibliography

Ball, S.J., Vincent, C., 2007. Education, class fractions and the local rules of spatial relations. Urban Studies 44, 1175–1189.

Bondi, L., Matthews, M.H. (Eds.), 1988. Education and Society: Studies in the Politics, Sociology and Geography of Education. Routledge, London.

Bunar, N., 2010. The geographies of education and relationships in a multicultural city. Enrolling in high-poverty, low-performing urban schools and choosing to stay there. Acta Sociologica 53, 141–159.

Dupin, C., 1826. Carte figurative de l'instruction populaire de la France (Thematic Map of Popular Education in France). Jobard, Paris.

Fletcher, J., 1849. Moral and educational statistics of England and Wales. Journal of the Statistical Society of London 12, 151–176, 189–335.

Freytag, T., 2003. Bildungswesen, Bildungsverhalten und kulturelle Identität. Ursachen für das unterdurchschnittliche Ausbildungsniveau der hispanischen Bevölkerung in New Mexico (Educational System, Educational Attainment and Cultural Identity. Reasons for the below Average Educational Achievement of Hispanics in New Mexico). Heidelberger Geographische Arbeiten 118. Department of Geography, Heidelberg.

Furet, F., Ozouf, J. (Eds.), 1977. Lire et ecrire. L'Alphabétisation des Français de Calvin à Jules Ferry (Reading and Writing. The Alphabetization of the French Population from Calvin to Jules Ferry), vol. 2. Les Editions de Minuit, Paris.

Gamerith, W., 2006. Ethnizität und Bildungsverhalten. Ein kritisches Plädoyer für eine "Neue Kulturgeographie" (Ethnicity and educational attainment. A critical appeal for a "New cultural geography"). In: Kempter, K., Meusburger, P. (Eds.), Bildung und Wissensgesellschaft. Heidelberger Jahrbücher, vol. 49. Springer, Heidelberg, Berlin, pp. 309–332.

Geipel, R., 1966. Angewandte Geographie auf dem Feld der Bildungsplanung (Applied geography on the field of educational planning). In: Monheim, F., Beuermann, A. (Eds.), Tagungsbericht des Deutschen Geographentages 1965 in Bochum. Steiner, Wiesbaden, pp. 448–457.

Giband, D., Lacquement, G. (Eds.), 2007. La ville et ses marges scolaires: Retour d'expériences sur l'éducation prioritaire et la rénovation urbaine en France et à l'étranger (The City and Its Educational Margins: Lessons Learned in Priority Education and Urban Renewal in France and Abroad). Presses Universitaires de Perpignan, Perpignan.

Gibbons, S., Telhaj, S., 2007. Are schools drifting apart? Intake stratification in English secondary schools. Urban Studies 44, 1281–1305.

Gordon, I., Monastiriotis, V., 2007. Education, location, education: a spatial analysis of English secondary school public examination results. Urban Studies 44, 1203–1228.

Gulson, K.N., 2007. Repositioning schooling in Inner Sydney: urban renewal, an education market and the 'absent presence' of the 'middle classes'. Urban Studies 44, 1377–1391.

Hamnett, C., Ramsden, M., Butler, T., 2007. Social background, ethnicity, school composition and educational attainment in East London. Urban Studies 44, 1255–1280.

Hanson Thiem, C., 2009. Thinking through education: the geographies of contemporary educational restructuring. Progress in Human Geography 33, 154–173.

Heffernan, M., 1989. Literacy and geographical mobility in nineteenth century provincial France: some evidence from the Département of Ille-et-Vilaine. Local Population Studies 42, 32–42.

Holloway, S.L., Hubbard, P., Jöns, H., Pimlott-Wilson, H., 2010. Geographies of education and the significance of children, youth and families. Progress in Human Geography 34, 583–600.

Holloway, S.L., Jöns, H., 2012. Geographies of education and learning. Transactions of the Institute of British Geographers 37, 482–488.

Hoyler, M., 1996. Anglikanische Heiratsregister als Quellen historisch-geographischer Alphabetisierungsforschung (Anglican marriage registers as sources of historical-geographical research on literacy). Heidelberger Geographische Arbeiten 100. Department of Geography, Heidelberg, pp. 174–200.

Jahnke, H., 2005. Der italienische Mezzogiorno auf dem Weg in die europäische Wissensgesellschaft. Eine Untersuchung der Erwerbssituation und der regionalen Mobilität junger Akademiker am Beispiel Siziliens (The Italian Mezzogiorno on Its Way to the European Knowledge Society. An Analysis of the Employment Situation and Regional Mobitlity of Young University Graduates of Sicily). Berliner Geographische Arbeiten 101. Department of Geography, Humboldt University, Berlin.

Kramer, C., 1993. Die Die Entwicklung des Standortnetzes von Grundschulen im ländlichen Raum: Vorarlberg und Baden-Württemberg im Vergleich (The Location Development of Elementary Schools in Rural Areas. The Example of Vorarlberg and Baden-Württemberg). Heidelberger Geographische Arbeiten 93. Department of Geography, Heidelberg University, Heidelberg.

Livingstone, D., 1995. The spaces of knowledge: contributions towards a historical geography of science. Society and Space 13, 5–34.

Marsden, W.E., 1987. Unequal Educational Provision in England and Wales: The Nineteenth-Century Roots. Woburn, London.

Matthews, H., Limb, M., 1999. Defining an agenda for the geography of children: review and prospect. Progress in Human Geography 23, 61–90.

Mayr, A., Nutz, M. (Eds.), 2002. Bildung und Kultur. Nationalatlas Bundesrepublik Deutschland (Education and culture. National atlas of the Federal Republic of Germany), vol. 6. Spektrum Akademischer Verlag, Heidelberg, Berlin.

Meusburger, P., 1980. Beiträge zur Geographie des Bildungs- und Qualifikationswesens. Regionale und soziale Unterschiede des Ausbildungsniveaus der österreichischen Bevölkerung (Contributions to a Geography of Knowledge and Skills. Regional and Social Inequalities of Educational Achievement among the Austrian Population). Innsbrucker Geographische Studien 7. Department of Geography, Innsbruck University, Innsbruck.

Meusburger, P., 1998. Bildungsgeographie. Wissen und Ausbildung in der räumlichen Dimension (Geography of Education. Knowledge and Learning in the Spatial Dimension). Spektrum Akademischer Verlag, Heidelberg.

Meusburger, P., Schuch, T. (Eds.), 2012. Wissenschaftsatlas of Heidelberg University. Spatio-Temporal Relations of Academic Knowledge Production. Bibliotheca Palatina, Knittlingen.

Schmude, J., 1988. Die Feminisierung des Lehrberufs an öffentlichen, allgemeinbildenden Schulen in Baden-Württemberg. Eine raum-zeitliche Analyse (The Feminization of the Teaching Profession at Public Elementary Schools in Baden-Württemberg. A Spatio-Temporal Analysis). Heidelberger Geographische Arbeiten 87. Department of Geography, Heidelberg.

Taylor, C., 2009. Towards a geography of education. Oxford Review of Education 35, 651–669.

Waters, J.L., 2006. Emergent geographies of international education and social exclusion. Antipode 38, 1046–1068.

Webber, R., Butler, T., 2007. Classifying pupils by where they live: how well does this predict variations in their GCSE results? Urban Studies 44, 1229–1253.

Education, History of

Jurgen Schriewer, Humboldt University, Berlin, Germany
António Nóvoa, Universidade de Lisboa, Lisboa, Portugal

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 4317–4323, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

While education as a social practice has occurred at all times and social settings, this article focuses, in Section The Rise of the Modern Educational System, on a crucial stage in the specifically modern transformation of that practice, namely its crystallization, from the late eighteenth century onward into fully institutionalized and increasingly inclusive 'national systems.' This process of educational system formation gave rise to the emergence not only of educational theory more generally, but also of history of education as a particular field of study. As a consequence of these origins, the field has oscillated up to the recent past between education and history, i.e., between systemic self-reflection committed to educational ideals and reform and more detached research committed to the rules of historical method. Section History of Education in Connection with Systemic Self-Reflection examines the institutional conditions and disciplinary models that have contributed to shaping the history of education into either an adjunct to educational theory or a specialized branch of historical study. Finally, Section History of Education as a Field of Social Science Study reviews the important renewal and wide-ranging diversification the field has experienced from the 1960s onward as a consequence of the influx of concepts and methods stemming from innovations both in social and intellectual history and in the social sciences more generally.

'History of education' is an ambiguous term in at least two respects. First of all, the term has come to designate different levels of reality. On the one hand, it refers to education as a universal social practice and to the manifold forms and successive transformations that this practice has experienced over time. On the other hand, it refers to a field of intellectual activity, namely a branch of historical study, devoted to describing, interpreting, and analyzing this practice. Obviously, both levels are not disconnected from each other. Rather, it was a crucial stage in the transformation of the former, namely the crystallization of education, into fully institutionalized and increasingly inclusive 'national education systems,' which gave rise to the emergence, not only of educational theory in general but also of the historical study of education in particular. This historical connection, however, gave rise to yet another ambiguity, an ambiguity inherent in the emergent field of study as such. From its very origins, this field has oscillated between different disciplinary orientations, namely education and history, i.e., between systemic self-reflection committed to educational ideals and reform, and more detached research committed to the rules of historical method. Against this background, the article briefly describes the process of educational system formation and some of its key features (Section The Rise of the Modern Educational System) in order then to set out the institutional conditions and disciplinary models that have contributed to shaping the history of education into either an adjunct to educational theory (in terms of, e.g., *Geschichte der Pädagogik* or *histoire de la pédagogie*) or a specialized field of historical study (in terms of, e.g., *Historische Bildungsforschung* or *histoire 'historienne' de l'éducation*) (p. 2). The article then goes on to describe in greater detail the important renewal the field has experienced, especially from the 1960s onward, as a consequence of the influx of concepts and methods stemming from innovations both in social and intellectual history and in the social sciences more generally (Section History of Education as a Field of Social Science Study).

The Rise of the Modern Educational System

Although education has occurred, and continues to do so, in manifold forms and at all times – in families, through catechizing, at the workplace, in youth clubs and organizations, or in borstals – education in modern times has by and large become institutionalized in the form of large-scale public systems of education. These constitute the modern type of educational organization *par excellence* (Thomas et al., 1987). Processes of the formation and eventual consolidation of 'the modern educational system,' while starting toward the end of the eighteenth century, came to completion only toward the turn from the nineteenth to the twentieth century (Müller et al., 1987). Moreover, educational system formation was only possible as a component of more encompassing processes of social transformation, and of concomitant processes of state and nation-building, which led to the specifically modern pattern of societal organization explained as 'functional differentiation' by recent social theory (Luhmann, 1982). Drawing on related concepts, one may characterize educational system formation by key features such as *inclusion*, *nationalization*, and the system's *internal articulation* as well as *systemic self-reflection*.

Inclusion refers to one of the central aspects of modern social structures, namely the access of a population as a whole to all of the functionally specific subsystems constituted at the level of society. More precisely, inclusion means the universalization, not of so-called 'service roles' such as those of a politician, a physician, or a teacher, but of the corresponding 'service-receiving roles' specific to each of society's subsystems, namely the roles of the electorate, patients, or pupils, respectively. Thus, the educational correlate to the inclusion process at the level of society at large was, in terms common to present-day debates on education and development, the 'universalization of primary education.' This process started in the Protestant countries of central and northern Europe as early as the

seventeenth century; was largely achieved, in most of western Europe, by the end of the nineteenth century; and came to an end, in some countries of southern Europe, Latin America, and Asia, only in the middle of the twentieth century. And yet, it still constitutes an unmet challenge in large parts of the Southern Hemisphere. At the level of national legislation, important steps toward universalizing primary education often went hand in hand with measures undertaken to assure greater inclusion also in the political and legal systems. As illustrated by the historical experience, not only of England and Italy, but in varying forms also of France and Prussia or to some extent Russia, such measures consisted in, e.g., the extension of the franchise or, more generally, steps designed to turn hitherto dependent subjects into national citizens. At the local level, however, national school laws were gradually implemented as an answer to the radically altered communication requirements that were attendant on social structures increasingly characterized by new patterns of differentiation and ensuing processes of rationalization, mobilization, and industrialization. In this sense, the process of inclusion in the emerging educational system corresponded to the gearing of that system to its actual function, for this function consists precisely in laying the foundations of the capability for universal communication across all subsystems of society. In sociohistorical terms, thus, it was essentially the achievement of universal mass schooling that marked the decisive transition toward the constitution of institutionalized education as a modern 'system.'

Nationalization refers to another process that came to fruition from the eighteenth century onward. To the extent that the traditional union of schools and church(es) gave way to the differentiation between education and religion, which were more and more conceived of as distinct spheres of social action and meaning, it was the increasingly rationalized and secularized nation-state that was thought to serve as the natural support and guardian of the education of its citizens. This view was underscored by the fact that the eventual achievement and continuous running of an inclusive educational system presupposed organizational resources and required an administrative infrastructure that no corporate social actor other than the state was deemed able to provide. Alternatively, the state and state-supported agencies were ready to use compulsory mass schooling also for pushing through the realization of more general state-related goals such as social mobilization and equalization, national integration and collective identity-building, or political socialization and indoctrination. Thus, whether in Europe, in the Americas, or in Asia, the realization of an all-inclusive educational system by way of universalizing compulsory mass schooling went hand in hand with the increasing intervention of central state agencies and the building of comprehensive administrative frameworks. Education ministries were established for securing the financial resources and building up the administrative infrastructure necessary in order not only to guarantee generalized supply but also to successfully run standardized networks of institutionalized schooling. Just like the realization of compulsory mass schooling, this process took more than a century to come to completion. It was a long way, indeed, from late-eighteenth-century programs of establishing a coherent state-run scheme of educational institutions modeled on Catholic Church organization (in the pioneering Italian principalities of Parma and Piedmont and, subsequently, in imperial France) to the all-encompassing organization of institutionalized schooling as a 'public legal organism' (as this accomplishment was acclaimed in late-nineteenth-century social analysis).

The emergent system's *internal articulation* was the inevitable complement to inclusion. Since education systems in Europe were not created in a vacuum, their system formation processes started from quite an array of existing institutions that were stemming, on their part, from premodern establishments such as medieval cathedral schools and Renaissance universities, church-run colleges and academies for the sons and daughters of the nobility, endowed grammar schools and *Gymnasien* established by enlightened princely rulers, as well as municipal petty schools and the nineteenth-century forms of drawing schools, science schools, higher-grade, technical, and vocational schools. It was, then, a significant feature of the differentiation process that, during the initial stage of inclusion and due to such origins, the emergent educational systems were rather decentralized, uncoordinated, and open in their organization. The more the continuing process of inclusion advanced, however, the more acute the need for internal specification and rational articulation became apparent. Specification, on the one hand, implied the removal of ambiguities and unregulated competition between institutions that were to become clearly defined alternative tracks or types of schooling. The systems' articulation, on the other hand, raised the issue of connecting hitherto uncoordinated institutional components with the aim of constructing a rational 'sequence,' or 'ladder,' of successive levels of education. The outcomes of such processes of defining educational systems' institutional structures can then be conceived of as ways of coping with divergent expectations and requirements: with inclusion (by organizing universal access to primary schooling for an entire age-group) and selection (by consecutively channeling pupils into different tracks and levels), or with qualifying for universal communication across all fields of society and inculcating differential 'habitus' as appropriate to particular groups, trades, and areas of activity. As a consequence, the issue of defining an educational system's career sequence has always been closely associated with a nation's public policy in general and with its social and economic policy in particular. This applied especially to the construction of sequence between the 'lower' or vernacular schools (e.g., the so-called 'English' or 'German' schools) and the 'learned' or Latin-based establishments, or between, to take up a graphic phrasing coined by Prost (1968), *l'école du peuple* and *l'école des notables*. Thus, from the late-nineteenth-century 'systematization' of the major western European education systems over the definition, during the first half of the twentieth century, of examination-based transition procedures from the primary to the secondary level of education up to the successive introduction, from the 1960s onward, of comprehensive types of schooling in many, but not all, countries throughout the world, contrasting models and programs for educational sequencing have been at the very core of educational policy debate. At the same time, these models and programs have fostered again and again the production of both educational reflection – particularly reformative reflection – and educational research.

History of Education in Connection with Systemic Self-Reflection

It is a significant feature of modernity that in subsystems of society that have crystallized in connection with the far-reaching transformation processes explained in terms of 'functional differentiation,' system-specific self-descriptions intensify to such an extent that they take the form of explicitly stated reflection theories. This common pattern of self-referential theories of a societal subsystem developed within that system holds true also for educational theorizing pursued within the emergent educational system (Luhmann, 1982). The tremendous increase, from the late eighteenth century onward, in the production of educational ideologies, educational theories, and educational knowledge more generally has to be seen in connection with the growing acceptance of programs and policies fueling the establishment of state-run 'national systems of education.' In particular, the conceptual model of systemic self-reflection intrinsic to the major subsystems of society makes intelligible the institutionalization of educational theory as a field of academic study in close temporal and ideological connection with the passing of the decisive laws on compulsory schooling in most of the western European nations. The historically unprecedented scale of universal mass education, its inclusive character, and, especially in the Catholic countries of southern Europe, its secularized organization not only presupposed an administrative infrastructure of a previously unknown magnitude, but also required the intensification of reflection in new and more sophisticated forms. Moreover, the transition to universal mass education also entailed the need for public legitimization of the new system, not least in the form of throwing into relief distinctive national traditions of education and schooling. Salient examples include the foundation of the *Musée Pédagogique* in France (1879) and of its counterpart in Spain (1882), which were meant to serve as depositories of their respective nation's educational memoirs; the virtual invention of national educational tradition in monumental works such as Gabriel Compayré's *Histoire critique des doctrines de l'éducation en France* (1879) and Ferdinand Buisson's *Dictionnaire de pédagogie* (1887–88); or the launching of the *Monumenta Germaniae Paedagogica* meant to collect 'the materials, scattered over archives and libraries, that are relevant to the history of German education and schooling' (1886–1938).

In addition, the necessities of systematic teacher training and attendant considerations of social utility and responsibility impinged heavily on the emergent forms of the historical interpretation of educational traditions. These necessities and considerations favored a genre of educational history that remained subordinated to the categories and problematiques of systematic – i.e., philosophico-reflective – theorizing, a genre in other words that gave preference to narratives of the great educators and philosophers of the past, and to interpretations of their respective doctrines realized in terms of normative hermeneutics. Interestingly enough, this genre has been common to most European countries. It can be traced from early British works over, for instance, Dutch and Scandinavian examples as well as the impressive range of, in particular, German and Italian histories of educational thought up to recent specimens in quite a number of countries (Heinemann, 1979–85). This genre as well as a good many of the studies commonly assigned to the history of education have aimed at producing – and in fact continue to produce – not primarily historical research in the manner of historians, but, instead, justification and legitimization, the construction of enlightened norm consensuses, and the reinterpretation and reactualization with respect to issues of present-day educational reform and of the normative and theoretical substance contained in historical texts, models, or institutions. In other words, this tradition of an 'educational' history of education – of *Geschichte der Pädagogik* or *histoire de la pédagogie* – has not aimed at a detached historicization of lines of educational thought or of past experiences, but at certain forms of theoretical and normative 'traditionalization.'

Thus, seen from the vantage point of sociohistorical analysis, the history of education closely associated, in terms of institutional and disciplinary affiliation, with educational theory has from its origins shared the ambiguities proper to the latter field. Just as educational studies in general, history of education has oscillated between rigorous research and systemic self-reflection. Accordingly, the field has given room to a considerable variety of constructions of reality, the varying presuppositions of which ensue from their being related to distinct system logics and problem perspectives: to either issues of objective method and explanation as debated from the vantage point of social science research, or issues of meliorism and reform immanent in the dynamics of modern education, and to attendant needs for interpretation, orientation, and professional knowledge. The respective proportions, however, in which these component modes of reasoning have given shape to historical studies on education, arise from an intricate interplay of institutional conditions, intellectual traditions, and disciplinary models (Schriewer et al., 1993). Thus, recent analyses of the state of the art have – on the grounds of reviews of substantive interests, research topics, and methodological approaches that are given preference to as well as of the field's disciplinary status, specialized journals, and associations – contrasted two major models. These are, on the one hand, the German model, which is well represented also in countries such as Switzerland, Austria, and Hungary as well as Italy and Scandinavia, and, on the other hand, the institutional-*cum*-epistemological model prevalent in France and the United Kingdom, respectively (Compère, 1995). In the former model, due to a long-standing tradition of educational-*cum*-philosophical theorizing as well as to the institutional and intellectual integration of the history of education into education as a discipline firmly established in academia, the system-reflection style of educational history described above has by and large prevailed up to the 1970s and still has a strong impact up to the present. By contrast, the tradition of education conceived as a philosophico-hermeneutic field of theorizing gave way to educational studies based on the social and human sciences both in France, from the 1930s onward, and in the United Kingdom, in the aftermath of World War II. Consequently, these developments also facilitated the alignment of historical studies on education with history and sociology proper, and the intellectual opening of such studies to wider sociological and anthropological frames of reference. Thus, in France and the United Kingdom, just as in the United States, history of education has turned into an academic discipline in

its own right, which is fairly independent of educational studies in general, based on its own specialized journals, and elaborated by scholars who are (also) trained as historians. This *histoire 'historienne' de l'éducation* has in turn encouraged initiatives leading to the emergence of a social history-related *Historische Bildungsforschung* also in countries traditionally influenced by the German model.

History of Education as a Field of Social Science Study

Fueled by debates on a 'new history of education,' which arose in the United States at the beginning of the 1960s, conceptual models and methodological approaches heavily influenced by modern social history (*see* Social History), intellectual history (*see* Intellectual History), and the social sciences more generally have since then considerably altered the field. The publication, in 1960, of Philippe Ariès' *L'enfant et la Vie Familiale sous l'Ancien Régime* and Bernard Bailyn's *Education and the Forming of American Society* can be seen as a turning point. A stimulating and intellectually demanding period was about to begin. From the 1960s to the 1980s, the 'new history of education' brought a critical perspective into a discipline that had been dominated by parochialism and myths of progress. Ever since, its repertoire has come to include the *longue durée*, a focus on structures and mentalities, a special attention to serial and quantitative methods, and the ambition of 'total history.' These changes, moreover, are reflected in corresponding shifts from narrative history to problem-centered history, from descriptive to explanatory approaches, and from conceptions of unilinear evolution to a vision of historical processes that emphasize their often conflict-laden, asynchronous nature. While accomplishing such shifts, scholars like Michael Katz and David Tyack, Antoine Prost and Dominique Julia, and Lawrence Stone and Harold Silver have been deeply influenced by social history perspectives. They have turned the legitimizing narratives of the past upside down, providing new arguments about the emergence and development of educational systems, the role of schooling in social reproduction, the educational strategies of various social classes, and the action taken by the nation-state in establishing mass schooling. They accordingly shifted the attention from the evolution of educational ideas to the relationships between education and society.

While in the early 1980s, a sense of accomplishment was shared by a majority of the researchers involved in rewriting the history of education, only a few years later, the resources of the 'new' history of education seemed to be exhausted. The accumulation of data, the considerable increase in well-conducted studies based on extended series of archival sources, and the growing specialization of historical inquiry nevertheless left the impression that educational historians had fallen short of addressing actually meaningful issues. In addition, efforts meant to raise new questions, related, for instance, to minority and gender issues, did not lead to an epistemological breakthrough, but rather to an even wider fragmentation of the field. Innovative solutions were to be found, then, in the adoption of theoretical frameworks and methodological approaches that focus not only on the organization of practices and discourses over time, but also on the elucidation of how these practices and discourses have constructed and reconstructed the individual lives of people and, thereby, social reality. In this sense, educational historians are facing the same challenges and dilemmas as social scientists more generally. To the extent that the difficulties of coping with the plurality of meanings henceforth present in the analysis of social phenomena became apparent, the 'new' complexity of historical work has nourished old suspicion as far as theory is concerned (Lagemann, 2000). The transition from 'history' to 'histories' implies a reconfiguration of the field that invites educational historians to reconceptualize their work and to adopt a more comprehensive approach, shifting from a social history of culture to a cultural history of society (Chartier, 1998). Still, there is no consensus regarding the changes that are occurring in historical studies in education (Cohen and Depaepe, 1996). Nevertheless, drawing on an often used conceptual division between educational actors, school practices, educational ideas, and educational systems, it is possible to identify major tendencies that have come to shape historical research in education (Nóvoa, 1998).

The History of Educational Actors and the Rediscovery of Experience

For a long time, historians have described the history of educational institutions rather than that of educational processes or interactions. Consequently, the history of educational actors has been one of the most neglected areas of research. The rediscovery of the concept of 'experience' opened new possibilities, well illustrated by numerous texts on autobiographies, teachers' professional careers, or students' life stories. One of the reasons why feminist theories have been so powerful in educational history is that they help to reveal histories silenced by traditional narratives. From these points of view, modernity dispossessed the actors of the educational process from their subjectivity through the imposition of a structural logic and a populational reasoning. People were classified as categories (the teachers, the learners, the children at risk, etc.) and managed as populations. However, after a long tradition of conceptualizing the world in terms of structure and representation, time had come to see the world as 'experience.' This implied a new 'epistemology of the subject' that, by dissolving the 'natural' divisions between public and private, discourse and practice, or text and context, prepared the way for a shift toward a self-conscious study of pupils and teachers, and for bringing the individual actor in education back into the historical portrayal. Moreover, new conceptions of the subject and of experience allow attention to be paid not only to living experience but also to the way people elaborate it. Finally, the concept of 'experience' has to be seen not only at an individual level but also in its collective dimension, which is characterized by a redefinition of identities and belongings through the participation of people in different communities of meaning.

The History of School Practices and a New Concept of Culture

The history of school practices is another field that has traditionally been neglected by educational historians. Explanations of this 'black box' phenomenon are directly related to the concept of 'culture.' Schools have always been regarded as places of culture: at first, in an idealistic vision of putatively

universal values and knowledge to be transmitted to the young generations; afterward, in a critical perspective of ideological inculcation and social reproduction. In both the cases, educational historians failed to look at the internal production of a distinctive school culture that is not independent of social struggles and conflicts, but cannot be explained as being merely determined by the external world either. More recent conceptualizations, in contrast, have defined school culture as both an ensemble of norms (which organize the branches of knowledge and behaviors that must be brought into schooling) and an ensemble of practices (which allow the transmission of such knowledge and the integration of such behaviors). Moreover, it is possible to identify tendencies of historical research that successfully try to refocus historical inquiry onto the internal functioning of the school by providing thoroughgoing analyses of organizational structures, curriculum issues, school subjects, and day-to-day realities. Thus, one of the most interesting lines of work tries to understand the permanence of a 'grammar of schooling' that resists changes and reforms. Another line of research focuses on the history of the curriculum, viewed as the outcome of a complex interplay of competing values and traditions. Several authors are voicing the need for deconstructing a 'natural history' conception of the curriculum and therefore seek to problematize the curriculum as text and historicize its evolution over time. A third strand of recent research worth mentioning is concerned with the history of school subjects. Historians are asking salient questions regarding the formation of school subjects, the institutionalization of school knowledge, and the ways in which it helps to organize our understanding of the world. Finally, one should not fail to point out a research trend concerned with the 'silence' observed within the traditional history of education about meaning and culture of classrooms, or the everyday education practice (Grosvenor et al., 1999; Berrio, 2000).

From the History of Educational Ideas to a Sociohistory of Knowledge

In contrast to traditional histories of educational ideas, the 'new' cultural history of education is interested in the production, diffusion, and reception of educational discourses throughout time and space. In other words, it is interested in "a historical imagination in the study of schooling that focuses on knowledge as a field of cultural practice and cultural production" (Popkewitz et al., 2001). Therefore, the research interest has shifted to the discursive practices that regulate schooling, especially in situations of conflict or rupture. In this sense, several authors are calling for a historiography, not of the general terms, but of the conceptual structures that determine educational discourses, and for analyses, not of prominent ideas, but of fields of historically defined discursive problematiques. Such an approach is also important for analyses of the roles played by educational experts in providing 'legitimate' or 'scientific' interpretations of school realities. It helps to understand the original institutionalization of educational studies, not as a narrative of intellectual progress, but as a form of 'governmentality.' The question is not how to identify different ideas, but how to apprehend the rise of expert systems of knowledge (Drewek et al., 1998). The same holds true for the analysis of the complex transformation procedures that change academic subjects (integrated into their distinctive conceptual spaces) into school subjects (integrated into the arena of schooling), or for the study of the production and diffusion of textbooks that are not simply 'delivery systems' of factual information, but the results of cultural constructions imposing particular ways of seeing and interpreting the world. In this sense, knowledge has increasingly come to be taken as a text that redefines subjectivity and identities, elaborates rules and behaviors, and configures meanings and beliefs.

Comparative-Historical Research on Education

The publication of Fritz Ringer's *Education and Society in Modern Europe* (1979) heralded a virtual breakthrough of yet another paradigmatic approach to the field, namely the comparative-historical research on education. Although some forms of cross-national analysis or international account had not completely been missing in previous periods, the programmatic views expressed by Ringer laid claim to new and ambitious objectives: "There is simply no other means of arriving at explanations, and not just descriptions, of change in education than the comparative approach." Substantive research inspired by such views has remarkably increased from the 1970s onward. It covers, for instance, quasi-experimental investigations of the relationships between education and economic growth, large-scale theory-based analyses of the varying processes of educational system formation, as well as cross-national analyses of the emergence and varying discipline formation patterns of education as an academic field of study in the context of European university systems and the emerging social sciences more generally. Comparative-historical research has also shed new light on the role that nation-specific social meanings, widely accepted interpretative schemes, or 'symbolic structures' deeply rooted in a particular civilization have played in shaping distinctive academic cultures and educational institutions (Archer, 1979; Müller et al., 1987; Schriewer et al., 1993).

Not always, however, have attempts at providing convincing explanation through the combination of historical research and cross-national, or cross-civilizational, analysis been successful. Consequently, much debate has been devoted to issues of explanation (*see* Comparative History). While traditionally, and to some extent up to the present, mainstream models of explanation have followed lines of reasoning as defined in the orthodox philosophy of science, far-reaching shifts in both problem awareness and conceptual models have taken place in the wake of recent theory developments (Schriewer, 2000). Among these are, for instance, world-system models and analyses that, by drawing on Braudelian concepts, not only reintroduced a macrohistorical perspective into comparative research, but also tended to replace the latter by analyses of global diffusion processes and of the working of an emerging 'world polity' (Thomas et al., 1987). New forms of reconciling history and comparison are also suggested, however, by insights from historical sociology as well as by interdisciplinary developments discussed under headings such as 'self-organization,' 'morphogenesis,' or 'evolutionary complex systems research' (Gulbenkian Commission, 1996). Thus, comparative-historical research on education informed by concepts and models of this kind seems to be particularly suited, not only to radically historicizing

educational phenomena, but also to linking historical analysis up with insight relevant to social theory.

See also: Anthropological Approaches to the Study of Education: The United States and Beyond; Education (Primary and Secondary Schools) and Gender; Education and Economic Growth; Education, Economics of; Educational Institutions, History of; Life Course in History; Nationalism, Historical Aspects of: The West.

Bibliography

Archer, M.S., 1979. Social Origins of Educational Systems. Sage, Beverly Hills, CA.

Ariès P.L., 1960. Éducation et la Vie Familiale sous l. Ancien Régime, Plon, Paris.

Bailyn, B., 1960. Education and the Forming of American Society—Needs and Opportunities for Study. Vintage Books, New York.

Berrio, J.R. (Ed.), 2000. La Cultura Escolar de Europa. Tendencias Históricas Emergentes. Biblioteca Nueva, Madrid, Spain.

Buisson, F. (Ed.), 1887–88. Dictionnaire de Pédagogie et D'Instruction Primaire. Hachette, Paris, 2 parts in 5 vols.

Bowen, J., Bowen, J., Bowen, J., Bowen, J., 2003. A History of Western Education. Routledge, London.

Chartier, R., 1998. Au Bord de la Falaise: L'histoire Entre Certitudes et Inquiétude. Albin Michel, Paris.

Cohen, S., Depaepe, M. (Eds.), 1996. History of Education in the Postmodern Era. C.S.H.P., Ghent, Belgium (Paedagogica Historica 32 (2)).

Compayré, G., 1879. Histoire Critique des Doctrines de L'Éducation en France depuis le Seizième Siècle. Hachette, Paris (Reprint edition 1970. Slatkine, Geneva).

Compère, M.M., 1995. L'Histoire de l'Éducation en Europe. Peter Lang, Bern, Switzerland.

Drewek, P., Lüth, C., Aldrich, R., Scholtz, H., Schriewer, J., Tenorth, H.E. (Eds.), 1998. History of educational studies. Geschichte der Erziehungswissenschaft. Histoire des Sciences de l'éducation. Paedagogica Historica, Supplementary Series, vol. III. C.S.H.P, Ghent, Belgium.

Dekker, J.J.H., 2010. Educational Ambitions in History: Childhood and Education in an Expanding Educational Space from the Seventeenth to the Twentieth Century. Lang, Frankfurt am Main.

Duke, B.C., 2009. The History of Modern Japanese Education: Constructing the National School System, 1872–1890. Rutgers University Press, New Brunswick, NJ.

Grosvenor, I., Lawn, M., Rousmaniere, K. (Eds.), 1999. Silences and Images. The Social History of the Classroom. Peter Lang, New York.

Gulbenkian Commission, 1996. Open the Social Sciences. Report on the Gulbenkian Restructuring of the Social Sciences. Stanford University Press, Stanford, CA.

Heinemann, M. (Ed.), 1979–85. Die historische Pädagogik in Europa und den USA, 2 vols. Klett-Cotta, Stuttgart, Germany.

Lagemann, E.C., 2000. An Elusive Science: The Troubling History of Education Research. University of Chicago Press, Chicago.

Luhmann, N., 1982. The Differentiation of Society (S. Holmes, C. Larmore, Trans.). Columbia University Press, New York.

Monumenta Germaniae Paedagogica. Schulordnungen, Schulbücher und pädagogische Miscellaneen aus den Landen deutscher Zunge, 62 vols, 1886–1938. Weidmann, Berlin. Launched by Kehrbach K, edited by Gesellschaft für deutsche Erziehungs- und Schulgeschichte.

Müller, D.K., Ringer, F.K., Simon, B. (Eds.), 1987. The Rise of the Modern Educational System. Structural Change and Social Reproduction 1870–1920. Cambridge University Press and Editions de la Maison des Sciences de l'Homme, Paris, Cambridge, UK.

McCulloch, G., 2011. The Struggle for the History of Education. Routledge, London.

Nóvoa, A., 1998. Histoire et Comparaison. Essais sur l'Éducation. Educa, Lisbon, Portugal.

Popkewitz, T., Franklin, B., Pereyra, M. (Eds.), 2001. Cultural History and Education: Critical Essays on Knowledge and Schooling. Routledge, New York.

Prost, A., 1968. Histoire de l'Enseignement en France 1800–1967. Armand Colin, Paris.

Patterson, J.T., 2001. Brown v. Board of Education: A Civil Rights Milestone and Its Troubled Legacy. Oxford University Press, Oxford.

Ringer, F.K., 1979. Education and Society in Modern Europe. Indiana University Press, Bloomington, IN.

Reese, W.J., 2007. History, Education, and the Schools. Palgrave Macmillan, New York.

Roest, B., 2000. A History of Franciscan Education (c. 1210–1517). Brill, Leiden, the Netherlands.

Schriewer, J. (Ed.), 2000. Discourse Formation in Comparative Education. Peter Lang, New York.

Schriewer, J., Keiner, E., Charle, C. (Eds.), 1993. Sozialer Raum und Akademische Kulturen/A la recherche de l'espace universitaire européen. Peter Lang, New York.

Thomas, G.M., Meyer, J.W., Ramirez, F.O., 1987. Institutional Structure. Constituting State, Society, and the Individual. Sage, Newbury Park, CA.

Thelin, J.R., 2004. A History of American Higher Education. Johns Hopkins University Press, Baltimore.

Urban, W.J., Wagoner, J.L., 2009. American Education: A History. Routledge, New York.

Education in Latin America and the Caribbean: Systems and Research

Dante J Salto, University at Albany, State University of New York, Albany, NY, USA; and Universidad Nacional de Córdoba, Argentina

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Post, volume 12, pp. 8395–8399, © 2001, Elsevier Ltd.

Abstract

The reality of educational systems and research in Latin America and the Caribbean (LAC) depicts a complex, heterogeneous panorama that includes great disparities across the region and within each country. LAC education greatly differentiates by support from government measured by investment in education, coverage through enrollments, and quality through regional and international tests. When considered these indicators, LAC locates in the mid tier of developing regions in the world. Educational research follows similar patterns. Research quality varies greatly but there are notable improvements, as well as increased production. A handful of countries dominate the scene in production of research in education.

Introduction to Latin American Educational Systems

LAC Educational Systems: Between Complexity and Heterogeneity

Educational systems in Latin America and the Caribbean (LAC) have a longstanding tradition going back to the formation of their nation-states. Education became a default feature of nation-states in LAC, incorporated in their constitutions after independence. In this phase, most educational institutions were public and conceived of as part of the civil service, teachers and principals being public servants and curricular content being determined top-down hierarchically (Schwartzman and Cox, 2010).

Although it is possible to refer to common features of LAC educational systems, the reality depicts a complex, heterogeneous panorama that includes great disparities within the region and within each country measured against most social, economic, and development indicators, including educational measures. LAC education greatly differentiates in support from government measured by the share of Gross Domestic Product (GDP) to education, coverage through enrollments, and quality through regional and international tests.

LAC countries were instrumental in the early development of their educational systems, in giving a priority to primary education in order to educate citizens of the newly formed nation-states. In this sense, in the beginning of the twentieth century, countries such as Argentina increased enrollment rates in primary education to reach a threshold close to universalization of that level. Today, primary education is legally compulsory in all LAC countries. A once elite-oriented secondary education is facing a transformation and various countries are extending compulsory education at least to the lower secondary education level as well. Some countries are advancing in policies toward including preprimary and upper secondary school as compulsory too (SITEAL, 2010). Higher education (university and nonuniversity levels) has already achieved thresholds of massification (defined as enrolling 15% of the age cohort) in most countries and universalization (defined as enrolling 50% of the age cohort) in a few of the countries. Still the region faces challenges related to retention, illiteracy, and overall quality in its educational systems.

Despite the overall regional trends and certain shared features, LAC is a region characterized by enormous differences among countries in terms of educational development. UNESCO's Education for All (EFA) Development Index (EDI) provides a portrait of the status of basic education in that region encompassing four indicators: (1) primary education net enrollment ratio, (2) adult literacy, (3) gender parity and equality, and (4) quality of education calculated as survival rate. In 2007, only six LAC countries ranked 'high' in this global index: Cuba, Argentina, Uruguay, Chile, Mexico, and Venezuela. Fourteen LAC countries ranked in the middle, and one LAC country, Nicaragua, fell into the lowest category of the index. The fact that most LAC countries have only reached a medium level in the EDI composite means that they have not met minimum benchmarks, showing poor levels in adult literacy and quality indicators (e.g., Guatemala) and low retention rates in countries such as Brazil, the Dominican Republic, El Salvador, and Guatemala (UNESCO, 2010).

Another measure of assessment is the allocation of public budget to education. UNESCO and its EFA partners have consistently advocated an increase in the allocation of funding as measured by the share of GDP (4–6%) allotted to education (UNESCO, 2006). Most regions worldwide allocate above 4% of their GDP to education. Only high income countries reach averages close to 6%, including all European Union countries, the United States, and Canada. On average, LAC countries spend 4.75% – just above developing countries in other regions such as Middle East and North Africa (4.69%), East Asia and Pacific (4.44%), and Sub-Saharan Africa (4.33%). South Asian countries lag behind, allocating only 2.85%. This measure also shows great heterogeneity within LAC region. By 2009, 12 countries invested at least 4% of their GDP in education, with countries such as Cuba, Bolivia, Costa Rica, and Argentina investing at least 6% of their GDP. Those countries along with Brazil and Mexico have increased their investments by at least 1.5 points in a decade. However, eight other countries languish with public expenditure in education lower than 4%, with specific cases, such as the Dominican Republic, Guatemala, Uruguay, and Peru, falling below 3% (World Bank, 2009).

Trends and Persistent Issues in LAC Educational Systems

As described earlier, educational systems in LAC region have made progress in certain educational indicators while many issues remain poorly handled and others arise as more complex challenges.

Illiteracy is one of the prevalent issues that educational systems face in the developing world. With 91% adult literacy rates, LAC stands as one of the regions with higher literacy rates than other developing regions, mostly due to proactive policies adopted to reduce the number of illiterates in that region. Only the East Asian and the Pacific region reach 94%, while the Arab States (75%), Sub-Saharan Africa (63%), and South and West Asia (63%) regions lag behind (UNESCO, 2012). Within LAC, the Economic Commission for Latin America and the Caribbean (CEPAL) data on literacy, show that illiteracy is still a prevalent problem in some countries while in others it is reaching eradication thresholds. By 2010, the countries with the highest illiteracy rates in the LAC region were Haiti (48.7%), Guatemala (75.2%), Nicaragua (78%), El Salvador (84.5%), and Honduras (84.8%). Moreover, in countries where literacy rates are low, women are disproportionately under-represented. For instance, in Guatemala, 70.3% of women are literates compared to 80.6% of men (CEPAL, 2012). The Latin American Information System on Educational Trends (SITEAL) reports that in certain countries, differences between urban and rural settings and among regions, are strikingly high when considering illiteracy rates. At the other end of the continuum, 11 countries including Cuba, Chile, Uruguay, Argentina, Costa Rica, and Venezuela show comparatively high literacy rates (SITEAL, 2010). According to UNESCO and the Organization of Ibero-American States (OEI), literacy rates above 95% correspond to eradication of illiteracy. However, it is crucial to highlight that literacy rates are calculated through conventional mechanisms that only reveal extreme cases of illiteracy, and do not include those cases where written and oral language are minimal; thus illiteracy numbers could be higher than those showed in reported databases.

Most regions in the developing world have made sustained progress to incorporating most of their populations in primary education. Net enrollment rates by region show that LAC reached in 2010 95% coverage, similar to East Asia and the Pacific (96%), Central Asia (94%), and South and West Asia (93%) and above Arab States (88%) and Sub-Saharan Africa (77%). However, the regions that already achieved by 1999 percentages over 90%, have not made drastic improvements. Coverage growth in a decade (1999–2010) was only 1% for the LAC region, while there was growth of 18% in South and West Asia and 16% Sub-Saharan Africa during the same decade. This situation may indicate the existence of a glass ceiling in this indicator. Universalization of primary education is one of the earliest achievements of LAC countries. By 2010 net enrollment rates represented a regional average of 95%. Argentina, Ecuador, Uruguay, Cuba, and Panama led with net enrollment rates above 98%. On the other hand, countries such as the Dominican Republic and Paraguay enrolled between 82 and 91% of their potential primary education student population (CEPAL, 2012). In 2011, primary education gross enrollment rates in LAC represented 112.5%, indicating the prevalent efficacy problems (UNESCO Institute for Statistics, 2012).

Since most countries have already reached universality of primary education, issues of quality and efficacy become predominant. Most LAC countries suffer badly from issues related to timely entry, age heterogeneity, and academic achievement. Repetition, the main cause of age heterogeneity, affected mostly primary education. However, progress has been made in most of these areas in the last decades, and the main issues are now found in secondary education (Wolff et al., 2002).

Secondary education is increasingly perceived as the new starting point in educational systems. Overall, LAC region enrollment rates in that level reach 73.5%, being below developing countries in Europe and Central Asia (82.5%) and ahead of East Asia and Pacific (72%), Middle East and North Africa (68%), and Sub-Saharan Africa (26%) (Latest data available 2007) (World Bank, 2010). Within LAC, most countries extend compulsory education to, at least, the lower secondary level. This aim represents a challenge for a level of the educational systems that was conceived to educate the elites of LAC countries and prepare them to continue studies in higher education institutions. In 2010 net enrollments rates for secondary education in LAC countries averaged 73.5%, ranging from the lower end with enrollments between 45 to 63% in countries such as Nicaragua, Guatemala, El Salvador, Paraguay, and the Dominican Republic to the upper-end with net enrollments between 84 to 86% in countries led by Cuba, Bolivia, Chile, and Argentina (CEPAL, 2012). By 2010, gross enrollment rates in LAC countries represented, in average, 90% – a difference revealing efficacy issues.

Efficacy issues in secondary education translate into inequality in secondary level completion rates. This disproportionately affects populations in the lower end of the socio-economic scale: students who do not enter the secondary education level and those leaving secondary education before obtaining the degree. In countries such as Guatemala, Colombia, and Honduras completion rates are highly unequal when considered by quintiles. In contrast, countries such as Chile and to a lesser extent Peru and Argentina, show higher completion rates and lesser inequalities in secondary education. Overall, in LAC the ratio of high-SES (socio-economic status) students graduating from the secondary level is three times higher than the number of low-SES students completing that educational level (CINDA, 2011).

Current trends in secondary education enrollments may signal the existence of a ceiling that impedes universalization of that level. It is expected that limitations on the growth of enrollment rates in LAC countries will affect more seriously those who benefited the most from the universalization of primary education: indigenous populations, African descendants, low-SES adolescents, and people living in rural areas (López, 2013).

LAC region has been involved in a series of educational reforms starting in the 1980s to adapt the systems to new realities. Educational reforms in primary and secondary education have ushered in several changes. One of them relates to changes in the structure of educational systems based on the incorporation of more years of compulsory schooling and taking into account psychological principles on developmental stages. A second type of change relates to curricular reforms that tend to transition from a discipline-based curriculum to an area/integrated-type of curriculum. A third innovation has impacted the management and governance of educational institutions through increasing decentralization and autonomy, outcome assessment, and resources allotted on a competitive basis (Tedesco and López, 2004).

Trends and issues found in the other levels of education are largely echoed when it comes to higher education. In comparison to other regions, LAC positions itself in a similar standing as in primary and secondary education enrollments. Developing countries in Europe and Central Asia lead the scene enrolling 56% of their age grade group, followed by LAC (41%), Middle East and North Africa (30%), and East Asia and Pacific (26%). South Asia (15%) and Sub-Saharan Africa (7%) locate at the bottom of the continuum. As in most educational indicators, the gap between developed and developing nations is striking. High income OECD-member countries, in average, enroll 75% of the age grade group in higher education.

Within LAC, as of 2010, with more than 6.5 million students enrolled in higher education, Brazil has easily the largest higher education system, followed by almost 3 million students in Mexico and 2.5 million in Argentina (PROPHE, forthcoming). In the age grade group enrolled in institutions of higher education (18–24 years), five higher education systems in the LAC region enroll at least 50% of the age grade group, thus reaching the universality threshold set by Trow's typology (1973): Cuba (95%), Venezuela (78%), Argentina (75%), Chile (66%), and Uruguay (63%). Most of the other LAC higher education systems enroll between 15% and 50% of the age grade group, including the two largest higher education systems in the region, Mexico (28%) and Brazil (26%) (World Bank, 2010).

Most of the issues faced by LAC higher education systems relate to their foundations. Systems were originally designed to train the political elite. In general, the establishment of universities in the colonial period followed a model that privileged the training of professionals, focusing the functions of the institutions on teaching activities rather than research. The roots of the emphasis on professions can be traced back to the Napoleonic (French) model. That model based institutional efforts toward the development of professional fields of study at the undergraduate level, such as law and medicine. Thus, research was rarely a preeminent activity in the early days of universities in LAC (Brunner, 1990).

To a greater or lesser extent, massification and universalization brought issues related to effectiveness, quality, autonomy, and efficiency to the forefront (Castro and Levy, 2000). LAC higher education systems are attempting to increase the number of minority and low-SES students accessing and completing higher education. However, they still face structural issues such as low graduation rates. Higher education systems in LAC are highly ineffective, graduating a small proportion of those entering the level, and also inefficient since those graduating may spend as twice as much time, than the time scheduled to finish their studies. To tackle quality issues, reforms have focused on the establishment of quality assurance processes through accreditation agencies. Today, most countries in LAC have implemented some type of accreditation system, for institutions and programs.

From Quantity to Quality: Educational Assessments in LAC

Although educational coverage indicators are still used as a proxy to measure educational systems' progress, increasingly international organizations and governments are utilizing indicators of quality. During the 1980s and 1990s, most LAC countries developed their technical capabilities and established or improved their systems designed for measuring and assessing student academic performance mostly through two different modalities: indigenous creation of tests for highly specific purposes, such as those utilized to regulate admission to higher education or those created with support provided by international financial agencies such as the Inter-American Development Bank (IADB) and the World Bank (Ferrer, 2006). The status of educational assessments in LAC tends to vary significantly by country. Countries such as Chile and Colombia have long experience in administering tests with a high degree of legitimacy and continuity, while other systems in place in countries such as Bolivia, Ecuador, Guatemala, and Peru have been subjected to political changes, making data more difficult to be publicly accessed, with abrupt changes in technical staff (Ferrer, 2006).

LAC countries are increasingly participating in regional and international educational tests, such as the Trends in International Mathematics and Science Study (TIMSS), the Program for International Student Assessment (PISA), and the Regional and Comparative Regional Study (ERCE). Around 20 countries have administered at least one comparative international assessment (Ferrer, 2006). One of the main proxies to assess the comparative quality level of educational systems worldwide is the PISA assessment designed and administered by the Organization for Economic Co-operation and Development (OECD). More than half of the participating countries are highly developed ones. Various Latin American countries such as Chile, Uruguay, Mexico, Colombia, Brazil, Argentina, Panama, and Peru have participated in different years in the administration of the international test, which measures competences developed by 15 year-old students in reading, mathematics, and science. As expected by the large number of OECD-member/developed countries included in the test, LAC countries performed below the OECD average in every competency measure, ranking in the bottom third section out of 65 developed and developing countries. Measurements also show major performance differences within countries, indicating striking gaps between those students in different quintiles of the population (OECD, 2010).

Besides the international assessments described earlier, different regions in the world have developed their own assessment regimes. In LAC, ERCE is a comparative assessment of primary education students in their third and sixth grades that tests their knowledge in mathematics, language (reading and writing), and natural sciences. The different stages of ERCE tests have been designed by the Latin American Laboratory for Assessment of the Quality of Education (LLECE), an initiative that promotes regional tests and is coordinated by the Regional Bureau for Education in Latin America Latina and the Caribbean (OREALC-UNESCO) branch in Chile. Implemented for the first time in 1997, the test is the oldest initiative of this nature in the LAC region. The SERCE test or Second ERCE administered in 2006 shows relevant differences in the performance of students in different LAC countries. Cuba is the single country reaching mean scores significantly higher than the regional average in mathematics (3rd and 6th grades), reading (3rd grade), and natural sciences (6th grade). A second group of countries with identical and higher scores than the regional average are Argentina, Brazil, Chile, Colombia, Costa Rica, Mexico, and Uruguay. Finally, a third group exhibits mean

scores lower than the regional average: Guatemala, Ecuador, El Salvador, Nicaragua, Panama, Paraguay, Peru, and the Dominican Republic (LLECE, 2008).

Although LAC countries have been increasingly willing to participate in regional and international educational assessments, information exchange between assessment entities and system users remains weak. Connections are also missing between test results and policies/decisions and this is related to the limited involvement of the major educational stakeholders (policymakers, teachers, principals, parents, etc.) in devising the tests, designing the reports, and defining strategies for dissemination (Ferrer, 2006).

Educational Research in and About LAC

Investment in R&D in the Region and Educational Research Indicators

Educational research in LAC is part of a broader context of development and investment in Research and Development (R&D) patterns worldwide. When compared to other regions, LAC expenditure in R&D in 2010 locates in the fourth place with 3% of the worldwide expenditures in that sector, behind Asia (34.1%), North America (United States and Canada) (33.1%), and Europe (26.9%) and just ahead of Oceania (2%), and Africa (0.8%) (RICYT, 2012). These patterns reflect structural limitations to research productivity in the LAC region that may impact on educational research, which is part of it.

As described in the previous section, LAC is a region characterized by a great internal heterogeneity that makes difficult to generalize on the status of educational research in the region. Data shows that three LAC countries dominate the R&D scene, investing more than 90% of the R&D resources. By 2010, Brazil led the list with 63.3% of the total investment in R&D in the region, followed by Mexico (19.2%), and Argentina (9.7%). The other LAC countries count just for the remaining 7.8% (RICYT, 2012). These shares can be seen in educational research productivity, where Brazil and Mexico are the key players in the area.

One of the most evident differences between countries within the LAC region relates to research productivity measured through human resources and bibliometric indices. The consolidation of the R&D systems as well as Science and Technology (S&T) statistics systems are appropriate proxies. R&D systems include number of university graduates as percentage of population and the number of publications in international databases. The application of R&D measurements generates three categories when applied to LAC countries (Arber et al., 2008): (1) A first group of countries with consolidated R&D and S&T statistics systems includes Argentina, Brazil, Mexico, and Chile; (2) A second group with consolidated R&D and less developed S&T statistical systems includes Colombia, Costa Rica, Cuba, Panama, Uruguay, and Venezuela; finally, a third group with R&D and S&T systems in early developmental stages embraces Nicaragua, Peru, El Salvador, Paraguay, Guatemala, Bolivia, Ecuador, Dominican Republic, and Honduras.

To a certain extent, S&T statistic systems reflect the overall availability and collection of educational data in LAC countries. Quality and availability of educational data has broad implications and may impact educational research. Some countries in that region are advancing their data generation, collection, and publication mechanisms, and one of the most relevant cases is Brazil. The *Microdados* a web-based, public, and free-access secondary database reported by the Institute of Educational Statistics (INEP) in Brazil makes available disaggregated educational data on primary, secondary, and higher education indicators, educational assessments, and research conducted by INEP from 1995 onward. Raw and non-identifiable data is made available in databases ready to be inputted in quantitative analysis software. This represents a step forward in LAC countries, where the traditional and still predominant method to show data on educational systems consists of reports which show already processed data. Developments on data availability in LAC have usually been promoted by international funding agencies through their loan systems, to develop capacity through training of technical staff and the establishment of entire subunits within educational bureaus at the national and state/provincial levels.

As described, educational research follows patterns similar to those already existing in R&D indicators overall. A handful of countries dominate the scene on production of research in education. A useful although not comprehensive proxy to assess educational research productivity refers to the availability of peer-reviewed academic journals. In recent decades, there have been a series of initiatives to generate digital catalogs of peer-reviewed journals available in LAC such as Latindex and Scielo. The UNAM-based Latindex is conceived as the largest LAC journal index system and compiles almost 5000 scientific journals in that region of which almost 1200 correspond to education journals. Brazil leads the number of educational journals indexed by Latindex with a total of 377 journals, followed by Mexico (167), Chile (144), Venezuela (47), and Colombia (47). Latindex fixes common standards in international journal publications, only including those that meet minimum criteria, so the real number of journals in each country may double or triple those in indexed databases. Still, there is a pattern toward the incorporation of greater number of journals to be included in indexes since they raise their legitimacy and visibility.

All the journals indexed in Latindex as well as Scielo and other regional databases are published either in Spanish or Portuguese; thus limiting visibility outside the region. Journals have usually been published only in hardcopy versions limiting the accessibility and dissemination of LAC educational research. This situation is progressively changing, and journals that were available only in hardcopy are turning to digital versions. This trend shows a turn from local communities of researchers to more cosmopolitan research communities, at least at a regional level.

Research Traditions in Education

Tracing back, beginning in the 1950s and 1960s, educational research in LAC countries was greatly influenced by the human capital theory. The main assumption surrounding educational research was the idea that educational systems would act as the main drivers of change in overall society. For instance, secondary education was portrayed as an agency for economic development and most official documents accepted the existence of a relationship between increasing the number of secondary graduates and economic progress (Watson, 1967).

Not only research but also policy making developed in this stream considered education as a major institution to shape overall development. This shared paradigm on educational research and policy was coupled with studies on the universalization of compulsory education and the struggles to turn a once elite secondary level into a mass/universal level.

Coupled with human capital theory, Durkheim and its structural-functionalist approach in sociology influenced educational researchers and their understanding that quality of education could be assimilated to enrollment rates, and the possibilities of fulfilling the functions of educational systems. In education this era was characterized as 'pedagogical optimism.' In other words, both society and researchers deposited trust in the transformative effects of education and its role in promoting development.

Within the first tradition, educational research was in its first stage, characterized by a more descriptive and normative approach corresponding to an underdevelopment of the social sciences and of their research methodologies. Researchers outside the region were producing sound research about LAC but its impact in the region was minimal at that time (Watson, 1967).

Mostly starting in the 1970s and 1980s, critical theories in education began to criticize the idea that education could be a predominant way to change society, allowing for social and economic mobility. Based on sociological literature such as that produced by Bourdieu and Passeron (1977), and Baudelot and Establet (1971), researchers fell in what is known as the 'pedagogical pessimism.' In other words, researchers focused on what seemed to be reproductive functions of education rather than transformative ones. According to these theoretical developments, educational systems would be responsible for maintaining social inequalities in societies, and even legitimize them rather than promote upward mobility and social change. Moreover, according to Marxist approaches to education, schools were seen as agencies of the dominant (capitalist) ideology that served the interest of the ruling class.

The 1980s and 1990s came accompanied with the restoration of democratic governments in LAC and most countries began to discuss possible ways to overhaul their educational systems. In this period, international funding agencies such as the World Bank and IADB, acted as the main funding agencies for those reforms across Latin America, promoting the dissemination of ideas that had the focus on the decentralization and privatization of educational systems. The development of these trends pushed the research agenda toward the study of educational systems in LAC in terms of policy, governance, structure, curriculum, and teacher professionalization. These developments were understood as part of a neoliberal agenda in a scenario guided by the Washington Consensus.

The 1990s reforms brought new developments to educational systems in LAC from primary to higher education, such as decentralization, new teaching and research conditions (e.g., merit-pay systems), new curriculum developments, overall quality assurance processes, and governance systems. During this period, research began to analyze the so-called NPM-based (New Public Management-based) reforms aimed to generate more efficiency and promote quality in the public sector through practices already consolidated in the private sector. This movement toward policies oriented by deregulation, regulation, and liberalization of education systems takes place in a broader context of discussion on public and private educational services, and its incorporation in the WTO's (World Trade Organization) GATS (General Agreement on Trade in Services) as a tradable commodity. These policies have been usually labeled as neoliberal policies and researchers have adopted a critical stance against the introduction of market-based policies in the educational sector.

Research focusing on policy analysis from different social sciences (political science, sociology, anthropology, and economics) has dominated educational research in LAC countries in the last decades. Research results published in academic journals in between 1993 and 2001 show the dynamic change in educational research theoretical and epistemological paradigms described earlier. In the last decades researchers studying educational policy have predominantly used three main epistemological approaches: neo-Marxism, poststructuralism, and pluralism. A small number of published articles in LAC journals utilize structural-functionalist approaches (Tello and Mainardes, 2012).

First, the neo-Marxist perspective allowed for relative autonomy and individual agency on the relationship between schools and capitalist relations of production that Marxist theory already claimed more recently, studies on educational policy in Latin America have followed Dale's (2000) neo-Marxist approach. Second, educational studies in Europe and the United States (Ball, 1994; Popkewitz and Fendler, 1999) triggered the adoption of poststructuralist perspectives in Latin America, based on the idea that actors occupy a central role in the understanding of policies and that power flows among multiple and different actors. Finally, the pluralist approach focuses on the understanding of diverse interpretations about an educational policy issue (Tello and Mainardes, 2012). All of these perspectives reject the possibility of understanding education as an objective reality, in clear opposition to positivist approaches.

Another characteristic of educational research in LAC countries is its focus on local developments directed toward a local community of researchers. Only in recent decades, and through the expansion of societies on comparative and international studies, researchers are showing more interest in generating comparative research at regional and international level. Comparative education societies in Argentina (2001), Brazil (1983), Cuba (1989), Mexico (2003), and the recently created in Uruguay (2009) promote the production and dissemination of comparative research through the organization of conferences and the publication of specialized peer-reviewed journals.

Latin American studies centers and associations are active in most developed countries. The US-based Latin American Studies Association (LASA) is one of the main Latin American studies association worldwide and the largest in the United States with over 8000 members. Their annual conference papers reflect, to a certain extent, the visibility of educational research in Latin American studies in the United States. The largest comparative educational society, the United States-based Comparative and International Education Society (CIES) also attracts scholars who study educational systems in LAC countries.

The assessment of educational articles in Latinamericanist journals may signal that educational researchers are not the mainstream of Latinamericanist studies and that educational scholars may choose other publication outlets, usually

discipline-based journals. For instance, although not a Latinamericanist journal, the US-based Educational Policy Analysis Archives (EPAA) journal publishes educational policy papers in English, Spanish, and Portuguese reaching academics in the United States and the LAC region, thus generating a hybrid between American scholars studying LAC educational issues and their counterparts from the scholarly community in LAC.

Educational Research by Level

Research on primary and secondary education is mainstream with scholars in LAC, although it is hardly a distinctive feature of the region, since most educational research is oriented to the study of those educational levels worldwide. National associations of researchers in education are increasing in number but are still a rarity in the region. Crucial exceptions are the Brazilian National Association of Educational Research (ANPED) and the Mexican Council on Educational Research (COMIE), dating back to 1976 and 1993, respectively. Countries such as Argentina have not developed yet an overarching research community in education through a generalized affiliating association/society. Instead, they have developed discipline-based associations. For instance, in Argentina societies are established on the bases of specific topics and issues or in the study of education from specific social sciences, e.g., SAHE, the Argentine society for the study of the history of education.

As described in the previous section, educational issues in LAC transitioned from enrollment rates in primary and secondary education to what nowadays is predominantly a concern related to quality and its relationship to equity, stratification, and social mobility. Studies are increasingly centering the analysis of education in LAC countries on how students are performing and the implications that different levels of quality within countries affect educational systems and generate concerns on social equity (Schwartzman and Cox, 2010).

Nonformal education related to education of (illiterate) adults as well as educational issues in rural areas has usually been overlooked by traditional research. The UNESCO center for the study of adult education, CREFAL, has been instrumental in the development of research through specific funding and training of doctoral students and researchers in issues on nonformal education. Founded in 1950, the center produces reports and disseminates research with a focus on the research-action perspective.

Research on higher education is still in a developing stage in LAC, mostly when compared to theoretical and methodological developments in Europe and the United States. Most publications produced in that region have a strong normative, propositional, and ideological content although that situation is changing over time due to new developments and more researchers trained for research in this specific level of education (Krotsch, 2001). Research on higher education began to change its focus and ways of approaching issues starting from the 1980s in concomitance with the growth of public and private universities and their competition for students, professors, and ultimately for prestige. Public policies oriented to report information on higher education systems; international organizations and foundations in charge of financing educational research; and the emergence of graduate level and the issues linked to management and teaching derived from a new sublevel (Krotsch, 2001).

Three different traditions have permeated the study of higher education in LAC. Durkheim's functionalist perspective, mostly utilized in the study of the higher education systems in the region following similar analysis to those carried out by Durkheim on the historical development of higher education in Europe in the medieval times. Bourdieu's analysis of fields in order to uncover their principles of social regulations and reproduction. Finally, and most importantly the sociological studies of educational systems in comparative perspective elaborated by Clark (Krotsch, 2001). The developments from the latter are still the most widely utilized frameworks for the analysis of higher education systems in LAC, based on the idea of a coordination triangle between the state, the market, and the academic oligarchy.

The single most proactive research community in developing research on higher education issues in LAC is Mexican encompassing the UNAM's Center for the Study of Higher Education (CESU) now called Institute for the Study of University and Education (IISUE) one of the most longstanding and prestigious research centers producing research in that level. UNAM also created the Higher Education Researchers Network (RISEU), a network of Latin American higher education scholars that promotes the dissemination of research in that specific level. The North American country also counts with specific higher education research centers in the Montemorelos University (UM), the Center for Research and Advanced Studies (CINVESTAV) of the National Polytechnic Institute (IPN). The Ibero-American University (UI), a private Jesuit university also created its own and well known research center on higher education.

Brazil has consolidated various centers on the study of higher education recently. The longstanding Higher Education Research Center (NUPES) now called Public Policy Research Center (NUPP) at University of Sao Paulo, and many others in the public and the private sector (*Observatorio Universitario*) play a critical role in the production of knowledge specific for this level of education. In Chile, the Latin American School of Social Sciences (FLACSO) and a group of centers specializing in the study of higher education such as the Development University Center (CINDA) and the Corporation of University Promotion (CPU) not only produce research targeted to scholars and also policy makers. In recent years, CINDA has been at the forefront of regional comparisons of higher education systems in the LAC region.

The Regional Center for Higher Education in Latin American and the Caribbean (CRESALC) later renamed as the International Institute for Higher Education in Latin America and the Caribbean (IESALC) is another relevant institution in the promotion of the study of higher education in LAC. An UNESCO institution in Venezuela, CRESALC was first established in 1978 with the aim of contributing to the cooperation and improvement of higher education systems in the region. Now, IESALC serves as a hub to promote the development of national reports on higher education systems, as well as comparative studies, and in certain cases, the compilation of statistics.

International funding agencies also played a key role in the dissemination of knowledge on higher education systems in LAC. In its own review of higher education policies to promote educational reform, the World Bank edited reports on LAC higher education systems: *Higher Education: The Lessons of Experience* published in 1994 and *Higher Education in*

DevelopingCountries: Peril and Promise edited in 2000. The IADB also published the report *Myth, Reality, and Reform: Higher Education Policy in Latin America* by two Latin American experts in higher education. To a greater or lesser extent, these reports called for an overhaul of the higher education systems, based on the conviction that institutions of higher education in that region were highly ineffective and inefficient, and necessarily required organizational and structural change and diversification. These reports brought research concerns into the educational agenda, mostly from the United States, that were absent or overlooked by researchers in the region and generated controversies. Those reports emphasized certain discussions on autonomy, financing, and evaluation, from a perspective of equality of opportunity that did not fit the still predominant LAC approach of a highly state subsidized higher education system with highly autonomous institutions.

Policy recommendations emanating from international funding agencies' reports, specifically those produced by the World Bank, generated widely divergent reactions from two sectors in the higher education research community. One view considered the reports as part of a neoliberal agenda in educational reform overall that tended to highlight alternative approaches to higher education finance, autonomy, and governance. Another sector of researchers were promoting comparative studies of higher education systems and assessing the limitations of the 'Latin American model' of autonomy and governance.

Heterogeneity and unimpressive quality signal the current status of LAC educational systems and educational research. Enormous disparities exist with a few countries consistently outperforming on different educational indicators, the majority with mediocre performances. When compared to other regions in the world, this highly heterogeneous situation leads LAC to remain located below Central Europe and East Asia and above South Asia and Sub-Saharan Africa.

See also: Comparative Research in Education: Iea Studies; Cross-National Comparisons in Education: Findings from Pisa; Data Bases and Statistical Systems: Education, Statistical Systems; Education (Primary and Secondary Schools) and Gender; Education and Economic Growth; Education in South Asia; Education in Sub-Saharan Africa; Education in the Middle East and North Africa; Educational Planning Worldwide; Educational Research, The Field of; Educational Systems: North America; Equity and Education; Gender and Education; Higher Education Research; International Organisations and Education; Literacy Education; Social Inequality and Schooling; Social Studies Education; Socioeconomic Status and Academic Achievement.

Bibliography

Arber, G., Barrere, R., Anilo, G., 2008. Measuring R&D in Developing Countries: Measurement Priorities and Methodological Issues in Latin America (Working paper prepared for the UNESCO Institute for Statistics, Montreal).

Ball, S.J., 1994. Education Reform: A Critical and Post-structural Approach. Open University Press, Buckingham, England; Philadelphia.

Baudelot, C., Establet, R., 1971. L'ecole capitaliste en France. F. Maspero, Paris.

Beech, J., 2002. Latin American education: perceptions of linearities and the construction of discursive space. Comparative Education 38, 415–427.

Bourdieu, P., Passeron, J.C., 1977. Reproduction in Education, Society and Culture. Sage Publications, London; Beverly Hills, CA.

Brunner, J.J., 1990. Educación Superior en América Latina. Cambios y desafíos. Fondo de Cultura Económica, Santiago, Chile.

Castro, C.D.M., Levy, D.C., 2000. Myth, Reality, and Reform: Higher Education Policy in Latin America. Johns Hopkins University Press for the Inter-American Development Bank, Washington, DC.

CEPAL, 2012. Statistical Yearbook for Latin America and the Caribbean. Comisión Económica para América Latina y el Caribe, Santiago, Chile.

CINDA, 2011. Educación Superior en Iberoamérica. Informe 2011. Centro Interuniversitario de Desarrollo, Santiago, Chile.

Dale, R., 2000. Globalization and education: demonstrating a "Common World Educational Culture" or locating a "Globally Structured educational agenda"? Educational Theory 50, 427–448.

Ferrer, G., 2006. Educational Assessment Systems in Latin America: Current Practice and Future Challenges (Programa de Promoción de la Reforma Educativa en América Latina y el Caribe, Washington, DC).

Krotsch, P., 2001. Educación Superior y Reformas Comparadas [Higher Education and Comparative Reforms]. Universidad Nacional de Quilmes, Buenos Aires.

Krotsch, P., 2002. Higher education research: reflections on Argentina and Latin America. In: Schwarz, S., Teichler, U. (Eds.), The Institutional Basis of Higher Education Research. Kluwer Academic Publishers, Dordrecht, pp. 137–155.

La Belle, T.J., 2000. The changing nature of non-formal education in Latin America. Comparative Education 36, 21–36.

Levy, D.C., 1986. Higher Education and the State in Latin America. The University of Chicago Press, Chicago, IL.

LLECE, 2008. Student Achievement in Latin America and the Caribbean. Results of the Second Regional Comparative and Explanatory Study (SERCE). OREALC/UNESCO, Santiago.

López, N., 2013. La educación primaria y secundaria en América Latina: balance y desafíos. Latin American Studies Association Forum 44, 5–6.

OECD, 2010. PISA 2009 Results: Executive Summary. OECD, Paris.

Popkewitz, T.S., Fendler, L., 1999. Critical Theories in Education: Changing Terrains of Knowledge and Politics. Routledge, New York.

PROPHE. Latin America's Private and Public Higher Education Shares [WWW Document]. http://www.albany.edu/dept/eaps/prophe/international_databases.html, forthcoming.

RICYT, 2012. El Estado de la Ciencia: Principales Indicadores de Ciencia y Tecnología Iberoamericanos/Interamericanos. RICYT, Buenos Aires.

Schwartzman, S., Cox, C. (Eds.), 2010. Políticas educacionais e coesão social: uma agenda latino-americana. Elsevier, Rio de Janeiro.

SITEAL, 2010. Informe sobre tendencias sociales y educativas en América Latina 2010. Sistema de Información de Tendencias Educativas en América Latina, Buenos Aires.

Tedesco, J.C., 1987. Paradigms of socioeducational research in Latin America. Comparative Education Review 31, 509–532.

Tedesco, J.C., 1994. Changes in managing education: the case of Latin American countries. International Journal of Educational Research 21, 809–815.

Tedesco, J.C., López, N., 2004. Algunos dilemas de la educación secundaria en América Latina. Revista Electrónica Iberoamericana Sobre Calidad, Eficacia Y Cambio en Educación 2, 1–20.

Tello, C., Mainardes, J., 2012. La posición epistemológica de los investigadores en Política Educativa: debates teóricos en torno a las perspectivas neomarxista, pluralista y pos-estructuralista. Education Policy Analysis Archives 20, 9.

Torche, F., 2010. Economic crisis and inequality of educational opportunity in Latin America. Sociology of Education 83, 85–110.

Trow, M., 1973. Problems in the Transition from Elite to Mass Higher Education. Carnegie Commission on Higher Education, Berkeley, CA.

UNESCO, 2006. Final Communiqué of the Sixth Meeting of the High-level Group on Education for All. UNESCO, Cairo.

UNESCO, 2010. Education for All Global Monitoring Report. UNESCO, Paris.

UNESCO, 2012. Education for All Global Monitoring Report. UNESCO, Paris.

UNESCO Institute for Statistics, 2012. Data Center [WWW Document]. http://www.uis.unesco.org/.

Watson, P.E., 1967. The status of research in secondary education in Latin America. Latin American Research Review 3, 31–43.

Wolff, L., Castro, C.D.M., 2000. Secondary Education in Latin America and the Caribbean. The Challenge of Growth and Reform. Inter-American Development Bank, Washington, DC.

Wolff, L., Schiefelbein, E., Schiefelbein, P., 2002. Primary Education in Latin America: The Unfinished Agenda. Sustainable Development Department Technical Papers Series. Inter-American Development Bank, Washington, DC.

World Bank, 2009. Open Data [WWW Document]. http://data.worldbank.org/.

World Bank, 2010. Open Data [WWW Document]. http://data.worldbank.org/.

Education in Psychiatry

Brian Lunn, School of Medical Education, The Medical School, Newcastle University, Newcastle upon Tyne, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H.J. Walton, volume 6, pp. 4227–4234, © 2001, Elsevier Ltd.

Abstract

Medical education has been leading the way with rapidly evolving pedagogic methodologies. While psychiatric education has been slightly behind the curve this evolution has made significant changes to teaching and assessment of the subject in recent years. Curricula are evolving and being exported. An evidence base is informing practice. There has been a shift to emphasizing measurable variables such as reliability in assessment. Regulators are taking increasing interest in the output of both undergraduate and post graduate training programs. This comes with a risk that the rush to standardization and reliability will result in increasing homogenization and loss of validity.

Introduction

In the previous edition of this article Walton (2001) highlighted the 'creative ferment' that was the state of medical education worldwide. The intervening years have done nothing to decrease this ferment. What has developed is an increasing focus on measurable variables, standardization, and regulatory control.

What has been true for medical education in general is reflected in psychiatry but it would appear that in some respects we have been behind the curve (Gask et al., 2011). In many parts of the world a traditional knowledge base, apprenticeship model still dominates while in others rapidly evolving pedagogic methodologies are leading to a disparity of approach.

Many universities across the world still rely on traditional curricula with divisions into preclinical and clinical phases but there has been an increasing shift to curricula that integrate these elements. Such curricula can be 'blended'; with a mix of didactic lectures and various small group formats, or 'problem based.' Alongside this shift of undergraduate pedagogy has been a shift in teaching in postgraduate environments and specialty training.

With a growing realization that the separation of medical training into undergraduate, general post-qualification, and specialty training was impacting adversely on recruitment into psychiatry, there has been a drive in some countries to introduce psychiatric posts much earlier in training. This development has been alongside bodies such as the Royal College of Psychiatrists putting emphasis on developing curricula that allow for integration of psychiatry early in undergraduate courses (Royal College of Psychiatrists. Report of the Royal College of Psychiatrists' Scoping Group on Undergraduate Education in Psychiatry, 2009).

Curriculum Development

Undergraduate

As highlighted above in some areas of the world traditional curricula as described by Walton (2001) remain the norm, but the trend internationally has been to see a shift from curricula being set by their academic clinicians to being designed by educationalists with an eye to research evidence. The biggest change is to base teaching around how patients present. For psychiatry this means that students are introduced to how patients with mental health problems present at an early stage of their course. The focus has also shifted to what the pluripotential doctor needs to know rather than aiming at protospecialists. That 'physical' health problems such as ischemic heart disease may coexist with mental health problems reflects the reality of clinical practice and directs learning that is focused on understanding a patient in the context of their physical, social, and psychological/psychiatric well-being.

General Post-Qualification Training

Most doctors on qualifying proceed then to 1 or 2 years of general training that may or may not include psychiatry. The international concern about recruitment into psychiatry (covered in more detail below) has led to calls for psychiatry jobs to be made available or integrated into this period of training at as an early stage as possible. That the focus of such early jobs should be training rather than service delivery has gained greater traction recently as has the awareness that a broad base of clinical experience cannot only improve recruitment into low uptake specialisms but improve quality of care for patients.

Specialty Training

The duration of specialty training in Europe averages 5 years (Mihai, 2006) but content and approach vary significantly (**Figure 1**).

Quality assurance of training, when those who deliver it are often widely distributed and not always fully engaged with governing bodies, can be challenging. European Union legislation and the European Board of Psychiatry (Hohagen, 1996) require that national authorities own national quality assurance of training and are responsible for establishing training programs. The European Federation of Psychiatric Trainees (EFPT) have also set out their recommendations for inspection of training. These elements include working conditions and all clinical and academic aspects (European Union of Medical Specialists Report of the European Board of Psychiatry – Quality Assurance in Specialist Training in Psychiatry, 2002). Key to their effectiveness was felt to be the ability to insist on changes to be made or, where appropriate removal of approval from training posts or schemes.

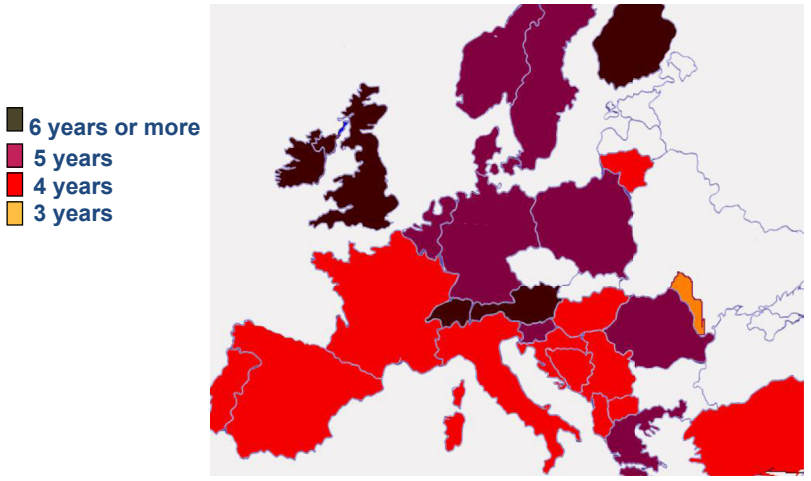

Figure 1 Length of training in psychiatry in Europe.

Training elements considered by quality assurance teams
Accessibility to clinical supervision
Individual supervision
Supervision in psychotherapy
The qualification of trainer
Number of trainees allocated to one supervisor
Structure of theoretical courses
Cases presentation opportunities
Problem solving cases
Workload
Numbers of duty hours
Number of patients examined by each trainee per day
Presence of a trainees' room and a consultation room

Reproduced from European Union of Medical Specialists Report of the European Board of Psychiatry – Quality Assurance in Specialist Training in Psychiatry, 2002.

A common core of knowledge and skills was specified by the European Board of Psychiatry who recommended that this be required of all trainees (European Union of Medical Specialists Report of the European Board of Psychiatry – Quality assurance in specialist training in psychiatry, 2002).

Compulsory training in various settings:
Inpatient units
Outpatient/community settings
Liaison
Emergency psychiatric settings (e.g., emergency departments and police stations)
Compulsory specialisms:
General adult psychiatry
Old age psychiatry
Psychiatric aspects of substance misuse
Psychotherapy (as a minimum psychodynamic and cognitive/behavioral models)
Highly recommended specialisms:
Developmental psychiatry
Forensic psychiatry

Reproduced from European Union of Medical Specialists Report of the European Board of Psychiatry – Quality Assurance in Specialist Training in Psychiatry, 2002.

Training should include structured training (lectures, seminars, and so on) over 4 years for, on average, 4 hours per week.

Subjects core to training
Scientific basis of psychiatry: biological, social, and psychological aspects
Psychopathology. Examination of the psychiatric patient. Diagnosis and classification. Psychological tests and laboratory investigations
Specific disorders and syndromes
Psychotherapies. Psychopharmacology and other biological treatments. Multidimensional clinical management. Community psychiatry
Child and adolescent psychiatry. Mental handicap. Psychiatric aspects of substance misuse. Old age psychiatry. Women and psychiatry. Psychiatry of ethnic minorities. Legal and ethic issues in psychiatry
Research methodology. Epidemiology of mental disorders. Psychiatric aspects of public health and prevention
Medical informatics and telemedicine. Administration, management and economics

Reproduced from van Diest, R., van Dalen, J., Bak, M., et al., 2004. Growth of knowledge in psychiatry and behavioural sciences in a problem-based learning curriculum. Medical Education 38, 1295–1301.

(*Continued*)

Training in psychotherapy was seen as vital and on a par with biological and social elements. This was felt to enhance the ability of psychiatrists to understand and provide balanced treatment for their patients. The key is, as with other treatment modalities, to use methodologies based on scientifically proven systems and the available evidence base.

Whether due to marketing of curricula as western universities move to a business model or in the hope of developing an internationally mobile workforce, there has been a drive to export/import curricula. This is not without its own problems. Typically the direction of movement is from West to East but this is not with any evidence to say that western models work in eastern countries (Bleakley et al., 2011; Rao, 2006a,b). Of some possible concern is that any failure of such transplantation leads to the assumption that this reflects problems with the local society reflected by the "apologetic stance taken by authors in the east about their slowness in adopting western methods, even though … those methods will demand 'intense re-socialisation of learners into metropolitan Western mindsets'" (Bleakley et al., 2011: p. 177).

In addressing this Grant (2014) has argued that "we should be clear about the necessary components of a curriculum designed to be sensitive to the local context." She suggests that this requires consideration of those aspects of the skill and knowledge set required for clinical practice in that locale. Teaching needs to reflect local clinical priorities, the practice setting including how services are provided, how patients present with problems and how cultural factors affect communication and decision making.

In western countries regulators such as the General Medical Council (GMC) in defining their curriculum requirements not only list the learning outcomes they require medical schools to teach but also that quality assurance and stakeholder engagement be factored in (General Medical Council, Tomorrows' Doctors, 2009). They follow up on this with periodic visits.

All of this has led to a shift to defining the purpose and context of a curriculum from content. This could not have been more starkly highlighted than when a post-qualification training reorganization (Modernizing Medical Careers in the UK) required almost immediate post implementation review that reasserted the importance of the apprenticeship model (Took, 2008).

Patient Involvement

The role of patients in medical student teaching has waxed and waned over the millennia (Rosher, 1997). To an outsider in the twenty-first century it must seem incredible that patients would not be front and center in medical education yet there was a significant period when this was not the case and it was only gradually that the patient returned. Even then, while medicine has emphasized teaching at the patient's bedside for many centuries now, the patient "was often no more than a medium, 'an interesting case'" (Spencer and McKimm, 2014).

In the latter half of the twentieth century there was an intense debate as to the nomenclature to be used to describe those with whom doctors came into contact with in the line of duty; patients, clients, service users, or consumers? Often this debate focused on the power dynamic and was driven by vocal activists yet an unexpected benefit was that it usefully highlighted the involvement of patients (for simplicity the most widely and commonly used term is used here) in educating students.

Spencer et al. (2000) in their review of roles that patients hold in medical education suggested a framework for involvement based around four contexts as follows:

- Why (… the patient's background, culture, experience and education enhanced learning)?
- How (… patients may be involved)?
- Where (… to involve patients)?
- What (… to involve patients in)?

One challenge is how to involve patients education beyond the bedside (e.g., assessment policy) while ensuring that this is not mere tokenism. One problem with the term 'patient' is that it implies active illness, when in reality we may all become (or have been) patients. It is therefore possible to have 'patient' involvement in the most technical aspects of medical education by involving people with particular expertise (the 'Why?' suggested by Spencer et al. above) who are not involved in health care. Particular examples of this can be seen in the involvement of nonmedical academics from the humanities with expertise in assessment pedagogy in quality assurance of assessment as is the case with the Royal College of Psychiatrists and Newcastle University.

In psychiatry there can be particular issues around the capacity of patients to consent to involvement in teaching when they are at their most unwell. One solution is the use of simulated patients. Well trained simulated patients can, for medical students and trainees, be difficult to distinguish from 'real' patients (Kurtz et al., 2005; Thistlethwaite and Ridgeway, 2006). The quality of training and support for simulated patients is of particular importance in teaching in psychiatry due to the potential harm when playing roles that are emotionally intense (Eagles et al., 2007) but with training and support the benefits have been argued to outweigh potential disadvantages (Spencer and Dales, 2006).

Selection

For those wishing to become psychiatrist there are many selection hurdles to get over. In the UK this includes as a minimum:

- selection to medical school
- allocation to a foundation training program
- acceptance into core specialty training
- appointment to a higher specialty training post
- appointment to a consultant post

These various selection processes use a heterogeneous group of processes and tools with some part of national programs and some very local.

Traditionally the approach to undergraduate selection has focused on high academic achievement with the emphasis on science-based subjects. In some areas the humanities are totally disregarded when considering academic performance for selection. What is changing is a shift to personal characteristics. One example would be that seen in the UK where there is a multiplicity of approaches across medical schools. Common

to all is that via a centrally administered university application process candidates need to demonstrate that they have achieved an academic threshold (or are expected to do so). From there different universities use personal statements, references, aptitude tests, and interview to varying degrees.

The use of aptitude tests varies from country to country and even when the same test is used in a country how it is used between institutions may vary.

Admission aptitude tests	
Test	*Countries using test*
United Kingdom Clinical Aptitude Test (UKCAT)	United Kingdom
Medical College Admission Test (MCAT)	United States of America and Canada
Graduate Australian Medical School Admission Test (GAMSAT)	Australia and Ireland
Undergraduate Medical and Health Sciences Admission Test (UMAT)	Australia and New Zealand

The use of such tests is only relatively recent so there is, on the whole, little in the way of published evidence about their validity. Data from studies of the MCAT (Tutton and Price, 2002) and UKCAT (Wright and Bradley, 2010; Sartania et al., 2014) provide some support for their use but it has been questioned whether these provide any more useful data than traditional educational assessment (McManus et al., 2005).

Use of interviews to select has become a focus for argument, not least because of the inherent subjectivity. As a result the Multiple Mini Interview (MMI) has begun to be used in selection. Developed at McMaster University in Canada from 2001, it proved to be a reliable predictor of future performance (Eva et al., 2004). In a typical MMI each candidate rotates through the circuit of interviews, with one interviewer per station in that circuit. The interview stations focus a variety of objectives required by the selecting institution. Examples of domains include ethics, professionalism, interpersonal relationships, working with others (e.g., managing people, communication, and collaborating), as well as task performance. A common number of stations is around nine or more, each of about 8 min.

Again selection to initial post-qualification varies significantly throughout the world (Crane and Ferraro, 2000). In most areas of the world traditional interviews will hold sway but there is a push to use more objective methods. In the US the National Resident Matching Program (NRMP) (Miller et al., 2003) gives the United States Medical Licensing Examination (USMLE) primary importance. In the UK a program has been put into place that feeds through from leaving medical school to selection for specialty training in areas such as psychiatry. For selection into Foundation Program jobs all graduates are ranked nationally based on performance at medical school. A wide variety of metrics are used in this ranking process. Before submitting their application all candidates take a Situation Judgment Test (SJT) (Patterson et al., 2012; McDaniel et al., 2001). From this, jobs are allocated based on the candidates' preference for area of the country and type of program.

After 2 years of Foundation training selection to core specialty training in psychiatry is from a national process using selection centers. First used by the military in World War II such selection centers were adopted by industry and then pioneered in the UK for family practitioner selection (Patterson et al., 2001, 2005) before being picked up by other specialisms, including psychiatry. In 2014, for the first time, Higher Education England piloted a Specialty Selection Test (SST) for application to specialty posts, including psychiatry. This pilot test included a Clinical Problem Solving Test (CPS) and an SJT. The results of this test are awaited (Higher Education England. Specialty Selection Test (SST) pilot study for first year applicants to specialty and core training).

Recruitment into Psychiatry

Recruitment into psychiatry is an issue that has been recognized internationally (Pidd, 2003; Feifel et al., 1999; Malhi et al., 2003; Strebel et al., 2000; Fazel and Ebmeier, 2009; Bulbena et al., 2005). In the UK the Royal College of Psychiatrists' annual census has regularly shown that 10-15% of consultant posts are vacant alongside medical student interest in a career in psychiatry resting around 4-5%. Over the years the changes in the demographics of those entering medical courses has had no effect on these figures. Neither have changes in undergraduate curricula. This may be, in part, due to the stability of selection criteria with primacy of sciences over humanities (Brockington and Mumford, 2002) and in some medical schools that physics and chemistry are rated more highly than biology. Silvermann et al. (1983) in a study showed that medical students in Boston were unlikely to consider psychosocial factors in the presentations of patients with chest or abdominal pain and posited that it may actually be that medical curricula reinforce a bias toward such an approach.

A study looking at doctors in the post-qualification training period identified three broad areas that appeared to impact positively on choice of psychiatry as a career (Shah et al., 2011). These were early medical school experience, the influence of more senior doctors, and various aspects related to the working environment.

Shah et al. were not unique in identifying the first of these items (Brown et al., 2007; Baxter et al., 2001; Holm-Petersen et al., 2007). Unfortunately it seems that the interest kindled in medical school declined following qualification (Baxter et al., 2001; Maidment et al., 2003).

What influences choice of psychiatry as a career?	
Early medical school experience	(Shah et al., 2011)
The influence of more senior doctors	
Aspects related to the working environment	
Interest in, and concern for, the mentally ill	(Brown et al., 2007)
Being more interested in people than diseases	
A greater interest in social aspects of medicine	
An ability to tolerate ambiguity	
Undergraduate medical experience	

While some studies encouragingly show that teaching of and exposure to psychiatry improve medical students' views about psychiatry (Baxter et al., 2001; McParland et al., 2003; Holm-Petersen et al., 2007), Shah et al. (2011) showed that as junior doctors progress through their careers the influence of their undergraduate training diminishes. Worryingly just 2 months after qualifying interest in a potential career in psychiatry had fallen to 1.9%. Holding a training post had a positive impact and by the end of training interest had improved to where 4.8% of those studied would consider it a career option.

In understanding why interest in psychiatry declines a number of authors looked at students' views of psychiatry (Feifel et al., 1999; Aghukwa, 2010; Rajagopal et al., 2004) and why students who held positive views about the specialty subsequently chose not to pursue it as a career (Lambert et al., 2006).

One study (Wigney and Parker, 2007) found that a perceived lack of scientific underpinning and the attitudes of doctors from other branches of medicine were hugely influential. The impact of the negative views of other medical professionals have been replicated in several studies (Feifel et al., 1999; Shah et al., 2011; Brown et al., 2007; Lambert et al., 2006) along with the belief that there was a lack of science to back up practice and treatment (Malhi et al., 2003; Silvermann et al., 1983; Shah et al., 2011; Lambert et al., 2006) – somewhat ironically perhaps at a time when commentators outside the medical profession continue to attack the perceived reliance on biological models.

In addressing how to deal with this recruitment challenge while some authors have suggested targeting those who are more likely to choose psychiatry as a career (Pidd, 2003; Malhi et al., 2003) it has been argued that this is unlikely to have any significant effect in the light of the relatively small numbers involved and the effect of training post-qualification (Lunn, 2011). More important is likely to be increasing exposure to psychiatry in these first two training years and demonstrating to colleagues the effectiveness of treatment and the evidence base behind practice.

Assessment

The assessment of knowledge has been the primary mode of assessment in medicine due to its relative ease of development, implementation, and high reliability but the importance of assessing skills while still achieving high reliability has led to significant work in assessment over the last 50 years. As educational theory has developed there has been a shift in the view of educationalists as to the purpose of assessment with greater focus on the needs of the learner rather than the teacher or assessing institution. That being said the need of the regulator, and ultimately the patient, has primacy at the end of the day. This becomes the *de facto* 'finishing line' for candidates and gives focus to the aphorism that assessment drives learning.

Assessment is always done with a purpose. That purpose may be to inform the learner/teacher as to how they are progressing (formative assessment) or to come to a decision about competency at a moment in time (summative assessment). The competence may relate to selection, progression, or qualification.

Depending on the purpose of the assessment test selection will vary. High reliability testing will tend to be structured and standardized, for example, using one of the varieties of multiple choice type questions or 'objective' clinical tests (of which more is discussed later) and these often provide little usable feedback to the candidates. Formative testing, because it does not define a change in state for the candidate (and is unlikely to be subject to challenge), can be more varied and should have a primary focus of providing feedback to the candidates. It is commonly argued that assessments are either summative or formative but in reality all assessments will have a mix of both, with the balance of the mix shifting according to the purpose.

It is important for bodies responsible for regulation of the medical profession, such as national regulatory bodies, to be able to provide credible evidence that those certified by them have achieved defined educational outcomes. This has led to the development of national licensing examinations such as the United States Medical Licensing Examination (USMLE) and in the UK developments overseen by the Medical Schools Council Assessment Alliance (MSCAA), which are covered below. The primary focus of such assessment is summative and feedback is limited.

The USLME has three steps. Step 1 is an 8 hour online multiple choice examination that aims to assess understanding and application of important concepts of those sciences basic to the practice of medicine. Special emphasis is placed on the principles and mechanisms underlying health, disease, and modes of therapy. Step 2 has two components – another MCQ and a clinical examination using 12 simulated patients. It aims to assess application of medical knowledge, skills, and understanding of clinical science essential for the candidate to be able to provide patient care under supervision and includes emphasis on health promotion and disease prevention. Step 3 uses formats similar to Step 2 and aims to assess application of medical knowledge essential for the unsupervised practice of medicine, with emphasis on patient management and leads to a license to practice medicine without supervision.

In the UK the MSCAA was set up in 2010 to share assessment items among the 33 UK medical schools as an alternative to the introduction of a national licensing examination. Along with developing this shared question bank the MSCAA works in partnership with the Foundation Schools to deliver the national SJT and British Pharmacological Society to develop and deliver the Prescribing Safety Assessment (PSA). For the first time the PSA was taken by all UK final year medical students in 2014. In most universities this was used as a formative examination but if a student does not pass she/he either has to pass at re-sit, or pass an alternative, locally administered prescribing assessment to be able to prescribe and progress to full registration.

Assessment for specialty qualifications in psychiatry varies significantly between nations with assessment of knowledge being the only common element across all licensing bodies. At specialty level there is significant variation in who provides recognition of attaining specialty status. This ranges from nationally administered bodies, through the profession itself in

specialty bodies such as the Royal College of Psychiatrists in the UK to individual universities.

Many of the bodies that confirm specialist qualifications have sought to define specific competencies in the hope of standardizing requisite behaviors and skills (Accreditation Council for Graduate Medical Education (ACGME), European Board of Psychiatry (2009), Royal College of Psychiatrists, Royal College of Physicians and Surgeons of Canada) that define competent practice. The Royal College of Physicians and Surgeons of Canada in one of the early attempts to do this introduced the CanMEDS Competencies in 1996 (Frank, 2005; Zaretsky, 2009). This defined a series of roles that were thought to be central to doctors providing optimal health care. These not only drive competency based examinations but inform work place based assessments in a number of countries, not just Canada.

CanMEDS roles
The doctor as a…
Medical expert (central role)
Communicator
Collaborator
Manager
Health advocate
Scholar
Professional

Clinical Skills Assessment

Reliable and valid clinical assessment, assessment of what doctors do, is perhaps the key assessment that medical students and trainees undergo. The traditional 'long-case,' the assessment of a doctor in a setting as close to clinical practice as is possible, has been in decline as it was not without its pedagogic issues. There has therefore been considerable effort in medical education to address this and find suitable alternatives.

Modes of Clinical Assessment

The Long Case

As briefly mentioned in the preceding section, the long case has been the primary mode of clinical skills assessment. In this form of examination the candidate is placed in a room with a patient and required to assess her/him and from the findings develop a diagnosis and management plan. The examiner(s) would only tend to see the candidate after this phase was completed and so assessment is largely a viva voce examination, perhaps including a brief practical element where some clinical signs would be demonstrated in front of the examiner(s).

While the long case appears to have high face validity, as it involves assessment of real patients and requires 'higher order' clinical skills such as the integration of clinical findings and knowledge, a primary problem is that it is unusual for the candidate's interview/examination of the patient to be more than briefly observed. As a consequence key clinical skills are often assessed by inference *post hoc* in the *viva voce* phase. An obvious solution is to observe the whole assessment (Newble, 1991) but this is hugely resource intensive and realistically assessment can only be on one or two cases in total.

The use of real patients is the strength of this form of examination but introduces other difficult-to-control variables. One such variable is the element of 'luck,' which may play to candidates' strengths or weaknesses. This will principally be from the case selection, which will depend on what patients are available to the organizers at the examination center, determined by the nature of the local clinical service. No two 'real' patients with the same disorder will present with the same symptoms and signs. Not limited to psychiatry but perhaps most obvious in this clinical setting, the readiness or otherwise of patients to disclose their history can be independent of a candidate's performance. Of course not only patients are subject to variation – examiners bring a variable skill mix and their own sets of prejudices and preferences. To control for the variability of this form of examination requires multiple cases with multiple examiners. A single long case of an hour has a reliability of only around 0.60 rising to 0.86 after 4 h of testing (Wass and Jolly, 2001); a logistical impossibility for many assessment boards. This can also mean that limited sampling impacts both reliability and validity.

The Modified Long Case

Efforts to address the issues that affect reliability, as described in the last section, led to the development of a modified long case format, the Objective Structured Long Examination Record (OSLER) (Gleeson, 1997). This requires the observation of the candidate's clinical skills and codifies assessment of the presentation of the case, their conclusions, and recommendations by use of a structured oral examination. Typically this still leaves the issues of logistics. A variation of the OSLER the Multiple Objective Structured Long Examination Record (MOSLER) was developed to try and address this later problem. Cases are shorter and focused, which allows for multiple cases and examiners. Nesting of multiple variables makes assessment of reliability difficult but there is some evidence that this is within the range desired by regulators (Wright et al., 2013) with high validity and acceptability to patients, candidates, and examiners alike.

The Objective Structured Clinical Examination

The Objective Structured Clinical Examination (OSCE) was developed out of a desire for a clinical examination that allowed objective assessment of clinical skills (Harden and Gleeson, 1979). A circuit of 'stations' that allow assessment of a variety of clinical scenarios where candidates' skills in dealing with the chosen assessment objectives are tested. With multiple scenarios and examiners this form of assessment can achieve much higher reliability in the same time it would take to carry out two long cases. Additional benefits are that one can define a blueprint of assessment objectives that matches learning outcomes (Lunn, 2005) and control of patient variables with the use of simulated patients (see Section Patient Involvement).

The OSCE wasn't used in psychiatry until relatively recently (Famuyiwa et al., 1991) but has now an evidence base for its use in undergraduate and postgraduate assessment. In the UK the Royal College of Psychiatrists developed an OSCE that was used for the first clinical assessment (Tyrer and Oyebode, 2004). Although this was discontinued following an overhaul of their assessment processes it lives on in all but name in the current final clinical examination, the Clinical Assessment of Skills and Competencies (CASC).

The ability, consequent on the use of simulated patients in the OSCE, to allow structured, replicated, assessment of skills in such a way as would not be possible if real patients were used, is a real advantage of this assessment tool. That said, concerns have been voiced regarding the verisimilitude of the clinical encounter and empathic engagement of candidates. Studies have suggested that this is not a particular issue (Sanson-Fisher and Poole, 1980) and that OSCEs and long cases allow assessment of clinical competence 'no worse and no better' than each other (Wass et al., 2001) and that OSCE skill assessment correlates reasonably with workplace performance (Melding et al., 2002).

The qualitative data from OSCEs has allowed assessment between various demographic groups. As is the norm in most examinations women tend to do better than men. It is interesting to note, however, that in the CASC this is not statistically significant when UK primary medical qualification (PMQ) candidate gender groups are compared. It remains highly significant for the non-UK PMQ groups. A number of factors are probably contributing to this. One is that the UK PMQ candidates' training in medical school closely fits to what is being assessed in the post-graduate arena. The clinical context, which carries its own cultural character, will advantage those with prior knowledge and who are comfortable with it as will language, particularly colloquial language. While congruous with real clinical practice these variables are independent of clinical knowledge and, perhaps, clinical skill. Accepting that the MRCPsych is an assessment of the ability to practice as a specialist in the UK puts this in context, as does that similar patterns of performance have been found in Canadian and Australasian specialty examinations. This emphasizes that the culture of society and psychiatric practice and the idiom of language all play a role in the ability to carry out a specialist clinical role.

The original OSCEs were structured around highly operationalized checklists. This checklist based approach carries problems with it for psychiatry. It has been highlighted that more experienced candidates, with higher order skills, can do less well than less experienced candidates when such checklists are used (Hodges et al., 1999) and that more global 'expert' judgment has higher value in psychiatry.

Workplace-Based Assessments

Miller classified clinical skills according to a pyramidal hierarchy (Miller, 1990). In this structure, 'knowledge' forms the base on which the clinical competence is built. Traditional assessment methodology has placed emphasis on the lower tiers of his pyramid and only with objective driven assessments do candidates begin to be assessed on the ability to show how they do something (performance). Yet there is a need to assess how doctors actually do something (action in Miller's hierarchy). This has led to the use of workplace based assessments (WPBA) (**Figure 2**).

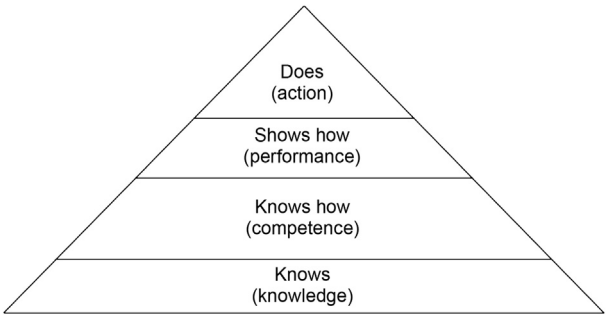

Figure 2 Framework for clinical assessment. From Miller, G.E., 1990. The assessment of clinical skills/competence/performance. Academic Medicine. 6 (Suppl.), S63–S67.

Initial enthusiasm began to be tempered by the realization that when doctors formed working relationships with those they subsequently had to assess there was a 'failure-to-fail.' In the UK this led, in the Foundation years, for the Academy of Medical Royal Colleges to shift the emphasis from these assessments being summative to formative. Candidates are then assessed against their engagement remediating any identified learning needs. While with independent assessment high reliability can be achieved with such assessments (Norcini et al., 2003) the logistics of such methodologies tend not to support such independent assessment.

The driver for regulators in recent years has been to ensure assessment is reliable. This does carry with it a shift in emphasis to methods that can diverge from real clinical practice. Authors such as Hodges have begun to argue that there needs to be a reappraisal of the value of subjectivity. Hodges argues that multiple assessments addressing the single question of whether the individual marking the candidate would be happy for them to treat someone they care about have greater validity and can be reliable when sufficient numbers of such assessments are performed. Key to this is the provision of comprehensive feedback explaining why the assessor made the decision they did. In this form individual assessments have high value as formative assessment and low summative value. Cumulatively the formative value remains and the summative value increases to the point where it has real value and is reliable (Hodges, 2013). Most importantly this methodology has a greater chance of influencing student learning.

Life-Long Learning

Of course learning does not end when formal education ends. If a psychiatrist does not continue to learn then very quickly she/he will fall behind the curve of best medical practice and as a consequence will be a poorer doctor for it. This has long been recognized by the profession and lifelong learning has been supported by professional organizations such as the Royal College of Psychiatrists and the American Psychiatric Association (Bourgeois et al., 2009) through conferences, workshops, and courses and lately by provision of online learning materials.

Regulators have caught on to the importance of this and through licensing and revalidation requirements now assess doctors' engagement with lifelong learning. Tools include peer review, involvement of patients, students, trainees, and nonmedical colleagues through 360° feedback and mandatory participation in learning. Eventually as with everything regulators do, this will become increasingly focused on reliability at the expense of validity and structured examinations will be imposed.

Conclusion

Therefore, Walton's 'creative ferment' has not died down and it is not likely to in the near future. For the moment there are two competing drives in the pedagogy of psychiatry. On one hand regulators seek greater reassurance with a focus on reliable, standardized assessments and curricula that can support high stakes decisions and can be defended in court. On the other, educationalists are arguing for a return to valuing validity of and making the learner the center of the pedagogic process. These need not be mutually exclusive but those who hold the purse strings are unlikely to dispense the largesse that would support, for example, high validity, reliable assessments. Such investment is justified, however, and as Robinson has argued (2006, 2010), our current system reduces education to an industrialized 'fastfood' output model where we need and want individualized, customized product.

Considering that the problem of recruitment in psychiatry seem universal and perennial it would be easy to become disheartened, but there has been a fundamental change in that all the issues raised are being discussed openly and there appears to be a desire from regulators, governments, and from within the profession to find creative solutions to these issues.

See also: Continuing Professional Development; Education: Anthropological Aspects; Educational Research, The Field of; Globalization in Education; Medical Education; Neuroscience of Education.

Bibliography

Aghukwa, N.C., 2010. Attitudes towards psychiatry of undergraduate medical students at Bayero University, Nigeria. South African Journal of Psychiatry 16, 147–152.

Association of American Medical Colleges. Medical College Admission Test. Available at: https://www.aamc.org/students/applying/mcat/ (accessed 06.14.).

Australian Council for Educational Research. Graduate Australian Medical School Admission Test. Available at: http://www.gamsat-ie.org (accessed 06.14.).

Baxter, H., Singh, S., Standen, P., Duggan, C., 2001. The attitudes of "Tomorrow's doctors" towards mental illness and psychiatry: changes during the final undergraduate Year. Medical Education 35, 381–383.

Bleakley, A., Bligh, J., Browne, J., 2011. Medical Education for the Future. Identity Power and Location. Springer, London.

Bourgeois, J.A., Hales, R.E., Young, J.S., Yuodsky, S.C. (Eds.), 2009. The American Psychiatric Publishing Board Review Guide for Psychiatry. American Psychiatric Publishing, Arlington.

Brockington, I.F., Mumford, D.B., 2002. Recruitment into psychiatry. British Journal of Psychiatry 180, 307–312.

Brown, T., Addie, K., Eagles, J., 2007. Recruitment into psychiatry: views of consultants in Scotland. The Psychiatrist 31, 411–413.

Bulbena, A., Pailhez, G., Coll, J., Balon, R., 2005. Changes in the attitudes towards psychiatry among Spanish medical students during training in psychiatry. European Journal of Psychiatry 19, 79–87.

Crane, J.T., Ferraro, C.M., 2000. Selection criteria for emergency medicine residency candidates. Academic Emergency Medicine 7, 54–60.

van Diest, R., van Dalen, J., Bak, M., et al., 2004. Growth of knowledge in psychiatry and behavioural sciences in a problem-based learning curriculum. Medical Education 38, 1295–1301.

Eagles, J.M., Calder, S.A., Wilson, S., et al., 2007. Simulated patients in undergraduate education in psychiatry. The Psychiatrist 31, 187–190.

European Board of Psychiatry, 2009. European Framework for Competencies in Psychiatry Available at: http://uemspsychiatry.org/wp-content/uploads/2013/09/2009-Oct-EFCP.pdf (accessed 06.14.).

European Union of Medical Specialists, 2002. Report of the European Board of Psychiatry – Quality Assurance in Specialist Training in Psychiatry.

Eva, K.W., Reiter, H.I., Rosenfeld, J., Norman, G.R., 2004. An admissions OSCE: the multiple mini-interview. Medical Education 38, 314–326.

Famuyiwa, O.O., Zachariah, M.P., Ilechukwu, S.T., 1991. The objective structured clinical examination in undergraduate psychiatry. Medical Education 25, 45–50.

Fazel, S., Ebmeier, K.P., 2009. Specialty choice in UK junior doctors: is psychiatry the least popular specialty for UK and international medical graduates? BMC Medical Education 9 (1), 77.

Feifel, D., Moutier, C.Y., Swerdlow, N.R., 1999. Attitudes toward psychiatry as a prospective career among students entering medical school. American Journal of Psychiatry 156, 1397–1402.

Frank, J.R., 2005. The CanMEDS 2005 Physician Competency Framework. Better Standards. Better Physicians. Better Care. The Royal College of Physicians and Surgeons of Canada, Ottawa. Available at: http://www.royalcollege.ca/portal/page/portal/rc/common/documents/canmeds/resources/publications/framework_full_e.pdf (accessed 06.14.).

Gask, L., Baron, D., Coskun, B., 2011. Overview: the need for improvements in psychiatric education. In: Gask, L., Baron, D., Coskun, B. (Eds.), Teaching Psychiatry: Putting Theory into Practice. Wiley, Hoboken.

General Medical Council, 2009. Tomorrow's Doctors Available at: http://www.gmc-uk.org/education/undergraduate/tomorrows_doctors_2009.asp (accessed 06.14.).

Gleeson, F., 1997. AMEE medical education guide no 9: assessment of clinical competence using the objective structured long examination record (OSLER). Medical Teacher 19, 7–14.

Grant, J., 2014. Principles of curriculum design. In: Swanick, T. (Ed.), Understanding Medical Education. Evidence, Theory and Practice, second ed. Wiley Blackwell, Chichester.

Harden, R.M., Gleeson, F.A., 1979. Assessment of clinical competencies using an objective structured clinical examination (OSCE). Medical Education 13, 39–45.

Higher Education England, 2014. Specialty Selection Test (SST) Pilot Study for First Year Applicants to Specialty and Core Training. Available at: http://specialtytraining.hee.nhs.uk/specialty-recruitment/specialty-selection-test-sst-pilot-study-for-first-year-applicants-to-specialty-and-core-training/ (accessed 06.14.).

Hodges, B., Regeher, G., Mcnaughton, N., et al., 1999. OSCE checklists do not capture increasing levels of expertise. Academic Medicine 74, 1129–1134.

Hodges, B., 2013. Assessment in the post-psychometric era: learning to love the subjective and collective. Medical teacher 35 (7), 564–568.

Hohagen, F., 1996. Training in psychiatry in Europe - recommendations of the European board of psychiatry. European Psychiatry 11, s4–248s.

Holm-Petersen, C., Vinge, S., Hansen, J., Gyrd-Hansen, D., 2007. The impact of contact with psychiatry on senior medical students' attitudes toward psychiatry. Acta Psychiatrica Scandinavica 116, 308–311.

Kurtz, S., Silverman, J., Draper, J., 2005. Teaching and Learning Communication Skills in Medicine, second ed. Radcliffe Medical Press, Oxford.

Lambert, T.W., Turner, G., Fazel, S., Goldacre, M.J., 2006. Reasons why some UK medical graduates who initially choose psychiatry do not pursue it as a long-term career. Psychological Medicine 36, 679–684.

Lunn, B., 2005. The OSCE blueprint and station development. In: Rao, R. (Ed.), OSCEs in Psychiatry. Gaskell, London.

Lunn, B., 2011. Recruitment into psychiatry: an international challenge. Australian and New Zealand Journal of Psychiatry 45, 805–807.

Maidment, R., Livingston, G., Katona, M., Katona, C., 2003. Carry on shrinking: career intentions and attitudes to psychiatry of prospective medical students. The Psychiatrist 27, 30–32.

Malhi, G.S., Parker, G.B., Parker, K., et al., 2003. Attitudes toward psychiatry among students entering medical school. Acta Psychiatrica Scandinavica 107, 424–429.

McDaniel, M.A., Morgenson, F.P., Finnegan, E.B., et al., 2001. Use of situational judgment tests to predict job performance: a clarification of the literature. Journal of Applied Psychology 86 (4), 730–740.

McManus, I.C., Powis, D.A., Wakeford, R., et al., 2005. Intellectual aptitude tests and A levels for selecting UK school leaver entrants for medical school. British Medical Journal 331, 555–559.

McParland, M., Noble, L.M., Livingston, G., McManus, C., 2003. The effect of a psychiatric attachment on students' attitudes to and intention to pursue psychiatry as a career. Medical Education 37, 447–454.

Medical Schools Council Assessment Alliance. Available at: http://www.medschools.ac.uk/MSCAA/Pages/default.aspx (accessed 06.14.).

Melding, P., Coverdale, J., Robinson, E., 2002. A 'fair play?' Comparison of an objective structured clinical examination of final year medical students training in psychiatry and their supervisors' appraisals. Australian Psychiatry 10, 344–347.

Mihai, A., 2006. European federation of psychiatric trainees EFPT and harmonisation in training programs. In: New Tools for Clinical Practice the 28th Nordic Congress of Psychiatry Abstracts, pp. 28–29.

Miller, G.E., 1990. The assessment of clinical skills/competence/performance. Academic Medicine 6 (Suppl.), S63–S67.

Miller, J.B., Schaad, D.C., Critenden, R.A., et al., 2003. Communication between programs and applicants during residency selection: effects of the match on medical students' professional development. Academic Medicine 78, 403–411.

Newble, D.I., 1991. The observed long case in clinical assessment. Medical Education 25, 369–373.

Norcini, J.J., Blank, L.L., Duffy, D., et al., 2003. The mini-CEX: method for assessing clinical skills. Annals of Internal Medicine 138, 476–481.

Patterson, F., Ferguson, E., Lane, P., Norfolk, T., July 2001. A new competency based selection system for general practitionersThe 10th International Conference of the International Society for the Study of Individual Differences Edinburgh.

Patterson, F., Ferguson, E., Norfolk, T., Lane, P., 2005. A new selection system to recruit general pravtitioners: preliminary findings from a validation study. British Medical Journal 330, 711–714.

Patterson, F., Ashworth, V., Zibarras, L., et al., 2012. Evaluations of situational judgment tests to assess non-academic attributes in selection. Medical Education 46, 850–868.

Pidd, S., 2003. Recruiting and retaining psychiatrists. Advances in Psychiatric Treatment 9, 405–411.

Rajagopal, S., Rehill, K.S., Godfrey, E., 2004. Psychiatry as a career choice compared with other specialties: a survey of medical students. Psychiatric Bulletin 28, 444–446.

Rao, R.H., 2006a. Reflections on the state of medical education in Japan. The Keio Journal of Medicine 55 (2), 41–51.

Rao, R.H., 2006b. Perspectives in medical education. Implementing a more integrated, interactive and interesting curriculum to improve Japanese medical education. The Keio Journal of Medicine 56 (3), 75–84.

Robinson, K., 2006. How Schools Kill Creativity Available at: http://www.ted.com/talks/ken_robinson_says_schools_kill_creativity (accessed 06.14.).

Robinson, K., 2010. Bring on the Learning Revolution! Available at: http://www.ted.com/talks/sir_ken_robinson_bring_on_the_revolution (accessed 06.14.).

Rosher, L., 1997. The growth of medical education and the medical profession. In: Louden, I. (Ed.), Western Medicine. Oxford University Press, Oxford, pp. 147–157.

Royal College of Psychiatrists, 2009. Report of the Royal College of Psychiatrists' Scoping Group on Undergraduate Education in Psychiatry Available at: https://www.rcpsych.ac.uk/pdf/Final%20Education%20in%20Psychiatry%20Scoping%20group%20report%20May%202009.pdf (accessed 06.14.).

Sanson-Fisher, R.W., Poole, A.D., 1980. Simulated patients and the assessment of medical students' interpersonal skills. Medical Education 14, 249–253.

Sartania, N., McClure, J.D., Sweeting, H., Browitt, A., 2014. Predictive power of UKCAT and other pre-admission measures for performance in a medical school in glasgow: a cohort study. BMC Medical Education 14 (116) Available from: http://www.biomedcentral.com/1472-6920/14/116 (accessed 06.14.).

Shah, P., Brown, T., Eagles, J., 2011. Choosing psychiatry: factors influencing career choice among foundation doctors in Scotland. In: Brown, Eagles (Eds.), Teaching Psychiatry to Undergraduates. RCPsych Publications, London, pp. 255–263.

Silvermann, D., Gartnell, L.N., Aronson, M., et al., 1983. In search of the bio-psychosocial perspective: an experiment with beginning medical students. American Journal of Psychiatry 140, 1154–1159.

Strebel, B., Obladen, M., Lehmann, E., Gaebel, W., 2000. Attitude of medical students to psychiatry: a study with the German translated, expanded version of the ATP-30. Nervenarzt 71, 205–212.

Spencer, J., Blackmore, D., Heard, S., et al., 2000. Patient-oriented learning: a review of the role of the patient in the education of medical students. Medical Education 34, 851–857.

Spencer, J., Dales, J., 2006. Meeting the needs of simulated patients and caring for the person behind them? Medical Education 40, 3–5.

Spencer, J., McKimm, J., 2014. Patient involvement in medical education. In: Swanick, T. (Ed.), Understanding Medical Education. Evidence, Theory and Practice, second ed. Wiley Blackwell, Chichester.

Thistlethwaite, J., Ridgeway, G., 2006. Making it Real. A Practical Guide to Experiential Learning. Radcliffe Publishing, Oxford.

Took, J., 2008. Aspiring to Excellence Findings and Final Recommendations of the Independent Inquiry into Modernising Medical Careers Available at: http://www.medschools.ac.uk/AboutUs/Projects/Documents/Final%20MMC%20Inquiry%20Jan2008.pdf (accessed 06.14.).

Tutton, P., Price, M., 2002. Selection of medical students. British Medical Journal 324, 1170–1171.

Tyrer, S., Oyebode, F., 2004. Why does the MRCPsych examination need to change? The British Journal of Psychiatry 184, 197–199.

United Kingdom Clinical Aptitude Test. Available at: http://www.ukcat.ac.uk (accessed 06.14.).

United States Medical Licensing Examination. Available at: http://www.usmle.org (accessed 06.14.).

Universities Admissions Centre. Undergraduate Medical and Health Sciences Admission Test. Available at: http://www.uac.edu.au/undergraduate/admission/umat.shtml (accessed 06.14.).

Walton, H.J., 2001. Education in psychiatry. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Elsevier, Oxford.

Wass, V., Jolly, B., 2001. Does observation add to the validity of the long case? Medical Education 35, 729–734.

Wass, V., Jones, R., van der Vleuten, C., 2001. Standardized or real patients to test clinical competence? The long case revisited. Medical Education 35, 321–325.

Wigney, T., Parker, G., 2007. Medical student observations on a career in psychiatry. Australian and New Zealand Journal of Psychiatry 41, 726–731.

Wright, S.R., Bradley, P.M., 2010. Has the UK Clinical Aptitude Test improved medical student selection? Medical Education 44 (11), 1069–1076.

Wright, S., Lunn, B., Barton, R., Myers, A., 2013. Challenges of assessing the reliability of the MOSLER clinical examination15th Ottawa Conference 9-13 March 2012 Kuala Lumpur, Malaysia Available at: http://media.wix.com/ugd/bd1cfc_c54c42a6e92b626fbcb65c15e4ad90f4.pdf (accessed 06.14.).

Zaretsky, A., 2009. Evaluation of resident competency in psychiatry: a Canadian perspective. In: Leverette, J.S., Hnatko, G.S., Persad, E. (Eds.), Approaches to Postgraduate Education in Canada: What Educators and Residents Need to Know. Canadian Psychiatric Association, Ottawa.

Education in South Asia

R Malatesha Joshi, Texas A & M University, College Station, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

According to the South Asian Association for Regional Cooperation, eight membership countries are Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka form South Asia. After a brief background of these countries, the literacy rate is provided based on the Central Intelligence Agency World Fact Book (2009). The literacy rate is low in most of the countries compared to the international average due to poverty and lack of educational facilities. Despite the nations' emphasis on free and compulsory education as well as child labor laws, many children still do not go to school and work in the field because of economic situations. Males generally have a higher rate of literacy rate than females. The educational system in various countries is reviewed and surprisingly similar in terms of the classification of levels into primary, secondary, and high school levels. Most of the nations support teaching two languages or more from early grades.

Introduction

The concept of a group of countries that could be classified as South or Southern Asia may be attributed to Ziaur Rahman, former president of Bangladesh. President Rahman, in 1977, recommended that a group, known as South Asian Association for Regional Cooperation (SAARC) be formed to demonstrate self-reliance in technology, economic, as well as social and cultural development. However, SAARC did not come to fruition till December 1985 and at the present time, is headquartered in Kathmandu, Nepal. The original seven founding members were Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka. The number of countries was increased to eight in 2005 when Afghanistan joined the group. Similar to Group of 8, heads of state usually meet once a year. United Nations Statistical Division (2011), on the other hand, designated the following countries as 'South Asia': Afghanistan, Bangladesh, Bhutan, India, Iran (Islamic Republic of), Maldives, Nepal, Pakistan, and Sri Lanka. It should be noted that the United Nations specifically points out that the countries designated as 'South Asia' is purely for statistical convenience rather than to any political, religious, or linguistic affiliations. South Asia, sometimes is also referred to as 'Indian subcontinent' (McLeod, 2002), however, according to Bose and Jalal (2004), South Asia is preferred rather than Indian subcontinent. It appears that there is no clear classification for the area defined as 'South Asia.' For our purpose, we shall use the term South Asia throughout this article, and would include the eight countries (Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka) included in SAARC.

Even though 'education' can include both formal and informal and can include from preschool or kindergarten to university and postgraduate education, this article addresses only the formal education up to high school level and not discuss university and college education. Further, this article does not discuss about religious schools. Further, I am mostly concentrating on the aspect of literacy and not go into detail about aspects like science or math education. As it becomes clear from the explanation below, there is a lack of disagreement about the definition of literacy as well as the rate of literacy from different reports.

Literacy, according to United Nations Educational, Scientific and Cultural Organization (UNESCO, 2004), "is the ability to identify, understand, interpret, create, communicate and compute, using printed and written materials associated with varying contexts. Literacy involves a continuum of learning in enabling individuals to achieve their goals, to develop their knowledge and potential, and to participate fully in their community and wider society" (p. 13). As can be seen, this is a broad definition that includes computation, communication, and participation in community and wider society. The National Assessment of Adult Literacy (White and Dillow, 2005) defines literacy more from the perspective of reading and writing skills and provides two types of definitions: 'task-based (conceptual)' and 'skills-based (operational).' Task-based definition is mainly related to performing literacy tasks and is defined as the "ability to use printed and written information to function in society, to achieve one's goals, and to develop one's knowledge and potential" (White and Dillow, 2005). As the name indicates, skills-based definition refers to the successful use of printed material at the word-level reading skills and higher level literacy skills. For our purpose in this article, we shall define 'literacy' broadly to refer to reading and writing skills.

The definition of literacy as the ability to read and write is also similar to the one suggested by the World Fact Book (2009), published by Central Intelligence Agency (CIA) of the United States, except that the literacy rate is based on anyone above 15 years of age who can read and write. Even though the CIA definition may be acceptable, the method used in assessing the literacy rate may be questionable. CIA might have assessed the literacy rate by asking individuals whether they can read and write; or might have asked to complete a survey. Naturally, people might have answered affirmatively for the question whether they can read and write and perhaps, only those who can read and write might have completed the survey. For instance, according to the World Fact Book, the literacy rate of the US is listed as 99% for the total population and the percentage is the same for both males and females as well. However, according to the report by the National Center for Education Statistics (NCES) (1999), one-third of fourth graders cannot comprehend simple texts and in the case of students from the minority groups almost two-third of the students have difficulty in reading and writing. Additionally,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92143-3

one-fourth of the adults have difficulty in reading a newspaper. It is under this light that the statistics has to be interpreted that I have reported in this article. Due to lack of empirical research in many of the nations in South Asia, I have relied on the information provided by the World Fact Book.

Afghanistan

The official name of Afghanistan is the Islamic Republic of Afghanistan and is a landlocked county bordering Pakistan and Iran, with a population of about 29 million people. The two official languages of Afghanistan are Persian, also known as Dari, and Pashto. Both languages belong to the Indo-European family and Iranian subfamily. The writing system of both Dari and Pashto may be traced back to the Arabic system even though there may be differences from region to region.

According to the World Fact Book (2009), the literacy rate, based on the 2001 census, for the total population is 28.1%, with 43.1% of males and 12.6% of the females being literate. Even though women made up 40% of the doctors and 60% of the teachers at Kabul University in 1988, due to the wars, the number, especially education of females, has been seriously neglected. According to the report from International Institute for Educational Planning (UNESCO, 2011), in 10 years from 2001 to 2010, primary school enrollment increased from 1 million to 7 millions with an increase of 37% among females. Even though some improvements have been made in school buildings, only half of the buildings can be considered as usable. Beginning 2010, the United States started several schools known as Lincoln learning centers in Afghanistan. The purpose of these centers is to offer English language classes, provide library facilities, Internet access, educational and other counseling services. There are approximately 20 such centers and aim to reach at least 4000 Afghan citizens per month per location (U.S. Dept. of State, 2011).

In addition to the problem of wars, poverty, lack of school building, an Afghan student might face another problem in learning to read and write – the diglossic nature of the language. Diglossia, according to Ferguson (1959), is when one form of the language is used for ordinary and informal conversations and the other form of the language is learned largely by formal education and is used for mostly written and formal spoken purposes. Perhaps, the initiatives from UNESCO and the US could improve the education and literacy development for both boys and girls in Afghanistan.

Bangladesh

The official name of Bangladesh is the People's Republic of Bangladesh and is located in South Asia bordered by India and Burma with a population of about 160 million. Original split from British India in 1947, it was called East Pakistan, but became an independent country in 1971. The capital city is Dhaka and the official language is Bengali and English is quite common. Bengali belongs to the Indo-Aryan language family and is the sixth most spoken language in the world. The written language is fairly transparent.

According to the World Fact Book (2009), the literacy rate, based on the 2001 census, for the total population is 47.9%, with 54% of males and 41.4% of the females being literate. Educational levels in Bangladesh can be classified into three segments – primary level ranging from grades 1 to 5; junior secondary level from grades 6 to 8; secondary level from grades 9 to 10; and higher secondary level from grades 11 to 12. There are options to attend government-supported schools as well as private schools, where English instruction may be more popular. The government funds most of the schools and may provide some funding for private schools also. Bangladesh conforms to education for all principles and the primary level, (grades 1–5; 6–10 years of age) is free and compulsory for all students in government schools. Further, the government encourages education for females by providing free education up to grade 10 as well as stipends to continue education. Even though education in Bangladesh is supported by the government and has a good model, because of poverty, there are various problems. The problems include high teacher–pupil ratio in primary schools, perhaps on the average of 45 students per teacher, high teacher absence ratio, high dropout rates, as well as grade repetition (UNESCO, 2007).

Bhutan

Kingdom of Bhutan is the official name and is bordered by the Himalayan mountains and India, with a population of approximately 700 000 (World Fact Book, 2009). Despite poverty, Bhutan is considered the happiest country in Asia and the eighth happiest in the world (Business Week, 2006). The capital of Bhutan is Thimphu and the official language is Dzongkha, even though English is widely used. Dzongkha belongs to Sino-Tibetan language family and the orthography is fairly transparent. The World Fact Book (2009) reports the literacy percentage rate of Bhutan, based on 2003 estimation, as 47, 60, and 34, for the total population, males, and females, respectively. The Government of Bhutan provides free and compulsory education at the primary level, which consists of grades 1–5. Grades 6–8 are called the junior high school and 9–11 make up the high school. Due to the distance, most of the junior high and high schools are boarding schools. English is the medium of instruction in majority of the schools at the junior high and high school levels. Recently, some private schools have been started in big cities like Thimphu, which are mostly boarding schools and the medium of instruction is English. The primary aim of many of the private schools is to prepare students to attend universities. Due to the proximity, many students also go to study in India.

India

The official name of India is The Republic of India and is bordered by Pakistan, China, Burma, Bangladesh, Nepal, and Bhutan. It is the second most populous country with over 1.2 billion people and the most populous democracy in the world. The capital city is New Delhi and the official languages of India are Hindi and English. Hindi is an Indo-European language, written in Devanagiri script and fairly transparent. Additionally, there are 22 official languages in India mainly based on the different states yet, the speakers of these languages

may be in the millions, like, Bengali, spoken in West Bengal, is the sixth most spoken language in the world. Generally, most of the major languages from Northern India can be considered as belonging to the Indo-European family while most of the major South Indian languages can be considered as belonging to the Dravidian family, yet both of them have fairly transparent orthographies. The scripts of the two families are different as well. Additionally, there are thousands of dialects. According to the CIA World Fact Book (2009), the literacy rates of India, based on the 2001 census, are 61%, 73.4%, and 47.8%, respectively for the total population, males, and females. India defines literacy as anyone of 7 years of age and above who can read and write with understanding. Based on this definition, according to the recently completed 2011 census data, the rates were 75.06%, 82.14%, and 65.46% for the total population, males, and females, respectively (The Hindu, 2012). Because of the size of the country and different initiatives, there are differences in the literacy rate among the 28 states, with the state of Kerala reporting a literacy rate of 94% and the state of Bihar reporting 64%. Even though, there may be some minor differences in the classification of different levels, broadly 1–12 grades can be divided into primary level, from grades 1 to 4; middle school level, from grades 5 to 8; and high school level, from grades 9 to 11. However, beginning 2013, the levels will be changed to 12 years of formal education with an additional year at the high school level. Majority of the schools are run by the government and the small number of private schools are very expensive. Because of the infrastructure of the country, many schools, especially in rural areas are overcrowded, lack text books and writing materials like pencils and books. However, the private schools have smaller classrooms and equipped with modern facilities like computers. There are also several international schools primarily for expatriates even though many native Indians attend them. India follows the three-language formula in schools, which means, by the time students enter high school, all of them should have learned three languages – Hindi, English, and a regional or another language.

Maldives

The official name of Maldives is the Republic of Maldives and is also known as the Maldive Islands, with about 1200 islands. It is an island nation in the Indian Ocean and has several distinctions such as one of the smallest countries in the world in terms of area as well as in population, with about 328 000 people (World Fact Book, 2009). It is also one of the world's most geographically dispersed countries. Other distinctions include, planet's lowest country and the country with the lowest natural highest point in the world. The capital city is Malé and the official language is Dhivehi, which is an Indo-Aryan language. Dhivehi, similar to Arabic, is also a diglossic language, where there is a clear difference between the spoken and written formats of the language. The CIA World Fact Book (2009) places Maldives having a high literacy rate. Based on the 2006 census data, the rates were 93.8%, 93%, and 94.7%, respectively, for the total population, males, and females. Maldives follows the universal primary education and is free; even though most of the schools offer primary and secondary education, higher education is continued in other countries, such as India, Sri Lanka, and England. English is widely used in schools even though Dhivehi and Arabic are also included in the school curriculum. Based on the CIA World Fact Book, apparently education is successful in the Maldives with such high literacy rate.

Nepal

The official name of Nepal is the Federal Democratic Republic of Nepal and is a landlocked country bordering India, China, and the Himalayan mountain range. The country has a population of about 27 million people and the capital city is Kathmandu. The official language is Nepali or Nepalese, which belongs to the Indo-European language family. It is written in Devanagari script and the orthography is fairly transparent. According to the CIA World Fact Book (2009), the literacy rates, according to 2001 census data, are 48.6%, 62.7%, and 34.9%, respectively for the total population, males, and females. Beginning with 6 years of age, grades 1–5 are referred to as primary level, grades 6–8 are referred to lower secondary, grades 9–10 referred to as secondary school, and grades 11–12 referred to as higher secondary level. There are both private schools and government-supported schools, with most of the private schools having better facilities. Similar to many countries in the region, Nepal also has problems with overcrowded schools, lack of facilities in terms of books and school buildings, as well as easy access to schools.

Pakistan

The Islamic Republic of Pakistan is the official name of Pakistan and is surrounded by India, Afghanistan, and China. The population of Pakistan is approximately 177 million and the capital city is Islamabad. Pakistan has two official languages: Urdu and English. Urdu language belongs to the Indo-European family and the script is written from right to left and is influenced by Persian and Arabic writing systems. As Urdu language is influenced by Arabic language, it exhibits 'diglossic' characteristics having separate forms of formal literary language and a less formal spoken language. According to the CIA World Fact Book (2009), the literacy rates, according to the 2005 census, were 49.9%, 63.0%, and 36.0%, respectively, for the total population, males, and females. The ministry of education of the Government of Pakistan defined literacy as "anyone age 10 and over who can read a newspaper and write a simple letter, in any language" (Ministry of Education, 2009). According to this criterion, the literacy rates, based on the Pakistan Social and Living Standards Measurement Survey 2008–09 census, were 57%, 69%, and 45%, respectively, for the total population, males, and females. Similar to the educational system in other areas of South Asia, the education system of Pakistan has primary level, from grades 1 to 5; middle school level, from grades 6 to 8; high school level, from grades 9 to 10; and higher secondary school, from grades 11 to 12. Both Urdu and English are taught in schools from the beginning. The Ministry of Education (2009) report provided a bleak report on the institutional facilities in public or governmental schools: about 60% of the schools did not

have electricity; about 37% did not have boundary walls and bathroom facilities; about 35% did not have drinking water facilities. The percentages may be even higher in rural areas and girls' schools (p. 148). Memon (2007) further strongly criticized the status of the educational system by commenting about the shortage of teachers, poor laboratory equipments, overcrowded classrooms, and the poor quality of teachers.

Sri Lanka

The official name of Sri Lanka is Democratic Socialist Republic of Sri Lanka. It is an island nation in Indian Ocean near India, with a population of approximately 20 million inhabitants. The capital of Sri Lanka is Colombo. Formerly known as Ceylon till 1972, Sri Lanka has two official languages – Sinhala and Tamil. Sinhala, also referred to as Sinhalese, Singhalese, and locally as Helabasa, belongs to the Indo-Aryan branch of the Indo-European languages. Similar to Dhivehi and Arabic languages, Sinhala language also exhibits 'diglossia,' where there is a distinction between literary (formal written) language and the spoken language and the orthography is fairly transparent. Tamil language belongs to the Dravidian language family and also exhibits 'diglossia,' and the orthography is transparent. Tamil is one of the most widely spoken languages in the world and, according to Harte and Heifetz (1988), has one of the richest literatures in the world. According to the CIA World Fact Book (2009), the literacy rates, according to the 2001 census, were 90.7%, 92.3%, and 89.1%, respectively, for the total population, males, and females.

Similar to the many educational systems in the region, educational system in Sri Lanka can be classified as primary education, from grades 1 to 5; junior secondary level, also referred to as middle school, ranging from grades 6 to 9; and the senior secondary school from grades 10 to 11. Sri Lanka has the universal education for free till grade 9, till 14 years of age, even though the government encourages students to continue schooling. The medium of instruction is either in Sinhala or Tamil languages. However, there are many international and private schools where English is the medium of instruction (Ministry of education, 2012). Since the end of Civil War in 2009, there has been much improvement in the educational system in Sri Lanka. These improvements include establishing schools in the war-torn places and permitting opening of international schools. The private and international schools have modern facilities but the fees are high and hence can be expensive. The provincial and national schools are less expensive as they are funded by the government but suffer from high teacher–pupil ratio and lack of text books and facilities. There have been some concerns expressed by parents and academicians that the educational system is rigid and competitive, not encouraging creative thinking but encouraging rote memory learning.

Conclusions

South Asia, comprising of eight countries, Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka, is one of the most populous regions and perhaps one of the poorest regions in the world, with high malnutrition and high rate of infant mortality (Ibn live, 2010). These reasons also are reflected in the poor educational opportunities for all school-age children. The educational system from grades 1–12 is very similar among South Asian countries with 3–4 broad levels of classification beginning with the primary level lasting for 4–5 years and completing in 2–3 years of high school or higher secondary school. All the countries have free and compulsory education till at least 10 years of age, yet many countries have not strictly followed the rules due to socioeconomic conditions and child labor practices. In general, males have higher percentage of literacy rate and also complete higher grade compared to females. Even though the literacy rate is obtained from the CIA World Fact Book, the rate has to be verified again as how the data were collected is not very clear and may not be based on empirical research. It is interesting, however, to note that all countries encourage learning of two or more languages from early school years.

See also: Asia, Sociocultural Overviews: South Asia; Child Labor; Cross-Cultural Study of Education; Early Education and Care in the World Regions; Education in Sub-Saharan Africa; Education in the Middle East and North Africa; Educational Planning Worldwide; Fertility Transition: South Asia; Illiteracy, Sociology of; Literacy Education; Literacy and Illiteracy, History of; Literacy and Orality; Second Language Learning.

Bibliography

Bose, S., Jalal, A., 2004. Modern South Asia. Routledge, New York.
Business Week, 2006. The World's Happiest Countries. http://images.businessweek.com/ss/06/10/happiest_countries/index_01.htm (accessed 21.04.12.).
Ferguson, C., 1959. Diglossia. Word 15, 325–340.
Harte, G.L., Heifetz, H., 1988. The Forest Book of the Ramayana of Kampan. University of California Press, Berkeley, CA.
The Hindu (2012). http://www.thehindu.com/news/national/article1588270.ece (accessed 19.04.12.).
Ibn live (2010). http://ibnlive.in.com/news/child-mortality-rate-highest-in-india-unicef/57174-17.html (accessed 19.04.12.).
McLeod, J., 2002. The History of India. Greenwood Publishing Group, Westport, CN.
Memon, R.G., 2007. Education in Pakistan: the key issues, problems and the new challenges. Journal of Management and Social Sciences 3, 47–55.
Ministry of Education, Government of Pakistan, 2009. http://finance.gov.pk/survey/chapter_10/10_Education.pdf (accessed 16.04.12.).
Ministry of education, Government of Sri Lanka, 2012. http://www.moe.gov.lk/ (accessed 16.04.12.).
National Center for Education Statistics (NCES), 1999. Reading Report Card for the Nation and the States. Author, Washington, DC.
United Nations Educational, Scientific and Cultural Organization (UNESCO), 2004. The Plurality of Literacy and Its Implications for Policies and Programs. UNESCO Press, Paris, France.
United Nations Educational, Scientific and Cultural Organization (UNESCO), 2007. Institute for Statistics Data. UNESCO Press, Paris, France.
United Nations Educational, Scientific and Cultural Organization (UNESCO), 2011. On the Road to Resilience: Capacity Development with the Ministry of Education in Afghanistan. UNESCO Press, Paris, France.
United Nations Statistical Division, 2011. http://millenniumindicators.un.org/unsd/methods/m49/m49regin.htm (accessed 21.03.12.).
U.S. Dept. of State, 2011. http://photos.state.gov/libraries/afghanistan/231771/PDFs/RFP-Lincoln-Learning-Centers.pdf (accessed 04.04.12.).
White, S., Dillow, S., 2005. Key Concepts and Features of the 2003 National Assessment of Adult Literacy. NCES 2006-471. U.S. Department of Education, Washington, DC. National Center for Education Statistics.
The World Fact Book, 2009. Central Intelligence Agency, Washington, DC. https://www.cia.gov/library/publications/the-world-factbook/index.html.

Education in Sub-Saharan Africa

Sylvia C Kalindi, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

© 2015 Elsevier Ltd. All rights reserved.

Nomenclature

AAF Asante Africa Foundation
EFA Education for All
GER Gross enrollment ratio
GPI Gender parity index
IMF International Monetary Fund
IOB Policy and Operations Evaluation Department, Netherlands Ministry of Foreign Affairs
MDGs Millennium development goals
SACMEQ Southern and Eastern Africa Consortium for Monitoring Educational Quality
UNESCO United Nations Educational Scientific and Cultural Organisation
UNICEF United Nations Children's Fund

Abstract

International declarations, conferences on education as well as international financial institutions have greatly influenced education in Sub-Saharan Africa. This article looks at the current trends of education in the region such as high enrollment rates after the introduction of free basic education, challenges of retaining pupils in the system, poor levels of attainment as well as the prevailing factors in the region affecting the teaching process and profession.

This article also highlights the countries in the region that are faring well on the different issues raised, as well those that are really struggling and later gives some recommendations.

Introduction

Education has a lot of benefits, which the developed as well as developing countries have been enjoying. This includes improved health, increased per capita income, and sustained democracies. Furthermore, education enables people to live with dignity and improve the quality of their lives (IOB, 2011; Yahie, 2000; Spaull, 2011; UNESCO, 2006). Inasmuch as education is a human right, only a few people in Sub-Saharan Africa have the privilege of being educated and so the region cannot really tap benefits such as increased economic growth since the masses are largely steeped in poverty, disease, and other social ills. Although world conferences such as the 1990 Jomtien Conference on Education for All (EFA) and the 2000 World Education Forum in Dakar have helped to highlight the need for concerted efforts in actually making education a human right, 'Education for All' still remains great a challenge in Sub-Saharan Africa. In spite of commitment to international treaties and declarations such as the second and third Millennium Development Goals (MDGs): 'ensuring that all boys and girls *complete* a full course of primary schooling by the year 2015' and '*eliminating gender disparities in primary and secondary schools*', respectively, all education indicators are below world and developing country averages. Most children especially girls are still deprived of the right to education (World Bank, 2008; UNESCO, 2009).

This article will review the performance and challenges of different countries in Sub-Saharan Africa with regard to salient current features of basic education. This includes enrollment rates, attainment levels, and how teachers play a key role in all this.

Background

The 1960s and early 1970s were postindependence era for most countries in Sub-Saharan Africa and they were marked with economic boom. The economic gains were, however, doomed as prices of primary commodities fell, this was further worsened by the oil shocks of the early 1980s as it increased the government deficits and foreign debt (Yahie, 2000). In the 1980s, attempts were made by major financial institutions, that is, the World Bank and International Monetary Fund (IMF) to increase financial stability of developing nations through introduction of Structural Adjustment Programmes (SAP). SAP, however, had negative implications on the public sector, as Kruijer (2010: p. 20) points out that education provision "… was confronted with neo-liberal policies that limited the budget for social expenditures and governments were forced to make choices that restricted public education".

The World Bank (2009) further pointed out that due to the economic crises, governments in developing countries had to reduce the public spending per pupil in primary and secondary

education through stringent austerity measures such as freezing teachers' salaries, increasing class size, and reintroducing school fees. As a result, developing countries experienced a sharp decline in the quality of education provided in institutions of learning subsequent to the 1980s. For example, in Zambia, this led to dilapidated and grossly overcrowded classrooms, that had few or no instructional materials, and demoralized well-trained teachers (Kelly, 1991). Such devastating conditions generally contributed to the impoverishment of the people in the whole Sub-Saharan region. According to Sinyolo (2007), Kelly (1991), as well as Kruijer (2010) the international sanctions imposed on the public sector wage bills forced governments to employ unqualified, volunteer, and contract teachers, which ultimately damaged the quality of education.

With the 1990s came less stringent humane adjustments as both the IMF and the World Bank advocated for good governance and national governments' participation in poverty reduction. After the mid-1990s, national governments like Malawi and Uganda began once again the process of abolishing school fees, often because of promises made during election campaigns (Sinyolo, 2007). In subsequent publications, the World Bank emphasized the need for good quality fee-free primary education, stating it as the basis in governments' poverty reduction strategies (World Bank, 2009).

Clearly, a lot of external as well as internal efforts and energies have been combined in working toward meeting the immediate demands of EFA by 2015 together with those of the MDGs. The results of the policies that focus on meeting EFA and MDG, such as Universal Primary Education are varied but generally influence critical areas of education such as enrollment and retention, as well as teacher supply and literacy.

Enrollment and Retention

Preprimary Schooling

In the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) working paper, Spaull (2011) emphasizes that preschools play an important role in readying the student for primary education by imparting the emotional, intellectual, and social skills necessary to succeed at school. It is further argued that children who attend preschool are more likely to succeed at primary school and less likely to repeat and drop out. In spite of this positive outlook, preprimary education is still viewed as a luxury for nearly all children. Anyhow, access to preprimary education has steadily increased over the years in Sub-Saharan Africa. UNESCO (2009) reports indicate that in 2006, nearly 9 million children were enrolled in preprimary education, up from some 5 million in 1999. In spite of this, the average gross enrollment ratio remained low at 14% in 2006 with huge regional disparities. Spaull (2011) also further indicates that some countries, such as Equatorial Guinea, Ghana, Liberia, and South Africa, have recorded significant enrollment improvements while a few nations like Seychelles and Mauritius have records of over 100% improvement.

Generally, preschool education is mostly provided by private institutions and is concentrated in urban and wealthy areas as it tends to be very expensive. However, in countries like "South Africa and Namibia the government plays an active role in the provision and monitoring of preschool education" (Spaull, 2011: p. 54). (Refer to Figure A.1 in the Appendix for the enrollment trend between 1999 and 2006.)

Primary Education

In Sub-Saharan Africa, primary education usually stretches over a period of 6–7 years, starting from grade 1 to 6, in places like Nigeria, and all the way to grade 7 in Zambia Mozambique or Zimbabwe.

For the majority of children, this is the first time they set foot in a classroom and enrollment age across the region varies between 6 and 7 years. For example, in Namibia, Benin, and Cameroon children start school at age 6 while in most other countries the formal enrollment age is 7 (World Bank, 2008). It is, however, worth noting that issues of early and late school entry are quite widespread with some countries having extreme cases of underage or overage children. For instance, household surveys in Tanzania, Ethiopia, and Ghana indicated that more than 60% of the children in primary school were above the required age for their grade while for countries like Niger, Cameroon, and Uganda underage pupils made up a large share of first graders (UNESCO, 2009). A recent study by Fink et al. (2012) conducted in Zambia on early childhood development also cited early enrollment in school as common practice, with about 25% of children starting school at age 6 or earlier. Countries in the Sub-Saharan region have recorded a sharp increase in the number of children entering primary school since the Dakar conference. The average primary net enrollment between 1999 and 2006 increased from 56 to 70% with some countries such as Tanzania, Zambia, Madagascar, and Benin registering very impressive progress moving from enrollment ratios of 50–70% in 1999 to levels in excess of 80% in 2006 (UNESCO, 2008b).

Tremendous improvement has also been noticed in some countries like Rwanda where between 1998 and 2009 "enrolment in primary school almost doubled with an annual growth rate of 5.4%" (World Bank, 2011: p. 59). Ethiopia is another country that managed to double its enrollment ratio from 34% in 1999 to 71% in 2007 (UNESCO, 2009). Successful strategies have included ambitious school construction programs in rural areas, abolition of school tuition fees and targeting disadvantaged groups like orphans and vulnerable children as has been noticed in several countries. For instance, in Zambia, meeting the increased demand for primary education has also meant constructing more schools, that is from having approximately 5300 basic schools in 2000 to about 8400 in 2010 (with the largest increase in the number of community schools) (World Bank, 2011; IOB, 2011).

On the other hand, there are some West African countries were the enrollment results have been less positive, especially for girls. For example, in spite of some progress made since 1999, Nigeria only showed a net enrollment rate of 63% with gender parity index of 0.83 in 2005. According to UNESCO (2008b), this implies that 48% of girls in the primary school age were out of school.

Nonetheless, enrollment disparities within countries are also very pronounced and sometimes even greater than between countries. Generally, children living in the urban areas

are enrolled at higher rates than their rural counterparts and the gaps tend to increase as they move up the educational ladder. "Disparities within the countries based on wealth, gender, race, language, ethnic group or disability hinder progress towards Universal Primary Education" (Kruijer, 2010: p. 19). In Burkina Faso, Chad, Ethiopia, Mali, and the Niger, Kruijer (2010) further indicates that children from the richest 20% are from three to about four times more likely to attend school than children from the poorest quintile.

So inasmuch as there is a general increment in the enrollment rates of the Sub-Saharan region, it is worth noting that there is still a large number of children below the age of 7 who are still not enrolled in school because of perceived dangers on the way and lack of nearby schools (AAF, 2011; IOB, 2011; UNESCO, 2009). IOB (2011) also observes that in 2006 almost one-third of the school age population in Sub-Saharan Africa, about 35 million children were not enrolled in schools.

Compared to other regions in the world, countries in Sub-Saharan Africa have low enrollment rates and strong gender disparities and inequalities. Significant gender disparities in primary enrollment rates are still reported in many countries. For instance, as is indicated in **Figure 1** below, the Central African Republic, Chad, Cote d'Ivoire, Mali, and Niger have fewer than 80 girls enrolled in primary school for every 100 boys in 2006 (UNESCO, 2009).

Primary Completion Rates

Although there have been marked improvements in access to education for most countries in the Sub-Saharan Africa, there are still large differences in dropout and retention rates. According to Lloyd and Hewitt (2009) Sub-Saharan Africa has the lowest primary completion rates compared to other regions of the world. In instances where there is high dropout and repetition rate many children have a low survival rate to the end of primary school.

"Excessive grade repetition is a common problem in Sub-Saharan Africa" (Spaull, 2011: p. 53). Students who do not achieve the required grade are held back and repeat that same grade in the hope that they will acquire the skills the second time around. UNESCO (2009) indicated that in 2006, the median percentage of repeaters in the region was just above 13%.

Factors such as disease, hunger, child labor, and distant schools make children to be absent from school and eventually drop out. For example, AAF (2011) reveals that in Kenya, 4 out of 10 children miss school on a daily basis mainly due to family support demands and distant schools.

SACMEQ studies done in some countries in Southern Africa have shown variations in reaching grade 6 (as indicated in the Figure A.2 in the Appendix), where Mozambique has a survival rate of 60% while South Africa, Namibia, and Botswana all have high rates of about 90%. In this case, countries like Mozambique have high dropout rates affecting the number of children reaching grade 6 and eventually secondary school (Spaull, 2011).

It is, however, interesting to note that over the past 30 years, the trends in the percentage of boys and girls completing primary school in the Sub-Saharan region have shown improvements (Lloyd and Hewitt, 2009), as can be seen in **Figure 2** below.

Figure 2 clearly shows that in early 2000, approximately 60% of boys aged 10–14 had completed primary school, and over the past 30 years, there has been a slow rise in the completion rate for boys. Over the same period, girls' completion rates have risen steadily almost doubling from 30 to 56% indicating that the gender gap in the region has narrowed steadily and consistently.

Current primary completion rates in the region vary from as low as 8% for girls and 19% for boys in Niger to as high as 90% for girls and 86% for boys in South Africa (Lloyd and Hewitt, 2009).

Secondary

Depending on the country, young people are generally expected to enroll into secondary school from the age of 12 or 14 and are expected to be in the system for 6 or 5 years, respectively. While participation in primary education is expanding, access to secondary education remains limited for most young people in Sub-Saharan Africa. UNESCO (2009) monthly reports suggest that overall enrollment in the region is rising with over 12 million more students in 2006, up from 20.6 million in 1999. Despite this significant trend, 78 million of the region's secondary school age children were not enrolled in secondary school in 2006.

There are a number of factors causing low participation in secondary schools in this region, such as low primary enrollment, low transition from primary to secondary, or high dropout rate. In addition, a large proportion of the schooling capacity is held by students who are repeating grades (World Bank, 2008).

There exists a close relation between secondary and primary enrollments. On average, countries with very low secondary enrollment also have a relatively low primary enrollment and in addition also, a significant gender imbalance. The World Bank (2008) working paper further shows that for countries like South Africa and Namibia with high primary completion rates, secondary enrollments are equally high whereas transition rate from primary to secondary in Tanzania is 21%, Uganda 29%, and Kenya at 53%. (Refer to Table A.1 in the Appendix for a classification). The under representation of girls is a common characteristic of countries with low enrollments rates like Eritrea, Mali, Senegal, and Tanzania. (Refer to Figure A.3 in the Appendix for details).

Transition from lower secondary to higher secondary is a critical dropout point for most countries in sub-Saharan Africa education systems. In 2006, the regional average gross enrollment rate was higher in lower secondary (38%) than upper secondary (24%) with differences in participation rate between the two levels (UNESCO, 2009). Like primary education, within-country inequalities in secondary education are even more marked than those among countries. Secondary school attendance and survival rates to the last grade are marked by disparities related to household wealth and language. UNESCO (2009) also shows that attendance rates of richer households are much more than those of poorer households for instance in Burkina Faso, Guinea, Mali, and Niger.

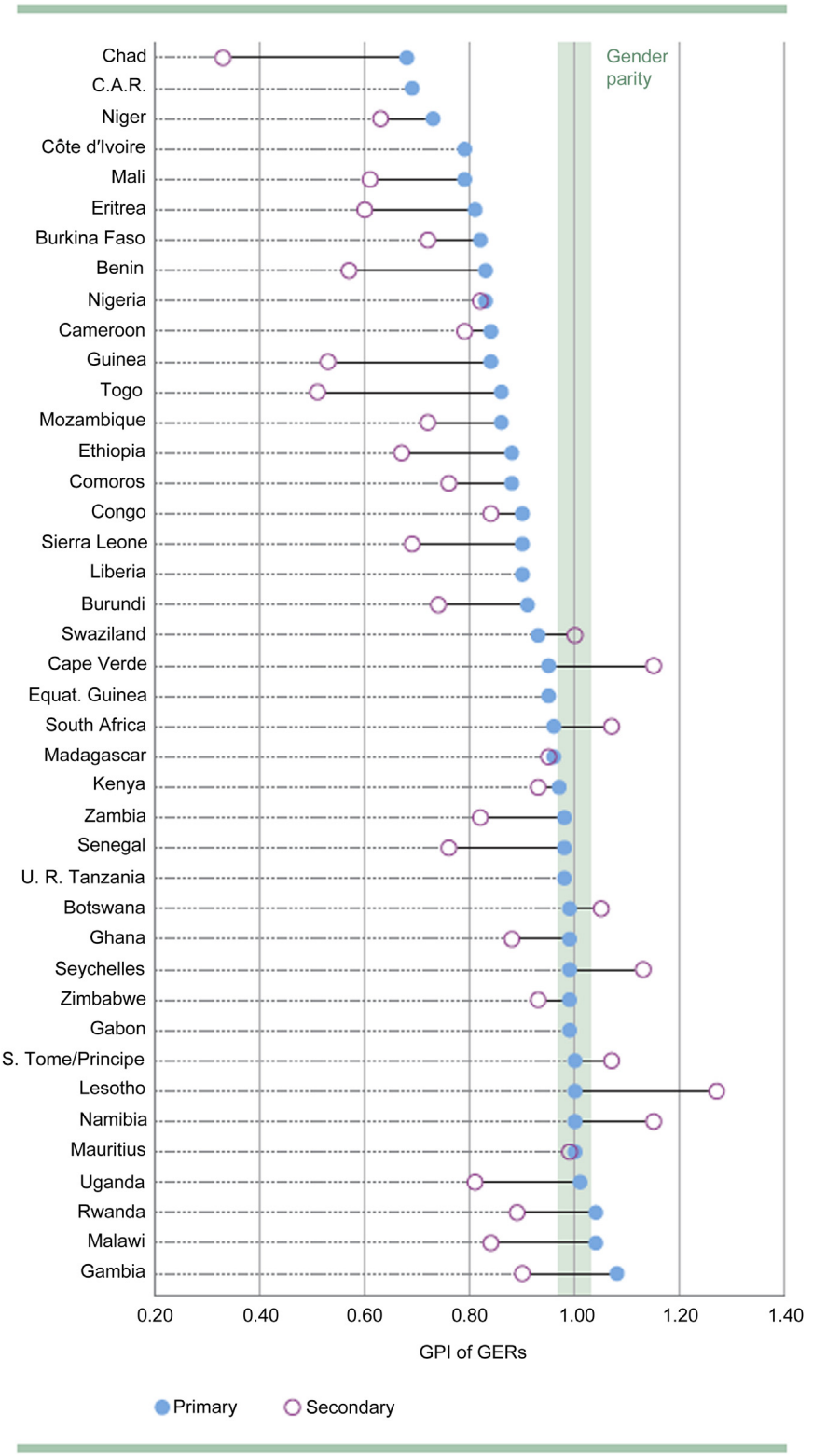

Figure 1 Changes in gender disparities in primary and secondary enrollment ratios, 2006. GPI, Gross enrollment ratio; GER, Gender parity index. Source: UNESCO, 2009. Regional Overview of Sub-Saharan Africa – EFA Global Monitoring Report, Overcoming Inequality: Why Governance Matters. UNESCO, Paris, p. 8.

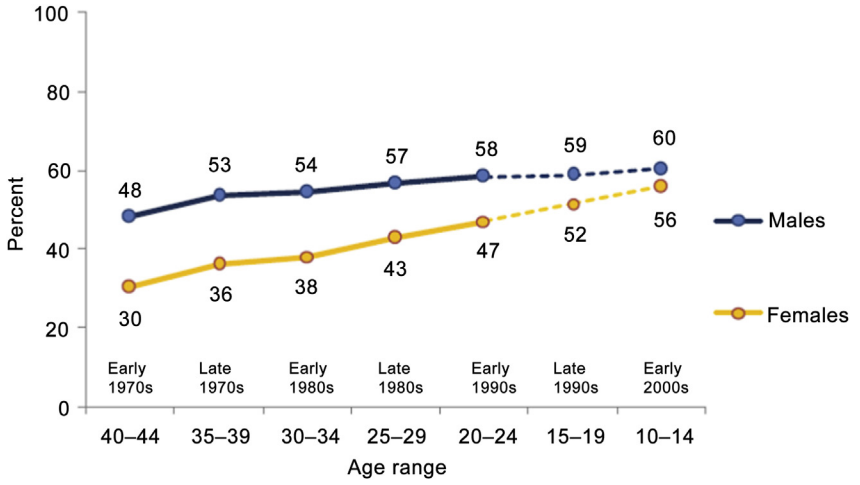

Figure 2 Trends in completion of primary school, 33 African countries (population weighted). Source: DHS, MICS-2 data in Lloyd, C.B., Hewitt, P.C., 2009. Poverty Gender and Youth, Educational Inequalities in the Midst of Poverty: Diversity across Africa in Educational Outcomes, Working Paper. Population Council, New York, p. 17.

Inasmuch as enrollment rates and issues of attendance and grade progression are good indicators of achieving universal primary education, they are not entirely adequate because it is not easy to know how well students are performing in the classroom.

Learning Achievement/Educational Quality

High enrollments brought on by the abolishment of school tuition fees in Sub-Saharan Africa have compromised the quality of basic education. One area that has been very crucial is with regard to 'what children learn' and their ability to use the acquired skills. According to the World Bank (2008), the low enrollment in secondary education has been related to broader problems in school attainment and quality of schooling. Student competence levels in numeracy and literacy as well as national examinations have been good indicators of quality with regard to school attainment.

While literacy improves with schooling, the correlation between the two is far from perfect in most countries in Sub-Saharan Africa. In Tanzania, for instance, 1 out of 5 pupils who finish primary education cannot read at a second grade level in Kiswahili and only 5 out of 10 children can read in English upon completion of primary school, which is a debilitating factor as secondary school is taught in English only. In Kenya, 7 out of 10 children in third grade cannot do second grade mathematics and reading work (AAF, 2011).

UNESCO (2006) further highlights the link between poverty and illiteracy in countries with particularly low levels of literacy such as Cote d'Ivoire, Guinea Bissau, Rwanda, Senegal, Togo, and Sierra Leone where the literacy gap between the poorest and wealthiest households is more than 40%.

Furthermore, in some countries like Ethiopia with low overall literacy rates, UNESCO (2006) points out that they are also faced with large urban/rural disparities where urban literacy rates are 74% while rural literacy rates are 23%.

Results from regional assessments have highlighted poor learning outcomes in Sub-Saharan Africa, for example, in 1999 Uganda carried out assessments to determine overall achievement levels in grades 3 and 6 in English literacy and numeracy. Results indicated that in spite of the general increase in achievement levels over time, fewer than half of the children reached defined competence levels in English literacy (UNESCO, 2008a).

Generally, it has been observed that large numbers of children who reach the last grade in primary school have weak literacy and numeracy skills. Results from the SACMEQ study of 2000–02 conducted in 13 Southern Africa countries indicate that hardly any grade 6 students in Lesotho, Malawi, and Zambia reached one of the four highest levels of the numeracy scale while more than a third did so in Kenya, Mauritius, and Seychelles (UNESCO, 2006).

Other SACMEQ assessments conducted in 2000 and 2007 on grade 6 children in various countries indicate that Namibia, Swaziland, Lesotho, and Tanzania registered substantial improvements for both reading and mathematics (SACMEQ, 2010). On the other hand, Mozambique registered substantial deterioration in achievement in both reading and mathematics, as indicated in **Table 1** below.

Table 1 highlights the performance of pupils in reading and mathematics for the years 2000 and 2007, with the arrows indicating either an increase or decrease in performance.

UNESCO (2009) has also observed that girls and boys achieve very different outcomes in school not just in overall performance but also by subject. Recent studies have shown that girls continue to outperform boys in reading literacy and language arts. This is evident in a diverse group of countries including those with great gender disparities in school participation such as Burkina Faso. Although boys have generally outperformed girls in mathematics in all grades of primary and secondary education, there is a general observation that this trend is slowly changing as the girls are increasingly performing at levels equal to or better than boys, for instance in

Table 1 Levels and trends in pupil achievement for SACMEQ countries

	Pupil reading score			Pupil mathematics score		
	2000	2007		2000	2007	
Botswana	521.1	534.6	▲	512.9	520.5	▶
Kenya	546.5	543.1	▶	563.3	557.0	▶
Lesotho	451.2	467.9	▲	447.2	476.9	▲
Malawi	428.9	433.5	▶	432.9	447.0	▲
Mauritius	536.4	573.5	▲	584.6	623.3	▲
Mozambique	516.7	476.0	▼	530.0	483.8	▼
Namibia	448.8	496.9	▲	430.9	471.0	▲
Seychelles	582.0	575.1	▶	554.3	550.7	▶
South Africa	492.3	494.9	▶	486.1	494.8	▶
Swaziland	529.6	549.4	▲	516.5	540.8	▲
Tanzania	545.9	577.8	▲	522.4	552.7	▲
Uganda	482.4	478.7	▶	506.3	481.9	▼
Zambia	440.1	434.4	▶	435.2	435.2	▶
Zanzibar	478.2	533.9	▲	478.1	486.2	▶
Zimbabwe[a]	504.7	507.7	▶	××	519.8	××
SACMEQ	**500.0**	**511.8**	▲	**500.0**	**509.5**	▶

▲ Increased by 10 points or more ▶ Minimal change (less than ±10) ▼ Decreased by 10 points or more

[a]Zimbabwe did not participate in the SACMEQ II Project (2000) and the value given in grey (reading) is from the SACMEQ I Project (1995).

Source: SACMEQ, 2010. Policy Issues – What are the Levels and Trends in Reading and Mathematics Achievement? SACMEQ, p. 4.

Francophone Africa. SACMEQ results of sixth graders have also shown significant male advantages in mathematics for Zambia, Kenya, Mozambique, Zanzibar, and Tanzania but not for Seychelles (UNESCO, 2009).

The deep learning deficits in the region are also confirmed by national examination results. For example, in Zambia with the government's objective that every child completes 9 years of basic education, grade 9 examinations have become more important over time (IOB, 2011). This is reflected in the growth in the number of candidates as is indicated in **Figure 3** below.

Since 2003, the grade 9 examination results have indicated a slight improvement with girls, although on the overall boys achieved significantly better in mathematics. However, in 2009

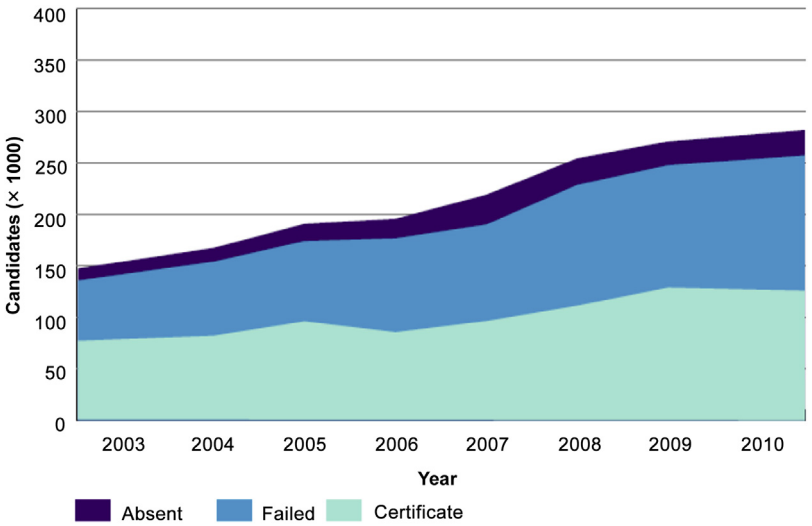

Figure 3 Grade 9 examination results, 2003–10. Source: Examinations Council of Zambia, 2010. Learning Achievement at the Middle Basic Level, Zambia's National Assessment Survey Report in IOB, 2011. Unfinished Business: Making a Difference in Basic Education. An Evaluation into the Impact of Education Policies in Zambia and the Role of Budget Support. Ministry of Foreign Affairs of the Netherlands, The Hague, p. 98.

girls showed larger improvements than boys both in English and mathematics (IOB, 2011). (See Table A.2 in Appendix for low performance since maximum score is 100 as well as Table A.3 for the grade 7 examinations, which have a total score of 60 points.)

According to Ahonen and Aro (2011), there are a number of factors that may explain the poor attainment levels not just in Zambia but the entire Sub-Saharan region. These include teacher shortages, overcrowded classrooms, poor teaching, primary language interference, lack of printed material in local languages, illiterate parents, and lack of parental support in the learning process.

Teachers and Factors Affecting the Teaching Process

Teachers are the main providers of quality education. In this regard, the way they are recruited, trained, and deployed across the schools is vital for improving learning outcomes. As already alluded to, the introduction of free basic education in Sub-Saharan Africa automatically resulted in more children in the classrooms thereby increasing the pupil teacher ratio. This situation is compounded by the sharp teacher shortages especially in the rural areas. In Tanzania and Kenya, the AAF (2011) reported the pupil teacher ratio to be 72:1 and 47:1, respectively.

UNESCO studies indicate that in 2006 the average pupil teacher ratio across the region was 45:1, up from 41:1 in 1999. However, in countries like Chad, Mozambique, and Rwanda, UNESCO (2009) stresses that the national primary pupil teacher ratio exceeded 60:1.

Larger disparities of pupil teacher ratios exist within countries and they are influenced by location, wealth, and type of school, for instance in the Bayelsa state of Nigeria the pupil teacher ratio is five times higher than Lagos, while Zambian rural schools have a pupil teacher ratio higher than 80 (UNESCO, 2009; IOB, 2011).

Trained teachers are in short supply in many countries in the region and the teacher shortage is often an indication of inadequate investment in education. At primary level, the teacher education is often fragmented, incomplete and in some cases nonexistent. Although governments have realized the need to recruit additional teachers to meet the increasing enrollment in primary education it is difficult to train new teachers when resources are so scarce. For many countries in the region investment in education is inadequate because the governments heavily depend on external funding to support basic education policy and practice. Financial institutions such as the World Bank and IMF are some of the major founders of education programs in developing countries. Kruijer (2010) indicates that in countries like Chad, Togo, and Madagascar fewer than 40% of teachers are trained. To speedily reduce the pupil teacher ratio governments especially in West Africa have decided to recruit contract teachers to work outside the civil service structure. In addition to this, the author highlights that the cost of contract teachers is much lower than civil servant teachers and the length of their education is much shorter.

Due to financial constraints as well as agreements reached with international financial institutions like World Bank, some countries in the region could not deploy teachers coming out of colleges and universities for a long time. For instance, Sinyolo (2007) explains that Kenya and Zambia had 40 000 and 15 000 unemployed qualified teachers respectively, and yet because of the high pupil teacher ratio resorted to using volunteer, unqualified as well as contract teachers. In 2006, Zambia had over 7000 unqualified teachers working in schools while other countries in the region like Gambia had 1600, Lesotho had over 6600 with Tanzania having over 10 000 unqualified teachers (Sinyolo, 2007).

Nonetheless, the use of unqualified teachers brings severe deficiencies which outweigh the financial advantages. Those with limited professional education are likely to provide lower quality education and the fact that they are deployed in the rural understaffed schools means the rural children's attainment levels are adversely affected, further exasperating the rural urban inequalities. For example, in Ghana, untrained teachers are concentrated in the north, which has the lowest economic development and the most out of school children (Kruijer, 2010; UNESCO, 2009).

A number of countries, including those were the IMF had put caps on the public sector wage bill in the region, are, however, making positive progress toward having and recruiting well-trained teachers in the primary schools. Huge progress has been made for instance in Namibia where more than 90% of the teachers had the required training in 2005, up from 29% in 1999. The sharp increase resulted from the policy to upgrade the teacher qualification and replace untrained teachers with trained ones (Sinyolo, 2007; UNESCO, 2008a). A number of countries in Sub-Saharan Africa have been involved in the upgrading program of unqualified teachers, for example, Malawi, Tanzania, and Nigeria. Anyhow, as Kruijer (2010) has pointed out it is not clear if the participation in an upgrading program leads to improved classroom practices. Important to note is the fact that the average academic level of unqualified teachers is low.

Kruijer (2010) further highlights a research report on in-service education in Namibia which revealed that more than three-quarters of the sample of teachers did not demonstrate the desired teaching approaches. Possible reasons for this situation could be the poor teaching conditions that teachers are exposed to such as lack of accommodation, physically inadequate classrooms, a lack of furniture and teaching material as well as high pupil teacher ratio which make it impossible to use the newly learnt teaching skills. Furthermore, when teachers are faced with scarcity of resources and numerous limitations they tend to fall back on their previous traditional schooling practices, such as lecturing (Kruijer, 2010; Kelly, 1991; UNESCO, 2008a).

The World Bank (2008) suggests that other factors that affect the quality of teaching include low salaries which leads to low teacher morale and also results in teachers taking extra teaching loads or other paid work to secure a decent economic outcome. **Table 2** above shows teachers' salaries in 6 countries.

Table 2 Teacher's gross salaries per month in US Dollars (2005–06)

Level	*The Gambia*	*Kenya*	*Lesotho*	*Uganda*	*Tanzania*	*Zambia*
Primary	60	152	177	115	20	200
Lower secondary	75	272	265	144	95	250
Upper secondary	90	305	638	260	125	325

Source: Sinyolo, D., 2007. Education International. Teacher Supply, Recruitment and Retention in Six Anglophone Sub-Saharan African Countries in The Gambia, Kenya, Lesotho, Tanzania, Uganda and Zambia. Education International, Brussels, p. 56.

Due to low salaries and poor conditions of service in most of Sub-Saharan Africa, the status of the teaching profession is declining. This ultimately leads to severe retention problems coupled with high rates of absenteeism. UNESCO (2009) points out that teacher absenteeism rates in primary schools averaged 27% in Uganda. The author further highlights that in other countries such as Ghana, South Africa and Tanzania teacher absenteeism was more pronounced in public schools, schools with poorer infrastructure, in rural areas, in poorer states and in schools serving children from lower socioeconomic backgrounds.

Teacher absenteeism is also associated with low student performance because of the inadequate covering of the curriculum and shorter time on task. SACMEQ studies found that grade 6 South African mathematics and reading teachers were absent for twice as many days as Botswana and Namibian teachers while three times as many days as the average Mozambican mathematics teacher (Spaull, 2011).

Illness and financial problems linked to irregular pay are some of the reasons for high absent rates of teachers in Sub-Saharan Africa. Although teacher mortality rates due to HIV/AIDS are decreasing and reasonably stable, the epidemic is damaging lives and educational systems. For example, in Kenya, about 14 500 are estimated to be HIV positive and 4–6 teachers die each day due to AIDS and opportunistic infections (UNESCO, 2009).

Conclusion

It has been observed that the abolishment of primary school tuition fees in the mid-1990s has had a very positive effect with regard to allowing more children access basic education. Once the policy of free basic education was introduced in the Sub-Saharan region there was a 'new' demand for education from the poor masses that were initially disadvantaged with the introduction cost sharing measures by international financial institutions in the 1980s. This led to increased enrollment rates at primary school across the region (UNESCO, 2009; World Bank, 2009).

Ironically, most governments in the region were not ready for the new demand and so experienced a lot of serious challenges in trying to adequately cater for the increased number of pupils (SACMEQ, 2010). The challenges included lack of classrooms and appropriate furniture, lack of resources in the form of student books and teachers' manuals and lack of trained teachers which ultimately led to questioning the quality of the education offered (IOB, 2011; Kruijer, 2010). What further compounded the challenge was the lack of morale teachers experienced due to low salaries, lack of incentives for progression in the profession and the general poor working conditions they were subjected to. This led to high rates of attrition and absenteeism making the shortage of teachers more pronounced. To fill the gap governments resorted to cheap labor in the form of volunteers and contract teachers who are easy to fire and hire or even send to remote parts of countries which had marked problems of teacher shortages. Seeing most contract teachers were untrained they lacked preservice training and governments in the region quickly provided in-service training to fill their knowledge gap (Sinyolo, 2007). However, the intermittent training has not been very effective as it has been observed that the pupils have been negatively affected in that even after completing primary education they are not ready for tasks in secondary school. This is mainly because of the low levels of literacy and numeracy as well as the low passing rate in examinations.

Recommendations

To completely achieve universal primary education, it is cardinal for the governments in the Sub-Saharan region to set up good collective bargain structures and ultimately allocate more funding to construction of new classrooms that are inclusive even to disabled children. This will also help to improve the teacher pupil contact hours and will eventually reduce the pupil teacher ratio. Furthermore, infrastructure should be advanced with resources such as solar power, electricity, and laboratories. Teacher training also needs beefing up to ensure the presence of qualified teaching personnel who can teach not only the academic but also vocational subjects in order to prepare dropouts with some life skills. Alongside this, the general working conditions of teachers including salaries, housing, and other incentives needed to attract teachers to rural areas should be improved. As regards, the in-service component of the upgrading program, it is important that it be closely monitored and consistently evaluated to allow for adequate mentoring of the students. Additionally, the remedial programs for repeaters both in primary and secondary schools equally need consistent evaluation to make the repeating process more effective.

Appendix

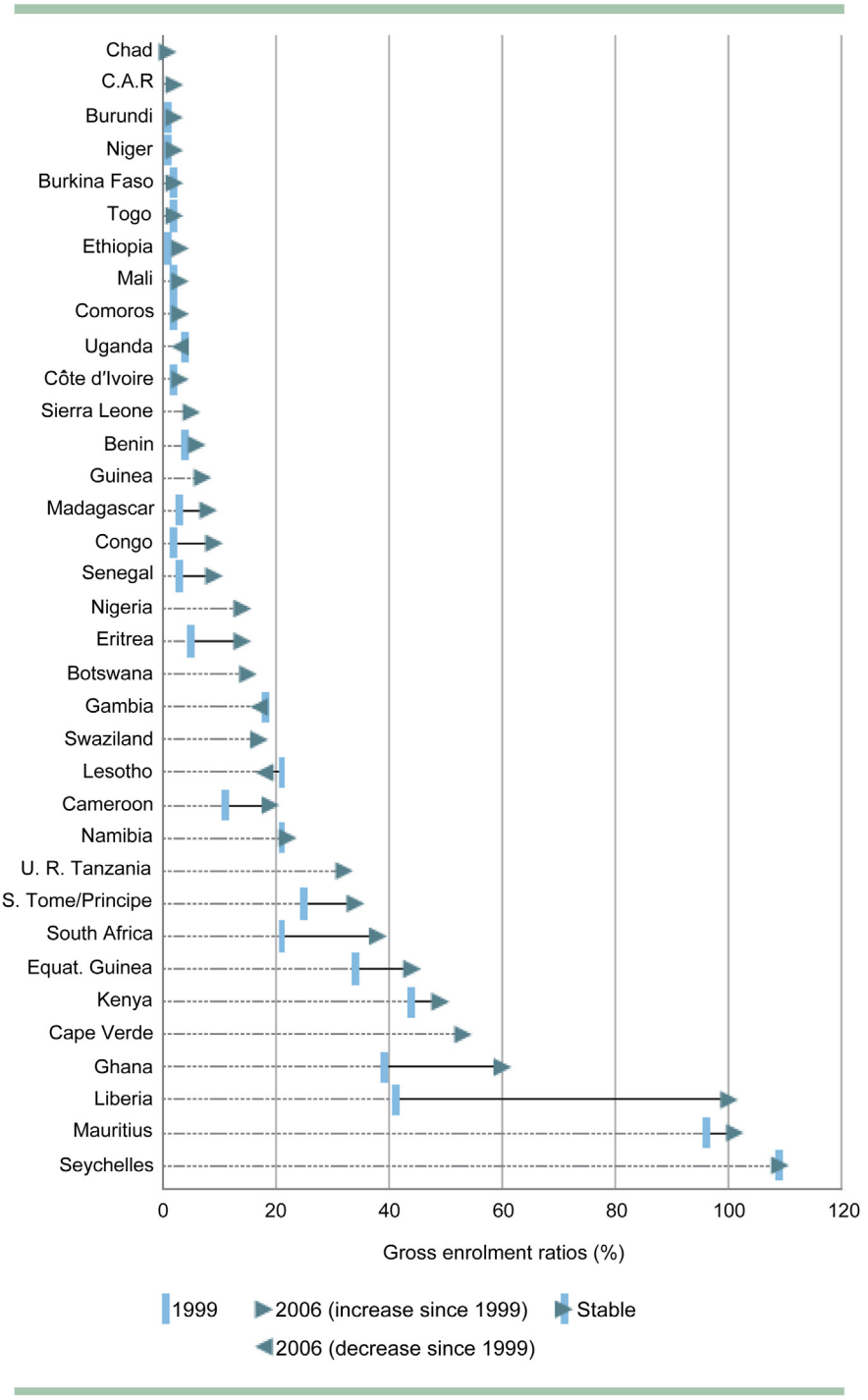

Figure A.1 Changes in preprimary gross enrollment ratios between 1999 and 2006. Source: UNESCO, 2009. Regional Overview of Sub-Saharan Africa – EFA Global Monitoring Report, Overcoming Inequality: Why Governance Matters. UNESCO, Paris, p. 2.

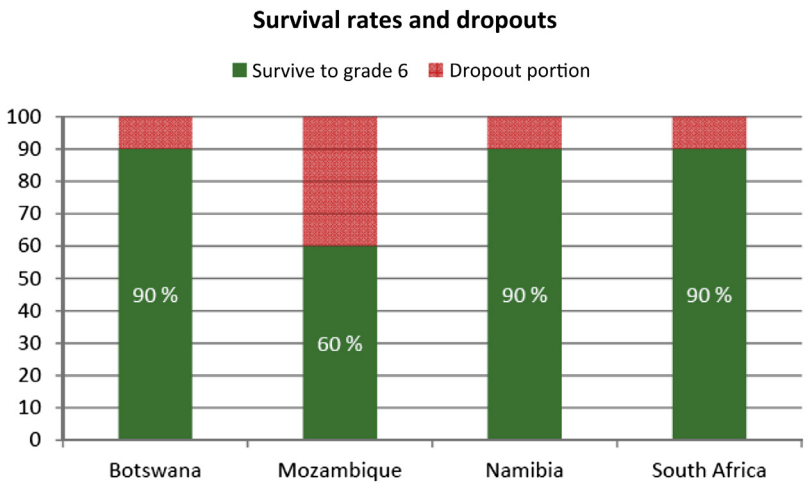

Figure A.2 Survival rates and Dropouts in Mozambique, Namibia, Botswana, and South Africa. Source: Spaull, N., 2011, SACMEQ Working Paper: Primary School Performance in Botswana, Namibia, Mozambique and South Africa. South and Eastern Africa Consortium for Monitoring Educational Quality SACMEQ, p. 42.

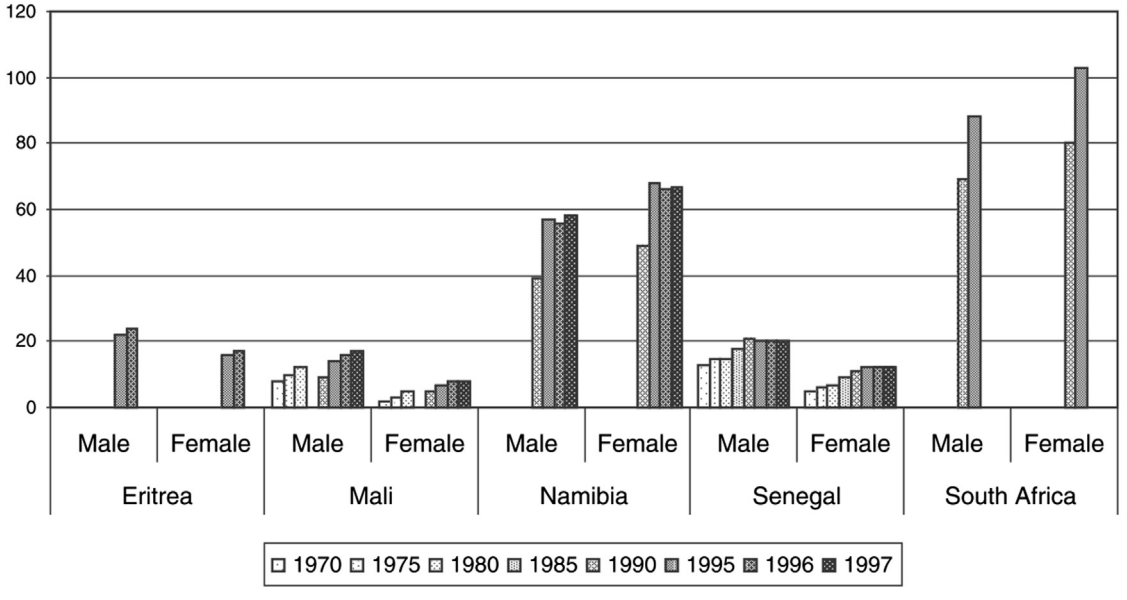

Figure A.3 Gross enrolment rates of males and females in secondary schools in five different countries. Source: World Bank, 2008. Transitions in Secondary Education in Sub-Saharan Africa, Equity and Efficiency Issues. World Bank Working Paper. World Bank, Washington, p. 6.

Table A.1 African countries classified by level of secondary enrollment in education, 1996

Below 10%	*10–24%*	*25–49%*	*50% and above*
Burkina Faso[a]	Benin	Cameroon	Botswana
Burundi[a]	Eritrea	Cote d'Ivoire	Cape Verde
Chad	Ethiopia	Democratic Republic of Congo	Congo
Mozambique	Kenya	Gambia	Mauritius
Niger	Madagascar	Ghana[a]	Namibia
Tanzania	Malawi	Lesotho	South Africa
Rwanda[a]	Mali	Togo	Swaziland
	Nigeria	Zambia	
	Senegal	Zimbabwe	
	Sudan		
	Uganda		
	Angola[a]		
	Central African Republic		
	Guinea		
	Sierra Leone[a]		
	Guinea		

[a]*Data from 1990.*
Source: UNESCO, 2000. World Education Report in World Bank, 2008. Transitions in Secondary Education in Sub-Saharan Africa, Equity and Efficiency Issues, World Bank Working Paper. World Bank, Washington, p. 4.

Table A.2 Grade 9 results for English and mathematics, 2007–09

	2007	*2008*	*2009*
English			
Male	45.0	43.6	45.3
Female	45.1	44.0	46.1
Total	45.0	43.8	45.7
Mathematics			
Male	38.5	36.7	42.8
Female	33.7	32.5	40.5
Total	36.3	34.7	41.7

Source: Examinations Council of Zambia, 2010. Learning Achievement at the Middle Basic Level, Zambia's National Assessment Survey Report in IOB, 2011. Unfinished Business: Making a Difference in Basic Education. An Evaluation into the Impact of Education Policies in Zambia and the Role of Budget Support. Ministry of Foreign Affairs of the Netherlands, The Hague, p. 99.

Table A.3 Examination results for English and mathematics (grade 7), 2001–09

		All pupils	*Rural*	*Urban*
English	2001	29.9	29.2	31.0
	2003	28.1	27.0	30.0
	2005	29.9	28.1	33.1
	2006	29.6	27.8	32.7
	2008	30.6	28.2	34.5
	2009	29.0	26.3	33.1
Mathematics	2001	30.6	31.0	30.1
	2003	28.6	28.8	28.2
	2005	28.3	28.0	28.6
	2006	26.3	26.0	26.8
	2008	29.7	28.8	31.0
	2009	27.7	26.6	29.2

Source: Examinations Council of Zambia, 2010. Learning Achievement at the Middle Basic Level, Zambia's National Assessment Survey Report In IOB, 2011. Unfinished Business: Making a Difference in Basic Education. An Evaluation into the Impact of Education Policies in Zambia and the Role of Budget Support. Ministry of Foreign Affairs of the Netherlands, The Hague, p. 87.

See also: Conflict: Education and Youth; Early Education and Care in the World Regions; Education (Primary and Secondary Schools) and Gender; Education in the Middle East and North Africa; Illiteracy, Sociology of; Literacy Education; Motivation, Learning, and Instruction; Teacher Behaviours and Student Outcomes; Teacher Education.

Bibliography

AAF, 2011. Annual Report Educating Children, Transforming Worlds. Asante Africa Foundation, Oakland, CA.

Ahonen, T., Aro, T. (Eds.), 2011. Assessment of Learning Disabilities: Cooperation Between Teachers, Psychologists and Parents. Uniprint, Jyvaskyla. African Edition.

Fink, G., Matafwali, B., Moucheraud, C., Zuilkowski, S.S., 2012. The Zambia Early Childhood Development Project: 2010 Assessment Final Report. Harvard University, Cambridge, MA.

IOB, 2011. Unfinished Business: Making a Difference in Basic Education. An Evaluation into the Impact of Education Policies in Zambia and the Role of Budget Support. Ministry of Foreign Affairs of the Netherlands, The Hague.

Kelly, M.J., 1991. The financing and costing of basic education for all. In: Kelly, M.J. (Ed.), National Conference On Education for All, vol. 2. Conference Papers. National Task Force on Education For All, Lusaka, pp. 87–90, 116, 132.

Kruijer, H., 2010. Education International. Learning How to Teach, the Upgrading of Unqualified Primary Teachers in Sub-Saharan Africa. Education International, Brussels.

Lloyd, C.B., Hewitt, P.C., 2009. Poverty Gender and Youth, Educational Inequalities in the Midst of Poverty: Diversity Across Africa in Educational Outcomes. Working Paper. Population Council, New York.

Matafwali, B., Munsaka, E., 2011. An evaluation of community based early childhood education programmes in Zambia: a case of four selected districts. Journal of Early Childhood Development 5, 109–131.

SACMEQ, 2010. SACMEQ Policy Issues – What are the Levels and Trends in Reading and Mathematics Achievement? SACMEQ.

Sinyolo, D., 2007. Education International. Teacher Supply, Recruitment and Retention in Six Anglophone Sub-Saharan African Countries in The Gambia, Kenya, Lesotho, Tanzania, Uganda and Zambia. Education International, Brussels.

Spaull, N., 2011. SACMEQ Working Paper: Primary School Performance in Botswana, Namibia, Mozambique and South Africa. South and Eastern Africa Consortium for Monitoring Educational Quality. SACMEQ.

UNESCO, 2006. Regional Overview Sub-Saharan Africa, EFA Global Monitoring Report – Literacy for Life. UNESCO, Paris.

UNESCO, 2008a. Regional Overview Sub-Saharan Africa, EFA Global Monitoring Report – Education for All by 2015 Will We Make it? UNESCO, Paris.

UNESCO, 2008b. EFA Global Monitoring Report 2009: Overcoming Inequality: Why Governance Matters. UNESCO, Paris.

UNESCO, 2009. Regional Overview of Sub-Saharan Africa – EFA Global Monitoring Report, Overcoming Inequality: Why Governance Matters. UNESCO, Paris.

World Bank, 2008. Transitions in Secondary Education in Sub-Saharan Africa, Equity and Efficiency Issues. World Bank Working Paper. World Bank, Washington, DC.

World Bank, 2009. Abolishing School Fees in Africa. Lessons from Ethiopia, Ghana, Kenya, Malawi and Mozambique. World Bank/UNICEF, Washington, DC.

World Bank, 2011. Rwanda Country Status Report, Toward Quality Enhancement and Achievement of Nine Year Basic Education. World Bank, Washington, DC.

Yahie, A.M., 2000. Poverty Reduction in Sub-Saharan Africa: Is There a Role for the Private Sector?. Economic Research Papers Africa Development Bank, Cote d'Ivoire.

Education in the Middle East and North Africa

Abdeljalil Akkari, University of Geneva, Genève, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article analyzes the educational systems in the Middle East and North Africa and highlights the similarities and specificities of countries in the region. The first part describes the structures of the educational systems. The second part identifies the characteristics of the educational systems in four distinct groups of countries in the region with common challenges. The third part analyzes the development of access to different levels of education in the region and discusses the development of mass schooling and higher education. The fourth part deals with the impact of education on societies in the region, analyzing the high rate of youth unemployment and the persistence of illiteracy. The place of religion and cultural diversity in the educational systems of countries in the region are addressed.

Introduction

The region of the Middle East and North Africa (MENA (The MENA region does include Iran. The educational problems of this country are in some ways similar to those of the region. However, this article focuses on the educational situation in Arab countries.)) is characterized by its sociopolitical diversity. The region consists of 22 countries with diverse political regimes (republic, monarchy). The level of educational development varies between those countries with low school enrollment and those which manage to get all of their children into school. Despite disparities in income and literacy level, the region enjoys a fair degree of homogeneity in other respects and is well integrated historically and culturally. A common language (Arabic) and religion (Islam) connect the vast majority of the region's population (Salehi-Isfahani, 2010).

Arab countries concentrate a heavier expenditure on education, as a share of gross domestic product, than the world's average. They have made great strides in eradicating mass illiteracy; in boosting primary, secondary, and university enrollment; and in reducing gender gaps in education. Yet, two recent reports published by the World Bank and the Rand Corporation point out that Arab schools continue to propose programs that are irrelevant to the socioeconomic needs of the population and that do not effectively prepare students for the global economy of twenty-first century (World Bank, 2008; Gonzalez et al., 2008). The teaching approach is mainly based on rote learning and little links exist between school curricula and work settings. The region is very youthful, with 60% of the population aged under 30 years and one-third under 15 years. This demographic fact is both an asset and a challenge. Education, it must be kept in mind, is vital to shaping societies and governments in the near future. Thus, investing in educational reform, today, would encourage democratic citizenship and would help the current quest for an Arab democracy.

The Structures

Contemporary educational structures in MENA countries are linked to colonization. Indeed, prior to European colonization of the nineteenth and twentieth centuries, education in the region was mainly linked to religious Koranic schools, with the exception of the Mediterranean cities where large Jewish minorities and European inhabitants were living. In cities like Tunis, Cairo, Alexandria, Beirut, and Damascus, some modern, Western-style schools emerged before European colonization, offering classes including sciences and foreign languages.

The independence of the Arab countries in the middle of the twentieth century has seen the emergence of educational structures inherited from the former European colonial powers (United Kingdom, France). North African countries (Algeria, Morocco, Tunisia, and Mauritania) and the Levant (Lebanon, Syria) have copied the structures of the French educational system with the use of Arabic as a language of instruction, at different degrees of intensity. Postcolonial bilingualism (French–Arabic) in North Africa has given rise to much debate and controversy until now. The rest of the Arab countries have adopted the structures of the British system with Arabic as the main language of instruction.

Current educational structures of the MENA countries are characterized by the coexistence of public and private institutions. The private sector has, in many contexts, the monopoly of schooling the elites, and reaps the benefits of the associated social prestige.

Another feature of educational structures is the deep centralization of educational structures. Textbooks, teachers, and school legislation are set at the national level. Educational reforms require good governance. Unfortunately, this notion is lacking in the region at the level of the central government, regional administration, and school. Ministries of education assume a highly centralized role and are dominated by top-down authoritarian management systems. Furthermore, most ministries lack long-term vision, appropriate strategic planning, and competent human resources.

Because of wide disparities in income, MENA forms a heterogeneous group of countries from the standpoint of per capita income and educational development. Therefore, it is important that the discussion of their education systems, of which income is one of the most important determinants, be conducted with three groups of countries in mind: (1) High income, oil-rich countries of the GCC (The Gulf Cooperation Council: Oman, United Arab Emirates, Bahrain, Qatar, Kuwait,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92149-4

and Saudi Arabia), and Libya. For these countries, infrastructure and school coverage are similar to those of developed countries. Enrollment rates are high at all levels of schooling. In these countries, higher education is marked by the strong presence of foreign, for-profit universities. A lack of qualified local teachers and managers is observed. (2) Middle income, mainstream MENA countries (Morocco, Algeria, Tunisia, Egypt, Syria, Lebanon, Palestine, Jordan, and Iraq). The educational systems of these countries have a decent infrastructure and enrollment. Residual problems exist in some areas (rural and remote regions). Enrollment rates are high at primary and secondary levels and increasing at the tertiary level. (3) Low-income, periphery MENA countries (Sudan, Mauritania, Yemen, Somalia, and Comoros). These countries need to address the problem of educational infrastructure. Challenges exist at different levels of the educational system such as illiteracy rates, children not in school, and the gender gap.

Quantitative Expansion of Education

Arab countries have gradually improved their school enrollment in the last three decades. According to the United Nations Educational, Scientific and Cultural Organization (UNESCO) statistics, access to primary school has become the norm for most children in the region. Thus, on average, 87% of boys and 82% of girls at each relevant age are currently enrolled in primary education in Arab countries. This has led to significant progress toward universal access to school.

Another element to measure the development of education is to examine school life expectancy. According to the Institute for Statistics (2012), school life expectancy is 14.5 years in Tunisia, 13.7 years in Saudi Arabia, and only 5.1 years in Djibouti. According to UNESCO (2007), the average years of schooling in 2005 exceeded 12 years in half of the Arab countries, and exceeded 10 years in the majority of countries.

This improvement in school enrollment in the region is related to two main factors. The first is linked to the building of the postcolonial states. National elites in positions of power saw education as the top priority, believing that a well-instructed nation can access economic development and reach technological independence. The second factor is linked to oil revenues that have enabled many states in the region to invest in the construction of schools and teacher recruitment.

By national law, most countries in the MENA region offer an average of 9 years of compulsory education, while it is 12 years in the Organisation for Economic Co-Operation and Development (OECD) countries like the United States and United Kingdom (Figure 1).

Despite progress in schooling, 6 million primary school children remain out of school, the majority of whom are girls. Between 1999 and 2005, Yemen, one of the world's poorest countries, witnessed a large increase in enrollment. However, in 2008, it recorded over 1 million out-of-school children, the highest number in the region (UNESCO, 2011). This increase in children out of school has slowed over the past few years. However, that number could increase again as a consequence of the resent political revolts in the country.

The Problem of Educational Quality in the Region

While government expenditure on education in some MENA countries, such as Tunisia (6.9%), Jordan (4.9%), Saudi Arabia (5.6%), and Morocco (5.6%), is higher than in the OECD countries (5.4%) and North America (5.1%) (Al Masah, 2011), all MENA countries have weak educational outcomes:

> Qualitative criteria that deal with such issues as the quality of teachers, curricula, teaching methodologies, facilities, evaluation tools, guidance and counseling services, special services for individuals of special needs, and the nature of out-of-class activities. They also include the extent to which life-long educational services and facilities are available, especially within non-formal education. The progress achieved in the Arab countries according to the qualitative criteria has been modest.
>
> Masri, 2009: p. 134

If the problems of educational quality in the region are not new and are partly related to the legacy of weak colonial

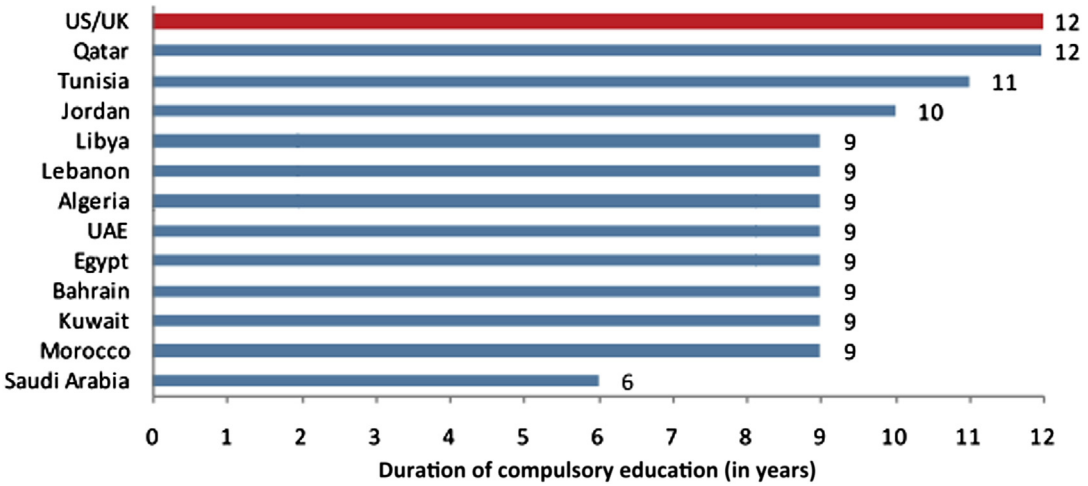

Figure 1 Duration of compulsory education.

schooling (Akkari, 2010) and poverty (Rebouha and Pochet, 2011), what is worrying is the persistence of this situation in comparative perspective (Heyneman, 1997). All of the countries in the region are facing a serious problem, that of the quality of education provided in schools. One of the most convincing indicators to realize the existence of this common challenge is to analyze literacy rate in the region. Most Asian and Latin American countries do better.

Illiteracy rates are particularly high in densely populated countries such as Egypt, Morocco, and Yemen. The illiteracy rate for women exceeds 50% in Mauritania, Morocco, and Yemen. While the literacy rate has been rising in the region, it has not risen fast enough to meet the population growth. The number of adult illiterates is still rising, reaching 60 million in 2008, 18 million of whom live in Egypt. Women have been disproportionately affected in most countries, with women representing over 70% of illiterates in Yemen, Libya, Jordan, and the occupied Palestinian territories. Bahrain, Jordan, Kuwait, Lebanon, Qatar, the United Arab Emirates, and the Occupied Palestinian Territories all had literacy rates above 90% (UNESCO, 2011; **Table 1**).

Another way to approach the quality of education in Arab countries is to situate their ranking in international comparative surveys. When looking at this indicator, even rich Arab gulf countries need to improve the quality of educational institutions at the primary, secondary, and tertiary levels. A common problem that arises across the region is that the educational institutions do not teach young people the skills necessary to succeed in the private sector (Davis and Hayashi, 2007).

The Progress in International Reading Literacy Study, an international comparative study of the reading literacy of young students, ranked all Arab countries, including the rich Gulf States, among the worst performers. In Morocco, for example, a majority of students had not acquired basic reading skills after 4 years of primary education (UNESCO, 2011).

Table 1 Adult (15+ years) literacy rate

Country	*Literacy rate (%)*
Algeria (2006)	72.6
Bahrain (2009)	91.4
Egypt (2006)	66.4
Iraq (2009)	78.7
Jordan (2007)	92.2
Kuwait (2008)	93.9
Lebanon (2007)	89.6
Libya (2009)	88.9
Mauritania (2009)	57.5
Morocco (2009)	56.1
Palestine (2009)	94.6
Oman (2008)	86.6
Qatar (2009)	94.7
Saudi Arabia (2009)	86.1
Sudan (2009)	70.2
Syria (2009)	84.2
Tunisia (2008)	77.6
United Arab Emirates (2005)	90
Yemen (2009)	62.4

Source: UIS. 2012. Data Centre: UNESCO Institute for Statistics.

A survey called Trends in International Mathematics and Science Study, which comes out every 4 years, revealed in its latest report, in 2007, that out of 48 countries tested, all 12 participating Arab countries fell below the average. More disturbingly, less than 1% of students aged 12–13 years in 10 Arab countries reached an advanced benchmark in science, compared with 32% in Singapore and 10% in the United States. Only one Arab country, Jordan, scored above the international average, with 5% of its 13-year-olds reaching the advanced category. Education in the Arab world lags behind that in other regions. The region had not witnessed the positive changes seen in Asia and Latin America, particularly in literacy rates (World Bank, 2008).

Internationally comparable reading test scores at grade four show that, in Kuwait, Qatar, and Morocco, over 90% of the students scored lower than the minimum benchmark, which indicates that the students have not acquired basic reading comprehension after at least 4 years of schooling. The results do not improve over time either. Math and science scores for grade eight show that the majority of students in Algeria, Egypt, Saudi Arabia, Syria, and Qatar are below the minimum threshold (Adams and Winthrop, 2012).

The low quality of education in the region is directly related to teacher quality. However, the number of trained teachers and their professional development are still neglected by policy makers:

> The current shortage of qualified teachers, especially at the primary and secondary levels, is due to several factors. These include low teacher wages and inefficient school management, which do not entice the best and the brightest to enter into the academic profession.
>
> Maroun et al., 2008: p. 14

UNESCO pointed out that the future shortage of quality teachers threatens the goal of education for all, and requires that the Arab States create 450 000 new teaching posts, mainly in Egypt, Iraq, Morocco, and Saudi Arabia (UN News Center, 2006). Furthermore, the causal link between the recruitment of female teachers and girls' enrollment and retention is direct and very significant. There is also evidence that the lack of a school close to the home raises concerns for the safety and security of girls (Alim, A., et al., 2007).

A pedagogical revolution is needed to improve the quality of learning because teaching in most Arab states continues to be highly didactic, teacher-directed, and not conducive to fostering analytical or critical thinking. As pointed out by Whitaker (2009), if this makes young Arabs well equipped for anything at all, it is how to survive in an authoritarian system: just memorize the teacher's words, regurgitate them as your own, avoid asking questions – and you will stay out of trouble. In the same way, the suppression of the critical faculties of the students turns some of them into gullible recipients for religious ideas that would collapse under serious scrutiny while making them ill equipped for roles as active citizens and contributors to their countries' development.

Despite the inability of Arab education to develop critical thinking among young people, the year 2011 saw the mobilization of youth to challenge many ossified political powers. The contribution of information and communication

technology including social networks has been crucial. Now, as large parts of the Arab world (Bahrain, Syria, Egypt, Yemen, Libya, and Tunisia) start the long process of democratization, a self-evident but often ignored fact is that democracy will thrive only in a culture that accepts diversity, respects different points of view, regards truths as relative rather than absolute, and tolerates – even encourages – dissent thinking. Schools hold particular responsibility on this issue.

Youth Employment, Minority Languages, and Religious Education

The issue of youth employment, minority languages, and religious education will guide the possibilities of improving educational systems in the region over the coming decades.

Education is often suggested as a solution to unemployment, but in MENA, educated youth are often less likely to be employed than the less educated. For example, in Egypt, those with the lowest education have the lowest unemployment (El-Hamidi and Wahba, 2006).

Current youth unemployment levels in the region are approximately 25%. First-time job seekers, mostly between 15 and 24 years of age, make up more than 50% of the unemployed – the highest regional average worldwide. Young people with secondary and postsecondary education face special challenges. The average duration of unemployment for youth with university or vocational education is relatively high. In Morocco, it is 3 years and in Egypt, it is two and a half. In the past, a secondary school degree made graduates eligible for a public sector job with lifetime security and social protection. Today, secondary degrees are at best guarantors of an informal wage job, demonstrating a dramatic devaluation of educational credentials and, as a result, tremendous dissatisfaction among the educated (Dhillon and Yousef, 2007).

In the twenty-first century, a typical youth in the Middle East spends an average of 8 years being educated, followed by a period of unemployment, or 'wait-hood,' which can last several years. During this period, however, pathways to independent life become unpredictable and inaccessible (Dhillon and Yousef, 2007).

Singerman (2007) argues that "the consequences of the youth bulge in the region can only be fully understood by an examination of the political economy of youth through the lens of the 'marriage imperative'." Economists speak of the phenomenon of 'wait unemployment,' of enduring long periods of unemployment in order to secure a high paying 'permanent' position with good benefits. In a similar vein, many young people in Egypt, and throughout the region, experience 'wait adulthood' or 'wait-hood' as they negotiate their prolonged adolescence and remain single for long periods of time while trying to save money for marriage. The author argues that there has been silence around the economic dynamics of marriage. Thus, many youth and parental attitudes and decisions, as well as policy interventions surrounding the school-to-work transition, make little sense without incorporating marriage into the equation. Policy makers must, therefore, consider the school-to-work-to-marriage transition when designing strategies to improve the lives of young people.

It is important to emphasize that the issue of lack of employment cannot be solved simply by schooling or education. At the same time, it is necessary that the economy of the region can be more dynamic in terms of job creation. Therefore, it is the overall development strategy that must be revised entirely if one hopes to find a solution to the problem of youth unemployment.

For decades, the main linguistic and ethnic minorities in the region had no right to use their language in schools. However, the two main linguistic minorities (Tamazight and Kurds) have seen their language rights improve during the last decade.

Until recently in Morocco, Berber languages, such as Tamazight, were not used formally in school and were only used in informal settings in areas where people considered themselves ethnically Berber:

> After independence, Berber was excluded from schools in the name of state-building and in search of a unified national identity. This forced a whole generation of children to enter school in a language they had never spoken before, contributing to a higher dropout rate among Berber children. Trouble for Berber-only speakers did not stop in the educational system. Many continued to face other difficulties communicating in hospitals and the court system, where Arabic and French dominated.
>
> Ennaji, 2009: p. 19

In the year 2000, King Mohammed VI established the National Charter of Education and Training, providing for reforms in the area of education. The Charter acknowledges the existence of Morocco's linguistic issues in Act 9, Article 110, where the stated goal is to strengthen the teaching and use of the Arabic language, empower foreign languages, and be open to the teaching of Tamazight (Daniel and Ball, 2009). In 2007, nearly 300 000 students – native Arabic speakers as well as Berber speakers – were enrolled in Berber courses, according to the Ministry of Education (Errihani, 2008).

In Iraq, article 34 of the Constitution of 2005 stipulates that education is a fundamental factor for the progress of society and is a right guaranteed by the State. Article 4 specifies that the Arabic language and the Kurdish language are the two official languages of Iraq. The right of Iraqis to educate their children in their mother tongue, such as Turkmen, Assyrian, and Armenian, shall be guaranteed in government educational institutions (UNESCO-IBE, 2011). Today, Kurdish children in Kurdistan Iraq are taught Kurdish as their means of instruction in all subjects (Taylor and Skutnabb-Kangas, 2009).

The question of religious education arises at several levels in the MENA region. First, because of the religious origin of traditional educational institutions before colonization, religion is present in schools. Some reports point to the importance of the amount of time devoted to religion compared to other subjects and compared to the situation in other developing countries. Primary schools in Saudi Arabia devote 31% of the classroom time to religion (Rivard and Amadio, 2003), compared with just 20% of time allotted to mathematics and science. Furthermore, research has shown that the time spent on academic tasks is one of the key factors in learning outcomes. In fact, if the countries of the region want to improve their performance in international comparative studies, they must reflect on the time allocated for each subject including religion.

Another more controversial issue in religious education is related to the content of this teaching. Groiss (2003) focused

on the biased attitude that Saudi Islamic education textbooks construct about topics such as Judaism and Christianity, Western civilization and imperialism more generally, and the way certain historical events such as the Crusades and the Palestinian struggle are taught. Yet, as rightly suggested by Labidi (2010), Groiss addresses with great simplicity the complex issues that lie beneath the question of Western representation. The relationship between the West and the Arab World are shaped by complex colonial, geopolitical, and historical factors.

> In order to improve Arab education, mutually deconstruct Arab/Muslim and Western stereotypes, diffuse the tension between the West and Arab populations, and put both sides on better grounds of negotiation, over simplifications, inaccuracies, distortions, and de-contextualizations of certain facts ought to be addressed. Teaching a selective and sanitized history that centers on oppression, silence, and denial leads to ignorance and naïve judgments.
>
> Labidi, 2010: p. 213

Conclusion

Education in the MENA region has undergone vast changes in recent decades. Access to school, which was limited in the aftermath of independence, has now become the rule for most Arab children. However, this quantitative leap has not been followed by the desired effect. Not only the quality of young Arabs' learning, which is not well rated by international standards, but also the expected benefits of schooling, have been slow to materialize. The unemployment rate for graduates is even higher than for those who did not attend school. Structural changes are urgent. Educational governance of the education systems of the countries of the region is inevitable. Consequently, reform of the teaching profession should be a top priority.

See also: Affordable Private Schools in Developing Countries; Anthropological Approaches to the Study of Education: The United States and Beyond; Bilingual Education: International Perspectives; Comparative Research in Education: Iea Studies; Cross-National Comparisons in Education: Findings from Pisa; Early Education and Care in the World Regions; Education in South Asia; Education in Sub-Saharan Africa; Gender and Education; International Organisations and Education.

Bibliography

Adams, A., Winthrop, R., 2012. The Role of Education in the Arab World Revolutions. Brookings. http://www.brookings.edu/.

Akkari, A., 2010. La escuela en el Magreb: Desarrollo inacabado. Icaria Editorial-Biblioteca Casa arabe, Barcelona, Madrid.

Al Masah, 2011. MENA: Education Sector Report. Al Masah, Dubai.

Alim, A., Ben Abdallah, K., Ramaroson, S., Sidikou, M., Van de Wiel, L., 2007. Accelerating Girl's Education in Yemen: Rethinking Policies in Teachers' Recruitment and School Distribution. UNICEF, Sanaa.

Daniel, M.C., Ball, A.F., 2009. The Moroccan educational context: evolving multilingualism. International Journal of Educational Development 30, 130–135.

Davis, N., Hayashi, C., 2007. The Gulf cooperation council (GCC) countries and the world: scenarios to 2025: implications for competitiveness. In: Drzeniek Hanouz, M., El Diwany, S., Yousef, T. (Eds.), The Arab World Competitiveness Report 2007. WEF, Geneva, pp. 129–141.

Dhillon, N., Yousef, T., 2007. Inclusion: Meeting the 100 Million Youth Challenge, Middle East Youth Initiative. The Wolfensohn Center for Development at Brookings & The Dubai School of Government, Dubai.

Dhillon, N., Yousef, T., 2009. Generation in Waiting: The Unfulfilled Promise of Young People in the Middle East. The Brooking Institution Press, Washington, DC.

El-Hamidi, F., Wahba, J., 2006. Why Does the MENA Region Have Such High Unemployment Rates? Working Paper 270. University of Pittsburgh, Department of Economics.

Ennaji, M., 2009. Multiculturalism, citizenship, and education in Morocco. Mediterranean Journal of Educational Studies 14 (1), 5–26.

Errihani, M., 2008. Language Policy in Morocco: Implications of Recognizing and Teaching Berber. Vdm Verlag Dr Müller, Saarbrücken.

Gonzalez, G., Karoly, L.A., Constant, L., Salem, H., Goldman, C.A., 2008. Facing Human Capital Challenges of the 21st Century: Education and Labor Market Initiatives in Lebanon, Oman, Qatar, and the United Arab Emirates. RAND, Santa Monica, CA.

Groiss, A. (Ed.), 2003. The West, Christians and Jews in Saudi Arabian Schoolbooks. American Jewish Committee and the Center for Monitoring the Impact of Peace, New York.

Heyneman, S.P., 1997. The quality of education in the Middle East and North Africa. International Journal of Educational Development 17 (4), 449–466.

Labidi, I., 2010. Arab education going medieval: sanitizing Western representation in Arab schools. Journal for Critical Education Policy Studies 8 (2), 196–219.

Maroun, N., Samman, H., Moujaes, C.N., Abouchakra, R., 2008. How to Succeed at Education Reform: The Case for Saudi Arabia and the Broader GCC Region. Booz & Company, Abu Dhabi.

Masri, M.W., 2009. Policy process and education Reform in the Arab world. Mediterranean Journal of Educational Studies 14 (1), 129–144.

Rebouha, F., Pochet, P., 2011. Pauvreté et accès à l'éducation dans les périphéries d'Oran. Autrepart 3 (59), 181–198.

Rivard, J.-F., Amadio, M., 2003. La place accordée à la religion dans les grilles horaires officielles. Prospects XXXIII (2), 102–109.

Salehi-Isfahani, D., 2010. Human Development in the Middle East and North Africa. United Nations Development Programme, New York.

Singerman, D., 2007. The Economic Imperatives of Marriage: Emerging Practices and Identities among Youth in the Middle East. The Wolfensohn Center for Development at Brookings & The Dubai School of Government, Dubai.

Taylor, S.K., Skutnabb-Kangas, T., 2009. The educational language rights of Kurdish children in Turkey, Denmark, and Kurdistan (Iraq). In: Ayers, W., Quinn, T., Stovall, D. (Eds.), Handbook of Social Justice in Education. Routledge, New York, pp. 171–190.

UIS, 2012. Data Centre. UNESCO Institute for Statistics.

UN News Centre, April 25, 2006. Global Teacher Shortages Threaten Goal of Quality Education for All.

UNESCO, 2007. Education for All by the Year 2015 (In Arabic). UNESCO, Paris.

UNESCO, 2011. UNESCO Education for All Global Monitoring Report 2011: The Hidden Crisis: Armed Conflict and Education. UNESCO, Paris.

UNESCO-IBE, 2011. World Data on Education. UNESCO-BIE, Genève.

Whitaker, B., 2009. What's Really Wrong with the Middle East. Saqi Books, London.

World Bank, 2008. The Road Not Traveled. Education Reform in the Middle East and North Africa. The World Bank, Washington, DC.

Education: Skills Training

Roy Canning, School of Education, University of Stirling, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R. LaLonde, volume 6, pp. 4243–4249, © 2001, Elsevier Ltd.

Abstract

Defining the concept of skills can be somewhat problematic. The term is basically an Anglo-Saxon one derived from the structure of jobs in the labor market. Other countries from around the world avoid the term skills, preferring instead the notion of occupations and/or competencies. In this article, I will explore these tensions in how skills are defined and measured in different economies. This will involve examining the distinctions between generic and basic skills and how skills are acquired and utilized. Finally, the links between qualifications and skills will be explored with reference to National Qualification Frameworks and skills training in initial and continuing vocational education systems.

Defining the concept of skill is notoriously difficult. The term is based upon the notion of a cluster of activities surrounding a job. The actual dictionary definition refers to ability, forte, gift, strength, and/or know-how. There is also a suggestion that it involves a judgment of competence. They can be referred to as hard and soft skills, generic skills, and basic skills. Indeed, in many ways the "concept of skill has become bigger, broader and much fuzzier around the edges" (Warhurst et al., 2004: p. 14). The article will provide a general introduction to the notion of skills, one that problematizes the concept and sets it within different cultural contexts. It will be argued that skills are difficult to define and measure. However, it will also be acknowledged that they have become much more important in the labor market and consequently within education and training systems around the world.

Defining Skills

The notion of 'skills' can really only be understood within a context, that context being the vocational education and training system of the country in question. There is, in other words, a strong cultural imperative that underpins the notion of skills. For instance in Anglo/Saxon countries, skills are often defined very narrowly:

> A skill is a specific ability of an individual, typically manual and/or coordinative features, which is geared to a task – or, more precisely, a task type – itself quite a narrow category, involving the application of a technique.
>
> Winch, 2011: p. 88

They are to be found within jobs in the workplace and are essentially individual rather than collective activities. They can also be measured by proxy through qualifications and classified into occupational groupings to determine how they will change over time.

However, in other countries the term skill is often avoided. In Germany the concept of *Beruf* is preferred, denoting an occupational identity and status. It is concerned with the individual acting in a certain socially defined role that has been arrived at through negotiations with social partners (Brockmann et al., 2011). This is a much broader concept than skills and forms the basis of the much admired Dual System of work-based education for intermediate occupations. In France the VET system is underpinned by the concept of *savoir*, meaning a body of knowledge. Although knowledge in action is important, savoir-faire, vocational formation is primarily concerned with acquiring and mastering knowledge (Méhaut, 2011).

One of the main determinants of how skills are conceptualized is the context within which they are situated. Anglo-Saxon countries will tend to locate skills within the individual and measure them through employer's allocated jobs in the workplace. Alternatively, in countries with a long history of social partnerships, skills will only form part of an overall framework based upon occupations and a collective sense of social division and citizenship.

Within the European Union the concept of *competence* has become an overarching term for denoting capability. Interestingly, competence is not defined narrowly within this context. Rather it includes the notions of autonomy and responsibility and incorporates both skills and knowledge requirements. This broader and more integrated concept of competence is used within the European Qualification Framework (EQF). However, the notion of competence is difficult to quantify and is rarely used when measuring skills.

Generic and Basic Skills Training

Within the policy-making discourse, generic skills are generally taken to mean a set of discrete clusters of skills, normally at a foundation level, that are transferable across different work contexts. They are usually distinguished from basic skills and, it is claimed, can be identified and assessed at varying levels of ability. Although these clusters of skills change over time, they have become increasingly identified with the work-based practices of communication, working with others, numeracy, problem solving, and information technology. The theoretical argument in support of the concept of generic skills rests mainly on the twin ideas that skill sets reside within the individual and are transferable

across boundaries. This notion of a 'transporting' phenomenon seems intuitive enough, particularly when applied to contexts that are identical in nature. The theoretical arguments underpinning generic skills have, however, come under criticism in recent years. Firstly, the notion of 'transfer' has been seen to be highly problematic. There is little convincing evidence to support the argument that skills can be transported across contexts even at the most rudimentary of levels (Cree et al., 1998). The idea that skills are not easily transferred across contexts may seem, at first glance, to be somewhat counter-intuitive. However, it is difficult to identify even basic replicative knowledge that is transportable in this manner. The suggestion that more complex skills are, somehow, embedded within an individual and automatically transferred across a range of contexts remains highly implausible. Indeed, the concept of 'transfer' is probably better understood as a 'metaphorical' discourse than an empirical phenomenon.

One of the most heated debates around the notion of skills is that of basic skills or foundation skills. In most countries this often refers to literacy and numeracy skills. However, other 'softer' skills are now being included under this heading. For instance, communication skills, teamwork, and information technology are recognized as essential skills. Most education systems acknowledge that these skills are required before progressing on to a broader academic or vocational curriculum. However, not everyone agrees on how to teach basic skills. Recent research has tended to contextualize our understanding of foundation skills within a broader context (Barton et al., 2000). Here literacies are viewed as part of social practices and learned through interactions between 'real texts and the associated practices with their use.'

Skill Acquisition and Utilization

Generally speaking, the policy rhetoric on skills tends to prioritize skill acquisition over skill utilization. In initial vocational education and training the emphasis is on acquiring skills in order to compete in the labor market. This can refer to entry level skills for young people and/or returners to the workforce following a period of unemployment. Governments from across the world have also tended to set challenging targets for the level of skills held by their workforce. Most claim to be pursuing a 'high-skill' policy agenda, although in practice Anglo-Saxon labor markets seem to exhibit more of an 'hour glass' phenomenon with a combination of high and low skills and a squeezed middle of intermediate skills. Indeed, even this squeezed middle remains problematic as the emerging 'sunrise' associate professional and technical occupations are growing in number against a backdrop of a decline in the 'sunset'-skilled trades. In comparison, less attention has been paid to how skills are utilized (Keep and Mayhew, 2001). There are of course ongoing debates on the implications of an 'over-educated' workforce given the number of graduates entering the labor market. However, far less attention has been paid to the 'overworked' population and the shifting patterns of employment and occupations within current skill-sets.

Measuring Skills

How are skills measured? Most national frameworks for measuring skills are informed by the International Standard Classification of Occupations, ISCO-883. For instance in the United Kingdom the Standard Occupational Classification structure (SOC) is based on two main concepts: the concept of the kind of work performed or 'job' and the concept of 'skill.' Of these two concepts, that of a 'job' is well recognized. However, skills can be more problematic:

> The concept of skill is widely used, but it not defined or regulated in a consistent fashion. For classification purposes, skill is defined as the ability to carry out the tasks and duties of a job in a competent, thorough and efficient manner. Two aspects of this definition of skill are distinguished in the SOC: *skill level* is defined in terms of complexity of the tasks and duties to be performed; *skill specialisation* is defined as the field of knowledge required for competent, thorough and efficient conduct of the tasks.
>
> Anderson, 2009

Within the ISCO, four skill levels are broadly defined. Interestingly, the measurement of skills using the standard classification allows us to identify and explore the changing patterns of occupations and employment. This is a particularly valuable aspect of the categorization process as it allows us to see how jobs change over time. What is striking here is how occupations and jobs emerge and disappear, historically speaking, in relatively short time periods. In some countries these shifts are quite dramatic, while in others labor markets are more stable. How a country competes internationally can often be ascertained through shifting occupational structures and the rapidly changing nature of jobs in the economy.

Qualifications and Skills Training

Qualifications tend to act as a proxy for skills. This is the case within the OECD and the European Union. A recent trend has been to encourage the establishment of National Qualifications Frameworks (NQFs) in order to facilitate the transfer of skills across nation-states. The EQF for instance sets out to establish equivalences of qualifications based upon different levels of academic and vocational achievement. These levels (eight in total) are, in turn, matched with three separately identified components of ability: knowledge, skills, and competence. This 'metalevel' framework then acts as a benchmark for comparison with individual countries' own NQFs. These frameworks have been developed elsewhere in the world, most notably in Australia, New Zealand, and South Africa. There has also been an attempt to standardize individual qualifications using the concepts of learning outcomes and competencies. The former is based upon what is notionally required of an experienced worker in terms of knowledge and skills as a result of undertaking a recognized qualification. The competency element relates to the type of skill required (e.g., hard and soft skills) and the knowledge necessary to carry out the task successfully and is an indicator of the degree of autonomy and responsibility exercised by the worker.

Skills training can be categorized as initial vocational education and training or continuing vocational education and training. Initial vocational training can either be work-based or work-related. Many European countries have vocational training as part of their upper-school curriculum, often combined with liberal studies. This normally involves an element of work placement activities and attendance at a school or college. However, a number of countries operate a work-based apprenticeship model, e.g., Germany, Austria, and Switzerland. There has been a renewed interest in modern apprenticeships for younger and older workers, typically to be found in the new service sectors. Interestingly, the gender balance is also shifting, with the majority of new apprentices being women. Nonetheless, apprenticeships still account for a small proportion of the number of occupations in many countries. In fact work-related skills are more often than not taught in colleges of further education or increasingly within Higher Education. Continuing provisional education and training normally takes place in organizations and workplaces. It is typically informal and unaccredited and concentrated amongst younger and newer employees. Most labor market studies indicate that access to continuing education and training at work is highly gendered and based upon occupational status and previous educational attainment.

See also: Education and Economic Growth; Education and Employment; Educational Productivity; Vocational Education.

Bibliography

Anderson, P., 2009. Intermediate occupations and the conceptual and empirical limitations of the hour glass economy thesis. Work, Employment and Society 23, 169–180.

Barton, D., Hamilton, M., Ivanic, R. (Eds.), 2000. Situated Literacies: Reading and Writing in Context. Routledge, London.

Boreham, N., 2004. A theory of collective competence: challenging the neo-liberal individualisation of performance at work. British Journal of Educational Studies 52, 5–17.

Brockmann, M., Clarke, L., Winch, C., Hanf, G., Méhaut, P., Westerhuis, A. (Eds.), 2011. Knowledge, Skills and Competence in the European Labour Market: What's in a Vocational Qualification? Routledge, Abingdon.

Canning, R., 2007. Reconceptualising core skills. Journal of Education and Work 20 (1), 17–26.

Cree, V., Macaulay, C., Loney, H., 1998. Transfer of Learning: A Study. Scottish Office Central Research Unit, Edinburgh.

Keep, E., Mayhew, K., 2001. Globalisation, Models and Competitive Advantage and Skills. University of Oxford: SKOPE.

Méhaut, P., 2011. Savoir: the organizing principle of French VET. In: Brockmann, M., Clarke, L., Winch, C., Hanf, G., Méhaut, P., Westerhuis, A. (Eds.), Knowledge, Skills and Competence in the European Labour Market: What's in a Vocational Qualification? Routledge, Abingdon, pp. 36–49.

Warhurst, C., Grugulis, I., Keep, E., 2004. The Skills that Matter. Palgrave, London.

Winch, C., 2011. Skill: a concept manufactured in England. In: Brockmann, M., Clarke, L., Winch, C., Hanf, G., Méhaut, P., Westerhuis, A. (Eds.), Knowledge, Skills and Competence in the European Labour Market: What's in a Vocational Qualification? Routledge, Abingdon, pp. 85–101.

Educational Effectiveness Theory and Research: Recent Advances

Leonidas Kyriakides, University of Cyprus, Nicosia, Cyprus

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides a critical analysis of the main approaches to modeling educational effectiveness (economic, psychological, and sociological). During the 1990s an integrated approach to modeling effectiveness has been developed. Studies testing the validity of multilevel models emerged from this approach reveal the importance of considering the dynamic nature of education. Thus, a dynamic model of educational effectiveness has been developed. The main characteristics of this model and findings of studies providing support to its validity are presented. Finally, the importance of investigating the extent to which the dynamic model can be used for school improvement purposes is stressed.

Introduction

Educational effectiveness research (EER) addresses the question what works in education and why. EER has expanded rapidly during the last three decades. Methodological advances have enabled more efficient estimates of teacher and school differences in student achievement to be obtained. In regard to the theoretical component of the field, progress was made by a more precise definition of the concepts used and the relations between the concepts (Scheerens and Bosker, 1997). However, there is a shortage of rational models from which researchers can build theory. The problem is aggravated by infrequent use of whatever models exist (Kyriakides, 2005). As a consequence, most effectiveness studies are concerned with the establishment of statistical relationships between variables rather than with the generation and testing of theories which could explain those relationships.

In this article, a critical analysis of current approaches to modeling educational effectiveness is provided and the importance of considering the dynamic nature of educational effectiveness is stressed. It is argued that the dynamic model of educational effectiveness (Creemers and Kyriakides, 2008), which has recently been developed in order to address the weaknesses of the integrated models, could contribute in establishing a theory-driven and evidence-based approach to school improvement. Thus, the essential characteristics of the dynamic model are presented and the main findings of studies testing the validity of the model are outlined. Finally, suggestions for research in order to develop the theoretical framework of EER are drawn.

A Critical Analysis of the Models of EER

In the literature of educational effectiveness modeling, three basic approaches have been used. First, the economic approach is focused on estimating the relationship between the "supply of selected purchased schooling inputs and educational outcomes controlling for the influence of various background features" (Monk, 1992: p. 308). Resource input variables such as student/teacher ratio, teacher salary and overall measures of per student expenditure were of primary interest in the earlier studies. The emerging 'education production' models (e.g., Brown and Saks, 1986; Coates, 2003; Elberts and Stone, 1988) are based on the assumption that increased inputs will lead to increments in outcomes and their main characteristics are concerned with (1) the selection of resource inputs as the major type of selection of antecedent condition, (2) the measurement of direct effects, and (3) the use of data at only one level of aggregation (i.e., either at micro (e.g., student) level or aggregated (e.g., school) level). However, the research done using these models revealed that the relation between input and outcomes is more complex than was assumed. For example, studies from Hanushek and Hedges (e.g., Hanushek, 1986, 1989; Hedges et al., 1994) show that reducing student/teacher ratio and/or increasing the amount of funding education per student does not necessarily result in higher student outcomes.

The second approach to effectiveness modeling is similar to the economic approach but is focused on a different choice of antecedent conditions since it is mainly focused on variables at student level which are assumed to predict student outcomes. Some attention is also paid on processes from two different perspectives concerning learning and school as organizations. Within this approach, educational psychologists focused on student background factors such as 'learning aptitudes,' 'personality,' and 'motivation' and on variables measuring the learning processes which take place in classrooms. On the other hand, the sociological perspective is focused on factors that define the educational background of students such as socio-economic status, gender, social-capital, and peer group. This perspective does not only examine student outcomes, but also the extent to which schools manage to reduce the variance in student outcomes compared to prior achievement. In this respect, studies on the effects of contextual factors (Opdenakker and Van Damme, 2006) and on the extent to which teachers and schools are equally effective with different groups of students (i.e., differential school effectiveness) have been conducted (e.g., Campbell et al., 2004; Strand, 2010). Moreover, the sociological perspective raises attention for process variables emerged from organizational theories which were treated as school-level factors associated with student achievement (Scheerens, 2013).

Finally, the models of the third approach emerged by researchers attempt to integrate the findings of school effectiveness research, teacher effectiveness research, and the early

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92102-0

input–output studies. Thus, the models of this approach (e.g., Creemers, 1994; Scheerens, 1992; Stringfield and Slavin, 1992) have a multilevel structure, where schools are nested in contexts, classrooms are nested in schools, and students are nested in classrooms or teachers. Although these models make use of both organizational theories and theories of learning and refer to multiple factors at different levels, each of them is either focused on the classroom or the school level. Depending on this, more emphasis is given either to theories of learning (e.g., Creemers, 1994) or to organizational theories (e.g., Scheerens, 1992).

The comprehensive model of educational effectiveness (Creemers, 1994) is considered as one of the most influential models in the field (Teddlie and Reynolds, 2000). Creemers' model is based on four assumptions. First, time-on-task and opportunity used at the student level are directly related to student achievement. Second, quality of teaching, the curriculum, and the grouping procedures influence the time on task and opportunity to learn. For example, some teachers spend more time actually teaching than others who spend more time on classroom management and keeping order. Teachers are, therefore, the central component in instruction at the classroom level. Third, teaching quality, time, and opportunity at the classroom level are influenced by factors at the school level that may or may not promote these classroom factors. Thus, quality, time, and opportunity are not the key concepts at the classroom level only, but also at the school level.

Six studies examined the validity of Creemers' model and provided some empirical support to the model (Kyriakides, 2008). These studies revealed that the influences on student achievement are multilevel. This finding is in line with the findings of most effectiveness studies (Teddlie and Reynolds, 2000) and provides support to the argument that the theoretical models of EER should be multilevel in nature. However, the analyses of the results of these studies also revealed that next to the multilevel nature of effectiveness the relationship between factors at different levels might be more complex than assumed in the integrated models. This is especially true for interaction effects among factors operating at classroom and student level which reveal the importance of investigating differential effectiveness (Campbell et al., 2004). A synthesis of these studies has also revealed suggestions for further development of the model, especially by taking into account the dynamic nature of educational effectiveness (Kyriakides, 2008). Teaching and learning are dynamic processes that are constantly adapting to changing needs and opportunities. Effective schooling should, therefore, be treated as a dynamic, ongoing process. This idea is also consistent with the contingency theory (Donaldson, 2001; Mintzberg, 1979). In this context, Creemers and Kyriakides (2008) have developed a dynamic model of educational effectiveness which attempts to define the dynamic relations between the multiple factors found to be associated with student achievement.

The Dynamic Model of Educational Effectiveness: An Overview

The dynamic model is based on the following three main assumptions. First, the fact that most of the effectiveness studies are exclusively focused on language or mathematics rather than on the whole school curriculum aims (cognitive, metacognitive, and affective) reveals that the models of EER should take into account the new goals of education and related to this their implications for teaching and learning. This means that the outcome measures should be defined in a more broad way rather than restricting to the achievement of basic skills. It also implies that new theories of teaching and learning are used in order to specify variables associated with the quality of teaching. Second, an important constraint of the existing approaches of modeling educational effectiveness is the fact that the whole process does not contribute significantly to the improvement of school effectiveness. Thus, the dynamic model is established in a way that helps policy makers and practitioners to improve educational practice by taking rational decisions concerning the optimal fit of the factors within the model and the present situation in the schools or educational systems. Finally, the dynamic model should not only be parsimonious but should also be able to describe the complex nature of educational effectiveness. This implies that the model could be based on specific theory, but at the same time some of the factors included in the major constructs of the model are expected to be interrelated within and/or between levels.

The main characteristics of the dynamic model are as follows. First, the dynamic model is multilevel in nature and refers to factors operating at the four levels shown in **Figure 1**. Teaching and learning situation is emphasized and the roles of the two main actors (i.e., teacher and student) are analyzed. Above these two levels, the dynamic model refers to school-level factors which are expected to influence the teaching–learning situation by developing and evaluating the school policy on teaching and the policy on creating the School Learning Environment (SLE). The system level refers to the influence of the educational system through a more formal way, especially through developing and evaluating the educational policy at the national/regional level. The teaching and learning situation is also influenced by the wider educational context in which students, teachers, and schools operate. Factors such as the values of the society for learning and the importance attached to education play an important role both in shaping teacher and student expectations as well as in the development of the perceptions of various stakeholders about effective teaching practice.

Second, the dynamic model supports that factors at the school and system level have both direct and indirect effects on student achievement since they are able to influence not only student achievement but also the teaching and learning situations. Third, the model assumes that the impact of the school- and system-level factors has to be defined and measured in a different way than the impact of classroom-level factors. Policy on teaching and actions taken to improve teaching practice must be measured over time and in relation to the weaknesses that occur in a school. The assumption is that schools and educational systems, which are able to identify their weaknesses and develop a policy on aspects associated with teaching and their SLE, are also able to improve the functioning of classroom-level factors and their effectiveness status. Only changes in those factors for which schools face significant problems are expected to be associated with the improvement of school effectiveness. This implies that the

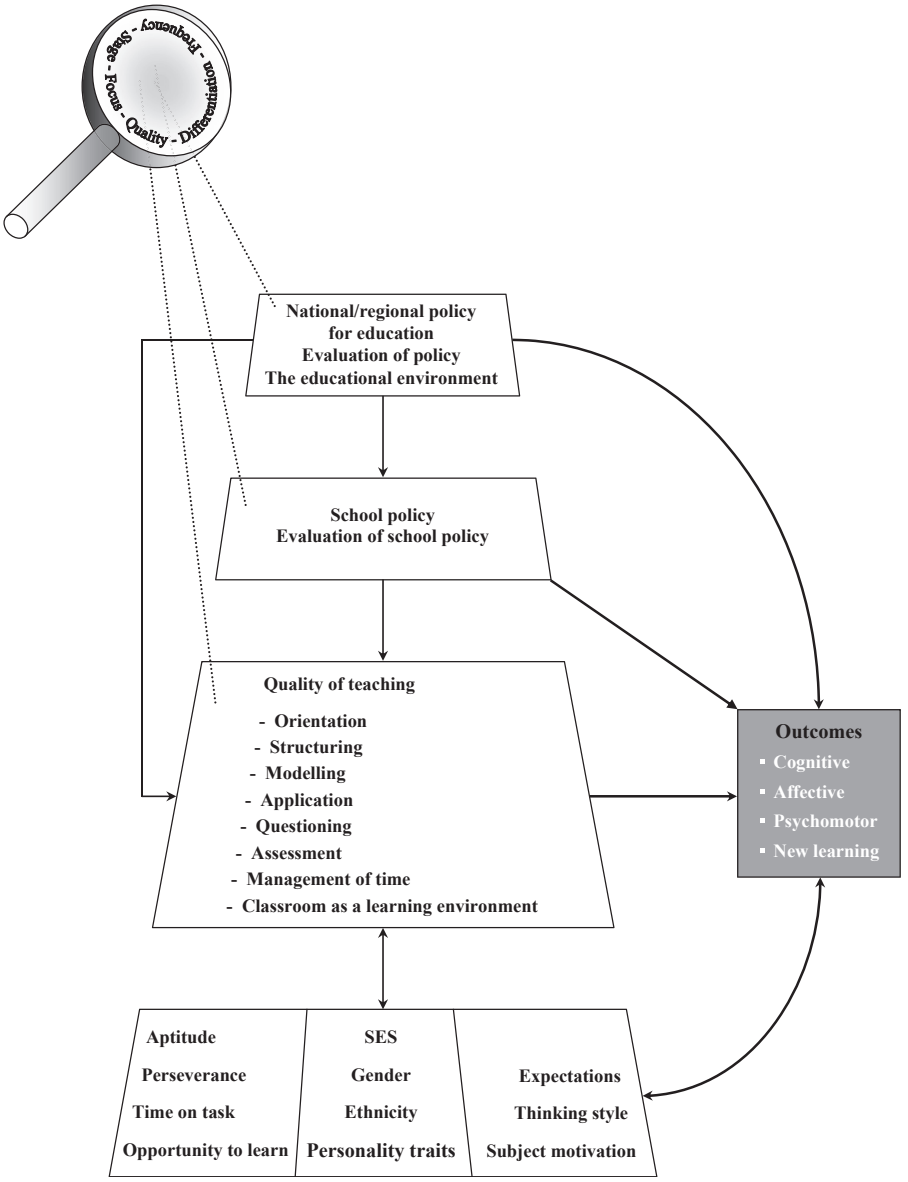

Figure 1 The dynamic model of educational effectiveness.

impact of school- and system-level factors depends on the current situation of the objects under investigation.

Fourth, the model assumes that there is a need to carefully examine the relationships between the various effectiveness factors which operate at the same level. Walberg's (1984) model, which is one of the most significant educational productivity models, attempts to illustrate such relationships. Aptitude, instruction, and the psychological environment are seen as major direct causes of learning. They also influence one another and are in turn influenced by feedback on the amount of learning that takes place. The Walberg's model was tested as a structural equation model on science achievement, indicating more complex, indirect relationships (Reynolds and Walberg, 1990). This implies that there is a need to refer to the relationships between the effectiveness factors which operate at the same level. Such approach to modeling educational effectiveness may reveal grouping of factors that make teachers and schools effective. Therefore, strategies for improving effectiveness which are comprehensive in nature may emerge.

Finally, the dynamic model is based on the assumption that each factor can be defined and measured by using five dimensions: *frequency*, *focus*, *stage*, *quality*, and *differentiation*. *Frequency* is a quantitative way to measure the functioning of each effectiveness factor whereas the other four dimensions examine qualitative characteristics of the functioning of each factor at the system/school/classroom level. A brief description of the four dimensions concerned with the qualitative characteristics of effectiveness factors is given below.

Two aspects of the *focus* dimension are taken into account. The first one refers to the specificity of the activities associated with the functioning of the factor whereas the second one to the number of purposes for which an activity takes place. The

measurement of the focus of an activity, either in terms of its specificity or in terms of the number of purposes that it is expected to achieve, may be related in a nonlinear way with student achievement. For example, guidelines on parental involvement which are very general may not be helpful at all in establishing good relations between parents and teachers, which, when good, can result in supporting student learning. On the other hand, a school policy which is very specific in defining activities may restrict teachers and parents from being productively involved and creating their own ways for implementing the school policy. The *stage* at which tasks associated with a factor take place is also examined. Factors need to take place over a long period of time to ensure that they have a continuous direct or indirect effect on student learning. The *quality* refers to the properties of the specific factor itself, as these are discussed in the literature. Finally, *differentiation* refers to the extent to which activities associated with a factor are implemented in the same way for all the subjects involved with it (e.g., all the students, teachers, schools). It is expected that adaptation to specific needs of each subject or group of subjects will increase the successful implementation of a factor and will ultimately maximize its effect on student learning outcomes. The use of different measurement dimensions reveals that looking at just the frequency of an effectiveness factor does not help us identify those aspects of the functioning of a factor which are associated to student achievement. Considering effectiveness factors as multidimensional constructs not only provides a better picture of what makes teachers and schools effective but may also help us develop specific strategies for improving educational practice.

Based on the main findings of teacher effectiveness research (e.g., Brophy and Good, 1986; Muijs and Reynolds, 2001; Rosenshine and Stevens, 1986), the dynamic model refers to factors which describe teachers' instructional role and are associated with student outcomes. These factors refer to observable instructional behavior of teachers in the classroom rather than on factors that may explain such behavior (e.g., teacher beliefs and knowledge and interpersonal competences). The eight factors included in the model are as follows: orientation, structuring, questioning, teaching-modeling, application, management of time, teacher role in making classroom a learning environment, and *classroom assessment*. These eight factors do not refer only to one approach of teaching such as structured or direct teaching (Joyce et al., 2000) or to approaches associated with constructivism (Schoenfeld, 1998). An integrated approach in defining quality of teaching is adopted.

In regard to the school factors, the dynamic model gives emphasis to the following two main aspects of the school policy which affect learning at both the level of students and teachers: (1) *school policy for teaching* and (2) *school policy for creating a learning environment at school*. Guidelines are seen as one of the main indications of school policy and this is reflected in the way each school-level factor is defined. However, in using the term guidelines we refer to a range of documents, such as staff meeting minutes, announcements, and action plans, which make the policy of the school more concrete to the teachers and other stakeholders. These two factors do not imply that each school should simply develop formal documents to install its policy. The factors concerned with the school policy mainly refer to the actions taken by the school to help teachers and other stakeholders have a clear understanding of what is expected from them to do. Support offered to teachers and other stakeholders to implement the school policy is also an aspect of these two school factors.

Based on the assumption that the essence of a successful organization in the modern world is the search for improvement, the dynamic model is also concerned with the processes and the activities which take place in the school in order to improve the teaching practice and its learning environment. For this reason, the processes which are used to evaluate the school policy for teaching and the SLE are investigated. Thus, the following four overarching factors at the school level are included in the model:

1. school policy for teaching and actions taken for improving teaching practice,
2. evaluation of school policy for teaching and of actions taken to improve teaching,
3. policy for creating an SLE and actions taken for improving the SLE, and
4. evaluation of the SLE.

Testing the Validity of the Dynamic Model

Some supportive material for the validity of the dynamic model has been provided. Specifically, a longitudinal study measuring teacher and school effectiveness in three different subjects (i.e., mathematics, language, and religious education) was conducted in order to test the main assumptions of the model. Using Structural Equation Modeling (SEM) techniques, it was possible to demonstrate that classroom and school factors can be defined by reference to the five dimensions of the dynamic model (Kyriakides and Creemers, 2008). The added value of using these five dimensions of the classroom and school level factors to explain variation on student achievement in both cognitive and affective outcomes of schooling was also demonstrated. Finally, it was possible to generate evidence supporting the assumption that the impact of school factors depends on the current situation of the school and on the type of problems/difficulties that the school is facing. Specifically, school factors were found to have situational effects. The development and evaluation of a school policy for teaching were found to have stronger effects in schools where the quality of teaching at classroom level was low (Creemers and Kyriakides, 2009).

Second, a study investigating the impact of teacher factors on achievement of students at the end of preprimary school was conducted (Kyriakides and Creemers, 2009). By comparing the results of this study with the findings of the first study testing the validity of the model, it was shown that almost all teacher factors were associated with achievement in language and mathematics at both phases of schooling (see Kyriakides and Creemers, 2009). Some factors were also found to be more important for one age of schooling indicating the possibility of having differential effects. For the purpose of testing the generic nature of the model, this difference does not question the importance of teacher factors within the model. These differences in effect sizes might be attributed to

differences in the developmental stages of the two groups of students and related to that to the functioning and the curriculum of each phase of schooling.

Third, the validity of the dynamic model at the school level was supported by the results of a quantitative synthesis of 67 studies exploring the impact of school factors on student achievement (Kyriakides et al., 2010). This meta-analysis revealed that effective schools are able to develop policies and take actions in order to improve their teaching practice and their learning environment. Factors excluded from the dynamic model were found to be weakly associated with student achievement.

Finally, a follow-up study testing the validity of the dynamic model was conducted (Creemers and Kyriakides, 2010). The methods used were identical to those followed by the original study testing the validity of the model. This study provided support to the generalizability of the original study. Very similar results on the impact of teacher and school factors upon student achievement emerged from both the original and the follow-up study. Since the follow-up study took place in the same schools where the original study took place, changes in the effectiveness status of schools and in the functioning of effectiveness factors were also identified. Discriminant function analysis reveals that changes in the functioning of school factors can help us classify the schools into those which improved their effectiveness status or remained equally effective or even reduced their effectiveness status (see Creemers and Kyriakides, 2010). Thus, this study was able to test one of the essential differences of the dynamic model which has to do with its attempt to relate changes in the effectiveness status of schools to the changes in the functioning of school factors.

Conclusions and Suggestions for Further Research

In this article, it is demonstrated that useful contributions were already being made in the area of theory in 1990s but there is a need to take them further. Studies testing the validity of the comprehensive model revealed that the variation in 'what worked,' if it could be explained and theoretically modeled, would force the field toward the development of more complex and multifaceted accounts than the 'one size fits all' mentality that had hitherto existed in the field during the 1990s. In this context, the dynamic model of educational effectiveness has been developed. This model attempts to illustrate the complex and dynamic nature of educational effectiveness and takes into account the findings of research on differential effectiveness (Scheerens, 2013). In this article, the dynamic model is outlined and studies supporting the validity of the model are briefly presented. Since the dynamic model was designed in order to establish stronger links between EER and improvement of practice, experimental studies and/or case studies are currently being conducted to identify the extent to which schools can make use of the dynamic model for improvement purposes (see Creemers and Kyriakides, 2012). These studies may help us identify when and under what conditions schools can make use of the dynamic model and establish an evidence-based and theory-driven approach to school improvement.

See also: Educational Effectiveness, The Field of; Environments for Education; School Effectiveness Research; School Improvement.

Bibliography

Brophy, J., Good, T.L., 1986. Teacher behavior and student achievement. In: Wittrock, M.C. (Ed.), Handbook of Research on Teaching, third ed. MacMillan, New York, pp. 328–375.
Brown, B.W., Saks, D.H., 1986. Measuring the effects of instructional time on student learning: evidence from the beginning teacher evaluation study. American Journal of Education 94, 480–500.
Campbell, R.J., Kyriakides, L., Muijs, R.D., Robinson, W., 2004. Assessing Teacher Effectiveness: A Differentiated Model. Routledge Falmer, London.
Coates, D., 2003. Education production functions using instructional time as an input. Education Economics 11 (3), 273–292.
Creemers, B.P.M., 1994. The Effective Classroom. Cassell, London.
Creemers, B.P.M., Kyriakides, L., 2008. The Dynamics of Educational Effectiveness: A Contribution to Policy, Practice and Theory in Contemporary Schools. Routledge, London.
Creemers, B.P.M., Kyriakides, L., 2009. Situational effects of the school factors included in the dynamic model of educational effectiveness. South African Journal of Education 29 (3), 293–315.
Creemers, B.P.M., Kyriakides, L., 2010. Explaining stability and changes in school effectiveness by looking at changes in the functioning of school factors. School Effectiveness and School Improvement 21 (4), 409–427.
Creemers, B.P.M., Kyriakides, L., 2012. Improving Quality in Education: Dynamic Approaches to School Improvement. Routledge, London and New York.
Donaldson, L., 2001. The Contingency Theory of Organizations: Foundations for Organisational Science. Sage, Thousands Oaks, CA.
Elberts, R.W., Stone, J.A., 1988. Student achievement in public schools: do principles make a difference? Economics Education Review 7, 291–299.
Hanushek, E.A., 1986. The economics of schooling: production and efficiency in public schools. Journal of Economic Literature 24, 1141–1177.
Hanushek, E.A., 1989. The impact of differential expenditures on student performance. Educational Research 66 (3), 397–409.
Hedges, L.V., Laine, R.D., Greenwald, R., 1994. Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes (An exchange: Part 1). Educational Researcher 23 (3), 5–14.
Joyce, B., Weil, M., Calhoun, E., 2000. Models of Teaching. Allyn & Bacon, Boston.
Kyriakides, L., 2005. Extending the comprehensive model of educational effectiveness by an empirical investigation. School Effectiveness and School Improvement 16 (2), 103–152.
Kyriakides, L., 2008. Testing the validity of the comprehensive model of educational effectiveness: a step towards the development of a dynamic model of effectiveness. School Effectiveness and School Improvement 19 (4), 429–446.
Kyriakides, L., Creemers, B.P.M., 2008. Using a multidimensional approach to measure the impact of classroom level factors upon student achievement: a study testing the validity of the dynamic model. School Effectiveness and School Improvement 19 (2), 183–306.
Kyriakides, L., Creemers, B.P.M., 2009. The effects of teacher factors on different outcomes: two studies testing the validity of the dynamic model. Effective Education 1 (1), 61–85.
Kyriakides, L., Creemers, B.P.M., Antoniou, P., Demetriou, D., 2010. A synthesis of studies for school factors: Implications for theory and research. British Educational Research Journal 36 (5), 807–830.
Mintzberg, H., 1979. The Structuring of Organizations. Prentice Hall, Englewood Cliffs, NJ.
Monk, D.H., 1992. Education productivity research: an update and assessment of its role in education finance reform. Educational Evaluation and Policy Analysis 14 (4), 307–332.
Muijs, D., Reynolds, D., 2001. Effective Teaching: Evidence and Practice. Sage, London.
Opdenakker, M.C., Van Damme, J., 2006. Differences between secondary schools: a study about school context, group composition, school practice, and school effects with special attention to public and Catholic schools and types of schools. School Effectiveness and School Improvement 17 (1), 87–117.
Reynolds, A.J., Walberg, H.J., 1990. A Structural Model of Educational Productivity. Northern Illinois University, IL.

Rosenshine, B., Stevens, R., 1986. Teaching functions. In: Wittrock, M.C. (Ed.), Handbook of Research on Teaching, third ed. Macmillan, New York, pp. 376–391.
Scheerens, J., 1992. Effective Schooling: Research, Theory and Practice. Cassell, London.
Scheerens, J., 2013. The use of theory in school effectiveness research revisited. School Effectiveness and School Improvement 24 (1), 1–38.
Scheerens, J., Bosker, R.J., 1997. The Foundations of Educational Effectiveness. Pergamon, Oxford.
Schoenfeld, A.H., 1998. Toward a theory of teaching in context. Issues in Education 4 (1), 1–94.
Strand, S., 2010. Do some schools narrow the gap? Differential school effectiveness by ethnicity, gender, poverty, and prior achievement. School Effectiveness and School Improvement.
Stringfield, S.C., Slavin, R.E., 1992. A hierarchical longitudinal model for elementary school effects. In: Creemers, B.P.M., Reezigt, G.J. (Eds.), Evaluation of Educational Effectiveness. ICO, Groningen, pp. 35–69.
Teddlie, C., Reynolds, D., 2000. The International Handbook of School Effectiveness Research. Falmer Press, London.
Walberg, H.J., 1984. Improving the productivity of America's schools. Educational Leadership 41 (8), 19–27.

Educational Effectiveness, The Field of

Bert PM Creemers, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands
Leonidas Kyriakides, University of Cyprus, Nicosia, Cyprus

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides a historical review of Educational Effectiveness Research (EER). Four sequential phases in the field which address different types of research questions and promote the theoretical development of educational effectiveness are identified. It is shown that research on effectiveness moved from investigating the extent to which teachers and schools influence learning outcomes to establishing theoretical models which refer to the complexity and the dynamic nature of education. Methodological developments that promoted the development and testing of the theoretical framework of EER are also presented. Thus, the knowledge base of EER is currently used for establishing an evidence-base and theory-driven approach to promote quality in education.

Introduction

Educational Effectiveness Research (EER) can be seen as an overarching theme that links together a conglomerate of research in different areas, including research on teacher behavior and its impacts, curriculum, student grouping procedures, school organization, and educational policy. The main research question underlying EER is the identification and investigation of which factors in the teaching, curriculum, and learning environments (operating at different levels such as the classroom, the school, and above-school) can directly or indirectly explain measured differences (variations) in the outcomes of students. Further, such research frequently takes into account the influence of other important background characteristics, such as student ability, socioeconomic status (SES), and prior attainment. Thus, EER attempts to establish and test theories which explain why and how some schools and teachers are more effective than others in promoting better outcomes for students. However, it is also important to note that the three terms – school effectiveness, teacher effectiveness, and educational effectiveness – are used inconsistently in the literature and that these are themselves interrelated. In this article, school effectiveness is taken to mean the impact that school-wide factors, such as a school policy for teaching, school climate, and the 'mission' of a school, have on students' cognitive and affective performance. On the other hand, teacher effectiveness is taken to mean the impact that classroom factors have on student performance, and includes teacher behavior, teacher expectations, classroom organization, and use of classroom resources. Teddlie (1994) argued that most teacher effectiveness studies have been concerned with only the processes that occur within classrooms to the exclusion of school-wide factors, whereas most school effectiveness studies have involved phenomena that occur throughout the school with little emphasis on particular teaching behaviors within individual classrooms. This weakness has begun to be addressed in studies conducted during the last three decades. The attempts to deal with both teacher and school influences can be seen as a significant development in EER since joint studies on school and teacher effectiveness reveal that neither level can be adequately studied without considering the other. In this context, we are using the term educational effectiveness rather than teacher and/or school effectiveness to emphasize the importance of conducting joint school and teacher effectiveness research which can help us identify interactions between the school, classroom, and student levels and their contributions in explaining variation in students' outcomes, both academic and noncognitive. Finally, it is important to note that EER also refers to the functioning of the educational system as a whole, and this research can therefore also be used to support the development and testing of different models of effectiveness. In turn, these models of effectiveness ultimately attempt to explain why educational systems and their subcomponents perform differently, toward the aim of providing relevant evidence for policy makers.

In this context, this article provides a historical review of educational effectiveness. This article helps readers identify the theoretical and methodological progress that the field managed to make during the last decades. It is shown that studies conducted in various countries managed to contribute to the development and testing of complex and dynamic theories of effectiveness. In the last part of this article, we discuss how the knowledge base of EER can be used to establish an evidence-based and theory-driven approach to school improvement and refer to experimental studies searching for the impact of such an approach in promoting student learning outcomes.

Historical Review of Educational Effectiveness

The origins of EER stem from reactions to the work on equality of opportunity undertaken by Coleman et al. (1966) and Jencks et al. (1972). The first two effectiveness studies undertaken, independently, by Edmonds (1979) and Rutter et al. (1979) during the 1970s were concerned with examining evidence and making an argument about the potential power of schooling to make a difference in students' life chances. This was an optimistic point of view because many studies published in that period showed that teachers, schools, and maybe even education in general did not make much of a difference at all. The early existence of two independent research projects in different countries asking similar questions and drawing, to

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92101-9

a certain extent, on similar quantitative methodologies demonstrated the potential for establishing a scientific domain dealing with effectiveness in education (Creemers et al., 2010). Thus, the publications by Brookover et al. (1979) and Rutter et al. (1979) were followed by numerous studies in different countries on educational effectiveness and the development of international interest and collaboration through the creation of the International Congress for School Effectiveness and Improvement in 1990 (Teddlie and Reynolds, 2000). Looking at the history of EER, we see four sequential phases in the field which address different types of research questions and promote the theoretical development of EER.

First Phase: Establishing the Field by Showing That Teachers and Schools Matter

The studies of the first phase of EER were mainly concerned with revealing that teachers and schools differ among themselves in their impact on student performance and thereby that student progress in part depends on who is their teacher and which school they attend. The extent to which schools differ was the next question raised by researchers in the field with a more precise version concerning how much schools differ in terms of student outcomes when they are more or less equal in terms of the innate abilities and socioeconomic background of their students (using statistical controls for variations in student intake characteristics). EER aimed to make fair comparisons between teachers and schools in order to assess the impact of schooling on student achievement that can be uniquely attributed to, for example, teacher A or school X, rather teacher B or school Y. Such research was enabled through the use of multilevel models that allowed for clustering of the data at the teacher (class) and at the school level which enabled more precise estimates of variation between schools and the identification of individual school effects (through residual estimates and their associated confidence limits). By the end of this phase, a clear message about the important role of teachers and schools had emerged from a large number of studies conducted in various countries and these provided a strong argument against critics who had argued that teachers and schools did not matter for student outcomes. However, the issue of educational effectiveness did not end by assessing the differences between schools and teachers in terms of their effectiveness. Rather, this was merely a prelude to exploring what matters in schools. In this context, the so-called 'five-factor model' (Edmonds, 1979) was presented which was based on outlier and correlation studies. This initial model has since been criticized on methodological and also conceptual grounds. But it has an impact on moving the field from the question whether teacher and/or school matters to the identification of factors that may be related with student outcomes, as is demonstrated below.

Second Phase: Searching for Factors Associated with Student Outcomes

The main research question of the second phase of EER refers to identifying those factors that help to explain differences in the effectiveness of teachers and schools. The results of studies conducted during this phase produced lists of correlates that were associated with better student achievement and which were treated as key effectiveness factors. A review of the results of the studies conducted during this phase (Levine and Lezotte, 1990; Sammons et al., 1995) resulted in numerous correlates for effective classrooms, schools, and above-school levels (districts, states, country). Together, these studies emphasized once more the importance of further developing the relatively limited theoretical foundation of EER by including the combination of correlates into categories.

Third Phase: Development of Theoretical Models

The third phase of EER saw researchers use several theoretical orientations to help explain why certain characteristics might contribute to educational effectiveness (Scheerens and Bosker, 1997). Generally speaking, there are three perspectives within EER in this phase which attempted to explain why and how certain characteristics contribute to educational effectiveness, and relevant theoretical models emerged from these approaches. These perspectives stem from economics, psychology, and sociology of education and are described in the article Theory of Educational Effectiveness (*see* Educational Effectiveness Theory and Research: Recent Advances). The limitations of using a single approach to describe educational effectiveness have also been identified. As a consequence, integrated multilevel models of educational effectiveness have been developed. Specifically, researchers attempted to integrate the findings of school effectiveness research, teacher effectiveness research, and the early input–output studies, and the models which emerged (e.g., Creemers, 1994; Scheerens, 1992; Stringfield and Slavin, 1992) have a multilevel structure, where schools are nested in contexts, classrooms are nested in schools, and students are nested in classrooms or teachers.

Fourth Phase: Analyzing in More Detail the Complex Nature of Educational Effectiveness

During the fourth phase, researchers attempted to respond to a major criticism of earlier EER that was concerned with the failure of the field to contribute significantly to the establishment of strong links between research on effective factors and developmental work so as to improve the quality of education. However, a dynamic perspective of education is now being taken into account more explicitly in theoretical and empirical EER studies (Creemers and Kyriakides, 2008). Thus, in this phase, teaching and learning are seen as dynamic processes that are constantly adapting to changing needs and opportunities. Thus, studies investigating the process of change in schools gradually emerged (Opdenakker and Van Damme, 2006). Implications of the findings of these studies for modeling educational effectiveness in a way that takes into account the complex nature of education (Creemers and Kyriakides, 2010) can be drawn. Moreover, such studies helped us look at the functioning of each effectiveness factor using a dynamic rather than an instrumental perspective. An example of such a theoretical model is presented in the article Theories of Educational Effectiveness namely the dynamic model of educational effectiveness (*see* Educational Effectiveness Theory and Research: Recent Advances). This also implies that specific developments in the methodology of EER are needed since this research does

not support the traditional approach of modeling effectiveness as a static feature or attribute of schools or teachers (Creemers et al., 2010). Modeling effectiveness or change should not be restricted into fitting conditional models in which measures of student learning outcomes (adjusted for background characteristics such as SES, gender, and prior knowledge) are regressed on a set of explanatory variables. Further, measures of change based on only one or two time points are also unreliable (Goldstein, 1997) and so provide an inadequate basis for studying change (Gustafsson, 2013). Therefore, the approach promoted in the current phase of EER does not give undue emphasis on measuring the short-term outcomes of the immediate effect of schools and teachers on student achievement gains during a single school year. In contrast, this approach instead reveals a need for longitudinal research to study results of schools and classrooms and their functioning over a longer period, modeling the growth in student outcomes over at least three time points.

Methodological Developments Promoting Theory and Research on Effectiveness

This section attempts to show that much of the progress of EER is due to advances in methodology made during the last 30 years. During the first phase of EER, major emphasis was given to conducting outlier studies and comparing the characteristics of more effective schools with those of less effective schools such as the Louisiana Study (Teddlie and Stringfield, 1993). However, the emphasis on identifying outliers was criticized for both conceptual and methodological reasons (Goldstein, 1997). During the 1980s, researchers moved to the use of mainly cohort and longitudinal designs involving larger numbers of schools and students and such studies multiplied in the 1990s. In addition, the development of hierarchical regression approaches that are involved in multilevel modeling techniques enabled researchers in the area to take the multilevel structure of educational systems into consideration and thereby deal with the methodological weaknesses of earlier studies which used only unilevel regression analysis (the 15 000 h study by Rutter et al. (1979)). Early examples of studies which made use of hierarchical regression approaches include the School Matters (Mortimore et al., 1988) and the Inner City Infant School research (Tizard et al., 1988) in England. Gradually, the way of measuring the links between inputs, outcomes, and processes became more sophisticated particularly through the development of contextual value-added models that controlled for student level prior attainment and background as well as contextual measures of school or class composition. Moreover, progress in the area of developing tests and other assessments helped researchers establish better measures of achievement not only in the area of basic skills such as reading, writing, and mathematics but also in higher-level cognitive outcomes and even in the affective and social domains.

During the second and third phases of EER, a large number of reviews were conducted and their main purpose was to provide the research community and policy makers with the latest developments from the field (Levine and Lezotte, 1990; Sammons et al., 1995). However, although these reviews were usually based on a collection of studies which were seen by the authors as providing good examples of research, their judgments of the methodological deficiency of studies that were excluded were not necessarily systematic and were often based on implicit rather than explicit criteria. On the other hand, the reviews which were not selective resulted in a huge number of factors for which not much information about their empirical support was provided (Levine and Lezotte, 1990). As a consequence, the results of these early reviews can be questioned. In this context, carrying out meta-analyses using advanced quantitative approaches can be seen as a major methodological development which promotes the theoretical development in the field and enables researchers to identify generic and more specific factors whose impact is dependent on the educational setting in which they are operating (Scheerens and Bosker, 1997).

During the third and fourth phases of EER, emphasis was given to searching for not only predictors with direct effects on student outcomes but also those with indirect effects. Moreover, the theoretical models that have been developed during this phase refer to relations among factors situated at different levels (Creemers and Kyriakides, 2008; Scheerens and Bosker, 1997). In this context, the development of multilevel structural equation modeling (SEM) approaches enables researchers to search for indirect effects and/or test the validity of the current models of EER in relation to this assumption.

During the fourth phase of EER, emphasis is given to modeling the dynamic nature of effectiveness. This implies, among other things, that longitudinal studies which last for at least 3 years should be conducted not only to measure the long-term effect of schools and teachers but also to find out how changes in the functioning of factors are associated with changes in educational effectiveness (Creemers and Kyriakides, 2010). For this reason, developments in advanced quantitative research methods such as the use of growth modeling techniques are to be welcomed because they can help us answer such research questions. Moreover, conducting longitudinal studies enables researchers to search for reciprocal relations which, according to current theoretical developments in the field of EER, are expected to exist. Such relations are often included in relevant statistical models and refer to the relation of student factors that are likely to change with achievement. For example, advanced SEM techniques can be used to search for reciprocal relations between motivation or academic self-concept and student achievement by making use of data collected at different points of time (Marsh et al., 2006).

At this point, we would like to claim that a challenge for EER is to make better use of current developments and progress in research methodology and to provide empirical support for new ways of conceptualizing the dynamic nature of educational effectiveness. At the same time, it is acknowledged that the knowledge base of educational effectiveness and its attempt to establish theoretical models challenge the development of methodologically appropriate studies and ways of analyzing data. An example is the development of SEM and multilevel techniques which help EER to search not only for linear but also for nonlinear relations between effectiveness factors and student achievement. Furthermore, a challenge for EER research methodology is to establish and use advanced quantitative techniques which will identify meaningful grouping of factors

(Kyriakides et al., 2009). Such a development will contribute to the establishment of stronger links between EER and those who are involved in school/teacher development or improvement projects that seek to promote better practice (defined as that which promotes better outcomes for students) and offers the hope that more comprehensive strategies for school improvement will eventually emerge (Creemers et al., 2013).

Improving Quality in Education: The Contribution of Educational Effectiveness

In the first parts of this article, it was made explicit that the main research question underlying most educational effectiveness studies is to identify factors at different levels (e.g., school, class/teacher) that are associated directly or indirectly with students' learning outcomes and to seek to explain (through reference to theory) why these factors influence students' learning outcomes. This implies that EER attempts to establish and test theories which explain why and how some schools and teachers are more effective than others in promoting better outcomes for their students. Thus, the first area of investigation for this field is concerned with the development and validation of the theoretical framework of educational effectiveness. By doing so, researchers attempt to establish and validate a knowledge base that can be used for improving educational practice and that is of relevance to both policy makers and practitioners.

The relevance of the knowledge base generated by researchers for use by policy makers and practitioners suggests that beyond modeling effective practices, a second area of investigation that needs to be examined is the use of this knowledge base for improvement purposes. The importance of this second area of investigation reflects the fact that the ultimate aim of EER is to improve practice and it argued that improvement is most likely to be achieved by helping schools and teachers make use of the research-generated knowledge base of EER in order to develop their improvement strategies and actions (Scheerens, 2013). In this article, we move a step forward and suggest that the research agenda of the current phase of educational effectiveness may be expanded not only by attempting to develop a valid theoretical framework but also by investigating under which conditions this framework can be used for improvement purposes. This argument is based on the fact that experimental studies are mainly conducted by researchers in this field in the USA. These studies not only help us to demonstrate cause and effect relations in a more systematic way but also to design interventions that are concerned with the improvement of factors associated with student achievement. Moreover, an attempt to make use of theories in designing school improvement interventions can be identified in several European projects (see Creemers and Kyriakides, 2012).

In this context, the dynamic model of educational effectiveness was used in developing an approach for teacher and school improvement purposes. The proposed dynamic approach promotes the design of improvement projects that are based on a theory which has been systematically tested. This approach is described in the article on school improvement (*see* School Effectiveness Research). This implies that relying only on school stakeholders' experiences can be limiting in terms of their school development and improvement, since there appears to be little, if any, evidence from the published literature that assisting the school stakeholders to engage in any developmental or improvement program without providing them with a validated framework to meet their personal needs and identify priorities for improvement necessarily improves their effectiveness. Although each school is treated as a professional community responsible for designing and implementing its own improvement strategies and action plans, school stakeholders are not left alone to design and implement their strategies and actions, but are encouraged to make use of a research and advisory team to design their action plans. Therefore, a systematic research-based approach to design, implement, and evaluate improvement efforts is promoted by the dynamic approach (Creemers and Kyriakides, 2012). More specifically, school stakeholders are expected to collaborate with the research and advisory team and develop evaluation mechanisms in order to identify priorities for improvement. During the implementation of the interventions, formative evaluation mechanisms are also established in order to help schools take decisions on how to improve their improvement strategies and action plans. A number of experimental studies demonstrated the added value of the dynamic approach for school and teacher improvement purposes (Antoniou and Kyriakides, 2011; Demetriou and Kyriakides, 2012). This implies that educational effectiveness can contribute not only to the development of theories explaining the complexity of education but also to the establishment of intervention programs that are based on a theory-driven and evidence-based approach.

The review of EER presented in this article seems to show that educational effectiveness remains a dynamic field of enquiry and, if no longer in its adolescence, is still certainly in an early stage of maturity. Educational effectiveness remains an exciting area with many possibilities to influence not only policy and practice but also the development of advanced methodological approaches to study education. There are numerous opportunities to consolidate, refine, and extend the existing knowledge base of school and teacher effectiveness and, in particular, to develop its methodology further. Moreover, the need for better links with (and contribution to) thoughtful school improvement and evaluation studies remains urgent.

See also: Classroom Level in Educational Effectiveness Research; Educational Effectiveness Theory and Research: Recent Advances; Educational Evaluation: Overview; School Effectiveness Research.

Bibliography

Antoniou, P., Kyriakides, L., 2011. The impact of a dynamic approach to professional development on teacher instruction and student learning: results from an experimental study. School Effectiveness and School Improvement 22 (3), 291–311.

Brookover, W.B., Beady, C., Flood, P., Schweitzer, J., Wisenbaker, J., 1979. School Systems and Student Achievement: Schools Make a Difference. Praeger, New York.

Coleman, J.S., Campbell, E.Q., Hobson, C.F., McPartland, J., Mood, A.M., Weinfeld, F.D., York, R.L., 1966. Equality of Educational Opportunity. US Government Printing Office, Washington, DC.

Creemers, B.P.M., 1994. The Effective Classroom. Cassell, London.
Creemers, B.P.M., Kyriakides, L., 2008. The Dynamics of Educational Effectiveness: A Contribution to Policy, Practice and Theory in Contemporary Schools. Routledge, London and New York.
Creemers, B.P.M., Kyriakides, L., 2010. Explaining stability and changes in school effectiveness by looking at changes in the functioning of school factors. School Effectiveness and School Improvement 21 (4), 409–427.
Creemers, B.P.M., Kyriakides, L., 2012. Improving Quality in Education: Dynamic Approaches to School Improvement. Routledge, London and New York.
Creemers, B.P.M., Kyriakides, L., Antoniou, P., 2013. Teacher Professional Development for Improving Quality in Teaching. Springer, Dordrecht, The Netherlands.
Creemers, B.P.M., Kyriakides, L., Sammons, P., 2010. Methodological Advances in Educational Effectiveness Research. Routledge, London and New York.
Demetriou, D., Kyriakides, L., 2012. The impact of school self-evaluation upon student achievement: a group randomization study. Oxford Review of Education 38 (2), 149–170.
Edmonds, R.R., 1979. Effective schools for the urban poor. Educational Leadership 37 (1), 15–27.
Goldstein, H., 1997. The methodology of school effectiveness research. School Effectiveness and School Improvement 8 (4), 369–395.
Gustafsson, J.-E., 2013. Causal inference in educational effectiveness research: a comparison of three methods to investigate effects of homework on student achievement. School Effectiveness and School Improvement 24 (3), 275–295.
Jencks, C., Smith, M., Acland, H., Bane, M.J., Cohen, D., Gintis, H., Heyns, B., Michelson, S., 1972. Inequality: A Reassessment of the Effects of Family and Schooling in America. Basic Books, New York.
Kyriakides, L., Creemers, B.P.M., Antoniou, P., 2009. Teacher behaviour and student outcomes: suggestions for research on teacher training and professional development. Teaching and Teacher Education 25 (1), 12–23.
Levine, D.U., Lezotte, L.W., 1990. Unusually Effective Schools: A Review and Analysis of Research and Practice. National Center for Effective Schools Research and Development, Madison, WI.
Marsh, H.W., Wen, Z., Hau, K.-T., 2006. Structural equation models of latent interaction and quadratic effects. In: Hancock, G.R., Mueller, R.O. (Eds.), Structural Equation Modeling: A Second Course. Information Age Publishing, Greenwich, CT, pp. 225–265.
Mortimore, P., Sammons, P., Stoll, L., Lewis, D., Ecob, R., 1988. School Matters: The Junior Years. Open Books, Somerset, UK.
Opdenakker, M.C., Van Damme, J., 2006. Differences between secondary schools: a study about school context, group composition, school practice, and school effects with special attention to public and catholic schools and types of schools. School Effectiveness and School Improvement 17 (1), 87–117.
Rutter, M., Maughan, B., Mortimore, P., Ouston, J., Smith, A., 1979. Fifteen Thousand Hours: Secondary Schools and Their Effects on Children. Harvard University Press, Cambridge, MA.
Sammons, P., Hillman, J., Mortimore, P., 1995. Key Characteristics of Effective Schools: A Review of School Effectiveness Research. Office for Standards in Education and Institute of Education, London.
Scheerens, J., 1992. Effective Schooling: Research, Theory and Practice. Cassell, London.
Scheerens, J., 2013. The use of theory in school effectiveness research revisited. School Effectiveness and School Improvement 24 (1), 1–38.
Scheerens, J., Bosker, R.J., 1997. The Foundations of Educational Effectiveness. Pergamon, Oxford.
Stringfield, S.C., Slavin, R.E., 1992. A hierarchical longitudinal model for elementary school effects. In: Creemers, B.P.M., Reezigt, G.J. (Eds.), Evaluation of Educational Effectiveness. ICO, Groningen, pp. 35–69.
Teddlie, C., 1994. The integration of classroom and school process data in school effectiveness research. In: Reynolds, D., Creemers, B., Nesselrodt, P.S., Shaffer, E.C., Stringfield, S., Teddlie, C. (Eds.), Advances in School Effectiveness Research and Practice. Pergamon, Oxford, pp. 111–133.
Teddlie, C., Stringfield, S., 1993. Schools Make a Difference: Lessons Learned from a 10-year Study of School Effects. Teachers College Press, New York.
Teddlie, C., Reynolds, D., 2000. The International Handbook of School Effectiveness Research. Falmer Press, London.
Tizard, B., Blatchford, P., Burke, J., Farquhar, C., Plewis, I., 1988. Young Children at School in the Inner City. Lawrence Erlbaum, Hove.

Relevant Websites

www.earli.org/special_interest_groups/educational_effectiveness – EARLI SIG 18 – Educational Effectiveness.
www.ucy.ac.cy/esf – European Science Foundation project "establishing a knowledge base for quality in education: testing a dynamic theory for education".
www.ucy.ac.cy/jls – European Commission's Daphne III program "designing evidence-based strategies and actions to face bullying by considering socio-ethnic diversities in school population and evaluating their effects".
www.highreliabilityschools.co.uk – High Reliability Schools.
www.icsei.net – International Congress of School Effectiveness and Improvement (ICSEI).
www.sree.org – Society of Research on Educational Effectiveness (SREE).

Educational Evaluation: Overview

Geoff Lindsay, Centre for Educational Development, Appraisal and Research (CEDAR), University of Warwick, Coventry, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H. Wottawa, D. Pult, volume 6, pp. 4255–4259, © 2001, Elsevier Ltd.

Abstract

Evaluation of educational programs requires evidence of both efficacy, typically from randomized controlled trials, and effectiveness of program implementation. In order to provide useful evidence to support policy, program evaluations must demonstrate that the program can work under optimal controlled conditions and that it will work when rolled out on a large scale in community settings. Evaluation benefits from a combined methods strategic plan to address these different aspects of effectiveness.

Educational Evaluation

Educational evaluation addresses the effectiveness of program interventions. Within education these typically concern learning, e.g., reading; behavioral, emotional, and social development, e.g., antibullying programs; or broader topics, e.g., whole school system changes such as inclusive education. Debates have raged within the research community about methodology, in particular the place of qualitative methods in the evaluation of program effectiveness compared with quantitative methods. There has also been strong political involvement with some governments taking positions on the type of evidence required for evaluation studies, in particular the emphasis on randomized controlled trials (RCTs).

The first aim of program evaluation is to discover whether the intervention is effective. This may be small scale, e.g., a school investigating the introduction of a new reading scheme, but evaluation can also be large scale, at school, district, state (local authority (LA)), or national level. The availability of national or state data sets such as the U.K. Government's National Pupil Database and its student-level School Census provides opportunities for large-scale evaluations of educational interventions, including curricular reforms (e.g., Bhatt and Koedel, 2012) or the differential development of groups of students (e.g., the relationship between identification of special educational needs and ethnicity: Strand and Lindsay, 2009). Increasingly, however, the importance is being recognized of the need to evaluate both the effectiveness of the program itself and its implementation.

Methods

In the past there was a strong commitment to experimental methods for evaluation research. The favored 'gold standard' method was the RCT, a design whereby participants are allocated randomly to an intervention or nonintervention group. Comparison of results indicates the relative effectiveness of the program compared with either no intervention (or 'provision as usual') or an alternative intervention, such as an existing program. This approach continues to be regarded as the most effective method for examining causal relationships, i.e., whether improvement is as a result of the program, but see Shadish et al. (2001) for a comprehensive review of methods.

The use of RCTs is increasingly recognized as necessary but not sufficient. Ultimately, a program demonstrated to be highly effective by RCTs will only be useful if its implementation in 'real life' is also effective. RCTs are typically small scale in order to maximize their rigor, which is facilitated by very strong control being exercised by the researchers. This includes high levels of specificity of both independent and dependent variables, and strong adherence to the required protocols for implementation, including number and length of sessions, who may administer the program, training requirements, and data collection. The evaluation of education programs will typically need to address complex issues. For example, schools differ in type, staff interest in the research, their levels of relevant competence and motivation. This may be at the level of the program itself or scale of implementation. For example, a program to develop reading may focus on a particular element such as letter sound-symbol or whole word learning; the development of reading, however, requires both skills and is influenced by many other factors. Rigorously controlling implementation of a program may be possible in a small-scale trial but impractical across a large number of schools. Consequently, fidelity to the program may vary: schools may change key elements of implementation, and to different degrees.

Other methods contribute to improve the evaluation of implementations. Meta-analyses enable the weight of evidence across studies to be examined: consistency of strong effects indicates support for the program's effectiveness and generalizability. However, meta-analyses typically indicate a range of effects across studies which raise the question, why do results vary? A third approach, therefore, is to examine the reasons not only for the effectiveness of a program, but also its variability and even its noneffectiveness. Qualitative methods are appropriate for this, including interviews with key participants to explore in depth what was successful about the program (perhaps not all elements were effective or necessary), for whom (variations across different participants) and, primarily, the reasons for different levels of effectiveness for different participants.

Increasingly, evaluations are designed as combined (or mixed) methods studies. Here, research methods are combined to address different research questions within the overall

examination of effectiveness, drawing on the strengths of different approaches. A number of combined methods designs are possible taking into account the relative proportion of quantitative and qualitative methods, the number and the arrangement of stages – concurrent or sequential – and, if the latter, the nature of the sequence (Leech and Onwuegbuzie, 2009). See also Tashakkori and Teddlie (2010) for a comprehensive overview.

Context

Both the micro and macro contexts of a program implementation are important. Introducing and running a program in a school will be assisted or hindered by the institution's view of the program itself, the priority of the topic in the school's development planning, and school leaders' views of supporting research given other priorities. Macro-level issues include the professional debates about a particular program and different methodologies; the political views on the importance and priority of the program domain; and the nature of evaluation. For example, recently in both the United States and England there has been a strong government commitment to evidence-based (or evidence-informed) practice, although the implications of that commitment have varied. Politics, national and local, may have an impact on evaluation in terms of technical/methodological positions, for example, the nature of appropriate evidence; or through broader values and the social policies that follow from these, and hence particular areas of intervention, particularly where funding is controlled or influenced politically.

Intervention and Implementation

There is increasing recognition that evaluation must address both the program and its implementation. Both technical/methodological factors and values, at the levels of both social policy and methodological approaches, are also important. For example, there is now appreciation of the importance and benefits of engaging research participants in the research enterprise, not simply as passive providers of data (Lewis and Lindsay, 2000): Enabling participants' voices to be heard may take many forms including research design, gathering their perspectives as data, and their involvement in dissemination (Roulstone and Lindsay, 2012).

Evidence for the effectiveness of programs has increased in education and other fields. Web-based resources provide researchers and practitioners with access to evidence for the effectiveness of interventions for learning (e.g., What Works Clearing House: http://www.ies.ed.gov/ncee/wwc/), parenting programs (United Nations Office on Drugs and Crime: http://unodc.org/), and the recent What Works for SLCN? (Speech, Language and Communication Needs: https://www.thecommunicationtrust.org.uk/whatworks). These resources are also available to the public. Hence a parent may check the evidence on an intervention proposed for their child.

The research base for evaluating implementation is less well developed than RCTs for efficacy trials (Fixsen et al., 2005) but is growing. Meyers et al. (2012) have produced a conceptual overview and proposed a Quality Implementation Framework of four phases comprising 14 critical steps:

- Initial considerations concerning the host settings.
- Creating a structure for implementation.
- Ongoing structure once implementation begins.
- Improving future applications.

Evidence: From Efficacy to Effectiveness

Educational evaluations may comprise particular research studies of an individual program. However, in order to develop evidence-based practice the evaluation enterprise must have a broader conceptualization, demonstrating that:

- Interventions are effective.
- A policy for implementation has been approved, at a local, regional, or national level.
- Systems are available for ensuring that teachers or other relevant practitioners and policy makers concerned with implementation have the required knowledge and skills to interpret the evidence.
- Practitioners and policy makers have the knowledge and skills to implement the program with fidelity.
- The context in which to host the implementation is supportive.
- There is understanding of the likely cost effectiveness and cost benefits of the program.
- There is the will to support the program with appropriate leadership and resources.

In support of achieving this comprehensive approach, a staged approach is necessary:

- The program must be derived from sound theoretical analysis for which there is also an evidence base.
- A program must be developed which operationalizes the theoretical underpinnings such that it reflects the relevant principles derived from this analysis.
- The program requires an evidence base for its content and its implementation including target population; amount, frequency, and intensity of implementation ('dosage').
- Initial trials of the program to test out the content and practical elements.
- An efficacy trial that meets quality standards, typically an RCT (Society for Prevention Research: http://www.preventionresearch.org/) and demonstrates positive effects.
- Further efficacy trials with positive results to demonstrate generalizability.
- Efficacy trials that demonstrate sustainability of the program, with effects maintained over time.
- Effectiveness studies to examine whether the program continues to show effectiveness when it is implemented on a large scale in community settings.

This approach to evaluation must be fleshed out taking into account the issues addressed above, including the use of combined methods, ensuring the voice of key participants is accessed and reported, and the importance of both practical and political in addition to technical factors. Generalization of findings to other populations must also be examined, not

assumed, including variants by ethnicity, socioeconomic disadvantage, and intellectual status (Hastings, 2013).

Example

The Parenting Early Intervention Programme (PEIP) was implemented by the U.K. Government (2006–11), initially in 18 LAs (school districts). It was one response to concerns about antisocial behavior of children and young people (Respect Task Force, 2006). A commissioned review identified parenting programs shown to be effective, typically by RCTs (Moran et al., 2004). An effectiveness study in community settings in 18 LAs (over 2000 parents: 2006–08) demonstrated that after attending one of the three programs used in the PEIP parents had lower levels of less-effective parenting skills (laxness and overreactivity) and improved mental well-being, with borderline large effect sizes (Lindsay et al., 2011a). The parents also reported their children having lower levels of behavior problems (with borderline small to moderate effect sizes).

Interviews with parents, group facilitators, and LA staff were used to identify the factors that supported or hindered successful implementation across the LAs (Lindsay et al., 2008). On the basis of the evaluation data, the Government funded a national rollout to all English LAs (2008–11). This too was evaluated (over 6000 parents) and demonstrated similarly positive effectiveness. A 1-year follow-up of over 1000 parents indicated that the improvements in parenting skills and children's behavior were largely maintained (Lindsay and Strand, 2013). The qualitative strand provided supportive evidence for the main factors for successful implementation including strong leadership, effective day to day management and organization, and a clear policy for parenting support (Lindsay et al., 2011b). Together, the evidence from quantitative and qualitative methods in this large-scale combined methods evaluation influenced national policy.

Conclusion

Implementation evaluation is a key element in determining the effectiveness of an intervention: RCTs may provide evidence of efficacy of the program, often under particularly favorable conditions, but for the development of evidence-based policy and rollout of programs on a large scale, it is essential also to evaluate the factors that increase the likelihood of success and reduce the possible adverse impact of factors likely to impair effectiveness. An evaluation strategy is needed which collects evidence on program efficacy, and on effectiveness when implemented on a large scale, and with a range of populations, allowing the development of a strong evidence base to support policy and practice.

Bibliography

Bhatt, R., Koedel, C., 2012. Educational Evaluation and Policy Analysis 34, 391–412.

Fixsen, D.L., Naoom, S.F., Blase, K.A., Friedman, R.M., Wallace, F., 2005. Implementation Research: A Synthesis of the Literature. University of South Florida, Tampa, FL.

Hastings, R., 2013. Running to catch up: rapid generation of evidence for interventions in learning disability services. British Journal of Psychiatry 203, 245–246. http://dx.doi.org/10.1192/bjp.bp.113.127605.

Leech, N.L., Onwuegbuzie, A.J., 2009. A typology of mixed methods research designs. Quality and Quantity 43, 265–275.

Lewis, A., Lindsay, G., 2000. Researching Children's Perspectives. Open University Press, Buckingham, UK.

Lindsay, G., Strand, S., 2013. The evaluation of a national roll-out of parenting programmes across England: the Parenting Early Intervention Programme (PEIP). BMC Public Health 2012 11, 962. http://dx.doi.org/10.1186/1471-2458-11-962.

Lindsay, G., Strand, S., Davis, H., 2011a. A comparison of the effectiveness of three parenting programmes in improving parenting skills, parent mental wellbeing and children's behaviour when implemented on a large scale in community settings in 18 English local authorities: the Parenting Early Intervention Pathfinder (PEIP). BMC Public Health 2011 11, 962. http://dx.doi.org/10.1186/1471-2458-11-962.

Lindsay, G., et al., 2011b. Parenting Early Intervention Programme Evaluation: DFE-RR121A. DfE, London. https://www.education.gov.uk/publications/eOrderingDownload/DFE-RR121A.pdf.

Lindsay, G., et al., 2008. Parenting Early Intervention Pathfinder Evaluation DCSF-RW054. https://www.education.gov.uk/publications/standard/publicationdetail/page1/DCSF-RW054.

Meyers, D.C., Durlak, J.A., Wandersman, A., 2012. The quality implementation framework: a synthesis of critical steps in the implementation process. American Journal of Community Psychology. http://dx.doi.org/10.1007/s10464-012-9522-x.

Moran, P., Ghate, D., van der Werwe, A., 2004. What Works in Parenting Support? A Review of the International Literature. http://www.prb.org.uk/publications/.

Ogilvie, et al., 2011. Assessing the evaluability of complex public health interventions: five questions for researchers, funders and policymakers. The Millbank Quarterly 89, 206–225.

Respect Task Force, 2006. Respect Action Plan. UK Government Home Office, London. http://tna.europarchive.org/20100413151441/http://www.asb.homeoffice.gov.uk/uploadedFiles/Members_site/Articles/About_Respect/respect_action_plan.pdf.

Roulstone, S., Lindsay, G., 2012. The Perspectives of Children and Young People Who Have Speech, Language and Communication Needs, and Their Parents. DfE, London. https://www.education.gov.uk/publications/standard/publicationDetail/Page1/DFE-RR247-BCRP7.

Shadish, W.R., Cook, T.D., Campbell, D.T., 2001. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin, Boston.

Strand, S., Lindsay, G., 2009. Ethnic disproportionality in special education: evidence from an English population study. Journal of Special Education 43, 174–190.

Tashakkori, A., Teddlie, G., 2010. SAGE Handbook of Mixed Methods in Social and Behavioral Research, second ed. SAGE, Thousand Oaks, CA.

United Nations Office on Drugs and Crime. 2009, Compilation of Evidence-based Family Skills Training Programmes. http://www.unodc.org/.

Relevant Websites

http://www2.warwick.ac.uk/fac/soc/cedar/ – CEDAR, University of Warwick, UK Evaluation Reports (accessed 09.10.13.).

http://www.earlyinterventionfoundation.org.uk/ – Early Intervention Foundation: Graduations of Early Intervention Programs (accessed 09.10.13.).

http://educationendowmentfoundation.org.uk/ – Education Endowment Foundation: Evaluations of programs to Improve the Educational Achievement of Disadvantaged Children (accessed 09.10.13.).

http://www.preventionresearch.org/ – Society for Prevention Research: Standards of Evidence (accessed 09.10.13.).

http://www.education.gov.uk/researchandstatistics – UK Government's Department of Education's annual statistics (accessed 09.10.13.).

http://www.unodc.org/ – United Nations Office on Drugs and Crime: Evidence-Based Family Skills Programmes (accessed 09.10.13.).

http://www.ies.ed.gov/ncee/wwc/ – What Works Clearing House (accessed 09.10.13.).

https://www.thecommunicationtrust.org.uk/ – What works for Speech, Language and Communication Needs (accessed 07.10.13.).

Educational Institutions, History of

Peter M Roeder, Max Planck Institute for Human Development, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 6, pp. 4266–4272, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Treatment of the topic in this article is restricted to institutions of general higher, secondary, and elementary education. It is important, however, to be aware that a clear distinction between these levels was not established before the eighteenth century. In the introductory section of the article a distinction between nonformal and formal educational institutions is proposed to characterize educational practice during the classical Greek period, compared with the schools established much earlier in the Sumerian city states or those later founded in the towns bordering the Mediterranean and Aegean Seas during the Hellenistic period. The liberal arts curriculum evolving in these latter schools had a strong impact on the course of studies in the cathedral schools and in the liberal arts faculties of the universities up to the early modern era. This is illustrated for the Middle Ages, focusing on the development of schools and of institutions of higher learning in France and somewhat later in Central Europe (Sect. 2). Section 3 examines the impact of the humanistic movement and of the Reformation on developments in the field of education. Changes in curricular contents and objectives are noted in the teaching of classical literature and languages. The role of the state in the control of educational institutions is considerably strengthened, although their relation to the church remains a close one. In the Protestant regions of northern and Central Europe a spurt in the spreading of literacy is to be observed, but even there, obligatory schooling was fully enforced only in the second half of the nineteenth century. Whereas, in France, universities and colleges of the Ancien Régime were suppressed by imperial order in 1793 and, in their stead professional schools were organized, universities in Central Europe consolidated their dominant position in higher education and succeeded in safeguarding a sphere of autonomous decision making: Section 4 recounts the evolution of the modern university between 1694 (foundation of the University of Halle), 1734 (foundation of the University of Göttingen), and 1810 (foundation of the University of Berlin), a process characterized by, for example, the substitution of Latin by the vernacular as medium of instruction, the introduction of new fields of knowledge, an increasing emphasis on research relative to teaching increasing specialization in both teaching and research, with the consequence of the emergence of new disciplines. These developments affect the role of both teachers and students. Throughout the nineteenth century, remarkable changes are discernible also at the level of elementary education (rising enrolment rates, creation of institutes for the education of elementary teachers) and secondary schools (establishment of different types of secondary schools varying in curricular emphasis, rising enrollment rates for female students, although parity was not achieved until the second half of the twentieth century, expansion of regional and central bureaucracies for both elementary and secondary schools (see Sects. 5 and 6 for parallel developments in England France and Germany)). At the turn of the twentieth century coherent systems of education had emerged in the more developed European nations, representing what has more recently been termed a 'global model of schooling.' The thesis that we have been witnessing the 'formation of a world educational system' is discussed in the final section, with reference to the activity of international agencies in educational policy and to the situation in developing countries.

Origins

Any attempt to specify a starting point for the history of educational institutions is necessarily arbitrary. This is due to the uneven and contingent distribution of early historical sources across time and space, as well as to the fuzziness of the term 'educational institutions.' Broadly speaking, the term refers to a set of rules that govern the aims and activities of the adult generation in socializing the young generation into more or less age-graded roles. Knowledge about how to profit from the extended period of human infancy in order to further physical and moral cognitive development must have been part of the everyday wisdom of the older generation long before precepts concerning education appeared in religious texts. And even after the introduction of educational institutions in a more narrow sense of the term (schools), these rule-sets governing everyday socializing educational practice remained of prime importance as these formal institutions were highly specialized and, for centuries to come, reached only a small minority of a population.

The boundaries between religious or general cultural practice and educational practice and the respective institutions in these fields are not clearly marked. Thus, for ancient Jewish education the basic set of texts is identical with the literature guiding religious practice: the Pentateuch, the oldest part of the Old Testament, i.e., the books Moses 1–5. Of course, those parts of the text referring directly to education are clearly recognizable as such but, obviously, their interpretation (e.g., in the Talmud) depends on the religious context and vice versa (Schwenk, 1996, pp. 153–177).

For example, we are most familiar with an educational practice that is not primarily channeled through educational institutions in the narrow sense of the term but rather builds on participation in cultural, social, and political activities in a highly dynamic society, namely that prevailing in ancient Greece during the classical period (for the following, see Marrou, 1948; Schwenk, 1996). This does not hold, of course, for Sparta, where the male child left the family at the age of 7 years to join a state-controlled educational system for the next 12 years bent on producing fearless and obedient warriors.

Athens at that time had not developed a public system of schools; military training of 2 years' duration was introduced only in the late fourth century BC. Education of the young (boys) was the responsibility of the family, which availed itself of the services of private teachers. Boys of fathers having full citizenship had the right, at age 10 years, to be present at most cultural events. Although the works of the poets and historians were available in writing and many of the (male) citizens must have been literate to be able to fulfill their public duties, Greek culture of the classical period was to a large degree an oral culture. Festivals, the theater, and competitions in presentations of dance, music, poetry, drama, and athletics provided ample opportunity to familiarize the young generation with the rich heritage of Greek civilization. This experience was understood to be more than a means to achieve esthetic refinement; it exposed the young to reflections on human destiny and justice, concepts of virtue and models of conduct and of good governance. It was preparation for citizenship and moral education. Physical education and music education, too, were considered fundamental for the formation of the free citizen. Especially the former was seen as part of the preparation for military service, and the communes were responsible for providing and controlling localities for athletic activities. The growing prestige of the Panhellenic games may, however, have served to weaken the nexus with military training.

Many elements of the (nonformal) curriculum of the classical period reappeared in the curriculum of the public schools provided by the larger communes during the Hellenistic era, normally dated from 338 BC when Philipp II of Macedonia defeated Athens and their confederates in the battle of Chaironeia. In both periods the curriculum was strictly oriented toward general as opposed to vocational education. The adult role emulated in the schools is not that of the skilled professional but that of the citizen potentially responsible for different general tasks in and for his community and in need of general moral as well as cognitive competences.

Educational institutions of a more restricted vocational orientation had already been introduced long before the apex of Greek civilization in the Sumerian city states of the third millennium BC. These were highly differentiated societies facing complex administrative challenges and had important trading relations. To document and control their diverse activities they needed systems of book-keeping for the different administrative fields and functions. Considering the state of development of the sign system for writing, mastering this task required systematic learning. Consequently, writing schools for trainees of the administration were established (Feldbusch, 1988, pp. 1473ff). Comparable institutions specializing in the education of scribes, administrators, and probably also artisans had already existed in ancient Egypt (Craig, 1981; Adick, 1992, pp. 164ff).

The Hellenistic system of postelementary education which prospered not only in the Greek communes but also in many regions bordering on the Aegean and Mediterranean Seas, and which functioned in the eastern part of the empire until the fall of Byzantium in1453, is important for further developments during the Middle Ages in Europe in several respects: it provides the context in which early Christian concepts of education were developed in close contact with classical pre-Christian educational theory. The curriculum of the Enkyklios Paideia Septem Artes liberales remains a point of reference in the debate on the contents and objectives of postelementary education, even gaining in importance with the growth of first-hand knowledge of the classical texts in central and Western Europe.

The Middle Ages

The cathedral schools first founded toward the end of the eighth century and spreading throughout the Carolingian Renaissance provide evidence of this type of long-term curricular continuity (Verger, 1986). Although their primary purpose was the education of the clergy via an in-depth study of the Holy Scripture, they also introduced their students to a version of the trivium grammar, rhetoric and – around the end of the tenth century – dialectic. Thus, these schools offered a relatively broad general education, including the study of classical nonreligious texts, as a basis for the professional preparation for the clergy. Some of these schools developed into universities in the first half of the thirteenth century, the largest and, with respect to the quality of its liberal arts and theological faculty, perhaps the most important among them the University of Paris, which attracted large numbers of students from England, the German-speaking regions of Europe and Italy. Among its famous teachers were Albertus Magnus (1240–48), Bonaventura (1248–57), and Thomas of Aquino (1252–58 and 1268–72) (Verger, 1986, p. 421). Their names, like Duns Scotus, Roger Bacon, and Occam, also point to the importance of the Christian orders, foremost among them Dominicans and Franciscans, as centers of the rich intellectual life and scholarly learning of the twelfth and thirteenth centuries. The attractiveness of the University of Paris was comparable to that of law schools in northern Italy (e.g., Bologna and Pavia), which served as models for the rest of Europe. The hierarchical relation between professional schools of theology, law (Orleans), or medicine (Montpelier) at the top and the ancillary and preparatory functions of the liberal arts faculty, typical of the university up to the end of the eighteenth century, is already discernible in some of the French cathedral schools of the eleventh and twelfth centuries. This may have helped to preserve the idea that higher education should be more than just a rather restricted professional training. When, by the late Middle Ages, much of the knowledge acquired in classical antiquity and preserved in the centers of learning of the eastern part of the Roman Empire had been recovered, this offered the opportunity to widen the scope of scholarly learning to approximate the full range of the classical liberal arts curriculum by including parts of the quadrivium: arithmetic, geometry, music, and astronomy. To be sure, this was already the vision of Alcuin – teacher and close counselor of Charlemagne, learned author of textbooks on the seven liberal arts and very active in the propagation of cathedral schools in the early ninth century. In most of the French cathedral schools, however, teaching of the liberal arts seems to have been restricted to the trivium.

When the first German universities were founded in the second half of the fourteenth century (beginning with Prague in 1348), they followed the organizational pattern developed in the older European universities: the liberal arts faculty was

by far the largest but ranked lowest. The faculties of theology and law enjoyed the highest prestige and the best remuneration of their teachers. In most of these universities, however, the arts faculties covered the full range of the arts, including natural science on the basis of Aristotle's writings (Paulsen, 1919, Vol. I; Ruegg, 1993, 1996, Vol. 1).

The larger and more successful European universities were international in the composition of their student bodies and teaching faculties, which, of course, was rendered possible by the use of Latin as the lingua franca of higher learning. Students were attracted by famous teachers and these, too, traveled much in search of better conditions for their work. Thus, knowledge spread fairly quickly, but was accessible only to that small minority who mastered the medium of communication. Schools that could have transmitted the necessary skills were small and widely dispersed. Demand for this type of instruction was low for a host of reasons. Until the end of the eighteenth century there was no clear demarcation between the respective functions and objectives of schools and universities.

By the end of the fourteenth century, most European towns of some importance had one or more schools. They were attached either to a parish, a convent, or a chapter, or directly sponsored and controlled by the communes. The latter form, with the supervision of the town's council, seems to have prevailed in Germany. But here, too, the relation between school and church remained a tight one. In many cases the minister of the church was himself the teacher or he may have acted as rector of the school, responsible for the selection of a teacher who, in turn, chose his own teaching assistants. The pupils were taught in two or more groups formed on the basis of age and learning progress. The main objective of the town school was to teach the pupils to read and to write Latin. It seems remarkable that these schools took in not only the children of well to-do families but sometimes also a sizable proportion of poor children living on charity (Paulsen, 1919, Vol. I, 18ff).

The Humanistic Movement and the Reformation

Around the turn of the sixteenth century, universities and schools were shaken by two interdependent events: the humanistic movement, originating in the Italian Renaissance, and the Reformation (Ruegg, 1993, 1996, Vol. 2). The humanistic movement was a polemic revolt against the culture and learning of the Middle Ages, which it condemned as a dark, narrow, uninspired, mechanistic, unnatural, and bookish educational tyranny; and it was a movement for a fresh reading of the poets and philosophers of classical antiquity in their original language – Greek, Latin, Hebrew – a re-reading freed from the pressure to prove this literature's compatibility with Christian values and dogma. It was, last but not least, a plea for a study of (human) nature, unfettered by the authority of classical literature (for a contrasting evaluation, compare Paulsen, 1919, Vol. I, Chapters 3–5; Garin, 1964, pp. 25ff).

The movement's immediate effect on the schools seems to have been rather limited, depending on the conditions encountered locally: the prestige of the classical languages was raised, new textbooks written or recommended by devotees of the movement were introduced, some schools hired a poet to teach the basics of Latin poetry and eloquentia. In a small number of schools, Greek was introduced as a new subject. The long-term effects, on the other hand, were important to both schools and universities: The curriculum of the liberal arts faculty was considerably extended to include Greek language and literature which was, like Latin literature, to be represented in a fairly wide selection of poetry philosophical, historical, mathematical, and other scientific texts. Most arts faculties also offered courses in Hebrew. This increase in service functions opened teaching positions for humanists and doubtless raised the prestige of the arts faculties. In the larger schools created in the course of the sixteenth century by combining several smaller Latin schools, too, Greek was offered as a new subject; in some of them even an introductory course of Hebrew was taught. Several of these larger academic schools in fact covered areas of knowledge that were traditionally part of the curriculum of higher education.

As for the effects of the Reformation on the development of educational systems, only two aspects will be mentioned here: it supported the relatively fast spreading of literacy in the mainly Protestant areas of northern Europe (the concept of universal priesthood combined with the accessibility of the sources of faith translated into the vernacular to every literate person). Sweden seems to have been most successful in this respect. Already in the seventeenth century, admission to full church membership was tied to the ability to read. Prussia was the first state to initiate steps toward compulsory elementary education early in the eighteenth century. Other Protestant states followed suit early in the nineteenth century. Fairly reliable quantitative data on the development of attendance at primary schools and on the spreading of literacy are available only for the decades from about 1830 onwards. In both respects, they show a considerable advantage for the Protestant territories (Flora, 1975; Flora et al., 1983).

In central Europe, in particular, the Reformation changed the relation between the university and the polity. The universities became part of a territorial state with a distinctly denominational character that assumed control over the reorganization of the religious system and of the educational institutions still closely connected with it. This process entailed a loss of autonomy on the part of the universities, a loss which increased with the growth of the central state administration This is clearly visible in the Protestant territories, where the ruling dukes and princes were actively engaged in the process of reorganizing the universities and of creating a new type of academic school (Landesschule, Furstenschule), financed directly by the state and recruiting its students more or less statewide. The control of the elementary schools, financed by the communes with few additional resources provided by the state, remained mostly in the hands of the church. Towns continued to be responsible for the establishment and financing, mostly via fees, of Latin schools or different types of intermediate schools without foreign language instruction for children of well-to-do families (Leschinsky and Roeder, 1976). To remit (partly) fees for motivated pupils from poor families seems to have been widely practiced.

In England, the impact of the Reformation on school development might have been even stronger than in central Europe: "modernization of the educational system dates from

the Reformation. It was during this period – from the early mid-sixteenth to the mid seventeenth century – that grammar schools were established in most market towns, with separate school buildings being erected and teaching being developed as a profession, and a systematic approach to the structuring of the content of education … took root" (Muller et al., 1987, p. 93).

Despite the changes observable in the schools and universities during the sixteenth century, they still had more in common with their precursors of the late Middle Ages than with the modern system of education which slowly evolved at the end of the eighteenth century and which developed with increasing rates of change and growth from the middle of the nineteenth century onwards. The main function of the university still is the transmission of a cultural tradition legitimized by traditions. Where such traditions were not established as in the case of Protestant theology, this led to bitter and fruitless quarrels between adherents of dissenting factions of Protestantism, who had their strongholds in different universities.

There seems to have been little room for creative minds in the fields of philosophy, mathematics, and natural science. Descartes, Spinoza, Newton, and Leibnitz stayed out of its bounds. Hopes for the progress of knowledge and for innovations with the potential of improving the lot of mankind concentrated not on the university but on the Academy, an association of men of proven ability, especially in mathematics, science, and technology (Stichweh, 1984, pp. 67ff). That the academies established in Germany did not pose a threat to the position of the universities is mainly due to the fact that the latter, from the start of the eighteenth century, had embarked on the road to an ambitious reform.

Reforming the University

The University of Halle, founded in 1694, acted as forerunner and model in this development. It succeeded in attracting a number of outstanding teachers, among them August Hermann Francke, one of the heads of the Pietistic movement and an original pedagogical thinker, Christian Thomasius, a jurist influenced by Grotius and Pufendorf, who also taught and published in the German language, Christian Wolf, mathematician and an eminent advocate of rationalistic philosophy, who also published textbooks in German, and J.S. Semler, a theologist (1752–91) who was among the first to introduce the historical–critical method into the study of the Holy Scripture. The curriculum of the faculty of philosophy was extended to include history, geography, mathematics, and the natural sciences; the law school offered lectures on, among others politics and natural law. Halle clearly pointed the way to a modern university that was no longer just a teaching institution but had begun to accept systematic research as a major function and service to society.

The second model university of the eighteenth century was that at Gottingen, founded in 1734. Its focus was on the modern disciplines: political sciences, history, law, mathematics, and natural sciences; however, it was equally important for the development of the study of the civilizations and literature of classical antiquity. Lectures were given regularly in German. There existed something like a common understanding that it was an obligation of the university to extend knowledge in the respective disciplines by doing original research. Like A.H. Francke in Halle, both Gesner and Heyne, the dominant figures in the field of classical studies in Gottingen, were influential in improving the preparation of teachers and the educational practice in schools.

Whereas in France the universities and colleges of the Ancien Régime were suppressed by imperial order in 1793 and, in their stead, a series of professional schools were organized (Verger, 1986), universities in Germany consolidated their dominant position in higher education and succeeded in safeguarding a sphere of autonomous decision making in the organization of their teaching and research. The pattern of the further development throughout the nineteenth century and beyond was already clearly discernible in the conceptual framework elaborated by Fichte, Schleiermacher, and W.v. Humboldt for the foundation of the Berlin university in 1810: the traditional ranking of the faculties was to be abolished; a central role was ascribed to philosophy, in the sense that a philosophical spirit ought to pervade work in all fields of study. The mission of the university was defined as science in the service of truth, i.e., doing research to extend the boundaries of knowledge. This mission explicitly included the students. They were expected, fairly early in the course of their studies, to choose topics of interest where they could do original work on a restricted scale, the results of which were to be presented and discussed in the seminar. The seminar provided a new setting for teaching and learning, adapted to the new mission of the university: a regular meeting of a professor with his more advanced students for the purpose of discussing original research presented either by the professor or by a student.

A growing specialization of both teachers and students and the formation of subdisciplines was one important effect of these developments, regarded as ambivalent by those concerned with the preparation of teachers for academic secondary schools, that profession to which most students of the philosophical faculty aspired (Paulsen, 1919, Vol. II, pp. 259–278). For a comparative perspective on the "transformation of higher learning … expansion, diversification, social opening, and professionalization in England, Germany, Russia and the United States" see Jarausch (1982).

Schooling in the Eighteenth and Nineteenth Centuries

The first decades of the nineteenth century were not only a period of change for the universities. They also saw the evolution of the dominant type of academic secondary schools (Gymnasium), the curriculum of which centered on the classical languages Latin and Greek and, to a lesser degree, on mathematics and German, without completely neglecting the study of history, geography, and science. Relations between secondary schools and universities were clearly articulated with the introduction of the final examination (Abitur) and the examination of teachers qualifying for employment (Muller and Zymek, 1987). Parallel developments were observable in France with the evolution of the Lycee and the introduction of the Baccalaureat in 1820.

With respect to the development of elementary education, wide regional variations in the actual enrollment of that age group for which compulsory education was mandated are observable in the first half of the nineteenth century. The highest proportion of enrollment around 1820 was reached in Saxony, totaling more than 80%. An average for the whole of Prussia in 1816 was (probably over) estimated at 60%. Three decades later, the least developed province of Prussia topped this mark with 70% of the age group under compulsory schooling actually enrolled. Obviously, there must have been a rather steady growth of the provision of elementary education over the period (Leschinsky and Roeder, 1976, pp. 137ff). The expansion of the provisions for teacher training seems even more remarkable: 17 new seminars for the education of elementary school teachers were established during the period, raising the total number to 38. Their yearly output of qualified teachers seems to have been sufficient to meet the growing demand.

In England, compulsory attendance was not introduced until the final decades of the nineteenth century. But there, too, a rapid increase in literacy throughout the century is to be observed, much of it due to the willingness of working class parents to pay for the education of their children in one of the many local private schools, which they often preferred to the cheaper schools established by the churches or by associations crusading for the education of the lower classes, as Thomas W. Laqueur contends (Stone, 1976, pp. 192ff).

The Emergence of Educational Systems

At the start of the twentieth century, coherent systems of education emerged in the more developed European nations. The elementary schools no longer stood apart from the rest of the system, but developed their function as feeder institutions for the secondary schools. Within the secondary system or between secondary and primary schools, new types of schools were established, varying in curricular focus prestige and entitlements conferred upon them. Well into the twentieth century, the classical academic secondary school with its clear focus on Latin and Greek languages and literature – lycees, colleges in France, grammar schools and, of course, the small number of highly prestigious boarding schools, the so-called 'Public Schools,' in England, and the Gymnasium in Germany – succeeded in defending its top position. Below this type ranked secondary schools that did not offer Greek or neither Latin nor Greek but, instead, taught modern languages and reserved more hours of instruction for the sciences: enseignement special in France, higher-grade schools in England, Realschulen, Oberrealschulen, Realgymnasien in Germany. Many of these schools emerged owing to the activities of local pressure groups convinced that the classical curriculum did not meet the educational needs of their children. The ambitious blueprints for their modernized curricula, integrating elements of vocational and of general education, however, had already been proposed from the second half of the seventeenth century onwards (Leschinsky and Roeder, 1976, pp. 168–208). Some of these types of secondary schools eventually achieved parity with the classical academic secondary schools – if not in prestige, at least in entitlements to matriculation (Prost, 1969; Muller et al., 1987; Muller and Zymek, 1987). Central and regional bureaucracies, responsible for the development and functional control of the system, were expanded early in the nineteenth century in France and Germany. This happened considerably later in England, where an effective substitute had been developed in the form of the Royal Commissions "with the brief of examining and making recommendations for the reform of all levels of education" (Muller et al., 1987, p. 88). Curricular frameworks and standards of examinations for different types of schools and teacher training were specified by the central administration, traditionally leaving most leeway for curricular decisions on the level of the individual school in England. The inclusion of females had progressed considerably throughout the nineteenth century, although parity was not achieved until the second half of the twentieth century.

A Global Model of Schooling

In reviewing the long-term evolution of educational institutions in the more developed countries of western, northern, and central Europe and in the USA, one gains the impression that, despite national peculiarities and the weight of regional traditions, the systems emerging at the turn of the twentieth century are basically similar in structure and functioning. They seem to represent what has more recently been termed a 'global model of schooling.' The thesis that we are witnessing the "formation of a world educational system" (Adick, 1995) implies that the dynamics fueling the development of systems of mass education presupposes the existence of the modern nation state and the concept of citizenship but that it is to an important degree transnational. This thesis has received substantial support in empirical research conducted by John W. Meyer, one of its principle proponents, and his collaborators (Meyer et al., 1992). Thus, an analysis of constitutions enacted since about 1870 shows that practically all of them include statements concerning the individual right of access to education and the authority and obligation of the state to provide and regulate the educational institutions (Boli-Bennet in Meyer and Hannan, 1979). The Prussian constitution of 1794 (Allgemeines Landrecht) can, at least in some respects, be cited as an important early precursor of these later developments. A study of primary school curricula enacted between 1920 and 1986 also reveals a high degree of communality with respect to, e.g., the subjects taught, hours of instruction per subject, and basic changes in the content of subjects across a large sample of countries (Meyer et al., 1992). Further support comes from case studies of educational developments in colonial and postcolonial West Africa (Adick, 1992, 1995) that document autochthonous initiatives in the course of the nineteenth and early twentieth centuries to establish a school system for the native population. The problem of whether the curriculum should cultivate native African traditions or emulate the best international models was, and still is, openly debated (see Reichmuth (1995) for the influence of Christian and Muslim communities on education in Nigeria and Wiegelmann (1999) for the development of schooling in Francophone Senegal, with special attention to the problem of the language of instruction, afflicting many of these countries).

More recently, the evolution of a global model of schooling has been actively promoted by international agencies: the Organization for Economic Cooperation and Development (OECD)'s 'Reviews of National Policies for Education' are

based upon a continuously refined set of common criteria and standards used in the evaluation of national educational systems. The European Union, too, may induce a further rapprochement between educational systems of member states; mutual recognition of qualifications certified in schools and universities can be viewed as a step in this direction. Large international quantitative studies of school achievement presuppose communalities in the curricula of primary and secondary schools, their success in constructing achievement tests, developed in international groups of researchers and valid in countries otherwise different, offers proof that such communalities exist in fact; and decisions of national educational administrations taken in reaction to test results may again push the trend toward globalization. This trend has recently accelerated, encouraged by developments such as the McKinsey report "How the world's best school systems keep getting better" (Mourshed et al., 2010), which purports to draw broadly generalizable conclusions from international studies, which would allow education systems to develop to international standards of excellence (Mourshed et al., 2010).

To be sure, the global model still leaves room for variation between countries, e.g., with respect to ensuring elite education for minorities of high achievers, in the degree of comprehensiveness of higher secondary education, in the organization of different types and levels of vocational and higher education, and in the importance assigned to specific subjects like, for example, religion.

A final remark addresses the problem of large gaps in the global model: both the quality and the inclusiveness of educational provisions are far below accepted international standards even on the level of primary education in a number of developing countries, despite considerable progress achieved in the decades between 1960 and 1980. The report prepared for the 1990 'World Conference on Education for All' presents a long list of deficiencies in, e.g., equity of access, completion of primary cycle preparation of teachers, quality of instruction, availability and quality of instructional materials, and managerial competences of personnel at the central, regional, and school level (Lockheed and Verspoor, 1990). The recommendations, elaborated in the report, for action to overcome these deficiencies may have been effective in some countries. In others, the situation may well have deteriorated further in the course of the 1990s. This, at least, is the impression gained from a survey of the conditions of primary schools conducted in 14 of the least-developed countries in 1995 (Postlethwaite, 1998). In significant parts of our world, primary education for all still constitutes a vision rather than a reality.

See also: Educational Sociology; Social Mobility, History of; Social Science and Universities.

Bibliography

Adick, C., 1992. Die Uni ersalisierung der modernen Schule. Eine theoretische Skizze zur Erklarung der weltweiten Verbreitung der modernen Schule in den letzten 200 Jahren mit Fallstudien aus Westafrika. Ferdinand Schoning, Paderborn, Germany.

Adick, C., 1995. Formation of a world educational system. In: Roeder, P.M., Richter, I., Fussel, J.-P. (Eds.), Pluralism and Education. Current World Trends in Policy, Law, and Administration. Institute of Governmental Studies Press, University of California, Berkeley, CA.

Craig, J.E., 1981. The expansion of education. Review of Research in Education 9, 151–210.

Feldbusch, E., 1988. Entstehung der geschriebenen Sprache. In: Ammon, U., Dittmar, N., Mattheier, K.J. (Eds.), Sociolinguistics. An International Handbook of the Science of Language and Society, vol. II. Walter de Gruyter, Berlin, pp. 1469–1479.

Flora, P., 1975. Indikatoren der Modernisierung. Ein historisches Datenhandbuch. Westdeutscher Verlag, Opladen, Germany.

Flora, P., et al., 1983. State, Economy, and Society in Western Europe. The Growth of Mass Democracies and Welfare States, vol. I. Campus Verlag, Frankfurt, Germany. Macmillan Press, London; St. James Press, Chicago.

Garin, E., 1964. Geschichte und Dokumente der abendlandischen Padagogik I. Mittelalter. Rowohlt Taschenbuch verlag, Reinbek, Germany.

Jarausch, K.H. (Ed.), 1982. The Transformation of Higher Learning 1860–1930. Klett-Cotta, Stuttgart, Germany.

Leschinsky, A., Roeder, P.M., 1976. Schule im historischen Prozess.Zum 'Wechsel erhaltnis on institutioneller Erziehung und gesellschaftlicher Entwicklung. E. Klett Verlag, Stuttgart, Germany.

Lockheed, M.E., Verspoor, A.M., 1990. Improving primary education in developing countries. A review of policy options. Draft prepared for World Conference on Education for All, Bangkok, 1990.

Marrou, H.I., 1948. Histoire de l'Education dans l'Antiquite. Editions du Seuil, Paris.

Meyer, J.W., et al., 1992. School Knowledge for the Masses: World Models and National Primary Curricular Categories in the Twentieth Century. Falmer Press, Washington, DC.

Meyer, J.W., Hannan, M.T., 1979. National De Elopment and TheWorld System, Educational, Economic, and Political Change, 1950–1970. University of Chicago Press, Chicago.

Mourshed, M., Chijioke, C., Barber, M., 2010. How the world's most improved school systems keep getting better. McKinsey & Co, London.

Muller, D.K., Ringer, F., Simon, B., 1987. The Rise of the Modern Educational System: Structural Change and Social Reproduction 1870–1920. Cambridge University Press, Cambridge,UK. Editions de la Maison des Sciences de l'Homme, Paris.

Muller, D.K., Zymek, B., 1987. Datenhandbuch zur deutschen Bildungsgeschichte. Bd.II, 1. Teil. Sozialgeschichte und Statistik des Schulsystems in den Staaten des Deutschen Reiches1800–1945. Vandenhoeck und Ruprecht, Gottingen, Germany.

Paulsen, F., 1919. Geschichte des gelehrten Unterrichts auf den deutschen Schulen und Uni ersitaten om Ausgang des Mittelalters bis zur Gegenwart, 2 vols. Veitand, Leipzig, Germany.

Postlethwaite, N., 1998. The conditions of primary schools in least developed countries. International Review of Education 44, 289–317.

Prost, A., 1969. Histoire de l'Enseignement en France, 1800–1967. Colin, Paris.

Reichmuth, S., 1995. Christian and Muslim communities and their influence on education in Nigeria. In: Roeder, P.M., Richter, I., Fussel, H.-P. (Eds.), Pluralism and Education. Current World Trends in Policy, Law, and Administration. Institute of Governmental Studies Press, University of California, Berkeley, CA.

Roeder, P.M., Richter, I., Fussel, H.-P. (Eds.), 1995. Pluralism and Education. Current World Trends in Policy, Law, and Administration. Institute of Governmental Studies Press, University of California, Berkeley, CA.

Ruegg, W. (Ed.), 1993, 1996. Geschichte der Uni ersitat in Europa, 4 vols. C. H. Beck, Munich, Germany.

Schwenk, B., 1996. Geschichte der Bildung und Erziehung on der Antike bis zum Mittelalter. Deutscher Studien Verlag, Weinheim, Germany.

Stichweh, R., 1984. Zur Entstehung des modernen Systems wissen schaftlicher Disziplinen. Physik in Deutschland 1740–1890. Suhrkamp Verlag, Frankfurt Main, Germany.

Stone, L. (Ed.), 1976. Schooling and Society. Studies in the History of Education. Johns Hopkins University Press, Baltimore.

Verger, J. (Ed.), 1986. Histoire des Uni ersites en France. Bibliotheque Historique Privat. France, Toulouse.

Wiegelmann, U., 1999. Alphabetisierung und Grundbildung inSenegal. Ein empirischer Vergleich zwischen modernen undtraditionellen Bildungsgangen und Schulen. IKO-Verlag fur interkulturelle Kommunikation, Frankfurt Main, Germany.

Educational Learning Theory

Kelly YL Ku, Hong Kong Baptist University, Hong Kong
Sivanes Phillipson and Shane N Phillipson, Monash University, Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A broad orientation to the principles and mechanisms underlying three major views of learning, namely, the behavioral perspective, the cognitive perspective, and the sociocultural perspective are presented. The focuses are on how these perspectives differ in their grounding assumptions and beliefs on how learning occurs and what constitutes the results or products of learning, as well as their impacts on research, practices, and educational policies.

This article presents a broad orientation to the principles and mechanisms underlying different views of learning, with an emphasis on how different representations of learning have affected and will continue to affect educational research and practice.

A brief account of the development trends of the paradigms of learning is first presented. Three major perspectives of learning are discussed, namely, the behavioral perspective, the cognitive perspective, and the sociocultural perspective. The focus is on how these perspectives differ in their assumptions and beliefs about how learning occurs, what constitutes the results or products of learning, and how they impact research and instructional approaches.

Introduction

Modern learning theories originated in the behavioral view of learning, which took a straightforward approach to learning as a stimulus–response relationship. The behavioral perspective on learning has been applied in many educational settings since the 1920s. Instructional strategies that are now widely adopted in classrooms, specifically those of classroom management and behavioral modification, are rooted in behavioral principles. How the mind makes it possible to understand, know, and think became a revolutionary focus in the 1950s as cognitivism challenged the behavioral view of the mind as a 'blank slate.' The cognitive perspective views knowledge as the key learning outcome. In learning, cognitive processes and functioning determine knowledge acquisition and construction. Cognitive processes are, therefore, important for educators to know and to incorporate into the design of instruction to best facilitate learning.

Learning can also be conceptualized as a socially engrained activity. Sociocultural perspectives have been used to explain this kind of learning. Modern learning theory recognizes that learning depends on the interaction of components within a developing system. In other words, learning depends on the maturation of the individual in his or her social context. As a consequence, research in educational learning theory seeks to understand the interactions between these components and the immediate and extended environment. Modern learning theories take into account the special educational needs of students with emotional and behavioral disorders, and those with the potential for exceptional academic achievement. Increasingly, research focuses on the link between learning needs and culture, particularly within the ethnically diverse classroom (Phillipson, 2011b). Hence, current research on learning recognizes the important contributions of parents, teachers, peers, and wider sociocultural influences and seeks to understand how the social environment can be manipulated to enhance learning outcomes. Table 1 gives an overview of the attributes of the behavioral, cognitive, and sociocultural learning theories.

The remaining sections elaborate on the three major paradigms in education, beginning with the behavioral and cognitive. This is followed by a description of the sociocultural view. The final section describes individual differences underlying learning processes and learning outcomes. The article closes with a brief account of current trends and related educational policies.

The Behavioral Perspective: Learning as Responding to Environmental Stimuli

Behaviorism, which originated in psychology, asserts all human behaviors result from interacting with the environment. Behaviorists regard a learning outcome as a relatively permanent change in a person's behavior caused by environmental influences such as experience or practice. J.B. Watson (1878–1958), an American psychologist and the founder of behaviorism, regarded learning as fundamentally physiological and mechanistic. In other words, learning is an automatic and involuntary reaction to external stimuli. Later, B.F. Skinner (1904–90) expanded the behavioral perspective of learning into the investigation of voluntary action. What is central to both involuntary and voluntary learning is the rejection of introspective examination of the mind; cognition, feeling, purpose, aim, and will are excluded in seeking to understand the mechanism of how we learn, as these internal processes cannot be publicly observed, controlled, and studied. This view encouraged and emphasized the importance of formulating principles of learning through the empirical study of observable behavior.

Pavlov's theory of classical conditioning and Skinner's operant conditioning characterized the behavioral principles of learning involuntary and voluntary actions respectively. They both describe the stimulus–response relationship that results in a person learning to perform an act or to respond in a certain

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92150-0

Table 1 Overview of the three major learning theories

	Paradigm		
	Behavioral	*Cognitive*	*Sociocultural*
Principles of learning	Learning occurs as a result of stimulus–response relationship. Motivation to learn comes from external sources. Internal mental processes are irrelevant.	Learning occurs as a result of the development of cognitive abilities and use of cognitive strategies in remembering, integrating, and making sense of the world. Learning might take place without observable changes.	Learning occurs as a result of the maturation of an individual in his or her social context through interaction with others who are more skilled and knowledgeable in the system. Learning takes place in reciprocal conditions.
Factors affecting learning	Environmental influences, reward, punishment, etc.	Cognitive development, environmental influences, cognitive strategies, beliefs, self-efficacy, etc.	Cognitive development, significant other environmental influences, cognitive strategies, beliefs, etc.
Role of learners	Passive and have little awareness of the learning process	Engage mentally in active and purposeful construction of new information and experience	Engage with others and the environment in active and purposeful construction of new information and experience
Products of learning	Overt behavior changes	Intellectual growth and knowledge acquired	Intellectual growth, adaptation to the environment, and knowledge acquired
Key theories	Classical conditioning (Pavlov), operant conditioning (Skinner)	Model of cognitive development (Piaget), information processing theory (Atkinson and Shiffrin), levels of processing (Craik and Lockhart)	Zone of proximal development (Vygotsky), cultural behavior (Wertsch), apprenticeship (Rogoff), ecological system (Bronfenbrenner), social cognitive theory (Bandura)
Educational implications	Widely adopted in creating learning environment that shapes behavior in desirable ways; strategies derived include token economy, response cost, and time-out procedures, etc.	Essential to contemporary understanding of how learners store, retain, and reason information; allow educators to create pedagogical strategies that facilitate the organization and transmission of knowledge	Widely adopted as effective ways in structuring learning where teachers are active facilitators in scaffolding strategies in classroom teaching and assessments; pedagogical strategies target individuals to reach his or her potential in learning.

way. In particular, commonly used strategies, such as the token economy, response cost, and time-out procedures, arise from the theory of operant conditioning. These strategies aim at managing observable behaviors through consequences. If applied consistently, they can be effective and powerful in inducing learning motivation and modifying inappropriate behaviors.

Involuntary Learning: The Theory of Classical Conditioning

Classical conditioning demonstrated involuntary or reflexive learning through association with stimuli. The founder of the theory, Ivan Pavlov (1849–1936), was a Nobel Prize-winning Russian physiologist and neurologist whose earlier research focused on involuntary reflexes of animals and later extended to children. In Pavlov's (1927) experiment on animals' digestive systems, he found that dogs learned to salivate when the sound of a bell was repeatedly paired with the presentation of food. Later, the learned response to salivate upon the sound of the bell continued even without the presentation of food. This automatic form of learning is what is called 'conditioning.' Watson extended Pavlov's finding into the study of more complex reflexes, such as emotional response. In a controversial experiment, he conditioned a young child named Albert to develop a fear of rats and objects that resembled them.

In education, the theory of classical conditioning provides the basis for understanding and treating test anxiety, school phobia, and similar conditions. For instance, test anxiety is a physiological response that can be unlearned if the presentation of tests is repeatedly associated with a stimulus that elicits pleasant emotion in students (Legge and Harari, 2000); teachers should consistently present nonthreatening testing environments to induce a positive reaction in their students.

Voluntary Learning: The Theory of Operant Conditioning

The word 'operant' refers to what we do to operate and function. Edward Thorndike (1894–1949) formulated the Law of Effect, which states that a positive consequence will lead to repetition of a behavior and that a negative one will discourage a behavior from being repeated. This stimulus (consequence)–response association forms the basic mechanism behind operant conditioning. The theory, which was named by Skinner in 1930s, describes the effect of consequences on the learning of voluntary behaviors. Based on the principle that a behavior is contingent on the desirability of consequences, two concepts are central to operant conditioning: reinforcement and punishment.

Strengthening a Behavior

Reinforcement is the addition of desirable, or the removal of undesirable, stimuli from the environment immediately after a behavior is performed so as to cause the particular behavior to happen again. Academic performance can be reinforced by praise, teacher recognition, opportunity to engage in a certain school activity, and tangible rewards such as food and prizes. The same reinforcer may not appeal to all students. For instance, the token economy is a reinforcement strategy whereby students are rewarded with a token (e.g., a sticker) that can be used to exchange things with an inherent value, such as extra activity time or access to computer learning materials. In contrast, the strategy of response cost involves taking away an earned reward for misbehavior. It maximizes the effect of a reinforcer when used in conjunction with the token economy. Providing reinforcement at the right time is the key to sustaining a learned behavior. In general, reinforcement needs to be provided consistently. Yet the schedule of reinforcement can also be varied in terms of interval or ratio so that a behavior is not reinforced every time.

Recent research has been exploring the effect of a combination of adding desirable and removing undesirable stimuli on task performance (Bouxsein et al., 2011). Moreover, research has also begun to examine the physical and neurological effects of reward and punishment on learning different tasks.

Weakening a Behavior

In contrast, punishment reduces the probability of a behavior being repeated. When a behavior is followed by an unpleasant stimulus (e.g., public reprimand), or when a pleasant stimulus is removed (e.g., recess time), the behavior is less likely to happen again. Although punishment can immediately stop or suppress a behavior, it can instill fear, anxiety, and avoidance of the punisher. A popular strategy that is used by teachers and psychologists to modify disruptive classroom behaviors in young students is the time-out. A time-out is a form of punishment whereby a student is temporarily removed from an environment that reinforces a bad behavior. There are several ways to impose a time-out, and factors such as its duration, the location, and the student's understanding of the reasons underlying the time-out may affect its effectiveness.

Punishment is not considered effective for promoting learning as it very often leads to negative emotions, which are counterproductive to learning (Baumrind, 1997). Psychologists have called for caution in using physical punishment on young children as it could lead to aggressive behaviors in childhood and adulthood. Meta-analyses (Gershoff, 2002) on parental punishment and children's behavior have revealed that parental corporal punishment is related to higher levels of immediate compliance, but also to aggression and lower levels of moral internalization.

The Cognitive Perspective: Learning as Mental Operations and Reasoning of Information

The study of cognition can be traced to Gestalt psychology in the late nineteenth century – a school of psychology that studied how the mind perceives and organizes pieces of information into a structured whole, and derives from it a meaningful representation of the world. Leading cognitivists Jean Piaget (1896–1980), George Miller (1920–2012), Allen Newell (1927–92), and Herbert Simon (1916–2001) argued that mental activities define a human as an active learner who selects, filters, evaluates, and predicts according to his or her needs and goals.

Cognitive learning theories demonstrate a variety of types of cognitive functioning including how we remember, how

new information is integrated with prior knowledge to create new knowledge, how cognition develops, and how meaningful learning occurs. Together, they exert a strong influence on modern education as they reveal how knowledge is transmitted from teachers to students, and how students transfer knowledge from one domain to solve new problems in another.

Knowledge Acquisition

Piaget's view on human cognition has greatly influenced educators' understanding of children's minds. When a child learns, he or she navigates the world through two mutually dependent processes: assimilation and accommodation. In assimilation, previously learned information is used to make sense of new information. When the new information fails to fit in to existing understanding, accommodation takes place by altering and adding to prior understanding. In addition, the acquisition of knowledge requires our mind's readiness. Piaget's (1970) model of cognitive development describes children's changes in cognitive functioning as they grow. He regarded human cognitive advancement as a function of biological maturation and proper environmental stimulation. There are individual differences in the speed of development yet all children go through universal stages in the same sequence. Each of Piaget's stages (sensorimotor, preoperational, concrete operational, and formal operational) presents distinct cognitive characteristics for teachers to take into consideration when preparing learning materials. For instance, abstract mental manipulation does not take place until the child has reached the formal operational stage (around 11 years of age); hence mathematical and logical operations that require deductive inferences are not within the developmental stage of very young children. Piaget's model highlights the importance of teaching children what is within their developmental stage. He opposed teaching children beyond their mental capability.

Knowledge acquired must be properly retained for effective retrieval and application. The analogy of how a computer works best describes how the mind processes information. Atkinson and Shiffrin's (1968) multistore model describes a three-stage information-processing sequence of encoding, storage, and retrieval. As human memory is selective and has a limited storage capacity, instructional aids have been designed to facilitate the transformation of information into long-term storage. For instance, students' attention can be engaged by combining visual and verbal forms of teaching materials and using advance organizers that enable teachers to present information in hierarchical and graphical ways that would enable students to sort through what is relevant and what is not. The use of analogies and concept maps is also helpful for making connections between new and prior concepts explicit and concrete.

Teaching to Facilitate Knowledge Construction

The levels of processing theory, proposed by Craik and Lockhart (1972), describe knowledge retention as a function of the depth of analysis. Elaborated perceptual analysis results in less forgetting. Material that relates to a person's own interests is better memorized and more effectively transferred to the solving of new problems. Therefore, learning materials must be understood in a meaningful manner, and not through mere repetition, if true learning is to take place. This view has prompted modern education to move from knowledge retention to knowledge construction. The nature of knowledge construction is purposeful and dynamic, with the aim for students to reason about subject knowledge critically and independently. Its process emphasizes the integration of new information with prior knowledge, metacognitive controls, and strategic use of cognitive skills.

The traditional direct instructional method has been challenged by inquiry-based learning (see Dewey, 1938; Lashley et al., 2002), which encompasses many instructional strategies, including, but not limited to, inquiry learning, expository learning, discovery learning, problem-based learning, and experiential learning. These modes of learning are characterized by elements of collaborative learning activities, investigation of ill-structured problems, hands-on field experience, and an interactive classroom context. These elements call for students to work together to define, examine, analyze, evaluate, and present their own ideas. This practice allows students to use their higher order thinking skills to make decisions and judgments, and solve problems. At the core of inquiry-based learning is the principle that students as learners do not merely process and repeat information but actively make sense of it to generate new knowledge. The use of inquiry-based methods does not dismiss the significance of direct teaching as some base of knowledge must be taught systematically for inquiry-based learning to take place.

Constructing Beliefs about Knowledge

Students' conception about the nature of knowledge is essential to knowledge construction. Personal epistemology refers to a person's beliefs about and representation of the nature of knowledge and knowing. Kuhn (1992) has posited a developmental model that describes four stages of epistemological development: realist, absolutist, multiplist, and evaluativist. Kuhn's stages reflect a progression from an entirely objective regard of knowledge to an entirely subjective one, and eventually to a relative view of knowledge. In the early stages, learners rely on external sources to make decisions; those in the later stages view knowledge as less absolute and rigid. In line with Kuhn's model, a five-dimension model of epistemological beliefs developed by Schraw et al. (2002) is also commonly used instrument for assessing individual's personal epistemology. The five dimensions include *simple knowledge* (knowledge as isolated information vs complex), *certain knowledge* (knowledge as absolute vs tentative), *omniscient authority* (knowledge is drawn from authority), *quick learning* (learning occurs quickly or not at all), and *innate ability* (intelligence is a fixed entity). Together, they describe a person's understanding of what knowledge is and how it is acquired. Individuals with sophisticated beliefs of knowledge tend to hold a more open-minded view of conclusion and judgment, and would value more the use of reason and evidence in supporting one's point of view. Epistemological beliefs have been shown to influence students' approaches to argument analysis and critical thinking, and their academic performance.

Controlling Cognitive Processes

Efficacy in learning is a function of how much we learn as well as how we go about learning. Self-regulatory learning consists of monitoring, evaluating, and reflecting on one's task progress and performance. These strategies are called metacognitive strategies. Metacognition is the control of cognition (Flavell, 1979). It includes applying a set of metacognitive strategies to the use of cognitive skills on a given task. Self-regulatory learners devote their attention to completing the task to their best ability, make plans for how to go about completing a task, seek alternatives, and constantly evaluate their own path of thought. Self-regulatory learning can be developed in young children. Effective instruction should include guidelines on self-monitoring and reflect on the way to go about learning, which will help students to link the learning method, as opposed to ability, to performance. The responsibility of learning is borne by the learner, and supported by the teacher, which creates independent learners who are clear about their goals, strengths, and weaknesses, and strategic in their approach to learning.

The Sociocultural and Systems Perspectives: Learning as Social Interaction

Sociocultural theory highlights the importance of society's contribution to individual development by building upon the interaction between developing and learning individuals, and the society and culture in which they live. The basic assumption of sociocultural theory in relation to learning is that a learning context should be observed and interpreted in the broader social system in which the learning is taking place. This would allow a sociocultural theorist to understand an individual's thinking and learning based on his or her participation in a socially and culturally integrated activity. Sociocultural theory draws from Lev Vygotsky's (1896–1934) work from the early 1930s, which has been elaborated on by theorists such as Aleksei Leontiv (1903–79), James Wertsch (1985), Michael Cole (2010), Harry Daniels (2001, 2008), and Barbara Rogoff.

Parallel to sociocultural theory is the system type of approach to learning, specifically the Bronfenbrenner model and the social imitation theory of social cognitive theory. These two theories again make society the nesting place where interactions or links established within a society provide the basis for learning.

Sociocultural Theory

Sociocultural theory, according to Vygotsky, requires the researcher to probe the environment surrounding students in order to understand their mental development. Two overlapping aspects of Vygotsky's sociocultural theory are important for the purposes of explaining learning: the observation of students' cognition within a sociocultural plane; and mediation and internalization, which take place between the social environment and the individual to facilitate learning and development.

Vygotsky (1978: p. 57) viewed students' cognitive development as existing on two planes: the social or 'between the people' plane, and the individual or 'inside the child' plane. The 'between the people' or interpsychological plane is the catalyst of the development of students' higher mental functions. The interpsychological plane is the platform for the development of mental functions through social interactions, which take place between students and other people who are more competent, including parents and peers.

The way in which a student learns and develops is determined by mediation between the student's psychological self and his or her social or cultural environment. Vygotsky (1978) regarded mediation as central to the acquisition and progression of students' higher mental functions. Mediation occurs initially between the student and the social plane, and is followed by internalization within the student's intrapersonal self. Together with signs and symbols, language is seen as a tool of culture and psychology that is part of the mediation process.

Vygotsky argued that mediation and internalization exposed students to 'mature cultural forms of behavior' that linked the external and internal planes in a 'genetic' relationship. In practice, this means that parents, teachers, peers, and others who have more knowledge and experience of the sociocultural environment become filters for students' learning and cognitive development. With guidance and instructions from more skilled partners, students can then absorb and internalize the information and thinking tools provided to them at different levels and in different ways.

Neo-Vygotskians such as Wertsch and Rogoff extended the premises of learning that happens on the social and personal planes to include aspects of cultural behavior and apprenticeship, respectively.

Wertsch used Vygotsky's thesis to demonstrate that without the social and cultural interaction, minimal learning happens. Wertsch, however, illustrated the learning that happens in classrooms. He showed that the cultural aspects of language and the associated psychological tools or mediatory individuals such as teacher or parent mediate classroom learning. Wertsch demonstrated that when language and cultural competence are integrated into learning, optimal learning results (Wertsch, 2007).

Rogoff expanded Vygotsky's thesis of the significant other in the process of learning. Rogoff's research centered on the idea that a young apprentice could learn constructively and productively by being mentored (Rogoff, 1990). Rogoff also showed how apprentice learning could happen on the social, personal, and institutional planes. She stressed that for learning to happen, the interaction between apprentice and mentor must transcend the social and personal within the institutions from which they are learning.

Research using sociocultural theory as a framework has adopted two approaches to explain the interactional and mediatory aspects of the learning process. The first approach is observation of the processes between student and teacher within the learning context. Through her intrapersonal, interpersonal, and cultural lenses, Rogoff applied the methodology of observation as a valid method of documenting learning. Other researchers have used statistical modeling to show the mediation process of learning. This is usually the case when the learning is within a context that has embedded social, cultural, and personal planes.

Bronfenbrenner's Ecological System

In Urie Bronfenbrenner's (1979) bioecological systems theory, it is possible to see the influence of Vygotsky's sociocultural theory where the child's social setting is an important aspect of his or her development. Bronfenbrenner's model focuses on the child and interactions between the child and the various social contexts. Some of these interactions will encourage the child's learning; others will inhibit it.

This system theory is represented as a series of environmental levels with the child in the first level known as the microsystem, which represents the most direct influences on the child's development. In addition to biological factors, these factors include the child's family, friends, classmates, and teachers. Here, interactions add to the child's experiences and development, which in turn affects the people around the child.

The next level, the mesosystem, represents the interactions among the components of the microsystem. For example, parents and teachers may work together to support the child's education. However, the child's immediate friends or classmates may discourage the child from participating in school.

The third environmental level is the exosystem, which describes the link between the child's immediate contacts and the social setting of these contacts. For example, the child's teacher attends a university course as part of his or her continuing professional development and learns about a new teaching strategy that may indirectly affect the child's learning experience. Other elements in the exosystem are the interaction between the child's parents and the parents' occupations.

The final level in the model is the chronosystem. This level explains the changes that occur over time. For example, the parents may divorce and remarry, the child moves from Grade 2 to Grade 3, or there are political or social changes in the country where the child lives.

Social Cognitive Theory

Albert Bandura's (1971) social cognitive theory explains how children can learn by imitating other modeled behaviors through reciprocal interaction among cognitive, behavioral, and environmental influences. Bandura stressed that for learning to occur, several cognitive and affective elements must exist in the reciprocal interaction of modeling between a person such as a teacher and a child as a student. These elements are attention paid to learning, retention of what is remembered and learned, reproduction or application of learned concepts, and the motivation or intention to learn (Bandura, 2001).

Similar to Vygotsky's view of dynamic interaction, Bandura believed in 'reciprocal determinism,' which assumes that a person's behavior is a result of his or her social world and that the world is a result of that person's behavior and action. Ultimately, when we think of learning in terms of Bandura's theory, we see it as a consequence of one's attention to what is being taught and one's intention to remember what is attended to. Following on from this is the person's motivation to learn and ability to apply what is learned. In many instances, this is seen as an imitation process of the person who is teaching by the person who is being taught.

Educational Implications of Sociocultural Perspectives

Vygotsky suggested that meaningful and dynamic instructions that form the social and cultural nature of interactive and reciprocal learning encourage the psychological development of a child. From this notion, Vygotsky conceptualized the zone of proximal development (ZPD), which is the difference between a child's 'actual development level as determined by independent problem solving' and the increased level of "potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky, 1978: p. 86). The ZPD focuses on the interaction of learning and development that emphasizes thinking and speech as two essential components of verbal thought that dynamically work together for children to make meaning and sense during the learning process.

Vygotsky (1960) saw the dynamic activity as a social process where the participants bring to the process their own knowledge and experience. He suggested that children are capable of solving problems beyond their actual developmental level or of engaging in higher order thinking when interacting with a teacher or parent. This process can also be situated within the imitation process of social cognitive theory and the broader system process. From this notion, dynamic assessment is a probing procedure that is constructed on an empirical assumption that the beginning and the last levels of a learning task are better remembered and learned than the ones in the middle. Structured scaffolding or hinting and probing take place when a student fails at each level of learning task until retaining enough of the learning to reach the next level or his or her full potential.

Individual Differences in Learning

Differences in educational outcomes might be attributable to individual differences in intelligence, creativity, motivation, and aspects of *self* such as self-concept, self-esteem, and self-efficacy. Research concerning the bases of individual differences has been highly influential in education. For example, Strenze (2007) concluded that psychometric (g) intelligence is a powerful predictor of educational, occupational, and financial success. Although teachers do not routinely test their students' intelligence, a number of alternative models of intelligence, such as Howard Gardner's multiple intelligence theory, have become popular among educators because they are inclusive and can provide a framework for curriculum development.

Creativity research notes the link between cognition and learning. Gelman and Gottfried (2006), for example, have argued that creative processes are necessary for the development of language and the construction of new knowledge.

Many models of learning are based on the relationship between *academic* self-concept, self-esteem, self-efficacy, and academic achievement (Hattie, 2009). Hattie concluded that academic self-concept and self-efficacy begin with past academic achievement, but that self-esteem does not. In the classroom, teachers focus on raising the academic self-concept and self-esteem of students.

Achievement goal theory focuses on the relationships between goal type (learning goals and performance goals),

cognitive processes, and academic achievement. Although there is broad agreement that learning goals lead to better educational outcomes, research suggests that performance goals can also be effective, particularly for Chinese students (Phillipson, 2011a). The causal link between the classroom structure and the development of the different goal orientations is an important area of educational research.

Changing Paradigms and Related Educational Policies

Educational policies are based on paradigms of 'quality' education, with quality defined in terms of economic indicators such as efficiency, effectiveness, economy and accountability, and educational outcomes. The development of educational policy is influenced not only by our understanding of the psychology of learning but also by advances in the psychometric instruments that measure student learning. For example, the No Child Left Behind Act of 2001 in the United States relies on research on psychometrics to define standards and to measure student academic achievement.

Educational systems around the world are more responsive than ever to international tests of academic achievement, including the Program for International Student Assessment, coordinated by the Organization for Economic Co-operation and Development, and the Progress in International Reading Literacy Study and Trends in International Mathematics and Science Study, both coordinated by the International Association for the Evaluation of Educational Achievement. Again, the results of these tests impact the development of educational systems and policy on student achievement, and are dependent on advances in psychometrics.

In response to these changing paradigms, many educational systems are implementing educational reforms. In Hong Kong, for example, these reforms include a reorganization of the secondary and tertiary school structure, the reintroduction of Liberal Studies into the curriculum, the adoption of an outcomes-based curriculum, and the selective implementation of English as the medium of instruction. In broad terms, the purpose of these reforms is to enable all students to 'develop their potentials (sic) and lead a rich life in the new century' and places no numerical limit on the number of students capable of achieving exceptional levels of performance.

With a greater focus on accountability in education, teachers have become increasingly aware of the importance of research about assessment theory. With a foundation in statistics, educational assessment takes into account the validity and reliability of assessment results and the need for measurement objectivity. Accordingly, many teacher education courses include training in assessment theory and practice, and advances in item response theories such as Rasch modeling.

Modern learning theories have recognized that learning outcomes are affected by the interactions between students and their immediate and wider environments. Based on Bronfenbrenner's bioecological systems theory of human development, the immediate environment includes peers, parents, and teachers. Attempts to model the interactions between students and their environment include systems models of learning such as Ziegler's (2005) actiotope model of giftedness (AMG). Although focusing on the development of exceptionality, the AMG can be used to understand the impact of the environment on academic achievement more broadly as well as to identify the basis of underachievement. In particular, the AMG integrates research across a number of paradigms, including motivation; aspects of 'self' such as self-concept, self-esteem, and self-efficacy; and the educational and cultural environment.

Hattie's (2009) meta-analysis of the effects of influence on student learning concluded that the most important predictors of student learning originated from the teacher, including feedback related to attributions of success, instructional quality and type, and classroom environment. In terms of learners, the most important predictor of student learning apart from past success was their learning disposition. At the same time, parental expectations of academic achievement were the most important predictor of learning originating from the home. However, the discrepancy between research and classroom practice is often significant, and teacher education programs need to strengthen the relationship between research and practice. In addition, mechanisms need to be developed to ensure that teachers remain cognizant of research on learning.

See also: Cognitive Development: Mathematics Learning and Instruction; Education and Learning: Lifespan Perspectives; Learning Theories and Educational Paradigms; Metacognitive Development: Educational Implications; Piaget's Theory of Human Development and Education; Vygotsky's Theory of Human Development and New Approaches to Education.

Bibliography

Atkinson, R.C., Shiffrin, R.M., 1968. Human memory: a proposed system and its control processes. In: Spence, K.W., Spence, J.T. (Eds.), The Psychology of Learning and Motivation, vol. 2. Academic Press, New York.

Bandura, A., 2001. Social cognitive theory: an agentic perspective. Annual Review of Psychology 52, 1–26.

Bandura, A., 1971. Social Learning Theory. Prentice-Hall, Englewood Cliffs, NJ.

Baumrind, D., 1997. Necessary distinctions. Psychological Inquiry 8, 176–182.

Bronfenbrenner, U., 1979. The Ecology of Human Development: Experiments by Nature and Design. Harvard University Press, Cambridge, MA.

Bouxsein, K.J., Roane, H.S., Harper, T., 2011. Evaluating the separate and combined effects of positive and negative reinforcement on task compliance. Journal of Applied Behavior Analysis 44, 175–179.

Cole, M., 2010. What's culture got to do with it? educational research as a necessarily interdisciplinary enterprise. Educational Researcher 39 (6), 461–470.

Craik, F.I.M., Lockhart, R.S., 1972. Levels of processing: a framework for memory research. Journal of Verbal Learning and Verbal Behavior 11, 671–684.

Daniels, H., 2001. Vygotsky and Pedagogy. Routledge/Falmer, London.

Daniels, H., 2008. Vygotsky and Research. Routledge, London.

Dewey, J., 1938. Experience and Education. Collier Books, New York.

Flavell, H., 1979. Metacognition and cognitive monitoring: a new era of cognitive developmental inquiry. American Psychologist 34, 906–911.

Gelman, S.A., Gottfried, G.M., 2006. Creativity in children's thoughts. In: Kaufman, J.C., Baer, J. (Eds.), Creativity and Reason in Cognitive Development. Cambridge University Press, New York, pp. 221–243.

Gershoff, E.T., 2002. Corporal punishment by parents and associated child behaviors and experiences: a meta-analytic and theoretical review. Psychological Bulletin 128 (4), 539–579.

Hattie, J., 2009. Visible Learning: A Synthesis of over 800 Meta-analyses Relating to Achievement. Routledge, New York.

Kuhn, D., 1992. Thinking as argument. Harvard Educational Review 62 (2), 155–178.

Lashley II, T.J., Matczynski, T.J., Rowley, J.B., 2002. Instructional Models: Strategies for Teaching in a Diverse Society, second ed. Wadsworth/Thomson Learning, Belmont, CA.

Legge, K., Harari, P., 2000. Psychology and Education. Heinemann, Oxford.
Marsh, H.W., Craven, R.G., McInerney, D. (Eds.), 2008. Self-Processes, Learning, and Enabling Human Potential: Dynamic New Approaches. Information Age Press, Greenwich, CT.
Pavlov, I.P., 1927. Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. Oxford University Press, London.
Phillipson, S.N., 2011a. Motivation and effective learning among Chinese learners. In: Phillipson, S.N., Lam, B.-h. (Eds.), Learning and Teaching in the Chinese Classroom: Responding to Individual Needs. Hong Kong University Press, Hong Kong, pp. 175–202.
Phillipson, S.N., 2011b. The social context of learning. In: Phillipson, S.N., Lam, B.-h. (Eds.), Learning and Teaching in the Chinese Classroom: Responding to Individual Needs. Hong Kong University Press, Hong Kong, pp. 231–263.
Piaget, J., 1970. The Science of Education and the Psychology of the Child. Orion Press, New York.
Rogoff, B., 1990. Apprenticeship in Thinking: Cognitive Development in Social Context. Oxford University Press, New York.
Schraw, G., Bendixen, L.D., Dunkle, M.E., 2002. Development and validation of the Epistemic Belief Inventory (EBI). In: Hofer, B.K., Pintrich, P.R. (Eds.), Personal; Epistemology: The Psychology of Beliefs about Knowledge and Knowing. Erlbaum, Mahwah, NJ, pp. 121–141.
Strenze, T., 2007. Intelligence and socioeconomic success. Intelligence 35, 401–426.
Vygotsky, L.S., 1960. Razvitie Vysshikh Psikhicheskikh Funktsii [The Development of Higher Mental Functions]. APN, Moscow.
Vygotsky, L.S., 1978. Mind in Society: The Development of Higher Mental Process. Harvard University Press, Cambridge, MA.
Wertsch, J.V., 1985. Vygotsky and the Social Formation of Mind. Harvard University Press, Cambridge, MA.
Wertsch, J.V., 2007. Mediation. In: Daniels, H., Cole, M., Wertsch, J.V. (Eds.), The Cambridge Companion to Vygotsky. Cambridge University Press, Cambridge, pp. 178–192.
Ziegler, A., 2005. The actiotope model of giftedness. In: Sternberg, R.J., Davidson, J.E. (Eds.), Conceptions of Giftedness. Cambridge University Press, New York, pp. 411–436.

Educational Media: Potentials for Learning

Deborah L Linebarger, N278 Lindquist Center, University of Iowa, Iowa City, IA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.E. Clark, volume 6, pp. 4279–4283, © 2001, Elsevier Ltd.

Abstract

Using mass forms of media for educational purposes is not a new idea. Mass media have been used for centuries to deliver educational content to children. From penny books, film projectors, radios, televisions, cassette recorders, and video machines to newer forms of media like computers, CD-ROMS and DVDs, the Internet, tablets (e.g., iPads), and mobile devices (e.g., smart phones), mass media have been introduced, enthused over by some advocates for their ability to reform education, and bemoaned by others who argue that mass media have led to declining educational and moral standards. This article reviews research, theories, and controversies concerning the use of mass media in educational settings, investigates a few well-known cases (e.g., Sesame Street), and summarizes current and emerging directions for further research.

Introduction

Using mass forms of media for educational purposes is not a new idea (Fisch, 2012; Wartella and Robb, 2008). Mass media have been used for centuries to deliver educational content to children. From penny books, film projectors, radios, televisions, cassette recorders, and video machines to newer forms of media like computers, CD-ROMS and DVDs, the Internet, tablets (e.g., iPads), and mobile devices (e.g., smart phones), mass media have been introduced, enthused over by some advocates for their ability to reform education, and bemoaned by others who argue that mass media have led to declining educational and moral standards. Marshal McLuhan wrote in 1964 that "Each new technology creates an environment that is itself regarded as corrupt and degrading. Yet the new one turns its predecessor into an art form" (p. ix). Media are a pervasive part of children's lives. Children under 8 years of age spend just over 3 h per day using media (Commonsense Media, 2011), while children between 8 and 18 years of age spend 7.5 h per day with media when multitasking is considered (Rideout et al., 2010). In recent years, there has been an explosion of media across multiple platforms.

Definitions of Educational Media

Educational media include media designed to intentionally or incidentally educate or enlighten (Kirkorian and Anderson, 2008). At present, there is no consensus or federally constructed definition of educational media; however, a definition for educational television that was developed as part of the creation of the Children's Television Act of 1990 (CTA) is a useful starting point. The CTA was created by the Congress in response to data indicating that children spend more than 3 h per day watching TV and that there is clear evidence that exposure to TV influences their development. The CTA states that "core programming [should be] specifically designed to serve the educational and informational needs of children 16 years of age and under, including the child's intellectual/cognitive or social/emotional needs" and must serve "the educational and informational needs of children as a significant purpose" (CTA, 1990). Advocates of the use of media as a potential learning tool argue that all media content is educational; that is, children will learn from whatever content with which they engage regardless of the format through which it is delivered (e.g., Skoll et al., 2006). For instance, there is considerable evidence that using or viewing violent media content (e.g., violent TV programs or video games) is associated with violent and aggressive behavior (e.g., Anderson and Bushman, 2001). Educational media, on the other hand, are typically created with the purpose to educate and inform. As such, research reviewed in this article must be specific to media that are "intentionally curriculum-driven with a deliberate plan to teach" (Kirkorian and Anderson, 2008: p. 188).

The CTA definition is purposely broad. It is much more elusive to define what characteristics make a product labeled 'educational' both high quality and effective. Jordan and Woodward (1997) defined the quality of children's educational television using six criteria created in consultation with experts in the field as well as after a review of research findings indicating that programs designed to be educational are generally beneficial to children. These criteria include:

1. Lesson clarity: how directly or explicitly the primary lesson is presented.
2. Lesson integration: the extent to which the primary lesson is repeated or incorporated throughout the program.
3. Lesson involvement: the extent to which the primary lesson is engaging and absorbing.
4. Lesson applicability: the extent to which the primary lesson is shown to be connected to the real world.
5. Lesson importance: the overall value or usefulness of the primary lesson.
6. Lesson reinforcement: the extent to which learning (including effort and mastery) of the lesson is rewarded within the episode.

Other advocacy organizations and researchers have adapted both the CTA language and these criteria for multiple media forms. For instance, Common Sense Media, an organization whose mission is to improve the lives of kids and families by providing trustworthy information about media to parents and educators, uses these definitions to evaluate content across TV,

video games, computer games, Internet sites, and mobile applications (see http://www.commonsensemedia.org/about-us/our-mission/about-our-ratings). Within a research context, Barr and Linebarger (2010) defined media as educational if the product materials made some type of educational claim. Educational claims were defined as visual or verbal content on packaging, product Web sites, and other product materials that suggested exposure to a particular product "could assist children in learning important information, skills, values, and behavior while entertaining them and exciting their curiosity to learn" (CTA, 1990: p. 303a; Fenstermacher et al., 2010: p. 561). This definition was used as a guide for the selection of materials that the creators of that content deemed educational whether or not evidence existed to support such claims. This was done purposely with the intent that these 'claims' would be further analyzed to evaluate whether the program content both matched the claims and matched the ways in which young children were known to learn best.

Models and Theories of Learning from Educational Media

The process of acquiring new information from media is a complex one involving attention to and subsequent comprehension of these stimuli. When children interact with media, they integrate the various stimuli into meaningful, comprehensible bits of information by attending to important or interesting aspects of the stimuli (Huston and Wright, 1989). Although a great deal of knowledge about young children has been gained from past research on media use individually, the obvious step of merging the ways in which media technologies are used for learning to understand and enhance cognitive development is limited. Models of learning developed from the television literature provide a useful framework for understanding how media might impact children's attention and learning through media.

Anderson and his colleagues (Anderson and Lorch, 1983; Anderson et al., 1981) proposed a model of attention to television where perceptually salient features of the stimulus initially draw a child's attention to the screen. When a child has little experience with television, sounds and unusual visual effects trigger the basic orienting mechanisms built into the perceptual system (Miron et al., 2001). These production techniques ('formal features') provide structure and give meaning to the sensory images contained in the programs (Calvert et al., 1982; Campbell et al., 1987). Comprehension is then improved when formal features are used to denote key moments and critical content for young children's focal attention.

Formal features constitute the means by which information is conveyed in a particular medium that, in turn, influences how that information is processed (Neuman, 1995). For television, the symbol system comprises perceptually salient events (i.e., sound effects, loud music, and rapid cuts), reflection (e.g., long zooms, singing, and moderate levels of physical activity), and character dialogue or speech (e.g., female or peculiar voices; Calvert et al., 1982; Huston and Wright, 1997). For computer and mobile gaming applications, the symbol system includes icons for repetitive functions (e.g., go back, save, help), navigational elements, music, animations during transitions from one state to another, drop-down menus, etc. For console video games, the symbol system includes the input device and its associated components (e.g., for XBox, the controller includes buttons labeled A, B, X, and Y), setup requirements (things that must be done once at the beginning of a game), the progression of play (what happens during game play including the leveling system), and resolution (what conditions cause the game to end, how an outcome is determined based on the game state).

Symbol systems are carriers of information; however, once the individual has received the perceptually salient and relevant stimuli, information will be processed similarly across media. Initially, children are more dependent on a medium's symbol system for decoding purposes. For example, children who are beginning to read spend enormous amounts of time learning the correspondence between printed letters and word recognition. Once children have learned to decode efficiently (when decoding has become automatic), they can read for content. Media operate in the same manner. Children must learn some rules about the forms to become efficient processors of its messages (Huston and Wright, 1983). Therefore, Salomon (1979) proposed that symbol systems might play a role initially, at lower levels of processing. As children become adept at using a medium's symbol system, the same kinds of higher-order cognitive skills used effectively in other media will then be used for deeper processing in their most recently mastered medium.

Huston and Wright (1989) and Rice et al. (1982) proposed a model of attention that elaborates this idea (see **Figure 1**). At various points while viewing a program, children make a series of attentional decisions. These decisions are based on cues of perceptual salience, comprehensibility of the program, and interpretability of the content. This model also predicts that attention is a joint function of the stimulus features, both form and content, and the dispositions of the viewer, including experience with the material, the medium, and general world knowledge. Interest and attention are a function of the stimulus characteristics. A stimulus perceived as "moderately novel, of intermediate complexity, integratable, somewhat regular, partially ordered, and recognizable" (Huston and Wright, 1989: p. 117) should elicit the greatest amounts of interest and attention. When the stimulus characteristics no longer arouse interest or attention because of familiarization and habituation (i.e., familiar, simple, wholistic, redundant, repetitive, and expected), children no longer attend to them. Likewise, if the stimulus characteristics fall at the incomprehensible end of these continua (i.e., novel, complex, incongruous, inconsistent, unpredictable, and surprising), children will not attend because the stimulus is immediately perceived as incomprehensible. High-end stimuli become more familiar and hence more comprehensible with repetition and thus gain attention while the low-end stimuli undergo further habituation with repetition and thus lose attention. The resultant of these two kinds of changes over time is that the high point of interest and attention slowly migrate toward the higher end. With age and viewing experience, the child's focus will continually move toward more cognitively challenging stimuli, hence the stimuli that are initially incomprehensible will "gradually move toward and through the child's focal lens of

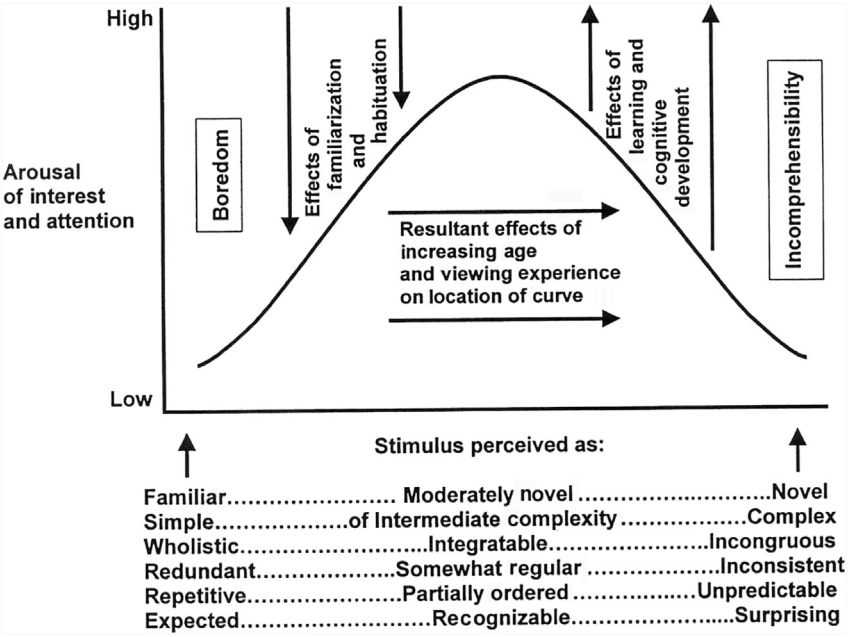

Figure 1 Traveling lens model. Reproduced from Rice, M.L., Huston, A.C., Wright, J.C., 1982. The forms and codes of television: effects of children's attention, comprehension, and social behavior. In: Pearl, D., Bouthilet, L., Lazar, L. (Eds.), Television and Behavior: Ten Years of Scientific Progress and Implications for the Eighties. US Government Printing Office, Washington, DC, pp. 24–38.

maximum interest, and then lose attention as they are habituated and become old hat" (Huston and Wright, 1989: p. 118).

In addition to these processing models of media, there are two media-specific theories of how children learn from media, both with roots in information-processing approasches: Fisch's Capacity Model (2000, 2004) and Mayer and Moreno's Cognitive Theory of Multimedia Learning (CTLM, 2003). Broadly, these two theories focus on the sequence and execution of cognitive events that occur when children are engaged with media. Children attend to environmental stimuli, encode these stimuli as information, store the new information in memory, and subsequently retrieve this information when needed.

Originally developed from the empirical base of studies examining learning from educational television, Fisch's Capacity Model (2000, 2004) proposes that the sequence and execution of cognitive events are influenced by two forms of content: narrative (i.e., storyline) and educational (i.e., intended lesson). Information-processing approaches argue that working memory resources are limited. For effective processing, cognitive demands on working memory cannot exceed available resources. Cognitive demands come from three components of the media content: processing of narrative content, processing of educational content embedded within the narrative, and the distance between these two content types. The distance reflects the degree to which the educational content is tangential to or integral to the narrative. When tangential, both content types compete for limited working memory resources. Because the narrative content is easier to process, children will default to processing the narrative at the expense of processing the educational content. Conversely, when the content types are integral to one another (i.e., the educational content is embedded within the story and helps to advance the storyline), competition for working memory resources is reduced and children are more effective at processing both content types and learning the educational content. This model predicts that children will comprehend educational content more completely when the cognitive demands of processing both content types are low and when the distance between the content types is small.

An added difficulty arises when considering learning beyond the educational content embedded within the program, or transfer of learning to situations outside the original context. Transfer requires that the viewer learns the educational content within the program, constructs a mental representation of that content, and then applies that content in a novel situation (Bransford et al., 1999). Educational content that is overly contextualized when initially presented can impede transfer (Bransford and Schwartz, 1999). The capacity model argues for contextualization of educational content within the narrative (Fisch, 2004), creating difficulty in determining the best presentation of the educational content within the narrative. Fisch (2004) proposed that media content must be contextualized when presented, but that it could also be presented multiple times using varied formats across different contexts (referred to as 'varied practice'; e.g., Piotrowski et al., 2013).

Mayer and Moreno's (2003) Cognitive Theory of Multimedia Learning (CTML) describes learning from media using three assumptions of cognitive science: (1) there are two separate channels (visual/pictorial and verbal/auditory) for processing information; (2) each channel has a limited capacity; and (3) learning is an active process that entails filtering, selecting, organizing, and integrating information using prior knowledge. Learning from multimedia is a demanding process that requires five cognitive processes: selecting relevant words,

selecting relevant images, organizing those words to create a coherent verbal model, organizing the images to create a coherent pictorial model, and then integrating the verbal and pictorial models together as well as with prior knowledge (Mayer, 2005). Layered within these cognitive processes are five different forms that words and pictures assume depending on the stage of processing: (1) the words and pictures actually presented in the stimuli; (2) the acoustic and iconic representations of these words and pictures that a child holds in sensory memory; (3) sounds and images that a child selectively attends to and then holds in working memory to be acted upon (i.e., beginning of knowledge construction); (4) a verbal model and a pictorial model that a child constructs in working memory (i.e., the child organizes these two models into coherent representations and integrates them together); and (5) knowledge in long-term memory (i.e., prior knowledge a child possesses that is accessed and used to make sense of the incoming stimuli). Consequently, CTML predicts that children learn content more deeply and effectively when words and pictures are combined than when one or the other alone conveys content.

Both the Capacity Model and the CTML have been used to examine the cognitive processing associated with the integration of incoming stimuli and the constraints that are associated with working memory, as this processing relates to educational television specifically (Capacity Model) or multimedia texts more broadly (CTML). While both approaches predict that integration of incoming stimuli can be difficult as the complexity of the stimuli increase, the nature of stimulus complexity varies (see Fisch, 2012). The Capacity Model predicts that learning challenges arise from difficulties with integration of different types of content. Specifically, when the educational content or curriculum embedded in a TV program is not well linked to the narrative or storyline in that program, learners will default to the easier of the two content types (i.e., the storyline) at the expense of learning the educational content. In contrast, learner challenges predicted by CTML arise from difficulties in processing multiple streams of information received across multiple modalities (i.e., words, pictures). Because learners have a limited working memory, they are able to process only a few chunks of information at any one time. As the complexity and volume of both words and pictures increase, the ability to adequately process, store, and later retrieve information is degraded.

Knowledge Base on Learning from Educational Media

From the accumulated evidence, high-quality educational media can be powerful resources for learning. Areas where beneficial effects have been found include general cognitive skills, school readiness, language acquisition, literacy, math, science, civics, social studies, general knowledge, prosocial behavior, executive functioning, problem solving, creativity, etc. (see Fisch, 2004; for a detailed review). Across these domains, researchers have determined that understanding educational media effects involves understanding independently and in combination the ways in which a child's individual abilities and prior knowledge and experiences, attributes of the stimulus, and the environmental and contextual features that are available during the processing of a specific media content (e.g., Bandura, 2001; Barr and Linebarger, 2010; Huston and Wright, 1989; Linebarger and Vaala, 2010). This orientation to media effects examines learning as a set of cognitive processes in an effort to elucidate the causal mechanisms through which effects are produced (Barr and Linebarger, 2010; Clark, 1983; Dupré and Cartwright, 1988; Hedstrom and Ylikoski, 2010; Kozma, 1994; M. Mackey and E. Ellerman, personal communication, 12 August 2013). Media can then be divided according to the type of cognitive activity associated with its use: viewing (TV, DVDs), listening (music, audio books), interacting (computers, video games, cell phones, and applications), producing (digital video/camera), integrating (comprehensive intervention, transmedia storytelling), and understanding (issues with technology use in the classroom; M. Mackey and E. Ellerman, personal communication, 12 August 2013).

Examples of Learning from Educational Media

Media-based interventions may be designed broadly (e.g., to provide caregivers or parents with resources to promote literacy) or narrowly (e.g., to promote a particular topic such as nutrition; Fisch, 2004). These interventions typically combine ancillary materials (e.g., manipulatives, books, scripts, posters, puppets, felt boards, the Internet, or mobile applications) and teacher guides with various TV episodes from an educational series. The program acts as an anchor for the intervention. Because children have reported confidence in learning from television (e.g., Linebarger, 2001), coupling existing educational television programs with ancillary materials offers opportunities for enhanced learning. In the three cases described below, learning has been increased with the addition of intervention strategies. Subtle differences in the development of the materials can make a large difference. For instance, when creating materials for an intervention designed around Between the Lions (BTL), research demonstrated stronger effects when the materials were developed in collaboration with the target audience using both developmentally appropriate and culturally relevant materials (Linebarger and Wainwright, 2007a) compared with the materials developed in-house and given to the target audience (Prince et al., 2002). Although learning effects were found in both interventions, greater effects were associated with the more specifically tailored intervention. In other nonmedia interventions, this tailoring to the target audience also leads to more enhanced effects (Barrera and González Castro, 2006).

Sesame Street

In the late 1960s, Sesame Street was developed to serve the educational and social needs of preschool children, especially those living in poverty (Mielke, 1994) with great success (e.g., Wright et al., 2001), both in the United States and all over the world (Cole et al., 2001). The program was a pioneer in the field of educational media for several reasons. First, it is one of the few programs that created and implemented a detailed and comprehensive set of curricular goals. Second, it used research to inform the development of content to improve production (formative research) as well as to test the overall effectiveness of

the program (summative research). Over time, this program has been successful in helping generations of preschool children learn all sorts of skills ranging from the alphabet and numbers, to helping and empathy, and even self-regulation and executive functioning (e.g., Fisch and Truglio, 2001; Wright et al., 2001). These early advantages have also been longitudinally linked to adolescent outcomes. Specifically, early viewers of Sesame Street obtained better grades in high school, read more books for fun, reported greater academic self-concepts, and detailed fewer aggressive attitudes when compared with teens who rarely or never viewed Sesame Street at 5 years of age (Anderson et al., 2001).

Interventions involving Sesame Street vary from instructing the child to pay attention to particular content prior to viewing (Sell et al., 1995) to using sophisticated curricula involving episodes, books, and extension activities (Yotive and Fisch, 2001), texting Sesame Street-themed parenting tips, audio messages from Elmo, and literacy-related Sesame Street videos to parents' cell phones (Horowitz et al., 2006). Sell et al., (1995) randomly assigned children to conditions where they either received instruction prior to viewing or repeated viewings, or both. Previewing instruction did not influence children's comprehension or program content whereas repeated viewings were successful at helping children comprehend the plot. The authors suggested that repeatedly viewing the videotape allowed the children to process the content more effectively.

Outreach efforts involving Sesame Street were initially conducted to increase the viewership of the program. Once its popularity was firmly established, outreach projects changed from viewership efforts to educational enhancement. It was hoped that the popularity of Sesame Street characters could be exploited to deliver educational messages to the target audience more effectively than the show on its own. The Preschool Education Program was composed of the TV series, children's books related to the series, and developmentally appropriate activities related to the series and books. Activities included extensions of viewing and book-related experiences (e.g., equivalent words in different languages, foods from different cultures). In addition to the various materials, there was also a teacher training component to help providers learn about the materials and to discuss key elements of the program with other professionals. In addition, newsletters containing information regarding upcoming episodes were distributed on a regular basis to provide opportunities for providers to integrate the intervention into their daily activities. Limited research on the intervention provided evidence that it had an impact on both children and their providers (Yotive and Fisch, 2001). Specifically, the children's social behavior, motivation for learning, verbal skills, and early literacy skills improved from pre- to posttest. Providers reported more self-confidence and improved instructional skills, especially skills related to joint book-reading episodes. Growth in skills was greater for providers and children in home-based care situations when compared with center-based child care.

In recent years, Sesame Street has begun to investigate the use of new technologies to deliver content that targets both parents and their preschoolers. Parents reported that after receiving the tips and materials via their cell phones, they felt more competent at scaffolding their children's learning as well as greater enthusiasm for using such content (Horowitz et al., 2006). Sesame Street has always been on the leading edge of using media to promote young children's academic and social learning. Advances in mobile technologies are now making it possible to provide educational opportunities for learning and discovery in ways that capitalize on the integral nature of such technologies in people's lives (Shuler, 2009).

Between the Lions

In an initial examination of BTL, kindergarten children who watched the program performed significantly better on almost all outcome measures of reading achievement when compared to similar children who did not (Linebarger et al., 2004). Specifically, the kindergarten children who watched 17 episodes from the first season of BTL were able to identify more letters and read more words than those who did not. Their understanding of concepts of print was enhanced and their phonemic awareness and letter–sound correspondence skills, skills highly predictive of latter, fluent reading, were significantly accelerated over those who did not watch. BTL was also the focus of a study conducted by researchers in Mississippi (Prince et al., 2002). More than 50 teachers and nearly 1000 preschool, kindergarten, and first grade students from the Choctaw Indian Reservation and the Delta region of the state (i.e., children who were low income and/or English Language Learners (ELL) students) were assigned to either a treatment group (i.e., received videotapes and ancillary materials) or a control group (i.e., did not receive the materials; however, they might see the program on their local Public Broadcasting Service (PBS) station). This investigation differed from the initial evaluation in two important ways. First, the evaluation period lasted 9 months (vs just over 1 month in Kansas). In addition, this project combined extensive ancillary intervention materials, training, and follow-up (e.g., teachers' guides, handbooks, posters, a CD-ROM with content from the Web site, sets of classroom books and take-home books stickers, bookmarks, and information for parents) with videotapes of the episodes.

Researchers found children who watched half-hour episodes of BTL regularly, and whose teachers carried out related activities, significantly outperformed control groups on several key reading skills including knowledge of print conventions, identification of initial word sounds, ability to read nonsense words, and the ability to identify spoken vocabulary words (Prince et al., 2002). Important questions remained regarding its applicability for multiple subgroups of children in different contextual situations. Several implementation issues arose in this Mississippi project. For instance, although logs were kept, it is unclear how well these correlated with student outcomes as fidelity of implementation was not examined. Investigators did not systematically collect information regarding teachers' perceptions of the intervention materials or satisfaction with and usability of the materials. Finally, it is possible that training related to the delivery of the intervention was ineffective with some of the educators, especially the poorly trained child care providers. Considerable variation in training and education levels of teachers may have impacted teachers' ability to deliver an effective intervention. Again, these data were not formally collected.

Lessons learned from the first Mississippi study were resolved in several follow-up intervention studies. The original intervention materials were made primarily by BTL content creator. This posed problems when the materials were taken into the field that might have been avoided had the end users been more fully consulted during the process. BTL project staff worked closely with the local cosmmunities to develop more culturally sensitive and appropriate literacy materials. Teacher training was provided along with technical assistance and mentoring throughout the interventions. Children's skills improved in direct relation to the amount of intervention experienced for oral language knowledge, letter names, phonemic awareness (i.e., isolating sounds in words and blending words), and general literacy abilities (Linebarger and Wainwright, 2007; Linebarger and McMenamin, 2010).

The Ready to Learn Initiative

In recent research funded by the US Department of Education, researchers investigated the creation of a 'media-rich curriculum' comprised of multiple media properties and platforms (Penuel et al., in press). Researchers devised a 10-week early literacy curriculum that featured episodes of the preschool literacy series Super Why, smaller self-contained segments from both Sesame Street and BTL, books related to all program content, games created for the PBS Kids Web site, and teacher-led viewing and postviewing activities (e.g., pausing the video, reflecting on specific segments after viewing). Teachers received training, supplemental materials, and ongoing mentoring prior to and during implementation of the intervention. Preschoolers' knowledge of letter names, phonemic awareness, story and print concepts, and specific letters in their own names were significantly better for those in the media-rich curriculum compared with preschoolers in a science-oriented control curriculum (Penuel et al., 2009).

Changes in Focus/Emphasis over Time

With the introduction of most forms of mass media, there have been advocates for the power and reach of the new medium to cynics who argue that this new medium will promote intellectual passivity or social problems (Wartella and Robb, 2008). Most studies that investigate new forms of media examine how exposure alone without any differentiation by the content of that exposure affects developmental outcomes. This orientation resulted in a preponderance of research investigating the negative effects of media (e.g., aggression, obesity). Focusing on the media delivery system or platform (e.g., TV, educational application, book) leads to simple cause and effect explanations. In contrast, examining cognitive processing, content, the structural features used to convey that content, and the contexts surrounding use of media offer stronger explanatory models for observed effects (Anderson et al., 2001; Barr and Linebarger, 2010; Clark, 1983; Huston and Wright, 1983; Kozma, 1994; Metzger, 2009). Over time, there has been a greater emphasis on understanding which types of content support positive developmental outcomes and which types of content exacerbate negative developmental outcomes and whether and how individual differences coupled with the ways in which media are used mediate or moderate the relations between exposure and outcomes (e.g., Jordan, 2004; Linebarger and Vaala, 2010). As such, researchers are better able to describe the most effective ways to deliver content, who is more or less likely to learn this content and why, and whether the presence of certain features in the environment support or diminish learning.

Current and Emerging Directions in Theory and Research

Transmedia

Transmedia refers to 'across media.' It describes any number of possible relationships that might exist between and among the various media that constitute a contemporary entertainment franchise. Kinder (1991) referred to transmedia as the 'entertainment supersystem' that emerged around various TV-based media properties (e.g., Teenage Mutant Ninja Turtles). Transmedia storytelling extended the use of characters and properties to include the "systematic unfolding of elements of a story world across multiple media platforms, with each platform making a unique and original contribution to the experience as a whole" (Herr-Stephenson et al., 2013: p. 6). Content and characters are repeated and expanded across platforms. As described by Fisch (2004), these types of learning opportunities promote varied practice with specific content thereby increasing both content-specific learning as well as transfer of that learning to other situations or problems. Proponents for transmedia learning argue that content delivered via multiple platforms promotes additive comprehension because children attend to and learn content as they follow or track the same story across these different platforms (Herr-Stephenson et al., 2013; Jenkins, 2006). Transmedia differentiates itself from cross-media (use of other media platforms to deliver the same content) and multimedia (which uses multiple kinds of media like words, pictures, sounds, videos in a single package) in that the user is incited to search out dispersed content and reassemble it into a meaningful mental model. It requires a creative reworking of stories and experiences as children reimagine, play, and learn from the media culture surrounding them (Herr-Stephenson et al., 2013). Empirical evidence testing the effectiveness of a transmedia approach is limited. Fisch and colleagues reported that children's use of mathematical learning strategies after viewing and using online games linked to the children's program Cyberchase was more sophisticated and their use of problem-solving strategies was more consistent when compared with children who engaged only in gaming (Fisch et al., 2011). The authors argued that exposure to the characters in multiple modalities facilitated both content and character connections across the two platforms (Fisch et al., 2011).

A Focus on Neuroscience

Anderson and colleagues (Anderson, 2007; Anderson et al., 2006a,b) proposed that technological advances in methodologies like eye tracking and magnetic resonance imaging should be considered as potentials for investigating questions related to the study of children and media. Researchers have used eye-tracking techniques to investigate how children process

onscreen text (Scinto, 2001; Vaala et al., 2009); whether and how infants understand the use of audiovisual production techniques like cuts (Kirkorian et al., 2012); and as a means to understanding basic processing of program content (Flagg, 1978; Chernin and Linebarger, 2005). Magnetic resonance imaging has been used to investigate both children and adult's attentional and comprehension capabilities during online processing of media content (Anderson et al., 2006b) as well as their emotional responses to violent media content (Mathews et al., 2005; Murray et al., 2006; Weber et al., 2006). Anderson (2007) describes the inherent difficulties in using neuroscience techniques to investigate media use and media effects but also argues that the use of such techniques can explicate the influence of media on brain structure and function; promote new modes of thinking about the problems that media effects researchers face; and provide new perspectives on the influence of media on children.

Methodological Problems

Measuring Media Exposure

Media use is pervasive among children. Capturing both the quantity and quality of exposure to media is a fundamental yet challenging measurement problem. Without adequate media exposure measures, efforts to tie the amount of time exposed to different media forms and content types to various outcomes will be difficult (Jordan et al., 2007). Exposure data have been obtained via a number of assessment strategies including interviews with children (if old enough) or their parents and caregivers (e.g., Jordan et al., 2007), diaries (e.g., Lapierre et al., 2012), neurological and psychophysiological measures during exposure (e.g., Anderson et al., 2001; Anderson, 2007), experience sampling methods (e.g., Moreno et al., 2012), electronic monitoring systems (e.g., Christakis et al., 2009), and video or direct observation (Anderson and Field, 1991). There is limited but increasing research that systematically and empirically compares different methods of capturing exposure (e.g., Anderson and Field, 1991; Jordan et al., 2007). Researchers who use exposure assessments across a number of disciplines (e.g., media, environmental toxins, dietary intake, time use) advocate the use of different collection methods to validate data and reduce error in measurement (dos Santos Silva, 1999; Fave et al., 2009; Wartella et al., 2009; Vandewater and Lee, 2009) and suggest that researchers carefully consider how much detail in exposure is needed as well as the kinds of data needed to answer specific research questions (Vandewater and Lee, 2009).

Links to Databases and Software

Center on Media and Child Health

The Center on Media and Child Health (CMCH) at Children's Hospital Boston, Harvard Medical School, and Harvard School of Public Health is dedicated to understanding and responding to the effects of media on the physical, mental, and social health of children through research, translation, and education. CMCH created a searchable database to access scientific research on media effects. Available at: http://www.cmch.tv/research/searchcitations.asp.

Common Sense Media

An independent nonprofit organization dedicated to improving the lives of kids and families by providing trustworthy information needed to thrive in a world of media and technology. Available at http://www.commonsensemedia.org/.

Federal Communications Commission/CTA

Guide to the CTA and what is considered educational on TV. Available at http://www.fcc.gov/guides/childrens-educational-television.

Joan Ganz Cooney Center

The Joan Ganz Cooney Center is an independent research and innovation lab that focuses on the challenges of educating children in a rapidly changing media landscape. Available at http://www.joanganzcooneycenter.org.

PBS Parents

PBS Parents is a parent-directed resource containing information on child development, early learning, and educational media. Available at http://www.pbs.org/parents/.

See also: Learning Theories and Educational Paradigms; Lifelong Learning and Its Support with New Technologies; Mass Media, Representations In; Mass Media: Introduction and Schools of Thought; Media and Child Development; New Media and Social Capital; Public Broadcasting; Social Media; Television: General; Visual Images in the Media.

Bibliography

Anderson, D.R., 2007. A neuroscience of children and media? Journal of Children and Media 1 (1), 77–85.

Anderson, C.A., Bushman, B.J., 2001. Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: a meta-analytic review of the scientific literature. Psychological Science 12, 353–359.

Anderson, D.R., Field, D.E., 1991. Online and offline assessment of the television audience. In: Bryant, J., Zillmann, D. (Eds.), Responding to the Screen: Reception and Reaction Processes. Erlbaum, Hillsdale, NJ, pp. 199–216.

Anderson, D.R., Lorch, E.P., 1983. Looking at television: action or reaction? In: Bryant, J., Anderson, D.R. (Eds.), Children's Understanding of Television: Research on Attention and Comprehension. Academic Press, New York, pp. 1–33.

Anderson, D.R., Lorch, E.P., Field, D.E., Sanders, J., 1981. The effects of TV program comprehensibility on preschool children's visual attention to television. Child Development 52, 151–157.

Anderson, D.R., Huston, A.C., Schmitt, K., Linebarger, D.L., Wright, J.C., 2001. Early childhood television viewing and adolescent behavior: the recontact sudy. Monographs of the Society for Research in Child Development 66, 1, Serial No. 264.

Anderson, D.R., Bryant, J., Murray, J.P., Rich, M., Rivkin, M.J., Zillmann, D., 2006a. Brain imaging: an introduction to a new approach to studying media processes and effects. Media Psychology 8 (1), 1–6.

Anderson, D.R., Fite, K.V., Petrovich, N., Hirsch, J., 2006b. Cortical activation while watching video montage: an fMRI study. Media Psychology 8 (1), 7–24.

Bandura, A., 2001. Social cognitive theory of mass communication. Media Psychology 3, 265–298.

Barr, R., Linebarger, D.L., 2010. Introduction to the special issue: content and context in infant-directed media. Infant and Child Development 19 (6), 553–556.

Barrera Jr., M., González Castro, F., 2006. A heuristic framework for the cultural adaptation of interventions. Clinical Psychology: Science and Practice 13 (4), 311–316.

Bransford, J.D., Schwartz, D.L., 1999. Rethinking transfer: a simple proposal with multiple implications. In: Iran-Nejad, A., Pearson, P.D. (Eds.), Review of Research in Education, vol. 24. American Educational Research Association, Washington, DC, pp. 61–100.

Bransford, J., Brown, A., Cocking, R., 1999. How People Learn: Brain, Mind, Experience, and School. National Academy of Science, Washington, DC.

Calvert, S.L., Huston, A.C., Watkins, B.A., Wright, J.C., 1982. The relation between selective attention to television forms and children's comprehension of content. Child Development 53, 601–610.

Campbell, T.A., Wright, J.C., Huston, A.C., 1987. Form cues and content difficulty as determinants of children's cognitive processing of televised educational messages. Journal of Experimental Child Psychology 43, 311–327.

Chernin, A.R., Linebarger, D.L., April 2005. The Effects of Active Mediation on Children's Attention to and Comprehension of Television: An Eye-tracking Study (Poster session presented at the biennial meeting of the Society for Research in Child Development, Atlanta, GA).

Children's Television Act of 1990. 1990. Publ. L. No. 101-437, 104 Stat. 996-1000, codified at 47 USC Sections 303a, 303b, 394.

Christakis, D.A., Gilkerson, J., Richards, J.A., Zimmerman, F.J., Garrison, M.M., Xu, D., Gray, S., Yapanel, U., 2009. Audible television and decreased adult words, infant vocalizations, and conversational turns: a population-based study. Archives of Pediatric and Adolescent Medicine 163 (6), 554–558.

Clark, R.E., 1983. Reconsidering research on learning from media. Review of Educational Research 53, 445–459.

Cole, C.F., Richman, B.A., McCann-Brown, S.A., 2001. The world of Sesame street research. In: Fisch, S.M., Truglio, R.T. (Eds.), "G" Is for Growing: Thirty Years of Research on Children and Sesame Street. Lawrence Erlbaum, Mahwah, NJ, pp. 147–179.

Common Sense Media, 2011. Zero to Eight: Children's Media Use in America. A Common Sense media research study. Common Sense Media, San Francisco, CA.

Dupré, J., Cartwright, N., 1988. Probability and causality: Why Hume and indeterminism don't mix. Nôus 22, 521–536.

Fave, G., Beckmann, M.E., Draper, J.H., Mathers, J.C., 2009. Measurement of dietary exposure: a challenging problem which may be overcome thanks to metabolomics. Genes & Nutrition 4 (2), 135–141.

Fenstermacher, S.K., Barr, R., Garcia, A., Salerno, K., Shwery, C., Pempek, T.A., Calvert, S., Linebarger, D.L., 2010. Infant-directed media: an analysis of product information and claims. Infant and Child Development 19 (6), 557–576.

Fisch, S.M., 2000. A capacity model of children's comprehension of educational content on television. Media Psychology 2, 63–91.

Fisch, S.M., 2004. Children's Learning from Educational Television: Sesame Street and beyond. Erlbaum Publishers, Mahwah, NJ.

Fisch, S.M., Lesh, R., Motoki, E., Crespo, S., Melfi, V., 2011. Children's mathematical reasoning in online games: Can data mining reveal strategic thinking? Child Development Perspectives 5, 88–92.

Fisch, S.M., 2012. Learning from educational television. In: Lemish, D. (Ed.), The Routledge International Handbook of Children, Adolescents, and Media. Taylor & Francis, London, pp. 403–409.

Fisch, S.M., Truglio, R.T., 2001. "G" Is for Growing: Thirty Years of Research on Children and Sesame Street. Erlbaum Publishers, Mahwah, NJ.

Flagg, B.N., 1978. Children and television: Effects of stimulus repetition on eye activity. In: Senders, J.W., Fisher, D.F., Monty, R.A. (Eds.), Eye movements and the higher psychological functions. Erlbaum, Hillsdale, pp. 279–291.

Hedstrom, P., Ylikoski, P., 2010. Causal mechanisms in the social sciences. Annual Review of Sociology 36, 49–67.

Herr-Stephenson, B., Alper, M., Reilly, E., Jenkins, H., 2013. Transmedia: Learning through Transmedia Play. Joan Ganz Cooney Center, New York.

Horowitz, J.E., Sosenko, L.D., Hoffman, J.L.S., Ziobrowski, J., Tafoya, A., Haagenson, A., Hahn, S., 2006. Evaluation of the PBS Ready to Learn Cell Phone Study: Learning Letters with Elmo (A report prepared for the PBS Ready to Learn Grant).

Huston, A.C., Wright, J.C., 1983. Children's processing of television: the informative functions of formal features. In: Bryant, J., Anderson, D.R. (Eds.), Children's Understanding of Television: Research on Attention and Comprehension. Academic Press, New York, pp. 35–68.

Huston, A.C., Wright, J.C., 1989. The Forms of Television and the Child Viewer, vol. 2. Academic Press, New York.

Huston, A.C., Wright, J.C., 1997. Mass media and children's development. In: Damon, W. (Series Ed.), Sigel, I., Renninger, K.A. (Eds.), Child Psychology in Practice. Handbook of Child Psychology, fifth ed. Wiley, New York, pp. 999–1058.

Jenkins, H., 2006. Convergence Culture: Where Old and New Media Collide. New York University Press, New York.

Jordan, A.B., 2004. The role of media in children's development: an ecological perspective. Journal of Developmental and Behavioral Pediatrics 25 (3), 196–206.

Jordan, A.B., Woodward, E., 1997. The 1997 State of Children's Television Report: Programming for Children over Broadcast and Cable Television. Rep. No. 14. The Annenberg Public Policy Center of the University of Pennsylvania, Philadelphia, PA.

Jordan, A.B., Trentacoste, N., Henderson, V., Manganello, J., Fishbein, M., 2007. Measuring the time teens spend with media: challenges and opportunities. Media Psychology 9 (1), 19–42.

Kinder, M., 1991. Playing with Power in Movies, Television and Video Games: From Muppet Babies to Teenage Mutant Ninja Turtles. University of California Press, London.

Kirkorian, H.L., Anderson, D.R., 2008. Learning from educational media. In: Calvert, S.L., Wilson, B.J. (Eds.), The Blackwell Handbook of Children, Media, and Development. Blackwell, Boston, MA, pp. 319–360.

Kirkorian, H.L., Anderson, D.R., Keen, R., 2012. Age differences in online processing of video: an eye movement study. Child Development 83, 497–507.

Kozma, R., 1994. Will media influence learning: reframing the debate. Educational Technology Research and Development 42, 7–19.

Lapierre, M., Piotrowski, J.T., Linebarger, D.L., 2012. Background television in the homes of US children. Pediatrics 130 (5), 1–8.

Linebarger, D.L., 2001. Learning to read from television: the effects of using captions and narration. Journal of Educational Psychology 93 (2), 288–298.

Linebarger, D.L., McMenamin, K., 2010. Evaluation of the between the Lions Mississippi Literacy Initiative, 2008–2009. Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA.

Linebarger, D.L., Vaala, S., 2010. Infants and toddlers, screen media, and language development: an ecological perspective. Developmental Review 30 (2), 176–202.

Linebarger, D.L., Wainwright, D.K., April 2007a. American Indian children, their home media and literacy environments, and their literacy abilities in relation to participation in a between the Lions literacy intervention. In: Barr, R. (Ed.), The Context of Media Viewing: The Influence of Television on Infants and Preschoolers (Paper symposium to be presented at the biennial meeting of the Society for Research in Child Development, Boston, MA).

Linebarger, D.L., Wainwright, D.K., 2007b. Learning while viewing: urban myth or dream come true? In: Mazzarella, S.R. (Ed.), 20 Questions about Youth and the Media. Peter Lang, New York, pp. 179–196.

Linebarger, D.L., Kosanic, A., Greenwood, C.R., Doku, N.S., 2004. Effects of viewing the television program between the Lions on the emergent literacy skills of young children. Journal of Educational Psychology 96 (2), 297–308.

Mathews, V.P., Kronenbrug, W.G., Wang, Y., Lurito, J.T., Lowe, M.J., Dunn, D., 2005. Media violence exposure and front lobe activation measured by functional magnetic resonance imaging in aggressive and nonaggressive adolescents. Journal of Computer Assisted Tomography 29, 287–292.

Mayer, R.E., 2005. Cognitive theory of multimedia learning. In: Mayer, R. (Ed.), The Cambridge Handbook of Multimedia Learning. Cambridge University Press, Cambridge, UK, pp. 31–48.

Mayer, R.E., Moreno, R., 2003. Nine ways to reduce cognitive load in multimedia learning. Educational Psychologist 38 (1), 43–52.

McLuhan, M., 1964. Understanding Media. Routledge, London.

Metzger, M.J., 2009. The study of media effects in the era of internet communication. In: Nabi, R., Oliver, M.B. (Eds.), Handbook of Media Effects. Sage, Thousand Oaks, CA, pp. 561–576.

Mielke, K., 1994. Sesame street and children in poverty. Media Studies Journal 8, 125–134.

Miron, D., Bryant, J., Zillman, D., 2001. Creating vigilance for better learning from television. In: Singer, D.G., Singer, J.L. (Eds.), Handbook of Children and the Media. Sage, Thousand Oaks, CA, pp. 153–181.

Moreno, M.A., Jelenchick, L., Koff, R., Eikoff, J., Diermyer, C., Christakis, D.A., 2012. Internet use and multitasking among older adolescents: an experience sampling approach. Computers in Human Behavior 28 (4), 1097–1102.

Murray, J.P., Liotti, M., Ingmundson, P.T., Mayberg, H.S., Pu, Y., Zamarripa, F., et al., 2006. Children's brain activations while viewing televised violence revealed by fMRI. Media Psychology 8, 25–38.

NAEYC & Fred Rogers Center for Early Learning and Children's Media, 2012. Technology and Interactive Media as Tools in Early Childhood Programs Serving Children from Birth through Age 8. Joint position statement. NAEYC, Washington, DC. Fred Rogers Center at Saint Vincent College, Latrobe, PA.

Neuman, S., 1995. Literacy in the Television Age: The Myth of the TV Effect, second ed. Ablex Publishing Corporation, Norwood, NJ.

Penuel, W.R., Pasnik, S., Bates, L., Townsend, E., Gallagher, L.P., Llorente, C., et al., 2009. Preschool teachers can use a media-rich curriculum to prepare low-income children for school success: results of a randomized controlled trial. Education Development Center, Inc. SRI International. Menlo Park, CA, New York, NY.

Penuel, W.R., Bates, L., Gallagher, L.P., Pasnik, S., Llorente, C., Townsend, E., Domínguez, X., VanderBorght, M., 2012. Supplementing literacy instruction with a media-rich intervention: results of a randomized controlled trial. Early Childhood Research Quarterly 27 (1), 115–127.

Piotrowski, J.T., Jennings, N.A., Linebarger, D.L., 2013. Extending the lessons of educational television with young children. Journal of Children and Media 7 (2), 216–234.

Prince, D.L., Grace, C., Linebarger, D.L., Atkinson, R., Huffman, J.D., 2002. Between the Lions Mississippi Literacy Initiative: A Final Report to Mississippi Educational Television. Report prepared for Mississippi Educational Television and WGBH Educational Foundation. The Early Childhood Institute, Mississippi State University, Starkville, MS.

Rice, M.L., Huston, A.C., Wright, J.C., 1982. The forms and codes of television: effects of children's attention, comprehension, and social behavior. In: Pearl, D., Bouthilet, L., Lazar, J. (Eds.), Television and Behavior: Ten Years of Scientific Progress and Implications for the Eighties. US Government Printing Office, Washington, DC, pp. 24–38.

Rideout, V., Foehr, U.G., Roberts, D.F., 2010. Generation M^2: Media in the Lives of 8–18 Year Olds. Kaiser Family Foundation, Menlo Park, CA.

dos Santos Silva, I., 1999. Cancer epidemiology: Principles and Methods. International Agency for Research on Cancer, World Health Organization, Lyon, France.

Salomon, G., 1979. Interaction of Media, Cognition, and Learning. Jossey-Bass, San Francisco, CA.

Scinto, L.F.M., 2001. Children's Allocation of Attention and Viewing Responses to between the Lions as Measured by Eye Movements. A report prepared for WGBH Educational Foundation. Brigham & Women's Hospital & Harvard Medical School, Boston, MA.

Sell, M.A., Ray, G.E., Lovelace, L., 1995. Preschool children's comprehension of a Sesame street video tape: the effects of repeated viewing and previewing instructions. Educational Technology Research and Development 43, 49–60.

Shuler, C., 2009. Pockets of Potential: Using Mobile Technologies to Promote Children's Learning. Joan Ganz Cooney Center, New York.

Skoll, J. (Producer), Goldstein-Knowlton, L. (Producer/Director), Hawkins Costigan, L. (Producer/Director), 2006. The world according to Sesame Street: a global documentary of local proportions [Documentary]. Las Lindas, United States.

Vaala, S.E., Lapierre, M.A., Linebarger, D.L., May 2009. Eye Movements of Children and Adults Reading Onscreen Print in Educational Programming (Poster presented at the annual convention of the Association for Psychological Science, San Francisco, CA).

Vandewater, E.A., Lee, S.-J., 2009. Measuring children's media use in the digital age: issues and challenges. American Behavioral Scientist 52 (8), 1152–1176. http://dx.doi.org/10.1177/0002764209331539.

Wartella, E., Robb, M., 2008. Historical and recurring concerns about children's use of the mass media. In: Calvert, S.L., Wilson, B.J. (Eds.), The Handbook of Children, Media, and Development. Blackwell, Oxford, UK, pp. 7–26.

Wartella, E., Huston, A.C., Rideout, V., Robb, M., 2009. Studying media effects on children: improving methods and measures. American Behavioral Scientist 52 (8), 1111–1114.

Weber, R., Ritterfeld, U., Mathiak, K., 2006. Does playing violent video games induce aggression? Empirical evidence of a functional magnetic resonance imaging study. Media Psychology 8, 39–60.

Wright, J.C., Huston, A.C., Murphy, K.C., St Peters, M., Pinon, M., Scantlin, R., Kotler, J., 2001. The relations of early television viewing to school readiness and vocabulary of children from low-income families: the early window project. Child Development 72 (5), 1347–1366.

Yotive, W., Fisch, S.M., 2001. The role of Sesame street-based materials in child-care settings. In: Fisch, S.M., Truglio, R.T. (Eds.), "G" Is for Growing: Thirty Years of Research on Children and Sesame Street. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 181–196.

Educational Philosophy

Gert Biesta, University of Luxembourg, Walferdange, Luxembourg

© 2015 Elsevier Ltd. All rights reserved.

Abstract

While philosophy of education is in one sense as old as philosophy itself, it emerged as a particular specialism of educational and philosophical scholarship in the English-speaking world in the twentieth century. This article discusses the philosophical engagement with education through time, highlighting issues concerning the definition of its two key terms – philosophy and education – discussing some classics in the philosophical literature on education, in order to arrive at a sketch of the development of philosophy of education in the twentieth century.

Introduction

In her widely used book *Philosophy of Education*, the American philosopher of education Nel Noddings states that philosophy of education "is the philosophical study of education and its problems" (Noddings, 2012: xiii). "Philosophers of education," she writes, "study the problems of education from a philosophical perspective" (Noddings, 2012: xiv). She therefore concludes that the "subject matter [of philosophy of education] is *education*, and its methods are those of *philosophy*" (Noddings, 2012: xiii; emphasis added). While many would agree with this definition – and in a sense philosophy of education can be nothing but the philosophical study of education – it nonetheless raises a number of further questions. One question is what philosophy actually *is*, which has been an ongoing topic of debate ever since philosophy came into existence as a distinctive intellectual 'genre.' However, for the philosophy of education there is not only the question what philosophy is, but there is also the question of what education is or might be – and here again we find a wide array of answers, not in the least from those who see themselves as philosophers of education. Rather, therefore, than seeing philosophy of education as the philosophical study of education, it is more productive to see philosophy of education as the encounter between philosophical and educational thinking, and it is this line that I will pursue in what follows.

What 'Is' Philosophy?

To think of philosophy of education as an encounter between philosophy and education makes it necessary to say something about the meaning of both concepts, that of 'philosophy' and that of 'education.' Our contemporary – and as some might emphasize: Western idea of philosophy emerged in Greek civilization between about 600 and 150 BC. While the Greek word 'philosophy' originally had a rather broad meaning, referring to a thirst for learning or a thirst for education, over time it came to describe the more specific tradition of thinking and writing that we now recognize as philosophy. One common definition of philosophy – a literal translation of the word – is that of 'love of wisdom.' In the *Phaedrus*, a dialogue written by Plato (427–347 BC), we can find Socrates (439–399 BC) arguing for a distinction between, on the one hand, being wise and having knowledge of the truth, and, on the other hand, *aspiring to* wisdom, knowledge, and truth. Socrates then argues that only a god can be called wise in the former sense, so that human beings who aspire for wisdom, knowledge, and truth should more fittingly be called *lovers* of wisdom, that is, philosophers. This line of thought already hints at a prominent characteristic of contemporary philosophy, in that philosophers nowadays would seldom lay claim to having knowledge – in this regard the domain of philosophy and the domain of science have over time become clearly separated areas of intellectual activity – but rather see it as their task to raise critical questions about everyday and scientific beliefs and arguments and their underlying assumptions.

Behind the simple definition of philosophy of education as "the philosophical study of education and its problems" (Noddings) we can find a whole universe of philosophical ideas and traditions and of ways and styles of doing philosophy, which makes it quite difficult if not impossible to reduce this variety to some kind of 'essence' of what philosophy is or is about – which is one of the reasons why the American philosopher Richard Rorty has suggested that we should better think of philosophy as a literary genre or 'a kind of writing,' as he has put it, held together not by form or matter but by tradition (see Rorty, 1978). While few, for example, would challenge the suggestion that philosophy is about asking critical questions, it makes all the difference whether in asking such questions one takes inspiration from, say, Socrates, Kant, Marx, or Derrida, to name but a few of the critical philosophers. (On the particular issue of what it means to be critical see Biesta and Stams, 2001.) While few would challenge the suggestion that philosophy is about the exploration and exposure of underlying assumptions, it makes all the difference whether in doing so one takes inspiration from, say, Plato, Hume, Hegel, or Foucault. And while few would challenge the idea that philosophers have important things to say about knowledge, belief, and truth, it again makes all the difference whether one does so from the perspective of, say, Aristotle, Descartes, Locke, or Dewey. Philosophy thus presents us with a vast arsenal of ideas, approaches, traditions, and ways of thinking that cover a wide range of areas, including logic, epistemology, ethics, esthetics, metaphysics, anthropology, social and political philosophy, philosophy of language, and philosophy of science. In this way philosophy provides much more than

a huge resource for the study of education and its problems. Many of the topics that have been explored by philosophers throughout the centuries are actually highly relevant for the study of education (see below).

What 'Is' Education?

The word education has Latin roots (a combination of 'ex' and 'ducere') and literally means 'to lead out.' Some have taken this to argue that education not only *is* a process that leads people out of their current beliefs and ways of doing, but also that education *ought to* lead people 'beyond the present and the particular' (Bailey, 1984). When we look at the word 'education' in more detail, we can distinguish at least five different ways in which it is being used. First, it is used to refer to *institutions* such as schools, colleges, and universities. Hence we talk about primary, secondary, further, continuing, and higher education. Secondly, partly related to this, 'education' is used to refer to *domains of practice*, such as adult education, school education, vocational education, technical education, or university education. Thirdly, the word is used to refer to the *processes* that go on in such institutions and practices. Fourthly, the word is used to refer to the *activity of educating*. Teaching is one obvious example of this use of the word, and parenting another. Fifthly, the word is also used to refer to the *outcomes* of education, such as in the idea of the educated person, a topic extensively discussed in the philosophy of education literature (see, e.g., Peters, 1970; Martin, 1981).

While the English word 'education' thus contains multiple meanings, other languages are sometimes slightly more helpful in that they use different words for these different meanings (see also Biesta, 2011). One such distinction can be found in the German language, where the word 'Bildung' is used to refer to processes of education whereas the word 'Erziehung' is more often used to refer to the activity of educating. 'Bildung' (sometimes translated into English as 'formation' or 'edification') refers to the ways in which, through engagement with culture and tradition, we develop our human capacities. It thus comes with an idea of education as cultivation and enculturation and, more importantly, with an explicit idea of what it means to be educated. This obviously means that not *any* engagement with culture and tradition necessarily contributes to the 'Bildung' of a human person, so that one important question with regard to 'Bildung' concerns the selection of those aspects of culture and tradition that are deemed to be beneficial for one's 'Bildung.' The idea of 'Erziehung' is less about processes and more about the activity of educating. 'Erziehung,' therefore, implies intention on the side of the educator (an educator has the explicit intention to have some impact on the one to be educated) and the articulation of what the activity of 'Erziehung' is supposed to bring about, that is, a sense of the aim or aims of 'Erziehung' (see Oelkers, 1993, 2001). Most importantly, 'Erziehung' implies an orientation toward the education of the person as a free human being. This means that the one being educated is never just seen as an object for influence and manipulation but always also as a subject (or an emerging subject) in its own right.

The conceptual complexities surrounding the word 'education' have led several authors to claim that 'education' is an 'essentially contested concept' (Gallie, 1956). The idea of essentially contested concepts in not so much that they are difficult to define or that people have differing opinions about their meaning, but that this difficulty has to do with the fact that such disputes cannot be settled by just referring to 'the facts' or just by means of argument or logic, because the discussion is ultimately not about what the concept means but what it *should* mean. The discussion is, in other words, *normative* and is concerned with people's values and beliefs. We can see this quite clearly in the case of education because on closer inspection many definitions of education are actually statements of what the person defining the concept considers to be *good or worthwhile* education. This is why Richard Peters' definition of education as implying "that *something worthwhileis being or has beenintentionally transmittedin amorally acceptable manner*" (Peters, 1966: p. 25) is not so much a definition as a statement of his views about *good* education. There is actually nothing wrong with the fact that discussions about what education is or ought to be are thoroughly normative and thus always require engagement with questions of values. After all, if we take education in the sense of educating (the fourth meaning) toward becoming an educated person (the fifth meaning), it is always about deliberate attempts to effect change in someone's knowledge, beliefs, values, attitudes, skills, or personhood – which always raises the question of justification, that is, the question to what extent such changes can be considered desirable.

This is a complicated discussion with a long history that became even more complicated as a result of the rise of Enlightenment ideas of human self-determination and freedom. If, to put it briefly, before the Enlightenment the educational keyword was that of 'perfection' – that is, making the human being more perfect through its engagement with the best aspects of culture and tradition – from the Enlightenment onward the educational keyword became that of 'autonomy,' that is, using one's rational capacities for responsible self-determined action. This shift has resulted in what has become known in the literature as the 'educational paradox,' very aptly summarized by Immanuel Kant in the question "How do I cultivate freedom through coercion?"

The Question of Purpose in Education

The normative dimensions of any discussion about education reveal one further characteristic of the concept of 'education,' which is the fact that 'education' is what in the philosophical literature is known as a *teleological* concept; that is, a concept that implies a *telos*, which is, an aim or purpose. The idea here is that education is not an activity we undertake just for its own sake, but always because we want to achieve something or bring something about. Talking about education thus always raises the question as to what education is *for*. There are quite different views about what education should be for, not only depending on people's understandings of what education 'is' but also, and more importantly, depending on the values they hold and the views they have about what education ought to achieve. Viewed from this angle it is actually quite difficult to make any clear statement about what the purpose or purposes of education should be. This does not mean, however, that it is

not possible to bring some more clarity and precision in discussions about the purposes of education.

A first step toward such clarification is to acknowledge that educational systems perform a number of different functions (see Biesta, 2010a). One major function of education – particularly of formal education – is that of *qualification*. This has to do with providing children, young people, and adults with the knowledge, skills, and understandings and often also with the dispositions and forms of judgment that allow them to 'do something' – a 'doing' which can range from the very specific (such as in the case of vocational or professional education) to the much more general (such as the role of education in preparing children and young people for life in highly complex societies). A second major function of education is that of *socialization*, which has to do with the ways in which, through education, we become initiated into and part of existing social, cultural, political, professional, or religious communities, practices, and traditions. In some cases, socialization is actively pursued – for example, where the aim is to inculcate the values of democratic citizenship or the ways of doing and being of particular professional communities. But even if socialization is not an explicit aim of educational programs and practices, education will still have a socializing 'effect' on those who take part in it – a phenomenon well documented in the idea of the 'hidden curriculum.' Through its socializing function education inserts individuals into existing ways of doing and being and thus plays an important role in the continuation of culture and tradition. Education does, however, not only contribute to qualification and socialization but always also impacts on the human person. Here, education performs the function of subjectification, that is, of becoming a subject of action and responsibility rather than remaining an object of the demands and directions of others.

The distinction between qualification, socialization, and subjectification is not only useful in order to get a clearer understanding of the different ways in which education *functions*, but we can also consider them as *three different domains of educational purpose*; that is, three domains we can – and in a sense ought to – take into consideration when we try to articulate what education should be for.

The Encounter between Philosophy and Education: Some 'Classics'

Having started from the simple definition of philosophy of education as the philosophical study of education and its problems, and having tried to give an answer both to the question of what philosophy is and to the question of what education is, we can now see that the encounter between philosophy and education contains a vast array of possibilities, not only because of the variety of philosophical positions, traditions and ideas, but also because of the complexity of the very notion of education itself. What makes this encounter even more rich and complex is the fact that, unlike in the domain of science where new theories and insights tend to replace older ways of thinking, many if not all of the ideas generated by philosophers over the centuries remain valid or at least interesting. This is because in philosophy later views tend not so much to *replace* earlier insights and ideas as that they add depth, detail, complexity, and controversy to the discussion.

When we look at the actual work done in the philosophy of education, two things stand out. One is that throughout the whole history of philosophy, philosophers themselves have been discussing educational issues and in doing so have generated a number of real 'classics.' In Plato's text *The Republic* (Plato, 1955) we can, for example, find a detailed account of the roles of education in the ideal state, with a clear distinction between the kind of education that the guardians of this state should receive and the kind of education that the rulers of the state, the so-called 'philosopher-kings,' should have. Plato not only provides much 'technical' detail but also discusses extensively the aims and methods of such education. Interestingly, Plato connects the education of the rulers of such a society strongly with philosophy – in a sense he proposes a philosophical education for them – highlighting the crucial importance of knowledge of the good. In *The Republic* which, as many of Plato's writings, is written in the form of a dialogue – in this case one in which Socrates is the main protagonist – Plato introduces a number of ideas and metaphors, such as the idea of education as a process of 'turning the soul' toward the light of knowledge, that are still being discussed today (see, e.g., Haroutunian-Gordon, 2010).

It is, however, not only the *content* of Plato's writings that has made a lasting impact on educational thought. Perhaps even more famous and influential is its dialogical *form* which itself can be understood as a particular view of education. Here we often find Socrates, through a refined play of questions and answers, bringing his interlocutors to insights and understandings including, often, an acknowledgment of the limits and limitations of one's own knowledge on the assumption that the first step to wisdom is recognition of one's own ignorance. Socrates often claimed that he actually knew nothing. This approach has become known in the literature as the Socratic method or Socratic teaching and often compares teaching to the practice of midwifery, that is, a process of bringing out what is already inside – that is, a process of helping students to come to clearer insight and understanding – rather than as a process where the teacher would transmit something new to the student. Aspects of the Socratic method can also be found in what is known as Philosophy for Children, an approach to engage children in philosophical enquiry developed by the American philosopher Matthew Lipman (1922–2010) that has become a popular program around the world (see, e.g., Lipman et al., 1980).

Another important and at the time highly influential text on education written by a philosopher is John Locke's *Some thoughts concerning education*, which was published in 1693 (see, e.g., Locke, 2008). In this book Locke presents, among other things, his idea of the mind as a 'tabula rasa,' that is, as a clean slate upon which experience makes impressions, thus highlighting the crucial important of education on the formation of the human being. Locke also highlighted the importance of instilling good habits in children before embarking on more academic and intellectual forms of education, and discussed extensively the formation of virtue and reason, a process that should focus on treating children as rational beings and on using esteem and disgrace rather than rewards and

punishments as the latter would arouse 'passions' rather than reason. That Locke's text was highly popular and potentially highly influential can be seen from the fact that in the eighteenth century alone it was published in at least 53 editions in English, French, Italian, German, Dutch, and Swedish (see Axtell, 1968: pp. 100–104).

The third major work to mention is Jean-Jacques Rousseau's *Émile, or on Education* which was published in 1762 (see, e.g., Rousseau, 1979). The book – which contains the famous opening sentence "Everything is good as it leaves the hands of the author of things, everything degenerates in the hands of man." ("Tout est bien sortant des mains de l'Auteur des choses, tout dégénère entre les mains de l'homme.") – uses the device of a story about the education of the child and young man Émile by his tutor and also contains a (shorter) section on the education of Sophie, Émile's wife-to-be. One aspect of what Rousseau puts forward in the book is the question of how to balance what is 'of the child' with what is 'of society.' In the psychological literature this issue is often discussed as the 'nature-nurture' debate and concerns the question to what extent what we become is the result of internal (nature) factors or the result of external (nurture) factors – nowadays most people would argue that it is an interaction between the two rather than one or the other. The educational question is, however, not about explaining the factors that contribute to what we are or become, but raises the issue how we can, in terms of Rousseau's argument, maintain the original goodness of the human being while at the same time becoming good and responsible members of society. The educational question here, in other words, is about how to deal with the fact that any engagement with education – either through our own enculturation ('Bildung') or through the efforts of others to impact upon our growth ('Erziehung') – is not just a gain but also potentially a loss (for a poignant discussion of this experience see Kerdeman, 1999; also Mollenhauer, 1983).

Another classical text in the philosophy of education is a short publication attributed to Immanuel Kant (though probably authored by someone else based on notes from a lecture) called *On Education* (*Über Pädagogik*), published in 1803 (see Kant, 1803, 1982). In *On Education*, Kant highlights the crucial role of education in helping human beings to develop their rational powers in order to achieve rational autonomy – that is autonomy based on the use of reason. In his short text "An answer to the question: What is Enlightenment?" (originally published in 1784; see Kant, 1992) Kant had defined the 'motto' of Enlightenment as 'sapere aude,' that is, 'dare to think' or have the courage to use your own capacity for reason. What is interesting from an educational perspective is that Kant did not conceive of this capacity as a contingent historical possibility – something that individuals could either acquire or not – but saw it instead as something that was a central part of human nature. He described the "propensity and vocation to free thinking" as man's "ultimate destination" and as the "aim of his existence" (Kant, 1982: p. 701), which meant that to block progress in enlightenment would be "a crime against human nature" (Kant, 1992[1784]: p. 93). Interestingly enough, however, Kant also argued that the 'propensity to free thinking' could *only* be brought about through education (see Kant, 1982: p. 710). Kant not only wrote that man "is the only creature that has to be educated" (Kant, 1982: p. 697), but he also argued that the human being can only become human – that is, a rational autonomous being – "through education" ("Der Mensch kann nur Mensch werden durch Erziehung"; Kant, 1982: p. 699). With Kant the rationale for the educational 'project' became founded on the idea "of a certain kind of subject who has the inherent potential to become self-motivated and self-directing," while the task of education became one of bringing about or releasing this potential "so that subjects become fully autonomous and capable of exercising their individual and intentional agency" (Usher and Edwards, 1994: pp. 24–25).

The fifth 'classic' to be mentioned here (although this list is far from exhaustive) is John Dewey's *Democracy and education*, a book that was first published in 1916 (see, e.g., Dewey, 1966). Although the subtitle of the book is *An Introduction to the Philosophy of Education*, the book does much more than just introducing the philosophy of education. In the context of an impressive account of the development of educational thought throughout the centuries, the book not only provides a detailed account of Dewey's own 'theory of education' – a theory that, looked at from a Continental perspective, is perhaps more a theory of 'Bildung' than a theory of 'Erziehung'' (see also Siljander et al., 2012) – but also includes a statement of his still influential ideas about the relationship between education and democracy (Biesta, 2010b).

There are of course many more books in which philosophers have engaged with educational questions, and one could argue, looking at the history of such publications, that for a long time major thought about education was indeed generated by philosophers. It is important to mention here – although in a sense it is difficult to have missed this in the discussion so far – that the history of philosophy is entirely dominated by men even if, such as in the case of Rousseau, the discussion is about the education of women. Mary Wollstonecraft (1759–97), now considered to be one of the early feminist philosophers, actually wrote a strong critique of Rousseau's ideas in her *A Vindication of the Rights of Women* (published in 1792; see e.g., Wollstonecraft, 2004). It has only been in recent decades, particularly in North America through the work of such authors as Jane Roland Martin (see, e.g., Martin's landmark book *Reclaiming a Conversation: The Ideal of the Educated Woman* from 1985) and Nel Noddings (see e.g., her *Caring: A Feminine Approach to Ethics and Moral Education* from 1984), that feminist thought has begun to influence the philosophical conversation about education (see also Stone, 1994; Weiner, 1994).

The Contemporary 'Field' of Philosophy of Education

While this work draws inspiration from wider developments in feminist philosophy and theory – and while the impact of feminist thought on education has also been achieved through, for example, feminist psychology, sociology, and cultural theory – feminist philosophy of education should be understood in the context of an important development in the second half of the twentieth century, that of the establishment of philosophy of education as an own 'field' or

specialism, institutionally more strongly located in university schools and departments of education than emanating from schools and departments of philosophy. In the first half of the twentieth century, as Phillips (2001) makes clear, philosophy of education existed mainly of what he characterizes as the 'isms' approach, a way of conducting philosophy of education that locates itself within a particular philosophical school, tradition or approach – such as realism, idealism, Thomism, existentialism, or pragmatism – and then asks what the implications for education of taking such a philosophical stance would be. Phillips suggests that the 'isms' approach suffers from two problems, one being the fact that there is no "simple one-to-one correspondence between a person's deep philosophical commitments ... and his or her everyday actions, or actions in the classroom" (Phillips (2001): p. 4290) and the other being that this approach relies on an "oversimple notion of implication" (Phillips, 2001).

The major way in which philosophy of education established itself from the 1950s onward was in the form of *analytic* philosophy of education – sometimes known as 'APE' both by its admirers and its critics. The work of Israel Scheffler in North America and that of Richard Peters and Paul Hirst in the United Kingdom played a major role in the development of analytic philosophy of education as the dominant approach in what it meant to do philosophy of education. Many others followed in their footsteps and it can be argued that up to the late 1970s or early 1980s analytic philosophy of education – which took its philosophical inspiration mainly from philosophy of language, logic, and epistemology – was the dominant approach in the field. Analytic philosophers of education partly focused their efforts on the analysis and clarification of key educational concepts such as 'teaching' (see, e.g., the discussion of Peters' analysis of the concept in chapter 1; or Macmillan and Nelson, 1968) and 'learning' (see, e.g., Phillips and Soltis, 1985), or of the wider language of education (see, e.g., Scheffler, 1980; see also Peters, 1966; Dearden et al., 1972). In addition to the analysis of key educational concepts (see also Soltis, 1968), the analytic philosophers of education were also interested in exploring the nature of knowledge, truth, and rationality, which resulted in a strong interest within analytic philosophy of education in questions of critical thinking and its education (see, e.g., Siegel, 1988 and, for a critique and reformulation inspired by feminism and pragmatism, Thayer-Bacon, 2000). (For an intriguing and highly readable account of the history of analytic philosophy through a number of intellectual 'self-portraits' see Waks, 2008.)

From the 1980s onward, philosophy of education has become increasingly diverse and plural, engaging not only with a much wider array of philosophical ideas and fields – including social and political philosophy, ethics and moral philosophy, and esthetics – but also with a much wider range of philosophical traditions – including feminist philosophy, postmodern and post-structural philosophy, critical theory, and increasingly also with Aristotelian thought (for an early overview see Kohli, 1995). Another important line within philosophy of education was the rediscovery of pragmatism as a philosophy and a philosophy of education (e.g., Garrison, 1995). While in some regard the range of topics and issues that philosophers of education focus on has expanded together with the widening of philosophical orientations from which the field draws, there is also clear evidence of a return of the 'isms' approach, as many publications focus either on the educational implications of particular philosophical schools and traditions – such as pragmatism, postmodernism, and post-structuralism – or of particular philosophers. The proliferation of philosophical and theoretical orientations within contemporary philosophy of education has, on the one hand, raised questions about what the specific identity of philosophy of education (still) is, but has, on the other hand, opened up new alliances and opportunities, particularly with those who, while working in such fields as curriculum theory and cultural studies of education, draw inspiration from similar philosophical and theoretical resources. While some have expressed concerns about the relatively marginal status of philosophy of education within the wider field of educational research and scholarship, some of the key international journals in the field of philosophy of education – such as *Studies in Philosophy and Education*, *Educational Philosophy and Theory*, *Educational Theory*, and *The Journal of Philosophy of Education* – not only show the high level of activity within the international (English-speaking) philosophy of education community but also provide evidence of the still growing interest in the writings of philosophers of education.

Concluding Comments

Philosophers have always had an interest in educational questions and in this regard philosophy of education can be considered to be as old as philosophy itself. As a distinct field or specialism within educational research and scholarship philosophy of education emerged in the twentieth century and gained its identity perhaps first and foremost through the rise of analytic philosophy of education from the early 1950s onward. More recently, the field of philosophy of education has become more diverse and plural, drawing from a wide range of philosophical traditions and engaging with an equally wide range of educational topics and concerns (for a recent example see Biesta, 2012). One question that remains unresolved with the field of philosophy of education is whether philosophy of education should be understood as a form of applied philosophy that has its intellectual and disciplinary 'roots' within the field of philosophy, or whether it is a form of educational scholarship that draws inspiration from philosophy. Finally, it is important not to forget that the very idea of philosophy of education emerged in the English-speaking world and is therefore closely connected to how the field of educational scholarship has developed more generally in this context, namely in the form of the inter- or multidisciplinary study of education rather than as an academic discipline in its own right. It is, therefore, important not to assume that philosophy of education is a universal phenomenon that has its equivalent in all countries and languages. There remain important and significant differences between how more theoretical and philosophical dimensions of the study of education have gained shape and form in different contexts (see Biesta, 2011), and the idea of philosophy of education is

just one approach within a wider array of ways of engaging theoretically and philosophically with education.

See also: Educational Institutions, History of; Educational Research, The Field of; Educational Sociology.

Bibliography

Axtell, J.L., 1968. Introduction. In: Axtell, J.L. (Ed.), The Educational Writings of John Locke. Cambridge University Press, Cambridge.

Bailey, C., 1984. Beyond the Present and the Particular: A Theory of Liberal Education. Routledge & Kegan Paul, London.

Biesta, G.J.J., 2010a. Good Education in an Age of Measurement: Ethics, Politics, Democracy. Paradigm, Boulder, CO.

Biesta, G.J.J., 2010b. "The most influential theory of the century." Dewey, democratic education and the limits of pragmatism. In: Troehler, D., Schlag, T., Osterwalder, F. (Eds.), Pragmatism and Modernities. Sense Publishers, Rotterdam, pp. 197–213.

Biesta, G.J.J., 2011. Disciplines and theory in the academic study of education: a comparative analysis of the Anglo-American and continental construction of the field. Pedagogy, Culture and Society 19 (2), 175–192.

Biesta, G.J.J. (Ed.), 2012. Making Sense of Education. Fifteen Contemporary Educational Theorists in Their Own Words. Springer, Dordrecht, Heidelberg, New York.

Biesta, G.J.J., Stams, G.J.J.M., 2001. Critical thinking and the question of critique. Some lessons from deconstruction. Studies in Philosophy and Education 20 (1), 57–74.

Dearden, R.F., Hirst, P.H., Peters, R.S. (Eds.), 1972. Education and the Development of Reason. Routledge & Kegan Paul, London.

Dewey, J., 1966. Democracy and Education: An Introduction to the Philosophy of Education. Free Press, New York.

Gallie, W.B., 1956. Essentially contested concepts. Proceedings of the Aristotelian Society 56, 167–198.

Garrison, J., 1995. The New Scholarship on John Dewey. Kluwer Academic Publishers, Dordrecht, Boston, London.

Haroutunian-Gordon, S., 2010. Learning to Teach through Discussion: The Art of Turning the Soul. Yale University Press.

Kant, I., 1803. In: Rink, D.F.T. (Ed.), Über Pädagogik. Friedrich Nicolovius, Koningsberg.

Kant, I., 1982. Über Pädagogik. In: Kant, I. (Ed.), Schiften zur Anthropologie, Geschichtsphilosophie, Politik und Pädagogik. Insel Verlag, Frankfurt am Main, pp. 695–761.

Kant, I., 1992[1784]. An answer to the question 'What is Englightenment?'. In: Waugh, P. (Ed.), Post-modernism: A Reader. Edward Arnold, London, pp. 89–95.

Kerdeman, D., 1999. Between Interlochen and Idaho: hermeneutics and education for understanding. In: Tozer, S. (Ed.), Philosophy of Education 1998. Philosophy of Education Society, Urbana, IL, pp. 272–279.

Kohli, W. (Ed.), 1995. Critical Conversations in Philosophy of Education. Routledge, New York.

Lipman, M., Sharp, A.M., Oscanyan, F.S., 1980. Philosophy in the Classroom. Temple University Press, Philadelphia.

Locke, J., 2008. Some Thoughts Concerning Education. BiblioBazaar, Charleston, SC.

Macmillan, C.J.B., Nelson, T. (Eds.), 1968. Concepts of Teaching. Rand McNally, Chicago.

Martin, J.R., 1985. Reclaiming a Conversation: The Ideal of the Educated Woman. Yale University Press, New Haven, CT.

Martin, J.W., 1981. The ideal of the educated person. Educational Theory 31 (2), 97–109.

Mollenhauer, K., 1983. Vergessene Zusammenhänge. Juventa, München.

Noddings, N., 1984. Caring: A Feminine Approach to Ethics and Moral Education. University of California Press, Berkeley, CA.

Noddings, N., 2012. Philosophy of Education, third ed. Westview Press, Boulder, CO.

Oelkers, J., 1993. Influence and development: two basic paradigms of education. Studies in Philosophy and Education 13 (2), 91–109.

Oelkers, J., 2001. Einführung in die Theorie der Erziehung. Beltz, Weinheim, Basel.

Peters, R.S., 1966. Ethics and Education. Routledge & Kegan Paul, London.

Peters, R.S., 1970. Education and the educated man. Journal of Philosopy of Education 4 (1), 5–20.

Phillips, D.C., 2001. Educational philosophy: historical perspectives. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Pergamon/Elsevier, Oxford, pp. 4286–4292.

Phillips, D.C., Soltis, J., 1985. Perspectives on Learning. Teachers College Press, New York.

Plato, 1955. The Republic (H.P.D. Lee, Trans.). Penguin, London.

Rorty, R., 1978. Philosophy as a kind of writing: an essay on Derrida. New Literary History 10 (1), 141–160.

Rousseau, J.-J., 1979. Emile, or on Education (A. Bloom, Trans.). Basic Books, New York.

Scheffler, I., 1980. The Language of Education. Charles C. Thomas, Springfeld, IL.

Siegel, H., 1988. Educating Reason: Rationality, Critical Thinking, and Education. Routledge, New York.

Siljander, P., Kivelä, A., Sutinen, A. (Eds.), 2012. Theories of 'Bildung' and Growth. Connections and Controversies between Continental Educational Thinking and American Pragmatism. Sense, Rotterdam, Boston, Taipei.

Soltis, J., 1968. An Introduction to the Analysis of Educational Concepts. Addison-Wesley, Reading, MA.

Stone, L. (Ed.), 1994. The Education Feminism Reader. Routledge, London, New York.

Thayer-Bacon, B., 2000. Transforming Critical Thinking. Teachers College Press, New York.

Usher, R., Edwards, R., 1994. Postmodernism and Education. Routledge, London, New York.

Waks, L.J. (Ed.), 2008. Leaders in Philosophy of Education. Intellectual Self-portraits. Sense Publishers, Rotterdam, Boston, Taipei.

Weiner, G., 1994. Feminisms in Education. Open University Press, Buckingham.

Wollstonecraft, M., 2004. A Vindication of the Rights of Woman. Penguin, Harmondsworth.

Educational Planning Worldwide

Françoise Caillods, Formally Deputy Director of the International Institute for Educational Planning, Currently Senior Advisor at the International Social Science Council, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Hallak, volume 6, pp. 4292–4296, © 2001, Elsevier Ltd.

Abstract

Educational planning is the activity that allows the public authorities to orient educational development and identify priority interventions. After going through a major crisis of confidence in the 1980s, educational planning has undergone major transformation: it has become more participatory, more flexible, less technocratic, and more diverse. It has gone beyond what its main focus was for a long time – planning infrastructures, increasing access, and increased efficiency – to become more strategic and addressing a variety of key issues of the educational system, such as quality, inequality, and factors influencing demand for schooling.

Educational planning is the activity which allows the public authorities to orient educational development, set priorities in consultation with other actors, identify priority interventions, and extend funding support to achieve the country's economic and social objectives. It is about exploring the future and developing a vision of what the education system should look like in the next 5–10 years, and identifying what needs to be done to reach this desirable situation.

Educational planning involves a series of steps: analyzing the conditions of the educational system at base year, identifying its major strengths but also its problems and constraints, projecting enrollment on the basis of objectives set by the policy makers, identifying possible strategies to address problems identified, and projecting the necessary human and financial resources required to reach the desirable situation at a reasonable cost. After going through a major crisis of confidence in the late 1970s and 1980s, educational planning has undergone major transformation: it has become more participatory, more flexible, less technocratic, and more diverse. It has gone beyond what its main focus was for a long time – planning infrastructures, increasing access, and increased efficiency – to become more strategic and addressing a variety of key issues of the educational system, such as quality, inequality, and factors influencing demand for schooling. Educational planning is no longer practiced in the planning unit of the Ministry of Education only: it is practiced at all levels from the central to regional level down to the institution and school level. We shall first review what educational planning was early on when it became a major activity of ministries of education, the reasons why it was criticized, the factors that contributed to shaping the way it is now, and the way it is practiced now in a range of countries.

A Bit of History

Socialist states have prepared directive plans ever since the Soviet revolution in 1917. After World War II nonsocialist countries in Europe such as France and the Netherlands also prepared plans to guide their postwar recovery. Japan and the Republic of Korea did the same. But it is in the developing countries that most educational planning activities took place. After independence, in the 1950s and 1960s, numerous countries engaged in rapid expansion of their education system with a view to educating their population and fostering economic development. Several countries had the ambition to universalize access to primary education following the various regional conferences sponsored by UNESCO in Karachi in 1960, Addis Ababa in 1961 and Santiago in 1962. The greatest constraints to development were nevertheless perceived as being the lack of skilled manpower: engineers, doctors, and managers. Most countries created an educational planning unit within their ministry, which were entrusted with the task to estimate their economy's skilled manpower requirement and plan the physical, financial, and human resources needed to support the expansion of the system.

The human capital theory, which was developed in the 1960s, established a link – presented as causal – between years of education, knowledge and skills, productivity, higher earnings, reduced inequality, and faster economic growth. By increasing the people's productive capacity, education was meant to help increase national production and the country's future wealth, while at the same time providing other benefits such as reduced fertility, better health, and nutrition (Schultz, 1961). This theory provided a powerful – albeit often criticized – framework for making choices between investment options (Blaug, 1976). It was used to justify the huge investment made in education both in developing and developed countries. Theoretical discussions concerned the criteria that should guide the development of education at postprimary level. Bid debates existed among economists of education concerning whether planners should use the manpower planning approach (that explicitly gears the degree and type of educational expansion to the anticipated demand for different categories of labor), the investment approach (linking the rate of expansion of one educational level to the rate of return of investments made at different levels), or the social demand approach (essentially projecting past trends). According to the approach followed, more priority was given – or not – to specialized education (technical and vocational education; higher education). Debates continued for many years but on the whole, the three approaches led to the rapid expansion of education (Bertrand, 2004).

Thirty years later, the great hopes that were placed in education and in educational planning were seriously shaken (Caillods, 1989). The economic crisis of the 1970s and 1980s led to rising unemployment among graduates; poverty and

income inequalities within and between countries had not disappeared, on the contrary; as more and more credentials were required to have access to the same jobs, the idea emerged that education could be nothing else but a screening device to select best candidates, without necessarily contributing to productivity. These developments shed doubts on the fact that education did drive economic growth.

With the introduction of structural adjustment programs in several countries, many plans did not get implemented. A number of recommendations were made which came to be known as the Washington consensus (promoted by big international organizations such as the World Bank, the IMF, and the US treasury department all based in Washington) which emphasized cutting on government expenditures and relying more on market forces in education provision and elsewhere in the economy. Planning lost much of its prestige. Several countries stopped preparing plans. These were replaced by policy analysis and a multiplication of projects financed by different donors, largely uncoordinated (Caillods, 1989).

The combination of the economic crisis, the cutting of government expenditures, and the lack of appropriate planning eventually led to a serious deterioration of the education coverage and quality in developing countries, even in a region where education was quite advanced as Latin America. This is why UNESCO, UNICEF, the World Bank, and UNFPA decided to convene the World Conference on Education for All (EFA), which took place in Jomtien, in 1990. The World Declaration on EFA was then adopted, which stressed that education is a fundamental human right and pushed countries to strengthen their efforts to improve education and meet the basic learning needs for all. The international community met again 10 years later at the World Education Forum in Dakar and the participants reaffirmed their commitment to achieving EFA by the year 2015 (UNESCO, 2000).

These two events put educational planning back on the agenda. But planning as it became practiced was quite different from what it used to be.

Traditional planning has been replaced by strategic planning, which involves a variety of different plans from long-term plans to medium-term plans, 3-year operational plans, and 1-year plans, linked to the preparation of the budget. Planning is no longer a one-shot operation leading to the preparation of a document which remains on a shelf: it is now a continuous cyclical process that includes different phases from diagnosis, policy formulation, plan preparation, and costing to plan implementation, monitoring, and evaluation. The implementation stage, and the design of appropriate programs and projects, is part and parcel of the planning process. Also planning is no longer a technocratic exercise prepared by a small group of experts: it involves a large number of actors at central, regional, and local levels.

From Traditional Planning to Strategic Planning

1. Several factors contributed to shaping educational planning as it is practiced now in many developing countries (Caillods, 1997; UNESCO PROAP, 2005; Carron et al., 2010):
   a. To start with the international commitment in favor of EFA. Numerous commitments were made by national governments in the 1990s and 2000s which engaged all governments, those in developing and donor countries (UNESCO, 2000). The commitment in favor of EFA was shortly afterward followed by another international commitment in favor of poverty reduction (the Millennium Development Goals). These two international commitments influenced national policy choices in favor of primary and basic education, girls' education, and to a lesser extent preschool, literacy, and nonformal education. They also influenced donor agencies who gradually shifted their development assistance away from support for higher skill development programs in favor of basic education. It finally strongly emphasized planning: every country had to prepare a national EFA plan, in order to get external funds.
   b. Changes in the practices of external agencies. The concern of development partners for a proper implementation of the EFA goals has had the effect of putting planning back on stage. The existence of a 'credible plan' in favor of basic education became a precondition for external financing. New aid practices emphasize the leadership of the host country, and the existence of locally owned programs of development. Consultation of local stakeholders was taken as an indicator of the country's commitment and a possible guarantee for its successful implementation. New aid modalities increasingly included sector-wide approaches and budget support (sector or full budget support) in the framework of which the resources of several development partners are coordinated. In such schemes, funds are provided to the Ministry of Finance on the basis of an approved sector policy document, with an agreed strategic framework which specifies clear measurable outcomes to be attained within a given time frame. These become the year by year milestones against which progress is measured and funding tranches are released. Access to other international funding schemes (such as the Global Partnership for Education, formally called the EFA Fast Track Initiative) and possibility of debt relief also required that heavily indebted low-income countries prepare a country-driven and -owned Poverty Reduction Strategy Paper, which is based on long-term perspectives and broad consultation of the civil society (Caillods and Hallak, 2004).
   c. Decentralization: As systems of education expanded, it became less and less feasible for central governments to administer education directly. Decentralization of responsibilities to a lower level of governments occurred in every sector as an attempt to improve the efficiency in state governance. In education, the decentralization process often started as a way of tackling the financial crisis passing the burden of financing primary and secondary education to local authorities. It has also been a way of correcting the inefficiency of central administration, placing the locus of decision-making closer to where action is really taking place. The decentralization process took many forms transferring authority to lower administrative levels (deconcentration), to elected bodies accountable to local population (decentralization), or to the private sectors (Mc Ginn and Welsh, 2000). In many countries of South East Asia and Latin America, provincial or state education authorities have thus been entrusted with new

responsibilities in terms of allocation and utilization of resources. Yet the decentralization process is not devoid of problems as richer provinces/states and communities, having more resources and more capacities, often develop faster than others. Central planning remains nevertheless essential to ensure coherence between the different regional plans, to train and support staff in all provinces, to monitor regional developments, to assign resources according to different criteria, and to propose measures to reduce disparities.

d. Modernization of public sector management (UNESCO PROAP, 2005).
Education is the largest public sector in many countries. It is therefore directly affected by public service reforms. It is also often chosen as a pilot sector for such reforms. Decentralization just described is the first type of reform in public sector management which affected education, but also other sectors. Other reforms concern budget management. No plan, no project can be implemented unless it is duly taken into consideration in the national budget. Ministries of finance have started preparing a 3-year budget so as to be able to anticipate what the longer-term impact of various sector policies is and check that they will have the necessary resources to finance them. This reform started in developed countries and spread to developing countries. Increasingly multi-annual budgets, called medium-term expenditure frameworks (MTEFs), get prepared. They are meant to increase the predictability of the budget planning process by providing more reliable estimates of expenses and revenues over a 3-year-period. It is based on the various 3-year plans prepared by different ministries. The better the sectorial plan the more likely it is that it will be properly reflected in the MTEF. Finally, the modernization of public sector management influenced planning by promoting result-based management. What is to be monitored is not so much whether activities have taken place as planned but that expected results and outcomes have been reached.

e. A final factor which influenced planning practices is the spreading of new information technologies and availability of cheap and powerful computers. Preparing a plan, estimating enrollment, and assessing resources requirement used to take ages 30 years ago and limited the number of options that could be analyzed. The new information technology made it very easy to analyze a large number of data and develop simulation models, and to test various hypotheses. Several models exist which allow one to prepare various scenarios of enrollment expansion and costing them. These facilitate the policy dialogue and the presentation of various options to different audiences of stakeholders.

2. Characteristics of strategic planning or modern educational planning.

Several critiques were addressed to the old model of educational planning in the 1980s and 1990s (Hallak, 1997; Carron et al., 2010). According to these, educational planning, as it was practiced in many developing countries, failed primarily because it was too technocratic, too much conceived as a top-down exercise (thus not allowing to build ownership), too rigid (unable to adapt to changing circumstances), too focused on investments and infrastructures and not enough on quality of learning, and not paying attention enough to implementation (such as linking planning with budget preparation for example).

Strategic planning – a concept which was elaborated for the business sector – was widely adopted in education as a reaction to the above criticisms (Caillods, 1989, 1997; Carron et al., 2010). By contrast to what was occurring before, strategic planning:

a. Is guided by a vision of where the country wants to go, what it wants to achieve in the long term, and how this translates into major political orientations for the educational system. The vision and the long-term goals are meant to mobilize the efforts of the various actors and make sure that this is the direction they will all strive for.

b. Is built on knowledge: strengthening of the educational management information system, analysis of indicators, conduct of surveys, review of existing research on key educational issues, evaluations, wide consultations of stakeholders to ascertain what the main strengths and problems of the educational systems are, what the main constraints and blockage in the implementation of previous policies were; what do the stakeholders – families, students, employers – expect from the educational system; and which strategies are most likely to address the problems encountered.

c. Is sensitive to the environment and starts with a thorough analysis of the context: demographic, sociocultural, economic, and political. It is aware of possible risks and tries to adapt to new situations as they emerge, not losing sight of the results the countries want to achieve.

d. Is flexible in its implementation. A strategic plan defines clearly what its final goals and expected outcomes are and which strategies will be used to get there but it does not enter into the details of the different activities and steps to be followed. The identification of major programs and projects, and plan monitoring are integral parts of the planning process, allowing changes to be introduced if the implementation of a certain strategy does not lead to expected results.

e. Recognizes that educational change is the sum of thousands of microdecisions. What explains the success or the failures of a policy may not necessarily be attributed to one particular intervention but to the demand and perception of many independent actors making decisions. It is therefore essential to identify who the different actors are and what is likely to influence their actions and decisions (Kemmerer and Windham, 1997). Stakeholder analysis is a common feature of many educational plans.

The Practice of Educational Planning Today

The situation is different in aid-supported countries and in middle-income countries. Advanced countries no longer practice educational planning at the central level but some of them continue to develop a long-term vision, identify priority programs in the framework of the preparation and implementation of an overall reform, and widely consult stakeholders. Planning is more practiced at the decentralized level and at the institutional levels (universities and schools).

1. Strategic and operational planning in aid-supported countries.

Most countries that are dependent on external assistance have developed a medium-term plan, based on a long-term vision. These plans pretty much follow the same approach. They start with a situation analysis at base year reviewing for the different educational levels performance in terms of access, equity, and quality improvement. On the basis of such diagnosis and of evaluation of the implementation of past policies they define the overall development objectives of the plan and its specific objectives. The list of priority programs and projects that will lead to expected outcomes is then given. The resource implications of the plan are measured with the help of a simulation model, and the possible financing gap assessed. The objectives, the expected outcomes, and output with specific targets are presented in a logical framework. (Several aid agencies (e.g., USAID, World Bank, EU) have adopted the logical framework approach (logframe) as the core technique for managing the complete project cycle from design to implementation, monitoring, and evaluation. Models of logframes can be found on their respective Web sites.) These are monitored by a set of indicators, which, if the country is on track, act as a trigger for the funds to be released. The last part of the plan is the presentation of the monitoring schemes.

Some plans cover the whole education system from primary to higher education; others cover basic education only from preschool to secondary education. Separate plans are sometimes prepared for higher education or for technical and vocational education. Aid-dependent countries have therefore moved from a situation where planning was largely discredited to one where many plans are prepared (long term, medium term, and short term) and planning has become a high-stake process, which determines the funds they receive from development partners.

Planning has become more sophisticated thanks to the use of the latest technology. The planning process has also become more participatory with more and more stakeholders being consulted including the development partners. The implementation is ensured by the preparation of the programs and projects, which are closely monitored. Programs and projects generally concern increased access and equity (with measures which address both supply and demand for education), improvement of the quality of education (reforms of curriculum and pedagogical approaches, teacher training, teaching materials, and use of new technologies), administration, and governance (improvement of management information systems, recruitment and deployment practices of teachers, financial management, and transparency). (The various EFA monitoring reports give a fair summary of the various complex issues which are on the educational planners agenda (UNESCO, 2005, 2009).)

Closer links are established with the budget preparation, which is facilitating implementation. The medium-term plan is linked to the medium-term expenditure framework, while yearly operational plans are prepared, which outlines the detailed activities to be carried out and which are linked to the yearly budget. When readjustments to the medium-term plan are needed, they are introduced in the annual operational plan.

Medium-term plans and yearly operational plans are being increasingly prepared at regional levels. Various educational institutions also prepare plans: universities, technical institutions, research institutions, examination boards, and schools.

Eventually, the planning unit at the Ministry of Education will become less involved in plan preparation and spend more time coordinating the work of other directorates in the Ministry of Education, launching research, organizing the policy dialogue with national stakeholders and with development partners, advising the ministerial authorities on different policy options to solve particular issues, etc., as is the case in several middle-income countries. Many of them already do it.

2. Planning education in middle-income countries and emerging economies.

The situation is much more diverse in middle-income countries and the kind of planning that takes place may be quite different from one country to the next, and from one region to the next; it also varies with time. In Latin America, a region which has a long tradition in educational planning, planning activities have developed in the past 10–15 years inside and outside the Ministry of Education, at central, regional, and institutional levels. Countries have considerably increased the number of plans which they prepare. Different departments or institutions within the Ministry of Education develop their own plans. Planning activities have been decentralized to lower levels and provinces/regions/states increasingly prepare their own plan, as do the educational institutions. As a result, central educational planners get more time to engage in a policy dialogue with different stakeholders; they advise ministry officials on how to address the continuously burning issues of improving quality, reducing inequalities, and preparing the necessary workforce to remain competitive in an increasingly competitive world. They also advise on the preparation of educational reforms.

One of the main tasks of the planning unit in many countries is to organize a strong management information system to inform decision-makers at different levels and also actors and users of the education systems. The planning unit is in charge of producing reliable and timely educational statistics; relevant and up-to-date indicators; organize surveys on a regular basis such as standardized evaluations of students' achievements; evaluate plans and programs; and commission studies on key policy issues. Planning units have to play the role of a knowledge broker facilitating the multiway exchange of information between knowledge producers and those who use knowledge, such as policy makers, teachers, headteachers, or the general public (Poggi, 2013).

In so doing, educational planners become actors in the process of educational change. They no longer are simple technicians; they assume a political role. This is not new: this was always the case for top educational planners.

Conclusions

We are approaching the year 2015, the date fixed for reaching the Dakar goals for EFA. Clearly not all countries will reach all objectives. And a good number of primary school age children are still out of school: 61 million according to the 2012 EFA Global

Monitoring Report (UNESCO, 2012). Getting the last 10% in school will require a lot of effort and innovative approaches. Yet a good number of countries have significantly increased their net enrollment ratios at the primary level. This is largely due to the mobilization of the international community, the introduction of effective policy measures (such as free primary education, conditional cash transfers, crash programs for training, and recruiting teachers), and effective educational planning.

What will happen to educational planning when there is no need for building a large number of new schools, for training a large number of new teachers etc.? The experience of middle-income countries shows that there is still a lot to do for educational planning. New challenges have emerged such as preparing countries to enter the new knowledge economy, and promoting innovation and knowledge-based processes in an increasingly competitive and market-oriented environment. This brings back to the fore the need to invest in technical and vocational education and in higher education in addition to focusing on quality and training at the primary and secondary level (Hanushek and Wössman, 2007). Planning will no longer be the exclusive responsibility of one unit or directorate in the Ministry of Education, nor will it be the responsibility of the Education Ministry alone. There are a host of other actors that need to be engaged from trade unions to universities, companies, and businesses. Ministries of education will remain responsible for ensuring high-quality, efficient, equitable, and innovative education. More than ever planning units will have to monitor educational developments paying particular attention to equity and to learning achievements. They will increasingly have to focus on the design of new governance mechanisms based on incentives so as to promote desired behavior of key actors (Kemmerer and Windham, 1997).

See also: Cross-National Comparisons in Education: Findings from PISA; Education and Economic Growth; Education in South Asia; Education in Sub-Saharan Africa; Education in the Middle East and North Africa; Education, Economics of; Education: Skills Training; Educational Effectiveness, The Field of; Educational Evaluation: Overview; Educational Policy, Mechanisms for the Development of; Equity and Education; Family and Schooling; Globalization in Education; Higher Education Management; School Effectiveness Research; Social Inequality and Schooling; Socioeconomic Status and Academic Achievement.

Bibliography

Blaug, M., 1976. The empirical status of human capital theory: a slightly jaundiced survey. Journal of Economic Literature 14 (3). American Economic Association Publications, Pittsburgh, PA.

Bertrand, O., 2004. Planning Human Resources: Methods, Experiences and Practices. Fundamentals of educational planning No. 41. IIEP-UNESCO, Paris.

Bienaymé, A., 1995. Does company strategy have any lessons for educational planning. In: Hallak, J., Caillods, F. (Eds.), Educational Planning: The International Dimensions. IBE/UNESCO, Geneva. Garland Publishing, New York.

Bray, M., Varghese, N.V., 2011. Directions in Educational Planning. International Experiences and Perspectives. IIEP-UNESCO, Paris.

Caillods, F., 1989. The Prospects for Educational Planning. A Workshop Organised by IIEP on the Occasion of Its XXVth Anniversary. IIEP–UNESCO, Paris.

Caillods, F., 1997. The new educational environment: planning for participation. In: Windham, D., Kemmerer, F. (Eds.), Incentives Analysis and Individual Decision-Making in the Planning of Education. UNESCO, Paris.

Caillods, F., Hallak, J., 2004. Education and PRSPs: A Review of Experience. IIEP UNESCO, Paris.

Carron, G., Mahshi, K., De Grauwe, A., Gay, D., Choudhuri, S., 2010. Strategic planning: concept and rationale. (vol. 1 and Strategic Planning: Techniques and Methods (vol. 3). Education Sector Planning Working Papers. IIEP-UNESCO, Paris UNESCO.

Coombs, P., 1970. What Is Educational Planning? Fundamentals of Educational Planning No. 1. IIEP-UNESCO, Paris.

El Hachem, T., 1992. A Quoi Sert le Plan? Un Regard sur le Système Educatif. Economica, Paris.

Hallak, J., 1997. Preface to Incentives Analysis and Individual Decision-making in the Planning of Education. UNESCO, Paris.

Hanushek, A.E., Wössman, L., 2007. The role of Education Quality in Economic Growth. World Bank Policy Research Working Paper. World Bank, Washington DC.

Kemmerer, F.N., Windham, D.M., 1997. Linking information to incentives. In: Incentives Analysis and Individual Decision-making in the Planning of Education. UNESCO, Paris.

Mc Ginn, N., Welsh, T., 2000. Decentralization of Education: Why, When, What and How. Fundamentals of Educational Planning No. 64. IIEP-UNESCO, Paris.

Poggi, M., 2013. Some Reflections on Two Key Planning Functions: Information and Evaluation; a Latin American Perspective. Power Point. IIEP Buenos Aires Office.

Schultz, T.W., March 1961. Investment in human capital. The American Economic Review 51 (1). American Economic Association Publications, Pittsburgh, PA.

UNESCO, 2000. The Dakar Framework for Action, Education for All: Meeting Our Collective Commitments. UNESCO, Paris.

UNESCO Bangkok Asia and Pacific Regional Bureau for Education, 2005. Implementing National EFA Plans. Handbook for Decentralized Education Planning.

UNESCO, 2005. Education for All. The Quality Imperative. EFA Global Monitoring Report. UNESCO, Paris.

UNESCO, 2009. Overcoming Inequality: Why Governance Matters. EFA Global Monitoring Report. UNESCO, Paris.

UNESCO, 2012. Youth and Skills: Putting Education to Work. EFA Global Monitoring Report. UNESCO, Paris.

Educational Policy, Mechanisms for the Development of

Sandra Vergari, State University of New York at Albany, Albany, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Key factors that shape development of education policy include actors located inside and outside of government, ideas and values, institutions, and information. Education policy materializes in part when policymakers are persuaded to accept a particular definition of a problem and respond with an authoritative decision such as a law, regulation, rule, or order. Education policy is also shaped by dynamics that occur during implementation. Contemporary education policy is developed in a context of global economic competition.

Education politics consists of a struggle among competing actors to secure government policies that reflect their respective values. Education policy materializes in part when policymakers are persuaded to accept a particular definition of a problem and respond with an official decision such as a law, regulation, rule, or order. Education policy is also shaped by dynamics that occur during implementation. Technical and political factors impact the extent to which policy implementation is faithful to the policy as designed. When government chooses not to respond to claims that an education problem exists, this inaction may be viewed as implicit policy. This article discusses key aspects of the education policy process including actors located inside and outside of government, ideas and values, institutions, and information.

Context, Purposes, and Values

Education has numerous consequences for individuals and societies. Education policy is linked to economic growth, social cohesion, democracy, and the overall well-being of individuals and communities. Education is viewed by many as a fundamental human right essential for the exercise of other human rights (UNESCO, 2013). The World Bank (2014) asserts: "Education is a powerful driver of development and one of the strongest instruments for reducing poverty and improving health, gender equality, peace, and stability."

The historical, economic, social, cultural, and governance contexts of a jurisdiction influence education policy. Education policy in a community with a relatively wealthy, homogenous population that has historically favored government activism is likely to differ from policy in an economically depressed community where the native language of many students differs from the dominant language and there is a history of government distrust.

Education policy is increasingly developed in a global context. Factors contributing to this trend include links between education and economic competition in the global economy, the power of ideas, and the sophistication and speed of modern technologies of information gathering and dissemination. Government officials around the world speak about education in terms of economic strength and competitiveness. The President of the United States Barack Obama declared "countries that out-educate us today will out-compete us tomorrow" (Obama, 2010: 1).

The development of education policy is contentious partly because interests disagree about the purposes of education. There are political tensions between interests who value broad, liberal arts education and interests who favor education policy that emphasizes narrower, technical training closely aligned with labor market needs. Some interests view education largely in terms of workforce development, ensuring jobs for individuals, and advancing collective prosperity. Others favor broader purposes including developing students as well-rounded citizens who think critically and are prepared and motivated to contribute to civil society and democracy.

Some interests portray schools as settings for solutions to social problems. They urge policymakers to require schools to address matters such as childhood obesity, teen pregnancy, drug abuse, and bullying and violence. Debates about curriculum policy extend beyond concerns that students master core academic skills. Interests disagree on the values, norms, and historical events to be included in official curricula (Ravitch, 2004).

Education policy development typically emphasizes one or more of the following values: excellence, equity, efficiency, and freedom. Excellence is often discussed in terms of policies intended to ensure a high-quality education for students. Policies focused on excellence address matters such as curriculum standards and teacher quality. Advocates concerned about equity avow that all students are entitled to a high-quality education. Debates about equity often focus on the allocation of financial and human resources to schools and students. Efficiency refers to whether an education policy yields optimal results from the resources invested in the policy. Freedom refers to elements of autonomy and choice on the part of education producers and consumers. Advocates of deregulation in education are concerned about freedom.

Communities disagree on how these values should be embodied in education policy. Education policy is shaped by how policymakers prioritize and weigh trade-offs among the values. For example, while class size reduction may enable teachers to provide higher quality education to each individual student, delivering education to larger classes is more efficient.

Moreover, there may be policies other than class size reduction that balance preferences for both excellence and efficiency and yield more favorable outcomes.

Governance

The governance system of a jurisdiction refers to the institutions and actors that have authority and power to adopt policies. Education policy is developed within classrooms, schools, subnational governments, and national governments. These institutions may be influenced by ideas and policies developed by international institutions such as the Organisation for Economic Co-operation and Development (OECD), United Nations (UN), and World Bank.

The greater the number of government layers and actors involved, the more complex the education policy process. Government complexity can inhibit citizen participation in education policy as it can be difficult to sort out the institutions, policies, and actors that are responsible for particular outcomes. Thus, government complexity inhibits accountability in education policy. On the other hand, multilayered governance offers multiple options for securing policy change (Baumgartner and Jones, 1993).

There is not consensus among education policy researchers, politicians, and stakeholders regarding optimal forms of education governance. Participants in education policy networks debate the merits of centralization versus decentralization and democratic control versus privatization. The public benefits of education justify public funding (Friedman, 1962). Privatization advocates such as Milton Friedman emphasize the freedom value. They prefer that schools be publicly funded yet operated by nongovernment actors free from democratic control and most government regulations (Chubb and Moe, 1990).

There is a lack of research evidence regarding whether particular forms of education governance yield better student outcomes than others (Manna and McGuinn, 2013). Given that student outcomes are determined by a multitude of in-school and out-of-school factors it is difficult to isolate the independent impact of governance. In some cases, the objectives of an education policy are defined by a central government and the means by which to achieve the objectives are decided by decentralized institutions. This blend of centralized and decentralized policymaking permits adaptation of policy to local conditions and local innovations that can inform future centralized policymaking.

In systems of federalism, it is useful to distinguish between federal education policy imposed by the federal government and national education policy that results from subnational units independently adopting similar policies. Canada is unusual among industrialized federations because it does not have a federal department of education and it has relatively few federal education policies. The provinces and territories have adopted many similar education policies and education equity and student outcomes in Canada rank well internationally. Subnational governments are motivated to develop and adopt national education policy due to global economic competition, pressure for subnational units to rank well against each other, policy learning, the power of policy ideas based on logic rather than partisan calculations, and efficiency concerns (Vergari, 2010).

Education Policy Actors

Actors engaged in the education policy process can be classified according to their location inside or outside of government, whether they hold their positions due to election or appointment, and the extent to which they are visible. Visible actors often fulfill prominent roles in placing problems on the policy agenda while less visible and hidden actors are typically more involved in developing proposed policy solutions for problems (Kingdon, 2011).

Government actors include elected and appointed officials and civil servants at the national and subnational levels. Nongovernmental actors include political parties, interest groups, foundations and think tanks, and the mass media. Some of the most active interests in education politics include teachers, business, and foundations. Parents are key stakeholders in education but are not well organized for sustained, broad involvement in education policy. A threat of a school closure or a proposal to cancel a beloved school sports program can mobilize parents for temporary engagement in the education policy process. Once such issues are resolved, parent involvement in policy typically recedes. While schools exist to serve students and prepare them for life in their communities, their voices and potential roles of students in school systems are widely neglected in education policy (Mitra, 2009).

Philanthropic foundations are located outside of government and, thus, lack authority to adopt government policy. However, foundations can wield significant influence and power over education policy. Foundations distribute private funds to encourage government actors to support foundation policy agendas. Contemporary foundations have engaged in two key strategies. First, they provide direct financial support to initiatives such as school choice and online education in order to help ensure project success. Second, foundations fund information dissemination and policy advocacy. In both strategies, foundation funding is a mechanism for persuading policymakers to allocate public funds in accordance with foundation agendas.

Many foundation leaders have backgrounds in the business sector and promote policies that incorporate business practices in public education. In the United States, the Bill & Melinda Gates Foundation, Eli and Edythe Broad Foundation, and the Walton Family Foundation are prominent actors in the education policy arena. Sometimes called 'the Big Three,' their education reform activism is welcomed by some interests and opposed by others. The Big Three favor market-based approaches in education policy including school choice, competition, deregulation, and merit pay for teachers.

Voluntary state applications to the federal 'Race to the Top' competitive grant program in the United States offer recent evidence that states will change their education policies in order to win federal government funds. Just as jurisdictions typically do not want to forego public funds, they may also be willing to align their policies with the priorities of private foundations in return for foundation grants. Opponents of Race to the Top can participate in democratic processes for

holding the federal and state governments accountable. However, as private institutions, foundations are not subject to the accountability mechanisms of publicly funded, democratic government institutions. Critics concerned about the influence of private foundations assert that foundations are able to shape education policy development free from public accountability (Ravitch, 2010).

Some analysts argue that top-down foundation efforts to develop education policy are less likely to take hold and endure than policies developed at the grassroots level (Reckhow, 2013). When stakeholders participate actively in policy development, as opposed to having policy developed externally and imposed upon them, they are more like to view the policy as legitimate, develop a sense of ownership and commitment to robust implementation, and act to protect the policy from threats over the long-term.

Another type of education policy actor is the policy entrepreneur. Located inside or outside of government, the policy entrepreneur is energetic and persistent in advocating a particular idea or policy. Policy entrepreneurs invest time, energy, reputation, and money in pursuit of their policy objectives. They have a claim to a hearing based on their expertise or leadership position. Policy entrepreneurs are politically savvy and skilled at networking and negotiating (Kingdon, 2011). Education policy entrepreneurs have been successful in promoting market-based policy ideas such as school choice (Mintrom, 2000).

In recent years, there has been an increase in the activities of international and transnational actors attempting to influence education policy development around the world. International actors are intergovernmental entities such the OECD, UN, and World Bank. Transnational actors are nongovernmental entities that operate internationally including philanthropic foundations (e.g., Bill & Melinda Gates Foundation, Open Society Foundations), advocacy networks (e.g., Global Campaign for Education), and businesses (e.g., Hewlett Packard).

Some analysts observe that OECD transitioned rapidly from a forum for information dissemination and learning to an international policy actor purposefully engaged in judging school systems and advising education policymakers. Critical observers have raised questions about OECD influence on global education policy development and the willingness of governments to develop education policy in accordance with OECD values and priorities (Meyer and Benavot, 2013). Similar to philanthropic foundations, OECD is not accountable to voting publics. The precise roles of international and transnational actors – including when, where, and why they are able to influence specific education policies – merit greater attention from the education research community (Mundy and Ghali, 2009).

Problem Definition and Agenda Setting

Persuasion is a key mechanism in the education policy process. In order for an education issue to gain agenda status, policymakers must be convinced that the issue reflects not merely a condition but rather a public problem that merits a government response. Values, interest groups, and information shape problem definition (Rochefort and Cobb, 1994). Policy entrepreneurs and other interests make frequent use of education, economic, and social indicators to persuade policymakers to accept their claims about an education problem. Examples of such indicators include data on graduation rates, poverty, and school funding. Indicators do not speak for themselves and education policy activists are strategic in how they interpret and use indicators. Activists advancing claims about a particular problem often use comparisons and rankings as persuasive mechanisms. Among the most frequently referenced indicators in education policy debates are student scores on standardized assessments.

Interests disagree on whether policymakers should address education problems with policies focused on what happens in schools or with policies that address out-of-school factors. There is widespread agreement that the most important in-school factor shaping student performance is teacher quality. Student outcomes are also determined by many additional factors. A few examples include the quality of the curriculum and instructional tools (e.g., texts, computer technology, and science laboratories), student will and capacity to learn, and infrastructure matters such as optimal seating, lighting, and air quality.

The famous 'Coleman Report' established the significant impact of home, neighborhood, and peer environment on student performance (Coleman et al., 1966). Students may be distracted by concerns about being victims of crime on the way to and from school. They may arrive at school hungry and with untreated health problems, in clothing that is unclean or inadequate for seasonal conditions, and they may be suffering from unstable and troubled home situations. Each of these out-of-school factors impact student readiness and ability to learn while in school.

Comparisons

Problem definition often involves comparisons of policies and outcomes within or across nations (Kingdon, 2011). Advocates of school finance reform point to funding disparities in education and appeal to policymakers by emphasizing values of equity, efficiency, and excellence. Student scores on standardized assessments compose one of the most common, prominent items of comparison in education policy development. Education analysts in the United States compare state rankings on the National Assessment of Educational Progress, a standardized student testing system sponsored by the federal government. In Canada, education analysts rank provinces and territories according to student performance on the Pan-Canadian Assessment Program, a voluntary effort of the provinces and territories coordinated by the Council of Ministers of Education, Canada.

Every 3 years, the OECD administers the Program for International Student Assessment (PISA) to 15-year-olds in participating countries and subnational jurisdictions. The assessment measures reading, math, and science knowledge and skills. First administered in 2000, PISA enjoys credibility among many in the mass media and education policy arena. Indeed, PISA is now a prominent indicator in education policy discourse around the globe. Some interest groups and members of the research community view PISA as a useful mechanism for the development of evidence-based education

policy. Other analysts question the validity and reliability of PISA (Meyer and Benavot, 2013). Two other international assessments in which many nations participate are the Progress in International Reading Literacy Study and Trends in International Mathematics and Science Study, both sponsored by the International Association for the Evaluation of Educational Achievement.

The OECD compares PISA results across countries with very different populations, histories, cultures, values, governance structures, and education systems. Based on its high ranking in several administrations of PISA, Finland became a global symbol of educational excellence and the envy of education policymakers in countries ranking lower on PISA. Critics questioned the usefulness of comparing PISA results of the large, diverse United States with those of Finland, a small country with a much more homogeneous population. Given its subsequent decline on the 2012 administration of PISA, Finland may become a symbol of imperfect global assessments and rankings and the international quest for simple answers to complex problems in education policy.

Dramatic Cases and Crisis Rhetoric

Education activists also use dramatic focusing events, symbols, and crisis rhetoric to convince policymakers that an education problem exists (Kingdon, 2011). For example, policy entrepreneurs refer to school shootings as evidence of problems with school safety or to student suicides as evidence of problems with school mental health services or inadequate attention to bullying prevention. After the Soviet Union became the first country to launch a satellite into space in 1957, 'Sputnik' was a focusing event for education policy development in the United States. In 1958, the federal government adopted the National Defense Education Act to improve science, math, and foreign language training with the aim of competing with the Soviet Union in science and technology fields.

In 2000, member countries of the UN endorsed eight 'Millennium Development Goals' to be achieved by 2015. These goals include achieving universal primary education and promoting gender equality. According to UNESCO (2013), despite increased access to education, up to 250 million children have weak literacy and numeracy skills and 775 million adults lack basic literacy skills. Based on these and other indicators, UNESCO asserted that there is a 'global learning crisis' and called for increased global attention to student access to high-quality learning.

Education Policy Design

Prior to the time that a problem gains status on the government agenda, policy specialists have been working on proposed policy solutions. Brand new ideas do not typically receive substantial attention during the policy design process. Instead, policy designers usually combine elements of prior, familiar options into a new policy proposal (Kingdon, 2011). Policy design is shaped by the power of ideas and interests. Education policy designers consider policy goals and the target population, policy instruments, and implementation arrangements.

Policy proposals must be technically feasible and acceptable on value grounds in order to succeed. Technical feasibility refers to logistics including funding. For example, a viable proposal to introduce year-round public schooling in a community would address technical issues such as funding, scheduling, curriculum, staffing and professional development, and installation of air conditioners in schools without such. Policymakers would also need to be convinced that the public favors departing from a longstanding tradition of school-free summers for most students. Constituents might value use of summer months for family vacations rather than for year-round schooling. Summer-oriented business owners might object on economic grounds arguing that they would suffer from a loss of teen labor and decline in summer business.

The power of ideas and policy learning contributes to education policy design. Policy ideas float across borders. Policy specialists in different jurisdictions and levels of government communicate via professional networks. The Internet and other technology advances have led to more rapid collection, analysis, and dissemination of education data and research findings that shape policy discourse and design.

Common problems and pressures in education lead different jurisdictions to design common policy responses. Developers of the Common Core State Standards Initiative in the United States examined the academic standards of countries with high-ranking performance on international student assessments such as PISA. They aimed to align the Common Core standards with the standards of such countries.

Policy Types and Instruments

Education policies can be classified according to whether they are regulatory, redistributive, or distributive (Lowi, 1964). Distributive policies are attractive to policymakers who want to claim credit for broadly distributed benefits (e.g., grants for school building improvements). Regulatory and redistributive policies are more controversial. Regulatory policies require or prohibit certain actions (e.g., teacher licensing processes, secondary school graduation requirements, bans on smoking on school premises). Redistributive policies shift resources or power among groups (e.g., contracting with private providers for school services).

There are five types of education policy instruments or tools (McDonnell and Elmore, 1987; Schneider and Ingram, 1990). First, a mandate is a requirement akin to Lowi's regulatory policy. The success of mandates depends on the will and capacity of the target population to comply and on government will and capacity to enforce compliance. For example, a mandate might require schools to include character education in the curriculum. Policy outcomes are shaped by whether schools embrace or resist the mandate and whether and how government chooses to enforce the mandate.

Second, an inducement is the provision of money or other benefit in return for the production of a good or service. Education grants with policy conditions attached are inducements. Third, capacity building policies are similar to inducements but focus on long-term investment.

Fourth, system change policies are similar to redistributive policies because they transfer authority and power from one group to another. Mayoral control of school districts in some

United States cities is an example of a system change policy. Privatization advocates favor system change policies that shift power over education delivery away from the public sector and toward private providers. Finally, hortatory or persuasive tools distribute information, communicate intended norms, and attempt to convince people that a policy is congruent with their values. Signs posted in a school encouraging students and staff to recycle or be kind to one another are examples of hortatory policy.

Policy Adoption

Education policy is adopted by legislative, executive, and judicial institutions at national and subnational levels. Education policy is also adopted in schools and classrooms. Education interests may engage in venue shopping in a strategic effort to find the optimal government setting for pursuit of their goals (Baumgartner and Jones, 1993). While this article does not focus on courts, it is important to note that the judiciary is an important arena for education policy development and enforcement. For example, advocates of school finance reform and civil rights often pursue their agendas in judicial venues. Judicial education policy is shaped by the constitutional and legal context of a jurisdiction.

Actors with authority to adopt policy assess both the logical merit of a given policy proposal and the strength of interest group support or opposition. Policy change is more likely when there is not strong, organized opposition to a proposed policy. However, the power of an idea can trump interest group power. In the United States, policymakers in a majority of the states demonstrated receptivity to the charter school idea and adopted laws permitting charter schools despite strong opposition from traditionally powerful teacher interests.

Policy change is generally more likely to occur during times of robust budgets. However, the economic context may be of little importance when there is policymaker consensus that a policy addresses a strongly held value or right (e.g., ensuring student safety). An intangible yet significant factor that influences policy adoption is policymaker perception of a general climate or 'mood' in a jurisdiction (Kingdon, 2011). For example, policymakers may perceive consensus in favor of increased accountability or competition in the public sector.

Policy Implementation

Policy development does not end with policy adoption. Adoption of an education policy does not ensure robust implementation of the policy as designed. Opponents can alter the policy during the implementation process. Policy implementation outcomes depend partly on the will and capacity of implementers to act in accordance with the original policy design. In federal systems, subnational actors enjoy expertise and implementation advantages that enable them to resist and reshape federal education policies that threaten their perceived interests (Nugent, 2009; Vergari, 2012).

Implementers opposed to a policy may do the minimum necessary to avoid penalties for noncompliance, and they may not want the policy to gain long-term legitimacy. These factors make it important for policy advocates to secure stakeholder input and 'buy-in' during policy design and adoption. Implementation outcomes are also determined by the will and capacity of policy adopters and administrators to enforce the original policy. Effective policy implementation requires a strategic balance of pressure and support. Pressure focuses implementer attention and support provides resources necessary for robust implementation (McLaughlin, 1987).

When policy implementation proves technically or politically difficult, policymakers and administrators may be tempted to dilute the policy or revise timelines for implementation. For example, if a new policy on secondary school graduation requirements results in large numbers of students unable to graduate on time, policymakers may revise elements of the policy or extend implementation time frames.

Research and Policy Development

Despite frequent rhetoric touting data-drive decision-making, the power of values, beliefs, and preferences can rival the impact of research evidence in the education policy process. Stakeholders and policymakers may neglect high-quality studies that yield mixed or cautionary findings in favor of less rigorous research that supports their policy agendas (Goldhaber and Brewer, 2008). Philanthropic foundations that fund research are often less interested in careful, peer-reviewed research than in shaping research findings that support foundation agendas (Henig, 2008). The Internet and other technology advances have led to rapid collection, analysis, and dissemination of education data and research findings that shape policy discussions. Some researchers rush to share their findings with the mass media and policymakers before the studies have undergone scrutiny by experts in a peer review process (Hess, 2008).

The Drug Abuse Resistance Education (DARE) program was created in 1983 in Los Angeles, California. According to DARE, the program has been implemented in a majority of the United States school districts and in more than 40 countries. The program uses trained, uniformed police officers in schools to deliver a 10-week series of structured lessons aimed at developing student skills to resist drugs, alcohol, gangs, and violence. Two decades after DARE was founded, a series of studies and government reports indicated that there were not significant differences in drug use between students who had completed DARE and students who had not.

Despite the volume of research suggesting that DARE was not producing intended policy results, DARE has remained popular. Based on personal experiences and perceptions, many people simply believe that DARE works. They also like the idea that schools are doing something to prevent youth drug abuse and violence, in accordance with the value preference that schools should fulfill such a role. Supporters suggest that DARE has benefits other than reducing drug use, such as promoting good relationships between police officers and students and enabling educators to focus their time and energy on other parts of the curriculum. While some districts discontinued use of DARE, stakeholders defended the program from threats. In

response to criticism, the DARE curriculum has been revised and evaluations of the new program may inform future policy debates (Birkland, 2011).

Class size reduction is another popular policy despite research indicating significant costs and unintended consequences. It requires that resources be allocated to hiring more teachers and providing adequate classroom space. The costs and benefits of these measures may not compare well to other potential uses for the funds such as teacher professional development and curriculum improvements. Interests such as teachers and parents favor class size reduction as a policy reflecting the value of excellence in education. However, other policies aimed at excellence may be more efficient in terms of yielding better returns on the public funds invested.

Multilayered governance and multiple education reform policies implemented simultaneously make it difficult to isolate the effects of a given education policy and assign accountability for outcomes. When a policy does not yield intended results, there may be several reasons for this outcome. There may be flaws in the policy as designed; the causal theory linking policy design and the implementation process may be faulty; implementers may lack will and capacity to implement the policy in a robust manner; actors overseeing implementation may not provide sufficient monitoring, support, and pressure to ensure successful implementation; and, the policy may have been watered down in response to challenges encountered during implementation.

Future of Education Policy Development

Several education policy trends are likely to continue for the foreseeable future. First, education policy development will continue to be influenced by global economic competition and concerns about workforce preparation. Political portrayals of education policies as straightforward solutions to economic problems tend to disregard additional policy areas that contribute to economic conditions (e.g., monetary, trade, tax, wage, and health care policies).

Second, international and transnational organizations will continue to influence national and subnational education policy discussions. Some analysts are concerned that the global education policy discourse is being guided by narrow conceptions of learning. They contend that easily measured traditional school knowledge and basic skills are emphasized to the neglect of a broader range of learning areas such as ethics, aesthetic appreciation, and civic and environmental responsibility (UNESCO International Bureau of Education, 2013). Global movement toward similar education practices could jeopardize the benefits of policy diversity and innovation (Meyer and Benavot, 2013). While economic goals will remain prominent in education debates, advocates of additional education purposes will continue to influence policy agendas.

Finally, developments in computer and mobile technologies will influence policy agendas pertaining to curriculum and education delivery. Technology advances will also facilitate timely collection and analysis of education data and dissemination of information and ideas.

See also: Agendas: Political; Education and Economic Growth; Education, Economics of; Educational Planning Worldwide; Globalization in Education; Implementation: Political; School Choice; School Improvement; Schools, Micropolitics of.

Bibliography

Baumgartner, F.R., Jones, B.D., 1993. Agendas and Instability in American Politics. University of Chicago Press, Chicago, IL.

Birkland, T.A., 2011. An Introduction to the Policy Process: Theories, Concepts, and Models of Public Policy Making, third ed. M.E. Sharpe, Armonk, NY.

Chubb, J.E., Moe, T.M., 1990. Politics, Markets, and America's Schools. Brookings, Washington, DC.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., York, R.L., 1966. Equality of Educational Opportunity. U.S. Government Printing Office, Washington, DC.

Friedman, M., 1962. Capitalism and Freedom. University of Chicago Press, Chicago.

Goldhaber, D.D., Brewer, D.J., 2008. What gets studied and why: examining the incentives that drive education research. In: Hess, F.M. (Ed.), When Research Matters: How Scholarship Influences Education Policy. Harvard Education Press, Cambridge, MA, pp. 197–217.

Henig, J.R., 2008. The evolving relationship between researchers and public policy. In: Hess, F.M. (Ed.), When Research Matters: How Scholarship Influences Education Policy. Harvard Education Press, Cambridge, MA, pp. 41–62.

Hess, F.M. (Ed.), 2008. When Research Matters: How Scholarship Influences Education Policy. Harvard Education Press, Cambridge, MA.

Kingdon, J.W., 2011. Agendas, Alternatives, and Public Policies, updated second ed. Pearson Longman, Boston, MA.

Lowi, T.J., 1964. American business, public policy, case studies, and political theory. World Politics 16, 677–715.

Manna, P., McGuinn, P. (Eds.), 2013. Education Governance for the Twenty-first Century: Overcoming the Structural Barriers to School Reform. Brookings, Washington, DC.

McDonnell, L.M., Elmore, R.F., 1987. Getting the job done: alternative policy instruments. Educational Evaluation and Policy Analysis 9, 133–152.

McLaughlin, M.W., 1987. Learning from experience: lessons from policy implementation. Educational Evaluation and Policy Analysis 9, 171–178.

Meyer, H.D., Benavot, A. (Eds.), 2013. PISA, Power, and Policy: The Emergence of Global Educational Governance. Symposium, Oxford.

Mintrom, M., 2000. Policy Entrepreneurs and School Choice. Georgetown University Press, Washington, DC.

Mitra, D., 2009. Student voice and student roles in education policy and policy reform. In: Sykes, G., Schneider, B., Plank, D.N., Ford, T.G. (Eds.), Handbook of Education Policy Research. Routledge, New York, pp. 819–830.

Mundy, K., Ghali, M., 2009. International and transnational policy actors in education: a review of the research. In: Sykes, G., Schneider, B., Plank, D.N., Ford, T.G. (Eds.), Handbook of Education Policy Research. Routledge, New York, pp. 717–734.

Nugent, J.D., 2009. Safeguarding Federalism: How States Protect Their Interests in National Policymaking. University of Oklahoma Press, Norman, OK.

Obama, B., 19 January 2010. Remarks by the President on Race to the Top at Graham Road Elementary School, Falls Church, Virginia. The White House Office of the Press Secretary, Washington, DC.

Ravitch, D., 2004. The Language Police: How Pressure Groups Restrict What Students Learn. Vintage, New York.

Ravitch, D., 2010. The Death and Life of the Great American School System: How Testing and Choice Are Undermining Education. Basic Books, New York.

Reckhow, S., 2013. Follow the Money: How Foundation Dollars Change Public School Politics. Oxford University Press, Oxford.

Rochefort, D.A., Cobb, R.W. (Eds.), 1994. The Politics of Problem Definition: Shaping the Policy Agenda. University Press of Kansas, Lawrence, KS.

Schneider, A., Ingram, H., 1990. Behavioral assumptions of policy tools. Journal of Politics 52, 510–529.

United Nations Educational, Scientific and Cultural Organization, 2013. The Global Learning Crisis: Why Every Child Deserves a Quality Education. UNESCO, Paris, France.

UNESCO-International Bureau of Education, 2013. Learning in the Post-2015 Education and Development Agenda. UNESCO-IBE, Geneva, Switzerland.

Vergari, S., 2010. Safeguarding federalism in education policy in Canada and the United States. Publius: The Journal of Federalism 40, 534–557.

Vergari, S., 2012. The limits of federal activism in education policy. Educational Policy 26, 15–34.

World Bank, 2014. Education Overview. http://www.worldbank.org (accessed 02.03.14.).

Educational Productivity

Susan J Paik, School of Educational Studies, Claremont Graduate University, Claremont, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Educational productivity is a chief concern for all nations as it impacts human capital and economic growth. Referred to as human productivity, every nation also has a keen interest in developing exceptional talent. An interdisciplinary lens can provide insights as to how homes, schools, time, and other factors affect learning. A 10-factor productivity model provides a new mastery approach in understanding how multiple factors influence productive outcomes. The alterable factors in the model show how constructive changes can help optimize achievement, excellence, and productive giftedness. Helping students to be productive in school and life early on can increase individual and national productivity.

Introduction and Significance of the Topic

Researchers, practitioners, and policymakers have long been interested in promoting educational productivity and striving for national excellence. Educational productivity is an international concern for all nations as it relates to the importance of developing human capital – an essential goal to maintain economic growth. Sometimes referred to as human productivity, every nation also has a keen interest in developing exceptional talent at every level. While colleges and universities need to help train and provide the necessary skills for the labor market, targeting students early on is the key to keeping talent in the academic and career pipeline. Children's success in school and life are the two most important goals for educators and parents, and their overall productivity will undoubtedly affect national excellence (Paik, 2004).

For these reasons, educational productivity continues to be at the forefront of almost every political debate. Educational productivity has been typically measured by achievement tests generally in math, reading, and science. Although standardized achievement tests have been a source of contention in themselves, achievement test scores in light of economics and politics are undeniably pressing issues. Learning, however, may come in many forms, such as academic, creative, behavioral, cognitive, social, emotional, adult success, and other outcomes. All nations are interested in maximizing learning while minimizing costs, time, and other resources. Costs and benefits are important considerations to any endeavor, especially in education. While costs are typically measured by dollars, time and other resources should also be considered in the productivity equation. For example, developing new educational programs, creating an infrastructure to carry out a new program, and training staff can be costly and time-consuming. While these ideas may be innovative, are they worth the costs, time, and resources? Or are there other ways to achieve productivity? Trillions of dollars are spent everywhere across the world on the latest educational fads, new curriculum and instruction, professional development workshops, and teacher-training programs, but the relationship to achievement is not always clear.

For-profit as well as nonprofit organizations try to produce the best results with little resources. While there are some ways to increase educational productivity, many solutions simply require a lot of time and money. Cutting-edge programs may be helpful, but we often overlook the factors that have the most direct impact – the immediate home and school environments of children, especially in the first 18 years of life (Walberg and Paik, 1997). Investing in the home and school in these years may be an economic strategy with a high rate of returns. With little monetary expenses, learning outcomes can be directly impacted by constructive habits, practices, and attitudes. For example, homework time, supportive home environments, good teaching, direct mentoring, self-regulation skills, productive behavior, etc. are simple suggestions, but powerfully generalizable across all students. Moreover, they add value over time. The home environment is considered more powerful than socioeconomic status, especially in the first 6 years of life. The next 12 years in school can also have a tremendous impact in their schooling years from good teaching, curriculum, and other school factors (Walberg and Paik, 1997). Eighteen years of attentive support and skill-building for every child before they go to college could dramatically influence the overall labor market. Supportive learning environments, practices, and conditions in these critical years should not be underestimated. Research has shown that early investments in children can make for large differences over time and can increase educational and human productivity.

The purpose of this article is to (1) discuss educational productivity and the importance of achieving excellence at the individual and national level, (2) provide an interdisciplinary lens (i.e., economic, educational, psychological, etc.) in understanding how home and school environments can impact productivity, (3) present an international perspective in terms of time investments, (4) describe a new mastery model of educational and human productivity, and to (5) conclude with research, practice, and policy on alterable factors that encourage students to be productive in school and life.

Home and School Environments: Economic Investments

Interdisciplinary perspectives provide a wide lens in assessing problems and finding strategies. This article presents primarily

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92076-2

educational and psychological research, as well as other disciplines including economics. Educational productivity viewed through an economic lens can provide great insights in understanding and explaining human behavior (Becker, 1976). Economic perspectives can help illuminate what is required for excellence for individual and national productivity. More specifically, economists provide a practical understanding of time, money, energy, and other resources. From the early teachings of Adam Smith's *Wealth of Nations* in 1776, economic principles can help explain how human, social, and cultural capital contribute to economic growth and educational productivity.

Common sense tells us if we spend more, we get more … So, does more spending mean higher educational productivity? Many researchers would disagree, including James Coleman, Eric Hanushek, and James Heckman. James Coleman, a famous sociologist, is known for the 'Coleman Report' in 1966. Based on one of the largest studies with more than 650 000 students, he found that learning does not depend on the amount of money spent on schools, but what students bring to school that makes the difference. Referred to as social capital, he found that student background and socioeconomic status were more contributing factors. The study also found that good teachers and schools also impacted learning. Decades later, Eric Hanushek (1994), a leading economist and scholar in the field of educational productivity, found that school learning should not be underestimated. While home environments are indeed essential, quality schools and better use of resources can make a large difference in educational productivity. Hanushek studied the effects of educational spending on schools (i.e., teachers, class sizes, and other school resources) and found little to no influence on student outcomes.

More recently, Nobel Prize recipient James Heckman and Dimitry Masterov (2007) analyzed data from several preschools including the well-known Perry Preschool Project, a 40-year longitudinal study. The experimental study randomly assigned 123 at-risk African-American 3- and 4-year olds into a treatment and control group from Ypsilanti, Michigan. The treatment group received early childhood care and education, while the control group was not exposed to any treatment. The study found positive long-term effects for those in the treatment group. The results showed more individuals were employed, more individuals graduated from high school (particularly females), and crime was lower, resulting in fewer arrests. Moreover, the study found that schooling and workforce outcomes had higher economic returns. Some additional findings for education included reduced tuition subsidies and pupil–teacher ratios, and less demand for adult literacy programs, public job training, convict rehabilitation programs, and police expenditures. Heckman's productivity argument is that investing in young children even before kindergarten is an economic growth strategy.

Since the family unit has deteriorated in the US and other countries in the past few decades, early childcare support is one way to supplement the role of the family, especially for disadvantaged communities. Based on the cost–benefit analysis of their findings, early investments not only provide benefits to society, but can also enhance economic vitality and international competitiveness.

Educational Productivity in International Context: Time Investments

Given the interest in educational productivity, much research on this topic comes from the US perhaps because of the efforts in trying to improve achievement and costs over the years. International and national trends tell us that American productivity continues to remain stagnant though per-student schooling costs are higher in the US than in most other industrialized countries. Elementary and secondary schooling in the US might be one explanation for the achievement differences among industrialized countries. By the end of secondary school, the US generally falls behind many of their Asian and European counterparts. The achievement gap widens from elementary to secondary school. Researchers have examined school expenditures, class sizes, teachers, teaching, curriculum, and other indicators in understanding elementary, middle, and secondary schools with differing results.

Based on international comparisons, East Asian countries appear to fare well. Japan, Korea, Hong Kong, Singapore, and Taiwan have all been top performers in past international achievement tests. The latest 2009 Program of International Student Achievement (PISA) results reported that China was the top performer out of 65 countries though it was their first time to compete internationally. Researchers have sought to explain the achievement differences in the learning environments and practices of East Asian and U.S. students. One striking difference is the use of time in and out of school. U.S. President Obama stated, "Our children spend over a month less in school than children in South Korea. That is no way to prepare them for a 21st century economy" (Edu in Review, 2009). U.S. policymakers have since considered extending instructional time per school day and year.

Time usage can indeed impact individual and school productivity. In the first 18 years, home environments play a strong role as parents nominally control roughly 92% of the student's waking time, and schools strongly influence the remaining 8% of time spent in school (Walberg and Paik, 1997). However, East Asian students spend more time in their school year. In the case of South Korea, Korean students are in school for about 222 school days compared to 180 U.S. days. For U.S. school years, 12 years of schooling (180 6-h days) is equivalent to 12%. In comparison, 12 years of Korean schooling (222 8-h days including Saturdays) is equivalent to 19%. The estimated time in school is about 50% more for Korean students (Paik, 2001).

After-school hours also amount to different learning experiences for East Asian and U.S. students. Homework time, private tutoring, and television viewing appear to influence learning outcomes. While the focus of many after-school programs is on remediation in the US, after-school programs or private tutoring have a different meaning in East Asia. Considered the norm for most East Asian students, Korean 'hagwuan,' Japanese 'juku,' and Chinese 'buxiban' focus on academic enrichment or getting ahead. This additional study time with extra classes or homework time is estimated to be 4 h/day for East Asian students, while U.S. students average one extra hour per day. With minus 40 school days and 3 h less study time, U.S. students watch television on average of 3 h/day (Walberg and Paik, 1997). In calculating total time in and

out of school, many East Asian students, in particular Korean students averaged 100% more learning time from ages 6 to 18 (Paik, 2001). Developing skills in the first 18 years is bound to have some impact on educational productivity nationwide. Time is another precious resource and how it is spent can make for large differences in individuals, schools, and nations.

Human Productivity: Developing Exceptional Talent

In order to be globally competitive with the growing demands of technology, specialized talent is more important than ever to all nations. Nobel prize winning Gary Becker's (1976) work informs us that not only education, but specialized training enhances human capital. Although Adam Smith's economic principles in 1776 on specialization and division of labor referred to a different context of efficient production, the general notion of developing expertise as critical to productivity can be broadly applied here. There is an economic need for specialization in all disciplines. Human productivity (i.e., achievement, accomplishment, eminence) can provide a national platform for excellence. Given the rapid modernization in industrialized countries, a highly skilled workforce is in demand. Encouraging students to enter technical careers, such as sciences, technology, engineering, mathematics, etc., and other specialized fields are important to today's economy. Moreover, despite the discipline, exceptional talent development must be developed strategically.

When and how do we encourage a steady pipeline of talent to enter specialized fields? Generally, for most students, the pipeline ends at college. Therefore, students entering pipelines before or during elementary school provide the most promising outcomes. Research has found that many students start thinking about careers and professions as early as elementary and middle school. By now it is no secret that investing in our students early on will undoubtedly impact individual and national productivity. Referred to as 'Matthew Effects' or accumulative learning by Robert Merton in 1968, early investments will yield greater dividends and higher productivity. In Heckman and Masterov's (2007) study, they describe preschools as "the best job training program" where they learn 'soft skills' such as paying attention, learning social skills and new experiences. Since skills and abilities are linked to economic and social problems, the study showed that early job training did have high benefit–cost ratios and rates of return. Early investment is a strategy of efficiency as talent specialization can later achieve even greater productivity.

Children spend the most time at home and school before they go off to college. How we encourage students during these years can make a tremendous impact. A comprehensive model to understand the factors that influence achievement is needed for some additional strategies. Are student success and talent development scalable? Though not a simple answer, any scalable success can only happen when parents, teachers, and key stakeholders play a supportive role in children's education. There are many strategies, but a comprehensive model that includes individual, home, school, time, and other factors might provide an initial start in understanding productivity.

Productive Giftedness: An Early Investment Mastery Model

The productive giftedness model (a 10-factor productivity model) is a comprehensive and systematic model to understand educational and human productivity (see **Figure 1**) (Paik, 2013). In borrowing economic language, how can we make learning more 'productive?' or make giftedness, talent, or excellence more productive? While the model is largely psychological, it provides an interdisciplinary lens as illustrated earlier through economic, educational, and other perspectives. From an economic standpoint, research studies show that early investments of learning and skill-building can produce positive outcomes over time. In the same vein, the model advocates that accumulative learning can be more advantageous as it can impact productive outcomes. However, it is important to note that learning and talent development are not static, but an ongoing process for all individuals. While developing skills early on is ideal, learning can happen at any time and still be made productive. In cases where students may not have optimal conditions or environments, other factors in the model can be distinctively applied to supplement learning at different stages. Learning must also be considered in the context of the area of specialization.

The productive giftedness model is a mastery model where individual skills and expertise can be nurtured (Paik, 2013). The 'nature versus nurture' debate continues and more recently has been referred to as 'mystery versus mastery' (Matthews and Foster, 2006). The mystery model uses a traditional lens to explain innate abilities, but the mastery model focuses on nurturing and developing ongoing skills. Mastery learning has traditionally been derived from the classroom experience or school learning. For example, mastering each step or concept in the curriculum is required before moving to the next level. With allocated time on task, students can pace their own learning. In the current model, the definition of mastery learning or skill-building extends beyond the walls of the classroom. Time in school and out of school can also affect development, learning, and expertise. Although teachers play an important role,

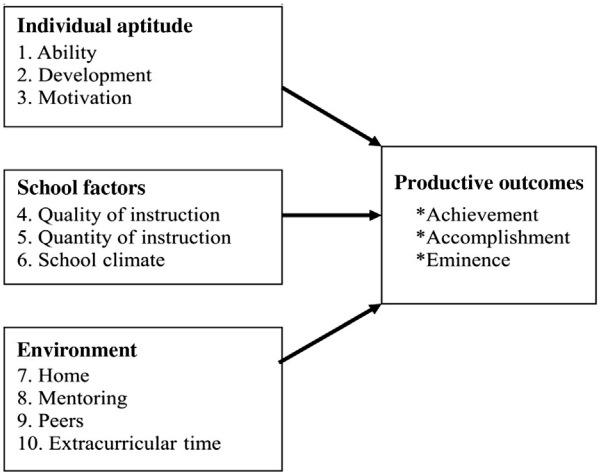

Figure 1 Productive giftedness model.

parents, mentors, experts, and peers also affect skills, habits, and attitudes.

Given its practical approach and accessibility, the model suggests that achievement and accomplishment are possible for all students, even those from disadvantaged communities. The nature argument suggests that individuals are born with talent, while others argue that one must find and develop their own talent. In both cases, nurturance is necessary for fruition. The factors in the model can still be applied distinctively across individual students since learning and development are an ongoing process. Based on the author's previous work and other studies including Walberg's earlier work, the model is generalizable across demographic variables, such as socioeconomic status, age, gender, race, and diverse cultures (i.e., the US, East Asia, Europe, and African-American, Latino, white, other groups in the US) (Paik, 2001; Walberg, 1984; Walberg and Paik, 2005). In general, the model provides a comprehensive and practical approach in assessing individual differences. Mentoring or support programs were also added in the new model as extra measures of support as they can be helpful especially for students coming from less than optimal conditions (Paik, 2013). The model calls for different stakeholders in different disciplines to work together and encourage productive giftedness among students.

Defining Productive Giftedness

Productivity is important in every field, especially in advancing education and developing talent in specialized fields. There are many students with potential, but too few individuals successfully stay in the academic and career pipeline. Giftedness is defined broadly in this context, and such terms as talent, expertise, success, or excellence have been used interchangeably. Despite the use of terminology, 'giftedness' is best understood when made productive (Walberg and Paik, 2005). Productive giftedness is not just potential; it is defined as achievement, accomplishment, and can lead to adult eminence (Paik, 2013). As defined by the author, productive giftedness is inclusive of both high achievement (intelligence or IQ) and high ability (creativity or other domains). Leadership, another form of productive giftedness, is also a productive outcome where researchers have now begun linking early experiences to leadership development (Reichard and Paik, 2011). For the purposes of this article, while creative accomplishment and leadership are important areas of research, the focus of this article will be on academic achievement and high-achieving individuals.

Given the mystery versus mastery debate, the term, 'productive giftedness' is representative of both effort and ability respectively (Paik, 2013). The two terms represent somewhat contradictory ideas, and in fact, some cultures value these concepts differently (i.e., Eastern vs Western ways of learning – effort vs ability). Despite the differing terms, together they represent an important concept that ability can be nurtured. Consider, 'unproductive giftedness' for a moment. One might understand these terms as talent without any direction or efforts. Since learning and development are an ongoing process, all forms of giftedness and talent need to be nurtured. Productive giftedness can be developed and sustained by hard work, perseverance, focused motivation, and other traits (Paik, 2013). Educators and other stakeholders need to take the perspective that all students can achieve when provided with opportunity, support, and resources (Paik and Walberg, 2007). A commitment of their own time and efforts are also keys to their own growth. Teaching children the value of productive giftedness or sense of accomplishment early on can indeed impact educational and human productivity.

Walberg's Productivity Model: School Learning in Context

Although influenced by Carroll, Bloom, and Glaser's school-learning models, Walberg's nine-factor productivity model is inclusive of more environmental factors. The nine factors include the following: (1) development, (2) ability, (3) motivation, (4) quality of instruction, (5) quantity of instruction, (6) home curriculum, (7) social climate of the classroom, (8) peers, and (9) television viewing. The productivity factors have been found to be consistent and generalizable in a number of studies. Specific definitions and concepts of the model can be found in Walberg (1984).

The productive giftedness model builds upon Walberg's earlier work, as well as other seminal research in this field. Walberg's model was selected as the basis of the current model for several reasons. First, compared to other models, the productivity model is one of the most comprehensive models. A multiple-factor approach can help explain individual differences. Second, the model was empirically developed based on a meta-analysis of roughly 3000 studies. Few models and theories have been empirically developed. Third, research studies including the author's own work have found the model to be generalizable with regard to aptitude, instruction, and learning environments (Paik, 2001; Walberg, 1984; other studies). Fourth, the multiple-factor model consists of alterable factors that can impact learning outcomes. Walberg's model provides an empirical basis in understanding educational productivity. While there is no perfect model and none are without limitations, a ten-factor model provides a new mastery approach to understand productivity.

Productive Giftedness Model: 10 Productivity Factors

Along with other theories and models, Walberg's model was intended as a school-learning model. The productive giftedness model has been revised to be inclusive of learning and talent development for early and later success. Several factors share similar terms with Walberg's model, but the underlying concepts have been broadened and adapted to the new model. New terms, concepts, and definitions have also been included in the model. To provide an overview of the model, the factors are briefly described in this section. While all factors of the model are important and interrelated, given the centrality of family and schooling in this article, the following section will conclude with the importance of alterable practices in the home and school.

The multiple-factor model is considered largely alterable, which refers to practices that can be constructively changed, especially in home or school environments. To some degree, adults and students themselves can alter some of the factors. When optimized or altered, they can influence productive outcomes. Alterable practices, attitudes, and habits can include

such examples as homework, study time, curriculum, goal-setting, productive behavior, reading more, self-regulation skills, attitudes, etc. Parents, teachers, and other adults can provide supportive and conducive learning environments for all students.

The model shows that a comprehensive approach is likely to be the best approach, but it is important to note that all factors need not be present to influence productive outcomes. Learning is dependent on all factors, but factors can compensate for one another. For example, use of time (quantity of instruction or extracurricular time) will be different for students given the range of motivation or ability across students. Development or age is also an important factor in terms of the type of instruction they may receive in the classroom. All factors are interrelated and work together, but may be present with differing rates or processes depending on students' interests, abilities, developmental stages of growth, and environmental factors. Environmental factors may also directly and indirectly supplement some of the factors. Homes, schools, peers, and the presence of other adults differ for all students.

Productive Outcomes

Student success is described in the model as *productive outcomes*, and it is further operationalized as *achievement* (generally IQ/intelligence) and *accomplishment* (creativity, other domains). Productive giftedness can be further defined as distinction among peers; they can be broadly operationalized as test scores, awards, honors, competitions, occupation, affiliation, status, etc. Other productive outcomes can include *eminence* or *leadership* (relevant to all domains). Though rare, achievement and accomplishment in the school years can lead to eminence in adulthood. Adult success and other outcomes can also be characterized as productive outcomes. Although *leadership* is not included in **Figure 1**, leadership is another productive outcome for many high-achieving and high-ability individuals. Since the focus of this article is on achievement, the remaining sections will discuss how the model can increase learning and achievement.

Individual Aptitude

There are three factors under individual aptitude: *development, ability, and motivation*. From a lifespan perspective, *development* and learning are continuous. Defined as age or stage of maturation, this factor can be helpful in studying early and later success at different stages. The model can support prospective or retrospective studies for childhood, adolescent, or adult outcomes (Pre-K to post-college) depending on the population and research inquiry. *Ability* has been expanded to incorporate achievement and *other* ability tests (if applicable) depending on the domain-specific area. For academic achievement, standardized test scores have been typically used to measure learning outcomes. Grades and other measurable productive outcomes as noted earlier can also be considered when assessing ability. *Motivation* or *focused motivation* can impact productive outcomes. Specifically, *focused motivation* is defined as "undeterred, intentional perseverance with an end goal or product in mind" (Paik, 2013). Focused motivation is described as intrinsic and/or extrinsic depending on the goals or product. Perseverance, mind-sets, resilience, self-regulation, personality traits, and attitudes are important factors that should be examined. The psychological model naturally lends itself toward a positive psychology lens in studying the strengths of successful individuals, such as perseverance, determination, hard work, and other positive traits.

School Factors

There are three factors in this category: *quality of instruction, quantity of instruction, and school–classroom climate*. The *quality of instruction* refers to excellent teaching, curriculum, content matter, assessment, evaluation, cooperative learning, and any other relevant information that adds to the quality of the school and classroom experience. Much research shows that teachers and what they do in the classroom can directly make a difference. The *quantity of instruction* is defined by the amount of time spent on learning or years in schooling. School hours per day and year, as well as instruction time can make a difference in learning. Schooling years and degrees can also impact learning outcomes. The *school climate* is defined as any school or classroom experiences or characteristics that affect the overall morale of the school or classroom environment. Sociocultural factors, such as type of school, neighborhood, or community, can also provide helpful information.

Environmental Factors

There are four environmental factors: *home, mentoring, peers, and extracurricular time*. The *home environment or* 'curriculum of the home' is a central factor to learning. Parental involvement, practices, and any other home-related information are central to productive outcomes. Demographic information should also be considered, such as culture, race, language, socioeconomic status, parents' education, gender, and any other relevant information. These characteristics are important as they inform us about the role of opportunity for many students, especially those coming from disadvantaged communities. *Mentoring* can involve formal or informal coaching, expertise, programs, or other supportive opportunities (Paik, 2013). Guidance, knowledge, or skill-building in any domain-specific area can be helpful. Mentoring is more common among older students or adults, but there are exceptions and can occur at any age. In the case of disadvantaged or marginalized students, mentoring provides extra support for individuals to stay on track and receive additional guidance. *Peers* can influence attitudes, beliefs, and experiences, especially during adolescence. Peer relationships can provide valuable information at different stages of learning and talent development. *Extracurricular time* is broadly defined as time spent on activities mainly outside of school learning. These activities can be academic, recreational, domain-specific activities, hobbies, sports, participation in places of worship, supportive programs, clubs, etc. Television viewing, gaming, and other forms of technology (i.e., Internet, electronic communication, blogging, etc.) can be time-consuming activities. How students spend their time after school can be informative in understanding individual productivity.

Nurturing Productive Giftedness: Alterable Home and School Practices

Research has found that stimulating home and school conditions and experiences during childhood were strong indicators of later success (Bloom, 1985). Excellence and expertise cannot be fully developed without practice, parental support, dedicated teachers, and good mentoring. Bloom (1985) studied the talent development of world-class neurologists, mathematicians, pianists, sculptors, swimmers, and tennis champions. From the early years, he studied the involvement of parents, teachers, and expert instruction. Bloom later reported, "What we found were exceptional conditions, not exceptional kids." Learning environments vary from home to home and from school to school, but parents, teachers, and other adults should recognize that these environments are alterable and significant on early development and learning.

Altering the Home Curriculum: The Role of Parents

Research has consistently found parental involvement as a significant factor regarding learning, well-being, and other productive outcomes. International studies, such as the Trends in International Math and Science Study and PISA, have confirmed the home environment as strongly related to student success across all countries. Given the concerns of educational productivity, family environments continue to be an important factor in the productivity equation. In the US, national reforms have emphasized the importance of parental involvement, family–school partnerships, and healthy families. Parents are the first and most influential teachers where the curriculum of the home matters more than socioeconomic status. Motivation and learning are also significantly correlated to families and home environments (Walberg, 1984).

National reforms in the US have emphasized not only the importance of parents, but also the value of time for students. In the home, parents can engage students constructively, provide the necessary guidance and feedback, and create a learning environment conducive to learning. Alterable factors include providing stability and routine of family life and healthy parent–child relationships to help support and guide their children, clear expectations and supervision, discipline and affection, extra lessons and monitored time, family field trips, parent–child conversations about school or peer activities, encouragement of leisure reading, engaging them in interactive projects, family time, and discussion of books and television, self-regulation skills and goal-setting. The constructive habits, practices, and conditions optimize learning and long-term goals (Redding, 2003; Walberg, 1984).

Altering the School and Out-of-School Curriculum: The Role of Teachers and Mentors

Given the importance of teachers, billions of dollars have been given to teacher reform initiatives across the world. Similar to home environments, children need conducive classroom environments. Good teachers matter. Teachers have great influence on their students, especially for students who may be coming from homes with less than optimal conditions. Research studies consistently show that teachers have a significant impact on achievement. While educational expenditures or other school policy-related matters are less direct, teachers can indeed directly impact student outcomes.

Teachers can use curriculum and any other classroom experiences to engage and expose students to various topics. Engaging students in projects with real-life problems can also enhance their problem-solving skills. Alterable factors include teaching methods and curriculum, book selection, constructive homework, projects, or programs that enhance learning and other skills, career day for early exposure to different disciplines or occupations, interactive group work or projects, high expectations for each child, and communicating and developing relationships with parents for extra support. What teachers do in the classroom can motivate and encourage students to be engaged in their own learning.

Parents and teachers obviously play an important role, but high-achieving students may need differentiated or expert instruction to provide the appropriate knowledge and skills that parents or teachers may not be able to provide. Research shows that mentors and other supportive programs are extremely helpful for all students, especially disadvantaged students. Mentorships can provide the extra support in a domain-specific area and life in general. Mentorships can take on a formal or informal role providing expert knowledge, skills, and advice at different stages (Clasen and Clasen, 2003; Jacobi, 1991). Mentoring practices may vary in different fields, but encouragement and guidance are helpful to all students. Bloom's (1985) study found that every student worked closely with parents, teachers, and mentors. These supportive communities were critical at every stage during the early, middle, and later years. Clear expectations and encouragement by at least one caring adult can make a tremendous impact on student success.

Conclusion and Implications

Educational productivity requires achieving excellence at the individual and national level. An interdisciplinary and international perspective can provide new insights into how homes, schools, and time factors impact productivity. The productive giftedness model provides a new mastery approach in understanding how multiple factors influence productive outcomes. The alterable factors in the model show how constructive changes can help optimize excellence, expertise, and productive giftedness. Helping students to be productive in school and life early on can increase individual and national productivity.

Children spend the most time with their families (0–6 years) and in schools (6–18 years) in their first 18 years. The accumulative advantage of Matthew Effects shows that early investments of time can optimize long-term productivity in life. Parents, teachers, and other adults need to foster learning in the home and school learning environments. Mentors, experts, and other supportive programs can provide extra support and accountability for student success. While this article did not focus on peers, their role can also be critical in the learning and development process. Supportive communities require all stakeholders to work together where resources, support, and opportunities can impact not just one, but larger numbers of children and youth. The productive giftedness

model suggests that achievement and excellence are possible on an even larger scale.

See also: Academic Self-Concept and Achievement; Cognitive Development: Child Education; Comparative Research in Education: IEA Studies; Education and Learning: Lifespan Perspectives; Education for the Gifted and Talented; Education, Economics of; Environments for Education; Family and Schooling; Intelligence in Cultural, Social and Educational Context; School Effectiveness Research.

Bibliography

Becker, G.S., 1976. The Economic Approach to Human Behavior. The University of Chicago Press, Chicago, IL.

Bloom, B.S., 1985. Generalizations about talent development. In: Bloom, B.S. (Ed.), Developing Talent in Young People. Ballantine, New York, NY.

Clasen, D.R., Clasen, R.E., 2003. Mentoring the gifted and talented. In: Colangelo, N., Davis, G.A. (Eds.), Handbook of Gifted Education. Pearson Education, Boston, MA, pp. 254–267.

Edu in Review, 2009. Obama Proposes Longer School Days, Extended School Year. http://www.eduinreview.com/blog/2009/03/obama-proposes-longer-school-days-extended-school-year/ (retrieved 05.01.11.).

Hanushek, E.A., 1994. Making Schools Work: Improving Performance and Controlling Costs. The Brookings Institution, Washington, DC.

Heckman, J., Masterov, D., 2007. The productivity argument for investing in young children. Review of Agricultural Economics 29 (3), 446–493.

Jacobi, M., 1991. Mentoring and undergraduate academic success. A literature review. Review of Educational Research 61, 505–532.

Matthews, D.J., Foster, J.F., 2006. Mystery to mastery: shifting paradigms in gifted education. Roeper Review 28, 64–69.

Paik, S.J., 2001. Educational productivity in South Korea and the United States. Research monograph: International Journal of Educational Research 35, 535–607.

Paik, S.J. (Ed.), 2004. Advancing Educational Productivity: Policy Implications from National Databases. Information Age Publishing, Greenwich, CT.

Paik, S.J., 2013. Nurturing talent, creativity, and productive giftedness: a new mastery model. In: Kim, K.H., Kaufman, J.C., Baer, J., Sriraman, B. (Eds.), Creatively Gifted Students Are Not like Other Gifted Students Research, Theory, and Practice. Sense Publishers, The Netherlands.

Paik, S.J., Walberg, H.J. (Eds.), 2007. Narrowing the Achievement Gap: Strategies for Educating Latino, Black, and Asian Students. Springer, New York.

Redding, S., 2003. Parents and Learning. Educational Practices Series-2. International Academy of Education, Geneva, Switzerland.

Reichard, R.J., Paik, S.J., 2011. Developing the next generation of leaders: research, policy, and practice. In: Murphy, S.E., Reichard, R.J. (Eds.), Early Development and Leadership: Building the Next Generation of Leaders. Taylor and Francis, New York, pp. 309–328.

Walberg, H.J., 1984. Improving the productivity of America's schools. Educational Leadership 41, 19–27.

Walberg, H.J., Paik, S.J., 1997. Home environments for learning. In: Walberg, H.J., Haertel, G.D. (Eds.), Psychology and Educational Practice. McCutchan Publishing, Berkeley, CA, pp. 356–368.

Walberg, H.J., Paik, S.J., 2005. Making giftedness productive. In: Sternberg, R.J., Davidson, J.E. (Eds.), Conceptions of Giftedness. Cambridge University Press, New York, pp. 395–410.

Educational Research, The Field of

Felice J Levine and Lori Diane Hill, American Educational Research Association, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article focuses on the social organization of education research as a social science field in the United States and considers its development and scope worldwide. It examines the emergence of education research, the structure and context of training in the field, doctoral degree production, characteristics of education researchers and their location in the labor force, the formal organizations of professionals in the field, and patterns of knowledge dissemination. The article also includes salient comparisons to other social science fields.

Introduction to the Field of Education Research

This article focuses on the social organization of education research as a field of study in the social sciences. The emphasis is on education research in the United States, although attention is paid to the development and scope of this field worldwide. The distinctive identity of education research is reflected in its history, presence in higher education institutions, the theoretical perspectives and modes of inquiry that inform knowledge production, characteristics of advanced degree training, the formal organization of professionals in the field and their locations in the labor force, and patterns of knowledge dissemination.

Education Research as a Social Science

Education research[1] is the social science field that studies education and learning processes and the human attributes, interactions, organizations, and institutions that shape educational outcomes. The field embraces the full spectrum of social scientific methods used in other fields and also has driven the development of social science tools and methods.[2] Scholarship in this arena seeks to describe, understand, and explain how learning takes place throughout the life cycle and how formal and informal processes of education affect students' learning and attainment. While considerable attention is paid in education research to the processes of developing critical reasoning skills and knowledge, the field also addresses how educational contexts affect social skills, beliefs, and behaviors.

Like other social science fields, education research is directed to building fundamental knowledge through rigorous methods and the sound application of that knowledge. Its focus is on a major social institution of society – education – in its various organizational forms and structures. It also, however, emphasizes education and learning at all levels of analysis and contexts from early childhood through adult and workforce education. In that sense, education research is analogous to some other fields in the social sciences such as economics or political science that similarly focus on major social institutions and processes; that is, the economy and polity (Levine, 2003).[3]

Emergence of Education Research

Education research as an arena of social scientific inquiry can be traced back to the last decade of the nineteenth century and the first decade of the twentieth century. Historians have observed that the development of the social sciences was influenced by the social reform and social service impulses of the post-Civil War era (Furner, 1975; Haskell, 1977; Ross, 1991). Education research, along with other evolving social science fields and disciplines at the time, emphasized producing useful information to address the challenges of industrialization and a society coping with issues of poverty, race relations, and new immigrant communities. Sociology, for example, emerged during this period to "put 'social science' to work" on social problems (Calhoun, 2007), just as there was a call for a 'science of education' and a 'new profession' to produce facts essential to helping public schools deal with their new roles and responsibilities (Mershon and Schlossman, 2008).

The professionalization of the social sciences essentially occurred within the academy at a time of growth in higher education. As Brint (1994) pointed out, the period of urban industrialization during the late 1800s was ripe for all types of professionalization activity, including the development of modern social science. At that time, departments of education were established in universities (e.g., New York University, 1890; Stanford University, 1891; University of California, Berkeley, 1892), with University of California, Berkeley conferring its first PhD degree in 1898. Education research began to flourish in universities during this period, but it did not develop as distinctive an identity as did other social science professions.[4]

In these early years, the location of influence and leadership for education research was not as exclusively situated within universities as it was for other social science fields. For example, social science associations like the American Political Science Association and the American Sociological Association (ASA) were founded in 1903 and 1905, respectively, largely by academics. In contrast, the establishment of an educational research association in 1916, the National Association of Directors of Educational Research (NADER)[5], was largely driven by directors of bureaus of education research primarily located in city public schools. These founders placed strong emphasis on surveys, systematic information, quantification, and measurement as did their counterparts in universities.

While they worked closely with academics, their interest in having public school administrators use scientific research led them to conceive of and press for the establishment of NADER.

Over time, along with the growth of education faculty, university-based research training, and research bureaus located in universities, the role and influence of academics within NADER expanded. NADER was initially founded as an organization within the Department of Superintendence of the National Education Association (NEA). By 1922, it officially changed its name to the Educational Research Association of America and, by 1928, to the American Educational Research Association (AERA). During the next decades, the culture and identity of the organization and the field became much more academic – with AERA officially separating from NEA in 1968.[6]

Disciplinarity and Interdisciplinary Linkages

From both inside and outside of the field, education research is generally depicted as an interdisciplinary social science (see, e.g., Lagemann and Shulman, 1999). In addition to indicators of its freestanding emergence, in its early development, it was both shaped by and embedded in other social science fields, especially psychology. While education research is situated in the social sciences, it also has strong interdisciplinary links to the humanities, in particular history and philosophy, and to mathematics and the physical sciences (in part because of the emphasis that these sciences have placed on unraveling the dynamics of science learning).

Overall, based on its history and functional properties, education research might best be described as both disciplinary and interdisciplinary in its origins and trajectories. Permeable at its boundaries and broad in its networks, education research includes features of traditional disciplines (e.g., dedicated journals, scholarly organizations, graduate training) that help to support and advance building cumulative bodies of knowledge (Levine, 2003). Education research, however, also benefits from its adjacency and synergism with other disciplines and subfields within them. For example, work on expectancy effects in psychology and policy implementation in political science has, respectively, informed studies of student achievement and of the impact of educational reforms.

In the twenty-first century, much attention is anchored on knowledge integration and building knowledge across all fields of science. In general, in social science fields, distinctions between interdisciplinary and disciplinary are neither well understood nor readily resolved. As Calhoun and Rhoten (2010: p. 103) observed, "The distinctions among the social science disciplines are historically forged and largely arbitrary … They may be more matters of style than method – characteristic structures of attentions, values, and ways of solving problems – but they are jealously defended and also maintained simply by habit and social networks."

Structure and Context of Training

Training in education research essentially occurs at the doctoral level. At the undergraduate and master's levels, students primarily pursue education degrees in order to enter the teaching profession in elementary or high school settings. In general, undergraduates in the social sciences are less anchored on scientific careers within their fields than in other fields of science (Levine and Bell, 2015). At the bachelor's level in education, programs are focused on teacher preparation, and thus even a lower proportion of students receive initial exposure to research or to becoming research professionals. Over the last 10 years, some growing interest in undergraduate programs is evident.[7] The University of California, Berkeley and the University of California, Irvine are examples of institutions that offer concentrations in education research, with the former having a minor and the latter introducing a major in education sciences commencing in 2014. Master's degree programs in graduate schools of education are also largely directed to professional preparation for teaching. In some fields of education research, however, master's programs prepare graduates for applied research positions (e.g., MA programs in educational measurement or testing or in educational policy).

Most education research professionals receive their preparation and training in graduate schools of education.[8] Education is also a specialty concentration within other social science (e.g., sociology of education) and humanities (e.g., history of education) disciplines, but the vast proportion of doctoral degree production occurs within schools of education. The number of US universities offering a doctorate in an education research field is not precisely known, but was estimated at 133[9] by Levine and Shepard in undertaking the most comprehensive study of education research doctorate programs to date in the United States (DocStudy).[10]

Education Research Doctoral Preparation

Education research doctorate programs prepare a substantial number of doctorates each year. One important distinction between doctorate education in education research and other social science fields relates to the size of degree production. Disciplines like sociology, political science, or economics have general preparation programs where students specialize in one or another area. In education research, where doctoral production is much more sizable, the structure of training is typically more organizationally complex, with dedicated doctorate programs in identifiable fields.

Levine and Shepard emphasize that the vast majority of graduate schools of education have separate programs in more than one field, although they vary in how they structure their doctorate programs. The analog they use for doctoral training in a field as large as education research is to the life sciences where overall there is also high doctoral production and where dedicated doctorate programs can span fields from genomics to entomology (Levine and Shepard, 2008).

Taxonomy of Education Research Doctorate Programs

An important window on education research comes from a specification of the fields that train doctorates through identifiable programs. Levine and Shepard (2008, 2013) developed an empirical taxonomy as part of their assessment of education research doctorate programs. Based on systematic research of all identifiable education research doctorate programs, they arrived at a taxonomy of 17 fields that produce

almost all of the education research doctorates in the United States.

Levine and Shepard (2013) point out that the number of research doctorate programs in each of the 17 fields varies – from fields like curriculum and instruction with a very large number to those with a smaller number of programs like educational policy.[11] They report that the fields engaged in doctoral preparation range from those closely aligned to teaching and the curriculum components of education (including in reading/literacy, mathematics, and science) to those more anchored on learning and developmental processes and the broader contexts that shape education or in which learning occurs. They also observed dedicated doctorate programs in educational measurement and statistics.

Doctorate Degree Production

Overall Patterns

Dedicated education research doctorate programs support the production of a very large number of doctorates each year. Table 1 shows doctorates awarded in education research and six other social science-related fields from 1987 to 2012. Over this 25-year period, the number of education research doctorates produced remained sizable and steady at approximately 2500 per year across fields.[12] Adjusting for clinical psychology, fields of psychological research were generally comparable to education research in doctorate production at approximately 2000 per year. In comparison, the other social sciences produced 1000 or fewer doctorates per year over the same 25-year period, with only modest growth in certain fields over time.

Diversity and Doctoral Production

Over the 10-year period from 2002 to 2012, women earned approximately 70% of the doctorates in education research (S&ED, 2012: Table 15). This high proportion reflects an overall trend in the expansion of doctorates earned by women since the 1980s. The National Science Foundation (NSF) does not report on historical trends in the gender composition of those earning doctorates in education research. Overall patterns within the social sciences show an increase of women receiving doctorates over time – from slightly more than one-third in 1982 to nearly 60% in 2012 (S&ED, 2012: Table 14). There is also substantial variation across fields. In 2012, women earned 63.4% of doctorates awarded in sociology, and 71% of doctorates in psychology. In 2012, in education research, women earned 70.4% of the education research doctorates (S&ED, 2012: Table 16).

In 2012, approximately 25% of recent doctorate recipients in education research were persons of color (S&ED, 2012: Table 24).[13] African-Americans were the largest racial/ethnic minority group, earning 12.9% of the doctorates, followed by Hispanics earning 6.3% of the doctorates, and Asians earning 4.6%. In comparison, across the social sciences, approximately 22% of the 2012 doctorates were persons of color. As reported by Levine and Bell (2015), there is considerable variation in minorities earning doctorate degrees in the social sciences by discipline. Social science fields ranged from those with overall low numbers of minorities earning doctorates (anthropology and political science with 15.9% and 15.7%, respectively) to those more similar to education research (psychology, economics, and sociology with 21.4%, 23.5%, and 27%, respectively) (S&ED, 2012: Table 24).

Table 1 Doctorate degrees awarded in seven social science-related fields for selected years, 1987–2012

	Year					
Field	*1987*	*1992*	*1997*	*2002*	*2007*	*2012*
Anthropology	352	320	430	496	512	546
Economics	796	885	982	894	974	1155
Political science	404	513	654	606	588	726
Sociology	423	495	571	547	576	631
Statistics (social sciences)	49	29	56	54	29	18
Psychology	3173	3263	3366	3206	3276	3614
Psychology (nonclinical)	1959	1954	2105	1985	2116	2385
Education	6454	6677	6497	6508	6448	4802
Education research	2602	2503	2664	2776	2671	2571

Sources: National Science Foundation, 1999. Doctorate Recipients from United States Universities: Summary Report 1997, Appendix Table B-1; S&ED, 2012. Doctorate Recipients from U.S. Universities, Science and Engineering Doctorates: Detailed Statistical Tables, Table 13.

Characteristics of the Profession

As noted above, the education research profession is sizable. It is composed of social scientists, students, and other scholars engaged in research in arenas reflecting both the scope of the field and the numbers trained to pursue professional research careers. Currently data are limited on the size, scope, and parameters of this research community; preparation for the education research profession, career patterns, and trajectories; and areas of stability and change over time (Levine, 2003; Ranis and Walters, 2004).[14] However, imperfect, the AERA membership database provides the broadest single overview available of the education research profession as currently constituted.[15]

General Overview

Just taking the United States, the total number of active education research professionals is not known but might be estimated at 50 000.[16] The AERA membership is an undercount of active professionals because some proportion of the community does not join professional organizations or joins specialized research organizations. In 2012, AERA reported over 25 000 members, with 63% regular members, 27.4% graduate students, and 4.8% international affiliates, among other categories.[17] The sizable proportion of graduate students (and the stability of these numbers over time) reflects an attraction to the field and to the continuing engagement of successive cohorts in it.

In any field, one indicator of an identifiable research community is a shared professionalization experience within a single field. The AERA membership data suggest that to be the

case. The vast majority – 74.9% – reports their primary discipline as education rather than a separate social science or other discipline.[18] Further, professionals in this field dominantly have a doctorate degree. Within AERA, the 66.9% with doctorate degrees and the 27.8% with master's degrees generally reflect the relative proportion of nonstudent and student members.

Also, the patterns of participation of education researchers within AERA reflect the major areas of research emphases. AERA is structured into 12 divisions. The four largest divisions in terms of membership are teaching and teacher education, learning and instruction, the social context of education, and measurement and research methodology. Each has memberships ranging from approximately 3000 to almost 6000. As would be expected, these generally correspond to some of the major fields within education research where there are also dedicated doctorate programs (Levine and Shepard, 2013).

Demographic Profile

The demographic backgrounds of AERA members also provide a window on the composition of the field. It is consistent with the profile presented above for recent doctorates. Of the AERA membership, approximately 65% are female and 35% male, reflecting general patterns and trends in some (e.g., sociology and psychology), but not all, of the social sciences (Levine and Bell, 2015). In terms of race and ethnicity, 67.4% of the AERA membership is non-Hispanic white, 10.7% is Asian/Pacific Islander, 8.4% is African-American, and approximately 6.1% is Hispanic. As seen in other fields, less than 1% is American Indian or other Native American.

Education Research Professionals and the Workplace

Sectors of Employment

Constructing a clear and complete picture of education researchers in the US labor force remains challenging. An important limiting factor is the way in which education fields are classified in national surveys and statistical profiles of employment in scientific fields. For example, the NSF's Scientists and Engineers Statistical Data System classifies all education fields (except math and science teacher education) as 'nonscience and engineering fields.' Thus, education researchers are not represented in social science occupation statistics.

The NSF data on postgraduation plans of doctorate recipients provide some useful information on employment sectors (S&ED, 2012). Like recent doctorates in other social science fields, a substantial majority of education researchers are employed in academic institutions. An examination of the 2012 data shows that, among those who had secured employment and planned to enter the workforce immediately, 62% planned to work in postsecondary institutions. The employment destination for the second largest segment of this cohort (22%) was primarily elementary or secondary schools. Much smaller sectors of employment are nonprofit organizations (which include research and development institutions), industry/business, and government with 6.4%, 5.6%, and 3.9%, respectively (S&ED, 2012: Table 65). In the same year, 60.8% of new social science doctorates moved into the postsecondary education sector although concentration in this sector is considerably higher for some disciplines than for others.[19]

The AERA membership database helps to complete this picture. These data reflect the entire membership and not just those entering employment after their doctorate. Education research professionals are dominantly in universities (79.2%), with the next categories being considerably lower (school systems at 7.2%, R & D organizations at 3.9%, government at 1.4%, and industry/business at 0.6%). The presence of education research professionals in universities is higher than for professionals overall in science and engineering fields (45%) and more comparable to some (though not all) fields of social science, including political science, sociology, and anthropology (see Levine and Bell, 2015). Indeed, the AERA data are consistent with the 2012 membership data from the ASA where 79.6% of employed members were in the academy.[20]

Education Research in Universities and Research and Development Organizations

As set forth above, universities are the primary employment location for education researchers, including those with faculty positions and those working in university-based research settings. Also, research bureaus, research institutes, R & D organizations, and research agencies in some states or school districts are major sites of employment. The social organization of the work sectors for education research might be viewed as a complex and overlapping set of entities that is largely anchored by faculty in education research doctorate programs[21] but also includes prominent units inside and outside of universities.

Within universities, education research centers vary in size, scope, and purpose. Some are adjacent to and very much reflect the programmatic research interests of specific faculty at an institution. Others are major entities in their own right with quite sizable research teams, projects, and streams of funding involving faculty as well as other education research professions. Examples of long-term prominent centers are the Learning Research and Development Center at the University of Pittsburgh founded in 1963; the University of Wisconsin Center for Education Research founded in 1964; the Center for the Study of Evaluation (now called the National Center for Research on Evaluation, Standards, and Student Testing) at the University of California, Los Angeles founded in 1966; and the Center for the Social Organization of Schools at Johns Hopkins University also founded in 1966. Some of these centers themselves have collaborative partnerships, for example, the Hopkins Center includes a Center for Research on the Education of Students Placed At Risk, which is a partnership with Howard University.

As noted earlier, university-based bureaus appeared early in the twentieth century both as contexts for providing research-based service and research. In 1964, Lazarsfeld and Sieber in their study of organizing education research observed that university units are variously called bureaus, institutes, centers, and laboratories – noting that in some universities

a tension between research and service led to a splitting of the two (Lazarsfeld and Sieber, 1964: p. 37). The Bureau of Educational Research at the University of Illinois, Urbana-Champaign is among the oldest bureaus and is primarily directed to the support of university faculty and students in the development of their research and the identification of funding sources. The Bureau of Educational Research and Field Services at the University of Nebraska, Lincoln in contrast undertakes research and research dissemination in response to requests by the state.

Over the last several decades, an increasing number of research entities on university campuses have adopted a model that establishes a context for structured, long-term collaboration among university researchers, school districts, and community stakeholders. A well-known example of this kind of campus-based education research center is the University of Chicago's Consortium on Chicago School Research, which was established in 1990 and conducts research intended to build the capacity of the Chicago Public Schools to address their most important and pressing challenges.

Beyond the university setting, there are a number of nonprofit research organizations that also have long-standing histories as producers of education research, data, and evaluations. Among the most visible nationally and internationally are the American Institutes for Research (AIR) with a strong emphasis on education research from its founding in 1946 and the Education Program at the Rand Corporation which dates back to the 1970s (Rand was founded in 1948). Some of the most visible companies engaged in educational testing and assessments also have long-term R & D programs. Three notable examples are ETS, the College Board, and ACT. All of these nonuniversity sectors include postdoctoral training and engage graduate students on research teams although, with the exception of Rand, none includes graduate degree conferring programs.

Also private sector organizations include education research units. Publishers like Pearson that have expanded their portfolio to assessment services include R & D units and research teams. Also, state agencies and school districts continue to have research units doing important work. In addition, currently 10 Regional Educational Laboratories (RELs) are federally funded to assist state, school districts, and others to use data and research to improve practices and outcomes.[22] While typically not directly undertaking research, public agencies (e.g., Institute of Education Sciences) and private foundations (e.g., Spencer Foundation) include education research professionals on staff that contribute to the overall capacity of the field. Despite the importance of all of these work sectors, universities and their faculties and research centers are central to knowledge production and to the training of master- and doctoral-level graduates working in these other sectors.

Professional Associations

Education research associations are an important part of the formal infrastructure of the field. As is the case in all scientific and scholarly fields, from their inception, research associations have provided an organizational framework for scholarly interaction, communication, and dissemination. In the early part of the twentieth century, most associations embraced this purpose through convening meetings and publishing journals. In the last quarter of the twentieth century, large national research associations began taking on other programmatic functions that support and advance their fields and professionals within them. Along with new functions came a gradual shift from volunteer-driven organizations to the emergence of an employed staff, although the role, functions, and leadership of elected officers, councils, and committees continue to be central to such associations.

As the largest professional association in education research, AERA evidenced this transformation. The association's functions include the articulation of ethical standards, professional development courses at largely annual meetings, doctoral dissertation and research grant support for early career scholars, initiatives directed to the inclusion of racial and ethnic minorities in the field, and government relations activities to press for research funding and sound R & D policies. While social science societies changed at somewhat different paces, AERA and associations such as the ASA and the American Political Science Association showed similar patterns of development and understandings of their missions by the end of the first decade of the twenty-first century.

With education research being as large a field as it is in the United States, other education research associations evolved and operate to support education research professionals in the field. There are some 24 state and regional education research associations (e.g., Eastern Educational Research Association, California Educational Research Association). In addition, education research within the United States includes research associations in specialized areas such as the Association for the Study of Higher Education, the National Council on Measurement in Education, the History of Education Society, and the Society for Research on Educational Effectiveness. Like AERA, many of these associations include international members. The National Association for Research in Science Teaching explicitly states that it is international. Also, the Comparative and International Education Society has by virtue of its aims an inherently international perspective.

In addition to education research associations, a number of national social science associations (e.g., American Anthropology Association, American Psychological Association, ASA) have well-established subfield specialties in education research with dedicated sections or divisions reflecting this field. These education research-focused subgroups have a strong presence in their respective organizations, and, through the publication of journals (e.g., *Anthropology and Education Quarterly*, *Sociology of Education*) and other activities, they also play an important role in advancing education research as a social science field.

Journals and Knowledge Dissemination

Peer-reviewed journals are the primary mechanism for research dissemination. For the major associations in the United States, journal publishing is a foundational function. As can be seen in Table 2, in the United States, associations

Table 2 Top 100 education and education research journals by year founded, region, and association affiliation

	Year founded				
	Before 1950	*1950–75*	*1976–2000*	*2001–Present*	*All years*
Region: US					
Affiliated	4	14	9	7	34
Nonaffiliated	6	2	7	1	16
US total	10	16	16	8	50
Region: Non-US					
Affiliated	0	6	12	1	19
Nonaffiliated	2	8	19	2	31
Non-US total	2	14	31	3	50

Source: Thomson Reuters, 2013. Journal Citation Reports, Social Sciences Edition. Additional data on year founded, region, and association affiliation retrieved from journal Web sites. Full source data table available upon request from the first author.

publish the most visible and highly ranked journals. Of the US journals ranked among the top 100 by Thomson Reuters Institute for Scientific Information (TR) in 2012, 68% (34 of 50) were affiliated with research associations. The fact that approximately 70% of the top-ranked US journals founded up through 1975 (18 of 26) were association journals further underscores the early emphasis placed by associations on high-quality publishing, since these journals currently remain highly ranked and cited.[23]

The TR data also provides a window on the overall impact of education research journals. Impact factors for scientific journals are calculated as a ratio of the total number of published items that could be cited to the total number of citations from a journal during a specific period. Figure 1 shows the average impact factors for 2012 and for the 5-year period ending in 2012 for the top 100 ranked journals in education/education research in comparison to other social sciences. These data indicate that the impact of education research journals is comparable to journals in most other social sciences – 1.35 for 2012 and 1.62 over 5 years. As can be seen in Figure 1, the impact of education research journals is much like those in sociology and statistics/probability and higher than in political science and anthropology. Economics and psychology journals have a higher impact than journals in the other social science fields.

International Footprint of Education Research

The international scope of education research as a field of study is long-standing and has broadened substantially in recent years. Table 3 shows national and regional education associations that have been established outside of the United States by founding year. Five associations were established prior to 1960, three in Asia and one each in Europe and South Africa. Between 1960 and 1980 the total number of education research associations more than doubled. In 1970, the World Council of Comparative Education Societies was established. Also, a substantial number of Western European education associations were established in the 1970s.

Table 3 indicates that, during the last two decades of the twentieth century, the growth of education research associations extended to other areas of Asia, Europe, and America. Also, two major international associations – the European Association for Research on Learning and Instruction and the European Educational Research Association – were established in 1989 and 1994, respectively. Since 2000, the global expansion of associations that focus on the advancement of education research as a social scientific field of study has continued. Most prominently, the World Education Research Association was established in 2009. The founding of the Pakistan Association for Research in Education in 2006, the Peruvian Educational Research Association in 2009, and the Slovenian

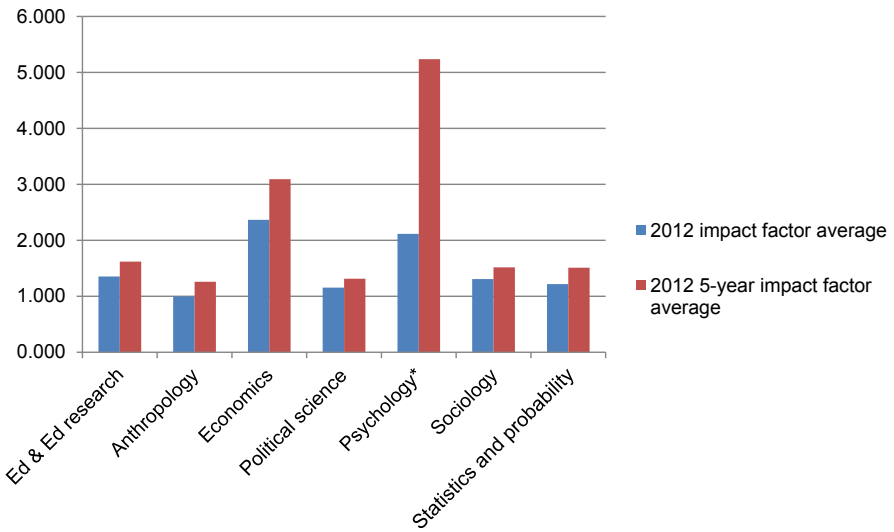

*Excludes clinical psychology and psychoanalysis.

Figure 1 Impact factors for top 100 journals in selected social science-related fields. Source: 2012 Journal Citation Reports® Social Sciences Edition; Reuters, T., 2013. Journal Citation Reports, Social Science Edition. http://thomsonreuters.com/journal-citation-reports/ (accessed 05.10.14.).

Table 3 Selected non-US international education research associations by year of founding

Associations founded before 1960	*Associations founded from 1960 to 1979*
China Education Society (1933)	German Educational Research Association (1964)
Japanese Educational Research Association (1941)	Finnish Educational Research Association (1967)
Spanish Society of Pedagogy (1949)	Australian Association for Research in Education (1970)
Korean Educational Research Association (1953)	World Council of Comparative Education Societies (1970)
Education Association of South Africa (1958)	Association of Teachers and Researchers in Education Sciences (France) (1971)
	Canadian Society for the Study of Education (1972)
	Nordic Educational Research Association (1972)
	British Educational Research Association (1974)
	Scottish Educational Research Association (1974)
	Educational Studies Association of Ireland (1975)
	Netherlands Educational Research Association (1975)
	Swiss Society for Research in Education (1975)
	International Group for the Psychology of Mathematics Education (1976)
Associations founded from 1980 to 1999	*Associations founded from 2000 to 2013*
Hellenic Educational Society (1981)	Austrian Association for Research and Development in Education (2000)
Polish Pedagogical Association (1981)	Belgian Francophone Association of Educational Researchers (2000)
Hong Kong Education Research Association (1984)	The International Society of the Learning Sciences (2002)
Educational Research Association of Singapore (1986)	Slovak Society for Educational Sciences (2002)
All India Association for Educational Research (1987)	Cyprus Educational Sciences Association (2004)
Cyprus Pedagogical Association (1987)	Pakistan Association for Research in Education (2006)
Interuniversity Educational Research Association (Spain) (1987)	Turkish Educational Research Association (2008)
Estonian Academic Educational Society (1989)	Peruvian Educational Research Society (2009)
European Association for Research on Learning and Instruction (1989)	Taiwan Education Research Association (2009)
Hungarian Educational Research Association (1989)	World Education Research Association (2009)
Italian Society of Pedagogy (1989)	Multidisciplinary Education Research Association (Spain) (2011)
Portuguese Society of Education Sciences (1990)	Slovenian Educational Research Association (2011)
Czech Educational Research Association (1992)	South African Education Research Association (2013)
Mexican Council of Education Research (1993)	
European Educational Research Association (1994)	
European Science Education Research Association (1995)	
Flemish Forum for Educational Research (1995)	
Lithuanian Educational Research Association (1999)	

Source: World Educational Research Association Web site (www.weraonline.org).

Educational Research Association in 2011 are further indicators of the ongoing expansion of the field of education research in South Asia, Latin America, and Eastern Europe.

The pattern in the emergence of the top journals in education research also provides an indicator of the international footprint of this field. Table 2 shows the year of establishment of education research journals rated in the top 100 worldwide by TR in 2012. There was an early emergence of these top journals in the United States, with 10 being established before 1950 and comparable numbers of top journals (16) established in each of the last two-quarters of the twentieth century. In contrast, the establishment of what are now top journals outside of the United States occurred later in time (only two before 1950), with 31 of 50 top ranked journals founded in the last quarter of the twentieth century. That the top 100 journals are distributed in equal numbers inside and outside of the United States reflects the global reach of this field.

The proliferation of education research globally is further illustrated in Figure 2 that shows the pattern of growth in the number of citable documents in education research journals in four regions of the world between 1996 and 2012. As noted previously, the impact of an individual journal is measured in part by the total number of published items that could be cited in a given period. Although North America, heavily influenced by contributions from the United States, leads the world in the total number of citable documents published in education journals, Figure 2 shows an overall positive trend in the production of education scholarship across all four world regions. During the first decade of the twenty-first century Asia, North America, and Western Europe showed similar trajectories of growth in the total number of citable documents in education journals. Between 2005 and 2011 the number of citable documents in Latin American education journals also increased. However, in Latin America, both the overall number and the rate of growth in citable documents have consistently fallen below the other regions.

Conclusion

Education research as a field encounters many of the stresses and strains that can beset the social sciences as scientific fields. These sciences are generally not treated comparably to other

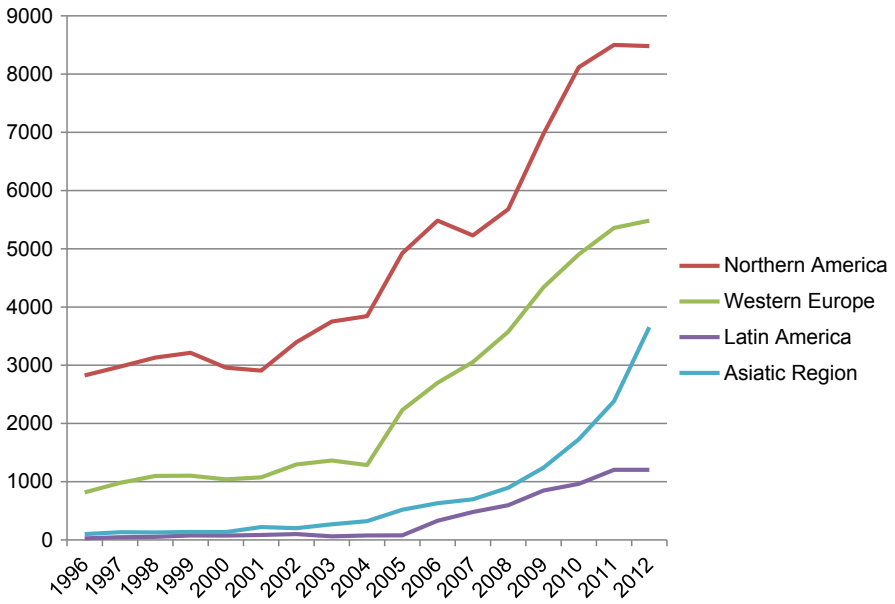

Figure 2 Number of citable documents in education journals by region (1996–2012). Source: SJR, SCImago Journal & Country Rank. http://www.scimagojr.com (accessed 05.02.14.).

fields in the natural sciences. The reasons for this status hierarchy are quite complex, but in part may be explained by the fact that social phenomena are often subjects about which laypersons have opinions or perceptions that are strongly held and are not seen to require special training or expertise. The fact that the study of social and behavioral phenomena are not part of elementary or, in the main, secondary school science education reinforces the perception that these fields are lesser sciences (see Levine et al., 2004).

The investment by the US federal government in education research is one indicator of the level of support. As with the social sciences more generally, the federal investment in education research is comparatively low. The overall nondefense budget for research and development in the United States in 2012 was approximately $64 billion (White House Office of Science and Technology, 2013: Table 1). In 2012, the Institute of Education Sciences had a budget of $190 million for research (excluding the regional laboratories) and an additional $110 million for data and statistics through the National Center for Education Statistics (excluding the National Assessment of Educational Progress) (U.S. Department of Education, 2013). The Educational and Human Resources Directorate at the NSF, the primary source of support for education research at the NSF, had a similar budget allocation of $220 million in 2012 (NSF, 2013). Even considering other sources of support at the NSF and the National Institutes of Health, the investment in education research is likely no more than 1% of the total nondefense R & D.

Yet, the study of education throughout the last century and into this century continues to be an arena of pervasive interest. There is wide recognition across countries and cultures of the social and scientific importance of understanding the dynamics underlying education and learning. Education is generally understood to be central to the critical thinking and skills of a population, to a productive and internationally competitive economy, and to a committed citizenry. Even in countries where education may not be universally available or is closed in access to certain segments of a population, leaders recognize that understanding how to educate is an essential function of society.

This pervasive interest can be seen in the social organization of education research as a field and profession and in the quite separable interest in research on education often evident in other fields and disciplines. Just as education is ubiquitous in society so too is education research. Nonetheless, the presence of a dedicated field offers a venue for incremental building of bodies of knowledge, interaction within and across networks of scholars, and a framework for the intergenerational transmission and refinement of methods and substance that would not readily evolve across disparate disciplines or fields.

Education research because of its commitment to an educative process is in many respects a grand discipline within social science research fields (see also Kalantzis and Cope, 2014). Some of the current arenas of inquiry that are capturing the imagination of those engaged education research include open access to knowledge and its effects on the speed with which knowledge is produced and used, education as a measurable societal asset and as a producer of social equality and inequality in society, learning analytics in a world increasingly dominated by digital environments, observational and time-use recording to unravel the teaching and learning exchange, the effects of learning environments (including through the use of technology) on building analytic skills and brain capacity, and the use of information that is produced and available in rapid and real time (known by some as 'big data'). The importance of some of these scientific issues is shaping the emergence or transformation of education research doctorate programs, specialized associations, and journals (e.g., the emergence just in the last decade of learning sciences programs, dedicated associations, and journals).

Part of the measure of any field is its capacity to build and produce useable knowledge, to create opportunities for interaction even where crossovers do not yet exist, and to itself be reflective and capable of change. As emphasized in this article, education research has always been defined and marked by a porous identity. While at times a point of concern and critique, this dynamic portends well for this field as it moves into its second century as a 'work in progress' worldwide.

End Notes

1. During the twentieth century, the field was dominantly called 'educational' research, but 'education' research has become the more contemporary term in the twenty-first century, and thus is used in this article, even when addressing the history of the field.
2. Two examples of areas of methodological contributions that emanated largely from education research are meta-analysis (Glass, 1976) and hierarchical linear modeling (Raudenbush and Bryk, 1985).
3. Both scholarly (Lagemann, 1997; Shulman, 1999) and institutional reports (NRC, 2002) have noted that education research has been critiqued as lacking a 'core' knowledge base and for the absence of consensus on theory and method. Some attribute these critiques to the interdisciplinary nature of the field and the complexity of building a 'distinctive research community' (Lagemann, 1999: 10–11). While observing these challenges, others (Feuer et al., 2002; Levine, 2003; Ranis and Walters, 2004) have emphasized that the strengths and weaknesses of education research are not unique to it or to the social sciences.
4. The reasons why are not clear, although there is considerable conjecture that the continued engagement of academics in training teachers and the low status of the teaching profession may have affected perceptions of education research or distinguishing it from education (Lagemann, 1997, 2000; Labaree, 2003).
5. NADER was not the first education research association in the United States. In 1901, the National Society for the Study of Education (NSSE) was founded to promote scholarly and scientific research on education through convening meetings and publishing a yearbook. NSSE generally held meetings with the National Education Association (NEA), but was not part of NEA. It was dissolved in 2008. See https://nsse-chicago.org.
6. This overview draws upon the history of AERA by Mershon and Schlossman, 2008.
7. Since 2009, the AERA has held an undergraduate conference within the AERA Annual Meeting, with close to 200 applicants for approximately 20 openings.
8. This article focuses on the field of education research and the preparation of research professionals through education research doctorate degrees. Also, graduate schools of education variously offer practice doctorate degrees, but the doctoral training of practitioners is not the focus on this article.
9. During the last quarter of the twentieth century, graduate schools of education increased their attention to offering education research training. In a 1964–65 survey by Sieber and Lazarsfeld (1966) only a small number of graduate schools (less than 30%) emphasized research.
10. The study, An Assessment of Education Research Doctorate Programs, funded by the National Science Foundation (NSF DRL-0836856), is directed by Felice J. Levine and Lorrie A. Shepard, under the auspices of the AERA and the National Academy of Education. Prior studies are few in number. For example, in 1964 Sieber and Lazarsfeld undertook an important set of studies of education research in universities that included doctoral education (see, e.g., Lazarsfeld and Sieber, 1964; Sieber and Lazarsfeld 1966). In 1982 Schneider led a study of doctoral programs in education under the auspices of the then Deans' Network (Schneider, 1987; Schneider et al., 1984).
11. As Levine and Shepard (2008) noted, any empirically based taxonomy will change over time as new areas emerge or others recede in a field. The DocStudy taxonomy is based on the configuration and frequency of the dominant types of education research doctorate programs in 2010.
12. See *Doctorate Recipients from U.S. Universities: 2012* (S&ED, 2012). These data derive from the Survey of Earned Doctorates (SED) based of an enumeration of all individuals receiving a research doctorate from an accredited institution in a given academic year. The SED designates certain fields as education research. The DocStudy (Levine and Shepard, 2008, 2013) offers a more refined classification and estimation. Nevertheless, the SED data provide an aggregate estimate over time that is informative.
13. Table 24 reports on US citizens and permanent residents only.
14. The AERA, for example, has established a programmatic emphasis on research directed to producing knowledge about the field and the education research profession.
15. AERA membership data tables are available from the first author.
16. This estimate is based on the number of doctorates produced in education research annually (2500) adjusted for a 30% attrition (750) of those who do not enter or leave the field over an estimated career span of 30 years.
17. A large number of international members are not international affiliates, but regular or affiliate members or student members. Over the period 2009–13, approximately 15% of AERA members were from locations outside of the United States, with approximately 5% in Canada and 10% in other countries.
18. While, as noted above, there are distinct doctorate education programs within the fields of education research, they share in common being part of an overarching field.
19. For example, 77.2% of political scientists and 82.1% of sociologists who earned doctorates in 2012 were employed in the postsecondary education sector, while only 47.8% of psychologists and 59.4% of economists in the same cohort moved into this sector (S&ED, 2012: Table 61).

20. Membership data were provided by Karen Edwards, ASA Director of Membership and Publications.
21. The number of faculty in such programs was estimated at about 8000 by Levine and Shepard (2008).
22. See http://ies.ed.gov/ncee/edlabs. RELs are operated under contracts with universities (e.g., REL Southeast by Florida State University) and research organizations (REL Midwest by AIR).
23. As Table 2 further indicates, associations outside of the United States do not evidence the same pattern. This pattern may reflect publishers displaying greater interest in establish journals to reach non-US audiences in the last quarter of the twentieth century.

See also: Associations, Scholarly; Discipline Formation in the Social Sciences; Doctoral Education; Higher Education Research; International Organisations and Education; Journals, Scholarly; Social Science Infrastructure: North America (Resourcing and Funding); Social Science Infrastructure: North America (Social Science Academies and Related Organizations).

Bibliography

Brint, S., 1994. In an Age of Experts: The Changing Role of Professionals in Politics and Public Life. Princeton University Press, Princeton, NJ.

Calhoun, C., 2007. Sociology in america: an introduction. In: Calhoun, C. (Ed.), Sociology in America: A History. University of Chicago Press, Chicago, pp. 1–38.

Calhoun, C., Rhoten, D., 2010. Integrating the social sciences: theoretical knowledge, methodological tools, and practical applications. In: Frodeman, R., Thompson Klein, J., Mitcham, C. (Eds.), The Oxford Handbook of Interdisciplinarity. Oxford University Press, Oxford, pp. 103–118.

Feuer, M., Towne, L., Shavelson, R., 2002. Scientific culture and educational research. Educational Researcher 31, 4–14.

Furner, M.O., 1975. Advocacy and Objectivity: A Crisis in the Professionalization of American Social Science, 1865–1905. The University Press of Kentucky, Lexington, KY.

Glass, G.V., 1976. Primary, secondary, and meta-analysis of research. Educational Researcher 5, 3–8.

Haskell, T.L., 1977. The Emergence of Professional Social Science: The American Social Science Association and the Nineteenth Century Crisis of Authority. University of Illinois Press, Urbana, IL.

Kalantzis, M., Cope, W., 2014. 'Education is the new philosophy', to make a meta-disciplinary claim for the learning sciences. In: Reid, A.D., Hart, E.P., Peters, M.A. (Eds.), A Companion to Research in Education. Springer, New York, pp. 101–115.

Labaree, D.F., 2003. The peculiar problems of preparing education researchers. Educational Researcher 32, 13–22.

Lagemann, E.C., 1997. Contested terrain: a history of education research in the United States, 1890–1990. Educational Researcher 26, 5–17.

Lagemann, E.C., 1999. An auspicious moment in educational research. In: Lagemann, E.C., Shulman, L.S. (Eds.), Issues in Education Research: Problems and Possibilities. Jossey-Bass, San Francisco, CA, pp. 3–16.

Lagemann, E.C., 2000. An Elusive Science: The Troubling History of Education Research. University of Chicago Press, Chicago.

Lagemann, E.C., Shulman, L.S. (Eds.), 1999. Issues in Education Research: Problems and Possibilities. Jossey-Bass, San Francisco, CA.

Lazarsfeld, P.F., Sieber, S.D., 1964. Organizing Educational Research. Prentice-Hall, Englewood Cliffs, NJ.

Levine, F.J., 2003. Doctoral Programs in Education Research: Comparative Perspectives in Defining the Territory, unpublished presentation before the Committee on Research in Education of the National Research Council, Washington, DC, November 12, 2003.

Levine, F.J., Abler, R.F., Rosich, K.J., 2004. Education and Training in the Social, Behavioral, and Economic Sciences: A Plan of Action. Report to the National Science Foundation, Washington, DC. http://www.nsf.gov/pubs/2004/nsf0442/nsf0442.pdf (accessed 06.11.14.).

Levine, F.J., Bell, N.E., 2015. Social Science Professions and Professionalization. In: The International Encyclopedia of Social and Behavioral Sciences, second ed. http://mrw.elsevier.com/isb2/menu.htm.

Levine, F.J., Shepard, L.A., 2008. An Assessment of Education Research Doctorate (DRL-0836856). Proposal Submitted by the American Educational Research Association (AERA) and the National Academy of Education (NAEd) to the National Science Foundation, May 19, 2008.

Levine, F.J., Shepard, L.A., 2013. Study Overview and a Topographic Map of Research Programs in U.S. Institutions, unpublished powerpoint presentation to the Council of Academic Deans from Research Education Institutions (CADREI), Naples Florida, September 30, 2013.

Mershon, S., Schlossman, S., 2008. Education, science and the politics of knowledge: The American Educational Research Association, 1915–1940. American Journal of Education 114, 307–340.

National Research Council (NRC), 2002. Scientific research in education. In: Committee on scientific Principles for education research, Shavelson, R.J., Towne, L. (Eds.), Center for Education. Division of Behavioral and Social Sciences and Education. National Academy Press, Washington, DC.

National Science Foundation (NSF), 1999. Doctorate Recipients from United States Universities: Summary Report 1997. National Science Foundation, Washington, DC. http://www.nsf.gov/statistics/doctorates/summ97/ (accessed 05.20.14.).

National Science Foundation (NSF), 2013. National Science Foundation Budget Request. National Science Foundation, Washington, DC. http://nsf.gov/about/budget/ (accessed 06.10.14.).

Ranis, S., Walters, P.B., 2004. Education as a contested enterprise: the deliberations of the SSRC-NAE Joint Committee on Education Research. European Educational Research Journal 3, 795–806.

Raudenbush, S.W., Bryk, A.S., 1985. Empirical Bayes meta-analysis. Journal of Educational Statistics 10, 75–98.

Ross, D., 1991. The Origins of American Social Science. Cambridge University Press, Cambridge, UK.

Reuters, T., 2013. Journal Citation Reports, Social Science Edition. http://thomsonreuters.com/journal-citation-reports/ (accessed 05.10.14.).

S&ED, 2012. Doctorate Recipients from U.S. Universities, Science and Engineering Doctorates: Detailed Statistical Tables. National Science Foundation, Division of Science Resources Statistics, Washington, DC. http://www.nsf.gov/statistics/sed/2012/archives.cfm (accessed 05.19.14.).

Schneider, B., 1987. Tracing the provenance of teacher education. In: Popkewitz, T. (Ed.), Critical Studies in Teacher Education. Falmer Press, Philadelphia, PA, pp. 211–242.

Schneider, B.L., Brown, L., Denny, T.B., Mathis, C., Schmidt, W., 1984. The deans' perspective on the status of doctoral programs in schools of education. The Phi Delta Kappan 65, 617–620.

SJR, 2014. SCImago Journal and Country Rank. http://www.scimagojr.com (accessed 05.02.14.).

Shulman, L., 1999. Professing educational scholarship. In: Lagemann, E.C., Shulman, L.S. (Eds.), Issues in Education Research: Problems and Possibilities. Jossey-Bass, San Francisco, CA, pp. 159–165.

Sieber, S.D., Lazarsfeld, P.F., 1966. The Organization of Educational Research in the United States. Columbia University Bureau of Applied Social Research, New York.

U.S. Department of Education, 2013. Education Department Budget History Table, FY 1980-2014. Washington, DC. http://www2.ed.gov/about/overview/budget/history/index.html (accessed 06.10.14.).

White House Office of Science and Technology Policy, April 10, 2013. The 2014 Budget: A World-Leading Commitment to Science and Research. Washington, DC. http://www.whitehouse.gov.ostp (accessed 06.10.14.).

Educational Sociology

Lawrence J Saha, Sociology, College of Arts and Social Sciences, Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Educational sociology focuses attention on the social factors that both cause and are caused by education. It includes the study of factors relating to education, such as gender, social class, race and ethnicity, and rural–urban residence. Educational sociology has developed a range of sociological theories to explain and guide research into the various levels and types of education, and it has also contributed to the development of methodological and statistical techniques. Three current issues are discussed as examples, namely the relationship between education and social inequality, school violence and bullying, and teacher burnout and the accountability movement.

Introduction

Educational sociology is one of the major subfields in sociology and also in the development and production of educational research and teaching. Because of the enormous breadth and scope of what is called 'education' in modern society, the sociological study of education covers a wide field and is perhaps one of the largest subfields within sociology. As Ballentine and Hammack (2012) point out, education starts with birth and does not end until death. In one form or another, either formal or informal, it takes place throughout one's life. However, sociologists tend to direct their attention to the type of education that takes place in formal institutions which are designated by society to learning in an organized, systematic manner. Thus the sociological study of education, and what is called 'educational sociology,' is generally restricted to education in a formal sense, namely schools, colleges, and universities. Through these institutions, society ensures that a standard body of knowledge, skills, and values get transmitted to future generations.

Emile Durkheim, one of the first sociologists who studied formal education in a systematic manner, argued that formal education is the way that society guarantees its own survival (Saha, 2001). This does not mean that the process of education takes place seamlessly, and that the desired body of knowledge and values do not change, or that it is passed on without contestation. Durkheim himself understood that few societies are in agreement about what knowledge should be passed on through formal education. The divisions within society, such as social class, religion, gender, and race/ethnic groups, usually have some unique preferences about what knowledge should be passed on to the young. As a result, educational policy, particularly in modern multiethnic societies, is usually a source of much disagreement. These disagreements not only mean that much compromise often takes place in the establishment of formal educational systems, but that often two or more different types of schools become established, each passing on at least some unique knowledge and values to those students who attend them. This explains why many modern societies consist of a government school system and one or more private school systems. These are some of the kinds of phenomena that sociologists attempt to understand through their research.

Educational Sociology or Sociology of Education

Two labels have dominated the sociological study of education through its academic development in the English-speaking world. Perhaps the older of the two is 'educational sociology.' Indeed this was the label which was used for the title of the first dedicated journal in the United States, *The Journal of Educational Sociology*. The journal originally was established outside of the official professional academic sociological association, the American Sociological Association (ASA). The journal content reflected the two directions that sociologists took in their study of education. Some were concerned about educational policy and how education 'ought to be' in American society. Others, however, thought that the sociological study of education should be more 'scientific,' and thus more concerned with 'what was actually happening' in educational institutions. However, when a move was made to incorporate the journal into the official publications of the ASA, the members of the latter perspective prevailed, and the name of the journal was changed to *Sociology of Education*. The two labels continue to be used, sometimes interchangeably. For the most part, however, the former tends to be used by sociologists in academic departments of education, while the latter tends to be used in departments of sociology. In the end, both approaches are necessary, as sociologists need to address both the factual and 'scientific' operations of educational systems as well as the policies needed to administer them (Dworkin and Ballantine, 2012). In this article, the labels are used interchangeably.

The sociological study of education experienced rapid growth in the early history of the discipline. All of the 'founding fathers' of sociology, namely Emile Durkheim, Max Weber, and Karl Marx, had something to say about education, although it was Durkheim who was most responsible for establishing the field.

There have been critics of educational sociology throughout its development. For example, some have referred to it as "ivory tower nonsense" (Hansen, 1967) and have noted that it has been characterized by the "sterility of the questions that it asked" (Corwin, 1965). However, more recently it has been considered as "extremely rich, vibrant and diverse" with "an exciting range of substantive issues for investigation, as well as an equally exciting range of theoretical considerations and

methodological tools for undertaking these investigations …" (Noblit and Pink, 1995: p. 27).

But as the field continues to expand into the twenty-first century, better theories and methodological tools are opening the door to more rigorous research. Hallinan, for example, contends that "sociologists of education must conduct theoretically rich research that identifies mechanisms that link exogenous and endogenous factors to student outcomes" (Hallinan, 2011: p. 2).

Theories and Methods in Educational Sociology

Like other areas in sociology, the development of theory and research methods implies a continuing vitality in the field. This has certainly been true for educational sociology. Although the traditional theories emanating from the 'founding fathers' of sociology continue to be important in the study of education, new perspectives continually emerge which identify, clarify, and explain educational phenomena which are unique in an increasingly geographical mobile and technologically sophisticated world population.

Theories

The early classical sociological theories, primarily from Durkheim and Weber, made it clear that formal education, that is schools and universities, played a vital role not only in passing on the social and cultural heritage of society, but also in the maintenance and formation of social inequalities. In the case of Durkheim, these inequalities could be seen as functional insofar as they helped recognize innate talent and motivation among students. For Weber these inequalities which emerged from schooling had more to do with the cultural and political factors which already divided society. Although Marx never developed a coherent theory of education in his critique of capitalism, his ideas eventually led to a very important neo-Marxist educational perspective which focused on education as a form of 'symbolic violence' whereby the social class structure is reproduced through educational processes (Althusser, 1971; Bourdieu and Passeron, 1977). More importantly, this reproduction process has subsequently been found to occur with the corporation and compliance of the students from disadvantaged backgrounds (for example, see Giroux, 1983; Willis, 1977).

These theoretical perspectives focused primarily on the structures, both formal and informal, which divide society. There emerged a completely different (and opposite) theoretical perspective which pointed to micro-level interactions and subjective constructions related to the educational processes. Drawing on seminal work of the early interactionist theorists, such as George Herbert Mead, Charles Horton Cooley, Herbert Blumer, and Irving Goffman, a major sociological perspective was launched by Berger and Luckman (1967), who argued that reality was constructed through people's creative action. In the study of education from a sociological perspective, labeling theory became a major theory which explained educational processes and outcomes in terms of the interactions between all parties involved in the education of young people. For example, teachers and students have expectations of one another's behavior, and these have consequences on one another's behavior. The important work of Rosenthal and Jacobson (1968) has generated a large body of research which has investigated the impact of teacher labeling. However, as Rubie-Davies (2009) has pointed out, the characteristics of the teachers are as important in understanding this process as are the characteristics of the students. Most importantly, however, this rich body of research illustrates the importance of the interactionist perspective for research, which has helped understand educational processes in the classroom. Because these theories focus on the interaction of two or relatively few persons, the body of theories which fall into this category are referred to as micro-theories. Those which focus on organizational structures, or on the major group divisions in society, are referred to as macro-theories.

The sociological study of education has been influenced by many newer and more general theoretical perspectives. Indeed, it can be said that almost every theory in the social sciences has been used to investigate educational processes of one kind or another. These theories often occur in schools or interrelated clusters which are dominated by a particular underlying perspective. For example, in the 1970s a perspective called the 'New Sociology of Education' emerged which shifted the focus from educational student outcomes to the knowledge content of school curricula and who controlled it (Young, 1971). This critical approach provided not only the dual purpose of 'exposing' what really occurred in the structure and process of schooling, but also the liberation and empowerment of those students (and sections of society) who were dominated and disadvantaged by it. This perspective capitalizes on a "critical reflection which liberates or emancipates actors from false beliefs and subsequently leads to concrete proposals for overcoming oppression" (Lakomski, 1997: p. 57). Thus there is not one critical theory, but a range of critical theories, each with its own focal point depending on what aspect of schooling is being examined. These theoretical versions appear in educational sociological subfields such as the curriculum, educational administration, teacher education, educational policy, and planning, to name just a few.

One important critical theoretical perspective is reflected in feminist theories. Feminist theories not only challenge the male bias of most educational research, but they argue that as a consequence of this bias women's experiences in schools are often neglected. Some of these concern the issues of social justice and differential treatment and opportunities in both educational experiences, but more fundamentally the focus has been on educational policies (Arnot and Weiler, 1993).

What is important, however, is that these approaches not only are designed to 'uncover' hidden features of what takes place in schools, but by so doing, they provide the tools to bring about change and reform. These theories stand in contrast to research and explanations emerging from the functional, conflict, and interactional perspectives which attempt to describe and explain educational processes.

A final influential theoretical perspective has influenced many researchers, and it focuses on the wider macro-processes which impact on the increasing expansion of standard organizational and structural schooling models which have rapidly spread, and continue to spread, throughout the world. The institutional and global approach, according to Benevot

(1997), consists of three themes: (1) a focus on the origins and expansion of mass secular systems of schooling throughout the world in the twentieth century; (2) a focus on the institutional aspects of modern education, such as its selection function, its importance in the allocation of social status; and finally (3) the impact of education on other structures and institutions of society, such as gender roles, the nature of work, and even aspects of the nation-state, and the relations between nation-states. Because of the historical and macro-foci of the perspective, it is closely related to theories of globalization.

The spread of mass secular schooling suggests the emergence of a world culture which includes universal models of education systems, and in particular schooling. Meyer (1997) argues that the modern nation-states, and their subsequent educational systems, are shaped by common values relating to education, the nation, and the state. Robertson and Dale (2008), however, argue that these globalizing forces include additional fundamental factors such as international economic competitiveness and an increasingly interconnected global economy. Thus the social and economic values placed on education and mass schooling make this social institution vulnerable to these forces, which in turn are manifested through increasing standardization and accountability measures.

It has been argued that the institutionalization and globalization of schooling impacts not only the structures and processes of school systems, but also the values and expectations of the students within them. Sikora and Saha (2010) argue that the high levels of occupational expectations found across societies, in particular the developing societies, are a reflection of the globalization of the link between education and jobs. Using Programme for International Student Assessment (PISA) data from 50 countries, they found that high levels of occupational expectations were particularly noticeable in the less developed countries. Yet these countries are the recipients of considerable aid through international and nation-state agencies. Ironically, Sikora and Saha argue that this aid might itself be a conduit of the globalizing process which drives the expansion of a standard model of an educational system.

Looked at from a broad perspective, the theories which guide research in educational sociology are broad in their scope, but rich in the insights that they make possible in understanding the many faces of educational systems and the processes which occur within them.

Research Methods

The sociological study of education has always used research methods adopted from mainstream sociological research. However, because of the unique challenges in the study of educational systems and processes, given that students are nested in classes, and in schools, the discipline has also been the center for the development of new research methods, particularly in quantitative data analysis, which have eventually made their way into mainstream sociology. The following section provides an overview of the main characteristics of research methods used in this field.

From the very beginning, the methods used in educational research incorporated both qualitative and quantitative approaches. One only has to examine the methods of the early 'classics' in educational sociology to observe this practice. Early studies of education processes tended to be 'holistic' in nature, such as a study of a school, a particular social group, or education in a particular community, and in which both qualitative and quantitative data were used. This was the case in studies in many countries. For example, Hollingshead's *Elmstown's Youth* (1947) and Coleman's *Adolescent Society* (1961) represent path-breaking sociological studies of American education over 50 years ago, and they are regarded as relevant even today. Similarly, in the UK Hargreaves' *Social Relations in a Secondary School* (1967) and Ford's *Social Class and the Comprehensive School* (1969) used multiple methods in their holistic studies of single schools, and they also produced research which has had long-lasting impact on British educational sociology. In France a classical monograph of this period was Bourdieu and Passeron's *Reproduction in Education, Society and Culture* (1977), which not only introduced the concept of culture capital to sociology, but was also responsible for the development of reproduction theory, which was to become one of the dominant theories in educational sociology. These, and other similar studies in the early years of educational sociology, were to highly influence research for several decades, until the emergence of more sophisticated computer technology and statistical packages made possible the analysis of educational systems and processes in an entirely different manner. But qualitative research methods also continued to develop, but in a very different manner.

In her article of trends in qualitative research in educational sociology, LeCompte (1997) describes how new approaches developed even while the old methods, largely influenced by anthropological and evaluation research, remained frequently used. However, according to LeCompte, what fundamentally changed was a move away from positivistic qualitative studies, based on methods such as observation and other forms of field work, to research based on critical theory and postmodernism. With this shift, the new focus has become the perspective and narratives of the subjects, as the objects of study, rather than that of the researcher. In other words, by focusing on the 'voices' of the subjects of research, the categories and interpretations are those of the teachers and students, for example, rather than those of the researcher. However the assumption is that teachers and students have 'stories' or narratives to tell and to analyze, and the interpretation of their presence or absence becomes problematic. Examples of these kinds of problems are found in studies where the informant is low income and mentally handicapped or where the informants have no 'stories' at all to tell (LeCompte, 1997).

Thus, research based on stories or the life experiences of individuals, and the analysis of narratives and biographies have opened new paths and challenges for qualitative research. According to LeCompte (2009) these shifts in qualitative methodologies can be traced in the series of handbooks of research on teaching produced under the sponsorship of the American Education Research Association from 1963 to 2001. While reflecting a 'sea change' during these years in theory and research in most social science disciplines, new epistemological issues emerged as researchers began to ask 'Whose story is being told?' and whether the researchers have got the stories right (LeCompte, 2009: p. 39). One of the driving forces of

continued attempts to develop qualitative research methods has been desired to avoid a 'top-down' imposition of meanings and understandings of educational phenomena and replace it with new ways of researching which are more "collaborative, participatory, and action-oriented" (p. 41).

While in recent years qualitative research methods have continued to make a valuable contribution to research in educational sociology, so too have the developments in quantitative methods. For a period of time, particularly in the 1980s and 1990s, developments in qualitative methods tended to overshadow those in quantitative research. As LeCompte observed, "Qualitative research was embraced by the educational community with such enthusiasm that in some academic quarters, quantitative research was relegated to the background, caricatured as unable to address real issues concerning real people" (LeCompte, 2009: p. 43).

However, during that period users of quantitative methods were not standing still, and developments in statistical techniques and modeling began to address some of the challenging contexts that educational research presented, namely, how do you study education processes and outcomes of students who are nested in classrooms, sometimes taught by different teachers throughout the day, all within different schools, and across different countries? To further complicate this challenge, quantitative techniques also had to be able to incorporate the educational effects of a myriad of family background characteristics as well as the effects of friendship groups and community contexts. Fortunately, breakthroughs in statistical packages, which emerged in the 1970s, began to reap returns just as qualitative methods were reaching a kind of epistemological endpoint in pushing that research deeper into the conscious world of individual students.

Quantitative research in education is divided into two main groups, depending on whether the research design is experimental or not. Researchers who come through a more psychological or educational psychological tradition tend to favor analyses which involve experimental designs based on a treatment group and a control group. The advantage of this approach is that a single variable can be manipulated in such a way that its causal effect on the outcome can be tested. Researchers from this perspective tend to use analysis of variance, or a variation of it, as their primary statistical tool. This statistical procedure has persisted for generations of researchers, and continues to be popular today.

However, for researchers who take a more sociological approach in the use of quantitative data, the use of survey designs is more typical. The primary difference between experimental and survey designs is that the latter tend to have only one study group, the research participants tend to be larger in number, and there are more variables included in the analysis. However, unlike experimental designs with control and treatment groups, survey designs are multivariate and the isolation of a single causal variable is replaced by multivariate and assumed multicausal models. Nevertheless, the development of new statistical techniques and the increasing frequency of longitudinal survey designs make both the identification of the multicausal effects easier to identify and the disaggregation of the between-student, between-class, between-school, and between-country effects possible. While correlational analysis and multiple regression techniques remain widely used, they are often limited by conditions where demonstrated causality or the disaggregation of effects between levels is less important.

One attempt to solve the problem of multicausality emerged in the 1960s with the development of path analysis. The use of path models was stimulated by an interest in the factors which explained occupational attainments. Following Blau and Duncan's study of occupational attainment among adults in the United States, Sewell and his colleagues at the University of Wisconsin used path analysis to study the attainments of high school graduates who were then followed through university and subsequent occupational careers. What made the work of the so-called Wisconsin model so important for educational sociology was the inclusion of educational variables. For over a decade the use of path analysis to study the determinants of both educational and occupational attainments became dominant among quantitative educational sociologists. Further developments of this approach, using partial least squares (PLSpath) and maximum likelihood techniques (LISREL), made it possible to create a latent variable from a number of observed variables, thus increasing the number of variables and refining their relationships and meanings for path models. However, for the study of school students, there remained the problem of nested data and problems of differentiating between the student, classroom, school, and, in the case of comparative studies, country-level effects.

Major breakthroughs in solving these latter statistical challenges began to occur in the 1980s with the development of hierarchical linear modeling through which the overall ability to explain educational or occupational attainments could be disaggregated into the various nested levels. Further developments which incorporated both path analysis and multilevel modeling have emerged, and procedures such as AMOS, MPLUS, and STREAMS have been developed and are increasingly being used by quantitative educational sociologists, with the result that there have been pressures to collect higher quality educational data which these procedures demand. Overall, these developments have greatly improved the understanding of the complex relationships and multiplex determinants of educational processes and attainments.

Perhaps equally important has been the gradual adoption of these highly sophisticated and complex procedures into mainstream sociological research. Thus advances in the quantitative research into education, to meet unique conditions in educational research, have been of benefit to sociology as a whole. For a more detailed description of the history and development of these procedures, and their importance for educational sociology, see Keeves (1997), Saha and Keeves (2003), and Keeves and Darmawan (2009).

Apart from the above types of research techniques, there are other approaches which have investigated education from a completely different perspective, namely the importance of the interconnection between individuals and the benefits or disadvantages they bring to educational attainments. Network analysis is a growing type of statistical analysis which makes possible the study of the effect of friendship and peer groups among students on educational attainments (Klovdahl, 1994; Saha, 1994). Theoretically, network analysis can be linked with the notion of social capital, a concept which originated in the

work of Coleman, which focuses on the human resources that individuals possess that can be used for various social attainments, including education (Coleman, 1988). For example, Rizzuto et al. (2009) found, in their study of American university students, that the connectivity (network density) of students in a university classroom had a significant, though not the most important, effect on academic performance. In an innovative use of archival data, Eckles and Stradley (2012) found that students who were closely interconnected through residence and classes were more likely to remain at university, compared to those who dropped out, who were less socially connected. These two recent examples illustrate the utility and importance of social network analysis in educational research.

As this exposition has made clear, the theoretical and methodological approaches within educational sociology are varied and diverse. At times this diversity has resulted in conflict within the field. On the other hand, recent discussions of this diversity have seen it as a strength rather than a weakness. For example, Weis and her colleagues, in documenting the struggles between competing theories and methodologies, argue that working across the divisions, rather than maintaining boundaries, will be more productive for the discipline in the long run (Weis et al., 2009).

Some Current Topics in Educational Sociology

As should already be apparent, educational sociology is a broad and inclusive subarea of sociology. Because of the importance of education in society, hardly any aspect of social life is unaffected by it. The topics which are studied and researched are almost endless, ranging from early forms of education through socialization and preschool experiences, through formal educational experiences, both academic and vocational, and finally to forms of lifelong learning which follows the completion of formal educational studies. Furthermore, these educational experiences are often studied in the context of family background characteristics, friendship and peer group experiences, and other similar social contexts. It would be impossible here to summarize issues relating to these topical areas in this brief exposition. Therefore, three current key topics will be examined here to illustrate recent work by educational sociologists. These are education and social inequalities, school violence and bullying, and teacher burnout and the standardized testing and accountability movement.

Education and Social Inequalities

The direct relationship between education and various forms of social inequality has probably been the earliest and most widely theorized and researched topic in educational sociology. The general pattern in all societies is that the more education a person has attained, the greater will be other valued social attainments. But educational attainment is not equally distributed, so that those groups which acquire more education also enjoy more social benefits. The dimensions of these attainment inequalities generally are many, ranging from gender, ethnic and race origins, rural–urban residential locations, government and private school attendance, types of family socialization practices, and social and cognitive disabilities, to name just a few. What has been most striking is that, with the exception of gender inequalities, most of the traditional inequalities have persisted in spite of much research on education and educational reforms. Furthermore, these inequalities exist across societies.

Early theorizing about education and inequality was based on the functional perspective which argued that, because a social role of education is to sift, sort, and allocate the most talented and motivated students into the most important and lucrative social positions, the inequalities are a product of merit, and a sign that the education system is working. The complex process by which this comes about has been labeled the status attainment model. Basically the model assumes that there are many determinants of how much education a person attains, including home background factors, aspirations and expectations, and prior attainments. Then the amount and type of education acquired lead to other occupational and social attainments.

However, because the meritocratic process, studied in this way, always seemed to favor students from privileged backgrounds, countertheories argued that the continued existence of social inequalities in educational systems represented a form of social reproduction which was not merit based, but one based on hegemony, privilege, and power over the education system itself. Starting with Bernstein's research between class-based language codes in England, and their importance in academic achievement (Bernstein, 1971), Bourdieu and Passeron (1977) also found the mechanisms of the reproduction of social class in education in France, not because of a meritocracy but because of the transmission of what he later called 'cultural capital.' Thus the concept of cultural capital, and Coleman's related concept of social capital, provided an alternative explanation for the pervasiveness of social inequality and education.

Empirically, the quantitative studies of the link between education and social inequality found that the relationship existed in most western industrial societies (Erikson and Goldthorpe, 1992), and that class of social origin was a main contributor, although other research found that its effects were diminishing (Dronkers, 1993). Nevertheless the differences between the social classes and education have been found to be persistent across countries (Shavit and Blossfield, 1993).

The direct relationship between education and social inequality has also been found to vary between countries, depending on characteristics of the political and economic systems and also on characteristics of the educational systems. For example, countries with more social democratic systems tend to have lower educational and social inequalities than countries with liberal market economies (Beller and Hout, 2006). Similarly countries which have early selection of students into tracks based on presumed innate ability, or on vocational or academic occupational destinations, compared to countries with nontracked comprehensive education systems, tend to have stronger links between educational attainments and other social attainments (Marks, 2005). Furthermore, it has also been argued that even as the general levels of educational attainment in some societies have appeared to become more equal, for example in the expectation that everyone can attend university and obtain a university degree, there remain hidden factors which bring about negative

and unequal consequences. These emerge when the realistic possibilities among students of attaining goals are not recognized or acknowledged, with the result that failures are worse than if realistic lower level goals were pursued to begin with (Rosenbaum et al., 2011). Thus, ironically, the ideal or perfectionist views about how to reduce or eradicate the link between education and inequality actually have been shown to maintain them, although through hidden processes. Clearly much sociological research remains to be done, and policies need to be developed, at both the school and the political and economic level to reduce the link between education and social inequality, if that is, indeed, the goal of a society.

School Violence and Bullying

Recent years have seen a rapid increase in various forms of violence taking place in schools in many countries of the world. Because schools serve young children and teenage youth, they are seen as places that should be free and safe from various forms of psychological and physical harm. Apart from the disruption that these harmful forms of behavior cause to the teaching and learning process, school students are extremely vulnerable and powerless to protect themselves when confronted with them. Similarly, it has been demonstrated that teachers suffer negative consequences, such as burnout, when they are made to work in an unsafe and threatening school environment. Therefore, sociologists, parents, and education administrators and policymakers have contributed much to our understanding of the factors which contribute to unsafe schools.

Violence in schools can take many forms, ranging from the 1999 shooting at Columbine High School in the United States to physical and emotional abuse between students. While the Columbine High School incident may have been the result of individuals with social and emotional problems, the frequency and diversity of incidents such as these, not only in the United States but elsewhere, poses important sociological research challenges and subsequent policies. For example, the experience of violence among youth generally, whether domestic or in the school, can have disruptive consequences for academic performance among students, as well as their relationships with other, and their own psychological well-being (Huang and Mossige, 2012).

Bullying, in particular, has attracted considerable research attention from both sociologists and psychologists. Although school bullying had been regarded by many as a normal part of school life, ever since the seminal and influential work of Olweus, who began to study bullying in 1970 in Scandinavia, it has become regarded as an important social problem in schools. Olweus' research has focused on both the bully and the victim, and takes into account both psychological and sociological factors (Olweus, 1993). Recent researchers on bullying have also included both sets of variables. For example Cassidy (2009), in his study of adolescent youth in the United Kingdom, found that bully victims were more likely to be girls, were more likely to come from less supportive families, and were more likely to have a lower group identity and to be poorer in problem-solving skills. Bibou-Nakou et al. (2012), who studied Greek secondary school students using a qualitative approach, found that the students themselves thought the school was partly responsible because of the competitive environment it fostered. Thus programs directed at lowering or eliminating the incidence of bullying in schools, such as that put forward in Norway, include primarily sociological variables, such as the promotion of more empathetic relationships between students and strong leadership by the teachers in fostering a more cooperative and an antibullying environment (Roland et al., 2010).

Bullying continues to be considered a serious issue in educational systems throughout the world. Rigby and Smith (2011) argue that, overall, the incidence of bullying has declined, largely because of the various intervention programs which have been put into place to reduce or prevent it. However, because of the growing dominance of electronic technology, new forms of bullying, such as outside-of-school cyber bullying, appear to be replacing the traditional in-school bullying. Research into both in-school and out-of-school bullying continues to expand our knowledge of the causes and consequences of this type of violence among youth. Sociological research has made considerable contribution to the better understanding of bullying in schools, and how to treat and prevent it.

The Accountability Movement and Teacher Burnout

The teacher counterpart to the consequences of bullying among students is teacher burnout. Teachers play a crucial role in educational processes, but at the same time they are caught between a number of conflicting pressures, in particular those relating to the expectations of parents, the indifference or even the hostility of students, and the authoritarianism or unsupportive posture of school principals and other administrators. The teacher has been the object of study virtually since the beginning of the sociological study of education itself, beginning with Durkheim's study of education, especially what he called 'moral education.' Durkheim contended that a fundamental purpose of education was to form children into adults who were fully integrated into society and its norms. For him, the teacher played a central role in this process, especially in providing discipline, integration into social groups, and the gradual development of autonomy (Durkheim, 1961(1925)). Two important early sociological works on teachers were Waller's *Sociology of Teaching* (1932) and Lortie's *Schoolteacher: A Sociological Study* (1975).

Lortie (1975) foreshadowed some of the current problems facing the teaching profession and teachers when he observed that teachers, at least in America, were 'both honored and disdained,' partly because they are deemed to be professionals, but at the same time are paid relatively low wages (p. 10). According to him, they occupy a 'shadowed social standing.' This insight seems to have been predictive of many of the pressures on teachers today in light of their loss of autonomy and professional status, and their increasing subjection to accountability policies about which they have little control. These kinds of pressures have produced what contemporary sociologists and psychologists call 'burnout.'

Although psychologists consider burnout to be primarily a problem internal to the individual, sociologists see it as a product of a work environment which removes the meaning and power from the teacher. Thus Dworkin (1987) has argued for a sociological approach to the study of teacher burnout

which focuses more on the organizational structure of schools and the relationship between teachers and school administrators rather than with students and parents. Thus, rather than focus on the psychological characteristics of the teachers, a sociological approach would focus on aspects of the school, and in particular the administrative style of the principal and other aspects of the relationships between the teachers themselves. Indeed, Dworkin et al. (2003) found that teachers in schools with a more democratic environment and more teacher autonomy were less likely to experience burnout symptoms.

In contrast to the above, one feature of the current globalization process is the increase in the use of standardized testing of students to assess not only the students' performance, but also that of teachers and principals. As Natriello (2009) points out, there are negative consequences for teachers with the introduction of standard testing, with costs to teachers for the low performance of students, thus the name 'high stakes testing.' The introduction of this form of standards-based policy is part of an increasing global movement which is removing teacher professional autonomy and reducing the profession to that of perfunctory bureaucratic personnel in the system. Sociologists who study education are increasingly turning their attention to the changing nature of the teacher profession under these new organizational circumstances.

Conclusion

The sociological study of education is a broad-ranging part of the educational and sociology disciplines. Because of the pervasive nature of education into all sectors of the lives of individuals, and also into the functioning of society as a whole, the discipline is extremely inclusive and there is hardly a sector of society where it is not relevant. Educational sociology has changed extensively since Emile Durkheim began to lecture about it to his students at the Sorbonne. But just as during the early days of sociology, so too today the study of education from a sociological perspective addresses important contemporary issues across societies. Some of these issues are common to all societies, such as those relating to inequality, while others are more unique to the characteristics of individual societies, for example issues relating to social and economic development. Educational sociology occupies a central place in the disciplines of education and sociology. In terms of both theory and research methods, it has made important contributions to the studies of education and also to wider fields of sociological research generally.

See also: Burnout, Psychology of; Coeducation and Single-Sex Schooling; Cognitive Development, Educational Theories of; Coleman, James Samuel (1926–95); Comparative Research in Education: IEA Studies; Defiant Behavior during Adolescence and Cultural Variations; Durkheim, Émile (1858–1917); Education: Anthropological Aspects; Effects of Stress on Development during Childhood; Ethnicity and Educational Achievement; Family and Schooling; Globalization in Education; Marx, Karl (1818–83); Motivational Development, Systems Theory of; Social Inequality and Schooling; Socioeconomic Status and Academic Achievement; Teacher Stress and Burnout: Methodological Perspectives.

Bibliography

Althusser, L., 1971. Ideology and ideological state apparatuses. In: Althusser, L. (Ed.), Lenin and Philosophy and Other Essays. New Left Books, London, pp. 121–176.
Arnot, M., Weiler, K., 1993. Feminism and Social Justice in Education. Falmer Press, London.
Ballentine, Jeanne H., Hammack, Floyd M., 2012. The Sociology of Education: A Systematic Analysis. Pearson, Boston.
Beller, E., Hout, M., 2006. Welfare states and social mobility: how educational and social policy may affect cross-national differences in the association between occupational origins and destinations. Research in Social Stratification and Mobility 24, 353–365.
Benevot, A., 1997. Institutional approach to the study of education. In: Saha, Lawrence J. (Ed.), International Encyclopedia of the Sociology of Education. Pergamon Press, Oxford.
Berger, P., Luckman, T., 1967. The Social Construction of Reality. Penguin, Harmondsworth.
Bernstein, Basil B., 1971. Class, Codes and Control. In: Theoretical Studies towards a Sociology of Language, vol. 1. Routledge and Kegan Paul, London.
Bibou-Nakou, I., Tsiantis, J., Assimopoulos, H., Chatzilambou, P., Giannakopoulou, D., 2012. School factors related to bullying: a qualitative study of early adolescent students. Social Psychology of Education: An International Journal 15, 125–145.
Bourdieu, P., Passeron, J., 1977. Reproduction in Education, Society and Culture. Sage, London.
Cassidy, T., 2009. Bullying and victimization in school children: the role of social identity, problem-solving and family and school context. Social Psychology of Education: An International Journal 12 (1), 63–76.
Coleman, James S., 1961. The Adolescent Society. The Free Press of Glencoe, New York.
Coleman, James S., 1988. Social capital in the creation of human capital. American Journal of Sociology 94, S95–S120.
Corwin, R.G., 1965. A Sociology of Education. Appleton-Century Crofts, New York.
Dronkers, J., 1993. Educational reform in the Netherlands: did it change the impact of parental occupation and education? Sociology of Education 66, 262–277.
Durkheim, E., 1961(1925). Moral Education. The Free Press of Glencoe, New York.
Dworkin, A. Gary, Ballantine, Jeanne H., 2012. Sociology of education in the United States. In: Osipov, A. (Ed.), Sociology of Education around the World. State University of Novgorod Press, Novgorod.
Dworkin, A. Gary, Saha, Lawrence J., Hill, Anette N., 2003. Teacher burnout and perception of a democratic school environment. International Education Journal 4 (2), 108–120.
Dworkin, Anthony G., 1987. Teacher Burnout in the Public Schools: Structural Cause and Consequences for Children. State University of New York Press, Albany.
Eckles, James E., Stradley, Eric G., 2012. A social network analysis of student retention using archival data. Social Psychology of Education: An International Journal 15, 165–180.
Erikson, R., Goldthorpe, J., 1992. The Constant Flux: A Study of Class Mobility in Industrial Society. Clarendon Press, Oxford.
Ford, J., 1969. Social Class and the Comprehensive School. Routledge and Kegan Paul, London.
Giroux, H., 1983. Theory and Resistance in Education. Heinemann, London.
Hallinan, Maureen T., 2011. Improving the interaction between sociological research and educational policy. In: Hallinan, Maureen T. (Ed.), Frontiers in Sociology of Education. Springer, Dordrecht, pp. 1–8.
Hansen, Donald A., 1967. The uncomfortable relation of sociology and education. In: Hansen, Donald A., Gerstl, Joel E. (Eds.), On Education: Sociological Perspectives. John Wiley & Sons, New York, pp. 3–35.
Hargreaves, David, 1967. Social Relations in a Secondary School. Routledge and Kegan Paul, London.
Hollingshead, A.B., 1947. Elmstown's Youth. Wiley, New York.
Huang, L., Mossige, S., 2012. Academic achievement in Norwegian secondary schools: the impact of violence during childhood. Social Psychology of Education: An International Journal 15 (2), 147–164.
Keeves, John P., 1997. Trends in quantitative research methods. In: Saha, Lawrence J. (Ed.), International Encyclopedia of the Sociology of Education. Pergamon Press, Oxford.
Keeves, John P., Darmawan, I., Giusti Ngurah, 2009. In: Saha, Lawrence J., Dworkin, A. Gary (Eds.), Developments in Quantitative Methods in Research into Teachers and Teaching. Springer Science+Business Media LLC, New York.
Klovdahl, Alden S., 1994. Social network analysis. In: Husén, Torsten, Postlethwaite, T. Neville (Eds.), The International Encyclopedia of Education, second ed. Pergamon Publishers, Oxford.

Lakomski, G., 1997. Critical theory and education. In: Saha, Lawrence J. (Ed.), International Encyclopedia of the Sociology of Education. Pergamon Press, Oxford.

LeCompte, Margaret D., 2009. Trends in research on teaching: an historical and critical overview. In: Saha, Lawrence J., Dworkin, A. Gary (Eds.), International Handbook of Research on Teachers and Teaching. Springer Science+Business Media LLC, New York.

LeCompte, Margaret D., 1997. Trends in qualitative research methods. In: Saha, Lawrence J. (Ed.), International Encyclopedia of the Sociology of Education. Pergamon Press, Oxford, pp. 246–263.

Lortie, D., 1975. Schoolteacher. University of Chicago Press, Chicago.

Marks, Gary N., 2005. Cross-national differences and accounting for social class inequalities in education. International Sociology 20 (4), 483–505.

Meyer, John W., 1997. World society and the nation-state. American Journal of Sociology 103 (1), 144–181.

Natriello, G., 2009. High stakes testing and teaching to the test. In: Saha, Lawrence J., Dworkin, A. Gary (Eds.), International Handbook of Research on Teachers and Teaching. Springer, Dordrecht, pp. 1101–1111.

Noblit, George W., Pink, William T., 1995. Mapping the alternative paths of the sociology of education. In: Pink, William T., Noblit, George W. (Eds.), Continuity and Contradiction: The Futures of the Sociology of Education. Hampton Press, Cresskill, NJ.

Olweus, D., 1993. Bullying at School: What We Know and What We Can Do. Blackwell, Oxford.

Rigby, K., Smith, Peter K., 2011. Is school bullying really on the rise? Social Psychology of Education: An International Journal 14 (4), 441–455.

Rizzuto, Tracey E., LeDoux, J., Hatala, John P., 2009. It's not just what you know, it's who you know: testing a model of social networks to academic performance. Social Psychology of Education: An International Journal 12 (2), 175–189.

Robertson, Susan L., Dale, R., 2008. Researching education in a globalising era: beyond methodological nationalism, methodological statism, methodological educationism, and spatial fetishism. In: Resnick, J. (Ed.), The Production of Educational Knowledge in the Globalising Era. Sense Publishers, Rotterdam.

Roland, E., Bru, E., Midthassel, Unni V., Vaaland, Grete S., 2010. The zero programme against bullying: effects of the programme in the context of the Norwegian manifesto against bullying. Social Psychology of Education: An International Journal 13 (1), 41–55.

Rosenbaum, James E., Rosenbaum, Janet E., Stephan, Jennifer L., 2011. Perfectionist dreams and hidden stratification: is perfection the enemy of the good? In: Hallinan, Maureen T. (Ed.), Frontiers in Sociology of Education. Springer, Dordrecht.

Rosenthal, R., Jacobson, L., 1968. Pygmalion in the Classroom: Teacher Expectations and Pupils' Intellectual Development. Holt, Rinehart & Winston, New York.

Rubie-Davies, C., 2009. Teacher expectations and labeling. In: Saha, Lawrence J., Dworkin, A. Gary (Eds.), International Handbook of Research on Teachers and Teaching. Springer, Dordrecht.

Saha, Lawrence J., 2001. Durkheim's sociology of education: a critical assessment. Education and Society 19 (2), 19–31.

Saha, Lawrence J., 1994. Sociometric methods. In: Husén, Torsten, Postlethwaite, T. Neville (Eds.). Pergamon Press, Oxford.

Saha, Lawrence J., Keeves, John P., 2003. Leading the way: the development of analytic techniques in the sociology of education. In: Torres, Carlos A., Antikainen, A. (Eds.), The International Handbook on the Sociology of Education. Rowman & Littlefield Publishers, Inc, Lanham.

Shavit, Y., Blossfield, H.-P., 1993. Persistent Inequality: Changing Educational Attainment in Thirteen Countries. Westview Press, Boulder.

Sikora, J., Saha, Lawrence J., 2010. New directions in national education policymaking: student career plans in international achievement studies. In: Wiseman, W. (Ed.), The Impact of International Achievement Studies on National Education Policymaking, International Perspectives on Education and Society. Emerald Group Publishing Limited, Bingley.

Waller, Willard W., 1932. The Sociology of Teaching. Russell & Russell, New York.

Weis, L., Jenkins, H., Stich, A., 2009. Diminishing the divisions among us: reading and writing across theory and method in the sociology of education. Review of Educational Research 79 (2), 912–945.

Willis, P., 1977. Learning to Labour: How Working Class Kids Get Working Class Jobs. Gower, Aldershot.

Young, Michael F.D. (Ed.), 1971. Knowledge and Control: New Directions for the Sociology of Education. Collier-Macmillan, London.

Educational Systems: North America

Gerald K LeTendre, The Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Education in North America, since colonial times, has been shaped by a cultural dynamic that promotes education as an essential good that is constantly being expanded while also seeing education as potentially disruptive, subversive, and decisive – hence something that must be controlled and regulated. The national systems of public, compulsory education in Canada and the USA have been largely shaped by this cultural dynamic with some differences in historical and cultural forces. In this article, shared cultural dynamics are discussed and an overview of other subsystems of education, such as military schools, tribal and historically black colleges, and religious schools, is provided.

North America is home to two national systems of public, compulsory education that differ in size, degree of centralization, and organizational structure but have been shaped by similar historical and cultural forces. In this article, shared cultural dynamics are discussed and an overview of other subsystems of education, such as military schools, tribal and historically black colleges, and religious schools, is provided.

The Effect of Religion and Language on Early Forms of Schooling

Early in the seventeenth century, Father Superior Paul Le Jeune sat down with 'a small Indian boy on one side and a small negro on the other' and attempted to instruct the children in the Catholic faith. While brief and ineffective, 'as neither of the three understood the language of the others,' Le Jeune's efforts epitomize the tenor and focus of schooling in North America during the seventeenth and eighteenth centuries (Parkman, 1983: p. 413). This early encounter is also a good example of the cultural dynamic that persists to this day: education continues to be offered to an ever-expanding set of people, but groups battle for who will control the content and access to various types of education.

In both the French and English colonies, schools were used to maintain existing social hierarchies and assimilate non-European populations by disseminating religious views, inculcating moral virtues (European), and instilling literacy (in English). Early colleges (e.g., the Jesuit College in Quebec or Harvard College) provided instruction in religious dogma and the Christian sacral languages. Literacy was tied to religious indoctrination, and the bible was probably the most common school 'text' (see Kett, 1977 or Sutton, 1988). As Cremin noted: "During the first phases of the Virginia settlement, for example, education was in the hands of the ministers, who used it both as an instrument for promoting discipline and order among the colonists and as a device for winning the loyalty of the natives" (Cremin, 1970: p. 9).

The impact of schooling was limited, though, because during the seventeenth, eighteenth, and early nineteenth centuries most of the population received little if any formal education. There were early attempts at vocational (or domestic) education – for example, the Ursuline nuns of Montreal received "a thousand francs to buy wool, and a thousand more for teaching girls to knit" (Parkman, 1983; Cremin, 1970: p. 1357), but most 'education' consisted of skills learned in the family or apprenticeships. Even slaves participated in these forms to a wide extent: O. Equiano notes that he was trained as a ship's pilot as a slave (Equiano, 1995: p. ix). Education in the colonies was unsystematic, polymorphous, and above all, practical.

Emerging Systems: Education to Transform Society

In the USA, networks of public schools developed rapidly in the early 1800s. Early architects of the public education system, such as Thomas Jefferson, believed that a secular, public system of schools would serve as the foundation of democracy in the new republic. Benjamin Rush lobbied (albeit unsuccessfully) to create a state-funded system of education in the state of Pennsylvania. Horace Mann worked tirelessly in the cause of the 'common school' in Massachusetts and worked to establish 'normal schools' to train teachers – a group whose educational accomplishments were often astonishingly meager in the 1800s.

Driving these movements for educational reform was a widespread belief in the transformative power of education and a desire to build a system of education that would change society – a cultural tendency that has been piquantly labeled as a propensity to 'tinker toward utopia' (Tyack and Cuban, 1995). Rush was one of the first to articulate these ideas in a systematic way. He noted: "Our schools of learning … will render the mass of people more homogenous, and thereby fit them more easily for uniform and peaceable government." Enumerating "a few advantages of learning upon mankind," he wrote "A free government can only exist in an equal diffusion of literature … and where learning is confined to a *few* people, we always find monarchy, aristocracy and slavery" (Rush, 1787: pp. 1–2, 14).

Yet, during this same period, many state legislatures began to systematically prohibit literacy and other forms of schooling to both slave and free African populations by denying them access to education. "By 1835, every southern stat had a law prohibiting the schooling of slaves" (Smith, 1984: p. 688). During the same period that industrial education experiments

for white, working-class girls were occurring in northern towns like Lowell, Massachusetts, education for African-Americans saw a brief flowering after the Civil War (1870–80) that was suppressed by the institution of segregated schools.

It was not until the late 1800s that educational progressive like the philosopher John Dewey gained widespread political ascendancy and began to 'tinker' with public schools in order to transform the social order. From the late 1800s through the end of the Second World War, scholars and politicians from a wide range of backgrounds worked to extend and systematize public schooling. Yet these reformers typically assumed, as did Rush, that schooling would homogenize new waves of immigrants and make them 'good citizens' of Canada or the US.

Conflict of Ideals: Maintaining Society or Transforming It through Schooling?

The historical development of the North American school systems was affected by conflicting ideals about what schools should do for society. One end of the continuum is the view that school should be a vehicle for instilling morals, proper behavior, and transmitting a stable social order. At the other is the view that schools should allow students the widest range of self-determination and exploration and eliminate old prejudices, thereby paving the way for a new, and better, social order.

Colonization, Native Peoples, and Schooling

Early in the development of both the US and Canadian systems of education, special schools were created for the children of Native or First Peoples that stripped them of their culture and language and attempted to replace them with English, European moral sensibilities, and 'Christian' virtue. These subsystems of education, including Bureau of Indian Affairs schools in the USA and church-run boarding schools in Canada, continued to function in this manner well into the twentieth century. At the start of the twenty-first century, some groups in Canada are still pursuing legal action against such schools for the abusive treatment of children. And from the 1960s on, there has been considerable educational reform among Native communities, including the development of a system of tribal colleges in the US.

From Segregation to *a Nation at Risk*: The Federal Role in US Public Schools

In the former Confederate States of the US, freed slaves made significant educational advances during reconstruction until 1873, when the political imposition of Jim Crow laws and subsequent legalized segregation (flowing from the Supreme Court decision of Plessy vs Ferguson, 1896) reversed the educational advancement of many African-Americans. A small, but influential, number of students enrolled in historically black colleges and universities which continue to play a significant role in the USA today (Drewry, 2003). Beginning with the early Civil Rights movement in the US, renewed political battles occurred over the role of the nation's schools. The US Supreme Court decision of Brown versus the Board of Education (1954) overturned the legal support for school segregation, fomenting widespread social unrest and unprecedented school reform. In one of the few instances where federal troops have been used to countermand the orders of state leaders, National Guardsmen opened a high school in Birmingham, Alabama and allowed a handful of black teenagers to enter.

The US federal role in public elementary and secondary schools has continued to grow, and has increasingly become the focus of intense political conflict. Events such as the launch of the Russian satellite Sputnik and the publication of *A Nation at Risk* set off tremendous public debate about how schools were failing children and what could be done about it. While many have interpreted school reform since 1980 as an 'attack on the public schools,' attempts to create charter schools, provide vouchers for parents to send children to private schools, or pass laws making home schooling easier are evidence of a continued belief that education is the best tool for creating a better society.

Achieving Equity: Conflict over Language, Religion, and the Curriculum

A concern with educational equity has dominated educational concerns since the 1960s but the focus has shifted from assuring equal access to ethnic or racial minorities to one of who ultimately controls what values the school will transmit. These battles – over language, the teaching of religion (e.g., offering Creation Science), and general course content – have been far more vociferous in the US than in Canada. In the US constitution, there is no clear enumeration of educational rights comparable to that found in the Canadian Charter of Rights and Freedoms which specifically details the educational rights of language minority parents. This leaves open the possibility of political referendums, like the California Act ending bilingual education, to substantially change the focus of schooling.

The Systems Today

Despite the progress made in establishing comprehensive, egalitarian systems of public education through the secondary level, the public school systems in the US and Canada have not yet been able to fully rectify persistent problems of social inequality and conflicting cultural values. Yet the history of tension and educational experimentation has resulted in highly decentralized systems of free, public education often emulated by other industrialized nations. In the Canadian system, under the British North America Act of 1867, provinces have wide discretion in setting educational policy, and there is no national ministry of education. In the US, individual states have broad powers to set educational standards but locally elected school boards exert considerable control on the day-to-day operation of the schools. Federal control over schools in the US is indirect, typically using laws that require states to meet certain standards in schools in order to obtain federal funds for education to insure compliance.

In both nations education is compulsory from about ages 6 or 7 to about age 16. In Canada, students make the transition from elementary to secondary schooling between sixth and eighth grade (with a slightly different system in Quebec). In the

US, despite the highly successful junior high school movement of the early twentieth century, there is no standard form of early secondary school organization. About 75% of students in the USA graduate on time from public high schools (Chapman et al., 2010: p. 52), compared with about 84% of Canadian students (http://www.statcan.gc.ca). Although dropout rates have been falling in both countries, these statistics obscure the fact that minority students have statistically significant lower mean achievement scores, higher dropout rates, and lower rates of college enrollment than majority students do.

The cultural emphasis on public school as the means of promoting social change or maintaining a peaceful and orderly society has often led scholars to ignore other nationwide systems of education. In any broad assessment of systems of education in North America, the role of tertiary education, military schools, and religious schools must be addressed.

The Tertiary Sector: Dominance in the World System

At the university level, in the US, and to some extent Canadian, the cultural oppositions that affect elementary and secondary education have largely disappeared. If we compare two institutions that have had a profound impact on the development of college education in general – Harvard University and the University of Virginia – we find two distinct historical roles. For much of its history, Harvard educated the New England social elite and the Protestant clergy. The University of Virginia, while also transmitting social privilege, had a secular orientation that was replicated in the land-grant colleges set up around the US. Although land-grant colleges were sometimes modeled on elite private colleges (e.g., the University of Georgia and Yale University), these institutions focused on improving the conditions of the state populace and offered courses in such pedestrian subjects as agricultural or domestic science. Enrollment grew rapidly; most states now boast a network of high-quality public colleges or universities accessible to a large segment of the populace.

Over time the differences between highly ranked private and public universities have declined, giving rise to the 'research university' that combines undergraduate training with advanced research in virtually every discipline currently studied. Many US and Canadian universities have eclipsed the European institutions on which they were originally modeled as international centers of knowledge production, and the US is widely recognized as the world leader in production of PhD recipients.

Military Schools and Education

Over half of high school graduates go on to some form of college in the US and Canada, but a large percentage of the population still uses the military as a means for educational advancement. For children of working-class families, military training and military service offer low-cost educational opportunities. Not only does the military provide basic instruction to its enlisted men and women, in both nations it also runs an independent system of elementary and secondary schools for the children of military personnel.

The US military also has a substantial tertiary presence. It operates highly regarded military colleges such as the Army Academy at West Point and funds students in public and private colleges through the officer training (i.e., the Reserve Officer Training Corps). Less visible, but not less important to tertiary education, the US Department of Defense funds extensive research projects, providing substantial external sources of revenues for many colleges.

Religious Schools and Churches

Churches and religious schools form another diverse system outside of public schooling. "As in colonial times, the churches continued to serve as centers of formal and informal education" (Cremin, 1970: p. 383). The role of churches as centers for disseminating education and promoting social change was dramatically seen in the US, when in the 1960s, Sunday sermons melded into political calls for ending segregation. At present, US and Canadian churches are host to a wide array of meetings that have an explicitly educational component. Large suburban churches typically offer a range of weekly classes and meetings ranging from independent piano lessons to church-sanctioned marriage classes. These activities are distinct from, and constitute a considerable educational impact over and beyond, the activities of private (US) or state-related (Canada) religious schools.

Religious schools (primarily, but not exclusively Christian) are found throughout North America, ranging from local kindergarten programs on up to world-class universities like Notre Dame. Both nations experienced considerable political conflict over the role government should play in supporting these schools. In the US political battles and court challenges (such as the conflict in the 1840s regarding Catholic schools) have generally been resolved to reinforce the separation of church and state, although groups like the Amish have mounted successful court challenges that have limited the state's ability to enforce compulsory schooling laws. In Canada many school boards were organized along denominational lines, and there is considerable variation in the relationship between provincial educational authorities and church-supported schools.

Experimentation and Innovation: The North American Educational Dynamic

Mr. Fredrick Douglas, a former slave, penned one of the most compelling statements of North American beliefs in the transformative power of education. Douglas wrote:

> Mr. Auld [his owner] … at once forbade Mrs. Auld to instruct me further, telling her, … that it was unlawful, as well as unsafe, to teach a slave to read. The very decided manner with which he spoke, … served to convince me that he was deeply sensible of the truths he was uttering … and the argument which he so warmly urged, against my learning to read, only served to inspire me with a desire and determination to learn.
>
> Huggins, 1980: pp. 42–43

The focus on who shall be educated has shifted to higher education as attendance through secondary schooling became standardized. Jerome Karabel (2005) showed the desperate

lengths that administrators at some Ivy League schools went to in order to prevent immigrant, Jewish, and working-class children from entering. The end of the twentieth century saw repeated legal challenges to affirmative action and other forms of race-conscious admissions policies, despite strong evidence that these programs were effective in opening access to more students (Bowen et al., 1998).

North Americans share a culture that places extreme emphasis on the power of education to transform both the individual and society and typically view schools as important institutions of social change or control. However, there is considerable cultural conflict over how schools are to be used: many citizens hold the view that a good government should mandate a system of free and comprehensive public education that is uniform for all, while others view any government intrusion into educational matters with extreme aversion. This dynamic is currently expressed in concerns about universities becoming too much like corporations (Bok, 2004) and whether colleges actually educate students in ways that lead them to useful occupations (Arum and Roksa, 2010).

North Americans also share a cultural tendency to engage each new educational innovation as a chance to move one step forward toward social utopia. This is true for groups that have long faced social oppression and bans on schooling. For example, having endured the 'cultural genocide' promulgated in schools like the Carlisle Indian Industrial School, many US Indian nations have launched their own tribal colleges, and organized the American Indian Higher Education Consortium. This facet of North American culture may appear to others in the world as an obsession with educational reform, yet, this cultural dynamic has served to create a system of public education that is widely emulated. In the 1980s and 1990s, the performance of US and Canadian students on international tests of academic achievement caused many to declare the failure of North American schools and praise the educational culture of East Asia. Few saw the irony of the situation: the Japanese school system was heavily influenced by educational reformers of US and educational policy makers in East Asia continue to look to North America as a center of educational reform and research.

Educational reform initiatives originating in North America will continue to influence educational systems around the globe. Looking toward the future, the most significant changes in systems of education will be wrought by the historical confluence of new electronic communication networks and a resurgence of political efforts to decentralize the control of schools. This can be seen in both the home schooling and charter school movements.

Widespread access to immense databases and electronic communication now allows parents to provide children with access to a world-class curriculum within the home, thereby weakening old claims that 'home learning' is greatly inferior to 'schooling.' While the majority of home-educated children belong to fundamentalist, Christian groups, a wide range of parents, anywhere from 300 000 to over half a million around 1995, now opt to school their children at home (U.S. Department of Education, 2000).

Charter schools, in the US, have also increased in the wake of the educational restructuring movement of the 1980s and 1990s. These schools are publicly funded and subject to governmental regulation but operate under fewer restrictions than ordinary public schools. There is great variation in state laws regarding the establishment of these schools, and only a small percentage of the US school-age population attends charter schools. But again, the ability of a small number of adults to access an immense amount of literature, music, and art computer-based networks has made the establishment of such schools quite viable.

The demand for these, and other alternatives to public education, will continue to arise from the important cultural role that education has in North American society. The early part of the twentieth century saw the dramatic expansion of an organized public school system, then viewed as the best way to move society forward. The early part of the twenty-first century will likely see dramatic decentralization as families push for more autonomy over their child's education and individuals of all ages continue to resist any attempt to control what they can, or cannot, learn.

See also: Comparative Research in Education: IEA Studies; Creativity Training; Cross-Cultural Study of Education; Education: Anthropological Aspects; Educational Institutions, History of.

Bibliography

Arum, R., Roksa, J., 2010. Academically Adrift: Limited Learning on College Campuses. University of Chicago Press, Chicago, IL.

Bok, D., 2004. Universities in the Marketplace. Princeton University Press, Princeton, NJ.

Bowen, W., Bok, D., Shulman, J., 1998. The Shape of the River. Princeton University Press, Princeton, NJ.

Chapman, C., Laird, J., KewalRamani, A., 2010. Trends in High School Dropout and Completion Rates in the United States: 1972–2008. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Cremin, L., 1970. American Education: The Colonial Experience 1607–1783. Harper and Row, New York.

Drewry, H., 2003. Stand and Prosper: Private Black Colleges and Their Students. Princeton University Press, Princeton, NJ.

Equiano, O., 1995. The Interesting Narrative. Penguin Books, New York.

Huggins, N., 1980. Slave and Citizen: The Life of Frederick Douglas. Little, Brown and Company, Boston.

Karabel, J., 2005. The Chosen: The Hidden History of Admission and Exclusion at Harvard. Yale and Princeton. Houghton Mifflin, San Francisco.

Kett, J., 1977. Rites of Passage Adolescence in America, 1790 to the Present. Basic Books, New York.

National Commission on Excellence in Education, 1983. A Nation at Risk. Washington, DC.

Organisation for Economic Co-Operation and Development (OECD), 1998. Education at a Glance, OECD Indicators. OECD, Paris.

Parkman, F., 1983. France and England in North America, vol. I. Library of America, New York.

Rush, B., 1787. A Plan for the Establishment of Public Schools and Diffusion of Knowledge in Pennsylvania. Thomas Dobson, Philadelphia.

Smith, J., 1984. Race and human capital. The American Economic Review 74, 685–698.

Sutton, J., 1988. Stubborn Children. University of California Press, Berkeley.

Tyack, D., Cuban, L., 1995. Tinkering toward Utopia: A Century of Public School Reform. Harvard University Press, Cambridge, MA.

U.S. Department of Education. National Center for Education Statistics, 2000. Issues Related to Estimating the Home-Schooled Population in the United States with National Household Survey Data, NCES 2000–311. Washington, DC.

Efficacy, Political

Thomas J Scotto and Carla Xena, University of Essex, Colchester, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E. Riedel, volume 7, pp. 4353–4356, © 2001, Elsevier Ltd.

Abstract

Since Campbell et al. (1954) first defined political efficacy, numerous studies emerged placing emphasis on: (1) the theoretical relevance of the concept for the functioning of liberal democracies; (2) its relationship with other political attitudes and behaviors, and (3) the (unresolved) controversy around its measurement. *This article provides a comprehensive* overview of the nature and evolution of the literature on political efficacy by presenting how disciplines such as political science, sociology, communication and media studies and psychology *tackle these questions.*

Political efficacy is defined by Campbell et al. as "the feeling that individual political action does have, or can have, an impact upon the political process, i.e., that it is worthwhile to perform one's civic duties" (1954: 187). It is an important prerequisite for the conventional forms of political participation – voting and membership in party and civic organizations – viewed as vital for maintaining functioning and stable democratic societies. Debates over the definition of political efficacy, its measurement on standard surveys, its cross-national and cross-temporal comparability, and its causes and effects fill numerous pages of scholarly journals and monographs.

The Concept of Political Efficacy

In early quantitative studies, measures of political efficacy are used to separate those who believe "the affairs of government can be understood and influenced by individual citizens" from those who believe "politics is a distant and complex realm that is beyond the power of the common citizen to affect" (Campbell et al., 1960: 104). Early empirical work treats political efficacy as a single concept and errs toward tapping an individual's belief about their own ability to be effective citizens in the public sphere (Pateman, 1970; Thompson, 1970; Finkel, 1985). Citizens who have such abilities are seen as possessing what Almond and Verba (1963: 257) call "subjective competence," and their presence is vital to the stability and functioning of representative democracies.

However, significant scholarly disagreement exists, even in the early literature. As an example, Muller (1970) argues that Almond and Verba's concept of political competence really has three dimensions: (1) the attention one pays to politics and public affairs as measured by following the news, knowledge of political leaders, etc.; (2) the ability to influence government or a belief that one would get involved in a civic matter and be effective at affecting change; and (3) political efficacy which contains measures tapping citizen beliefs as to how responsive they believe the government is to their concerns.

Rosenberg's (1954–55) qualitative work from the early 1950s identifies respondents reluctant to discuss politics because of their feelings of inadequate expertise or knowledge of the subject. These individuals are both lacking in political awareness and do not have faith in their ability to participate. Other respondents are shown to disengage from politics because of feelings that it is 'futile' to participate because the political class is seen to be unresponsive to the beliefs of ordinary citizens. Many who feel inadequate also feel the system is unresponsive, but there is also a group of interviewees who, regardless of their own agency, feel that politicians and the system will never be responsive. This differentiation is picked up by Robert Lane (1959: 149) who notes that political efficacy "combines the image of the self and the image of democratic government. [It] contains the tacit implication that the image of the self as effective is intimately related to the image of democratic government as responsive to the people." Beliefs that politicians are responsive to the needs of ordinary citizens is dubbed external political efficacy, and peoples' subjective assessments of their abilities to affect political change is coined internal political efficacy.

There is an expectation of interplay between citizen levels of internal and external efficacy, so even though the concepts are viewed as theoretically distinct, high levels of correlation between the two types of efficacy are expected (Clarke et al., 2010). Coleman and Davis (1976: 191–193) summarize: "[I]ndividuals who believe the system is responsive to people like themselves will be more likely to believe that they personally have the skills to induce government officials to act" (quoted in Craig and Maggiotto, 1982: 86).

Research from the 1980s brings questions about construct and statistical conclusion validity as it relates to the concept of political efficacy. Balch (1974) empirically demonstrates that survey questions measuring efficacy break down into two separate dimensions. Craig and Maggiotto (1982) argue that high levels of correlation between the efficacy dimensions should not be assumed because internal efficacy is a measure of internal 'political effectiveness' while external efficacy has more to do with attitudes toward 'system responsiveness.' According to Craig et al. (1990: 305), the correlation that does exist occurs "presumably because beliefs about one's own competence help to shape beliefs about the potential for citizens generally to play a meaningful role in political decision making." But there is a counterargument. Research dating back to the early 1960s (e.g., Agger et al., 1961) suggests that those low in internal efficacy can be more likely to believe that politicians are doing what they can to help the citizenry. In contrast, those high in

internal efficacy believe that they will do a better job in office than those elected and this mentality is reflective of or fosters low levels of external efficacy. The fact that internal and external efficacy can, in some circumstances, be mutually reinforcing while in others is unrelated underscores the need to divide the concept of efficacy into two. After decades of research, scholars agree that internal efficacy concerns "citizens' feelings of personal competence to understand and to participate effectively in politics" (Craig et al., 1990: 290) while external efficacy focuses on citizens' perceptions of the responsiveness of political bodies and actors to citizens' demands (Balch, 1974; Converse, 1972; Miller et al., 1980).

Pollock, III (1983) and Zimmerman (1989) break people into one of four groups depending on their internal and external efficacy levels. Group 1, those with low levels of internal and external political efficacy, eschew all forms of political participation, and are seen as 'uninvolved.' Those in group 2, who do not feel personally competent enough to participate in politics but feel the system is responsive to their needs, are more likely than those in the first group to vote or participate in a campaign but just as unlikely to protest or be active in political groups. Those in this group are happy to follow, for the most part leaving politicians alone as group members lack confidence in their own expertise to play a more active role in politics. Group 3 contains individuals with high levels of both internal and external efficacy. They are far more likely than those in group 1 to vote, campaign, contact others to discuss politics, and be members of groups. Zimmerman (1989) also notes that this is the group most likely to produce future politicians. Finally, those in the last group, who have high levels of internal but low levels of external efficacy, are more likely than those in the other groups to participate in protest activity and almost as likely as respondents in Group 3 to participate in a campaign or contact people to talk politics. However, Group 4 respondents are reluctant to vote or join conventional political groups.

The Relevance of Political Efficacy

The conceptualization of political efficacy is of paramount salience due to its implications for political disengagement and participation in liberal representative democracies. (Examples of literature include: Almond and Verba, 1963; Balch, 1974; Barnes, 1966; Barnes and Kaase, 1979; Campbell et al., 1960; Craig and Maggiotto, 1982; Craig et al., 1990; Easton and Dennis, 1967; Gamson, 1968; Niemi et al., 1991; Rosentstone and Hansen, 1993; Smetko and Valkenburg, 1998; Siegelman and Feldman, 1983.) Scholars consider subjective or internal political efficacy to be an essential indicator of the health of democratic regimes since sentiments of system legitimacy underlie feelings of efficacy (Finifter, 1970; Pateman, 1970; Smetko and Valkenburg, 1998), and high levels of political efficacy are a potential defense against citizen disillusionment with democratic structures of governance (Easton and Dennis, 1967). Citizens in democracies theoretically have higher levels of external political efficacy than those living in authoritarian or dictatorial regimes because of the sense that representative governments are attentive to their needs. This responsiveness facilitates citizen attentiveness and interest in politics and personal political empowerment, which, in turn increases internal efficacy. In semi-authoritarian and dictatorial regimes, this link is severed, partially because of the state's ability to curry favor with chosen groups rather than via the democratic process (Coleman and Davis, 1976).

Psychologists who study individuals' feelings of personal self-worth and efficacy note that people usually undertake collective action or join group activities only when they believe their participation is effective in producing outcomes they support (Bandura, 1995). Participation in politics is predicated on the belief that citizens can affect change and/or politicians and public officials are responsive to their participatory actions. Citizens' sense of political efficacy is highly correlated with the decision to turn out to vote (Campbell et al., 1960; Clarke et al., 2004) and become active in political parties (Whiteley, 2011). Declines in turnout over time are linked to falling levels of external political efficacy because a lack of perceived governmental responsiveness to people's problems reduces the belief that participation is a useful activity (Abramson and Aldrich, 1982).

Internal and external efficacy positively correlates with voting behavior in panel studies of political participation, but external efficacy is a stronger predictor of more high intensity participation such as attending campaign events or donating to political causes (Finkel, 1985). In a study of the Canadian public, Clarke et al. (2010) find that political trust has a strong correlation with efficacy and note that both internal and external efficacy are strongly related to the levels of satisfaction with democracy and the influence people believe they have on politics. In contrast to the findings of Finkel (1985), they find that most forms of political participation such as voting, the decision to contact elected officials, and boycotting and petitioning government actions have a stronger relationship with internal than with external efficacy. This can arise because political interest manifests itself into structured discussions and deliberations over political matters that can in turn improve internal political efficacy, although the relationship may be indirect (Morrell, 2005). When discussion about politics arises, more high intensity forms of political action follow.

The concept of political trust, defined by Arthur Miller (1974) as the citizen's belief that the government and politicians are "producing outputs in accord in accord with individual expectations" often appears with political efficacy as a co-determinant of political action. Gamson (1968) notes that participation in mainstream political participation requires *both* high levels of efficacy and trust, and those lacking in trust, often labeled as 'politically cynical,' but high on levels of efficacy are prone to mobilization for protest activity. In contrast, those lacking both trust and efficacy are apathetic but unlikely to do anything to ameliorate their situation. Finally, those trusting but lacking in efficacy are willing to see the system continue to function without their active involvement.

Seligson's (1980) study of peasants in Costa Rica suggests that high levels of political cynicism or distrust is more strongly related than efficacy at predicting anti-systemic political activity while efficacy is a stronger predictor than trust at predicting the more standard forms of political participation, but the evidence is mixed (cf Pierce and Converse, 1989). Standard forms of political participation among the distrusting but politically efficacious is possible if education levels are high (Chan, 1997)

and potential voters believe are trustworthy opposition candidates waiting in the wings (Nilson and Nilson, 1980). (For an overview, see also Levi and Stoker (2000).)

A disconnect between levels of trust and efficacy are not the only potential drivers of unconventional participation. Converse (1972) notes a decline in political efficacy during the 1960s, a point that, in light of increasing levels of education in the (American) population is surprising to scholars. However, a clear picture emerges when distinguishing between the internal and external components of efficacy – during times of political upheaval, efficacy gaps may emerge. Many citizens of the late 1960s and early 1970s possess high levels of internal efficacy but come to believe politicians and the system is unresponsive to their demands. The unconventional forms of political participation seen in the United States during the 1960s and 1970s are a function of high internal efficacy and low external efficacy (Craig, 1980).

The Antecedents of Political Efficacy

Perhaps the best-known finding concerning differences in efficacy levels among citizens within a nation concerns the link between political efficacy and education. Agger et al. (1961) notes that the completion of a course of higher education is strongly and positively associated with an individual's propensity to believe they are politically efficacious. Prewitt (1968: 226 as cited in Aberbach, 1977) argues that education may be the key contributor to higher levels of political participation. Studies of community action in the late 1960s action posit that low status groups and racial minorities may not effectively organize because of their low levels of education and empowerment (Lipsky, 1968). Children in deprived environments are noted to exhibit higher degrees of hostility toward the political system and politicians and lag behind wealthier students in their levels of efficacy (Lyons, 1970). Recent findings (e.g., Gimpel et al., 2003) describe how African-American youth continue to lag behind both white children and Asians when it comes to their levels of both internal and external efficacy.

Education is seen as so vital to the development of political efficacy that political scientists place emphasis on determining what makes young people efficacious and interventions are designed to raise levels of political efficacy via school curricula or civics education (e.g., Langton and Jennings, 1968). The idea is that increased civics courses increase political efficacy, which is positively related to political interest and the propensity to participate in political discussions, both important precursors for emerging democratic citizens. Although education is the main driver of efficacy and a child's political development, cross-cultural studies show that other social demographic factors are also related to higher levels of efficacy. For example, a study of students in India, the United Kingdom, and the United States finds that parental income is related to students' sense of efficacy (Shrivastava, 1989).

Political efficacy also varies by personality type. Mondak and Halperin (2008) observe a number of relationships between the so-called 'big-five' personality traits and efficacy. Those open to new experiences are likely to have higher levels of internal political efficacy while those more agreeable and willing to follow are less confident in their own political effectiveness but have higher levels of external efficacy. Those high on conscientiousness or who act according to plan have lower levels of both internal and external efficacy, and extroversion is positively associated with efficacy (Vecchione and Caprara, 2009). Another strand of work goes further back in the causal chain and links both self and political efficacy to a person's genetic characteristics (McGue and Bouchard, 1998; Fowler et al., 2008).

New media also has a role in developing political efficacy. Weaver's (1994) analysis of talk radio in the United States finds that those who listen to talk radio shows where listeners were able to call in and give their political opinions are less alienated, even if the direction of the news is biased (see also Banducci and Karp, 2003). In a study of college students, Lee (2006) finds that accessing traditional news sources such as newspapers increases internal efficacy. However, and unexpectedly, accessing sources of information online, particularly government sites are correlated with lower levels of external efficacy. Social network use among American and Chinese students facilitates political discussion and efficacy, particularly by those who are not otherwise civically minded or engaged (Chan and Guo, 2013).

The presence of new media may make people, both young and old alike, more aware of their political rights and increase citizens' internal efficacy as their competence with accessing political information on new media grows (Stoker, 2006). However, this more individualistic form of political activity might leave traditional political avenues out of the picture, and lower external efficacy levels as citizens come to believe that parties, campaigns, and interest groups cannot respond to the public's demands in a timely manner (Coleman et al., 2008).

The (Un)Virtuous Circle? Reciprocal Relationships and Political Efficacy

The disentanglement of the relationship between political participation and political efficacy is a complex exercise, as reciprocal effects exist between the two. In contrast to early findings suggesting levels of political efficacy develop early in adulthood and remain stable over time (cf Iyengar, 1980), elections and other political activities are found to influence efficacy levels, regardless of whether citizens are actually participating in such events (cf Clarke and Acock, 1989). The analysis of panel data from the United States and West Germany in the 1970s demonstrates that campaign activity and external political efficacy have a reciprocal relationship. Finkel's (1985, 1987) analysis of these data concludes that those who partake in high intensity political participation may grow to become more involved over time. Building on the work of Thompson (1970: 66), he notes that those active in campaigns become more confident in their own internal political efficacy which leads people to look for additional avenues of involvement in the political process (see also Finkel et al., 2000). In contrast, those protesting or embroiled in illegal political activity see their levels of political alienation grow, which can reduce their external political efficacy. This may further divide activists into two camps, those supportive of the system and those disaffected.

In his analysis of panel data, Finkel (1985, 1987) did not find that voting increased political efficacy but subsequent work establishes that external political efficacy *does* increase among the subgroup that backs winning candidates. For example, Pattie et al. (2004) find that British respondents who back the Labour Party during the late 1990s and in their 2001 re-election efforts see their political efficacy levels grow. Hence, efficacy may increase not just as a function of political participation but as a result of what Leighley (1991: 199) notes is the "perceived success of such activity" (see also Clarke and Acock, 1989; Clarke and Kornberg, 1992; Valentino et al., 2009).

Campaigns can also raise efficacy levels by paying attention to voters. Those with a history of voting and high levels of political interest are likely to engage with campaigns. Banducci and Karp (2003) find that Americans who report being contacted by a candidate or party during a campaign display higher levels of political efficacy. Some work argues that ballot initiative campaigns engage voters and may improve levels of efficacy as they are a direct way for citizens to have input into the lawmaking or constitutional design of the political system (Mendelsohn and Cutler, 2000). The number of initiatives on a ballot in an American state at a given time is also linked to improved efficacy levels among voters (Bowler and Donovan, 2002). However, Dyck and Lascher's (2009) replication of these results using more sophisticated statistical techniques cast doubt on early findings.

Reciprocal models linking efficacy to a wider variety of forms of political participation include those employed by Stenner-Day and Fischle (1992) on data from Australia and Gastil and Xenos (2010), who look at a three-wave panel study of Americans in the Pacific Northwest and uncover relationships between political efficacy and a wider range of political activities that go in both directions. The latter authors find internal, not external, efficacy is a key driver of political action other than voting, but this relationship is not reciprocated. Media use does not predict internal efficacy but the citizens' engagement with news did enhance internal efficacy. Stenner-Day and Fischle (1992) examine the reciprocal relationships between efficacy and partisan and community activism and political extremism and find considerable feedback between perceived effectiveness, perceptions of the degree to which government listens to citizens and these three forms of political and civic engagement.

Measuring Political Efficacy

Despite its importance as a precursor to political behavior in representative democracies and a long literature claiming it is something to be nurtured, social scientists have difficulty agreeing on how best to measure both internal and external efficacy. Questions measuring the concept vary, but five, whose origins date to the 1952 American National Election Study (ANES) remain prominent. They are agree-disagree questions, which read as follows: (1) "People like me have no say in what the government does"; (2) "Voting is the only way that people like me can have any say about how the government runs things"; (3) "Sometimes politics and government seem so complicated that a person like me cannot really understand what is going on."; (4) "Public officials don't care what people like me think."; and (5) "The way people vote is the main thing that decides how things are run in this country." (The fifth question did drop out of favor with those attempting to validate efficacy batteries in the 1970s (cf Balch, 1974).) In both national and cross-sectional surveys, derivatives of these questions are found, even on surveys fielded in the twenty-first century. For example, the 2002 European Social Survey, a cross-national instrument with respondents from more than 20 nations, contains derivatives of the politicians caring what people think and politics being too complicated questions (see Xena, unpublished).

Despite some consistency in the questions selected over time, variation in the questions employed to tap internal and external across surveys in the 1960s and 1970s is considerable. Balch (1974) is the first to seriously subject common indicators to tests of measurement validity and reliability, and states that questions 2 and 3 from the original ANES battery are reflections of internal efficacy and items 1 and 4 are manifestations of external efficacy. Wright (1975) and Clarke et al. (2010) examine the four items for response set or acquiescence bias, the tendency of survey respondents to agree or disagree to an item regardless of its content, and find little evidence of the problem.

Questions over which items to designate as indicators of internal and external efficacy arise after the ANES updated and enhanced the battery of questions designed to tap both types of efficacy (cf Miller et al., 1980). Specifically, Craig and Maggiotto (1982) find that it is unclear whether item 1 from the original ANES battery is an indicator of external or internal efficacy, conclude that the conceptual validity of most items are suspect, and call for the development of substitutes. In a study examining the suitability of the traditional indicators across multiple countries, Acock et al. (1985) disagree and note that standard indicators employed by the ANES are cross-nationally valid as long as sophisticated latent variable modeling is employed to derive factors of internal and external efficacy (see also Acock and Clarke, 1990). (Acock et al. (1985) designate items 2, 3, *and* 1 as indicators of internal efficacy. Indicators of external efficacy are item 4 and two questions that first appear on the 1968 ANES: "Parties are only interested in people's votes but not in their opinions," and "Generally speaking, those we elect lose touch with the people pretty quickly." The lack of clarity as to whether certain traditional indicators of efficacy are best reflective of the internal or external component continues to this day (Xena, unpublished).)

In the 1980s, the investigators of the ANES respond to the controversy over the efficacy indicators by experimenting with alternative question batteries on pilot studies in the field during the period. Testing a large item battery of potential substitutes, Craig et al. (1990) conclude with a recommendation of a new 10 question item battery, six to measure internal efficacy, and four to measure external efficacy: (1) "I feel that I could do as good of a job in public office as most other people."; (2) "I think I am as well-informed about politics and government as most people."; (3) "I don't often feel sure of myself when talking with other people about politics and government." (4) "I feel that I have a pretty good understanding of the

important political issues facing our country."; (5) "I consider myself well-qualified to participate in politics."; and (6) "Sometimes politics and government seem so complicated that a person like me can't really understand what's going on." are agree-disagree questions designed to measure internal efficacy. External efficacy is measured by ascertaining respondents' agreement or disagreement with the following statements: (7) "There are many legal ways for citizens to successfully influence what the government does."; (8) "Under our form of government, the people have the final say about how the country is run, no matter who is in office."; (9) "If public officials are not interested in hearing what the people think, there is really no way to make them listen."; and (10) "People like me don't have any say about what the government does."

Confirmatory factor analytic techniques on American data show that two distinct latent dimensions drive responses to questions 1–6 and questions 7–10. The alternative measures of internal efficacy are further validated by Morrell (2003), who recommends using items 1, 2, 4, and 5 if a four-question battery is desired, and notes that the Craig et al.'s (1990) and Niemi et al.'s (1991) measures are employed in studies by, among others, Greenberg et al. (1996), Hansen (1997), and Soss (1999). Despite the use of sophisticated measurement techniques to validate the Craig et al. (1990, 1991) measures, questions designed to tap internal efficacy that have their origins in the ANES from the 1950 and 1960s are still in widespread use, particularly in comparative studies (e.g., Anderson and Tverdova, 2001).

The measurement of external efficacy remains contentious, with items 7 and 8 from the revised efficacy battery employed by Craig et al. (1990) often considered to tap regime support and not government responsiveness to the needs of citizens. This calls into question whether these items are truly reflective of external efficacy as a concept or simply fit well with the other two indicators in measurement models. Another problem is that, depending on the study and question battery, item 10 from the revised battery is designated as reflective of both internal and external efficacy. Despite its importance as a concept, it appears that social scientists are far from agreeing on a standardized question battery to measure internal and external efficacy.

Future Research

As a longstanding topic of interest to political scientists concerned with civic life and democratic citizenship, discussions of the causes, effects, and measurement of political efficacy are unlikely to wane. The increasing capabilities of scholars to conduct surveys across space (nations) raises the possibility that proper attention can be paid to the role political institutions have in determining a citizen's views about their own political agency and how responsive are the politicians that result from the voting decisions that are translated via electoral institutions. Social scientists view majoritarian systems as power concentrating, and although they offer electorates clarity in terms of who is governing, the fact that such systems create clear winners and losers tend to leave segments of the population feeling excluded. Karp and Banducci (2008) note that representative democracies nations employing first-past-the-post electoral systems see lower levels of efficacy than those employing more proportional schemes. Further, supporters of smaller political parties are more efficacious in the former than in the latter. More such studies are likely in the future.

The improved capabilities of panel surveys, often conducted via the Internet (cf Johnston, 2008), suggests that the ability of scholars to determine how political efficacy responds to a host of 'shocks' (elections, economic downturns, wars, etc.) will improve and be an interesting avenue for future research. Better panel data can also facilitate ongoing efforts to determine how citizens in emerging and newer democracies perceive their own abilities to participate and the responsiveness of the state to their needs. This is a particularly relevant question as cross-national research finds survey respondents in developing democracies to have lower levels of external efficacy than their more established counterparts (Torcal and Montero, 2006). Finally, surveys conducted in countries that are not yet democracies ponder whether efficacy levels in these nations are strong enough to facilitate self-rule (e.g., Shi, 2000), and whether a loss in efficacy leads democracies to slip back into authoritarian or semi-authoritarian rule (Lussier, 2011; Pietsch et al., forthcoming).

Scholars do have a more nuanced view of efficacy in long-standing representative democracies, noting that it is important not to overplay the notion that falling or low levels of efficacy will lead to certain democratic decline. Although low levels of external efficacy and turnout seem to represent an important challenge for the functioning of democratic regimes and do warrant attention (see Macedo et al., 2005), they will not necessarily result in negative consequences for contemporary political systems. Along these lines, Gamson (1968) argued that certain levels of political efficacy and political cynicism are key elements for political mobilization. When internal-external efficacy gaps occur, protest may follow, potentially raising the probability that governments will pay more attention to those previously disenfranchised, which in turn can lead to civic revival. Locating potential cut points and whether this occurs in practice is an area ripe for research with emerging panel data.

The links between efficacy, regime support, and conventional political participation encourage school-based civics curriculum with designs to enhance efficacy. This is not without controversy. Modern social scientists interested in civics education advise that tasks simply designed to increase efficacy, particularly external efficacy may reduce skills acquisition. Kahne and Westheimer (2006) warn that focusing only on efficacy may mask students' exposure to the reality of power inequalities that lead to political disempowerment of marginalized groups. Further, they note that many civics-based efficacy enhancing exercises have a tendency to emphasize volunteerism with apolitical groups. In the long run, this type of activity might neither generate efficacy nor lead individuals to work on issues that fundamentally are political in nature.

Political efficacy is a concept of importance with a rich history in the social science literature that attracts attention from scholars in political science, sociology, and psychology. Many of the questions that continue to get asked are enduring – what are the antecedents and consequences of the levels of internal and external efficacy we observe among people and

how do we measure these two concepts? Others are less well established, particularly observing the role political institutions have in facilitating or inhibiting efficacy in large N, multi-level, cross-national surveys. The fact that both longstanding and emerging questions surrounding political efficacy remain vital to how scholars and pundits view the health of civic life in established and emerging democracies will act to ensure that studies of political efficacy will continue to be discussed at professional conferences, in peer-reviewed journals, and in well-received books on the topic of political participation for the foreseeable future.

See also: Citizen Participation; Data Bases and Statistical Systems: Political Behavior and Elections; Electoral Systems; Interpersonal Trust across the Lifespan; Politics and Participation; Self-Efficacy; Voting and Direct Democracy.

Bibliography

Aberbach, J.D., 1977. Power consciousness: a comparative analysis. American Political Science Review 71 (4), 1544–1560.

Abramson, P.R., 1972. Political efficacy and political trust among black schoolchildren: two explanations. The Journal of Politics 34 (4), 1243–1275.

Abramson, P.R., Aldrich, J.H., 1982. The decline of electoral participation in America. American Political Science Review 76 (3), 502–521.

Acock, A.C., Clarke, H.D., Stewart, M.C., 1985. A new model for old measures: a covariance structure analysis of political efficacy. The Journal of Politics 47 (4), 1062–1084.

Acock, A.C., Clarke, H.D., 1990. Alternative measures of political efficacy: models and means. Quality and Quantity 24 (1), 87–105.

Agger, R.E., Goldstein, M.N., Pearl, S.A., 1961. Political cynicism: measurement and meaning. The Journal of Politics 23 (3), 477–506.

Aish, A., Jöreskog, K.G., 1990. A panel model for political efficacy and responsiveness: an application of LISREL 7 with weighted least squares. Quality and Quantity 24 (1), 405–426.

Almond, G., Verba, S., 1963. The Civic Culture. Princeton University Press, Princeton.

Anderson, C.J., Tverdova, Y.V., 2001. Winners, losers and attitudes toward government in contemporary democracies. International Political Science Review 22 (4), 321–338.

Balch, G.I., 1974. Multiple indicators in survey research: the concept 'sense of political efficacy'. Political Methodology 1 (2), 1–43.

Banducci, S.A., Karp, J.A., 2003. How elections change the way citizens view the political system: campaigns, media effects and electoral outcomes in comparative perspective. British Journal of Political Science 33 (3), 443–467.

Bandura, A., 1995. Exercise of personal and collective efficacy in changing societies. In: Bandura, A. (Ed.), Self-efficacy in Changing Societies. Cambridge University Press, Cambridge, England.

Barnes, S.H., 1966. Participation, education and political competence: evidence from a sample of Italian socialists. American Political Science Review 60 (2), 348–353.

Barnes, S.H., Kaase, M. (Eds.), 1979. Political Action: Mass Participation in Five Western Democracies. Sage, Beverly Hills.

Bowler, S., Donovan, T., 2002. Democracy, institutions, and influence on citizen influence on government. British Journal of Political Science 37 (2), 371–390.

Campbell, A., Gurin, G., Miller, W.E., 1954. The Voter Decides. Row Peterson, Evanston, IL.

Campbell, A., Converse, P.E., Miller, W.E., Stokes, D.E., 1960. The American Voter. John Wiley & Sons, Inc, New York.

Chan, S., 1997. Effects of attention to campaign coverage on political trust. International Journal of Public Opinion Research 9 (3), 286–296.

Chan, M.C.M., Guo, J., 2013. The Impact of Political Identity, Efficacy, and Selective Media Exposure on Political Participation: A Comparative Study of Young Adults in the United States and Hong Kong [Online]. School of Journalism and Communication, The Chinese University of Hong Kong. Available from: http://jcmotion.com.cuhk.edu.hk/eshop/index.php?route=product/product&product_id=105 (accessed 28.07.14.).

Clarke, H.D., Acock, A.C., 1989. National elections and political attitudes: the case of political efficacy. British Journal of Political Science 19 (4), 551–562.

Clarke, H.D., Kornberg, A., 1992. Do national elections affect perceptions of MP responsiveness? A note on the Canadian case. Legislative Studies Quarterly 17 (2), 183–204.

Clarke, H.D., Sanders, D., Stewart, M.C., Whiteley, P., 2004. Political Choice in Britain. Oxford University Press, Oxford.

Clarke, H.D., Kornberg, A., Scotto, T.J., 2010. Accentuating the negative? A political efficacy question-wording-experiment. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences 6 (3), 107–117.

Coleman, K.M., Davis, C.L., 1976. The structural context of politics and dimensions of regime performance: their importance for the comparative study of political efficacy. Comparative Political Studies 9 (2), 189–206.

Coleman, S., Morrison, D., Svennevig, M., 2008. New media and political efficacy. International Journal of Communication 2, 771–791.

Converse, P.E., 1972. Change in the american electorate. In: Campbell, A., Converse, P.E. (Eds.), The Human Meaning of Social Change. Rusell Sage Foundation, New York, pp. 263–373.

Craig, S.C., 1979. Efficacy, trust, and political behaviour: an attempt to resolve a lingering conceptual dilemma. American Politics Research 7 (2), 225–239.

Craig, S.C., 1980. The mobilization of political discontent. Political Behaviour 2 (2), 189–209.

Craig, S.C., Maggiotto, M.A., 1982. Measuring political efficacy. Political Methodology 8 (3), 85–110.

Craig, S.C., Niemi, R.G., Silver, G.E., 1990. Political efficacy and trust: a report on the NES pilot study items. Political Behavior 12 (3), 289–314.

Dalton, R.J., 1988. Citizen Politics in Western Democracies. Public Opinion and Political Parties in the United States, Great Britain, West Germany and France. Chatham House, Chatham, NJ.

Dalton, R.J., 1999. Political support in advanced industrial democracies. In: Norris, P. (Ed.), Critical Citizens. Global Support for Democratic Governance. Oxford University Press, Oxford.

Dyck, J., Lascher, E., 2009. Direct democracy and political efficacy reconsidered. Political Behaviour 31 (3), 401–427.

Easton, D., Dennis, J., 1967. The child's acquisition of regime norms: political efficacy. The American Political Science Review 71 (1), 25–38.

Finifter, A.W., 1970. Dimensions of political alienation. American Political Science Review 64 (2), 389–410.

Finkel, S.E., 1985. Reciprocal effects of participation and political efficacy: a panel analysis. American Journal of Political Science 29 (4), 891–913.

Finkel, S.E., 1987. The effects of participation on political efficacy and political support: evidence from a West German panel. Journal of Politics 49 (2), 441–464.

Finkel, S.E., Sabatini, C.A., Belvis, G.G., 2000. Civic education, civil society, and political mistrust in a developing democracy: the case of the Dominican Republic. World Development 28 (11), 1851–1874.

Fowler, J.H., Baker, L.A., Dawes, C.T., 2008. Genetic variation in political participation. American Political Science Review 102 (2), 233–248.

Gamson, W.A., 1968. Power and Discontent. Dorsey Press, Homewood.

Gastil, J., Xenos, M., 2010. Of attitudes and engagement: clarifying the reciprocal relationship between civic attitudes and political participation. Journal of Communication 60 (2), 318–343.

Gimpel, J.G., Celeste Lay, J., Schuknecht, J.E., 2003. Cultivating Democracy: Civic Environments and Political Socialization in America. The Brookings Institution, Washington, DC.

Greenberg, E., Grunberg, L., Daniel, K., 1996. Industrial work and political participation: beyond 'simple spillover'. Political Research Quarterly 49 (2), 287–304.

Hansen, S.B., 1997. Talking about politics: gender and contextual effects on political proselytizing. Journal of Politics 59 (1), 73–103.

Iyengar, S., 1980. Subjective political efficacy as a measure of diffuse support. The Public Opinion Quarterly 44 (2), 249–256.

Jackman, R.W., 1970. A note on intelligence, social class, and political efficacy in children. The Journal of Politics 32 (4), 984–989.

Johnston, R., 2008. Modeling campaign dynamics on the web in the 2008 National Annenberg Election Study. Journal of Elections, Public Opinion, and Parties 18 (4), 401–412.

Kahne, J.E., Westheimer, J., 2006. The limits of efficacy: educating citizens for a democratic society. PS: Political Science & Politics 39 (2), 289–296.

Karp, J.A., Banducci, S.A., 2008. Political efficacy and participation in twenty-seven democracies: how electoral systems shape political behaviour. British Journal of Political Science 38 (2), 311–334.

Lane, R.E., 1959. Political Life: Why and How People Get Involved in Politics. The Fee Press, New York.

Langton, K., Jennings, M.K., 1968. Political socialization and the high school civics curriculum in the United States. American Political Science Review 62 (3), 852–867.
Lee, K.M., 2006. Effects of Internet use on college students' political efficacy. Cyberpsychology & Behavior 9 (4), 415–422.
Leighley, J., 1991. Participation as a stimulus of political conceptualization. The Journal of Politics 53 (1), 198–211.
Levi, M., Stoker, L., 2000. Political trust and trustworthiness. Annual Review of Political Science 3, 475–507.
Lipset, S.M., Scheider, M., 1983. The Confidence Gap. The Free Press, New York.
Lipsky, M., 1968. Protest as a political resource. The American Political Science Review 62 (4), 1144–1158.
Lussier, D.N., 2011. Mobilisation or resignation? The role of political efficacy in Russia's authoritarian reversal. In: Annual Meeting of the American Political Science Association. Seattle, 1st September to 4th September.
Lyons, S.R., 1970. The political socialization of ghetto children: efficacy and cynicism. Journal of Politics 32 (2), 288–304.
Macedo, S., et al., 2005. Democracy at Risk: How Political Choices Undermine Citizen Participation, and What We Can Do about It. Brookings Institution, Washington, DC.
McGue, M., Bouchard, T.J., 1998. Genetic and environmental influences on human behavioral differences. Annual Review Neuroscience 21, 1–24.
Mendelsohn, M., Cutler, F., 2000. The effect of referendums on democratic citizens: information, politicization, efficacy and tolerance. British Journal of Political Science 30 (4), 669–698.
Miller, A.H., 1974. Political issues and trust in government: 1964–1970. The American Political Science Review 68 (3), 951–972.
Miller, W.E., Miller, A.H., Schneider, E.J., 1980. American National Election Studies Data Sourcebook, 1952–1978. Harvard University Press, Cambridge, MA.
Mondak, J.J., Halperin, K.D., 2008. A framework for the study of personality and political behaviour. British Journal of Political Science 38 (2), 335–362.
Morrell, M.E., 2003. Survey and experimental evidence for a reliable and valid measure of internal political efficacy. Public Opinion Quarterly 67 (4), 589–602.
Morrell, M.E., 2005. Deliberation, democratic decision-making and internal political efficacy. Political Behavior 27 (1), 49–69.
Muller, E.N., 1970. Cross national dimensions of political competence. The American Political Science Review 64 (3), 792–809.
Niemi, R.G., Craig, S.C., Mattei, F., 1991. Measuring internal political efficacy in the 1988 national election study. The American Political Science Review 85 (4), 1407–1413.
Nilson, D.C., Nilson, L.B., 1980. Trust in elites and protest orientation: an integrative approach. Political Behavior 2 (4), 385–404.
Nye, J.S.J., 1997. Introduction: the decline of confidence in government. In: Nye, J.S.J., Zelikow, P.D., King, D.C. (Eds.), Why People Don't Trust Government. Harvard University Press, Cambridge, MA.
Pateman, C., 1970. Participation and Democratic Theory. Cambridge University Press, Cambridge.
Pattie, C., Seyd, P., Whiteley, P., 2004. Citizenship in Britain: Values, Participation and Democracy. Cambridge University Press, Cambridge.
Pharr, S.J., Putnam, R.D., 2000. Disaffected Democracies. What's Troubling the Trilateral Countries. Princeton University Press, Princeton.
Pierce, R., Converse, P.E., 1989. Attitudinal roots of popular protest: the French upheaval of May 1968. International Journal of Public Opinion 1 (3), 221–241.
Pietsch, J., Miller, M., Karp J. Public opinion towards democracy: citizens in transnational regimes. Journal of Elections, Public Opinion and Parties, forthcoming.
Pingree, R., 2009. Is there truth out there? Neutral reporting and epistemic political efficacy. In: Annual Meeting of the International Communication Association. Chicago, 21st May to 25th May.
Pollock III, P.H., 1983. The participatory consequences of internal and external political efficacy: a research note. The Western Political Quarterly 36 (3), 400–409.
Prewitt, Kenneth, 1968. Political efficacy. In: International Encyclopedia of the Social Sciences, vol. 12. MacMillan, London.
Rosenberg, M., 1954. Some determinants of political apathy. Public Opinion Quarterly 18 (4), 349–366.
Rosentstone, S.J., Hansen, J.M., 1993. Mobilization Participation, and Democracy in America. Macmillan, New York.
Seligson, M.A., 1980. Trust, efficacy and modes of political participation: a study of Costa Rican peasants. British Journal of Political Science 10 (1), 75–98.
Shi, T., 2000. Political culture: a prerequisite for democracy? American Asian Review 18 (2), 53–83.
Shrivastava, R., 1989. Political efficacy: a comparative study of the United States, the United Kingdom and India. Youth and Society 21 (2), 170–195.
Siegelman, L., Feldman, S., 1983. Efficacy, mistrust and political mobilization: a cross-national analysis. Comparative Political Studies 16 (1), 118–143.
Smetko, H.A., Valkenburg, P.M., 1998. The impact of attentiveness on political efficacy: evidence from a three-year German panel study. International Journal of Public Opinion 10 (3), 195–210.
Soss, J., 1999. Lessons of welfare: policy design, political learning, and political action. American Political Science Review 93 (2), 363–380.
Stenner-Day, K., Fischle, M., 1992. The effects of political participation on political efficacy: a simultaneous equations model. Australian Journal of Political Science 27 (2), 282–305.
Stoker, G., 2006. Why Politics Matters. Palgrave, Basingstoke.
Thompson, D., 1970. The Democratic Citizen. Cambridge University Press, Cambridge.
Torcal, M., Montero, J.R., 2006. Political Disaffection in Contemporary Democracies: Social Capital, Institutions and Politics. Routledge, London.
Valentino, N.A., Hutchings, V.L., Banks, A.J., Davis, A.K., 2009. Is a worried citizen a good citizen? Emotions, political information seeking, and learning via the Internet. Political Psychology 29 (2), 247–273.
Vecchione, M., Caprara, G.V., 2009. Personality determinants of political participation: the contribution of traits and self-efficacy beliefs. Personality and Individual Differences 46 (4), 487–492.
Weaver, D., 1994. Media agenda setting and elections – Voter involvement or alienation. Political Communication 11 (4), 347–356.
Whiteley, P., 2011. Political Participation in Britain: The Decline and Revival of Civic Culture. Palgrave Macmillan, Oxford.
Wright, J.D., 1975. Does acquiescence bias the index of political efficacy? Public Opinion Quarterly 39 (2), 219–226.
Xena, C. Does the Concept of Political Efficacy Travel across National Borders?, unpublished.
Zimmerman, M.A., 1989. The relationship between political efficacy and citizen participation: construct validation studies. Journal of Personality Assessment 53 (3), 554–566.

Egalitarianism

Christian Hiebaum, Department of Legal Philosophy, Sociology of Law, and Legal Informatics, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Equality is a notoriously controversial value and highly contested political concept. In this article, the aim is to introduce some basic problems and different ways of thinking about equality as the main ingredient of any egalitarian conception of social justice. Section The Beginnings of Modern Egalitarian Social Thought gives a rough sketch of the early history of modern egalitarian theorizing; Section Post-Rawlsian Egalitarianism presents two paradigms of recent egalitarian reasoning about the meaning of equality; Section The Site of Egalitarian Demands addresses the debate about the proper site of equality; and Section The Scope of Egalitarian Demands deals with different egalitarian approaches to the question of the scope of equality.

Introduction

'Egalitarianism,' as it is understood here, applies to left-liberal, social-democratic, socialist, and anarchist political and social theories that focus on, and proceed from the assumption of, equality as the core of social justice. Though it is perfectly reasonable to say that one may be an egalitarian in this but not in that social context (for example in the realm of health care or family or sexual relations but not with regard to the distribution of income and wealth), in this article, egalitarianism is treated as a broader (often implicit) approach to questions of how society should and could be organized. As different social systems, institutions, and practices depend on each other, however, such a broader approach also pertains to the analysis of social microcontexts.

Equality, above basic equality of humans as moral subjects, is a notoriously controversial value and highly contested political concept. Who is or should be equal to whom? In what sense, or in what respect, should people be equal? How does equality relate to other values such as freedom (or liberty), solidarity, peace, and efficiency? Do norms of equality as principles of social justice only apply to social institutions or also to individual conduct? How much inequality does actually exist within different nation-states and at the global level, and how is it to be explained? Is equality a viable social state, and is it achievable? While the former issues are primarily dealt with by ethicists and political philosophers, the answer to the questions of the degree of existing inequalities and of viability and achievability more obviously depends on knowledge only social and behavioral sciences can provide.

Thus egalitarianism also inspires empirical research and analysis, particularly in the fields of economics, sociology, and psychology. As a matter of fact, egalitarian, or 'critical' or 'progressive' or 'emancipatory,' theorizing and research shows a strong tendency to blur the boundaries between empirical and explanatory disciplines as well as between social science and normative reflection without, however, fully abandoning the division of academic labor as it is suggested by the distinction between the issues of desirability, viability, and achievability (see Wright, 2010: 20–25).

In this article, the main focus lies not on equality itself. The aim is rather to introduce some basic problems and different ways of thinking about equality as the main ingredient of any egalitarian conception of social justice, as a ground of social criticism, and as a goal of, as well as a constraint on, political practice. Section The Beginnings of Modern Egalitarian Social Thought gives a rough sketch of the early history of modern egalitarian theorizing; Section Post-Rawlsian Egalitarianism presents two paradigms of recent egalitarian reasoning about the meaning of equality; Section The Site of Egalitarian Demands addresses the debate about the proper site of equality and social justice; and Section The Scope of Egalitarian Demands deals with different egalitarian approaches to the question of the scope of equality and social justice.

The Beginnings of Modern Egalitarian Social Thought

Egalitarian thought has a long history (see Phelps Brown, 1991). The history of egalitarianism as a social attitude and practice is even longer (see Boehm, 2000). In the following, however, we will only take a look at the debates and theoretical frameworks right before the development of the most recent egalitarian theories of justice.

The beginning of political modernity is marked by freedom and equality together with solidarity becoming political values (that is, values to be realized in this world and not just in afterlife). In the early days it used to be mainly natural rights doctrines that grounded criticism of social hierarchies and privileges and policies of liberalization and democratization. Social thinkers of the Enlightenment (among others Baruch de Spinoza (1632–77), John Locke (1631–1704), Pierre Bayle (1647–1706), Baron de Montesquieu (1689–1755), Voltaire (1694–1778), and Jean-Jacques Rousseau (1712–87)) provided for the rationalization of discontent that drove movements such as the Levellers and later on the revolutionaries in France and other countries. Roughly at the same, though not exactly for the first time, feminists *avant la lettre* such as de Olympe de Gouge (1748–93) and Mary Wollstonecraft (1759–97) criticized the marginalization and exclusion of women in social life as well as in human rights thought.

The nineteenth century brought about another broadening of the focus of social criticism. It became clear that equality

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03127-5

requires more than equal (formal) liberty rights, that it is incompatible with huge economic, political, and other social differences – particularly, with large parts of the population living in misery and poverty, exploited and without access to political decision-making. The theoretical frameworks most useful to egalitarian movements were utilitarianism and socialism. Both utilitarian ethics and economics and socialist theory suggested more or less radical reforms toward more equal socioeconomic conditions or, in the case of Marxism, even predicted a revolution supposed to implement equality or 'real freedom' (as opposed to mere formal liberty).

Utilitarianism pervaded economics and became a major precursor of modern welfare economics. It may be characterized as proto-egalitarian. While utilitarians did not consider equality as an end in itself, they proceeded from the assumption that happiness or welfare in society should be maximized without distinguishing according to preexisting status and wealth. Furthermore, they had a powerful argument for more equal distribution of income and wealth: the 'law of diminishing marginal utility' of money which supports redistribution, for example, by way of progressive taxation, from the rich to the poor as a means to enlarge, directly as well as indirectly (through various side effects), the aggregate welfare of the society.

Whereas utilitarianism as a social, political, and economic theory thought of equality as a means to increase total utility, socialism treated equality as an end in itself, identical with real freedom. However, in contrast to other forms of socialist thought (referred to as 'utopian'), the most systematic and thoroughly theorized version of socialism did not conceive itself as a normative doctrine, let alone an ethical conception. Still, there are good reasons to ascribe an implicit egalitarian theory of justice to Karl Marx (1818–83) and his followers.

To be sure, there have always been serious opponents to egalitarian thought: conservative, libertarian, and social Darwinist political, social, and economic theories. At the beginning of the twentieth century, another intellectual threat to theoretical frameworks that were, at least, friendly to egalitarians appeared: skepticism about the possibility of interpersonal comparisons of well-being together with the idea that scientific analysis ought to be value-neutral. Later on, however, Keynesian macroeconomics provided for a new paradigm attractive to moderate egalitarians by giving them new arguments to keep economic inequalities within limits. More radical egalitarians without sympathy for a Sowjet-style economy built their hopes on almost simultaneously emerging theories of market socialism. Yet Keynesianism proved to be politically more effective in the Western world, that is, a framework for policy debates and bargaining between social partners antiegalitarians had a much harder time to reject.

For quite a long period of time, however, systematic reasoning about social justice, about its meaning and principles, was not considered a worthwhile academic exercise. This period came to an end when John Rawls's *Theory of Justice* (Rawls, 1971) appeared. The question of what it means to live in a society of free and equals became a theoretical issue again – dealt with not only by philosophers but also by economists and other social scientists. Though Rawls's liberal conception of social justice was quite commonly considered egalitarian, doubts among sympathetic critics arose as to whether it was properly or sufficiently egalitarian. Hence the question what 'equality' and 'egalitarian' actually mean became an important issue in political theory (while at the same time Hayekian neoliberalism gained momentum in extra-academic politics). Even pronounced socialists who had lost faith in the traditional modes of Marxist analysis participated in the debate reconstructing (parts of) Marxist social theory by making use of the methods of neoclassical economics, game theory, and analytic philosophy (Roemer, 1986). Egalitarianism ceased to be mainly instrumental egalitarianism or a prediction. It became a paradigm of its own in normative theory and its associated fields of empirical research (Bowles and Gintis, 1998; Holtug and Lippert-Rasmussen, 2007). In contrast to the other major debate following *A Theory of Justice*, namely the debate between liberals and communitarians, the controversy about egalitarianism continues to flourish.

Post-Rawlsian Egalitarianism

Today, not only is there much disagreement between egalitarians and non- or antiegalitarians, we do not have much agreement between professed egalitarians about the proper form of egalitarianism either. But genuine egalitarians seem to have one thing in common: they believe in equality not just for instrumental reasons. They consider equality also intrinsically valuable or morally required regardless of consequences, though social scientists with an egalitarian political outlook naturally tend to focus on the consequences of real or hypothetical institutions of equality (asking which egalitarian institutions work and which do not, and why some such institutions work or fail or do not come into existence in the first place).

On the other hand, bad consequences of what appears to be an egalitarian social arrangement may raise doubts among egalitarians about their conception of equality. Social facts as well as facts about the genetic endowment and psychological capabilities or propensities of human beings may influence beliefs of what equality, properly understood, consists in. They may be taken as specifications of what Rawls, following David Hume, has termed "the circumstances of justice." However, not all egalitarians accept the idea that empirical research can contribute to the determination of the content of fundamental moral values – just like positivistically minded social researchers typically want to keep their work free from 'unscientific' and purportedly necessarily subjective reflection about moral values. At least it remains undisputed that the application of norms of social justice and the realization of egalitarian social arrangements require lots of empirical knowledge. For example, whatever we may think of the Rawlsian difference principle, whether it is fulfilled or whether the capitalist welfare state is capable of maximizing the minimum, certainly is, to a considerable degree, an empirical question dealt with by social scientists as well as philosophers.

Egalitarians agree that many political and economic inequalities of the modern capitalist society cannot be plausibly justified by any reference to social equality. Beyond that, the views of egalitarians diverge. We can roughly distinguish two sorts of egalitarians: those who cling to a narrow,

distributive paradigm of equality and those who are committed to a more comprehensive, relational understanding. The former approach seems to prevail in egalitarian analytic political philosophy and welfare economics, while the latter is more popular among egalitarians that prefer to blend philosophy, sociology, and anthropology.

Narrow Egalitarianism

Narrow egalitarianism, which, for some time from the 1980s onward, almost epitomized egalitarianism, focuses not on relations between people but on the distribution of goods. Accordingly, 'equality' means equal distribution of goods. Given such a definition, the question arises as to which goods are to be distributed equally. And given that some differences and inequalities are unavoidable or obviously justified, we have to look for an abstract cardinal good: (access to) welfare, opportunities, (access to) advantages, capabilities, resources, or something else. Narrow egalitarians have spent many years searching for such a cardinal good and discussing various proposals (Arneson, 2011; Hausman and McPherson, 2002: 135–159). Furthermore, they maintain that the cardinal good should be distributed in a luck-neutralizing way. This line of reasoning constitutes what is now termed 'luck egalitarianism.' It is inspired by Rawls's suggestion "to regard the distribution of natural talents as a common asset" and to not let the morally arbitrary natural lottery determine the distribution of social goods (Rawls, 1971: 101). According to luck egalitarianism, "an unequal distribution whose inequality cannot be vindicated by some choice or fault or desert on the part of (some of) the relevant affected agents is unfair, and therefore, *pro tanto*, unjust, and nothing can remove that particular injustice" (Cohen, 2008: 7). In other words, the distribution should be ambition-sensitive and endowment-insensitive (Dworkin, 2000: chapter 1 and 2).

Whatever the cardinal good that is supposed to be distributed in such a way, it seems clear that this account of equality relies heavily on a distinction between choice and (unchosen) natural as well as social circumstances. By that, luck egalitarianism retains some connection to widespread intuitions about why, for example, racist and sexist discrimination is wrong. For, roughly speaking, being discriminated is commonly conceived of as being disadvantaged (by authorities, norms, and institutions) just because of an unchosen property one cannot (reasonably be expected to) dispose of. Luck egalitarianism also appears compatible with the belief that humans are responsible agents, and that individual choices and efforts should make a difference in the distribution of goods, especially of material wealth. At least in theoretical discourse, egalitarianism thus takes a powerful idea away from its opponents and incorporates it into a critique of precisely those inequalities that used to be defended through reference to individual responsibility. It seems to take responsibility and individual desert more seriously than most self-declared proponents of meritocracy on the antiegalitarian side.

However, according to other egalitarians (and, needless to say, of non- and antiegalitarians), the fact that luck egalitarians lay so much stress on the distinction between choice and circumstance amounts to a serious shortcoming of their conception of social justice. For this distinction not only opens up the field for empirical research, it may also lead us deeper into metaphysical terrain than social analysts as well as political thinkers and actors are willing to go. Furthermore, by itself luck egalitarianism looks sociologically sterile. It seems to be mainly concerned with the question of desirability (less with feasibility, and still less so with achievability) and tends to separate distributive equality all too neatly from other politically relevant values. For luck egalitarians, says G.A. Cohen, "are interested in the very nature of distributive justice, not in the different question as to what principles, to be influenced by distributive justice *but also by other things*, a society should adopt as its basic ones" (Cohen, 2008: 301).

Just think of persons who remain paralyzed after being injured due to imprudence or as a consequence of a risky undertaking such as mountain climbing. Most people believe that they have a right to special therapy and care. And many, at least in Europe, would hesitate to bill them for the full costs of such provisions and to consider social disadvantages resulting from so caused disability not in need of further justification just due to their being 'self-inflicted.' Now, luck egalitarians might reply that they are not that stonehearted either, that there are other values beside equality to be taken into account. Often equality has to be traded off against solidarity, the right to health care, or subsistence, etc. Equality is just one important value among others. Some luck egalitarians even hold that to attribute intrinsic value to equality implies the belief that a society consisting only of blind people is better in at least one respect than a society in which not all are blind. It's just other values which, in the latter society, prevent egalitarians from putting out the eyes of the sighted (Temkin, 1993: 257).

Comprehensive Egalitarianism

In contrast to luck egalitarianism, comprehensive egalitarianism is primarily concerned with equality as an ideal referring to relations between people free from oppression and dominance (Anderson, 1999, 2012; Scheffler, 2003; Walzer, 1983; Young, 1990). Though to some degree a reaction to luck egalitarianism and what is perceived as its shortcomings, comprehensive egalitarianism seems, in spirit and conceptually, closer to historical and current political movements against, and empirical analyses of, social inequality. Accordingly, egalitarians are supposed to take aim at "different types of social hierarchy, including slavery, serfdom, debt patronage, feudalism … oligarchy, caste and class inequality, racism, patriarchy, colonialism, and stigmatization based on sexuality, disability, and bodily appearance" (Anderson, 2012: 40). It is not denied that social relations are shaped by distributions of goods and that, among the dimensions of social justice, distributive justice has a special relevance. Yet according to comprehensive egalitarians, it is a variety of social goods the distribution of which brings about relations of equality or inequality, and some goods, such as basic legal rights, have to be distributed quite equally regardless of choice and luck.

Thus comprehensive egalitarians, while granting that in some contexts justice has a lot to do with neutralizing luck, do not aim at determining a cardinal good whose equal (or ambition-sensitive/endowment-insensitive) distribution defines equality. The questions guiding their research and reflection are: What is the role of social equality in evaluating

institutions and actions? Which set of principles of distribution and of the treatment of human beings are compatible with, or conducive to, social equality under which circumstances, or rather in which contexts? When, for example, does the argument of differential effort or productivity outweigh the argument of need? What is the relevance of legitimate expectations? How are egalitarian societies organized, or how could they be organized? Which principles are egalitarian as such, that is, just in virtue of their content, is not an issue comprehensive egalitarians are much interested in.

This version of egalitarianism is comprehensive in two respects: first, it is not just concerned with economic distributive justice but also with transactional justice, political justice, and corrective justice; and second, it refers more to all-embracing social orders and less to distributional microcontexts. Thus, it implies, or demands for, some kind of second-order theory, even of distributive justice. Assuming that there will always be disagreement about the justice of particular distributions, comprehensive egalitarians deny that social equality amounts to distributive justice all the way down. They maintain that social equality should not be identified with perfect (and notoriously contested) economic distributive justice – that, up to a certain degree, it allows for deviation from distributive ideals provided that there remains a political space for challenging ideals and what is taken as deviations.

So, comprehensive egalitarians conceive of social equality as a feature of persons living in various overlapping social relationships that are partly characterized by distributions of goods, not as a feature of distributional patterns themselves. Whatever social equality amounts to in detail, comprehensive egalitarians define it as a distributional state together with certain (though perhaps not fully compatible) public justifications. For equals have to have good reasons to consider themselves equal. An accidental distribution which is compatible with equality at a particular point of time is not enough. Neither is the mere belief that we are all equals. Accordingly, any strict distinction between egalitarian distribution and its justification is a priori impossible. Distribution relates to its justification roughly like equality of outcome to the treatment of people as equals, or like equality as a political goal to equality as a side constraint of political action.

Not only do luck egalitarians disagree with each other, there is also lots of disagreement among comprehensive egalitarians. A good deal of disagreement has its sources not just in different views of what is just or unjust, but in different perceptions of how society is best described and in competing accounts of existing hierarchies. Take the debate about multiculturalism. It is now commonplace that 'progressive' or leftist social science all too long ignored the fact that there are noneconomic dimensions of social life that cannot be treated as mere epiphenomena of the division of classes. One such dimension is supposed to be culture. Though most political thinkers and social scientists agree that, in modern society, no social sphere is in principle more basic than all the others, there is still some debate about whether culture constitutes such a sphere, whether 'culture' is an analytically useful concept, what 'cultural equality' means and how cultural identity relates to other aspects of individual identity, and whether collective rights to 'cultural identity' should rank high on the egalitarian political agenda (see Barry, 2001; Kymlicka, 1995).

Debates such as this one (with a notorious propensity to go on forever) might be viewed as a serious weakness of comprehensive egalitarianism. For they suggest that it lacks conceptual, normative, and methodological coherence, at least compared to luck egalitarianism. Comprehensively egalitarian theories and analyses seem to display little more than family resemblances. To luck egalitarians, comprehensive egalitarianism looks like a set of theories that are not properly characterized as egalitarian at all, but rather as prioritarian (such as the theory of Rawls, 1971), or even merely sufficientarian (such as the conception of 'democratic equality' of Anderson, 1999). It appears prioritarian insofar as it proposes principles that demand for special consideration of the fate of the least advantaged. And it appears sufficientarian to the degree that it only requires us to make sure that no member of society falls below a certain threshold of well-being (whatever that threshold may be). That comprehensive egalitarianism is less vulnerable to antiegalitarian challenges (as it incorporates a plurality of values into the framework of equality) thus does not seem to be much of an asset. On the other hand, the looseness of comprehensive egalitarianism might simply be the price of not restricting political thinking and social research to conceptual analysis, questions of desirability, and abstract modeling. That is, it might just be a reflex of the complexity of social reality itself with its multiple layers and intersections of differences and inequalities.

The Site of Egalitarian Demands

Another question is whether egalitarianism should be concerned exclusively with institutions or whether it should also cover personal behavior, or rather: whether norms that apply to institutions and are supposed to guarantee equality should equally guide us in our private lives. This, of course, is an issue every theory of social justice (not just egalitarian ones) has to deal with.

Historically, liberalism as well as socialism (particularly orthodox Marxism) used to show a strong bias toward institutions and structures and to be highly skeptical about moralizing. Such a bias even seems to be typical of secular social thought. Religious social ethics has always been more open to the idea of individual duties of social justice and solidarity transcending the duty to create and preserve just institutions.

So what, if anything, does equality as the central value of social justice require from us in terms of personal behavior? Those who believe that equality applies equally to institutions and personal conduct use to be referred to as 'monists,' those who insist on a more or less strict separation of the sites of social justice are called 'dualists.'

Monism

Monists claim that norms of equality should also motivate us to do more than egalitarian institutions demand from us (see Cohen, 2000). The wealthy, for example, should also give part of their after-tax income to the poor, directly or indirectly via

charity organizations or by way of unsalaried community work, in order to further reduce economic inequality. Or to put it differently: a widespread individual ethos is part of the basic institutions of society. In this sense, the personal is considered political.

That is, monists reject the idea that if we get the economic and legal institutions right individuals may always act selfishly, provided they comply with the institutionalized rules. On this view, the attitude of selfish individuals who still affirm, and comply with, egalitarian institutions, though psychologically possible, is less than fully coherent.

Dualism

Dualists insist that it does not make good sense to require individuals to comply with principles such as Rawls's difference principle for pragmatic, epistemic, and moral reasons (Pogge, 2000). Sharing more of one's wealth than egalitarian institutions require with the poorer members of society may be virtuous but it is not a clear duty that follows from a reasonable egalitarian conception of social justice (be it a luck-egalitarian or a comprehensive-egalitarian conception).

Dualists concede that the proper design of the institutions may, to some degree, depend on the ethos spread in a population. For after all, how progressive a tax system should be varies with what financial incentives are necessary for the talented to make efforts which also benefit the others. However, due to the fact that not only the poor but also the rich and productive face collective action and free rider problems, attempts at blackmailing the rest need not to be taken very seriously. That is, claims of the wealthy to the contrary notwithstanding, we could, at least through international cooperation, impose significantly higher taxes on high income and wealth and thus create much more economic equality without necessarily reducing the size of the cake that is to be distributed.

The Scope of Egalitarian Demands

As to the issue of the proper scope of egalitarian thinking as thinking about social justice, we can distinguish between particularistic and cosmopolitan approaches. This distinction cuts across the distinction between luck egalitarianism and comprehensive egalitarianism. (Though, at first glance, luck egalitarians might be expected to naturally prefer the cosmopolitan account of justice since they use to skip over the question about the kinds of social relationships that have to obtain for equality to become an issue at all. Thus they appear to be committed to extend the scope of equality beyond relations between compatriots or members of a community of another type.)

Particularists argue that egalitarian conceptions of social justice or norms of equality only apply to already existing political communities, while cosmopolitans believe that social justice with equality at its center has a significant transnational or global dimension. Obviously, particularism corresponds to a long tradition in social thought of roughly (and often implicitly) equating society with something organized as state, sometimes referred to as the 'container theory of society' or "methodological nationalism" (Beck and Sznaider, 2006). In the field of empirical research and theorizing, this tradition is reflected by the (once quite strict) disciplinary division of political science and international relations theory. In contrast, cosmopolitanism seems to be more in line with antistatist, internationalist, or globalist accounts of society (which, to be sure, also have a history that goes back to the beginnings of the social sciences).

Particularistic Egalitarianism

According to particularistic egalitarianism, equality is a domestic or intracommunal issue or something primarily concerning the treatment of citizens by their government (Blake, 2001; Miller, 2007). There may be moral norms whose scope transcend communal boundaries and state borders, but either (1) they are not norms of social justice, as opposed to, for example, norms that impose humanitarian duties, or (2) if they are norms of justice, they do not derive from the value or some norm of equality, as opposed to, for example, the principle of sufficiency.

Particularistic egalitarians, just like many non- or antiegalitarians, do think that the affluent have moral duties toward the global poor, or, at least, a collective responsibility. However, they deny that the ideal of equality makes good sense at the global level. In addition to considerations of feasibility and achievability and (occasionally) relativistic assumptions about the origins, existence, and objectivity of moral values, particularistic egalitarians put forward, at least, three arguments: (1) the argument from cooperation; (2) the argument from solidarity; and (3) the argument from coercion.

1. Egalitarian demands presuppose a certain type of cooperation, that is, not just 'ordinary' social cooperation but also a cooperative determination of the terms of social cooperation through democratic decision-making. At the global level, we may have interaction that can be described as cooperation, yet we lack institutions of political cooperation and democratic decision-making. Without such basic political institutions, it does not make sense to raise claims to equality against each other.
2. The argument from cooperation is often supplemented by the claim that political cooperation depends on a sense of solidarity and belonging, on loyalty that cannot be directed at the whole of mankind but is naturally much more exclusive. Social equality transcends mere equal moral standing and thus requires a closer bond between individuals than mutual respect. This argument may be backed by anthropological evidence about the limits of our inborn capability to identify with socially or geographically distant others.
3. Egalitarian concerns are triggered by the use of political force directly constraining individual autonomy. For such coercion creates a demand for justification that also requires the application of 'relative' principles of justice. Since there is no such direct political coercion of individuals at the global level, the issue of equality remains restricted to the domestic level.

Cosmopolitan Egalitarianism

According to cosmopolitan egalitarianism, equality is not only relevant to relations between members of a particular political

community, or to the distribution of goods and burdens among them, but also to the design of global political and economic institutions. It may even be a ground of individual duties toward people in other parts of the world. To the degree that society is globalized, social justice should have a global dimension that transcends mere international justice (Beitz, 1999; Brock, 2009; Caney, 2005).

With regard to social theory and empirical analysis, cosmopolitans reject what they refer to as "explanatory nationalism" (Pogge, 2002: 139–144), that is, the tendency to deny or ignore the impact of global (particularly economic) institutions on the distribution of resources and on democracy within the state. With regard to the particularistic arguments mentioned above, most cosmopolitans concede that such arguments give us good reasons not to believe that social equality at the global level means exactly the same as domestic social equality or that we should hope for particular political communities or states to give way to a world state. Still, cosmopolitans deny that these arguments convincingly define the scope of social equality:

1. While it is true that social equality as a matter of fact depends on political cooperation, it may also create a demand for new or modified political institutions. What is more, we already have firmly institutionalized global political decision-making, for example, within the World Trade Organization. But it is highly and unfairly biased in favor of the most powerful states and economic actors. Moreover, some global institutions, such as the borrowing and resource privilege (Pogge, 2002: 112–116), create strong incentives for governments to not even try to represent the interests of their citizens or, prior to that, for particular groups within a country to seize power by undemocratic or violent means.
2. Solidarity between compatriots is solidarity between strangers, consisting, to a large degree, not in mutual sympathy but in accepting, and complying with, institutions and the underlying quite abstract values. Patriotic solidarity is certainly valuable, but it is also a highly 'artificial' thing (compared to egalitarianism in hunter-gatherer societies) partly generated precisely by the political, educational, and welfare state institutions it is supposed to support.
3. Apart from the fact that some global or transnational institutions may constrain individual autonomy quite directly, we have also lots of indirect political coercion within the state whose justification includes considerations of equality and requires the application of relative principles of justice. Take as an example tax law which needs to be justified not only to those who are taxed but also to those who may be indirectly affected (through higher prices, less capital investment, etc.).

Conclusion

What seems clear is that reasonable egalitarianism applying to modern society, whether national or transnational, differs in many ways from egalitarianism as an attitude and practice in much less complex communities or in social microcontexts. It certainly requires a significant widening of moral concern and the ability to get motivated by more abstract conceptions of justice. One might thus believe that, today, egalitarianism deserving its name demands too much from people. As a matter of fact, early 'gene-centered' sociobiology, sometimes contrary to the intentions of its exponents, showed a tendency to nourish skepticism about egalitarianism beyond kin groups. However, more recent research in evolutionary psychology, experimental economics, and game theory, though far from supporting a view of human beings as fully constituted by culture and social institutions and therefore radically malleable, let modern egalitarian hopes appear much less incompatible with human nature (Bowles and Gintis, 2011).

See also: Analytical Marxism; Discrimination and the Law; Enlightenment: Impact on the Social Sciences; Equality of Opportunity; Equality: Philosophical Aspects; Inequality Analysis: Overview; Justice: Philosophical Aspects; Keynesianism and the Question of State Interventionism; Liberalism: Impact on Social Science; Multiculturalism; Political Economy, History of; Political Theory, History of; Science and Politics: Value Neutrality; Social Justice: Historical and Theoretical Considerations; Social Question: Impact on Social Thought; Socialism; Utilitarian Social Thought, History of.

Bibliography

Anderson, E.S., 1999. What is the point of equality? Ethics 109, 287–337.
Anderson, E.S., 2012. Equality. In: Estlund, D. (Ed.), The Oxford Handbook of Political Philosophy. Oxford University Press, Oxford, pp. 40–57.
Arneson, R.J., 2011. Luck egalitarianism – a primer. In: Knight, C., Stemplowska, Z. (Eds.), Responsibility and Distributive Justice. Oxford University Press, Oxford, pp. 24–50.
Barry, B., 2001. Culture and Equality: An Egalitarian Critique of Multiculturalism. Polity Press, Cambridge.
Beck, U., Sznaider, N., 2006. Unpacking cosmopolitanism for the social sciences: a research agenda. The British Journal of Sociology 51, 1–23.
Beitz, C., 1999. Political Theory and International Relations. Princeton University Press, Princeton.
Blake, M., 2001. Distributive justice, state coercion, and autonomy. Philosophy and Public Affairs 30, 257–296.
Boehm, C., 2000. Hierarchy in the Forest: The Evolution of Egalitarian Behavior. Harvard University Press, Cambridge.
Bowles, S., Gintis, H., 1998. Recasting Egalitarianism: New Rules for Communities, States and Markets. Verso, London, New York.
Bowles, S., Gintis, H., 2011. A Cooperative Species: Human Evolution and Its Evolution. Princeton University Press, Princeton, Oxford.
Brock, G., 2009. Global Justice: A Cosmopolitan Account. Oxford University Press, Oxford.
Caney, S., 2005. Justice Beyond Borders: A Global Political Theory. Oxford University Press, Oxford.
Cohen, G.A., 2000. If You're an Egalitarian, How Come You're So Rich? Harvard University Press, Cambridge, MA, London.
Cohen, G.A., 2008. Rescuing Justice and Equality. Harvard University Press, Cambridge.
Dworkin, R., 2000. Sovereign Virtue: The Theory and Practice of Equality. Harvard University Press, Cambridge, London.
Fraser, N., 1997. From redistribution to recognition? Dilemmas of justice in a "postsocialist" age. In: Fraser, N. (Ed.), Justice Interruptus: Critical Reflections on the "Postsocialist" Condition. Routledge, New York, London, pp. 11–39.

Hausman, D.M., McPherson, M.S., 2002. Economic Analysis and Moral Philosophy. Cambridge University Press, Cambridge.
Holtug, N., Lippert-Rasmussen, K. (Eds.), 2007. Egalitarianism: New Essays on the Nature and Value of Equality. Clarendon Press, Oxford.
Kymlicka, W., 1995. Multicultural Citizenship: A Liberal Theory of Minority Rights. Clarendon Press, Oxford.
Miller, D., 2007. National Responsibility and Global Justice. Oxford University Press, Oxford.
Phelps Brown, H., 1991. Egalitarianism and the Generation of Inequality. Clarendon, Oxford.
Pogge, T., 2000. On the site of distributive justice: reflections on Cohen and Murphy. Philosophy and Public Affairs 29, 137–169.
Pogge, T., 2002. World Poverty and Human Rights. Polity, Cambridge.
Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge, MA.
Roemer, J.E. (Ed.), 1986. Analytical Marxism. Cambridge University Press, Cambridge.
Scheffler, S., 2003. What is egalitarianism? Philosophy and Public Affairs 31, 5–39.
Temkin, L.S., 1993. Inequality. Oxford University Press, New York; Oxford.
Walzer, M., 1983. Spheres of Justice: A Defense of Pluralism and Equality. Basic Books, New York.
Wright, E.O., 2010. Envisioning Real Utopias. Verso, London; New York.
Young, I.M., 1990. Justice and the Politics of Difference. Princeton University Press, Princeton.

Eisenberg's Theory of Prosocial Reasoning

Bilge Yagmurlu and Hilal Sen, Koc University, Istanbul, Turkey

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, we present information about Eisenberg's prosocial moral reasoning. Prosocial moral reasoning focuses on the moral justifications behind prosocial actions intended to benefit another person. There are five levels of prosocial moral reasoning, progressing from self-centered concerns to other-oriented and value-related concerns. This change in prosocial moral reasoning is partly related with improving empathic, sympathetic, and perspective-taking skills. Finally, there is a positive association between prosocial moral reasoning and prosocial behavior, but motivational and situational factors are also important in this relationship.

In everyday life, people encounter different dilemmas in which they decide whether to help a person or not; sometimes they may even have to decide whether to violate a law in an emergency situation. The way that people resolve moral conflicts is considered to be related to their level of moral reasoning. Eisenberg defines prosocial moral reasoning as "reasoning about conflicts in which the individual must choose between satisfying his or her own wants or needs and those of others in contexts in which laws, punishments, authorities, formal obligations, and other external criteria are irrelevant or deemphasized" (Eisenberg-Berg, 1979: p. 128). In this sense, *prosocial moral reasoning* focuses on moral judgments that pertain to positive social behaviors intended to benefit another person. The dilemmas that tap such situations are widely experienced in real life.

Eisenberg (1986) proposes a scheme that portrays age-related changes in prosocial moral reasoning, from childhood to adulthood. This scheme is presented in **Table 1**.

According to Eisenberg (1986), there is a progress from self-centered concerns to other-oriented and value-related concerns in prosocial moral reasoning. At the lowest level of prosocial justifications, self-focused (hedonistic) concerns are central; engaging in prosocial behaviors is determined by selfishness (i.e., gaining reward or reciprocity). At the second level, the needs of others are emphasized but not in a sophisticated way. That is, individuals may refer to other's needs without any reference to empathic concerns for the welfare of the needy one. At the third level, engaging in prosocial behavior is decided by the consideration of the approval of others. There is a motivation for having a good impression in others' eyes. At the fourth and fifth levels, more mature forms of reasoning are apparent. At the fourth level, the emphasis is on empathic concerns for other people and taking their perspective. Then, there is a progress to internalized values, norms, and goals to improve the conditions of individuals at the fifth level.

Although the rudimentary needs-oriented orientation in moral reasoning is a relatively low level of prosocial moral reasoning, Eisenberg (1986) argues that needs-oriented moral reasoning is different from self-oriented or pragmatic categories of reasoning, as highlighted by the negative correlation between needs-oriented and self-oriented moral reasoning. She also claims that needs-oriented reasoning overlaps with empathic reasoning categories. Moreover, in their pencil-and-paper measure of moral reasoning used with adolescents and adults, Eisenberg and colleagues (e.g., Carlo et al., 1992; Eisenberg et al., 2002) argue that approval-oriented reasoning is at a lower level than needs-oriented reasoning.

Issues in Kohlberg's Assumptions: Eisenberg's Approach to Moral Reasoning

Eisenberg (1986) questioned Kohlberg's assumptions about moral judgment. In his model of moral reasoning, Kohlberg (1969) proposed that stages of moral reasoning are hierarchically integrated, structurally whole, invariant, and qualitatively different. The assumption of hierarchical integration suggests that when an individual is at higher stages of moral reasoning (stage 4 or 5), he or she reorganizes lower levels of reasoning into his or her current level, so the individual cannot make reasoning at lower levels (stage 1, 2, or 3). Eisenberg (1986) argued that a given individual's moral reasoning can occur at a variety of levels. Within any level of cognitive development, individuals can reason at different levels. She found that adolescents produced both higher and lower types of moral reasoning at the same time for different dilemmas (Eisenberg-Berg, 1979). Therefore, Eisenberg (1986) argued, hierarchical integration is not correct for the nature of moral judgment. Kohlberg's structural wholeness assumption considers each stage as a reflection of an underlying thought organization; hence, responses at a stage are not random, and they reflect an underlying systematic thinking or reasoning. This is in part due to the scoring system of Kohlberg. In his coding scheme, the variation in moral responses is minimized by eliminating infrequent levels of moral reasoning, emphasizing only consistently produced levels of moral reasoning. This procedure loses nuances in moral judgments and results in an artificial homogeneity in moral responses, indicating that individuals generate similar responses as if they are at one stage. According to Eisenberg (1986), however, a structural wholeness assumption is not accurate because people produce reasoning at a variety of levels.

Eisenberg (1986) also argued that invariance in the sequence of stages (i.e., the sequence of stages does not change) does not depict the nature of moral development

Table 1 Eisenberg's levels of prosocial moral reasoning

Level	*Explanation*	*Age*
Level 1. Hedonistic, self-focused orientation	The individual is concerned with self-oriented consequences rather than moral considerations. Reasons for assisting or not assisting another include consideration of direct gain to the self, further reciprocity, and concern for others because one needs and/or likes the other (due to affectional ties).	Predominant mode primarily for preschoolers and younger elementary-school children
Level 2. Needs-oriented orientation	The individual expresses concern for the physical, material, and psychological needs of others even though the other's needs conflict with his or her own needs. This concern is expressed in the simplest terms, without clear evidence of self-reflective role taking, verbal expressions of sympathy, or reference to internalized affect such as guilt.	Predominant mode for many preschoolers and many elementary-school children
Level 3. Approval and interpersonal orientation and/or stereotyped orientation	Stereotyped images of good and bad persons and behaviors and/or considerations of others' approval and acceptance are used in justifying prosocial or nonhelping behaviors.	Predominant mode for some elementary and many high school students
Level 4a. Self-reflective empathic orientation	The individual's judgments include evidence of self-reflective, sympathetic responding or role taking; concern with the other's humanness; and/or guilt or positive emotion related to the consequences of one's actions.	Predominant mode for a few older elementary school children and a moderate number of high school students
Level 4b. Transitional level	The individual's justifications for helping or not helping involve internalized values, norms, duties, or responsibilities; show concern for the condition of the larger society; or refer to the necessity of protecting the rights and dignity of other people. These ideas, however, are not clearly or strongly stated.	Predominant mode for a minority of people high school age or older
Level 5. Strongly internalized stage	Justifications for helping or not helping are based on internalized values, norms, and responsibilities; the desire to maintain individual and societal contractual obligations or improve the condition of society; and the belief in the dignity, right, and equality of all individuals. Positive or negative emotion related to the maintenance of self-report for living up to one's own values and accepted norms also characterizes this stage.	Predominant mode for only a small minority of high school students and no elementary school children

Source: Reproduced from Eisenberg, N., 1986. Altruistic Emotion, Cognition and Behavior, Lawrence Erlbaum, Hillsdale, NJ.

well. Moral reasoning does not follow a strict path, and individuals can show variation in their moral reasoning development. Eisenberg also asserted that moral stages might not be qualitatively different from each other, implying that some stages could be parts of an underlying general theme. As Carlo et al. (2008) demonstrated, adolescents' stage 4 and 5 prosocial moral reasoning responses loaded on the same factor, suggesting that they are parts of an overarching value such as *internalized concerns for other's welfare*. In summary, according to Eisenberg, the stages of prosocial moral reasoning are not invariant, structured whole, qualitatively different, and hierarchically integrated, although they do tend to emerge in a given order due to constraints on children's socio-cognitive abilities. In addition, she argues that individual differences in prosocial moral reasoning are affected by individuals' socio-cognitive skills such as empathy and perspective taking, socialization practices, and situational factors (see subsequent sections of this article for the details).

Age and Gender Differences in Prosocial Moral Reasoning

A series of cross-sectional and longitudinal studies were conducted in order to delineate age-related changes in prosocial moral judgment. Eisenberg and colleagues reported that young children are prone to use lower levels of reasoning; and older children and adolescents tend to use higher levels of moral reasoning (e.g., Eisenberg, 1986; Eisenberg et al., 1995). In other words, the justifications of young individuals often focus on self-related issues and personal gains or future reciprocity. Older individuals, in contrast, rely on more advanced levels of reasoning and emphasize internalized values, sympathetic concern, and perspective taking with regard to engaging in prosocial behaviors. Many other findings have also supported similar age-related differences in other cultures like Turkey (Kumru et al., 2012) and Brazil (Carlo et al., 1996).

Specifically, the findings of Eisenberg's longitudinal studies (Eisenberg et al., 1987, 1991, 1995, 1999, 2002, 2005) indicate that the most dominant response in prosocial moral reasoning interviews is hedonistic (stage 1, including self-oriented responses) in the preschool period, although needs-oriented responding is fairly frequent. During the preschool and elementary school years, the frequency of hedonistic responses decreases, while at the same time need-oriented concerns increase. Approval-oriented responses emerge in elementary school and are most frequent in the secondary school years. In the high school years, empathic concerns, perspective taking, and internalized values become more common, although adolescents more frequently verbalize needs-oriented, stereotypic, and approval-oriented reasoning. Needs-oriented reasoning declined in adolescence, whereas there is evidence of fluctuations in stereotypic and approval-oriented reasoning across adolescence and early adulthood. Overall, those results demonstrated that there is an increase in the sophistication in the level of moral reasoning from childhood to adulthood.

Findings with regard to gender differences are more mixed. A meta-analysis on gender differences in moral orientations indicated that females performed higher on caring and prosocial moral reasoning compared to males, but the difference was small (Jaffee and Hyde, 2000). However, studies also revealed that the marked difference between females and males in higher forms of moral reasoning that emerges in early adolescence (Eisenberg et al., 1991, 1995) begins to diminish by late adolescence (Carlo et al., 1992), probably because maturation in perspective taking and empathy, which are significant correlates of higher forms of moral reasoning, minimizes the gap between females and males. Nonetheless, gender differences favoring women were still evident in the 20s and early 30s, although the higher level of hedonistic reasoning used by males dropped off across the 20s (Eisenberg et al., 2005, 2013).

The reported difference in moral justifications between females and males might be due to gender-specific socialization practices (Eisenberg, 1986; Gilligan, 1982). From very early on, females are encouraged to care for the other, both the younger and older, and are expected to be empathetic with others' concerns. All of these practices endorsed by different parties in society (e.g., family, friends, neighbors, or school) foster the skills of taking others' perspective and understanding their needs. This may underlie females' use of more needs-oriented and empathy-oriented reasoning as a motive for moral justifications.

Socio-Cognitive Skills in Prosocial Moral Reasoning

Eisenberg (1986) emphasized the importance of empathy and perspective-taking skills in moral judgment. Previous theoretical arguments (Kohlberg, 1969; Underwood and Moore, 1982) basically focused on the advancement of cognitive skills as a precursor to sophistication in moral reasoning. In this respect, maturation in taking the perspective of another person cognitively is regarded as central to moral understanding. Nevertheless, along with cognitive aspects, Eisenberg (2005) argued for the role of affective aspects in moral judgment, like empathy and sympathy. In general, improvements in empathy and perspective taking are linked to advanced levels of moral reasoning. As individuals master these skills, they produce more needs-oriented reasoning and internalized values, and lower levels of self-directed or hedonistic reasoning (Eisenberg et al., 1999, 2002).

Affective Side: Empathy

Empathy is considered to activate internalized moral values such as concern for other people's benefit and welfare (Hoffman, 2000). Eisenberg (2005) also highlighted the role of empathy in prosocial moral reasoning and responding. She defined empathy "as an affective response that stems from the apprehension or comprehension of another's emotional state or condition and is similar to what the other person is feeling or would be expected to feel in the given situation" (Eisenberg, 2005: p. 75). In this conceptualization, empathy often evokes two different emotional reactions: *sympathy* and *personal distress*. These two emotional reactions include a certain degree of cognitive aspect such as perspective taking and might be experienced separately or

simultaneously. *Sympathy* refers to "an emotional response stemming from the apprehension of another's emotional state or condition that is not the same as the other's state or condition but consists of feelings of sorrow or concern for the other" (Eisenberg, 2005: p. 75). Thus, feeling concerned about a person, but not experiencing the same feeling, is the characteristic of sympathy. *Personal distress*, on the other hand, is a self-directed emotional reaction, which might be a result of empathic overarousal. Hence, personal distress is related to aversive reactions and the motivation to dampen the intensity of one's own empathic arousal.

Eisenberg claimed that empathy, sympathy, and personal distress are differentially associated with prosocial moral reasoning. Several research findings indicated that empathy (Eisenberg et al., 1987; Eisenberg-Berg and Mussen, 1978) or sympathy was positively linked to higher levels of prosocial moral reasoning and/or less hedonistic reasoning concurrently (Carlo et al., 1992; Eisenberg et al., 2001; Miller et al., 1996) and longitudinally (Eisenberg et al., 1991, 1995). That is, individuals with empathic or sympathetic tendencies expressed more needs-oriented moral reasoning and less benefit-oriented reasoning. Given that empathy-related responding includes 'the apprehension or comprehension of another's emotional state' and 'feeling the same emotional experience' (only for empathy) or 'feeling of concern for other' (only for sympathy), these characteristics are considered the motives for a higher, self-reflective level of moral judgment.

As personal distress is the motive for self-oriented tendencies, it is expected to be related to hedonistic reasoning at the theoretical level (Eisenberg, 1986, 2005). However, the findings with regard to the association between personal distress and prosocial moral reasoning are contradictory. Some findings demonstrated that personal distress was negatively related to higher levels of moral reasoning (i.e., stereotypic and internalized) (Miller et al., 1996) and positively related to approval-oriented reasoning (Carlo et al., 1992). But other findings indicated no significant association between personal distress and prosocial reasoning (Eisenberg et al., 1991, 2002).

Cognitive Side: Perspective Taking

Cognitive capabilities are considered important for moral justifications (Kohlberg, 1969). For example, maturation in perspective taking and other cognitive skills (e.g., abstract thinking) has been considered to be related to age-related changes in prosocial moral reasoning (Carlo et al., 1992) such that as perspective-taking skills improve, prosocial moral reasoning also becomes less hedonistic and more other oriented, needs oriented, and internalized (Eisenberg et al., 1991, 1995).

According to Eisenberg (1986), this association between perspective taking and prosocial moral reasoning might not be very strong or consistent because the empathy or sympathy level may be a source of variation in prosocial reasoning. Eisenberg et al. (2001) found that regardless of their perspective-taking level (low, medium, or high), adolescents with higher levels of sympathy expressed more advanced levels of prosocial moral reasoning. However, for adolescents with low or moderate sympathy, the perspective-taking level was a significant predictor of moral reasoning. This interaction indicated that for individuals with lower or moderate levels of sympathy, cognitive factors contributed to higher levels of prosocial moral reasoning. Moreover, Eisenberg et al. (2001) found that cognitive perspective taking did not have a direct relation to adolescents' prosocial behavior; its relation to prosocial behavior was mediated through prosocial moral reasoning and sympathy, indicating that perspective taking has an indirect influence via either sympathy or prosocial moral reasoning. Eisenberg et al. (2001) concluded that perspective taking fosters prosocial acts only if it promotes sympathy or mature levels of prosocial moral reasoning.

Cross-Cultural Similarities and Differences in Prosocial Moral Reasoning

Cross-cultural studies in prosocial moral judgment mostly focus on the individualism–collectivism continuum. Individualistic cultures value personal goals over communal goals; and they endorse independence, self-reliance, initiative, and economic freedom more. On the other hand, group harmony, cohesion, interdependence, and obedience are the attributes valued more highly in collectivistic societies (Triandis, 1995). So, people in individualistic societies are expected to promote values of self-gain more highly, whereas people in collectivistic societies are expected to emphasize other-oriented concerns more.

Studies have revealed both similarities and differences in prosocial moral reasoning among individualistic societies and between collectivistic and individualistic cultures. The overall patterns of reasoning of German (Eisenberg et al., 1985) and Israeli children (Fuchs et al., 1986) were mostly similar to those of US children (Eisenberg, 1986). German children expressed more direct reciprocity in their moral reasoning, while American children referred to stereotypic reasoning more than did German children (Eisenberg et al., 1985). In the study of Fuchs et al. (1986), moral justifications of Israeli children were found to be higher and more diverse than those of American children. Moreover, it was shown that the responses of Israeli Kibbutz children involved more internalized values and needs-oriented aspects than did the responses of Israeli city and American children. This difference was consistent with their cultural orientation of Kibbutzim, where personal ties and contact are highly emphasized.

Cross-cultural differences in the level of moral reasoning have been noted, especially in the expression of internalized values or affective states. The frequency of this reasoning was lower in the Maisin (Tietjen, 1986) and Brazilian people (Carlo et al., 1996) compared to their American counterparts, and lower in Turkish compared to Spanish adolescents (Kumru et al., 2012). Researchers have argued that this difference between individualistic and collectivistic societies might be due to the educational system in those countries (Carlo et al., 1996; Kumru et al., 2012). It was claimed that the Westernized educational system, which encourages critical understanding and abstraction, might foster more internalized moral reasoning.

The Association between Moral Judgment and Moral Behavior

Does having a certain level of prosocial moral judgment mean that it would result in performing moral behavior such as donating, helping, sharing, or volunteering? An early review of the studies examining the relation between moral cognition and behavior highlighted that there was a positive association between them (Underwood and Moore, 1982). Eisenberg and colleagues have also demonstrated that in the preschool period, spontaneous sharing was negatively related to hedonistic reasoning (Eisenberg et al., 1984a) and positively related to needs-oriented reasoning (Eisenberg et al., 1985; Eisenberg-Berg and Hand, 1979). Other types of prosocial behaviors, including spontaneous helping and peer-initiated sharing and helping (Eisenberg et al., 1984), and helping and comforting behaviors (Eisenberg-Berg and Hand, 1979), were not associated with the levels of prosocial moral reasoning. In early childhood, donating behavior was negatively related to hedonistic-oriented reasoning and positively related to needs-oriented reasoning, whereas helping did not covary with the forms of prosocial reasoning (Eisenberg et al., 1985). In middle childhood, hedonistic reasoning was negatively related to donating behavior, and affectional relationship reasoning was positively related to helping in middle childhood (Eisenberg et al., 1984b). Helping behavior was not associated with higher levels of moral reasoning, but donating was negatively correlated with hedonistic-oriented reasoning and positively with higher levels of moral reasoning in adolescence (Eisenberg et al., 1991, 1995) and in late adolescence and adulthood (Eisenberg et al., 2002, 2013).

These findings suggest that the strength of the association between moral reasoning and behavior changes as a function of the motivations for engaging in prosocial behavior. In other words, not every prosocial behavior includes altruistic motivation, that is, self-sacrifice or genuine voluntary behavior. Eisenberg (2005) suggested that there are several motives for performing prosocial behaviors, including egoistic concerns (e.g., the expectation of reciprocity, award, or approval), practical concerns (e.g., preventing an unwanted situation or helping), other-oriented concerns (e.g., showing sympathy), and moral values (e.g., internalized values about the world and people). Egoistic and practical concerns elicit nonaltruistic behaviors, whereas other-oriented concerns and moral values promote altruistic behaviors (Eisenberg, 1986). This distinction between altruistic and nonaltruistic prosocial behavior is important in understanding the association between prosocial reasoning and acting.

Several findings indicated that sympathetic reactions and personal distress are differentially related to prosocial *behaviors* such that sympathy was correlated with higher levels of prosocial behaviors, while personal distress was correlated with lower levels of prosocial responding (Eisenberg and Fabes, 1990; Eisenberg et al., 1990; Miller et al., 1996). Eisenberg (2005) suggested that sympathy is associated with altruistic behaviors because it includes other-oriented concerns, which are likely to promote prosocial acts. However, personal distress is considered to be related to nonaltruistic prosocial behaviors because people might engage in helping behavior with practical concerns (i.e., dampening the stress level). Moreover, Miller et al. (1996) found that children high in both needs-oriented reasoning and sympathetic concerns (i.e., self-reported sadness to peer-distress) were more likely to exhibit helping behavior to distressed peers, indicating that the levels of prosocial moral reasoning and sympathy are important for prosocial responding.

In addition to different motives in prosocial behavior, situational factors also contribute to the magnitude of the relation between cognition and behavior. Eisenberg and Shell (1986) found that the cost of the prosocial behavior mediated the association between prosocial moral reasoning and behavior. In this study, children were asked to donate to poor children in two different donating occasions (the donation of high- and low-cost items). It was shown that needs-oriented reasoning was positively related to donating in high-cost situations, but the link between moral reasoning and donation was not significant in low-cost situations. Eisenberg (1986) argued that "if a given behavior is low in cost to the actor, it frequently may be performed without much cognitive reflection, moral or otherwise. Furthermore, if the benefit to the other is small, moral concerns related to other's welfare are unlikely to arise. Of course, if the price of helping is too high, there also may be little relation between moral reasoning and behavior because of limited variation in helping behavior (few individuals may help)" (p. 155).

Summary

Eisenberg's prosocial moral reasoning focuses on the justifications behind prosocial actions in more real-life conflicting situations. She suggests five levels of prosocial moral reasoning, ranging from hedonistic orientation to empathic or internalized orientation. Age-related changes in prosocial moral reasoning are evident such that lower levels are more pronounced in early childhood, whereas higher levels are common (often in tandem with lower levels) in adolescence. This change in prosocial moral reasoning is partly related with improving empathic, sympathetic, and perspective-taking skills. Finally, there is a positive association between prosocial moral reasoning and prosocial behavior, but motivational and situational factors are also important in this relationship.

See also: Prosocial Behavior During Adolescence; Prosocial Behavior During Childhood and Cultural Variations; Prosocial Behavior During Infancy and Early Childhood: Developmental Patterns and Cultural Variations; Prosocial Behavior, Effects of Parenting and Family Structure On.

Bibliography

Carlo, G., Eisenberg, N., Knight, G.P., 1992. An objective measure of adolescents' prosocial moral reasoning. Journal of Research on Adolescence 2, 331–349.

Carlo, G., Koller, S.H., Eisenberg, N., Da Silva, M.S., Frohlich, C.B., 1996. A cross-national study on the relations among prosocial moral reasoning, gender role orientations, and prosocial behaviors. Developmental Psychology 32, 231–240.

Carlo, G., McGinley, M., Roesch, S.C., Kaminski, J.W., 2008. Measurement invariance in a measure of prosocial moral reasoning to use with adolescents from the USA and Brazil. Journal of Moral Education 37 (4), 485–502.

Eisenberg, N., 1986. Altruistic Emotion, Cognition and Behavior. Lawrence Erlbaum, Hillsdale, NJ.

Eisenberg, N., 2005. The development of empathy-related responding. In: Carlo, G., Edwards, C.P. (Eds.), Moral Development through the Life Span: Theory, Research, and Application. University of Nebraska Press, Lincoln, NE, pp. 73–117.

Eisenberg, N., Boehnke, K., Schuhler, P., Silbereisen, R.K., 1985. The development of prosocial behavior and cognitions in German children. Journal of Cross-Cultural Psychology 16 (1), 69–82. http://dx.doi.org/10.1177/0022002185016001006.

Eisenberg, N., Carlo, G., Murphy, B., Van Court, P., 1995. Prosocial development in late adolescence: a longitudinal study. Child Development 66, 1179–1197.

Eisenberg, N., Cumberland, A., Guthrie, I.K., Murphy, B.C., Shepard, S.A., 2005. Age changes in prosocial responding and moral reasoning in adolescence and early adulthood. Journal of Research on Adolescence 15 (3), 235–260. http://dx.doi.org/10.1111/j.1532-7795.2005.00095.x.

Eisenberg, N., Fabes, R.A., 1990. Empathy: conceptualization, measurement and relation to prosocial behavior. Motivation and Emotion 14, 131–149.

Eisenberg, N., Fabes, R.A., Miller, P.A., 1990. The evolutionary and neurological roots of prosocial behavior. In: Ellis, L., Hoffman, H. (Eds.), Crime in Biological, Social, and Moral Contexts. Praeger, New York, pp. 247–260.

Eisenberg, N., Guthrie, I.K., Cumberland, A., Murphy, B.C., Shepard, S.A., Zhou, Q., Carlo, G., 2002. Prosocial development in early adulthood: a longitudinal study. Journal of Personality and Social Psychology 82 (6), 993–1006. http://dx.doi.org/10.1037/0022-3514.82.6.993.

Eisenberg, N., Guthrie, I.K., Murphy, B.C., Shepard, S.A., Cumberland, A., Carlo, G., 1999. Consistency and development of prosocial dispositions: a longitudinal study. Child Development 70, 1360–1372. http://dx.doi.org/10.1111/1467-8624.00100.

Eisenberg, N., Hofer, C., Sulik, M.J., Liew, J., 2013. The development of prosocial moral reasoning and a prosocial orientation in young adulthood: concurrent and longitudinal correlates. Developmental Psychology 50 (1), 58–70.

Eisenberg, N., Miller, P.A., Shell, R., McNalley, S., Shea, C., 1991. Prosocial development in adolescence: a longitudinal study. Developmental Psychology 27, 849–857.

Eisenberg, N., Pasternack, J.E., Cameron, E., Tryon, K., 1984a. The relation of quantity and mode of prosocial behavior to moral cognitions and social style. Child Development 55, 1479–1485.

Eisenberg, N., Pasternack, J.E., Lennon, R., March 1984b. Prosocial development in middle childhood. Paper presented at the biennial meeting of the Southwestern Society for Research in Human Development, Denver, CO.

Eisenberg, N., Shell, R., 1986. Prosocial moral judgment and behavior in children: the mediating role of cost. Personality and Social Psychology Bulletin 12 (4), 426–433. http://dx.doi.org/10.1177/0146167286124005.

Eisenberg, N., Shell, R., Pasternack, J., Lennon, R., Beller, R., Mathy, R.M., 1987. Prosocial development in middle childhood: a longitudinal study. Developmental Psychology 23 (5), 712–718. http://dx.doi.org/10.1037/0012-1649.23.5.712.

Eisenberg, N., Zhou, Q., Koller, S., 2001. Brazilian adolescents' prosocial moral judgment and behavior: relations to sympathy, perspective taking, gender-role orientation, and demographic characteristics. Child Development 72 (2), 518–534.

Eisenberg-Berg, N., 1979. Development of children's prosocial moral judgment. Developmental Psychology 15 (2), 128–137. http://dx.doi.org/10.1037/0012-1649.15.2.128.

Eisenberg-Berg, N., Hand, M., 1979. The relationship of preschoolers' reasoning about prosocial moral conflicts to prosocial behavior. Child Development 50 (2), 356–363. http://dx.doi.org/10.1111/1467-8624.ep12421678.

Eisenberg-Berg, N., Mussen, P., 1978. Empathy and moral development in adolescence. Developmental Psychology 14 (2), 185–186. http://dx.doi.org/10.1037/0012-1649.14.2.185.

Fuchs, I., Eisenberg, N., Hertz-Lazarowitz, R., Sharabany, R., 1986. Kibbutz, Israeli city, and American children's moral reasoning about prosocial moral conflicts. Merrill-Palmer Quarterly 32 (1), 37–50.

Gilligan, C., 1982. In a Different Voice: Psychological Theory and Women's Development. Harvard University Press, Cambridge, MA.

Hoffman, M.L., 2000. Empathy and Moral Development: Implications for Caring and Justice. Cambridge University Press, Cambridge.

Jaffee, S., Hyde, J.S., 2000. Gender differences in moral orientation: a meta-analysis. Psychological Bulletin 126, 703–726.

Kohlberg, L., 1969. Stage and sequence: the cognitive developmental approach to socialization. In: Goslin, D. (Ed.), Handbook of Socialization Theory and Research. Rand McNally, Chicago, IL, pp. 347–480.

Kumru, A., Carlo, G., Mestre, M.V., Samper, P., 2012. Prosocial moral reasoning and prosocial behavior among Turkish and Spanish adolescents. Social Behavior and Personality: An International Journal 40 (2), 205–214. http://dx.doi.org/10.2224/sbp.2012.40.2.205.

Miller, P.A., Eisenberg, N., Fabes, R.A., Shell, R., 1996. Relations of moral reasoning and vicarious emotion to young children's prosocial behavior toward peers and adults. Developmental Psychology 32 (2), 210–219.

Tietjen, A.M., 1986. Prosocial reasoning among children and adults in a Papua New Guinea society. Developmental Psychology 22 (6), 861–868. http://dx.doi.org/10.1037/0012-1649.22.6.861.

Triandis, H.C., 1995. Individualism and Collectivism. Westview, Boulder, CO.

Underwood, B., Moore, B., 1982. Perspective-taking and altruism. Psychological Bulletin 91, 143–173.

Eisenstadt, Shmuel N. (1923–2010)

Luis Roniger, Wake Forest University, Winston–Salem, NC, USA; and Hebrew University of Jerusalem, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Shmuel Noah Eisenstadt was a major figure of macro- and comparative-historical sociology. His work connected the comparative study of institutions, a Weberian interest in world religions, and Martin Buber's lead on human creativity, intertwining structural differentiation and institutional analysis with a cultural-interpretative approach to the dynamics of continuity and change in historical and modern societies. He connected the study of power and institutions with the search for trust and meaning. Among his contributions are studies stressing the role of agency, structure, and culture in processes of change; a critical assessment of modernization studies; institutional and civilizational analyses, among them of Japan and Jewish civilization; and the approach of multiple modernities.

Shmuel Noah Eisenstadt (1923–2010) was one of the towering figures of twentieth-century macro- and comparative-historical sociology. In his work, which includes close to 100 books and 600 academic journal articles and book chapters, Eisenstadt brought together the comparative study of institutions, a Weberian interest in world religions and civilizations, and Martin Buber's lead on human creativity, intertwining structural differentiation and institutional analysis with a cultural-interpretative approach to the dynamics of continuity and change in historical and modern societies.

Born in Warsaw, Shmuel Noah Eisenstadt migrated to Mandatory Palestine with his widowed mother when he was young. After completing his academic studies at The Hebrew University of Jerusalem, Eisenstadt served as a Faculty member of the university in Israel from 1949 until his retirement in 1990. A worldwide academic traveler, he was a visiting professor at many other universities and research institutions, including Chicago, Harvard, Zurich, Vienna, Bern, Stanford, Berlin, and Uppsala. Eisenstadt received many awards, among them the McIver Prize of the American Sociological Association (1964), the Rothschild Prize in Social Sciences (1970), the Israel Prize (1973), the International Balzan Prize (1988), the Max Planck Award for Social Sciences (1994), the Amalfi Prize for Sociology and Social Sciences (2001), the EMET Prize in Sociology (2005), and the Holberg International Memorial Prize from the Norwegian Parliament (2006). He also held honorary PhDs from Harvard and Warsaw. Eisenstadt was a member of the Israeli Academy of Sciences, a Honorary Foreign Member of the American Philosophical Society, Foreign Associate of the National Academy of Sciences in the USA, Honorary Foreign Member at the Academy of Arts and Sciences, Honorary Foreign Research Fellow at the Institute of Sociology of the Chinese Academy of Social Sciences, and Honorary Fellow of the London School of Economics and Political Science.

A student of Martin Buber in Jerusalem, his work projected Buber's approach on human creativity into a venue of comparative-historical sociology inspired by the approach of Max Weber and comparative institutional analysis, as developed by British sociology and social anthropology. Bridging social sciences and world history, Eisenstadt stressed the interplay of social structure, culture, agency, and change. He was particularly interested in tracing the multiple ways in which human societies have confronted structural challenges while organizing power, building trust, and searching for meaning. His early work dealt with the sociology of immigration and youth from the perspective of elites, primary groups, and social differentiation. He then analyzed the political system of empires and comparative studies of modernization; the study of strata formation, charisma, and institution building. Starting in the 1990s, he increasingly moved to the comparative study of civilizations, through which he stressed the interplay between the order-maintaining and transformative dimensions of culture, and completed a transition from comparative institutional analysis to comparative civilizational analysis.

His early works were influenced by structural-functionalism, sharing its drive at systematization and its emphasis on structural differentiation, particularly as he studied processes of modernization. Yet, already at that stage in his career, he combined such emphasis with an interest in human agency and change, i.e., in the active role played by human beings in shaping their own history under conditions not of their own choice, as he used to indicate following Marx's famous statement in *The Eighteen Brumaire of Louis Bonaparte*. Increasingly, Eisenstadt broke with the premises of other studies of modernization, in which he was involved in the 1960s. In his later writings on modernization, he became highly critical of theories that expected societies to converge around a homogenous model of modernity driven by urbanization, industrialization, or secularism, supposedly replicating the model led by the 'rise of the West.' Rather, he stressed how modernization unfolded differently throughout the world as the result of multiple cultures and their different ways of interpreting and shaping human existence. While exploring processes of construction of institutions, he devoted efforts to trace the open aspects of social life, the ephemeral and contingent, the charismatic and value-laden transcendent dimensions of existence – all of which led to different constellations of power, trust, and meaning. This ultimately led him to stress the open code of modernity, while rephrasing and elaborating a systematic research program around the idea of not one, but rather Multiple Modernities.

From Generation to Generation (1956, new edition 2002) and *The Political System of Empires* (1963, new edition 1992) stand out among his early works. The former is his first broad comparative work in which he analyzed age groups and youth groups in different societies, connecting their emergence to issues of transference and transformation of trust from family and kinship groups to broader social settings. In the latter book, which gained him recognition beyond the realm of sociology particularly by creating long-lasting dialogues with historians, Eisenstadt suggested that the formation and institutionalization of empires were not automatically determined by structural differentiation but was rather contingent upon certain additional conditions. Among such conditions, he identified the development of free resources, i.e., resources not embedded in ascriptive settings, and the emergence of political entrepreneurs, elites, and coalitions able to construct imperial systems out of their internal contradictions and external processes and forces, all of which generated struggle, change, and the eventual demise or transformation of such systems. This perspective led to studies of structural differentiation and modernization, in which he likewise stressed the role of protest, fragility, and change and the possible breakdown of modernization (e.g., *Essays on Comparative Institutions*, 1965; *Modernization, Protest and Change*, 1966; *Tradition, Change and Modernity*, 1973a). Likewise, he increasingly moved to recognize how traditions shape processes of modernization and are embedded in them, a line that crystallized in his analysis of post-traditional societies (Eisenstadt, 1973b) and years later became integral to his work on multiple modernities (Fisher, 2012).

In the 1970s, Eisenstadt connected this early work with the strong criticisms directed at structural-functionalism by different approaches that gathered momentum in sociology and related disciplines. All these approaches shared an unwillingness to accept any institutional setting as given due to 'systemic needs,' thus setting the stage for controversies on the salience of structure and agency and pulling interventions that crystallized around the opposing poles of 'negotiated order' versus that of 'deep structures,' and that of culture or the symbolic dimension of human interaction versus the emphasis on material or power dimensions. In *The Form of Sociology: Paradigms and Crises* (1976, with M. Curelaru), Eisenstadt engaged in a discussion of the very construction of institutions as a central problem of sociological analysis, suggesting the need to systematically address the relations between agency and structure, and between culture and social structure. He then moved to emphasizing the autonomy of culture and individual choices as part of his reappraisal 'of the processes of reconstitution of social order' and the dynamics of social change.

The perspective of social exchange, as derived from Emile Durkheim, Marcel Mauss, and others provided a means for such reappraisal. First elaborated in *Social Differentiation and Stratification* (Eisenstadt, 1971) and in unpublished articles with E.O. Schild, in which he addressed the importance of various media of exchange, including money, power, and prestige as crucial for center formation, elite status, and stratification, he fully developed this perspective in the 1980s, particularly in the comparative work on clientelism and friendship undertaken with Luis Roniger. In *Patrons, Clients and Friends* (1984), the systematic study of interpersonal relations in different historical and contemporary societies revealed various modes in which such relations connect specific forms of resource flow and access to resources within broader forms of exchange and the structuring of trust in society. The work indicated how specific and generalized exchange can be combined in diverse ways, which then affect how social order is structured and negotiated in society, and also what countervailing social forces emerge to contest the presence of interpersonal relations such as patron–client relations or various forms of friendship. He himself considered such endeavor as part of re-examining "the processes through which social formations are crystallized" and in that sense, as somehow akin to Anthony Giddens' theory of structuration, Pierre Bourdieu's emphasis on 'habitus' and reflexive praxis, and Alain Touraine's work on social movements (Eisenstadt, 1995: 20–22). Yet, in Eisenstadt's case, the concern with social exchange was led by a comparative-historical optic and the drive to move away from reductionist approaches that often saw culture as epiphenomenal to power or economic interests. His approach is singled out by a search for identifying the ground rules of social interaction, which he saw as driven by cultural assumptions and codes, themselves anchored in diverse cosmological visions of the world.

Eisenstadt joined such inquiry into social exchange with a long-term interest in three additional venues of research. First was his interest in center formation and charisma, triggered during his postdoctoral studies with Edward Shils at the London School of Economics. Such interest matured in his studies on empires, and later on was translated into the study of comparative urbanization, undertaken together with social geographer Arie Shachar (*Society, Culture and Urbanization*, 1987) and again in his introduction to the collected articles of Martin Buber on charisma that Eisenstadt put together in 1992.

Second, was his interest in the study of intellectuals and other carriers of change, exemplified in the book co-edited with R. Graubard on *Intellectuals and Tradition* (1973) and in a more political venue in his book on *Revolution and the Transformation of Societies* (1978). Third, was his work on the sociology of knowledge, leading to the co-edited publication with Ilana Friedrich Silber of *Knowledge and Society: Studies in the Sociology of Culture Past and Present* (1988), a book stressing the recognition of humans as cognitively oriented and reflexive beings seeking to construct symbolic configurations geared to understanding and making sense of the world, and underscoring the role of diverse soteriological orientations in particular in structuring the ensuing 'worlds of knowledge' (see also Silber, 2011).

These various macrosociological and comparative institutional studies facilitated the reexamination of the relations between cultural visions and institutional patterns and between agency and historical contingency. The work on revolutions, for example, indicated that structural factors may explain the breakdown of regimes but not the specific character of such breakdowns; that is, whether the change was revolutionary or not. In order to explain revolutions, Eisenstadt suggested the need to study the combined effects of civilizational and structural conditions coming together along with historical contingencies. The above macrosociological and comparative institutional analyses led him to move increasingly toward the comparative study of civilizations.

In his view, civilizations crystallized around overall visions that entail attempts to construct or reconstruct social life according to transcendental premises. Major civilizations developed in the period which Max Weber called the era of the 'great world religions' and Karl Jaspers and others defined as 'Axial civilizations' (Eisenstadt, 1982). At the Truman Research Institute in Jerusalem, Eisenstadt drafted by the mid-1980s a comprehensive sociological approach to comparative civilizations (1986) that eventually found its way into full-fledged books such as *The Origins and Diversity of Axial Civilizations* (1986), *Civilta comparate* (1990), and *Kulturen der Achsenzeit* (1991). In the 1990s, as he retired from Hebrew University and continued working at the Van Leer Jerusalem Institute, he launched a comprehensive research program on collective identities, political order, and public spheres, conducting various discussion groups and organizing national and international seminars, among them on comparative political thought, on Islam, and on Latin America (see Eisenstadt and Giesen, 1995; Barth and Enzell, 1999; Roniger, 2011). He also combined research efforts with Wolfgang Schluchter, Björn Wittrock, Jóhann Páll Arnason, and other colleagues to engage in the comparative study of axial – and later on postaxial – civilizations (e.g., Eisenstadt, 1998, 2000; Arnason et al., 2005).

During that stage in his career he became increasingly interested in Japan as a civilization. In Eisenstadt's view, fully developed in *Japanese Civilization: A Comparative View* (1996), Japan was one of the most interesting foci of comparative analysis due to its unusual combination of a very high level of structural differentiation combined with a low degree of autonomy of its main elites, which he attributed to the fact that Japan did not experience an Axial Age transformation leading to a chasm between the transcendental and the mundane order, as typical of monotheistic, Chinese, or Buddhist civilizations. In Japan, he identified a rather unusual combination of institutional similarities and very high levels of philosophical, literary, and ideological discourse and self-reflexivity, which resemble those in Western Europe, yet came together with institutional change disconnected from ideological claims couched in transcendental, universalistic terms, as carried out by autonomous elites in other civilizations. In his analysis, Eisenstadt stressed how Japan 'subdued' as it were the strong universalistic orientations inherent in Buddhism and Confucianism and did not become accordingly "a nation carrying God's universal mission" (Eisenstadt, 1995: 29; 1996). In that context, he claimed that Japanese modernity developed by premises that were a mirror image of the religious premises underlying the development of North American civilization (Eisenstadt, 2002).

Eisenstadt also engaged in a *longue durée* study of Jewish civilization, as reflected in *Jewish Historical Experience: The Civilizational Dimension* (2004). He pointed out that, as the first monotheism that called for the reconstruction of the world, the contours of Jewish civilization resulted from the concrete circumstances of its historical experience. Such experience was especially shaped by dispersion and Diaspora and thus prone to develop simultaneously through multiple political, economic, and ecological environments and through multiple encounters with other, especially but not only Axial civilizations. Under such circumstances, tensions grew in the unique configuration of a collectivity which entailed the concomitant interweaving of universalistic and particularistic orientations and multiple contesting interpretations and confrontations, led by dynamic carriers of identity and change. In Eisenstadt's analysis, the dispersion and lack of centralized authority enabled the development of multiple arenas for the rise of relatively autonomous, even quite anarchic, 'carriers of meaning' among Jews. This also resulted in multiple definitions of Jewish collective consciousness; in the emergence of heterodoxies and sectarianism; in a high reflexivity of Jews as a marginalized minority for two millennia; and in ambivalent modes of participation of the Jews in broader civilizational frameworks, primarily in its interactions with the Christian and Islamic worlds. This, in turn, opened the way for radical changes and innovation once conditions allowed it, but also posed anew the problem of Jewish survival and continuity in modern times. In claiming this civilizational perspective rooted in Jewish historical experience and stressing that Jews were a creative part of an intense transcivilizational interaction and interreligious competition for centuries, Eisenstadt could challenge well-established alternative interpretations such as Weber's characterization of a 'pariah people' or Arnold Toynbee's view that once dispersed, they were transformed into a 'fossilized' civilization. This also allowed him to explain – against the claims of Christian theology and some strands of modern historiography – the continuity of Jewish history in spite of territorial and political discontinuity, exile, and dispersion (see Eisenstadt, 1997a; 2009).

Eisenstadt also contributed multiple articles and chapters on Confucianism, Islam, India, and other civilizations, although not in the same detail as his analyses of Japanese, Jewish, and European civilizations (on the latter see Eisenstadt, 1987). Rather, he became the comparative voice that served as a trigger of reflexivity for a wide array of historians, social scientists, and area specialists, leading to the publication of a series of books in which sociologists, historians, political scientists, and students of culture found common ground for explorations of the encounters between world history and comparative studies. Illustrative works are *The Early State in African Perspective* (1988), co-edited with Michel Abitbol and Naomi Chazan or his contributions to collective volumes such as *Globality and Multiple Modernities: Comparative North and Latin American Perspectives* (2002, edited by Luis Roniger and Carlos H Waisman, and see also Eisenstadt, 2013) in which he suggested directions of research on the various civilizations of the Americas in terms of comparative analysis and multiple modernities.

Living in Jerusalem, throughout his life Eisenstadt also contributed studies on major aspects in the development of Israeli society. From his early studies of immigrants arriving from both Western and Middle Eastern and Asian countries to his later comprehensive analyses of Israeli society and of its transformations (1967, 1985), Eisenstadt retained a deep interest in shifts occurring in Israel. Being the leading Israeli sociologist for several decades, later generations of Israeli sociologists disputed some of the premises of his approach, accusing him of projecting a structural-functionalist approach that supposedly downplayed or displaced conflicting aspects beyond its borders (see Kimmerling's (1992) assessment). Those addressing his sociology of Israeli society in critical terms usually failed to

relate to his later writings in general, and on civilizations and multiple modernities in particular. Still, even those critics recognized his contributions to the development of sociology as an established discipline connected to comparative perspectives, with macroinstitutional concerns and addressing key questions of development. An extensive interview with anthropologist Danny Rabinovich in a festschrift in his honor provides a glimpse of Eisenstadt's ability to both understand the changes underwent by Israeli society and continue contributing a critical eye on its development as well as on the challenges that the social sciences face in a small country at the semiperiphery of the world system (Herzog et al., 2007; Weil, 2010).

As indicated, in the 1990s Eisenstadt returned with new vigor to the study of modernity, albeit with a very critical reappraisal of the classical visions of modernization studies, stressing instead the multiple, open-ended, and uncertain character of modernity, as reflected in the dynamics of change, of shifting boundaries of political membership and shifting patterns of citizenship and participation. He engaged in institutional analyses of modernity and democracy (e.g., *Paradoxes of Democracy: Fragility, Continuity and Change*, 1999a). He equally devoted energies to analyze how the tensions between orthodoxies and heterodoxies were transformed in modern times, stressing that fundamentalisms are basically modern reformulations of patterns of contestation rooted in the various soteriological premises of civilizations while challenging alternative interpretations of modernity (see e.g., his book on *Fundamentalism, Sectarianism and Revolution*, 1999b). To his last works, he remained fully aware not only of the constructive dimensions of modernity, but he also emphasized the potential destructive tendencies of "barbarism lurking within the very core of modernity" (Eisenstadt, 2013: 52).

Eisenstadt's approach on Multiple Modernities was his last major contribution to comparative social research. His view of multiple modernities entailed a basic tension between seeing the expansion of modernity as "the crystallization of a new type of civilization not unlike the expansion of the great religions or the great imperial expansions in past times," while at the same time emphasizing that even as it spread to most of the world, "modernity did not give rise to a single civilization or to one institutional pattern, but to the development of several modern civilizations, or at least civilizational patterns," i.e., of civilizations which share common characteristics, but which tend to develop different, even cognate ideological and institutional dynamics" (1998: 38–40). He did not try to downplay this tension, but rather saw it as part of the basic tensions and antinomies that have developed within the program of modernity. Among these he singled out the tensions and antinomies between totalistic and more diversified or pluralistic conceptions about many of the components of that program of modernity; namely, the conception of reason and its place in human life and society and how the construction of nature, human society, and history was conceived; the place of reflexivity and active construction of society and nature; different evaluations of major dimensions of human experience; and tensions between control and autonomy or freedom (Eisenstadt, 2000, 2007). Specifically addressing the political arena, he identified tensions between constructivist approaches that view politics as the active construction of society against those accepting society as given; between the autonomy of civil society and state power; between the civil adherence to rules and the utopian tenets of modernity; and between normal and revolutionary politics, entailing Jacobin trends (1997b, 1999b). Eisenstadt stressed that it was around these tensions that the critical discourse of modernity developed in a multifaceted and ever-changing confrontation between the claims to enhance freedom and the strong tendency to control, putting modernity in general and particularly modern democratic regimes 'on endless trial,' as he used to indicate following Leszek Kolakowski's term. All this led to continuously developing new questions, potential crises, and manifold reinterpretations of modernity interwoven with different cultural agendas and different modes of connecting instrumental rationality and value rationality; in short, to the rise of multiple modernities sustaining far-reaching reformulations of the initial cultural program of modernity (see Preyer, 2012). Eisenstadt claimed that many fundamentalist and postmodern movements, even when attempting to dissociate themselves from Westernization and in some cases rejecting the Western cultural program of modernity, yet claim modernity for themselves by reinterpreting their traditions and civilizations in an ambivalent rephrasing of modernity's premises.

In terms of sociological theory, this perspective is connected to Eisenstadt's persistent inquiry into order-maintaining and order-transforming dimensions of culture, aspects he saw as closely related to the extent of mutual determination of culture, social structure, and human behavior. In his view, these two aspects are but 'two sides of the same coin.' Distancing himself from reductionist, idealist, or materialistic interpretations of social action and social structure that give primacy to studies of culture or studies of social structure, leading to an artificial dissociation between both dimensions, Eisenstadt called to reunite studies of culture and social structure. He stressed as well that culture is neither static and homogenous nor entirely open and malleable. Focusing on dissent, protest, and liminal situations, he stressed how societies become aware of the arbitrariness of any social order, and social research becomes conscious that institutions are symbolically constructed and can be affected at different historical moments, in tribal societies, axial civilizations, and in modernity. In his own words,

> No social order or pattern of social interaction can be constituted without the symbolic dimension of human activity, especially of some basic cultural and ontological visions. It is such visions that constitute the starting point for articulating the premises and institutional contours of any patterns of social interaction and especially of institutional and macrosocietal formations. Such a construction of social order, of the interweaving of social and cultural dimensions, is effected by processes of intensive social interaction and conflict among different elites and influentials and among them and broader groups and strata. Such processes are effected by coalitions and counter-coalitions of different social actors … [who promulgate] different ontological visions [and] activate various processes of control …. [T]hese very processes give rise to orientations and movements of protest, to the structuring of liminal situations, some of which may become harbingers of change. Such developments are not accidental or external to the realm of culture. They are given in the basic fact of the inherent interweaving of culture and of social structure as two constitutive elements of the construction of social order.

> … Such seeds of transformation are indeed common to all societies. Yet the way in which they grow – the concrete constellations of different liminal situations, of different orientations and movements of protest and their impact on societies within which they develop – greatly vary among different societies, giving rise to different patterns of social and cultural dynamics.
> (Eisenstadt 1995: 323–324)

The wide impact of Eisenstadt's work in general and his perspective on Multiple Modernities in particular is likely to persist in future generations. In 'S.N. Eisenstadt: A sociological giant,' Roland Robertson has called attention to traits that transformed him into a galvanizing figure of many research programs in sociology and beyond: his amazing capacity for intellectual and emotional empathy that led to an extensive range of connections all over the world and a profound acquaintance "with virtually every region of the contemporary, early modern, primal and ancient worlds" (Robertson, 2011: 304). Edward Tiryakian has defined him recently as "the major comparative sociologist of our times" (2011: 248). Willfried Spohn considered Eisenstadt to be a "major visionary for the sociological challenges of the twenty-first century," suggesting that his work will have a lasting impact on the sociology of civilizations in the context of world history as well as in the multicivilizational conception of global modernity, since he brought to fruition a multidimensional and intercivilizational methodology that stresses the internal dynamic forces of civilizations and their specific encounters, transfers, and interrelations – a challenge that is increasingly relevant for the future of humankind (Spohn, 2011). Likewise, Donald Levine has identified Eisenstadt's legacy as championing awareness to the complexities of different civilizational traditions and sub-traditions and the potentialities of intercivilizational dialogue (Levine, 2011). Central to Eisenstadt's research program since his early cooperation with Stein Rokkan (1973) has been the identification of the inner dynamism, multiplicity of contesting centers, and increasing autonomy of European institutions (Eisenstadt, 1987), leading to its worldwide projection and generation of multiple processes of contestation and new civilizational encounters and amalgams. Thus, it is almost natural that Eisenstadt's influence has been prominent in the milieu of European social sciences – and to some extent in developing countries – rather than in the USA, as reflected, for example, in an extensive presentation of his research in *Protosociology*, volume 24 ('Multiple Modernities – A Paradigm of Cultural and Social Evolution, 2007,' and also in *Protosociology*, 2009) and in the works by Plake and Schultz (1993) and Preyer (2011). In parallel, Eisenstadt's work has continued to be translated into many languages, including recent publications in Greek and Spanish, which add to previous publications in various languages, including Arabic, Chinese, Japanese, and Portuguese.

The books published in his honor are testimony to the wide range of topics and ideas that his work has generated; see in this connection the festschrifts edited by Cohen et al. (1985), Plake and Schultz (1993), Ben-Rafael and Sternberg (2005), Herzog et al. (2007) and Gazit and Yair (2010). Forthcoming books edited by Michael Sussman and Gerhard Preyer (forthcoming) and by Benjamin Keidar et al. (forthcoming) are likely to project the impact of Eisenstadt's research program on the interface of comparative sociology and world history into new fields of inquiry, including transnational politics, human rights, or the role of contingency in human history.

See also: Civilizations; Comparative Method in Anthropology; Differentiation: Social; Durkheim, Emile (1858–1917); Europe: Sociocultural Overview; Ideologies, Institutions, and the New Institutionalism; Japan: Sociocultural Aspects; Modernization, Sociological Theories of; Nations and Nation-States in History; Patron–Client Relations, Social and Anthropological Study of; Religion and Stratification; Revolutions, Sociology of; Social Stratification; Weber, Max (1864–1920); Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Arnasson, Jóhann Páll, Eisenstadt, Shmuel N., Wittrock, Björen (Eds.), 2005. Axial Civilizations and World History. Brill, Leiden and Boston.

Barth, Theodor, Enzell, Magnus (Eds.), 1999. Collective Identity and Citizenship in Europe. Fields of Access and Exclusion. ARENA and The Van Leer Institute, Oslo and Jerusalem. Report No. 3.

Ben-Rafael, Eliezer, Sternberg, Yitzhak (Eds.), 2005. Comparing Modernities: Pluralism versus Homogeneity. Essays in Homage of Shmuel N. Eisenstadt. Brill, Leiden.

Cohen, Erik, Lissak, Modshe, Almagor, Uri (Eds.), 1985. Comparative Social Dynamics: Essays in Honor of S.N. Eisentadt. Westview, Boulder and London.

Eisenstadt, Shmuel N., 1956 (2002). From Generation to Generation. Free Press, New York.

Eisenstadt, Shmuel N., 1963 (1992). The Political System of Empires. Free Press, New York.

Eisenstadt, Shmuel N., 1965. Essays on Comparative Institutions. Wiley, New York.

Eisenstadt, Shmuel N., 1966. Modernization, Protest and Change. Prentice Hall, Englewood Cliffs, NJ.

Eisenstadt, Shmuel N., 1967. Israeli Society. Weidenfeld & Nicolson, London.

Eisenstadt, Shmuel N., 1971. Social Differentiation and Stratification. Scott, Foresman, Glenview, Ill.

Eisenstadt, Shmuel N., 1973a. Tradition, Change and Modernity. Wiley, New York.

Eisenstadt, Shmuel N., 1973b. Post traditional societies and the continuity and the reconstruction of tradition. Daedalus 102 (1), 1–27 (in the special issue on Post-Traditional Societies that he edited).

Eisenstadt, Shmuel N., 1978. Revolution and the Transformation of Societies. Free Press, New York.

Eisenstadt, Shmuel N., 1982. The axial age: the emergence of transcendental visions and the rise of clerics. European Journal of Sociology 23 (2), 294–314.

Eisenstadt, Shmuel N., 1985. The Transformation of Israeli Society: An Essay in Interpretation. Weidenfeld & Nicolson, London.

Eisenstadt, Shmuel N., 1986. Origins and Diversity of Axial Civilizations. State University of New York Press, Albany.

Eisenstadt, Shmuel N., 1987. European Civilization in Comparative Perspective. Oslo University Press, Oslo.

Eisenstadt, Shmuel N., 1990. Civiltà Comparate. Le Radici Storiche Della Modernizzazione. Liguori, Napoli.

Eisenstadt, Shmuel N. (Ed.), 1991. Kulturen der Achsenzeit, vols 3. Suhrkamp, Frankfurt/Main.

Eisenstadt, Shmuel N., 1995. Power, Trust and Meaning. Essays in Sociological Theory and Analysis. University of Chicago Press, Chicago.

Eisenstadt, Shmuel N., 1996. Japanese Civilization: A Comparative View. University of Chicago Press, Chicago.

Eisenstadt, Shmuel N., 1997a. Sectarianism and heterodoxy in Jewish history: some comparative civilizational notes. Jewish Studies 37, 7–59.

Eisenstadt, Shmuel N., 1997b. Les antinomies de la modernité : Les composants jacobines de la modernité et du fundamentalisme. L'Arche, Paris.

Eisenstadt, Shmuel N., 1998. Multiple modernities in an age of globalization. In: Grenzenlose Gesellschaft? Vehandlungen des 29. Kongresses der Deutschen Gesellschaft fur Sociologie, pp. 37–50.

Eisenstadt, Shmuel N., 1999a. Paradoxes of Democracy: Fragility, Continuity and Change. Johns Hopkins University Press, Baltimore.

Eisenstadt, Shmuel N., 1999b. Fundamentalism, Sectarianism and Revolution. The Jacobin Dimension of Modernity. Cambridge University Press, Cambridge.
Eisenstadt, Shmuel N., 2000. Multiple modernities. Daedalus 129 (1), 1–29.
Eisenstadt, Shmuel N., 2002. Mirror-image modernities: contrasting religious premises of Japanese and U.S. Modernity. In: Madsen, Richard, Sullivan, William M., Swidler, Ann, Tipton, Steven M. (Eds.), Meaning and Modernity: Religion, Polity and Self. University of California Press, Berkeley, pp. 56–77.
Eisenstadt, Shmuel N., 2004. Explorations in Jewish Historical Experience: The Civilizational Dimension. Brill, Leiden and Boston.
Eisenstadt, Shmuel N., 2007. Multiple modernities – a paradigm of cultural and social evolution (Selected articles). ProtoSociology 24, 20–308.
Eisenstadt, Shmuel N., 2009. Tensions as resources in Jewish historical civilizational experience. European Journal of Sociology 50 (2), 233–271.
Eisenstadt, Shmuel N., 2013. Latin America and the problem of multiple modernities. In: Sznajder, Mario, Roniger, Luis, Forment, Carlos A. (Eds.), Shifting Frontiers of Citizenship: The Latin American Experience. Brill, Leiden and Boston, pp. 43–54.
Eisenstadt, Shmuel N., Abitbol, Michel, Chazan, Naomi (Eds.), 1988. The Early State in African Perspective. Brill, Leiden.
Eisenstadt, Shmuel N., Curelaru, Miriam, 1976. The Form of Sociology: Paradigms and Crises. John Wiley and Sons, New York.
Eisenstadt, Shmuel N., Giesen, Bernhard, 1995. The construction of collective identity. European Journal of Sociology 36, 72–102.
Eisenstadt, Shmuel N., Graubard, Stephen R. (Eds.), 1973. Intellectuals and Tradition. Humanities Press, New York.
Eisenstadt, Shmuel N., Rokkan, Stein, 1973. Building States and Nations. Sage, Beverly Hills, Calif.
Eisenstadt, Shmuel N., Roniger, Luis, 1984. Patrons, Clients and Friends. Interpersonal Relations and the Structure of Trust in Society. Cambridge University Press, Cambridge.
Eisenstadt, Shmuel N., Schachar, Arie, 1987. Society, Culture and Urbanization. Sage, Newbury Park, Calif.
Eisenstadt, Shmuel N., Silber, Ilana F. (Eds.), 1988. Knowledge and Society: Studies in the Sociology of Culture Past and Present. JAI Press, Greenwich, CT.
Fisher, Shlomo, 2012. Multiple modernities and Shmuel Noah Eisenstadt's sociological vision [in Hebrew] Soziologia Israelit 14 (1), 185–193.
Gazit, Orit, Yair, Gad (Eds.), 2010. Collective Identities, States and Globalization. Essays in Honor of S.N. Eisenstadt. Magnes Press, Jerusalem.
Herzog, Hannah, Zelnicker, Shimshon, Kochavi, Tal (Eds.), 2007. Generations, Spaces, Identities: Perspectives on the Configuration of Society and Culture in Israel. Van Leer Institute and Hakibutz Hameuhad, Jerusalem [in Hebrew].
Kedar, B.Z., Ilana Silber, Adam Klin-Oron (Eds.), 2015. Dynamics of Continuity, Patterns of Change: Between World History and Comparative Historical Sociology. Israel Academy of Sciences and Humanities and the Van Leer Jerusalem Institute, Jerusalem.
Kimmerling, Baruch, 1992. Sociology, ideology, and nation-building: the Palestinians and their meaning in Israeli sociology. American Sociological Review 57, 446–460.
Levine, Donald N., 2011. The dialogue of civilizations: an Eisenstadt legacy. Journal of Classical Sociology 11 (3), 313–326.
Plake, Klaus, Schultz, Wolfgang K. (Eds.), 1993. Entillusionierung als Programm: Beitrage zur Soziologie von Shmuel N. Eisenstadt. Deutscher Studien Verlag, Weinheim.
Preyer, Gerhard, 2011. Zur Aktualität von Shmuel N. Eisenstadt. VS Verlag Sozialwissenschaft, Wiesbaden.
Preyer, Gerhard, 2012. The perspective of multiple modernities. On Shmuel N. Eisenstadt's sociology. Theory and Society 4, 187–225.
Protosociology, 2009. Multiple modernization (various articles). Special Issue on Modernization in Times of Globalization 26.
Robertson, Roland, 2011. S.N. Eisenstadt: a sociological giant. Journal of Classical Sociology 11 (3), 303–311.
Roniger, Luis, 2011. Connected histories, power and meaning: transnational forces in the construction of collective identities. Journal of Classical Sociology 11 (3), 251–268.
Roniger, Luis, Waisman, Carlos H. (Eds.), 2002. Globality and Multiple Modernities. Comparative North and Latin American Perspectives. Sussex Academic Press, Brighton.
Silber, Ilana F., 2011. Deciphering transcendence and the open code of modernity: S.N. Eisenstadt's comparative hermeneutics of civilizations. Journal of Classical Sociology 11 (3), 269–280.
Spohn, Willfried, 2011. An appraisal of Shmuel Noah Eisenstadt's global historical sociology. Journal of Classical Sociology 11 (3), 282–302.
Sussman Michael, Gerhard Preyer (Eds.), 2015. Multiple Modernities in the Contemporary Scene. A Continuation of the Multiple Modernities Project.
Tiryakian, Edward A., 2011. A sociological odyssey: the comparative voyage of S.N. Eisenstadt. Journal of Classical Sociology 11 (3), 241–250.
Turner, Bryan S., Susen, Simon, 2011. Special issue: Shmuel Noah Eisenstadt. Journal of Classical Sociology 11 (3), 229–335.
Weil, Shalva, 2010. On Multiple Modernities, Civilizations and Ancient Judaism. An Interview with Shmuel N. Eisenstadt. http://www.academia.edu/348048/Interview_with_Prof._S._N._Eisenstadt.

Elder Abuse

Jacki Pritchard, Jacki Pritchard Ltd, Sheffield, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Elder abuse, while not a new phenomenon, has gradually gained more recognition as a social problem, with terminology and definitions changing during the past 30 years. Research has tended to be undertaken in more developed countries, but cultural, social, and economic differences between societies affect how researchers, organizations, and citizens view abuse. Consequently, the findings of incidence and prevalence studies differ. Elder abuse can occur for many different reasons; it can be the result of carers' stress but there is now more acknowledgment that perpetrators of abuse can deliberately target and groom their victims.

Introduction

Elder abuse is not a new phenomenon, having existed for many centuries. Awareness about it increased during the twentieth century after two British geriatricians wrote about their patients (Baker, 1975; Burston, 1975, 1977). However, elder abuse has never had the same level of attention or interest as child, family, or domestic abuse. For many years, it remained a taboo subject for a number of reasons: the ageism that exists in some societies, older people not being seen as emotive as children, one's fear of growing old, and acknowledging that abuse could happen to oneself.

During the past 30 years, terminology has changed, with various terms being connected to the current term 'elder abuse':

- Granny bashing/battering.
- Old age abuse.
- Adult abuse (this term was introduced in the 1990s and is used for all vulnerable adults aged over 18 years, so that elder abuse would be incorporated into this term).
- Safeguarding adults (introduced in the United Kingdom in 2005 (ADSS, 2005) and includes older people).

Professionals are confronted with many practice dilemmas when working with elder abuse. It is vital that organizations in the statutory, voluntary, and independent sectors define elder abuse in their policies and also gives detailed guidance in procedures so that practitioners are clear about how and when to intervene. Training is also essential to develop safe and best practices when investigating, empowering, and supporting victims of abuse.

Definitions

The term 'elder abuse' can be interpreted to mean harm, mistreatment, or neglect and may be carried out in many different ways including definitions and then specific categories of abuse. Researchers and policymakers can define elder abuse differently, but since the beginning of the twenty-first century there has been more usage of a definition that originally was developed by the UK Charity Action on Elder Abuse but then adopted by the World Health Organization (WHO) and other important organizations worldwide that focus on the prevention of elder abuse (AEA Bulletin, 1995; WHO, 2002, 2011):

> Elder abuse is a single or repeated act, or lack of appropriate action, occurring within any relationship where there is an expectation of trust which causes harm or distress to an older person. (http://www.who.int/ageing/projects/elder_abuse/en/ – accessed 16.06.13.)

This definition like many others focuses on the older person being abused by someone in a position of trust. It does not consider that an older person can be targeted, befriended, and groomed for financial or sexual exploitation by strangers, which will be discussed below. The root of this stems from studies conducted from the 1980s onward that indicated that older persons were likely to be abused by their carers and consequently carers' stress dominated that literature (O'Malley et al., 1983; Steinmetz, 1978; Eastman, 1984; Pritchard, 1990).

Another common theme in definitions of elder abuse regards the violation of human rights

> …a violation of an individual's human and civil rights by any other person or persons. (Department of Health, 2000: p. 9)

> Elder abuse is a violation of human rights and a significant cause of injury, illness, lost productivity, isolation and despair. Confronting and reducing elder abuse requires a multisectoral and multidisciplinary approach. (Australian Network for the Prevention of Elder Abuse: www.anpea.com.au – accessed 16.06.13.)

> *Violation of Human Rights* is the denial of fundamental rights e.g. dignity, freedom of speech, religion, thought, expression, or cruel and degrading treatment. (Action on Elder Abuse South Africa: www.elderabuse.co.za – accessed 16.06.13.)

Around the world, definitions of elder abuse will differ because of cultural differences; behaviors acceptable in one society may be considered abusive in another. How older people are regarded within society will also affect how people identify and work with elder abuse. There may be denial that

such a social problem exists, especially in societies that have great respect for their older citizens (Shankardass, 2009). Research has been limited in Asia, yet some countries are beginning to highlight the issues. For example, HelpAge India undertook a survey in 24 cities across 20 states in 2013 and reported that 23% of older people had experienced abuse (HelpAge India, 2013).

(http://www.hindustantimes.com/India-news/NewDelhi/One-fifth-of-India-s-elderly-abused-at-home/Article1-1076551.aspx – accessed 18.06.13.).

In Africa, elder abuse is gradually being recognized as a social problem (Mba, 2007):

> Despite their acknowledged social and economic contributions, as well as their role in traditional and cultural affairs, many older adults in Africa experience abuse and are largely excluded in socio-economic and political affairs. The processes of modernization and urbanization are beginning to erode the traditional social welfare system of Africa, the extended family. The abuse and violence against older persons and their vulnerability to financial exploitation is a key challenge faced by many African countries. Older persons, particularly women, have to care for their dying children and orphaned grandchildren. They suffer from financial hardship and social isolation thereby endangering their health (p. 230).

Regarding South Africa, Ferreira and Lindgren (2008) discuss:

> The problem of elder abuse and neglect in South Africa is widespread but the definition of abuse remains problematic. An expanded typology is needed to classify certain types of abuse reported commonly such as marginalization, disrespect, exploitation and violence (p. 91).

Research into elder abuse in some countries of the world is still very limited, whereas in more developed countries, it can be difficult to identify elderly victims within black minority ethnic groups for research purposes because they do not ordinarily access statutory services (Bowes, Avan and Macintosh 2008; Patel 2003) but in Europe may access black and minority ethnic services (Patel, 2003).

Definitions can change and develop over time as knowledge increases. For example, at one time it was thought that elder abuse had to be ongoing:

> The systematic maltreatment, physical, emotional, or financial, of an elderly person … this may take the form of physical assault, threatening behavior, neglect and abandonment or sexual assault. (Eastman, 1984: p. 3)

Later it was recognized that it could be a one-off incident:

> Abuse … causes harm to the older person, either temporarily or over a period of time. (Department of Health, 1993: p. 3)
>
> Abuse may consist of a single act or repeated acts. (Department of Health, 2000: p. 9)

It is important to have a clear definition of elder abuse but there are additional definitions (institutional abuse, stranger abuse, and domestic violence) that help further understanding of how abuse might be inflicted. Institutional abuse is defined as follows:

> Neglect and poor professional practice also need to be taken into account. This may take the form of isolated incidents of poor or unsatisfactory professional practice, at one end of the spectrum, through to pervasive ill treatment or gross misconduct at the other. Repeated instances of poor care may be an indication of more serious problems and this is sometimes referred to as **institutional abuse**. (DH, 2000: p. 10)
>
> …where the environment, practices and rules become abusive in themselves. (Decalmer, 1993: p. 59)

In this context, consideration is given to abuse in actual institutions rather than in the community, e.g.,

- Residential care home
- Day care facility
- Hospital
- Prison

Like abuse in the community, there can be perpetrators from different backgrounds. Carer's stress has been discussed at length again in regard to this type of abuse as Pillemer and Moore (1989, 1990) highlighted during the late 1980s. Care staff can become abusive when overwhelmed by having to care for residents/patients with high dependency needs (Pritchard, 1996, 2007). Other possible causal factors are

- Lack of education/training for staff.
- Staff groups/factions developing (staff groups becoming competitive).
- Ineffective management.
- Existing rules and regimes, resulting in restriction of choice and rights, e.g., having to get up at a certain time, being made to go to the toilet before meals, and not being allowed to do something.
- Poor environment/conditions (damaged furniture, old decoration, few windows, and poor ventilation).
- Low staff morale.
- Low staffing levels.
- Staff characteristics (little patience, not caring, not suited to the job, or doing it because they need employment rather than wanting or choosing to work in the care sector).
- Characteristics of residents (high dependency needs).

A key question regularly put forward is 'When does bad practice cross that fine line and become abuse?' The debate will continue just as many scandals will be highlighted in the media. But there may also be staff members who have deliberately chosen to work in the care sector to gain access to vulnerable older people, whom they can abuse. On one level, money and/or possessions can be stolen or a person can be persuaded to sign documents that transfer assets; on another level, sexual abuse can occur.

Many people visit institutions and can have the opportunity to perpetrate abuse including

- family members,
- friends,

- professionals,
- volunteers, and
- advocates.

Also, many older people in prisons can be abused by fellow inmates or prison officers. Additionally, their physical needs can be neglected (Hayes and Fazel, 2008).

Stranger Abuse – The Debate

Much has been written about the fact that older people are abused by people who they know. However, they can also be targeted and groomed by complete strangers (Pritchard, 1993). Some organizations see this as a criminal activity that is outside the definition of elder abuse. However, others see parallels with child abuse and believe that it should also be considered.

Older people can become very lonely and isolated when family and friends die, so much so that it can make them vulnerable and susceptible to exploitation. The older person may be willing to hand over money to someone to have them visit the house. Children of a very young age can be involved in this type of abuse as can groups of teenagers. Sometimes an adult will be organizing gangs of children to engage in this activity (Pritchard, 1993).

In the United Kingdom, the Department of Health introduced the definition of stranger abuse into the *No Secrets* guidance:

> Stranger abuse will warrant a different kind of response from that appropriate to abuse in an ongoing relationship or in a care location. Nevertheless, in some instances it may be appropriate to use the locally agreed inter-agency adult protection procedures to ensure that the vulnerable person receives the services and support that they need. Such procedures may also be used when there is the potential for harm to other vulnerable people. (Department of Health, 2000: p. 11)

Domestic Abuse – To Include or Not to Include in Elder Abuse

Many types of abuse are interlinked; within families abuse can happen across the generations. In the United States and Canada, a great deal of attention has been given to the concept of family violence by researchers (Straus et al., 1980; Gelles and Cornell, 1985; Gelles, 1997). Much debate over the years concerns whether domestic abuse should be considered a form of elder abuse and how it should be addressed (Ingram, 2001). It was recognized that older women could have been living in domestic violence situations for most of their married lives. More recently, domestic abuse has been acknowledged as a form of elder abuse. Research is now being undertaken regarding intimate partner violence against older women (Nagele et al., 2010). Attention has been given to the role reversal, i.e., where the male abuser becomes the victim, the typical scenario being when the man becomes dependent on the woman later in life and the woman gets her own back because she knows he can no longer physically attack her (Friday et al., 2007).

Men can also be victims of domestic violence throughout a marriage. Although not a new concept, this remains a taboo subject. Steinmetz (1978) highlighted the problem back in 1978 and as a result had her life threatened, while the controversy continued for the next two decades (Straus, 1999). When undertaking research into the needs of older women who had been abused, Pritchard (2000, 2001) was approached by male victims who wanted to disclose the abuse they had experienced. In 2007, there was a special issue of the *Journal of Abuse and Neglect* devoted to the abuse of older men (Kosberg, 2007).

In domestic situations, mutual abuse can occur where both the older man and older woman are abusive toward each other.

> Domestic Abuse is any incident of threatening behavior, violence or abuse (psychological, physical, sexual, financial or emotional) between adults, aged 18 and over, who are or have been intimate partners or family members regardless of gender and sexuality.
>
> Family members include mother, father, son, daughter, brother, sister, grandparents, whether directly related, in-laws and stepfamily.
>
> The definition includes 'same sex' partners and ex-partners (irrespective of how long they have been apart). (Association of Chief Officers of Police and National Policing Improvement Agency, 2008: p. 7)

How Victims Define Abuse

It is crucial that when developing definitions of any type of abuse, the views and opinions of victims are taken into consideration. Eight countries – Argentina, Austria, Brazil, Canada, India, Kenya, Lebanon, and Sweden – participated in the study *Missing Voices* (WHO, 2002) to address this issue. Their analysis of the major themes revealed remarkable similarities across participating countries. Older people perceived abuse under three broad areas:

- Neglect – isolation, abandonment, and social exclusion.
- Violation – of human, legal, and medical rights.
- Deprivation – of choices, decisions, status, finances, and respect.

(WHO, 2002: p. v)

Categories of Abuse

Some definitions make a statement about elder abuse that includes what some may consider categories or forms of abuse. Thus, there can be some disparity depending on how people view what actions or behaviors actually constitute elder abuse and their importance.

It is helpful for professionals and workers to have clear guidance regarding categories of abuse to work from; however, there is some overlap between the individual categories. When investigating cases of abuse, it is useful to consider on the balance of probabilities which categories of abuse have been substantiated. Policies and procedures around the world may differ in format but there is some general consensus about what constitutes abuse, i.e., what actions and behaviors may cause an older person to be physically or emotionally harmed

(DH, 2000; Payne, 2013). In England, for many years, elder abuse was considered to have just three categories:

> The systematic maltreatment, physical, emotional or financial, of an elderly person … this may take the form of physical assault, threatening behavior, neglect, abandonment or sexual assault. (Eastman, l984: p. 3)

In the early 1990s, after the Department of Health had undertaken research into elder abuse (DH, 1992), neglect and sexual were added as separate categories:

> Abuse may be described as physical, sexual, psychological or financial. It may be intentional or unintentional, or the result of neglect. (Department of Health, l993: p. 3)

Then, in 2000, a further category was added – discriminatory abuse:

> …including racist, sexist, that based on a person's disability, and other forms of harassment, slurs or similar treatment. (Department of Health, 2000: p. 9)

Around the world, it would seem that other forms of abuse are often highlighted within definitions or as stand-alone categories of abuse. For example, the WHO states:

> This type of violence constitutes a violation of human rights and includes physical, sexual, psychological, emotional; financial and material abuse; abandonment; neglect; and serious loss of dignity and respect. (WHO *Fact Sheet*, August 2011 – http://www.who.int/mediacentre/factsheets/fs357/en/ – accessed 16.06.13.)

Sometimes particular abuses are highlighted as separate categories because organizations think they are very important or particularly significant in the country concerned. Examples include

Social: such as restricting the older person's social freedom and isolating him or her from family and friends (Australia).

Witchcraft: occurs mainly in the more rural provinces of South Africa. Women, especially black women, living alone who are particularly dark skinned and wrinkled, are accused of being witches. This is usually prompted by disasters in the area, e.g., storm or lightning damage and crop failures. The victims are in danger of suffering physical harm and/or having their homes (huts) burnt down.

Systemic abuse: this is denying older persons the right to be treated with respect and dignity when accessing services, both those belonging to the state or the private sector (South Africa).

General Overview of the Categories of Elder Abuse

Physical Abuse

- Being physically attacked/hurt/assaulted with hand or implement: slapping, hitting, punching, pushing, kicking, shaking, pinching, hair pulling, spitting and burning
- Force feeding
- Restraint
- Rough handling
- Imprisonment: keeping in confined space
- Medication: overmedicating, withdrawal, hiding medication, not giving correct dosage, being given another person's medication, and not reviewing medication regularly

Emotional/Psychological Abuse

- Verbal: shouting, swearing, name-calling, misuse of names, use of nicknames, and derogatory comments
- Threats/coercion: causing fear/anguish, unjust emotional pressure
- Humiliation
- Ridicule
- Degradation
- Manipulation
- Harassment
- Bullying
- Lack of acknowledgment/ignoring
- Exclusion
- Abandonment
- Overprotection: not being allowed to do things
- Lack of social interaction/contact
- Denial of human rights

Financial/Material Abuse

- Theft
- Fraud
- Mismanagement of money/pension/benefits/assets
- Denying access to money/assets
- Taking/spending money without consent (e.g., out of the automated teller machine)
- Shopping: saying things cost more than they did
- Selling without consent: objects, assets
- Using own reward points cards for personal financial gain
- Not paying bills
- Falsifying documents
- Forgery/forcing/obtaining signatures
- Forcing transfer of money, accounts, assets
- Forcing changes to bank accounts, wills
- Blackmail
- Charging for services that should be free or are included in fees
- Abusing position of trust, especially as attorney or appointee
- Obtaining a credit card in the older person's name and using it for oneself
- Writing of wills

Neglect

- Failure to meet physical and/or emotional needs
- Act of omission: failure to do something
- Deprivation/lack of basic care
- Lack of light/heat
- Lack of food, drink, appropriate diet
- Lack of appropriate/new/desired clothing, no access to washing facilities
- Leaving older person sitting/lying in urine/feces

- Not providing appropriate accommodation/environment
- Absence of mobility aids (e.g., walking frame), movement being restricted
- Not aiding communication
- Not allowing social contact
- Lack of stimulation
- Ignoring
- Excluding
- Imprisonment, kept in confined area
- Not replacing essential items (e.g., batteries for hearing aid, ill-fitting dentures, and old/broken spectacles)
- Denying access to services (e.g., medical/health/community) when needed
- Not providing transport
- Not providing access
- Not providing information/education/options in order to make choices
- Not assessing/reassessing when needs arise or change
- Religious/spiritual/cultural needs not being met

Sexual Abuse

- Being forced into sexual activity without consent
- Penetration of mouth, anus, or vagina (with finger, penis, object)
- Indecent assault
- Rape
- Incest
- Inappropriate touching/kissing/fondling
- Inappropriate/sexualized language or conversation
- Masturbation: being forced to masturbate or to watch someone else
- Sexual harassment
- Pornography: forced to engage in, look at magazines, watch digital video disc
- Exposure
- Voyeurism
- Prostitution
- Sex trafficking
- Bestiality
- Ritual/satanic abuse

Discriminatory Abuse

- Making discriminatory/derogatory comments
- Harassment
- Hate crime: attacks on person/property
- On grounds of age, gender, race, culture, religion, sexual orientation, or disability not allowing access to:
  - a place
  - a service
  - treatment
- Forced marriage
- Honor Based Violence

Incidence and Prevalence

Researchers' and organizations' differing definitions of 'abuse' can affect the statistics presented in incidence and prevalence studies and the way in which people work with victims and abusers (including which cases they will/will not take on). Studies may also differ regarding the age of an 'older' person. Organizations working with older people can define old age as 50, 55, 60, 65 years or more. These differences cause problems in estimating how much elder abuse is being perpetrated. Also significant is the large amount of underreporting for various reasons. An older person may feel shame or embarrassment that someone close to them is harming them. Victims may love their perpetrator and choose to stay in the situation, although they may want the abuse to stop because maintaining their relationship with the perpetrator is more important than leaving or doing something about it. They may feel that no one will believe them if they disclose abuse. The victim may have a communication problem (e.g., physical disability – deafness, speech impairment after a stroke, or a learning disability) or a cognitive problem (dementia: brain injury) that could impede disclosure. An older person living in a care setting may be scared to speak out, fearing reprisals from staff caring for them. Workers themselves may not report abuse they see perpetrated by colleagues because they need their job and do not want to be considered whistle-blowers.

Researchers can face a range of problems when conducting studies (either quantitative or qualitative): developing definitions, gaining access to agencies, and identifying victims to obtain data. It is still useful to read some of the early prevalence studies (Finkelhor and Pillemer, 1988; Podnieks and Pillemer, 1990; Podnieks, 1992; NCEA, 1998) alongside more recent ones (Naughton et al., 2012; Lifespan of Greater Rochester, Inc., 2011; O'Keeffe et al., 2007; NCEA, 2006; Eisikovits et al., 2004; Teaster and Otto, 2007). Prevalence figures vary significantly – from 1 to 10% (Lachs and Pillemer, 2004), making actual rates for prevalence and incidence unclear.

The incidence and prevalence of elder abuse may remain unknown for the following reasons:

- Inconsistency in defining what constitutes abuse
- How statistics are collated
- Differences between countries (cultural): how society and the victims see abuse
- Underreporting because of fear of the abuser, the repercussions, not being believed, shame, and embarrassment.

Despite the vast literature on the quality of care in institutions, no global prevalence or incidence data are available. A survey of nursing home personnel in one US state disclosed that 36% of nursing and aide staff reported having seen at least one incident of physical abuse by other staff members in the preceding year and 10% admitting to having committed at least one act of physical abuse themselves. At least one incident of psychological abuse against a resident was observed by 81% of the sample and 40% admitted to committing such an act (Pillemer and Moore, 1989, 1990). The lessons from this are that all organizations should have whistle-blowing policies in place together with training programs that encourage the reporting of abuse.

Practice Issues and the Law

> …elder abuse is an issue to be considered in all domains of social work practice. Intervention in cases of elder abuse is often fraught with ambiguity and ethical dilemmas as the application of professional principles is less than straightforward, bringing to the fore personal, legal, and ethical concerns in the management of the client's safety and well-being. (Donovan and Regehr, 2010: p. 174)

Practice dilemmas can occur for anyone working with a victim of elder abuse. To promote best practice, workers need knowledge, time for reflection, and experience in dealing with cases. Comprehensive training programs need to be developed so that workers are clear about: what constitutes abuse (acknowledging cultural differences), how to identify and investigate abuse, and how to support a survivor of abuse in the long term. Adequate funding is required to improve professional support and interventions. Working with elder abuse is very different from child abuse because adults have the human right to live their life as they wish. Therefore, it is important to listen to and empower a victim. Workers may recognize that an older person is at risk of harm but she or he chooses to remain in that situation or the victim contacts an agency for help but then refuses to accept services (Linzer, 2004).

Key questions for continued debate concern:

- Mandatory reporting
- Statutory intervention
- Promoting self-determination
- Breaking confidentiality

Adults have human rights to live their life as they wish and the capacity to make decisions (including unwise ones) for themselves unless proven otherwise. However, there will be some older people who become vulnerable and susceptible to abuse because they lack mental capacity regarding some aspects of their life and may need protection. The law covers two main issues regarding abuse: whether there should be mandatory reporting of abuse and whether there should be specialist legislation to protect vulnerable adults. Many abusive acts are crimes and so practitioners across the professions should be aware of the criminal law that needs to be adhered to and utilized. However, some policymakers believe that elder abuse should not be criminalized and a welfare model should be followed (Bennett and Kingston, 1993; Mellor and Brownell, 2006; Nerenberg, 2008).

In certain parts of the world, specialist laws have been developed to protect adults, which include older people. For example, in Scotland, three civil statutes were introduced to form the legal framework for inquiry, assessment, and intervention (Mackay, 2008):

- Adults with Incapacity Act (Scotland) 2000
- Mental Health (Care and Treatment) (Scotland) Act 2003
- Adult Support and Protection (Scotland) Act 2007

Legislation that protects an older person who is being abused will usually cross different parts of the law: criminal, family, or civil. In the United States, older people are protected by federal and state statutes (which differ from state to state) under three categories of law:

- Adult protective services
- Institutional abuse
- Long-term care ombudsman program (Steigel and Klem, 2007).

In Canada, four main types of law cover elder abuse cases:

- Family violence laws
- Criminal law
- Adult protection laws
- Adult guardianship laws

In the province of Quebec, human rights legislation can also be used (Canadian Network for the Prevention of Elder Abuse www.cnpea.ca/canadian_laws_on_abuse_and_negle.htm#) – accessed 16.06.13.

Japan has introduced a specific elder abuse law – the Elder Abuse Prevention and Caregiver Support Law (April 2006) – which:

1. Legally defines and categorizes elder abuse for the first time.
2. Clarifies the responsibilities of national and local governments, designating the municipal government as the central agency in charge of elder abuse prevention.
3. Establishes the mandatory reporting system for both domestic and institutional abuse cases.
4. Emphasizes the importance of caregiver support (Watanabe, Y., International Longevity Center Japan, http://longevity.ilcjapan.org/t_stories/0703.html – accessed 16.06.13.).

A study concerned with the impact of the law concludes "it was a significant step in reporting systems and activities for increasing awareness among service providers. Further policy should address how to establish intervention teams and multi-agency networks, how to conduct home visits to investigate reports, and how to approach cases resistant to outside support." (Nakanishi et al., 2009, http://www.ncbi.nlm.nih.gov/pubmed/19041151 – accessed 16.06.13.)

In developed countries, policies, procedures, and guidance exist for organizations regarding how to prevent, report, and investigate cases of elder abuse. This is often seen as a social problem with social care agencies taking the lead in conjunction with the police. In Britain, the review of the Department of Health's guidance *No Secrets* considered whether the police or social services should take the lead in future: "The Police Service believed that national leadership should lie with Adults Social Care, while the Department of Health and, on a local level, felt responsibility should sit with Safeguarding Adult Boards." (DH, 2009: p. 7)

Conclusion

Progress is being made around the world in dealing with elder abuse. However, governments and organizations across the statutory, voluntary, and independent sectors ought to continue to raise awareness about the issue. Everyone needs education about how prevalent the problem is, who might

become a victim, who perpetrates harm and mistreatment, why this might happen, and where it occurs. Preventative work should be ongoing, while policies, procedures, and guidance should be in place for anyone working with an older person.

See also: Aging, Theories of; Child Neglect; Community Care in England; Elder Care; Life Course: Sociological Aspects; Retirement and Encore Adulthood: The New Later Life Course; Sexuality Over the Life Course.

Bibliography

Action on Elder Abuse, 1995. Bulletin, No 11. AEA, London.

Association of Chief Officers of Police and National Policing Improvement Agency, 2008. Guidance on Investigating Domestic Abuse. NPIA, London.

Association of Directors of Social Services, 2005. Safeguarding Adults: A National Framework of Standards for Good Practice and Outcomes in Adult Protection Work. ADSS, London.

Baker, A.A., 1975. Granny battering. Modern Geriatrics 5 (8), 20–24.

Bennett, G., Kingston, P., 1993. Elder Abuse: Concepts, Theories and Interventions. Chapman and Hall, London.

Bowes, A., Avan, G., Macintosh, S., 2008. They Put up with It What Else Can They Do? Mistreatment of Black and Minority Ethnic Older People and the Service Response. Age Concern Scotland, Edinburgh.

Burston, G.R., September 1975. Granny battering. British Medical Journal 3 (5983), 592–593.

Burston, G.R., May 1977. Do your elderly patients live in fear of being battered? Modern Geriatrics 7 (5), 54–55.

Decalmer, P., 1993. Clinical presentation. In: Decalmer, P., Glendenning, F. (Eds.), The Mistreatment of Elderly People. Sage Publications, London.

Department of Health, 1992. Confronting Elder Abuse. HMSO, London.

Department of Health, 1993. No Longer Afraid: The Safeguard of Older People in Domestic Settings. HMSO, London.

Department of Health, 2000. No Secrets: Guidance on Developing and Implementing Multi-agency Policies and Procedures to Protect Vulnerable Adults from Abuse. DH, London.

Department of Health, Criminal Justice System and Home Office, 2009. Safeguarding Adults: Report on the Consultation on the Review of 'No Secrets'. DH, London.

Donovan, K., Regehr, C., June 2010. Elder abuse: clinical, ethical, and legal considerations in social work practice. Clinical Social Work Journal 38 (2), 174–182.

Eastman, M., 1984. Old Age Abuse. Age Concern, London.

Eisikovits, Z., Winterstein, T., Lowenstein, A., December 2004. The National Survey on Elder Abuse and Neglect in Israel. University of Haifa: Faculty of Social Welfare and Health Studies.

Ferreira, M., Lindgren, P., 2008. Elder abuse and neglect in South African: a case of marginalization, disrespect, exploitation and violence. Journal of Elder Abuse and Neglect 20 (2), 91–107.

Finkelhor, D., Pillemer, K., 1988. The prevalence of elder abuse: a random sample survey. The Gerontologist 28 (1), 51–57.

Friday, P., Hartman, J., Lord, V., Exum, M., 2007. Batterers and the Battered: Role Reversals in Domestic Violence. Paper presented at the annual meeting of the American Society of Criminology, Atlanta Marriott Marquis, Atlanta, Georgia, 13 November 2007. http://citation.allacademic.com/meta/p_mla_apa_research_citation/2/0/0/7/0/p200700_index.html (accessed 18.06.13.).

Gelles, R.J., 1997. Intimate Violence in Families, third ed. Sage, Thousand Oaks, CA.

Gelles, R.J., Cornell, C.P., 1985. Intimate Violence in Families. Sage, Beverley Hills, CA.

Hayes, A.J., Fazel, S., 2008. Older adults in prison: vulnerability, abuse and neglect. In: Pritchard, J. (Ed.), Good Practice in Safeguarding Adults: Working Effectively in Adult Protection. Jessica Kingsley Publishers, London.

HelpAge India, 2013. Report on Elder Abuse in India 2013. HelpAge India, New Delhi.

Ingram, R., 2001. Is it a domestic? Or is adult abuse? In: Pritchard, J. (Ed.), Good Practice with Vulnerable Adults. Jessica Kingsley Publishers, London.

Kingsley, B., Johnson, S., 1993. Elder Abuse: The Ethical Dilemma [Unedited. Paper as presented at Conference of 23–25 February 1993]. Australian Institute of Criminology, Canberra. http://www.aic.gov.au/media_library/conferences/olderpeople/johnson.pdf (accessed 19.06.13.).

Kosberg, J., 2007. Abuse of Older Men. The Haworth Maltreatment and Trauma Press, Binghamton, NY co-published simultaneously as Journal of Elder Abuse and Neglect 19, 1–2.

Lachs, M., Pillemer, K., 2004. Elder abuse. The Lancet 364 (9441), 1192–1263.

Liao, S., 2006. Ethical Dilemmas in Elder Abuse: Successful Resolutions of Challenging Cases. Program and abstracts of the American Geriatrics Society 2006 Annual Scientific Meeting; 3–7 May 2006; Chicago, Illinois. Session 0413. http://www.medscape.org/viewarticle/532943 (accessed 19.06.13.).

Lifespan of Greater Rochester, Inc., May 2011. Under the Radar: New York State Elder Abuse Prevalence Study. Weill Cornell Medical Center of Cornell University, New York City Department for the Aging, New York.

Linzer, N., 2004. An ethical dilemma in elder abuse. Journal of Gerontological Social Work 43 (2–3), 165–173.

Mackay, K., 2008. Scottish legislative framework for supporting and protecting adults. In: Pritchard, J. (Ed.), Good Practice in the Law and Safeguarding Adults: Criminal Justice and Adult Protection. Jessica Kingsley Publishers, London.

Mba, C.J., 2007. Elder abuse in parts of Africa and the way forward. Gerontechnology 6 (4), 230–235. http://dx.doi.org/10.4017/gt.2007.06.04.006.00 (accessed 16.06.13.).

Mellor, J., Brownell, P. (Eds.), 2006. Elder Abuse and Mistreatment: Policy, Practice and Research. The Haworth Press Inc, Binghamton.

Nagele, B., Bohm, U., Gorgen, T., Toth, O., December 2010. Intimate Partner Violence against Older Women – Summary Report. Deutsche Hochschule der Polizei and Zoom, Gottingen.

Nakanishi, M., Hoshishiba, Y., Iwama, N., Okada, T., Kato, E., Takahashi, H., May 2009. Impact of the elder abuse prevention and caregiver support law on system development among municipal governments in Japan. Health Policy 90 (2–3), 254–261.

National Center on Elder Abuse, September 1998. The National Elder Abuse Incidence Study. NCEA, Washington, DC.

National Center on Elder Abuse, April 2006. Fact Sheet: Elder Abuse Prevalence and Incidence. www.ncea.aoa.gov/Main_Site/pdf/publication/FinalStatistics050331.pdf (accessed 19.06.13.).

Naughton, C., Drennan, J., Lyons, I., Lafferty, A., Treacy, M., Phelan, A., O'Loughlin, A., Delaney, L., 2012. Elder abuse and neglect in Ireland: results from a national prevalence survey. Age and Ageing 41 (1), 98–103. http://ageing.oxfordjournals.org/content/41/1/98.abstract. aff-1.

Nerenberg, L., 2008. Elder Abuse Prevention: Emerging Trends and Promising Strategies. Springer Publishing Co. LLC, New York.

O'Keeffe, M., Hills, A., Doyle, M., McCreadie, C., Scholes, S., Constantine, R., Tinker, A., Manthorpe, J., Biggs, S., Erens, B., 2007. UK Study of Abuse and Neglect of Older People Prevalence Survey Report. Comic Relief and Department of Health, London.

O'Malley, T.A., Everitt, D.E., O'Malley, H.C., Campion, E.W., 1983. Identifying and preventing family mediated abuse and neglect of elderly persons. Annals of Internal Medicine 98 (6), 998–1005.

Patel, N., 2003. Minority Elder Care in Europe: Country Profiles. PRIAE Policy Research Institute on Ageing & Ethnicity, Bolton.

Payne, B.K., 2013. Physical and Emotional Abuse of the Elderly. Community Orientated Policing Services, US Department of Justice, Washington, DC.

Pillemer, K., Finkelhor, D., 1988. The prevalence of elder abuse: a random sample survey. Gerontologist 28 (1), 51–57.

Pillemer, K., Moore, D., 1989. Abuse of patients in nursing homes: findings from a survey of staff. Gerontologist 29 (3), 314–320.

Pillemer, K., Moore, D., 1990. Highlights from a study of abuse of patients in nursing homes. Journal of Elder Abuse and Neglect 2 (1&2), 5–29.

Podnieks, E., 1992. National survey on abuse of the elderly in Canada. Journal of Elder Abuse and Neglect 4 (1&2), 5–58.

Podnieks, E., Pillemer, K., 1990. Survey on Abuse of the Elderly in Canada. Ryerson University, Toronto.

Pritchard, J., 19 October 1990. Charting the hits. Care Weekly, 10–11.

Pritchard, J., 8 July 1993. Gang warfare. Community Care, 22–23.

Pritchard, J., 1996. Working with Elder Abuse: A Training Manual for Home Care, Residential and Day Care Staff. Jessica Kingsley Publishers, London.

Pritchard, J., 2000. The Needs of Older Women: Services for Victims of Elder Abuse and Other Abuse. The Policy Press, Bristol.

Pritchard, J., 2001. Male Victims of Elder Abuse: Their Experiences and Needs. Jessica Kingsley Publishers, London.

Pritchard, J., 2007. Working with Adult Abuse: A Training Manual for People Working with Vulnerable Adults. Jessica Kingsley Publishers, London.
Shanhardass, M.K., 2009. No One Cares about Elder Abuse in India. One World South Asia, New Delhi. http://southasia.oneworld.net/peoplespeak/no-one-cares-about-elder-abuse-in-india (accessed 19.06.13.).
Steinmetz, S., 1978. The battered husband syndrome. Victimology 2 (3–4), 499–509.
Stiegel, L., Klem, E., 2007. Adult Protective Services, Institutional Abuse and Long Term Care Ombudsman Program Laws: Citations by State. American Bar Association Commission on Law and Aging for the National Center on Elder Abuse, Washington, DC. http://www.americanbar.org/content/dam/aba/migrated/aging/about/pdfs/APS_IA_LTCOP_Citations_Chart.authcheckdam.pdf (accessed 02.02.14.).
Straus, M., 1999. The controversy over domestic violence by women. In: Arriaga, X.B., Oskamp, S. (Eds.), Violence in Intimate Relationships. Sage Publications, Thousand Oaks, CA.
Straus, M.A., Gelles, R.J., Steinmetz, S.K., 1980. Behind Closed Doors: Violence in the American Family. Anchor, Garden City, NY.
Sukosky, D.G., 1987. Role reversal of the elderly: intervention and prevention. Family Life Educator 6 (1), 14–17.
Teaster, P.B., Otto, J.M., 2007. The 2004 Survey of State Adult Protective Services: Abuse of Adults 60 Years of Age and Older. National Center on Elder Abuse, Washington, DC. http://www.ncea.aoa.gov/Resources/Publication/docs/APS_2004NCEASurvey.pdf (accessed 02.02.14).
World Health Organization, 2002. Missing Voices: Views of Older Persons on Elder Abuse. WHO, Geneva.
World Health Organization, August 2011. Elder Maltreatment – Fact Sheet 357. WHO, Geneva. www.who.int/mediacentre/factsheets/fs357/en/index.html (accessed 16.06.13.).
Watanabe, Y. The Elder Abuse Prevention and Caregiver Support Law. International Longevity Center Japan: http://longevity.ilcjapan.org/t_stories/0703.html (accessed 16.06.13.).

Relevant Websites

www.elderabuse.org.uk – Action on Elder Abuse (accessed 16.06.13.).
www.elderabuse.co.za – Action on Elder Abuse South Africa (accessed 16.06.13.).
www.anpea.com.au – Australian Network for the Prevention of Elder Abuse (accessed 16.06.13.).
www.preventelderabuse.eu – The European Reference Framework Online for the Prevention of Elder Abuse (accessed 16.06.13.).
www.helpage.org/elderabuse – HelpAge International (accessed 16.06.13.).
www.inpea.net – International Network for the Prevention of Elder Abuse (accessed 16.06.13.).
www.ncea.aoa.gov – National Center on Elder Abuse (accessed 16.06.13.).
www.who.int/en/ – World Health Organization (accessed 16.06.13.).

Elder Care

Mary P Sullivan, Brunel University, Uxbridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Both increasing life expectancy and declining fertility rates across the globe mean that by 2045 the number of older people will exceed the number of children. The implications of population aging are made more complex by diverse family structures, disease, and disability, particularly for the oldest old, and slow economic growth. The management of these and other consequences impact upon social work's participation and contributions to interdisciplinary approaches to elder care. Adopting person-centered approaches to care, social work promotes citizenship; dignified care to minimize the impact of transitions, decline, or disability; and protect the most vulnerable.

Introduction

Health and social care for an aging population is drawing increasing attention around the globe. This article discusses social work's contribution to elder care. It briefly sets the context for gerontological social work by reviewing the changing global demographic profile, introduces social and economic policy considerations that directly impact on how care is delivered, and characteristic health and social care needs of older people. Social work's key roles in assessment of need, care management, and therapeutic interventions are then reviewed to draw attention to the profession's contribution to intervening in and facilitating caring environments for older people with complex and changing needs. Finally, gerontological social work's capacity to address the needs of an aging population is considered and recommendations for further development made. This article draws on practice examples from various countries, but a thorough cross-national analysis is beyond its reach.

Setting the Context for Social Work with Older People

Population Aging

The world's population is aging. Both increasing life expectancy and declining fertility rates in the global north and south means that by 2045 the number of older people will, for the first time, outnumber the number of children. The global percentage of people over the age of 60 years is expected to rise from 11% in 2009 (700 million) to 22% in 2050 (2 billion). And while the global north has already experienced the impact of a transformed age structure, the growth of an aging population is far greater in the global south and these countries are expected to draw level to the global north by 2050 (UN, 2009). The resultant dependency or potential support ratio is what concerns nations about how they will respond to these challenges as the size of the working-age population (16–64 years) shrinks in relation to the increasing size of the dependent or nonworking population (65 years or older). In 2009, there were nine potential workers for every person 65 or older. By 2050, this is expected to change to four workers for every older person (UN, 2009). While countries around the globe are exploring ways to extend older people's participation in the workforce, relatively little is known about older worker health and well-being, the feasibility of an older workforce in varied work environments, or the larger scale social and economic impacts of extending working lives (Bossaert, 2012; Kalache et al., 2005). Thus, the far-reaching social, political, and economic impact of this 'demographic transition' is unprecedented and presents important opportunities and challenges for all nations.

The implications of population aging are both complex and numerous, and have substantial relevance for social work. Some key issues are highlighted here to enhance the critical lens that social workers adopt and that contest the social construction of ageing as a 'problem.' First, the provision of social support for older people within the family is being confronted by a reduction in intergenerational assistance and/or filial piety as family size reduces, as women engage in work outside the home, and as children migrate for employment (Silverstein and Giarrusso, 2010). Second, as people live longer beyond retirement and the working-age population shrinks, the current levels of provision of health and social welfare, including pensions, are no longer economically sustainable and ultimately demand radical transformation (Bossaert, 2012; Marshall and Taylor, 2005). In the global south where welfare structures remain scarce, the prospect of any formal care system significantly diminishes despite need (e.g., less informal care exchanges due to declining fertility rates) (Kalache et al., 2005).

Third, the fastest growing segment of the population is people over the age of 80 years (UN, 2009). The World Health Organisation (2004) reported that 46% of people over the age of 60 years were living with moderate to severe disability, and levels of disability were higher in low- to middle-income countries. This percentage gradually increases with age, for example, 55% of people aged 85–89 years are disabled and 90%of people aged 95–99 years are disabled (Poon et al., 2005). Poon et al. (2005) estimate that one-third of the oldest old are able to live independently despite health conditions, one-third are functionally impaired, and one-third are either extremely frail or disabled. While disease in the global north is commonly associated with cardiovascular and respiratory disease and cancer, it is worth noting the number of people

who will experience dementia is likely to increase. The World Alzheimer Report (2010) projects the number of people across the globe who will be affected by dementia will reach 67.5 million by 2030.

Policy Context

The current global financial crisis has had a dramatic impact on policy direction as countries grapple with economically and socially viable alternatives to support a population that is still commonly viewed as a burden. These demographic fears (Mullun, 2000) easily contribute to the oversimplification of an ideal solution to what is otherwise very complex and do little to tackle issues such as ageism in health and social care structures and/or intergenerational tensions. For example, the concern about health care costs due to the expansion of morbidity fails to consider health inequalities earlier in life. Older people who engage in either paid or charitable work beyond typical retirement age do not confirm society's construction of dependency. And, grandparents who provide care for grandchildren while the middle generation pursue careers fail to corroborate the absence of thriving interdependent relationships and informal social exchanges.

The United Nations Madrid International Plan of Action on Ageing (MIPAA) (2002) identified three priorities to assist governments and other nongovernmental organizations to "reorient the ways in which their societies perceive, interact with and care for their older citizens" (UN, 2012: p. 5). These priorities included (1) older people and development, (2) advancing health and well-being in older age, and (3) ensuring enabling and supportive environments. While these priorities may be realized in individual countries, the essential thrust is promoting social inclusion, economic security, good health, and safe and accessible environments. The International Federation of Social Work's (IFSW) 'International Policy on Ageing and Older Persons' (2009) drew on the MIPAA to translate salient implications for social work around the globe. Briefly, the IFSW acknowledges social work's strategic position to enable the full social integration of older people. This includes promotion of their social, economic, and intellectual contributions and considered macrolevel and microlevel advocacy to tackle any social conditions that hinder security, health, and well-being (Hokenstad and Roberts, 2011).

While a genuine commitment to pursue these objectives was likely given by the nations involved, no one could predict a global financial crisis and subsequent contraction of welfare programs (including pensions) in so many countries, or persistent political instability and conflict in others. Such an environment does not bode well for those who are already marginalized or disadvantaged due to socioeconomic inequalities. Policy makers, however, have to make choices.

Alongside the principles embedded in the MIPPA, the concepts of active, healthy, or successful aging have become familiar in Western policy discourses. More recently, the term active aging has taken a stronghold in global social policy as a means of enhancing quality of life as people age. The principles broadly promote activity throughout the life course as a protective measure from decline and isolation, and activity in later life as a means to encourage pursuits that contribute to well-being, family and community engagement and participation, intergenerational solidarity, and respect for diversity (Walker, 2009). However, active aging easily leads policy makers into a quagmire, given this cannot be achieved by individual effort alone. Determinants of health unequivocally draw attention to the role of structural or contextual factors in optimizing good health or quality of life at any age (Holstein et al., 2011; Lloyd et al., 2013). In the event that a poor economy only permits a very conservative or tentative reform of welfare to support active aging, it follows that it becomes an individual responsibility to age well. For those who are very old and unavoidably dependent, further marginalization denies them an authentic caring environment (Boudiny, 2012; Holstein et al., 2011).

Health and Social Care Needs of Older People

A diverse range of biological, psychological, and sociological theories of aging provide the backdrop for understanding the determinants associated with longevity or decline, psychological development in later years, and the experience of being old. While this is the lens by which researchers have come to understand older age, the purpose here is to present what is commonly understood in relation to health and social care problems, given their consequences are directly relevant to social work practice.

Nevertheless, in the social sciences, the life course perspective has become integral to understanding how human lives are intimately linked to and embedded within historical events, social and cultural institutions, and technological advancements (Green, 2010). To briefly elaborate, the principles of this perspective progress an understanding of how lives may take different paths, despite similar events, and the ever-constant tension between continuity and change over the life course. Thus, age and its various meanings are lifelong processes that are situated within (1) an individual's interconnectedness with others, (2) historical time and place, (3) transitions and their timing, and (4) individual agency (Grenier, 2012). Moreover, the utility of the life course perspective also lays in its recognition of "blurred and ragged transitional periods rather than clear-cut ages and stages" (Green, 2010: p. 24). Accordingly, this lens is particularly applicable in social work, given its person-in-environment orientation and attention to lifelong development.

Physical and Mental Health

Gradual physical changes to the body are a normal part of ageing. People tend to 'slow down' as a result of changes in bones, skin, muscle, and circulation, resulting in energy levels, flexibility, and strength that may be lesser than in earlier life. For many, these normal changes are accompanied by some form of disease and disability, which will affect an individual's quality of life. Disease and disability develop from a complex combination of intrinsic or extrinsic factors such as genes, environment, behavior, and lifestyle and access to health care (Victor, 2010).

Good health, or the absence of disease or illness, is commonly viewed as central to maintaining quality of life in older age (Victor, 2005). For older people, however, there is a tendency to evaluate their health and well-being using much broader criteria such as their ability to perform normal roles and responsibilities (Ray and Phillips, 2012). From a health care cost and provision perspective, the aim is to minimize those factors that cause mortality and morbidity. These factors will vary between and within countries owing to the various determinants of health associated with socioeconomic and political conditions (Victor, 2005), or what is commonly referred to as accumulated advantage or disadvantage. Determinants of health are reflected in the leading cause of death in various countries. For high-income countries, Eurostat (2009) lists the major cause of deaths for those aged 65 years and older as ischemic heart disease, cerebrovascular disease, diseases of the respiratory system, lung cancer, and colorectal cancer. For lower income countries, the overall leading causes of death include coronary heart disease, lower respiratory conditions, human immunodeficiency virus/acquired immunodeficiency syndrome, perinatal conditions and stroke, and other cerebrovascular conditions (WHO, 2008).

For older people, it is not uncommon to cope with comorbidity (i.e., conditions occurring because of a leading disease) or multimorbidity (i.e., simultaneous existence of several conditions). Common chronic problems include cardiovascular disease, diabetes mellitus, chronic obstructive pulmonary disease, asthma, arthritis, and unipolar depressive disorders (Pomerleau et al., 2008). While there is some support that morbidity is being compressed to fewer years before death (Freedman et al., 2002), the picture is complex and some would suggest that there may be more older people living with disability longer (Lafortune et al., 2007).

The growing number of older people with mild or severe mental health problems reinforces the complexity of health problems in this population. Similarly, as care continues to be delivered using community-based approaches, innovative models of support are required to respond to the new challenges and the impact of long-term inequality, stigma, and disempowerment (Sullivan et al., 2004). Approximately one in five people over the age of 65 years experience a mental health problem. This includes dementia, depression, psychosis, bipolar disorders, schizophrenia, and anxiety disorders (Jeste et al., 1999). Prevalence rates in institutional care are of greater concern and it is estimated that 80% of all residents are experiencing depression or behavioral disturbances as a result of dementia (Brodaty et al., 2003).

Equally disabling, but often overlooked, is sensory loss. In the United Kingdom, for example, 6.4 million people aged 65 years and older have some form of hearing loss. Almost 700 000 are considered to have severe or profound loss (Action on Hearing Loss, 2011). According to the Royal National Institute for the Blind (2012), of the 2 million in the United Kingdom with sight loss, 80% are over age 60 years. And while for many this is an anticipated feature of their older age, sensory loss can lead to depression, low self-esteem, and social isolation (Stuen, 2006).

The impact of health conditions on an older person's daily living activities, the person's ability to manage associated risks (e.g., falls and drug reactions), and his/her individual and family life (i.e., participation, adjustment, and caring) have key relevance for social work practice. The management of these and other consequences of aging and older age also have relevance for social work's contribution to interdisciplinary approaches to elder care.

Social Care and Support

The heterogeneity of the aging population reinforces the necessity to adopt a person-centered approach to social care and support, given that gender, family status, social class, ethnicity, occupation, life history, and health and disability directly influence an older person's ability to manage change and transitions at this stage in life. In addition to the possibility of disease and disability, this stage may be characterized by changes in identity, roles, family structure, and loss and stress. The features of someone's life course and the impact of, for example, the loss of a spouse, becoming a grandparent, or living in poverty will also be affected by an individual's access to intrapersonal, interpersonal, and external resources. Thus, the older population are not easily categorized as being independent or dependent.

Social workers often intervene to coordinate an older person's natural support systems and bolster these with external supports if necessary. Social work's expertise in this area tends to focus on minimizing vulnerability or risk by the provision of practical support at home. In this role, the practitioner may work with communities to cultivate relationships and supports, and with individuals to balance "conflicting needs and aspirations, working creatively with risk, seeing to preserve autonomy, and working with systems with diverse and contradictory opinions and views" (Ray and Phillips, 2012: p. 7).

Critical Gerontological Social Work Practice

Professional social work is uniquely positioned, usually as part of interdisciplinary teams, to provide a valuable contribution to well-being in later life through its leadership in relation to psychosocial issues and holistic approaches to problem solving and intervention. As Kaplan and Berkman (2011) have argued, social work values, skills set, and resources, and roles that span across a variety of organizations, enable practitioners to provide assessment and supportive, therapeutic, and educational services to individuals and their families.

Social work's contribution also extends much further. Foremost, older service users highly regard social work's involvement in the process of obtaining guidance and assistance (Beresford et al., 2011). Research by Manthorpe et al. (2007) found that "… social workers' knowledge about specialist services, persistence, commitment, reliability, and being supportive, sympathetic and prepared to listen" was repeatedly identified as essential in the helping process. Given the prominence of the social and person-in-environment perspective within the profession, social work recognizes individual need as located in the broader social context. Extending far beyond the biomedical model, individual need and/or threats to well-being are considered in relation to rights and citizenship and how an individual's life experiences are shaped by structures governing society. Social work has the capacity, when bureaucratic systems

so easily exclude vulnerable older people, to engage in advocacy and empowerment work to ensure older people have the right to make choices, the right to control their own lives, and the right to high-quality care (Hokenstad and Roberts, 2011; McDonald, 2010; Lloyd, 2006) and ensure that systematic discrimination does not deny any older person an enabling environment (Hokenstad and Roberts, 2011). Social work's role then naturally extends to strategically positioning itself to influence policy and system development.

Assessment

Using a combination of a biopsychosocial and strengths approach, a social worker characteristically adopts an assessment of need that systematically considers functioning in the following domains (Ray and Phillips, 2012; Nelson-Becker and Sullivan, 2012; Mellor and Lindeman, 1999):

1. Physical – medical history, medications, functional abilities, appearance, behavior, and substance use
2. Psychological – personality characteristics, identity, affect, and cognitive functioning
3. Social – family structure, other relationships and support networks, social roles, occupation, financial status, leisure, and community engagement
4. Environment – living conditions, risk and vulnerability, and degree of independence within the environment
5. Cultural – values, beliefs, communication, and patterns of interaction within cultural domain
6. Spiritual – religion, personal meaning and purpose, life aspirations, wholeness in relationships, and self-determination

Social workers may use standardized measures during the assessment. In specialized services such as mental health, for example, social workers may use measures to screen or narrow down the possible problem, assist in care planning, communicate assessment outcomes to other professionals, determine an optimal setting for care, or evaluate the outcomes of interventions.

The 'process' of the assessment is as important as the assessment itself. Typically grounded in a problem-solving approach, a social worker has a significant amount of power to direct the assessment and its procedural aspects and identify agency or professionally determined outcomes versus service-user choices (Sullivan, 2009). To moderate professional power and control, particularly problematic when service users are frail or cognitively impaired, a social worker must adopt a critically reflective approach. In addition, a social worker makes use of effective communication skills to cultivate a rapport with an older person and ensure their full participation (Harwood, 2007). Furthermore, the social worker simultaneously negotiates questioning and information giving, the involvement of significant others for additional data (e.g., family caregivers), and the handling of difficult or complex individual systems (Ray et al., 2009; Geron, 2006).

Care Management and Therapeutic Interventions

Following the outcomes of the assessment process, a social worker's role varies depending on the agency and the nature of the services provided – for example, community-based services, acute care, hospice, mental health or substance abuse settings, prisons, and nursing homes. Generally speaking, gerontological social work interventions are "directed at enhancing dignity, self-determination, personal fulfilment, quality of life, optimal functioning and ensuring the least restrictive living environment possible" (Mellor and Lindeman, 1999: p. 3). More recently, intervention research has contributed to a growing body of evidence of effectiveness to inform the design and delivery of psychosocial interventions (Cummings and Kropt, 2009; Burnette et al., 2003).

Familiar in the European Union, support planning and care management play the key role in community-based gerontological practice. Under the guise of community-based rather than institutional care, care managers act on behalf of government to carefully administer welfare resources for those who meet clearly defined eligibility criteria (Sullivan, 2009; Lymbery, 2005). Usually focusing on disability and dysfunction, the core functions in this model of delivery involve both the development and coordination of a support plan, access and linkage to a variety of resources (e.g., home carers, meal on wheels, transport services, and financial or legal assistance), and regular monitoring for changes (Ray and Phillips, 2012; Naleppa, 2006). Personal safety and security are also essential considerations. In the United Kingdom, recent legislation has strengthened a push toward individual or personal budgets whereby need is translated into a carefully calculated budget that is then transferred to the service user for purchasing and managing their own care (Carr and Robbins, 2009).

Not surprising, this model of practice is highly criticized. Its administrative rather than therapeutic approach has reduced social work to rationing agents who construct need as stable and treatable (Lymbery, 2012). Decision making is undoubtedly justified using discourse that places autonomy and independence as essential for all older people. A life course or person-in-environment approach is inconsistent with this system, thereby further reducing opportunities for social work intervention. Older people with complex and changing needs (i.e., those in greatest need) and the development of humane and interdependent caring environments are unfortunately abandoned (Lloyd, 2010; Sullivan, 2009).

Social workers positively engage in therapeutic models of intervention focusing on many of the critical challenges that older people encounter such as health issues and end-of-life care, mental health problems, and social roles (e.g., grandparents raising grandchildren). There is a range of theoretically and research-informed intervention approaches for individuals, families, groups, or communities. It is not possible to elaborate on all developments in evidence-informed gerontological practice here, but it is worth noting a few examples to demonstrate social work's contribution.

According to Arthritis Research UK (2008), more than one-third of the population over the age of 50 years have pain that interferes with normal activities of daily living. Its impact can be far reaching and include limited mobility, isolation due to restricted activities, and feelings of despair. Cognitive-behavioral and psychoeducational interventions are widely used, given their cost, effectiveness, and flexibility, to increase coping skills, reduce emotional distress, and maximize self-care

behaviors (Yoon and Doherty, 2009). The effectiveness of cognitive-behavioral and psychoeducational interventions for depression and anxiety disorders (McInnis-Dittrich, 2006) and substance abuse (Cummings et al., 2009) is also well documented.

Psychosocial problems as a result of dementia are also common at all stages of the disease. Gerontological social workers are involved in supporting the older person and his/her family and/or caregivers. Personality and mood changes and decline in functional abilities may cause increasing levels of stress and strain (Judge et al., 2009). Support or mutual aid group interventions for family caregivers are commonly provided. A mixed range of evidence suggests effectiveness using psychotherapeutic, cognitive, or problem-solving approaches to manage stressors (Ray et al., 2009; Greene and Riley, 2006). A variety of interventions with individuals with dementia are also being used by social work with varying levels of evidence of effectiveness to reduce the negative impact of illness-related stressors and improve quality of life (Kaplan and Berkman, 2011). Specifically, reminiscence and life review appear to provide successful approaches that are also flexible, cost-effective, and nonstigmatizing (Ray et al., 2009; Lai et al., 2004). A variety of approaches to reminiscence and life review are available. However, the essential thread is the use of memories to explore events, unresolved issues and conflicts, and inner strengths to build a positive sense of control and self-efficacy (Ingersoll-Dayton and Bommarito, 2006).

Social Work's Capacity to Respond to the Needs of an Aging Population

Social work's leadership in systematically addressing the breadth of age-related issues ranging from age discrimination and socioeconomic disadvantage to complex co-morbidityor multimorbidity and elder abuse requires continual progress to keep pace with rapid social change and the correlate ethical problems that arise. Yet, the social work profession, as a whole, has long overlooked the needs of an aging population (Sullivan, 2009). According to Kaplan and Berkman (2011), 4% of social workers in the United States have a geriatric specialism but 81% of social workers are working with older people. Social work education has not championed gerontological practice and social work intervention research has not kept pace with an increasing number of older people with complex needs. Evidence from a variety of countries continually reveals that social work students do not give preference to future practice with older people (Richards et al., 2013). As previously described, the IFSW's policy on aging and older people outlines social work's responsibility to respond to global aging and is testament that it is time to invigorate the profession's capacity to take action.

High-profile gerontological social work initiatives are occurring in North America. Since 1999, the Geriatric Social Work Initiative (GSWI), funded by the John A. Hartford Foundation in the United States, collaborates with various social work programs and organizations to raise the visibility and desirability of gerontological social work and prepare the profession to deliver high-quality care (GSWI, 2012). The GSWI's three related programs have transformed social work with older people in the United States. The Scholar's Program trains academic champions to teach future generations of social workers and engage in cutting-edge research with the Doctoral Fellow Programs, which recruit doctoral students to sustain leaders in teaching and research. The National Centre for Gerontological Social Work is 'gerontologizing' education by embedding gerontological competencies in the curriculum and thus ensuring the workforce is prepared for practice. Finally, the Hartford Partnership Program for Aging Education is developing models of field education, including the dissemination of new knowledge into the practice arena (GSWI, 2012).

Canada's National Initiative for the Care of the Elderly (NICE), founded in 2005, is a nonprofit charitable organization and social work–led network of researchers, practitioners, and students committed to improving care for older people. Initially funded by the Canadian government, the approach here is to bring research to practice and is purposively interdisciplinary owing to the research evidence that suggests best practice in geriatric care is interdisciplinary. Its three goals include to (1) help close the gap between evidence-based research and actual practice; (2) improve the training of existing practitioners, geriatric educational curricula, and student interest in geriatric care; and (3) effect positive policy changes for the care of older people (NICE, 2012). Using a series of themes (e.g., dementia care, end-of-life care, elder abuse, and caregiving), teams review the research evidence to develop and disseminate user-friendly interdisciplinary practice tools (paper and digital). NICE is currently undertaking a large-scale evaluation of the format and effectiveness of their tools (NICE, 2012).

Gerontological social work in other countries around the globe does not appear to be supported by a formal agenda for development. In the European Union, for example, welfare delivery and approaches to social work vary from country to country. Very generally speaking, social work is typically government funded, although it may be delivered by nongovernmental organizations. Private social work delivery such as that found in North America is not a trend. Practitioners assume the role of care manager or coordinator of care provision in the home – family care, private care, and state care (Soderberg et al., 2012; Crossland, 2008). In this primarily administrative model of practice, social workers have become agents charged with rationing and prioritizing state-funded support (Topo, 2011; Phillips, 2006). However, in some states in the European Union, it seems that social work has found a strong footing in areas of service-user empowerment and participation, promoting citizenship and advocacy, and playing a role in integrating health and social care systems (Emilsson, 2009).

In Australia, social work with older people appears to align itself with a more therapeutic model with various roles beyond that of care management. However, here too, practitioners tend to view gerontological practice as routine and unfulfilling (Ozane, 2006; Hughes and Heycox, 2005). These social workers are practicing in the community, hospitals, geriatric mental health teams, and dementia and end-of-life care. Private practice is also flourishing in areas such as the management of major life transitions and intergenerational

family conflict. Despite social work's visibility in these settings, Ozane (2006) urgently recommends an articulated gerontological action strategy to strengthen education and training, research capacity, and leadership.

China has witnessed an unprecedented rate at which the population is aging. The approach to care remains largely traditional in that care of older people is a normative family duty. Filial piety deeply penetrates all aspects of society, and no less is it felt than in the political realm as alternatives are considered to mitigate the effects of longer life expectancy, smaller extended families, migration, new market economies, and so forth (Xu and Chow, 2011; Chi, 2006; Zhan et al., 2006). In mainland China, social work is delivered primarily by civil society, nongovernmental organizations, and the private sector (Chi, 2006). The two key developments appear to be the growth of institutional care and various community-based initiatives to supplement family care. Preferences appear to be those that permit the older person to remain at home. Generating funds from a variety of sources (e.g., individual and corporate donations and government sources), meal programs, self-help groups, day centers, in-home care, and recreational programs are flourishing (Xu and Chow, 2011). Social work's direct contribution in these newer developments appears to be community-based practice. However, Chi (2006) states that gerontological social work is urgently required to address the psychosocial well-being of older people and their families, advance indigenous evidence-based interventions, and contribute to interdisciplinary approaches to care.

Conclusion

The implications of an aging population across the globe are numerous. All nations and their citizens are being confronted with both slow-moving economies and an increasing number of older people with complex needs. The subsequent challenges for elder care in social work are layered in social and economic policies, the competent provision of care management and therapeutic interventions, education and research, and the profession's ability strategically to interact with social institutions to remove the barriers for the creation of inclusive, caring, and enabling environments.

The aging population is diverse. Individual life course patterns are accumulated in later years and reflected in health and social care needs and a person's ability to draw on resources for support. The meaning of aging and older age is also experienced within particular contexts that may contribute added burden. Thus, person-centered approaches to care are necessary to promote citizenship; dignified care to minimize the impact of transitions, decline, or disability; and protect the most vulnerable.

More recently, critical gerontological social work initiatives have articulated the profession's unique contribution to supporting older people and their families. Practice is characterized by both macro- and microlevel engagement, specialized skills to address complexity and co- or multimorbidity, evidence-informed interventions, and high-level commitment to working with systems to encourage ethical responses to constraining features in contexts of practice. Social work is about working with those who are not privileged to foster human flourishing and supporting individual capabilities. For critical gerontological social workers, the time is now for compassionate thinking and considered responses.

See also: Community Care in England; Elder Abuse; Gendered Care and Social Work Practice.

Bibliography

Action on Hearing Loss, 2011. Facts and Figures on Hearing Loss and Tinnitus. http://www.actiononhearingloss.org.uk/supporting-you/factsheets-and-leaflets/deaf-awareness.aspx (accessed 16.12.12.).

Arthritis Research UK, 2008. Arthritis in the UK – The Key Fact. http://www.arthritisresearchuk.org/arthritis-information/data-and-statistics.aspx (accessed 16.12.12.).

Beresford, P., Bewley, C., Branfield, F., Flemming, J., Glynn, M., 2011. Supporting People: Towards a Person-Centred Approach. Policy Press, Bristol.

Bossaert, D., 2012. The Impact of Demographic Change and Its Challenges for the Workforce in the European Public Sectors: Working Paper 2012/W/01. European Institute of Public Administration, Maastricht.

Boudiny, K., 2012. Active ageing: from empty rhetoric to effective policy tool. Ageing and Society. http://dx.doi.org/10.1017/S0144686X1200030X. Firstview published 10 July 2012.

Brodaty, H., Draper, B.M., Millar, J., Low, L.F., Lie, D., Sharah, S., Paton, H., 2003. Randomized controlled trial of different models of care for nursing home residents with dementia complicated by depression or psychosis. Journal of Clinical Psychiatry 64 (1), 63–72.

Burnette, D., Morrow-Howell, N., Chen, L., 2003. Setting priorities for gerontological social work research: a national Delphi study. The Gerontologist 43 (6), 828–838.

Carr, S., Robbins, D., 2009. Research Briefing 20: The Implementation of Individual Budget Schemes in Adult Social Care. Social Care Institute for Excellence, London.

Chi, I., 2006. Older adults in China. In: Berkman, B. (Ed.), Handbook of Social Work in Health and Aging. Oxford University Press, Oxford, pp. 953–958.

Crossland, J., 2008. Care in the community for older people: an English perspective on the German care system. European Journal of Social Work 11 (3), 313–316.

Cummings, S.M., Kropt, N.P., 2009. Overview of evidence-based practice with older adults and their families. In: Cummings, S.M., Kropt, N.P. (Eds.), Handbook of Psychosocial Interventions with Older Adults. Routledge, New York, pp. 1–10.

Cummings, S.M., Bride, B., Cassie, K.M., Rawlins-Shaw, A., 2009. Substance abuse. In: Cummings, S.M., Kropt, N.P. (Eds.), Handbook of Psychosocial Interventions with Older Adults. Routledge, New York, pp. 206–233.

Emilsson, U., 2009. Health care, social care or both? A qualitative exploration of different focuses in long term care of older people in France, Portugal and Sweden. European Journal of Social Work 12 (4), 419–434.

Eurostat, 2009. Major Causes of Death for Persons Under 65 Years and Aged 65 and More – Standardised Death Rates (SDR) per 100 000 Inhabitants. http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Causes_of_death_statistics_-_people_over_65 (accessed 16.12.12.).

Freedman, V., Martin, L., Schoeni, R., 2002. Recent trends in disability and functioning among older adults in the United States: a systematic review. Journal of the American Medical Association 288 (24), 3137–3146.

Geriatric Social Work Initiative, 2012. GSWI Programs. http://www.gswi.org/programs/index.html (accessed 15.12.12.).

Geron, S.M., 2006. Assessment. In: Berkman, B. (Ed.), Handbook of Social Work in Health and Aging. Oxford University Press, Oxford, pp. 721–727.

Green, L., 2010. Understanding the Life Course: Sociological and Psychological Perspectives. Polity Press, Cambridge.

Greene, R.R., Riley, J., 2006. Family and group interventions. In: Berkman, B. (Ed.), Handbook of Social Work in Health and Aging. Oxford University Press, Oxford, pp. 757–764.

Grenier, A., 2012. Transitions and the Lifecourse. The Policy Press, Bristol.

Harwood, J., 2007. Understanding Communication and Aging: Developing Knowledge and Awareness. Sage Publications, Thousand Oaks, CA.

Hokenstad, M.C., Roberts, A.R., 2011. International policy on ageing and older persons: implications for social work practice. International Social Work 54 (3), 330–343.

Holstein, M.B., Parks, J.A., Waymack, M.H., 2011. Ethics, Aging and Society. Springer Publishing, New York.

Hughes, M., Heycox, K., 2005. Promoting reflective practice with older people: learning and teaching strategies. Australian Social Work 58 (4), 344–356.

Ingersoll-Dayton, B., Bommarito, A., 2006. Reminiscence and life review. In: Berkman, B. (Ed.), Handbook of Social Work in Health and Aging. Oxford University Press, Oxford, pp. 781–787.

Jeste, D.V., Alexopoulos, G.S., Bartels, S.J., Cummings, J.L., Gallo, J.J., Gottlieb, G.L., Halpain, M.C., Palmer, B.W., Patterson, T.L., Reynolds, C.F., Lebowitz, B.D., 1999. Consensus statement on the upcoming crisis in geriatric mental health: research agenda for the next two decades. Archives of General Psychiatry 56, 848–853.

Judge, K.S., Menne, H.L., Whitlatch, C.J., 2009. Stress process model for individuals with dementia. The Gerontologist 50 (3), 294–302.

Kalache, A., Barreto, S.M., Keller, I., 2005. Global ageing: the demographic revolution in all cultures and societies. In: Johnson, M.L. (Ed.), The Cambridge Handbook of Age and Ageing. Cambridge University Press, Cambridge, pp. 30–46.

Kaplan, D.B., Berkman, B., 2011. Dementia care: a global concern and social work challenge. International Social Work 54 (3), 361–373.

Lafortune, G., Balestat, G., Disability Study Expert Group Members, 2007. Trends in Severe Disability among Elderly People: Assessing the Evidence in 12 OECD Countries and Future Implications. Organisation for Economic Cooperation and Development, Paris.

Lai, C.K.Y., Chi, I., Kayser-Jones, J., 2004. A randomized controlled trail of a specific reminiscence approach to promote the well-being of nursing home residents with dementia. International Psychogeriatrics 16 (1), 33–49.

Lloyd, L., 2006. A caring profession? The ethics of care and social work with older people. British Journal of Social Work 36, 1171–1185.

Lloyd, L., 2010. The individual in social care: the ethics of care and the 'personalisation agenda' in services for older people in England. Ethics and Social Welfare 4 (2), 188–200.

Lloyd, L., Tanner, D., Milne, A., Ray, M., Richards, S., Sullivan, M.P., Beech, C., Phillips, J., 2013. Look after yourself: active ageing, individual responsibility and the decline of social work with older people in the UK. European Journal of Social Work. European Journal of Social Work. http://dx.doi.org/10.1080/13691457.2013.829805.

Lymbery, M., 2005. Social Work with Older People: Context, Policy and Practice. Sage, London.

Lymbery, M., 2012. Social work and personalisation: fracturing the bureau-professional compact?. Advance Access published November 4, 2012. British Journal of Social Work. http://dx.doi.org/10.1093/bjsw/bcs165.

Manthorpe, J., Moriarty, J., Rapaport, J., Clough, R., Cornes, M., Bright, L., Illiffe, S., OPRSI, 2007. "There are wonderful social workers but it's a lottery": older people's views about social workers'. British Journal of Social Work 38, 1132–1150.

Marshall, V.W., Taylor, P., 2005. Restructuring the lifecourse: work and retirement. In: Johnson, M.L. (Ed.), The Cambridge Handbook of Age and Ageing. Cambridge University Press, Cambridge, pp. 572–582.

McDonald, A., 2010. The impact of the 2005 Mental Capacity Act on social workers' decision-making and approaches to the assessment of risk. British Journal of Social Work 40 (4), 1229–1246.

McInnis-Dittrich, K., 2006. Cognitive-behavioural interventions. In: Berkman, B. (Ed.), Handbook of Social Work in Health and Aging. Oxford University Press, Oxford, pp. 749–756.

Mellor, J.M., Lindeman, D., 1999. The role of the social worker in interdisciplinary geriatric teams. Journal of Gerontological Social Work 30 (3/4), 3–7.

Mullun, P., 2000. The Imaginary Time Bomb: Why an Aging Population Is Not a Social Problem. I.B. Tauris, New York.

Naleppa, M.J., 2006. Case management services. In: Berkman, B. (Ed.), Handbook of Social Work in Health and Aging. Oxford University Press, Oxford, pp. 521–527.

National Initiative for the Care of the Elderly, 2012. About NICE: What We Do. http://www.nicenet.ca/detail.aspx?menu=31&app=170&cat1=536&tp=2&lk=no (accessed 15.12.12.).

Nelson-Becker, H., Sullivan, M.P., 2012. Social care and spirituality. In: Cobb, M., Puchalski, C., Rumbold, B. (Eds.), The Textbook of Spirituality in Healthcare. Oxford University Press, Oxford, pp. 409–416.

Ozane, E., 2006. Older adults in Australia. In: Berkman, B. (Ed.), Handbook of Social Work in Health and Aging. Oxford University Press, Oxford, pp. 945–952.

Phillips, J., 2006. Older adults in Europe. In: Berkman, B. (Ed.), Handbook of Social Work in Health and Aging. Oxford University Press, Oxford, pp. 959–966.

Pomerleau, J., Knai, C., Kolte, E., 2008. The burden of chronic disease in Europe. In: Nolte, E., McKee, M. (Eds.), Caring for People with Chronic Conditions: A Health Systems Perspective. Open University Press, Maidenhead, Berkshire, pp. 15–42.

Poon, L.W., Jang, Y., Reynolds, S.G., McCarthy, E., 2005. Profiles of the oldest-old. In: Johnson, M.L. (Ed.), The Cambridge Handbook of Age and Ageing. Cambridge University Press, Cambridge, pp. 346–353.

Ray, M., Bernard, M., Phillips, J., 2009. Critical Issues in Social Work with Older People. Palgrave Macmillan, Baskingstoke, Hampshire.

Ray, M., Phillips, J., 2012. Social Work with Older People, fifth ed. Palgrave Macmillan, Basingstoke, Hampshire.

Richards, S., Sullivan, M.P., Tanner, D., Beech, C., Milne, A., Ray, M., Phillips, J., Lloyd, L., 2013. On the edge of a new frontier: is gerontological social work in the UK ready to meet 21st century challenges?. Advance Access published May 10, 2013. British Journal of Social Work. http://dx.doi.org/10.1093/bjsw/bct082.

Royal National Institute for the Blind, 2012. Older people. http://www.rnib.org.uk/livingwithsightloss/copingwithsightloss/pages/older_people.aspx (accessed 16.12.12.).

Silverstein, M., Giarrusso, R., 2010. Ageing and family life. Journal of Marriage and the Family 72, 1039–1058.

Soderberg, M., Stahl, A., Emilsson, U.M., 2012. Independence as a stigmatizing value for older people considering relocation to a residential home. European Journal of Social Work. http://dx.doi.org/10.1080/13691457.2012.685054.

Stuen, C., 2006. Older adults with age-related sensory loss. In: Berkman, B. (Ed.), Handbook of Social Work in Health and Aging. Cambridge University Press, Cambridge, pp. 65–77.

Sullivan, M.P., 2009. Social workers in community care practice: ideologies and interactions with older people. British Journal of Social Work 39, 1306–1325.

Sullivan, M.P., Kessler, L., LeClair, J.K., Stolee, P., Whitney, B., 2004. Defining best practices for specialty geriatric mental health outreach services: lessons for implementing mental health reform. Canadian Journal of Psychiatry 49 (7), 458–466.

Topo, P., 2011. Social protection of older in Finland from the eighteenth to the twenty-first centuries: messages for current policy and practice from an historical analysis. British Journal of Social Work 41 (5), 876–893.

United Nations Population Division, 2009. World Population Ageing. Department of Economic and Social Affairs, United Nations, New York.

United Nations Population Division, 2012. Changing Levels and Trends in Mortality: The Role of Patterns and Cause of Death. Department of Economic and Social Affairs, United Nations, New York.

Victor, C.R., 2005. The epidemiology of ageing. In: Johnson, M.L. (Ed.), The Cambridge Handbook of Age and Ageing. Cambridge University Press, Cambridge, pp. 95–105.

Victor, C.R., 2010. Ageing, Health and Care. The Policy Press, Bristol.

Walker, A., 2009. Commentary: the emergence and application of active aging in Europe. Journal of Aging and Social Policy 21 (1), 75–93.

World Health Organisation, 2008. The Global Burden of Disease Update 2004. World Health Organisation, Geneva.

Xu, Q., Chow, J.C., 2011. Exploring the community-based service delivery model: elderly care in China. International Social Work 54 (3), 374–387.

Yoon, E., Doherty, J.B., 2009. Arthritis pain. In: Cummings, S.M., Kropt, N.P. (Eds.), Handbook of Psychosocial Interventions with Older Adults. Routledge, New York, pp. 77–101.

Zhan, H.G., Guan, X., Bai, H., 2006. Recent developments in institutional elder care in China: changing concepts and attitudes. Journal of Ageing & Social Policy 18 (2), 85–108.

Elections: Apportionment

Bernard Grofman, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Apportionment refers to the determination of the number of representatives to be allocated to preexisting political or geographic units, while *districting* refers to how lines are drawn on a map within those units to demarcate the geographic boundaries of individual constituencies. *Malapportionment* refers to differences in the ratio of the number of voters/electors to the number of representatives across different constituencies. *Gerrymandering* refers to the drawing of districting lines for purposes of political (e.g., partisan or ideological or ethnic) advantage/disadvantage. Apportionment and districting practices are reviewed and key issues identified. While there is a particular focus on data from the United States, the general approach is a comparative one.

Most contemporary students of democratic theory take for granted that the basis of political representation will be geographic. There are two key components of any geographic system of representation: apportionment and districting. While the two terms are often used synonymously, formally, *apportionment* refers to the determination of the number of representatives to be allocated to preexisting political or geographic units, while *districting* refers to how lines are drawn on a map within those units to demarcate the geographic boundaries of individual constituencies. *Malapportionment* refers to differences in the ratio of the number of voters/electors to the number of representatives across different constituencies. *Gerrymandering* refers to the drawing of districting lines for purposes of political (e.g., partisan or ideological or ethnic) advantage/disadvantage.

Use of geographical districting leaves open many key questions: How many and how large are the districts to be? Will seats be allocated to whole political units, such as provinces or towns, or will district lines be permitted to cut across existing political subunit boundaries? Will district lines be required to satisfy standards of compactness or contiguity? To what extent will apportionment and districting lines be based (entirely or almost entirely) on total population? Or on population of (eligible) voters?

The United States has been a leader in defining standards of apportionment and districting to implement the principle of popular sovereignty. The U.S. House of Representatives was intended by its founders to be the representative chamber of a bicameral legislature and its apportionment rules were set up to require a purely population-based allocation of House districts to the states, with changes in seat allocations made after each decennial census. Indeed, the various apportionment methods that have been used for the House over the past several centuries are mathematically identical to proportional representation methods such as D'Hondt and Ste. Laguë (Balinski and Young, 1982; also *see* Electoral Systems). These methods differ in how they handle the problem of rounding.

In Baker vs Carr, 369 U.S. 186 (1962), the U.S. Supreme Court held that failure to redraw district lines when new census data were available was unconstitutional and that courts could fashion appropriate remedies. In subsequent landmark districting decisions, such as Reynolds vs Sims, 377 U.S. 533 (1964), the U.S. Supreme Court went much further, proclaiming *one person, one vote* as the only appropriate standard for both districting and apportionment. While one person, one vote notions of representation have had a profound influence throughout the world, by and large, the United States remains extreme among nations in its insistence on strict adherence to one person, one vote standards. For state legislative and local redistricting plans, where the one person, one vote standard is derived primarily from the equal protection clause of the Fourteenth Amendment, Supreme Court cases in the United States have established a 10% total deviation as prima facie evidence of constitutionality, though evidence that districting choices lack a rational justification for deviations from ideal can lead to a plan that is under the 10% total deviation being struck down. (*Total deviation* is the sum of the absolute values of the differences between actual district size and ideal district size of the largest district and the smallest districts, normalized by dividing through by ideal district size.) For Congressional districting, where standards are based directly on the supposed meaning of language in Article I of the Constitution, the Supreme Court has held that districts must be as close to zero deviation as is practicable. For example, in Karcher vs Daggett, 462 U.S. 725 (1983), a Congressional plan with a total deviation of only 0.698% was invalidated. In reaction to this ruling, in some 1990s Congressional plans, districts were drawn that were equal in population to within a handful of persons. In contrast, in other countries, especially those using plurality elections, no such strict population requirements exist. Many countries require (or even only just suggest) that differences should be no greater than ±25 or ±50% of ideal (Butler and Cain, 1992).

However, the notion that near-perfect equality of political representation has been achieved in the United States is misleading. The grossly malapportioned U.S. Senate tends to be omitted from international comparisons despite the fact that it is a coequal chamber. The U.S. House requires that each state has at least one representative – a rule that usually gives representation to several small states who would not otherwise be entitled to seats. Also, since states are the units, 'rounding rules' create variation in average House district population across states. For example, while the average size of a Congressional district based on the 2010 Census apportionment

 http://dx.doi.org/10.1016/B978-0-08-097086-8.93008-3

population was 710 767, the largest House district in the 2010 apportionment was the single district in Montana, with 994 416 people, nearly two times the size of the smallest House district, the single seat in Rhode Island, with 527 624 people. Thus, the total deviation in the entire House of Representatives came to 66% (283 649 in Montana plus 118 465 in Rhode Island). The population discrepancies have been comparable in recent apportionments and even greater in earlier apportionments.

Moreover, even districts that are equal in population need not be equal in terms of (eligible) voters, much less in terms of actual voters. Perhaps most importantly, unless we somehow regard voters as completely interchangeable units, neither population nor voter equality across constituencies, however perfect, guarantees equality of effective representation of the disparate groups and interests within a society. The degree and geographic locus of malapportionment and differential turnout across groups interact with how a group's voting strength is distributed across districts to affect the translation of a group's voting strength into actual electoral impact (Grofman et al., 1997). Indeed, malapportionment is sometimes referred to as a form of *silent gerrymander*, since malapportionment can easily translate into the political disadvantage of groups whose influence has been diminished because their members are disproportionally concentrated in constituencies whose voters have been underrepresented relative to their numbers. Even without conscious gerrymandering, the way in which districting lines are drawn will necessarily have an impact on the representation of different parties or groups (Dixon, 1968).

The term gerrymandering comes from wordplay on the last name of Elbridge Gerry, Governor of Massachusetts. In 1812, Gerry signed into law a districting plan for the Massachusetts Senate, allegedly designed to maximize the electoral successes of Republican-Democrat candidates and minimize the electoral successes of Federalist candidates, which included some rather strangely shaped districts. In a map in the Boston Gazette of 26 March 1812, the strangest of these districts was shown as a salamander, given tongue and teeth (**Figure 1**). Perhaps, the most pernicious aspect of this figure is that it has led to a potentially misleading association of gerrymanders with oddly shaped districts. The defining aspect of a gerrymander is the political consequences it entails, not its shape. Political disadvantage can come about even when districts look like squares or hexagons (Grofman, 1990). In fact, however, the 1812 Senate plan did achieve partisan advantage for the Republican-Democrats; in the next election they won 29 of 40 seats, even though they received less than half of the votes (Hardy, 1990).

Gerrymanders can be classified as partisan, bipartisan (often called 'incumbent gerrymanders'), racial, and personal, depending on who can be expected to be harmed or helped. In the United States, for example, the debate about gerrymandering has been fought largely over racial rather than partisan issues, e.g., over the extent to which plans should seek to place members of historically disadvantaged groups such as African-Americans into districts where they comprise the majority of the population, even if doing so meant drawing districts that were irregular in appearance or cut across municipal and other political unit boundaries (Grofman, 1998).

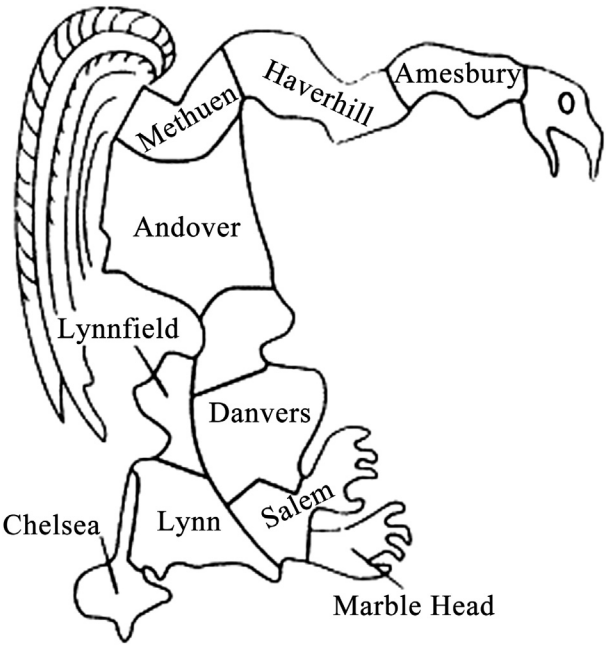

Figure 1 The infamous original gerrymander (as artistically rendered).

There are two basic techniques of gerrymandering: (1) 'packing' members of the group that is to be disfavored into districts that are won by very large majorities, thus 'wasting' many of that group's votes and (2) 'cracking' the voting strength of members of the group by dispersing the group's population across a number of districts in such a fashion that the group's preferred candidates will command a majority of the votes in as few districts as possible. In addition, if elections are held under plurality, a group's voting strength may be submerged in multimember districts that use bloc voting – a technique sometimes called 'stacking.' A fourth technique has been called 'kidmapping' (Grofman, 1985). Here district lines are moved so that the incumbent previously representing the district now finds his or her home in a different district, perhaps paired with an incumbent of his or her own party.

The terms 'affirmative action gerrymander' and 'benign gerrymander' have been used to denote districting done to advantage members of a historically disadvantaged group. However, it is important to distinguish between plans that are drawn with an aim to create a level-playing field by avoiding unnecessary fragmenting of minority population concentrations, but that otherwise generally take into account the usual districting criteria such as respect for natural geographic boundaries and historical communities of interest, and that seek to specially privilege particular groups by totally disregarding features other than race (or party) in drawing lines.

Because the way in which lines are drawn can be expected to matter, an important issue has to do with who draws the lines. In most democracies, especially those electing under plurality, nonpartisan boundary commissions are responsible for drawing district lines (Butler and Cain, 1992; Handley and Grofman, 2008). In the United States, the preponderant pattern is for a legislature to be responsible for its own redistricting and for each state legislature to be responsible for the

drawing of Congressional district lines for its state. However, in most U.S. legislatures, no plan can be passed without gubernatorial agreement. Because of divided party rule and other factors, states may be unable to reach agreement on plans, thus throwing decision-making into the courts.

One way in which districting practices in the United States are distinct from those in other countries is the extent to which courts play a critical role as arbiter. In recent decades, all but a handful of states have had a legislative or Congressional plan challenged in court, and many plans have been rejected – in the 1960s and 1970s, mostly for reasons having to do with population inequalities across districts; in the 1980s and 1990s, for reasons having to do with racial representation (Grofman, 1998). Indeed, throughout these decades, courts themselves were responsible for drawing some of the legislative or Congressional districting plans that were actually used. Another peculiarity of U.S. districting practices is the role of the U.S. Department of Justice under Sections 2 and 5 of the Voting Rights Act of 1965 as amended in 1982 (Grofman, 1998) in enforcing racial fairness.

From a comparative perspective, we may say that, generally speaking, gerrymandering is more important in plurality elections than in elections under proportional or semiproportional rules. In particular, when there are more than two candidates or political parties competing, districting can have a dramatic impact on outcomes in plurality elections (Taylor et al., 1986). *Ceteris paribus*, for elections under proportional or semiproportional methods, the larger the average *district magnitude* (the number of representatives to be elected from the constituency), the lesser the probable impact of districting choices on outcomes; in contrast, for elections under plurality voting, the greater the average *district magnitude*, the greater is the expected impact on outcomes, since *plurality bloc voting* (the extreme case of which is an *at-large election*) can result in the virtual submergence of the views of those in the minority. However, even under proportional representation, expected outcomes can still be manipulated by districting choices, especially choices as to district magnitude (Mair, 1986).

See also: Electoral Geography; Electoral Systems; Political Geography; Political Parties, History of; Political Representation; Proportional Representation.

Bibliography

Balinski, M.L., Peyton Young, H., 1982. Fair Representation: Meeting the Ideal of One Man. One Vote. Yale University Press, New Haven, CT.

Butler, D., Cain, B., 1992. Congressional Redistricting. Macmillan, New York.

Dixon, R.G., 1968. Democratic Representation: Reapportionment in Law and Politics. Oxford University Press, New York.

Grofman, B., 1985. Criteria for districting: a social science perspective. UCLA Law Review 33 (1), 77–184.

Grofman, B., 1990. Toward a coherent theory of gerrymandering: Bandemer and Thornburg. In: Grofman, B. (Ed.), Political Gerrymandering and the Courts. Agathon, New York, pp. 29–63.

Grofman, B. (Ed.), 1998. Race and Redistricting. Agathon, New York.

Grofman, B., Koetzle, W., Brunell, T., 1997. An integrated perspective on the three potential sources of partisan bias: malapportionment, turnout differences, and the geographic distribution of party vote shares. Electoral Studies 16 (4), 457–470.

Handley, L., Grofman, B. (Eds.), 2008. Redistricting in Comparative Perspective. Oxford University Press, Oxford.

Hardy, L., 1990. The Gerrymander: Origin, Conception and Re-emergence. Claremont McKenna College, Rose Institute of State and Local Government, Claremont, CA.

Mair, P., 1986. Districting choices under the single-transferable vote. In: Grofman, G., Lijphart, A. (Eds.), Electoral Laws and Their Political Consequences. Agathon, New York, pp. 289–308.

Taylor, P., Gudgin, G., Johnston, R.J., 1986. The geography of representation: a review of recent findings. In: Grofman, B., Lijphart, A. (Eds.), Electoral Laws and Their Political Consequences. Agathon, New York, pp. 183–192.

Electoral Geography

Ron Johnston, University of Bristol, Bristol, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Electoral geography provides a spatial perspective to the study of the organization, conduct, and outcome of elections. Its main focus is on the nature of, and interactions between, three maps. Two refer to the 'inputs' – the geography of support for parties and/or candidates (themselves in many cases a function of the geography of the social groups from which each party draws its main support); and the geography of the constituencies (electoral districts, ridings, etc.) within which votes are cast and counted. The third map shows the outputs, the geography of representation that emerges when the votes are counted and the winners and losers determined. Much research in electoral geography involves the analysis of two types of data. The first is ecological, or aggregate, data; the second is data on individual electors – traditionally the products of voter surveys. Using evermore sophisticated analytical procedures, the two are increasingly combined, enabling the study of individual behavior patterns in their local contexts.

Electoral geography provides a spatial perspective to the study of the organization, conduct, and outcome of elections (Johnston and Pattie, 2012b; discuss its origins and development). No general synthesizing text has appeared since 1979 (Taylor and Johnston, 1979) – apart from one that focuses on British elections only (Johnston and Pattie, 2006) – although a collection of essays has assessed recent developments (Warf and Lieb, 2011); the journal *Political Geography* publishes a number of papers on electoral geography each year.

The main focus of electoral geography is the nature of, and interactions between, three maps (of which, given advances in computer cartography, there is an increasing number of excellent examples – as in Brunn et al., 2011). Two refer to the 'inputs' – the geography of support for parties and/or candidates (themselves in many cases a function of the geography of the social groups from which each party draws its main support); and the geography of the constituencies (electoral districts, ridings, etc.) within which votes are cast. The third map shows the outputs, the geography of representation that emerges when the votes are counted and the winners and losers determined. (Gudgin and Taylor's (1979) book is a classic treatment of these three maps: it was republished, with an extensive introduction, in 2012.) Although most electoral geographers focus on these maps (and others linked to them, as discussed below), there are considerable variations in the relative attention given to each (as Johnston, 2005, illustrates in a US/UK comparison).

Much research in electoral geography involves the analysis of two types of data. The first is ecological, or aggregate, data – the number of votes cast for each candidate in an area, for example; the second is data on individual electors – traditionally the products of voter surveys. Using evermore sophisticated analytical procedures, the two are increasingly combined, enabling the study of individual behavior patterns in their local contexts.

The Geography of Votes

André Siegfried's pioneering essays on *géographie électorale* established cartographic correlations between patterns of party support and French physical, social, and economic landscapes, sophisticatedly linking environment to voting through agrarian systems, land ownership, and rural social stratification and tradition: parties drew their support from different sociospatial milieux (e.g., Siegfried, 1949; Dogan, 1967; Lacoste, 1986, follow in that tradition.)

Siegfried's work showed substantial continuity over time in these patterns of support, as did a further seminal work. Key (1955) demonstrated that continuity in American voting maps reflected stable electoral cleavages, with parties sustained within the same milieux across a sequence of elections. That continuity was occasionally broken, however. An interruption might be only temporary – producing a geography that differed from the 'normal pattern' when substantial numbers of voters deviated from their usual partisan leanings reflecting the particular circumstances of a contest. On other occasions the change was more permanent and a new geography established that was then sustained over a sequence of elections; this reflected realignment in the socioeconomic and sociodemographic support for one or more parties and was represented in the geography of their voting strength. Archer and Taylor (1981) used this framework of normal, deviating, and realigning elections in analyses of the changing electoral mosaics of support for American presidential candidates (Archer and Shelley, 1986).

For the majority of Western European countries, Lipset and Rokkan (1967) identified four electoral cleavages that underpinned support for most of their political parties. Two of those cleavages were initiated following the 'national revolution' and the emergence of nation-states from the seventeenth century on: the first was between the dominant and minority cultures within a state's territory (the core-periphery cleavage); and the second reflected the contest between church and secular interest groups (the church-state cleavage). The other two were products of the 'industrial revolution': the first (the class cleavage) reflected conflicts between employers and workers and the second those between an economy's primary and secondary sectors (the urban–rural cleavage). Of these four, the class cleavage was present in most European countries by the mid-twentieth

century; the relative strength of the others reflected local histories and cultures. The importance of those cleavages has however declined recently – more so in some countries than others. Electoral decisions are less driven by ideological differences – between classes, for example – and more by what are termed 'valence politics' (Clarke et al., 2009): voters, from whatever backgrounds, make their decisions on which party they think can best manage the country on the most important issues of the day, which are usually dominated by the economy and national prosperity.

If cleavages are the main sources of variations in electoral support within a country, then if you know a party's social bases of support the geography of voting for it should be readily 'predicted': if a party gets most of its support from the 'working class,' it should be strongest where members of that class are most numerous. This suggests a uniformity to a country's electoral geography, as illustrated in a classic study of voting in Britain (Butler and Stokes, 1969, 1974). Similarly, if valence politics dominate electoral choice, a party that gets its strongest support from those who are positive about the future economically should attract most votes in areas where optimism prevails.

In both cases, the geography of voting would be relatively straightforward to uncover – if you know a country's economic or social geography, you can readily 'forecast' its electoral geography. But that is rarely completely the case, as Johnston et al. (1998) showed in an analysis of three UK general elections: each political party performed better than expected in some areas and worse in others. In part this reflected the country's uneven geography of prosperity (the 'north–south divide'), but work by Agnew (1987, 1996), building on Cox's (1970) study of the geography of party support in Wales, provided a stronger – inherently geographical – argument for the existence of electoral geographies that are more complex than a simple mapping of national cleavages. Agnew's argument comprised six claims listed below:

1. Class and community affiliations take on different meanings in particular places (as illustrated in Johnston's (1991) essays on British coalfields and the American south).
2. Local communication networks, through which political affiliations are mobilized, may be relatively isolated in particular territories, leading to separate voting patterns.
3. There are continuing tensions between some areas within a state and the national political institutions, engendering local political cultures (as with intrastate secessionist movements).
4. All social cleavages are locally constituted – there is no national class cleavage, for example, only a myriad local ones – which may lead to local variations in electoral support among people with similar backgrounds.
5. Some political parties' histories are rooted in particular places, which become their main areas of electoral support and whose interests they promote in their manifestos.
6. The microgeography of everyday settings – home, neighborhood, work, leisure, etc. – can stimulate local distinctiveness that is reflected in voting patterns.

His own work on Italy (Agnew, 2002; Agnew and Shin, 2008) illustrates these themes.

Local Influences on Voting Decisions

Agnew's sixth claim is the basis for much research that has explored the extent of, and reasons for, variations in voting patterns at local scales that cannot be accounted for by general theories such as those of cleavages and valence politics. Some of the earliest work on this theme derived from Key's (1949) book on elections where party cues are irrelevant. The Democratic Party dominated politics in the American South for nearly a century, where voters' main choices were among the contenders for its nomination. Many candidates performed best in their home areas – creating a 'friends and neighbors effect,' formalized by Reynolds (1969) and Cox (1969).

Cox's (1969) pioneering essay – which in many ways 'launched' the subdiscipline of quantitative electoral geography (Johnston and Pattie, 2012b) – extended the argument underpinning the 'friends and neighbors effects' to the more general concept of a 'neighborhood effect' whereby an area's majority opinion is accentuated if some of their residents who initially favored minority views are won over through a process of 'conversion by conversation.' The thesis that 'people who talk together vote together' was established using survey data (e.g., Huckfeldt and Sprague, 1995; Pattie and Johnston, 2000). Establishing that such conversations underpin observed voting patterns has not been straightforward, however. Many analyses of putative neighborhood effects have relied on aggregate data: their findings are consistent with the argument that a party tends to perform better than expected in areas where its supporters are in the majority, for example, but cannot uncover the processes involved. Research combining survey and aggregate data has gone some way to resolving this problem, however, as in studies using 'bespoke neighborhoods' – small areas separately defined as the immediate milieux to each respondent's home address – to represent the local context (e.g., MacAllister et al., 2001). These have shown that the more that people are involved in their locale, the more open they are likely to be to neighborhood influences (Johnston et al., 2005), while analyses using contextual data at a variety of scales have suggested that variations reflecting neighborhood influences are strongest at more local scales (Jones et al., 2007).

Another set of local influences on voting behavior involves the activities of parties in their search for votes – the local campaign maps. Alongside national election campaigns – which focus through the national media on such issues as policy manifestos, leadership potential, and government performance – are locally organized and focused campaigns. These have three main goals: to build a candidate's profile locally; to identify voters who are likely to support the party/candidate; and then to ensure that those supporters turn out to vote. Early studies of such activities in the United Kingdom used data on the amounts spent on such campaigns as an index of their intensity, and showed that the more that candidates spent (especially challengers rather than incumbents) the better their performance (Johnston, 1987). These findings have been complemented by analyses using other indices of campaign intensity (Denver and Hands, 1997), and confirmed by studies using survey data which show that those contacted by a party (especially those who are not firm in their voting intention: Johnston and Pattie, 2012b) are

more likely to vote for it than those who are not contacted. Since parties tend to concentrate such campaigning not only in the relatively marginal constituencies where either victory or defeat could be the outcome but also in parts of those constituencies where their likely supporters' homes are concentrated (Johnston et al., 2005a), such successful geographically targeted campaigns enhance the processes producing neighborhood effects. US presidential elections are similarly highly focused on the small number of 'swing states' where both candidates have a chance of victory, with the others getting little personal attention from them because the outcome there is taken for granted (Brunn et al., 2011: 29–42).

The Geography of Constituency Definition

The electoral system in most countries involves the use of territorially defined constituencies; the contest for votes is split into a large number of separate areas (Farrell, 2011). Those districts vary greatly in their size depending on the electoral system deployed: in some countries, notably those using single-member constituency systems, the details of their definition can crucially impact on an election outcome because the overall result reflects not just the total number of votes cast but also in how many separate districts a party (or a candidate in a presidential election) wins the largest number of votes.

This issue of where to draw district boundaries has attracted much attention among American electoral geographers, because redistricting there – of Congressional Districts, for example – has traditionally been undertaken by the political party in power in the relevant state, which has sought boundary configurations that promote its electoral interests and penalize its opponents' (Morrill, 1973; Winburn, 2009; Bullock, 2010). Two main 'cartographic abuses' have been identified. The first is malapportionment, whereby constituencies are either created with substantial variations in their population or such variations are allowed to develop with differential population changes (for example, rural to urban and central city to suburban migrations). A party whose electoral strength is in areas where on average constituencies are small will tend to perform better – i.e., win more seats relative to its vote share in the relevant legislature – than one whose strength is in areas where they are larger. Malapportionment was outlawed in the 1960s 'redistricting revolution' and American courts now require that all districts have equal populations. But the second 'cartographic abuse' has been much more difficult to remove. Gerrymandering involves the drawing of constituency boundaries to benefit one party rather than another, and can be practiced even where equality of electorates is the paramount consideration. Much gerrymandering involves the drawing of odd-shaped constituencies, so as to create as many constituencies as possible where the party involved should obtain a majority of the votes (as illustrated by Monmonier, 2002).

Although gerrymandering is a deliberate strategy to promote partisan ends, Gudgin and Taylor (1979) showed that even where the boundaries are drawn by strictly neutral, nonpartisan bodies, nevertheless the equivalent of gerrymandering can be the outcome. They showed – through sophisticated mathematical and statistical reasoning – that, for a given geography of party support, different constituency configurations for an area could result in very different electoral outcomes. Extending their work in the United Kingdom, for example, Johnston and Rossiter (1983) showed that in Sheffield, a city with six constituencies, although the most likely outcome was Labour winning five seats and the Conservatives one, some configurations would give Labour six seats whereas others would see it win only four. Disproportional outcomes – with one party getting a larger share of the seats than of the votes – could thus eventuate even where neither malapportionment nor gerrymandering was practiced.

Further work on this issue in the United Kingdom has identified not only disproportionality in election results as the product of nonpartisan districting but also bias – defined as differential treatment of the parties even when they have the same share of the votes. At the 1997 general election, for example, if – under certain assumptions – Labour and the Conservatives had won an equal share of the votes, Labour could have gained 82 more seats than its opponent (out of 652). Four years later, Labour's advantage in a similar situation could have been 142 seats; in 2005 it could have been 112; and in 2010 – when the Conservatives won 7% points more of the votes – if the two parties had been equal Labour could have won 54 more seats. Johnston et al. (2001; Johnston and Pattie, 2009, 2011a) used a mathematical procedure to decompose this bias, finding that part of it was due to malapportionment, part to an equivalent of unintentional gerrymandering (Labour's votes were more efficiently distributed across the constituencies than were the Conservatives'), part to variations in turnout (there tend to be more abstentions in Labour-held seats), and part to the differential impact of smaller parties. That methodology, which focuses on the two largest parties only (and has been applied to the US Electoral College: Johnston et al., 2005b), has been extended for the analysis of three-party systems (Thrasher et al., 2011; Borisyuk et al., 2010). As a consequence of that bias, the British government changed the rules for redistributions in 2011 to remove any malapportionment effects (Johnston and Pattie, 2012a; Rossiter et al., 2012).

Conclusions

Elections are based on geographies (the maps of party support and the territories within which votes are cast and counted); their conduct involves geographical processes (such as the campaigns conducted by political parties and the informal processes within neighborhoods through which interpersonal influence on how people vote is transmitted); and their outcomes are geographies – a party in government may have much more support in some areas than others, which may well be reflected in its policies, in who gets what, where, in the allocation of public goods and services. These features are present in some form, with a small number of exceptions only, whatever electoral system is used (Johnston and Pattie, 2011b). It is impossible to take the geography out of elections; it is fundamental to the operation of representative democracies.

See also: Campaigning: Political; Data Bases and Statistical Systems: Political Behavior and Elections; Elections: Apportionment; Electoral Systems; Political Geography; Spatial Analysis in Geography; Spatial Data; Territoriality: Geographical; Thematic Maps in Geography.

Bibliography

Agnew, J.A., 1987. Place and Politics: The Geographic Mediation of State and Society. Allen & Unwin, Boston.

Agnew, J.A., 1996. Mapping politics: how context counts in electoral geography. Political Geography 15, 129–146.

Agnew, J.A., 2002. Place and Politics in Modern Italy. University of Chicago Press, Chicago.

Agnew, J.A., Shin, M.E., 2008. Berlusconi's Italy: Mapping the Second Republic. Temple University Press, Philadelphia.

Archer, J.C., Shelley, F.M., 1986. American Electoral Mosaics. Association of American Geographers, Washington, DC.

Archer, J.C., Taylor, P.J., 1981. Section and Party: A Geography of American Presidential Elections from Andrew Jackson to Ronald Reagan. Research Studies Press, Chichester.

Borisyuk, G., Johnston, R.J., Thrasher, M., Rallings, C., 2010. A method for measuring and decomposing electoral bias for the three-party case, illustrated by the British case. Electoral Studies 29, 733–745.

Brunn, S.D., Webster, G.R., Morrill, R.L., Shelley, F.M., Lavin, S.J., Archer, J.C. (Eds.), 2011. Atlas of the 2008 Elections. Lanham MD: Rowman & Littlefield.

Bullock III, C.S., 2010. Redistricting: The Most Political Activity in America. Rowman & Littlefield, Lanham, MD.

Butler, D.E., Stokes, D., 1969. Political Change in Britain: Forces Shaping Electoral Choice. Macmillan, London.

Butler, D.E., Stokes, D., 1974. Political Change in Britain: The Evolution of Electoral Choice. Macmillan, London.

Clarke, H.D., Sanders, D., Stewart, M.C., Whiteley, P., 2009. Performance Politics and the British Voter. Cambridge University Press, Cambridge.

Cox, K.R., 1969. The voting decision in a spatial context. Progress in Human Geography 1, 81–117.

Cox, K.R., 1970. Geography, social contexts and voting behaviour in Wales, 1861–1951. In: Allardt, E., Rokkan, S.E. (Eds.), Mass Politics. The Free Press, New York, pp. 117–159.

Denver, D.T., Hands, G., 1997. Modern Constituency Electioneering: Local Campaigning at the 1992 General Election. Frank Cass, London.

Dogan, M., 1967. Political cleavage and social stratification in France and Italy. In: Lipset, S.M., Rokkan, S.E. (Eds.), Party Systems and Voter Alignment. The Free Press, New York, pp. 129–195.

Farrell, D.M., 2011. Electoral Systems: A Comparative Introduction. Palgrave-Macmillan, Basingstoke.

Gudgin, G., Taylor, P.J., 1979. Seats, Votes and the Spatial Organization of Elections. Pion, London (Republished, with an introduction by Ron Johnston – ECPR Press, Colchester, 2012).

Huckfeldt, R., Sprague, J., 1995. Citizens, Politics and Social Communication: Information and Influence in an Election Campaign. Cambridge University Press, Cambridge.

Johnston, R.J., 1987. Money and Votes: Constituency Campaign Spending and Election Results. Croom Helm, London.

Johnston, R.J., 1991. A Question of Place? Exploring the Practice of Human Geography. Blackwell, Oxford.

Johnston, R.J., 2005. Anglo-American electoral geography: same roots and same goals, but different means and ends? The Professional Geographer 57, 580–587.

Johnston, R.J., Pattie, C.J., 2006. Putting Voters in Their Place: Geography and Elections in Great Britain. Oxford University Press, Oxford.

Johnston, R.J., Pattie, C.J., 2009. Geography: the key to recent British elections. Geography Compass 3, 1865–1880.

Johnston, R.J., Pattie, C.J., 2011a. The British general election of 2010: a three-party contest or three two-party contests? The Geographical Journal 177, 17–26.

Johnston, R.J., Pattie, C.J., 2011b. Electoral systems, geography, and political behaviour: United Kingdom examples. In: Warf, B., Lieb, J. (Eds.), Revitalizing Electoral Geography. Ashgate, Farnham, pp. 31–58.

Johnston, R.J., Pattie, C.J., 2012a. From the organic to the arithmetic: new redistricting/redistribution rules for the United Kingdom. Election Law Journal 11, 70–89.

Johnston, R.J., Pattie, C.J., 2012b. Kevin Cox and electoral geography. In: Jonas, A.E.G., Wood, A. (Eds.), Territory, the State and Urban Politics. Ashgate, Farnham, pp. 23–44.

Johnston, R.J., Pattie, C.J., Allsopp, G., 1998. A Nation Dividing? The Electoral Map of Great Britain, 1979–1987. Longman, Harlow.

Johnston, R.J., Pattie, C.J., Dorling, D., Rossiter, D.J., 2001. From Votes to Seats: The Operation of the British Electoral System since 1945. Manchester University Press, Manchester.

Johnston, R.J., Propper, C., Sarker, R., Jones, K., Bolster, A., Burgess, S., 2005a. Neighbourhood social capital and neighbourhood effects. Environment and Planning A 37, 1443–1461.

Johnston, R.J., Rossiter, D.J., Pattie, C.J., 2005b. Disproportionality and bias in US Presidential elections: how geography helped Bush defeat Gore but couldn't help Kerry beat Bush'. Political Geography 24, 952–968.

Johnston, R.J., Rossiter, D.J., 1983. Constituency building, political representation and electoral bias in urban England. In: Herbert, D.T., Johnston, R.J. (Eds.), Geography and the Urban Environment, vol. 3. John Wiley, Chichester, pp. 113–156.

Jones, K., Johnston, R.J., Propper, C., Burgess, S.M., 2007. Region, local context, and voting at the 1997 general election in England. American Journal of Political Science 51, 640–654.

Key, V.O., 1949. Southern Politics in State and Nation. A.A. Knopf, New York.

Key, V.O., 1955. A theory of critical elections. Journal of Politics 17, 3–18.

Lacoste, Y., 1986. Géopolitiques des Régions Françaises, 3 vols. Fayard, Paris.

Lipset, S.M., Rokkan, S.E., 1967. Cleavage structures, party systems and voter alignments: an introduction. In: Lipset, S.M., Rokkan, S.E. (Eds.), Party Systems and Voter Alignments. The Free Press, New York, pp. 3–64.

MacAllister, I., Johnston, R.J., Pattie, C.J., Tunstall, H., Dorling, D., Rossiter, D.J., 2001. Class dealignment and the neighbourhood effect: Miller revisited. British Journal of Political Science 31, 41–59.

Monmonier, M.S., 2002. Bushmanders and Bullwinkles: How Politicians Manipulate Electronic Maps and Census Data to Win Elections. University of Chicago Press, Chicago.

Morrill, R.L., 1973. Ideal and reality in reapportionment. Annals of the Association of American Geographers 63, 548–556.

Pattie, C.J., Johnston, R.J., 2000. 'People who talk together vote together': an exploration of the neighbourhood effect in Great Britain. Annals of the Association of American Geographers 90, 41–66.

Pattie, C.J., Johnston, R.J., 2012. Talking to the converted or reaching out to the uncommitted? Who do political campaigns influence? International Journal of Political Marketing 11, 1–34.

Reynolds, D.R., 1969. A spatial model for analyzing electoral behaviour. Acta Sociologica 12, 122–130.

Rossiter, D.J., Johnston, R.J., Pattie, C.J., 2012. Representing people and representing places: community, continuity and the current redistribution of constituencies in the UK. Parliamentary Affairs 66, 856–886.

Siegfried, A., 1949. Géographie Électorale de l'Ardèche sous la IIIème République. Armand Colin, Paris.

Taylor, P.J., Johnston, R.J., 1979. Geography of Elections. Penguin Books, London.

Thrasher, M., Borisyuk, G., Rallings, C., Johnston, R.J., 2011. Electoral bias at the 2010 general election: evaluating its extent in a three-party system. Journal of Elections, Public Opinion and Parties 21, 279–294.

Warf, B., Lieb, J. (Eds.), 2011. Revitalizing Electoral Geography. Ashgate, Farnham.

Winburn, J., 2009. The Realities of Redistricting. Rowman & Littlefield, Lanham, MD.

Electoral Systems

Alan Ware, Oxford University, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 7, pp. 4378–4382, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

An electoral system is a set of rules governing an election, and many scholars distinguish between three main elements of any system: ballot structure (how and 'for what' a vote is cast); constituency structure (whether, and how, the electorate is divided into territorially defined constitutencies), and the electoral formula (in an assembly election this is the method of translating votes into seats). The main types of electoral formulae are majoritarian, plurality, and proportional. The most common way of classifying electoral systems is to distinguish between proportional (PR) and nonproportional systems. The two main variants of the former are party-list systems and the single transferable vote (STV); the three main variants of the latter are plurality, majoritarian, and (so-called) semiproportional systems. Much debate has focused on the effects of electoral systems, and on 'Duverger's law.' However, most political scientists now argue that the electoral system has only a limited effect on a regime's party system, and hence on its potential for instability. However, the timing of the adoption of PR systems does have an effect, with its early introduction increasing the likelihood of fragmentation in the party system. Moreover, at particular junctures, changes in the electoral system can affect which parties are likely to govern a regime.

An electoral system is a set of rules governing an election. The study of electoral systems has tended to be restricted in a number of respects. First, most research focuses on elections to legislative assemblies and to political executives, even though, of course, many nongovernmental agencies also hold elections as the means of filling various offices. Second, many issues of central concern in the study of electoral systems – such as paradoxes emanating from the application of particular voting rules – are usually studied within the 'theory of voting' branch of rational choice analysis; in some respects, therefore the study of electoral system can be understood as more of an 'applied' social science, while the 'theory of voting' is 'pure' theory.

Elements of Electoral Systems

Following Rae (1971), many scholars distinguish between three main elements of an electoral system: ballot structure, constituency structure, and the electoral formula.

Ballot Structure

There are several different ways in which this can vary, including:

1. The number of votes each voter is permitted. In some systems the voter casts as many votes as there are positions to be filled (e.g., plurality systems with multimember constituencies – as employed in many state legislatures in the USA before the 1960s); in others, voters may cast only one vote irrespective of how many positions are being filled (e.g., the Single Non-Transferable Vote system used in Japan before 1994).
2. The kind of information the voter can reveal in voting. Some systems permit two different kinds of vote to be cast; for example, in the German Additional Member System (AMS) the voter votes for a party candidate in a constituency, as well as for a party for the party list component of the ballot. In many versions of party list systems the voter can reveal both which party they prefer, and which of that party's candidates they prefer. Again, by combining two separate elections, a system may make it possible for voters to reveal extra information about their preferences. For example, in Israel since 1996, voters reveal not merely which party they prefer, but also which party they wish to form the government: the Israeli voter can vote both for a party in the election for the legislature, and for a particular candidate to be prime Minister following that election. Furthermore, in some systems, voters can reveal only their first preference (whether for a party or a candidate), while in others they can provide a rank ordering of the different parties (or candidates); the plurality system used in British parliamentary elections is an example of the former, while the Single Transferable Vote (STV) employed in the Republic of Ireland is an example of the latter. Finally, some electoral systems limit the expression of voter preference to those candidates or parties listed on the ballot, while others allow the 'writing in' of names not on the ballot, or allow the voter to register opposition to all those listed on the ballot by voting for 'none of the listed candidates.'
3. Whether the ballot is open or secret. In the twentieth century it has become the norm for national elections to be contested under a secret ballot, but until the late nineteenth century balloting was a public act. In the USA, parties in the nineteenth century printed their own ballot papers, and in several states the parties used different colors of paper for the ballot so they could tell who had voted for them. Whether the ballot should be secret or not was a major controversy in nineteenth-century political debates, and some of those who advocated an extensive franchise (e.g., John Stuart Mill) still favored an open ballot.

Constituency Structure

There are three main ways in which electoral systems vary in this regard.

a. The election can be conducted on an 'at large' basis, or an election can be held in each of several constituencies. Generally, though with some exceptions, 'at large' elections are held only in smaller territories and when a proportional mechanism is used to allocate seats in an assembly (e.g., The Netherlands). Most national and subnational assembly elections are divided into constituencies.
b. Whether the constituencies are organized on a territorial or some other basis. Although there have been debates about the desirability of a functional, rather than a territorial, basis for representation, most electoral systems use the latter. The odd exceptions include the 'university seats' of the British Parliament that were abolished before the 1950 election. Any territorial subdivision of a polity involves the drawing of boundaries for the constituencies, and how the boundaries are drawn may affect the overall election result. This issue becomes most significant in systems using non-proportional means for allocating seats, but it is still an issue in those using proportional mechanisms.
c. Whether the constituencies are single-member or multi-member. Systems that employ proportional means for allocating seats in an assembly have to use multimember systems, as does the SNTV; other systems do not. Generally, polities using the plurality system for allocating seats have moved towards the exclusive use of single-member constituencies during the twentieth century.

Electoral Formula

This is the way in which votes are connected to the office(s) being contested – in an assembly it is those rules governing how votes are translated into seats. There are three main kinds of formulae:

1. *Majoritarian.* Under these systems, the winning candidate or party must obtain more than 50 percent of the vote. The two most commonly used means for achieving this are the Alternative Vote System (AV), and the double Ballot System (DB). The AV, which is used in Australia, involves voters showing their second, and subsequent, preferences on the ballot form; when no candidate obtains more than 50 percent of the vote, the lowest-placed candidate is eliminated and his/her voters have their second-choice votes counted. This process continues until one candidate attains more than 50 percent of the vote. The DB requires a second election to be held in all constituencies in which no candidate obtains more than 50 percent of the vote; this second contest includes only the two leading candidates from the first election.
2. *Plurality.* Here the candidate(s) with the largest number of votes wins. In relation simply to its electoral formula, the SNTV system is merely a plurality system; however, its very different ballot structure – voters having only one vote each in multimember constituencies – mean that it is a radically different system in its operation.
3. *Proportionality.* This formula attempts to provide a 'fit' between the share of the votes that the candidates or parties obtain, and the share of the seats the parties get in the assembly. There are two main devices for achieving this. One is to use a formula, of which there are a number of variants including d'Hondt and St. Lagüe, that distributes seats to parties on the basis of their share of the vote. The other method is to permit voters to rank-order all candidates, and to redistribute the 'wasted votes' of those candidates who have more votes than they need to get elected; that, broadly speaking, is the basis of STV. Disproportionalities in supposedly proportional systems can occur for three main reasons:

   1. the ways in which electoral boundaries are drawn, if the system is divided into constituencies, thereby giving an advantage to particular parties;
   2. relatively small district magnitude – that is, a small number of seats being contested in each constituency; the fewer the seats in each constituency, the more likely it is that, in aggregate the allocation of seats between parties will diverge from the distribution of total votes. (STV has to be employed in conjunction with relatively small district magnitude; it becomes unwieldy with large district magnitude because of the complexities involved in expressing preferences between a large number of candidates.); and
   3. in party-list systems, through the use of thresholds which require that parties failing to secure a minimum share of the vote will fail to win any seats at all. A threshold of between 2 and 3 percent of the total vote usually reduces the number of parties represented in a legislature quite dramatically, while a threshold of 5 percent (as used in Germany) restricts seriously the number of parties represented.

Classification of Electoral Systems

There are a very large number of possible electoral systems, including many hybrids of the most common types. (An example of such a hybrid is the system used for the French National Assembly, which is a variant of the DB; a second ballot is used, but it is not confined to the two leading candidates from the first ballot. Instead, any candidate with more than 12.5 percent of the vote is permitted, but not required, to take part in the second ballot.) Given this large variety, considerable attention has been given to classifying the main types of system. It is widely agreed that the main distinction is between 'proportional' systems and other types, but there is some disagreement on two matters.

The first is whether the principal division is a bifurcation between 'proportional' and 'non-proportional' systems, or whether a more complex main division is appropriate. The second disagreement concerns how to treat a variety of electoral systems that are sometimes grouped together under the heading 'semiproportional'; these include the SNTV and the Cumulative Vote (CV). (The latter is a system used in Illinois for many years that, in a three-member constituency would give three votes to each voter but would allow the voter to give all three votes to one candidates, give two votes to one candidate and one to another, or to give a single vote to each of three candidates.) Sartori (1994), for example, argues in favour of treating the SNTV as a form of proportional representation (PR), but most scholars do not do so.

The most common classification is to distinguish between those systems that have mechanisms for generating proportional results that are internal to the systems themselves, and which do not depend for their proportionality on the behavior or strategies of voters or candidates, from those that do not.

Proportional Electoral Systems

Within the class of proportional systems there are two main variants – the *party-list* and the *STV* systems, and within the party-list subclass there are four main sub-subclasses:

Party-List

1. *Pure party-list.* In these systems the order of the party's candidates on the list, and hence their likelihood of being elected, is determined exclusively by the party. The voter can vote only for a party and not for a candidate. (Such systems are relatively uncommon, but this system is used in Israel and was used in France in 1945–96, and in a slightly less pure form during the rest of the fourth Republic.)
2. *Voter-choice from party-list.* These systems allow the voter to express a choice between candidates on the party-list, though great variations in the procedures are evident. In some systems it is actually difficult for voters to propel a popular candidate from a low position on the list to a higher one; in other systems it is much easier to do so. These types of system are the most common form of the party-list, having been used in many European countries.
3. *Mixed systems.* The AMS is the principal variant of this type. A proportion of the seats (sometimes one half) are contested in single-member constituencies under plurality voting. The remaining seats are contested using a party-list, and the allocation of seats from the party-list is used so as to correct for any disproportionalities resulting from the elections in the constituencies. This type of system, which is used in Germany and New Zealand, produces election results that are rather similar to those that would occur under other party list systems.
4. *Panachage.* Used in Switzerland, this system permits voter choice between candidates of more than one party. Some experts argue that it is not really a form of party-list system because of this rule, although its operation more closely resembles such systems than any other types.

STV

Unlike party-list systems, STV achieves proportionality in the distribution of seats in a legislature through a mechanism that allocates seats to particular candidates. A party acquires seats through given candidates of theirs receiving sufficient votes to win seats, rather than the parties being allocated votes through a formula that distributes seats directly to the parties.

Nonproportional Electoral Systems

Among the nonproportional systems, there are three main subclasses:

1. *Plurality.* The first-past-the-post system has been used in many Anglo-Saxon democracies, and it is still the system used in Canada, and most of the USA, as well as for elections to the UK Parliament.
2. *Majoritarian.* The most common variants of systems requiring the winning candidate (or party) to win more than 50 percent of the vote are the AV and DB. Outside national elections, there are some majoritarian systems that require the winning candidate to receive more than 50 percent of the vote – one example being the Democratic National Convention which, until the 1930s, required that the nominee for the presidency receive two-thirds of the delegates' votes. Since such systems can produce many rounds of balloting and still fail to produce a result, they are wholly unsuitable for elections at which electors cannot interact personally with each other in an effort to achieve a definite result.
3. *Semi-proportional.* This is a subclass containing several rather different kinds of system, including the CV, SNTV and the Limited Vote (LV). The last is rather similar to the SNTV in that there are multimember constituencies, but, whereas in SNTV voters have only one vote, under LV they have more than one vote, but fewer votes than there are seats being contested. It is employed for some seats in the Spanish Senate, and was used in a few urban constituencies in Britain between 1867 and 1883. Semiproportional systems acquire their (rather misleading) name from the fact that they tend to generate greater representation for smaller parties than does the plurality system. However, the objection to classifying them as a form of proportional system is not that they tend to produce disproportional results – though in fact they do tend to do so – but rather that how proportional they are depends on the particular behaviour of voters and parties. In some situations they may not be proportional at all. By contrast, proportional systems are designed to generate a degree of proportionality.

Consequences of Electoral Systems

Public and academic interest in electoral systems tends to be much greater when the franchise is being extended (as in Britain in the 1860s), or new legislative assemblies are introduced (Britain in the 1990s) when democratic regimes have collapsed or are unstable (Europe from the 1920 to the 1950s), and when regimes are democratizing (Eastern Europe in the 1990s), or redemocratizing (Latin America in the 1980s and 1990s). One of the main focal points of attention is the likely consequences that one form of electoral system will have compared with others.

In the 1950s, and in light of the experience of collapsed democracies in the interwar years and with the potentially unstable regimes in postwar France and Italy, Duverger (1959) propounded an argument that has come to be known as Duverger's Law: 'the simple-majority single ballot system favours the two-party system'. He also argued that PR tended to be associated with multipartism. His argument was often linked to the widely held view in Britain and the USA at that time that PR was responsible for unstable governments with multiparty systems, of the kind found in the fourth Republic, whereas plurality

systems were a recipe for stable two-party democratic government.

Since Duverger's Law was first propounded it has had its critics, and at the time of writing many political scientists would point to its limitations. It does not apply when there are strong regional cleavages, for example; they also reject the popular view that PR was causally responsible for unstable government. That view does not fit well with the evidence: most European countries since 1945 have used some form of PR, and most have had multiparty systems and coalition governments, yet most have not had unstable regimes. Although there are some exceptions, notably Sartori (1994), many political scientists now argue that the electoral system has only a limited effect on a regime's party system, and hence on its potential for instability. Increases in the number of parties in a system, for instance, tend not to follow the introduction of PR, but rather tend to precede it; moreover, the timing of the introduction of PR is significant in relation to whether party fragmentation will occur. An early switch to PR following democratization does tend to produce fragmentation – because the larger parties have had limited opportunities to develop loyal electorates. A switch much later does not lead to fragmentation, because the established parties now have loyal electorates who are not likely to defect to new or small parties.

Nevertheless, the precise form of electoral system used can have an impact on election results: changing the rules does change the political game. For example, the switch from plurality voting to AV in British Columbia in the early 1950s probably did prevent the socialist CCF party from gaining power in that decade. Similarly, the switch to PR from DB for the 1986 Assembly elections in France helped to minimize the Socialists' losses and to establish the National Front as a credible party that could win seats in the Assembly. Electoral systems cannot transform the social bases of electoral politics, but they can make it more likely that some parties will win office, or share power in government, and make it less likely that others will do so. For that reason, politicians are tempted sometimes to change them. For the most part, though, they are not changed frequently – partly because those in office are the people who usually benefit from the working of the existing system, and partly because the full consequences will be unknown, and hence might actually run counter to the interests of those seeking change.

Editor's Note

Updated statements on electoral systems by the author of this article include Alan Ware, 2009 and The American Direct Primary: Party Institutionalization and Transformation in the North. Cambridge University Press, 2009.

See also: Data Bases and Statistical Systems: Political Behavior and Elections; Democratic Theory; Elections: Apportionment; Parliamentary Government; Party Systems; Pluralism; Political Representation; Proportional Representation; Voting, Sociology of; Voting: Turnout.

Bibliography

Blais, A., Massicotte, L., 1996. Electoral systems. In: LeDuc, L., Niemi, G., Norris, P. (Eds.), Comparing Democracies. Sage, Thousand Oaks, CA, pp. 49–82.

Duverger, M., 1959. Political Parties, second ed. Methuen, London.

Lijphart, A., 1994. Electoral Systems and Party Systems. Oxford University Press, Oxford, UK.

Rae, D.W., 1971. The Political Consequences of Electoral Laws, second ed. Yale University Press, New Haven, CT.

Sartori, G., 1994. Comparative Constitutional Engineering. Macmillan, Basingstoke, UK.

Ware, A., 2009. The Dynamics of Two Party Politics. Oxford University Press, Oxford.

Electroencephalography: Basic Principles and Applications

David B MacDonald, King Faisal Specialist Hospital & Research Center, Riyadh, Saudi Arabia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E.R. John, volume 7, pp. 4385–4391, © 2001, Elsevier Ltd.

Abstract

Electroencephalography is the discipline of analyzing and interpreting electrical waveforms recorded from the scalp or cortex, usually in a clinical context. Physiologic, technical, and interpretive principles must be mastered for meaningful results. Diagnostic applications include epilepsy, disturbances of consciousness, brain death, dementia, cerebrovascular or structural brain disease, and to a limited extent, psychiatric disorders. Long-term monitoring indications include epilepsy, coma, and certain surgical procedures. Computer processing techniques may compliment expert visual waveform analysis in selected brain diseases and efforts to apply them to psychiatric or cognitive assessments show promise but presently remain investigational.

Introduction

An electroencephalogram (EEG) is a recording of fluctuating electrical waveforms at the scalp. An electrocorticogram (ECoG) is a similar recording made directly from the cortex. Electroencephalography is the discipline of analyzing and interpreting these waveforms, usually in a clinical context. This article describes some basic principles and applications.

Physiologic Principles

A synapse on a cerebral neuron produces an excitatory (depolarizing) or inhibitory (hyperpolarizing) postsynaptic potential that generates electric current inside and outside the cell. By integrating many sequential or simultaneous synaptic inputs, each neuron exhibits continuously changing postsynaptic potentials. The resulting extracellular currents volume conduct through tissue, but their strength declines with the square of distance. Thus, single neuron electrical activity recordable with a microelectrode in or very near the cell is too weak to register farther away. Fortunately, synchronous extracellular currents of many neurons can spatiotemporally summate to generate aggregate potentials strong enough to record at the cortical surface or even the scalp (Buzsaki et al., 2002; Speckmann and Elger, 2005).

The uniform radial geometry of cortical pyramidal neurons enables spatiotemporal summation, while the potentials of other more randomly oriented cortical or subcortical cells tend to cancel out (Buzsaki et al., 2002; Speckmann and Elger, 2005). Thus, EEG selectively records synchronous electrical activities of pyramidal neuron populations. However, oscillating subcortical–cortical synaptic circuits modulate pyramidal neurons, thereby contributing to rhythmic waveforms that may indirectly reveal subcortical influences (Buzsaki et al., 2002; Speckmann and Elger, 2005).

Synchronous potentials of many pyramidal neurons are needed to generate scalp signals. For example, registering a spike at the scalp requires a synchronous neuronal discharge under a cortical surface area of about 5–6 cm^2 (Ebersol, 2002a). Cortical potentials diffuse as they propagate through intervening tissues, especially the skull, thereby spreading out over large scalp areas. Each of the resulting scalp fields has a peak surrounded by declining voltage, like a hill (Ebersol, 2002a,b; Connolly and Sharbrough, 2002). Many fields are dipolar, having negative and positive poles with voltage gradients between and around them, like a hill and valley (Ebersol, 2002a,b). Depending on the source, the poles may be nearby, widely separated or on opposite sides of the head, and they may have similar or different amplitude and gradients. Focal superficial cortical sources generate relatively localized peaks and steep gradients, while deep or diffuse cortical sources generate broad gradually sloping fields. Localizing an EEG potential involves analyzing its field as sampled by all scalp electrodes in order to estimate the generating cortical region (Ebersol, 2002a,b). The inevitable overlap of fields from different cortical regions can complicate localization.

Because ECoG involves less spatiotemporal summation and diffusion, one obtains more focal recordings: a high-voltage discharge at one electrode may be absent a centimeter away. Localization therefore considers each electrode separately and the activity of unrecorded cortex is unknown. The difference between EEG and ECoG can be likened to viewing a forest vs. viewing a sample of individual trees.

Some Technical Principles

Modern EEGs are digital (Krauss and Webber, 2005). The instruments have multiple solid-state amplifiers, or channels. The analog output of each amplifier is digitally sampled and stored to disk as a time series of discrete signed integer values in a binary file. To accurately record high frequencies, the analog to digital (AD) conversion sampling rate must exceed twice the highest frequency of interest. For standard clinical EEG the highest frequency is 70 Hz and 200–256 Hz sampling is sufficient. When higher frequencies are desired, the sampling rate must be increased proportionally. The vertical resolution depends on the number of AD converter bits, typically 12 ($2^{12} = 4096$ points), 16 ($2^{16} = 65\ 536$ points), or 24 ($2^{24} = 16\ 777\ 216$ points).

The size of a digital recording varies with its duration, channel number, AD bits and sampling rate. For example, a 16-channel 20-min EEG with 12-bit 200 Hz sampling may be only a few

megabytes, whereas a 256-channel 24-h recording with 24-bit 4000 Hz sampling may be several gigabytes. Patient digital video further increases size. The system's processing and storage capacity has to match its intended purpose.

Each differential amplifier measures and amplifies the voltage difference between two inputs (I_1 and I_2). This is done by electronically subtracting I_2 voltage relative to ground (G) from I_1 voltage relative to ground, so that $(I_1–G) – (I_2–G) = I_1–I_2$. By convention, the deflection is up in the event that I_1 is more negative than I_2 or I_2 is more positive than I_1, and down in the event that I_1 is more positive than I_2 or I_2 is more negative than I_1. Ideally there is no deflection when the difference is zero, but in reality a small degree of common-mode gain does occur with voltages common to both inputs. The common-mode rejection ratio (CMRR) is the ratio between differential and common-mode gain. Modern EEG amplifiers have very high CMRR typically exceeding 110 dB. Other important specifications include high input impedance and low amplifier noise.

Recordings are usually made from each electrode (E_1, E_2, …, E_n) as I_1 to a recording reference (R) as I_2. Thus, $E_1–R$, $E_2–R$, …, $E_n–R$ are recorded to disk and to see what $E_1–E_2$ was, one simply computes $(E_1–R) – (E_2–R) = E_1–E_2$. In this way, any desired derivations can be reconstructed from the stored raw data during review. Display sensitivity, montage and digital filter settings are also freely adjustable. These benefits are a major advance over older unmodifiable analog recordings written to paper.

Another digital benefit is the ability to process the EEG with computer algorithms (Lopes da Silva, 2005). For example, interpolation estimates what scalp potentials might have been in-between the recording sites at any one moment. Mapping the results with a color gradient or contour line plot on a cartoon of the head may compliment visual analysis or help to communicate findings. However, one must be aware that the interpolated values are estimates based on assumptions and subject to error. Recording from more electrodes to increase spatial sampling of real data improves accuracy (Gevins, 1998). Although EEG maps are graphic, they are not anatomical brain images.

Source dipole localization attempts to estimate the location, direction, and strength of a cortical dipole source that could explain an interpolated scalp field (Ebersol, 2002b). This is done with mathematical head models that vary from simple spherical approximations to complex individualized models based on the patient's neuroimaging. Results vary with the number of electrodes, model sophistication, and the algorithm used.

Fast Fourier transformation converts selected EEG segments into the frequency domain (Lopes da Silva, 2005). The results quantify signal power at every frequency for each channel and are often divided into bands, typically delta (<4 Hz), theta (4 to <8 Hz), alpha (8–13 Hz), and beta (>13 Hz). Beta may be subdivided into beta (13–30 Hz) and gamma (>30 Hz), although distinguishing gamma waves from cranial muscle potentials is problematic because their frequencies overlap. Quantification can compliment expert waveform interpretation and makes statistical comparisons to normative data possible, although this type of analysis is mostly a research tool (Nuwer, 1997). Again, interpolated frequency content maps can aid interpretation or communication. Expert selection of appropriate EEG segments is critical to avoid misleading results (Nuwer, 1997).

Spike and seizure detection algorithms attempt to automatically flag epileptic events, but are imperfect and require expert review and revision (Gotman, 2002). Autocorrelation, cross-correlation and chaos theory calculations are some other processing techniques having various and so far mostly research uses (Lopes da Silva, 2005).

Recording Principles

Personnel

Properly recording an EEG requires training and experience. Technologists undergo a 2-year program followed by a 1-year practicum and become qualified after board certification; some jurisdictions also stipulate continuing medical education and periodic recertification. These requirements underscore the principle that technologist skill is critical for the quality and validity of EEG results.

Patient Preparation

The patient is greeted and brought into a private, relaxed recording room. The technologist explains the procedure and summarizes the patient's medical data taken from the referral, chart, and interview. This is important for anticipating special recording procedures and for enhancing subsequent interpretation. Some patients need to be partially sleep deprived the night before. Some may require sedation that must follow hospital policies and procedures. Very ill patients may need nurse attendance and critically ill patients must have a portable bedside EEG.

Head Measurement

The international 10–20 system for electrode placement defines 21 standard scalp coordinates derived from four anatomical landmarks: the nasion, the inion, and the two preauricular points (**Figure 1**; Tyner et al., 1983). The technologist locates and marks each 10–20 coordinate using a millimeter tape measure and skin pencil. Accuracy takes practiced skill and is critical for valid results and comparisons across time or between patients. Electro-caps consisting of an elastic cap with embedded electrodes approximating 10–20 sites are considered too inaccurate for routine clinical use. An international 10–10 system defines additional coordinates midway between and below 10–20 sites and may be applied to increase spatial sampling. Some laboratories routinely include one of variously defined sites below F_8 and F_7, or a row of inferior temporal sites. The recording reference is typically between C_z and P_z, and the ground is usually on the forehead; both should be on the head, but their exact locations are less critical.

Electrode Application

Standard precautions (hand washing, gloves, personal protective barriers) apply to all patients. Additional transmission-based precautions apply to patients with active infectious disease. Scalp EEGs are normally recorded with reusable 0.5- to 1-cm

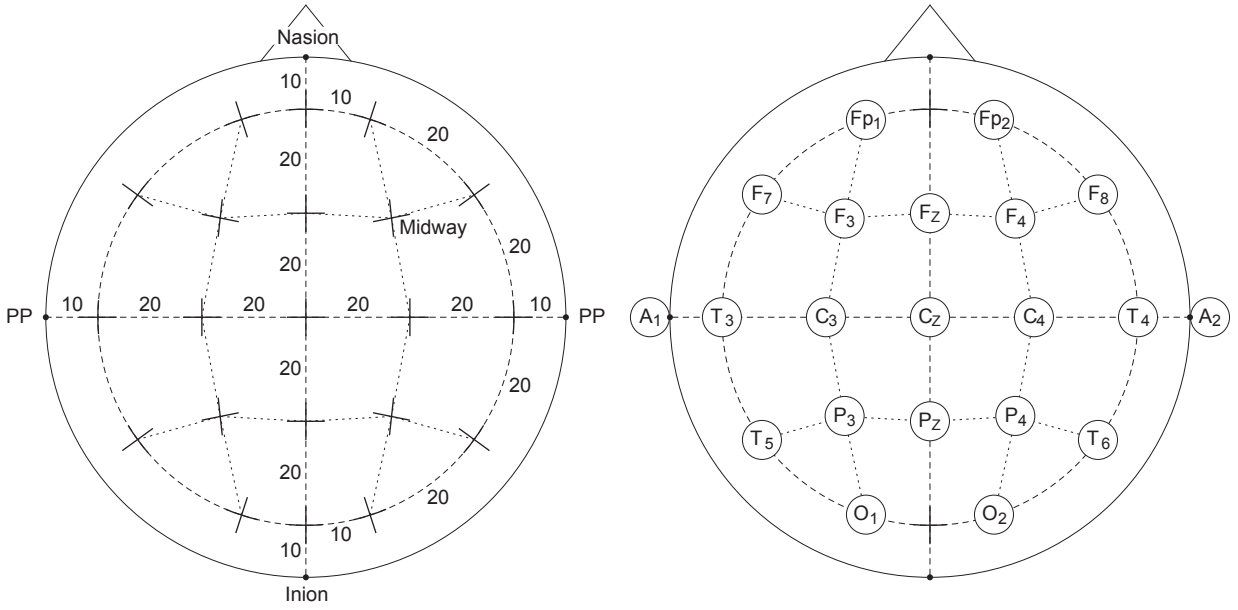

Figure 1 The international 10–20 system. Fp, frontopolar; F, frontal; C, Central; T, temporal; P, parietal; O, occipital; A, auricular; odd numbers are left, even numbers are right and z (zero) is the sagittal midline. Ten or 20% of the distance between the nasion and inion, between the two preauricular points (PP), and along each hemi-circumference determines most coordinates. Half the distance along the parasaggital planes (Fp_2 to O_2 through C_4 and Fp_1 to O_1 through C_3), the frontal plane (F_8 to F_7 through F_Z) and the parietal plane (T_6 to T_5 through P_Z) completes F_Z, C_4, C_3 and P_Z coordinates. F_4 is at the intersection of lines drawn midway between Fp_2 and C_4 and between F_8 and F_Z, thus making F_8–F_4 and F_4–F_Z distances equal and also making Fp_2–F_4 and F_4–C_4 distances equal; the same approach defines F_3, P_4 and P_3. A_2 and A_1 are on the earlobes or adjacent mastoids.

diameter metallic cup electrodes that provide excellent recording and safety; needle electrodes are not used because of infection risks. The hair and scalp should be clean and dry. The skin at each measured site must be rubbed with a mildly abrasive gel using a cotton-tipped applicator in order to reduce epidermal impedance to below 5 kΩ.

Each electrode is pressed onto to the prepared skin with a conductive paste to adequately secure it in place for a routine EEG. In the case of long-term or intraoperative EEGs needing firm attachment, electrodes filled with conductive paste or gel are glued to the skin with collodion that is subsequently dissolved with acetone for removal.

Lead wires are bundled or braided together to reduce extraneous electromagnetic interference. Tight braiding is crucial in the operating room and other electrically noisy environments.

Calibration

Because amplifiers exhibit performance variations and degrade or fail over time, the instrument must be internally calibrated before making patient connections. This ensures equal time axis, filter response and output amplitude for a known input voltage across all amplifiers. Additional biologic calibration optionally done after making patient connections ensures that each amplifier produces an identical output waveform in response to a single patient input.

Recording

A routine EEG of a conscious patient primarily assesses the relaxed awake state with the patient supine on a stretcher. Maneuvers to alert or relax patients may be needed. Segments with eyes open and closed assess occipital rhythms generated by visual cortex. Similarly, voluntary hand movements test central rhythms generated by sensorimotor cortex. Segments of hyperventilation, photic stimulation at several flash frequencies and of sleep if time permits are routine activation procedures mainly intended to bring out epileptiform abnormalities. Additional activation procedures such as reading, watching television, playing a video game, standing, listening to music, etc. may be indicated by the patient's history. The technologist adapts procedures to the history and adjusts technique according to observed waveforms and patient behavior. This includes testing responsiveness, memory, language and motor function during a suspected seizure. Frequent annotations of eye opening and closing, swallowing, movements, state and behavior are important. Patient video recording is a useful adjunct.

The minimum standard recording time is 20 min; longer recordings are encouraged. The standard bandwidth is 0.3–70 Hz with the 50 or 60 Hz notch filter off unless absolutely necessary. The minimum standard number of channels is 16, but more are encouraged, including additional EEG channels as well as ECG, eye movement and other adjunctive monitors as indicated. Display gain is adjusted to utilize dynamic range while avoiding trace overlap or clipping.

A sleep EEG may be indicated for suspected epilepsy after a negative routine EEG, or occasionally for excessively agitated patients impossible to record while awake. Partial sleep deprivation promotes sleepiness and sedation may also be needed. These recordings primarily attempt to record as much sleep as possible including slow wave sleep. Sometimes an awake and sleep EEG are both done during the same visit.

EEGs of stuporous or comatose patients are adapted to the level of consciousness and include tests of reactivity to auditory, visual, and sensory stimulation. There are special technical requirements for ancillary EEG testing of clinically diagnosed brain death (Chatrian and Turella, 2002). These include a minimum 30-min recording time, high display gain, long interelectrode distance montages, recording system integrity confirmation, and excluding hypothermia or sedation.

Long-term video-EEG or ECoG monitoring lasting days or weeks may be done for the investigation of epilepsy. The technologist trims these recordings through visual scanning assisted by computer-flagged patient push-button events and imperfect software detection of possible epileptic patterns. The trimmed segments are presented to the electroencephalographer for interpretation.

Monitoring is also done in the intensive care unit or operating room in an effort to detect adverse changes that may prompt intervention to prevent brain injury or titrate medical therapy of status epilepticus. These recordings need real-time interpretation by sufficiently trained personnel. Intraoperative ECoG may be used to tailor cortical resections for intractable epilepsy; the electroencephalographer must be in the operating room because the decisions are irrevocable.

After Recording

After recording, the technologist removes scalp electrodes and wipes off paste or gel while inspecting for rare skin reactions or abrasions; residual materials wash away with normal shampooing. The electrodes, leads, and exposed instrument surfaces are cleaned and disinfected for subsequent use. An exception occurs after recording a patient with suspected Creutzfeldt–Jakob disease, when the electrodes are sterilized or destroyed. Direct cortical and depth electrodes are always destroyed after use. The technologist reviews the recording and ensures everything is in order for clinical interpretation.

Interpretative Principles

Personnel

Properly interpreting an EEG requires extensive training and experience. Prerequisites for clinical EEG fellowship training are being a physician with fellowship training and board certification or eligibility in neurology or another relevant specialty. Fellows study EEG full-time for at least six to 12 months and often longer. In some countries, EEG is part of a comprehensive clinical neurophysiology residency after medical internship. Full qualification requires board certification; some jurisdictions also stipulate continuing medical education and periodic recertification. These requirements stress the principle that broad in-depth expertise is essential for clinically meaningful interpretation.

Artifact Discrimination

Scalp EEGs are highly amplified to register signals in the range of tens of microvolts. Consequently, many potentials are not actually from the brain, but rather from other biologic or nonbiologic sources and these ever-present artifacts must be consistently excluded from interpretation. In fact, there are no truly artifact-free segments and it can be helpful to think of the EEG as an 'electroscalpogram.' In extreme cases, brain activity can be so obscured that no interpretation is possible. After proper training and experience it becomes second nature to recognize artifacts, but some can still be challenging. Inexpert readers can easily mistake artifacts for brain waves.

Biologic artifacts include electromyographic signals from scalp, extraocular, pharyngeal and neck muscles, blinking and other eye movements, tongue movements, ECG, pulse and balistocardiographic artifact, sweat artifact, and others. Nonbiologic artifacts include 50 or 60 Hz and other extraneous electromagnetic interference including cell phones, electrode artifacts consisting of transient 'pops' or more sustained rhythmic potentials, movement artifacts, electrostatic discharges, dissimilar metal artifact from dental fillings, amplifier noise, and others.

Sleep EEGs contain fewer artifacts because of patient stillness and cranial muscle relaxation, but are not artifact free. ECoG contains relatively few artifacts because of lower amplification to register signals in the range of hundreds of microvolts. Nevertheless, artifacts still arise.

Processed EEG methods accentuate the problem because computer algorithms do not distinguish artifacts from brain signals. Consequently, meaningful results require expert selection of relatively clean EEG segments to process, while keeping in mind that there is always some residual artifact (Nuwer, 1997).

Normal and Benign Variant Patterns

The electroencephalographer must master the full range of normal and benign variant EEG patterns at all ages and states (Kellaway, 2002; Niedermeyer, 2005a,b; Westmoreland, 2002). Normal 'background' consists of several overlapping, waxing-and-waning rhythms, and transient waveforms generated by different cortical regions. Momentarily higher amplitude or isolated background waveform components may suggest abnormality, but morphologically fit in with ongoing rhythms in other segments, thereby being normal. Normal intermittent transient waveforms may simulate abnormality either alone or in random combination with other background components. All these normal features exhibit highly variable amplitude and commonly have sharp contours.

Criteria based on frequency, distribution, morphology, polarity, state, reactivity, amplitude, and symmetry define normal patterns and their variations. For example, the occipital alpha rhythm is an 8–13 Hz posterior sinusoidal, notched, or sharply contoured rhythm of variable regulation, persistence, and amplitude that may differ by up to 50% between the two sides of the head. It is present in relaxed wakefulness with eyes closed and 'reacts' (attenuates or desynchronizes) bilaterally to eye opening or mental effort. It may persist into early drowsiness, but disappears in sleep. The fast alpha and slow alpha variants consist of occipital rhythms having approximately twice or half the dominant alpha frequency and showing identical reactivity. Another

normal variation consists of persistently low voltage (<20 μV) or even absent occipital alpha rhythm.

Similarly, the central mu rhythm is an alpha-frequency rhythm of variable persistence and amplitude that may differ by up to 100% between sides and is present in a minority of normal subjects. It has an arciform morphology with negative sharp peak and rounded positive trough. It appears in relaxed motionless wakefulness with eyes open or closed, attenuates with movement planning or execution or with somatosensory stimuli and may persist into early drowsiness but disappears in sleep.

Normal beta (>13 Hz) activity is generalized but frontal dominant, sinusoidal, sharply contoured or bursting with variable <50-μV amplitude that may differ by up to 35% between sides. It is present in wakefulness, may persist or increase in drowsiness, subsides in deep sleep and shows no specific reactivity.

Some other criteria-defined normal patterns include lambda waves, certain slow activities such as posterior slow waves of youth, hyperventilation, and photic stimulation responses, drowsy patterns and vertex waves, sleep spindles, k-complexes, and positive occipital sharp transients of sleep.

Normal background evolves from infancy to old age (Niedermeyer, 2005b). Infants' EEGs are dominated by delta and theta activity that gradually declines through childhood to adult appearance dominated by alpha and beta frequencies. The occipital dominant rhythm first appears around 3 months and is at least 5 Hz by 1 year, 6 Hz by 4 years, 7 Hz by 5 years, and 8 Hz at 8 years of age or older (Lüders and Noachtar, 2000). Sleep spindles are typically asynchronous and sharply contoured between 3 and 12 months of age. Posterior slow waves of youth are prominent in children and teenagers but occasionally persist into the 20s or rarely even 30s. A few intermittent temporal and generalized slow activities may arise during normal senescence.

Furthermore, several infrequent benign variant patterns resembling epileptiform discharges (small sharp spikes, wicket waves, 14 and 6 Hz positive bursts, and 6 Hz 'phantom' spike-and-wave), or seizure-like patterns (rhythmic mid-temporal theta, subclinical rhythmic electrographic discharges of adults, midline theta, and frontal arousal rhythm) must be recognized by their criteria and excluded from clinical diagnosis (Westmoreland, 2002).

Abnormal EEG Classification

Structured classification of EEG abnormalities is recommended to encourage precision, enhance communication, facilitate research, provide a framework for trainees, and make EEG an objective criteria-based discipline. There are various systems, mostly unpublished and institution-specific. The author recommends the published classification of Lüders and Noachtar (2000). It provides well-selected abnormal categories (**Table 1**) defined by criteria along with localization and grading rules, as well as established clinical correlations and excellent illustrations. Fully describing all these abnormalities is beyond the scope of this article. Another published system that may gain prominence is 'SCORE' (standardized computer-based organized reporting of EEG) (Beniczky et al., 2013).

Table 1 An EEG classification system

Normal
Abnormal
Slow activity
Background slow
Intermittent slow
Continuous slow
Epileptiform patterns
Spikes
Sharp waves
Benign epileptiform discharges of childhood
Spike-and-wave complexes
3 Hz spike-and-wave complexes
Slow spike-and-wave complexes
Polyspikes
Hypsarrhythmia
Photoparoxysmal response
Seizure pattern
Status pattern
Artifact-obscured EEG
Special patterns
Excessive fast activity
Asymmetry
Sleep-onset rapid eye movement
Periodic pattern
Triphasic waves
Periodic lateralized epileptiform discharges
Burst-suppression
Background suppression
Special patterns used only for patients in coma or stupor
Alpha coma/alpha stupor
Spindle coma/spindle stupor
Beta coma/beta stupor
Theta coma/theta stupor
Delta coma/delta stupor
Electrocerebral inactivity

Adapted from Lüders, H.O., Noachtar, S., 2000. Atlas and Classification of Electroencephalography. W.B. Saunders, Philadelphia, With permission.

Approach to Clinical Interpretation

The EEG is first read without knowing the patient's clinical history to prevent bias and ensure independent assessment. The electroencephalographer classifies the findings based on criteria. Then the patient's clinical information is reviewed. Finally, the classification and clinical data are merged to synthesize an interpretation setting forth the likely meaning of the EEG classification for that patient. Adherence to this approach progressively hones skill though clinical practice.

Care must be taken to avoid over interpretation that can lead to misdiagnosis, especially of epilepsy. Spikes and sharp waves are the interictal epileptiform abnormalities that correlate with epilepsy (Niedermeyer, 2005c; Lüders and Noachtar, 2000; Zifkin and Cracco, 2002). They are sharp, brief (20–200 ms) cerebral transients that are unequivocally distinct from background and do not conform to criteria for any other normal, benign variant or abnormal waveforms. Thus, the reader must have a full command of EEG to reliably distinguish epileptiform abnormalities. Inexpert readers can easily over-report or miss them. In fact, trainees develop through three

predictable stages. In the first stage they perceive almost anything sharp as an epileptiform abnormality. In the second stage they perceive almost nothing sharp as abnormal because there are so many normal sharp waveforms. In the third stage they find objective balance and exhibit progressively reliable discrimination. The same learning process applies to other EEG abnormalities.

Montages

A montage is a combination of channel organization and analysis technique intended to facilitate localization and comparisons between homologous regions or channels and across the coronal planes (Connolly and Sharbrough, 2002). The problem of presenting time-varying signals from the three-dimensional surface of the head on a two-dimensional display is similar to the cartographic problem of mapping the globe on a sheet of paper: spatial distortion is inevitable, but can be minimized. To this end, montages should be anatomically logical, organized front to back (longitudinal) or side-to-side (transverse) and consistent with the laboratory's chosen left over right or right over left convention.

Longitudinal channels are usually organized into midline, parasaggital and temporal groups. Homologous groups of the two sides may be paired together in the display to aid regional symmetry assessment, or unpaired to minimize spatial distortion and enhance the display of peri-sagittal features or hemispheric asymmetry. Alternating longitudinal montages display sequential channel pairs to facilitate homologous channel comparison, but increase spatial distortion. Transverse montages facilitate coronal comparisons with minimal spatial distortion and are usually organized into frontopolar, frontal, central, parietal, and occipital groups. Thus, there are four basic channel arrangements: paired group longitudinal, unpaired group longitudinal, alternating longitudinal, and transverse (**Figure 2**). Cartographers would prefer unpaired group and transverse displays, but there are valid indications for the others.

The three analysis techniques are bipolar, referential and source derivation. Each has strengths and weakness for displaying and localizing various EEG signals. Bipolar technique measures sequential potential differences across electrode chains. Localization considers mainly 'phase reversal' and 'end of chain' effects. Bipolar montages can provide reasonable background feature differentiation, good symmetry

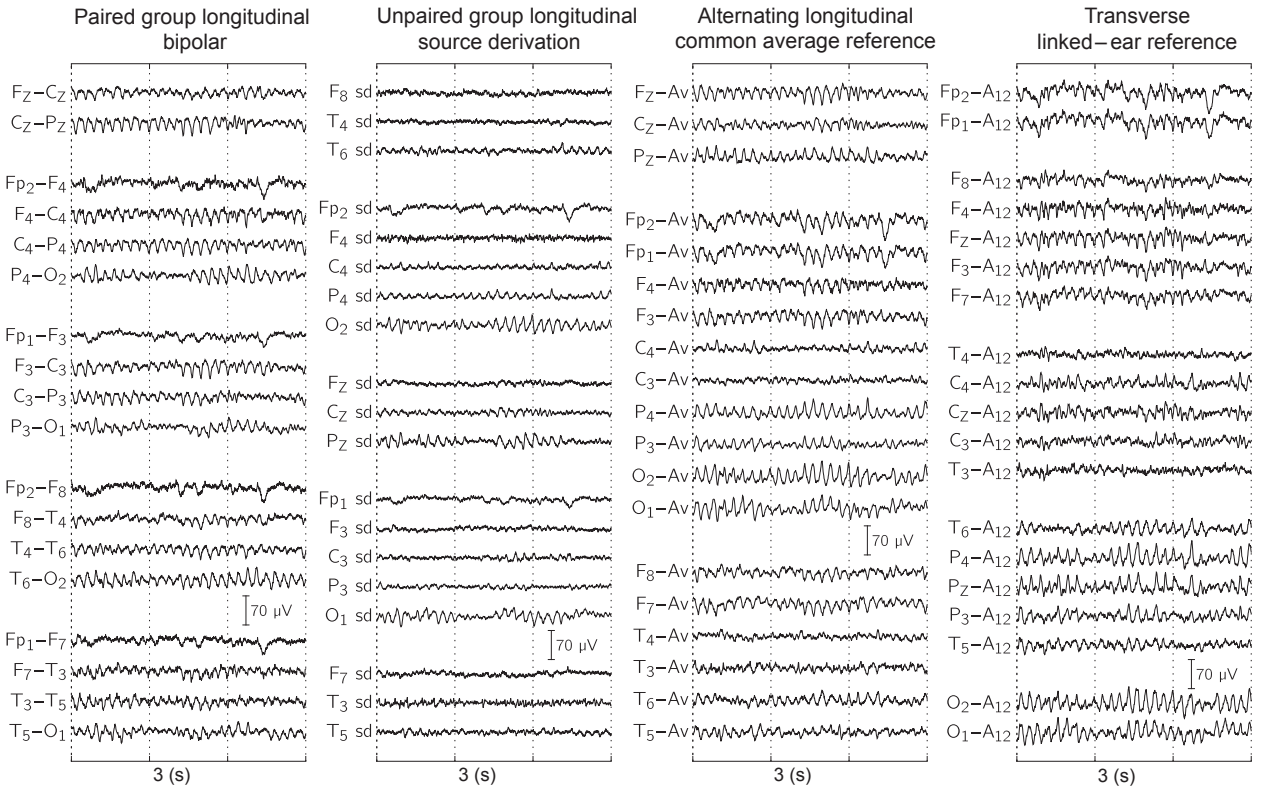

Figure 2 The same EEG segment displayed with four different montages. Av, average; sd, source derivation; $A_{12} = (A_1 + A_2)/2$. Paired group montages facilitate homologous group comparison, unpaired group displays minimize spatial distortion, alternating montages aid homologous channel comparison, and transverse displays enhance coronal comparison. In this example, bipolar technique provides reasonable background differentiation, but blurs together some features, while source derivation better separates background patterns, such occipital (O_2, O_1) alpha and central (C_4, C_3) mu rhythms. Both referential techniques exhibit background rhythm reference contamination causing spurious frontal derivation alpha activity and poor background differentiation. Common average is useful for analyzing focal abnormalities or dipolar scalp fields and linked-ears can enhance broad peri-sagittal fields, such as sleep potentials, generalized epileptiform abnormalities, or fronto-central beta activity as seen in this example.

assessment, and generally display focal features well. However, they distort broad scalp fields and are subject to spurious asymmetry due to uneven electrode placement, as well as in-phase cancellation when two electrodes equally near to a scalp field peak are in the same channel.

Referential technique measures potential difference between each electrode and a reference common to all channels. The reference may be the ipsilateral or linked ears, the recording reference, another strategically chosen electrode, or the common average of all electrodes usually excluding frontopolar and ear leads. The principle is to select the least active reference for the waveform of interest. Localization considers mainly amplitude. Referential montages can display localized and broad scalp fields well and are less subject to spurious asymmetry. However, they are subject to reference contamination when the reference is active for the waveform of interest. The common average reference can aid focal and dipolar scalp field analysis, but is often contaminated by background rhythms and broad fields, so is suboptimal for assessing these features. The linked-ear reference enhances broad perisagittal scalp fields, but can be contaminated by posterior background rhythms or temporal scalp potentials, so is often suboptimal for analyzing these features.

Source derivation technique (Hjorth, 1991) measures potential difference between each electrode and a weighted average of its three or four closest neighboring electrodes or alternatively, all the other electrodes. The weighting coefficients are the squares of the reciprocal distances from the central electrode and normalized to a total value of -1. The sum of the partial products is then added to the central electrode weighted $+1$. The effect is a high-pass spatial filter that tends to 'de-blur' scalp potential fields. Localization considers mainly amplitude. Source derivation facilitates background differentiation, symmetry assessment and focal scalp field analysis. However, it reduces overall amplitude, distorts broad fields and is subject to spurious phase reversals.

Thus, the four display methods and three analysis techniques combine into 12 fundamentally different montage types. Electroencephalographers naturally develop individual preferences, but must be fluent with all montages and able to select one that optimally displays and localizes the waveform of interest.

Applications

Epilepsy

Epilepsy is the main EEG indication. Epileptic neurons exhibit abnormally large and prolonged membrane discharges called paroxysmal depolarizing shifts. When these occur synchronously in a large enough number of epileptic pyramidal neurons, spatiotemporal summation generates an EEG spike or sharp wave discharge (Niedermeyer, 2005c). Thus, epileptiform abnormalities are uniquely relevant to epilepsy and are not disclosed by any other test apart from the related magnetoencephalogram.

Interictal epileptiform abnormalities correlate very highly with epilepsy, but infrequently arise in patients who have not had a clinical seizure and are not always found in epileptics. Thus, their presence can support, but not prove a clinical diagnosis of epilepsy, and their absence cannot rule out epilepsy. Regional epileptiform abnormalities suggest focal epilepsy (Figure 3), while diffuse ones suggest generalized epilepsy that may be primary if background is normal (Figure 4), or symptomatic of an underlying brain disease if background is abnormal. These diagnostic refinements can guide treatment decisions. For example, different anticonvulsant medications may be indicated for generalized or focal seizures, and surgery may be considered for medically intractable focal epilepsy.

Some types of epileptiform abnormality can point to certain syndromes, such as West syndrome, Lennox-Gastaut syndrome, benign Rolandic epilepsy of childhood, juvenile myoclonic epilepsy, absence epilepsy, or temporal lobe epilepsy (Niedermeyer, 2005c). Syndromic diagnoses may further refine treatment and prognosis. For example, Lennox-Gastaut syndrome is characterized by lifelong intractable seizures and mental retardation, juvenile myoclonic epilepsy is usually lifelong but often easily controlled by medication, and benign Rolandic epilepsy of childhood predictably remits after puberty.

Seizures are infrequent during routine EEGs, but are regularly obtained by long-term monitoring, often done with anticonvulsant tapering. The recording of an EEG seizure pattern is diagnostic and can localize the cerebral region of seizure onset, or demonstrate a generalized onset (e.g., Figure 5). The recording of an ECoG seizure pattern with intracranial electrodes can more precisely localize epileptogenic cortex and may be necessary for selected patients considering epilepsy surgery. The absence of a seizure pattern during a clinical attack resembling an epileptic seizure can help to establish an alternative diagnosis, such as convulsive syncope or psychogenic events. Thus, monitoring is important for the evaluation of epilepsy surgery candidates as well as for seizure classification and differentiation between epileptic and nonepileptic attacks.

Some research frontiers in epilepsy include analysis of infra slow and gamma activities (high-frequency oscillations) that might have localizing value, as well as efforts to treat epilepsy with implanted electrodes and microchips attempting to predict seizure onset and then initiate electrical stimulation hoping to abort a seizure.

Disorders of Consciousness

Impaired consciousness is another common EEG indication. Bihemispheric and/or brainstem malfunction is necessary to reduce consciousness. The clinical manifestations range from mild confusion and disorientation to stupor with impaired responsiveness to unresponsive coma; encephalopathy is a generic clinical term applied to these states. The EEG typically shows generalized nonspecific abnormalities, most often consisting of various types and degrees of slow activity roughly correlating to clinical severity (Lüders and Noachtar, 2000; Chatrian and Turella, 2002; Zifkin and Cracco, 2002). When sedative drugs are contributing, there may be intermingled excessive fast activity. With severe encephalopathy there may be triphasic waves, periodic discharges, burst-suppression or background suppression. A monorhythmic unresponsive

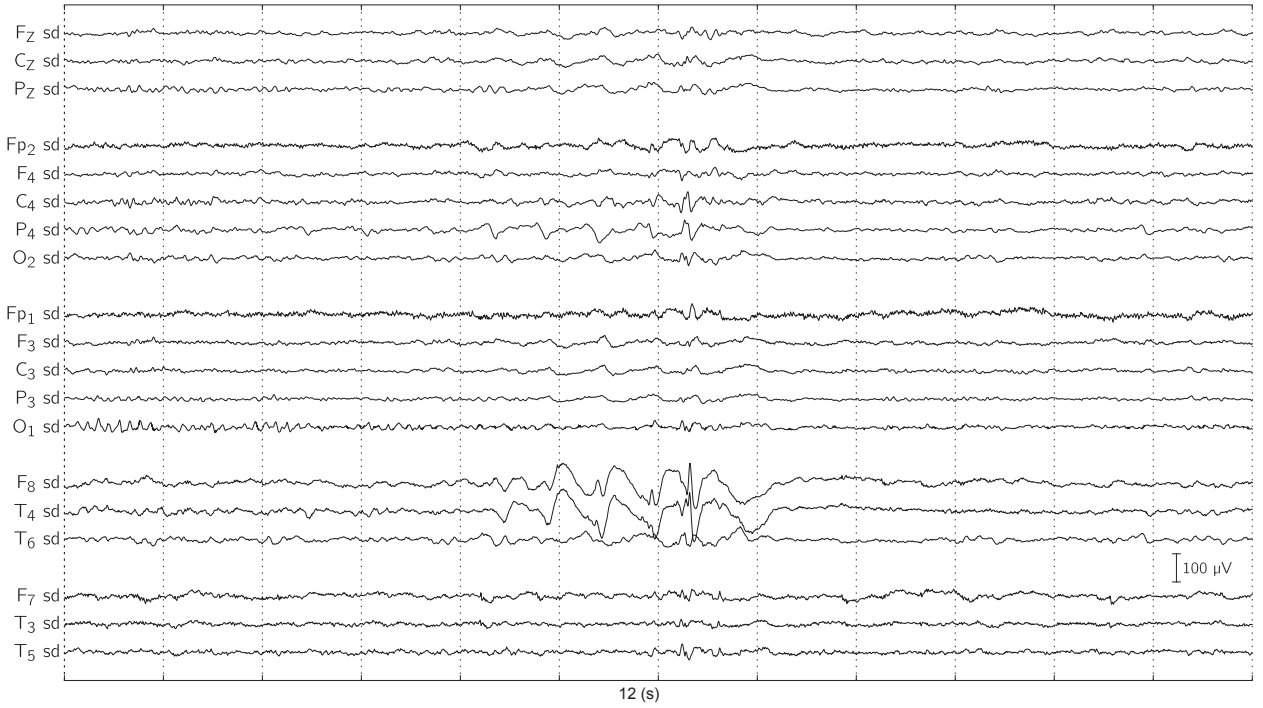

Figure 3 Right anterior to mid temporal (F_8, T_4) sharp waves displayed with a paired group longitudinal source derivation (sd) montage. Based on this focal interictal epileptiform abnormality, the patient's clinical diagnosis was revised from generalized epilepsy with tonic-clonic seizures to temporal lobe epilepsy with secondarily generalized seizures.

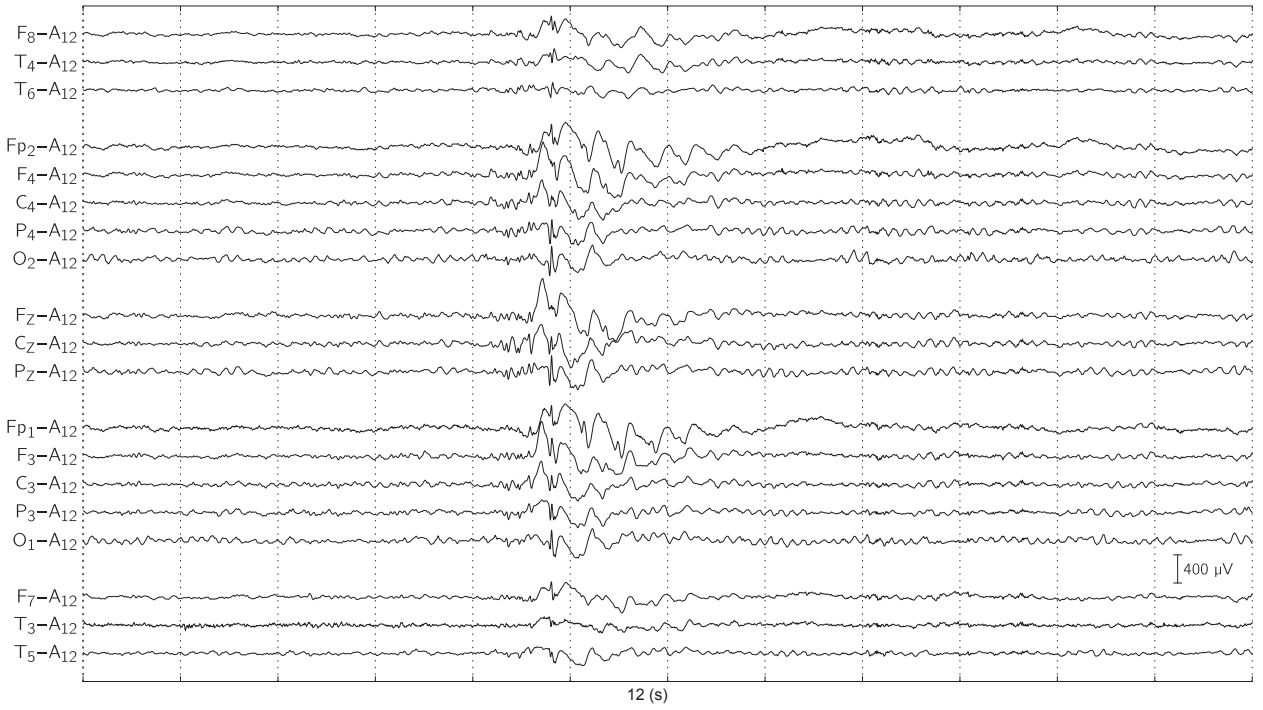

Figure 4 Generalized polyspikes and spike-and-wave complexes displayed with an unpaired group longitudinal linked-ear (A_{12}) referential montage. The diffuse interictal epileptiform abnormalities arising from normal background supported the patient's clinical diagnosis of primary generalized epilepsy with myoclonic and tonic-clonic seizures.

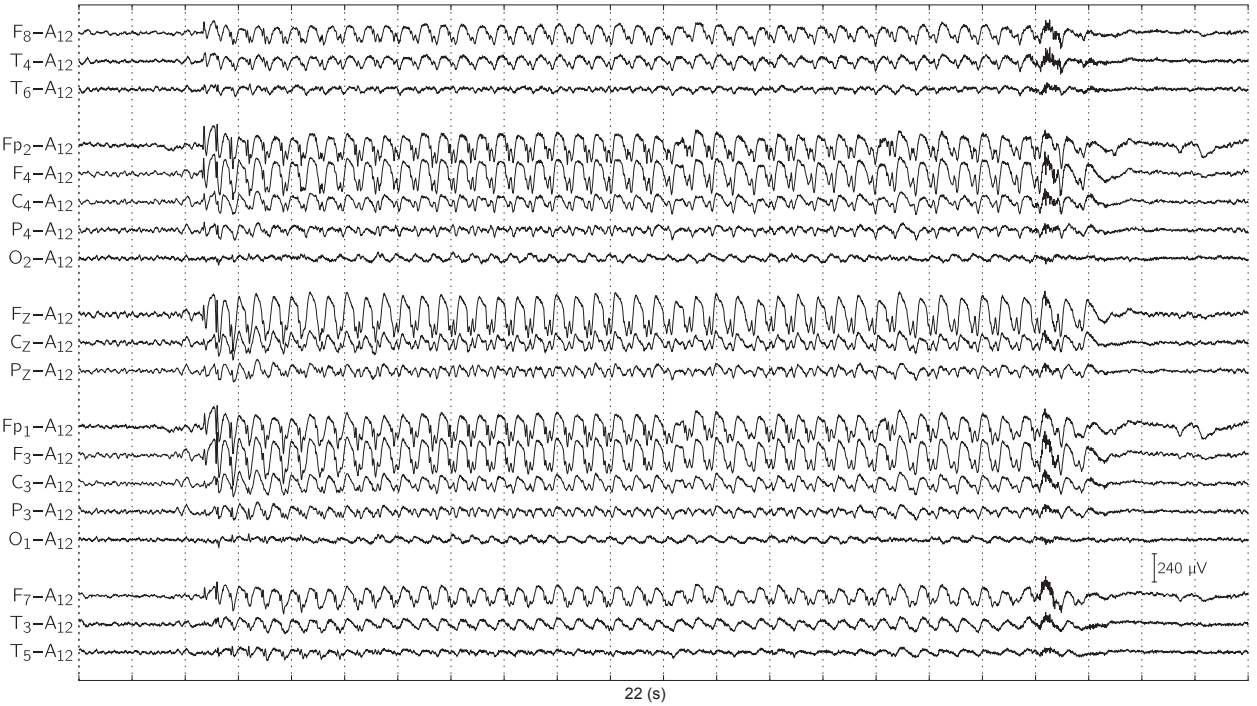

Figure 5 A generalized 3 Hz spike-and-wave seizure pattern displayed with an unpaired group longitudinal linked-ear (A_{12}) referential montage. The patient was unresponsive with rhythmic eye blinking and amnesia during the 17-s event. The recording is diagnostic of primary generalized absence epilepsy and corrected a mistaken clinical diagnosis of temporal lobe epilepsy.

alpha coma pattern is sometimes seen as an agonal rhythm in hypoxic–ischemic encephalopathy.

Typically the EEG confirms encephalopathy, but does not disclose the etiology or prognosis. However, in certain clinical contexts, severe abnormalities such as burst-suppression or suppression can indicate a grave prognosis and thereby assist clinical management decisions (Chatrian and Turella, 2002). Sometimes the EEG reveals a status pattern leading to a diagnosis of potentially reversible nonconvulsive status epilepticus causing or contributing to the impairment of consciousness. In this case, anticonvulsant treatment might promote recovery. Focal slow activity or asymmetry may suggest other treatable etiologies, such subdural hematoma or focal brain lesion, but these possibilities are better evaluated with neuroimaging. Long-term monitoring may increase detection of potentially treatable conditions or complications and can also guide the titration of medical therapy for status epilepticus. Rarely a surprisingly normal EEG in an unresponsive patient may suggest psychogenic unresponsiveness, catatonia or locked-in syndrome (motor paralysis with preserved awareness and sensation).

Brain Death

Brain death is a clinical diagnosis based on unresponsive coma with persistently absent brainstem reflexes due to irreversible brain disease (Chatrian and Turella, 2002). Formal protocols must be followed to establish this very serious diagnosis. In some jurisdictions, clinical criteria are sufficient and ancillary testing is optional; in others an ancillary test is mandatory. The ancillary test can be an angiogram or other study demonstrating absent cerebral blood flow, or an EEG showing electrocerebral inactivity. Note that electrocerebral inactivity by itself is not diagnostic of brain death, but may also occur with sedation, hypothermia and severe cortical damage with preserved brainstem function. Hence, the clinical diagnosis must be made first.

Because of the serious consequences, only properly qualified personnel should record and interpret brain death EEGs. The recordings must satisfy all technical requirements. These EEGs are never 'flat': interpretation must recognize and account for all artifacts before concluding that there is no detectable cortical activity (Chatrian and Turella, 2002).

Dementia

EEGs are often done for dementia that consists of a progressive deterioration of memory, insight, cognition, personality and behavior due to diffuse or multifocal cortical and/or subcortical brain pathology. Alzheimer's disease is the most widely known etiology, but there are many others. The EEG typically shows nonspecific background slow activity and other slow activities roughly corresponding to dementia severity. However, occasionally it may reveal more specific abnormalities, such as periodic patterns suggesting Creutzfeldt-Jacob disease or subacute sclerosing panencephalitis. Rarely, focal slow activity and/or asymmetry may raise the possibility of a potentially treatable structural cause, such as chronic subdural hematoma or brain tumor, but these possibilities are better evaluated by neuroimaging. Epileptiform

abnormalities may appear in dementia with or without a history of clinical seizures; rarely, seizure or status patterns may be encountered and modify medical therapy. A normal EEG might suggest pseudodementia due to depression, but can also occur in the early phases of organic dementias (Chatrian and Turella, 2002; Markand and Brenner, 2002).

Cerebrovascular and Structural Brain Disease

Stroke and brain tumor were common EEG indications in the past, but today modern neuroimaging provides superior anatomical assessment. Regional continuous slow activity and asymmetry consisting of reduced background rhythm amplitude are the hallmarks of focal structural brain lesions (Lüders and Noachtar, 2000; Zifkin and Cracco, 2002). However, the same abnormalities can be found with reversible functional disturbances, such as during a transient ischemic attack (TIA), or following a focal seizure or classical migraine with aura. Thus, it is important to remember that the EEG is a functional study and not an anatomical imaging test. Regional continuous slow activity and asymmetry with increased background rhythm amplitude ('breach rhythm') are characteristic findings after craniotomy. Deep midline brain lesions may be manifest by nonspecific generalized intermittent or continuous slow activity.

Sometimes the EEG can help differentiate between TIA and seizure. It may also identify epileptiform abnormalities or seizure patterns complicating stroke or brain tumor. Intraoperative EEG monitoring during carotid endarterectomy can guide selective shunting to minimize the risk of perioperative stroke.

Psychiatry, Psychology and Cognition

The concept that an EEG could objectively measure psychiatric disorders, psychological states or cognition is intuitively appealing since the underlying biochemical and synaptic processes might be expected to produce characteristic EEG patterns. However, while routine EEG can disclose epilepsy or other brain disease mimicking, complicating or contributing to psychiatric disorders and may show certain adverse psychotropic drug effects (Small, 2005), it has little more to offer psychiatric diagnosis or cognitive assessment because simply visualizing signals from 21 scalp sites cannot possibly reveal the vast and mysterious complexity of human thought and emotion (Gevins, 1998).

Nevertheless, a growing body of research indicates that more detailed analyses could measure some aspects of thought, emotion, and their disorders (Gevins, 1998; Altenmuller et al., 2005; Lopes da Silva, 2005). These methods involve computer processing, such as statistical frequency content analysis, autocorrelation, cross-correlation, cognitive evoked potentials, and other techniques best applied to 'high-resolution' recordings from many more scalp electrodes to increase spatial sampling (Gevins, 1998).

These fascinating approaches do show promise and might ultimately achieve some of their objectives, but none have yet reached mainstream clinical use for individual patient diagnosis. For example, they are not part of the criteria for mental or behavioral disorders in either the Diagnostic and Statistical Manual or the International Classification of Diseases (Small, 2005). In addition, according to the American Academy of Neurology and the American Clinical Neurophysiology Society, computer processed EEG remains investigational for mental and behavioral disorders and risks serious misdiagnosis when clinically applied by inexpert practitioners (Nuwer, 1997).

Conclusion

Properly recording and interpreting an EEG or ECoG requires suitable training and qualification. The basic principles and criteria for normal, benign variant, and abnormal patterns must be fully mastered and adhered to. The indications for diagnostic recordings include epilepsy, disturbances of consciousness, brain death, dementia, cerebrovascular or structural brain disease, and to a limited extent, psychiatric disorders. Monitoring is useful for selected patients with epilepsy or impaired consciousness and for some surgical procedures such as carotid endarterectomy. Some quantitative computer EEG analyses are useful adjuncts to expert waveform interpretation in selected neurological brain diseases. Efforts to apply quantitative EEG to psychiatric or cognitive assessments show promise but presently remain investigational.

See also: Alzheimer's Disease, Neural Basis of; Cerebral Cortex; Consciousness, Neural Basis of; Epilepsy; Eye Movements in Reading; Functional Brain Imaging of Language processes; Kindling; Sleep: Neural Systems.

Bibliography

Altenmuller, E.O., Munte, T.F., Gerloff, C., 2005. Neurocognitive functions and the EEG. In: Niedermeyer, E., Lopes da Silva, F. (Eds.), Electroencephalography: Basic Principles, Clinical Applications and Related Fields. Lippincottt Williams & Wilkins, Philadelphia, pp. 661–682.

Beniczky, S., Aurlien, H., Brøgger, J.C., et al., 2013. Standardized computer-based organized reporting of EEG: SCORE. Epilepsia 54 (6), 1112–1124.

Buzsaki, G., Traub, R.D., Pedley, T.A., 2002. The cellular basis of EEG activity. In: Ebersol, J.S., Pedley, T.A. (Eds.), Current Practice of Clinical Electroencephalography, third ed. Lippincott Williams & Wilkins, Philadelphia, pp. 1–11.

Chatrian, G.E., Turella, G.S., 2002. Electrophysiological evaluation of coma, other states of diminished responsiveness, and brain death. In: Ebersol, J.S., Pedley, T.A. (Eds.), Current Practice of Clinical Electroencephalography, third ed. Lippincott Williams & Wilkins, Philadelphia, pp. 405–462.

Connolly, M.B., Sharbrough, F.W., 2002. Electrical fields and recording techniques. In: Ebersol, J.S., Pedley, T.A. (Eds.), Current Practice of Clinical Electroencephalography, third ed. Lippincott Williams & Wilkins, Philadelphia, pp. 72–99.

Ebersol, J.S., 2002a. Cortical generators and EEG voltage fields. In: Ebersol, J.S., Pedley, T.A. (Eds.), Current Practice of Clinical Electroencephalography, third ed. Lippincott Williams & Wilkins, Philadelphia, pp. 12–31.

Ebersol, J.S., 2002b. EEG voltage topography and dipole source modeling of epileptiform potentials. In: Ebersol, J.S., Pedley, T.A. (Eds.), Current Practice of Clinical Electroencephalography, third ed. Lippincott Williams & Wilkins, Philadelphia, pp. 732–752.

Gevins, A., 1998. The future of electroencephalography in assessing neurocognitive functioning. Electroencephalography and Clinical Neurophysiology 106 (2), 165–172.

Gotman, J., 2002. Automatic detection and analysis of seizures and spikes. In: Ebersol, J.S., Pedley, T.A. (Eds.), Current Practice of Clinical Electroencephalography, third ed. Lippincott Williams & Wilkins, Philadelphia, pp. 713–731.

Hjorth, B., 1991. Principles for transformation of scalp EEG from potential field into source distribution. Journal of Clinical Neurophysiology 8 (4), 391–396.

Kellaway, P., 2002. Orderly approach to visual analysis: elements of the normal EEG and their characteristics in children and adults. In: Ebersol, J.S., Pedley, T.A. (Eds.), Current Practice of Clinical Electroencephalography, third ed. Lippincott Williams & Wilkins, Philadelphia, pp. 100–159.

Krauss, G.L., Webber, R.S., 2005. Digital EEG. In: Niedermeyer, E., Lopes da Silva, F. (Eds.), Electroencephalography: Basic Principles, Clinical Applications and Related Fields. Lippincottt Williams & Wilkins, Philadelphia, pp. 797–814.

Lopes da Silva, F., 2005. Computer-assisted EEG diagnosis: pattern recognition and brain mapping. In: Niedermeyer, E., Lopes da Silva, F. (Eds.), Electroencephalography: Basic Principles, Clinical Applications and Related Fields. Lippincottt Williams & Wilkins, Philadelphia, pp. 1233–1264.

Lüders, H.O., Noachtar, S., 2000. Atlas and Classification of Electroencephalography. Saunders, Philadelphia.

Markand, O.N., Brenner, R.P., 2002. Organic brain syndromes and dementia. In: Ebersol, J.S., Pedley, T.A. (Eds.), Current Practice of Clinical Electroencephalography, third ed. Lippincott Williams & Wilkins, Philadelphia, pp. 378–404.

Niedermeyer, E., 2005a. The normal EEG of the waking adult. In: Niedermeyer, E., Lopes da Silva, F. (Eds.), Electroencephalography: Basic Principles, Clinical Applications and Related Fields. Lippincottt Williams & Wilkins, Philadelphia, pp. 167–192.

Niedermeyer, E., 2005b. Maturation of the EEG: development of waking and sleep patterns. In: Niedermeyer, E., Lopes da Silva, F. (Eds.), Electroencephalography: Basic Principles, Clinical Applications and Related Fields. Lippincottt Williams & Wilkins, Philadelphia, pp. 209–234.

Niedermeyer, E., 2005c. Epileptic seizure disorders. In: Niedermeyer, E., Lopes da Silva, F. (Eds.), Electroencephalography: Basic Principles, Clinical Applications and Related Fields. Lippincottt Williams & Wilkins, Philadelphia, pp. 209–234.

Nuwer, M., 1997. Assessment of digital EEG, quantitative EEG, and EEG brain mapping: report of the American Academy of Neurology and the American Clinical Neurophysiology Society. Neurology 49 (1), 277–292.

Small, J., 2005. Psychiatric disorders and EEG. In: Niedermeyer, E., Lopes da Silva, F. (Eds.), Electroencephalography: Basic Principles, Clinical Applications and Related Fields. Lippincottt Williams & Wilkins, Philadelphia, pp. 639–660.

Speckmann, E.J., Elger, C.E., 2005. Introduction to the neurophysiological basis of the EEG and DC potentials. In: Niedermeyer, E., Lopes da Silva, F. (Eds.), Electroencephalography: Basic Principles, Clinical Applications and Related Fields. Lippincottt Williams & Wilkins, Philadelphia, pp. 17–30.

Tyner, F.S., Knott, J.R., Mayer, W.B., 1983. Electrode placement. In: Tyner, F.S., Knott, J.R., Mayer, W.B. (Eds.), Fundamentals of EEG Technology, Basic Concepts and Methods, vol. 1. Raven Press, New York, pp. 136–145.

Westmoreland, B.F., 2002. Benign electroencephalographic variants and patterns of uncertain clinical significance. In: Ebersol, J.S., Pedley, T.A. (Eds.), Current Practice of Clinical Electroencephalography, third ed. Lippincott Williams & Wilkins, Philadelphia, pp. 235–245.

Zifkin, B.G., Cracco, R.Q., 2002. An orderly approach to the abnormal electroencephalogram. In: Ebersol, J.S., Pedley, T.A. (Eds.), Current Practice of Clinical Electroencephalography, third ed. Lippincott Williams & Wilkins, Philadelphia, pp. 288–302.

Electronic Waste

Josh Lepawsky, Memorial University, St. John's, NL, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Electronic waste (e-waste) can be understood as the detritus of the information age. Yet, there are important historical and geographical reasons to consider a more nuanced understanding of e-waste. Historically, the electronic objects that are today treated as waste in places such as the United States and the European Union were explicitly treated as a source of strategic resources. Contemporary international trade in e-waste does not conform to a simple geography of dumping. A fuller understanding of e-waste entails fundamental ethical questions about learning how to live well together.

Introduction

Electronic waste (e-waste) can be understood as the detritus of the information age (a period of human history corresponding, roughly, with the rise of mass digital computing in the 1970s). Yet, such a definition hinders understanding of the historical contingency and geographical variation of practices through which electronics become waste. One might more accurately approach e-waste *qua* waste as an indexical possibility, that is, like the words 'here' and 'there' in the English language, e-waste as waste only comes to be waste in a situation in which it is enacted as such. Understood this way, e-waste is a contingent result of a conjuncture of specific and situated practices of ridding and the affordances, both material and informational, of digital things. What this means is that discarded electronics need not necessarily become waste.

The most common representations of e-waste today offered by environmental nongovernmental organizations (ENGOs), media, and the large majority of scholarly studies, frame the issue as a postconsumption problem characterized by electronics being dumped by rich Northern consumers on the poor of the global South. This seemingly straightforward narrative has arguably become the dominant story line of environmental politics (Hajer, 1995) surrounding e-waste. This dominant story line is predominantly framed by variants of the pollution haven hypothesis. The pollution haven hypothesis proposes that polluting industries have a propensity to agglomerate in those jurisdictions where environmental regulation is least costly and/or least stringent (J. Clapp, 2002). The dominant e-waste story line about rich countries and their citizens dumping their toxic e-waste on the poor countries such as Nigeria, Ghana, China, India, and Pakistan trades on a popular imagination of the pollution haven hypothesis. However, scholarly investigations of the hypothesis have shown inconclusive results (Lepawsky and McNabb, 2010). Some studies find support for the hypothesis while others do not, and some have documented improvements in various measures of environmental quality as foreign direct investment has increased over time.

Collectively, then, the assessment of the pollution haven hypothesis remains equivocal. Part of the reason for this equivocation relates to how the pollution haven hypothesis is typically framed (J. Clapp, 2002). Few studies, for example, examine waste processing and disposal industries. Instead the focus is on manufacturing and then often there is no comparative differentiation between industries that are more pollution intensive than others. Pollution haven studies also tend to measure direct pollution abatement costs while foregoing other important related costs (such as insurance, certification, and public relations) but which are difficult to measure across and between industrial sectors.

An examination of recently available data highlights the complexity of the international trade in electronic discards (**Figure 1**). For example, the largest flows are dominated by trade from the United States to Mexico, Canada, and the Republic of Korea (South Korea). The United States also exhibits a host of much smaller trade outflows to countries in Europe, but also in Central and South America (e.g., Nicaragua, Ecuador, and Brazil) and Africa (e.g., Nigeria, Congo, and Zambia) (**Figure 2**). However, a comparison with China highlights additional complexity to the story (**Figure 3**). China exports to several European countries (e.g., Germany, the United Kingdom, and Ireland), all of North America (Canada, the United States, and Mexico), but also to several Asian (e.g., Vietnam, Thailand, and Cambodia) and African countries (e.g., Nigeria, Ethiopia, and Tanzania). Despite China having a per capita income considerably higher in 2011 (The World Bank, n.d.) than some of its Asian and all of its African trade partners (making China the 'richer' country), all of these trades are permissible under the Basel Convention (as of 2013), the primary international legal arrangement governing transboundary waste flows. This is because China's trade with its Asian and African partners occurs between what the Basel Convention defines as non-Annex VII countries (non-Annex VII countries include all signatories to the Convention except for the Organization for Economic Co-operation and Development (OECD), European Union (EU), and Lichtenstein).

If attention is shifted back to North America, to Mexico, it is by far the poorest member of the North American Free Trade Agreement, yet it is also an OECD country and, as such, an Annex VII country under the Basel Convention, making it a legitimate trade partner with all other Annex VII countries (OECD, EU, and Lichtenstein) (**Figure 4**). Though dominated by imports from the United States, Mexico is also an exporter (though in much smaller volumes) to Canada, the United Kingdom, Qatar, and Guatemala – the latter two countries being non-Annex VII and thus, barring extant bilateral trade agreements with Mexico, illegal recipients of e-waste exports from Mexico under the Basel Convention.

These examples show that the international trade in discarded electronics does not conform to a simple image of

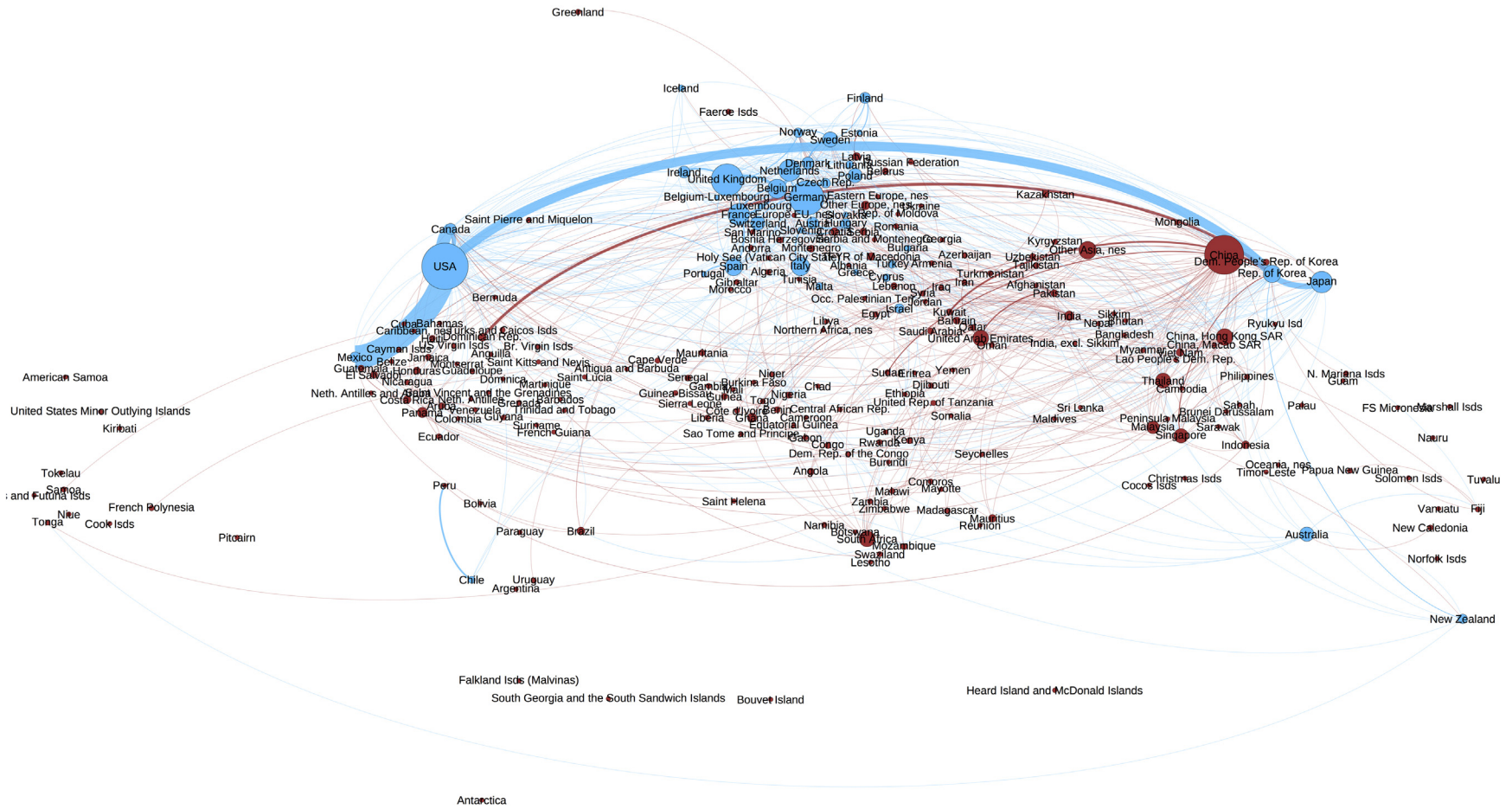

Figure 1 International trade flows of e-waste in 2010. Flow lines are read clockwise from source country to trade partner(s). Width of flow lines is proportional to weight in metric tons. Size of countries is proportional to number of export partners. Color of countries refers to their status under the Basel Convention: blue are Annex VII, red are non-Annex VII. Color of flow lines indicates the Annex VII status of a given flow's source country. See Lepawsky, J., McNabb, C., 2010. Mapping international flows of electronic waste. The Canadian Geographer 54 (2), 177–195 and Lepawsky, J., forthcoming. The changing geography of global trade in electronic discards: time to rethink the e-waste problem. The Geographical Journal. http://dx.doi.org/10.1111/geoj.12077 for discussion of data source, issues, and limitations.

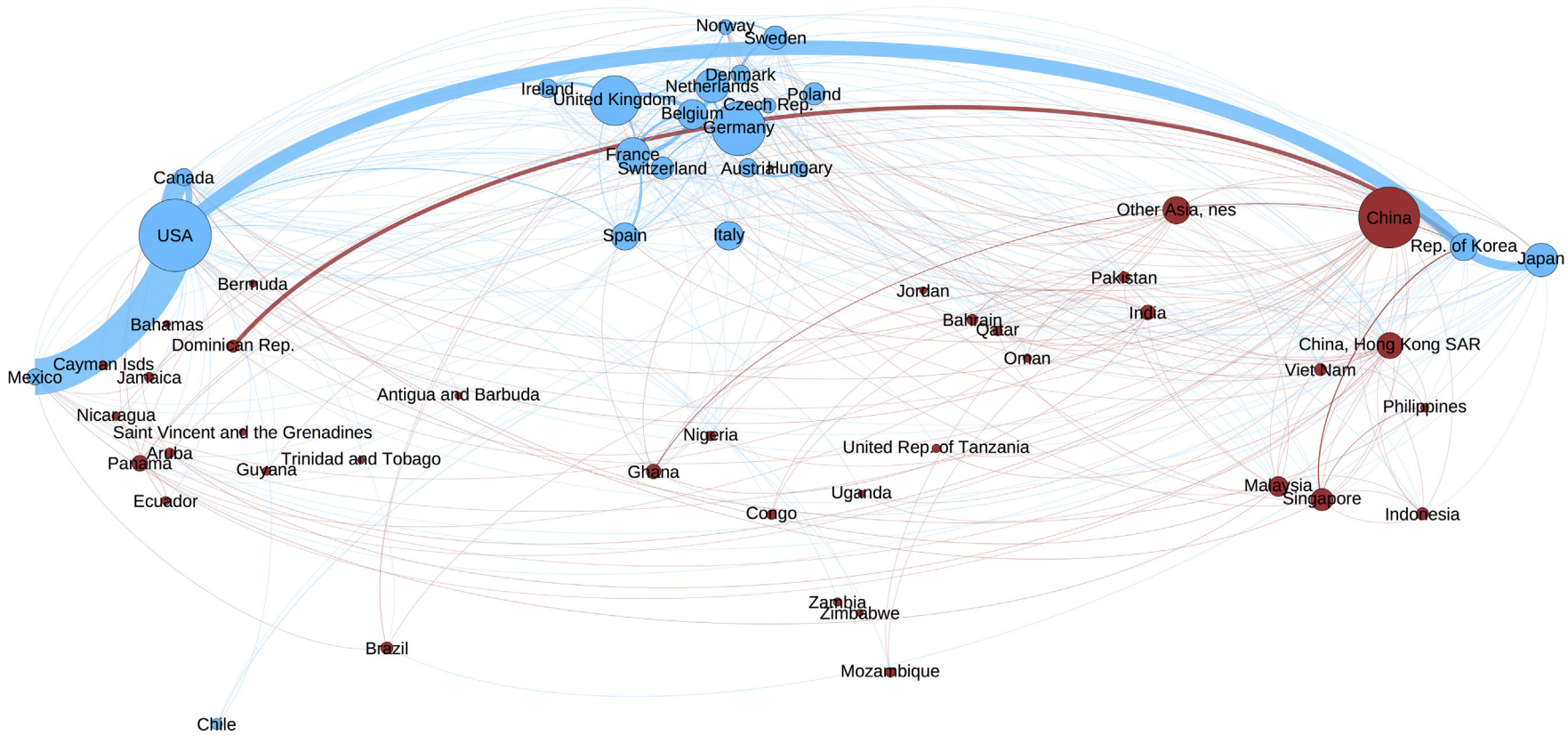

Figure 2 USA's e-waste trade network, 2010. See Lepawsky, J., McNabb, C., 2010. Mapping international flows of electronic waste. The Canadian Geographer 54 (2), 177–195 and Lepawsky, J., forthcoming. The changing geography of global trade in electronic discards: time to rethink the e-waste problem. The Geographical Journal. http://dx.doi.org/10.1111/geoj.12077 for discussion of data source, issues, and limitations.

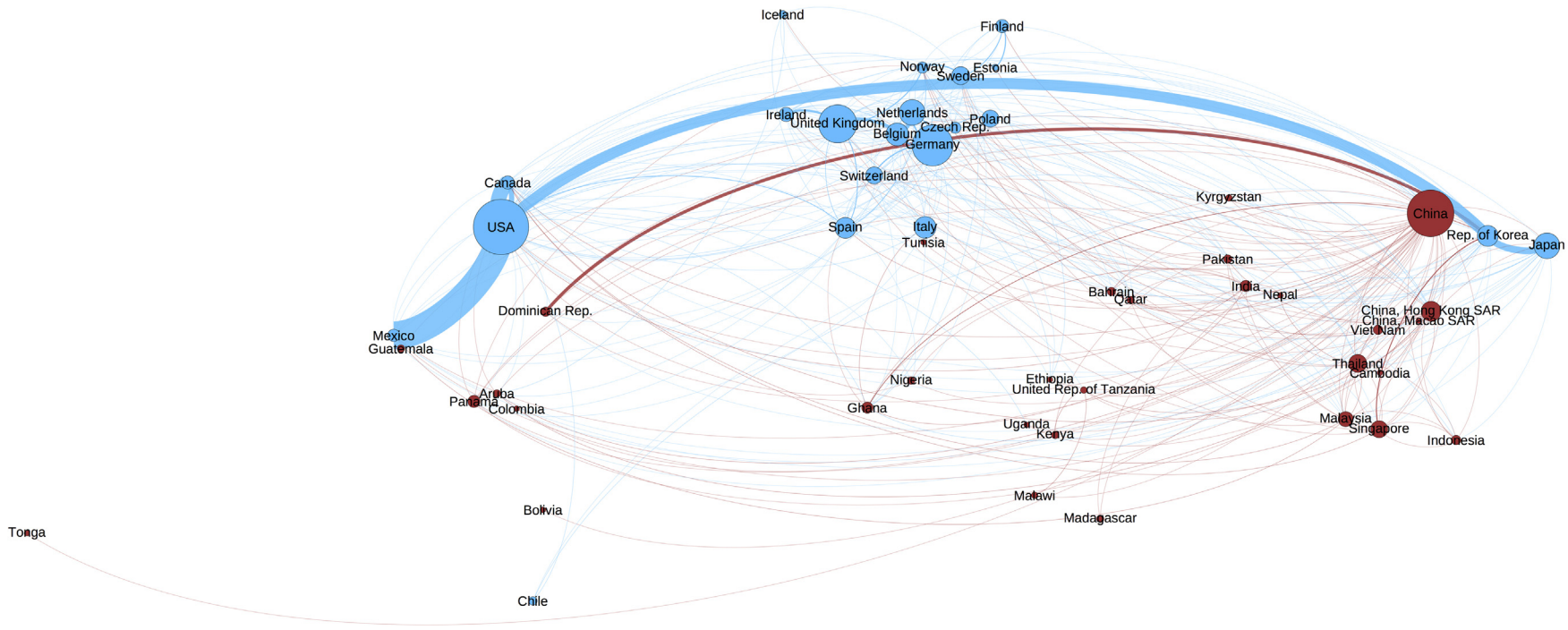

Figure 3 China's e-waste trade network, 2010. See Lepawsky, J., McNabb, C., 2010. Mapping international flows of electronic waste. The Canadian Geographer 54 (2), 177–195 and Lepawsky, J., forthcoming. The changing geography of global trade in electronic discards: time to rethink the e-waste problem. The Geographical Journal. http://dx.doi.org/10.1111/geoj.12077 for discussion of data source, issues, and limitations.

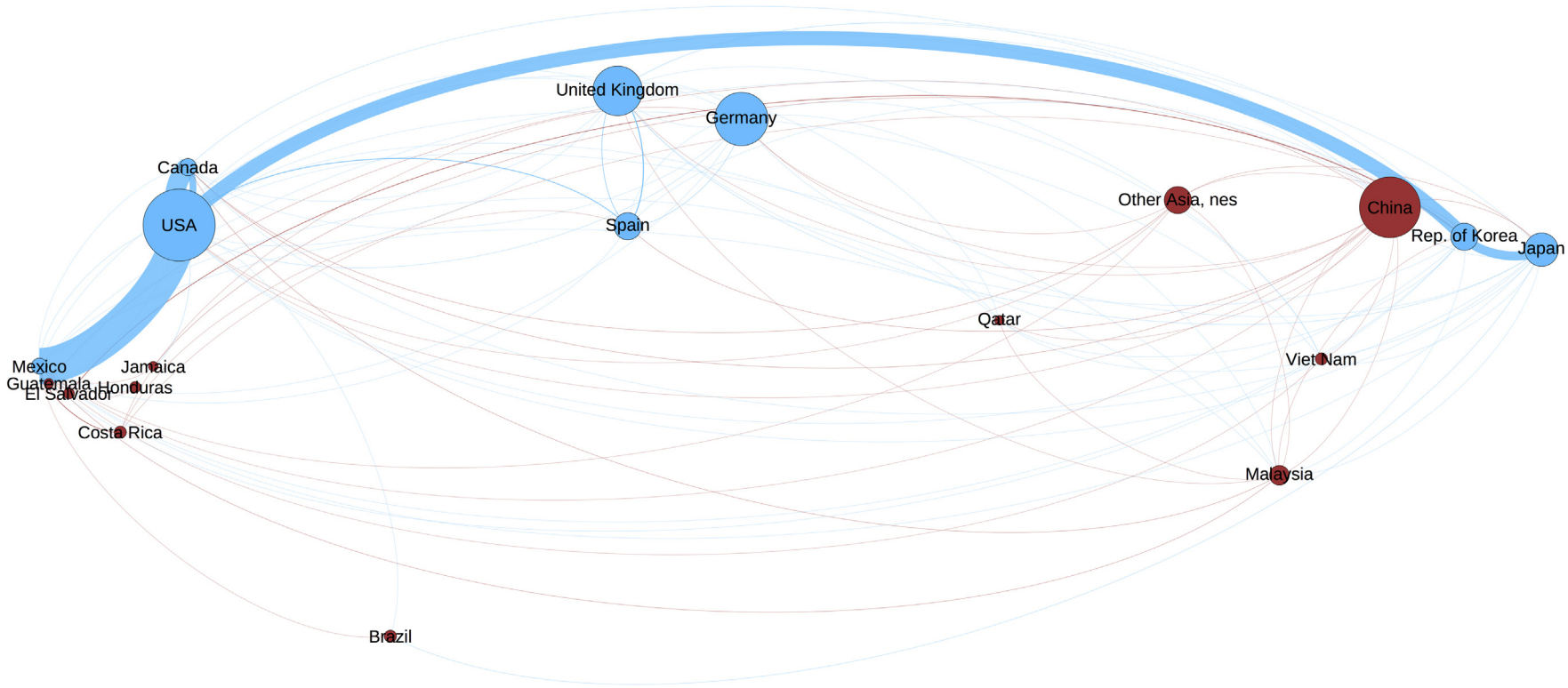

Figure 4 Mexico's e-waste trade network, 2010. See Lepawsky, J., McNabb, C., 2010. Mapping international flows of electronic waste. The Canadian Geographer 54 (2), 177–195 and Lepawsky, J., forthcoming. The changing geography of global trade in electronic discards: time to rethink the e-waste problem. The Geographical Journal. http://dx.doi.org/10.1111/geoj.12077 for discussion of data source, issues, and limitations.

high-income countries dumping their waste in low-income countries. The next section of this article further explores the complexities involved.

Situated Histories and Geographies of 'the E-Waste Problem'

Discarded electronics have a history as 'e-waste' that is quite recent. It was reports from ENGOs that brought widespread attention to the issue (e.g., Basel Action Network, 2002; Toxic Link India, 2003). There was little, if any, public discussion of a thing called e-waste prior to these reports. Instead, electronics that were rubbish in Thompson's (1979) sense of the term were seen as a potential source of strategic metals and treated as scrap, rather than waste. As early as the 1970s, discarded electronics were being investigated not as waste but as sources of recoverable common and precious metals (e.g., aluminum, copper, and silver; Dannenburg et al., 1972; Kenahan et al., 1973; Kleespies et al., 1970). During this period, materials engineers were publishing technical papers on metal recovery from electronic scrap and the US Bureau of Mines experimented with precious metal recovery from electronic scrap from military and civilian sources.

Multiple factors motivated this research, including the passage in the United States of the Solid Waste Disposal Act of 1965 as well as concerns over resource scarcity. An important geographical and historical shift in what discarded electronics are occurred: what was once practiced as a resource in the United States decades later came to be practiced as waste there and elsewhere (e.g., EU). How and why this discursive shift occurred has yet to be studied in relation to e-waste, but it deserves scholarly attention since such an inquiry is likely to reveal much about how modes of wasting work and about the cultural, political, and economic practices that shape them.

There are recent hints that the nature of electronic discard practices is turning once again toward enacting it as resource (Schluep et al., 2009; Wang et al., 2012), but the organization of these practices is much different than those of the 1970s. Whereas then, manufacturing, consumption, disposal, and resource recovery was being done predominantly within North America and Europe, today a more complicated geography of value capture and creation is emerging vis-à-vis e-waste. Instead of approaching e-waste as a linear and necessary endpoint in a chain of production, consumption, and subsequent disposal, electronic discards circulate in nonlinear webs of possible wasting and valuing (Lepawsky and McNabb, 2010). Trade in such discards is international and largely undermines the common images associated with e-waste of it flowing from high-income countries to low-income countries. The bulk of such trade occurs within, rather than between these regions. It also sustains vast recovery economies (or 'good enough' markets, see Gadiesh et al., 2007) of resale, repair, refurbishment, remanufacturing, and reuse. In Dhaka, Bangladesh, for example, estimates suggest that as many as 60 000 people are directly engaged in the electronic discard recovery economy and the bulk of those discards are derived from within Asia (Lepawsky and Billah, 2011). Processing electronic discards under certain conditions can indeed be risky for workers' health and the environment, but this activity also comes with a wide range of benefits that cannot be ignored. Such benefits include broader access to technology, technological upgrading, skill and knowledge transfer, income, and the reuse of electronics, which dampens demand for raw material extraction. These potential benefits do not necessarily outweigh the risks, but they should be considered in conjunction with concerns about poverty, inequality, and justice.

Much evidence indicates that electronic discards from countries in North America and Europe end up in poorer countries of the global South. The dominant e-waste story line narrated by ENGOs and much media in North America and Europe (e.g., CBS, 2009; Höges, 2009; Klein, 2009) mutually reinforce this understanding of e-waste. This version of the story relies on evidence such as asset tags (labels of ownership that remain attached to discarded electronics) and forensic data analysis of hard drives found in processing sites. Without disputing the veracity of such evidence, questions are being raised about how representative it is of the range of sources from which electronic discards are derived. Yu et al. (2010), for example, offer projections showing that by 2015 the developing world will surpass the developed world in e-waste generation; and a recent empirical study in Ghana estimates that of the 'second hand' electronics imported into the country 15% are nonworking and enter the informal recycling system, whereas 85% are working and move into the resale, repair, and refurbishment markets (Amoyaw-Osei et al., 2011: pp. x–xi). The report also suggests that more than 50% of the total e-waste in the country is derived from domestic sources in Ghana. Similar results have been found in Dhaka, Bangladesh, where intraregional imports from elsewhere in Asia exceed those from North America and Europe (Lepawsky and Billah, 2011). Moreover, these imports sustain dynamic electronics discards recovery economies premised on resale, repair, refurbishment, and remanufacturing – an issue also highlighted by recent research at Agbogbloshie in Accra, Ghana (Grant and Oteng-Ababio, 2012; Oteng-Ababio, 2012), one of the sites singled out by ENGOs and media to support the dominant e-waste story line that focuses on dumping.

None of this is to suggest the dominant story line about e-waste narrated by ENGOs and media is entirely false, but it does trade on a number of unwarranted assumptions about e-waste as a matter of concern. These assumptions include that North American and European sources of electronic discards are more important than intraregional and/or domestic sources for the countries in which e-waste processing is occurring. There is, then, good reason to reconsider the dominant e-waste story line about rich Northern countries dumping e-waste in poor Southern ones.

Moreover, there is no necessary reason to restrict our understanding of e-waste to what happens to electronic devices after they are consumed and discarded. Indeed, doing so blinds us to crucial issues of waste generation from electronics manufacturing that occur long before consumers even purchase the devices of which they may eventually rid themselves. As early as the 1980s, for example, important work on occupational health and safety in the electronics manufacturing sector began to emerge that raised serious concerns about the health consequences of working with chemicals commonly used in the industry that make their way into workers' bodies and groundwater (Beall et al., 2005; Bender et al., 2007;

R.W. Clapp, 2006; LaDou, 1984, 1985, 1986). These chemicals do not become part of the manufactured item and as such are a remainder of manufacturing that must be managed.

Concerns over occupational health effects from chemicals used in the manufacture of computer chips led to the formation of the Santa Clara Center for Occupational Health (SCCOSH) in the late 1970s. SCCOSH later initiated a project called the Silicon Valley Toxics Coalition (SVTC, which later became an independent organization) in 1982 after evidence began to emerge that semiconductor manufacturing plants were releasing toxic chemicals into the valley's groundwater (Gabrys, 2011; see also Smith et al., 2006). As a consequence of SCCOSH's and SVTC's advocacy, the US Environmental Protection Agency eventually identified more than 20 sites in the valley that met the criteria to be defined as Superfund sites for cleanup. Silicon Valley thereby gained the dubious distinction of the most Superfund sites of any locale in the country.

In the 1990s, the advocacy work of SVTC expanded to include the Campaign for Responsible Technology (now the International Campaign for Responsible Technology), which expanded activism not only around toxins, but also groundwater extraction for chip manufacturing, especially in the expanding industry of the US Southwest. The 'clean-industry' image that had to that point been part of the Silicon Valley mythology was crumbling as a consequence, but notably these issues of toxic release and groundwater extraction were not framed as 'e-waste' as that term is typically used today. Others have examined the material and energy inputs of electronics manufacture (Kahhat et al., 2012; Kuehr and Williams, 2003; Williams et al., 2002, 2004) and note that this industry is an energy and materially intensive one, the products of which exceed by orders of magnitude the energy and material inputs of products in other industries, such as automobile manufacturing. Again, however, this use of energy and materials is not framed as wasteful, nor is it visible in the dominant e-waste story line.

Scholarly Debates

The contemporary scholarly debates about e-waste have developed several prominent foci including toxicology, management and processing techniques, and regulation. The toxicology literature has, in the first decade of the twenty-first century, documented a wide range of negative health and environmental outcomes associated with certain techniques of processing discarded electronics for their parts and materials. Potential health outcomes include toxic body loading (Ni et al., 2010; Tue et al., 2010), neonatal behavioral change (Y. Li et al., 2008, 2010), and diffusion of toxins into food supplies (Chan et al., 2012; X.-B. Zheng et al., 2012). Examples of negative environmental impacts include accumulation of toxins in proximate and distant ecosystems (F. Wang et al., 2009; J.P. Wu et al., 2011; H. Yang et al., 2012).

While there is no doubt that when processed under particular conditions discarded electronics are a source of toxic hazards, there are important controversies around this issue. For example, the processing and management literature typically confines itself to high-technology industrial processing techniques. Meanwhile, almost no research exists documenting and comparing different artisanal (often called 'primitive' in the dominant e-waste story line) processing techniques. Some of these techniques, such as open burning, do indeed result in highly toxic remainders. Images of open burning and open container hydrometallurgy (in which acid solutions are used to extract metal ores and which can easily lead to spills of toxic substances) often provide the pictorial evidence of the negative aspects of e-waste processing in ENGO documents and journalistic accounts. Yet such techniques are not used universally. Other forms of artisanal smelting such as the *bhati* process used in Bangladesh, for example, adapt technologies originally developed for smelting and metalworking to processing discarded electronics. This process involves small enclosed clay furnaces which, because they result in much higher temperatures than open burning, may lead to less toxic residues and emissions being produced (though this possibility has yet to be studied).

Questions also arise about the desirability and environmental consequences of industrial electronics recycling (Barba-Gutiérrez et al., 2008). Transport for collection and processing has its own environmental footprint. So too do the energy and material requirements of industrial scale recycling which typically shreds electronics in order to sort them into recoverable materials. Such practices destroy what, given the right conditions, may be perfectly workable equipment and components. In the process, they also result in the loss of all the energy embodied in manufactured electronics derived from, among other things, the mining of raw materials, the supply logistics of manufacturing inputs, and labor that went into building the electronics that are, after discard, subsequently run through a shredding machine and destroyed.

Another important theme in the e-waste literature is how it is or should be regulated and how transitions to more sustainable forms of e-waste management can be made. The main debates focus on what constitutes proper legal compliance, how responsibility for managing e-waste should be distributed, and how power inequities may be better adjudicated in transitions to sustainable e-waste management (Lawhon, 2012). The focus of the compliance literature is often the Basel Convention. Much controversy persists around interpretations of the Basel Convention, in large part due to ambiguities created by the Basel Convention itself. For example, contradictions exist between the definition of hazardous waste used in the Basel Convention and the definitions used by signatory countries. The Basel Convention contains provisions that allow transboundary shipments of hazardous waste between signatories as well as signatories and nonsignatories under certain conditions (i.e., when bilateral or multilateral trade agreements exist between countries). It simultaneously lists materials considered to be hazardous, many of which are relevant to electronics, and exempts many of those same materials because no standard tests exist to decide their toxic potential.

Further, the Basel Convention imagines a world divided up into two monolithic blocs of countries (Annex VII countries comprising the OECD, EU, and Lichtenstein vs non-Annex VII countries comprising all other signatories) in which the latter, non-Annex VII countries, are construed as all equally vulnerable to hazardous waste shipments from the former. Yet, the world of trade no longer conforms to such a division. Today (2013), far more discarded electronics are traded

between non-Annex VII countries than between Annex VII and non-Annex VII countries. Consequently, under the Convention countries as radically different as China, India, Singapore, and Kenya are lumped together as equally vulnerable to trade from Annex VII countries but not trade amongst themselves. Thus, under the Basel Convention it is perfectly legal for Kenya to receive hazardous waste from China, India, or Singapore.

Adjudicating the responsibility for managing discarded electronics is another important area of controversy in the scholarly literature (Lepawsky, 2012). Laws premised on extended producer responsibility (EPR) have been passed in many jurisdictions specifically to manage e-waste, though it is not a form of waste governance unique to managing electronic discards. EPR is supposed to shift the financial and/or physical responsibility for managing discards to those that produce them, in other words, the brand manufacturers of electronic products. The logic behind the approach is that shifting responsibility to producers will force them to internalize the costs of managing disposal of their products and thus incentivize them to create products that last longer and/or can be more readily recycled. In practice, however, EPR legislation for electronic discards rarely if ever achieves such cost internalization. Instead, the costs are borne by consumers of such technologies who, in an increasing number of jurisdictions, are required to pay fees at the time of purchase to cover the costs of subsequent disposal of their devices. This result is not necessarily negative in itself since an argument can be made that those who consume a particular class of manufactures should share some responsibility for the costs of their eventual disposal. Yet, it is not a coincidence that these costs have successfully been kept external from brand manufacturers who are not interested in seeing such costs negatively effect their profit margins.

While the distribution of responsibility for managing electronics discards is one source of controversy around their broader management, there is another more subtle but arguably more important issue at stake: what is and is not open to democratic debate. Despite the variety of e-waste legislation all but a narrow range focus exclusively on managing waste once it is already produced. In other words, the contemporary public discussion of what to do about e-waste is largely confined to what to do after it already exists. Almost no public discussion or regulation about what happens in the design or manufacturing of electronics occurs. An important, though limited, exception to this is the European Union's Reduction of Hazardous Substances (RoHS) directive. The RoHS directive does mandate against the use of specific materials with toxic potential (e.g., lead and mercury) in electronic products and so does affect how brand manufacturers design and build their products to a certain extent. But, mandating against such materials is a limited form of ameliorating the problem of electronic discards. For example, it does nothing to increase the likelihood that electronics will be longer lasting, more reusable, more repairable, or more recyclable.

The limitations of regulatory approaches like RoHS or the myriad e-waste laws enacted over the last decade is illustrated by recent design changes to Apple laptops highlighted by iFixit (Kyle, 2012), a website dedicated to do-it-yourself repair and reuse. In a step-by-step 'teardown' (i.e., disassembly to illustrate repairability) of a recent model of MacBook the website notes design changes from older models that include proprietary screws to prevent user access to the machine's internal parts, RAM chips soldered to the logic board making it impossible to swap in upgrades, a battery glued into the case (making it at best difficult to replace the battery when it fails), and a display that is a completely fused unit making it impossible to repair or replace parts of the display should they fail. It is notable that all of these design changes, which reduce the repairability and upgradability of these laptops were made after RoHS and EPR e-waste legislation have been in place in the EU and some North American jurisdictions for a minimum of 10 years.

Conclusion

Whereas the dominant story line about e-waste in the early twenty-first century is one of dumping in low-income countries by high-income countries, its historical geography is quite different. As a potential source of strategic metals, emphasis was placed on sources of electronics in North America, particularly in cities (Kenahan et al., 1973). As waste from electronics manufacturing, attention was concentrated on Silicon Valley, California, and some parts of the Southwest United States, especially in terms of groundwater extraction and contamination. All of this is to say that the things that now typically comprise e-waste as a matter of concern are rather different than those that comprised its historical–geographical antecedents. Discarded electronics are not the same thing now as they were in the past. Nor do they have the same geographies. Their ontology as waste is fluid, rather than fixed. What these different enactments of waste from electronics suggest is that the contemporary framing of e-waste as necessarily a postconsumption phenomenon is a highly particular framing of it, rather than a necessary one. Paying attention to the way that e-waste is framed as a matter of concern is critical, because that framing plays an important role in how the problem is construed as a problem and, consequently, in framing the solutions proposed to alleviate it.

Framing e-waste as a postconsumption problem means discussion and management practices are restricted to the issue of what to do with waste after the products that will constitute it have already been manufactured (see Gille, 2007). More effective solutions to the e-waste problem, even as it is so framed, requires reorganizing relations between production, consumption, and disposal. If the public discussion of e-waste remains confined to what to do with electronic discards after they already exist, then the prospects for enacting a safer world will be very limited. Individual consumers, households, businesses, and industries may divert more and more electronics away from landfills, but the longevity of those electronics, the ease with which they can be reused, repaired, refurbished, or recycled, the toxicity of the materials with which they are made, the conditions under which people labor to produce them, or the material remainders of manufacturing itself are unlikely to improve. Indeed, if the public discussion remains oriented toward end-of-pipe solutions, then, as Gille (2007: p. 210) argues, "the most that democracy can achieve is to regulate what to do with the wastes already produced." If a goal in solving the e-waste problem is to enact a world that is safer in terms of toxics and fairer in terms of labor conditions,

then the public discussion about e-waste *qua* waste must escape the end of the pipe and flow through the factory gates and into a discussion about how and with what materials that which will eventually be discarded is manufactured before it ever leaves the factory floor. Though such democratization of production, as Gille (2007) calls it, might sound utopic, experiments with regulating production and its discards along these lines are underway in other industries (such as fair trade certification of foods).

Though certainly problematic, such regulatory experiments demonstrate that the entrance of public discussion into conditions of production can and does occur. What such politics of production, consumption, and disposal tell us is that practices that enact electronic discards as waste (or as resource or as something else altogether) are 'worldings' (Haraway, 2008; Roy and Ong, 2011; Tsing, 2005), that is, actions that gather together otherwise unlike things – such as circuit boards, monitors, and CPUs with corporations, consumers, citizens, workers, labor conditions, and moral injunctions ('do the right thing') – and enacts them as if they all inhabit the same common world. Such worldings vis-à-vis waste cannot help but raise fundamental questions about what is good or what is to be done (Hawkins, 2006). E-waste *qua* waste, then, is an entry point into fundamental issues about learning how to live together well.

See also: Economics and Ethics; Justice: Philosophical Aspects; Responsibility: Philosophical Aspects.

Bibliography

Amoyaw-Osei, Y., et al., 2011. Ghana e-Waste Country Assessment, Secretariat of the Basel Convention. Available at: http://www.ewasteguide.info/Amoyaw-Osei_2011_GreenAd-Empa (accessed 19.04.11.).

Barba-Gutiérrez, Y., Adenso-Díaz, B., Hopp, M., 2008. An analysis of some environmental consequences of European electrical and electronic waste regulation. Resources, Conservation and Recycling 52 (3), 481–495.

Basel Action Network, 2002. Exporting Harm: The High-tech Trashing of Asia. Available at: http://www.ban.org/main/library.html (accessed 27.10.05.).

Beall, C., et al., 2005. Mortality among semiconductor and storage device-manufacturing workers. Journal of Occupational and Environmental Medicine 47 (10), 996–1014.

Bender, T.J., et al., 2007. Cancer incidence among semiconductor and electronic storage device workers. Occupational and Environmental Medicine 64 (1), 30–36.

CBS, 2009. Following the Trail of Toxic E-waste – 60 Minutes. CBS News. Available at: http://www.cbsnews.com/stories/2008/11/06/60minutes/main4579229.shtml (accessed 25.03.11.).

Chan, J.K.Y., et al., 2012. Dietary intake of PBDEs of residents at two major electronic waste recycling sites in China. Science of the Total Environment 463–464, 1138–1146.

Clapp, J., 2002. What the pollution havens debate overlooks. Global Environmental Politics 2 (2), 11–19.

Clapp, R.W., 2006. Mortality among US employees of a large computer manufacturing company: 1969–2001. Environmental Health: A Global Access Science Source 5 (30), 1–10.

Dannenburg, R.O., Maurice, J.M., Potter, G.M., 1972. Recovery of Precious Metal from Electronic Scrap. United States Department of the Interior, Bureau of Mines, Salt Lake City, UT.

Gabrys, J., 2011. Digital Rubbish: A Natural History of Electronics. University of Michigan Press, Ann Arbor.

Gadiesh, O., Leung, P., Vestring, T., 2007. The battle for China's good-enough market. Harvard Business Review 85 (9), 80–89.

Gille, Z., 2007. From the Cult of Waste to the Trash Heap of History: The Politics of Waste in Socialist and Postsocialist Hungary. Indiana University Press, Bloomington, IN.

Grant, R., Oteng-Ababio, M., 2012. Mapping the invisible and real "African" economy: urban e-waste circuitry. Urban Geography 33 (1), 1–21.

Hajer, M.A., 1995. The Politics of Environmental Discourse. Oxford University Press, Oxford, New York.

Haraway, D., 2008. When Species Meet. University of Minesota Press, Minneapolis, MN.

Hawkins, G., 2006. The Ethics of Waste: How We Relate to Rubbish. Rowman & Littlefield, Lanham, MD.

Höges, C., 2009. The Children of Sodom and Gomorrah: How Europe's Discarded Computers are Poisoning Africa's Kids. Available at: http://www.spiegel.de/international/world/the-children-of-sodom-and-gomorrah-how-europe-s-discarded-computers-are-poisoning-africa-s-kids-a-665061.html (accessed 18.02.14.).

Kahhat, R., Hieronymi, K., Williams, E. (Eds.), 2012. E-Waste Management: From Waste to Resource (Hardback). Taylor & Francis, Routledge; New York.

Kenahan, C.B., et al., 1973. Bureau of Mines Research Programs on Recycling and Disposal of Mineral-, Metal-, and Energy-Based Wastes. US Bureau of Mines. Inforamtion Circular, (8595), Washington, DC.

Kleespies, E.K., Bennetts, J.P., Henrie, T.A., 1970. Gold recovery from scrap electronic solders by fused-salt electrolysis. Journal of Metals 22 (1), 42–44.

Klein, P., 2009. Frontline/World Ghana: Digital Dumping Ground. PBS. Available at: http://www.pbs.org/frontlineworld/stories/ghana804/resources/ewaste.html (accessed 17.02.10.).

Kuehr, R., Williams, E., 2003. Computers and the Environment: Understanding and Managing Their Impacts. Kluwer Academic Publishers, Dordrecht.

Kyle, 2012. MacBook Pro with Retina Display Teardown – IFixit. Available at: http://ifixit.org/2753/macbook-pro-with-retina-display-teardown/ (accessed 21.08.12.).

LaDou, J., 1984. The not-so-clean business of making chips. Technology Review 87 (4), 22–36.

LaDou, J., 1985. Water contamination in the Silicon Valley. Journal of Occupational and Environmental Medicine 27 (4), 304.

LaDou, J., 1986. Health issues in the microelectronics industry. Occupational Medicine-State of the Art Reviews 1 (1), 1–11.

Lawhon, M., 2012. Relational power in the governance of a South African e-waste transition. Environment and Planning A 44 (4), 954–971.

Lepawsky, J., 2012. Legal geographies of e-waste legislation in Canada and the US: jurisdiction, responsibility and the taboo of production. Geoforum 43 (6), 1194–1206.

Lepawsky, J., forthcoming. The changing geography of global trade in electronic discards: time to rethink the e-waste problem. The Geographical Journal. http://dx.doi.org/10.1111/geoj.12077.

Lepawsky, J., Billah, M., 2011. Making chains that (un)make things: waste-value relations and the Bangladeshi rubbish electronics industry. Geografiska Annaler: Series B, Human Geography 93 (2), 121–139.

Lepawsky, J., McNabb, C., 2010. Mapping international flows of electronic waste. The Canadian Geographer 54 (2), 177–195.

Li, Y., et al., 2008. Monitoring of lead load and its effect on neonatal behavioral neurological assessment scores in Guiyu, an electronic waste recycling town in China. Journal of Environmental Monitoring 10 (10), 1233–1238.

Li, Y., et al., 2010. Assessment of cadmium exposure for neonates in Guiyu, an electronic waste pollution site of China. Environmental Monitoring and Assessment 177 (1–4), 343–351.

Ni, H.-G., et al., 2010. Environmental and human exposure to persistent halogenated compounds derived from e-waste in China. Environmental Toxicology and Chemistry 29 (6), 1237–1247. http://dx.doi.org/10.1002/etc.160.

Oteng-Ababio, M., 2012. When necessity begets ingenuity: e-waste scavenging as a livelihood strategy in Accra, Ghana. African Studies Quarterly 13 (1–2), 1–21.

Roy, A., Ong, A., 2011. Worlding Cities: Asian Experiments and the Art of Being Global. Wiley-Blackwell, West Sussex, Chichester; Malden, MA.

Schluep, Mathias, et al., 2009. Recycling – from E-Waste to Resources. United Nations. Available at: http://www.unep.org/PDF/PressReleases/E-Waste_publication_screen_FINALVERSION-sml.pdf (accessed 24.02.10.).

Smith, T., Sonnenfeld, D.A., Pellow, D.N., 2006. Challenging the Chip: Labor Rights and Environmental Justice in the Global Electronics Industry. Temple University Press, Philadelphia, PA.

The World Bank. GDP per capita (current US$) – Data – Graph. Available at: http://data.worldbank.org/indicator/NY.GDP.PCAP.CD/countries/CN-VN-TH-KH-NG-ET-TZ?display=graph (accessed 22.08.12.), n.d.

Thompson, M., 1979. Rubbish Theory: The Creation and Destructison of Value. Oxford University Press, Oxford, New York.

Toxic Link India, 2003. Scrapping the High-Tech Myth: Computer Waste in India. Toxic Link India, New Delhi.

Tsing, A.L., 2005. Friction: An Ethnography of Global Connection. Princeton University Press, Princeton, NJ.

Tue, N.M., et al., 2010. Accumulation of polychlorinated biphenyls and brominated flame retardants in breast milk from women living in Vietnamese e-waste recycling sites. Science of the Total Environment 408 (9), 2155–2162.

Wang, F., et al., 2009. Chemical and ecotoxicological analyses of sediments and elutriates of contaminated rivers due to e-waste recycling activities using a diverse battery of bioassays. Environmental Pollution 157 (7), 2082–2090.
Wang, F., et al., 2012. The Best-of-2-Worlds philosophy: developing local dismantling and global infrastructure network for sustainable e-waste treatment in emerging economies. Waste Management 32, 2134–2146.
Williams, E., Ayres, R.U., Heller, M., 2002. The 1.7 kilogram microchip: energy and material use in the production of semiconductor devices. Environmental Science & Technology 36 (24), 5504–5510.
Williams, E., Ayres, R.U., Heller, M., 2004. Response to comment on "The 1.7 kilogram microchip: energy and material use in the production of semiconductor devices". Environmental Science & Technology 38 (6), 1916–1917.
Wu, J.P., et al., 2011. Bioaccumulation potentials of polychlorinated biphenyls in bottom fishes from an e-waste recycling site in South China. China Environmental Science 31 (4), 637–641.
Yang, H., et al., 2012. Distribution, historical trends and inventories of polychlorinated biphenyls in sediments from Yangtze River Estuary and adjacent East China Sea. Environmental Pollution 169, 20–26.
Yu, J., et al., 2010. Forecasting global generation of obsolete personal computers. Environmental Science & Technology 44 (9), 3232–3237.
Zheng, X.-B., et al., 2012. Halogenated flame retardants in home-produced eggs from an electronic waste recycling region in South China: levels, composition profiles, and human dietary exposure assessment. Environment International 45 (1), 122–128.

Relevant Websites

http://www.ban.org/ – Basel Action Network.
http://www.electronicstakeback.com/home/ – Electronics Takeback Coalition.
http://retroworks.blogspot.ca/ – Ethical Electronics Exports, Fair Trade Recycling.
http://ewasteguide.info/ – Ewasteguide.
http://ifixit.org/ – iFixit.org.
http://www.icrt.co/ – International Campaign for Responsible Technology.
http://www.r2solutions.org/ – R2 Solutions.
http://svtc.org/ – Silicon Valley Toxics Coalition.
http://www.step-initiative.org/ – Solve the E-Waste Problem.

Eliade, Mircea (1907–86)

Robert A Segal, Lancaster University, Lancaster, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 7, pp. 4405–4409, © 2001, Elsevier Ltd.

Abstract

Mircea Eliade was a Romanian-born historian of religions who spent his last three decades at the University of Chicago Divinity School. He was one of the last great comparativists and was celebrated for his writings on the myths, rituals, and symbols of religions around the world. He was most celebrated for his lifelong defense of the approach to religion that characterizes the discipline of religious studies. Eliade maintains that religion originates to serve a distinctly religious end: providing contact with God. He assumes that, by contrast, all of the social sciences account for religion non-religiously: as serving some sociological, anthropological, psychological, or economic end. Eliade advocates the religious account over any social scientific account on two grounds: that the religious account coincides with the account offered by believers themselves, and, even more, that it alone satisfactorily accounts for, so to speak, the religious nature of religion. Eliade was influential outside of religious studies as well as within it. His attempt to find religiosity in the beliefs and activities of professedly atheistic moderns was especially influential.

Mircea Eliade was the best-known contemporary theorist of religion from the discipline of religious studies. He was one of a handful of scholars within the discipline who influenced scholars outside it. Unlike specialists in a single religion, he was influential as a theorist of religion as a whole. While he wrote specialized tracts on Australian religions, Indian religions, alchemy, yoga, shamanism, and the religion of his native Romania, he was most celebrated as one of the last comparativists and, even more, as the latest defender of the defining tradition of religious studies. Eliade sought to demonstrate the universality of religion by finding it in even seemingly secular domains. He also wrote fiction, and in Romania was a popular novelist and short story writer.

Life

Mircea Eliade was born in Bucharest on March 9, 1907. He began writing for Romanian weeklies as a teenager. In 1925, after completing the *lycée*, he entered the University of Bucharest and received an MA for his thesis on Italian philosophy from Ficino to Bruno. His mentor was the philosopher Nae Ionescu. While at university, he continued writing regularly – short stories, scholarly articles, reviews, and translations. He also founded a journal.

In 1928 Eliade went to India to study Sanskrit and Indian philosophy with Surendranath Dasgupta of the University of Calcutta. He aspired to become sufficiently trained linguistically to secure a professorship in comparative religion at the University of Bucharest – an especially ambitious goal since no chair in the field existed at the University at the time. Daring to hope that one would be created for him, he immediately immersed himself in Sanskrit grammar. He also discovered the world-affirming variety of yoga known as Tantrism. This outlook came to express Eliade's view of religion universally: that the sacred is to be found within the profane rather than outside it.

Impressed by Eliade's diligence and enthusiasm, Dasgupta soon took him into his household and treated him almost as an adopted son. When Dasgupta discovered that Eliade was having an affair with his 16-year-old daughter, Maitreyi, under his own roof, he threw Eliade out of his house, and the two parted ways forever. Undeterred, Eliade proceeded to spend six months in the ashram of Rishiskesh, learning the practice, not merely the philosophy, of yoga. Even during his time in India, which constituted his field-work, he recognized that his heart lay in comparative religion rather than in any single religion. While in India, he continued writing, turning out newspaper columns and magazine articles about his travels, scholarly articles on yoga, and several novels. After three years in India, he returned to Romania, pressured by his father to undertake his already deferred year of military service. He intended to return to India but never did.

Back home, Eliade worked on his doctoral thesis on yoga, which he had begun in India, and simultaneously published articles, wrote novels, worked for journals, gave radio talks, and organized public lectures – all on a host of topics. He came to be considered the leading cultural figure of his generation of young Romanians. In 1932 he did his year of military service, largely doing translations. In 1933 he received his doctorate. In the same year he published the novel *Maitreyi*, about his love affair with Dasgupta's daughter. The book won a prize and became a best seller. It was the Romanian literary sensation of 1933, and Eliade became famous.

At the end of 1933 Eliade began teaching at the University of Bucharest. He had hoped to be offered a lectureship but had to settle for serving as an assistant to Ionescu. He held the post for five years, lecturing on various topics in both religion and philosophy. Ionescu, whom he continued to idolize, was the intellectual mentor of the Iron Guard, the Romanian counterpart to the Nazis , and wrote diatribes against Jews. Proclaiming that Jews deserved to suffer because they had rejected Jesus as their savior, he demanded that Romania rid itself of all Jews by any possible means. In his autobiography Eliade self-servingly emphasizes his own break with Ionescu over the status of Jews, but recent research has uncovered scores of viciously anti-Semitic newspaper articles written by Eliade himself (see Berger, 1989; see, in reply, Rennie, 1996, Chap. 13). Eliade associated Jews with everything foreign to true Romanian

culture. In stock anti-Semitic fashion, Jews were blamed for communism, materialism, democracy, and modernity. Eliade's official biographer, M L Ricketts, uncritically accepts Eliade's own presentation of his position (see Ricketts, 1988, II pp. 903–912). According to both, the fervently Romanian nationalism that Eliade espoused was spiritual rather than political. The Government nevertheless arrested and detained Eliade in 1938 because of his zealous support for the Iron Guard. Eliade was subsequently transferred to a sanitarium.

In 1940 Eliade was allowed to leave the country to serve in the Roman Legation in London, and in 1941 he was transferred to the Legation in Lisbon, where he remained until the end of World War II. He returned only once to Romania. Had he returned after the war, he would have faced arrest as a collaborator. He left for Paris in 1945, where he spent 10 years writing the scholarly tomes that made him internationally famous. In 1946 he served as a visiting professor at the *Ecole des Hautes Etudes* of the Sorbonne. In 1956 he was a guest lecturer at the Divinity School of the University of Chicago, and the next year he became a permanent Professor of the History of Religions. In 1962 he was named the Sewell S. Avery Distinguished Service Professor. He co-founded the journal *History of Religions.* He became the best-known scholar of religious studies in the world and the leading advocate of the field. He retired in 1983, at the age of 76, and was honored with the establishment of an endowed chair in his name. He died in Chicago on 22 April 1986. Shortly afterwards, there appeared the culminating opus of his career: the 16-volume *Encyclopedia of Religion*, edited by him.

Theory of Religion

The tradition perpetuated by Eliade goes back at least to the nineteenth-century comparativists Friedrich Max Müller (1823–1900) and C. P. Tiele (1830–1902) and was continued by such figures as P. D. Chantepie de las Saussaye, Nathan Söderblom, Gerardus van der Leeuw, W. Brede Kristensen, and Raffaele Pettazzoni. This tradition, alternatively called comparative religion, phenomenology of religion, the history of religions, and, most simply, religious studies, maintains that religion is religion and little, if anything, else. Accordingly, religion is not the subject of anthropology, sociology, psychology, or economics. It is the subject of religious studies alone. The focus of religious studies is less on the truth of religion than on its origin and function. Religion, according to this tradition, arises for the distinctly religious purpose of making contact with God.

The tradition is, however, ambiguous about its aims. Sometimes it claims to be merely *presenting* the adherent's point of view, which purportedly deems religion a route to God. Other times the tradition claims to be *endorsing* the adherent's point of view. Now religion really does serve to make contact with God. The tradition thereby ceases to offer a mere description of religion and now offers an explanation, or a theory, of religion. But then the issue of the truth of religion proves difficult to sidestep. For if religion succeeds in making contact with God, God must exist.

Eliade's own position is unclear. He himself wavers between presenting and endorsing the believer's point of view. Typically, he declares modestly that 'The ultimate aim of the historian of religions is to understand, and to make understandable to others, religious man's behavior and mental universe' (Eliade, 1959a, p. 162). In contrast to the theologian, from whom he repeatedly distinguishes himself, the historian seeks only to explicate, not to advocate, the adherent's point of view: 'The procedure of the historian of religions is just as different from that of the theologian. All theology implies a systematic reflection on the content of religious experience, aiming at a deeper and clearer understanding of the relationships between God–Creator and man–creature. But the historian of religions uses an empirical method of approach. He is concerned with religio-historical facts which he seeks to understand and to make intelligible to others' (Eliade, 1959a, p. 88).

Yet in actuality Eliade ventures far beyond the mere presentation of the believer's point of view. He does not hesitate to discuss dead religions, whose adherents are scarcely available for consultation. His chief source of information for even living religions is not informants but their texts and practices. Moreover, he is prepared to grant that the meaning of all religions is partly, even largely, unconscious. Of religious symbols, for example, he states that 'we do not have the right to conclude that the message of the symbols is confined to the meanings of which a certain number of individuals are fully conscious, even when we learn from a rigorous investigation of these individuals what they think of such-and-such a symbol belonging to their own tradition. Depth psychology has taught us that the symbol delivers its message and fulfills its function even when its meaning escapes awareness' (Eliade, 1959b, pp. 106–107). Eliade goes as far as to assert that even most professedly non-religious moderns are in fact religious: 'The majority of the "irreligious" still behave religiously, even though they are not aware of the fact .... [T]he modern man who feels and claims that he is non-religious still retains a large stock of camouflaged myths and degenerated rituals' (Eliade, 1959a, pp. 204–205). Eliade outright bars any exceptions: '[E]ven the most avowedly non-religious man still, in his deeper being, shares in a religiously oriented behavior' (Eliade, 1959a, p. 211). In imputing to avowed atheists a religiosity of which they are wholly unconscious and which contradicts their own conscious one, he is doing more than describing the actor's point of view. His defense of his view of religion as irreducibly religious cannot therefore be that it is so for actors themselves. It must be that religion is itself irreducibly religious. He must be offering a theory, not just a description, of religion.

The aspects of religion on which Eliade focuses are myths, rituals, and symbols. All are means of encountering God. Religion for Eliade arises not from an encounter with God but from a *yearning* for an encounter. Religion provides both places and times for the encounter. It demarcates as 'sacred' those places, or spaces, where God has previously appeared. Adherents assume that God, who is elusive, will more likely appear where God has already been than in a new place. Even if adherents believe that God is everywhere, they still go to church to seek out God. Similarly, religion sets off as sacred those times, always long past, when God most often appeared. Because adherents believe that God was closest at hand during or just after creation, they use myths about creation not merely

to tell them when God was nigh but, more, to carry them back to that time: 'It is easy to understand why the memory of that marvelous time haunted religious man, why he periodically sought to return to it. *In illo tempore* the gods had displayed their greatest powers. The cosmogony is the supreme divine manifestation, the paradigmatic act of strength, superabundance, and creativity. Religious man ... seeks to reside at the very source of primordial reality, when the world was *in statu nascendi*' (Eliade, 1959a, p. 80). Myth works like a magic carpet.

In the tradition of religious studies, Eliade maintains that the quest for God, or the sacred, is not a means to an end but the end itself. Adherents seek the sacred because they seek the sacred, not because they seek something else through it. Religion can have anthropological, sociological, psychological, and economic benefits, but those benefits are considered to be only the consequences of a quest for the sacred and not the purpose of the quest. Furthermore, Eliade tolerates no reciprocal causality, the way, for example, Max Weber does. Religion can have non-religious effects, but not vice versa. Other scholars in the field of religious studies do allow for multiple causes as well as for multiple effects of religion, but they still insist that the main cause is the religious one. For the uncompromising Eliade, the sole cause of religion is religious. Somehow religion cannot arise for both religious and non-religious reasons.

Because Eliade pits the religious cause against all other proposed ones, he pits the field of religious studies against the social sciences, which he continually faults for accounting for religion non-religiously, or 'reductively': '... a religious phenomenon will only be recognized as such if it is grasped at its own level, that is to say, if it is studied *as* something religious. To try to grasp the essence of such a phenomenon by means of physiology, psychology, sociology, economics, linguistics, or any other study is false; it misses the one unique and irreducible element in it – the element of the sacred' (Eliade, 1958a, p. xiii). Eliade takes for granted that, as disciplines, the social sciences stand committed to a non-religious account of religion.

Only once, in all his writings, does Eliade qualify his condemnation of the social sciences as resolutely reductive in their approach to religion. In fact, the qualification comes in an otherwise unbending indictment of anthropologists for spurning the adherent's point of view, which he assumes to be religious: 'Indeed, as it is well known, the main interest of the western scholars has been the study of material cultures and the analysis of family structure, social organization, tribal law, and so forth. These are problems, one may say, important and even urgent for western scholarship, but of secondary importance for the understanding of the *meaning of a particular culture*, as it was understood and assumed *by its own members* (broader, sympathetic and intelligent studies like those of Marcel Griaulle, E. G. [sic] Evans-Pritchard or R. G. Lienhardt, and a few other cultural anthropologists are rather exceptional' (Eliade, 1967, pp. 502–503 [italics in original]).

Eliade's characterization of anthropology was outdated even when he wrote the essay, and it has become ever more outdated ever since. At best, his characterization, which borders on caricature, applies to the British brand of social anthropology influenced by A. R. Radcliffe-Brown. But the interest of British social anthropologists in social structure has rarely been at the expense of an interest in religion, as the cases of Evans-Pritchard and Lienhardt, as well as that of Radcliffe-Brown himself, conspicuously attest.

More important than the accuracy of Eliade's depiction of anthropology and of other social sciences is his argument. For him, the social sciences wrongly disregard the believer's point of view. But in fact they do not. Far from ignoring, let us even suppose, the religious nature of religion for adherents, social scientists begin with it. The actor's point of view constitutes the data to be accounted for. If social scientists ignored 'the meaning of a particular culture, as it was understood and assumed by its own members', they would have nothing left to account for. Why religion has meaning for adherents is what social scientists want to know.

Yet if social scientists necessarily turn to adherents for the data to be accounted for, they do not necessarily turn to adherents for an account of the data. Believers are akin to patients; social scientists, akin to doctors. It is not that religion is a disease to be cured but that religion is better understood by doctors than by patients. Patients may provide their own diagnoses of their ailments, but the final diagnosis rests with the doctor. That a patient rather than the doctor has the disease does not make the patient an expert on it. A patient's own diagnosis may turn out to be correct, but not because it is the patient's.

Eliade's indictment of anthropologists for spurning the actor's point of view is even more tenuous because as noted, he himself spurns it. If he is entitled to find a religious function in the professedly non-religious activities of avowed non-believers, why are social scientists not entitled to find non-religious functions in the professedly religious activities of avowed believers? In truth, both are entitled to venture beyond the actor's point of view.

Eliade's additional argument for accounting for religion religiously is that an account of any phenomenon must match the phenomenon being accounted for. Just as an account of a work of literature must be literary, so an account of a religious belief or practice must be religious. Eliade is right to insist on symmetry between an account and the subject of the account. But he is wrong to assume that the phenomenon being accounted for can be characterized and thereby accounted for in only one way. A 'literary' work such as a novel can alternatively be characterized as a work of history, sociology, or autobiography, and can then be accounted for accordingly. A Freudian account of religion characterizes a believer's relationship to God as a child's relationship to the child's father, and then attempts to account for that relationship psychoanalytically.

For Eliade to reply that God is more than a human being would miss the point. Freud never claims to be able to account for all aspects of God, only for father-like ones. He never claims that his characterization of the relationship between believer and God as filial is exhaustive. As a whole, social scientific accounts of religion claim at most to be offering necessary, not sufficient, accounts of religion. Freud recognizes the gap between a father and a God, and he is prepared to turn to religious studies to close it. He claims only to be demonstrating how father-like God is. Eliade, however, would bar Freud from offering an account, whether necessary or sufficient, of any aspects of God, and would do so on the question-begging grounds that religion is religion and not psychology.

See also: Anthropology and History; Myth in Religion; Religion, Sociology of; Ritual and Symbolism, Archaeology of; Ritual; Symbolism in Anthropology.

Bibliography

Allen, D., 1998. Myth and Religion in Mircea Eliade. Garland/Taylor and Francis, New York.

Berger, A., 1989. Fascism and religion in Romania. Annals of Scholarship 4, 455–465.

Eliade, M., 1954. The Myth of the Eternal Return. Trans. Trask W R. Pantheon Books, New York (also published 1955 Routledge and Kegan Paul, London; reprinted as 1959 Cosmos and History, Harper Torchbooks, New York; reprinted as 1971 The Myth of the Eternal Return, Princeton University Press, Princeton, NJ; 1991 Princeton University Press Mythos Series).

Eliade, M., 1958a. Patterns in Comparative Religion. Trans Sheed R. Sheed and Ward, New York (reprinted 1963 Meridian Books, Cleveland; reprinted 1974 New American Library, New York).

Eliade, M., 1958b. Yoga: Immortality and Freedom. Trans. Trask W R. Pantheon Books, New York. Routledge and Kegan Paul, London (second ed. 1969 Princeton University Press, Princeton, NJ; reprinted 1970 Princeton University Press, Princeton, NJ; 1990 Princeton University Press Mythos Series).

Eliade, M., 1959a. The Sacred and the Profane: The Nature of Religion. Trans. Trask W R. Harcourt, Brace, New York (reprinted 1961 Harper Torchbooks, New York; reprinted 1968 Harvest Books, New York).

Eliade, M., 1959b. Methodological remarks on the study of religious symbolism. In: Eliade, M., Kitagawa, J.M. (Eds.), The History of Religions: Essays in Methodology. University of Chicago Press, Chicago.

Eliade, M., 1963. Myth and Reality. Trans. Trask W R. Harper, New York (also published 1964 Allen and Unwin, London; reprinted Harper Torchbooks, New York, 1968; reprinted 1975 Harper Colophon, New York).

Eliade, M., 1964. Shamanism: Archaic Techniques of Ecstasy. Trans. Trask W R. Pantheon Books, New York. Routledge and Kegan Paul, London (reprinted 1970 Princeton University Press, Princeton, NJ).

Eliade, M., 1967. On understanding primitive religions. In: Müller, G., Zeller, W. (Eds.), Glaubej/Geist/Geschichte: Festschrift für Ernst Benz. Brill, Leiden, The Netherlands.

Eliade, M., 1969. The Quest: History and Meaning in Religion. University of Chicago Press, Chicago (reprinted 1975 University of Chicago Press Phoenix Books, Chicago).

Eliade, M., 1972. Zalmoxis: The Vanishing God: Lomactive Studies in the Religious and Folklore of Dacia and East Europe. Trans. Trask W R. University of Chicago Press, Chicago (reprinted 1978 University of Chicago Press Phoenix Books, Chicago).

Eliade, M., 1978–85. Trans. Trask W R vol. 1 and 2 Hiltebeitel A, Apostolos-Cappadona D vol. 3. A History of Religious Ideas, vol. 3. University of Chicago Press, Chicago (also published 1979–85 Collins, London).

Eliade, M. (Ed.), 1987. Encyclopedia of Religion, Vol. 16. Macmillan, New York.

Eliade, M., 1981. Trans. Ricketts M L. Autobiography, I: 1907–1937, Journey East, Journey West. Harper and Row, San Francisco (reprinted 1988 University of Chicago Press, Chicago).

Eliade, M., 1988. Autobiography, II: 1937–1960, Exile's Odyssey. Trans. Ricketts M L. University of Chicago Press, Chicago.

Idinopoulos, T.A., Yonan, E.A. (Eds.), 1994. Religion and Reductionism: Essays on Eliade, Segal and the Challenge of the Social Sciences for the Study of Religion. Brill, Leiden, The Netherlands.

Leach, E.R., 1966. Sermons by a man on a ladder. New York Review of Books 7, 28–31. October 20.

Rennie, B.S., 1996. Reconstructing Eliade: Making Sense of Religion. State University of New York Press, Albany, NY.

Ricketts, M.L., 1988. Mircea Eliade: The Romanian Roots, 1907–1945, vol. 2. East European Monographs, Boulder, CO.

Elias, Norbert (1897–1990)

Wolfgang Ludwig-Mayerhofer, University of Siegen, Siegen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Starting with a biography of Norbert Elias (1897–1990), this article proceeds to outline his major work, *Über den Prozeß der Zivilisation*. While influenced by different strands in sociology, history, and psychology, the unique and major achievement of this work consists in linking developments on the level of individuals, i.e., the 'civilization' of conduct, manners, or display of emotions, to large-scale processes on the macro level (monopolization of state violence, longer chains of interdependence among human beings). Finally, the resonance this work and its author have found both within sociology and a wider academic audience is discussed.

Born on 22 June 1897 to Jewish parents in Breslau (now Wrocław, Poland), then part of Germany, Elias, after completing his secondary education, first set out to study medicine, but soon changed to philosophy. Only after having completed his dissertation in that discipline in 1924 did he turn to sociology, pursuing his studies at Heidelberg, then one of the leading universities in Germany. He was accepted by Alfred Weber, Max Weber's brother, as a candidate for the *Habilitation* (conferring qualification as a university lecturer or professor), but when his colleague and friend Karl Mannheim obtained a chair at the University of Frankfurt, Elias followed him in 1930 as his assistant. His thesis for the *Habilitation* was completed by the beginning of 1933, but with the Nazis having taken over power in Germany, Elias had to flee, leaving Germany first for France, and then, in 1935, for London, on a grant from a philanthropic organization. Between 1935 and 1938, he completed his now famous two-volume *Über den Prozeß der Zivilisation* (*The Civilizing Process*), which, as a book published in the German language on the eve of World War II, did not receive much attention before republication in 1969. Elias remained in England where he obtained a post as Lecturer in Sociology at the University of Leicester but not before 1954. After retirement in 1962, he spent 2 years as professor of sociology in Ghana. Later on, he moved first to Amsterdam and then to Bielefeld (Germany), and was also invited to a number of European universities as visiting professor during the 1970s. In the mid-1980s, he moved back to Amsterdam where he died on 1 August 1990 (on his biography, see Elias, 1990; Mennell, 1989). Only during the 1970s, after a few sporadic articles, had he begun to publish extensively, first in English and later on, during his Amsterdam and Bielefeld years, drawing much on older manuscripts, in German. Some of the manuscripts were published only after his death.

There are few sociologists for whom knowledge about their biography is as central in understanding both their fate as a sociologist and their influence on the discipline as in the case of Elias. His first major book, and by far the most influential one, was not recognized as a masterpiece before his retirement as a university teacher. During the 1950s and 1960s, Elias seemingly was cut off, and cut himself off, from his former German contacts, such as Theodor W. Adorno. Neither Elias nor those very few sociologists who acknowledged the outstanding achievement of *Über den Prozeß der Zivilisation* (e.g., Johan Goudsblom or Dieter Claessens) were successful in campaigning for this book. It only became famous in the early 1970s, when copies of the 1969 edition began to circulate among students and when a best-selling German paperback edition appeared in 1976. Translations into about 15 languages testify to the wide interest the book still attracts. Elias later complained about the insufficient recognition of the other works he published, among other subjects, on the sociology of sciences (Elias, 1974), on time (Elias, 1988), on relationships between the established and the outsiders (Elias and Scotson, 1965), on sports (Elias, 1970a), and even on topics such as Mozart (Elias, 1991a) or *The Loneliness of the Dying* (Elias, 1987a). Yet, without any doubt, *Über den Prozeß der Zivilisation* is by far his most important, and probably also his only truly original contribution to sociology.

This is not to say that this work, like his others, does not draw on many concepts and notions developed by fellow sociologists. The idea of conceiving of society in terms of processes, instead of states, the absence of which from contemporary sociology was a theme reiterated in Elias's writings from the 1960s onward, was firmly rooted not only in the work of those American and English sociologists, such as M. Ginsburg and W. Osgood, whom Elias began to study in London in 1935 (van Krieken, 1998: pp. 26–28), but also in important strands of German sociology, not least two thinkers whose work he was surely well acquainted with, Karl Mannheim and Alfred Weber. From the latter, he probably borrowed the notion of the civilizing process (Weber, 1920), but giving it an entirely different content, as Weber by that term referred to the continual accumulation of knowledge, conceived of as a transition from 'naiveté' to reflection and to mastering of nature (a theme that is not entirely absent from Elias's late writings). Other important influences that are easily recognized were Sigmund Freud and the Dutch historian Johan Huizinga, whose *Herfstij der middeleeuwen* (Huizinga, 1919) was Elias's dominant source on the late Middle Ages. But the blend Elias formed from these, and other ideas (such as Simmel (see Tabboni, 1993), or traces of Darwinism), together with his own research, was decidedly unique.

The main achievement lies in the rigor with which Elias portrays the development of certain forms of conduct or manners and ties this development to large-scale processes of state formation and modernization, on the one hand, and the

concomitant effects on the psychic level, above all the display of emotions, on the other. The first volume of *Über den Prozeß der Zivilisation*, drawing on a variety of etiquette manuals, lucidly describes how norms of conduct we now think as self-evident, such as our eating habits, or norms concerning bodily functions including nakedness and sexuality, gradually developed from the Middle Ages onward (with the selection of this 'starting point' not meant to imply that the process had not begun before that time). The Middle Ages are described as a time when man could answer needs and desires more spontaneously, when emotions were displayed with much less restraint. The gradual replacement of the use of fingers with forks and knives at the table, the diffusion of the idea that spitting or blowing one's nose onto the ground or into the fingers are to be considered embarrassments to others, or the refinement and suppression of violence – all these are examples of the change of outward, overt behavior in the direction of increasing 'civilization.' These changes are accompanied by a gradual shift of the psychic apparatus toward an increasing influence of the superego (especially feelings of shame and embarrassment), and also of the ego, over the human drives, or the id, and by a restriction of affects and impulses, a 'damping' of the amplitude of the swings of emotion.

These developments were, in the second volume of *Über den Prozeß der Zivilisation*, analyzed as embedded in large-scale historical processes. The rise of the European empires, especially French absolutism, and the concomitant development of monopolies of violence on the state-level restrained aggressive behavior in everyday conduct. The French court was also the site where the nobility developed those manners that were continually refined in the courtiers' competition for the king's favor (a process described in more detail in Elias, 1969). These manners also served as elements of distinction of the entire nobility vis-à-vis the *bourgeoisie* – but to no avail, as the latter, and later on also other social strata, began to imitate and to take over the nobility's patterns of behavior. These interdependencies are part of the overall process of modernization, especially of social differentiation, which requires of modern man a degree of reflection, foresight, and calculation unknown to medieval man. The distinctive overall characteristic of this process is the increasing internalization of social constraints, or the gradual replacement of 'outer' constraints by self-constraint – a self-constraint we are only sometimes conscious of, as it is routinized and habitualized, or incorporated, in our everyday conduct.

The main thrust of the book thus resonates well with a theme that has become central to the social sciences and the humanities in the 1970s: the investigation of domination, or constraint, not (only) on the (macro) level of relationships between social groups, but rather as it makes itself felt on the (micro) level of emotions and of individual conduct. It is a topic that can be found in the writings of Herbert Marcuse, or of Michel Foucault of the mid-1970s, who both echoed, albeit with very different emphasis (and seemingly without relating to Elias), the 'discontents of civilization' Sigmund Freud had already articulated in the late 1920s. Elias's account was perhaps received with even more sympathy as it nonetheless retained an optimistic tone. On the one hand, it was sensitive to the 'costs of civilization' that were felt by many at that time. On the other, Elias regarded the process of civilization as possibly leading to a balance between individual needs and the requirements of social life and in the long run to a more peaceful world society. The uniqueness of *Über den Prozeß der Zivilisation* has two (related) dimensions: In terms of content, it lies in its blend of sociology with historical and psychological insight. In terms of method, it lies in its linking of grand theory with careful scrutiny of detail.

Many sociologists would readily acknowledge that sociology has to look at historical processes; however, few, if any, have actually delivered a picture that so vividly describes the changes in people's manners and habits that have occurred from the Middle Ages onward. Elias gave a detailed description of how 'medieval man' (women are indeed largely absent from his account) gradually became 'modern man'; that is, where others have rendered snapshots, Elias has produced a movie, as it were. We can indeed understand that medieval man, as depicted by Elias (i.e., more impulsive, aggressive, and yielding more readily to emotions than modern man) is alive today, but in a more suppressed, constrained way.

This, together with its often lucid language, largely free of sociological jargon, has rendered *Über den Prozeß der Zivilisation* popular among a wider audience of nonspecialists. It has also been received well by historians such as Robert Chartier or Robert Muchembled, and also in psychology and other humanities. This is not meant to imply that Elias has triggered new research agendas in these disciplines; rather, his work was considered as congenial and as providing an overarching framework for research that had been devoted to similar topics, but without his broad, all-encompassing approach. At the same time, it is often concluded that by assuming a long-term trend toward 'more civilized' behavior, Elias possibly has drawn too far-reaching conclusions. While accusations of Eurocentrism in his work can easily be refuted, the observation that other cultures have developed comparable standards of civilization (in Elias's sense) remains a challenge to the theory (see, however, Mennell's (2007) recent attempt at analyzing 'The American Civilizing Process' in a manner quite analogous to that of Elias). It is also questioned whether medieval man indeed was so different. Certainly some of what Elias says on the display of emotions is valid; but the Middle Ages also had highly developed cultural rules for constraining and guiding behavior. A related problem is that Elias in the final analysis was not able to account for the cruelties of the twentieth century, not least those of Nazism. Elias certainly was aware of the ambivalences of civilization, but they were recognized by him mostly on the individual level, to the neglect of the dangers inherent in such seemingly civilized (and civilizing) forces such as modern technology or the state monopoly of violence. It must be added that during his late life, he became more sensitive to these issues, but still arguing fervently in favor of the possibility (not the unavoidability) of worldwide peace.

Elias's account exhibits serious lacunae also on the more factual level (van Krieken, 1989). Among the deficiencies of *Über den Prozeß der* Zivilisation is the absence of an account of the processes through which the manners of conduct, described by Elias only with reference to the etiquette manuals, were effectively brought about by (private, and later on public) educational practices. Even more importantly, some presumably decisive conditions of modern civilization (in Elias's understanding of that term) are entirely lacking.

Among these, most conspicuous are, first, the disciplinary mechanisms that were invented or became more widespread during the eighteenth and nineteenth centuries, especially the school and the factory, an observation that also points to the almost complete absence of the working class from Elias's work. Second, going further back in history, we also find a neglect of religion, certainly one of the most important sources of (European) civilization in the form of monastic discipline and more generally the religious prescriptions (and proscriptions), not least of Protestantism.

Elias's later theoretical work, claiming to develop a general 'theory of human beings' (*Menschenwissenschaft*), exerted little influence on sociology in general. The earlier focus on social processes in these later writings was complemented by more emphasis on the 'figurational' nature of society, a term coined by Elias to point out that man and society could and should be understood as relational entities (see especially Elias, 1970b, 1987b). With that concept, he claimed to overcome the alleged dichotomy of individual and society that he saw prevailing in most other sociologists' thought – an idea borne not so much from in-depth acquaintance with sociological theorizing but rather from his desire, obvious in most of his late works, to portray himself as a unique innovator of sociological theory. Figurations were analyzed mainly in terms of power – or more exactly of power differentials, of the dynamics resulting from these, and finally of the outcomes which, as he often argued, cannot be understood as a realization of the purposes of social agents but rather as unintended consequences. In other words, his sociology was mainly a structural-relational one, with emphasis on the dynamics of social relations (in contrast to much of the 1930s–60s structuralism). The price was a neglect of individual motivation as well as of broader cultural systems of meanings, such as values or norms. The absence of such aspects from Elias's sociology is perhaps a consequence of his reluctance, acquired very early in his grappling with philosophy, to concede that ideas have an existence of their own. Only in his latest writings did he start to acknowledge the influence of systems of meaning and language (Elias, 1988, 1991b). With respect to some sociologists' claim to apply or to extend Elias's figurational sociology, it must be borne in mind that he never developed anything approaching a systematic account. Therefore, references to this theory must be viewed as partly gratuitous, as the idea of a 'relational' sociology certainly is not Elias's original contribution.

In contrast, fruitful inspirations continue to come from his theory of the civilizing process. Research in this spirit is accomplished above all in those places where he spent the last two decades of his life, that is, the Netherlands, especially Amsterdam, and also Germany, whereas in the United States, in spite of a few admirers such as Lewis A. Coser or Thomas J. Scheff, his work has attracted wide attention. There have been fine extensions of the theory to specific phenomena such as eating practices (Mennell, 1985), the history of marriage (Schröter, 1985), the domestication of fire (Goudsblom, 1992), or the use of tobacco (Hughes, 2003), demonstrating how Elias's ideas are especially applicable to aspects of everyday cultural practices. But there are also attempts at rendering his approach useful to macroanalyses, such as in de Swaan's (1988) work on the development of modern welfare states. There is still much debate concerning the idea of civilization as an irreversible process and its relationships to modernity. While some researchers have explored the possibilities, and the conditions, of decivilizing processes, others have investigated processes of informalization, that is, the development of less restricted, destandardized norms of behavior (Wouters, 2007). Overall, it seems to be difficult to use Elias's ideas on the civilizing process as a starting point for really new developments; rather, much research focuses on applications of his insights to as yet unexplored phenomena, or on refining the theory in order to render it capable to account for phenomena that seem to contradict it at first sight.

The continuing interest in Elias's work is fed by the publication of his *Gesammelte Schriften* in German (Suhrkamp Verlag, completed in 2010) and his Collected Works in English (University College Dublin Press, not yet completed at the time of writing). Elias' correspondence and manuscripts are housed at the *Deutsches Literaturarchiv* in Marbach (Germany). The Norbert Elias Foundation, established by him in 1983, holds the copyright of all his published work and maintains a newsletter that reports biannually about activities related to Elias. These are entertained by a dedicated group of adherents to Elias, rightly described as a 'clan' or a 'family' by one of its members (Mennell, 2006). One can truly say that Elias has succeeded in establishing a 'school' of followers who both try to defend the achievements of their master (and their own) and to devalue other ways of doing sociology. At the same time, in some European countries there is still a market for Elias's work amongst a broader academic public; for example, a book published in France (Elias, 2010) catches attention with a red bellyband announcing '*Norbert Elias inédit*.' Thus, while Elias's influence on sociology as a whole may remain restricted, his work, and work in his tradition, will continue to constitute an important, if small, branch of sociology.

See also: Freud, Sigmund (1856–1939); Mannheim, Karl (1893–1947); Simmel, Georg (1858–1918).

Bibliography

de Swaan, A., 1988. In Care of the State. Health Care, Education and Welfare in Europe and the USA in the Modern Era. Polity Press, Cambridge.

Elias, N., 1969. Die Höfische Gesellschaft. Luchterhand, Neuwied (1983. The Court Society. Pantheon, New York).

Elias, N., 1970a. The genesis of sport as a sociological problem. In: Dunning, E. (Ed.), The Sociology of Sport. Frank Cass, London, pp. 88–115.

Elias, N., 1970b. Was Ist Soziologie? Juventa, München (1978. What is sociology? Hutchinson, London).

Elias, N., 1974. The sciences: towards a theory. In: Whitley, R. (Ed.), Social Processes of Scientific Development. Routledge & Kegan Paul, London, pp. 21–42.

Elias, N., 1976[1939]. Über den Prozeß der Zivilisation, 2 vols. Suhrkamp, Frankfurt a.M (1978/1982. The Civilizing Process. Blackwell, Oxford, UK).

Elias, N., 1987a. Über die Einsamkeit der Sterbenden in unseren Tagen. Suhrkamp, Frankfurt a.M (1985. The Loneliness of the Dying. Blackwell, Oxford, UK).

Elias, N., 1987b. Die Gesellschaft der Individuen. Suhrkamp, Frankfurt a.M (1991. The Society of Individuals. Blackwell, Oxford, UK).

Elias, N., 1988. Über die Zeit. Suhrkamp, Frankfurt a.M (1992. Time. An Essay. Blackwell, Oxford, UK).

Elias, N., 1990. Über Sich Selbst. Suhrkamp, Frankfurt a.M (1994. Reflections on a Life. Polity, Cambridge, UK).

Elias, N., 1991a. Mozart. Zur Soziologie eines Genies. Suhrkamp, Frankfurt a.M (1993. Mozart. Portrait of a Genius. Polity, Cambridge, UK).

Elias, N., 1991b. The Symbol Theory. Sage, London.
Elias, N., 2010. Au-delà de Freud. Sociologie, psychologie, psychanalyse. La Découverte, Paris.
Elias, N., Scotson, J.L., 1965. The Established and the Outsiders. A Sociological Enquiry into Community Problems. Cass, London.
Goudsblom, J., 1992. Fire and Civilization. Allen Lane, London.
Huizinga, J., 1919. Herfsttij der Middeleeuwen. Studie over levens – en gedachtenvormen der veertiende en vijftiende eeuw in Frankrijk en de Nederlanden (1924. The Waning of the Middle Ages. Edward Arnold, London; 1996. Newer Translation: The Autumn of the Middle Ages. University of Chicago Press, Chicago).
Hughes, J., 2003. Learning to Smoke: Tobacco Use in the West. University of Chicago Press, Chicago, London.
Mennell, S., 1985. All Manners of Food: Eating and Taste in England and France from the Middle Ages to the Present. Blackwell, Oxford.
Mennell, S., 1989. Norbert Elias: Civilization, and the Human Self-image. Blackwell, Oxford.
Mennell, S., 2006. Elias and the counter-ego: personal recollections. History of the Human Sciences 19, 73–91.
Mennell, S., 2007. The American Civilizing Process. Polity, Cambridge.
Schröter, M., 1985. "Wo zwei zusammenkommen in rechter Ehe…". Sozio – und psychogenetische Studien über Eheschließungsvorgänge vom 12. bis 15. Jahrhundert. Suhrkamp, Frankfurt a.M.
Tabboni, S., 1993. Norbert Elias. Un Ritratto Intellettuale. Il Mulino, Bologna.
van Krieken, R., 1998. Norbert Elias. Routledge, New York, London.
van Krieken, R., 1989. Violence, self-discipline and modernity. Beyond the civilizing process. Sociological Review 37, 193–218.
Weber, A., 1920. Prinzipielles zur Kultursoziologie (Gesellschaftsprozeß, Zivilisationsprozeß, und Kulturbewegung). Archiv für Sozialwissenschaft und Sozialpolitik 47, 1–49.
Wouters, C., 2007. Informalization: Manners and Emotions since 1890. Sage, London.

Relevant Website

http://www.norberteliasfoundation.nl/ – The Norbert Elias Foundation.

Elicitation of Probabilities and Probability Distributions

Lara J Wolfson, Brigham Young University, Provo, UT, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 7, pp. 4412–4417, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Under confidentiality, identifiable data provided for statistical purposes is protected from unauthorized disclosure. Organizations acting as brokers between respondents and data users seek to disseminate useful data products while keeping low the risk of confidentiality disclosure. Recognizing that deidentification of each data record is generally inadequate to protect its confidentiality against attack by a data snooper, agencies restrict the data they release for general use. Typically, these restricted data procedures have involved transformation or masking of the original collected data through such devices as adding noise, topcoding, data swapping, and recoding. Another approach is to use the original data to determine a statistical model and use it to generate synthetic data. Generically, statistical disclosure limitation is a body of restricted data procedure that transforms data so that release of the transformed data adequately limits disclosure risk. Desirably, statistical analysis of the transformed data leads to inferences similar to that obtained by analysis of the original data. The technical procedures for implementation of disclosure limitation involve a range of mathematical and statistical tools.

Bayesian statistical and decision analyses require several inputs aside from observed data. Among these inputs are *prior distributions* and *utility functions* (Barnett, 1982), which typically reflect the opinion(s) of an individual or group. *Elicitation* is the process of encoding the subjectively held beliefs of an individual (called the *elicitee*) or group into prior distributions or utility functions. Elicitation is thus a complex (and imprecise) mapping from human knowledge processing to a quantitative metric. As such, the study of elicitation methods has two primary divisions – first, the psychological aspects of how humans are able to articulate knowledge, and second, the mathematical aspects of how such articulated knowledge is captured for use in a Bayesian analysis.

Subjective Opinion within a Bayesian Framework

Subjective opinion is actually employed in several parts of any statistical analysis, Bayesian or frequentist (Lad, 1996). The Bayesian model of decision making and inference is that prior beliefs about a particular attribute or state of nature are updated through data, and then used together with utilities to decide on a course of action. Formally, θ represents a (possibly multiattributed) state of nature such that $\theta \in \Theta$; $\mathbf{x}$ represents observed data; and

$f_\Theta(\theta)$	Prior
$f_X(\mathbf{x}\|\theta)$	Likelihood
$f_\Theta(\theta\|\mathbf{x}) = \dfrac{f_X(\mathbf{x}\|\theta)f_\Theta(\theta)}{\int_\Theta f_X(\mathbf{x}\|\theta)f_\Theta(\theta)d\theta}$	Posterior
$L(\theta, \mathbf{a}) = -U(\theta, \mathbf{a})$	Loss Function $(=-\text{Utility})$
$D(\mathbf{x}) = \min_a \int_\Theta L(\theta, \mathbf{a})f_\Theta(\theta\|\mathbf{x})d\theta$	Optimal decision minimizes posterior expected loss

Under a personal and subjective view of Bayesian inference sometimes called subjective expected utility (Savage, 1954), a decision maker will choose the course of action, a, that, based on prior beliefs updated by collected data, will have the least adversarial consequences when compared with other possible decisions.

Elicitation is concerned with mapping subjectively held opinions into prior probabilities, $f(\theta)$, and (subjective) preferences and value judgments into utilities (or exchangeably, loss functions) $L(\theta|a)$. These prior distributions and utilities represent an amalgamation of information available in addition to the data collected for a particular problem. As such, the most valuable elicited opinions are those of experts in the area of the problem, who may (or may not) have any training in statistical methods. Meyer and Booker (2000) and Chaloner (1996) outline reasons why subjective expert opinions are valuable in Bayesian analysis. These include the a priori information that can be quantified in the investigation of new, rare, complex, or poorly understood phenomena; when collecting large amounts of data is not feasible; when collected data alone is insufficient evidence for decision making; in designing experiments (*see* Experimental Design: Overview); in forecasting future events; and incorporating historical data and experiences that may not be well documented.

Spetzler and Staël von Holstein (1975) identify three stages of an elicitation process: the *deterministic phase*, which corresponds to choosing a functional form for a statistical model, the *probabilistic phase* where the subjective assessments are made, and the *informational phase* where verification checks are carried out. The primary focus of elicitation methods is on the middle phase (probabilistic). It is assumed generally, at least in statistical practice, that the first phase is fixed by the likelihood chosen for the data; the third phase is addressed more in the research on cognitive aspects of elicitation than the statistical aspects of encoding subjective beliefs and preferences.

Cognitive Aspects of Elicitation

In reporting subjectively held beliefs and preferences, there are several psychological heuristics that can lead to misrepresentation (*see* Cognitive Psychology: History). For more

detailed discussion on these, early work on the subject is found in Kahneman et al. (1982), Kyberg and Smokler (1980), Hogarth (1987); updated coverage is detailed in Poulton (1994) and in Wright and Ayton (1994). The most common problems in eliciting subjective opinions come from:

1. *Overconfidence*. Probability assessors tend to underestimate variability and the tails of the distribution.
2. *Hindsight bias*. Most assessors believe they would have predicted correctly the outcome of an event; thus only the outcomes that actually occurred are viewed as having nonzero probability of occurrence.
3. *Representativeness*. Expert judgments can be based on the synthesis of previously observed data. If that data came from small samples, it may not be representative. Other terms often used in conjunction with this heuristic are *base-rate neglect*, *small-sample fallacy*, and *misperception of randomness*. Another well-known aspect of representativeness is the *conjunction fallacy*, where higher probability is given to a well-known event that is a subset of an event to which lower probability is assigned.
4. *Availability*. The probability of an event is judged by the frequency with which an event can be recalled in memory. The classic example of this is in the elicitation of beliefs about likely causes of death; botulism, which typically gets a great deal of press attention, is usually overestimated as a cause of death, whereas diabetes, which does not generate a great deal of media attention, is underestimated as a cause of death.
5. *Adjustment and anchoring*. When an initial assessment is made, elicitees often make subsequent assessments by adjusting from the initial anchor, rather than using their expert knowledge.

To overcome possible biases introduced in the elicitation of probabilities and utilities by these heuristics, Kadane and Wolfson (1998) summarize several principles for elicitation:

1. Expert opinion is the most worthwhile to elicit.
2. Experts should be asked to assess only observable quantities, conditioning only on covariates (which are also observable) or other observable quantities.
3. Experts should not be asked to estimate moments of a distribution (except possibly the first moment); they should be asked to assess quantiles or probabilities of the predictive distribution.
4. Frequent feedback should be given to the expert during the elicitation process.
5. Experts should be asked to give assessments both unconditionally and conditionally on hypothetical observed data.

Eliciting Prior Distributions

Eliciting prior distributions entails quantifying the subjective beliefs of an individual into a statistical distribution (*see* Distributions, Statistical: Approximations). Kadane and Wolfson (1998) discuss the distinction between *structural* and *predictive* methods of doing this. In structural methods, an expert is asked to directly assess parameters or moments of the prior distribution; in predictive methods, elicitees are asked questions about aspects of the prior predictive distribution; the predictive distribution is

$$f_Y(\mathbf{y}|\mathbf{x}) = \int_{\Theta} f_Y(\mathbf{y}|\theta) f_{\Theta}(\theta|\mathbf{x}) \mathrm{d}\theta$$

and when no data has been observed (i.e., $\mathbf{x} = \{\varnothing\}$), this is known as the prior predictive distribution. In most complex problems, the predictive method is preferable, since it is more suited to the prescription for good elicitation methods outlined in the section 'Cognitive Aspects of Elicitation'. The functional form of the prior distribution is usually selected to be conjugate to the likelihood chosen to model the data.

The elicitation is generally executed by asking assessors to identify quantile–probability pairs (q, p) in some sense, where $p = P_X(X < q)$. Questions can be asked by specifying either a point or interval value for p or q, and responses can be given directly, as values, probabilities, or odds, or indirectly where preferences are stated between two alternatives.

To illustrate, consider various ways in which a prior distribution for the probability parameter in a binomial model might be elicited. The beta distribution is the usual choice for modeling the prior distribution, and functionally, the model is

$f_P(p\|\alpha, \beta)$	$\sim$ *Beta* (a, β)	Prior
$f_X(x\|p, n)$	$\sim$ *Binomial* (n, p)	Likelihood
$f_P(p\|\alpha, \beta, x, n)$	$\sim$ *Beta* $(\alpha + x, \beta + n - x)$	Posterior

In this instance, x is the number of successes observed in n trials; from the way in which the prior parameters are updated by the posterior distribution, $\alpha + \beta$ can be viewed as a 'prior sample size' or measure of prior strength, and $\alpha/(\alpha + \beta)$ as a point estimate of the mean prior probability of a success. So one way to structurally elicit a prior would be to ask an elicitee to specify a prior sample size and probability of success, and this would uniquely define the parameters of the prior distribution. This, however, does not provide any feedback to the elicitee, and will almost surely be overconfident because of the way in which the variability of the distribution is never elicited directly. It also does not address the issue of the skewness of the distribution. Chaloner and Duncan (1983) propose a predictive method for eliciting this same prior distribution, where the elicitee is asked to specify the modal number of successes for given prior samples sizes, and then is asked several questions about the relative probability of observing numbers of successes greater than or less than the specified mode. The elicitee is also shown (graphically) the implied predictive distributions for future observations specified by their current answers, and then walked through a bisection algorithm to adjust the spread of the distribution.

While the predictive method just described is more difficult to implement, it is also more appropriate in terms of having the elicitee answer questions that will effectively allow them to communicate both their knowledge and their uncertainty within the constraints of the model. To illustrate, suppose that in the structural method, an elicitee specified a prior sample size of 15 and a prior mean of 1/3. In the predictive method, the same elicitee specified, for a sample size of 15, a predictive mode of five successes; the odds of observing four rather than five successes was specified as 0.8, and the odds of observing six rather than five success was specified as 0.7. The two priors, in the structural case, a beta (5, 10) distribution, and in the predictive case, a beta (9.67, 18.74), are shown in **Figure 1**. This illustrates that way in

which the predictive method allows for better representation of the actual state of knowledge the elicitee has.

Existing protocols for eliciting prior distributions are described in some detail in Kadane and Wolfson (1998). Elicitation protocols are described as being either *general* or *application-specific*. General methods are those that are meant for specific classes of models, such as the beta-binomial, the normal linear regression model, the dirichlet-multinomial, and others. Application-specific methods are ones tailored to the problem under study; several examples of these are described both in Kadane and Wolfson (1998) and French and Smith (1997). Application-specific methods are used typically for complex situations, where there may be no conjugate or standard prior that is suitable, or where the elicitation process must be *decomposed* in order to provide sensible assessments.

Eliciting Utilities

The ultimate goal of eliciting utilities is to make decisions that minimize expected loss, under the constraints of uncertainty specified by the prior distribution and likelihood function. Determining expected loss means constructing loss functions that consist of outcomes based on a quantity of interest, $\mathscr{Q}$, and the loss associated with each outcome. The loss can be fixed or random. In general, there is a set $\mathscr{A}$ of possible actions that can be taken, and for each element of $\mathscr{A}$, there are losses that may or may not depend on $\mathscr{Q}$. The set $\mathscr{A}$ must be exhaustive; in other words, all possible actions must be enumerated, and each action must be exclusive of every other action.

The first stage of the elicitation process should be to choose the action set $\mathscr{A}$ and the quantity of interest $\mathscr{Q}$. The set of actions should be well defined and must contain at least two elements. These actions are usually dependent on some quantity of interest. For example, if the action set consists of choosing which family planning program to implement, the quantity of interest might be some quantification of the state of infant mortality. In choosing the quantity of interest, it is important to make sure that $\mathscr{Q}$ is a quantity that incorporates the uncertainty inherent in the problem under study.

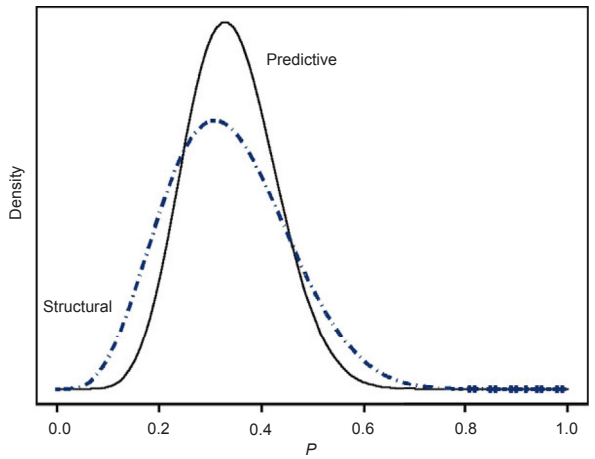

Figure 1 Structural and predictive elicited distributions for the beta-binomial model.

Generally, $\mathscr{Q}$ will be a random quantity defined by a statistical model, and as such, it can have several dimensions.

Once the action set $\mathscr{A}$ and the quantity of interest $\mathscr{Q}$ have been determined, the second stage of the elicitation is to construct loss functions based on them. This is the most difficult stage of the elicitation process. If several stakeholders are involved in the decision-making process, they may construct these separately or as a group; regardless, those assessing the loss functions must give careful thought to the consequences resulting from each action in $\mathscr{A}$ for any value of $\mathscr{Q}$. For useful overviews of loss function elicitation strategies, see works by Clemen (1996), Zeleny (1982), Edwards (1992), and Saaty (1980). In general, it is necessary to specify the form the loss functions should take; this is sometimes specified in advance, but it is advisable to consider this as part of the elicitation process.

There are several 'standard' loss functions. One of the best-known examples of a loss function is the 0–1 loss function generally applied to statistical hypothesis testing. In this case, the set of actions is given by $\mathscr{A} = \{a_1\text{: reject the null hypothesis}, \ a_2\text{: do not reject the null hypothesis}\}$ (Berger, 1985). An extension of this is the 0–k loss function, which is typically written (letting $\mathscr{Q} = \theta$)

$$L(\theta, a_i) = \begin{cases} 0 & \theta \in \Theta_i \\ k_i & \theta \notin \Theta_i \end{cases}, \quad i = 1, 2$$

There are many possible variants on this loss function that could be considered, and identifying the 'form' of the loss function is the more controversial of the tasks that must be performed in eliciting utilities. Given the conceptual framework of a 0–k loss function, consider the next elicitation task – identifying the values of k_1, k_2. Because k_1, k_2 can represent rather intangible quantities, it is often advantageous to specify their relative rather than absolute values. Consider that the 'optimal' decision in this example is to choose a_1 when

$$\frac{k_1}{k_2} > \frac{P(\theta \in \Theta_2|\mathbf{x})}{P(\theta \in \Theta_1|\mathbf{x})} = \frac{P(\theta \in \Theta_2|\mathbf{x})}{1 - P(\theta \in \Theta_2|\mathbf{x})}$$

Then the remaining elicitation task needed for decision-making is to elicit the ratio k_1/k_2, which simplifies the elicitation task from eliciting k_1, k_2 directly or independently. The ratio on the right-hand side above can be viewed as the ratio of the posterior probabilities that the parameter lies outside the 'rejection region' ($\theta \in \Theta_2$) to the posterior probability that the parameter lies within the region (or $\theta \in \Theta_1$). From this, a general relationship can be drawn between the functional form of the model and the two quantities for which preferences need to be elicited; this is illustrated in Figure 2.

Exploiting the relationships between the functional form of the model and loss function and the quantities about which preferences and value judgments must be elicited is a useful strategy. Many examples of the application of evaluating utilities and preferences are found in journals in the fields of risk analysis, accounting, and economics.

Other Issues

It is sometimes advisable or necessary to elicit the opinion of more than one individual. Genest and Zidek (1986) review

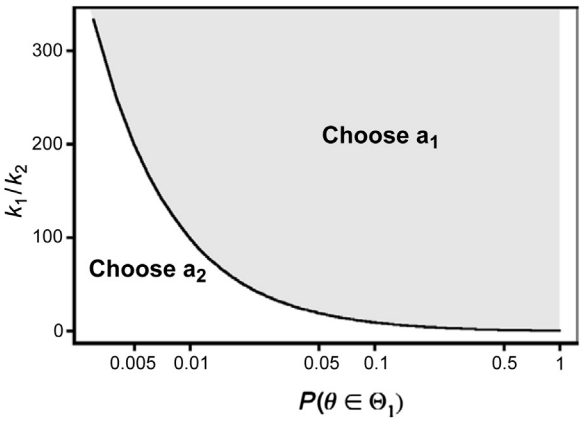

Figure 2 A decision maker should choose a_1 or a_2 based on the values of $P(\theta \in \Theta_1)$ and k_1/k_2.

methods for combining the opinions of several experts; Cooke (1991), Meyer and Booker (2000), and Morgan and Henrion (1990) provide guidelines for structuring elicitation tasks with groups of experts. French and Smith (1997) have several examples of analyses where elicitation of more than one expert is used. In general, there are two approaches used in group elicitation – the collective opinion of the experts is elicited in the whole or the opinions of several experts can be combined in some fashion.

Another issue that is discussed in the literature (see Kadane and Wolfson, 1998 for further discussion) is that of validating and comparing elicitation methods. There are several criteria that might be considered; the most important of these is *coherence*, that the elicited opinions obey the laws of probability (*see* Probability Theory: Formal). In general, this should be considered a minimal requirement for any elicitation method. Other considerations are the *reliability* of the elicitation methods in producing stable reflections of the elicitee's beliefs and preferences; also, the elicitation method can be validated by an examination of whether it obeys *proper scoring rules* (i.e., does not systematically encourage the misrepresentation of beliefs and preferences). Some practitioners of elicitation, particularly in the field of meteorology, advocate the *calibration* of expert opinion, whereby elicited opinions are adjusted to reflect the elicitee's ability to correctly predict events. Subjective Bayesians typically object to calibration because subjectively held beliefs and preferences are not meant to be infallible. A final consideration given in the validation and comparison of elicitation methods should be given to the modeling of errors-in-elicitation. Some cognitive aspects of this are described in Poulton (1994), but very little statistical literature is available on the subject.

See also: Bayesian Statistics; Multivariate Analysis: Discrete Variables (Overview); Spatial Autocorrelation; Spatial Sampling.

Bibliography

Chaloner, K.M., 1996. The elicitation of prior distributions. In: Berry, D.A., Stangl, D.K. (Eds.), Case Studies in Bayesian Biostatistics. Marcel-Dekker, New York, pp. 141–156.

Chaloner, K.M., Duncan, G.T., 1983. Assessment of a beta prior distribution: PM elicitation. The Statistician 32, 174–180.

Clemen, R.M., 1996. Making Hard Decisions, second ed. Duxbury Press, Belmont, CA.

Cooke, R.M., 1991. Experts in Uncertainty: Opinion and Subjective Probability in Science. Oxford University Press, New York.

Edwards, W. (Ed.), 1992. Utility Theories: Measurements and Applications. Kluwer Academic Publishers, Boston.

French, S., Smith, J.Q. (Eds.), 1997. The Practice of Bayesian Analysis. Arnold, London.

Garthwaite, Paul H., Kadane, Joseph B., O'Hagan, Anthony, 2005. Statistical methods for eliciting probability distributions. Journal of the American Statistical Association 100 (470), 680–701.

Genest, C., Zidek, J.V., 1986. Combining probability distributions: a critique and an annotated bibliography. Statistical Science 1, 114–148.

Hogarth, R.M., 1987. Judgment and Choice. Wiley, New York.

Kadane, J.B., Wolfson, L.J., 1998. Experiences in elicitation (with discussion). The Statistician 47, 3–19.

Kahneman, D., Slovic, P., Tversky, A. (Eds.), 1982. Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, UK.

Karni, Edi, 2009. A mechanism for eliciting probabilities. Econometrica 77 (2), 603–606.

Kyberg Jr., H.E., Smokler, H.E. (Eds.), 1980. Studies in Subjective Probability, second ed. Krieger, New York.

Lad, F., 1996. Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction. Wiley, New York.

Meyer, M.A., Booker, J.M., 2000. Eliciting and Analyzing Expert Judgement: A Practical Guide. Society for Industrial and Applied Mathematics and American Statistical Association, Philadelphia, PA.

Morgan, M.G., Henrion, M., 1990. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press, New York.

O'Hagan, Anthony, Buck, Caitlin E., Daneshkhah, Alireza, Eiser, J Richard, Garthwaite, Paul H., Jenkinson, David J., Oakley, Jeremy E., Rakow, Tim, 2006. Uncertain Judgements: Eliciting Experts' Probabilities. John Wiley & Sons, West Sussex, UK.

Poulton, E.C., 1994. Behavioral Decision Theory: A New Approach. Cambridge University Press, Cambridge, UK.

Saaty, T.L., 1980. The Analytic Hierarchy Process. McGraw-Hill, New York.

Savage, L.J., 1954. The Foundations of Statistics. Wiley, New York.

Spetzler, C.S., Staël von Holstein, C.A.S., 1975. Probability encoding in decision analysis. Management Science 22, 340–358.

von Winterfeldt, D., Edwards, W., 1986. Decision Analysis and Behavioral Research. Cambridge University Press, Cambridge, UK.

Wright, G., Ayton, P. (Eds.), 1994. Subjective Probability. Wiley, New York.

Zeleny, M., 1982. Multiple Criteria Decision Making. McGraw-Hill, New York.

Elites, Anthropology of

Hugh Gusterson, Elliott School of International Affairs, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Elites have been little studied by anthropologists because of practical and temperamental barriers to fieldwork. Early work on elites (in Africa, Oceania, South Asia, and Latin America) situated them in the context of broader social wholes and emphasized the importance of kinship in defining and maintaining elites. Subsequent work highlighted the roles of colonialism and capitalism in creating new elites – based on access to capital, education, and bureaucratic perquisites – that challenged and hybridized with traditional elites. While bureaucratic industrial societies present themselves as meritocracies, elites devise strategies of heredity dependent on their privileged access to the educational system and cultural capital. Anthropologists have only recently become interested in 'studying up' in the US, having earlier ignored the public debate of the 1950s and 1960s on the alleged 'power elite' in the US. Some of the most striking new work on elites has been of scientists, bankers, journalists, evangelists, psychiatrists, and development bureaucrats. Such work often requires innovative approaches to fieldwork.

'Elites' have been undertheorized and understudied in the social sciences, and the term has a certain vagueness: both 'traditional' and 'modern' societies are said to have elites, despite their divergent political and economic structures, and the privileges of elites may be grounded in genealogy, land ownership, control of capital, bureaucratic position, educational attainment, or religious status, for example. Analysts influenced by Marxism tend to emphasize the material underpinnings of elite status while those tilting toward Weber are more interested in elites' command of authority and prestige (Bottomore, 1993; Khan, 2012; Milner, 2014). Meanwhile some commentators treat the term 'elite' as broadly interchangeable with 'ruling class,' while others – making elite status more contextual – lend the term a fractal quality: subordinate classes and groups may have their own internal elites (skilled laborers, for example), so that society is honeycombed with elites all the way down. Thus one can forgive Richard Fox (1985: 931–932) for complaining that "the breadth of the elite concept comes at the expense of theoretical integrity."

The theorization of elites is often traced back to the Italians Vilfredo Pareto (1935) and Gaetano Mosca (1960), each working independently in the late nineteenth and early twentieth centuries to develop an alternative to Marx's theory of hierarchy in modern society. Pareto in particular saw the hereditarian elites of old Europe as prone to degeneration over time, and hoped that elites in capitalist societies would provide a meritocratic basis for social stratification. Despite their influence on sociological discussions of elites, especially as a counterpoint to Marxian discussions, the work of Pareto and Mosca has been of less interest to anthropologists.

In fact, anthropologists have historically shown relatively little interest in theorizing or studying elites. Early anthropologists were more interested in developing holistic descriptions of entire societies than of fractional segments, and they tended to be drawn to societies with relatively little division of labor. However, even as anthropology moved away from its early emphasis on holistic representations of non-Western societies, elites have still received relatively little ethnographic attention. There are many reasons for this, including practical obstacles, the historical conventions of the discipline itself, and the temperamental preferences of anthropologists. Elites are busy, important people who have the power to elude or obstruct anthropologists, and whose lives are usually unsuited to the anthropologist's classic (and time-intensive) technique of investigation – participant observation (Hertz and Imber, 1995). (While it is easy to imagine anthropologists helping peasants with the harvest, spending hours with poor womenfolk while they cook and clean, or idling with urban street-corner society, it is harder to imagine them sitting unobtrusively at board meetings, loitering in the hallways of the upper floors of the Pentagon, or being repeatedly invited to dinner with the Rockefellers). Anthropologists have also continued to neglect the study of elites for reasons of convention: the traditional division of labor in the social sciences has given anthropologists dominion over the geographically and socially marginal, and anthropology has historically developed as a discipline that explains to the center the logic of lifeways on the periphery. This convention retains a certain force even in the era of repatriated and postcolonial anthropology. There is also, finally, for many anthropologists a temperamental obstacle to fieldwork with elites: the traditional anthropological methodology of participant observation has historically demanded of anthropologists a reflexive sympathy for the people they have studied – largely poor and marginal people. This, combined with the liberal politics of most anthropologists, makes fieldwork with elites psychologically unappealing and conflictual (Gusterson, 1997).

Insofar as earlier generations of anthropologists wrote about elites, they often studied them not so much as topics in their own right but as components of larger social wholes or as products of social processes that were themselves the principal object of analysis. Examples include Dumont's (1970) work on caste, Barth's (1959) monograph on patron–client relations among the Swat Pathans of Pakistan, Kuper's (1961) study of an African aristocracy, and Sahlins' (1958) work on chiefs and followers in Polynesia. Insofar as this made elite status a processual outcome rather than a thing in itself, this was all to the good. Marcus (1983) observes that such work on elites tended

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12061-6

to cluster in particular ethnographic zones where kinship – the core subject matter of early anthropology – was particularly important in understanding the dynamics of elite status: South Asia, marked by its caste system; the Pacific Islands, home to hereditary chiefs and 'big men'; Africa, peopled by hereditary chiefs and kings; and the oligarchies of Latin America. Work by first wave anthropologists emphasized the importance in sustaining elite status of kinship relations, marriage strategies, land inheritance, patron–client relations, and ritual – what we might call the preindustrial sinews of elite power.

As colonization, industrialization, and missionization transformed societies around the world, new sinews of elite power emerged. Accumulation of capital (not just land) became increasingly important, as did access to education, geographical mobility, and positions in the expanding bureaucracies of new states. In many colonial societies the power of traditional, usually hereditary, elites was challenged by the rise of new elites who owed their standing to Western education, positions in emergent state bureaucracies, and trading opportunities provided by the colonial encounter. The processes of decolonization and globalization produced rivalries between emergent and established elites with different relationships to metropolitan power centers but, over time, old and new elites often hybridized through, for example, intermarriage, the conversion of capital into land, and a turn to Western education (Billig, 2007). (This process broadly reenacted a similar development that took place in the industrializing societies of Western Europe where the ancient aristocracies and rising bourgeoisies learned over time to accommodate to one another through marriage, education, and financial alliance). Processes of elite formation and mutation under and after colonialism are discussed by Ferguson (1990), Geertz (1963), Leach and Mukherjee (1970), Luhrmann (1996), and Vincent (1971).

As societies become more complex, bureaucratized and marketized, the bases of elite power shift and new kinds of specialized elites (from physicists and lawyers to investment bankers) emerge. One of the distinguishing features of complex industrial societies is that they have a more pluralistic constellation of elites than do their predecessor societies. Moreover, if the dominance of elites in 'traditional' societies was often expressed in direct, face-to-face contexts, elite dominance in industrial society is increasingly mediated by bureaucracies, investment markets, and educational systems, distancing elites from those below them. And if the privileges of elites in 'traditional' societies were often brazenly hereditary, reproduction of elites in industrial societies is, in theory at least, less regulated by kinship. Thus, Marcus (1989) and Marcus and Dobkin (1992),who has studied traditional elites on the Pacific Island of Tonga and dynastic families in the US, argues that, while kinship appears to define the transmission and reproduction of elite status for the Gettys just as much as it does for the Tongan aristocracy, the analogy is superficial and deceptive: for families such as the Gettys kin relations of inheritance constitute a shell that conceals the corporate and bureaucratic character of the contemporary dynastic family dependent, as it is, on lawyers and money managers for its functioning and reproduction.

Cohen (1981) points out that in industrial, bureaucratic societies recruitment to elites is generally advertized as meritocratic, yet principles of heredity often find a way to suborn processes of promotion ostensibly based on merit as members of elites transmit their privileges to their children. For example, private schools become a means of converting the wealth of one elite generation into the superior educational attainment of its children, allowing the circumvention of the meritocratic prohibition on the simple inheritance of elite position (Baltzell, 1987; Khan, 2011). Moreover, as Bourdieu (1986) and Lamont (1992) argue, elites come to value taste and manners as signs of distinction when recruiting a new generation of leaders, and accomplishment in this realm is asymmetrically transmitted to the children of existing elites as they are subjected to esoteric training in manners and taste in informal social contexts within the family, in private clubs, in peer groups, and at private schools. Secret societies, such as the freemasons and the fraternities at elite universities in the United States, can also play a role in short-circuiting the processes of meritocracy (Armstrong and Hamilton, 2013).

Anthropologists engaged in fieldwork in Western societies for many years tended to ignore Western elites. Thus, for example, anthropologists in the 1950s and 1960s kept out of the lively, at times bitter, debate in sociology and among public intellectuals provoked by the publication of C. Wright Mills' (1956) argument that, far from being a meritocratic society governed by a plurality of separate elites, the United States was ruled by a single 'power elite.' According to Mills this elite class was loosely integrated by the backstage functioning of private schools, elite universities, private clubs, and the marriage market and had become adept at disguising from the mystified masses its privileges and its degree of integration. Instead of joining this debate on elites in US society, American anthropologists engaged in fieldwork at home at this time were busy writing about 'the culture of poverty' or, in a logical sequel to earlier fieldwork abroad, studying processes of acculturation and economic struggle in Western enclaves of ethnic migrants from Third World countries.

A milestone in the recognition of asymmetries in the anthropological gaze came with the publication of Laura Nader's celebrated (1969) article calling on anthropologists to 'study up.' Nader drew attention to the problems of imposing a special visibility on the poor and powerless and appealed to anthropologists to shine an ethnographic spotlight on the functioning of power in the United States. Her article is often misrepresented as a call for anthropologists to engage in bounded studies of elite cultures, where in fact she called for studies of 'vertical slices' of social life that would enable the analysis of hierarchical relationships between elites and others and dramatize the essential nature of those relationships for the grounding of elite status and power. This was a way of reworking the traditional holism of anthropology. However, if the first generation of anthropologists had subordinated the study of elites to the analysis of societies as social wholes, Nader's is a call for holism that puts the privileges and power of elites front and center.

Nader observed that 'studying up' would probably require modifications of anthropology's fetishized research technique of face-to-face participant observation, and this has indeed proved to be the case (Stryker and Gonzalez, 2014). Given that elites are often mobile, busy, protected by minders, resistant to being studied, and expensive to socialize with, it is unlikely that

they can be studied using the same techniques anthropologists have developed for peasants, hunter-gatherers, and the urban poor. In order to study elites effectively anthropologists are having to rethink traditional notions of bounded field sites (Marcus, 1995), and to weave naturalistic personal interactions with their subjects together with archival and media research, use of big statistical data, online interaction, and formal interviews with informants too busy to meet more than once – a research strategy I refer to as "polymorphous engagement" (Gusterson, 1997). Examples of ethnographic studies of elites that depart from conventional methodological norms include Harding's (2000) study of the powerful American preacher, Jerry Falwell, which evoked the world he had created without access to Falwell and his aides; Rose's (1989) highly literary response to Nader's call to study 'vertical slices' of American life, juxtaposing miniaturized vignettes from the lives of poor African-Americans in Philadelphia with scenes from the lives of business elites and the country gentry in Pennsylvania; Vine's (2011) study of decisions by the US and the UK military elites to take a Pacific Island from its inhabitants, which makes use of archival sources and formal interviews, as well as participant observation with the displaced islanders; Aldrich's (1997) autoethnographic study of the Rockefeller dynasty, of which he himself is a member; and the sociologist Jennifer Pierce's (1995) study of a law firm by means of covert fieldwork conducted under cover of employment as a paralegal – a strategy that violates the ethics code of the American Anthropological Association if not the American Sociological Association.

Some of the most interesting work in recent years in the ethnography of elites has explored the practices of financial elites, elite taste cultures, and the world of scientific elites. This work has been undertaken in a context where, as many theorists of neoliberalism have observed, the global capitalist economy is increasingly financialized, increasingly dependent on a burgeoning new class of knowledge producers, and increasingly marked by levels of social and economic inequality that have not been seen in the West for a century (Castells, 2009; Gouldner, 1979; Harvey, 2007; Piketty, 2014).

One might think that bankers and financial traders would be the most inaccessible of elites to ethnographic study, but three fine studies, all by younger female anthropologists, take readers deep inside the highly masculine world of financial elites, exploring how they are recruited and socialized and what they do at work, how their work has been transformed by digital technology, how they have remade the corporate world through leveraged buyouts, how they come to believe they deserve multimillion dollar bonuses, and how their practices helped produce the near meltdown of the global financial system of 2008 (Ho, 2009; Tett, 2010; Zaloom, 2010).

Financial elites sit at the top of an increasingly unequal socioeconomic system where the wealthy are withdrawing into enclaves that are spatially segregated (Low, 2004) and aesthetically marked. Building on the work of Bourdieu (1986) and Lamont (1992) on the use of taste cultures by the wealthy to build networks and invisible barriers to socioeconomic mobility, a number of studies have explored elite taste cultures. Two of the most widely read have been sharply observed quasi-ethnographic studies by the journalists David Brooks (2001) and Robert Frank (2008). Beal (2000) and Czegledy (1998) analyze the mansion-building strategies of Jordanian and Hungarian elites respectively, parsing their careful hybridization of Western and local aesthetics, while Steiner (1994) and Myers (2002) track the international art market that transforms indigenous art work from Africa and Aboriginal Australia into objets d'art that signal the worldly sophistication of the global elites that purchase them. Wulff (1999), also working at the intersection of elite culture and aesthetics, writes on the culture of international ballet dancers.

Some of the most rigorous and innovative ethnographic work on Western elites has appeared in the new field of science studies. Perhaps because of their academic culture, scientific elites have proved unusually amenable to ethnographic study. Many ethnographic studies of scientists probe the symbolic prestige of science and the means by which scientists exercise their special license to declare what is true. Latour and Woolgar (1979) show how biologists make their claims authoritative; Traweek (1988) explores the cultural authority of particle physicists in Japan and the United States, while Gusterson (1996) examines the culture of American nuclear weapons scientists, and Helmreich (1998) documents the world of scientists at the elite Santa Fe Institute, and the way seemingly obscure scientific experiments may anchor and replicate dominant political ideologies. Luhrmann (2000) has written a powerful study of the socialization of American psychiatrists – an increasingly important category of expert in the production of American culture and identity. Meanwhile, exploring the scientific endeavor most productive of new wealth, Rabinow (1996, 2002) and Sunder Rajan (2006) have penned studies of the conjoining of venture capital and scientific entrepreneurship in the biotech industry's remaking of the world. This new wave of studies of elites still leaves gaps. One gap concerns political elites. Weatherford's (1981) study of the US Congress is now over 30 years old. While Wedel (1999, 2011) has written important studies of the transnational elites that manage the global development machinery and of patronage practices among the American political class, we need more studies of the way political elites function in the United States and elsewhere. That means studies of lobbyists, regulators, elected representatives, political consultants, and others. And, Bickford (2011) notwithstanding, we need more studies of military elites. In a world that is powerfully constructed by media, we also need studies of media elites and their relationship to financial, military, and other elites. Pedelty (1995) has written a vivid account of the journalistic elite that shaped Western understandings of wartorn Central America in the 1980s, exploring the ways in which these journalists' articles are shaped in response to the interests of elites both in Central America and in the United States, and Hannerz (2004) explores the culture of foreign correspondents more generally, but we also need studies of television and radio journalists, Hollywood productions, and of the elites who own media.

The ethnographic study of elites has the potential to stretch anthropology methodologically and theoretically, and to illuminate the hidden mechanics of an increasingly unequal global society. Anthropologists have historically gone into opaque cultural settings and rendered them transparent to their audience. If one of the crutches of elite power is invisibility, anthropologists have an important potential contribution to make to democracy and public discourse, as well as to

anthropological scholarship, in adapting and redeploying their traditional methodology for the study of elites.

See also: Anthropology at Home; Art, Anthropological Aspects of; Big Man, Anthropology of; Caste; Colonialism, Anthropology of; Economic Anthropology; Ethnography; Fieldwork in Social and Cultural Anthropology; Finance, Anthropology of; Globalization, Anthropology of; Hegemony: Anthropological Aspects; Inequality: Comparative Aspects; Kinship in Anthropology; Marriage; Modernity: Anthropological Aspects; Money: Anthropological Aspects; North America: Sociocultural Aspects; Patron–Client Relations, Social and Anthropological Study of; Political Anthropology; Political Economy in Anthropology; Postcoloniality; Science and Technology, Anthropology of.

Bibliography

Aldrich, N.W., 1997. Old Money: The Mythology of America's Upper Class. Allworth Press, New York.

Armstrong, Elizabeth A., Hamilton, Laura T., 2013. Paying for the Party: How College Maintains Inequality. Harvard University Press, Cambridge, MA.

Baltzell, E.D., 1987. The Protestant Establishment: Aristocracy and Caste in America. Yale University Press, New Haven.

Barth, Fredrik, 1959. Political Leadership among Swat Pathans. Althone Press, London.

Beal, E. Anne, 2000. Real Jordanians don't decorate like that! The politics of taste among Amman's elites. City and Society 12 (2), 65–94.

Bickford, Andrew, 2011. Fallen Elites: The Military Other in Post-Unification Germany. Stanford University Press.

Billig, Michael S., 2007. The interests of competing elites: fighting over protectionism and "free markets" in Phillipine sugar. Culture, Agriculture, Food and Environment 29 (2), 70–77.

Bottomore, Tom, 1993[1964]. Elites and Society. Routledge, NY.

Bourdieu, Pierre, 1986. Distinction. Harvard University Press, Cambridge, MA.

Brooks, David, 2001. Bobos in Paradise: The New Upper Class and How They Got There. Simon & Schuster, NY.

Castells, Manuel, 2009. The Rise of the Network Society: The Information Age: Economy, Society, and Culture Volume 1. Wiley Blackwell, Oxford.

Cohen, Abner, 1981. The Politics of Elite Culture: Explorations in the Dramaturgy of Power in a Modern African Society. University of California Press, Berkeley.

Czegledy, Andre, 1998. Villas of wealth: a historical perspective on new residences in post-socialist Hungary. City and Society 10 (1), 245–268.

Dumont, Louis, 1970. Homo Hierarchicus. University of Chicago Press, Chicago.

Ferguson, James, 1990. The Anti-Politics Machine: Development, Depoliticization, and Bureaucratic Power in Lesotho. Cambridge University Press, New York.

Fox, Richard G., 1985. Review of Elites: Ethnographic Issues by George Marcus. American Anthropologist 87 (4), 930–932.

Frank, Robert, 2008. Richistan: A Journey through the American Wealth Boom and the Lives of the New Rich. Three Rivers Press.

Geertz, Clifford (Ed.), 1963. Old Societies and New States. Free Press, Glencoe, IL.

Gouldner, Alvin, 1979. Future of Intellectuals and the Rise of a New Class. Seabury Press, New York.

Gusterson, Hugh, 1996. Nuclear Rites: A Weapons Laboratory at the End of the Cold War. University of California Press, Berkeley.

Gusterson, Hugh, 1997. Studying up revisited. Political and Legal Anthropology Review 20 (1), 114–119.

Hannerz, Ulf, 2004. Foreign News: Exploring the World of Foreign Correspondents. University of Chicago Press, Chicago.

Harding, Susan, 2000. The Book of Jerry Falwell. Princeton University Press, Princeton.

Harvey, David, 2007. A Brief History of Neoliberalism. Oxford University Press, Oxford.

Helmreich, Stefan, 1998. Silicon Second Nature: Culturing Artificial Life in a Digital World. University of California Press, Berkeley.

Hertz, Rosanna, Imber, Jonathan (Eds.), 1995. Studying Elites Using Qualitative Methods. Sage, Beverly Hills, CA.

Ho, Karen, 2009. Liquidated: An Ethnography of Wall Street. Duke University Press.

Khan, Shamus Rhaman, 2011. Privilege: The Making of an Adolescent Elite at St. Paul's School. Princeton University Press, Princeton.

Khan, Shamus Rhaman, 2012. The sociology of elites. Annual Review of Sociology 38, 361–377.

Kuper, Hilda, 1961. An African Aristocracy: Rank Among the Swazi. Oxford University Press, London.

Lamont, Michele, 1992. Money, Morals, Manners: The Culture of the French and the American Upper-Middle Class. University of Chicago Press, Chicago.

Latour, Bruno, Woolgar, Steve, 1979. Laboratory Life: The Social Construction of Scientific Facts. Sage, Beverly Hills.

Leach, Edmund, Mukherjee, S.N. (Eds.), 1970. Elites in South Asia. Cambridge University Press, Cambridge.

Low, Setha, 2004. Behind the Gates: Life, Security and the Pursuit of Happiness in Fortress America. Routledge, New York.

Luhrmann, Tanya, 1996. The Good Parsi: The Fate of a Colonial Elite in a Postcolonial Society. Harvard University Press, Cambridge, MA.

Luhrmann, Tanya, 2000. Of Two Minds: The Growing Disorder in American Psychiatry. Knopf, New York.

Marcus, George (Ed.), 1983. Elites: Ethnographic Issues. University of New Mexico Press, Albuquerque.

Marcus, George (Ed.), 1989. The problem of the unseen world of wealth for the rich: toward an ethnography of complex connections. Ethos 17 (1), 114–123.

Marcus, George (Ed.), 1995. Ethnography in/of the world system: the emergence of multi-sited ethnography. Annual Review of Anthropology 24, 95–117.

Marcus, George, Dobkin, Peter, 1992. Lives in Trust: The Fortunes of Dynastic Families in Late Twentieth Century America. Westview Press, Boulder, CO.

Mills, C. Wright, 1956. The Power Elite. Oxford University Press, New York.

Milner, Murray, 2014. Elites: A General Model. Polity Press, London.

Mosca, Gaetano, 1960. The Ruling Class. McGraw-Hill, New York.

Myers, Fred, 2002. Painting Culture: The Making of Aboriginal High Art. Duke University Press.

Nader, Laura, 1969. Up the anthropologist: perspectives gained from studying up. In: Hymes, Dell (Ed.), Reinventing Anthropology. Random House, New York.

Pareto, Vilfredo, 1935. The Mind and Society. Harcourt, Brace, New York.

Pedelty, Mark, 1995. War Stories: The Culture of Foreign Correspondents. Routledge, New York.

Pierce, Jennifer, 1995. Gender Trials: Emotional Lives in Contemporary Law Firms. University of California Press, Berkeley.

Piketty, Thomas, 2014. Capital in the Twenty-First Century. Belknap Press, Cambridge, MA.

Rabinow, Paul, 1996. Making PCR: A Story of Biotechnology. University of Chicago Press, Chicago.

Rabinow, Paul, 2002. French DNA: Trouble in Purgatory. University of Chicago Press, Chicago.

Rose, Dan, 1989. Patterns of American Culture: Ethnography and Estrangement. University of Pennsylvania Press, Philadelphia.

Sahlins, Marshall, 1958. Social Stratification in Polynesia. University of Washington Press, Seattle.

Steiner, Christopher, 1994. African Art in Transit. Cambridge University Press, New York.

Stryker, Rachel, Gonzalez, Roberto (Eds.), 2014. Studying Up, Down and Sideways: Anthropologists Trace the Pathways of Power. Berghahn Books, Brooklyn, NY.

Sunder Rajan, Kaushik, 2006. Biocapital: The Constitution of Post-Genomic Life. Duke University Press, Durham.

Tett, Gillian, 2010. Fool's Gold. Free Press, New York.

Traweek, Sharon, 1988. Beamtimes and Lifetimes: The World of High Energy Physics. Harvard University Press, Cambridge.

Vincent, Joan, 1971. African Elite: The Big Man of a Small Town. Columbia University Press, New York.

Vine, David, 2011. Island of Shame: The Secret History of the U.S. Military Base on Diego Garcia. Princeton University Press, Princeton, NJ.

Weatherford, Jack, 1981. Tribes on the Hill. Rawson, Wade, New York.

Wedel, Janine, 1999. Collision and Collusion: The Strange Case of Western Aid to Eastern Europe, 1989–1998. St. Martin's, New York.

Wedel, Janine, 2011. Shadow Elite: How the World's New Power Brokers Undermine Democracy, Government and the Free Market. Basic Books, New York.

Wulff, Helena, 1999. Ballet across Borders: Career and Culture in the World of Dancers. Berg, Oxford.

Zaloom, Caitlin, 2010. Out of the Pits: Traders and Technology from Chicago to London. University of Chicago Press, Chicago.

Elites: History of the Concept

Philipp Korom, Max Planck Institute for the Study of Societies, Cologne, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term 'elite' was introduced to the vocabulary of the social sciences in the late nineteenth century by Vilfredo Pareto. Pareto, Gaetano Mosca, and Robert Michels developed classic elite theories in opposition to the Marxist belief that a popular proletarian regime would yield democratization. A ruling minority, so the basic theorem goes, will always out-organize large majorities. Contrary to the classics, the social sciences since World War II have mostly treated elites and democracy as reconcilable. Technical terms such as 'strategic elites' (Keller) were widely used to refer to incumbents of top positions in the various social sectors. The 'functionalist' or 'pluralist' school of thought departs from the idea of heterogeneity. 'Critical elite theory' claims that, even in modern democracies, economically based social hierarchies are reproduced, thereby giving rise to a single dominant class.

Long before the term 'elites' was coined, the concept had been used by various thinkers. Already Plato (*c.* 380–360 B.C.E./2002) imagined in *Republic* (book VII) a state ruled by philosopher-kings ('kings of a hive, better and more perfectly educated than the rest'). Plato's justification of such a rule is that philosophers only have knowledge of what really is, e.g., just or good. That genuine philosophical knowledge is needed for ruling, is the primary thrust of his argument.

Niccolò Machiavelli (1469–1527), an Italian diplomat who is commonly regarded as a predecessor of the classic elite theories, used the terms 'nobles,' 'aristocrats,' or 'the great' interchangeable to describe society's elite. In *The Prince*, Machiavelli (1531/2005) makes several recommendations for acquiring, governing, and preserving principalities in response to the political crisis of Italian city-states, including his own city-state of Florence: a prince should be both loved and feared; he needs able advisers; his best fortress is the loyalty of his subjects; and cruelty can be legitimate to maintain order. In a famous passage, Machiavelli advises the prince to emulate both the fox and the lion, cleverly outmaneuvering his foes ('find out the Snares') and courageously standing up against enemies ('fright away the Wolves'). In this first European secular treatise on politics, elites are considered as a threat to democracy, in the sense that they will bias the actual enacting of laws in their own interest. Machiavelli advises princes to base their power on the majority of citizenry rather than on the wealthy elite. His early ideas on power-checking mechanisms to constrain elites have been taken up in modern democratic theory (MacCormick, 2011). Classical elite theory adopted, among other things, not only the lion and fox metaphor, but also the idea that different styles of leadership were related to the rise and fall of elites as well as the conviction that self-ishness is a natural property of the human species.

Henri de Saint-Simon (1760–1825), a French utopian socialist, developed his ideas on those at the apex of social hierarchy with the French Revolution fresh in mind. He argued that the landed aristocracy ceased to be the dominant class and that its place was taken by entrepreneurs, financiers, and scholars – people whose importance derived from function rather than status. Saint-Simon illustrated this vital point when speculating on the loss to society if the best 'three thousand scientists, artists, and artisans of France' were taken away: "The nation would become a lifeless corpse … It would immediately fall into a position of inferiority compared with the nations which it now rivals, and would continue to be inferior until this loss had been replaced, until it had grown another head" Saint-Simon (1819/1964: 72).

Like Saint-Simon, other leading aristocratic critics of mass tendencies did not refer to 'elites,' although their ideas on large-scale processes of massification were established along the juxtaposition of the 'noble man' and the 'mass man' (Ortega y Gasset). The French social psychologist Gustave le Bon (1841–1931) and the Spanish philosopher José Ortega y Gasset (1883–1955) were both concerned with the crowd/the mass of the mediocre entering fields of decision-making (Le Bon, 1895/1897; Ortega y Gasset, 1929/1957). The 'era of crowds' (Le Bon) was suspected to result in cultural degradation and deterioration. Human excellence, it was ascertained, was only to be found in a few and not the many. Society itself was seen as aristocratic in its very nature: it can never be the many, the average, the mass, who rule. Mass, then, referred primarily to the quality of human beings and their numbers rather than to social class. Elites, in contrast to the masses, share a superior moral and psychological makeup.

The term 'élite' first appears in an encyclopedia published in France between 1751 and 1772 and edited by Denis Diderot (*Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers*). The encyclopedia entry suggests that the word was originally used to describe exquisite, top-of-the-line products and that its usage was later extended to designate membership in a superior social group ('*hommes d'élite*'). The term was, however, only introduced to the vocabulary of the social sciences in the late nineteenth century. A first broad definition of 'elite' in the social sciences is to be found in *Trattato di sociologia generale* of 1915–19 (translated as *A Treatise of General Sociology*) by the Italian economist Vilfredo Pareto (1848–1923):

> Let us assume that in every branch of human activity each individual is given an index which stands as a sign of his capacity, very much the way grades are given in the various subjects in examinations in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03079-8

> school. The highest type of lawyer, for instance, will be given 10. The man who does not get a client will be given 1 – reserving zero for the man who is an out-and-out idiot …. So let us make a class of the people who have the highest indices in their branch of activity, and to that class give the name … of élite.
>
> Pareto (1916/1935: 1422–23)

Pareto himself did not make any further use of this abstract concept. However, the idea that what counts are the skills and activities that people bring to their work was later taken up by functionalism, making Pareto their main precursor.

Classic Elite Theorists

The classics in elite theory were fierce opponents of Marxism (Bottomore, 1993). First, they criticized it for establishing a theory of socioeconomic development upon a single principle, which is the struggle between the unpropertied working class and the property-owning class. In their view, one has to take a multicausal approach to the study of human history instead of pursuing any deterministic 'single-track doctrine' (Mosca). Second, they denied that a classless society, as envisaged by most socialists, could ever be attained by arguing that minority domination is an inescapable feature of human life. Third, they challenged the view that primarily economics and not politics were the determining force in history.

Mosca: The Ruling Class

Gaetano Mosca (1858–1941) posits in his work entitled *Elementi di scienza politica* of 1896 (translated as *The Ruling Class* in 1939) that social change is determined not by class structure but by the political apparatus and an organized minority in control of this apparatus. Regardless of the type of political organization, this minority imposes its will on the disorganized majority. Mosca's argument hinges crucially on the paired terms 'organized' and 'disorganized.' He acknowledges that the ruling class in a parliamentary system will normally be divided into two or more competing parties and that the election outcome may have an effect on government policy. It is simply, however, that there will always be a discrepancy between *de iure* and *de facto* authority, or between formal political structure and informal political power, whatever democratic mechanisms may be employed (Femia, 1993). For Mosca (1896/1939: 53), the key to elite control lies in the minority's organizational capacity: "A hundred men acting uniformly in concert, with a common understanding, will triumph over a thousand men who are not in accord." He adds, however, that ruling minorities are endowed by 'a certain material, intellectual, or even moral superiority' (Mosca, 1896/1939: 53). It is not genetic heredity that contributes to the superiority of character, but special personal qualities, manners, habits, and wealth, which are inherited from similarly endowed ancestors. The intergenerational transfer of wealth and educational privileges especially tends to transform the rulers into a hereditary caste.

Despite this hereditary basis of power, the political class does not simply rule by force but always endeavors to justify its power in the eyes of those being governed by a set of values, the 'political formula.' In a Machiavellian fashion, Mosca thus emphasized the need for a moral basis to maintain power. He was convinced that political change originated from conflicts within the elite and not between elite and mass. It is, however, only in his criticism of Marx that Mosca draws a sharp distinction between the mass and an organized minority, while for the most part he presents a more subtle depiction in which he distinguishes between the top governmental leaders and a lower stratum consisting of high-level civil servants and others. The latter serves as a bridge between the ruling class and the rest of society. Despite a preeminent 'aristocratic' tendency to recruit members from the ruling class itself, parts of the lower stratum are granted access to positions of power and amalgamate with the elite ('democratic' tendency). The most enduring social organization is thought to be a mixed government (partly aristocratic, partly democratic), so that the political class does not atrophy or become unstable through a lack of continuity. In that way, the efficacy of government would go hand in hand with a certain measure of accountability to the interests of the people as a whole. A stable society could only be maintained if all 'social forces,' that is, all major institutional spheres such as religion or education, are represented within the uppermost class of society.

Pareto: The Governing éLite

Instead of 'ruling class,' Vilfredo Pareto (1848–1923) introduced the term 'élite' in his 1916 work *Trattato di sociologia generale* (translated as *A Treatise of General Sociology* in 1935), in order to avoid the Marxist intellectual baggage of the former term. In an initial, broad definition, 'élite' encompasses all individuals with the highest ranking in their social activities based on their skills and regardless of their moral and social qualities. Pareto, however, does not indicate where he thinks a feasible cutoff exists for separating the elite from the nonelite. Most likely, he abandons this formal definition because, in reality, a certain ascriptive status, along with wealth and birth, give the less skilled an advantage for attaining leading positions. In a more narrow definition, he subdivides the class of maximum performers into 'governing' and 'nongoverning' sections, the first of which includes all individuals who 'play directly or indirectly a noteworthy role at the highest levels of power' (Pareto, 1916/1935: 1423). How the governing elite, which is not necessarily marked by superior ability, goes about its job, is not specified. Like Mosca, Pareto is convinced that all societies are ruled by organized minorities, also in parliamentary democracies, and that only the elites are responsible for change, not the masses.

Pareto places particular emphasis on psychological characteristics as the basis of elite rule. Notably, Pareto employs Machiavelli's lion and fox typology by applying it not only to alternative control strategies but also to changing elite psychologies. Famously, Pareto theorized that fundamental distinctions can be made among the governing elites according to the predominate 'residues,' with residues being – roughly speaking – nonlogical propensities rooted in people's basic aspirations and drives. Within an elite, these are classified as either 'Class I' residues ('instincts of combinations') – innovative men skilled in intrigue and possessing cleverness in negotiation, the 'foxes' in Machiavelli's language – or 'Class II' residues ('persistence of aggregates') – conservative men with

strong sentiments of order and discipline, the 'lions.' Just as individual atoms form molecules in nature, so, too, did Pareto group individuals upon the basis of common 'residues' (Marshall, 2007: 33). Pareto specifically conceived of residues as manifestations of human sentiment. Residues are thus not thought of as concrete realities, but as concepts created by the observer, which implies, for example, that a man's (hidden) taste for pork chops does not fall into the analytical category 'residues.' If, however, a man constructs a theory according to which the Chinese cuisine is superior to the American, Pareto would suggest analyzing the (irrational) residues underlying this theory (Coser, 1977: 391).

While politics requires both sets of qualities, 'Class II' residues tend to weaken over time. One of Pareto's key arguments is that there is a constant circulation of individuals within the governing elite. Elites decline, degenerate, and die, sometimes at a fast rate, sometimes slowly: "It flows on like a river, never being today what it was yesterday. From time to time sudden and violent disturbances occur. There is a flood - the river overflows its banks. Afterwards, the new governing élite again resumes its slow transformation" Pareto (1916/1935: 1431). History, in Pareto's words, is a 'graveyard of aristocracies.' Typically, new elites seizing power consist of lions who are eventually replaced by more cunning foxes who, in turn, meet their destiny in the form of new lions, and so on. Change for Pareto is cyclical, not linear, and it is change that maintains the equilibrium in a system.

Michels: Political Leaders and the 'Iron Law of Oligarchy'

"It is organization which gives birth to the dominion of the elected over the electors, of the mandataries over the mandators, of the delegates over the delegators. Who says organization, says oligarchy." This postulate, spelled out in Robert Michels' 1911 seminal book *Political Parties* (1911/1999: 365), became encapsulated in the thesis of the 'iron law of oligarchy.' It essentially posits that democracy contains an oligarchical nucleus because even the most democratically committed organizations, such as the Social Democratic Parties in Europe studied by Michels (1876–1936), become divided into elite fractions who pursue their own interests within the organization by merely exploiting the labor and resources of the average party members. By declaring that the exercise of political power is an inherently oligarchic enterprise, Michels implicitly assumes that the elite, or rather the oligarchs, are characterized by what Meisel (1958) called the '3 Cs': group consciousness; group coherence; and conspiracy, that is, a common will to action. Structural forces, or more precisely, the 'technical indispensability of leadership' (Michels, 1911/1999: 364), push parties, like any other large-scale organization that has moved out of amateurishness, toward oligarchy. A political party needs to organize its vote, canvass supporters, raise contributions, attend to its financial structure and legal standing; all these activities require hard-to-acquire expertise and scarce organizing talents that limit the capacity of the party's rank and file to challenge elected leaders' decisions or to replace the leadership, while the leaders recognize a common interest in steering the organization and maintaining their position. Due to the '3 Cs,' lower level protest is hobbled, conservative loyalists promoted to the party staff, and power inexorably concentrated in the higher echelons of bureaucratic hierarchies. Psychological forces, such as the oligarchs' desire for power and the masses' psychological need for leadership, supplement the resulting estrangement between the leaders and the led. Michels puts himself clearly in a Machiavellian tradition when he writes that the 'the desire to dominate, for good or evil, is universal' (Michels, 1911/1999: 206), and he suggests a modified version of Pareto's circulation of elites when speaking of the 'réunion des elites' (Michels, 1911/1999: 182): one group of elites does not replace another. Rather, there is an ongoing process of power-sharing pacts, with established elites absorbing and assimilating newcomers.

Despite the fact that classic elite theorists had not developed a clear elite concept and most of their arguments were overly general and lacking concrete substance, they succeeded in providing bearings for empirical elite research that became centered primarily on four major topics: elite profiles, elite recruitment, elite change, and elite cohesiveness (Aron, 1960: 272). Mosca, Pareto, and Michels, who regarded themselves as empirical social scientists rather than social philosophers, were either ambivalent or outright critical of democratic movements: from their standpoint, an elite-dominated 'democracy,' in which career politicians impose themselves on the voters, is the most that is possible (Best and Highley, 2010). The degree to which these thinkers discredited themselves by their attitude toward fascism differs. Mosca, a member of the Italian Senate, took to the floor on different occasions to argue against law bills endorsed by the fascist leader Benito Mussolini. Pareto, an engineer who turned to economics later in life, welcomed the advent of fascism, although fascist ideologists did not find much merit in his works (Cirillo, 1983). The sociologist Michels converted from a revolutionary socialist to a fascist ideologue and accepted a chair of politics from Mussolini's government. Later defenders of elite theory clearly distanced themselves from the 'Machiavellians' and Michels by primarily arguing that the role of elites is not irremediably opposed to a substantial degree of democracy.

Pluralist Elite Theories

Mannheim, Dahrendorf, Dahl, Keller: Functional, Strategic Elites

After World War II, the social sciences generally shared the basic functionalist assumption that modern societies are characterized by structural differentiation and thus focused predominantly on the pluralist elite structure, in which the various meritocratic subsystems (the economy, politics, media, etc.) develop their own distinctive status hierarchies, with different elite positions linked to the apex of power. Elites are thus primarily those with power and resources at their disposal, thanks to professional achievement (Khan, 2012). The emergence of these autonomous elites was perceived as central to the foundation of democracy (Etzioni-Halevy, 1990). Arguably, the most important precursor of pluralist elite theories is the German sociologist Karl Mannheim (1893–1947). Although Mannheim noted that it was "the important contribution of modern democracy that the achievement principle increasingly becomes the criterion of social success" (Mannheim, 1935/1940: 86), he interpreted,

somehow paradoxically, the general public participation as having a deleterious effect on societal elites. Mannheim feared that the increased size of the elite, the 'proletarization of the intelligentsia,' as well as inter- and intraelite conflicts could deprive society of the continuity it needs; therefore, he offered an elitist solution to the threat of 'negative democratization.' Education should promote unity in the age of massification by instilling a sense of group cohesiveness and a common *Weltanschauung* within future democratically orientated elites.

Similar to Mannheim, Ralf Dahrendorf (1929–2009) contended that there are multiple elites in industrial societies. In his theoretical approach, society consists of a plurality of interest groups, which engender a plurality of conflicts. Elites are theorized as 'quasi-groups' (Dahrendorf, 1962: 310) and thus as a loose formation of individuals who occupy positions that have the same role interests. If latent common interests become conscious goals, then elites can turn into 'conflict groups.' Originally, Dahrendorf (1962) distinguished between the prestige upper class, the economic upper class, and the ruling class, or power elite. In his later work, however, all three quasi-groups are classified as elites without any differentiation (Dahrendorf, 1972: 126–136).

The American political scientist Robert A. Dahl (1915–2014) moved beyond classical elitism by discrediting any ruling elite model as quasimetaphysical; instead, he supported that of a polyarchy. For Dahl, social structure was not analogous in shape to a single pyramid, but to a range of pyramids. He argued that only a careful examination of a series of concrete decisions could establish the dominance of a specific group within a nation or community. Consequently, in his work *Who Governs?* (Dahl, 1961), he studied three political issues in New Haven, Connecticut, namely, party nominations for local elective offices, public education, and urban redevelopment. Dahl found no signs of a single elite operating behind the scenes but rather many lines of cleavage and politicians who were responsive to the desires of the citizenry.

Finally, Suzanne Keller (1927–2010) coined the term 'strategic elites' to refer to multiple, nonintegrated groups with power in the diverse arenas of modern society. Her underlying orientation is akin to the structural functionalism of Talcott Parsons. Strategic elites compromise "not only political, economic, and military leaders, but also moral, cultural, and scientific ones. Whether or not an elite is counted as strategic does not depend on its specific activities but on the scope of its activities, that is, on how many members of society it directly impinges upon and in what respects" (Keller, 1963/1991: 20). Elites, in this sense, are those with sustained social impact. These 'strategic elites' overwhelmingly act in accordance with the functional requirements of their status (i.e., the political elite is primarily responsible for deciding the goals to be pursued). Individual achievement and not class background is the chief principle for recruitment to strategic elite positions. Keller stresses cohesiveness regarding values as the key to the effectiveness of 'strategic elites.' Goals may differ, but value consensus among elites is theorized as a prerequisite for prosperity.

The majority of social scientists today endorse pluralist theories of elites, with various nuances and variations. Common to all is the argument that power is dispersed rather than concentrated in the hands of a few.

Critical Elite Theory

Mills, Bourdieu: Power Elite, Field of Power

Not surprisingly, pluralist elite theories have come under fire. In his provocative work *The Power Elite* (1956), C. Wright Mills (1916–62), a radical in spirit, argued that American society is controlled by single cohesive elite that constitutes the leadership of three major areas of society: military, business, and government. These three elites only appear to be separate. Due to the similarity of their social origins, close personal and family relationships, overlapping cliques between the principal power holders in each sector, and institutional proximity, those in different higher circles "conceive of themselves to be, and are felt by others to be, the inner circle of the 'upper social classes'" (Mills, 1956: 11). Mills preferred the term 'power elite' over 'ruling class,' because the latter term was judged to be a 'badly loaded phrase' as it did not give autonomy to the political order (Mills, 1956: 277). In a Marxist interpretation, the ruling capitalist class not only controls society's wealth but also decides on a society's political policy. By contrast, Mills believed that power has shifted from the owners of the means of production to managers and functionaries of the principal institutions of American society, such as high-ranking politicians. Members of the power elite are consequently defined as those 'in command of the major hierarchies and organizations of modern society,' who have the power to make decisions of national and international consequence (Mills, 1956: 4). Moreover, Mills subdivides the American power structure as a whole into three levels: (1) the close-knit power elite, an alliance between the highest military, corporate, and political leaders; (2) the middle level of power, meaning local elites, members of Congress, judges, pressure groups, and other organized groups; and (3) the virtually powerless, economically dependent and unorganized masses. In essence, those on top issue orders, which obedient administrators at the second level transmit to the population at the third level. Nearly all of Mills' assertions on the American power elite have been criticized as exaggerated or unproven (Dahl, 1958).

The French sociologist Pierre Bourdieu (1930–2002) discards the substantialist notion of 'ruling class' or 'elite' in favor of the relational concept of 'field of power' (Bourdieu and Wacquant, 1993). The field of power is the locus of struggle between agents who already hold dominant positions in their respective social fields, such as art, politics, or the economy. In this general space, France is the state that particularly but not exclusively exercises influence over the interrelations between specific fields because it holds a monopoly on the educational field and thus controls the entry points into all other fields. Within Bourdieu's theoretical framework of capital forms, the topological notion of 'field of power' enables various internal divisions resulting from struggles over which particular type of capital should be recognized as the most important. Drawing on extensive empirical studies, Bourdieu (1989/1996) creates in *State Nobility* a topology of the French educational system, especially the elitist *grandes écoles*, and relates it to the internal structure of the field of power. By applying correspondence analysis to 84 institutions, he discovers two striking homologies between both fields. First, the difference between major and minor schools reflects contrasts between the *haute bourgeoisie* (industrialists, private sector executives) and the *petty*

bourgeoisie (business owners, technicians, school teachers). Second, the animosity between the more academic schools (such as the *École normale supérieure*) and schools explicitly established to train its students to assume power positions (such as the *École des hautes etudes commerciales*) reproduces a fissure within the dominant class that separates the cultural from the economic bourgeoisie.

In general, Bourdieu identifies the *grandes écoles* not only as power institutions of the elite, but also as fortresses of isolation, which are just as difficult to leave as to enter. He shows that the social and family factors dictate not only the choice of more or less esteemed educational institutions, but also how the criteria of university examiners reflect the values of the dominant classes. From an analysis of a survey of winners of a national high school academic competition (*Concours Général*) and other sources, Bourdieu finds that these gate keepers in elite institutions rely on vague contrasts (brilliant versus dull, gifted versus motivated) that systematically discriminate against lower class pupils and thus contribute to the reproduction of class inequalities via the educational system.

Synopsis

The concept of elites has a long and checkered history. Initially, it was used to describe the rule of a minority over the majority of the population. Associated with it were usually rather vague meanings as 'fewness' and 'excellence.' Precursors of classic elite theory like Ortega (1929/1957) used dichotomies such as 'noble man' versus 'mass man' to refer to the few of noble spirit who place great demand upon themselves in the pursuit of excellence, on the one side, and the many ordinary men who are not concerned about individuality and consider themselves equal to everyone else, on the other. The category of 'mass' is not socioeconomic and class based; it is primarily intellectual excellence that marks the elite. Mosca gives the concept a different twist by introducing the term 'ruling class,' which essentially refers to an organized minority who monopolizes the decision-making power in society. He imagined the members of the 'ruling class' as leaders of excellence who could promote the moral and material interests of a society. Pareto defined elites as persons with the highest capacities in their field. These elites have to be efficient to stay at the top, that is, they must make optimal use of scarce resources to best benefit society. Such efficiency demands different styles of governing that range from adaptive politics (elites strong in class I residues) to aggressive and authoritarian measures (elites strong in class II residues). Finally, Michels was concerned with the emergence of oligarchs within complex, large-scale organizations that allow people in top positions to monopolize organizational means.

Most contemporary concepts of elites depart from a simplified juxtaposition of 'mass' and 'elite.' The pluralist school of thought emphasizes that, with the increasing differentiation of societies, power is dispersed among multiple leadership groups who bargain and compromise over key societal decisions. To identify a unique elite becomes impossible. Society is conceptualized instead as a number of parallel pyramids, each capped by an elite who hold the largest share of a given resource. These elites are accountable to public scrutiny and open to people from all social strata. Mills' concept of the 'power elite' differs from the prevailing pluralist vision on elites because it implies that few upper class people are essentially able to control very different societal sectors. Bourdieu also stands apart because he posits that the dominant class is able to maintain its traditional claim to key leadership positions in society. His work on the French 'state nobility' shows that the acquisition of exclusive educational degrees opens the door to top positions in companies for the few who are already in privileged social positions and discriminates against the many who are less privileged.

All elite concepts, although different, have some similarities. Each concept deals with the criteria of distinction and the sources of power of the elites. Table 1 offers a systematic overview of these rather hidden commonalities.

Table 1 Different Meanings of the Elite Concept

	Criterion of distinction	*Source of power*	*Domain*
Ortega	Intellectual excellence	No power	Society
Mosca	Excellence	Number of organizations	Society
Pareto	Efficiency	Force and consent	Society
Michels	Position within the organization	Monopoly in the use of organizational means	Complex organizations
Pluralists	Possession of high degree of resources	Various resources	Society
Mills	Social status and position within institutions	Monopoly over every source of power in society	Society
Bourdieu	Social status and exclusive educational degrees	Transformation of cultural capital into economic capital	Society

Note: Partly based on Zannoni, Paolo, 1978. The concept of elite. European Journal of Political Research 19, 1–30.

See also: Bourdieu, Pierre (1930–2002); Elites: Sociological Aspects; Field Theory; Functionalism, History of; Mannheim, Karl (1893–1947); Mills, Charles Wright (1916–62); Pareto, Vilfredo (1848–1923).

Bibliography

Aron, R., 1960. Classe sociale, classe politique, class dirigeante. European Journal of Sociology 1, 260–282.

Best, H., Highley, J., 2010. Democratic Elitism: New Theoretical and Comparative Perspectives. Brill, Leiden.

Bottomore, T.B., 1993. Elites and Society. Routledge, London/New York. Stanford University Press, Stanford.

Bourdieu, P., 1996. The State Nobility: Elite Schools in the Field of Power (Originally Published in 1989).

Bourdieu, P., Wacquant, L.J.D., 1993. From ruling-class to field of power – an interview with Pierre bourdieu on La Noblesse d' Etat. Theory, Culture & Society 10, 19–44.

Cirillo, R., 1983. Was Pareto, Vilfredo really a "precursor" of fascism? American Journal of Economics and Sociology 42, 235–246.
Coser, L., 1977. Masters of Sociological Thought. Ideas in Historical and Social Context. Harcout, San Diego.
Dahl, R.A., 1958. A critique of the ruling elite model. American Political Science Review 52, 463–469.
Dahl, R.A., 1961. Who Governs? Democracy and Power in an American City. Yale University Press, New Haven.
Dahrendorf, R., 1962. Eine neue deutsche Oberschicht? Notizen über die Eliten der Bundesrepublik. Die Neue Gesellschaft 9, 18–31.
Dahrendorf, R., 1972. Konflikt und Freiheit. Auf dem Weg zur Dienstklassengesellschaft. Piper, München.
Etzioni-Halevy, E., 1990. Democratic-elite theory: stabilization versus breakdown of democracy. European Journal of Sociology 31, 317–350.
Femia, J., 1993. Mosca revisited. European Journal of Political Research 23, 145–161.
Keller, S., 1991. Beyond the Ruling Class. Strategic Elites in Modern Society. Transaction Publications, New Brunswick (Originally Published in 1963).
Khan, S.R., 2012. The sociology of elites. Annual Review of Sociology 38, 361–377.
Le Bon, G., 1897. The Crowd: A Study of the Popular Mind. Unwin, London (Originally Published in 1895).
Machiavelli, N., 2005. The Prince. Oxford University Press, Oxford/New York (Originally Published in 1531).
MacCormick, J.P., 2011. Machiavellian Democracy. Cambridge University Press, Cambridge.
Mannheim, K., 1940. Man and Society in the Age of Reconstruction. Harcourt, Brace, New York (Originally Published in 1935).
Marshall, A., 2007. Vilfredo Pareto's Sociology: A Framework for Political Psychology. Ashgate, Aldershot.
Meisel, J.H., 1958. The Myth of the Ruling Class: Gaetano Mosca and the 'Elite'. University of Michigan Press, Ann Arbor.
Michels, R., 1999. Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracy. Transaction Publishers, New Brunswick, NJ (Originally Published in 1911).
Mills, C.W., 1956. The Power Elite. Oxford University Press, New York.
Mosca, G., 1939. The Ruling Class. McGraw-Hill, New York/London (Originally Published in 1896).
Ortega y Gasset, J., 1957. The Revolt of the Masses. Norton, New York (Originally Published in 1929).
Pareto, V., 1935. The Mind and Society: A Treatise on General Sociology. Harcourt Brace Jovanovich, New York (Originally Published in 1916).
Plato, 2002. Republic. In: Cahn, S.M. (Ed.), Classics of Political and Moral Philosophy. Oxford University Press, New York, Oxford, pp. 32–167 (Originally Published *c.* 380-360 B.C.E.).
Saint-Simon, H., 1964. Social Organization, the Science of Man and Other Writings. Harper & Row, New York (Originally Published in 1819).
Zannoni, Paolo, 1978. The concept of elite. European Journal of Political Research 6, 1–30.

Relevant Websites

http://hyperbourdieu.jku.at/hyperbourdieustart.html – A comprehensive, contextual, and referential bibliography and mediagraphy of all works and public statements by Pierre Bourdieu.
http://encyclopedie.uchicago.edu/ – Electronic edition of Diderot and d'Alembert's Enyclopédie.
http://www.cwrightmills.org/ – The official C. Wright Mills website, maintained by his son.
http://www2.ucsc.edu/whorulesamerica/ – A site by G. William Domhoff about how power is distributed and wielded in the United States.

Elites: Sociological Aspects

Michael Hartmann, TU Darmstadt, Darmstadt, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E. Etzioni-Halevy, volume 7, pp. 4420–4424, © 2001, Elsevier Ltd.

Abstract

The term 'elites' refers to those people who, due to their positions, are capable of considerably influencing social developments. Classical elite theories assume an unavoidable opposition of ruling elite and dominated masses; functionalist elite theories assume competing performance and sectoral elites; critical elite theories identify power elites recruited on the basis of social origin. Current elite research discusses most of all the homogeneity of elites, the role they play in the context of an ever more unequal distribution of incomes, and their internationality.

In the current language, 'elite' refers to a small group of people standing out from the overall population due to particular characteristics. The term is used both for top politicians and top managers as well as for outstanding athletes or scientists. Both outstanding achievements and an unusual degree of influence as well as unusual power are considered the essential criteria for considering somebody an elite member. The social sciences apply a somewhat stricter definition. According to the latter's criteria, only those people belong to an elite who are capable of essentially influencing social developments. It is debated, however, which positions must be taken into consideration in this context. This refers first of all to individuals whose social influence is not due to occupying hierarchical power positions but to their personal reputation or prestige. What is also not clear is the limits which must be drawn concerning the rank of the respective position.

Most social scientists include only the top ranks of the classical fields of politics, business, administration, justice, armed forces, media and sciences into their studies. Among them, apart from office bearers at the national level in federally constituted states such as Germany or the United States, there count the top politicians, high ranking public officials or judges of the individual federal states. Mayors are counted only if they govern leading cities of the respective country. Other regional or local elites are usually not taken into consideration. For elite research, top politicians, top managers, high ranking administrative civil servants and judges, top army officers, the publishers, and leading editors of the most important media as well as the top representatives of the big scientific institutions and elite universities form the important social elites.

Individual authors go clearly beyond. This holds, e.g., for the fundamental analyses by Dreitzel (1962) and Keller (1963). For example, Dreitzel states explicitly that not only those occupying top positions, thus immediately contributing to preserving or changing society, must be counted among the elite but also those who, due to their prestige, are capable of acting as role models, thus normatively influencing the behavior of the population (Dreitzel, 1962). Also, Keller includes not only the top ranks of politics, business, sciences, or armed forces into her 'strategic elites,' which she calls 'external elites.' Among elites she also counts outstanding representatives of philosophy and churches as well as famous artists, authors, entertainers, film stars, or athletes, so-called 'internal elites,' who in her opinion are responsible for the formulation of moral standards and for maintaining everyday morality (Keller, 1963). However, for extended empirical elite studies this second group plays hardly any role at all. They deal almost exclusively with the top representatives of the classical sectors. If other fields, such as culture, are represented at all, this does not refer to famous artists, writers, etc., but to cultural managers, i.e., those occupying leading positions of publishing houses and the arts sections of the media.

Accordingly, in modern industrial societies the crucial elites come from the fields of business, politics, administration, and justice. In this respect they are the most influential groups. If the boards or owners of big trusts make decisions on their enterprises' future orientations, if a government decides about austerity measures or reforms of the educational system, if high ranking ministry officials formulate draft bills, or if the judges of federal courts make their verdicts – in all these cases it is a small group of people who makes crucial decisions concerning the life of the broad population. For the elites from other important sectors, such as the media, the armed forces or the sciences, this definition is true only in a limited sense. Indeed also they are capable of influencing social developments, but concerning crucial aspects they depend on the decisions most of all of political and business elites who, after all, decide about available financial means and overall direction. In individual countries however, such as most of all the United States, also the armed forces count among the crucial elites. Concerning the top representatives of trade unions, churches or NGOs, it is debated if they count among the elites. Most studies from Europe do count them, some studies from the United States do not. The reason is mostly that, other than elites from business, politics, administration, justice, the media or the sciences, they actively or passively (in the form of membership fees) depend on support from their members. The less this support, the less power they have. This may reach as far as to being mostly or even completely insignificant. The power of actual elites, on the other hand, is not connected to being actively supported by the population but solely to occupying an office and/or disposition of property in the form of enterprises. It is structurally rooted in the power structure of society. This holds also for the political elite, as is demonstrated by low and constantly declining turnouts and party

memberships which do not at all restrict the scope of action of the political elite.

The crucial questions, when it comes to the analysis of elites, are How do elites recruit themselves, and what about their inner cohesion? In other words, it is about the vertical and horizontal integration of elites. The vertical dimension concerns the connection between elites and population, i.e., most of all the social origin of elites and the thus connected attitudes. Horizontal integration refers to the way in which the individual elites are connected to each other, by way of social origin, educational, and career paths as well as common attitudes and values.

Classical Elite Theories

All of the classical elite theories, those by Mosca (1896), Pareto (1916), and Michels (1911), were formulated at the turn of the nineteenth to the twentieth century. Their development and great impact is due the bourgeosie's widespread fear of the urban masses and the organized working class in those days. In this extent, the urban masses were a historically new phenomenon. They developed only in the second half of the nineteenth century, as a result of comprehensive industrialization and urbanization (Hartmann, 2007). Theoretically, the elite theories were the alternative to Marxism, which dominated the organized working class. The predecessor of the elite theories was Le Bon's very influential psychology of the masses (1895). There, Le Bon states that the masses, most of all due to being driven by their instincts, are incapable of creating cultures but could only destroy them. Cultures could be created only by small, intellectual aristocracies of superior minds. Accordingly, he says, the masses need leaders who guide them by use of force.

The contradiction between small, ruling minorities, and vast majorities who are ruled is also the bottom line of classical elite theories. For them the division into a ruling class or elite – Mosca calls it the ruling class, in Pareto it is the ruling class but also the ruling or governing elite – and a mass which is ruled is the basic principle which is valid for all societies and forms of government. The elite rules due to its superior material and organizational possibilities as well as to its intellectual capabilities and by use of force. In Mosca's and Pareto's opinion, all elites tend toward monopolization and bequeathing their power positions. Thus, there is a constant conflict between these aspirations and the upward aspirations of parts of the masses. Either these upward aspirations are taken into consideration by way of the permanent integration of individual elements of the dominated class, or the ruling class will be overthrown in the course of revolutions. Pareto calls this process the elite cycle. Michels, whose analysis refers basically to political parties, completes Pareto's and Mosca's explanations by stating that all human organizations inevitably tend toward an oligarchy.

The Functionalist Elite Theories

Due to their both topical and (in the case of Michels and Pareto) also personal closeness to Fascism, after 1945 the classical elite theories were deeply discredited. They were replaced by the still today dominating functionalist or pluralist elite theories, from Lasswell and Kaplan (1950) via Dahl (1961), Dreitzel (1962), and Keller (1963) as far as to Field and Higley (1980). Their two core statements are clearly different from the classical elite theories. In their opinion, there is no more ruling class or elite but only individual, competing subelites in the various fields of society. Also, they say, there is no lasting hierarchy of these subelites. Secondly, access to these elites happens primarily via individual performance. The still observed predominance of the upper classes among the elites is said to be due to their better access to education instead of heredity and will step by step disappear with the extension of the educational system.

However, according to the functionalist elite theories, the existence of subelites and access via performance causes a grave problem. The heterogeneity of elites, due to their division into different sectors and their recruitment from different classes, is said to endanger the consensus among the elites. Whereas in traditional societies it was created by way of the existence of a ruling elite and a common social origin, these days it cannot be taken for granted. However, such a consensus on basic values and the way of settling conflicts is said to be indispensable for modern societies. To establish it, as is the opinion of most functionalist elite researchers, the elites must support their positions offensively. Only the elites, they say, are capable of securing the stability of parliamentary democracies. Stability is never, as Field and Higley say, the result of the voluntary and peaceful cooperation of all social actors but always only the result of prudent decision-making by the elites. Thus, the danger does not come from the elites but from too much influence of the masses. Field and Higley support this position particularly harshly, like Lasswell and Kaplan before them, but also almost all other functionalist elite sociologists see things similarly. For example, Keller believes too strong demands for democracy, equality, and elite responsibility to be one of the greatest dangers for democracy in modern societies. After all, all these elite researchers agree that the elites' scope of action must not be restricted by the influence of the broad population.

The Critical Elite Theories

Already at a very early stage, in the 1950s, Mills (1956) criticized the functionalist elite theories. Then most of all Bourdieu (1979, 1989) discussed them critically. Both theoreticians agree that the basic assumptions of the functionalist elite theories are wrong. In their opinion, there are neither subelites of equal rank nor recruitment according to the performance principle. Concerning the former point, still Bourdieu in most of his works even uses the term 'ruling class.' However, in contrast to the classical elite theories he does not equal it with the elite but refers to that of the three great social classes from which those occupying top positions are predominantly recruited. Mills uses the term 'power elite.' This way he wants to make obvious that for him 'class' is an economic concept, but rule is a political one, and also he wants to avoid the risk of underestimating the autonomy of the political elite.

For Mills, the power elite consists of the top representatives of politics, business, and armed forces. In his opinion, these three sectors dominate all other sectors because those decisions as being decisive for the development of society are made by those being at the top of the highly centralized power apparatuses of politics, business, and armed forces. Most influential in this context, he says, are the top managers of the big trusts, whose power nobody can resist in the long run. The interwoveness of the three sectors as well as common interests, together with a mostly identical origin from the upper classes and a mostly common worldview, provide a high degree of stability and cohesion within the power elite. Apart from growing up in the upper classes, he states, also being educated at a country's elite universities and the recruitment criteria and values common among these circles are the reason for social homogeneity.

Concerning the significance of social origin and elite educational institutions, Bourdieu mostly agrees with Mills. This holds also for another crucial point. Also Bourdieu considers the owners of economic capital to play the dominating role within the ruling class, thus calling them the ruling faction of the ruling class. However, in contrast to Mills, he is not interested in the question of who makes the important socially relevant decisions and how they are made but first of all in the reproduction of the ruling class. Most of all, in his more recent publications he even explicitly distances himself from elite theories because, as he says, he is interested in the analysis of power structures and not in analyzing power actors. In his opinion, a class-specific habitus and elite universities are crucial for the reproduction of the ruling class. They make sure that the important power positions can be handed over to the offspring even without being directly bequeathed. By way of the class-specific habitus, he says, an individual's experiences from growing up in a certain family environment are impersonated. It is a system of dispositions determining perception, thought and behavior, and it mediates between an individual's position in the social space and his/her lifestyle. This habitus, he states, decides essentially about the chances of being accepted at one of the elite universities and of acquiring their exclusive academic grades. The latter, on the other hand, are then the precondition for rising to the important power positions. This way the reproduction of the ruling class is secured. In contradiction to its claims, Bourdieu says, the educational system is not characterized by equality of opportunity but by subtle mechanisms of the reproduction of class structures. Thus, access to top positions is indeed not primarily based on individual performance but on the individuals' class-specific habitus.

Bourdieu and Mills do not only share their rejection of the basic theses of functionalist elite theories: In contrast to the latter, as essential explanatory elements they always include the class situations in the respective societies into their analyses. In Bourdieu, this is obvious because it is one, if not even the essential, aspect of his entire work. But also in Mills the class situation plays a crucial role. This becomes obvious most of all when he explains the declining influence of the political elite, in particular of professional politicians, in comparison to business and military elites. In his opinion, two processes are responsible for this: The decline of the traditional middle class of economically rather independent farmers and freelancers in favor of the new, economically dependent, middle class of white-collar workers and the loss of power of the organized working class. The significance of the traditional middle class and (even more) the organized working class, he states, were the basis of the (most of all during the New Deal) unusually big scope of action of the political elite. The mass of the population does not contribute anymore to the political life but is manipulated by the mass media. Thus, for Mills mass society and power elite form an inseparable unity.

Recent Developments in Elite Analysis

In the field of sociology the interest in the topic of elites has clearly grown in the past two decades. Since the turn of the millennium alone there have been quite a number of publications dealing first of all with the analysis of European elites, although in single cases also the elites of non-European countries such as Japan (Colignon and Usui, 2003) and the United States (Domhoff, 2009) are discussed. There exist both compilations of smaller country studies, such as on the political elites of Southern Europe (de Almeida et al., 2002) or the economic elites of East Europe (Lane et al., 2007), and publications which, on the basis of the authors' own empirical surveys, deal with the elites of Germany (Hartmann, 2013), Denmark (Christiansen et al., 2001; Ellersgard et al., 2013), France, and Great Britain (McLean et al., 2006), Norway (Gulbrandsen et al., 2002; Hjellbrekke and Korsnes, 2009) or of all of Europe (Hartmann, 2010). Usually it is about the social profile of elites, i.e., their social recruitment, their educational and career paths and their connections to each other. The social–political attitudes of elites are included into the surveys only in the cases of Germany and Norway (Gulbrandsen et al., 2002; Hartmann, 2013). Theoretically, the functionalist concept of elites is still dominating; the majority of current empirical studies (Colignon and Usui, 2003; Domhoff, 2009; Ellersgard et al., 2013; Hartmann, 2013; Hjellbrekke and Korsnes, 2009; McLean et al., 2006), however, takes a critical attitude following the tradition of Mills and Bourdieu.

The great interest in elites is basically due to two grave social changes: The fall of the Eastern Block and the rapidly growing gap between the poor and the rich in most industrial states. For the first time after the end of World War II, the end of the Eastern Block has raised the urgent question in Europe if and in how far there happens an exchange of traditional elites. The fast growing differences between the poor and the rich raise the question of the role of the elites in this context.

Concerning the development in the countries of the former Eastern Bloc, regarding the elites the situation looks much varied. There are countries where rather soon the old elites were mostly replaced. With the exception of the special case of the GDR, where the old elites were almost completely replaced by people from the old Federal Republic of Germany, this holds most of all for the Baltic states, Poland, Czech, and Hungary. In Poland, Czech, and Hungary this exchange concerns most of all the sectors of politics, administration, justice, armed forces, and media, whereas in the business sector often members of the higher management ranks simply moved one step further, to the top of enterprises. Thus, in this field the elites have only

been partly replaced. In a number of other countries, such as most of all the successor states of former Yugoslavia and Russia, this process started only later and most of all with the political elite. There, the members of the former political elite did not simply leave or join the opposition parties, but after having been replaced quite often they took over top positions in business (Eyal et al., 1998; Lane et al., 2007). In some states, such as Bulgaria, Romania, or White Russia, the exchange of elites has been only very limited or hardly at all.

The rapidly growing differences in income in many industrial countries have resulted in an intensive debate on the question if and in how far the elites, whose scopes of action have clearly been extended due to the massive weakening of trade unions as a result of globalization, are responsible for this development. What is most of all of significance in this context is the question of how far the elites have been closing off from the rest of the population. Current studies pursue this question most of all by asking about the social recruitment and the educational and career paths of elites. Their results can be shortly summarized as follows: In most European countries the social recruitment of elites has become more exclusive, and also regarding their educational and career paths the elites have become more homogeneous. However, in this respect there are great differences in West Europe. On the whole, there exist three basic types of elite formation: firstly France, which knows both institutions of elite education and elite corps of state administration; secondly Great Britain and Spain, who have institutions of elite education or elite corps, respectively; and thirdly the other countries which have neither the former nor the latter (Hartmann, 2010).

In case of the first two types, the elites are more closed off from the population than in the third case. The essential reason is the elite institutions, which are attended by the majority of their members. These institutions do not only provide for social exclusivity, due to their access procedures, but also for a high degree of internal homogeneity due to the time their members have spent together when attending them. The overwhelming majority of the elites in these countries come from the upper or upper middle class, i.e., the upper *c.*4% of the population in the generation of their fathers. Concerning the business elite, about four-fifths come from these classes, with the elites of justice, armed forces, politics, and administration it is between three-fifths and three-quarters. Due to their origin and the time they have spent together at institutions of elite education and/or elite state institutions, these elites achieve an outspokenly high degree of internal homogeneity. In other West European countries, the degree of elite homogeneity is traditionally clearly lower. There exist neither institutions of elite education nor common career paths. Furthermore, the individual subelites are very much different concerning their social origin. On the one hand, there is usually the business elite, whose social recruitment is frequently the most exclusive one; on the other hand, there are most of all the political and/or administrative elites which are the most open ones in this respect (de Almeida et al., 2002; Hartmann, 2010; McLean et al., 2006). However, social origin varies much from country to country. Socially most exclusive are the elites in Germany, and this exclusiveness has partly even been growing in the past decade. In business, their recruitment is as exclusive as it is in the case of the elites in France or Great Britain, in the justice and administrative sectors it is almost as exclusive, with a share of two-thirds of upper or upper middle class children. Also with the political elite, which for half a century was clearly dominated by people from the broad population, social recruitment has clearly changed since the turn of the millennium. These days, about one half of top politicians come from the upper or upper middle class, whereas in the past it was only one-third. Furthermore, changing from elite positions in one sector to those in another sector has become much more frequent. The homogeneity of German elites has thus much increased (Hartmann, 2013). The Scandinavian states form the antipole in Europe. They are the least socially exclusive ones. There, people from the broad population are represented much more frequently than with the other European elites. This holds in particular for those coming from the working class. Almost in every sector elite members from the upper or upper middle class are in the minority. Furthermore, in the past decades the elites have somewhat opened up. However, the appropriate data for the Danish business elite are debated, as becomes obvious by the debate on the result of the Danish elite study by Christiansen et al. (2001) (Ellersgard et al., 2013; Gulbrandsen et al., 2002).

Furthermore, for Norway Gulbrandsen (2008) analyzed in more detail the mutual trust among elites as an essential precondition for elite consensus. He presents four essential results. Firstly, mutual trust is lower than originally assumed by the researchers. Secondly, the subelites mostly trust the elites of comparable sectors, e.g., the military elite has confidence in the elites of justice and police, or the administrative elite has confidence in the political elite. Thirdly, the basic political attitude toward the welfare state plays a crucial role for mutual trust. This is an essential, if not the only, reason why, fourthly, of all elites the business elite has the least confidence in the political elite and why at the same time all other elites have comparatively little confidence in the business elite. In both respects, the latter is an outsider.

On the German business elite there exists a study which, by way of the careers of 6500 engineers, jurists, and economists with a doctoral grade, explains the exclusive social recruitment of this elite and doubts the functionalist thesis of performance-oriented recruitment. It shows that social origin and the thus connected class-specific habitus influence the access to elite positions, not only by way of achieving high educational degrees but quite directly. Even in case of identical preconditions concerning educational features, the probability for those with a doctoral grade coming from the upper class to reach the uppermost floor of a big enterprise is more than double than for their mates from the broad population. If their fathers were board members, their probability is even 17 times as high as for their mates whose fathers were working class (Hartmann, 2002).

These differences in the social recruitment of elites come hand in hand with differences in the distribution of incomes. In this respect, the Scandinavian countries show the least differences, Great Britain and the big continental-European countries the biggest ones. The statistical relation between the distribution of incomes and elite recruitment can be explained if one looks at the relation between the social origin of elites and their social–political convictions and

attitudes. The most recent study on German elites (Hartmann, 2013) has analyzed this connection in more detail.

The attitude of the elites, as is the result, is clearly characterized by their social origin, most of all when it comes to judging on social differences and the reasons for the current financial crisis. Whereas almost three quarters of the population consider the social differences in Germany to be unjust, among the elites only a minority is of this opinion. However, even among their ranks there are considerable differences. Whereas working class children consider social inequalities to be unjust almost two and a half times as those considering them just, among upper class children the relation is almost exactly the opposite. The answers, if it is about important finance- and labor market-political measures, are appropriate. Whereas among the population two-thirds support a rise of the top tax rate, two-thirds of elite members are against it. In this context, again and again they answer according to their social origin. Two and a half time as often as others, elite members with a working class background believe tax increases for high incomes and great fortunes to be important. This difference is basically found with all subelites, if at different levels respectively. Concerning the essential reasons for the financial crisis, the situation is the same. Those elite members as coming from working class families clearly believe the deregulation of the financial markets to be the most important reason, a four-to-one majority of those who grew up in upper class families, on the other hand, believed the national debts to be the most important reason. The majority of working class children identified a responsibility of deregulated markets, upper middle, and upper class children saw a responsibility of the state and, most of all, social benefits. This attitude is found across all sectors. Thus, the connection between the social recruitment of elites and the development of incomes is not only one of statistical data but is explained by the convictions and attitudes of elite members which are influenced by their social origins.

One question elite researchers have increasingly, be dealing with in the past one and a half decade is that of the internationalization or globalization of elites. In this context, for the time being research focusses on business elites because as a result of globalization they are much more internationalized than the other elites which, after all, in their great majority are tied to the institutions of the national state. The question is answered very differently. Some researchers state that in business meanwhile there exists a global elite. They point out that the big multinational enterprises are increasingly interwoven, that networks become more intensive, and that accordingly their boards are ever more international. Among the worldwide business elites, as is their position, transnationalists increasingly replace national networkers. Thus, as, e.g., Carroll (2009) states in his study on the boards of the 500 biggest enterprises in the world, we may rightly so speak of a 'global corporate elite.' Kentor and Jang (2004) in their study on the boards of the Fortune Global 500 argument similarly. Also they come to the result that the emergence of a transnational business community is a matter of fact. However, both Carroll and Kentor and Jang admit (in contrast to some other, less empirically working authors) that for the time being this process is mostly restricted to Europe and North America and only marginally includes the business elites of Asia or South America.

The low degree of internationalization among Asian top managers is demonstrated by Pohlmann (2009) in a study on the CEOs of the biggest enterprises in Germany, the United States, China, Korea, and Japan. Among the CEOs of the 300 biggest enterprises of the three Asian countries there are just four foreigners. From this he draws the conclusion that the thesis of global business elites does not stand the test of empirical evaluation. Also Hartmann's study (2011) on the CEOs of the 100 biggest enterprises in Germany, France, Great Britain, Italy, Spain, the United States, China, and Japan points in this direction. With 9 and 18%, respectively, only in Germany and Great Britain a considerable share of CEOs comes from foreign countries. In other countries, on the other hand, it is just 5% or even nobody at all. As furthermore, two-thirds or even nine-tenths of national CEOs have no foreign experience of more than 6 months, one may not speak of an internationalization of business elites or even of a global elite. The still dominating significance of national career paths and institutions of elite education must be considered the decisive reason for this.

See also: Aristocracy, Nobility, and Gentry, European History of; Bourgeoisie and Middle Classes, History of; Caste; Class: Social; Elites, Anthropology of; Income Distribution and Inequality: Measurement Issues; Mobility, Social; Pluralism; Power in Society; Social Inequality in History (Stratification and Classes); Social Mobility, History of; Social Stratification; Wealth Distribution.

Bibliography

de Almeida, P.T., Pinto, A.C., Bermeo, N. (Eds.), 2002. Who governs Southern Europe? Regime Change and Minsterial Recruitment, 1850-2000, South European Society & Politics, 7 (2), (Special Issue).

Bourdieu, P., 1979. La Distinction, Critique sociale du jugement. Éditions de Minuit, Paris, Trans. Distinction: A Social Critique of the Judgement of Taste. Harvard University Press, Cambridge, MA.

Bourdieu, P., 1989. La noblesse d'Etat. Grandes écoles et esprit de corps. Éditions de Minuit, Paris, Trans. The State Nobility. Elite Schools in the Field of Power. Polity Press 1996, Cambridge, MA.

Carroll, W.K., 2009. Transnationalists and national networkers in the global corporate elite. Global Networks 9 (3), 289–314.

Christiansen, P.M., Møller, B., Togeby, L., 2001. Den Danske Elite. Hans Reitzels Forlag, København.

Colignon, R.A., Usui, C., 2003. Amakudari: The Hidden Fabric of Japan's Economy. Cornell University Press, Ithaca, NY.

Dahl, R., 1961. Who Governs? Democracy and Power in an American City. Yale University Press, New Haven, CT.

Domhoff, G.W., 2009. Who Rules America? Power and Politics, sixth ed. McGraw-Hill, Boston.

Dreitzel, H.P., 1962. Elitebegriff und Sozialstruktur. Eine soziologische Begriffsanalyse. Ferdinand Enke, Stuttgart.

Ellersgard, C.H., Larsen, A.G., Munk, M.D., 2013. A very economic elite: the case of the Danish top CEOs. Sociology 47 (6), 1051–1071.

Eyal, G., Szelényi, I., Townsley, E., 1998. Making Capitalism without Capitalists. Class Formation and Elite Struggles in Post-communist Central Europe. Verso, London.

Field, G.L., Higley, J., 1980. Elitism. Routledge & Kegan Paul, London.
Gulbrandsen, T., 2008. Elite integration an institutional trust in Norway. In: Sasaki, M. (Ed.), Elites: New Comparative Perspectives. Brill, Leiden, pp. 171–192.
Gulbrandsen, T., Engelstad, F., Klausen, T.B., Skjeie, H., Teigen, M., Østerud, Ø., 2002. Norske Makteliter. Gyldendal Norsk Forlag, Trondheim.
Hartmann, M., 2002. Der Mythos von den Leistungseliten. Spitzenkarrieren und soziale Herkunft in Wirtschaft, Politik, Justiz und Wissenschaft. Campus, Frankfurt.
Hartmann, M., 2007. The Sociology of Elites. Routledge, London/New York.
Hartmann, M., 2010. Elites and power structure. In: Immerfall, S., Therborn, G. (Eds.), Handbook of European Societies. Springer, New York, pp. 291–323.
Hartmann, M., 2011. Internationalisation et spécifités nationales des élites économiques. Actes de la Recherche en Sciences Sociales 37 (190), 10–23.
Hartmann, M., 2013. Soziale Ungleichheit – kein Thema für die Eliten? Campus, Frankfurt.
Hjellbrekke, J., Korsnes, O., 2009. Quantifying the field of power in Norway. In: Robson, K., Sanders, C. (Eds.), Quantifying Theory: Pierre Bourdieu. Springer, Heidelberg/New York, pp. 31–46.
Keller, S., 1963. Beyond the Ruling Class: Strategic Elites in Modern Society. Random House, New York.
Kentor, J., Jang, Y.S., 2004. Yes, there is a (growing) transnational business community: a study of global interlocking directorates, 1983–1998. International Sociology 19 (3), 355–368.
Lane, D., Lengyel, G., Tholen, J. (Eds.), 2007. Restructuring of the Economic Elites after State Socialism. ibidem Verlag, Stuttgart.
Lasswell, H.D., Kaplan, A., 1950. Power and Society: A Framework for Political Inquiry. Yale University Press, New Haven, CT.
Le Bon, G., 1895. Psychologie des foules. Alcan, Paris, Trans. The Crowd: A Study of the Popular Mind. Macmillan, New York, 1896.
McLean, M., Harvey, C., Press, J., 2006. Business Elites and Corporate Governance in France and the UK. Palgrave Macmillan, Houndmills and New York, NY.
Michels, R., 1911. Zur Soziologie des Parteiwesens in der modernen Demokratie. Untersuchungen über die oligarchischen Tendenzen des Gruppenlebens. Verlag Werner Klinkhardt, Leipzig, Trans. Political Parties. Hearst's International Library, New York, 1915.
Mills, C.W., 1956. The Power Elite. Oxford University Press, Oxford/New York.
Mosca, G., 1896. Elementi di Scienza Politica. Bocca, Turin, Trans. The Ruling Class. McGraw-Hill, New York, 1939.
Pareto, V., 1916. Trattato di Sociologia generale. Barbera, Firenze, Trans. The Mind and Society. A Treatise on General Sociology. Dover Publications, New York, 1935.
Pohlmann, M., 2009. Globale ökonomische Eliten? Eine Globalisierungsthese auf dem Prüfstand der Empirie. Kölner Zeitschrift für Soziologie und Sozialpsychologie 61 (4), 513–534.

Elster, Jon (1940–)

Charles HG Crothers, School of Social Sciences and Public Policy, AUT University, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Jon Elster has contributed to critiquing and also extensively constructing the standards against which to judge social science explanations, and rational choice theory, extending into institutional designs ensuring that analytical work obeys the injunction of methodological individualism and identified plausible mechanism, which explain outcomes, adding indeterminacies, emotions, and norms to the rational choices actors make under limited conditions. His multidisciplinary work has dissected Marxian theories but investigated an array of societal issues of modern times.

General

Jon Elster (born 22 February 1940, Oslo) is a Norwegian, currently Paris-domiciled social theorist who has contributed to the philosophy of social science; rational choice, game and collective or public choice theories; behavioral economics; institutional design models; the theory of distributive justice; and the history of social thought. He has also been involved with Analytical Marxism, most recently 'analytical sociology' and the critique of rational choice and public choice theory, especially on the grounds that they inadequately capture social realities. His more than 24 books to date (many published by Cambridge University Press) plus many edited volumes have been translated into many (17) languages. In addition, he has written *c*.200 articles. Elster's work has been multidisciplinary, covering the full array of the social sciences and extending into literature and history. He has never felt the need to seriously connect up with the associational life of particular academic disciplines. His work is studded with numerous examples from contemporary life, literature, and history.

Intellectual Trajectory

Both Elster's parents were writers: his mother (Magli Elster) wrote poetry while his father (Torolf Elster) wrote novels and short stories and was a political journalist for much of his career and then CEO of the Norwegian Broadcasting Corporation. His father had university degrees in mathematics, mechanics, and astronomy. The effect of this family situation was constant exposure to rigor and clarity, imagination, creativity and moral integrity together with left politics. His mother's family had lived in Paris where her father was a translator and writer of a Norwegian-French dictionary, so this provided some French ambience. These maternal links with France were reinforced through an exceptionally stimulating French teacher in high school. After his high school degree in 1958 and a summer job, Elster went to Paris for 3 months in order to learn French. At the Pris Alliance Française, there were many foreigners which allowed improvement of his English. He was Vice-President of the French students' Association, and married its President.

Mathematics and literature were early passions and these were supplemented by reading French and philosophy at the University of Oslo. After several years of study there, Elster wanted to study Marx, to establish more solid foundations for his intuitive socialism. But this, in turn, required study of Hegel. So, in 1966 he went to Paris to work with prominent Hegel scholar Jean Hyppolite. However, Hyppolite directed him instead to study with Pere Regnier, Gaston Fessard, and other Hegel scholars in a Jesuit monastery at Chantilly. The resulting dissertation (partway between a Masters and PhD in size) "Prise de conscience dans la Phenomenologie de l'Esprit de Hegel," was defended at the University of Oslo. While Elster found some parts of Hegel inspirational – such as his stunning analysis of the master–slave dialectic – he found most of the texts nearly impenetrable.

However, by 1968 he again wished to go to Paris, to now study Marx more generally and again he was sponsored by Hyppolite (then Director of the Ecole Normale Superieure) to obtain a place as 'pensionnaire etranger' and to supervise a dissertation for a Doctorat d'Etat. However, Hyppolite died so he turned instead to Raymond Aron (Directeur d'Etudes at the Ecole Pratique des Hautes Etudes) to be his supervisor. Elster found Aron's common sense and discussions at his seminar (including the interventions of classmates) were helpful. Outside of these seminars Aron and Elster only met once or twice a year to discuss drafts of dissertation chapters, and this left Elster to lay the foundation for much of his later work. Working backward from Marx meant coming to grips with Marxian economics and then to understand economic theory more generally, and especially its foundations in rational choice theory (RCT). Understanding Marx also required delving into economic history. Luce and Raiffa's *Games and Decisions* was an especially important source. Elster wrote a book in Norwegian (which was then included in his dissertation) on methodological problems of economic history, which led shortly thereafter to an academic post in Oslo. The thesis ('Production et reproduction: Essai sur Marx') was defended in 1972 in front of a jury consisting of Paris intellectual leaders Aron, Raymond Boudon, Jean-Claude Casanova, and Alain Touraine.

From 1971 to 1973, Elster held a University fellowship in Oslo and then taught Sociology for a year at the Universite de Paris VIII (Vincennes) with his work mainly focusing on Leibniz while he taught courses in Mathematical Sociology and Marxism.

In 1975, Elster became associate professor at the University of Oslo, with a joint appointment between the departments of

philosophy and history. Having failed to find a French publisher for his dissertation, he, therefore, ceased working on Marx, and instead drew on his readings in the rational-choice literature and on earlier training in logic to develop his work on rationality and its limitations. *Sour Grapes* (1983b) claims that there is a large class of states that are essentially by-products of action undertaken for other purposes and cannot be brought about intentionally. For example, one cannot achieve spontaneity or forgetfulness, or overcome insomnia, impotence or stuttering, by pure acts of will. Another claim is that there is a large class of preference structures, especially 'adaptive preferences' that depend on the options available to the agents rather than being independent of them, as is assumed in standard economic theory. For example, in the well-known fable the fox found the grapes sour only because he could not get to them. Counteradaptive preferences are the opposite case; e.g., "Forbidden fruit tastes best."

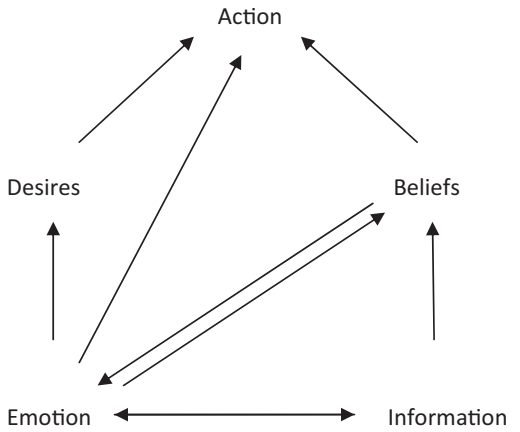

Explaining Technical Change (1983a) focused on explanation in the social sciences, with particular attention to the important topic of technology, which is widely seen as the driver of modern 'progress,' and was further developed in *Nuts and Bolts* (1986f). Having distinguished three main forms of scientific explanation: causal, intentional, and functional, Elster relates these to each of four theories of technical change: neoclassical theories, evolutionary theories, Marxian theories, and Schumpeter's theory. Elster insists on methodological individualism and attacks functionalist explanation.

G.A. Cohen's work on *Karl Marx's Theory of History* (1978) provided the opportunity for a reconsideration of Elster's Marxist work, after he had rested this interest for over a decade. Elster, Cohen, John Roemer, and others founded a small group of 'analytical Marxists' in 1979 that met annually (or biannually), although he bailed out in early 1990s. Its members produced a series of books published by Cambridge University Press, including as its first a wholly revised version of Elster's dissertation, *Making Sense of Marx* (1985). Elster criticized many areas of Marx's writing including his philosophy of history, economic theories, theory of the class struggle, and theory of the state. Nevertheless, Elster retained respect for Marx's normative theories a criticism of capitalist alienation, and Marx's holding out a life of active and communal self-realization as the positive alternative.

Elster then returned to issues of rationality in particular, reacting to a provocative newspaper editorial arguing that in child custody litigation the choice of a custodial parent should not always follow 'the best interest of the child' principle. Elster argued that it might be in the better interests of the child if a simple mechanical rule such as a presumption for maternal custody were to be uniformly followed or, more radically, choosing the custodial parent by the toss of a coin. Thus, emerged 'Solomonic judgements,' later (1989) published as a book with the same title. Following on from Otto Neurath, Elster argued that in many cases where reason is powerless to yield determinate outcomes, the best one can do is to leave the decision to chance.

Around the same time the Swedish Association of Employers asked Elster to write on the Swedish system of collective wage bargaining, which seemed to defy explanation in terms of RCTs of bargaining. Stronger labor unions, despite being able to realize their rational self-interest, which was based in bargaining from strength, choose not to do so because they felt constrained by powerful norms of equality. This drew Elster's attention to the role of norms.

In the mid-1970s, Elster had been visiting professor in the philosophy departments of Berkeley and Stanford a few times, and had made occasional guest appearances at the University of Chicago. He was invited in 1979 to come as a visiting professor to the Political Science Department at the University of Chicago, and received tenure there in 1984. Also, the thrust of his work was changing in this period from philosophy of the social sciences toward social science proper.

A Range of Books Were Produced Over the Following Decade

Ulysses & the Sirens (1984/1979) is composed of studies in a descending sequence from perfect rationality (characterized by an actor's capacity to relate strategically to the future), through imperfect and problematical rationality, to irrationality.

The Multiple Self (1986b) considers whether the self is a unity or as a divided 'multiple self' with the latter conception helping to understand problems of self-deception and weakness of will.

The Cement of Society (1989) addresses the contrast between rationality and norms more systematically, adding many other examples to that on wage bargaining. Elster argues that there are two extreme cases: the high-stakes e.g., norms of revenge, feuding or dueling and the low-stakes norms of e.g., etiquette, dress, speech, and points out that any unified theory of norms would have to include and explain both.

Local justice (starting around 1987 and published in 1992 a,b) was a project arising out of work on lotteries, which was a mechanism that had been used (in some societies but not in others) to allocate a large variety of goods and burdens e.g., admission to medical school or to military service in wartime. The allocative practices Elster studied operate in a very decentralized way, constrained – but far from determined – by central administrations. Elster also found that the actors' perceptions of what constitutes a fair allocation are a major determinant of the allocative system that emerges.

Political Psychology (1990/1993) shows how belief and preference formation in the realm of politics are shaped by social and political institutions.

By the late 1980s, Elster's immediate interests changed dramatically to comparing constitution-making processes (especially those of the late eighteenth century – the Federal Convention in Philadelphia and the Assemblee Constituante in Paris). Also, some 200 years later Communism began to crumble, leading to Quasiconstitutional Round Table Talks between the government and the opposition that place in many Eastern Europe countries, followed by constitutional amendments and, in some of these countries, election of assemblies with the task of writing wholly new democratic constitutions, which would then be significant in shaping political development. Elster reacted to these opportunities by being instrumental in establishing a Center for the Study of Constitutionalism in the Eastern Europe at the University of Chicago Law School, which would gather documentation about constitutional, legal, and political developments in the ex-Communist countries and elaborate a causal analysis of the process of constitution making and of political transitions more generally. Drawing on the work of Jurgen Habermas and Thomas Schelling, Elster claims that constitution-making processes contain elements of impartial arguing as well as of interest-driven bargaining (see Elster, 1993). Many apparently impartial claims are actually based on interest: e.g., when large parties argue for electoral systems based in plurality voting in single member districts, where they do not refer to the fact that this system benefits large parties, but appeal instead to the impartial values of efficiency and governability that are promoted by this procedure. *Deliberative Democracy* (1998) explores decisions by discussion or deliberation as alternatives to more formal voting. *Alchemy of the Mind* (1999) extends Elster's treatment of the role of the emotions in human behavior.

Elster's next set of research projects, in the new Millennium, included:

1. A book on de Tocqueville (arising out of a paper cotaught with Stephen Holmes), which argues that he ought to be read for his sociological, and above all psychological, insights which retain value despite being written some 150 years earlier.
2. The program 'Retroactive justice in New Democracies' addresses the question of how countries that have undergone a transition to democracy handle the transgressions of their prior elites. The scope of the study included the former Communist countries, the post-World War II European war crimes trials and recent transitions to democracy in Argentina, Chile, and other Latin American countries.
3. In 1992, Elster used funding from the Norwegian Research Council to carry out a 5-year study of addictions. Addiction poses, in an acute form, the problem of how voluntary, self-destructive behavior is possible.
4. Revisiting earlier work on the nature of the emotions and their place in the theory of action, Elster read every article on the emotions in the leading journal in the field (*Journal of Personality and Social Psychology*) from 1987 to 1992, and following up references. He found that many of the articles using American college students report the inconclusive results so, instead, in his study introspection, casual observation, novels, philosophy, and social science were drawn on, with the most useful insights coming from the classical French moralists – Montaigne, Pascal, La Rochefoucauld, La Bruyere – together with de Tocqueville and Proust.

Elster transferred to Columbia in 1995 in part because most of his friends and colleagues in Chicago had preceded him to East Coast Universities and he wanted to experience living in New York ('the world's capital'). In turn, his Columbia job allowed him a year off every 3 years, which he spent in Paris. Shortly after going to New York, Elster married a French-American literature scholar, whom he had met in Chicago.

A Further Slew of Studies (Still Ongoing) Has Eventuated

Closing the books (2004) examines how after a change of political system (especially transitions from an autocratic to a democratic, or at least constitutional, regime) a process of transitional justice emerges in which wrongdoers from the previous regime are judged responsible and victims are compensated. In addition to the numerous transitions after 1945 in Western Europe and after 1989 in Eastern Europe, transitional justice arrangements had taken place in classical Greece, the English, and French restorations.

Collective Wisdom (2012) proposes a normative theory of collective decision-making inspired by Jeremy Bentham that in designing democratic institutions the impact of self-interest, passion, prejudice, and bias on the decision-makers should be reduced as much as possible. This rule is prudent given that there is no independently defined good outcome that institutions can be organized to achieve and nor is there any way of reliably selecting good decision-makers.

These Institutional Design considerations are taken up further in *Securities Against Misrule: Juries, Assemblies, Elections* (2013). The obstacles Elster is concerned to guard against include:

- bad sheep in the process (domineering individuals or spoilers);
- ignorance, secrecy, and publicity;
- the effects of strategic behavior by participants;
- cyclical voting outcomes (Arrow paradox);
- dictatorship (institutional arrangements that permit one actor to determine the outcome);
- sensitivity of the collective outcome to the order of the agenda;
- procedural indeterminacy;
- power differentials;
- capture of the process by special interests;
- deadlock through requirements of super majorities; and
- information asymmetries among participants.

Among his recent work has been a trilogy – *Le désintéressement: Traité critique de l'homme économique* – which has consolidated his critique of economic man, mainly through publications flowing from a lecture series given in Paris.

Influences

Overall Elster has not been particularly influenced by extant 'masters' although collegial work with an array of contemporaries has been important. Crucial intellectual seedbeds have been the works of Leibnitz and Marx. But books have been particularly influential include: Robert Duncan Luce and Howard Raiffa's *Games and Decisions*, Thomas Schelling's *Strategy and Conflict*, Joseph Levenson's *Confucian China and its Modern Fate*, Keith Thomas's *Religion and the Decline of Magic*, the first volume of *Das Kapital*, the books by Paul Veyne (e.g., *Bread and Circuses: Historical Sociology and Political Pluralism*), Zinoviev's *The Yawning Heights*, de Tocqueville's several studies, John Rawls's *A Theory of Justice*, and Montaigne's *Essays*. These books had an impact as they offered perspectives on the world. Nonacademic books, of considerable importance, particularly include all the novels of Stendhal and Jane Austen.

Overview of Work

Elster is fluent in Norwegian, English, and French with a reading knowledge of German. Being multilingual, Elster has been sensitive to the potential of different languages for better expressing clear ideas. Moreover, he has lived not only in several urban regions but quite often simultaneously in two, although retaining affection for Oslo as his home town. While these commutes stretched his family life, it also absolved him from participating in academic administrative tasks.

Elster is a member of the American Academy of Arts and Sciences, the Norwegian Academy of Science, and Academia Europaea and is a Corresponding Fellow of the British Academy and a member of Institut Jean Nicod. Prizes and awards have included Prize of the Norwegian Research Council for excellence in research, 1992; Prix Jean Nicod, Paris 1997; John von Neumann award, Budapest University, 2002; Fridtjof Nansen Prize for Excellence in Research, Norwegian Academy of Science 2006. His John Simon Guggenheim Fellowship was awarded for 1989. He has held visiting professorships at many universities and is an Honorary Doctor at many. Elster has facilitated the establishment and (co)directed centers at the University of Chicago (the Center for Ethics, Rationality and Society, and then the Center for the Study of Constitutionalism in Eastern Europe). He has edited several books series, given many named lectures and been on the Editorial boards of a range of journals.

Elster's scholarship could be described as having been broadly very consistent while pursuing a wide array of topics. The direction of this work has not been much planned or foreseen, but he makes attempts to retrospectively rework to conjoin later explorations with earlier material and to rework the earlier material to better fit with emerging conceptualizations. Often opportunities have arisen, which he has pursued while knowing little about the topic, but which has provided opportunities to construct broadly generalized knowledge. He has been assiduous in reaching back to the writings of historical predecessors and then reconstructing their work to elucidate those parts, which work well, while also signposting areas where their work is weak. Some of his work is not systematic, at least during earlier stages of development when there is a tendency to generate a list of elements and then to work through these while more systematic frameworks that might allow more tightly developed schema have yet to be developed.

There is only sparse extended commentary on Elster's work, although his work is cited often. There are two essay collections (Cuypers, 1995; Leschke, 2008); several essay-length reviews, a few autobiographical accounts (Elster, 1995a; n.d.), and various interviews (e.g., with Swedberg, 1990; Guénard and Landemore, 2008) and YouTube appearances.

Philosophy of Social Science

A continuing concern for Elster has been not just to generate social theories but also to reflect on the methodology of social science. Early campaigns denigrated the convoluted looseness of much received theoretical work, especially from various European theorists. However, Elster has also been concerned with the overly common resort to functionalist explaining, especially among US theorists. More generally, he has argued that useless theories create waste and harm – deflecting from the hard work necessary to create useful theories and at worst providing wrong understandings. He sees much of social science as falling into either hard or soft obscurantism: making unrealistic assumptions and relying on very limited research designs on the one hand and falling back onto merely descriptive work on the other.

He believes (and offers this as an explanatory account) that there is a deep-seated human need to perceive meaning and order in the universe, and an equally deep-seated aversion to admitting the existence of chance and accident, sound and fury. Ways in which to satisfy this need for order, shared too by social scientists, include: blaming the victim (to reduce any appearance of injustice), finding conspiracies, appealing to unconscious intentions, transforming into the goal of behavior what can only be a by-product of behavior – and appealing to functional explanations.

Elster regards the usual assumption that the goal of social science is to produce general, lawlike theories, that can yield not only explanations after the fact but also predictions before the fact, to be too ambitious since individuals and societies are just too complicated. However, social science does not need to fall back on being confined to narratives or descriptions. An intermediate strategy for the social sciences focuses on mechanisms: "causal patterns of a simple, general kind that occur in many different situations under unknown triggering conditions or with indeterminate consequences. Ex post, they can illuminate and explain many forms of behavior but they have little predictive power, since we cannot tell ex ante when one will be triggered" (Elster, 2007a: p. 36). There are two general kinds of mechanism: type A are triggered under unknown conditions, whereas by contrast, type-B mechanisms have indeterminate consequences. In a situation of type A, at the beginning of the causal sequence, one or other causal chain is triggered, but we do not know which one. For instance, the prey detects the predator and sometimes this leads to escape, but sometimes the prey ignores the predator. However, in a type-B situation, two (or more) chains are triggered with an indeterminate overall result (net effect) at the end of the evolving causal

sequence. For example, swimming training seems to activate two causal chains so that the net effect of training on risk is often indeterminate: in one, skill tends to depress the risk of misdemeanor; in the other, confidence tends to increase it. Sometimes, confidence in swimming skill increases more rapidly than the skill itself, although the trajectory can be the other way round.

Elster has also upheld the principle of 'methodological individualism' or the principle of 'the need for micro-foundations.' Since only actors can create social outcomes, Elster has resolutely clung to models, which are at base concerning individual activities.

Theoretical Overview

The conceptual architecture of Elster's work has been an extended and grounded version of RCT, in which he has added conceptual elements outside of the traditional framework (especially emotions and norms) but has also explicitly identified limitations of this approach. He sees RCT models as far too abstract and limited, needing to be limited by more realistic pictures. By ignoring the calculative burdens of actors in implementing fully rational actions, RCT theorists produce a science fiction.

Nevertheless, rationality is seen as a basic human conception, although always culturally specific in detail. So, a guiding hypothesis should always be to assume rationality as a first step in understanding what is happening in a situation. Rationality, emotions, and social norms are three main motivations, which can account for most of human behavior. They should be built into explanations, and be considered in that order. Social norms can override rationality, and emotions may override both. For example, when I take revenge, is it because I want to build up a reputation for vindictiveness, because I live in a society with strong norms of revenge, or because I lash out in passion without pausing to think about the consequences? When I help my neighbor, is it because I expect him to reciprocate (instrumental rationality), because it is expected by third parties (social norms) or out of spontaneous empathy (passion)?

The standard model of rational choice relates the four elements: actions (or action tendency), beliefs, desires (or preferences), and information. These relationships can be seen as either explanatory or optimizing. The rational actor optimizes by choosing the action that best realizes his/her preferences, given beliefs about what the options are and about the consequences of choosing one or the other of them. These beliefs are, in turn, inferred from the available evidence by the procedures that are most likely (in the long run, and on average) – to yield true beliefs. On the standard model belief formation is prior to preferences and prior to belief and optimal evidence is gathered (in relation to the preferences and to the expected costs/benefits of gathering more information). There is a feedback loop focused on information that reflects that expected benefits may be modified by the ongoing results of the search process. To these links, Elster adds emotions, of which – he argues – there are some two dozen, which can be well-established. Beliefs can generate emotions, which in turn have consequences for behavior.

There are wider structural limitations that effect behaviors. Some outcomes are indeterminist, there is also irrationality and uncertainty is widespread (i.e., you cannot give a probability for different outcomes). Moreover, some interventions are 'Brute': e.g., disasters. In particular, it is difficult to know numbers and preferences of future generations.

Over the last two or so decades, Elster has also explored similar issues at a more collective (or collective action) level, although he has yet to tie this material together or to relate it to his more individual-level conceptions. The vocabulary used in posing these issues is usually that of 'institutional design.' This is an alternative to the less action-orientated functionalist approach in that the organizational principles are entirely explicit and forward-looking, whereas in functionalism they are implicit and concerned with consequences. Institutional design principles provide the context of individual's situated actions and decisions and the key design idea is to remove obstacles to good decisions. Elster's work has explored three distinct areas: collective bargaining, scarce goods allocation, and finally retributive justice (including issues arising with political transitions).

Elster has not been able to satisfactorily develop his ideas on distributive justice, because of the difficulty of issues than need to be faced. This is in contrast to the idea of rationality, which is relatively unambiguous albeit fenced around with paradoxes and conundrums since in almost all situations almost everybody will agree on what it is rational to do, subject to the problem of indeterminacy. The notion of justice is far more complex since although there is widespread (but not universal) agreement on the importance of distributive principles such as need, rights, efficiency and equality, but ideas about the tradeoffs among these values are very vague. It is also hard to articulate ideas about how the good(s) are to be distributed according to these principles.

In sum, Elster might be labeled (although labels are not important to him) an 'Analytical realist' who sits in the middle of the interplay between the messiness of social reality and the abstract models necessary to explain it, between endeavoring to understand individual decision-making and the institutional contexts shaping this, and between the social science tasks of explanation and the wider tasks of bringing normative assessments to bear.

Effects

Elster's contributions to social science are spread across a wide range, which has given him an equally widespread audience but perhaps has meant that the cumulative effect has not been as great in any of the disciplines covered. Moreover, Elster has made little attempt to link his work with existing conceptual apparatuses of academic disciplines. Writing on Marx during periods when this was a heated topic was not always well-received. In his review essay, Daniel Little wrote a view with which other commentators would undoubtedly concur:

> Elster has made important contributions to several fields. The breadth and depth of his writings are striking in a time of high specialization; he is read and discussed by political scientists,

economists and philosophers. His work is difficult to summarize in a slogan, but it is generally informed by a broad and deep acquaintance with relevant literature in economics, political science, history, philosophy, and psychology.

1992

Conclusion

Elster's intellectual development has both partly negative and largely constructive trajectories. More negatively, his successive disillusionments have included concern over the lack of clarity and soundness of reasoning of many social theorists, including Marx and his predecessors, and then a more generalized loss of belief in the predictive power of the social sciences in general, and that of RCT, more specifically. However, Elster has very considerably extended the microfoundations of RCT by indicating the ways in which emotions and social norms supplement rationality and how individual decision-making is contextualized within institutions whose specific design features have particular consequences. Rather than searching for lawlike theory, Elster has sought mechanism-based explanations and has identified and exposed the workings of a number of these: 'reculer pour mieux sauter,' self-binding, states that are essentially by-products, adaptive preferences, the view of weakness of will as an intrapersonal collective action problem, the tendency for reason to go beyond its natural limits and degenerate into hyperrationality, the civilizing force of hypocrisy, and others. In building up the known repertoire of mechanisms, Elster has often acknowledged earlier theorists and his role has often been to collate a larger range of applications of these mechanisms and then to make explicit pertinent conceptual distinctions. Moreover, while eschewing social commentary and critique his work has often held out implications of considerable policy relevance. But all this is still a work in progress.

See also: Analytical Marxism; Analytical Sociology; Explanation: Conceptions in the Social Sciences; Field Theory; Functional Explanation: Philosophical Aspects; Individualism versus Collectivism: Philosophical Aspects; Intentionality and Rationality; Justice, Conflicts, and the Justice of Conflict Resolution; Marx, Karl (1818–83); Political Theory, History of; Rational Choice Explanation: Philosophical Aspects; Rational Choice and Organization Theory; Social Mechanism; Subjective Wellbeing, Psychology of.

Bibliography

Callebaut, Werner (Ed.), 1993. Taking the Naturalistic Turn, or How Real Philosophy of Science Is Done. University of Chicago Press.

Cohen, G.A., 1978. Karl Marx's Theory of History: A Defence. Princeton University Press, Princeton, NJ.

Cuypers, S. (Ed.), 1995. Indirecte Rede: over Rationaliteit en Irrationaliteit. Leuven/Amersfoort, Acco.

Elster, J., Herpin, N. (Eds.), 1992. Ethics of Medical Choice Poitiers. Actes Sud (Translated into English and Spanish).

Elster, J., Hylland, A. (Eds.), 1986. Foundations of Social Choice Theory. Cambridge University Press.

Elster, J., Le Pillouer, A. (Eds.), 2013. A quoi servent les elections. Presses Universitaires Françaises, Paris.

Elster, J., Moene, K.O. (Eds.), 1989. Alternatives to Capitalism. Cambridge University Press (Chinese and Spanish translations).

Elster, J., Roemer, J. (Eds.), 1991. Interpersonal Comparisons of Well-Being. Cambridge University Press.

Elster, J., Skog, O.-J. (Eds.), 1999/2007. Getting Hooked: Rationality and Addiction. Cambridge University Press.

Elster, J., Slagstad, R. (Eds.), 1988. Constitutionalism and Democracy. Cambridge University Press (Chinese and Spanish translations).

Elster, Jon, Offe, Claus, Preuss, Ulrich K., Boenker, Frank, Goetting, Ulrike, Rueb, Friedbert W., 1998. Institutional Design in Post-communist Societies: Rebuilding the Ship at Sea. Cambridge University Press, Cambridge, UK; New York, NY, USA.

Elster, J., et al. (Eds.), 2006. Understanding Choice, Explaining Behavior: Essays in Honor of Ole-jørgen Skog. Oslo Academic Press.

Elster, J., 1975. Leibniz et la Formation de l'Esprit Capitaliste. Aubier-Montaigne, Paris.

Elster, J., 1978. Logic and Society: Contradictions and Possible Worlds. Wiley, Chichester (Translated into German, Spanish and Chinese).

Elster, J., 1979. Ulysses and the Sirens: Studies in Rationality and Irrationality. Cambridge University Press (Translated into French (partially), German (partially), Italian, Spanish).

Elster, J., 1983a. Explaining Technical Change, a Case Study in the Philosophy of Science. Cambridge University Press (Translated into Spanish).

Elster, J., 1983b. Sour Grapes: Studies in the Subversion of Rationality. Cambridge University Press (Translated into French (partially), German (partially), Italian, Spanish, Slovenian, and Turkish).

Elster, J., 1985. Making Sense of Marx. Cambridge University Press (Translated into Turkish, French, Korean and Chinese).

Elster, J., 1986a. An Introduction to Karl Marx. Cambridge University Press (Translated into Norwegian, Portuguese and Spanish).

Elster, J. (Ed.), 1986b. The Multiple Self. Cambridge University Press (Translated into Italian).

Elster, J. (Ed.), 1986c. Rational Choice. Blackwell, Oxford.

Elster, J. (Ed.), 1986d. Karl Marx: A Reader. Cambridge University Press.

Elster, J., 1986e. Solomonic Judgments: Studies in the Limitations of Rationality. Cambridge University Press (Translated into Spanish).

Elster, J., 1986f. Nuts and Bolts for the Social Sciences. Cambridge University Press (Translated into Spanish, Italian, Greek, Croatian, Lithuanian, Portuguese, and Japanese).

Elster, J., 1989. The Cement of Society a Study of Social Order. Cambridge University Press (Translated into Chinese, Spanish, Italian and Czech).

Elster, J., 1992. Local Justice. Russell Sage Foundation (In Co-Publication with Cambridge University Press), New York (Translated into Spanish and Italian).

Elster, J., 1993. Political Psychology. Editions de Minuit. English Trans. (revised). (Translated into Spanish and Italian).

Elster, J., 1995a. Going to Chicago. In: Cuypers, S. (Ed.), Jon Elster Over Rationaliteit en Irrationaliteit. Acco, Leuven/Amersfoort.

Elster, J. (Ed.), 1995b. Local Justice in America. Russell Sage, New York.

Elster, J. (Ed.), 1996. The Round Table Talks in Eastern Europe. University of Chicago Press.

Elster, J. (Ed.), 1998. Deliberative Democracy. Cambridge University Press (Spanish translation).

Elster, J. (Ed.), 1999a. Addiction: Entries and Exits. Russell Sage, New York.

Elster, J., 1999b. Alchemies of the Mind: Rationality and the Emotions. Cambridge University Press (Chinese, Spanish and Portuguese translations. Partial French translation).

Elster, J., 1999c. Strong Feelings: Emotion, Addiction and Human Behavior. MIT Press (Translated into Spanish, Italian and Japanese).

Elster, J., 2000. Ulysses Unbound: Studies in Rationality, Pre-commitment, and Constraints. Cambridge University Press (Translated into Italian and Spanish).

Elster, J., 2004. Closing the Books: Transitional Justice in Historical Perspective. Cambridge University Press (Translated into Serbian, Norwegian, Spanish, Italian and German).

Elster, J. (Ed.), 2006. Retribution and Restitution in the Transition to Democracy. Cambridge University Press.

Elster, J., 2007a. Explaining Social Behavior, More Nuts and Bolts for the Social Sciences. Cambridge University Press (Translated into Spanish, Chinese, Russian, Arabic, Italian, Romanian and Turkish).

Elster, J., 2007b. Agir Contre Soi. Editions Odile Jacob, Paris (Translated into Italian).

Elster, J., 2009a. Le Désintéressement. Seuil, Paris (Translated into English and Spanish).

Elster, J., 2009b. The First Social Scientist. A Study of Alexis de Tocqueville. Cambridge University Press.
Elster, J., 2010. L'irrationalité. Seuil, Paris (Translated into Spanish).
Elster, J., 2013. Securities against Misrule: Juries, Assemblies, Elections. Cambridge University Press.
Elster, J. (Ed.). Publicity and Privacy in Votes and Debates. Cambridge University Press, forthcoming.
Elster, J, n.d. Paris round Trip, (in Norwegian).
Guénard, Florent, Landemore, Hélène, 2008. When the Lottery Is Fairer than Rational Choice. Interview with Jon Elster (text&video), laviedesidees.fr.
Hollis, Martin, January 1991. Why Elster is stuck and needs to recover his faith. London Review of Books 13.
Landemore, H., Elster, J. (Eds.), 2012. Collective Wisdom. Cambridge University Press.
Leschke, Martin (Ed.), 2008. Conzepte Der Gesellschaftstheorie #14: Jon Elster's Theorie Rationaler Bindungen (Jon Elster's Theory of Rational Self-binding). Mohr Kurier, Germany.
Little, Daniel, 1992. Jon Elster. In: Samuels, Warren (Ed.), New Horizons in Economic Thought: Appraisals of Leading Economists. Edward Elgar Publishing.
Loewenstein, G., Elster, J. (Eds.), 1992. Choice over Time. Russell Sage Foundation, New York.
Novak, Stéphanie, Elster, Jon (Eds.), 2014. Majority Decisions: Principles and Practice. Cambridge University Press.
Swedberg, Richard, (Ed.),1990. John Elster. Economics and Sociology: Redefining Their Boundaries: Conversations with Economists and Sociologists. Princeton University Press, Princeton, NJ, pp. 233–248.
Williams, Melissa, Elster, Jon, Nagy, Rosemary (Eds.), 2012. Transitional Justice. New York University Press.

Embodied Social Cognition

Caterina Suitner, University of Padova, Padova, Italy
Mauro Giacomantonio, University of Rome 'Sapienza,' Rome, Italy
Anne Maass, University of Padova, Padova, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Embodied social cognition refers to the relation between bodily states and higher cognitive processes regulating social behavior. It may operate bottom-up (with sensory-motor system affecting cognition) or top-down (with higher cognitive processing affecting motor responses and bodily feelings). After providing specific examples for both types of embodiment, the main theoretical approaches are presented, together with moderator variables that facilitate or constrain embodied social cognition.

Definition

Research examining the relationship between bodily states and higher cognitive processes that regulate our social behavior (such as making decisions or understanding the intentions of others) can be included under the umbrella of embodied social cognition.

The main goal and principal strength of this research domain and of the related embodied theory of social cognition resides in the attempt to integrate and link different aspects and levels of human experience, from perception to social processes, from motor system to knowledge organization (for a recent overview on the topic see Meier et al., 2012).

The embodied cognition approach stresses the situated nature of cognitive processes, which are influenced by bodily states, by the physical properties of the target object, and by the physical environment in which they are carried out, even when such features are not directly linked to the task. To cite an example that has frequently been interpreted as evidence of embodied cognition, we are more likely to evaluate a humorous target stimulus as funny when we hold a pen horizontally with our teeth (therefore activating the muscles typically involved in smiling) compared to when we block our facial muscles by holding a pen with our lips. What makes this finding interesting is the fact that there is no intrinsic logical reason why facial muscle activity should make visually presented stimuli any more enjoyable.

The idea that bodily states or actions contribute to cognitive processing has received growing interest in social cognition, especially in recent years. As a consequence, cognitive processes that are crucial for social relations (ranging from attitude formation to person perception, up to the bodily grounding of social stereotypes) have been increasingly investigated in relation to the motor system. In doing so, social scientists have primarily focused on two complementary processes, namely (1) the environmental and bodily influences on social cognitive processes (from body to mind) and (2) the embodied display of grounded social cognition (from mind to body).

The first line of research deals with the role of environmental features and congruent bodily states on overtly unrelated higher cognitive tasks and may be referred to as bottom-up embodiment. A multitude of research has provided evidence for the situated nature of cognition by showing that sensory/motor inputs are included in, and exert a direct influence on, the information process. The most important characteristic of this type of studies is that the relation between the bodily state and the higher cognitive process is not directly relevant for the task, as illustrated by the above shift in evaluation of humorous stimuli as a function of facial muscle activity.

The second line of studies takes into consideration the effects of higher order cognitive process on bodily states. This line of studies can be classified as top-down embodiment, in the sense that the body is 'used' by cognition to perform its tasks. According to Barsalou (1999), this type of embodied phenomenon can be explained by a simulation mechanism, namely the activation of a relevant motor modality. For other scholars this is simply a motor resonance of the activation of a semantic network. For instance, while smiling we activate the concept of enjoyment and this semantic activation is responsible for the pen in the mouth effect previously mentioned (see Section Theoretical Perspective).

In the following section, we will review empirical research pertaining to embodiment and its social implications by summarizing few illustrative examples of recent relevant work. We will first examine studies in which sensory-motor system affects cognition, with sensory-motor system being all the sensory processes involved in perception (e.g., seeing, smelling, tasting) as well as motor behavior (e.g., flexing an arm, running). Subsequently, the influence of higher cognitive processing on motor responses and bodily feelings will be considered. Together the findings reported below suggest that social cognition is deeply grounded in the environment in which it takes place, and that social processes and bodily states are linked closely and in a bidirectional fashion. We will then discuss different theoretical perspectives that have been advanced to explain embodied phenomena and the boundary conditions of embodiment.

From Body to Mind: How Sensory-Motor Activation Relates to Higher Cognitive Processes

One of the primary concerns of research in the embodiment tradition has been to demonstrate that environmental features and congruent bodily states contribute to logically unrelated higher cognitive functioning. In other words, individuals use

information embedded in the environment or in their own body to make judgments and evaluations concerning a wide array of domains, even when such information is not directly enlightening about the target to be evaluated. Obviously, as we will see later, this process also applies to and has important implications for social phenomena.

What are the sources of information individuals rely on when forming embodied judgments? As specified by the term embodiment, any bodily state of the person who is performing an evaluative process may in principle influence his or her cognitive processes. Thus, whether the body is leaning to left or right or whether an arm is flexing or extending might be elements capable to affect unrelated judgments, such as political opinions or liking of different social groups. In this case, the motor system of the cognizer is responsible for embodied phenomena in a strict sense. A different, yet related source resides in the physical features characterizing the context in which the cognitive process is taking place, such as – for example – the smell or the temperature of a room. These perceptive characteristics may also pertain to objects that are present in the environment (e.g., a cup, or a clipboard) or even a specific feature of the target itself (e.g., size or brightness). For instance, the height of a person may affect the power we ascribe to him or her. This perceptual source of information is not directly connected to bodily states as compared to the example above. Nevertheless, contextual characteristics might affect bodily feelings and perceptions, thereby being indirectly responsible for embodied phenomena. In this second case, the body functions is a means through which environmental variables affect apparently unrelated higher order cognitive processes. In the following section, we will review relevant work according to the type of source outlined above.

Body as Source

Robust empirical evidence supports the idea that body movements or states (postures) influence social judgments, perceptions, and relationships. The work of Chandler and Schwarz (2009) constitutes a paradigmatic example of this line of research. The authors wanted to test whether hand movements could affect judgments concerning a target person. By means of a cover story concerning the effect of motor activation on cognitive activities, participants were asked to extend either their middle or index finger while reading about a target person described as ambiguously hostile. As hypothesized, the target was seen as more hostile when the description was read while the middle, rather than index finger, was extended. Likewise, a subsequent study showed that a more favorable impression was created when thumb rather than index finger was extended. Therefore, different finger movements affected impression formation of an ambiguous person. Interestingly, the association of the finger movement with a positive or hostile impression is entirely arbitrary and culturally dependent, thus indicating that also bodily states, actions or feelings that are not *hardwired* and connected with evolutionary needs (e.g., cold, carrying a heavy weight, extending vs. flexing an arm) can affect higher order cognition. In other words, also actions and bodily states that we arbitrarily associate with a specific meaning (i.e., extending the middle finger to communicate contempt) and that vary across time, location, and culture can shape our reasoning and social judgments.

The role played by bodily influences on impression formation is not confined to the positive vs. negative dimension, as illustrated by the fact that stereotypes can be promoted by specific information rooted in the body. More specifically, Mussweiler (2006) reasoned that stereotypic movements may trigger the corresponding stereotype, which, in turn, is used to judge a target verbally described with information that is unrelated to body weight. Consistently with the reasoning, the author made participants move similarly to overweight people. This goal was accomplished by asking participants to wear a life west and weights around ankles and wrists and to engage in a series of movements. In a second allegedly unrelated task, a target was judged using both stereotypic (e.g., friendly, unhealthy, lazy) and nonstereotypical dimensions (e.g., musical, creative, forgetful) concerning overweight individuals. Results showed that ratings were more stereotypical when participants' movements were hampered by the vest and the weights.

Besides shifting stereotypic impressions of ambiguously described others, can bodily feedback also affect relatively stable and frequently used schemata? According to Schubert and Koole (2009), it is possible that certain bodily states can influence one's own social self-concept. By signaling the intention or proneness to undertake a certain kind of action, bodily states favor a variation of the self-concept in line with the nature of the action to be taken. Specifically, if a person performs a motor action (such as making a fist), which signals readiness to use physical force against someone else, people should personally experience an increased sense of power. Consistently, the authors found that individuals making a fist rated themselves as more powerful and assertive than participants in a control group both on implicit and explicit measures. Interestingly, these findings emerged only for male participants. A possible explanation is that males more than females associate physical force to power over others; as a consequence, only males' power concept was affected by gestures strictly associated with the use of force. Together, the above examples demonstrate that motor actions or states affect not only how we perceive others, but also personal self-construal.

Contextual Characteristics as Source

In the previous paragraph, judgments relied on the body as source of information. What changed between conditions was mainly a position, a gesture or a movement performed by participants with their own motor system. Conceptually similar findings have been obtained with sources of information that are external to the body, thereby involving the sensory system. As we mentioned above, it is possible to distinguish at least three instances of contextual sources of information that trigger embodied processes.

The first contextual source resides in the general characteristics of the surrounding environment. Thus, the extent to which a room is chaotic or clean might be relevant in determining how we judge the behavior of a person that is not even present in the room or the proclivity to socialize and trust others. Ijzerman and Semin (2009), for example, building on the idea that the abstract concept of social closeness is based on the concrete experience entailing warm sensations, such as the hug, found that the warmth of a room can influence the

perception of social relationships. More specifically, participants in one study were placed in a warm (22–24 °C) or cold room (16–18 °C). Subsequently, relational focus of participants was measured in different ways. Results showed that those in the warm room, as compared to those in the cold room, expressed a stronger relational concern.

The second sensory source of information pertains to the physical features of the object with which the individual is interacting. Surprisingly, physical characteristics of objects we encounter in our everyday activities can shape the way we evaluate unrelated individuals. A study supporting this claim was conducted by Ackerman et al. (2010). The authors asked participants to solve a simple puzzle whose pieces were, or were not, covered with rough sandpaper. Subsequently, an ambiguous social interaction between two strangers was evaluated. As a consequence of this subtle manipulation, participants in the 'rough' conditions evaluated the interaction as less friendly than those in the 'smooth' condition. In a related study, a different object and a different physical characteristic were examined, further reinforcing the idea that embodiment can stem from the artifacts we handle in a mindless way. Ackerman and colleagues tested whether the weight of a clipboard would affect the perceived importance of the social issues presented on that clipboard. It was found that a heavy clipboard (about 1.5 kg) increased respondent's willingness to spend government funding on a variety of social issues as compared to a light clipboard (less than 1/2 kg).

Finally, an unrelated physical feature of the evaluation target may be the source of the information used for the evaluation, as shown in a study by Meier et al. (2004) in which participants had to categorize words (*kiss*, *justice*, *danger*, etc.) as either positive or negative. People found it easier to identify good concepts (e.g., *justice*) as positive when presented in bright (vs. dark) colors, and vice versa for negative concepts (e.g., *danger*), suggesting that responses are facilitated when there is a match between the valence of the concept and the brightness of the stimulus word. The authors explained this effect by noting that, in Western cultures, darkness is metaphorically associated with evil and death, whereas light is associated with good.

Together, the empirical evidence reviewed so far supports the situated nature of cognition. Sensory/motor inputs originating from the body or from the situation are included in the information processing and have a direct influence on it. What is important in this specific line of inquiry is that there is no logical relation between the bodily state and the task. For instance, the height of a target is obviously a key feature to estimate the target's weight, and its influence would not provide any meaningful insight in the understanding of embodied cognition. Differently, the fact that height is a critical factor for estimating an apparently irrelevant characteristic of the target such as social power is taken as evidence for the situated nature of cognition.

From Mind to Body: Higher Cognitive Processes Automatically Activate Motor Responses

In the previous section, we reviewed research examining how context and body affect higher order cognition. This path is particularly important, especially in social life, because individuals generally make decisions and take actions in a situated way that is in a specific context filled with social and perceptual stimuli that affect, in an indirect and surprising fashion, the way they think and judge others.

However, the research reviewed above does not speak to whether this perceptual conceptual path can be reversed. In other words, can the content of our higher cognitive processes affect the way we feel and perceived the environment or even our bodily states (such as our body temperature)? A great deal of research indicates that often our postures or our bodily feelings are a consequence of how we reason and think.

Several examples can be driven from the literature of nonverbal communication, which widely shows that the posture, bodily movements and facial expressions, reveals the psychological state of the person, such as his or her emotions. For example, happiness is associated with the upright position and – of course – with a smiley face. Similar results can be obtained when a feeling is made salient. For example, Zhong and Liljenquist (2006) asked participants to recall unethical (vs. ethical) episodes of their life. When unethical episodes were recalled that presumably threatened the participants' sense of moral purity, they felt the urge for cleansing. This was translated into increased preference for cleansing products over other products and increased likelihood to accept an antiseptic wipe rather than a pen as a free gift.

Previously, we discussed how sensorimotor experience of warmth can shape unrelated social judgments. Importantly, this connection can work in the opposite direction and social experiences can shape our body feelings or perceptions of the surrounding context. The recall of a situation in which a person was socially excluded led participants to perceive a room temperature about 3 °C colder than after recalling a situation in which they were socially included (Zhong and Leonardelli, 2008). In a related fashion, DeWall and Baumeister (2006) drew a parallel between the pain related to social exclusion and to physical injury. The authors argued that, as a consequence of social exclusion, individuals should be less responsive to physical pain. Indeed, in order to cope with exclusion, the capacity of the body pain system to detect both emotional and physical pain is altered and pain thresholds are raised. Empirical tests fully supported this reasoning. In one experiment, for example, a bogus feedback on a personality questionnaire indicating that participants were likely to be alone in their future life increased physical pain tolerance and pain threshold.

Although the mind to body link is well established, more research is needed to fully understand the underlying processes. Recently, some theoretical and empirical efforts have been made in this direction. Neuroscience research, for example, has shown that the mere act of reading words expressing physical actions such as *lick* or *kick* activated neurons of the motor cortex that are also activated when the target action is actually performed (Hauk et al., 2004).

The Theoretical Perspective

So far we have presented embodied social cognition as a line of research that empirically investigates social cognition in relation to the motor system. The practical implications of

these findings are evident. For example, understanding how embodiment works offers practical indications on how to promote specific cognitive outcomes by altering the environmental context. For instance, varying the weight of an object may be a subtle strategy for increasing the perceived importance of such object and changing the room temperature may be a tool for facilitating conflict resolution. All the findings presented above can be easily used for persuasive purposes in numerous applied contexts.

Whereas the practical importance of this line of research is generally accepted, there is considerable debate about the theoretical underpinnings of such effects. Embodied theories are often juxtaposed to the so-called disembodied theories, according to which the motor system is activated as a byproduct of the activation of a superordinate system (often equated to the semantic network). According to this view, the activation of the motor system is simply a side effect of the semantic activation of related concepts. Among the theories that explicitly hypothesize a strong relation between the motor system and higher cognitive processes, we can cite *in primis* the broad theory of Grounded Cognition (see below for a brief description) and the Ecological Approach (Gibson, 1979) according to which environmental or physical features (such as the position of the handle of a target object) affect the elaboration process of the target stimuli even if these features are not directly relevant to the task at hand. For example, if the picture of a cup is oriented with the handle on the right (rather than on the left) side of the visual field, participants more quickly recognize it when they respond with their right hand, showing a facilitating effect of the spatial features of the stimulus. This has been interpreted as functional to action planning, in this case to grasping that is facilitated when handle and hand are on the same side (Phillips and Ward, 2002). Affordances refers to the relation between the environment and the person acting in it, and specifically to the environmental or object characteristics that are relevant for the interaction (e.g., the handle is an affordance offered to people by a cup). Such characteristics of the object are therefore included even in tasks (such as the categorization of the color of the cup) that do not require an action in which the affordance (in the example the handle) is relevant (Phillips and Ward, 2002). Another frequently cited approach is Conceptual Metaphor Theory (Lakoff and Johnson, 1999), according to which the use of conceptual metaphors based on physical features (e.g., the vertical metaphor to indicate power, see 'upper status') creates a motor resonance that explains the relation between cognition and bodily states. One way to reconcile the theoretical definitions of embodiment is to distinguish between a hard and a soft version of the embodied cognition perspective.

Soft vs Hard Approach of the Embodiment

The *hard view* on embodied cognition is primarily based on Barsalou's (1999) Grounded Cognition Theory, according to which any cognitive process or representation is deeply embedded in a specific bodily modality. The grounded cognition hypothesis suggests that sensory-motor and intellectual activities do not represent two distinct systems, rather they are thought to be the constituents of embodied cognition. Put simply, cognition *is* a sensory-/motor-based process. According to this view, the architecture of the human mind is made of direct connections among modality-specific representations. For example, the perception of size is the modality that is connected to the concept of social power, therefore observing a tall person directly promotes the attribution of power. In off-line situations, the bodily state precedes the task, and a simulation mechanism reenacts relevant physical modalities to fulfill thinking. For example, when we feel powerful, we activate the sensory state that is relevant to the concept of power, namely straightening our back and thereby increasing our body size. In the example, the self-evaluation consists in the motor simulation. Accordingly, we activate our own mirror and motor system when understanding actions performed by another person. The mirror neuron system provides a correspondence in the neural activation during the actual execution of an action and the observation of other person's action and this system is supposed to allow several key social processes, such as understanding other persons' intentions and behaviors, as well as self-awareness and learning through imitation. According to the hard view, the sensory-motor system is involved in any type of information processing. This perspective also implies that cognition is always grounded in the body and that the motor system is automatically implied during any cognitive activity. The *chain of causality* is bottom-up, from body to cognition. The hard view has been criticized by scholars arguing that the empirical findings that have been interpreted as evidence for embodiment can be easily interpreted from a disembodied theoretical perspective.

Whereas the hard version implies the lack of an a-modal system (being it semantic, symbolic, or statistical) that orchestrates such representations, the *soft view* grants the necessity of an a-modal superordinate system that structures and organizes modal and embodied representations. The role of bodily states is therefore a crucial aspect of cognition, but cognition is not equated to the motor/sensory system. Thus, whereas the hard view conceives cognitive processes entirely as a product of bodily states and of their simulations, the softer approach views cognition as bidirectionally related to bodily states, but does not embrace the idea that cognition is totally and exclusively rooted in the body. Rather motor and sensory states can be seen as instantiations of cognition. In this softer view of embodiment, cognition is made of symbolic, abstract and a-modal representations fed by bodily modalities that represent at the same time (1) the ground in which cognition is embedded and (2) the means through which cognition expresses the output of its processes. We perceive through our bodies, we act through our bodies, we rely on perceptions of our bodily states. Both at input and output level in encoding and decoding processes, the modalities of our body constrain the cognitive processes. This soft approach allows for both modality-specific representations of concepts and a superordinate system managing such modal representations.

Moderators

After an initial emphasis on demonstrating the existence of embodied phenomena, research on embodied cognition is now turning to the analysis of related underlying processes. As

in many areas of psychological research, this attention on processes is reflected not only in direct assessment of mediating variables but also in the analysis of moderators that define the boundary conditions under which embodied phenomena are likely to emerge. Although currently in an early stage of development, this newly developing line of research deserves attention for two main reasons. First, it represents a theoretical perspective that is rapidly gaining ground, namely that embodied cognition is a process that is activated under specific conditions and that can reach different degrees of activation depending on the situation and on needs. The second reason is that such preliminary findings are likely to have a strong impact in the understanding of embodied cognition. We will briefly describe some of the major moderators of embodiment that have been individuated at this stage.

Familiarity

Familiarity has been proposed as a moderator in the evaluation of importance based on weight. Chandler et al. (2012) hypothesize that the embodied process requires the presence of prior knowledge that serves as an anchor. For instance, greater weight of a book will make a known book appear more important, but weight will not be influential when the book is unknown. Similarly, the weight of coins will affect the estimation of currency value, but only to the extent that participants are already familiar with the currency.

Consistency

Another moderator is the consistency between the motor activity on one side and the information to be processed or a primed psychological state on the other. For example, arm flexion (extension) movements facilitate the processing of and memory for positive (vs. negative) information, because flexing is an approach behavior that is more congruent with positive elements, in contrast with the complementary association between avoidance and negative elements (Forster and Strack, 1998). Similarly, the memory of presented information is improved when the affective valence of the motor activity (e.g., extending vs. flexing the arm, or smiling vs. frowning) matches the affect of a primed concepts (e.g., positive vs. negative words, Centerbar et al., 2008). This match supports the idea that online motor activity interacts with previous knowledge and suggests that it is subject to selection biases, similarly to any other type of information.

Modality match

In addition, embodied phenomena are stronger when the modality of the action involved during the cognitive process matches the modality through which the embodied concept has been acquired. Testing the role of physical washing in cleaning the guilt experienced after an immoral act, Lee and Schwarz (2010) made their participants lie using their mouth (lying via voice mail) or using their hands (lying via e-mail) and asked participants to rate the desirability of a mouthwash and of a hand sanitizer. After lying by e-mail participants rated hand sanitizer more positively, but after lying using their voice participants rated the mouthwasher as more desirable. Participants seemed to search for a specific behavior able to clean their immoral act, confirming that the bodily modality of experience is not a-modal or a-specific.

Construal Level

A final, somewhat controversial moderator is construal level. Construal level depends on the psychological distance of the object or events such that distal objects are represented with higher level of construal. Higher level of construals correspond to abstract, decontextualized, and schematic representations of objects or events. In contrast, lower level of construals correspond to concrete, contextualized and detailed representations. According to Maglio and Trope (2012), embodied phenomena are more likely to occur at low levels of construal, namely when people elaborate information attending to peculiar and local characteristics. Carrying a heavy backpack affected visual length estimation only when participants were primed to attend to local characteristics. To the contrary, Suitner and Giacomantonio (2012) suggest that embodiment is more likely to emerge under low construal only when the embodied phenomenon is bodily driven (such as the body to mind type of embodiment discussed earlier). The authors indeed investigated a top-down (or mind to body) phenomenon, i.e., the Spatial Agency Bias,which is a cognitive horizontal asymmetry that is related to writing direction. Left-to-right readers usually prefer, visualize, and represent left-to-right acting targets. The agency of a target can be physically defined (e.g., moving targets) or socially defined (e.g., high status targets). The Spatial Agency Bias was stronger under high construal level, namely when the information is processed in a global fashion, without attending to contextual details. Specifically, when primed with high construal level, participants drawings were more aligned to their habitual writing direction, namely they placed more often the agentic target spatially before the receiving target in representations of dyadic interactions. Additional research is needed to clarify the role of this moderator.

Together, these new lines of research suggest that embodied cognition is less general and pervasive than originally thought. Rather than being an undifferentiated, universal mechanism, embodied cognition seems to be constrained by a number of variables including prior learning history, coherence of modality, and the perceiver's current mind-set.

See also: Heuristics in Social Cognition; Implicit Social Cognition; Social Cognition; Social Psychology.

Bibliography

Ackerman, J.M., Nocera, C.C., Bargh, J.A., 2010. Incidental haptic sensations influence social judgments and decisions. Science 5986 (328), 1712–1715. http://dx.doi.org/10.1126/science.1189993.

Barsalou, L.W., 1999. Perceptual symbol systems. Behavioral and Brain Sciences 22, 577–660.

Centerbar, D.B., Schnall, S., Clore, J.L., Garvin, E.D., 2008. Affective incoherence: when affective concepts and embodied reactions clash. Journal of Personality and Social Psychology 94 (4), 560–578.

Chandler, J., Schwarz, N., 2009. How extending your middle finger affects your perception of others: learned movements influence concept accessibility. Journal of Experimental Social Psychology 45, 123–128.

Chandler, J., Reinhard, D., Schwartz, N., 2012. To judge a book by its weight you need to know its content: knowledge moderates the use of embodied cues. Journal of Experimental Social Psychology 48 (4), 948–952.

DeWall, C.N., Baumeister, R.F., 2006. Alone but feeling no pain: effects of social exclusion on physical pain tolerance and pain threshold, affective forecasting, and interpersonal empathy. Journal of Personality and Social Psychology 91, 1–15.

Forster, J., Strack, F., 1998. Motor actions in retrieval of valenced information: II. Boundary conditions for motor congruency effects. Perceptual and Motor Skills 86, 1423–1426.

Gibson, J., 1979. The Ecological Approach to Visual Perception. Houghton Mifflin Co, Boston.

Hauk, O., Johnsrude, I., Pulvermuller, F., 2004. Somatotopic representation of action words in human motor and premotor cortex. Neuron 41, 301–307.

Ijzerman, H., Semin, G., 2009. The thermometer of social relations. Mapping social proximity on temperature. Psychological Science 20 (10), 1214–1220.

Lakoff, G., Johnson, M., 1999. Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. Basic Books, New York.

Lee, S.W.S., Schwarz, N., 2010. Dirty hands and dirty mouths: embodiment of the moral-purity metaphor is specific to the motor modality involved in moral transgression. Psychological Science 21, 1423–1425.

Maglio, S.J., Trope, Y., 2012. Disembodiment: abstract construal attenuates the influence of contextual bodily state in judgment. Journal of Experimental Psychology: General 141 (2), 211–216. http://dx.doi.org/10.1037/a0024520.

Meier, B.P., Robinson, M.D., Clore, G.L., 2004. Why good guys wear white. Automatic inferences about stimulus valence based on brightness. Psychological Science 15 (2), 82–87.

Meier, B.P., Schnall, S., Schwarz, N., Bargh, J.A., 2012. Embodiment in social psychology. Topics in Cognitive Science 4, 705–716. http://dx.doi.org/10.1111/j.1756-8765.2012.01212.x.

Mussweiler, T., 2006. Doing is for thinking! Stereotype activation by stereotypic movements. Psychological Science 17, 17–21.

Phillips, J.C., Ward, R., 2002. S–R correspondence effects of irrelevant visual affordance: time course and specificity of response activation. Visual Cognition 9, 540–558.

Schubert, T.W., Koole, S.L., 2009. The embodied self: making a fist enhances men's power-related self-conceptions. Journal of Experimental Social Psychology 45 (4), 828–834.

Suitner, C., Giacomantonio, M., 2012. Seeing the forest from left to right: how construal level affects the spatial agency bias. Social Psychological and Personality Science 3, 180–185. http://dx.doi.org/10.1177/1948550611414591.

Zhong, C.B., Leonardelli, G.J., 2008. Cold and lonely: does social exclusion feel literally cold? Psychological Science 19, 838–842.

Zhong, C.B., Liljenquist, K., 2006. Washing away your sins: threatened morality and physical cleansing. Science 313, 1451–1452.

Embodiment and Culture

Gabe Ignatow, University of North Texas, Denton, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Influenced by developments in cognitive neuroscience and by internal disciplinary debates, social scientists are rethinking what culture is, how it is learned, and how it ought to be studied. Rather than imagining culture as systems of abstract symbols contained within the minds of individuals, scholars are increasingly analyzing culture in terms of embodied practices (including linguistic and cognitive practices) learned as techniques and ingrained as habits. Some of the major theoretical and methodological developments that have attended this conceptual shift are reviewed here.

Introduction

Research programs in the social and behavioral sciences are often categorized in terms of several key dimensions. These include the *scale of analysis* of research, from microlevel studies of the mind and brain to macrolevel research on global economic systems and social trends. A second dimension is *universalism versus particularism*: while some (*nomothetic*) research seeks to identify universal principles and mechanisms underlying behavior, more culturally oriented and *idiographic* research explores the heterogeneity of human beliefs and practices over historical time and across social groups. Other categorizations include *critical versus applied* research, and research employing *qualitative versus quantitative* methods. Another, somewhat less frequently referenced dimension useful for categorizing social research programs is their *degree of embodiment*. This refers to the extent to which social and behavioral research takes into account bodily dimensions of, and influences on, social behavior, or else treats the body, and research on the body (generally from cognitive neuroscience, biology, and psychology) as irrelevant or orthogonal to social research. The embodiment dimension might be assumed to be equivalent to the universalism–particularism dimension, but this is not the case, as there has been a substantial amount of both nomothetic and more culturally oriented and idiographic social science research that takes into account bodily phenomena. In this article, I review recent developments in the latter of these categories of research: studies of cultural phenomena such as language, symbols, rituals, ideologies, and institutions that take into account, in one way or another, bodily processes.

Social scientists are divided in terms of their interest in results from fields such as cognitive neuroscience, biology, and psychology. Many sociologists and anthropologists take their cue from the anthropologist Clifford Geertz in treating culture as an emergent macroorder social fact irreducible to psychological and biological processes (Geertz, 1973: pp. 92–93). According to Geertz, though humans' capacity for culture is a product of natural selection, culture is so extraordinarily malleable and adaptive that it is effectively functionally autonomous from biological processes. For Geertz culture is a self-contained universe of symbols and social meanings, and as such contemporary research from biology, neuroscience, cognitive science, and psychology is assumed to be of little relevance to cultural analysis. Research from these fields might help social scientists to understand something about the raw materials from which cultures are constructed, but Geertz's position implies that these fields cannot produce insights into the group- and individual-level processes that generate, reproduce, and transform culture's varied forms.

Today, the growing ranks of social scientists who are expropriating insights from cognitive neuroscience, biology, and psychology (e.g., Cerulo, 2002) have put Geertzian arguments for the strict separation of cultural social science from biology and cognitive neuroscience on the defensive. Within sociology in particular, a touchstone for post-Geertzian cultural analysts (who are empowered by sociology's disciplinary position as an 'irremediably interstitial' field (Abbott, 2001: p. 6) that regularly exchanges ideas with many other fields of inquiry) is Pierre Bourdieu's treatment of the ancient idea of *habitus* (Wacquant, 2004; Lizardo, 2013). A shorthand term for these scholars, whose work has intellectual roots in the French tradition of psychologically informed sociology of Durkheim, Mauss, and Bourdieu, is 'practice theorists.'

'Strong' practice theory in sociology (Lizardo and Strand, 2010) conceptualizes culture in terms of socially learned, intertwined bodily, cognitive, and linguistic *skills* rather than in terms of units or systems of abstract knowledge (Bloch, 1986). More precisely, such 'strong' practice theorists are part of a 'third phase' of Bourdieu's reception in the United States (Lizardo, 2012) in which younger scholars are drawing on Bourdieu's practice theory, and especially his concept of *habitus*, while emphasizing experiential, ethnographic methodologies. Lizardo and Strand (2010) refer to this movement as both 'strong practice theory' and the 'French branch' of practice theory. Such practice theorists have drawn on Bourdieu's theoretical work on practice (Bourdieu and Wacquant, 1992), as well as on research from cognitive neuroscience, biology, and psychology, to challenge widely used dialectical models of cultural internalization and externalization. Instead of intuitively appealing 'container models' in which cultural 'software' is seen as stored within somatic 'hardware,' practice theorists rely (explicitly or implicitly) on anti-Cartesian *procedural models* of learning (Lizardo, 2012).

In container models learning is assumed to operate via a process of *transfer* (Lizardo, 2012) or *transduction* (Barsalou, 1999) of sensory experience into disembodied mental information. Sensory experience is transferred/transduced into

cultural software, which is thought to be stored in long-term memory much as a hard drive stores data: in the form of a nonsensory language akin to a binary computer language that is qualitatively different from the sensory experiences that are encoded in it. In contrast, procedural models of learning emerged well after the cognitive revolution of the 1960s–70s as part of a transdisciplinary shift away from epistemological models that treated knowledge in terms of abstract information, and toward models that conceptualize knowledge in terms of embodied skills and abilities. This 'bodily turn' or 'somatic revolution' has taken different forms across disciplines, but it has been pronounced in anthropology, linguistics, philosophy, and sociology. Research within this bodily turn shares a view of learning as situated, embodied enskillment, as *techne* rather than only *episteme*. But for social scientists, possible implications for cultural analysis of this broad shift from theories of culture as *episteme* to theories of culture as *techne* and *habitus*, and from container to procedural models of learning, have yet to be systematically considered. Several programmatic scholarly movements that are reflective of these shifts have emerged in recent years, so in the following Section Theoretical Strategies review some of the recently developed theoretical and also methodological strategies for cultural analysis that are compatible with the turn to embodied epistemologies. I do so with the understanding that in cultural social science as in all social science, theory and method are often interdependent: theories focus researchers' attention on empirical phenomena that are more amenable to analysis with some methodological strategies than with others, while methods contain within themselves metatheoretical assumptions about how social phenomena can be observed and measured. Nevertheless, for heuristic purposes my review of recent cultural research includes separate sections for theoretical and methodological strategies, and I emphasize points of compatibility, potential compatibility, and incompatibility between these strategies.

Theoretical Strategies

How do we learn culture? How is action shaped by culture? And how should social scientists' theories of culture react to recent research on learning, perception, and memory from cognitive neuroscience, biology, and psychology? Many culturally oriented social scientists (including social scientists influenced by Bourdieuian practice theory) simply ignore findings from these fields, while others explicitly reject such findings because of concerns over disciplinary imperialism, theoretical reductionism, and genetic determinism. On the other hand, a steadily increasing number of cultural researchers are choosing to selectively integrate insights from cognitive neuroscience into their work, balancing the potential for reductionism against their desire to ground cultural research on empirically defensible theoretical microfoundations. If culture influences behavior, they contend, then we ought to know something about how culture is learned and remembered in the brain and elsewhere in the body, and how it thereby informs decisions and actions. There are today several contending sets of analytic strategies – that is, sets of theoretical categories and related research strategies – that are informed by ideas from cognitive neuroscience and biology, and that have clear links to sociological practice theory. Three of the most influential such strategies are: (1) distinguishing between 'cold' and 'hot' cognition; (2) distinguishing between deliberative and automatic forms of cognition; and (3) holistic theoretical strategies that seek to identify 'physio-psycho-sociological' (Wacquant, 2004: p. 17) sequences of empirical phenomena.

'Cold' and 'Hot' Cognition

Early proponents of integrating the social and cognitive sciences promoted the importation into social science research of cognitive scientists' categories of 'different kinds of cognition' (DiMaggio, 1997), in particular the categories of 'hot' (emotionally saturated) vs 'cold' (affect-free) cognition. Cerulo (1995, 2010) has argued that this distinction is useful because it opens new avenues for social scientists to study the social structures and contexts that activate hot vs cold cognitive processes, rather than treating those processes as stable traits of individuals. Cerulo finds evidence of sociologically interesting cognition–situation interactions in studies of social movements (e.g., Goodwin et al., 2001) and decision making, and in her own study of national flag designs (Cerulo, 1995).

'Discursive' and 'Practical' Consciousness

A second binary set of types of cognition imported from cognitive science into social science is a distinction between 'deliberative' and 'automatic' forms of cognition. Deliberative cognition (in sociological parlance 'discursive consciousness') is a cognitive system that is slow and deliberate that processes consciously stated values and beliefs. Automatic cognition (or 'practical consciousness') is a system of intuitions that is fast, automatic, and largely unconscious that allows people to operate in ordinary situations without having to constantly reevaluate their goals and strategies (DiMaggio, 2002). Vaisey has been perhaps the foremost proponent of the applicability of 'dual-process' (deliberative vs automatic) models of cognition to social science research. He has demonstrated the applicability of dual-process theory to sociological research in an analysis of survey data on youths' moral values (Vaisey, 2009), while Srivastava and Banaji (2011) have used the Implicit Association Test developed by cognitive psychologists to tease apart deliberative and automatic cognitive processes operating among members of a biotechnology firm (see also Shepherd, 2011).

Maussian Holism

Practice theory and compatible ideas from cognitive neuroscience have led some social scientists not to import binary theoretical categories directly from these fields, but rather to use cognitive science research in support of their arguments for studying social phenomena holistically, from multiple levels of analysis and in non-Cartesian fashion. Drawing from the French strand of practice theory represented by Marcel Mauss (Mauss, 1979/1935, 2006; Ignatow, 2012), Pierre Bourdieu, and Loic Wacquant (Lizardo and Strand, 2010), as well as from cognitive neuroscience research on 'mirror neurons' that shows that people have a capacity to understand the meaning

and intended purpose of others' actions "at an implicit, bodily level, without recourse to an explicit 'theory of mind' of other agents" (Lizardo, 2009). Ethnographers from anthropology and sociology have developed holistic ethnographic strategies that allow them to explore the implicit, embodied pedagogies (Bourdieu, 1990) that characterize both formal institutions and informal social settings (e.g., Desmond, 2006; Pálsson, 1994; Pagis, 2009; Wacquant, 2003) without necessarily distinguishing between different 'types' of cognition or between mental and corporeal phenomena.

Evaluation

The relations between human embodiment and cultural learning and action are complex, and there is no reason to imagine that one theoretical framework or set of theoretical categories would be capable of capturing this complexity in anything like its totality. Yet it is worth noting a few of the inherent strengths and weaknesses of the theories and categories that have developed so far in this area. For theories of embodiment and culture, the key elements that determine these theories' analytic value in empirical research have tended to be their degree of theoretical reductionism (i.e., their conceptual simplicity and elegance) as vs their ecological validity (i.e., their ability to explain real-world social phenomena).

The distinction between hot and cold cognition has proven its analytic value in a number of sociological studies, particularly within the field of social movements research, which for years was dominated by theories that emphasized strategic and instrumental rational action with little consideration of the role of emotions or noninstrumental values and ideals such as justice or dignity. But the turn from 'cold' to 'hot' theories risks amounting to a turn *back* to the century-old crowd psychology of Gustave Le Bon, who analyzed collective behavior as emotionally driven and therefore, ultimately, irrational. This oscillation in intellectual fashion from Le Bon's approach rooted in the body, to mentalistic rational choicelike theories and back again, discourages researchers from building upon older ideas to develop new insights, theories, and methods (Dobbin, 2001). A second, more fundamental critique is that the categories of hot and cold cognition are ultimately rooted in a Cartesian strand of cognitive theory – 'cognitivism' per se – that 'seems to have run its course' (Dreyfus, 2005; Lakoff, 2012). Ethnographic and laboratory studies have demonstrated that bodily, emotional, cognitive, and linguistic capacities do not operate as independent modules, but are rather inextricably intertwined: 'cool reason' depends on bodily and emotional information (see Damasio, 1994), while 'hot cognition' relies on all sorts of cognitive processing (Storbeck et al., 2006; Storbeck and Clore, 2007). Theories of hot cognition may be valuable as a corrective to rationalist/mentalist theories, but in the end they represent an effort to integrate emotions into cognitive sociology within a basically Cartesian framework.

Like theories of hot and cold cognition, dual-process theories potentially exaggerate the modularity (the functional independence) of different types of cognition. This exposes dual-process theories to critiques on grounds of ecological validity, because in practice, deliberative and automatic cognition are not readily identified or distinguished – though it should be recognized that analytic dualism may allow researchers to probe "whether and when the (more) 'mental' becomes (more) 'corporeal' and vice versa" (Vaisey and Frye, forthcoming). A second critique that has been leveled at dual-process theories is that they may be too psychological – too focused on the individual rather than social groups and interactions – for social science research conducted on groups and institutions. Fundamentally, dual-process theories treat cognitive schemas as internal to individuals rather than as products of person–context interactions (Shepherd, 2011: p. 123).

Analytic holism has demonstrated its value as an ethnographic strategy having impressive ecological validity, but a holism that is 'at once corporeal and mental' (Wacquant, 2004: p. 16) also has its weaknesses. In particular, it may not provide theoretical leverage on examining how the 'corporeal' and the 'mental' can influence each other sequentially over time via analytically distinct types of processes. Vaisey notes that if social learning is only ever analyzed, as Wacquant suggests, as *simultaneously* corporeal and mental, researchers cannot investigate the social conditions that cause cognition to become more or less deliberative (Vaisey and Frye, forthcoming). It appears at least at first blush that holistic analytic approaches may be useful mainly for ethnography, a methodology that has great ecological validity but is relatively weak in terms of its ability to identify generalizable patterns within data.

Methodological Strategies

Cognitive neuroscience and sociological practice theory have numerous implications for methods of social science research, having given rise to numerous revisions and reappropriations of extant research methods including ethnographic fieldwork methods, survey techniques, and content analysis methods. These revised and reappropriated methods are in general only loosely and contingently linked with the three theoretical strategies outlined above. Below I review a few methods that have been developed for the study of culture as embodiment, and focus on some of these methods' compatibilities, incompatibilities, and potential compatibilities with the three analytic strategies that have been developed in this area (hot and cold cognition, dual-process theory, and holism).

Surveys

Vaisey claims that one implication of dual-process theory is to cast doubt on the value of ethnographic interview methods for cultural analysis, because, Vaisey contends, such methods "engage with discursive consciousness alone" (2009: p. 1687) and thus do not give scholars much leverage on unconscious cognitive processes. Vaisey promotes fixed-response, forced-choice survey items as a superior method of cultural analysis because such items are more akin to the experience of solving everyday practical problems: forced-choice items potentially tap into practical consciousness, which "has to *make* (as opposed to *discuss*) many such decisions each day" (pp. 1688–1689): "When we hear a survey question, we simply

have to pick the response our practical consciousness prefers, the response that 'feels right' or 'sounds right' to us" (p. 1689).

Vaisey has analyzed survey data from the National Survey of Youth and Religion to demonstrate that people are often unable to articulate reasons for their moral judgments and decisions, but that moral schemas that are a part of practical consciousness are nevertheless predictive of future behavior. In another innovative study based on dual-process theory, Srivastava and Banaji (2011) used forced-choice survey items, supplemented by a laboratory experiment, to examine the interplay of cognition and social networks in a corporation with norms that emphasize cross-boundary collaboration. They found that social desirability concerns induced a disparity between how people viewed themselves in deliberative vs automatic cognition, and that these differences had implications for the resulting pattern of intraorganizational collaborative ties.

Ethnography

Within anthropology and sociology, the study of culture as embodied practice is associated primarily with ethnographic methods. Cultural anthropologists who make use of this framework have studied enskillment, apprenticeship, and socialization within indigenous communities. Pálsson (1994), for instance, uses his ethnographic studies of Icelandic fishermen to argue for a holistic and socially contextualized analysis of enskillment, which he finds has important implications for the economic success of communities as well as for differential success within communities (see also Bloch, 1998; Marchand, 2003; Harris, 2005; Simpson, 2006). Sociological ethnographic studies have focused on embodied dimensions of socialization into modern groups and institutions, such as a boxing gym (Wacquant, 2003), a fire fighting crew (Desmond, 2006), investment banks (Michel, 2011), a congressional political campaign (Mahler, 2011), and the US military (Lande, 2007). Sociologists have distinguished between the primary and secondary habitus, and analyzed how the primary habitus (developed in childhood and adolescence) influences later socialization into an occupational field and the development of an adult occupational habitus (Desmond, 2006). Sociologists have also emphasized the holistic character of shared memories of traumatic events (Spillman and Conway, 2007), the role of the bodily habitus in the negotiation of gender roles (Wade, 2011), and self-transformative practices such as meditation (Pagis, 2009) and religious conversion (Winchester, 2008).

Content Analysis

Content analysis refers to a collection of methods that include both qualitative and quantitative methods for analyzing themes in textual and visual data. These methods are commonly used in sociology, communications research, and political science. The possible implications of practice theory and cognitive neuroscience for content analysis methods in these disciplines have only begun to be explored. Linguists, however, have developed several new subfields that share with practice theory and cognitive neuroscience a holistic view of social actors, and that explore connections between human capacities for language, cognition, vision, emotion, learning, and memory (Feldman, 2006). These subfields include cognitive linguistics (Lakoff, 2012), construction theory (Goldberg, 1995), cognitive grammar (Langacker, 2008), and embodied construction grammar (Bergen and Chang, 2005). Linguists have also developed methods of formal text analysis based on these theoretical programs, such as collostructional analysis (Stefanowitsch and Gries, 2003).

Within the social sciences, Ignatow's quantitative text analysis studies are among a small number of published studies to incorporate practice theory and cognitive neuroscience into the theory and method of text analysis. Ignatow studied embodied metaphors used by programmers through a Durkheimian theoretical lens (2003), collocations of thematically related metaphors and syntactic patterns in transcripts from meetings of shipyard union workers in order to address theoretical questions about cultural coherence and fragmentation (2004), and semantic networks within the moral discourses of religious and secular overeaters support groups in order to investigate the effects of culture on patterns of group membership and participation (2009). While these studies demonstrate the applicability of practice theory and cognitive neuroscience to social science text analysis, the theoretical logic of this approach has not yet been explicated or formalized. As social media makes massive collections of user-generated text readily available to researchers, and as automated text analysis programs increase in power and sophistication, the value to researchers of text analytic strategies that are based on cutting-edge and empirically grounded theories of culture is likely to increase.

Summary

The study of culture as embodied practice rather than as systems of abstract symbols and signs is rapidly entering the mainstream of cultural analysis in anthropology and sociology. It appears to be only a matter of time before social scientists working on topics related to culture – such as learning, memory, knowledge, ritual, and socialization – will habitually take into account the bodily dimensions of their research topics. This article has suggested that up to this point cultural researchers have accounted for the body either by importing somewhat reductionist categories and mechanistic models from cognitive neuroscience, or else by insisting on using holistic, non-Cartesian, and multilevel analytic strategies that risk generating research with very limited generalizability.

As in all social science research, cultural researchers interested in the relations between culture and the body will need to choose the theoretical frameworks and methods best suited to the task at hand, and will preferably do so based on clear-eyed assessments of the tradeoffs involved, particularly tradeoffs between theoretical parsimony, ecological validity, and generalizability. At the same time, it is worth noting that there are voices within disciplines such as cognitive science (Strauss and Quinn, 1998) and psychology (Haidt and Graham, 2009) calling for more holistic and multilevel analyses of cognitive and social psychological phenomena. The sheer complexity and multidimensionality of culture,

particularly when studied as *techne* and *habitus* rather than only as *episteme,* may yet bring the long hoped for consilience of the social, cognitive, and biological sciences a step closer to reality. In the meantime, social scientists are better positioned than ever to understand some of the mechanisms through which culture is learned and reproduced without losing sight of the whole.

See also: Culture, Cognition and Embodiment; Culture: Contemporary Views; First Language Acquisition, Developmental Psychology of; Formal Methods of Cultural Analysis; Gene–Culture Coevolution; Habitus: History of a Concept; Mind–Body Dualism.

Bibliography

Abbott, Andrew, 2001. Chaos of Disciplines. University of Chicago Press, Chicago.
Barsalou, Lawrence, 1999. Perceptual symbol systems. Behavioral and Brain Sciences 22, 577–609.
Bergen, Benjamin K., Chang, Nancy, 2005. Embodied construction grammar in simulation-based language understanding. In: Östman, J.-O., Fried, M. (Eds.), Construction Grammar(s): Cognitive and Cross-language Dimensions. John Benjamins (Reprinted in 2007. Evans, V., Bergen, B., Zinken, J. (Eds.), The Cognitive Linguistics Reader. Equinox.).
Bloch, Maurice, 1986. From cognition to ideology. In: Fardon, R. (Ed.), Knowledge and Power: Anthropological and Sociological Approaches. Scottish University Press, Edinburgh, pp. 21–48.
Bloch, Maurice, 1998. Language, anthropology and cognitive science. In: How We Think They Think: Anthropological Approaches to Cognition, Memory and Literacy. Westview Press, Boulder, CO, pp. 3–21.
Bourdieu, Pierre, 1990. The Logic of Practice. Polity Press, Cambridge, UK.
Bourdieu, Pierre, Wacquant, Loïc J.D., 1992. An Invitation to Reflexive Sociology. The University of Chicago Press, Chicago.
Cerulo, Karen A., 1995. Identity Designs: The Sights and Sounds of a Nation. The Arnold and Caroline Rose Book Series of the American Sociological Association. Rutgers University Press, New Brunswick, NJ.
Cerulo, Karen A., 2002. Culture in Mind: Toward a Sociology of Culture and Cognition. Routledge, New York.
Cerulo, Karen A., 2010. Mining the intersections of cognitive sociology and neuroscience. Poetics 38 (2), 115–132.
Damasio, Antonio R., 1994. Descartes' Error: Emotions, Reason, and the Human Brain. G.P. Putnam's Sons, New York.
Desmond, Matthew, 2006. Becoming a firefighter. Ethnography 7, 387–421.
DiMaggio, Paul, 1997. Culture and cognition. Annual Review of Sociology 23, 263–287.
DiMaggio, Paul, 2002. Why cognitive (and cultural) sociology needs cognitive psychology. In: Cerulo, Karen (Ed.), Culture in Mind: Toward a Sociology of Culture and Cognition. Routledge, New York, pp. 274–281.
Dobbin, Frank, 2001. The Business of social movements. In: Goodwin, J., Jasper, J., Polletta, F. (Eds.), Passionate Politics. University of Chicago Press, Chicago, pp. 74–80.
Dreyfus, Hubert L., 2005. Overcoming the Myth of the Mental: How Philosophers Can Profit from the Phenomenology of Everyday Expertise. APA Pacific Division Presidential Address.
Feldman, Jerome, 2006. From Molecule to Metaphor. MIT Press, Cambridge, MA.
Geertz, Clifford, 1973. The Interpretation of Cultures. Basic Books, New York.
Goldberg, Adele E., 1995. Constructions. University of Chicago Press, Chicago.
Goodwin, Jeff, Jasper, James M., Polletta, Francesca (Eds.), 2001. Passionate Politics: Emotions and Social Movements. University of Chicago Press, Chicago.
Haidt, Jonathan, Graham, Jesse, 2009. Planet of the Durkheimians, where community, authority, and sacredness are foundations of morality. In: Jost, J., Kay, A.C., Thorisdottir, H. (Eds.), Social and Psychological Bases of Ideology and System Justification.
Harris, Mark, 2005. Riding a wave: embodied skills and colonial history on the Amazon floodplain. Ethnos 70 (2), 197–219.
Ignatow, Gabriel, 2003. Idea hamsters on the bleeding edge: profane metaphors in high technology Jargon. Poetics 31, 1–22.
Ignatow, Gabriel, 2004. Speaking together, thinking together? Exploring metaphor and cognition in a shipyard union dispute. Sociological Forum 19, 405–433.
Ignatow, Gabriel, 2009. Culture and embodied cognition: moral discourses in internet support groups for overeaters. Social Forces 88 (2), 643–669.
Ignatow, Gabriel, 2012. Mauss's lectures to psychologists: a case for holistic sociology. Journal of Classical Sociology 12 (1), 3–21.
Lakoff, George, 2012. Explaining embodied cognition results. Topics in Cognitive Science 4 (4), 773–785.
Lande, Bryan, 2007. Breathing like a soldier: culture incarnate. The Sociological Review 55 (s1), 95–108.
Langacker, Ronald W., 2008. Cognitive Grammar: An Introduction. Oxford University Press, Oxford.
Lizardo, Omar, 2009. Is a "special psychology" of practice possible? From values and attitudes to embodied dispositions. Theory & Psychology 19 (6), 713–727.
Lizardo, Omar, 2011. Pierre Bourdieu as a post-cultural theorist. Cultural Sociology 39, 651–688.
Lizardo, Omar, 2012. The three phases of Bourdieu's U.S. reception: comment on Lamont. Sociological Forum 27 (1), 238–244.
Lizardo, Omar, 2013. Habitus. In: Byron, Kaldis (Ed.), Encyclopedia of Philosophy and the Social Sciences. Sage Publications, London.
Lizardo, Omar, Strand, Michael, 2010. Skills, toolkits, contexts and institutions: clarifying the relationship between different approaches to cognition in cultural sociology. Poetics 38, 205–228.
Marchand, Trevor H.J., 2003. A possible explanation for the lack of explanation; or 'why the master builder can't explain what he knows': introducing the informational atomism against a 'definitional' definition of concepts. In: Pottier, J., Bicker, A., Sillitoe, P. (Eds.), Negotiating Local Knowledge. Pluto Press, pp. 30–50.
Mahler, Matthew, 2011. The day before election day. Ethnography 12 (2), 149–173.
Mauss, Marcel, 1979/1935. Sociology and Psychology: Essays (B. Brewster, Trans.). Routledge & Kegan Paul, London.
Mauss, Marcel, 2006. Techniques, Technology and Civilization. Durkheim Press/Berghahn Books, New York.
Michel, Alexandra, 2011. Transcending socialization: a nine-year ethnography of the body's role in organizational control and knowledge workers' transformation. Administrative Science Quarterly 56 (3), 325–368.
Pagis, Michal, 2009. Embodied self-reflexivity. Social Psychology Quarterly 72 (3), 265–283.
Pálsson, Gisli, 1994. Enskilment at sea. Man 29, 901–927.
Shepherd, Hana, 2011. The cultural context of cognition: what the implicit association test tells us about how culture works. Sociological Forum 26 (1), 121–143.
Simpson, Edward, 2006. Apprenticeship in Western India. Journal of the Royal Anthropological Institute 12, 151–171.
Spillman, Lyn, Conway, Bryan, 2007. Texts, bodies, and the memory of bloody sunday. Symbolic Interaction 30 (1), 79–103.
Srivastava, Sameer B., Banaji, Mahzarin R., 2011. Culture, cognition, and collaborative networks in organizations. American Sociological Review 76 (2), 207–233.
Stefanowitsch, Anatol, Gries, Stefan T., 2003. Collostructions: investigating the interaction between words and constructions. International Journal of Corpus Linguistics 8 (2), 209–243.
Storbeck, Justin, Clore, Gerald L., 2007. On the interdependence of cognition and emotion. Cognition & Emotion 21 (6), 1212–1237.
Storbeck, Justin, Robinson, Michael D., McCourt, Mark E., 2006. Semantic processing precedes affect retrieval: the neurological case for cognitive primacy in visual processing. Review of General Psychology 10 (1), 41–55.
Strauss, Claudia, Quinn, Naomi, 1998. A Cognitive Theory of Cultural Meaning. Cambridge University Press, Cambridge, UK.
Vaisey, Stephen, 2009. Motivation and justification: toward a dual-process theory of culture in action. American Journal of Sociology 114, 1675–1715.
Wacquant, Loïc, 2003. Body & Soul: Notebooks of an Apprentice Boxer. Oxford University Press, Oxford.
Wacquant, Loïc, 2004. Habitus. In: Beckert, Jens, Zafirovski., Milan (Eds.), International Encyclopedia of Economic Sociology. Routledge, London, pp. 315–319.
Wade, Lisa, 2011. The emancipatory promise of the habitus: Lindy Hop, the body, and social change. Ethnography 12 (2), 224–246.
Winchester, Daniel, 2008. Embodying the faith: religious practice and the making of a Muslim moral habitus. Social Forces 86 (4), 1753.

Relevant Website

http://sociology.rutgers.edu/cultcog/cultcog.html – Culture and Cognition Network.

Embodiment Theory

Anna M Borghi, University of Bologna, Bologna, Italy; and Institute of Cognitive Sciences and Technologies, Italian National Research Council, Rome, Italy
Fausto Caruana, University of Parma, Parma, Italy; and Istituto Italiano di Tecnologia (IIT), Brain Center for Social and Motor Cognition, Parma, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Embodied cognition (EC) views propose that cognition is shaped by the kind of body that organisms possess. We give an overview of recent literature on EC, highlighting the differences between stronger and weaker versions of the theory. We also illustrate the debates on the notions of simulation, of representation, and on the role of the motor system for cognition, and we address some of the most important research topics. Future challenges concern the understanding of how abstract concepts and words are represented, and the relationship between EC and other promising approaches, the distributional views of meaning and the extended mind views.

Definition of Embodied Cognition

Embodied cognition (EC) theory is intended as a response to the increasing dominance of the classic representational and computational theories of mind (RCTM) in cognitive science. Despite many versions of embodied theories, there are at least two commonalities between all EC approaches. The first is the view that cognitive processes are constrained by perception and motor processes, and therefore that the kind of body possessed by organisms shapes their cognition. The second is the refusal of the information processing model of the mind, and of the metaphor of the mind equated with software that manipulates symbols.

The view that perception and action contribute to cognition is in contrast with the classic view of cognition as separated from sensorimotor control, and that perception, cognition, and action are temporally and functionally independent processes. In contrast to this "sandwich model of the mind" (Hurley, 1998), proponents of EC underline the circularity of these processes, inspiring research programs aimed at showing that action influences both perception (Creem-Regehr and Kunz, 2010; Proffitt, 2006; Witt, 2011) and abstract thought (Goldin-Meadow and Beilock, 2010).

The second tenet, that is, the refusal of the view that cognitive processes involve computations on amodal representations (RCTM), leads to two different variants of EC. The critique advanced by proponents of the first variant is limited to the view that representations are amodal symbols, expressed in propositional format and arbitrarily linked to their referents (Fodor, 1975). In contrast, EC suggests that all concepts rely on the reactivation (or the simulation) of the sensorimotor experience with objects or events they refer to. The propositional approach has indeed to face the symbol-grounding problem (Harnad, 1990). In a nutshell, arbitrary symbols can be grounded only in other symbols, but in order to understand what objects are, we need to exit this vicious circle, 'grounding' the word meaning. The proponents of the first EC variant however, accept the computational side of the RCTM, bridging the gap between RCTM and EC. Accordingly, many EC theorists assume that cognition consists in computations on representation, but endorsing action-oriented representations expressed in bodily formats, including visuomotor, somatosensory, affective, and interoceptive formats (Goldman and de Vignemont, 2009). This EC variant is an improved version of the computational functionalisms on which RCTM is based. A more drastic critique to the second tenet comes from radical versions of EC (REC; Chemero, 2009; Hutto and Myin, 2013; van Elk et al., 2010; Wilson and Golonka, 2013) that refuse the assumption that cognition requires content of any kind. This antirepresentational version of EC abandons the idea that cognition requires content-involving representations and, given the RCTM mantra "no computation without representation", it also discards the computational thesis. Accordingly, REC considers cognition as a dynamical system characterized by continuously and interdependently changing variables that are better described by dynamic system theory (e.g., Spivey, 2007) than by representational explanations.

EC studies cover most areas of psychology and cognitive neuroscience (Borghi and Pecher, 2011; Davis and Markman, 2012; Chatterjee, 2010; Gentner, 2010), including development (Thelen and Smith, 1993), social cognition and emotions (Niedenthal, 2007; Semin and Smith, 2008; Becchio et al., 2010), attention, memory, and language (Barsalou, 2008; Pecher and Zwaan, 2005; Meteyard et al., 2012). Intriguing combinations between different research areas are emerging, including tighter relations between studies on attention, action, and social cognition (Galantucci and Sebanz, 2009; Knoblich et al., 2011). Aside from psychology and cognitive neuroscience, EC theories have developed in a variety of areas, including robotics and computer science (Ziemke, 2002; Arbib, 2006), linguistics (Lakoff, 2012), and philosophy (Chemero, 2009; Nöe, 2004; Hutto and Myin, 2013; Prinz, 2002; Shapiro, 2011).

Perception versus Action?

All EC approaches ascribe a crucial relevance to perception and action. However, the role played by perception or by action for

 http://dx.doi.org/10.1016/B978-0-08-097086-8.56025-5

cognition has been differently emphasized depending on the cultural tradition on which the various approaches relied. The theoretical background of EC can be found in American empiricism and pragmatism and in European phenomenological tradition. Even if some EC approaches underline the role of perception and other stress the importance of action, the two views are not incompatible; in addition, most EC theories emphasize the role of both, arguing that cognition is grounded in the sensorimotor system (see Borghi, 2005 for discussion).

The role of perception has been underlined in particular in research influenced by phenomenology. The "primacy of perception" was stressed by early phenomenologists, including Husserl and Merleau-Ponty, and inherited by contemporary phenomenologists (Gallagher and Zahavi, 2008). Furthermore, phenomenology impressed some of the early proponents of embodiment theories (Varela et al., 1991; Thompson, 2007). Finally, the interest in phenomenology offered new tools to study the body and its extensions (e.g., Berlucchi and Aglioti, 1997; Umiltà et al., 2008) and the relationship between the body and the self, stressing the dichotomy between body schema and body image, and that between sense of body ownership and sense of agency (Tessari et al., 2010; Tsakiris et al., 2007). The role of perceptual aspects has also been emphasized relying on the empiricist philosophical tradition, in particular on the work of Hume and Locke. Work by Barsalou and, on the philosophical side, by Prinz (2002), is in continuity with this tradition. According to the Perceptual Symbols Theory (Barsalou, 1999), no transduction process from sensorimotor experience to symbolic knowledge is necessary. Perceptual symbols, the building block of knowledge, have the combinatorial and productivity characteristics of arbitrary symbols, but they are modal rather than amodal, because they reflect the sensory qualities of the perceived entities. Stressing the role of perception has led some authors to remark the fact that bodily states do play a role but not an exclusive one. For this reason Barsalou and collaborators have proposed using the label 'grounded cognition' instead of 'embodied cognition' to underline the fact that cognition is grounded in a variety of situations, situated simulations, and not only in bodily states (Barsalou, 2008; Pezzulo et al., 2011).

The American pragmatist tradition and the ecological psychology of Gibson (1979) represent the theoretical background of the approaches that put a strong emphasis on action for cognition. Furthermore, an emphasis on the role of the motor system and of the overt behavior comes from the ordinary language philosophers, such as Gilbert Ryle and the late Wittgenstein. As nicely summarized by Wilson (2002) at the beginning of the EC wave, the principle underlying EC is that "knowledge is for action" – knowledge is both grounded and oriented to action. Accordingly, many authors described EC as a pragmatic turn in cognitive science, according to which cognition should not be understood as providing models of the world but as subservient to action, being grounded in sensorimotor coupling and in the ongoing interaction with the external world (Engel et al., 2013). In psychology, work on memory and on language grounding by Glenberg exemplifies this motor-oriented approach. In his BBS paper of 1997, significantly entitled "What Memory Is For," Glenberg argued that memory has an important adaptive role to support us in situated actions.

In neuroscience, the role played by the motor system in perception and cognition was emphasized by the research on the mirror and canonical neurons (Gallese et al., 1996; reviews: Rizzolatti and Craighero, 2004; Rizzolatti and Sinigaglia, 2010). Mirror and canonical neurons are located in the motor system and contribute to action execution. However, they support different cognitive processes. Mirror neurons are activated during the execution of a specific action and the observation of the same action performed by another agent, thus suggesting that the passive observation of actions recruits the corresponding motor representation in the observer motor system. After their discovery in different stations of the motor system, including the primary motor cortex, other similar mirror mechanisms have been found in emotional and sensory areas, suggesting that the perception and comprehension of others' intentional actions, emotional expressions, and somatosensory experiences depend on an implicit and automatic simulation within the sensorimotor system (see Buxbaum and Kalenine, 2010 for an embodied view according to which mirror neurons are not sufficient for comprehension of actions). The interpretation of the mirror neuron system offered by the Embodied Simulation theory (Gallese, 2001), according to which the observation of others' behavior triggers an automatic, subpersonal, and preconceptual simulation, provides an embodied account of intersubjectivity, in antithesis to the two cognitivist explanations of the theory of mind abilities (the theory–theory and the simulation theory). Furthermore, the mirror mechanisms opened new important perspectives in research on cognitive processes such as language comprehension and imitation, as well as a new key to interpret the social deficit in autism. Canonical neurons are premotor neurons that discharge both when interacting with objects and when passively observing them. Their discovery supports the view that perception involves the motor system and, in particular, the preparation of possible interactions with objects. This discovery contributed to launching studies and research on object affordances.

Ideomotor theories (Prinz, 1997) represent an important family of theories investigating the relationship between perception and action, and the goal-directed character of action. According to one of the most influent ones, the Theory of Event Coding (TEC) (Hommel et al., 2001), perceived events and actions rely on common representational structures, that is, they are represented by a network of feature codes. TEC predicts that perception is facilitated in case of overlap between the action an organism perceives and the action he or she is able to perform. Evidence in support of ideomotor theories ranges from imitation to attention to social cognition. For example, the compatibility between observed actions and responses influences the attribution of personal traits: finger-key responses lead to faster identification of individuals typing rather than performing sporty activities (Bach and Tipper, 2007). The mirror neuron system represents the neural underpinnings of ideomotor theories: our brain resonates more when observing actions we can reproduce than unknown action. Motor resonance has been documented in a variety of areas; a notable one concerns

empathy for others' pain (e.g., Avenanti et al., 2010; Iacoboni, 2009).

The Notion of Simulation

The concept of simulation was originally introduced in the field of social cognition, to account for our folk psychological abilities to predict others' behavior by ascribing mental states. This account was intended as a response to the theory–theory perspective, which conceives folk psychology as a theoretical and observational enterprise, suggesting that mindreading is rather based on using ones' own subpersonal models as the measure of everyone else's. In a seminal paper, Gallese and Goldman (1998) suggested a possible role of mirror neurons in mental simulation. More recently, theorists differentiated between an explicit (non-embodied) simulation, based on the view that simulation depends on a deliberative inference, versus an implicit, embodied simulation involving subpersonal activation of mirror mechanisms and shared representations (Gallese, 2005). Both versions were criticized by proponents of REC, suggesting that the resonance processes, including the mirror mechanisms, are part of the processes that underlie intersubjective 'direct perception' rather than simulation (Gallagher and Zahavi, 2008).

Besides of its use in social cognition, the concept of simulation was also used to account for the recruitment of the same perceptual, motor, and emotional system during the interaction with objects and entities in the world, as well as during imagery and language comprehension. Starting from this general idea, this notion now has a variety of connotations, also due to the different nuances that EC has assumed. Some authors intend simulation as a form of reenactment "of perceptual, motor, and introspective states acquired during experience with the world, body, and mind" (Barsalou, 2008), while others, underlining its predictive aspects, prefer to use the term 'motor simulation', intending it as a form of action preparation (Gallese, 2005; Grush, 2004). In addition, while some authors underline the implicit, automatic aspect of simulation (Jeannerod, 2006), others tend to equate it with a form of (motor) mental imagery, which may occur *a posteriori* (Decety and Grèzes, 2006). Finally, in their attempt to highlight dynamic and flexible aspects of cognition and to reject representations, REC avoids using the notion of simulation, particularly if intended as a form of reenactment.

The Agenda

EC theories have received ample support in a variety of domains. We will refer to some important areas, with no pretense of being exhaustive.

Affordances

The notion of affordance (Gibson, 1979) refers to the fact that objects invite organisms to act and suggest actions. For example, apples suggest grasping to humans, but not to worms. Affordances are therefore interactive properties, emerging from the relationship between organisms and environment, and pertain at the same time to perception and action. In the last years many studies have focused on affordances and 'micro-affordances', a term introduced by Ellis and Tucker (2000) to highlight both continuities and discontinuities with Gibson. Microaffordances are more specific that Gibson's affordances: they refer to action components (e.g., grasping) suitable for specific objects, thus implying object recognition. Furthermore, while Gibson's view is externalist and not focused on brain representations, microaffordances are neural representations corresponding to patterns in the brain of conjoint perception and action experiences. Numerous studies have demonstrated that observing objects activates affordances (e.g., grasping actions); recent work on affordances highlights their flexibility and contextual dependency (review: Thill et al., 2013).

Concepts and Words

According to all EC views, concepts and language are grounded in perception, action, and emotional systems (reviews: Coello and Bartolo, 2012; Fischer and Zwaan, 2008; Gallese, 2008; Jirak et al., 2010; Willems and Hagoort, 2007; special issues: Cappa and Pulvermüller, 2012; Cangelosi and Borghi, in press).

The influential Perceptual Symbols Theory (Barsalou, 1999) states that understanding concepts depends on the ability to form and use simulations. Thus, to decide whether a telephone rings we need to form a multisensory simulation. The phone ring is not stored in propositional terms, as a mental word; rather, it is the acoustic record of previous experiences with phones. This view received great impulse thanks to studies on representation of categories in the modal areas of the brain (Martin, 2007). EC views contrast with domain specific approaches, according to which in the brain there are innate categorical subsystems. Rather, concepts are intended as modality-specific distributed and flexible representations (Kiefer and Pulvermüller, 2012).

In continuity with this view, Zwaan (2004) proposes that comprehending language implies activating a simulation involving the motor system (see also Kemmerer, 2006). Many experiments have demonstrated that the activation in perception and motor areas during language comprehension is sensitive to the effectors (foot, hand, mouth), to kinds of affordances, to the bodily space (Ferri et al., 2011), to the direction implied by sentences (toward vs away from the body) (Glenberg and Kaschak, 2002), and to the emotional connotations of words (we tend to attract positive words and reject negative ones, Chen and Bargh, 1999); furthermore, it is highly flexible and sensitive to the context (van Dam et al., 2012).

In the strong EC perspective proposed by Glenberg (1997), concepts support us in selectively focusing on affordances allowing us to interact appropriately with objects. Hence, concepts are constrained by the kind of body we possess; for example, we conceive cups in terms of their distance from us, weight, and so forth – more generally, in terms of what we can do with them. Gallese and Lakoff (2005) argued that concepts and language exploit at a more sophisticated and higher level the multimodal character and the basic structures of the sensorimotor system (Anderson, 2010). In continuity with these views, Glenberg and Gallese (2012) have proposed that mechanisms of motor control have been exploited for language learning, comprehension, and

production (on the involvement of the same systems during language comprehension and production, see also Pulvermüller and Fadiga, 2010; Pickering and Garrod, 2013; on the idea that language evolves from manual gestures, Gentilucci and Corballis, 2006).

Body and Emotion

Given its disembodied view, cognitivism rarely paid due attention to the study of the body. In contrast, EC recovered some conceptual distinctions from phenomenology and neuropsychology, thus allowing neuroscientific investigations on the bodily experience and providing new insights on topics such as tool use, agency, and emotion.

The recovery of the distinction between body image and body schema, the former being a mental construct or a set of beliefs about the body, and the latter being a nonconscious model that monitors posture and movement (Gallagher and Zahavi, 2008), brings to a deeper understanding of different psychological disorders. At the same time, it shows that a number of objects, including prostheses and tools, can be incorporated into the body schema and perceived as part of one's own body. Notably, neurophysiological and neuropsychological evidence showed that the use of tools operating in the far space entails the updating of the map of the body schema and, as a consequence, it extends in the new operational space the peripersonal space anchored to the body (Maravita and Iriki, 2004). The above-mentioned distinction between sense of ownership and sense of agency (Tsakiris et al., 2007) is a further phenomenological distinction adopted by EC researchers, which has been largely used in the study of the bodily self, becoming a useful tool in the study of mental disorders such as schizophrenia (Ferri et al., 2012).

Finally, the attention to the bodily representations breathed new life into the debate on emotion. EC retrieved and renewed the somatic theory proposed by William James, according to which the emotional experience consists in the central representation of the bodily response to an emotional stimulus. New evidence revealed the existence of an afferent neural system that represents the physiological condition of the physical body (interoception), thus constituting a central representation of the 'material self' founding subjective feelings and emotions (Craig, 2002). Remarkably, this new link between interoception and emotion suggested that different psychological disorders, including neuroticism, social phobia, anxiety, and eating disorders, could be dependent on anomalous interpretations of the bodily interoceptive feedback. At the same time, the link between body and emotion contributed to the formulation of the 'facial feedback hypotheses' (Niedenthal, 2007), according to which the facial feedback affects the emotional experience; it follows that the motor production of emotional expression is part of the emotional experience, as predicted by early American pragmatists (Caruana and Gallese, 2012).

Critiques and Debates

EC has been the object of a lot of criticism. One of the most debated issues concerns the role played by the motor system in concepts and word processing. In an influential paper, Mahon and Caramazza (2008) argued that, even though the evidence showing that the motor system is involved in conceptual and language processing is indisputable, it can be explained through activation spreading from 'disembodied concepts' to the sensorimotor system interfacing with them. Hence, the role played by the sensorimotor system would be only epiphenomenal and not necessary for language comprehension. If it was necessary, then disruptions of the sensorimotor areas should impair language processing.

In contrast with this view, EC proponents argue that the motor system involvement is constitutive of the process. This is supported by evidence showing that the motor system activation occurs early, in an automatic, bottom-up fashion, and that it is somatotopic, that is, different areas of the brain are activated during processing of words related to action with different effectors, such as arms, legs, and mouth (pick, kick, lick) (Pulvermüller, 2005). The discussion is still open, however, because there are some controversial results. For example, not all studies have found early activation of the motor and premotor cortices during language processing; in addition, even if the motor system is modulated during language comprehension, there is evidence of both interference and facilitation (Willems and Franken, 2012); finally, evidence on patients is clear (review: Pulvermüller and Fadiga, 2010), but some clinical studies suggest that motor system lesions do not always lead to comprehension impairments. An interesting solution to this debate is advanced by enactivists (van Elk et al., 2010). As it has been proposed, the issue of necessity implies the idea that words are characterized by a core meaning, but enactivists underline that words are action oriented and contextual dependent, and that the human brain is plastic. This makes it possible to predict that comprehension is also possible when sensorimotor activation during language comprehension does not imply forms of reenactment.

Future Challenges

Some of the most important challenges for EC concern the explanation of abstract concepts and the relationship with other approaches, as the distributional approach and the extended mind view. One of the main challenges of EC is to account for concepts and words without concrete referents, such as 'phantasy'. Several EC theories of abstract concepts and words (ACWs) have been proposed (review: Pecher et al., 2011; special issues: Borghi and Pecher, 2011; Tomasino and Rumiati, 2013). Some views underline that the systems involved in concrete word representation are responsible also for representation of ACW, since not only concrete but also abstract words activate the motor system. Other EC theories underline the differences between concrete words and ACWs. The most influential theory claims that abstract concepts are represented in terms of concrete ones, through the mediation of metaphors (Lakoff and Johnson, 1999); for example, the abstract notion of time is explained in terms of the concept of space (Casasanto and Boroditsky, 2008). According to another view, ACWs differ from concrete ones due to their content: ACWs rely more on

situations and introspective features (Barsalou, 1999), and activate more emotions (Vigliocco et al., 2013). Recently multiple representation theories have been proposed (Dove, 2011; Barsalou et al., 2008). For example, according to the Words As Social Tools (WAT) view (Borghi et al., 2011; Borghi and Binkofsky, in press), both concrete words and ACWs are grounded in sensorimotor systems; however, for ACWs linguistic information plays a major role, due to their acquisition modality: they are primarily acquired through words rather than through interaction with their referents, and verbal labels help assemble sparse sensorimotor experiences.

Promising future directions concern the relationship between EC and two views emerging in different areas (computer science and philosophy), that is, distributional approaches (DA) to meaning and extended mind (EM) views.

DA and EC have been typically considered as antithetical. For example, according to the embodied Indexical Hypothesis (Glenberg and Robertson, 2000), words' referents evoke affordances, and these affordances constrain the relationships between words. In DA, meaning derives from the statistical co-occurrence of words in text corpora: from the relationship between associate words (real words, not mental words), not between words and their referents. Therefore words are not embodied, and the symbol-grounding problem is present. This distributive information is, however, able to account for many empirical findings, such as semantic priming effects. Recently some hybrid approaches have been proposed (review: Andrews et al., 2013). In hybrid views, conceptual processing is both linguistic and embodied, depending on the task (Louwerse, 2011). In EC multiple representation accounts, instead, access to meaning occurs only through simulation, but linguistic shallow tasks such as lexical decision can be performed without necessarily using full embodied simulations: linguistic distributional information can be used as a shortcut, allowing fast responses (Barsalou et al., 2008; Connell and Lynnot, 2013).

According to the extended mind (EM) view, the mind is not limited within the boundaries of our head/brain but is rather distributed in our brain, body, and external devices, which possess the power to augment our computational abilities (Clark and Chalmers, 1998). Even if many proponents of the EM view favor an embodied view, EC and EM views are typically considered independent perspectives, both because EM is more widespread in philosophy than in neuroscience and because of its functionalist flavor (Kiverstein and Clark, 2009). However, interesting convergences between these two areas are emerging (see Borghi et al., 2013; Borghi and Cimatti, 2010; Clark, 2008), particularly in research on the sense of the body and on bodily extensions.

See also: Cognition, Evolution of; Embodied Social Cognition; Health and Illness: Mental Representations in Different Cultures; Human Cognition, Evolution of; Mental Imagery: Visual Cognition; Mind, Theories of; Mirror Neurons, Theory of.

Bibliography

Anderson, M.L., 2010. Neural reuse as a fundamental organizational principle of the brain. Behavioral and Brain Sciences 33, 245–266.

Andrews, M., Frank, S., Vigliocco, G., 2013. Reconciling embodied and distributional accounts of meaning in language. Topics in Cognitive Science.

Arbib, M.A., 2006. From Action to Language via the Mirror System. Cambridge University Press, Cambridge, UK.

Avenanti, A., Sirigu, A., Aglioti, S.M., 2010. Racial bias reduces empathic sensorimotor resonance with other-race pain. Current Biology 20, 1018–1022.

Bach, P., Tipper, S.P., 2007. Implicit action encoding influences personal-trait judgments. Cognition 102, 151–178.

Barsalou, L.W., 1999. Perceptual symbol systems. Brain and Behavioural Sciences 22, 577–660.

Barsalou, L.W., 2008. Grounded cognition. Annual Review of Psychology 59, 617–645.

Barsalou, L.W., Santos, A., Simmons, K.W., Wilson, C.D., 2008. Language and simulations in conceptual processing. In: De Vega, M., Glenberg, A.M., Graesser, A.C. (Eds.), Symbols, Embodiment and Meaning. Oxford University Press, Oxford, pp. 245–283.

Becchio, C., Sartori, L., Castiello, U., 2010. Towards you: the social side of actions. Current Directions in Psychological Science 19, 183–188.

Berlucchi, G., Aglioti, S., 1997. The body in the brain: neural bases of corporeal awareness. Trends in Neurosciences 20, 560–564.

Borghi, A.M., 2005. Object concepts and action. In: Pecher, D., Zwaan, R.A. (Eds.), Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking. Cambridge University Press, Cambridge, pp. 8–34.

Borghi, A.M., Binkofsky, F. An Embodied View on Abstract Concepts. Springer, Berlin, New York, in press.

Borghi, A.M., Cimatti, F., 2010. Embodied cognition and beyond: acting and sensing the body. Neuropsychologia 48, 763–773.

Borghi, A.M., Flumini, A., Cimatti, F., Marocco, D., Scorolli, C., 2011. Manipulating objects and telling words: a study on concrete and abstract words acquisition. Frontiers in Psychology 2, 15. http://dx.doi.org/10.3389/fpsyg.2011.00015.

Borghi, A.M., Pecher, D., 2011. Introduction to the special topic embodied and grounded cognition. Frontiers in Psychology 2, 187. http://dx.doi.org/10.3389/fpsyg.2011.00187.

Borghi, A.M., Scorolli, C., Caligiore, D., Baldassarre, G., Tummolini, L., 2013. The embodied mind extended: words as social tools. Frontiers in Psychology 4, 214. http://dx.doi.org/10.3389/fpsyg.2013.00214.

Buxbaum, L.J., Kalenine, S., 2010. Action knowledge, visuomotor activation, and embodiment in the two action systems. Annuals of the New York Academy of Science 1191, 201–218.

Cangelosi, A., Borghi, A.M., in press. Special issue on action-language integration. Topics in Cognitive Science.

Cappa, S.F., Pulvermüller, F., 2012. Cortex special issue: language and the motor system. Cortex 48 (7), 785–787.

Caruana, F., Gallese, V., 2012. Overcoming the emotion experience/expression dichotomy. Behavioral and Brain Sciences 35 (3), 145–146.

Casasanto, D., Boroditsky, L., 2008. Time in the mind: using space to think about time. Cognition 106, 579–593.

Chatterjee, A., 2010. Disembodying cognition. Language and Cognition 2 (1), 79–116.

Chemero, A., 2009. Radical Embodied Cognitive Science. MIT Press, Cambridge, MA.

Chen, M., Bargh, J.A., 1999. Consequences of automatic evaluation: immediate behavioral predispositions to approach or avoid the stimulus. Personality and Social Psychology Bulletin 25, 215–224.

Clark, A., 2008. Supersizing the Mind. Embodiment, Action, and Cognitive Extension. Oxford University Press, Oxford.

Clark, A., Chalmers, D., 1998. The extended mind. Analysis 58, 7–19.

Coello, Y., Bartolo, A. (Eds), 2012. Language and action in cognitive neuroscience. Contemporary Topics in Cognitive Neuroscience Series. Psychology Press.

Connell, L., Lynnot, D., 2013. Flexible and fast: linguistic shortcut affects both shallow and deep conceptual processing. Psychonomic Bulletin & Review 20 (3), 542–550.

Craig, A.D., 2002. How do you feel? Interoception: the sense of the physiological condition of the body. Nature Reviews Neuroscience 3, 655–666.

Creem-Regehr, S.H., Kunz, B.R., 2010. Perception and action. Wiley Interdisciplinary Reviews: Cognitive Science 1 (6), 800–810.

Davis, J.I., Markman, A.B., 2012. Embodied cognition as a practical paradigm: introduction to the topic, the future of embodied cognition. Topics in Cognitive Science 4, 685–691.
Decety, J., Grèzes, J., 2006. The power of simulation: imagining one's own and other's behavior. Brain Research 1079, 4–14.
Dove, G., 2011. On the need for embodied and dis-embodied cognition. Frontiers in Psychology 1, 242.
Engel, A.K., Maye, A., Kurthen, M., König, P., 2013. Where's the action? The pragmatic turn in cognitive science. Trends in Cognitive Sciences 17 (5), 202–209.
Ferri, F., Frassinetti, F., Mastrangelo, F., Salone, A., Ferro, F.M., Gallese, V., 2012. Bodily self and schizophrenia: the loss of implicit self-body knowledge. Consciousnes and Cognition 21 (3), 1365–1374.
Ferri, F., Riggio, L., Gallese, V., Costantini, M., 2011. Objects and their nouns in peripersonal space. Neuropsychologia 49 (13), 3519–3524.
Fischer, M.H., Zwaan, R., 2008. Embodied language: a review of the role of the motor system in language comprehension. The Quarterly Journal of Experimental Psychology 61, 825–850.
Fodor, J.A., 1975. The Language of Thought. Harvard University Press, Cambridge.
Galantucci, B., Sebanz, N., 2009. Joint action: current perspectives. Topics in Cognitive Science 1, 255–259.
Gallagher, S., Zahavi, D., 2008. The Phenomenological Mind: An Introduction to the Philosophy of Mind and Cognitive Science. Routledge, London.
Gallese, V., 2001. The "Shared Manifold" hypothesis: from mirror neurons to empathy. Journal of Consciousness Studies 8 (5–7), 33–50.
Gallese, V., 2005. Embodied simulation: from neurons to phenomenal experience. Phenomenology and the Cognitive Sciences 4, 23–48.
Gallese, V., 2008. Mirror neurons and the social nature of language: the neural exploitation hypothesis. Social Neuroscience 3, 317–333.
Gallese, V., Fogassi, L., Fadiga, L., Rizzolatti, G., 1996. Action recognition in the premotor cortex. Brain 119, 593–609.
Gallese, V., Goldman, A.I., 1998. Mirror neurons and the simulation theory. Trends in Cognitive Science 2, 493–501.
Gallese, V., Lakoff, G., 2005. The brain's concepts: the role of the sensory-motor system in conceptual knowledge. Cognitive Neuropsychology 22, 455–479.
Gentilucci, M., Corballis, M.C., 2006. From manual gesture to speech: a gradual transition. Biobehavioral Reviews 30, 949–960.
Gentner, D., 2010. Psychology in cognitive science: 1978–2038. Topics in Cognitive Science 2, 328–344.
Gibson, J.J., 1979. The Ecological Approach to Visual Perception. Houghton Mifflin, Boston, MA.
Glenberg, A.M., Kaschak, M.P., 2002. Grounding language in action. Psychonomic Bulletin & Review 9, 558–565.
Glenberg, A.M., 1997. What memory is for. Behavioral and Brain Sciences 20, 1–55.
Glenberg, A.M., Gallese, V., 2012. Action-based language: a theory of language acquisition, comprehension, and production. Cortex 48, 905–922.
Glenberg, A.M., Robertson, D.A., 2000. Symbol grounding and meaning: a comparison of high-dimensional and embodied theories of meaning. Journal of Memory and Language 43 (3), 379–401.
Goldin-Meadow, S., Beilock, S.L., 2010. Action's influence on thought: the case of gesture. Perspectives on Psychological Science 5, 664–674.
Goldman, A., de Vignemont, F., 2009. Is social cognition embodied? Trends in Cognitive Science 13 (4), 154–159.
Grush, R., 2004. The emulation theory of representation: motor Control, imagery, and perception. Behavioral and Brain Sciences 27, 377–442.
Harnad, S., 1990. The symbol grounding problem. Physica D 42, 335–346.
Hommel, B., Müsseler, J., Aschersleben, G., Prinz, W., 2001. The theory of event coding (TEC): a framework for perception and action planning. Behavioral and Brain Sciences 24, 849–878.
Hurley, S., 1998. Consciousness in Action. Harvard University Press, Cambridge.
Hutto, D.D., Myin, E., 2013. Radicalizing Enactivism. Basic Minds without Content. MIT Press.
Iacoboni, M., 2009. Imitation, empathy, and mirror neurons. Annual Review of Psychology 60 (3), 653–670.
Jeannerod, M., 2006. Motor Cognition: What Actions Tell the Self. Oxford University Press, Oxford.
Jirak, D., Menz, M., Buccino, G., Borghi, A.M., Binkofski, F., 2010. Grasping language. A short story on embodiment. Consciousness and Cognition 19, 711–720.
Kemmerer, D., 2006. Action verbs, argument structure constructions, and the mirror neuron system. In: Arbib, M. (Ed.), From Action to Language via the Mirror Neuron System. Cambridge University Press, Cambridge, UK.
Kiefer, M., Pulvermüller, F., 2012. Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. Cortex 48 (7), 805–825.
Kiverstein, J., Clark, A., 2009. Introduction: mind embodied, embedded, enacted: one church or many? Topoi 28, 1–7.
Knoblich, G., Butterfill, S., Sebanz, N., 2011. Psychological research on joint action: theory and data. In: Ross, B. (Ed.), The Psychology of Learning and Motivation, vol. 54. Academic Press, Burlington, pp. 59–101.
Lakoff, G., 2012. Explaining embodied cognition results. Topics in Cognitive Science 4 (4), 773–785.
Lakoff, G., Johnson, M., 1999. Philosophy in the Flesh. Basic Books.
Louwerse, M.M., 2011. Symbol interdependency in symbolic and embodied cognition. Topics in Cognitive Science 3, 273–302.
Mahon, B., Caramazza, A., 2008. A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. Journal of Physiology 102, 59–70.
Maravita, A., Iriki, A., 2004. Tools for the body (schema). Trends in Cognitive Science 8, 79–86.
Martin, A., 2007. The representation of object concepts in the brain. Annual Review of Psychology 58, 25–45.
Meteyard, L., Cuadrado, S.R., Bahrami, B., Vigliocco, G., 2012. Coming of age: a review of embodiment and the neuroscience of semantics. Cortex 48, 788–804.
Niedenthal, P.M., 2007. Embodying emotion. Science 316, 1002–1005.
Nöe, A., 2004. Action in Perception. MIT Press, Cambridge, MA.
Pecher, D., Boot, I., Van Dantzig, S., 2011. Abstract concepts: sensory-motor grounding, metaphors, and beyond. In: Ross, B.H. (Ed.), The Psychology of Learning and Motivation. Academic Press, Burlington, pp. 217–248.
Pecher, D., Zwaan, R.A., 2005. Grounding Cognition. The Role of Perception and Action in Memory, Language, and Thinking. Cambridge University Press, Cambridge, MA.
Pezzulo, G., Barsalou, L.W., Cangelosi, A., Fischer, M.A., McRae, K., Spivey, M., 2011. The mechanics of embodiment: a dialogue on embodiment and computational modeling. Frontiers in Cognition 2 (5), 1–21.
Pickering, M.J., Garrod, S., 2013. An integrated theory of language production and comprehension. Behavioral and Brain Sciences 36 (4), 329–347.
Prinz, W., 1997. Perception and action planning. Europena Journal of Cognitive Psychology 9, 129–154.
Prinz, J.J., 2002. Furnishing the Mind. Concepts and Their Perceptual Basis. MIT Press, Cambridge, MA.
Proffitt, D.R., 2006. Embodied perception and the economy of action. Perspectives in Psychological Science 1 (2), 110–122.
Pulvermüller, F., 2005. Brain mechanisms linking language and action. Nature Reviews Neuroscience 6 (7), 576–582.
Pulvermüller, F., Fadiga, L., 2010. Active perception: sensorimotor circuits as a cortical basis for language. Nature Reviews Neuroscience 11 (5), 351e360.
Rizzolatti, G., Craighero, L., 2004. The mirror neuron system. Annual Review of Neuroscience 27, 169–192.
Rizzolatti, G., Sinigaglia, C., 2010. The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. Nature Reviews Neuroscience 11, 264–274.
Shapiro, L., 2011. Embodied Cognition. Routledge Press, New York.
Semin, G.R., Smith, E.R. (Eds.), 2008. Embodied Grounding: Social, Cognitive, Affective, and Neuroscientific Approaches. Cambridge University Press, New York.
Spivey, M.J., 2007. The Continuity of Mind. Oxford University Press, New York.
Tessari, A., Tsakiris, M., Borghi, A.M., Serino, A., 2010. The sense of body: a multidisciplinary approach to body representation. Introduction to the special number on the sense of body. Neuropsychologia 48, 643–644.
Thelen, E., Smith, L.B., 1993. A Dynamic Systems Approach to Development. MIT Press, Cambridge, MA.
Thill, S., Caligiore, D., Borghi, A.M., Ziemke, T., Baldassarre, 2013. Theories and computational models of affordance and mirror systems: an integrative review. Neuroscience and Biobehavioral Reviews 37, 491–521.
Thompson, E., 2007. Mind in Life: Biology, Phenomenology and the Sciences of Mind. Harvard University Press, Cambridge, MA.
Tomasino, B., Rumiati, R.I., 2013. Special topic on: What does neuropsychology say about the role of sensorimotor processes in conceptual knowledge and abstract concepts. Frontiers in Human Neuroscience 7, 498. http://dx.doi.org/10.3389/fnhum.2013.00498.
Tsakiris, M., Schütz-Bosbach, S., Gallagher, S., 2007. On agency and body-ownership: phenomenological and neurocognitive reflections. Consciousness and Cognition 16, 645–660.

Umiltà, M.A., Escola, L., Intskirveli, I., Grammont, F., Rochat, M., Caruana, F., Jezzini, A., Gallese, V., Rizzolatti, G., 2008. When pliers become fingers in the monkey motor system. Proceedings of the National Academy of Sciences of the United States of America 105, 2209–2213.

van Dam, W.O., van Dijk, M., Bekkering, H., Rueschemeyer, S.A., 2012. Flexibility in embodied lexical-semantic representations. Human Brain Mapping 33, 2322–2333.

van Elk, M., Slors, M., Bekkering, H., 2010. Embodied language understanding requires an enactivist paradigm of cognition. Frontiers in Psychology 1, 234.

Varela, F.J., Thompson, E., Rosch, E., 1991. The Embodied Mind: Cognitive Science and Human Experience. MIT Press, Cambridge, MA.

Vigliocco, G., Kousta, S.T., Della Rosa, P.A., Vinson, D.P., Tettamanti, M., Devlin, J.T., Cappa, S.F., 2013. The neural representation of abstract words: the role of emotion. Cerebral Cortex 24 (7), 1767–1777.

Willems, R.M., Franken, J.C., 2012. Embodied cognition: taking the next step. Frontiers in Psychology. http://dx.doi.org/10.3389/fpsyg.2012.00582.

Willems, R.M., Hagoort, P., 2007. Neural evidence for the interplay between language, gesture and action: a review. Brain and Language 101, 278–289.

Wilson, A.D., Golonka, S., 2013. Embodied cognition is not what you think it is. Frontiers in Psychology. http://dx.doi.org/10.3389/fpsyg.2013.00058.

Wilson, M., 2002. Six views on embodied cognition. Psychonomic Bulletin & Review 9 (4), 625–636.

Witt, J.K., 2011. Action's effects on perception. Current Directions in Psychological Science 20, 201–206.

Ziemke, T. (Ed.), 2002. Situated and Embodied Cognition. Special Issue of the Journal of Cognitive Systems Research 3 (3).

Zwaan, R., 2004. The immersed experiencer: toward an embodied theory of language comprehension. In: Ross, B.H. (Ed.), Psychology of Learning and Motivation, vol. 44. Academic, New York, pp. 35–62.

Relevant Websites

http://laral.istc.cnr.it/borghi/.
http://www.iit.it/people/fausto-caruana.html.

Emergency Relief and Social Work Responses

Desley Hargreaves, School of Social Work and Human Services, University of Queensland, Brisbane, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article is a practice-based reflection on the role social workers can play at the strategic and service intervention level when responding to disasters and emergency events. It discusses two significant events which impacted Australia – the Bali bombings in October 2002 and Victorian bushfires in February 2009 – and shaped policy responses by government. It outlines the lessons learned by social workers from this involvement. The profession needs to examine how it can take a stronger leadership role in emergency response and recovery.

Introduction

On 12 October 2002, 88 Australians died as the result of a terrorist attack at two popular night spots for tourists in Bali, Indonesia. Social workers in the Australian government agency Centrelink (now incorporated into the Commonwealth Department of Human Services – DHS) became a critical part of the onshore response. In May 2003, Centrelink social workers began providing support offshore to families and survivors attending the trials of the bombers in Bali. In February 2009, the Australian State of Victoria burned: 173 people died and over 2000 homes were destroyed. Some communities suffered major loss of critical infrastructure, homes, and community gathering places. Social workers found themselves on the front line of the nation's and the state's response. This article tells their story to illustrate what social workers can offer in disaster response. Building on insights gained from involvement in a range of disasters and emergencies, and knowledge the profession has developed since the 2004 Indian Ocean Tsunami and 2005 Hurricane Katrina in the United States, I illustrate how in times of national trauma, social work can position itself at the center of relief and recovery efforts. I suggest that the profession should examine how it can be more strategically represented in emergency response and recovery.

There is much debate about definitions of disaster (Dominelli, 2012; Kreps, 1995) albeit within a consensus that they tend to be overwhelming, often sudden in onset, cause significant damage, and possible loss of life, beyond the capability of the affected area to respond and therefore requiring outside assistance in response and recovery. For this article, I use the following:

> A serious disruption to community life which threatens or causes death or injury in that community, and damage to property which is beyond the day-to-day capacity of the prescribed statutory authorities and which requires special mobilisation and organisation of resources other than those normally available to those authorities.
> Emergency Management Australia Glossary, 1998: pp. 22–23

Disasters can happen anywhere and occur often without warning. Many natural disasters are being attributed to the effects of climate change and global warming. Since the beginning of the twenty-first century, many parts of Australia have experienced floods, bushfires, cyclones, severe weather events, equine influenza outbreaks, severe drought, and minor earthquakes. Australia experienced the impact of terrorism when Australians died in the World Trade Center attack in 2001. However, the Bali bombings outraged Australians who felt their traditional sense of safety and security threatened in a country that had long been a favorite holiday destination. This event set a new direction for the Australian government's response to survivors and families of deceased, and those killed or injured overseas as a result of terrorist events or natural disasters.

Overview of the Literature: Social Work and the Response to Disasters

There is a strong body of literature on the human response to and experience of disasters and emergencies. There is also an increasing body of literature about social work and disasters (Rowlands and Tan, 2008; Taubman and Weisz, 2011; Dominelli, 2012). Raphael (1986) researched the human responses to the different phases of disaster and discussed in some detail the recovery trajectory for individuals and community. This seminal work was based on her research on the social disconnection which occurred with the evacuation of families from Darwin following Cyclone Tracy. Her findings on this issue link closely with the work of Hobfoll et al. (2007). In their review of the empirical evidence of the five essential elements of immediate and mid-term mass trauma intervention, they identify the 'Promotion of Connectedness' (pp. 296–298) as of central importance. They discuss the body of research that identifies the constructive role that social support can play in combating stress and trauma.

Mathbor (2007) has written about the role social workers can and should play in preparation and planning for response as well as in the recovery effort. In this article, he discusses social capital and the challenges of mobilizing networks, contacts, and cohesion after a disaster has occurred. He also identifies the work needed at times to focus community collaboration, coordination, and use of existing networks to support effective service delivery. Mathbor (2010) further highlights skills and training that social workers need to work effectively and within their role in the recovery effort.

Dominelli (2012) proposes that boundaries between natural and hu(man) made disasters are becoming increasingly blurred. She examines different kinds of disasters, the geopolitics behind them, and the unsustainability of current forms of industrialization, and challenges the social workers to become more strategically involved.

Zakour (2005) discusses the varied roles that social workers play in disasters and reflects on their contribution to various research agendas. He notes the focus on responding to the social disruption brought about by disasters and emergencies and discusses the concepts of vulnerability at both the individual and community level. Social structural factors including poverty and social isolation would traditionally be a focus for social workers. He notes that in ensuring improved access to services and resources, social workers engage in a range of activities such as case management, brokerage, advocacy, and outreach. He also discusses the contribution social workers make in organizational and interagency networks.

Themes such as grief and loss, trauma, strengths-based interventions, and resiliency have received much attention. Tedeschi and Kilmer (2005) highlight the move in thinking from a deficit to a strengths approach in clinical practice. They further discuss concepts of resilience and the possibility of growth after trauma. This reflects work by Richardson (2002) where his review of articles focuses on the work that examines individual strengths that can be used to assist people to cope with and develop through their experience of trauma. This is further reinforced by the work of Walsh (2007) who brings those concepts into the field of traumatic loss and major disasters. Concepts of hope building are also identified as important.

Cronin et al. (2007) offer a useful analysis of the American Red Cross staff support program and comment on the lead role that social workers could play in providing support programs not only in disasters but also in caregiver support. Alston (2013) and Enarson and Meyreles (2004) bring a focus to the gender dimension of emergency response both from a historical and a geographical perspective. Both articles identify the risks of ignoring this dimension in planning for and responding to the effects of disasters.

Social Work and Disaster Services in Australia

Australia is fortunate to have a Social Work Service located in a critical Commonwealth (Federal) government service delivery agency. The Social Work Service comprises approximately 700 professionally trained social workers within an agency of approximately 35 000 staff, delivering income support, child support, health, and related services to millions of Australians. These social workers are located in customer service centers, call centers, and other national teams across Australia. In the two examples cited earlier, the Social Work Service was able to position itself in the very early days as an integral part of the response from the strategic to the tactical and operational levels. This was contrary to the structural arrangements previously in place in the Department.

The concept of social work involvement in responding to emergencies and disasters is not new. Social work literature on this work has increased significantly over recent years (Rowlands and Tan, 2008; Mathbor, 2007; Dominelli, 2012). Social workers play a critical and often unrecognized role in the early recovery period, as well as in supporting individuals, families, and communities in the longer term. In Australia, social workers are involved at the Commonwealth, State, and local level as part of the psychosocial recovery arrangements. In the UK, social workers are statutory second responders after people have been evacuated, along with health providers (Child et al., 2008). Tan et al. (2011) highlight the involvement of social workers in both China and Taiwan in responding to various disasters but express concern at their lack of training for the task. Iravani and Ghojavand (2005) discuss the positive role social workers have played in Iran and Taiwan following disasters.

Each country has a process for responding to disasters whether formalized or reactive. In Australia, a formalized approach has been refined over many years, since Cyclone Tracey struck Darwin on Christmas Day in 1974. This led to the development of national, state, local government, and local district planning guidelines for emergency response and recovery. Australia has been supported in this by a formal agency – Emergency Management Australia – which sits within the Commonwealth Attorney General's Department.

Policy Responses

Tarrant (2006) notes that a disaster is a very complex policy problem which cuts across all levels and functions of government. The Australian government has shaped its policy response to disasters and other emergencies based on its experience of different types of disasters over many years. Feedback from constituents, those impacted by disasters, workers involved in community recovery, and evidence-based research has also shaped the changing policy response in Australia. Formal arrangements exist between the Commonwealth and State and Territory governments which activate the flow of funds and other assistance depending on the scale and nature of the disaster. It also provides a Commonwealth payment known as the Australian Government Disaster Relief Payment. This is made available to individuals through the Commonwealth DHS based on a declaration by the Attorney General. A similar arrangement exists in the US (FEMA, 2012).

With offshore crisis responses, the nature of the assistance has been determined by government specific to each circumstance. The Bali bombings resulted in a package of assistance including medical, funeral, financial, and support services onshore, as well as the provision of financial and support services offshore when the families of victims and survivors were funded to attend the trial of the Bali bombers. In 2012, the Australian Parliament passed a legislation which provides financial compensation for victims of offshore terrorist events – a further step by government to fine tune its policy response to a specific cohort of victims and survivors.

The Family Support Team – Support to Australians Impacted by Offshore Disasters or Terrorist Events

Following the Bali bombings, the Centrelink Social Work Service established a case management service (the Family Liaison Service) which provided a coordinated whole of

Commonwealth government response to families of the deceased and survivors in Australia. Initially this service played the role of a safety net service as Red Cross Australia provided a service for the first 12 months after the bombings. The safety net was intended to cater for those affected who did not wish to engage with Red Cross or other service providers but required support to access the range of assistance available. Centrelink social workers in locations where survivors and families of the deceased lived, undertook this family liaison officer role in addition to their usual role within the agency.

At the request of the Interdepartmental Emergency Taskforce (IDETF) chaired by the Department of Foreign Affairs and Trade (DFAT), social workers were deployed to Bali for the first time in May 2003. This was an entirely new type of work. Centrelink had never delivered services offshore previously and there was considerable organizational risk, pushing the boundaries of the agency's remit. Staff operated with the ongoing threat of further bombs. They worked in another country – different jurisdictions, different cultures, very humid climate, health risks (SARS), and at least 8 h flight from their support service. Being Australian potentially made them a target for attack; so engaging appropriate security protocols with the advice of the Australian Federal Police (AFP) who had a long-standing relationship with Indonesian National Police was mandatory.

The deployment was for a period of 7 months to provide personal and psychosocial support to those attending the trials. These staff members worked in gender-balanced pairs and were rotated every 3 weeks throughout that period. The role was developed, clarified, and negotiated within a very short time frame by Centrelink's chief social worker during an urgent trip to Bali, prior to the first social workers being deployed. Because this work fell outside the normal operations of the agency, the Social Work Service needed to negotiate access to funding, develop guidelines for deployment, remuneration, occupational health and safety arrangements, administrative arrangements, communication options, and set up daily supervisory and immediate support access arrangements for the staff offshore.

Subsequently a Family Support Team (FST) was established comprising social workers who could be deployed offshore as part of the Australian Government's Emergency Response team. They worked alongside consular and other staff from DFAT, the AFP, and a range of staff from other key departments who were deployed as part of the broader response team. A critical element of FST work is based on the social work practice principle – 'Start where the client is' (Goldstein, 1983). Bearing witness is often the starting point. In their article 'After the Wave,' Manning et al. (2007) note that practical assistance is a key component of effective recovery. This creates an opportunity to start trust building. This was particularly important when the FST had a period of 7 days to work with affected people. This was the period funded through government assistance, for people to travel offshore to be with, locate, or commemorate their loved ones. The FST provided supportive intervention to assist people to make necessary decisions within a short time frame and develop some basic coping strategies until they could return to their usual support network in Australia. Such decisions could have been about whether to stay or go home when it became clear that the Disaster Victim Identification, a long and painstaking process, could stretch into weeks. It could have been about where the body of a loved one should be returned to, whether cremation was the preferred option, or whether to visit the site of the tragedy.

The initial deployment informed the role undertaken by social workers for the first anniversary commemorations where the government funded over 500 families and individuals to attend the services in Bali on 12 October 2003 as well as in Canberra a few days later. Centrelink sent 11 social workers to Bali for this event. Centrelink's chief social worker headed up a team of 28 staff from four different departments and agencies who provided support to family members and survivors. For the Canberra service, a senior manager in the Social Work Service led the family support arrangements. Social workers were involved at the strategic (IDETF and agency Executive) and tactical level in influencing the development of the policy response, providing feedback on anomalies, sourcing and providing feedback from families, and shaping and providing the service delivery response.

Subsequent to that, social workers at all levels have been an integral part of recovery efforts onshore and critical response services offshore. These have included deployments to Thailand and Sri Lanka following the Indian Ocean Tsunami, to London and Bali after the 2005 bombings in both locations, and associated anniversary events including the tenth anniversary of the 2002 bombings in October 2012. Social workers were deployed to Cyprus and Turkey to support evacuated Australians during the Lebanon crisis in 2006, to Yogyakarta, Indonesia, and Papua New Guinea following crashes of civilian aircraft, and to Jakarta and Mumbai following terrorist events there. Further deployments included those to Frankfurt to meet evacuated Australians during the crisis in Cairo (2011), to New Zealand following the mine collapse at Pike River in 2010, Christchurch following the earthquake in 2011, and Japan following the earthquake, tsunami, and nuclear event in 2011. The social work role has remained consistent with that originally developed for Bali. In all of these deployments, social workers played significant leadership roles in providing advice about psychosocial impacts, delivering services on the ground, and contributing to the shaping of the government's policy response. Providing feedback on the impact of policies and policy anomalies has been an essential and valued role.

Service integration is critical. The FST social workers' short-term offshore intervention is supported by the longer-term onshore Family Liaison Service which links people to appropriate support services in their local community. The staff who deploy offshore facilitate direct referrals of families or individuals to their onshore colleagues so that people do not have to tell their story more than once to get help, all Commonwealth assistance is coordinated through one person, and people are linked in to local services that can respond to their needs. These local referrals can vary from assistance with the execution of wills and finalization of estates through to interventions from qualified trauma professionals. This integration also includes the other agencies with which staff work offshore in the response. There are always challenges in managing such relationships, but doing so is essential to ensuring a seamless service for those affected.

The literature review highlighted the fact that social connection is crucial to effective recovery. The experience of FST

social workers working offshore has been that it is important to identify those who lack support or where support networks are very fragile, and prioritize time with them. Some people seek contact with others who have been affected by the same disaster. One part of the social work role has been to facilitate such contacts – usually informal gatherings at a café or hotel. Others have preferred to manage their grief in private away from the scrutiny of the media or well-intentioned support services. Social workers need to be wary about mistaking solitude for desperate aloneness. Men and women demonstrate different help-seeking behaviors in these circumstances. Responding appropriately can be challenging.

The Victorian Bushfires

Following the devastating bushfires on 7 February 2009, the then Prime Minister of Australia and the Premier of Victoria announced on 10 February that everyone affected by the fires would be allocated a case manager within 24 h of registration with what became known as the Victorian Bushfire Case Management Service (VBCMS). The service did not exist at the time of the announcement nor was a registration process in place. Social workers took a leadership role in the development and delivery of VBCMS. It had to be operational within 3 days of the announcement. It was unclear how many people would require assistance. The announcement was quite open-ended – anyone affected by the fires. Under the emergency management arrangements, the Victorian government was in charge of the recovery arrangements. The Commonwealth government supported in many ways but was not in charge.

Two Senior Executive staff from the Victorian DHS (DHS Vic) were assigned to manage the establishment of the service. At the Commonwealth level, Centrelink's chief social worker was re-assigned to work with the two Victorian Executives to develop and operationalize the service. All three staff were social workers with backgrounds in disasters, human resources, and case management. The potential for sensitivities in Commonwealth/State relationships could have interfered with the task but a very constructive, collaborative working relationship developed. This was supported by the deployment of a number of staff with social work skills to key positions to support the three Executives in the development of VBCMS. Challenges included developing a shared understanding of what a case management service would deliver (including standards, quality assurance, registration process) and finding staff with appropriate skills to undertake case management at very short notice. At a practical level, it was easier for the Commonwealth to mobilize its employees quickly. The first case managers were deployed on 13 February as the case management service structure developed around them. At the end of the first week of operation, a client database had been deployed, 2428 registrations had been received, 100 case managers were employed, and 1207 cases were assigned. A resource kit and core training had been developed and was being delivered. By 31 March, all cases were finally allocated.

Such a large-scale and combined Commonwealth/State case management response had never been undertaken before in Australia. Staff redeployment from across the country to many affected communities within Victoria required highly organized and structured management arrangements. DHS Vic negotiated with its service partners that were funded to deliver a range of social services on its behalf to also undertake case management roles. Many convincing arguments were required as the Executives met with Regional Executive groups and Community Service Organizations. This was not ordinary community recovery work.

The three Executive staff remained fully involved for the first 10 weeks before starting to hand over responsibility to other senior staff. This hands on approach assisted in risk mitigation and management and enabled troubleshooting to occur quickly, in a highly volatile political environment. Governance structures were established quickly and daily meetings at all levels of the service operations ensured communication of key messages was reinforced and feedback could be responded to urgently. Because of the size and complexity of the Commonwealth response, a Commonwealth Professional Services structure was established, identifying lines of accountability, professional support, and supervision. The Commonwealth also operated the VBCMS call center for registering clients, staffed by 25 social workers. A number of affected people moved interstate. To ensure consistency in case management provision for these people, Centrelink social workers in the area closest to where the person had relocated, provided this service under the auspice of VBCMS.

The VBCMS was designed to operate for 2 years. Over 70 agencies worked under its banner providing over 450 case managers. The Commonwealth provided the initial surge capability with case managers building up to 105 and withdrawing in stages as the local community service organizations successfully recruited case managers. The Commonwealth officially withdrew its case managers at the end of December 2009 as cases were handed over to the local community service organizations for follow up. Leadership, relationship management, stakeholder management, and communication were essential to the development and continuing operation of the service. Working collaboratively – being respectful of each other's role – was critical.

It is well understood in the disaster literature that most people will recover from disaster without professional intervention. Most people turn first to their normal social networks for support (Kaniasty and Norris, 2004; Hobfoll et al., 2007). Connecting them becomes a critical role when these normal networks are disrupted by the event. The Victorian bushfires created some major disruptions as whole communities were damaged and people were relocated and disconnected from their normal social networks. The role of case managers was crucial in reconnecting, ensuring community information was provided, and facilitating attendance at meetings where the future of communities was being discussed.

Lessons Learned

The role of social workers in leadership is an important one. Social work leadership takes place at all levels in an emergency from strategy to delivery on the ground. A clear decision making and reporting structure needs to be established quickly. This is often described as a command and control leadership style and important in helping to bring some sense of order for staff to the environment in which they will be working. This

was very evident in the response to the Victorian bushfires where the task seemed overwhelming. Consensus decision making did not lend itself to the need to respond to the situational ambiguity as fires continued to burn, and policy and political imperatives kept changing. This style is one that does not sit easily with social workers but experience over many years has shown that once these arrangements are in place, staff can focus more on the actual service delivery rather than the uncertainty of the environment within which they are operating. Role clarity is equally important as are the staff support arrangements. Daily debriefs and operational reviews enable the service to be tailored to constantly changing needs and staff members are key contributors to these decisions. The conduct of these is a clear leadership responsibility.

For social workers returning to their normal work and home environment after the intensity of involvement in an emergency response, the adjustment back to what used to be normal can be very challenging. The disaster becomes a point of reference in people's lives and all involved are impacted and changed in some way by what they have seen, heard, experienced, and done. This highlights the importance of quality professional supervision, good staff support arrangements, and appropriate follow up.

Social workers operate as part of a broader team in such recovery efforts. They bring a unique perspective to the assessment of need which is ever changing as the disaster unfolds and the recovery effort ramps up. The concept of integrated, collaborative service delivery is critical to supporting those impacted.

The purpose of the intervention in these situations is to help people to create some order out of chaos and regain some sense of control of their lives, and over time to find some meaning and hope out of something catastrophic. It is critical that social workers understand their role, the context in which they are working, and the stage of the response in which they are involved.

Building social connectedness is an important intervention but it can be challenging to connect and create a way forward. Traditional divisions in families and the community become evident fairly quickly, comparisons are made about the level of assistance received, and notions of 'deserving' and 'undeserving' emerge (Kaniasty and Norris, 2004). The real challenge is in identifying those people who have lost or never had any social support networks and supporting them to re-engage with or access mainstream support services. These are often individuals or families who have poor help-seeking behaviors and who can be socially marginalized.

Support needs can increase when the disaster has gone off the front page and generally people think that it is time for those affected to get on with their lives. Even some of the strongest family support networks can break down after a period where they feel they have nothing left to give emotionally to the affected person. It is critical that the role of the helping professional becomes one of enabler or facilitator, assisting people to regain control of their lives and to restore or improve their level of functioning.

The intervention model is built on a strengths-based approach (Brun and Rapp, 2001). Helping people to identify how they have coped with a difficult situation in the past can give them confidence to cope with the challenge of the current situation and to have a belief that they will get through or survive this experience. The strengths-based approach aids empowerment and recognizes the individual or community's ability to contribute to recovery. Social workers' knowledge of how groups and communities work is invaluable in understanding the dynamics, where the fractures might be, and where the strengths and social capacity might lie.

Ethical dilemmas are prevalent in this work. Confidentiality, for example, can be significantly compromised in the context of an evacuation center, a recovery center, or working out of a hotel or other tourist locale offshore. Providing services to affected Australians in other countries where local people do not receive the same level of support is ethically challenging. FST staff have been creative in devising ways of assisting local people.

Conclusion

Social workers in leadership positions in key Commonwealth and State government agencies influenced and shaped the policy and service delivery response in a catastrophic disaster. The Australian system of having social workers in the primary national service delivery agency made the responses illustrated in the case studies possible. Social workers made a significant contribution to building resilience and reconnection not just at the individual level but also at the family and the wider community level. Emergency response and recovery are complex, and there will always be criticism and aspects which could be improved or more appropriately responsive. These have been highlighted in the inevitable reviews and Royal Commission postdisaster. In my view, the Australian people's trust in government at various levels during emergencies has been greatly assisted by their response to the offshore disasters and the fires.

The social work profession has much to offer in emergency response. Mathbor (2007) makes the point that more needs to be done to include emergency work in social work curricula. Social work needs to claim legitimacy in this field. It needs to move beyond the clinical and direct service intervention to think more about strategic positioning. The opportunity to influence and shape the response and recovery effort can be challenging when many key players involved come from very different perspectives. They may be focused on physical, economic, and infrastructure aspects of recovery and do not necessarily have visibility of the social, emotional, psychological, cultural, and spiritual aspects. While these aspects are often covered within community recovery planning at all levels of government, they can sometimes appear to take a lesser priority. Bringing a focus to these matters at the strategic and operational level is a critically important role for social work.

Bibliography

Alston, M., 2013. Gender and climate change in Australia and the Pacific. In: Alston, M., Whittenbury, K. (Eds.), Research, Action and Policy: Addressing the Gendered Impacts of Climate Change. Springer, Dordrecht, pp. 175–188.

Alston, M., Whittenbury, K. (Eds.), 2013. Research, Action and Policy: Addressing the Gendered Impacts of Climate Change. Springer, Dordrecht.

Brooks, J., Bodeau, D., Fedorowicsz, J., 2011. Network Management in Emergency Response: Articulating Practices of State Level Managers Interweaving Up, Down and Sideways. Administration and Society. Sage. www.mitre.org/work/tech_papers/2011/10_3770/ (accessed 20.05.13.).

Brun, C., Rapp, R.C., 2001. Strengths based case management. Individuals' perspective on strengths and case manager relationship. Social Work 46 (3), 278–288.

Chamberlain, E., Lievesley, S., 1984. Welfare policy and disaster. In: Report of Proceedings of a Research Workshop on Human Behaviour in Disaster in Australia. Australia Counter-Disaster College, Mt Macedon, pp. 151–174.

Child, C., Clay, D., Warrington, C., Das, J., 2008. Caring in a Crisis: The Contribution of Social Care to Emergency Response and Recovery. Tavistock Institute. www.scie.org.uk (accessed 18.08.13.).

Cronin, M.S., Ryan, D.M., Brier, D., 2007. Support for staff working in disaster situations: a social work perspective. International Social Work 50 (3), 370–382.

Dominelli, L., 2012. Green Social Work. Polity Press, Cambridge.

Emergency Management Australia, 1998. Australian Emergency Management Glossary. Emergency Management Australia, Canberra, pp. 22–23. www.ema.gov.au (accessed 15.04.13.).

Enarson, E., Meyreles, L., 2004. International perspectives on gender and disaster: differences and possibilities. International Journal of Sociology and Social Policy 24 (10), 49–93.

Federal Emergency Management Agency, 2012. Disaster Survivor Assistance – Disaster Declaration. Retrieved from: www.fema.gov/declaration-process.

Goldstein, H., 1983. Starting where the client is. Social Work 64 (5), 267–275.

Gordon, R., Wraith, R., 1993. Responses of children and adolescents to disaster. In: Wilson, J.P., Raphael, B. (Eds.), International Handbook of Traumatic Stress Syndromes. Plenum Press, New York, pp. 561–575.

Hobfoll, S.E., Watson, P., Bell, C.C., Bryant, R.A., Brymer, M.J., et al., 2007. Five essential elements of immediate and mid-term mass trauma intervention: empirical evidence. Psychiatry 70 (4), 283–315.

Iravani, M.R., Ghojavand, K., 2005. Social work skills in working with survivors of earthquake: a social work intervention – Iran. Social Work and Society 3 (2), 265–272.

Kaniasty, K., Norris, F.H., 2004. Social support in the aftermath of disasters, catastrophes, and acts of terrorism: altruistic, overwhelmed, uncertain, antagonistic and patriotic communities. In: Ursano, R., Norwood, A., Fullerton, C. (Eds.), Bioterrorism: Psychological and Public Health Interventions. Cambridge University Press, Cambridge, pp. 200–229.

Kreps, G.A., 1995. Disaster as systemic event and social catalyst: a clarification of subject matter. International Journal of Mass Emergencies and Disasters 13 (3), 255–284.

Manning, C., Millar, S., Newton, T., Webb, S., 2007. After the wave – the Centrelink social work response offshore. Journal of Social Work in Disability and Rehabilitation 5 (3&4), 81–95.

Mathbor, G.M., 2007. Enhancement of community preparedness for natural disasters: the role of social work in building social capital for sustainable disaster relief and management. International Social Work 50 (3), 357–369.

Mathbor, G.M., 2010. Surviving disaster: the role of invisible assets of communities. In: Gillespie, D., Danso, K. (Eds.), Disaster Concepts and Issues – a Guide for Social Work Education and Practice. CSWE Press, Virginia, pp. 145–161.

McFarlane, A.C., Williams, R., 2012. Mental Health Services Required after Disaster: Learning from the Lasting Effects of Disasters. Retrieved from: www.hindawi.com/journals/drt/2012/970194/.

Murphy, R., 2009. Leadership in a Disaster. Learning for a Future with Global Climate Change. McGill-Queen's University Press, Montreal.

Raphael, B., 1986. When Disaster Strikes: A Handbook for the Caring Professions. Basic Books, New York.

Raphael, B., Dunsmore, J., Wooding, S., 2005. Terror and trauma in Bali. Australia's mental health response. Journal of Aggression, Maltreatment and Trauma 9 (1–2), 245–256.

Richardson, G.E., 2002. The metatheory of resilience and resiliency. Journal of Clinical Psychology 58 (3), 307–321.

Rowlands, A., Tan, N.T., 2008. Social redevelopment following the Indian Ocean Tsunami: an international social work response through the FAST project. Social Development Issues 30 (1), 47–58.

Saleeby, D., 2006. The Strengths Perspective in Social Work Practice, fourth ed. Pearson Education Inc., Boston.

Schonfield, D.J., 2002. Almost one year later: looking back and looking ahead. Developmental and Behavioural Pediatrics 23 (4), 292–294.

Tan, N.T., 2013. Emergency management and social recovery in different contexts. Journal of Social Work in Disability and Rehabilitation 12 (1–2), 8–18.

Tan, N.T., Huang, Y., Wang, L., 2011. Disaster management in China and Taiwan: models, policies, programs for social recovery. Journal of Global Social Work 4 (1). Retrieved from: www.globalsocialwork.org/vol4no1/.

Tarrant, M., 2006. Risk and emergency management. The Australian Journal of Emergency Management 21 (1), 9–14.

Taubman, A., Weisz, A., 2011. Emerging concerns for international social work and disaster response: from relief to development and sustainability. Columbia Social Work Review 2, 37–48.

Tedeschi, R.G., Kilmer, R.P., 2005. Assessing strengths, resilience and growth to guide clinical interventions. Professional Psychology: Research and Practice 36 (3), 230–237.

Walsh, F., 2007. Traumatic loss and major disasters: strengthening family and community resilience. Family Process 46 (2), 207–227.

Zakour, M.J., 2005. Social Work and Disasters. Emergency Management Institute. www.training.fema.gov/emiweb/ (accessed 20.02.13.).

Relevant Websites

www.dhs.gov.au/ – Information on Evaluation of Case Management in 2009 Bushfires.

www.dse.vic.gov.au/fire_and_other_emergencies.

www.em.gov.au – Australian Government Emergency Management Website.

www.em.gov.au/library/ – Information on Bali Bombings.

www.fema.gov/ – US Government Emergency Management Website.

www.med.monash.edu.au/glass/ – Gender, Leadership, and Social Sustainability.

www.royalcommission.vic.gov.au/finaldocuments/ – Investigation into 2009 Bushfires.

www.smh.com.au/specials/Bali.

www.socialworkers.org.

Emergent Properties

Marlis Buchmann, University of Zurich, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 7, pp. 4424–4428, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article details varied concepts of emergent properties, including their relationship to systems theories. After reviewing the historical background of these concepts, the article focuses on the sociological tradition of emergence and then on more recent perspectives. Finally, methodological issues concerning emergent properties are discussed, such as the problem of their measurement.

The concept of emergent properties is the key element of a theoretical and methodological tradition claiming that the whole is more than the sum of its parts. This implies that the properties of a whole (systemic qualities) cannot be deduced by summing or averaging the properties of its components. Emergence theories are most elaborate in philosophical thought. However, an influential strand of twentieth-century social theory (Emile Durkheim, Talcott Parsons, Niklas Luhmann) borrowed some elements of this philosophical tradition and adopted the idea that the social (i.e., social organization at any level) should be regarded as an emergent property. By asserting that society is a reality of its own, nonreducible to the rules governing individual behavior, this sociological tradition stands in stark opposition to microreductionist thought. More recent approaches to the understanding of emergent properties (e.g., theories of self-organization and self-referentiality; chaos theory) are interdisciplinary in character, addressing research questions that allude to problems raised in classic emergence theory. However, the concept of emergent properties is used less as an analytical tool and more as a guiding research idea. In conclusion, methodological issues related to the measurement and the inference of emergent properties are discussed.

Emergence and emergent properties are concepts that are used across many scientific disciplines. They are the key elements of a theoretical and methodological tradition – most prominently represented in the 1920s – claiming that the whole is more than the sum of its parts. This amounts to saying that the properties of a whole (systemic qualities) cannot be deduced by summing or averaging the properties of its components (nondeducibility thesis). Vice versa, the appearance of genuinely novel properties characterizing compounds or systems as wholes (i.e., emergent properties) cannot be solely predicted from knowing the properties of the constituents and their relations. Emergence theories are most elaborate in philosophical thought. However, an influential strand of twentieth-century social theory (Emile Durkheim, Talcott Parsons, Niklas Luhmann) borrowed some elements of this philosophical tradition and adopted the idea that the social (i.e., social organization at any level) should be regarded as an emergent property. Arguing against the powerful approach of reductionism, this sociological tradition asserts that systemic properties of the social world cannot be deduced from properties of the constituents (i.e., individuals) and their respective interactions. They need to be accounted for at the system level. In the latter part of the twentieth century, the scientific discourse about emergence has regained prominence in many scientific disciplines – not to speak of a veritable renaissance. More recent approaches to the understanding of emergent properties include, among others, the theory of deterministic chaos, complex-system theory, and theories of self-organization and self-referentiality. Interdisciplinary in orientation, they focus on the complex dynamics between interactions and relations at different organizational levels (multilevel analysis) and stress the coexistence of system concepts and constituent concepts in explaining composition outcomes. After a brief history of the concept in philosophical thought, the major theoretical discourses about emergence in the context of social theory will be discussed. Next, there is a brief discussion of the recent interdisciplinary approaches that incorporate the idea of emergence. Finally, methodological issues related to the measurement and the inference of emergent properties are discussed.

The Philosophical Tradition of Emergence

The concepts of emergence and emergent properties have a long philosophical tradition (see Stephan (1999) and Beckermann et al. (1992) for extensive reviews). Unanimously, the origins of the idea of emergence are attributed to John Stuart Mill (1974). Considered to be the classic source, Mill distinguished between two types of joint effect produced by the interplay of various causes (composition of causes). In the first type, the joint effect of several component causes is identical with the sum of the separate effects. With respect to the second type, the joint effect is heterogeneous to the separate causes. The properties of a system or a compound generated by the second type of joint effect were later to be labeled emergent. Offering an alternative to mechanism and vitalism, the 1920s may be regarded as the heyday of the philosophical tradition that grew out of Mill's fundamental insight.

It was in this decade that those later to become *classic* authors – Samuel Alexander, Conwy Lloyd Morgan, Charles Broad, and Roy Sellars – published their influential works. According to Stephan (1999), 'weak' versions of emergentism may be distinguished from 'strong' ones. The former loosen some of the assumptions deemed to be essential for defining emergent properties in the latter versions. Both strands of

theory assert that emergent characteristics of wholes are not the sum of the properties of their parts (nonadditivity). They cannot be understood by simply summing (or averaging) the behaviors of the constituents; rather, they must be described in terms of their own. The novelty thesis claims that in evolution, genuinely novel (systemic) properties emerge. Examples are the advent of mind or reflective thought. In more abstract terms, this amounts to saying that emergent system properties are different in type from the properties of the system's constituent components. The two strands differ, however, with regard to the core assumptions of emergence theory – deducibility and predictability of systemic properties. Strong versions maintain that emergent properties are not deducible or predictable from the properties of the constituents and their relations (nonpredictability and nondeducibility). In this view, emergence may be characterized as the antithesis of microreductionism (Auyang, 1998). By contrast, in weak variants of emergence theory, all systemic properties that distinguish a system from the qualities characterizing its elements must be regarded as candidates for representing emergent properties. The late-1960s and early 1970s saw a renaissance of ideas of emergence. Centering on the fundamental issue of the mind–brain relationship – the psycho–physical problem – Bunge (1977) and Popper and Eccles (1977) offered theories strongly drawing on the idea of emergent properties.

Sociological Conceptualizations of Social Reality

From its inception, sociological thought has been characterized by two dramatically different conceptualizations of social reality. In particular, these differences pertain to the fundamental understanding of the relationship between the individual and social organization. Methodologically, this problem may be couched in terms of the relation between the micro- and macrolevels – in other words, the micro–macro link. One tradition, most commonly referred to as the individualistic–reductionist position, asserts that the properties of the social at any level may ultimately be explained by the interplay between the properties of individuals (e.g., motives, preferences), who are the constituents of society. Having strong roots in the Scottish moral philosophers (David Hume, Adam Smith, Adam Ferguson), the development of twentieth-century social theory has been characterized by various attempts to explain social phenomena by the rules governing the behavior of individuals. Basically, this approach asserts that social organization results from individual actions. Fundamentally opposed to this view is the other sociological tradition, often classified as the collectivist position. Its basic assumption is that the social world constitutes a reality of its own. Historically, this position is closely intertwined with the demands of sociology as a young scientific discipline to claim its territory and to define its own true subject matter vis-à-vis already well-established disciplines. By asserting that society is a reality of its own, nonreducible to the rules governing individual behavior, sociology distinguishes itself especially from economics and psychology. Searching for the particular subject matter of sociology, proponents of this position asked about the novel qualities and the specificity of the social vis-à-vis other types of reality, such as the mental realm, the biological or physical realm. It is with this latter tradition in sociology that the ideas of emergence and emergent properties have been most commonly associated. For this reason, the historical development of the latter tradition is discussed below.

The Sociological Tradition of Emergence: Society as an Emergent Property

In late-nineteenth and early twentieth-century sociology, the ideas of emergent properties are reflected in the conceptualization of society as representing a reality of its own endowed with separate characteristics nonreducible to the properties of its constituents (i.e., individuals). The sociological tradition that incorporates this assumption strongly draws on the thesis of emergence as nondeducibility (Bausch, 2001). In the early days of sociology, Emile Durkheim (1938) most prominently advocated this approach to the understanding of social phenomena. In Durkheim's view, society is not solely the sum of individuals; rather, it is the *association* of individuals. Hence, it is the *nature of this association* that constitutes the decisive factors of social life and determines the proper characteristics of social phenomena ('faits sociaux'). This theoretical stance was epitomized in the famous methodological rules introduced by Durkheim (1938) according to which the social realm needs to be explained by social phenomena.

In modern sociological thought, developed in the second half of the twentieth century, general systems theory has advanced to one of the most influential currents, implicitly or explicitly advocating ideas of emergence (Parsons, 1937, 1951; Luhmann, 1995). With the Durkheimian tradition, systems theory shares the fundamental methodological conviction of the nondeducibility of the social world. Social phenomena at different levels of social organization (i.e., society, organizations, social groups) show novel properties that cannot be explained by reducing them to the properties of and interactions between their constituents (i.e., behavioral properties of individuals). Hence, a genuine *social* theory is needed to account for social reality as an emergent property.

This line of thought, advocated by Durkheim in late-nineteenth-century sociology, was continued in the mid-twentieth century by Talcott Parsons, albeit with some major modifications (Parsons, 1951). For Parsons, the old problem raised by Hobbes about the origin of social order and social integration constitutes the fundamental sociological issue to which social theory needs to provide a satisfactory answer. The core of Parsons' answer to this fundamental problem is his thesis on the independence of institutionalized values and norms vis-à-vis the reality of individuals. The common system of ultimate values, constituting a reality *sui generis*, helps to establish order at various levels of social life and to prevent a free-for-all. In essence, Parsons declares common-value integration as the emergent property of social action systems. Parsons derives this conclusion from his systems-theoretical approach to the relationship between individual actors and social organization, whereby the starting point is the assumption that all reality realms constitute systems (i.e., the biophysical, personality, cultural, and social systems). Despite fundamental differences between various types of systems, they all share the system character as a common feature. By drawing a sharp distinction

between the social system (the social reality) and the personality system (the human being), Parsons maintains that each system needs to solve its own particular functional problem (i.e., system stability), for which each system relies on self-regulating mechanisms. In the case of the social system, social order and social integration are the functional problems, and processes of socialization and social control are the self-regulating mechanisms. Despite their own system logics, the two systems interact with each other in such a way that each system constitutes the environment for the other. The interactions between systems and their environment include the mutual exchange of resources, necessary for each system to keep up its functional capacity (i.e., system maintenance). Within this systems-theoretical framework, the social system and the personality system mutually presuppose each other; they are of equivalent status. This implies that there is no ontological priority of one system over the other. Hence, the properties of the social world cannot be deduced from the rules governing the personality system. The society is thus an emergent property.

Within the tradition of systems theory, Niklas Luhmann (1995) has provided the latest elaborate attempt to conceive of social reality as a separate, nonreducible system, reproducing itself solely on the basis of its own elements. Sharing some common ground with Parsons' theoretical approach, Luhmann designs a systems theory that makes use of fundamental ideas developed in molecular biology. In particular, Luhmann applies the concept of autopoiesis advocated by Maturana to social reality. Arguing at a very high level of abstraction, social systems are autopoietic systems in as much as they are self-reproduced by their own components, which are defined as acts of communication. This amounts to saying that social systems operate on the basis of communicative acts engendering subsequent acts of communication. Hence, the social system is an independent, self-reproductive whole. Like Parsons' theory, Luhmann's approach attributes great significance to the strict distinction between the social system and the psychological system. These two separate systems reciprocally constitute the environment of the other, for any system needs appropriate environmental relations for its self-reproduction. Given the system–environment relationship, the autopoiesis of the social system is not conceivable without the autopoiesis of the psychological system and vice versa. In other words, the two autopoietic systems enable each other to operate. From this it follows that the two systems are equivalent (and not hierarchically dependent), which precludes the reductionist assumption that systemic properties of one system can be deduced from those of any other system. Consequently, social systems are emergent properties.

The Renaissance of Ideas of Emergence in Late-Twentieth-Century Theoretical Thought

In the latter part of the twentieth century, ideas of emergence have again attracted considerable attention in various scientific fields. Several new approaches, such as synergetics (Haken, 1978) and other versions of self-organization theory (Krohn and Küppers, 1992; Krohn et al., 1990; Conte et al., 2001) or chaos theory (Newman, 1996), address interdisciplinary research questions at a relatively high level of abstraction that allude to problems raised in classical emergence theory. Examples of these research problems are the investigation of the spontaneous and self-confined emergence of new structures and systemic properties or the examination of predictability in chaotic dynamic processes. Interestingly enough, this body of research seldom refers to the tradition of emergence theory, and the notion of emergence itself remains rather vague. This suggests that, for these theoretical approaches, the concept of emergent properties is used less as an analytical research tool and more as a guiding research idea.

This is most apparent in the booming field of theories of self-organization. Krohn and Küppers (1992), for example, provide a rather vague concept of emergent properties. They define emergence as the sudden appearance of novel properties that can be explained neither by the properties themselves nor by the relations of the elements involved, but rather by specific self-organizing process dynamics. To complicate matters, the notion of self-organization is also a shifting concept. Depending upon the strand of system theory, the meaning attributed to self-organization turns out to be completely different. Searching for the least common denominator, self-organization may be understood as interactions between the components of a system and its organization that lead to modifications on both sides (Hejl, 1990: 115). It follows that the dynamics of self-organization engender new structures endowed with emergent systemic properties. In theories of self-organization then, self-referential organization of systems and emergent properties are causally related. Within this theoretical framework, social phenomena, such as social networks and social movements, easily lend themselves to the conceptualization of emergent properties generated by self-organizing process dynamics.

The issues addressed in theories of self-organization discussed above may be generalized to include the problem of how to conceive of the relationship between microinteractions and macrostructures – in short, the micro–macro link. With respect to social theory, the great bulk of previous work adheres to either a bottom-up method (microreductionism) or a top-down method (systems theory). Both methods, taken by themselves, have their limits. The reductionist tradition in social theory must at least acknowledge systemic properties as marginal conditions of individual behavior. Vice versa, the collectivist tradition is often short of explaining many systemic properties. Against this background, the need for a circular movement of theory building when conceptualizing the micro–macro link (i.e., interlevel explanations) is evident (Sawyer, 2003; Raub et al., 2011). Hence, approaches that attempt to couple systems concepts with constituent concepts are in demand. A theoretical framework of this type "simultaneously accounts for the behaviors of systems on one organizational level, the behaviors of their constituents on a lower level, and the relations between them" (Auyang, 1998: 5). In this respect, the approach of synthetic microanalysis proposed by Auyang (1998) may be regarded as one example.

Emergent Properties: Methodological Issues

Methodologically, the question arises of how to assess emergent properties. The key to answering this question lies in the

exploration of the link between the micro- and macrolevels of social phenomena. From this perspective, the claim that the whole is more than the sum of its parts amounts, first, to furnishing proof of the *specific assumptions* implied in either the 'strong' or 'weak' versions of emergence theory (see Section The Philosophical Tradition of Emergence). According to the 'strong' version, properties at the higher levels of aggregation (i.e., the macrolevel) qualify as emergent, if they *cannot* be deduced or predicted from the properties of units at lower levels (i.e., the microlevel). In the 'weak' version, they qualify, as long as they can be expressed as complex functions of the lower-level characteristics. Second, the assessment of emergent properties must be based on evidence regarding the *impact* of system properties (i.e., contextual properties) on lower-level units and/or relationships that operate independently of the effects of any lower-level characteristics. Couched in more technical terms, the second requirement implies that the attributes of units operating at the microlevel are a function of context (i.e., the macrolevel). The two tasks identified here to assess emergent properties address a measurement issue on the one hand and an issue about modeling techniques on the other. Each issue will be addressed in turn.

The Measurement Issue

'Strong' and 'weak' versions of emergence theory agree that higher-level properties that can be expressed as simple functions of lower-level attributes do not qualify as emergent. Examples include contextual characteristics defined as the sum, as a proportion, or as the mean of lower-level properties (e.g., percentage of community residents above the age of 65 years; average income of regions). The two strands of emergence theory differ, however, with regard to the theses of deducibility and predictability. The 'strong' version claims that emergent properties are *global* attributes that are not expressible as functions of lower-level properties. This implies that system properties need to be measured at the system level as opposed to the level of system elements. One example of such an approach would be the description of political systems characterized by the degree of corporatism, whereby the measurement of corporatism would be based on expert judgment.

The measurement issue involved in 'weak' versions of emergence theory includes the application of complex rules of aggregation. Expressed as complex functions of lower-level attributes, system properties can be deduced or predicted from lower-level units. The income distribution characteristic of regional contexts – expressed by a complex function of individual incomes – serves as an example. A methodological perspective that is well suited to identifying emergent properties in line with 'weak' versions of emergence theory is *network analysis*, most prominently advocated by scholars such as Ronald S. Burt, Mark Granovetter, and Barry Wellmann. This technique truly addresses the fundamental idea of the whole being more than the sum of its parts by focusing on *relational* characteristics and by introducing complex rules of aggregation. The social closure of elites, for example, cannot be inferred from the attributes of individual members of the elite. It is a property inferred from the *relational* attributes of the network members. Hence, the relational order of network elements (i.e., the structure) is a system property that truly qualifies as an emergent property. The second basic assumption of network analysis is the embeddedness of lower-level units in higher-level systems (i.e., networks) (see Granovetter, 1985). This implies that lower-level outcomes (e.g., preferences, tastes, action orientations, performance at the individual level) are mediated by higher-level system properties. This is to say that individual and collective actors respond to their social context. Couched in more technical terms, the notion of embeddedness addresses the issue of the additive and interactive effects of context on lower-level outcomes. This remark leads up to the second issue raised above, namely the modeling techniques required to assess emergent relationships.

The Impact Issue

To assess the effects of system properties measured at various levels of a hierarchical structure on lower-level outcomes, multilevel models (or contextual models) are the method of choice. A variety of models are used to specify the link between the macro- and microlevels of social phenomena. While the names of the models and their statistical properties do differ (e.g., hierarchical linear model, random coefficients model, hierarchical mixed linear model), the basic idea is "to analyze data that consist of multiple macro units (contexts) and multiple micro units within each macro unit" (DiPrete and Forristal, 1994: 334). Within these models, two types of effect can be distinguished: main or additive effects and interactive effects. Main effects of higher-level properties on lower-level outcomes (e.g., behavior) provide evidence that context matters over and above the impact of any lower-level attributes. This indicates that the relationships identified through the effect analysis cannot be reduced to individual differences. Examples of substantive research along these lines include the contextual effects of schools and classes on student achievement or the impact of macrocontexts, such as organizational structure and labor market segmentation, on career outcomes. The second type of effect involves interaction effects of properties measured at various levels of a hierarchical structure. The existence of cross-level interaction effects provides evidence of how higher-level properties (i.e., the macrocontext) affect the impact of lower-level attributes (i.e., those at the microlevel) on the outcome of interest. Variation in the strength of the effect of microlevel properties on the outcome when system properties are included indicates the presence of relationships that can be qualified as emergent. The reason is that interaction effects involving cross-level components show that they are more than the sum of their parts. Substantive research employing this complex modeling strategy includes, for example, the ways in which country characteristics (e.g., economic development; policy programs) affect the influence of maternal education on the reproductive behavior of women.

See also: Clustering and Cohesion in Networks: Concepts and Measures; Data Bases and Statistical Systems: Culture; Self-Organizing Dynamical Systems; Social Network Analysis.

Bibliography

Auyang, S.Y., 1998. Foundations of Complex-System Theories in Economics, Evolutionary Biology, and Statistical Physics. Cambridge University Press, Cambridge, UK.

Bausch, K.C., 2001. The Emerging Consensus in Social Systems Theory. Springer, New York.

Beckermann, A., Flohr, H., Kim, J. (Eds.), 1992. Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism. Walter de Gruyter, Berlin.

Bunge, M., 1977. Emergence and the mind. Neuroscience 2, 501–509.

Conte, R., Edmonds, B., Moss, S., Sawyer, R.K., 2001. Sociology and social theory in agent based social simulation: a symposium. Computational & Mathematical Organization Theory 7, 183–205.

DiPrete, T.A., Forristal, J.D., 1994. Multilevel models: methods and substance. Annual Review of Sociology 20, 331–357.

Durkheim, E., 1938[1895]. The Rules of Sociological Method. Free Press, New York [Les Règles de la Méthode Sociologique. Presses Universitaires de France, Paris].

Granovetter, M., 1985. Economic action and social structure: the problem of embeddedness. American Journal of Sociology 91, 481–510.

Haken, H., 1978. Synergetics, second ed. Springer, Heidelberg, Germany.

Hejl, P.M., 1990. Self-regulation in social systems. In: Krohn, W., Küppers, G., Nowotny, H. (Eds.), 1990 Self-organization: Portrait of a Scientific Revolution. Sociology of the Sciences, vol. XIV. Kluwer, Dordrecht, The Netherlands, pp. 269–292.

Krohn, W., Küppers, G. (Eds.), 1992. Emergenz: die Entstehung von Ordnung, Organisation und Bedeutung. Suhrkamp, Frankfurt am Main, Germany.

Krohn, W., Küppers, G., Nowotny, H., 1990. In: Self-organization: Portrait of a Scientific Revolution. Sociology of the Sciences, vol. XIV. Kluwer, Dordrecht, The Netherlands.

Luhmann, N., 1995[1984]. Social Systems. Stanford University Press, Stanford, CA [Soziale Systeme: Grundriss einer allgemeinen Theorie. Suhrkamp, Frankfurt am Main].

Mill, J.S., 1974[1843]. In: A System of Logic: Ratiocinative and Inductive. Collected Works, vols. VII and VIII. University of Toronto Press, Toronto, ON.

Newman, D.V., 1996. Emergence and strange attractors. Philosophy of Science 63, 245–261.

Parsons, T., 1937. The Structure of Social Action. McGraw-Hill, New York.

Parsons, T., 1951. The Social System. Free Press, New York.

Popper, K.R., Eccles, J.C., 1977. The Self and Its Brain. Springer, Heidelberg, Germany.

Raub, W., Buskens, V., van Assen, M.A.L.M., 2011. Micro-macro links and micro-foundations in sociology. Journal of Mathematical Sociology 35, 1–25.

Sawyer, R.K., 2003. Artificial societies: multiagent systems and the micro-macro link in sociological theory. Sociological Methods & Research 31, 325–363.

Stephan, A., 1999. Emergenz: von der Unvorhersagbarkeit zur Selbstorganisation. Dresden University Press, Dresden, Germany.

Emotional Development, Effects of Parenting and Family Structure on

Suzanne Bester and Marlize Malan-Van Rooyen, University of Pretoria, Pretoria, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article's primary objective is to review the literature on the effects of parenting and family structure on the emotional development of children. Parenting styles play an important role in how children are socialized into society and it determines children's emotional, social, cognitive, and behavioral development. Parents are the most significant role players who determine the family environment which will directly influence their children's emotional development. In the twenty-first century there are many definitions of what constitutes a family. In this article, we will review traditional, divorced, single, step, adoptive, extended, and gay and lesbian family structures and how each one of these very different family structures influence children's emotional development. Specific attention will be given to risk and resilience factors which give each one of the aforementioned family structures a unique identity.

Introduction

Having children and raising them is part of our existence as human beings. Apart from ensuring the continuation of the human race, the act of becoming a parent and raising a child also has existential significance for many people. The profound experience of creating a life and taking on the responsibility of caring for and nurturing a child who will one day become living proof of their legacy motivates and drives many parents to commit fully to parenting their offspring. Parenting takes place in the context of a specific family structure where children are helped to achieve important social skills and developmental goals. In this context, parents play a key role in raising children to become emotionally competent and functional members of society. The individual characteristics of the particular child, parents, family, and society culminate in a context-specific, unique, dynamic, and varying experience for each parent and child.

Parenting and the Emotional Development of Children

Meaningful relationships are important in the parenting process as they help children develop emotionally and cope with problems. A parent–child relationship characterized by engagement, warmth, security, consistency, and appropriate parental responses is key to a child's emotional development (Walsh, 2012; Vélez et al., 2011; Brown, 2010; Stack et al., 2010; Fiese and Winter, 2009).

The parents' relationship often affects their parenting abilities. Parents who are aware of one another's needs, who support one another, and who display fondness and mutual respect are likely to handle the demands of parenthood with relative ease. Conversely, conflict between parents is likely to affect the parent–child relationship adversely often leading to emotional problems in the child. Ideally, parents should share responsibilities in a supportive and complementary manner (Driver et al., 2012; Waldfogel et al., 2010; Fiese and Winter, 2009; Schwartz and Finley, 2009).

By structuring a child's environment, a parent protects the child from negative environmental influences and creates positive influences. Parental knowledge is parents' knowledge of their child; knowledge of where their child is; knowledge of what their child is doing, and knowledge of how their child is coping with life (Padilla-Walker et al., 2011; Neumann et al., 2010).

Parental knowledge is considered one of the strongest protective factors against negative environmental influences. Sources of parental knowledge include child disclosure (children informing parents), parental socialization (parents questioning their child and their child's friends), and parental control (controlling where their child goes and what he (The feminine and masculine pronouns (she, he, her, him, hers, his) should be regarded as interchangeable in this article.) does) (Racz and McMahon, 2011; Padilla-Walker et al., 2011).

A warm and accepting parent–child relationship contributes to higher levels of parental control because it increases access to parental knowledge resources such as child disclosure and parental solicitation. Lower levels of parental knowledge can often be linked to inadequate family processes (Padilla-Walker et al., 2011; Bumpus and Rodgers, 2009).

A structured environment promotes feelings of self-efficacy in children and the ability to control their surroundings. Risk factors for children include economic instability and inconsistent parenting. Routines that foster emotionally and physically safer environments promote child well-being (Brown, 2010; Fiese and Winter, 2009). Abbott (2011) maintains that positive family structures can promote healthy child development. Protection from harmful environmental factors includes exposing the child to a nurturing, safe, and stable environment.

A safe and stable environment has social and physical components and is created by responding to the child's physical, developmental, and emotional needs. The family's role is to protect the child from physical harm and to create a stable environment characterized by consistency and predictability (Mercy and Saul, 2009).

Negative transitions or events (such as divorce, illness, or death) can undermine the process of providing a child with a nurturing, safe, and stable environment. Families need to be able to adapt to these transitions by making optimal use of resources such as positive relationships (i.e., positive parent–child and parental relationships); awareness of the child's interaction with the environment (i.e., parental

knowledge); and structuring of the family environment (i.e., economic resources, social resources, and routines) (Walsh, 2012).

The importance of the continuity of parental influence is self-evident, yet other factors and variables also contribute to the child's functioning. A bioecological systems perspective (Bronfenbrenner, 1979) holds that the individual characteristics of a child and her parents (e.g., temperament, cognition, biology) and the broader societal influences (e.g., culture, community), all contribute to a holistic picture of the child as an individual. Cummings et al. (2000) stress the importance of an integrative biopsychosocial approach in understanding parenting.

Brooks (2008a) identifies three major influences on the parenting process: the child's personal characteristics; the parents' personal history, psychological characteristics, and resources; and the social context of stressors and supports.

Parenting Styles and the Emotional Development of Children

Parenting styles refers to the 'how' of parenting, that is, how parents interact, discipline, communicate, and respond to the behavior of a child while socializing the child into their group. Baumrind (1991) originally identified two main parenting dimensions, namely acceptance/responsiveness and demandingness/control. These two dimensions are still the subject of much discussion (Rinaldi and Howe, 2012).

Berns (2007c) maintains that parents who are accepting of and responsive to their children tend to give them affection, provide encouragement, and show sensitivity to their needs. Conversely, parents who are unaccepting of and unresponsive to their children tend to reject them, to criticize them, and to be insensitive to their needs. Parents who are demanding and controlling tend to set rules and monitor their children's behavior, while parents who are undemanding and controlling tend to allow their children more autonomy and to make fewer demands on them. Uninvolved parents are neither responsive nor demanding. Four parenting styles emerge from these two dimensions: authoritative, authoritarian, permissive, and uninvolved.

Authoritative parents know how to place demands on their children and how to maintain control that is appropriate for their children's age. They are affectionate, communicate clearly, and respond appropriately to the needs of their children. They know how to balance demandingness and responsiveness. They also set clear standards for and monitor their children's behavior effectively (Alegre, 2011).

The authoritative parent–child relationship is characterized by warmth and monitoring resulting in generally higher levels of self-competence, self-esteem, social competence, and academic success in their children. It is also reported that lower levels of internalizing (e.g., depression, withdrawal, anxiety, somatization) and externalizing behaviors (e.g., aggression, attention problems, hyperactivity) occur in these children. Warm and supportive parents are regarded as good resources for their children that lead to positive developmental outcomes (Camp, 2012; Rinaldi and Howe, 2012).

Parent–child relationships characterized by warmth are also related to higher levels of emotional sensitivity, perspective taking (i.e., being aware of and understanding other people's situations), and prosocial behaviors (i.e., sharing and donating resources, comforting others, volunteering for charitable activities, and helping the needy) (Carlo et al., 2010).

Children of authoritative parents tend to outscore children from the other three parenting styles on measures of adjustment, attachment, resilience, school achievement, social and school competence, and prosocial behavior (Alegre, 2012).

Authoritarian parents expect their children to be obedient and not to question or challenge their authority. They are status driven, demanding, and often unresponsive to the needs of their children. Communication between these parents and their children is generally limited (Camp, 2012).

Authoritarian parents are strict and controlling. This, together with the high demands they place on their children, may result in behavior outcomes characterized by high levels of anger and self-centeredness. Disciplinary practices such as verbal scolding and physical punishment often also lead to aggressive tendencies and antisocial behavior (Camp, 2012).

Parental demandingness is a complex phenomenon marked by positive as well as negative responsiveness. Positive responsiveness includes monitoring and supervision, control, autonomy granting, making/having age appropriate demands, and expectations, and applying inductive discipline. This is often associated with higher academic functioning, life satisfaction, confidence, and prosocial behavior with less risk-taking behavior. Negative responsiveness includes psychological control and inconsistent, punitive and harsh discipline. The result is often higher internalizing and externalizing behaviors and cognitive anxiety with also a higher risk of personality disorders (Alegre, 2011).

Permissive and neglectful parents make few demands on and exercise little control over their children. Permissive parents are often responsive to and affectionate toward their children while neglectful or uninvolved parents often show little affection toward and rarely communicate with their children (Alegre, 2011).

The behavior outcomes of children with permissive parents seem to correlate strongly with poor self-reliance and self-control with a greater tendency toward impulsive and aggressive behavior and poor exploration behavior. The children of uninvolved parents tend to experience deficits in attachment, cognition, and emotional and social skills. They also often exhibit poor self-control and low self-esteem (Berns, 2007a).

Effect of Family Structure on Emotional Development

Traditional Families

Like other families, traditional families face risk factors that can compromise their ability to provide their children with a nurturing, safe, and secure environment. Most of these risk factors are also found in other family structures. These risk factors include difficulty in integrating work and family life, marital problems and conflict associated with preconceived ideas about gender role responsibilities (Fraenkel and Capstick, 2012; Brown, 2010; Kelly, 2010; Perrone et al., 2009).

A healthy marital relationship increases parents' well-being, confidence, and parenting skills. Marital problems can have

a negative effect on family functioning and specifically on children's social, emotional, and behavioral health. Marital problems are characterized by distressed spouses, negative and criticizing conversations, and disappointment in the relationship itself (Driver et al., 2012; Mannering et al., 2011; Brooks, 2008b). Children who are exposed to high conflict environments often present with emotional conditions such as depression, anxiety, and social withdrawal (Driver et al., 2012; Fraenkel and Capstick, 2012; Brooks, 2008b).

A common struggle in many families is integrating work and family life. Spending too much time and energy at work compromises a parents' ability to build quality familial relationships. Struggling to balance work and family life is also associated with higher levels of interpersonal conflict as well as lower levels of parental knowledge and monitoring (Fraenkel and Capstick, 2012).

Changes in gender role responsibilities have generally undermined parents' ability to balance family and work life. Traditionally, a women's role was being a full-time mother rather than a breadwinner. Changes in gender role responsibilities are linked to changes in work and family roles that can be beneficial in terms of financial resources but taxing in terms of role strain, which often leads to marital conflict (Perrone, 2009).

Various protective factors associated with traditional nuclear families can, however, help increase family resilience. When two parents are present in a family, certain requirements for improved child development can be met more easily such as more monitoring, more money, two role models, and a larger support network. Identity struggles are easier to avoid, and children have a wider variety of role models. If the parents are happily married, the children can generally experience a safer and more stable environment. Children in traditional families often have better education, employment, childbearing, and psychological outcomes making this still the optimal environment for the raising of children (Blau and Van der Klaauw, 2013; Abbott, 2011).

Divorced Families

The disruption and stress that accompanies unstable family processes before, during, and after divorce can lead to various behavioral, emotional, interpersonal, and academic problems in children. Stressors associated with divorce can compromise parenting abilities and often lead to irritability and impatience with the children (Amato, 2010; Berns, 2007b; Brooks, 2008b).

Risk factors that have an influence on the parent–child relationship during and after divorce include visitation arrangements; possible loss of a familiar neighborhood, school, and friends; reduced income; absence of a coparent; emotional stress and continuing parental conflict; as well as the possibility that the parents may become less emotionally available to the children during this process (Amato, 2010; Waldfogel et al., 2010; Schwartz and Finley, 2009; Brooks, 2008b).

Protective resources include effectively reorganizing the family structure by distributing responsibilities evenly, negotiating a workable relationship with the ex-spouse, and strengthening existing networks for support. If the parent–child relationship remains healthy, the negative impact of the divorce on the children's development will be ameliorated (Vélez et al., 2011; Amato, 2010; Odom et al., 2010).

Single Parent Families

Single parenthood can come about in various ways including the death of a parent, nonmarital births, and adoption (Anderson, 2012). Divorce is also considered a pathway to single parenthood, but since the existence of an ex-spouse often implies coparenting, it will not be discussed in this section.

Risk factors associated with single parent families include lower income, less time for the children, lower education, poorer neighborhoods, younger parents, less emotional support, and higher levels of child victimization, especially when the parent has a nonfamily cohabiting partner. In instances where the abovementioned risks are overwhelming, the children tend to present with emotional problems. These families often have fewer resources to meet the children's needs and to provide a safe, secure, and nurturing environment (Anderson, 2012; Falana et al., 2012; Huston and Bentley, 2010; Waldfogel et al., 2010; Turner et al., 2007).

The absence of one parent is often associated with less economic and socioemotional support. The economic hardship linked with single parenting can often impair parenting skills. Single parents frequently have either to work long hours or take on more than one job to provide additional income – this leads to less time investment in the parent–child relationship and accordingly affects the children's emotional well-being. Fulfilling the roles of two parents can result in role overload thereby impairing a single parent's parenting abilities and lowering her parental knowledge and monitoring (Brown, 2010; Padilla-Walker et al., 2011; Bumpus and Rodgers, 2009; Berns, 2007b).

Lower levels of parental knowledge could be attributable to less child disclosure and fewer solicitations. Parental knowledge, as mentioned earlier, can serve as protection against negative environmental factors. The provision of protection against such factors is a key role of parents (Bumpus and Rodgers, 2009).

Instability tends to be a characteristic of single parent families as such families often undergo more transitions than married parent families. Instability is manifested in frequent changes in the family structure (e.g., when a single mother's boyfriend lives with the family), and more frequent moving around can undermine existing social networks (Brown, 2010; Waldfogel et al., 2010; Turner et al., 2007).

Protective factors that can promote the resilience of a single parent mother include an involved father, an extended family, good socioeconomic status, psychological strength, and cultural support. A single parent can also extend her social network by joining community groups and faith congregations (Anderson, 2012; Walsh, 2012).

Stepfamilies/Remarried Families

Stepfamilies face numerous challenges inherent in adjusting to a new family life, which often leads to higher levels of conflict and instability. Challenges faced by stepfamilies generally revolve around complex familial relationships (Brown, 2010).

Ambiguous relationships, which are characteristic of stepfamilies, contribute to the complexity of such families making it difficult to foster sound relationships. The complexity of such relationships lies in redefining roles, questioning stepparents' authority, managing lower quality relationships, and increased levels of interpersonal conflict and deciding on the custodial rights of nonresident parents (Pasley and Garneau, 2012; Schrodt, 2011; Brown, 2010; Bumpus and Rodgers, 2009; Turner et al., 2007; Sweeney, 2010).

Research evidence indicates that the relationship between a stepparent and a child is more challenging than one between a biological parent and a child. Children often view the stepparent as an outsider leading to boundary ambiguity. The relationship between a child and a nonresident parent also becomes more complex after remarriage. The existence of an ex-spouse further complicates the relationship between the married couple. Negotiating parental roles and power and relational boundaries can lead to tension that can affect the quality of the couple's relationship. Endorsement of autonomy is a common problem, especially in respect of childrearing (Pasley and Garneau, 2012; Schrodt, 2011; Sweeney, 2010).

Stepfamilies often display lower parental knowledge than intact biological families. Parental knowledge, as mentioned earlier, is crucial in protecting children from negative environmental influences. Lower levels of parental knowledge can be attributed to lower level parent–child relationships and, accordingly, lower levels of solicitation of information (Racz and McMahon, 2011; Bumpus and Rodgers, 2009).

Despite the challenges facing stepfamilies, the existence of protective factors can help foster a more safe, stable, and nurturing environment. These factors include open communication, flexibility, realistic expectations, and being in tune with one another's experiences. Supportive coparenting between stepparents and nonresidential ex-spouses can help minimize the complexity of relationships and associated conflict. Effective readjustment of roles, especially in respect of childcare responsibilities, can enable a mother to spend more time with her children and thereby provide them with a further positive adult role model (Pasley and Garneau, 2012; Schrodt, 2011; Sweeney, 2010).

Extended Family – Kinship Care

Extended families consist of several generations of people and can include biological parents and their children as well as in-laws, grandparents, aunts, uncles, and cousins. Extended families are typical of collective cultures where all family members are interdependent and share family responsibilities including childrearing roles (Waites, 2009; Strong et al., 2008).

Extended family members usually live in the same residence where they pool resources and undertake familial responsibilities. Multigenerational bonds and greater resources increase the extended family's resiliency and ability to provide for the children's needs, yet several risk factors associated with extended families can decrease their well-being. Such risk factors include complex relationships, conflicting loyalties, and generational conflict (Engstrom, 2012; Waites, 2009).

Complex intergenerational relationships can complicate the child–parent relationship as they can cause confusion regarding the identity of the primary parent. Such confusion can result in a child undermining the authority of her existing parent (Anderson, 2012) and feeling uncertain about her environment.

Extended families often value the wider kin group more than individual relationships, which can lead to loyalty issues within the family and also cause difficulties in a couple's relationship where a close relationship between a husband and wife may be seen as a threat to the wider kin group. Another factor that can add to the complexity of relationships in an extended family is the need to negotiate the expectations and needs of each family member. Complex extended family relationships can also detract from the parent–child relationship (Strong et al., 2008; Langer and Ribarich, 2007).

The literature points to various protective factors associated with extended families that can help the parents and family meet the children's various needs. Extended families usually have more resources at their disposal that can be used to ensure the well-being of the children. Also, when the family functions as a collaborative team, has strong kinship bonds, is flexible in its roles, and relies on cultural values to sustain the family, the family itself serves as a lifelong buffer against stressful transitions (Engstrom, 2012; Waites, 2009).

Kinship care as a cultural value in extended families is associated with positive child outcomes, yet this may not be the case when such families have to take responsibility for a child because his parents are unable to do so. In such cases, kinship care becomes similar to foster care. Situations like the latter usually arise from substance abuse, incarceration, abuse, homelessness, family violence, illness, death, or military deployment (Langosch, 2012).

Although children in kinship care often fare better than children in foster care, various risk factors can have a negative impact on the children's well-being. Risk factors include low socioeconomic status, inability to meet children's needs properly, unhealthy family dynamics, older kin, less-educated kin, and single kin (Langosch, 2012; Palacios and Jiménez, 2009; Harris and Skyles, 2008; Metzger, 2008; Winokur et al., 2008).

Kinship care as foster care is often characterized by complex relationships and the trauma caused by the loss of an able parent. The family member who assumes the role as parent often finds it difficult to balance his former relationship with his new role as the person responsible for the child's well-being. For instance, a grandmother may have to adapt to the idea of being a strict parent instead of a loving, indulgent grandmother (Engstrom, 2012; Langosch, 2012).

The extended family member who steps into the parenting role is often overwhelmed by the stress caused by new parental responsibilities, attachment difficulties, and possible feelings of resentment and anger toward the biological parent, as well as having to deal with traumatic transitions after the loss of an able parent. The relationship between the new parent and other family members may also experience strain due to loyalty issues. Besides complex relationships, changes in the child's environment call for new routines, the setting of new limits, and sometimes coparenting with the biological parent, all of which can contribute to a less stable environment (Engstrom, 2012; Langosch, 2012).

An extended family member who takes on kinship care faces many challenges, although positive experiences associated with such care can also serve as a protective factor buffering the child against the negative effect of traumatic transitions. The new

parent may find this transition meaningful in the sense that it adds purpose to her life, and the child may also experience a sense of security, consistency, continuity in family identity, emotional ties, and familiarity (Langosch, 2012; Harris and Skyles, 2008; Metzger, 2008).

Adoptive Families

Adoptive families face all the challenges of contemporary nuclear families and more. Attachment difficulties are common among adoptive families and become even more challenging in cases where the adopted child was exposed to abuse or neglect. Adoption implies that the child will lose existing social and family structures, and frequent changes in caregivers and the loss of existing social networks can increase stress and anxiety in adopted children, making their environment seem unstable and unsafe (Carnes-Holt, 2012; Rampage et al., 2012).

Developing a safe and trusting relationship between the parents and an adopted child is challenging. Adoptive parents are often concerned about the quality of attachment that will develop between them and the adopted child. When children are adopted from institutions, such as orphanages, there is an increased possibility of disturbances in interpersonal relationships. Children who were exposed to abuse/neglect or anything that compromised initial attachment may need additional attention and care to foster sound attachment between them and the new parents. The challenges adoptive families face are often influenced by characteristics of the child such as age, race, and country of origin (Carnes-Holt, 2012; Garvin et al., 2012; Rampage et al., 2012; Tirella et al., 2012).

Adopted children whose race differs from that of their new parents may experience difficulties with their sense of ethnic identity. Parents who adopt from outside their own country may have limited access to information such as medical histories, prenatal care, and early living conditions. Racial and cultural differences may also compromise the parents' ability to provide relevant care, for instance, parents from a Western culture may promote independence while their child may be used to a culture that promotes interdependence. Children adopted within the first 12 months of their lives seem to have attachment relationships similar to those of nonadopted children, but older children often present more of a challenge (Carnes-Holt, 2012; Rampage et al., 2012; Docan-Morgan, 2010).

Protective factors that help adoptive parents build good attachments include strong parental relationships, thorough parental preparation, sharing responsibilities, and having a good support network. Structuring the environment on the basis of routine can help adopted children feel safe thereby also promoting attachment. Lastly, parents who encourage self-disclosure and provide emotional support facilitate the transition for their adopted children (Tirella et al., 2012; Docan-Morgan, 2010).

Lesbian and Gay Families

Lesbian and gay couples have various pathways to parenthood including having children from a former heterosexual relationship or marriage, adoption or conception through different reproductive technologies. Like heterosexual parents, lesbian and gay parents choose parenthood because of their desire to nurture a child, to be parents and to enjoy the company of a child. Lesbian and gay parents also face the same challenges as heterosexual parents such as balancing work and family life and adapting to parenthood. Challenges faced by lesbian and gay families generally stem from social prejudice, which can lead to victimization of the parents and their child (Green, 2012; Lee, 2009).

Within the parent–child relationship, the parents may need to pay special attention to issues such as social judgment, prejudice, and rejection that may affect the child. Social prejudice can undermine the child's identity since the child may not be regarded as an individual but simply as the "child of a lesbian or gay couple" (Rabitaille and Saint-Jacques, 2009: p. 436). This may even influence the child's interpersonal relationships as certain heterosexual parents may forbid their own children to play with the child. Children of lesbian and gay families consequently sometimes feel hesitant to share information about the nature of their family structure. Lesbian and gay parents should carefully explain this structure to their child within the safety of the parent–child relationship (Green, 2012; Lee, 2009).

Lesbian and gay couples also often face legal and social discrimination that may complicate their relationship, which is often 'sexualized' by society, meaning that society sees their relationship as exclusively sexual. Other aspects of the relationship such as love and commitment are often overlooked. Some lesbian and gay parents report downplaying their relationship in front of other role players in their child's life in an attempt to 'normalize' their parenting. Other relationship stressors include the logistics of starting a family, unsupportive laws, and the possibility of losing family and community support. Exposure to these stressors often impacts parenting skills negatively thereby also affecting the child (Lee, 2009; Wall, 2011).

Lesbian and gay parents are just as capable as their heterosexual counterparts of providing nurturing, safe, and stable family environments. However, they do face sexual prejudice that can turn judges, legislators, professionals, and the public against them often resulting in loss of physical custody, restrictions on visitations, and prohibitions against adoptions. Several countries and states have granted same-sex couples' rights, which include lesbian and gay marriages as well as adoption. Homophobia and discrimination, however, still affect lesbian and gay families. For instance, research reveals that these families still face discrimination when taking their children to healthcare services where they often experience parental exclusion and other difficulties with the health system due to their particular family structure (Amato, 2012; Chapman et al., 2012; Marks, 2012; Biblarz and Savci, 2010).

Despite the challenges faced by lesbian and gay families, various factors associated with lesbian and gay couples often act as buffers against stressful influences. Lesbian and gay couples tend, for example, to share domestic chores, decision-making, and parenting more equally than do heterosexual couples. Because of the shared roles, lesbian and gay couples often report higher satisfaction with their relationship and less socioeconomic stress. Additional social and community support also serve as valuable protective resources (Green, 2012; Berns, 2007b; Brooks, 2008b).

Conclusion

Much has been written about parenting and the role of the parent in the child's life, and much more will probably be written in the future. Parenting a child is a dynamic and ongoing process that shifts and changes together with changes in families across generations and changes in the environment and the broader society. We can never hope to understand fully the complexity of each individual family, but we can endeavor to peel away, discover, appreciate, and understand the multiple layers that constitute family life.

See also: Attachment During Infancy and Early Childhood: Cultural Variations, Development of; Behavioral Problems, Effects of Parenting and Family Structure on; Early Emotional Development and Cultural Variability; Empathy During Early Childhood Across Cultures, Development of; Life Management, Developmental Psychology of; Parenting Attitudes and Beliefs across Cultures; Personality Development: Systems Theories; Prosocial Behavior During Infancy and Early Childhood: Developmental Patterns and Cultural Variations; Self and Identity Development During Adolescence across Cultures; Self in Culture: Early Development; Self-Development in Childhood and Adolescence; Self-Regulation During Early Childhood Across Cultures, Development of; Social Referencing During Infancy and Early Childhood across Cultures; Social and Emotional Development in the Context of the Family; Stress in Childhood: Effects on Development; Stress in Infancy and Early Childhood: Effects on Development; Temperament and Human Development.

Bibliography

Abbott, D.A., 2011. Is there an optimal environment for child development? Ave Maria Law Review 10 (1), 121–146.

Alegre, A., 2011. Parenting styles and children's emotional intelligence: what do we know? The Family Journal 19 (1), 56–62.

Alegre, A., 2012. Is there a relationship between mothers' parenting styles and children's trait emotional intelligence? Electronic Journal of Research in Educational Psychology 10 (1), 1696–2095.

Amato, P.R., 2010. Research on divorce: continuing trends and new developments. Journal of Marriage and Family 72, 650–666.

Amato, P.R., 2012. The well-being of children with gay and lesbian parents. Social Science Research 41, 771–774.

Anderson, C.M., 2012. The diversity, strengths, and challenges of single-parent households. In: Walsh, F. (Ed.), Normal Family Processes: Growing Diversity and Complexity, fourth ed. The Guilford Press, New York, pp. 128–148.

Baumrind, D., 1991. The influence of parenting style on adolescent competence and substance use. Journal of Early Adolescence 11 (1), 56–95.

Berns, R.M., 2007a. Ecology of socialization. In: Berns, R.M. (Ed.), Child, Family, School, Community: Socialization and Support, seventh ed. Thomson Wadsworth, USA, pp. 35–74.

Berns, R.M., 2007b. Ecology of the family. In: Berns, R.M. (Ed.), Child, Family, School, Community: Socialization and Support, seventh ed. Thomson Wadsworth, USA, pp. 75–116.

Berns, R.M., 2007c. Ecology of parenting. In: Berns, R.M. (Ed.), Child, Family, School, Community: Socialization and Support, seventh ed. Thomson Wadsworth, USA, pp. 117–158.

Biblarz, T.J., Savci, E., 2010. Lesbian, gay, bisexual, and transgender families. Journal of Marriage & Family 72, 480–497.

Blau, D.M., Van der Klaauw, W., 2013. What determines family structure? Economic Inquiry 51 (1), 579–604.

Bronfenbrenner, U., 1979. The Ecology of Human Development: Experiments by Nature and Design. Harvard University Press, Cambridge, Massachusetts, pp. 1–44.

Brooks, J.B., 2008a. Parenting is a process. In: Brooks, J.B. (Ed.), The Process of Parenting, seventh ed. McGraw Hill, New York, pp. 3–34.

Brooks, J.B., 2008b. Parenting in diverse family structures. In: Brooks, J.B. (Ed.), The Process of Parenting, seventh ed. McGraw Hill, New York, pp. 409–443.

Brown, S.L., 2010. Marriage and child wellbeing: research and policy perspectives. Journal of Marriage and Family 72 (5), 1059–1077.

Bumpus, M.F., Rodgers, K.B., 2009. Parental knowledge and its sources: examining the moderating roles of family structures and race. Journal of Family Issues 30 (10), 1356–1378.

Camp, N.H., 2012. The Influence of Parental Involvement in Single-father, Single-mother and Heterosexual Married Two-parent Family Systems on Adolescent Internalizing and Externalizing Behaviors. Counseling Psychology. Dissertations. Paper 30.

Carlo, G., Mestre, M.V., Samper, P., Tur, A., Armenta, B.E., 2010. The longitudinal relations among dimensions of parenting styles, sympathy, prosocial moral reasoning, and prosocial behaviors. International Journal of Behavioral Development 35 (2), 116–124.

Carnes-Holt, K., 2012. Child-parent relationship therapy for adoptive families. The Family Journal: Counseling and Therapy for Couples and Families 20 (4), 419–426.

Chapman, R., Wardrop, J., Freeman, P., Zappia, T., Watkins, R., Shields, L., 2012. A descriptive study of the experiences of lesbian, gay and transgender parents accessing health services for their children. Journal of Clinical Nursing 21 (7–8), 1128–1135.

Cummings, E.M., Davies, P.T., Campbell, S.B., 2000. Complex patterns of influence: risk and protective factors. In: Cummings, E.M., Davies, P.T., Campbell, S.B. (Eds.), Developmental Psychopathology and Family Process: Theory, Research, and Clinical Implications. The Guilford Press, New York, pp. 124–151.

Driver, J., Tabares, A., Shapiro, A.F., Gottman, J.M., 2012. Couple interaction in happy/unhappy marriages: Gottman laboratory studies. In: Walsh, F. (Ed.), Normal Family Processes: Growing Diversity and Complexity, fourth ed. The Guilford Press, New York, pp. 57–77.

Docan-Morgan, S., 2010. "They don't know what it's like to be in my shoes": topic avoidance about race in transracially adoptive families. Journal of Social and Personal Relationships 28 (3), 336–355.

Engstrom, M., 2012. Family processes in kinship care. In: Walsh, F. (Ed.), Normal Family Processes: Growing Diversity and Complexity, fourth ed. The Guilford Press, New York, pp. 196–221.

Falana, B.A., Bada, F.O., Ayodele, C.J., 2012. Single-parent family structure, psychological, social and cognitive development of children in Ekiti state. Journal of Educational and Developmental Psychology 2 (2), 158–164.

Fiese, B.H., Winter, M.A., 2009. Family influences. In: Benson, J.B., Haith, M.M. (Eds.), Social and Emotional Development in Infancy and Early Childhood. Elsevier Inc., Oxford, pp. 139–147.

Fraenkel, P., Capstick, C., 2012. Contemporary two-parent families: navigating work and family challenges. In: Walsh, F. (Ed.), Normal Family Processes: Growing Diversity and Complexity, fourth ed. The Guilford Press, New York, pp. 78–101.

Garvin, M.C., Tarullo, A.R., Van Ryzin, M., Gunnar, M.R., 2012. Postadoption parenting and socioemotional development in postinstitutionalized children. Development and Psychopathology 24, 35–48.

Green, R., 2012. Gay and lesbian family life: risk, resilience, and rising expectations. In: Walsh, F. (Ed.), Normal Family Processes: Growing Diversity and Complexity, fourth ed. The Guilford Press, New York, pp. 172–195.

Harris, M.S., Skyles, A., 2008. Kinship care for African American children: disproportionate and disadvantageous. Journal of Family Issues 29 (8), 1013–1030.

Huston, A.C., Bentley, A.C., 2010. Human development and societal context. Annual Review of Psychology 61, 411–437.

Kelly, J., 2010. Challenges to the traditional nuclear family. Social & Political Review xx, 11–17.

Langer, N., Ribarich, M., 2007. Aunts, uncles-nieces, nephews: kinship relations over the lifespan. Educational Gerontology 33 (1), 75–83.

Langosch, D., 2012. Grandparents parenting again: challenges, strengths, and implications for practice. Psychoanalytical Inquiry: A Topical Journal for Mental Health Professionals 32 (2), 163–170.

Lee, J., 2009. 'This is normal for us': resiliency and resistance among lesbian and gay parents. Gay and Lesbian Issues and Psychology Review 5 (2), 70–80.

Mannering, A.M., Harold, G.T., Leve, L.D., Shelton, K.H., Shaw, D.S., Conger, R.D., Neiderhiser, J.M., Scaramella, L.V., Reiss, D., 2011. Longitudinal associations between marital instability and child sleep problems across infancy and toddlerhood in adoptive families. Child Development 82 (4), 1252–1266.

Marks, L., 2012. Same-sex parenting and children's outcomes: a closer examination of the American psychological association's brief on lesbian and gay parenting. Social Science Research 41 (4), 735–751.

Mercy, J.A., Saul, J., 2009. Creating a healthier future through early interventions for children. The Journal of the American Medical Association 301 (21), 2262–2264.

Metzger, J., 2008. Resiliency in children and youth in kinship care and family foster care. Child Welfare 87 (6), 115–140.

Neumann, A., Barker, E.D., Koot, H.M., Maughan, B., 2010. The role of contextual risk, impulsivity, and parental knowledge in the development of adolescent antisocial behavior. Journal of Abnormal Psychology 119 (3), 534–545.

Odom, W., Zimmerman, J., Forlizzi, J., 2010. Designing for dynamic family structures: divorced families and interactive systems. In: Proceedings of the 2010 ACM Conference on Designing Interactive Systems. ACM Press, Aarhus, Denmark, pp. 151–160.

Padilla-Walker, L.M., Harper, J.M., Bean, R.A., 2011. Pathways to parental knowledge: the role of family processes and family structure. Journal of Early Adolescence 31 (4), 604–627.

Palacios, J., Jiménez, J.M., 2009. Kinship foster care: protection or risk? Adoption and Fostering 33 (3), 64–75.

Pasley, K., Garneau, C., 2012. Remarriage and stepfamily life. In: Walsh, F. (Ed.), Normal Family Processes: Growing Diversity and Complexity, fourth ed. The Guilford Press, New York, pp. 149–171.

Perrone, K.M., 2009. Traditional and nontraditional work and family roles for women and men. Journal of Career Development 36 (1), 3–7.

Perrone, K.M., Wright, S.L., Jackson, Z.V., 2009. Traditional and nontraditional gender roles and work-family interface for men and women. Journal of Career Development 36 (1), 8–24.

Rabitaille, C., Saint-Jacques, M., 2009. Social stigma and the situation of young people in lesbian & gay stepfamilies. Journal of Homosexuality 56 (4), 421–442.

Racz, S.J., McMahon, R.J., 2011. The relationship between parental knowledge and monitoring and adolescent conduct problems: a 10-year update. Clinical Child and Family Psychology Review 14, 377–398.

Rampage, C., Eovaldi, M., Ma, C., Foy, C.W., Samuels, G.M., Bloom, L., 2012. Adoptive families. In: Walsh, F. (Ed.), Normal Family Processes: Growing Diversity and Complexity, fourth ed. The Guilford Press, New York, pp. 222–248.

Rinaldi, C.M., Howe, N., 2012. Mothers' and fathers' parenting styles and associations with toddlers' externalizing, internalizing, and adaptive behaviors. Early Childhood Research Quarterly 27 (2), 266–273.

Schrodt, P., 2011. Stepparents' and nonresidential parents' relational satisfaction as a function of coparental communication in stepfamilies. Journal of Social and Personal Relationships 28 (7), 983–1004.

Schwartz, S.J., Finley, G.E., 2009. Mothering, fathering, and divorce: the influence of divorce on reports of and desires for maternal and paternal involvement. Family Court Review 47 (3), 506–522.

Stack, D.M., Serbin, L.A., Enns, L.N., Ruttle, P.L., Barrieau, L., 2010. Parental effects on children's emotional development over time and across generations. Infants and Young Children 23 (1), 52–69.

Strong, B., Devault, C., Cohen, T.F., 2008. The meaning of marriage and the family. In: Strong, B., Devault, C., Cohen, T.F. (Eds.), The Marriage and Family Experience: Intimate Relationships in a Changing Society, tenth ed. Thomson Wadsworth, USA, pp. 2–29.

Sweeney, M.M., 2010. Remarriage and stepfamilies: strategic sites for family scholarship in the 21st century. Journal of Marriage and Family 72 (3), 667–684.

Tirella, L.G., Tickle-Degnen, L., Miller, L.C., Bedell, G., 2012. Parent strategies for addressing the needs of their newly adopted child. Physical and Occupational Therapy in Pediatrics 32 (1), 97–110.

Turner, H.A., Finkelhor, D., Ormrod, R., 2007. Family structure variations in patterns and predictors of child victimization. American Journal of Orthopsychiatry 77 (2), 282–295.

Vélez, C.E., Wolchik, A., Tein, J.Y., Sandler, I., 2011. Protecting children from the consequences of divorce: a longitudinal study of the effects of parenting on children's coping processes. Child Development 82 (1), 224–257.

Waites, C., 2009. Building on strengths: intergenerational practice with African American families. Social Work 54 (3), 278–287.

Waldfogel, J., Craigie, T.A., Brooks-Gunn, J., 2010. Fragile families and child wellbeing. Future Child 20 (2), 87–112.

Wall, M., 2011. Hearing the voices of lesbian women having children. Journal of GLBT Family Studies 7 (1–2), 93–108.

Walsh, F., 2012. The "new normal": diversity and complexity in 21st-century families. In: Walsh, F. (Ed.), Normal Family Processes: Growing Diversity and Complexity, fourth ed. The Guilford Press, New York, pp. 4–27.

Winokur, M.A., Crawford, G.A., Longobardi, R.C., Valentine, D.P., 2008. Matched comparison of children in kinship care and foster care on child welfare outcomes. Families in Society: The Journal of Contemporary Social Sciences 89 (3), 338–346.

Emotional Intelligence and Competencies

Christopher TH Miners, Queen's School of Business, Queen's University, Kingston, ON, Canada
Ivona Hideg, School of Business and Economics, Wilfrid Laurier University, Waterloo, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Emotional intelligence (EI) has been described as everything from a panacea to an invalid concept. This article focuses on recent research that clarifies what does and does not constitute EI, as well as its relationships with a variety of outcomes. We emphasize the role of EI in the workplace, and we distinguish between the positive outcomes that it facilitates, the negative outcomes to which it contributes, and the outcomes for which its role is unclear. We conclude by identifying specific issues that require further attention to advance our understanding of EI.

Emotional intelligence (EI) has a short but turbulent history. While it has risen to the heights of the cover of Time magazine, it has also been subjected to scathing criticism by its detractors. According to some, it is a panacea for industry's and society's ills. For others, it is an invalid concept. EI is a set of four interrelated abilities: the abilities to perceive emotions in the self and others, use emotions to facilitate performance, understand emotions and emotional knowledge, and regulate emotions in the self and others (Mayer and Salovey, 1997). The academic interest in EI started with the seminal work of Salovey and Mayer (1990). The widespread interest in it, however, is in large part the result of the popular books published by Goleman (1995, 1998). The argument that people could be intelligent in a way that is different from how 'smart' has commonly been understood seemed to have an intuitive appeal, and provocative claims about the importance of EI enhanced the appeal. For example, Goleman suggested that EI is twice as important as the combination of IQ and technical expertise with respect to the job performance of most employees (see Salopek, 1998).

The early research on EI, however, yielded mixed results, even the most positive of which failed to live up to the provocative claims. There were multiple definitions of EI, and many of them were inconsistent with one another. There were also multiple measures of EI, but a number of them overlapped with measures of extant constructs. The imprecise definition and measurement of EI made it difficult to assert that it was a separate, unique construct and to demonstrate that it could predict important outcomes above and beyond other well-known abilities (e.g., cognitive intelligence) and personality traits. This invited timely criticisms of the scientific status of emotional intelligence (e.g., Landy, 2005).

We review the recent research to clarify the definition of EI and the nature of its relationships with a variety of important outcomes, with an emphasis on its role in the workplace. We start with an overview of the three streams of research that ostensibly focus on EI. We then review the positive outcomes that it facilitates, the negative outcomes to which it contributes, and the outcomes for which its role is unclear. We conclude by identifying specific issues that warrant further attention, including cultural variance in EI, its differential impact for men and women, the training of EI, and measurement and methodological issues.

Row, Row, Row Your Boat, Gently Down the (First) Stream

There are three streams of research on 'emotional intelligence' (Ashkanasy and Daus, 2005), but just one of them captures a construct that can meaningfully use the moniker (Côté and Miners, 2006). It should be clear that any construct that purports to be EI should concern emotions *and* intelligence. It should also be clear that the construct should be measured in a way that captures what might reasonably be considered a form of intelligence.

Stream 1 is predicated on the work of Salovey and Mayer (1990). It conceptualizes EI as a set of four interrelated abilities: the abilities to perceive, use, understand, and manage emotions. Abilities reflect a person's capacity to successfully complete a particular task or, in other words, a person's maximum performance (Carroll, 1993). The way in which a person typically performs (i.e., across situations and over time) is the domain of personality traits, which are distinct from abilities (Sackett et al., 1988). Stream 1 therefore conceptualizes EI in a fashion that it is consistent with a form of intelligence.

Stream 1 employs performance tests to assess EI. Performance tests are comprised of questions, the answers to which are right or wrong, or more or less right or wrong. Intelligence concerns the capacity to solve problems (Neisser et al., 1996), hence it is important to employ measures that enable the identification of respondents who have a greater or lesser body of the relevant knowledge and a greater or lesser capacity to spontaneously solve relevant problems. Stream 1 employs omnibus measures of EI, the most popular of which is the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT; Mayer et al., 2002), and performance tests that are specific to a branch of EI (see Roberts et al., 2010 for a description of the measures). Recent meta-analyses have demonstrated that the MSCEIT has a positive correlation with measures of other forms of intelligence, and that it does not correlate highly with measures of personality traits (Joseph and Newman, 2010; O'Boyle et al., 2011).

Stream 2 also conceptualizes EI as a set of interrelated abilities, but it employs self- and peer-reports to assess it. The measures often ask respondents to indicate the extent to which they (or another focal individual) successfully engage in various behaviors. There are a number of problems with using

self-reports to assess a form of intelligence. The common thread among them is that people tend to make inaccurate estimates of their own abilities (e.g., Dunning et al., 2004). Peer-reports are also problematic for a number of reasons. For example, the respondents may not have observed the employee in a wide enough range of emotionally demanding situations to determine whether his or her behavior is the manifestation of one level of EI or another (Côté and Miners, 2006). For these reasons and others, it is difficult-cum-impossible to assert that Stream 2 captures EI. The results of a number of meta-analyses bolster this argument: the measures employed in Stream 2 do not correlate with the measures of other forms of intelligence, and they have a stronger correlation with measures of a number of personality traits (Joseph and Newman, 2010; O'Boyle, Jr. et al., 2011).

Stream 3 does not conceptualize EI as a set of interrelated abilities. Instead, it conceptualizes EI in terms of a series of behavioral outcomes. For example, Bar-On (1997) conceptualized EI as a set of noncognitive capabilities, competencies, and skills that concern (successfully) dealing with environmental pressures. Other researchers explicitly conceptualize it as a personality trait (e.g., Petrides and Furnham, 2001), resulting in the oxymoronic 'trait emotional intelligence.' All of the above conceptualizations incorporate elements that are not abilities and thus, by definition, cannot capture a form of intelligence. Stream 3 also relies on the use of self- and peer-reports. The results of the same meta-analyses and research, unsurprisingly, yield a similar pattern of results to those for the measures that are used in Stream 2: a relationship to other forms of intelligence that is conspicuous by its absence, and a relationship to personality traits that is conspicuous by its presence.

A review of the literature, therefore, shows that Stream 1 captures EI, whereas Streams 2 and 3 do not capture it. Streams 2 and 3 capture important, often highly predictive individual differences, but individual differences for which the moniker 'emotional intelligence' is inappropriate. We believe that the seemingly inauspicious start to research on EI was, at least in part, due to the use of multiple definitions and measures, a number of which were (and are) unfit for purpose. The remainder of this article focuses on Stream 1, and hereafter 'EI' should be understood as such.

The Good, The Bad, and The Ugly

The relationships between EI and a wide variety of important outcomes have been examined, ranging from the quality of social relationships to job performance. The strength of these relationships varies, and while EI is often beneficial (*The Good*), it can also be harmful to individuals and organizations (*The Bad*), or it can occupy an as of yet unclear role (*The Ugly*). Contrary to its (early) reputation as a panacea to organizational and societal problems, a high level of EI is not always the solution to a problem.

The Good

Interpersonal Relationships

EI facilitates interpersonal interactions, such that people with a high level of EI establish and maintain strong relationships. Expressions of emotion contain important information about (other) people's goals, attitudes, and intentions (Frijda and Mesquita, 1994). The abilities to decode emotional expressions, understand the impact of specific emotions on decision-making processes, and manage other people's emotions therefore provide an important means of anticipating and responding to other people in a socially appropriate manner (Côté et al., 2006).

Past research has shown that men with a high level of EI are more likely to be rated as socially competent by the people with whom they interact and by independent observers than men with a low level of EI (Brackett et al., 2006). Furthermore, men and women with a high level of emotion regulation, a branch of EI, are more likely to exercise interpersonal sensitivity and prosocial tendencies, and to be viewed in favorable terms by their peers (Lopes et al., 2005). Past research has also shown that the association between EI and the quality of social relationships remains intact, even when cognitive intelligence and the Big Five personality traits are entered into the analyses (Rossen and Kranzler, 2009).

Leadership Emergence

The same mechanisms that link EI to (positive) social relationships also help us to understand its relationship with leadership emergence. Leadership emergence is the degree to which a person who is not in a formal position of authority influences the other members of a group (Taggar et al., 1999). For example, if a person is able to decode (other) people's expressions of emotion, then he or she can determine their reactions to a proposed course of action. The ability to manage emotions, in turn, can help the person to change the other people's reactions if it becomes necessary to build further support for the proposed course of action (see Huy, 1999).

Recent research has shown that EI does, in fact, predict leadership emergence, even when controlling for cognitive intelligence and the Big Five personality traits (Côté et al., 2010). Walter et al. (2012) obtained a similar pattern of results. Their studies demonstrated that a branch of EI, emotion recognition, had a positive relationship with leadership emergence through its effects on task coordination, at least for people who also had a high level of extroversion. Recent research by Emery (2012) has shown that the abilities to perceive and manage emotions facilitate the emergence of relationship-oriented leaders, and that the abilities to use and understand emotions facilitate the emergence of task-oriented leaders.

Job Performance

The (positive) relationship of EI with job performance is likely to be a complex function of the indirect benefits of social capital and the direct benefits of successfully incorporating emotions in (or excluding them from) decision-making processes. EI should help employees to establish a high level of social capital that, in turn, will provide greater access to valuable information and other resources that facilitate job performance (Sparrowe et al., 2001). EI might also facilitate job performance more directly, by enabling employees to create or up-regulate the emotions that facilitate a particular task and down-regulate the emotions that interfere with it. A

person with a high level of EI is aware of the associations between emotion and cognition, and capable of changing the former to facilitate the latter. This, in short, will help the person to make the best possible decisions and help other people to do the same, which is likely to be reflected in their level of job performance (see Côté and Miners, 2006).

There are a number of individual studies and, now, meta-analyses that confirm EI can facilitate job performance. For example, with respect to the former, Kluemper et al. (2013) found that the ability to manage emotions can facilitate task performance and organizational citizenship behavior directed at the individual, and reduce the likelihood of workplace deviance. The pattern of results held even when a proxy for cognitive intelligence and a measure of the Big Five personality traits were entered into the analyses. The most recent meta-analysis was conducted by O'Boyle Jr. et al. (2011), who used dominance analyses to estimate the relative contributions of EI, cognitive intelligence, and the Big Five personality traits to job performance. They found that EI demonstrated substantial, relative importance even in the presence of these other, robust predictors.

Studies conducted by other researchers have shown that it may be helpful to consider the roles of other individual differences to understand the relationship between EI and job performance. Côté and Miners (2006) found that the relationship between EI and job performance becomes increasingly positive as cognitive intelligence decreases. EI, in other words, compensated for lower levels of cognitive intelligence in helping to achieve a high level of job performance. EI also interacts with other individual differences in a similar fashion, such that its role becomes more prominent if other means of achieving a high level of job performance are absent, or present but in an insufficient quantity. For example, Doucet and Oldham (2006) found that EI had a stronger relationship with job performance for employees at a telephone service center who possessed low levels of agreeableness and conscientiousness, both of which should facilitate a high level of job performance when present to a greater extent.

The results of recent research also suggest that the role of situational characteristics and job characteristics, in particular, should be considered to better understand the relationship between EI and job performance. For example, Farh et al. (2012) found that EI predicted job performance for employees whose work entailed a high level of managerial demands (i.e., the extent to which a particular job involves the management of diverse individuals, functions, and lines of business), but not for employees whose work entailed a low level of such demands. A meta-analysis by Joseph and Newman (2010) provides further impetus for considering situational characteristics. Their results showed that EI had a much stronger, positive correlation with job performance among jobs with a high level of emotional demands, at least when EI was conceptualized and measured as a set of abilities.

The Bad

Counterproductive Work Behavior

EI was presented as a panacea by some of its proponents. It can, however, be harmful to individuals, and it can be used to help other people or to harm them, depending on the ambitions of the particular person. There is now some research that seeks to redress the balance by presenting a more complete picture of the ways in which EI can be used. Elfenbein and Ambady (2002), for example, found that people who are skilled at recognizing emotions can be distracted by even low-level, subtle displays of negative emotion that should pass unnoticed. The distraction consumes resources (i.e., attention) that could otherwise be used on the task at hand, and performance suffers as a result of it. Côté et al. (2011) demonstrated that EI can be leveraged to increase people's level of prosocial behavior, *and* to facilitate deviant behavior that harms other people. Kilduff et al. (2010) have put forward an agenda for future research to examine the strategic deployment of EI or, in other words, to consider when it is used to advance personal interests, even at the expense of other people's success and well-being.

The Ugly

Negotiation

EI should facilitate creating and claiming value during negotiations (Fulmer and Barry, 2004). People with a high level of EI should be able to learn about a negotiation partner's preferences by decoding the information that the partner provides through his or her displays of emotion. Furthermore, they should be able to prevent strong, negative emotions from impacting the exchange of information, and to generate positive emotions that facilitate creative problem-solving, reciprocal positive actions, and frank discussion that will help both parties to move beyond positional bargaining (Adler et al., 1998; Kopelman et al., 2006). This should help to create value in negotiations. A high level of EI should also help people to generate or at least display emotions that beget concessions (and to hide emotions that require making them; e.g., Côté et al., 2013). A high level of EI should also help people to detect when a negotiation partner has reached the reservation point for a particular issue or set of issues (i.e., the lowest value that is still acceptable), and, therefore, to prevent unnecessary concessions (see Elfenbein et al., 2007). This should help to claim value in negotiations.

In spite of the above arguments, the evidence linking EI to negotiation outcomes is mixed. Elfenbein et al. (2007) found that the ability to perceive emotions helped dyads who completed the negotiation exercise to create more joint value (i.e., to increase the size of the outcomes that would ultimately be claimed by the individual members of the dyad). The same ability, however, helped the negotiators who occupied just one of the two roles in the exercise to claim value: the ability to perceive emotions facilitated claiming value for the participants in the role of seller but not in the role of buyer. Other research has focused on the subjective outcomes of negotiations. For example, Mueller and Curhan (2006) have shown that the negotiation partners of people with a high level of EI feel better about the negotiation outcomes, even after controlling for their positive affect and the objective outcomes that they claimed.

The above research suggests that EI may be of limited help with respect to claiming value in a negotiation, but it remains

unclear if and under what circumstances this is true. EI is, however, likely to be beneficial in other ways. For example, if a negotiation is a part of an ongoing relationship between two individuals, then its (positive) impact on the negotiation partner's perception of their outcomes may be invaluable (Elfenbein et al., 2007). A negotiation partner who is pleased with the outcomes that he or she has obtained, even if the reaction is unwarranted in objective terms, is more likely to be willing to negotiate again, and perhaps even to show a measure of generosity to help balance the previous 'successes' and nurture the relationship. Future research will elucidate the nature of the relationship between EI and negotiation outcomes, as additional studies are conducted and longitudinal designs are employed.

Leadership

A number of theoretical arguments link EI to desirable leader behaviors and leadership effectiveness. George (2000), for example, argued that EI should facilitate five core elements of leadership effectiveness, such as creating and maintaining enthusiasm, confidence, optimism, cooperation, and trust. Arguments about the role of emotional aperture in leadership (i.e., the ability to recognize the composition of diverse emotions in a collective) further suggest that the abilities associated with EI should facilitate organizational processes that can provoke strong emotions in employees (Sanchez-Burks and Huy, 2009).

There are, however, arguments to the contrary. Some have argued that a high level of EI will not contribute to leadership or, perhaps, that it might even work against it. For example, Antonakis has argued that condition-action scripts (i.e., mental if-then representations of how to behave in certain situations) are sufficient to explain effective leadership, and to prevent EI from usurping the explanatory power of cognitive intelligence and well-known personality traits (Antonakis et al., 2009). He further argued that a high level of the abilities that constitute EI, such as emotion recognition, may even be harmful, because they may result in unnecessary distraction and prevent leaders from making decisions that will benefit the organization, because even the best possible decisions are unlikely to please everyone.

The empirical research reflects this disagreement among various authors, and suggests that EI can but does not always predict leadership effectiveness. Rubin et al. (2005) showed that leaders with a high level of emotion recognition obtain higher ratings of transformational leadership from their employees than leaders with a relatively low level of this branch of EI, but only if they also had a high level of extroversion. Rosete and Ciarrochi (2005) found a positive relationship between EI and leadership effectiveness, indexed by 360° feedback about the achievement of organizational goals. In contrast, Føllesdal and Hagtvet (2013) found that none of the branches of EI or overall EI predicted employees' ratings of their leader's transformational behavior. A meta-analysis yielded a correlation of .24 in 10 studies that examined the relationship between EI and self-assessed transformational leadership, but the correlation dropped to .05 in the four studies in which subordinates rated the transformational leadership of their supervisors (Harms and Credé, 2010).

The connection between EI and leadership effectiveness is intuitive, but the empirical evidence that is currently available is sparse and not entirely supportive of the connection. As EI has multiple dimensions and leadership effectiveness can be indexed in more or less instrumental or affiliative terms, it will be interesting to clarify the role that EI plays in each domain of outcomes, and to explore any negative impact that it might have on leadership.

Discussion

The evidence reviewed in this article suggests that EI can be helpful to individuals and organizations. It also suggests that we should be wary of the bold claims made by some of its proponents. People should also be attentive to the ways in which EI can be harmful, and mindful of the fact that there are likely to be outcomes to which it does not contribute. EI appears to be beneficial to interpersonal relationships, leadership emergence, and job performance. EI, however, also appears to facilitate the achievement of less salubrious goals. It will be exciting to further explore its dark side and, in turn, to present a more balanced view of it that gives due attention to its virtues and its vices. The role of EI in other outcomes, including negotiations and leadership, if nothing else, requires further research. The current review is not meant to be exhaustive, but instead representative and suggestive of the broader literature. Even so, it is clear that Stream 1 could do with greater breadth and depth to fulfill its potential.

Future Directions

In this section, we discuss areas of research that have not yet received attention in this article or in previous reviews of literature (e.g., Côté, 2014). The main controversies raised in previous reviews relate to the definition and measurement of EI. As we have argued, we believe that future research should employ an ability model of EI and the corresponding performance tests. We now move beyond these issues to provide novel directions for future research.

Cultural Variability

Many of the current definitions and measures of EI have been developed in Western cultures, and thus reflect the values rooted in them. There is, however, a rich body of cross-cultural research shows that there is variability in the perception, expression, and regulation of emotions. For example, past research has shown that people from collectivistic cultures (e.g., Japan) value and encourage the open expression of emotions less than people from individualistic cultures (e.g., United States), because the expression of (negative) emotions can be detrimental to maintaining social harmony (Markus and Kitayama, 1991).

The lack of integration of culture into current models and measures of EI is problematic. For theory, it means that our current models of EI may not be correctly specified, and that the effects of EI on a range of organizational outcomes may be moderated by culture. For example, the consequences of poor emotion regulation may be more severe in collectivistic cultures that value emotion suppression than in individualistic cultures

that encourage the open expression of emotion. For practice, it means that a number of the available measures of EI may have an adverse impact on (potential) employees. If the current measures identify correct answers based on what is appropriate in Western cultures, then they may discriminate against non-Westerners. Indeed, research by Moon (2011) has shown that East Asians score lower on the MSCEIT than Westerners do, and that this effect was due, in part, to the incongruence between East Asians' preference for interdependence and answers that are based on the value of independence.

Given the theoretical and practical challenges of the current research on EI and the rapid globalization of the workplace, we encourage researchers to incorporate the role of culture into their investigations of EI. It would also be helpful to show a similar level of sensitivity to this issue as the next generation of measures is developed, so that researchers can explore the role of EI in a wider variety of cultures with confidence, and practitioners can reduce the likelihood of adverse impact.

Differential Impact for Men and Women

Discussion of the role of gender in the literature on EI has been sparse and, for the most part, focused on the variability in EI test scores that shows that women often score higher than men (see Joseph and Newman, 2010). A more substantive integration of gender into the research on EI is a curious omission, given that EI could be seen as a stereotypical feminine skill and, as such, it may have different consequences for men and women, especially in the workplace.

Past research on gender roles and stereotypes has shown that women are often expected to be warm, affectionate, empathetic, sensitive to others' needs, and good listeners (Eagly and Karau, 2002). Furthermore, it has shown that women who violate gender norms (i.e., who do not display stereotypical feminine traits) are punished (e.g., Rudman, 1998). There appears to be a substantial overlap between EI and stereotypical feminine traits, so it is possible (even probable) that women who do not display a high level of EI may be punished, whereas women who display a high level of EI may not gain very much (especially compared to men). In other words, possessing and displaying a high level of EI may not help to advance women's careers, but a low level of EI (or the failure to display a high level of EI) may hinder them.

This discussion also raises a question about men: would men who display a high level of EI be punished because they are violating the expectations that are associated with stereotypical masculine traits? Past research has shown that men and women incur penalties when they violate their respective gender roles (e.g., Rudman, 1998). Thus, if we construe EI as a feminine set of abilities, men who display a high level of EI may be more likely to be punished than men who display a low level of EI. At the same time, past research has shown that men who display stereotypical feminine traits (e.g., empathy) coupled with stereotypical masculine traits (e.g., ambition) are perceived favorably (Prentice and Carranza, 2002). In other words, a high level of EI may present a unique advantage for men who aspire for leadership positions.

We have outlined a number of possible ways in which EI might differentially influence the careers of women and men, and whether our specific conjectures are right or wrong, we believe that this issue warrants further investigation.

Training

There is a lot of interest in training EI in organizational and educational contexts, but the research on it has been sparse and inconclusive. Some past research suggests that branches of EI can be improved, such as emotion recognition (Elfenbein, 2006). Other research on developing EI, however, has used self- or peer-report measures of EI, rather than a performance test (e.g., Kotsou et al., 2011), hence it is difficult to attest to much more than an improvement in declarative knowledge. Other research still has questioned whether it is even possible to develop EI, especially if EI is defined as a form of intelligence (see Côté, 2014 for a discussion of the trainability and heritability of EI). Furthermore, even if EI can be developed, there is another problem that needs to be solved. Sheldon et al. (2014) found that the people who need the training the most are the least motivated to take it!

If one overlooks the (academic) debate about whether EI *itself* can be developed, then it appears to be possible to develop skills that help people to behave in ways that could be described as 'emotionally intelligent.' Returning to the debate in question, current research has not provided conclusive evidence that EI can, in fact, be trained. As such, we believe that more systematic research on this issue is warranted, particularly given that there is already a large demand for such services.

Measurement and Methodological Considerations

Our review, among other things, has outlined an agenda for advancing our understanding of EI. We would be remiss, however, not to outline some important methodological considerations that should be in place before the research is conducted. The first consideration relates to the measurement of EI. The MSCEIT is the most widely used and extensively validated performance test, but some important concerns have been raised regarding its psychometric properties and its scoring system (e.g., Conte, 2005). To this end, other measures of EI have been developed, and should also be given due consideration (see Roberts et al., 2010 for a review of the available measures).

Furthermore, most research on EI has focused on overall EI rather than on the effects of its specific branches. Past research, however, has shown that different branches may be more or less helpful (or harmful) in one situation or another (e.g., Elfenbein et al., 2007). Assessing overall EI therefore may provide an incomplete or even inaccurate picture of its role. We therefore encourage researchers to continue developing branch-specific hypotheses, and to be mindful not just of which individual branch(es) matter(s) the most, but also of the way in which configurations of multiple branches might affect various outcomes (e.g., a high score on perceiving emotions and a low score on managing emotions might be especially detrimental to an individual's job performance).

Conclusion

EI is popular, but its popularity has done at least as much harm as good. The unrealistically high expectations set by some of its proponents, ironically, were not and are not endorsed by the very people who introduced the concept (i.e., Salovey and Mayer; see Mayer, 2004). These expectations, coupled with the

inconclusive early research, resulted in criticism that was overly pessimistic. We hope that this article has found the middle ground, somewhere closer to the truth: EI can be beneficial to individuals and organizations, but it can also be harmful to them, and its role in a number of outcomes will only be clarified through further research.

Acknowledgments

We thank Stéphane Côté, Paulo N. Lopes, Amanda C. Kesek, and the Section Editor, Richard D. Roberts, for their advice and insightful comments on earlier drafts of this manuscript. This work was supported by two research grants from the Social Sciences and Humanities Research Council of Canada, one to the first author and one to the second author.

See also: Culture and Emotion; Emotion in Cognition; Emotional Regulation; Work Motivation.

Bibliography

Adler, R.S., Rosen, B., Silverstein, E.M., 1998. Emotions in negotiation: how to manage fear and anger. Negotiation Journal 14, 161–179.

Antonakis, J., Ashkanasy, N.M., Dasborough, M.T., 2009. Does leadership need emotional intelligence? Leadership Quarterly 20, 247–261.

Ashkanasy, N.M., Daus, C.S., 2005. Rumors of the death of emotional intelligence in organizational behavior are vastly exaggerated. Journal of Organizational Behavior 26, 441–452.

Bar-On, R., 1997. The Emotional Quotient Inventory (EQ-I): A Test of Emotional Intelligence. Multi-Health Systems, Toronto, ON.

Brackett, M.A., Rivers, S.E., Shiffman, S., Lerner, N., Salovey, P., 2006. Relating emotional abilities to social functioning: a comparison of self-report and performance measures of emotional intelligence. Journal of Personality and Social Psychology 91, 780–795.

Carroll, J.B., 1993. Human Cognitive Abilities: A Survey of Factor-Analytic Studies. Cambridge University Press, New York, NY.

Conte, J.M., 2005. A review and critique of emotional intelligence measures. Journal of Organizational Behavior 26, 433–440.

Côté, S., 2014. Emotional intelligence in organizations. The Annual Review of Organizational Psychology and Organizational Behavior 1, 459–488.

Côté, S., DeCelles, K.A., McCarthy, J.M., Van Kleef, G.A., Hideg, I., 2011. The Jekyll and Hyde of emotional intelligence: emotion-regulation knowledge facilitates both prosocial and interpersonally deviant behavior. Psychological Science 22, 1073–1080.

Côté, S., Hideg, I., van Kleef, G.A., 2013. The consequences of faking anger in negotiations. Journal of Experimental Social Psychology 49, 453–463.

Côté, S., Lopes, P.N., Salovey, P., Miners, C.T.H., 2010. Emotional intelligence and leadership emergence in small groups. Leadership Quarterly 21, 496–508.

Côté, S., Miners, C.T.H., 2006. Emotional intelligence, cognitive intelligence, and job performance. Administrative Science Quarterly 51, 1–28.

Côté, S., Miners, C.T.H., Moon, S., 2006. Emotional intelligence and wise emotion regulation in the workplace. In: Zerbe, W.J., Ashkanasy, N.M., Härtel, C.E.J. (Eds.), Research on Emotion in Organizations, vol. 2. Elsevier, Oxford, UK, pp. 1–24.

Doucet, L., Oldham, G.R., 2006. I know how you feel, but it doesn't always help: emotional accuracy and the effectiveness of customer service representatives. In: Presented at "Service with a Smile, Regardless": Emotions and Customer Service Work Symposium, Annual Meeting for the Society for Industrial and Organizational Psychology, Dallas, TX.

Dunning, D., Heath, C., Suls, J.M., 2004. Flawed self-assessment: implications for health, education, and the workplace. Psychological Science in the Public Interest 5, 69–106.

Eagly, A.H., Karau, S.J., 2002. Role congruity theory of prejudice toward female leaders. Psychological Review 109, 573–598.

Elfenbein, H.A., 2006. Learning in emotion judgments: training and the cross-cultural understanding of facial expressions. Journal of Nonverbal Behavior 30, 21–36.

Elfenbein, H.A., Ambady, N., 2002. Predicting workplace outcomes from the ability to eavesdrop on feelings. Journal of Applied Psychology 87, 963–971.

Elfenbein, H.A., Foo, M.D., White, J., Tan, H.H., Aik, V.C., 2007. Reading your counterpart: the benefit of emotion recognition accuracy for effectiveness in negotiation. Journal of Nonverbal Behavior 31, 205–223.

Emery, C., 2012. Uncovering the role of emotional abilities in leadership emergence. A longitudinal analysis of leadership networks. Social Networks 34, 429–437.

Farh, C.I., Seo, M.G., Tesluk, P.E., 2012. Emotional intelligence, teamwork effectiveness, and job performance: the moderating role of job context. Journal of Applied Psychology 97, 890–900.

Føllesdal, H., Hagtvet, K., 2013. Does emotional intelligence as ability predict transformational leadership? A multilevel approach. Leadership Quarterly 24, 747–762.

Frijda, N.H., Mesquita, B., 1994. The social roles and functions of emotions. In: Kitayama, S., Markus, H.R. (Eds.), Emotion and Culture: Empirical Studies of Mutual Influence. American Psychological Association, Washington, DC, pp. 51–87.

Fulmer, I.S., Barry, B., 2004. The smart negotiator: cognitive ability and emotional intelligence in negotiation. International Journal of Conflict Management 15, 245–272.

George, J.M., 2000. Emotions and leadership: the role of emotional intelligence. Human Relations 53, 1027–1055.

Goleman, D., 1995. Emotional Intelligence. Bantam, New York, NY.

Goleman, D., 1998. Working with Emotional Intelligence. Bantam, New York, NY.

Harms, P.D., Credé, M., 2010. Emotional intelligence and transformational and transactional leadership: a meta-analysis. Journal of Leadership and Organizational Studies 17, 5–17.

Huy, Q.N., 1999. Emotional capability, emotional intelligence, and radical change. Academy of Management Review 24, 325–345.

Joseph, D.L., Newman, D.A., 2010. Emotional intelligence: an integrative meta-analysis and cascading model. Journal of Applied Psychology 95, 54–78.

Kilduff, M., Chiaburu, D.S., Menges, J.I., 2010. Strategic use of emotional intelligence in organizational settings: exploring the dark side. Research in Organizational Behavior 30, 129–152.

Kluemper, D.H., DeGroot, T., Choi, S., 2013. Emotion management ability predicting task performance, citizenship, and deviance. Journal of Management 39, 878–905.

Kopelman, S., Rosette, A.S., Thompson, L., 2006. The three faces of eve: strategic displays of positive, negative, and neutral emotions in negotiations. Organizational Behavior and Human Decision Processes 99, 81–101.

Kotsou, I., Nelis, D., Grégoire, J., Mikolajczak, M., 2011. Emotional plasticity: conditions and effects of improving emotional competence in adulthood. Journal of Applied Psychology 96, 827–839.

Landy, F.J., 2005. Some historical and scientific issues related to research on emotional intelligence. Journal of Organizational Behavior 26, 411–424.

Lopes, P.N., Salovey, P., Côté, S., Beers, M., 2005. Emotion regulation ability and the quality of social interaction. Emotion 5, 113–118.

Markus, H.R., Kitayama, S., 1991. Culture and the self: implications for cognition, emotion, and motivation. Psychological Review 98, 224–253.

Mayer, J.D., 2004. Be realistic. Harvard Business Review 82, 28.

Mayer, J.D., Salovey, P., 1997. What is emotional intelligence? In: Salovey, P., Sluyter, D.J. (Eds.), Emotional Development and Emotional Intelligence. Basic Books, New York, NY, pp. 3–31.

Mayer, J.D., Salovey, P., Caruso, D.R., 2002. Mayer-Salovey-Caruso Emotional Intelligence Test. Multi-Health Systems, Toronto, ON.

Moon, S., 2011. East Meets West: The Cultural Relativity of Emotional Intelligence (Unpublished doctoral dissertation). University of Toronto.

Mueller, J.S., Curhan, J.R., 2006. Emotional intelligence and counterpart mood induction in a negotiation. International Journal of Conflict Management 17, 110–128.

Neisser, U., Boodoo, G., Bouchard Jr., T.J., Boykin, A.W., Brody, N., et al., 1996. Intelligence: knowns and unknowns. American Psychologist 51, 77–101.

O'Boyle Jr., E.H., Humphrey, R.H., Pollack, J.M., Hawver, T.H., Story, P.A., 2011. The relation between emotional intelligence and job performance: a meta-analysis. Journal of Organizational Behavior 32, 788–818.

Petrides, K.V., Furnham, A., 2001. Trait emotional intelligence: psychometric investigation with reference to established trait taxonomies. European Journal of Personality 15, 425–448.

Prentice, D.A., Carranza, E., 2002. What women and men should be, shouldn't be, are allowed to be, and don't have to be: the contents of prescriptive gender stereotypes. Psychology of Women Quarterly 26, 269–281.

Roberts, R.D., MacCann, C., Matthews, G., Zeidner, M., 2010. Emotional intelligence: toward a consensus of models and measures. Social and Personality Psychology Compass 4, 821–840.

Rosete, D., Ciarrochi, J., 2005. Emotional intelligence and its relationship to workplace performance outcomes of leadership effectiveness. Leadership & Organization Development Journal 26, 388–399.

Rossen, E., Kranzler, J.H., 2009. Incremental validity of the Mayer–Salovey–Caruso Emotional Intelligence Test Version 2.0 (MSCEIT) after controlling for personality and intelligence. Journal of Research in Personality 43, 60–65.

Rubin, R.S., Munz, D.C., Bommer, W.H., 2005. Leading from within: the effects of emotion recognition and personality on transformational leadership behavior. Academy of Management Journal 48, 845–858.

Rudman, L.A., 1998. Self-promotion as a risk factor for women: the costs and benefits of counterstereotypical impression management. Journal of Personality and Social Psychology 74, 629–645.

Sackett, P.R., Zedeck, S., Fogli, L., 1988. Relations between measures of typical and maximum job performance. Journal of Applied Psychology 73, 482–486.

Salopek, J.J., 1998. Train your brain. Training and Development 52, 26–33.

Salovey, P., Mayer, J.D., 1990. Emotional intelligence. Imagination, Cognition and Personality 9, 185–211.

Sanchez-Burks, J., Huy, Q.N., 2009. Emotional aperture and strategic change: the accurate recognition of collective emotions. Organization Science 20, 22–34.

Sheldon, O.J., Dunning, D., Ames, D.R., 2014. Emotionally unskilled, unaware, and uninterested in learning more: reactions to feedback about deficits in emotional intelligence. Journal of Applied Psychology 99, 125–137.

Sparrowe, R.T., Liden, R.C., Wayne, S.J., Kraimer, M.L., 2001. Social networks and the performance of individuals and groups. Academy of Management Journal 44, 316–325.

Taggar, S., Hackett, R., Saha, S., 1999. Leadership emergence in autonomous work teams: antecedents and outcomes. Personnel Psychology 52, 899–926.

Walter, F., Cole, M.S., der Vegt, G.S.V., Rubin, R.S., Bommer, W.H., 2012. Emotion recognition and emergent leadership: unraveling mediating mechanisms and boundary conditions. Leadership Quarterly 23, 977–991.

Emotional Regulation

Min Y Kim, The Hebrew University in Jerusalem, Mount Scopus, Jerusalem, Israel and Keimyung University, South Korea
Yochanan Bigman and Maya Tamir, The Hebrew University in Jerusalem, Mount Scopus, Jerusalem, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Emotion regulation involves attempts to change the experience and expression of emotions. People engage in such attempts to satisfy either hedonic motives (i.e., to feel good in the moment) or instrumental motives (i.e., to effectively achieve their goals). To change their emotions, people can engage in a variety of strategies. While some strategies (e.g., cognitive reappraisal) have been found to be relatively effective, others (e.g., expressive suppression) have proven less effective. The quality and efficacy of emotion regulation depends, in part, on the mental resources people have at their disposal, on their beliefs about the ability to change emotions, and on their knowledge and understanding of emotions. The process of emotion regulation, in turn, carries important implications for subsequent emotional experiences, for cognition and decision-making, for social interactions, for mental and physical health, and for well-being.

Our daily life is imbued with emotions. Sometimes we savor them and sometimes we try to change them. Overcoming fear of public speaking, getting ourselves excited about a new project at work, stifling an inappropriate giggle – these are all examples of emotion regulation. Emotion regulation involves attempts to maintain, increase, or decrease the experience and expression of unpleasant or pleasant emotions (Gross, 1998). Intrinsic (or intrapersonal) emotion regulation involves attempts to change the experience or expression of our own emotions, whereas extrinsic (or interpersonal) emotion regulation involves attempts to change the experience or expression of emotions in others (Thompson, 1994).

Emotions are reactions to goal-relevant events that involve changes in subjective experience, which may be accompanied by changes in cognition, physiology, and behavior (*see* Emotions, Psychological Structure of). Given that emotions carry important implications, it is not surprising that emotion regulation has been implicated in important outcomes across multiple domains, including decision-making, organizational behavior, intergroup relations, education, psychological and physical health, and well-being (for recent reviews, see Gross, 2014).

Research on emotion regulation in the past few decades has addressed many key questions. These include why people regulate emotions, how people regulate emotions, what influences the way people regulate emotions, and whether and how emotion regulation influences subsequent outcomes. In what follows, we discuss each of these questions and review some of the research that speaks to them.

Why Do People Regulate Emotions?

People engage in emotion regulation for different reasons. To date, researchers have distinguished between two categories of motives that drive emotion regulation (e.g., Tamir, 2009) – namely, hedonic and instrumental. First, people regulate emotions to achieve hedonic benefits, that is, to increase pleasure and decrease pain (e.g., Larsen, 2000). Pleasant emotions, such as happiness or joy, feel good, whereas unpleasant emotions, such as fear, sadness, or anger, feel bad. By increasing pleasant emotions and decreasing unpleasant emotions, people can increase pleasure and decrease pain, respectively. Therefore, people are often motivated to increase pleasant emotions and decrease unpleasant ones to attain hedonic benefits.

Second, people regulate emotions to achieve instrumental benefits – namely, to increase utility or expected benefits (e.g., Tamir, 2009). Regardless of how pleasant they are, emotions can be useful or harmful for goal attainment. For instance, happiness is pleasant, but it may also promote successful collaboration. Similarly, anger may be unpleasant, but it may also promote successful confrontation, and fear may be unpleasant, but it may promote successful avoidance of threat (e.g., Frijda, 1986). Both pleasant and unpleasant emotions may be useful in certain contexts and harmful in others. By increasing useful emotions and decreasing harmful emotions, whether they are pleasant or unpleasant, people can increase utility (Parrott, 1993). Therefore, people are often motivated to increase useful emotions and decrease harmful emotions to attain instrumental benefits. Because emotions that are unpleasant may be useful in certain contexts, people may be motivated to experience even unpleasant emotions to promote the attainment of instrumental goals. For instance, people may be motivated to increase their anger in preparation for a fight (e.g., Tamir et al., 2008).

What people want to feel, in turn, determines the direction in which they regulate their emotions. People who are motivated to feel less angry will try to decrease their anger, whereas those who are motivated to feel angrier will try to increase their anger (Tamir et al., 2008). People who regulate their emotions to satisfy hedonic motives will likely try to increase positive emotions and decrease negative emotions. People who regulate their emotions to satisfy instrumental motives may try to change their emotions in whichever way they believe would promote the attainment of their goals (Tamir et al., 2014).

How Do People Regulate Their Emotions?

People can use various strategies to regulate their emotions. There have been two approaches to studying emotion-regulation strategies. One approach focuses on particular strategies, in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.25055-1

an attempt to identify and understand their unique antecedents and consequences. A second approach focuses on broader theoretical models that organize multiple types of regulation strategies.

One prominent example of the first approach to studying emotion-regulation strategies involves research on rumination, which involves repetitively thinking about the symptoms of one's distress, and on its possible causes and consequences (Nolen-Hoeksema et al., 2008). For instance, after having failed a test, a person who ruminates might think about the test over and over again, focusing on how they have done wrong and how this might affect them. Although people might engage in rumination in an attempt to resolve negative events, rumination is ineffective in decreasing negative emotions (e.g., Webb et al., 2012), and may even backfire (Nolen-Hoeksema et al., 2008). Indeed, there is evidence that rumination can actively contribute to psychopathology, and depression, in particular (e.g., Aldao et al., 2010).

Perhaps the best example of the second approach to studying emotion-regulation strategies is research on the process model of emotion regulation (Gross, 1998). According to this model, emotions unfold in a sequence of four stages: situation, attention, appraisal, and response. First, we encounter an emotion-eliciting situation. For instance, after having had dinner with friends, we might find ourselves walking alone back to our car, which is parked in a dark alley. Second, we attend to the emotion-relevant features of the situation. For instance, our attention is drawn to a few large men huddling in a corner. Third, we assign meaning to the situation, through a process of cognitive appraisal. For instance, we might tell ourselves that this group of men could be dangerous. Finally, we have an emotional response. For instance, we might feel fear and hurry to our car as fast as we can.

According to Gross' process model, emotion regulation can be activated in different stages of the emotion-generation process. Emotion-regulation strategies, therefore, can be distinguished by the specific stage they target in the emotion-generation process (Gross, 1998). The first two types of strategies involve attempts to alter the situation that elicits the emotional response. First, situation-selection strategies are initiated before people encounter an emotion-eliciting event. Such strategies involve attempts to influence the likelihood of encountering events that elicit desired (or undesired) emotions. For instance, we could decide to take a cab to dinner earlier in the evening, thereby eliminating the need to walk alone in a dark alley at night – a situation that may elicit fear. Second, situation-modification strategies are initiated when people encounter an emotion-eliciting event. Such strategies involve attempts to change the nature of the situation itself in a manner that changes its emotional impact. For instance, we could ask some friends to walk us to our car after dinner, thereby increasing our sense of safety, making it less likely that we would feel fear. Situation selection and situation modification involve attempts to change the situation that elicits emotions.

The next two types of strategies involve attempts to change the way we process the emotion-eliciting event. First, attention-modification strategies are initiated after the encounter with the emotion-eliciting situation. These strategies involve attempts to redirect attention away or toward emotion-eliciting features of the situation. For instance, if we are bound to walk alone to our car at night, we could decide to focus our attention on the elderly couple walking hand in hand rather than the group of huddling men, keeping our fear at bay. Second, cognitive-change strategies are initiated after we have attended to the emotion-eliciting features of the situation. These strategies involve attempts to change our interpretation or the meaning we assign to the situation, so that the emotional impact of the situation changes. For instance, assuming we have already attended to the group of huddling men, we could tell ourselves that these men are not much different than the friends with whom we have had dinner, who are enjoying each other's company after a hard day at work. This interpretation makes the situation less threatening, making it less likely that we would feel fear.

According to Gross (1998), the strategies discussed above are antecedent-focused strategies, because they are initiated before the emotional response. Response-focused strategies are initiated after the emotional response has occurred and target features of the emotional response, including experiential, physiological, behavioral, and expressive features. For example, as we get scared watching the group of men huddling in the dark alley, we may try to hide the expression of fear on our faces and continue walking at a steady, calm pace toward our car.

Gross' process model gives rise to predictions regarding the relative efficacy of different regulation strategies. According to Gross (1998), antecedent-focused strategies should be more effective than response-focused strategies, because they are initiated earlier in the process. To test this prediction, researchers have focused on two specific regulation strategies – namely, cognitive reappraisal and expressive suppression (e.g., Gross, 1998). Cognitive reappraisal is an example of a cognitive-change strategy, which involves changing the appraisal of a situation. Expressive suppression is an example of a response-modulation strategy, which involves inhibiting the overt expression of emotion. There is now considerable evidence showing that cognitive reappraisal is relatively effective in changing emotional experience, whereas expressive suppression is relatively ineffective (Webb et al., 2012).

In addition to identifying specific emotion-regulation strategies and studying their relative efficacy, researchers have examined the manner in which regulation strategies are employed. Recent research on self-regulation, broadly construed, has demonstrated that although people often use conscious effort when using regulation strategies, they can also activate such strategies unconsciously (e.g., Mauss et al., 2007). For instance, participants who were implicitly primed with the concept of reappraisal were as effective in decreasing their physiological arousal as those who were explicitly instructed to reappraise (e.g., Williams et al., 2009). Although there is some evidence that people can regulate their emotions without conscious intention, there is still some debate about the nature, antecedents, and consequences of automatic emotion regulation.

What Are the Antecedents of Emotion Regulation?

In addition to why and how people regulate their emotions, there are various factors that influence whether and how successfully they do so. In this section, we review several key factors that influence the nature and efficacy of the emotion-regulation process. These include the regulatory resources that

enable people to regulate their emotions, people's beliefs about the feasibility of emotion regulation, and people's knowledge about emotions.

Ability

All forms of self-regulation require mental resources, and emotion regulation is no exception (*see* Self-Regulated Learning). There is evidence that emotion regulation can be depleting. Effortful emotion regulation impairs performance on subsequent effortful tasks, is impaired when the individual is placed under high cognitive load, and in individuals with impaired executive functioning skills (for a review, see Zelazo and Cunningham, 2007). Therefore, to regulate emotions effectively, individuals need to have sufficient regulatory resources.

Beliefs

To engage in emotion regulation, people need to believe that emotions can be controlled. People have implicit theories (i.e., beliefs about the malleability of attributes) about emotions (Tamir et al., 2007). Some people hold entity beliefs about emotions and believe that emotions cannot be controlled, whereas others hold incremental beliefs about emotions and believe that emotions can be controlled. The more people believe that emotions are controllable, the better they are in controlling their emotions and the more pleasant and the less unpleasant emotions they experience over time (Tamir et al., 2007).

To engage in emotion regulation, however, people must believe not only that emotions can be controlled, but that they personally are capable of regulating their emotions. Self-efficacy refers to the belief that one has the ability to exert control (*see* Self-Efficacy). People vary in their self-efficacy in emotion regulation. People who believe that emotions can be controlled experience more self-efficacy in emotion regulation, and this sense of personal efficacy is associated with the experience of more-pleasant and less-pleasant emotions and with more effective emotion regulation (Tamir et al., 2007). Therefore, to attempt emotion regulation people may need to believe that emotions can be controlled and that they personally can regulate their emotions.

Knowledge

To effectively regulate emotions, people need to know what emotions are, what gives rise to them, what their consequences are, and how they can be modified (Barrett and Gross, 2001). Individuals who have more elaborate knowledge about emotions are more effective at emotion regulation. For instance, people who were better able to differentiate between discrete emotional experiences reported using a wider range of regulation strategies (Tugade et al., 2004). Therefore, to regulate emotions effectively people need to know and understand emotions.

What Are the Consequences of Emotion Regulation?

Emotion regulation carries important implications in multiple domains. As might be expected, emotion regulation influences emotional experience. In addition, however, emotion regulation influences cognition and decision-making, social relationships, health, and well-being. We review each of these domains below.

Emotion Regulation Influences Emotion

People engage in emotion regulation in order to regulate their emotions. Indeed, emotion regulation can change the experience and expression of emotions. People who are trained and instructed to use effective regulation strategies are successful at reducing negative emotional reactions to distressing stimuli (see Webb et al., 2012). For example, cognitive reappraisal, distraction, and distancing are effective in decreasing negative emotional reactions. When emotion regulation is used effectively, it can change emotional experiences in both adults and in children (e.g., Thompson, 1994).

Emotion regulation can change not only the experience, but also the expression of emotions. For instance, expressive suppression involves decreasing the overt expression of emotions, whereas venting involves the overt expression of negative emotions. Both are effective ways of controlling the expression of emotions.

Emotion Regulation Influences Cognition

Some forms of emotion regulation carry cognitive costs (for a review, see Suri et al., 2013). For instance, certain strategies impair the processing of event-related information. In particular, distraction leads to more superficial processing of the emotion-eliciting event, because it involves intentionally diverting attention away from it. Expressive suppression also impairs memory of the emotion-eliciting event, potentially because of the mental resources it requires. Other forms of emotion regulation carry cognitive benefits (for a review, see Suri et al., 2013). For instance, there is evidence that some strategies (e.g., cognitive reappraisal) promote enhanced recall of explicit memory. Emotion regulation, more generally, could promote effective processing of information. For instance, children who were instructed to decrease their emotional responses were better able than others to recall educational material.

Emotion Regulation Influences Decision-Making

Some forms of emotion regulation may promote certain aspects of decision-making. For instance, emotion regulation may influence risk-perception. Participants who actively increased their anger became less aversive to losses, and consequently made riskier financial choices (e.g., Tamir et al., 2014). Emotion regulation can also influence utilitarian reasoning. When facing an unfair financial offer, people usually reject the offer and prefer to get nothing rather than agree to an unfair division. Participants who actively decreased their anger were more likely to accept an unfair offer to increase personal financial gain (Wout et al., 2010).

Emotion Regulation Influences Social Interactions

There is considerable evidence that emotion regulation influences social interactions, from intimate bonds to group

processes. First, emotion regulation plays an important role in close relationships (Shaver and Mikulincer, 2007). According to attachment theory, individuals are motivated to maintain proximity to attachment figures. When faced with potential separation, individuals experience distress and are motivated to reestablish proximity. This process initiates the regulation of emotion. Emotion regulation is recruited to deal with attachment threat. Securely attached individuals experience their distress, notice it, and tolerate it. Anxiously attached individuals engage in hyperactivation strategies, increasing the experience and expression of distress, to increase the chances that the attachment figure will attend to them. Avoidantly attached individuals engage in deactivating strategies, decreasing the experience and expression of distress, given that they view the attachment figure as nonresponsive (Shaver and Mikulincer, 2007). These regulation strategies are assumed to influence the nature and quality of close relationships.

Second, emotion regulation plays an important role in social relationships, more broadly. In particular, decreasing emotions that promote interpersonal violence (e.g., anger, hatred) can result in more adaptive social outcomes. There is evidence that aggressive behavior is associated with impaired emotion regulation. Children who have difficulties in regulating their emotions are less socially competent and behave more aggressively toward others. Similarly, adults who have difficulties regulating their emotions are more likely to be violent toward intimate partners. In contrast, effective emotion regulation is associated with more stable and positive relationships (for a review, see English et al., 2013).

Both the regulation of emotional experience and the regulation of emotional expression influence social relationships. For example, sharing emotions with other people not only improves one's ability to cope with negative events, but it also strengthens the social bond between the sharer and the person with whom she/he shares emotions (Rimé, 2007). While the expression of emotions to others carries social benefits, the suppression of emotional expression carries social costs. People who suppress the expression of their emotions are less effective in communicating with others, are liked less by others, and are perceived by others as inauthentic (see English et al., 2013). Suppression of emotions can also have deleterious effects in the workplace (e.g., Côté and Morgan, 2002).

Finally, emotion regulation plays an important role not only in dyadic relationships, but in group contexts as well (for a review, see Halperin, 2014). In the context of the Isreali–Palestinian conflict, Israeli participants who were trained to decrease their anger using cognitive reappraisal were more willing to provide humanitarian aid to Palestinians and less likely to support aggressive policies. Such studies demonstrate that emotion regulation can be used to enhance less-aggressive and more-peaceful social relationships, at the dyadic and group levels.

Emotion Regulation Influences Mental Health

Impairments in emotion regulation actively contribute to psychopathology. Difficulties in regulating negative emotions and the use of ineffective regulation strategies contribute to longer episodes of depression in depression-prone individuals (see Joormann and Siemer, 2014). Similarly, difficulties in regulating positive emotions seem to contribute to the development of mania in bipolar disorder (Gruber, 2011). More generally, dysfunctional emotion regulation has been implicated in depression, bipolar mood disorder, borderline personality disorder, generalized anxiety disorder, post-traumatic stress disorder (PTSD), substance-use disorders, eating disorders, and somatoform disorders (see Berking and Wupperman, 2012).

Emotion dysregulation in psychopathology may involve two possible factors. First, individuals who suffer from psychopathology tend to use ineffective regulation strategies (e.g., rumination, suppression) and fail to use more effective ones (e.g., reappraisal). Second, individuals who suffer from psychopathology have difficulties in using emotion regulation in a flexible and context-dependent manner (Bonanno and Burton, 2013). Instead, they are likely to use few regulation strategies in a rigid manner, regardless of what the situation calls for, thus impairing their success in regulation.

Emotion Regulation Influences Physical Health

There is some evidence linking emotion regulation to physical health. Using less-effective emotion-regulation strategies (e.g., suppression) has been associated with a greater risk of chronic health outcomes, such as heightened cortisol reactivity in response to stressors (Lam et al., 2009). In contrast, effective emotion regulation has been related to a lower risk of metabolic syndrome (Liukkonen et al., 2006). Similarly, writing about personally upsetting topics has been found to have beneficial influences on immune function, including t-helper cell growth, antibody response to Epstein–Barr virus, and antibody response to hepatitis B vaccinations (e.g., Pennebaker et al., 1988). Taken together, these findings suggest that emotion regulation impacts physical health. However, the mechanisms by which they do so remain to be examined.

Emotion Regulation May Influence Well-being

Emotions comprise an important component of well-being. Therefore, it is conceivable that the regulation of emotions can contribute to well-being (see Nyklicek et al., 2011). Individuals who can effectively regulate their emotions report higher levels of life satisfaction and psychological well-being, compared to those who do not engage in emotion regulation or regulate their unpleasant emotions ineffectively (e.g., John and Gross, 2007). Similarly, people who cultivate pleasant emotions are more satisfied with their lives and are less likely to suffer from depressive symptoms. Emotion regulation also mediates the effects of life stress on well-being. For instance, people who use cognitive reappraisal are more likely to recover from emotional stress and less likely to suffer from depressive symptoms (Troy et al., 2010). Much of the research linking emotion regulation to well-being is correlational, however, and whether emotion regulation influences well-being or well-being influences emotion regulation is yet to be determined.

Summary: Healthy Emotion Regulation

Emotion regulation is a process by which people change the experience or expression of their emotions. It can be used to

make people feel better or to achieve a variety of both personal and interpersonal goals. It can be accomplished using a variety of strategies, some more effective and some less effective than others. It is influenced by the regulatory resources people have at their disposal, by their beliefs about emotions, and by their knowledge of emotions. Finally, it impacts how people feel, how they think, how they make decisions, how they interact with one another, as well as their mental health, physical health, and well-being. To promote adaptive outcomes across these diverse domains, people need to engage in healthy emotion regulation.

Healthy emotion regulation involves changing emotions in a useful and context-sensitive manner. Healthy emotion regulation has several key features. First, it involves regulating emotions in a manner that is consistent with one's goals. Second, it involves using effective regulation strategies and avoids using ineffective ones. Third, to be healthy, emotion regulation should be used in a flexible, variable, and context-sensitive manner. Indeed, it is important to know not only when to regulate, but also when not to regulate and instead embrace emotional experiences as they are (Hayes et al., 2006).

See also: Chronic Illness, Psychosocial Coping with; Chronotype; Culture and Emotion; Early Emotional Development and Cultural Variability; Emotion in Cognition; Emotions, Psychological Structure of; Stress, Coping and Health.

Bibliography

Aldao, A., Nolen-Hoeksema, S., Schweizer, S., 2010. Emotion-regulation strategies across psychopathology: a meta-analytic review. Clinical Psychology Review 30, 217–237. http://dx.doi.org/10.1016/j.cpr.2009.11.004.

Barrett, L.F., Gross, J.J., 2001. Emotional intelligence: a process model of emotion representation and regulation. In: Mayne, T.J., Bonanno, G.A. (Eds.), Emotions: Current Issues and Future Directions. Guilford Press, New York, NY, pp. 286–310.

Berking, M., Wupperman, P., 2012. Emotion regulation and mental health: recent finding, current challenges, and future directions. Current Opinion in Psychiatry 25, 128–134. http://dx.doi.org/10.1097/YCO.0b013e3283503669.

Bonanno, G.A., Burton, C.L., 2013. Regulatory flexibility: an individual differences perspective on coping and emotion regulation. Perspectives on Psychological Science 8 (6), 591–612.

Côté, S., Morgan, L.M., 2002. A longitudinal analysis of the association between emotion regulation, job satisfaction, and intentions to quit. Journal of Organizational Behavior 23, 947–962. http://dx.doi.org/10.1002/job.174.

English, T., John, O., 2013. Understanding the social effects of emotion regulation: the mediating role of authenticity for individual differences in suppression. Emotion 13, 314–329. http://dx.doi.org/10.1037/a0029847.

English, T., John, O.P., Gross, J.J., 2013. Emotion regulation in close relationships. In: Simpson, J.A., Campbell, L. (Eds.), The Oxford Handbook of Close Relationships. Oxford University Press, pp. 500–513.

Frijda, N., 1986. The Emotions. Cambridge University Press, London.

Gross, J.J., 1998. Antecedent- and response-focused emotion regulation: divergent consequences for experience, expression, and physiology. Journal of Personality and Social Psychology 74, 224–237.

Gross, J.J. (Ed.), 2014. Handbook of Emotion Regulation, second ed. Guilford, New York, NY.

Gruber, J., 2011. When feeling good can be bad: positive emotion persistence (PEP) in bipolar disorder. Current Directions in Psychological Science 20, 217–221.

Halperin, E., 2014. Emotion, emotion regulation, and conflict resolution. Emotion Review 6 (1), 68–76.

Hayes, S.C., Luoma, J.B., Bond, F.W., Masuda, A., Lillis, J., 2006. Acceptance and commitment therapy: model, processes and outcomes. Behaviour Research and Therapy 44, 1–25.

John, O.P., Gross, J.J., 2007. Individual differences in emotion regulation. In: Gross, J.J. (Ed.), Handbook of Emotion Regulation. Guilford Press, New York, NY, pp. 351–372.

Joormann, J., Siemer, M., 2014. Emotion regulation in mood disorders. In: Gross, J.J. (Ed.), Handbook of Emotion Regulation, second ed. Guilford Press, New York, pp. 413–427.

Lam, S., Dickerson, S.S., Zoccola, P.M., Zaldivar, F., 2009. Emotion regulation and cortisol reactivity to a social-evaluative speech task. Psychoneuroendocrinology 34, 1355–1362. http://dx.doi.org/10.1016/j.psyneuen.2009.04.006.

Larsen, R.J., 2000. Toward a science of mood regulation. Psychological Inquiry 11, 129–141. http://dx.doi.org/10.1207/S15327965PLI1103_01.

Liukkonen, T., Silvennoinen-Kassinen, S., Jokelainen, J., Räsänen, P., Leinonen, M., Meyer- Rochow, V.B., Timonen, M., 2006. The association between C-reactive protein levels and depression: results from the northern Finland 1966 birth cohort study. Biological Psychiatry 60, 825–830.

Mauss, I.B., Bunge, S.A., Gross, J.J., 2007. Automatic emotion regulation. Social and Personality Psychology Compass 1, 146–167. http://dx.doi.org/10.1111/j.1751-9004.2007.00005.x.

Nolen-Hoeksema, S., Wisco, B.E., Lyubomirsky, S., 2008. Rethinking rumination. Perspectives on Psychological Science 3, 400–424.

Nyklicek, I., Vingerhoets, A.J.J.M., Zeelenberg, M., 2011. Emotion regulation and well-being: a view from different angles. In: Nyklicek, I., Vingerhoets, A.J.J.M., Zeelenberg, M. (Eds.), Emotion Regulation and Well-being. Springer, New York, pp. 1–9.

Parrott, W.G., 1993. Beyond hedonism: motives for inhibiting good moods and for maintaining bad moods. In: Wegner, D.M., Pennebaker, J.W. (Eds.), Handbook of Mental Control. Prentice-Hall, Inc., Englewood Cliffs, NJ, USA, pp. 278–305.

Pennebaker, J.W., Kiecolt-Glaser, J., Glaser, R., 1988. Disclosure of traumas and immune function: health implications for psychotherapy. Journal of Consulting and Clinical Psychology 56, 239–245.

Rimé, B., 2007. Interpersonal emotion regulation. In: Gross, J.J. (Ed.), Handbook of Emotion Regulation. Guilford, New York, pp. 466–485.

Shaver, P.R., Mikulincer, M., 2007. Adult attachment strategies and the regulation of emotion. In: Gross, J.J. (Ed.), Handbook of Emotion Regulation. Guilford Press, New York, pp. 446–465.

Suri, G., Sheppes, G., Gross, J.J., 2013. Emotion regulation and cognition. In: Robinson, M.D., Watkins, E.R., Harmon-Jones, E. (Eds.), Handbook of Cognition and Emotion. Guilford, New York, NY, pp. 195–209.

Tamir, M., 2009. What do people want to feel and why? Pleasure and utility in emotion regulation. Current Directions in Psychological Science 18, 101–105. http://dx.doi.org/10.1111/j.1467-8721.2009.01617.x.

Tamir, M., Bigman, Y.E., Rhodes, E., Salerno, J., Schreier, J., 2014. An Expectancy-Value Model of Emotion Regulation: Implications for Motivation, Emotional Experience, and Decision Making. Emotion. Advance online publication. http://dx.doi.org/10.1037/emo0000021.

Tamir, M., John, O.P., Srivastava, S., Gross, J.J., 2007. Implicit theories of emotion: affective and social outcomes across a major life transition. Journal of Personality and Social Psychology 92, 731–744. http://dx.doi.org/10.1037/0022-3514.92.4.731.

Tamir, M., Mitchell, C., Gross, J.J., 2008. Hedonic and instrumental motives in anger regulation. Psychological Science 19, 324–328. http://dx.doi.org/10.1111/j.1467-9280.2008.02088.x.

Thompson, R.A., 1994. Emotion regulation: a theme in search of definition. In: Fox, N.A. (Ed.), The Development of Emotion Regulation: Biological and Behavioral Considerations, Monographs of the Society for Research in Child Development, vol. 59, pp. 25–52.

Troy, A.S., Wilhelm, F.H., Shallcross, A.J., Mauss, I.B., 2010. Seeing the silver lining: cognitive reappraisal ability moderates the relationship between stress and depressive symptoms. Emotion 10, 783–795. http://dx.doi.org/10.1037/a0020262.

Tugade, M.M., Feldman Barrett, L., Gross, J.J., 2004. Matters of feeling precisely: examining emotional granularity and emotion regulation. In: Paper presented at the annual meeting of the American Psychological Society, Chicago, Illinois.

Webb, T.L., Miles, E., Sheeran, P., 2012. Dealing with feeling: a meta-analysis of the effectiveness of strategies derived from the process model of emotion regulation. Psychological Bulletin 138, 775–808. http://dx.doi.org/10.1037/a0027600.

Williams, L.E., Bargh, J.A., Nocera, C.C., Gray, J.R., 2009. The unconscious regulation of emotion: nonconscious reappraisal goals modulate emotional reactivity. Emotion 9, 847–854. http://dx.doi.org/10.1037/a0017745.

Wout, M. Van, Chang, L.J., Sanfey, A.G., 2010. The influence of emotion regulation on social interactive. Emotion 10, 815–821. http://dx.doi.org/10.1037/a0020069.

Zelazo, P.D., Cunningham, W.A., 2007. Executive function: mechanisms underlying emotion regulation. In: Gross, J.J. (Ed.), Handbook of Emotion Regulation. Guilford, New York, pp. 135–158.

Emotion and Expression

Paul E Griffiths and Elena Walsh, University of Sydney, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The scientific study of emotional expression began with studies of facial expression in the nineteenth century. Studying these expressions has been an important site for disputes over the existence of a universal human nature and disputes about whether emotions are naturally divided into a manageable number of discrete kinds with innate patterns of behavior evolved by natural selection. A more recent dispute concerns the extent to which facial expressions are voluntary, performative strategies for social gain, as opposed to involuntary expressions of subjective feeling.

What Is at Stake in Discussions of Facial Expression?

Emotion can be expressed in the face, voice, posture, and gesture. Since Darwin, most researchers have concentrated on facial expressions. However, in recent years interest in vocal expression has increased, partly due to advances in acoustic technologies (Scherer, 2003). The nature and origins of facial expressions of emotion may not seem at first glance to be a matter of momentous concern for the social sciences and humanities. However, the emotions are widely believed to be a critical feature of moral agency, and are even more widely believed to be a critical part of aesthetic response. If all healthy people display, recognize, and respond to the same limited number of emotional expressions, this suggests a similar uniformity in emotion itself. Conversely, if human expressions of emotion are as diverse as human languages and if humans can only understand the expressive repertoire of their own cultural group, this might support a view of emotions themselves as culturally specific phenomena.

The issue of whether emotions fall into a limited number of discrete kinds has also been debated for its own sake, as a foundational issue in the psychology of emotion. Evidence about facial expression has played a critical role in these debates.

The Origins of Research on Facial Expression

The scientific study of facial expression began with the nineteenth-century anatomical investigations of Sir Charles Bell (1844) and G.B. Duchenne (1867/1990). The most influential work of the nineteenth century, however, was undoubtedly Charles Darwin's *The Expression of the Emotions in Man and Animals* (1872). The book started life as a chapter of Darwin's *Descent of Man* (1871), based on notes that Darwin had been accumulating since the 1830s. The chapter grew so large that it became a separate work. In *Descent of Man*, Darwin aimed to show evolutionary continuity between human ethical behavior and animal behavior and also, via his theory of sexual selection, between the human aesthetic sense and the aesthetic sense of animals. In *The Expression of the Emotions in Man and Animals*, he aimed to show evolutionary continuity between the human and animal emotions by demonstrating continuity between the facial expressions of humans and animals. Darwin explicitly targeted Bell's claim that the muscles of the human face were created by God to express human emotions. This led him to argue that many movements that now express emotion were vestiges of previous ways of life: "With mankind some expressions, such as the bristling of the hair under the influence of extreme terror, or the uncovering of the teeth under that of furious rage, can hardly be understood, except in the belief that man once existed in a much lower and animal-like condition" (Darwin, 1872: p. 12). Some authors have denied that Darwin ascribed any current adaptive functions to emotional expressions (e.g., Fridlund, 1994). However, to take just one example, Darwin argued that humans (unlike chimpanzees) open their mouths when terrified because (unlike chimpanzees) humans must breathe through their mouth when fleeing. Many of Darwin's specific claims about the facial movements associated with particular emotions are supported by contemporary research. He also pioneered the use of photographs to test the ability of observers to infer emotional state from facial expression (see **Figure 1**; **Figure 2**).

The Affect Program Theory of Facial Expression

Darwin's work on emotional expression was neglected for the first half of the twentieth century. Anthropologists such as Weston La Barre (1947) and Ray Birdwhistell (1963) argued that culturally specific emotional states were signaled in a culturally specific code of facial expressions and gestures acquired by the individual during their upbringing. This culturalist tradition was displaced in the late 1960s by a powerful revival in the Darwinian approach. Many of Darwin's specific findings about facial expression were confirmed and extended, most notably by the psychologist Paul Ekman and his collaborators (Ekman, 1972). One famous experiment used subjects from the Fore language group in New Guinea with a minimum of prior contact with Westerners and their cultural products. These subjects were given three photographs, each showing a face, and told a story which was designed to involve only one emotion. They were asked to pick the photograph showing the person in the story. This method has the advantage that no translation of the names of emotions is needed. The subjects were highly successful in picking the photograph intended to express the emotion appropriate for each story. The New Guinean subjects were also asked to act out the facial behavior

Figure 1 Photographs used by Darwin to test the recognition of emotion from facial expression. 1–3, Degrees of mild amusement; 4, impassivity; 5, true smile; 6, false smile. Note the electrodes used by Duchenne to stimulate the muscles.

Figure 2 Some photographs used in Ekman and his collaborators' research on facial expression. From top left: anger, fear, disgust, surprise, happiness, and sadness. Source: Ekman, P., Friesen, W.V., 1975. Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues. Prentice-Hall, Englewood Cliff, New Jersey, © Paul Ekman. Reproduced with permission.

of the people described in the stories. Videotapes of their responses were shown to US college students. The students were generally accurate in their judgments of the emotion intended by the New Guineans. The six facial expressions that figured in these experiments were labeled surprise, joy, sadness, fear, anger, and disgust. These and a seventh initially not distinguished from disgust but later labeled contempt, became the basis of the revival of the Darwinian approach to expression.

Ekman conceptualized facial expressions as key components of 'affect programs,' complex patterns of facial and other actions controlled by the central nervous system and triggered in response to some specific stimulus (Ekman, 1984). This conceptualization derived, ultimately, from the idea of a 'fixed action pattern' as it figures in the classical ethology of Konrad Lorenz and Niko Tinbergen. These founders of the modern study of animal behavior rediscovered Darwin's work on the emotions and interpreted it as prefiguring many of their own ideas (Lorenz, 1965). Emotions, they argued, were innate patterns of behavior that had evolved by natural selection. Darwin's emphasis on the vestigial nature of many human facial expressions was interpreted using the ethological concept of ritualization (Tinbergen, 1952). Patterns of behavior that once served a practical function, such as opening the mouth and retching in disgust, were retained and modified by natural selection because of their value as signals. Hence, the primary biological function of facial expressions in humans is as a signal of the person's emotional state. The distinctive experimental paradigm of classical ethology was the 'deprivation experiment,' which aimed to show that an animal will produce a pattern of behavior even when it is deprived of the opportunity to learn that behavior. The human ethologist Irenaus Eibl-Eibesfeldt (1973) examined the facial movements of children born deaf and blind and showed that these children used the same patterns of facial action as other children to display their emotions. He concluded that these behaviors were evolved features of human nature.

Ekman and his collaborators explained cultural differences in the expression of emotion with the concept of a 'display rule.' This concept is exemplified by an experiment in which American and Japanese college students were shown neutral and stress-inducing films while alone in a room. The repertoire of facial behaviors shown during the stress phase by the two sets of subjects was very similar. However, when an experimenter was introduced into the room and allowed to ask questions about the subject's emotions as the stress film was shown again, the facial behavior of the Japanese diverged radically from that of the Americans. Video analysis showed the momentary occurrence of negative emotional expressions and their replacement with polite smiles. An important feature of the display rule conceptualization of cultural differences is that the evolved expressions remain intact but interact with

culturally specific behaviors to determine the observable pattern of facial action.

Ekman and collaborators' theory of discrete facial expressions of emotions retains a special status in the study of emotion and its expression because of the close relationship between the discrete emotions model and the widely used technology for coding facial expression (FACS – Facial Action Coding System) whose use was pioneered by Ekman and collaborators. FACS is a set of objective, anatomical measures that need not be used to classify emotions into a small number of discrete categories; but that use of the method is, for example, integrated into the training manuals for FACS coding.

Universality and Cultural Specificity

The research leading to the affect program theory and subsequent discussion have helped to clarify issues about the universality of emotion. The affect programs have the same output across cultures, but they do not have the same input. There are some universal elicitors of affect programs in childhood, such as very loud noises, which elicit fear. There are also systematic biases in the conditioning of affect program responses that could lead to a convergence in the eliciting conditions for adult responses. The general picture, however, is that affect programs come to be associated with whatever stimuli locally fulfill a broad functional role, so that the fear affect program comes to be associated with whatever locally constitutes a threat, the disgust response with whatever locally appears noxious or unclean, and so forth (for a good introduction to work in this area, see Ekman and Davidson, 1994: pp. 144–178). Distinguishing between the universality of outputs and that of inputs makes it clear that at least some disputes over the universality of emotion is merely semantic. Universalists regard two cultures as manifesting the same emotion when they respond in the same way to different stimuli. Their critics insist that having the same emotion means responding in the same way to the same things.

Further clarification results from distinguishing the question of whether emotions are pancultural (found in all cultures) from the question of whether emotions are universal (found in all healthy individuals). The types of evidence normally gathered by universalists are designed to show that emotions are pancultural and have little bearing on the question of universality. Emotions might be pancultural but still be like blood type or eye color, with several different types of individual in each population. Models of the evolution of social emotions sometimes predict that competing types will be maintained in the same population through competition. It is surprising that the issue of whether emotions have evolved is still so strongly linked to the issue of whether there is a single, universal, human emotional nature.

Finally, the debate over universality could be clarified by abandoning the last vestiges of the traditional dichotomy between learned and innate behaviors. Some critics have argued that a biological perspective on emotion is inappropriate merely because the emergence and maintenance of emotional responses depends upon environmental factors. Conversely, evidence that emotions are pancultural and thus likely to be the products of evolution is still thought to imply that these emotions are genetically determined and resistant to modification by environmental changes. These inferences ignore the fact that the environment plays a rich and constructive role in the development of even the most stereotypically biological traits, such as bodily morphology or sexual behavior. Evolved emotions, like the rest of evolved psychology, will likely make use of many reliable features of the environment of the developing child in order to construct and maintain themselves.

If facial expressions of emotions depend on both genetic and environmental factors in development, as it is commonly assumed all phenotypes do, then they will be open to cultural and individual variation as a result of environmental variation, as well as genetic variation. Some recent work in the psychology of facial expression explicitly adopts this perspective: Elfenbeim et al. (2007) suggest the evidence base favors both universality and cultural specificity, and that research should focus on the ways in which an underlying 'universal grammar' of emotional expression is affected by culturally specific emotional dialects. Parkinson (2012) suggests that research on emotional expression should articulate more precisely the ways in which environmental factors influence phylogenetic, ontogenetic, and moment-by-moment social transactions, rather than continuing to fixate on the question of the universality of expressions. Griffiths (1997) suggested that questions of universality can often be usefully reframed in terms of the biological concept of 'homology': Are two emotions in different individuals or cultures homologous, that is, are they modified forms of a response in a common ancestor of those individuals or cultures?

Contemporary Alternatives to the Affect Program Theory

Skepticism about the affect program view of emotion, as well as of emotional expression, has been articulated most clearly in recent years by James A. Russell (2003) and Lisa Feldman Barrett (2006a,b) under the label of 'psychological constructionism.' Russell and Barrett object to affect program theory for taking emotion words like 'anger' and 'fear' as psychological primitives. They claim that the fundamental building block of emotion is 'core affect'; a subjective awareness of one's bodily state along two dimensions, valence (pleasure to displeasure) and arousal (activation to deactivation). Core affect becomes emotion only when top-down processes of categorization couple one's affective bodily state to an external event. For example, when a bushwalker couples his sudden sweating and increased heart rate with the fact that he can see a snake nearby, he labels his experience as one of 'fear.' If he did not categorize his experience in this way, he would not be feeling the emotion of fear. Because this ability to categorize requires social learning, and because social learning varies by culture, core affect theory opposes universalism about discrete emotions and about their expressions. However, Russell and Barrett think that the underlying processes of valence and arousal are universal, even if their means of expression are necessarily mediated by culturally specific factors.

Barrett and Russell are part of a wider group of emotion researchers who have questioned the affect program theory's

account of the production and interpretation of facial expression (Russell and Fernandez-Dols, 1997). However, not all of these authors subscribe to Russell and Barrett's psychological constructionist account of emotion. Some, for example, are more concerned with the interpersonal nature of emotions, and with the idea that expressions are primarily signals, rather than expressions of an internal, motivational state (Fridlund, 1991; Parkinson et al., 2005). Some critics have offered methodological criticisms of the work of Ekman and his collaborators. They are critical of the use of forced choices from a list of emotion terms dictated by the experimenter. This procedure forces a range of responses to a range of facial signals into a few boxes defined by the experimenter. It cannot provide evidence for the existence of a limited number of discrete facial expressions, since this assumption is built into the experimental procedure. On the positive side, critics of the affect-program view have produced considerable evidence of 'audience effects' on the production of emotional expressions. This is interpreted as evidence that expressions are produced for the benefit of the audience rather than to express the feelings of the person making the display. This suggests that the affect program theory exaggerates the automated nature of emotional expressions and neglects their voluntary, performative aspect.

While they criticize the experiments usually taken to establish the universality of emotion, many critics accept that emotional expressions have evolved and would therefore presumably expect the flexible signaling system they describe to be at least substantially pancultural. Alan Fridlund (1994) argues that an evolutionary perspective actually favors the view that emotional expression is a form of paralanguage, since involuntary signals of true emotional state would be subverted by the evolution of dissimulation and deceit. While this is an important perspective on the evolution of emotional expressions, it does not constitute the decisive argument that Fridlund seems to suppose. Veridical signals do evolve in nature, often by making use of the so-called hard-to-fake signals. The cost of being unable to suppress a signal of emotional state may be balanced by the advantage of being believed. A purely theoretical argument based on the evolutionary dynamics of signaling systems seems more likely to support the view that "examples of emotional behavior lie along a continuum from expression to negotiation" (Hinde, 1985: p. 989) than a purely expressive or purely manipulative picture of emotional expression.

Conclusion

Facial expressions of emotion are an important site for the debate between universalism and cultural specificity about emotion because both sides believe that here they can offer substantial evidence for their positions. An examination of this evidence need not serve only to support one or other side of the debate. The actual details of emotional expression, its evolution, and its development in the individual can be used to question the validity of the dichotomies between biology and culture, fixity and malleability, universality and idiosyncraticity which are assumed by both sides.

See also: Affect-Regulation Motivation; Avoidance and Approach Motivation: A Brief History; Culture and Emotion; Emotion, Evolution of; Face Recognition Models: Computational Approaches; Face Recognition, Psychological and Neural Aspects; Facial Emotion Expression, Individual Differences in; Flow in Motivational Psychology; Motivation and Actions, Psychology of; Passion and Motivation.

Bibliography

Barrett, L.F., 2006a. Are emotions natural kinds? Perspectives on Psychological Science 1, 28–58.
Barrett, L.F., 2006b. Solving the emotion paradox: categorisation and the experience of emotion. Personality & Social Psychology Review 10 (1), 20–46.
Bell, C., 1844. Anatomy and Philosophy of Expression as Connected with the Fine Arts, third ed. John Murray, London.
Birdwhistell, R.L., 1963. The kinesic level in investigations of the emotions. In: Knapp, P.H. (Ed.), Expression of the Emotions in Man. International Universities Press, New York, pp. 123–139.
Darwin, C., 1871. The Descent of Man. John Murray, London.
Darwin, C., 1872. The Expressions of Emotions in Man and Animals, first ed. Philosophical Library, New York.
Duchenne, G.B., 1867/1990. The Mechanism of Human Facial Expression. Cambridge University Press, Cambridge.
Eibl-Eibesfeldt, I., 1973. Expressive behavior of the deaf and blind born. In: von Cranach, M., Vine, I. (Eds.), Social Communication and Movement. Academic Press, London & New York, pp. 163–194.
Ekman, P., 1972. Emotions in the Human Face. Pergamon Press, New York.
Ekman, P., 1984. Expressions and the nature of emotions. In: Scherer, K.R., Ekman, P. (Eds.), Approaches to Emotion. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 319–344.
Ekman, P., Davidson, R.J. (Eds.), 1994. The Nature of Emotion: Fundamental Questions. Oxford University Press, New York, Oxford.
Ekman, P., Friesen, W.V., 1975. Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues. Prentice-Hall, Englewood Cliff, NJ.
Elfenbeim, H.A., Beaupre, M., Lévesque, M., Hess, U., 2007. Toward a dialect theory: cultural differences in the expression and recognition of posed facial expressions. Emotion 7 (1), 131–146.
Fridlund, A.J., 1991. Evolution and facial action in reflex, social motive and paralanguage. Biological Psychiatry 32, 3–100.
Fridlund, A.J., 1994. Human Facial Expression: An Evolutionary View. Academic Press Inc., San Diego.
Griffiths, P.E., 1997. What Emotions Really Are: The Problem of Psychological Categories. The University of Chicago Press, Chicago.
Hinde, R.A., 1985. Was 'expression of the emotions' a misleading phrase? Animal Behavior 33, 985–992.
La Barre, W., 1947. The cultural basis of emotions and gestures. Journal of Personality 16, 49–68.
Lorenz, K., 1965. Preface. In: Darwin, C. (Ed.), The Expression of the Emotions in Man and Animals. University of Chicago Press, Chicago.
Parkinson, B., Fischer, A.H., Manstead, A.S.R., 2005. Emotions in Social Relations: Cultural, Group and Interpersonal Processes. Psychology Press, New York.
Parkinson, B., 2012. Piecing together emotion: sites and timescales for social construction. Emotion Review 4 (3), 291–298.
Russell, J.A., Fernandez-Dols, J.M. (Eds.), 1997. The Psychology of Facial Expression. Cambridge University Press, Cambridge.
Russell, J.A., 2003. Core affect and the psychological construction of emotion. Psychological Review 110 (1), 145–172.
Scherer, K., 2003. Vocal communication of emotion: a review of research paradigms. Speech Communication 40 (1–2), 227–256.
Tinbergen, N., 1952. Derived activities: their causation, biological significance, origin and emancipation during evolution. Quarterly Review of Biology 27, 1–32.

Emotion in Cognition

Julia A Harris, Harvard University, Cambridge, MA, USA
Derek Isaacowitz, Northeastern University, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by K. Oatley, volume 7, pp. 4440–4444, © 2001, Elsevier Ltd.

Abstract

This article addresses emotion and cognition by first introducing basic models where emotions are thought to be discrete entities (e.g., anger, sadness, happiness). Appraisal theories of emotion are introduced next, including psychological construction approaches, where emotions are classified as emerging from specific ingredients of the mind, and moves away from the basic emotions approach and emphasizes the variability and variety of emotion categories across cultures. Recent and past empirical evidence is provided throughout the article, providing support for both basic and nonbasic theories of emotion. Finally, clinical implications on emotion and cognition are discussed along with providing a life-span perspective.

Introduction

Traditionally, an encyclopedia entry on 'the role of emotion in cognition' might consider different emotions, such as anger and fear, and present theory and some data considering how each of these different emotional states might influence (and perhaps 'direct') cognitive operations like attention, memory, and decision-making. In this updated article, we will consider this approach as well, but hope to embed that approach within larger conceptual frameworks that provide complementary and alternative ways of considering how emotion and cognition may relate to each other. In the first part of the article, we consider how different definitions of emotion lead to different views of the role of emotion in cognition (and the place of cognition in emotion). In the second part, we present some relevant empirical findings. In the third part, we discuss how a clinical and life-span perspective may also inform the study of emotion–cognition links.

Approaches to Emotion and Their Implications for Emotion–Cognition Links

Basic Emotion Models

Asking simple questions about how emotion influences cognition may be most associated with a view of emotions known as the 'basic emotions' view. In a basic emotion model, each discrete emotion triggers a set of evolutionarily derived biological responses. A discrete emotion when triggered by stimuli in the environment initiates a cascade of behavioral, expressive, autonomic, and neuroendocrine responses that are specific to that emotion. The classic example is as follows: You are approached by a bear (stimulus). The bear triggers an emotion (likely fear). The fear response goes off like a bomb and sets in motion a set of programmed responses (e.g., widened eyes, increased heartbeat, sweaty palms) that are biologically programmed to help you survive the bear chase.

An important feature of basic emotion models is that each emotion (e.g., 'fear,' 'anger,' 'sadness') is a unique mechanism that causes a unique mental state with unique, measurable outcomes. Further, the mental states and measurable outcomes (associated with each emotion) occur *consistently* across every instance of that emotion and *specifically* only in that emotion (Ekman, 1972; Panksepp, 1998; Izard, 1993; for recent reviews, see Ekman and Cordaro, 2011; Gendron and Barrett, 2009; Gross and Barrett, 2011).

From a basic emotion perspective, the answer to the question 'What are the cognitive prerequisites for emotion?' would likely be 'minimal.' Basic emotion theorists postulate that low-level sensory information is processed for emotional relevance, quickly and automatically, activating a dedicated neural pathway that, in turn, triggers a cascade of responses to organize behavior and physiology. Further, theorists propose that this processing occurs very rapidly and does not require any conscious awareness or higher-order cognitive processing (Adolphs, 1999; Ekman and Davidson, 1994; LeDoux, 2000; Levenson, 2011; Panksepp, 1998; Tracy and Randles, 2011). To the extent that discrete emotions reflect different antecedent events and orchestrate distinct behavioral responses, it is logical that certain emotions would have certain effects on attention and memory aligned with the behaviors associated with that particular emotional response.

Appraisal Models

Appraisal theories view emotional experiences as a process of changes between components. These components can include environment evaluation, motivation, peripheral physiological response, expressive and instrumental behavior, and subjective feelings. In nearly all appraisal views of emotion, cognition is a necessary prerequisite for emotional experience. Cognitive processes or 'appraisals' are considered antecedent events that must occur in order for a response to take place. Some appraisal models (particularly early models, e.g., Arnold, 1960; Lazarus, 1998) refer to appraisals as specific cognitive antecedents of emotion that make meaning from the world. In these models, appraisals can be thought of as a set of switches, that, when configured in certain patterns, trigger biologically basic emotional responses characterized by behavioral or physiological outputs or by a strong, almost inescapable tendency to interact with the world in a particular way, an 'action tendency' (for recent reviews, see Moors, 2013). Other appraisal models

do not view appraisals as causes of an emotion. Rather, they are loosely coordinated action tendencies that are configured based on context and may or may not come to fruition.

Whereas some appraisal theories offer a more general mechanism of appraisal in which attention is generally captured by stimuli that are intrinsically relevant to the individual and an emotional response follows that relevance appraisal, there are also versions that provide a more detailed account of what particular appraisals may be critical. For example, the component process model posits specific evaluations that result in particular emotional responses such as novelty, intrinsic pleasantness, and goal relevance. These evaluations include determining the implications and possible consequences of the stimulus or event, potential to cope with the consequences, and whether or not the event is aligned with self and/or cultural standards, in that sequential order (Scherer, 1984, 2001, 2013). All stimulus evaluation checks in this model depend on the valence determination from the previous check; thus, this entire process can be thought of as a filter system to determine how much a particular stimulus or event will be appraised by an individual and what emotion will result.

Most appraisal theorists agree that appraisals are automatic and often unconscious and that the appraisal process begins immediately following the onset of an emotion-eliciting stimulus. A few appraisal models, however, view emotions as a way of experiencing the world. In these models, cognition and emotion emerge and interact to form the content of that meaning (e.g., Clore and Ortony, 2008; Schwarz, 2010). Often called 'affect as information' or 'feelings as information,' these theories propose that we use our current affective state as information about the way the world is. From this perspective, emotions and cognition are bidirectional and influence each other.

In one classic experiment, participants were asked about their life satisfaction on sunny and rainy days. Participants who were not asked about the weather reported significantly less life satisfaction on rainy days than on sunny days. For participants who were reminded about the weather, sun or rain made no difference in their ratings of life satisfaction (Schwarz and Clore, 1983). These studies (and others) suggest that people can use their momentary affective states as information about themselves and the world around them.

Psychological Construction Model

Psychological construction models view emotions as emerging from more basic elements and/or systems of the mind, rather than emotions being the basic building blocks of the mind. Instead of breaking down emotions into specific emotion categories (anger, fear, happiness, etc.), psychological construction theories emphasize the variability in the number and variety of emotion categories both across individuals and across cultures. In psychological construction models, cognition and emotion are not separate processes that occur sequentially. Rather, basic psychological ingredients (such as language, affect, and executive attention) combine to form discrete emotion experiences. Psychological construction theories do not subscribe to a basic emotions account that all humans express specific emotions that fall into specific emotion categories. Instead, psychological construction approaches view emotions as being emergent, where emotion instances are created when afferent information from the body is made meaningful in relation to external surroundings. In order to be made meaningful, 'core systems' or 'ingredients' of mental states are utilized where a human has a system(s) in place for representing sensations related to the body (e.g., pounding of the heart) and a separate system for making sense of these sensations in relation to the situation (e.g., giving a speech in front of a large crowd) (Barrett, 2013, 2009a). One psychological construction model, the conceptual act model, proposes that humans are constantly picking up cues from the body and the external world and simultaneously categorizing them according to their preexisting conceptual knowledge and past experiences (Barrett, 2006, 2009b; Lindquist, 2013).

In terms of emotions having specific neural components, the iterative reprocessing model (an additional psychological construction theory) emphasizes that many different patterns of neural activity can create a single mental state or emotional episode. Some theories (specifically the conceptual act theory) claim that emotions have no localization to any one particular brain network, while iterative reprocessing model believes in a 'processing framework' in which sensory information flows through a hierarchical organization of the brain but is continually reinterpreted and shaped according to an individual's goals, values, desires, and intentions (Cunningham et al., 2013; Cunningham and Zelazo, 2007).

In summary, psychological construction approach theories emphasize that emotions cannot be reduced to specific components (such as facial expression, valence, arousal, neural localization, physiological response) and instead emotions are continuously redefined and categorized at an individual level depending on the situational context and the state of the body (Wilson-Mendenhall et al., 2011). In terms of cognition, psychological construction models assert that specific mechanisms are necessary to construct an emotional episode, such as memory and categorization (learning), and can be guided by attention or perception of visual images and auditory sounds.

From Conceptual Models to Research Findings

Clearly, there are different ways of conceptualizing the role of discrete emotions in attention and cognition, and these vary in somewhat fundamental ways as a function of how emotions themselves are defined. Nonetheless, they each provide a way to investigate how different types of emotional states may influence various kinds of cognitive processing. Below, we review some recent research in this area.

Some theorize that discrete emotions, or emotions that are short-lived, intense, and arise from stimulus-specific events, have clear effects on attention and cognition (Ekman and Cordaro, 2011; Izard, 2011; Levenson, 2011; Panksepp and Watt, 2011). These findings tend to be most consistent with the basic emotions approach. Emotions of negative valence (e.g., disgust, fear, sadness, anger, etc.) differ to the extent that they influence cognition, particularly on appraisals that involve judgment and decision-making. For instance, disgust, an emotion often recognized from a feeling in the 'gut' (often inducing feelings of nausea or other gastroenteric nervous system responses) and from its distinct facial expression

(Ekman, 1972; Tybur et al., 2013), has been found to influence cognition in terms of shaping moral judgments. In empirical designs, subjects induced to experience feelings of disgust (e.g., via smell, taste, etc.) tend to judge acts or situations (e.g., first-cousin marriage) as being more wrong or less acceptable compared to subjects not experiencing disgust (Schnall et al., 2008; Eskine et al., 2011). Thus, researchers theorize that disgust intuitions serve as an 'input to systems' that serve to aid in decision-making about what is morally appropriate versus what is not (Pizarro et al., 2011; Tybur et al., 2013).

Fear has been thought to have deep evolutionary origins stemming from the need to survive and to be constantly aware of potentially threatening surroundings. Due to this, the experience of fear places a strong control on attention, demonstrated in a study conducted by Öhman et al. (2001) in which participants were instructed to detect a target stimulus in the face of distractor stimuli. Participants more quickly identified the target stimulus if it was fear-relevant target (e.g., pictures of snakes and spiders) compared to nonfearful target (e.g., mushrooms, flowers), providing evidence that fearful target objects direct visual attention. Furthermore, the emotional experience of fear leads to the need for protection and escape and often leads to higher estimations of the likelihood of risky events occurring (Angie et al., 2011).

Sadness, often evoked by a substantial loss to an individual (e.g., the death of a loved one), turns attention inward and promotes detail-oriented information processing. The emotional experience of sadness allows for a 'pause' and a cognition update to revise goals and plans through the decrease of physiological arousal, which facilitates problem-solving (Welling, 2003; Bonanno and Keltner, 1997; Oatley and Johnson-Laird, 1996). Anger, an emotion often arising from events or people threatening an individual's outcome in a particular situation, has been found to be associated with a sense of individual control and evokes risk-seeking choices and subsequent behaviors. The experience of anger often triggers physiological responses resulting in increased blood flow to facial regions, thus linking perception of the color red to anger. In fact, research has found that activating anger concepts in participants (e.g., by categorizing anger-related words such as 'furious' or 'irate') led to the interpretation of perceiving ambiguous color objects as being red (Fetterman et al., 2011). In addition to the experience of anger resulting in 'seeing red,' anger can also lead to biased information processing, which can have detrimental effects on individuals. For example, anger initiates cognitive errors, such as misattributing causation, or "making negative assumptions while ignoring the possibility of other interpretations" (Martin and Dahlen, 2007: pp. 156) and overgeneralization, or usage of broad concepts and language such as 'everyone,' 'never,' or 'always' (Martin and Dahlen, 2007).

Generally, emotions of a positive valence have been theorized to initiate exploration or the broadening of attention, expanding cognition and creativity (Fredrickson, 1998, 2001; Gasper and Clore, 2002; Isen, 1987; Rowe et al., 2007). In an eye-tracking study conducted by Wadlinger and Isaacowitz (2006), participants induced into a positive mood state fixated more on peripheral stimuli (e.g., images of a similar valence) than central stimuli compared to control participants, providing evidence that positive emotional states promote attentional breadth via visual attention. For example, the experience of happiness has been found to promote creativity in studies assessing problem-solving task accuracy, suggesting that the experience of positive affect "increases the tendency to combine material in new ways and to see relatedness between divergent stimuli" (Isen et al., 1987: p. 1130). Joy, often termed as happiness (Lazarus, 1991), initiates the thought–action experience by increasing the urge to 'play,' involving increased exploration, invention, imagination, and can have long-standing positive effects on building an individual's physical, intellectual, and social skills (Fredrickson, 1998). Gratitude, or the experience of feeling grateful or thankful, is a positive emotion that has recently received more attention in emotion research, particularly in the domain of social functioning (Algoe, 2012; Algoe et al., 2013). Cognitive-emotion theories emphasize that the feeling of gratitude is associated with the recognition and appreciation of an altruistic gift or action and results in increased cognitive capacity to empathize with others (Lazarus and Lazarus, 1994; McCullough et al., 2001). Contentment, a low-arousal positive emotion also termed as 'tranquility' or 'serenity,' arises from appraisals of situation safety, certainty, and requiring minimal effort (Ellsworth and Smith, 1988). The experience of contentment prompts individuals to cognitively integrate recent life events and circumstances into their overall self and worldview, resulting feelings of connection or 'oneness' with the world around them (De Rivera et al., 1989; Fredrickson, 1998; Izard, 1977).

In addition to negative (e.g., anger, fear, etc.) and positive (e.g., happiness, joy, etc.) emotions, the experience of awe (otherwise termed as 'astonishment,' 'interest,' 'amazement,' 'surprise') can be evoked by the occurrence of rare events perceived by the viewer as 'fascinating,' including viewing something unknown for the first time, or experiencing what one would consider to be a 'miracle,' and can be both positive and negative in nature (Ekman, 1992; Frijda, 1986; Keltner and Haidt, 2003). The emotional experience of awe encompasses feelings of both perceived vastness (e.g., larger than the self) and the need for accommodation (e.g., the adjusting of mental structures to assimilate an entirely new experience) (Piaget and Inhelder, 1966/1969). The process of accommodation can leave one feeling confused, small, and powerless if the conception of the vast experience is not made sense of, or enlightened if the individual succeeds during the accommodation process and makes sense of the vastness (Keltner and Haidt, 2003).

Other research has suggested that whether or not cognition is impaired or enhanced by emotional information depends on the nature of the emotional stimuli. In one such study, Pessoa et al. (2012) found that participants exhibited strong response inhibition when presented with fearful and happy facial expressions relative to neutral faces, suggesting that in one case emotional stimuli (regardless of valence) serve to facilitate cognition (via increased ability to not respond to a stimulus in a stop-signal task). However, follow-up studies manipulating the nature of the fearful facial expression by a conditioned shock to the response stimulus showed decreased response inhibition in that participants struggled to stop responding. This finding suggests that arousal, but not valence, may play

a primary role in emotion–cognition interactions and provides evidence that the type of emotional stimuli and the effect that it has on the individual (and perhaps the appraisal) will either facilitate or inhibit cognition.

Furthermore, emotionally salient events/stimuli are thought to "burn longer and brighter in the mind's eye" (Todd and Anderson, 2014: p. 12) by capturing visual attention and enhancing visual cortex activation. Due to this, emotionally evocative information has been thought to have privileged perceptual processing by grabbing attention even during states of diminished attentional resources (e.g., situations of limited memory consolidation capacity). However, there is a plethora of visual stimuli in the environment both emotionally salient and neutral, so the question becomes at what point does the visual information reach consciousness, and just how and why does some information reach conscious awareness, while other information passes through, unfiltered? Some research has suggested effects of particular emotional expressions on attention, in ways that would then have implications for cognitive processing (Ekman et al., 1969; Tracy and Matsumoto, 2008; Susskind et al., 2008). For example, when one encounters a spider, a fearful facial expression is often produced that includes elongation of the muscles surrounding the eyes, nose, and mouth. Thus, the sensory organs are essentially 'open' and the individual will experience increased sensory vigilance to the environment. During the emotional experience of disgust, the facial expression produced has been prototyped by researchers to have the exact opposite muscle activity as fear, resulting in clenched muscles surrounding the eyes, nose, and mouth (Susskind et al., 2008). During the facial expression of disgust, individual experiences sensory rejection of the surrounding environment, resulting in diminished attention to the environment. It has been proposed that facial expressions might, in fact, serve the physical purpose to modify perception by visually restricting access to sensory information during contamination (disgust) or increasing access to sensory information during threat (fear) (Todd and Anderson, 2014; Tybur et al., 2013).

Beyond effects of particular emotional states on attention and cognition, positive and negative states more generally may have effects as well. Appraisals resulting in sustained emotional states (e.g., positive or negative emotional states) have been found to alter neurotransmitters function and thus influence subsequent cortical processes responsible for motivation, attention, etc. For example, negative emotional states have been hypothesized to narrow perceptual processing causing decreased detection and peripheral exploration, whereas positive emotional states have been hypothesized to broaden attention to the periphery or to task-irrelevant stimuli in an unbiased, flexible, exploratory manner (Todd and Anderson, 2014).

Other evidence further suggests that the affective state of an individual can actually predict visual attention toward stimuli in the environment. One such study induced positive or negative affective states in participants followed by quickly flashing images of facial expressions containing different affective states (scowling, smiling, or neutral) or images of neutral houses to each eye. Affective expression states dominated visual awareness (measured by keystroke response times) compared to the images of houses to participants in negative affective states, while images of smiling faces dominated visual awareness compared to neutral expression or scowling to participants in positive affective states (Anderson et al., 2011).

Research conducted on positive affective states in particular has shown that positive affect leads to the broadening of visual attention through the relaxation of attentional selectivity. Rowe et al. (2007) had participants induced into happy, sad, and neutral affective states complete a flanker task in which they were prompted to identify a central letter that was surrounded by congruent or incongruent similar-looking letters (e.g., 'H and N') and complete a word solving/generation task. Happy participants were able to generate more words in the word problem-solving task compared to those in sad or neutral states, underlining previous work suggesting that positive affective states lead to an increased capacity for creativity (Isen et al., 1987). Additionally, participants in happy moods also demonstrated greater flanker interference, meaning that participants in positive states were slower to respond to the central letter in the task because they were likely attending to the surrounding letters. This decreased response time does provide further evidence for attentional broadening in positive states; however, the researchers do point out that positive moods may only facilitate tasks that require a more global processing style (e.g., problem-solving tasks) but may actually impair tasks requiring selective visual attention. In summary, this work provides evidence for top-down influences of affective states on attention and furthermore suggests that affective states may serve to bias visual attention by shifting the breadth of attention in the environment, particularly while in positive affective states.

Clinical Implications

Affective disorders such as depression and anxiety may also influence attention and cognition, thus providing an additional window onto how emotion and cognition may be related. In research studies, clinically anxious individuals have been found to possess a more extreme attention bias compared to non-anxious individuals in a dot-probe task in that the clinically anxious individuals respond more to mildly threatening emotion words compared to nonanxious individuals. Since fear is theorized to have an evolutionary component, it is reasonable to assume that all individuals will respond to more extreme threat stimuli such as a large spider or a skidding automobile. However, fear turns to anxiety when this 'danger threshold' is exceeded and an individual begins responding with equal fear intensity to nonthreatening stimuli (MacLeod et al., 1986). Additionally, people who have experienced trauma and suffer from posttraumatic stress disorder (PTSD) demonstrate attentional biases toward trauma-related stimuli, including trauma words in a Stroop task (Beck et al., 2001; Buckley et al., 2000; Vythilingam et al., 2007). For example, Beck et al. (2001) administered a Stroop task containing accident words, pain words, positive words, and neutral words to participants who have experienced motor vehicle accidents. Participants who had PTSD symptoms as a result of the accident showed decreased response times to accident and pain-related words in the Stroop task, providing evidence for an attentional bias toward trauma-related stimuli. Furthermore, sadness turns to depression when the 'pause' or cognitive reset that occurs during a typical emotional

experience of sadness becomes prolonged. In these cases, individuals engage in what is termed rumination, or preservative thinking about upsetting symptoms and the causes and consequences of the symptoms without initiating problem-solving strategies that may resolve the distress (McLaughlin and Nolen-Hoeksema, 2011). Clinically depressed individuals have also been shown to possess an attentional bias toward negative information in that they have a difficult time disengaging from negative material once it captures their attention (Gotlib and Joorman, 2010).

Emotion in Cognition across the Life span

Another approach to considering the dynamic and bidirectional interrelationships between emotion and cognition is to investigate changes in the both across development. This has been an area of particular interest since Laura Carstensen proposed a motivational theory, socioemotional selectivity theory, which suggests fundamental shifts in how emotions and cognitions may be linked across adulthood (Carstensen et al., 1999; Carstensen, 2006). According to the theory, limited-time perspective leads individuals to prioritize emotional over other types of goals; in other words, limited time makes people care about feeling good. As older adults have a more limited-time perspective, they are thought to prioritize emotional goals, and this goal shift is hypothesized to lead older adults to display 'age-related positivity effects' in their cognition, focusing attention and memory on positive rather than negative stimuli (see also Reed and Carstensen, 2012). This is then an effect of an emotional goal on cognition and leads to the suggestion that older adults' cognition is directed by their desire to achieve positive emotional states. Interestingly, there is also some evidence that engaging in emotion regulation leads to more downstream cognitive costs for younger rather than older adults (Scheibe and Blanchard-Fields, 2009). While these findings about age differences in emotion–cognition links do not speak directly to the general issues raised above, they suggest that developmental considerations may broaden the scope of research on how emotion and cognition may reciprocally interact in a way that may also feed back to the more general theoretical models that guide researchers in the field.

Conclusions

Clearly, individuals' emotional states are related to how they attend to information in their environment and also influence cognitive processing of information from the environment. There is evidence that specific emotional states impact attention and cognition, and also that more general affective states such as positive and negative affects do as well. To some degree, the type of research that has been done and the questions that have been asked in particular studies vary as a function of the conceptual model of the nature of emotion held (either implicitly or explicitly) by the study authors. Researchers who subscribe to a basic emotions or appraisal approach tend to ask about the cognitive effects of specific emotional states (such as fear, anger, disgust), whereas psychological construction approaches that view core affect as a key psychological primitive would instead focus on how valence and arousal influence attention and cognition. Future work will need to determine how research findings can either be used to test conceptual models against each other or can be integrated across models to provide a more comprehensive account of how emotion, broadly construed, can influence cognition, also broadly construed. That these processes also seem to vary by clinical diagnosis, as well as by age, further suggests the need for a nuanced conceptual model, as well as empirical investigations that consider individual differences as well as the effects of specific emotional or affective states.

See also: Emotion, Perception and Expression of; Emotional Regulation; Emotions, Psychological Structure of; Emotions: Methods of Assessment; Facial Emotion Expression, Individual Differences in; Selfconscious Emotions, Psychology of.

Bibliography

Adolphs, R., 1999. Social cognition and the human brain. Trends in Cognitive Sciences 3, 469–479. http://dx.doi.org/10.1016/S1364-6613(99)01399-6.

Algoe, S.B., 2012. Find, remind, and bind: the functions of gratitude in everyday relationships. Social and Personality Psychology Compass 6, 455–469. http://dx.doi.org/10.1111/j.1751-9004.2012.00439.x.

Algoe, S.B., Fredrickson, B.L., Gable, S.L., 2013. The social functions of the emotion of gratitude via expression. Emotion 13, 605–609. http://dx.doi.org/10.1037/a0032701.

Anderson, E., Siegel, E.H., Barrett, L., 2011. What you feel influences what you see: the role of affective feelings in resolving binocular rivalry. Journal of Experimental Social Psychology 47, 856–860. http://dx.doi.org/10.1016/j.jesp.2011.02.009.

Angie, A.D., Connelly, S., Waples, E.P., Kligyte, V., 2011. The influence of discrete emotions on judgment and decision-making: a meta-analytic review. Cognition and Emotion 25, 1393–1422.

Arnold, M.B., 1960. Emotion and Personality. Columbia University Press, New York, NY, USA.

Barrett, L., 2006. Solving the emotion paradox: categorization and the experience of emotion. Personality and Social Psychology Review 10, 20–46. http://dx.doi.org/10.1207/s15327957pspr1001_2.

Barrett, L.F., 2009. The future of psychology: connecting mind to brain. Perspectives in Psychological Science 4, 326–339.

Barrett, L., 2009. Variety is the spice of life: a psychological construction approach to understanding variability in emotion. Cognition and Emotion 23, 1284–1306. http://dx.doi.org/10.1080/02699930902985894.

Barrett, L.F., 2013. Psychological construction: a Darwinian approach to the science of emotion. Emotion Review 5, 379–389.

Beck, J., Freeman, J.B., Shipherd, J.C., Hamblen, J.L., Lackner, J.M., 2001. Specificity of Stroop interference in patients with pain and PTSD. Journal of Abnormal Psychology 110, 536–543. http://dx.doi.org/10.1037/0021-843X.110.4.536.

Bonanno, G.A., Keltner, D., 1997. Facial expressions of emotion and the course of conjugal bereavement. Journal of Abnormal Psychology 106 (1), 126–137. http://dx.doi.org/10.1037/0021-843X.106.1.126.

Buckley, T.C., Blanchard, E.B., Neill, W., 2000. Information processing and PTSD: a review of the empirical literature. Clinical Psychology Review 20, 1041–1065. http://dx.doi.org/10.1016/S0272-7358(99)00030-6.

Carstensen, L.L., 2006. The influence of a sense of time on human development. Science 312, 1913–1915. http://dx.doi.org/10.1126/science.1127488.

Carstensen, L.L., Isaacowitz, D.M., Charles, S.T., 1999. Taking time seriously: a theory of socioemotional selectivity. American Psychologist 54, 165–181.

Clore, G.L., Ortony, A., 2008. Appraisal theories: how cognition shapes affect into emotion. In: Lewis, M., Haviland-Jones, J.M., Barrett, L. (Eds.), Handbook of Emotions, third ed. Guilford Press, New York, NY, USA, pp. 628–642.

Cunningham, W.A., Zelazo, P.D., 2007. Attitudes and evaluations: a social cognitive neuroscience perspective. Trends in Cognitive Sciences 11, 97–104.

Cunningham, W.A., Dunfield, K.A., Stillman, P.E., 2013. Emotional states from affective dynamics. Emotion Review 5, 344–355.

De Rivera, J., Possel, L., Verette, J.A., Weiner, B., 1989. Distinguishing elation, gladness, and joy. Journal of Personality and Social Psychology 57, 1015–1023.

Ekman, P., 1972. Universals and cultural differences in facial expressions of emotion. In: Cole, J. (Ed.), Nebraska Symposium on Motivation, vol. 19. University of Nebraska, Lincoln, pp. 207–282.

Ekman, P., Davidson, R.J., 1994. The Nature of Emotion: Fundamental Questions. Oxford University Press, New York, NY, USA.

Ekman, P., Sorenson, E., Friesen, W.V., 1969. Pan-cultural elements in facial displays of emotion. Science 164, 86–88. http://dx.doi.org/10.1126/science.164.3875.86.

Ekman, P., 1992. Are there basic emotions? Psychological Review 99, 550–553. http://dx.doi.org/10.1037/0033-295X.99.3.550.

Ekman, P., Cordaro, D., 2011. What is meant by calling emotions basic. Emotion Review 3, 364–370. http://dx.doi.org/10.1177/1754073911410740.

Ellsworth, P.C., Smith, C.A., 1988. Shades of joy: patterns of appraisal differentiating pleasant emotions. Cognition and Emotion 2, 301–331.

Eskine, K.J., Kacinik, N.A., Prinz, J.J., 2011. A bad taste in the mouth: gustatory disgust influences moral judgment. Psychological Science 22, 295–299. http://dx.doi.org/10.1177/0956797611398497.

Frijda, N., 1986. The Emotions. Cambridge University Press, Cambridge, UK.

Fetterman, A.K., Robinson, M.D., Gordon, R.D., Elliot, A.J., 2011. Anger as seeing red: perceptual sources of evidence. Social Psychological and Personality Science 2, 311–316. http://dx.doi.org/10.1177/1948550610390051.

Fredrickson, B.L., 1998. What good are positive emotions? Review of General Psychology 2, 300–319. http://dx.doi.org/10.1037/1089-2680.2.3.300.

Fredrickson, B.L., 2001. The role of positive emotions in positive psychology: the broaden-and-build theory of positive emotions. American Psychologist 56, 218–226. http://dx.doi.org/10.1037/0003-066X.56.3.218.

Gasper, K., Clore, G.L., 2002. Attending to the big picture: mood and global versus local processing of visual information. Psychological Science 13, 34–40.

Gendron, M., Barrett, L., 2009. Reconstructing the past: a century of ideas about emotion in psychology. Emotion Review 1, 316–339. http://dx.doi.org/10.1177/1754073909338877.

Gotlib, I.H., Joorman, J., 2010. Cognition and depression: current status and future directions. Annual Review of Clinical Psychology 6, 285–312.

Gross, J.J., Barrett, L., 2011. Emotion generation and emotion regulation: one or two depends on your point of view. Emotion Review 3, 8–16. http://dx.doi.org/10.1177/1754073910380974.

Isen, A.M., 1987. Positive affect, cognitive processes, and social behavior. Advances in Experimental Social Psychology 20, 203–253.

Isen, A.M., Daubman, K.A., Nowicki, G.P., 1987. Positive affect facilitates creative problem solving. Journal of Personality and Social Psychology 52, 1122–1131.

Izard, C.E., 1977. Human Emotions. Plenum Press, New York.

Izard, C.E., 1993. Organizational and motivational functions of discrete emotions. In: Lewis, M., Haviland, J.M. (Eds.), Handbook of Emotions. Guilford Press, New York, NY, USA, pp. 631–641.

Izard, C.E., 2011. Forms and functions of emotions: matters of emotion-cognition interactions. Emotion Review 3, 371–378.

Keltner, D., Haidt, J., 2003. Approaching awe, a moral, spiritual, and aesthetic emotion. Cognition And Emotion 17, 297–314. http://dx.doi.org/10.1080/02699930302297.

Lazarus, R.S., 1991. Emotion and Adaptation. Oxford University Press, New York.

Lazarus, R.S., 1998. Coping from the perspective of personality. Zeitschrift Für Differentielle Und Diagnostische Psychologie 19, 213–230.

Lazarus, R.S., Lazarus, B.N., 1994. Passion and Reason: Making Sense of Our Emotions. Oxford University Press, New York.

LeDoux, J.E., 2000. Emotion circuits in the brain. Annual Review of Neuroscience 23, 155–184. http://dx.doi.org/10.1146/annurev.neuro.23.1.155.

Levenson, R.W., 2011. Basic emotion questions. Emotion Review 3, 379–386.

Lindquist, K.A., 2013. Emotions emerge from more basic psychological ingredients: a modern psychological constructionist model. Emotion Review 5, 356–368.

MacLeod, C., Mathews, A., Tata, P., 1986. Attentional bias in emotional disorders. Journal of Abnormal Psychology 95, 15–20.

Martin, R.C., Dahlen, E.R., 2007. The angry cognitions scale: a new inventory for assessing cognitions in anger. Journal of Rational-Emotive & Cognitive-Behavior Therapy 25, 155–173. http://dx.doi.org/10.1007/s10942-006-0033-2.

McCullough, M.E., Kilpatrick, S.D., Emmons, R.A., Larson, D.B., 2001. Is gratitude a moral affect? Psychological Bulletin 127, 249–266. http://dx.doi.org/10.1037/0033-2909.127.2.249.

McLaughlin, K.A., Nolen-Hoeksema, S., 2011. Rumination as a transdiagnostic factor in depression and anxiety. Behaviour Research and Therapy 49, 186–193. http://dx.doi.org/10.1016/j.brat.2010.12.006.

Moors, A., 2013. On the causal role of appraisal in emotion. Emotion Review 5, 132–140.

Oatley, K., Johnson-Laird, P.N., 1996. The communicative theory of emotions: empirical tests, mental models, and implications for social interaction. In: Bonanno, G.A., Goorin, L., Coifman, K.G. (Eds.), Sadness and Grief. The Guilford Press, Inc., New York, pp. 797–810.

Öhman, A., Flykt, A., Esteves, F., 2001. Emotion drives attention: detecting the snake in the grass. Journal of Experimental Psychology: General 130, 466–478.

Panksepp, J., 1998. Affective Neuroscience: The Foundations of Human and Animal Emotions. Oxford University Press, New York, NY, USA.

Panksepp, J., Watt, D., 2011. What is basic about basic emotions? Lasting lessons from affective neuroscience. Emotion Review 3, 387–396.

Pessoa, L., Padmala, S., Kenzer, A., Bauer, A., 2012. Interactions between cognition and emotion during response inhibition. Emotion 12, 192–197. http://dx.doi.org/10.1037/a0024109.

Piaget, J., Inhelder, B., 1969. The Psychology of the Child (H. Weaver, Trans.). Basic Books, New York (Original work published 1966).

Pizarro, D.A., Inbar, Y., Helion, C., 2011. On disgust and moral judgment. Emotion Review 3, 267–268. http://dx.doi.org/10.1177/1754073911402394.

Reed, A.E., Carstensen, L.L., 2012. The theory behind the age-related positivity effect. Frontiers in Psychology 3. http://dx.doi.org/10.3389/fpsyg.2012.00339.

Rowe, G.G., Hirsh, J.B., Anderson, A.K., Smith, E.E., 2007. Positive affect increases the breadth of attentional selection. Proceedings of the National Academy of Sciences of the United States of America 104, 383–388. http://dx.doi.org/10.1073/pnas.0605198104.

Scheibe, S., Blanchard-Fields, F., 2009. Effects of regulating emotions on cognitive performance: what is costly for young adults is not so costly for older adults. Psychology and Aging 24, 217–223. http://dx.doi.org/10.1037/a0013807.

Scherer, K.R., 1984. Emotion as a multicomponent process: a model and some cross-cultural data. Review of Personality & Social Psychology 537–563.

Scherer, K.R., 2001. Appraisal considered as a process of multilevel sequential checking. In: Scherer, K.R., Schorr, A., Johnstone, T. (Eds.), Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, New York, NY, USA, pp. 92–120.

Scherer, K.R., 2013. The nature and dynamics of relevance and valence appraisals: theoretical advances and recent evidence. Emotion Review 5, 150–162. http://dx.doi.org/10.1177/1754073912468166.

Schnall, S., Haidt, J., Clore, G.L., Jordan, A.H., 2008. Disgust as embodied moral judgment. Personality and Social Psychology Bulletin 34 (8), 1096–1109. http://dx.doi.org/10.1177/0146167208317771.

Schwarz, N., 2010. Meaning in context: metacognitive experiences. In: Mesquita, B., Barrett, L., Smith, E.R. (Eds.), The Mind in Context. Guilford Press, New York, NY, USA, pp. 105–125.

Schwarz, N., Clore, G.L., 1983. Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. Journal of Personality and Social Psychology 45, 513–523. http://dx.doi.org/10.1037/0022-3514.45.3.513.

Susskind, J.M., Lee, D.H., Cusi, A., Feiman, R., Grabski, W., Anderson, A.K., 2008. Expressing fear enhances sensory acquisition. Nature Neuroscience 11, 843–850. http://dx.doi.org/10.1038/nn.2138.

Tiedens, L.Z., Linton, S., 2001. Judgment under emotional certainty and uncertainty: the effects of specific emotions on information processing. Journal of Personality and Social Psychology 81, 973–988. http://dx.doi.org/10.1037/0022-3514.81.6.973.

Todd, R.M., Anderson, A.K., 2014. Salience, state, and expression: the influence of specific aspects of emotion on attention and perception. In: Ochsner, K.N., Kosslyn, S.M. (Eds.), The Oxford Handbook of Cognitive Neuroscience, vol. 2: The Cutting Edges. US: Oxford University Press, New York, pp. 11–31.

Tracy, J.L., Matsumoto, D., 2008. The spontaneous expression of pride and shame: evidence for biologically innate nonverbal displays. Proceedings of the National Academy of Sciences of the United States of America 105, 11655–11660.

Tracy, J.L., Randles, D., 2011. Four models of basic emotions: a review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. Emotion Review 3, 397–405. http://dx.doi.org/10.1177/1754073911410747.

Tybur, J.M., Lieberman, D., Kurzban, R., DeScioli, P., 2013. Disgust: evolved function and structure. Psychological Review 120, 65–84. http://dx.doi.org/10.1037/a0030778.

Vythilingam, M.M., Blair, S.S., McCaffrey, D.D., Scaramozza, M.M., Jones, M.M., Nakic, M.M., Blair, R.R., 2007. Biased emotional attention in post-traumatic stress disorder: a help as well as a hindrance? Psychological Medicine 37, 1445–1455. http://dx.doi.org/10.1017/S003329170700092X.

Wadlinger, H.A., Isaacowitz, D.M., 2006. Positive mood broadens visual attention to positive stimuli. Motivation and Emotion 30, 89–101.

Welling, H., 2003. An evolutionary function of the depressive reaction: the cognitive map hypothesis. New Ideas in Psychology 21 (2), 147–156. http://dx.doi.org/10.1016/S0732-118X(03)00017-5.

Wilson-Mendenhall, C.D., Barrett, L., Simmons, W., Barsalou, L.W., 2011. Grounding emotion in situated conceptualization. Neuropsychologia 49, 1105–1127. http://dx.doi.org/10.1016/j.neuropsychologia.2010.12.032.

Emotion, Evolution of

Elena Walsh and Paul E Griffiths, University of Sydney, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P.E. Griffiths, volume 7, pp. 4433–4437, © 2001, Elsevier Ltd.

Abstract

Here we outline the development of the evolutionary study of emotion. We begin with the work of Charles Darwin, who initiated the study of the evolution of emotion. Since then, at least three different approaches to the subject have emerged in the literature. The orthodox 'basic emotions' or 'affect program' view developed in the 1960s out of human ethology and psychological and physiological studies of emotion. Both the Evolutionary Psychology movement and the transactional approach to emotion developed in the 1960s and 1970s, amending the orthodox approach in different ways. In recent years, particular attention has been paid to the evolution of moral emotions, such as shame, pride, and moral disgust.

The Darwinian Legacy

Darwin did not see his *Expression of the Emotions in Man and Animals* (1872) as a theory of the evolution of emotion itself, because he defined emotions as subjective feelings. Following Herbert Spencer, Darwin distinguished emotions as feelings caused by external states, in contrast to feelings caused by bodily states, such as hunger and pain. Later evolutionary theorists have preferred behavioral and physiological definitions of emotion, with the result that Darwin's theory of the expression of emotions has been interpreted as a theory of the emotions themselves.

Darwin proposed three principles of the evolution of expression. The most important of these was the 'principle of serviceable associated habits,' which is a straightforward application of Darwin's theory of instincts to the case of emotion. Darwin believed that instinctive behaviors derive from habits. The consistent acquisition of the same habit for many generations causes it to become hereditary, or instinctive, as a result of the inheritance of acquired characteristics, of which Darwin was a firm believer. Most of the distinctive behaviors associated with particular emotions, such as the erection of the hair in fear, reflect long since vanished lifestyles in which those behaviors were rewarded and reinforced in each generation until they were finally incorporated into the hereditary material as instincts. Darwin supplements this principle with two others, the 'principle of antithesis' and the 'principle of direct action.' The antithesis principle postulates an intrinsic tendency for opposite states of feeling to produce opposite behaviors. Darwin remarks of a submissive dog that,

> Not one of the movements, so clearly expressive of affection, are of the least direct service to the animal. They are explicable, as far as I can see, solely from their being in complete opposition or antithesis to the attitude expressive of anger. (Darwin, 1872: p. 51)

Darwin explains the behaviors left over after the application of these two principles as the results of the 'direct action' of the nervous system. Excess nerve energy built up in an emotional episode is released in behaviors such as sweating and trembling for no other reason than that it must go somewhere and that these channels are physiologically available for its release.

The Ethological Tradition

The founders of ethology saw themselves as the direct heirs of Darwin's work on mental evolution (Lorenz, 1965). They retained Darwin's principles, but reinterpreted them to fit the theory of evolution as it emerged in the 1930s in the 'modern synthesis' of Darwinism and Mendelian genetics. The principle of serviceable associated habits was transformed into the ethological concepts of 'ritualization' and 'derived activity' (Tinbergen, 1952). Derived activities are behaviors that originally evolved for one purpose but were later selected for another purpose. Ritualized behaviors are derived activities that originally evolved to fulfill some practical function but which were later selected to function as signals. Thus, although piloerection in fear and rage does little to make a human being appear larger to an opponent, it does communicate their emotional state. Derived activities require a distinctive, two-stage form of evolutionary explanation. They cannot be understood purely in terms of the function they currently perform and the selection pressures that currently maintain them in the population. This is particularly obvious in the case of signals. Piloerection is not intrinsically better as a signal of fear than smiling or laughing. This particular behavior was selected as a signal only because it was already associated with fear in the distant past. It was not associated with fear because it was a signal of fear, but because it made the animal appear larger.

The concept of ritualization allowed ethology to reconstruct Darwin's principle of serviceable associated habits while avoiding his commitment to the inheritance of acquired characteristics. Darwin's descriptions of the psychological rewards that led to the reinforcement of emotional behaviors are equally plausible as descriptions of the original selective advantage of those behaviors. Darwin's other two principles were similarly open to reinterpretation. The principle of antithesis was explained by the selective value of unambiguous signals: it may be important for an animal to signal clearly that it is not aggressive. Hence there can be selection of behaviors

merely because they look different from the behaviors that signal aggression. The principle of direct action was transformed into the ethological concept of displacement activity. Early ethologists shared Darwin's view that instinctive motivations cause a build up of mental energy that must be released in some behavior or other. Displacement activities "serve a function as outlets, through a safety valve, of dangerous surplus impulses" (Tinbergen, 1952: p. 23). An example commonly given is that of an angry cat that is unwilling to attack and begins to wash itself. This wholesale reinterpretation of Darwin's three principles worked so smoothly and allowed the retention of so much of the detail of Darwin's work that the early ethologists seemed almost unaware of the differences between Darwin's theory and their own (Lorenz, 1965).

The Affect Program Theory of Basic Emotions

Darwin's detailed accounts of human facial expressions were confirmed and extended during the 1960s and early 1970s and many of his suggested homologies between human and primate facial expressions were confirmed. It became generally accepted that some emotional expressions, the so-called basic emotions, are found in all cultures and have an evolutionary history at least as long as the history of the human species (Ekman, 1973). The basic emotions are usually referred to as fear, anger, disgust, contempt, joy, sadness, and surprise, although in normal usage each of these words has a much wider meaning. During the same period, the hydraulic metaphors that the ethologists had used to describe psychological processes were replaced with a computational model of emotion. Each basic emotion corresponds to an 'affect program' stored somewhere in the brain. When activated, this program coordinates a complex of actions that include facial expression, autonomic nervous system changes, expressive vocal changes, and muscular–skeletal responses such as flinching or orienting. The concept of an affect program inherits many of the features of the ethological concept of a 'fixed action pattern.' Both concepts suggest that certain apparently complex behaviors are, in reality, atomic units of behavior that unfold in the same, stereotyped sequence whenever they are triggered. The affect programs are controlled by an 'automatic appraisal mechanism.' This is a specialized neural system that applies its own distinctive rules for stimulus evaluation to a limited set of data derived from the earliest stages of the processing of perceptual information. Considered together, the appraisal mechanisms and affect programs form a cognitive 'module' in the sense favored by more recent evolutionary psychologists (Barkow et al., 1992).

The affect program theory retains two distinctive features of the earlier, ethological model of emotions. First, emotional behaviors are an inevitable and involuntary accompaniment to the underlying emotional state and as such are reliable signals of that state. Second, deliberate attempts to disguise emotions are subject to 'leakage' from the operation of the involuntary emotional response. Voluntary attempts to suppress emotional behavior can only operate by simultaneously using the muscles involved in the expression for some other purpose. They cannot interfere with the actual operation of the emotion module.

The affect program theory was accompanied by an account of the evolution of the emotion system as an ancient form of cognition that had originally operated on its own and later been supplemented by higher cognitive functions. This view was supported by MacLean's theory of the 'triune brain,' according to which the emotions are located in the 'paleomammalian' portions of the brain, while higher cognitive functions are realized in newer, 'neomammalian' structures (MacLean, 1952). The survival of these ancient forms of behavior control in primates was explained by their value as fail-safe responses ensuring that vital behaviors are performed whenever necessary even if that means they are performed too often. Although MacLean's 'triune brain' is no longer accepted in its original form, this view of the emotion system as a collection of primitive but reliable fail-safe mechanisms remains influential (Panksepp, 1998).

Sociobiology and the Emotions

Sociobiology brought a new perspective to bear on the evolution of emotion in the 1970s and 1980s. It also moved the focus of investigation from the basic emotions to the moral and quasimoral emotions involved in human social interaction. Emotions such as trust, loyalty, guilt, and shame play an obvious role in mediating the competitive social interactions that were the focus of most research in human sociobiology. Numerous sociobiologists made brief comments to the effect that moral emotions must have evolved as psychological mechanisms to implement evolutionarily stable strategies of social interaction (Weinrich, 1980). Robert A. Frank (1988) suggested that the moral emotions evolved as solutions to 'commitment problems.' A commitment problem arises when the winning strategy in an evolutionary interaction involves making a binding but conditional commitment to do something that would be against one's own interests if the condition were ever met. If such a commitment is to be credible, some special mechanism is needed that would cause the organism to act against its own interests. Frank suggests that emotions such as rage and vengefulness evolved to allow organisms to engage in credible deterrence, threatening self-destructive aggression to deter a more powerful aggressor. Conversely, emotions such as love and guilt evolved to allow organisms to engage in reciprocal altruism in situations where no retaliation is possible if one partner fails to reciprocate.

Sociobiologists criticized ethology because it lacked a theoretical framework with which to predict how humans would behave, accusing it of being little more than descriptive natural history (Barash, 1979; Barkow, 1979). In contrast, sociobiology seemed to make strong predictions that clashed with some aspects of the affect program theory of basic emotions. From a sociobiological perspective it makes no sense for organisms to possess involuntary expressive behavior that freely communicates information about the animal's motivational state to anyone who has a use for it. The application of evolutionary game theory to emotional behavior predicts that this behavior will be designed to manipulate the expectations of other organisms rather than to transparently 'express' emotional states. This theoretical argument has been used by some to reject the affect program view of basic emotions.

Alan Fridlund (1994) has argued that emotional behavior should be regarded as a paralanguage of social signals whose production depends at least as much on an organism's social context as on its emotional state. Fridlund and other paralanguage theorists have documented 'audience effects' on the production of the basic emotions and have argued that this is inconsistent with the affect program theory. They have also tried to show that the classical facial expressions of emotion are not strongly correlated with the underlying emotional state and that people rely more on context than on behavior in attributing emotions to others (Russell and Fernández-Dols, 1997). It is unclear, however, that this data contradicts the predictions of the affect program theory. Signaling behavior exhibits audience effects in many organisms, such as domestic chickens, in which that behavior is presumably part of a highly stereotyped behavior sequence controlled by a relatively simple cognitive mechanism.

Mark Hauser has suggested that Fridlund's arguments bear on questions about the biological function of emotional behavior while the affect program model is concerned with the mechanisms that produce this behavior (Hauser, 1996: pp. 495–496). In some places, however, Fridlund does seem to be discussing the nature of the underlying emotional processes and not merely their biological function. Fridlund argues that since emotions are signals produced because of the impact they will have on other organisms, a particular emotion defined behaviorally need not correspond to any particular motivational state (emotion) of the organism producing it.

Evolutionary Psychology and the Emotions

The Evolutionary Psychology movement of the late 1980s and 1990s has been less inclined than sociobiology to question the accepted view of basic emotions. The evolutionary psychologists believe that the mind as a whole is made up of modules that operate independently of one another. Like the emotion system, these modules operate on specific kinds of data using algorithms that differ from those used by other modules. Hence, evolutionary psychology endorses the affect program theory of basic emotions, but wants to go further, both by adding to the complexity of the known affect programs and by finding modular mechanisms underlying other emotional behaviors. John Tooby and Leda Cosmides (1990) suggest a number of cognitive and behavioral systems that they would expect to receive output from the affect programs, in addition to those currently known to receive such output. They also suggest that the affect programs will be elicited by simple, perceptual stimuli that have played a recurrent role in human evolutionary history. Current empirical research suggests, however, that few simple perceptual stimuli elicit affect programs in the absence of any individual experience of those stimuli. Most evolutionary influences on sensitivity to simple perceptual stimuli seem to take the form of learning preparedness – the relative ease with which one association is acquired rather than another. Genuine pancultural antecedents of emotion seem to be relatively abstract states of affairs having to do with the significance of a situation for the organism and requiring local knowledge to translate them into perceptible features of the stimulus situation (Ekman and Davidson, 1994: Section 4, pp. 144–177).

The leading evolutionary psychologist David Buss has argued for the existence of a module for sexual jealousy, one of the additional modules predicted by Tooby and Cosmides. Buss argues that sexual jealousy has simple perceptual elicitors such as unusual scents, changed sexual behavior, excessive eye contact, and violation of rules governing personal space. Buss claims that "these signals alert us to the possibility of infidelity, since they have been statistically linked with relationship loss over the long course of human evolutionary history" (2000: p. 45). The jealousy module uses special-purpose algorithms and, like the basic emotions, it functions as a fail-safe mechanism: "Failing to detect an infidelity was more costly than mistakenly accusing an innocent partner of betrayal. Evolution, as a consequence, forged a hypersensitive defense system, designed to sound the alarm not just when an infidelity has been discovered, but also when the circumstances make it slightly more likely" (Buss, 2000: p. 224). When triggered, the module produces various forms of violence against female sexual partners, including, under certain environmental conditions, murder. A noticeable contrast between these recent theories and more traditional accounts of the evolution of emotion is the absence of the idea that emotions represent a more primitive form of behavioral control that can be contrasted to rational, planned action. The emotions are seen as just another cognitive module reflecting details of the environment of evolutionary adaptedness.

Evolution of the Moral Emotions

The important domain of human behavior that we call 'morality' is closely related to a number of emotions, sometimes collectively referred to as the moral emotions. These include guilt, shame, pride, embarrassment, indignation, and moral disgust. The role of these emotions in morality is part of 'moral psychology,' a field that bridges philosophy and psychology. Many of the research traditions discussed above have had views on the evolution of the moral emotions, but recent years have seen an increase in interest, particularly in moral disgust, shame, and pride.

Daniel Kelly (2011) presents a synthesis of much recent research on disgust. He argues that this emotion combines a taste-aversion mechanism with a mechanism for avoiding infection and parasitism. Kelly argues that disgust is uniquely human. It is an example of the co-option and developmental entanglement that occurs as an evolving lineage confronts new adaptive challenges and must solve them by modifying existing features. Both elements of the disgust response have homologues or analogs in other species. It is their entanglement that creates a uniquely human emotion. Moral disgust represents a further piece of evolutionary tinkering in which the emotion of disgust was co-opted to help enforce conformity with in-group norms.

Although competing explanations of the evolution of pride and shame are extant in the literature, a consensus is emerging. Behaviors homologous to those we associate with pride and shame are components of displays employed by nonhuman

primates (and other mammals) during the negotiation or affirmation of relative rank. We can label the emotions that underlie such displays as 'protopride' and 'protoshame.' In a competitive environment, the expression of protoshame signals the actor's acceptance of a subordinate position in a hierarchy. It functions to deflect aggression from the dominant party. The expression of protopride, in contrast, signals the actor's dominance. Importantly, protopride is a pleasant feeling, while protoshame is unpleasant. Daniel Fessler (1999) argues that protopride and protoshame were selected for in primates because they promoted dominance-seeking behaviors, which increased the chances of reproductive success. These emotions are the phylogenetic ancestors of pride and shame in human beings (Fessler, 1999, 2004). However, Jason A. Clark (2012) has argued that humans retain the ability to experience protopride and protoshame alongside their more sophisticated descendants, pride and shame.

Pride and shame are more cognitively complex than protopride and protoshame. They require that the organism has 'theory of mind' – the capacity to attribute mental states to other individuals. They also serve a different adaptive function in the relatively cooperative, nonhostile environments human beings inhabit. In such contexts, pride and shame are experienced in response to the actual or imagined assessment of an actor's behavior by an audience. If the assessment inspires the contempt or pity of the audience, the actor feels shame. If the assessment inspires the envy or admiration of the audience, the actor feels pride. In the cooperative context, then, pride and shame encourage the actor to seek the admiration and envy of others, which in turn encourages them to act in ways that support the well-being of the group (Fessler, 1999). In this way, these emotions inspire adherence to socially sanctioned norm-governed behavior.

There is broad agreement that the capacity to adhere to norms is an evolved capacity in humans that requires, among other things, the capacity for appropriate emotional responses. Whether this means that the capacity for morality has itself evolved is a matter of some controversy. Jesse Prinz (2007) is a prominent exponent of the view that morality has not evolved. He opposes nativists who think humans are born with a set of universal moral principles like *'Don't harm innocent people*.' However, most self-described moral nativists agree with Prinz that it is unlikely that universal moral judgments of this sort evolved. What such nativists do endorse is the more plausible view that the *capacity* to form moral judgments is the result of natural selection (Joyce, 2006). Evidence supporting nativism of this latter sort frequently appeals to the universality of moral systems in human cultures. All cultures have moralities, even if those cultures disagree about which universal moral principles are correct. Nativists also point to evidence that young children appear to recognize what is distinctive about moral rules (as opposed to merely conventional rules) without formal tutelage. Prinz (2007) argues that this latter piece of evidence fails to consider the rich resources of the learning environment. He suggests that the transgression of moral rules is punished more severely than the transgression of merely conventional rules. As a result, children learn to associate emotions like guilt and shame with the transgression of moral norms, and in this way become able to discern the distinctive character of moral judgments.

Jonathan Haidt and Craig Joseph (2007) have recently outlined a nativist explanation of the origins of five sets of moral concerns: harm, justice, loyalty, authority, and sanctity. Haidt and Joseph argue that humans have learning modules responsible for emotions whose appearance depends on the detection of moral patterns in the social world. They think that each learning module corresponds to one of the five sets of moral concerns, but that the specific content, and the relative importance, of each moral concern is modified by experience. Their work moves away from moral psychology's overarching focus on the moral concerns of modern, secular, liberal, and Western societies (namely, harm and justice). It advocates inclusion of the moral concerns of non-Western and traditional cultures and Western religious conservative culture (namely, loyalty, authority, and sanctity).

The dispute between nativists and nonnativists about the evolution of morality, like most nature/nurture disputes, is made more complex by the multiple ambiguities of the term 'innate' (Mameli and Bateson, 2006). A trait may be 'innate' in the sense of being designed by natural selection, but nevertheless depend on features of the normal developmental environment, such as normal parenting, if it is to develop. So even if some aspect of morality is not 'innate' in the sense that its development requires the right kind of environment, it may still be the result of adaptive evolution. Natural selection can also produce phenotypic plasticity, in which organisms develop a range of different traits depending on the particular environments in which they develop. So even if some aspect of morality is not 'innate' in the sense that it is part of a universal human nature, the range of features that develop in different environments may still be the result of adaptive evolution. Distinguishing the question of whether morality has been designed by natural selection from these and other alternative senses in which morality might be 'innate' is essential when assessing the evidence for the evolution of different aspects of morality and of the moral emotions. Haidt and Joseph try to finesse this problem by defining 'innate' with the apparently minimal condition of "organized, to some extent, in advance of experience" (2007: p. 367), but this still does not separate the phylogenetic issue of evolutionary design from the ontogenetic issue of the role of experience.

Conclusion

Theories of the evolution of emotion have an intellectual lineage stretching back to Darwin himself. The development of the field has reflected more general trends in the history of Darwinian approaches to mind and behavior. Important current issues include the number of separate, evolved emotions; the nature of the stimuli that produce evolved emotions; and a cluster of issues about the evolutionary function of emotional behaviors and their relationship to underlying emotional states, issues which are often discussed in terms of whether emotional behaviors should be primarily regarded as expressions or as signals. Perhaps the most significant recent development in the field is an upsurge of

work on the evolution of moral emotions. This article has substantial implications for ideas about the evolution of morality itself.

See also: Cognition, Evolution of; Darwinism; Disgust, Psychology of; Emotion and Expression; Evolutionary Social Psychology; Mind, Theories of; Sociobiology: Overview.

Bibliography

Barash, D.P., 1979. Human ethology and human sociobiology. Behavioral and Brain Sciences 2 (1), 26–27.

Barkow, J.H., 1979. Human ethology: empirical wealth, theoretical dearth. Behavioral and Brain Sciences 2 (1), 27.

Barkow, J.H., Cosmides, L., Tooby, J. (Eds.), 1992. The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, Oxford.

Buss, D.M., 2000. The Dangerous Passion. Simon and Schuster, New York.

Clark, J., 2012. Integrating basic and higher-cognitive emotions within a common evolutionary framework: lessons from the transformation of primate dominance into human pride. Philosophical Psychology 26 (3), 437–460.

Darwin, C., 1872. The Expressions of Emotions in Man and Animals. Philosophical Library, New York.

Ekman, P. (Ed.), 1973. Darwin and Facial Expression: A Century of Research in Review. Academic Press, New York and London.

Ekman, P., Davidson, R.J. (Eds.), 1994. The Nature of Emotion: Fundamental Questions. Oxford University Press, New York.

Fessler, D.M.T., 1999. Toward an understanding of the universality of second order emotions. In: Hinton, A. (Ed.), Beyond Nature or Nurture: Biocultural Approaches to the Emotions. Cambridge University Press, New York, pp. 75–116.

Fessler, D.M.T., 2004. Shame in two cultures: implications for evolutionary approaches. Journal of Cognition and Culture 4, 207–262.

Frank, R.H., 1988. Passions within Reason: The Strategic Role of the Emotions. Norton, New York.

Fridlund, A., 1994. Human Facial Expression: An Evolutionary View. Academic Press, San Diego, CA.

Haidt, J., Joseph, C., 2007. The moral mind: how five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In: Carruthers, P., Laurence, S., Stich, S. (Eds.), The Innate Mind, Foundations and the Future, vol. 3. Oxford University Press, New York, pp. 367–392.

Hauser, M.D., 1996. The Evolution of Communication. MIT Press, Cambridge, MA.

Joyce, R., 2006. The Evolution of Morality. MIT Press, Cambridge, MA.

Kelly, D., 2011. Yuck! The Nature and Moral Significance of Disgust. MIT Press, Cambridge, MA.

Lorenz, K., 1965. Preface. In: Darwin, C. (Ed.), The Expression of the Emotions in Man and Animals. University of Chicago, Chicago, IL.

MacLean, P.D., 1952. Some psychiatric implications of physiological studies on frontotemporal portions of the limbic system (visceral brain). Electroencephlaography & Clinical Neurophysiology 4, 407–418.

Mameli, M., Bateson, P.P.G., 2006. Innateness and the sciences. Biology and Philosophy 21 (2), 155–188.

Panksepp, J., 1998. Affective Neuroscience: The Foundations of Human and Animal Emotions. Oxford University Press, Oxford and New York.

Prinz, J., 2007. The Emotional Construction of Morals. Oxford University Press, New York.

Russell, J.A., Fernández-Dols, J.M., 1997. The Psychology of Facial Expression. Cambridge University Press, Cambridge.

Tinbergen, N., 1952. Derived activities: their causation, biological significance, origin and emancipation during evolution. Quarterly Review of Biology 27 (1), 1–32.

Tooby, J., Cosmides, L., 1990. The past explains the present: emotional adaptations and the structure of ancestral environments. Ethology and Sociobiology 11, 375–424.

Weinrich, J.D., 1980. Towards a sociobiological theory of emotions. In: Plutchik, R., Kellerman, H. (Eds.), Emotion: Theory, Research and Experience, Theories of Emotion, vol. 1. Academic Press, New York, pp. 113–140.

Emotion: History of the Concept

Katharina Scherke, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Mandler, volume 7, pp. 4437–4440, © 2001, Elsevier Ltd.

Abstract

In the past decades, in several scientific disciplines we observe growing interest in emotion – from neurobiology, philosophy, history, and anthropology to as far as sociology. In the course of the twentieth century, under the influence of a rationalist concept of man pushing through, the value of emotions had declined in the social sciences. Contrasting rationality and emotionality as mutually excluding attitudes looks back to a long history and has only recently increasingly been questioned. The development of this rationalist attitude is sketched in this article. In addition, the relationship between physiological aspects, cognitive appraisal, and emotion is described (most of all in reference to psychological literature and those of the neurosciences) and an overview of current sociology of emotions is given.

In the past decades, in several scientific disciplines we observe a growing interest in emotion. This covers a whole range – from neurobiology, philosophy, history, anthropology, as far as to sociology. If we consider the history of emotion research, it is conspicuous that over long periods emotions were considered the realm of individual disciplines (such as psychology) and were, thus, in tendency neglected by other disciplines (Scherke, 2009: 11–15). Only recently, an attitude has become more prevailing, which assumes that for the understanding of human emotions we need an interdisciplinary approach. In this article, we present some important steps of the historical development of emotion being a concept in the social sciences (sociology, in particular) – with a special focus also on the changing importance of this concept for various disciplines.

Already, the classical authors of sociology pay attention to human emotions, as has been shown in recent years by referring to the works of people such as Max Weber (1864–1920), Emile Durkheim (1858–1917), or Georg Simmel (1858–1918) (Kemper, 1990: 5; Williams and Bendelow, 1998: XV; Flam, 2000: 154). However, the analyses of the classical authors of sociology concerning the topic of emotions were not immediately continued. In the course of the twentieth century, emotions were gradually excluded from the sociological debate, which continued into the 1970s (Scherke, 2009: 291–294). Only after that time one can speak of sociology rediscovering the topic – most of all in the English-speaking countries (on the various trends of the special field of the sociology of emotions see Turner and Stets, 2005).

In the course of the twentieth century, under the influence of a rationalist concept of man pushing through, the value of emotions declined in debates of the social sciences. The development of this rationalist attitude shall be sketched here briefly. Contrasting rationality and emotionality as mutually excluding attitudes looks back into a long history and has only recently been questioned increasingly. For a structuring of the debate on rationality and emotionality, the division into three of the possible attitudes as suggested by Barbalet (Barbalet, 1998: 29) proves to be very useful: the *conventional approach*, which in the past few years has increasingly been questioned; the *critical approach* which since most recently is characterized, most of all, by the inclusion of neuroscientific insights and postulates a close connection of rationality and emotionality; and the *radical approach* which completely gives up on any distinction between emotionality and rationality.

Rationality and Emotionality

From a conventional point of view, rationality and emotionality are two separate fields of human 'being-in-the-world.' According to this approach, emotions are in most cases an obstacle for rational action. First roots of such a way of understanding emotions are found already in antiquity. The negative judgment of the Pre-Socratics, advised to avoid the stimulation of emotions, is also supported by Plato (fifth/fourth century BC), for example, although already he draws a differentiated, not exclusively negative, image of emotions and considers pleasure together with reason, an essential part of happiness in life. Also Aristotle (fourth century BC) does not strive for being completely liberated from emotions but considers the right measure a desirable virtue (Lanz, 1971: 89–95). Then, the rather negative view of emotions as unreasonable impulses that are to be avoided is found most of all among the representatives of the Stoa. While, for example, Zenon (fourth/third century BC) considers wrong judgments of reason the starting point for emotions which, however, he considers phenomena that must basically be distinguished from reason, for Chrysippos (third century BC) emotions cannot be separated from reason but must be judged on as exercising them in the wrong way. In the Middle Ages, this attitude is taken up again and sharpened, as far as asking in how far emotions are capable of influencing conscious decisions. Thomas Aquinas (1225–74) emphasizes the physical dimensions of emotion. As such, he says, emotions are neither good nor bad; if they are controlled by reason, they count among the virtues, if not, then they lead toward sin. Basically, they are not completely subject to the human will. The comprehensive way of understanding emotions, which in antiquity and the Middle Ages is still found to a certain degree – emotions both as suddenly triggered or long-time states of the soul and efforts by the soul – gradually becomes more differentiated in the debate of Modern Age, by more clearly distinguishing between the rather lasting desires of man and his spontaneous emotions. Also, the idea of body and soul being mostly one, which is still found in the writings of some ancient

authors and also in Thomas Aquinas, is then replaced by the idea of the two being strictly separated.

Thus, Barbalet describes the conventional understanding of the relation of rationality and emotionality by reaching back to René Descartes (1596–1650). Descartes' 'cogito ergo sum' makes reason the center of human existence, whereas now emotions are considered most of all by their relation to reason and by their possible destructive effects on it. The will, he states, must master the passionate impulses coming from the body. In this point of view emotions and reason are understood to be contradicting phenomena, in the context of which the realm which is associated with man's free will – that is, reason – is appreciated, whereas emotions must be suppressed as bodily phenomena that are obstacles for reason. However, when explaining Descartes' view of emotions it may not be overlooked that he dealt in very much detail with the physical processes of emotion. Thus, appreciating reason was not connected to neglecting (or lacking interest in) emotion. Also, we must note that the soul is capable, by imagining situations which contradict currently existing passions, of starting appropriate *physical* mechanisms, and thus of controlling unwanted passions. Often the perception of Descartes overlooked this fundamental connection of body and soul (which only allowed also for mastering the passions), thus one-sidedly emphasizing his explanation of the passions as phenomena that contradict reason. In the past centuries conventional view proved to be the predominant understanding of the relation of rationality and emotionality (despite the existence of other views) (Barbalet, 1998: 54–61).

In contrast to the conventional concept, the critical view of the relation of rationality and emotionality emphasizes that only emotions make cognition, that is, rational consideration, purposeful. Also this view considers emotion and reason as two different phenomena; however, one assumes a closer interaction of both. Also for this concept we find predecessors already in antiquity and the Middle Ages. In the eighteenth century, the Cartesian concept of man was criticized most of all by the Scottish moral philosophers who strived for a rehabilitation of man's senses and passions. David Hume (1711–76) assumed that passions control the will, thus directing human behavior toward a goal. Furthermore, already in Hume we find the basic idea of the radical approach, which is presented in the following, according to which rationality itself may be considered a certain kind of emotion. Thus, from the point of view of the critical approach emotions are necessary to direct reason, which may be of significance most of all in respect to a revision of modern Rational-Choice theories. By a kind of division of labor between the emotional and the rational components of human 'being-in-the-world' it becomes possible to define goals and then to try to achieve them by way of reasonable action. Often, humans are confronted with contradicting goals between which a rational decision is hardly possible. Furthermore, in many situations they have only insufficient information, which makes reasonable action even more difficult. In such cases emotions may contribute significantly to decision making, by allowing for a clear judgment on the various goals and for selecting one of them, despite lacking background information. This way they also guarantee long-term social cooperation – by avoiding egotistic, short-term goals. Most recently, most of all insights of the biosciences have produced evidence for the critical approach which have meanwhile also found response in the field of economics (Damasio, 1994; Barbalet, 1998: 39; Elster, 1998: 47–74; Fehr and Fischbacher, 2004). In contrast to the conventional understanding, which considers too much emotion an obstacle for rational decision making, *lack* of emotions is now seen as counterproductive for rational behavior.

The radical view assumes that rationality and emotionality are basically not different from each other, but that rationality itself may be considered a conglomerate of particular emotions.

As mentioned above, the idea of an inseparable connection between reason and emotion is already found in Chrysippos. Sometimes, this connection of reason and emotion is also found among the thinkers of Enlightenment, together with a positive judgment on the knowledge of the world guided by emotion. Such an idea is found in Alexander Gottlieb Baumgarten (1714–62), for example. For him, insight resulting from an emotional, aesthetic understanding of the world may be considered to be in analogy with reasonable insight (Franke and Oesterle, 1974: 83–86). Also, in the debate on so-called moral emotions, emotion is said to have a cognitive function in the widest sense. There, emotions become the actual bearer of human decision making and are thus attributed a degree of significance which is sometimes believed to be superior to reason-based insight, for example as described in the works of Anthony Ashley Cooper Shaftesbury (1671–1713) or in Francis Hutcheson (1694–1746). Also in Max Scheler's (1874–1928) phenomenology some cognitive achievements are based on such an emotional ground. Emotions are important most of all for concluding on values, and the thus resulting moral decisions (Scheler, 1954: 260–272). This independent relation to reality by way of emotions or feelings is crucial for Scheler's revaluation of emotion compared to reason.

We find evidence for the radical position also in David Hume and William James (1842–1910). Hume was of the opinion that spiritual activities carried out prudently and calmly are called reason. However, he said, this name hides the work of calm passions which, no less than violent passions, count among emotions (Barbalet, 1998: 31). Furthermore, James pointed out to the physical changes happening when experiencing certain emotions. In his opinion, man's spirit cannot be considered without these physical states. Insofar as rationality is characterized by certain physical sensations. According to James, the emotion of rationality is characterized most of all by being oriented at the present moment which does not seem to require further justification or explanation, thus allowing for focusing thought (James, 1956: 63–64). A 'feeling of rationality' provides security, thus guaranteeing the continuation of the action, which otherwise would not have happened. According to this view, rationality is just a special kind of emotional orientation toward the future (which may be described as being characterized by composure and calmness); however it cannot in principle be distinguished from other emotional orientations toward the future.

Body and Emotion

As sketched above, already philosophy in antiquity and the thinkers of Enlightenment as well as those of the nineteenth

century dealt with emotions. In the course of the differentiation of the system of sciences, which in the nineteenth century resulted in psychology being separated from philosophy, emotions were systematically dealt with in the context of this 'newly' developing discipline. In the course of the past 150 years psychology has assessed – sometimes by way of experiments – a variety of different theories on the question of how emotions are created. The following overview focuses on the question of which meaning has been attributed to physiological factors when it comes to emotional processes. Various points of view shall be sketched by way of some 'milestones' on the road to today's concept of emotions which is much influenced by the neurosciences, and thus by physiological factors. However, we have to take into account that this was not at all a linear process and even in psychology there exist theories that do not refer to physical aspects in a narrow sense.

There is, for example, the James–Lange theory that emphasized the significance of physiological experience for emotional feeling. At the end of the nineteenth century William James (like at about the same time as Carl Lange (1834–1900)) assumed that initially we never judge on a situation in a cognitive way before there happens an emotional reaction (we see a bear, recognize the danger, become afraid, and run away), but that it is our physiological experience that determines our emotions (we see a bear, run away and are thus afraid, i.e., we interpret our running away as a reaction of fear). Thus, the conscious experience of an emotion comes after the physiological processes. Often this was interpreted as stating that any emotion is characterized by a specific physiological signature (LeDoux, 2001: 48–51).

In the 1920s this assumption was taken up and modified by the Cannon–Bard theory. According to this theory, developed by Walter Cannon (1871–1945) and Philip Bard (1898–1977), it is only general conditions of agitation (due to reactions by the sympathetic nervous system, which are assumed to be consistent) which may become subject to a cognitive appraisal. Cannon and Bard assumed that physiological agitation and the conscious perception of an emotion happen parallel and that the consciously perceived emotional state is not a consequence of physical agitation only, as it was still assumed by James and Lange. The physical reactions, they stated, are most of all responsible for the intensity and urgency of a perceived emotion but do not determine its specific nature.

Then, in the course of the predominance of behaviorism in psychology, as far as into the 1950s, emotions were hardly discussed. This changed in the course of the 'cognitive revolution' which paid increasingly more attention to the conscious processing of information by the human mind. In contrast to the behaviorist stimulus–response model, man's mind was 'rehabilitated' and became once again a matter for psychology. However, as LeDoux criticizes, this sometimes happened by neglecting the physical (most of all neuronal) preconditions of the human information processing (LeDoux, 2001: 29, 58). For example, in the 1960s Stanley Schachter (1922–97) and Jerome Singer (1934–2010) joined the Cannon–Bard theory of the general state of agitation, however by enriching it by the thesis that apart from perceiving physical agitation also the perception of the actual environment influences the finally resulting emotion (Schachter and Singer, 1962). Schachter and Singer described in particular the social environment as being crucial for the kind of created emotions. According to this view, cognitive processes are thus responsible for the development of different emotional states. As Magda Arnold (1903–2002) found out in the 1960s, it is also cognitive appraisal that provides the starting point for the emotional process, after all. She made the triggers of these physical processes the focus of her considerations (Arnold, 1960). In the course of a process of cognitive appraisal, she said, the individual judges on his/her current (environmental) situation, and that is concerning the possible damage or benefit it may cause.

Arnold and other representatives of the appraisal theory assumed that those processes as being responsible for triggering an emotion are basically accessible to consciousness, which is why reports of patients based on introspection may be used as sources for the research of emotional processes. In contrast to this, LeDoux follows Robert Zajonc's (1923–2008) considerations, who, in the 1980s, had moved away from the long predominance of cognition and postulated a more complex relation of cognitive and emotional processes, for which most of all processes of unconscious information processing play an important role. Decisive for Zajonc's considerations were experiments during which test persons had developed preferences for stimuli that were offered only for a short time (and were thus not consciously perceived) (Zajonc, 1980). This unconscious information processing was the starting point for LeDoux's examinations (LeDoux, 2001: 58–69). Already, since the 1950s certain parts of the brain (the so-called limbic system) were considered the central switchboard of emotional perception. From LeDoux's point of view, not those parts of the brain which are called the limbic system are responsible for emotions, but there exists a complex process of processing of the incoming sensual information, in the course of which different physical (and thus also neuronal) systems are activated. For his work, LeDoux focuses on fear. The essential result, in short, is that there is so to speak a double process of information processing in the brain. Sensual perceptions reach the thalamus in the form of nerve impulses and from there go to the neocortex (where there happens cognitive assessment) and also to the amygdala (which controls physical reactions). In this context, those nerve impulses reaching the amygdala may sometimes already trigger physical reactions even before the sensual perceptions have been cognitively assessed. Thus, this approach enriches the cognitive appraisal theory, insofar as it offers an explanation also for the rapid, unconscious assessment of certain situations.

In the context of the debate on the emotional process, the concept of evolution has played an important role time and again – apart from the question of how physiology and cognition work together – that is, human emotions were analyzed in view of their functions as adjustment advantages (Turner, 2000: 19–21). Closely connected to these evolutionary considerations is also the idea of emotions finding universal expression. This assumption, which Charles Darwin (1809–82) had already expressed in 1872, was taken up by Paul Ekman in the 1960s. He gave evidence to it by several culture-comparative studies, in the course of which, however, it also became clear that there are culture-specific rules of presentation which demand the open expression or hiding of certain emotions (Ekman, 1984). Linked to the debate on the cross-cultural status of emotions and their evolutionary functions is also the idea of the existence of certain basic emotions or primary emotions, the opinions concerning

their kind and number being widely apart still today (Ulich and Mayring, 1992: 36–37, 132–133).

Since the 1960s, the complexity of the human emotional process, which includes biological, social, and cultural components, has also produced system–theory-oriented theories of emotion – such as in Carroll E. Izard or Silvan S. Tomkins (Tomkins, 1962; Izard, 1971) – which attempt descriptions of the interaction of these different components. More recently, the analysis of neuronal processes has played an important role also for system–theory-oriented theories of emotions.

In contrast to the debates of the 1960s and 1970s, which were characterized by the cognitive sciences, currently the physical dimension of the emotional process is in the fore of scientific attention (most of all in the neurosciences) again. However, since the 1970s also a 'sociology of emotions' has established as a special field within sociology, trying to shed light on the significance of social factors for the emotional process and to find out about their interaction with other factors (such as physiological ones).

Sociology of Emotions

Two main lines can be distinguished within the *sociology of emotions*: one line making the *social development* of emotions a topic of discussion, and the other investigating the *social effects* of emotions (Scherke, 2009: 58–62). This is only a very rough structure which, however, is suitable for illustrating the basic direction of studies.

Authors considering them to be characterized or changed by social structures analyze emotions for the way in which they are created. In this context, emotions are not only considered something biologically predetermined, but also as a phenomenon that is essentially (co)influenced by the social and/or cultural environment (like already Schachter/Singer have shown). The question about the 'nature' of emotions, that is, about their fundamental characteristics and ways of appearance, is in the fore for these studies. This is a topic also dealt with by other disciplines, which is why this field shows a number of interdisciplinary interfaces.

However, in the context of the *sociology of emotions* we also find studies analyzing the effect of emotions on social situations. According to this approach, emotions are assumed to be mostly given (without any implicit decision about their biological and/or social nature). Here the question of how emotions are created is postponed, and instead one discusses the effects of emotions on the structuring of social contexts. Basically, this approach also shows interfaces with other disciplines; however, in terms of research practice they may be more neglected to a certain degree than it is the case with the former approach. In other words, new bioscientific insights concerning the way in which emotions are created do not necessarily result in a revision of sociological explanations for the social effects of emotional behavior. The social consequences of emotions may, so to speak, be considered to belong to tasks of sociology in the *stricter* sense.

Among the approaches dealing with how emotions are characterized by social structures and interactions we may distinguish a positivist and a constructivist approach. Between these two there was an extensive debate in the Anglo-Saxon countries in the 1970s and 1980s. For both approaches the macro-level is in the fore. The constructivist approach refers to the macro-level by way of normative feeling rules regulating the way in which emotions are expressed, thus serving as a link between the social or cultural context (as well as its historical development) and the emotions of the individuals. By way of interactions between participants there occur negotiations about which emotions are felt or may be felt in each case. Aspects of social control play an important role for this modulation of emotions (in the form of sanctions by others or of internalized self-control learned in the course of socialization). It is the individuals who create and redesign emotions – in the sense of emotion work or emotion management (Hochschild, 1983) – adjusted to the respective social context. As has been shown also by recent ethnological studies, the respective cultural context is quite decisive for the way in which emotions are expressed. Furthermore, it must be taken into consideration that feeling rules change in the course of history. Quite a number of studies deal with the development or change of emotional concepts – such as the concept of Romantic love which is mostly a 'product' of the nineteenth century and has essentially contributed to the way in which partner relationships have been lived since then (Illouz, 1997).

The positivist approach assumes that emotions are more or less directly created by social–structural positions or relations of interaction partners, without considering processes of emotion management. Partly, most of all in Theodore D. Kemper, this idea also implies that social–structural positions, physiological mechanisms, and perceived emotions are closely linked (Kemper, 1978). Also, this approach refers predominantly to the micro-level, by sketching in a model possible interaction situations between two individuals or among small groups concerning their power and status relations and by analyzing them for their emotional consequences. However, also for this approach we may state a reference to the meso- or macro-level, as the social position of social groups (thus also a model of social segregation) can be included into the analysis.

Those approaches as analyzing the influence of emotions on the social plane can be further divided into those discussing emotional states as general preconditions for human behavior (and thus also social behavior) and those dealing with the various social consequences triggered by concrete emotions. From the point of view of the interpretative approach within the *sociology of emotions*, emotions are attributed a considerable role for the construction of reality. Emotions influence social situations in a way as to contributing to a (mostly nonverbal) structuring of social situations, by reducing the range of possible meanings in the context of interactions. Thus, emotions are part of the unquestioned preconditions for social situations, thus allowing for everyday processes. By way of facial gestures – and emotions are expressed most of all by facial expressions – fundamental attitudes toward the world and the current situation are communicated, thus reducing the possible leeway for interpretation, as interactions are always connected to a certain degree of insecurity. In other words, a sentence which is said with a smile is attributed a different meaning than the same sentence expressed in an angry voice and by the appropriate facial expression. Without analyzing the content in detail, the respective interaction partner will already react to the emotional tone of the message, which will steer the

course of the interaction further into a certain direction. In so far, emotions and the emotional way of creating meaning can contribute to behavioral certainty, and thus to efficient acting in the context of interaction. Among the concrete emotions whose effects on social behavior or social structures have been analyzed there count for example trust (Barbalet, 1998: 84–94), fear (Flam, 2000: 54–65), and shame (Scheff, 1990: 284–289). According to Collins, apart from short-term emotions (e.g., fear or anger) it is most of all long-term emotions or emotional moods that influence social structures (Collins, 2004: 3–25). These long-term emotions are created in the context of rituals. For this, there must be the direct physical encounter of at least two humans whose attention is directed to the same object or at a common activity and who finally become aware of this common direction. By way of directing their attention at the same object, the actors are capable of adjusting their emotional moods. Here, Collins refers to the state of collective effervescence as a result of group rituals which was already emphasized by Emile Durkheim. Collins speaks of rhythmic entertainment, by which he means the adjustment of the participants' moods due to physiological processes in the context of a successful ritual. If this adjustment to a common mood succeeds, not only the emotions of each individual are increased but there also results group solidarity. For example at burials, where the individual actors express emotions of mourning, the common ritual may not only increase the individual's emotions of mourning, they may also create or consolidate solidarity among the mourners. Thus, the long-term result of the interactions of a ritual is on the one hand the revaluation of group solidarity (including the positive charging of group symbols and group-related norms) and on the other hand an increase of the individuals' so-called emotional energy.

Among the *sociology of emotions* there exist a number of refined theoretical concepts and meanwhile a multitude of empirical studies on a variety of social fields (from social movements, politics, the Internet, as far as to business enterprises; Flam and King, 2005; Wettergren and Sieben, 2010; Pixley, 2012; Demertzis, 2013; Benski and Fisher, 2014). There has been a slight tendency to emphasize the social constructedness of emotions in order to overcome the predominance of psychology (or nowadays of the neurosciences) in emotions research and to make emotions a research topic for sociology again. This has led to a neglect of bodily aspects in tendency (with some exceptions like Kemper). This gap has to be closed by future research, which should also strive for a stronger connection of theoretical and empirical approaches.

See also: Affect and Emotion, Anthropology of; Affect-Regulation Motivation; Culture and Emotion; Early Emotional Development and Cultural Variability; Emotion and Expression; Emotion in Cognition; Emotion, Evolution of; Emotion, Neural Basis of; Emotion, Perception and Expression of; Emotional Development, Effects of Parenting and Family Structure on; Emotional Regulation; Emotions and Intergroup Relations; Emotions and Work; Emotions, History of; Emotions, Psychological Structure of; Emotions, Sociology of; Emotions: Methods of Assessment; Selfconscious Emotions, Psychology of.

Bibliography

Arnold, Magda, 1960. Emotion and Personality. Columbia University Press, New York.
Barbalet, J.M., 1998. Emotion, Social Theory and Social Structure. A Macro-sociological Approach. Cambridge University Press, Cambridge; New York; Melbourne.
Benski, Tova, Fisher, Eran (Eds.), 2014. Internet and Emotions. Routledge, New York; London.
Collins, Randall, 2004. Interaction Ritual Chains. Princeton University Press, Princeton; Oxford.
Damasio, Antonio, 1994. Descartes' Error: Emotion, Reason, and the Human Brain. Putnam, New York.
Demertzis, Nicolas (Ed.), 2013. Emotions in Politics: The Affect Dimension in Political Tension. Palgrave Macmillan, Basingstoke.
Ekman, Paul, 1984. Expression and the nature of emotion. In: Scherer, Klaus R., Ekman, Paul (Eds.), Approaches to Emotion. Erlbaum, Hillsdale; London, pp. 319–343.
Elster, Jon, 1998. Emotions and economic theory. Journal of Economic Literature XXXVI, 47–74.
Fehr, Ernst, Fischbacher, Urs, 2004. Social norms and human cooperation. Trends in Cognitive Sciences 8 (4), 185–190.
Flam, Helena, 2000. The Emotional Man and the Problem of Collective Action. Lang, Frankfurt am Main.
Flam, Helena, King, Debra (Eds.), 2005. Emotions and Social Movements. Routledge, London; New York.
Franke, Ursula, Oesterle, G., 1974. Gefühl. In: Joachim, Ritter, Karlfried, Gründer, Gottfried, Gabriel (Eds.), Historisches Wörterbuch der Philosophie, band 3. Schwabe, Basel, pp. 82–89.
Hochschild, Arlie R., 1983. The Managed Heart. The Commercialization of Human Feeling. University of California Press, Berkeley.
Illouz, Eva, 1997. Consuming the Romantic Utopia. Love and the Cultural Contradictions of Capitalism. University of California Press, Berkeley.
Izard, Carroll E., 1971. The Face of Emotion. Appleton-Century-Crofts, New York.
James, William, 1956. The sentiment of rationality. In: James, William (Ed.), The Will to Believe and Other Essays in Popular Philosophy, Human Immortality. Dover Publ, New York, pp. 63–110.
Kemper, Theodore D., 1978. A Social Interactional Theory of Emotions. Wiley, New York.
Kemper, Theodore D., 1990. Themes and variations in the sociology of emotions. In: Kemper, Theodore D. (Ed.), Research Agendas in the Sociology of Emotions. State University of New York Press, Albany, pp. 3–23.
Lanz, J., 1971. Affekt. In: Joachim, Ritter, Karlfried, Gründer, Gottfried, Gabriel (Eds.), Historisches Wörterbuch der Philosophie, band 1. Schwabe, Basel, pp. 89–100.
LeDoux, Joseph, 2001. Das Netz der Gefühle. Wie Emotionen entstehen. Dt. Taschenbuch-Verl, München (The emotional brain. The mysterious underpinnings of emotional life, New York 1996).
Pixley, Jocelyn (Ed.), 2012. New Perspectives on Emotions in Finance: The Sociology of Confidence, Fear and Betrayal. Routledge, London; New York.
Schachter, Stanley, Singer, Jerome E., 1962. Cognitive, social, and physiological determinants of emotional states. Psychology Review 69, 379–399.
Scheler, Max, 1954. Der Formalismus in der Ethik und die materiale Wertethik. Neuer Versuch der Grundlegung eines ethischen Personalismus. In: Scheler Maria (Ed.), Max Scheler. Gesammelte Werke, vol 2. fourth ed. Francke, Bern.
Scherke, Katharina, 2009. Emotionen als Forschungsgegenstand der deutsch-sprachigen Soziologie. VS Verl. für Sozialwissenschaften, Wiesbaden.
Scheff, Thomas, 1990. Socialization of emotions: pride and shame as causal agents. In: Kemper, Theodore D. (Ed.), Research Agendas in the Sociology of Emotions. State University of New York Press, Albany, pp. 281–304.
Tomkins, Silvan S., 1962. Affect, Imagery, Consciousness. Springer, New York.
Turner, Jonathan H., 2000. On the Origins of Human Emotions. A Sociological Inquiry into the Evolution of Human Affect. Stanford University Press, Stanford.
Turner, Jonathan H., Stets, Jan E., 2005. The Sociology of Emotions. Cambridge University Press, Cambridge.
Ulich, Dieter, Mayring, Philipp, 1992. Psychologie der Emotionen. Kohlhammer, Stuttgart; Berlin; Köln.
Wettergren, Asa, Sieben, Barbara (Eds.), 2010. Emotionalizing Organizations and Organizing Emotions. Palgrave Macmillan, Basingstoke.
Williams, Simon J., Bendelow, Gillian, 1998. Introduction: emotions in social life. Mapping the sociological terrain. In: Bendelow, Gillian, Williams, Simon J. (Eds.), Emotions in Social Life. Critical Themes and Contemporary Issues. Routledge, London; New York, pp. XV–XXX.
Zajonc, Robert, 1980. Feeling and thinking: preferences need no inferences. American Psychologist 33 (2), 151–175.

Emotion, Neural Basis of

Edmund T Rolls, Oxford Centre for Computational Neuroscience, Oxford, UK; and University of Warwick, Coventry, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Emotions can usefully be defined as states produced by rewards and punishers. Emotions prepare the body for action, but perhaps their most important function is to enable appropriate actions to be produced to stimuli and events that are rewards or punishers. The rewarding stimuli or events can be innate (gene-specified) or learned. The orbitofrontal cortex is especially important in emotion, because it represents rewards and punishers, and learns about which previously neutral stimuli are associated with rewards and punishers, and when these associations change. For this fundamental reason, parts of the brain such as the orbitofrontal cortex are important in social and emotional behavior. The anterior cingulate cortex receives from the orbitofrontal cortex and is involved in learning goal-directed actions to obtain reward or avoid punishers. The amygdala is an evolutionarily older system also involved in some responses to emotional stimuli.

In order to provide a fundamental basis for understanding the neural basis of emotion, it is useful to have a clear view of the events that give rise to emotions and the functions of emotion, for these help to define the input systems in the brain involved in producing emotions, and the output systems in the brain that are involved in expressing emotion. With this basis, developed fully by Rolls (2005, 2013a, 2014), the neural systems that underlie emotions are then considered.

The Events that Give Rise to Emotions

Emotions can usefully be defined as states produced by rewards and punishers. Rewards are stimuli for which an animal (including humans) will work, and punishers are stimuli that an animal will work to escape from or avoid. Rewards and punishers are instrumental reinforcers in that they alter the probability of behavior. Some stimuli are unlearned or primary reinforcers (e.g., the taste of food if the animal is hungry, or pain). Other stimuli may become reinforcing by learning, because of their association with such primary reinforcers, thereby becoming 'secondary reinforcers.' For example, fear is an emotional state which might be produced by a sound (the conditioned stimulus) that has previously been associated with a painful stimulus (the primary reinforcer).

The different emotions can be described and classified according to whether the reinforcer is positive or negative, and by the reinforcement contingency. An outline of such a classification scheme, which applies to each reinforcer, is shown in Figure 1.

The operation of the following factors enables a very wide range of different emotions to be described:

1. The reinforcement contingency (see Figure 1).
2. The intensity of the reinforcer (see Figure 1).
3. Any environmental stimulus might have a number of different reinforcement associations. For example, a stimulus might be associated both with the presentation of a reward and of a punishment, allowing states such as conflict and guilt to arise.
4. Emotions elicited by stimuli associated with different primary reinforcers will be different.
5. Emotions elicited by different secondary reinforcing stimuli will be different from each other (even if the primary reinforcer is similar).
6. The emotion elicited can depend on whether an active or passive behavioral response is possible. For example, if an active behavioral response can occur to the omission of a previously rewarded stimulus, then anger might be produced, but if only passive behavior is possible, then sadness, depression, or grief might occur.

By combining these six factors, it is possible to account for a very wide range of emotions. It is also worth noting that

1. emotions can be produced just as much by the recall of reinforcing events as by external reinforcing stimuli;
2. cognitive processing (whether conscious or not) is important in many emotions, for very complex cognitive processing may be required to determine whether environmental events are reinforcing or not;
3. emotions normally consist of cognitive processing which determines the reinforcing valence of the stimulus, and an elicited mood change if the valence is positive or negative;
4. and that stability of mood implies that absolute levels of reinforcement must be represented over moderately long time spans by the firing of mood-related neurons, a difficult operation which may contribute to 'spontaneous' mood swings, depression which occurs without a clear external cause, and the multiplicity of hormonal and transmitter systems which seem to be involved in the control of mood.

The Functions of Emotion

The functions of emotion can be summarized as follows.

1. The *elicitation of autonomic responses* (e.g., a change in heart rate) *and endocrine responses* (e.g., the release of adrenaline). These prepare the body for action.
2. *Flexibility of behavioral responses to reinforcing stimuli.* Emotional (and motivational) states allow a simple interface between sensory inputs and motor outputs. The new understanding is that some genes specify which stimuli are rewarding and punishing, and these stimuli are

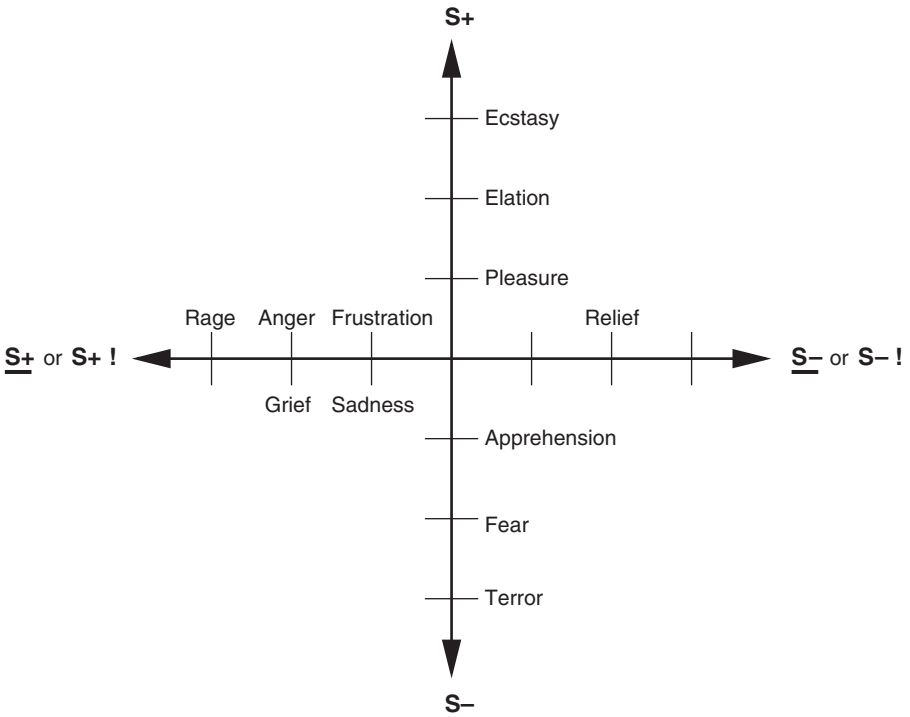

Figure 1 Some of the emotions associated with different reinforcement contingencies are indicated. Intensity increases away from the center of the diagram, on a continuous scale. The classification scheme created by the different reinforcement contingencies consists of (1) the presentation of a positive reinforcer (S+), (2) the presentation of a negative reinforcer (S−), (3) the omission of a positive reinforcer (S+) or the termination of a positive reinforcer (S+ !), and (4) the omission of a negative reinforcer (S−) or the termination of a negative reinforcer (S− !). It should be understood that each different reinforcer will produce different emotional states: this diagram just summarizes the types of emotion that may be elicited by different contingencies, but the actual emotions will be different for each reinforcer.

the goals for flexible actions. For example, genes specify that strong somatosensory stimuli produce a state of pain which animals (including humans) are built to perform actions to avoid or escape. The state produced by the painful stimulus is an emotional state which has the adaptive function of identifying and maintaining the goal for an action. This is an efficient design, for the genes need only specify the goals for actions, and not the actions themselves, so that the actions can be selected flexibly to meet the goals. It would be very inflexible and inefficient for the genes to specify actions themselves. In this way, emotions provide an efficient interface between sensory systems and actions. The hypothesis is that our emotions correspond to the goal-specifying states arising from stimuli that genes have found it adaptive and possible to specify. Rolls (2014) provides a table of some of the primary reinforcers that genes appear to specify, and thus an idea of how a wide range of different emotions have evolved because of their adaptive implications. This approach provides a fundamental approach to understanding emotions in terms of their evolutionary adaptive value.

3. Emotion is *motivating*. For example, fear learned by stimulus–reinforcement association formation provides the motivation for actions performed to avoid noxious stimuli. Rolls' evolutionary approach provides a unifying theory of motivation and emotion, for motivational states occur when we try to obtain (gene-specified) goals, and emotional states occur when those goals are, or are not, obtained. The goal in both cases is defined by the same gene-based reward and punishment systems in the brain.
4. *Communication*. For example, monkeys may communicate their emotional state to others, by making an open-mouth threat to indicate the extent to which they are willing to compete for resources, and this may influence the behavior of other animals. This aspect of emotion was that emphasized by Darwin (1872).
5. *Social bonding*. Examples are the emotions associated with the attachment of the parents to their young, and the attachment of the young to their parents.
6. The current mood state can affect the *cognitive evaluation of events or memories* (Pessoa, 2010; McIntyre et al., 2012; Rolls, 2013b), and this may have the function of facilitating continuity in the interpretation of the reinforcing value of events in the environment.
7. Emotion may facilitate the *storage of memories* (Rolls, 2008, 2014). One way in which this occurs is that episodic memory (i.e., one's memory of particular episodes) is facilitated by emotional states. This may be advantageous in that storing many details of the prevailing situation when a strong reinforcer is delivered may be useful in generating appropriate behavior in situations with some similarities in the future. A second way is that the current emotional state may be stored with episodic memories,

providing a mechanism for the current emotional state to affect which memories are recalled. A third way in which emotion may affect the storage of memories is by guiding the cerebral cortex in the representations of the world which are set up. For example, in the visual system, it may be useful to build perceptual representations or analyzers that are different from each other if they are associated with different reinforcers, and it may be less likely to build them if they have no association with reinforcement.

8. Another function of emotion is that by enduring for minutes or longer after a reinforcing stimulus has occurred, it may help to produce *persistent and continuing motivation and direction of behavior*, to help achieve a goal or goals.
9. Emotion may trigger the *recall of memories* stored in neocortical representations. Amygdala backprojections to the cortex could perform this for emotion in a way analogous to that in which the hippocampus could implement the retrieval in the neocortex of recent (episodic) memories (Rolls, 2008).
10. Rewards and punishers, and the emotional states they produce, are represented on a common scale that facilitates choices and decisions between different rewards, and between actions and their costs to obtain these rewards (Grabenhorst and Rolls, 2011; Rolls, 2014).

Brain Mechanisms of Emotion

Overview

Animals are built with neural systems that enable them to evaluate which environmental stimuli, whether learned or not, are rewarding and punishing, that is will produce emotions and will be worked for or avoided. Sensory stimuli are normally processed through several stages of cortical processing to produce a sensory representation of the object before emotional valence is decoded. For example, in the taste system, taste is analyzed in primates to provide a representation of what the taste is in the primary taste cortex, and this representation is independent of the reward value of the taste in that it is not affected by hunger. In the secondary taste cortex, in the orbitofrontal region (see **Figure 2**), the reward value of the taste is represented, in that neurons only respond to the taste if the primate is hungry (Rolls, 2012a). In another example, in the visual system representations of objects which are view-, position- and size-invariant are produced in the inferior temporal visual cortex after many stages of cortical processing, and these representations are independent of the emotional valence of the object (Rolls, 2012b). Then in structures such as the orbitofrontal cortex and amygdala, which receive from the inferior temporal visual cortex, associations are learned between the objects and the primary reinforcers associated with them, by the process of stimulus–reinforcement association learning. In the orbitofrontal cortex and amygdala, emotional states are thus represented (Rolls, 2014).

Consistent with this, electrical stimulation of the orbitofrontal cortex and amygdala is rewarding, and damage to these structures affects emotional behavior by affecting stimulus–reinforcement association learning. These brain regions influence the selection of behavioral actions through brain systems such as the ventral striatum and other parts of the basal ganglia (Rolls, 2005).

The Amygdala

The amygdala receives information about primary reinforcers (such as taste and touch), and also information about visual and auditory stimuli from higher cortical areas (such as the inferior temporal cortex) that can be associated by learning with primary reinforcers (**Figure 2**). Bilateral removal of the amygdala in monkeys produces tameness, a lack of emotional responsiveness, the excessive examination of objects, often with the mouth, and the eating of previously rejected items such as meat (the Kluver-Bucy syndrome). In analyses of the bases of these behavioral changes, it has been observed that there are deficits in learning to associate stimuli with primary reinforcement, including both punishments and rewards. The association learning deficit is present when the associations must be learned from a previously neutral stimulus (e.g., the sight of an object) to a primary reinforcing stimulus (such as the taste of food).

Further evidence linking the amygdala to reinforcement mechanisms is that monkeys will work in order to obtain electrical stimulation of the amygdala, that single neurons in the amygdala are activated by brain-stimulation reward of a number of different sites, and that some amygdala neurons respond mainly to rewarding stimuli, and others to punishing stimuli (Rolls, 2014). The association learning in the amygdala may be implemented by associatively modifiable synapses from visual and auditory neurons on to neurons receiving inputs from taste, olfactory or somatosensory primary reinforcers (LeDoux and Damasio, 2013). Consistent with this, at least one type of associative learning in the amygdala can be blocked by local application to the amygdala of an *N*-methyl-D-aspartate (NMDA) receptor blocker, which blocks long-term potentiation, which is a model of the synaptic changes that underlie learning (Davis, 2006). The amygdala is involved in some conditioned responses to fear, including startle, freezing, and autonomic responses (Phelps, 2006; LeDoux and Damasio, 2013).

An interesting group of neurons in the amygdala (in, for example, the basal accessory nucleus) responds primarily to faces (Leonard et al., 1985). They are probably part of a system which has evolved for the rapid and reliable identification of individuals from their faces, and of facial expressions, because of the importance of this in primate social behavior. Consistent with this, activation of the human amygdala can be produced in neuroimaging studies by some face expressions, and lesions of the human amygdala may cause difficulty in the identification of some face expressions.

The Orbitofrontal Cortex

The orbitofrontal cortex receives inputs from the inferior temporal visual cortex and superior temporal auditory cortex; from the primary taste cortex and the primary olfactory (pyriform) cortex (see **Figure 2**); from the amygdala, and from the midbrain dopamine neurons. Damage to the caudal orbitofrontal cortex in the monkey produces emotional changes.

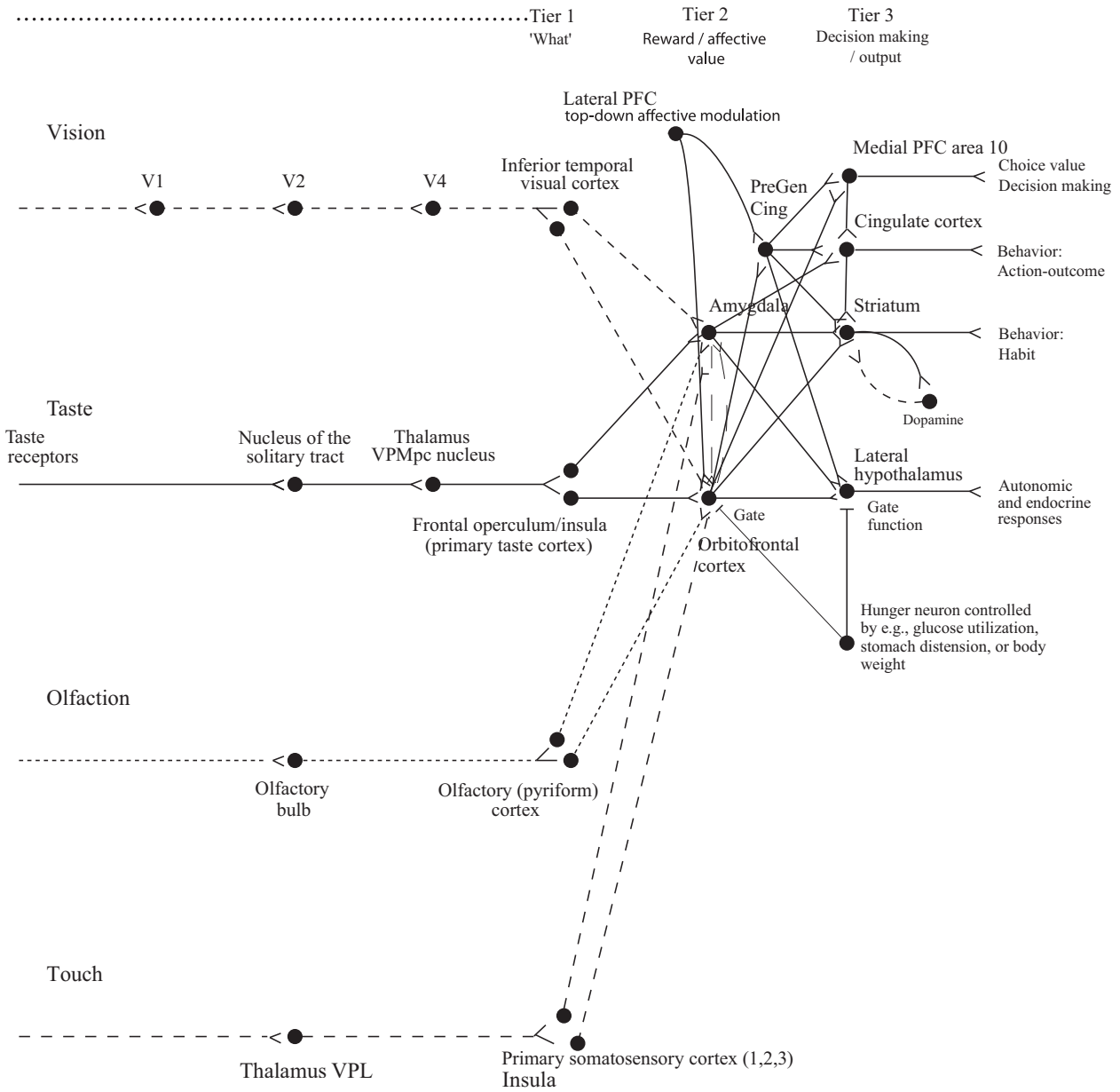

Figure 2 Schematic diagram showing some of the connections of the taste, olfactory, somatosensory, and visual pathways in the brain. V1, primary visual (striate) cortex; V2 and V4, further cortical visual areas. PFC, prefrontal cortex. VPL, ventro-postero-lateral nucleus of the thalamus, which conveys somatosensory information to the primary somatosensory cortex (areas 1, 2 and 3). VPMpc, ventro-postero-medial nucleus pars parvocellularis of the thalamus, which conveys taste information to the primary taste cortex. Pregen Cing, pregenual cingulate cortex. For purposes of description, the stages can be described as Tier 1, representing what object is present independently of reward value; Tier 2 in which reward value is represented; and Tier 3 in which decisions between stimuli of different values are taken, and in which value is interfaced to behavioral output systems.

These include decreased aggression to humans and to stimuli such as a snake and a doll, and a reduced tendency to reject foods such as meat. These changes may be related to a failure to react normally to and learn from nonreward in a number of different situations. This failure is evident as a tendency to respond when responses are inappropriate, for example, no longer rewarded.

For example, monkeys with orbitofrontal cortex damage are impaired on Go/NoGo task performance (in which they should make a response to one stimulus to obtain a reward, and should not make a response to another stimulus in order to avoid a punishment), in that they Go on the NoGo trials. They are also impaired in an object-reversal task in that they respond to the object which was formerly rewarded with food. They are also impaired in extinction in that they continue to respond to an object which is no longer rewarded. Further, the visual discrimination learning deficit shown by monkeys with orbitofrontal cortex damage, may

be due to the tendency of these monkeys not to withhold responses to nonrewarded stimuli.

The primate orbitofrontal cortex contains neurons which respond to the reward value of taste (a primary reinforcer), in that they only respond to the taste of food when hunger is present (which is when food is rewarding). It also contains neurons which learn to respond to visual stimuli associated with a primary reward such as taste, and which reverse their responses to another visual stimulus in one trial when the rewards and punishers available from those visual stimuli reverse. Further, these visual responses reflect reward, in that feeding the monkey to satiety reduces the responses of these neurons to zero. Moreover, in part of this orbitofrontal region, some neurons combine taste and olfactory inputs, in that they are bimodal, and are in 40% of cases affected by olfactory-to-taste association learning and by feeding the monkey to satiety, which reduces the reward value. In addition, some neurons in the primate orbitofrontal cortex respond to the sight of faces. These neurons are likely to be involved in learning which emotional responses are currently appropriate to particular individuals, and in making appropriate emotional responses given the face expression.

It has been found that another class of neurons in the orbitofrontal cortex of the monkey responds in certain non-reward situations (Thorpe et al., 1983). For example, some neurons respond in extinction, immediately after a lick had been made to a visual stimulus which had previously been associated with fruit juice reward, and other neurons responded in a reversal task, immediately after the monkey had responded to the previously rewarded visual stimulus, but had obtained punishment rather than reward. Another class of orbitofrontal neuron responded to particular visual stimuli only if they were associated with reward, and these neurons showed one trial stimulus–reinforcement association reversal. Another class of neuron conveyed information about whether a reward had been given, responding, for example, to the taste of sucrose, or for other neurons of saline.

These types of information may be represented in the responses of orbitofrontal neurons because they are part of a mechanism which evaluates whether a reward is expected, and generates a mismatch (evident as a firing of the nonreward neurons) if reward is not obtained when it is expected (Rolls, 2014). These neuronal responses provide further evidence that the orbitofrontal cortex is involved in emotional responses, particularly when these involve correcting previously learned reinforcement contingencies, in situations which include those usually described as involving frustration.

It is of interest and potential clinical importance that a number of the symptoms of frontal lobe damage in humans appear to be related to this type of function: altering behavior when stimulus–reinforcement associations alter, as described next. Thus, humans with frontal lobe damage can show impairments in a number of tasks in which an alteration of behavioral strategy is required in response to a change in environmental reinforcement contingencies (Rolls et al., 1994). Some of the personality changes that can follow frontal lobe damage may be related to a similar type of dysfunction. For example, the euphoria, irresponsibility, lack of affect, and lack of concern for the present or future which can follow frontal lobe damage may also be related to a dysfunction in altering behavior appropriately in response to a change in reinforcement contingencies. At one time, following a report by Moniz (1936), prefrontal lobotomies or leucotomies (cutting white matter) were performed in humans to attempt to alleviate a variety of problems, and although irrational anxiety or emotional outbursts were sometimes controlled, intellectual deficits and other side effects were often apparent (Valenstein, 1973). Thus these operations have been essentially discontinued.

To investigate the possible significance of face-related inputs to orbitofrontal visual neurons described above, the responses to faces that were made by patients with orbitofrontal damage produced by pathology or trauma were tested. Impairments in the identification of facial and vocal emotional expression were demonstrated in a group of patients with ventral frontal lobe damage who had socially inappropriate behavior (Hornak et al., 1996). The expression identification impairments could occur independently of perceptual impairments in facial recognition, voice discrimination, or environmental sound recognition. Thus the orbitofrontal cortex in humans appears to be important not only in the rapid relearning of stimulus–reinforcement associations, but also in representing some of the stimuli, such as face expression, which provide reinforcing information. Consistent with this, neuroimaging studies in humans are showing representations which reflect the pleasantness of the taste and smell of food, the pleasantness of touch, and also of quite abstract rewards and punishers such as winning or losing money (O'Doherty et al., 2001; Grabenhorst and Rolls, 2011).

The behavioral selection system must deal with many competing rewards, goals, and priorities. This selection process must be capable of responding to many different types of reward decoded in different brain systems that have evolved at different times, even including the use, in humans, of a language system to enable long-term plans to be made. These many different brain systems, some involving the implicit (unconscious) evaluation of rewards, and others explicit, verbal, conscious, evaluation of rewards and planned long-term goals, must all enter into the selector of behavior (Rolls, 2014). Although poorly understood, the issue of emotional feelings is part of the much larger problem of consciousness, and may involve the capacity to have thoughts about thoughts, that is higher order thoughts (Rolls, 2012c).

See also: Emotion and Expression; Emotion in Cognition; Emotion, Evolution of; Emotions, Psychological Structure of; Emotions, Sociology of; Infancy and Childhood: Emotional Development; Self and Emotional Development in Adulthood and Later Life; Self-Determination Theory.

Bibliography

Darwin, C., 1872. The Expression of the Emotions in Man and Animals, third ed. University of Chicago Press, Chicago.

Davis, M., 2006. Neural systems involved in fear and anxiety measured with fear-potentiated startle. American Psychologist 61, 741–756.

Grabenhorst, F., Rolls, E.T., 2011. Value, pleasure, and choice in the ventral prefrontal cortex. Trends in Cognitive Sciences 15, 56–67.

Hornak, J., Rolls, E.T., Wade, D., 1996. Face and voice expression identification in patients with emotional and behavioural changes following ventral frontal lobe damage. Neuropsychologia 34, 247–261.

LeDoux, J.E., Damasio, A.R., 2013. Emotions and feelings. In: Kandel, E.R., Schwartz, J.H., Jessell, T.M., Siegelbaum, S.A., Hudspeth, A.J. (Eds.), Principles of Neural Science, fifth ed. McGraw Hill, New York, pp. 1079–1094.

Leonard, C.M., Rolls, E.T., Wilson, F.A.W., Baylis, G.C., 1985. Neurons in the amygdala of the monkey with responses selective for faces. Behavioural Brain Research 15, 159–176.

McIntyre, C.K., McGaugh, J.L., Williams, C.L., 2012. Interacting brain systems modulate memory consolidation. Neuroscience and Biobehavioral Reviews 36, 1750–1762.

Moniz, E., 1936. Tentatives Opératoires dans le Traitement de Certaines Psychoses. Masson, Paris.

O'Doherty, J., Kringelbach, M.L., Rolls, E.T., Hornak, J., Andrews, C., 2001. Abstract reward and punishment representations in the human orbitofrontal cortex. Nature Neuroscience 4, 95–102.

Pessoa, L., 2010. Emergent processes in cognitive-emotional interactions. Dialogues in Clinical Neuroscience 12, 433–448.

Phelps, E.A., 2006. Emotion and cognition: insights from studies of the human amygdala. Annual Review of Psychology 57, 27–53.

Rolls, E.T., 2005. Emotion Explained. Oxford University Press, Oxford.

Rolls, E.T., 2008. Memory, Attention, and Decision-Making: A Unifying Computational Neuroscience Approach. Oxford University Press, Oxford.

Rolls, E.T., 2012a. Taste, olfactory, and food texture reward processing in the brain and the control of appetite. Proceedings of the Nutrition Society 71, 488–501.

Rolls, E.T., 2012b. Invariant visual object and face recognition: neural and computational bases, and a model, VisNet. Frontiers in Computational Neuroscience 6 (35), 1–70.

Rolls, E.T., 2012c. Neuroculture. On the Implications of Brain Science. Oxford University Press, Oxford.

Rolls, E.T., 2013a. What are emotional states, and why do we have them? Emotion Review 5, 241–247.

Rolls, E.T., 2013b. A biased activation theory of the cognitive and attentional modulation of emotion. Frontiers in Human Neuroscience 7 (74), 1–15.

Rolls, E.T., 2014. Emotion and Decision-Making Explained. Oxford University Press, Oxford.

Rolls, E.T., Hornak, J., Wade, D., McGrath, J., 1994. Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. Journal of Neurology, Neurosurgery and Psychiatry 57, 1518–1524.

Thorpe, S.J., Rolls, E.T., Maddison, S., 1983. Neuronal activity in the orbitofrontal cortex of the behaving monkey. Experimental Brain Research 49, 93–115.

Valenstein, E.S., 1973. Brain Control. A Critical Examination of Brain Stimulation and Psychosurgery. Wiley, New York.

Emotion, Perception and Expression of

Daisung Jang and Hillary Anger Elfenbein, Washington University in St. Louis, Saint Louis, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Emotions evolved to serve our need to communicate quickly and efficiently. Expressions serve as symptoms of our internal states, a signal appealing to others for action, and a symbol to convey information about an event. Expression produces outwardly visible cues interpreted by perceivers for their valuable information. Expression and perception is largely automatic, yet also results from norms in expressing and perceiving emotions. Although expressions may have originated from reflexive actions, they have evolved into a complicated system providing information about other people's reactions, attitudes, and likely behaviors. Reading these messages confers an advantage for interpersonal interactions.

The human experience is characterized by interaction with others. Above and beyond verbal communication, communicating via emotion serves as a quick and efficient means to coordinate social action by conveying information that provides insight into others' internal states. Humans are uniquely equipped to display emotional expressions, via highly developed facial and vocal structures, body language, and artifacts such as written communication, music, and art. Likewise, humans are well equipped to perceive these emotional expressions, and those who can do so effectively have an advantage in their social lives and even work lives.

We begin by stepping back from emotional expression and perception to discuss more broadly the emotion process, which is responsible for the emotional experience that typically sets expression and perception into motion. We discuss theoretical perspectives and empirical findings regarding their functions, individual differences, and deliberate regulatory processes. We then discuss the concept of emotional contagion, whereby people 'catch' each other's emotions in a complex interplay between expression and perception, as well as discuss cultural differences.

The Emotion Process

The antecedent to expressing an emotion is a combination of emotional experience and deliberate attempts to manipulate expressions for strategic purposes. We summarize the two major theoretical perspectives on emotional experience.

Although intuition may suggest emotions arise irrationally, evidence shows that emotions arise out of an orderly and clearly sequenced process (Frijda, 1986; Smith and Lazarus, 1990). The most common scientific definition of emotion is that they are psychological and physiological reactions to the stimuli that people observe in the world around them – with stimuli including other people, events, internal thoughts, written messages, music, and works of art, to name a few (Frijda, 1988).

This means emotions are about *something* – something that catches our attention, even if we cannot immediately identify its source. Major theorists argue that we continually and automatically engage in a process called cognitive appraisal (Frijda, 1986; Smith and Lazarus, 1990). Cognitive appraisal involves asking a series of questions in an ordered checklist to evaluate any noteworthy stimulus (e.g., Frijda, 2007; Scherer, 1988; Smith and Ellsworth, 1985). Although there are slightly differing models, most include at least these five dimensions, among others: (1) *valence*, also called *pleasantness*, which is the extent to which the stimulus is positive, negative, or neutral – e.g., distinguishing happiness from sadness; (2) *novelty*, which is the extent to which the stimulus is familiar versus unfamiliar – e.g., distinguishing a positive surprise from serenity; (3) *goal obstruction*, or the extent to which the stimulus indicates that there is a concern in reaching one's goals – e.g., distinguishing challenge from enjoyment; (4) *control*, or an evaluation of whether the stimulus is controllable by the self, another person, or no one – e.g., distinguishing pride from anger from sadness; and (5) *fairness*, also called *norms*, or the extent to which the stimulus is consistent with social or personal standards – e.g., distinguishing contempt from irritation. The pattern of evaluations acts like a formula to determine what emotion we should feel (see **Table 1** for examples). For example, the dimension of control distinguishes many related emotions, particularly negative emotions – e.g., if a team project gets canceled, then a sense of personal responsibility over the cancellation is associated with guilt or shame,

Table 1 Dimensions of emotional appraisal

	Pleasantness	*Novelty*	*Goal obstruction*	*Control*	*Fairness*
Happiness	High	High	No		High
Sad	Low	Low	Yes	No one	
Angry	Low	High	Yes	Another person	Low
Fear	Low	High	Yes	Another person or no one	
Surprise		High			
Serenity	High	Low			
Disgust	Low				
Guilt	Low		Yes	Oneself	Low
Pride	High		No	Oneself	High
Relief	High		No		

Sources: Ellsworth, P.C., Scherer, K.R., 2003. Appraisal processes in emotion. In: Davidson, R.J., Scherer, K.R., Goldsmith , H.H. (Eds.), Handbook of Affective Sciences. Oxford University Press, New York, pp. 572–595. Scherer, K.R., Schorr, A., Johnstone, T. (Eds.), 2001. Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, New York.

others' influence over the cancellation is associated with anger, and no one being responsible is associated with sadness or fear.

In describing this cognitive appraisal process, we emphasize that the tight mapping between appraisal dimensions and emotional experience creates opportunities for communication – through emotional expression and perception. In a process called *backtracking* (Elfenbein, 2007; Frijda, 2007), witnessing emotional displays gives us information because it allows us to step back and infer how other people appraise their situations. Even when emotional displays are strategically manipulated, backtracking provides useful information because it allows us to infer what expresser would like us to believe.

For balance, we note that there has been controversy between the *basic emotions* approach that is described here versus the *circumplex model* of emotion (Barrett and Russell, 1999). According to the circumplex model, our internal experience consists only of the pleasantness dimension of appraisal plus a sense of intensity level – and that the remainder of the appraisal process described above is a social construction. Because these two competing perspectives concern internal processes, our discussion of outward expressions can be consistent with either of these models.

This description of the appraisal process is intended to set the stage as context for discussing the expression and recognition of emotional displays. One does not only experience emotions based on these appraisals, but conveys these appraisals as information to others.

Expression of Emotion

Functions of Emotional Expression

Emotional experience creates urges that need to be released or contained. The first type of urge is to physical action; the root of the word emotion comes from the Latin word for motion. Each distinct emotion is related to specific *action tendencies* (Frijda, 1986), which are psychological and physiological responses that ready the individual to respond adaptively to the most pressing issues in their environments (Cosmides and Tooby, 2000; Frijda, 1986; Scherer et al., 2013). For example, happiness promotes social bonding, anger leads us to address problems in a relationship, and contempt leads us to reinforce a social hierarchy. The *social functional* perspective on emotion argues that these action tendencies serve functions in promoting group living (Morris and Keltner, 2000). Even if an emotion does not appear functional at every instance, it remains available in the repertoire because in some cases it helps social groups, like family units, to thrive or even simply remain intact. The second type of urge created by emotional experience is emotionally expressive behavior.

Why do we express our emotions? The Organon model (Bühler, 1934), summarized in **Table 2**, articulates three distinct functions: (1) *symptoms* convey the authentic internal experience of a speaker, which provides information to others about their states, intentions, attitudes, and likely future behaviors; (2) *signals* convey requests for actions from other people; and (3) *symbols* convey information about an event or object, in the way that spoken language represents our thoughts and environment. Although there is debate about the extent to which affective display is involuntary (i.e., symptoms, called *push factors*) or deliberate communication (i.e., signals and symbols, also called *pull factors*) (Ekman, 1971; Fridlund, 1994; Parkinson, 2005; Russell, 1994), these three functions are not meant to compete with each other. All three are valid reasons for emotional expression, and each of the three can be used singly or in combination – e.g., a wince can result from pain, can elicit helpful responses from those nearby, and can warn others to stay away from a dangerous situation (Singer et al., 2004). Separate psychological processes appear to distinguish push versus pull processes, which use separate neuroanatomical pathways guided by different tracts of facial nerves (Rinn, 1984).

Research and theorizing about emotional expression has been most common for the function of symptom; indeed, Ekman's (1971) influential *neurocultural theory* argues that our facial expressions are authentic readouts of emotional experience unless we are paying conscious attention to managing our faces. Even taking a less extreme perspective, it is clear that emotion displays are generated quickly and often without our awareness, which makes them an efficient form of communicating (e.g., Shariff and Tracy, 2011).

The signal function of emotion has also been an active focus of academic work, in terms of the social functions of emotion that lead emotional displays to induce reactions from others (Fridlund, 1994; Keltner and Haidt, 1999). By eliciting actions from others, emotions help to enhance the coordination of social groups. Positive expressions encourage others to

Table 2 The Organon model of interpersonal communication

Function	*Definition*	*Examples through emotional expression*
Symptom	Conveying the internal state of the expresser. Conveys emotions, intentions, and attitudes.	Sincere and automatic displays include an embarrassed blush when slipping on a banana peel, a surprised gasp when witnessing a car accident, and a joyful shriek when hearing good news.
Signal	Conveying an appeal to the observer that requests an action from them.	Deliberate appeals include a look of fear that begs for help, a look of anger that commands to stop, and a display of shame that tells others a punishment is no longer needed.
Symbol	Representing an object or event.	Deliberate communication of cognitive appraisal includes a scowl to show displeasure, a look of pride to convey who won the contest, or a smile to assure that a large animal is actually friendly.

Sources: Bühler, 1934/1990. Theory of Language: The Representational Function of Language (D.F. Goodwin, Trans.). Foundations of Semiotics; J. Benjamins Pub. Co., Amsterdam; Philadelphia; Scherer, K.R., 1988. Criteria for emotion-antecedent appraisal: a review. In: Hamilton, V., Bower, G.H., Frijda, N.H. (Eds.), NATO Advanced Study Institutes Series. Series D, Behavioural and Social Sciences. Kluwer, New York, pp. 89–126.

continue their actions, whereas negative expressions serve as deterrence (Blair, 2003).

Relatively little work has been done on the symbolic function of emotional expressions. Laukka and Elfenbein (2012) found that appraisal dimensions can be inferred from listening to the acoustical cues posed by professional actors through their voice. As such, we can backtrack from emotional expressions and thereby reveal information about how the expresser evaluated their situation.

Modes of Emotional Expression

Humans are uniquely equipped to express a variety of emotions, having the specialized neural circuitry and muscular structures that allow for a range of expressions that stand unrivaled compared to other species (Burrows, 2008; LeDoux, 2012). Initially, emotional expressions may have been adaptive in readying the physiological systems to react to stimuli that have survival implications such as a fear response that facilitates heightened sensitivity of the sensory organs (Susskind and Anderson, 2008). These simple reflexes produced reliable signals that conveyed valuable information to others, and over time, began to be used deliberately (Russell et al., 2003) and evolved to communicate social and abstract concepts (Blair, 2003; Shariff and Tracy, 2011).

Emotional expression takes on many forms. First, push factors create outwardly observable cues. Long-standing theories point out that emotions are composed of both subjective feeling states and physiological responses (Darwin, 1872; James, 1884). Physiological markers of emotion experience include facial expressions, nonverbal components of vocal expressions, galvanic skin response, respiration, heart rates, bodily posture, and muscular movement (Banse and Scherer, 1996; Barrett et al., 2011; DePaulo and Friedman, 1998; Ekman and Rosenberg, 2005). Any of these changes can provide visible cues to others, even if they were not necessarily meant for an audience. Second, we evolved pull factors via the ability to manipulate the most outwardly visible of nonverbal cues, namely facial expressions, vocal tone, and body posture. Facial expressions are the most salient mode of expressions for humans, especially because humans are uniquely equipped with the complex facial muscular structure that can produce many differentiated emotional expressions. Perhaps for this reason, the bulk of research on emotional expression has focused on the face (Russell et al., 2003). However, humans also have expressive control via vocal chords and bodily poses, which makes emotion expressions multimodal in nature (Rosenthal et al., 1979).

In examining channels of communicating emotion, Ekman and Friesen (1969) proposed a controllability-leakage hierarchy of channels of communication. They suggested that the more controllable channels, such as the face, are subject to high accountability on the part of the expresser, elicit external feedback from others, and provide the expresser with internal feedback. This is in contrast to the less controllable 'leaky' channels such as the body and the voice. As such, facial expressions generally express the information that we choose to volunteer (DePaulo, 1992). By contrast, information expressed through less controllable channels such as body movements and the voice may provide a truer window into a person's feelings. This information can still be controlled, but it is more difficult and requires more conscious, deliberate effort (Rosenthal and DePaulo, 1979). In a dramatic illustration of leaky communication, expressions of contempt have been found years later to predict the deterioration of marriages (Gottman, 1998).

In examining how we communicate emotion, there is an often-cited statistic that in communicating one's feelings and attitudes, 93% of the information conveyed is through nonverbal channels versus our words (Mehrabian and Ferris, 1967). However, this is based on a laboratory study that does not account for the wide range of methods available to express ourselves. This is especially true in cases where expressers and perceivers are removed in time and space – such as when one writes a letter to a distant friend – where the verbal channel is the only medium of communication.

Individual Differences in Emotion Expression

In examining individual differences in emotional expression, there have been two distinct operational definitions.

The first approach has been to examine individual differences in expressivity, or how much people tend to 'wear their hearts on their sleeves.' Some individuals have more intense levels of emotional experience and/or more intense urges to express their emotions regardless of their underlying experience (Gross and John, 1998). Expressivity has been conceptualized as a personality trait (Gross and John, 1998; Halberstadt, 1986) because it refers to individual differences in one's way of being rather than producing correct versus incorrect expressions.

The second approach has been to examine individual differences in the ability to express one's emotions clearly so that they can be understood accurately by others. This treats expressivity as an ability – testing if the emotions intended were communicated to others. Emotional expression accuracy – as well as emotion perception accuracy, which is discussed below – is typically considered central to models of *emotional intelligence* (Mayer et al., 2008), as well as the related constructs of *affective social competence* (Halberstadt et al., 2001) and *social skill* (Riggio, 1986). We have an intuitive understanding that controlling emotional displays can be beneficial, e.g., terms like *poker face* exist to explain the benefits of hiding one's evaluation of a situation. Indeed, expressive accuracy appears to be an asset, with greater expression ability associated with positive life outcomes such as workplace performance (DiMatteo et al., 1986).

It is noteworthy that these two distinct ways to conceptualize individual differences in emotion expression – that is, expressiveness versus expression accuracy – tend to be positively correlated, albeit with mixed findings (Halberstadt, 1986; Tucker and Riggio, 1988). An interpretation could be that we become clearer in conveying emotions that we tend to convey intensely – in other words, perhaps practice makes perfect.

Regulation of Emotional Expression

From a young age, we regulate our emotional expressions to be appropriate for the social environment (for a detailed review, see Gross, 2007). However, this regulation comes at a cost – namely, the frustration of holding in the action tendencies and expressive energy that evolved for each emotion (Gross and Levenson, 1993). Notably, e-mail existed relatively briefly before the expressive channel of *emoticons* was invented

(Associated Press, 2007), presumably to reduce the frustration involved with expressing oneself in a text-only medium.

Humans are potentially unique from other species for following norms about displaying affect that match expectations about emotional experience in social contexts (Gallois, 1994; Hochschild, 1983), with display rules to deintensify, intensify, neutralize, and mask displays with qualitatively different displays (Ekman, 1972). This makes the display of emotions an interactive process between the physiological responses and the cultural, social, and interpersonal expectations that shape emotional expression (Scherer et al., 2013). The discussion above of the Organon model emphasizes that pull factors are a matter of deliberate communication – and, as with verbal communication, the messages we choose are shaped by social norms. Returning to appraisal theory provides a theoretical framework to guide what emotions tend to be appropriate in given situations. For example, people expect positive displays when an event is supposed to stay positive, such as when losing, beauty contestants must smile for the winner.

As mentioned above, regulation can come at a cost, if it is accompanied by feelings of inauthenticity or expressive urges being thwarted. In the working world, the concept of emotional labor (Hochschild, 1983) refers to the requirement to express specific emotions on the job, whether or not they are felt, and particularly toward clients, e.g., flight attendants displaying happiness or judges displaying neutrality (Rafaeli and Sutton, 1989). In regulating one's expressions, two major strategies have been identified: surface level, which involves simply changing the appearance of displays, versus deep level, which involves *reappraising* by returning to the appraisal model and reconsidering one's original answers (Grandey, 2003). For example, one can 'look on the bright side' to reappraise valence or find a way to change one's circumstances to reappraise control.

Perception of Emotion

Functions of Emotion Perception

Although the push factors described above emphasize that emotional expression can occur for its own sake, ultimately expressions serve as communication. Each of the modalities for expression described above produces cues that are visible to the social environment, and visible cues can be perceived by others. Brunswik's (1955) general model of perception has been central to research on emotional perception by emphasizing that a psychological state produces cues and cues are recognized, but there is room for error at each of these stages. Looking at the nature of the errors, perception is imperfect not only because we make mistakes but also by design; there may be a kind of 'arms race' between expressers who may wish to shield their emotional state and those who wish to uncover it (Schmidt and Cohn, 2001).

We appear to be hardwired to perceive others' emotions. Encounter with emotional stimuli triggers a rapid neural response that attempts to identify the emotion being perceived. Neural activity identifiable as emotion perception occurs around 100 ms after initial perception of the stimulus, with about 300 ms required before conceptual knowledge of the specific emotion being perceived is activated (Adolphs, 2002). Accumulated evidence suggests neural processes related to emotion are not confined to particular brain regions but involve interaction from various structures in the brain (Lindquist et al., 2012). Although detailed discussion is beyond the scope of the article, perceiving emotional stimuli involves sensory pathways to detect the stimulus, regions of the brain associated with retrieval of memories and conceptual information, and activation of neural systems that facilitate simulation of emotion (Adolphs, 2002; LeDoux, 2012).

Revisiting the Organon model, perception of emotion in others is valuable because expressions of emotions are not only clues about others' internal experience of emotions, but serve as a social signal that offers clues about another person's appraisal of their situation and their likely actions (Campos et al., 1994). Achieving insight into the emotional states of others furnishes one with a survival advantage because a functioning society requires coordination among individuals (Burrows, 2008). Decoding the emotions of others facilitates achievement of shared goals, notably in distinguishing the critical decision to approach versus avoid objects, and the establishment and maintenance of social hierarchies (Keltner and Haidt, 1999; Rosenthal et al., 1979; Schmidt and Cohn, 2001).

Individual Differences in Emotion Perception

A long-standing research program examines individual differences in emotion recognition accuracy and its correlates (Hall et al., 2009; Nowicki and Duke, 1994; Rosenthal et al., 1979). For example, women are better at emotion recognition, as are people of higher education levels and socioeconomic status (Hall et al., 2009; Izard, 1971). As mentioned above, accuracy with emotional expression and perception fall within the larger umbrella of emotional intelligence (Mayer et al., 2008), which has expanded interest in the topic.

There is an advantage to people who can recognize emotions well – better social adjustment, better school performance, and even better workplace success across a wide range of industries and job types (for reviews, see Elfenbein et al., 2007; Hall et al., 2009). That said, being exceptionally good at recognizing emotions can also make people uncomfortable if they feel others are *eavesdropping* on their private feelings (Blanck et al., 1981; Elfenbein and Ambady, 2002).

Given their importance in emotional functioning, are people good at expressing emotions generally good at recognizing emotions? The answer appears to be yes. Accumulated evidence suggests a positive association between expression and perception when expression is measured in terms of an ability (Elfenbein and Eisenkraft, 2010).

Regulation of Emotion Perception

Just as expressers can regulate their displays, perceivers can regulate their interpretations. Decoding rules (Matsumoto, 1989) refer to norms about the appropriate recognition of others' emotions. Perceivers can choose to decode a message inaccurately, or they can decode a message accurately but not allow themselves to respond. Decoding rules can protect perceivers' interests when sensitivity to others' cues may be detrimental. For example, a parent might pretend they did not

witness a display of contempt between two siblings, or tears after their child loses a chess tournament.

Emotional Contagion

Although we have discussed emotion expression and perception as related but separate constructs, and largely as dyadic phenomena, they act together in concert and across multiple individuals. That is, emotions are contagious (Barsade, 2002; Hatfield et al., 1994). Relatively little is known about the mechanisms for *emotional contagion*. The most influential account is *primitive efference* (Zajonc, 1998) – notably that we mimic others' emotional expressions and, through facial feedback, come to experience the same emotions ourselves (Hatfield et al., 1994). However, this mechanism and its magnitude when tested cannot satisfactorily explain empirical findings (Parkinson, 2011). Numerous other mechanisms for the social sharing of emotion have been proposed (Elfenbein, 2014). These include (1) *social comparison*, also called *social appraisal*, in which people compare their feelings with compatriots in order to learn how they are supposed to feel (Barsade, 2002; Parkinson, 2011); (2) *emotional interpretation*, where others' expressive displays serve as information that feeds into the emotional appraisal process and thereby induces emotions (Hareli and Rafaeli, 2008); and (3) *empathy*, or imagining another person's feelings, which is itself emotionally evocative (Davis, 1983; Hatfield et al., 1994). Although most work on contagion focuses on sharing a convergent state, perceivers can also take on a complimentary state (Hess and Fischer, 2013) – e.g., joy in the face of opposing sports fans' disappointment, fear in the face of rage, or contempt in the face of someone's shame.

Just as we may need some minimal perception of other's emotions to converge with them, the reverse can be true in that sharing others' emotions aids the emotion recognition process. The perceiver's own facial expression or activation of facial muscles associated with emotional expression can facilitate or inhibit perception of congruent affective stimuli and also influence the accuracy of emotion perception (Havas et al., 2010; Strack et al., 1988). Along these lines, we may be able to feel our way into perceiving others' expressions.

Cultural Differences in the Expression and Perception of Emotion

Two major mechanisms underlie the cultural differences in emotion that have long been observed, namely norms and styles. Norms govern our display rules and decoding rules, and both of these vary substantially across cultures. Notably, cultures that value greater social harmony tend to inhibit the expression of negative emotion (Mesquita and Frijda, 1992). Styles of emotional expression also vary. According to dialect theory, the specific cues we use to express our emotions – e.g., the configuration of facial muscles, acoustical properties of the voice – differ subtly across cultures (Elfenbein, 2013). Dialect theory was developed to explain the finding of *in-group advantage*, namely that accuracy tends to be higher when perceiving emotions expressed by members of the same versus different cultural group due to greater familiarity.

See also: Disgust, Psychology of; Emotion and Expression; Emotion in Cognition; Emotion, Neural Basis of; Emotional Intelligence and Competencies; Emotional Regulation; Emotions and Aging; Emotions and Health; Emotions and Intergroup Relations; Emotions and Work; Emotions, Psychological Structure of; Emotions: Methods of Assessment; Facial Emotion Expression, Individual Differences in; Facial Perception; Self and Emotional Development in Adulthood and Later Life.

Bibliography

Adolphs, R., 2002. Neural systems for recognizing emotion. Current Opinion in Neurobiology 12, 169–177. http://dx.doi.org/10.1016/S0959-4388(02)00301-X.

Associated Press, 2007. Turns 25 – CNN.com.

Banse, R., Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70, 614–636. http://dx.doi.org/10.1037/0022-3514.70.3.614.

Barrett, L.F., Russell, J.A., 1999. The structure of current affect: controversies and emerging consensus. Current Directions in Psychological Science 8, 10–14. http://dx.doi.org/10.2307/20182545.

Barrett, L.F., Mesquita, B., Gendron, M., 2011. Context in emotion perception. Current Directions in Psychological Science 20, 286–290. http://dx.doi.org/10.1177/0963721411422522.

Barsade, S.G., 2002. The ripple effect: emotional contagion and its influence on group behavior. Administrative Science Quarterly 47, 644–677.

Blair, R.J.R., 2003. Facial expressions, their communicatory functions and neuro–cognitive substrates. Philosophical Transactions of the Royal Society of London. Series B 358, 561–572. http://dx.doi.org/10.1098/rstb.2002.1220.

Blanck, P.D., Rosenthal, R., Snodgrass, S.E., DePaulo, B.M., Zuckerman, M., 1981. Sex differences in eavesdropping on nonverbal cues: developmental changes. Journal of Personality and Social Psychology 41, 391–396. http://dx.doi.org/10.1037/0022-3514.41.2.391.

Brunswik, E., 1955. Representative design and probabilistic theory in a functional psychology. Psychological Review 62, 193.

Bühler, K., 1934. Theory of Language: The Representational Function of Language (D.F. Goodwin, Trans.). Foundations of Semiotics; J. Benjamins Pub. Co., Amsterdam; Philadelphia.

Burrows, A.M., 2008. The facial expression musculature in primates and its evolutionary significance. BioEssays 30, 212–225. http://dx.doi.org/10.1002/bies.20719.

Campos, J.J., Mumme, D.L., Kermoian, R., Campos, R.G., 1994. A functionalist perspective on the nature of emotion. Monographs of the Society for Research in Child Development 59, 284–303. http://dx.doi.org/10.2307/1166150.

Cosmides, L., Tooby, J., 2000. Evolutionary psychology and the emotions. Handbook of Emotions 2, 91–115.

Darwin, C., 1872. The Expression of the Emotions in Man and Animals (Penguin Classics), Original ed. Penguin Classics.

Davidson, R.J., Scherer, K.R., Goldsmith, H.H. (Eds.), 2002. Handbook of Affective Sciences. Oxford University Press, Oxford.

Davis, M.H., 1983. Measuring individual differences in empathy: evidence for a multidimensional approach. Journal of Personality and Social Psychology 44, 113–126. http://dx.doi.org/10.1037/0022-3514.44.1.113.

DePaulo, B.M., 1992. Nonverbal behavior and self-presentation. Psychological Bulletin 111, 203–243. http://dx.doi.org/10.1037/0033-2909.111.2.203.

DePaulo, B.M., Friedman, H.S., 1998. Nonverbal communication. In: Gilber, D.T., Fiske, S.T, Lindzey, G. (Eds.), The Handbook of Social Psychology, vols 1 and 2, fourth ed. McGraw-Hill, New York, pp. 3–40.

DiMatteo, R.M., Hays, R.D., Prince, L.M., 1986. Relationship of physicians' nonverbal communication skill to patient satisfaction, appointment noncompliance, and physician workload. Health Psychology 5, 581–594. http://dx.doi.org/10.1037/0278-6133.5.6.581.

Ekman, P., Friesen, W.V., 1969. Nonverbal leakage and clues to deception. Psychiatry: Journal for the Study of Interpersonal Processes 32, 88–106.

Ekman, P., Rosenberg, E.L., 2005. What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). Oxford University Press, Oxford; New York.

Ekman, P., 1971. Universals and cultural differences in facial expressions of emotion. Nebraska Symposium on Motivation 19, 207–283.

Ekman, P., 1972. Universals and cultural differences in facial expressions of emotion. In: Cole, J. (Ed.), Nebraska Symposium on Motivation, 1971. University of Nebraska Press, Lincoln, pp. 207–283.

Ekman, P., 2003. Emotions Revealed. Henry Holt and Co., New York.

Elfenbein, H.A., Ambady, N., 2002. Predicting workplace outcomes from the ability to eavesdrop on feelings. Journal of Applied Psychology 87, 963–971. http://dx.doi.org/10.1037//0021-9010.87.5.963.

Elfenbein, H.A., Eisenkraft, N., 2010. The relationship between displaying and perceiving nonverbal cues of affect: a meta-analysis to solve an old mystery. Journal of Personality and Social Psychology 98, 301–318.

Elfenbein, H.A., Foo, M.D., White, J., Tan, H.H., Aik, V.C., 2007. Reading your counterpart: the benefit of emotion recognition accuracy for effectiveness in negotiation. Journal of Nonverbal Behavior 31, 205–223. http://dx.doi.org/10.1007/s10919-007-0033-7.

Elfenbein, H.A., 2007. Emotion in organizations: a review and theoretical integration. Academy of Management Annals 1, 371–457.

Elfenbein, H.A., 2007. Emotion in organizations. The Academy of Management Annals 1, 315–386. http://dx.doi.org/10.1080/078559812.

Elfenbein, H.A., 2013. Nonverbal dialects and accents in facial expressions of emotion. Emotion Review 5, 90–96. http://dx.doi.org/10.1177/1754073912451332.

Elfenbein, H.A., 2014. The many faces of emotional contagion: An Affective Process Theory for affective linkage. Organizational Psychology Review.

Ellsworth, P.C., Scherer, K.R., 2003. Appraisal processes in emotion. In: Davidson, R.J., Scherer, K.R., Goldsmith, H.H. (Eds.), Handbook of Affective Sciences. Oxford University Press, New York, pp. 572–595.

Fiske, S.T., Gilbert, D.T., Lindzey, G. (Eds.), 2010. Handbook of Social Psychology, fifth ed. Wiley, Hoboken, NJ.

Fridlund, A.J., 1994. Human Facial Expression: An Evolutionary View. Academic Press, San Diego, CA.

Frijda, N.H., 1986. The Emotions. Cambridge University Press, Cambridge, UK.

Frijda, N.H., 1988. The laws of emotion. American Psychologist 43, 349–358. http://dx.doi.org/10.1037/0003-066X.43.5.349.

Frijda, N.H., 2007. The Laws of Emotion. Lawrence Erlbaum Associates, Mahwah, NJ.

Gallois, C., 1994. Group membership, social rules, and power: a social-psychological perspective on emotional communication. Journal of Pragmatics 22, 301–324. http://dx.doi.org/10.1016/0378-2166(94)90114-7.

Gottman, J.M., 1998. Psychology and the study of marital processes. Annual Review of Psychology 49, 169–197. http://dx.doi.org/10.1146/annurev.psych.49.1.169.

Grandey, A.A., 2003. When "the show must go on": surface acting and deep acting as determinants of emotional exhaustion and peer-rated service delivery. The Academy of Management Journal 46, 86–96. http://dx.doi.org/10.2307/30040678.

Gross, J.J., John, O.P., 1998. Mapping the domain of expressivity: multimethod evidence for a hierarchical model. Journal of Personality and Social Psychology 74, 170–191. http://dx.doi.org/10.1037/0022-3514.74.1.170.

Gross, J.J., Levenson, R.W., 1993. Emotional suppression: physiology, self-report, and expressive behavior. Journal of Personality and Social Psychology 64, 970–986. http://dx.doi.org/10.1037/0022-3514.64.6.970.

Gross, J.J., 2007. Handbook of Emotion Regulation. Guilford, New York.

Halberstadt, A.G., Denham, S.A., Dunsmore, J.C., 2001. Affective social competence. Social Development 10, 79–119. http://dx.doi.org/10.1111/1467-9507.00150.

Halberstadt, A.G., 1986. Family socialization of emotional expression and nonverbal communication styles and skills. Journal of Personality and Social Psychology 51, 827–836. http://dx.doi.org/10.1037/0022-3514.51.4.827.

Hall, J.A., Andrzejewski, S.A., Yopchick, J.E., 2009. Psychosocial correlates of interpersonal sensitivity: a meta-analysis. Journal of Nonverbal Behavior 33, 149–180. http://dx.doi.org/10.1007/s10919-009-0070-5.

Hareli, S., Rafaeli, A., 2008. Emotion cycles: on the social influence of emotion in organizations. Research in Organizational Behavior 28, 35–59. http://dx.doi.org/10.1016/j.riob.2008.04.007.

Hatfield, E., Cacioppo, J.T., Rapson, R.L., 1994. Emotional Contagion, Studies in Emotion and Social Interaction. Cambridge University Press; Editions de la Maison des sciences de l'homme, Cambridge, England; New York, Paris.

Havas, D.A., Glenberg, A.M., Gutowski, K.A., Lucarelli, M.J., Davidson, R.J., 2010. Cosmetic use of botulinum toxin-a affects processing of emotional language. Psychological Science 21, 895–900. http://dx.doi.org/10.1177/0956797610374742.

Hess, U., Fischer, A., 2013. Emotional mimicry as social regulation. Personality and Social Psychology Review 17, 142–157. http://dx.doi.org/10.1177/1088868312472607.

Hochschild, A.R., 1983. The Managed Heart: Commercialization of Human Feeling. University of California Press, Los Angeles.

Izard, C.E., 1971. The Face of Emotion. Appleton-Century-Crofts, New York.

James, W., 1884. What is an Emotion? Mind 9, 188–205. http://dx.doi.org/10.1037/10735-001.

Keltner, D., Haidt, J., 1999. Social functions of emotions at four levels of analysis. Cognition & Emotion 13, 505–521. http://dx.doi.org/10.1080/026999399379168.

Laukka, P., Elfenbein, H.A., 2012. Emotion appraisal dimensions can be inferred from vocal expressions. Social Psychological and Personality Science 3, 529–536. http://dx.doi.org/10.1177/1948550611428011.

LeDoux, J., 2012. Rethinking the emotional brain. Neuron 73, 653–676. http://dx.doi.org/10.1016/j.neuron.2012.02.004.

Lindquist, K.A., Wager, T.D., Kober, H., Bliss-Moreau, E., Barrett, L.F., 2012. The brain basis of emotion: a meta-analytic review. Behavioral and Brain Sciences 35, 121–143. http://dx.doi.org/10.1017/S0140525X11000446.

Matsumoto, D., 1989. Cultural influences on the perception of emotion. Journal of Cross-Cultural Psychology 20, 92–105. http://dx.doi.org/10.1177/0022022189201006.

Matthews, G., Zeidner, M., Roberts, R.D. (Eds.), 2007. The Science of Emotional Intelligence: Knowns and Unknowns. Oxford University Press, New York.

Mayer, J.D., Roberts, R.D., Barsade, S.G., 2008. Human abilities: emotional intelligence. Annual Review of Psychology 59, 507–536. http://dx.doi.org/10.1146/annurev.psych.59.103006.093646.

Mehrabian, A., Ferris, S.R., 1967. Inference of attitudes from nonverbal communication in two channels. Journal of Consulting Psychology 31, 248–252. http://dx.doi.org/10.1037/h0024648.

Mesquita, B., Frijda, N.H., 1992. Cultural variations in emotions: a review. Psychological Bulletin 112, 179–204.

Morris, M.W., Keltner, D., 2000. How emotions work: the social functions of emotional expression in negotiations. Research in Organizational Behavior 22, 1–50.

Nowicki Jr., S., Duke, M.P., 1994. Individual differences in the nonverbal communication of affect: the diagnostic analysis of nonverbal accuracy scale. Journal of Nonverbal Behavior 18, 9–35. http://dx.doi.org/10.1007/BF02169077.

Parkinson, B., 2005. Do facial movements express emotions or communicate motives? Personality and Social Psychology Review 9, 278–311. http://dx.doi.org/10.1207/s15327957pspr0904_1.

Parkinson, B., 2011. Interpersonal emotion transfer: contagion and social appraisal. Social and Personality Psychology Compass 5, 428–439. http://dx.doi.org/10.1111/j.1751-9004.2011.00365.x.

Rafaeli, A., Sutton, R.I., 1989. The expression of emotion in organizational life. In: Staw, B.M., Cummings, L.L. (Eds.), Research in Organizational Behaviour: An Annual Series of Analytical Essays and Critical Reviews, vol. 11. JAI Press, Greenwich, CT, pp. 1–42.

Riggio, R.E., 1986. Assessment of basic social skills. Journal of Personality and Social Psychology 51, 649–660. http://dx.doi.org/10.1037/0022-3514.51.3.649.

Rinn, W.E., 1984. The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions. Psychological Bulletin 95, 52–77. http://dx.doi.org/10.1037/0033-2909.95.1.52.

Rosenthal, R., DePaulo, B.M., 1979. Sex differences in eavesdropping on nonverbal cues. Journal of Personality and Social Psychology 37, 273–285. http://dx.doi.org/10.1037/0022-3514.37.2.273.

Rosenthal, R., Hall, J.A., DiMatteo, M.R., Rogers, P.L., Archer, D., 1979. Sensitivity to Nonverbal Communication: The PONS Test. Johns Hopkins University Press Baltimore.

Russell, J.A., Bachorowski, J.-A., Fernández-Dols, J.-M., 2003. Facial and vocal expressions of emotion. Annual Review of Psychology 54, 329–349. http://dx.doi.org/10.1146/annurev.psych.54.101601.145102.

Russell, J.A., 1994. Is there universal recognition of emotion from facial expressions? A review of the cross-cultural studies. Psychological Bulletin 115, 102–141. http://dx.doi.org/10.1037/0033-2909.115.1.102.

Scherer, K.R., Mortillaro, M., Mehu, M., 2013. Understanding the mechanisms underlying the production of facial expression of emotion: a componential perspective. Emotion Review 5, 47–53. http://dx.doi.org/10.1177/1754073912451504.

Scherer, K.R., 1988. Criteria for emotion-antecedent appraisal: a review. In: Hamilton, V., Bower, G.H., Frijda, N.H. (Eds.), NATO Advanced Study Institutes Series. Series D, Behavioural and Social Sciences. Kluwer, New York, NY, pp. 89–126.

Scherer, K.R., Schorr, A., Johnstone, T. (Eds.), 2001. Appraisal Processes in Emotion. Theory, Methods Research. Oxford University Press, New York.

Schmidt, K.L., Cohn, J.F., 2001. Human facial expressions as adaptations: evolutionary questions in facial expression research. American Journal of Physical Anthropology (Suppl. 33) S3–S24.

Shariff, A.F., Tracy, J.L., 2011. What are emotion expressions for? Current Directions in Psychological Science 20, 395–399. http://dx.doi.org/10.1177/0963721411424739.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R.J., Frith, C.D., 2004. Empathy for pain involves the affective but not sensory components of pain. Science 303, 1157–1162. http://dx.doi.org/10.1126/science.1093535.

Smith, C.A., Ellsworth, P.C., 1985. Patterns of cognitive appraisal in emotion. Journal of Personality and Social Psychology 48, 813–838. http://dx.doi.org/10.1037/0022-3514.48.4.813.

Smith, C.A., Lazarus, R.S., 1990. Emotion and Adaptation. In: Pervin, L.A. (Ed.), Handbook of Personality: Theory and Research. Guildford Press, New York, pp. 609–637.

Strack, F., Martin, L.L., Stepper, S., 1988. Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. Journal of Personality and Social Psychology 54, 768–777. http://dx.doi.org/10.1037/0022-3514.54.5.768.

Susskind, J.M., Anderson, A.K., 2008. Facial expression form and function. Communicative & Integrative Biology 1, 148–149.

Tucker, J.S., Riggio, R.E., 1988. The role of social skills in encoding posed and spontaneous facial expressions. Journal of Nonverbal Behavior 12, 87–97. http://dx.doi.org/10.1007/BF00986927.

Zajonc, R.B., 1998. Emotions. In: Gilbert, D.T., Fiske, S.T., Lindzey, G. (Eds.), The Handbook of Social Psychology, fourth ed., vols. 1 and 2. McGraw-Hill, New York, pp. 591–632.

Emotions and Aging

Tammy English, Washington University in St. Louis, St. Louis, MO, USA
Laura L Carstensen, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Emotional experience and regulation continue to change in a number of ways through adulthood. In many aspects, such changes appear to benefit emotional functioning. The aim of this article is to review the literature on adult development about emotion, including empirical findings and conceptual accounts of age differences in functioning. We focus on four aspects of emotional functioning: (1) subjective experience, (2) physiology, (3) behavioral expression, and (4) regulation. The discussion of theoretical mechanisms underlying age differences in emotion covers a range of different approaches, including biological, environmental, and motivational.

Life span theories of psychology view development as an adaptive process that continues throughout life and entails both gains and losses (Baltes and Baltes, 1990). Research on psychological aging has primarily focused on loss, particularly cognitive decline (e.g., memory, speed of processing; Craik and Salthouse, 2008). When research on emotion began to accumulate, however, a starkly different picture began to emerge, which suggests a relatively positive trajectory.

In this article, we describe the age-related changes in emotion that occur across adulthood, including shifts in the three core components of emotion, namely, emotional experience, physiology, behavioral expression, as well as emotion regulation. We discuss potential explanations for these changes, including biological, environmental, and motivational accounts. Finally, we propose some promising directions for future research.

Age Differences in Emotional Experience

There has been a great deal of work focused on documenting and understanding how emotional processes develop in childhood and adolescence (Campos et al., 2004; Denham, 1998). Far less research has focused on normative emotional changes in adulthood. When researchers initially began to study emotion and aging, the strong expectation was that emotional experience in late life would be characterized by sadness and anxiety. Despite real losses that come with age, such as the deaths of loved ones and health problems, older adults tend to experience similar or higher levels of emotional well-being than do younger adults (see reviews by Charles and Carstensen, 2014; Consedine and Magai, 2006; Scheibe and Carstensen, 2010). This pattern of findings is often referred to as the 'paradox of aging.'

When study participants are asked about emotional experience, responses typically suggest improvements in emotional well-being across adulthood. Most research on emotion and aging has been cross-sectional and subsequently open to alternative explanations. Older adults today, for example, may be uniquely resilient because of their experiences that are specific to the historical period in which they lived (e.g., Great Depression, World Wars). Longitudinal studies, however, have revealed similar patterns. One of the first longitudinal studies on emotion in adulthood focused on global reports using the Bradburn Affect Balance Scale (Charles et al., 2001). Four birth cohorts were assessed between 1971 and 1994. Consistent with cross-sectional work, self-reported negative experience declined over time. A 10-year longitudinal study using experience sampling at 5 year intervals revealed similar findings (Carstensen et al., 2011). It should be noted that some researchers have found that negative affect is stable or slightly increased in very old age (Davey et al., 2004; Diener and Suh, 1998; Fiske et al., 2003), yet when physical health is controlled, age is reliably associated with reduced negative affect (Gatz et al., 1996; Kunzmann et al., 2000) and Gerstorf et al. (2010) found that closeness to death, as opposed to age per se, better accounts for the increase in negative affect. Thus, the trajectory of emotional experience across adulthood is largely positive until, not surprisingly, illness just prior to death frequently takes an emotional toll.

Even though older adults experience negative emotions less frequently than do younger adults, no age differences have been observed in the intensity of emotional experience (Carstensen et al., 2011; Labouvie-Vief et al., 2003; Levenson et al., 1991; Uchino et al., 1999). That is, once emotions are elicited they are felt as intensely in older as younger ages. The combination of findings about frequency and intensity speak against a general flattening of emotional experience. In addition, as individuals grow older, they are more likely to experience positive and negative emotions at the same time, rather than feeling exclusively positive or negative (Carstensen et al., 2011; Ersner-Hershfield et al., 2008), suggesting comparatively richer and more complex experiences. Importantly, mixed emotional experience is positively associated with resilience (Coifman et al., 2007; Ong and Bergeman, 2004) and good physical health (Hershfield et al., 2013).

Most of the improvement in emotional experience appears to be related to a reduction in the experience of negative emotions. Evidence about changes in positive experience has been mixed. For instance, Diener and Suh (1998) observed a slight decrease in positive emotions with age, whereas Mroczek and Kolarz (1998) observed a modest increase in positive emotions across adulthood. Lawton et al. (1992) suggested that disparate findings may reflect the fact that high

arousal emotions become less frequent with age and many measures of positive emotion tap positive emotions with items like excitement and enthusiasm. Indeed, age-related decreases in negative emotion are most apparent for high arousal states (e.g., anger, rage, despair; Lawton et al., 1992; Phillips et al., 2008; Ross and Mirowsky, 2008; Schieman, 1999). Additionally, Scheibe et al. (2013) recently reported an age-related increase in daily low arousal positive emotions in a sample spanning young adulthood to old age, while high arousal positive emotions were age invariant. They also asked participants about *preferences* for specific emotions and found that older people preferred low arousal positive states (e.g., calm) over high arousal positive states. Compared with younger adults, older adults are generally less focused on increasing positive emotion and more interested in maintaining their existing levels of positive emotion (Riediger et al., 2009). Relatedly, older adults have less moment-to-moment variability in emotional experience, suggesting an age-related increase in emotional stability (Carstensen et al., 2011). Charles (2010) has argued that older people are less able to recover physiologically from prolonged and inescapable arousal and thus may avoid situations that may elicit strong emotions.

Overall, the evidence to date suggests that emotional life becomes less negative and more stable as people age. With the exception of the dementias, older adults have a lower incidence of all psychiatric disorders (Jorm, 2000; Kessler et al., 2005; see review by Piazza and Charles, 2006). Additional research is needed to better elucidate how emotional experience changes with age but the overall profile is clearly positive (English and Carstensen, 2014a).

Aging- and Emotion-Related Physiology

Few studies have examined age differences in emotion-related physiological reactivity. One study found that older adults showed a similar pattern of autonomic nervous system activation as younger adults when posing emotional expressions or recalling past emotional experiences (Levenson et al., 1991). The magnitude of activation, however, was lesser for older adults despite the absence of age differences in emotion expression or subjective reports of emotion experience while recalling past events.

There is evidence that general physiology may change with age in ways that make regulation more difficult. In particular, there is reduced physiological flexibility in both the cardiovascular system and the neuroendocrine system (Graham et al., 2006; Sapolsky et al., 1986; for review see Charles, 2010). As a result, older adults are slower to become physiologically aroused and slower to recover. Consistent with this idea, stressors have been found to elicit less of an increase in heart rate and skin conductance among older adults (Labouvie-Vief et al., 2003; Levenson et al., 1991; Tsai et al., 2000). In addition, in studies of the hypothalamic–pituitary–adrenal (HPA) axis, older adults are slower to return to their baseline physiological state after being exposed to a stressor (for reviews see Björntorp, 2002; Otte et al., 2005). Future research is needed to understand when physiological reactivity is reduced and when it is heightened, as well as whether such changes reflect improved abilities to regulate emotional responses (Magai et al., 2006) or more general changes in physiological reactivity.

Age Differences in Emotional Expression

Aging may impact emotional expressions in at least two ways: the ability to detect emotional expressions in others and the ability to convey one's own emotions. In a study that examined perceptions of posed expressions, older adults attributed less negative emotion to expressions of anger, disgust, fear, and sadness and attributed more unintended positive emotion to these negative expressions (Riediger et al., 2011). Similarly, other studies have found that older adults are poorer at detecting negative emotions in faces than younger adults, whereas there is no age difference in perception of happy faces (Ebner and Johnson, 2009; Ebner et al., 2010). These findings are consistent with the age-related positivity effect, whereby older adults tend to focus on the positive aspects of a stimulus more so than the negative aspects compared to younger adults (Reed et al., 2014; Reed and Carstensen, 2012).

Lab studies that instruct individuals to show specific emotions have not observed age differences in the expression of emotion (Levenson et al., 1991). However, a number of studies have found that others have a more difficult time reading the emotional expressions of older adults (Ebner and Johnson, 2009; Malatesta et al., 1987; Riediger et al., 2011). There is some evidence to suggest that this may be somewhat attenuated in older adult perceivers – that is, older adults may be better at reading the expressions of other older adults (Malatesta et al., 1987; Riediger et al., 2011). However, this advantage for age-matched faces is not a strong or consistent pattern across studies (Ebner and Johnson, 2009; Ebner et al., 2010).

Age Differences in Emotion Regulation

People are not simply at the mercy of their emotions; rather they persistently attempt to influence how they feel and what they express to others. The ability to effectively control emotions is crucial for optimal functioning. Emotion regulation involves a broad range of conscious and subconscious strategies that can be used to control different aspects of emotion (Gross, 1998). Each strategy requires specific resources (e.g., executive functioning, knowledge of the social environment) in order to be optimally successful. With repeated experience, individuals may develop a better understanding of how to use these strategies and which strategies to deploy in which situations. That is, emotion regulation can be considered a knowledge-based skill that individuals hone over time. Given older adults' vast life experiences, there is reason to expect that emotion regulation improves with age.

Older adults do report being better able to regulate their emotions well, especially the experience of emotion (Gross et al., 1997). Consistent with such reports, when asked to regulate their emotions under controlled laboratory conditions, older adults are typically equally or more effective than younger adults (Kunzmann et al., 2005; Phillips et al., 2008), especially when using strategies that draw on past life

experiences (e.g., positive reappraisal; Shiota and Levenson, 2009). There is also some evidence to suggest that emotion regulation becomes less effortful with age, such that regulation requires fewer cognitive resources for older adults. For instance, older adults show less memory impairment after being instructed to regulate their emotions than do younger adults (Emery and Hess, 2011; Scheibe and Blanchard-Fields, 2009).

When older adults choose how to regulate their emotions (in the lab or in daily life), they tend to use more effective emotion regulation strategies (Blanchard-Fields et al., 2004). For instance, they habitually rely less on relatively unhealthy strategies, such as expressive suppression, and more on relatively healthy strategies such as cognitive reappraisal (John and Gross, 2004). Compared to younger adults, older adults tend to have more positive appraisals (Almeida and Horn, 2004; Lachman et al., 2008; Story et al., 2007), they are more likely to tailor their regulation efforts to the specific situation (Blanchard-Fields, 2007) and they use more flexible strategies (Diehl et al., 1996). Older adults are also more likely to meet their emotional goals in daily life (Scheibe et al., 2013).

The literature on social preferences and social networks points to ways that antecedent emotion regulation may be a central advantage that older people have in regulating their emotions. Social networks grow smaller as people grow older, primarily due to a reduction in peripheral social partners (Carstensen et al., 2003). Emotionally close partners are retained so networks have a greater ratio of emotionally close partners as people age. It appears that the overall decrease in network size reflects an active pruning process whereby older adults show an increased preference to spend time in emotionally satisfying and predictable interactions and avoid potentially aversive interactions (Lang, 2000). The shaping of social networks reduces exposure to negative stimuli (Birditt and Fingerman, 2005; Folkman et al., 1987) and facilitates conflict avoidance (Blanchard-Fields et al., 2007; Carstensen et al., 1995; Charles et al., 2010).

Theoretical Approaches to Adult Development and Emotion

Theoretically, aging could benefit emotional experience for a number of reasons. As noted above, aging is associated with changes in biological reactivity (Cacioppo et al., 2011) involving structural decline in emotion-related centers of the brain and slowing of the autonomic system (Cacioppo et al., 1998). Therefore, emotion regulation may become easier because emotions are felt less intensely. Although reasonable theoretically, recent research points to well-preserved functioning in regions of the brain that are involved with emotion memory, threat detection, and reward anticipation (Samanez-Larkin and Carstensen, 2011). Thus, biological and cognitive change cannot fully explain change in emotion across adulthood.

Normative changes, like retirement, may make it easier for older adults to maintain more positive, balanced emotional states (Ginn and Fast, 2006). Some have suggested that social interactions are more pleasant for older adults because others treat them with more respect and are less willing to express negative emotions toward them (Fingerman et al., 2008). One recent study, however, found that increased freedom and leisure time did not fully account for older adults' increased emotional well-being (Riediger and Freund, 2008), and the linear improvements in emotional well-being beginning in early adulthood (e.g., Carstensen et al., 2011) speak against other old-age-specific explanations for improvements.

Another related explanation focuses on life experience (e.g., Blanchard-Fields, 2007; Magai et al., 2006). This perspective places more emphasis on the strengths of older adults and their active involvement in creating their positive outcomes. Older adults have had more opportunities to try to control unwanted emotions and see how different experiences impact their emotional state. With growing experience, individuals may develop emotional expertise that allows them to better regulate their emotions and more flexibly apply strategies in an appropriate manner (i.e., choose the strategy that best fits the situation). Consistent with this idea, older adults have been found to have a better sense of the emotional impact of certain future events than do younger adults (Scheibe et al., 2011).

Socioemotional selectivity theory posits that emotion-related goals are increasingly prioritized across adulthood due to shortening time perspectives (Carstensen, 2006; Carstensen et al., 1999). With age, as increased constraints are placed on time left in life, individuals focus their resources on goals that can be realized in the moment (emotional satisfaction and meaning) rather than more future-oriented goals (exploring, expanding horizons). Many studies have provided support for an age-related shift in motivation (for a review, see Carstensen, 2006). Socioemotional selectivity theory postulates that as people age they become increasingly selective in their choice of social partners and activities, investing increasingly in a smaller set of important people. A recent longitudinal analysis of social network change over time supports this claim. The reduction in network size appears to be driven by reductions in acquaintances while the inner circle of close family and friends is retained and concomitant improvements in daily emotional experience accompany these selective reductions (English and Carstensen, 2014b).

Because goals direct cognitive resources, older adults may focus more on emotional information (given its goal relevance, based on socioemotional selectivity theory), especially positive information (given its emotional benefits). Tests of this hypothesis led to the identification of the *positivity effect* in attention and memory; Older people reliably attend to and remember positive information (Mather and Carstensen, 2005; Reed and Carstensen, 2012) and this selective cognitive processing may benefit emotional experience and regulation (Isaacowitz et al., 2008).

The model of strength and vulnerability integration (SAVI; Charles, 2010) combines motivational and knowledge perspectives, positing that older adults are often better able to regulate their emotions (and thus have higher emotional well-being) because of their accumulated experience and self-knowledge, as well as their heightened emotional goals. Due to their physiological vulnerability (i.e., reduced flexibility), however, the model predicts that older adults may not regulate as well in situations that are emotionally intense or prolonged. A similar line of argument suggests that older adults avoid negative and arousing stimuli in an effort to conserve limited cognitive and physical resources (Heckhausen and Schulz,

1995; Labouvie-Vief, 2003). Fortunately, older adults are often able to avoid these kinds of stressors in daily life and as a result they tend to experience less negative affect (Charles et al., 2010).

Future Directions and Conclusions

Research on emotion and aging began only about two decades ago. Initial expectations that old age would be associated with sadness, fear, and despair have not been supported. On the contrary, emotional experience seems to improve with age. Recent studies have also revealed age-related improvements in the ability to regulate emotion. However, many questions remain. We end this article by discussing two promising areas of future research: (1) Intraindividual change in emotion across adulthood and (2) Mechanisms underlying age-related changes in emotion.

There have been a number of cross-sectional studies comparing different aspects of emotion in younger and older adults. These studies have revealed some positive trends, such as improved emotional well-being and ability to regulate emotion later in life. The longitudinal studies that exist also tend to find similar age-related improvements in emotion when tracking individuals as they age. More longitudinal studies are needed to better understand how emotional processes develop across adulthood. In particular, there is a pronounced lack of longitudinal studies focused on the trajectory of emotion regulation, emotional expressivity, and physiological reactivity. These kinds of studies not only can provide better insight into normative trends (i.e., how individuals typically develop emotionally) but also can provide some information about individual differences in change across adulthood. Despite the fact that, on average, individuals show improvements in certain aspects of emotional well-being and regulation, there are individuals who do not show this pattern (e.g., perhaps those who are more neurotic or socially isolated). Future work that is focused on intraindividual shifts in emotion is critical in order to improve the ability to predict who may be at risk for adverse emotional outcomes later in life.

Another important direction for future work is isolating the mechanisms underlying the age-related shifts that occur in emotion. Theoretical explanations span experience-based improvements in emotion regulation to biological deterioration. The motivational account offered by socioemotional selectivity theory (Carstensen, 1992) has been most widely examined. SAVI (Charles, 2010) offers an integration of socioemotional selectivity theory and evidence that in some contexts aging may no longer be protective. Importantly, extant conceptual accounts are not mutually exclusive. In all likelihood, motivation, experience, and biological changes each contribute to emotional experience in later adulthood. Additional studies are needed that address each of these potential mechanisms and examine the role they may play in driving changes in emotional experience, expression, physiology, and regulation.

See also: Aging, Theories of; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Emotions and Health; Emotions, Psychological Structure of; Selfconscious Emotions, Psychology of.

Bibliography

Almeida, D.M., Horn, M.C., 2004. Is daily life more stressful during middle adulthood? In: Brim, O.G., Ryff, C.D., Kessler, R.C. (Eds.), How Healthy Are We? A National Study of Well-Being at Midlife. University of Chicago Press, Chicago, IL, pp. 425–451.

Baltes, P.B., Baltes, M.M., 1990. Psychological perspectives on successful aging: the model of selective optimization with compensation. In: Baltes, P.B., Baltes, M.M. (Eds.), Successful Aging: Perspectives from the Behavioral Sciences. Cambridge University Press, New York, pp. 1–34.

Birditt, K.S., Fingerman, K.L., 2005. Do we get better at picking our battles? Age group differences in descriptions of behavioral reactions to interpersonal tensions. Journals of Gerontology: Series B. Psychological Sciences and Social Sciences 60, P121–P128.

Björntorp, P., 2002. Alterations in the ageing corticotropic stress – response axis. Novartis Foundation Symposium 242, 46–58.

Blanchard-Fields, F., 2007. Everyday problem solving and emotion—An adult developmental perspective. Current Directions in Psychological Science 16, 26–31.

Blanchard-Fields, F., Mienaltowski, A., Seay, R.B., 2007. Age differences in everyday problem-solving effectiveness: older adults select more effective strategies for interpersonal problems. Journals of Gerontology: Series B. Psychological Sciences and Social Sciences 62, P61–P64.

Blanchard-Fields, F., Stein, R., Watson, T.L., 2004. Age differences in emotion-regulation strategies in handling everyday problems. Journals of Gerontology: Series B. Psychological Sciences and Social Sciences 59, 261–269.

Cacioppo, J.T., Berntson, G.G., Bechara, A., Tranel, D., Hawkley, L.C., 2011. Could an aging brain contribute to subjective well being? The value added by a social neuroscience perspective. In: Tadorov, A., Fiske, S.T., Prentice, D. (Eds.), Social Neuroscience: Toward Understanding the Underpinnings of the Social Mind. Oxford University Press, New York, pp. 249–262.

Cacioppo, J.T., Berntson, G.G., Klein, D.J., Poehlmann, K.M., 1998. Psychophysiology of emotion across the life span. In: Schaie, K.W., Lawton, M.P. (Eds.), Annual Review of Gerontology and Geriatrics, Focus on Emotion and Adult Development, vol. 17. Springer, New York, pp. 27–74.

Campos, J.J., Frankel, C.B., Camras, L., 2004. On the nature of emotion regulation. Child Development 75, 377–394.

Carstensen, L.L., 1992. Social and emotional patterns in adulthood: support for socioemotional selectivity theory. Psychology and Aging 7, 331–338.

Carstensen, L.L., 2006. The influence of a sense of time on human development. Science 312, 1913–1915.

Carstensen, L.L., Fung, H., Charles, S., 2003. Socioemotional selectivity theory and the regulation of emotion in the second half of life. Motivation and Emotion 27, 103–123.

Carstensen, L.L., Gottman, J.M., Levenson, R.W., 1995. Emotional behavior in long-term marriage. Psychology and Aging 10, 140–149.

Carstensen, L.L., Gross, J., Fung, H., 1997. The social context of emotional experience. In: Lawton, M.P., Schaie, K.W. (Eds.), Annual Review of Geriatrics and Gerontology. Springer, New York, pp. 325–352.

Carstensen, L.L., Isaacowitz, D.M., Charles, S.T., 1999. Taking time seriously: a theory of socioemotional selectivity. American Psychologist 54, 165–181.

Carstensen, L.L., Mikels, J.A., 2005. At the intersection of emotion and cognition: aging and the positivity effect. Current Directions in Psych Logical Science 14, 117–121.

Carstensen, L.L., Pasupathi, M., Mayr, U., Nesselroade, J.R., 2000. Emotional experience in everyday life across the adult life span. Journal of Personality and Social Psychology 79, 644–655.

Carstensen, L.L., Turan, B., Scheibe, S., Ram, N., Ersner-Hershfield, H., Samanez-Larkin, G.R., Brooks, K.P., Nesselroade, J.R., 2011. Emotional experience improves with age: evidence based on over 10 years of experience sampling. Psychology and Aging 26, 21–33.

Charles, S.T., 2010. Strength and vulnerability integration: a model of emotional well-being across adulthood. Psychological Bulletin 136, 1068–1091.

Charles, S.T., Carstensen, L.L., 2014. Emotion regulation and aging. In: Gross, J.J. (Ed.), Handbook of Emotion Regulation, second ed. Guilford Press, New York, pp. 203–218.

Charles, S.T., Luong, G., Almeida, D.M., Ryff, C., Sturm, M., Love, G., 2010. Fewer ups and downs: daily stressors mediate age differences in negative affect. Journals of Gerontology: Series B. Psychological Sciences and Social Sciences 65, 279–286.

Charles, S.T., Reynolds, C.A., Gatz, M., 2001. Age-related differences and change in positive and negative affect over 23 years. Journal of Personality & Social Psychology 80, 136–151.

Coifman, K.G., Bonanno, G.A., Rafaeli, E., 2007. Affect dynamics, bereavement and resilience to loss. Journal of Happiness Studies 8, 371–392.

Consedine, N.S., Magai, C., 2006. Emotional development in adulthood: a developmental functionalist review and critique. In: Hoare, C. (Ed.), Handbook of Adult Development and Learning. Oxford University Press, New York, pp. 123–148.

Craik, F.I.M., Salthouse, T.A., 2008. The Handbook of Aging and Cognition, third ed. Psychology Press, New York.

Davey, A., Halverson Jr., C.F., Zonderman, A.B., Costa Jr., P.T., 2004. Change in depressive symptoms in the baltimore longitudinal study of aging. Journals of Gerontology: Series B. Psychological Sciences and Social Sciences 59, P270–P277.

Denham, S.A., 1998. Emotional Development in Young Children. Guilford Press, New York.

Diehl, M., Coyle, N., Labouvie-Vief, G., 1996. Age and sex differences in strategies of coping and defense across the life span. Psychology and Aging 11, 127–139.

Diener, E., Suh, E.M., 1998. Subjective well-being and age: an international analysis. In: Schaie, K.W., Lawton, M.P. (Eds.), Annual Review of Gerontology and Geriatrics: Vol. 17. Focus on Emotion and Adult Development. Springer, New York, pp. 304–324.

Ebner, N.C., Johnson, M.K., 2009. Young and older emotional faces: are there age group differences in expression identification and memory? Psychology and Aging 9, 329–339.

Ebner, N.C., Riediger, M., Lindenberger, U., 2010. FACES – a database of facial expressions in young, middle-aged, and older women and men: development and validation. Behavior Research Methods 42, 351–362.

Emery, L., Hess, T.M., 2011. Cognitive consequences of expressive regulation in older adults. Psychology and Aging 26, 388–396.

English, T., Carstensen, L.L., 2014a. Emotional experience in the mornings and the evenings: consideration of age differences in specific emotions by time of day. Frontiers in Psychology 5, 185.

English, T., Carstensen, L.L., 2014b. Selective narrowing of social networks across adulthood is associated with improved emotional experience in daily life. International Journal of Behavioral Development 38, 195–202.

Ersner-Hershfield, H., Mikels, J.A., Sullivan, S.J., Carstensen, L.L., 2008. Poignancy: mixed emotional experience in the face of meaningful endings. Journal of Personality and Social Psychology 94, 158–167.

Fingerman, K.L., Miller, L., Charles, S.T., 2008. Saving the best for last: how adults treat social partners of different ages. Psychology and Aging 23, 399–409.

Fiske, A., Gatz, M., Pedersen, N., 2003. Depressive symptoms and aging: the effects of illness and non-health-related events. Journals of Gerontology: Series B. Psychological Sciences and Social Sciences 58, P320–P328.

Folkman, S., Lazarus, R.S., Pimley, S., Novacek, J., 1987. Age differences in stress and coping processes. Psychology and Aging 2, 171–184.

Gatz, M., Kasl-Godley, J.E., Karel, M.J., 1996. Aging and mental disorders. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, fourth ed. Academic Press, San Diego, CA, pp. 365–382.

Gerstorf, D., Ram, N., Mayraz, G., Hidajat, M., Lindenberger, U., Wagner, G.G., Schupp, J., 2010. Late-life decline in well-being across adulthood in Germany, the UK, and the US: something is seriously wrong at the end of life. Psychology and Aging 25, 477–485.

Ginn, J., Fast, J., 2006. Employment and social integration in midlife: preferred and actual time use across welfare regime types. Research on Aging 28, 669–690.

Graham, J.E., Christian, L.M., Kiecolt-Glaser, J.K., 2006. Stress, age, and immune function: toward a lifespan approach. Journal of Behavioral Medicine 29, 389–400.

Gross, J.J., 1998. The emerging field of emotion regulation: an integrative review. Review of General Psychology 2, 271–299.

Gross, J.J., Carstensen, L.L., Pasupathi, M., Tsai, J., Gotestam-Skorpen, C., Hsu, A.Y.C., 1997. Emotion and aging: experience, expression, and control. Psychology and Aging 12, 590–599.

Heckhausen, J., Schulz, R., 1995. A life-span theory of control. Psychological Review 102, 284–304.

Hershfield, H.E., Scheibe, S., Sims, T., Carstensen, L.L., 2013. When bad can be good: mixed emotions benefit physical health outcomes across the life span. Social Psychological and Personality Science 4, 54–61.

Hess, U., Adams, R.B., Simard, A., Stevenson, M.T., Kleck, R.E., 2012. Smiling and sad wrinkles: age-related changes in the face and the perception of emotions and intentions. Journal of Experimental Social Psychology 48, 1377–1380.

Isaacowitz, D.M., Toner, K., Goren, D., Wilson, H.R., 2008. Looking while unhappy: mood-congruent gaze in young adults, positive gaze in older adults. Psychological Science 19, 848–853.

John, O.P., Gross, J.J., 2004. Healthy and unhealthy emotion regulation: personality processes, individual differences, and life span development. Journal of Personality 72, 1301–1333.

Jorm, A.F., 2000. Does old age reduce the risk of anxiety and depression? A review of epidemiological studies across the adult life span. Psychological Medicine 30, 11–22.

Kessler, R.C., Berglund, P., Demler, O., Jin, R., Merikangas, K.R., Walters, E.E., 2005. Lifetime prevalence and age-of-onset distributions of DSM–IV disorders in the national comorbidity survey replication. Archives of General Psychiatry 62, 593–602.

Kunzmann, U., Kupperbusch, C.S., Levenson, R.W., 2005. Behavioral inhibition and amplification during emotional arousal: a comparison of two age groups. Psychology and Aging 20, 144–158.

Kunzmann, U., Little, T.D., Smith, J., 2000. Is age-related stability of subjective well-being a paradox? Cross-sectional and longitudinal evidence from the Berlin aging study. Psychology and Aging 15, 511–526.

Labouvie-Vief, G., 2003. Dynamic integration: affect, cognition, and the self in adulthood. Current Directions in Psychological Science 12, 201–206.

Labouvie-Vief, G., Lumley, M.A., Jain, E., Heinze, H., 2003. Age and gender differences in cardiac reactivity and subjective emotion responses to emotional autobiographical memories. Emotion 3, 115–126.

Lachman, M., Röcke, C., Rosnick, C., Ryff, C.D., 2008. Realism and illusion in Americans' temporal views of their life satisfaction: age differences in reconstructing the past and anticipating the future. Psychological Science 19, 889–897.

Lang, F.R., 2000. Endings and continuity of social relationships: maximizing intrinsic benefits within personal networks when feeling near to death. Journal of Social and Personal Relationships 17, 155–182.

Lawton, M.P., Kleban, M.H., Rajagopal, D., Dean, J., 1992. Dimensions of affective experience in three age groups. Psychology and Aging 7, 171–184.

Levenson, R.W., Carstensen, L.L., Friesen, W.V., Ekman, P., 1991. Emotion, physiology, and expression in old age. Psychology and Aging 6, 28–35.

Magai, C., Consedine, N.S., Krivoshekova, Y.S., Kudadjie-Gyamfi, E., McPherson, R., 2006. Emotion experience and expression across the adult life span: insights from a multimodal assessment study. Psychology and Aging 21, 303–317.

Malatesta, C.Z., Izard, C.E., Culver, C., Nicolich, M., 1987. Emotion communication skills in young, middle-aged, and older women. Psychology and Aging 2, 193–203.

Mather, M., Carstensen, L.L., 2005. Aging and motivated cognition: the positivity effect in attention and memory. Trends in Cognitive Sciences 9, 496–502.

Mroczek, D.K., Kolarz, C.M., 1998. The effect of age on positive and negative affect: a developmental perspective on happiness. Journal of Personality and Social Psychology 75, 1333–1349.

Ong, A.D., Bergeman, C.S., 2004. The complexity of emotions in later life. Journal of Gerontology: Psychological Sciences 59B, P55–P60.

Otte, C., Hart, S., Neylan, T.C., Marmar, C.R., Yaffe, K., Mohr, D.C., 2005. A meta-analysis of cortisol response to challenge in human aging: importance of gender. Psychoneuroendocrinology 30, 80–91.

Phillips, L.H., Henry, J.D., Hosie, J.A., Milne, A.B., 2008. Effective regulation of the experience and expression of negative affect in old age. Journal of Gerontology: Psychological Sciences 63B, 138–145.

Piazza, J.R., Charles, S.T., 2006. Mental health among the baby boomers. In: Whitbourne, S.K., Willis, S.L. (Eds.), The Baby Boomers Grow up: Contemporary Perspectives on Midlife. Erlbaum, Mahwah, NJ, pp. 111–146.

Reed, A.E., Carstensen, L.L., 2012. The theory behind the age-related positivity effect. Frontiers in Emotion Science 3, 1–9.

Reed, A.E., Chan, L., Mikels, J.A., 2014. Meta-analysis of the age-related positivity effect: age differences in preferences for positive over negative information. Psychology and Aging 29, 1–15.

Riediger, M., Freund, A.M., 2008. Me against myself: motivational conflicts and emotional development in adulthood. Psychology and Aging 23, 479–494.

Riediger, M., Schmiedek, F., Wagner, G.G., Lindenberger, U., 2009. Seeking pleasure and seeking pain: differences in prohedonic and contra-hedonic motivation from adolescence to old age. Psychological Science 20, 1529–1535.

Riediger, M., Voelkle, M.C., Ebner, N.C., Lindenberger, U., 2011. Beyond "happy, angry, or sad?": age-of-poser and age-of-rater effects on multi-dimensional emotion perception. Cognition and Emotion 25, 968–982.

Ross, C.E., Mirowsky, J., 2008. Age and the balance of emotions. Social Science & Medicine 66, 2391–2400.

Samanez-Larkin, G.R., Carstensen, L.L., 2011. Socioemotional functioning and the aging brain. In: Decety, J., Cacioppo, J.T. (Eds.), The Handbook of Social Neuroscience. Oxford University Press, New York, pp. 507–521.

Sapolsky, R.M., Krey, L.C., McEwen, B.S., 1986. The neuroendocrinology of stress and aging: the glucocorticoid cascade hypothesis. Endocrine Reviews 7, 284–301.

Scheibe, S., Blanchard-Fields, F., 2009. Effects of emotion regulation on cognitive performance: what is costly for young adults is not so costly for older adults. Psychology and Aging 24, 217–223.

Scheibe, S., Carstensen, L.L., 2010. Emotional aging: recent findings and future trends. Journals of Gerontology: Psychological Sciences 65, 135–144.

Scheibe, S., English, T., Tsai, J.L., Carstensen, L.L., 2013. Striving to feel good: ideal affect, actual affect, and their correspondence across adulthood. Psychology and Aging 28, 160–171.

Scheibe, S., Mata, R., Carstensen, L.L., 2011. Age differences in affective forecasting and experienced emotion surrounding the 2008 U.S. Presidential election. Cognition and Emotion 25, 1029–1044.

Schieman, S., 1999. Age and anger. Journal of Health and Social Behavior 40, 273–289.

Shiota, M., Levenson, R., 2009. Effects of aging on experimentally instructed detached reappraisal, positive reappraisal, and emotional behavior suppression. Psychology and Aging 24, 890–900.

Story, T.N., Berg, C.A., Smith, T.W., Beveridge, R., Henry, N.J.M., Pearce, G., 2007. Age, marital satisfaction, and optimism as predictors of positive sentiment override in middle-aged and older married couples. Psychology and Aging 22, 719–727.

Tsai, J.L., Levenson, R.W., Carstensen, L.L., 2000. Autonomic, subjective, and expressive responses to emotional films in older and younger Chinese Americans and European Americans. Psychology and Aging 15, 684–693.

Uchino, B.N., Uno, D., Holt-Lunstad, J., Flinders, J.B., 1999. Age-related differences in cardiovascular reactivity during acute psychological stress in men and women. Journals of Gerontology: Series B. Psychological Sciences and Social Sciences 54, P339–P346.

Emotions and Health

Ayse K Uskul, School of Psychology, University of Kent, Canterbury, UK
Andrea B Horn, University of Zuerich, Zuerich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G.D. Bishop, volume 7, pp. 4454–4459, © 2001, Elsevier Ltd.

Abstract

Experiencing health problems can produce strong emotional responses that may impact psychological well-being and disease outcome. Also emotional states have been found to be associated with the development of disease. Negative emotions have been found to be associated with health complaints. Conversely, positive emotions have been found to be associated with better health. Emotions may affect health indirectly by influencing behaviors known to be associated with health outcomes or through physiological mechanisms such as physiological reactivity and changes in immune functioning. Approaches for helping people deal more effectively with negative emotions and maximize positive ones are discussed.

Emotions have long been believed to play a significant role in health. However, definitive evidence on the nature of this role and the mechanisms involved is still being developed. On the one hand, patients often have emotional reactions to the diagnosis of serious or life-threatening illness that can influence psychological as well as physical well-being. On the other hand, emotions and their expression or inhibition have been linked with the development of various disease conditions. This article explores what is currently known about these links between emotions and health and the implications for both theory and intervention.

Following a period dominated by behaviorism and stimulus–response approaches to behavior, there has, of late, been an explosion of interest in the emotions, their nature, antecedents, and consequences. Beginning in the 1980s, this interest has been brought about by a growing acceptance of a cognitive meditational approach to human behavior, which goes beyond a simple stimulus–response model of human behavior to seeing behavior as being strongly mediated by internal psychological processes of which, of course, emotions are a part (Lazarus, 1993). However, there still remains a dispute on the nature of emotions and their definition, which is beyond the scope of this article. For the purposes of this article, we follow the general consensus of the definition in which emotions are seen as the affective states, which can be differentiated into identifiable valenced qualities, are embedded in a social and cultural context, are experienced internally (subjective feeling), manifest themselves on a behavioral level (e.g., as emotional expression), and are associated with identifiable patterns of physiological activity.

There is consensus on the functional note of emotional states in the first place – negative emotions are in many situations adequate and serve important psychological and social functions. However, in many cases emotions might interfere with goals and need to be regulated. In recent years, *emotion regulation* has been subject to an increasing amount of research. It is defined as the shaping of which emotion is experienced and expressed when and to what extent (Gross, 2014). Emotion regulation can be automatic or occur under awareness and may engage in different phases of the experience of an emotion (e.g., before the full blown emotional experience evolves or afterward). (Dys-)functional emotion regulation is seen as a core feature of mental health and of high relevance in the context of physical health. Prolonged negative affect resulting in depressive or anxiety symptoms can be conceptualized as a dysregulation of mood. Recently, the importance of interpersonal processes in emotion regulation has been receiving increasing attention in research (e.g., Debrot et al., 2013). The examination of the links between socioaffective factors and physical health is a promising venue in the search for psychological pathways to physical health.

We focus on emotions as both consequences and antecedents of physical illness. Thus, we begin our discussion with a consideration of the emotional responses to illness (and how they are regulated) and then move on to a discussion of emerging evidence for the role of emotion in the development of disease.

Emotional Responses to Illness

Recognizing the importance of emotions in illness states, Leventhal and his colleagues developed the self-regulation model with a parallel-response framework, which combines a cognitive and an emotional path (e.g., Leventhal et al., 2001). The cognitive path consists of the psychologically 'objective' representation of the health threat with its coping procedures and evaluative processes, whereas the emotional path consists of the 'subjective' or emotional processing system, which creates feeling states, and the coping procedures and appraisal rules for the management of the system.

A wide range of emotional responses can appear at various stages of a health problem in response to related symptoms, diagnoses, and treatments. Emotional responses experienced in the face of illness often lessen with time as the person comes to grips with the condition but can lead to serious emotional maladjustment. Key determinants of quality of emotional adjustment appear to be age, the severity of the condition, coping and emotion regulation strategies, and social support. Coping strategies appear to be critical determinant of emotional adjustment. For example, positive adaptation to breast cancer appears to be facilitated by finding a sense of

meaning in the experience of cancer as well as developing a sense of mastery and control over the disease and using downward comparison to help one feel good about how one is doing in the face of the disease (Taylor, 1983). Finally, the sociointerpersonal context is considered as highly relevant for the emotional adjustment to stress and traumatic experiences (Maercker and Horn, 2013; for example, having supportive social relationships is critical to positive adaptation).

Emotional responses to serious illness can also have important implications for the person's physical well-being. For example, 10-year survival rates for breast cancer patients who responded to their diagnosis with feelings of helplessness, denial, or stoic acceptance have been found to be substantially lower than those for patients who responded to their diagnosis with what has been termed a 'fighting spirit,' which included optimism in the face of their disease and a determination to do everything in their power to overcome the disease (Pettingale, 1984). Furthermore, affective symptoms after a cardiac event are highly predictive not only of psychosocial adjustment but also of physical health-related outcomes. For example, depression following myocardial infarction is associated with a more than twofold increased risk for all cause of mortality (Meier et al., 2011). Particularly, depressive and anxiety symptoms and their associated mood dysregulation are highly linked with the progression of cardiac diseases. Studies show that treating mood disorders after a cardiac event can improve the adjustment process and save health care costs (Ladapo et al., 2012). In general, it can be stated that physical conditions often lead to affective response that, if dysregulated, can translate into affective symptoms and mental health problems such as depression, anxiety, adjustment disorder, and traumatic symptoms. In a meta-analysis in the context of cancer, up to one-third of the researched patients showed clinically significant disordered mood (Mitchell et al., 2011). Thus, life-threatening events such as acute coronary symptoms can provoke posttraumatic affective symptoms, which in turn are associated with worse clinical outcomes (Edmondson et al., 2012).

Emotional responses to physical health problems do not emerge in a social vacuum and may be shaped as a function of individuals' social identities and their relationships with others. For example, the collective self predicts the experience of emotions such as shame and embarrassment in the context of physical health problems, which are likely to result from one's failure to participate successfully in an interdependent relationship, or from doing harm to a relationship (Uskul and Hynie, 2014). Thus, individuals whose self-definitions are derived from their memberships in in-groups or social categories are more likely to experience health problems in terms of how this situation may impact social others and their relationships with them. This is further supported by findings indicating that the endorsement of collective self is associated with the experience of social concerns, concerns that relate to how the health problem will impact significant others' lives or one's relationships with them (Uskul and Hynie, 2007).

Emotions and the Development of Disease

Emotions have also been linked to the development of a variety of disease conditions. In what follows, evidence is examined concerning emotions that are associated with a higher risk of disease as well as others, which appear to be protective against ill health.

Emotions as Health Risk

Not surprisingly, emotions associated with increased risk of disease are negative ones. Evidence has accumulated linking various negative affect to the experience of ill health as well as the development of various disease conditions (Kiecolt-Glaser et al., 2002). Further, there is a tendency for the experience of different negative emotions to be correlated with each other and as such it makes sense to refer to a general tendency toward the experience of negative emotions. This tendency is known as negative affectivity and is closely related to the personality trait of neuroticism. A number of studies have found negative affectivity to be common ingredient running through various health-related measures of stress, emotion, coping, and illness experience. As such it has been argued that findings relating such measures as life stress, pessimism, loneliness, low-life satisfaction, and other measures of psychological distress to poor health can be explained, at least in part, as being a reflection of this general tendency to experience negative emotions and interpret experience in negative terms (Clark and Watson, 1991). Other evidence indicates that it is not just the experience of negative emotions that is associated with poorer health but more specifically the lack of expression of these emotions. Since the 1980s, research evidence has been accumulating on the negative health effects of emotional inhibition. In particular, people with repressive coping styles as well as those actively inhibiting the expression of emotion have been found to have generally poorer health and to show negative psychophysiological as well as immunological changes as a result of suppressing the expression of negative feelings (Pennebaker, 1995).

Although these general variables of negative affectivity and emotional inhibition are associated with poor health generally, there is also good reason to believe that a full understanding of the relationship of emotion to health requires examination of the ways in which specific emotions relate to health status and to specific disease conditions. In particular, the relationships of different emotions to specific health conditions are somewhat variable, suggesting at least a certain amount of specificity in the role of emotions in health. Numerous studies have found significant associations between anger/hostility and deaths from coronary heart disease (CHD) as well as mortality from all causes, an association that remains significant when other risk factors are statistically controlled for (Siegman and Smith, 2013). Anger and hostility also appear to be related to other disease conditions such as cancer and arthritis but the strength of these relationships tends to be somewhat weaker than that found with CHD. Depressive mood is a second emotional state with significant health implications. The most obvious association between depression and health is with suicide. However, depression has been found to have an association with a number of other medical conditions as well. A bidirectional association between depression and medical conditions is assumed. On the one hand, there is evidence that depression is associated with an increased likelihood of conditions such as heart disease and cancer as well as with overall mortality

(Wulsin et al., 1999). On the other hand, depression has been found to be associated with poorer prognosis for patients diagnosed with conditions such as AIDS, cancer, and heart disease. For example, meta-analyses show depression to be a significant predictor of mortality among cardiac patients (Barth et al., 2004) whereas similar findings have been obtained for cancer patients with a 25% increased mortality rate for patients reporting even subclinical depressive symptoms (Satin et al., 2009). Relatedly, grief has been found to have significant health implications. Several studies have shown increased illness as well as mortality among individuals who have been recently bereaved (Stroebe et al., 2007). Socio-affective pathways are also suggested to play a role in the established link between divorce and mortality (Sbarra et al., 2011). More recently, the experience of more social emotions have been linked to physical health. For example, individuals who experience the highest levels of shame in situations involving social-evaluative threat also report the highest increase in cortisol (Kemeny et al., 2004). Moreover, individuals who experience shame in a task that induces self-blame show the greatest elevations in proinflammatory cytokine activity, while guilt and general negative emotions are not associated with cytokine changes (Dickerson et al., 2004).

Emotions as Protector of Health

The flip side of the coin is that positive emotions may be protective of health (Salovey et al., 2000). Evidence for this is not as strong as for the effects of negative emotions but links are being established. For example, positive emotional experience in early life is associated with longevity 6 years later (Danner et al., 2001) and positive affect predicts 2-year survival rates in an elderly population (Ostir et al., 2000).

A well-established piece of evidence for positive emotions as a protector of health comes from work on optimism. Optimism refers to a person's tendency to look on the positive side of events and to believe that events will work out for the best. As such optimists can be expected to experience positive emotions more often than more pessimistic individuals. Research on optimism indicates that optimists tend to live longer, healthier lives than more pessimistic individuals and also to be less susceptible to viral infections, such as the common cold. In addition, optimistic cancer patients tend to live longer than more pessimistic ones and optimism is associated with greater immune system competence as indicated by the ratio of T4 to T8 cells (e.g., Kubansky et al., 2001).

More recent work has investigated two other positive emotions and health, namely hope and curiosity, and shown that, across hypertension, diabetes mellitus, and respiratory tract infection, higher levels of hope and curiosity were associated with lower levels of disease development (Smart Richman et al., 2005). This evidence further contributes to past research showing the protective role of positive emotions in the development of disease.

Until recently, research examining the role of emotions in health outcomes has been limited to studies in developed countries in the West. Asking whether the emotion–health connection is a first-world problem, Pressman et al. (2013) have used survey data from individuals from 142 countries asking about positive and negative emotions and self-reported health and found that both positive and negative emotions have a unique and moderate association with self-reported health. These associations remain when controlling for data on food, shelter, and safety. Associations between positive emotions and health are stronger in low-GDP countries compared to high-GDP countries. Thus, the emotion–health link seems not to be limited to developed countries, but can be observed across the globe.

In a more fine-grained cultural analysis of the emotional–health link, Miyamoto and Ryff (2011) show that experiencing both positive and negative emotions moderately frequently, thus experiencing both types of emotions in a balanced manner, is associated with fewer illness symptoms among Japanese, but not among Americans. Such balanced experience of positive and negative emotions among Japanese has been claimed to assist them in showing a better adjustment to both positive and negative cues in their social environment, a stronger cultural imperative among Japanese than Americans.

Mechanisms Linking Emotions and Health

As noted above there is accumulating evidence that emotions are significantly related to health and indications that emotional states may play a detrimental or a protective role. How can these relationships be explained? What are the mechanisms linking emotions and health? The answers to these questions are still emerging but a number of plausible hypotheses have been proposed. These hypotheses relate to both indirect and direct effects that emotions may have on disease processes.

Indirect Links between Emotions and Health: The Health Behavior Connection

One possibility is that emotions influence health indirectly through their effects on behavior, particularly health behaviors. Numerous studies have noted that stress, which often includes emotional upset, is associated with poorer health behaviors. For example, higher levels of stress have been associated with higher levels of smoking as well as increased alcohol and drug usage and poorer eating habits. Specific emotions also appear to be associated with health behaviors. For example, there is evidence that higher levels of hostility are associated with greater consumption of cigarettes and alcohol as well as less physical exercise and poorer self-care (Smith, 1992). Another example concerns the link demonstrated between negative affect and treatment nonadherence in the case of HIV (Gonzales et al., 2011). Supportive evidence for this possible link also comes from research investigating pathways between positive emotions and health demonstrating that positive emotion is associated with greater health-enhancing behavior such as better diet and regular exercise (Pressman and Cohen, 2005; Steptoe et al., 2009).

Direct Links between Emotions and Health: The Physiological Connection

There also appear to be direct physiological links between emotions and health. In addition to their experiential aspects

emotions are very much physiological events. There has been much controversy over the years about the precise relationship of emotions to physiology, whether there are specific patterns of physiological changes that correspond to specific emotions, and whether the phenomenological experience of emotion precedes or is the result of such physiological changes, but there is no doubt that emotions are associated with a variety of physiological changes. These changes occur in a number of different bodily systems including the nervous, cardiovascular, respiratory, endocrine, and muscular systems and vary considerably in intensity. A substantial number of studies have attempted to chart these changes and relate them to specific emotions. Results from such studies have been mixed in that physiological changes found for specific emotions tend to vary considerably between studies and with the methods used for evoking emotion as well as measuring physiological changes. In many respects, this variation is not surprising given the complexity of physiological events and the methodological hurdles in establishing their relationship to particular emotions. Although there is considerable controversy about the precise relationship between emotions and physiology, progress has been made in charting likely physiological pathways between emotions and specific diseases. Some of the most first evidence came from work on the relationship of anger and hostility to cardiovascular disease. A number of studies have found that individuals high in anger and hostility tend to show heightened physiological, and particularly cardiovascular, responses to various stressors. Although results have not been entirely consistent, there is evidence that individuals who are high in hostility, and who are thus more prone to the experience of anger, show larger stress-related increases in blood pressure, heart rate, and stress-related hormones than do individuals lower in hostility. This heightened physiological reactivity appears particularly likely when the person is faced with social stressors such as harassment or other social challenge and is considered by some to be at least a marker if not a causal agent for the future development of cardiovascular disease (Smith, 1992). Recently, the focus of emotion research has shifted from researching general physiological indicators of activation to affective neuroscience (Davidson et al., 2000).

In addition to heightened reactivity in the face of stressors, it is also quite possible that high hostile individuals experience more situations evoking anger, either because they gravitate toward such situations and/or because their actions tend to evoke hostile responses from others. For example, there is evidence that high levels of hostility are associated with generally high levels of interpersonal conflict and low levels of social support. In addition, high hostility has been associated with the experience of more frequent and severe daily hassles as well as major negative life events. This tends to suggest a synergism between poor interpersonal relationships and high physiological reactivity, which would further increase the high hostile individual's vulnerability to disease (Smith, 1992).

Another important link concerns the relationship of emotion to immune functioning. In this respect, there is a growing literature showing that the tendency to experience certain emotions can affect immunocompetence, as can transient moods. First, indices of immunocompetence have been found to be lower among individuals with depressive symptomotology as well as among recently bereaved individuals who strongly suggest that dysphoric moods may lead to the suppression of immune functioning (Kiecolt-Glaser et al., 2002). One possible explanation for this is that withdrawal responses and the generally more passive coping responses associated with helplessness and depression are associated with the activation of the hypothalamic–pituitary–cortical axis, which leads to increases in cortisol output, which in turn have the effect of suppressing immune functioning. Such processes have important implications for a variety of immune-related diseases including cancer, AIDS, and autoimmune diseases. Recently, inflammation has been identified as a crucial immunological process linking negative affective states and a variety of physical health risks such as cardiovascular disease. There is evidence that adaptive emotion regulation is associated with lower and accordingly maladaptive emotion regulation with higher levels of inflammation markers (Appleton et al., 2013).

There is also growing evidence showing that changes in physiological systems represent a pathway between positive emotion and health in adulthood (Ong, 2010). For example, trait positive emotion is associated with lower salivary cortisol output, lower ambulatory heart rate, and reduced fibrinogen responses (Steptoe et al., 2005). Finally, there is increasing support for the suppression and expression of emotion having both physiological and immunological effects. Inhibition of emotion has been shown to lead to changes in electrodermal as well as cardiovascular activity. Also, in addition to evidence cited above demonstrating that suppression of negative emotions is related to reduced immune functioning, there is evidence from a growing number of studies that the expression of negative feelings as a form of catharsis can result in improved immune functioning. For example, several studies have now demonstrated that having people write or talk about traumatic experiences they have not disclosed to other people previously results in an increase in immune responsiveness (Pennebaker and Chung, 2011).

Recent comparative research has put to test the assumption that previously observed physiological pathways might be universal. Examining the link between negative emotions and proinflammatory markers (specifically interleukin-6, IL-6) among Americans and Japanese, Miyamoto et al. (2013) have shown that, as expected, the negative emotions predicted higher IL-6 among Americans, whereas such a prediction was absent among Japanese for whom the experience of negative emotions are seen as more acceptable (and sometime expected) than for Americans (Eid and Diener, 2001; Uchida and Kitayama, 2009). This pattern remained after controlling for a wide range of demographic, psychological, and health-related variables.

Common Underlying Factors between Emotions and Health: Constitutional Vulnerability

A third possible link between emotions and health is that there may be common underlying factors for both emotionality and the occurrence of disease. For example, there may be genetic factors that influence, on the one hand, the emotions a person experiences and how those emotions are expressed and, on the other hand, the occurrence of ill health in general as well as specific disease conditions. For example, recently a common underlying genetic mechanism of inflammation and depressed

mood as a behavioral manifestation has been suggested (Raison et al., 2006).

Reducing Health Risk

The emerging pattern of linkages between emotions and health raises the question of what can be done to reduce the health risk from negative emotions and boost the benefits from positive ones. Three possibilities are considered here but these are best seen as only representative of interventions that may be effective. First, psychotherapeutic techniques can be used to help people deal more effectively with negative emotions as well as boost positive emotions; replacing maladaptive emotion regulation strategies by more adaptive ones is the goal of many recent intervention approaches (Gross, 2014). Numerous techniques have been developed, and there is reason to believe that at least some of these may have positive benefits for physical health as well as mental health. For example, psychotherapy interventions have been shown to be effective in boosting immune responsiveness and improving physical health. Recently, psychotherapy with cardiac patients has been shown to lead to better outcomes (Ladapo et al., 2012), which is consistent with data showing that psychotherapy with a variety of medically ill patients can lead to improved health outcomes (Fawzy and Fawzy, 1994). Of course, psychotherapy does more than simply reduce dysfunctional negative emotions and boost positive ones but such data, in conjunction with the evidence on the role of emotions in health, suggest the potential of psychotherapy for reducing the negative effects of emotions on health.

Closely related to these effects of psychotherapy in general is work showing the psychological and physical benefits of writing or talking about traumatic events. Research by Pennebaker and others has produced substantial evidence for the positive psychological and physical benefits of writing as well as talking about traumatic events. Overall, when people, ranging from university students to holocaust survivors, are induced to write or talk about traumatic experiences positive psychological benefits are found and there are reductions in harmful arousal as well as improvements in immune functioning and subsequent reductions in the number of times the person seeks medical attention. Such changes are particularly evident when the person writes or talks about traumas that had not been previously disclosed to others and when he or she writes about the emotions involved and not just the facts of the situation (Pennebaker and Chung, 2011). This strongly argues for the health benefits of disclosure of negative experiences to others and suggests an apparently powerful way of reducing the harmful effects of emotion on health. It has to be noted that a series of individual, social, and cultural factors may moderate this link between expression of emotions and physical health (Consedine et al., 2002).

Finally, evidence on the role of anger and hostility in heart disease strongly argues for interventions directed toward reducing these potentially harmful emotions. A number of strategies have been suggested and there is evidence that teaching people to be aware of their feelings and then utilizing a series of strategies for dealing with difficult life situations can have positive effects both on the person's relationships and psychological adjustment and possibly on their physical health as well (Suls, 2013).

Implications and Future Directions

Findings concerning the relationships of emotion to health are a prime example of the way in which psychological, social, and cultural factors are closely intertwined with biological processes in determining one's health status. Disease is clearly a biological phenomenon but it is one that is strongly influenced by psychosocial and cultural factors. Also, phenomena such as emotions that are generally viewed in the realm of the psychological have a definite basis in biological processes. As such one cannot speak of psychological processes without consideration of the biological and vice versa.

Although the evidence relating emotions to health has been growing rapidly, there are still many questions that remain to be answered. One major question concerns the degree to which specific emotions are related to specific disease conditions. There has long been speculation about such relationships but as yet there has been no clear resolution of this question. Whereas work on such topics as negative affectivity tends to suggest a general relationship between negative emotions and ill health, other work tends to suggest at least a degree of specificity. This is a question that needs further work and one which will benefit from a clearer specification of how one can best differentiate between different emotions both psychologically and physiologically. Relatedly, there are still many questions about the mechanisms linking emotions and health. Evidence for the roles of physiological reactivity and immunocompetence in this linkage is growing but much remains to be learned. Exactly how is it that different emotions produce these physiological changes? What other biological processes operate in linking emotion to health? How are genetic factors involved in this relationship? A growing, but still limited area of research is the role of cultural factors in the emotion–health link. Evidence from diverse regions of the world with different physical, religious, economic, and interpersonal environments is desperately needed to reach a more generalizable body of knowledge. Finally, much remains to be done with respect to capitalizing on what has been learned in this area for improving health. Current interventions have been described but these would appear to only scratch the surface of what might be possible with respect to harnessing psychological methods for improving physical health.

See also: Coping, Personality and; Culture and Emotion; Culture and the Self: Implications for Psychological Theory; Depression, Pessimism, and Health; Emotion and Expression; Emotion, Neural Basis of; Emotional Regulation; Neuroticism.

Bibliography

Appleton, A.A., Buka, S.L., Loucks, E.B., Gilman, S.E., Kubzansky, L.D., 2013. Divergent associations of adaptive and maladaptive emotion regulation strategies with inflammation. Health Psychology 32, 748–756.

Barth, J., Schumacher, M., Herrmann-Lingen, C., 2004. Depression as a risk factor for mortality in patients with coronary heart disease: a meta-analysis. Psychosomatic Medicine 66, 802–813.

Clark, L.A., Watson, D., 1991. General affective dispositions in physical and psychological health. In: Synder, C.R., Forsyth, D.R. (Eds.), Handbook of Social and Clinical Psychology: The Health Perspective. Pergamon Press, New York.

Consedine, N.S., Magai, C., Bonanno, G.A., 2002. Moderators of the emotion inhibition-health relationship: a review and research agenda. Review of General Psychology 6, 204–228.

Danner, D.D., Snowdon, D.A., Friesen, W.V., 2001. Positive emotions in early life and longevity: findings from the nun study. Journal of Personality and Social Psychology 80, 804–813.

Davidson, R.J., Jackson, D.C., Kalin, N.H., 2000. Emotion, plasticity, context, and regulation: perspectives from affective neuroscience. Psychological Bulletin 126, 890–909.

Debrot, A., Schoebi, D., Perrez, M., Horn, A.B., 2013. Touch as an interpersonal emotion regulation process in couples' daily lives the mediating role of psychological intimacy. Personality and Social Psychology Bulletin 39, 1373–1385.

Dickerson, S.S., Kemeny, M.E., Aziz, N., Kim, K.H., Fahey, J.L., 2004. Immunological effects of induced shame and guilt. Psychosomatic Medicine 66, 124–131.

Eid, M., Diener, E., 2001. Norms for experiencing emotions in different cultures: inter- and intranational differences. Journal of Personality and Social Psychology 81, 869–885.

Edmondson, D., Richardson, S., Falzon, L., Davidson, K.W., Mills, M.A., Neria, Y., 2012. Posttraumatic stress disorder prevalence and risk of recurrence in acute coronary syndrome patients: a meta-analytic review. PLoS One 7, e38915.

Fawzy, F.I., Fawzy, N.W., 1994. Psychoeducational interventions and health outcomes. In: Glaser, R., Kiecolt-Glaser, J.K. (Eds.), Handbook of Human Stress and Immunity. Academic Press, San Diego, CA.

Gonzales, J.S., et al., 2011. Depression and HIV/AIDS treatment nonadherence: a review and meta-analysis. Journal of Acquired Immune Deficiency Syndromes 58, 181–187.

Gross, J.J. (Ed.), 2014. Handbook of Emotion Regulation, second ed. Guilford, New York.

Kemeny, M.E., Gruenewald, T.L., Dickerson, S.S., 2004. Shame as a basic emotional response to threats to the social self: behavioral, physiological, and health effects (commentary). Psychological Inquiry 15, 153–160.

Kiecolt-Glaser, J.K., McGuire, L., Robles, T., Glaser, R., 2002. Emotions, morbidity, and mortality: new perspectives from psychoneuroimmunology. Annual Review of Psychology 53, 83–107.

Kubansky, L.D., Sparrow, D., Vokonas, P., Kawachi, I., 2001. Is the glass half empty or half full? A prospective study of optimism and coronary heart disease in the normative aging study. Psychosomatic Medicine 63, 910–916.

Ladapo, J.A., Shaffer, J.A., Fang, Y., Ye, S., Davidson, K.W., 2012. Cost-effectiveness of enhanced depression care after acute coronary syndrome: results from the coronary psychosocial evaluation studies randomized controlled trial. Archives of Internal Medicine 172, 1682–1684.

Lazarus, R.S., 1993. From psychological stress to the emotions: a history of changing outlooks. Annual Review of Psychology 44, 1–21.

Leventhal, H., Leventhal, E.A., Cameron, L., 2001. Representations, procedures, and affect in illness self-regulation: a perceptual-cognitive model. In: Baum, A., Revenson, T., Singer, J.E. (Eds.), Handbook of Health Psychology. Lawrence Erlbaum, Mahwah, pp. 19–47.

Maercker, A., Horn, A.B., 2013. A socio-interpersonal perspective on PTSD: the case for environments and interpersonal processes. Clinical Psychology & Psychotherapy 20, 465–481.

Meier, A., Conradi, H.J., Bos, E.H., Thombs, B.D., van Melle, J.P., de Jonge, P., 2011. Prognostic association of depression following myocardial infarction with mortality and cardiovascular events: a meta-analysis of 25 years of research. General Hospital Psychiatry 33, 203–216.

Mitchell, Alex J., et al., 2011. Prevalence of depression, anxiety, and adjustment disorder in oncological, haematological, and palliative-care settings: a meta-analysis of 94 interview-based studies. The Lancet Oncology 12, 160–174.

Miyamoto, Y., Boylan, J.M., Coe, C.L., Curhan, K., Levine, C.S., Markus, H.R., Park, J., Kitayama, S., Kawakami, N., Karasawa, M., Love, G.D., Ryff, C.D., 2013. Negative emotions predict elevated interleukin-6 in the United States but not in Japan. Brain, Behavior, and Immunity 34, 79–85.

Miyamoto, Y., Ryff, C., 2011. Cultural differences in the dialectical and non-dialectical emotional styles and their implications for health. Cognition and Emotion 25, 22–30.

Ong, A.D., 2010. Pathways linking positive emotion and health in later life. Current Directions in Psychological Science 19, 358–362.

Ostir, G.V., Markides, K.S., Black, S.A., Goodwin, J.S., 2000. Emotional wellbeing predicts subsequent functional independence and survival. Journal of the American Geriatrics Society 48, 473–478.

Patterson, T.L., Shaw, W.S., Semple, S.J., Cherner, M., McCutchan, J.A., Atkinson, J.H., Grant, I., Nannis, E., 1996. HIV Neurobehavioral Research Center (HNRC) group relationship of psychosocial factors to HIV disease progression. Annals of Behavioral Medicine 18, 30–39.

Pennebaker, J.W., 1995. Emotion, disclosure, and health. American Psychological Association, Washington, DC.

Pennebaker, J.W., Chung, C.K., 2011. Expressive writing and its links to mental and physical health. In: Friedman, H.S. (Ed.), Oxford handbook of Health Psychology. Oxford University Press, New York, pp. 417–437.

Pettingale, K.W., 1984. Coping and cancer prognosis. Journal of Psychosomatic Research 28, 363–364.

Pressman, S.D., Cohen, S., 2005. Does positive affect influence health? Psychological Bulletin 131, 925–971.

Pressman, S.D., Gallagher, M., Lopez, S., 2013. Is the emotion-health connection a "first world problem"? Psychological Science 24, 544–549.

Raison, C.L., Capuron, L., Miller, A.H., et al., 2006. Cytokines sing the blues: inflammation and the pathogenesis of depression. Trends in Immunology 27, 24–31.

Salovey, P., Rothman, A.J., Detweiler, J.B., Steward, W.T., 2000. Emotional states and physical health. American Psychologist 55, 110–121.

Satin, Jillian R., Linden, Wolfgang, Phillips, Melanie J., 2009. Depression as a predictor of disease progression and mortality in cancer patients. Cancer 115 (22), 5349–5361.

Sbarra, D.A., Law, R.W., Portley, R.M., 2011. Divorce and death. A meta-analysis and research agenda for clinical, social, and health psychology. Perspectives on Psychological Science 6, 454–474.

Siegman, A.W., Smith, T.W. (Eds.), 2013. Anger, Hostility, and the Heart. Psychology Press.

Smart Richman, L., Kubzansky, L.D., Maselko, J., Kawachi, I., Choo, P., Bauer, M., 2005. Positive emotion and health: going beyond the negative. Health Psychology 24, 422–429.

Smith, T.W., 1992. Hostility and health: current status of a psychosomatic hypothesis. A review. Health Psychology 11, 139–150.

Steptoe, A., Dockray, S., Wardle, J., 2009. Positive affect and psychobiological processes relevant to health. Journal of Personality 77, 1747–1775.

Stroebe, M., Schut, H., Stroebe, W., 2007. Health outcomes of bereavement. Lancet 370, 1960–1973.

Steptoe, A.P.A., Wardle, J., Marmot, M., 2005. Positive affect and health-related neuroendocrine, cardiovascular and inflammatory processes. Proceedings of the National Academy of Sciences of the United States of America 102, 6508–6512.

Suls, J., 2013. Anger and the heart: perspectives on cardiac risk, mechanisms and interventions. Progress in Cariovascular Diseases 55, 538–547.

Taylor, S.E., 1983. Adjustment to threatening events: a theory of cognitive adaptation. American Psychologist 38, 1161–1173.

Uchida, Y., Kitayama, S., 2009. Happiness and unhappiness in east and west: themes and variations. Emotion 9, 441–456.

Uskul, A.K., Hynie, M., 2007. Self-construal and concerns elicited by imagined and real health problems. Journal of Applied Social Psychology 37, 2156–2189.

Uskul, A.K., Hynie, M., 2014. The role of self-aspects in emotions elicited by threats to physical health. Psychology & Health 29, 199–217.

Williams, V., Williams, R.B., 1997. Life Skills. Times Books, New York.

Wulsin, L.R., Vaillant, G.E., Wells, V.E., 1999. A systematic review of the mortality of depression. Psychosomatic Medicine 61, 6–17.

Emotions and Intergroup Relations

Colin W Leach, Fouad Bou Zeineddine, and Atilla Cidam, University of Connecticut, Storrs, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Emotion has long been considered important to the relations between groups because emotions are powerful collective experiences that people share with like-minded others. As such, emotion enables coordinated and concerted action as a group, vis a vis other groups. As such, the emotions that people feel about intergroup relations are useful to social and behavioral science because they are potent symbols for the meaning people give to their experience. In this article, we focus on recent social psychological work on anger, empathy, pride, shame, and guilt.

The (social, political, economic) relations between groups have long been considered passionate affairs. Since antiquity, observers have focused on pride in one's nation, hate for one's enemies, and shame for the foibles of one's fraternity. Because emotions were long considered irrational and impulsive states, a great deal of classical thought viewed emotion suspiciously as a tempest in the otherwise tranquil sea of social life. This view of emotion continued in much of early twentieth-century social scientific work on antagonistic social relations evidenced in crowds, war, protest, revolution, prejudice, caste, and conflict. For example, in much of the psychological work on stereotypes, prejudice, and discrimination before the 1960s, feelings of indignation, disgust, envy, and even hatred were thought to provide the psychological animus behind the oppression of societally devalued or disadvantaged groups. Influenced by psychoanalytic notions that destructive feelings and impulses submerged below the surface were the primary drivers of intergroup antagonism, many social and behavioral scholars viewed groups as something like Freud's primal horde blinded by their bloodlust and rapacious desire for power (for reviews, see Duckitt, 1992; Mackie and Smith, 2002). Even the seemingly benevolent feeling of love and pride for one's own group was refracted into a latent ethnocentrism that positioned own as superior and other as inferior when viewed through the prevailing lens.

Parallel developments in approaches to emotion and intergroup relations in the 1960s ushered in a radically different view, however. At the same time as cognitive, phenomenological, and semiotic (or meaning-based) approaches to emotion began to replace psychodynamic notions, sociological and psychological approaches to intergroup relations began to increasingly understand emotion as a marker of the social and psychological quality of the actively constructed relations between groups. Nowhere is this convergence of ideas more clear than in the sociological and psychological work on the concept of relative deprivation in the 1960s and 1970s. Much of this work was an attempt to understand the social, political, and economical bases of the feelings of anger and resentment expressed by people around the world disenfranchised by colonialism, poverty, segregation, sexism, and other large-scale forces that promote inequality between groups (for a review, see Walker and Smith, 2002). What sociologists, psychologists, and economists came to see is that the anger in the streets, in Jakarta, in Watts, in Algiers, in Paris, was not the spitting, envious, infantile, and irrational anger of the mob described by thinkers such as Freud, Le Bon, or Tarde. Rather, the anger in the streets, in the music, and in the posture of the people, was an anger of resistance. And, the action of this anger was often calculated, targeted, and the result of organized efforts to rebel. Although many everyday accounts, and some scholarly ones, continue to describe collective rebellion as 'riots' of rage and 'mob mentality,' this view of emotion in intergroup relations is increasingly regarded as misguided (van Zomeren et al., 2012; Walker and Smith, 2002; see also Scott, 1977).

Anger in Intergroup Relations

At the heart of the revolution in thinking about group-based emotion and its role in rebellion was the idea that individuals can feel anger, resentment, and dissatisfaction about grievances that they share with other members of their in-group. This is called (a sense of) relative deprivation (for a review, see Walker and Smith, 2002). And, this shared anger is a basis for shared action to coordinate in-group members to rebel against those viewed as responsible for illegitimate deprivation. In fact, a comprehensive empirical summary by Smith and Ortiz (2002) showed peoples' feelings of group deprivation to be a potent predictor of the desire to protest, as well as actual protest. In contrast, feeling dissatisfied about individual grievances has little link to protest, instead predicting psychological difficulties such as stress and poor well-being. Recent integrative work on relative deprivation in sociology and social psychology shows that this form of anger can operate in conjunction with quite careful consideration of the material and other resources at one's disposal and with deliberative estimations of the effectiveness of various strategies of protest. Thus, rather than being an irrational and impulsive basis of destructive antagonism, anger about a group grievance is often part of a larger constellation of feelings, thoughts, and motivations aimed at best addressing the grievance (van Zomeren et al., 2012). A growing body of research on anger in intergroup relations more generally also shows that anger is a common response to a sense of grievance and that it routinely motivates people to resist, rebel, or otherwise take antagonistic action aimed at redressing a perceived wrong (for reviews, see Mackie and Smith, 2002; Parkinson et al., 2005). This is consistent with appraisal theories of emotion, which argue that

 http://dx.doi.org/10.1016/B978-0-08-097086-8.24057-9

the emotion of anger is based in part on appraisals (or interpretations) of an unpleasant event as a wrong that should be righted by some form of antagonistic action toward those responsible (see Lazarus, 1991).

Although the experience of anger may be seen as implying that one has some modicum of power or control to act (Lazarus, 1991), the calculus of how to act most effectively appears to be a distinguishable process (van Zomeren et al., 2012). In fact, impulsive or otherwise poorly planned resistance can undermine the apparent purpose of anger by leading people to engage in unwise or ineffective action. Combining anger with strategic concerns for the efficacy of possible action may be especially important in intergroup relations marked by dramatic differences in power, status, and moral commitment. Like in all relationships, in intergroup relationships the aggrieved party must gauge what can and should be done with its anger. This determination should be based in its assessment of the solidarity of feeling in the group and its ability to act in the concerted and coordinated way required for effective protest (van Zomeren et al., 2012). Of course, even the most organized group must also strategize in response to the coordination, power, status, and likely response of its adversary. Thus, it is important to be aware that emotions such as anger do not necessarily lead to antagonistic action irrespective of the quality of the intergroup relation in which the anger develops. Emotions are, after all, relational. As emotions emerge within relationship their expression in feeling and action also emerges within the relationship and is thus rooted in it (Lazarus, 1991; Tiedens and Leach, 2004). A group of workers, aggrieved at a cut in pay, is likely to express and act on its anger in the context of a cooperative relationship with benevolent management in a way quite different to another in a competitive relationship with blatantly exploitative or callous management. Anger in a cooperative or otherwise benevolent relationship is more likely to manifest as a complaint oriented toward constructive efforts at changing the other party or the relationship itself for the better. In contrast, anger in a competitive or otherwise malevolent relationship is more likely to be a protest oriented toward destructive efforts at punishing the other party, weakening their ability to do further wrong, or reducing the material or psychological importance of the relationship itself (for discussions, see Parkinson et al., 2005; Tiedens and Leach, 2004).

Other Examples of Emotions in Intergroup Relations

Although work on anger and relative deprivation in intergroup relations continues, social psychological work in the last 15 years has come to focus on a wider range of emotions in a wider range of intergroup relations than the gross inequality typically studied in work on relative deprivation (for reviews, see Mackie and Smith, 2002; Parkinson et al., 2005; Tiedens and Leach, 2004). This recent effort was aided by Smith's (1993) important ideas integrating the social psychology of group identity, emotion, and intergroup relations. Since then, social psychologists have developed emotion-based approaches to relative group advantage (Leach et al., 2002), stereotypes (Fiske et al., 2002), prejudice (Cottrell and Neuberg, 2005), and a host of other phenomena that occur within intergroup relations (for reviews, see Iyer and Leach, 2008; Mackie and Smith, 2002; Parkinson et al., 2005).

Empathy

In many ways, empathy is the most general way in which we can understand how the relations between parties affect them emotionally. Indeed, empathy appears to be the psychological process by which the boundaries between people are transcended to increase understanding and responsiveness. A prominent view in social psychology is that there are different kinds of empathy. Three types are most often discussed: perspective-taking, empathic concern (i.e., sympathy, compassion), and empathic distress (Davis, 1983). However, empathy can also be thought of simply as the ability to simulate, or mirror, the cognitive, motivational, affective, and/or behavioral state of another organism (Iacoboni, 2009). Both perspectives seem valid. According to the ethologist De Waal (2008), empathy and imitation share a hard-wired simulation mechanism observable in humans and many other social organisms. This mechanism allows for simple and effortless types of empathy such as emotional contagion, mirroring, and motor mimicry. But, as cognitive capacities increased over evolutionary history, this state-matching became more complex among social species to include more elaborate forms of empathy (e.g., concern for other and perspective-taking) and imitation (e.g., coordination, shared goals, emulation).

Many empathy researchers are concerned with the question of how empathy translates into altruism and prosocial behavior. De Waal's (2008) examination of the evolution of empathy argues that the adaptiveness of empathy is determined by the degree of self–other differentiation, trust in reciprocity, and the value given reciprocal altruism. Thus, where self and other are highly differentiated – perhaps because of tribal, ethnic, national, or other group distinctions – there is likely to be less empathy. Less empathy is also expected in structural, political, and cultural contexts where reciprocity is not prioritized or is not possible. Consistent with this view of empathy as rooted in social relations, psychological research has shown that empathy is less likely when others are not valued (Batson et al., 2007), and when others are seen as unfair (Singer et al., 2006).

Self–other distinctions in power are also associated with lower empathy. For example, powerful people engage less in perspective-taking (Galinsky et al., 2006), and show less empathic concern and more egocentric allocation than the less powerful (Dana et al., 2007). Where 'market' exchange relations and competitiveness are inherent 'rules,' the less powerful are seen as responsible for their own advancement. Thus, the less powerful are 'fair game' for the more powerful to exploit without concern for their welfare. Importantly, many of these trends can be eliminated by changing the rules of the game to espouse reciprocity (van der Weele et al., 2010). And inducing empathy can also eliminate the egocentrism that the powerful display in a 'business transaction' environment (Batson and Moran, 1999). However, in such environments the *completely* powerless are not seen as viable competitors for resources nor as responsible for their own advancement. This evokes among the powerful feelings of social responsibility and altruistic

behavior instead of competitive dominance and egocentric behavior (Handgraaf et al., 2008).

Although much of the above work focuses on individuals' empathy toward other individuals, it is clear that similar processes occur in the empathy that individuals experience toward members of out-groups (for a review, see Weiner, 1995). For example, Fong (2007) found that poor people, who are seen as having *some* power and as responsible for their poverty, invite less empathy and help from others. Similarly, empathy for groups victimized by war or oppression is increased by valuing the group and by not seeing the group as responsible for its own suffering (e.g., Cehajic et al., 2009). At an even simpler level, intergroup distinctions between self and other affect the level of empathy that people experience. Thus, whether in the majority or the minority, ethnic groups show greater empathic neural responses to in-group than to out-group members (for a review, see Chiao and Mathur, 2010).

Although people appear to be especially empathetic to fellow in-group members, this too appears to be modulated by power and status. In fact, members of low power or otherwise subordinate groups show an even greater bias in favor of empathizing with in-group members than do members of dominant groups (Chiao and Mathur, 2010). Consistent with De Waal's (2008) more general view, this is likely the result of the lesser value and trust given to intergroup reciprocity by subordinate groups, who have more reason to wary of reciprocity in unequal intergroup relations. Indeed, members of subordinate groups distrust dominant groups and authorities more than do others, often with good reason (for a review, see Bou Zeineddine and Pratto, 2014). Especially for subordinate groups then, being more empathic toward the in-group than the dominant out-group reduces the individual costs and risks associated with nonreciprocated empathy and altruism. In contrast, for powerful or dominant groups, there is little cost, little risk (and little value) in empathizing with the less powerful even if they are unlikely to reciprocate (see also Leach et al., 2002). The greater prevalence of opportunity-hoarding, dominance, and exploitation among advantaged groups who have overcome the problem of subsistence has strengthened their orientation toward competition and profit (see Scott, 1977). When combined with recent shifts toward greater individualism and narcissism, such dominance may be expected to reduce empathy (see Konrath et al., 2011).

The literature on empathy also shows many ways in which empathy can be used to improve intergroup relations. In their review, Batson and Ahmad (2009) identify perspective-taking as especially promising in reducing stereotyping, increasing situational explanations of suffering, and fostering positive attitudes, empathic concern, and willingness to help out-groups. Because of this, many interventions aim to improve intergroup relations by inducing or encouraging empathy for the disadvantaged. For example, the encouragement of empathy is used in conflict resolution and reconciliation workshops and antidiscrimination and bullying educational programs. In many of these interventions, personalizing contact with individual members of disadvantaged groups, sympathy-provoking examples, and/or the identification of common goals initiate perspective-taking. Perspective-taking, and the associated situational explanations of suffering, generate empathic concern for the out-group member. Concern for the out-group member translates to concern for the out-group as a whole. It is unclear, however, how such efforts at using empathy to reduce prejudice and improve understanding might translate into larger scale change at the cultural, political, or structural level. There is a danger that the effects of empathy interventions remain psychological and local, never translating to the core aspects of the intergroup relation that are likely to have reduced empathy in the first place. Future work would do well to better integrate the macro- and microelements of empathy so that this important concept can be better used to examine and improve intergroup relations.

Pride and Other Positive Emotions

Likely because unpleasant emotions are considered more common or more important in intergroup relations, there is decidedly less work on pleasant feelings, such as pride, joy, and satisfaction (for reviews, see Iyer and Leach, 2008; Parkinson et al., 2005). This parallels work on emotion more generally, which also tends to parse unpleasant emotions more finely and to thereby focus on a larger variety of unpleasant emotions (see Lazarus, 1991; Tiedens and Leach, 2004). Often, when pleasant emotions are examined, they are taken to be problematic for intergroup relations. For example, some recent studies examine Schadenfreude – satisfaction at seeing an out-group suffer a defeat or other misfortune. Gloating over the inferiority of out-groups is another emotion that is implied in work on expressions of in-group superiority, nationalism, collective narcissism, and the excessive glorification of the in-group (for a discussion, see Leach et al., 2002). This has some similarity to notions of individual narcissism and hubristic pride, as such positive feelings are thought to be exaggerated and unnecessarily antagonistic (Tracy et al., 2007).

There is surprisingly little work on the seemingly more prevalent and pervasive feeling of pride that people feel about their national, ethnic, political, and other in-group memberships (for a review, see Leach et al., 2002). However, the experience of pride is often implied in a lot of theory and research on the experience and expression of positive in-group identity. Whether higher or lower in power and status, in-groups that afford positive identification are presumed to lead individuals to feel proud of their membership in the group. Such pride is expected to lead members to view the group and its future more optimistically and it is expected to lead members to be more committed to the group and be more willing to sacrifice for the group's benefit. However, more work is needed to determine whether, when, and why pride in one's in-group is related to the perceived in-group superiority and hostility toward out-groups that are more obviously troublesome for intergroup relations.

Shame and Guilt

In an effort to understand perpetrator groups' acknowledgment of their mistreatment of out-groups and their possible role in processes of restitution and reconciliation, researchers have recently focused on self-critical emotions about in-group wrongdoing like shame and guilt (for reviews, see Branscombe and Doosje, 2004; Iyer and Leach, 2008; Parkinson et al., 2005). Although the words 'shame' and 'guilt' are often used

interchangeably in everyday English, they are typically conceptualized as two distinct emotions that have distinct implications for intergroup relations in light of group wrongdoing. The prevalent view in emotion theory and research is that guilt is a feeling of regret and blameworthiness for causing another party harm. As such, guilt is thought to be linked to empathy for the other's suffering and thus a desire to make restitution for it (e.g., apology, compensation). Shame is thought to be a more intense and aversive state of self-criticism focused on the perpetrator's flaw rather than the harm caused the victim. Thus, shame is thought to be less moral than guilt, motivating perpetrators to avoid their misdeeds and those harmed by them (Tracy et al., 2007).

There is now ample evidence that individuals can feel shame and guilt about the wrongdoing of their group (for reviews, see Branscombe and Doosje, 2004; Parkinson et al., 2005). Like most other such group-based emotions, shame and guilt are greater when individuals feel psychologically connected to their in-group and recognize its role in wrongdoing against undeserving victims (for a review, see Iyer and Leach, 2008). For example, Iyer et al. (2007) found students in the US and the UK to report feeling a little guilty and a little more ashamed about their country's occupation of Iraq and its detrimental effects on the citizenry. Both feelings were associated with holding the in-group responsible and viewing its actions as illegitimate. Consistent with this, both shame and guilt predicted a greater desire to see the US and the UK withdraw from Iraq and to compensate Iraqis for the damage done. Although anger was most associated with wanting to confront the respective governments for the wrongdoing, shame and guilt were also associated with this sort of antagonistic political action. These findings are consistent with a great deal of other work showing that individuals' guilt about an in-group's wrongdoing is tied to wanting to make up for the wrongdoing, typically by apologizing or compensating (for a review, see Branscombe and Doosje, 2004). Although the positive effects of shame in Iyer et al. (2007) run in contrast to the general view of this emotion as oriented to the avoidance of wrongdoing, numerous recent studies also show shame to be linked to the constructive approach of in-group misdeeds (for a review, see Gausel and Leach, 2011).

One reason that shame may be linked to wanting to make restitution for in-group wrongdoing is that the feeling of shame can be based in a concern for the in-group's image in the eyes of others. In fact, in Iyer et al. (2007), participants' shame was linked to a concern that the country's social image in the eyes of others was threatened by their occupation of Iraq. One way to reduce this threat to social image is to make restitution to the out-group in an attempt to reestablish the in-group's moral standing (for a review, see Gausel and Leach, 2011). Many countries, for example, make material or symbolic restitution to out-groups they have mistreated in an effort to regain acceptance in 'the international community' after being condemned by them. Indeed, membership in international organizations, such as the United Nations or the European Union, is sometimes predicated on aspirants atoning for past wrongs. The European Union pushing Turkey to acknowledge the mass killing of Armenians in the early twentieth century is only one example of this.

Shame about in-group wrongdoing can also be linked to support for restitution because group members wish to improve the collective self-image damaged by their wrongdoing (Gausel and Leach, 2011). Because shame is more focused on peoples' identity as good, shame may motivate change of the self in a way that guilt's focus on the harm of wrongdoing may not. Where shame is the result of an intense self-criticism for a failing in the self that can be repaired, shame is likely to lead to concerted effort to improve the self (Gausel and Leach, 2011). By improving the group's failing, shame may offer a more profound change that is more promising for the in-group's relation to the out-group relations and is thus more satisfying to the affected out-group. Of course, where shame is felt about an irreparably damaged social image or self-image, such shame is likely to lead to the kind of avoidance of wrongdoing and those affected by it that emotion theory and research consider detrimental to social relations. Thus, the latest theory and research on shame and guilt about in-group perpetrators of wrongdoing suggest that the reparability or irreparability of the wrongdoing is more important to understanding what group members will do than whether they experience shame or guilt. As feelings of self-criticism, both shame and guilt may motivate constructive efforts when this criticism is aimed at a wrongdoing that is seen as reparable through concerted action. On the other hand, feelings of self-criticism, especially the more intense emotion of shame, may motivate an avoidance designed to defend self-image and social image when it appears that little can be done to redress the in-group's wrongdoing. Taken together, this line of thinking suggests that those who wish to encourage constructive efforts to repair wrongdoing and to reconcile perpetrator and victim would do well to encourage perpetrator and victim to view the wrongdoing and the intergroup relation as reparable.

Conclusion

Emotion has long been considered important to the relations between groups. This is because emotions can be powerful collective experiences that people share with like-minded others. Such shared emotion is a sign of the solidarity that enables coordinate and concerted action as a group, such as in a protest or a parade. Although emotions were frequently considered an irrational and impulsive basis of conflict between groups, it should be clear that emotions are now typically thought of as the social and psychological meaning that individuals give to their membership in a group and to its relation with other groups. As such, the emotions that people feel about intergroup relations are useful to social and behavioral science because they are potent symbols for the meaning people give to their experience (Tiedens and Leach, 2004). To know that a group's members are angry about a court decision is to know that they feel that the decision is unfair and that they may be prepared to act against it. In contrast, to know that the group is sad is to know that they are disappointed about the court decision and that they are unsure if they have the wherewithal to respond. Social and behavioral scientists use emotion terms to describe these experiences for the same reason that the people they study use emotions to characterize their own experience; emotions are pregnant with meaning and

motive. Thus, compassion, pride, joy, shame, anger, envy, disappointment, and the like allow us to convey a great deal about what we think, what we feel, and what we wish to do about the world we live in and those we live with. Poets, priests, and politicians all recognize this power in emotion, and thus work at moving people to common feeling for a common purpose. Those who wish to understand such collective mobilization would do well to examine and understand the social psychological processes at the heart of emotion.

See also: Collective Behavior, Social Psychology of; Culture and Emotion; Deindividuation, Psychology of; Intergroup Relations; Social Psychology; Stereotypes in Social Psychology.

Bibliography

Batson, C.D., Ahmad, N.Y., 2009. Using empathy to improve intergroup attitudes and relations. Social Issues and Policy Review 3, 141–177.

Batson, C.D., Eklund, J.H., Chermok, V.L., Hoyt, J.L., Ortiz, B.G., 2007. An additional antecedent of empathic concern: valuing the welfare of the person in need. Journal of Personality and Social Psychology 93, 65–74.

Batson, C.D., Moran, T., 1999. Empathy-induced altruism in a prisoner's dilemma. European Journal of Social Psychology 29, 909–924.

Bou Zeineddine, F., Pratto, F., 2014. Political distrust: the seed and fruit of popular empowerment. In: van Prooijen, J.W., van Lange, P.A.M. (Eds.), Power, Politics, and Paranoia. Cambridge University Press, Cambridge, UK.

Branscombe, N.R., Doosje, B. (Eds.), 2004. Collective Guilt: International Perspectives. Cambridge University Press, New York.

Cehajic, S., Brown, R., Gonzalez, R., 2009. What do I care? Perceived ingroup responsibility and dehumanization as predictors of empathy felt for the victim group. Group Processes and Intergroup Relations 12, 715–729.

Chiao, J.Y., Mathur, V.A., 2010. Intergroup empathy: how does race affect empathic neural responses? Current Biology 20, 478–480.

Cottrell, C.A., Neuberg, S.L., 2005. Differential emotional reactions to different groups: a sociofunctional threat-based approach to "prejudice." Journal of Personality and Social Psychology 88, 770–789.

Dana, J., Weber, R.A., Kuang, J.X., 2007. Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. Economic Theory 33, 67–80.

Davis, M.H., 1983. The effects of dispositional empathy on emotional reactions and helping: a multidimensional approach. Journal of Personality 51, 167–184.

De Waal, F.B.M., 2008. Putting the altruism back into altruism: the evolution of empathy. Annual Review of Psychology 59, 279–300.

Duckitt, J., 1992. The Social Psychology of Prejudice. Praeger, New York.

Fiske, S.T., Cuddy, A.J.C., Glick, P., Xu, J., 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology 82, 878–902.

Fong, C.M., 2007. Evidence from an experiment on charity to welfare recipients: reciprocity, altruism and the empathic responsiveness hypothesis. The Economic Journal 117, 1008–1024.

Galinsky, A.D., Magee, J.C., Inesi, M.E., Gruenfeld, D.H., 2006. Power and perspectives not taken. Psychological Science 17, 1068–1074.

Gausel, N., Leach, C.W., 2011. Concern for self-image and social-image in the management of moral failure: rethinking shame. European Journal of Social Psychology 41, 468–478.

Handgraaf, M.J., Van Dijk, E., Vermunt, R.C., Wilke, H.A., De Dreu, C.K., 2008. Less power or powerless? Egocentric empathy gaps and the irony of having little versus no power in social decision making. Journal of Personality and Social Psychology 95, 1136–1149.

Iacoboni, M., 2009. Imitation, empathy, and mirror neurons. Annual Review of Psychology 60, 653–670.

Iyer, A., Leach, C.W., 2008. Emotion in inter-group relations. European Review of Social Psychology 19, 86–125.

Iyer, A., Schmader, T., Lickel, B., 2007. Why individuals protest the perceived transgressions of their country: the role of anger, shame, and guilt. Personality and Social Psychology Bulletin 33, 572–587.

Konrath, S.H., O'Brien, O.H., Hsing, C., 2011. Changes in dispositional empathy in American College students over time: a meta-analysis. Personality and Social Psychology Review 15, 180–198.

Lazarus, R.S., 1991. Emotion and Adaptation. Oxford University Press, New York.

Leach, C.W., Snider, N., Iyer, A., 2002. "Poisoning the consciences of the fortunate": the experience of relative advantage and support for social equality. In: Walker, I., Smith, H.J. (Eds.), Relative Deprivation: Specification, Development, Integration. Cambridge University Press, New York, pp. 136–163.

Mackie, D.M., Smith, E.R. (Eds.), 2002. From Prejudice to Intergroup Emotions: Differentiated Reactions to Social Groups. Psychology Press, New York.

Parkinson, B., Fischer, A.H., Manstead, A.S.R., 2005. Emotion in Social Relations: Cultural, Group, and Interpersonal Processes. Psychology Press, New York.

Scott, J.C., 1977. The Moral Economy of the Peasant: Rebellion and Subsistence in Southeast Asia. Yale University Press, New Haven, CT.

Singer, T., Seymour, B., O'Doherty, J.P., Stephan, K.E., Dolan, R.J., Frith, C.D., 2006. Empathic neural responses are modulated by the perceived fairness of others. Nature 439, 466–469.

Smith, E.R., 1993. Social identity and social emotions: toward new conceptualizations of prejudice. In: Mackie, D.M., Hamilton, D.L. (Eds.), Affect, Cognition, and Stereotyping: Interactive Processes in Group Perception. Academic Press, San Diego, CA, pp. 297–315.

Smith, H.J., Ortiz, D., 2002. Is it just me? The different consequences of personal and group relative deprivation. In: Walker, I., Smith, H.J. (Eds.), Relative Deprivation Theory: Specification, Development, and Integration, pp. 91–115.

Tiedens, L.Z., Leach, C.W., 2004. The Social Life of Emotions. Cambridge University Press, New York.

Tracy, J.L., Robins, R.W., Tangney, J.P., 2007. The Self-conscious Emotions: Theory and Research. Guilford Press, New York.

Walker, I., Smith, H.J., 2002. Relative Deprivation: Specification, Development, and Integration. University Press, Cambridge.

van der Weele, J.J., Kulisa, J., Kosfeld, M., Friebel, G., 2010. Resisting Moral Wiggle Room: How Robust Is Reciprocity?. IZA Discussion Paper No. 5374. http://ssrn.com/abstract=1728331.

Weiner, B., 1995. Inferences of responsibility and social motivation. Advances in Experimental Social Psychology 27, 1–47.

van Zomeren, M., Leach, C.W., Spears, R., 2012. Protesters as "passionate economists": a dynamic dual pathway model of coping with collective disadvantage. Personality and Social Psychology Review 16, 180–199.

Emotions and Work

Neal M Ashkanasy, University of Queensland Business School, The University of Queensland, Brisbane, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although neglected until the 1990s, the study of emotions in workplace settings is now recognized as a legitimate and important field of research in industrial and organizational psychology. In this article, I review the field of emotion in the workplace from five perspectives, corresponding to five discrete levels of analysis: (1) within-person temporal effects, (2) between-person (personality and attitudes) factors, (3) interpersonal behaviors (perception and communication of emotion), (4) group level (leadership and teams), and (5) organizational level (culture and climate). Within these perspectives, I address the importance of affective events theory, deal with positive and negative affect (state vs trait), as well as emotional intelligence, labor, and contagion. The article concludes with suggestions for future research.

Introduction

Emotions may be defined as the experience of a form of biological response to environmental stimulus, resulting in physical and psychological changes and subsequent readiness for action (Frijda, 1986). While early scholars of industrial and organizational (IO) psychology and organizational behavior (OB) seemed to recognize the importance of studying emotional dimensions (Brief and Weiss, 2002), post-WWII research tended to concentrate on behavioral and cognitive aspects of work, where emotions and affect were subsumed under the broader heading of job satisfaction (Barsade et al., 2003).

This situation began to change, however, following publication of Hochschild's (1983) sociological analysis of emotional labor, which suggested that employees forced to display emotions at odds with what they truly feel, suffer burnout and loss of productivity. The concept of emotional labor sparked increased interest in emotions and soon found its way into mainstream OB literature (e.g., see Rafaeli and Sutton, 1989). Subsequent seminal reviews by Pekrun and Frese (1992) and Ashforth and Humphrey (1995) served to broaden interest in this field further. Subsequently, publication in 1996 of Weiss and Cropanzano's idea of 'affective events' provided a useful framework for studying emotions in the workplace as a dynamic phenomenon. This period marked "the affective revolution in organizational behavior" (Barsade et al., 2003: p. 3). The ongoing level of interest in studying emotions in organizational settings is today reflected in the existence of an annual book series in this topic, *Research on Emotion in Organizations* and an associated biannual conference, the *International Conference on Emotions and Worklife* (http://www.emotionsnet.org). More recently, Ashkanasy and Humphrey (2011a) provided a comprehensive review of emotions in organizational settings, which forms an appropriate basis for this article.

The Multilevel Model of Emotion in Organizations

Ashkanasy and Humphrey (2011a) based their discussion around an integrated model of emotions in organizations developed by Ashkanasy (2003a). The model crosses five levels of analysis and, as such, provides an appropriate framework to structure this article. At the base of the model, Level 1, are temporal variations in emotions and behavior, referred to as 'within-person' variability. Level 2 refers to between-person variability, such as personality and emotional intelligence. Level 3 refers to interpersonal relationships, including perception and communication of emotion, including emotional labor. At Level 4, analysis shifts to group phenomena including team leadership. Finally, at Level 5 of the model, the focus is on the organization as a whole including emotional culture and climate.

Level 1: Within-Person

Ashkanasy and Humphrey (2011a) point out that "Level 1 is best understood in terms of Affective Events Theory (AET: Weiss and Cropanzano, 1996)" (p. 215) and distinguish between positive and negative mood effects (Ashton-James and Ashkanasy, 2005). In this regard, Isen (1987) and Fredrickson (2001) stress the role of *mild positive affect* as a catalyst of creativity and cognitive flexibility. A further noteworthy aspect of within-person variation is that, under this perspective, empirical procedures need to take account in real time of variations in affect and behavior (Fisher, 2008). Research, for example, can use daily diary data (e.g., see Weiss et al., 1999) or experience sampling methods (ESM: Larson and Csikszentmihalyi, 1983; also known as ecological momentary assessment, Weiss and Beal, 2005), where measurements are taken several times over a day. For example, Amabile et al. (2005) found in an ESM-based field study that (mild) positive affect is related to improved product design and innovation.

Other evidence tells us that negative affect can also lead to increased creativity and innovation under particular circumstances. For example, To et al. (2012) found ESM data to suggest that whether positive or negative mood leads increased creative behavior depends on an interaction of situation and personality. This finding is consistent with the 'sadder but wiser' hypothesis put forward by Staw and Barsade (1993). Forgas and George (2001) point out further that employees in

a negative mood are likely to be less susceptible to bias and less likely to be swayed by persuasion (Forgas and East, 2003).

In workplace settings, where tasks are likely to involve substantive cognitive processing, Forgas and George (2001) argue that employees are likely to be subjected to 'affect infusion' (Forgas, 1995), whereby their affective states color their decision making. In demonstration of this effect, Mittal and Ross (1998) found that risk taking under conditions of uncertainty is higher for decision makers in a positive rather than a negative mood. Ashton-James and Ashkanasy (2008) argue that this is because managers in a positive mood tend to be more optimistic in their situation appraisals than when they are in a negative mood, and therefore are more prone to taking risks. Note, however, that risk taking, especially in today's dynamic business environments is not necessarily a bad thing. For example, risks often need to be taken when the organization needs to move quickly to respond to a changed regulatory or market environment. This effect has subsequently been demonstrated in field research by Amabile et al. (2005).

Researchers have also employed diary methods and ESM to unravel some of the enduring mysteries of job satisfaction. Thus, while job satisfaction has long been thought to be an affective variable, Weiss et al. (1999) used a daily diary method to demonstrate that job satisfaction comprises both affective and cognitive components. Fisher (2002) employed ESM to show moreover that, once (cognitive) attitudinal factors are accounted for, affect is in fact unrelated to job satisfaction. Fisher also demonstrated that affect is related to affective commitment, helping behavior, and role conflict over and above the effects of job satisfaction (see also Fuller et al., 2003). In a more recent ESM study, Fisher et al. (2013) found that momentary emotions depend on task appraisal and confidence.

Another widely accepted misconception about job satisfaction is that it is only weakly related to job performance. For example, Judge et al. (2001) reported the results of a meta-analysis of job satisfaction–performance relationship that the correlation was only 0.30. However, Fisher and Noble (2004), in an extension of Fisher's (2002) earlier ESM studies using programmable watches, found that affect and job satisfaction and performance were strongly related within-person (correlations in the 0.70 range). Thus the within-person relationship is considerably stronger than the between-person job satisfaction–job performance.

In summary of Level 1, it seems safe to conclude that AET and ESM have essentially revolutionized our understanding of the role emotions play in the workplace. AET has been consistently supported across numerous diary and ESM studies (see Weiss and Beal, 2005; for a more extensive review).

Level 2: Between-Person

The dominant variable studied at the between-person (or individual difference) level of analysis is emotional intelligence, which comprises four basic abilities or 'branches': (1) recognition of emotions, both in self and others; (2) use of this information in cognitive decision making; (3) understanding of the effects of emotions; and (4) using and managing emotions in behavioral decision making (Mayer and Salovey, 1997). The construct was popularized by *NY Times* journalist Daniel Goleman (1995), however, resulting in exaggerated claims and led to stringent criticism of emotional intelligence (e.g., see Locke, 2005). On the other hand, Ashkanasy and his colleagues (see Ashkanasy and Daus, 2005; Daus and Ashkanasy, 2005; Jordan et al., 2008) resolutely defended the validity and utility of emotional intelligence, claiming that it is a valid individual difference, similar to cognitive intelligence.

In an effort to clarify the controversy, Ashkanasy and Daus (2005) identified three approaches to emotional intelligence, which they referred to as 'streams.' *Stream 1* is based on the Mayer–Salovey four-branch model and is measured using an IQ-style ability measure called the Mayer–Salovey–Caruso Emotional Intelligence Test (Mayer et al., 2002). *Stream 2* research also uses the Mayer–Salovey definition of emotional intelligence, but is measured using self- or peer-report measures (Brackett and Mayer, 2003). *Stream 3* models of emotional intelligence do not use the Mayer–Salovey definition, and are usually measured using self-reports (e.g., Bar-On, 1997). Ashkanasy and Daus discourage the use of Stream 3 models on the basis that they are usually overly correlated with personality. Nonetheless, in terms of establishing a relationship between emotional intelligence and work performance, Joseph and Newman (2010) and O'Boyle et al. (2011) found in independent meta-analyses that all three streams of emotional intelligence demonstrate incremental validity over and above cognitive ability and personality. More recently, Schlaerth et al. (2013) reported meta-analytic results supporting a positive relationship between all streams of emotional intelligence and skill in conflict resolution.

In summary of the role of emotional intelligence at Level 2 of the Ashkanasy (2003a) model, and despite the controversy, the emerging consensus is that emotional intelligence is an important and valid personality characteristic that is positively associated with work performance.

Level 3: Interpersonal Emotions

This level focuses on how emotions are perceived and communicated between dyadic interactions. In this regard, Ekman (1972) found that emotional expression is a basic human property, independent of race or culture. What is culture specific, however, is the set of rules people employ regarding perception and display of emotions. In this instance, Elfenbein and Ambady (2003) argue that there are strong intercultural differences. Subsequently, Elfenbein et al. (2007) coined the term 'emotional dialect' and used this idea in developing the "Integrated Interpersonal Process Framework for Emotion in Organizations" (Elfenbein, 2007). More recently, Allen et al. (2014) reported finding cross-cultural differences in emotional labor (see below).

As Ashkanasy (2003a) stresses, however, most studies of emotional communication in the context of OB refer to emotional labor. The term was coined by sociologist Arlie Hochschild (1983), who observed that service employees are required as a part of their terms of employment to display appropriate (positive or negative) emotions to customers or clients. For example, service providers in the retail, food, travel,

and entertainment industries are expected to display positive emotional expressions or provide "service with a smile" (Pugh, 2001). Other employees in enforcement industries (e.g., police, debt collection) are expected to display negative emotions.

Hochschild (1983) identified in particular two types of emotional labor. These comprise *surface acting* (outward display of emotion that are at odds with the emotion the actor feels) and *deep acting* (where actors try to experience the emotion). Morris and Feldman (1996) subsequently conceptualized surface and deep acting and their effects in terms of three dimensions: intensity, variety, and duration. In additional to emotional labor, Ashforth and Humphrey (1993) recognized that *genuine* emotions might also be displayed. Diefendorff and Gosserand (2003) subsequently demonstrated empirically that genuine emotion is distinct from emotional labor.

There is also a considerable body of evidence to show that performing emotional labor can have detrimental health effects, especially when associated with 'dissonance' between displayed and felt emotions (Van Dijk and Kirk-Brown, 2006). Moreover, the harmful effects of surface acting appear to be much stronger than for deep acting (Kammeyer-Mueller et al., 2013), even carrying over to home life (Wagner et al., 2014), although Grandey et al. (2012) recently found that the harmful effects of emotional labor are ameliorated in a "climate of authenticity" (p. 1).

On the other side of the service encounter, it appears that deep or genuine emotion is also more effective than surface acting. For example, Groth et al. (2009) found that, compared to surface acting, service provider deep acting resulted in improved customer satisfaction (see also Grandey, 2003).

Level 4: Teams and Leadership

At Level 4 of his model, Ashkanasy (2003a) identified leadership as having a major effect on the moods and feelings of team members. A similar conclusion was reached by Humphrey (2002), who argued the mood management may well be the most critical element of team leadership. This idea is also supported in empirical findings reported by Seong and Choi (2014). If this is so, then it would also follow that a leader's emotional intelligence may be the key. In this regard, George (2000) reasoned that emotionally intelligent leaders engender enthusiasm among team members. This idea was supported in empirical research by Pescosolido (2002), who found that effective leaders tended to perform better than their less effective colleagues when it came to dealing with workplace events that involved strong emotions, and that this contributes to development of stronger team harmony and cohesion.

Although controversial (Antonakis et al., 2009), scholarly evidence is tending increasingly to support the notion that emotional intelligence is related to leadership effectiveness. Thus, in a recent review of 10 years of research into the relationship of emotional intelligence and leader effectiveness, Walter et al. (2011) found overwhelming support across all three 'streams' of emotional intelligence identified in Ashkanasy and Daus (2005).

In groups and teams, the principal mechanism for spreading a particular emotional state via emotional contagion (Hatfield et al., 1992), whereby members of a group are 'infected' by others' emotional states through mimicking other members' facial expressions, body language, and vocal tone. Given the modeling role of leaders, it follows therefore that contagion should be a major mechanism for leaders to transfer emotional states to team members. In this regard, Goleman et al. (2002) argue that the most effective leaders (which they refer to as 'primal' leaders) engender 'emotional synchronization' or resonance with their followers. In essence, followers 'tune in' to their leader and adopt whatever emotional state the leader is modeling. Empirical support for the role of emotional contagion in groups comes from a study conducted among MBA study groups by Barsade (2002), while Sy et al. (2005) found evidence for leader-to-member contagion. More recently, Tee et al. (2013) found that the effect can go the other way: from followers to the leader.

A corollary of this is that, to be effective, leaders need to manage emotional contagion lest it spirals out of control. In this regard, Humphrey (2008: p. 1) argues that leaders need to "lead with emotional labor" (see also Ashkanasy and Humphrey, 2011b) as a means to regulate their own emotional feelings. Thus effective leaders engage in genuine emotional expression or (at least) deep acting to model the emotions suited to a particular situation; this emotion is then likely to be picked up by group members (via contagion) leading to the whole team adopting the (appropriate) emotion (Sy et al., 2005).

In addition to the direct role of emotion in leader group/team situations, it seems that emotions also play a key role in broader models of leadership. Indeed, there is evidence that a key skill of transformational leaders lies in an ability to help followers to deal with negative emotional events. For example, in a study of R&D teams, Pirola-Merlo et al. (2002) found that leaders could boost performance by helping their followers deal with everyday frustrations and negative moods. McColl-Kennedy and Anderson (2002) found similarly that transformational leaders engendered positive emotion leading to optimism, improved performance, and goal attainment (see also Tsai et al., 2009).

Finally, it is worth noting that there may also be a potential 'dark side' to emotional intelligence in the context of leadership. For example, Kilduff et al. (2010) recently suggested that emotional intelligence can be used 'strategically' to manipulate others.

Level 5: The Organizational Level

At Level 5 of his five-level model, Ashkanasy (2003a) considers the role of emotions at the organizational level, and argues the need for organizational managers to work toward a "healthy emotional climate" (Ashkanasy and Daus, 2002). In this regard, Härtel and Ashkanasy (2010) posit that a healthy climate is characterized by positive emotions, created and then sustained across the whole organization.

This brings us to the notion of an organizational emotional climate, defined by de Rivera (1992; see also Yurtsever and de Rivera, 2010) as "an objective group phenomenon that can be

palpably sensed – as when one enters a party or a city and feels an attitude of gaiety or depression, openness or fear" (p. 197). In this sense, emotional climate represents a particular form of organizational climate (Schneider et al., 2010), which deals specifically with the collective mood of organizational members, and their attitudes toward their peers and leaders, as well as the organization as a whole. Note that climate there, although related to organizational culture, is differentiated from culture in that it is not a function of organizational members' beliefs, values, and assumptions (Schein, 2004).

There is, however, some debate as to whether emotional climate is a team (i.e., Level 4) phenomenon, rather than something that exists organization wide (Level 5). For example, Ashkanasy and Nicholson (2003) examined a 'climate of fear' into Australian restaurant chains and found that it varied by subunit within the two organizations. This was in contrast to measures of organizational culture, which they found to vary across (but not within) organizations.

In fact, there is evidence that "Level 5 is qualitatively different from the other levels of the model" (Ashkanasy and Ashton-James, 2005: p. 221) insofar as this level subsumes the five lower levels. In this instance, Level 5 emotions are generated as a result of the accumulation of affective events at Level 1, employees' emotional intelligence abilities at Level 2, which, in turn, impacts their interactions with coworkers, subordinates, supervisors, and clients at Level 3, also affecting group emotion and leadership at Level 4. Consistent with this idea, Dasborough et al. (2009) argue that "a leader's behavior to subordinates (Level 4) is reflected in team-member relationships (Level 3) that in turn reflect on the leader's performance via emotional contagion processes, leading to an organizational management response to the leader (Level 5)" (Ashkanasy and Humphrey, 2011a: p. 219). Finally, it should also be noted that organizational culture is often reflected in organizational policies. From the perspective of emotions, this refers to an organization's policies regarding emotional displays required by employees (i.e., emotional labor, see Beyer and Niño, 2001; Diefendorff and Greguras, 2009; Gosserand and Diefendorff, 2005; Fineman, 2000; Hochschild, 1983; van Maanen and Kunda, 1989; Rafaeli and Sutton, 1987, 1989).

Discussion and Conclusions

Although the role of emotion in organizations was neglected for many years, this state of affairs has changed over the past 20 years, to the point where the study of emotions has become a priority in IO research. Major published reviews of the field by leading authors in the field (e.g., Ashforth and Humphrey, 1995; Ashkanasy and Daus, 2002; Brief and Weiss, 2002; Elfenbein, 2007) suggest that this is an area with "a bright future" (Ashkanasy and Ashton-James, 2005: p. 221).

In this brief overview of the field, I employed Ashkanasy's (2003a) structure that includes five levels: (1) within-person temporal effects, (2) between-person factors, (3) interpersonal behaviors, (4) group and team leadership, and (5) the organization-wide view. I also stressed the interconnectedness of the levels, suggesting in Ashkanasy (2003b) that the 'glue' facilitating this lies in the biological basis of emotion.

Looking forward to the future of research in this field, four issues are noteworthy (Gooty et al., 2009). First, there is a need for scholars to be more consistent in their definitions of basic terms, especially when it comes to differentiating affect, emotion, and mood. Second, scholars have generally restricted their studies to positive and negative affect, rather than addressing the effects of discrete emotions. For example, anger, fear, and sadness are all examples of negative affect, but are associated with different action tendencies (Frijda, 1986). Third, while studies of within-person effects using diary or ESM methods are becoming more prevalent, there continues to be a pressing need for more Level 1 research. Fourth, research on emotions needs to pay more attention to the role played by the context in studies of emotion in organizations. In addition to the foregoing points, Ashkanasy (2009) also call for research in OB, in general, to make more use of multilevel analysis and to take into consideration the neurobiological nature of behavior and emotions.

See also: Authentic Leadership; Behavior Analysis, Applied; Job Satisfaction; Leader–Member Exchange Theory; Managerial and Organizational Cognition; Organizational Citizenship Behavior; Organizational Climate in the Work Setting; Organizational Commitment; Personality and Values at Work.

Bibliography

Allen, J.A., Diefendorff, J.M., Ma, Y.F., 2014. Differences in emotional labor across cultures: a comparison of Chinese and US service workers. Journal of Business and Psychology 29, 21–35.

Amabile, T.M., Barsade, S.G., Mueller, J.S., Staw, B.M., 2005. Affect and creativity at work. Administrative Science Quarterly 50, 367–403.

Antonakis, J., Ashkanasy, N.M., Dasborough, M.T., 2009. Does leadership need emotional intelligence? The Leadership Quarterly 20, 247–261.

Ashforth, B.E., Humphrey, R.H., 1993. Emotional labor in service roles: the influence of identity. Academy of Management Review 18, 88–115.

Ashforth, B.E., Humphrey, R.H., 1995. Emotion in the workplace: a reappraisal. Human Relations 48, 97–125.

Ashkanasy, N.M., 2003a. Emotions in organizations: a multilevel perspective. In: Dansereau, F., Yammarino, F.J. (Eds.), Research in Multi-level Issues, vol. 2. Elsevier Science, Oxford, UK, pp. 9–54.

Ashkanasy, N.M., 2003b. Emotions at multiple levels: an integration. In: Dansereau, F., Yammarino, F.J. (Eds.), Research in Multi-level Issues, Volume 2: Multi-level Issues in Organizational Behavior and Strategy. Elsevier Science, Oxford, UK, pp. 71–81.

Ashkanasy, N.M., 2009. After thirty years: what does the future hold? Journal of Organizational Behavior 30, 15–20.

Ashkanasy, N.M., Ashton-James, C.E., 2005. Emotion in organizations: a neglected topic in I/O psychology, but with a bright future. In: Hodgkinson, G.P., Ford, J.K. (Eds.), International Review of Industrial and Organizational Psychology, vol. 20. Wiley, Chichester, UK, pp. 221–268.

Ashkanasy, N.M., Daus, C.S., 2002. Emotion in the workplace: the new challenge for managers. Academy of Management Executive 16 (1), 76–86.

Ashkanasy, N.M., Daus, C.S., 2005. Rumors of the death of emotional intelligence in organizational behavior are vastly exaggerated. Journal of Organizational Behavior 26, 441–452.

Ashkanasy, N.M., Härtel, C.E.J., 2014. Emotional climate and culture: the good, the bad, and the ugly. In: Schneider, B., Barbera, K. (Eds.), The Oxford Handbook of Organizational Culture and Climate. Oxford University Press, New York, pp. 136–152.

Ashkanasy, N.M., Humphrey, R.H., 2011a. Current research on emotion in organizations. Emotion Review 3, 214–224.

Ashkanasy, N.M., Humphrey, R.H., 2011b. A multi-level view of leadership and emotions: leading with emotional labor. In: Bryman, A., Collinson, D., Grint, K.,

Jackson, B., Uhl-Bien, M. (Eds.), Sage Handbook of Leadership. Sage Publications, London, pp. 363–377.
Ashkanasy, N.M., Nicholson, G.J., 2003. Climate of fear in organizational settings: construct definition, measurement, and a test of theory. Australian Journal of Psychology 55, 24–29.
Ashton-James, C.E., Ashkanasy, N.M., 2005. What lies beneath? A deconstructive analysis of affective events theory. In: Ashkanasy, N.M., Zerbe, W.J., Härtel, C.E.J. (Eds.), Research on Emotion in Organizations, vol. 1. Elsevier/JAI Press, Oxford, UK, pp. 23–50.
Ashton-James, C.E., Ashkanasy, N.M., 2008. Affective events theory: a strategic perspective. In: Zerbe, W.J., Härtel, C.E.J., Ashkanasy, N.M. (Eds.), Research on emotion in organizations, volume 4: emotions, ethics, and decision-making. Emerald Group Publishing, Bingley, UK, pp. 1–34.
Bar-On, R., 1997. Bar-On Emotional Quotient Inventory (EQ-i): Technical Manual. Multi-Health Systems, Toronto, Canada.
Barsade, S.G., 2002. The ripple effect: emotional contagion and its influence on group behavior. Administrative Science Quarterly 47, 644–675.
Barsade, S.G., Brief, A.P., Spataro, S.E., 2003. The affective revolution in organizational behavior: the emergence of a paradigm. In: Greenberg, J. (Ed.), Organizational Behavior: The State of the Science. Lawrence Erlbaum and Associates, Mahwah, NJ, pp. 3–52.
Barsky, A., Kaplan, S.A., 2007. If you feel bad, it's unfair: a quantitative synthesis of affect and organizational justice perceptions. Journal of Applied Psychology 92, 286–295.
Beyer, J., Niño, D., 2001. Culture as a source, expression, and reinforcer of emotions in organizations. In: Payne, R.L., Cooper, C.L. (Eds.), Emotions at Work: Theory, Research, and Applications for Management. Wiley, Chichester, UK, pp. 173–197.
Brackett, M.A., Mayer, J.D., 2003. Convergent, discriminant, and incremental validity of competing measures of emotional intelligence. Personality and Social Psychology Bulletin 29, 1147–1158.
Brief, A.P., Burke, M.J., George, J.M., Robinson, B.S., Webster, J., 1988. Should negative affectivity remain an unmeasured variable in the study of job stress? Journal of Applied Psychology 73, 193–198.
Brief, A.P., Weiss, H.M., 2002. Organizational behavior: affect in the workplace. Annual Review of Psychology 53, 279–307.
Côté, S., Lopez, P.N., Salovey, P., Miners, C.T.H., 2010. Emotional intelligence and leadership emergence in small groups. The Leadership Quarterly 21, 496–508.
Dasborough, M.T., Ashkanasy, N.M., Tee, E.E.J., Tse, H.H.M., 2009. What goes around comes around: how meso-level negative emotional contagion can ultimately determine organizational attitudes toward leaders. The Leadership Quarterly 20, 571–585.
Daus, C.S., Ashkanasy, N.M., 2005. The case for an ability-based model of emotional intelligence in organizational behavior. Journal of Organizational Behavior 26, 453–466.
Diefendorff, J.M., Greguras, G.J., 2009. Contextualizing emotional display rules: examining the roles of targets and discrete emotions in shaping display rule perceptions. Journal of Management 35, 880–898.
Diefendorff, J.M., Gosserand, R.H., 2003. Understanding the emotional labor process: a control theory perspective. Journal of Organizational Behavior 24, 945–959.
Ekman, P., 1972. Universals and cultural differences in facial expressions of emotion. In: Cole, J. (Ed.), Nebraska Symposium on Motivation, 1971, vol. 19. University of Nebraska Press, Lincoln, pp. 207–282.
Elfenbein, H.A., 2007. Emotion in organizations: a review and theoretical integration. The Academy of Management Annals 1, 315–386.
Elfenbein, H.A., Ambady, N., 2003. Universals and cultural differences in recognizing emotions. Current Directions in Psychological Science 12, 159–164.
Elfenbein, H.A., Beaupré, M.G., Lévesque, M., Hess, U., 2007. Toward a dialect theory: cultural differences in the expression and recognition of posed facial expressions. Emotion 7, 131–146.
Fineman, S. (Ed.), 2000. Emotion in Organizations, second ed. Sage, London.
Fisher, C.D., 2002. Antecedents and consequences of real-time affective reactions at work. Motivation & Emotion 26, 3–30.
Fisher, C.D., 2008. What if we took within-person performance variability seriously? Industrial and Organizational Psychology: Perspectives on Science and Practice 2, 185–189.
Fisher, C.D., Minbashian, A., Beckmann, N., Wood, R.E., 2013. Task appraisals, emotions, and performance goal orientation. Journal of Applied Psychology 98, 364–373.
Fisher, C.D., Noble, C.S., 2004. A within-person examination of correlates of performance and emotions while working. Human Performance 17, 145–168.
Forgas, J.P., 1995. Mood and judgment: the affect infusion model (AIM). Psychological Bulletin 117, 39–66.
Forgas, J.P., East, R., 2003. Affective influences on social judgments and decisions: implicit and explicit processes. In: Forgas, J.P., Williams, K., Von Hipple, W. (Eds.), Social Judgments: Implicit and Explicit Processes. Cambridge University Press, UK, pp. 198–226.
Forgas, J.P., George, J.M., 2001. Affective influences on judgments and behavior in organizations: an information processing perspective. Organizational Behavior and Human Decision Processes 86, 3–34.
Frijda, N.H., 1986. The Emotions. Cambridge University Press, New York.
Fredrickson, B.L., 2001. The role of positive emotions in positive psychology: the broaden-and-build theory of positive emotions. American Psychologist 56, 218–226.
Fuller, J.A., Stanton, J.M., Fisher, G.G., Spitzmüller, C., Russell, S.S., Smith, P.C., 2003. A lengthy look at the daily grind: time series analysis of events, mood, stress, and satisfaction. Journal of Applied Psychology 88, 1019–1033.
George, J.M., 2000. Emotions and leadership: the role of emotional intelligence. Human Relations 53, 1027–1055.
Goleman, D., 1995. Emotional Intelligence: Why It Can Matter More than IQ. Bantam Books, New York.
Goleman, D., Boyatzis, R., McKee, A., 2002. Primal Leadership: Realizing the Power of Emotional Intelligence. Harvard Business School Press, Boston.
Gooty, J., Gavin, M., Ashkanasy, N.M., 2009. Emotions research in OB: the challenges that lie ahead. Journal of Organizational Behavior 30, 833–838.
Gosserand, R.H., Diefendorff, J.M., 2005. Emotional display rules and emotional labor: the moderating role of commitment. Journal of Applied Psychology 90, 1256–1264.
Grandey, A., 2003. When "the show must go on": surface and deep acting as predictors of emotional exhaustion and service delivery. Academy of Management Journal 46, 86–96.
Grandey, A., Foo, S.C., Groth, M., Goodwin, R.E., 2012. Free to be you and me: a climate of authenticity alleviates burnout from emotional labor. Journal of Occupational Health Psychology 17, 1–14.
Groth, M., Hennig-Thurau, T., Walsh, G., 2009. Customer reactions to emotional labor: the roles of employee acting strategies and customer detection accuracy. Academy of Management Journal 52, 958–974.
Härtel, C.E.J., Ashkanasy, N.M., 2010. Healthy human cultures as positive work environments. In: Ashkanasy, N.M., Wilderom, C.E.P., Peterson, M.F. (Eds.), The Handbook of Culture and Climate, second ed. Sage, Thousand Oaks, CA, pp. 85–100.
Hatfield, E., Cacioppo, J., Rapson, R.L., 1992. Emotional contagion. In: Clark, M.S. (Ed.), Review of Personality and Social Psychology: Emotion and Social Behavior. Sage Publications, Newbury Park, CA, pp. 151–177.
Hochschild, A.R., 1983. The Managed Heart: Commercialization of Human Feeling. University of California Press, Berkeley, CA.
Humphrey, R.H., 2002. The many faces of emotional leadership. The Leadership Quarterly 13, 493–504.
Humphrey, R.H., 2008. The right way to lead with emotional labor. In: Humphrey, R.H. (Ed.), Affect and Emotion: New Directions in Management Theory and Research. Information Age Publishing, Charlotte, NC, pp. 1–17.
Isen, A.M., 1987. Positive affect, cognitive processes and social behavior. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 20. Academic Press, San Diego, CA, pp. 203–253.
Jordan, P.J., Ashkanasy, N.M., Daus, C.S., 2008. Emotional intelligence: rhetoric or reality? In: Cartwright, S., Cooper, C.L. (Eds.), The Oxford Handbook of Personnel Psychology. Oxford University Press, Oxford, UK, pp. 37–58.
Joseph, D.L., Newman, D.A., 2010. Emotional intelligence: an integrative meta-analysis and cascading model. Journal of Applied Psychology 95, 54–78.
Judge, T.A., Thoresen, C.J., Bono, J.E., Patton, G.K., 2001. The job satisfaction-job performance relationship: a qualitative and quantitative review. Psychological Bulletin 127, 376–407.
Kammeyer-Mueller, J.D., Rubenstein, A.L., Long, D.M., Odio, M.A., Buckman, B.R., Zhang, Y.W., Halvorsen-Ganepola, M.D.K., 2013. A meta-analytic structural model of dispositional affectivity and emotional labor. Personnel Psychology 66, 47–90.
Kilduff, M., Chiaburu, D.S., Menges, J.I., 2010. Strategic use of emotional intelligence in organizational settings: exploring the dark side. Research in Organizational Behavior 30, 129–152.
Larson, R., Csikszentmihalyi, M., 1983. The experience sampling method. In: Reis, H.T. (Ed.), Naturalistic Approaches to Studying Social Interaction, New Directions for Methodology of Social and Behavioral Science, vol. 15. Jossey-Bass, San Francisco, pp. 41–56.
Locke, E.A., 2005. Why emotional intelligence is an invalid concept. Journal of Organizational Behavior 26, 425–431.
van Maanen, J., Kunda, G., 1989. "Real feelings": emotional expression and organizational culture. Research in Organizational Behavior 11, 43–103.
Mayer, J., Salovey, P., 1997. What is emotional intelligence? In: Salovey, P., Sluyter, D. (Eds.), Emotional Development and Emotional Intelligence: Implications for Educators. Basic Books, New York, pp. 3–31.
Mayer, J.D., Salovey, P., Caruso, D.R., 2002. Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT) Users Manual. MHS Publishers, Toronto, Canada.

McColl-Kennedy, J.R., Anderson, R.D., 2002. Impact of leadership style and emotions on subordinate performance. The Leadership Quarterly 13, 545–559.

Mittal, V., Ross, W.T., 1998. The impact of positive and negative affect and issue framing on issue interpretation and risk taking. Organizational Behavior and Human Decision Processes 76, 298–324.

Morris, J.A., Feldman, D.C., 1996. The dimensions, antecedents, and consequences of emotional labor. Academy of Management Review 21, 986–1010.

O'Boyle Jr., E.H., Humphrey, E.H., Pollack, J.M., Hawver, T.H., Story, P.A., 2011. The relation between emotional intelligence and job performance: a meta-analysis. Journal of Organizational Behavior 32, 788–818.

Pekrun, R., Frese, M., 1992. Emotion in work and achievement. International Review of Industrial and Organizational Psychology 7, 153–196.

Pescosolido, A.T., 2002. Emergent leaders as managers of group emotion. Leadership Quarterly 13, 583–599.

Pirola-Merlo, A., Härtel, C.E.J., Mann, L., Hirst, G., 2002. How leaders influence the impact of affective events on team climate and performance in R&D teams. The Leadership Quarterly 13, 561–581.

Pugh, S.D., 2001. Service with a smile: emotional contagion in the service encounter. Academy of Management Review 44, 1018–1027.

Rafaeli, A., Sutton, R.I., 1987. Expression of emotion as part of the work role. Academy of Management Review 12, 23–37.

Rafaeli, A., Sutton, R.I., 1989. The expression of emotion in organizational life. Research in Organizational Behavior 11, 1–42.

de Rivera, J., 1992. Emotional climate: social structure and emotional dynamics. International Review of Studies of Emotion 2, 197–218.

Schein, E.H., 2004. Organizational Culture and Leadership, third ed. Jossey-Bass, San Francisco.

Schlaerth, A., Ensari, N., Christian, J., 2013. A meta-analytical review of the relationship between emotional intelligence and leaders' constructive conflict management. Group Processes & Intergroup Relations 16, 126–136.

Schneider, B., Ehrhart, M.G., Macey, W.H., 2010. Organizational climate research: achievements and the road ahead. In: Ashkanasy, N.M., Wilderom, C.E.P., Peterson, M.F. (Eds.), The Handbook of Culture and Climate, second ed. Sage, Thousand Oaks, CA, pp. 29–49.

Seong, J.Y., Choi, J.N., 2014. Effects of group-level fit on group conflict and performance. The initiating role of leader positive affect. Group & Organization Management 39, 190–212.

Staw, B.M., Barsade, S.G., 1993. Affect and managerial performance: a test of the sadder-but-wiser vs happier-and-smarter hypotheses. Administrative Science Quarterly 38, 304–328.

Sy, T., Côté, S., Saavedra, R., 2005. The contagious leader: the impact of the leader's mood on the mood of group members, group affective tone, and group processes. Journal of Applied Psychology 90, 295–305.

Tee, E.Y.J., Ashkanasy, N.M., Paulsen, N., 2013. The influence of follower mood on leader mood and task performance: evidence for an affective, follower-centric perspective of leadership. The Leadership Quarterly 24, 496–515.

To, M.L., Fisher, C.D., Ashkanasy, N.M., Rowe, P.A., 2012. Within-person relationships between mood and creativity. Journal of Applied Psychology 97, 519–612.

Tsai, W., Chen, H., Cheng, J., 2009. Employee positive moods as a mediator linking transformational leadership and employee work outcomes. The International Journal of Human Resource Management 20, 206–219.

Van Dijk, P.A., Kirk-Brown, A., 2006. Emotional labor and negative job outcomes: an evaluation of the mediating role of emotional dissonance. Journal of Management & Organization 12, 101–115.

Wagner, D.T., Barnes, C.M., Scott, B.A., 2014. Driving it home: how workplace emotional labor harms employee home life. Personnel Psychology 67, 487–516.

Walter, F., Cole, M.S., Humphrey, R.H., 2011. Leadership and emotional intelligence: new directions for a controversial topic. The Academy of Management Perspectives 25 (1), 45–59.

Weiss, H.M., Beal, D.J., 2005. Reflections on affective events theory. In: Ashkanasy, N.M., Zerbe, W.J., Härtel, C.E.J. (Eds.), Research on Emotion in Organizations, vol. 1. Elsevier/JAI Press, Oxford, UK, pp. 1–22.

Weiss, H.M., Nicholas, J.P., Daus, C.S., 1999. An examination of the joint effects of affective experiences and job beliefs on job satisfaction and variations in affective experiences over time. Organizational Behavior and Human Decision Processes 78, 1–24.

Yurtsever, G., de Rivera, J., 2010. Measuring the emotional climate of an organization. Perceptual and Motor Skills 110, 501–516.

Emotions, History of

Peter N Stearns, George Mason University, Fairfax, VA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 7, pp. 4466–4472, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

Historians have long considered the role of emotions in the past but the explicit study of past emotions and emotional standards has only been studied since 1980. The main emphasis in this field rests on determining points of change in emotional values, and explaining why they occur and what impact they have on individual behaviors and social institutions. Emotions history interacts strongly with anthropological and sociological study. While major areas of inquiry remain open, important findings have focused on key shifts in emotional standards during the eighteenth century and also during the second quarter of the twentieth century. Exploring the consequences of these changes, not only in personal but also in political life, and embellishing analysis through such additional techniques as comparative study, offer important channels for future research.

While historians have long considered the role of emotions in the past, for example in biographies, the explicit study of past emotions and emotional standards has only been studied since 1980. The main emphasis in this field rests on determining points of change in emotional values, and explaining why they occur and what impact they have on individual behaviors and social institutions. Emotions history interacts strongly with anthropological and sociological study. While major areas of inquiry remain open (particularly in the history of societies other than Western Europe and the USA), important findings have focused on key shifts in emotional standards during the eighteenth century (for example, the rise of new definitions and valuations of love) and also during the second quarter of the twentieth century (in campaigns for a reduction of intensity in such emotions as grief, anger, and even love). Exploring the consequences of these changes, not only in personal but also in political life, and embellishing analysis through such additional techniques as comparative study, offer important channels for future research.

Using the Past for New Understanding

This field of study builds on two assumptions. First, historians of emotion believe that emotion, the response to stimuli that combines physiological reactions and cognitive evaluations, plays a significant role in historical action. In this they differ from some historians who have argued, for example in interpreting rioting, that purely rational calculations underlie behaviors even when there are admitted emotional overtones such as anger or fear. Second, historians of emotion believe, and demonstrate, that emotions are subject to change, according to shifts in historical context. In this they differ from some researchers who argue that emotions are human constants, based on inherent biological and psychological factors, varying only by personality. Historians of emotion join anthropologists and many sociologists in seeing emotional standards and reactions as group variables, at least in part. Whereas anthropologists mainly study different emotional configurations across contemporary cultures, historians of emotion deal with changes over time, venturing a periodization of key developments. Historians of emotion also contribute to emotions-study, a concern for the social results of emotional standards, as opposed to purely personal qualities.

Major Themes

Historical research has become a significant contributor to the exploration of emotion and the role of emotion in larger social interactions. Although there are a number of different approaches to emotions history, and several different rationales, most historians agree in the focus on 'changes' in emotional standards and experience. They seek to chart what the changes were, granting that explicit standards, or what is often called emotional culture or emotionology, are easier to delineate than actual individual or collective experience. However, they demonstrate that standards themselves have important results both in emotional life and in social and institutional behavior. Here, indeed, is one of the chief methodological pillars of the field. Historians also try to explain what the 'causes' of change were, which helps connect emotions to larger social developments. They are also eager to show what the 'results' were, not just in individual perceptions of emotions and capacities for display, but in other institutions and interactions.

How the Field Became Explicit

History cannot be written without dealing with emotional issues and expressions. Even before the rise of psychohistory – the application of psychological theories, particularly Freudian psychoanalytic concepts, to historical data – biographers often referred to emotional qualities of their subjects. Classic attempts to convey the larger spirit of a historical period, like Johan Huizinga's characterization of the waning Middle Ages, often included emotional dynamics, as did descriptions of reactions to wars or crowd behavior. With all this, however, the principal focus was not on emotion per se, but on emotion as a component of some other phenomenon.

The history of emotions received a considerable boost in the 1930s. The rise of social history in France, seeking to understand the experience of ordinary people in the past, involved exploring topics not previously part of the historical repertoire. Innovators like Lucien Febvre (1973) specifically called for research on emotions as part of the larger social history effort. At about the same time, German sociologist Norbert Elias sketched his theory of a 'civilizing process,' (Elias, 1978, 1982) arguing that standards of self-control became more rigorous in Europe's upper classes in the seventeenth and eighteenth centuries. His primary focus involved physical behavior, such as eating habits, but his attention to more refined manners also involved discipline over emotion. Neither Febvre nor Elias generated an immediate response; indeed, Elias' theories are used currently more widely than ever before.

The development of psychohistory in the 1970s certainly called new attention to the historical exploration of emotion. Historians of childhood like Philip Greven (1977) and David Hunt (1970) thus explored the emotional impact of characteristic styles of childrearing in the seventeenth and eighteenth centuries, using psychoanalytic theory; Greven focused on the consequences of new styles of discipline associated with religious movements in colonial America. Most psychohistorians, however, concentrated primarily on biography, as in the classic study of Luther by Erik Erikson (1962). Still, historians influenced by Freudianism have continued to contribute, particularly when they have turned to larger cultural standards and individual experience around these standards. Generational emotional results of the experience of World War I have thus figured in some explanations of the rise of German Nazism (including the impact of father absence).

However, most explicit attention to emotions history resulted from a slightly different set of developments, in social history and related social sciences, as a framework established in the 1970s translated into explicit research from the mid-1980s onward. Social historians, now dominant in many areas of historical research, continued to widen their exploration of how ordinary people functioned in the past. A major extension of these inquiries, launched in France but spreading more widely, argued for inquiry into deeply held values and assumptions – *mentalités* – concerning subjects such as death, community, or the forces of nature. Not surprisingly, attitudes about emotion quickly figured into this mix; sweeping studies of fear were one key result. English and American historians, for their part, broadened the investigation of family patterns into qualitative relationships among family members. In this vein they explored shifts in the definition and valuation of love, both marital and parental, and also changes in the ways anger was viewed in family life. Mentalities and family historians alike painted a distinctive picture of emotional standards in early modern – sixteenth to seventeenth century – European and American societies, compared to more modern norms. They also paid explicit attention to processes of change, particularly by the eighteenth century.

Two related strands of social history inquiry thus produced a new understanding of the extent to which past culture, and institutions like the peasant village or the family, could not be grasped without assessing the emotional dynamics involved. The same inquiry underlined the extent to which emotional dynamics varied over time – with the historical period – and were therefore subject to explainable change.

Concurrently, various social scientists exploring emotion, began to probe the historical record as well, in this case with primary, though not exclusive, attention to more recent developments. American sociologists and social psychologists began to look at new emotional standards at work, as service sector jobs expanded in the twentieth century. They also investigated changing definitions of family related emotion. While most anthropologists concentrated on painting rich, but fairly static pictures of emotional standards in various societies, some began to build in attention to change. Thus Robert Levy, in his classic study of the Tahitians (1973), dealt with the impact of Western colonialism on Tahitian emotional culture, while Catherine Lutz (1988) in a variety of studies looked at shifts in emotion over time, while relying heavily on contemporary field work descriptions for her basic categories. A number of European sociologists, some strongly influenced by Elias' theories, began to examine changes in 'emotion management,' particularly in the twentieth century but in some cases over a longer sweep of time, focusing on ways emotional impulses were to be controlled. A Dutch group, in particular, worked on the emotional standards that accompanied more informal manners and relationships in the twentieth century. By the 1980s, explicit work in the history of emotion was well underway (see Wouters, 1992).

Gains and Gaps in Historical Research on Emotion

Once launched, emotions history inevitably gained additional vantage points and emphases. The list of emotions historically explored now includes envy, disgust, joy (tentatively), grief, shame, and guilt, as well as the initial roster.

The takeoff of emotions history has also widened in terms of place and time – crucial indications of a project's significance in this empirically specific discipline. Thus several studies deal with changes and continuities in emotional standards and expressions in Chinese history. Historians addressing medieval Europe have generated an impressive literature, as in an important set of studies on anger and projects on wonder and empathy (Rosenwein, 1998). Classicists are also dealing with love, in the context of further explorations of family history. The growing range still begs for expansion and connection. Changes within the Middle Ages need more explicit linkage to the patterns noted in the early modern centuries, while the Renaissance, as a possible period in emotions history, has not clearly surfaced at all. Geographical gaps are at least as obvious as the challenges in chronology and linkage. Specialists dealing with a number of key areas – Eastern Europe, Africa, Latin America, additional parts of Asia – have yet to generate focused histories of emotion, as opposed to more haphazard overlaps and connections with other topics like family life. Relatedly, key theories about emotional change – such as the civilizing process idea – have yet to be seriously tested beyond the Western context, so despite some speculation there is no consensus about possible applicability. Nevertheless, it is fair to say that the process of moving emotions history toward a full historical canvas is underway.

Two Turning Points

The most elaborate work, however, continues to emphasize developments in the USA and Western Europe. Here, in addition to broadening the varieties of emotions considered and the contexts in which they operate, historians and kindred social scientists have particularly emphasized two points of particularly revealing transformation. An important set of changes redefined emotional culture between the late seventeenth and early nineteenth centuries, with timing somewhat varying with the emotional configuration involved. A subsequent turning point intervened in the second quarter of the twentieth century, possibly a bit earlier in the USA than in Western Europe. Within this periodization framework a host of questions remain disputed or unanswered, but the framework itself has proved widely useful in explaining significant recasting of emotional guidelines and experience, and in beginning to amass evidence about why emotions change and with what larger results.

In the early modern emotionological transformation, two factors propelled change. First, growing commercialization helped place new value on the family, as communities became somewhat more competitive. Family ties were alternatives to older, now frayed communitarian bonds. Same-sex friendships, by the same token, declined. Along with this, Protestantism, attacking the idea of special credit for celibacy, greatly improved the religious status of marriage (and, in destroying convents, made marriage more essential for women). By the seventeenth century, Protestant writers urged emotional sharing in marriage. From these twin bases, actual emphasis on love as a key ingredient in forming and evaluating marriage measurably expanded by the mid-eighteenth century. A concomitant, noteworthy by this point, was a more systematic attack on anger in the family, as contradictory to love and also to the greater equality in human relationships being promoted by Enlightenment ideology. Jealousy was also redefined and more systematically reproached, particularly as the idea of a jealous defense of honor came under attack as uncivilized. Growing emphasis on parent–child love, combined with further erosion of community controls, helped generate the growing reliance on guilt over shame in childrearing (Stearns and Stearns, 1988). Techniques such as isolation increasingly supplanted both shaming and physical punishments, convincing children that love was temporarily interrupted and could be restored only by sincere repentance – often establishing guilt as a durable emotional enforcer of character standards. Grief, though more fully amplified in the nineteenth century, also followed from a greater emphasis on love, along with new beliefs that deaths should be brought under greater control. The apparatus available to commemorate death changed in order to accommodate growing grief needs (Rosenblatt, 1983).

Emotional change also developed outside the family. The acceptability of sadness and melancholy declined, with people (initially, particularly men) urged to present a cheerful demeanor. More broadly, emotions were increasingly seen as operating under individual control, rather than sweeping through the body as an outgrowth of physical humors. Correspondingly, responsibility for personal emotion management increased. Words like temper began to suggest a type of anger that should be controlled, while outright neologisms, such as tantrums, indicated a childish level of rage that could disqualify an adult from serious consideration (Stearns and Stearns, 1986). The rise of new standards of disgust followed from cultural redefinitions of emanations from the body and from growing social inequality, with disgust used to mark and justify boundary lines between respectable and unrespectable people and situations (Corbin, 1986).

Historians are still working out some of the implications of these changes, in altering the context of family life but also spilling over into wider public arenas. The transformation of criminal justice, shifting from shaming to imprisonment, clearly reflected the new commitment to guilt, for example, along with growing disgust at crowd reactions to public punishments. Ramifications of emotional change could reach into a wide spectrum.

Redirections in emotional culture between 1900 and 1950 were also extensive. Here, the dominant motifs are greater control over intensity plus growing informality in social relationships. Some historians also add an increasing democratization of emotional standards, with fewer systematic social divisions between respectable groups and others. Suspicion of intensity showed in several arenas. Anger, previously approved for men as a motivation for business and professional life (though dangerous in the home), was now more systematically reproved, with new techniques designed to defuse it. Grief was increasingly disallowed. Even more revealing attacks on intensity turned against too much stress on guilt, which might damage children's self-confidence. Maternal love, praised without major qualification in the nineteenth century, now fell victim to extensive criticism because of its overpowering qualities; mothers should keep their affection under tighter bounds. Romantic love, which had evolved toward increasingly idealistic statements in the nineteenth century, seemed excessive as well by contemporary standards; many prescriptive writers urged more limited, realistic emotional connections, while also freeing sexuality from some of the emotional commitment promoted in nineteenth-century advice literature (Seidman, 1991; Stearns, 1994).

The new concern about intensity was no accident, and it had some of the same range of effects noted in the earlier Western transformation. Causes centered on changes in economic structure, and therefore emotional functions, along with the rise of a new, psychologically informed expertise. Growing consumerism urged emotional investment in acquisition and goods, and this diverted from some of the emotional investments that figured powerfully in the eighteenth and nineteenth centuries. Corporate and service employment required higher levels of emotional self-manipulation, in order to turn on a friendly smile and avoid bureaucratic confrontations. 'Impersonal, yet friendly' became a literal watchword in business training programs, with considerable emotional implications. At the same time, popularizing experts now referred to their psychological, rather than moral credentials, creating a picture of a more fragile mental structure than had previously prevailed. A recent study (Pfister and Schnog, 1997) further probes the particular American fascination with psychological explanations and remedies, as a framework for contemporary emotional life.

The consequences of change involved, as with earlier redirections, both personal and social features. Americans became more nervous about emotional display, convinced that intensity reflected immaturity on their part; hence a recent study emphasizes a proclivity to check with others about onsets of jealousy, in order to provide a peer mirror to guard against lack of restraint. Other studies (Salovey, 1994) show the consequences of rigorous emotional control on the job, in distorting responses even in private relationships. Legal changes, including the rise of no-fault divorce, incorporated the growing belief that moderns should be able to keep anger and grief in check, in the most trying circumstances. Even changes in protest rates – the marked decline of labor strikes after 1958 – derived in part from successful efforts to persuade workers that anger was childish, a reflection of personal flaw rather than a valid emotional response to objective problems. The reach of this new change in basic emotional style, in sum, was considerable.

Finally, amid growing constraints on certain emotional expressions, a recreational culture arose that provided some symbolic contrasts. Intense spectatorship, for example at sporting events, played off against normal emotion management, and athletes and actors were allowed displays of anger, grief, or vivid camaraderie that were not usually countenanced – providing, in other words, what has been described as a compensatory surrogate culture.

Exploration of the two great transformations has involved the work of a substantial number of historians and historical sociologists. Inevitably, even in a new field, some disagreements have surfaced. Specific emphases and attributions of cause differ in the approaches to the early modern transformation. French work divides between explorations of distinctively Catholic features (as in the work on fear) and a sense that French emotional culture moved in harmony with other parts of Europe (as in the work on familial affections; see Delumeau, 1978). Divisions, though not outright clashes, also inform the work on the twentieth-century turning point. European sociologists have worked more on the process of 'informalization' than have their American counterparts, arguing that contemporary emotional relations work because people can assume successful incorporation of the rules of restraint (Wouters, 1992). American historians accept this conclusion, in part, but additionally point to important new regulations and tensions. No coherent schools have emerged as yet, but there are fruitful debates.

Both of the great changes in emotional standards raise a host of questions, headed by the problem of social class. The rise of new standards of love in the eighteenth century involved a variety of social groups; proletarians, free from the constraints of property arrangements, may have moved early to emphasize a type of love more fully (Gillis, 1985). But some of the standards particularly attached to the rising middle class. Emphasis on heightened emotional sensitivities between parents and children, for example, while no monopoly of a single class, moved into a bourgeois lexicon; and many observers criticized workers for failure to display the finer parental feelings. Changes in the twentieth century did have a democratic air, in that nineteenth-century attempts to downplay certain groups as impossibly crude noticeably declined. Romantic love was a case in point. American jurists in the late nineteenth century had been willing to allow lack of love as a valid reason for divorce for middle-class people, under the heading of mental cruelty, but they would not do the same for the working class because such emotional refinement was impossible at this level. In the twentieth century, however, the now-modified romantic love ideal was proclaimed for all. Still, issues of how particular emotional standards affected social classes are only now beginning to receive needed attention. One study, for example, shows how new hostility to the use of fear in childrearing, an American middle-class staple by 1900, captured Catholic working-class culture only in the 1950s. In other instances, middle-class prescriptions were not simply more slowly assimilated by other groups, but were positively defied, as in some of the more emotionally exuberant facets of twentieth-century African American culture (Stearns and Lewis, 1988).

Current Issues and Future Directions

The success of historical work to date considerably adds to available information on varieties of emotional formulations and the ways values and emotions interact. It thus amplifies the constructionist approach to research on emotion, arguing that a significant part of emotional codes and responses vary with context – by place, but also by time period. The exploration of change and causation builds a further ingredient into constructionism, often missing from anthropological studies, in showing how one emotions variant may transform into another, responding to wider factors in economic, political, and cultural structure.

The historical agenda also raises problems, even aside from the range of areas and periods, and types of emotions, still open to this kind of analytical inquiry. The relationship between widely prescribed cultural standards and other aspects of emotional experience – aspects that usually leave less explicit past records – is an obvious issue. Standards count in themselves: they affect social rules and practices, based on expected emotional behavior; they color evaluations of others and of self; and they unquestionably shape actual experience. Thus a culture that discourages romantic love – a common pattern in premodern peasant societies – almost certainly has less of it. But there are disjunctures as well, and historians must be sensitive to these. Among other things, emotional experience may vary somewhat less than standards do. Historians who initially assumed that the absence of explicit encouragement (indeed, the considerable explicit hostility) to marital love, in prescriptive materials before the seventeenth century, meant that there was in fact little, were assuming too much, and more recent researchers have improved on the evidence available in generating more nuanced statements of contrast and change.

A related issue, now happily beginning to gain constructive attention, involves subcultures. As noted previously, most emotions history to date has focused on mainstream standards issuing from dominant social groups. Various social classes, religious contingents, even age categories form emotional cultures of their own. Research in American history, particularly, is beginning to reveal patterns of emotionology in various religious and racial groups, and in the working class.

These patterns sometimes link to mainstream trends, if with a certain chronological lag, but they sometimes deliberately run counter – as in the enthusiastic religious sects that arose around 1900 as an antidote to growing mainstream hostility to emotional intensity.

Comparison also deserves more attention. The extensive focus on large transformations invites fine-tuning. Most research on both the early modern and the twentieth-century turning points has been either strictly national (just the USA or just France) or has assumed large transatlantic, Western trends. Yet we know that in the year 2000, in key particulars, regional patterns vary – the French get angry when jealous, the Dutch sad, while Americans scurry to get other people's opinions. We need explicit historical comparisons to determine when, and above all why, distinctions of this sort emerged.

History can also establish middle-level generalizations about emotional change, based on the accumulation of cases. We can begin to use partial models of the factors most likely to cause change, and about the length of time normally required between the introduction of widely shared new standards, and actual changes in social arrangements and behaviors (including childrearing). It took about a century, for example, for warnings against scaring children with bogeymen stories to gain sufficient acceptance that the prescriptive emphasis could be cut back (after the 1920s). Thirty years elapsed between new strictures against anger in the workplace, and incorporation of the same standards in parenting advice; and almost 40 years between the new culture's inception and clear implementation in laws and protest levels. These are straws in the wind, toward more systematic statements about timing.

In what is still a new field of explicit research, despite promising results to date, dealing with empirical and theoretical gaps inevitably forms a challenging agenda for the future. Identifying subcultures has barely begun, for example in dealing with racial differences over emotional goals in the USA. Geographical omissions are troubling in themselves, and they limit opportunities for comparison. Even for the Atlantic world, where the literature is richest, comparative opportunities have barely been taken up. With all this, two additional challenges remain: further integration with history, and further integration with the other disciplines that examine emotion.

The number of historians interested in emotion has expanded steadily. But findings about emotional change are not usually incorporated into mainstream historical study. Family historians come closest to recognizing the importance of this topic, as an element for example, along with birthrates and the roles of family members, in establishing periodization. But the day when historians of protest or politics consider the role of shifts in emotional standards has not yet arrived. The connection is possible: at a time when the emotional reactions of American politicians are carefully scripted in light of the dominant emotionology, there is obvious invitation to explore mutual interactions between culture and political reactions in the twentieth century (Kofler, 1997). The same holds for studies of sports or labor relations. While the record to date is tentative at best, there is reason to assume that fuller integration will occur, as historians become more comfortable with the idea that emotional standards shift and that shifts have important consequences.

The even greater challenge for the future is interdisciplinary. Historians of emotion have found ready audience among anthropologists, sociologists, and some social psychologists (particularly, of course, those theoretically wedded to a discursive, cultural-constructionist approach). But actual interdisciplinary projects, which would deal with change and impact in relation to current emotional manifestations, have yet to emerge. Mainstream psychology focuses primarily on invariant aspects of emotional display such as facial expressions, and ignores issues of cultural variance and change (Harré and Stearns, 1995). We lack the collaboration necessary to use historical analysis to help test the boundary lines between inherent emotional components and the role of cultural standards, yet precisely this collaboration constitutes a vital next step. Projects that would identify recent change and subject the emotional results to standard psychological scrutiny are possible – for example, the tentative signs, since the 1970s, of partial modification of the contemporary hostility to grief. But constructing the interdisciplinary apparatus that can handle such issues is no easy task, given the growing split between cultural analysis and scientistic psychology. The desirable agenda is obvious; its realization, distinctly problematic.

Even as things stand, the history of emotions is contributing not only to an understanding of emotion, but also to a wider assessment of emotion's impact on behaviors and social institutions. Too often, emotions research focuses on individual experience. Even in the short time explicit emotions topics have commanded their attention, historians have advanced the exploration of why particular emotional articulations matter – how they affect the experience of family life and friendship, how they connect to regimens at work, even – in some extensions currently being developed – how they can help shape new political styles. Historians will continue to probe emotional characteristics in the past and particularly cases of significant change; but they will also be eager to fold emotion actively into our larger understandings of how societies function. Momentum continues to build in this active new disciplinary partner in emotions research.

See also: Biography: Historical; Body, History of; Collective Memory, Psychology of; Cultural History; Cultural Psychology; Elias, Norbert (1897–1990); Emotion: History of the Concept; Emotions, Sociology of; Family and Kinship, History of; Love and Intimacy, Psychology of; Psychoanalysis, History of; Psychohistory.

Bibliography

Alberti, F.B., 2010. Matters of the Heart: History, Medicine, and Emotion. Oxford University Press, Oxford.

Corbin, A., 1986. The Foul and the Fragrant: Odor and the French Imagination. Harvard University Press, Cambridge, MA.

Delumeau, J., 1978. La Peur en Occident, XIVe-XVIIe Siècles: Une Cité Assiégée. Fayard, Paris.

Elias, N., 1978. The Civilizing Process: The History of Manners. Blackwell, Oxford, UK.

Elias, N., 1982. The Civilizing Process: State Formation and Civilization. Blackwell, Oxford, UK.

Elvin, M., 1989. Tales of the Shen and Xien: body-personal and heart-mind in China during the last 150 years. Zone 4, 266–349.

Erikson, E.H., 1962. Young Man Luther: A Study in Psychoanalysis and History. Norton, New York.
Febvre, L., 1973. A New Kind of History. Harper and Row, New York (Original work published 1933).
Frevert, U., 2011. Emotions in History: Lost and Found. Central European University Press, Budapest, Hungary.
Gay, P., 1984–1998. The Bourgeois Experience: Victoria to Freud, 2 vols. Oxford University Press, New York.
Gillis, J.R., 1985. For Better, for Worse: British Marriages, 1600 to the Present. Oxford University Press, New York.
Greven, P., 1977. The Protestant Temperament: Patterns of Child Rearing, Religious Experience and the Self in Early America, first ed. Knopf, New York.
Gross, D.M., 2006. The Secret History of Emotion: From Aristotle's Rhetoric to Modern Brain Science. University of Chicago Press, Chicago, IL.
Harré, R., Stearns, P.N. (Eds.), 1995. Discursive Psychology in Practice. Sage, London.
Hunt, D., 1970. Parents and Children in History: The Psychology of Family Life in Early Modern France. Basic, New York.
Kofler, A. (Ed.), 1997. Emotion and culture. Innovation: The European Journal of Social Sciences 10. Special Issue.
Kagan, J., 2007. What Is Emotion?: History, Measures, and Meanings. Yale University Press, New Haven, CT.
Lantz, H.R., 1982. Romantic love in the pre-modern period: a social commentary. Journal of Social History 15, 349–370.
Levy, R.I., 1973. Tahitians: Mind and Experience in the Society Islands. University of Chicago Press, Chicago, IL.
Lutz, C., 1988. Unnatural Emotions: Everday Sentiments on a Micronesian Atoll and Their Challenge to Western Theory. University of Chicago Press, Chicago, IL.
Lystra, K., 1989. Searching the Heart: Women, Men and Romantic Love in Nineteenth-Century America. Oxford University Press, New York.
Laffan, M.F., Weiss, M., 2012. Facing Fear: The History of an Emotion in Global Perspective. Princeton University Press, Princeton.
Pfister, J., Schnog, N. (Eds.), 1997. Inventing the Psychological: Toward a Cultural History of Emotional Life in America. Yale University Press, New Haven.
Paster, G.K., Rowe, K., Floyd-Wilson, M., 2004. Reading the Early Modern Passions: Essays in the Cultural History of Emotion. University of Pennsylvania Press, Philadelphia.
Rieber, R.W., 2012. Encyclopedia of the History of Psychological Theories. Springer, New York.
Reddy, W.M., 2001. The Navigation of Feeling: A Framework for the History of Emotions. Cambridge University Press, Cambridge, UK.
Rosenblatt, P.C., 1983. Bitter, Bitter Tears: Nineteenth Century Diarists and Twentieth-Century Grief Theories. University of Minnesota Press, Minneapolis.
Rosenwein, B.H., 1998. Anger's Past: The Social Uses of an Emotion in the Middle Ages. Cornell University Press, New York.
Salovey, P., 1994. The Intelligent Emotions. Brenzel Publishing Company, Florence, KY.
Seidman, S., 1991. Romantic Longings: Love in America, 1830–1980. Routledge, New York.
Stearns, C., Stearns, P., 1986. Anger: The Struggle for Emotional Control in America's History. University of Chicago Press, Chicago, IL.
Stearns, C., Stearns, P., 1988. Emotion and Social Change: Toward a New Psychohistory. Holmes and Meier, New York.
Stearns, P., 1994. American Cool: Constructing a Twentieth-Century Emotional Style. New York University Press, New York.
Stearns, P., Lewis, J., 1988. Emotional History of the United States. New York University Press, New York.
Walton, S., 2004. A Natural History of Human Emotions. Grove Press, New York.
Wouters, C., 1992. On status competition and emotion management. Journal of Social History 29, 699–718.

Emotions: Methods of Assessment

Marcello Mortillaro, Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland
Marc Mehu, Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland; and Webster Vienna, Private University, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Emotions are brief episodes during which several functional components are synchronized in response to an eliciting event. Emotions can be assessed by measuring responses in each of these components: subjective feeling, physiology, expression, and motivation. In this article we review the most advanced methods that can be used to assess emotions in each of the four components: self-report measures and questionnaires, measures of physiological activation, instruments for the analysis of nonverbal expressive behavior, and behavioral tasks.

Introduction

Even if the exact definition of 'emotion' is still a matter of debate between theorists, there is general consensus about the main components of an emotion episode (Frijda and Scherer, 2009). An emotion episode is originated by the appraisal of an eliciting event that causes synchronized changes in four main components: a physiological component, an expressive component, a motivational component, and the subjective feeling (Mortillaro et al., 2013). The changes in these four components are, therefore, informative of the state of the person and can be used as indicators of an emotional experience. Emotion assessment methods, indeed, can be grouped based on which of these four components they target.

In this article we review some of the measures that have been proposed to assess emotions and group them according to the component to which they refer. We focus on methods that are elaborated for nonclinical adult human samples. We will not consider all the instruments that were used for the purpose of better understanding the key elements of an emotional experience, namely cognitive antecedents (e.g., the GRID questionnaire; Fontaine et al., 2007), correlates of emotions, and emotional personality traits.

Two components have received particular attention in the context of emotion assessment: subjective feeling, by means of questionnaires and self-report measures; and expression, by means of coding instruments built to objectively measure facial, vocal, and bodily expressions. Physiology has been the object of several studies but researchers mostly focused on the existence of emotion-specific physiological patterns and clear indications about how using physiology to assess emotions are yet to come. The motivational component is the least studied and we will discuss it only briefly.

Subjective Feeling: Emotion Assessment through Questionnaires and Self-Report Measures

The subjective feeling is the holistic cognitive representation of all the changes elicited by the emotional event; this representation allows the individual to be aware of his/her state and name it with a particular emotion label (Scherer, 2009). Self-report measures are probably the best (if not the only) way to get the individuals' conscious representation of their emotional experience. Nevertheless, when using a self-report measure, the researcher should always consider the risk of social desirability of the responses and the variability that exists between individuals with respect to the ability or the willingness to report on emotional states.

There is not one standard self-report measure or procedure that can be used for any purposes in any context, but rather there are several self-report measures, each of which has specific contexts of application, and theoretical implications. As a general remark, state of the art self-report measures are based on clear theoretical grounds, are empirically validated, and are used to assess current emotional experience rather than reports about past or future events, and stable emotion-related traits (Mauss and Robinson, 2009).

Instruments to assess the emotional subjective feeling can be divided into four groups, defined along two major axes. The first axis is the theoretical model they refer to emotions as discrete categories (Ekman et al., 1983; Izard, 1977) or emotions as states defined by dimensions (Russell, 1980). The second axis is the visual format of the instrument: purely verbal (only text) or instruments that include pictorial elements. Generally speaking, purely verbal measures – in the format of adjective checklists – have the advantage of being more specific and precise when measuring the individuals' subjective feeling. The pictorial instruments, however, are generally faster to administer and easier to understand for the respondents, and, more importantly, they do not rely on verbal knowledge and, therefore, do not need translation to be used in different cultures and can be administered to populations that do not possess complex verbal knowledge (e.g., children).

The discrete emotion approach is used to categorize the experience into separate states that are labeled with emotion words. Most frequently, this has been done through text-based instruments that ask people to rate the extent to which they experienced certain emotional state in a specific context, period of time, or in response to a stimulus. The exact form in which the ratings are given differs across instruments. Several studies assess emotions by simply asking to choose one emotion label from a predefined list of emotion words (nominal level). Some adjustments have been suggested in order to improve the

reliability of this procedure; in particular, respondents should have the possibility of choosing an option 'no emotion' and to be allowed to indicate another emotion label that is not included in the list (Bänziger et al., 2012; Scherer, 2005). One-item nominal answers provide quick access to a global estimate of the subjective feeling of the person, but the reliability remains rather low. Using multiple items to assess an emotional state has greater reliability and can assess a larger part of the emotional spectrum. These procedures require the respondents to rate on a scale (ordinal or interval level) the extent to which each one of a list of emotions was experienced. In the following paragraph we describe some of the most commonly used text-based measures that adopt a discrete perspective.

The most frequently used instrument in this tradition is the *Differential Emotion Scale* (DES; Izard, 1977; Izard et al., 1993). The DES "is a standardized instrument that reliably divides the individual's description of emotion experience into validated, discrete categories of emotion" (Izard, 1977: p. 124). Subjects indicate on 5-point scales (from 'not at all' to 'very strongly') the strength of their feelings using an adjective checklist made of 30 items. These 30 items are grouped in 10 discrete emotion categories, or factors, and were derived from cross-cultural studies on emotion–expression labeling (Izard, 1971). The 10 factors are: anger, surprise, interest, contempt, fear, guilt, shame/shyness, joy, sadness, and disgust. Other versions of the instrument have been proposed over the years. In the DES II, the item content was refined and the instructions changed to allow the assessment of emotions experienced over extended time periods (Izard, 1972, 1977). The DES III is a version of the DES adapted for children and adolescents and includes phrases that describe the subjective feeling associated with each primary emotion (Kotsch et al., 1982). The DES IV (Izard et al., 1993) includes 12 factors instead of 10, with separate scales for Shame and Shyness, and a new scale for Self-Hostility.

Another 'text-only' instrument is the emotion part of the *Stress Questionnaire* by Foikman and Lazarus (1985). This questionnaire was developed in the context of stress research and includes a section about emotion. The subjects are asked to indicate on a 5-point Likert scale the extent to which they felt each of 15 emotions, grouped in 4 scales. The score for each scale is computed by summing the ratings for each item. The four scales are based on the main appraisal theme of the emotions: Threat emotions (worried, fearful, and anxious); Challenge emotions (confident, hopeful, and eager); Harm emotions (angry, sad, disappointed, guilty, and disgusted); and Benefit emotions (exhilarated, pleased, happy, and relieved).

Discrete emotions can also be assessed through instruments that include graphical elements that are designed to simplify the task for the respondents. The Geneva Emotion Wheel (GEW; Scherer, 2005; Scherer et al., 2013a) consists of several discrete emotion labels arranged in a circle. The disposition of the labels is determined by the position of the emotions in a two-dimensional space defined by valence and control/power according to both theoretical modeling and empirical data (Scherer et al., 2013a). The intensity of the emotion is mapped on the distance between the center and the rim of the wheel. Respondents can choose the emotion and indicate its intensity by choosing one of the circles that connect the emotion label to the center of the wheel: the size of each circle corresponds to the intensity of the respective emotion. The specific choice of emotions included in the GEW is not fixed. Users can modify the choice of terms for their specific research purposes but they should respect the underlying theoretical structure.

An example of a purely pictorial instrument to assess discrete emotions is The Product Emotion Measuring Instrument, an instrument that was specially developed for measuring emotional responses to products (PrEmo; Desmet et al., 2007). PrEmo is a self-report instrument that uses cartoon characters to represent emotions. The character expresses seven positive emotions (inspiration, desire, satisfaction, pleasant surprise, fascination, amusement, and admiration), and seven negative emotions (disgust, indignation, contempt, disappointment, dissatisfaction, boredom, and unpleasant surprise.) The interface includes still pictures of the character's animations with a three-point scale: "I feel the emotions," "to some extent I feel the emotion," and "I do not feel the emotion."

The measures built on the basis of a dimensional account describe the emotion experience on a number of underlying dimensions. The specificity of the instruments that adopt a dimensional approach is in the output that they provide more than in their format – adjective checklist. These instruments vary in terms of the emotional dimensions that they are targeting.

The Positive and Negative Affect Schedule (PANAS; Watson et al., 1988) is the most frequently used adjective checklist in the dimensional tradition. The PANAS is an adjective checklist made of 20 emotion items that are used to estimate the emotional experience of the respondents in terms of two orthogonal dimensions: Positive Affect and Negative Affect. Positive Affect represents the extent to which a person is enthusiastic, active, and alert; Negative Affect represents general distress and unpleasant engagement. Respondents are asked to use a 5-point unipolar intensity scale for each of the 20 items to describe their emotional experience. The PANAS has published normative and clinical data; it provides an interval scale level, and has been used in several research domains. A more recent version is the PANAS-X, which includes 60 items that measure the same two higher order scales of the original PANAS (Positive Affect and Negative Affect) in addition to 11 specific affects (Fear, Sadness, Guilt, Hostility, Shyness, Fatigue, Surprise, Joviality, Self-Assurance, Attentiveness, and Serenity; Watson and Clark, unpublished manuscript).

An alternative dimensional model is the Pleasure-Arousal-Dominance model (PAD), which assumes that these three independent dimensions are necessary and sufficient to describe emotional states (Mehrabian and Russell, 1974; Russell and Mehrabian, 1977). Unlike the PANAS, the PAD model considers valence as one dimension. Mehrabian and Russell (1974) devised semantic differential scales to directly assess these three dimensions. Each of these three scales is made of six bipolar adjective pairs that are rated using a 9-point scale. More recently, Russell et al. (1989) suggested a single-item scale to measure two of the three dimensions of the PAD model, pleasure and arousal. This instrument, called Affect Grid, is a 9 by 9 grid in which the horizontal dimension represents pleasure and the vertical dimension

arousal: the respondents check one of the 81 cells to report their level of arousal and pleasure.

The Brief Mood Introspection Scale is an instrument that integrates both dimensional perspectives in the same adjective checklist (BMIS, Mayer and Gaschke, 1988). Indeed, the BMIS can be used to measure the mood in four continuous scales: Pleasant-Unpleasant scale; Arousal-Calm scale; Positive-Tired scale; and Negative-Relaxed scale. The checklist consists of 16 adjectives, 2 from each mood states: happy, loving, calm, energetic, fearful/anxious, angry, tired, and sad. The respondents rate each adjective using a 4-point scale.

Researchers in the dimensional approach frequently use the Self-Assessment Maniking (SAM; Bradley and Lang, 1994), a nonverbal pictorial assessment technique that directly measures pleasure, arousal, and dominance. The SAM is both a computerized and paper-and-pencil instrument that consists of three sets of graphic illustrations for each of the three major dimensions. The pleasant dimension is represented by 5 illustrations, with at one extreme a smiling figure and at the other end a frowning and inverted-U mouth shaped figure; the arousal dimension ranges from a wide-open-eyed 'excited' figure to a sleepy one (i.e., eyes are closed); the Dominance dimension is metaphorically represented in terms of the size of the figure (the bigger the size, the higher the feeling of dominance, control).

In the paper-and-pencil version of the SAM, the respondents can, for each dimension, mark one of the pictures or the space between any two of them to report about their feeling – resulting in a 9-point scale. This instrument is relatively easy to understand and faster to use compared to any adjective checklist. However, the output of this scale is based on a single item response per dimension.

Physiology: Autonomic Measures of Emotion

All emotion theories postulate that physiological changes are an inherent aspect of the emotional experience (Frijda and Scherer, 2009). Physiological changes happen in different organismic subsystems, but it is the autonomic nervous system (ANS) that attracted most of the researchers' attention. The ANS is the set of peripheral nerves and ganglia that modulate peripheral functions and it is made of two branches, the sympathetic system – generally associated with activation, – and the parasympathetic system – generally associated with relaxation. The activity of the ANS is implicated in several bodily functions that are not related to the emotion process per se; however, the emotional state of the individual can alter some physiological indices.

Most physiological measures used as indicators of emotion can be grouped in three functional groups: respiratory functioning (e.g., respiration rate), cardiac functioning (e.g., heart rate), and electro-dermal activity (e.g., galvanic skin response). Several studies followed the pioneering work of Ekman et al. (1983; Levenson et al., 1990) and aimed at identifying emotion-specific autonomic patterns. Recent reviews, however, do not explicitly confirm the existence of emotion specificity (Kreibig, 2010; Larsen et al., 2008). Although stable differences can be found between pairs of emotions (e.g., anger is associated to higher diastolic blood pressure than fear), the search for emotion-specific autonomic patterning is still inconclusive. Kreibig (2010) proposed that findings should be interpreted in the light of the motivational dispositions underlying the emotions. For example, decreased cardiac activity was associated with emotions characterized by an approach motivation, while an increase is found in emotions that are related to an active coping response to negative situations. Similarly, increased electro-dermal response happens in almost all emotions and may be a symptom of the motor preparation that is an integral part of the emotion episode (action tendency, Frijda, 1986), while decrease is found for all emotions in which passivity is the shared motivational state.

Researchers who did not focus on the specific associations between discrete emotion and physiological signals investigated, instead, physiological correlates of emotional dimensions like arousal and valence. Several studies found that physiological indices can be reliably used under controlled conditions to estimate the level of arousal/activation associated with an emotional experience. In particular, electro-dermal response was repeatedly suggested as a proxy to estimate the arousal level of the person experiencing an emotion: for example, in one very influential study, Bradley, Lang, and colleagues (2000) showed that the level of skin conductance is linearly correlated to the perceived arousing properties of emotional stimuli. In connection to the valence dimension, research findings seem to be less consistent. Several studies used facial Electromyography (EMG) to assess valence: In these cases, researchers placed electrodes on muscles that are supposedly involved in valenced responses, namely *zygomaticus mayor*, the muscle that pulls the lip corners upward producing a smile (usually associated with positive responses), and the *corrugator*, the muscle that lowers the eyebrow while pulling them together, producing the frown (usually associated with negative responses).

Expression: Emotion Assessment from Nonverbal Behavior

The study of nonverbal behavior has a particular status in emotion research, as it is at the core of some of the most influential emotion theories. For example, basic emotion theory (e.g., Ekman, 1992) was largely inspired by Darwin's (1872) seminal work on the evolutionary bases of emotional expression. Dimensional models of affect derived from studies on the perception of facial behavior (Schlosberg, 1952; Russell and Bullock, 1985), and componential theories of emotion view nonverbal emotional expression as a major component of affect (Scherer and Ellgring, 2007a; Smith and Scott, 1997). The dynamic systems approach of socio-emotional development is based on the study of nonverbal exchanges between infants and caretakers (Fogel et al., 1992). The study of nonverbal behavior also fueled major controversies about the role of emotion in communication and about the universality of emotional expressions (Fridlund, 1994; Owren and Bachorowski, 2003; Russell, 1994; Scherer, 1992).

To understand the connection between emotion and nonverbal behavior, it is useful to consider the former as a process of interrelated changes in several components of

psychobiological functioning (Frijda, 2007; Roseman and Smith, 2001; Scherer, 2001; Smith, 1989). In this view, emotional communication belongs to the motor expression component and would function to convey reactions and behavioral intentions to conspecifics (Scherer, 2005). Emotional expressions could also function to modify the internal states of perceivers in a way that is advantageous to signalers (Owren and Bachorowski, 2003), or to ensure the honesty of social signals as indicators of adaptive dispositions (Brown et al., 2003; Mehu and Scherer, 2012). It is believed that, from an evolutionary point of view, the function of emotional expression is better served by multimodal patterns of behavioral cues including facial, vocal, and bodily signals (Mortillaro et al., 2013; Scherer and Ellgring, 2007b). The measurement of these signals may therefore provide an objective, though indirect, basis to assess emotional experience (Ekman et al., 1980).

Face

Facial behavior has dominated the research on emotional communication, and different models have been proposed to account for the association between the face and affect.

The continuous development of coding schemes for the systematic measurement of facial movements testifies of the great interest in this research area. The most recent coding scheme to date, the Facial Action Coding System (or FACS, Ekman and Friesen, 1978), is a comprehensive, anatomically based system to measure facial behavior, which allows researchers to compare their results across individuals and cultures. Each functional muscular movement that happens on the face can be labeled with a specific code (Action Unit, AU). Combinations of action units can be used to describe, unequivocally, any possible facial expression. In addition, explicit rules for the coding of action units boundaries and phases (onset, apex, and offset) are included in the coding system. Because the use of FACS does not require interpretation about the mechanisms responsible for the activation of facial muscles, all researchers can use it, independently of their attachment to particular theoretical frameworks.

The most influential theory in facial expression research postulates that a limited number of discrete emotions – or families of emotion – have specific signatures in configurations of facial movements (Basic Emotion Theory; Ekman, 1992; Izard, 1971; Tomkins and Carter, 1964). In this view, specific configurations of facial actions indicate the presence of discrete basic emotions such as joy, anger, fear, or disgust, to name a few. The reliability of emotional signaling would be ensured by the presence, within these facial configurations, of movements that are not easy to manipulate voluntarily (Ekman, 2003; Mehu et al., 2012). Researchers who adopted this approach developed the Emotion FACS (EMFACS) in which coding is directly related to the facial movements and configurations that are considered signals of discrete emotions. Despite the popularity of the evidence in support of Basic Emotion Theory, a recent review suggests that the coherence between discrete emotions and facial behavior is relatively weak (Reisenzein et al., 2013).

Dimensional models of affect (e.g., Russell, 1980) postulate that facial movements are indicative of fundamental affective dimensions such as valence and arousal. In this view, specific facial movements do not represent discrete emotion categories but underlying dimensions of affect. For example, the pulling of lip corners (most commonly known as smiling) indicates positive valence, whereas eyebrow raise and eye opening would reflect emotional arousal (Snodgrass, 1992). Recent research has extended the number of underlying dimensions to four: valence, arousal, power/control, and predictability/novelty (Fontaine et al., 2007), and the association between these dimensions and facial behavior remains to be systematically tested. The links between emotional dimensions and facial action units are, however, unlikely to be specific, as a single facial movement may be associated with several dimensions (Mehu and Scherer, under review).

A third group of theoretical models, called Componential Models, suggests that the different elements of facial behavior (action units) reflect the outcome of cognitive appraisal processes, such as the evaluation of novelty, pleasantness, or coping potential (Scherer and Ellgring, 2007b; Smith and Scott, 1997). Appraisal models of facial expressions received empirical support in several studies. For example, the eyebrow frown is related to perceived goal-obstacles (Smith, 1989), and movements related to eye opening (such as eyebrow raise, and upper eyelid raise) are associated with the appraisal of novelty or unpredictability, namely, greater eye opening indicates perceived novelty (Mortillaro et al., 2011). The later study also revealed that cheek raise is associated with the appraisals of intrinsic pleasantness and goal conduciveness. Evidence in support of this model also comes from psychophysiological research in which a variety of AUs have been related to the outcome of appraisal processes (for a brief review of this literature, see Scherer et al., 2013b).

Voice

Voice has always been considered an important channel of emotional communication (Scherer, 1986). Research shows that infants as young as 5 months of age can discriminate happy from sad voices (Caron et al., 1988). People are able to reliably infer valence of nonlinguistic vocalizations (Fecteau et al., 2005), acoustic properties of speech can be used as an index of emotional intensity (Bachorowski and Owren, 1995), and mood change can be traced using vocal cues (Ellgring and Scherer, 1996). The voice is a powerful medium to communicate emotion, as it has a larger reach than signals that require the visual attention of perceivers in order to operate.

Although the relationship between voice and discrete emotion categories has not been researched extensively, a recent study suggests that basic emotions might have distinct vocal signatures (Sauter and Scott, 2007). People have been shown to accurately discriminate among different emotion categories on the basis of vocal cues only (Bänziger et al., 2012; Hawk et al., 2009), and there is some evidence that vocal emotional expressions are universal (Bryant and Barrett, 2008). The acoustic information used by perceivers to classify emotions is extremely varied and includes different measures of pitch (the proximal percept of fundamental frequency), vocal intensity, spectral measures, speech rate, and the duration of voiced periods (Banse and Scherer, 1996; Scherer, 1986). It is likely that the computation of new vocal parameters will,

ultimately, enlarge the list of vocal features related to emotional experience.

Like for faces, vocal expressions also convey information related to underlying dimensions of emotional experience. Arousal or activation is the main dimension that has been linked to voice (Banse and Scherer, 1996; Goudbeek and Scherer, 2010) and acoustic of speech can be used as an index of emotional intensity (Bachorowski and Owren, 1995). The vocal channel is also believed to convey information relative to the 'power dimension,' that is, the ability to control environmental situations (Goudbeek and Scherer, 2010). Emotion expressed in the voice varies according to the circumstances in which the emotion is produced as well as dispositional differences in positive and negative emotionality (Bachorowski and Owren, 1995), suggesting that cognitive appraisal of situations as well as emotional valence are connected to the vocal channel. Finally, it appears that the acoustic parameters involved in the production and perception of emotion share a common basis with sounds used in musical performance to express and elicit emotions (Juslin and Laukka, 2003).

Body

Body movements are of prime importance in the judgments of emotional experience (Montepare et al., 1999; Wallbott, 1998), psycho-affective disorders (Ekman and Friesen, 1974; Wallbott, 1985), and emotional arousal (Mehu and van der Maaten, 2014). Two aspects are important in the evaluation of bodily cues: Static configurations (or body postures) and movement quality. Body postures can be relatively stable over time and change much more slowly than facial and vocal cues. As a result, they have been considered as indicators of 'gross affective states' such as the feeling of pleasantness (Ekman and Friesen, 1969). On the other hand, the complexity of movement offered by the different body parts (e.g., hand and head movements) makes the body an important hub for the communication of subtle and transient emotional states. These ideas suggest that different properties of bodily cues could reflect different emotional information.

Despite the idea that the body is a rich source of emotional information (Darwin, 1872; Ekman, 1965), research on bodily cues of affect has been much less frequent than research on facial and vocal cues. The difficulty inherent to the development of systematic coding systems for bodily cues is probably the main reason for the paucity of research on body-affect relationships (Harrigan, 2005). In recent years, however, we have assisted new attempts at developing comprehensive coding schemes for body movements and postures, notably the Body Action and Posture coding system (BAP; Dael et al., 2012a). The BAP, inspired by the FACS, adopts an integrative approach to the dynamic description of body movement on an anatomical level (different articulations of body parts), a form level (direction and orientation of movement), and a functional level (communicative and self-regulatory functions).

In the same manner as in facial and vocal expression research, scholars interested in bodily cues of emotion have tried to isolate movements that are specific to discrete emotional states (Coulson, 2004; Tracy and Robins, 2008; Wallbott, 1998). However, specificity in movement patterns to emotional categories has not received conclusive support, as the same combination of movements often occurs in different emotional states (Dael et al., 2012b). For example, raising the arms up is found in expressions of both pride (Tracy and Robins, 2008) and joy (Boone and Cunningham, 1998; Wallbott, 1998). Instead, recent evidence suggests that patterns of body movements are associated with functional components of emotions (e.g., modes of appraisal and action readiness) that show considerable overlap between emotion categories (Dael et al., 2012b). Details about the bodily cues associated with emotion categories can be found in several studies (Dael et al., 2012b; Montepare et al., 1999; Wallbott, 1998).

Motivational Component: Assessing Emotions through Motor Actions

Despite the importance that all theories attribute to the motivational and action component of emotions (Frijda, 1986), there are very few studies that suggested to use motor actions (that are not considered as communicative units) to assess emotions.

Russell and Mehrabian (1978) suggested that the behavior of a person changes as a function of the emotion-eliciting quality of the environment, but did not consider the issue of emotion assessment. Motivational behaviors such as approach and avoidance are frequently used to estimate whether a stimulus is positive or negative and indirectly the valence of the reaction of the person (Chen and Bargh, 1999). However, several authors suggested that anger, a negative emotion, is also characterized by an approach motivation (e.g., Carver and Harmon-Jones, 2009; Wilkowski and Meier, 2010) and, therefore, motivational direction and affective valence should be dissociated (Coombes et al., 2007).

The most frequently used paradigm to estimate approach and avoidance reactions uses the movements of the arm of the respondent. Approach is associated with moving objects closer to a person, while avoidance is associated with pushing them away. For example, Marsh et al. (2005) found that angry facial expressions facilitated avoidance behavior that is, pushing a joystick away from oneself more quickly than pulling it closer.

In another version of the paradigm, the effect of pulling the joystick closer to the self or pushing it away from the self has the consequence to, respectively, magnify (i.e., increase the size) or shrink (i.e., reduce the size) the picture that is observed (Rinck and Becker, 2007). The perceptual effect is that the pictures are coming closer or moving away depending on the movement of the joystick. Nevertheless, recent work questioned the absolute validity of such tasks by showing that a modification in the instructions would cause different results than predicted (Eder and Rothermund, 2008). This led these authors to argue that evaluative implications of the instructions assign affective codes to motor behavior that interact with stimulus evaluation.

Conclusion

Past research showed that several instruments can be used to assess the emotional state of a person. Nevertheless, our review

showed there is not one single method for emotion assessment that is ideal under all circumstances. The choice of the method depends on the research question and on the resources available for the realization of specific research designs. Most importantly, researchers should look for the method that best allows them to test their hypotheses, and consider using multiple assessment measures at the same time in order to increase the internal validity of their studies.

Acknowledgments

This research was supported by the National Centre of Competence in Research (NCCR) Affective Sciences financed by the Swiss National Science Foundation (51NF40–104897) and hosted by the University of Geneva.

See also: Culture and Emotion; Early Emotional Development and Cultural Variability; Emotion and Expression; Emotion in Cognition; Emotion, Perception and Expression of; Emotional Intelligence and Competencies; Emotional Regulation; Emotions and Aging; Emotions and Health; Emotions, Psychological Structure of; Facial Emotion Expression, Individual Differences in; Selfconscious Emotions, Psychology of.

Bibliography

Bachorowski, J.-A., Owren, M.J., 1995. Vocal expression of emotion: acoustic properties of speech are associated with emotional intensity and context. Psychological Science 6 (4), 219–224.

Banse, R., Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70, 614–636.

Bänziger, T., Mortillaro, M., Scherer, K.R., 2012. Introducing the Geneva Multimodal Expression corpus for experimental research on emotion perception. Emotion 12 (5), 1161–1179.

Boone, R.T., Cunningham, J.G., 1998. Children's decoding of emotion in expressive body movement: the development of cue attunement. Developmental Psychology 34 (5), 1007.

Bradley, M.M., Lang, P.J., 1994. Measuring emotion: the self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry 25 (1), 49–59.

Bradley, M.M., Lang, P.J., 2000. Measuring emotion: behavior, feeling and physiology. In: Lane, R., Nadel, L. (Eds.), Cognitive Neuroscience of Emotion. Oxford University Press, New York, pp. 242–276.

Brown, W.M., Palameta, B., Moore, C., 2003. Are there nonverbal cues to commitment? An exploratory study using the zero-acquaintance video presentation paradigm. Evolutionary Psychology 1, 42–69.

Bryant, G.A., Barrett, H.C., 2008. Vocal emotion recognition across disparate cultures. Journal of Cognition and Culture 8, 135–148.

Caron, A.J., Caron, R.F., MacLean, D.J., 1988. Infant discrimination of naturalistic emotional expressions: the role of face and voice. Child Development 59 (3), 604–616. http://dx.doi.org/10.2307/1130560.

Carver, C.S., Harmon-Jones, E., 2009. Anger is an approach-related affect: evidence and implications. Psychological Bulletin 135 (2), 183–204.

Chen, M., Bargh, J.A., 1999. Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. Personality and Social Psychology Bulletin 25, 215–224.

Coombes, S.A., Cauraugh, J.H., Janelle, C.M., 2007. Emotional state and initiating cue alter central and peripheral motor processes. Emotion 7 (2), 275–284.

Coulson, M., 2004. Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence. Journal of Nonverbal Behavior 28 (2), 117–139.

Dael, N., Mortillaro, M., Scherer, K.R., 2012a. The body action and posture coding system (BAP): development and reliability. Journal of Nonverbal Behavior 36 (2), 97–121. http://dx.doi.org/10.1007/s10919-012-0130-0.

Dael, N., Mortillaro, M., Scherer, K.R., 2012b. Emotion expression in body action and posture. Emotion 12 (5), 1085–1101.

Darwin, C., 1872. The expression of the emotions in man and animals. John Murray, London, England. http://dx.doi.org/10.1037/10001-000.

Desmet, P.M.A., Porcelijn, R., van Dijk, M., 2007. Emotional design; application of a research based design approach. Journal of Knowledge, Technology & Policy 20 (3), 141–155.

Eder, A.B., Rothermund, K., 2008. When do motor behaviors (mis)match affective stimuli? An evaluative coding view of approach and avoidance reactions. Journal of Experimental Psychology – General 137 (2), 262–281.

Ekman, P., 1965. Differential communication of affect by head and body cues. Journal of Personality and Social Psychology 2 (5), 726–735.

Ekman, P., 1992. An argument for basic emotions. Cognition & Emotion 6 (3/4), 169–200.

Ekman, P., 2003. Darwin, deception, and facial expression. In: Campos, J.J., Davidson, R.J., de Waal, F.B. (Eds.), Emotions inside out, vol. 1000. New York Academy of Sciences, New York, pp. 205–211.

Ekman, P., Friesen, W.V., 1969. The repertoire of nonverbal behavior: categories, origins, usage, and coding. Semiotica 1 (1), 49–98.

Ekman, P., Friesen, W.V., 1974. Nonverbal behavior and psychopathology. The Psychology of Depression: Contemporary Theory and Research 203–224.

Ekman, P., Friesen, W.V., 1978. The Facial Action Coding System, first ed. Consulting Psychologists Press, Palo Alto, CA.

Ekman, P., Friesen, W.V., O'Sullivan, M., Scherer, K., 1980. Relative importance of face, body, and speech in judgments of personality and affect. Journal of Personality and Social Psychology 38 (2), 270–277.

Ekman, P., Levenson, R.W., Friesen, W.V., 1983. Autonomic nervous system activity distinguishes among emotions. Science 227 (4616), 1208–1210.

Ellgring, H., Scherer, K.R., 1996. Vocal indicators of mood change in depression. Journal of Nonverbal Behavior 20 (2), 83–110.

Fecteau, S., Armony, J.L., Joanette, Y., Belin, P., 2005. Judgement of emotional nonlinguistic vocalizations: age-related differences. Applied Neuropsychology 12 (1), 40–48.

Fogel, A., Nwokah, E., Dedo, J.Y., Messinger, D., Dickson, K.L., Matusov, E., Holt, S.A., 1992. Social process theory of emotion: a dynamic systems approach. Social Development 1 (2), 122–142. http://dx.doi.org/10.1111/j.1467-9507.1992.tb00116.x.

Foikman, S., Lazarus, R.S., 1985. If it changes it must be a process: a study of emotion and coping during three stages of a college examination. Journal of Personality and Social Psychology 48, 150–170.

Fontaine, J.R.J., Scherer, K.R., Roesch, E.B., Ellsworth, P.C., 2007. The world of emotions is not two-dimensional. Psychological Science 18, 1050–1057.

Fridlund, A.J., 1994. Human Facial Expression: An Evolutionary View. Academic Press, San Diego, CA.

Frijda, N.H., 1986. The Emotions. Cambridge University Press, Cambridge, UK.

Frijda, N.H., 2007. The Laws of Emotion. Erlbaum, Mahwah, NJ.

Frijda, N.H., Scherer, K.R., 2009. Emotion definitions (psychological perspectives). In: Sander, D., Scherer, K.R. (Eds.), The Oxford Companion to Emotion and Affective Sciences. Oxford University Press, New York, NY, pp. 142–144.

Goudbeek, M., Scherer, K.R., 2010. Beyond arousal: valence and potency/control cues in the vocal expression of emotion. Journal of the Acoustical Society of America 128 (3), 1322–1336.

Harrigan, J.A., 2005. Proxemics, kinesics, and gaze. In: Harrigan, J.A., Rosenthal, R., Scherer, K.R. (Eds.), The New Handbook of Methods in Nonverbal Behavior Research. Oxford University Press, New York, NY, pp. 137–198.

Hawk, S.T., van Kleef, G.A., Fischer, A.H., van der Schalk, J., 2009. "Worth a thousand words": absolute and relative decoding of nonlinguistic affect vocalizations. Emotion 9 (3), 293–305. http://dx.doi.org/10.1037/a0015178.

Izard, C.E., 1971. The Face of Emotion. Appleton-Century-Crofts, New York.

Izard, C.E., 1972. Patterns of Emotions: A New Analysis of Anxiety and Depression. Academic Press, San Diego, CA.

Izard, C.E., 1977. Human Emotions. Plenum Press, New York.

Izard, C.E., Libero, D.Z., Putnam, P., Hayes, O.M., 1993. Stability of emotional experiences and their relations to traits of personality. Journal of Personality and Social Psychology 64, 847–860.

Juslin, P.N., Laukka, P., 2003. Communication of emotions in vocal expression and music performance: different channels, same code? Psychological Bulletin 129 (5), 770–814. http://dx.doi.org/10.1037/0033-2909.129.5.770.

Kotsch, W.E., Gerbing, D.W., Schwartz, L.E., 1982. The construct validity of the differential emotions scale as adapted for children and adolescents. In: Izard, C.E. (Ed.), Measuring Emotions in Infants and Children. Cambridge University Press, New York, NY, pp. 251–278.

Kreibig, S.D., 2010. Autonomic nervous system activity in emotion: a review. Biological Psychology 84, 394–421.

Larsen, J.T., Berntson, G.G., Poehlmann, K.M., Ito, T.A., Cacioppo, J.T., 2008. The psychophysiology of emotion. In: Lewis, R., Haviland-Jones, J.M., Feldman-Barrett, L. (Eds.), The Handbook of Emotions, third ed. Guilford, New York.

Levenson, R.W., Ekman, P., Friesen, W.V., 1990. Voluntary facial action generates emotion-specific autonomic nervous system activity. Psychophysiology 27, 363–384.

Marsh, A.A., Ambady, N., Kleck, R.E., 2005. The effects of fear and anger facial expressions on approach- and avoidance-related behaviors. Emotion 5, 119–124.

Mauss, I.B., Robinson, M.D., 2009. Measures of emotion: a review. Cognition & Emotion 23 (2), 209–237.

Mayer, J.D., Gaschke, Y.N., 1988. The experience and meta-experience of mood. Journal of Personality and Social Psychology 55, 102–111.

Mehrabian, A., Russell, J.A., 1974. An Approach to Environmental Psychology. The MIT Press.

Mehu, M., Scherer, K.R., 2012. A psycho-ethological approach to social signal processing. Cognitive Processing 13 (2), 397–414. http://dx.doi.org/10.1007/s10339-012-0435-2.

Mehu, M., Scherer, K.R., n.d. Emotional categories and dimensions in the facial communication of affect: an integrated approach. Manuscript Under Review.

Mehu, M., van der Maaten, L.J., 2014. Multimodal integration of dynamic audio-visual cues in the communication of agreement and disagreement. Journal of Nonverbal Behavior 38, 569–597.

Mehu, M., Mortillaro, M., Bänziger, T., Scherer, K.R., 2012. Reliable facial muscle activation enhances recognizability and credibility of emotional expression. Emotion 12 (4), 701–715. http://dx.doi.org/10.1037/a0026717.

Montepare, J., Koff, E., Zaitchik, D., Albert, M., 1999. The use of body movements and gestures as cues to emotions in younger and older adults. Journal of Nonverbal Behavior 23, 133–152.

Mortillaro, M., Mehu, M., Scherer, K.R., 2011. Subtly different positive emotions can be distinguished by their facial expressions. Social Psychological and Personality Science 2 (3), 262–271.

Mortillaro, M., Mehu, M., Scherer, K.R., 2013. The evolutionary origin of multimodal synchronization in emotional expression. In: Altenmüller, E., Schmidt, S., Zimmerman, E. (Eds.), Evolution of Emotional Communication: From Sounds in Nonhuman Mammals to Speech and Music in Man. Oxford University Press, New York, pp. 3–25.

Owren, M.J., Bachorowski, J.-A., 2003. Reconsidering the evolution of nonlinguistic communication: The case of laughter. Journal of Nonverbal Behavior 27 (3), 183–200.

Reisenzein, R., Studtmann, M., Horstmann, G., 2013. Coherence between emotion and facial expression: evidence from laboratory experiments. Emotion Review 5 (1), 16–23.

Rinck, M., Becker, E.S., 2007. Approach and avoidance in fear of spiders. Journal of Behavior Therapy and Experimental Psychiatry 38, 105–120.

Roseman, I.J., Smith, C.A., 2001. Appraisal theory: overview, assumptions, varieties, controversies. In: Scherer, K.R., Schorr, A. (Eds.), Series in Affective Science: Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, London, pp. 3–19.

Russell, J.A., 1980. A circumplex model of affect. Journal of Personality and Social Psychology 39 (6), 1161–1178.

Russell, J.A., 1994. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. Psychological Bulletin 115 (1), 102–141.

Russell, J.A., Bullock, M., 1985. Multidimensional scaling of emotional facial expressions: similarity from preschoolers to adults. Journal of Personality and Social Psychology 48 (5), 1290–1298. http://dx.doi.org/10.1037/0022-3514.48.5.1290.

Russell, J.A., Mehrabian, A., 1977. Evidence for a three-factor theory of emotions. Journal of research in Personality 11 (3), 273–294.

Russell, J.A., Mehrabian, A., 1978. Approach-avoidance and affiliation as functions of the emotion-eliciting quality of an environment. Environment and Behavior 10 (3), 355–387.

Russell, J.A., Weiss, A., Mendelsohn, G.A., 1989. Affect grid: a single-item scale of pleasure and arousal. Journal of Personality and Social Psychology 57 (3), 493–502.

Sauter, D., Scott, S., 2007. More than one kind of happiness: can we recognize vocal expressions of different positive states? Motivation and Emotion 31 (3), 192–199. http://dx.doi.org/10.1007/s11031-007-9065-x.

Scherer, K.R., Shuman, V., Fontaine, J.R.J., Soriano, C., 2013a. The GRID meets the wheel: assessing emotional feeling via self-report. In: Fontaine, J.R.J., Scherer, K.R., Soriano, C. (Eds.), Components of Emotional Meaning: A Sourcebook. Oxford University Press, Oxford, pp. 281–298.

Scherer, K.R., Mortillaro, M., Mehu, M., 2013b. Understanding the mechanisms underlying the production of facial expression of emotion: a componential perspective. Emotion Review 5 (1), 47–53.

Scherer, K.R., 1986. Vocal affect expression: a review and a model for future research. Psychological Bulletin 99 (2), 143–165. http://dx.doi.org/10.1037/0033-2909.99.2.143.

Scherer, K.R., 1992. What does facial expression express?. In: Strongman, K.T. (Ed.), International Review of Studies of Emotion, vol. 2. Wiley, Chichester, UK, pp. 139–165.

Scherer, K.R., 2001. Appraisal considered as a process of multilevel sequential checking. In: Schorr, A., Johnstone, T. (Eds.), Appraisal Process in Emotion: Theory, Method, Research. Oxford University Press, New York & Oxford, pp. 92–120.

Scherer, K.R., 2005. What are emotions? And how can they be measured? Social Science Information 44 (4), 695–729.

Scherer, K.R., 2009. The dynamic architecture of emotion: evidence for the component process model. Cognition & Emotion 23, 1307–1351.

Scherer, K.R., Ellgring, H., 2007a. Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? Emotion 7 (1), 113–130.

Scherer, K.R., Ellgring, H., 2007b. Multimodal expression of emotion: affect programs or componential appraisal patterns? Emotion 7 (1), 158–171.

Schlosberg, H., 1952. The description of facial expressions in terms of two dimensions. Journal of Experimental Psychology 44 (4), 229–237. http://dx.doi.org/10.1037/h0055778.

Smith, C.A., 1989. Dimensions of appraisal and physiological response in emotion. Journal of Personality and Social Psychology 56 (3), 339–353. http://dx.doi.org/10.1037/0022-3514.56.3.339.

Smith, C.A., Scott, H.S., 1997. A componential approach to the meaning of facial expression. In: Russell, J.A., Fernandez-Dols, J.-M. (Eds.), The Psychology of Facial Expression. Cambridge University Press, Cambridge, UK, pp. 229–254.

Snodgrass, J.D., 1992. Judgment of Feeling States from Facial Behavior: A Bottom-up Approach. Ph.D. thesis. University of British Columbia.

Tomkins, S.S., Carter, R.M., 1964. What and where are the primary affects? Some evidence for a theory. Perceptual and Motor Skills 18, 119–158.

Tracy, J.L., Robins, R.W., 2008. The nonverbal expression of pride: evidence for cross-cultural recognition. Journal of Personality and Social Psychology 94 (3), 516–530.

Wallbott, H.G., 1985. Hand movement quality: a neglected aspect of nonverbal behavior in clinical judgement and person perception. Journal of Clinical Psychology 41 (3), 345–359.

Wallbott, H.G., 1998. Bodily expression of emotion. European Journal of Social Psychology 28 (6), 879–896. http://dx.doi.org/10.1002/(SICI)10990992(199811028:6<879::AID-EJSP901>3.0.CO;2-W.

Watson, D., Clark, L.A., 1994. The PANAS-X: Manual for the Positive and Negative Affect Schedule-Expanded Form. University of Iowa, Iowa City. Retrieved online on 21.03.14 at: http://www2.psychology.uiowa.edu/faculty/Watson/PANAS-X.pdf.

Watson, D., Clark, L.A., Tellegen, A., 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. Journal of Personality and Social Psychology 54, 1063–1070.

Wilkowski, B.M., Meier, B.P., 2010. Bring it on: angry facial expressions potentiate approach-motivated motor behavior. Journal of Personality and Social Psychology 98 (2), 201–210.

Emotions, Psychological Structure of

Vera Shuman, University of Lausanne, Lausanne, Switzerland
Klaus R Scherer, Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Emotions are short-term episodes involving feelings, cognitive appraisals, motivational action tendencies, motor expressions, and physiological changes. External or internal events elicit emotions. Emotions have been shaped by evolutionary adaptation and by sociocultural contexts. Theorists disagree about the relative importance of and the interrelations between emotion components, the differentiation of emotional from nonemotional episodes, the relative extent of evolutionary and sociocultural impact (more universality vs more cultural specificity), and the dimensionality of the emotion space. Future directions in emotion research focus on aesthetic and positive emotions, emotions in social contexts, brain correlates of emotions, and interdisciplinary approaches to studying emotions.

Psychological research has identified particular characteristics of emotions that distinguish the scientific use of the term 'emotion' from its use in lay terminology, and that are used to differentiate emotions from other affective phenomena, such as moods (Section Characterizing Emotions). Beyond this relatively consensual view, however, different schools of thought in emotion psychology propose alternative answers to central questions about emotions (Section Alternative Answers to Four Central Questions Proposed by Four Families of Emotion Theories). Research programs resulting from the different theoretical perspectives have brought forth specific, and to some extent complementary, bodies of knowledge (Section Complementarities of Different Perspectives). Current developments in emotion research include the continued attempts to resolve ongoing controversies about the nature of emotions, research on the previously understudied aesthetic and positive emotions, emotions in social contexts, brain correlates of emotions, and the interdisciplinary study of emotions (Section Future Directions; see also Shuman and Scherer, 2014).

Characterizing Emotions

Emotions Are Triggered by Events

Emotions are generally elicited by actually occurring, remembered, or imagined events (e.g., failing at a task, remembering past failures, or worrying about potential failures). In some cases emotions may also be caused by direct neurophysiological factors (hormones, drugs) or entrainment (e.g., music; see Ellsworth and Scherer, 2003). A defining characteristic of emotions is that they are about something: emotions have an "event focus" (Mulligan and Scherer, 2012: p. 346). However, individuals are not always consciously aware of the stimuli or situations that have aroused the emotion.

Emotions Are Short Term

Emotions have a short duration. Typical anger and joy episodes, for example, last between 15 and 60 min and fear episodes up to 15 min (Verduyn et al., 2009). The short duration can be attributed to the prevalence of rapidly changing situations: changes occur in the object of the emotion, the individual experiencing the emotion, and the relation between the object and the individual (e.g., fear aroused by a growling dog may change when the dog stops growling, when one finds a big stick, and when one increases the distance from the dog, respectively). Situations are also altered after habituation (e.g., one gets used to the dog's growling). Personality and situational variables influence how long everyday emotional episodes last.

Emotions Are Multicomponential

Emotions involve multiple components that are associated with different functions. A subjective feeling component, a motor expression component, and a physiological component have long been considered central to emotions. Later, an additional action tendency component was advocated. Furthermore, during the cognitive revolution in the social and behavioral sciences, a cognitive appraisal component was added. However, some argue that the appraisal component is not a part of emotions but that it is rather an antecedent of emotions (see Moors, 2009).

The subjective feeling component is often used synonymously with 'emotion' or 'affect.' However, many emotion researchers regard affect as a superordinate category that includes emotions but also moods and other affective phenomena, and they define subjective feelings as one of several emotion components. The subjective feeling component is evident in verbal expressions such as 'feeling good,' 'feeling happiness,' 'feeling anger,' or 'feeling gratitude.' The function of this component is to monitor the individual's state. For example, feeling anger allows the individual to take regulatory efforts to reduce anger and feeling less anger gives the individual feedback that the regulation was successful. The monitoring function is important for regulating emotions when an individual is alone (at the intraindividual level) and when interacting with others (at the interindividual level; see Keltner and Haidt, 1999).

The function of the cognitive appraisal component is to determine the meaning of a situation for the individual regarding the situation's causes, consequences, and required

behavioral reactions. In particular, appraising an event as relevant to the self is considered a defining feature of emotions. Additional appraisals (e.g., who is responsible for the situation or whether the individual can adapt to or change the situation) enable the individual to better understand the causes of a situation and its consequences, and to assess the most promising alternatives for behavioral adaptation. Appraisals can be processed unconsciously or consciously by an expressive motor, schematic, and conceptual system (e.g., Leventhal and Scherer, 1987). Knowledge about appraisals enables the individual to regulate the occurrence of emotions in the self and others by predicting when emotions will arise. At the interindividual level, appraisals also facilitate the communication of emotional knowledge.

The action tendency component refers to the prioritization and preparation of behavior based on the appraised meaning of the situation. This motivational component allows for flexibility as to which behavior will be carried out, if any. For example, feeling afraid is associated with an urge to avoid the situation, but an organism can decide whether or not to follow that urge and how to do so. 'Control precedence' is a defining feature of emotions (Frijda, 2007) and is particularly important at the intraindividual level.

Observable facial, postural, and vocal expressions are part of the motor expression component. This component serves a communicative function at the interindividual level (e.g., a smile can communicate happiness). Additionally, the motor expression component supports action tendencies. For example, Susskind et al. (2008) showed that disgust expressions lead to sensory closure, which may support the action tendency to dispel the source of the disgust. The motor expression component is therefore also relevant at the intraindividual level.

Changes in the subjective feeling, appraisal, action tendency, and motor expression component have neurophysiological underpinnings. For example, feelings are generated by an integration of different neural structures and circuits, including somatovisceral proprioception (Kreibig et al., 2010). Appraising an event as relevant to the self is associated with increased activity in the amygdala and different areas in the frontal cortex (Sander et al., 2005). Other appraisals, such as moral judgments, are processed in the neocortex. Furthermore, action tendencies generate changes in the endocrine and autonomic nervous systems. Motor expression commands generated in the brain are executed by the somatic motor system. The ubiquity of underlying neurophysiological processes in the subjective feeling, appraisal, action tendency, and motor expression component notwithstanding, researchers have postulated a separate 'physiological' or 'somatic' component – mostly referring to autonomic and visceral phenomena – to highlight the importance of the respective symptoms independently of overt emotional behavior and facial, postural, and vocal expressions.

Emotions Are Particular Kinds of Affective Phenomena

The characteristics of emotions (i.e., evoked by events, of short duration, and having specific components) distinguish them from related affective phenomena. In particular, moods, attitudes, preferences, and affect dispositions differ from emotions (Scherer, 2005).

First, the event focus of emotions distinguishes them from moods, attitudes, and preferences. Unlike emotions, moods generally do not have an object. Moods have causes, such as hormones, but they are rarely focused on any specific cause. The event focus also differs for attitudes. Although attitudes involve beliefs about an attitude object, the attitude object need not be present in reality or in imagination to trigger the attitude. Similarly, preferences are relatively stable and do not depend on triggering events.

Second, the short duration of emotions distinguishes them from longer lasting moods. Similarly, attitudes, preferences, and affect dispositions typically last longer than emotions.

Third, emotions have more specific action tendency components than attitudes and preferences that tend to be associated with general approach–avoidance motivations. Across situations, attitudes, and preferences may be good predictors of behavior, but within a particular situation, ongoing emotions may be stronger and more specific determinants of behavior.

Finally, the subjective feeling component of emotions is more differentiated than the affective component of attitudes and of preferences. Valence (positive–negative) similarly characterizes emotions, attitudes, and preferences. However, the subjective feeling component of emotions is additionally described by a power, arousal, and novelty dimension (see Fontaine et al., 2013). The feeling component also distinguishes emotions from affect dispositions (also called trait emotions). A disposition is not an affective experience itself, but an enduring tendency to more readily experience a specific mood or emotion (e.g., an angry or an anxious personality trait, Spielberger and Reheiser, 2009).

Emotions Are Shaped by Evolution and by Sociocultural Contexts

Fundamental debates in psychology center on the question to what extent a phenomenon is the result of evolutionary or social influences. An implication of an evolutionary view on emotions is that emotion triggers, experiences, and consequences are similar across cultures (and, to some extent, across species). In contrast, an implication of a social constructionist view is that emotion triggers, experiences, and consequences vary across cultures. There is increasing evidence that both of these factors determine emotion jointly and that their relative importance may vary depending on the nature of the emotion and the respective situation.

Emotions evolved because they have been adaptive in several ways. First, the action tendency associated with particular emotions can foster survival (e.g., fear from a threat is associated with an impetus to run away and avoid the threat). At the same time, the link between emotions and actions is sufficiently flexible to allow the organism not to carry out an action tendency (e.g., standing one's ground in a threatening situation, if need be). In contrast to reflexes, the behavioral repertoire of organisms with emotions is therefore more flexible. Second, the motor expression component associated with emotions is a means to communicate to others that a situation is desirable, threatening, and so forth, and can thereby increase

the fitness of one's in-group. Third, the feeling component allows the individuals to regulate their emotions, improve the match of situational requirements and actions, and ensure group harmony; as a result, personal and group fitness is enhanced. Fourth, the appraisal component allows the organism to make meaning of a situation, to learn for the future, and to communicate to in-group members, thereby increasing personal and in-group fitness.

In addition to evolutionary influences, emotions are shaped by sociocultural factors (attitudes, values, or norms in families, peer groups, nations, generations, etc.) in several ways (e.g., De Leersnyder et al., 2013). First, social groups partially determine which stimuli an organism is exposed to (e.g., certain foods are more common and arouse more positive emotions in one culture than another). Second, social groups influence how events are appraised (e.g., whether agency is typically attributed to the self or not). Third, social groups influence which and how much of an emotion is seen as a normal/deviant and as a desirable/undesirable experience for particular people in particular situations (e.g., big boys do not cry); in other words, culture influences how emotions are normatively evaluated. Finally, social groups influence how emotions should be communicated to others, including display rules and emotion terminology (e.g., men should express their love by proposing marriage and women by saying 'yes').

Alternative Answers to Four Central Questions Proposed by Four Families of Emotion Theories

Beyond the convergent views on the characteristics of emotions described in Section Characterizing Emotions, emotion theories propose different answers to central questions about the psychological structure of emotions. Regarding the question of how emotion components relate to each other, the received opinion among philosophers and scholars in the Darwinian tradition has been that the components are lawfully organized, driven by the evaluation of the triggering situation. In contrast, some early emotion researchers suggested that emotions are the perception of emotion-specific bodily activity, or that emotions result when unspecific arousal is attributed to a situational context (James, 1884; Schachter and Singer, 1962). To date, basic emotion theorists, appraisal theorists, social constructionist theorists, and nonlinear dynamic systems theorists still propose alternative answers to this question (**Figure 1**). Theorists from these four schools of thought also respond differently to the question of how emotional episodes differ from nonemotional episodes, of how strongly an emotion is influenced by evolution and social contexts, and of how to structure the emotion space (for further details, see, e.g., Moors, 2009).

What Are the Respective Roles of and the Interrelations between Emotion Components?

In basic/discrete emotion theories (e.g., Ekman, 1992; Izard, 2007; Panksepp, 1998; Plutchik, 2001), no one emotion component is more central than another. The components are seen as jointly triggered. Component patterns occur in particular types of situations. For example, the component pattern of a fearful facial expression and an urge to run away is triggered in a threat situation, whereas the component pattern of an angry facial expression and an urge to attack is triggered in a hostile interaction.

In contrast, appraisal theories (e.g., Ellsworth and Scherer, 2003; Lazarus, 1968; Ortony et al., 1990; Roseman et al., 1996; Scherer, 2009) propose that appraisals precede and cause changes in the other four emotion components. Appraisals may be processed simultaneously and yet have an interpretable output only in a specific sequence. For example, logically goal obstructiveness appraisal outputs should precede coping potential outputs and not vice versa, as coping depends on the nature of the obstruction (Scherer, 2009). Subjective feelings are seen as an integrative representation of changes across emotion components.

In constructionist theories of emotions (Barrett, 2006; Russell, 2003), a combination of the subjective feeling component with none, few, or all of the other emotion components forms an emotion. The subjective feeling component is specified as core affect consisting of valence (positive–negative) and arousal (calm-excited). An emotion results when core affect is combined with emotional knowledge about emotion terms and scripts that describe or prescribe the appropriate pattern of emotion components for particular situations. Subjective feeling with no other emotion component can be experienced as an emotion when, for example, a particular emotion term is extremely salient (e.g., due to laboratory manipulations). As the salient conceptual knowledge generally varies across situations, individuals, and groups, so do the emotion components that are integrated with the subjective feeling to form an emotion.

Nonlinear dynamic systems theories (Camras, 2011; Fogel et al., 1992; Lewis, 2005) propose a nonlinear relation between the emotion components. Emotions result when attractor states are reached through the self-organization among the emotion components by nested positive and negative feedback loops. Any emotion component may draw the other emotion components into an attractor state. For example, smiling may cause feelings, appraisals, and physiological changes to take on the attractor state of happiness.

To summarize, basic emotion theories propose that all emotion components occur jointly, appraisal theories view appraisals as causing changes in the other four emotion components, psychological constructionist theories suggest that emotions are constructed from subjective feeling (core affect) in combination with none or all of the other components, and nonlinear dynamic systems theories argue that any emotion component can cause changes in the other components.

How Are Emotional Episodes Distinguished from Nonemotional Episodes?

A second controversy among emotion theories concerns the question of how emotions can be differentiated from nonemotional states. In basic emotion theories, an emotion sequence consists of a baseline, a stimulus, a discrete emotion, and a return to the baseline. Emotions that occur in response to an external (real or imagined) stimulus are differentiated from homeostatic drives (e.g., hunger, thirst) that occur in response to internal signals from the body and motivate the organism to maintain the body's internal milieu. Therefore, the baseline of

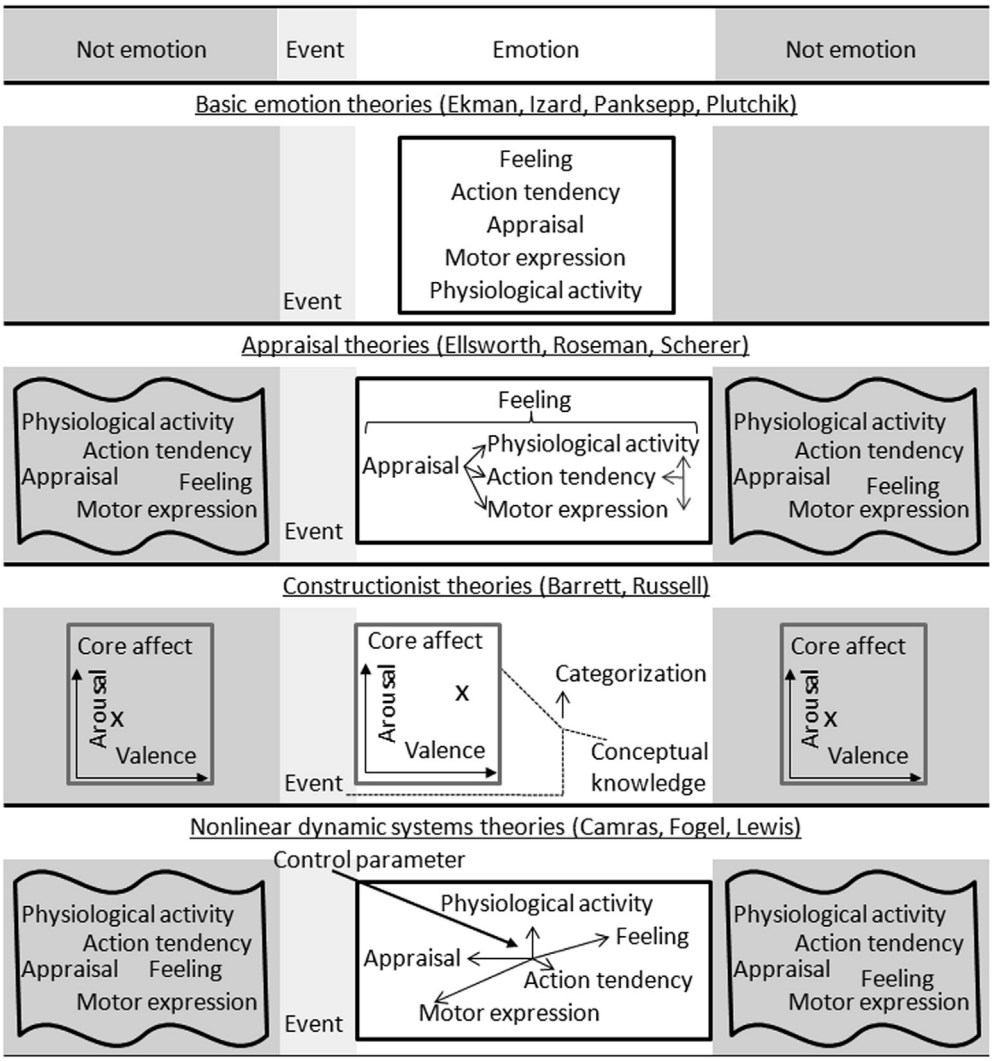

Figure 1 Theoretical perspectives on the relation among emotion components and the differentiation of emotional from nonemotional episodes. First published in Shuman, V., Scherer, K.R., 2014. Concepts and structures of emotions. In: Pekrun, R., Linnenbrink-Garcia, L. (Eds.), International Handbook of Emotions in Education. Routledge, London, pp. 13–35.

homeostatic drives is called homeostasis, but the baseline of emotions is not.

Appraisal theories propose that individuals constantly appraise their environment. As a result, states that are not considered full-blown emotions are characterized by variable activities in all emotion components, including appraisals. An emotion occurs when an event is appraised as highly relevant to the individual's goals and needs; other appraisals then differentiate among emotions. An additional criterion to distinguish emotions from nonemotions is that changes in appraisals and other components are synchronized in time (Scherer, 2009).

In psychological constructionist theories, core affect can lead to an emotion when emotional knowledge is either attributed to perceived core affect (Russell, 2003) or integrated with it (i.e., in a conceptual act; Barrett, 2006). The boundaries of an emotion are entirely dependent on the conceptual knowledge about emotions.

With nonlinear dynamic systems theories, an emotion is seen as a particular kind of state in a succession of attractor states. Emotional attractor states are seen as special because they are "motivationally charged" (Fogel et al., 1992: p. 133). As attractor states are locally stable energy minima, a large change in one or more of the emotion components may be needed before a change from one to another attractor state occurs (i.e., phase shift). For example, a spouse may stay calm when faced with an increasingly frustrating situation, but once the spouse is angry, it may take a large change in the situation to reduce the anger. As a result, the relation between a stimulus and an emotion is not linear (the degree of frustration from the event does not correspond one-to-one with the anger), but depends on the previous state of the system (i.e., calm or angry). In other words, knowledge about preceding states is needed to predict future states.

To summarize, basic emotion theories suggest that emotions differ from nonemotional states because of the joint activity in

all emotion components, appraisal theories propose that the appraisal of an event as highly relevant and component synchronization characterize emotions, psychological constructionist theories suggest that emotions are differentiated by conceptual knowledge about emotions, and nonlinear dynamic systems theories associate emotions with motivational charge.

To What Extent Are Emotions Influenced by Evolution and Sociocultural Contexts?

Emotions are subjected to the influences of evolution and social contexts, but theorists disagree about the relative importance of these influences. Building on Darwin's work (1872/1998), basic emotion theorists suggest that for a limited number of basic emotions (sometimes described as 'affect programs'), the entire package of emotion components has evolved jointly to meet the fundamental needs of survival that occur throughout evolution. Therefore, basic emotions are thought to be universally shared by humans and, to varying degrees, by other species such as the great apes. Some basic emotion researchers suggest that throughout ontogenetic development, basic emotions become influenced by cultural pressures through stimulus selection and display rules (e.g., Ekman, 1992); others suggest that when basic emotions are influenced by culture, they give rise to complex emotion schemas (e.g., Izard, 2007). Social influences on appraisal tendencies of events are less emphasized but also considered by basic emotion theorists (e.g., Ekman, 1992).

Unlike basic emotion theories, appraisal theories suggest that evolutionary pressures shape in particular the links between appraisals and other emotion components. Across cultures, the same appraisals lead to the same emotions. Social contexts influence the triggering stimuli, the appraisal tendencies in a situation, and the evaluation and communication of emotions.

In psychological constructionist theories, only core affect has evolved. Constructionist theories propose that there are no natural relations between situation types and emotions (as in basic emotion theories), or between appraisals and other emotion components (as in appraisal theories). Social pressures determine conceptual knowledge about emotions, which define the boundaries of an emotion. Systematic relations among emotion components can be attributed to shared conceptual knowledge among individuals.

Nonlinear dynamic systems theories suggest that the capacity to reach systematic patterns of emotion component activities and the sensitivity of such patterns to the context of the organism results from evolution. The patterns are neither arbitrary (as in psychological constructionism) nor organized by an affect program (as in basic emotion theory) or by a particular emotion component (as in appraisal theory). Instead, they are self-organized. Furthermore, the contexts are not general themes (such as threat in basic emotion theory), but refer more subtly to the organism's previous state in combination with the relation of the organism to its current environment. The perception of dynamics and their contingencies, as well as the ability to integrate such information in self-organized attractor states, has evolved (Fogel et al., 1992). Social pressures in interaction with ontogenetic maturation lead to the differentiation and integration of the emotion components to specific attractor states (Camras, 2011). The interaction between a context and maturation may be similar across cultures and results in the appearance of similar emotion attractor states.

To summarize, basic emotion theories propose that basic emotions have evolved, appraisal theories suggest that links between appraisals and other emotion components are universally shared, psychological constructionist theories argue that only subjective feeling (core affect) results from evolution, and nonlinear dynamic systems theories propose that evolution led to sensitivities for dynamics and contingencies.

How Should the Emotion Space Be Described?

A fourth controversy concerns the characterization of the relation among emotions. Basic emotion theories propose that emotions are discrete events. These theories identified between 3 and 11 discrete basic emotions (e.g., joy, sadness, and anger). Other emotions are seen as variants within an emotion family (e.g., Ekman, 1992), as mixtures of the more basic emotions (e.g., Plutchik, 2001), or as other types of emotions (e.g., emotion schemas; Izard, 2007). Systematic relations between discrete emotions have been proposed (e.g., Plutchik, 2001).

Appraisal theorists suggest that emotions are structured on the basis of an underlying appraisal space. Appraisals are commonly viewed as continuous (but see Roseman et al., 1996) and contain relevance appraisals and additional appraisals such as agency and coping potential. For example, the component process model (e.g., Scherer, 2009) has 14 appraisal dimensions. The resulting appraisal space contains an unlimited number of emotions, but researchers often focus on 'modal' patterns of emotion components that occur with greater frequency in human interaction and are labeled with commonly known emotion terms (e.g., happiness, anger). Given the large appraisal space, emotions are not typically represented in terms of appraisals only.

Similar to appraisal theorists, psychological constructionist theorists propose the existence of affective states that do not conform to the common component patterns described in natural language; however, the underlying dimensions are seen in core affect and not in appraisals. Focusing on the valence and arousal dimension, psychological constructionists build on dimensional models that date back to Wundt (1897).

However, the valence and arousal model is not applicable to the assessment of mixed emotions. As an alternative to describing emotions on a bipolar valence dimension (i.e., positive–negative), others have therefore proposed a bivariate valence space with independent positive and negative dimensions (e.g., Larsen and McGraw, 2011).

Similar to appraisal theories, nonlinear dynamic systems theories suggest that an unlimited number of emotion component patterns are possible and that some are more likely to occur than others (these are the so-called attractor states; Camras, 2011). Furthermore, control parameters determine whether the relation between input and output is linear or not. As a result, the relation between emotions may be linear at some levels of the control parameter, but show abrupt phase shifts at other levels of the control parameter. The structure of the emotion space is therefore described in relation to control parameters.

To summarize, basic emotion theories characterize the emotion space by discrete emotions, appraisal theories propose

underlying appraisal dimensions, psychological constructionist theories describe underlying valence and arousal dimensions, alternative dimensional theories propose independent positive and negative valence dimensions, and nonlinear dynamic systems theories characterize the emotion space in relation to specific control parameters.

Complementarities of Different Perspectives

To some extent, the alternative views about emotions proposed by different schools of thought can be reconciled. Furthermore, it can be argued that in focusing on different research agendas, the competing emotion theories have led to distinct insights that result in a broader understanding of the complex mechanisms underlying emotional reactions.

The Emotion Space as a Coin with Two Sides

With regard to the emotion space, in particular, the discrete and dimensional approaches taken by basic emotion theorists, appraisal theorists, and psychological constructionist theorists do not necessarily contradict each other (Izard, 2007; Scherer, 2005). For example, discrete emotions can be projected onto a dimensional model, such as Russell's circumplex model (see Russell, 2003). Indeed, researchers often aggregate discrete emotions on the basis of superordinate dimensions, factors, or clusters. Based on a new semantic profile analysis of emotion words (GRID; Scherer, 2005), Fontaine et al. (2013) found that the emotion domain can be represented from the perspectives of both emotion components and emotion dimensions and furthermore, that each of these perspectives lead to unique insights. For example, the dimensions underlying the subjective feeling component most closely resemble the dimensions underlying all emotion components jointly (i.e., valence, power, arousal, novelty; in that order of importance), whereas the dimensions underlying the physiological component are rather different. Based on this study, a self-report measure of emotions that reflects a combined discrete and dimensional approach has been refined (Geneva Emotion Wheel; Scherer et al., 2013).

Complementary Bodies of Knowledge about Emotions

The bodies of knowledge that have been generated by proponents of different emotion theories add specific insights to our understanding of emotions. Basic emotion researchers focus on the adaptive function of an emotion and examine continua across phylogenetic and ontogenetic development. In much of this work, scientists examine emotion expression and perception. An example of the research knowledge gained from this focus is that enacted expressions of happiness, anger, sadness, fear, disgust, and surprise are similarly perceived across cultures (e.g., Ekman, 1992). Also, differences between natural and simulated emotion expressions (e.g., in deception) have been identified.

Appraisal theories focus on the process that leads to an emotion. Researchers have converged on a short list of the most important appraisals, namely, appraisals of novelty, valence, certainty, goal significance, agency, coping potential, and compatibility with norms (Ellsworth and Scherer, 2003). Further, studies have shown support for many of the proposed links between appraisals and subjective feelings, action tendencies, motor expressions, and physiological activities (e.g., Scherer, 2009). Underlying appraisals have proven useful to describe cross-cultural differences in emotion experience (Scherer, 1997).

Psychological constructionist theories focus on the construal of an emotion category. Scientists working within this research tradition have identified context factors in emotion categorization such as salient emotion terms and previously seen emotion expressions. They have sensitized researchers to the influence of response formats in the study of emotions (e.g., Russell, 2003). They also gathered evidence to further establish that the valence dimension explains the most variance in the classifications of affective stimuli and experiences (e.g., see review in Barrett, 2006).

Nonlinear dynamic systems theories focus on the context-dependent temporal dynamics within emotional states and on the dynamics between emotional and nonemotional states. Researchers have found that throughout ontology, emotional components are increasingly differentiated and integrated to attractor states in relation to particular contexts (e.g., Camras, 2011; Fogel et al., 1992). Studies have been conducted to identify the temporal dynamic of the attractor state for specific emotions (e.g., Hoeksma et al., 2007), and researchers have examined phase shifts between attractor states (e.g., in emotion perception; Sacharin et al., 2012).

Future Directions

The four questions posed above (Sections What Are the Respective Roles of and the Interrelations between Emotion Components?, How Are Emotional Episodes Distinguished from Nonemotional Episodes?, To What Extent Are Emotions Influenced by Evolution and Sociocultural Contexts? and How Should the Emotion Space Be Described?) have been extensively discussed by emotion researchers in the past. These questions concern the very nature of emotion and they remain central to the field. Future research will therefore continue to address, and disagree, about these issues. There is a tendency toward more integrative approaches that acknowledge the need to take into consideration many different aspects of the emotion process. In particular, the need to examine the appraisal of emotion-eliciting stimuli and situations is now widely recognized, although there is continuing debate about the causal role of appraisals in terms of changing other components. Similarly, few emotion researchers would deny that the conceptual knowledge of emotion categories plays a major role in verbal labeling, and that this process is strongly influenced by cultural and linguistic factors. However, because emotion processes in daily life are only rarely clearly categorized and labeled with an emotion term, many emotion researchers question the utility of emotion 'construction' based on purely conceptual grounds.

A more recent development in emotion research is an increased research focus on previously understudied types of emotions such as aesthetic emotions (e.g., rapture during listening to music), epistemic emotions (e.g., wonderment

about a scientific breakthrough), and different types of positive emotions (e.g., gratitude, love, pride). These types of emotions differ from the more commonly studied set of negative emotions (anger, fear, disgust, sadness) in that associated action tendencies – other than maintaining attention and approach tendencies – cannot be easily recognized. Researchers from the growing field of positive psychology have identified the utilitarian value of positive emotions by studying their long-term outcomes, for example, for health and life satisfaction (Fredrickson, 1998). A convincing functional account of aesthetic emotions has yet to be proposed.

Additionally, researchers are increasingly paying attention to the social functions of emotion. Thus, in addition to studying emotions at the individual level, emotion scientists are developing the tools needed to study emotions in complex social contexts. This includes research on the influence of culture and social class on emotional experiences, on emotions in social media, and on emotions in interpersonal interaction (e.g., von Scheve, 2013).

Furthermore, neuroscience has a strong influence on emotion research and led to the institutionalization of the affective neurosciences (Panksepp, 1998). Many affective neuroscientists have started to abandon the quest for emotion-specific brain structures and increasingly examine emotion processes in the brain by identifying the circuitry underlying dynamic processes. Affective neuroscience provides new input for emotion research regarding the emotion mechanism as it provides new tools for the precise measurement, for example, with respect to the sequence and timing of appraisal processes (Gentsch et al., 2013; van Peer et al., 2014).

The neuroscience turn in emotion research is part of a general trend toward an increased reliance on interdisciplinary approaches to examining emotions. The new field of affective sciences (see Davidson et al., 2003; Sander and Scherer, 2009) brings together psychologists, biologists, clinicians, neuroscientists, philosophers, linguists, lawyers, historians, and economists. Very recently, the affective sciences have been extended to include engineers and computer scientists studying affective computing, which includes computer recognition of emotion and modeling of artificial emotions through avatars (see Scherer et al., 2010). This development allows researchers to study emotions in novel ways that have hitherto been impossible (e.g., due to ethical concerns). The field of affective sciences provides, for the first time in the history of emotion research, the theoretical constructs, and the methodological tools required for the comprehensive study of the phenomenon.

See also: Affect and Emotion, Anthropology of; Culture and Emotion; Emotion and Expression; Emotion, Evolution of; Emotion, Perception and Expression of; Emotion: History of the Concept; Emotions, Sociology of.

Bibliography

Barrett, L.F., 2006. Solving the emotion paradox: categorization and the experience of emotion. Personality and Social Psychology Review 10, 20–46.

Camras, L.A., 2011. Differentiation, dynamical integration and functional emotional development. Emotion Review 3, 138–146.

Darwin, C., 1998. In: Ekman, P. (Ed.), The Expression of Emotions in Man and Animals. HarperCollins, London (Original work published 1872).

Davidson, R., Scherer, K.R., Goldsmith, H. (Eds.), 2003. Handbook of Affective Sciences. Oxford University Press, New York and Oxford.

De Leersnyder, J., Boiger, M., Mesquita, B., 2013. Cultural regulation of emotion: individual, relational, and structural sources. Frontiers in Psychology 5.

Ekman, P., 1992. An argument for basic emotions. Cognition and Emotion 6, 169–200.

Ellsworth, P.C., Scherer, K.R., 2003. Appraisal processes in emotion. In: Davidson, R.J., Goldsmith, H.H., Scherer, K.R. (Eds.), Handbook of Affective Sciences. Oxford University Press, New York, pp. 572–595.

Fogel, A., Nwokah, E., Dedo, J.Y., Messinger, D., Dickson, K.L., Matusov, E., Holt, S.A., 1992. Social process theory of emotion: a dynamic systems approach. Social Development 1, 122–142.

Fontaine, J.R.J., Scherer, K.R., Soriano, C., 2013. Components of Emotional Meaning: A Sourcebook. Oxford University Press, Oxford.

Fredrickson, B.L., 1998. What good are positive emotions? Review of General Psychology 2, 300–319.

Frijda, N.H., 2007. The Laws of Emotion. Lawrence Erlbaum, Mahwah, NJ.

Gentsch, K., Grandjean, D., Scherer, K.R., 2013. Temporal dynamics of event-related potentials related to goal conduciveness and power appraisals. Psychophysiology 50 (10), 1010–1022.

Hoeksma, J.B., Oosterlaan, J., Schipper, E., Koot, H., 2007. Finding the attractor of anger: Bridging the gap between dynamic concepts and empirical data. Emotion 7, 638–648.

Izard, C.E., 2007. Basic emotions, natural kinds, emotion schemas, and a new paradigm. Perspectives on Psychological Science 2, 260–280.

James, W., 1884. What is an emotion? Mind 9, 188–205.

Keltner, D., Haidt, J., 1999. Social functions of emotions at four levels of analysis. Cognition and Emotion 13, 505–521.

Kreibig, S.D., Schaefer, G., Brosch, T., 2010. Psychophysiological response patterning in emotion: implications for affective computing. In: Scherer, K.R., Bänziger, T., Roesch, E. (Eds.), Blueprint for Affective Computing: A Sourcebook. Oxford University Press, Oxford, pp. 105–130.

Larsen, J.T., McGraw, A.P., 2011. Further evidence for mixed emotions. Journal of Personality and Social Psychology 100, 1095–1110.

Lazarus, R.S., 1968. Emotions and adaptation: conceptual and empirical relations. In: Nebraska Symposium on Motivation, vol. 16. University of Nebraska Press, Lincoln, NE, pp. 175–270.

Leventhal, H., Scherer, K.R., 1987. The relationship of emotion to cognition: a functional approach to a semantic controversy. Cognition and Emotion 1, 3–28.

Lewis, M.D., 2005. Bridging emotion theory and neurobiology through dynamic systems modeling. Behavioral and Brain Sciences 28, 169–194.

Moors, A., 2009. Theories of emotion causation: a review. Cognition and Emotion 23, 625–662.

Mulligan, K., Scherer, K.R., 2012. Toward a working definition of emotion. Emotion Review 4, 345–357.

Ortony, A., Clore, G.L., Collins, A., 1990. The Cognitive Structure of Emotions. Cambridge University Press, Cambridge.

Panksepp, J., 1998. Affective Neuroscience. Oxford University Press, New York.

Plutchik, R., 2001. The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American Scientist 89, 344–350.

van Peer, J.M., Grandjean, D., Scherer, K.R., 2014. Sequential unfolding of appraisals: EEG evidence for the interaction of novelty and pleasantness. Emotion 14 (1), 51–63.

Roseman, I.J., Antoniou, A.A., Jose, P.E., 1996. Appraisal determinants of emotions: constructing a more accurate and comprehensive theory. Cognition and Emotion 10, 241–278.

Russell, J.A., 2003. Core affect and the psychological construction of emotion. Psychological Review 110, 145–172.

Sacharin, V., Sander, D., Scherer, K.R., 2012. The perception of changing emotion expressions. Cognition and Emotion 26, 1273–1300.

Sander, D., Scherer, K.R. (Eds.), 2009. The Oxford companion to emotion and the affective sciences. Oxford University Press, Oxford, UK.

Sander, D., Grandjean, D., Scherer, K.R., 2005. A systems approach to appraisal mechanisms in emotion. Neural Networks 18, 317–352.

Schachter, S., Singer, J.E., 1962. Cognitive, social, and physiological determinants of emotional state. Psychological Review 69, 379–399.

Scherer, K.R., 1997. The role of culture in emotion-antecedent appraisal. Journal of Personality and Social Psychology 73, 902–922.

Scherer, K.R., 2005. What are emotions? And how can they be measured? Social Science Information 44, 695–729.

Scherer, K.R., 2009. The dynamic architecture of emotion: evidence for the component process model. Emotion 23, 1307–1351.

Scherer, K.R., Bänziger, T., Roesch, E.B. (Eds.), 2010. Blueprint for Affective Computing: A Sourcebook. Oxford University Press, Oxford.

Scherer, K.R., Shuman, V., Fontaine, J.R.J., Soriano, C., 2013. The GRID meets the wheel: assessing emotional feeling via self-report. In: Fontaine, J.R.J., Scherer, K.R., Soriano, C. (Eds.), Components of Emotional Meaning: A Sourcebook. Oxford University Press, Oxford, pp. 281–298.

von Scheve, C., 2012. Emotion regulation and emotion work: two sides of the same coin? Frontiers in Psychology 3, 1–10.

Shuman, V., Scherer, K.R., 2014. Concepts and structures of emotions. In: Pekrun, R., Linnenbrink-Garcia, L. (Eds.), International Handbook of Emotions in Education. Routledge, London, pp. 13–35.

Spielberger, C.D., Reheiser, E.C., 2009. Assessment of emotions: anxiety, anger, depression, and curiosity. Applied Psychology: Health and Well-Being 1 (3), 271–302.

Susskind, J.M., Lee, D.H., Cusi, A., Feiman, R., Grabski, W., Anderson, A.K., 2008. Expressing fear enhances sensory acquisition. Nature Neuroscience 11, 843–850.

Verduyn, P., Delvaux, E., Van Coillie, H., Tuerlinckx, F., Van Mechelen, I., 2009. Predicting the duration of emotional experience: two experience sampling studies. Emotion 9, 83–91.

Wundt, W., 1897. Outlines of Psychology. Stechert, New York.

Emotions, Sociology of

Kathryn J Lively, Dartmouth College, Hanover, NH, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D.D. Franks, volume 7, pp. 4477–4481, © 2001, Elsevier Ltd.

Abstract

Since its introduction some 30 years ago, sociological interest in the study of emotion has flourished. Today, it is no longer necessary to defend the role of emotion in social life. Indeed, emotions and their consequent processes have become recognized not only as relevant, but also as crucial to many – if not all – levels of society. Sociological scholarship on emotion includes a range of topics, including, but certainly not limited to, the role that feelings play in face-to-face interactions, mental health, identity maintenance, group processes, the creation and reification of stratification hierarchies, social change, social movements, and time.

Introduction

Prior to the formal development of the sociology of emotion, emotion had, historically, been left to the purview of psychologists. With few exceptions, psychological scholarship on emotion has focused more closely on cognitive and neurological processes as predictors of emotional experience than it has on the social environments in which individuals are embedded. In particular, psychologists have concerned themselves with how emotions relate to cognition and personality and, to a lesser degree, social and physical health, without taking into consideration the degree to which every aspect of emotion – from its initial constitution to its consequence – are profoundly social (see Lewis et al., 2007 for a recent review of the field). Over the last three decades, however, sociological scholarship on emotion has begun to fill this gap. That is, sociological observations reveal that feelings are not only social constructs, but also socially patterned and play a vital role in social life.

At their most basic, sociological definitions of emotion reveal the assumption that emotions are inherently social. Hochschild (1983), for one, likens emotions to senses that signal what is personally relevant about surrounding events. Similarly, Lively and Heise (2014) have defined emotion as responses to events that are linked with corporeal manifestations. Consistent with both these definitions, is Thoits' (1990) four-factor model of emotion. Building off the classic two-factor model of emotion first developed by Schachter and Singer (1962), which included physiology and cognition, Thoits' analyses suggest that emotional experience also includes expression (Ekman, 1973) and emotion words (Ritchie and Barker, 2006). Notably, each of these four factors – that is, physiology, cognition, expression, and labeling – are interdependent and most sociologists assume that a change in one may lead to a change in the others, suggesting multiple pathways for emotion management.

Origins of the Sociology of Emotions: 1975–85

Particular emotions were, indeed, important to classical sociologists (Marx and alienation, Weber and disenchantment, Durkheim and anomie, etc.). Their treatment, however, was not concerned with the light they shed on the broader social–emotional process in general (Smith-Lovin, 1995).

It was not until 1975 that American sociologists called for a study of emotion as a social process in its own right. In that year, Arlie Hochschild (1975) argued that a purely cognitive view of human life was not sufficient for sentient humans. Her insistence that emotion was socially patterned was partly explained by her concept of feeling rules, i.e., culturally and historically variable norms defining what should be felt in various situations. Her conceptualization of 'emotion work' described the individual processes that maintained these norms, which occurs when actors experience a discrepancy between what they actually feel and what they should feel. Techniques used to change emotions include surface acting to mask or enhance the expression of emotion and deep acting wherein the actual emotion is changed.

Although Hochschild focused primarily on the cognitive strategies that individuals used to accomplish deep acting, later studies suggest that emotion-management specialists (that is therapists (Francis, 1997), psychodrama leaders (Thoits, 1995), and social movement activists (Britt and Heise, 2000)) also use interpersonal emotion strategies to manage the emotions of others directly. Many of these interpersonal strategies target one or more of the four factors in Thoits' model of emotion – as detailed above – simultaneously (also see Simon and Nath, 2004).

Taking a more positivist approach to the study of emotion, Kemper's (1978) social interactional theory argued that two fundamental dimensions – power and status –determine all emotions across a variety of social contexts. Drawing on Weber's classic definitions, 'social power' refers to one's ability to force another's compliance against their will, whereas 'social status' refers to favors freely given. When individuals experience the levels of status or power they expect during the course of a social interaction, they are predicted to feel content or happy. Experiences of excesses or deficits along these dimensions, however, automatically produce certain distressful emotions depending on whether one attributes the cause to one's self or others as well as the relationship between actors. For instance, if an individual loses status, as a result of his or her incompetence, embarrassment is expected. However, if he or she loses

status as a result of another's actions, anger is the expected response. These predictions are conditioned further depending upon the degree that the other is well liked.

At the same time that Kemper was developing his somewhat deterministic model, Shott (1979) introduced an interactionist view of emotion that stressed the importance of emotion in social control. According to the interactionist perspective, role taking is the capacity to see oneself as do others and to use their anticipated response as a guide to further conduct. Although Mead's (1967) discussion of role taking was almost entirely cognitive (MacKinnon, 1994), Shott, like Hochschild before her, argued that cognition, alone, was not enough to motivate and to sustain normative social control. However, role-taking emotions, such as embarrassment, shame, guilt, as well as other directed empathetic emotions of sympathy and pride could. She suggests that individuals' desire to avoid negative emotions (including shame and embarrassment) and to experience positive ones (such as pride and empathy) encourages them to engage in self-control, which translates, writ large, into social control.

By 1985, the sociology of emotions was formally recognized as a subsection in the American Sociological Association. Since that time, there have been numerous reviews of the field (Thoits, 1989; Smith-Lovin, 1995; Turner and Stets, 2006), including two full-length handbooks (Stets and Turner, 2014) and a series of volumes dedicated to social perspectives on emotion (e.g., Franks and Gecas, 1992). During this same period, several social psychological theories within sociology were also expanded to make considerations of emotion more central (Smith-Lovin, 1990; Ridgeway and Johnson, 1990; Lively and Heise, 2014; Stets, 2005; Lawler and Thye, 1999; Lively et al., 2010; Hegdvedt and Parris, 2014; among others).

Cultural Shaping of Emotion

Today there is little debate that certain cultures encourage and discourage certain emotional experiences, which lead to important behavioral dispositions. In her seminal work on sympathy in everyday life, Clark (1997) begins by describing the Ik, a tribal community known for their marked lack of sympathy. Given their propensity to laugh when older group members fall down or to simply watch when an unattended child walks toward an open fire, the Ik provide a stark contrast to most Western cultures, whose members are, for the most part, sympathetic toward the elderly and children or others whose misfortunes seem largely beyond their control.

Clark's analysis of the US emotion culture details a sympathy economy, under which individuals are automatically granted a certain amount of sympathy (known as sympathy margins) that are, in large part, a function of their age, gender, social class, etc. As individuals move through daily life, they draw upon – and invest in – these accounts by asking for and providing sympathy to others. Clark identifies four sympathy rules to which individuals are expected to adhere in order to avoid becoming overdrawn; these include not making undue claims to sympathy, not asking for too much sympathy, asking for some sympathy, and reciprocating once sympathy has been received, either by providing sympathy to your sympathizer when necessary, being sympathetic to others, or simply expressing gratitude.

Historical accounts of various emotions also illustrate not only how feelings, and expectations regarding them are culturally patterned, but also how they change over time (Stearns and Stearns, 1989; Stearns and Lewis, 1998). Content analyses of self-help books and women's magazines, for example, reveal that the emotion norms – particularly those surrounding love and anger – have changed considerably for wives over the last several decades (Cancian and Gordon, 1988). Additionally, love in the United States has grown increasingly feminized; despite heterosexual couples' move toward more egalitarian/androgynous gender roles (Cancian, 1987; Gerson, 2010).

Other examples of the cultural shaping of emotion include the literature on gender and emotion, leading some scholars to argue that there are distinct emotion cultures for women and men (Simon and Nath, 2004). Important here is the devaluation of emotion (as irrational or unprofessional, for example) and the belief that women are inherently more emotional than men. Such beliefs routinely disadvantage women in the workplace or cause their labor at home to remain largely invisible. Recent survey research suggests that there are actually very few differences in the self-reported emotional experience of women and men once structural considerations are accounted for; however, differences in self-reported expressions of emotion persist (Simon and Nath, 2004).

Scholars have also turned their attention to how race and ethnicity shape emotional experience and expression. Qualitative studies of historically white organizations and occupations reveal that African-Americans often encounter different emotion rules than their white peers and that they are keenly aware of the necessity of avoiding racial stereotypes, particularly those involving the experience and expression of anger (Harlow, 2003; Wingfield, 2010).

Studies of subcultures are also instructive. Individuals participating in a polyamorous community, for example, have created their own language as a way to dispel disruptive feelings and to create a deeper sense of community. Indeed, in a subculture where jealously is looked down upon, members are encouraged to use less emotionally charged words to describe their feelings, such as 'wibbly.' Similarly, polyamorists also have words to describe the love they feel for their partners' lover, another concept that has no real meaning in societies where monogamy is the hegemonic ideal. Similarly, fans of jam bands, such as the Grateful Dead and Phish have developed different emotional cultures, particularly regarding such social identities as narcotics agents and law enforcement officials (Hunt, 2008, 2013).

Structural Shaping of Emotion

Emotional experience, as well as its expression, is also shaped by structure or, to put it a different way, status. The literature on gender and emotion and race and emotion are useful here as well.

In her now classic work, *The Managed Heart: The Commercialization of Human Feeling*, Hochschild first introduced the

notion of 'status shields,' i.e., shared interpretive dispositions that protect those of privilege against (1) the displaced feelings of others, (2) the negative interpretation of one's own emotional expressions, and (3) the conviction of others that one's emotions are unwarranted and unimportant. The undermining of women's anger by interpreting them as 'crabby' because of their menstrual cycle or otherwise 'hysterical,' just as undermining African-American men's anger as a function of being 'an angry Black man,' are obvious examples of a structured lack of status shield. The aggregate of such interpretations serve white male dominance, not only in the workplace, but also at home and in small task-focused groups, such as juries.

Emotions are also shaped by the ways whole societies are organized, as well as the organizations within them. For instance, Hochschild and Machung, 1989 demonstrated how macrolevel changes in women's entry into the workforce affected microlevel gender imbalances in feelings of gratitude between couples in the household. Because of traditional gender roles, men's relatively small contributions to housework evoked more gratitude than the larger contributions from their working wives. This resulted in a 'second shift' for women and interpersonal tension within the couple. Drawing on her previous ideas of emotion work, Hochschild argued that husbands and wives cocreated elaborate family myths as a way to bridge the gap between their own gender ideologies (that is, traditional, transitional, and egalitarian) and the household arrangements within their marriages. These 'myths' tended to obscure ongoing inequities, as well as challenges to individuals' identities as feminist, egalitarian, or, in some cases, breadwinners. Scholarship in this area is likely to prove fruitful going forward, as many women – particularly those in the working class families – now outperform their male partners in terms of education, labor market participation, and wages (Rosin, 2012).

Although emotions have historically been viewed as antithetical to rational behavior and professionalism, businesses and organizations have become increasingly adept – unwittingly as well as consciously – at organizing emotional processes. While most accept that feelings can impede organizational efficiency, it is also true that they are necessary for organizations to operate effectively. Any workplace must encourage loyalty and manage disenchantment, stirrings about power, status, and justice; anxieties about keeping control and being controlled; pride; defensiveness; fear of change; and the comforts of the familiar.

Emotional labor, according to Hochschild (1983), is the commercial exploitation of the principle of emotion work (see above). The flight attendants she studied were paid to 'be nice' regardless of insults and paradoxically asked to be genuine. Bill collectors, on the other hand, were taught to be angry, mean, and impatient, in order to bring about fear and compliance in debtors. Unlike 'emotion work,' which typically occurs in individuals' private lives (such as trying to feel sad when one has adopted the role of mourner or trying to feel happy when one has adopted the role of bride) and has a 'use value,' 'emotional labor' is sold for a wage and typically benefits the corporation by whom the actor is employed or the customers, clients, or patients with whom the actor is engaged. Since its introduction, there have been numerous studies attempting to determine the socioemotional effects of providing emotional labor (Wharton, 2009).

The emotional labor necessary of students (Cahill, 1999; Smith and Kleinman, 1989), fast food workers and insurance salesmen (Leidner, 1993), amusement park employees (Van Maanen and Kunda, 1989), male and female rescue workers (Lois, 2003), Korean nail salon owners (Kang, 2010), professors (Bellas, 1999; Harlow, 2003), mixed martial arts fighters (Vaccaro et al., 2011), home health care aids (Stacey, 2011), health care providers (Erickson and Grove, 2008), direct home sales consultants (Mullaney and Shope, 2012), homeschooling mothers (Lois, 2013), and Black college men (Jackson, 2013), among others, have all proved of interest to sociologists interested in the nuances, as well as the consequences of, emotional labor.

Emotional Deviance and the Emotions of Deviants

In addition to studying how culture and social structure shapes emotional experience and behavior, sociologists are also interested in the ways in which individuals – and their emotions – deviate from cultural and structural expectations. To date, deviance – or the inability or unwillingness to adhere to societal expectations – has been dealt with in two ways. The first refers to emotional deviance (Thoits, 1989); the second refers to the emotions of those individuals who hold deviant, or stigmatized, identities (Lively and Heise, 2014).

In an attempt to better understand the relationship between emotion and mental health problems, Thoits (1990) coined the term 'emotional deviance.' Building on Becker's (1963) labeling theory of deviance, Thoits argues that individuals are often labeled – by self and others – as mentally ill, as a result of their inability to bring their emotions in line with preexisting emotion rules or norms. According to this perspective, sources of discrepant feelings include multiple roles, role transitions, and unrealistic role demands. Individuals who self-label as emotionally deviant are increasingly likely to seek professional help, despite the historical stigma associated with mental illness. In this regard, Thoits' analysis drew attention not only to the emotional dimensions of coping behaviors and social support but also the effect of the deinstitutionalization of mental health care on individuals' willingness to voluntarily seek help for emotional issues. Since its introduction in the late 1980s, the concept of emotional deviance has been applied to numerous populations, most often to mothers (Taylor, 1996; Lois, 2013).

More recently, scholars have also begun examining the emotions of deviants – that is, individuals who hold socially devalued or otherwise stigmatized social identities, such as rapist or murderer. Based on computational models, Heise found that the experience of positive emotions do not benefit those with stigmatized identities to the same degree as they do those with socially valued identity. For example, a happy mother is viewed as much nicer than a mother who experiences or expresses no affect, whereas a happy rapist is viewed only as slightly less negative than a rapist who appears affectively neutral (Lively and Heise, 2014). Further analysis suggests that extremely negative role identities, such as rapist, mugger, child molester, etc., are also more emotionally volatile than those that are less negative. Future scholarship – perhaps

qualitative or ethnographic in nature – may be useful in better understanding the emotional lives of society's most stigmatized members.

Identity, Emotion, and Social Structure

According to sociologists, one's sense of self, or one's identity, is connected to social structures and institutions (MacKinnon and Heise, 2010; also see Stryker and Serpe, 1994, 1992; Burke, 1991). This is, in part, because identities are embedded in social roles. According to most theories of identity, emotions help determine what roles are chosen as central to how one sees oneself. Emotions affect role performance (e.g., confidence may propel us forward, whereas fear may cause us to choke). Emotions also arise from role performances (e.g., pride when things go well or shame when they do not).

The emotional aspects of identity work are, in effect, commitments to role performances. That is, when actors conform to the emotional and behavioral expectations of social roles they are not only managing their sense of self, but they are also reifying the institutions and organizations in which these roles are embedded. For example, when individuals enact the emotional and behavioral expectations of a physician, they are also reifying the institution of medicine. They are also perpetuating the relative status positions of physicians and patients and physicians and other health care providers.

According to affect control theory (ACT), individuals create events that confirm their fundamental affective sentiments regarding themselves, others, and settings (Heise, 2007). Moreover, when they find themselves in situations that disconfirm these sentiments, they feel deflection, which is often experienced as emotion. Fundamental sentiments capture evaluation (how good or bad something is), potency (how powerful or weak something is), and activation (how lively or quiescent something is); these dimensions of affective meaning have been documented across numerous cultures and, within a single culture, represent shared meaning regarding identities, behaviors, settings, attributes, and emotion Heise 2007). Drawing on a cybernetic model, ACT predicts that when transient sentiments contradict fundamental sentiments, individuals are motivated to alter their behavior in order to minimize deflection and reduce the likelihood of reidentification of self, other, or both.

Much of the theoretical work within the sociology of emotion has been based on qualitative methods, such as ethnography and in-depth interviews. In contrast, ACT incorporates a unique computer program – Interact – that uses mathematical models in order to make probabilistic predictions regarding behaviors, emotions, and labeling, many of which have been replicated using ethnography, survey data, and experiments (Francis, 1997; Lively and Powell, 2006; Robinson et al., 1994).

Although ACT was initially developed as a theory of identity, where emotions were essentially a signal of the degree to which one's identity was being confirmed or disconfirmed during a course of daily interaction, emotions have become increasingly more central. Lively and Heise (2014) recently summarized the contributions of ACT to the sociological literature on emotion and have suggested a number of new avenues of ACT-based scholarship in emotion, including emotional segueing (Lively and Heise, 2004; Lively, 2008), emotional stations (Lively and Heise, 2004; MacKinnon and Heise, 2010), and the cross-cultural absence of emotions that can be viewed of as simultaneously positive, weak, and active. Rogers et al. (2014) have suggested that ACT may provide long-overdue launching point for cross-disciplinary collaboration on emotion because of its empirically grounded conceptualization of social mechanisms operating at the level of interaction, relationship, and culture, as well as its specification of processes linking social and individual aspects of emotion.

Identity control theory, as articulated by Burke and Stets (Burke, 1991; Stets, 2005), also illustrates that emotions are implicated when individual's identities are disconfirmed. Unlike ACT, which calculates deflection using culturally shared affective meaning of social roles, behaviors, and settings, identity control theory is focused primarily on the correspondence between individuals' self-conceptions and others' reactions to their identity performances.

While both identity control theory and ACT assume that emotions result from identity processes, there is some disagreement as to what forms these emotions will actually take. Regardless, identity control theory also predicts that individuals are motivated to bring others' reactions to their identities in line with their own. Burke and Harrod's (2005) analysis of identity processes among intimate couples, for example, reveals that individuals within committed relationships are more content when their partner's evaluations of them match their own, compared to individuals' at the beginning stages of relationships, who would rather their partners' see them as slightly better than they see themselves.

Emotion and Equity, Justice, and Exchange

Other social psychological theories – including equity theory (Walster et al., 1978), justice theory (Hegdvedt and Parris, 2014), and exchange theory (Lawler, 2001; Lawler and Thye, 1999) – also detail the coconstitutive relationship between emotion and social structure.

Equity theory, for instance, predicts that individuals are also likely to experience negative emotion when they perceive an inequity in their interactions with others. Originally, the theory predicted that individuals would experience ungeneralized distress, whether they were underbenefited or overbenefited. However, later scholars specified that anger was more likely in the former case, whereas shame was more likely experienced in the latter.

Most of the early scholarship on equity and discrete everyday emotion was based on experiments, based on college student samples. Later survey-based studies of married and cohabitating adults, however, found that a broad range of emotions – including positive ones – are affected by perceptions of inequity within the context of long-term romantic relationships. Further, in the case of the household division of labor, men were more likely to experience anger and rage when faced with perceptions of under benefiting (that is, doing more than their fair share), whereas women were more susceptible to feelings of self-reproach and fear when confronted with

perceptions of overbenefiting (that is, doing less than their fair share). These findings suggest that women may willingly do more than their fair share of the household division of labor in order to manage their own – and their male partners' – negative feelings. Needless to say, such endeavors – while the managing unpalatable emotions of both self and other – do little to ameliorate the conditions that arise to perceptions of inequity in the home that led to these emotions to begin with.

The justice framework also has grown to include the study of emotion (see Hegdvedt and Parris, 2014, for a complete review). Justice theorists generally consider two processes: procedural justice and distributive justice. The former focuses on the means by which outcomes are distributed, whereas the latter is more concerned with the actual allocation of outcomes, often taking into account such issues as need, equity, and equality. Generally speaking, people are happy when they feel like procedural decisions are just and the outcomes seem fair. They are less sanguine, however, when they believe the procedures to be unfair and the outcomes – as a result – fail to meet their expectations. Indeed, most people assume that fair procedures will lead to fair outcomes. However, even when individuals experience less than they expected, their negative feelings are somewhat alleviated, if they believe the procedure governing the decision is fair (Hegdvedt and Killian, 1999) or if the decision maker is legitimate (Clay-Warner, 2006).

Finally, exchange theorists, too, have been making advances in conceptualizing the role of emotions in exchange interactions. A framework for a number of more specific theories (e.g., the theory of social commitments (Lawler et al., 2009), relational cohesion theory (Lawler and Yoon, 1996), and the affect theory of social exchange (Lawler, 2001)), exchange theory, at its most basic, suggests that individuals who perceive equity in their social arrangements, who see these arrangements as inherently just, and who believe that their contributions are successful, are not only happier with their outcomes, but also experience more group cohesion and are more affectively committed to participate in ongoing exchanges. Conversely, individuals who feel that their costs outstrip their rewards, their contributions are not valued, or the exchange is not successful are likely to exit the exchange – whether within the context of a friendship, a marriage, or a job, or a small task-focused group – especially in light of better viable alternatives.

Taken together, these social psychological perspectives reveal that emotions are shaped by existing social and structural arrangements. They also illustrate emotions' ability to contribute to – and disrupt – the very conditions that gave rise to them initially. Equity theory, justice theory, and exchange theory all rely on the assumption that it is the perception of equity, fairness, or net gains over the course of a single – or ongoing – interaction that affect emotions, rather than the actual, obdurate conditions in which the individuals are actually embedded.

Emotions and Social Change

Emotions have always been recognized as part of social movements (Goodwin et al., 2001). Until the 1970s, emotions were viewed as essential to movements but were, nonetheless, viewed pejoratively – that is, as something that transforms protestors into angry, irrational mobs. Then, from the 1970s to the 1990s, scholars routinely overlooked the emotions of protesters – mainly out of fear that documenting strong emotions would paint the protestors as irrational – as opposed to rational – actors. It was not until the 1990s – bolstered by the rise of cognitive-appraisal theories – that emotion began to reappear as a normal part of political action (see Jasper, 2014).

Since making emotions more central, social movement scholars have identified several emotional dynamics that help explain social and political protest. These include acts of persuasion, motivations for action, the solidarities of groups, the impact of morality on political action, and the gendered division of labor in social movement organizations.

In addition to the role that emotions play in social movements (e.g., pride or righteous anger), scholars have also turned their attention to the role that emotion management – particularly interpersonal emotion management – plays in motivating social movement participation. Identity-based social movements face a unique challenge in motivating potential constituents to participate in collective action (Britt and Heise, 2000). Because identity-based movements often involve stigma, and thus shame, constituents may wish to limit their involvement. As such, inspiring individuals to take action can be challenging. According to Britt and Heise, movement activists use identity-based emotion-management strategies that maneuver constituents through fear and anger in order to get them from an affective space largely characterized by shame and into one of pride. When potential constituents feel pride (a relatively positive, powerful, and active emotion), they are more likely to participate in positive, powerful, and active behaviors, such as social protest, marching, chanting, etc. (Also see Taylor, 1996.)

Gould's recent analysis of the rise, success, and eventual decline of ACT UP (AIDS coalition to unleash power) organizations during the beginning of the US acquired immune deficiency syndrome (AIDS) epidemic documents the role of both emotion and emotion management in creating and sustaining social change (2009). According to Gould, the gay and lesbian community actively encouraged members to feel and express anger toward the federal government following a Supreme Court decision that upheld a state ruling deeming homosexual sodomy in the privacy of one's own home illegal. The successful transformation of the groups' collective shock and grief that had arisen during the early years of the AIDS epidemic into anger and rage after the Supreme Court decision resulted in the proliferation of ACT UP organizations across the country, which led to both short-term upheaval and long-term social change. Gould's analysis suggests that in order for emotions and, subsequently, protest to ignite, a certain emotional habitus – that is, a template for what and how to feel – is necessary. Prior to the Supreme Court decision, the gay and lesbian community was lacking such a template that would not only allow, but also encourage, its constituents to feel anger and to engage in sustainable protest.

Emotion and Time

Over the last decade, emotion scholars have become increasingly interested in time. To date, sociologists have utilized three

approaches when thinking about how time influences emotion and emotion management.

The first approach, which deals specifically with emotion, has to do with how time – usually the shortage thereof – gives rise to certain feelings. Family scholars in particular have documented how a marked lack of time leads to an increase in negative affect and decreased familial satisfaction, particularly for women who routinely have less 'leisure' time compared to men (Hochschild and Manchung, 1989).

The second approach, which deals more with the temporality of emotion management, speaks to issues of timing – that is, how emotion management unfolds before, during, and after a triggering event (e.g., participating in a search and rescue mission (Lois, 2003), passing as a member of the opposite sex (Schrock et al., 2009), or engaging in mixed martial arts (Vaccaro et al., 2011)).

The final approach, also dealing with emotion management, reveals how individuals purposefully manipulate their own sense of time in order to manage their emotions. Homeschooling mothers, for example, purposefully focus on the fleeting nature of childhood or focus how much they are going to miss homeschooling once their children are grown so that they can more easily appreciate the present (Lois, 2013). The Home Direct Sales industry, among others, have also gotten very sophisticated at managing time in order to generate enthusiasm and loyalty among their consultants (Mullaney and Shope, 2012).

See also: Alienation, Sociology of; Anomie; Exchange: Social; Identity Movements; Labor, Division of; Norms; Social Identity in Sociology; Social Movements: A Social Psychological Perspective; Social Psychology.

Bibliography

Becker, H., 1963. Outsiders: Studies in the Sociology of Deviance. Free Press, New York.

Bellas, M., 1999. Emotional labor in academia: the case of professors. Annals of the American Academy of Political and Social Science 561, 96–110.

Britt, L., Heise, D.R., 2000. From shame to pride in identity politics. In: Stryker, S., Owens, T.J., White, R.W. (Eds.), Self, Identity, and Social Movements. University of Minnesota Press, Minneapolis, MN.

Burke, P.J., 1991. Identity processes and social stress. American Sociological Review 56, 836–849.

Burke, P.J., Harrod, M.M., 2005. Too much of a good thing? Social Psychology Quarterly 68 (4), 359–374.

Cahill, S.E., 1999. Emotional capital and professional socialization: the case of mortuary science students (and me). Social Psychology Quarterly 62 (2), 101–116.

Cancian, F.M., Gordon, S.L., 1988. Changing emotion norms in marriage: love and anger in U.S. women's magazines since 1900. Gender & Society 2 (3), 308–342.

Cancian, F.M., 1987. Love in American: Gender and Self-Development. Cambridge University Press, Cambridge.

Clark, C., 1997. Misery and Company: Sympathy in Everyday Life. University of Chicago Press, Chicago, IL.

Clay-Warner, J., 2006. Procedural justice and legitimacy: Predicting negative emotional reactions to workplace injustice. Advances in Group Processes 23, 207–227.

Durkheim, E., 1951. Suicide. Free Press, New York.

Lawler, E.J., Yoon, J., 1996. Commitment in exchange relations: test of a theory of relational cohesion. American Sociological Review 61, 89–108.

Ekman, P., 1973. Darwin and Facial Expression. Academic Press, New York.

Erickson, R.J., Grove, W.J.C., 2008. Emotional labor and health care. Sociology Compass 2 (2), 704–733.

Francis, L.E., 1997. Ideology and interpersonal emotion management: redefining identity in two support groups. Social Psychology Quarterly 60 (2), 153–171.

Franks, D., Gecas, V. (Eds.), 1992. Social Perspectives on Emotion, vol. 1. JAI Press, Greenwich, CT.

Gerson, K., 2010. The Unfinished Revolution: How a New Generation is Reshaping Family, Work, and Gender in America. Oxford University Press, New York.

Goodwin, J., Jasper, J.M., Polletta, F. (Eds.), 2001. Passionate Politics: Emotions in Social Movements. University of Chicago Press, Chicago, IL.

Harlow, R., 2003. 'Race doesn't matter, but…': the effect of race on professors' experiences and emotion management in the undergraduate college classroom. Social Psychology Quarterly 66 (4), 348–363.

Hegdvedt, K.A., Killian, C., 1999. Fairness and emotions: reactions to the process and outcomes of negotiations. Social Forces 78, 269–303.

Hegdvedt, K., Parris, C.L., August 2014. Justice theory and emotions. In: Stets, J.E., Turner, J.H. (Eds.), Handbook of Sociology of Emotions, vol. II. Springer, New York.

Heise, D.R., 1979. Understanding Events: Affect and the Construction of Social Action. Cambridge University Press, New York.

Heise, D.R., 2007. Expressive Order: Confirming Sentiments in Social Actions. Springer, New York.

Hochschild, A., 1975. The sociology of feeling and emotion: selected possibilities. Sociological Inquiry 45 (2–3), 280–307.

Hochschild, A.R., 1979. Emotion work, feeling rules, and social structure. American Journal of Sociology 85 (3), 551–575.

Hochschild, A.R., 1983. The Managed Heart: Commercialization of Human Feeling. University of California Press, Berkeley, CA.

Hochschild, A.R., Machung, A., 1989. The Second Shift: Working Parents and the Revolution at Home. Viking, New York.

Hunt, P.M., 2008. From festies to tourrats: examining the relationship between jam band subculture involvement and role meanings. Social Psychology Quarterly 71, 356–378.

Hunt, P.M., 2013. Where the Music Takes You: The Social Psychology of Music Subcultures. Cognella, San Diego, CA.

Jackson, B., 2013. The bonds of brotherhood: emotional and social support among college black men. The ANNALS of the American Academy of Political and Social Science 642 (1), 61–71.

Jasper, J., July 2014. Constructing indignation: anger dynamics in protest movements. Emotion Review.

Kang, M., 2010. The Managed Hand: Race, Gender, and the Body in Beauty Service Work. University of California Press, Berkeley, CA.

Kemper, T.D., 1978. A Social Interactional Theory of Emotions. John Wiley & Sons, New York.

Lawler, E.J., 2001. An affect theory of social exchange. The American Journal of Sociology 107 (2), 321–352.

Lawler, E.J., Thye, S.R., 1999. Bringing emotions into social exchange theory. Annual Review of Sociology 25, 217–244.

Lawler, E.J., Yoon, J., 1996. Commitment in exchange relations: test of a theory of relational cohesion. American Sociological Review 61, 89–108.

Lawler, E.J., Thye, S.R., Yoon, J., 2009. Social Commitments in a Depersonalized World. Russell Sage, New York.

Leidner, R., 1993. Fast Food, Fast Talk. Service Work and the Routinization of Everyday Life. University of California Press, Berkeley, CA.

Lewis, M., Haviland-Jones, J.M., Barrett, L.F. (Eds.), 2007. The Handbook of Emotions, third ed. The Guildford Press, New York.

Lively, K.J., 2008. Emotional segueing and the management of emotion, by women and men. Social Forces 87, 911–936.

Lively, K.J., 2013. Age and its effect on the experience and management of emotion. Advances in Group Processes 30, 231–265.

Lively, K.J., Heise, D.R., 2004. Sociological realms of emotional experience. American Journal of Sociology 109 (5), 1109–1136.

Lively, K.J., Heise, D.R., August 2014. Emotions in affect control theory. In: Stets, J.E., Turner, J.H. (Eds.), Handbook of Sociology of Emotions, vol. II. Springer, New York.

Lively, K.J., Powell, B., 2006. Emotional expression at work and at home: domain, status or individual characteristics? Social Psychology Quarterly 69, 17–38.

Lively, K.J., Steelman, L.C., Powell, B., 2010. Equity, emotion, and household division of labor. Social Psychology Quarterly 73 (4), 358–379.

Lois, J., 2003. Heroic Efforts: The Emotional Culture of Search and Rescue Volunteers. New York University Press, New York.

Lois, J., 2013. Home is Where the School is: The Logic of Home Schooling and the Emotional Labor of Mothering. New York University Press, New York.

MacKinnon, N.J., 1994. Symbolic Interactionism as Affect Control. State University of New York Press, Albany.

MacKinnon, N.J., Heise, D.R., 2010. Self, Identity, and Social Institutions. Palgrave, New York.

Marx, K., 1988. The Economic and Philosophic Manuscripts and the Communist Manifesto. Prometheus Books, New York.
Mead, H., 1967. In: Morris, C.W. (Ed.), Mind, Self, and Society from the Standpoint of a Social Behaviorist. University of Chicago Press, Chicago, IL.
Mullaney, J.L., Shope, J.H., 2012. Paid to Party: Working Time and Emotion in Direct Home Sales. Rutgers University Press, New Brunswick, NJ.
Pierce, J.L., 1995. Gender Trials: Emotional Lives in Contemporary Law Firms. University of California Press, Berkeley, CA.
Ritchie, A., Barker, M., 2006. 'There aren't words for what we do or how we feel so we have to make them up': constructing polyamorous languages in a culture of compulsory monogamy. Sexualities 9 (5), 584–601.
Ridgeway, C.L., Correll, S.J., 2004. Unpacking the gender system: a theoretical perspective on gender beliefs and social relations. Gender & Society 18 (4), 510–531.
Ridgeway, C.L., Johnson, C., 1990. What is the relationship between socio-emotional behavior and status in task groups? American Journal of Sociology 95, 1189–1212.
Robinson, D.T., Smith-Lovin, L., Tsoudis, O., 1994. Heinous crime or unfortunate accident? Effects of remorse on responses to mock criminal confessions. Social Forces 73, 175–190.
Rogers, K.B., Schröder, T., von Scheve, C., April 2014. Dissecting the sociality of emotion: a multi-level approach. Emotion Review 6, 124–133.
Rosin, H., 2012. The End of Men: And the Rise of Women. Riverhead Books, New York.
Schachter, S., Singer, J., 1962. Cognitive, social, and physiological determinants of emotional state. Psychological Review 69 (5), 379–399.
Shott, S., 1979. Emotion and social life: a symbolic interactionist analysis. American Journal of Sociology 84, 1317–1334.
Schrock, D.P., Boyd, E.M., Leaf, M., 2009. Emotion work in the public performances of male-to-female transsexuals. Archives of Sexual Behavior 38, 702–712.
Simon, R.W., Nath, L.E., 2004. Gender and emotion in the U.S.: do men and women differ in self-reports of feelings and expressive behavior? American Journal of Sociology 109 (5), 1137–1176.
Smith III, A.C., Kleinman, S., 1989. Managing emotions in medical school: students' contacts with the living and the dead. Social Psychology Quarterly 52 (1), 56–59.
Smith-Lovin, L., 1990. Emotion as the confirmation and disconfirmation of identity: an affect control model. In: Kemper, T.D. (Ed.), Research Agendas in the Sociology of Emotions. State University of New York Press, Albany, NW.
Smith-Lovin, L., 1995. The sociology of affect and emotion. In: Cook, K., Fine, G., House, J. (Eds.), Sociological Perspectives on Social Psychology. Allyn and Bacon, New York.
Stacey, C., 2011. The Caring Self: Work Experiences of Home Health Care Aides. Cornell University Press, Ithaca, NY.
Stearns, C.Z., Stearns, P.N., 1989. Anger: The Struggle for Emotional Control in America's History. University of Chicago Press, Chicago, IL.
Stearns, P.N., Lewis, J. (Eds.), 1998. An Emotional History of the United States. New York University Press, New York.
Stets, J.E., 2005. Examining emotions in identity theory. Social Psychology Quarterly 68, 39–56.
Stets, J.E., Turner, J.H. (Eds.), 2006. Handbook of the Sociology of Emotion, vol. I. Springer, New York.
Stets, J.E., Turner, J.H. (Eds.), August 2014. Handbook of the Sociology of Emotion, vol. II. Springer, New York.
Stryker, S., Serpe, R.T., 1994. Identity salience and psychological centrality: equivalent, overlapping, or complementary concepts. Social Psychology Quarterly 57, 16–35.
Stryker, S., Serpe, R.T., 1982. Commitment, identity salience and role behavior: theory and research example. In: Ickes, W., Knowles, E. (Eds.), Personality, Roles and Social Behavior. Springer-Verlag, New York, pp. 192–216.
Sutton, R.I., 1991. Maintaining norms about expressed emotions: the case of bill collectors. Administrative Science Quarterly 36 (2), 245–268.
Taylor, V., 1996. Rock-a-by Baby: Feminism, Self-help and Post-Partum Depression. Routledge, New York.
Thoits, P.A., 1995. Managing the emotions of others. Symbolic Interaction 19 (2), 85–109.
Thoits, P.A., 1990. Emotional deviance: research agendas. In: Kemper, T.D. (Ed.), Research Agendas in the Sociology of Emotions. State University of New York Press, Albany.
Thoits, P.A., 1985. Self-labeling processes in mental illness: the role of emotional deviance. American Journal of Sociology 92 (2), 221–249.
Thoits, P.A., 1989. The sociology of emotions. Annual Review of Sociology 15, 317–342.
Turner, J.H., Stets, J.E., 2006. Sociological theories of human emotions. Annual Review of Sociology 32, 25–52.
Vaccaro, C., Schrock, D., McCabe, J., 2011. Managing emotional manhood: fighting and fostering fear in mixed martial arts. Social Psychology Quarterly 74 (4), 414–437.
Van Maanen, J., Kunda, G., 1989. Real feelings: emotional expression and organizational culture. In: Cummings, L., Staw, B. (Eds.), Research in Organizational Behavior. JAI, Greenwich, CT.
Walster, E., Walster, G.W., Bershcheid, E., 1978. Equity: Theory and Research. Allyn and Bacon, Inc., New York.
Weber, M., 1958. The Protestant Ethic and the Spirit of Capitalism. Charles Scribner's Sons, New York.
Wharton, A.S., 2009. The sociology of emotional labor. Annual Review of Sociology 35, 147–165.
Wingfield, A.H., 2010. Are some emotions marked 'whites only'? Racialized feeling rules in professional workplaces. Social Problems 57 (2), 251–268.

Empathy: A Social Neuroscience Perspective

Jean Decety and Greg J Norman, The University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of empathy is applied to various phenomena which cover a broad spectrum ranging from experiencing emotions that match another individual's emotions, feelings of concern for other people and knowing what the other is thinking or feeling. This conceptual diversity explains the difficulties in measuring empathy. None of the attempts to quantify it with self-reports, peer ratings, or rating scales of observed behavior have been able to capture the entire range of affective, cognitive, and behavioral components of empathy. In this article, we will expound that human empathy is better understood with reference to the phylogeny of the social brain. Empathy has deep evolutionary, biochemical, and neurological underpinnings. Even the most advanced forms of empathy in humans remain connected to core mechanisms associated with affective communication, social attachment, and parental care. Empathy is mediated and moderated by multiple physiological and brain systems as well as social context. Finally, drawing from patients–doctor relationships in medicine, we will show that empathy is adaptive for oneself and for others.

Empathy—the natural capacity to share and understand others' feelings and thoughts—plays a pivotal role in human social interaction. Empathy-related processes motivate prosocial behavior, inhibit aggression, and provide the foundation of moral reasoning. Empathy is such a necessary component of healthy coexistence that its absence leads to serious social-emotional dysfunctions, including psychopathologies such as psychopathy. Empathy and other prosocial behaviors have been a focus of speculation in philosophical, psychological, and physiological investigations throughout written history (Batson, 2009). However, over the last decade there has been an explosion of research into empathy and its underlying evolutionary, psychological, and neurobiological foundations that have taken the topic from the realm of philosophical speculation to the confines of rigorous scientific investigation (Mason & Ben Ami-Bartal, 2010).

In the present article, we propose that the topic of human empathy is best understood within the context of its underlying evolutionary, neurobiological, and psychological mechanisms. We will show that empathy has deep evolutionary, biochemical, and neurobiological underpinnings associated with affective communication, social attachment and parental care, characteristic of all mammals. Finally, drawing from research in medicine on patient–doctor relationships, we will expound that the application of current research into empathy has the potential to improve physician efficacy and patient outcomes.

Evolution of Empathy

Social species, particularly primates, are adapted to live in groups and their members are naturally driven toward establishing and maintaining social bonds. Adults who formed attachments would be more likely to reproduce and long-term relationships would increase the chances that the offspring would reach maturity and reproduce in turn. Empathy increases the inclusive fitness of organisms by enhancing survival in a variety of ways (e.g., providing increased defense against predators) and it serves to bond individuals to one another, especially mothers to offspring (Plutchik, 1987). Mammalian reproductive fitness and survival depend crucially on the ability of conspecifics to communicate with each other, sharing information about their affective and motivational states and appropriately responding to their allies and relative's needs. Importantly, empathy entails more than just understandings or expectations. When individuals empathize, they vicariously feel the emotions of others, which can promote affective communication and can motivate individuals to behave prosocially toward conspecifics (Decety, 2011).

The psychological and neurobiological components that underlie empathy and empathy-related processes have been conserved across evolutionary history and are apparent to some extent in all mammals. Indeed, the capacity for empathy in humans emerges from an elaboration and rerepresentation of functions apparent in other social mammals. Thus, the neurobiological substrates of empathy-related processes are present at many levels of the neuraxis. The rerepresentation of empathy-related processes allows for the rapid expression of empathic-like responses, such as orofacial mimicry, while simultaneously generating more elaborated emotional and cognitive empathy-related processes, such as perspective taking, that rely on dynamic interactions between limbic and cortical structures (Buchanan et al., 2012; Decety and Jackson, 2004; de Waal, 2008). Research in animals has shown that the capacity to share and be affected by the emotional state of conspecifics is generated by basic neurobiological systems subserving attachment-related processes, involving the brainstem, preoptic area of the thalamus, and paralimbic areas (Panksepp, 1998; Watt, 2000). As a result of the importance of social connections for mammalian survival, these attachment systems appear to exploit the well-developed physical pain system, borrowing the aversive signals associated with pain to indicate when relationships are threatened (Eisenberger, 2011; MacDonald and Leary, 2005). Therefore, just as the physical pain system alerts organisms to the presence of a noxious environmental stimulus and motivates them to rapidly escape the source of pain, so too does the social pain system; the experience of social pain alerts an individual to

potential threats in their social environment and can induce various coping strategies to attempt to mitigate the threat (e.g., increase motivation to strengthen relationships) (MacDonald and Leary, 2005). In addition to the pain system mentioned above, the dopaminergic reward system also appears to have been partially coopted for attachment and caring, and thus for empathic concern. Indeed, mammals are highly motivated to care for their offspring and experience this interaction as a highly rewarding experience (Mayes et al., 2009). Although empathy in humans is assisted by other general high-level cognitive abilities such as executive functions and language, which introduce contextual control and expand the repertoire of behaviors that can be driven by empathy and emotional connection, it also operates on more primitive reward processes and is highly dependent upon these lower level processes to achieve higher order goals associated with affective communication, social attachment, parental care, and motivation to cooperate.

Motivation to Care for Others

One likely source of empathic responses in mammals comes from the phylogenetically ancient practice of caring for one's offspring for extended periods of time that can be observed in other organisms (Cockburn, 2006; Goodwin et al., 1998). Mammals in particular have developed intricate parenting behaviors. These behaviors tend to take the form of large energetic investments in a fewer number of offspring. The ability to provide offspring with nourishment (i.e., lactation) and safety over long periods of time facilitates the development of more complex neurobiological and behavioral processes (i.e., social attachment) that improve the offspring's ability to pass on its genes into future generations. While there are clear advantages to this strategy, it also comes with significant risks. Indeed, this approach is less resilient to predation and disease as the loss of a single offspring has the ability to render the parental energy expenditure useless. In order to counteract this risk, mammals have developed a strong emotional attachment to their offspring that predisposes individuals to remain hypervigilant for potential threats thereby minimizing their risk of losing the ability to pass their genes onto future generations due to predation or disease.

The neurobehavioral manifestations of parental care also provide the means through which individuals within a social group are able to cooperate and care for one another. Kin selection models of social behavior presumes that the prosocial tendencies within groups resulted from the fact that organisms that assist their genetic relatives are better able to propagate a fraction of their own genetic legacy into future generations (e.g., assisting sibling's offspring which share 25% of one's genes) (Wilson, 1975). However, kin selection models of prosocial behaviors within groups do not actually specify how these behaviors are motivated and do not assume that individuals must be able to detect genetic relatedness (Bell, 2001). Thus, any evolved motivational system that increases the fitness of kin will produce the same results in terms of fitness regardless of fact that this motivational system may also apply to nonrelatives (Bell, 2010). The ability to model the emotions of nonrelatives and react appropriately within in a social group would likely confer some fitness advantages (e.g., better able to communicate and detect distress in group members). For example, the motivational systems that may have originally been developed to care for offspring have likely been coopted and used in the service of facilitating positive relationships between unrelated group members. In humans for example, our relationships with spouses, friends, and coworkers are highly valued and require vast amounts of psychological resources to maintain. In fact, the perception that such relationships are threatened engenders profound emotional and physiological stress responses (Norman et al., 2012a). Conversely, feeling well connected with friends and family provides a strong behavioral and physiological buffer that can diminish stress responses (Uchino et al., 1996). Therefore, while the motivational components of prosocial behaviors such as empathy may have originally developed in service of parental care, they have now become invaluable tools for the formation and maintenance of strong social bonds between unrelated individuals who provide the foundation for much of our successes as a social species. Importantly, the evolution of a caregiving system has produced social preferences. Kin selection and group selection may have occurred as an outcome of the evolution of a caregiving system.

Neuroevolution of Empathy-Related Behaviors

At the behavioral level, empathy-like responses can be observed in numerous species, for example, mammalian and avian species. For instance, the presence of specific behavioral (e.g., decreased preening behavior) and physiological (eye temperature) changes in hens observing their chicks being exposed to a mildly aversive stimulus indicates a responsive capacity that is distinguishable from the hens' own experiences of the same stimulus (Edgar et al., 2011). A number of empirical studies with rodents have demonstrated that affective arousal and emotional contagion can prompt efforts to alleviate the distress of a conspecific. For instance, rats that had learned to press a lever to obtain food stop doing so if their action is paired with the delivery of an electrical shock to a visible neighboring rat (Church, 1959) and rats will press a bar to lower another rat suspended in mid-air (Rice and Gainer, 1962). Similarly, rats will work to free a cagemate locked in a restrainer even when social reward was prohibited (Ben-Ami Bartal, Decety and Mason, 2011). This latter study found that rats that were made to choose between liberating a cagemate or receiving a highly palatable food (chocolate), they typically chose to release their cagemate and subsequently share the chocolate. Furthermore, pain sensitivity in mice is modulated by the presence of other mice displaying pain behaviors (Langford et al., 2006). Interestingly, this relationship is conditional upon the identity of the target mouse such that observing pain behaviors in conspecifics only influences pain behavior when the target mouse is their cagemate. Similarly, female mice show more freezing behavior when exposed to the pain of a close relative than when exposed to the pain of a more distant relative (Jeon et al., 2010). To investigate whether such pain behavior can serve the function of soliciting a primitive

form of empathic concern, Langford et al. (2010) used a social approach paradigm to test mice in various dyadic or triadic conditions. Some conditions involved restrained mice that were in pain as a result of intraperitoneal injection of acetic acid and test mice free to approach or avoid the restrained mice. Results revealed a gender-specific effect whereby female, but not male, mice approached a familiar same-gender conspecific in pain at higher rates than an unaffected familiar or unfamiliar, but affected, conspecific. These results can be interpreted as evidence of an effect of social support on the experience of pain, a finding consistent with human literature. Indeed, the presence of an individual who provides passive or active support reduces experimental pain in humans (Brown et al., 2003). It has been reported that perceived empathy of a partner can modulate subjective and autonomic responses to physical pain (Sambo et al., 2010).

Empirical reports of empathic reaction in apes indicate they too have an appreciation for the situation of conspecifics (Warneken et al., 2007). Primitive aspects of empathy, such as emotion contagion and parental care, have been clearly demonstrated in chimpanzees (de Wall, 2008). For example, an experiment in which peripheral skin temperature was measured in chimpanzees while they viewed an emotionally laden video demonstrated a decrease of skin temperature, indicative of sympathetic arousal, when they viewed videos of conspecifics injected with needles or videos of the needles themselves, whereas these changes were not observed when the chimpanzees viewed videos of a conspecific chasing the veterinarian (Parr, 2001). Thus, when chimpanzees perceive a conspecific exposed to painful stimuli, they show physiological changes similar to those observed in humans (Hatfield et al., 2009). Providing comfort to distressed conspecifics is an expression of empathic concern, and such other oriented behavior is now well documented in chimpanzees (Romero et al., 2010). Therefore, various social organisms display aspects of empathy and understanding of how these processes function in related animals provides an important insight into their role in human social behavior.

Neural Circuits Associated with the Perception of Others' Pain and Distress

The long history of mammalian evolution has shaped maternal brains (and paternal brains to a lesser extent) to be sensitive to signs of suffering in one's own offspring (Haidt and Graham, 2007). Even very young infants of 6 months of age are capable of preverbal evaluations when viewing social interactions. For example, babies are more likely to smile when viewing prosocial events, and to frown in response to antisocial events (Bloom, 2012). The earliest forms of empathic concern in humans appear between 8 and 16 months of age and continue to develop until adulthood (Roth-Hanania et al., 2011; Vaish and Warneken, 2012). In many primates and other mammals, this sensitivity has extended beyond the mother–child relationship; so all typically developed individuals generally find the suffering of others as aversive. Pain serves adaptive functions not only by warning the suffering individual, but also by impelling expressive behaviors that attract the attention of others that may provide assistance (Craig, 2009). Crying for instance is one prototypical attachment behavior. It is only an advantage to the extent that crying attracts a caregiver faster than it attracts a predator. Thus, pain can be considered an emotion triggered by the activation of a mental or neural representation of actual or potential tissue damage. This representation involves somatic sensory features, as well as affective–motivational reactions associated with the promotion of protective or recuperative visceromotor and behavioral responses. It is the affective experience of pain that signals an aversive state and motivates behaviors to terminate, reduce, or escape exposure to the source of noxious stimulation. Furthermore, the expression of pain also provides a crucial signal that can motivate soothing and caring behaviors in others (Jackson et al., 2005). Therefore, in addition to removing the potential threat, the expression of pain behaviors allows individuals to obtain assistance from other individuals within the group.

In humans, a number of neuroimaging studies have documented reliable activation of a neural network involved in the processing of pain, including the anterior midcingulate cortex (aMCC), anterior insular cortex (AIC), supplementary motor area (SMA), and periaqueductal gray (PAG) when individuals watch facial expressions of pain (Botvinick et al., 2005; Lamm et al., 2007) or body parts being injured (Jackson et al., 2005; Cheng et al., 2008; Decety and Michalska, 2010), imagine the pain of others (Jackson et al., 2006), or simply observe a signal indicating that someone will be hurt (Singer et al., 2004). It is worth noting that vicariously instigated activations of the pain matrix are not specific to the sensory qualities of pain, but instead are associated with more general survival mechanisms such as aversion and withdrawal when exposed to danger and threat (Decety, 2010). Of particular importance, the aMCC, a region that implements a domain-general process that is integral to negative affect, pain and cognitive control contains pain-responsive neurons that are activated by both anticipation of pain and instrumental escape from pain (Shackman et al., 2011). Activation of the AIC is nearly ubiquitous in studies of pain empathy: this response can even be elicited automatically as shown when participants viewed color photographs depicting human body parts in painful or nonpainful situations and performed either pain judgment (painful/nonpainful) or laterality judgment (left/right) of the body parts in the absence of explicit task requirements and attentional demands (Gu et al., 2010). The AIC and aMCC with the amygdala are parts of a salient network that segregates the most relevant among internal and extrapersonal stimuli in order to guide behavior (Menon and Uddin, 2010).

These findings demonstrate the ease with which an individual can integrate the physical and social pain of another into their own bodily somatic representation. Consistent with Jackson's notion of neuroevolutionary rerepresentation of function (Jackson, 1884), it has been proposed that the anterior insula, and its array of projections, serves to generate a higher order metarepresentation of primary interoceptive activity, which is related to the feeling of pain and its emotional awareness (Craig, 2007). These representations play an important role in the learning and adaptation of social behavior in addition to basic decision-making and homeostatic processes and play a critical role in empathy.

Furthermore, the neural overlap between the first-hand experience of pain and its third-person perception or

imagination is consistent with neural reuse theories which posit that neural circuits continue to acquire new use after an initial or original function is established (Anderson, 2010).

Altogether, results from these studies have bolstered the theory that shared representations between self and others rely on similar neural computations between first-person experience and third-person observation (Decety and Sommerville, 2003). Whether this overlap in activation between self and others reflects an identical neural population or originates from the intersection of adjacent and partially distinct activations is far from being established (see Box 1). One of the few studies specifically designed to address the extent of the overlap between felt pain and seen pain in others did not find identical activation in the anterior cingulate cortex, and thus did not support a mirrorlike interpretation (Morrison and Downing, 2007). Rather, the findings are more in line with general neural mechanisms involved in aversion-related processing, in which the ACC and anterior insula play a critical role.

Aspects of Empathy Specific to Humans

Humans have the capacity to feel empathic concern in response to a wide range of contexts in which others are in need, and even routinely extend this concern beyond conspecifics and onto other species. The development of human empathy has been elaborated through the integration of other abstract and domain-general high-level cognitive abilities such as executive functions, language, and theory of mind, underpinned by the prefrontal cortex (PFC), which expand the range of situations that can elicit empathy and the range of behaviors that can be driven by empathy (Stone, 2006).

Human language is a powerful tool to create and share emotional states, both by its content and prosody, which is not found in other species. By its very nature, language acts at a distance across space and, in the case of written language, across time (Harris, 2000). Conversation helps to develop empathy, for it is often here that people learn of shared experiences and feelings. Several empirical studies have now demonstrated that affective language dampens amygdala response (Lieberman et al., 2007) and facilitates exposure-related attenuation of autonomic reactivity to aversive stimuli (Tabibnia et al., 2008). Thus, the ability to communicate emotional states with others through speech and linguistic expression provides humans with a powerful means to share their emotional states with others and to extend the reach of empathic concern across space and time. Importantly, language production and comprehension are grounded in internal bodily states, and seem to involve circuits long associated with motor control functions (Pulvermuller, 2005), suggesting a rerepresentation of function and not the evolution of entirely new neurobiological processes.

Another seemingly unique aspect of empathy in humans is perspective taking, the cognitive ability to explicitly put oneself into someone else's shoes to represent his or her knowledge or emotional experience as compared with one's knowledge or affective experience. Such a capacity requires additional computational mechanisms (working memory and inhibitory control to hold simultaneously two perspectives in mind) needed for its development, which plays a critical role in promoting empathic concern. Neuroimaging studies have revealed a specific network of brain regions that are recruited when one takes the perspective of another individual to represent their knowledge or emotional experience as compared with one's own. Adopting a third person perspective, as opposed to a first person perspective has been associated with neural activation in the posterior superior temporal sulcus (pSTS) and medial prefrontal cortex (medial PFC) (Jenkins and Mitchell, 2011; Ruby and Decety, 2003, 2004). Studies in social psychology have demonstrated that empathic concern is enhanced when participants are told of another individual's plight and asked to imagine how that person feels (Batson et al., 1997). However, when the same individuals were asked to imagine instead how they would feel in the place of the other person, feelings of anxiety and personal distress were evoked. These results suggest that perspective taking helps keep feelings of personal distress at a distance, thereby permitting the emergence of empathic concern. To test this hypothesis, a functional magnetic resonance imaging (fMRI) study was performed in which participants viewed videoclips of individuals exhibiting facial expressions of pain, purportedly the result of a painful medical treatment (Lamm et al., 2007). Participants were instructed to either imagine how they would themselves feel if they were in the patient's

Box 1 The contribution of multisensory processing to the experience of empathy

Sparked by the discovery of sensorimotor neurons (AKA mirror neurons) in the monkey premotor and posterior parietal cortices that discharge both when a particular action is performed (such as grasping) and when the monkey observes the same action performed by another individual, a number of functional neuroimaging studies have documented that shared brain circuits are involved when processing one's own somatosensory and affective experiences and when one perceives these states in other individuals. For instance, participants scanned while watching video clips of a leg being touched or having their own leg being touched activates the secondary somatosensory cortex (Keysers et al., 2004). In another study, subjects observed a hand either being touched or not touched by a brush presented from either egocentric or allocentric perspectives (Schaefer et al., 2009). While both perspectives of viewing touch were associated with activation of the somatosensory, they also showed differences. The egocentric perspective yielded activation of the anterior part of SI and the allocentric perspective in the posterior part of SI. In addition, only the former perspective induced self-attribution of being touched oneself. The response in regions of the interoceptive sensory system, such as the posterior insula, processing the internal affective feeling states, is specifically modulated by the visual perception of other individuals' social or affective cutaneous experiences (Ebish et al., 2011).

Together, these data support models of embodied simulation proposing that multimodal cortical areas underlying first-person bodily experiences are also involved with an implicit, bottom-up understanding of the actions and emotions of other individuals, thus playing a role in intersubjectivity and empathic experiences (Decety and Grèzes, 2006; Gallese, 2003). Several authors have interpreted these studies as self-other overlap. At the neural level, the perception of another affective state activates neural regions or circuits in the observer that are also activated when he/she experiences the same affective state.

situation, or imagine what the patient was feeling. The former perspective was associated with a strong hemodynamic increase in the amygdala, AIC, and anterior cingulate cortex (ACC), as well as reports of anxiety and personal distress. When the participants were imagining what the patients were feeling, a significant reduction in activity was detected in the amygdala, with reduced feelings of anxiety combined with increased reports of empathic concern.

As illustrated by the example above, the evolutionary development of more complex neural systems, such as the regions of the medial PFC involved in perspective taking, has thus conferred greater behavioral flexibility and contextual control. For instance, Lamm et al. (2010) used fMRI to examine how participants empathize with the feelings of patients who reacted with no pain to surgical procedures but with pain to a soft touch. Results showed that empathizing with someone whose bodily and affective representations are distinct from our own recruited areas of the PFC involved in self-other distinction (medial PFC) and cognitive control (right inferior frontal gyrus), and these regions increased their functional connectivity with more phylogenetically older regions such as the amygdala, brainstem, insula, and PAG, that serve basic protective functions, including affective arousal.

Cognitive, sensorimotor, and somatovisceral mechanisms are thus intimately connected, and in line with neural reuse theories (Anderson, 2010) share similar computational resources. One salient example comes from work on humans that are experimentally induced to feel they have been ostracized. As expected, rejected participants reported and increase in negative emotions. However, when the identical study was done in an MRI scanner something far more interesting was detected. The act of being rejected actually brought about an increase in the activity of brain structures known to mediate the physiological response to pain (Eisenberger, 2011; Kross et al., 2011). Just as pain is a response to remove the organisms away from a noxious stimulus, the aversive feelings associated with social rejection represent a personal alarm alerting one that their status within the group may be compromised and motivates the individual to rebuild and reconnect social relationships (Cacioppo and Patrick, 2009). In this way, evolutionary processes are able to exploit an already online system that is fully capable of detecting and avoiding threats. Accordingly, knowledge regarding primitive neurobehavioral processes (e.g., pain, thirst) can provide important insights into higher level processes such as empathy that operate through an elaboration of these evolutionarily ancient and highly effective survival mechanisms.

Empathy as an Adaptive Response

Empathy is a flexible phenomenon, and its induction and expression are dependent upon various social and contextual factors. Clearly, it would be maladaptive to respond with care and to extend one's empathic concern to all. As noted above, rodents do not react indiscriminately to other conspecifics in distress (Langford et al., 2010). Recent neuroimaging studies in humans reveal that the neural network implicated in empathy for pain is modulated by various social and contextual factors. For instance, one fMRI study demonstrated that empathic arousal is moderated early in information processing by a priori attitudes toward conspecifics (Decety et al., 2009). Study participants were significantly more sensitive to the pain of individuals who had contracted AIDS as a result of a blood transfusion as compared with individuals who had contracted AIDS as a result of their illicit drug addiction (sharing needles), as evidenced by higher pain and empathy ratings and greater hemodynamic activity in areas associated with pain processing (i.e., AIC, aMCC, PAG). Furthermore, activity in the pain matrix network is enhanced when people viewed their loved ones in pain compared with strangers (Cheng et al., 2010). Another neuroimaging study demonstrated that the failures of an in-group member are painful, whereas those of a rival out-group member gives pleasure – a feeling that may motivate harming rivals (Cikara et al., 2011).

Empathic arousal is also modulated by an individual's knowledge and experience with pain. Two neuroimaging studies directly investigated how physicians react to the perception of others' pain. One study compared the neuro-hemodynamic response in a group of physicians and a group of matched control participants while they viewed videoclips depicting mild bodily harm (i.e., needle pricks) or being touched by a cotton swab (nonpainful situations) (Cheng et al., 2007). The results demonstrated activation of the pain matrix in the controls when they attended to the painful situations relative to the nonpainful ones. A different pattern of signal change was detected in the physicians; cortical regions underlying executive functions and executive attention were found to be activated, and unlike in the control group, no signal increase was detected in the pain matrix. A second study recorded event-related potentials (ERPs) from physicians and matched controls as they were presented with the same visual stimuli (Decety et al., 2010). The results showed early N110 differentiation between pain and no pain, reflecting negative arousal, over the frontal cortex, as well as late P300 over the centroparietal regions in control participants. In contrast, no such early ERP response was detected in the physicians.

Consequences of Being Empathetic

As empathy has evolved in the context of parental care and social bounds, it comes as no surprise that clinical empathy is an essential element of quality care. This interpersonal skill is an acknowledged essential component of professional competence (Epstein and Hundert, 2002). Patients desire it, and physicians who are empathic are more satisfied and have less burnout (Larson and Yao, 2005). There is evidence that perceived empathy is associated with improved patient satisfaction, adherence to treatment, as well as fewer malpractice complaints (Gleichgerrcht, & Decety, 2012). In contrast to models of 'detached concern,' physicians who attempt to understand what their patient is feeling and communicate their concern achieve a number of valuable outcomes for their patients and themselves (Halpern, 2012). Physicians' own emotions may help them attune to and empathically understand patients' emotional states, which may help produce a positive therapeutic impact. Sincere empathy is recognized as a major constituent in the practice

styles of excellent 'healing' clinicians (Schattner, 2012). Research shows that physicians' empathy actually improves patient outcomes for chronic diseases such as diabetes. One study including 891 patients with diabetes whose doctors had high empathy scores were better able to regulate blood glucose levels than patients whose doctors had low empathy scores (Hojat et al., 2011). Clinician empathy, as perceived by patients with the common cold, significantly predicts subsequent duration and severity of illness and is associated with immune system changes (Rakel et al., 2009). Doctors who show empathy have less malpractice complaints and better patient satisfaction and compliance (Huntington and Kuhn, 2003). Therefore, physician empathy can improve the patient's psychological and physiological adjustment to disease; a fact that should be earmarked for better communication throughout medical school training.

Conclusions

Even the most advanced forms of empathy in humans are built on more basic forms and remain connected to core mechanisms associated with affective communication, social attachment, and parental care. Empathy is not an all-or-none phenomenon, nor is it automatic or reflexive, as many social, interpersonal and contextual factors affect its induction and expression. The incoming sensory information is constrained and filtered by appraisal and reappraisal processing, which shapes the emergence of the experience of empathy and behavioral outcomes. As discussed above, empathy-related behaviors have coopted more primitive homeostatic processes involved in reward and pain systems in order to facilitate various social attachment processes. This evolutionary perspective is compatible with the nested brain-mind hierarchies view developed by Panksepp (2011), which posits that similar to other mammalian species, human emotions may spread across conspecifics, and at times these shared emotions may facilitate empathic concern, which promotes prosocial behaviors and altruism. Primary emotional processes, where sources of empathy may arise (i.e., feeling what other organisms are feeling), coordinate with secondary-process learning and memory mechanisms (i.e., knowing what others are feeling). Both of these then interact with higher mental processes, which can exert a variety of top-down influences on the regulation of empathic tendencies (i.e., desires to respond compassionately to others' distress). Once the capacity for empathy evolved, it came to be expressed outside the parental care context (de Waal, 2008). For instance, when people send money to distant earthquake victims in Haiti, or petition to support a bill that would contribute to curb the violence in Darfur, empathy reaches beyond its evolutionary origins, extending beyond inclusive fitness among kin and social group members. Along with enculturation, the development of empathy and caring is then further assisted by domain-general and high-level cognitive abilities such as executive functions, theory of mind, rational cognition, and language, which combine to expand the range and the scope of empathically driven behavior. Furthermore, the development of human empathy across history may have been facilitated by advancements and innovations from our past. Notably, Singer (1986) argued that over the course of history, people have enlarged the range of beings whose interests they value as value of their own.

Overall, the evolutionary conceptual view of empathy that we have developed in this article is compatible with the hypothesis that advanced levels of social cognition have arisen as an emergent property of complex interactions between the neurobiological structures underlying executive functioning and affect. Importantly, these cognitive functions are likely to rely and interact with preexisting neural circuits already possessing suitable resources. Similar neurobiological mechanisms that regulate parental behavior, attachment, and affective processing in all mammals interact with more recent cortical systems shaped by social, cultural, and educational contexts to produce the flexible and generalized forms of nurturant care found among humans. These evolutionary newer developments help explain why humans care not only for their offspring but also for strangers, and why they are motivated to uphold moral principles such as justice and fairness (Batson, 2011).

See also: Adaptation, Fitness, and Evolution; Cognition, Evolution of; Comparative Evolutionary Psychology: Current Status and a Proposal for a More Integrated Future; Culture and Emotion; Emotion in Cognition; Emotional Regulation; Empathy During Early Childhood Across Cultures, Development of; Evolutionary Social Psychology; Human Cognition, Evolution of; Morality: Evolution of; Occupational Health; Primates, Evolution of; Prosocial Behavior and Empathy.

Bibliography

Anderson, M.L., 2010. Neural reuse: a fundamental organizational principle of the brain. Behavioral and Brain Sciences 33, 245–313.
Batson, C.D., 2009. These things called empathy: eight related but distinct phenomena. In: Decety, J., Ickes, W. (Eds.), The Social Neuroscience of Empathy. MIT Press, Cambridge, pp. 3–15.
Batson, C.D., 2011. Altruism in Humans. Oxford University Press, New York.
Batson, C.D., Early, S., Salvarani, H., 1997. Perspective taking: imagining how another feels versus imagining how you would feel. Personality and Social Psychology Bulletin 23, 751–758.
Bell, D.C., 2001. Evolution of parental caregiving. Personality and Social Psychology Review 5, 216–229.
Bell, D.C., 2010. The Dynamics of Connection. Lexington Books, Plymouth, UK.
Ben-Ami Bartal, I., Decety, J., Mason, P., 2011. Empathy and pro-social behavior in rats. Science 334, 1427–1430.
Bloom, P., 2012. Moral nativism and moral psychology. In: Mikulincer, M., Shaver, P.R. (Eds.), The Social Psychology of Morality. APA, Washington DC, pp. 71–89.
Botvinick, M., Jha, A.P., Bylsma, L.M., Fabian, S., Solomon, P., Prkachine, K.M., 2005. Viewing facial expressions of pain engages cortical areas involved in the direct experience of pain. Neuroimage 25, 312–319.
Brown, J.L., Sheffield, D., Leary, M., Robinson, M.E., 2003. Social support and experimental pain. Psychosomatic Medicine 65, 276–283.
Buchanan, T.W., Bagley, S., Stansfield, R.B., Preston, S.D., 2012. The empathic, physiological resonance of stress. Social Neuroscience 7 (2), 191–201.
Cacioppo, J.T., Patrick, B., 2009. Loneliness: Human Nature and the Need for Social Connection. W. W. Norton and Company, New York.
Cheng, Y., Chen, C.Y., Lin, C.P., Chou, K.H., Decety, J., 2010. Love hurts: an fMRI study. NeuroImage 51, 923–929.
Cheng, Y., Lin, C., Liu, H.L., Hsu, Y., Lim, K., Hung, D., Decety, J., 2007. Expertise modulates the perception of pain in others. Current Biology 17, 1708–1713.
Cheng, Y., Yang, C.Y., Lin, C.P., Lee, P.R., Decety, J., 2008. The perception of pain in others suppresses somatosensory oscillations: a magnetoencephalography study. NeuroImage 40, 1833–1840.

Church, R.M., 1959. Emotional reactions of rats to the pain of others. Journal of Comparative and Physiological Psychology 52, 132–134.
Cikara, M., Botvinick, M.M., Fiske, S.T., 2011. Us versus them: social identity shapes responses to intergroup competition and harm. Psychological Science 22, 306–313.
Cockburn, A., 2006. Prevalence of different modes of parental care in birds. Proceedings of the Royal Society London, B, Biological Sciences 273, 1375–1383.
Craig, A.D., 2007. Interoception and emotion: a neuroanatomical perspective. In: Lewis, M., Haviland-Jones, J.M., Barrett, L.F. (Eds.), Handbook of Emotion, third ed. Guilford Press, New York, pp. 272–288.
Craig, K.D., 2009. The social communication model of pain. Canadian Psychology 50, 22–32.
De Waal, F.B.M., 2008. Putting the altruism back into altruism: the evolution of empathy. Annual Review of Psychology 59, 279–300.
Decety, J., 2010. To what extent is the experience of empathy mediated by shared neural circuits? Emotion Review 2, 204–207.
Decety, J., 2011. Dissecting the neural mechanisms mediating empathy. Emotion Review 3, 92–108.
Decety, J., Echols, S.C., Correll, J., 2009. The blame game: the effect of responsibility and social stigma on empathy for pain. Journal of Cognitive Neuroscience 22, 985–997.
Decety, J., Grèzes, J., 2006. The power of simulation: imagining one's own and other's behavior. Brain Research 1079, 4–14.
Decety, J., Jackson, P.L., 2004. The functional architecture of human empathy. Behavioral and Cognitive Neuroscience Reviews 3, 71–100.
Decety, J., Michalska, K.J., 2010. Neurodevelopmental changes in the circuits underlying empathy and sympathy from childhood to adulthood. Developmental Science 13, 886–899.
Decety, J., Sommerville, J.A., 2003. Shared representations between self and others: a social cognitive neuroscience view. Trends in Cognitive Sciences 7, 527–533.
Decety, J., Yang, C.Y., Cheng, Y., 2010. Physicians down regulate their pain empathy response: an event-related brain potential study. NeuroImage 50, 1676–1682.
Ebisch, S.J.H., Ferri, F., Salone, A., Perrucci, M.G., D'Amico, L., Ferro, F.M., Romani, G.L., Gallese, V., 2011. Differential involvement of somatosensory and interoceptive cortices during the observation of affective touch. Journal of Cognitive Neuroscience 23, 1808–1822.
Edgar, J.L., Lowe, J.C., Paul, E.S., Nicol, C.J., 2011. Avian maternal response to chick distress. Proceeding of the Royal Society, Biological Sciences 278, 3129–3134.
Eisenberger, N., 2011. Why rejection hurts: what social neuroscience has revealed about the brain's response to social rejection. In: Decety, J., Cacioppo, J.T. (Eds.), The Oxford Handbook of Social Neuroscience. Oxford University Press, pp. 586–598.
Epstein, R.M., Hundert, E.M., 2002. Defining and assessing professional competence. JAMA 287, 226–235.
Gallese, V., 2003. The roots of empathy: the shared manifold hypothesis and the neural basis of intersubjectivity. Psychopathology 36, 171–180.
Gleichgerrcht, E., Decety, J., 2012. The costs of empathy among health professionals. In: Decety, J. (Ed.), Empathy: From Bench to Bedside. MIT Press, Cambridge, pp. 245–261.
Goodwin, N.B., Balshine-Earn, S., Reynolds, J.D., 1998. Evolutionary transitions in parental care in cichlid fish. Proceeding of the Royal Society, Biological Sciences 265, 2265–2272.
Gu, X., Liu, X., Guise, K.G., Naidich, T.P., Hof, P.R., Fan, J., 2010. Functional dissociation of the frontoinsular and anterior cingulate cortices in empathy for pain. Journal of Neuroscience 30, 3739–3744.
Haidt, J., Graham, J., 2007. When morality opposes justice: conservative have moral intuitions that liberals may not recognize. Social Justice Research 20, 98–116.
Halpern, J., 2012. Clinical empathy in medical care. In: Decety, J. (Ed.), Empathy – From Bench to Bedside. MIT Press, Cambridge, MA, pp. 229–244.
Harris, P.L., 2000. Understanding emotion. In: Lewis, M., Haviland-Jones, J.M. (Eds.), Handbook of Emotions. Guilford Press, New York, pp. 281–292.
Hatfield, E., Rapson, R.L., Le, Y.C., 2009. Emotional contagion and empathy. In: Decety, J., Ickes, W. (Eds.), The Social Neuroscience of Empathy. MIT Press, Cambridge, MA, pp. 19–30.
Hojat, M., Louis, D.Z., Markham, F.W., Wender, R., Rabinowitz, C., Gonnella, J.S., 2011. Physicians' empathy and clinical outcomes for diabetic patients. Academic Medicine 86, 359–364.
Huntington, B., Kuhn, N., 2003. Communication gaffes: a root cause of malpractice claims. Baylor University Medical Center Proceedings 16, 157–161.
Jackson, J.H., 1884. Evolution and dissolution of the nervous system (Croonian Lectures). Reprinted In: Taylor, J. (Ed.), 1958, Selected Writings of John Hughlings Jackson. Basic Books, New York.
Jackson, P.L., Brunet, E., Meltzoff, A.N., Decety, J., 2006. Empathy examined through the neural mechanisms involved in imagining how I feel versus how you feel pain: an event-related fMRI study. Neuropsychologia 44, 752–761.
Jackson, P.L., Meltzoff, A.N., Decety, J., 2005. How do we perceive the pain of others? A window into the neural processes involved in empathy. NeuroImage 24, 771–779.
Jenkins, A.C., Mitchell, J.P., 2011. Medial prefrontal cortex subserves diverse forms of self-reflection. Social Neuroscience 6 (3), 211–218.
Jeon, D., Kim, S., Chetana, M., Jo, D., Ruley, H.E., Lin, S.Y., Rabah, D., Kinet, J.P., Shin, H.S., 2010. Observational fear learning involves affective pain system and Ca(v)1.2 Ca2+ channels in ACC. Nature Neuroscience 13, 482–488.
Keysers, C., Wicker, B., Gazzola, V., Anton, J.L., Fogassi, L., Gallese, V., 2004. A touching sight: SII/PV activation during the observation and experience of touch. Neuron 42, 335–346.
Kross, E., Berman, M.E., Mischel, W., Smith, E.E., Wager, T.D., 2011. Social rejection shares somatosensory representations with physical pain. PNAS 108, 6270–6275.
Lamm, C., Batson, C.D., Decety, J., 2007. The neural basis of human empathy: effects of perspective-taking and cognitive appraisal. Journal of Cognitive Neuroscience 19, 42–58.
Lamm, C., Meltzoff, A.N., Decety, J., 2010. How do we empathize with someone who is not like us? Journal of Cognitive Neuroscience 2, 362–376.
Langford, D.J., Crager, S.E., Shehzad, Z., Smith, S.B., Sotocinal, S.G., Levenstadt, J., et al., 2006. Social modulation of pain as evidence for empathy in mice. Science 312, 1967–1970.
Langford, D.J., Tuttleb, A.H., Brown, K., Deschenes, S., Fischer, D.B., Mutso, A., Root, K.C., Sotocinal, S.G., Stern, M.A., Mogil, J.S., Sternberg, W.F., 2010. Social approach to pain in laboratory mice. Social Neuroscience 5, 163–170.
Larson, E.B., Yao, X., 2005. Clinical empathy as emotional labor in the patient-physician relationship. JAMA 293, 1100–1116.
Lieberman, M.D., Eisenberger, N.I., Crockett, M.J., Tom, S.M., Pfeifer, J.H., Way, B.M., 2007. Putting feelings into words: affect labeling disrupts amygdala activity to affective stimuli. Psychological Science 18, 421–428.
MacDonald, G., Leary, M.R., 2005. Why does social exclusion hurt? The relationship between social and physical pain. Psychological Bulletin 131, 202–223.
Mason, P., Ben Ami-Bartal, I., 2010. How the social brain experiences empathy: summary of a gathering. Social Neuroscience 5, 252–256.
Mayes, L.C., Magidson, J., Lejuez, C.W., Nichols, S.S., 2009. Social relationships as primary rewards. In: De Haan, M., Gunnar, M.R. (Eds.), Handbook of Developmental Social Neuroscience. Guilford Press, New York, pp. 342–377.
Menon, V., Uddin, L.Q., 2010. Saliency, switching, attention and control: a network model of insula function. Brain Structure and Function 214, 655–667.
Morrison, I., Downing, P.E., 2007. Organization of felt and seen pain responses in anterior cingulate cortex. Neuroimage 37, 642–651.
Norman, G.J., Hawkley, L.C., Cole, S.W., Berntson, G.G., Cacioppo, J.T., 2012. Social neuroscience: the social brain, oxytocin, and health. Social Neuroscience 7 (1), 18–29.
Panksepp, J., 1998. Affective Neuroscience: The Foundations of Human and Animal Emotions. Oxford University Press, London.
Panksepp, J., 2011. Empathy and the laws of affect. Science 334, 1358–1359.
Parr, L.A., 2001. Cognitive and physiological markers of emotional awareness in chimpanzees (Pan troglodytes). Animal Cognition 4, 223–229.
Plutchik, R., 1987. Evolutionary bases of empathy. In: Eisenberg, N., Strayer, J. (Eds.), Empathy and Its Development. Cambridge University Press, New York, pp. 37–46.
Pulvermuller, F., 2005. Brain mechanisms linking language and action. Nature Reviews Neuroscience 6, 576–582.
Rakel, D.P., Hoeft, T.J., Barrett, B.P., Chewning, B.A., Craig, B.M., Niu, M., 2009. Practitioner empathy and the duration of the common cold. Family Medicine 41, 494–501.
Rice, G.E., Gainer, P., 1962. Altruism in the albino rat. Journal of Comparative and Physiological Psychology 55, 123–125.
Romero, T., Castellanos, M.A., de Waal, F.B.M., 2010. Consolation as possible expression of sympathetic concern among chimpanzees. PNAS 107, 12110–12115.
Roth-Hanania, R., Davidov, M., Zahn-Waxler, C., 2011. Empathy development from 8 to 16 months: early signs of concern for others. Infant Behavior and Development 34, 447–458.
Ruby, P., Decety, J., 2003. What you believe versus what you think they believe? A neuroimaging study of conceptual perspective taking. European Journal of Neuroscience 17, 2475–2480.
Ruby, P., Decety, J., 2004. How would you feel versus how do you think she would feel? A neuroimaging study of perspective taking with social emotions. Journal of Cognitive Neuroscience 16, 988–999.

Sambo, C.F., Howard, M., Kopelman, M., Williams, S., Fotopoulou, A., 2010. Knowing you care: effects of perceived empathy and attachment style on pain perception. Pain 151, 687–693.

Schaefer, M., Xu, B., Flor, H., Cohen, L.G., 2009. Effects of different viewing perspectives on somatosensory activations during observation of touch. Human Brain Mapping 30, 2722–2730.

Schattner, A., 2012. Who cares for empathy? Quaterly Journal Medicine 105, 287–290.

Shackman, A.J., Salomons, T.V., Slagter, H., Fox, A.S., Winter, J.J., Davidson, R.J., 2011. The integration of negative affect, pain and cognitive control in the cingulate cortex. Nature Review Neuroscience 12, 154–167.

Singer, P., 1986. The Expanding Circle: Ethics and Sociobiology. Princeton University Press, Princeton, NJ.

Singer, T., Seymour, B., O'Doherty, J., et al., 2004. Empathy for pain involves the affective but not the sensory components of pain. Science 303, 1157–1161.

Stone, V., 2006. The moral dimensions of human social intelligence. Philosophical Explorations 9, 55–68.

Tabibnia, G., Lieberman, M.D., Craske, M.G., 2008. The lasting effect of words on feelings: words may facilitate exposure effects to threatening images. Emotion 8, 307–317.

Uchino, B.N., Cacioppo, J.T., Kiecolt-Glaser, J.K., 1996. The relationship between social support and physiological processes: a review with emphasis on underlying mechanisms and implications for health. Psychological Bulletin 119, 488–531.

Vaish, A., Warneken, F., 2012. Social-cognitive contributors to young children's empathic and prosocial behavior. In: Decety, J. (Ed.), Empathy – From Bench to Bedside. MIT Press, Cambridge, pp. 131–146.

Warneken, F., Hare, B., Melis, A.P., Hanus, D., Tomasello, M., 2007. Spontaneous altruism by chimpanzees and young children. PLoS Biology 5, e184.

Watt, D.F., 2000. The centrencephalon and thalamocortical integration: neglected contributions of periaqueductal gray. Emotion and Consciousness 1, 93–116.

Wilson, E.O., 1975. Sociobiology: The New Synthesis. Belknap, Cambridge, MA.

Empathy during early Childhood Across Cultures, Development of

Elena Geangu, Fylde College, Lancaster University, Lancaster, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, we will highlight what is currently the most accepted definition of empathy, as well as definitions of similar aspects that are relevant for our understanding of empathy but which are conceptually different. This will be followed by a discussion of the underlying mechanisms of affect sharing, the specific component of empathy. In the last sections, we will present what we know about empathy development during the very first 2 years of life across cultures.

Our abilities to cooperate and understand others are most likely key to our success as a species. Among these, the ability to empathize with others is quintessential. Empathy is a multifaceted and complex psychological construct encompassing many subprocesses that have different developmental trajectories. During development, these components become progressively integrated, and some of the incipient forms are likely to disappear and not be observable at later ages. Hence, investigating and understanding early empathy development are invaluable. Looking at the very beginnings of empathy allows us to understand how the components of a system interact in ways that are not possible in adulthood, when the components are fully mature and functional (de Haan and Gunnar, 2009). Understanding how this ability develops is highly relevant for promoting adaptive social development and for an early diagnosis and intervention in the case of those pathologies characterized by atypical manifestations of empathy, like autism and psychopathy.

Defining Empathy and Related Concepts

After a brief overview of the past decade of literature on empathy (development or otherwise), one quickly becomes puzzled by the multitude of related but somewhat different concepts: empathy, cognitive empathy, emotional empathy, personal distress, emotional contagion, sympathy, compassion, altruism, prosociality, helping, and cooperation, to mention just a few. This multitude of concepts reflects scientists' attempts to differentiate between many aspects of a complex phenomenon. While this differentiation is useful for unrevealing such complexity and guiding research into the human ability to respond empathically to others and behave prosocially, clear definitions are necessary in order to avoid interchangeable use of the concepts and conceptual ambiguity, as well as an adequate approach into identifying the underlying mechanisms. The rise of social cognitive (and affective) neuroscience has been particularly responsible for such a movement toward formulating narrower definitions of empathy and related concepts, mostly due to its aims of establishing clear psychological and functional inferences about underlying processes of social behavior (Ochsner, 2007).

Recently, there seems to be a consensus toward defining empathy primarily as an emotional response (De Vignemont and Singer, 2006) triggered by observing or imagining another individual's affective state. Unlike other emotions, empathic responses are isomorphic with observed ones, and it appears as if we 'share' the affect we see in others, an experience that others have referred to as 'feeling with' the other. The isomorphism in empathic responses mostly refers to its valence, implying also that we can empathize with both the plight and the joy of another (Light et al., 2009). 'Feeling with' the other requires some level of awareness of the difference between our own emotions and those observed in another (Eisenberg, 2000). Depending on the intensity of the response, empathy can further lead to sympathy and motivate prosocial behavior or to personal or empathetic distress (Eisenberg and Eggum, 2009; Zahn-Waxler and Van Hulle, 2012). When not sufficiently regulated, these feelings with the other can be too intense and distressing for the observer, leading to withdrawal and a focus on the self rather than the other and hence being a potential precursor for nonadaptive and pathological responses (Zahn-Waxler and Van Hulle, 2012). Sometimes, in the literature the term 'empathy' is further divided into emotional and cognitive empathy. Emotional empathy refers mainly to the affective component, our ability to feel what others are feeling, while cognitive empathy refers to our ability to attribute mental states to others. As we will see Section Mechanisms of Affect Sharing, attributing mental states to others can be an important process involved in generating empathic emotional responses. However, it has different neural underpinnings from those functionally associated with affect sharing (De Vignemont and Singer, 2006), and it can be present outside the context of empathy response. We can, for example, try to understand what our friend is feeling without experiencing any emotion ourselves. Emotional contagion is another concept that is sometimes used to refer to empathy. It refers to the tendency to automatically mimic and synchronize facial expressions, vocalizations, postures, and movements with those of another person (Hatfield et al., 1994). Like the ability to attribute mental states, emotional contagion is an important process that could be involved in generating empathic responses (De Vignemont and Singer, 2006) but could also operate independently of such a context. For example, we can automatically mimic the emotional expressions of a person we interact with as a means of acknowledging that we have understood his or her emotions but without necessarily feeling the affect whose expression we mirror (Hess and Blairy, 2001).

When efficiently regulated, negative empathic responses can lead to feelings of concern or sympathy for the well-being

of the other, which represent a significant motivator for adaptive prosocial behaviors like helping and altruistic compassion (Knafo et al., 2008). Sympathy has always a negative valence, but does not need to be isomorphic with the observed emotion. A friends' sadness over the loss of a significant other may not lead to sadness in us, but to a great concern for this friend's well-being and a desire to see them well. At the brain level, it seems that sympathy is associated with the activation of the septal area, a brain area that has been functionally linked to the motivation for action (Morelli et al., 2012). Septal area is associated with individuals' prosocial behavior in their natural environment during daily life, and is activated when observing people experiencing various emotions like pain, anxiety, and happiness. Given this motivational value for behaving in the benefit of another, sympathy is also referred to as 'feeling for' the other.

While empathy, sympathy, and empathic distress are all emotional responses, prosocial behaviors like cooperating, sharing, helping, and comforting are acts that can sometimes rely on the motivational value of these emotions, but not always. For example, some forms of prosocial behavior could be the result of respecting preestablished social norms or of the joy associated with performing interesting activities in a social context. Also, some prosocial behaviors can be more altruistic in nature, while others are more instrumental, but most altruistic behaviors have as an ultimate goal the well-being of the other, with no evident benefit for the self and at least some cost to the actor, and in the absence of any external reward (Svetlova et al., 2007).

Mechanisms of Affect Sharing

Usually, three main groups of processes are involved in generating empathic responses: those related to the elicitation of affect sharing, emotion regulation, and self–other awareness. Undoubtedly, affect sharing is the component that is specific for empathy, the one that gives its unique characteristics among other emotional responses.

One of the major questions in empathy research concerns the mechanisms underlying our ability to 'catch' others' emotions. How can we feel what someone else feels when he or she experiences emotions such as sadness or happiness, or bodily sensations such as pain, in the absence of any emotional or sensory stimulation to our own body? Probably the most influential proposal is the perception–action model formulated by Preston and de Waal (2002). According to this model, when the observer attends to another person's emotional states, the observer's representations of similar states are automatically activated together with the associated autonomic and somatic responses, allowing him or her to experience the same emotion. This theoretical model has received some support from neuroimaging studies of empathy for pain, showing that part of the neural networks involved in firsthand experiences of pain are also activated when we simply observe someone in pain (Singer et al., 2004), and also from studies showing activation within the mirror neural system when observing emotions like sadness and happiness (Perry et al., 2012; Rameson and Lieberman, 2009). First discovered in the macaque premotor cortex, mirror neurons are a specific type of neurons that fire both when a monkey performs a hand action itself and when it merely observes another monkey or a human performing the same hand action (Rizzolatti et al., 1996), and it has been suggested that this is the main mechanism involved in imitative behavior. In humans, the same neural network involved in execution of an action (the supplementary motor area (SMA), pre-SMA, premotor cortex, supramarginal gyrus, intraparietal sulcus, and superior parietal lobe) is also activated when merely observing someone else performing the same action (Grezes and Decety, 2001). For what concerns affect sharing, it has been proposed that mirroring mechanisms are automatically activated when we observe others emoting. Both phenomenological observations and empirical evidence suggest that this is not always the case. We do not empathize with everyone we see or think of, and not always. This can actually have an adaptive value. If, for example, others' emotional signals are confusing, it would be not efficient for us to empathize with such an ambiguous state.

Appraisal of other contextual information like the personality characteristics of the observed person (e.g., fairness) or their group affiliation (e.g., in- vs out-group) has been shown to modulate the activation of the neural network involved in empathy for pain (Lamm et al., 2011; Singer and Steinbeis, 2009) and speak against a fully automatic bottom-up driven empathy. Although mimicry mechanisms could be involved in generating affect-sharing responses, for some type of stimuli the affect programs could be triggered by the perception of emotional stimuli independent of conscious visual recognition and even when the emotion is communicated via other motor means than those through which it is reproduced. Tamietto and De Gelder (2008) have shown that when adults are presented very briefly with bodily expressions of fear, their facial muscles specific for expressing fear become activated as in expressing the emotion, even though they have not seen the actual facial expression. This speaks against a perception–action mechanism, as the observer reproduces an expression that has not been perceived. It could thus be that for some emotional stimuli like those signaling danger, affect sharing is the result of triggering affect programs by the mere perception of the emotional stimuli.

There are situations in which the observable cues about the emotions that others experience are missing, either because we are not in direct contact with them or because these cues are ambiguous. In these instances, top–down processes like inferences about mental states (both cognitive and emotional) are involved in generating affect-sharing responses. Rameson and Lieberman (2009) have asked adults to observe people displaying various emotions and also to try to empathize with them while their brain activity was recorded. Participants' subjective reports of felt empathy were associated with increased activation in the brain areas involved in inferring the mental states of others (the medial prefrontal cortex and precuneus) and recognizing emotional expressions (the ventrolateral prefrontal cortex) alongside parietal mirror regions. This activity was particularly stronger when participants made an effort to empathize with the observed others, and also when certain emotions were more ambiguous (e.g., anxiety or sadness) and required reference to contextual cues.

Development of Empathy in the First Years of Life

Although the literature documenting the development of empathy in the first 2 years of life is scarce and many questions regarding the underlying mechanisms are yet to be answered, most current evidence converges to suggest that the ability to respond empathically to others' emotions is present very soon after birth.

Negative Empathy

When being exposed to vocal expressions of distress, like the cry of another infant, 18- to 72-hour-old newborns manifest facial and vocal negative emotions themselves, associated with increases in arousal as indexed by heart rate (Sagi and Hoffman, 1976; Simner, 1971). Given that initially newborns were found to cry whenever hearing another infant crying, these reactions have been suggested to indicate automatic and undifferentiated emotional contagion, akin to emotional mimicry, rather than regulated empathic responses associated with some awareness of self as different from the perceived other. This interpretation was supported by the view that newborns and young infants are not proficient in regulating their emotions, and do not have an explicit representation of self. However, we now know that from birth, infants have at least some incipient abilities to regulate the arousal they experience. They also have an implicit sense of self based on multisensory integration of information about their own body (Rochat and Striano, 2001). These abilities are highly relevant for infants' empathic responses, and the existent evidence suggests that they are already involved to some extent in generating affect-sharing responses in newborns. For example, newborns do not cry or manifest facial distress so intensely when exposed to their own prerecorded cry or when presented with the cry of a newborn from a different species (i.e., a newborn chimpanzee), suggesting that these negative affect-sharing responses are empathic in nature (although more empathic distress) rather than undifferentiated emotional contagion.

Later, during the first year of life, similar empathic responses persist, with 2-week- to 24-month-old infants manifesting negative emotional responses to another in distress (Field et al., 2005; Geangu et al., 2010; Nichols et al., 2009; Roth-Hanania et al., 2011). The cues to others' emotions that can elicit such responses are varied, including auditory (most often, a cry sound), visual (facial and bodily expressions of distress), and contextual. These affect-sharing responses are associated with stable levels of implicit self–other awareness, and infants engage age-specific strategies for regulating these empathic responses (Geangu et al., 2011a). For example, while younger 3-month-old infants use more rudimentary behaviors known to have a regulatory function, like thumb sucking, in order to delay or reduce in intensity their responses, older 6-month-old infants use more advanced attentional strategies known to become increasingly efficient at this age. Infants' success in regulating these affect-sharing responses to an adaptive level, and hence to manifest negative empathy rather than empathic distress, seems to be dependent on the intensity and the duration of the emotional cues they are exposed to (Roth-Hanania et al., 2011). Although no systematic investigation has tested this proposition, existent empirical data suggest that infants' attempts to regulate even the more intense emotions they observe in others are successful depending on their duration. For example, 3- to 9-month-old infants start crying only after they have heard another infant crying for approximately 2–3 min (Geangu et al., 2011a).

At the age of 9 months, infants' ability to discriminate between their own body and that of a peer is associated with a reduced empathic distress toward another infant's cry (Geangu et al., 2011a). This bodily self-awareness is based on an implicit intermodal body schema that integrates information related to bodily space, intermodal timing, and intermodal forms of action (Rochat and Striano, 2001), and probably allows infants to distinguish between experiences that are genuine to their own body and those of others. Toward the age of 9 months, infants become better able to differentiate between self and others more than in terms of bodily sensations. They begin to understand that others have different intentions from their own, intentions that motivate their behaviors (Tomasello, 1995).

Infants' increasing abilities to experience varied emotions are probably also relevant to their diversity of responses to others' emotions. For example, sadness is an emotion that is usually displayed later in the first year of life, and this disposition is related to how much infants respond with sad facial expressions to the cues of a peer in distress (Geangu et al., 2011a).

Positive Empathy

Infants are able to resonate not only to others' negative emotions but also to their positive ones. Although evidence in this respect is limited, it is highly informative. Young 2-month-old infants respond with matched facial expressions of happiness when their mothers display vocal and facial expressions of positive emotions (Haviland and Lelwica, 1987). Similar patterns of results have been recorded in 4- to 6-month-old and 7- to 9-month-old infants, who responded with positive behaviors to a happy female face and with negative behaviors to an angry face (Serrano et al., 1995). While these emotional expressions may represent just the use of adequate social interaction means, they may also indicate empathic-like responses. When the cues for social interaction were reduced, Geangu et al. (2011b) found that both 6- and 12-month-old infants show an increased emotional arousal, as indexed by increased pupil dilation, for another's joy but for shorter duration than for perceived distress. The increase in arousal to positive emotions compared to neutral states has a longer delay for 6- than for 12-month-old infants. When such differentiation occurs, it follows a pattern specific for a negativity bias, with increased arousal for negative compared to positive emotions. This is similar to adult brain activation during empathy for positive and negative emotions. While similar networks are activated in both cases, during negative empathy their level of activation is higher (Perry et al., 2012).

The mechanisms underlying infants' abilities to share others' positive and negative emotions and ultimately feel empathy toward them are still issues of debate. Until recently, the most influential hypothesis was the one formulated by Andrew Meltzoff (2002) which relied heavily on the principles of

simulation theories. It assumes that infant empathic responses have their origins in infants' innate abilities to imitate, to map the actions of other people onto actions of their own body. Most of the supporting evidence comes from studies suggesting that newborns can imitate simple orofacial acts like mouth opening, pouting, and tongue protrusion, as well as some manual acts like index finger extension and sequential finger movements (Jones, 2009). However, empirical evidence has emerged to suggest that alternative mechanisms like exploration-associated arousal could explain such effects, while the same effects observed in newborns have not been replicated with older infants (Jones, 2009; Ray and Heyes, 2011). Evidence is even less compelling regarding newborns' and infants' abilities to mimic emotional facial expressions, which raises serious doubts that the affect-sharing responses we observe in infants are actually mediated by emotional mimicry mechanisms. Obviously, more investigation is needed before drawing conclusions regarding the mechanisms of affect sharing, at least for the first year of life. At the moment, it seems that alternative views, including those referring to subcortical routes of emotional processing (probably specific for negative emotions), are equally plausible (De Gelder et al., 2004).

Sympathy

Although both infants and adults seem to respond empathically stronger to negative rather than to positive emotions, the fact that infants are able to empathize with others' joy is of particular interest. Older children who empathize with another's joy are more likely to show an increased frontopolar activation known to be engaged in experiencing positive emotions (Light et al., 2009).

It has been suggested that this increased frontopolar activity enables children to integrate information about their own emotional states with the information about other's emotional states (Light et al., 2009). It may be thus possible that the empathic happiness we record in older infants could index, among other things, more engagement of the frontal areas of the brain. As suggested in this article, frontal areas of the brain are recruited in many social cognitive processes. At the behavioral level, we do have evidence that older infants and toddlers do engage more in trying to understand the situation that led to another's distress and their emotions (Roth-Hanania et al., 2011), and that this ability, and not the affect sharing per se, keeps developing beyond infancy and toddlerhood (Knafo et al., 2008). Infants and toddlers' abilities to engage social cognitive processes in understanding another's plight are strongly related to their ability to manifest sympathetic concern and with the emergence of their other-oriented prosocial behavior (Nichols et al., 2009; Roth-Hanania et al., 2011). Already at the age of 8 months, infants manifest clear, although moderate, signs of sympathetic concern for another's distress. These gradually increase toward the age of 20 months and characterize children's responses across various situations, involving both family members and strangers (Knafo et al., 2008; Roth-Hanania et al., 2011). Importantly, children who tend to manifest sympathetic concern at an early age show this tendency consistent across the following years (Knafo et al., 2008). Although sympathetic concern is present to a certain degree very early (at the age of 8 months), it seems to become an important motivator for prosocial behaviors only beginning with the second year of life; this is probably dependent on the development by children of a repertoire of adequate behaviors and their ability to functionally link these with the needs and desires of their recipient (Brownell et al., 2009). It is at first more prominent for a child's familiar environment, as children seem to more readily help their mother than a stranger. But, by their third birthday, sympathetic concern motivates prosocial behavior in more varied contexts, including those involving strangers (Knafo et al., 2008).

Cross-Cultural Aspects of Empathy Development

Although children's tendencies to manifest sympathetic concern during their first 3 years of life are to a large extent due to inherited factors, the social environment is equally important and seems to shape the expression of such genetic components (Knafo et al., 2008). That empathy becomes in adulthood strongly linked to social context has received strong support from both behavioral and neuroimaging studies that showed, for example, that adults who watch someone in pain will manifest more empathy if the person they see is a member of their own group rather than one belonging to an adverse one (Hein et al., 2010).

The development of many of the underlying mechanisms of empathy and prosocial behaviors has been shown to be sensitive to the cultural environment in which children live. Parents and siblings are important socialization agents at earlier ages. It is thus expected that the final product, the complex empathic response, will also be shaped by culture.

Infants' abilities to express and regulate emotions develop on different trajectories depending on the culture in which they grow up. When exposed to frustrating or painful situations, young infants from East Asian cultures like Japan manifest longer latencies to the first signs of distress and are more quickly soothed than the North American ones (Camras, 1992), even though they manifest higher arousal (Camras et al., 1998). Such differences in the early development of the ability to express emotions can be even observed within the same macro-cultural environment. For example, Chinese infants manifest less facial motility specific to emotional expressions in response to frustrating events than Japanese infants.

The way in which parents interact with their own children, including their socialization practices, is shaped by the traditions and norms of the culture in which they live. For example, the mothers in the Nso tribes in Cameroon carry their infants closely tied to their bodies and do not engage much in face-to-face interaction or encourage them to explore the objects in their close environment. This seems to be related to a relatively delayed development in children's explicit representations of self as indexed by the mirror self-recognition test when compared to children from Western countries (Keller et al., 2004). Empathic emotional responses are, by definition, associated with a minimal awareness of the difference between our own emotions and bodily sensations and those observed in another. Yet we do not know to which extent

this close bodily contact influences the development of all levels of body awareness beyond explicit representations and whether they have an impact on their ability to express empathic responses. If this is indeed the case, could it be that Nso infants are more prone to less self–other differentiated affect sharing and empathy? Societies like Cameroonian Nso tribes have communal social structures in which people live in close bodily proximity. In this case, higher tendencies to respond with a low level of self–other differentiated affect sharing may increase the sensitivity to many others' emotional cues and help build multiple attachment bonds (Kärtner et al., 2010; Keller, 2012).

See also: Meltzoff's "Like Me" Theory; Mirror Neurons, Theory of; Neural Correlates of Social Development; Prosocial Behavior During Adolescence; Prosocial Behavior During Childhood and Cultural Variations; Social Competence During Adolescence Across Cultures.

Bibliography

Brownell, C.A., Svetlova, M., Nichols, S., 2009. To share or not to share: when do toddlers respond to another's needs? Infancy 14, 117–130.

Camras, L.A., 1992. Expressive development and basic emotions. Cognition & Emotion 6, 269–284.

Camras, L.A., Oster, H., Campos, J., Campos, R., Ujiie, T., Miyake, K., Wang, L., Meng, Z., 1998. Production of emotional facial expressions in European American, Japanese, and Chinese infants. Developmental Psychology 34, 616–628.

De Gelder, B., Snyder, J., Greve, D., Gerard, G., Hadjikhani, N., 2004. Fear fosters flight: a mechanism for fear contagion when perceiving emotion expressed by a whole body. Proceedings of the National Academy of Sciences of the United States of America 101, 16701–16706.

De Vignemont, F., Singer, T., 2006. The empathic brain: how, when and why? Trends in Cognitive Science 10, 435–441.

Eisenberg, N., 2000. Emotion, regulation, and moral development. Annual Review of Psychology 51, 665–697.

Eisenberg, N., Eggum, N.D., 2009. Empathic responding: sympathy and personal distress. In: Decety, J., Ickes, W. (Eds.), The Social Neuroscience of Empathy. MIT Press, Cambridge, pp. 71–83.

Field, T., Nadel, J., Hernandezreif, M., Diego, M., Vera, Y., Gil, K., Sanders, C., 2005. Depressed mothers' infants show less negative affect during non-contingent interactions. Infant Behavior and Development 28, 426–430.

Geangu, E., Benga, O., Stahl, D., Striano, T., 2010. Contagious crying beyond the first days of life. Infant Behavior and Development 33, 279–288.

Geangu, E., Benga, O., Stahl, D., Striano, T., 2011a. Individual differences in infants' emotional resonance to a peer in distress: self-other awareness and emotion regulation. Social Development 20, 450–470.

Geangu, E., Hauf, P., Bhardwaj, R., Bentz, W., 2011b. Infant pupil diameter changes in response to others' positive and negative emotions. PloS One 6, e27132.

Grezes, J., Decety, J., 2001. Functional anatomy of execution, mental simulation, observation, and verb generation of actions: a meta-analysis. Human Brain Mapping 12, 42004.

de Haan, M., Gunnar, M.R., 2009. Handbook of Developmental Social Neuroscience. Guilford Press, New York.

Hatfield, E., Cacioppo, J., Rapson, R.L., 1994. Emotional contagion. Cambridge University Press, New York.

Haviland, J., Lelwica, M., 1987. The induced affect response: 10-week-old infants' responses to three emotional expressions. Developmental Psychology 23, 97–104.

Hein, G., Silani, G., Preuschoff, K., Batson, C.D., Singer, T., 2010. Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. Neuron 68, 149–160.

Hess, U., Blairy, S., 2001. Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy. International Journal of Psychophysiology 40, 129–141.

Jones, S.S., 2009. The development of imitation in infancy. Philosophical Transactions of the Royal Society of London 364, 2325–2335.

Kärtner, J., Keller, Heidi, Chaudhary, N., 2010. Cognitive and social influences on early prosocial behavior in two sociocultural contexts. Developmental Psychology 46, 905–914.

Keller, H., 2012. Attachment and culture. Journal of Cross-cultural Psychology 44, 175–194.

Keller, H., Yovsi, R., Borke, J., Kärtner, J., Jensen, H., Papaligoura, Z., 2004. Developmental consequences of early parenting experiences: self-recognition and self-regulation in three cultural communities. Child Development 75, 1745–1760.

Knafo, A., Zahn-Waxler, C., Van Hulle, C., Robinson, JoAnn L., Rhee, S.H., 2008. The developmental origins of a disposition toward empathy: genetic and environmental contributions. Emotion 8, 737–752.

Lamm, C., Decety, J., Singer, T., 2011. Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. NeuroImage 54, 2492–2502.

Light, S.N., Coan, J.A., Zahn-Waxler, C., Frye, C., Goldsmith, H.H., Davidson, R.J., 2009. Empathy is associated with dynamic change in prefrontal brain electrical activity during positive emotion in children. Child Development 80, 1210–1231.

Meltzoff, A.N., 2002. Elements of a developmental theory of imitation. In: Meltzoff, A.N., Prinz, W. (Eds.), The Imitative Mind: Development, Evolution, and Brain Bases. Cambridge University Press, Cambridge, pp. 19–41.

Morelli, S.A., Rameson, L.T., Lieberman, M.D., 2012. The neural components of empathy: predicting daily prosocial behavior. Social Cognitive and Affective Neuroscience. http://dx.doi.org/10.1093/scan/nss088. Epub Sep 29.

Nichols, S.R., Svetlova, M., Brownell, C.A., 2009. The role of social understanding and empathic disposition in young children's responsiveness to distress in parents and peers. Cognition, Brain, Behavior: An Interdisciplinary Journal 13, 449–478.

Ochsner, K.N., 2007. Social cognitive neuroscience and future promise. In: Kruglanksi, A., Higgins, E.T. (Eds.), Social Psychology: A Handbook of Basic Principles. Guilford Pres, New York, pp. 39–66.

Perry, D., Hendler, T., Shamay-Tsoory, S.G., 2012. Can we share the joy of others? Empathic neural responses to distress vs joy. Social Cognitive and Affective Neuroscience 7, 909–916.

Preston, S.D., de Waal, F.B.M., 2002. Empathy: its ultimate and proximate bases. Behavioral and Brain Sciences 25, 1–71. http://dx.doi.org/10.1017/S0140525X02000018.

Rameson, L.T., Lieberman, M.D., 2009. Empathy: a social cognitive neuroscience approach. Social and Personality Psychology Compass 3, 94–110.

Ray, E., Heyes, C., 2011. Imitation in infancy: the wealth of the stimulus. Developmental Science 1, 92–105.

Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D., Fazio, F., 1996. Localization of grasp representation in humans by PET: 1. Observation versus execution. Experimental Brain Research 111, 246–252.

Rochat, P., Striano, T., 2001. Perceived self in infancy. Child Development 23, 513–530.

Roth-Hanania, R., Davidov, M., Zahn-Waxler, C., 2011. Empathy development from 8 to 16 months: early signs of concern for others. Infant Behavior and Development 34, 447–458.

Sagi, A., Hoffman, M.L., 1976. Empathic distress in the newborn. Developmental Psychology 12, 175–176.

Serrano, J.M., Iglesias, J., Loeches, A., 1995. Infants' responses to adult static facial expressions. Infant Behavior and Development 18, 477–482.

Simner, M.L., 1971. Newborn's response to the cry of another infant. Developmental Psychology 5, 136–150.

Singer, T., 2006. The neuronal basis and ontogeny of empathy and mind reading: review of literature and implications for future research. Neuroscience and Biobehavioral Reviews 30, 855–863.

Singer, T., Steinbeis, N., 2009. Differential roles of fairness- and compassion-based motivations for cooperation, defection, and punishment. Annals of the New York Academy of Science 1167, 41–50.

Svetlova, M., Nichols, S.R., Brownell, C.A., 2007. Toddlers' prosocial behavior: from instrumental to empathic to altruistic helping. Child Development 81, 1814–1827.

Tamietto, M., De Gelder, B., 2008. Emotional contagion for unseen bodily expressions: evidence from facial EMG. 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, 1–5.

Tomasello, M., 1995. Joint attention as social cognition. In: Moore, C., Dunham, P.J. (Eds.), Joint Attention: Its Origin and Role in Development. Lawrence Erlbaum, Hillsdale, NJ, pp. 103–130.

Zahn-Waxler, C., Van Hulle, C., 2012. Empathy, guilt, and depression. In: Oakley, B., Knafo, A. (Eds.), Pathological Altruism. Oxford University Press, New York, pp. 321–344.

Empirical Legal Studies

Christopher Zorn, Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Empirical legal studies (ELS) is a burgeoning field of scholarly inquiry situated at the intersection of law and the social sciences. The contemporary ELS movement began in the 1990s, and was primarily the result of an interdisciplinary group of scholars working in U.S. law schools. ELS is broadly characterized by a commitment to a positivist epistemology, an emphasis on quantitative methods of inquiry, and a substantive focus on matters of legal policy. By the first decade of the twenty-first century, ELS had created a professional association, an annual meeting, a quarterly peer-reviewed journal, and a widely viewed weblog.

Introduction

It is tempting to begin any discussion of empirical legal studies (ELS) with Justice Holmes' famous line from *The Common Law*, where he notes that "(T)he life of the law has not been logic: it has been experience." Holmes' quote claims both too little and too much. On the one hand, the law is clearly much more than its operation. At the same time – as any good historian knows – not all experiences contribute equally to the evolution of society; the 'experience' that mattered to Holmes in *The Common Law* was not comprehensive, and was in fact a different (and substantially more attenuated) version than that deemed influential by contemporary scholars.

In a similar fashion, 'ELS' is both broader and narrower than one first encountering the term is likely to realize. For those steeped in one of the traditional social sciences, it is easy to think of ELS in the confines of one's own discipline; upon seeing the broader, interdisciplinary picture of the subject, one is often struck by the scope and diversity of social scientific work on legal phenomena. In the early twenty-first-century United States, however, the term 'ELS' also comprises a well-defined movement in the legal and social scientific academy. While the sense of ELS as a movement is the central focus of this article, I hope also to make clear how the movement relates to the broader meaning of the term.

Origins

In its contemporary sense, ELS refers most directly to a professional association (the Society for Empirical Legal Studies, or SELS), the primary activities of which are to publish a peer-reviewed journal (the *Journal of Empirical Legal Studies*, or *JELS*) and to host an annual conference (the Conference on Empirical Legal Studies, or CELS). ELS was not the first coordinated effort on the part of a group of scholars to address questions of the law empirically. It was preceded most recently by the law and society movement of the 1960s and 1970s (as constituted, inter alia, by the Law and Society Association (LSA), the *Law and Society Review* (*LSR*), and the National Science Foundation's Law and Social Sciences program), by the law and economics movement begun in the 1950s, and – before that – by the legal realists of late nineteenth and early twentieth century. In challenging conventional approaches to the study of the law, and in demonstrating the value of empirical (rather than purely theoretical) analyses to that study, those influential enterprises laid the groundwork for ELS.

Existing accounts date the origins of the contemporary ELS movement to the decade of the 1990s. Suchman and Mertz, for example, note that "Harvard opened what may have been the nation's first self-proclaimed Program on Empirical Legal Studies" in 1996 (2010: 556–557), and Kritzer's review of more than a century of empirical work on the law dates the term's widespread use to "around 2000" (2009: 925). While the exact instant of the movement's origins probably cannot be known, what is clear is that, beginning in the last decade of the twentieth century, scholars in the social sciences and the legal academy who were investigating legal questions using empirical methods began to coalesce into a new organization.

The roots of this new movement were different in many respects from those of its predecessors. Whereas the institutionalization of law and economics was predominantly (and unsurprisingly) the task of economists, and the origins of law and society were largely the disciplines of sociology and criminology, ELS's founders were in a disciplinary sense more diverse, hailing from political science, psychology, economics, public policy, and a smattering of other fields. Chambliss notes that "(A)mong the eight founders of SELS, four have Ph.D.s: one in economics, one in psychology, one in education and social policy, and one in political science" (2008: 32–33). And as Eisenberg (2011) notes, in addition to their close disciplinary ties, both LSA and the American Law and Economics Association owe their origins at least in part to funding by external organizations: the former, to the Russell Sage and Walter E. Meyer Foundations, and the latter to the Olin Foundation. In contrast, SELS was founded without either a strong disciplinary affiliation or any external foundation support; instead, as Eisenberg (2011) details, it received initial support from law schools, especially Cornell. As a result, it lacked both the clear disciplinary and ideological orthodoxies of other similar associations.

At the same time, the ELS movement was also narrower than those organizations in other respects. It was, at least initially, almost entirely made up of faculty with appointments in elite law schools, rather than in the social and behavioral

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86135-8

sciences. It was also almost exclusively positivist in its epistemological outlook, quantitative in its methodological approach, and policy-oriented in its substantive focus. In an influential early essay, Michael Heise (one of the founders of ELS) defined ELS as "the subset of empirical legal scholarship that uses statistical techniques and analyses" (Heise, 2002: 821). Similarly, the editor's introduction to the first issue of *JELS* noted that "The central purpose of *JELS* is to add to knowledge of the legal system based on observation or empirical analysis, including experimental analysis," and that "*JELS*'s editorial board is committed to providing expert statistical evaluation of manuscripts with the goal of raising the level of statistical sophistication in empirical legal scholarship." While this emphasis has changed somewhat over time, the predominant methodological tools adopted by ELS's adherents remain quantitative (primarily statistical studies of observational data, but also laboratory and field experiments), and the central topics addressed by research appearing at CELS and in *JELS* are of a policy nature, albeit those familiar to a legal audience.

As a professional association, ELS began with the creation of *JELS* at Cornell University Law School, under the editorship of Professors Theodore Eisenberg, Jeffrey Rachlinski, Stewart Schwab, and Martin Wells. The first issue of *JELS* was published in 2004. Shortly thereafter, a group of organizers hosted the first CELS at the University of Texas – Austin in October 2006. That conference drew more than 200 participants, and was followed by annual meetings at NYU (2007) and Cornell (2008). Subsequent CELS gatherings were hosted by USC, Yale, Northwestern, Stanford, the University of Pennsylvania, and UC–Berkeley. In February of that same year (2006), an affiliated group of scholars launched the Empirical Legal Studies Weblog. Soon after that initial conference, Eisenberg and seven others founded SELS, an umbrella organization responsible for the administration of *JELS* and CELS. SELS was registered in New York as a federal 501(c)(3) organization in 2011. SELS membership is open; annual membership dues include a subscription to *JELS*.

ELS Today

In the decade or so since its creation, SELS has grown considerably in size and influence. SELS currently has nearly 500 individual members; attendance at the 2013 CELS was 350. Unlike other similar meetings, presentation of research at CELS is refereed and selective; the acceptance rate for papers submitted to the 2012 CELS conference was 28%. *JELS* is published quarterly, and is available at 3480 academic and other institutions worldwide via license, and at another 5200 institutions in the developing world via philanthropic efforts. *JELS*'s 2013 impact factor (the average number of times articles from the journal published in 2011–12 were cited in 2013) was 1.067; Eisenberg's analysis of the Washington and Lee Law Library Database also shows that JELS ranks "in the top five" among social science journals indexed, and "in the top 5%" of all refereed journals considered (2011: 1725 and Table 2). Eisenberg (2011: Table 1) also demonstrates that, at least during the first decade of the twenty-first century, the disciplinary affiliation of *JELS* authors was more disciplinarily diverse than either that of authors in *Law and Society Review* or the *Journal of Legal Studies*, two long-running and prominent journals with their roots in the social sciences. Based on the data reported in Eisenberg's (2011) Table 1, the 2004–10 disciplinary Herfindahl index (a measure of the concentration of authors within disciplines) for JELS was 0.23; the corresponding statistics for *JLS* and *LSR* were 0.70 and 0.97, respectively.

The development of ELS over the past two decades has paralleled (and, to varying degrees, propelled) a more general empirical turn among legal scholars. In 2006, the Association of American Law Schools took as the organizing theme of its annual meeting 'Empirical Scholarship: What Should We Study, and How Should We Study It?' And as numerous scholars have noted, research with an empirical component is now commonplace in law reviews. In one recent example, the *Stanford Law Review* dedicated an issue to 'The Empirical Revolution in Law'; in their introduction to that issue, Ho and Kramer show that the incidence of articles using the word 'empirical' in that journal have risen from effectively none in the 1950s and 1960s to (on average) 60% of all research published there (Ho and Kramer, 2013: Figure 1).

On the pedagogical front, courses on empirical methods (primarily, but not exclusively, quantitative) are now a regular part of the curriculum at many law schools; a current (and very incomplete) list of major law schools offering such courses includes Chicago, Cornell, Harvard, Illinois, Northwestern, Stanford, UCLA, Vanderbilt, Washington University, and Yale. Law schools are also increasingly hiring faculty who hold both a JD and a PhD, and in some instances have made prominent appointments of empirical scholars with a social science PhD alone.

Disciplinary reactions to the rise of ELS – and of empirical work on the law more generally – have varied widely. Among conventional doctrinal legal scholars, responses have varied from enthusiasm through skepticism to outright disdain. The nature of the reactions track broadly with the degree of involvement with empirical work on the law; scholars focusing on areas of the law that are intrinsically amenable to empirical analysis (e.g., antitrust, election law, and so forth) have tended to be more receptive to such work than those working in less empirically relevant areas (such as legal philosophy).

Among the social sciences, disciplinary reactions also generally correlated with the degree of involvement in ELS by members of that discipline. On the one hand, scholars of legal phenomena in political science, economics, and psychology – those fields with substantial representation and involvement in the ELS community – generally welcomed the new organization. While ELS's impact on some other fields (notably anthropology) has been minimal, perhaps the strongest reaction has been among sociologists, particularly those associated with the 'law and society' tradition of sociolegal research. As early as 2006, scholars in those areas wondered aloud whether ELS "will become a more complacent, insular, and technocratic endeavor" than the law and society movement, and whether "the sociology of law, as we know it, may become even more marginal in the legal world than it is today" (Suchman, 2006: 3; see also Suchman and Mertz, 2010: 574–575). Chambliss characterized ELS as "a movement by a group of social-science-trained law professors … to build their

reputations and careers by creating a market demand – 'empirical legal studies' – and filling it" (2008: 31); this was possible because "scientism sells, especially when it is pitched by faculty from prestigious universities" (2008: 36). She recommends a policy of careful disengagement, where "epistemological, theoretical, and political debates" take place in other contexts, leaving ELS "to fight the common enemy of junk science and pseudo-research" (2008: 24).

These responses, both within and outside of the legal academy, are in many respects unsurprising. In addition to their conceptual distance from empirical work, doctrinal scholars in law schools are also the least likely to have the training critically to evaluate empirical work (or the inclination to seek it). In the social sciences, the absence of any substantial 'critical turn' in the fields of psychology, economics, or political science leaves little cause for epistemological disagreement with members of the ELS community. In contrast, sociologists of law and members of the law and society community have embraced that turn, and as a result are deeply – and sometimes irredeemably – skeptical about ELS's central tenets of positivism, quantification, and policy emphasis.

The Future of ELS

The institutionalization of ELS into the contemporary legal and social science academies is largely complete. In addition to the growth of SELS itself, empirical work on legal phenomena of the sort undertaken by ELS members constitutes large and in some cases dominant bodies of research on the law in economics, political science, psychology, and other fields. Empirical scholars are also increasingly integrated into legal education; conversely, legal academics with an empirical bent are finding an increasing willingness to collaborate on the part of scholars in the social sciences.

Research by members of the ELS community has also been notable for its impact. An early review by Eisenberg (2004) noted that ELS research received significant attention from mainstream media; for example, by the time of publication of its first issue, research findings appearing in *JELS* had been featured in the *New York Times*, the *Economist*, the *Financial Times*, the *Wall Street Journal*, and the *International Herald Tribune*. The movement has also had a visible impact in legal education. Beyond offering courses in empirical methods, ELS adherents have called for – and in some cases prompted – changes to the ways in which legal publications are edited (encouraging the adoption either of a system of peer review or of mandatory data disclosure and replication for empirical work appearing in student-edited law reviews).

The popular attention to and increasing emphasis of empirical work in law schools has in turn contributed to the growing use of quantitative data and analytics in the legal services industry, particularly among large U.S. law firms. It has also influenced the development of private-sector companies that provide analytical and empirical expertise to actors in the legal industry, including law firms, in-house legal departments, and law schools. While some of these changes undoubtedly reflect the larger move toward greater and more sophisticated use of data in business, their development in the notoriously slow-to-adapt legal services industry was arguably fostered by the visibility of quantitative, empirical research on legal issues, and on the increased exposure of young attorneys to such research during law school.

Beyond its intended impacts, ELS has also prompted developments in the larger field of social scientific work on law. One of these is the creation of what its adherents term 'New Legal Realism' (NLR), which has been variously characterized as a parallel movement (Eisenberg, 2011; Lempert, 2013), a counterpoint (Suchman and Mertz, 2010), or a competitor (Chambliss, 2008: 24) to ELS. To the extent that it possesses a coherent outlook, NLR exalts 'law in action,' 'bottom-up' approaches to the study of law, and interdisciplinarity. In part because of the hegemonic influence of the older sociological 'law and society' tradition upon it, NLR is also infused with a progressive political agenda that distinguishes it from ELS. No similar countermovements have yet been observed in economics, psychology, political science, or the other social sciences to which ELS bears a connection.

ELS itself has also evolved, at least partly in response to the issues raised by its critics. While the research presented at CELS and published in *JELS* is still predominantly positivist, quantitative, and policy-focused, its institutionalization has allowed the movement and its adherents to become broader and more self-reflective. The fraction of work employing qualitative and mixed methods that appears in *JELS* has increased somewhat since the journal's creation; while maintaining its disciplinary diversity, current issues also exhibit a wider range of methodological approaches. In recent years the movement's leaders have also welcomed critical reflection; for example, the aforementioned *Stanford Law Review* special issue is comprised of a series of essays by prominent non-ELS scholars responding to and critiquing bodies of work presented at the 2012 CELS.

At the same time, the central tenets of the movement remain fundamentally unchanged. The movement's greater acceptance of qualitative and historical methods is consistent with its continuing dedication to a positivist (rather than relativist/constructivist) epistemology. And the movement so far shows no particular inclination to try to "escape the pull of the policy audience" (Suchman and Mertz, 2010: 575). This is also somewhat unsurprising; as Chambliss (2008) notes, the intersection of ELS's three central characteristics – a commitment to a positivist epistemology, an emphasis on quantitative methods of inquiry, and a substantive focus on matters of legal policy – make it particularly well positioned to inform and contribute to policy making.

Going forward, a number of larger societal phenomena point to the continued viability and growth of ELS. In particular, the ever-increasing availability of digital data on socio-legal phenomena, the growing power of computers to visualize and analyze those data, and the rising prominence of legal academics in government and policy circles would all seem to be consistent with the ELS mandate. So too does the increasing demand for evidence-based and data-driven decision-making in the public and private sectors. Together with the increasing demand for empirical research in the legal academy, these trends augur the ongoing vibrancy of ELS as a scholarly endeavor.

See also: Law and Economics: Empirical Dimensions; Law and Society: Development of the Field; Law: History of Its Relation to the Social Sciences; Law: The Socio-Legal Perspective; Legal Positivism; Legal Process and Social Science: United States; Legal Realism; Legal Scholarship.

Bibliography

Chambliss, E., 2008. When do facts persuade? Some thoughts on the market for empirical legal studies. Law and Contemporary Problems 71, 17–40.

Eisenberg, T., 2004. Why do empirical legal scholarship? San Diego Law Review 41, 1741–1746.

Eisenberg, T., 2011. The Origins, Promise, and Nature of Empirical Legal Studies and a Response to Concerns. University of Illinois Law Review, 1713–1738.

Heise, M., 2002. The past, present, and future of empirical and experimental legal scholarship: judicial decision making as a case study. University of Illinois Law Review 4, 819–850.

Ho, D.E., Kramer, L., 2013. Introduction: the empirical revolution in law. Stanford Law Review 65, 1195–1202.

Holmes Jr., O.W., 1881. The Common Law. Little, Brown and Co., Boston.

Kritzer, B., 2009. Empirical legal studies before 1940: a bibliographic essay. Journal of Empirical Legal Studies 6, 925–968.

Lempert, R., 2013. Growing up in law and society: the pulls of policy and methods. Annual Review of Law and Social Science 9, 1–32.

Suchman, M., 2006. Empirical legal studies: sociology of law, or something ELS entirely? Amici 13 (1), 1–4.

Suchman, M.C., Mertz, E.M., 2010. Toward a new legal empiricism: empirical legal studies and new legal realism. Annual Review of Law and Social Science 6, 555–579.

Relevant Websites

http://www.elsblog.org/ – Empirical Legal Studies Weblog.

http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1740-1461 – Journal of Empirical Legal Studies.

http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504727 – Program for Law and Social Sciences, National science foundation.

http://www.lawschool.cornell.edu/sels/ – Society for Empirical Legal Studies.

Empirical Social Research, History of

Hynek Jerabek, Institute of Sociological Studies, Faculty of Social Sciences, Charles University in Prague, Prague, Czech Republic

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Empirical social research (ESR) uses empirical data on social formations and their members as a basis for explaining social phenomena. Historically it developed in three stages. The first stage was represented by the changes introduced into scientific procedure by the Chicago School of Sociology between 1895 and 1929. The second stage began outside the field of sociology in the late 1920s with the use of standardized data on social units (individuals and households) for marketing and polling purposes, and the approach was then applied in sociology starting in the later 1930s; most social data processing at this time were descriptive in nature and purpose. The third stage involved use of the same mass data in the survey analysis model of research to test hypotheses on causal relations between selected variables. The aim was to develop 'theory-based ESR.' ESR refers to empirical research activities grounded in social theory and enriching social theory.

For some decades now, empirical social research (ESR) has been contributing to knowledge through the collection and evaluation of social data and the testing and enriching social theories. Alongside the history of sociological thought and in parallel with the history of social theories we therefore need to study the history of ESR. This will help us to understand how important discoveries were made, and explain the methods by which significant theories were confirmed or challenged.

Definition

'ESR' must be distinguished from the related terms 'survey,' 'social survey,' 'survey research,' and 'survey analysis.' 'Survey' or 'survey research' is a general term for standardized mass questioning of a representative sample of individual members of a population under study. The 'history of the social survey movement' refers to the emergence and development of the use of survey techniques, primarily in the work of British charities in the nineteenth and early twentieth centuries, and then in similar activities in the United States in the first two or three decades of the twentieth century. This research work was conducted mainly in urban conditions, and was aimed at identifying and addressing negative social phenomena such as poverty, bad working conditions, suicide, discrimination, racial segregation, social exclusion, and crime. The concept of 'ESR' refers to survey activity that has clear scientific applications in either basic or applied research, but it also includes scientific activities carried out by social scientists with the same or similar objectives but not using the survey technique. *The history of ESR is thus the history of research activities that have used empirical data on social formations and their members to identify and then offer explanations for social phenomena relating to social groups, communities, social institutions, or social processes.* As we understand the term, ESR generally means 'theory-based ESR,' i.e., research grounded in social theory and yielding results that enrich social theory.

While it may include some activities that did not have social scientific applications, the 'history of survey research' is nevertheless largely a part of the 'history of ESR.' Indeed, the most valuable achievements of the social survey movement, which influenced the very nature of the emergent social sciences in the first decades of the twentieth century, form a very important part of the history of ESR. In the late 1930s and 1940s survey analysis was based on use of an abundant body of data and careful formulation of problems in an effort to discover causal explanations for the social phenomena studied, and as such it pushed ESR toward ever greater applicability in the construction or enrichment of sociological theory.

The Main Trajectories and Stages in the History of ESR

It is possible to distinguish three stages in the institutionalization of ESR. The first stage is most comprehensively represented by the series of changes introduced into scientific procedure by the Chicago School of Sociology between 1895 and 1929. Its new research model made the 'Chicago school' the most important center of scientific sociological research in the United States in its day. The second stage started with the widespread use of standardized data on social units that emerged outside the field of sociology in the late 1920s and came to be applied in sociology in the second half of the 1930s and later. Most of the data processing conducted at that time were of a descriptive nature, and so the third stage began when the same mass data that had been used earlier just for descriptive purposes were employed by the creators of the survey analysis model of research to test hypotheses about the causal relations between the variables studied. (cf. Arguments on the three stages of institutionalization of ESR in Jerabek (2012) and Converse (1987).)

The Origins of the Collection and Analysis of Social Data – Europe

In the period from the seventeenth to the nineteenth century the first steps were taken in the collection and analysis of data on society. The collection and use of such data were driven by two different purposes. One was administrative, and a matter of the increasing demands on the government. The other was statistical and a response to the practical needs of the emerging insurance industry and science; the aim was to predict and explain patterns of mass phenomena in society.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03217-7

The next period, from the 1890s to the end of the World War II, is generally regarded as the time when ESR acquired firmer contours. During this period ESR became fully institutionally established, i.e., it gradually became a legitimate part of academic sociology. This process culminated in the late 1940s in the United States and subsequently in Europe and other parts of the world.

The Administrative Use of Empirical Social Data

The administrative use of empirical social data was connected with government, the drafting of legislation, and territorial administration. This line of activity was based on the work of William Petty, who in 1672 wrote a report on the economic and social structure of Ireland (*The Political Anatomy of Ireland*). The socioadministrative line then progressed in centralized France, where Jean-Baptiste Colbert, controller general of finances, and, after him Marquis de Vauban, collected data necessary for the administration of the country. This stage culminated in a survey describing 32 administrative districts of France, 'The Grande Enquete,' initiated in 1697 (Lecuyer and Oberschall, 1968, p. 38). The line was developed further in the eighteenth century in the work of the Encyclopedistes Jean S. Bailly, Antoine L. Lavoisier, and Pierre S. Laplace, when they reorganized the network of hospitals in France in the years 1785–89. Pierre J.G. Cabanis continued in this spirit in the next century as part of the 'hygiene movement' (Lecuyer and Oberschall, 1968, pp. 39–40). The reform movement in Great Britain from the end of the eighteenth to the middle of the nineteenth century was represented by John Howard (1777), Frederick Morton Eden (1797), J.P. Kay-Suttleworth (1832), and Edwin Chadwick (1842) (Lecuyer and Oberschall, 1968, pp. 40–41). A similar line of research was then supported by the British Parliament, which ordered numerous reports on poverty, the health of the population, child labor, working conditions in factories and so on. Martin Bulmer refers to this trend as *Blue Book Sociology* (Bulmer, 1985b: 6). The data obtained in the preparation of these reports between 1833 and 1897 were used as a basis for legislation in Great Britain. (One of the longest reports, the 'Report on Labor,' drawn up in 1893–97, is remarkable for the enormous amount of material it compiled; it drew on interviews with 583 witnesses, posed 97 336 questions, and distributed 5350 questionnaires, 2100 of which were returned completed (Zeisel, 1974, p. 103).)

Charles Booth is generally considered the most prominent representative of the administrative line of collection of empirical social data. From 1886 to 1897 he organized the collection of a large amount of empirical material as a basis for his study of poverty and the living conditions of London's working class (Bales, 1991). In the course of more than a decade he published the successive volumes of his series *Life and Labour of the People in London* (Booth, 1889, 1891, 1892–1897, 1902–1903). He made it his goal to provide an accurate map of the full scale of poverty in the population of London. Booth decided to undertake a laborious survey (conducted door-to-door, and in subsequent surveys street by street) (Bales, 1991, p. 97) in which he collected empirical data with the aid of a dense network of school inspectors, who were well acquainted with the workers' families (Bales, 1991, p. 74).

Booth was unquestionably a source of inspiration for many other researchers, especially in the United States. The latter carried out their own investigations, for which the term 'social survey' was introduced. Booth's immediate successor in Britain was Benjamin Seebohm Rowntree. In *Poverty. A Study in Town Life* he investigated the population of York and used his results as a basis for a distinction between primary and secondary poverty. Primary poverty meant living below a certain minimum income level, whereas secondary poverty involved the imprudent or uneconomical use of resources, with consequences that were often worse than the effects of primary poverty (Kent, 1985, pp. 52–61; Hennock, 1991, pp. 190–195).

The next studies of similar phenomena, initiated and conducted by Arthur L. Bowley (Bowley and Burnett-Hurst, 1915; Bowley and Hogg, 1925), used sampling procedures in data collection. In 1928, Llevellyn Smith published the *New Survey of London Life and Labour*. This was a replication of the studies done by Charles Booth in London, allowing historical comparison of the new findings with Booth's observations and at the same time a methodological comparison of Booth's exhaustive survey technique and Bowley's 'sampling' procedure (Kent, 1985, pp. 52–59).

The Statistical Use of Empirical Social Data

The purpose and goal of the second line of data collection, i.e., the statistical evaluation of collected empirical data, were to identify 'quantitative social laws' or at least patterns observed in figures on births, deaths, and later women's fertility, and on other mass social phenomena. This line of research was initiated by John Graunt, who published the first 'mortality tables' in *Observations upon the Bills of Mortality* in 1662 (Zeisel, 1974, p. 106; Lazarsfeld, 1982a/1961). He drew attention to certain recurring patterns and uniformities in these tables and thus took the first step toward using mass data to describe some underlying laws in the life of society. What primarily distinguished this line of development was the emphasis on standardized types of information. The data collected on comparable units were identical in structure. So it could be summarized in tables, compared with subsamples, and employed to say something about recurring patterns in observed social phenomena. Initially this line of research was called 'political arithmetic.' The insurance industry needed information on the average length of life in the population. Some programs to meet this demand were developed by the astronomer Edmund Haley, who in 1693 published the very first 'life tables' (on the population of Breslau in Silesia) (Lecuyer and Oberschall, 1968, p. 37). Unlike Graunt's data on London, these data additionally contained information on how old people were when they died. As a result Haley was able to calculate the 'chances of survival to any given age' (Lecuyer and Oberschall, 1968, p. 37).

This line then continued in France, where in 1750–89 the government left the initiative 'to learned societies and private individuals' (Lecuyer and Oberschall, 1968, p. 38). The challenge was accepted by the Encylopedistes. In an entry headed *Arithmétique politique*, Denis Diderot gave a detailed account of the work of William Petty in 1751. Deparcieux published 'mortality tables' in 1746. The abbé Expilly investigated numbers of births in 9000 parish records between 1762 and 1770 and in his *Geographic, Historical and Political Dictionary of*

France-based estimates of the population of France on the number of families, producing a figure of 22 million for the population in France (Lecuyer and Oberschall, 1968, p. 39). These calculations prompted further research by M. Moheau, who produced more accurate demographic calculations in the year 1778 using parish records on the number of births, broken down by sex, and parental status and with consideration of the natural and social conditions influencing maternal fertility (Lecuyer and Oberschall, 1968, p. 39). After the fall of Robespierre the French government showed a revived interest in scientific institutions and in 1795 founded the Institut National des Sciences et des Arts (Lecuyer and Oberschall, 1968, p. 40).

Meanwhile, in the second half of the eighteenth century, the ideas of Herman Conring, who had worked at the University in Helmstedt in Brunswick in the seventeenth century, began to attract attention in Germany. Conring's ideas were initially recorded by his students in Latin as *Collegium Politicalstatisticum*. He was interested in how the state functioned, and offered a description of its economic, administrative and legal system, and formal and informal relations in organizations (Lazarsfeld, 1982b, pp. 110–115). Conring's system was first introduced in the German language by Gottfried Achenwall only many decades later in 1749.

In France Pierre Simon Laplace (*Mechanique céleste* 1798), Jean Baptiste Fourier, Nicolas de Condorcet (*Mechanique sociale*), and others developed the basic elements of the theory of probability. This work motivated the young Adolph Quételet to apply the general rules of probability theory to the analysis of the human population, first in relation to physical size, then to crime rates broken down by region and age of offenders, and later to other social phenomena. Quételet talked of 'moral statistics.'

Adolphe Quételet, starting with his study *Sur l' homme et le développement de ses facultés, ou Essai de physique sociale*, *A Treatise on Man and the Development of His Aptitudes, or Essay on Social Physics* originally published in 1835, in English translation from 1842 (Quetelet, 1835/1842; Sarton, 1935), demonstrated that clear patterns can also be found in social data. He was the first to show the value of comparative statistics and the application of probability to moral and social issues (Eknoyan, 2007, p. 49). In his view there existed 'social laws' as well as natural laws. His measurements of the attributes of the human body provided evidence that mass data on humans could exhibit the same patterns as data in the natural sciences. His statistics on criminal acts classed by the sex and age of offender for individual judicial districts showed similar patterns. The concept of the 'average man' became a meaningful new idea in the social sciences, i.e., that mass data collected on a group of people and statistically processed could yield a mean (average) and margin of variation around the mean as a meaningful statement about that group (Landau and Lazarsfeld, 1968, p. 251).

The German Historical School in Economics as the Inspiration for Empirical Sociology

The German historical school in economics led by Gustav Schmoller made a crucial contribution to the development of empirical sociology. When these ideas crossed the Atlantic they proved to have a major influence, for example, on W.E.B. Du Bois' *The Philadelphia Negro* (Bulmer, 1991, pp. 171–172).

Du Bois studied economics in Berlin under Gustav Schmoller and after completing his doctorate in history at Harvard University carried out a unique study, *The Philadelphia Negro*, of the black population in a slum in Philadelphia. He drew inspiration from the method of Charles Booth, but went further, obtaining standardized data on 2400 black families in a slum in Philadelphia 'face to face' using house-to-house interviews with the inhabitants of the slum alone and without the help of inspectors (Bulmer 1991, pp. 177–179). Du Bois was an Afro-American and therefore enjoyed a high degree of trust from his respondents. In addition to this extraordinary research monograph, between 1898 and 1910 he published 12 independent and collective research studies (altogether more than 2000 pages), which dealt successively with the 10 most pressing problems of the black population in the southern US states. He established the first historically recorded sociological school with a research program at Atlanta University, organized regular scientific conferences, and founded a publication series called *Atlanta University Studies* (Bulmer, 1991, p. 184; Wortham, 2005).

The influence of the economic historical school in Germany was evident in the early stages of sociography in the works of the first systemizers of empirical evidence Gustav Schnapper-Arndt, Adolph Levenstein (Petersen, 2012, pp. 217–224), and Max Weber (Oberschall, 1965).

Another important branch of study based on empirical evidence drawn from the social world was the research of Fréderic Le Play. He based his conclusions about the social world on detailed accounts of individual families and households. He wrote very accurate and thorough monographs on the families of miners and laborers, first in France and later in other countries in Europe (Le Play, 1877–79, 1859–60), and his work has therefore been frequently considered to be the 'foundation of the monographic method.' Data about a specific miner's family, including his family budget, information on his housing, number of children, and the management of the household were all relevant for Le Play.

The Beginnings of ESR in the United States: *The Polish Peasant*, the 'Social Survey' Movement, and the Studies of the Chicago School

In the United States, ESR emerged in direct connection with the transformation of top American universities into institutions of science and research. In sociology, the first and strongest focus on empirical research emerged at the University of Chicago. As early as 1892 a Department of Sociology and Anthropology was founded at Chicago and soon afterward a program of doctoral studies was introduced by the department's head sociologist, Albion W. Small. In 1895 the department began to publish the *American Journal of Sociology* – the very first journal of sociology. In 1905 A.W. Small also cofounded the American Sociological Society (later called the American Sociological Association).

In 1908 William I. Thomas obtained what was at the time the extraordinary sum of 55 000 USD to study migration to the United States. From 1908 to 1914 he made a number of

research trips to Europe. He focused on Polish migrants, and in Poland began working with social psychologist Florian Znaniecki. Together they produced a monumental five-volume study entitled *The Polish Peasant in Europe and in America* (Bulmer, 1984, pp. 45–63). This was to provide models and inspiration for the next generation of Chicago sociologists. W.I. Thomas developed ideas about social attitudes, while F. Znaniecki made an analysis of values. In these volumes Thomas's theory of disorganization and reorganization was applied to an extensive body of empirical evidence in the *Polish Peasant* study and it was also a contribution to methodology. The biographic method used in the story of Wladek Wiszniewski, the analysis of personal documents such as letters of migrants to their families in Poland and from Poles to relatives in America, and Thomas's theory of 'four wishes,' elaborated in the book *The Unadjusted Girl*, all had a formative and long-term influence on empirical sociology at the University of Chicago. In 1938 the *Polish Peasant* study was selected by the sociological community as representative of the field and put forward for special discussion and critical analysis (Blumer, 1939).

In 1914 Robert Ezra Park joined Chicago's Department of Sociology, bringing with him experience as a journalist. In 1916 the *AJS* published his article *The City* in which he presented his vision of Chicago as a social laboratory. He founded a doctoral seminar in which he supervised doctoral students undertaking empirical field work. His closest colleague was Ernest Watson Burgess, whose strength was his concern for methodology and practicability in doctoral studies (Bulmer, 1984, pp. 94–95, 117). In 1921 R.E. Park and E.W. Burgess wrote their *Introduction to the Science of Society*, which was christened the 'green bible' by students and represented an entirely new concept in sociology textbooks. In 1925 they published *The City*, an important collective monograph that, in addition to Park's profile article *The City*, also included E.W. Burgess' theories of the concentric zones and of succession (Bulmer, 1984, pp. 154–155).

In the early twentieth century, sociology in the United States was very frequently associated with the charity work of institutions like the Russell Sage Foundation, Hull House in Chicago, and the Settlement Houses in many large cities on the east coast of the United States. Their main model for the collection and use of social data was Charles Booth and his studies. The dozens of social surveys produced by these institutions, with results of varying quality, were motivated by the desire to combat poverty, eliminate blatant social differences, and above all protect America's middle classes from the criminality and disorder that originated in the slums. Social surveys were the most common type of instrument applied in social work. In most cases social reformers used maps to identify areas in which most crime, poverty, unemployment, suicide, and other negative social phenomena were concentrated, and presented the results of their studies directly to the American public. They were consciously seeking to win public support for the adoption of legislation and measures aimed at reducing the incidence of negative social phenomena. One exceptional study in terms of its quality and its influence on society at the time was *Hull House Maps and Papers*, which was written and organized by Jane Addams and Florence Kelley. A later example was the series of Pittsburgh studies directed by Paul Kellogg (Oberschall, 1972, p. 216nn).

The principal problem of these social surveys was that none of these researchers knew how to statistically process a massive amount of individual data to arrive at demonstrable conclusions about a social group or subgroup under scrutiny. The 'technological' conditions for collecting, processing, and evaluating social survey data would only be developed in later decades.

In the meantime, the sociologists in Chicago who studied the urban population and published scientific papers and the first empirically oriented monographs conducted data collection and made generalizations from their findings without the use of social survey techniques. Focusing on firsthand data about natural social groupings, they were highly skeptical about information from individuals or individual households or families and did not use such data. In cooperation with other social sciences the Department of Sociology progressively established three monograph series, most of them published by the University of Chicago Press: a general social sciences series, a sociological series, and a series specializing in crime and youth. Thanks to the empirical research work of Chicago doctoral students coordinated by Park and Burgess, between 1923 and 1935 two dozen doctoral monographs and other work by graduates of the Chicago School of Sociology were published. The first was Nels Anderson's *The Hobo* (1923), followed by Ernest Mowrer's *Family Disorganization* (1927), Frederic Thrasher's *The Gang* (1927), *Suicide* by Ruth Cavan (1928), *The Ghetto* by Louis Wirth (1928), *The Gold Coast and the Slum* by Harvey W. Zorbaugh (1929), *Delinquency Areas* by Clifford Shaw and Henry D. McKay (1929), *The Jack-Roller* by Clifford Shaw (1930), *The Natural History of a Delinquent Career* and *Brothers in Crime* by Clifford Shaw (1931, 1932), *The Taxi-Dance Hall* by Paul Cressey (1932), and *Vice in Chicago* by Walter C. Reckless (1933). All these studies and other publications testify to the success of the Chicago School of Sociology. It had developed theories to guide its work, planned a sensitive research program, and became the first major school of sociology in the history of the field. (More details and arguments about the role of the Chicago school in the development of ESR can be found in Jerabek (2012).)

Community Studies in the United States and Europe: Middletown, Middletown in Transition, Yankey City Series, Marienthal

The year 1929 saw the publication of the first sociological bestseller in the United States: *Middletown*, by Robert Staughton Lynd and his wife Helen. This study of a small American town became the model for community studies. In 1924 and 1925 Robert and Helen Lynd spent a year in Muncia in Indiana, where they participated in the life of the community. They studied six areas of small-town life: 'getting a living,' 'making a home,' 'training the young,' 'using leisure,' 'engaging in religious practices,' and 'engaging in community activities' (Lynd and Lynd, 1929/1957, p. 4).

They used five different groups of methods and techniques to observe people's lives and study the town: (1) participation in the lives of people in the town; (2) the study of documents; (3) the independent collection and compilation of statistics; (4)

interviews; and (5) questionnaires (Lynd and Lynd, 1929, pp. 507–512). Living with families, they attended social events, parties, balls, charity events, private celebrations, and joined religious societies. They were invited into people's homes informally for tea or dinner. They or their assistants attended school assemblies, judicial proceedings, political debates and conventions, missionary meetings, lectures, working meetings, official lunches, and the annual dinners of various clubs and civic associations. They used participant observation and took detailed field notes of all of these research activities.

This book provided a comprehensive, complex portrait of a small American town. Unlike the Chicago School, the Middletown study was not concerned with urban problems. And unlike the social reform movement it was not motivated by the desire to improve social conditions. It therefore had more potential to serve as a model for university-based academic research, especially since readers of the book liked the pleasant picture of the world of a small American town (Neurath, 1987, pp. 72–75). The study increased the popularity of social research among the wider American public, and in doing so it certainly made it much easier for many universities to introduce sociology and social research into their curricula. In this sense Middletown marked an important step toward the inclusion of ESR as a legitimate part of academic sociology.

Robert and Helen Lynd returned to the same town a decade later to observe the changes that the Great Depression had brought about in the community. In 1937 they published *Middletown in Transition* (Lynd and Lynd, 1937). This time, they drew more heavily on social stratification theory. The picture of the community that they produced was much less idyllic than in the 1920s. This is evident, for instance, in a chapter devoted to a prominent family in the town and the way its influence had increased during and as a consequence of the economic depression.

The series of research surveys headed by William Lloyd Warner was another significant contribution to empirical research in the 1930s. The results of these surveys were subsequently published in five volumes between 1939 and 1960 and are most often cited collectively as the *Yankey City Series*. Compared to the clearly defined concept in *Middletown* or even the concepts of the Chicago School, the model of data collection and analysis employed in this series was less clear-cut. Nevertheless, it likewise exhibited the important feature of teamwork which, following tradition established by the Chicago School, was now a regular aspect of every empirical study as a tried and tested part of research methodology.

Some individual empirical studies also emerged in Europe. One was *Marienthal*, a study of an Austrian village with a high unemployment rate. It was carried out in the early 1930s during the great economic crisis (Jahoda et al., 1974/1933). *Marienthal* was one of the first empirical social studies to combine qualitative and statistical methods in its research design (Fleck and Stehr 2011: 26-28). The researchers carried out tasks like measure the pace at which 50 Marienthal inhabitants walked across the main square, ask inhabitants to fill in 80 individual time budgets, record 40 weekly household menus of unemployed families, and ask children about their wishes for Christmas presents. They conducted statistical analyses of the sale of expensive and cheap food items in the local shop using time series, and recorded the narratives of unemployed people about the past. Findings were collated and corroborated using a variety of different methods. Most of the data were drawn from families and individuals, but the findings were used to characterize the unemployed community of Marienthal as a whole. This study was published in 1933 in Germany, but since it was the work of three Jewish researchers it was not distributed via bookstores or libraries under the Nazi regime. It received recognition much later, when it was republished in German in 1960 by the Institut für Demoskopie in Allensbach, and then in 1971 and 1974 in English translation in London and New York (Jerabek, 2006, pp. 70–85). As an appendix, the book included Hans Zeisel's study *On the History of Sociography* (Zeisel, 1974/1933).

Street Corner Society – An Analysis of Relations in Groups of Young People

A different approach was developed by William Foote Whyte in his *Street Corner Society*. Whyte used 'participant observation' as almost his sole method of data collection. He then summarized his 4 years of observations made between 1937 and 1940 and published the book in Chicago in 1943. The monograph enriched sociological methodology by demonstrating that the almost exclusive use of qualitative research methods has the capacity to shed light on patterns in the structure of social groups. Whyte's *Cornerville* was an invented name for the Italian neighborhood in which he carried out his observations, and is almost wholly constructed from information about a gang of boys that hung out on street corners and members of an Italian community club in a slum in Northern Boston. It does not therefore present a complex picture of the urban community in that area; there is no description of family life, the way children are raised in the family, the way people make a living, or the status hierarchy in the whole population, nor is there anything about the religious life of the town or the neighborhood. The entire study is built on analysis of the relations within small social groups. It shows that no unsurpassable divide exists between 'the world of qualitative data' and 'the world of social patterns.' Qualitative methods can uncover social patterns, just as statistical methods do. They can identify frequently repeated relations among members of social groupings, and according to W.F. Whyte, "… we begin to see a pattern that we have not seen before" (Whyte, 1993/1943, p. 280). These relations in "the case study can discover uniformities that can be checked further in other case studies or experimentally and/or quantitatively" (Whyte, 1993/1943, p. 368). This was the contribution of *Street Corner Society* and W.F. Whyte to the history of ESR.

Market Research, Public Opinion, and Communication Research as Models for the Further Development of Empirical Social Research

In the 1920s large files of standardized data began to be used in market research, preelection polls, and generally in the study of public opinion and in media research. In this kind of

'survey research,' 'primary input data' on individual consumers or households, individual voters, listeners, or readers are gathered and the findings then applied to the whole social unit.

It was in market research and election polls that statistical methods were first regularly applied to process large standardized data sets about social phenomena. The approach involves interpreting the overall statistics for an aggregate of units that together make up a society or just a clearly identifiable population. These numbers can serve as indicators or values of variables for the aggregated characteristics of voters of the Republican Party, or attitudes of the readers of *Life* magazine, or attributes of behavior of Coca-Cola consumers.

This fundamental advance, which was initiated outside sociology, transformed ESR in the 1930s and 1940s. Unified data on the units that make up social formations were converted with the aid of statistical methods into a relevant statement about the whole social formation. Hitherto ESR had focused on characterization of concrete social formations – usually communities, towns religious, or ethnic communities. Rarely had it employed summarized data obtained from statistics on a set of units of a given social population. Verbal descriptions rather than statistical tabulations had been the usual practice. Robert and Helen Lynd did not calculate averages for 'working class' and 'business class' households. Even researchers in the social survey movement who worked with primary data on individual inhabitants and their households tended to use these data only in a socially cartographic way, so as to produce color banding for different rates of poverty or the different ethnic backgrounds of the majority of inhabitants in a given neighborhood.

Once the statistical processing of mass data had been introduced into ESR, the nature of the field began to change. One of the first examples of this kind of social research with scientific results was the 'RAVAG Study,' a radio audience study conducted by P.F. Lazarsfeld in 1931 and 1932 in Vienna. Lazarsfeld used a questionnaire survey. He did not employ any sophisticated sampling procedure, but nor did he use the data to represent, or as it were map, the whole audience of listeners. He received responses from 110 000 individual listeners and processed their responses by calculating 7000 simple indexes, each for an individual group of listeners and type of radio program. These (partially nonrepresentative) data were then analyzed on a comparative basis, interpretations being based on the pattern of the different values of indexes in relation to the groups of listeners (Jerabek, 2006, p. 65–70).

The frequent use of individualized data from large samples of units, usually consumers, voters, the population of a town, etc., led to the gradual development of standardized methods of questioning, to the improvement of sampling techniques, and to the development of statistical indicators to facilitate comparisons and adequate formulation of conclusions about the relationships between the studied characteristics of subsamples. Data collection methods also improved. In the 1920s and 1930s scaling methods and techniques were developed (Emory Bogardus, Rensis Likert, Charles Osgood, Louis Thurstone, Louis Guttman, Paul Lazarsfeld). Researchers gradually acquired punching (Hollerith) machines for classifying data, tabulations, and simple calculations (since World War II also used in academic social research). Computers and advanced software for social research were developed in the last decades of the twentieth century.

ESR Reformulates Its Research Questions: From Survey Research to Survey Analysis

The use of individualized data for scientific purposes in social research is associated with two large research projects conducted in the late 1930s and in the 1940s. Paul Lazarsfeld started his 'Princeton Radio Project' in 1937, and followed up his investigation at *The Bureau of Applied Social Research* in the 1940s. Samuel A. Stouffer directed and carried out the 'American Soldier' project between 1941 and 1950 (Stouffer et al., 1949).

Between 1939 and 1949 the 'Princeton Radio Project' generated a series of books: *Radio & the Printed Page* by P.F. Lazarsfeld; *Invasion from Mars* by Hadley Cantril, Hazel Gaudet, and Herta Herzog (1940); *Mass Persuasion* by Robert K. Merton; *The People's Choice* by P.F. Lazarsfeld, B. Berelson, and Hazel Gaudet (1944); and *Communications Research 1948–1949* by P.F. Lazarsfeld and F. Stanton (1949).

'American Soldier' was published in three stages. The first two parts and part three on the psychological experiments in 1949 (Stouffer et al., 1949; Hovland et al., 1949). Part four on the methodology of the research in 1950 (Stouffer et al., 1950). The fifth part titled 'Continuities in Social Research' by Robert K. Merton and Paul F. Lazarsfeld concentrated on the basic theoretical and methodological studies of R.K. Merton, P.F. Lazarsfeld, S.A. Stouffer, and other colleagues (Merton and Lazarsfeld, 1950).

Both projects were based on the idea that it is necessary to uncover, identify, and systematically verify causal relations in the collected data. The 'vote counting' model was transformed into the survey analysis model in conjunction with fundamental changes in data analysis (Coleman, 1972, p. 402). The new research question was why? – investigating the causes of observed phenomena led to a search for the real causes hidden behind the manifest data.

The Principles of Survey Analysis

The social sciences by their nature cannot engage in experimental research. Paul F. Lazarsfeld and Samuel A. Stouffer made up for this limitation with a new model of survey analysis and elaboration (Stouffer and Lazarsfeld, 1972/1937). 'Survey analysis' uses empirical evidence on the simultaneous effects of multiple independent variables on an observed dependent variable.

Survey analysis reveals the *causes* of observed differences. This approach, which distinguishes arbitrary relations from genuine (fundamental) causal relations between variables, was named and codified by Paul F. Lazarsfeld as the 'elaboration model' (Lazarsfeld, 1955/1946; Kendall and Lazarsfeld, 1950). The principle and methods of 'survey analysis' are thus only a generalization of the elaboration formula. 'Survey analysis' can essentially be described as the attempt to discover true, genuinely valid, fundamental connections between observed phenomena.

Important Empirical Social Studies in the 1930s and 1940s

In the 1930–1940s 'social survey' research advanced at other universities and state-funded and government-founded institutions. Key scholars in this sphere included George Gallup, Rensis Likert, Archibald Crossley, and Elmo Roper. Important institutions where research work that influenced the future nature of ESR was undertaken included the Gallup Organization, National Opinion Research Center in Chicago, Institute for Social Research at University of Michigan, and The Roper Center for Public Opinion Research. These institutions mainly conducted large representative surveys. Studies and publications drawing on local tradition in social survey were produced in many countries. Social research in France has been discussed by Terry N. Clark (Clark, 1967), and British investigation is mentioned by Martin Bulmer (Bulmer, 1985a, pp. 11–22). In Germany important studies were undertaken by Theodor Geiger, Ferdinand Tönnies, and the Frankfurt School, while in the United States prominent researchers in this area included Elton Mayo, Fritz J. Roethlisberger and William J. Dikson, and Gunnar Myrdal (Swedish origin). In the postwar period, it was the 'Kinsey Reports' that attracted the most interest, and the first of these was published in the period covered by this article (Kinsey et al., 1948).

Conclusion: The Relationship between Sociological Theory and ESR

ESR eventually became a recognized source of contributions to sociological theory. The process by which ESR became an established part of academic sociology took 50 years. In the initial phase of the process we find sociologists at the Chicago School using empirical data to improve social theory. The empirical data marshaled by H.W. Zorbaugh about 'natural areas' in *Gold Coast and the Slum*, Nels Anderson's observations on the roles of 'hoboes' in *Hobo* and the detailed description of the institution of dance halls in Paul Cressey's *The Taxi-Dance Hall* helped R.E. Park and E.W. Burgess in the development of their theories.

In the next phase of the evolution of ESR, statistically processed mass data on individuals and their families provided opportunities for generalizations about whole social units. Although often no link was made between empirical data and sociological theory in this phase, this phase was the essential starting point for future generalizations on the higher level of social units such as communities, audiences, etc. and the precondition for using and establishing a new type of data on lower order units, individuals and their attitudes, households and their characteristics. The 'survey research' stage assisted this development by standardizing the observed indicators, introducing sampling techniques, supporting the use of mathematical methods, and immensely advancing techniques of data collection and analysis.

The third phase of development made causal analysis its objective. Equipped with a clear idea of the goals and methods of data use, sociologists were already able to demonstrate that aggregate data were a tool and an instrument for social theory. We can understand this next stage as a kind of 'scientification of previously spontaneous development.' The model explicitly expressed the relations between the studied groups (or persons) and structured attributes of the phenomenon to be explained. Particular data obtained from individuals are a component part of the whole to which social theory refers. These data could then refer to and become instrumental in the development of the social sciences (Fleck and Stehr, 2011). Robert K. Merton made the point with P.F. Lazarsfeld, in the introduction to their jointly edited monograph *Continuities of Social Research*, when they described the combination of ESR and sociological theory as follows: "The appropriate arrangement of such data stemming from individuals, then, can provide systematic evidence on the cultural norms and social organization of groups" (Merton and Lazarsfeld, 1974/1950, p. 13).

Today the truth and legitimacy of scientifically obtained findings is often challenged. The challenge has been launched from the positions of the interpretative and constructivist visions of the world. Those who embrace these visions use stories and suggestive pictures, which they then elaborate, as a mode of persuasion. To expose these appearances as unsubstantiated is one of the tasks of science. In this situation it is becoming immensely important to defend and develop the cumulative structure of sociology as an integrated system of theoretically grounded empirical research and individual sociological theories built on or enriched by the findings of empirical research. Study of the history of ESR has been drawing attention to examples of the meaningfulness of this cumulative building of academic sociology.

Acknowledgment

This article was sponsored by the Grant Agency of the Czech Republic, grant no 404/11/0949.

See also: Academic Research and Employment: Recent Changes in Europe and the United States; African Studies: History; Communication Research and Media Studies, History of; Definition of the situation: History of the Concept; Demography, Early History of; Discipline-Building in the Social Sciences: Collective Memory, Biography and Autobiography; Funding of Social Science, History of; Psychology, History of (Twentieth Century); Schools in the Social and Behavioral Sciences: Concepts and Historical Relevance; Sociology, History of; Statistical Analysis Packages; Statistics, History of.

Bibliography

Bales, K., 1991. Charles Booth's survey of life and labour of the people in London 1889–1903. In: Bulmer, M., Bales, K., Kish Sklar, Kathryn (Eds.), The Social Survey in Historical Perspective 1880–1940. Cambridge University Press, Cambridge, pp. 66–110.

Barton, A., 1982. Paul Lazarsfeld and the invention of the university institute for applied social research. In: Holzner, B., Nehnevajsa, J. (Eds.), Organizing for Social Research. Schenkman, Cambridge, MA, pp. 17–83.

Barton, A., 2001. Paul Lazarsfeld as institutional inventor. International Journal of Public Opinion Research 13 (3), 245–269.

Blumer, H., 1939. An Appraisal of Thomas and Znaniecki's the Polish Peasant in Europe and America. Transaction Books, New Brunswick, NJ.

Bowley, A.L., Burnett-Hurst, A.R., 1915. Livelihood and Poverty. Ratan Tata Foundation, London.
Bowley, A.L., Hogg, M.H., 1925. Has Poverty Diminished? A Sequel to 'Livelihood and Poverty'. King, London.
Bulmer, M., 1984. The Chicago School of Sociology. Institutionalization, Diversity, and the Rise of Sociological Research. Chicago University Press, Chicago.
Bulmer, M. (Ed.), 1985a. Essays on the History of British Sociological Research. Cambridge University Press, Cambridge.
Bulmer, M., 1985b. The development of sociology and of empirical social research in Britain. In: Bulmer, M. (Ed.), Essays on the History of British Sociological Research. Cambridge University Press, Cambridge, pp. 3–36.
Bulmer, M., 1991. W.E.B. Du Bois as a social investigator: the Philadelphia Negro 1899. In: Bulmer, M., Bales, K., Kish Sklar, Kathryn (Eds.), The Social Survey in Historical Perspective 1880–1940. Cambridge University Press, Cambridge, pp. 170–188.
Bulmer, M., Bales, K., Kish Sklar, Kathryn (Eds.), 1991. The Social Survey in Historical Perspective, 1880–1940. Cambridge University Press, Cambridge.
Calhoun, Craig (Ed.), 2007. Sociology in America. A History. University of Chicago Press, Chicago.
Clark, T. Nichols, 1967. Empirical Social Research in France, 1850–1914. Columbia University Press, New York.
Coleman, J.S., 1972. Paul Lazarsfeld's work in survey research and mathematical sociology. In: Lazarsfeld (Ed.), Qualitative Analysis. Allyn & Bacon, Boston, pp. 395–409.
Converse, Jean M., 1987. Survey Research in the United States. Roots and Emergence 1890–1960. University of California Press, Berkeley.
Eknoyan, G., 2007. Adolphe Quetelet (1796–1874) – the average man and indices of obesity. Nephrology Dialysis Transplantation 23 (1), 47–51.
Easthope, G., 1974. A History of Social Research Methods. Longman, London.
Fleck, Ch., Stehr, N., 2011. Introduction. In: Lazarsfeld, P.F. (Ed.), An Empirical Theory of Social Action: Collected Writings. (Edited by Ch. Fleck and N. Stehr.). The Bardwell Press, Oxford, Cambridge, pp. 1–46.
Haas, H., Jerabek, H., Petersen, T. (Eds.), 2012. The Early Days of Survey Research and Their Importance Today. Braumüller Verlag, Wien.
Hennock, E.P., 1991. Concept of poverty in the British social surveys from Charles Booth to Arthur Bowley. In: Bulmer, M., Bales, K., Sklar, Kish Karthryn (Eds.), The Social Survey in Historical Perspective 1880–1940. Cambridge University Press, Cambridge, pp. 189–216.
Jahoda, Marie, Lazarsfeld, P.F., Zeisel, H., 1974[1933]. In: Marienthal. The Study of Unemployed Community, first Engl. ed. Tavistock, London. Aldine-Atherton, Chicago 1971.
Jerabek, H., 2006. Paul Lazarsfeld's Research Methodology. Biography. Methods. Famous Projects. The Karolinum Press, Prague.
Jerabek, H., 2012. Three stages in the institutionalization of empirical social research. In: Haas, H., Jerabek, H., Petersen, T. (Eds.), The Early Days of Survey Research and Their Importance Today. Braumüller Verlag, Wien, pp. 100–118.
Kendall, P.L., Lazarsfeld, P.F., 1950. Problems of Survey Analysis. In: Merton, R.K., Lazarsfeld, P.F. (Eds.), Continuities in Social Research: Studies in the Scope and Method of 'The American Soldier.' Arno Press, New York (reprint 1974 The Free Press, Glencoe, Ill).
Kent, R., 1985. The emergence of the sociological survey, 1887–1939. In: Bulmer, M. (Ed.), Essays on the History of British Sociological Research. Cambridge University Press, Cambridge, pp. 52–69.
Kern, H., 1982. Empirische Sozialforschung. Ursprünge, Ansetze, Entwicklungslinien. Verlag C.H. Beck, München.
Kinsey, A.C., et al., 1948. Sexual Behavior in the Human Male. Saunders, Philadelphia.
Landau, D., Lazarsfeld, P.F., 1968. Quételet, Adolphe. In: International Encyclopedia of the Social Sciences, vol. XIII, pp. 247–257.
Langenbucher, W.R. (Ed.), 1990. Paul F. Lazarsfeld. Die Wiener Tradition der empirischen Sozial- und Kommunikationsforschung. Ölschläger Vrlg, München.
Lautman, J., Lécuyer, B.P. (Eds.), 1998. Paul Lazarsfeld (1901–1976) La sociologie de Vienne à New York. L'Harmattan, Paris – Montréal.
Lazarsfeld, P.F., 1949. The American soldier – an expository review. Public Opinion Quarterly 13, 377–404.
Lazarsfeld, P.F., 1955/1946. The interpretation of statistical relations as a research operation. In: Lazarsfeld, P., Rosenberg, M. (Eds.), The Language of Social Research. A Reader in the Methodology of Social Research. The Free Press, New York, pp. 115–125 (Orig. Isis (1961), 52, 277–333.).
Lazarsfeld, P.F., 1982a[1961]. Notes on the history of quantification in sociology – trends, sources, and problems. In: Kendall, Patricia L. (Ed.), The Varied Sociology of Paul F. Lazarsfeld. Columbia University Press, New York, pp. 97–167 (Orig. Isis (1961), 52, 277–333.).
Lazarsfeld, P.F., 1972. Qualitative Analysis. Historical and Critical Essays. Allyn and Bacon Inc., Boston.
Lazarsfeld, P.F., 1982b. The varied sociology of Paul F. Lazarsfeld. In: Kendall, Patricia L. (Ed.). Columbia University Press, New York.
Lecuyer, B., Oberschall, A., 1968. The early history of social research. Section III in article 'Sociology'. In: International Encyclopedia of the Social Sciences. MacMillan, New York.
Lombardo, C., 2004. Lazarsfeld, Merton e la scuola sociologica della Columbia. Sociologia e ricerca sociale 74, 127–149.
Lowery, Shearon, De Fleur, M.L., 1983. Milestones in Mass Communication Research. Media Effects. Longman, New York-London.
Lynd, R.S., Lynd, Helen, 1929. Middletown. A Study in Contemporary American Culture. Constable, London.
Lynd, R.S., Lynd, Helen, 1937. Middletown in Transition. Harcourt, Brace and Co., New York.
Madge, J., 1974/1962. The Origins of Scientific Sociology. The Free Press, New York.
Maus, H., 1973. Zur Vorgeschichte der Empirischen Sozialforschung. In: König, R. (Ed.), Handbuch der empirischen Sozialforschung, pp. 23–33. Enke, 3. Aufl., Stuttgart, Bd.1.
Merton, R.K., Lazarsfeld, P.F. (Eds.), 1950. Continuities in Social Research: Studies in the Scope and Method of 'The American Soldier'. The Free Press, Glencoe, IL (Reprint 1974 Arno Press, New York).
Merton, R.K., Coleman, J., Rossi, P. (Eds.), 1979. Qualitative and Quantitative Social Research. Papers in Honor of Paul F. Lazarsfeld. Free Press, New York.
Neurath, P., 1981. Paul F. Lazarsfeld and the institutionalization of empirical social research. In: Paper given before a joint session of the Columbia University Seminars on Content and Method of the Social Sciences and Mathematical Methods in the Social Sciences, November 11, 1981. Queens College, City University of New York and University of Vienna (Archive of the author).
Neurath, P., 1987. Vom Sinn der empirischen Sozialforschung. Ihre Entstehung im 18ten und 19ten Jahrhundert. Akademie für Sozialarbeit der Stadt Wien, Wien.
Neurath, P., 1988. Paul Lazarsfeld und die Institutionalisierung der empirischen Sozialforschung: Ausfuhr und Wiedereinfuhr einer Wiener Institution. In: Srubar, I. (Ed.), Exil, Wissenschaft, Identitat. Die Emigration Deutscher Sozialwissenschaftler 1933–1945. Suhrkamp, Frankfurt, pp. 67–105.
Oberschall, A., 1965. Empirical Social Research in Germany 1848–1914. Mouton & Co., The Hague–Paris.
Oberschall, A. (Ed.), 1972. The Establishment of Empirical Sociology: Studies in Continuity, Discontinuity and Institutionalization. Harper & Row Pbls., New York, Evanston.
Oberschall, A., 1978. Paul Lazarsfeld and the history of empirical social research. Journal of the History of the Behavioral Sciences 14 (3), 199–206.
Odum, H. (Ed.), 1927. American Masters of Social Science. Holt, Rinehart & Winston, New York.
Le Play, F., [1850]1877–1879. Les Ouvriers Européens, second ed., 6 vols. A. Mame et fils., Tours.
Le Play, F., 1859–1860. Les ouvriers des deux mondes. Société d'économie sociale, Paris.
Petersen, T., 2012. The art of discovering something simple and new. The role of outsiders in the history of social research. In: Haas, H., Jerabek, H., Petersen, T. (Eds.), The Early Days of Survey Research and Their Importance Today. Braumüller Verlag, Wien, pp. 213–230.
Platt, Jennifer, 1969. A History of Sociological Research Methods in America 1920–1960. Cambridge University Press, Cambridge.
Quetelet, A., 1835/1842. Sur l' homme et le développement de ses facultés, ou Essai de physique sociale. Academie Royale de Belgique, Brussels. 1835; A Treatise on Man and the development of his aptitudes, or Essay on Social Physics. Burt Franklin, New York, 1842/1968.
Sarton, G., 1935. Quetelet. Isis 23, 6–24.
Schad, Susanne P., 1972. Empirical Social Research in Weimar-Germany. Mouton, Paris–The Hague.
Smith, T.W., White, L.D. (Eds.), 1968[1929]. Chicago. An Experiment in Social Science Research, first ed. The University of Chicago Press, Chicago.
Stouffer, S.A., et al., 1949. Studies in Social Psychology in World War II. The American Soldier: Adjustment During Army Life. vol. I., Combat and Its Aftermath. vol. II. Princeton University Press, Princeton.
Stouffer, S.A., et al., 1950. Studies in Social Psychology in World War II. The American Soldier. Measurement and Prediction. vol. IV. Princeton University Press, Princeton.

Stouffer, S.A., Lazarsfeld, P.F., 1972/1937. Research Memorandum on the Family in the Depression. Arno Press, New York (first ed., 1937. Social Science Research Council, New York).

Whyte, W.F., 1993/1943. Street Corner Society, fourth ed. The Chicago University Press, Chicago (first ed. in 1943).

Wortham, R.A., 2005. The early sociological legacy of W.E.B. Du Bois. In: Blasi, A.J. (Ed.), Diverse Histories of American Sociology. Bril, Boston, pp. 74–95.

Young, Pauline V., 1949. Scientific Social Surveys and Research, second ed. Prentice Hall, New York.

Zeisel, H., 1974[1933]. Toward a history of sociography. In: Jahoda, Marie, Lazarsfeld, P.F., Zeisel, H. (Eds.), Marienthal. The Study of Unemployed Community, first Engl. ed. Tavistock, London, pp. 99–125. Aldine-Atherton, Chicago 1971.

Empiricism, History of

Peter Lipton[†], Cambridge University, Cambridge, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 7, pp. 4481–4485, © 2001, Elsevier Ltd.

Abstract

Empiricism in its many forms is united by an emphasis on the importance of experience to the formation of concepts and to the acquisition of knowledge. The foil to empiricism is rationalism, which emphasizes instead the importance to thought and knowledge of material that is in some sense independent of experience. Empiricism has been a widely discussed philosophical approach from the ancient period to the present day, providing accounts of the nature of thought and of the sources of knowledge. Empiricists have typically seen themselves as the scientist's friend, since they hold that knowledge of the world requires observation and experiment, but insofar as empiricists maintain that experience is the only source of factual knowledge, they leave themselves open to powerful skeptical arguments that seem to show that there can never be good reason to believe any claim about unobservable events or processes, and perhaps not even any claim that goes beyond what has actually been observed.

There are almost as many empiricisms as there are empiricists, but what these views or approaches have in common is an emphasis on the importance of experience to the formation of concepts and to the acquisition of knowledge. The foil to empiricism is rationalism, which emphasizes instead the importance to thought and knowledge of material that is in some sense independent of experience. The range of empiricist positions is vast, from the shocking view that all we can think or know about are our sensations to the mundane claims that experience plays some role in the formation of some of our concepts and in the justification of some of our beliefs. Empiricism of some form may seem both obviously correct and obviously the correct philosophy for science on the grounds that it is clear that one can only find out about the world by observing it; but this innocuous looking thought has been disputed, and in any case many forms of empiricism go far beyond it in their claims. It is also unclear whether empiricism ultimately supports or undermines claims to scientific knowledge.

History

Empiricism is a hardy perennial in the history of philosophy. It was extensively discussed in antiquity, often with considerable hostility. Plato disdained it, partly on the grounds that observation can only provide information about a realm of appearances less important and indeed less real than a more abstract realm that we may be able to grasp through a form of thinking that does not depend on experience. Mathematics was a model of knowledge for many ancient thinkers, and this is perhaps the area where empiricism is least plausible, since mathematical proofs do not seem to depend on observation. Nevertheless, empiricism did have defenders in the ancient world, with Epicurus often cited as its first clear exponent. He held not only that observation is the only source of knowledge about the world but that, insofar as they are taken as sources of information about the appearances of things, the senses never lie.

Empiricism also had its advocates in the medieval period. St. Thomas Aquinas wrote that there is nothing in the intellect which was not previously in the senses, though by this he meant to endorse only concept empiricism and not a strong form of knowledge empiricism. Concepts can only be formed with the help of sensory images, but some knowledge – for example, some knowledge about God – is available by means of logical argument independent of particular observations. Empiricism is perhaps most strongly associated with three philosophers of the seventeenth and eighteenth centuries: John Locke, George Berkeley, and David Hume. These British empiricists developed versions of both concept and knowledge empiricism. In some respects Berkeley (1710, 1713) can be seen as attempting to purify Locke's (1700) view of its nonempiricist elements and Hume (1739, 1777) as attempting the same for Berkeley. Both Locke and Hume appealed to Newton's achievements in physics for inspiration: Locke presented himself as Newton's underlaborer, 'clearing ground a little, and removing some of the rubbish, that lies in the way to knowledge.' Part of that ground-clearing operation consisted of a vigorous rejection of the claim that we are born with innate knowledge, a characteristic point of dispute between empiricists and rationalists (cf Mackie 1976). Hume characterized his own project more ambitiously, as an attempt to do for the human sciences ('moral philosophy') what Newton had done for the physical sciences ('natural philosophy'), by providing a theory that explained the creation and behavior of ideas. Berkeley was less influenced by the model of natural philosophy, aiming instead to combat skepticism and to support theism. The result was an exceptionally lucid statement and defense of idealism, the view that the only things that exist are minds and their contents. Locke had not taken this step, hoping to hold on to a material world even if our knowledge of its nature must be severely limited, but Berkeley argued that only idealism was compatible with the principles of empiricism. Hume was an idealist too, but made less of a fuss than Berkeley about this aspect of their philosophies, in part because Hume wished to avoid metaphysics, even of a negative sort.

In the nineteenth century, John Stuart Mill carried forward the tradition of British Empiricism from Hume, including his

[†] Deceased

idealism, though in Mill's (1872) hands the position is normally referred to as 'phenomenalism.' Mill held that physical objects are 'permanent possibilities of sensation': to say that there is a tree in the quad at a certain time is to say either that someone is having an appropriate arboreal experience at that time or that they would have had one if they had been in the quad. Berkeley famously preferred to account for the existence of the tree when no human is observing it by appealing to God's constant observation, but in one or two passages he clearly presages Mill's counterfactual alternative, which appeals to possible human observations rather than to actual divine observations.

Mill went beyond his British precursors, however, in his treatment of mathematics, the most obvious candidate for nonempirical knowledge. Most empiricists have adopted the strategy of admitting that pure mathematics is nonempirical but denying that it provides worldly knowledge, claiming instead that it tells us only about the logical consequences of our definitions. Mill, however, maintained that mathematics, however pure, is empirically based, an inductive inference from pervasive regularities in our experience, such as those involving the manipulation of objects.

Twentieth-century empiricism remained inspired by science. Logical positivists and logical empiricists saw empiricism both as a way to understand scientific claims and as a way to make philosophy itself more scientific and less metaphysical (cf. Ayer 1936). They especially pushed the conceptual side of empiricism, for example, in the verifiability theory of meaning, whose motto was that 'the meaning of an expression is its method of verification.' On this view, the experiences that would confirm a claim also tell us what it means, and a claim which no course of observation could either confirm or refute is either a mere definition or metaphysical nonsense. Empiricism in this century has also been promoted by scientists themselves, in movements such as operationalism in the physical sciences and behaviorism in psychology. Another influential empiricist strain in this period has been Pragmatism, especially as developed by the American triumvirate of William James, Charles Saunders Pierce, and John Dewey, and more recently in a radical form that rejects definitions and meanings by Willard Van Orman Quine (1951).

Concept Empiricism

Empiricism about concepts is the view that our concepts derive in some sense from our experiences. Hume offered what has come to be a canonical version of this position, arguing that all our ideas are copies of preceding impressions, meaning roughly that every concept is a faint image of the experiences that cause it. Thus, to take a favorable example, our idea of the color blue is supposed to a faint image of the particular blue experiences we have enjoyed. For Hume, the copy principle was not just a fundamental fact about our cognitive economy but also a powerful philosophical tool. Constructively, the copy principle is supposed to allow us to analyze the content of an idea or concept, since by tracing ideas back to their source impressions we determine what that content really is. Destructively, the copy principle is supposed to expose empty talk or thought, since if a putative idea has no possible source impression, it must be devoid of content. The copy principle obviously requires some general demarcation between impressions and ideas. Hume draws it subjectively, in terms of the relative 'force and vivacity' of perceptions, recognizing the vagueness of this distinction, but trusting that we will all understand the difference between the vividness of what one sees as compared with faintness of memory and imagination.

Hume's most famous applications of his copy principle were destructive, part of his program to expose the emptiness of metaphysics. Thus, although he allowed that we do have an idea of the 'necessary connection' that is supposed to glue causes to their effects, the source impression reveals this idea to be purely subjective. For that impression is merely the feeling of expectation of a familiar effect when we see a familiar cause, something that is obviously only in our minds. We cannot conceive one thing making another thing happen: all we can conceive is a pattern of objects. Indeed although Hume often did write as if we could conceive of external objects, he seems to have held that the copy principle rules out conception of these too.

Hume's antimetaphysical fervor was taken up by the logical positivists of the first half of the twentieth century. They took statements rather than ideas as the unit of analysis, and held that a statement is meaningful only if there is some specifiable experience that would verify it. They made an exception of mathematics and logic, which were allowed to count as meaningful even though not empirical, but only because they consisted of definitions and their consequences rather than substantive claims. Since philosophical truths have often been supposed to consist precisely of nonempirical truths about the world, it is not surprising that the verification principle was supposed to sweep away great swaths of traditional philosophical doctrine, not on the grounds that such claims are false, but that they are literally meaningless, just noise.

As with Hume, the logical positivist program also had a positive side: for those statements that did manage to be meaningful, their meaning comprised the experiences that would confirm them. From this it follows that all claims are claims about experiences. Some tempered this extreme view by speaking of observable states of affairs rather than experiences, thus making the data public. In the hands of some behaviorists in the philosophy of psychology, this brought about a kind of reversal of the positivists' original, experiential doctrine, since now rather than analyzing all claims as claims about experiences, experiences and other mental states are analyzed in physical terms. On this view, the real meaning of the claim that someone is happy is not that the person is enjoying a certain kind of inner life, but that they are disposed to behave in characteristically happy ways. Hostility to this position is the source of the story of the behaviorist who, upon meeting another behaviorist in the street, proclaims 'You're fine, how am I?'

Concept empiricism has been extensively criticized in both its negative and positive aspects. The demarcation that concept empiricism attempts to draw between genuine concept or meaningful statement on the one hand and nonsense on the other depends on making reasonably precise the notion of a copy of an impression or the conditions under which evidence confirms a hypothesis. This has proven remarkably difficult so that, for example, familiar accounts of confirmation have been shown to have the consequence that everything is evidence for anything else, a result that would make all statements meaningful and is anyway obviously false. The other conspicuous

weakness of concept empiricism concerns the specific meanings it attributes to statements that it rightly classes as meaningful. The attempt to reduce the meaning of a statement to the evidence that would support it typically both attributes meaning the statement does not have and leaves out meaning it does have. This comes out clearly in the case of scientific hypotheses that traffic in unobservable entities and processes. A hypothesis about the behavior of subatomic particles does not include as part of its *meaning* the diverse and indirect kinds of empirical evidence in its favor, while that meaning does include claims about unobservable particles.

Knowledge Empiricism

Knowledge empiricism is the view that all knowledge of the world is based on experience. Put negatively, this is the denial of a *priori* knowledge of the world. Empiricism here opposes rationalism, according to which the fundamental source of knowledge of the world is reason, not experience.

Critics of knowledge empiricism generate putative examples of a *priori* knowledge. Empiricists deal with these cases in several ways, claiming either that the statement in question cannot be known a *priori*, because it is unknowable altogether or knowable but only through experience, or that that the statement does not really provide knowledge *of the world*. Thus, when René Descartes, the great seventeenth century rationalist, claimed to know *a priori* of his own existence, Hume could respond that the only knowledge one has of oneself is of one's experiences, which are known through themselves. Mathematical knowledge is another popular area for rationalists' examples of the *a priori* although, as we have seen, Mill claimed that even this knowledge is ultimately empirical, resting on observations of the result of adding together and dividing apart physical objects. The more common empiricist treatment of mathematical knowledge, however, is to admit it does not depend on specific experiences but to deny that it is about the world. Both Hume and the logical empiricists admit that definitions or analytic truths – such as that a bachelor is an unmarried man – can be known to be true without experience, but maintain that these statements have no factual content. They then analyze mathematical axioms as definitions and treat everything that is derived from them as analytic truths which can be known without experience but tell us nothing about the world.

Empiricism and Scientific Knowledge

Alongside the question of whether there can be any knowledge independent of experience is the question of how far empirical knowledge may extend beyond the experience upon which it is based. Since Hume's seminal discussion of the problem of induction, it has become difficult to see how knowledge can extend beyond actual experience at all. It is neither surprising nor worrying that any nondemonstrative inference from experience might in principle go wrong; but Hume gave a remarkably resilient argument that we can have no reason whatever for extrapolating in one direction rather than in another, even when dealing with everyday claims about the immediate future or the unobserved past and present. The nub of the problem is that any justification for the future success of our methods of prediction or extrapolation seems itself to rely on such a method, leaving us arguing in what is apparently a vicious circle, not unlike trusting someone solely on the grounds that he or she claims to be trustworthy. Thus empiricism has been a friend to the skeptic: if Hume is right, we have no more reason to trust the most sophisticated scientific predictions than tea leaves or entrails.

Even empiricists who, ignoring the Humean problem of induction, help themselves to inferences from past to future observations have had the greatest difficulty in showing how scientific knowledge is possible, since scientific theories appear to traffic in claims about entities and processes that are un observable, not just unobserved. As we have seen, this is denied by the concept empiricists, for whom no claims about the unobservable can even be formulated, but few recent empiricists have been willing to go down this road. Instead, they have had to face the problem of the underdetermination of theory by data, developed with special force in the work of Pierre Duhem (1954) and Willard Van Orman Quine (1975).

The problem is that no matter how much evidence there is in favor of a particular scientific theory, there exist in principle many other theories, incompatible with the first yet compatible with all the evidence. This creates an epistemic stalemate that seems to block warranted inference to any claim that goes beyond what is observable.

Empiricists have attempted to deal with the underdetermination problem in a number of ways. Karl Popper (1959) adopted a radical solution that was designed to solve the Humean problem of induction as well. It was to deny that scientists need to indulge in nondemonstrative inferences at all. Instead, they are to attempt to refute their theories and then replace them by better ones, where this process relies only on a deductive inference from the falsity of a prediction to the falsity of the theory that entailed the prediction. This is empiricism in the sense that the rejection of a theory is determined by observational data. Popper is one of the few philosophers of science to have attracted the respect of a significant number of practicing scientists, probably because scientists appreciate being reminded of the importance of considering what sort of results would tend to undermine a hypothesis, but Popper's view has been severely criticized by philosophers. Falsifying an interesting and successful scientific theory is not nearly as straightforward as Popper would have scientists pretend. Moreover, even if Popper's methodology could be put into practice, it would still be a form of skepticism, since according to him we never have any reason whatsoever to think that any scientific theory or prediction is even approximately correct, nor can we have any reason to believe any claim about the future.

Empiricism continues to play an important role in the study of knowledge within science studies. One example is the 'strong program' in the sociology of knowledge, which rests on a kind of psychological empiricism. As David Bloor (1991) puts the matter, the strong program assumes that perception and thought are distinct but asymmetrically related: perception influences thought more than thought influences perception. Within the philosophy of science, a recent empiricist approach to scientific knowledge and the underdetermination problem is Bas Van Fraassen's constructive empiricism. Van Fraassen (1980) introduces a technical sense of 'acceptance,' where to accept a scientific theory is not to believe that it is true, but only that it is empirically adequate, where a theory is empirically

adequate just in case everything it says about what is observable is true.

Constructive empiricism is the view that the aim of science is to discover theories that are empirically adequate, not theories that are true. The content of the theory that goes beyond its empirical claims is to be regarded as a model which may serve its function of generating true predictions without itself being a true representation of an invisible realm. Constructive empiricism thus appears to solve the underdetermination problem, since although the theories underdetermined by the data will be incompatible with each other, their empirical consequences, which is all the constructive empiricist would have us believe, may be compatible.

Constructive empiricism can thus be seen as a form of moderate skepticism, but it is still too skeptical for those who think that knowledge or at least reasonable belief about unobservable entities and processes is possible. Constructive empiricism also faces the difficulty of providing a nonarbitrary distinction between what is observable and what is not, a tricky problem faced in one form or other by almost all forms of empiricism.

See also: Darwin, Charles Robert (1809–82); Experimentation in Psychology, History of; Galton, Sir Francis (1822–1911); Helmholtz, Hermann Ludwig Ferdinand von (1821–94); Hume, David (1711–76); Logical Positivism and Logical Empiricism; Positivism, History of; Positivism, Sociological.

Bibliography

Ayer, A.J., 1936. Language, Truth and Logic. Victor Gollancz, London.
Bennett, J., 1971. Locke, Berkeley, Hume: Central Themes. Oxford University Press, Oxford, UK.
Berkeley, G., 1710/1975. A treatise concerning the principles of human knowledge. In: Berkeley, G. (Ed.), Philospohical Works. Dent, London.
Berkeley, G., 1713/1975. Three dialogues between Hylas and Philonous. In: Berkeley, G. (Ed.), Philosophical Works. Dent, London.
Bloor, D., 1991. Knowledge and Social Imagery, second ed. University of Chicago Press, Chicago.
Duhem, P., 1954. The Aim and Structure of Physical Theory. Princeton University Press, Princeton, NJ.
Fraassen, B.V., 1980. The Scientific Image. Oxford University Press, Oxford, UK.
Hume, D., 1739/1975. In: Selby-Bigg, L.A., Nidditch, P.H. (Eds.), A Treatise of Human Nature. Oxford University Press, Oxford, UK.
Hume, D., 1777/1975. In: Selby-Bigg, L.A., Nidditch, P.H. (Eds.), An Enquiry Concerning Human Understanding. Oxford University Press, Oxford, UK.
Locke, J., 1700/1975. In: Nidditch, P.H. (Ed.), An Essay Concerning Human Understanding. Oxford University Press, Oxford, UK.
Mackie, J.L., 1976. Problems from Locke. Oxford University Press, Oxford, UK.
Mill, J.S., 1872. An Examination of William Hamilton's Philosophy. Longmans, London.
Popper, K., 1959. The Logic of Scientific Discovery. Hutchinson & Co, London.
Quine, W.V.O., 1951. Two dogmas of empiricism. Philosophical Review 60, 20–43.
Quine, W.V.O., 1975. On empirically equivalent systems of the world. Erkenntnis 9, 313–328.

Employment and Labor, Regulation of

Katherine VW Stone, UCLA School of Law, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Employment regulation consists of the legal rules and institutions that constitute, govern, and structure the employment relationship. Until recently, most industrial countries had employment regulation that provided most employees with job security and an adequate package of social protection. However, these regulations have been relaxed in recent years as employers have moved away from hiring employees into long-term employment relationships and have created many types of short-term employment relationships instead. The decline of the standard form of employment and the regulatory regime that supported it has given rise to a number of controversies involving employment regulation. In addition, the spread of global trade and the diffusion of production around the world have put pressure on industrial countries to dilute their employment regulation and lower their labor standards. However, as global production and trade continue to proliferate, social pressures are building for more labor law protections both in the developed and in the developing world. Hence employment regulation is not likely to disappear, but it will be transposed as part of the emerging international regime.

Employment regulation consists of the legal rules and institutions that constitute, govern, and structure the employment relationship. Employment regulation potentially affects all aspects of employment – it determines whether an employee has job security, the minimum amount an employee must be paid, the safety and health conditions at the workplace, the right to vacation pay, sick leave pay, or other benefits, protection from discrimination on the basis of gender, race, ethnicity, religion, age, or personal attributes, the minimal age for employment, the age and terms of retirement, the provision of job training, protection from intrusive surveillance on or off the job, and many other aspects of the employment relationship.

The content of employment regulation also determines the parties' bargaining power vis-à-vis each other (Hale, 1923). For example, legal rules determine whether workers are free to join together into unions or whether such actions constitute unlawful conspiracies, whether and under what circumstances they may lawfully strike and picket, and whether they can use secondary pressure or appeals to the general public to support their position. Because bargaining power ultimately determines employees' wages, the legal rules that govern the employment relationship have powerful distributional consequences. For this reason, the content of employment regulation has always been a hotly contested and intensely politicized issue.

Employment regulation occurs by means of legislative action, judicial ruling, administrative action, contract, or custom. It is often assumed that employment regulation refers to affirmative interventions by the state into the employment relationship by means of statutes, administrative regulations, executive orders, or other acts of positive law. That assumption is incomplete. Employment regulation also takes place through the background contracting rules with which a buyer and a seller of labor power – that is, and an employer and an employee – form an employment relationship. For example, some countries treat the individual employment contract as a conventional contract that imposes affirmative duties on both the employer and the employee. For this reason, in many parts of Europe and Latin America, an employer violates the employment contract by firing an employee without good cause. In the United States, in contrast, the background default form of the employment contract is an at-will contract, that is, an arrangement in which an employee can be fired at any time for any reason (Stone, 2007).

In most industrial nations, there are two types of labor regulations – regulations that establish specific terms of the wage bargain and regulations that legally empower labor and management to engage in self-regulation. Direct employment regulation includes statutes that set minimum wages and maximum hours of work, mandate certain health and safety standards, regulate employer-based pensions and health insurance, and provide anti-discrimination protections. Legally empowered self-regulation involves statutes or doctrines that permit employees to form unions and bargain with employers for the terms and conditions of employment. Most Western countries have both types of employment regulation, but they differ as to which type is dominant in their labor relations system (Kahn-Freund, 1983). For example, most continental European countries have detailed legal codes that specify many terms of the employment relationship such as mandated hours of work, safety conditions, and childcare leave policies. Most also have laws that establish criteria for a lawful labor union and authorize such unions to engage in collective bargaining to determine certain terms and conditions of work. Many also have specialized labor courts to enforce both collective and individual employment rights. Moreover, in Germany and France, terms negotiated between unions and employers can be transformed into regulations applicable to entire industries or regions of the country (Weiss, 1987). In contrast, Great Britain relies primarily on the common law to regulate the employment relationship, and leaves many important employment terms to the parties to negotiate. In some instances, this leaves agreements outside the law altogether (Wedderburn, 1986).

Countries also differ in terms of what kinds of rights, privileges, and protections they give to organized labor. Some countries give very broad protection to collective action such as organizing, strikes, and picketing, while some do not.

While employment regulation creates the respective bargaining power of labor and management, it also reflects each side's relative power in the political process. Employers and unions operate as pressure groups, lobbying directly or indirectly galvanizing public opinion to shape the substantive law of employment in ways that will give them advantage. Judges and legislatures are, at least in part, responsive to the competing demands of employers and employees when they decide cases and enact employment-related statutes.

Employment Regulation of the Standard Contract of Employment

Despite differences between countries, a standard contract of employment became the norm in large manufacturing firms throughout much of the twentieth century. This standard contract provided employees with job security, decent wages, a rising trajectory for promotions and raises, and a generous package of health, retirement, and other job-linked benefits. While the standard contract was never universal, it covered large portions of the workforce in industrialized countries and became the prevailing norm by the middle of the twentieth century (Stone and Arthurs, 2013).

The standard contract of employment did not emerge fully formed: It was devised by industrial relations theorists, implemented by large manufacturing firms, bolstered by pressures from unions, and fortified by the various labor law regimes in the industrial nations. Gradually over the first half of the twentieth century, legal regimes that regulated employment emerged in various countries that adopted this standard employment model as their template. As a result, by the mid-twentieth century, most industrial nations had labor law systems that provided extensive protection for job security either by statute, contract, or custom. Most countries also had laws that required employment relationships to conform to the standard contract and that placed strict limits on the ability of firms to utilize temporary workers, short-term workers, or independent contractors. In addition, most had laws fostering company and/or industry-level collective bargaining that assumed that unions represented members who had long-term jobs.

To be sure, the standard employment contract was never universal. In each country, it covered different proportions of the workforce, and within the workforce, different sectors and segments. It was more characteristic of large manufacturing enterprises and the public sector than it was of small enterprises or the service sector, was more likely to cover men than women, and was more available to well-established populations than to recent immigrants or racial and ethnic minorities. Nonetheless, allowing for all these variations, the standard employment contract was both the paradigm that informed much labor policy and practice and the ideal to which these aspired.

Although the basic shape of the standard employment contract was prevalent in industrial countries for the last part of the twentieth century, the legal regimes that emerged to support it differed markedly from country to country. In some countries, the state mandated job security by statute. In others, the standard contract of employment was imposed by mandatory terms that were incorporated into the individual contract of employment. In yet others, job security was negotiated as part of a state-supported system of collective bargaining. And in some, it was a product of custom reinforced by legal mechanisms.

Until recently, extensive job security protections in certain countries such as France, Italy, and Spain gave workers nearly impregnable job security and made it extremely difficult for employers to fire workers or to downsize their workforces, even in dire economic exigencies. Furthermore, in Germany, the Netherlands, and other continental European countries, statutory job security protection was fortified by strong centralized unions, whose power was also protected by law as a matter of state policy.

In contrast, in the United Kingdom and the United States, the individual contract of employment was practically an oxymoron. Under the at-will contract rule that prevails in those countries, an employee can be fired at any time for any reason. Furthermore, the employer can change the terms of the contract at any time. Because the employment relationship has neither fixed duration nor fixed terms, it is somewhat misleading to call it a contract at all. The United Kingdom mitigated the harshness of the at-will rule with legislation that gave compensation to workers who were unfairly dismissed. In the United States, a few states have adopted common law doctrines to provide some unfair dismissal protection, but these are the exceptions, not the rules. Rather, unions in the United States often provided job security for their members by negotiating contractual provisions that specify that an employer can only dismiss an employee for 'just cause.' However, because union density has been comparatively low and steeply declining, this type of contractual protection is becoming rare (Stone, 2013a).

Countries also differ as to the way in which aspects of employment other than job security are regulated. In some countries, the state uses general tax revenues to provide a broad spectrum of work-related benefits such as skills training, employment insurance, job placement centers, day care, as well as universal benefits to workers and others, such as health care, social housing, and education. In other countries, some of these benefits are provided through contributory arrangements to which both employers and employees are required to contribute. In some, these benefits are provided by employers under compulsion of law, or in response to pressures brought by unions through collective bargaining or as part of a strategy to recruit and retain a loyal and productive workforce.

The Changing Nature of Employment

Recently employers throughout the industrial world have moved away from hiring employees into long-term employment relationships and have created many types of short-term employment relationships instead. They have increased the use of 'atypical' employees – those lacking some or all the elements of the standard contract. They have also increased the use of independent contractors and individuals leased from temporary employment agencies. Furthermore, in all industrial countries, job security has been declining (Stone, 2013b); in fact, many of these regimes recently have

changed their laws specifically to relax job security protections (Stone and Arthurs, 2013).

The decline of the standard form of employment and the regulatory regime that supported it have given rise to a number of controversies involving employment regulation. Some of the current controversies concern the following issues:

1. Who is an 'employee' under the labor laws, and hence entitled to employment protection? Should some employment protections be accorded to individuals who have no discernible employer, and if so, which protections, and to which individuals should they apply?
2. What are the respective roles of the state and employers in ensuring job security and social protection?
3. Can unions operate effectively in the workplace if employers no longer provide long-term jobs? Are there new roles for unions at the political level, and if so, can unions be effective in the political rather than the workplace realm?
4. Should there be transnational labor regulation to prevent firms from undermining national labor standards by relocating production to countries that have the lowest levels of labor standards?

Who is an 'Employee' under the Labor Laws, and Hence Entitled to Employment Protection?

Traditionally employment regulation has been applicable only to individuals who are 'employees.' The term 'employees' refers to an individual in an employment relationship with an employer. The law assumes that the employment relationship is well defined. Hence employment regulation does not apply to individuals who provide services for a fee, whether as business entities or independent contractors. When disputes arise, courts apply either a 'right to control' or an 'economic dependency' test to determine whether an individual was an 'employee' for purposes of the labor laws (Davidov, 2005; Stone, 2006).

In recent years, the line between employee and independent contractor has become blurred. Increasingly, employers utilize the services of individuals whom they characterize as independent contractors, even when the employer exercises considerable control over the work tasks of the individual so retained. Numerous other work relationships have arisen that do not fit the traditional employee–independent contractor dichotomy. For example, on-call workers, zero-hours workers, agency temporary workers, unpaid interns, volunteers, franchisees all pose problems when using traditional analyses to determine employee status (Hyde, 2006).

Adding to the difficulty in defining employees is the fact that many employers no longer seek to control the precise method of task execution by their 'employees.' Under current human resource practices, employers often delegate increasing amounts of discretion to employees, thus defying any mechanistic application of the 'right to control' test used for determining employee status (Stone, 2004).

Often employers today seek to label individuals as independent contractors to avoid the obligation to comply with employment regulations. For example, several years ago, trucking companies in the United States began to convert their drivers into owner-operators, but often required the drivers to lease the company's trucks, wear company-provided uniforms, drive routes the company orders them to drive in the sequence and on the timetable the company mandates, and serve exclusively the company's clients. However, as independent contractors, the drivers were not entitled to minimum wage and overtime pay, were not covered by unemployment insurance, workers compensation, or social security, and were legally precluded from belonging to a union.

The use of phony 'independent contractors' has spread far beyond the trucking industry, both in the United States and elsewhere. Today the practice is prevalent for firms that use the services of janitors, construction workers, IT workers, secretarial workers, hospital orderlies, and even machinists inside factories. In the United States, the claims by employees that they have been 'misclassified' as independent contractors are the most frequently litigated employment disputes today.

In recent years, courts, legislators, and scholars in many countries have debated the issue of who is an 'employee.' In the United Kingdom, Sweden, Germany, and Canada, courts have created an intermediate category between employee and independent contractor called 'semi-dependent worker' or 'dependent contractor' that enables individuals who are dependent on a single employing entity to enjoy some employment protections (Davidov, 2005). Some scholars are also attempting to redefine the concept of employment to embrace not only an immediate employment status, but also an individual's entire employment relationship (Freedland and Kountouris, 2011).

What Are the Respective Roles of the State and Employers in Providing Job Security and Social Protection?

Given the decline of long-term jobs and the rapid growth of short-term and atypical forms of employment, some commentators have proposed that there be legal regulation to ensure job security. Some, for example, propose that the right to a decent job be treated as ahuman right, enforceable under the International Covenant on Economic, Social and Cultural Rights (Mantouvalou, 2014). Others advocate a return to the twentieth century labor law regimes of continental Europe, in which employees had strong dismissal protection even in the face of severe employer financial distress (Njoya, 2012). However, many have criticized rigid job security protection because it deters employers from hiring employees. There have been efforts to enact legislation that would guarantee work to all who want it by making the government serve as employer as last resort, but these efforts have not been succeeded (Stone, 2014a).

Because long-term jobs are declining, social protections that are linked to employment are also declining in prevalence and scope. Until now, countries have differed as to which types of social protection are provided by employers and which are provided by the state. There are two basic types of systems for providing for unemployment, health insurance, old age security, and other social protection. One is a contributory social insurance program to which employees, employers, and sometimes the state, contributes. An example is found in the Nordic countries, where unemployment insurance funds, administered by the unions, provide generous benefits. Employees voluntarily contribute to the funds, and the state provides additional revenues. The other commonly used mechanism for providing social protection is by means of

universal schemes paid out of general tax revenue. In many countries in Europe health insurance, industrial accident insurance, paid pregnancy leave, and pensions are provided by the state as a universal benefit. However, both contributory and universalistic systems are typically based on employment status. That is, an individual has to have been employed for a specific period of time to qualify for benefits.

The employment-linked nature of social protection has become problematic in a labor market in which individuals no longer have long-term jobs. Because many employees now make frequent moves between employers, and transition periodically from employment to self-employment, training, unemployment, and care work, the employment-linked nature of the social safety net is no longer adequate. Many scholars have proposed alternative arrangements for providing social benefits that do not depend upon long-term employment (Supiot, 2001). For example, Australia has devised a system of social protection that is not linked to employment at all. There the state provides a generous means-tested flat-rate benefit, financed by general tax revenues raised through a progressive income tax, to individuals who are out of work for any reason (O'Donnell, 2013).

Can Unions Operate Effectively in the Workplace If Employers No Longer Provide Long-Term Jobs? If Not, are There New Roles for Unions at the Political Level?

For most of the twentieth century, labor unions have focused on improving wages and conditions for workers. In many places, unions operated at the workplace level and negotiated specific terms for workers of a particular firm. In some countries, unions have also operated at the national or industry level to negotiate terms and conditions that apply across numerous firms. Overall, unions were created by, and supported by, employees who had a long-term investment in their jobs. Thus, because long-term employment has declined, unions have lost much of their effectiveness in the workplace.

New types of labor organizations are developing to further the interests of workers across workplaces or for workers who have multiple and changing employers. For example, in many cities in the United States, unions have formed coalitions with community groups to enact living wage ordinances to improve labor standards for low-wage public sector employees citywide. Presently there are living wage ordinances in over 40 United States cities. Some of these ordinances not only require employers to pay minimum wages at a level that is above the national statutory rate but also require employers to provide their workers with health insurance or other benefits (Stone and Cummings, 2011).

In Japan, there has been an explosion of nontraditional forms of labor organizations. In many places there are city unions that pressure employers in an entire metropolitan area around work-related concerns. There are also women's unions organized to further interests of women workers, and ethnically specific unions to assert pressure on behalf of specific ethic groups. A particularly dynamic form of nonconventional union in Japan is the community union. These unions include part-time workers, temporary workers, immigrant workers, and others who are generally excluded from the traditional enterprise-based unions. In 1990, the Community Union National Network was formed to bring together 72 community unions that were operating in 30 different prefectures. These unions have continued to grow in membership and numbers, and some have affiliated with the National Confederation of Trade Unions or the Japanese Trade Union Congress (Nakamura and Nitta, 2013).

Some commentators have argued that unions have lost much of their effectiveness at the workplace level, and so in spite of that they should devote more efforts to advocating for the interests of workers in the political action. Although unions have always been involved in politics to some extent – more directly in some countries than in others – these commentators advocate that unions shift their emphasis because many important issues of employment regulation are decided at the political level.

In the United States, unions have attempted to compensate for decreased membership and power in the workplace by increasing their expenditures and efforts in the political domain in recent years. However, the ability of unions to operate effectively in the political process has been constrained both by court decisions and by legislative action at the state level (Sachs, 2013).

It is not necessarily the case that unions must choose between seeking power in the bargaining arena and power in the political arena. In the Nordic countries, for example, unions play a large role in the political process and in shaping employment legislation, and they use their power in the political arena to strengthen and reinforce their role in the workplace (Due and Madsen, 2008; Jorgensen, 2002).

Should There be Transnational Labor Regulation to Prevent Firms from Undermining National Labor Standards by Relocating Production to Countries That have the Lowest Levels of Labor Standards?

At present, one of the most pressing issues for employment regulation lies at the transnational level. Businesses that are faced with increased competitive pressures from global trade or presented with enhanced opportunities for direct foreign investment are eager to reduce their labor costs. When they can relocate their operations from restrictive regulatory environments less restrictive ones to lower their labor costs, they do so. To prevent capital flight and to attract jobs, countries engage in regulatory competition by lowering their labor standards regulations and by nonenforcement of the employment regulations they have. Thus firms move around the globe like checkers, jumping from place to place to avoid onerous employment laws, and generating a downward spiral of deregulatory reforms. This dynamic is termed the 'race to the bottom' for labor standards.

Labor organizations and policy-makers in many countries have made numerous proposals for a regime of transnational labor regulation that could impose some minimum global labor standards. Some types of transnational labor regulation have emerged through trade agreements such as the European Union and the Side Agreement to the North American Agreement on Free Trade. In addition, some states have taken unilateral action to condition trading privileges on labor standards, and some companies have adopted voluntary company codes of conduct to prevent labor abuses

(Stone, 1999, 1995). Although each of these mechanisms has had some success in raising labor standards in specific instances, they have not deterred companies from diverting jobs to low-wage companies nor protected workers in those countries against unsafe conditions or abusive labor practices. In part, the reason for the partial success is that each of these efforts relies on a specific country, trade block, or multinational company to serve as the promulgator and enforcer of labor standards; they do not build a universal transnational consensus.

There are several approaches for developing a more universal mechanism for enforcing labor standards at the international level. First, some have called for more powerful transnational institutions that have the power and authority to impose hard law solutions – something like the International Labor Organization (ILO) but with teeth. The argument is that the ILO, or the World Trade Organization (WTO), or some kind of global labor tribunal with sufficient power and legitimacy, should set minimal standards, define fundamental labor rights, and set up an enforcement mechanism akin to the International Criminal Court to adjudicate and remedy allegations of breaches (Barry and Reddy, 2006; Hyde, 2009).

Second, some have argued not for establishing a floor for labor standards, but rather for policies to bolster the safety net so that everyone can have a decent standard of living regardless of trends in labor markets. One scheme of this type is the proposal for a basic income, sometimes called a 'citizenship grant,' which is a socially adequate level of support made available to every citizen without any work requirement or other type of strings attached (Standing, 1999).

A third approach is to advocate policies that encourage a race to the top that can counterbalance a race to the bottom. There have been several different proposals for mechanisms to facilitate a race to the top (Trubek and Trubek, 2005).

Given today's political landscape, each of these proposals faces daunting obstacles. Nonetheless, each of these approaches has not only been theorized and debated in the scholarly literature – each has actually been implemented in some places (Stone, 2014b).

Nonetheless, most trade experts remain adamant that labor considerations should not enter into trade agreements (Bagwati, 1994). Thus the world trading system is evolving without any effective mechanism to protect labor standards, thereby creating an unregulated global labor market in which companies can easily avoid domestic labor unions and national labor regulation. The resulting international race to the bottom will not only depress labor standards in the developed countries, it will also weaken the political influence of labor unions to sustain high domestic labor standards or to press for international ones (Stone, 1995). Given the wide disparities in labor standards across nations, workers in the developed world will be the first ones to suffer from the unfettered quest for corporate profits. However, workers in the developing world will ultimately experience the harshness of a labor market unmediated by redistributive and empowering employment regulation.

The present-day commitment of free trade advocates to separate social concerns from trade is a new formalism that has operated to prevent transnational agencies from addressing social concerns, raising standards, and redressing inequalities. However, as global production and trade continue to proliferate, social pressures are building for more labor law protections both in the developed and in the developing world. Hence employment regulation is not likely to disappear, but it will be transposed as part of the emerging international regime (Stone, 2014b).

See also: Business Law; Insurance and the Law; Poverty Law: United States; Regulation and Administration; Social Insurance: Legal Aspects.

Bibliography

Bagwati, Jagdish, 1994. Fair trade, reciprocity, and harmonization: the new Challenge to the theory and policy of free trade. In: Deardorff, A., Stern, R. (Eds.), Analytical and Negotiating Issues in the Global Trading System. University of Michigan.

Barry, Christian, Reddy, Sanjay G., 2006. Global justice and international economic arrangements: international trade and labor standards: a proposal for linkage. Cornell International Law Journal 39, 545.

Davidov, Guy, 2005. Who is a Worker? Industrial Law Journal 34, 57.

Due, Jesper, Madsen, Joren Steel, 2008. The Danish model of industrial relations: erosion or renewal? Industrial Relations Journal 50, 513.

Freedland, Mark, Kountouris, Nicola, 2011. The Legal Construction of Personal Work Relations. Oxford University Press, London.

Hale, Robert, 1923. Coercion and distribution in a supposedly non-coercive state. Political Science Quarterly 38, 470.

Hyde, Alan, 2009. The international labor organization in the stag hunt for global labor rights. Law & Ethics of Human Rights 3, 154.

Hyde, Alan, 2006. What is labor law? In: Davidov, G., Langille, B. (Eds.), Boundaries and Frontiers of Labour Law. Hart Publishers, Oxford, pp. 37–61.

Jorgensen, Henning, 2002. Consensus, Cooperation, and Conflict: The Policy-making Process in Denmark. Edward Elgar, Cheltenham.

Kahn-Freund, Otto, 1983. Labor and the Law, third ed. Stevens & Sons.

Mantouvalou, Virginia, 2014. The right to non-exploitative work. In: Mantouvalou, V. (Ed.), The Right to Work: Legal and Philosophical Perspectives. Hart Publishing.

Nakamura, Keisuke, Nitta, Michio, 2013. Organizing nonstandard workers in Japan: old players and new players. In: Stone, K., Arthurs, H. (Eds.), Rethinking Workplace Regulation. Russell Sage Foundation Press, New York.

Njoya, Wanjiru, 2012. Job security in a flexible labor market: challenges and possibilities for worker voice. Comparative Labor Law and Policy Journal 33, 459.

O'Donnell, Anthony, 2013. Safety nets and transition assistance: continuity and change in a liberal welfare state. In: Stone, K., Arthurs, H. (Eds.), Rethinking Workplace Regulation. Russell Sage Foundation Press, New York.

Sachs, Benjamin I., 2013. The unbundled union: politics without collective bargaining. Yale Law Journal 123, 149.

Standing, Guy, 1999. Global Labour Flexibility: Seeking Distributive Justice. Macmillan, Basingstoke.

Stone, Katherine V.W., 2014a. A right to work in the United States. In: Mantouvalou, V. (Ed.), The Right to Work: Legal and Philosophical Perspectives. Hart Publishing, Oxford.

Stone, Katherine V.W., 2014b. Globalization and the middle class. In: de Búrca, G., Kilpatrick, C., Scott, J. (Eds.), Critical Perspectives in Global Governance. Hart Publishing, Oxford.

Stone, Katherine V.W., 2013a. The decline of the standard employment contract in the United States: a socio-regulatory perspective. In: Stone, K., Arthurs, H. (Eds.), Rethinking Workplace Regulation. Russell Sage Foundation Press, New York.

Stone, Katherine V.W., 2013b. The decline of the standard employment contract: a review of the evidence. In: Stone, K., Arthurs, H. (Eds.), Rethinking Workplace Regulation. Russell Sage Foundation Press, New York.

Stone, Katherine V.W., 2007. Revisiting the at-will employment doctrine: imposed terms, implied terms, and the normative world of the workplace. Industrial Law Journal 36, 84–101.

Stone, Katherine V.W., 2006. Legal protections for atypical employees: employment law for workers without workplaces and employees without employers. Berkeley Journal of Employment and Labor Law 27, 251.

Stone, Katherine V.W., 2004. Widgets to Digits, Employment Regulation for the Changing Workplace. Cambridge University Press, London.

Stone, Katherine V.W., 1999. To the Yukon and beyond: local laborers in a global labor market. Journal of Small and Emerging Business Law 3, 93.

Stone, Katherine V.W., 1995. Labor and the global economy: four approaches to transnational labor regulation. Michigan Journal of International Law 16, 987.

Stone, Katherine V.W., Arthurs, Harry, 2013. Rethinking Workplace Regulation: Beyond the Standard Contract of Employment. Russell Sage Foundation Press, New York.

Stone, Katherine V.W., Cummings, Scott, 2011. Labor activism in local politics. In: Davidov, G., Langille, B. (Eds.), The Idea of Labour Law. Oxford University Press, London, pp. 273–292.

Supiot, Alain, 2001. Beyond Employment, Changes in Work and the Future of Labour Law in Europe. Oxford University Press, London.

Trubek, David M., Trubek, Louise G., 2005. Hard and soft law in the conception of social Europe: the role of the open method of coordination. European Law Journal 11, 343.

Wedderburn, William, 1986. The Worker and the Law. Pelican Books, London.

Weiss, Manfred, 1987. Labor Law and Industrial Relations in the Federal Republic of Germany. Klewer, Boston.

Empowerment Evaluation

David M Fetterman, San Jose, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Empowerment evaluation is an evaluation approach designed to help communities monitor and evaluate their own performance. It is also used to help groups accomplish their goals. Empowerment evaluation focuses on fostering self-determination and sustainability. It is particularly suited to the evaluation of comprehensive community-based initiatives or place-based initiatives.

Definition

Empowerment evaluation is the use of evaluation concepts, techniques, and findings to foster improvement and self-determination (Fetterman, 1994). An expanded definition is: Empowerment evaluation is an evaluation approach that "aims to increase the probability of achieving program success by (1) providing program stakeholders with the tools for assessing the planning, implementation, and self-evaluation of their program, and (2) mainstreaming evaluation as part of the planning and management of the program/organization" (Wandersman et al., 2005).

Background

Empowerment evaluation was introduced to the field of evaluation in 1993 (Fetterman, 1994). It has been used in remote Amazonian regions as well as corporate offices of Hewlett-Packard in the Silicon Valley. Empowerment evaluation has been used by NASA/Jet Propulsion Laboratory to educate youth about the prototype Mars Rover, townships in South Africa to create sustainable community health initiatives, the US Department of Education's Office of Special Education and Rehabilitation Services to foster self-determination, and with Native American tribes to build technological and economic infrastructures on reservations. It has also been used to address academic distress (Fetterman, 2005b), accreditation in higher education (Fetterman, 2001, 2012), minority tobacco prevention (Fetterman, 2005b; Fetterman et al., 2014a), and medical education (Fetterman, 2009; Fetterman et al., 2010). Empowerment evaluation is international in scope. It is applied in over 16 countries. A sample of the literature situating empowerment evaluation in the field of evaluation is listed below: Alkin and Christie (2004), Altman (1997), Brown (1997), Cousins (2005), Donaldson (2005), Fetterman (1995, 1997, 2001), Fetterman et al. (1996, 2014b), Fetterman and Wandersman (2005, 2007), Fetterman et al. (2010), Patton (1997a, 2005), Scriven (1997, 2005), Sechrest (1997), Stufflebeam (1994), and Wild (1997).

Theories

Empowerment evaluation is guided by empowerment theory, self-determination theory, process use, and theories of use and action. *Empowerment theory* focuses on gaining control of resources in one's environment. It also provides a guide for the role of the empowerment evaluator. *Self-determination* theory highlights specific mechanisms or behaviors that enable the actualization of empowerment. *Process use* cultivates ownership by placing the approach in community and staff members' hands. Finally, the alignment of *theories of use and action* explains how empowerment evaluation helps people 'walk their talk' and produce desired results.

Empowerment Theory

Empowerment theory is about gaining control, obtaining resources, and understanding one's social environment. It is also about problem solving, leadership, and decision making. It operates on many levels, and distinguishing between empowering processes and outcomes is critical. According to Zimmerman (2000):

> Empowerment processes are ones in which attempts to gain control, obtain needed resources, and critically understand one's social environment are fundamental. The process is empowering if it helps people develop skills so they can become independent problem solvers and decision makers. Empowering processes will vary across levels of analysis. For example, empowering processes for individuals might include organizational or community involvement, empowering processes at the organizational level might include shared leadership and decision making, and empowering processes at the community level might include accessible government, media, and other community resources.

Empowerment theory processes contribute to specific outcomes. Linking the processes to outcomes helps groups specify their chain of reasoning. Zimmerman (2000) provides additional insight into the outcome level of analysis to further explicate empowerment theory:

> Empowerment outcomes refer to operationalization of empowerment so we can study the consequences of citizen attempts to gain greater control in their community or the effects of interventions designed to empower participants. Empowered outcomes also differ across levels of analysis. When we are concerned with individuals, outcomes might include situation-specific perceived control, skills, and proactive behaviors. When we are studying organizations, outcomes might include organizational networks, effective resource acquisition, and policy leverage. When we are concerned with community level empowerment, outcomes might include

evidence of pluralism, the existence of organizational coalitions, and accessible community resources.

Role

Zimmerman's (2000) characterization of the community psychologist's *role* in empowerment activities is similar to the role of the empowerment evaluator.

> An empowerment approach to intervention design, implementation, and evaluation redefines the professional's role relationship with the target population. The professional's role becomes one of collaborator and facilitator rather than expert and counselor. As collaborators, professionals learn about the participants through their cultures, their worldviews, and their life struggles. The professional works *with* participants instead of advocating *for* them. The professional's skills, interest, or plans are not imposed on the community; rather, professionals become a resource for a community. This role relationship suggests that what professionals do will depend on the particular place and people with whom they are working, rather than on the technologies that are predetermined to be applied in all situations.

Additional literature on empowerment theory is provided by Zimmerman (2000), Zimmerman et al. (1992), Zimmerman and Rappaport (1988), and Dunst et al. (1992).

Self-Determination

Mithaug's work in this area is typically referred to as self-regulation theory, with self-determination as a guiding concept in his work. For the purposes of clarity and as it relates to instructing empowerment evaluation, self-determination is being used as the umbrella term in this discussion. Self-determination is defined as the ability to chart one's own course in life. It consists of numerous interconnected capabilities, such as the ability to identify and express needs; establish goals or expectations and a plan of action to achieve them; identify resources; make rational choices from various alternative courses of action; take appropriate steps to pursue objectives; evaluate short- and long-term results, including reassessing plans and expectations and taking necessary detours; and persist in the pursuit of those goals. A breakdown at any juncture of this network of capabilities – as well as various environmental factors – can reduce a person's likelihood of being self-determined (see also Bandura (1982) for more detail on issues related to self-efficacy and self-determination). Dennis Mithaug (1991, 1993) has provided extensive work on the topic, focusing on individuals with disabilities. Fetterman and Mithaug's Department of Education work on self-determination and individuals with disabilities provided additional clarity to the concept. Self-determination mechanisms help program staff members and participants implement an empowerment evaluation.

Process Use

Process use assumes that the more that people conduct their own evaluation, the more they own them. The greater the sense of ownership, the more likely people are to consider their findings credible and act on their own recommendations. Empowerment evaluation places evaluation in the hands of community and staff members to facilitate ownership, enhance credibility, and promote action. In addition, a byproduct of this experience is that people learn to think evaluatively (Patton, 2002). This makes them more likely to make decisions and take actions based on their evaluation data.

Theories of Action and Use

Theories that enable comparisons between action and use are essential. Empowerment evaluation relies on the reciprocal relationship between theories of action and use at every step in the process. A *theory of action* is the espoused operating theory about how a program or organization works. This theory of action is compared with a theory of use. The theory of use is the actual program reality, the observable behavior of stakeholders (see Argyris and Schon, 1978; Patton, 1997b). People engaged in empowerment evaluations create a theory of action and test it against the existing theory of use. Because empowerment evaluation is an ongoing and iterative process, stakeholders test their theories of action against theories of use during various microcycles to determine whether their strategies are being implemented as recommended or designed.

These theories are used to identify gross differences between the ideal and the real. For example, communities of empowerment evaluation practice compare their theory of action with their theory of use to determine whether they are even pointing in the same direction. Three common patterns that emerge from this comparison include *in alignment*, *out of alignment*, and *alignment in conflict*. In alignment is when the two theories are parallel or pointed in the same direction. They may be distant or close levels of alignment, but they are on the same general track. Out of alignment occurs when actual practice is divergent from the espoused theory of how things are supposed to work. The theory of use is not simply distant or closely aligned, but actually off target or at least pointed in another direction. Alignment in conflict occurs when the theories of action and use are pointed in diametrically opposite directions. This signals a group or organization in serious trouble or self-denial.

After making the first-level comparison, a gross indicator, to determine whether the theories of action and use are even remotely related to each other, communities of empowerment evaluation practice compare their theory of action with their theory of use in an effort to reduce the gap between them. This assumes they are at least pointed in the same direction. The ideal progression is from distant alignment to close alignment between the two theories. This is the conceptual space where most communities of empowerment evaluation practice strive to accomplish their goals as they close the gap between the theories. The process of empowerment embraces the tension between the two types of theories and offers a means for reconciling incongruities.

Principles

Empowerment evaluation is guided by 10 specific principles (Fetterman and Wandersman, 2005: pp. 1–2, 27–41, 42–72). They include:

1. Improvement – Empowerment evaluation is designed to help people improve program performance; it is designed to help people build on their successes and reevaluate areas meriting attention.
2. Community ownership – Empowerment evaluation values and facilitates community control; use and sustainability are dependent on a sense of ownership.
3. Inclusion – Empowerment evaluation invites involvement, participation, and diversity; contributions come from all levels and walks of life.
4. Democratic participation – Participation and decision making should be open and fair.
5. Social justice – Evaluation can and should be used to address social inequities in society.
6. Community knowledge – Empowerment evaluation respects and values community knowledge.
7. Evidence-based strategies – Empowerment evaluation respects and uses the knowledge base of scholars (in conjunction with community knowledge).
8. Capacity building – Empowerment evaluation is designed to enhance stakeholders' ability to conduct evaluation and to improve program planning and implementation.
9. Organizational learning – Data should be used to evaluate new practices, inform decision making, and implement program practices; empowerment evaluation is used to help organizations learn from their experience (building on successes, learning from mistakes, and making midcourse corrections).
10. Accountability – Empowerment evaluation is focused on outcomes and accountability; empowerment evaluation functions within the context of existing policies, standards, and measures of accountability; empowerment evaluations ask: did the program accomplish its objectives?

Empowerment evaluation principles help evaluators and community members make decisions that are in alignment with the larger purpose or goals associated with capacity building and self-determination. The principles of inclusion, for example, remind evaluators and community members to include rather than exclude members of the community, even though fiscal, logistic, and personality factors might suggest otherwise. The capacity building principle reminds the evaluator to provide community members with the opportunity to collect their own data, even though it might initially be faster and easier for the evaluator to collect the same information. The accountability principle guides community members to hold one another accountable. It also situates the evaluation within the context of external requirements and credible results or outcomes (see Fetterman, 2005a: p. 2).

Concepts

Key concepts guiding empowerment evaluation include critical friends, cultures of evidence, cycles of reflection and action, communities of learners, and reflective practitioners. These concepts are influenced by traditional organizational development and transformation theorists including Argyris and Schon (1978) and Senge (1994), as well as evaluators associated with organizational learning (Preskill and Torres, 1999). A critical friend is an evaluator who facilitates the process and steps of empowerment evaluation. They believe in the purpose of the program, but provide constructive feedback. They help to ensure the evaluation remains organized, rigorous, and honest.

Empowerment evaluators help cultivate a culture of evidence by asking people why they believe what they believe. Community members and program participants are asked for evidence or documentation at every stage, so that it becomes normal and expected to have data to support one's opinions and views. Cycles of reflection and action involve ongoing phases of analysis, decision making, and implementation (based on evaluation findings). It is a cyclical process. Programs are dynamic, not static, and require continual feedback as they change and evolve. Empowerment evaluation is successful when it is institutionalized and becomes a normal part of the planning and management of the program.

Empowerment evaluation is driven by a group process. It is a community of learners. The group learns from each other, serving as their own peer review group, critical friend, resource, and norming mechanism. Individual members of the group hold each other accountable concerning progress toward stated goals. Finally, empowerment evaluations help create reflective practitioners. Reflective practitioners use data to inform their decisions and actions concerning their own daily activities. This produces a self-aware and self-actualized individual who has the capacity to apply this worldview to all aspects of their life. As individuals develop and enhance their own capacity, they improve the quality of the group's exchange, deliberation, and action plans.

Steps

There are many ways in which to implement an empowerment evaluation. In fact, empowerment evaluation has accumulated a warehouse of useful tools. The three-step (Fetterman, 2001) and ten-step (Chinman et al., 2004) approaches to empowerment evaluation are the most popular tools in the collection.

Three-Step

The three-step approach includes helping a group: (1) establish their mission; (2) take stock of their current status; and (3) plan for the future. The popularity of this particular approach is in part a result of its simplicity, effectiveness, and transparency.

Mission

The group comes to a consensus concerning their mission or values. This gives them a shared vision of what is important to them and where they want to go. The empowerment evaluator facilitates this process by asking participants to generate statements that reflect their mission. These phrases are recorded on a poster sheet of paper (and may be projected on an LCD projector depending on the technology available). These

phrases are used to draft a mission statement (crafted by a member of the group and the empowerment evaluator). The draft is circulated among the group. They are asked to 'approve' it and/or suggest specific changes in wording as needed. A consensus about the mission statement helps the group think clearly about their self-assessment and plans for the future. It anchors the group in common values.

Taking Stock

After coming to a consensus about the mission, the group evaluates their efforts (within the context of a set of shared values). First, the empowerment evaluator helps members of the group generate a list of the most important activities required to accomplish organizational or programmatic goals. The empowerment evaluator gives each participant five dot stickers, and asks the participants to place them by the activities they think are the most important to accomplish programmatic and organizational goals (and thus the most important to evaluate as a group from that point on). They can put one sticker on five different activities or all five on one activity if they are concerned that activity will not get enough votes. The top 10 items with the most dots represent the results of the prioritization part of taking stock. The 10 activities represent the heart of part two of taking stock: rating.

The empowerment evaluator asks participants in the group to rate how well they are doing concerning each of the activities selected, using a 1 (low) to 10 (high) scale. The columns are averaged horizontally and vertically. Vertically, the group can see who is typically optimistic and/or pessimistic. This helps the group calibrate or evaluate the ratings and opinions of each individual member. It helps the group establish norms. Horizontally, the averages provide the group with a consolidated view of how well (or poorly) things are going. The empowerment evaluator facilitates a discussion and dialogue about the ratings, asking participants why they gave a certain activity a 3 or 7.

The dialogue about the ratings is one of the most important parts of the process. In addition to clarifying issues, evidence is used to support viewpoints and 'sacred cows' are surfaced and examined during dialogue. Moreover, the process of specifying the reason or evidence for a rating provides the group with a more efficient and focused manner of identifying what needs to be done next, during the planning for the future step of the process. Instead of generating an unwieldy list of strategies and solutions that may or may not be relevant to the issues at hand, the group can focus its energies on the specific concerns and reasons for a low rating that were raised in the dialogue or exchange.

Planning for the Future

Many evaluations conclude at the taking stock phase. However, taking stock is a baseline and a launching off point for the rest of the evaluation. After rating and discussing programmatic activities it is important to do something about the findings. It is time to plan for the future. This step involves generating goals, strategies, and credible evidence (to determine if the strategies are being implemented and if they are effective). The goals are directly related to the activities selected in the taking stock step. For example, if communication was selected, rated, and discussed, then communication (or improving communication) should be one of the goals. The strategies emerge from the taking stock discussion, as well, as noted earlier. For example, if communication received a low rating and one of the reasons was because the group never had agendas for their meetings, then preparing agendas might become a recommended strategy in the planning for the future exercise.

Many programs, projects, and evaluations fail at this stage for lack of individual and group accountability. Individuals who spoke eloquently and/or emotionally about a certain topic should be asked to volunteer to lead specific task forces to respond to identified problems or concerns. They do not have to complete the task. However, they are responsible for taking the lead in a circumscribed area (a specific goal) and reporting the status of the effort periodically at ongoing management meetings. Similarly, the group should make a commitment to reviewing the status of these new strategies as a group (and be willing to make midcourse corrections if they are not working). Conventional and innovative evaluation tools are used to monitor the strategies, including online surveys, focus groups, interviews, as well as the use of a quasi-experimental design (if appropriate). In addition, program-specific metrics are developed, using baselines, benchmarks or milestones, and goals (as deemed useful and appropriate). For example, a minority tobacco prevention program empowerment evaluation in Arkansas has established:

1. Baselines (the number of people using tobacco in their community)
2. Goals (the number of people they plan to help stop using tobacco by the end of the year)
3. Benchmarks or milestones (the number of people they expect to help stop using tobacco each month)
4. Actual performance (they record the number of people they help to stop using tobacco and compare their figures with their goals and benchmarks to determine if they are making progress or need assistance)

These metrics are used to help a community monitor program implementation efforts and enable program staff and community members to make midcourse corrections and substitute ineffective strategies for potentially more effective ones as needed. These data are also invaluable when the group conducts a second taking stock exercise (3–6 months later) to determine if they are making progress toward their desired goals and objectives. Additional metrics enable community members to compare, for example, their baseline assessments with their benchmarks or expected points of progress, as well as their goals. See also Fetterman (2003a, 2010) for the use of descriptive or ethnographic techniques.

The 10-Step

The 10-step approach is another useful tool. It is called Getting-to-Outcomes (GTO®) (Chinman et al., 2004). GTO is a method for planning, implementing, and evaluating programs in order to achieve success. GTO provides a framework of 10 accountability questions. When answered with quality it provides a pathway to help programs achieve results and demonstrate accountability (see **Table 1**).

Table 1 GTO accountability questions

1. Needs assessment	What are the needs and resources in your organization/school/community/state?
2. Goal setting	What are the goals, target population, and desired outcomes (objectives) for your school/community/state?
3. Science and best practices	How does the intervention incorporate knowledge or science and best practices in this area?
4. Collaboration; cultural competence	How does the intervention fit with other programs already being offered?
5. Capacity building	What capacities do you need to put this intervention into place with quality?
6. Planning	How will this intervention be carried out?
7. Implementation/process evaluation	How will the quality of implementation be assessed?
8. Outcome and impact evaluation	How well did the intervention work?
9. Total quality management; continuous quality improvement	How will continuous quality improvement strategies be incorporated?
10. Sustainability and institutionalization	If the intervention is (or components are) successful, how will the intervention be sustained?

These steps and guiding questions are designed to build evaluation capacity and improve the probability of program success and sustainability. There are many GTO manuals with worksheets designed to address how to answer each of the 10 questions (e.g., Chinman et al., 2004). GTO has been used in multiple domains, including substance abuse prevention, underage drinking (Imm et al., 2006), promoting positive youth development (Fischer et al., 2006), and home visiting programs (Mattox et al., 2013).

Clarifying Issues in Empowerment Evaluation

A sample of issues concerning the use of empowerment evaluation is presented below to provide additional conceptual clarity. They range from the locus of control to the audience for empowerment evaluation and are briefly discussed below.

1. *Empowerment*. People empower themselves. A common misconception about empowerment evaluation is that it empowers either individuals or groups. Empowerment evaluation cannot empower anyone. Empowerment evaluation simply provides the tools and environment conducive to empowering oneself.
2. *Objectivity and advocacy*. Empowerment evaluation is transparent, brings bias to the surface, and generates meaningful data to inform decision making. These findings are used by staff, community members, and other relevant parties to advocate for their programs or communities as data merit.
3. *Consumer focus*. Consumers (community members, program staff, and participants) are the driving force or focus in empowerment evaluation. However, evaluators and donors remain an integral part of empowerment evaluation
4. *Internal versus external evaluation*. Empowerment evaluation (internal evaluation) and traditional forms of evaluation (typically external evaluation) can be mutually reinforcing. They are not mutually exclusive. However, external evaluations should be rooted in internal concerns; otherwise, they may divert program staff, participants, and resources from the most relevant issues given the organization's stage of development.
5. *Purpose*. Empowerment evaluation's most significant contribution is to program or community development. However, it makes a strong contribution to accountability by cultivating internal accountability. This contribution remains long after an episodic and often anticipated external examination.
6. *Bias*. Empowerment evaluations are typically more critical of their own programs than external examinations (which are often influenced by their interest in returning for a follow-up or extended engagement). Empowerment evaluations provide people with a window of opportunity to address long-standing issues of dysfunction and inefficiency in their own organizations. In addition, the process is inclusive and transparent, open to critique and review, and makes it difficult to keep people from publicly 'speaking their truth.'
7. *Outcomes*. Empowerment evaluations are highly collaborative and participatory in nature. However, the bottom line remains: did you accomplish the desired results? Empowerment evaluations are conducted within the context of what people are already being held accountable for in their communities or workplaces. This makes the entire process more credible and authentic.
8. *Audience*. Empowerment evaluators believe that everyone can benefit from being more empowered. Although much of empowerment evaluation's focus has been on disenfranchised populations, empowerment evaluators assume all people can benefit from taking greater control of their lives, particularly from a positive psychological growth perspective.

Additional points providing greater conceptual clarity, as well as additional discussion about methodological specificity and outcomes, are discussed in detail in the literature (Fetterman, 2001; Fetterman and Wandersman, 2005, 2007). Christie (2003) has also revealed a unique contribution that empowerment evaluation has made to the field and how to differentiate its contribution from similar approaches (see also Fetterman, 2003b).

Conclusion

Empowerment evaluation is firmly rooted in the field of evaluation. The approach has been applied internationally, debated in refereed journals, and is the subject of longitudinal research. The first empowerment evaluation book *Empowerment Evaluation: Knowledge and Tools for Self-assessment and Accountability* (Fetterman et al., 1996) was controversial since it proposed turning evaluation on its head – allowing community and staff members to conduct their own evaluations. That book helped launch the approach, focusing on breadth and scope of

empowerment evaluation, as well as useful tools. The second empowerment evaluation book *Foundations of Empowerment Evaluation* (Fetterman, 2001) consolidated arguments and pertinent discussions in the field. It also applied the utility, feasibility, propriety, and accuracy standards (based on the Joint Committee on Standards for Educational Evaluation). The third empowerment evaluation book *Empowerment Evaluation Principles in Practice* (Fetterman and Wandersman, 2005) introduced the 10 principles guiding empowerment evaluation. The fourth book *Empowerment Evaluation: Knowledge and Tools for Self-Assessment, Evaluation Capacity Building, and Accountability* (Fetterman et al., 2014b) represents the product of over 20 years of experience, providing additional conceptual clarity and methodological specificity. Empowerment evaluation continues to be a self-assessment approach designed to help people help themselves, using evaluation to build their capacity and accomplish their goals.

See also: Academic Research and Employment: Recent Changes in Europe and the United States; Applied Social Research, History of; Applied Sociology; Case Studies; Cultural Participation, Trends in; Egalitarianism; Ethnography in Applied Social Research; Evidence-Based Practice; Implementation Science; Organizations and Culture; Prevention Research.

Bibliography

Alkin, M., Christie, C., 2004. An evaluation theory tree. In: Alkin, M. (Ed.), Evaluation Roots: Tracing Theorists' Views and Influences. Sage, Thousand Oaks, CA, pp. 381–392.

Altman, D., 1997. Review of the book empowerment evaluation: knowledge and tools for self-assessment and accountability. Community Psychologist 30 (4), 16–17. Retrieved from: http://www.davidfetterman.com/AltmanBookReview.htm.

Argyris, C., Schon, D.A., 1978. Organizational Learning: A Theory of Action Perspective. Addison-Wesley, Reading, MS.

Bandura, A., 1982. Self-efficacy mechanism in human agency. American Psychologist 37, 122–147.

Brown, J., 1997. Review of the book empowerment evaluation: knowledge and tools for self-assessment and accountability. Health Education & Behavior 24 (3), 388–391. Retreived from: http://www.davidfetterman.com/BrownBookReview1.htm.

Chinman, M., Imm, P., Wandersman, A., 2004. Getting to Outcomes: Promoting Accountability through Methods and Tools for Planning, Implementation, and Evaluation. RAND Corporation, Santa Monica, CA. http://www.rand.org/pubs/technical_reports/TR101/.

Christie, C.A., Spring 2003. What guides evaluation? A study of how evaluation practice maps onto evaluation theory. In: Christie, C.A. (Ed.), The Practice-Theory Relationship in Evaluation (No. 97). Jossey-Bass, San Francisco.

Cousins, B., 2005. Will the real empowerment evaluation please stand up? A critical friend perspective. In: Fetterman, D.M., Wandersman, A. (Eds.), Empowerment Evaluation Principles in Practice. Guilford, New York, pp. 183–208.

Donaldson, S., 2005. Book Review of Empowerment Evaluation Principles in Practice. Amazon. http://www.amazon.ca/Empowerment-Evaluation-Principles-Practice-Fetterman/dp/1593851146.

Dunst, C.J., Trivette, C.M., LaPointe, N., 1992. Toward Clarification of the Meaning and Key Elements of Empowerment.

Fetterman, D.M., 1989. Ethnography: Step by Step. Sage, Newbury Park, CA.

Fetterman, D.M., 1994. Empowerment evaluation. Evaluation Practice 15 (1), 1–15.

Fetterman, D.M., 1995. In response to Dr. Daniel Stufflebeam's empowerment evaluation, objectivist evaluation, and evaluation standards: where the future of evaluation should not go, where it needs to go, October, 1994, 321–338. American Journal of Evaluation 16, 179–199. Retrieved from: http://www.davidfetterman.com/dfresponsetostufflebeam.pdf.

Fetterman, D.M., 1997. Empowerment evaluation: a response to Patton and Scriven. American Journal of Evaluation 18, 253–266. Retrieved from: http://www.davidfetterman.com/Fettermanresponsepattonscriven.pdf.

Fetterman, D.M., 2001. Foundations of Empowerment Evaluation. Sage, Thousand Oaks, CA.

Fetterman, D.M., 2003a. Ethnography. In: Lewis-Beck, M., Bryman, A., Futing Liao, T. (Eds.), Encyclopedia of Social Science Research. Sage, Thousand Oaks, CA.

Fetterman, D.M., 2003b. Fetterman-House: A Process Use Distinction and a Theory. In: New Directions for Evaluation, vol. 97. Jossey-Bass, San Francisco.

Fetterman, D.M., 2005a. A window into the heart and soul of empowerment evaluation: looking through the lens of empowerment evaluation principles. In: Fetterman, D.M., Wandersman, A. (Eds.), Empowerment Evaluation Principles in Practice. Guilford Publications, New York.

Fetterman, D.M., 2005b. Empowerment evaluation: from the digital divide to academic distress. In: Fetterman, D.M., Wandersman, A. (Eds.), Empowerment Evaluation Principles in Practice. Guilford Publications, New York.

Fetterman, D.M., 2009. Empowerment evaluation at the Stanford University School of Medicine: using a critical friend to improve the clerkship experience. Ensaio: Aval. Pol. Publ. Educ., Rio de Janeiro 17 (63), 197–204.

Fetterman, D.M., 2010. Ethnography: Step by Step, third ed. Sage, Thousand Oaks, CA.

Fetterman, D.M., 2012. Empowerment evaluation and accreditation case examples: California Institute of Integral Studies and Stanford University. In: Secolsky, C. (Ed.), Measurement and Evaluation in Higher Education. Routledge, London.

Fetterman, D.M., Deitz, J., Gesundheit, N., 2010. Empowerment evaluation: a collaborative approach to evaluating and transforming a medical school curriculum. Academic Medicine 85 (5), 813–820.

Fetterman, D.M., Delaney, L., Tremain, B., Evans-Lee, M., 2014a. Empowerment evaluation and evaluation capacity building in a 10-year tobacco prevention initiative. In: Fetterman, D.M., Kaftarian, S., Wandersman, A. (Eds.), Empowerment Evaluation: Knowledge and Tools for Self-Assessment, Evaluation Capacity Building, and Accountability. Sage, Thousand Oaks, CA.

Fetterman, D.M., Kaftarian, S., Wandersman, A., 2014b. Empowerment Evaluation: Knowledge and Tools for Self-Assessment, Evaluation Capacity Building, and Accountability. Sage, Thousand Oaks, CA.

Fetterman, D.M., Kaftarian, S., Wandersman, A., 1996. Empowerment Evaluation: Knowledge and Tools for Self-Assessment and Accountability. Sage, Thousand Oaks, CA.

Fetterman, D.M., Pitman, M.A., 1986. Educational Evaluation: Ethnography in Theory, Practice, and Politics. Sage, Beverly Hills, CA.

Fetterman, D.M., Wandersman, A., 2005. Empowerment Evaluation Principles in Practice. Guilford Publications, New York.

Fetterman, D.M., Wandersman, A., 2007. Empowerment evaluation: yesterday, today, and tomorrow. American Journal of Evaluation 28 (2), 179–198.

Fischer, D., Imm, P., Chinman, M., Wandersman, A., 2006. Getting to Outcomes with Developmental Assets: Ten Steps to Measuring Success in Youth Programs and Communities. Search, Minneapolis, MN.

House, E.R., Howe, K.R. (Eds.), 2000. Deliberative Democratic Evaluation. New Directions for Evaluation, vol. 85. Jossey-Bass, San Francisco.

Imm, P., Chinman, M., Wandersman, A., Rosenbloodm, D., Guckenburg, S., Leis, R., 2006. Preventing Underage Drinking: Using Getting to Outcomes with the SAMHSA Strategic Prevention Framework to Achieve Results. RAND Corporation, Santa Monica, CA.

Mattox, T., Hunter, S.B., Kilburn, M.R., Wiseman, S.H., 2013. Getting to Outcomes for Home Visiting: How to Plan, Implement, and Evaluate a Program in Your Community to Support Parents and Their Young Children. RAND Corporation, Santa Monica, CA. http://www.rand.org/pubs/tools/TL114.html.

Mithaug, D.E., 1991. Self-determined Kids: Raising Satisfied and Successful Children. Macmillan, New York.

Mithaug, D.E., 1993. Self-Regulation Theory: How Optimal Adjustment Maximizes Gain. Praeger, New York.

Patton, M.Q., 1997a. Toward distinguishing empowerment evaluation and placing it in a larger context. Evaluation Practice 15 (3), 311–320. Available on-line at: http://www.davidfetterman.com/pattonbkreview1997.pdf.

Patton, M.Q., 1997b. Utilization-Focused Evaluation: The New Century Text, third ed. Sage, Thousand Oaks, CA.

Patton, M.Q., 1980. Qualitative Evaluation Methods. Sage, Beverly Hills, CA.

Patton, M.Q., 2002. Qualitative Research and Evaluation Methods. Sage, Thousand Oaks, CA.

Patton, M.Q., 2005. Toward distinguishing empowerment evaluation and placing it in a larger context: take two. American Journal of Evaluation 26, 408–414.

Preskill, H., Torres, R., 1999. Evaluative Inquiry for Organizational Learning. Sage, Thousand Oaks, CA.

Sechrest, L., 1997. Review of the book Empowerment evaluation: knowledge and tools for self-assessment and accountability. Environment and Behavior 29 (3), 422–426. Retreived from: http://www.davidfetterman.com/SechrestBookReview.htm.

Senge, P., 1994. The Fifth Discipline: The Art and Practice of the Learning Organization. Doubleday, New York.

Scriven, M., 1997. Empowerment evaluation examined. Evaluation Practice 18 (2), 165–175. Available online at: http://www.davidfetterman.com/scrivenbkreview1997.pdf.

Scriven, M., 2005. Review of the book: empowerment evaluation principles in practice. American Journal of Evaluation 26 (3), 415–417.

Stufflebeam, D., 1994. Empowerment evaluation, objectivist evaluation, and evaluation standards: where the future of evaluation should not go and where it needs to go. Evaluation Practice 15 (3), 321–338. Retrieved from: http://www.davidfetterman.com/stufflebeambkreview.pdf.

Wandersman, A., Snell-Johns, J., Lentz, B., Fetterman, D.M., Keener, D.C., Livet, M., Imm, P.S., Flaspohler, P., 2005. The principles of empowerment evaluation. In: Fetterman, D.M., Wandersman, A. (Eds.), Empowerment Evaluation Principles in Practice. Guilford Publications, New York, p. 27.

Wild, T., 1997. Review of empowerment evaluation: knowledge and tools for self-assessment and accountability. Canadian Journal of Program Evaluation 11 (2), 170–172. Available on-line at: http://www.davidfetterman.com/WildBookReview.htm.

Zimmerman, M.A., 2000. Empowerment theory: psychological, organizational, and community levels of analysis. In: Rappaport, J., Seldman, E. (Eds.), Handbook of Community Psychology. Kluwer Academic/Plenum, New York, pp. 2–45.

Zimmerman, M.A., Israel, B.A., Schulz, A., Checkoway, B., 1992. Further explorations in empowerment theory: an empirical analysis of psychological empowerment. American Journal of Community Psychology 20 (6), 707–727.

Zimmerman, M.A., Rappaport, J., 1988. Citizen participation, perceived control, and psychological empowerment. American Journal of Community Psychology 16 (5), 725–750.

Encyclopedias, Handbooks, and Dictionaries

Alan Sica, Pennsylvania State University, State College, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews revolutionary changes, which have occurred since 2000 regarding the creation and publication of reference books and online sources. Factors that have altered the way that such books, especially those pertaining to the social sciences, include the increased speed of composition due to word-processing technology, the downsizing of libraries as they switch from printed to electronic reference sources, and the increased pressure upon academics to publish rapidly, especially in Britain. The article also offers a concise history of social science reference works in several disciplines, beginning in the eighteenth century and ending with the present.

The Old World of Reference Publishing

Historic Overview of Reference Works

According to one informed estimate, if a copy of every important reference work known in Western history were set side by side in chronological order, some two miles of shelf space would be required for their display. Ever since Plato and Aristotle first called for the assembling of what they named 'encyclopedia' – a challenge quickly accepted by one of Plato's nephews, Speusippus (d. 339 BC) – intellectuals have tried to summarize and codify what they knew, either for general literate audiences or for specialized groups of colleagues. Of these two miles of systematized learning, those volumes that treat the social sciences would appear in only the last several hundred yards of material, in the material produced since the Enlightenment, since the social sciences in their modern form are largely products of the Enlightenment. Clearly, social science was not part of the trivium/quadrivium arrangement of classical and renaissance learning, and official academic status was accorded these fields only during the last century in most research universities.

The history of encyclopedias, handbooks, and dictionaries (designations which until at least the eighteenth century were used interchangeably) is distinguished not only by sheer quantity of printed material, but also by the quality of thinkers and writers who created and contributed to them. Pliny the Elder's *Natural History*, for instance, served as a model of organizing such material for 1700 years after its appearance in AD 77, despite an odd mixing of fact and fancy throughout its 37 books. Francis Bacon devised in 1620, a huge encyclopedia of 130 sections that was never completed, but his mode of organizing knowledge was still consulted by Diderot and later compilers 130 years later. In 1817, Samuel Taylor Coleridge also put forth an ambitious reorganizing of all human knowledge, which he interestingly titled *Encyclopædia Metropolitana*.

The urge, then, to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West. Even in AD 551 when the love of knowledge and external conditions for its preservation reached an apparent nadir, Cassiodorus saw to it that the labors of the monastery he founded at Vivarium be dedicated principally to the preservation of worthwhile information, producing the 2-volume *Institutes of Divine and Secular Literature*. Vincent of Beauvois has long been identified as the finest encyclopedist of the Middle Ages, producing his *Speculum majus* (*The Greater Mirror*) in 1244, made up of 10 000 chapters in 80 books, an achievement that held the field for sheer scope until the eighteenth century. Alphabetical as opposed to categorically arranged encyclopedias began to appear at the turn of the first millennium, though most have followed the latter rather than the former method of organization. And the first grouping of all knowledge within a format that is recognizably modern, along alphabetical lines, was the *Grand Dictionnaire Historique* produced by Louis Moreri in 1674. The audience for such a work was already developed, for his venture went through six editions in the following 17 years, followed by still more editions in English.

Growing Availability of Reference Works

It was not always so difficult to name the leading works to which scholars or novices could turn when searching for basic information regarding the social sciences, especially if one strays beyond the confines of the English language. There was a time, of course, when dictionaries, handbooks, and encyclopedias (which in Greek means 'general education') could have been viewed as a special genre of scholarly display, with Samuel Johnson's *Dictionary* (1755), Diderot's *Encyclopedia* (1751), or Voltaire's *Philosophical Dictionary* (1764) initially setting the tone for many modern versions that followed. As noted above, for some centuries prior to the Enlightenment, the words 'dictionary' and 'encyclopedia' were used synonymously, lacking precise meaning, e.g., when Rabelais around 1534 introduced the word into French literature, in *Pantagruel*: "he has opened for me the true well and abyss of encyclopedic knowledge" (*The Complete Works of François Rabelais*, University of California Press, 1991, p. 201). Though Rabelais meant simply 'knowledge' when using the term, within only 25 years a German, Paul Scalich, had published his *Encyclopaedia*, the scope of which conformed to today's usage. Since then, of course, untold thousands of such works have appeared in every major language.

But unlike scholarly articles and books, rolling off the presses with merciless regularity, substantial reference works have until recently been rarely enough published that they eluded the standard procedure of critical appraisal brought to

bear on most nonreference works. And they are hardly the innocent, bland summaries, which the naive reader might take them to be. Each new entry into this crowded market must either claim to bring together for the first time a body of knowledge which to that point had not been assembled by quite the same rules as those which constitute the 'new' rubric, or it must offer some other, less lofty advantage, such as lower price or ease of use. All of this is currently going in the world of social science publishing, and at an unprecedented rate, mostly owing, one suspects, to the huge composing advantages supplied by personal computers.

The positive result of the new technology lies in the superabundance of suddenly available reference resources, some of which will surely define for the first time entire realms of intellectual inquiry. A less sanguine prospect, however, is the burden this torrent of innovative materials places on those who most evaluate or use them. As the range of choices widens, the provision of usable criticism becomes ever more costly, and probably less reliable.

This was not always the case, at least regarding the social sciences. A fairly limited selection of well-known reference tools were at the disposal of researchers in the late nineteenth and early twentieth centuries, which gave this growing general zone of study a tidy coherence that today is all but gone. And whereas such an orderly presentation of knowledge, then common to a given discipline or subdiscipline, might today seem either unnecessary or epistemologically indefensible, for the founders of the social sciences (*Geisteswissenschaften*) 100 years ago, the task of proving to more established disciplines that intellectual coherence indeed existed within their upstart fields became a truly vital effort at self-preservation.

Systemization of Reference Works

Within what we could now broadly call sociology, for example, systematizations appeared fairly early. This is interesting for a number of reasons, not least of which is the fact that sociology as a discipline was defined and taught (particularly in the U.S.) some time before its sister disciplines within the social sciences. It often played the role of avant-garde as this new way of carving up the academic world began rearranging disciplinary boundaries. While many of today's practitioners are familiar, at least with the titles, of the *Encyclopedia of the Social Sciences* and *The International Encyclopedia of the Social Sciences*, or Robert E.L. Faris's (Ed.) *Handbook of Modern Sociology* (1964) or Neil Smelser's (Ed.) *Handbook of Sociology* (1988), most are not aware of this genre's substantial tradition, especially as produced in Europe. It is now commonly supposed that Faris' was the first handbook of American sociology to be published, since it was widely distributed at the time and was fortunate in appearing just as sociology became a 'growth industry' on campuses worldwide. But even some reviewers at the time knew better, e.g., Joseph Gittler's opening line in his review: "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one" (Gittler, 1965: p. 335). But it must have seemed like much less of a 'milestone in the field of sociology' – as the publisher had characterized the *Handbook* through its mailed advertisements to American Sociological Association (ASA) members – if one considers the substance of the reviews it received. The book was probably taken as a handy cribbing device for grad students and young instructors, glanced at perfunctorily, then parked on the shelf with similar volumes which Rand McNally had become good at marketing.

As Gittler was hinting, the history of sophisticated handbooks that sought to introduce audiences to new trends of analysis is in fact long and distinguished, though it is quite true that only few of them were in English. Carl von Rotteck and Carl Welcker edited the *Staats-Lexikon oder Encyclopädie der Staatswissenschaften* (1834–43), the oldest I have found, adding up to 15 volumes at 800 pages each. (Even though the social sciences hardly existed at that time, formally speaking, the subject matter in the *Encyclopädie*, as well as the other sources named below, included much of what we would now recognize as such.) In 1852 and 1853, Charles Coquelin oversaw production of the *Dictionaire de L'economie Politique* in 2 volumes (1858 pages). Meanwhile in Italy, Gerolamo Boccardo brought out *Dizionario della economnia politica e del commercio: Cosi teorico come practico* (1857) in 4 volumes and 2800 pages, an equally lavish item.

The predecessor to Weber's famous *Grundriss der Sozialökonomik* was Gustav Schoenberg's *Handbuch der politischen oekonomie* (1882–85) in 3 volumes. Surely the most important of these early handbooks was edited by Conrad et al., *Handwörterbuch der Staatswissenschaften* (1895), an enormous work in 7 volumes and about 6000 pages. Supplements were added in 1895 and 1897 of 2000 pages, and later editions (1898 and 1909–11) included two versions of Weber's famous *Agrarverhältnisse im Altertum* (first translated in 1976 as *The Agrarian Sociology of Ancient Civilizations*), the initial version 27 pages long, the second 136 pages. Ludwig Elster, part of Conrad's team, published a shorter version of the monumental work in only 2 volumes (1898), merely 2100 pages in length. The French were not idle, with Léon Say and Joseph Chailley's *Nouveau Dictionnaire d'Economie Politique* (1891/1892), amounting to 2-fat volumes of 2800 pages, and a supplement in 1900 of 271 more. Finally, the famous Robert Harry Inglis Palgrave edited *Dictionary of Political Economy* (1894, 1900–1901), which in its last version came to 3 volumes and 2500 pages. This was soon followed in 1897 by William D.P. Bliss' *Encyclopedia of Social Reform*, 1 volume of 1439 pages (1908). What all this shows is that the notion of social science compendia has been around quite a while, and that most of the nineteenth-century European models were enormous publishing achievements. With the new century, ambitions and costs both shrank considerably.

The New World of Reference Publishing

Recent Changes in Reference Publishing

In the dozen years since the first version of this article was written, the world of reference volumes has undergone a revolution more profound than any in living memory, and perhaps more radical in its ultimate effect than any since Diderot released his 'long-delayed *Prospectus*' for the *Encyclopédie* in 1750 (Furbank, 1992: p. 74). Diderot, of course, did not invent the idea of a scholars' encyclopedia, but he improved existing models to such an extent that his, in print to this day, is regarded as 'the first.' The sea-change that recently overtook reference work production is of an even greater magnitude than that registered by the appearance of Diderot's great work – so

great, in fact, that the Era of the Reference Book as known to scholars of the past 200 years may be at its end. That *The International Encyclopedia of the Social and Behavioral Sciences*, 2nd edition, was initially to be marketed exclusively in an electronic edition, whereas its first version was also available as a set of books, hints at some of these changes, although it is certainly a pertinent observation that this decision was rescinded and the current work has been made available in both hard copy and electronic formats.

Several disparate forces conjoined to alter the nature of social science reference works such that the very idea of 'reference' has inalterably shifted from one format, accepted for centuries, to another that is truly new. Prior to the global adoption of Wikipedia, Google Scholar, Google Books, and their many offspring, reference works were typically compiled over many years, the creative province of expert scholars and scientists, offered mainly in printed form as books, very costly to produce, and rarely revised. Cheap word-processing and its ultimate gift to writers, the 'copy and paste' function, has made it easy to borrow creatively from a world of source material, including even plagiarizing oneself for disparate occasions. Naturally, these technical changes have sped up and simplified production by a very large factor.

The Shift toward Electronic Libraries

Though the beloved 11th edition of the *Encyclopædia Britannica* (1910) did not include an article about its own origins and greatness, the same cannot be said for its replacement, the self-reflexive Wikipedia. As the entry for Wikipedia says about itself: it was "launched on January 15, 2001 … [its] departure from the expert-driven style of encyclopedia building and the presence of a large body of unacademic content received extensive attention in print media." Meanwhile, Google Books, launched in October 2004 amidst great ambivalence and anxiety from publishers, authors, and scholars, had by March 2012 scanned over 20 million books into its globalized database. What began with obliging help from only a handful of major research libraries in the Anglophone world soon grew into a surprising flood of participating institutions. This development also contributed to the wholesale destruction of printed materials in libraries pressed, as they always are, for space, since once a book was scanned by Google Books, it was assumed to be forever available in that form. The fact scanning a book is not always done perfectly (omitted pages, blurred images, lack of color reproduction, etc.), or that not everyone has access to printers should they need a paper copy, has not stopped libraries from ridding themselves of their printed books at an astonishing rate, including, of course, their reference materials. In fact, discarding 'old' reference materials became a priority among avant-garde officials at large libraries, ever hungry for more shelf space. Only a Don Quixote fighting for the printed form could object to this apparently 'rational' welcoming of digitization and few have appeared to rescue 'old' books from landfills across the country.

Digitization of Content and Labor

While Wikipedia developed its early articles by importing into its site the nonpareil 11th edition of the *Britannica*, Google Books originated no new content, instead scanning millions of volumes at scores of worksites into a form that allows readers selective, and often very limited, access to pages within scanned volumes, and to search for terms within each book. Despite legal battles in the U.S. and abroad (especially in France), the 'googlization' of everything in print proceeds apace. Books still under copyright, however, cannot be read or searched in their entirety, which causes endless frustrations for scholars looking for material sequestered from their sight. This becomes particularly distressing when university libraries discard their printed volumes or fail to buy them to begin with, relying on 'someone else' to supply needed materials to their researchers. Thus, the ancient 'problem of the commons' has surfaced in a new form peculiar to academic libraries, and especially as it pertains to reference materials.

Yet aside from technical changes of speed and access introduced through the Internet, other alterations of reference materials occurred as libraries worldwide cut their labor costs by firing or 'repurposing' reference librarians *en masse*. These trained professionals, often women and stereotypically maternal and catholic in their approach to helping all library patrons, had been the foundation of public and university libraries for at least 125 years, ever since communities and campuses collectively decided that any town worth its name must have a library staffed by helpful, literate librarians. That career as such is almost gone. In its place, at even the finest university libraries, one discovers that 'the reference desk' is either absent entirely in lieu of self-service computers by the dozen, or is staffed by a transient young person without special knowledge who simply 'googles' any question put to them. So, instead of a mature librarian with proper educational credentials finding answers to questions by means of large, expensive, and trusted reference volumes, well thumbed and well known to their users, a young, inexpensive (even unpaid) amateur refers to computerized reference sources just as any citizen might who has access to their own computer.

The Shift toward Quanity

Abundance of Reference Materials

Moreover, and perhaps just as important over the last dozen years, academics and their publishers (most noticeably in the United Kingdom) have fallen under increasing pressure to 'produce' published works on an unforgiving timetable. The conventional 'factory speed-up' in manufacturing has spread to the 'academic factory.' This form of 'labor discipline' as applied to professors originated during the regime of Margaret Thatcher, who disliked the pampered 'chattering classes,' and wished to put them on a work schedule not unlike those applied to ordinary white-collar laborers in other 'industries.' British publishers responded to the resulting scholars' need to write ever more, and ever more swiftly, by issuing 'print on demand' monographs at extraordinary prices, hoping to sell most of them to libraries in North America, a far larger market than the United Kingdom's. Along with this manufactured pressure for scholars to publish rapidly, and for publishers to meet their sales goals in an ever-shrinking market for scholarly monographs, a number of well-known British publishers (followed hesitantly by their American counterparts) began to

issue dozens of handbooks, companions, encyclopedias, and dictionaries of social science material. This concatenation of forces produced an avalanche of reference materials the like of which had never been seen in libraries prior to about 2000.

It would require dozens of pages even to list all the handbooks, companions, dictionaries, encyclopedias, and guides, which have been 'cranked out' since 2000. When offered in printed form, their high prices and physical heft have consigned them almost exclusively to libraries rather than individuals. Few are offered in affordable paperback format. No one knows how often they are now consulted, given that the students and scholars educated over the last 20 years or so have become almost entirely reliant upon online sources of ready information. In today's campus phraseology, the notion of 'trudging' to the 'brick and mortar' library, 'hauling down' a 'fat tome,' which claimed when published to offer authoritative overviews of a given topic, is as alien to today's students as learning Latin or Greek. Since many of these reference volumes to not circulate, the articles of interest must be photocopied, another ancient technology happily superseded by ubiquitous printers connected to billions of personal computers all over the planet. Since there are no longer any clearly superior or inclusive reference sources in the social sciences as there once were (e.g., *The Encyclopedia of the Social Sciences*, 1933; *The International Encyclopedia of the Social Sciences*, 1968), it becomes difficult to identify those of particular merit for several reasons. The standard production format for today's reference volumes usually prohibits long-term investment of time or money, since revisions of those that sell well must occur rapidly. Editors at publishing companies which specialize in reference volumes typically seek academic advice when identifying a marketable topic. They then select a fairly senior or visible specialist in that area, pay them a very modest amount when gauged by the time they will likely invest in the project, and then leave it up to the volume editor(s) to find subeditors or authors for particular areas that need to be covered.

Influence on Authorship

Authors of the actual articles, which usually vary from 500 to 1500 words in a dictionary-like volume, all the way to 10 000 words or more for chapters in a handbook, are often paid little or nothing, it being assumed that they (being often younger scholars) need to publish rapidly, so will eagerly agree to work for free merely to get into print. This means, of course, that the most qualified scholars, those who have been working the longest on a given topic – unless, of course, it is brand new, or seems to be – are not as often found in such handbooks or guides as one might wish they were. The obvious trade-off lies between rapidly written pieces by junior scholars and researchers versus the work of seasoned experts whose presence in such volumes would require more time and money than the publisher is willing to invest. One such volume before me now, published in 2012 and aimed at the U.S. market, features two editors – Associate Professor (UK) and Professor (UK) – and authors from the following ranks: Postdoctoral Fellow (Australia), Assistant Professor (US), Postdoctoral Fellow (UK), Assistant Professor (US), Associate Professor (UK), Professor (UK), Senior Lecturer (UK), Lecturer (UK), doctoral student (Sri Lanka), private scholar (UK), and so on. Most of these writers are young, and wonderful as their work may well be, they are not recognized authorities on the topic in question. The standard response of reviewers when and if they evaluate such volumes is to say that 'the chapters/entries are uneven in quality,' as one would expect. And because such publications, rightly or wrongly, are regarded by administrators and other professional arbiters as not being peer-reviewed, they carry much less scholarly or reputational weight than more conventional publications. All of this notwithstanding, the flood of such books (in printed and electronic forms) shows no signs of retreating.

It may be worthwhile to mention a few of the many hundred recently published reference sources, if for no other reason than to indicate the vast topical range they now cover. This broad landscape of coverage is as much due it seems to the business needs of publishers as to any actual or perceived need on the part of working researchers. Unlike with expensive monographs or other special volumes, library acquisitions of such volumes are seldom the result of faculty input; in fact, few faculties are aware what reference volumes their libraries acquires since their presumed market is the student body rather than professors. This means that libraries purchase such works either under 'blanket order plans' or randomly, as one or another subject bibliographer so chooses. The purchase prices are so high and the production costs so low that publishers can profit from relatively few sales, particularly given print-on-demand technology.

Main Publishers of Reference Works for the Social Sciences

The main publishers today of reference works in the social sciences (and other areas of learning) are Ashgate (Surrey, UK and Vermont), Cambridge University Press (Cambridge, UK and New York City), Elsevier (Amsterdam and Oxford), Oxford University Press (Oxford, UK and New York City), Routledge (London and New York City), Sage (London) and Sage (California), and Wiley-Blackwell (Oxford, UK and Massachusetts). Among these several and more, hundreds of volumes have appeared in the last half-dozen years.

For example, Sage has recently published 'handbooks' on the following topics: qualitative research (2011), grounded theory (2010), international higher education (2012), mix methods in social and behavioral research (2010), social psychology (2007), interview research (2012), action research (2007), social network analysis (2011), qualitative research (2005), organizational research methods (2009), persuasion (2012), intercultural competence (2009), quantitative methodology for the social sciences (2004), leadership (2011), organizational institutionalism (2008), social cognition (2012), interpersonal communication (2011), innovation in social research methods (2011), qualitative methods in health research (2010), punishment and society (2012), political communication (2012), sociology of religion (2007), and organization studies (2006), to name only a few from their complete list.

Oxford and Cambridge have been creating handbooks and 'companions' for decades, long before Ashgate or Sage existed as reference publishers. Their imprimatur continues to lend to

their books a scholarly legitimacy, which may have outlived its actuality, given the ferocious competition among publishers for this market.

Oxford's famous handbook series, which varies somewhat in tone from Sage's, includes treatments of political institutions (2008), comparative politics (2009), political economy (2008), international relations (2010), causation (2012), positive psychology (2011), law and politics (2010), public policy (2008), political methodology (2010), political psychology (2003), Internet studies (2013), leadership (2012), intergroup conflict (2012), sociology and organization studies (2010), corporate social responsibility (2009), stress, health, and coping (2010), organization theory (2005), history of consumption (2012), organizational psychology (2012; 2 volumes), gender and politics (2013), urban politics (2012), traumatic stress disorders (2012), and American bureaucracy (2012), again, the tip of a large iceberg.

Cambridge University Press reference volumes have become indispensable in certain areas of research. The 'companions' series has become virtually a small world unto themselves, e.g., on science fiction, Kant, American novelists, atheism, utopian literature, gothic fiction, Jesus, Bob Dylan, Wittgenstein, and so on. Most are more firmly anchored in the humanities than the social sciences, although the companions to Tocqueville, Marx, Durkheim, Weber, Foucault, or Levi-Strauss each speak unambiguously to sociological interests. Cambridge also publishes handbooks, for instance, on 'the study of mental health' (2009).

Less august British companies have also leapt into the reference book trade with some success. Ashgate has invested heavily in its 'research companion' series, covering avant-garde topics like critical geopolitics (2013), war (2012), political violence (2012), queer theory 2009), globalization of health (2012), heritage and identity (2008), modern warfare (2010), multiculturalism (2010), cosmopolitanism (2011), democratization in Europe (2013), and biosocial theories of crime (2011), a few among many others. It has obviously worked hard to distance itself from the output of its rivals, whose participation in this market has been much longer.

The same can be said for Routledge, although an older company overall, has only recently begun to generate large and costly reference works whereas in its past, serious monographs were its forte. The market for such books having shrunk very considerably since neither scholars nor libraries are able any longer to spend large sums on single volumes, Routledge like so many of its peers has shifted to reference works. Its 'companion' series, imitating Cambridge, covers postmodernism (2011), critical theory (2006), religion (2009), philosophy of psychology (2011), social and political philosophy (2012), digital consumption (2013), identity and consumption (2012), social theory (2009), postcolonial studies (2007), feminism and postfeminism (2002), organizational change (2011), bioethics (2013), multicultural education (2009), postmodernism (2007), decolonization (2000), video game studies (2013), and emotional and behavioral difficulties (2012), to list a few from a large set. Routledge also offers 'handbooks,' e.g., on contemporary social and political theory (2011), identity studies (2011), cosmopolitan studies (2012), cultural sociology (2010), and climate change and society (2010), among others.

Imitation among publishers is obviously a factor in the production of these volumes as one can see from multiple entries under the most pressing current concerns: cosmopolitanism, social media, climate change, postcolonialism, and so on. It seems that eventually all reference materials will become electronic, and the production of long, heavy, expensive printed volumes created by many hands will cease. Even large academic libraries have begun to resist acquiring them as a matter of course, opting instead for a subscription to electronic versions, forgoing the purchase of its printed companion. Some libraries in computer-linked systems are sharing costly reference volumes through interlibrary loan, a practice which for decades was inconceivable since such materials by definition did not circulate. All this, like so much else in the world of scholarly publishing, is changing too rapidly for even seasoned researchers to absorb. Audiophiles are returning in significant numbers to vinyl reproductions of sound, and one wonders if a 'return to the printed book' might not also be on the horizon after the novelty of digitized knowledge sources wears thin.

Reference Works by Discipline

Sociology

Although in the space allotted here it is impossible to offer even a succinct history of social science reference books, some sense of how they evolved over the last century or so might be conveyed simply by means of mentioning some titles which seemed to define a given discipline at a particular moment in its development. I will begin with my own field whose history is more familiar to me than the other disciplines.

Despite its senior status as a university discipline of independent identity relative to most of the other social sciences, reference works have not been as numerous for sociology as one might therefore expect. Thomas Nixon Carver's *Sociology and Social Progress: A Handbook for Students of Sociology* (Sumner, 1907: 810 pp.) was remarkable for its earliness. Yet the first real encyclopedic effort in English came only in 1992 under Edward Borgatta's editorship (Macmillan: 4 volumes), a full century after academic departments bearing the discipline's name appeared in several midwestern universities. (Two previous efforts, one edited by the noted scholar Michael Mann, did not live up to their titles as 'encyclopedias' proper.) Other useful compendia include Smelser and Swedberg's *The Handbook of Economic Sociology* (1994) and Frank Magill's *International Encyclopedia of Sociology* (1995: 2 volumes).

The first handbook expressly designed for sociology was Alfred Vierkandt's *Handwörterbuch der Soziologie* (1931), a mere 690 pages long, followed in 1956 by Werner Ziegenfuss' *Handbuch der Soziologie,* at double the length. And in 1962 Rene König edited the famous *Handbuch der Empirischen Sozialforschung* in 2 volumes of 2000 pages. In the United States one could begin with *An Introduction to Sociology* edited by Jerome Davis and H.E. Barnes (1927, 1931), a 900-page collabora by seven authors. *Trends in American Sociology* followed in 1929, edited by George Lundberg, Read Bain, and Nels Anderson, with substantial chapters on the history of sociology, theory, social psychology, culture, rural sociology, urban sociology, educational sociology, social work, applied sociology, and

methods. In 1937 Prentice-Hall sold *Man and Society* for $3.75, a 787-page work edited by Emerson Schmidt that tried to introduce students to all the social sciences simultaneously, covering everything from human values to income distribution to social psychology (Herbert Blumer's memorable chapter). The Gurvitch-Moore volume, *Twentieth Century Sociology*, published just following the war (1945), seemed to have widespread influence, despite its execrable typos and layout. Its most distinctive feature, in addition to chapters on all the subfields of sociology at the time, was a second section of 250 pages that dealt with sociology in France, the United States, Britain, Germany, Latin America, Italy, Spain, Russia, Poland, Czechoslovakia, Rumania, and Yugoslavia. This should have inspired later handbooks, and may account for the volume's popularity, since a distinctly ecumenical flavor swept the academy worldwide following World War II. Its authors – Burgess, Parsons, H.P. Becker, Sorokin, Maclver, Znaniecki, Gurvitch, Roscoe Pound, Merton, Wach, Levi-Strauss, Faris, Albert Salomon, Bastide, and so on – comprise a list of distinguished scholars that has not been matched since in a collaborative volume. The first recognizably 'modern' compendium was Joseph Gittler's *Review of Sociology* (1957), with chapters by recognized authors on the topics that soon became the staples of introductory textbooks. (Georges Gurvitch edited *Traité de sociologie* in 1960, 2 volumes and 1000 pages, which seemed to have little impact in the United States.) And finally, Parsons himself had a go with his *American Sociology* (1968), with chapters by many scholars now in the twilight of their careers but who then were the bedrock of the mainstream, as it were.

One could say many things along these lines about the Smelser *Handbook*, which, given its size and unique position within sociology at present, deserves more than passing attention. This is especially so regarding the 'theories' that are put forth in the opening chapters as constituting 'the' new streams. For instance, neither Walter Wallace nor Jeffrey Alexander, the chapters' authors, wishes to consider postmodernism as a fact of existential and scholarly life, and therefore each throws himself or herself into a strange time warp. Peterson had characterized Faris' introduction to his earlier compendium as 'blindly benign' (Peterson, 1965: p. 254), which seems perfectly correct, even when rereading Faris today. One could say something similar about the theoretical terrain that is gingerly walked over in opening chapters of the Smelser *Handbook*. A range of old beliefs, now diminished by criticism from philosophers and literary critics – some of whom, incidentally, number among our most influential 'social' theorists – are trotted out as if nothing substantial had occurred to Western culture since 1964. It is as if the 'self' about which Parsons wrote with such conviction is still out there doing what selves have 'always' done. The same could be said about the precritical notion of 'science' and 'rationality' that figures in both chapters, but especially in Alexander's, as if definitions of rational action and societal rationality are not now in themselves enormously contentious topics. Such limitations begin to indicate the intellectual boundaries of compendia of this sort, even those written by top scholars, for the adventurousness that might pervade a monograph composed by the same authors is diminished when they turn their hands to 'general' statements. Since reference sources, by common consent, are supposed to offer a summary of 'established knowledge' in a specific field or subfield, their authors typically draw back from the perimeters of discovery and innovation. Therefore, such works almost always lag behind the leading edges of thinking and offer their users a snapshot of what was confidently held to be true a dozen or more years before they are published. And with the ever-increasing rate of change in global intellectual culture, this poses a problem, particularly for printed reference sources, which may be met only via constantly updated electronic versions.

Comparison of Smelser's *Handbook*, the most recent entry into the genre, with earlier versions must remain fairly rough. Its predecessors employed other sorts of authors, aimed at different audiences (those not yet effectively lobotomized by videocretinization), and seem to have been composed in an atmosphere of genuine wonder and hopeful discovery regarding what social scientific study might yield for intellectuals and for the social order that supported them. This kind of childlike hope decomposed after the Gurvitch–Moore volume in 1945 and was replaced by a somewhat defensive, stuffy scientism in later compendia that reflected sociology's struggle to survive McCarthyism on one hand, literary criticism on the other. But such books, while high on scientific rhetoric, seemed correspondingly low on political self-examination and savvy, while also revealing a foreshortened understanding of the philosophical problems necessarily part of sociological study. Still, it is uplifting to read these older books because their authors were usually straightforward and unpretentious. They tended to avoid enormous claims for the discipline, and offered up its fruits with a direct and unembattled character that nowadays is harder to find. It is probably no accident that after years of sociology bashing at the highest levels of government, both here and in Europe, the field should have put out a summary package for broad consumption in hopes of regaining lost ground, both within the cloister and beyond its high walls. We shall see if it succeeds as part of the long process of recertifying the discipline for a public which has been told repeatedly that nothing good can come of thinking about social reality in the way 'we' do.

Dictionaries for sociology began with Constantine Panunzio's pamphlet, *A Student's Dictionary of Sociological Terms* (California, 1937), continuing with Henry Fairchild's *Dictionary of Sociology* (1944). Only in the late 1960s were a number of competing volumes added to this small shelf of sociological reference books, including those of Geoffrey Mitchell (1968, 1979), George and Achilles Theodorson (1969), and Thomas Hoult (1969). With Raymond Boudon and François Bourricaud's *Critical Dictionary of Sociology* (1989), celebrated scholars from Europe entered the field for the first time, followed in the 1990s with new dictionaries by David Jary (1991), Nicholas Abercrombie (1994), Gordon Marshall (1994, 1998), and Allan Johnson (1995), the former three being well-known British scholars. American sociologists have not of late contributed substantially to this genre.

Anthropology

Of all the social sciences, anthropology has been to date the least well served in terms of general reference books. This seems odd given the nineteenth century's penchant for vast

compilations of quasi-anthropological data, e.g., Frazer's *Golden Bough* or Herbert Spencer's *Descriptive Sociology* (in fact comparative anthropology) in 12 volumes. The earliest dictionary of sociology that I have found has now become a standard, Charles Winick's *Dictionary of Anthropology* (1956), with several later editions. *Anthropology A to Z* (1963), edited by Gerhard Heberer, Carleton Coon et al., was based on an earlier German compilation of several hands. A recent trend among publishers, to name reference books after their firm, is typified in *The Macmillan Dictionary of Anthropology* (1986) edited by Seymour-Smith. The implication of this general practice – that a source's value to the user is more or less guaranteed by virtue of a firm's prestige – is not necessarily realized in practice. The most recent entry is Barfield's *Dictionary of Anthropology* (1997).

The discipline came into its own so far as references sources go when Honigmann et al. produced the 1300-page *Handbook of Social and Cultural Anthropology* (1973) at a period when the social sciences reached a zenith of appeal among college undergraduates in the U.S. Two misnamed volumes are Hunter and Whitten's *Encyclopedia of Anthropology* (1976) and Barnard and Spencer's *Encyclopedia of Social and Cultural Anthropology* (1996), both books of modest size. An interesting twist appeared with Bock's *Handbook of Psychological Anthropology* (1994), an instance, ever more common, of a reference book seemingly anticipating an intellectual or disciplinary trend before it gains acceptance at the institutional level. Such a rubric might well surprise both psychologists and anthropologists, relatively few of whom typically relate their two fields in any intimate way.

Cognitive Sciences and Related Fields

As soon as it became apparent in the 1970s to policy makers, governmental budget specialists, and academics that improvements in psychopathic or sociopathic conditions could be more cheaply and quickly realized through the use of drugs and related somatic measures – as opposed to 'talking cures' or other socially based interventions – those fields that stress the linkage between individual physiology and the social order experienced a sudden renaissance. What seemed a neo-Darwinian rage for simplification overtook large zones of inquiry within the social and behavioral sciences. This has become so much the case during the last decade that physiological reductionism often seems the order of the day, nearly extinguishing the behavioral component in explanations for an ever-broadening range of ailments, from alcoholism to schizophrenia to general criminality. Given this intellectual environment, it is hardly surprising that a range of new reference books appeared which positioned themselves along a continuum of biobehavioral explanation of social action, from crude sociobiological reductionism to subtle neurological analysis of brain function and behavior. Much of this new wave of work stirs up controversy because of its real or imagined political consequences, most famously exemplified by the raucous reception given Herrnstein and Murray's *The Bell Curve* (1994), both in the ordinary as well as the academic press.

Among many such books, which have recently been published, several might be mentioned for their innovative qualities. Michael Gazzaniga edited *Handbook of Cognitive Neuroscience* (1984), complemented by Dunlop and Fetzer's *Glossary of Cognitive Science* (1993), brought to even grander fruition in Wilson and Keil's *The MIT Encyclopedia of the Cognitive Sciences* (1999). Not unrelatedly are Wyer and Srull's *Handbook of Social Cognition* (1984, 1994) and Van Hasselt and Hersen's *Handbook of Social Development* (1992). More traditional, and connected with a field the roots of which go back at least to Saussure's posthumous masterpiece, *Course in General Linguistics* (1916), is Florian Coulmas's *The Handbook of Sociolinguistics* (1997), and what amounts conceptually to its complementary volume, Morton Gernsbacher's *Handbook of Psycholinguistics* (1994). These interestingly hybrid fields will undoubtedly grow in the new century, and it seems now that a biosocial focus will stimulate at least as much theorizing and research as did the more purely sociopolitical analyses during the preceding half-century.

Economics

Unlike some of the newcomers mentioned above, economics (or 'political-economy' as it was first known) has a long history relative to other social sciences, dating at least to 1776 and the publication of Smith's *Wealth of Nations*. Reference sources therefore abound. Before economics became an academic area, many guides to financial matters existed, of course, e.g., *The Handbook of Trade and Commerce, or, A concise dictionary of the terms and principles of trade, commerce, manufactures, commercial and common law, etc., with tables of money, weights, and measures*, published in London in 1840. We also know from sources like Marx's *Theories of Surplus Value*, his 3-volume anthology of annotated excerpts from the history of economic theory, that the literature on economic matters was by the 1850s quite well developed. Reference works, however, did not apparently surface until somewhat later. Palgrave's *Dictionary of Political Economy* (1894) is perhaps the most famous, and John Joseph Lalor's *Cyclopedia of Political Science, Political Economy, and the Political History of the United States: By the Best American and European Writers* (1881–84) is another early example. So was Frank Bower's *A Dictionary of Economic Terms for the Use of Newspaper Readers and Students* (1905), kept in print until 1951 under subsequent editors.

More obviously pertinent to the discipline as it now exists was Herwith Price's *Economic Dictionary*, 2 volumes (1926–29), published in Berlin, as well as John Todd's *The Mechanism of Exchange: A Handbook of Economics* (3rd edition, 1927) (which became *The Science of Prices* in a 4th edition (935) before reverting to its original title in its 5th). Frank Bower's brief *Dictionary of Economic Terms* reached its 10th edition in 1936, while Horton's *Dictionary of Modern Economics* appeared in 1948. Beginning in the early 1950s, the number and variety of dictionaries and handbooks concerning economics exploded, keeping pace with undergraduate enrollments at large public universities in the U.S. and elsewhere. Sloan and Zurcher's went through five editions between 1951 and 1970, while J.L. Hanson's reached its 6th edition in 1986, having been in print for 21 years. The latest dictionaries of economics include those by John Black (Oxford, 1997) and Bannock et al. (Wiley, 1998). On a larger scale, encyclopedias of economics began to appear relatively late, with the Guilford *Encyclopedia of Economics* in 1973/74, the *McGraw-Hill Encyclopedia of Economics* in 1982 and 1994 (2nd edition), and *The Fortune*

Encyclopedia of Economics in 1993. O'Hara has reverted in his choice of title to an earlier time, perhaps for political as much as intellectual reasons, with *Encyclopedia of Political Economy* in 2 volumes (1999). Another new entry is K. Liou's *Handbook of Economic Development* (1998), which broadens the discipline's terrain by including the global dimension.

Political Science

One of the earliest general reference works in this area is *Political Dictionary: Forming a work of universal reference, both constitutional and legal; and embracing the terms of civil administration, of political economy and social relations, and of all the more important statistical departments of finance and commerce* (London, 1845–46) in 2 volumes. It is instructive that those fields, which in the mid-nineteenth century seemed closely enough linked, empirically and practically, to be treated within a single work, have since the mid-twentieth century been torn asunder into mutually incompatible and incomprehensible camps of specialists, who have little to say to one another except at ceremonial occasions. This situation is not necessarily an advance in social knowledge. The 14-volume *American Annual Cyclopaedia and Register of Important Events: Embracing Political, Civil, Military, and Social Affairs* (New York, 1862–75) does not take in political science per se, though would surely have uses to the historian of the discipline. Already mentioned, John Lalor's 1882 *Cyclopedia of Political Science* […] is among the first to use the term in the modern sense. Another amalgamated volume that typifies an era unafraid to mix disciplines was William Bliss's *Encyclopedia of Social Reform: Including political economy, political science, sociology and statistics* (1897), some 1439 pages in length.

Dictionaries and lexicons of political science are as numerous as encyclopedias or handbooks are scarce. William Connolly's *Terms of Political Discourse* (1974) and Maurice Cranston's *Glossary of Political Terms* (1966) were produced by scholars who subsequently became well known for other work. Joseph Dunner's *Dictionary of Political Science* (1964), Jack Plano's *Political Science Dictionary* (1973), and Roger Scruton's *A Dictionary of Political Thought* (1982) are several among a score of similar works, which, as in neighboring fields, have been increasing in number, length, and complexity during the last decade. Works of larger scope, in terms of length as well as conceptual breadth, include David Miller's *The Blackwell Encyclopedia of Political Thought* (1987, 1991), Hawkesworth and Kogan's *Encyclopedia of Government and Politics* (1992: 2 volumes), and Goodin and Klingemann's *A New Handbook of Political Science* (1996).

Psychology

Just as the number of undergraduate majors in psychology typically overwhelms those of all the other social sciences, with ratios often running five or six to one or higher, so with reference books attached to this enormous field. The most distinguished early survey was surely William James's *Principles of Psychology* (1890: 2 volumes), in print continuously ever since, in full and abridged versions. Clark Murray's *Handbook of Psychology* (1885, 1888), James Sully's *Teacher's Handbook of Psychology* (1886, 1909: 5th edition), and the *Handbook of Psychology* (1889–91: 2 volumes) by William James's equally famous contemporary, James Mark Baldwin, were each substantial compilations, and helped give psychology a head-start toward producing a unified view of the social world vis a vis its sister disciplines. More eclectic viewpoints were expressed in Geoffrey Rhodes's *The Mind at Work: A Handbook of Applied Psychology* (1914) (with help from the noted sociologist, L.L. Bernard) and in Leander Keyser's *A Handbook of Christian Psychology* (1928: 2nd edition). By 1931, however, the field had become thoroughly professionalized, as exemplified by Anderson, Buhler, and A. Freud et al., *A Handbook of Child Psychology* (1933: 2nd edition), and its companion volumes in the Clark University series, H. Banister et al., *A Handbook of General Experimental Psychology* (1934: 1125 pp.) as well as Allee, Allport, et al., *A Handbook of Social Psychology* (1935: 1195 pp.). Somewhat less innovative were W.B. Pillsbury's *Handbook of General Psychology* (1942), Kimball Young's *Handbook of Social Psychology* (1946), or Lindner and Selger's *Handbook of Correctional Psychology* (1947). Specialization within the discipline had already reached such a pitch that in 1950 Norman Munn offered *Handbook of Psychological Research on the Rat* (first published in 1933 as *A Introduction to Animal Psychology*).

The 1950s were a boom period for psychology reference works of the larger scale, reflecting the discipline's ever-widening influence in Western societies, with at least six more 'handbooks' entering the market. In more recent times innovation has once again become commonplace, with books like M. Gazzaniga's *Handbook of Psychobiology* (1975), Raaig, Veldhoven, et al., *Handbook of Economic Psychology* (1988), and Crawford and Krebs's *Sociobiology and Psychology* (1987), which recently became the *Handbook of Evolutionary Psychology* (1998). The last volume intimates a strong connection currently being forged between researchers with refreshed enthusiasm for neoevolutionary doctrines on the one hand, and sociobiological explanations for behavior on the other. This general tendency – in some ways a repudiation of earlier causal models for social action and psychological states – has been given substantial impetus by the genome project and related studies that attempt to link physiology with behavior in an unmediated fashion.

Encyclopedias have also been a more regular feature of psychology's disciplinary life than in other fields. Early efforts included McKean's *The Encyclopedia of Psychology: A series of graphic descriptions of human characteristics* (Brussels, 1928) and Philip Harriman et al., *Encyclopedia of Psychology* (1946). The 1970s exploded with new versions varying substantially in size and scope. Robert Goldenson's *Encyclopedia of Human Behavior: Psychology, Psychiatry, and Mental Health* (1970: 2 volumes) brought together supplementary fields, just preceding the herculean effort by Benjamin Wolman et al. (300 editors and 1500 authors), *International Encyclopedia of Psychiatry, Psychology, Psychoanalysis and Neurology* took up 12 initial volumes (1977) plus a supplement (1983) and surely set a new standard for comprehensiveness of these allied research areas. Wolman also edited a 1-volume *Encyclopedia of Psychiatry, Psychology, and Psychoanalysis* in 1996. Another half-dozen encyclopedias of psychology proper appeared during the late 1970s and 1980s that of Raymond Corsini's running to 4 volumes (Wiley, 1984) and Frank Magill's reference book production team brought out the *Survey of Social Science: Psychology Series* (1993) in

6 volumes, grandly renamed the *International Encyclopedia of Psychology* when reissued and reset in 1996. (Magill and his specialty editors have produced like volumes for all the social sciences in his *Survey* series.)

Ever since James Mark Baldwin published *Dictionary of Philosophyy and Psychology* (1901–05: 3 volumes; new edition, 1928), there has never been a shortage of dictionaries for the discipline. I have identified 35 in English alone, some of them of continuing interest simply for the peculiarity of their design, at least by today's standards, e.g., William Taylor's *Taylor's Dictionary of Bio-Psychology, Genetics, and Various Branches of Applied Psychology* (1925: 148 pp.), or Vergilius Ferm's *A Dictionary of Pastoral Psychology* (1955). Of the more standard fare – a score or more appear in most standard bibliographies – many were compiled by the same scholars who oversaw production of the encyclopedias named above, with the most lavish entries all appearing during the 1990s.

Conclusion

Were Moreri and his contemporary compilers alive today, they would probably be astonished at the level of continuing interest in reference sources. Informed academic reference librarians and professional bibliographers now claim that we are living in a golden era of reference materials, especially if one uses sheer quantity of output as the vital measure of the gold standard. One director of acquisitions wrote of "So Many Choices, So Little Money" regarding new reference books (Wilkinson, 1997), and the editor of *The Book Review* of *The Library Journal* recently analyzed 'reference for a new age' with particular attention to the mesh of electronic and printed materials suddenly dominating publishers' catalogs (Hoffert, 1996). A third expert notes that flashy graphics and other user-friendly alterations have attracted many first-time users to reference materials who before would have shunned their perceived complexity (Mantell, 1997). Even a cursory look at the current availability of dictionaries, handbooks, or encyclopedias in the social sciences indicates that in some fields, for which before 1990 virtually no reference books (let alone CD-ROMs) were available, now have at their disposal a half-dozen fresh entries in the genre. And even at considerable cost, they are being bought by research libraries almost as a matter of course, just as the electronic forms have begun to encroach on printed versions. The latter may continue to exist into the upcoming century, but will probably be relegated to the status of secondary resources, subordinate to the newer formats. Or they might persist as specialized bodies of information for ever-smaller audiences. Overall, it is fair to say that the reference materials 'business' is thriving as never before, especially when measured by library acquisition budgets aimed in its direction, and the ultimate effect this outpouring may have on the way social science is conceived and used remains difficult to predict.

That said, however, no one seems to be arguing that the latest entries into the reference book market are supplanting in quality of information or breadth of coverage some of the old standards. *The Encyclopedia of the Social Sciences* from the early 1930s will never be rendered entirely obsolete simply because so many of its contributors have themselves become the subjects of articles in later works. What was true in 1910 of the 11th edition of the *Encyclopaedia Britannica*, still the nonpareil standard of reference works in English, is similarly the case for a range of earlier works. When authors close to the origin of a given field of research accepted invitations by reference publishers to write about their own scholarly innovations (e.g., Werner Sombart on 'Capitalism' in the *ESS*, Benedetto Croce's article on 'Aesthetics' or Edmund Husserl's on 'Phenomenology,' both in the 14th edition of the *Britannica*), later authors have been obliged to accept such statements as first-order expert testaments, beginning their reformulations from there. Surely some of this is going on today, for instance, in newer fields like 'cognitive neuroscience,' 'psycholinguistics,' 'psychobiology,' or 'economic psychology,' and a century hence these may well be accepted as foundational statements. (Certainly, all of these 'reformulations' are accorded extensive treatment in the current work!) But for the time being, it is hard to distinguish the ephemeral from the more or less permanent when evaluating the raft of reference works that have appeared during the last decade or so.

See also: E-Communications Platforms and E-learning; Libraries; Publishing as Medium; Publishing: Academic and University Presses; Social Science and Civil Society; Social Science and Universities.

Bibliography

Bliss, William (Ed.), 1897. Encyclopedia of Social Reform. Funk & Wagnalls, New York.
Boccardo, Gerolamo (Ed.), 1857. Dizionario della economia politica e del commercio. Franco e Figli, Torino.
Burnett, John, 1984. Useful Toil: Autobiographies of Working People from the 1820s to the 1920s. Penguin Books, New York.
Chinoy, Ely, 1966. Review of Faris' handbook of modern sociology. American Sociological Review 31, 270–271.
Conrad, et al. (Eds.), 1895. Handwörterbuch der Staatswissenschaften. Gustav Fischer, Jena.
Coquelin, Charles (Ed.), 1852/1853. Dictionaire de L'economie politique. Guillaumin, Paris.
Cupp, William L., 1959. Review of Merton et al., sociology today. Social Forces 38, 77–78.
Davis, Jerome, Barnes, Harry Elmer (Eds.), 1927. An Introduction to Sociology. D.C. Heath, Boston.
Elster, Ludwig (Ed.), 1898. Worterbuch der Volkswirtschaft. Gustav Fischer, Jena.
Faris, Robert E. Lee (Ed.), 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Furbank, Paul, 1992. Diderot: A Critical Biography. Alfred A. Knopf, New York.
Gittler, Joseph (Ed.), 1957. Review of Sociology: Analysis of a Decade. John Wiley & Sons, New York.
Gittler, Joseph (Ed.), 1965. Review of Faris' handbook of modern sociology. American Journal of Sociology 71, 335–336.
Gurvitch, Georges, 1960. Traite de sociologie. Presses Universitaires de France, Paris.
Gurvitch, Georges, Moore, Wilbert (Eds.), 1945. Twentieth Century Sociology. Philosophical Library, New York.
Hoffert, Barbara, 15 November 1996. Reinventing reference for a new age. Reference 97: Supplement to Library Journal, S4–S7.
House, Floyd, 1959. Review of Merton et al., sociology today. American Journal of Sociology 45, 211.
König, Rene (Ed.), 1962. Handbuch der Empirischen Sozialforschung. Ferdinand Enke, Stuttgart.
Lundberg, George, et al. (Eds.), 1929. Trends in American Sociology. Harper & Brothers, New York.

Mantell, Suzanne, September 29, 1997. Looking it up is looking up. Publishers Weekly 244, 51–53.
Merton, Robert K., et al. (Eds.), 1959. Sociology Today. Basic Books, New York.
Moore, Wilbert, 1959. The whole state of sociology. American Sociological Review 24, 715–718.
Palgrave, Robert H.I. (Ed.), 1894. Dictionary of Political Economy. Macmillan, London.
Palgrave, Robert H.I. (Ed.), 1900–1901. Dictionary of Political Economy. Macmillan, London.
Parsons, Talcott (Ed.), 1968. American Sociology. Basic Books, New York.
Peterson, Richard, 1965. Review of Faris' Handbook of modern sociology. Social Forces 44, 253–254.
Rotteck, Carl von, Welcker, Carl (Eds.), 1834–1843. Staats-Lexikon oder Encyclopedie der Staatswissenschaften. Ultona, Hammerich.
Say, Léon, Chailley, Joseph (Eds.), 1891/1892. Nouveau Dictionnaire D'economie Politique. Felix Alcan, Paris.
Schmidt, Emerson (Ed.), 1937. Man and Society. Tübingen: H. Laupp, New York, Prentice-Hall.
Sumner, W.G., 1907. Folkways: a Study of the Sociological Importance of Usages, Manners, Customs, Mores, and Morals. Ginn & Company, Boston.
Vierkandt, Alfred (Ed.), 1931. Handwörterbuch der Soziologie. Ferdinand Enke, Stuttgart.
Wilkinson, Frances E., September 1997. Reference materials: so many choices, so little money: librarians and publishers speak out! Against the Grain 9 (4), 1, 18, 20, 47, 167.
Ziegenfuss, Werner (Ed.), 1956. Handbuch der Soziologie. Ferdinand Enke, Stuttgart.

End of Life Choices

Frances Norwood, George Washington University, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

As advancements in medicine, pharmaceuticals, and technology at the end of life have proliferated, particularly in Western industrialized nations around the world, the range of choices for people at the end of life has not. Nature and culture strike a delicate balance at the end of life today. Choice is not simple, but instead follows a highly structured, cultural script. Ultimately, death is biological and in one sense of that word, inevitable. Importantly, however, death is also a cultural event, and 'choices' are those options culturally scripted and publically mediated that allow people to die within and outside of culturally appropriate ways of dying in places around the world. In the Foucauldian sense, choice is an illusion. But even that is not absolute, as individuals creatively push against and transform even the most sacred traditions around the end of life. This article is about end-of-life choices in contemporary societies, and predominantly Western industrialized countries, and the role of culture in shaping those choices.

It is ironic that as advancements in medicine, pharmaceuticals, and technology at the end of life have proliferated, particularly in Western industrialized countries around the world, choices for people at the end of life have not. Ultimately, death is biological and in one sense of that word, inevitable. Clearly, the body is an organism that is affected by natural processes that determine life, health, disease, decline, and death. Importantly, however, death is also a cultural event, and 'choices' are those options culturally scripted and publically mediated that allow people to die within and outside of socially appropriate ways of dying in places around the world. Nature and culture strike a delicate balance at the end of life today. Choice about the end of life is never a simple choice, but instead may follow a highly structured, cultural script about how to die.

According to philosopher Michel Foucault (1972, 1991), choice is somewhat of an illusion and discourses are scripts that emerge in societies – cultural forms that shape the way we think, feel, and act. Discourses are not simply spoken exchanges, but limit what can be spoken, what is constituted as taboo, what is held in collective conscience, what is reconstituted from the past, and who in society has access (Foucault, 1972: 115; Foucault, 1991: 58–60). Dominant discourses – like those of modern-day medicine – limit the ability we have to make independently derived decisions and choices in cultural practice. As Foucault states, "in every society the production of discourse is at once controlled, selected, organised and redistributed according to a certain number of procedures, whose role is to avert its power and its dangers, to cope with chance events, [and] to evade its ponderous, awesome materiality" (Foucault, 1972: 216). Add to that the limitations of biology, and the decline of the human body and 'end-of-life choices' become a phrase not well suited to describe the end of life today.

End-of-life discourses among scholars, policy makers, health care providers, families, and the media are highly mediated by concepts of nature and culture, where players at the end of life – physicians, patients, families, medical associations, law, the media, and powerful industries such as pharmaceutical, hospital, and insurance industries – invoke 'nature' and 'culture' as needed to further their oftentimes competing agendas. Yet, neither of these concept categories or the scripts behind them are absolute as individuals creatively push against and transform even the most sacred traditions around end of life. From euthanasia in the Netherlands to the overtreatment of patients at the end of life in the US to the inequalities of access in India, individuals, culture, and nature strike a delicate and complicated balance at the end of life. This article is not a comprehensive, legal history of end-of-life choices in places in all parts of the world, but rather focuses on the interplay between nature and culture at the end of life, highlighting the role of culture in shaping prevalent end-of-life choices in a few contemporary societies.

Nature and Culture in End-of-Life Choices

In social science, death is often at the debated intersection between nature and culture, where 'nature' is often represented by some combination of natural process, materials from God, or biological fact and is opposed to 'culture,' alternately defined as human manipulation of nature (a cultural intervention) or as what it is not (not pristine, not 'natural'). Scholars who have examined this dichotomy closely have suggested that both nature and culture are permeated by aspects of the other to such an extent that the opposition is no longer a useful tool (Nader, 1996; Strathern, 1992; Yanagisako and Delaney, 1995). Marilyn Strathern in her book *After Nature* (1992), for example, takes on postmodern notions of nature and culture, which suggest that both are cultural constructions where nature is consumed by a sea of pluralistic, fragmentary, and ever-emerging process of cultural construction. Nature vanishes in postmodern thought, replaced and subsumed by culture. Strathern argues that nature and culture are neither in competition (dualistic), nor consumed one by the other, but instead coexist one within the other. She writes,

> [c]ulture and nature may be connected together as each operates in a similar way according to laws of its own; at the same time, each is also connected to a whole other range of phenomena which differentiate them – the activities of human beings, for instance by contrast with the physical properties of the universe. This second connection makes the partial nature of the analogy obvious. It presupposes that one thing differs from another insofar as it belongs to or is part of something else.
>
> Strathern, 1992: 73

So what is 'after nature' is at the same time connected to it, but in modern-day practice nature has come to be trivialized, a product of the consumer's choice. Strathern writes, "[i]ts image may be borrowed here and there, slapped on to products as a new dimension, a kind of marketed contextualization that reinvents human responsibility as a matter of discriminating between products" (Strathern, 1992: 197).

In the debates around death and dying, these varying conceptions of nature and culture permeate – nature and culture as discrete dichotomous entities; nature as cultural construction; and nature and culture as derivatives of the other. Is death the place where culture ends and nature reasserts itself or is the biology of death and dying inextricably reliant on culture for meaning? The difficulty of pinpointing even the changing biology of death, as demonstrated by different conceptions and receptions of brain death and organ transplantation for instance, demonstrates the inextricable connections that nature and culture share (Hogle, 1999; Lock, 2002; Scheper-Hughes, 2004, 2005).

Nature, Culture, and Brain Death in Japan

It is often assumed that death is merely a biological or 'natural' event. Anthropologist Margaret Lock takes issue with that in her ethnography *Twice Dead* (2002), where she finds that concepts of brain death do not fit well with pervading cultural understandings of life and death in Japan. She points out that until the advent of life-prolonging technologies, such as cardiopulmonary respiration (CPR) and the artificial respirator, and advances in medicine around organ and tissue transplantation, did it become necessary to reassign death from failure of the heart and lungs to 'brain death.' She writes, "[i]t was essential that the new death be a diagnosable event and that it be timed to allow the removal of organs while they remained 'fresh' and reasonably well oxygenated. No longer based on the commonsense notion of the end of life – a failure of heart and lungs – the new death is determined by the condition of the brain alone, even though the body clearly remains biologically alive, albeit with mechanical assistance" (Lock, 2002: 78).

Reassignment of death from the heart and lungs to the brain was necessary in places around the world in order to facilitate organ and tissue removal and reuse. The new demarcation of death also served to avert legal complications and to address public fears. Lock writes, "[t]he very existence of ventilator-dependent patients raised two concerns: first, the fear of being counted dead before one's time and over hastily designated an organ donor; second, the fear of being kept alive too long, as a 'vegetable,' with severe, irreversible brain damage, but not technically brain-dead" (Lock, 2002: 78).

The rollout of the new death received different cultural reactions in places around the world. In Germany, for instance, the legacy of the Holocaust has manifested in a general reluctance toward organ procurement (Hogle, 1999) and in some nonindustrialized countries evidence is showing that the commodification of body parts is a burden not shared equally (Scheper-Hughes, 2004). In Japan, organ transplantation and brain death continue to be viewed as 'unnatural.' Cultural notions of tradition and modernity, Japan and the Western 'other,' religion and medicine, individual and society, and distrust of the medical establishment come into play in Japan. Lock writes, "[d]eath in Japan remains above all a social event; the family is in control, and death with dignity requires family participation… The concept of brain death is seen by many as a clumsy creation of technological medicine, devised to hasten the end of life [and] because the concept of 'person' is not usually lodged exclusively in the brain, locating death in the brain causes discomfort" (Lock, 2002: 369).

Denying Death: Cultural Constructions at the End of Life

Many social scientists have focused on the cultural aspects of death and dying, either arguing against the reification of biology and science or to delineate the powerful role of culture at the end of life. A common theme that has shaped how social scientists and cultural historians view the end of life is the focus on modernity, or industrialized cultures, as death denying. In *Western Attitudes toward Death* (1974), historian Philippe Ariès argues that attitudes toward death in Western societies (predominantly Europe and the US) have changed over time and across cultures. He identifies death in the Middle Ages as a 'tame death' in which the dying person assembles his or her family, friends, neighbors, doctors, and priests bedside, orchestrating the ritual and its protocol, and ceremoniously making his or her final exit. Death during this time was viewed as a fact of life and the types of end-of-life choices that people of this time had were largely to die quickly due to injury or disease and in the relative comfort of home and family. Death in the Middle Ages was inevitable. In Ariès words, "in death man encountered one of the great laws of the species, and he had no thought of escaping it or glorifying it" (Arìes, 1974: 28). By the twelfth century, this resignation to the 'collective destiny' of dying shifted to a more individualistic approach to death. The dying person became more concerned with his or her own death, viewing it as a type of judgment day. This shift toward individualism, he argues, has remained throughout the present time.

By the eighteenth century, death became a more dramatized event, and the focus shifted from the dying person to the grieving survivors who now saw death as a disruption of their own economic and social lives. During this time, the family was given more power over the event of death; mourning became much more expressive; and a 'cult of memory' was created to mark the event for those most affected by it: the survivors (Arìes, 1974: 66–67). Today, according to Ariès and a number of other researchers, modern death has undergone a transformation marked by a type of denial of death (Arìes, 1974; Becker, 1973; Giddens, 1991; Glaser and Strauss, 1968; Gorer, 1955; Walter, 1994). Ariès explains, "death, so omnipresent in the past that it was familiar, would be effaced, would disappear … it would become shameful and forbidden" (Arìes, 1974: 85). Death moved out of the home to be cleaned, sterilized, and sequestered in institutions of medicine, moved from the private, family-dominated space of home into hospitals where death could largely be erased under layers of disease-focused, body-compartmentalizing, and technology-driven medical intervention.

In *The Denial of Death* (1973), anthropologist Ernest Becker disagrees with the timing of denial, arguing that denial and fear of death is universal to human beings of all cultures and times. The world is a terrifying place, he says, marked by creation and destruction; life, violence, and the inevitability of death. Human beings naturally react with anxiety or fear, which gives rise to repression of that anxiety or denial of death. Denial of death is not enough to deal with the terror of life, however, so society creates a 'hero system,' designed to help us transcend death in this life by showing that we are part of something worthwhile (Becker, 1973: xiii).

Sociologist Anthony Giddens, in contrast, argues that denial of death is not something universal to all cultures across time,

but something new. He links denial to the onset of modernity (1991). Giddens argues that death, sickness, and other 'non-normal' types of human experience (e.g., criminality or mental illness) are hidden or sequestered from daily social life. The 'sequestration of experience' is a process intrinsic to modernity in which nonnormal experience is physically removed from ordinary social life, housed in institutions of control, such as prisons, asylums, or hospitals, and then socially reconstituted via a modern predisposition toward self-reflexivity and the resulting internalization of society's rules, values, and norms (Giddens, 1991: 1–34, 156–162). Sequestration allows a separation "from 'original' nature and from a variety of experiences bearing on existential questions and dilemmas" and from "direct contact with events and situations which link the individual lifespan to broad issues of morality and finitude" (Giddens, 1991: 8). In this sense, sickness, dying, and death in modernity are found to be outside of the social system, sequestered and hidden from view, resulting in existential disconnect. He writes, "personal meaninglessness – the feeling that life has nothing worthwhile to offer – becomes a fundamental psychic problem in circumstances of late modernity" (Giddens, 1991: 9).

A growing number of social scientists are challenging this view of modern society as death denying. Both Giddens and sociologist Clive Seale question Ariès' view of the 'tame' death as a contrast to modern-day denial, citing evidence from Elias which suggests that death in the Middle Ages was not as tame as Ariès shows it to be (Giddens, 1991: 161; Seale, 1998: 52–54). Seale also questions theories put forth by Becker and Ariès, which rely on the presence of fear as the explanation for denial of death. Seale argues that the evidence does not necessarily support the premise that fear was alleviated by religion prior to modern-day society or that medical science offers alleviation from fear today (Seale, 1998: 76).

In *Constructing Death* (1998), Seale suggests that those who view modern society as death denying are confounding psychological denial with sociological denial of death. He argues that at the psychological level modern-day societies are seldom death denying. He writes,

> [d]eath is in fact actively managed in modernity and thus can be understood as the result of a characteristically modern, full square and unflinching facing of death at the personal level. No doubt there are individuals who are helped to suppress their personal fears by the fact that dying people are less visible, but there are many others who engage in a rather practical acceptance of the realities of death. Not least, the host of 'experts' in health care who are, after all, also a part of 'society', are very often remarkably accepting of the facts of death, supported by the routinisation of care that modern bureaucratic organisations provide.
> Seale, 1998: 54

Death is 'hidden away' in modern, industrialized societies, but not because of a psychological denial of death, but because the dead have little power to disrupt social bonds in modern society. Seale explains,

> [t]he continuing social existence of the dead, intervening in the affairs of the living, as ghosts, spirits or ancestors, is a good indicator of the degree to which death is disruptive of social order. Ghosts can be understood, largely, as reifications of the unfinished business of people who die when still socially engaged. In general, the society of the dead can be understood as socially controlling the moral life of the living… [T]he dead, in modern societies, are relatively powerless.
> Seale, 1998: 52

Seale argues that the aged and the dying are sequestered from everyday life, devalued, segregated, and controlled, but while death in modern times may be suppressed from immediate social physicality and awareness, everyday life is at the same time permeated by social, symbolic, and embodied enactments of the 'renewal of life against death' (Seale, 1998: 70).

Anyone flipping through prime-time television with the proliferation of violent crime dramas and supernatural walking-dead entertainment might agree with Seale, questioning whether we truly are a collection of death-denying societies. Anthropologist Geoffrey Gorer (1955) makes a compelling argument that takes into account the proliferation of mediated death in modern-day societies. He suggests that as Western societies rejected some of the rituals around death, at the same time they became more open to sexuality, and ultimately, he argues, death replaced sex as the new taboo in many places. While violent, nonnatural deaths took on an almost pornographic tone in mediated forms, 'natural' or ordinary, everyday death from illness, disease, and old age and the rituals and mourning practices associated with those were denied (Gorer, 1955). Several authors since then have looked at the rise in graphic, mediated forms of death around the world (Bogard, 2008; Foltyn, 2008; Gibson, 2007; Hanusch, 2008), including sociologist William Bogard who suggests that our current preoccupation with unnatural, graphic, and violent death marks a shift from social control via the corpse to social control via modulation of hybrid life forms (Bogard, 2008: 191). The corpse, Bogard argues, once pivotal in Foucault's (1973) concept of biopower, has been replaced. New biopower relies on control over life and choices at the end of life as new global discourses aimed at "technologies for generating organs, growing hybrid bodies, and creating life *in vitro*" now dominate (Bogard, 2008: 188).

Medicalizing End-of-Life Choices: Hospital Death and Its Options

In the backdrop of these changing cultural constructions of death and dying, discourses of medical science have emerged to largely overtake the end of life, making death and end-of-life choices increasingly medical concerns (Porter, 2001; Starr, 1982). By the twentieth century, physicians and medicine replaced chaplains and religion at the bedside of the sick and the dying. Key advances in medicine, such as the advent of antisepsis techniques in surgery and the ability to maintain oxygen flow via the artificial ventilator, allowed many people to live longer lives, but that also meant that suffering was prolonged (Lock, 2002: 57–75; Lynn, 2004: 4–31). A hundred years ago, the average life expectancy was 47 years in the US and most people died at home from infections or accidents after experiencing at most hours or weeks of decline before death. Today, life expectancy and living with chronic illness

have both increased dramatically. In 2000, US life expectancy averaged around 75 years and people were most likely to die in the hospital with an average of 2 years of chronic conditions and decline (Lynn, 2004: 1–10). Death and dying, once the domain of the priest, family, and home has largely become medicalized, overtaken by the physician, the hospital, and medical interventions with powerful industries and interests – medical, hospital, insurance, and pharmaceutical – vying for the declining body and pervading end-of-life choices.

Research suggests that in many cultures, certainly not all, the preference is to die at home (Kaufman, 2005; McWhinney et al., 1995; Townsend et al., 1990). Yet, in most Western industrial nations many continue to die in the hospital. In the US in 2005, for example, approximately 45% of Americans died in hospitals, 22% in nursing homes, and only 25% died at home where most say they would prefer to be when they die (Brumley, 2002; CDC, 2005; Tang, 2003). In a recent study comparing death in six European countries, the majority of people died in hospitals ranging from 63% in Sweden and Wales, 59% in Scotland, 58% in England, and 52% in Belgium (Cohen et al., 2008). The Netherlands, known for its more progressive policies at the end of life and one of the lowest rates of hospital deaths, stands out from the rest reporting only 34% hospital deaths (Cohen et al., 2008). The Dutch Central Bureau for Statistics reports that in the Netherlands more than one-fourth of persons each year die in their own home and one-fifth die in residential or nursing home care (CBS, 2004).

Hospital-Mediated Dying: A Typical Case

Jack Carter is an 84-year-old patient with congestive heart failure taken to the intensive care unit (ICU) after a heart attack and by day 10 in the ICU he is not improving. His cardiologist has been treating him for the last 8 years and knows Mr Carter and his advance care wishes well. Mr Carter would take CPR and ventilator support if he has a reversible medical problem, but not if his condition could not improve. As anthropologist, Sharon Kaufman, writes, Mr Carter had "reached a crossroads on the heroic intervention pathway, and now he is stuck there."

An angiogram has been ordered to investigate the extent of the artery damage around the heart even though the contrast dye used could also cause kidney failure and the cardiologist saw it as a last ditch, and unlikely, attempt to help him live longer. Mr Carter was unsure, asking his long-time cardiologist, "what would you do?" to which his physician answered, "I'm not eighty-four. I don't know what I would do …" The nurse on duty also thought the angiogram was not the best choice, explaining that many patients like Mr Carter come in for a 'heart tune-up' and it usually takes several hospitalizations before they are ready to face death. Mr Carter is alert and animated and tells Kaufman that he has weighed the risks and is ready to go forward with the angiogram. Later, he tells the chaplain something different, fearing that he might die from the procedure. The day before the procedure, Mr Carter's blood pressure drops and soon after his cardiologist performs electric shock to stop an irregular heartbeat. It does not work and Mr Carter's heart stops beating entirely. Kaufman describes what follows: "[t]he medical team performs CPR in an attempt to resuscitate him. For nearly two hours they do chest compressions, apply electric shock, and give blood pressure medications intravenously, hoping to restore his heartbeat and blood pressure. 'He wanted aggressive treatment,' the cardiologist explains afterward." Mr Carter, 84 years of age, dies after 10 days in the ICU.

Kaufman argues that this case falls between what some would consider an inhumane death – dying after a prolonged ICU stay following the violence of CPR – and a death according to the patient's wishes – his wish to explore life-saving treatment until it was clear that sustaining life was no longer possible. She attributes Mr. Carter's problematic death to the heroic medical pathway that promotes aggressive treatment in the face of unlikely positive results and on the lack of clear guidance offered by the cardiologist. When the patient showed doubt about continuing on the heroic pathway, the doctor failed to answer his question, or discuss it further, instead putting the responsibility for deciding about death on to his patient who clearly did not have the information or time to process whether continued treatment was the best choice for him.

Sharon Kaufman, 2005. And a Time to Die: How American Hospitals Shape the End of Life. Scribner Publishers, New York, pp. 161–166.

Medical anthropologist Sharon Kaufman (2005) studied hospital death in the US, and her findings echo others who have examined hospital-dominated dying. She suggests that American medicine has attempted to stave off death with the most sophisticated and expensive treatments and technologies. Kaufman points out that as many as one-fourth of all hospitalized patients are treated in an ICU of a hospital before they die and when they enter the hospital system they enter a system that due to a complex set of circumstances – including incentives for overtreatment by specialists and the ever prevalent threat of legal action against health-care practitioners in the US – pushes life-saving treatments even when none of the participants (hospital staff, patients, nor families) view that as the best choice (Kaufman, 2005: 319). The result is a way of dying that medicalizes the end of life, alienates all involved, and results in overly aggressive interventions and treatment and, ultimately, a lack of real choices for people as they die. Studies from around the Western industrialized world demonstrate that even over the last decade, there has been a general failure of hospital medicine to change and aggressive and often futile treatments continue to pervade care for the dying even up to the time of death.

Advance Directives

Kaufman points out that one of the most prevalent 'choices' made at the end of life in the hospital is around code status, or the choice of whether or not to receive resuscitation or ventilation in the event of heart or lung failure. Advance directives, or living wills, are preferences that patients can put in writing to detail what medical actions should and should not be taken to maintain their life in the event of a medical emergency in which they would otherwise not be able to communicate their wishes. CPR is performed in the United States on anyone who does not have a written order, or advance directive, stating his or her code status. This order is called a Do Not Resuscitate order and is one of several advanced care directives that patients near the end of life can make prior to or during a hospitalization to attempt to influence the course of their treatment while in the hospital. Emergency CPR typically entails insertion of a breathing tube, chest compressions, and electronic shocks to the heart. On frail bodies near the end of life, this default procedure can be traumatic for all involved and can include breaking ribs, damaging windpipes, and postponing what could have been

a more peaceful death (Kaufman, 2005: 48–49). Ventilators are used to maintain lung functions when a patient cannot do so on his or her own. Both CPR and use of ventilators are two prevalent emergency interventions routinely used at the end of life that have great consequences on how people die, and thus are an important part of any advance directive.

Advance directives usually come in two forms: (1) various health instructions in the event of a medical emergency and (2) designation of a durable power of attorney or someone who can serve as a health proxy in the event that they are not able to advocate their end-of-life preferences for themselves. Precipitated by high-profile cases of persons left partially alive in persistent vegetative states around the world, advance directives emerged often as a patient-led response to overtreatment at the end of life and in the context of public discussions about what constitutes 'good' or 'bad' forms of dying (Brown, 2003; Emanuel and Emanuel, 1998; Hirai et al., 2006; Ishihara et al., 2008; Norwood, 2006b; Seale and Geest, 2004; Van der Geest, 2004).

In the US, for example, advance directives arose as a patient-led response to the case of Karen Quinlan in 1975 and Nancy Cruzan in 1983, both young women kept alive for many years in a persistent vegetative state. The patient's rights movement has been strong in the US and ultimately led to the Patient Self-Determination Act of 1990, which among other things recognizes the right of individuals to refuse or accept medical treatment (Brown, 2003; Norwood, 2009: 94–95). Many nations in Europe also have some variation of living wills or legislation regarding patient's rights to refuse treatment, including Austria, Belgium, Finland, Hungary, the Netherlands, Spain, Switzerland, and the United Kingdom. Bulgaria, Greece, Italy, Norway, Serbia, and Turkey do not regulate living wills and/or they are not legally binding in these countries (Brown, 2003). France, Germany, and Slovakia have policies that partially recognize either refusal of treatment or advance directives (Brauer et al., 2008).

Advance directives and how they are employed also vary quite a bit among subcultures. Many African-Americans in the US, for example, have a troubled history with the medical community following the public exposure of the Tuskegee Syphilis Study between 1932 and 1972 in which poor African-American sharecroppers from Alabama were allowed to die of syphilis even after penicillin had been identified as a cure (Kenny, 1972). More recently, the story of Henrietta Lacks was exposed, highlighting the exploitation of a poor, African-American family for medical gain (Skloot, 2010). Henrietta Lacks was an African-American woman who in the early 1950s was treated for and subsequently died of cancer at Johns Hopkins University Hospital. Physicians treating her at Hopkins took Lacks' cells, without her or her family's knowledge, which at the time was common practice. Using Lacks' cells, scientists were able for the first time to replicate an 'immortal' line of human cells grown in culture, paving the way for medical experiments previously impossible on live human subjects and resulting in a number of very profitable industries that included development of the polio vaccine and advances in the treatment of cancer, *in vitro* fertilization, cloning, and gene mapping to name only a few. Meanwhile, members of her family were only recently told of how the cells of Lacks live on, contributing to medical science but no longer owned by the woman or the family who live in poverty and without access to health care (Skloot, 2010). These and other significant historical events affect the rates at which African-Americans in the US, for example, use advance directives and have trust that clinicians will help them to receive equitable treatment at the end of life (Perkins et al., 2002).

The problems with advance directives are many. In the US, for example, too few persons think to establish advance directives prior to an emergency and clinicians too often are not having effective communications with high-risk patients and their families prior to or during an emergency. For the few with advance directives, many of these are not effectively communicated in the heat of a medical emergency. Rarely do patients with advance directives actually bring the written directive with them to the emergency department and without the availability of shared electronic health records in many places in the US the directive goes unseen. In the absence of a written advance directive, family members present with the patient are called on to recall the person's wishes, but conflicting opinions from family members, perceptions and feelings of guilt around being the one to end their loved ones life, or lack of previous communication prior to the emergency event can derail a person's end-of-life preferences. In some cases where written directives are available, clinicians will fail to heed the directive for any number of reasons including conflicting family preferences, doubts about the validity of the directive, misinterpreting the directive, conflicting hospital policy, or simply that clinician's preferences to treat (Brown, 2003; Ishihara et al., 2008; Perkins, 2007; Teno et al., 1997).

Ultimately, advance directives are problematic because as they are currently used, they represent a single point in a decision-making process that is very fluid and, ideally, must occur over time. In a qualitative study with persons dying in the Netherlands, for example, it was found in practice that people make and change their end-of-life wishes frequently. The fluid nature of advance planning in practice manifested as a 'line in the sand,' where dying individuals would say, 'I do not want to go on if I cannot do my art anymore' or 'I will want euthanasia once they remove my other leg,' but invariably as that line in the sand approached, many incorporated their new limitation only to draw another line in the sand for another day (Norwood, 2007, 2009: 75–76). In addition to shifting preferences, it is difficult to detail and plan for the various scenarios that might occur at the end of life, thus life-saving treatment may be desired in one scenario, but not in the next.

There is no quick fix for the problems with advance directives, but they are a step in the right direction. The next step to make advance directives more effective would include supporting on-going conversations that occur among patients, family members, and their clinicians regarding advance care planning and end-of-life preferences; these conversations must be periodically documented in writing in electronic medical record systems that can be readily updated and available in an emergency situation, and policy must exist both at the governmental and health system levels that facilitate the use of patient's preferences at the end of life.

Refusal of Treatment

Patients' rights legislation has emerged in places around the world to detail rights in terms of access to health care. In some places, this also includes language related to a patient's rights to refuse life-sustaining treatments (Erbay et al., 2013; Ryan and Callaghan, 2010; Saks, 2010). In addition to refusal of emergency medical treatment, such as CPR or artificial ventilation, refusal of life-sustaining treatment can also take the form of refusal of food and hydration or a treatment or procedure, such as chemotherapy, that might prolong life. Rights to refuse treatment vary quite a bit from hospital policy, to state policy, to individual preferences and once someone is on life-sustaining supports it may be difficult for families to have supports removed, which makes the decision of refusal prior to the use of artificial supports an important choice to make. In the United Kingdom where end-of-life choices frequently make headlines, for example, they passed the Mental Capacity Act of 2005, which partially allows patients to refuse treatments under limited conditions, but does not allow refusal of food, water, warmth, or hygiene by advance directive (Brauer et al., 2008: 97). In Serbia, on the other hand, the Health Care Act of 2005 allows the patient to "forgo any kind of medical treatment even that of life-saving or life-sustaining" as long as they do so in written form (Brauer et al., 2008: 76).

Medical Behaviors That Potentially Shorten Life

End of life today rarely occurs as a completely intervention-free process, thus the majority of end-of-life choices are typically medical ones. For this reason, the term medical behaviors that potentially shorten life (MBPSL) has emerged to encompass a number of medical practices that contribute to the shortening of life in response to the medicalization of death and dying. There are a number of types of MBPSL including withholding and withdrawal of support (WH/WD), euthanasia and assisted dying (or assisted suicide), terminal or palliative sedation, and life-terminating acts without an explicit request (LAWER) by the patient.

Withholding and Withdrawing Support

Withholding of life-sustaining supports, sometimes referred to as passive euthanasia, includes such medical practices as withholding artificial nutrition and hydration; withholding emergency life resuscitation, such as CPR or ventilation; and withholding life-sustaining treatments or medications, such as supplemental oxygen or antibiotics. It is typically delineated from withdrawing life-sustaining supports, which often involves an additional legal step to have someone actively removed from life-sustaining supports. Withdrawal from supports typically includes such practices as removing someone from a ventilator, nutrition, or any number of other life-sustaining supports, medications, or practices (Smedira et al., 1990).

Much of the public debate around WH/WD emerged as a reaction, again, to the high-profile cases of persons being kept partially alive in persistent vegetative states around the world. In the US, for instance, there was the case of 21-year-old Karen Ann Quinlan in 1976. After a night of drinking and drug use, Karen Quinlan collapsed into a coma at a roadside bar and for almost a year after, her family fought for the right to withdraw her from life-sustaining treatment. The family eventually won their case and Quinlan died 9 years after being disconnected from a respirator (Dowbiggin, 2005: 121–122). In 1989, the Nancy Cruzan case was heard by the US Supreme Court, which ruled that the Constitution supports a patient's right to refuse treatment if there is 'clear and convincing evidence' of the refusal. More recently, the Terri Schiavo case made headlines from 1998 to 2005. Schiavo was just 26 years old when she suffered irreversible brain damage due to complications from an 'iced tea diet' and years of damage from bulimia. She lay in a persistent vegetative state for a year before her husband petitioned for the right to remove her feeding tube. What ensued was an almost 7-year battle between her husband, her parents (who opposed withdrawing life support), the Florida legal system, and eventually the US Supreme Court. Fueled by the media, the Schiavo case tapped into deep-seated concepts of who has the right to decide (parents, husband, or state) and what constitutes life. After several attempts at state and federal legislation to block Schiavo's husband from withdrawing support, the courts determined that Terri Schiavo's wishes prior to her illness were to not continue life-prolonging measures and in March 2005 her feeding tube was removed. She died 13 days later (NNDB, 2008). More recently in India, the Supreme Court ruled in favor of WH/WD under certain circumstances after hearing the case of Aruna Shanbaug, a young woman kept alive for over 30 years on a ventilator (Goel et al., forthcoming).

Withholding and Withdrawing Support in India

In 1973, Aruna Shanbaug, a nurse in one of the leading teaching hospitals in Mumbai was raped and strangled by a hospital custodian whose advances she had spurned, leaving her brain-dead and in a persistent vegetative state. For more than 30 years, Shanbaug, shunned by her family, has been kept alive on a ventilator taken care of by former coworkers at the hospital where she worked. In March 2011, in response to a petition by a journalist on Shanbaug's behalf, India's Supreme Court ruled that Shanbaug should live, while at the same time stating that passive euthanasia – the WH/WD – be permissible in certain cases based on merit. Following the decision, it is no longer illegal to remove supports for persons who are brain-dead or in the case of some persistently vegetative patients with the knowledge and intention of ending their life.

This case sparked a new debate in India on just what kind of end-of-life options physicians and families want available. It also raises important moral and ethical conflicts regarding all types of euthanasia, active and passive, and the maintenance of life in patients whose life span has been prolonged at the cost of their quality of life without adding real hope of recovery.

In August 2011, the Indian Medical Association in association with researchers from the University College of Medical Sciences assembled a team of Indian and international physicians, policy makers, and end-of-life experts to discuss and develop policy recommendations. The group deliberated on issues related to passive euthanasia, the economics of WH/WD, the related economics of keeping vegetative patients alive using invasive technology, legal issues in physician-assisted death, and the preparedness of Indian society for advance directives. While most issues still remain in discussion, it is noteworthy that India has taken up these issues and demonstrates how important it is for societies to work out health policy in relation to their own individual cultural contexts.

With special thanks to Ashish Goel, Assistant Professor in Medicine, University College of Medical Sciences and the Indian Medical Association (Goel et al., forthcoming).

Looking at some of the rates for WH/WD shows how much these practices vary country to country. In a 2001 study of MBPSL across Europe, for instance, the EURELD study revealed that as a percentage of all deaths, WH/WD was the highest in Switzerland (41%) and the lowest in Italy (6%), with the Netherlands (30%), Belgium (27%), Denmark (23%), and Sweden (22%) somewhere in the middle (Griffiths et al., 2008: 490). Clearly, WH/WD is intricately tied to the cultural history of a place and the pervading religious or cultural views toward the end of life that vary among and between nations and their subcultures over time (Devictor and Latour, 2011; Khater et al., 2011; Smedira et al., 1990; Vogel, 2014).

Euthanasia and Assisted Dying (or Assisted Suicide)

Euthanasia and assisted dying, or assisted suicide (depending on what side of the debate the author falls), are two of the most controversial MBPSL that occur today. Euthanasia and assisted dying have existed at least as far back as we have oral history accounts and, according to anecdotal accounts, continue in medical practice today even in places where they are not legally regulated, including in the United States and places across Europe (Emanuel et al., 2000; Heide et al., 2003; Magnusson, 2004; Meier et al., 1998; Morita et al., 2002). Only recently, however, have these practices begun to fall under the realm of law.

Euthanasia (terminating a person's life at that person's explicit request typically for reasons of incurable illness) and assisted dying (providing a person with the means to terminate his or her own life typically for reasons of incurable illness) occur legally in only a few places. The Netherlands has the longest standing legal practice (defined as both legal precedence and transparency of practice) of euthanasia and assisted suicide, legalizing these practices by court decision in 1984 and by law in 2002. Today, euthanasia is legally practiced in the Netherlands (since 1984), Belgium (since 2002), and Luxembourg (since 2009). Assisted dying or assisted suicide is legally practiced in Switzerland (since 1937), the Netherlands (since 1984), Belgium (since 2002), Luxembourg (since 2009), and in the United States in Oregon (since 1997), Washington (since 2008), Montana (since 2009), and Vermont (since 2013). Most recently, the right to assisted dying was upheld by a Second District Court in New Mexico (2014) (Steck et al., 2013). The Netherlands is not the first country in the modern era to legalize euthanasia by law (Australia legalized euthanasia for a brief period in 1995), but it is the first country to legally practice euthanasia and, with nearly 30 years of legal practice, it is the country most often held as the model for or against euthanasia policies in other countries (Cohen et al., 2013; Goldstein, 2008; Heide et al., 2003; SCBI, 2002; Scherer and Simon, 1999).

While euthanasia and assisted dying have been widely debated and researched around the world (Bittel et al., 2002; Cohen et al., 2013; Davis et al., 1993; Downie, 2004; Foley and Hendin, 2002; Førde et al., 1997; Hendry et al., 2013; Plattner, 1996; Quill and Battin, 2004; Televantos et al., 2013), very few comprehensive, ethnographic studies of what euthanasia and assisted dying are in practice exist (Norwood, 2009; Pool, 2000; The, 1997). Using ethnographic research between 1999 and 2001, my research revealed something not previously understood about euthanasia (Norwood, 2009). First, my study highlights that euthanasia in practice is predominantly a discussion that only rarely culminates in a euthanasia death. Statistics from 2001 indicated that approximately 34,700 people initiated euthanasia discussions with their physicians in the Netherlands, yet only 1 in 10 who initiated requests died a euthanasia death (Wal et al., 2003: 46). In 2005, 28,600 people initiated euthanasia discussions and less than 1 in 10 who initiated requests died by euthanasia. Of total deaths, only 2.8% died by euthanasia or assisted suicide in 2001 and by 2005 that number had dropped to 1.8% (Heide et al., 2007; Onwuteaka-Philipsen et al., 2007; Wal et al., 2003). My findings suggest that 'euthanasia talk' – or communications that were initiated among patients, families, and physicians in the wake of a euthanasia request – serves as a kind of Foucauldian discourse grounded in Dutch cultural and historical norms that shape how people think, feel, and act at the end of life. This led to my second major finding, which exposes an interesting irony – that euthanasia talk in many ways serves a palliative function at the end of Dutch life. People use it in part to plan a euthanasia death, but by engaging in these discussions there is also an important by-product to this type of advance planning, which serves to prolong social connections and stave off social death by providing participants with a venue for processing meaning, giving voice to suffering, and reaffirming social bonds and self-identity at the end of Dutch life (Norwood 2009, 2013).

Euthanasia Talk as Palliative Care in the Netherlands

The Netherlands legalized euthanasia and assisted dying by court decision in 1984 and by law in 2002. The Netherlands has the longest standing legal practice of assistance in dying and is typically held as the model for or against euthanasia and assisted dying policies around the world. Between 1999 and 2001, I conducted an ethnographic study of euthanasia and other end-of-life practices in the Netherlands. What I found was not at all what I expected. I found euthanasia to be a practice largely based in talk – a discussion that rarely ends in a euthanasia death. The Dutch home at the end of life is already full with the presence of home care employees and home visits by general practitioners, but once 'euthanasia talk' is invoked, the Dutch public presence in the home increases – with frequent home visits and discussions orchestrated by the general practitioner with patients and their families around a person's request for euthanasia.

Consider the following typical case of euthanasia from the Netherlands. Mrs Veenstra (not her real name), 72 years of age, was an artist, a mother, and a wife when she was diagnosed with inoperable cancer of the lower intestine, 18 years after multiple successful treatments for cancer. This time her cancer had metastasized and she was given approximately 3 months to live. With the support of her husband and adult children, Mrs Veenstra discussed euthanasia with her general practitioner, making both a verbal and written request in the event that her suffering became too great. When I met her she was living on the ground floor of her home and was being taken care of around the clock by her husband who had refused the help of *Thuiszorg*, the home health agency, because he preferred to care of her himself. She had completed her final course of chemotherapy to prolong her life and had chosen to minimize the use of morphine for as long as possible. She told me she wanted to live "as long as I have pleasure." Mrs Veenstra lived 10 months after receiving her terminal prognosis and after initiating her written request for euthanasia with her doctor. She chose euthanasia after multiple conversations with her family and her doctor at the point when she was no longer able to sit up to do her artwork and when she had nearly constant nausea, vomiting, incontinence, and pain.

The consequence of how euthanasia talk has emerged is powerful. In practice, euthanasia talk, even when it does not end in euthanasia death, offers dying persons an alternate route through the end of life. For the dying person, euthanasia talk offers a way to emotionally manage fear, uncertainty, and isolation caused by the often chaotic course of an end-of-life illness. It gives all participants (patients, families, and physicians) an active role at a time when roles are often at best unclear and at worst diminishing. Euthanasia talk offers an ideal image of death that can be used by families to manage the difficult realities of daily decline and loss and to reaffirm family and societal connections. Ultimately, euthanasia talk has emerged as a kind of palliative practice that affirms and maintains social bonds, keeping people connected to Dutch life even as they die at home.

Terminal and Palliative Sedation

Another category of MBPSL includes cases in which death occurs following an increase of medication to alleviate pain symptoms typically accompanied by an explicit request by the patient. This practice is often called terminal or palliative sedation. As a relatively new category of MBPSL, but not necessarily a new practice, this term is defined differently from country to country and researcher to researcher (Maltoni et al., 2012; Morita et al., 2002; Rietjens et al., 2004). It is defined generally as a palliative practice, offering a patient continuous and deep sedation near the end of life and in response to obvious suffering. Griffiths et al. use a more specific definition, defining terminal sedation as "continuous, deep sedation [that] is accompanied by withholding or withdrawing [of] artificial nutrition and hydration for more than two weeks" and palliative sedation as continuous, deep sedation accompanied with or without artificial nutrition and hydration for less than 2 weeks (Griffiths et al., 2008: 76). However it is defined, alleviating pain at the request of patients near the end of life has been occurring in practice for a long time and, like cases of LAWER, these can be viewed as part of the process of good end-of-life care or not depending on the details of each case (Battin, 2008; Gillick, 2004; Tännsjö, 2004).

Life-Terminating Acts without Explicit Request

Of growing concern is a category of MBPSL called LAWER (Jochemsen and Keown, 1999; Pijnenborg and van der Maas, 1993; Rietjens et al., 2007). Legal sociologist, John Griffiths, describes these as cases in which a patient's life is terminated when he or she is close to death, no longer able to communicate, and appearing to be suffering severely (Griffiths et al., 2008: 181). Not every country collects data on this category of MBPSL and not all deaths within this new category are potentially problematic. In the Netherlands, one of the few countries that regularly collect comprehensive statistical data on LAWER cases, they include in this category: deaths of babies under 1 year of age due to withdrawal of treatments, WH/WD cases and cancer cases in which a lethal drug is used to alleviate suffering, and a small number of long-term coma or persistent vegetative state patients. LAWER appear to be a small portion of deaths compared to WH/WD cases (0.4% LAWER vs 16% WH/WD in The Netherlands 2005, for example), but will require on-going scrutiny as nations work out for themselves the boundaries of good care at the end of life (Griffiths et al., 2008: 154, 181).

Arguments around MBPSL

One recurring criticism of the Dutch euthanasia policy and some other MBPSL policies from around the world is the argument popularly known as the 'slippery slope' argument that claims that allowing physicians to conduct euthanasia or assisted dying may eventually lead to patients being killed against their will (Foley and Hendin, 2002; Gillick, 2004; Hendin, 1997; Keown, 2002). Arguments suggesting the presence of a slippery slope in the Netherlands contend that cases of LAWER and terminal sedation where there is an intention by the physician to hasten death are examples of a slippery slope in effect (Hendin, 2002: 97–106). Others suggest that these practices are similar to rates found elsewhere in Europe, that not all cases of LAWER and MBPSL are the same, and that many of these life-terminating behaviors constitute long-standing 'normal medical practice' in the Netherlands and elsewhere (Griffiths et al., 2008: 180–181).

The problem with many of these arguments for and against MBPSL is that they tend to cite data that originate largely from retrospective quantitative surveys and interview-based studies that do not allow researchers an understanding of these behaviors in day-to-day practice. If you look more closely at the types of choices that people are making at the end of a medicalized life, you see that end-of-life discourses exist in many cultures in many places, shaping how people participate in the end of life. Euthanasia talk is just one Dutch discourse at the end of life. End-of-life talk takes many forms in many places and these discourses can be powerful in shaping choice.

Looking at what the Dutch do at the end of life demonstrates one cultural response to restore a 'natural' ideal at the end of life, and a trend that is beginning to gain momentum around the Western industrialized world. Euthanasia policy and euthanasia talk in the Netherlands have allowed Dutch people to replace the image of an isolating, highly medicalized trajectory with an ideal image where family members are assembled bedside in the home and the loved one slips painlessly into death with the help of a trusted physician. The fact that euthanasia deaths rarely occur in reality does not seem to matter. But while this image may be based on a 'natural' ideal where machines and interventions do not dominate the end of life, the act of euthanasia clearly is not natural. It is more than simply unplugging machines and withdrawing interventions; it is replacing one cultural intervention with another (Norwood, 2013).

Future Directions and Alternative Choices at the End of Life

While medicalization of the end of life may dominate, there are other alternative choices that are working to transform how people around the world experience the end of life. Some would argue that less medical attention and more caring for the social person is what is called for at the end of life and there are

a number of nonmedical approaches to dying that are gaining momentum. Palliative care and hospice are on the rise in the US, Europe, and in other places around the world and are beginning to offer a real alternative to medicalized ways of dying. Palliative care, also called comfort care, varies extensively from place to place (Broom et al., 2012; Bruera et al., 2000; Hendry et al., 2013; Higginson et al., 2003; Kellehear, 2000; Lawton, 2000). In some places, palliative care may simply be a retooling of medical care to focus on relieving symptoms and pain management just in the final hours and days of life once all curative measures have ceased; in other places, however, it encompasses something broader – prolonged practices that strive to address the needs of the whole person, including spiritual, existential, emotional, as well as physical needs at and near the end of life. Palliative care is limited, however, in places like the US where it is largely a retooling of medical care, and hospice is equated with the final hours and days of dying. When this occurs, palliative care and hospice are accessed too late in the life course to offer a real alternative to medicalized dying. In places like the Netherlands, however, where whole-patient tenets of palliative care were never extricated from what primary care physicians do, broader forms of caring within the medical establishment exist, offering Dutch people of all ages access to more holistic care throughout the life course (Norwood, 2009: 93–94).

A resurgence of hospital chaplains and spirituality in health care in recent times offers another alternative to the medical pathway at the end of life (Aldridge, 2000; Cadge, 2013; Kellehear, 2000; Puchalski, 2006; Young and Koopsen, 2010). The modern-day chaplains are immersed in a world of hospital medicine, but at the margins they have been able to reestablish a place at the bedside of the sick and dying. In an ethnographic study of hospital chaplaincy in 1997, I found chaplains who were able to forge a strategic, if ambivalent, existence within hospital medicine by alternately embracing and distancing themselves from competing discourses of religion and medicine in order to offer patients what medicine cannot. They bear witness; they bring skin-to-skin touch and comfort through religious ritual at a time when medical treatments are no longer viable. Processing what it means to die, once largely the domain of priests, chaplains, and religion, now rests in many hospitals in the hands of all-faith chaplains who (when skilled) listen without proselytizing and witness without imposing their own agenda (Norwood, 2006a).

Finally, home- and community-based supports for persons who are aging and living with chronic illness and decline are on the rise in Europe and in the US. In the United States, for example, new supports are emerging boosted in part by the passage of the Patient Protection and Affordable Care Act of 2010. These supports are designed to keep more people at home or in community-based settings and are designed increasingly around holistic, person-directed models of long-term care (Gleckman, 2009; SAL, 2001; Thomas, 1996). In places around the US, for instance, there are a number of grassroots organizations designed to help keep people as they age in their own home or community. The Village movement is a community-driven initiative where neighborhoods of persons who are aging band together, paying annual membership fees to support an executive director and office space. Members then have only to request assistance – a ride to the grocery store, assistance picking up prescriptions, or help accessing home health services or supports – and organized volunteers from the community provide these services (Gleckman, 2009: 122–124). When persons can no longer live at home on their own, assisted living has emerged in places around Europe and the US to offer a very different model of residential care compared to the more traditional medical model nursing home. Pioneered in the US by Paul and Terry Klaassen, founders of *Sunrise Senior Living* and expanded by Bill Thomas through the *Eden Alternative* and the *Green House Project*, not all assisted living are created alike. Assisted living can vary from small, eight-person group homes to large facilities, but most are predicated on a founding philosophy exemplified by the Eden Alternative that emphasizes 'person-centered' care, choice, companionship, and dignity over the more traditional medical-model of long-term care (Thomas, 1996: 9). Unfortunately in the US, not everyone has equal access to these home- and community-based alternatives. Home health services are not utilized enough and government-subsidized long-term care for the indigent, Medicaid, cannot currently be used to pay for most assisted living so it remains for the time being for those who can afford to pay.

As people age, they need supports that will help them maintain not only their physical body, but their social being. In a recent study of innovations in long-term care, I found that the ability of persons to participate in reciprocating relationships through the exchange of 'gifts' is a critical part of maintaining social bonds and social life in long-term care. Using anthropologist Marcel Mauss' concept of gift exchange (1967), I found that long-term care environments that foster self-determination, social connection, and reciprocity in relationships help residents maintain a good quality of social life and alternately prevent premature social death – or death of the social being. Even giving the simplest of gifts can foster social connection – from participating in cutting vegetables for dinner, to helping a fellow resident cross a room, to listening to someone tell a story of their life. It is these types of 'gifts' that help people maintain social life and make long-term care settings more like home, particularly near the end of life (Norwood, forthcoming).

For those who cannot (or do not wish to) avoid medicalized ways of dying, choices are also being transformed within medicine. Around the world, health systems are feeling the pressure as more baby boomers age and live longer with chronic conditions. With the increasing cost of end-of-life care – some of the costliest care persons receive (Lynn, 2004: 1–10) – policy makers and other key stakeholders are being forced to make difficult choices in the provision of health and long-term care. In the US, for example, the Patient Protection and Affordable Care Act, or Obamacare as it is popularly named, brings with it a culture of medical change as insurers, hospitals, clinics, and providers around the country are called on under a number of initiatives to improve health care – improving access to health care, improving health outcomes, improving quality of care, while reducing overlap and overall cost of care. While the full impact of this legislation is not

clear, what is clear is that medical practices in the US are changing as a result. At the end of life, this is likely to mean shifts toward home- and community-based solutions and a growing shift away from institutionalized care, such as the traditional nursing home model.

Natural process and cultural intervention pervade end-of-life choices and nature and culture really cannot exist without the other. End-of-life choices around the world exist, but largely exist among a culturally restricted set of options. This is what causes people like Mr Carter and the many others like him to die deaths that are not what they or their family envisioned and worse – hooked up to machines, dying in hospitals, experiencing CPR, and other aggressive forms of intervention in the final days of what they had hoped might be a more peaceful passing. That death – or at least ordinary, everyday dying – is denied to some extent in many places in the modern world is probably true. It is what leaves people like Mr Carter without a medically or culturally supported place to contemplate how he ideally wanted to live prior to his death. It left him with few people to talk to, with little time to process, and without a clear understanding of his medically constructed options. The result for many who die similar deaths around the world is a death without real choices.

Sociologist William Bogard may be right – that choice at the end of life is in part an illusion, and instead how many people in highly industrialized societies die is largely a factor of medical discourses exerting social control to modulate the course of modern life (Bogard, 2008). We die these highly medicalized deaths because that is the current dominant discourse on dying. In many Western industrialized nations our choices are medically mediated choices, offering dying persons and their families mostly institutionalized, aggressive, futile, and painful options for sustaining dying bodies often without regard to the physical, emotional, spiritual, and social suffering that is caused. Death may be a natural event, however treatment of death and dying and the choices that flow from that are largely cultural. Still, at the borders of dominant discourses, human agency and competing discourses are forging alternative pathways. While choice may always be somewhat restricted by cultural mediation, new choices are emerging and competing with dominant discourses to alter and change the practices and meanings of how people around the world die.

Acknowledgments

Portions of this article were reprinted from Norwood, Frances, 2009. The Maintenance of Life: Preventing Social Death through Euthanasia Talk and End-of-Life Care – Lessons from The Netherlands. Carolina Academic Press, Durham, NC, and Norwood, Frances, 2013. A window into Dutch life and death: euthanasia and end-of-life in the public-private space of home. In: Lynch, Caitrin, Danely, Jason (Eds.), Transitions and Transformations: Cultural Perspectives on Aging and the Life Course. Berghahn Books, New York.

Bibliography

Aldridge, David, 2000. Spirituality, Healing, and Medicine: Return to the Silence. Jessica Kingsley Publishers, London.

Arìes, Philippe, 1974. Western Attitudes toward Death: From the Middle Ages to the Present (P.M. Ranum, Trans.). Johns Hopkins University Press, Baltimore.

Battin, Margaret P., 2008. Terminal sedation: pulling the sheet over our eyes. Hastings Center Report 38 (5), 27–30.

Becker, Ernest, 1973. The Denial of Death. Free Press, New York.

Bittel, Nadine, Neuenschwander, Hans, Stiefel, Friedrich, 2002. "Euthanasia": a survey by the Swiss Association for Palliative Care. Supportive Care in Cancer 10 (4), 265–271.

Bogard, W., 2008. Empire of the living dead. Mortality 13 (2), 187–200.

Brauer, Susanne, Biller-Andorno, Nikola, Andorno, Roberto, February 2008. Advance Directives: Towards a Coordinated European Perspective? Institute of Biomedical Ethics, University of Zurich, Switzerland.

Broom, Alex, et al., Feb. 2012. The art of letting go: referral to palliative care and its discontents. Social Science & Medicine 78, 9–16.

Brown, Barbara A., 2003. The history of advance directives. A literature review. Journal of Gerontological Nursing 29 (9), 4.

Bruera, Eduardo, et al., 2000. Attitudes and beliefs of palliative care physicians regarding communication with terminally ill cancer patients. Palliative Medicine 14 (4), 287–298.

Brumley, Richard D., 2002. Future of end-of-life care: the managed care organization perspective. Journal of Palliative Medicine 5 (2), 263–270.

Cadge, Wendy, 2013. Paging God: Religion in the Halls of Medicine. University of Chicago Press.

CBS, 2004. One in Three People in the Netherlands Die in Hospital. Centraal Bureau voor de Statistiek. Web Magazine. http://www.cbs.nl (accessed 24.01.08.).

CDC, 2005. Deaths by place of death, age, race and sex: United States, 2005. In: National Vital Statistics System, Mortality Tables. N.C.f.H.S. Centers for Disease Control and Prevention.

Cohen, J., et al., 2008. Population-based study of dying in hospital in six European countries. Palliative Medicine 22 (6), 702–710.

Cohen, Joachim, et al., 2013. Public acceptance of euthanasia in Europe: a survey study in 47 countries. International Journal of Public Health. http://dx.doi.org/10.1007/00038-013-0461-6.

Davis, Anne J., et al., 1993. An international perspective of active euthanasia: attitudes of nurses in seven countries. International Journal of Nursing Studies 30 (4), 301–310.

Devictor, Denis J., Latour, Jos M., 2011. Forgoing life support: how the decision is made in European pediatric intensive care units. Intensive Care Medicine 37 (11), 1881–1887.

Dowbiggin, Ian, 2005. Euthanasia: A Concise History of Life, Death, God, and Medicine. Rowman & Littlefield Publishers, Inc., Lanham, MD.

Downie, Jocelyn G., 2004. Dying Justice: A Case for Decriminalizing Euthanasia and Assisted Suicide in Canada. University of Toronto Press.

Emanuel, Ezekiel J., Emanuel, Linda L., 1998. The promise of a good death. The Lancet 351, SII21–SII29.

Emanuel, Ezekiel J., et al., 2000. Attitudes and practices of U.S. oncologists regarding euthanasia and physician-assisted suicide. Annals of Internal Medicine 133 (7), 527–532.

Erbay, Hasan, Alan, Sultan, Kadioglu, Selim, 2013. Attitudes of Prehospital Emergency Care Professionals toward Refusal of Treatment: A Regional Survey in Turkey. Nursing Ethics: 0969733013505311.

Foley, Kathleen, Hendin, Herbert (Eds.), 2002. The Case against Assisted Suicide: For the Right to End-of-life Care. Johns Hopkins University Press, Baltimore, MD.

Foltyn, J.L., 2008. Dead famous and dead sexy: popular culture, forensics, and the rise of the corpse. Mortality 13 (2), 153–173.

Førde, Reidun, Gjerløw Aasland, Olaf, Falkum, Erik, 1997. The ethics of euthanasia—attitudes and practice among Norwegian physicians. Social Science & Medicine 45 (6), 887–892.

Foucault, Michel, 1972. The Archaeology of Knowledge (A.M.S. Smith, Trans.). Pantheon Books, New York.

Foucault, Michel, 1973. The Birth of the Clinic: An Archaeology of Medical Perception. Pantheon Books, New York.

Foucault, Michel, Burchell, Graham, Miller, Peter, Gordon, Colin, 1991. The Foucault Effect: Studies in Governmentality with Two Lectures by and an Interview with Michel Foucault. University of Chicago Press, Chicago, p. x, 307.

Gibson, Margaret, 2007. Death and mourning in technologically mediated culture. Health Sociology Review 16 (5), 415–424.

Giddens, Anthony, 1991. Modernity and Self-identity: Self and Society in the Late Modern Age. Stanford University Press, Stanford, CA.

Gillick, Muriel R., 2004. Terminal sedation: an acceptable exit strategy? Annals of Internal Medicine 141 (3), 236–237.

Glaser, Barney G., Strauss, Anselm L., 1968. Time for Dying. Aldine, Chicago.

Gleckman, Howard, 2009. Caring for Our Parents: Inspiring Stories of Families Seeking New Solutions to America's Most Urgent Health Crisis. St. Martin's Press, NY.

Goel, A., et al. Physicians' Role in Euthanasia and End-Of-Life Care Issues in India: A Consultation Report of Stakeholders from a Consultative National Workshop, forthcoming.

Goldstein, Jacob, 2008. Washington Passes Initiative 1000, Legalizing Physician-assisted Suicide. Wall Street Journal Blogs. http://blogs.wsj.com/health/2008/11/05/washington-passes-initiative-1000-legalizing-physician-assisted-suicide/ (accessed 05.11.08.).

Gorer, Geoffrey, 1955. The pornography of death. Encounter 5, 49–53.

Griffiths, John, Weyers, Heleen, Adams, Maurice, 2008. Euthanasia and Law in Europe. Hart Publishing, Oxford.

Hanusch, F., 2008. Graphic death in the news media: present or absent? Mortality 13 (4), 301–317.

Heide, Agnes van der, et al., 2003. End-of-life decision-making in six European countries: descriptive study. The Lancet 362 (9381), 345–350.

Heide, Agnes van der, et al., 2007. End-of-life practices in the Netherlands under the Euthanasia Act. The New England Journal of Medicine 356 (19), 1957–1965.

Hendin, Herbert, 1997. Seduced by Death: Doctors, Patients and Assisted Suicide. W.W. Norton and Company, New York.

Hendin, Herbert, 2002. The Dutch experience. In: Kathleen Foley, M.D., Herbert Hendin, M.D. (Eds.), The Case against Assisted Suicide: For the Right to End-of-life Care. The Johns Hopkins Press, Baltimore, pp. 97–121.

Hendry, Maggie, et al., 2013. Why do we want the right to die? A systematic review of the international literature on the views of patients, carers and the public on assisted dying. Palliative Medicine 27 (1), 13–26.

Higginson, Irene J., et al., 2003. Is there evidence that palliative care teams alter end-of-life experiences of patients and their caregivers? Journal of Pain and Symptom Management 25 (2), 150–168.

Hirai, Kei, et al., 2006. Good death in Japanese cancer care: a qualitative study. Journal of Pain and Symptom Management 31 (2), 140–147.

Hogle, Linda, 1999. Recovering the Nation's Body: Cultural Memory, Medicine, and the Politics of Redemption. Rutgers University Press, New Jersey.

Ishihara, Kristina K., et al., 2008. Advance directives in the emergency department: too few, too late. Academic Emergency Medicine 3 (1), 50–53.

Jochemsen, Henk, Keown, John, 1999. Voluntary euthanasia under control? Further empirical evidence from the Netherlands. Journal of Medical Ethics 25 (1), 16–21.

Kaufman, Sharon, 2005. And a Time to Die: How American Hospitals Shape the End of Life. Scribner, New York.

Kellehear, Allan, 2000. Spirituality and palliative care: a model of needs. Palliative Medicine 14 (2), 149–155.

Kenny, Stephen C., 1972. Tuskegee Syphilis Experiment.

Keown, John, 2002. Euthanasia, Ethics, and Public Policy: An Argument against Legislation. Cambridge University Press, Cambridge, New York.

Khater, Wejdan A., et al., 2011. The practice of withholding and withdrawing life-support measures among patients with cancer in Jordan. International Journal of Palliative Nursing 17 (9), 440.

Lawton, J., 2000. The Dying Process: Patients' Experiences of Palliative Care. Routledge, London.

Lock, Margaret M., 2002. Twice Dead: Organ Transplants and the Reinvention of Death. University of California Press, Berkeley.

Lynn, Joanne, 2004. Sick to Death and Not Going to Take it Anymore! Reforming Health Care for the Last Years of Life. University of California Press, Berkeley.

Magnusson, R.S., 2004. Euthanasia: above ground, below ground. Journal of Medical Ethics 30, 441–446.

Maltoni, Marco, et al., 2012. Palliative sedation in end-of-life care and survival: a systematic review. Journal of Clinical Oncology 30 (12), 1378–1383.

Mauss, Marcel, 1967. The Gift; Forms and Functions of Exchange in Archaic Societies. Norton, New York.

McWhinney, I.R., Bass, M.J., Orr, V., 1995. Factors associated with location of death (home or hospital) of patients referred to a palliative care team. Canadian Medical Association Journal 152 (3), 361.

Meier, Diane E., et al., 1998. A national survey of physician-assisted suicide and euthanasia in the United States. The New England Journal of Medicine 338, 1193–1201.

Morita, Tatsuya, et al., February 2002. Practices and attitudes of Japanese oncologists and palliative care physicians concerning terminal sedation: a nationwide survey. Journal of Clinical Oncology 20 (3), 758–764.

Nader, L., 1996. Naked Science: Anthropological Inquiry into Boundaries, Power, and Knowledge. Routledge, London.

NNDB, 2008. Terri Schiavo: NNDB, Soylent Communications. http://www.nndb.com/people/435/000026357/ (accessed 28.03.08.).

Norwood, Frances, 2006a. The ambivalent chaplain: negotiating structural and ideological difference on the margins of modern-day hospital medicine. Medical Anthropology 25, 1–29.

Norwood, Frances, 2006b. A hero and a criminal: Dutch huisartsen and the making of good death through euthanasia talk in the Netherlands. Medische Antropologie 18 (2), 329–347.

Norwood, Frances, 2007. Nothing more to do: euthanasia, general practice, and end-of-life discourse in the Netherlands. Medical Anthropology 26, 139–175.

Norwood, Frances, 2009. The Maintenance of Life: Preventing Social Death through Euthanasia Talk and End-of-life Care – Lessons from the Netherlands. Carolina Academic Press, Durham, NC.

Norwood, Frances, 2013. A window into Dutch life and death: euthanasia and end-of-life in the public-private space of Home. In: Lynch, C., Danely, J. (Eds.), Transitions and Transformations: Cultural Perspecties on the Life Course. Berghahn Books, New York.

Norwood, Frances. Transforming Long Term Care: Fostering Social Reciprocity and Preventing Social Death Through Form and Practice in The Netherlands. Healthy Aging Research. Barcelona, Spain, forthcoming.

Onwuteaka-Philipsen, B.D., et al., 2007. Evaluatie: Wet toetsing levensbeëindiging op verzoek en hulp bij zelfdoding [Evaluation of the Termination of Life on Request and Assisted Suicide (Review Procedure) Act of 2002]. ZonMw, The Hague.

Perkins, Henry S., 2007. Controlling death: the false promise of advance directives. Annals of Internal Medicine 147, 51–57.

Perkins, Henry S., et al., 2002. Cross-cultural Similarities and Differences in attitudes about Advance care planning. Journal of General Internal Medicine 17 (1), 48–57.

Pijnenborg, Loes, van der Maas, P.J., 1993. Life-terminating acts without explicit request of patient. Lancet 341 (8854), 1196–1199.

Plattner, Andrew L., 1996. Australia's Northern Territory: the first jurisdiction to legislate voluntary euthanasia, and the first to repeal it. DePaul Journal of Health Care Law 1 (3), 645–653.

Pool, Robert, 2000. Negotiating a Good Death: Euthanasia in the Netherlands. The Haworth Press, New York.

Porter, Roy, 2001. The Cambridge Illustrated History of Medicine. Cambridge University Press.

Puchalski, Christina M., 2006. A Time for Listening and Caring: Spirituality and the Care of the Chronically Ill and Dying. Oxford University Press, Oxford.

Quill, Timothy E., Battin, Margaret P. (Eds.), 2004. Physician-assisted Dying: The Case for Palliative Care and Patient Choice. Johns Hopkins University Press, Baltimore, MD.

Rietjens, Judith A.C., et al., 2004. Physician reports of terminal sedation without hydration or nutrition for patients nearing death in the Netherlands. Annals of Internal Medicine 141 (3), 178–185.

Rietjens, Judith A.C., et al., 2007. Using drugs to end life without an explicit request of the patient. Death Studies 31 (3), 205–221.

Ryan, Christopher J., Callaghan, Sascha, 2010. Legal and ethical aspects of refusing medical treatment after a suicide attempt: the Wooltorton case in the Australian context. The Medical Journal of Australia 193 (4), 239–242.

Saks, Elyn R., 2010. Refusing Care: Forced Treatment and the Rights of the Mentally Ill. University of Chicago Press.

SAL, 2001. The Sunrise Story: Lessons Learned, Lives Touched and Love Shared. Sunrise Assisted Living, Inc, McLean.

SCBI, 2002. Euthanasia and Physician-assisted Suicide: Recent Developments and Ethical Analysis. Southern Cross Bioethics Institute.

Scheper-Hughes, Nancy, 2004. Parts unknown undercover ethnography of the organs-trafficking underworld. Ethnography 5 (1), 29–73.

Scheper-Hughes, Nancy, 2005. The last commodity: post-human ethics and the global traffic in "fresh" organs. In: Global Assemblages: Technology, Politics, and Ethics as Anthropological Problems, pp. 145–167.

Scherer, Jennifer M., Simon, Rita James, 1999. Euthanasia and the Right to Die: A Comparative View. Rowman & Littlefield Publishers, Lanham.

Seale, Clive, 1998. Constructing Death: The Sociology of Dying and Bereavement. Cambridge University Press, Cambridge, New York.

Seale, Clive, Geest, Sjaak van der, 2004. Good and bad death: introduction. Social Science and Medicine 58 (2004), 883–885.

Skloot, Rebecca, 2010. The Immortal Life of Henrietta Lacks. Crown.

Smedira, Nicholas G., et al., 1990. Withholding and withdrawal of life support from the critically ill. New England Journal of Medicine 322 (5), 309–315.

Starr, Paul, 1982. The Social Transformation of American Medicine: The Rise of a Sovereign Profession and the Making of a Vast Industry. Basic Books, New York.

Steck, Nicole, et al., 2013. Euthanasia and assisted suicide in selected European countries and US states: systematic literature review. Medical Care 51 (10), 938–944.

Strathern, Marilyn, 1992. After Nature. Cambridge University Press, Cambridge.

Tang, Siew Tzuh, 2003. When death is imminent: where terminally ill patients with cancer prefer to die and why. Cancer Nursing 26 (3), 245–251.

Tännsjö, Torbjörn (Ed.), 2004. Terminal Sedation: Euthanasia in Disguise? Kluwer Academic Publishers, Dordrecht.

Televantos, Anastasios, et al., 2013. Attitudes towards euthanasia in severely ill and dementia patients and cremation in Cyprus: a population-based survey. BMC Public Health 13 (1), 878.

Teno, Joan M., et al., 1997. Do advance directives provide instructions that direct care? SUPPORT Investigators. Study to understand prognoses and preferences for outcomes and risks of treatment. Journal of the American Geriatrics Society 45 (4), 508.

The, Anne-Mei, 1997. 'Vanavond om 8 uur…': verpleegkundige dilemma's bij euthanasie en andere beslissingen rond het levenseinde. Houten/Diegem: Bohn Stafleu van Loghum.

Thomas, William H., 1996. Life Worth Living: How Someone You Love Can Still Enjoy Life in a Nursing Home: The Eden Alternative in Action. VanderWyk & Burnham, Acton, MA.

Townsend, Joy, et al., 1990. Terminal cancer care and patients' preference for place of death: a prospective study. British Medical Journal 301 (6749), 415.

Van der Geest, Sjaak, 2004. Dying peacefully: considering good death and bad death in Kwahu-Tafo, Ghana. Social Science & Medicine 58 (5), 899–911.

Vogel, Lauren, 2014. Texas forced life-support case could happen in Canada. Canadian Medical Association Journal, 109–4711.

Wal, Gerrit van der, et al., 2003. Medische besluitvorming aan het einde van het leven: De praktijk en de toetsingsprocedure euthanasie. de Tijdstroom, Utrecht.

Walter, Tony, 1994. The Revival of Death. Routledge, London.

Yanagisako, Sylvia, Delaney, Carol (Eds.), 1995. Naturalizing Power: Essays in Feminist Cultural Analysis. Routledge, New York.

Young, Caroline, Koopsen, Cyndie, 2010. Spirituality, Health, and Healing: An Integrative Approach. Jones & Bartlett Publishers.

Endocrinology and Psychiatry

Rand Mahmoud, Program in Neuroscience, University of British Columbia, Centre for Brain Health, Vancouver, BC, Canada
Liisa AM Galea, Program in Neuroscience, Centre for Brain Health, Vancouver, BC, Canada; and University of British Columbia, Vancouver, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by F. Holsboer, volume 7, pp. 4504–4510, © 2001, Elsevier Ltd.

Abstract

The nervous and endocrine systems are involved in a large degree of cross-talk, and are intricately linked in both structure and function. The brain regulates the secretion of endocrine hormones, which in turn have profound effects on brain and behavior through the abundant hormone receptors in the central nervous system. Therefore, it is not surprising that steroid hormones are implicated in the etiology, manifestation, and treatment of psychiatric illness. Here, we focus on stress and gonadal hormones and their roles in depression, bipolar disorder, and schizophrenia.

Despite decades of research, our knowledge around the etiology and pathophysiology of psychiatric disorders is incomplete. More importantly, available pharmacotherapies are effective in only a subpopulation of patients, and therefore efforts to discover new and improved drug targets are ongoing. However, considerable evidence has emerged in support of the involvement of the neuroendocrine system in various psychiatric disorders. In particular, there is a growing recognition of the sex differences in incidence, symptoms, and efficacy of treatment, which have been largely ignored in the literature thus far. Given these sex differences, it is perhaps not surprising that gonadal hormones are implicated in many neuropsychiatric disorders. Here, we focus our discussion on two major neuroendocrine systems: the hypothalamic–pituitary–adrenal (HPA) axis, and the hypothalamic–pituitary–gonadal (HPG) axis, both of which are implicated in the etiology, manifestation, and treatment of several psychiatric disorders, including depression, bipolar disorder, and schizophrenia.

Cross-Talk between the Nervous and Endocrine Systems

Both structurally and functionally, the nervous and endocrine systems are intimately connected. Largely through the hypothalamus, the central nervous system (CNS) regulates the release of hormones from the pituitary gland at the base of the brain. In turn, pituitary hormones regulate a variety of physiological processes directly by acting on target tissue, or indirectly by inducing the secretion of hormones from other endocrine glands. Besides acting on target tissue, hormones can bind to receptors at various levels of the neuroendocrine system, allowing for the recognition and regulation of circulating hormone levels, as further described below in the case of the HPA axis. As a result of this relationship between the nervous and endocrine systems, we are able to gain unique insights into the functioning of the relatively inaccessible CNS through circulating levels of hormones.

The HPA Axis and Negative-Feedback Regulation

In response to a real or perceived stressful event, the HPA axis is activated, and as a result, there is an increased release of glucocorticoids from the adrenal glands into the bloodstream. Cortisol and corticosterone, in humans and rodents, respectively, are main glucocorticoid hormones involved in the biological stress response. In the face of threat, activation of the HPA axis is essential, as glucocorticoids mobilize energy stores to prepare to deal with the stressors such as escaping from danger. The stress-induced HPA activation is a multistep hormonal process, initiated at the activation of neurons in the paraventricular nucleus (PVN) of the hypothalamus. When activated, the PVN secretes corticotropin-releasing hormone (CRH), which in turn acts on the anterior pituitary gland, inducing the release of adrenocorticotropic releasing hormone (ACTH) into the bloodstream. Finally, ACTH acts on the medulla (the inner portion) of adrenal glands, stimulating the release of glucocorticoids. The action of glucocorticoids is not limited to the activation of the stress response; they also play a critical role in the termination of that same response. In a negative-feedback loop, elevations in glucocorticoid levels are detected, in large by the hippocampus, which in turn signals the hypothalamus to suppress further release of CRH. This negative-feedback inhibition is essential as it prevents prolonged exposure of tissue to high levels of glucocorticoids. Prolonged exposure to high levels of glucocorticoids is grossly detrimental to health.

The HPG Axis

In a multistep hormonal cascade similar to that described for the HPA axis, the HPG axis is a neuroendocrine system that regulates the release of gonadal hormones from the gonads (ovaries or testes). The functions of gonadal hormones extend well beyond regulating sexual and reproductive functions, to include effects on mood and cognition among other CNS processes. These effects are primarily mediated through gonadal hormone receptors that are found in abundance throughout the brain, and in the limbic system in particular, including the hippocampus. It is important to note here that sex differences in disease prevalence, symptomology, or treatment can point to an involvement of gonadal hormones in the disease process. Implications of gonadal hormones in neuropsychiatric disorders may be due to their involvement in the pathophysiology at early stages (during development), later in life (during adolescence or adulthood), or both.

Neuroendocrine Involvement in Mood Disorders

In the following section, we discuss the significant neuroendocrine changes that are associated with depression and bipolar disorder, two of the most prevalent and serious psychiatric mood disorders. We summarize findings from the literature suggesting that HPA and/or HPG changes may be linked to the risk for, symptom profile, and treatment of these disorders.

Stress, HPA Function, and Depression

While the etiology of depression is not fully understood, mounting evidence suggests that chronic stress exposure is intimately linked to the development of the disorder (reviewed in Tennant, 2002). The HPA axis plays a key role in regulating the biological response to stress, and as such, it is not surprising that HPA axis dysfunction is a biomarker of depression (reviewed in Stetler and Miller, 2011). There are major differences in HPA function between depressed and nondepressed individuals, the first of which is elevated circulating cortisol levels in depression. Additionally, the characteristic diurnal rhythm of cortisol secretion, which features distinct cortisol peaks in the morning and evening, is dampened in depressed individuals. Importantly, the ability of the HPA axis to self-regulate, that is, to recover after a stress response via the negative-feedback system, is also compromised in depression. The integrity of the HPA negative-feedback system can be evaluated using the dexamethasone (DEX) suppression test. This medical test utilizes DEX, a synthetic glucocorticoid, which essentially mimics the endogenous cortisol in the negative-feedback system of the HPA axis. Therefore, when administered to healthy individuals, DEX reduces HPA activity and suppresses the release of cortisol. In many depressed individuals, however, DEX fails to suppress cortisol release, indicating a faulty negative-feedback system (Ising et al., 2007).

Because HPA dysfunction is the best described biological endophenotype associated with depression, restoration of its function could be a central drug target. In fact, commonly prescribed antidepressant medications are effective in restoring HPA disruption only in patients who respond favorably to treatment. In other words, antidepressant alleviation of depressive symptoms seems to be linked to a restoration of HPA dysfunction. Temporally, this restoration coincides with, or just precedes, the relief from depressive symptoms (Ising et al., 2007). This indicates that in treatment responsive patients, HPA dysfunction may be central to the pathophysiology of the disorder. In order to gain a better understanding of the relationship between stress, depression, and HPA dysfunction, researchers use animal models of depression to assess relevant brain, behavioral, or neuroendocrine phenotypes (reviewed in Yan et al., 2010). Two well-established animal models that show considerable validity and reliability are the chronic variable stress model and the chronic corticosterone administration model. Interestingly, the neuroendocrine changes demonstrated in these models are similar to those observed in depression patients. Such experimental models are fundamental in understanding the etiology of the disorder, the mechanisms of antidepressant action, and in the development of novel treatments.

Sex Differences in Depression

Of the most robust epidemiological findings in depression is the higher prevalence of the disorder in women, consistent across cultures studied, typically here is a prevalence ratio of at least 2:1 for women to men for major depressive disorder (Gutierrez-Lobos et al., 2002). This sex difference in prevalence arises at puberty, declining after the age of 50, and is most striking during women's reproductive years. Specifically, the postpartum and perimenopause are two periods in a women's life wherein the risk of depression is largely increased (Cohen et al., 2006; Hendrick et al., 1998). Interestingly, these two periods are associated with dramatic fluctuations in gonadal and/or stress hormones. While many hypotheses, both biological and psychosocial, have attempted to explain the sex differences in depression, the best supported are neuroendocrine theories highlighting the role of gonadal hormones (Hammarstrom et al., 2009). In addition to the higher prevalence of depression in women, there are differences in the presenting symptomology between men and women. For example, women are more likely to develop atypical depression, which is characterized by excessive fatigue, hypersomnia, and/or hyperphagia (Young et al., 1990; Silverstein, 2002), and are more likely to have comorbid anxiety (Sloan and Kornstein, 2003; Keers and Aitchison, 2010).

Gonadal Hormone Levels Are Associated with Depression and Its Treatment

Perimenopause is a 10-year period of transition into menopause, involving dramatic hormonal changes leading to a permanent state of hypogonadism, that is, diminished levels of gonadal hormones. Women with or without a previous history of depression are at an increased risk of developing the disorder in this period (Freeman et al., 2006), suggesting that the associated hormonal changes are involved in the disease. However, it should be noted that there is rarely a direct association between serum levels of ovarian hormones and depression ratings. In men, levels of gonadal hormones are also associated with depression. This is evident from correlational studies showing that both younger and older depressed men have lower levels of testosterone, the main testicular hormone in men (McIntyre et al., 2006). Furthermore, the prevalence of depression is greater among hypogonadal men. Additionally, in both men and women, gonadal hormone levels are linked to antidepressant efficacy. For example, in postmenopausal women diagnosed with depression, selective serotonin reuptake inhibitors (SSRIs) are more effective in relieving depression symptoms when prescribed alongside hormone therapy (HT) (Thase et al., 2005). Similarly, testosterone as an adjunct therapy in depressed hypogonadal men is effective in boosting the efficacy of antidepressants, as shown by a meta-analysis (Zarrouf et al., 2009). Taken together, it is clear that low levels of gonadal hormones in both men and women are associated with depression, and possibly resistance to antidepressant treatment. Moreover, in depressed hypogonadal men or women, HT may be an effective adjunct therapy.

Interestingly, the ability of antidepressants to ameliorate depression symptoms is dependent on sex, where SSRIs are more effective in women, and tricyclic antidepressants (TCAs) are more effective in men (Keers and Aitchison, 2010). This

implies that gonadal hormones, and/or chromosomal differences, may be important for antidepressant efficacy beyond the context of hypogonadism. Moreover, the capability of antidepressants to restore HPA axis function shows a disparity by sex, in which the ability to normalize HPA negative-feedback dysregulation is better associated with remission in women as compared to men (Binder et al., 2009). Although beyond the scope of this article, this difference may be in part modulated by the well-established interactions between the HPA and HPG axes.

Gonadal and Stress Hormones in Postpartum Depression

The postpartum period marks the time of greatest risk for developing depression in women. In fact, postpartum depression (PPD) affects a striking estimate of 15% of women (Wisner et al., 2013). Apart from the delirious effects on the well-being of mothers, PPD is associated with poorer cognitive, social, and emotional outcomes for children of affected mothers. Importantly, pregnancy and postpartum are coupled with dramatic neuroendocrine changes, mainly in HPG and HPA hormones, and some of which are amplified by the placenta. In terms of HPA function, pregnancy and postpartum are associated with elevated cortisol levels, and blunted diurnal patterns of peak cortisol release. Intriguingly, this hormonal profile is similar to that commonly observed in depressed patients, as we discussed previously. This suggests that pregnancy-related changes in HPA function may contribute to the increased risk of depression for women in this period. On the other hand, changes in ovarian hormones include dramatic increases of estrogens (notably estradiol and estriol) throughout gestation, with estradiol reaching levels 200 times normal level at week 20 of gestation, and a sudden prolonged drop in estrogens after delivery of the placenta (itself an endocrine gland). Indeed, postpartum women are considered hypogonadal for up to 180 days after parturition. These dramatic fluctuations in hormones led to the development of the 'ovarian hormone-withdrawal' hypothesis of PPD. This hypothesis proposes that the elevated levels of estrogens during pregnancy, followed by the sudden hypogonadal state post-delivery may play a role in the etiology of PPD. In support of this hypothesis, studies have shown that hormone-stimulated estradiol withdrawal led to a significant increase in depressive symptomology and enhanced cortisol release in nondepressed women with a history of PPD when compared to nondepressed women without a history of PPD (Bloch et al., 2000, 2005). This illustrates that women who develop PPD may be more vulnerable to the effects of pregnancy-related changes in HPG hormones, although what underlies this vulnerability remains unclear. Animal models of PPD are important for enhancing our understanding of the unique female physiology related to pregnancy and postpartum, and how it relates to the increased risk for depression. In an attempt to mimic the human condition, such models emphasize the role of stress and sex hormones. For example, the chronic administration of corticosterone to rats in the postpartum period increases depressive-like behavior in the mothers, and decreases maternal behavior toward her pups, while withdrawal from a hormone-simulated pregnancy increases depressive-like behavior and decreases hippocampal plasticity (reviewed in Brummelte and Galea, 2010).

HPA and HPG Involvement in Bipolar Disorder

Bipolar disorder has a lifetime prevalence of approximately 4%, and is characterized by episodes of mania or hypomania, alternating with periods of depression. In a similar, but more exaggerated pattern to that seen depression, individuals with bipolar disorder present with indicators of HPA dysfunction (reviewed in Daban et al., 2005). This includes elevated basal cortisol and negative-feedback dysregulation as measured by the failure of DEX to suppress cortisol. However, while treatment-associated alleviation of symptoms is linked to the restoration of HPA function in depression, the same relationship is not apparent in bipolar disorder. Lithium-treated bipolar patients in remission still exhibit elevated levels of cortisol and negative-feedback dysregulation, suggesting that HPA dysregulation may be at the core of the disorder, or alternatively that lithium treatment acts on a different disease-associated system. Interestingly, the pattern of HPA dysfunction may be predictive of future manic/hypomanic relapse in patients with bipolar disorder. This has been shown in a CRH stimulation study carried out in stable medicated bipolar patients, where exogenous CRH resulted in higher ACTH levels in patients that relapsed 6–12 months following the test (Vieta et al., 1999), indicating a higher sensitivity of the pituitary to CRH prior to relapse.

Unlike depression, which is a unipolar mood disorder, there does not seem to be a clear sex disparity in the lifetime prevalence rates of bipolar disorder. However, there are differences in the manifestation of the disorder between men and women, and in periods of vulnerability, suggesting that gonadal hormones are still important to consider for a holistic understanding of the disease underpinnings. As reviewed by Diflorio and Jones (2010), sex difference in bipolar disorder are apparent in the subtypes and clinical presentation, including a higher prevalence of type II diagnosis (hypomania predominant), and a higher likelihood of mixed and rapid cycling episodes in women. Interestingly, in women with bipolar disorder, gonadal hormone levels may be linked to the severity of symptoms or the recurrence of bipolar episodes. Suggestive of this are findings showing a significant number of women with bipolar disorder having increased symptoms in the perimenopausal period (Marsh et al., 2009), when ovarian hormones begin to decline prior to the onset of menopause. Additionally, postmenopausal women with bipolar disorder who are not taking HT are more likely to experience a worsening of symptoms after menopause in comparison to women taking HT (Freeman et al., 2002). More evidence linking levels of female gonadal hormones to the disease manifestation comes from the literature studying bipolar women during postpartum, which is a period associated with an especially high risk of severe episode recurrence (Jones and Craddock, 2001).

Neuroendocrine Involvement in Schizophrenia

Schizophrenia is a serious psychiatric disorder that affects approximately 1% of the population. It is characterized by positive symptoms (including hallucinations and delusions), negative symptoms (such as social withdrawal, flat affect, and poverty of speech), and cognitive impairments (such as

memory, executive function, and attention deficits). The etiology of schizophrenia is complex and heterogeneous; however, amassing evidence suggests an interaction of genetic vulnerabilities and environmental factors in the pathogenesis of the disorder. In the following section we present evidence of significant HPA and HPG involvement in schizophrenia.

Stress, HPA Function, and Schizophrenia

Unlike in depression, large population studies do not show a clear relationship between exposure to stress and risk for developing schizophrenia. That is, on average, individuals with a diagnosis of schizophrenia do not have a history of more severe or frequent stress exposure than the healthy population. However, as reviewed in Walker et al. (2008), there is strong evidence illustrating that stressful life events often precede relapse of psychotic episodes and exacerbate symptoms in schizophrenia patients. Moreover, in individuals with a high genetic risk of developing schizophrenia, that is, those with a family history of the disorder, exposure to stressful life events is linked to a higher likelihood for a future diagnosis of schizophrenia. Taken together, this suggests that in individuals predisposed to the disorder, stress-induced activation of the HPA axis may trigger vulnerable neural substrates related to schizophrenia.

A body of literature shows signs of HPA dysfunction in schizophrenia patients (reviewed in Walker and Diforio, 1997), the first being elevated baseline levels of cortisol and ACTH. This effect is seen in first-episode and chronic patients, and is especially evident in nonmedicated patients. Moreover, elevations in cortisol are correlated with an exacerbation of negative symptoms. Both typical and atypical antipsychotics have a restorative effect of HPA function, whereby they have been shown to reduce elevations in cortisol and ACTH. Interestingly, antipsychotic drugs are more effective in alleviating symptoms in patients presenting with more profound cortisol elevations prior to the initiation of treatment (Ritsner et al., 2005). This indicates that antipsychotic drugs may target HPA dysfunction through the suppression of the HPA axis. In addition, HPA dysfunction may be at the root of the disorder in a subpopulation of patients, further emphasizing that schizophrenia is a largely heterogeneous disease.

Sex Differences in Schizophrenia and the Role of Gonadal Hormones

While the onset of schizophrenia is typically in late adolescence/early adulthood, individuals who go on to develop the disorder often present with a period of social, behavioral, and cognitive decline during early adolescence, suggesting that the manifestation of the disorder coincides with a time of considerable changes for HPG function and maturation of HPA axis. Interestingly, there are clear sex differences in the lifetime prevalence rates of schizophrenia, in which men are approximately 40% more likely to develop the disorder than women (McGrath et al., 2004). Additionally, the course of illness shows distinct differences between men and women. Although the onset of schizophrenia is typically late in adolescence/early adulthood for both sexes, the average age of onset in men is 3–4 year prior to that in women (Munk-Jorgensen, 1987). In other words, the peak onset is earlier and steeper in men than in women. Intriguingly, women also have a second peak in incidence rate during late perimenopause, or the mid-late 40s, which is not seen in men around the same age. This pattern of incidence rates across the lifespan leads to hypotheses suggesting that estrogens, such as estradiol, which are more abundant in women, may play a protective role against the development of schizophrenia. There is also evidence that in schizophrenia patients, temporal lobe volume is decreased in men but not in women (Bryant et al., 1999), which may be a neural correlate of the fact that men with schizophrenia are more likely to present with more severe cognitive disturbances than women with schizophrenia.

Providing support for the hypothesis that estrogens play a protective role in the disease is a plethora of data linking levels of ovarian hormones to the onset and severity of symptoms. For example, in patients diagnosed with schizophrenia, an earlier onset of puberty is linked to a delayed onset of the disorder in women only (Cohen et al., 1999). Interestingly, pregnancy is associated with a reduction of symptoms, and this effect may be attributed to the dramatic increase of estrogens throughout gestation. Furthermore, the psychotic symptom severity in women diagnosed with schizophrenia varies alongside the menstrual cycle (Bergemann et al., 2007). That is, psychotic symptoms are exacerbated around menstruation, which is the period during a women's menstrual cycle with the lowest circulating levels of estradiol. Additionally, levels of ovarian hormones may have important implications in the treatment of the disorder. In fact, there is evidence suggesting an interaction between gonadal hormone levels and antipsychotic efficacy in women. For example, in antipsychotic-treated postmenopausal women with schizophrenia, HT was associated with a lower prescribed dose of antipsychotic treatment, and independent of antipsychotic dose, HT was also associated with less severe negative symptoms (Lindamer et al., 2001). Hence, as an adjunct therapy, HT may improve prognosis for postmenopausal women with schizophrenia. Altogether, these data suggest that there is an inverse relationship between levels of female gonadal hormones and severity of schizophrenia symptoms, with higher levels of gonadal hormones correlating with less severe symptom presentation.

If ovarian hormones in women reduce the vulnerability threshold for developing schizophrenia, then it would be reasonable to expect that women with schizophrenia will have lower mean levels of gonadal hormones when compared to women without the disorder. In fact, this has been supported by studies in which circulating estradiol levels were measured across several menstrual cycles in schizophrenia patients, and were found to be lower than the control population across all phases of the menstrual cycle (Riecher-Rossler et al., 1994). From these data, however, it is not possible to interpret the directionality of this relationship. That is to say, it is not clear whether lower levels of estradiol precede the illness and constitute a risk factor, or represent a consequence of the disorder. Building on this evidence of hypogonadism in premenopausal women with schizophrenia, several placebo-controlled clinical trials tested the efficacy of estradiol as an adjunct therapy in the alleviation of symptoms. These studies showed that estradiol paired with antipsychotic treatment resulted in a more rapid and more effective alleviation of

symptoms in comparison with antipsychotic treatment alone (Kulkarni et al., 1996; Akhondzadeh et al., 2003). Therefore, this indicates that HT may be an effective adjunct therapy in both pre- and post-menopausal women with schizophrenia.

From the statistics showing a higher prevalence of schizophrenia in men, it can be speculated that, rather than estrogens being protective in women, the more abundant testosterone in men may contribute to their vulnerability. For this speculation to hold merit, we should expect to see elevated levels of testosterone in men with schizophrenia; however, the evidence suggests otherwise. In fact, men with schizophrenia are reported to have lower or equal mean testosterone levels in comparison to men without the disorder. Moreover, in men with schizophrenia, there is a correlation between the severity of negative symptoms and circulating testosterone levels, where lower testosterone levels are found in patients presenting with more severe negative symptoms (Shirayama et al., 2002). Therefore, similar to the data in women, hypogonadism seems to be associated with schizophrenia in men, and the question of whether this effect precedes or result from the disorder is similarly unclear. Moreover, there is evidence suggesting that testosterone is effective as an adjunct therapy as it boosts the effects of antipsychotic medication to alleviate negative symptoms of schizophrenia in men (Ko et al., 2008).

Endocrine Disorders and the Risk for Psychiatric Illness

So far in our discussion, we provide examples of endocrine dysfunction associated with a primary diagnosis of psychiatric illness. However, the inverse relationship also exists, where a primary diagnosis of endocrine illness can be associated with an increased risk of developing psychiatric disorders. A prime example of this relationship comes from Cushing's syndrome, which regardless of etiology is characterized by chronically elevated levels of glucocorticoids. The most common cause of Cushing's syndrome is pituitary tumor (adenoma), resulting in an excess secretion of CRH. As reviewed in Sonino and Fava (2001), Cushing's syndrome is associated with a high prevalence rate of depression, episodes of mania or hypomania, and anxiety disorders. Interestingly, a restoration of HPA function, for example through surgery in the case of pituitary adenoma, is effective in the alleviation of psychiatric symptoms. Thus, findings from the psychiatric comorbidities in Cushing's syndrome provide important support for the strong relationship between endocrinology and psychiatry.

Conclusions

The literature we overview here, with a focus on depression, bipolar disorder, and schizophrenia, suggests that significant neuroendocrine changes are associated with psychiatric disorders, and that neuroendocrine dysfunction may lie at the core of the etiology and pathophysiology of these disorders, in at least a subpopulation patients. The functional interaction between the nervous and endocrine systems provides a unique window into the CNS, and allows for endocrine hormones to be useful as biological markers for psychiatric disorders, ultimately aiding our understanding of disease underpinnings and enhancing treatment prospects.

We summarize our discussion in three major themes:

- Stress is a risk factor for the development of psychiatric disorders in vulnerable individuals, and is associated with relapse and exacerbation of symptoms. Additionally, a hyperactive HPA axis and HPA negative-feedback dysregulation are well-documented neuroendocrine changes associated with all psychiatric disorders discussed here. Therefore, HPA dysfunction is an important target for the development of novel pharmacotherapies.
- There are many sex differences in the prevalence rates, clinical presentation, and treatment outcomes of psychiatric disorders, and the unique physiology of women is important to consider in this context. As such, understanding the role of gonadal hormones in various aspects of the disease mechanism is essential to the development of individualized, effective, and safe pharmacotherapies for both men and women.
- Gonadal hormones have protective effects on mental health in both men and women, and are implicated in the treatment of psychiatric illness. This was demonstrated in the discussion of periods where low levels gonadal hormones in women are associated with an increased risk for, or exacerbated symptoms of psychiatric disorders (i.e., perimenstruation, postpartum, and post-menopause). Similarly, hypogonadism is associated with psychiatric disorders in men, and may be associated with more severe symptoms. Finally, in hypogonadal men and women, there is evidence that HT is an effective adjunct therapy alongside traditional treatment for psychiatric disorders.

See also: Antidepressant Drugs; Bipolar (Manic-Depressive) Disorder; Depression; Gene–Environment Interactions in Health and Well-Being; Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of; Hypothalamus; Mental Disorders, Epidemiology of; Mental Illness, Etiology of; Mental Illness, Genetics of; Neuroendocrinology; Placebo Studies (Double-Blind Studies); Schizophrenia, Treatment of; Schizophrenia: Metabolic Consequences; Schizophrenia.

Bibliography

Akhondzadeh, S., Nejatisafa, A.A., Amini, H., Mohammadi, M.R., Larijani, B., Kashani, L., Raisi, F., Kamalipour, A., 2003. Adjunctive estrogen treatment in women with chronic schizophrenia: a double-blind, randomized, and placebo-controlled trial. Progress in Neuro-Psychopharmacology and Biological Psychiatry 27 (6), 1007–1012.

Bergemann, N., Parzer, P., Runnebaum, B., Resch, F., Mundt, C., 2007. Estrogen, menstrual cycle phases, and psychopathology in women suffering from schizophrenia. Psychologie Medicale 37 (10), 1427–1436.

Binder, E.B., Kunzel, H.E., Nickel, T., Kern, N., Pfennig, A., Majer, M., Uhr, M., Ising, M., Holsboer, F., 2009. HPA-axis regulation at in-patient admission is associated with antidepressant therapy outcome in male but not in female depressed patients. Psychoneuroendocrinology 34, 99–109.

Bloch, M., Rubinow, D.R., Schmidt, P.J., Lotsikas, A., Chrousos, G.P., Cizza, G., 2005. Cortisol response to ovine corticotropin-releasing hormone in a model of pregnancy and parturition in euthymic women with and without a history of postpartum depression. Journal of Clinical Endocrinology and Metabolism 90, 695–699.

Bloch, M., Schmidt, P.J., Danaceau, M., Murphy, J., Nieman, L., Rubinow, D.R., 2000. Effects of gonadal steroids in women with a history of postpartum depression. American Journal of Psychiatry 157 (6), 924–930.

Brummelte, S., Galea, L.A., 2010. Depression during pregnancy and postpartum: contribution of stress and ovarian hormones. Progress in Neuro-Psychopharmacology and Biological Psychiatry 34 (5), 766–776.

Bryant, N.L., Buchanan, R.W., Vladar, K., Breier, A., Rothman, M., 1999. Gender differences in temporal lobe structures of patients with schizophrenia: a volumetric MRI study. American Journal of Psychiatry 156 (4), 603–609.

Cohen, L.S., Soares, C.N., Vitonis, A.F., Otto, M.W., Harlow, B.L., 2006. Risk for new onset of depression during the menopausal transition: the Harvard study of moods and cycles. Archives of General Psychiatry 63, 385–390.

Cohen, R.Z., Seeman, M.V., Gotowiec, A., Kopala, L., 1999. Earlier puberty as a predictor of later onset of schizophrenia in women. American Journal of Psychiatry 156 (7), 1059–1064.

Daban, C., Vieta, E., Mackin, P., Young, A.H., 2005. Hypothalamic-pituitary-adrenal axis and bipolar disorder. Psychiatric Clinics of North America 28 (2), 469–480.

Diflorio, A., Jones, I., 2010. Is sex important? Gender differences in bipolar disorder. International Review of Psychiatry 22 (5), 437–452.

Freeman, E.W., Sammel, M.D., Lin, H., Nelson, D.B., 2006. Associations of hormones and menopausal status with depressed mood in women with no history of depression. Archives of General Psychiatry 4, 375–382.

Freeman, M.P., Smith, K.W., Freeman, S.A., McElroy, S.L., Kmetz, G.E., Wright, R., Keck Jr., P.E., 2002. The impact of reproductive events on the course of bipolar disorder in women. Journal of Clinical Psychiatry 63, 284–287.

Gutierrez-Lobos, K., Scherer, M., Anderer, P., Katschnig, H., 2002. The influence of age on the female/male ratio of treated incidence rates in depression. BMC Psychiatry 2, 3.

Hammarstrom, A., Lehti, A., Danielsson, U., Bengs, C., Johansson, E.E., 2009. Gender-related explanatory models of depression: a critical evaluation of medical articles. Public Health 123, 689–693.

Hendrick, V., Altshuler, L.L., Suri, R., 1998. Hormonal changes in the postpartum and implications for postpartum depression. Psychosomatics 39, 93–101.

Ising, M., Horstmann, S., Kloiber, S., Lucae, S., Binder, E.B., Kern, N., Kunzel, H.E., Pfennig, A., Uhr, M., Holsboer, F., 2007. Combined dexamethasone/corticotropin releasing hormone test predicts treatment response in major depression – a potential biomarker? Biological Psychiatry 62, 47–54.

Jones, I., Craddock, N., 2001. Familiality of the puerperal trigger in bipolar disorder: results of a family study. American Journal of Psychiatry 158, 913–917.

Keers, R., Aitchison, K.J., 2010. Gender differences in antidepressant drug response. International Review of Psychiatry 22 (5), 485–500.

Ko, Y.H., Lew, Y.M., Jung, S.W., Joe, S.H., Lee, C.H., Jung, H.G., Lee, M.S., 2008. Short-term testosterone augmentation in male schizophrenics: a randomized, double-blind, placebo-controlled trial. Journal of Clinical Psychopharmacology 28 (4), 375–383.

Kulkarni, J., de Castella, A., Smith, D., Taffe, J., Keks, N., Copolov, D., 1996. A clinical trial of the effects of estrogen in acutely psychotic women. Schizophrenia Research 20 (3), 247–252.

Lindamer, L.A., Buse, D.C., Lohr, J.B., Jeste, D.V., 2001. Hormone replacement therapy in postmenopausal women with schizophrenia: positive effect on negative symptoms? Biological Psychiatry 49 (1), 47–51.

Marsh, W.K., Ketter, T.A., Rasgon, N.L., 2009. Increased depressive symptoms in menopausal age women with bipolar disorder: age and gender comparison. Journal of Psychiatric Research 43, 798–802.

McGrath, J., Saha, S., Welham, J., El Saadi, O., MacCauley, C., Chant, D., 2004. A systematic review of the incidence of schizophrenia: the distribution of rates and the influence of sex, urbanicity, migrant status and methodology. BMC Medicine 2, 13.

McIntyre, R.S., Mancini, D., Eisfeld, B.S., Soczynska, J.K., Grupp, L., Kinarski, J.K., Kennedy, S.H., 2006. Calculated bioavailable testosterone levels and depression in middle-aged men. Psychoneuroendocrinology 31 (9), 1029–1035.

Munk-Jorgensen, P., 1987. First-admission rates and marital status of schizophrenics. Acta Psychiatrica Scandinavica 76 (2), 210–216.

Riecher-Rossler, A., Hafner, H., Stumbaum, M., Maurer, K., Schmidt, R., 1994. Can estradiol modulate schizophrenic symptomatology? Schizophrenia Bulletin 20 (1), 203–214.

Ritsner, M., Gibel, A., Maayan, R., Ratner, Y., Ram, E., Biadsy, H., Modai, I., Weizman, A., 2005. Cortisol/dehydroepiandrosterone ratio and responses to antipsychotic treatment in schizophrenia. Neuropsychopharmacology 30 (10), 1913–1922.

Shirayama, Y., Hashimoto, K., Suzuki, Y., Higuchi, T., 2002. Correlation of plasma neurosteroid levels to the severity of negative symptoms in male patients with schizophrenia. Schizophrenia Research 58 (1), 69–74.

Silverstein, B., 2002. Gender differences in the prevalence of somatic versus pure depression: a replication. American Journal of Psychiatry 159, 1051–1052.

Sloan, D.M., Kornstein, S.G., 2003. Gender differences in depression and response to antidepressant treatment. Psychiatric Clinics of North America 26 (3), 581–594.

Sonino, N., Fava, G.A., 2001. Psychiatric disorders associated with Cushing's syndrome. Epidemiology, pathophysiology and treatment. CNS Drugs 15 (5), 361–373.

Stetler, C., Miller, G.E., 2011. Depression and hypothalamic-pituitary-adrenal activation: a quantitative summary of four decades of research. Psychosomatic Medicine 73 (2), 114–126.

Tennant, C., 2002. Life events, stress and depression: a review of recent findings. Australian and New Zealand Journal of Psychiatry 36 (2), 173–182.

Thase, M.E., Entsuah, R., Cantillon, M., Kornstein, S.G., 2005. Relative antidepressant efficacy of venlafaxine and SSRIs: sex-age interactions. Journal of Womens Health (Larchmt) 14 (7), 609–616.

Vieta, E., Martinez-de-Osaba, M.J., Colom, F., Martinez-Aran, A., Benabarre, A., Gasto, C., 1999. Enhanced corticotropin response to corticotropin-releasing hormone as a predictor of mania in euthymic bipolar patients. Psychologie Medicale 29, 971–978.

Walker, E., Diforio, D., 1997. Schizophrenia: a neural diathesis-stress model. Psychological Review 104 (4), 667–685.

Walker, E., Mittal, V., Tessner, K., 2008. Stress and the hypothalamic pituitary adrenal axis in the developmental course of schizophrenia. Annual Review of Clinical Psychology 4, 189–216.

Wisner, K.L., Sit, D.K., McShea, M.C., Rizzo, D.M., Zoretich, R.A., Hughes, C.L., Eng, H.F., Luther, J.F., Wisniewski, S.P., Constantino, M.L., Confer, A.L., Moses-Kolko, E.L., Famy, C.S., Hanusa, B.H., 2013. Onset timing, thoughts of self-harm, and diagnoses in postpartum women with screen-positive depression findings. JAMA Psychiatry 70 (5), 490–498.

Yan, H.C., Cao, X., Das, M., Zhu, X.H., Gao, T.M., 2010. Behavioral animal models of depression. Neuroscience Bulletin 4, 327–337.

Young, M.A., Scheftner, W.A., Fawcett, J., Klerman, G.L., 1990. Gender differences in the clinical features of unipolar major depressive disorder. Journal of Nervous and Mental Disease 178, 200–203.

Zarrouf, F.A., Artz, S., Griffith, J., Sirbu, C., Kommor, M., 2009. Testosterone and depression: systematic review and meta-analysis. Journal of Psychiatric Practice 15 (4), 289–305.

Endogeneity: Methodology

Trond Petersen, University of California, Berkeley, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 7, pp. 4511–4513, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Explanation and examples are given of the distinction between endogenous and exogenous variables. The concept of endogeneity arises when a distinction is made between a dependent variable, to be explained, and independent variables, used as explanatory factors or inputs. A variable is often labeled exogenous or independent if it is used to explain variations in the endogenous or dependent variable. There are various forms of endogeneity, such as in recursive versus interdependent systems. The role of endogeneity in social science empirical research is discussed. Examples of recursive and interdependent systems are given.

Endogenous versus Exogenous Variables

In several fields in the social sciences a distinction is made between endogenous and exogenous variables. A simple example may clarify. In studies of social inequality researchers try to explain differences in the level of education that people achieve. Education can be measured, for example, as years of schooling or as less than high school, a high school degree, and so forth. The goal of social science analysis will then be to figure out how the level of education achieved varies with other characteristics, such as a person's social background, sex, race, and more. Social background can be measured in terms of parents' socioeconomic status or income. An early example of this kind of analysis in sociology is Blau and Duncan (1967).

In this example, level of education is the endogenous variable. The variables used to explain variations in the level of education are called exogenous. More generally, the variables that show differences we wish to explain are called endogenous, while the variables used to explain the differences are called exogenous. Often this goes along with a causal imagery. Exogenous variables are thought of as causes, endogenous as their effects. But there is no necessary connection; one may use the terms without implying causality (Sobel, 1995). A simpler terminology for exogenous and endogenous is independent and dependent variables.

This distinction between exogenous and endogenous variables is used almost universally in quantitative social science. It arises in statistical data analysis, as in large parts of sociology, political science, and psychology. It arises also in mathematical analysis of theoretical ideas, as in large parts of economics and some parts of political science. But the distinction clearly is of relevance also in more qualitative research, such as in fieldwork and historical research, though less often used explicitly in those contexts. The distinction is simply relevant whenever there is some phenomenon that the researcher is trying to explain, which then is the endogenous variable, and there are some explanatory factors that are used to explain the phenomenon, which then are the exogenous variables. For example, in Skocpol's (1979) study of social revolutions, two exogenous variables are the state and agrarian structures in a society, while the endogenous is whether a revolution occurs or not.

Recursive Systems: Causal Chains

The example above can be elaborated. Consider what happens once people have achieved a particular level or type of education. They then often enter the labor market hoping to find a job. But the kinds of jobs people can expect to get depend strongly on their education. Psychologists are not hired to build bridges and engineers are rarely employed to provide mental-health services to children in despair. So in a next step, in explaining the kinds of jobs people find, education becomes an important independent or exogenous variable. It was endogenous, but is now an exogenous variable for explaining the type of job, which then is the endogenous variable. But the job may in turn become an important factor, thus exogenous, in explaining the salary received, which then is the endogenous variable.

And so it may continue. A variable, which once was endogenous, may be exogenous in another setting. There is a technical term for such reasoning, not important in itself, but sometimes encountered (from Bentzel and Wold, 1946; see Hausman, 1998). The variables entering into the reasoning are said to constitute a 'recursive system': parents' socioeconomic status influences education, education influences the type of job, the type of job influences earnings and career opportunities, and so forth. In causal imagery, they make up a causal chain. And it may also be that parents' socioeconomic status not only influences education, but also the job their offspring gets, for example through the network ties parents may have to employers or others who make hiring decisions. So an exogenous variable may be exogenous not only to one of the endogenous variables but to several of them.

For a set of variables to constitute a recursive system, there can be no feedback effects between them. Education may influence the type of job. But the type of job can have no influence on education itself. That such is the case is almost always tautologically true, provided the problem is carefully framed. The level of education is determined chronologically prior to the type of job, so there can be no feedback effect. This does not rule out that expectations and desires about future jobs will influence the education sought. But expectations about the future are not the same as the future. Enough of us have desired to be glamorous rock stars, outstanding violinists,

famous war generals, or influential academics. Some may even have sought educational tracks correspondingly. A few succeed, most end up other places than aimed for. But whatever education and training we acquire, it usually has some impact on where we end up, though not always enabling us to get where we aspired to.

Statistical versus Mathematical Analysis

Often the analysis of the relationships between variables is statistical. Researchers then typically use multivariate statistical methods. One tries to assess the joint influence of, say, sex and race on earnings. The goal is to see how the average level of earnings varies with sex, race, and so forth, not to explain the outcome for each individual. Similarly, if the endogenous variable is the type of job, say blue-collar, white-collar, professional, etc., then the goal of the analysis is to figure out how the percentages falling into each of these groups vary with education, sex, race, etc. Statistical social science analysis is eminently Holmesian, taking its cue from Sherlock Holmes. He argued that "while the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician" (Doyle, 1986: 175). And though few quantitative social scientists would accept the statement that aggregate relationships are a mathematical certainty or that we can successfully predict aggregate behaviors, the gist of Holmes' comment is correct and agreed upon: We investigate what on average goes on.

If the analysis is purely mathematical, where one tries to explore theoretical ideas by means of mathematical tools, then the writer traces out the logical implications of the relationships between the variables. For instance, from the knowledge of how producers and consumers act and interact in modern economies, economists will specify several mathematical equations that try to capture elements of these actions and interactions. A goal can be to investigate, using the mathematical equations as an aid for reasoning, what the impact of an increase in the interest rate will be for consumption, production, and finally the national product. For example, increasing the interest rate may lead to a decline in home sales, leading to a decline in building of new homes, with correspondingly increased unemployment in the home construction sector, and so forth.

Interdependent Systems: Mutual Dependencies

In some settings two or more variables may mutually influence each other, sometimes referred to as being interdependently related, constituting an interdependent system. For example, the amount of schooling a person obtains may depend on her family behavior: whether she is married or not, and the presence or absence of children. Those who marry early and get children may decide not to pursue lengthy education. But vice versa, a person's family behavior may depend on her educational behavior, whether she is in school or not (e.g., Marini, 1984). Being in school may delay family formation and getting children. In this example, family status is endogenous to schooling and schooling is similarly endogenous to family status. They mutually influence each other.

In the social sciences, many attempts have been made to assess the strengths and even possible directions of such mutual influences. These vary greatly in how successful they are.

A number of approaches have attempted to use cross-sectional data, with information about variables at one point in time, to determine the strength and direction of causation between them. One may obtain information on the amount of schooling and family status of say people who are 30 years old. One then tries, using advanced statistical techniques, to disentangle how the two mutually depend on each other. Such attempts suffer from the drawback that they do not take into account the sequencing of and the temporal relationships between the variables. They can hence not be considered particularly successful (see Alexander and Reilly, 1981).

An example may help. A woman may get pregnant at 18 years of age, after finishing high school. This may lead to a shorter subsequent education. But the shorter subsequent education did not influence the event of giving birth at 18 years of age. The expectation or intent of not obtaining additional education may have influenced whether to go through with a pregnancy at 18 years of age, but as noted above, an expectation about the future is not the same as the future. The shorter subsequent education may, however, shorten the time before the person has a second child, which in turn may delay further education. At 30 years of age, we may observe only the family status at that age, the number of children, and the education obtained so far. But unless one also collected data on the sequence of events as they unfolded over the life cycle, one cannot determine how childbirth influenced education and vice versa how education delayed or led to additional childbirths. Without a temporal relationship between values on two or more processes, mutual dependencies are hard if not impossible to assess (Petersen, 1995).

It is therefore widely recognized that such interdependencies are best studied by means of dynamic models using over-time data, where the sequence and types of transitions are studied as they occur over the life cycle. This was already pointed out by Lazarsfeld (1948).

In the context of the present example, with over-time data on several individuals, one can investigate how family status at 18 years of age influences schooling status at 19 years of age, and vice versa how schooling status at 18 years of age influences family status at 19 years of age. Suppose that family status has no impact on schooling status, so that marital status does not influence whether one goes to school or not in the next year. But suppose that schooling status has an effect on family status, so that it does matter whether you are in school or not for whether you get married or not in the next year. One would then say that schooling is an exogenous variable relative to family status. This is the same as saying that family status is endogenous to schooling status. If they both impact each other, then both are endogenous to each other. Having figured out what happens at 19 years of age, one can continue with analyzing the mutual influences of family and schooling status at 19 years of age on the same variables at 20 years of age, and so forth. One can sequentially unravel how the relationships between the variables evolve over time. Bentzel and Hansen

(1954) argued that mutual dependencies between two or more variables should be studied in such a manner.

There is an extensive statistical literature that addresses how to investigate which variables mutually influence each other, requiring access to over-time data as described above (Engle et al., 1983; Bergstrom, 1984; Petersen, 1995). Probably the best way to advance knowledge about which variables are endogenous and which are not is to collect careful data on sequences of event or variables (e.g., Wooldridge, 2010; Allison, 2013). No statistical technique can compensate for deficient data.

But no matter the kind of approach used or even favored, quantitative or more qualitative, the distinction between endogenous and exogenous variables is essential. Usually, there is some phenomenon in which we try to explain variations: the endogenous variable. Usually, there are some factors used to explain those variations: the exogenous variables.

See also: Causal Inference: Overview; Comparative Studies: Method and Design; Control Variables in Research; Experimental Design: Bayesian Designs; Experimental Design: Overview; Experimental Design: Randomization and Social Experiments; External Validity; Internal Validity; Latent Structure and Causal Variables; Longitudinal Causal Inference; Longitudinal Data: Event–History Analysis in Discrete Time; Mechanism-Based Causal Analysis; Moderator Variable: Methodology; Nonequivalent Group Designs; Observational Studies: Overview; Panel Surveys: Uses and Applications; Quasi-Experimental Designs; Regression Discontinuity Design; Selection Bias, Statistics of; Social Experiments.

Bibliography

Allison, P.D., 2013. Event History Analysis. Sage, Thousand Oaks.

Alexander, K.L., Reilly, T.W., 1981. Estimating the effects of marriage timing on educational attainment: some procedural issues and substantive clarifications. American Journal of Sociology 87 (1), 143–156.

Bentzel, R., Hansen, B., 1954. On recursiveness and interdependency in economic models. Review of Economic Studies 22 (3), 153–168.

Bentzel, R., Wold, H., 1946. Statistical demand analysis from the viewpoint of simultaneous equations. Scandinavian Actuarial Journal 29, 95–114.

Bergstrom, A.R., 1984. Continuous time stochastic models and issues of aggregation over time. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. II. North-Holland, Amsterdam, pp. 1145–1212.

Blau, P.M., Duncan, O.D., 1967. The American Occupational Structure. Wiley, New York.

Doyle, A.C., 1986. Sherlock Holmes: The Complete Novels and Stories, vols. 1 and 2. Bantam Books, New York.

Engle, R.F., Hendry, D.F., Richard, J.-F., 1983. Exogeneity. Econometrica 51 (2), 277–304.

Hausman, D.M., 1998. Causal Asymmetries. Cambridge University Press, New York.

Lazarsfeld, P.F., 1948. The use of panels in social research. Proceedings of the American Philosophical Society 92 (5), 406–410.

Marini, M.M., 1984. Women's educational attainment and the timing of entry into parenthood. American Sociological Review 49 (4), 491–511.

Petersen, T., 1995. Models for interdependent event history data: specification and estimation. Sociological Methodology 25, 317–375 (In: Peter V. Marsden (Ed.). Basil Blackwell, for the American Sociological Association Oxford, UK).

Skocpol, T., 1979. States and Revolutions. Cambridge University Press, New York.

Sobel, M.E., 1995. Causal inference in the social and behavioral sciences. In: Arminger, G., Clogg, C.C., Sobel, M.E. (Eds.), Handbook of Statistical Modeling for the Social and Behavioral Sciences. Plenum Press, New York, pp. 1–38 (Chapter 1).

Tuma, N.B., Hannan, M.T., 1984. Social Dynamics. Models and Methods. Academic Press, Orlando, FL.

Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data. MIT press, Cambridge, MA.

Energy and Society

Kirby Calvert, The Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The energy–society relationship is coevolutionary. On one hand, the material properties of energy inputs influence the structure and population size of a society and establish the parameters around, and expectations for, social and economic life. On the other hand, a physical entity is repositioned as an 'energy resource' only under specific technological and social conditions, and the evolution of energy systems represents a set of social choices shaped by political–economic (power) relations, cultural norms, and previous social decisions. The transition toward sustainable energy is as much a cultural and political project as it is a technological and economic one, and will (require societies to) reconfigure and reimagine sociotechnical structures and society–environment relations.

Introduction

The energy–society relationship is best described as coevolutionary and coproductive. On one hand, energy – i.e., the 'ability to do work' or 'capacity to influence the state of a system' (Smil, 2008) – is requisite for all human activities. Energy is derived from three primary sources: nuclear fusion reactions and subsequent radiant energy from the sun, gravitational force, and geothermal activity. These undergo a variety of transformations on Earth so that energy is also available in many secondary forms – e.g., stored in biomass and fossil fuels; flowing in water and wind (*see* Energy Flow Analysis). The quantity, quality, form, and spatiotemporal distribution of energy inputs into society establishes limitations on the population size and complexity of social systems and shapes the way in which institutions, conversion devices, and built infrastructure – or *sociotechnical metabolic pathways* – are organized in order to transfer energy between ecological and social systems (Luten, 1971; Cook, 1976; Smil, 2008). On the other hand, the types and quantity of energy used by a social system are structured by social norms, values, beliefs, knowledge, and previous social choices embedded in prevailing systems of energy provision (e.g., built infrastructure). Collective and individual energy choices are the means to achieve a certain kind of society and lifestyle, while social and technical conditions form the criteria by which specific energy resources are chosen and shape the way in which energy systems are incorporated into ongoing and new social functions (Lutzenhiser, 1993; Sovacool and Brown, 2007; van der Kroon et al., 2013).

In order to balance these 'energy deterministic' and 'socially constructed' perspectives on the energy–society relationship, this discussion is organized into two sections, which review well-established themes prominent in the academic literature. The first section explores the 'societal energetics' theme. Here, society is conceptualized as a "physical input-output system drawing material and energy from its environment, maintaining internal physical processes and dissipating wastes, emissions, and low-quality energy to the environment" (Haberl, 2006: 87). This section explores the relationship between social structure and energy use, as well as how new energy inputs (re)configure sociotechnical metabolic pathways. The second section explores the 'social construction of energy.' Here, the energy–society relationship is conceptualized as more than a physical process. This section explores the cultural and political economic processes that (re)position a physical entity as an energy input and, which shape sociotechnical metabolic pathways as well as desired energy services (end uses). As we unpack this relationship, we learn that the development of a sustainable energy system will include not only new resources and new technologies, but a new kind of society with fundamentally different attitudes toward, and ways of structuring and interacting with, the natural and built environment.

Societal Energetics

Humans are not *autotrophic* (self-nourishing) organisms and, like other *heterotrophs*, must redirect stocks and flows of energy in the environment toward basic physiological maintenance and growth. As such, available energy and more specifically edible biomass (i.e., stored chemical energy in plants and animals) is the primary physical mediator in the human–environment relationship (Cook, 1976). Energy provision is more efficient through collective rather than individual pursuits so that a social system might be conceived as a human strategy to extract energy from ecological systems (see Boehm, 1999). (A brief review of the first and second laws of thermodynamics brings precision to this notion: (1) energy available in a system is never created nor destroyed, only converted or transferred to another system; (2) all closed systems tend toward a state of increased entropy or thermodynamic equilibrium (less productive energy). Since energy is always available but degraded, it is more accurate to state that societies extract *productive energy*, otherwise known as *exergy* (Rosen, 2002), which is made available when high concentrations of pressure, temperature, or mass tend toward thermodynamic equilibrium (i.e., dissipate) within their environment. According to the 'maximum power principle,' those societies designed to dissipate these concentrations (maximize entropy) through productive social activities and physical infrastructure (embodied energy) are most likely to reproduce and grow (Odum and Odum, 1976; Rebane, 1995).) Energetic studies of social evolution reveal a correlation between the material properties of societal energy inputs and the size and structure of

a society. In other words, variations in energy inputs influence the organization of sociotechnical metabolic pathways and lead to very different kinds of societies in terms of size, structure, and energy services. For the sake of illustration, the following generalizes otherwise spatiotemporally diverse societies into three distinct regimes: hunter-gatherer societies, agrarian societies, and industrial societies (see Sieferle, 2001; Haberl, 2002; Goudsblom, 2012; Fischer-Kowalski et al., 2014).

Early hunter-gatherer societies relied primarily on recent photosynthetic activity as they foraged for edible plants and, when necessary, hunted animals to attain stored caloric energy (especially high-fat meat). The extent of daily and seasonal travel, as well as group population size, varied based on local spatiotemporal density of nutritional energy, which is typically higher in grassland biomes than in forest biomes. Societies living near, and supplementing terrestrial productivity with, marine ecosystems tended to be larger, more complex, and less nomadic. In fact, marine ecosystems were likely the first source of surplus energy. Salmon runs along the Pacific coast and Baleen whale migration in the Arctic Ocean are exemplary in this regard (Smil, 2008).

Hunting and gathering requires significant expenditures of energy, so that survival and growth is dependent on deliberate techniques to increase net energy gains. This was initially achieved in two ways. First, societies developed technologies and practices to store their food – e.g., freezing whale fat in underground permafrost shelters (Smil, 2008). Storage techniques liberated time spent hunting and gathering, allowing societies to spend their time in more interesting ways; a point we will return to later in the discussion. Second, humans physically preprocessed their food to improve the efficiency of the digestive process. Most notably, cracking, pounding, and grinding techniques were used to reduce the particle size of food intake and thereby expand the surface area exposed to digestive enzymes (Carmody and Wrangham, 2009).

These techniques notwithstanding, the energetic maturation of hunter-gather societies was marked by their ability to control fire for hunting and cooking. Cooking with heat further improves net energy gain from food by externalizing (using *exosomatic* energy for) the initiation of the digestive process, which (1) reduces the internal work (*endosomatic* energy) required for digestion and (2) improves the overall digestibility of useful energy from otherwise recalcitrant food (e.g., high-starch foods such as potatoes) (Carmody and Wrangham, 2009). Combined with the development and use of clay pots, fire also enabled the addition of soups to regular diets, which extracted nutritional value from otherwise inedible portions of a kill or a harvest. In addition, fire was used to direct large herds of animals to a more convenient location for hunting purposes and, later, smoke curing became a common way to preserve meat and fish where temperatures were not ideal for cold storage. More broadly speaking, the use of fire represents the first branch in fuel-mix diversity within social systems. Social groups now commanded two primary energy inputs – food (biomass with nutritional value for humans) and fuel (biomass with no nutritional value for humans) – and expanded the social functions requiring an energy input. To the latter point, the controlled burning of fire meant that humans no longer had to wait for the midday sun to find warmth and light. A controlled fire was also a focal point for social gathering, where ideas were shared and problems resolved. For these and other reasons, fire has in many ways come to symbolize the dawn of civilization (Goudsblom, 1992).

Generally speaking, hunter-gatherer societies are powered by, and structured around, 'uncontrolled solar energy' (Sieferle, 2001). Agrarian societies, in contrast, deliberately concentrate the flow of solar energy into edible biomass through two predominant systems: crop agriculture and pastoralism. Where precipitation and soil quality is sufficient, agrarian societies clear land of inedible phytomass (trees and grasses) to instead cultivate edible phytomass such as cereals, legumes, and tubers. While this activity reduces global net primary productivity (NPP) (because agroecosystems are less productive than the ecosystems they replace), it greatly increases useful energy density per unit area and the overall appropriation of local NPP by societies (Haberl, 2006). Where geophysical conditions are insufficient to support annual crop growth, such as in arid regions, solar energy is instead concentrated through pastoralist practices whereby the health and activity of large herds of animals is managed to ensure a steady conversion of inedible (lignocellulosic) biomass such as grasses into edible fat and proteins. Agrarian societies grew larger than foraging societies in large part because of higher yields of edible biomass; greater security of food supply; and the realization that increasing labor meant increasing returns on investment.

A sort of 'preindustrial' agrarian society began to emerge with organizational and technical innovations to agrarian energy production systems. An early example involves the concentration of otherwise diffuse sources of physical power in human and animal muscles – humans by force in slave nations or through coercion in feudal societies, and draught animals (horse, oxen) by domestication. These forces were multiplied efficiently using various combinations of harnesses, ropes, levers, pulleys, treadmills, and gears (Smil, 2008); technologies made increasingly available after the combined development of charcoal and controlled blast furnaces, which enabled, among other things, metalworking at large-scales. Preindustrial agrarian societies are also marked by greater fuel-mix diversity. Using what Smil (2010) calls new 'prime movers' – i.e., devices that capture and convert energy through a productive process – wind and water (*inaminate energy*) began to supplement human and animal muscles (*animate energy*). Windmills were used to pump water to clear or irrigate land and therefore expand the area that could be actively farmed. Turbines converted falling water into mechanical energy to process agricultural products (e.g., mill grain) at large scales. These new prime movers diversified energy inputs into societal functions, increased the overall energy appropriated from ecological systems for social use, and, in so doing, had profound impacts on where and how people (could) live (Pasqualetti, 2012).

The shift toward inanimate sources of energy initiated a trend toward the geographic separation of social activities. Agricultural communities expanded on flat and fertile soils producing surplus nutritional energy that was distributed to mill communities, which developed where kinetic energy in hydrological systems was concentrated at a waterfall. Trading hubs were established at calm natural bays along coastlines as well as the junction or mouth of major inland rivers, which also became strategic places. Specialization of social activities and

divergent attitudes toward space and time emerged in this new geoeconomy. Agricultural life was (is) very task-oriented and organized around seasonal variations of natural energy flows, while life in mill communities was (is) increasingly organized around clock time and shift work. Trading communities, through interaction with visitors (merchants), became increasingly cosmopolitan.

Wind and water energy systems also underpinned the emergence of more complex political economies and connected societies in new ways. The production of surplus goods using hydrological power combined with more extensive transport by sail were precursors to regional (and later global) markets. Otherwise distant social groups became increasingly connected through trade networks, and consumers began to value exotic goods by virtue of their foreign (locally scarce) nature. Furthermore, wind and water represent early 'energies of conquest' to the extent that they were the energetic basis of colonization (Nye, 2001). Indeed, early political maps were (re)drawn by those nations that managed their energy resources in a way that enabled the control of large populations over great distances, even in relatively small numbers; e.g., in the construction of naval infrastructure and weaponry. In many cases, the conquered society was forced into slavery, thereby reinvesting human energy (multiplication of smaller forces, discussed above) into the fitness and growth of the dominant society.

'Fossil energy' refers to subterranean concentrations of dead biomass in solid (coal), liquid (oil), and gaseous (natural gas) forms. These energy resources provided the quantity and form of inanimate energy that is requisite to an industrial society. The discovery of coal meant that manufacturing and metalworking activities could expand in size and spatial distribution even as forests were massively depleted of their wood. New prime movers were significant here as well: the invention of the steam engine highlights the mutually reinforcing nature of energy and technology from which industrial societies emerge and are maintained. Coal-powered steam engines slowly replaced falling water as the primary source of mechanical power. Since coal is a dense stock of energy that can be transported in bulk, manufacturing output became more reliable because it was no longer negatively impacted by periods of drought. Moreover, production centers could now be colocated with trade centers to maximize market exposure; indeed, the transition toward coal explains why many trading hubs became manufacturing centers (e.g., Pittsburgh, Pennsylvania) (Nye, 2001). (Prior to the introduction of coal and steam engines into manufacturing, production centers and trade centers were dislocated from each other because the river systems used to generate power (narrow, powerful) were not the same river systems that could be used for navigation (wide, calm).) Meanwhile, fossil energy systems allowed productive activities, and even entire cities, to expand in size and complexity even as they became increasingly disconnected from primary energy sources.

All inanimate energy systems – fossil energy systems chief among them – are more capital and less labor intensive. This underpins a number of profound transformations in social organization and the way societies perceive and behave toward their environment. A pipeline system that carries cooking gas to homes changes the social and household roles of women and children who traditionally gathered fuelwood (van der Kroon et al., 2013). A winter storm, once dreadful or inconvenient, can be perceived as delightful and oddly comforting with access to a natural gas furnace to keep warm. Capital intensive energy systems simultaneously reduce strain on muscles; increase leisure time; and contribute to periods of energy surplus. This combination has increased 'disposable' energy, which can be invested into nonproductive uses, some of which become new priorities and new economies – e.g., spending time and energy reducing one's golf handicap. As a corollary, the efficiency with which societies metabolize food decreases. So where high-fat foods were once an energetic necessity they are in some societies an energetic problem manifested as an obesity endemic. What's more, the capital intensive nature of modern energy systems has underpinned unprecedented social movements, especially within high-energy subgroups. Where endurance running to chase animals to exhaustion was once an investment of stored nutritional energy to improve hunting success (Smil, 2008: 139), many people now jog as a hobby or to maintain a body image. Where seasonal disruptions to agricultural output were once compensated with animal meat, diets can maintain diversity of vegetarian foods shipped over long distances in shorter time spans within refrigerated transport mediums. Access to cheap energy has significantly increased buying power, leading to pleasure shopping (high consumerism). The development of the internal combustion engine and modern transport systems based on liquid fuels underpins the desire for low-density living in North America, in contrast to high-density living within societies that developed with access to less flexible energy resources and transport systems (Pasqualetti and Brown, 2014).

The unique qualities of fossil energy resources have magnified dislocation between production and consumption. Because they are unevenly distributed and can be easily transported, societies are increasingly reliant on external rather than internal (domestic) sources of energy. This has social implications across scales. At the international scale, the uneven distribution of fossil energy resources strongly mediates the ways in which societies interact, underpinning 'resource wars' and at the same time reinforcing the strategic value of (international) sociopolitical institutions that facilitate amicable relations between producer and consumer nations such as treaties, trade agreements, and international agencies. Access to a natural endowment of energy resources is indeed no longer sufficient to explain the energetic evolution of a given society; Switzerland and Japan offer a case-in-point (Luten, 1971). At smaller scales, energy users have become disconnected from the direct environmental impacts of their energy choices, as their interaction with the actual ecological impacts of energy production is mediated through light switches, thermostats, and outlets. Because fossil energy resources are subterranean and higher in energy density than any other energy resource, the correlation between rising energy consumption and cultivated or deforested land area is drastically weakened. As such, a city can stay warm during winter while choosing to designate the surrounding forested area as a preservation area; the forest, once a source of fuelwood, can now be used for recreational purposes.

The development of electricity was also transformative, and especially to cities. As Nasaw (1992) explains, new kinds of

public spaces were made possible with the control of electric current. Electric lighting made possible imaginations of the 'city as a spectacle' through a welcoming blaze of 'big city lights,' and street lighting in suburban neighborhoods enabled a stronger sense of security, which encouraged 'pleasure walking.' Private space changed as well, increasingly decorated with bright rather than dark colors since the soot residue from gas lighting was no longer a concern (Nye, 2001). Electric motors could operate refrigerant technology, enabling localized cold storage while changing expectations around thermal comfort in the summer. Economies of scale in fossil energy extraction and hydrodam construction along with technological advances that control electric currents at the microscale in information and communication technology innovations have given rise to telecommuting (i.e., blurring the lines between work space and home space) and a culture of 24/7 connectivity or 'always-on-ness' (Spinney et al., 2012); new features of what might be called the 'information society.'

In sum, the process of reinvesting surplus energy into making land and industrial activity more productive is, for many theorists, the process through which societies are able to increase in size and complexity. As societies transition from labor to capital intensive energy systems they undergo profound changes to the way in which built and natural environment is organized, perceived, and engaged. Our prevailing energy infrastructure, land-use patterns, environmental ethics, and expectations and priorities related to energy use have been shaped by the use of dense, subterranean, flexible, and transportable stocks of fossilized energy. But we must be careful to avoid a sense of energy determinism here; although energy is the 'mother' and technology the 'father' of social growth and complexity and human–environment relations, social values are the 'midwife' (Rosa et al., 1988: 152).

The Social Construction of Energy

While energetic theories of society emphasize how forms of energy – as ability to do work – shape sociotechnical metabolic structures and social practices, social theories of energy focus on the social work (political, economic, cultural) that is done to turn a physical entity into an energy resource, as well as the social factors, which prioritize how physical work is expended and who benefits in the process of energy production, distribution, and use. From this perspective, the ways in which sociotechnical metabolic pathways are organized, and the types and quantity of energy that is used, involves a set of social choices. Stated differently, resources have no power outside of particular social and technical conditions. In addition to undermining energy determinism, this perspective challenges the 'energy ladder' model, which posits that energy choices are based on the energetic value of a given resource relative to the cost of acquisition. Surely, there is a noticeable trend within and between societies to convert energy into more useful forms – e.g., grass to meat; fuel into heat or electricity – as efficiently as possible. But economic choices are socially structured, meaning that noneconomic factors are important in understanding energy choices (Lutzenhiser, 1993). Indeed, deviations from the energy ladder model are prevalent. Despite access to gas and electricity, for example, wood and peat are still burned (for nostalgic reasons) in modern households and wood is in some cases preferred as a cooking fuel for a variety of traditional dishes (Taylor et al., 2011; van der Kroon et al., 2013). Under certain conditions raising animals for food results in a net energy loss, but as Smil (2008: 152) asks, how else could the provision of 'tasty' combinations of fat and protein be controlled? In other words, society does not follow a natural preordained order toward higher value fuels or more efficient energy systems, nor is energy exploited simply as a function of its existence. Energy exploitation, rather, is rooted in the cultural, political, and economic trajectories whereby (1) social meaning is attached to a physical entity, as well as a particular use of that entity; and (2) the skills, institutions, and built infrastructure through which an energy resource can be used (in particular ways) is established.

Not all individuals or cultures see the world the same way nor do all societies share the same goals. As such, different social groups will not equally rank the attributes of specific energy systems and the services that they provide. Firstly, since a decision to extract energy is necessarily a decision to alter the environment in some way, energy exploitation is directly related to social perceptions of and ethical positions toward the nonhuman world. As Nye (2001: 249) explains, the desire among early European settlers to make land more productive through intensive management was not shared by indigenous populations; this cultural difference rather than inherent technological ingenuity or access to coal underpinned European status as a 'high energy society' and dominance of the new world. Related to this point, Bell et al. (2005) identify a national local 'social gap' problem in the context of renewable energy development, whereby national level polls show strong support for renewable energy due to concerns related to climate change, while many local citizens are upset by, among other things, the way in which renewable energy infrastructure spoils the aesthetic qualities of local landscapes. Secondly, energy is requisite to all social activity and is therefore a means to a certain kind of society (Sovacool and Brown, 2007), so that energy production and distribution is wrapped up in ideas about social equality and social development. Although concentrated human labor certainly has energetic benefits, the misguided idea of inherent racial inferiority justified slavery toward this end; it was not some energetic imperative or a natural extension of having tapped the 'energies of conquest' (Nye, 2001). Huber (2013) demonstrates that (particular uses of) energy resources are intimately wrapped up in notions about life, liberty, and social organization; e.g., just-in-time delivery of goods, the single-family dwelling, and expectations around personal transport. Lizot (1977) suggests that energy demand is related to a cultural attitude toward work and ideas about what constitutes 'progress.' This is based on the observation that one group in close proximity to an animal-rich forest consumed less than half as much food energy and protein as a neighboring group, the former preferring instead to spend more time in leisure. Overall, these and other sociocultural factors help to explain why energy demands and energy systems between households and nations vary in space and time, even in the case where all other factors are equal (e.g., climate, resource endowments, technology, demographics).

In the same way that specific energy systems shape social and political movements, new social and political movements play

a role in reconfiguring energy systems. Those surrounding environmental values including localization of food diets and energy sustainability, which are driving investments into distributed renewable energy systems, as well as political movements aimed at democratic freedom or state control in the Middle East, which disrupt global supply chains of fossil fuels, are exemplary in this regard. At the same time, energy choices are both constrained and enabled by past social choices, which underpin prevailing institutional and technical contexts. Nye (2001) suggests that an industrial revolution could emerge only in the context of an *industrious* revolution among certain (members of) societies, whereby more members of the household were driven toward paid labor by commercial incentives, and consumer tastes were enhanced due to rising purchasing power. Another example comes from attempts at large-scale rural electrification programs in developing nations, many of which failed because rural citizens in these countries could not afford the appliances necessary to utilize electricity, nor had they developed a set of cultural practices that necessitated or valued electricity consumption in the first place (Taylor, 2005).

Shove and Walker (2010) point out that energy might best be conceived as one ingredient in an individual 'performance,' which affirms or resists many of the sociocultural conventions discussed above. They take showering as an example: ideas about 'freshness' and how it relates to social acceptability or social status, combined with the interaction of material elements such as shampoo and indoor plumbing at spaces dedicated for bathing, perpetuate energy-intensive hygiene. From this perspective, energy use is as much involved in a struggle for social fitness and legitimacy as it is a matter of economic efficiency. The diffusion of particular energy resources or technologies, then, is the diffusion of 'multiple, multisited' social practices (Nicholls and Strengers, 2014). What's more, the process of normalization of these practices, as well as the resources and systems through which they are supplied, is rapid. The peculiar affordances of fossil energy systems, for example, ranging from greater convenience in basic provisions (e.g., high-yield food production; internal room temperature of 20 °C) to expanded access to luxury (e.g., timely exotic travel and year-round vegetarian diets), are perceived by many as imperative to social and economic life even though these features are unique to what Shaffer (2009) calls 'hydrocarbon man,' an era equivalent to only 0.01% of humanity's 'terrestrial tenure' (Cook, 1976). This coembedded and mutually reinforcing relationship between social institutions and individual practices with energy systems means that deconstructing prevailing fuel-mixes, sociotechnical metabolic pathways, and use profiles is a sluggish process (Unruh, 2000). This is true despite potential to achieve gains in energy efficiency; e.g., uptake of agricultural practices was delayed in places where foraging yields were sufficient to satisfy lifestyle demands (Smil, 2008) and slave labor systems delayed uptake of steam power in the southern U.S. agricultural systems despite clear signs that steam power was beginning to rearrange the competitive landscape (Nye, 2001). Currently, although renewable energy is socially accepted in principal prevailing livelihood strategies, built and natural environments, and vested political economic interests (e.g., social status) that are wrapped up into prevailing energy systems are recalcitrant. Furthermore, there are clear tensions between proconservation values and the desire for convenience, comfort, and social status that comes with energy consumption (Lutzenhiser, 1993). These inconsistencies underpin the notion that renewable energy is unlikely to support the 'modern lifestyle' (Bryce, 2010; Trainer, 2013). Since the modern lifestyle is unlikely to be voluntarily redefined, the argument goes, renewable energy will never receive political support unless of course a geological imperative forces such a change. To the latter point, it is important to consider the fact that energy demands related to the construction of a 'modern lifestyle' in less developed countries are now driving the demand curve for energy, and have contributed to nearly 80% of the global demand growth for fossil resources in the last decade (Simon, 2005; IEA, 2012).

With all of this in mind, theories are divided on how to encourage societal change in terms of energy production and consumption. Three approaches are common and are most effective when targeted or creatively combined: (1) appeals to individual costs and benefits; (2) appeals to intrinsic values, as when renewable energy is connected to notions of 'energy democracy' and 'the health of the community/public'; or (3) by deconstructing prevailing norms and their supportive infrastructure and institutions and replacing them with new ones, such as redefining what thermal 'comfort' means (Rosa et al., 1988; Moloney et al., 2010; Nicholls and Strengers, 2014). Coercion through command-and-control (tax) or market mechanisms (pricing) is also an option. Responses to price signals are differential based on willingness and ability to pay. This approach also comes with significant ethical baggage because it restricts access to what is perceived to be, or at least rhetorically framed as, a social imperative – i.e., cheap (fossil) energy – and it will certainly disproportionately burden low-income groups. Another solution is to focus on improvements to the efficiency of energy production, distribution, and consumption technologies, which can, in theory, dramatically reduce energy demands without a change in lifestyle. Critics of this approach are quick to invoke Jevons' Paradox (see Sorman and Giampietro, 2013), which posits that improvements in energy efficiency, by themselves, do not lead to absolute reductions of energy use because people tend to enroll surplus energy into new services (e.g., outdoor garden lighting) or into enhancing existing ones (e.g., when heating systems are made more efficient, internal temperature is increased).

Conclusion

Energy is a physical entity that underpins all societal functions. The quality, quantity, form, spatial distribution and timing of energy availability, and the sociotechnical metabolic pathways through which they are recovered sets parameters around what kinds of societies are possible in terms of structure (size, complexity, political economic organization) as well as culture (livelihoods and lifestyles). To this point, terms such as 'hydrocarbon man' and 'fossil capitalism' are deliberatively provocative and meant to elicit the notion that fossil fuels are embodied at the individual and the collective level. But this embodiment is not inevitable, nor the result of a natural or preordained progression along an 'energy ladder.' The types of physical entities that become an energy resource, the way in which systems of energy provision are organized, and the use

profiles toward which they are oriented are all social network effects. In other words, they are material expressions of a dynamic interplay between common collective goals, accepted forms of interpersonal interaction, identity formation, power imbalances, knowledge, attitudes toward the nonhuman world, and prior social decisions. As such, societies and energy systems are coproductive, and they evolve together in nonlinear and spatiotemporally heterogeneous ways.

Growing concerns around current systems of energy provision related to resource availability and environmental impact, combined with key technological advancements, have redirected policy and industry attention toward unconventional fossil fuels (e.g., shale energy and oil sands) as well as renewable energy resources. What we have learned above is that these energy transitions are not natural progressions, nor are they simply a change in fuel-mix. They are cultural as well as political projects, which ultimately (require societies to) reconfigure sociotechnical structures, society–environment relations, everyday routines, and the social institutions and norms by which all of these are structured. Debates about our collective energy future are not only conversations about how to redesign infrastructure or economic systems, or about which technologies and resources are best; they are also about what kind of a society we envision and our individual roles therein.

See also: Energy Flow Analysis; Fire, Human Use, and Consequences; Oil Age: Energy in Transition.

Bibliography

Bell, D., Gray, T., Haggett, C., 2005. The 'social gap' in wind farm siting decisions: explanations and policy responses. Environmental Politics 14, 460–477.

Boehm, C., 1999. Hierarchy in the Forest: The Evolution of Egalitarian Behavior. Harvard University Press, Cambridge.

Bryce, J., 2010. Power Hungry: The Myths of Green Energy and the Real Fuels of the Future. PublicAffairs, New York.

Carmody, R.N., Wrangham, R.W., 2009. The energetic significance of cooking. Evolutionary Anthropology 57, 379–391.

Cook, E.F., 1976. Man, Energy Society. Freeman and Co., San Francisco.

Cottrell, F., 1955. Energy and Society. McGraw-Hill, New York.

Fischer-Kowalski, M., Krausmann, F., Pallua, I., 2014. A socio-metabolic reading of the Anthropocene: modes of subsistence, population size, and human impact on Earth. Anthropocene Review 1 (1), 6–31.

Goudsblom, J., 2012. Energy and civilization. International Review of Sociology 22 (3), 405–411.

Goudsblom, J., 1992. Fire and Civilization. Penguin, London.

Haberl, H., 2002. The energetic metabolism of societies part II: empirical examples. Journal of Industrial Ecology 5 (2), 71–88.

Haberl, H., 2006. The global socioeconomic energetic metabolism as a sustainability problem. Energy 31, 87–99.

Huber, M., 2013. Lifeblood: Oil, Freedom, and the Forces of Capital. University of Minnesota Press, Minneapolis.

International Energy Agency, 2012. World Energy Outlook 2012: How Will Global Energy Markets Evolve to 2035? (Paris, France). http://www.worldenergyoutlook.org/media/weowebsite/2012/factsheets.pdf (accessed 20.03.13.).

van der Kroon, B., Brouwer, R., van Beukering, P.J.H., 2013. The energy ladder: theoretical myth or empirical truth? Results from a meta-analysis. Renewable and Sustainable Energy Reviews 20, 504–513.

Lizot, J., 1977. Population, warfare, and resources among the Yanomami. Man 12, 497–517.

Luten, D.R., 1971. The economic geography of energy. The Scientific American 224, 164–175.

Lutzenhiser, L., 1993. Social and behavioural aspects of energy use. Annual Review of Energy and Environment 18, 247–289.

Moloney, S., Horne, R.E., Fien, J., 2010. Transitioning to low carbon communities –, from behaviour change to systemic change: lessons from Australia. Energy Policy 38, 7614–7623.

Nasaw, D., 1992. Cities of light, landscapes of pleasure. In: Ward, D., Zunz, O. (Eds.), The Landscape of Modernity: Essays on New York City, 1900–1940. Russel Sage Foundation, New York, pp. 273–286.

Nicholls, L., Strengers, Y., 2014. Air conditioning and antibiotics: demand management insights from problematic health and household cooling practices. Energy Policy 67, 673–681.

Nye, D.E., 2001. Consuming Power: A Social History of American Energies. MIT Press, Cambridge, MA.

Odum, H.T., Odum, E.C., 1976. Energy basis for Man and Nature. McGraw Hill, New York.

Pasqualetti, M.J., 2012. Reading the changing energy landscape. In: Stremke, S., van den Dobbelsteen, A. (Eds.), Sustainable Energy Landscapes: Designing, Planning and Development. CRC Press, Baton Rouge, FL, pp. 11–44.

Pasqualetti, M.J., Brown, M.A., 2014. Ancient discipline, modern concern: geographers in the field of energy and society. Energy Research & Social Science 1, 122–133.

Rebane, K.K., 1995. Energy, entropy, environment: why is protection of the environment objectively difficult? Ecological Economics 13, 89–92.

Rosa, E.A., Machlis, G.E., Keating, K.M., 1988. Energy and society. Annual Review of Sociology 14, 149–172.

Rosen, M.A., 2002. Energy crisis or exergy crisis? Exergy, an International Journal 2, 125–127.

Shaffer, B., 2009. Energy Politics. University of Pennsylvania Press, Philadelphia, PA.

Shove, E., Walker, G., 2010. Governing transitions in the sustainability of everyday life. Research Policy 39, 471–476.

Sieferle, R., 2001. The Subterranean Forest: Energy Systems and the Industrial Revolution. White Horse Press, Cambridge.

Simon, C.A., 2005. Alternative Energy: Political, Economic and Social Feasibility. Rowman and Littlefield Publishers, Lanham, MD.

Smil, V., 2008. Energy in Nature and Society: General Energetics of Complex Systems. MIT Press, Cambridge.

Smil, V., 2010. Energy Transitions: History, Requirements and Prospects. Praeger, Toronto, ON.

Sorman, A.H., Giampietro, M., 2013. The energetic metabolism of societies and the degrowth paradigm: analyzing biophysical constraints and realities. Journal of Cleaner Production 38, 80–93.

Sovacool, B.K., Brown, M.A., 2007. Introduction – the compelling tangle of energy and American society. In: Sovacool, B.K., Brown, M.A. (Eds.), Energy and American Society – Thirteen Myths. Springer, Dordrecht, The Netherlands, pp. 1–22.

Spinney, J., Green, N., Burningham, K., Cooper, G., Uzzell, D., 2012. Are we sitting comfortably? domestic imaginaries, laptop practices, and energy use. Environment and Planning A 44, 2629–2645.

Taylor, M.J., 2005. Electrifying rural Guatemala: central policy and rural reality. Environment and Planning C: Government and Policy 23, 173–189.

Taylor, M.J., Moran-Taylor, M.J., Castellanos, E.J., Elias, S., 2011. Burning for sustainability: biomass energy, international migration, and the move to cleaner fuels and cookstoves in Guatemala. Annals of the Association of American Geographers 101, 918–928.

Trainer, T., 2013. Can the world run on renewable energy? A revised negative case. Humanomics 29, 88–104.

Unruh, G.C., 2000. Understanding carbon lock-in. Energy Policy 28, 817–830.

Energy, Anthropology of

Dominic Boyer, Rice University, Houston, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews anthropological research on energy dating back to the 1940s. Three generations of research are identified and explored: the first was Leslie White's pathbreaking effort in the 1940s and 1950s to rethink models of cultural evolution through the lens of energy use. Then, in the 1970s, a second generation of anthropological research took shape in response to the global oil shocks and the beginning of serious political debates over the need for transition to renewable energy resources. Those conversations were largely silenced by a political recommitment to carbon and nuclear fuel across the world in the 1980s. However, in the past several years, a third generation of anthropological research has rapidly developed, spurred by mounting evidence of anthropogenic climate change and its impacts of human and nonhuman life across the world.

As viewed from the perspective of rapidly accumulating anthropological research on energy impacts (e.g., Crate and Nuttall, 2009) and energy transition (e.g., Howe, 2011) across the world, the anthropology of energy would seem to be a research area of rather recent and specific vintage, a response above all to growing scientific and political attention to anthropogenic climate change. Yet energy is a topic that has long been addressed in social–cultural anthropology, even if largely outside of mainstream research trends. By the author's count, the lively research field that has formed in the past decade is no less than the third generation of the anthropology of energy. In this article, the author reviews past and present anthropological research on energic forces and forms and offer some guarded speculation about where the field will head in the coming decade.

Energy and Cultural Evolution

The anthropology of energy's first generation was defined foremost by the work of Leslie White, a maverick thinker who granted energy a prominent place in his efforts to resurrect evolutionary theory in American anthropology (1949, 1959). For White, energy was not just one research topic among others; it was the conceptual key to understanding everything about human life and history. In what is arguably his most influential work, '*Energy and the Evolution of Culture*,' he outlined a literally universal theory: "Everything in the universe may be described in terms of energy. Galaxies, stars, molecules and atoms may be regarded as organizations of energy. Living organisms may be looked upon as engines which operate by means of energy derived directly or indirectly from the sun. The civilizations or cultures of mankind, also, may be regarded as a form or organization of energy" (1943: p. 335). White's key equation was that "cultural development varies directly as the amount of energy per capita per year harnessed and put to work" (p. 338). Thus, at the beginning of cultural history, with only the energy of its bodies with which to operate, humanity's cultural development remained at a very low level. To satisfy needs and to improve capabilities, both physical and intellectual, humanity sought to harness new sources of energy, first in the domestication of animals and plants, increasing "tremendously the amount of energy per capita available for culture-building" (p. 343). But after several thousand years of steady improvement, humanity once again plateaued until the advancements of the eighteenth and nineteenth centuries allowed the widespread harnessing of fuel and the invention of engines. Fuel and engines were the technological-cum-sociological basis of modern civilization, White said, and the source of all its great cultural leaps forward from machinery to the arts.

Although in many respects a teleological narrative of cultural advancement, White also emphasized a dialectical process. The social system optimalized for a given energy regime typically resisted new technologies designed to harness even greater magnitudes of energy. So when "cultural advance" ceased under a given energy regime, "it can be renewed only by tapping some new source of energy and by harnessing it in sufficient magnitude to burst asunder the social system which binds it" (p. 348). Thus, the fuel regime exploded the social apparatus of the agricultural regime in the modernization of society, a story that political economy interpreted as the struggle between two political orders (capitalism and feudalism). Encrypted here is the kernel of an idea that had great, but mostly undeveloped, critical potential: the notion that modern capitalist society was a fuel society to its core; its achievements were fundamentally predicated on fuel consumption such that rampant consumption had become archetypal throughout its culture. It is possible that White did not pursue this avenue of cultural critique because he felt the fuel regime was nearing its end anyway. He wrote of peak oil and peak coal, of dwindling reserves of fuel for a world demanding more and more energy. A committed but closeted socialist, White doubtless found his energy theory a reassuring rationale for impending revolution. He invoked the Second Law of Thermodynamics frequently in his writings to establish a Manichean struggle between the dissipation and concentration of energy in the universe, history, and culture (e.g., "All life is a struggle for free energy," 1987: p. 118). Life being, in his view, the fundamental struggle to concentrate energy against entropy, it could not cease with the dimming of fuel. Instead, new energies would eventually be harnessed and new social systems and cultural advancements would accrete around them.

White's sense of an energic revolution in the making deserves some context. In the late 1930s, successful experiments in nuclear fission set off a great deal of scientific and popular speculation about what a possible several million-fold increase in the energy at humanity's disposal would mean for all aspects of social life (see, e.g., O'Neill, 1940). Then, just months before the publication of White's landmark 1943 article, Enrico Fermi and his Manhattan Project team accomplished, although in secret, the first self-sustaining nuclear chain reaction – the pioneering controlled operationalization of nuclear energy that would pave the way toward 'the atomic age.' White noted the terrific potentiality of nuclear energy, "To be able to harness sub-atomic energy would, without a doubt, create a civilization surpassing sober imagination of today" (1943: p. 351), although he ultimately seemed more attracted to the possibility that the sun would become "directly our chief source of power in the future" (1943: p. 351).

Although prescient in these and other ways, White probably was as much an obstacle to future anthropological research on energy as an inspiration. His politics and personality won him a few friends but many more enemies, enemies who succeeded in isolating him and minimizing his work for many productive years. Not only did his universalist model stand in sharp contract to the Boasian historicism and individualism that dominated American anthropology of the 1940s and 1950s, but also White compounded this contrast with a strident insistence that the evolutionist thinking of figures like Lewis Henry Morgan and Edward B. Tylor was more coherent and theoretically superior to the work of Boas. It was many years before some rapprochement could be found between White and the Boasians, by which time other, less controversial figures like Julian Steward, Marshall Sahlins, and Elman Service had become more central to the movement that had come to be known as 'cultural materialism,' pushing White's thermodynamic and energetic focus into the shadows of mainstream social–cultural anthropology.

Politics of Energy Development and the Culture of Energy Experts

There is thus less genealogy than one might expect between the first generation of anthropological interest in energy and the second generation, which emerged in the 1970s and 1980s. Some Whitean influence remained, particularly through the networks of Michigan anthropology (see, e.g., Adams, 1975; Rappaport, 1975) but there were concerns as to how to develop White's energy theory further. In his 1977 Presidential Address to the American Anthropological Association, Richard Adams recalled his tutelage under White and the problems that arose from importing physical and chemical conceptualizations of energy into social analysis, "Basically, we relied on the first and second laws of thermodynamics and followed the example of community ecologists. But the fact that the second law was applicable only to closed systems posed substantial barriers to its use in analyzing human societies. They were, after all, clearly open systems. I think the dilatory development of energy study in social anthropology is in part the result of an inadequate theoretical basis provided by physics and chemistry" (1978: p. 298). Adams did not view this problem as insurmountable, seeing great promise, for example, in the chemist Ilya Prigogine's "foundation for an understanding of thermodynamic open systems" (p. 302), but he noted that road ahead required a shift to the "holistic study of larger societies" (p. 301) for which, in his judgment, anthropological theory was as yet poorly equipped. Adams's address ended with a rousing and once again prescient call for anthropologists to take a more active role in the critical investigation of the magnitudes and sources of energy flow with an ultimate aim "to develop more direct dependence on solar energy and to reduce both the use of nonrenewable energy forms and the derangement of the natural ecological processes that construct the energy forms we need for our own nutrition and that of other species" (p. 307).

Yet, by and large, this call was not taken up immediately. Second-generation anthropologists of energy typically did stand in a critical relationship to the energy forms and norms of the developed world. But they were significantly less interested in advancing energetic and thermodynamic cultural theory than they were interested in highlighting the cultural and social impacts of energy development for indigenous peoples (Nordstrom et al., 1977) especially in terms of nuclear power (Robbins, 1980), uranium mining (Robbins, 1984), and oil extraction (Kruse et al., 1982; Jorgensen, 1990). The second generation thus drew energy into important wider debates over the rights of indigenous communities, environmental impacts and resource exploitation, debates that remain central features of the anthropology of energy today (e.g., Love and Garwood, 2011; Powell, 2010; Sawyer, 2004; Sawyer and Gomez, 2012). But, at the same time, much as Adams feared, the mainstream discipline's interest in the topic of energy remained 'dilatory.' The AAA flagship journal, *American Anthropologist*, for example, published scarcely any energy-related research after the mid-1970s. In part, this gap can be attributed to the 'applied' character of many second-generation projects that were designed foremost to help improve relationships among indigenous groups, corporations, and governments as regarded energy development (see, e.g., Jorgensen et al., 1978). But it also expressed the sense of theoretical impasse to which Adams had alluded as well as the progressive retreat from anthropological holism that was occasioned by growing subspecialization of anthropological research and the waning capacity of the subfields to communicate and collaborate effectively with each other.

An important exception to this trend was the work of Laura Nader, whose participation in the late 1970s in the US National Academy of Science's Committee on Nuclear and Alternative Energy Systems (CONAES) served as an inspiration to Adams as to what anthropologists might be capable of accomplishing (1978: p. 307). Through her research on behalf of CONAES, Nader became interested in the contribution of worldviews, both popular and scientific, to the rejection of ideas of energy conservation and energy transition (Nader and Beckerman, 1978). As the first anthropologist to research the energy policy community and energy scientists, Nader became both fascinated and deeply disturbed by "the culture of energy experts" (2004: p. 775). She saw energy policy as "grounded in fear of deleterious change in life-styles and options" despite considerable evidence of "the very wide range of choices of

life-styles that is available in any plausible energy future" (Nader, 2010: p. 241). Energy science meanwhile was plagued by an "inevitability syndrome" (2004: p. 775) that resisted models not predicated upon ever-increasing resource use and energy expenditure:

> Also striking was the omnipresent model of unilinear development (a concept that anthropologists had left in the dust decades earlier), with little general understanding of macro-processes. For example, the recognition that civilizations arise but that they also collapse was missing from the thinking about the present. Prevalent was the nineteenth century belief that technological progress was equivalent to social progress. In such a progressivist frame science too could only rise and not fall or wane. Furthermore, the possibility that experts might be part of the problem was novel to the expert who thought that he stood outside of the problem. (2004: p. 776)

Nader ended up laying much blame at the foot of the specialization and hierarchization of scientific communities in which 'standardization and conformity' ruled the day (ibid) and prevented much needed creative thinking about energy futures. Among anthropologists, Nader remains the only scholar to have researched the expert imagination of energy futures in such depth. Her work also paved the way for more recent ethnographies of energy experts (Mason and Stoilkova, 2012) and for political anthropologies of carbon (Coronil, 1997) and nuclear (Gusterson, 1996; Masco, 2006) statecraft.

If there is a lesson to be drawn from the timing of the first two generations of anthropological research to energy it is that they have accompanied vulnerable or transitional moments in dominant regimes of 'energopower' (Boyer, 2011). In White's case, his positioning of energy as the key to understanding all human culture (and indeed all existence), accompanied the nuclear energy revolution and its new magnitudes of creative and destructive power. In case of the second generation, the context was what Nader termed "the energy decade" of the 1970s. Beginning with the oil shocks of 1973 (that signaled the end of a certain phase of Western imperial control over carbon fuel resources, particularly in the Middle East) and followed by a short-lived political willingness to explore alternative energy sources, fissures and tremors spread in the energopolitical apparatus that attracted anthropological attention. The political recommitment to carbon and nuclear energy across the industrialized world in the 1980s blunted the aspirations and urgency of energy research in anthropology as elsewhere in the human sciences. Indeed, between the mid-1980s and the mid-2000s, anthropological research on energy seemed to go into a kind of hiatus (see, however, Dawson, 1992; Coronil, 1997), before displaying signs of renaissance (Henning, 2005; Love, 2008; Mason, 2007; Sawyer, 2004; Wilhite, 2005). Over the past several years, in response to growing public and political attention to questions of climate change and energy transition, a mounting body of intriguing case studies has begun to generate more profound theoretical challenges (e.g., Reyna and Behrends, 2008; Sawyer, 2007; Winther, 2008). Some 70 years after White's landmark paper, it seems as though energy has at long last become a subject worthy of serious attention in mainstream social–cultural anthropology.

Energy Anthropology in the Era of Anthropogenic Climate Change

This inevitably raises the question as to where the anthropology of energy is headed next. Since, in its third generation, the anthropology of energy is expanding and subspecializing rapidly, it is likely that it will have not one but several paths ahead of it. The author comments here briefly on three that seem particularly promising:

1. Critical political and environmental ethnography of carbon fuel and its impacts. As the demands of human civilization and the expansion of human population increasingly impact the planetary ecology, one finds anthropology sharpening its critical attention to carbon fuel economies and their social and environmental impacts (see, e.g., Behrends et al., 2011; McNeish and Logan, 2012). This literature aims at surfacing the political-economic importance of apparatuses of carbon fuel extraction and challenging common Northern models of resource management (in which constant growth of demand and supply is not only tolerated but actively encouraged). Scholars are investigating how projects of carbon fuel development can be at once a blessing and a curse to communities across the world, not only attracting new wealth but also cultivating forms of social exclusion and inequality alongside environmental toxicity. Although attentive to contemporary dimensions of resource demand and contention, this literature is deeply connected to earlier generations of energy anthropology, especially to the second generation's interest in the politics and impacts of energy extraction in indigenous communities and to White's inchoate critique of the social forms of fuel society.
2. Studies of the social formation of climatological and energic expertise. At the intersection of energy anthropology, the anthropology of knowledge, and anthropological science and technology studies, we find very vibrant conversations emerging over the social formation of climatological (e.g., Hastrup, 2012) and energic (e.g., Mason and Stoilkova, 2012) expertise and the processes through which forms of energy and environmental authority (or, 'ecoauthority,' see Howe, forthcoming) are constituted and defended. What animates this area of research is not, for example, skepticism regarding the facticity of climate change so much as recognition that theories of climate change, like other modes of expert knowledge, operate in complex and often competitive fields of epistemic authorization. Scholars in this research area thus argue we need to better understand how climatological and energic knowledges originate; in what kinds of expert communities, institutions, and practices; how and why some models thrive and others fail; and how expert models are received, challenged, and/or elaborated by other (e.g., political and economic) authorities and by publics. Like the resource and fuel studies noted above, this is a research area that is both highly innovative (especially in terms of its leveraging of analytic methods drawn from the anthropology of knowledge and STS) and also extensional of earlier research, particularly Nader's initial spadework on the culture of energy experts, in its effort to illuminate

dominant and emergent forms of expertise at the juncture of energy and the environment.

3. Critical reevaluation of anthropological models of political power through the lens of energic forces and infrastructures. Finally, in the past few years, it has seemed increasingly urgent to rethink, in a more fundamental way, anthropological models of political power through what the author has termed 'energopower,' power seen through the twin analytics of electricity and fuel (Boyer, 2011, forthcoming; Rogers, forthcoming). Although this line of thinking certainly is connected to White's, it does not seek to simply import the truth propositions of the physical sciences into anthropology, although for somewhat different reasons than the concerns raised by Adams. Anthropology remains necessarily attentive to the mediation of truth claims such that it cannot ignore the social and historical contingencies of even mathematical and physical sciences. Thus, energopolitical analysis is not premised upon the universal truth of energic theory. Rather, it utilizes energy as lens through which to bridge and balance historical, material, and political dimensions of power. Topically, this research area communicates actively with the broader 'ontological turn' in the human sciences and with rising consideration of energoenvironmental infrastructures like electricity grids, pipelines, and water and waste systems. The theoretical inspiration for this move to rethink political power through the lens of energic forces and infrastructures comes most directly from engaging the work of scholars such as Timothy Mitchell (2009, 2011) and Hermann Scheer (2004). Mitchell revealed how twentieth-century Keynesian social democracy was enabled and later disabled by changing infrastructures of carbon energy. Scheer meanwhile analyzed how the longer energy supply chains materially intrinsic to carbon and nuclear fuel systems supported centralized modes of political authority and governance and theorized that the shorter supply chains of solar energy could generate new democratic political possibilities. Anthropological research on energopower and energopolitics is interested in both expanding the range of ethnographic case studies that discuss these entanglements of energic and political powers, and also in developing more general theoretical propositions. The hope is that both can help the human sciences to more adequately criticize our current patterns of energy flow and use (as they are clearly fomenting ecological crisis), and to imagine and advocate for new and more responsible institutions.

See also: Industrial Societies; Infrastructures of Society, Anthropology of; Knowledge, Anthropology of; Materiality and Culture; Modernity: Anthropological Aspects; Political Anthropology; Science and Technology, Anthropology of.

Bibliography

Adams, Richard Newbold, 1975. Energy and Structure: A Theory of Social Power. University of Texas Press, Austin.

Adams, Richard Newbold, 1978. Man, energy, and anthropology: I can feel the heat, but where's the light? American Anthropologist 80, 297–309.

Behrends, Andrea, Reyna, Stephen, Schlee, Guenther (Eds.), 2011. Crude Domination: The Anthropology of Oil. Berghahn, New York.

Boyer, Dominic, 2011. Energopolitics and the anthropology of energy. Anthropology Newsletter 5 (May), 7.

Boyer, Dominic, forthcoming. Energopower: an Introduction. Anthropological Quarterly.

Coronil, Fernando, 1997. The Magical State: Nature, Money, and Modernity in Venezuela. University of Chicago Press, Chicago.

Crate, Susan, Nuttall, Mark (Eds.), 2009. Anthropology and Climate Change. Left Coast Press, Walnut Creek, CA.

Dawson, Susan E., 1992. Navajo uranium mining workers and the effects of occupational illnesses: a case study. Human Organization 51 (4), 389–397.

Gusterson, Hugh, 1996. Nuclear Rites: A Weapons Laboratory at the End of the Cold War. University of California Press, Berkeley.

Hastrup, Kirsten, 2012. The Icy Breath: modalities of climate knowledge in the Arctic. Current Anthropology 53 (2), 226–244.

Henning, Annette, 2005. Climate change and energy use. Anthropology Today 21 (3), 8–12.

Howe, Cymene, 2011. Grasping the wind: renewable energy development in Oaxaca. Anthropology Newsletter 5 (May), 7.

Howe, Cymene, forthcoming. Anthropocenic Ecoauthority: The Winds of Oaxaca. Anthropological Quarterly.

Johnston, Barbara Rose, Dawson, Susan E., Madsen, Gary E., 2007. Uranium mining and milling: Navajo Experiences in the American Southwest. In: Smith, Sherry, Frehner, Brian (Eds.), Indians and Energy. SAR Press, Santa Fe, pp. 97–116.

Jorgensen, Joseph G., et al. (Eds.), 1978. Native Americans and Energy Development. Anthropology Resource Center, Cambridge.

Jorgensen, Joseph G., 1990. Oil Age Eskimos. University of California Press, Berkeley.

Kruse, John, Kleinfeld, Judith, Travis, Robert, 1982. Energy development on Alaska's North Slope: effects on the Inupiat population. Human Organisation 41 (2), 95, 97–106.

Love, Thomas, 2008. Anthropology and the fossil fuel era. Anthropology Today 24 (2), 3–4.

Love, Thomas, Garwood, Anna, 2011. Wind, sun and water: complexities of alternative energy development in rural northern Peru. Rural Society 20, 294–307.

Masco, Joseph, 2006. Nuclear Borderlands: The Manhattan Project in Post-cold War New Mexico. Princeton University Press, Princeton.

Mason, Arthur, 2007. The rise of consultant forecasting in liberalized natural gas markets. Public Culture 19 (2), 367–379.

Mason, Arthur, Stoilkova, Maria, 2012. Corporeality of consultant expertise in Arctic natural gas development. Journal of Northern Studies 6 (2), 83–96.

McNeish, John-Andrew, Logan, Owen (Eds.), 2012. Flammable Societies: Studies on the Socio-economics of Oil and Gas. Pluto, London.

Mitchell, Timothy, 2009. Carbon democracy. Economy and Society 38 (3), 399–432.

Mitchell, Timothy, 2011. Carbon Democracy: Political Power in the Age of Oil. Verso, New York.

Nader, Laura, 1980. Energy Choices in a Democratic Society. A Resource Group Study for the Synthesis Panel of the Committee on Nuclear Alternative Energy Systems for the U.S. National Academy of Sciences. National Academy of Sciences, Washington, DC.

Nader, Laura, 1981. Barriers to thinking new about energy. Physics Today 34 (9), 99–104.

Nader, Laura, 2004. The Harder path – shifting gears. Anthropological Quarterly, 771–791.

Nader, Laura (Ed.), 2010. The Energy Reader. Wiley-Blackwell, Oxford.

Nader, Laura, Beckerman, Stephen, 1978. Energy as it relates to the quality and style of life. Annual Review of Energy 3, 1–28.

Nordstrom, Jean Maxwell, et al., 1977. The Northern Cheyenne Tribe and energy developments in Southeastern Montana. In: Social and Cultural Investigations. Lame Deer: Northern Cheyenne Research Project, vol. 1.

O'Neill, John J., 1940. Enter atomic power. Harper's 181 (June), 1–10.

Powell, Dana E., 2010. "Landscapes of power: renewable energy Activism in Diné Bikéyah, " with D.J. Long. In: Smith, S., Frehner, B. (Eds.), Indians & Energy: Exploitation and Opportunity in the American Southwest. SAR Press, Santa Fe.

Rappaport, Roy, 1975. The flow of energy in agricultural society. In: Katz, S. (Ed.), Biological Anthropology: Readings from Scientific American. W.H. Freeman, San Francisco, pp. 371–387.

Reyna, Stephen, Behrends, Andrea, 2008. The crazy curse and crude domination: toward an anthropology of oil. Focaal 52 (Winter), 3–17.

Robbins, Lynn, 1980. The Socioeconomic Impacts of the Proposed Skagit Nuclear Power Plant on the Skagit System Cooperative Tribes. Lord and Associates, Bellingham.

Robbins, Lynn, 1984. Energy developments and the Navajo nation: an update. In: Jorgensen, Joseph (Ed.), Native Americans and Energy Development. Anthropology Resource Center, Boston, MA.
Rogers, Douglas, forthcoming. Energopolitical Russia: Reframing State Formation in an Oil Region. Anthropological Quarterly.
Sawyer, Suzana, 2004. Crude Chronicles: Indigenous Politics, Multinational Oil, and Neoliberalism in Ecuador. Duke University Press, Durham, NC.
Sawyer, Suzana, 2007. Empire/multitude–state/civil society: rethinking topographies of power through transnational connectivity in Ecuador and beyond. Social Analysis 51 (2), 64–85.
Sawyer, Suzana, Gomez, Terence (Eds.), 2012. The Politics of Resource Extraction: Indigenous Peoples, Corporations and the State. Palgrave Macmillan, London, UK.
Scheer, Hermann, 2004. The Solar Economy: Renewable Energy for a Sustainable Global Future. Earthscan, London.
Strauss, Sarah, Orlove, Ben (Eds.), 2003. Weather, Climate, Culture. Berg, Oxford.
White, Leslie, 1943. Energy and the evolution of culture. American Anthropologist 45 (3), 335–356.
White, Leslie, 1949. The Science of Culture: A Study of Man and Civilization. Farrar, Straus and Giroux, New York.
White, Leslie, 1959. The Evolution of Culture: The Development of Civilization to the Fall of Rome. McGraw-Hill, New York.
White, Leslie, 1987. In: Dillingham, Beth, Carniero, Robert L. (Eds.), Ethnological Essays. University of New Mexico Press, Albuquerque.
Wilhite, Harold, 2005. Why energy needs anthropology. Anthropology Today 21 (3), 1–3.
Winther, Tanja, 2008. The Impact of Electricity: Development, Desires and Dilemmas. Berg, Oxford.

Energy Flow Analysis

Helmut Haberl, Institute of Social Ecology, Alpen-Adria Universität Klagenfurt, Vienna, Austria; and Integrative Research Institute on Transformations in Human Environment Systems, Humboldt-Universität zu Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Energy is a key concept for understanding the functioning of any biophysical system. In physics, energy is defined as the ability to do work. The physical notion of energy encompasses mechanical (potential and kinetic) energy, radiant energy, chemical energy, heat, electric energy, and nuclear energy. Biophysical structures of society are intrinsically linked with the biosphere through the energy flows required for their growth, sustenance, and functioning. Socioeconomic energy flows include food for humans and livestock as well as all the energy required for artifacts (infrastructures and machinery). Energy balances published by statistical offices trace the flow of energy through an economy. They include data on primary energy, energy conversion processes, final energy, and sometimes useful energy, usually in sectoral disaggregation. These energy balances include 'technical energy,' i.e., energy used in machinery and artifacts but neither food and feed required to nourish humans and livestock nor their physical work. More inclusive methods of energy flow analysis can be used to derive a comprehensive picture of society's energetic metabolism. This is particularly important when looking at socioecological transitions, e.g., from agrarian to industrial society due to the large importance of food, feed, and physical work in preindustrial societies. In addition to measuring the energy throughput of societies, methods exist to evaluate the efficiency with which energy is gained and used. For example, the amount of energy society needs to invest per unit of energy obtained from a particular source – its 'energy return on investment' – is a key characteristic of any primary energy source. This article explains some basics of energetics and discusses methods to account for energy use (conventional energy balances, comprehensive energy flow analysis) as well as the efficiency of energy provision and energy use.

Energy is an abstract, multifaceted notion. It is of paramount importance for the functioning of all biophysical systems (Smil, 1991), including the biosphere, and of course also the biophysical structures of social systems (Fischer-Kowalski, 1998). Energy flows between ecosystems and socioeconomic systems are an important part of the socioeconomic metabolism. Energy is of pivotal importance for sustainability because many sustainability problems of human societies are strongly related with the quantity and quality of energy used (Haas et al., 2008). Energy is an indispensable economic resource because almost all technologies used in production, transport, distribution, and consumption activities are somehow related with the availability and conversion of energy (Warr and Ayres, 2012). There is a long-standing, highly controversial discussion in the social sciences to what extent energy can be used to explain social phenomena, ranging from outright denial of any explanatory power of energy to the assumption that energy were one of the most important drivers of socioeconomic change (Martinez-Alier, 1987).

The purpose of this article is to discuss how energy flows in societies can be measured, accounted for, and analyzed. It discusses several perspectives and indicators that have been proposed and used to account for socioeconomic energy use.

What Is Energy?

The notion of energy as used in physics – which is how the term will be used here – defines energy as the ability to do work, i.e., to exert a defined amount of force over a defined distance. This abstract definition covers a large range of phenomena:

- Mechanical energy (kinetic and potential energy): the energy related to the movement or location of objects with a defined mass;
- Chemical energy: the energy stored in chemical compounds and released in chemical reactions, e.g., in combustion;
- Heat energy: the energy related to the temperature of physical objects;
- Electric energy: the energy related to a current of electrons through a conductor if an electric potential is applied;
- Radiant energy: the energy in electromagnetic radiation such as sunlight or X-rays; and
- Nuclear energy: the energy in the nuclei of atoms that can be released by fusion or fission of nuclei.

Energy Units

All these different kinds of energy can be measured in the same units. Unfortunately, there is a plethora of units used to measure energy. The SI unit is the joule (J). One joule is defined as a force of one newton (N) exerted over a distance of one meter (m). Because 1 N is a very small amount of force, 1 J is also a very small amount of energy. In almost all practical contexts, one will be concerned with multiples such as kilojoule (1 kJ = 10^3 J), megajoule (1 MJ = 10^6 J), gigajoule (1 GJ = 10^9 J), terajoule (1 TJ = 10^{12} J), petajoule (1 PJ = 10^{15} J), and exajoule (1 EJ = 10^{18} J). For example, humanity's current primary energy use is a bit over 500 exajoules per year (EJ y^{-1}). Unfortunately, a large number of units is commonly used for energy. Table 1 gives information on some often-used units and provides conversion factors for converting those into joules.

Thermodynamics

One of the important characteristics of energy is that it cannot be destroyed (which, strictly speaking, also means that the

Table 1 Some common energy units

Abbreviation	*Unit*	*Conversion to Joules*
kWh	Kilowatt-hour	1 kWh = 3.6 MJ
Cal	Calorie[a]	1 cal = 4.186 8 J
Btu	British thermal unit	1 Btu = 1.055 kJ
quad	Quadrillion Btu	1 quad = 10^{15} Btu = 1.055 EJ
toe	Tons of oil equivalent	1 toe = 41.868 GJ[b]
tce	Tons of coal equivalent	1 tce = 29.3 GJ[b]

[a]Note that 'food calories' always refer to kilocalories. 1 kcal = 1000 cal (10^3 cal).
[b]Indicative value: different conventions exist. If possible, always check for conversion factor assumed.

commonly used phrase 'energy consumption' is flawed). Physicists have formulated this as the first law of thermodynamics, which states that in a closed system, the total amount of energy is constant.

That all kinds of energy can be expressed in the same units, i.e., are on a certain level equivalent, does not imply that they were equal in terms of quality. The decisive question here is concerned with energy conversions, i.e., under which conditions one kind of energy can be converted into another. This is governed by the second law of thermodynamics, sometimes also called the entropy law. It states that in a closed system, energy conversions can only occur if entropy increases. Entropy is related with order: with rising entropy, order decreases. Consider the kinetic (mechanical) energy of a moving object, e.g., a falling stone, all atoms in the stone move in the same direction, i.e., their order is very high (low entropy). When the stone hits the ground, the kinetic energy is released, usually resulting in a deformation of the surface. The kinetic energy results in a deformation, i.e., a change in the potential energy of the surface material. In addition, during the impact, heat was released due to friction. If this happens in a closed system, the total energy is of course constant (first law of thermodynamics) but the entropy has increased (second law) because some of the highly ordered energy (kinetic energy) was converted to a very low-order energy, heat, i.e., the random movement of atoms and molecules within the stone, the surface, and the air above them; and perhaps some ordered structures were also destroyed during the impact.

What is meant by measuring all different kinds of energy in one unit, therefore, is that they are equivalent with respect to the lowest quality energy in the list, i.e., heat, into which all other kinds of energy can be converted without loss. In other words, what is being measured is the 'heat equivalent' or enthalpy. Because of the quality differences of different kinds of energy, it makes sense to introduce a second measure to characterize energy flows, which is exergy. Exergy is a measure of energy quality, i.e., the fraction of an energy flow that could be converted into mechanical energy, i.e., the highest quality energy. For heat, exergy depends on the temperature difference of a physical object compared to its environment. At ambient temperature, exergy is zero because no temperature flow exists that could be used to generate mechanical energy. The higher the temperature difference, the larger the temperature flow, and hence also the fraction of the heat energy that could (theoretically) be converted to mechanical energy.

Power

Another important distinction is that between energy and power. Energy refers to a defined amount of work performed, regardless of the time used to do so. Whenever energy is flowing or converted, this happens at a certain rate, i.e., a certain amount of energy is converted per unit of time. The physical notion to characterize this is power, defined as energy flowing per unit of time. The SI unit of power is the watt (W): one watt is defined as one joule per second ($1\,\mathrm{W} = 1\,\mathrm{J\,s^{-1}}$). If an energy flow of 1 W persists for 1 h (3600 s), the total energy flow amounts to 1 W h.

Energy Flows in the Biosphere

The whole biosphere can be described as a giant system of material flows, materially closed but driven by a continuous influx of energy from the sun (as well as some geothermal energy from the hot interior of the earth). The influx of radiant energy from the sun drives the earth's climate system, including all atmospheric processes (evaporation of water, cloud formation, wind, etc.), as well as almost all life processes (except for chemosynthetic organisms, e.g., deep-sea bacteria living near hydrothermal vents; Odum, 1971).

The basic process capturing solar energy and turning it into chemical energy available for all living organisms is photosynthesis. Through photosynthesis, plants transform inorganic inputs (carbon, water, nitrogen, phosphorous, etc.) into the myriads of different energy-rich organic materials of which all organisms consist. This process is denoted as 'primary production' and can be described using indicators such as gross and net primary production. The latter is abbreviated as NPP and is defined as the total amount of carbon fixed through photosynthesis (i.e., gross primary production), net of the plant's own energy demand.

Organisms capable of photosynthesis are denoted as 'autotrophic,' i.e., they can feed themselves from inorganic supplies. All other organisms – animals, fungi, and many microorganisms – are called 'heterotrophic' because their energy supply depends on the intake of organic materials from either living or dead organisms. An important branch of ecology (often called synecology or ecosystem ecology) is concerned with analyzing the flow of trophic energy through ecosystems, using distinctions such as herbivores (animals eating plants) or carnivores (animals eating other animals, which may be subdivided into several levels, depending on their prey) and analyzing the structure of food chains or food webs, as well as the processes of primary production and nutrient cycling in ecosystems.

Through hunting, gathering, cropping, livestock rearing, or forestry, humans are able to tap into this flow of trophic energy (Boyden, 1992), thereby appropriating a fraction of the NPP for their purposes (*see* Land Change: The Merger of Land Cover and Land use Dynamics; Land Use Indicators). Since the Neolithic revolution, humans actively manage or 'colonize' (Fischer-Kowalski and Haberl, 1993) ecosystems, thereby affecting many important patterns and processes in ecosystems, including soil, hydrology, the species composition of vegetation, nutrient and carbon cycles, and many more, in order to be able to harness the productivity of ecosystems for human purposes.

Energy is also of paramount importance at the level of the individual organism. The metabolism of any organism requires a continuous flux of energy, either as radiant energy or chemical energy in the form of food. All the intricate biochemical reactions within an organism critically depend on this energy supply for growth, maintenance of body functions, including their ability to do work, and reproduction. This of course also applies to the food supply of humans and livestock, and hence to socioeconomic metabolism.

Energy Flows in Socioeconomic Systems

Energy in Socioeconomic Metabolism

The concept of socioeconomic metabolism includes all stocks and flows of materials, energy, or defined substances (e.g., carbon) related to a society (*see* Energy and Society). This concept entails defining biophysical structures of society (physical stocks, usually thought to include human bodies, livestock, and artifacts) that require a continuous inflow of materials and energy to be reproduced and to function the way they do (Ayres and Simonis, 1994; Fischer-Kowalski, 1998).

This physical inflow is either taken from the biosphere (resource extraction) or imported from other socioeconomic systems. During production and consumption processes, materials are transformed from raw materials to interim products (often in many stages), final products, and eventually wastes and emissions that are generated throughout the production chain and at the end of the lifetime of products. A considerable fraction all materials used by a society ends up as infrastructures, including not only transport or manufacturing infrastructures but also housing and commercial buildings.

All these processes require energy, be it in the form of physical work of humans or livestock or produced through various kinds of machinery, heat, and chemical energy. Indeed, a considerable fraction of the materials that make up a modern society's metabolism are used for energy supply, e.g., coal, petroleum, natural gas, or biomass, which is used not only for combustion but also as food supply for humans and livestock. Use of materials and energy is therefore closely linked: Production processes and the use of products and infrastructures require energy (e.g., houses need to be heated, cooled, and lighted), but causality also runs in the opposite direction: available infrastructures increase the attractiveness of certain technologies, e.g., the more roads there are the higher the utility of a car.

Conventional Energy Statistics and Balances

Nutrition of humans and livestock is a relevant energy conversion process even under industrialized conditions but even more so in agrarian societies. Nevertheless, modern conventions for compiling energy statistics and energy balances for national economies exclude these biological energy flows and report only energy used in artifacts, i.e., in technical structures such as motors, power plants, furnaces, and boilers (IFIAS, 1974).

The notion of 'energy statistics' refers to any collation of energy data. Such data are usually collected by statistical offices and often presented in a multitude of units for different kinds of energy (e.g., fossil fuels are often reported in units such as 'oil equivalents,' whereas electricity is reported in kilowatt-hour). By contrast, an energy balance presents a consistent picture of inflows, conversion, and final use of energy within a defined socioeconomic system, usually a national economy. Energy balances are based on data from energy statistics but require also modeling related to incomplete or missing data for consistency reasons. Energy balances usually use only one unit.

The following notions are used to distinguish different stages of energy flows through an economy (Haas et al., 2008; Haberl, 2001a; IFIAS, 1974).

Primary energy is energy in the form in which it is extracted from the biosphere. This includes not only wood, coal, crude oil, natural gas, and heat derived from nuclear fission but also hydropower, wind power, solar energy, and heat from geothermal sources. In energy balances, primary energy use is generally approximated through the indicator 'total primary energy supply,' abbreviated TPES, which is the sum of the domestic extraction of primary energy plus the net import (=import minus export) of energy carriers. Therefore, TPES must not be equated with primary energy, in particular for countries for which the fraction of final energy carriers in energy import is substantial, but for many countries it is a reasonable, readily available proxy.

A multitude of conversion processes exist that convert primary energy into final energy carriers as used in households, service and commercial sectors, agriculture, manufacture, and industry. Prominent examples include the production of electricity in power plants, sometimes combined with cogeneration of heat for district heating and refineries that convert crude oil into a plethora of petroleum products.

Final energy is energy sold to final consumers. Final consumers are consumers that use the energy to generate useful energy and eventually energy services (see definitions below). Note that 'final energy' is an economic notion: Both households or companies may be final consumers; however, if energy is bought for conversion into another form of energy that is then sold, this is not considered part of the final energy use but of the conversion sector. Final energy carriers include fuel wood, wood chips, coal briquettes, coke, a variety of petroleum products such as gasoline, diesel, or heating oil, gas, electricity, or heat from a district heating plant. Energy balances usually report figures for final energy use, and the ratio of TPES to final energy use is to a certain extent informative with respect to an economy's energy efficiency.

Useful energy is the energy actually performing a certain task required to provide an energy service. Useful energy includes the mechanical energy needed to shape, move, or lift objects (stationary motors, vehicles); the heat required for heating or processing purposes; chemical energy needed in the chemical industry; light (radiant) energy; or the data processing services of computers. The transformation of final to useful energy is generally carried out by final consumers.

Energy services are the services for the sake of which energy is actually used (Lovins, 1979). Examples are manifold and include transport; establishment of suitable conditions within a building in terms of temperature, humidity, or illumination; cleaning of clothes; shaping of objects; chemical conversions; and many more. In contrast to primary, final, and useful energy, energy services cannot be measured in energy units – so

far, all attempts to define aggregate indicators for energy services have not produced satisfactory results.

There are several conventions that need to be considered when interpreting data from energy statistics and energy balances. In general, statistical offices usually report these together with the data or in separate methodological notes; failure to consult the respective documents in order to understand the definitions used may result in quite substantial misinterpretations. The two most important issues are noncombustible primary energy used in electricity generation and the distinction between gross and net calorific value. Different conventions are used to account for noncombustible energy (e.g., nuclear, wind, or water power, or geothermal energy) used in electricity generation; e.g., one may approximate the heat equivalent of the energy used or one may estimate that amount of fuel that would have been used in a thermal power plant producing the same amount of energy. Results from different methods may differ by large margins (factors of three or more). Gross calorific value (GCV) is the entire amount of energy released during combustion, including the energy released by condensation of water vapor in the flue gas; net calorific value (NCV) does not include this latter fraction. NCV is 5-15% smaller than GCV, depending on fuel.

The Energetic Part of Socioeconomic Metabolism

Preindustrial societies depended to a large extent on muscular power of humans or draught animals for their supply of mechanical energy. Additional primary energy inputs were wind and water power harnessed through mills as well as sailing boats (Smil, 1994). Despite their importance in terms of work supply for important activities such as grinding, forging, or transportation, these sources were quantitatively rather small. For example, in 1830, approximately 99% of Austria's primary energy supply consisted of biomass, most of which was used as food and feed (Krausmann and Haberl, 2002). Although the transition from agrarian to industrial society was driven by new energy sources, in particular fossil energy (Sieferle, 2001), biomass remained indispensable for socioeconomic metabolism as food and feed for humans and livestock. More inclusive concepts than those used in modern energy balances are hence required when analyzing preindustrial societies or long-term energy transitions (Fischer-Kowalski and Haberl, 2007).

Methods to account for the energetic aspects of socioeconomic metabolism in a manner that is conceptually consistent with material flow analysis (MFA) are primarily concerned with flows of biomass and muscular power. In principle, they follow the logic of energy balances as outlined in the previous section; i.e., they trace the flow of energy through the socioeconomic system from primary energy supply to useful energy (**Figure 1**). However, in doing so, these methods consider not only technical energy but also biomass and physical work (Haberl, 2001a,b). In contrast to conventional energy balances that account for biomass only if it is used in technical structures (e.g., in heating systems and power plants, or more recently as vehicle fuel), all 'metabolized' biomass is accounted for.

In short, these methods consider all energy crossing the nature–society boundary based on their physical energy content. Biomass is accounted for using the gross calorific value. Nutrition of humans and livestock is considered as an energy conversion process within society. Processing of primary biomass to food and feed is regarded as a conversion from primary to final energy. Food and feed are defined as final energy, and human as well as animal labor as useful energy (Haberl, 2001a). This allows construction of an energy flow analysis (EFA) framework that can be used to generate indicators such as

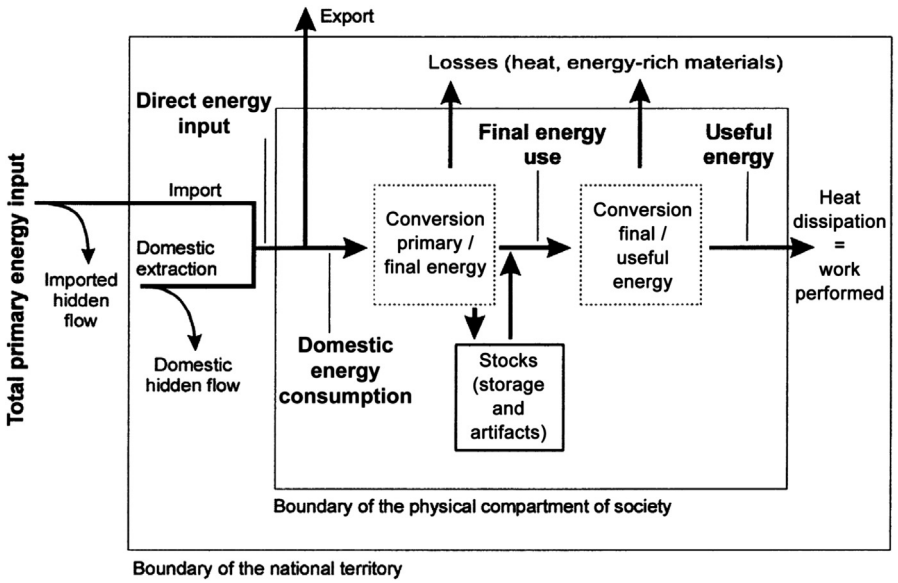

Figure 1 Flow chart of an energy flow analysis consistent with the material flow analysis approach as used in socioeconomic metabolism studies. Reprinted from Krausmann, F., Haberl, H., 2002. The process of industrialization from the perspective of energetic metabolism: Socioeconomic energy flows in Austria 1830–1995, Ecological Economics, 41 (2), pp. 177–201, Figure 1, with permission from Elsevier.

- Direct energy input: all energy entering a society (the same concept as 'direct material input' as used in MFA).
- Domestic energy consumption (DEC) is defined as direct energy input minus export (the same concept as 'domestic material consumption' as used in MFA).
- Indicators of final and useful energy as they are used in conventional energy balances.
- 'Hidden flows,' i.e., energy mobilized during energy extraction but not actually entering socioeconomic metabolism (e.g., crop residues).

EFA complements MFA methods to form the combined 'material and energy flow analysis' or the MEFA framework, which has been shown to be useful for studying sustainability (Haberl et al., 2004) and socioecological transitions (Haberl, 2001b; Krausmann and Haberl, 2002).

Embodied Energy and Emergy

The energy expended during production of a good and service, or in all products and services consumed in a national economy (or any other socioeconomic subsystem) is denoted as 'embodied energy' (Costanza, 1980). Embodied energy can be estimated using either input–output methods or life cycle assessment methods. Embodied energy is calculated using the system boundaries applied in conventional energy balances and EFA, whereas the concept of 'emergy' (Odum, 1996) traces all energy inputs back to one kind of available energy (exergy), usually solar energy. In this accounting framework, all biospheric energy conversions that are needed to produce any resource needed in a production process are traced back to the solar energy eventually required for them to happen and measured in 'solar equivalent joules,' abbreviated seJ.

Measures of Energy Efficiency and Quality

In physics, energy efficiency is defined as the relation between the input and output of energy in an energy conversion process. Energy efficiency (η) is defined as the ratio of energy output to energy input. η is dimensionless and may also be given as output in percent of input. This straightforward concept can, however, get a lot more complex when applied to socioeconomic systems due to the above-discussed questions of energy quality. Energy quality in this sense has largely two meanings (Cleveland et al., 1984; Hall et al., 1986): (1) the quality of energy in terms of its utility for performing certain tasks (which is often related to exergy) and (2) the quality of an energy source used by society (related to the efficiency with which it can be recovered from a source, i.e., its 'energy return on investment' or EROI).

Technoeconomic Concepts of Energy and Exergy Efficiency

An often-used measure of aggregate energy efficiency of a national economy is the ratio between the energy used and economic output, often measured as gross domestic product (GDP; *see* Data Bases and Statistical Systems: Economics (Statistical Systems); Environmental Economics). Relating these two variables results in an indicator of the form 'energy per GDP,' e.g., J US$$^{-1}$ or J €$^{-1}$. Of course, all issues related to the measurement of GDP need to be considered, e.g., whether it is measured in constant or nominal prices or how different currencies are converted (exchange rates or purchasing power parities).

Any of the above-discussed energy flow indicators (TPES, DEC, final energy, etc.) can be (and most likely has been) divided by GDP, but of course the meaning of the resulting efficiency measure, and accordingly its interpretation, will be different. For example, if measures of primary energy are used (DEC, TPES), improvements in the efficiency of conversion of primary to final energy show in the results, which is of course not the case if final energy data are analyzed. Such efficiency measures have been used in cross-country as well as longitudinal analyses, which almost always have shown quite substantial improvements in energy efficiency over time as well as along agrarian–industrial gradients (Fischer-Kowalski and Haberl, 2007).

However, as has been noted long ago (Cleveland et al., 1984; Hall et al., 1986), improvements of technoeconomic energy efficiency are only partially driven by improved technologies such as more efficient motors, heating systems, or power plants. While improvement of technical energy conversion efficiency has undoubtedly happened, long-term energy transitions (Grübler et al., 1999; Grübler, 2003) are also characterized by switches from thermodynamically poorer to superior kinds of energy, e.g., from biomass to coal to oil products to natural gas and electricity. Exergy is a useful concept for taking different qualities of energy sources into account (Ayres et al., 2003). This is particularly relevant when analyzing the role of energy for economic growth (Ayres and Warr, 2009).

Exergy is also relevant when considering technological measures to improve energy efficiency. For example, switching from oil to electricity for space heating may seem like an improvement of final energy efficiency because electrical resistance heating is 100% efficient, whereas the efficiency of heating systems based on fuel oil may be some 60–80%. In addition, electrical heating can be controlled a lot better than an oil-fired heating, saving additional amounts of final energy. However, switching from oil to electricity makes a lot less sense because converting the most valuable and versatile form of energy (electricity) into the lowest quality energy (low-temperature heat) is a huge waste of exergy. A much better way of using the exergy in the oil would be to switch to cogeneration of electricity and heat because any conversion of combustibles to electricity results in considerable energy losses in the form of waste heat due to the second law of thermodynamics. Waste heat can, however, be used for space heating if distributed through district heating grids, thereby raising the overall efficiency of the system in terms of both exergy and energy (Rosen et al., 2005).

Energy Return on Investment

One important consideration related to energy sources is how much energy society has to invest per unit of energy gained. The ratio of energy gained per unit of energy invested has been denoted as EROI. The calculation of EROI of different energy

sources is often denoted as 'net energy analysis' (Cleveland et al., 1984; Giampietro and Pimentel, 1991; Hall et al., 1986). Obviously, if EROI falls below one for a particular source of energy, then the process turns into an energy-losing activity instead of an activity delivering net energy to society. Industrial societies accept that in certain cases, e.g., in the case of many agricultural production activities that require more energy inputs than they deliver in terms of food calories (Krausmann, 2004; Pimentel et al., 1973). This is only possible because of the availability of other sources of net energy that can be used to energetically 'subsidize' food production.

Decarbonization

The notion of decarbonization relates to the reduction of carbon inputs to socioeconomic metabolism or of greenhouse gas (GHG) emissions such as CO_2 or CH_4. Indicators such as GHG emissions per unit of technical energy, e.g., per TPES, are widely used in climate-related discourses. They are often expressed as kilogram CO_2 equivalent per year. Different GHGs are converted into CO_2 equivalents according to their respective global warming potentials. Partial decarbonization can be achieved by increasing the share of low-GHG fuels such as hydropower in total energy supply. In theory, it is possible to reduce net carbon emissions to zero or even achieve net negative emissions if technologies such as bioenergy with carbon capture and sequestration are used on a large scale. However, on the input side, socioeconomic metabolism cannot be entirely decarbonized because biomass is indispensable for socioeconomic metabolism, at least as food and feed, and the carbon content of dry matter biomass amounts to 45–50% (Odum, 1971).

Conclusions

Methods of EFA are an important component of the toolkit of MEFA (Haberl et al., 2004) required for the empirical analysis of socioeconomic metabolism. While conventional energy balances and statistics provide a wealth of important information on the socioeconomic use of energy, for many instances they need to be complemented by additional data and methods to analyze important questions related to sustainability and society–nature interaction. EFA methods are available to trace socioeconomic energy flows in a manner that is consistent with MFA approaches. They are particularly important when analyzing preindustrial societies, but biomass and work flows not accounted for in conventional energy balances may be quantitatively important even for industrial economies. Measures of energy efficiency are required to optimize energy use and judge the quality of energy resources. Embodied energy is a useful concept to analyze indirect energy demand related to many products with no obvious relation to energy. Exergy is relevant for understanding differences in the quality of different kinds of energy that need to be taken into account when analyzing the role of energy for production processes and economic growth. Overall, EFA provides an indispensable framework and a variety of methods and tools for analyzing biophysical aspects of sustainability and society–nature interaction.

See also: Agrarian Political Economy; Data Bases and Statistical Systems: Economics (Statistical Systems); Eco-cities and Sustainable Urbanism; Energy, Anthropology of; Energy, Geography of; Environmental Economics; Land Change: The Merger of Land Cover and Land use Dynamics; Land-Use Indicators; Planning Issues and Sustainable Development; Sustainability and Sustainability Science; Sustainability: A Dialog of Values; Sustainable Development: An Economic Perspective; Sustainable Production and Consumption.

Bibliography

Ayres, R.U., Ayres, L.W., Warr, B., 2003. Exergy, power and work in the US economy, 1900–1998. Energy 28, 219–273.
Ayres, R.U., Simonis, U.E., 1994. Industrial Metabolism: Restructuring for Sustainable Development. United Nations University Press, Tokyo, New York, Paris.
Ayres, R.U., Warr, B., 2009. The Economic Growth Engine: How Energy and Work Drive Material Prosperity. Edward Elgar, Cheltenham, UK; Northampton, MA.
Boyden, S., 1992. Biohistory: The Interplay between Human Society and the Biosphere, Past and Present. UNESCO and the Parthenon Publishing Group, Paris.
Cleveland, C.J., Costanza, R., Hall, C.A.S., Kaufmann, R., 1984. Energy and the U.S. economy: a biophysical perspective. Science 225, 890–897.
Costanza, R., 1980. Embodied energy and economic valuation. Science 210, 1219–1224.
Fischer-Kowalski, M., 1998. Society's metabolism: the intellectual history of materials flow analysis, part I, 1860–1970. Journal of Industrial Ecology 2, 107–136.
Fischer-Kowalski, M., Haberl, H., 1993. Metabolism and colonization. Modes of production and the physical exchange between societies and nature. Innovation in Social Sciences Research 6, 415–442.
Fischer-Kowalski, M., Haberl, H., 2007. Socioecological transitions and global change. Trajectories of social metabolism and land use. In: van den Bergh, Jeroen (Series Ed.), Advances in Ecological Economics. Edward Elgar, Cheltenham and Northampton.
Giampietro, M., Pimentel, D., 1991. Energy efficiency: assessing the interaction between humans and their environment. Ecological Economics 4, 117–144.
Grübler, A., 2003. Technology and Global Change. Cambridge University Press, Cambridge.
Grübler, A., Nakićenović, N., Victor, D.G., 1999. Dynamics of energy technologies and global change. Energy Policy 27, 247–280.
Haas, R., Nakicenovic, N., Ajanovic, A., Faber, T., Kranzl, L., Müller, A., Resch, G., 2008. Towards sustainability of energy systems: a primer on how to apply the concept of energy services to identify necessary trends and policies. Energy Policy 36, 4012–4021.
Haberl, H., 2001a. The energetic metabolism of societies part I: accounting concepts. Journal of Industrial Ecology 5, 11–33.
Haberl, H., 2001b. The energetic metabolism of societies: part II: empirical examples. Journal of Industrial Ecology 5, 71–88.
Haberl, H., Fischer-Kowalski, M., Krausmann, F., Weisz, H., Winiwarter, V., 2004. Progress towards sustainability? What the conceptual framework of material and energy flow accounting (MEFA) can offer. Land Use Policy 21, 199–213.
Hall, C.A.S., Cleveland, C.J., Kaufmann, R., 1986. Energy and Resource Quality: The Ecology of the Economic Process. John Wiley & Sons, New York.
IFIAS, 1974. Energy Analysis Workshop on Methodology and Conventions. International Federation of Institutes for Advanced Study (IFIAS), Stockholm.
Krausmann, F., 2004. Milk, manure, and muscle power. Livestock and the transformation of preindustrial agriculture in Central Europe. Human Ecology 32, 735–772.
Krausmann, F., Haberl, H., 2002. The process of industrialization from the perspective of energetic metabolism: socioeconomic energy flows in Austria 1830–1995. Ecological Economics 41, 177–201.
Lovins, A.B., 1979. Soft Energy Paths: Towards a Durable Peace. Friends of the Earth International, San Francisco, CA; Cambridge, MA.
Martinez-Alier, J., 1987. Ecological Economics. Energy, Environment and Society. Blackwell, Oxford, UK.
Odum, E.P., 1971. Fundamentals of Ecology. Saunders, Philadelphia.

Odum, H.T., 1996. Environmental Accounting, Emergy and Environmental Decision Making. Wiley, New York.
Pimentel, D., Hurd, L.E., Bellotti, A.C., Forster, M.J., Oka, I.N., Sholes, O.D., Whitman, R.J., 1973. Food production and the energy crisis. Science 182, 443–449.
Rosen, M.A., Le, M.N., Dincer, I., 2005. Efficiency analysis of a cogeneration and district energy system. Applied Thermal Engineering 25, 147–159.
Sieferle, R.-P., 2001. The Subterranean Forest: Energy Systems and the Industrial Revolution. The White Horse Press, Cambridge.
Smil, V., 1991. General Energetics: Energy in the Biosphere and Civilization. Wiley Interscience, New York.
Smil, V., 1994. Energy in World History. Westview Press, Boulder, CO.
Warr, B., Ayres, R.U., 2012. Useful work and information as drivers of economic growth. Ecological Economics 73, 93–102.

Relevant Websites

http://www.aps.org/policy/reports/popa-reports/energy/units.cfm – American Physical Society: Energy Units.
http://physics.about.com/od/thermodynamics/a/lawthermo.htm – About.com Physics: Thermodynamics.
http://www.eoearth.org/ – Encyclopedia of Earth: A Plethora of Entries on Energy, Environment, Ecology, Food, Etc.
https://www.iea.org/stats/ – OECD/International Energy Agency (IEA): Energy Statistics.
http://www.physicsclassroom.com/Class/energy/ – The Physics Classroom: Energy Tutorial.
http://unstats.un.org/unsd/energy/ – United Nations (UN): Energy Statistics.

Energy, Geography of

Johannes Hamhaber, Cologne University of Applied Sciences, Institute for Technology and Resources Management in the Topics and Subtropics (ITT), Köln, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W. Brücher, volume 7, pp. 4520–4523, © 2001, Elsevier Ltd.

Abstract

The geography of energy analyzes societies' uses of energy and their relation to space. Rooted in historical transitions, the geography of energy follows some ordering principles such as the production chain and specific energy system dimensions. Using concepts of Euclidian space (distribution, distance, etc.), as well as specific spatialities, the geography of energy extends well beyond industrial geography, with social, political, agricultural, and environmental geography contributing to a deeper integrated view of the geographies of single fuels, energy technologies, energy regions, or even fundamental energy system transitions.

The Geography of Energy analyzes society's use of energy – as one of its fundamental resources – and its relation to spaces. This encompasses the natural systems as they provide energy in the form of fossil and renewable resources, the systems of energy production, conversion and provision to the users, the consumption of energy in all relevant sectors as well as the repercussions for and impacts on social and ecological systems. Starting from some fundamental ordering principles, mainly the production chain and the energy systems' dimensions, geographical conceptualizations of space and spatiality allow for specific approaches to analyze and explain the geographies of single fuels, energy technologies, energy regions and even fundamental energy system transitions.

Historical Energy Transitions

Energy is one of the primary building blocks of the functioning of all societies, be they prehistorical, historical, or contemporary, each era with their specific sociometabolic processes. The provision of functions like food production and processing, the manufacturing of raw materials and goods, and the operation of mobility, communication, and information technologies, among others, are based on the continuous and reliable supply of energy in its different forms. Regarding the geography of energy, all these options of supplying and using energy are organized in energy systems that have specific technical, economic, and political conditions and are directly linked with the environment and society. Thus, energy systems are inextricably embedded with physical space as well as with social spaces.

Historically, energy systems were based on the flow of energy provided by radiation intake from the sun, most of the energy forms we today summarize under the term of renewable energies. Thus, before the age of industrialization, most energy was derived mainly from wind, water, and biomass, and transformed by fire or animal or human muscle in an agrarian solar energy regime (Sieferle, 2001). This applies for mechanical energy and heat for processing, as well as for mobility through carts and ships. Further forms of renewable energies, such as geothermal, wave, or tidal energy, were only used under specific circumstances (**Table 1**).

With the industrial revolution, emerging mass production as well as mobility through railroads and steam ships were based on energy forms with a very high energy density. Traditional biomass as dominant primary energy source was replaced by fossil fuels like coal and – after the second half of the nineteenth century – also mineral oil. Similarly, steam and later electricity stepped in as higher quality secondary energy sources. They have accompanied fundamental social changes, including urbanization and mass consumerism, and also still fuel today's postmodern information and knowledge societies (Bridge et al., 2013).

Today, most global societies are in this fossil stage, with natural gas currently taking larger shares of the global fuel mix, notwithstanding some nations with considerable shares of nuclear power in their primary energy or the still relevant share of traditional biomass in many countries of low to medium development.

The reconsideration of renewable energies in recent times, based on technologies such as photovoltaic or concentrated solar power (CSP), or high-tech windmills, has not yet been able to replace the dominant share of fossil fuels of some two-thirds of global commercial energy demand; nor is it expected to be able to do so within the next few decades (IEA, 2012). The rise of emerging economies with their large populations and high dynamics of consumption patterns pushes global energy

Table 1 Overview of concepts of energy transitions (own compilation)

	Author		
Era	*Georgescu-Roegen (1982)*	*Sieferle (2001)*	*Brücher (2009)*
Fire	"Prometheus I"	Hunter-gatherer societies	Energy from space (fire, wood, food)
Food/Plants		Agricultural societies	
Fossils	"Prometheus II"	Modern (industrial) societies: fossil fuels interpreted as transformation phase	Energy for space
Post-fossil			Energy from space (renewables)

demand and leads these nations to rely largely on affordable and (domestically) available fuels and technologies, namely coal and oil; while the poorer parts of the Global South still struggle with sufficient supply and full coverage in replacing wood fuel and other forms of traditional biomass.

Organizational Principles and the Geography of Energy

Some fundamental characteristics of energy and the implemented energy systems place them squarely into geographical interest (Hamhaber, 2010).

- Being defined as a natural resource, energy straddles the connecting point between natural and social systems.
- Energy resources are distributed very unevenly, on all scales from global to local, as is the demand for it. These disjunct distributions of supply and demand pose the technological challenge to transport energy over ever increasing distances.
- Energy, energy production and conversion as well as energy consumption are highly relevant economically, as they fuel all activities and create respective markets on the supply and demand side.
- Energy use generates wide reaching environmental impacts on all scales from local emissions up to global climate change.
- This creates a societal challenge regarding access and affordability, be it local energy poverty or limited regional economic growth options due to unaffordable fuel supply. Obviously, the distribution of costs and benefits related to existing energy systems poses fundamental questions of social equity.
- Therefore, energy resources and their use are politically deeply contested – this reaches from local distributive conflicts of detrimental effects (such as 'not in my backyard,' NIMBY) up to hot conflicts to outright war ('blood for oil').

Given these characteristics of energy systems of obvious geographical interest, it seems surprising that the discipline has until recently not fully taken account of energy and energy systems as research topics. Despite the few available monographs (Manners, 1964; Sevette, 1976; Calzonetti and Solomon, 1985; Chapman, 1989; partly also Clark, 1990; lately Brücher, 2009) and quite some articles (e.g., ter Brugge, 1984), energy was never established as a subdiscipline or a research topic in its own right within (human) geography. Despite ample motivation for closer analysis provided by historical developments, such as oil crises, or single events like the nuclear accident in Chernobyl, geographers tended to scrutinize energy topics mainly from their subdisciplinary view.

Until recently, the geography of energy was typically considered a task of economic geography at large, and industrial geography in particular. Brücher (2009), however, considers several major differences of energy systems to other (production) industry sectors.

- Energy product variation is limited: Energetic resources mainly stay in the sector and will finally become energy products or energy services for the customer. Other industrial processes are not as inextricably linked to the given resource, or – vice versa – the resource can be converted into a multitude of products (e.g., steel into spoons, cars, construction material etc.). Thus, with the large exception of oil derivatives for chemical industries, energy production chains appear to be largely self-contained.
- The energy production chains are intensely vertically integrated: Despite their structure being similar to production chains in other industries, the energy industries, however, follow largely a narrow path with little deviations, called the imperative cohesion of the energy chain. Therefore, all energy markets show deeper links between the primary, secondary and tertiary sector than most other industries.
- Energy systems show market failures: the fordist character of most energy products creates massive economies of scale and, in centralized systems, typically requires large scale investments. Combined with technical conditions such as network-based distribution this has created mainly oligopolies and even monopolies. The transport and distribution grids of gas and power are considered natural monopolies, and even under the current deregulation regimes, the grid owner might be a single buyer (monopsist) or a monopolist service provider. To a lesser extent, also oil and coal industries show a tendency to oligopolistic structures.
- Energy is intensely regulated: an industry so important to society and with embedded market failures demands for close scrutiny and control by the state regulators. There is hardly any other industry under a similar political influence and interest in all countries. In many cases, the energy industries have been nationalized (electricity companies, oil companies, etc.) or remain in the ownership of smaller territorial political-administrative units. Often, cities have their own energy and service providers, as this often generates revenues, allows direct control of the sector and is a means for service provision by the states to their citizens.
- Energy harbors large conflict potential: as energy is not only a production factor for a single industry, but also a fundamental natural resource, conflict lines similar to those of water, land and minerals apply (see, e.g., the emerging debate on the energy-water food security nexus (Nexus, 2013)). More than for most other industries, many activities along the energy production chain are contested, be it mining, conversion in power plants or refineries, the large scale transport infrastructures, as well as the environmental repercussions.

Thus, the geography of energy may well be tackled from a perspective of industrial geography, among others, but has emerged as an interdisciplinary field that spans all three economic sectors with an extraordinary political influence, and deep social and environmental repercussions. Accordingly, other subdisciplines have found and lately renewed their interest in energy topics, among them urban geography, the geography of agriculture/forestry, political geography, and – to a lesser extent – social geography.

Regarding ordering principles, classical approaches toward a 'geography of energy' would apply an approach from industrial geography, along the production chain of energy systems. It reaches from primary energy (fossil fuels, solar radiation, uranium, etc.) through its conversion into secondary energy (electricity, diesel, etc.) to the end energy, and finally, provides

the energy service demanded (heat, light, mobility). On this production chain, specific activities are performed such as mining, processing, storing, transporting, enriching, distributing, and consuming. In each of these processes useful energy is lost; even the 'consumer' stage is nothing but a conversion and thus reduction of quality and energy contained in the fuel. This chain, however, is infused with other dimensions of the energy systems, and specifically, with spatial connotations.

Chapman (1989) stated that in order to mobilize and deliver energy, all societies created energy supply systems that respond to a specific spatiotemporal situation, the available technologies and other determinants. Such energy systems encompass six major dimensions.

- The functional dimension is defined largely by technological choices.
- The organizational one by the setup of major actors and stakeholders, their roles and the underlying regulation.
- The economic dimension focusses mainly on costs and benefits of such systems.
- The variables of the environmental dimension are the available resources and the residuals produced along the production chain.
- The social dimension encompasses values, perceptions, but also questions of equity.
- The social dimension encompasses values, perceptions, but also questions of equity.
- Finally, Chapman adds the spatial dimension regarding the location, distances and densities of and in such systems.

These dimensions are intrinsically linked: technology, for example, defines the energy resources and reserves of the globe, of nations, and of regions. Economic viability of resources is largely subdued to technological progress: drilling technologies to access deep-sea oil reserves have extended the peak-oil, fracking technology of shale gas collapsed the U.S. natural gas price and turned its role as net importer, or liquefying natural gas has allowed for marine transport compared to only transcontinental gas pipelines and thus created a new, global market. The dwindling price for photovoltaic solar modules is largely due to technological progress and economies of scale, and thus has created a globally marketable product which today even changes rural livelihoods in the least developed countries. Similarly, technological choices influence the nations' fuel mixes, as nuclear power, for example, delivers up to three quarters of electricity in France while other countries struggle for the right to access this technology as an energy source in politically contested arenas.

It may be argued that the political dimension is hidden in the organizational one, and Chapman insists on its large influence, but it seems largely underdefined, as the power of actors, their territorialities, or conflict potentials regarding energy systems are not covered by the organizational dimension. Brücher (2009) on the other hand even claims the political realm to be very gearbox linking all other dimensions of energy systems (see **Figure 1**), mediating between the energy systems proper (production chain), and their external socio-economic and spatial conditions.

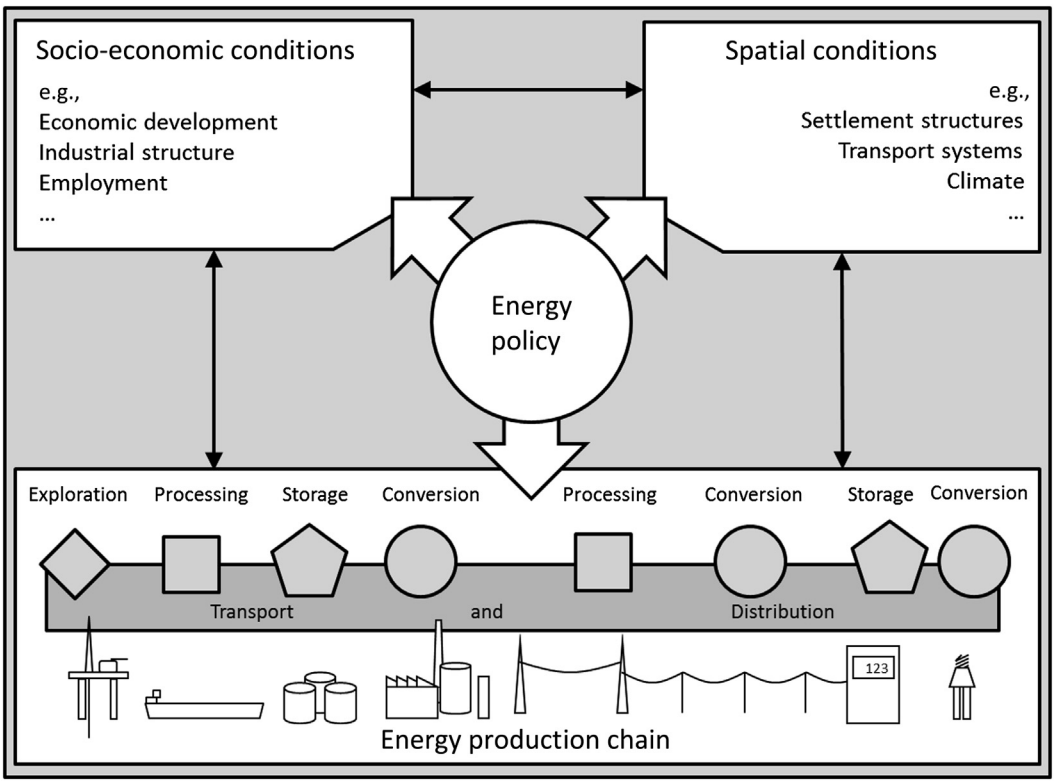

Figure 1 Energy systems model: the production chain in its interrelation to socio-economic and spatial conditions (after Brücher, W., 2009. Energiegeographie – Wechselwirkungen zwischen Ressourcen, Raum und Politik. – Studienbücher der Geographie. Borntraeger, p. 280, modified).

Here, however, the question remains how the energy systems – including all their dimensions – are spatial phenomena. In Figure 1, such spatial relations are twofold: the spatial allocation of the energy systems' production chain is specific for each energy source from primary to secondary to end energy, and externally, explicitly spatial conditions are named as being influenced and influencing the energy system such as other infrastructures like transport, climate and climate change, or settlement patterns. Beyond the debate, whether industrial structures may not also be seen as spatial conditions to the energy system, this opens the discourse on how to relate energy systems with different connotations of space, before these aspects are integrated into some specific approaches of a geography of energy.

Energy Systems and Euclidian Space

The geography of energy in a classical sense typically is understood as all spatial expressions of the energy systems in space, as well as the adaptation of such systems to spatial conditions. Therefore, instead of describing the ever changing spatial status of energy systems here by fuels, rather the different gradations of Euclidian space are explored in their explanatory value to energy systems.

Spatial Distribution: As shortly sketched above, the distribution patterns of both energy resources and demand are massively uneven. With some 70% of all conventional oil and gas reserves allocated in the so-called strategic ellipse, encompassing the Gulf region and stretching north into Central Asia and Western Siberia, the basic layout of the global oil and gas markets was determined for decades. With the rise of new exploration and exploitation technologies (deep-sea drilling, fracking, etc.) energetic resources such as arctic, deep-sea and nonconventional oil and gas have turned into economically viable reserves, shifting the playing field and possibly turning the tables for oil-exporters and importers ('golden age of gas,' IEA, 2012). Also, new players have entered the game, mainly with large new resources in Latin American and Sub-Saharan African countries.

Similarly, the potentials for renewable energies are distributed according to the natural endowments of geology, climate, and ecology. As the hydroenergy potentials of higher latitudes are largely exploited, the focus moves to the inner tropics and the foot range of the Tibetian plateau. The solar potential, however, is highest in the dry belts along the tropics, especially for CSP, similarly, wind, tidal, deep geothermal, and also bioenergies have their specific geographies (Geni, 2013).

On the other hand, the demand distribution is also massively lopsided, with the countries of the classical economic triad being the largest consumers, in total numbers as well as in per capita demand. Nowadays, a more multipolar world finds its expression also in a wider distribution of energy demand, with the emerging economies, above all China, emerging as major consumers and therefore shifting the flows of energy on a global scale (IEA, 2012). As obviously 'resources are not, they become,' the global fossil fuel map is constantly been redrawn.

Spatial allocation: The allocation of activities along the energy production chain is determined by the characteristics of different fuels. Accordingly, industries have emerged that produce and provide fossil fuels on a global scale. The oligopoly of the oil-producing seven sisters has been replaced and counteracted by the production oligopoly of the oil-exporting countries. During this shift also the allocation of the related physical infrastructure such as pipelines, refineries, and terminals has been relocated from the consumer's countries to the producers. Similarly, the transportation infrastructure starts to specialize toward pipelines and tankers for refined products. Hence, large parts of the value chain have been reallocated to the oil-producing countries which before had been mere exporters of raw products (Clark, 1990).

Scale: Accordingly, depending on the fuel source, energy markets of different sizes have evolved. Based on the energy density, the transportability and related costs, and the ease of use and conversion of the fuels, their markets are either global (as oil) to continental (for pipeline based natural gas networks) or national as is typical for electricity networks and markets down to simply local networks for wood fuel and wood chips. As this is a function of physical distance it fits very well into the concepts of Euclidean space and traditional economic geography.

This leads to new concepts of rescaling, as illustrated in integrated continental to trans-continental energy systems. The 'Desertec' scheme, for example, envisioned to provide electricity from CSP mega-projects in the Sahara not only to Northern Africa, but to Europe as well. Combined on continental scale with hydropower from Scandinavia and the Alps, wind from the maritime climate regions along the Atlantic coast and bioenergies produced in Eastern Europe, this concept aims at capitalizing on the qualities of each energy source to compensate for the respective weaknesses through a high-voltage transcontinental and circum-Mediterranean transmission grid, thus creating a new, continental geography of electrical energy in Europe. Similarly, on smaller scale, the deep transition to renewables in Germany ('Energiewende') demands for new grid capacity to absorb the shifts toward a new geography of power production, as the distribution of production capacity shifts with predominant wind and bioenergy resources in the North while some nuclear power plants in the South are being decommissioned.

Centrality: A special, spatial characteristic of energy systems is their level of centrality. As discussed above, most energy systems have emerged as being large-scale monopolistic or oligopolistic structures. Especially the network-based electricity system, where the fuel is not storable economically in large amounts, requires an immediate adaptation of the supply to the demand curve. Based on the high investment, potentially sunk costs, and the regional integration of the system grid, the electricity industry historically has been organized as fully vertically integrated systems. With the move toward nuclear power production, a technology that has large security considerations, a fundamental debate about central versus decentral energy systems emerged. Amory Lovins, a protagonist of that debate in the United States even stated a pending decision between decentral, community-based, low to medium technology, environmentally safe systems that fostered democratic participation and a sense of ownership on the one hand, and centralized, high tech, expert based, high risk technologies that would demand also a high-security state, alienate the

citizens and thus threaten democratic structures. As these two system typologies were perceived as being mutually exclusive, such a decision was assumed as irreversible (Lovins, 1977). Today, during the fundamental transition from fossil to renewable energy systems, both, central and decentral system elements are embedded in the current markets, such as in Germany with a large share of consumer owned small-scale solar photovoltaics side-by-side with base load nuclear and lignite power plants. Nevertheless, the centrality of energy systems remains a fundamental category of geographical description and analysis.

Based on the concept of centrality and spatial distribution patterns, Brücher (2009) conceived major phases of paradigmatic spatial characteristics of energy systems, regarding their dispersion and concentration in space. Where, before the industrial age, energy was derived from largely dispersed natural sources ('energy from space'), the highly concentrated fossil fuels and the emergent technologies growing with economies of scale created specific highly centralized production and distribution systems ('energy for space'). With the return to renewable energy sources in their low energy density (e.g., solar radiation), the principles of a distributed production and potentially decentralized supply systems (and thus 'energy from space') regain importance, including the conflict potentials due to competing uses of that same space (food, fodder, fiber, fuel).

New Spatialities of Energy Systems

The Geography of Energy as based on Euclidian space can clearly describe the distribution of resources and the allocation of functions of the energy production chains as original geographical contributions to the debate. Using sheer physical distance as determinant of the emergent systems, however, is unsatisfactory at least. Certainly, transport technology and related costs play a major role in decision making. Beyond the (economics of) physical space, further dimensions of the energy systems have to be considered. Energy systems are always steered largely by their organizational structure and thus the stakeholders' interests. Also, their repercussions in the social and environmental realms have gained importance in discourse and decision making, rendering energy a sector that is highly politically contested.

According to this understanding, the Geography of Energy also considers the social construction of space(s) as analytical framework for energy systems. This is in line with a paradigmatic theoretical turn of the wider discipline, where socially constructed spaces have complemented, if not largely replaced Euclidian space as reference for analysis in human geography. Thus, the approach goes well beyond the description on distributions or centrality: if e.g., politics is assumed as a 'gearbox' for all societal decisions regarding energy systems, a political–geographical approach has to include social science approaches to space. Similarly, economic geography, where the Geography of Energy traditionally would have been subsumed, has moved widely beyond classical allocational quests and has largely left Euclidian space behind. Thus, it should be ensured that the application of physical space should not extend beyond the physical world, whereas the mental and social spheres demand for their specific spatial reference systems (e.g., spatiality according to Soja, 1989; Harvey, 1996, and others). To comprehensively analyze energy systems, therefore, they have to be localized also in socially constructed space(s). From this standpoint, energy has clearly emerged with new urgency as a research subject in many subdisciplines of geography, and has been infused with spatial connotations beyond physical space. The approaches of a reformed (e.g., relational) economic geography (see, e.g., Bathelt and Glückler, 2011) help explain the shift of energy systems from a modern, mass producing energy supply infrastructure based on economies of scale to a postmodern, customer oriented service industry where power is 'green' and customers are willing to pay a primer on electrons delivered from windmills and solar systems. With the theoretical turn enter further conceptual geographical approaches to analyze energy systems with higher explanatory value. A selection will be explored in the following paragraphs.

Relative location: Other than absolute location, relative location is based on the concepts of relations of elements in a system, and thus shows the production chain not only as an interlinked, sequential set of physical locations, but rather as interdependent positions subdued to economic and time distances. A change in transport technology creates new conversion locations at the sea, be it refineries in Rotterdam, energy dependent industries such as aluminum smelters in Iceland and Quebec, or, quite recently, new terminals for liquified natural gas (LNG) tankers. New relative positions will therefore create an effect in physical space, shifting, and reallocating the product chain's elements of the world's map. Similarly, the upstream–downstream position in a network has deep repercussions on the reliability of supply, as has been proven by the Russian–Ukrainian gas dispute, effectively shutting down the supply to Europe as customer 'behind' the transit country (similarly for the UK position, see Bridge et al., 2013).

Centrality, Contiguity, and Connectivity: the concept of centrality may not fully grasp this interdependency of elements in a system. Going beyond, contiguity may grasp dispersion, whereas connectivity measures the intensity of interlinkages in a system. In a typology provided by Bridge et al. (2013), these characteristics are juxtaposed for different forms of energy systems, and by that display the fundamentally different character of systemic choices of potential energy futures, societies have to take. This also illustrates very well Lovins' (1976) argument of incompatibility of central and decentral systems. It also unveils how the debates that emerge about central or decentral systems during deep transitions to low-carbon societies hide deeper, relational characteristics, and mask distributive conflicts among actors in the field (**Table 2**).

Evolutive approaches and path dependency: structures of energy systems tend to display a high inertia to change. Many of the decision and actions in such systems may be contingent in specific structure-agency dynamics, often reconstructing existing structures whereas new, emergent structures would have to be implemented against the prevailing dominant form (Bathelt and Glückler, 2011). As an example, only due to the systematic crisis of the power industries in the 1980s and 1990s the formerly accepted vertical integration of the electricity production chain was put under scrutiny. At the same time, information and communication technology provided for

Table 2 Spatialities of energy infrastructure networks

Illustrative spatial form	*Contiguity*	*Connectivity*	*Centralization of capacity and supply decisions*
District heating system	High	High	Locally high Nationally low
National electricity grid (monopolistic)	Low	High	High
Continental gas pipeline	Low	Low	High
National electricity grid (with independent suppliers)	Low	High	Low
Off-grid 'island'/decentral electricity generation	High	Low	Internally high Externally low

After Bridge, G., Bouzarovski, S., Bradshaw, M.J., Eyre, N., 2013. Geographies of energy transition: space, place and the low-carbon economy. Energy Policy 53, 331–340.

a more complex coordination of supply and demand. This allowed for a vertical disintegration of the production chain, created new stakeholders (such as Independent Power Producers and Bulk Electricity Traders) and paved the way for the creation of a new market of electricity trade with platforms and new products such as futures and options. This unbundling has in many countries changed also the market mechanisms for price building and investment allocations. This immense repercussion on the economics of the power system shifted the whole market organization, and thus the roles and relations of all actors, who in turn again influence the design of future energy policies.

In this line of thought, also technological development remains a major driver for the emergent geographies of energy also under relational perspectives. It has for example deeply influenced the environmental debate, where for a long time negative externalities have been relocated spatially, temporally and between social groups. With a shift to integrate such externalities, end-of-pipe-technologies of fossil-fuel based power production are being (partly) replaced by integrated system approaches of green houses gases (GHG) emission mitigation, including the efficiency of (individual) mobility, homes, and household appliances (UNFCCC, 2013).

Also, economic geography has scrutinized the technology markets from a relational approach, as these technologies, e.g., solar energy, have created industrial clusters and regions of research and production and, recently, with a reduced pace of innovation, a massive shift of production to low-wage countries. As for another example, technological progress may be subdued to a strong path-dependency: in centralized and mass-production oriented systems, technological change demands a coevolution of production and demand side. This is clearly visible in the automotive industry where a shift toward electrical cars or – even more revolutionary – to hydrogen is retarded by a technological lock-in of the coupling of the existing car fleet with the dominant (fossil fuel) supply networks.

Spatial images and narratives: As we have seen, energy policy moderates between – and sometimes dominates – the energy production chain on the one hand, and the larger societal framework and environments of the energy system on the other. This includes an understanding of what may be called a 'double impact of space,' as societal decisions form the spatial structure of energy systems consciously or unwillingly, and – reciprocally – the spatial structure of the systems (pre-) determines the options for further societal choices. This figure of thought, however, remains rooted in approaches of physical space, as it tends to assign an active power or action to physical space and its conditions. Therefore, geography's understanding of energy systems has moved beyond and complemented this by approaches such as the construction of spatial images. The allocation of large infrastructures for example is based on the narratives of the different actors regarding the affected space: in the case of hydropower megaprojects in Northern Canada, it is depicted as 'nordic desert' and thus practically 'empty' space from promoters of the project, whereas the opponents refer to 'mother nature' as an imagination of the values of indigenous people and of nature itself (Hamhaber, 2004).

Social networks and interaction: Even more consequently, constructivist approaches not only explain the respective actors' images and definitions of, in short, their perspectives on energy systems in the process of societal negotiation. They also allow it to map the actors themselves within their respective social spaces, including their socially defined barriers, their options and their freedom to decide and to act. Active strategies of territoriality, regionalization, and scaling have been applied to explain energy-system driven conflicts as processes of transgression, as exclusion and inclusion strategies and efforts to gain and prohibit access to energy and to the participation in strategic societal decisions on future energy systems (Hamhaber, 2004). Obviously, there are many more options for a political science approach toward negotiations and conflict about energy systems, be it on renewable energy infrastructure such as wind parks, grid expansion, or, more traditionally, on oil production and pipelines. Certainly, not in all cases, social science approaches of large theoretical reach have to be applied, there is, however, a demand for theoretically informed approaches to analyze and explain energy systems and their dynamics.

Geographies of Energy Systems and Their Transition

The dimensions of energy systems allow for selective and focused research, often from the perspective of specific geographical subdisciplines. Going beyond, more integrated perspectives describe and explain the complexity of energy systems, their spatialities and their transitions.

Geographical Energy System Analyses

Geographical analyses of (commercial) energy system represent the current status of the resources potentials of a given region, the supply structures along the production chain through the primary, secondary, and tertiary sector, and the specific demands for energy services, with their technological, economic, social, and environmental dimensions, in order to allow for an integrated view of a regionally defined energy system. Here, energy geography investigates the global energy flows and their specific repercussions with social and environmental spheres, including a global energy and climate governance. On continental levels, the continuing integration of

markets (such as the network based natural gas and electricity markets in Europe or the regional efforts for grid connections in South-East Asia or in Eastern Africa) are understood as scalar shift and thus one form of energy system transitions (see below). However, most trans-boundary cooperation to day still is in the making and rather limited with the notable exception of the supranational energy policy of the European Union.

Thus, the national level still remains as the most relevant scale for energy system analyses as most policy and economic decisions refer to that spatial reference, at best avoiding the risk of falling back into an understanding of nations as territorial 'container.' Such analyses incorporate demand sectors, namely traffic, industrial processes, and buildings with their energy service demands for mobility, high and low temperature heat, light, and information and communication technologies (ICT), in their historical and current structure as well as with forecasts and scenarios, based on demographic, economic, and further indicators. Regarding national scale energy systems, the territorial frame dominates the debates: therefore, concepts of autarky and import/export energy flows gain massive relevance to policies, economics and also social issues. Thus, some countries such as Mexico or Ecuador are net exporters of crude oil, but import oil derivates. This may put a massive burden on national budgets especially when combined with energy subsidies. Obviously, social and political implications are included in such energy system analyses. Also, spatial repercussions of forecasted trends or policies are in the focus of the geography of energy. These include land use shifts related to altering primary fuel mixes.

Energy Landscapes and Energy Regions

Other than the territorial approach to energy systems, another line of argument follows the line of 'energy landscapes' as an integrated view of a region predominantly determined by energy resources exploitation, transformation and/or use (e.g., the Ruhr valley in Germany, sometimes also referred to as an 'energy region').

Here, fossil fuel landscapes may be distinguished from landscapes of renewable energies (Zimmerer, 2011), as they impose themselves on space in fundamentally different ways, and – finally – find expression in physical structures. As the success of fossil fuels is based on their high energy density, their local impact in different steps along the production chain can be massive, with open-pit exploitation literally eating away landscapes and geo-engineering creating new, post-mining ones. Also, urbanization during industrialization was largely determined in its location and development by the availability of (fossil) energy resources (see agglomerations based on coal mining in central Europe and England, Sieferle, 2001). Similarly, the current structural change after deindustrialization partly was triggered by those resources becoming exhausted or noncompetitive. In that context, some regions today reinvent themselves as postindustrial energy regions in reinterpreting their industrial heritage as cultural part of a new regional definition.

In comparison, renewable energies are largely dispersed in space and thus contribute to landscape transformation by new infrastructural installations. Transitions in agriculture and forestry triggered by energy transitions are partly massive. These are widely debated in geography of energy in regards to their impact on existing (cultural) landscapes, these impacts being one of the major reasons for conflicts about widespread installations of e.g., wind power or pumped storage systems or the impact of large-scale biofuel production in tropical rainforests. Finally, there may also be emergent 'landscapes of energy consumption' (for detailed examples see Zimmerer, 2011).

As this label, however, includes assessments of energy poverty or regulation regimes, the term 'landscape' is partly used in a traditional, integrated way as the 'character' of a given region (an energy landscape in the sense that it was or is determined predominantly by energy systems), and partly in a metaphorical sense ('Electricity Landscape' for the sociopolitical regime of the power sector of a country). Thus, the term 'energy landscape' often seems more like a simplistic ordering principle or rather an oscillating label than a scientifically sustained and fully elaborated spatial terminology (a comprehensive debate of the term landscape/'Landschaft' and its explanatory value was at the core especially of German geography).

The Geography of Energy (System) Transitions

Adding a distinctive temporal perspective leads to the analysis of transitions of energy systems, tackling the current paradigmatic shift of whole nations, of regions and also small-scale communities from fossil to postfossil energy systems. Bridge et al. (2013) define such energy transitions as "fundamentally a geographical process that involves reconfiguring current spatial patterns of economic and social activity." It has to be mentioned, however, that "there is no consensus on a desired end state" of such transition (Bridge et al., 2013), as the concurrent transition from traditional fuels to modern energy systems as depicted in rural electrification in developing countries is a similarly deep transformation and – according to the introducing arguments – also a paradigmatic change.

Such transitions are part and parcel of a profound restructuring of the societies affected. Post-oil societies that were deeply dependent on oil exports, mainly in the Gulf region, currently reinvent and redesign their national economies toward service and financial industries, tourism and airlines. Other former export nations and soon-to-be importers like Indonesia and Mexico have already diversified their economies and thus may be seen as half-way between postoil and postfossil societies. At the same time, some nations such as Angola, Ghana, and Ruanda currently experience the impacts of becoming oil producing countries with all related economic benefits and social and environmental challenges, including the deep disruptions of their social fabric labeled as the curse of the resource.

Industrialized nations on the other hand are challenged by the quest for postfossil societies, the total transformation of their energy systems in a shift from fossil-fuel import dependency to renewables and highly efficient energy systems. This becomes even more demanding when combined with a decision for a fast nuclear exit. Such transitions always come with disruptions, conflict potentials and negotiations on objectives, measures and consequences, in short, the distribution of costs and benefits that are specific for the system under transition. Thus, in Germany, as one of the wealthiest countries on the

globe and despite a large majority supporting the transition, debates still proliferate about misallocations of subsidies for biofuels and photovoltaics (PV), the costs and affordability of electricity for poor people, the equity of benefits distribution between companies, homeowners and tenants, or the obligations of market participants to cover costs of the transition. Given this complexity, is appears disconcerting that the International Energy Agency (IEA, 2012) claims the need for all nations to follow such a model of deep energy transition in order to mitigate greenhouse gases to reach the 2 °C target to prevent deep climate change.

Conclusion

The Geography of Energy in most reports of the international institutions involved in energy and/or climate change still is built upon a traditional Euclidian conceptualization of space. More often than not, reporting remains based on national statistics, understanding the energy system from a locational and traditional territorial viewpoint, and depicting energy facts in tables and maps supporting this view (IEA, 2012). Here, geographical analysis can contribute to a comprehensive understanding of energy systems, as it allows for a customary description of energy systems following accepted conventions, where the production chain, the scales and the relevant dimensions of energy systems remain valid as ordering principles.

Building upon this, and going beyond, examining the new spatialities of energy systems carries more explanatory value, as they do not put (physical) space in the center of analysis, but rather the interrelations, contextualities, interests, etc. In short, such analysis may unveil deeper drivers for the multiple spatial structures and processes that constituted the old and push forward the new emergent geographies (pl.) of energy and provide an integrated, systemic view of how societies negotiate, define and create their energy systems. The discipline may help to overcome bounded perspectives, and thus also normatively contribute to the transitions toward more sustainable geographies of energy.

See also: Energy and Society; Industrial Geography; Oil Age: Energy in Transition; Political Economy of Renewable Energy; Political Geography; Space and Social Theory in Geography.

Bibliography

Bathelt, H., Glückler, J., 2011. The Relational Economy. Oxford University Press, Oxford, p. 298.

Bridge, G., Bouzarovski, S., Bradshaw, M.J., Eyre, N., 2013. Geographies of energy transition: space, place and the low-carbon economy. Energy Policy 53, 331–340.

ter Brugge, R., 1984. Spatial structure in relation to energy production and consumption. Tijdschrift voor Economische en Sociale Geografie 75 (3), 214–222.

Brücher, W., 2009. Energiegeographie – Wechselwirkungen zwischen Ressourcen, Raum und Politik. – Studienbücher der Geographie. Borntraeger, p. 280.

Calzonetti, F.J., Solomon, B.D., 1985. Geographical dimensions of Energy. GeoJournal 5, 516. Dordrecht.

Chapman, J.D., 1989. Geography and Energy. Commercial Energy Systems and National Policies. Longman, Harlow, p. 260.

Clark, J.G., 1990. The Political Economy of World Energy. A Twentieth Century Perspective. Wheatsheaf Harvester, Hempstead, p. 397.

Geni, 2013. Global energy network institute; global renewable energy resources maps. http://www.geni.org/globalenergy/library/renewable-energy-resources/globalmaps.shtml (accessed 04.12.13.).

Georgescu-Roegen, N., 1992. The crisis of resources, its nature and its unfolding. In: Daneke, H. (Ed.), Energy, Economics, and the Environment. Lexington, pp. 9–24.

Hamhaber, J., 2004. Streit um Strom – eine geographische Konfliktanalyse New Yorker Elektrizitätsimporte aus Québec. In: KGA 84. University of Cologne, p. 131.

Hamhaber, J., 2010. Humangeographische Zugänge in der geographischen Energieforschung: Vom euklidischen Raum zu sozial konstruierten Raumbezügen. In: Schuessler, F. (Ed.), Geographische Energieforschung. Strukturen und Prozesse im Spannungsfeld zwischen Ökonomie, Ökologie und sozialer Sicherheit. Schriften zur Internationalen Entwicklungs- und Umweltforschung 27. Lang, Frankfurt, pp. 9–19.

Harvey, D., 1996. Justice, Nature & the Geography of Difference. Blackwell, p. 468.

IEA, 2012. World Energy Outlook 2012. International Energy Agengy, p. 690.

Lovins, A.B., 1977. Soft Energy Pathes. Toward a Durable Peace. Harper, New York, p. 231.

Manners, G., 1964. The Geography of Energy. London, p. 222.

Nexus, 2013. The Water, Energy & Food Security Nexus platform. Background Paper 'Understanding the Nexus'. http://www.water-energy-food.org/en/home.html (accessed 03.12.13.).

Sevette, P., 1976. Géographie et économie compare de l'énergie. Collection Connaissance des problèmes énergétiques. Université des sciences sociales, Grenoble, p. 272.

Sieferle, W.-P., 2001. The Subterranean Forest – Energy Systems and the Industrial Revolution. White Horse, Cambridge, p. 230 Available at: http://www.environmentandsociety.org/sites/default/files/key_docs/sieferle.pdf.

Soja, E., 1989. Postmodern Geographies. The Reassertion of Space in Critical Social Theory. Verso, New York, p. 266.

UNFCCC, 2013. United Nations Framework Convention on Climate Change. Clean Development Mechanism. http://cdm.unfccc.int/about/index.html (accessed 04.12.13.).

Zimmerer, K.S. (Ed.), July 2011. Geographies of energy. Annals of the AAG 101 (special issue).

Relevant Websites

Energypedia, 2013. Sustainable Enegry for All Information Website. https://www.energypedia.info/wiki/Main_Page (accessed 04.12.13.).

GENI, 2013. Global Energy Network Institute. http://www.geni.org/index.html (accessed 04.12.13.).

NEED, 2013. National Energy Education Development Project; Maps of the U.S. with Energy Data. http://www.need.org/maps (accessed 04.12.13.).

UNEP, 2012. Sustainable Energy for All. http://www.se4all.org/ (accessed 04.12.13.).

WBCSD, 2013. Business Perspectives and Solutions for a Sustainable World. http://www.wbcsd.org/home.aspx (accessed 04.12.13.), see e.g., Access to Energy. http://www.wbcsd.org/accesstoenergy.aspx (accessed 04.12.13.).

Engineering Studies

Gary L Downey, Virginia Tech, Blacksburg, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article highlights the growing need to understand the situated commitments and practices of engineers and engineering. It asks: What is engineering for? What are engineers for? Much research challenges images of normative holism and technical–social dualism. The first equates engineering practices with material advancements for humanity as a whole. The second demarcates technical contents of engineering work from social practices that engineers participate in as people. Critical analyses examine images of engineers serving a profession, private industry, and the nation, state, or country, as well as producing technology through design and protecting gender, racial, heterosexual, and developmental privilege. Each delineates pathways for critical participation.

We live in highly engineered worlds. Engineers play crucial roles in the normative direction of localized knowledge and social orders. This article highlights the growing need to understand the situated commitments and practices of engineers and engineering. It asks: What is engineering for? What are engineers for?

Drawing from a diverse arena of research, teaching, and outreach, this article raises awareness of how engineers imagine themselves in service to humanity, and how their service ideals impact the defining and solving of problems with multiple ends and variable consequences. It does so by examining relationships among technical and nontechnical dimensions, and how these relationships change over time and from place to place. Its researchers often are critical participants in the practices of education and work that they study.

Engineering studies scholars conduct historical, social, cultural, political, philosophical, rhetorical, and organizational research on engineers and engineering, paying particular attention to normative directionality in engineering epistemologies, practices, identities, and outcomes. Areas of concern include engineering formation, engineering work, engineering design, equity in engineering (gender, racial, ethnic, class, geopolitical), and engineering service to society.

This article thus pursues three related missions: (1) advance critical analysis and understanding of engineers, engineering, and outcomes of engineering work; (2) help build and serve communities of researchers and learners in engineering studies; and (3) expand critical participation in engineering education, research, practice, policy, and representation as important practices of knowledge production and scalable scholarship.

Building Engineering Studies

During the 2002 joint meeting of the Society for Social Studies of Science (4S) and the Society for History of Technology (SHOT), a gathering of 19 attendees expressed interest in establishing a scholarly network devoted to engineering studies. In the 1989 review essay, "The Invisible Engineer: How Engineering Ceased to Be a Problem in Science and Technology Studies," Gary Downey et al. (1989) documented how researchers in sociology, philosophy, and history had treated engineering and engineers as intellectually uninteresting, albeit for different reasons. Things had changed by 1994. In the more optimistic essay, 'Engineering Studies,' Downey and Juan Lucena traced expanding work by social scientists and humanists on questions of engineering knowledge, technical work, gender in engineering, and scholarly reflections by engineers on engineering (Downey and Lucena, 1994).

Despite the significant amount of available research, engineers and engineering still had limited visibility in the science and technology studies (STS) literature in 2002, except for playing cameo roles, serving some other purpose. One way the word 'engineering' frequently appears to this day is as a metaphor for the contingent work of construction. John Law's (1987) concept of heterogeneous engineering is not about engineers and engineering but the construction of technology. Another way is as a substitute or equivalent term for technology. Gabrielle Hecht, a prominent technology studies scholar, argued in 2005 (Valderrama et al., 2005: p. 13) against establishing a field of engineering studies because it demanded understanding how engineering might be different epistemologically from other forms of technological knowledge. This is a fair question. What if the answer is that engineering is more than a form of technological knowledge?

By 2002, many researchers were finding engineering and engineers to offer a significant collection of research sites for investigating forms of knowledge and material practices with normative commitments and directionality. The issue of normativity is central to studies of engineers and engineering. In 2008, the engineering studies scholar Ronald Kline (2008) argued that Edwin Layton's (1986[1971]) *The Revolt of the Engineers* became a classic that is still widely read because it reported and accounted for struggles between progressive and conservative engineers in the early twentieth century. In the formative work *Engineering the Revolution*, Ken Alder (1997: p. 60) found engineers across eighteenth-century France seeking "to describe quantitatively the relationship among measurable quantities, and then to use these descriptions to seek a region of optimal gain." What sorts of commitments live in optimal gain through engineering?

Despite growing momentum, many scholars in STS and history of technology found it difficult to advance research projects on engineers and engineering because academic

departments, publication outlets, grant agencies, and even professional societies continued to place primary emphasis on the terms 'science' and 'technology.' The work of building a networked arena of research in engineering studies started with the premise that engineers and engineering can and should be analyzed as central rather than peripheral actors and practices in worlds of science and technology. The International Network for Engineering Studies (INES) was established in Paris in 2004 at the joint meeting of 4S and the European Association for Studies of Science and Technology. Imagining a network whose participants had primary allegiances to many different organizations, INES would help advance research, teaching, and outreach in engineering studies.

Beginning in 2006, INES held workshops in Blacksburg, Virginia; Lisbon, Portugal; Grafton, New York; Cleveland, Ohio; and Copenhagen, Denmark. Also in 2006, INES members began organizing INES-affiliated paper sessions at 4S and SHOT meetings. The INES Web site hosts a large bibliography of research in engineering studies. In 2009, Taylor & Francis/Routledge began publishing *Engineering Studies: Journal of the International Network for Engineering Studies*. Morgan & Claypool began publishing the engineering studies series *Global Engineering* in 2011. In 2012, the MIT Press began publishing the capstone book series *Engineering Studies*.

INES emerged in the company of seven related initiatives in the humanistic and social science study of engineers and engineering:

1. In 2000, members of the Liberal Education Division of the American Society of Engineering Education began to respond to new accreditation criteria for engineering programs with research and learning practices that would go beyond bringing engineering students to liberal education and highlight the critical participation of liberal education practices within engineering education. Acknowledging a significant expansion of participation by members interested in critical participation within engineering education, members of that division changed its name in 2011 to the Liberal Education/Engineering & Society Division.
2. In 2002, in Europe, an international collection of historians began to organize meetings and publications for rethinking the history and current status of engineering formation and practice.
3. In 2004, scholar-activists interested in engineering and social justice began to establish a series of paper sessions, workshops, and conferences; the Morgan & Claypool book series *Engineers, Society, and Technology* in 2006; the *Engineering, Social Justice, and Peace* network in 2010; and the *International Journal of Engineering, Social Justice, and Peace* in 2011.
4. In 2005, historians of engineering in SHOT who were interested in reestablishing linkages to engineering education and training established the special interest group Prometheans.
5. In 2006, in a series of workshops and book collections, philosophers affiliated with the Society for Philosophy of Technology began mapping out a philosophy of engineering and its relations to the philosophy of technology. In 2009, Springer began publishing a series on Philosophy of Engineering and Technology.
6. In 2009, after several years of producing a semiannual publication dedicated to engineering studies, Chinese philosophers and social scientists launched the *Journal of Engineering Studies*. Across the country in which the main word for engineering means 'large technological projects,' the domain covered by engineering studies includes not only trained engineers, but everyone else involved in and affected by large technological projects.
7. Finally, an unprecedented scale of reform activities in engineering education across the planet during the 1990s and early 2000s stimulated the emergence of a scholarly field of engineering education research from within faculties of engineering. The field is built on normative commitments to change engineering education, on the premise that educators will be persuaded to alter pedagogy and curricula only when presented with rigorous research demonstrating better learning outcomes through alternative practices. Researchers in the American Society for Engineering Education, European Society for Engineering Education, and Australasian Association for Engineering Education have taken the lead, especially through the *Journal of Engineering Education*, *European Journal of Engineering Education*, and *Australasian Journal of Engineering Education*.

While many participants in these latter initiatives also contributed to the formal development of INES and engineering studies, many also did not identify their scholarship in the first instance with the label 'engineering studies.'

What Is Engineering for? What Are Engineers for?

Making visible the pursuits of optimal gain and the commitments that live in them requires one to confront the question: What is engineering for? Since one can find normative commitments in the personal orientations and service ideals of engineers, inquiring into what engineering is for necessarily involves examining the identities of engineers. It means that work in engineering studies pays much greater attention than other fields of STS to the formation – the education and training – of practitioners. It also requires asking: What are engineers for? (Downey, 2009).

Researchers in engineering studies regularly confront a dominant image of engineering work as a contribution to human advancement as a whole. Leaders who promote engineering throughout the world have long emphasized how the profession contributes directly to human progress. Although they look through different lenses and pursue variable objectives and material visions, they have tended to assert an equivalence between the technical contents of engineering practice and material benefits they achieve for a common humanity. The most prominent feature of these normative projects is their holism: as though the making of engineers produces people who contribute to human advancement as a whole (Downey, 2012).

The normative commitment to holism in engineering has far-reaching implications. It grounds what Wendy Faulkner (2007) has called technical–social dualism. Technical–social dualism is the sharp division engineers draw between the technical contents of their work and the social practices in

which they participate as people. Engineers routinely portray the core of engineering work as wholly technical in content, built on the engineering sciences, and culminating in technological designs. Both as individuals and as a collectivity, engineers learn to encounter and engage with flows of experience as endless sources of technical problems to solve, seeking optimal gain through technological design on behalf of humanity as a whole.

One implication of technical–social dualism is that distinctions among different fields and disciplines of engineering become functional technical differences that complement one another. This applies, for example, to the fields of civil, mechanical, electrical, and chemical engineering, as well as to the disciplines of mechanics, circuit theory, thermodynamics, and vibration analysis. While material practices performed on the job may range from design to manufacturing to sales, technical learning in the engineering sciences provides the core preparation for all of them.

The output of the engineering sciences that feeds into technological design is especially significant. Until fairly recently, engineers have been relatively free to claim jurisdiction over the design and development of new technologies, without competition from other knowledge workers. The absence of competition made it easy for both engineers and nonengineers to equate engineering with technology, frequently to the point of confusion. Engineers design technologies, and the making of engineers is always about the making of technologists. The image authorizes narrowness in the making of engineers and performance of engineering. Since quality engineering work produces technological outcomes that benefit humanity, engineers must be careful only to make sure that nothing interferes with the high-quality technical practices they learned in school and in subsequent training.

Much work in engineering studies begins, both explicitly and implicitly, with critical reactions to normative holism in engineering formation and practice. Most work, in fact, calls attention to the limitations of normative holism by demonstrating how it is inconsistent with actual features of engineering formation and practice. Practitioners of engineering studies then try to make visible the complex sociotechnical relationships that normative holism hides.

Engineers typically make technical–social dualism a basic assumption, an ontological given, for understanding, characterizing, and performing their work. Taking a practice as given grants it normative authority. As a result, research that examines relationships between the so-called technical and the nontechnical dimensions of engineering practice necessarily takes on normative contents of its own. It gains directionality, or assigned value, in relation to the image and practices of normative holism.

Even simple descriptions of relationships between technical and nontechnical agencies in the making of engineers raise the possibility that practices of engineering teaching, learning, and work might have been otherwise, could become otherwise, and, hence, possibly should be otherwise. Indeed, when one diligently pursues the fundamental questions – What is engineering for, and what are engineers for? – a resolute and narrowly defined commitment to normative holism can appear at best delusional, and at worst conspiratorial. It hides other normativities.

Critiques of engineering and engineers thus entail an openness to potential practices of critical participation. The scholarly work of conceptualizing and enacting critical participation is at least as difficult, if not more so, than critical analysis alone. It is knowledge production that requires accepting engineers, makers of engineers, future engineers, and advocates of engineering as audiences, along with fellow scholars in engineering studies. It involves calibrating theoretical and methodological agendas to include the possibility of nominating and scaling up, or achieving the status of givens, of alternative images and material practices.

Critical participation cannot happen without critique. Yet limiting research to critique risks limiting engineering studies to armchair analysis, with the hope that fellow scholars will embrace its virtuosity and then facilitate its diffusion to outside publics. It thus risks limiting itself to what the social critic Ivan Illich sarcastically characterized in 1968 as 'good intentions.' His argument was "To hell with good intentions" (Illich, 1968 [1990]).

A common scholarly competition in engineering studies is over how to formulate and scale up alternative images and practices by pairing critique with critical participation in such a way that makes visible what technical–social dualism hides and which assesses the implications of granting it more prominence. Neither the knowledge contents nor the normative contents of such work can be specified in advance. The competition is friendly. As long as normative holism and technical–social dualism remain invisible as givens in the making of engineers and engineering work, engineers will have no reason to reflect critically on either their commitments or the sociotechnical outcomes of their work.

The balance of this article briefly introduces a range of critiques of normative holism and technical social–dualism. It also highlights the possibilities generated by likely pathways for critical participation, and the difficulties posed by likely challenges.

For a Profession?

A persistent and vexed issue, especially in the United States and the United Kingdom, has been the tendency of engineers to claim the status of an autonomous profession. The claim itself is understandable. Embracing human progress as a whole grants the work of engineers a special kind of autonomy that seems analogous to holistic service in the classical professions of law, medicine, and divinity. Professionals serve the community as a whole. The difficulty with working engineers has been that they largely do not fit the image. This raises difficult questions for the project of building engineering formation around normative holism.

During the 1960s and 1970s, American sociologists struggled unsuccessfully to theorize the professions in a way that would include engineers. The problem was that engineers tended to work within large organizations. Separate career ladders for engineers were unpopular, and success for more elite engineers entailed success within the organization. Also, as recently as the mid-1950s, nearly one-tenth of the U.S. engineering labor force of 500 000 was unionized, with organizational concerns emphasizing salary levels and work

conditions. For American sociologists, examining engineers more closely actually contributed to a more general disciplinary shift to organizational studies.

In the mid-1980s, Robert Zussman (1985) unsuccessfully tried to revive sociological interest in engineers by examining them through the lenses of career and class rather than profession. The normative commitment of engineers to career mobility by changing jobs, he argued, could be traced to their middle-class status. But this move did not overcome the continuing problem that engineers faced in their effort to claim autonomy and normative holism for their profession while pursuing advancement within organizations.

Some Anglo-American historians of engineering implicitly accepted the assertion that engineers serve humanity as a whole by examining the organization of professional societies and professional institutions and their recruiting initiatives among students. Many historians, however, have long questioned the role of normative holism in the formation of the engineering profession precisely by calling attention to the way engineers were positioned within organizations. A few elite civil engineers in the late nineteenth-century United States did achieve a public reputation bordering on heroic status. The vast majority of U.S. engineers, even civil engineers on commission, had almost always been supervised within, while serving the interests of, governing organizations. Also, would-be professional engineers coexisted in the United States with a much larger number of nonprofessional engineers. Edwin Layton's *Revolt of the Engineers* traced a transitional process through which the leaders of professional engineering societies tended also to become leaders of industry. Peter Meiksins (1988) later showed that Layton's account masked struggles of rank-and-file engineers whose concerns were less about professional recognition and autonomy than humane employment and chances for promotion.

Taking a lead from the engineers' commitment to serving humanity as a whole, applied philosophers have devoted enormous effort to interrogating ethical tensions between normative holism and organizational loyalty. Early contributions explored engineers' rights and responsibilities as employees, examined the role of ethics training in engineering curricula, and assessed formal codes of ethics. I cannot do justice in this article to the evolution of research on engineering ethics over the past three decades. See Mitcham (2010) and Barry and Herkert (2014) for starting points.

The bottom line is that engineering studies scholarship has challenged the implication of normative holism that engineers are or should be treated as autonomous professionals. Explicitly positioning themselves within engineering studies, sociologists Carroll Seron and Susan Silbey (2009) showed that the U.S. accreditation process for engineering curricula enacts an instrumental logic that restricts the discretionary judgment characteristic of professions. Professional focus amounts to making sure engineering students master the instrumental logic.

Research in engineering studies that draws on the image of profession may challenge engineers to look past the bounded considerations of employees to consider identities focused on broader commitments. It points researchers in the direction of supporting or helping to create courses, curricula, and institutions devoted to making visible the involvement of engineers in issues such as climate change, water supply and usage, energy supply and usage, and so on. Critical participation of this sort can help students recognize and reflect on normative commitments implicitly built into existing curricula, e.g., that a key future responsibility will be to support the private interests of an industrial employer.

At the same time, work that judges engineering practice in relation to an image of the profession, whether explicitly or implicitly, may limit the ability of curricular interventions to move beyond the periphery of the new course, specialized track, or voluntary membership organization. The image of a profession serving humanity as a whole does not fit much of what engineers actually do.

For Private Industry?

That engineers in, especially, Anglo-American contexts have accepted private industry as the dominant venue for their work has been a primary source of interest and concern for engineering studies research on engineering formation. Monte Calvert (1967) famously attributed the bonding of mechanical engineers to industrial organizations to a contingent shift in educational practices from 'shop culture' to 'school culture.' This shift greatly expanded attention to formal education, especially at land-grant institutions. While mechanical engineers had been elite apprentices engaged in on-the-job preparation for leadership positions, they became tertiary school graduates prepared for entry-level positions and armed with a desire to move up the corporate ladder.

Are engineering students preparing to be servants of industrial corporations? David Noble (1977) put the issue most starkly in a classic Marxian study loaded with evidence about the emergence of engineering formation in the United States. He argued that normative holism was essentially false consciousness, because engineering education served as "a major channel of corporate power" (p. 47) by providing the "immediate manpower needs of industry and the long-range requirements of continued corporate development" (p. 170). Its output was a "domesticated breed" (p. 322) that "convinced themselves that they served the interests of society as a whole … [but] in reality served only the dominant class in society" (p. 324).

The influence, if not control, of normative projects from private industry remains a contentious issue in engineering studies. Many reactions have challenged the claim of domestication. How, for example, does one view capitalist relations of production as deterministic while also calling attention to historical contingencies that point engineers in a variety of directions (Meiksins and Smith, 1996)? Studies based in the United Kingdom have wrestled with such issues as employer scorn for formal engineering education, relative disinterest by government agencies, tension over the continuing role of craft training, and workers managing dual commitments inside and outside of industrial organizations. A surge of engineering studies scholarship on engineering work carried forward the normative questions and their implications for technical–social dualism. Anderson et al. (2010) documented the frustrations of engineers who tried to bring technical–social dualism onto the job and then

found that their actual responsibilities did not fit it. Gainsburg et al. (2010) found that workplace knowledge and commitments in an engineering occupation may derive more from on-the-job experience than from prior formal education. Meanwhile, Will-Zocholl and Schmiede (2011) argued that globalized work that was reliant on information technologies tended to marginalize purely technical tasks.

Taking the relationship between the formation of engineers and private industry as one's primary objective of study has different implications depending upon how one theorizes that relationship. David Noble's analysis of engineers as a breed domesticated by the dominant class depended upon his view that capitalism invariably produces exploitation of labor. Studies like Noble's which assume that class inequalities are structured by capitalism channel critical participation into revolutionary action; action that promises to transform engineers into an entirely different breed altogether, not to mention all of social life as well. The position effectively accepts resolute pessimism about changing the existing state of affairs.

Any alternatives ultimately require questioning the Marxian labor theory of value. Peter Lundgreen (1990), described below, articulated one prominent project that focused on the state rather than capitalist industry. Other approaches might build on the view that the labor theory of value insufficiently considers the range of normativities in industrial production, as well as the distinct roles of different technological agencies. Pursuing such agencies would require defining pathways for critical participation within the firm. In any case, Noble's thorough and consistent analysis continues to throw down a gauntlet that any research or attempt at critical participation through engineering studies ignores at its peril.

For Technology through Design?

An integral component of the engineer's commitment to humanity as a whole is that engineers learn their disciplines in order to design technologies. It is their real work. The flow is linear and the emphasis is technical. Much scholarship in engineering studies has problematized this flow by complicating each step.

Philosophers of technology have long wrestled with how engineering and the applied sciences manage distinctions between epistemic sources and technological outputs. Adopting normative stances that epistemic categories were freed from context, philosophers of technology in the 1960s debated whether the most important epistemological distinction in engineering was between pure and applied science or applied science and technology. One key outcome of this debate was the rise of interest in a philosophy of engineering.

Philosophers of engineering have tended to start with epistemics and then identify links to broader social projects that contribute to the actual identities of engineers. Walter Vincenti (1990) did not venture beyond epistemics when he posited universal, context-free categories of design knowledge, including fundamental design concepts, criteria and specifications, theoretical tools, quantitative data, practical considerations, and design instrumentalities. Positing context-free categories facilitated a continuing, implicit commitment to normative holism. Others maintained that the societal positioning of engineers as technological intelligentsia or of engineering as an occupation required one to plot the epistemological contents of engineering work in relation to a range of social and ethical contents and dimensions.

Questioning the linear thread from the engineering sciences to their jurisdiction over technology led researchers to also examine normativities in engineering design. Larry Bucciarelli (1994) found that engineering designers work in object worlds with distinct material commitments. Many philosophers emphasized that engineering design should be situated in context, introducing extraepistemological issues such as ethical ambiguities; diverse user perspectives; variable legal commitments and implications; and a variety of financial, environmental, and aesthetic considerations. Helpful starting points include Christensen et al. (2009) and van de Poel and Goldberg (2010).

Actor-network theorists examined design practices as themselves sources of broader material projects. John Law and Michel Callon (1988) portrayed practitioners in engineering as engineer-sociologists who become social activists by designing both machines and their social institutional contexts. Dominique Vinck, Eric Blanco, and colleagues (Vinck and Blanco, 2003) showed how engineering design practices organize things and people at the same time.

Finally, some scholars in engineering studies have examined frictions or discontents produced by competing claims of jurisdiction over technology itself. Matthew Wisnioski (2012) traced how the assumption that engineers enjoyed such jurisdiction encountered a crisis in elite engineering circles during the 1960s when technology itself lost its assumed link with progress. Gary Downey (1998) showed how engineering students and practitioners who sought control over computer-aided design technologies in the 1980s found that they also had to yield to them. Rosalind Williams (2002) explored how attempting to control increasingly diverse technologies during the 1990s threatened to produce an 'expansive disintegration' of engineering research and teaching.

The image of engineering in context fuels numerous projects that attempt to foster critical participation in the making of engineers and engineering practice. It is relatively straightforward to demonstrate to would-be engineers that they will have to participate in a variety of social contexts, including relations with employers, fellow employees, clients, professional associations, standards organizations, regulatory bodies, and so on. And making such contexts visible can help students recognize that technical problem solving will be only part of what they do and that they would do well to prepare for participation in a variety of contexts.

A difficulty facing this approach is that it risks leaving untouched the technical–social dualism that normative holism presumes and supports. This difficulty arises when *context* comes to refer to everything that does not have to do with the linear flow of technical practices from engineering science through design to the production of technology. The challenge is to take account of how engineers are already, as Wisnioski puts it following Bruno Latour, contextualizing themselves in their commitment to normative holism and technical–social dualism.

For the Nation, State, Country?

Peter Lundgreen's (1990) contribution to studies of normative projects in engineering resisted the image of the field as a response to the needs or demands of private industry. Lundgreen highlighted the way French engineers, since the eighteenth century, were committed to training through schools and government projects. The emergence of engineering education across Germany during the nineteenth century also could not be attributed simply to private sector needs. By emphasizing agencies of the state, Lundgreen called for research into how practices of supplying engineers go beyond responding to needs or demands, thus distinguishing between commitments in qualifying engineers and in putting them to work.

Enormous bodies of work now exist in histories of the development of engineering in different countries. Detailing what Antoine Picon (2007) called 'local specificities,' these works integrate histories of the formation of engineering with histories of the development of nation-states. Some studies presented engineers as developing distinct engineering 'styles' that 'reflect' or 'reinforce' host national cultures or grammars.

The most common form of analysis is to recount the relative role of state authority in both institutional and curricular developments in engineering education. I can provide here only starting points. Antoine Picon (1992) explained how eighteenth-century engineers overtook architects by acquiring quasi-judicial authority as they pursued technocratic ideals. Bruno Belhoste and Konstantinos Chatzis (2007) reviewed the preparation of state engineers in nineteenth-century France. Contributors to an edited volume by Ana Cardoso de Matos et al. (2009) examined how models for training engineers were disseminated, focusing on the role of key reference schools, and the distinctive normative programs in schools on the European periphery. Such interests in efforts to theorize appropriation, circulation, and transnationality have extended worldwide.

As should be clear by now, this article is informed by an approach that critiques normative holism and technical–social dualism. However, this approach is not only a matter of unmasking what the holistic and dualistic ideology hides, but is instead a matter of demonstrating multiplicity in the forms and formations of engineering. What has counted as progress or human advancement has varied dramatically across territories and over time. One reason why some scholars identify distinct cultural practices of engineering education and training is that such locates the technical contents of engineering in the same analytical and ontological category as practices of governmental action, political economy, higher education, activities of work, etc. They all become cultural phenomena. However, this point also presents a key challenge. Engineers tend to use 'culture' as a label that covers everything that is not technical, hence not authentic engineering. Attempting to persuade engineers that the technical contents of engineering are cultural may appear to remove technical rigor from engineering judgment and throw up a barrier to significant critical participation.

Documenting the role of the state in the making of engineers could, in principle, motivate critical participation in the operations of the state. The difficulty lies in accepting the multidimensional challenges of participating in governmental action at the national level. Studies of engineering which demonstrate that there are multiple versions of normative holism and technical–social dualism can inform those who educate and train engineers that what they take for granted is highly localized. A longer-term possibility is that of teaching engineering as a matter of both problem definition and solution, positioning every course and pedagogical practice as a normative project that makes some things visible and doable while hiding others (Downey, 2009). A limitation is that persuading engineering instructors to carry the vision of engineering as problem definition and solution to the heart of their pedagogical practices may require something akin to a mass social movement in engineering instruction.

For Gender, Racial, Heterosexual, Developmental Privilege

Numerous critiques of normative holism and technical–social dualism make visible other material normativities in engineering formation and engineering practice. Important areas explored to date include gender, racial, and heteronormativities, as well as issues of social justice in international development.

Extending Sally Hacker's (1989) early work on the masculinization of technology and technologists, Ruth Oldenziel (1999) found that the professionalization of engineering was a project that turned out to be a thoroughly male and middle-class endeavor. Karen Tonso (2007) showed how practitioner identities at an engineering school did not include women. Wendy Faulkner (2009) explored how engineering identities were congruent for men but not for women.

Race and sexual orientation have received less attention. Amy Slaton (2010) examined how conceptions of engineering talent and rigor have supported racial selectivity in U.S. engineering schools. Erin Cech and Thomas Waidzunas (2011) showed how mechanisms of heteronormativity produce experiences of marginalization and isolation for lesbian, gay, and bisexual students.

A fast-growing research area in engineering studies involves the way the involvement of engineering in industrial development has contributed to increased inequality and social injustice. Dean Nieusma and Donna Riley (2010) demonstrated how the priority accorded to final products in engineering for development could not only accentuate existing power asymmetries but also generate wholly new ones. For many more examples, see the sizable collection of volumes on engineering and social justice in the Synthesis Lectures on Engineers, Technology, and Society at Morgan and Claypool Press, edited by Caroline Baillie.

It is essential for engineering students and engineers to understand the workings of power and privilege in engineering, so that they can recognize the limitations in normative holism and technical–social dualism. Engineers and nonengineers who are disempowered or made invisible by dominant images of engineering can find lifelines and encouragement in critical analyses of sources of privilege in engineering. Engineers who learn how they benefit from privileges they did not earn cannot easily assert that engineering always serves humanity as

a whole. Understanding what dominant images hide is the first step in formulating alternatives. A significant challenge lies in carrying critique from the periphery into critical participation at the center. What would it take for routine pedagogy in engineering to make explicit the sources of gender, racial, heteronormative, and developmental privilege? What more must be learned about the barriers posed by normative holism and technical–social dualism in order to alter them?

What Is the Field of Engineering Studies for?

In his 1954 novel *The New Men*, C.P. Snow (1961[1954]: p. 176) blurred the two cultures of science and the humanities when he characterized engineers. "[E]ngineers," he wrote, "… were in nine cases out of ten … acceptant of any regime in which they found themselves." Although it is fiction, this characterization of engineers still stands out because it likely strikes the vast majority of readers as true. The key question it raises is: Why?

If engineers are bred to be wholly domesticated agents of powerful external forces, engineers are weak and their acceptance of the regime in which they find themselves is a kind of surrender. From this point of view, it makes little sense for critical scholars to go beyond critique to seek critical participation. Their work will be co-opted. The situation changes, however, if engineers are in a regime because they accept it; because, to paraphrase Snow, they are mainly interested in making their machines work and the regime facilitates that end. In that case, multiple normativities in engineering are more interestingly situated alongside its multiple epistemological and ontological contents. Initiatives to participate critically in engineering formation and practice also become more than exercises in engagement: they are sites for research, for expanding critical understandings of engineers and engineering. Critical participation in the service ideals and assumptions of engineering generates localized knowledge about the scope, operations, and accountability of engineering.

The questions – what is engineering for, and what are engineers for? – have guided engineering studies work. Yet the dominance of the images of normative holism and technical–social dualism means that no researcher in engineering studies can escape the additional question: What is engineering studies for?

Acknowledgments

The research reported here is based upon the work supported by National Science Foundation under award #DUE-1022898. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation. Thanks to Michael Lynch and the Oxford University Press staff for the invitation and helpful guidance that far exceeded the call of duty. Thanks to Matt Wisnioski for calling the C.P. Snow quote to my attention and for helping with the introduction. This article draws substantially from Downey (2012) and Downey (2014). Material from those publications is used with permission from the copyright holders, Springer and the author, respectively.

See also: Advocacy in Anthropology; Gender and Technology: From Exclusion to Inclusion?; Modernity: Anthropological Aspects; Organizational Culture, Anthropology of; Research and Development in Organizations; Science and Development; Science and Technology, Anthropology of; Science and The State; Technology and Organization; Technology, Social Construction of.

Bibliography

Alder, K., 1997. Engineering the Revolution: Arms and Enlightenment in France, 1763–1815. Princeton University Press, Princeton, NJ.

Anderson, K., Shaw Courter, S., McGlamery, T., Nathans-Kelly, T., Nicometo, C., 2010. Understanding engineering work and identity: a cross-case analysis of engineers within six firms. Engineering Studies 2 (3), 153–174.

Barry, B., Herkert, J., 2014. Engineering ethics. In: Johri, A., Olds, B. (Eds.), Cambridge Handbook of Engineering Education Research. Cambridge University Press, Cambridge, pp. 673–692.

Belhoste, B., Chatzis, K., 2007. From technical corps to technocratic power: French state engineers and their professional and cultural universe in the first-half of the 19th century. History and Technology 23, 209–225.

Bucciarelli, L.L., 1994. Designing Engineers. MIT Press, Cambridge, MA.

Calvert, M.A., 1967. The Mechanical Engineer in America, 1830–1910: Professional Cultures in Conflict. The Johns Hopkins Press, Baltimore, MD.

Cech, E., Waidzunas, T., 2011. Navigating the heteronormativity of engineering: the experiences of lesbian, gay, and bisexual students. Engineering Studies 3, 1–24.

Christensen, S.H., Delahousse, B., Meganck, M., 2009. Engineering in Context. Academica, Aarhus.

Downey, G.L., 1998. The Machine in Me: An Anthropologist Sits Among Computer Engineers. Routledge, New York.

Downey, G.L., 2009. What is engineering studies for? Dominant practices and scalable scholarship. Engineering Studies: Journal of the International Network for Engineering Studies 1, 55–76.

Downey, G.L., 2012. The local engineer: normative holism in engineering formation. In: Christensen, S.H., Mitcham, C., Li, B., An, Y. (Eds.), Engineering, Development and Philosophy: Chinese, American, and European Perspectives. Springer, Berlin, pp. 233–252.

Downey, G.L., 2014. The normative contents of engineering formation: engineering studies. In: Johri, A., Olds, B. (Eds.), Cambridge Handbook of Engineering Education Research. Cambridge University Press, Cambridge, pp. 673–711.

Downey, G.L., Donovan, A., Elliott, T.J., 1989. The invisible engineer: how engineering ceased to be a problem in science and technology studies. Knowledge and Society 8, 189–216.

Downey, G.L., Lucena, J.C., 1994. Engineering studies. In: Jasanoff, S., Markle, G., Petersen, J., Pinch, T. (Eds.), Handbook of Science and Technology Studies. SAGE Publications, Newbury Park, CA, pp. 167–187.

Faulkner, W., 2007. "Nuts and bolts and people": gender-troubled engineering identities. Social Studies of Science 37, 331–356.

Faulkner, W., 2009. Doing gender in engineering workplace cultures II: gender in/authenticity and the in/visibility paradox. Engineering Studies 1, 169–189.

Gainsburg, J., Rodriguez-Lluesma, C., Bailey, D., 2010. A "knowledge profile" of an engineering occupation: temporal patterns in the use of engineering knowledge. Engineering Studies 2 (3), 197–219.

Hacker, S., 1989. Pleasure, Power, and Technology: Some Tales of Gender, Engineering, and the Cooperative Workplace. Unwin Hyman, Boston, MA.

Illich, I., 1968[1990]. To hell with good intentions. In: Kendall, J. (Ed.), Combining Service and Learning. National Society for Internships and Experiential Education, Raleigh, NC, pp. 314–320.

Kline, R.R., 2008. From progressivism to engineering studies: Edwin T. Layton's revolt of the engineers. Technology and Culture 49 (4), 1018–1024.

Kranakis, E., 1997. Constructing a Bridge: An Exploration of Engineering Culture, Design, and Research in Nineteenth-Century France and America. MIT Press, Cambridge, MA.

Law, J., 1987. Technology and heterogeneous engineering: the case of Portuguese expansion. In: Bijker, W., Hughes, T.P., Pinch (Eds.), The Social Construction of Technological Systems. MIT Press, Cambridge, MA, pp. 111–134.

Law, J., Callon, M., 1988. Engineering and sociology in a military aircraft project: a network analysis of technological change. Social Problems 35, 284–297.
Layton, E.T., 1986[1971]. The Revolt of the Engineers: Social Responsibility and the American Engineering Profession. Johns Hopkins University Press, Baltimore and London.
Lundgreen, P., 1990. Engineering education in Europe and the U.S.A., 1750–1930: the rise to dominance of school culture and the engineering profession. Annals of Science 47, 33–75.
de Matos, A.C., Diogo, M.P., Gouzévitch, I., Grelon, A. (Eds.), 2009. Les enjeux identitaires des ingénieurs: entre la formation et l'action (The Quest for a Professional Identity: Engineers between Training and Action). Colibri, Lisbon.
Meiksins, P., Smith, C. (Eds.), 1996. Engineering Labour: Technical Workers in Comparative Perspective. Verso, London.
Meiksins, P., 1988. "The revolt of the engineers" reconsidered. Technology and Culture 29 (2), 219–246.
Mitcham, C., 2010. A historico-ethical perspective on engineering education: from use and convenience to policy engagement. Engineering Studies 1, 35–53.
Noble, D., 1977. America by Design: Science, Technology, and the Rise of Corporate Capitalism. Knopf, New York.
Nieusma, D., Riley, D., 2010. Designs on development: engineering, globalization, and social justice. Engineering Studies 2, 29–59.
Oldenziel, R., 1999. Making Technology Masculine: Men, Women and Modern Machines in America, 1870–1940. Amsterdam University Press, Amsterdam.
Picon, A., 1992. French Architects and Engineers in the Age of Enlightenment. Cambridge University Press, Cambridge.
Picon, A., 2007. French engineers and social thought, 18th–20th centuries. History and Technology 23 (3), 197–208.
van de Poel, I., Goldberg, D.E. (Eds.), 2010. Philosophy and Engineering: An Emerging Agenda. Springer, Berlin.
Seron, C., Silbey, S., 2009. The dialectic between expert knowledge and professional discretion: accreditation, social control, and the limits of instrumental logic. Engineering Studies 1, 101–127.
Slaton, A., 2010. Race, Rigor and Selectivity in U.S. Engineering: The History of an Occupational Color Line. Harvard University Press, Cambridge, MA.
Snow, C.P., 1961[1954]. The New Men. Scribner, New York.
Tonso, K., 2007. On the Outskirts of Engineering: Learning Identity, Gender and Power via Engineering Practice. Sense Publishers, Rotterdam.
Valderrama, A., Arias, R., Gómez, C., 2005. Engineering Studies: State of the Art, Knowledge and Practices. Report to Engineering Sciences, Division of Basic and Engineering Sciences, Natural Sciences Sector. UNESCO, Paris.
Vincenti, W.G., 1990. What Engineers Know and How They Know It: Analytical Studies from Aeronautical History. Johns Hopkins University Press, Baltimore.
Vinck, D., Blanco, E., 2003. Everyday Engineering: An Ethnography of Design and Innovation. MIT Press, Cambridge, MA.
Will-Zocholl, M.C., Schmiede, R., 2011. Engineers' work on the move: challenges in automobile engineering in a globalized world. Engineering Studies 3 (2), 101–121.
Williams, R.H., 2002. Retooling: A Historian Confronts Technological Change. MIT Press, Cambridge, MA.
Wisnioski, M., 2012. Engineers for Change: Competing Visions of Technology in 1960s America. MIT Press, Cambridge, MA.
Zussman, Robert, 1985. Mechanics of the Middle Class. University of California Press, Berkeley, CA.

Relevant Websites

Baille, C. (Ed.), Synthesis Lectures on Engineers, Synthesis Lectures on Engineers, Technology, and Society. Morton & Claypool Publishers. www.morganclaypool.com/toc/ets/1/1 (accessed 15.03.14.).
Downey, G. (Ed.), Synthesis Lectures on Global Engineering. Morgan & Claypool Publishers. www.morganclaypool.com/toc/ges/1/1 (accessed 15.03.14.).
Division of Liberal Education/Engineering & Society, American Society for Engineering Education. https://sites.google.com/site/aseelees/ (accessed 15.03.14.).
Engineering Studies Series, The MIT Press. http://mitpress.mit.edu/books/series/engineering-studies (accessed 15.03.14.).
Engineering Studies: Journal of the International Network for Engineering Studies. www.tandfonline.com/toc/test20/current (accessed 15.03.14.).
Engineering Cultures – Lectures on Engineers in China, France, Germany, Great Britain, Soviet Union/Russia, and United States. http://globalhub.org/resources/11; http://globalhub.org/members/1044/contributions (accessed 15.03.14.).
International Network for Engineering Studies. www.inesweb.org (accessed 15.03.14.).
International Journal of Engineering, Social Justice, and Peace. http://library.queensu.ca/ojs/index.php/IJESJP (accessed 15.03.14.).
Prometheans Special Interest Group, Society for the History of Technology. www.historyoftechnology.org/sigs.html (accessed 15.03.14.).
Philosophy of Engineering and Technology Series. Springer, www.springer.com/series/8657 (accessed 15.03.14.).

English Revolution, The

John Gurney, Newcastle University, Newcastle-upon-Tyne, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by P. Wende, volume 7, pp. 4531–4534, © 2001, Elsevier Ltd.

Abstract

The two decades from 1640 witnessed unprecedented change in England, including civil war between king and parliament, the trial and public execution of the king, and the abolition of monarchy. The term 'the English Revolution' has often been applied to this period, and the revolution has been portrayed as a forerunner of the eighteenth-century American and French revolutions. Many historians and other scholars working on the period have, however, doubted whether a revolution actually occurred in these years. This article explores the range of scholarly approaches to the period 1640–60 and suggests that in the realm of ideas, at least, we can see a revolution having taken place.

On 13 April 1640, an English parliament assembled at Westminster for the first time since 1629, at the end of the 11-year period of 'personal rule' when King Charles I had attempted to rule without the assistance of parliaments. The Short Parliament of 1640 sat for just 3 weeks before it was dissolved, but the parliament which gathered early in the following November was to sit uninterrupted until 1653. While this, the Long Parliament, was in being, England experienced prolonged periods of civil war and unprecedented political change. War between king and parliament broke out officially in August 1642, when Charles raised his standard at Nottingham. The first civil war concluded in 1646 with the defeat of the king's forces, but after a second civil war in 1648 parliament was purged of those MPs who had wished to continue negotiations with the king, and Charles was put on trial and convicted as a traitor to the people of England. In the months following the king's public execution on 30 January 1649, parliament legislated to abolish the monarchy and House of Lords, and England was declared to be a commonwealth and free state, to be governed by a unicameral parliament and council of state. The army, under the command of Oliver Cromwell (1599–1658), launched a bloody conquest of Ireland in the summer of 1649, and in the following year his forces invaded Scotland and, on 3 September, resoundingly defeated the Scots at the battle of Dunbar.

Relations between the army and parliament became increasingly fraught, and on 20 April 1653, Cromwell brought troops to Westminster and expelled the MPs. Power now resided with Cromwell and the army. In June, an assembly, made up of nominated rather than elected members, met at Westminster and soon declared itself to be a parliament. This assembly, which was popularly known as Barebone's Parliament, continued to sit until December, when, as a result of conflicts between its radical and more conservative members, it surrendered power back into the hands of Cromwell. Later that month, Cromwell was named Lord Protector of the Commonwealth of England, Scotland, and Ireland, his powers limited and defined by the Instrument of Government, England's first written constitution, which had been drawn up by some of Cromwell's closest supporters. The office of Lord Protector was to be elective rather than hereditary, and executive power was to be in the hands of a permanent council of 21 members. There was provision for triennial parliaments, and a system of checks and balances was intended to ensure that decision making on the important issues of war, finance, and appointment to senior offices could be shared between Protector, council, and parliament.

In 1655, frustrated by the slow process of godly reform, and in the wake of a series of royalist risings, Cromwell introduced the 'Rule of the Major-Generals,' in which England and Wales were divided into 10 regional associations, each presided over by a major-general. The major-generals were charged with maintaining security, upholding law and order, and advancing godly rule in the provinces. The scheme was to be financed by a decimation tax imposed on former civil war royalists, including those who had made their peace with the regime. Opposition to 'swordsmen and decimators' was a marked feature of the elections to the second Protectorate Parliament in 1656, and the major-generals scheme was abandoned after the militia bill, which would have placed the decimation tax on a permanent footing, was decisively voted down on 29 January 1657. In the following month, parliament presented Cromwell with the Humble Petition and Advice, a new draft constitution which sought to address what were seen as the limitations of the Instrument of Government. Cromwell was offered the crown, and provision was made for a second parliamentary chamber, or 'Other House.' It was not until May that Cromwell refused the offer of the crown, but a modified Humble Petition was accepted, with Cromwell being given the power to nominate his successor. Cromwell died on 3 September 1658, and was succeeded as Protector by his son Richard (1626–1712). His period of rule lasted only until the following May, when the Protectorate was swiftly brought to an end. Friction between the army and parliament had led to the dissolution of the third Protectorate Parliament in April 1659, and to the return on 7 May of the 'Rump' of the Long Parliament; with the Rump re-installed, Richard Cromwell had little option but to resign from office. The Rump remained in power until October, when it was expelled by the army, and there followed 2 months of rule by an army-backed committee of safety. After General George Monck, commander of the English forces in Scotland, declared in favor of parliament, the Rump was allowed to return and it met again on 26 December 1659. In January,

Monck marched into England, and arrived in London the following month. His decision to support the return to parliament of MPs secluded in the purge of December 1648 made the return of monarchy inevitable. The Long Parliament dissolved itself on 16 March 1660, and the Convention which followed invited Charles Stuart, son of the executed Charles I, to return to England as king. By many accounts, the revolution that had begun in 1640 was now at an end.

Historians have long argued over the significance of the events of the two decades from 1640 to 1660 and over whether they can actually be said to constitute a revolution. The term 'revolution' was of course in use in the seventeenth century, but its meanings were rather different from our modern understanding of the term (Snow, 1962; Harris, 2005: pp. 32–37). The events of 1640–60 were more often than not described as 'the Great Rebellion,' 'the Usurpation,' or 'the Troubles.' When more modern, and more precise meanings of the term revolution came into fashion, the term tended to be used to describe the events of 1688–89, when the rule of James II and the Stuart dynasty ended, rather than the mid-century events. For many years it was this 'Glorious Revolution' that was seen to constitute England's sole modern political revolution. The French Revolution of 1789 led to a reappraisal of the nature and significance of England's mid-seventeenth-century conflicts, and it was in France rather than England that 1640–60 first came to be identified as a period of revolution. It is usually thought that the first thinker to call the events of 1640–60 'the English Revolution' was the French politician and historian Francois Guizot (1787–1874), the first volume of whose *Histoire de la révolution d'Angleterre* appeared in 1826 (Guizot, 1826–27). It is, however, known that earlier French revolutionaries, recognizing the striking similarity between events in England in the 1640s and events in France after 1789, were in the 1790s already referring to 1640–60 as 'la Révolution Anglaise' (Hammersley, 2010: pp. 159–161). Guizot, moreover, never saw 1640–60 as a single, discrete revolution, but rather as part of a longer process of revolution which began in 1640 and culminated in the Glorious Revolution of 1688–89. For him, the English Revolution comprised 'three grand periods,' namely 1625–49, 1649–60, and 1660–88. It was, he maintained, only in 1688 that "England achieved the point she aimed at in 1640" (Guizot, 1846).

Karl Marx, despite his strong antipathy to Guizot the politician, was heavily influenced by his interpretation of seventeenth-century English history. Marx, as is well known, advanced the notion that England experienced a bourgeois revolution in the seventeenth century, and maintained that the English Revolution was among the most significant of such events. Both Karl Marx and Friedrich Engels acknowledged the interest and importance of what had taken place in the two decades from 1640, but they did not, as is often assumed, see the period 1640–60 as in itself a major turning point in English history. For Marx and Engels, like Guizot before them, the seventeenth-century English Revolution was a much longer process, one that began in 1640 and concluded only in 1688–89. The latter date was at least as significant as 1640 or 1649: it was then that 'the bourgeois transformation of English society' was accomplished and 'the first decisive victory of the bourgeoisie over the feudal aristocracy' achieved (Marx, 1979: p. 105; Pincus, 2009: p. 34). Eduard Bernstein was following established Marxist usage when he later described the revolution of 1688 as the 'last phase' of the English Revolution (Bernstein, 1930: p. 10).

It was the great Victorian historian S.R. Gardiner (1829–1902) who, in the 1880s and 1890s, was primarily responsible for focusing attention once more on 1640–60. Although he never lost sight of the longer perspective, he came to regard the two decades from 1640 as a period of revolution in its own right. In his Rankean multivolume histories of the period, which laid the foundations of modern scholarship on mid-seventeenth-century England, Gardiner adopted the term 'the Puritan Revolution' to describe the events of 1640–60 (Richardson, 1998: pp. 91–96; Worden, 2001: p. 270). Gardiner was – as his chosen term suggests – aware of the crucial importance of religion to the revolution, but much of his explanatory focus, like that of so many of his predecessors, was on issues of constitutional conflict. Earlier 'Whig' historians, like T.B. Macaulay (1800–59), had argued that the civil wars of the 1640s had been preceded, and brought about, by worsening relations between Crown and Parliament. For Macaulay, the blame lay squarely with the first two Stuart monarchs, James I and Charles I, who were accused of blindly pursuing absolutist policies, shored up by high-flown notions of 'divine right.' James was, for Macaulay, clearly "one of those kings whom God seems to send for the express purpose of hastening revolutions" (Macaulay, 1843: p. 435). It was Macaulay's contemporary Thomas Carlyle (1795–1881) who in his edition of *Oliver Cromwell's Letters and Speeches* (1845) fully grasped the importance of religion to the mid-seventeenth-century conflicts and captured the puritan spirit at the heart of the parliamentarian movement (Worden, 2001: pp. 269–277).

The high standards set by Gardiner were continued in the work of his close friend Sir Charles Firth (1857–1936), the author of major histories of the period and editor of significant memoirs and other collections of primary sources. Much of the focus of research in the first half of the twentieth century, particularly in the United States, was on early-Stuart parliamentary history and constitutional conflict. Among the most significant works to appear was Wallace Notestein's published version of his 1924 Raleigh Lecture, 'The winning of the initiative by the House of Commons' (Notestein, 1924), a work that had a tremendous influence in the field, and which too often came to be read without due attention to the qualifications that hedged the author's arguments about the rise in influence of the Commons (Russell, 1990: pp. 33–34). It was in 1940 that a startlingly new interpretation of the period 1640–60 appeared, in the work of the young Marxist historian Christopher Hill. In his contribution to his edited collection marking the tercentenary of the revolution, Hill boldly announced that "the English Revolution of 1640–60 was a great social movement like the French Revolution of 1789." The old, essentially feudal order was destroyed in a struggle waged by the bourgeoisie and progressive gentry, and 'a new and capitalist social order' was created in its place. The Restoration of 1660 was, he insisted, "by no means a restoration of the old order" (Hill, 1940: pp. 9, 13, 76). Hill's interpretation was a Marxist one, but one that was indebted less to Marx than to recent developments in Soviet

historiography. Just as Marxist and Marxist-influenced scholars in France had come to present the French Revolution as a necessary forerunner of 1917, so the English Revolution could be portrayed as one that anticipated the French and was comparable in its significance. The English had, it seemed, got there first. While Marx and Engels had always recognized the significance of 1688–89 as the point when England's bourgeois revolution reached its successful conclusion – the moment, as Marx remarked, when "Locke supplanted Habakkuk" (Marx, 1979: p. 105) – Hill argued that 1688 was little more than an afterthought, and that the crucial changes which might define a bourgeois revolution had taken place in England by 1660. England's eighteenth-century commercial prosperity, Hill maintained, derived from the Navigation Act of 1651, and this act, together with the sale of royalist lands, the ending of monopolies, the dismantling of the courts of Star Chamber and High Commission, and the abolition of feudal dues for landowners (though not, significantly, for their tenants) all combined to ensure that England's bourgeois revolution should be seen to have occurred in 1640–60 (Hill, 1940: pp. 70–71, 74–79). In his focus on 1640–60 at the expense of 1688, Hill was clearly following the lead set by Gardiner, but Gardiner's Puritan Revolution had now given way to a determinedly Marxist bourgeois revolution. In the years that followed, Hill's thesis shifted somewhat – the revolution was no longer said to have been consciously willed by a rising bourgeoisie – but his argument about the revolution's outcome remained essentially the same: the most significant impediments to capitalist development had, he insisted, all been removed in the revolutionary decades (Hill, 1980: pp. 134–135). The term 'the English Revolution' as a descriptor of the events of 1640–60 soon gained general currency, among Marxist and non-Marxist scholars alike. Only a few scholars held out against the adoption of the term, preferring instead to continue to refer to 'the Great Rebellion,' 'the Puritan Revolution,' or 'the Civil Wars.'

Hill was not alone in the 1940s in advancing a social interpretation of the revolution. In 1941, in the year following the publication of Hill's *English Revolution 1640*, there appeared R.H. Tawney's 'The rise of the gentry, 1558–1640,' which sought to demonstrate that a shift in the balance of property, from the nobility to the gentry, had occurred in the century before the civil war. Tawney's arguments, which appeared to many to confirm the philosopher James Harrington's belief that the civil wars and the breakdown of the old order were linked to long-term shifts in property holding, were for a time highly influential, and were soon supported by the work of the young historian Lawrence Stone. A formidable assault on Stone and Tawney was mounted in 1951 by Hugh Trevor-Roper, who argued that significant sections of the gentry were declining rather than rising in the years before the civil war, and that it was from among this group that a 'country' opposition, which was to rise up in rebellion and eventually defeat the king, drew its most vigorous supporters. Trevor-Roper's arguments were no less speculative than those of his opponents, but his demolition of the 'rising gentry' thesis was supported by detailed research by, among others, the historian J.P. Cooper. The so-called 'gentry controversy' continued on its destructive course for several years before petering out inconclusively, its main protagonists having withdrawn to heal their wounds or having moved on to less controversial areas of research (Richardson, 1998: pp. 116–121).

By the later 1960s, many of the old certainties had gone, without having been replaced by any new consensus. A bold attempt to offer a convincing and holistic explanation of the causes of the English Revolution was provided by Lawrence Stone in 1972. His work drew on research into other great revolutions, and on current sociological theory, as well as the most recent research into seventeenth-century British history, and he combined discussions of the typology of revolution with attempts to distinguish clearly between the preconditions, precipitants, and triggers of England's revolution. There was no doubt in Stone's mind that the English Revolution of 1640–60 was a genuine revolution, its revolutionary nature clearly demonstrated in both its deeds and its words. Stone's thesis was by no means a Marxist one, and he made clear his doubts about the applicability of 'the class war theory of the Marxists' to the seventeenth century (Stone, 1972: p. 54). In the same year, however, as the publication of Stone's *The causes of the English Revolution*, there appeared an even more influential work by the veteran Marxist historian Christopher Hill. Hill's *The World Turned Upside Down* focused on the radical fringe of the 1640s and 1650s, what Hill called the "revolt within the revolution." This was, for Hill, the radical revolution that threatened, but never quite managed, to take England's successful bourgeois revolution forward in unexpected directions. Hill's book, which was reissued 3 years later as a mass-market paperback, caught the counter-cultural spirit of the moment, and had an influence well beyond academia (Hill, 1972; Gurney, 2013: pp. 7–8, 123–124). Much of Hill's research in the 1950s and 1960s had focused on the relationship between puritanism, society, and literature in seventeenth-century England; his new, and widely read study of the prophets, radicals, and 'masterless men' of the revolutionary decades seemed to many to offer something more exciting: a radical history-from-below, one that demonstrated the truly revolutionary potential of the events of 1640–60. Just as Hill had sought in 1940 to prove that England had experienced a genuine, and successful revolution in the mid-seventeenth century, *The World Turned Upside Down* helped remould and revivify that revolution for a new generation (Reay, 1988: pp. 53–71).

Among the new generation of historians who came to prominence in the late 1960s and 1970s, many were unconvinced by conventional Whig and Marxist interpretations of seventeenth-century history, and they were impatient to tackle these head on. Although Christopher Hill retained widespread, though not universal, respect among younger historians, his broad arguments and conclusions were seen by many of them to be unpersuasive. Other scholars – many of them based in the United States – whose work continued to focus on long-term constitutional changes and on clashes over constitutional principle, were also found wanting. The greatest hostility was, however, directed against historians who aimed at providing overarching explanations: Stone's *The Causes of the English Revolution* was a particular target of the new generation (Morrill, 1993, 1999: p. 276, pp. 5–10). In place of ambitious attempts at synthesis, there appeared a plethora of influential works on a variety of subjects, united principally by their questioning of old certainties, their unraveling of long-accepted explanations,

and their drawing on sources that had been unavailable to, or largely ignored by, the older generation. The label loosely applied to these historians, and happily taken up by many of them, was 'revisionist.' The revisionists sought both to demolish the old edifice and rebuild afresh, but the rebuilding often proved more difficult than the initial act of demolition.

Some of the most significant research undertaken by revisionist historians addressed early-Stuart parliamentary history. Kevin Sharpe's *Faction and Parliament* was a key revisionist text, a collection of essays that challenged assumptions about the existence of binary divisions between government and opposition in pre-civil-war England, and drew attention to the politics of faction and interest in the workings of Stuart parliaments (Sharpe, 1978). The most influential critic of traditional accounts of pre-civil-war parliamentary history was undoubtedly Conrad Russell, whose searching 1976 essay 'Parliamentary history in perspective' did more than any other piece of work to undermine the assumption that Parliament, and particularly the House of Commons, was engaged in a long-term pursuit of power and was the main center of 'opposition' to the Crown. His *Parliaments and English Politics*, a forensic study of the parliaments of the 1620s, followed in 1979 (Russell, 1979, 1990: pp. 31–57). Mark Kishlansky's work on the emergence of adversary politics in the Long Parliament, and his important monograph on the rise of the New Model Army, helped do for the 1640s what Russell had done for the 1620s, questioning long-cherished assumptions about the nature of parliamentary politics and political divisions at the height of the civil wars (Richardson, 1998: pp. 222–223; Kishlansky, 1979).

Traditional arguments about the rise of Puritanism, which had featured so strongly in the older social interpretations of the English Revolution, also came under sustained attack at this time. Nicholas Tyacke's work on Arminianism challenged long-held views about the impact on the Church of England of an aggressive and increasingly self-confident Puritan opposition. Tyacke instead drew attention to the Calvinist consensus that lay at the heart of the early-Stuart Church, one that was, he argued, broken by the divisive policies pursued by Archbishop Laud and his allies. It was an Arminian rather than Puritan revolution that provoked instability in the Church: Charles and Laud were the true religious innovators (Tyacke, 1973: pp. 119–143). William Lamont's groundbreaking work on millenarianism, which succeeded in rescuing it from the radical fringe and restoring it to its central place in mainstream Protestant theology, also helped to undermine the notion that the causes of the conflicts of the 1640s were inextricably linked to the rise of revolutionary Puritanism (Lamont, 1969). J.C. Davis's forensic analysis of the nature of early modern radicalism also forced scholars to rethink what they had previously so often taken for granted (Davis, 1982, 1992). Together these works drew on, and contributed to the development of, the new approaches to early modern intellectual history pioneered by J.G.A. Pocock in his *The Ancient Constitution and the Feudal Law* (Pocock, 1957).

Detailed empirical research was a marked feature of revisionism. Blair Worden's *The Rump Parliament*, which drew attention to the limited ambitions of the participants in the revolutionary changes of 1648–49, and the unplanned and unexpected nature of events, displayed a command of the sources of which Gardiner and Firth would have been proud. Few of the MPs who remained at Westminster after Pride's Purge, or who returned in the months following the execution of the king, could, Worden demonstrated, be described as republicans. Many of them followed events rather than shaping them, and their continued participation in Westminster politics was as much about staving off military rule as about condoning the revolutionary developments of 1648 and 1649 (Worden, 1974). One area of research most closely associated with revisionism was the study of local society and relations between center and locality; several important county studies, among them Alan Everitt on Kent, John Morrill on Cheshire, and Anthony Fletcher on Sussex, appeared in the late 1960s and 1970s. Much of the focus of these studies was on the workings of the 'county community' and on the strengths of localist mentalities: for Everitt, England in the first half of the seventeenth century could plausibly be seen as a union of partially independent county-states, and the interests of the county often came before the interests of the nation. Not all of the county community historians went so far as Everitt, but cumulatively their work helped to complicate traditional narratives and undermine further the sense that battle lines across the nation had been drawn up well before the outbreak of war in 1642. There was also general consensus among county community historians that the success of Parliament in the civil war was due in part to its ability to override local sensibilities, its centralizing actions provoking further localist discontent.

Revisionist scholarship transformed understandings of the causes and nature of the conflicts of the revolutionary years. Teleological explanations were jettisoned, detailed empirical research opened up new areas of study, and the contingent nature of events was always emphasized. Nothing could be taken for granted any more, and it no longer seemed possible to fall back on the old social interpretations, or on interpretations that emphasized long-term constitutional or religious conflict, to explain the origins and course of the revolution. But the field of English Revolution studies was becoming a fragmented one, and the explanatory frameworks now adopted by historians would often appear more suited to showing what had not happened than what had. Older historians like Christopher Hill lamented the failure to replace discarded theories with any more persuasive replacements: during the 1980s Hill, placing his hopes as much in a younger generation of literary scholars as in historians now working on the seventeenth century, regularly called for the development of a 'new synthesis.'

Others too picked up on what they saw as the limitations of the revisionist approach. The American historians J.H. Hexter and Theodore K. Rabb both restated their belief in the centrality of parliament and its increasing importance in early-Stuart England. They were unconvinced by revisionist attempts to downplay parliament's importance, or to portray it as merely an irregular event rather than an institution, with only a peripheral role within government. For Rabb, the picture presented long ago by Gardiner was a remarkably durable one, and was in need rather of modification than abandonment (Hexter, 1978; Rabb, 1981). Derek Hirst joined these critics of

revisionism in arguing for the central importance of constitutional principle in shaping parliament's, and the wider public's, attitude to the Crown in the years before 1640 (Hirst, 1981). The 'county community' historians also came in for vigorous criticism, with Clive Holmes insisting that "the preeminence of local allegiance and the gentry's ignorance of, and lack of concern for national issues, goes beyond the evidence" (Holmes, 1980).

The need to replace older narratives of the English Revolution with more convincing new ones was addressed by a group of younger historians to whom the label 'postrevisionist' was often applied. The postrevisionists accepted many of the criticisms that revisionist historians had leveled against their predecessors, but argued that at too many points the revisionists had gone too far. In particular, they accused revisionists of exaggerating levels of political and religious consensus in pre-civil war England, and of ignoring or downplaying the plentiful evidence that existed of conflictual relations in church and state in the early-Stuart period. Their rallying-cry came in the influential volume *Conflict in Early Stuart England*, a volume of essays edited by Ann Hughes and Richard Cust which sought to demonstrate the ways in which significant religious and political divisions had emerged over the course of the sixteenth and early seventeenth centuries, and how ideological positions had hardened in this period. Such developments did not make the civil conflicts of the 1640s inevitable, but they could certainly help to explain why any pursuit of settlement in 1640–42 would have had little chance of success (Cust and Hughes, 1989).

New perspectives were also offered on the question of popular political consciousness. Historians like John Morrill and David Underdown, driven partly by their disquiet at what they saw as Christopher Hill's tendency to too readily equate the radical with the popular, had already sought to explore popular, non-radical responses to the burdens of civil war in the 1640s. Postrevisionists began looking afresh at the ways in which the spread of news and political ideas could influence the thinking of those below the gentry, and at how people were able to make positive political choices, independent of gentry leadership, at crucial moments (Cust, 1986). In some respects this work picked up from that of older historians like Brian Manning, whose *The English People and the English Revolution*, which had received a cool reception from revisionist historians, had sought both to place popular agency at the heart of 1640s politics and to demonstrate the ways in which a 'popular' movement had helped shape political developments in those years (Manning, 1976). Later work by historians such as John Walter, David Underdown, and David Cressy, would add immensely to understandings of popular responses to, and involvement in, political developments in the revolutionary decades (Walter, 2006; Underdown, 1985; Cressy, 2006).

Perhaps the most significant reappraisal of civil war history came from two historians who had long been associated with revisionism, John Morrill and Conrad Russell, who quite separately began exploring the *British* dimension to England's mid-century struggles. Drawing on J.G.A. Pocock's influential call for the creation of a new British history, Russell and Morrill both argued that England's civil war and revolution could not properly be understood independently of events elsewhere in the Atlantic archipelago. Charles I, as King of England, Ireland, and Scotland, was faced with immense, and possibly intractable, problems as ruler of multiple kingdoms. Conflicts in one kingdom invariably affected and became intertwined with developments elsewhere (Russell, 1990: pp. 231–251; Bradshaw and Morrill, 1996: pp. 1–38). The 'British civil wars,' or 'War (or wars) of the three kingdoms,' had begun in Scotland after Charles had attempted in 1637 to impose a new prayer book on a reluctant Scottish people. The Bishops' wars, in which the Scots covenanters had emerged victorious over their English opponents, had forced Charles to turn to parliament in 1640 and made him increasingly dependent upon that institution. The outbreak of Irish Rebellion of 1641 had widened the rift between king and parliament, particularly as rumors spread of Charles's sympathy for the Catholic rebels. Neither side trusted the other to command an army to suppress the rebellion, and the raising of forces by both sides helped bring about civil war. Parliament eventually emerged victorious, but it is doubtful whether the war could have been won without the military support provided by the Scots Covenanters. In the 1650s Scotland and Ireland effectively came under English rule, and it was the English army in Ireland, under Monck, which began the process that led to the restoration of monarchy in England in 1660.

The new British history, as applied to the events of 1640–60, encouraged historians to abandon wholly anglocentric approaches to the study of mid-seventeenth-century English history, and, perhaps unintentionally, helped raise further questions about the very existence of an 'English Revolution' in 1640–60. Some scholars continued to insist on the revolutionary nature of events in those two decades, while falling back on more specific dates – 1640–41, 1640–42, 1648–49, 1653, or 1659 – to locate the moments when truly revolutionary changes took place. Other leading historians of the 1640s and 1650s expressed doubts over the validity of the term 'the English Revolution' to describe what had happened in these decades. For Blair Worden, the civil wars had "transforming effects, albeit inadvertent ones," but they should not really be seen as "a 'revolution' of the kind which, in France and Russia – and in the American War of Independence – would transform, and be intended to transform, perceptions about the nature of politics or of liberty" (Worden, 2009: p. 156). Austin Woolrych accepted that "rebellion is a totally inadequate term with which to describe what happened in the 1640s and 1650s," yet while he acknowledged that the events of these years could usefully be described as a period of revolution, he welcomed the fact that "the designation of them as 'the English Revolution,' so widespread a generation ago, is fading out of use" (Woolrych, 2002: pp. 1–3, 792–796).

Robert Brenner's wide-ranging study of commercial change and political conflict in the period 1550–1653 offered a new and ambitious attempt at a social interpretation of the English Revolution. Building on his earlier work on the rise of agrarian capitalism in England, Brenner accepted the limitations of traditional social interpretations and sought to demonstrate instead the decisive role of 'new' colonial merchants, many of them drawn to political and religious militancy, as a driving force for change. But even he

acknowledged that much of what had been achieved in 1640–60 was reversed after 1660, and that it was only with the successful outcome of the 1688 revolution that earlier gains were consolidated (Brenner, 1993). More troubling for those who still wish to claim that the English Revolution took place in 1640–60 has been the recent revival of research interest in the events of 1688–89. For Tim Harris, it is to "the later seventeenth century that we need to look in order to discover the revolution – in the modern sense of the word – that truly transformed the early modern British polity." The old order was firmly restored in 1660; it was only after 1688 that a radical, and lasting transformation was possible (Harris, 2005: pp. 32–38). Steve Pincus has called the revolution of 1688–89 "the first modern revolution." It wins out over 1640–60 because so many of the changes experienced in the earlier period proved ephemeral; many men and women became revolutionaries in the middle years of the seventeenth century, "but they did not achieve a revolution" (Pincus, 2009: pp. 474–482).

Where does this leave 'the English Revolution' of 1640–60? Lawrence Stone, who was in no doubt that a genuine revolution had taken place in these years, was prepared to admit as long ago as 1972 that the revolution's success in permanently changing the face of England was slight. Rather, what was most important about the revolution was its intellectual legacy. The trial and public execution of the king, the abolition of monarchy and of the established church, the toleration of a wide range of beliefs, and the rise of groups proclaiming ideas of liberty, equality, and fraternity all had a radically transformative effect. Ideas, once expressed, could never totally disappear from view. It was, Stone argued,

> this legacy of ideas which makes it reasonable to claim that the crisis in England in the seventeenth century is the first 'Great Revolution' in the history of the world, and therefore an event of fundamental importance in the evolution of Western civilization.
>
> Stone, 1972: pp. 146–147

Current research on the revolution would seem to bear out Stone's conclusions. Jonathan Scott has suggested that we misunderstand the revolution if we think of it primarily in terms of institutional or constitutional change. Rather, the English Revolution was about ideas: "English radicalism ... *was* the English Revolution" (Scott, 2000: pp. 33–34). The reaction to the religious policies of Charles I manifested itself in religious anti-formalism, and this could be seen to have been the driving force of English Revolution radicalism. Radicals such as the Gerrard Winstanley, Abiezer Coppe, Laurence Clarkson, and George Fox demonstrated a marked anti-formalism in their writings, which enabled their ideas to develop in wholly unexpected ways. In Winstanley's case it helped lead him to a rejection not only of all organized religion but also of private property and monetary exchange. Constitutional anti-formalism led to the discarding of old-established institutions and to the adoption of new and innovative constitutional models, which, in turn, lasted for only so long as they were seen to serve their purpose. Leveller radicalism, with its emphasis on fundamental and inalienable rights, was a product of the revolutionary decades, and it was in the 1650s, particularly in the person of James Harrington, that England made some of its most significant contributions to the history of republican thought. Harrington's *The Commonwealth of Oceana*, which was published in 1656, was to have an influence far beyond England and well beyond the seventeenth century. Royalist writers too may be said to have laid the foundations of modern conservatism in the revolutionary decades (Wootton, 1986: p. 9). In the religious sphere, the breakdown of the established church in the 1640s led to a proliferation of sects and beliefs. The church was restored after 1660, but the religious settlement that followed was an exclusive one, and many remained outside of it. The nonconformist tradition, which was to have such a profound influence over the intellectual life of the nation in the following two centuries, was thus another product of the mid-seventeenth century conflicts. Finally, as Christopher Hill remarked, the fate of Charles I provided a reminder to all future kings that they had a joint in their neck – perhaps James II would not have taken flight so swiftly in 1688 had he not remembered what had happened to his father. Hill may have been overoptimistic in seeing the English Revolution of 1640–60 as a successful bourgeois revolution, in which the crucial changes in commerce, state, and society had all occurred by 1660. Guizot and Marx were, it seems, more correct in recognizing that fundamental change was not concluded before 1688–89. Nevertheless, in the realm of ideas, the English Revolution of 1640–60 had a lasting impact, one that helped shape the course and outcome of the revolution of 1688 and influenced the American and French revolutions of the following century.

See also: American Revolution, The; Bourgeoisie and Middle Classes, History of; French Revolution, The; Religion, History of; Russian Revolution, The; Social Movements, History of: General.

Bibliography

Bernstein, Eduard, 1930. Cromwell and Communism. George Allen & Unwin, London.
Bradshaw, Brendan, Morrill, John (Eds.), 1996. The British Problem, c.1534–1707: State Formation in the Atlantic Archipelago. Macmillan, Basingstoke.
Brenner, Robert, 1993. Merchants and Revolution: Commercial Change, Political Conflict, and London's Overseas Traders, 1550–1653. Princeton University Press, Princeton.
Cressy, David, 2006. England on Edge: Crisis and Revolution, 1640–1642. Oxford University Press, Oxford.
Cust, Richard, 1986. News and politics in early seventeenth-century England. Past and Present 112, 60–90.
Cust, Richard, Hughes, Ann (Eds.), 1989. Conflict in Early Stuart England: Studies in Religion and Politics, 1603–1642. Longman, London.
Davis, J.C., 1982. Radicalism in a traditional society: the evaluation of radical thought in the English commonwealth, 1649–60. History of Political Thought 3, 203.
Davis, J.C., 1992. Religion and the struggle for freedom in the English Revolution. Historical Journal 35, 507–530.
Guizot, F., 1826. Histoire de la révolution d'Angleterre depuis l'avenèment de Charles 1er jusqu'a la restauration de Charles II, Part 1, vol. 1. A. Leroux et C. Chantpie, Paris.
Guizot, F., 1846. History of the English Revolution of 1640 (William Hazlitt, Trans.). D. Appleton & Co., New York.
Gurney, John, 2013. Gerrard Winstanley: The Digger's Life and Legacy. Pluto Press, London.
Hammersley, Rachel, 2010. The English Republican Tradition and Eighteenth-Century France: Between the Ancients and the Moderns. Manchester University Press, Manchester.

Harris, Tim, 2005. Restoration: Charles II and His Kingdoms. Allen Lane, London.
Hexter, J.H., 1978. Power struggle, parliament and liberty in early modern England. Journal of Modern History 50, 1–50.
Hill, Christopher, 1940. The English Revolution 1640: Three Essays. Lawrence & Wishart Ltd, London.
Hill, Christopher, 1972. The World Turned Upside Down: Radical Ideas during the English Revolution. Temple Smith, London.
Hill, Christopher, 1980. A Bourgeois revolution? In: Pocock, J.G.A. (Ed.), Three British Revolutions, 1641, 1688, 1776. Princeton University Press, Princeton, pp. 109–139.
Hirst, Derek, 1981. Revisionism revised: two perspectives on early Stuart parliamentary history. The place of principle. Past and Present 92, 1.
Holmes, Clive, 1980. The county community in Stuart historiography. Journal of British Studies 19 (2), 54–73.
Kishlansky, Mark, 1979. The Rise of the New Model Army. Cambridge University Press, Cambridge.
Lamont, William M., 1969. Godly Rule: Politics and Religion, 1603–60. Macmillan, London.
Macaulay, Thomas Babington, 1843. Critical and Miscellaneous Essays, vol. 1. Longman, Brown, Green and Longmans, London.
Manning, Brian, 1976. The English People and the English Revolution. Heinemann, London.
Marx, Karl, 1979. Marx/Engels Collected Works, vol. 11. Progress Publishers, Moscow and London.
Morrill, John, 1993. The Nature of the English Revolution. Longman, London.
Morrill, John, 1999. Revolt in the Provinces: The People of England and the Tragedies of War, 1630–1648. Longman, London.
Notestein, Wallace, 1924. The Winning of the Initiative by the House of Commons. British Academy, London.
Pincus, Steve, 2009. 1688: The First Modern Revolution. Yale University Press, New Haven.
Pocock, J.G.A., 1957. The Ancient Constitution and the Feudal Law. Cambridge University Press, Cambridge.
Rabb, Thoedore K., 1981. Revisionism revised: two perspectives on early Stuart parliamentary history. The role of the commons. Past and Present 92 (1), 55–78.
Reay, Barry, 1988. The world turned upside down: a retrospect. In: Geoff, Eley, William, Hunt (Eds.), Reviving the English Revolution: Reflections and Elaborations on the Work of Christopher Hill. Verso, London, p. 64.
Richardson, R.C., 1998. The Debate on the English Revolution, third ed. Manchester University Press, Manchester.
Russell, Conrad, 1979. Parliaments and English Politics, 1621–1629. Clarendon Press, Oxford.
Russell, Conrad, 1990. Unrevolutionary England, 1603–1642. The Hambledon Press, London.
Scott, Jonathan, 2000. England's Troubles: Seventeenth-Century English Political Instability in European Context. Cambridge University Press, Cambridge.
Sharpe, Kevin (Ed.), 1978. Faction and Parliament: Essays on Early Stuart History. Methuen, London.
Snow, V.F., 1962. The concept of revolution in seventeenth-century England. Historical Journal 5, 167–174.
Stone, Lawrence, 1972. The Causes of the English Revolution, 1529–1642. Routledge & Kegan Paul, London.
Tyacke, Nicholas, 1973. Puritanism, Arminianism and counter-revolution. In: Russell, Conrad (Ed.), The Origins of the English Civil War. Macmillan, London.
Underdown, David, 1985. Revel, Riot and Rebellion: Popular Politics and Culture in England, 1603–1660. Clarendon Press, Oxford.
Walter, John, 2006. Crowds and Popular Politics in Early Modern England. Manchester University Press, Manchester.
Woolrych, Austin, 2002. Britain in Revolution, 1625–1660. Oxford University Press, Oxford.
Wootton, David, 1986. Divine Right and Democracy: An Anthology of Political Writing in Stuart England. Harmondsworth, Penguin.
Worden, Blair, 1974. The Rump Parliament. Cambridge University Press, Cambridge.
Worden, Blair, 2001. Roundhead Reputations: The English Civil Wars and the Passions of Posterity. Allen Lane, London.
Worden, Blair, 2009. The English Civil Wars, 1640–1660. Weidenfeld & Nicolson, London.

Enlightenment

Giuseppe Ricuperati, University of Turin, Turin, Italy; and Academic of Lincei, Rome, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The epochal concept of modernity, Enlightenment, describes the style of thinking that predominated in the eighteenth century. It is difficult to periodize strictly. It started with the 'crisis of the European mind' in the last decades of the seventeenth century, reaching its maturity in the 1750s to the 1760s. The Enlightenment made a strong contribution to the crisis of the Old Regime, creating some premises of the French Revolution, up to Romanticism. Reason, experience, public opinion, and cosmopolitanism were considered operative tools. This article stresses the 'reinvention of the Enlightenment' against cultures menacing democracy and liberty. It is difficult today to consider the Enlightenment as a conceptual unity (as Cassirer did), as differences and varieties prevail in recent interpretations. In this, national contexts have played an important role and so have new perspectives: not only the history of ideas, but also gender studies and cultural history. In this way, a previous, unitary conception has been challenged by different Enlightenments: it can be 'conservative' (Pocock) or 'radical' (Jacob and Israel). Debates since 2000 have included those surrounding Israel and his Radical Enlightenment, the five volumes by J. Pocock and his Conservative Enlightenment and its compatibility with Empire, as well as themes such as religious universalism, human rights, and Pierre Bayle's challenge concerning virtuous atheism, of particular relevance to a society like ours, which can be seen to be without religion, with few civic ethics and even postsecular. The last challenge concerns the continuity or fracture between the Enlightenment and the French Revolution.

Definition and Meaning

Enlightenment is an epochal concept of modernity. It can be compared with others that preceded it such as Renaissance and Baroque, or followed like Romanticism, the latter mostly as its antithesis. The word Enlightenment is molded exactly on its German original, *die Aufklärung*, the first conceptual definition of which was due to Immanuel Kant; in 1784, he had tried to explain the meaning of the word to the readers of a Berlin journal. The first point to be borne in mind therefore is that the term was condensed with theoretical precision in a German area and that Enlightenment (second half of the nineteenth century) or *Illuminismo* (beginning of the twentieth century) were created as translations of the German term. But the concept of *Aufklärung* has had several metaphoric equivalents in European culture since the eighteenth century. The best known are linked to the term light, which in its turn is connected with reason: therefore Enlightenment was conceived from the very beginning as the Age of Enlightenment, emancipating reason, the time of humanity's coming-of-age. This is what Kant meant. Reason is a critical tool allowing humans to get rid of prejudice, to change the world, dominating nature through science. Enlightenment is conceived as a culture of conscious change, that is, of reforms, the secularization of a term that was originally in the singular and expressed a religious concept of modernity.

It does not mechanically coincide with rationalism, but it is one of its essential and critical moments. In fact, a rationalistic conception of the world precedes the Enlightenment and really belongs to the seventeenth century, as philosophers like Descartes and Spinoza could show. With them, reason is a system, whereas with the Enlightenment it becomes a method. Reason becoming method is an essential point of enlightened epistemology, which not only uses Locke's empiricism, but also Newton's celebrated motto, 'Hypotheses non fingo.'

Times, Spaces, and Routes

As for all complex epochal categories, it is not easy to imprison Enlightenment in a definite period in time or space. The fact that the concept originated in a German area does not necessarily mean that the center of Enlightenment is linked to that environment. From a historiographic point of view, at least three great European cultures claimed to have given origin and identity to the Enlightenment: France, England, and the German world. As far as France was concerned, the capital of Enlightenment (light) was Paris, in the name of the *philosophes* and of the *Encyclopédie*, which synthesized its essential knowledge. London could not only boast Locke's and Newton's contribution, or the precocity of the freethinkers, and later the great lesson of Scottish intellectuals, but also its own political system, already based on the separation of powers, and above all on the rapid evolution of public opinion, favored by a free press. The Germans in their turn claimed not only that they had first created the conceptual condensation of the term, but also that they had given (through Leibniz, Wolff, and Kant) Enlightenment a philosophy. Actually, these claims implicitly conceal all the ideological processes of the following age, which perceived cultural identities on a national basis, in antithesis to Enlightenment, thus denying a fundamental presupposition, cosmopolitanism.

If one wants to understand the Age of Enlightenment today, one has to not only put aside any simplification identifying center and peripheries, but accept its polycentric nature, which should be considered as a European and worldwide circulation of ideas, to which Kant attributed the choice of universal emancipation. This does not mean that one should ignore the diversities that have induced some scholars to imagine the existence of various kinds of Enlightenment. A polycentric interpretation probably helps better than others to save the unity and the diversities of Enlightenment at the same time.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62004-4

It is never easy to define the period of time enclosing an epochal concept. If one compares the cultural events of France, England, and Germany with those of other areas from the Mediterranean sea to the north of Europe, one can take the 1680s, indicated by Paul Hazard as the beginning of the 'crisis of European conscience,' as a starting point, provided one does not fix dates or individual events. They were the years of the revocation of the Edict of Nantes, of the querelle between the Ancients and the Moderns, of the appearance in European culture of Etrangers Symboles ('foreigners symbols'), like the Chinese Mandarin, archetype of virtuous atheists. In England, it was the period not only of intellectual maturity with Locke and Newton, but also of freethinkers like Anthony Collins and John Toland. Thomasius and Leibniz were the protagonists of the *Frühe Aufklärung*. In Holland, refugees like Pierre Bayle and Jean Leclerc elaborated the tools of a great critical culture. Only slightly later, the adventure of Giambattista Vico, Pietro Giannone, Alberto Radicati, di Passerano, and Ludovico Muratori had its beginning in Italy. As for the Iberian Peninsula, the *décalage* was greater.

The 1730s and 1740s can be considered years of transition, with the important exception of Voltaire's and Montesquieu's France, which inaugurated the Age of Enlightenment. In England, the radical culture of freethinking died out, whereas the Augustan season asserted itself through Samuel Johnson. In Italy, Radicati and Giannone succumbed to the success of the érudites and of the enlightened Catholics (from Scipione Maffei to Muratori). The latter's Spanish equivalent was Father Fejio.

The 1750s and 1760s are the season of Enlightenment at the height of its glory. The circulation of ideas linked together all the European centers. Not only Paris, London, Berlin, or the cities of the great cultural mediation like Amsterdam and Geneva, but also Madrid, Lisbon, Milan, Naples, St Petersburg, and Stockolm became involved in a close dialogue. These were the days of the *Encyclopédie*, of the *Contrat Social*, of Cesare Beccaria's *Dei Delitti e Delle Pene*. Learned periodicals became journals of opinion. The first great reforms were introduced in the fields of economy, legislation, politics, and morals. Pombal's Portugal opened a confrontation with the Jesuits in which all the Bourbon states would become involved and which would have an echo in the great European enlightened culture.

The network of freemasonry developed the political language of a civil society, deeply entwining itself with all the other tools of European intellectual sociality. Reforming projects were most fruitfully nourished by Utopian ideals. Enlightened absolutism made its tentative appearance in Austria, Russia, Prussia, Tuscany, Denmark, and Sweden. If sovereigns were drawing near to the philosopher's projects, there was also a fruitful exchange between enlightened culture and the governmental experience of a new generation of officials.

The 1770s and 1780s, when little by little the generation of the great representatives of Enlightenment (Voltaire, Rousseau, Diderot, D'Alembert, and d'Holbach) disappeared, witnessed an Atlantic event – the American Revolution, which gave Europe firsthand ideological material on questions not yet approached, such as the self-determination of peoples, parliamentary representation, and federative connection within a republican model: liberty, democracy, and civil ethics as a shared and written constitutional project. Europe's boundaries widened with Russia entering the Mediterranean, and the inevitable modifications of the eastern balance. Historians like Edward Gibbon and law reformers like Gactano Filangieri took note of it. The models of reason changed. Freemasonry itself, which Lessing thought the ideal tool to diffuse Enlightenment, experienced a deep restlessness and the clash between operative rationality and a taste for the occult and mysticism. Raynal, who had enjoyed Diderot's collaboration, attacked the European colonial models, denouncing both their violence against other civilizations as well as slavery and their inclination toward despotism; he went so far as to assert that the torch of freedom would be carried to Europe by a black Spartacus.

It is not easy to find an *ad quem* term. The beginning of the French Revolution, a starting point of an epochal crisis, is plausible beyond a logic privileging the French course. Some frontier areas were soon involved in the revolutionary turmoil. The process seemed to take longer in Germany, England, Spain, Portugal, and Italy, not to mention in Eastern Europe. The French Revolution and war put a stop to the politics of enlightened absolutism. The crisis of enlightened culture and its transformation into something different (from Jacobinism to counterrevolutionary resistance) lasted through most of the 1790s. As for England – which experienced the first unsettling industrial revolution – it carried out an uncompromising war not only against the French Revolution, but against the Napoleonic Empire as well, yet this did not modify its constitutional evolution, nor did it alter substantially the role of public opinion. This is true also with regard to the United States, which would later appear to an observer like Alexis de Tocqueville (*De la démocratie en Amérique*, 1835) as the most mature and complete project of the Enlightenment, from which the Europe of *Restauration* would have to learn freedom and democracy anew.

Operative Tools and Fields

An analysis of the categories of space and time connected with the concept of the Enlightenment implicitly shows its operative tools (reason, experience, criticism, scientific models, reforms, Utopian tensions, and theories of progress and of public and private happiness). Each of these terms not only has its own internal dynamics, but also tends to adapt to highly differentiated realities. Enlightenment today does not appear so much as a clear unitarian philosophy, according to its reconstruction by Ernest Cassirer (1932), but as a process of general emancipation that moves hypothesis, possibilities, and strategies that might appear different and antithetical at first sight. This is clear from a brief analysis of its main fields of influence: religion, ethics, politics, economy, history, social transformation, and cultural change.

As regards religion, in the long term, secularization had its roots in the first period of modernity (humanism, renaissance, reformation, and libertinism). Criticism of revealed religions was an important element of English freethinking in the early eighteenth century, with repercussions not only in Holland, but also in France and Italy. Materialism, pantheism (the term was coined by Toland in 1704), and the negation of the immortality of the soul involved both the protestant and the catholic world, as the religious adventures, destined to be

defeated, of Jean Meslier, Giannone, and Radicati di Passerano demonstrate. The progressive reduction of magical, miraculous, and prophetic forecast practices were generalized as the refusal of superstition. With fruits of the application of critical models linked with philosophy and science, they created various intersections between faith and reason, from reasonable Christianity to enlightened Catholicism, to natural theology, to deism. Tolerance, which is not only the cohabitation of different religious confessions on the same territory, but also the possibility (enunciated by Pierre Bayle) of individual conscience to err, without any church arrogating to itself the power of constraint, was a good ground for testing important conquests. In the wake of a great juridical tradition, Voltaire would take this battle into the very core of the Enlightenment, elaborating for European public opinion the theoretical premises for the choices of Prussian, Austrian, and Tuscan enlightened absolutism.

D'Holbach's *coterie*, which represents the most radical experience of the French Enlightenment from a religious point of view (with two phases: the first an anti-Christian offensive; the second the elaboration of a social model where ethics and politics were separated from religion) is only one of the components, although the most aggressive and sensational, of the French and European Enlightenment. It is all too easy to underline its elitist nature, limited to a small group speculating with the complicity of European publishers interested in creating a flourishing market of clandestine books. One should not underestimate this aspect, after a scholar like Robert Darnton explained the relation between publishing and sedition, between forbidden books and a crisis of political, ethical, and religious social structures.

The most lasting result of the Enlightenment in this field was the widespread consciousness not only among laymen but among priests as well, including enlightened Catholics, that religious choice mainly belonged to the private sphere. In this sense, the most mature enlightened project had accepted and adapted for a larger civil society the rights of conscience proposed by Bayle at the time of the 'crisis of European conscience.'

Emancipation, Reform, and Utopia

In its most essential and creative moments, each innovative choice of the Enlightenment was preceded by an ethical reflection that cannot be mechanically reduced to an individual reckoning of pleasures and sorrows. The utilitarian moral itself had imagined with Bernard de Mandeville that private vice could become public virtue, but in the middle of the eighteenth century, it was perceived more as a tension to ensure the maximum happiness to the largest possible number of people. From this point of view, the analysis of Montesquieu's works, of the 'économists,' of Pietro Verri, Beccaria, and Gaetano Filangieri, is important because of the strong relation they establish with a new ethical project. The Scottish Enlightenment also started from morals to discover economy and sociology. In fact, one of the concepts most charged with future, that of 'civil religion' theorized by Rousseau in his *Contrat social*, was based on a profound understanding of natural law (above all through Jean Barbeyrac), which had tried to set the field of man's rights and duties free from the traditional links of Patristics and Scholasticism.

Rousseau and Mably had transformed this inheritance and founded a new system of the rights and duties of the citizen. The state took upon itself new tasks in the fields of instruction, cultural control, and assistance, thus widening its public dimension and secular action; as a consequence, religious choice tended to become a private question, involving the intimate and inviolable sphere. The most significant lesson of the Enlightenment was not so much a lay morality opposed to the religious one, as the possible dialogue on the field of citizenship between different kinds of ethics: a difficult project, continuously interrupted, which modernity has to tackle all the time.

Also, from the point of view of political models, the Enlightenment does not show a unified front, but a series of possible choices, from enlightened absolutism to the return of a republican culture, to the bureaucratic centralization of the 'well-ordered police state,' to constitutionalism, to debates on representation, to projects of direct democracy. Each of these terms in fact can be used to define very different experiences, unless there is the intervention of a contextual historical event. There is a sharp distinction, for example, between the forced modernization of Peter the Great, which is an example of reformism to the point that it attracted Voltaire's attention, and the strategies of 'enlightened' princes like Frederick II, Joseph II, Peter Leopold, and even Catherine the Great.

In the same way, with regard to republican thought, one should not confuse the patrician models of the republics, which survived the success of absolute power (Holland, Venice, and Genoa), or the mannered republicanism one learnt by reading the classics, with the new republican tensions that arose as a consequence of the claims for freedom of the Poles, the Corsicans, the Genevans, and above all the American colonies. But enlightened absolutism itself attracts or repels enlightened thinkers because of two equal possibilities. The first is the use of centralization in order to carry our reforming politics, where the accent was not so much on strength, however, necessary to win resistance, as on the involvement in the project of new fields: law, politics, economy, religion, and culture, together with the elimination of particularisms and unjust social relations. This implied the other necessity, that of transforming the model into constitutional forms. The process of emancipation, and of adaptation to the complexities of the real world one wants to change, Enlightenment moves between the two poles analyzed by Franco Venturi (1971): Utopia and reform, closely entwined. These two tensions vivify each other. But the course of the Enlightenment revolves round another two essential terms: freedom and democracy. The first term above all was the main object of Montesquieu's *L'esprit des lois* (1748). It is not possible here to go beyond a schematization recognizing virtue as the essential value of republics, as distinguished from honor, the real ethical and social cement of monarchies. Montesquieu did not choose among the different political forms. He only discarded Asian despotism and was very diffident against absolutism, which he wanted to temper not only by separating the three powers, or through the binding strength of the laws, but also through the concrete presence of intermediate bodies, including the nobility.

The other term of reference for the political theory of the Enlightenment is Rousseau. He opposed to eighteenth-century cosmopolitism the love for one's own country, which requires the practice of direct democracy. Rousseau's democracy is right for a small republican country. It does not admit representatives. Between majority and minority, only the former represents the general will, because sovereignty derived from the social pact cannot be divided. The Genevan's challenge was even more complex: he questioned a developmental model of civilization based not only on the progress of art and science, but also on the basis of social inequality. Democracy was the courage to modify ownership, justice, and instruction. A large part of later political thought is linked to a comparison between Montesquieu and Rousseau, to establish a connection between freedom and democracy, the aspiration to absorb the Genevan's Utopian drive and transform it into a spirit of reform. An exemplary text is *Dei delitti e delle pene*, which questions the right to punish in an unjust society because it is unequal.

Various other experiences belong to Enlightened politics: cameralism, an administrative model mainly used by Prussian and Austrian officials; the physiocratic theory of legal despotism, in which a monarchic power guarantees certain natural laws, among which is ownership. From the above considerations, it appears that the question of enlightened politics is narrowly linked with the question of law, which becomes the 'science of legislation.' The weaker field was feminine political and cultural identity. For this question, one would have to wait for Mary Wollstonecraft and her work on women's rights.

Economy became a science during the eighteenth century, evolving from late-mercantilist models. The physiocratic project established the essential points of the productive cycle based on agriculture and on free corn trade. If agriculture was the only wealth-producing sector, while transformation and circulation were the task of industry and commerce, taxation had to become progressive with respect to land, whose ownership was one of the fundamental natural rights. Adam Smith extended liberalism to sectors such as industry and commerce, which were not considered subsidiary, but highly productive. In this way, the first strong perception of an industrial revolution under way in England became a theory. During the last decades of the century, there was a return to protectionism.

What happened with the economy could well be verified with other sciences, not only social and natural, but human as well. The Scottish school created the premise for sociology, by studying social stratification or the concept of civil society. Economy itself, by defining a stages theory based on production, laid down the basis of anthropology. Consolidated disciplines like physics were renovated, as well as chemistry, geology, agronomy, and geography itself. Linneus and Buffon had reclassified nature, thus rendering it an open book. Natural history had given a common language to the natural and biological sciences.

As for history, one can only register a change in some fundamental paradigms: the loss of the superiority of sacred history, the transformation of ecclesiastical history into the history of religions, the emergence of the history of civilization within new models of universal history, the flourishing of civil history as institutional history, and the first appearance of the history of science.

Circulation and Practices: the Growth in Literacy

If one examines this process from the point of view of cultural history and of the general transformation of intellectual models, European and even the worldwide circulation of ideas can be measured through parameters that are apparently external to intellectual history, but not unimportant. The author refers first of all to the international book market, and to the creation of an incredibly large number of tools, thanks to which the consultation of books was made possible, from public and private libraries to bibliographies, catalogues, abstracts, and reviews. Books and newspapers gained an almost universal circulation. In Europe and America, there was an increase in literacy. The world of potential readers expanded not only thanks to the advent of the middle classes but because the distance between male and female instruction tends to diminish in urban spaces. Somebody called this the revolution of middle class readers. Even reading habits changed; there was a greater tendency for undirected reading, not linked to learned professions but to the consumption of leisure time. The demands of the new readers also brought about a change in genres. The novel had an incredible development, offering as it does to all classes means for education, identification, and escapism. Among the forbidden books of the second half of the eighteenth century, the *philosophique* and *pornographique* genres often followed the same semiclandestine circuits, thus contributing to the modification of the cultural and ethical general models of society. Music and theater as well were often fields for ideological formation and dispute. Journalism followed in the tracks of this expansion of the book market, and from the objectivity of abstracts, which are made for a public of learned people, moved to the subjectivity of journals, the concision of reports. Journals of opinion were essential for the formation of this new international public sphere, which discovered languages, interlocuters, and spaces. To this sphere belong the salons, public, and private spaces at the same time. Intellectual sociality took a number of forms, from the old and new academies, to clubs, to cafés, down to the taverns where conversation and gossip were a way for disseminating news. The Enlightenment–Academies relation is supported by a rich bibliography, starting with Reinhard Koselleck's (1959) work, which laid down the basis for a more through analysis of the links between Freemasonry and Enlightenment. The former is now considered the privileged center for the formation of the political language on which the assemblary model of the French Revolution would be created.

Many centers of intellectual communication reverse the prevailing male image one received when analyzing the concept from the point of view of intellectual history. Women were present in the famous French salons, where they have some personal control over an ephemeral though socially important genre like conversation. They started being accepted in some lodges and academies, their role thus changing from that of consumers of books and newspapers (also of a literature for women, produced by men) to that of writers. First, they occupied the space of women's literature, then they took an

active part in the journalistic debate. The inclusion of women in the public sphere does not alter the fact that their role was still subordinate. Women's writing often tends to self-identification, but it remains unfinished, secret. And this corresponds to a social model that is prevalent in the male-dominated Enlightenment. Rousseau, who had put the question of the public education of citizens, had kept women separate; Filangieri himself had followed suit. This explains the novelty of Wollstonecraft's denunciation. In fact, the public sphere created by the High Enlightenment seemed to be threatened by new classes joining the world of writing and communication, forming an international Grub Street and living on collaborations and translations. The creation of a more aggressive political language, and of the use of reason both as prophecy and announcement of the crisis, broke the balance of the previous generations, which had reached the highest levels of intellectual communication in the academies, in the teaching and in the publishing world. In this phase, women seemed to live less apart. The fact remains that on this question Enlightenment as emancipation was a partial project delegated to the future, because it had not fully accepted the other half of its universe.

Building an Epocal Concept

Finally, it should be said that the French Revolution, while celebrating the Enlightenment as its parent, strongly contributed to its interruption first of all because what had survived as *idéologie* had to adapt to the heavy Napoleonic tutelage and stabilization, and second because the behavior and the propaganda of Napoleonic France were experienced by other European cultures as a violence that forced them to find anew their national, religious, and identifying roots. This explains why part of Romanticism set itself up as an antithetic culture to the Enlightenment that was considered responsible for the political conditions in Europe. Friedrich Meinecke spotted the opposition between cosmopolitism and the national state (1908). While new problems required new kinds of culture, certain fundamental values, after being driven underground during the *Restauration*, after the Napoleonic age, sprang up again in all the moments of opening to freedom and democracy, which are scattered in the history of Europe. In French history, the ideological elaboration of the Third Republic renewed the strong link between Revolution and Enlightenment again, the latter being dragged by the former.

The twentieth century, however, really needed to reflect on the Enlightenment again, having a critical period troubled by nationalistic, fascist, and totalitarian cultures that threatened liberty, democracy, and equality. These threats were answered by Cassirer's (1932), Becker's (1932), Hazard's (1935, 1946) and many other projects. From these roots, there originated an important historiography (Venturi (1971), Gay (1967), Koselleck (1959), Pocock, Darnton, Jacob (1980), Roche, and many others) that to this day finds an echo in the International Society for Eighteenth Century Studies and in the dense network of National Societies. Margaret Candee Jacob has written about a 'Radical Enlightment,' which was masonic, pantheist, and republican. John Pocock (1999), more recently, has located in the English world a 'Conservative Enlightenment,' while Jonathan Israel (2001) widens the chronological terms of a European 'Radical Enlightenment' and 'its making of modernity,' starting from 1650.

There are also many publications on the limitations, deficiencies, and responsibilities of Enlightenment in the evils of modernity, to begin with T.W. Adorno and M. Horkheimer (1947). But in the light of the dramatic problems of our times, many evoke the necessity of a further Enlightenment, of a future of Enlightenment that will resume the interrupted project. One can well imagine that it will inevitably be a new one, because the past can give inspiration, but not necessarily return. It is true, however, that we must find our own answers to emancipation, freedom, and Utopia for the world otherwise risks being an icy globality without universalism.

Recent Discussions and New Themes about Enlightenment (2002–12)

Israel's work on the Radical Enlightenment and its prosecution has been the occasion for an interesting debate, which can be summarized by describing the conclusions of a meeting organized in France (Lyon) in 2004, an occasion to verify the terms implied, starting from the retrodating of the Enlightenment to the time of Spinoza: Israel considers spinozism the main root of the radical project. The organizers of the meeting agreed that Israel's work had reopened the discussion about the Enlightenment, but he faced serious criticism, not only about a solution that implies a key that is too restricted, but also one that is primarily resolved in terms of intellectual history. Also deemed to be of importance was his decision to produce a transnational history. The problems surrounding the chosen key are confirmed by the second volume, *Enlightenment Contested* (2006), which strangely enough returns at the first half of the eighteenth century. Also, the recent book on the revolution of the mind reveals that the connection between the Radical Enlightenment and the French Revolution, strongly defended in the previous volumes, is less linear than previously suggested. With his last volume, *Democratic Enlightenment* (2011), Israel not only has completed his huge opus concerning the Enlightenment and its roots, analyzing the temporal and spatial period between 1750 and 1790, but also answers his critics starting with Antoine Lilti, now leader of the 'Annales,' and an important scholar of the world of the Parisian salons in the eighteenth century, critically connected with social and cultural history. On the basis of his reading of the first two volumes, Lilti accused Israel of practicing only an old intellectual history. It is easy to perceive in the pages of the 'Annales' (3, 2009) a defense of a model of history, connected with great scholars such as Robert Darton, Roger Chartier, Daniel Roche, and others, but in some ways a sort of flagship of French historiography. Confirming his previous choice, Israel has good reasons not only to demonstrate that cultural and social history have not produced a general and transnational reconstruction of the Enlightenment as a worldwide 'revolution of the mind,' but also that his history is not the old intellectual history, focused solely on ideas. He considers also social processes of transmission, transformation, conflicts, and secret sects such as masons, all connected with a society not of the happy few, or the great intellectuals, but

rather collectivities of men and women. He also confirms his agreement with John Robertson about the unity of the Enlightenment, and its distance from the huge national project by Pocock, which separates Gibbon not only from the French, but also from the transnational Enlightenment. It is true that he insists on the keyword philosophy, but it incorporates religion, criticism, philology, as well as politics, science, and law. It is a fundamental tool of transformation, but only if it involves collective and conflicting social energies. Israel is open in his polemics against a model of history that reduces all to interpretation, literary artifact, or communication. But the best answers are on the terrain of the research. The first part on earthquakes involves not only Lisbon, but also South America and is a way to verify the differences between science and superstition, Nature and Providence, materialism and deism, philosophers and antiphilosophers, as social groups with their tools of propaganda. The challenge involves also Jean Jacques Rousseau and his growing distance from the *philosophes* and the 'Encyclopédie.' The chapters concerning the main tool of the French *philosophes* not only are clear and rich, but also incorporate the findings of Darnton. Less convincing and perhaps too simplified is what he has to say about Rousseau, considered more for his deism and antimaterialism than for his *Contrat social*. In the reconstruction of a transnational and complete image of the Enlightenment the second part, 'Rationalising the Ancien Regime,' is impressive, revealing a global vision, which connects the different European cultures, and political forms. Action and reaction, projects and resistances are clearly analyzed for all European countries. The third part, 'Europe and the Remaking of the World,' considers America, not only the North and the crucial event of the American Revolution, but also South America as well as Spain and Portugal India, China, Japan, and their relationships with different European cultures. This section reveals Israel to be perhaps the best pupil of the great John Eliott. He also accepts from Sankar Muthu that Enlightenment is against Empire. The fourth part is dedicated to the return of Spinoza: he defines in one of the central chapters *Pantheismusstrei*, involving France, but above all Germany where Lessing opened a Radical or Democratic Enlightenment, while Kant fought against Israel's hero, Spinoza, inventing an idealistic philosophy. The fifth part, dedicated to the French Revolution, appears a correct reconstruction of the glorious moment in which the Third Estate celebrated its triumph against Monarchy, Aristocracy, and the Church. It is easy to be persuaded by this part of the benefits of a Revolution, which erased the *Ancien Régime*, and invented Human Rights. The author is persuaded that a strong contribution to this field was offered by *philosophes* such as Rousseau and later Filangieri, in relation to the American Revolution. Also, Kant was an admirer of this phase of the French Revolution, but it is also true that Hegel, considering the Terror and the Napoleonic European authoritarian order, held the Enlightenment responsible. This book merits to be read with a great respect for its huge intellectual courage, and its capacity to answer many objections of his opponents. Respecting the project involves considering all three volumes. It would appear that for a long time this will be the main reference in the field. In this sense, the author thinks that *the lector in fabula* now has certainly the duty to correct single judgments, but doing that, must respect one of the intellectual adventures that can be compared only with Paul Hazard's opening a European Resistance, by Franco Venturi's continuation of the same Resistance, or now also by John Pocock.

Another great project is almost concluded. The author speaks about the five volumes by John Pocock, in which he utilizes the great historian Edward Gibbon not only to confront him with the proposals of the French Enlightenment, stressing the differences, but also to reconstruct the way in which the eighteenth century's English world debated great themes such as Empire and colonialism. The history of the past is here a means of considering the future. Pocock confirms the compatibility between a Moderate or Conservative Enlightenment and the imperial vocation of Great Britain, connected with the Union of 1707. He reads Roman history as a metaphor for the English colonial and imperial choice. The problem has been largely discussed not only in England and in the United States, but also in other countries. Sankar Muthu has demonstrated that Enlightenment and Empire were opposed (2003). The Scottish Enlightenment and its connection with the Union have been analyzed not only by J. Robertson, but also by G. Abbattista on Empire, D. Francesconi, considering 'the age of history' (2003), and S. Sebastiani, speaking about 'the limits of progress' (2008).

A new problem and field of research has emerged, as a creation and inheritance of the Enlightenment, around human rights. The terms, not coincidentally, emerged with Rousseau and Voltaire in the 1960s, becoming central in the 1970s and above all in the 1980s. The invention of human rights became central with the American Revolution and also in the Declaration of the Rights of the Man and the Citizen at the beginning of the French Revolution, connecting the birth of the United States with Europe. Not only did Diderot invoke liberty by a black Spartacus, but he also has a recent critical edition of Gaetano Filangieri's *Scienza della legislazione*. The study of its translations has confirmed the presence of human rights, republicanism, and constitutionalism in the 1980s in Europe. The Italian contribution has been emphasized by Ferrone in his research on Filangieri (2003), recently translated into French. Now he is writing about human rights, a problem announced in a chapter of his *Lezioni illuministiche* (2010). An attempt at studying human rights in the eighteenth century has been realized by L. Hunt (2007). She had formerly opened the question in 1996 with regard to the French Revolution. Her new fascinating enquiry explores the essential role of the novel to form general ideas of empathy, sympathy, identification, and knowledge of the Other, inventing a deep need for equality and preparing a diffused persuasion in the West that men, women, and also blacks had the same natural rights. If the function of the novel is well analyzed, it is possible to remark that the research about law, politics, and philosophy is instead reduced. Hunt explains the long self-evidence of the concepts connected with human rights and also the limits of their concrete application. If Rousseau (1762) and Voltaire (1763) individuated the concept and the words, the first attempt to define them emerged in 1776, to be deepened in 1789. Less explored here is the attempt at translating the seventeenth-century tradition of natural law realized for the future by Jean Barbeyrac, which echoed across Europe. Cesare Beccarla is quoted only through Morellet's translation. But the roots of human rights connected Vico, Genovesi, Filangieri, Pagano in Naples, and also Peter Leopold in Tuscany, as implicitly emerged in Furio

Diaz and Venturi's researches and now explicitly in Ferrone and Trampus. On empathy and the novel in the eighteenth century it is now also necessary to read R. Loretelli (2010).

Human rights have been connected with republicanism, as the mainstream (M. Viroli and others), but recent international research has verified that concepts such as liberty, public happiness, the common good, equal rights, patriotism, and also constitutionalism were present in *Monarchisms* (2007). England was not the only nation that evolved in this sense, but also Tuscany and the Hapsburg Empire, where Leopold II corrected Joseph II's centralism, a process interrupted by the French Revolution.

Another problem anticipated by researchers on the history of historiography, the history of philosophy and comparative studies again in Italy (G. Ricuperati, S. Berti, R. Minuti, and L. Simonutti) has been deepened recently by an international team of historians: L. Hunt, M.C. Jacob, and W.W. Mijnhardt, in two recent books exploring a huge work that was born in Amsterdam between 1723 and 1733 in seven folio volumes, involving J.F. Bernard, the editor, and an anonymous writer of the texts and B. Picart, one of the best engravers of his time, *Cérémonies et coutumes religieuses de tous les peuples du monde*. The researchers, who are connected with prestigious American and Dutch universities and international foundations such as the Getty Museum, have stressed the role of this creative compilation, described as *The Book That Changed Europe* (2010) and the first attempt to construct an innovative history of religions, which, through the confrontation of rituals, did not distinguish between monotheisms and polytheisms, opening in this field a true religious universalism in a preanthropological sense. The second book on Picart similarly involves a lot of problems of research and interdisciplinary competences (2010). The two works document not only the indifference of the anthropologists toward this creative and precocious approach, but also the ideological change presented in the later editions. If the Parisian edition by Antoine Banier and Jean Baptiste Le Mascrier was an attempt at orienting the project in a Catholic direction, the French and English editions of the nineteenth century were dominated by regressive and dangerous ideas of the superiority of the West and of Christian (Protestant) monotheism.

Also, the debate about the great challenge suggested by Pierre Bayle, in dialogue with Hobbes and Spinoza, about the 'virtuous atheist' and more generally about a society with civic virtue but without religion, has been reopened with regard to two different, even opposite, perspectives: the first is connected with the Radical Enlightenment and the roots of a cosmopolitan modernity; the second, which emerged in recent years as a result of a diffused consciousness of a strong return of institutional religions, considers our society as postsecular. The defeat of secularization is strongly connected with the pretension of Catholicism to have invented human rights, subtracting them aggressively from the secular Enlightenment. The same happened in other religions and sects that were also against science.

The last object of discussion is the connection between the French Revolution and the Enlightenment. If Israel, in his first two books (2001 and 2006), confirms a deep continuity between them, as the necessary root of our secular present and future, his recent book *A Revolution of the Mind* (2010) seems to like neither the Terror and Robespierre, or his presumed source of inspiration, Rousseau, holding them responsible for totalitarian democracy, and reading the general will, as J. Talmon did, as the main root of future evils. How these arguments can be coherent with continuity is a problem that is not resolved here. For a different analysis, one can cite Ferrone and his *Lezioni illuministiche* (2010), who openly challenges this connection, which was invented in a conservative and conditioning sense by Hegel, and united in the same condemnation Revolution and Enlightenment, without true historical evidence. The author has posed the same perspective in a meeting about the Enlightenment and the French Revolution in Italian debates (Vizille, 2007) the fruits of which were edited in 2010 by G. Bertrand and E. Neppi. Also, B. Baczko in his recent book about the politics of the French Revolution (2008) stresses the difficulty in finding an exit from the Revolution, confronting Washington and Bonaparte. The Napoleonic imperial adventure sounds very different from the correct choice of the American Hero and his renunciation. The Empire was a dramatic way to depart from the Revolution, paid for in Europe with tragic wars, and with a reaction constraining the conquests of the Enlightenment to an underground survival, conditioning *Restauration*, Romantism, and the Age of nations to be without cosmopolitanism, with a regression of liberty and human rights, dreamed for by all Enlightened philosophers as Utopia and reform.

See also: American Revolution, The; Democracy, History of; Democracy: Normative Theory; French Revolution, The; Political Economy, History of; Rationality in Society; Rousseau, Jean-Jacques (1712–78).

Bibliography

Adorno, T.W., Hoorkheimer, M., 1947. Die Dialektik der Aufklärung. Fischer, Frankfurt am Main, Germany (Engl. transl. 1972. Dialectic of Enlightenment. Allen, New York).

Becker, C.L., 1932. The Heavenly City of Eighteenth Century Philosophers. Yale, New Haven, CT, USA.

Benitez, M., 1996. La face cachée des Lumières. Recherches sur les manuscrits philosophiques clandestins de l'âge classique. Universitas Voltaire Foundation, Oxford/Paris.

Berti, S., 2012. Anticristianesimo e libertà. Studi sull'illuminismo radicale europeo. Il Mulino, for Istituto italiano studi storici of Naples, Bologna,.

Blom, H., Laursen, J.C., Simonutti, L. (Eds.), 2006. Monarchisms in the Age of Enlightenment. University of Toronto, Toronot, Buffalo, London.

Cassirer, E., 1932. Die Philosophie der Aufklärüng. Mohr, Tübingen, Germany.

Crocker, L.G., 1959. An Age of Crisis. Man and World in Eighteenth Century French Thought. Johns Hopkins University Press, Baltimore, MA, USA.

Delon, M. (Ed.), 1997. Dictionnaire européen des Lumières. PUF, Paris.

Diaz, F., 1986. Dal Movimento dei Lumi al Movimento dei Popoli. Il Mulino, Bologna.

Ferrone, V., 2003. La società giusta ed equa. Repubblicanesimo e diritti dell'uomo in Gaetano Filangieri. Laterza, Roma-Bari (Engl. transl. S.A. Reinert, 2012. The Politics of Enlightenment. Constitutionalism, Republicanism, and the Rights of Man in Gaetano Filangieri, Anthem Press, New York, London, Delhi).

Ferrone, V., 2007. Una scienza per l'uomo. Illuminismo e rivoluzione scientifia nell'Europa del Settecento. Utet Libreria, Torino.

Ferrone, V., 2010. Lezioni Illuministiche. Laterza, Roma-Bari.

Ferrone, V., Roche, D. (Eds.), 1997. L'Illuminismo, Dizionario Storico. Laterza, Roma.

Filangieri, G., 2003–2004. In: Ferrone, V. (Ed.), Scienza della legislazione. Centro G.Stiffoni sull'Illuminismo, Venezia.

Gay, P., 1967–1969. The Enlightenment. An Interpretation. Knopf, New York.

Geuna, M., Gori, G. (Eds.), 2011. I filosofi e la società senza religione. Il Mulino, Bologna.

Giarrizzo, G., 1995. Massoneria e Illuminismo nell'Europa del Settecento. Marsilio, Venezia.

Goodman, D., 1994, 2002. The Republic of Letters. A Cultural History of the French Enlightenment.

Guerci, L., 1986. L'Europa del Settecento. Permanenze Mutamenti. Utet, Turin.
Hazard, P., 1935. La crise de la conscience européenne. Boivin, Paris.
Hazard, P., 1946. La pensée européenne de Montesquieu à Lessing. Boivin, Paris.
Hunt, L., 2007. Inventing Human Rights. A History. Norton-London, New York.
Hunt, L., Jacob, M.C., Mijnhardt, W.W. (Eds.), 2010a. B. Picart and the First Global Vision of Religion. Getty Publications, Los Angeles, CA.
Hunt, L., Jacob, M.C., Mijnhardt, W.W., 2010b. The Book that Changed Europe. Picart and Bernard's Religious Ceremonies of the World. The Belknap Press of Harvard University Press, Cambridge Mass., London.
Israel, J.I., 2001–2006. Radical Enlightenment: Philosophy and the Making of Modernity 1650–1750. Oxford University Press, Oxford, UK. II. Enlightenment contested 1670–1752. Oxford University Press, Oxford, UK.
Israel, J.I., 2010. A Revolution of the Mind. Radical Enlightenment and the Origins of Democracy. Princeton University Press, Princeton.
Isarael, J., 2011. Democratic Enlightenment Philosophy, Revolution and Human Rights 1750–1790. Oxford University press, Oxford, UK.
Isarael, J., 2012. Enlightenment and Revolution: a critical Foray. Rivista storica italiana CXXIV 3, 1049–1091.
Jacob, M.C., 1980. The Radical Enlightenment. Unwin, London.
Jacob, M.C., 1994. Living Enlightenment: Freemasonry and Politics in Eighteenth Century Europe. Oxford University Press, New York.
Kors, A.C. (Ed.), 2003. Encyclopedia of the Enlightenment, vol. 4. Oxford University Press, Oxford/New York.
Koselleck, R., 1959. Kritik und Krise Zur Pathogenese der bürgerlischen Welt. Alber, Freiburg-München.
Loretelli, R., 2010. L'invenzione del romanzo. Dall'oralità alla lettura silenziosa. Laterza, Roma-Bari.
Minuti, R. Orientalismo e idée di tolleranza nella cultura francese del primo' 700. Olschki, Firenze.
Mortier, R., 1969. Clartés et ombres au siècle des Lumières. Droz, Geneva.
Muthu, S., 2003. Enlightenment Against Empire. Princeton University Press, (n.e. 2009).
Outram, D., 1995. The Enlightenment. Cambridge University Press, Cambridge, UK.
Pocock, J.G.G.A., 1999–2011. Barbarism and Religion, vol. 5. Cambridge University Press, Cambridge.
Porter, R., Teich, M. (Eds.), 1979. The Enlightenment in National Context. Cambridge University Press, Cambridge, UK.
Reinert, S.A., 2011. Translating Empire. Emulation and the Origins of Political Economy. Harvard University Press, Cambridge, MA.
Ricuperati, G., 1994. Le categorie di periodizzazione e il Setteceneto. Per un introduzione storigrapca. Studi Settecenteschi 14, 9–106.
Ricuperati, G. (Ed.), 2000. La reinvenzione dei Lumi. Percorsi storiografici del Novecento. Olschki, Firenze.
Ricuperati, G. (Ed.), 2003. Historiographie et usage des Lumières. Berlin Verlag Arno Spitz, Berlin (Concepts and Symbols of the Eighteenth Century in Europe, EFS).
Ricuperati, G., 2006. Frontiere e limiti della ragione. Crisi della coscienza europea e Illuminismo. Utet Libreria, Torino.
Ricuperati, G., 2009. Radical e Conservative Enlightenment: geografia, storia e critica di due concetti di periodizzazione del Settecento. Rivista storica italiana a. XXXI (1), 929–970.
Ricuperati, G., 2011. Un laboratorio cosmopolitico. Illuminismo e storia a Torino nel Novecento. ESI, Napoli (Quaderno della Rivista storica italiana).
Ricuperati, G., 2012. Le parole di Clio. Aperture, frontiere e confronti interdisciplinari dalla crisi della coscienza europea all'Illuminismo in. Rivista storica italiana CXXIV (1), 1–104.
Robertson, J., 2005. The Case for the Enlightenment. Scotland and Naples 1680–1760. Cambridge University Press, Cambridge.
Scheneiders, W. (Ed.), 1965. Lexikon der Aufklärung. Deutschland und Europa. Beck, München.
Schmidt, J. (Ed.), 1996. What Is Enlightenment? Eighteenth Century Answers and Twentieth Century Questions. University of California Press, Berkeley, CA.
Sebastini, S., 2008. I limiti del progresso. Razza e genere nell'Illuminismo scozzese. il Mulino, Bologna.
Secretan, C.-, Dagron, T., Bove, Laurent (Eds.), 2007. Qu'est-ce que Les Lumières radicales? Libertinage, athéisme et spinozisme dans le tournant ohilosophique de l'age classique. Presseus universitaires de Lyon, Lyon. The meeting had been in 2004 involving M. Jacob and J.I. Israel.
Todorov, T., 2006. L'esprit des Lumières. Laffont, Paris.
Tortarolo, E., 1999. L'illuminismo, Ragioni e Dubbi della Modernita. Carocci, Rome.
Trampus, A., 2008. Il diritto alla felicità. Storia di un'idea. Laterza, Roma-Bari.
Trampus, A., 2009. Storia del costituzionalismo italiano nell'età dei Lumi. Laterza, Roma, Bari.
Venturi, F., 1969–1990. Settecento Riformatore, vol. 5 (in 7 tomes), Torino, Einaudi. (1989, English transl. of the vol. III, The End of the Old Regime in Europe, 1768–1776: The First Crisis. Princeton University Press, Princeton; 1991, English transl. of vol. IV (in two tomes) The End of the Old Regime in Europe, 1776–1789, vols. 2. Princeton University Press, Princeton.).
Venturi, F., 1971. Utopia and Reform in the Enlightenment. Cambridge University Press, Cambridge.
Whatmore, R., 2012. Against War and Empire. Geneva, Britain and France in the Eighteenth Century. Yale University Press.

Enlightenment: Impact on the Social Sciences

John C Torpey, Graduate Center, City University of New York, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the relationship between the Enlightenment and the modern social sciences. It details the different national traditions of 'Enlightenment' and how these have affected the trajectory of the social sciences. It explores both the ways in which the social sciences have sought to carry out an 'Enlightenment project' and the doubts that have been raised about that project along the way.

The modern social sciences are children of the Enlightenment. The signature statement heralding that movement in European intellectual history was Immanuel Kant's famous dictum (1959b: 85–87), "*Sapere aude*! (Dare to know!) Have courage to use your own reason! – that is the motto of enlightenment." The use of one's own reason was to be contrasted, of course, with that dispensed by others, perhaps especially by the hierarchically organized and doctrinally authoritarian Catholic Church. Hence the Enlightenment was in certain respects a continuation of the Protestant Reformation's insistence on the more egalitarian 'priesthood of all believers' and on the ability of each individual to find his or her own path to god. People should use their own reason *publicly*, according to Kant, and all that was required for the public use of reason was political *freedom*. The idea of Enlightenment was thus bound up from its origins with an emancipatory political outlook that has generally continued to accompany it ever since. The anticlerical character of the Enlightenment was perhaps most apparent in the French case, where the Catholic Church comprised part of the 'old regime' and was part and parcel of the unfreedom that Kant saw as the chief barrier to Enlightenment. Not surprisingly, then, Kant was a strong supporter of the French Revolution. His stress on the difficulty of achieving truly moral action – reflected in his distinction between the 'is' and the 'ought' – would resonate throughout much subsequent social science thought.

The establishment of social 'science' as an endeavor separate from the wisdom of the churches would mean that it would be forced to distance itself from theology, philosophy, and other modes of knowing that lacked a 'scientific' foundation. That scientific foundation would be derived above all from the insistence on the this-worldly, empirical character of the social sciences. Hence Weber would later say (1949: 72) that social science is an "empirical science of concrete reality" (*Wirklichkeitswissenschaft*) and that definition has very much shaped the nature of the modern social scientific disciplines, even if these fields may in fact be profoundly different from the natural sciences. The need to distance it from religious modes of thought has driven the self-understanding of the social sciences as 'objective' and 'disinterested,' leading to a widespread divorce between *knowing about* the world and *doing something in* the world. The link between contemplation and action that had animated classical Greek thought was severed in the process. There is something ironic in this, given that some of the founding lights of the social science disciplines, such as Tocqueville, Marx, Weber, and Durkheim, were very much involved in the politics of their day, some of them holding high public office (Tocqueville was briefly foreign minister of France). The divorce between knowledge and action would later be intensified by the massive expansion of higher education from the 1960s onward, as a result of which academics came to constitute an audience for each other and were no longer required to write for broader 'lay' audiences. The extent of the segregation of thought and action, and of the 'is' from the 'ought,' has varied, however, in the different national traditions of the social sciences that emerged from the eighteenth- and nineteenth-century background. As compared to European and other variants of the social sciences, American social science has been seen as distinctively "modeled on the natural rather than the historical sciences and imbedded in the classical ideology of liberal individualism" (Ross, 1991: xiii). Indeed, the nature of the American social sciences is arguably another aspect of 'American exceptionalism' (see Lipset, 1996).

In addition to strengthening the role of human reason in public life generally, the Enlightenment 'project' had a good deal to do with finding ways to pacify human life and to overcome the scourge of war. Kant himself, of course, articulated a design for 'Perpetual Peace,' in which he proposed an idea for a sort of 'league of nations' that anticipated later developments along these lines. He also suggested that countries' predisposition toward violence beyond their borders was a function of their internal structure, and that republics, in particular, tend not to go to war. Kant thus laid the groundwork for the theory of the so-called 'democratic peace,' which holds that democracies do not go to war with one another. The theory generated considerable research and controversy in the 1990s and after. Kant also drew on a body of thought articulating the theory of '*doux commerce*' – the idea of the civilizing function of commercial activity. One version of the thesis, which was advanced by Montesquieu, Hume, and others, came from the American radical Thomas Paine. In his 1792 tract *The Rights of Man*, Paine wrote (1992: 172) that "commerce … is a pacific system operating to cordialize mankind, by rendering nations, as well as individuals, useful to each other. If commerce were permitted to act to the universal extent it is capable, it would extirpate the system of war…." The idea of commerce as a system for regulating the passions of human beings would prove a central theme in the Enlightenment-era thinking about economics, war, and modern society (see Hirschman, 1977).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03203-7

The Scottish moralists such as Adam Smith, Adam Ferguson, and others were preoccupied with related themes, examining the conditions under which republican government and civic humanism could exist under conditions of an advancing division of labor. Smith, perhaps most frequently associated with his 1776 treatise *The Wealth of Nations*, came to be regarded chiefly as a thinker concerned with economic activities, but – writing before the separation of social science and moral philosophy had been completed – Smith asked about the causes and consequences of those activities. "It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner," Smith wrote (1985: 119), "but from their regard to their own interest." Along with man's supposedly innate inclination to 'truck and barter,' Smith developed a rational model of economic motivation that would dominate the social sciences – whether as touchstone or as bête noire – thereafter. Meanwhile, Ferguson was concerned with the requirements of citizenship and the promotion of civic virtue, among other things with regard to the degree to which citizens should be involved in defending the state of whose acts they would be the authors. In his *Essay on the History of Civil Society*, Ferguson advances a conception of the citizen-soldier that would, in its turn, resonate through later social scientific thinking about the relationship between citizens and military institutions. The concerns of Smith and Ferguson would resurface over and over again in subsequent writing about modern society, economics, and politics.

Soon thereafter, in the aftermath of the revolutionary *levée en masse* and the Napoleonic wars, the German officer Carl von Clausewitz would endeavor to think systematically about the place of war in modern societies in which ordinary people were increasingly included in political decision-making. Clausewitz's dictum (1984: 87) that "war is … a continuation of political intercourse … with other means" was intended to reflect the nature of war in an age in which citizen armies were becoming the order of the day. Gradually, however, as the social science disciplines developed, sociology and political science went their separate ways, and political science – the study of 'high politics' – more or less exclusively held onto the portfolio of war and international conflict. Meanwhile, sociology – the study of 'low politics' – largely abandoned the study of war, and tended to operate as if Enlightenment dreams of ending war had been realized and as if war were not an endemic aspect of modern society (see Joas and Knöbl, 2013).

From a more systematic perspective, however, the origins of the modern social sciences are often traced back to the French thinker Auguste Comte, who sought to create a 'positive' science emphasizing the observation of concrete phenomena (see Lenzer, 1975). Imbued with the postrevolutionary aims of order and progress, Comte sought to understand what *is* in order to understand what *would be* – in other words, to predict the future and make of it the orderly unfolding of human progress. The emphasis on prediction and on general laws of social development would come to play a pervasive role in social scientific thinking far into the twentieth century. Comte developed a scheme for understanding human history, according to which there were three stages of human development – the theological, the metaphysical, and the positive or scientific. Comte thus introduced a scheme in which positive or scientific thought was sharply contrasted with that of religion. The legacy of the French Revolution and its anticlericalism is straightforwardly apparent. Comte's contemporary Alexis de Tocqueville did not share the harsh dismissal of religion that pervaded the Revolution, and indeed regarded religion as a mode of thinking that was never likely to die out entirely. 'Religion,' Tocqueville wrote in an 1857 letter to Sophie Swetchine, "is only a particular kind of hope, and it is also as natural to the human heart as hope itself" (quoted in Brogan, 2006: 53). Still, the connection between the Enlightenment, the social sciences, and what would later come to be known as the 'secularization thesis' – the idea that religiosity would decline as modernity advanced – was growing stronger.

The thesis could not have acquired its power from Tocqueville's analysis of the varying place of religion in the French Revolution and in the United States, however, as he strove to point out that religion played a very different role in the two contexts. As an element of the 'old regime' in France, according to Tocqueville (1998), the revolutionaries who wanted to dispense with the French monarchy were ill-disposed toward the Church not because they were antireligious but because they were opposed to the regime of which the Church was a part. In the United States, by contrast, religion and freedom marched hand-in-hand, the one reinforcing the other. The result, according to Tocqueville (2000), was that religion flourished in the United States – despite its constitutionally mandated disestablishment – whereas it withered in France due to its association with a power rendered illegitimate by the advance of democratic ideas and forces. As his relatively sympathetic comments on religion may suggest, Tocqueville was not a great admirer of the Enlightenment, which he regarded as the underlying source of the "abstract theories of government" (1998: 197) that led to the Revolution and to the extremes of the terror. Notwithstanding a guarded sympathy for 'democracy,' Tocqueville tended to believe that local people knew better than central rulers what was best for themselves, and that the abstractions underlying the French Revolution led to an unfortunate fate for postrevolutionary France.

The connection between Enlightenment-inspired views of religion and the social sciences would intensify in the work of Emile Durkheim and Max Weber. Durkheim distinguished sharply between religion and magic in terms of their capacity for generating institutions, noting in *The Elementary Forms of the Religious Life* (1915: 60) that "there is no church of magic." In that respect, for Durkheim, 'religion' was an advance over magic because it brought people together and enhanced human solidarity. Still, traditional religion was in the process of being supplanted by a form of religion more consistent with modern society. In his contribution to the Dreyfus affair in 1901, Durkheim wrote (1973: 52) that "man has become a god for man, and … he can no longer create other gods without lying to himself." The modern 'cult of the individual' was not about egoism in the negative sense, a celebration of utilitarianism, but rather a glorification of that dignity that human beings share in common. A member of the *Ligue des droits de l'homme*, Durkheim was instead invoking the idea of human rights as a replacement for traditional religion. Yet even if traditional religion was on its way out, in Durkheim's view, every society had and would always have a sort of religious basis. This idea of 'civil religion,' first adumbrated by Rousseau in *The Social Contract* (1973), was carried forward by one of

Durkheim's most important heirs, the sociologist Robert Bellah (1991).

Weber's position on the fate of religion was more ambiguous. He viewed religion as something of an endangered species as well, but in the sense that it would be shunted into the realm of the irrational rather than that it would disappear entirely. Weber took this position against the background of a larger commitment to the idea that human history could be understood above all in terms of a process of *rationalization*. This process took many forms and, in Weber's view, was fundamental to understanding the diverse routes taken by different civilizations into the modern world. Weber's comparative studies of the world religions were titled in German *Die Wirtschaftsethik der Weltreligionen* (*The Economic Ethics of the World Religions*), signaling the degree to which they were – like so much of Weber's writing – a response to Marx's analyses of the modern world. The studies of the world religions proposed to make sense of the 'rise of the West' – that is, why modern 'rational' capitalism emerged first in the West and not somewhere else – and paid less attention to political matters, except insofar as these were thought to foster or hamper the emergence of a 'rational' acquisitive ethos. Weber held Protestantism's stress on the idea of a 'calling' and especially Calvinism's dour view of frivolous self-enjoyment to be the outlook that harmonized – felicitously if unintentionally – with the workings of rational capitalism. Yet, with his typical proclivity for irony and tragedy, Weber argued that the elective affinity between Protestant ethical doctrines (*die protestantische Ethik*) and the spirit of capitalism would lead to the construction of an 'iron cage,' one that would promote bureaucratic rationalization and thus a further strengthening of the bars of the cage. For all its rationality, Weber was saying, rationalization was a trap.

For his part, Sigmund Freud continued the Enlightenment program of enthroning reason in human affairs by proposing to replace the chaos of the human instincts with rational self-control. (In the original German, Freud used the term *Triebe* or 'drives,' not *Instinkt*; the term 'instinct' was supplied by James Strachey, the English translator of much of Freud's oeuvre. Still, the term 'instinct' has become canonical in English.) To use the language of psychoanalysis that he developed, "where id was, there ego shall be" (Freud, 1965: 71). Freud sought to tame the irrational by mapping the human mind and developing a procedure for understanding unruly, unproductive human impulses. Freud's powerful analyses held out the promise that human unreason could be reduced, if not eliminated, by understanding the past experiences that unconsciously shaped present behavior. The technique, which made talk the vehicle of mental healing, came to be known as the 'talking cure.' With regard to the prospects of Enlightenment in human affairs, Freud himself lost some heart as a witness to the catastrophic carnage of World War I in Europe. Having previously posited a 'pleasure principle' as the chief driver of human behavior, in 1920 Freud (1961a) proposed that human beings also had a 'death instinct' that was an ineradicable part of their instinctual makeup. In his most sustained and expansive reflection on human civilization, Freud (1961b) argued against socialists and others that this aggressive drive could be lessened somewhat by greater social equality, but could never be eliminated entirely. Psychoanalysis took bourgeois Europe by storm, and – despite questions about its 'unscientific' status – profoundly influenced major twentieth-century social scientists such as Talcott Parsons (see Parsons, 1982), who absorbed European intellectual culture as a young man sojourning at the London School of Economics and the University of Heidelberg in the mid-1920s. Yet in a testament to the variable forms that 'Enlightenment' took in different social settings, Freud's texts had to be translated into a scientized (Greek) language in order to gain broader favor in American circles (see Bettelheim, 1983). For better or worse, Freud's oeuvre decisively shaped the *Triumph of the Therapeutic* (Rieff, 1966).

Weber's jaundiced view of rationalization would later be taken up by two sophisticated 'Western Marxist' thinkers, themselves deeply influenced by Freud, namely Max Horkheimer and T.W. Adorno. The two thinkers of the so-called Frankfurt School of 'Critical Theory' argued in a wartime tract that the Enlightenment had doubled back on itself, so to speak, and become a sort of curse, ensnaring humanity in a self-made web of unreason. The domination of 'instrumental reason' – the sort of rationality that Weber suggested was paramount in the 'iron cage' of modernity – had deprived humanity of the ability to think in terms of 'substantive reason,' the sort that treats human beings not as means but as ends in themselves (Horkheimer and Adorno, 1972). The irony, of course, is that one of Kant's central ideas – his 'practical imperative' – had been precisely that one should "treat humanity, whether in your own person or in that of another, always as an end and never as a means only" (1959a: 47). For Horkheimer and Adorno, soured on the Stalinist version of socialism but forced to flee an anti-Semitic behemoth they regarded as an outgrowth of capitalism, the world had lost the capacity to live up to Kant's imperative.

Horkheimer and Adorno's deeply pessimistic treatise from World War II era was relatively little known in the non-German-speaking world until it was translated into English in the 1970s. At that point, however, it resonated with a broader disenchantment with Western political dominance that was then spreading among the intellectual classes that had entered university life with the 1960s expansion of tertiary education. Consistent with the Frankfurt theorists' intuitions, the Holocaust and the atomic bomb fueled worries that the scientific advances once seen as advances stemming from the Enlightenment were turning out to be the means for the extermination of the species. One of the signal contributions was that of Michel Foucault, a French thinker whose Nietzschean orientation to philosophical questions spurred an exploration of the connections between power and knowledge. Whereas earlier social scientists had typically been driven by the belief that knowledge is worth acquiring because 'knowledge is power' and could be used to 'speak truth to power,' Foucault identified subterranean connections between 'regimes of truth' and 'discourses' that sustained a seamless link between 'power/knowledge.' Foucault's work was that of a disenchanted member of the French Communist Party, long known as a 'Stalinist' party and hence not particularly interested in (or actively opposed to) his concerns with madness, sexuality, and the politics of knowledge. Foucault's politics would thus drift toward the notion of 'biopolitics,' asserting the primacy of the discourses that governed the growth and regulation of populations in the modern period. His writings would thus

resonate strongly with emerging social and intellectual movements concerned with sexuality, identity, and the like. His critique of the 'universal' intellectual associated with the Enlightenment and his advocacy of a new 'specific' intellectual fit neatly with a postmodernist critique of 'grand narratives' (Lyotard, 1984) that coincided with the shift from class politics to identity politics in the 1970s and after.

Relatively little of this philosophical controversy influenced the main body of the political and social sciences, which emerged from World War II – especially in the United States – as increasingly empiricist disciplines. The impact on the more theoretical and interpretive margins of these fields were affected, however, as was anthropology, which proved to be more open to the 'interpretive turn' then sweeping the more 'humanistic' social sciences. Behind the leadership of Clifford Geertz (see esp. 1977), James Marcus and others had drawn the field in the direction of hermeneutics and the idea that societies could be 'read' as texts or discourses. These trends drew perhaps more on the work of Paul Ricoeur (1981) than on Foucault, but the shift away from a focus on social and kinship structures was pronounced. Geertz later regretted what had been done in his name (see Geertz, 2000), but the postmodern genie and its skepticism about the Enlightenment was out of the bottle.

Not everyone was prepared to concede that the 'grand narratives' associated with the Enlightenment destined for the junk heap of history, however. Perhaps the most influential and thoroughgoing counterattack was mounted by the German social philosopher Jürgen Habermas. In a series of lectures on *The Philosophical Discourse of Modernity* (1987), Habermas trained his critical attention on a variety of German and French antagonists of Enlightenment-style critical reason. One of his principal *bêtes noires* was Foucault, whom he understood as a crucial figure in the 'counter-Enlightenment' of the last part of the twentieth century. Foucault, Habermas wrote (1987: 282–283), "does not want to refine the language game of modern political theory (with its basic concepts of autonomy and heteronomy, morality and legality, emancipation and repression) and turn it against the pathologies of modernity – he wants to undermine modernity and its language games" *tout court*. As a student of Horkheimer and Adorno whose thought has been deeply shaped by what he regarded as Germany's catastrophic twentieth-century descent into irrationalism, Habermas had little sympathy for what he saw as a flirtation with antirationalism. That this flirtation came by way of absorption with the radical antidemocrat Friedrich Nietzsche only made it more suspect.

Criticism of the postmodern turn in the social sciences also came from a somewhat unlikely source, namely Edward Said. In the 1994 'Afterword' to his widely acclaimed volume *Orientalism*, Said (1994: 349) decries the 'Eurocentrism' and 'the almost decorative weightlessness of history' in postmodern writing, seeing in postcolonial studies at least a search for structures of domination and resistance. Postmodernism, by contrast, was marked, according to Said, by "a preponderance of theoretical and aesthetic emphasis stressing the local and contingent," an unpromising point of departure if one is interested in 'speaking truth to power,' as Said himself clearly did in his critique of 'orientalist' ways of thinking.

Soon thereafter, the anti-Enlightenment tendencies in the academy came in for a more visible challenge at the hands of a New York University physicist named Alan Sokal. Perturbed at the anti-Enlightenment trends in intellectual life and a self-described political progressive, Sokal set out to see whether he could persuade a major journal of the newly emerging field of 'cultural studies' to publish "an article liberally salted with nonsense if (a) it sounded good and (b) it flattered the editors' ideological preconceptions" (Sokal, 1996b). He could; *Social Text* accepted the article, apparently without serious evaluation by knowledgeable scholars as would be standard practice for a peer-reviewed scholarly journal. The aftershocks were not to remain confined to the ivy tower. The ensuing controversy was played out in the pages of academic journals and on airwaves. The editors of *Social Text*, among the more prestigious scholars in the world associated with cultural studies and post-modernism, suffered a rather serious black eye, as did the intellectual currents with which they were associated. That is hardly the same as saying the trends they represented and promoted would disappear, however.

Indeed, the discomfiture with the Enlightenment, with ('Western') modernity, and with the supposed failings of 'reason' has persisted. One of the most vigorous critics of this cultural vein in recent years has been the neuroscientist Steven Pinker. Best known for his writings on cognitive psychology and linguistics, in 2011 Pinker published a massive study arguing – against considerable evidence to the contrary – that violence had declined in the modern world as compared to the earlier societies in which no state held a "monopoly on the legitimate use of violence" (Max Weber's famous definition of the state). According to Pinker (2011: xxii), "a large swath of our intellectual culture is loath to admit that there could be anything good about civilization, modernity, and Western society." Pinker's book was perhaps as much directed at this audience as it was at students of violence in the modern world. Whether he is correct about the decline of violence on a world scale remains a matter of considerable debate; the judgment depends significantly on the reliability and generalizability of evidence of intrinsically recondite character (see Lawler, 2012).

But the effort to establish whether, against all appearances and our intellectual prejudices, violence has actually declined with the shift to state-dominated societies constitutes one of those quintessentially social scientific questions bequeathed to us by the Enlightenment. We cannot presume to know the answer without investigation and thoughtful weighing of the difficult evidence. We must dare to use our reason. That is surely what the modern social sciences understand themselves to be doing, even if there have been doubts along the way about whether that reason is as disinterested as it imagines or as helpful to the human species as its practitioners believe.

See also: American and French Revolutions: Impact on the Social Sciences; Applied Social Research, History of; Civil Religion; Enlightenment; Foucault, Michel (1926–84); Freedom, History of; French Revolution, The; Habermas, Jürgen (1929–); Kant, Immanuel (1724–1804); Postmodernism in Sociology; Postmodernism: Philosophical Aspects; Public Reason.

Bibliography

Bellah, Robert N., 1991[1967]. Civil religion in America. In: Beyond Belief: Essays on Religion in a Post-traditionalist World. University of California Press, Berkeley, pp. 168–189.

Bellah, Robert (Ed.), 1973. Individualism and the intellectuals. Emile Durkheim on Morality and Society. University of Chicago Press, Chicago, pp. 43–57.
Bettelheim, Bruno, 1983. Freud and Man's Soul. Alfred A. Knopf, New York.
Brogan, Hugh, 2006. Alexis de Tocqueville: A Life. Yale University Press, New Haven.
Clausewitz, Carl von, 1984[1832]. On War (M. Howard and P. Paret, Trans. and Ed.). Princeton University Press, Princeton.
Durkheim, Emile, 1915. The Elementary Forms of the Religious Life (J.W. Swain, Trans.). The Free Press, New York.
Ferguson, Adam, 1767. Essay on the History of Civil Society.
Foucault, Michel, 1980. Truth and power. In: Gordon, Colin (Ed.), Power/Knowledge: Selected Interviews & Other Writings, 1972–1977. Pantheon, New York, pp. 109–133.
Foucault, Michel, 1984. What is enlightenment? In: Rabinow, Paul (Ed.), The Foucault Reader. Pantheon, New York, pp. 32–50.
Freud, Sigmund, 1961a[1920]. Beyond the Pleasure Principle (J. Strachey, Trans.). W. W. Norton, New York.
Freud, Sigmund, 1961b[1930]. Civilization and Its Discontents (J. Strachey, Trans.). W. W. Norton, New York.
Freud, Sigmund, 1965[1933]. The dissection of the psychical personality (J. Strachey, Trans.). In: New Introductory Lectures on Psychoanalysis. W. W. Norton, New York, pp. 51–71.
Geertz, Clifford, 1977. The Interpretation of Cultures. Basic Books, New York.
Geertz, Clifford, 2000. Available Light: Anthropological Reflections on Philosophical Topics. Princeton University Press, Princeton.
Habermas, Jürgen, 1987[1985]. The Philosophical Discourse of Modernity: Twelve Lectures (F. Lawrence, Trans.). MIT Press, Cambridge, MA.
Hirschman, Albert O., 1977. The Passions and the Interests: Political Arguments for Capitalism before Its Triumph. Princeton University Press, Princeton.
Horkheimer, Max, Adorno, T.W., 1972[1944]. The Dialectic of Enlightenment (J. Cumming, Trans.). Continuum, New York.
Joas, Hans, Knöbl, Wolfgang, 2013[2008]. War in Social Thought: Hobbes to the Present (A. Skinner, Trans.). Princeton University Press, Princeton.
Kant, Immanuel, 1959a[1785]. In: Beck, Lewis White (Ed.), Foundations of the Metaphysics of Morals. Bobbs-Merrill, Indianapolis.
Kant, Immanuel, 1959b[1784]. What is enlightenment? In: Beck, Lewis White (Ed.), Foundations of the Metaphysics of Morals. Bobbs-Merrill, Indianapolis.
Lawler, Andrew, May 18, 2012. The battle over violence. Science 336 (6083), 829–830.
Lenzer, Gertrud, 1975. Auguste Comte and Positivism: The Essential Writings. Harper Torchbooks, New York.
Lipset, Seymour Martin, 1996. American Exceptionalism: A Double-Edged Sword. Norton, New York.
Lyotard, Jean-François, 1984[1979]. The Postmodern Condition: A Report on Knowledge (G. Bennington and B. Massumi, Trans.). University of Minnesota Press, Minneapolis.
Paine, Thomas, 1992[1792]. The Rights of Man. Hackett, Indianapolis, IN.
Parsons, Talcott, 1982[1952]. The superego and the theory of social systems. In: Mayhew, Leon (Ed.), Talcott Parsons on Institutions and Social Evolution. University of Chicago Press, Chicago, pp. 129–144.
Pinker, Steven, 2011. *The Better Angels of Our Nature: Why Violence Has Declined*. Viking, New York.
Ricoeur, Paul, 1981. Hermeneutics & the Human Sciences: Essays on Language, Action, and Interpretation (J. Thompson, Trans. and Ed.). Cambridge University Press, New York.
Rieff, Philip, 1966. The Triumph of the Therapeutic: Uses of Faith after Freud. University of Chicago Press, Chicago.
Ross, Dorothy, 1991. The Origins of American Social Science. Cambridge University Press, New York.
Rousseau, Jean-Jacques, 1973[1762]. The Social Contract and Discourses. E. P. Dutton & Co, New York.
Said, Edward, 1994[1978]. Orientalism. Vintage, New York.
Smith, Adam, 1985[1776]. The Wealth of Nations. Penguin, New York.
Sokal, Alan D., (Spring/Summer) 1996a. Transgressing the boundaries: towards a transformative hermeneutic of quantum gravity. Social Text 46/47, 217–252.
Sokal, Alan D., May 1996b. A physicist experiments with cultural studies. Lingua Franca. Available at http://linguafranca.mirror.theinfo.org/9605/sokal.html.
Tocqueville, de Alexis, 2000[1835/1840]. Democracy in America (H. Mansfield and D. Winthrop, Trans. and Ed.). University of Chicago Press, Chicago.
Tocqueville, de Alexis, 1998[1856]. In: Furet, François, Melonio, Françoise (Eds.), The Old Regime and the Revolution. University of Chicago Press, Chicago.
Weber, Max, 1949[1904]. 'Objectivity' in social science and social policy (E. Shils and H.A. Finch, Trans. and Ed.). In: The Methodology of the Social Sciences. The Free Press, New York, pp. 49–112.

Entertainment

Martin Zerlang, The University of Copenhagen, Copenhagen, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Entertainment is defined as play put on display. The etymology of the word points to the inner connection between entertainment and community. A map of the world of entertainment is sketched out, based on the oppositions between identification and alienation and between activity and passivity. It is shown how entertainment has a history leading from the festivals and spectacles of antiquity through medieval carnival to courtly diversions. The nineteenth century is described as the age of modern mass entertainment, and it is shown how in the twentieth and twenty-first centuries, the interplay between entertainment and urbanization gives rise to shopping malls and entertainment cities. The article ends with a note on the globalization of entertainment.

Entertainment is play, but play put on display. In the modern, Western or Westernized world, plays and games have become commodities for sale in that special branch of economic life called show business. There is no limit to the media, art forms, or genres to be marketed by the agents of show business, but the term 'light entertainment' marks the quintessence of entertainment: ease, glitter, and lightness. Entertainment has no other aim than to please as long as 'the music plays.' In 1844, *Tivoli-Avis*, the newspaper of the Tivoli Gardens, the amusement park that was established in 1843 on the military fortifications around Copenhagen, stated that "Tivoli produces only one thing: joy!" (Zerlang, 1997: p. 97). In 1957, an advertisement for the newly founded Disneyland called it "a new industry with happiness as its principal product" (Findlay, 1992: p. 75). Entertainment, in the modern sense of the word, may thus be defined as the commercialization of the pursuit of happiness.

Entertainment and Community

The pursuit of happiness has no doubt been a primary driving force of mankind since the beginning of time. One of the ancient Greek philosophers, Democritus, once stated that without festivity life is like a long road without an inn, and during the apogee of the Classical Athenian era, more than 40 feasts and festivals were celebrated every year. Arranging the festivals was a public office of the highest importance. Festivals also play a major role in Plato's description of the ideal state, because they are community building: "There is no greater good in a state than that the citizens should be known to one another" (Laws, 738).

The word entertainment actually refers to social exchanges. It is derived from Latin: *inter* means 'among' or 'between,' and *tenere* means 'to hold something mutually' or 'to keep something up.' Thus, entertainment is a social relationship and it is concerned with social relations. In premodern times, entertainment primarily indicated material support; with modernization, entertainment is about spiritual support. Support is another word for something uplifting. Listeners to happy jazz or pop songs expect 'to get carried away' by the ease of pop ballads or the swing of jazz music. And the meaning of dance is a temporary stepping out of the world of everyday life.

Happiness is uplifting; in an unhappy state of mind, one feels down. The individualism of modern culture has traditionally been accompanied by the problem of boredom. During the nineteenth century, writers repeatedly described and deplored boredom, spleen, ennui and Langeweile, and one of the pioneers of modern entertainment, the composer Jacques Offenbach, emphasized the importance of mutual support by declaring that his dream was "to found a share holding company of mutual insurance for the prevention of boredom" (Kracauer, 1976: p. 101). In his operetta *La vie parisienne* (1866), Offenbach celebrates and satirizes the collective festivity at railway stations, hotels, exhibitions, cafés, parties, and masked balls. Individual unhappiness disappears in the crowd.

Mapping Entertainment

Mapping the world of entertainment may start by a reflection on the expression: 'only entertainment.' If entertainment is 'only entertainment,' this indicates two things: that entertainment has become a specific field of activity and that entertainment has no importance; it is 'just fun.' As a specific field, as 'show business,' entertainment differs from play, although it is based on forms of play. And as a sphere of activities allegedly of no importance, entertainment differs from art, although it apparently enjoys the same autonomous status. Kant, who characterized the aesthetic attitude by its 'disinterested pleasure,' would consider the search for laughs, thrills, tears, and sex as 'interests' that would degrade entertainment in relationship to art. Others fulminated against the reduction of serious art to an entertaining commodity. A symptom of the great divide between art and entertainment that developed during the nineteenth century is Robert Schumann's reaction to the new, immensely popular waltz music: he wrote that he heard the gold rattling in Strauss Sr.'s pocket as he conducted.

The reduction of entertainment to 'only entertainment' may be interpreted as an effort to deny or exorcise the real importance of entertainment in modern society: as a substitute for religion, as a purveyor of a new mythology, as a political element, and as a key sector in the economy.

The question now is whether it is possible to see a pattern behind all the variety. In the same way as play may be divided

into a clear pattern of categories, the forms of entertainment may be organized according to the way in which they express the relation between individual and society. According to Roger Caillois (1967), games may be classified into four main headings based on different ways of handling the oppositions between identification and alienation and between activity and passivity.

1. The first pole is 'agon,' competitive entertainment based on active identification. Here one finds not only chariot races, chivalrous tournaments, and sports but also narratives of action and adventure whether told in novels, films, television (TV) shows, or computer games. The hero seeks recognition of his identity, and the audience identifying with the hero seeks recognition in his victory. The characteristic emotional experience of competitive entertainment is suspense, which gives the reader or spectator a vicarious, but intense bodily feeling of taking part in the action. Open spaces often characterize the world of 'agon.'
2. The second pole is 'alea,' the entertainment of chance based on passivity and identification. Here, one finds not only all games of chance but also stories of horror. The hero is more often than not a victim, and if he ends up as winner it is not by his own merit. Identification does not consist in recognition of his name, but identification of the number that makes him a winner – or victim. The characteristic emotional experience is the thrill that gives the reader or spectator a vicarious feeling of being victimized or saved by alien forces. Closed spaces often characterize the world of 'alea.'
3. The third pole is 'mimicry,' all kinds of theater, tableaux vivants, masquerades, variety shows, parodies, etc. The hero takes on a heroic, villainous, or just different persona, and if he gains a name, it is through this talent for impersonating. By definition, the whole gamut of emotions is invoked by this type of entertainment. As is obvious in melodrama, thresholds often energize this kind of entertainment.
4. The fourth pole is 'vertigo,' the high-spirited pleasures of a temporary suspension of one's everyday identity, promoted by roller coasters and swings, by waltzes and rock-n-roll, and by comedy. Here, the hero hands himself over to impersonal forces, but in contrast to the gambler, he does not look for any profit but the pleasure of vertigo. Even though emotional distance from the capers of clowns of all sorts is one of the enjoyments of comedy, the readers or spectators who give themselves over to laughter are also looking for the pleasure of letting go and losing control.

Bread and Circuses

In Western culture, the concepts of spectacle and circus go back to antiquity. The idea of the ludic, i.e., the playful, is derived from the Roman concept of *ludus* (chariot races); and the idea of the aleatory, i.e., gambling with chance, is derived from the Roman word for dice. In the course of Roman history, religious rituals – such as gladiatorial fights in connection with funeral rites – are restructured into cultural forms much closer to the modern idea of entertainment. In the tenth of his satires, Juvenal coined the phrase 'bread and circuses,' and this verdict on a social policy replacing the active political commitment of the citizens by their passive reception of material and mental entertainment has been an inspiration for modern cultural critique. According to Patrick Brantlinger, equating the social decay of the Roman Empire with the 'massification' of modern society constitutes a 'mythology of negative classicism.' However, seeking parallels between the ancient and modern worlds was not the preserve of critics. The founding fathers of modern entertainment also cultivated the Roman or classic connection: The great panorama building erected in London in 1827 was called the Colosseum. Las Vegas still boasts a Caesar's Palace.

Carnival

In the middle ages, Roman festivals such as the Saturnalia and Lupercalia metamorphosed into the extant tradition of carnival. Entertainment is often characterized as recreation, and recreation is what characterizes the carnival. The rural and pagan carnival would re-create the natural world, the Christian world would re-create the religious community, and the urban carnival would re-create the power and splendor of the city. The world is turned upside down, boys become bishops, men dress up as women, and processions carry gigantic sausages through the streets, and in this creative chaos, the sense of community is reanimated.

In contrast to commercial entertainment, those taking part in the carnival were not free to choose their leisure. Even if the carnival was entertaining, it was not pure entertainment. People were not individuals buying a ticket to a show, but members of a community obliged to follow the calendar. Shrovetide was a once-a-year experience.

From the perspective of the church or the city council presiding over the carnival, the aim of the event was to make the different groups of society accept the existing hierarchy. The creative chaos was also a deterrent. The blasphemous words, the mock battles, etc. would illustrate the consequences of real social change.

The authorities could use the carnival. That is the only plausible explanation of their pivotal role in staging the processions and plays of the carnival. But as long as the carnival lasted, the situation was potentially open, the functions and the future of society were put at stake, and the safety valve might turn into an explosive risk.

In his influential book *Rabelais and His World* (1968), Mikhail Bakhtin emphasizes the emancipatory aspects of the carnival. To him it exalts the people versus those in power, the lower social and bodily strata versus the repressive idealism of official culture. In *Le Carnaval de Romans* (1979), Emmanuelle Le Roy Ladurie closely analyzes one specific carnival, in the year 1580, and here the sources tell about the confrontation between disciplining forces and rebellious forces within the carnival.

Courtly Diversions

In the fifteenth and sixteenth centuries, a process started in Europe during which the paths of popular culture and elite culture diverged. Politically, as power became centralized, the powerful made a spectacle of themselves through fireworks,

pageants, and allegorical plays while reducing their subjects to spectators. Economically, the reformation and the rise of the middle class, who knew that time was money, was accompanied by a war against idle leisure.

To the aristocracy as well as the middle classes, the experience of time, the abstract time of the clock, became a new social experience, but while the aristocracy actively cultivated a culture of pastimes and diversions, the puritan middle classes rejected all pastimes that would divert their attention from work. While the aristocracy indirectly became the pioneers of modern entertainment, the puritans just as indirectly created the patterns of modern cultural critique.

The culture of *divertissements* at the French court developed as the process of centralization and the politics of absolutism transformed the 'nobility of the sword' into courtiers. Leisure time at the court became a psychological as well as a political problem. To aristocrats used to fighting, hunting, etc., life at court was inactive and boring, and they had to do something to pass their time. Diversions would compensate for the lack of activity and power, and entertaining discussions and dances, etc. would at the same time create a social bond between them. The intricate equestrian displays known as *carrousels* epitomized this domestication of a warrior class through entertainment. A trace of this phase in the history of entertainment may still be seen in the predilection for a lavish, baroque style in merry-go-rounds and other fairground architecture.

Commerce and Buying Fun

In the eighteenth century, the word entertainment (or in German: *Unterhaltung*) had a breakthrough, but this was in reference to private or domestic entertainment. Private evening entertainment would include music, dance, and the creation of *tableaux vivants*, but it was the expansion of the book market that really made the home a center of entertainment. This market became the great provider of entertainment, and these entertainments were open to everyone with sufficient means. In 1725, Daniel Defoe noted that writing had become "a very considerable Branch of the English Commerce" (Watt, 1957: p. 55), and an increasingly popular genre such as 'the penny dreadful' indicates the importance of the market. In the Scandinavian languages, the word commerce became synonymous with having fun.

Traditional entertainments did not disappear from 1 day to the next. At the beginning of the nineteenth century, Joseph Strutt wrote *The Sports and Pastimes of the People of England* (1801) in which it is obvious that the idea of public entertainment for the general public was only beginning to get a foothold. In his table of contents, the pastimes are divided according to time, place, and social group: 'Rural exercises practiced by persons of rank,' 'Rural exercises generally practiced,' 'Pastimes usually exercised in towns and cities, or places adjoining to them,' 'Domestic amusements of various kinds; and pastimes appropriated to particular seasons.'

In contrast to premodern entertainments, the break with hierarchy was a general feature of the modern culture of entertainment. When Philip Astley founded the first circus at Westminster Bridge in 1770 (Speaight, 1980) – gathering itinerant showmen, jugglers, clowns, and acrobats under the same roof, with riding masters as a tribute to aristocratic culture – he addressed his public 'the Nobility, Gentry and others.' But he, as well as other entrepreneurs within show business, soon learned that it would pay off to put the 'others' in the foreground. If the new entertainment industry used amusements of the old aristocratic culture, this would inevitably democratize the diversions of the past. Søren Kierkegaard made 'leveling' a key concept in his analysis of modern, urban culture just as Georg Simmel wrote of 'leveling' in his analysis of *The Metropolis and Mental Life* (1903).

The Mass Entertainment Industry of the Nineteenth Century

According to Asa Briggs, five conditions stand out as necessary for the development of a mass entertainment industry in the nineteenth century: (1) a large and concentrated urban population; (2) sufficient money to buy regular, cheap entertainment; (3) sufficient leisure time to make a commercial exploitation possible; (4) urban transport systems to take the public to the places and palaces of entertainment; and (5) the application of technology to entertainment.

Jacques Offenbach's *La vie Parisienne* may be said to illustrate these five preconditions for the rise of a mass entertainment industry: as the title indicates, the setting is Paris, and the action starts on one of the new railway stations of Paris. The cast of characters either has plenty of money or knows how to get money. The storyline revolves around a new social group, the tourists, and they of course have sufficient leisure time. To re-create the spectacular glitter of modern Paris on stage, the theater would have to apply the newest technologies of stage lighting, etc.

In the nineteenth century, the concentration of the urban population was preceded by an uprooting of people – from the countryside, from tradition, etc. – and the isolation of the individual in the middle of the urban crowd created a need for entertainment to provide psychological support. Just as the popular culture of fairy tales, carnival, etc. had given form and meaning to the culture of preindustrial society, the mass entertainment industry created a sense of community among urban dwellers who could no longer rely on time-honored traditions or preestablished social relations. Boredom may be termed the cultural disease of the nineteenth century. Similar to the boredom experienced at the courts of the seventeenth and eighteenth centuries, this was the psychological reflection of a loss of cultural meaning, a loss of a common cause. A recurrent theme in nineteenth century literature was the characterization of boredom (ennui, Langeweile, …) as a modern invention, a modern monster, the constitutional evil of the century (see Sagnes, 1969). In 1843, the same year that the cosmopolitan Georg Carstensen opened the Tivoli Gardens in Copenhagen, Søren Kierkegaard published his *Either-Or* (1987) with an ironical chapter on boredom as the source of all creation, including the Creation, and as a remedy to this boredom, he suggested public entertainments on a grand scale.

The grandest scale was represented by the city itself, and the crowd and urban life became a leitmotif in all kinds of modern entertainment: panoramas such as the London Colosseum showing a panorama of London as seen from the cupola of

Saint Paul's Cathedral, crime and detective stories, melodramas, and the 'sounds of the city' in the music halls.

Many of the new establishments of entertainment were even constructed as small cities, for instance, panoramas with their spectacular 360° representations of cityscapes, world exhibitions, as well as amusement parks and zoological gardens with gates, walls, streets, squares, and animals living in cages shaped like buildings from their native countries.

City life was entertaining in itself, as witnessed by cafés, promenades, show windows, etc., and the appeal to visual pleasures was noticed by most observers. Charles Lamb (1906) wrote of the spectacle as the essence of the city; Théophile Gautier (1858–59) declared that the era of purely visual spectacles had arrived, and in the 1880s, a new and abundant use of the terms 'show' and 'show business' (see Nasaw, 1993) marked the visual turn of entertainment, which has continued with film, TV, and even music videos into the twentieth and twenty-first centuries.

Variety

Distraction is an important aspect of entertainment, and diversity or variation has since time immemorial been a key to entertaining performances. In *A Short History of Western Performance Space* (2003), David Wiles traces the pleasures of variation back to the symposia of Ancient Greece. Thus, Xenophon describes a symposium hosted by a rich Athenian, where the audience was entertained by instrumental music, juggling hoops, acrobatics, a boy's athletic dance, a boy singing, and a mime play. In contrast to the art of the theater, which was characterized by its unity and its cathartic function, the symposium would offer a variety show in connection with culinary pleasures.

At the founding of music halls in the middle of the nineteenth century, there still was an inner connection between dining, drinking, and variety shows. As indicated by the word 'café,' coffee was part of the entertainment at the café concert and the café chantant, and similarly the posters by Toulouse-Lautrec show the role of sparkling champagne in 'la belle époque' (1890–1914) when the world of entertainment became an inspiration to the avant-garde.

Entertainment also contributed to the so-called postmodern art, and already in 1957, Roland Barthes with his essays in *Mythologies* sketched out some of the features that would inspire postmodernist artists. In his essays on the music hall, he gives a succinct version of the aesthetics of the spectacular. According to Barthes, variety emphasizes the pleasures of the moment, the present moment. The variety show celebrates the body and its gestures, and it cultivates the purely visual beauty of the perfect gesture. All the hard training preceding the 'turns' in a show is not part of the aesthetics. Lofty ideals do not count in this world of the perfect gesture of the acrobat and the imperfect gestures of the clown. And in this urban world of glittering diversions, attention is diverted from boredom, at least for the moment.

Funscapes

In *Sous couleur de jouer* (1989), a philosophical essay on the meaning of play, Jean Henriot argues that behind all the historical and cultural variations in meaning there is a common ground reflected in the use of 'jeu' or 'play' within a mechanism. Here play indicates a space with room for motion, especially rapid, brisk, and light movement, a more or less open interplay between forces.

As an illustration, Henriot quotes Friedrich Nietzsche, who in *Thus spake Zarathustra* talks about play as a "self-rolling wheel, a first movement" (Henriot, 1989: p. 91). In fact, the wheel, the round shape, and the curvilinear are ubiquitous in the world of entertainment. The ring of a circus, the curves of a balloon, the turns of a merry-go-round, the loops of a roller coaster, a Ferris wheel, and a panorama all celebrate relaxation and pleasure, in contrast to straight lines as symbols of strict rationality.

In 1785, William Cowper expressed his concern that balloons would cause 'the annihilation of all authority,' but in 1870, when Fulgence Marion wrote his book on *Wonderful Balloon Ascents*, all worries were overcome. He noted the 'familiar appearance' of the balloon, adding that this form afforded an appropriate symbol of the balloon's own characteristics: "… a harmless, placid, even attractive object, floating serenely and unconcernedly in the sky, subject always to the whims of changing currents or the mercy of strong winds, and carried here or there or (in a calm) nowhere …" (Freedgood, 1997: p. 62).

In his book *The Power of the Center*, Rudolf Arnheim observes that "roundness is the suitable shape for objects that belong nowhere and everywhere" (Arnheim, 1982: p. 115), and he calls the spherical and circular objects 'privileged foreigners in our midst.' In *Amusing the Million* (1978), John Kasson describes how the 'barrels of fun' that would take the immigrants to Coney Island, turning them upside down, and forcing them into close physical contact, would effect a very literal process of cultural integration.

The S- and U shapes of the arabesque also became a symbol of modern show business. The German writer Friedrich Schlegel declared the arabesque to be the oldest form of human imagination, and he claimed that "the suppleness of its line and its caprice are able to embrace the twists and turns, the spontaneity and chance encounters, of life itself" (quoted in Gordon, Ornament 33). The arabesque was of course closely connected to the Orientalism that during the nineteenth century became a way of framing entertainment – in theaters, music halls, etc. The arabesque combined escapism and experiment. The world of entertainment offered its visitors an escape from the routines of everyday life, but at the same time, it challenged them to deal with other cultures, other ways of life.

Owen Jones, who was one of the designers of the Great Exhibition in 1851, also published a book on *The Grammar of Ornament* (1868). Here he suggests that the arabesque may be interpreted as an invitation to stop and linger. In the midst of an ever more stressful urban life, these patterns would seduce the visitor or customer into enjoying an ambience of luxury and balance: "We have lines running horizontally, perpendicularly and diagonally again contrasted by circles in opposite directions. So that the most perfect repose is obtained, the tendency of the eye to run in any direction is immediately corrected by lines giving an opposite tendency, and wherever the eye strikes upon the patterns it is inclined to dwell."

Entertainment Cities

The nineteenth century was intoxicated with expositions, the French writer Gustave Flaubert observed, and nothing was more intoxicating than the world's fairs. During the summer and autumn of 1851, the Great Exhibition in London, especially the gigantic Crystal Palace, attracted more than 6 million visitors. According to Gustave Flaubert, the nineteenth century was delirious with exhibitions. Hans Christian Andersen, who visited the world's fair in Paris in 1867, wrote a 'modern fairytale' on this exposition, in which he introduced the concept of 'knick-knacks of greatness,' which very precisely indicates the principle governing world's fairs: scale reduction. If the world is put into a nutshell, it becomes perceptible and manageable; if the big world is the cause of fear, the little world is a source of fun. World's fairs became major institutions of entertainment, and at the same time they acquainted their huge audiences with new technologies of entertainment such as the Ferris wheel.

In nineteenth century Europe, around the middle of the century, the ineluctable process of urbanization made it necessary to tear down the walls and gates around the cities, but feelings of insecurity during the expansion were countered by the construction of amusement parks, zoological gardens, department stores, etc., as 'cities within the city.' Tivoli Gardens in Copenhagen was established on the military fortifications 15 years before the decision to demolish the ramparts, and in many respects, this amusement park may be interpreted as a harbinger of modernity. At the same time, however, the cultivation of the small scale, the rails around the park, and the impressive gates leading into the park, all pointed backward, nostalgically, to the era of walled cities.

Tivoli Gardens was an important inspiration to Walt Disney, whose Disneyland opened in 1955. In contrast to the diversity of the Tivoli Gardens, however, Disneyland was much more structured. It was the first theme park, based on the characters from Disney's cartoons and movies, but divided into five subthemes: Main Street, Adventureland, Fantasyland, Tomorrowland, and Frontierland. As banners make explicit, the motto of 'Disneyland' was and is 'Safe, Clean and Fun,' and the ideals of this entertainment utopia suggest its background in cultural currents of the 1950s. It capitalized on the ongoing 'White Flight' into the suburbs, on the fear of the impending atomic bomb, and it celebrated the nuclear family and 'the American Way of Life.' Disneyland was designed to wall off people from the outside world (Findlay, 1992: p. 54), and its creators, having learned from movie and TV designers, succeeded in creating what has been called "less an amusement park than a state of mind" (Findlay, 1992: p. 66). Critical commentators, however, have pointed to the interplay between entertainment and containment, between free fantasy and cold control in this utopia.

Disneyland was created at the same time as the Austrian-born architect Victor Gruen created the first shopping malls, and inside these closed architectural structures, shopping developed into a new kind of entertainment. In the words of Michael Sorkin, the shopping malls became 'variations on a theme park.' In the 1980s, Jon Jerde took over as the leading shopping mall environment architect, and he not only integrated various kinds of entertainment into shopping but also applied the playful and seductive architecture of amusement parks to his malls: curved balconies, crossing staircases, flashy colors, and reflecting surfaces.

Cities such as Los Angeles/Hollywood and Las Vegas base their economies on entertainment, but even here a kind of second-degree reference to the world of entertainment serves as an instrument of moneymaking. In 1993, Jon Jerde opened a miniature city, Universal City Walk, in Hollywood, based on film sets and well-known movie locations such as Melrose Boulevard, Sunset Strip, and Venice Beach. At the same time, in real cities, show business reshaped time-honored urban spaces. Disney put his stamp on Times Square in New York and in postwall Berlin Sony gave Potsdamer Platz its new identity.

Globalization of Entertainment

Entertainment has played a major role in the globalization of culture. Around 1900, Coney Island helped immigrants to adapt themselves to American culture, and at the same time American vaudeville – as Albert McLean puts it – became a ritual that allowed immigrants to articulate or interpret their experiences in the new country. In his book *Movie Made America* (1975), Robert Sklar follows this line of thought, but not only did Hollywood 'make' America. It also became a worldwide business. On one hand, Hollywood offered entertainment on a global scale and on the other hand, Hollywood appropriated entertainment from the rest of the world. An obvious example is the genre of Kung Fu movies. In 1974, Carl Douglas made a hit with his song 'Everybody was Kung Fu fighting,' and indeed since the 1970s the entertainment industry of Hong Kong has exported directors, actors, stories, and styles to Hollywood while Hollywood, at the same time, has influenced film making in Hong Kong and the rest of the world. Nicknaming the film industry in India 'Bollywood' reveals this close encounter of the first degree between local and global entertainment cultures.

See also: Aesthetic Education; Audiences, Media; Celebrity; Film History; Film and Video Industry; Film: Genres and Genre Theory; Globalization and World Culture; Leisure and Cultural Consumption: The European Perspective; Mass Media, Representations In; Public Sphere and the Media; Television: History.

Bibliography

Arnheim, R., 1982. The Power of the Center: A Study of Composition in the Visual Arts. University of California Press, Berkeley.
Bakhtin, M., 1968. Rabelais and His World. MIT Press, Cambridge, MA.
Baldau, A., 2008. Entertainment Cities: Stadtentwicklung und Unterhaltungskultur. Springer, Wien, New York.
Barthes, R., 1957. Mythologies. Paris.
Brantlinger, P., 1983. Bread and Circuses: Theories of Mass Culture as Social Decay. Cornell University Press, Ithaca, NY.
Briggs, A., 1960. Mass Entertainment: The Origins of a Modern Industry. Griffin Press, Adelaide, Australia.
Burke, P., 1978. Popular Culture in Early Modern Europe. Temple Smith, London.
Caillois, R., 1967. Les jeux et les hommes. Gallimard, Paris.

Corbin, A., et al., 1995. L'Avènement des loisirs (1850–1960). Flammarion, Paris.
Findlay, J.M., 1992. Magic Lands: Western Cityscapes and American Culture After 1940. University of California Press, Berkeley, Los Angeles, London.
Freedgood, E., 1997. Groundless optimism: regression in the service of egos, England and empire in victorian ballooning memoirs. Nineteenth Century Contexts. An Interdisciplinary Journal 20 (1), 61–80.
Gautier, T., 1858–59. Histoire de l'art dramatique en France depuis vingt-cinq ans, Édition Hetzel, vol. II. Bruxelles, Paris, p. 174.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Hannigan, J., 1998. Fantasy City: Pleasure and Profit in the Postmodern Metropolis. Routledge, London.
Harington, J., 1804. A treatise on playe. In: Harington, H. (Ed.), Nugae Antiquae, vol. 1. Vernor and Hood, Poultry and Cuthell and Martin, Middle Row, Holborn, London.
Harris, N., 1973. Humbug: The Art of P.T. Barnum. Little, Brown, Boston.
Heers, J., 1971. Fêtes, jeux et joutes dans les societies d'Occident à la fin du Moyen Age. Librairie J. Vrin, Paris.
Henriot, J., 1989. Sous couleur de jouer: La métaphore ludique. José Corti, Paris.
Huizinga, J., 1970. Homo Ludens: A Study of the Play Element in Culture. Temple Smith, London.
Jones, O., 1868. The Grammar of Ornament, Illustrated by Examples from Various Styles of Ornament. Bernard Quaritch, London.
Kasson, J.F., 1978. Amusing the Million. Coney Island at the Turn of the Century. Hill & Wang, New York.
Kierkegaard, S., 1987. Either/Or.
Kracauer, J., 1976. Jacques Offenbach und das Paris seiner Zeit. Suhrkamp, Frankfurt am Main, Germany.
Lamb, C., 1906. The Essays of Elia. The Macmillan Company, New York, p. 162.
Le Roy Ladurie, E., 1979. Le Carnaval de Romans: De la Chandeleur au Mercredi des Cendres 1579–1580. Gallimard, Paris.
McLean Jr., A.F., 1965. American Vaudeville as Ritual. University of Kentucky Press, Lexington, KY.
Mehl, J., 1990. Les Jeux au royaume de France du XIIIe au début du XVIe siècle. Fayard, Paris.
Metz, T., 2002. Fun! Leisure and Landscape. NAI Publishers, Rotterdam, NY.
Nasaw, D., 1993. Going Out. The Rise and Fall of Public Amusements. Basic Books, New York.
Plato, 1921. The Laws of Plato. Manchester University Press, Manchester.
Rearich, C., 1985. Pleasures of the Belle Epoque. Entertainment and Festivity-In-Turn of the Century France. Yale University Press, New Haven CT.
Roost, F., 2000. Die Disneyfizierung der Städte. Grossprojekte der Entertainmentindustrie am Beispiel des New Yorker Times Square und der Siedlung Celebration in Florida. Leske + Budrich, Opladen.
Sagnes, G., 1969. L'ennui dans la literature francaise de Flaubert à Laforgue (1848–1884). A Colin, Paris.
Schiller, F., 1967. On the Aesthetic Education of Men (E.M. Wilkinson, L.A. Willoughby, Trans.). Clarendon Press, Oxford.
Scott, D.B., 2008. Sounds of the Metropolis: The 19th Century Popular Music Revolution (In London, New York, Paris and Vienna). Oxford University Press, New York.
Simmel, G., 1950. The Sociology of Georg Simmel. Free Press, New York.
Sklar, R., 1975. Movie-Made America: A Social History of American Movies. Random House, New York.
Sorkin, M., 1992. Variations on a Theme-Park, the New American City and the End of Public Space. New York.
Speaight, G., 1980. A History of the Circus. Tantivy, London, UK.
Veyne, P., 1976. Le pain et le cirque: Sociologie historique d'un pluralisme politique. Seuil, Paris.
Watt, I., 1957. The Rise of the Novel: Studies in Defoe, Richardson and Fielding. University of California Press, Berkeley, CA.
Wiles, D., 2003. A Short History of Western Performance Space. Cambridge University Press, Cambridge.
Zerlang, M., 1997. Orientalism and modernity: tivoli in Copenhagen. Nineteenth-Century Contexts. An Interdisciplinary Journal 20 (1), 81–110.

Entrepreneurship and Economic Growth

Alexander S Kritikos, German Institute for Economic Research (DIW Berlin), Berlin, Germany; University of Potsdam, Potsdam, Germany; IZA, Bonn, Germany; and IAB, Nuremberg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Entrepreneurs, creators of new firms, benefit economies in several ways through innovative activities, new job creation, increased productivity and competition, or because they accelerate structural change. Without new entrepreneurs and the businesses they create, economies may stagnate. Therefore, new business formation by entrepreneurs is regarded as one source of innovation and economic growth.

Introduction

The activities of entrepreneurs are considered to be one important driver of economic dynamics. Entrepreneurs generate employment opportunities not only for themselves but also for others. Beyond the job creation, entrepreneurial activities may influence a country's overall economic performance in several ways: (1) entrepreneurs enter markets with new products, technologies, or production processes; (2) they increase productivity and competition; and (3) they accelerate structural change. Without new entrepreneurs, economies may stagnate.

Despite these apparently important roles in the economy, mainstream economic research generally ignored entrepreneurship for half a century (Barreto, 1989). Only in the 1990s it started that entrepreneurship was recognized as a critical part of successful economies. Researchers in economics, psychology, sociology, and management gradually became interested in entrepreneurship, which is mirrored not only in a growing number of field journals on entrepreneurship, but also in statements such as of Ed Lazear who claimed that "the entrepreneur is the single most important player in a modern economy" (Lazear, 2005: p. 649). This recognition did not happen by chance; it was triggered by substantial changes in the real economy where, in particular, the developments in the information and communication technology sector lead to an expansion of entrepreneurship in developed economies (Audretsch and Thurik, 2000).

At the same time, the media became enthusiastic about entrepreneurs, as well. Frequently incredible success stories of entrepreneurial heroes are featured on television and in magazines. The protagonists change, but the story line remains the same. Obsessed by their business idea, entrepreneurs make risky decisions every day and bear the consequences of these decisions. They are able to swallow any kind of setbacks and, if necessary, they are willing to mortgage their homes, and at the end of the story they become billionaires. The question arising from the scientific point of view is whether these icons serve as role models of the 'perfect entrepreneur' and, if so, whether such role models are helpful, as they may make the majority of potential entrepreneurs feel inadequate in comparison.

In this vein, only a minority of individuals become entrepreneurs: in highly developed economies like the United States, Sweden, and Germany, only one or two out of every 100 individuals in the working population will start entrepreneurial activities in a given year. Thus, the entrepreneur is rare and while some argue that they are important for economic growth and performance, others warn that we run the risk of overstating the impact of entrepreneurship and that we should not try just growing the number of entrepreneurs *ad infinitum* (Shane, 2009).

Given the increasing discussions about entrepreneurship, this contribution will start by defining who are entrepreneurs and what they do. Then the typical determinants of entrepreneurs are revealed, focusing on the socioeconomic, human capital, and personality characteristics that influence the decision to become an entrepreneur. This endeavor will also allow answering the question to what extent the above presented role model and personality characteristics are found in research findings. Based on these definitions and concepts, it is then shown how entrepreneurial activities may affect an economy.

Concepts and Definitions: Who Are Entrepreneurs and What Do They Do?

When describing what entrepreneurs do, we may come up with the following description: Individuals entering entrepreneurship need to identify and exploit opportunities. For the exploitation they typically produce a technology, a good, or service that may involve some form of innovation. More specifically, at every step of the production, marketing, and selling process of their specific product, the decisions they make will determine whether they start their business, and, assuming they do, whether it will succeed and create the maximum value possible. These decisions range from the basic quality and quantity of production, determining investments in the business, deciding upon a marketing strategy, understanding the competition, and, ultimately, delivery of their product to the client. Thus, entrepreneurs need to do many things, or, as Lazear (2005) notes, they need to be "jacks-of-all-trades." In this process, entrepreneurs make a large number of decisions, sometimes quickly, under risk or even uncertainty. Then, after decisions are made, they need to be sufficiently emotionally stable to tolerate the time span until the goods and services they have produced are actually sold. Hence, entrepreneurs need not just knowledge, expertise, and professional

competencies, but also a variety of skills and abilities that are influenced by their personality.

Over the last 300 years various aspects of entrepreneurial behavior have been captured by various scientists. The first author referring to the entrepreneur as an important actor in the economy is Cantillon (1755), who explains why a French word came to define the behavior. In his theoretical work, Cantillon identifies three types of economic agents who differ by the functions they perform: financially independent landowners, risk-taking entrepreneurs, and risk-averse hirelings who prefer stable income over active decision-making. In this setting, entrepreneurs engage in market exchanges to maximize their profit at their own risks.

Based on Cantillon's approach, a century later Say (1803) enlarged the definition of Cantillon and focuses more on the managerial capabilities, as he argues that an entrepreneur is an economic agent who unites all means of production (land, labor, and the capital) to make a product. By selling the product in the market he is able to pay for the production factors – rents, wages, and interest – and keeps the remains as his profit.

A more intensive discussion about definitions of entrepreneurship started in the twentieth century. Basically, there are three classes of definitions, also clarifying the variety of functions an entrepreneur may take: (1) those who focus on entrepreneurship as an occupational choice; (2) those who aim to capture entrepreneurial behavior; and (3) those who have the consequences of entrepreneurial activities for the economy in mind. Restricting to the most important contributions, the occupational choice definitions refer to the decision whether individuals prefer to work for somebody else or for their own account by becoming self-employed. In this context, a broad definition of entrepreneurs includes those persons who create new ventures, or more specifically who are the founders, owners, and managers of an organization under their own liability – mostly small firms where the decisions are made by the entrepreneurs themselves. Given that this definition covers anybody who starts a business, a further step is to distinct between 'replicative' and 'innovative' entrepreneurship, with 'innovative entrepreneurs' being those who bring new products and processes to the market, and introduce new kinds of services, marketing techniques, or business structures. 'Replicative entrepreneurs' are those who enter existing markets trying to have unique selling propositions in order to distinguish themselves from existing firms. Baumol (1993), one of the leading economists in this field, describes innovative entrepreneurs as the bold and imaginative deviators from established business patterns and practices.

Definitions of the second type originate with Knight (1921) and basically focus on the issue of decision-making. Knight argues that investment decisions of entrepreneurs are uncertain and that entrepreneurs have to be able to bear this kind of uncertainty. With this definition, Knight also established a central distinction between risk (which is calculable and, thus, could be covered by insurance, given that there is an underlying probability distribution for the entry of each possible outcome of an investment) and uncertainty (which is incalculable as it has to be handled with no further information) where the entrepreneur directs his actions to new and unpredictable paths. Given this definition, many scholars speak of the 'Knightian uncertainty.' Beyond issues of risk-taking behavior, another strand of behavioral approaches focuses on the specific talents an entrepreneur needs. Beyond Lazear's idea of the entrepreneur as 'jacks-of-all-trades,' earlier models like the one of Lucas (1978) highlight their different 'talents for managing' as the central explanatory variable in determining the choice of becoming an entrepreneur or, like Holmes and Schmitz (1990), who point to the ability to 'respond to opportunities for creating new products that arise from technological breakthroughs' as the crucial talent needed to become a *successful* entrepreneur.

The most important protagonist of the third class of definitions is Schumpeter, who asserts a central function to entrepreneurs in capitalist economies. According to his approach (Schumpeter, 1934), entrepreneurs are the innovators in an economy who 'carry out new combinations' and thus initiate what Schumpeter calls the process of creative destruction where old established firms and products are replaced with new firms with better or new products. More recent researchers, such as Leibenstein (1968) and Kirzner (1973), give a slightly different role to the entrepreneur in the economy. Here, the entrepreneur is able to identify and exploit opportunities by revealing inefficiencies, either among existing firms and organizations (Leibenstein), or from market imperfections (Kirzner). In their definitions, the entrepreneur is more of an arbitrageur who takes advantage of the observed gap by combining resources in a better way, by organizing the production in a more efficient way, or by producing new products or services (the latter definition being similar to Schumpeter's approach). In particular, Schumpeter's idea of creative destruction was later on formalized in models (as of Aghion and Howitt, 1992) where innovation evolves following a certain stochastic process dependent on the amount of investments in research and development (R&D). These models, however, replace the person of the entrepreneur by the amount of resources devoted to R&D activities.

Thus, there is a wide variety of definitions and concepts of entrepreneurship that are proposed. To condense these three classes of definitions, we may conceptualize entrepreneurs as individuals who are willing and able

1. to start and own an organizational entity with the purpose of creating value;
2. to innovate by identifying and exploiting (new) opportunities in the market; and
3. to bear the risk of introducing (new) products in the market as well as to manage their firm by deciding on the use of resources, production methods, and locations.

In this context, it should also be clarified that part (1) of this definition contains a limit. The establishment of a firm (an organizational entity) implies that other kinds of entrepreneurs like 'intrapreneurs' and public entrepreneurs are not included. These groups deserve separate approaches. Moreover, the purpose of creating value incorporates two groups of entrepreneurs, those who aim to maximize profit by creating value and another group that recently gained substantial attention, namely the so-called social entrepreneurs who are rather interested in value- instead of profit-maximizing.

Who Becomes an Entrepreneur and Who Succeeds?

Much effort has gone toward trying to uncover who will engage in entrepreneurial activity and who will not. Researchers have investigated which demographic and socioeconomic characteristics, as well as which personality characteristics influence the probability of becoming an entrepreneur.

Empirical analysis reveals some common characteristics in many countries when focusing on the influence of sociodemographic variables such as age, gender, and education. Starting with a person's age and the probability of becoming an entrepreneur, this relationship is inversely U-shaped. This means that the propensity of individuals to become an entrepreneur first increases with age, peaks at ages between 35 and 45 years, and then decreases again (Levesque and Minniti, 2006). While the increasing proclivity for entrepreneurship during younger ages can be explained with growing experience and an improved access to resources, the declining part may be induced by decreasing risk tolerance and the fact that individuals who have an openness for starting an own business, mostly decide to do so before they become 'too old.' This also makes clear that, for most individuals, the entry decision into entrepreneurship is not a once and for all decision at the beginning of their working career, it mostly happens after having gained 10 years of working experience (or more) as a dependently employed individual.

Second, females have a lower probability of becoming entrepreneurs than males, which may be explained by gender-specific role models, a difference in the willingness to take risks (Caliendo et al., 2014b), and the willingness to turn entrepreneurial intention into implementation (Haus et al., 2013). Third, the family surrounding influences transitions into entrepreneurship: Having self-employed parents significantly increase the likelihood of becoming an entrepreneur, perhaps because they are role models for occupational choices or because the children inherited the 'right skills and abilities' for entrepreneurial activities. Fourth, another source inducing changes to entrepreneurial activities is the amount of start-up capital available to the potential business founder. Blanchflower and Oswald (1998) show that there is a positive effect on this transition: Individuals with higher capital endowment have a higher probability of becoming an entrepreneur. Fifth, in many countries there is a positive relationship between entrepreneurship and working in the service sector, which reflects the relatively small average size of service-sector firms and lower barriers to entry in many parts of the service sector.

One important difference between countries and entrepreneurship is found with respect to educational levels. There are some countries with a linear positive relationship between having achieved higher formal qualifications, such as tertiary education or vocational training, and the propensity for starting entrepreneurial activities, while for other countries this relationship is U-shaped, meaning that lowly and highly educated individuals have a higher propensity to get involved with entrepreneurship than individuals with medium levels of education (e.g., Borjas and Bronars, 1989).

Beyond the question of how demographic and other exterior characteristics, such as capital endowment, guide the decision to start a business, there is also a discussion on the influence of personality traits on becoming an entrepreneur. This approach has been criticized by researchers for various reasons. One critique is that entrepreneurship is determined by a wide variety of factors and that the diversity in personality traits *among* entrepreneurs is much greater than differences *between* entrepreneurs and nonentrepreneurs such that it is impossible to determine the personality profile of a typical entrepreneur. Moreover, previous research tried to construct an average entrepreneurial profile, but failed to do so, as Aldrich (1999) made clear. He claimed that this kind of research had reached an empirical dead end.

Given these critical observations, recent research stopped comparing the stock of entrepreneurs with the stock of people in another employment status (see, e.g., Rauch and Frese (2007) for such comparison), but started to analyze to what extent personality influences entrepreneurial decisions at certain points in time, like the decision to quit a job and become an entrepreneur. These new approaches reveal how personality is, for instance, related to the selection into entrepreneurship without arguing that these characteristics are also important for other entrepreneurial decisions; thus they explicitly refrain from trying to construct an average entrepreneurial profile.

The most important personality characteristics that are found to influence the entry decision into entrepreneurship are the following: (1) risk attitudes, (2) openness for experience, (3) locus of control, and (4) extraversion. A closer look at these personality characteristics and their potential influence on the decision to start an own business makes clear that they all relate to the definitions of entrepreneurship discussed in the previous section.

Obviously, the outcome of each investment in particular in innovative activities is often unpredictable, meaning that every entrepreneurial decision is risky or even uncertain. Moreover, in contrast to ordinary managers, entrepreneurs often lose their own funds in case of investment failure. Entrepreneurs must, therefore, be willing to bear risks or even uncertainty and have accordingly high risk tolerances.

The second characteristic, openness for experience, describes an individual's ability to seek new experiences and to explore novel ideas. Persons scoring high on openness to experience are typically creative, innovative, and curious. These attributes are particularly crucial for individuals intending to start an innovative new venture, but also for replicative entrepreneurs aiming to create a unique selling proposition in their own firm.

The two further traits relate to another aspect discussed in the previous section, namely the managerial capacities of heading an own company. In this context, *locus of control* measures generalized expectations about internal and external control of reinforcement. People with an internal locus of control believe that they determine their future outcomes through their own actions. Persons with an external locus of control believe that their own future development is determined randomly or by the external environment, but not by their own actions. As entrepreneurs must continuously make decisions related to their business outcomes, a high level of internal locus of control is a relevant personality characteristic for all entrepreneurial decisions. The last trait, *extraversion*, consists of variables that describe the extent that individuals are

assertive, dominant, ambitious, energetic, and seek leadership roles. Moreover, extraverted individuals tend to be sociable, thus enabling them to develop social networks more easily, which may result in stronger partnerships with clients and suppliers. All parts of the factor – being assertive, seeking leadership, and developing networks – are important if an individual aims to become the head of a firm. Recent empirical research finds that individuals with these personality characteristics have a higher probability of becoming entrepreneurs (Caliendo et al., 2014a).

The important point now is that there are some personality constructs that affect entry into entrepreneurship. However, there are different ones – with the exception of locus of control – or different parameter values of the same construct that affect an exit decision from entrepreneurship and thus success or survival in entrepreneurship. More specifically, empirical research reveals that the most important personality characteristics influencing entrepreneurial success are the following four: lower levels of agreeableness, higher levels of need for achievement, higher levels of locus of control (as mentioned before), and rather medium levels of risk attitudes. Individuals scoring high on agreeableness are described as having a forgiving and trusting nature as well as being altruistic and flexible. Therefore, low scores of agreeableness might, in particular, help entrepreneurs survive as such low levels allow them to bargain for their own interest with their partners. Moreover, need for achievement expresses the motivation of entrepreneurs to realize new and better solutions than those given in the actual (market) environment, through their own performance in the market thus making high scores in this personality construct increasing entrepreneurial success (Holmes and Schmitz, 1990). The same holds for locus of control. Last, but not least, the relationship between risk tolerance and the probability of entrepreneurial success is inverse U-shaped and not linear (Caliendo et al., 2010). Entrepreneurs within a medium range of risk tolerance have the lowest exit probabilities. An explanation is that too low risk tolerance leads to low expected returns from low-risk projects and makes entrepreneurship unattractive in comparison to dependent employment, whereas entrepreneurs who are excessively risk tolerant engage in risky projects with higher failure rates. As to demographic characteristics, it can be further observed that the middle-aged and the well educated also have the highest survival probabilities.

In particular, the result on the impact of risk attitudes on entrepreneurial survival shows that stories about entrepreneurs and their willingness to mortgage their home (as mentioned in the introduction) while making high-risk decisions, with the tale ending that the person has become a billionaire, are good for media and may hold for a handful of people, but not for the majority of entrepreneurs, and that this uncontrolled risk-maximization might not be the best role model to follow. Entrepreneurial activities are more often determined to be successful if a reasonable risk assessment has been conducted.

As the decision to become an entrepreneur might also be influenced by public reports, role models are crucial, giving orientation. Bosma et al. (2012) address the occurrence, function, and characteristics of specific role models in entrepreneurship, in both the pre- and post-start-up phases of firms. Results show that role models matter for potential entrepreneurs: Most of the entrepreneurs have a role model in either of the two phases and view them as influential. One-third of the role model users claim that they would not have started their business without the presence of role models. The dominant functions of role models are 'learning by example,' followed by 'learning by support,' 'increasing entrepreneurial self-efficacy,' and 'inspiration or motivation.'

They also show that role models compensate for a lack of entrepreneurial experience. The individuals behind the role models are older and have much larger firms than their followers. However, icons from media reports are rarely actual role models; they rather tend to come from next-door.

Putting the puzzle together, it becomes clear that many different variables influence this kind of lifetime decision to switch from another employment form to entrepreneurship. Some demographic and some personality variables influencing the entry decision into entrepreneurship have been identified, but in particular with respect to personality traits these are different from the ones influencing entrepreneurial success. This also makes clear why we should refrain from trying to describe what the average entrepreneur should look like.

Why Entrepreneurs Are Important for the Economy?

In the last two decades many economies in the developed world have become more entrepreneurial. This transition is fostered by several factors. The costs and benefits of becoming an entrepreneur are affected, among others, by (1) shifts in demand from standardized mass production toward more individual products; (2) increasing importance of knowledge as an input factor for production; (3) new developments in information and communication technologies that facilitate the coordination of different labor divisions, thus allowing for more flexible small-scale production units; and finally (4) increasing globalization and opening new markets with new investment opportunities for a more productive division of innovative labor (Audretsch and Thurik, 2000). Those factors have led to a reduction of the minimum efficient size of production, to lower entry barriers and to new entrepreneurial opportunities.

What kind of effects can economies experience when entrepreneurs do enter markets?

First, and foremost, innovative entrepreneurs may create product or process innovations and open new markets. There are many examples of radical innovations introduced by entrepreneurs who founded firms like eBay, Google, Amazon, Oracle, SAP, Microsoft, Apple, and Skype. The companies emerging out of these innovations often affected economic growth (Valliere and Peterson, 2009). Innovative entrepreneurs thus offer a key value-generating contribution to economic progress. Their pronounced role with their new firms may be driven by the fact that compared with incumbent suppliers, new firms may invest more in searching for new opportunities. Incumbent suppliers are more reluctant to do so, either because of their failure to respond to changes in their environment due to organizational inertia, or because these newly invented goods compete with their established range of products. Therefore, established firms often miss the opportunity to adopt new ideas, and setting up one's own business appears to

be the only way for inventors to commercialize their knowledge (Klepper, 2009).

Job creation by entrepreneurs is another crucial issue. Entrepreneurs and their new firms can stimulate employment growth by generating new jobs. Such evidence is found for OECD economies (Carree and Thurik, 2008). However, when analyzing the employment effect of new firms over time, it is interesting to note that the effect for the economy must not be positive at all times. There is an S-shaped effect (see Fritsch, 2008), with entrepreneurship positive in the start-up phase and in the long run, while in the medium turn a decrease in employment can occur when new businesses fail to compete or crowd out incumbent firms.

Entrepreneurs entering the market with their new firms have several further effects to the economy. New firms increase the competition in the market by their pure number and thus decrease the market power of incumbents. They may also intensify competition between them and existing businesses. As this may push established firms to improve their performance, such kind of increased competition is particularly utility-increasing for consumers who realize lower prices and a greater product variety.

Competition between entrepreneurs and incumbent firms ideally leads to the 'survival of the fittest.' Accordingly, new firms should also foster productivity (see Geroski, 1989). This productivity-enhancing effect of new business formation should become evident shortly after entry because only entrepreneurs who perceive a gap in the market or are more efficient than existent market player will enter the market. Entry and the resulting selection process induce movements of the production frontier with less-efficient firms (both entrants and incumbents) dropping out. Empirical studies show that there is a positive relationship between the level of early-stage entrepreneurial activity and productivity. This relationship is particularly strong for entrepreneurs with high-growth and innovative ambitions while entrepreneurs with low-growth ambitions have a weaker but still positive impact on productivity. This pattern indicates that entrepreneurs generally increase the productive use of scarce resources in an economy with the strongest impact coming from innovative entrepreneurs.

Entrepreneurs may also be able to foster structural change: Incumbent firms often struggle to adjust to new market conditions and permanent changes. Thus failing to make necessary internal changes because they lack the ability for 'creative destruction' (Schumpeter, 1934). The entry of new businesses accompanied by an exit of some incumbents can help to free firms from this 'lock-in' situation. Moreover, they can "initiate the emergence of new markets and industries that may be drivers of growth processes" (Klepper, 2009).

Entrepreneurial Failure and Closure

It is an innate problem of entrepreneurship that during the first 5 years after starting their businesses, entrepreneurs have a substantial risk of failure or closure of their new firms. There are several reasons for failure and closure. First, not every individual is able to become a successful entrepreneur and may not have known prior to starting the business. Further reasons for failure include (1) restricted access to capital; (2) lack of customers; (3) poor managerial skills; and (4) the costs of regulatory burdens. These burdens can, for example, come in the form of lengthy registration procedures and complicated permit or license regulations. Typical failure rates are around 40–50% over the first 5 years of a firm, with the highest failure rate in the first year.

Summary

Entrepreneurs, while rarely existing in highly developed economies, serve several crucial functions in an economy. They create employment opportunities for themselves and others. Entrepreneurial activities may further influence the overall economic performance in a country because entrepreneurs do not just innovate but also influence productivity, competition, and structural change. Therefore, new business formation by entrepreneurs is regarded as a critical source of economic and employment growth.

Research in entrepreneurship is appealing for various reasons: First of all it is about the understanding of a rare species. Second, it combines research from all kinds of areas including economics, psychology, sociology, and management. Research bridging these disciplines improves understanding entrepreneurship and its effects. Third, entrepreneurial activities are particularly dynamic, which is why it is challenging to model these activities. Closing with the words of Barreto (1989: p. 141): "Simply put, entrepreneurship is above 'formalization' – it cannot be neatly packaged within a mechanistic, deterministic model." Future research has, on the one hand, the challenge of developing nonmechanistic models that bridge this gap. On the other hand, we are still far from fully explaining why some entrepreneurs are more successful than others. Finding better answers to the first question might open up new avenues to analyze the second.

See also: Education and Economic Growth; Innovation and Employment; Innovation and Technological Change, Economics of.

Bibliography

Aghion, P., Howitt, P., 1992. A model of growth through creative destruction. Econometrica 60, 323–351.

Aldrich, H.E., 1999. Organizations Evolving. Sage, London, UK.

Audretsch, D.B., Thurik, R., 2000. Capitalism and democracy in the 21st century: from the managed to the entrepreneurial economy. Journal of Evolutionary Economics 10, 17–34.

Barreto, H., 1989. The Entrepreneur in Micro-Economic Theory: Disappearance and Explanation. Routledge, London.

Baumol, W.J., 1993. Entrepreneurship, Management, and the Structure of Payoffs. MIT Press, Cambridge, MA.

Blanchflower, D., Oswald, A., 1998. What makes an entrepreneur? Journal of Labor Economics 16, 26–60.

Borjas, G.J., Bronars, S.G., 1989. Consumer discrimination and self-employment. Journal of Political Economy 97, 581–605.

Bosma, N., Hessels, J., Schutjens, V., Van Praag, M., Verheul, I., 2012. Entrepreneurship and role models. Journal of Economic Psychology 33, 410–424.

Caliendo, M., Fossen, F.M., Kritikos, A.S., 2010. The impact of risk attitudes on entrepreneurial survival. Journal of Economic Behavior and Organization 76, 45–63.

Caliendo, M., Fossen, F.M., Kritikos, A.S., 2014a. Personality characteristics and the decisions to become and stay self-employed. Small Business Economics 42, 787–814.
Caliendo, M., Fossen, F.M., Kritikos, A.S., Wetter, M., 2014b. The gender gap in entrepreneurship: not just a matter of personality. CESifo Economic Studies, published online on June 19; http://dx.doi.org/10.1093/cesifo/ifu023.
Cantillon, R., 1755. Essai sur la nature du commerce en général, ed. and with an English translation by Henry Higgs for the Royal Economic Society. Macmillan & Co, London.
Carree, M., Thurik, R., 2008. The lag structure of the impact of business ownership on economic performance in OECD countries. Small Business Economics 30, 101–110.
Fritsch, M., 2008. How does new business development affect regional development? Introduction to the special issue. Small Business Economics 30, 1–14.
Geroski, P.A., 1989. Entry, innovation, and productivity growth. Review of Economics and Statistics 71, 572–578.
Haus, I., Steinmetz, H., Isidor, R., Kabst, R., 2013. Gender effects on entrepreneurial intention: a meta-analytical structural equation model. International Journal of Gender and Entrepreneurship 5, 130–156.
Holmes, T., Schmitz, J.A., 1990. A theory of entrepreneurship and its application to the study of business transfers. Journal of Political Economy 98, 265–294.
Kirzner, I.M., 1973. Competition and Entrepreneurship. University of Chicago Press, Chicago.
Klepper, S., 2009. Spinoffs: a review and synthesis. European Management Review 6, 159–171.
Knight, F.H., 1921. Risk, Uncertainty and Profit. Houghton Mifflin, New York.
Lazear, E.M., 2005. Entrepreneurship. Journal of Labor Economics 23, 649–680.
Leibenstein, H., 1968. Entrepreneurship and development. American Economic Review 58, 72–83.
Levesque, M., Minniti, M., 2006. The effect of aging on entrepreneurial behavior. Journal of Business Venturing 21, 177–194.
Lucas, R.E., 1978. On the size distribution of business firms. Bell Journal of Economics 9, 508–523.
Rauch, A., Frese, M., 2007. Let's put the person back into entrepreneurship research: a meta-analysis on the relationship between business owners' personality traits, business creation, and success. European Journal of Work and Organizational Psychology 16, 353–385.
Say, J.B., 1803. A Treatise on Political Economy; or the Production Distribution and Consumption of Wealth. Translated from the 4th edition, Batoche Books Kitchener, 2001.
Schumpeter, J.A., 1934. The Theory of Economic Development. Harvard University Press, Cambridge.
Shane, S., 2009. Why encouraging more people to become entrepreneurs is bad public policy. Small Business Economics 33, 141–149.
Valliere, D., Peterson, R., 2009. Entrepreneurship and economic growth: evidence from emerging and developed countries. Entrepreneurship & Regional Development 21, 459–480.

Entrepreneurship and Organizations

Peter T Bryant, IE Business School-IE University, Madrid, Spain

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Martinelli, volume 7, pp. 4545–4552, © 2001, Elsevier Ltd.

Abstract

Entrepreneurship can be broadly defined as the identification and organized exploitation of novel opportunities for value creation and capture, where the value outcomes may be commercial, social, institutional, or cultural in nature. Hence, entrepreneurship has wide relevance for societies and economies. As a field of practice, entrepreneurship has grown continually since its emergence during the industrial revolution of the eighteenth and nineteenth centuries. More recently, entrepreneurship has taken on a wider meaning, as the process or activity whereby new organizational forms are created to exploit value-creating opportunities, even when the resources and capabilities to do so are not yet at hand.

Entrepreneurship can be broadly defined as the identification and organized exploitation of novel opportunities for value creation and capture, where the intended value outcomes may be commercial, social, institutional, or cultural in nature. Hence, entrepreneurship has wide relevance for societies and economies. Indeed, in the early twenty-first century, entrepreneurship is widely viewed as a major source of economic growth, job creation, and innovation. It has become a major topic of interest for governments, companies, educational institutions, and enterprising individuals. In parallel, scholarship on the subject has blossomed as well.

The modern concept of entrepreneurship has historic roots in European history (Hébert and Link, 2009). The origin of the word 'entrepreneur' can be traced to a time of warfare, rather than of economic change, in sixteenth-century France. During this period, entrepreneurs were defined as profit-seeking project workers, master builders, or manufacturers. Often the term referred to government contractors or the builders of military fortifications. Only later, in the eighteenth century, was the concept of entrepreneurship applied to economic actors as we understand them today. From this perspective, an entrepreneur was someone who either undertook contracts for public works, or introduced new agricultural techniques in their land, or risked their own capital in industry. More recently, the concept of entrepreneurship has taken on a wider meaning again, as the process or activity whereby new organizations are created to exploit value-creating opportunities, even when the resources and capabilities to do so are not yet at hand and the risk of failure is high.

Early Scholarship

Early scholarship on the topic of entrepreneurship reflected cultural and intellectual trends in eighteenth- and nineteenth-century Europe. During this period, Enlightenment thought and political revolution advocated cultural change, social mobility, and economic progress, providing entrepreneurial activity with both meaning and vitality. At the same time, early waves of industrialization and associated wealth redistribution were opening up new opportunities for entrepreneurship in the commercial sphere. Not surprisingly, therefore, French Enlightenment scholars such as Cantillon and Turgot advocated entrepreneurial activity. They stressed the need for individuals to accept the risk and uncertainty inherent in economic activity as its key distinctive feature.

During the same period, economic thought also began to incorporate conceptions of entrepreneurship (Hébert and Link, 2009). For example, Jean Baptiste Say defined the entrepreneur as the person who combines the factors of production for profit, putting him or herself at the center of several relations among capitalists, workers, technicians, merchants, and consumers; while the economist and philosopher John Stuart Mill discussed the entrepreneurial function in relation to the provision of risk capital. Later scholars built on these foundational thinkers, and their insights continue to infuse understanding about entrepreneurship.

However, it was not until the late nineteenth and early twentieth centuries that the sociologist Max Weber, and then the economist Joseph Schumpeter, provided deeper theoretical analyses of entrepreneurship as a force within society and the economy. Weber analyzed entrepreneurship in terms of the interplay between the religious mentality of Protestantism and modern rationalism in the rise of capitalism (Weber, 2007). In this context, the entrepreneur is distinguished from historical predecessors in traditional economies by virtue of a rational and systematic pursuit of economic gain and the subordination of consumption in the interest of capital accumulation. For Weber, entrepreneurship was an expression of this modern mentality.

A few decades later, Schumpeter provided one of the first and most thorough analyses of the entrepreneurial function from an economic viewpoint. In his early work, he conceptualized entrepreneurship as a driver of innovation within cycles of industrial change, thus giving it a dynamic quality lacking in earlier formulations (Schumpeter, 1934). Through entrepreneurial action and innovation, he argued, the novel combination of the factors of production, when combined with credit, disrupts the static equilibrium of economic life. Entrepreneurial action therefore changes the conditions of supply, combines existing resources in new ways, and creates new production functions. Waves of creative destruction are then released. For this reason, Schumpeter stressed the revolutionary character of entrepreneurial action.

During the following years of the twentieth century, the pace of global economic change quickened. In fact, on many dimensions – social, political, technological as well as economic – the pace of change during this period was as great as any time in human history. Rapid change continues in the early decades of the twenty-first century, breeding risk and opportunity. Reflecting these trends, entrepreneurship has become more prevalent and encouraged. Indeed, entrepreneurship is now widely promoted as an important factor in achieving economic growth and social development.

The Field of Entrepreneurship

Reflecting the social and economic trends of the late twentieth and early twenty-first centuries, entrepreneurship has emerged as a significant area of scholarly research, management practice, and government policy. Indeed, like many areas of scholarship about organizations and management, research into entrepreneurship has mushroomed during this period, resulting in a variety of perspectives. Chief among them are process theories, sociological, psychological and economic perspectives, and practice-oriented approaches. Following sections discuss each in turn.

It is important to note, however, that the status of entrepreneurship as an autonomous field of academic study remains controversial. On the one hand, some scholars argue that entrepreneurship as an area of study is correctly defined in terms of the phenomena it encompasses, rather than distinctive theories (e.g., Sorenson and Stuart, 2008). Relevant phenomena of interest include new ventures, investor networks, innovative business models, and the characteristics of the founders of new ventures. Advocates of this view do not regard entrepreneurship as a separate field of study, defined in terms of a distinct body of supporting theory. They argue instead that research into entrepreneurship must borrow from other disciplines and fields of study for explanatory theories and frameworks. Adopting this perspective, the study of entrepreneurship is distinguished by its phenomenal qualities.

On the other hand, an equally rigorous community of scholars does accord entrepreneurship the status of a distinct field of study (e.g., Busenitz et al., 2003; Shane, 2003). They argue that the processes and behaviors of entrepreneurship have special features which call for unique and specialized theories, above and beyond potential borrowings from other disciplines. From this perspective, any attempt fully to explain the phenomenal richness of entrepreneurship, embracing process, people, and profit, demands a form of synthesis that is not found in other fields. At the time of writing, this debate continues. It evidences the enduring complexity and breadth of entrepreneurship as both a phenomenon and topic of research.

Process Theories

The process of new venture creation is often seen as the archetype of entrepreneurial activity. In summary, an entrepreneur senses a market opportunity, creates a new organization, builds a team, and gathers the resources required to exploit the opportunity, perhaps raising venture capital investment, and then seeks to grow the business. This process perspective was clearly articulated by Shane and Venkataraman (2000) in a highly influential paper. They explained that entrepreneurship should be conceived as "the nexus of two phenomena: the presence of lucrative opportunities and the presence of enterprising individuals" (Shane and Venkataraman, 2000: 218). Following their advice, many scholars research the interaction of opportunities and individuals in the entrepreneurial process. Important phases of the process include opportunity recognition or creation, venture founding, team building, and resource accumulation.

Not all scholars agree, however, regarding the nature of the entrepreneurial process and its core phases. One major debate focuses on whether new opportunities exist objectively, and thus wait to be discovered; or whether opportunities come into being through entrepreneurial imagination and action (Alvarez and Barney, 2010). In other words, are opportunities discovered or created? In fact, there is evidence to suggest that opportunities may emerge via both routes. Sometimes, the market or social context exhibits clear needs that await entrepreneurial attention, while in other contexts, it requires original insight, imagination, and action to create and shape opportunities.

Another debate focuses on the role of the individual entrepreneur in the entrepreneurial process. Schumpeter exemplified the debate. In his early work, he emphasized the role of the individual entrepreneur in economic change. In his later work, however, he argued that industrial and economic forces were more important in the explanation of entrepreneurial activity. For the late Schumpeter, markets matter most, and while entrepreneurs play an important role, they are more the mediators than originators of entrepreneurial processes. Other scholars take the opposite view, and argue that founders are critical to the entrepreneurial process (Wasserman, 2012). They maintain that individual entrepreneurs create opportunities, found new ventures, and initiate industrial change. This debate also relates to the sociological and psychological perspectives on entrepreneurship, which are discussed in the next section.

Process theories also imply logics of progression across phases of stages of activity. Once again, approaches vary. Some argue it is possible to conceive of and explain cause and effect in the entrepreneurial process. For example, assuming causal relationships, entrepreneurs choose to pursue opportunities, build plans, and then create new ventures and act deliberately in the exploitation of their chosen opportunity. In contrast, Sarasvathy (2001) argues that causal explanations are not feasible in the uncertain, unpredictable realm of entrepreneurship. Following the general argument espoused by the economist Frank Knight (1942), Sarasvathy proposes that entrepreneurs effect the future, rather than cause it. In her 'effectuation theory,' one starts with only a set of means; in the process of deploying these means, goals gradually emerge. Using an effectual logic, future outcomes are not the result of causation, but rather emerge unpredictably from meaningful action. Indeed, most scholars agree that entrepreneurship involves high risk and uncertainty. The degree to which these conditions can be subject to causal or effectual reasoning may depend largely on the relative degrees of uncertainty in any given opportunity context.

Psychological and Sociological Approaches

Irrespective of the process logic one employs, entrepreneurship is clearly related to human thought and action, at both the individual and collective levels. Founders create firms and build teams, and these teams work together under pressure and uncertainty to pursue shared goals. They use networks and social relationships to gather resources and leverage influence. Hence, social, individual, and group factors play important roles. Many scholars therefore employ psychological and sociological theories and techniques to understand entrepreneurship.

From a psychological perspective, researchers employ numerous theories, including theories of personality, motivation, goal setting, decision making, perception, and self-regulation; and from the sociological perspective, theories of organizational populations, institutional legitimacy, organizational imprinting, and embeddedness. Major research questions in the psychological stream of research include the following: Why do some people choose entrepreneurship while others do not? Are entrepreneurs more alert to opportunities, and if so why? What are the characteristics or traits which drive some individuals to persist in the face of uncertainty and possible failure? and Are entrepreneurs more or less aware and tolerant of risk, when compared to non-entrepreneurs? (Baron, 2004). From a sociological perspective, typical research questions are the following: To what degree are new ventures reflective of their surrounding organizational context? How do new organizations emerge and evolve over time, and gain legitimacy? How do the early culture and routines of an entrepreneurial firm become imprinted and enduring as the organization grows, and with what effects? (Johnson, 2007).

Over two decades of research since the 1990s, many studies have investigated these and related questions, hoping to uncover the factors which help to explain and predict entrepreneurship (Grégoire et al., 2011; Sorenson and Stuart, 2008). However, the results have been mixed. To begin with, from the psychological perspective, personality trait theories are now broadly rejected; that is, few believe that entrepreneurs possess specific and distinctive personality traits. However, cognitive studies continue to be enlightening. There is good evidence that entrepreneurs think differently about risk, decision making, experimenting, planning, and goal setting (Bryant, 2007). They also appear to be more alert to opportunities, based on cognitive skills of pattern recognition and creativity (Baron, 2006). Evidence also suggests that such cognitive skills can be learned and enhanced over time.

Furthermore, in recent years, scholars have started to investigate entrepreneurial cognition using the tools and techniques of cognitive neuroscience. They hope to identify whether and how entrepreneurial thought and action are associated with individuals' neurophysiological characteristics (e.g., Laureiro-Martínez et al., 2010). While this research program is in its infancy, one can expect significant findings in future years, adding to the cognitive and psychological perspectives on the subject.

From a sociological perspective, research has shown that social networks, social capital, and integration into social structures are important for entrepreneurship (Busenitz et al., 2003). For example, social connections often help founders to attract talent and resources to new ventures. Social networks can open doors to opportunities and markets, and can be especially significant as sources of information and strategic intelligence. Indeed, from a social network perspective, entrepreneurship can be conceived as a boundary spanning or brokerage activity, whereby entrepreneurs link people, information, and resources across preexisting social and organizational boundaries (Burt, 2004). Studies also show that membership of social networks enhances status and trust in attracting investment and collaborators. For new ventures, such benefits are significant, given their liabilities of newness.

At another level of social organization, entrepreneurs are institutional actors. They hold governance roles in new organizations, influence the emergence of organizational cultures and behavioral norms, and imprint foundational norms and routines. Indeed, many entrepreneurs actively create new institutions and organizational forms, sometimes through leading the evolution of new social movements in response to unmet social needs or perceived injustice. Hence scholars investigate how institutional factors – such as culture, social norms, behavioral routines, and organizational structures – contribute to our understanding of entrepreneurship (Battilana et al., 2009). Moreover, as new organizational forms and social movements continue to emerge, institutional entrepreneurship may become even more salient in future years.

Economic Perspectives

Another rich tradition of research into entrepreneurship focuses on economic factors. As noted earlier, writing from within the Austrian tradition of economic thought, Schumpeter was the foundational theorist in modern times. He conceived of entrepreneurship as an important and inherent feature of economic change and development. More recently, Israel Kirzner (1998) has developed an influential theory of entrepreneurship, also based in the Austrian economic tradition. He emphasizes the role of entrepreneurial discovery and alertness to opportunities as key elements in the dynamics of nonequilibrium markets. In doing so, Kirzner reinvigorated interest in Schumpeter as well, by connecting the Austrian economic tradition to more contemporary microeconomic debates.

Working in a similar vein, evolutionary economists such as Nelson and Winter (1982) propose a sophisticated evolutionary theory of economic change which draws heavily on Schumpeter's ideas about innovation and industrial dynamics. Contrary to classical economic theories which assume utility maximization in near equilibrium states, Nelson and Winter argue that economies are in a continual state of evolutionary adaptation, driven by 'Schumpeterian competition' and supported by new combinations of routines at the firm level. In a subsequent paper, Winter (2006) continued this line of research by proposing a neo-Schumpeterian theory of the firm in which entrepreneurs play a central role as the leaders of new techniques and resource combinations. From this perspective, entrepreneurial action is central to the evolution of economic systems, the emergence of firms, and broader patterns of innovation. Assuming this perspective, entrepreneurship is best

understood in the context of industrial dynamics and economic change.

As Schumpeter also noted, technological innovation is a particularly important stimulus of economic change and entrepreneurship. Recent examples are compelling: companies such as Microsoft, Google, Apple, Genentech (all founded in the USA), Infosys (in India), Alibaba (in the People's Republic of China), and Samsung (in South Korea). In the early years of the twenty-first century, all are successful global companies that began as entrepreneurial ventures seeking to exploit opportunities based in innovative technologies. Indeed, technological innovation is now viewed as a major driver of entrepreneurship with its own principles of practice (Hsu et al., 2007). The 'Lean Startup' approach to new venture creation is a good example of this trend (Blank, 2013). In contrast to traditional theories of entrepreneurial ventures, the Lean approach advocates continual customer discovery, the rapid prototyping of minimum viable products, readiness to fail, and the necessity for adaptation or pivoting. It is especially applicable in the dynamic context of Internet-based entrepreneurship, but elements of the method are applied more widely as well.

Entrepreneurial Practice

Entrepreneurship is also an increasingly important area of management practice. Even in large established organizations, employee entrepreneurs (sometimes referred to as 'intrapreneurs') help to develop new business opportunities, open new markets, and foster organizational innovation. This significant area of practice has spawned a growing literature on corporate entrepreneurship and spinoff companies (e.g., Shepherd et al., 2009). Indeed, many established organizations now actively seek to encourage entrepreneurial activity by their employees, as a way to grow and adapt in an increasingly uncertain world. David Teece (2007) captures this trend in his theory about the dynamic capabilities which firms need in order to reconfigure their resources and activities in response to opportunities. From this perspective, even large companies must develop entrepreneurial capabilities and characteristics in order to compete in a dynamic world. In summary, entrepreneurship is not only about new startup ventures.

Social change agents can also be described in terms of entrepreneurship. For example, the founder of the Grameen Bank of Bangladesh, Dr Muhammad Yunus, was awarded the Nobel Peace Prize in 2006 for his entrepreneurship in the social sector. Imitating his ground breaking work, microfinance has emerged as a global industry, offering small affordable loans to people in the lowest income brackets. Comparable innovations can be observed in other areas of social life. In fact, many governments now look to social entrepreneurship and social businesses to help address many of the needs and opportunities in areas of social life, such as health, education, and environmental sustainability. As a consequence, in the early twenty-first century, social entrepreneurship and its close variant social business have become very significant areas of research and practice (Dees, 2007). Founders of this type of venture aim to solve pressing social challenges, and they often see opportunities in areas where the state has been inadequate or failed to act. Importantly, and in contrast to commercially driven entrepreneurship, social entrepreneurs hope to create value that members of society will capture and enjoy, rather than capturing the value for themselves as profits. These novel patterns of entrepreneurship also challenge the adequacy of purely economic accounts, because the goals of such activity are often more social than economic.

These developments pose a number of deep questions for scholars, policy makers, and practitioners. On the one hand, dynamic change and social challenges are stimulating more entrepreneurial action in a wide range of commercial, social, and institutional fields; especially as governments and social institutions often fail to respond to problems in a timely and effective manner. On the other hand, entrepreneurs as a group of actors are unlikely to fill the apparent void created by incapable or ineffective institutions – and scaling effective solutions is always difficult. The question then arises: How can and should social and political institutions nurture the entrepreneurial capabilities required to address the challenges and opportunities of the contemporary world? And what does this question imply for the future of public institutions and good governance? Such questions are already topics of study and no doubt will continue to attract significant attention (Christensen et al., 2006).

Global Trends

Ongoing globalization is having an equally profound impact on entrepreneurship. Across the developing world especially, rates of entrepreneurial activity and economic growth are unprecedented. Countries such as China, India, and Brazil are following the example set by other Asian economies in the late twentieth century (Mathews, 2006). Within a few decades, these nations have grown their economies and command an increasing share of global trade and wealth. Entrepreneurship has been critical to this success. The question is therefore posed, whether a new variant of entrepreneurship is emerging, or will developing economies exhibit the same entrepreneurial processes observed in older developed economies? A number of factors suggest there may be significant differences compared to the role of entrepreneurship in earlier phases of industrial change: in many rapidly emerging economies, government agencies are playing a more important role; imitative strategies are common, as opposed to more disruptive innovations; entrepreneurial action is also growing among the poorest members of society – at the so-called 'bottom of the pyramid' (Prahalad, 2005) – although wider social and political changes have not always been commensurate.

Entrepreneurship is also gaining increasing attention in relation to international business, competitive strategy, marketing, and other fields of management and organizational research. For example, in 2009 the prestigious Strategic Management Society founded a dedicated journal for such research titled the "Strategic Entrepreneurship Journal." Indeed, noting the dynamic nature of markets and industries in the early twenty-first century, some now argue that competitive strategy and entrepreneurship are deeply integrated and almost synonymous (Spender, 2014).

Conclusion

As markets, societies, and technologies continue to change, adapt, and transform, there will be a constant flow of opportunities to create and capture value in new and innovative ways. Moreover, these developments are globally pervasive, and entrepreneurship will be deeply relevant for organized responses to such change. We can therefore predict that entrepreneurship will continue to flourish, as both an area of research and a domain of practice.

See also: Behavioral Theories of Organization; Business Models; Cyclical Industrial Dynamics; Family Business; Innovation; International Business; Network Analysis, History of; Organization: Overview; Organizational Emergence and Firm Formation; Social Business; Social Innovation; Strategic Management; Strategizing; Venture Capital.

Bibliography

Alvarez, S.A., Barney, J.B., 2010. Entrepreneurship and epistemology: the philosophical underpinnings of the study of entrepreneurial opportunities. The Academy of Management Annals 4, 557–583.

Baron, R.A., 2004. The cognitive perspective: a valuable tool for answering entrepreneurship's basic "Why" Questions. Journal of Business Venturing 19 (2), 221–239.

Baron, R.A., 2006. Opportunity recognition as pattern recognition: how entrepreneurs "Connect the dots" to identify new business opportunities. Academy of Management Perspectives 20 (1), 104–119.

Battilana, J., Leca, B., Boxenbaum, E., 2009. How actors change institutions: towards a theory of institutional entrepreneurship. Academy of Management Annals 3 (1), 65–107.

Blank, S., 2013. Why the lean start-up changes everything. Harvard Business Review 91 (5), 63–72.

Bryant, P., 2007. Self-regulation and decision heuristics in entrepreneurial opportunity evaluation and exploitation. Management Decision 45 (4), 732–748.

Burt, R.S., 2004. Structural holes and good ideas. The American Journal of Sociology 110 (2), 349–400.

Busenitz, L.W., West, I., Page, G., Shepherd, D., Nelson, T., Chandler, G.N., Zacharakis, A., 2003. Entrepreneurship research in emergence: past trends and future directions. Journal of Management 29 (3), 285–308.

Christensen, C.M., Baumann, H., Ruggles, R., Sadtler, T.M., 2006. Disruptive innovation for social change. Harvard Business Review 84 (12), 94–101.

Dees, J.G., 2007. Taking social entrepreneurship seriously. Society 44 (3), 24–31.

Grégoire, D.A., Corbett, A.C., McMullen, J.S., 2011. The cognitive perspective in entrepreneurship: an agenda for future research. Journal of Management Studies 48 (6), 1443–1477.

Hébert, R.F., Link, A.N., 2009. A History of Entrepreneurship. Routledge, London, New York.

Hsu, D.H., Roberts, E.B., Eesley, C.E., 2007. Entrepreneurs from technology-based universities: evidence from MIT. Research Policy 36 (5), 768–788.

Johnson, V., 2007. What is organizational imprinting? Cultural entrepreneurship in the founding of the Paris opera. American Journal of Sociology 113 (1), 97–127.

Kirzner, I.M., 1998. How Markets Work: Disequilibrium, Entrepreneurship and Discovery. The Centre for Independent Studies, London.

Knight, F.H., 1942. Profit and entrepreneurial functions. The Journal of Economic History 2 (Suppl.), S126–S132.

Laureiro-Martínez, D., Brusoni, S., Zollo, M., 2010. The neuroscientific foundations of the exploration-exploitation dilemma. Journal of Neuroscience, Psychology, & Economics 3 (2), 95–115.

Mathews, J.A., 2006. Dragon multinationals: new players in 21st century globalization. Asia Pacific Journal of Management 23 (1), 5–27.

Nelson, R.R., Winter, S.G., 1982. An Evolutionary Theory of Economic Change. Harvard University Press, Cambridge, MA.

Prahalad, C.K., 2005. The Fortune at the Bottom of the Pyramid. Wharton School Publishing, Upper Saddle River, NJ.

Sarasvathy, S.D., 2001. Causation and effectuation: toward a theoretical shift from economic inevitability to entrepreneurial contingency. Academy of Management Review 26 (2), 243–263.

Schumpeter, J.A., 1934. The Theory of Economic Development. Harvard Business School Press, Boston, MA.

Shane, S., 2003. A General Theory of Entrepreneurship: The Individual-opportunity Nexus. Edward Elgar, Cheltenham, UK.

Shane, S., Venkataraman, S., 2000. The promise of entrepreneurship as a field of research. Academy of Management Review 25 (1), 217–226.

Shepherd, D.A., Covin, J.G., Kuratko, D.F., 2009. Project failure from corporate entrepreneurship: managing the grief process. Journal of Business Venturing 24 (6), 588–600.

Sorenson, O., Stuart, T.E., 2008. Entrepreneurship: field of dreams? Annals of the Academy Management 2, 517–543.

Spender, J.-C., 2014. Business Strategy: Managing Uncertainty, Opportunity, and Enterprise. Oxford University Press, Oxford.

Teece, D.J., 2007. Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance. Strategic Management Journal 28 (13), 1319–1350.

Wasserman, N., 2012. The Founder's Dilemmas: Anticipating and Avoiding the Pitfalls that Can Sink a Startup. Princeton University Press, Princeton, NJ.

Weber, M., 2007. Theory of Social and Economic Organization. A&C Black Publishers Ltd, London.

Winter, S.G., 2006. Toward a neo-Schumpeterian theory of the firm. Industrial & Corporate Change 15 (1), 125–141.

Entrepreneurship, Geography of

John R Bryson, Birmingham Business School, University of Birmingham, Birmingham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of entrepreneurship plays a central role in understanding the changing geography of the economy. Entrepreneurship is a process that involves the creation and growth of new companies. There is a well-developed literature on entrepreneurship and agglomerations or clusters and also ethnic entrepreneurship. Further research is urgently required to understand the emergence of 'born global firms' or companies that from their initial moment of inception seek to develop an international or global presence.

The study of entrepreneurship plays a central role in understanding the changing geography of the economy (de Groot et al., 2004; Scott, 2006; Stam, 2007). Entrepreneurship is a process or activity that involves the creation and growth of new companies. This process includes the identification of opportunities to supply new goods and services and to organize factor inputs in new ways (Bryson et al., 1993). Entrepreneurship includes developing products and services that may be new to the market or creating new firms that are intended to supply services and goods that already exist. This definition implies that the academic study of entrepreneurship includes a focus on understanding why, when, how, and where entrepreneurship occurs. It must also include a focus on the role gender (Hanson, 2009; Hanson and Blake, 2009), ethnicity, and age play in entrepreneurship. Two elements are central to this definition. First, an emphasis on self-employment in which an individual or group of individuals creates the conditions for their own employment. This is to highlight that entrepreneurship is about engaging in work for personal profit rather than salaried employment. Second, is an emphasis on understanding the processes and practices that result in the establishment of a new enterprise.

The on-going complexity of the division of labor continues to create new business and employment opportunities as often novel ways of dividing tasks are identified. This has led to the emergence of many new types of self-employment including personal trainers, image consultants, dog walkers, and life-style coaches. There are many different types of self-employment or entrepreneurship. This includes 'part-time' self-employment that may be facilitated by full-time employment or self-employment intended to provide flexibility to support caring for dependents. The former includes eBay entrepreneurs who buy and sell on the Web site either as a hobby or as a full-time business. Social entrepreneurs are a distinct type of entrepreneur who are interested in creating products or services to solve social needs and problems.

The development of the World Wide Web (WWW), e-commerce, and networked economies has greatly enhanced the ability of individuals to become entrepreneurs (Nijkamp, 2003). The WWW enables a spare bedroom to be converted into an online shop, but also facilitates the sourcing of goods from companies connected to WWW. This explains the growth in virtual manufacturing firms – firms that design and market manufactured goods but outsource the fabrication process to contract manufacturers. Selling customized products has become much easier as has identifying potential low-cost manufacturers. Innovations in manufacturing technology are also creating new business opportunities. Developments in 3D printing, additive manufacturing, or digital fabrication are enabling individuals to establish craft-based enterprises that have the ability to 3D print-customized products or to design products and outsource the 3D printing process to a specialist supplier. Technological innovations have led to the development of many new opportunities for entrepreneurial behavior using innovative business models with new geographies.

The academic literature on entrepreneurship has emphasized understanding the causes of entrepreneurship and its effects but has had limited success in explaining entrepreneurial behavior. In a small enterprise, the owner-manager plays four distinct roles: financier; orchestrator of capital, raw materials, labor, and financial and management information; innovator constantly searching for ways to improve output and productivity; and finally "a creator of jobs and a member and stimulator in the local and regional milieu" (Marchesnay and Julien, 1990: 269). Marchesnay and Julien suggest that it is unusual to find an entrepreneur who possesses all four characteristics, and more importantly, it is likely that an opposition exists between the "roles of discoverer, innovator, 'adventurer' or 'gambler' and those of organiser, manager and accountant" (1990: 269). The establishment of a new company, or the management of an existing small firm, requires a solution to what Baker and Faulkner have termed a 'combinational problem' (1991: 283). In their study of the Hollywood film industry, they describe the ways in which changing combinations of technical, capital, management, and support services are required at different stages in the evolution of an organization (Kenny and Patton, 2005). The combinational problem explains the existence of serial entrepreneurship; entrepreneurs who continually establish new firms and sell them or transfer the day-to-day management to someone else. Serial entrepreneurs have a special ability to identify gaps in a market, but may not have the skills or interest required to manage a business over a long time period.

Joseph A. Schumpeter was one of the most important social scientists to work on entrepreneurship. The entrepreneur plays a central role in all Schumpeter's writings (1947) with a particular emphasis on the relationship between innovation,

entrepreneurship, and creative destruction. Schumpeter was an economist who reclaimed the concept of the entrepreneur for economics by incorporating social and cultural approaches into his theory of entrepreneurship (Schumpeter, 1947). This theory defines the entrepreneurial process as one that involves the assemblage of new combinations of existing resources or assets. Entrepreneurs create new rules that have the potential to destroy existing processes and products as part of a process of creative destruction. Prior to Schumpeter, the cultural and social aspects of entrepreneurship had been neglected by economists. Nevertheless, Schumpeter was unable to persuade economists to develop a sociologically nuanced account of entrepreneurship as "economists are prone to treat as nonexistent what is not quantifiable and sometimes even what is not measureable" (Schumpeter, 1947: 149–159). It was geographers and sociologists who eventually made an important contribution to understanding the social and cultural aspects of entrepreneurship.

The study of entrepreneurship has become a central focus of the social sciences and has led to the emergence of a specialist academic subdiscipline that focuses on entrepreneurship. Much of this debate has been located in business schools, but geographers have also made important contributions to understanding the complex and evolving geographies of entrepreneurship (Bryson et al., 1993). The geographical debates on entrepreneurship have focused on understanding entrepreneurship and place (Parthasarathy and Aoyama, 2006; Aoyama, 2009; Reuschke, and Van Harn, 2013), new industrial districts or clustering (Scott, 1988), high-technology firms (Saxenian, 1994), entrepreneurship and services (Wood, 1991), ethnic entrepreneurship (Ram et al., 2002; Harris et al., 2012), and business support (Bryson and Daniels, 1998). It is worth noting that the study of the geography of entrepreneurship has never been a core research area in economic geography; at best it has been marginal. The *Oxford Handbook of Economic Geography*, for example, has only two entries in its subject index that refer to entrepreneurship – entrepreneurship and culture and entrepreneurship and location (Clark et al., 2000). The topic of entrepreneurship is present, but hidden in accounts of agglomeration and clustering. Entrepreneurship is a core part of geography, but the term and the literature of entrepreneurship has been largely sidelined within economic geography. Economic geographers have made many contributions to understanding firm behavior, firm formation, entrepreneurship, and cultural diversity (Nathan and Lee, 2013) and agglomeration or clusters (Delgado et al., 2010), but have tended not to frame their analysis within accounts of entrepreneurship. In this context, it is possible to argue that the geography of entrepreneurship is underdeveloped and still awaits the type of attention that has been placed on understanding the evolving geographies of global business (Dicken, 2003).

Entrepreneurship: From Fordism to Post-Fordism

The academic study of entrepreneurship emerged during the early years of the nineteenth century. During the late 1970s and early 1980s, entrepreneurial studies developed as a core focus of academics interested in business and management and of social scientists interested in understanding the geography or sociology of economic activity. Within the business school community, new chairs in entrepreneurship were established and a number of important new academic journals were established including *Frontiers in Entrepreneurship* (1981–Till present), *International Small Business Journal* (1982), *Journal of Business Venturing* (1985–Till present), *Small Business Economics* (1989–Till present), *Entrepreneurship and Regional Development* (1989–Till present), and the *International Journal of Entrepreneurship and Small Business* (2004).

The late 1960s and 1970s had seen a focus on understanding the developing geographies of large transnational corporations (TNCs) and the relative neglect of smaller firms. The post-1960s economy was going to be dominated by the activities of large TNCs and smaller firms were regarded as being of secondary or of only local importance. The period from the late-1970s heralded the rediscovery of the role small- and medium-sized enterprises (SMEs) played in national economies. This rediscovery of the importance of SMEs was partly based on an upsurge in the formation of new firms, but also a new policy interest in understanding SMEs as a possible solution to unemployment. In July 1969, the UK government established the *Committee of Inquiry on Small Firms*, which published the Bolton Report in 1971. Since the publication of the Bolton Report successive governments have introduced policies designed to encourage new firm formation and small firm growth (Bolton Report, 1971). The new emphasis placed by governments on entrepreneurship led to a revival in the academic study of small firms and the emergence of entrepreneurship as an academic subdiscipline. Within geography the new interest in small firms was reflected in a shift from studies that focused on the developing geographies of the Fordist regime of accumulation with a focus on large firms to a new focus on post-Fordism and an appreciation of the role SMEs played in flexible production systems. Within geography, the conceptual framework developed by French regulation theory was extremely influential in shifting the focus of research toward understanding the geography of smaller firms (Tickell and Peck, 1992).

Fordism was identified as a regime of accumulation based on mass production and consumption and the emergence of large firms. Wages were set by collective bargaining and a substantial proportion of employment was in manufacturing firms. Post-Fordism was a result of a crisis in Fordism that emerged in the early 1970s. A new regime of post-Fordism was identified based on flexible, specialized production. This included a revival in the role of SMEs and localized or clustered value chains (Wood, 1991). The identification of a shift from Fordism to post-Fordism highlighted the major transformations that were underway in the economies and societies of developed market economies. Nevertheless, the distinction also had its problems. The social sciences translated Henry Ford into Fordism and Frederick Taylor into Taylorism. These two individuals are identified with processes that are considered by many social scientists to have been critical in the evolution of advanced capitalism (Williams et al., 1992). Fordism became such a key concept that post-Fordism and neo-Fordism were developed as terms to define production systems that developed after the crisis of Fordism, identified as having occurred in the 1970s.

New Industrial Districts and Entrepreneurship

The shift in research emphasis toward post-Fordism was associated with a growth in research that explored the functioning of industrial districts or clusters (Scott, 1988). It is worth noting that economic geographers during the late 1960s and early 1970s focused on understanding the importance of 'industrial linkage.' This early literature reflects geographers' long-term interest in the spatial aspects of interfirm relationships. Industrial linkage is defined "as occurring when one manufacturing firm purchases inputs of goods and services from, or sells outputs to, another manufacturing firm" (Keeble, 1976: 61). Initially studies emphasized localized linkages, but "linkages can and do occur over considerable distances" (Keeble, 1976: 61). Industrial linkage is explained by the operation of a division of labor with firms specializing to minimize total production costs and to enhance profitability. The industrial linkage literature did include studies of industrial districts that explored "sharply demarcated 'swarms' of small firms in particular industries" (Keeble, 1976: 62) and perhaps the most famous study is Wise's analysis of the Birmingham jewelry and gun quarters (1949). An industrial district consists of a territorial system of SMEs. Industrial district studies shifted the research focus away from understanding the activities of single firms to understanding the social characteristics and functions of colocated firms. The research identified distinct agglomerations of firms that specialized in the creation of particular products. A central element of this research was a focus on embeddedness as a key analytical construct (Granovetter, 1985). This approach emphasized the embedding of business activities within much deeper and complex social processes. Much of this literature stressed the importance of localized learning and the transfer of forms of tacit rather than codified knowledge between individuals (Maskell and Malmberg, 1999). Codified knowledge could be relatively easily transferred between individuals and firms, but tacit knowledge could only be transferred through everyday practice and required various forms of face-to-face interaction. This literature also highlighted the relationship between place and entrepreneurship with a focus on cultural, social, economic, and political differences. There are thus important differences between places in terms of the emphasis different cultures place on entrepreneurship as well as differences in the type and availability of factor inputs (Aoyama, 2009; Barbu et al., 2013). Differences in factor inputs include access to finance, business support, and taxation systems intended to encourage entrepreneurial behavior (Scott, 2006).

The geographical literature on clustering is extensive (Scott, 1988; Taylor, 2010). Much of this research focused on what were identified as the three most important regions of the Third Italy: Veneto, Emila-Romagna, and Tuscany (Brusco, 1986). Industrial districts consist of two related processes. One set of processes reflects the functional characteristics of external economies. These include an extensive division of labor between colocated firms that reflect close intra- and intersectoral input–output relationships. Distinct product specialization at the level of the enterprise enables firms to work together as part of a localized value chain. Second, a local concentration includes a territorial- or place-based dimension. This includes local concentrations of skilled labor and of networks and local structures that facilitate local learning. These two processes also contribute to new firm formation or entrepreneurship. The existence of networks that facilitate localized learning and access to skilled labor as well as raw materials encourage new firms to be established. An industrial district is continually reinforced through the establishment of new firms. Employees working for firms within the district are able to access the resources required to establish their own companies. Such resources include access to localized learning, raw materials, suppliers, and potential customers. This implies that it may be easier to establish a new firm within a cluster. The density of small- and medium-sized firms in a cluster will reflect on-going processes of entrepreneurship and new firm formation. This means that the geographical literature on industrial districts or clusters is also a literature on entrepreneurship (Wood, 2004; Delgado et al., 2010).

Some industrial sectors have a greater propensity to form company spin-offs or new firms. An important enabler of new firm formation is the existence of limited barriers to firm formation. These barriers usually consist of access to capital required to purchase or lease land, machines, and raw materials and to cover the expenses related to firm formation. Establishing a service firm is much easier than establishing a capital-intensive manufacturing company. Many knowledge-based business and professional service firms require expertise and a reputation, but, in many instances, limited up-front capital investment.

Social science and economic geography is awash with networks. The 'network turn' in the social sciences has multiple origins. One root lies in actor-network theory (Law, 1999), and other roots are found in accounts of entrepreneurship and social networks (Emirbayer and Goodwin, 1994; Aldrich, 2004). The term 'network' had been transformed in to the dominant collective noun that is used to describe a collection of interactions that occur between organizations, firms, companies, enterprises, and people. Networks play an important role in the literature on agglomeration, entrepreneurship, and localized learning; networks constitute another conceptual framework that geographers have used to explore space, place, and entrepreneurial activity (Bryson et al., 1993).

Clusters have become an important element in many regional economic development strategies but regional strategies overemphasize the importance of clusters at the expense of other forms of policy intervention. Recently, Taylor (2010) has provided an important critique of both the academic and policy literature on clusters. In this critique, he highlights the dangers of adopting cluster development policies without understanding the complexity and diversity that exist within a local or regional economy.

Ethnic Entrepreneurship and Translocal or Transnational Relationships

Minority ethnic entrepreneurship is a major theme within the entrepreneurship literature. Within the UK, this focus can be traced back to Rex and Moore's (1967) classic account of racial conflict in Sparkbrook, Birmingham. Rex and Moore argue that urban society should be conceptualized as consisting of "a number of overlapping and sometimes contradictory systems

of social relations" (1967: 13). This type of analysis is echoed in Albrow's (1997) analysis of old established local people and an Asian community that both live in Tooting, an area in inner London. He argues that "these people inhabit co-existing social spheres, coeval and overlapping in space, but with fundamentally different horizons and time-spans. The reality of Tooting is constituted by the intermeshing and interrelating of these spheres" (Albrow, 1997: 47). These accounts of places that contain distinct and unrelated communities highlight both tensions that can exist between different communities as well as the advantages that can be associated with multiculturalism. Chan (2006) in an interesting account of multiculturalism in Birmingham, UK, based on the analysis of newspapers reveals the ways in which Birmingham City Council has 'rescripted' the character of Birmingham's population as an asset and opportunity rather than a problem or a threat. Part of this rescripting is based around the perceived importance of ethnic entrepreneurs as potential contributors to Birmingham's regeneration through their ability to compete by drawing upon their position in transnational networks. Chan argues that "the ethnic entrepreneur has become analogous with contributions garnered across international borders" (2006: 85) and that this type of activity while being undertheorized by academics had been identified in newspaper articles published in 1990.

Ethnic entrepreneurs provide places with new opportunities to exploit transnational networks. Ethnic entrepreneurs tend to concentrate their activities in sectors, such as catering, with low-entry barriers (Ram et al., 2002). Many of the sectors that ethnic entrepreneurs target suffer from intense competition and 'saturation.' A study of South Asian restaurants in Birmingham noted that this sector had reached saturation so that "any additional supplier renders the market unsustainable at existing levels of consumer demand" and that excessive competition existed (Ram et al., 2002: 27). Minority ethnic entrepreneurs enable places to experience 'globalization from below' – translocal relationships that link migrant communities with their ethnic homeland. This conceptualization of globalization as a process that can be facilitated from below though the articulation of ethnic networks is a well-established aspect of the literature on ethnic enterprise. Ethnic communities located in Western cities can produce ethnic products that are exported to ethnic communities located elsewhere. Thus, Rezaei, though an analysis of the Persian carpet industry, argues that "the formation of immigrant business networks across countries can be understood as a bottom-up globalisation process which complements the standard large firms/top-down one."

Business and professional services (BPS) that provide intermediate inputs to clients can function to connect different parts of the world together (Bryson et al., 1993). An excellent example of transnational BPS that link places is found in Zhou's (1998) analysis of Chinese-owned BPS firms based in Los Angeles. Zhou argues that small, privately owned, and decentralized Chinese accountants, banks, and computer companies located in Los Angeles play a key role in connecting immigrant economic activities with the international circulation of information, goods and capital and especially with their place of origin. In essence, these companies appear to act as enabling mechanisms that ease the expansion of Chinese-based client companies in particular localities in the United States. Since the 1970s, the number of Chinese immigrants and investments has grown rapidly in Los Angeles and it would appear that this growth is self-supporting. The driver behind this process is the operation of ethnic networks that extend access, and provide a mechanism for, regulating and coordinating long-distance interaction. The existing Los Angeles Chinese community utilizes its social relationships with Chinese-based enterprises and acts as an important bridge between Chinese and American business cultures and knowledge. Mitchell and Olds (2000) make the same point in a study of the investment and management practices of Hong Kong property tycoons in Vancouver; Chinese actors link both sides of the Pacific through blending knowledge and business practices. They argue that such knowledge sharing is more flexible than in non-Chinese business firms, and that the flexibility of Chinese firms may be a competitive advantage.

It is important not to 'other' minority ethnic entrepreneurs by suggesting that their business activities are in some way different to other members of a society or to other entrepreneurs. This is an important point in that it is important not to overstate the contribution ethnicity makes to the ways in which enterprises are established and managed (Ram et al., 2002). Much research on ethnic enterprise is constructed around notions of social capital, networks, and communal-familial solidarity as a competitive resource. Emphasizing cultural values in understanding ethnic enterprise is perhaps the oldest conceptual framework that has been applied to understanding ethnic business in the UK (Barrett et al., 2003: 112); this cultural approach tends to focus on relationships within the ethnic group and as such overlooks external relationships with customers and markets. An overemphasis on cultural attributes draws attention away from the institutional and market contexts within which all firms have to operate. Similarly, Ram et al. (2002: 22) note that this type of endogenous approach to ethnic enterprise neglects the role of urban context. Barrett et al. (2003: 133) observe that the cultural approach refuses to die and that "it has resurfaced in that part of human geography which has taken the 'cultural turn.'" Studies based upon "ethnic specificity culturalism may, at times, risk inadvertently promoting a form of stereotyping by highlighting inter-group cultural difference as an explanation of entrepreneurial performance" (Barrett et al., 2003: 113). It is important not to ignore the ethnic embeddedness perspective, but following Barrett et al. (2003) and Ram et al. (2002) the starting point for an analysis of ethnic entrepreneurship must be the analysis of a set of exogenous factors that are found within the local economic, social, and political systems.

Entrepreneurship and Born Global Firms – A New Research Agenda

Economic geography has yet to develop a major research or teaching agenda on entrepreneurship. A reading of the existing geographical literature on entrepreneurship leads to the assumption that entrepreneurship is about clustering or perhaps minority ethnic populations. More recently, an interesting geographical literature is developing on the relationship between place and entrepreneurship and with a particular focus on self-employment (Reuschke and Van Harn, 2013). The

WWW, new technologies and the on-going division of labor have created opportunities for new forms of entrepreneurship. This has created new business models with distinctive and often unusual economic geographies. The study of small firms and entrepreneurship has tended to be relegated to be of secondary importance compared to research into global value chains, TNCs, and international business. Nevertheless, entrepreneurship is a fundamental geographical process, which is both enabled and constrained by place-based relationships and translocal relationships (Yeung, 2009, 2010). There is an interesting opportunity for economic geography to pay more attention to entrepreneurs and entrepreneurship; the study of all types of entrepreneurship and entrepreneurial behavior should form one of the central research and teaching pillars of economic geography.

The world of the entrepreneur is being transformed with technological developments and the emergence of new business models. Developments in e-commerce have reduced firm formation barriers; new firms may be established in unusual locations as long as they have an effective Web presence. The rise of Web-based businesses has led to a major revolution in the geography of entrepreneurship. Nevertheless, this is a revolution that has yet to be explored by geographers but exploratory research has been undertaken in marketing studies (Chetty and Campbell-Hunt, 2004; Hashai, 2011). Conventional approaches to new firm formation highlight that firms are initially established to provide goods and services to local markets. Over time, a firm may begin to trade beyond its local market and eventually may begin to export goods or services. Over the last decade, new firms have been established, which have been 'born global.' A born global firm is a business that from its initial moment of inception seeks to develop an international or global presence. The born global firm begins with a 'borderless' view of the world economy and the formation process includes a strategy to target clients or customers located in other countries. The emergence of born global firms represents an important opportunity to understand the complex ways in which individuals and groups of individuals are developing novel business models that operate at the interface between a set of local and global processes.

See also: Economic Geography; Economic Restructuring: Geographic Aspects; Economic Sociology; Ethnicity, Sociology of; Finance, Geography of; Globalization: Geographical Aspects; Industrial Geography; International Business: Geographic Aspects; Local Economic Development; Service Economy, Geography of.

Bibliography

Albrow, J., 1997. Travelling beyond local cultures: socioscapes in a global city. In: Eade, John (Ed.), Living the Global City. Routledge, London, pp. 35–52.

Aldrich, H., 2004. Organizations Evolving. Sage, London.

Aoyama, Y., 2009. Entrepreneurship and regional culture: the case of Hamamatsu and Kyoto, Japan. Regional Studies 43 (3), 495–512.

Baker, W.E., Faulkner, R.R., 1991. Role as resource in the Hollywood Film Industry. American Journal of Sociology 97 (2), 279–309.

Barbu, M., Dunford, M., Liu, W., 2013. Employment, entrepreneurship, and citizenship in a globalised economy: the Chinese in Prato. Environment and Planning A 45 (10), 2420–2441.

Barrett, G.A., Jones, T.P., McEvoy, D., 2003. United Kingdom: severely constrained entrepreneurialism. In: Kloosterman, R., Rath, J. (Eds.), Immigrant Entrepreneurs: Venturing Abroad in the Age of Globalization. Berg, Oxford, pp. 101–122.

Brusco, S., 1986. Small firms and industrial districts: the experience of Italy. In: Keeble, D., Weever, E. (Eds.), New Firms and Regional Development in Europe. Croom Helm, London, pp. 184–202.

Bryson, J.R., Daniels, P.W., 1998. Business Link, strong ties and the walls of silence: small and medium-sized enterprises and external business service expertise. Environment and Planning C: Government and Policy 16, 265–280.

Bryson, J.R., Wood, P., Keeble, D., 1993. Business networks, small firm flexibility and regional development in UK business services. Entrepreneurship and Regional Development 5, 265–277.

Chan, W.F., 2006. Re-scripting the character of Birmingham's ethnic minority population: assets and others in the stories of a multicultural city. Area 38 (1), 79–88.

Chetty, S., Campbell-Hunt, C., 2004. A strategic approach to internationalization: a traditional versus a "Born-Global" approach. Journal of International Marketing 12 (1), 57–81.

Clark, G.L., Feldman, M.P., Gertler, M.S. (Eds.), 2000. The Oxford Handbook of Economic Geography. Oxford University Press, Oxford.

Delgado, M., Porter, M.E., Stern, S., 2010. Clusters and entrepreneurship. Journal of Economic Geography 10 (4), 495–518.

Dicken, P., 2003. Global Shift: Transforming the World Economy. Sage, London.

Emirbayer, M., Goodwin, J., 1994. Network analysis, culture, and the problem of Agency. American Journal of Sociology 99 (6), 1411–1454.

Granovetter, M., 1985. Economic action and social structures: the problems of embeddedness. American Journal of Sociology 91 (3), 481–510.

de Groot, H.L.F., Nijkamp, P., Stough, R.R. (Eds.), 2004. Entrepreneurship and Regional Economic Development: A Spatial Perspective. Edward Elgar, Cheltenham.

Hanson, S., 2009. Changing places through women's entrepreneurship. Economic Geography 85 (3), 245–267.

Hanson, S., Blake, M., 2009. Gender and entrepreneurial networks. Regional Studies 43 (1), 135–149.

Harris, C., Moran, D., Bryson, J.R., 2012. EU accession migration: national insurance number allocations and the geographies of Polish labour immigration to the UK. Tijdschrift voor economische en sociale geografie 103 (2), 209–221.

Hashai, N., 2011. Sequencing the expansion of geographic scope and foreign operations by "born global" firms. Journal of International Business Studies 42, 995–1015.

Keeble, D., 1976. Industrial Location and Planning in the United Kingdom. Methuen, London.

Kenny, M., Patton, D., 2005. Entrepreneurial geographies: support networks in three high-technology industries. Economic Geography 81 (2), 201–228.

Law, John, 1999. After ANT: complexity, naming and topology. In: Law, J., Hassard, J. (Eds.), Actor Network Theory and after. Blackwell, Oxford, pp. 1–15.

Maskell, P., Malmberg, A., 1999. Localised learning and industrial competitiveness. Cambridge Journal of Economics 23, 167–186.

Marchesnay, M., Julien, P., 1990. The small business: as a transaction space. Entrepreneurship and Regional Development 2, 267–277.

Mitchell, K., Olds, K., 2000. Chinese business networks and the globalisation of property markets in the Pacific Rim. In: Yeung, H., Olds, K. (Eds.), Globalization of Chinese Business' in Globalization of Chinese Business Firms. Macmillan Press, London, pp. 195–219.

Nathan, M., Lee, N., 2013. Cultural diversity, innovation, and entrepreneurship: firm-level evidence from London. Economic Geography 89 (4), 367–394.

Nijkamp, P., 2003. Entrepreneurship in a modern network economy. Regional Studies 37 (4), 395–405.

Parthasarathy, B., Aoyama, Y., 2006. From software services to R&D services: local entrepreneurship in the software industry in Bangalore, India. Environment and Planning A 38 (7), 1269–1285.

Ram, M., Jones, T., Abbas, T., Sanghera, B., 2002. Ethnic minority enterprise in its urban context: South Asian restaurants in Birmingham. International Journal of Urban and Regional Research 26, 24–40.

Report of the Committee of Inquiry on Small Firms, chaired by J.E. Bolton, Cmnd, 4811 Bolton Report, 1971. HMSO, London.

Reuschke, D., Van Harn, M., 2013. Testing the 'residential rootedness' hypothesis of self-employment for Germany and the UK. Environment and Planning A 45 (5), 1219–1239.

Rezaei, S., 2000. International Trade through Ethnic Networks: The Case of Persian Carpet Trade. Case study No. 1. LOK-study project (Management, Organisation & Competences). Copenhagen Business School.

Rex, J., Moore, R., 1967. Race Community and Conflict: A Study of Sparkbrook. Oxford University Press.

Saxenian, A., 1994. Regional Advantage: Culture and Competition in Silicon Valley. Harvard University Press, Cambridge.
Schumpeter, Joseph A., 1947. The creative response in economic history. The Journal of Economic History 7 (2), 149–159.
Scott, Alan, 1988. New Industrial Spaces: Flexible Production, Organisation and Regional Development in North America and Western Europe. Pion, England.
Shane, S., 2003. A General Theory of Entrepreneurship: The Individual-Opportunity Nexus. Edward Elgar, Cheltenham.
Scott, Alan, 2006. Entrepreneurship, innovation and industrial development: geography and the creative field revised. Small Business Economics 26, 1–24.
Stam, E., 2007. Why butterflies don't leave: locational behavior of entrepreneurial firms. Economic Geography 83 (1), 27–50.
Taylor, Michael, 2010. Clusters: a mesmerising mantra. Tijdschrift voor economische en sociale geografie 101 (3), 276–286.
Tickell, Adam, Peck, Jamie, 1992. Accumulation, regulation and the geographies of post-Fordism: missing links in regulationist research. Progress in Human Geography 16, 190–218.
Williams, K., Haslam, C., Williams, J., 1992. Ford versus 'Fordism': The Beginning of Mass Production? Work, Employment & Society 6 (4), 517–555.
Wise, Michael J., 1949. On the evolution of the jewellery and gun quarters in Birmingham. Transactions and Papers of the Institute of British Geographers 15, 59–72.
Wood, A.M., Watts, H.D., Wardle, P., 2004. Growth-orientated small firms and the nature and extent of local embeddedness: the case of a traditional metalworking cluster. Growth and Change 35 (4), 419–433.
Wood, Peter A., 1991. Flexible accumulation and the rise of business services. Transactions of the Institute of British Geographers 16, 160–172.
Yeung, H.W., 2009. Transnationalizing entrepreneurship: a critical agenda for economic geography. Progress in Human Geography 33 (2), 210–235.
Yeung, H.W. (Ed.), 2010. Globalizing Regional Development in East Asia: Production Networks, Clusters, and Entrepreneurship. Routledge, London. Regions and Cities Series No. 41.
Zhou, Y., 1998. Beyond ethnic enclaves: location strategies of Chinese producer service firms in Los Angeles. Economic Geography 74 (3), 228–251.

Environmental Archaeology

Nicholas Branch, School of Archaeology, Geography and Environmental Science, University of Reading, Reading, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S.E. Van Der Leeuw, volume 7, pp. 4585–4589, © 2001, Elsevier Ltd.

Abstract

Environmental archaeology, a subdiscipline of archaeology, is defined as the study of the environment and its relationship with people through time. To elucidate this definition, the article summarizes the historical development of the subject, and explores the main subfields of environmental archaeology, namely *geoarchaeology*, *zooarchaeology*, *archaeobotany*, and *geochronology*, with reference to specific case studies. The article also highlights some of the main research themes in environmental archaeology.

Introduction

Environmental archaeology has developed into a major subdiscipline of archaeology over the past five decades, and has attracted scientists from many different backgrounds of academia, including geography, geology, botany, zoology, physics, chemistry, and of course archaeology and anthropology, as well as providing specialist undergraduate and postgraduate training at universities throughout the world. We can define 'environmental archaeology' in archaeology, therefore, as the study of the environment and its relationship with people through time (see Branch et al., 2005); this has in turn led to the development of distinctive subdivisions of environmental archaeology called *geoarchaeology*, *zooarchaeology*, *archaeobotany*, and *geochronology* (see below).

History

During the eighteenth and nineteenth centuries, geologists, such as Lyell (1797–1875) and Hutton (1726–97), by examining the principles of uniformitarianism and laws of stratigraphy, and biologists, such as Darwin (1809–82), by contributing to evolutionary theory, laid the foundations of environmental archaeology. Uniformitarianism, in particular, is a key scientific philosophy. It refers to the concept that natural processes observed in the present day also occurred in the past. The idea applies to pedological (soil) and sedimentological processes, and the environmental parameters that govern the distribution of plants and animals. For example, we can observe today that *Alnus glutinosa* (common alder) inhabits wet areas, such as fens and marshes, and the edges of rivers, ponds, and lakes. If we find subfossils (e.g., flowers and seeds) of alder in either geological or archaeological deposits, then we can confidently assume that wetland habitats existed near to the site in the past. Although there are instances where this principle is questionable (e.g., history of *Taxus baccata* (yew) woodland; Branch et al., 2012), the application of uniformitarianism and other universal laws governing natural processes, means that environmental archaeology is a truly scientific discipline. The existence of universal laws directing human behavior is a matter of considerable debate, and principles such as uniformitarianism are undoubtedly less valuable when trying to explain archaeological phenomena. Nevertheless, recognition of the scientific basis of archaeology (see Pitt-Rivers (1827–1900), Morgan (1818–81), and Taylor (1832–1917)), especially from the 1930s, led eventually to the formulation of a new theoretical framework known as *processualism* in the 1960s (see Binford, 1968; Clarke, 1978). This paradigm shift in archaeology led to an innately scientific approach to research (Renfrew and Bahn, 2000), involving multidisciplinary studies and further complex theoretical developments, such as systems theory (Flannery, 1968; Brothwell and Higgs, 1969). Although succeeded by *postprocessualism* as the dominant paradigm in archaeology (see Archaeology, Theory in), environmental archaeology maintains a science-based approach, which seeks to understand the relationships between the biological and physical environment, and social and economic changes (Butzer, 1982).

Subdivisions of Environmental Archaeology

The fields of geoarchaeology, zooarchaeology, archaeobotany, and geochronology form the basis of environmental archaeology, and are part of a wider subject area known as archaeological science, which includes the study of human remains, materials science, conservation, and geophysics. These fields embrace the study of ancient sediments and soils, and fossil plant and animal remains, recovered from a wide variety of *archives*, which may be located in a diverse range of environments – terrestrial, marine, and ice. *Archaeological archives* are those located on settlement or occupation sites, irrespective of the length or nature of the human activities; they include rubbish and storage pits, postholes, drainage ditches, moats, wells, postholes, cesspits, hearths, and ancient buried soils, as well as the archives listed below. In contrast, *Geological archives* have no known evidence of occupation but may nevertheless provide important information on environmental change due to natural processes and human impact, and the broader environmental context of human activities; these typically include peat bogs and lakes, floodplains and valleys, as well as marine sediments and ice. Environmental archaeologists therefore utilize many of the field and laboratory methods used in biological and earth sciences, and data generated by

quaternary scientists, especially in the fields of paleoclimatology and paleoecology.

Geoarchaeology

Geoarchaeology, and the *geoarchaeologist*, is concerned with landscape and stratigraphic formation (e.g., sedimentary events and soil processes) and modification processes (i.e., deposit alteration), and draws mainly on geological, sedimentological, and soil science analytical techniques used in the field and laboratory (Dincauze, 2000; Branch et al., 2005). The primary role is therefore to record the physical characteristics (e.g., color, dryness, wetness) and composition (e.g., organic matter content and type (e.g., moss, herbaceous and wood peat, or lake sediment), and particle size (gravel, sand, silt, clay)) of stratigraphy in archaeological and geological archives. This information is then used to reconstruct and explain the events or processes responsible for the formation of the stratigraphic sequence, including postdepositional transformation. Recognition of the boundaries between *units* of stratigraphy (or *contexts*) is a critical part of this procedure because these may define changing depositional processes, both cultural and natural, and the relative rate of change, i.e., a *conformable boundary* (gradual change), and an *unconformable boundary* (rapid change).

Geoarchaeologists therefore utilize a wide range of analytical procedures to record and interpret stratigraphy, including standard approaches for sediment and soil description and classification, and laboratory methods, such as micromorphology, particle size analysis, magnetic susceptibility, and multielement geochemistry. Using these procedures, and applying uniformitarian principles, geoarchaeologists can assign sedimentary units to a specific natural process. For example, Riddiford and colleagues (2012) by analyzing a complex series of sediments in the Seille Valley (NE France) have been able to reconstruct the changing patterns of river channel formation, migration, and abandonment during an extended period of human settlement. The introduction of a human component into stratigraphic formation provides an additional complexity, and understanding the relationship between the event or process and the morphology of the deposit often presents a significant challenge even with the range of analytical procedures available. This is particularly apparent in archaeological sites where multiple phases of human occupation involving domestic and industrial activities (e.g., Tell Brak, Syria) result in a diverse range of deposits associated with ditches, pits, latrines, structures, and demolition debris (Matthews, 2003). Nevertheless, these procedures are particularly successful for a range of common depositional environments and their sedimentary deposits (**Table 1**).

Particle size and geochemical analyses, in particular, are two key approaches used to characterize stratigraphy in terms of not only formation processes but also past social and economic activities. Both techniques define the geological source of the sediments and soils examined, and hence susceptibility to weathering and erosion of parent material, and resultant mineral grain size (sand, silt, and clay) and elemental composition. Economically, these characteristics help to define the agricultural potential of soils, especially their cropping capability, drainage, level of biological activity, and nutrient status, as well as the management history of the soil profile (e.g., manuring regimes) and finally the sustainability of the farming system. For example, sandy soils (silica-rich rocks) with a low pH (acid) have less agricultural potential, especially for the growth of crops, than alkaline soils, derived from calcium carbonate-rich rocks, which have a higher pH (e.g., Gale and Hoare, 1991). For example, the study of ancient agricultural terraces in the Peruvian Andes using these methods has enhanced understanding of soil quality, manuring regimes, and nutrient status (e.g., Kemp et al., 2006). Phosphorus is a key elemental indicator, and enhanced values in soils (by absorption) on habitation sites suggest waste materials associated with human and animal activities (e.g., Parnell et al., 2002). The analysis of soils and sediments in thin section (micromorphology) permits examination of microscopic characteristics, especially minerals and inclusions, and is especially important

Table 1 Common depositional environments and their sedimentary deposits

Alluvium – water-borne sediments, mainly fine-grained (clay and silt), and deposited in lowland river systems either on the floodplain (over-bank deposits) or within former river and stream channels. The most extensive deposits are Holocene in age (last 11 700 years) but they also commonly occur in previous glacial and interglacial periods (e.g., Brown, 2003).
Colluvium – slope wash depositing unsorted sediment on the sides and bottoms of dry valleys mainly due to gravity and farming practices. The particle size is highly variable and greatly depends on the source parent material (e.g., Wilkinson, 2003).
Windblown sediments – sand (e.g., dunes) and silt (e.g., loess) deposited by wind action within open, poorly vegetated landscapes, and desert and dry temperate environments, as well as former periglacial zones.
Marine sediments – present within intertidal zones and coastal wetlands, and mainly fine-grained mineral sediments. These sediments have been used successfully to record periods of Holocene sea level rise (marine transgression), and to investigate the relationships with ecosystem and societal change (e.g., Bell et al., 2000).
Glacial sediments – there is a wide variety of sediments of glacial origin, many of which are unsorted and, therefore, composed of a range of particle sizes, including gravel, and associated with glacial landforms, such as moraines. Those sediments located within periglacial environments, such as river sands and gravels, and solifluction deposits, are occasionally associated with archaeological remains, such as Paleolithic tools (e.g., Hosfield, 2011).
Volcanic sediments – these are represented by a range of particle sizes (pumice, lapilli, ash) when deposited proximal to the source (e.g., Vesuvius and the Roman town of Pompeii; Sigurôsson et al., 1982) but often fine-grained tephra distal to the volcano. The latter, either occurring as visible tephra layers or as cryptotephra (nonvisible), is especially useful as chronological marker horizons in both archaeological and geological archives.
Organic sediments – mainly peat forming in fens, raised bogs, and blanket bogs and composed of mosses, herbaceous remains and wood and organic-rich lake sediments (gyttja). The waterlogged, anaerobic conditions provide an excellent environment for the preservation of archaeology, and subfossil biological remains (e.g., Branch et al., 2012).
Cave sediments – comprise a mixture of well-stratified water-borne and aeolian sediments, and rock fall. Caves are particularly attractive for human habitation during the Pleistocene and Holocene resulting in deep stratigraphy often composed of domestic waste, hearths, and stabling deposits.

for identifying human modification of soil sequences (e.g., manuring), use of space (e.g., stabling, flooring), and the inclusion of materials into cultural units (e.g., bedding, feces, fodder, charcoal, ash, plant material). Geoarchaeologists also utilize the magnetic susceptibility of soils and sediments to test for burning, the addition of heated materials, and the natural weathering of soils. It plays a useful role, therefore, in the detection of activity areas on habitation sites, as well as identifying erosional events in the catchment areas of lakes (see Thompson and Oldfield, 1986; Dearing, 1999).

Zooarchaeology

Zooarchaeology, and the *zooarchaeologist or archaeozoologist*, has a central focus on human subsistence of both wild and domesticated animals (*paleoeconomy*), but also broader environmental reconstruction (*paleoenvironment*), such as changes in climate and vegetation cover (*see* Zooarchaeology). The goal, therefore, is to enhance knowledge of the impact of humans on the natural environment, and the effects of natural environmental changes on humans, including the economy, diet, and daily life of human communities. This requires the use of a range of analytical methods, and the application of experimental and ethnographic data (*see* Ethnoarchaeology), to obtain independent lines of complementary scientific evidence from both archaeological and geological archives, and their sedimentary deposits (*see* Geoarchaelogy above).

The most common vertebrate remains found within archaeological and geological archives are bones (e.g., fish, small and large mammals). They can be identified to ascertain the genus or species, wild or domesticated, the type of bone (e.g., femur), and the sex and age of the animal. These data permit understanding of human economy and diet, in particular hunting and husbandry practices, pathologies, and enable reconstruction of past environments. In addition to the quantitative methods, such as bone measurement and weight, measurements of stable carbon ($\delta^{13}C‰$) and nitrogen ($\delta^{15}N‰$) isotope ratios of collagen, lipids, and cholesterol in human bone and tissue have improved understanding of the importance of terrestrial and marine food, the use of exotic food, and changing dietary patterns. In addition, analyses of oxygen isotopes ($\delta^{18}0$) in animal bone phosphate that correspond to local meteoric water values (drinking water of the animals) and mean annual air temperature have provided quantitative reconstructions of past climatic regimes.

Mollusca preserved on land (e.g., soil and mires), and in freshwater (e.g., lakes), brackish water (e.g., high salt marsh), and marine (e.g., estuaries) sediments provide three categories of information useful to environmental archaeology: (1) broad paleoenvironmental reconstruction, (2) human impact on the natural environment, and (3) human economy and diet, which is mainly (but not exclusively) confined to shellfish. For example, in the UK, hillslope sediment (e.g., Bell, 1983) and calcium carbonate-rich sediment (spring chalk) has provided rich sources of information on past human environments. The exploitation of marine mollusks has provided valuable information on subsistence practices, including diet, seasonality, technological developments, and trade (e.g., Rick and Erlandson, 2000; Mannino and Thomas, 2001; Cabral and da Silva, 2003). For example, at Ysterfontein (South Africa), a coastal cliff shelter, evidence has been provided for the exploitation of shellfish, including black mussel and granite limpet, as well as tortoise, cape dune molerat, and jackass penguin among many other animals, enabling the reconstruction of human exploitation strategies and population levels during the Middle Stone Age (*c*.60–40 000 years ago) (Halkett et al., 2003). Insect remains provide unique paleoclimatic records using the mutual climatic range method (Atkinson et al., 1986; Hellqvist and Lemdahl, 1996), as well as information on human and animal diet, the function of archaeological features, the condition of human and animal mummified remains, and the contents of offerings (e.g., Panagiotakopulu, 2001). For example, insects associated with urban archaeological deposits in York (UK) have provided precise data on local environmental living conditions and human activities. Investigators established six main groups of insects that commonly co-occurred in specific contexts in the past: (1) inside buildings and dry ('house fauna'), such as dry plant debris, wood borers, the human flea, and stored products; (2) foul decaying organic matter, rich in ammonia, commonly found in stable manure; (3) foul decaying organic matter, associated with moist ground; (4) very foul conditions; (5) cesspits; and (6) wool cleanings. These groups also highlighted the complex interpretational problems associated with single archaeological layers (e.g., floors) containing insect assemblages from several habitats, reflecting a range of human activities (e.g., Carrott and Kenward, 2001).

Several lesser groups of zooarchaeological remains are worthy of note. Ostracoda (aquatic invertebrates), like diatoms (below), are highly sensitive to changes in salinity with three main assemblages identifiable: freshwater ($<0.5‰$; lakes, rivers, and ponds), brackish water (0.5–30‰; lagoons and salt marshes), and marine (30–40‰; oceans and seas), as well as rainfall, temperature, and alkalinity. They are valuable in archaeology as indicators of changing coastline morphology and the height of relative sea level. Foraminifera are unicellular organisms with a well-known ecology. Different species occupy niches in marine environments according to food availability, levels of predation, type of sediment, light penetration, water temperature, and salinity, and thus are useful indicators of changes in water depth, salinity, and climate in coastal environments. Cladocera are microscopic animals (arthropods) with an ecology divided into two broad groups based upon their ecological preferences: (1) offshore planktonic species and (2) inshore littoral species. They are useful for reconstructing temperature and precipitation, and erosion due to human activity (e.g., Jeppesen et al., 2001). The application of Cladocera analysis to environmental archaeology is still in its infancy, but the technique clearly has enormous potential. For example, analyses of Cladocera from 26 lakes in Poland, and one in Croatia, have enabled reconstruction of climate change, lake level oscillations, nutrient status, and chemistry. It was concluded that lake water nutrient levels increased during phases of human settlement due to the deposition of refuse into the lake (Szeroczyńska, 2002).

Archaeobotany

Archaeobotany, and the *archaeobotanist*, also focuses on human economy and diet, but also environmental change. The most

common plant remains found on archaeological sites are seeds that have been charred (exposure to heat, e.g., hearth) or waterlogged (exposure to excess moisture, e.g., ditch, moat, well), and to a lesser extent mineralized (replacement by salts, e.g., cesspits), desiccated (exposure to extreme aridity), or preserved as pottery impressions. For example, Old World cereal grains (e.g., species of wheat and barley) might be found with charred or mineralized parts of the plant, such as straw. In geological archives, seeds are mainly preserved in a waterlogged state (e.g., lake sediments, peat). For example, remains found in peat indicate plants growing on the surface of the peat, while those found in lakes represent plants found within the wider landscape and the lake margin, and may provide valuable information on climate change or vegetation history. In contrast, remains found in archaeological archives provide essential information on the economy, diet, and daily life of the site's occupants, irrespective of whether they are hunters and gatherers or from agricultural communities. In the case of the latter, paleoeconomic, experimental, and ethnographic studies of Old World charred cereals and their associated components have made significant contributions to our understanding of the origin, spread, distribution, methods of cultivation, and processing techniques of these important food resources (e.g., Weiss and Kislev, 2004). Waterlogged seeds from archaeological deposits provide useful information on local flora, presence of standing water, animal husbandry, and the presence of cultivation (e.g., Ponel et al., 2000). For example, several lake-village sites in France and Switzerland have provided evidence for transportation to the site of plants exploited by humans for animal fodder, human food, and building materials (e.g., Rasmussen, 1993).

Palynology, the analysis of pollen grains and spores, is a widely used technique in environmental archaeology and provides valuable information on vegetation composition, structure and succession, plant migration, climate change, human modification of natural vegetation cover, and land use. They are highly resistant to decay and optimal conditions occur in acidic (low levels of microbial activity) or anaerobic (oxygen free), waterlogged archives, although pollen is also preserved in dry conditions, such as whole, desiccated coprolites (feces), and cave sediments. Identifying human impact is a key application of pollen analysis and requires the recognition of indicator species or groups of particular forms of land use, such as cultivation, pasture, and meadow, as well as deforestation and woodland management. For example, records showing a decline in tree pollen values or individual tree taxa, the presence of certain ruderal weed species strongly associated with cultivation or pastoralism, cultivated species of crops and taxa indicated managed woodland indicate a transition to agriculture, or the intensification of agriculture, in many world regions (Branch et al., 2005).

The presence of microscopic charred particles in archaeological and geological archives often complements pollen studies and provides useful information on natural wildfires (including climate history), human-induced woodland clearance, agricultural practices, and woodland management. For example, investigations of peat and lake sediments have frequently provided a rich source of information on human-induced woodland disturbance caused by burning (e.g., Moore, 2000). At North Gill, Yorkshire (UK), radiocarbon-dated fine pollen analysis combined with plant macrofossil and charred particle analysis has provided detailed information on woodland clearance (Innes and Blackford, 2003). Combined with fungal spore analysis, especially taxa that grow on the feces of domesticated animals (*Sporormiella*, *Cercophora*, *Coniochaeta*, and *Chaetomium*), the data have provided unequivocal evidence for linkages between woodland disturbance and forest grazing (Innes and Blackford, 2003). Macroscopic wood preserved by waterlogged conditions or burning is also often found in both archives, and provides primary data on woodland composition, and hence vegetation history, woodland management, agricultural practices (e.g., cultivation of plants for fodder and bedding for animals), woodland exploitation for domestic fires (fuel), and material culture (wooden artifacts) (e.g., Asouti, 2003).

Two further key categories of archaeobotanical remains epitomize the diverse range of methods available. Diatoms are unicellular algae with different species occupying the bottom of (benthic), or floating within (planktonic), water bodies. Preservation occurs in a wide range of archives of environmental archaeological interest, including lakes, alluvium, salt marshes, and ditches, and indicative of a variety of past environmental conditions (e.g., marine, brackish, or freshwater) that reflects prevailing temperature, salinity, pH, oxygen, and mineral content. For example, diatoms in lake sediments have provided valuable records of accelerated erosion due to deforestation during the last 6000 years, which in turn has caused changes in lake trophic (nutrient) status. Deep alluvial sequences in lowland river valleys (e.g., Lower Thames, UK) have also recorded changes in diatom assemblages due to fluctuations in the height of relative sea level during the Holocene (e.g., Battarbee, 1988). Finally, phytoliths are opaline silica bodies produced by plant cells and preserved in several archaeological archives, namely hearths and the sediment fills of caves, pits, and coprolites. Phytoliths may provide valuable information on past vegetation cover (e.g., the presence of meadow and pasture), the cultivation of crops (e.g., maize), and the presence of herbivores (e.g., within dung). For example, one of the most detailed recent studies of phytoliths has been concerned with the history of maize (*Zea mays*) domestication and cultivation in the New World, and especially extending the known range of prehistoric cultivation into southern parts of South America, e.g., southeastern Uruguay (Iriate, 2003).

Geochronology

To enable comparison of environmental archaeological data within and between archaeological and geological archives requires the application of dating methods that aim to provide both accurate and precise ages for the stratigraphy recorded, as well as establishing the duration of an event. By applying the universal law of superposition (a unit overlying another must have formed later), the *relative chronology* is often understood. Perhaps the most important of the so-called *absolute chronology* (precise age or date) techniques is radiocarbon dating (^{14}C; radiometric and accelerator mass spectrometry), but also optically stimulated luminescence, thermoluminescence, electron spin resonance, uranium–thorium ($^{234}U/^{230}Th$) and lead 210 (^{210}Pb) dating have all provided important chronological frameworks for

environmental archaeological evidence (e.g., Aitkin, 1990; *see* Chronology, Stratigraphy, and Dating Methods in Archaeology).

Research Themes

There are numerous excellent examples of where environmental archaeology has been enlightening, and readers are directed to the following key journals: *Environmental Archaeology*, *Journal of Archaeological Science*, *Geoarchaeology*, *Vegetation History and Archaeobotany*, *The Holocene*, and the following sample textbooks: Butzer, 1982; Branch et al., 2005. Case studies cover a broad range of themes and geographical areas, over a variety of spatial (local to global) and temporal scales, from the short term (<100 years) to the long term (>10 000 years). Collectively, site-specific data contribute to greater regional understanding, and occasionally, individual records provide something of unique, international importance. For example, peat cutting for fuel and horticulture has uncovered human bodies remarkably well preserved in the anaerobic, waterlogged conditions. These bodies provide a valuable insight into the daily lives of people in the past, and especially their diet, health, nutrition, and cause of death (see Stead et al., 1986). Data have emerged from geological archives that highlight the influence of climatic events over the past 10 000 years on socioeconomic developments in the Mediterranean (e.g., Drescher-Schneider et al., 2007; Berger and Guilaine, 2009; Mercuri et al., 2011), North and South America (e.g., deMenocal, 2001) and West Asia (e.g., Mithen and Black, 2011). For example, in northern Italy, climate change is considered to be a possible cause for initiation of terrace agriculture *c*.4200 years ago (Branch, 2013), while in northern Britain climate change may have caused the collapse of agricultural practices during the Bronze Age (Tipping et al., 2008). This theme, in particular, epitomizes one of the major ongoing challenges for environmental archaeology, namely understanding past human resilience and adaptation in the face of climate and environmental change.

A further illustration of a significant, long-term theme is the global origins and spread of agriculture (Thorpe, 1996; Harris, 2010). Environmental archaeological studies over several decades have demonstrated the existence of multiple centers of plant and animal domestication. While the timing of the transition to farming varied considerably, there is nonetheless unequivocal evidence in each location for morphological and genetic alteration of many of the food staples that remain important today (*see* Chronology, Stratigraphy, and Dating Methods in Archaeology). For example, the broad timing of plant domestication in West Asia (e.g., wheat, barley, rye, lentils, and peas) and Southeast Asia (e.g., millet and rice) was ~10 000 years ago (Kya), while in Australasia (e.g., yam and banana) and North Africa (e.g., millet and sorghum) it was ~7Kya and ~4–3Kya respectively. In North America (e.g., squash and sunflower) it was ~5Kya, Mesoamerica (e.g., squash and maize) ~10–8Kya, and South America (e.g., squash, arrowroot, yam, cotton, sweet potato, beans, peanut, manioc, chili pepper, potato, and quinoa) ~10–5Kya. The subsequent spread of agriculture across these geographical areas, traced by environmental archaeological data, has quantified the magnitude of impact on the natural environment, especially the creation of new ecosystems, due in many respects to increased human population levels and technological developments. This resulted, directly and indirectly, in animal extinctions, deforestation, soil degradation and erosion, afforestation, and the formation of mires, grasslands, and shrublands (heathlands and moorlands).

In Western Asia, arguably the most intensively studied region, agricultural origins have been linked to both climate change at the transition from the last glaciation to the present interglacial (Holocene), and a continuum of human interaction with plants and animals. In many areas, significant landscape modification accompanied the transition to farming, including woodland clearance and water management systems. During the subsequent spread of agriculture across Europe over ~6000 years, woodland clearance became increasingly apparent, complementing evidence for woodland exploitation for fuel, fodder, bedding, and construction. Although many records suggest that clearance was temporary ('shifting cultivation'), which permitted the woodland to recover, the general trend was progressive deforestation of mature coniferous (e.g., fir) and/or deciduous (e.g., elm, lime, oak) woodland, and the introduction of field systems and in some areas agricultural terracing. For example, the so-called 'elm (*Ulmus*) decline,' and later 'lime (*Tilia*) decline,' were two events linked to woodland clearance (e.g., 'slash and burn'), for both cereal cultivation and the harvesting of leaves, bark, and twigs for fodder and bedding. Both were part of a mixed economy involving animal husbandry and cultivation, and possible woodland management (pollarding). The subsequent development of grasslands and shrublands illustrate the long-term effects of degradation due to intensive and extensive human activities, with the development of Northern European heathlands, Mediterranean macchia/maquis, and lowland temperate grassland. The African savannah, American prairies, and Asiatic steppe have similarly been cited as plagioclimax communities (created and maintained by human activities) whose origins relate to sustained human intervention and management for many millennia.

The widespread impact of agriculture, in terms of environmental degradation, landscape modification, and societal resilience, therefore, remains a key global theme. In South America, for example, evidence for landscape modification is epitomized by the development of raised fields and terracing, with an estimated ~1 000 000 ha of terracing in Peru alone (see Branch et al., 2007). Terracing transformed the morphology of the landscape while also creating the basis for a highly innovative and sustainable agricultural system that primarily permitted the cultivation of maize and quinoa. Studies have permitted detailed investigations of the terrace structural organization and development, especially their irrigation and drainage. This work has revealed multiple phases of terrace construction during later prehistory, with evidence for possible soil enrichment through manuring and/or fallow periods, and localized landscape erosion triggered by the construction activities. Interestingly, there is evidence in several parts of Peru for the abandonment of terraces, such as those of the Wari civilization (AD 500–1000), and their subsequent reconstruction prior to the Colonial Period (AD 1533). While the persistence of terrace agriculture confirms the importance of

terraces in the agricultural system, the precise reason for their unsustainability remains unclear. In contrast, in Mesoamerica, the collapse of agriculturally intensive practices of raised field agriculture and terracing due to nutrient depletion and soil erosion has been strongly associated with the demise of the Maya civilization (Branch et al., 2005).

In summary, future environmental archaeological research will, therefore, focus on major methodological developments and key themes of regional and global significance. A primary research area will be improvements to the precision of key dating methods; this will ensure a better understanding of the timing and duration of major cultural and environmental changes, and will be crucial if we are to explain whether natural environmental change, such as climatic oscillations, caused social and economic change in the past, or if this was triggered by purely cultural factors. The origins and spread of agriculture, and the environmental impact of agriculture, will also continue to be major research areas, as will changing human diet, nutrition, and health; these linked themes permit better understanding of past food security, and the sustainability and resilience of societies in a changing world, which will be pertinent to present day concerns and debates surrounding global climate and environmental change.

See also: Archaeometry; Bioarchaeology; Chronology, Stratigraphy, and Dating Methods in Archaeology; Food Production, Origins of; Geoarchaeology; Hunting and Gathering Societies: Anthropology; Mesoamerica, Archaeology of; South America, Archaeology of; States and Civilizations, Archaeology of; Zooarchaeology.

Bibliography

Aitken, M.J., 1990. Science-based Dating in Archaeology. Longman, London.

Asouti, E., 2003. Woodland vegetation and fuel exploitation at the prehistoric campsite of Pinarbaşi, south-central Anatolia, Turkey: the evidence from the wood charcoal macro-remains. Journal of Archaeological Science 30, 1185–1201.

Atkinson, T.C., Briffa, K.R., Coope, G.R., Joachim, M.J., Perry, D.W., 1986. Climatic calibration of coleopteran data. In: Berglund, B.E. (Ed.), Handbook of Holocene Palaeoecology and Palaeohydrology. Wiley, Chichester, pp. 851–858.

Battarbee, R.W., 1988. The use of diatom analysis in archaeology: a review. Journal of Archaeological Science 15, 621–644.

Bell, M.G., 1983. Valley sediments as evidence of prehistoric land-use on the South Downs. Proceedings of the Prehistoric Society 49, 119–150.

Bell, M., Caseldine, A., Neumann, H., 2000. Prehistoric Intertidal Archaeology in the Welsh Severn Estuary. Council for British Archaeology Research Report 120, 269–270.

Bell, M., Walker, M.J.C., 2004. Late Quaternary Environmental Change. Longman, Harlow.

Berger, J.F., Guilaine, J., 2009. The 8200 cal BP abrupt environmental change and the Neolithic transition: a Mediterranean perspective. Quaternary International 200, 31–49.

Binford, L.R., 1968. Post Pleistocene adaptations. In: Binford, R., Binford, L.R. (Eds.), New Perspectives in Archaeology. Aldine Press, Chicago, pp. 313–341.

Branch, N.P., Kemp, R.A., Silva, B., Meddens, F.M., Williams, A., Kendall, A., Pomacanchari, C.V., 2007. Testing the sustainability and sensitivity to climatic change of terrace agricultural systems in the Peruvian Andes: a pilot study. Journal of Archaeological Science 34, 1–9.

Branch, N.P., 2013. Early–Middle Holocene vegetation history, climate change and human activities at Lago Riane (Ligurian Apennines, NW Italy). Vegetation History and Archaeobotany 22, 315–334.

Branch, N.P., Canti, M.G., Clark, P., Turney, C.S.M., 2005. Environmental Archaeology: Theoretical and Practical Approaches. Edward Arnold, London.

Branch, N.P., Batchelor, C.R., Cameron, N.G., Coope, G.R., Densem, R., Gale, R., Green, C.P., Williams, A.N., 2012. Holocene environmental changes in the Lower Thames Valley, London, UK: implications for our understanding of the history of Taxus (L.) woodland. The Holocene 22, 1143–1158.

Brothwell, D.R., Higgs, E. (Eds.), 1969. Science in Archaeology. Thames and Hudson, London.

Brown, A.G., 2003. Time, space and causality in floodplain palaeoecology. In: Howard, A., Macklin, M.G., Passmore, D.G. (Eds.), Alluvial Archaeology in Europe. A.A. Balkema, Lisse, pp. 15–24.

Butzer, K.W., 1982. Archaeology as Human Ecology. Cambridge University Press, Cambridge.

Cabral, J.P., Silva, da A.C.F., 2003. Morphometric analysis of limpets from an Iron-Age shell midden found in northwest Portugal. Journal of Archaeological Science 30, 817–829.

Carrott, J., Kenward, H., 2001. Species associations among insect remains from urban archaeological deposits and their significance in reconstructing the past human environment. Journal of Archaeological Science 28, 887–905.

Clarke, D.L., 1978. Analytical Archaeology. Methuen, London.

Dearing, J., 1999. Magnetic susceptibility. In: Walden, J., Oldfield, F., Smith, J. (Eds.), Environmental Magnetism: A Practical Guide. Quaternary Research Association Technical Guide 6, London, pp. 35–62.

Dincauze, D.F., 2000. Environmental Archaeology: Principles and Practice. Cambridge University Press, Cambridge.

Drescher-Schneider, R., de Beaulieu, J.-L., Magny, M., Walter-Simonnet, A.-V., Bossuet, G., Millet, L., Brugiapaglia, E., Drescher, A., 2007. Vegetation history, climate and human impact over the last 15 000 years at Lago dell'Accesa (Tuscany, Central Italy). Vegetation History and Archaeobotany 16, 279–299.

Flannery, K.V., 1968. Archaeological systems theory and early Mesoamerica. In: Meggers, B.J. (Ed.), Anthropological Archaeology in the Americas. Anthropological Society of Washington, Washington, pp. 67–87.

Gale, S.J., Hoare, P.G., 1991. Quaternary Sediments: Petrographic Methods for the Study of Unlithified Rocks. Belhaven Press, London.

Halkett, D., Hart, T., Yates, R., Volman, T.P., Parkington, J.E., Orton, J., Klein, R.G., Cruz-Uribe, K., Avery, G., 2003. First excavation of intact Middle Stone Age layers at Ysterfontein, Western Cape Province, South Africa: implications for Middle Stone Age ecology. Journal of Archaeological Science 30, 955–971.

Harris, D.R., 2010. Origins of Agriculture in Western Central Asia: An Environmental-Archaeological Study. University of Pennsylvania Museum of Archaeology and Anthropology, Pennsylvania.

Hellqvist, M., Lemdahl, G., 1996. Insect assemblages and local environment in the Mediaeval town of Uppsala, Sweden. Journal of Archaeological Science 23, 873–881.

Hosfield, R., 2011. The British Lower Palaeolithic of the early Middle Pleistocene. Quaternary Science Reviews 30, 1486–1510.

Innes, J.B., Blackford, J.J., 2003. The ecology of Late Mesolithic woodland disturbances: model testing with fungal spore assemblage data. Journal of Archaeological Science 30, 185–194.

Iriate, J., 2003. Assessing the feasibility of identifying maize through the analysis of cross-shaped size and three-dimensional morphology of phytoliths in the grasslands of southeastern South America. Journal of Archaeological Science 30, 1085–1094.

Jeppesen, E., Leavitt, P., Meester De, L., Jensen, J.P., 2001. Functional ecology and palaeolimnology: using cladoceran remains to reconstruct anthropogenic impact. Trends in Ecology and Evolution 16, 191–198.

Kemp, R.A., Branch, N.P., Silva, B., Meddens, F., Williams, A., Kendall, A., Pomacanchari, C.V., 2006. Pedosedimentary, cultural and environmental significance of palaeosols within Pre-Hispanic agricultural terraces in the southern Peruvian Andes. Quaternary International 158, 13–22.

Mannino, M.A., Thomas, K.D., 2001. Intensive Mesolithic exploitation of coastal resources? Evidence from a shell deposit on the Isle of Portland (southern England) for the impact of human foraging on populations of intertidal rocky shore molluscs. Journal of Archaeological Science 28, 1101–1114.

Matthews, W., 2003. Microstratigraphic sequences: indications of uses and concepts of space. In: Matthews, R.J. (Ed.), Excavations at Tell Brak. vol. 4. Exploring an Upper Mesopotamian Regional Centre, 1994-96. McDonald Institute for Archaeological Research and British School of Archaeology in Iraq, Cambridge, pp. 377–388.

deMenocal, P.B., 2001. Cultural responses to climate change during the Late Holocene. Science 292, 667–673.

Mercuri, A.M., Sadori, L., Uzquiano Ollero, P., 2011. Mediterranean and north-African cultural adaptations to mid-Holocene environmental and climatic changes. The Holocene 21, 189–206.

Mithen, S., Black, E. (Eds.), 2011. Water, Life and Civilisation: Climate, Environment and Society in the Jordan Valley. International Hydrology Series. Cambridge University Press, Cambridge.

Moore, J., 2000. Forest fire and human interaction in the early Holocene woodlands of Britain. Palaeoecology, Palaeoclimatology. Palaeoecology 164, 125–137.

O'Connor, T., Evans, J.G., 2005. Environmental Archaeology: Principles and Methods. Sutton Publishing, Stroud.

Panagiotakopulu, E., 2001. New records for ancient pests: archaeoentomology in Egypt. Journal of Archaeological Science 28, 1235–1246.

Parnell, J.J., Terry, R.E., Nelson, Z., 2002. Soil chemical analysis applied as an interpretive tool for ancient human activities in Piedras Negras, Guatemala. Journal of Archaeological Science 23, 379–404.

Ponel, P., Matterne, V., Coulthard, N., Yvinec, J.-H., 2000. La Tene and Gallo-Roman natural environments and human impact at the Touffréville rural settlement, reconstructed from coleopteran and plant macroremains (Calvados, France). Journal of Archaeological Science 27, 1055–1072.

Rasmussen, P., 1993. Analysis of goat/sheep faeces from Egolzwil 3, Switzerland: evidence for branch and twig foddering of livestock in the Neolithic. Journal of Archaeological Science 20, 479–502.

Renfrew, C., Bahn, P., 2000. Archaeology: Theory, Methods and Practice, third ed. Thames and Hudson, London.

Rick, T.C., Erlandson, J.M., 2000. Early Holocene fishing strategies on the California coast: evidence from CA-SBA-2057. Journal of Archaeological Science 27, 621–633.

Riddiford, N.G., Branch, N.P., Green, C.P., Armitage, S.J., Olivier, L., 2012. Holocene palaeoenvironmental history and the impact of prehistoric salt production in the Seille Valley, Eastern France. The Holocene 22, 831–845.

Sigurðsson, H., Standford, C., Sparks, S.R.J., 1982. The eruption of Vesuvius in AD 79: reconstruction from historical and volcanological evidence. American Journal of Archaeology 86, 39–51.

Stead, I.M., Bourke, J.B., Brothwell, D., 1986. Lindow Man the Body in the Bog. British Museum Publications, London.

Szeroczyńska, K., 2002. Human impact on lakes recorded in the remains of Cladocera (Crustacea). Quaternary International 95-96, 165–174.

Thompson, R., Oldfield, F., 1986. Environmental Magnetism. Allen and Unwin, London.

Thorpe, I.J., 1996. The Origins of Agriculture in Europe. Routledge, London.

Tipping, R., Davies, A., McCulloch, R., Tisdall, E., 2008. Response to late Bronze Age climate change of farming communities in north east Scotland. Journal of Archaeological Science 37, 2379–2386.

Weiss, E., Kislev, M.E., 2004. Plant remains as indicators of economic activity: a case study from Iron Age Ashkelon. Journal of Archaeological Science 31, 1–13.

Wilkinson, K.N., 2003. Colluvial deposits in dry valleys of southern England as proxy indicators of paleoenvironmental and land-use change. Geoarchaeology 18, 725–755.

Environmental Attitudes and Behavior: Measurement

Sebastian Bamberg, Bielefeld University of Applied Sciences, Bielefeld, Germany
Jonas Rees, Bielefeld University of Applied Sciences, Bielefeld, Germany; and Bielefeld University, Bielefeld, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The conservation of our planet's natural ecosystems and resources is one of the greatest challenges of the twenty-first century. As most environmental problems are caused by human activities, effective environmental policies depend on reliable and valid information about these behavioral patterns and their determinants. For this reason, the theoretical background, internal logic as well as (dis)advantages of two dominant strategies for assessing environmentally relevant behaviors are presented: the intent- and impact-oriented measurement approaches. We conclude by discussing the measurement of environmental concern and empirical findings regarding the environmental concern–pro-environmental behavior relation.

Introduction

John lives in a semidetached house in a small village just outside a medium-sized town. He owns a small car, but he tries to use his bike as much as possible. In his garden, he has a compost bin, and he makes sure that he always separates his glass, chemical, and paper waste. Usually, he buys organic food, and he does not eat meat more than twice a week. He recently replaced his old washing machine with a highly efficient one. He actively participates in demonstrations against the planned enlargement of a nearby airport and for this reason has voted for the Green party in the last elections. Because of his interest in African landscape and culture, in his holidays he frequently travels to this continent.

Mark lives in a small flat in the town center. He would like to own a car, but because parking space is rare, he does not buy one. Mark throws all his waste into the same bin. He prefers to buy frozen food, which he heats up in his microwave oven. Mark is not interested in politics and did not vote in the last elections. He hates traveling and uses his holidays to stay at home meeting friends in local pubs.

Who would you believe to be more environmentally friendly, John or Mark? Probably most people would label John 'more pro-environmentally' because he consciously tries to perform pro-environmental behaviors. Mark obviously does not care much about the environment.

However, from an objective impact-oriented point of view, Mark's household behavior may be less environmentally damaging because he does not own a car, lives in a small flat that requires less energy to heat, and most importantly, does not travel long distances by plane.

The example, adapted from Gatersleben et al. (2002), not only illustrates the complexities associated with the analysis of environmentally related behaviors but simultaneously reflects the two strategies we can find in the literature regarding the measurement of these behaviors: The intent-oriented approach uses the actor's standpoint for defining pro-environmental behavior. An action is viewed as pro-environmental behavior if it is taken with the intention to benefit the environment. It stresses an actor's subjective motivation as a defining characteristic of pro-environmental behavior.

In contrast, the impact-oriented approach associates all human activities with more or less negative environmental impact (e.g., energy and water use, waste production). As a consequence, the question is not whether a specific human action is 'pro-environmental' by intent but how serious its objective environmental impact is. Thus impact-oriented measures seek to identify target behaviors that significantly influence the environment and examine how the environmental impact of behavior patterns can be reduced.

Types of Environmentally Relevant Behavior

The introductory example also shows the multitude of behavioral domains where a person can articulate his or her intention to protect the environment. For this reason, Stern (2000) suggests to differentiate three main types of pro-environmental behavior: (1) pro-environmental activism (e.g., active involvement in environmental organizations and demonstrations), (2) nonactivist behaviors in the public sphere (e.g., petitioning on environmental issues, joining or contributing to environmental organizations, and supporting public policies), and (3) private-sphere environmentalism. Because of its immediate impact on the environment, researchers have focused mainly on the last behavioral type: the purchase, use, and disposal of personal and household products. It is useful to further subdivide these private-sphere behaviors according to the type of decision they involve: (1) the purchase of major household goods and services (e.g., cars, houses), (2) the use and maintenance of environmentally important goods (e.g., heating and cooling systems), (3) household waste disposal (e.g., composting), and (4) 'green' consumerism (e.g., purchasing organic food). Research based on such distinctions has revealed that the purchase of cars and major household appliances tends to have much greater environmental impact than changes in the level of use of the same equipment.

Led by the assumption that all these behavioral types reflect one common latent pro-environmental behavioral tendency, researchers for a long time expected a positive correlation between these different types of pro-environmental behavior. However, empirical research does not support this assumption. For example, Bratt (1999) found only weak, sometimes even

negative, correlations between different pro-environmental consumer activities. He concludes that there is no 'general' pro-environmental behavior: A person can behave very pro-environmentally in one behavioral domain (e.g., recycling) and very environmentally damaging in another (e.g., mobility). This behavioral heterogeneity and inconsistency has stimulated research. For example, Diekmann and Preisendörfer (1998) identified cognitive strategies that help individuals to harmonize and reconcile their seemingly incongruent behaviors. They called these strategies 'attention-shifting strategy' and 'low-cost strategy.' When attention shifting, people tend to emphasize the importance of some activities while simultaneously playing down the importance of other activities. The basis of the low-cost strategy is that "compliance to 'environmental morality' and 'ecological insight' in low-cost situations can be qualified as a cognitive mechanism to harmonize environmental attitudes and behavior" (Diekmann and Preisendörfer, 1998, p. 89). According to this logic, pro-environmental behaviors tend to be consonant with each other where the cost of pro-environment behavior is low. Bratt (1999) argues that people experiencing cognitive dissonance because of environmentally harmful behaviors that are difficult or costly to change (e.g., taking flights, eating meat) may engage in 'simple and painless' pro-environmental behavior as a means to relieve the discomfort created by these harmful behaviors. Holland et al. (2002) assume that this type of self-justification is based on the general acceptance of a 'contribution ethic.' This contribution ethic implies that neglecting a specific pro-environmental behavior is justified if one is already 'playing one's part' in other ways – leading one to 'rest on one's laurels.' For example, Thøgersen (1999) found that whenever people felt less obliged to do one specific thing (e.g., consider the waste consequences of their packaging choices when shopping), they often did something else to address the problem (i.e., sort household waste for recycling).

Measuring Private-Sphere Environmentalism – The Intent-Oriented Approach

Typically, intent-oriented private-sphere environmentalism measures incorporate behaviors that are seen as prototypical indicators of a person's latent motivation to act in a pro-environmental way. Respondents are provided with a list of such behaviors and asked to indicate how often (e.g., on a scale ranging from 'never' to 'always') they perform each of these behaviors. While some studies focus on one specific behavior, such as recycling (e.g., Schultz and Oskamp, 1996) or mobility (e.g., Van Lange et al., 1998), others deploy scales that combine different types of behavior (e.g., Kaiser, 1998).

A measure developed by Preisendörfer (1998) for the national survey 'Environmental Concern in Germany' provides a good demonstration of the logic underlying the intent-oriented approach as well as the results typically obtained within this approach. In this survey, face-to-face interviews with 2307 persons randomly drawn from the German population were conducted. Environmentally relevant behaviors were assessed using a list of 16 questionnaire items distributed across four subscales (**Table 1**): (1) four items assessed recycling behavior with respect to different materials, (2) four items assessed different aspects of pro-environmental shopping behavior, (3) four items assessed energy- and water-saving behavior, and (4) four items assessed aspects of mobility and transportation behavior. The results of the statistical analyses provide empirical evidence that the item sets used for assessing the behavior dimensions recycling, shopping behavior, and mobility/transportation behavior form homogeneous subscales. However, the statistical analyses indicate that the four items of the subscale energy- and water-saving behavior do not form one homogenous dimension, but two independent dimensions.

Table 1 Preisendörfer's (1998) intent-oriented 16-item pro-environmental behavior scale

Recycling behavior	*Shopping behavior*
Recycling of paper (v28a)	Buying products with eco-label (v12)
Recycling of glass (v28b)	Buying products with less packaging (v19)
Recycling of aluminum (v28e)	Buying seasonal fruits/regional vegetables (v20)
Recycling of batteries (v28f)	Buying organic food (v21)
Energy- and water-saving behavior	*Mobility and transportation*
Use of energy-saving bulbs (v36)	No car in household (v49d)
Reducing water-consumption under shower (v39)	Weekend trips without car (v66)
Use of water-saving installations (v40)	Last holiday trip without car/airplane (v68)
Turning down heating when going out (v44)	Shopping without car (v70)

Adapted from Jann, B.R.A., 1998. Zur Messung des Umweltverhaltens (Unpublished Master Thesis). Department of Sociology, University of Bern, Swiss, p. 21.

As discussed above, this study also suggests that participants' responses to the four subscales should be viewed as representing distinct behavioral tendencies as opposed to only one single homogeneous behavioral tendency: The intercorrelations of the four subscales are not only weak, but the mobility subscale actually correlates negatively with the other three subscales. A smallest-space analysis conducted by Jann (1998) with the intercorrelations of all 16 items as input offers detailed insights into the structure underlying participants' response patterns. **Figure 1** shows the spatial structure of the items: the mobility items are positioned at the left edge of the figure, the shopping and recycling items at the right edge. Furthermore, the items form a circle, however, one with significant gaps. With the low-cost hypothesis (Diekmann and Preisendörfer, 1998) as theoretical background, Jann (1998) interprets the circular structure as representing the costs associated with the performance of behaviors assessed by the four subscales. Consequently, the circle leads from situations with low behavioral costs (recycling) to situations with high behavioral costs (mobility) with energy saving in between.

The intent-oriented approach of measuring pro-environmental behavior has been criticized mainly for two reasons. One critical point is its reliance on self-reported behavior, which may not be an accurate measure of actual behavior. Instead, self-reported behavior tends to reflect people's perceptions or beliefs about their behavior, at least to some degree. Factors such as lack of knowledge, social

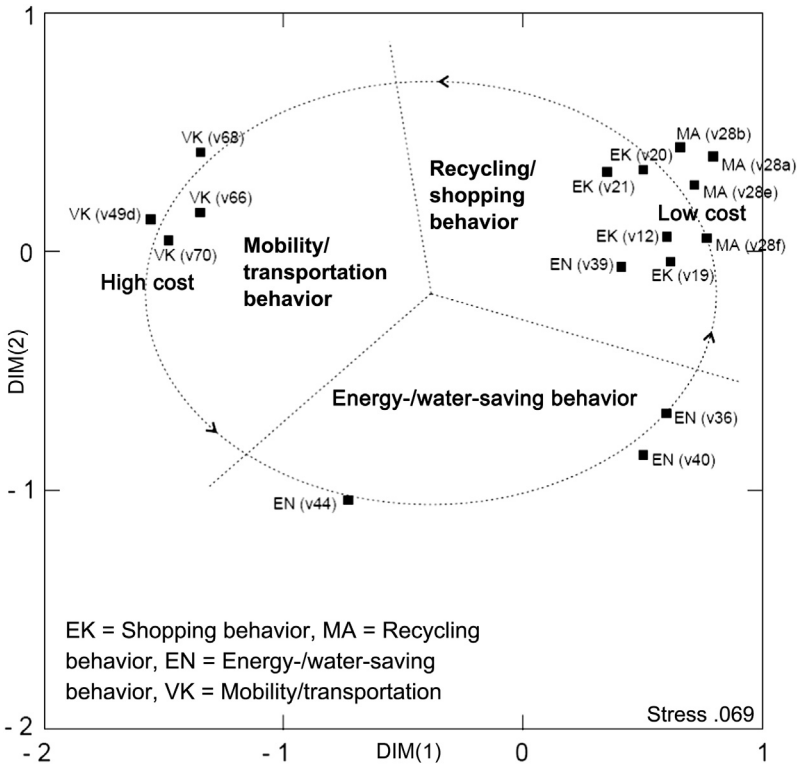

Figure 1 Two-dimensional smallest-space analysis plot of the intercorrelations of Preisendörfer's 16-item pro-environmental behavior scale. Adapted from Jann, B.R.A., 1998. Zur Messung des Umweltverhaltens (Unpublished Master Thesis). Department of Sociology, University of Bern, Swiss, p. 22.

desirability, and other types of (conscious or unconscious) response bias may result in inaccurate reports of actual behavior. For example, in Preisendörfer's (1998) survey, 24% of the participants reported that they always or often bought organically produced food. But at that time the share of organically produced food on the total food production in Germany was only about 2%, which questions the reliability and validity of participants' self-reported ecobehavior. Several studies, however, suggest that the difference between self-reported behavior and actual behavior is not systematic (Fuijii et al., 1985). Moreover, Kaiser et al. (1999) could show that people were only marginally tempted to give socially desirable answers in a study on pro-environmental behavior.

Another critique (e.g., Stern, 2000) relates to the way scales of pro-environmental behavior are usually constructed. Intent-oriented scales often include behaviors (e.g., refusing plastic bags in shops or buying recycled paper) that do not reflect the actual environmental impact of persons or households. As a consequence, respondents reporting a large number of these insignificant actions often receive a high total pro-environmental score, although such actions may only have a marginally positive impact on the environment.

Measuring Private-Sphere Environmentalism – The Impact-Oriented Approach

The impact-oriented approach is based on an analysis of the actual environmental load of household consumption activities. For this purpose, elaborated input–output energy analysis and life-cycle assessment methods are used to calculate the objective environmental load of household consumption activities, mainly in energy and/or greenhouse gas emission terms (for an overview of these methods, see Kok et al., 2006). The aim of these analyses is to both calculate average per capita energy use or CO_2 emissions and identify the consumption activities most significantly contributing to this total amount. For Germany, for example, this kind of analyses indicates an average yearly per capita output of approximately 11 t of CO_2 (UBA, 2007). On average, 25% of these CO_2 emissions come from heating and electricity, 23% from transportation, and 14% from food. That leaves 28% for other consumption and 10% for public infrastructure. The activities that most strongly affect a household's total CO_2 emissions are transportation – distances traveled by car and plane – and housing – size of living area and insulation standard. Food consumption also has a significant impact on CO_2 emissions, in particular the quantity of meat consumed and the purchase of organically produced food.

Based on these analyses, impact-oriented survey questionnaires cover the consumption activities (purchase and use) most relevant for a household's total CO_2 emissions. Work by Gatersleben et al. (2002) provides a good example of this strategy. In two large field studies among randomly selected Dutch households ($Ns = 2167$ and 1250), respondents were asked which household goods they possess (e.g., tumble driers, dishwashers, cars) and how often they use them per day, month, and year. Responses were weighted by estimates of

the average energy use needed for such behaviors. Using this measure, in the first study the average total energy use of respondents' households was estimated as 140 GJ, which is 61% of the mean total energy use of 229 GJ indicated by the official 1995 Dutch energy statistics. In the second study, the estimated average total energy use was 152.4 GJ, which is 67% of the actual total household energy use.

In another study, Gatersleben (2001) reports a more detailed analysis of the external validity of this measurement approach. For this purpose, 393 Dutch participants completed a computer-based questionnaire asking a total of 72 questions about their possession and use of household goods. Respondents were also asked for their annual gas and electricity use as indicated on their annual bill. Only 252 respondents (64%) were able to respond to this question. A correlation of $r = 0.45$ was found between gas use estimated by the 72 questions about their possession and use of household goods and the actual gas use reported in the bill, which means that 20% of the variance in actual gas use was measured by the indicator variables assessed in the computer-based questionnaire. Actual and estimated electricity use correlated $r = 0.51$, which means that 26% of the variance in actual electricity use was measured. Apparently, even the long list of 72 household consumption activities did not cover all the activities significantly contributing to a household's total energy use.

From the practitioner's perspective, another development of the impact-oriented measurement approach is interesting: the development of Internet-based personal CO_2 calculators. The German Federal Environment Agency (http://uba.klima-aktiv.de/) provides a German version of such a personal CO_2 calculator, Warwick University (http://carboncalculator.co.uk/users.php) offers a British one, and the Environmental Protection Agency (http://www.epa.gov/climatechange/ghgemissions/ind-calculator.html) has developed a calculator for the US. These calculators ask people to enter data regarding their household's energy consumption (heating and electricity) and transport as well as questions about their personal food consumption. All calculators report tons of CO_2 emitted per year by these household activities as the central outcome.

Summing up, how can we evaluate the impact-oriented approach of measuring pro-environmental behavior? Especially in the form of Internet-based individual CO_2 calculators, this approach provides a valuable instrument to examine the environmental impact of different household activities. For citizens and policy makers, the integration of knowledge from the natural and social environmental sciences results in more practically relevant information. By using dependent measures such as energy use in gigajoules or CO_2 emissions, the effects of current household activities as well as changes of these activities can be expressed in quantifiable units that have a meaning for policy makers and citizens.

One clear disadvantage of the impact-oriented approach is its high complexity. The information needed for calculating a household's total energy consumption requires elaborate and detailed questionnaires. Sometimes, people cannot supply the requested information. For example, many people are not able to report their annual consumption of natural gas in cubic meters or electricity in kilowatt-hour. The same is true for the reporting of travel and nutrition habits. As a consequence, many CO_2 calculators use raw national average estimates as substitutes. Furthermore, valid and sufficient indicators for assessing the indirect energy use resulting from the important 'other consumption' domain are currently missing. Other problems relate to the conversion factors used to calculate the energy use and CO_2 emissions resulting from the reported household activities. These conversion factors depend on steadily changing parameters such as the national or regional electricity mix (use of fossil, nuclear, or renewable energy sources for electricity production) or the energy efficiency of technology available to households. The results of different CO_2 emission calculators can thus vary considerably for the same input data, due to differences in calculating methodologies, behavioral estimates, or conversion factors (Padgett et al., 2008).

The Relation between Intent- and Impact-Oriented Measures

In the current literature, only very few studies explicitly analyze the empirical relationship of the two described strategies for measuring pro-environmental behavior. Two studies by Gatersleben et al. (2002) are rare exceptions. In their first study, they correlate two self-report measures of pro-environmental behavior (purchase of organic or regional food products and recycling behavior) with their estimate of household total energy use. The authors find a negative correlation between the estimated total household energy use and self-reported recycling behavior ($r = -0.24$, $p < 0.01$). However, there is no correlation between the purchase of environmentally benign food products and the estimated total household energy consumption ($r = -0.04$). In their second study, self-reported pro-environmental behavior was assessed via a scale consisting of 13 behaviors (e.g., turn down thermostat, travel short distances by bike, eat vegetarian food). In this second study, the correlation between the self-reported performance of the 13 pro-environmental behaviors and the estimated total household energy use was $r = 0.22$ ($p < 0.001$). This implies a significant but weak relation between self-reported frequency of pro-environmental behaviors and the household's total energy consumption. However, there may be a methodological reason for this weak relation: Whereas total energy use was measured on the level of the household, pro-environmental behaviors were measured on an individual level. Thus, it is not clear how much the total household energy consumption reflects the behavior of the respondent or that of other household members. It would be worthwhile to conduct studies in which self-reported pro-environmental behavior is also measured on the household level by collecting data on pro-environmental behaviors from all household members individually. This may then result in a closer correspondence of both measures.

Environmental Concern

In the 1970s, the notion that environmental problems are the consequences of 'maladaptive human behavior' (Maloney and Ward, 1973) motivated social scientists to analyze individual motives underlying this behavior. One central

assumption of this early research was that environmentally harmful behaviors reflect a person's low degree of environmental concern. Since then, most researchers conceptualize environmental concern as a general attitude, which comprises cognitive, affective, and conative judgments concerning the object of environmental protection (e.g., Weigel and Weigel, 1978). A measurement scale developed by Preisendörfer (1998) for the survey 'Environmental Concern in Germany' discussed previously exemplifies this approach. Based on substantive considerations and results of explorative factor analyses, this environmental concern scale consists of the nine statements presented in **Table 2** covering the cognitive (insight into the endangerment), the affective (feelings of fear, anger), and the conative component (willingness to do something). Preisendörfer constructed an additive index of environmental concern and standardized the range of this index to 0–20. The mean of the index in Preisendörfer's study was 13.1 and Cronbach's alpha was 0.72. Preisendörfer (1998) reports that the correlation between this environmental concern scale and the 16-item index of pro-environmental behavior discussed above is only $r = 0.22$. This low correlation fits with the results of meta-analyses summarizing the results for a great number of studies testing empirically the relationship between environmental concern and specific pro-environmental behaviors. In most of these studies, intent-oriented scales were used for assessing pro-environmental behavior. In their meta-analysis, Hines et al. (1986) report an average correlation of $r = 0.35$ (128 studies); Eckes and Six (1994) report an average correlation of $r = 0.26$ (17 studies). Taken together, the general attitude regarding environmental concern seems to explain not more than 10% of the variance of specific environmental behaviors.

Table 2 Preisendörfer's nine-item environmental concern scale

	−	−/+	++
Affective Dimension			
I am afraid when I think about environmental conditions for future generations.	8	18	74
If we continue our current style of living, we are approaching an environmental catastrophe.	11	23	66
Watching TV or reading in the newspaper about environmental problems, I am often embarrassed and angry.	11	26	63
Cognitive Dimension			
The great majority of German people do not act in an environmental responsible way.	13	29	58
There are limits of economic growth, which the industrialized world has already reached or will reach very soon.	10	34	56
* In my opinion, environmental problems are greatly exaggerated by proponents of the environmental movement.	54	27	19
Conative Dimension			
It is still true that politicians do much too little to protect the environment.	8	26	66
To protect the environment, we should be willing to reduce our current standard of living.	11	35	54
Environmental protection measures should be carried out, even if it reduces the number of jobs in the economy.	36	37	27

Note: Five-point response scale: − strongly disagree/disagree, −/+ partially disagree/partially agree, ++ strongly agree/agree; numbers related proportion of respondents in percent; * disagreement is seen as indicating higher environmental concern.
Adapted from Diekmann, A., Preisendörfer, P., 2003. Green and greenback: the behavioral effects of environmental attitudes in low-cost and high-cost situations. Rationality and Society 15, p. 455.

There are various theoretical approaches that have been developed for explaining the low relation between general environmental concern and specific pro-environmental behaviors. For example, Guagnano et al. (1995) introduced the A-B-C model, which combines insights from social psychological attitude theory and sociological rational choice theory. This model predicts an interaction effect "… where the main effect of attitudes [A] on behavior [B] depends on external conditions [C], with the effect of attitude approaching zero when external conditions are very strongly positive or negative" (p. 704). The A-B-C model was empirically supported in a field experiment with curbside recycling. The low-cost hypothesis by Diekmann and Preisendörfer (1998, 2003) states "… that environmental concern influences ecological behavior primarily in situations and under conditions connected with low costs and little inconvenience for individual actors. The lower the pressure of costs in a situation, the easier it is for actors to transform their attitudes into corresponding behavior" (p. 443). Diekmann and Preisendörfer (2003) provide evidence for the low-cost hypothesis using survey data from Germany. Their findings show that environmental concern has a stronger effect on low-cost behavior (such as recycling or buying products with eco-labels) than on high-cost behavior (such as shopping or going on holiday without a car).

According to Bamberg (2003), the weak direct relationship between general environmental concern and specific environmental behaviors is due to an inadequate understanding of how general attitudes influence specific behaviors. Because it is a situation-invariant general 'orientation pattern,' environmental concern cannot influence specific behaviors directly. Only the situation-specific cognitions concerning the consequences associated with a specific behavior are direct determinants of this behavior. Bamberg consequently suggests a framing model of the environmental concern–environmental behavior relationship (for a similar idea, see Lindenberg and Steg, 2007): This framing model assumes that as a general attitude, environmental concern influences specific behaviors only indirectly. More precisely, environmental concern influences a person's 'definition of the situation' that is how she or he frames a decisional problem, the relevant behavioral alternatives, and the personally important consequences associated with these alternatives. **Figure 2** presents the results of a study by Bamberg (2003) empirically supporting the framing model: Situation-specific cognition, conceptualized via Ajzen's (1991) theory of planned behavior, is a strong determinant of a specific pro-environmental behavior (actually requesting an information brochure about 'green' electricity products). After taking into account this situation-specific cognition, general environmental concern no longer has any direct effect on intention or behavior. But as expected, environmental concern has a substantive direct effect on the perception and evaluation of the situation-specific cognition, especially the personally salient behavioral consequences.

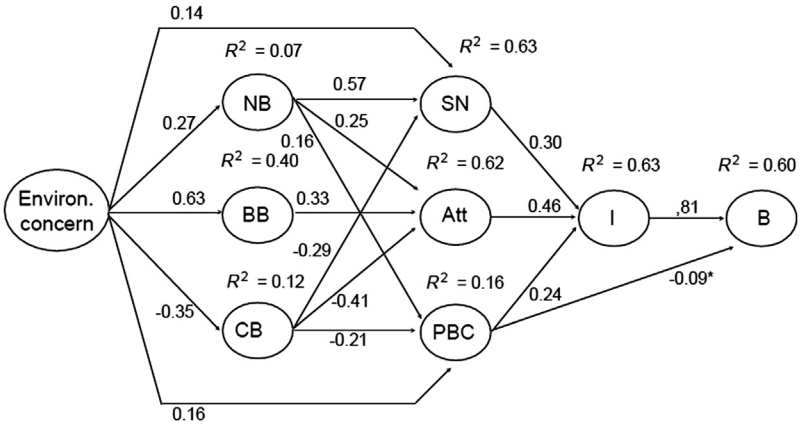

Figure 2 Behavior-specific constructs mediating the environmental concern–behavior relation. NB = normative beliefs; BB = behavioral beliefs; CB = control beliefs; Att = attitude toward the behavior; SN = subjective norm; PBC = perceived behavioral control; I = intention; B = behavior; * = not significant. Adapted from Bamberg, S., 2003. How does environmental concern influence specific environmentally related behaviors? A new answer to an old question. Journal of Environmental Psychology 23, 21–32; Note: Results are based on structural equation modeling (standardized path coefficients and explained variances); abbreviations relate to constructs from the theory of planned behavior (Ajzen, I., 1991. The theory of planned behavior. Organizational Behavior and Human Decision Process 50, 179–211.).

Conclusion

The ultimate goal of most studies on pro-environmental behavior is to provide information that can be helpful in reducing the negative environmental impact of human activities. In this context, it is important to be aware whether these studies use an intent- or impact-oriented approach. Intent-oriented measures focus on behaviors that are environmentally significant from an actor's point of view based on societal notions of environmentally significant behaviors. With these measures, what respondents deliberately do to benefit the environment, why some people act more pro-environmentally than others, and what can be done to persuade people to behave more pro-environmentally, for instance, could be examined. However, currently used intent-oriented measures tend to neglect behavior patterns with a strong objective environmental impact. Identifying the human activities with the strongest environmental impact is the focus of the impact-oriented measurement strategy. Here, knowledge from the natural and social sciences is integrated, more adequately reflecting the nature of most environmental problems. Furthermore, by using dependent measures such as CO_2 emissions, the impact-oriented approach usually yields more meaningful and thus more practically relevant information for citizens and policy makers. Impact-oriented measures are also more suitable for understanding how a household's total energy use is influenced by contextual factors (e.g., available technologies, infrastructures) whose impact is often hidden and outside of an individual actor's awareness or influence. Thus, impact-oriented measures are usually the better choice for studies interested in investigating how current living styles support or prevent sustainable behavior. The strategy used for assessing environmentally relevant behavioral patterns may also influence which predictors a study empirically identifies as the most important (Gatersleben et al., 2002). Using an intent-oriented measure, a study may find that psychological variables (e.g., perceptions, attitudes, values) are the most important determinants. Using an impact-oriented measure, a study may find that structural variables (e.g., income, type of car, house size) are most important. Research results regarding the environmental concern–pro-environmental behavior relation also underline the importance of these infra- or contextual factors. In the meantime, there seems to be general agreement to treat environmental concern as a general decisional precondition for considering the potential environmental impact of decisions rather than a direct predictor of actual behavior. Whether or not the more pro-environmental option is actually chosen depends on its subjectively perceived behavioral consequences (e.g., costs or inconvenience).

In closing, we want to mention that from our point of view, one central task of interdisciplinary environmental research will be to develop measures that integrate the advantages of the intent- and impact-oriented approaches. For this purpose, a list of reliable indicators would have to be developed that covers a substantial amount of variance in the actual environmental impact of a household without resulting in a questionnaire that is too long or demanding for the respondents.

See also: Attitude Measurement; Attitudes and Behavior; Climate Change and Global Poverty; Climate Change; Consumer Psychology; Data Bases and Statistical Systems: Environment and Urban Studies; Energy Flow Analysis; Energy and Society; Environmental Cognition, Perception, and Attitudes; Environmental Conservation and Development: Critical Perspectives; Environmental Economics; Environmental Planning; Environmental Sociology; Factor Analysis and Latent Variable Models in Personality Psychology; Global Public Policy; Life-style, History of the Concept; Material Flow Analysis; Political Ecology: Anthropological Perspectives; Sustainability and Sustainability Science; Sustainable Lifestyles; Sustainable Production and Consumption; Transportation, Psychology of.

Bibliography

Ajzen, I., 1991. The theory of planned behavior. Organizational Behavior and Human Decision Process 50, 179–211.

Bamberg, S., 2003. How does environmental concern influence specific environmentally related behaviors? A new answer to an old question. Journal of Environmental Psychology 23, 21–32.

Bratt, C., 1999. Consumers' environmental behavior: generalized, sector-based, or compensatory? Environment and Behavior 31, 28–44.

Diekmann, A., Preisendörfer, P., 1998. Environmental behavior: discrepancies between aspirations and reality. Rationality and Society 10, 79–102.

Diekmann, A., Preisendörfer, P., 2003. Green and greenback: the behavioral effects of environmental attitudes in low-cost and high-cost situations. Rationality and Society 15, 441–472.

Eckes, T., Six, B., 1994. Fakten und Fiktionen in der Einstellungs-Verhaltens-Forschung: Eine Meta-analyse. Zeitschrift für Sozialpsychologie 25, 253–271.

Fujii, E.T., Hennesy, M., Mak, J., 1985. An evaluation of the validity and reliability of survey response data on household electricity conservation. Evaluation Review 9, 93–104.

Gatersleben, B., Steg, L., Vlek, C., 2002. Measurement and determinants of environmentally significant consumer behavior. Environment and Behavior 34, 335–360.

Gatersleben, B.C.M., 2001. Sustainable household consumption and quality of life: the acceptability of sustainable consumption patterns and consumer policy strategies. International Journal of Environmental Pollution 15, 200–216.

Guagnano, G.A., Stern, P.C., Dietz, T., 1995. Influences on attitude-behavior relationships. A natural experiment with curbside recycling. Environment and Behavior 27, 699–718.

Hines, J.M., Hungerford, H.R., Tomera, A.N., 1986/87. Analysis and synthesis of research on environmental behavior: a meta-analysis. Journal of Environmental Education 18, 1–8.

Holland, R.W., Verplanken, B., Van Knippenberg, A., 2002. On the nature of attitude-behaviour relations: the strong guide, the weak follow. European Journal of Social Psychology 32, 869–876.

Jann, B.R.A., 1998. Zur Messung des Umweltverhaltens (Unpublished Master Thesis). Department of Sociology, University of Bern, Swiss.

Kaiser, F.G., 1998. A general measure of ecological behavior. Journal of Applied Social Psychology 28, 395–422.

Kaiser, F.G., Wölfing, S., Fuhrer, U., 1999. Environmental attitude and ecological behaviour. Journal of Environmental Psychology 19, 1–19.

Kok, R., Benders, R.M.J., Moll, H.C., 2006. Measuring the environmental load of household consumption using some methods based on input–output energy analysis: a comparison of methods and a discussion of results. Energy Policy 34, 2744–2761.

Lindenberg, S., Steg, L., 2007. Normative, gain and hedonic goal frames guiding environmental behavior. Journal of Social Issues 63, 117–137.

Maloney, M.P., Ward, M.P., 1973. Ecology: let's hear it from the people – an objective scale for measurement of ecological attitudes and knowledge. American Psychologist 28, 583–586.

Padgett, J.P., Steinemann, A.C., Clarke, J.H., Vandenbergh, M.P., 2008. A comparison of carbon calculators. Environmental Impact Assessment Review 28, 106–115.

Preisendörfer, P., 1998. Umweltbewusstsein in Deutschland 1998. Ergebnisse einer repräsentativen Bevölkerungsumfrage. Bundesministerium für Umwelt. Naturschutz und Reaktorsicherheit, Bonn.

Schultz, W.P., Oskamp, S., 1996. Effort as a moderator of the attitude-behaviour relationship: general environmental concern and recycling. Social Psychology Quarterly 59, 375–383.

Stern, P.C., 2000. Toward a coherent theory of environmentally significant behavior. Journal of Social Issues 56, 407–424.

Thøgersen, J., 1999. Spillover processes in the development of a sustainable consumption pattern. Journal of Economic Psychology 20, 53–81.

UBA (German Federal Environment Agency, 2007. Lifestyle: Consumers Lead the Way in Climate Protection. Dessau-Roßlau, Germany.

Van Lange, P.A.M., Van Vugt, M., Meertens, R.M., Ruiter, R.A., 1998. A social dilemma analysis of commuting preferences: the roles of social value orientations and trust. Journal of Applied Social Psychology 28, 796–820.

Weigel, R.H., Weigel, J., 1978. Environmental concern— the development of a measure. Environment and Behavior 10, 3–15.

Environmental Cognition, Perception, and Attitudes

Baruch Fischhoff, Carnegie Mellon University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by B. Fischhoff, volume 7, pp. 4596–4602, © 2001, Elsevier Ltd.

Abstract

Our decisions shape the environment, affecting what is preserved, destroyed, restored, and ignored. The environment shapes our decisions, affecting our health, economics, and well-being. These diverse decisions are often complex and uncertain. As a result, there is no unifying theory of environmental decisions. Rather, the social and behavioral sciences offer a repertoire of methods, results, and theories that, together, can describe environmental decisions with the requisite richness. They draw on two distinct research traditions: holistic approaches, seeking organizing patterns in those decisions and the people who make them, and reductionist approaches, seeking to assemble pictures by building on existing results. As a result, environmental decisions offer a venue for applying, testing, and extending the social and behavioral sciences, as well as integrating their work with that in the natural sciences.

Individuals' primary interaction with the natural environment is through their senses. They form feelings and impressions that both shape and are shaped by their basic cognitive structures and values. At deeper levels of processing, they form more explicit beliefs and attitudes, including theories and ideologies (Atran, 1990; Kahneman, 2011). These cognitions, perceptions, and attitudes affect both their own well-being and that of their environment. People feel better in a healthy environment and feel better knowing that their environment is healthy (De Young and Princen, 2012). Conversely, the health of that environment depends on choices that people make, as consumers, workers, citizens, officials, tourists, or industrialists (Garner and Stern, 1996; Kates et al., 2012). Indeed, even natural scientists studying the environment are increasingly realizing that understanding human behavior is essential to their work, both to identify the drivers and effects of environmental change, and to make that research useful (Levin and Pacala, 2005; Mooney et al., 2013).

Environment-related decisions are enormously varied, in the options, uncertainties, consequences, and decision-makers involved – not to mention their national, social, cultural, institutional, and economic settings. As a result, there is no simple summary of what people want, believe, feel, or do, relative to the environment. For example, over the past decade, in the United States, overt concern over climate change has increased, more or less, steadily for Democrats, while drifting downward for Republicans. On the other hand, overt behavior related to climate change has often tracked its own course, independent of those attitudes, with climate change 'deniers' often being relatively proactive in adopting energy conservation measures – but reflecting economic concerns, rather than environmental ones (Krosnick et al., 2006; Leiserowitz, 2006). Given the diversity of environmental decisions, this article focuses on approaches to studying *how* people address them, rather than on *what* they do about specific ones – except as illustrative examples.

Research Strategies for Describing Environmental Decisions

Environmental decisions are not only diverse, but also often complex, with many options, outcomes, uncertainties, and decision-makers. As a result, doing justice to any specific decision requires accommodating the details of its case study. Scholars grounded in the humanities have produced many such accounts, in fields such as environmental writing (Slovic and Dixon, 1992), retrospective technology assessment (Tarr, 1996), environmental history (Cronon, 1992), and cultural anthropology (Douglas, 1966). Their detailed observations shape and constrain the theories of social and behavioral scientists, in their quest for general accounts, capturing regularities in environmental behavior found across individual cases. Those scientists can adopt one of two basic compromises in their research approach.

1. *Holistic* approaches look for recurrent behavioral archetypes, in the sense of groups of people who respond in predictable ways (e.g., Cotgrove, 1982; Schwartz, 2006). For example, they might attempt to identify 'ecofreaks,' who resolutely oppose any intrusion into pristine areas, or 'technocrats,' who automatically dismiss such concerns. 'Cultural theory' pursues this approach, seeing whether people can be characterized by their positions on two dimensions and, if so, whether that characterization predicts their attitudes and behaviors (Dake, 1992; Douglas and Wildavsky, 1982). The ongoing 'Six Americas' project categorizes people by their positions on climate change, finding patterns of beliefs, attitudes, and behaviors, arrayed across the spectrum of Alarmed-Concerned-Cautious-Disengaged-Doubtful-Dismissive (Maibach et al., 2011). Moral Foundations Theory (Haidt, 2012) describes people by their orientation on five or six dimensions, such as their respect for 'authority' and their sanctification of 'purity' (as those general properties are defined by their social group).
2. *Reductionist* approaches assume that all people respond to the same factors, but interpret and weigh them differently, sometimes leading to different choices (Edwards, 1954). For example, all people may sometimes rely on the *availability heuristic*, inferring the likelihood of future events based on how easily they can remember or imagine them happening. However, they may reach different conclusions, if they have observed different events or have different powers of imagination (Tversky and Kahneman, 1974).

Similarly, all people learn how to extract the gist of recurrent events, but may be exposed to different events and receive different instruction in what to observe (Reyna, 2012). All people have sacred values, which they will only compromise under duress; however, which those are depends on their culture and social groups (Atran, 2002). Thus, from a reductionist perspective, ecofreaks and technocrats (if such archetypes exist) are otherwise similar people, who have been shaped by different social pressures and material conditions.

Holistic and reductionist accounts coexist in *dual-process* theories of attitude and cognition (Kahneman, 2011; Evans, 2003), which posit parallel processes involving a general orientation toward a stimulus (e.g., an environmental cause, scene, or insult) and a detailed analysis of its contents (Slovic, 2010). That general orientation might be guided by holistic general themes or by reductionist general concerns. The more detailed analysis might overturn those impressions or be constrained by them, as can happen when people fall prey to the many cognitive processes that favor initial beliefs. For example, people may disproportionately seek confirmatory, rather than contradictory evidence; resolve ambiguity in favor of current beliefs; or exaggerate the definitiveness of small samples of evidence (Gilovich et al., 2002). How hard people work to analyze an issue should depend on how important it seems, what return they expect from that effort, and how great is their general need for cognition – reflecting and shaping their opportunities to affect the environment.

Some researchers favor holistic accounts in principle, because they lament the richness lost when behavior is reduced to the expression of general processes. They may be especially skeptical of processes identified in the rarefied settings of experiments crafted to highlight phenomena that interest investigators. Conversely, some researchers favor reductionist accounts in principle, because they need the rigor of experimental evidence regarding clearly defined processes. They might have little patience with the 'grounded theories' that emerge from the immersion that one holistic methodology requires (Charmaz, 2006), unnerved by the limited intersubjectivity that such accounts require. When researchers respect both holistic and reductionist approaches, the tension between them can provide rich accounts, while posing healthy challenges to each. The next section discusses how to realize that potential.

Holistic Accounts

A general account of environmental behavior is useful only if researchers can unambiguously translate complex real-world situations into its necessarily abstract terms. For holistic theories that translation process may itself be holistic, attempting to identify specific cases with theoretical archetypes, such as a particular kind of intergroup conflict, stage of social-economic development, or level of civic engagement (Atran, 2002; De Young and Princen, 2012; Fischhoff, 1995; Kates et al., 2012). Holistic interpretation allows observers to use any available data to elaborate their account. Moreover, if those observers write well, the coherence of their narrative should help readers to retain and integrate the details of their analyses. Archetypes are a natural way to organize evidence, even if that sometimes means neglecting discordant details.

Unfortunately, holistic interpretation also allows observers to exploit the richness of the evidence, in order to fashion accounts that are more convincing than the evidence warrants. For example, they might exploit the privileged position of hindsight to select and interpret facts that affirm their theories (Fischhoff, 1975), perhaps arguing that particular decisions predictably led to success or ruin. Or, they might strategically identify people and groups with archetypes, obscuring inconsistent beliefs, attitudes, and behaviors. Moreover, the very act of categorizing other people is an exercise of power. It may reduce them to stereotypes, while flattering observers as possessing superior insight and even morals, as they reveal others' failings – feeding the natural tendency to create simplistic views of other people (Dawes, 1988; Nisbett and Ross, 1980).

These possibilities are not lost on the partisans in environmental conflicts, who value opportunities to demean their opponents and flatter their allies. As a result, they may seize on holistic accounts in the research literature, and then spin them to their own end, perhaps even loosely citing studies ('X has shown that the public is irrational about nuclear power'). Indeed, scientists can acquire a political following because of the conclusions they reach, independent of the strength of their evidence. For example, the British Conservative government was accused of establishing its Behavioral Analysis Unit, not because it understood the science (Thaler and Sunstein, 2009), but because it sought cover for antiregulatory policies, by claiming that people could be 'nudged' to protect themselves (House of Lords, 2011). The narrative character of holistic accounts means that both scientific and amateur ones have the same outward form, forcing readers to check their scientific pedigree (Funtowicz and Ravetz, 1990). What training did the authors have? What peer-review screens did their work pass? Who paid for the work? (Oreskes and Conway, 2008; Hoffman, 2010) As in other domains where science enters public discourse, environmental researchers bear a special responsibility to ensure that their work is interpreted appropriately.

If the behavioral archetypes that recur in holistic accounts are clearly defined, then it should be possible to operationalize them in individual-difference measures (e.g., Dake, 1992; Dunlap, 2013) If such measures pass conventional psychometric criteria (e.g., reliability, construct validity), then they allow testing predictions derived from the underlying theory (e.g., "corporate greens have deeper commitment to consumption-related issues"; "environmental pessimists have greater scientific knowledge and institutional distrust"). Like other empirical tests, these are only as sound as the *auxiliary assumptions* underlying them, such as how people perceive the costs and benefits of possible actions (e.g., Do pessimists and optimists interpret the same facts differently or possess different facts? Are people vulnerable to framing effects, such that their preferences depend on how the issues are presented?). Here, reductionist approaches can help, by assessing the plausibility of those assumptions ("that information is widely distributed, so most people should have seen it"; "people have stable preferences about such things, so framing shouldn't have much effect") or by measuring them directly

(e.g., finding similar preferences with alternative frames) (Fischhoff, 2005).

If archetypes cannot be measured authoritatively, then individuals may be characterized by observable features, such as their gender, race, income, or nationality – at the price of substituting general social theories for the theoretical force of archetype-based ones. For example, African-Americans typically express relatively great concern over environmental problems (Vaughan, 1993). That could reflect their greater exposure to those problems (meaning that they have more reason for concern), as documented by the environmental justice movement (Schlosberg, 2007). Or, it could reflect greater awareness of their immediate surroundings, greater physiological sensitivity, greater suspicion of social institutions, or poorer resources for self-protection, among other things. A similar array of hypotheses follows from Slovic's (1999) finding that most Americans have similar environmental attitudes, except for some white males, who trust technology more and worry about the environment less than other people. Characterizing the subset of white males responsible for this overall group difference is a starting point for a theory of archetypes, sorting out the experiences or circumstances that create these distinctive views.

Reductionist Approaches

Analogous measurement issues confront reductionist approaches, looking at how common behavioral patterns emerge in diverse settings. For example, approaches grounded in decision theory ask how individuals evaluate possible actions in terms of their expected effects on valued outcomes (Edwards, 1954; Fischhoff and Kadvany, 2011). Those outcomes might affect the well-being of individuals (e.g., taxes, wilderness opportunities, social status), their society (e.g., public health, justice, property rights), or their environment (e.g., biodiversity, reproductive success). In these accounts, actions are attractive to the extent that they increase the chances of good outcomes and decrease the chances of bad ones.

Although they adopt a rational-actor perspective for characterizing the elements of decisions, these approaches vary widely in terms of how far they expect such normative analyses to be descriptively accurate. At the one extreme, lie economic theories that envision well-informed individuals calculating the discounted lifetime stream of goods flowing from different choices. At the other extreme, lie psychological approaches that envision individuals driven by emotions and uncompromising sacred values. The contrast between those choices and analytically optimal ones shows the price that people pay for their nonrationality (Breakwell, 2014).

Such research readily accommodates the heterogeneity of environmental decisions and the importance of specific outcomes in them. For example, social norms may loom large for curbside recycling, where neighbors can see what one does, but not for roadside littering when no one is looking. Thus, while it may be tempting to make general statements of the form 'what matters to people, when it comes to the environment is X,' they always need to be qualified by specifying the context in question. For example, people may not litter even when alone, if they have internalized that behavior as a social norm, have made it a habit, or always fear watching eyes. Even for as important a topic as valuing human life, studies have found widely varying willingness to pay for reducing the probability of premature death (Tengs, 1995; Viscusi, 1992).

As a result, behavior appears most consistent in specific domains, such as home energy conservation (Gallagher and Randell, 2011). There, studies often find that conservation increases when people perceive greater benefits, lower costs, and fewer barriers to change. Although such results might seem just to affirm commonsense, they are needed to establish how general behavioral processes play out in specific settings: Are specific incentives (e.g., cost saving) strong enough to motivate specific energy conservation behaviors? How well did consumers understand those incentives (perhaps buried in their electricity bills)? To what extent did extrinsic incentives undermine intrinsic ones (e.g., environmental concern), by diluting the statement made by prosocial actions? Did consumers have realistic ways to save? Were they in a state of energy poverty, so that lower costs allowed them consume the energy needed for basic well-being (e.g., more tolerable indoor temperatures)?

Thus, as with holistic approaches, reductionist ones must assess the beliefs and values of the individuals whose behaviors they hope to explain, predict, or manipulate. Even a valid model may fail, if its terms are measured poorly. However, that problem is less acute for models with an additive, or 'compensatory,' structure, in which good and bad expectations can cancel one another out (e.g., the health belief model, the theory of reasoned action). Because such models are relatively insensitive to errors in measurement (Dawes, 1988) they will have some predictive validity as long as a study includes rough approximations of the factors affecting behavior (or variables correlated with them). Unfortunately, that very robustness limits these models' explanatory value, because many versions will produce similar predictions. As a result, it is hard to tell, which variables drive behavior and how important each is. That concern applies less with *noncompensatory* decisions, where one consideration overrides all others, as with sacred values. For example, there may be no way to overcome the stigma from contaminating part of the natural world – or to diminish the moral high ground from protecting it (Douglas and Wildavsky, 1982; Flynn et al., 2002; Baron and Spranca, 1997).

Evaluating Perceived Risks and Benefits of Environmental Changes

Assessing beliefs and values is a central activity for the disciplines of cognitive and social psychology. Assessing them for environmental issues has not only drawn on that basic science, but also extended it. For example, some environmental decisions involve unfamiliar processes (e.g., radiation, invasive species, climate changes) playing out unevenly over time and species (Levin and Pacala, 2005). Those decisions may pose fateful trade-offs, where individuals cannot rely on tradition or trial and error to decide what is right: What should we pay to protect endangered species, poor children from lead in their homes, or groundwater from contamination? How should we weigh what our children (and grandchildren) might think about our choices? Such decisions may leave people uncertain

about themselves as well as about their world, asking 'what do I want?' as well as 'what can I get?' Assessing and reducing both kinds of uncertainty has prompted social and behavioral scientists working on environmental issues to break new research grounds.

More or less formally, that research begins by analyzing the decisions that people face, in order to get the facts right and to focus on the right facts – in the sense of the ones that matter most (Baron, 2012). Those analyses take an 'inside view,' trying to see the world the decision-makers face (Kahneman, 2011), reducing the risk of researchers assuming that others share their environmental concerns or exaggerating their store of common knowledge (Dawes, 1988; Nickerson, 1999). Unlike science education, where experts determine which facts are worth knowing, the study of environmental decisions depends on individuals' circumstances, goals, and capabilities (Fischhoff, 2013).

Having characterized a decision, the research proceeds to assess how well individuals understand it, contrasting what they know with what they need to know, in order to make sound choices. That contrast may reveals gaps in their understanding about the world or about themselves. One strategy for avoiding decision-specific studies is to create tests of environmental 'literacy.' Their usefulness depends on whether they include facts that people *need to know* or just ones that it would be *nice to know*. For example, studies of energy conservation have found that people can know about many measures that make some difference without knowing which make a real difference or how to defend their choices against critics or temptations (Moser and Dilling, 2007). Researchers in science education have learned much about the knowledge and skills needed to master a domain (Klahr, 2000). Analogous studies are needed to identify core understandings that create the mental models that allow individuals to interpret diverse environmental decisions. Those might include having an intuitive grasp of how gases diffuse, how risks mount up through repeated exposure, how to use numbers, or how to apply decision-making rules.

Whereas environmental science can establish the clearest possible picture of how decision will affect outcomes that people value, it cannot determine which outcomes those should be. That depends on what matters to those making the decisions. Scientists can, however, ask how successful those people have been in constructing stable preferences for specific environmental choices, consistent with their basic values. That test recognizes that, even when a topic is familiar (e.g., declining fish populations, invasive species, suburban sprawl, vector-borne disease), the trade-offs posed by a specific choice often are not (e.g., a ballot measure, housing decision, a restaurant meal). As a result, individuals cannot just 'read off' a response, from a universal utility or value function, as naïve rational-actor models might hold. Rather, they must 'articulate' or 'construct' their preferences among the options on offer – or perhaps even create new ones (e.g., expanding their housing search).

In such cases, where individuals are not sure what they want, researchers seeking to elicit attitudes and preferences face a choice. They can either take people as they are, perhaps with fragmentary views, or help people to reflect on the issues so as to derive fuller positions. The former strategy risks a sin of omission, if people are swept along by the incomplete set of issues evoked by their initial thinking on the topic. The latter strategy risks a sin of commission, if suggesting alternative perspectives biases decision-makers' thinking, rather than deepening it. The two strategies draw on complementary research traditions. The former has roots in psychophysical research, the latter in decision theory.

Eliciting Environmental Beliefs and Attitudes

Psychophysical Measurement

Emerging from the natural sciences, early psychology focused on what came to be called *psychophysics*, determining the psychological equivalent of physical stimuli. These studies envisioned physiological and psychological mechanisms that translated stimuli into internal states of arousal. If asked, individuals can report on that state, with a word, number, or action (e.g., pressing on a handgrip to show their confidence, adjusting a rheostat to show the light intensity equivalent to the loudness of a tone) (Poulton, 1989).

The most straightforward applications of psychophysical approaches to environmental decisions pose questions such as how natural (or attractive or inspiring or calming) a natural scene is. Researchers might elicit direct ratings (Daniel and Vining, 1983) or similarity judgments, if they hope to capture attitudes that people have difficulty expressing, because they lack the words or hesitate to use them (Cantor, 1977). The less familiar a task, the greater the burden on researchers to ensure that it was interpreted as intended: which features were noticed – and overlooked? What did respondents believe about the *ecosystem services* (Daily, 1997) that the environment in question provides and the peril it faces? Did respondents sense (intended or inadvertent) hints regarding what to say? How real or hypothetical did the task seem? It takes a suite of converging studies, with suitable manipulation checks, to establish what environmental changes people believe they are evaluating, how deep their understanding is, and what they mean by their responses (Fischhoff and Furby, 1988).

Such interpretative issues have been at the center of the long-running controversy over *contingent valuation* (CV), a method advanced by some resource economists for monetizing 'goods' not traded in marketplaces where their economic value could be observed (e.g., atmospheric visibility, the survival of endangered species) (Carson and Hanneman, 2005; Mitchell and Carson, 1989). In the psychophysical tradition, CV interviewers ask people how much they are willing to pay in order to prevent an adverse environmental change (or how much they will demand as compensation for it). These responses are meant for use in cost–benefit analyses, where they could represent environmental changes that would otherwise be neglected in policy-making dominated by economic analyses. Whether a CV study captures respondents' preferences depends on how well they can understand and answer questions demanding explicit, quantitative evaluations of complex, novel stimuli with many potentially relevant details.

CV studies illustrate the strengths and weaknesses of eliciting attitudes toward environmental goods. Respondents' ability to provide some answer to any question suggests that they have some relevant feelings on any environmental issue.

However, feelings alone may not predict what they would pay where there is a market market for that environmental good (Kahneman et al., 1999). One focal research topic is whether CV judgments are properly sensitive to the amount (or 'scope') of the change being evaluated (Arrow et al., 1993). The death of one migratory bird saddens many people, while the death of many birds saddens them even more. However, the differences in those feelings may not be commensurate with the difference in their value (Slovic, 2010). Insensitivity to scope can be invoked as measurement failure. One defense is that people might give as much to 'adopt-a-bird' as to an avian conservation program, if they believe the former to be more effective or assume that other people will adopt the other birds.

However orderly such judgments might be, some critics question the propriety of asking them at all. They argue that monetization 'anaesthetizes moral feeling' (Tribe, 1972), by reducing everything to putative economic equivalents. As a result, CV may win some battles for the environment, by gaining recognition for some otherwise neglected changes, but lose the war, by failing to defend the environment's intrinsic value.

Decision Theoretic Measurement

The early days of the modern environmental movement confronted many technologies with public opposition that their proponents could not, and perhaps would not, understand. One natural response was (and still is) to claim that a technology's opponents overestimate its expected fatalities. Early studies examined that claim by eliciting fatality estimates, finding (1) a strong correlation between lay and statistical fatality estimates; (2) higher estimates for technologies with more 'available' deaths, relative to others with similar statistical frequency; (3) consistent ordinal estimates across response modes; and (4) inconsistent absolute estimates, across response modes. Thus, people seem to have a fairly robust feeling for relative fatalities, which emerge however such (unusual) questions are asked (Lichtenstein et al., 1978). However, they have a poorer feeling for what numbers to use, which makes their absolute judgments sensitive to the contextual cues that a response mode provides (Poulton, 1994).

However, quite different judgments emerged when lay respondents made judgments of 'risk,' rather than of 'average year fatalities.' Moreover, those judgments better predicted their attitudes (e.g., how strictly they should be regulated) (Slovic et al., 1979). Many studies have examined what other aspect of 'risk' affects people. One early candidate was that people give added weight to a technology's 'catastrophic potential,' in the sense that they see greater 'risk' when many lives could be lost at once. Anecdotal support for that hypothesis can be found in the attention drawn to plane crashes and public concern over nuclear power – despite lay respondents' recognition that, in an average year, the death toll from nuclear power is negligible. Studies found, however, that what really bothered people about catastrophic accidents was the uncertainty that made them possible (Slovic et al., 1984). Those two 'risk attributes,' catastrophic potential and uncertainty, embody quite different ethical principles. Avoiding catastrophes means caring how deaths are distributed (over time and space), hence preferring small-scale technologies (other things being equal). Avoiding uncertainty means being risk averse, hence requiring stronger evidence before accepting new technologies.

The study of attributes that define 'risk' can be traced to Starr's seminal (1969) claim that, for any level of benefit, the public tolerates a higher fatality rate for risks incurred voluntarily (e.g., skiing) than for risks imposed involuntarily (e.g., electric power). As evidence, Starr placed eight activities in a risk–benefit space, based on estimates of their annual fatalities and economic benefits, and then sketched parallel 'acceptable risk' lines, an order of magnitude apart, for voluntary and involuntary ones. Such a *revealed preference* analysis assumes that members of a society are satisfied with the risk–benefit trade-offs that they see around them. A study (Fischhoff et al., 1978) testing this hypothesis asked lay people to judge 30 technologies in terms of their current risks and benefits, as well as rating those risks in terms of their acceptability, voluntariness, and eight other attributes (e.g., catastrophic potential, dread, controllability, known to science, known to the public). It found:

1. a weak correlation between lay judgments of current risks and benefits (indicating that respondents saw no greater benefit from riskier technologies);
2. no greater correlation between judgments of current risks and benefits after controlling for judgments of the risk attributes (indicating no double standards for voluntary and involuntary risks, nor for any other attribute);
3. a belief that most technologies had unacceptably high risks (indicating that current trade-offs did not reveal respondents' preferences);
4. a strong correlation between judgments of current benefits and acceptable risks (indicating a willingness to make risk-benefit trade-offs);
5. a willingness to accept higher risk levels for voluntary risks, holding benefits constant, and for other risk attributes (indicating a double standard).

In addition to their substantive content, such studies demonstrate the importance of disciplining speculations about environmental beliefs, attitudes, and behaviors with direct empirical evidence.

Applying Behavioral Research

Setting Public Policy

If robust, such social and behavioral science results can guide public policies affecting the environment. The studies just mentioned suggest that the public will accept risk–benefit trade-offs – which might surprise critics who claim that the public demands zero risk. They show where lay and expert judgments of risk and benefit diverge; hence they must be reconciled through better analysis or better communication. They show risks attributes to consider when designing technologies and regulations. They clarify the ethics embedded in seemingly technical terms. For example, 'mortality risk' can be summarized as 'probability of premature death' or as 'days of lost life expectancy.' The latter measure puts a premium on deaths of young people (where many days are lost with each life), whereas the former treats deaths equally

(Crouch and Wilson, 1982; Fischhoff and Kadvany, 2011). Similarly, any measure of mortality risk (and any policy dependent on it) must either consider (or ignore) how equitably risks are distributed across population groups – and, if so, which groups matter and what 'equitably' means. By raising these questions, behavioral research reveals assumption hidden in the conventions of risk and benefit analysis, so that they can be actively considered in policy making.

Considering a large set of risk attributes makes policy-making more complicated and, potentially, too unwieldy to include public input. One simplification strategy looks for redundancy among risk attributes, and then focuses on a few canonical concerns. For example, if dread risks also tend to have catastrophic potential, then considering either attribute will lead to similar policies. Studies have, in fact, found that two factors can account for much of the variance in ratings of the attributes. One seems to reflect how well risks are understood, the second how much they are dreaded (Fischhoff & Kadvany, 2011). Characterizing multiple risks in terms of the same attributes allows comparisons across domains, and increases the chances for consistency (H.M. Treasury, 2005).

A second strategy accepts the complexity and works with people to master its details. The need for credible resolution of environmental disputes has prompted many consultative bodies to advocate proactive public involvement (e.g., Blue Ribbon Commission, 2012; Presidential-Congressional Commission, 1997). In one important initiative, the US Environmental Protection Agency (Davies, 1996) fostered some 50 state and regional consultations to set priorities among risks. In them, public representatives set the agenda, while technical experts gathered and explained relevant evidence. Dietz and Stern (2008) provide an authoritative summary of the science of public participation.

Ensuring Citizen Competence

Because behavioral and social science research seeks to understand and overcome the limits to lay decision-making, it also provides answers to a critical policy question: Are individuals competent to make sound choices – or should those be left to experts and officials (Lupia and McCubbins, 1998)?

Motivated answers to that question are common. A competent public serves the interests of those seeking active public participation or deregulation (and a marketplace where the public fends for itself). An incompetent public serves the interests of those who favor expert opinion and strong regulation (to protect a defenseless public). Behavioral research can provide a disciplined decision-specific answer, comparing what people know with what they need to know. Sometimes, fragmentary knowledge is enough to identify the best course of action. Sometimes, even extensive knowledge is inadequate and people need help.

Behavioral research can also expand the public's envelope of confidence by filling critical gaps in their knowledge. At times, people need specific facts (e.g., the payback period for improved home insulation, the probability of a nuclear core meltdown). In that case, there is a *supply curve* for information (to use the economic term), arranging facts in order of decreasing usefulness for distinguishing among decision options. Doing so uses recipients' attention wisely and respectfully, starting with the things that they most need to know. At other times, people need to understand the processes shaping environmental changes. In that case, communications must complete their mental model, so that they can follow the action, see when things have changed, and devise effective responses.

Behavioral research can also assess the competence of experts and provide ways to elicit their judgments more systematically (Kammen and Hassendzahl, 1999; O'Hagan et al., 2006). When even the most knowledgeable experts are highly uncertain, a case may be made for moving slowly – unless doing so allows even more poorly understood processes to go unchecked. For example, the debate over genetically manipulated organisms has revolved on whether anyone understands them well enough to make sound choices. Those who doubt that even the experts know enough to proceed often invoke a 'precautionary principle' – at times as sign of distrust, at times as a demand for deeper analyses of uncertainty (Löfstedt et al., 2002).

Conclusion

Environmental decisions are enormously diverse and often enormously complicated and uncertain. Individuals can have a hard time grasping the issues and deciding how they feel about them. Understanding and reducing such difficulties is part of the challenge and opportunity for researchers studying environmental decision-making. Driven by the importance of these problems, environmental social scientists have cast a wide net, adopting and inventing both holistic and reductionist approaches, drawing on diverse research streams. Their work makes it possible to go beyond speculation and anecdotes to systematic evidence regarding what people believe and want regarding their environment. Its fate and, with it, their own depend on how well they understand what is happening and realizing their desires for its future. Behavioral and social science research can help them, identifying new research topics along the way.

See also: Attitude Measurement; Attitudes and Behavior; Environmental Attitudes and Behavior: Measurement; Environmental History; Environmental Justice in the United States; Environmental Movements; Environmental Sciences; Environmental Sociology; Public Opinion: Social Attitudes.

Bibliography

Arrow, K., Solow, R., Portney, P., Leamer, E., Radner, R., Schuman, H., 1993. Report of the NOAA panel on contingent valuation. Federal Register 58, 4601–4614.

Atran, S., 1990. Cognitive Foundations of Natural History. Cambridge University Press, Cambridge.

Atran, S., 2002. In Gods We Trust. Oxford University Press, Oxford.

Baron, J., 2012. The point of normative models in judgment and decision making. Frontiers in Psychology 3 (577). http://dx.doi.org/10.3389/fpsyg.2012.00577.

Baron, J., Spranca, M., 1997. Protected values. Organizational Behavior and Human Decision Processes 70, 1–16.

Breakwell, G., 2014. The Psychology of Risk, second ed. Cambridge University Press, Cambridge.

Cantor, D., 1977. The Psychology of Place. Architectural Press, London, UK.

Carson, R.T., Hannemann, W.M., 2005. Contingent valuation. In: Mäler, K.-G., Vincent, J. (Eds.), Handbook of Environmental Economics. Elsevier, Amsterdam, pp. 821–936.

Charmaz, K., 2006. Constructing Grounded Theory: A Practical Guide through Qualitative Analysis. Sage Publications, Thousand Oaks, CA.
Cotgrove, S., 1982. Catastrophe or Cornucopia. Wiley, New York.
Cronon, W., 1992. Nature's Metropolis. Norton, New York.
Crouch, E.A.C., Wilson, R., 1982. Risk/Benefit Analysis. Balinger, Cambridge, MA.
Daily, G.C., 1997. Nature's Services: Societal Dependence on Natural Ecosystems. Island Press, Washington, DC.
Dake, J.A., 1992. Culture and the social construction of risk. Journal of Social Issues 48 (4), 21–37.
Daniel, T.C., Vining, J., 1983. Methodological issues in the assessment of landscape quality. In: Altman, I., Wohlwill, J.F. (Eds.), Human Behavior and Environment, vol. 16. Plenum Press, New York, NY, pp. 39–84.
Davies, C. (Ed.), 1996. Comparing Environmental Risks. Resources for the Future, Washington, DC.
Dawes, R., 1988. Rational Choice in an Uncertain World. Harcourt Brace Jovanovich, San Diego, CA.
De Young, R., Princen, T. (Eds.), 2012. The Localization Reader: Adapting to the Coming Downshift. Cambridge, MIT Press, MA.
Dietz, T., Stern, P.C. (Eds.), 2008. Public Participation in Environmental Assessment and Decision Making. National Academy Press, Washington, DC.
Douglas, M., 1966. Purity and Danger. Routledge & Kegan Paul, London.
Douglas, M., Wildavsky, A., 1982. Risk and Culture. University of California Press, Berkeley, CA.
Dunlap, R. (Ed.), 2013. Climate Change Skepticism and Denial. American Behavioral Scientist.
Edwards, W., 1954. The theory of decision making. Psychological Bulletin 41, 380–417.
Evans, J.StB.T., 2003. In two minds: dual-process accounts. Trends in Cognitive Science 7, 454–459.
Fischhoff, B., 1975. Hindsight ≠ foresight: the effect of outcome knowledge on judgment under uncertainty. Journal of Experimental Psychology: Human Perception and Performance 104, 288–299.
Fischhoff, B., 1995. Risk perception and communication unplugged: twenty years of process. Risk Analysis 15, 137–145.
Fischhoff, B., 2005. Cognitive processes in stated preference methods. In: Mäler, K.-G., Vincent, J. (Eds.), Handbook of Environmental Economics. Elsevier, Amsterdam, pp. 937–968.
Fischhoff, B., 2013. The sciences of science communication. Proceedings of the National Academy of Sciences 110, 14033–14039.
Fischhoff, B., Furby, L., 1988. Measuring values: a conceptual framework for interpreting transactions. Journal of Risk and Uncertainty 1, 147–184.
Fischhoff, B., Kadvany, J., 2011. Risk: A Very Short Introduction. Oxford University Press, Oxford.
Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S., Combs, B., 1978. How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. Policy Sciences 8, 127–152.
Flynn, J., Kunreuther, H., Slovic, P. (Eds.), 2002. Risk, Media and Stigma. Earthscan, London.
Funtowicz, S., Ravetz, J., 1990. Uncertainty and Quality in Science for Policy. Kluwer, London.
Gallagher, K.S., Randell, J.C., 2011. What makes the US energy consumers tick? Issues in Science and Technology 28 (4).
Garner, G.T., Stern, P.C., 1996. Environmental Problems and Human Behavior. Allyn & Bacon, Needham Heights, MA.
Gilovich, T., Griffin, D., D.Kahneman (Eds.), 2002. The Psychology of Judgment: Heuristics and Biases. Cambridge University Press, New York.
Haidt, J.H., 2012. The Righteous Mind. Pantheon, New York.
Hoffman, A.J., 2010. Climate change as a cultural and behavioral issue. Organizational Dynamics 39, 295–305.
House of Lords, 2011. Behavior Change. HL Paper 179. The Stationery Office, London.
Kahneman, D., 2011. Thinking, Fast and Slow. Farrar Giroux & Strauss, New York.
Kahneman, D., Ritov, I., Schkade, D., 1999. Economic preferences or attitude expression? Journal of Risk and Uncertainty 19, 203–242.
Kammen, D., Hassendzahl, D., 1999. Shall We Risk It? Princeton University Press, Princeton, NJ.
Kates, R.W., Travis, W.R., Wilbanks, T.J., 2012. Transformational adaptation when incremental adaptations to climate change are insufficient. PNAS. www.pnas.org/cgi/doi/10.1073/pnas.1115521109.
Klahr, D., 2000. Exploring Science: The Cognition and Development of Discovery Processes. MIT Press, Cambridge, MA.
Krosnick, J.A., Holbrook, A.L., Lowe, L., Visser, P.S., 2006. The origins and consequences of democratic citizens' policy agendas. Climatic Change 77 (1-2), 7–43.
Leiserowitz, A., 2006. Climate change risk perception and policy preferences. Climatic Change 77, 45–72.
Levin, S.A., Pacala, S.W., 2005. Contingent valuation. In: Mäler, K.-G., Vincent, J. (Eds.), Handbook of Environmental Economics. Elsevier, Amsterdam, pp. 61–95.
Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., Combs, B., 1978. Judged frequency of lethal events. Journal of Experimental Psychology: Human Learning and Memory 4, 551–578.
Löfstedt, R., Fischhoff, B., Fischhoff, I., 2002. Precautionary principles: general definitions and specific applications to genetically modified organisms (GMOs). Journal of Policy Analysis and Management 21, 381–407.
Lupia, A., McCubbins, M.D., 1998. The Democratic Dilemma: Can Citizens Learn What They Need to Know. Cambridge University Press, New York.
Maibach, E., Leiserowitz, A., Roser-Renouf, C., Mertz, C.K., 2011. Identifying like-minded audiences for climate change public engagement campaigns: an audience segmentation analysis and tool development. PLoS ONE 6 (3), e17571.
Mitchell, R.C., Carson, R.T., 1989. Using Surveys to Value Public Goods. Resources for the Future, Washington, DC.
Moser, S., Dilling, L. (Eds.), 2007. Creating a Climate for Change: Communicating Climate Change and Facilitating Social Change. Cambridge University Press, Cambridge.
Mooney, H.A., Duraiappah, A., Larigauderie, A., 2013. Evolution of natural and social science interactions in global change research programs. PNAS. www.pnas.org/cgi/doi/10.1073/pnas.1107484110.
Nickerson, R.A., 1999. How we know – and sometimes misjudge – what others know. Psychological Bulletin 125, 737–759.
Nisbett, R.E., Ross, L., 1980. Human Inference. Prentice-Hall, Englewood Cliffs, NJ.
O'Hagan, A., Buck, C.E., Daneshkhah, A., et al., 2006. Uncertain Judgements: Eliciting Expert Probabilities. Wiley, Chichester.
Oreskes, N., Conway, E.M., 2008. Merchants of Doubt. Bloomsbury Press, New York.
Poulton, E.C., 1989. Bias in Quantifying Judgment. Lawrence Erlbaum, Hillsdale, NJ.
Poulton, E.C., 1994. Behavioral Decision Making. Cambridge University Press, Cambridge.
Reyna, V.F., 2012. A new intuitionism: meaning, memory, and development in fuzzy-trace theory. Judgment and Decision Making 7, 332–339.
Schlosberg, D., 2007. Defining Environmental Justice: Theories, Movements, and Nature. Oxford University Press, Oxford.
Schwartz, S., 2006. A theory of cultural value orientations: explication and applications. Comparative Sociology 5 (2–3), 137–182.
Slovic, P., 1999. Trust, emotion, sex, politics, and science. Risk Analysis 19, 698–701.
Slovic, P., 2010. Feeling of Risk. Earthscan, London.
Slovic, P., Fischhoff, B., Lichtenstein, S., 1979. Rating the risks. Environment 21 (4), 14–20, 36–39.
Slovic, P., Lichtenstein, S., Fischhoff, B., 1984. Modeling the societal impact of fatal accidents. Management Science 30, 464–474.
Slovic, S., Dixon, T. (Eds.), 1992. Being in the World: An Environmental Reader for Writers. Macmillan, New York.
Starr, C., 1969. Societal benefit versus technological risk. Science 165, 1232–1238.
Treasury, H.M., 2005. Managing Risks to the Public. Author, London.
Tarr, J.A., 1996. The Search for the Ultimate Sink. University of Akron Press, Akron.
Tengs, T., 1995. Five-hundred life-saving interventions and their cost-effectiveness. Risk Analysis 15, 369–390.
Thaler, R., Sunstein, C., 2009. Nudge: Improving Decisions about Health, Wealth and Happiness. Yale University Press, New Haven, CT.
Tribe, L.H., 1972. Policy science: analysis or ideology? Philosophy and public affairs 2, 66–110.
Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics & biases. Science 185, 1124–1131.
Vaughan, E., 1993. Individual and cultural differences in adaptation to environmental risks. American Psychologist 48, 673–680.
Viscusi, W.K., 1992. Fatal Tradeoffs. Oxford University Press, Oxford.

Environmental Conservation and Development: Critical Perspectives

Scott Freeman, George Washington University, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Starting in the 1960s, environmental movements brought global attention to humanity's negative effects on the natural environment. Governments and international organizations became increasingly concerned with addressing these environmental problems. This article examines arguments about how and why environmental degradation occurs and discusses the critical literature at the intersection of environmental conservation and international development programs. While interventions are often well intended, many bring with them the unintended negative effects of population displacement, marginalization of indigenous communities, and questionable economic benefits. Future policies and programmatic interventions must be subject to rigorous research and critique in order to improve their effects on local populations and the environments.

Detrimental logging, mining, contamination of drinking water, desertification, and air pollution constitute widespread environmental degradation and have incited global concern. Not only have environmental problems become more visible and publicized since the 1960s, they have galvanized worldwide response. Environmental movements have inspired governments, individually and through international accords, to address degradation (Leach and Fairhead, 2000). Local and global nongovernmental organizations (NGOs), as well as multilateral entities such as the United Nations (UN) and the World Bank have made priorities of environmental sustainability and conservation. Resulting parks and protected areas, biodiversity conservation initiatives, and broad schemas to regulate and monitor environmental pollution and consumption have arisen in the late twentieth century. Many of these initiatives have aligned themselves with the simultaneous goal of development and the elimination of poverty.

Yet while many policies and programmatic interventions are promising on paper, they have proven to be ineffective, or worse, to have negative effects on the populations living in targeted conservation areas. Critiques of environmental conservation and conservation as development have highlighted the way in which populations have been forcibly displaced, and how, for example, indigenous rights have been violated. Examining state and transnational efforts to control resource exploitation highlights how the regulation of nature may introduce new social and material problems into targeted areas. In the age of increasing environmental awareness, researchers must examine not only the roots of environmental degradation but also the interventions themselves.

This article discusses how environmental degradation has been conceptualized and the initiatives that have arisen in response. I refer to environmental degradation as human instigated damage and destruction of what are commonly referred to as natural resources: water, land, trees, animals, and air. Degradation has been conceptualized in multiple ways, including neo-Malthusian explanations related to population growth, and explanations deriving from political ecology that site resource appropriation. As environmental conservation interventions sought to protect biodiversity throughout the 1980s and 1990s, governments and international NGOs were interested in melding conservation with the elusive goal of economic growth and development. This article reviews the literature on environmental degradation and the literature on development, as well as the intersection of the two. It covers trends in conservation such as parks and protected areas, the conservation and development paradigm, and neoliberal market-based environmental regulation.

Views of Environmental Degradation from Malthus to Political Ecology

Chronic degradation such as air pollution and short crisis events such as oil spills continue to galvanize local and global concern for the environment. The human effects on environmental change are so profound that a new geologic age has been proposed. The Holocene, the geologic era that has defined the past 10 000–12 000 years, has arguably given rise to a new geologic era, that of the Anthropocene (Crutzen and Stoermer, 2000; Crutzen, 2002). The fact that human action has altered the planet for the worse offers the possibility that a more enlightened humanity might alter the planet for the better, presenting a formidable sense of responsibility (Gibson-Graham and Roelvink, 2009).

In the first wave of environmentalism in the 1960s, scholars theorized this degradation as the result of population growth. Drawing on the work of Thomas Malthus, commentators considered population pressure as the cause of environmental destruction. Scholarship emerging in the late 1970s and 1980s challenged this perspective, offering contrary framings. The following sections outline environmental theories championed by neo-Malthusianism and political ecology.

Neo-Malthusianism

Post-World War II framings of environmental degradation were based on concerns of population growth and natural resource bases. This movement differed drastically from the conservation movement that Theodore Roosevelt headed in the 1890s and 1900s (Robbins, 2012). The post-War movement drew on the work of the philosopher Thomas Malthus, and focused on growing populations as the primary cause of degradation. Malthus' original writings were not immediately directed at the

environment, but rather theorized poverty. In the late eighteenth and early nineteenth centuries, Malthus saw population growth as the cause of the urban poverty he saw in England. His thesis held that it was the uncontrolled growth of the poor that was responsible for the increase in poverty. The poor, unable to restrain themselves from reproducing, would increase in number while the rich practiced measured self-restraint. Malthus' theories held that because of this, hunger and death were the forces that kept poor populations in check. For him, poverty was quite simply a condition of human existence (Rao, 1994).

In the 1960s, environmental movements drew on these theories. The association of population growth with poverty shifted to a connection between population growth and the degradation of natural resources and the environment (Peet and Watts, 1996). Population-oriented environmentalists drew on Malthus and warned about the increase in numbers of impoverished people and what this would mean for the environment. It was posited that growing populations would consume nature's assets beyond the point of self-renewal (Robbins, 2012). In his widely read book, *The Population Bomb* (1968), Paul Ehrlich raised public concern about population growth, especially in low-income countries, and pointed out the coming crisis that growing global pressure on natural resources was causing.

While this alarmist perspective lost momentum after the 1980s, the link between population growth and degradation has been maintained through the work of multiple international organizations. Multilateral institutions have continued to tie environmental degradation to population growth and therefore the need for population control and family planning. Accords such as the 1994 UN International Conference on Population and Development cemented relationships between family planning and environmental degradation (Hartmann, 2006). In the late 1990s, for example, the World Resources Institute cited population growth since 1950 as a primary actor in spurring anthropogenic degradation (Leach and Fairhead, 2000). As recently as 2013, the United States Agency for International Development (USAID) funded environment projects still assert relationships between population and the environment (USAID, 2013: 22–24).

Yet critics pointed out that environmental degradation can occur with or without population growth (Blaikie, 1985). Correlative relationships between population growth and degradation may misrepresent the relationships between people and forests (Leach and Fairhead, 2000). Concepts such as population 'carrying capacity' ignored the movement of resources and blamed those most proximate to resource extraction, ignoring detrimental consumptive practices that had tangible affects but were geographically distant. Political ecology arose as a response to these population based theories of degradation.

Political Ecology

Neo-Malthusian focus on precipitous population growth hid the realities of class divisions and inequalities of consumption that undergirded environmental degradation (Robbins, 2012). Political ecology combines concerns of power and inequality while considering questions of the environment. In doing so, political ecology refutes the population and resources view and asserts the centrality of poverty in degradation (Martinez-Alier, 1991). From the beginning of political ecology, a strong difference has been made between population pressure on resources and production pressure on resources (Blaikie and Brookfield, 1987).

Anthropologist Eric Wolf first used the term political ecology in 1972 to link land-use discussions with global political economy. The interdisciplinary field arose in the early 1980s from a combination of anthropological human ecology and Marxist-inspired world systems theory. Geographers Piers Blaikie (1981, 1985), and Harold Brookfield (Blaikie and Brookfield, 1987), famously authored the soil studies that would become the foundations of political ecology. They argued that previous studies of soil degradation had been myopic in their focus. Rather than study simply farmers, land, and soil scientists, Blaikie and Brookfield argued for analysis that included the land manager in a broader historical, political, and economic context. By using this inclusive focus, political ecologists argue that environmental degradation should be understood as a series of extractive relationships. Investigations of political ecology often highlight accumulation through dispossession, and focus on the marginalized position of rural populations. Because rural land managers may be merely negotiating the constraints of larger processes of accumulation that extract resources (crops, wood) from their lands, examinations that fail to take these perspectives into account are severely limited.

As political ecology grew, it continued to examine power and the environment, but also incorporated emerging theoretical perspectives from other social sciences. In the 1980s and 1990s, political ecology was complemented by feminist political ecology (Rocheleau et al., 2013), which analyzed gender dynamics in the environment. Others incorporated post-structuralism, arguing that nature was not a material 'given' but rather a discursive and social production (Peet and Watts, 1996). Contemporary political ecologists in the post-2000s have followed this thread, questioning assumptions about what the environment is. Many have pushed for understandings of the environment that blur the distinction between 'nature' and 'culture' paying attention to the existence of multiple world views, or ontologies (Escobar, 1999; Biersack and Greenberg, 2006; Kohn, 2013; Descola, 2005). Political ecologists not only challenge prevailing ideas about degradation, but raise questions about how degradation and the 'natural' environment is defined and addressed.

The Growth of Development and Development Aid

Since Blaikie and Brookfield's (1987) work, political ecology has focused not only on degradation but also the interventions that address that degradation. International development organizations have played a large part in such projects and initiatives. Here I refer to development as the drive for continued growth and the work done by development agencies and government actors in attempts to achieve economic growth. In a more basic sense, development is also the linear pursuit of progress that emerges with global capitalism (Leys, 1996). But as social scientist Wolfgang Sachs (1993) points

out, that very progress of industrialization comes at the cost of environmental resources. This unwitting destruction has been referred to as the second contradiction of capitalism: as capitalism and growth spreads, it needs resources in order to produce commodities for consumption. This process of growth therefore devours the very resources on which profits depend (O'Connor, 1998).

Development also refers to a set of theories that arose starting in the 1950s that address what was then deemed to be the problems in 'the third world.' The goal was to transfer Western knowledge and technology to make postcolonial countries more productive. Technical expertise wielded by Western Europe and the United States was called upon to benefit the so-called 'underdeveloped' nations. In those efforts, industrialization gained heightened importance and the so-called experts from the global north were sent to give technical expertise to the global south. Therefore, development was not only a theory of progress but also became defined by the imposition of planned social interventions (Escobar, 1995).

The results of the development paradigm and development aid have been severely critiqued on economic and social grounds. Economic development has been discussed as contributing to the improvement of the 'developed' countries at the expense of the 'underdeveloped.' Early development aid was little more than a way for previous colonial powers to benefit from the production of previous colonies (Leys, 1996). In some cases, planned social programs continued to propagate resource extraction and relocation in target countries (Greenough and Tsing, 2003).

Development aid has had profound social and discursive effects as well. Such a discourse of 'developed' and 'underdeveloped' creates blanket categories that eliminate difference and feed the needs of the development industry (Escobar, 1995). Development processes and documents construct particular needs, often quite divorced from reality (Ferguson, 1999). The constant drive and 'will to improve' justifies the work of experts (most often from the global north) and increases their control of nature (Li, 2007). Scholarship has also attempted to understand development as a particular type of practice (Mosse, 2013). Technical interventions create ideal models of development and cite expert opinion, but the realities 'on the ground' are quite different. Questions have arisen as to if good policy even matters, as the exigencies of development funding structures and the often stark disconnect between aid workers and beneficiaries create complications in moving from an ideal to a reality (Lewis, 2004).

Environmental Protection Strategies and Interventions

This section looks at how development interventions have intersected with ongoing environmental efforts. The environmental awareness of the mid-twentieth century translated into international agreements among countries in the 1980s intended to support sustainability and increase conservation (Peluso, 1993). 'Sustainability' emerged as a term highlighting the integration of development with environmental conservation. The term gained increased attention at the 1992 United Nations Conference on Environment and Development 'Earth Summit' in Rio de Janeiro (Sachs, 1993). If environmental protection was on the table, would it come at the expense of the economic growth of countries? In the end, the Rio conference held that economic development was of primary concern. It was the Brundtland Report in 1987 that cemented at Rio a simultaneous discourse of both growth and environmentalism. 'Sustainable development' would combine these ideas and become the champion phrase of the 1990s and into the 2000s.

Multilateral institutions such as the United Nations Environmental Programme, and large international NGOs such as Conservation International began to define the standards and patterns of interventions that would tackle both development and environmental conservation. USAID, the World Bank, the Department for International Development of the United Kingdom, the UN, and others contribute to goals of both development and environmental protection and conservation. These institutions plan and implement projects that aim to protect natural resources, conserve biodiversity, and achieve market-based environmental regulation.

While overt goals of environmental protection have been well publicized, the latent effects of programs and projects have often been eschewed. Two common strategies are parks and protected areas, and the conservation and development framework. In the 1990s and more fervently in the 2000s, neoliberal tactics of conservation have been developed. In each of these strategies, there are questionable patterns of engagement which have prompted the continued vigilance of scholars: "In any given conservation setting, few people may be interested in actual environmental protection, while many people may use the *idea* of protection to further other claims" (Haenn, 2005: 25).

Parks and Protected Areas

Parks and protected areas seek to protect nature and biodiversity by bounding particular spaces. They have long affected not only the environments that they seek to protect, but also the lives and communities that live in and around these areas. The number of protected areas has expanded greatly since World War II, and since the early 2000s the number has risen even more (Adams and Hutton, 2007; West and Brockington, 2006). By 2005, it was estimated that close to 12% of the land surface of the planet was covered by more than 100 000 protected areas (Chape et al., 2005). The establishment of these areas is in hopes of protecting and increasing biodiversity, yet the benefits of this extraordinary number of protected spaces has been called into question. The idea that a park or protected area preserves ecologies and communities in a state of suspended animation is assumed, but quite problematic. Rather, parks and protected areas take dynamic ecologies and societies and alter them, often radically, into a manageable and externally predetermined form (Holmes and Brockington, 2012; West and Brockington, 2006).

The widespread establishment of protected areas brought forth a wave of critiques of displacement in the late 1980s and into the 1990s. These analyses became even more sweeping and caustic in the early and mid-2000s (Agrawal and Redford, 2009). Alterations of natural and social environments have employed multiple forms of violence. Even early iterations of national parks, such as the national park system in the United

States, relied on techniques of displacement in order to achieve their goals. In fact, the US Calvary evicted long-time residents for the establishment of Yellowstone National Park (Jacoby, 2001; Holmes and Brockington, 2012). Parks and protected areas following US-derived models have followed this trend, evicting humans from spaces that will soon become 'protected.' Under this model, humans are seen (often incorrectly) as a priori threats to biodiversity (Adams and Hutton, 2007; West, 2006).

Displacement has been one of the most profound impacts of protected areas. Displacement has a direct affect on livelihoods and leads to risks of impoverishment (Brechin et al., 2002). In Kenya, 'legitimate' violence was used by the state to create protected areas with the goal of state resource sovereignty. In the end, the protected area restricted the movements of Masai nomads and validated lethal force against poachers, ignoring the context in which groups had been forced to assume alternative means of income (Peluso, 1993). Even if people are not removed from protected areas, these spaces alter land-use rights in general (West et al., 2006). While some argue for the increase in the size of protected areas and the rigor with which they are protected, these measures would likely result in even higher rates of displacement (Agrawal and Redford, 2009).

Often, resources from protected areas find their way into the hands of elites or other power brokers. Research in Guinea-Bissau showed that the benefits of conservation efforts do not always go to the 'beneficiaries' but are often controlled by the NGOs that act as intermediaries (Temudo, 2012). In effect, these areas are designed to exclude the majority of people from making decisions about an area and in practice this often translates into the maintenance of power within the hands of elites (Haenn, 2005). In Mexico, anthropologist Molly Doane (2012) found that large organizations had eschewed more local efforts. While local people were preserving land through communal control, organizations like the World Wildlife Fund redefined forests as 'rain forests,' appropriating the symbolic value of land and dismissing other efforts. Broad statistical reviews of parks and protected areas do not capture many of these dynamics, in-depth studies suggest that the extent of these negative effects may be more serious and widespread than reported.

Conservation and Development

Responding to the simultaneous calls for development and environmental protection, conservation has been intimately tied to development and poverty alleviation. The Integrated Conservation and Development Project model was fostered in 1980 by the International Union for the Conservation of Nature, the World Wildlife Fund, and the United Nations Environmental Programme (Dove, 2006). The model proliferated in the 1990s and has found global growth through commitments to raising the standards of living within protected areas (Wells, 1992).

This reworking of development toward environmentalism has prompted social science to follow with a steady and growing series of critiques (Sachs, 1993; West, 2006). Agrawal and Redford (2006) published a systematic survey of 37 projects attempting biodiversity conservation and poverty alleviation, finding little to no evidence of the synergy between the two goals. As the goals of development and poverty alleviation are applied to conservation goals, they may propagate a particular course of development that relies on natural resource exploitation, as in the Pacific region of Columbia (Escobar, 2008). However, conservation may also appropriate value from landscapes by transforming them into a type of capital that can circulate and generate further value at the global level (Garland, 2008). Conservation, through development interventions, has imposed changes on local livelihood strategies, turning 'target groups' into accessible sources of labor, for example, as workers on tea plantations in Rwanda (Masozera and Alavalapati, 2004).

Analyses have also highlighted how conservation and development imposes prefabricated definitions on the communities in question. 'Communities' are often erroneously treated as homogenous, territorially bounded, and sharing common norms and interests (Temudo, 2012). Often, this is the result of a 'rendering technical' that glazes over complex political and social realities in favor of technical solutions conceived of by aid institutions, states, and policy makers (Li, 2007). This rendering technical strips local areas of complexity and difference, relying on the assumption that local inhabitants are threats to resources (West, 2006). Often, these definitions produce myopic visions, ignoring evidence outside the area in question. For example, project documents in a cross-border park targeted the water-wasting practices of Lesotho agriculturalists, yet remained silent about the consumption of water in luxurious golf courses nearby (Büscher, 2012).

Much of the concern about conservation and development overlaps with the ways in which the market economy has been imposed on the environment. The very premise of conservation as development ties conservation to development, and the environment to the market economy (West, 2006; Escobar, 1999; Sachs, 1993).

Market Regulation and REDD+

Increasingly, markets are being used to regulate the environment, and as a response, political ecologists including geographers, anthropologists, and sociologists, are examining the results of neoliberal environmental regulation. Neoliberal approaches assert that "human well-being can best be advanced by liberating individual entrepreneurial freedoms and skills within an institutional framework characterized by strong private property rights, free markets and free trade" (Harvey, 2005: 2). Turning toward the dynamics of the market, these strategies seek to commodify environmental goals and often, the environment itself. Through market mechanisms the monetization of the environment will then ideally bring forth the 'rational economic actor.' It is hoped that this monetization will alter behaviors toward economic gain and away from economic loss.

While neoliberal economic policies ostensibly espouse market deregulation, evidence suggests that in fact neoliberalism entails 'reregulation' as well: government regulation is used to transform previously untradable things into tradable commodities (Igoe and Brockington, 2007). Many conservation interventions are designed on the assumption that the biggest obstacle to effective conservation is that nature has not

yet been adequately commodified (McAfee, 1999; Grandia, 2007). A key component of turning nature into buyable and sellable forms is the demarcation of territories in which this can occur. The demarcation and capture of resources not only affects individuals living in proximity, but also puts control in the hands of external agencies and institutions such as large international NGOs. Many of the indigenous groups affected by neoliberal conservation have therefore been on the losing side, finding their livelihood strategies criminalized by newly demarcated zones (Brockington and Igoe, 2006).

One such neoliberal environmental strategy has been carbon markets. Carbon offsets were developed as 'tools to manage anthropogenic climate change' by trading tons of carbon dioxide in international markets (Bumpus, 2011: 612). In effect, carbon credits are purchased by investment in carbon reduction projects, reductions that would not have occurred if the project had not taken place (Yamin, 2005).

Political ecologists have argued that the development of forest carbon networks has amplified existing exclusions (Filer and Wood, 2012). While carbon offsets may espouse market regulation of environmental contamination, the commodification of carbon markets has been critiqued for veiling problems of displacement and continued environmental degradation. By following a Dutch energy firm's carbon trade, Checker (2009) demonstrated the intricate way in which the company purchased carbon offsets to hide detrimental coal mining ventures in the United States, while simultaneously supporting a protected area in Uganda accused of propagating violence and displacement. A veneer of progress hid human rights violations in the form of displacement and allowed for ongoing extractive processes to be obscured behind supposedly proenvironmental efforts.

The REDD+ programme is a specific example of the carbon regulation markets. REDD+ (Reducing Emissions from Deforestation and forest Degradation; the '+' stands for protecting biodiversity and local livelihoods) aims to reduce carbon emissions caused by the cutting of forests. In 2007, the Subsidiary Body for Scientific and Technical Advice of the United Nations Framework Convention on Climate Change unveiled the concepts of REDD+. Broadly, REDD+ is a framework for the measurement and exchange of carbon stock, one that seeks to preserve forested space in developing countries by estimating carbon (tons per territory) and creating a market for that carbon among various actors abroad. While there is no single entity in charge of REDD+, several UN organizations and the World Bank are major actors (Thompson et al., 2011). One of the most serious critiques has been that REDD+ programs ignore people's relationships with the forest in favor of state and international accords. Some argue that REDD+'s efforts are toward building government capacity. However these efforts are often far from the interests of marginalized and indigenous peoples, and may exacerbate existing inequalities (Agrawal, 2010). While indigenous participation is facilitated by UN-REDD administrative structures, this participation has been critiqued for legitimizing the top-down project model without the meaningful consideration of stakeholders. Because the structure revolves around carbon accounting, there are opportunities for more heavily centralized forest management activities, activities that are not designed to incorporate more organic types of participation (Börner et al., 2010; Phelps et al., 2010).

While the REDD+ initiative envisions positive outputs to landowners in the area, these benefits may be problematic, unequitably rewarding larger landholders more than smaller landholders, and presenting far more challenges to lower-income families who practice, for example, swidden (shifting cultivation) agriculture on small plots (Börner et al., 2010). Similar to critiques of conservation more generally, Whiteman et al. (2010) argue that this type of framework continually privileges elites in developing countries. In doing so, it validates particular perspectives while marginalizing others. Böhm et al. (2012) argue that in practice, carbon markets are new creative modes of accumulation and are unlikely to transform dynamics in ways that form a more sustainable global economy and help the rural poor.

Alternative Visions of Conservation

Given the questionable results of top-down environmental and development interventions, other models of conservation have been proposed that incorporate the above critiques. Participatory conservation interventions have arisen that attempt to add more meaningful involvement of project beneficiaries. While broader goals take time, and participatory methods still need to consider the social dynamics within target groups, participatory interventions can bring about new types of conservation awareness (Khadka and Nepal, 2010). Research in the past year points to ways that participatory conservation might move away from 'quota'-based participation to develop more inclusive forms of incorporating marginalized groups (Dahal et al., 2014). Scholars focusing on participation note that community based models do not directly influence the degree of participation, and emphasize that an awareness of the local context is a far more important variable for future interventions to consider (Méndez-López et al., 2014). In one case in Mexico, participation was best achieved when NGOs were removed, allowing space for other types of groups to care for natural resources outside of the existing development paradigm (Martinez-Reyes, 2014).

As recommended by West and Brockington (2006), planners would be well served to recognize alternative understandings of the environment. Similarly, alternative visions of development, espoused by those living in the areas in question, might be taken into account (see Escobar, 2008). Interdisciplinary teams are arguing for more 'local' studies that embrace the knowledge produced by social scientists and critics (Gardner et al., 2013). The ongoing contemporary work of political ecologists and other scholars is to not only continually document and understand environmental degradation and interventions, but to document potential ways in which care for natural resources might occur.

See also: Anthropocene Epoch; Capitalism; Citizen Mapping; Climate Change; Climate, History of; Development Geographies; Development Studies; Environment and Development; Environmental Economics; Environmental History; Local Economic Development; Peasants and Rural Societies in History (Agricultural History); Poverty, Geography of; Rural Planning: General; Social Ecology; Social Inequality in

History (Stratification and Classes); Sustainable Development: An Economic Perspective; Sustainable Development.

Bibliography

Adams, William M., Hutton, Jon, 2007. People, parks and poverty: political ecology and biodiversity conservation. Conservation and Society 5 (2), 147.

Agrawal, Arun, 4 May 2010. Seeing REDD over Forest Peoples. BBC. http://news.bbc.co.uk/2/hi/science/nature/8660845.stm (accessed 09.07.14.).

Agrawal, Arun, Redford, Kent, 2009. Conservation and displacement: an overview. Conservation and Society 7 (1), 1.

Agrawal, Arun, Redford, Kent, 2006. Poverty, development, and biodiversity conservation: shooting in the dark? Ann Arbor, MI 48109 (734), 647–5948.

Biersack, Aletta, Greenberg, James B., 2006. Reimagining Political Ecology. Cambridge University Press.

Blaikie, Piers, 1981. Class, land-use and soil erosion. Development Policy Review 14 (2), 57–77.

Blaikie, Piers, 1985. The Political Economy of Soil Erosion in Developing Countries. Longman.

Blaikie, Piers, Brookfield, Harold, 1987. Land Degradation and Society. Methuen.

Böhm, Steffen, Ceci Misoczky, Maria, Moog, Sandra, 2012. Greening capitalism? A Marxist critique of carbon markets. Organization Studies 33 (11), 1617–1638.

Börner, Jan, Wunder, Sven, Wertz-Kanounnikoff, Sheila, Tito, Marcos Rügnitz, Pereira, Ligia, Nascimiento, Nathalia, 2010. Direct conservation payments in the Brazilian Amazon: scope and equity implications. Ecological Economics 69 (6), 1272–1282.

Brechin, Steven, Wilshusen, Peter, Fortwangler, Crystal, West, Patrick, 2002. Beyond the square wheel: toward a more comprehensive understanding of biodiversity conservation as social and political process. Society & Natural Resources 15 (1), 41–64.

Brockington, Daniel, Igoe, James, 2006. Eviction for conservation: a global overview. Conservation and Society 4 (3), 424.

Bumpus, Adam G., 2011. The matter of carbon: understanding the materiality of tCO2e in carbon offsets. Antipode 43 (3), 612–638.

Chape, Stuart, Harrison, Jeremy, Spalding, Mark, Lysenko, Igor, 2005. Measuring the extent and effectiveness of protected areas as an indicator for meeting global biodiversity targets. Philosophical Transactions of the Royal Society Series B 360, 443–455.

Checker, Melissa, 9 September 2009. Double Jeopardy: Carbon Offsets and Human Rights Abuses. Counterpunch. http://www.counterpunch.org/checker09092009.html (accessed 12.06.14.).

Crutzen, Paul J., 2002. The "Anthropocene". In Journal de Physique IV (Proceedings), 1–5. EDP sciences.

Crutzen, Paul J., Stoermer, Eugene F., 2000. The 'Anthropocene'. Global Change Newsletter 41, 17–18.

Dahal, Smriti, Nepal, Sanjay K., Schuett, Michael A., 2014. Examining marginalized communities and local conservation institutions: the case of Nepal's Annapurna conservation area. Environmental Management 53 (1), 219–230.

Descola, Philippe, 2005. Ecology as Cosmological Analysis. The Land within: Indigenous Territory and the Perception of the Environment, p. 22.

Doane, Molly, 2012. Stealing Shining Rivers: Agrarian Conflict, Market Logic, and Conservation in a Mexican Forest. University of Arizona Press.

Dove, Michael R., 2006. Indigenous people and environmental politics. Annual Review of Anthropology 35, 191–208.

Ehrlich, Paul R., 1968. The Population Bomb: Population Control or Race to Oblivion. Ballantine, New York.

Escobar, Arturo, 2008. Territories of Difference: Place, Movements, Life, Redes. Duke University Press.

Escobar, Arturo, 1999. After nature: steps to an antiessentialist political ecology. Current Anthropology 40 (1), 1–30.

Escobar, Arturo, 1995. Encountering Development: The Making and Unmaking of the Third World. Princeton University Press, Princeton, NJ.

Fairhead, James, Leach, Melissa, 2000. Desiccation and domination: science and struggles over environment and development in colonial Guinea. The Journal of African History 41 (01), 35–54.

Ferguson, James, 1999. Expectations of Modernity: Myths and Meanings of Urban Life on the Zambian Copperbelt. University of California Press.

Filer, Colin, Wood, Michael, 2012. The creation and dissolution of private property in forest carbon: a case study from Papua New Guinea. Human Ecology 40 (5), 665–677.

Gardner, Charlie J., Nicoll, Martin E., Mbohoahy, Tsibara, Oleson, Kristen, Ratsifandrihamanana, Anitry, Ratsirarson, Joelisoa, René de Roland, Lily-Arison, Virah-Sawmy, Malika, Zafindrasilivonona, Bienvenue, Davies, Zoe, 2013. Protected areas for conservation and poverty alleviation: experiences from Madagascar. Journal of Applied Ecology 50 (6), 1289–1294.

Garland, Elizabeth, 2008. The elephant in the room: confronting the colonial character of wildlife conservation in Africa. African Studies Review 51 (03), 51–74.

Gibson-Graham, J.K., Roelvink, Gerda, 2009. An economic ethics for the Anthropocene. Antipode 41 (s1), 320–346.

Grandia, Liza, 2007. Between Bolivar and bureaucracy: the Mesoamerican biological corridor. Conservation and Society 5 (4), 478.

Greenough, Paul, Tsing, Anna (Eds.), 2003. Nature in the Global South. Environmental Projects in South and Southeast Asia. Duke University Press, Durham, North Carolina.

Haenn, Nora, 2005. Fields of Power, Forests of Discontent: Culture, Conservation, and the State in Mexico. University of Arizona Press.

Hartmann, Betsy, 2006. Liberal ends, illiberal means national security, "environmental conflict" and the making of the Cairo consensus. Indian Journal of Gender Studies 13 (2), 195–227.

Harvey, David, 2005. A Brief History of Neoliberalism. Oxford University Press.

Holmes, George, Brockington, Dan, 2012. Protected areas-what people say about well-being. Biodiversity Conservation and Poverty Alleviation: Exploring the Evidence for a Link, 160–173.

Igoe, Jim, Brockington, Dan, 2007. Neoliberal conservation: a brief introduction. Conservation and Society 5 (4), 432.

Jacoby, Karl, 2001. Crimes against Nature, Poachers, Squatters, Thieves and the Hidden History of American Conservation. University of California Press, Berkeley.

Khadka, Damodar, Nepal, Sanjay K., 2010. Local responses to participatory conservation in Annapurna conservation area, Nepal. Environmental Management 45 (2), 351–362.

Kohn, Eduardo, 2013. How Forests Think: Toward an Anthropology beyond the Human. University of California Press.

Leach, Melissa, Fairhead, James, 2000. Challenging Neo-Malthusian deforestation analyses in West Africa's dynamic forest landscapes. Population and Development Review 26 (1), 17–43.

Lewis, David, 2004. On the difficulty of studying "civil society": reflections on NGOs, state and democracy in Bangladesh. Contributions to Indian Sociology 38 (3), 299–322.

Leys, Colin, 1996. The Rise & Fall of Development Theory. Indiana University Press.

Li, Tania Murray, 2007. The Will to Improve: Governmentality, Development, and the Practice of Politics. Duke University Press, Durham, NC.

Masozera, Michel K., Alavalapati, Janaki RR., 2004. Forest dependency and its implications for protected areas management: a case study from the Nyungwe Forest Reserve, Rwanda. Scandinavian Journal of Forest Research 19 (S4), 85–92.

Martinez-Alier, Joan, 1991. Ecology and the poor: a neglected dimension of Latin American history. Journal of Latin American Studies 23 (3), 621–639.

Martinez-Reyes, Jose E., 2014. Beyond nature appropriation: towards post-development conservation in the Maya Forest. Conservation and Society 12 (2), 162.

McAfee, Kathleen, 1999. Selling nature to save it? Biodiversity and green deveiopmentaiism. Environment and Planning 17, 133–154.

Méndez-López, María Elena, García-Frapolli, Eduardo, Pritchard, Diana J., Sanchez González, María Consuelo, Ruiz-Mallén, Isabel, Porter-Bolland, Luciana, Reyes-Garcia, Victoria, 2014. Local participation in biodiversity conservation initiatives: a comparative analysis of different models in south East Mexico. Journal of Environmental Management 145, 321–329.

Mosse, David, 2013. The anthropology of international development. Annual Review of Anthropology 42, 227–246.

O'Connor, James R., 1998. Natural Causes: Essays in Ecological Marxism. Guilford Press.

Peet, Richard, Watts, Michael, 1996. Liberation Ecologies: Environment, Development and Social Movements. Routledge, London.

Peluso, Nancy Lee, 1993. Coercing conservation?: the politics of state resource control. Global Environmental Change 3 (2), 199–217.

Phelps, Jacob, Webb, Edward L., Agrawal, Arun, 2010. "REDD+: When the Stakes Are High'' E-letter Responses to: Charles Palmer REDD+: Property Rights and Liability. Science. www.sciencemag.org/cgi/eletters/328/5982/1105-a (accessed on 05.05.14.).

Rao, Mohan, 1994. An imagined reality: Malthusianism, neo-Malthusianism and population myth. Economic and Political Weekly, PE40–PE52.

Rocheleau, Dianne, Thomas-Slayter, Barbara, Wangari, Esther, 2013. Feminist Political Ecology: Global Issues and Local Experience. Routledge.

Robbins, Paul, 2012. Political Ecology: A Critical Introduction, vol. 20. John Wiley & Sons, West Sussex, UK.

Sachs, Wolfgang, 1993. Global Ecology: A New Arena of Political Conflict. Zed Books, London.

Temudo, Marina Padrão, 2012. "The white men bought the forests": conservation and contestation in Guinea-Bissau, Western Africa. Conservation and Society 10 (4), 354.

Thompson, Mary C., Baruah, Manali, Carr, Edward R., 2011. Seeing REDD+ as a project of environmental governance. Environmental Science & Policy 14 (2), 100–110.

USAID, April 2013. Users Guide to USAID/Washington Health Programs.

Wells, Michael, 1992. Biodiversity Conservation, Affluence and Poverty: Mismatched Costs and Benefits and Efforts to Remedy Them. Ambio, Sweden.

West, Paige, Brockington, Dan, 2006. An anthropological perspective on some unexpected consequences of protected areas. Conservation Biology 20 (3), 609–616.

West, Paige, Igoe, James, Brockington, Dan, 2006. Parks and peoples: the social impact of protected areas. Annual Review of Anthropology 35, 251–277.

West, Paige, 2006. Conservation Is Our Government Now: The Politics of Ecology in Papua New Guinea. Duke University Press.

Whiteman, Gail, Dorsey, Michael, Wittneben, Bettina, 2010. Businesses and biodiversity: they would say that. Nature 466 (7305), 435.

Yamin, F., 2005. The international rules on the Kyoto mechanisms. In: Yamin, F. (Ed.), Climate Change and Carbon Markets: A Handbook of Emissions Reductions Mechanisms. Earthscan, London, pp. 1–74.

Environmental Determinism

Innes M Keighren, Royal Holloway, University of London, Egham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The doctrine of environmental determinism has, for more than two millennia, been used to explain the social organization and physical characteristics of populations. Classical scholars deployed humorism to define a causal relationship between environment and temperament, physiognomy, and intelligence. Enlightenment philosophers used environmental explanations to classify human variation and progress. During the nineteenth century, environmental theories were combined with biological metaphors of development to establish hierarchies of racial difference and to justify imperialism. Environmental determinism was central to the emergence of disciplinary geography and anthropology in the nineteenth century, but was subsequently subject to censure and rejection as a pseudoscience.

Environmental determinism, in its simplest expression, is a principle which states that human activity, culture, and physical and mental characteristics are, at once, informed and inhibited by the geographical and climatic conditions of the physical environment. As a causal mechanism used to describe and explain human development, and the qualities, successes, and failures of individual populations, environmental determinism has a venerable yet contested history. For more than two millennia the concept that geography matters to human affairs has been deployed for a variety of ideological purposes: to describe the geography of morality, to justify imperial processes of colonization, and to construct race as a scientific category.

The intellectual currency of environmental determinism in a variety of disciplinary contexts has varied, however, in relation to perceptions of its explanatory value and pejorative social and political implications. From being an intrinsic component of interpretative work in geography and social anthropology in the late nineteenth and early twentieth centuries, environmental determinism has become passé, even taboo. Notwithstanding the significance to, and validity of, environmental explanations in archaeology, paleontology, and allied disciplines, recent attempts by journalists and popular science writers to reappraise the significance of geographical and climatic influences on human societies have been subject to robust censure by scholars in the humanities and social sciences who regard such work as pseudoscientific at best, and racist at worst (Erickson, 1999; Judkins et al., 2008; Livingstone, 2012; Sluyter, 2003). To understand the vociferousness of contemporary criticism of neoenvironmental determinism, it is necessary to appreciate the concept's historical development and varied expression.

Classical Origins of Environmental Determinism

Environmental determinism, as a philosophy of natural law, has a prehistory which can be traced through the mythologies and theologies of a variety of ancient cultures and religions, not least those of Mesopotamia, the Indus Valley, and what is now China (Glacken, 1967). Philosophical and religious reflection on the role of the physical environment in shaping human culture and civilization has, in a variety of contexts, been central to understandings of societal growth and decline, material cause-and-effect relationships, and the divine (or otherwise) order of nature. The genealogy of environmental determinism in European thought can be traced to the historical, chorographical, and medical discourses of a variety of ancient Greek and Roman scholars who attempted to place understandings of the causal effects of the environment on a rational basis. Among these classical antecedents, Hippocrates of Cos (*c*.460–377 BC), Aristotle (384–322 BC), and Strabo of Amaseia (*c*.63 BC–AD 23) are common ancestors of later deterministic philosophy.

The physiological basis to environmental influence was articulated in its earliest form in a treatise on disease – *On airs, waters, and places* – which constituted part of a collection of medical scholarship (the *Corpus Hippocraticum*) attributed, at least in part, to Hippocrates (Jouanna, 1999). Most probably written to assist itinerant physicians in the correct diagnosis of the diseases they were likely to encounter in different geographical locations, *On airs, waters, and places* was based upon an elemental understanding of the natural world and on the differing effects of earth, air, fire, and water (and their corresponding characteristics: cold, dry, hot, and wet) on the human body. The treatise proposed that an individual's health – and, consequently, personality and physique – was determined by the physical environment's influence upon their bodily humors: the four fluids (blood, yellow bile, black bile, and phlegm) that serve to regulate mood, temperament, and physical well-being. Health was maintained by an equilibrium of the humors; disease promoted by their disequilibrium.

Combined with the persistent influence of climate, topography, soil, and water, humorism was deployed to explain variations in the physiology and mental and moral character not simply of different individuals, but of different populations. The relative superiority or inferiority of populations could thus be explained by reference to the favorable or detrimental characteristics of their physical setting (Wear, 2008). The Hippocratic formulation of environmental influence cast a long shadow across European understandings of human social and physical development and, in one modified form or another, was employed as the self-evident basis to subsequent discourses on law, politics, civilization, and race (among other

topics). The ideological potential (and consequence) of an environmental explanation of human development was evident in Greek scholars' attempts to understand the comparative success of their own civilization.

For Aristotle, the achievements and superiority of the Greek civilization were explained, in large part, by a favorable geographical location and a temperate climate: a 'golden mean' between the too-cold north and the too-hot south (Glacken, 1967: p. 93). In Aristotle's (2013: p. 199) flattering formulation, representing a Mediterranean-centered worldview, Greece was the geographical intermediate between unfavorable climatic binaries:

> The nations in cold locations, particularly in [northern] Europe, are filled with spiritedness, but relatively lacking in thought and art; hence they remain freer, but lack political governance and are incapable of ruling their neighbours. Those in Asia, on the other hand, have souls endowed with thought and art, but are lacking in spiritedness; hence they remain ruled and enslaved. But the stock of the Greeks shares in both – just as it holds the middle in terms of location. For it is both spirited and endowed with thought, and hence both remains free and governs itself in the best manner and at the same time is capable of ruling all, should it obtain a single regime.

Aristotle's interpretation in his *Politics* of Greek success thus served to define the moral boundaries of civilization and barbarism; a demarcation that would subsequently exert a long-standing influence on European conceptions of the association between climate, race, and intellectual capacity.

The relationship between geography and the state was later developed, and further refined, by Strabo in his historical-cultural account of the Roman Empire under Augustus. While Strabo attributed the success and influence of the Empire in part to the peninsular character of Italy (which secured it against attacks), the quality of its natural harbors (which facilitated trade and commerce), and its varied and temperate climate (which offered favorable conditions for agriculture and animal husbandry), he also emphasized the important role of human agency in shaping societal characteristics. In rejecting a straightforwardly climatic account, Strabo argued that social "institutions and education" served to shape "Arts, forms of government, and modes of life" irrespective of the climatic context in which they occurred (Strabo, 1854: p. 155). For Strabo, the physical environment was an explanatory (rather than a deterministic) mechanism to account for the development of culture, states, and their institutions. The differing significance attributed to geography by Strabo, Aristotle, and scholars in the Hippocratic tradition served, in certain respects, to define the lineaments of subsequent debate on environmental causation and determinism and on providence and free will.

Climate, Politics, and Human Types: Environmental Determinism from Magnus to Montesquieu

While the theological cosmology of the Christian church did much to constrain scholarly speculation in Europe during the Middle Ages on the relationship between the natural environment and human culture, echoes of classical understandings of climatic and geographical influence were nevertheless evident in, and often implicit to, the work of various medieval theologians and scholars. Writing in the seventh century, for example, Isidorus Hispalensis (*c.*560–636), archbishop of Seville, attributed the physiology, physiognomy, and character of different populations to their environmental setting: in his formulation, Romans were stately, Greeks shifty, Africans sneaky, and Gauls warlike (Friedman, 2000: p. 42). A tacit assumption as to the modifying influence of the physical environment was apparent, moreover, in the writings of Otto of Freising (*c.*1114–58), bishop and historian, who, in chronicling the invasion of northern Italy by the Lombards during the sixth century, noted that the raiders "gave up their wild customs and adopted Italian ways, partly because they married native women but partly as a result of the nature of the country and climate" (Wright, 1965: p. 232). A similar concern with climatic influence was evident in the work of the Dominican scholar Albertus Magnus (*c.*1200–80) who examined, in his treatise *De natura locorum* (On the nature of places), the causal relationship between the globe's climatic zones and the appearance, skin color, and temperament of their inhabitants.

In certain respects, philosophical reflection on environmental influence reached a particular epitome in the work of the French jurist and political theorist, Jean Bodin (*c.*1530–96). In seeking to provide a rational account of the relationship between history, geography, and the contours of contemporary society (with a view to identifying the system of law and sovereignty most suited to specific nations), Bodin synthesized the environmental theories of classical scholars and medieval theologians. Bodin sought in his wide-ranging text – *Six livres de la république* (The six books of the commonwealth, 1576) – to demonstrate that the earth *directed* the theater of human life. Bodin combined humorism with geographical and astrological explanations to account for human character and actions. Like Albertus Magnus, Bodin took as self-evident the existence of global climatic regions (frigid, temperate, and torrid zones) which, through humoral mechanisms, exerted specific influence on the physical and mental characteristics of those regions' inhabitants.

Bodin looked, moreover, beyond the terrestrial environment to assess the influence of the celestial realm upon society, culture, and individual disposition. To Bodin, the inhabitants of the southern parts of the globe, governed by Saturn, were, for example, predisposed to religious contemplation, while their northern counterparts, ruled over by Mars, were warlike. With strong echoes of the golden mean identified by Aristotle in his *Politics*, Bodin contended that the people of the intermediate latitudes – subject to the influence of Jupiter – were of even temperament, combined the most favorable characteristics of northern and southern populations, and were (by virtue of their geographical situation) inclined to rule over others. While Bodin considered environmental and astrological influences to be persistent, even inexorable, he, like Strabo, nevertheless allowed for the possibility that they could be resisted through determined and conscious social effort.

Bodin's deterministic philosophy had a variety of inheritors, not least the French political philosopher Montesquieu (1689–1755) – arguably the Enlightenment's most influential

proponent of environmental causation. Montesquieu's principal intellectual ambition – outlined in his treatise *L'esprit des lois* (The spirit of laws, 1748) – was to develop a rigorous framework by which human behavior could be explained by reference to morals, customs, and laws as well as to environmental context. From a systematic analysis of empirical data on the environmental and social characteristics of different populations and geographical locations – furnished, in part, by the Enlightenment's burgeoning travel literature – Montesquieu advanced a multicausal explanation which distinguished between environmental and cultural influences on human organization. Montesquieu drew attention, in so doing, to the 'internal (social and psychological)' and 'external (biophysical and technological)' milieux in which human life was enacted and by which it was shaped (Buttimer, 1971: p. 12).

Montesquieu's analysis was, in this respect, more nuanced than was that of Bodin, but he nevertheless considered the physical environment, and climate specifically, to be 'the first and most powerful' of explanatory mechanisms (Withers, 2007: p. 140). While the basic tenets of Montesquieu's theory were not new, his contribution to understandings of human society made it possible for Enlightenment scholars to see human progress as a fact of geography, not simply a question of history. Understood as a function of climate and of geographical situation, the assumed progression of human populations from savage and barbarous states, to commercial and polite (as they would later be categorized by Adam Ferguson (1723–1816)), could thus be seen, at each stage of development, contemporaneously in different locations on the terrestrial globe. Montesquieu's work thus exerted, in a variety of modified forms, an important influence on Enlightenment understandings of progress and of the assumed nature of human difference (Livingstone, 2011).

The Enlightenment concern with human difference and development reflected not only that era's classificatory impulse – the desire to impose order and regularity on the dizzying variety of the natural world – but also a more elemental interest in human origins. Quite apart from the physical and social mechanisms which might explain human difference, there was significant debate as to whether the indices by which that difference might be measured (skin color, stature, skeletal form, hirsuteness, among others) were innate characteristics or superfluous by-products (Hudson, 1996). In his *Systema naturae* (The system of nature, 1735), for example, Carl Linnaeus (1707–78) described a fourfold classification of human types based upon their continental distribution and skin color: white Europeans, red Americans, yellow Asians, and black Africans. While Linnaeus and many eighteenth-century theorists of race regarded these categories as fixed, bound to classical understandings of the humors and to defined geoclimatic zones, others, such as Buffon (1707–88), saw human difference as a rather more variable and dynamic phenomenon: the consequence of biological lineage and movement across space, as well as climatic influence.

For Buffon, who believed that all human populations shared a common (European) origin, dissimilarities were to be explained by the persistent influence of different climatic regimes. The relative superiority or inferiority of individual groups was not, then, the consequence of inherent qualities, but of the capacity of the environment to 'degrade' or 'elevate' those populations in moral and physical terms. A population's traits, both desirable and disagreeable, could be gained and lost under the influence of differing climatic contexts: Linnaean categories were not fixed, but mutable. Buffon's revision to Linnaeus's scheme notwithstanding, climatic understandings of human variation – developed further by Immanuel Kant (1724–1804), Johann Friedrich Blumenbach (1752–1840), among others – served to define a hierarchy of human types based upon the favorable (or otherwise) qualities of their specific geoclimatic situation, which always defined Europeans as its epitome. Although climatic explanations of human variety were not universally agreed upon by Enlightenment scholars – Voltaire (1694–1778), Henry Home (1696–1782), and David Hume (1711–76) were each critical in different ways of Montesquieu's attention to environmental influence – they were nevertheless central to marking out the contemporary globe's moral geography and to attempts during the nineteenth century to interpret race as a scientific category (Jackson and Weidman, 2004).

Biological Metaphors and Evolutionary Mechanisms in Environmental Determinism

Enlightenment-era debate concerned with human difference and progress was, at base, defined by the relative significance attributed to the social and the environmental as explanatory factors. While the interactions of populations and their environment had been considered at various stages during the eighteenth century as a reciprocal phenomenon (i.e., that humans has the capacity to modify their physical environment through processes of agriculture, deforestation, and so on), it was not until the nineteenth that this coconstitutive relationship was addressed systematically. It was done so, in part, by the Prussian naturalist and traveler Alexander von Humboldt (1769–1859). A five-year exploration of topical America between 1799 and 1804 served as the empirical foundation to Humboldt's philosophy of unity in nature; an approach which sought to describe the interrelatedness and codependence of natural and cultural phenomena through the biological metaphor of the organism.

Like earlier contributors to debates on environmental influence, Humboldt attributed significance to climatic zones (albeit defined in a more systematic manner than Magnus and Bodin were capable of). Humboldt's was, however, not a deterministic interpretation. Articulated most fully in his five-volume magnum opus *Kosmos* (published between 1845 and 1862), Humboldt viewed environmental factors as influences on human life rather than as determinants of it:

> Dependent, although in a lesser degree than plants and animals, on the soil, and on the meteorological processes of the atmosphere with which he is surrounded – escaping more readily the control of natural forces, by activity of mind and the advance of intellectual cultivation, no less than by his wonderful capacity of adapting himself to all climates – man everywhere becomes most essentially associated with terrestrial life.
>
> Humboldt, 1849: p. 360–361

In Humboldt's ecological rendering, the physical environment was not a yoke under which human populations labored,

but was the means by which they were nurtured and thus a resource to be revered and protected. Together with his near contemporary, Carl Ritter (1779–1859), Humboldt did much to place the study of human–environment relations on a systematic basis (Martin, 2005).

Humboldt's organismic understanding was further refined by the theories of biological evolution advanced by Jean-Baptiste Lamarck (1744–1829) and Charles Darwin (1809–82). While Lamarck and Darwin proposed distinct mechanisms to explain evolutionary change in species – and held dissimilar views as to progress, direction, and order in respect to that change – each attributed direct significance to environmental and climatic factors in accounting for transmutation. In simple terms, Lamarck ascribed evolutionary change to the inheritance of characteristics (i.e., adaptations to the environment) acquired during the lifetime of the individual organism. Darwin, meanwhile, understood evolution to be a rather slower, more stochastic process whereby random mutations might confer on the organism an advantage in dealing with its environmental context. Those advantages would increase the survival and reproductive success of the individual, and be passed on to the following generations more often than would less-favorable characteristics.

In various modified forms, the ideas of Lamarck and Darwin were applied, most particularly through the work of Herbert Spencer (1820–1903), to further develop understandings of human social and cultural evolution in relation to environmental circumstances. Spencer's characterization of natural selection as 'the survival of the fittest' encouraged further attention among scholars to questions of environmental causation, often with pejorative consequences. Social Darwinism, as Spencer's reading of evolutionary theory was characterized, was used, variously, to support laissez-faire economics, to justify imperialism, to define hierarchies of racial superiority, and to validate eugenics (Hawkins, 1997). While the tenets of Social Darwinism were not fundamentally dissimilar from those of previous theories of environmental causation, the application of what were assumed to be scientific principles to its investigation meant that its influence across geography, anthropology, zoology, and medicine, among other disciplines, was considerable (Livingstone, 2012; Peet, 1985).

Environmental Determinism and the Academy: From Humboldt to Huntington

The construction, during the late nineteenth century, of human evolution as a question of geographical and environmental determinants was central to the development of academic geography as the study of the causal relationships between organic and inorganic phenomena. Although many European geographers (or protogeographers) regarded the systematic elucidation of cultural–environmental relationships to be the discipline's scientific raison d'être, there was nevertheless considerable disagreement (reflecting Montesquieu's long-standing society–milieu distinction) regarding the relative significance attributed to social processes and to environmental contexts in explaining human character and change. This debate was characterized by an oppositional binary between deterministic explanations (associated with work of Friedrich Ratzel (1844–1904) and Ellen Churchill Semple (1863–1932), among others) and possibilistic interpretations (advanced, among others, by Paul Vidal de la Blache (1845–1918) and Carl Sauer (1889–1975)).

Drawing inspiration from the work of Oscar Peschel (1826–75) and Ferdinand von Richthofen (1833–1905) on the comparative study of landforms, and from Moritz Wagner (1813–87) on migration and organic diversity, Ratzel sought to put the study of human geography (*anthropogeographie*) on a scientific basis through a systematic analysis of the biotic and cultural features of different social groups in relation to their environments. For Ratzel, civilization was the product and consequence of the interrelationship of culture and environment, not a separate and independently arrived-at phenomenon. In this respect, Ratzel saw the evolution of human societies to depend upon a Spencerian struggle for survival, with those populations most adept at responding to the challenges and opportunities of their environmental circumstances being ultimately successful. Ratzel employed the concept of living space (*Lebensraum*) in describing an organismic model of human society that developed "within a frame (*Rahmen*), exploiting a place (*Stelle*), needing space (*Raum*) and finding limits (*Grenzen*)" (Wanklyn, 1961: p. 23). Ratzel's manifesto met with a varied response among contemporary scholars and engendered both deterministic and possibilistic readings of the environment's likely influence on society. Ratzel's concept of *Lebensraum* – which tied the health and success of a population to the size and quality of its habitat – was later adapted and distorted in Germany, under the banner of *geopolitik*, by the National Socialist Party in support of Nazi spatial ideology (Abrahamsson, 2013; Smith, 1980).

The transmission of Ratzel's theory of environmental influence to the Anglophone world was facilitated, in large measure, by his principal acolyte, the American geographer Ellen Churchill Semple. For Semple, the central appeal of Ratzel's model lay in its potential to define the academic legitimacy of geography – to transform it from an idiographic discipline (focused on the description of specific instances and phenomena) to a generalizing, nomothetic science, one capable of defining natural laws and rules. Where Ratzel had been subject to criticism for the inadequate empirical basis to his model, Semple supplemented her interpretation of it – published in 1911 as *Influences of geographic environment on the basis of Ratzel's system of anthropo-geography* – with more than 1000 citations of classical authorities and contemporary scholarship. In this respect, Semple was more systematic in her evidential support than was Ratzel and, in so being, presented an exhaustive range of environmental explanations to account for almost every conceivable aspect of human form, character, and social organization – from religion and architecture, to economic system and moral character (Cresswell, 2013; Keighren, 2010).

Notwithstanding the range of geographical influences which Semple invoked, her stated (although arguably unsuccessful) aim was to deny a straightforwardly deterministic relationship between the natural environment and human social and physiological organization. While her prefatory remarks made this ambition clear – "the writer speaks of geographic factors and influences, shuns the word geographic determinant, and speaks with extreme caution of geographic

control" – it is evident that for many contemporary and later readers, Semple's philosophy was interpreted quite differently and regarded as the epitome of environmental determinism (Semple, 1911: p. vii). Where Semple did distinguish herself from earlier scholars of environmental causation was, however, in her treatment of human different. Semple was disinclined (at least superficially so) to use race as an explanatory category. Semple's argument was that if peoples of different ethnic stock (a term she used to encompass a population's inherited traits), who were exposed to similar environmental circumstances, were seen to display similar social characteristics, these similarities were self-evidently the consequence of environmental influence and not of inherent racial differences. Semple's opinion, like that of Montesquieu and Buffon, was that environmental circumstances were ineluctably the determinant of human difference and thus had the potential to elevate or to degrade populations irrespective of their 'stock.' While Semple proceeded with caution in her discussion of race, several of her contemporaries in Anglophone geography were less circumspect.

The work of Ellsworth Huntington (1876–1947) and Thomas Griffith Taylor (1880–1963) is representative of a trend in Anglophone geography in the first half of the twentieth century to link, on an empirical and quantitative basis, climate, race, and the development of human civilization. Whilst there was little in such arguments to distinguish them, in a fundamental sense, from those of Bodin and his inheritors, what Huntington and Taylor attempted to do differently was to place such interpretations of human-environment relationships on a statistical and apparently objective basis. Huntington and Taylor each collected vast quantities of climatic and anthropological data in developing their understandings of the links between climate (which they regarded as the primary determinant) and human intelligence, physical appearance, and moral potential. In texts such as *The Character of Races* (Huntington, 1924) and *Environment and Race* (Taylor, 1927), laid the foundation to a positivistic and scientific construction of racial categories which – although subject to censure by contemporaries such as the sociologist Franz Boas (1858–1942) – retained currency throughout much of the first half of the twentieth century and was only permanently gainsaid as a consequence of developments in the science of genetics (Hannaford, 1996).

Environmental Determinism Redux?

For geographers, social anthropologists, and other scholars of the social sciences and humanities, environmental explanations for human difference and development were, by the middle decades of the twentieth century, largely considered to be obsolete – symbolic of scientific racism and profoundly tarnished by their deployment in support of Nazi geopolitics. Nevertheless, the doctrine of environmentalism (in its deterministic and possibilistic stripes) was never entirely eliminated and has found continued explanatory validity in biological anthropology and archaeology while also exerting a certain commonsensical appeal to those social scientists for whom the wholesale rejection of environmental explanations during the second half of the twentieth century was akin to throwing out the baby with the bathwater (Merrett, 2003).

Recent contributions to neoenvironmental determinism – such as Jared Diamond's *Guns, Germs, and Steel* (1997) and *Collapse* (2005), and Robert Kaplan's *The Revenge of Geography* (2012) – have sought to draw attention to the explanatory significance of geography and of climate in accounting for human populations' relative successes and failures. For their critics, Diamond and Kaplan are guilty of uncritically repackaging deterministic philosophy – reinvigorating ideas that previously were used to construct hierarchies of human difference and to justify damaging political ideologies – and of failing to "break free … from an environmental determinist straightjacket that sees geography as nothing more than a three-dimensional physical manifestation" (Johnston, 2013: p. 3). What such criticisms reveal is quite how polarizing environmental explanations (rendered toxic by their earlier socially and scientifically dubious iterations) remain for geographers, among others. Rather than heralding a revivified debate, the response to Diamond and Kaplan suggests that existing positions are being retrenched.

See also: Adaptation, Fitness, and Evolution; Archaeology, Theory in; Boas, Franz (1858–1942); Body, History of; Cultural Evolution: Theory and Models; Darwin, Charles Robert (1809–82); Darwinism: Social; Darwinism; Ecumene; Environmental History; Ethnic, Racial, and Nationalist Movements; Ethnicity and Ethnic Groups: Historical Aspects; Ethnocentrism; Eugenics, History of; Feudalism; Geographic Thought, History of; Geography: Overview; Human Ecology; Hume, David (1711–76); Kant, Immanuel (1724–1804); National Character; National Socialism and Fascism; Nature and Society in Geography; Progress: History of the Concept; Race and Racism in the Twenty-First Century; Racism, History of; Sauer, Carl Ortwin (1889–1975); Social Structure: History of the Concept; Spencer, Herbert (1820–1903); Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Abrahamsson, C., 2013. On the genealogy of Lebensraum. Geographica Helvetica 68, 37–44.

Aristotle, 2013. Aristotle's Politics (C. Lord, Trans.), second ed. University of Chicago Press, Chicago.

Buttimer, A., 1971. Society and Milieu in the French Geographic Tradition. Rand McNally and Company, Chicago.

Cresswell, T., 2013. Geographic Thought: A Critical Introduction. Wiley-Blackwell, Oxford.

Diamond, J.M., 1997. Guns, Germs, and Steel: The Fates of Human Societies. W.W. Norton, New York.

Diamond, J.M., 2005. Collapse: How Societies Choose to Fail or Succeed. Viking, New York.

Erickson, C.L., 1999. Neo-environmental determinism and agrarian 'collapse' in Andean prehistory. Antiquity 73, 634–642.

Friedman, J.B., 2000. The Monstrous Races in Medieval Art and Thought. Syracuse University Press, Syracuse.

Glacken, C.J., 1967. Traces on the Rhodian Shore: Nature and Culture in Western Thought from Ancient Times to the End of the Eighteenth Century. University of California Press, Berkeley.

Hannaford, I., 1996. Race: The History of an Idea in the West. Johns Hopkins University Press, Baltimore.

Hawkins, M., 1997. Social Darwinism in European and American Thought, 1860–1945: Nature as Model and Nature as Threat. Cambridge University Press, Cambridge.
Hudson, N., 1996. From "nation" to "race": the origin of racial classification in eighteenth-century thought. Eighteenth-Century Studies 29, 247–264.
Humboldt, A., 1849. Cosmos: A Sketch of a Physical Description of the Universe (E.C. Otté, Trans.). Henry G. Bohn, London.
Huntington, E., 1924. The Character of Races as Influenced by Physical Environment, Natural Selection and Historical Development. Charles Scribner's Sons, New York.
Jackson, J.P., Weidman, N.M., 2004. Race, Racism, and Science: Social Impact and Interaction. ABC-CLIO, Santa Barbara.
Johnston, R., 2013. Review of the Revenge of geography, by Robert Kaplan. The AAG Review of Books 1, 1–3.
Jouanna, J., 1999. Hippocrates. Johns Hopkins University Press, Baltimore.
Judkins, G., Smith, M., Keys, E., 2008. Determinism within human-environment research and the rediscovery of environmental causation. The Geographical Journal 174, 17–29.
Kaplan, R.D., 2012. The Revenge of Geography: What the Map Tells Us about Coming Conflicts and the Battle against Fate. Random House, New York.
Keighren, I.M., 2010. Bringing Geography to Book: Ellen Semple and the Reception of Geographical Knowledge. I.B. Tauris, London.
Livingstone, D.N., 2011. Environmental determinism. In: Agnew, J., Livingstone, D.N. (Eds.), The SAGE Handbook of Geographical Knowledge. SAGE, London, pp. 368–380.
Livingstone, D.N., 2012. Changing climate, human evolution, and the revival of environmental determinism. Bulletin of the History of Medicine 86, 564–595.
Martin, J., 2005. All Possible Worlds: A History of Geographical Ideas, fourth ed. Oxford University Press, Oxford.
Merrett, C.D., 2003. Debating destiny: nihilism or hope in Guns, Germs, and Steel? Antipode 35, 801–806.
Peet, R., 1985. The social origins of environmental determinism. Annals of the Association of American Geographers 75, 309–333.
Semple, E.C., 1911. Influences of Geographic Environment on the Basis of Ratzel's System of Anthropo-geography. Henry Holt, New York.
Sluyter, A., 2003. Neo-environmental determinism, intellectual damage control, and nature/society science. Antipode 35, 813–817.
Smith, W.D., 1980. Friedrich Ratzel and the origins of Lebensraum. German Studies Review 3, 51–68.
Strabo, 1854. The Geography of Strabo (H.C. Hamilton, W. Falconer, Trans.). Henry G. Bohn, London.
Taylor, G., 1927. Environment and Race: A Study of the Evolution, Migration, Settlement and Status of the Races of Man. Oxford University Press, New York.
Wanklyn, H., 1961. Friedrich Ratzel: A Biographical Memoir and Bibliography. Cambridge University Press, Cambridge.
Wear, A., 2008. Place, health, and disease: the airs, waters, places tradition in early modern England and North America. Journal of Medieval and Early Modern Studies 38, 443–465.
Withers, C.W.J., 2007. Placing the Enlightenment: Thinking Geographically about the Age of Reason. University of Chicago Press, Chicago.
Wright, J.K., 1965. The Geographical Lore of the Time of the Crusades: A Study in the History of Medieval Science and Tradition in Western Europe. Dover Publications, New York.

Environmental Economics

Vincent K Smith, Arizona State University, Tempe, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes the field of environmental economics. It outlines the conceptual problems posed by environmental externalities and public goods and describes how policy addresses them. The logic of benefit–cost analysis and its role in evaluating environmental regulations are discussed. The newest methods of quasi-experimental methods for estimating the effects of rules are summarized. The distributional effects of environmental change are also discussed along with a brief summary of the challenges in maintaining economic growth while assuring that environmental quality is maintained.

Environmental economics studies the effects that environmental systems have on the positive predictions and normative recommendations of economic models (see Solow, 1991; Arrow et al., 1995). Early in the history of economic modeling, Hotelling (1931) recognized the importance of intertemporal issues for economic choices involving renewable and nonrenewable natural resources. Nearly 50 years later, Dasgupta and Heal (1979) and numerous other economists (see Smith, 2012, for a review of the new economic research on fisheries) have extended this research. Today, there is extensive theoretical and empirical literature on these topics that are treated separately in the *Encyclopedia*.

Environmental economics assumes there are things people care about, either directly because they enhance individual well-being or indirectly because they influence the production of other commodities, that are (1) scarce; (2) available outside markets; and (3) influenced by the activities of others. These objects of environmental choice (OECs) have diverse characteristics and result from a variety of activities. Seven topic areas organize this overview: the allocation problems they pose, the design of policy, traditional and new approaches for measuring trade-offs that people would make to enhance environmental amenities, OECs as risks, the distributional implications of environmental changes, and interactions between growth and environmental resources.

Problems – Public Goods, Externalities, and Congestion

Pure public goods are nonrival and nonexcludable in consumption (Cornes and Sandler, 1996; Phaneuf and Requate, forthcoming). Each person's consumption of a public good does not reduce the amount available to others. Moreover, access to these goods cannot be limited. While it is difficult to provide examples satisfying these ideal conditions, it is easy to cite examples of goods and services with attributes close to a pure public good's features. Local air quality, measured in terms of the concentration of airborne pollutants, varies with location but within large areas it is approximately the same for all individuals living there. The global concentrations of greenhouse gases (e.g., CO_2, methane, etc.) can be described as examples of pure public 'bads' because emissions of them at any location have equivalent effects on their atmospheric stocks. However, the consequences of those accumulations for climate change and local weather conditions will depend on where people live. Perfect adherence to the ideal definition of a public good is not essential to consider the implications of 'publicness' for the efficiency of different resource allocation mechanisms.

Externalities correspond to specific types of interactions between two or more economic agents. One group's (or individual's) consumption, production, or exchange activities have unintended effects on the production or consumption of others. These external effects resemble public goods (or bads) in that they can have the same nonrival and nonexcludable attributes.

Congestion describes a phenomenon whose externality effects reduce the nonrival attributes of a public good. For example, over some levels of recreational use of a beach or a wilderness area (and the levels will be different in each case) further users can be supported without extra cost or diminution in the satisfaction each person realizes from his or her use. However, at some point, this simultaneous use diminishes each person's experience. The phenomenon linking those losses to the added users of a resource at the same time is described as congestion. While all three of these concepts are relevant outside environmental applications, they have found special meaning when environmental resources provide the services treated as OECs.

These concepts are important for at least two reasons. First, the incentives created by private markets do not assure that resources will be allocated to their highest valued uses when one or more of these problems is involved. Second, the information available from markets cannot be assured to signal and evaluate the performance of economic institutions. Much of environmental economics has been developed to address these issues.

Policies – Context, Design, and Performance

Economic analysis focuses on the incentive properties of each decision context. In the case of environmental problems these details can matter to the insights offered by economic analysis. For example, economic descriptions of production and consumption activities must be consistent with materials and energy balances (Kneese and Bower, 1968). When they are not,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71038-5

it is easy to overlook the residuals and the problems that must be faced in disposing them. This insight is important because most externalities arise from using the services of an environmental asset (i.e., the ambient atmosphere or a free flowing river) for residual disposal. The form and severity of environmental externalities depend on both the types of residuals involved and the functions of the ecosystem receiving them. Many ecosystems are available with unregulated access similar to an open access resource. At some levels of use, their services may well also be subject to depletion over time.

Economic prescriptions for reducing negative externalities (as well as for congestion) focus on creating the 'right' incentives. Private choices should reflect their full or social incremental cost (i.e., the private marginal cost of producing any commodity plus the marginal external costs due to any damages associated with a commodity's production, consumption, and ultimate disposal). When the marginal damages associated with the production or consumption of a good are measured per unit of each residual emitted, it is sometimes described as the price that should be charged for that emission. Pigouvian taxes or effluent charges set equal to these marginal damages are one type of policy response to externalities.

Under simplified conditions, efficient resource allocation in the presence of environmental externalities (or congestion) requires that, at the margin, agents face the full incremental cost (or incremental benefit) of their actions including the damages caused directly or indirectly from each choice. Those experiencing damages are assumed to take defensive actions, and costs are computed recognizing these choices, but the injured parties do not need to be compensated (Baumol and Oates, 1988; Phaneuf and Requate, forthcoming).

With full information, policy can take a variety of forms. These can be charges per equivalent unit of impact (taking account of the ecosystem) to the agents causing externalities; permits to emit pollution that can be traded; and even subsidies to abate emissions. Each of these policy measures can be made to appear equivalent in terms of the incentives for incremental choices. Each approach requires different implementation rules and a different set of adjustments to reflect their effects at the extensive margin of choice (e.g., the entry and exit decisions of firms). Charges vs subsidies will affect incentives to enter an industry (Carlton and Loury, 1980; Goulder and Parry, 2008). Tradable emission permits must be denominated in units reflecting the effects of emissions on the services associated with ambient environmental quality that defines the OECs. They must also consider the initial distributions of permits among firms, the rules for trading, and their ability to be saved (if temporal patterns of emissions are important) (Tietenberg, 2006). Interactions between pollutants when they are dispersed in environmental media can interfere with the performance of permit systems designed to respond to a single pollutant under specific conditions (Evans and Woodward, 2013).

Unfortunately, in the setting where actual decisions must take place, information is scarce and uncertainty abounds. If uncertainty affects the marginal control costs, then it can affect the preferred policy instrument. Comparing charges vs marketable permits (denominated in units of the OEC), Weitzman (1974) has shown that the choice depends on the relative steepness of the marginal benefit and cost schedules. Tradable permits are favored if the marginal benefit function is steeper than marginal abatement cost and effluent charges for the reverse. His results depend on the approximate linearity of these schedules. Moreover, the conclusion is a relative statement, not an absolute judgment. Composite strategies using price and quantity instruments together can be shown to be better in the sense that they enhance overall welfare (Roberts and Spence, 1976; Pindyck, 2007).

With perfect information either instrument can, in principle, lead to an efficient resource allocation. Charging for emissions or distributing permits to pollute seems to convey 'rights' to different economic agents to use ecosystems' services. The conclusion that they can be equivalent may seem counterintuitive. While surprising, it is one reflection of the general implications of the Coase (1960) theorem. That is, in a world of perfect information, no transaction costs and no income effects (from assigning rights to ecosystem services), the theorem implies that the assignment of rights will not affect the efficient outcome. It is often used in arguments that externalities can be resolved without public sector intervention by assigning property rights to ecosystem services. In most cases, the practical relevance of this conclusion is an empirical question. Nonetheless, the economic logic underlying the theorem also explains why the available policy instruments can be equivalent.

Environmental policy in the US is the result of rule making that translates the broad goals of environmental legislation into regulations. Prior to the 1990 Clean Air Act Amendments (CAAA), rules favored command-and-control (CAC) regulations. Specific technologies were prescribed to meet objectives such as using the best available control technology to establish the lowest emission rate. These types of engineering standards specified, at the unit process level, emission rates for the same pollutant that could turn out to be different at different stages of production in the same plant. The result was very different incremental costs for controlling the same pollutants even within a single facility. Abatement costs have been increasing over time as more stringent controls were imposed. Areas found to be in nonattainment with environmental standards faced restrictions on further economic activity. The result was a policy environment seeking some ways of increasing the flexibility in how firms could respond. The 1990 CAAA changed the US regulatory climate. It offered the first large-scale emission permit trading system as a way of realizing more stringent controls for sulfur dioxide. Ellerman et al. (2000) estimate that emission trading yielded cost savings by 33–67% of the most reasonable CAC alternative. Permit trading schemes have been used in small regional markets for an array of environmental resources – including air and water pollutants, wetlands, habitat for endangered species, and carbon dioxide (see Tietenberg, 2006; Evans and Woodward, 2013).

The socially efficient consequences, considering both the private costs and the external effects of production and consumptions, of most market goods and services cannot be analyzed without recognition of their environmental consequences (see Muller et al., 2011). This requirement is symmetrical. Environmental policy analysis cannot ignore the market and regulatory context for the products produced along

with the environmental externalities. Two sets of insights have been especially important to reaching this conclusion. The earliest, associated with Buchanan (1969), argued that a monopolistic product market may, by restricting output, serve environmental goals. In this context, imposing effluent charges or restricting emissions with tradable quotas could actually lead to welfare losses.

The logic underlying his result is more general. Judgments about the preferred incentives to realize an efficient resource allocation in an economy with externalities and multiple distortions are potentially complex. One initial stimulus to consider policy design in the presence of existing distortions was the promise of a 'double dividend' from taxes on CO_2 emissions. The logic would rely on using revenues from environmental taxes to reduce marginal tax rates of large, preexisting taxes with significant welfare distortions while maintaining tax revenue. Subsequent research (see Goulder et al., 1999; Bovenberg and Goulder, 2002; Parry et al., 2012) has established that, ignoring environmental benefits, welfare gains are possible with some revenue-neutral switching of environmental taxes for other factor related taxes. In a dynamic model, switches with capital taxes are more likely to lead to welfare gains than switches with labor taxes. Evaluating these revenue-neutral strategies to the use of environmental taxes within deficit reduction policy, Carbone et al. (2012) compare policies using increases in capital, labor, or consumption taxes alone vs using a carbon tax with one of these. Substituting a carbon tax for a capital tax is most efficient (i.e., smallest welfare losses) and for consumption tax the least. Barrage (2012) adds to these findings with a stochastic dynamic general equilibrium model focusing on optimal carbon taxes with preexisting distortions and finds that carbon taxes must be higher if they are undertaken with suboptional preexisting distortion taxes.

Amounts: The Role of Nonmarket Valuation and Benefit–Cost Analysis

Nearly all analyses of policy instruments assume that we have the information needed to decide how much environmental quality to provide. Answering this question involves more than deciding the amount of pollution permitted. It requires that we consider a wide range of natural assets whose amenity services are important to people. Everything from endangered species and their habitat to land and water resources that provide a wide range of outdoor recreation must be considered. Some choices must be made more carefully than others because some allocations of a natural environment will irreversibly alter its ability to provide amenity services (Krutilla, 1967; Krutilla and Fisher, 1975; Pindyck, 2007).

The extent of 'publicness' of the OECs compounds the difficulties in making these choices. Free-riding incentives arising from nonexclusion assure that private action alone cannot be relied upon to realize the efficient level of these types of resources (Cornes and Sandler, 1996; Phaneuf and Requite, forthcoming). Recovering information equivalent to what is provided when market exchanges correctly signal scarcity requires economic 'detective work'. Two strategies have been developed since Hicks (1943) first defined rigorously the monetary representations for the trade-offs defined by economic choices (Freeman, 2003). The first relies on the logic of revealed preference and uses choices of private goods to recover the trade-off information. For example, it recognizes that a consumer's decision to purchase (or rent) a home in an area free of air pollution reflects, in part, a payment for that ambient quality in comparison to the next best alternative (Palmquist, 2005). With a wide array of alternative conditions, the price gradients arising from different air quality conditions, with other housing attributes held constant, would reveal the incremental willingness to pay for improving air quality (Freeman, 2003). Similar logic implies that one could use the trips taken to recreation sites at locations with different travel costs to reveal the demand for the recreation these sites provide (Freeman, 2003). Given a set of recreation sites, varying in quality conditions, consumers' visitation decisions offer the potential to measure the incremental value of those quality dimensions. Decisions to avert or mitigate the negative consequences of pollution offer the same potential to reveal underlying trade-offs (Freeman, 2003).

The logic underlying each strategy relies on reconstructing the unobserved elements required to define the economic trade-off revealed with each choice. In the case of nonmarketed goods this process implies that a connection is also defined between the decision and the OEC of interest. For the trade-off to be informative about an individual's preferences for the nonmarketed good these connections must convey a degree of exclusivity (i.e., economic value of the amenity requires consumption of the linked private good). They must also specify how enhancements in an individual's ability to pay would change them. One way this control over the substitution and income effects of choice is implemented comes from the joint imposition of restrictions on the properties assumed in describing an individual's utility function. Weak complementarity (Mäler, 1974) and the Willig (1978) condition are among the most commonly used. Weak complementarity maintains that the environmental amenity is of value when it is used with the private good. Without some positive consumption of that good, an improvement in the amenity has no value to each person. The Willig condition adds to this restriction a requirement that the incremental value of increases in the amenity, per unit of the private good consumed, does not change with income. This condition together with weak complementarity is sufficient to allow willingness to pay for discrete changes in amenities linked to the private good. Bullock and Minot (2006) have demonstrated that numerical methods can, in some circumstances, allow it to be relaxed.

Stated preference approaches to recovering economic trade-offs offer the second nonmarket valuation strategy. They rely on presenting people believable (but hypothetical) choices with proposed financial consequences and interpreting their stated decisions as bounding economic trade-offs (Mitchell and Carson, 1989; Hanemann, 1994). An important motivation for proposing choices arises from the inability to link all the ways amenities enhance people's well-being to observable choices of private goods. Such unlinked reasons are sometimes referred to as nonuse or existence values (Krutilla, 1967). Stated preference methods are controversial

because they require that survey responses are to be treated as economically consequential (Carson, 2012; Kling et al., 2012; Hausman, 2012).

The primary use for nonmarket valuation estimates has been in benefit–cost analyses (BCAs) of projects involving environmental resources and regulations intended to improve environmental quality. BCAs assume that there are reasons why an existing resource allocation is not efficient. The analysis seeks to evaluate whether an aggregate choice improves the efficiency (Arrow et al., 1996). Four aspects of applying BCA are especially important. First, the net benefits from an aggregate choice are summarized as a present value and are influenced by the time horizon and discount rate used. Both are potentially influential and controversial judgments (see Portney and Weyant, 1999 Sterner and Persson, 2008; Pindyck, 2007). Second, large changes in relation to the overall economy have general equilibrium effects that may require evaluating the effect at the household level and aggregating (i.e., computing net benefits at the household level rather than aggregating benefits and costs separately, see Hazilla and Kopp, 1990; Carbone and Smith, 2013). Third, BCA will always face incomplete information about the changes being evaluated from a baseline to a new outcome. When an irreversible choice is being made and there is the possibility for resolving some of the uncertainty over time, the analysis should include the conditional value of any new information that would be consequential to the decision (Arrow and Fisher, 1974). Finally, monetization of all factors creating benefits or costs is not essential for BCA to be informative. Observed choices (at the individual level) define bounds for the trade-offs implied by a decision. This same logic can be used to delineate bounds for the missing elements in a BCA. That is, we define what the missing elements in a measured value for net benefits would need to be to support a specific choice (Krutilla and Fisher, 1975).

Research Design, Quasi-Experiments, and Measurement

An important lesson that economics brings to the evaluation of policy is that people's behavior matters. The newest literature on policy evaluation in environmental economics has emphasized the importance of research design (see Chay and Greenstone, 2003, 2005; Greenstone and Gallagher, 2008; Davis, 2004; Schlenker and Walker, 2011). Does the empirical strategy used to estimate the effects of pollution on some outcome, whether a health effect or a property value, distinguish this influence from others that may be related? This relationship can arise because other factors are spatially linked or because there are unobservable features of the unit defining each observation that could also affect the specific outcome of interest. This identification problem has been shown to be especially important for environmental applications.

There are two key questions identified in this literature: (1) how can the effects of amenities and disamenities be reliably measured? and (2) when can the estimates from reduced form quasi-experiments, designed to address the first question, be used to evaluate the net gains from new policies? (Kuminoff and Pope, 2014).

Risk: Perceptions and Values

One of the most important classes of OECs involves environmental sources for changes in the probabilities of undesirable outcomes. For example, air pollutants have been associated with premature death and increased episodes of serious illness. These decisions are highlighted for two reasons. First, the framework for evaluating private and public choice must consider attitudes toward risk and the opportunities available to adjust to it. Within a framework that evaluates decisions made before the uncertain events are resolved, there has been considerable discussion of how these risks should affect monetary measures of policy outcomes and discounting. There seems to be broad consensus that attitudes toward risk and opportunities for private and public adjustment should be built into the monetary measures of *ex ante* net benefits and not to the discount rates (Pindyck, 2007). Second, studies of laypersons' risk perceptions and behavior in both experimental and survey contexts suggest discrepancies between subjective beliefs and experts' recommendations. These differences are especially important for small probability and long latency outcomes (Slovic, 1987). Such findings raise questions with what constitutes an informed choice. It has also motivated a growing literature in behavioral economics applied to environmental problems. Much of this literature focuses on the effects of manipulations in the presentation of information or the presentation of default choices to induce specific 'preferred' outcomes (Thaler and Sunstein, 2008). Bernheim and Rangel (2007) have proposed an alternative structure for describing individual choice that seeks to explain seemingly irrational choices as well as the success of these manipulations. Their extension proposes that conventional descriptions of the individual choice process overlook elements of the context for choices and features of the choice alternatives that are important to consumers. In their framework, analysis should focus on understanding these *ancillary conditions*. These types of insights would offer a consistent basis for modifying revealed preference analysis and applied welfare theory. While it has not been used in environmental economics, it offers a promising framework for dealing with discrepancies between objective and subjective beliefs.

Distribution

The term distribution in most economic analyses of policies would lead to a discussion of equity effects, interpreted as evaluating the consequences of actions for different income groups. For environmental policy, distributional considerations are inherently more complex. At least four perspectives have been considered. First, the use of an aggregate net benefits criterion for evaluating policy itself is a distributional judgment. It implies departures from a strict efficiency standard and a willingness to admit policies where there exists the potential for improvements. Positive aggregate net benefits means those gaining could compensate those losing. It does not require the compensation to take place.

The second distributional perspective follows from the identification of long-lived, potentially irreversible externalities – with greenhouse gases and nuclear waste being two widely recognized examples. In these cases, the distribution of costs and benefits extends across generations and the equity issues are often framed in ethical terms. One important implication of this line of reasoning is that policies should treat all generations alike. As a result, this line of reasoning would seem to imply a zero discount rate should be used. However, there is an equally strong intuition from a different perspective suggesting it is not morally acceptable. It implies a decision criterion (i.e., the zero discount rate) that would demand excessively high savings rates of any one generation. This countervailing influence is important because Koopmans' (1960) classic analysis of choice rules over infinite dimensional consumption streams concluded that these rules must display impatience or some amount of positive discounting. Otherwise, it would require the first generation to sacrifice everything. With preference separability over time, Arrow concluded that irreversibility does not change conclusions in favor of intertemporal discounting (see Portney and Weyant, 1999). Of course to the extent preferences or production displays important substitutions or complementarities with irreplaceable environmental resources, then the guidance will depend on the details of those linkages (see Sterner and Persson, 2008).

Economic and environmental feedback effects can have significant distributional consequences within short time horizons. Several cases illustrate how this third perspective can arise. Deciding between CAC vs incentive-based policies can seem clear-cut, considering the total costs of meeting a uniform standard (Tietenberg, 2006). However, the environmental transfer system, linking emissions to ambient concentrations, can imply that the cost savings are realized by allowing the cleanest areas to become dirtier while those exceeding the standard become cleaner. As Evans and Woodward (2013) explain, these factors can also undermine the ability to sustain markets for tradable permits. That is, regulations on the ambient concentrations of other pollutants together with the atmospheric interactions of different pollutants undermine the implicit assumption of independence of the effects of emissions of one type of pollutant from the ambient concentrations of others that are chemically related.

Much less is known about the economic feedback loops connecting environmental policy in one sector to its impacts on others. The economic models used to compute general equilibrium responses usually describe activities as if they all took place at the same 'point in space' (Hazilla and Kopp, 1990). Recent adaptations to these models (Carbone and Smith, 2013) and to household equilibrium location models (Kuminoff et al., 2013) confirm the importance of relaxing these assumptions. Indeed, the results from the equilibrium location model suggest that spatially differentiated, but consistent, air quality improvements can induce housing price adjustments (within a closed system), which imply some households would actually be worse off after an improvement in their air quality. This result arises because the housing price increases induced by some households' relocations can offset the air quality gains for those households with a small taste for environmental public goods.

The last distributional perspective, paradoxically, receives the greatest attention in agencies' design of policies (in the US some legislation specifically identifies concern for the impacts of regulations on small businesses) and often the least in academic discussions. It concerns the impact of regulations on industries and different-sized firms within a sector. Empirical evidence confirms that environmental regulations can have effects on the timing, location, and scale of plant investments. Considering only air quality regulations in the US, they find that activity shifts to areas that meet ambient standards, not subject to stringent area-wide regulations. Less activity takes place in areas that are not in attainment with national ambient air quality standards (see Greenstone, 2002). Greenstone et al. (2012) found using over a million plant-level observations drawn from 1972 to 1993 Annual Survey of Manufacturing Activities that among surviving plants, stricter air quality regulations are associated with approximately a 2.6% decline in total factor productivity.

Growth and the Environment

One of the most durable questions posed in considering interactions between the economy and the environment concerns whether the environmental system poses some 'ultimate limit' on the scale of economic activity (see Barnett and Morse, 1963). Sometimes expressed in terms of raw materials (or natural resources) and other times in terms of ability to provide a sustainable food supply and waste absorptive services to a growing world population, the issue most recently has been expressed in terms of whether growth in income has a self-regulating property. That is, as household income grows, the taste for environmental amenities increases. Providing these services requires reducing pollution, protecting natural environments that provide habitat for endangered species, and enhancing absorptive capacity of ecosystems for some types of pollution.

At one scale, this issue has been posed using what is termed an 'environmental Kuznets curve' (Grossman and Krueger, 1995). Pollution is hypothesized to have an inverted U-shape with household income. It increases until income growth creates demand for amenities that ultimately reduces pollution and leads to policies that restore environmental quality. To date, the questions raised in this line of research have been based on empirical analyses, relying on the contemporaneous performance of economies varying in pollution and household income, as if they could be treated as descriptions of a single economy moving through a dynamic process (Carson, 2010).

This new formulation of the old question is potentially important. It casts the issue in its most relevant form. Economists are concerned about the services provided by environmental resources because market institutions cannot be assured to signal the scarcity of these resources. Understanding both the static distortions to patterns of use and the dynamic nonmarket incentives to bring about changes in the

utilization of these natural assets is important for designing new institutions, which recognize the role of environmental resources in assuring that it is possible to continue to enhance economic well-being.

See also: Carrying Capacity of the Environment; Ecological Economics; Economic Geography; Economic Growth; Environment and Development; Environment and Health; Environmental Planning; Environmental Risk; Kuznets Curves; Limits to Growth; Public Goods: International.

Bibliography

Anderson, David A., 2013. Environmental Economics and Natural Resource Management, fourth ed. Routledge.

Arrow, K.J., Fisher, A.C., 1974. Environmental preservation, uncertainty, and irreversibility. Quarterly Journal of Economics 88, 312–319.

Arrow, K.J., Bolen, B., Costanza, R., Dasgupta, P., Folke, C., Holling, C.S., Jansson, B.O., Levin, S., Mäler, K.G., Perrings, C., Pimentel, D., 1995. Economic growth carrying capacity and the environment. Science 268, 520–521.

Arrow, K.J., Cropper, M.L., Eads, G.C., Hahn, R.W., Lave, L.B., Noll, R.G., Portney, P.R., Russell, M., Schmalensee, R., Smith, V.K., Stavins, R., 1996. Is there a role for benefit cost analysis in environmental, health and safety regulation? Science 272, 221–222.

Barnett, H.J., Morse, C., 1963. Scarcity and Growth. Johns Hopkins University Press, Baltimore, MD.

Barrage, L., 2012. Optimal Dynamic Carbon Taxes in a Climate-Economy Model with Distortionary Fiscal Policy. Dept of Economics, Yale University, unpublished.

Baumol, W.F., Oates, W.E., 1988. The Theory of Environmental Policy, second ed. Cambridge University Press, Cambridge.

Bernheim, B.D., Rangel, A., 2007. Behavioral Public Economics: Welfare and Policy Analysis with Non-Standard Decision Makers. In: Diamond, P., Vartiainen, H. (Eds.), Behavioral Economics and its Applications, Princeton University Press, Princeton, NJ, pp. 7–84.

Bovenberg, L., Goulder, L.H., 2002. Environmental taxation and regulation. In: Auerbach, A., Feldstein, M. (Eds.), Handbook of Public Economics. Elsevier, Amsterdam.

Buchanan, J.M., 1969. External diseconomies, corrective taxes and market structure. American Economic Review 59, 174–177.

Bullock, David S., Minot, Nicholas, November 2006. On measuring the value of a nonmarket good using market data. American Journal of Agricultural Economics 88, 961–973.

Carbone, J.C., Smith, V.K., 2008. Evaluating policy interventions with general equilibrium externalities. Journal of Public Economics 92, 1254–1274.

Carbone, J.C., Smith, V.K., 2013. Valuing nature in a general equilibrium. Journal of Environmental Economics and Management 66 (1), 72–89.

Carbone, J.C., Morgenstern, R.D., Williams III, R.C., 2012. Carbon Taxes and Deficit Reduction, unpublished.

Carlton, D.W., Loury, G.C., 1980. The limitations of pigouvian taxes as a long run remedy for externalities. Quarterly Journal of Economics 95, 559–566.

Carson, R.T., 2010. The environmental Kuznets curve: seeking empirical regularity and theoretical structures. Review of Environmental Economics and Policy 4, 3–21.

Carson, R.T., 2012. Contingent valuation: a practical alternative when prices aren't available. Journal of Economic Perspectives 26, 27–42.

Chay, Kenneth Y., Greenstone, Michael, 2003. The impact of air pollution on infant mortality: evidence from geographic variation in pollution shocks induced by a recession. Quarterly Journal of Economics 118 (3), 1121–1167.

Chay, Kenneth Y., Greenstone, Michael, 2005. Does air quality matter? Evidence from the housing market. Journal of Political Economy 113 (2), 376–424.

Coase, R.H., 1960. The problem of social cost. Journal of Law and Economics 3, 1–44.

Cornes, R., Sandler, T., 1996. The Theory of Externalities, Public Goods, and Club Goods, second ed. Cambridge University Press, Cambridge.

Dasgupta, P.S., Heal, G.M., 1979. Economic Theory and Exhaustible Resources. Cambridge University Press, Cambridge.

Davis, L., 2004. The effect of health risk on housing values: evidence from a cancer cluster. American Economic Review 94 (5), 1693–1704.

Ellerman, A.D., Joskow, P.L., Schmalensee, R., Montero, J.P., Bailey, E.M., 2000. Markets for Clean Air: The U.S. Acid Rain Program. Cambridge University Press, Cambridge.

Evans, D.A., Woodward, R.T., 2013. What Can We Learn from the End of the Grand Policy Experiment? The Collapse of the National SO2 Trading Program and Implications for Tradable Permits as a Policy Instrument. Annual Review of Resource Economics 5, 325–348.

Freeman III, A.M., 2003. The Measurement of Environmental and Resource Values: Theory and Methods, second ed. Resources for the Future, Washington, DC.

Goulder, L.H., Parry, I.W.H., 2008. Instrument choice in environmental policy. Review of Environmental Economics and Policy 2, 152–174.

Goulder, L.H., Parry, I.W.H., Williams III, R.C., Burtraw, D., 1999. The cost-effectiveness of alternative instruments for environmental protection in a second-best setting. Journal of Public Economics 72, 329–360.

Greenstone, M., 2002. The impacts of environmental regualtions on industrial activity: evidence from the 1970 and 1977 clean air act amendments and the census of manufactures. Journal of Political Economy 110 (6), 1175–1220.

Greenstone, M., Gallagher, J., 2008. Does hazardous waste matter? Evidence from the housing market and the superfund program. Quarterly Journal of Economics 123 (3), 951–1003.

Greenstone, M., List, J., Syverson, C., September 2012. The Effects of Environmental Regulations on the Competitiveness of US Manufacturing. NBER Working Paper #18392.

Grossman, G.M., Krueger, A.M., 1995. Economic growth and the environment. Quarterly Journal of Economics 110, 353–377.

Hanemann, W.M., 1994. Valuing the environment through contingent valuation. Journal of Economic Perspectives 8 (4), 19–43.

Hanley, N., Shogren, J., White, B., 2013. Environmental Economics. Oxford University Press.

Hausman, J., 2012. Contingent valuation: from dubious to hopeless. Journal of Economic Perspectives 26, 43–56.

Hazilla, M., Kopp, R.J., 1990. Social cost of environmental quality regulations: a general equilibrium analysis. Journal of Political Economy 98, 853–873.

Hicks, J.R., 1943. The four consumer surpluses. Review of Economic Studies 11 (1), 31–41.

Hotelling, H., 1931. The economics of exhaustible resources. Journal of Political Economy 39, 137–175.

Kling, C.L., Phaneuf, D.J., Zhao, J., 2012. From exxon to BP: has some number become better than no number? Journal of Economic Perspectives 26, 3–26.

Kneese, A.V., Bower, B.T., 1968. Managing Water Quality: Economics, Technology and Institutions. Johns Hopkins University Press for Resources for the Future, Baltimore, MD.

Koopmans, T.C., 1960. Stationary ordinal utility and impatience. Econometrics 28, 287–309.

Krutilla, J.V., 1967. Conservation reconsidered. American Economic Review 57, 777–786.

Krutilla, J.V., Fisher, A.C., 1975. The Economics of Natural Environments: Studies in the Valuation of Commodity and Amenity Resources. Johns Hopkins University Press for Resources for the Future, Baltimore, MD.

Kuminoff, N.V., Smith, V.K., Timmins, C., 2013. The new economics of equilibrium sorting and its transformational role for policy evaluation. Journal of Economic Literature 51 (4), 1007–1062.

Kuminoff, N.V., Pope, J.C., 2014. Do "Capitalization Effects" for Public Goods Reveal the Public's Willlingness to Pay? International Economic Review, forthcoming.

Lawn, P., 2007. Frontier Issues in Ecological Economics. Edward Elgar Publishing Co., Abingdon.

Machina, M.J., 1990. Choice under uncertainty: problems solved and unsolved. In: Hammond, P.B., Coppock, R. (Eds.), Valuing Health Risks, Costs and Benefits for Environmental Decision Making. National Academy Press, Washington, DC.

Mäler, K.-G., 1974. Environmental Economics: A Theoretical Inquiry. John Hopkins University Press for Resources for the Future, Baltimore, MD.

Mitchell, R.C., Carson, R.T., 1989. Using Surveys to Value Public Goods, the Contingent Valuation Method. Resources for the Future, Washington, DC.

Muller, N.Z., Mendelsohn, R., Nordhaus, W., 2011. Environmental accounting for pollution in the United States economy. American Economic Review 101 (5), 1649–1675.

Palmquist, R.B., 2005. Property value model. In: Mäler, K.-G., Vincent, J. (Eds.), Handbook of Environmental Economics. vol. 2. Elsevier, Amsterdam, pp. 763–819.

Parry, I.W.H., Norregaard, J., Heine, D., 2012. Environmental tax reform: principles from theory and practice. Annual Review of Resource Economics 4, 101–125.

Phaneuf, D.J., Requate, T. A Course in Environmental Economics: Theory, Policy, and Practice. Cambridge University Press, New York, forthcoming.

Pindyck, R.S., 2007. Uncertainty in environmental economics. Review of Environmental Economics and Policy 1, 45–65.
Portney, P.R., Weyant, J.P. (Eds.), 1999. Discounting and Intergenerational Equity. Resources for the Future, Washington, DC.
Roberts, M.J., Spence, M., 1976. Effluent charges and licences under uncertainty. Journal of Public Economics 5, 193–208.
Schlenker, W., Walker, W.R., 2011. Airports, Air Pollution and Contemporaneous Health. NBER working paper #17684.
Slovic, P., 1987. Perception of risk. Science 235 (4799), 280–285.
Smith, M.D., 2012. The new fisheries economics: incentives across many margins. Annual Review of Resource Economics 4, 379–402.
Solow, R.M., 14 June 1991. Sustainability: An Economist's Perspective, 18th J. Seward Johnson Lecture, Marine Policy Center. Woods Hole Oceanographic Institution, Woods Hole, MA.
Sterner, T., Persson, U.M., 2008. An even sterner review: introducing relative prices into the discounting debate. Review of Environmental Economics and Policy 2 (1), 61–76.
Thaler, R.H., Sunstein, C.R., 2008. Nudge: Improving Decisions about Health, Wealth, and Happiness. Yale University Press, New Haven, CT.
Thampapillai, J., Sinden, J.A., Environmental Economics: Concepts, Methods and Policies. Oxford University Press.
Tietenberg, T.H., 2006. Emission Trading: Principles and Practice, second ed. Resources for the Future, Washington, DC.
Venkatachalam, L., 2008. Behavioral economics for environmental policy. Ecological Economics 67 (4), 640–645.
Weitzman, M.L., 1974. Prices versus quantities. Review of Economic Studies 41, 477–491.
Willig, R.D., 1978. Incremental consumer's surplus and hedonic price adjustment. Journal of Economic Theory 17 (2), 227–253.

Environmental Education and Education for Sustainable Development

Therese Hume, Institute of Technology, Ballinode, Ireland
John Barry, Queens University Belfast, Belfast, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article attempts a broad characterization of environmental education (EE) and education for sustainable development (ESD), and includes a short overview of the history of the field, key debates, the main approaches to ESD and EE, and a look toward the future. However, such a brief account should not be considered to be fully comprehensive, and can only be considered to provide a broad overview of the field from the authors' perspectives.

Introduction

The ecological crisis and its associated problems in the social, cultural, political, and economic spheres is, in part, the product of ignorance, and as such knowledge of our global dilemma is a prerequisite for addressing it. For many, what we need is nothing less than a new enlightenment, an ecologically based one orientated around the needs of people and planet, and a clear eyed recognition that it is 'unsustainability' not 'sustainability' that characterizes our current global reality (Barry, 2012). Thus, what is needed is to investigate the role of education (both formal and informal) and its role in the transition from current unsustainability.

Environmental education (EE) and education for sustainable development (ESD) are not, nor can they ever be, 'just another' subject to be taught, given the challenges and opportunities outside the classroom and lecture theater to which it addresses. Following the contours and character of the ecological crisis that has given rise to it, ESD necessarily must be interdisciplinary, action-orientated, and holistic and combine both cognitive and conative aspects, as well as integrating both ethical and political analyses. This is a tall order! It is no easy task, but ESD did not create the conditions and problems, which has given rise to its development. Thus, given the knowledge context of unsustainable development, it is vital that a role to be found within the formal education system and outside to recover a sense of purpose fit for the times we live in and the multifaceted challenges and opportunities in the transition from 'actually existing unsustainability' (Barry, 2012).

Definitions and Declarations: a Brief History

EE in its current form emerged in the mid to late 1960s as a response to environmental problems caused through processes of modernity such as industrialization, consumerism, and urbanization. Its precursors included traditions of nature study, natural history, and conservation education. In the UK, the Council for Environmental Education first met in 1968 (Sterling, 2004); the *Journal of Environmental Education* was first published in the United States in 1969, and a seminal article, 'The Concept of Environmental Education,' gave one of the earliest definitions of the field (Stapp, 1969). According to Stapp, the aim of EE was to produce a motivated and knowledgeable citizenry concerning the biophysical environment, its problems, and their solution. Another major objective of EE, in addition to fostering an understanding of the environment and environmental problems and motivation to deal with them, was to ensure that individuals acquired a clear understanding that

> man is an inseparable part of a system, consisting of man, culture [organizational strategies, social processes and social arrangements] and the biophysical environment and that man has the ability to alter the relationships of this system.
>
> Stapp, 1969: 34

The sexist language used in these early declarations has been commented upon by Gough (2013b: 16), who also notes the 'man-made' nature of the foundational discourses of EE – women were largely absent in their formulation, and the discourse is that of a modernist science that separates 'man' and 'nature.' The long association of EE with science education (Gough, 2013a: 10) only began to change in the 1990s with the emergence of more socioecological approaches, which saw the natural science perspective as a simplification and inadequate to address the complex multilayered environmental problems (Kyburz-Gruber, 2013). Gough also argues that this gendered and science-based discourse also had implications in terms of a heavily positivist epistemology and ontology, which influenced much early research in EE (Robottom and Hart, 1993) and also influenced practice. Sterling (2004: 51) notes the early (and still predominant) 'environmental responsibility' view of sustainability education where the major aim was instrumental: changing unsustainable behavior through addressing environmental ignorance. Later studies such as Kollmuss and Agyeman (2002) demonstrated the paucity of such a linear 'deficit' model, and this has been reflected in an increasing emphasis on processes of learning as well as outcomes seen in, for example, socioecological approaches to education involving constructive, reflective, critical, and participatory elements (Kyburz-Gruber, 2013: 24).

The United Nations Conference on the Human Environment in Stockholm (UNEP, 1972), marked one of the first international acknowledgments of human–environmental problems, recommended the establishment of an international program of EE. Following this, a series of international

declarations and initiatives provided a more solid foundation for the emerging 'interdiscipline.' It is interesting that the first of these, in Belgrade Charter (1975), very clearly situated EE within a framework, which acknowledged growing global inequality and the need for a holistic approach to address environmental problems in the context of other issues such as hunger, poverty, illiteracy, and class, racial, and gender exploitation and domination. Possibly reflecting the influence of the *Limits to Growth* study (Meadows et al., 1972), the Belgrade declaration also challenged an uncritical approach to orthodox economic growth and articulated a need for a new global ethic, which not only considered humanity's place within the biosphere, but also would enable change leading to "an equitable distribution of the earth's resources" (Belgrade, 1975: 1–2). The environmental goal for nations was to improve all ecological relationships through clarifying what was meant by such concepts as 'quality of life' and 'human happiness' (Belgrade, 1975: 3). Consequently, the goal of EE was "to develop a world population with the knowledge, skills, attitudes, motivations and commitment to work individually and collectively towards solution to current problems and the prevention of any new ones" (Belgrade, 1975: 3).

The objectives and guiding principles defined in Belgrade were further refined at the first major international intergovernmental conference on environmental education, held in Tbilisi, Georgia in 1977. The subsequent declaration (though not without criticism) was one of the most influential founding documents in the field and laid out the fundamental role, objectives, and characteristics of EE. Those characteristics included its necessarily multidisciplinary nature due to the biological, ethical, social, cultural, and economic aspects of environmental issues. Concepts such as complexity and interdependence were also introduced as key elements. The scope of EE was to encompass lifelong learning, critical and independent thinking within formal and informal learning contexts (UNESCO, 1977). These ideas were to heavily influence subsequent UN initiatives (see Tbilisi, 2012 for a detailed account of these).

Following the Brundtland report (WCED, 1988), which introduced the (much contested) concept of 'sustainable development,' the 1992 United Nations Conference on Environment and Development in Rio (UNEP, 2010) catalyzed a shift in emphasis toward ESD. Agenda 21 Chapter 36 laid out a set of recommendations and actions in education for national governments underpinned by the Tbilisi principles and proposals. Significantly, Agenda 21 stressed the *criticality* of education in "promoting sustainable development and improving the capacity of the people to address environment and development issues" (UNCED, 1992: Chapter 36, para 3). This also brought EE closer to the field of development education, which had emerged via a separate trajectory, involving NGOs and aid agencies and emphasizing individual and collective empowerment, participation, democratization, and social action (Sterling, 2004: 45). Development education brought a more radical set of influences such as the critical pedagogy of Freire (1972). More interpretative approaches to EE research became evident from the 1980s onward, increasing significantly in the 1990s (Gough, 2013b: 18). This coincided with a growth in more constructivist approaches to education generally. The influence of critical pedagogy has also become more prominent. For example, approaches using action research (Reason and Bradbury, 2006) and participatory action research (Tandon, 2008) have increased exponentially since the 1990s (Gough, 2013b: 19).

Further conferences included ECO-ED in Toronto (1992) and Thessaloniki (1997), where the concept of 'Education for a Sustainable Future' was introduced (UNESCO, 1997; Tbilisi, 2012). In Thessaloniki, a broad conceptualization of sustainability that encompassed poverty reduction, health, food, security, democracy, human rights, and peace was emphasized: "Sustainability is, in the final analysis, a moral and ethical imperative in which cultural diversity and traditional knowledge need to be respected" (UNESCO, 1997: para 10). These international initiatives added impetus to existing national strategies and catalyzed action in particular educational sectors. For example, Chhokar (2010) notes that, in 1992, the Indian Supreme Court ruled that EE should be a mandatory part of all higher education. The Talloires declaration (October 1990) included a commitment by university leaders to create cultures of sustainability within their institutions, ensure environmental literacy, and educate for environmentally responsible citizenship; to collaborate internally to promote interdisciplinarity and externally with relevant bodies to contribute to the global effort for sustainability. An organization, University Leaders for a Sustainable Future, was set up and a journal – the *International Journal for Sustainability in Higher Education* was established. The Copernicus Charter played a similar role in Europe.

The next significant development was the establishment of a UN Decade of Education for Sustainable Development (DESD) from 2005 to 2015, following a recommendation made at the World Summit on Sustainable Development in Johannesburg in September 2002 (Tbilisi, 2012). Each member country was encouraged to incorporate ESD into all relevant subjects in their formal education systems and to develop policies and practices to achieve this. Detailed recommendations and a framework for national implementation were outlined in the strategy document (UNECE, 2005). A network of Regional Centres of Expertise was also established to promote learning for a sustainable future (Mochizuki and Fadeeva, 2008). Other supporting initiatives included the Mainstreaming Environment and Sustainability in African Universities (MESA) programme, developed to support the mainstreaming of environment and sustainability concerns into teaching, research, community engagement, and management of universities in Africa.

Sterling (2004: 46) acknowledges that in the first 30 years from the 1960s, a huge amount of work has been done on the underlying philosophy, research methodologies, pedagogy, curricula, resources, communication, and dissemination strategies of EE/ESD, supported and resourced by international bodies such as UNESCO and other NGOs such as the WWF. The agenda has influenced national policies in many countries and many educational sectors, and there has been growing academic involvement and research networks (see also Wals and Blewitt, 2010). However, despite such sustained international attention and commitment, the 2007 Ahmedabad conference on Environmental Education explicitly acknowledged that, "the state of the planet has gone from bad to worse" (Ahmedabad, 2007: 3–4). Regarding EE/ESD (following

another series of international conferences and declarations Gothenburg, 2008; Bonn, 2009; Rio+20, 2012; Tbilisi, 2012), a recent review of the DESD (UNESCO, 2012) evidences mixed progress. ESD is increasingly seen as a means to renew education, teaching, and learning; however, despite a number of inspiring examples around the world, the reality is that, "ESD remains marginal if considered as a separate and distinct entity and commonly understood concept in schools' everyday conversations" (UNESCO, 2012: 85).

More than 40 years since the inception of the field, in the introduction to the *Handbook on Research in Environmental Education*, Stevenson et al. (2013: 2) provide a broad *consensus* on the characteristics of EE as currently manifest. For them, it embraces normative questions: it is interdisciplinary, with environmental, sociocultural, and economic dimensions; agency is important in addition to knowledge, understanding, attitudes, and values; it involves engaged, situated learning, which must also take place in the public domain; and has global and local aspects. In addition to this, perhaps above all else, it is explicitly (though unevenly) viewed as a political and transformative project. However, the breadth and, some would argue, necessary vagueness of terms used to define ESD have meant that the EE/ESD project has been widely criticized and to some extent fragmented (Sterling, 2004). The next section hence addresses key conflicts and contradictions in the field.

Problems and Key Contradictions

As discussed above, if we were to assess the effectiveness of the ESD project in terms of lessening what Barry (2012) has termed 'actually existing unsustainability,' we might conclude that there has been limited (if any) progress. One possible reason for this might be lack of implementation of the many declarations on ESD: Filho (2012), in a commentary preceding the Rio+20 summit, observes that the majority of declarations and action plans have never been fully implemented, particularly at regional and local levels. Martin et al. (2013: 1537), commenting generally on the DESD, note that most evidence indicates a rather ill-focused and half-hearted awareness raising campaign leading to a patchy impact and limited engagement. However, problems may run deeper than a failure to implement broadly agreed international strategies. Cook et al. (2010: 313–314) raise the question that "formal education, as it is currently organized is incapable *of* coming to grips with the problem?"

One core contradiction here is that ESD is mostly conducted within political and educational systems whose basic assumptions, particularly those underpinning how their economies are run, might fundamentally militate against ecological sustainability. As Wals and Corcoran (2012) note, it is impossible to examine questions of sustainability education without acknowledging the other processes taking place in society that are moving us toward unsustainability; these raise ethical and political issues of the unequal distribution of power and agency, social injustice, in addition to uncertainty and other complexities. This contradiction has been acknowledged as early as 1974 when the Institute for Earth Education was set up in the United States (Van Matre, 1990) on the basis that EE had been trivialized and diluted by both mainstream education and the agencies and industries who themselves are the main contributors to continuing unsustainability problems. David Selby points out the thorny problem of how a democratic citizenry, "immured in a pervasive consumerist ethic, can be squared with an environmental narrative predicated on the finiteness of the earth" Selby (2011: 10). The way that formal education systems themselves are structured and their fundamental epistemological objectives clearly militate against effective, that is, *transformative forms of ESD*. For example, one evaluation of the progress of the DESD (UNESCO, 2012: 86) suggests that colleges and universities around the world are starting to make more systemic changes toward sustainability, "amid educational reforms toward efficiency, accountability, privatization, and forms of management and control processes that often hamper their efforts" (UNESCO, 2012: 86). That is, the reform process within higher education institutions offers an unconducive context for ESD.

Other criticisms of ESD include the overemphasis of a tendency to view the student as an *individual* actor. As Scott and Gough put, this "neglects the picture of the student as a social and more public actor – as someone engaged in the messy but vital business of democratic citizenship that is so central to the received view of sustainable development and ESD" (Scott and Gough, 2010: 3737). This [individualized] view also lends itself to a serious dilution of a necessary politicized approach to learning and acting about current unsustainability and the transition to resilience and sustainability. However, the danger that such a politicized approach might turn prescriptive or dogmatic in certain cases has also been articulated: Sund and Lysgaard (2013) note that prescriptive, normative forms of ESD, which focus on learning outcomes as opposed to learning processes are problematic, as they often neglect educational best practice. Earlier criticisms of ESD concern politically prescriptive, dogmatic content using "emotionalism, myths and misinformation" (Kwong, 1997: 90). This raises the dilemma of how normative and ethical issues can be *effectively* explored in educational settings in a nondogmatic way, given the urgency of the need for change. The next section thus addresses a range of approaches to EE/ESD.

Approaches to EE/ESD: Moving Toward a Paradigm Shift

In 1993, three basic paradigms were identified by Robottom and Hart, roughly corresponding to the dominant research paradigms of the time: positivist, interpretivist, and critical. To these, Fien (2002) has added poststructuralism and more recently, Gough (2013b: 19) has noted the increasing influence of poststructuralism and multivoice, multicentred studies foregrounding indigenous knowledge, postcolonial and feminist perspectives, in addition to work on socioecological resilience. These basic paradigms or worldviews reflect particular approaches to education.

Vare and Scott (2007: 191) identify *instrumental* ESD, or ESD 1, as "the promotion of informed, skilled behaviours and ways of thinking," reflecting a more behavioristic, outcome-focused approach. The *intrinsic* ESD approach or ESD 2 is concerned with individual learning and involves "building capacity to think critically about what experts say and to test

ideas, exploring the dilemmas and contradictions inherent in sustainable living" (Vare and Scott, 2007: 191). This is based on a more constructivist educational philosophy and as such focuses more on the processes of learning (e.g., Wals, 2006). A third, more critical, approach to EE is concerned with social learning. Scott and Gough (2010: 3743) argue that "ESD can helpfully be seen as an education in citizenship: a responsive social learning process which is a preparation for informed, open-minded, social engagement with the main existential issues of the day that can be experienced in the family, the community and workplace, indeed, in all aspects of lifelong learning." More radical approaches derive from explicitly political and critical pedagogies (e.g., Freire, 1972), which place an emphasis on education for social and political transformation (see also Huckle, 2008). However, Freire's critical pedagogy, as well as other liberal education approaches such as Dewey's, has been criticized by Bowers (2005) who argues that the one key mistake these approaches have made is to assume that there is one 'true' approach to learning. To counter this, Bowers advocates examining "ecologically problematic cultural assumptions" (Bowers, 2005: 13) to dominant knowledge (including 'commonsense' taken for granted ideas of 'economic growth' as an agreed societal imperative), and a reappraisal of local knowledge systems and traditions reflects cultural and bioregional diversity.

The acknowledgment that content studied processes of learning, and diverse regional and political contexts of learning are *all* important, have led to more integrative models of ESD, utilizing social–ecological systems approaches. This shift can be seen in the work of educators such as Stephen Sterling (2003), who, echoing O'Sullivan (1999) note that (notwithstanding some essential differences) critical, transformative pedagogy and ESD can be complementary. For example, Sterling's (2003) model of 'sustainable education' essentially integrates and transcends a range of existing approaches and adopts a whole systems thinking approach, which he equates with a "postmodern ecological worldview" (Sterling, 2003: 39). Gregory Bateson's (1972) 'levels of learning' are used as a frame to enable the examination of learning in different contexts, and in different ways, all of which may have their uses. Learning 1 or learning within paradigm (a particular set of choices for action available within a particular worldview) can be seen as roughly corresponding to more instrumental forms of ESD or to "single-loop learning" (Argyris and Schön, 1996). Learning 2, more closely associated with critical thinking, and Argyris and Schon's "double-loop learning" involves awareness of alternative paradigms or worldviews that frame sets of choices. The third level of learning identified by Bateson, is essentially transformative, involving a more enlightened self-awareness of what frames choice in the individual. Thus, Sterling conceptualizes 'sustainable education' as:

> a change of educational culture which both develops and embodies the theory and practice of sustainability in a way which is critically aware. This would be a transformative paradigm which values, sustains and realizes human potential in relation to the need to attain and sustain social, economic and ecological well-being, recognizing that they are interdependent.
>
> Sterling, 2003: 233

This broader view incorporates individual, organizational, and social transformation and can also provide a pragmatic basis for change in existing formal education systems. For example, to engender change in higher education institutes, Jones et al. (2010), in common with many ESD practitioners advocate the use of a holistic definition of sustainability that brings into play dimensions of culture, economy, health, peace, and conflict, human rights, gender equality, science and technology, and social, global and intergenerational justice. This broader view also links to other more psychological approaches to how ESD can motivate behavioral change, i.e., move from critical thinking to action, such as the reasoned action approach (Fishbein and Ajzen, 2010). Multifaceted approaches inspire exploration of alternative problem definitions and solutions, and foster deliberation and debate, which require contribution from a wide range of disciplines.

Models such as Sterling's, by integrating and effectively transcending a number of approaches, thus fulfill the requirement articulated by the Ahmedabad conference (2007: 4) for "fundamental change in the purpose and practice of education." The American College and University Presidents' Climate Commitment, following Cortese (2003), describe what such change might look like in higher education:

> - The *context* of learning would change to make human/environment interdependence, values and ethics a seamless and central part of teaching of all the disciplines, campus operations should be sustainable, and partnership should occur between local, regional and global communities.
> - The *content* of learning would further reflect interdisciplinary systems thinking, dynamics and analysis for all majors and disciplines and bring lateral rigor across the disciplines as well as the vertical rigor currently within them.
> - The *process* of education would complement formal curriculum with active, experiential, inquiry-based learning and real-world problem solving both campus-based and within the larger community.
>
> ACUPCC, 2009: 6

Emerging areas in higher education, which reflect such a paradigm shift include the emerging field of sustainability science that: "seeks to understand the fundamental character of interactions between nature and society" (Kates et al., 2001: 641).

Moving toward the Future: ESD and the Building of Resilient Communities

A more recent movement in EE/ESD, reflecting the need to address emerging and potential environmental problems, has been in the area of learning for resilience and in examining the role of learning in the building of resilient communities in the context of exacerbating ecological and resource problems, such as climate change and 'peak oil.' This has been influenced by grassroots movements, such as the 'transition towns' movement (Hopkins, 2008; Barry, 2012), and studies of resilience and adaptive capacity in

socioecological systems (e.g., Folke et al., 2002). This also reflects a shift from more reductionist paradigms toward a complex adaptive systems worldview (Barry, 2012), which can be seen in fields such as education, ecology, ecosystem management, and reflexive governance for sustainability. This worldview sees reality as complex and dynamic, there is no 'one right answer,' and sustainability and resilience are seen as emergent properties of specific social–ecological systems. Thus sustainability and resilience (and therefore education for both) will differ in different socioecological contexts so that ESD/EE in Belfast will not necessarily be the same as EE/ESD in Boston, Berlin, or Beijing.

Thus, building resilience in bioregions involves "diagnosing, mapping and adapting to" (Giddens, 2009: 163) existing and *emerging* vulnerabilities. It also involves attending to "slowly-changing, fundamental variables that create memory, legacy, diversity, and the capacity to innovate in both social and ecological components of the system" (Folke et al., 2002: 8). This implies *on-going* processes of learning, innovation, and knowledge creation utilizing a diverse range of existing and historical knowledge and skills. In this way, learning for resilience occurs at different levels: individual, group, community, organization – and in different settings (home, school, work, street, community, formal and informal contexts, etc.). Sterling (2010) equates the resilient learner with the *intrinsic* approach to sustainability education (Vare and Scott, 2007), focused on building the capacity in the individual to be critically aware, inquire, and learn. How individuals collectively learn is also important in building resilience at community and societal levels. Lundholm and Plummer (2010), citing Glasser (2007), note that social learning depends on the maturity and competence of the individuals involved and the rules that guide particular organizational learning, public participation, and decision-making processes. Thus learning for resilience is lifelong, both responsive and anticipatory, and multifaceted. It requires a rethink of how learning and institutions of learning are structured, to reflect a more outward-looking social focus and provide a public resource for learning, action, and political transformation. And given the continuing unsustainability of human–nature relations, ESD must also be seen as an essential element in creating a "safe operating space for humanity" (Rockström et al., 2009).

Conclusion

At a fundamental level, Thomas Berry cites the relationship between humans and the other than human components that we live in as the "missing element" in education (Berry, in O'Sullivan, 1999: x). For Berry, educationalists have forgotten that "we and all our institutes of learning are primarily in a learning position in reference to the larger world around us," and that "every occupation must be established *within* the integral functioning of the planet" (O'Sullivan, 1999: xiv, emphasis added). Thus, there is an imperative to renew "our acquaintance with the sensuous world in which our techniques and technologies are all rooted" (Abram, 1996: x). To this end, Orr (2004) notes that education should be reshaped in a way that fosters cultivation of 'biophilia' ("the urge to affiliate with other forms of life" Wilson (1984)), a process that should begin in childhood and continue throughout life. Such an approach is evident in earth education (Van Matre, 1990), forest education (www.foresteducation.org), and the civic ecology (www.civicecology.org) movements – effectively outdoor and/or community-based education, which will draw out "our affinity for life" (Orr, 2004: 213), and foster a love of nature. The idea of 'sustainability competence' (Wals and Corcoran, 2012: 24) thus includes the ability to put oneself in the mind of others, even other species, having different backgrounds or living elsewhere, in addition to abilities to see relationships and interdependencies and to envision future implications.

However, to be effective in addressing existing forces of unsustainability, ESD must (in our view) be profoundly *political* and aimed at real world transformation (Barry, 2007). As mentioned above, almost 40 years ago, the Belgrade (1975) very clearly situated EE within an ethical and political framework, which would enable change leading to "an equitable distribution of the earth's resources […] more fairly satisfying the needs of all people" (Belgrade, 1975: 2) and improve socioecological relationships through clarifying what was meant by such concepts as 'quality of life' and 'human happiness' (Belgrade, 1975: 3). To grow such an ethical framework requires broader, systemic, and political conceptualizations of ESD incorporating individual and social learning, in addition to the acquisition of practical skills and diverse knowledge. As the 2007 Ahmedabad conference noted, "we need recommendations that help alter our economic and production systems, and ways of living radically. We need an educational framework that not only follows such radical changes, but can take the lead" (Ahmedabad, 2007: 4). And, in that way, we find that both EE and ESD are, properly speaking, concerned not so much with humanity 'learning' about nature, but learning about ourselves and nature within the central learning objective of how to understand and ultimately transform our species' foundational and ineliminable relationship with the nonhuman world.

See also: Communities of Practice; Education and Economic Growth; Educational Effectiveness Theory and Research: Recent Advances; Educational Philosophy; Environments for Education; Equity and Education; Gender and Education; Learning Theories and Educational Paradigms; Moral Development and Education; Science Education.

Bibliography

Abram, David, 1996. The Spell of the Sensuous. Random House.

ACUPCC, 2009. Education for Climate Neutrality and Sustainability. Guidance for ACUPCC Institutions. http://www.presidentsclimatecommitment.org/documents/EducationforClimateNeutralitySustainability_2009.05.07_finalWEB.pdf (accessed 16.12.11.).

Ahmedabad, 2007. The Ahmedabad Declaration 2007: A Call to Action. http://www.tbilisiplus30.org/Ahmedabad%20Declaration.pdf (accessed 16.12.11.).

Ahmedabad Recommendations, 2007. Moving Forward from Ahmedabad: Environmental Education in the 21st Century. http://www.tbilisiplus30.org/Final%20Recommendations.pdf (accessed 16.12.11.).

Argyris, C., Schön, D., 1996. Organisational Learning II: Theory, Learning and Practice. Addison-Wesley Publishing, Reading, MA.

Barry, J., 2007. Spires, plateaus and the infertile landscape of education for sustainable development: re-invigorating the university through integrating community, campus and curriculum. International Journal of Innovation and Sustainable Development 2 (3/4), 433–452.

Barry, John, 2012. The Politics of Actually Existing Unsustainability. Oxford University Press.

Bateson, G., 1972/1987. Steps to an Ecology of Mind. Aronson.

Belgrade, 1975. International Workshop on Environmental Education – Belgrade Charter. http://unesdoc.unesco.org (accessed 21.07.13.).

Bonn, 2009. Bonn Declaration, UNESCO World Conference on Education for Sustainable Development. www.esd-world-conference-2009.org (accessed 14.02.12.).

Bowers, Chet, 2005. Educating for a Sustainable Future: Mediating between the Commons and Economic Globalization. www.c-a-bowers.com (accessed 21.07.12.). (Refers to Bowers, Chet, 2005. Silences and Double Binds: Why the Theories of John Dewey and Paulo Freire Cannot Contribute to the Revitalization of the Commons. www.c-a-bowers.com (accessed 21.07.12.).)

Chhokar, Kiran Banga, 2010. Higher education and curriculum innovation for sustainable development in India. International Journal of Sustainability in Higher Education 11 (2), 141–152.

Cook, R., Cutting, R., Summers, D., 2010. If sustainability needs new values, whose values? Initial teacher training and the transition to sustainability. In: Jones, P., Selby, D., Sterling, S. (Eds.), Sustainability Education: Perspectives and Practice across Higher Education. Earthscan, London, pp. 313–328.

Cortese, Anthony D., 2003. The critical role of higher education in creating a sustainable future. Planning for Higher Education March-May, 15–22. http://www.scup.org/asset/48483/cortese.pdf (accessed 21.07.13.).

Fien, John, 2002. Advancing sustainability in higher education: issues and opportunities for research. International Journal of Sustainability in Higher Education 3 (3), 243–253.

Filho, Walter Leal, June 17, 2012. Future Challenges for Sustainable Development. University World News. http://www.universityworldnews.com (accessed 21.07.13.).

Fishbein, M., Ajzen, I., 2010. Predicting and Changing Behavior: The Reasoned Action Approach. Psychology Press.

Folke, C., et al., 2002. Resilience and Sustainable Development: Building Adaptive Capacity in a World of Transformations, Scientific Background Paper on Resilience for the Process of the World Summit on Sustainable Development. Environmental Advisory Council to the Swedish Government. www.mvb.gov.se (accessed January 2011).

Freire, P., 1972. Pedagogy of the Oppressed. Penguin, London.

Giddens, Anthony, 2009. The Politics of Climate Change. Newleaf.

Glasser, Harold, 2007. Minding the gap: the role of social learning in linking our stated desire for a more sustainable world to our everyday actions and policies. In: Wals, A.E.J. (Ed.), Social Learning towards a Sustainable World. Principles, Perspectives, and Praxis. Wageningen Publishers, Wageningen, The Netherlands, pp. 35–62.

Gough, Annette, 2013a. Historical, contextual and theoretical orientations that have shaped environmental education research. In: Stevenson, Robert B., Brody, Michael, Dillon, Justin, Wals, Arjen E.J. (Eds.), International Handbook of Research on Environmental Education. Routledge, pp. 9–12.

Gough, Annette, 2013b. The emergence of environmental education research: a 'history' of the field. In: Stevenson, Robert B., Brody, Michael, Dillon, Justin, Wals, Arjen E.J. (Eds.), International Handbook of Research on Environmental Education. Routledge, pp. 13–22.

Gothenburg, 2008. Gothenburg Recommendations on Education for Sustainable Development. University of Gothenburg.

Hopkins, Robert, 2008. The Transition Handbook from Oil Dependency to Local Resilience. Green Books.

Huckle, J., 2008. Sustainable development (Chapter 26). In: Arthur, J., Davies, I., Hahn, C. (Eds.), The Sage Handbook of Education for Citizenship and Democracy. Sage Publications, London, pp. 342–354.

Jones, P., Selby, D., Sterling, S., 2010. Sustainability Education: Perspectives and Practice across Higher Education. Earthscan, London.

Kates, R.W., Clark, W.C., Corell, R., Hall, J.M., Jaegar, C.C., Lowell, I., et al., 2001. Sustainability science. Science 291, 641–642.

Kollmuss, Anja, Agyeman, Julian, 2002. Mind the gap: why do people act environmentally and what are the barriers to pro-environmental behavior? Environmental Education Research 8 (3), 239–260.

Kwong, Jo, 1997. A Critique of Environmental Education in Environmental Education by Benjamin Aldrich-Moodie and Jo Kwong. IEA Education and Training Unit and IEA Environment Unit, London.

Kyburz-Gruber, H., 2013. Socioecological approaches to environmental education and research. In: Stevenson, Robert B., Brody, Michael, Dillon, Justin, Wals, Arjen E.J. (Eds.), International Handbook of Research on Environmental Education. Routledge, pp. 23–32.

Lundholm, C., Plummer, P., 2010. Resilience in social-ecological systems: the role of learning and education. Environmental Education Research 16 (5–6), 475–491.

Martin, S., Dillon, J., Higgins, P., Peters, C., Scott, W., 2013. Divergent evolution in education for sustainable development policy in the United Kingdom: current status, best practice, and opportunities for the future. Sustainability (5), 1522–1544.

Meadows, Donella, Meadows, Dennis, Jørgen, Randers, William, Behrens, 1972. Limits to Growth, Club of Rome 1972. New American Library, New York.

Mochizuki, Yoko, Fadeeva, Zinaida, 2008. Regional centres of expertise on education for sustainable development: an overview. International Journal of Sustainability in Higher Education 9 (4), 369–381.

O'Sullivan, Edward, 1999. Transformative Education. University of Toronto Press.

Orr, David, 2004. Earth in Mind, 10th anniversary ed. Island Press.

Reason, P., Bradbury, H., 2006. Handbook of Action Research. Sage.

Robottom, I., Hart, P., 1993. Research in Environmental Education: Engaging the Debate. Deakin University, Geelong.

Rockström, J., et al., 2009. Planetary boundaries: exploring the safe operating space for humanity. Ecology and Society 14 (2), 472–475. www.ecologyandsociety.org (accessed 26.03.11.).

Rio+20, 2012. Education for a Sustainable Future. http://www.unesco.org/new/en/rio-20/educating-for-a-sustainable-future/ (accessed 21.07.13.).

Scott, William, Gough, Stephen, 2010. Sustainability, learning and capability: exploring questions of balance. Sustainability (2), 3735–3746.

Selby, D., 2011. Education for sustainable contraction as appropriate response to global heating. Journal for Activism in Science & Technology Education 3 (1), 5.

Stapp, William, 1969. The concept of environmental education. Journal of Environmental Education 1 (1), 33–35.

Sterling, Stephen, 2003. Whole Systems Thinking as a Basis for Paradigm Change in Education: Explorations in the Context of Sustainability. University of Bath. www.bath.ac.uk/cree/sterling/sterlingthesis.pdf (accessed 20.07.13.).

Sterling, Stephen, 2004. An analysis of the development of sustainability education internationally: evolution, interpretation and potential transformation. In: Blewitt, John, Cullingford, Cedric (Eds.), The Sustainability Curriculum: The Challenge for Higher Education. Earthscan, pp. 43–62.

Sterling, Stephen, 2010. Resilience and education. Environmental Education Research 16, 5–6.

Stevenson, Robert B., Brody, Michael, Dillon, Justin, Wals, Arjen E.J. (Eds.), 2013. International Handbook of Research on Environmental Education. Routledge.

Sund, Per, Lysgaard, Jonas Greve, 2013. Reclaim "education". In: Environmental and Sustainability Education Research Sustainability, vol. 5, pp. 1598–1616.

Tandon, Rajesh, 2008. Search of Relevance: Higher Education for Participatory Research and Sustainable Development, in Reinventing Higher Education: Towards Participatory and Sustainable Development. UNESCO, Bangkok. http://unesdoc.unesco.org/images/0016/001631/163155e.pdf (accessed 28.07.13.).

Tailloires, October 1990. Tailloires Declaration. http://www.ulsf.org/talloires_declaration.html (accessed April 2009).

Tbilisi, 2012. Tbilisi+35 a History of Education for Sustainable Development. Ministry of Environment Protection of Georgia.

UNCED, 1992. Agenda 21. Chapter 36, para 3. http://www.un-documents.net/a21-36.htm (accessed 21.07.13.).

UNECE, March 23, 2005. UNECE Strategy for Education for Sustainable Development, United Nations. CEP/AC.13/2005/3/Rev.1.

UNEP, 1972. Declaration of the United Nations Conference on the Human Environment. http://www.unep.org/Documents.Multilingual/Default.asp?documentID=97&ArticleID=1503 (accessed 21.07.13.).

UNEP, 2010. Climate Change Factsheet. http://www.unep.org/climatechange/ (accessed 21.12.11.).

UNESCO, 1977. Intergovernmental Conference on Environmental Education: Final Report. UNESCO. http://unesdoc.unesco.org/images/0003/000327/032763eo.pdf (accessed 23.07.13.).

UNESCO, December 1997. Thessaloniki Declaration UNESCO-EPD-97KONF.40 IKLD.212. http://resources.spaces3.com/f49d326a-5930-45cf-a9ef-4d23a0670a6d.pdf (accessed 20.07.13.).

UNESCO, 2012. DESD Monitoring and Evaluation 2012. http://unesdoc.unesco.org/images/0021/002164/216472e.pdf (accessed 21.07.13.).

Vare, P., Scott, W.A.H., 2007. Learning for a change: exploring the relationship between education and sustainable development. Journal of Education for Sustainable Development September 1 (2), 191–198.
Van Matre, Steve, 1990. Earth Education: A New Beginning. Institute for Earth Education, Greenville, WV.
Wals, Arjen E.J., 2006. The end of ESD … the beginning of transformative learning – emphasizing the E in ESD. In: Cantell, Mikko (Ed.), Seminar on Education for Sustainable Development. Wageningen University, The Netherlands. Translation: Laura Murto, 2006. Publication Series of the Finnish National Commission for UNESCO, No. 83. Helsinki.
Wals, Arjen E.J., Blewitt, J., 2010. Third-wave sustainability in higher education, some (inter)national trends and developments. In: Jones, P., Selby, D., Sterling, S. (Eds.), Sustainability Education: Perspectives and Practice across Higher Education. Earthscan, London, pp. 55–74.
Wals, Arjen E.J., Corcoran, Peter B., 2012. Learning for Sustainability in Times of Accelerating Change. Wageningen Academic.
WCED, 1988. Our Common Future, Report of the World Commission on Environment and Development. Oxford University Press, Oxford, England.
Wilson, E.O., 1984. Biophilia. Harvard University Press, Harvard.

Relevant Websites

http://www.civicecology.org/ – Civic Ecology.
http://www.copernicus-alliance.org/ – Copernicus Alliance.
http://www.desd.org/ – Decade of Education for Sustainable Development.
http://www.ecojusticeeducation.org/ – Ecojustice Education.
http://www.foresteducation.org/ – Forest Education.
http://www.resalliance.org/ – Resilience Alliance.
http://www.transitionnetwork.org/ – Transition Network.
http://www.ulsf.org/ – University Leaders for a Sustainable Future.

Environmental Health and Safety: Social Aspects

Dean Bertolatti, Toni Hannelly, and Janis Jansz, Curtin University, Perth, WA, Australia

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by P. Brown, volume 7, pp. 4617–4621, © 2001, Elsevier Ltd.

Abstract

Environmental health factors play a significant role in human health and the protection of the environment. The health impacts resulting from exposure to these factors are closely linked to social aspects such as socioeconomic status and access to health care. Three current environmental case studies are discussed from both an environmental and a social aspect.

Introduction

As a cornerstone of public health, environmental health factors play a significant role in human morbidity and mortality in both developing and developed countries across the globe. The World Health Organization (WHO) recognizes the impact of these factors and has adopted a two-stranded definition of environmental health that embraces both the etiology of health problems and also the methodologies to manage them.

> Environmental health addresses all the physical, chemical, and biological factors external to a person, and all the related factors impacting behaviors. It encompasses the assessment and control of those environmental factors that can potentially affect health. It is targeted towards preventing disease and creating health-supportive environments.
> (WHO, 2013a)

In a 2006 WHO Report, "Preventing disease through healthy environments: Toward an estimate of the environmental burden of disease," Pruss-Ustun and Corvalan reported that an estimated 24% of the global burden of disease and 23% of all premature mortality can be attributable to environmental factors with the specific fraction of disease attributable to the environment varying widely across different disease conditions due to differences in environmental exposures and access across the regions. In 2013, WHO (2013a) listed 10 facts on preventing death, illness, and disability through healthy environments, highlighting the need for environmental management of potential hazards. These statements highlight the disproportionate impact of environmental factors on vulnerable populations, particularly the impoverished in developing countries.

- 'Worldwide, 13 million deaths could be prevented every year by making our environments healthier.
- In children under the age of five, one-third of all disease is caused by the environmental factors such as unsafe water and air pollution.
- Every year, the lives of four million children under 5 years of age – mostly in developing countries – could be saved by preventing environmental risks such as unsafe water and polluted air.
- In developing countries, the main environmentally caused diseases are diarrheal disease, lower respiratory infections, unintentional injuries, and malaria.
- Better environmental management could prevent 40% of deaths from malaria, 41% of deaths from lower respiratory infections, and 94% of deaths from diarrheal disease – three of the world's biggest childhood killers.
- In the least developed countries, one-third of death and disease is a direct result of environmental causes.
- In developed countries, healthier environments could significantly reduce the incidence of cancers, cardiovascular diseases, asthma, lower respiratory infections, musculoskeletal diseases, road traffic injuries, poisonings, and drownings.
- Environmental factors influence 85 out of the 102 categories of diseases and injuries listed in the 2004 World Health Report (WHO, 2013a).
- Much of this death, illness, and disability could be prevented through well-targeted interventions such as promoting safe household water storage, better hygiene measures, and the use of cleaner and safer fuels.
- Other interventions that can make environments healthier include: increasing the safety of buildings; promoting safe, careful use, and management of toxic substances at home and in the workplace; and better water resource management.'

Environmental determinants of health have traditionally focused on areas of underdevelopment such as inadequate water quality, inadequate sanitation and waste disposal, presence of disease vectors, and poor housing. While most of these areas have been managed successfully in developed countries, they are still major causes of morbidity and mortality in less developed counties (Liamputtong et al., 2012). Paradoxically, many countries are now faced with modern threats that have emerged due to overconsumption and overdevelopment, compounded by emerging threats such as climate change. These new hazards are the result of often uncontrolled development activities occurring without sufficient safeguards and unsustainable consumption of natural resources (Earl et al., 2012; SERI et al., 2009). The Health Council of the Netherlands (2012) also noted that in Western nations, increased health risks are usually related to an accumulation of unfavorable physical and social circumstances. Rapid urbanization and large-scale movements of people have introduced a whole new set of environmental health hazards while exacerbating others (Margai, 2012).

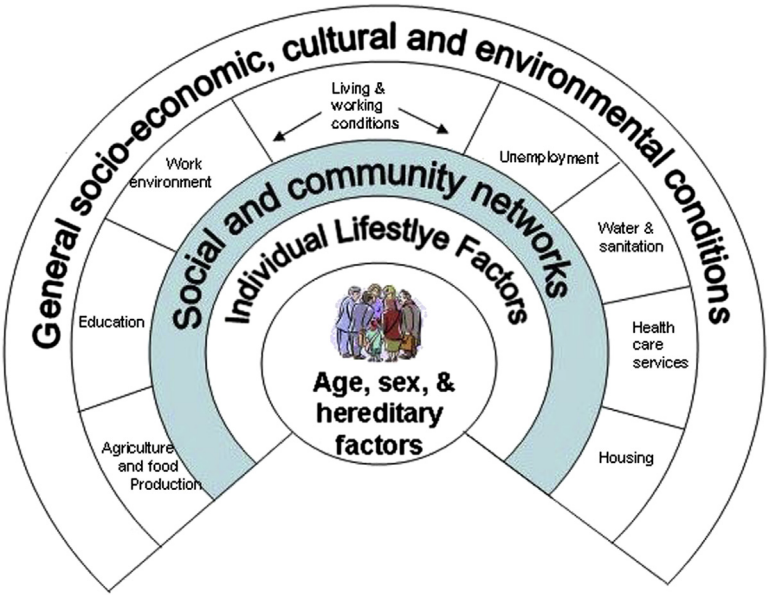

Figure 1 The main determinants of health. Dahlgren, G., Whitehead, M., 1991. Policies and Strategies to Promote Social Equity in Health. Institute for Further Studies, Stockholm.

In addition to environmental factors, human health is affected by many other interrelated and overlapping factors. One widely accepted model, which captures the interrelationships between various factors that affect human health is the Dahlgren and Whitehead (1991) 'Policy Rainbow' (**Figure 1**). This rainbow model conceptualizes the main determinants of population health and demonstrates the connections between health and socioeconomic, environmental, and cultural conditions. The inner layer represents personal factors and behaviors, the middle layer is the social and community networks that ideally support individuals as well as groups and families, and the outer layer is the environmental and social factors that play a more encompassing or overarching role in determining health. This reinforces the concept that despite favorable lifestyle factors and community support, environmental, cultural, and general economic conditions ultimately determine the state of human health.

In 2012, the Health Council of the Netherlands released a report on the "Social Aspects of the Living Environment in Relation to Environmental Health." They also noted that numerous relationships play a role in environmental-related health problems and that exposure to various environmental stressors or hazards can have a broad range of effects, with specific problems often the result of several stressors interacting on human health. The authors also noted that environmental stressors do not always have the same effect on health and that in addition to context and exposure, this is largely dependent on the social determinants of health which are the social, cultural, economic, environmental, and political factors that can impact on the health of individuals.

Social Determinants of Health

The WHO Commission on Social Determinants of Health broadly defines social determinants as "… the circumstances in which we are born; grow up, live, work and age, and the systems put in place to deal with illness. These circumstances are in turn shaped by a wider set of forces such as economics, social policies and politics" (WHO CSDH, 2008: p. 1). These are all factors that impact on our vulnerability or immediate coping capacity as well as our ability to adapt to changing situations. In the WHO publication, *The Solid Facts*, Wilkinson and Marmot (2003) and Marmot and Wilkinson (2006) list and discuss the 10 widely accepted social determinants of health, from social status differences to transportation. All of these determinants have environmental dimensions. For example, social inequality can be linked to overcrowding and substandard housing, which in turn may increase the risk of exposure to contaminated air or unclean water.

Vulnerability and Human Health

The impact of environmental hazards on human health is generally dependent on the vulnerability of the affected population. What becomes apparent when looking at the morbidity distribution of illness and disease caused by environmental factors is that the health impacts of these factors disproportionately affect the poor and isolated, both those in developing countries as well as the socially disadvantaged groups in developed countries. These people are vulnerable in that not only are they unable to respond appropriately to cope with the immediate effects of environmental hazards or incidents, but also they are less likely to be able to recover and adapt their circumstances and situation to minimize ongoing impacts and mitigate or prevent effects from future events.

The social aspects of vulnerability are recognized although it is acknowledged that little is known about them as they are difficult to quantify. Social vulnerability is partially a product of social inequalities and also includes place inequalities, i.e., differences in community such as level of urbanization,

infrastructure, and governance. In 2003, Cutter, Boruff, and Shirley constructed an index of social vulnerability called the Social Vulnerability Index by looking at conditions that contribute to making people or places vulnerable to extreme natural events. They also looked at communities' resistance or resilience to hazards. Using a factor analysis approach and based on US data, they found that there were 11 factors that accounted for over 75% of the variance in vulnerability. Factors identified included personal wealth (of countries), age (elderly and young especially vulnerable), density of the built environment (manufacturing and commercial establishments) housing, single sector economic dependence (especially agricultural as it is very climate dependent), housing stock (quality), tenancy (home ownership, renters, mobile homes, etc), race and ethnicity (marginalization with racial disparities), occupation and finally infrastructure dependence (economic vitality and revenue generating capacity).

Environmental Case Studies

Previous discussion has focused on the magnitude and scope of the possible effects social determinants may have on the outcomes of environmental conditions and incidents. It should be noted that environmental incidents themselves can also have a significant impact on social and psychosocial determinants such as employment, social exclusion, social support, socioeconomic status, and in particular, stress. This section will discuss three very significant twenty-first century examples of environmental phenomenon or *disasters* that have had far reaching social as well as environmental implications: Climate change, the 2010 Deepwater Horizon oil spill, and the 2011 nuclear accident in Fukushima, Japan. Climate change is a unique environmental phenomenon in that it has the potential to affect ecosystems and populations all over the world. The drivers of climate change have been occurring over decades, basically since the industrial revolution, with changes to our surroundings so subtle that they have gone relatively unheeded for many years. The other two examples involve technological failures that have caused immediate devastation over a more defined area and both continue to be the subject of long-term ongoing attempts at cleanup and restoration.

Health Impacts of Climate Change

Climate change is an emerging threat to public health that has the potential to have extremely significant health and social impacts, as both a direct and indirect consequence of the effects on the environment. The environmental manifestations range from mild fluctuations in weather patterns to the catastrophic destruction that may result from an increase in both the frequency and intensity of extreme weather events such as storms, cyclones, flooding, and heat and cold waves. While there may be some positive effects of climate change, such as lower cold-related mortality, changes in agricultural yields, or vector survival rates in some areas (Confalonieri et al., 2007), the majority of impacts are predicted to be negative. There are also predicted to be significant social effects from the potential displacement of large numbers of people both from coastal areas in low lying countries and agricultural areas that are no longer viable or able to sustain crops and livestock. This displacement may lead to the loss of social infrastructure and support, unemployment and poverty, as well as psychological distress.

Climate change has the potential to cause a broad range of impacts by various direct and indirect pathways (**Figure 2**). Human health may be affected directly through phenomena such as heat and cold waves, drought and extreme weather events such as storms, tsunamis, and cyclones. Indirect effects that may be detrimental to health generally result from changes to existing conditions such as an increase in the formation, dispersion, and deposition of air pollutants such as ozone and particulates and changes in water quality and availability. There may also be corresponding changes in vector ecology and ecosystems, with subsequent changes in communicable disease as well as changes to agriculture and food quality, industry, and settlements. Climate change can also cause social and economic disruption by forcing the displacement of populations following adverse weather conditions such as droughts and floods, extreme weather events, and also the loss of coastal regions due to rising water levels. There may also be a financial burden associated with property and personal loss as well as a reduction in personal wealth associated with displacement (Kreigler et al., 2012; WHO, 2013b).

As illustrated in **Figure 2**, environmental, social, and health system conditions all have the potential to have a modifying influence on the health impacts of climate change as they are all fundamental to determining adaptive capacity. From an environmental aspect, the prevailing environmental conditions will have a large bearing on the outcomes of any change. The higher the standard of water and air quality, the lesser the effect is likely to be. In countries that already have poor water, air, and soil quality, a further reduction in quality may have a devastating effect. The geography of a community will also be important with low lying or drought-prone areas being particularly vulnerable. Social conditions such as community organization, employment, and economic status will all be important in determining the resilience and adaptability of a community or county to the possible effects of climate change. Health system conditions such as access, affordability, and quality of health care will also be important determinants of health impacts (Confalonieri et al., 2007).

As a relatively recently recognized and studied phenomenon, there is little long-term epidemiological evidence about the effects of climate change. Research into climate change impacts and adaptation is particularly complex due to the many uncertainties surrounding the various ecosystems' responses to warming, as well as the difficulty associated with contextualizing the heterogeneity in impacts and adaptive capacity (Kreigler et al., 2012). In many parts of the world there has been very limited funding available especially where there has been a history of skepticism regarding the scope and magnitude of the predicted outcomes. Despite the limited research that has occurred, it is generally acknowledged that the key to mitigating the potential effects lies in a country or a population's adaptive capacity. Many developed countries appear to have refocused their priorities and rather than trying to reduce greenhouse emissions and slow the potential effects, they are instead investing in researching adaptation strategies to reduce the adverse consequences for various future scenarios.

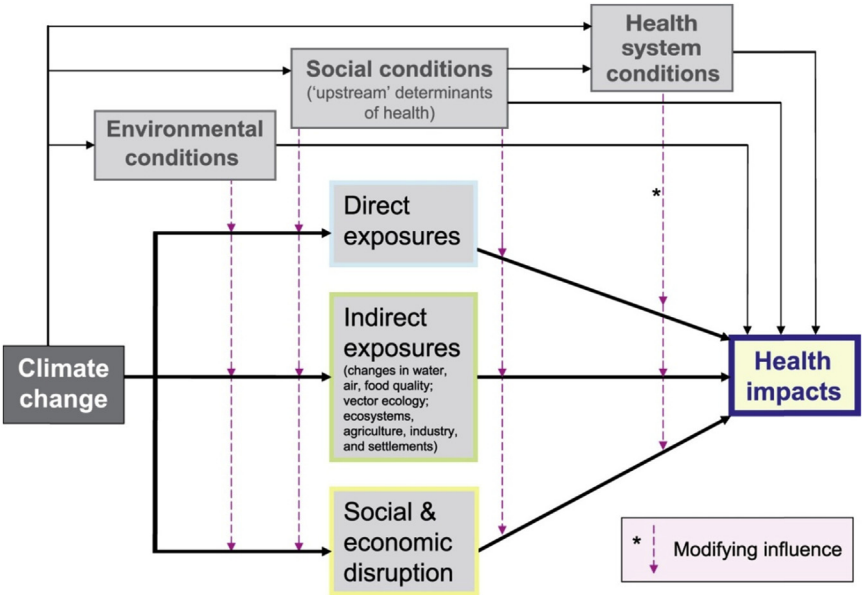

Figure 2 Schematic diagram of pathways by which climate change affects health, and concurrent direct-acting and modifying (conditioning) influences of environmental, social, and health system factors. Confalonieri, U., Menne, B., Akhtar, R., Ebi, K., Hauengue, M., Kovats, R., Revich, B., Woodward, A., 2007: Human health. Climate change 2007: impacts, adaptation and vulnerability. In: Parry, M., Canziani, O., Palutikof, J., van der Linden, P., Hanson, C. (Eds.), Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge.

Given that these strategies require significant financial investment, it follows that developing nations who have extremely limited finances and therefore a limited capacity to adapt will be impacted far more severely. Many of these countries already have high rates of disease and debility and are less able to cope successfully with stresses of all kinds, including the environmental and social impacts of climate change (NCCARF, 2013; WHO, 2013b).

Deepwater Horizon Oil Spill

In April 2010, a system malfunction on the Deepwater Horizon mobile offshore drilling rig, situated in the Gulf of Mexico, led to an explosion that killed 11 workers, injured 17 others, and caused large quantities of crude oil to be released. By the time the well was capped 87 days later, the equivalent of 4.9 million barrels of crude oil had escaped into the Gulf, making it the largest oil spill in recorded history (Ramseur and Hagerty, 2013). The oil slick covered 6500 km^2 of ocean and polluted hundreds of kilometers of beaches, wetlands, and estuaries on the Gulf leading to a complete loss of vegetation in some areas and the death of thousands of fish, birds, and mammals. The cleanup was also surrounded in controversy over the use and quantity (at least 6.8 million liters) of dispersants which contained both carcinogens and endocrine disrupting chemicals, and are now known to have increased the toxicity of the oil up to 50 times through a synergistic effect. There were also serious environmental concerns about in situ burning of the oil, which released dioxins and other harmful chemicals into the atmosphere (Rico-Martinez et al., 2013).

Following the oil spill, fishing areas were closed and fears about sea food safety lead to a dramatic reduction in local sea food harvesting even in areas away from where the oil spill reached resulting in people working in the commercial fishing industry losing their employment and sea food prices rising in the USA (Grattan et al., 2011). There was also a decrease in tourism in the coastal areas affected by the oil spill with a loss of over 400 000 tourism jobs (Diaz, 2013). Lack of fishing and tourism caused substantial economic hardship to the residents in the area (Cleveland, 2013).

Oil spills contain a mixture of chemicals that can include benzene, polycyclic aromatic hydrocarbons, toluene, xylenes, ethylbenzene, and other compounds that may be toxic to human health (Mendelssohn et al., 2012). An article by Levy and Nassetta (2011) analyzing previous studies on the health effects of oil spills identified that in addition to acute and chronic physical health effects, oil spill cleanup workers also suffered psychological disorders. In 2010, a study of 1200 residents who lived within 16 km of the affected coastline reported that one-third of parents indicated that their children experienced mental health distress or physical ill-health (mainly respiratory symptoms and skin irritations) as a direct result of the oil spill (Abramson et al., 2010). Twenty-five percent of the coastal residents interviewed were planning to move from their home and the coastal area due to the effects of the oil spill. Children's leisure activities were also changed with 84% of parents decreasing their children's swimming time, 81.8% decreasing fishing time, 73.2% decreasing boating time, and 65% of parents believing that local sea food was not safe to eat. Woodward (2010: p. 1292) reported that Gulf States residents who lived close to the areas of the oil spill, and people whose business had decreased or closed due to the oil spills "have been exhibiting signs of acute anxiety, depression, increased and excessive drinking, and

suicidal ideation" which may contribute to mental health issues in children.

Research conducted by Lee and Blanchard (2012) found that after the Deepwater Horizon disaster, community attachment, and the fear of losing members of the community, was a problem, particularly for residents who had lived in an area for generations working in the fishing industry or in the oil industry, or who were of Cajun descent, who were unwilling to relocate when there was no longer work available in their community. These residents were highly susceptible to psychological trauma when economic loss occurred. Community indignation about the Deepwater Horizon oil spill amplified resident's fear, anxiety, and worry (Lee and Blanchard, 2012). Grattan et al. (2011) conducted research comparing the psychological impacts of the Deepwater Horizon oil spill on people who had direct experience dealing with the spill with those that suffered economic loss due to the oil spill, but who were located away from the coast where the oil spill reached. They found that participants with oil spill related income loss had higher scores for being depressed, having tension, anxiety, fatigue, confusion, mood disturbances, and had lower scores for being resilient. They were more likely to use behavioral disengagement (withdrawal from society) as a coping strategy. Levy and Nassetta (2011) reported post-disaster changes in the community including struggling over where to place blame. Poor risk communication following the event with authorities being evasive and unresponsive led to the community, particularly affected individuals becoming suspicious and cynical.

Fukushima Nuclear Power Plant Accident

In March 2011 an earthquake of magnitude 9 in the Pacific Ocean (Tōhoku) triggered a tsunami that inundated the north east coast of mainland Japan. This caused the loss of cooling to four reactors at the Fukushima Daiichi nuclear power plant, which led to a partial meltdown and exposure of spent fuel stores and the consequent release of significant amounts of radiation into the surrounding environment (Murayama, 2012; Prati and Zani, 2012). The incident rated 7 on the International Atomic Energy Agency rating scale, which is the highest possible level and equal to the level given to the catastrophic meltdown that occurred at Chernobyl in 1986, even though the total amount of radiation released was considerably less. In a 2011, discussion paper questioning the ethical issues surrounding nuclear energy, Blowers (2011: p. 73) described the event as "a cataclysmic event, creating widespread ecological devastation, displacement of population, economic catastrophe, social disruption and psychological trauma." The author also noted that the effects are likely to persist far into the future. An independent investigation by a National Diet of Japan (2012) Commission into the accident reported that while the cause of the accident was clearly attributable to natural phenomena, the resulting nuclear accident was "a profoundly manmade disaster – that could have and should have been foreseen and prevented" (p. 9).

Although no short-term radiation exposure fatalities were reported, there were approximately 18 500 deaths directly related to the earthquake and tsunami. Estimates of the health and social impacts from exposure to radiation are only speculative, however, as based on the experiences with Chernobyl, there is likely to be a long-term, possibly permanent disruption and dislocation for people who lived and worked in what is now an exclusion zone. They have been forced to abandon their homes, communities, livestock, and livelihoods at short notice and are no longer able to live in their own homes or neighborhoods (Blowers, 2011).

The differing views on the health effects of low radiation doses made communication about radiological health risks particularly difficult. After the accident the level of risk communicated to the public differed significantly between local and national governments. Several standards for radioactive substances in food and drinking water and for radiation levels in school facilities were relaxed leading to distrust in the judgment of the national government (Murayama, 2012). The survivors felt powerless and lacking control over their future, especially when faced with unknown future health effects for themselves and future generations (Blowers, 2011). In the 5 months following the incident approximately 27 000 people moved out of the Fukushima Prefecture. The continued decrease in population in the area made it clear that anxiety about radiation risk had a significant social effect leading to the collapse of local communities in the Prefecture. Localized citizen activist groups also emerged and anxiety about radiation risks also appeared to have an adverse effect on the price of agricultural produce in the area.

Nuclear energy presents an ethical dilemma in that worldwide concerns about the effects of carbon dioxide levels on climate change have in fact driven the push for the development of alternative noncarbon dioxide emitting energy sources. The incident at Fukushima occurred at a time when people were beginning to accept nuclear power as safe, *clean*, and reliable. Twenty-five years on from Chernobyl, it appeared that it was now an economically viable alternative and proponents promote it as an extremely efficient, low-carbon emitting option to highly polluting fossil fuels. Environmental concerns have been pushed aside by fears of climate change, energy security, and concern about dependence on foreign energy supplies (Blowers, 2011). While it is unlikely that the Fukushima incident will ever be replicated, the demand for cheap, efficient, and reliable power puts increasing pressure on new high-risk technologies that are being developed on a larger and faster scale than ever before.

In addition to influencing the feelings of those most directly affected by the disaster, the Fukushima accident has also changed public perception in other countries not affected. In 2012, Prati and Zani hypothesized that a major nuclear accident may affect the most durable cognitive and cultural foundations that underpin public perceptions of nuclear power. Their studies included only Italian participants and reported decreases in nuclear trust, environmental organization trust and pronuclear attitudes as well as changes in risk perception, and antinuclear behavioral intentions. Despite a relatively small sample size, they concluded that major accidents may have the potential to influence values and proenvironmental beliefs. Their research provides support for the notion that a nuclear accident may have a deep impact on public views regarding nuclear power in the short term.

Sustainability and Future Challenges

The environment is a major determinant of human health and well-being. Despite advances in knowledge regarding the links between human health and the environment, there continues to be unprecedented exploitation of the natural environment to meet the needs and desires of the current population. Natural resources such as water, air, and fertile land which are fundamental to our existence continue to be exploited, with humans today extracting and using approximately 50% more of our natural resources than we did 30 years ago (SERI, 2009). There is growing concern that if scarce resources are depleted at an unsustainable rate they will not be available to future generations. A 2009 Sustainable European Research Institute report on the global use of resources noted the inequitable consumption of resources across the planet. People in rich countries were found to consume up to 10 times more natural resources than those in the poorest countries, even though many of the resources have actually been extracted from the poorer countries. In many cases the rush for more secure, reliable, and financially viable energy sources has often meant that the precautionary principle, as it applies to development, has been set aside, leading to potential exposure to both traditional and new hazards from technological advances such as nuclear energy and fracking.

Earlier this century Al Gore, a former US presidential candidate, took environmental issues to the world arena and in doing so raised awareness of the links between our actions, the health of the environment, and impacts on human health. At a grassroots level this increased awareness has empowered groups and communities and given rise to an increasing number of environmental or *green* groups whose agenda is fundamentally the protection of the environment and ultimately human health. Concern for the environment and human health is now firmly entrenched on the political agenda, with the governments of many countries now containing members from parties whose platform is based on socially responsible and humanitarian sustainable development.

Environmental and social determinants both play a very significant role in the health and well-being of all inhabitants of the earth, whether they live in affluent communities in developed countries, the slums of poorer developing countries, or belong to a vulnerable group. The challenge for the future is to recognize the strong connections between these determinants if sustained improvement in global community health and well-being, as well as protection of the environment, is to be achieved.

See also: Ecology and Health; Globalization and Health; Migration and Health; Natural Disasters: Health-Related Aspects; Social Epidemiology; Socioeconomic Status and Health; Urban Life and Health.

Bibliography

Abramson, D., Redlener, I., Stehling-Ariza, T., Sury, J., Banister, A., Park, Y., 2010. Impact on Children and Families of the Deepwater Horizon Oil Spill. Preliminary Findings of the Coastal Population Impact Study. National Centre for Disaster Preparedness, New York.

Blowers, A., 2011. Why Fukushima is a moral issue? The need for an ethic for the future in the debate about the future of nuclear energy. Journal of Integrative Environmental Sciences 8 (2), 73–80.

Confalonieri, U., Menne, B., Akhtar, R., Ebi, K., Hauengue, M., Kovats, R., Revich, B., Woodward, A., 2007. Human health. Climate change 2007: impacts, adaptation and vulnerability. In: Parry, M., Canziani, O., Palutikof, J., van der Linden, P., Hanson, C. (Eds.), Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge.

Cleveland, C., 2013. Deepwater Horizon Oil Spill. The Encyclopaedia of Earth. http://www.eoearth.org/view/article/161185/ (accessed 10.12.13.).

Cutter, S., Boruff, B., Shirley, W., 2003. Social vulnerability to environmental hazards. Social Science Quarterly 84 (2), 242–261.

Dahlgren, G., Whitehead, M., 1991. Policies and Strategies to Promote Social Equity in Health. Institute for Further Studies, Stockholm.

Diaz, J., 2013. The Deepwater Horizon Gulf oil spill disaster. In: Lansman, L. (Ed.), Cases in Public Health Preparedness and Response to Disasters. Jones & Bartlet Learning, Burlington, MA, pp. 149–193.

Earl, C., Terblanche, J., Patten, E., 2012. Health and the living environment. In: Liamputtong, P., Fanany, R., Verrinder, G. (Eds.), Health, Illness, and Well-being: Perspectives and Social Determinants South. Oxford University Press, Melbourne, pp. 154–173.

Grattan, L., Roberts, S., Mahan, W., McLaughlin, P., Otwell, W., Morris, J., 2011. The early psychological impacts of the Deepwater Horizon oil spill on Florida and Alabama communities. Environmental Health Perspectives 119 (6), 838–843.

Health Council of the Netherlands, 2012. Social Aspects of the Living Environment in Relation to Environmental Health. Health Council of the Netherlands, The Hague. Publication no. 2012/10E ISBN 978-90-5549-995-7.

Kreiger, E., O'Neill, B., Hallegatte, S., Kram, T., Lempert, R., Moss, R., Wilbanks, T., 2012. The need for and use of socio-economic scenarios for climate change analysis: a new approach based on shared socio-economic pathways. Global Environmental Change, 807–822.

Lee, M., Blanchard, T., 2012. Community attachment and negative affective states in the context of the BP Deepwater Horizon disaster. American Behavioural Scientist 56 (1), 24–47.

Levy, B., Nassetta, W., 2011. The adverse health effects of oil spills: a review of the literature and a framework for medically evaluating exposed individuals. International Journal of Occupational and Environmental Health 17 (2), 161–167.

Liamputtong, P., Fanany, R., Verrinder, G., 2012. Health, illness and well-being: an introduction. In: Liamputtong, P., Fanany, R., Verrinder, G. (Eds.), Health, Illness and Well-Being: Perspectives and Social Determinants. Oxford University Press, South Melbourne, pp. 154–173.

Margai, F., 2012. Environmental Health Hazards and Social Justice: Geographical Perspectives on Race and Class Disparities. Earthscan, London.

Marmot, M., Wilkinson, R. (Eds.), 2006. Social Determinants of Health, second ed. Oxford University Press, Oxford.

Mendelssohn, I., Andersen, G., Baltz, D., Caffey, R., Carman, K., Fleeger, J., Joye, S., Lin, Q., Maltby, E., Overton, E., Rozas, L., 2012. Oil impacts on coastal wetlands: implications for the Mississippi River Delta after the Deepwater Horizon oil spill. Bioscience 62 (6), 562–574.

Murayama, T., 2012. Social Impacts Induced by Radiation Risk in Fukushima. IAIA12 Conference Proceedings, Energy Future. The Role of Impact Assessment, 32nd Annual Meeting of the International Association for Impact Assessment. Porto, Portugal.

National Climate Change Adaptation Research Facility (NCCARF), 2013. http://www.nccarf.edu.au/ (accessed 10.12.13.).

National Diet of Japan, 2012. The Official Report of the Fukushima Nuclear Accident Independent Investigation Commission Executive Summary. http://www.nirs.org/fukushima/naiic_report.pdf (accessed 10.12.13.).

Prati, G., Zani, B., 2012. The effect of the Fukushima nuclear accident on risk perception, antinuclear behavioural intentions, attitude, trust, environmental beliefs, and values. Environment and Behaviour 45 (6), 782–798.

Pruss-Ustun, A., Corvalan, C., 2006. Preventing Disease through Health Environments: Towards an Estimate of the Environmental Burden of Disease. World Health Organisation, Switzerland.

Ramseur, J., Hagerty, C., 2013. Deepwater horizon oil spill: recent activities and ongoing developments. Congressional Research Service. http://www.fas.org/sgp/crs/misc/R42942.pdf (accessed 10.12.13.).

Rico-Martinez, R., Snell, T., Shearer, T., 2013. Synergistic toxicity of Macondo crude oil and dispersant corexit 9500A® to the Brachionus plicatilis species complex (Rotifera). Environmental Pollution 173, 5–10.

Sustainable Europe Research Institute (SERI), Global 2000, Friends of the Earth Europe, 2009. Overconsumption? Our use of the world's natural resources.

Wilkinson, R., Marmot, M. (Eds.), 2003. Social Determinants of Health. The Solid Facts, second ed. World Health Organisation, Denmark.

Woodward, C., 2010. Gulf oil spill exposes gaps in public health knowledge. Canadian Medical Association Journal 182 (12), 1290–1292.

World Health Organisation Commission on Social Determinants of Health (WHO CSDH), 2008. Closing the Gap in a Generation: Health Equity through Action on the Social Determinants of Health. Final report of the Commission on Social Determinants of Health. World Health Organisation, Geneva.

World Health Organisation, 2013a. Environmental Health. http://www.who.int/topics/environmental_health/en/ (accessed 10.12.13.).

World Health Organisation, 2013b. Climate Change and Human Health. http://www.who.int/globalchange/en/index.html (accessed 10.12.13.).

Environmental History

Vinita Damodaran, University of Sussex, Brighton, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 7, pp. 4621–4627, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article looks at the environmental history of the empire by looking at histories of environmental change in the tropics by pioneering colonial scientists and by contemporary historians who are carving out the domain of colonial environmental history. The article argues that the development of an environmental sensibility can be traced to the encounter of seventeenth and eighteenth century Western Europeans especially naturalists, medical officers and administrators with the startlingly unfamiliar environments of the tropics and with the damage done to these environments by them.

Introduction

What is environmental history and why should we do it? Environmental History has been described variously as the interdisciplinary study of the relations of culture, technology and nature through time by historians such as Donald Worster and as the historically documented part of the story of the life and death, not of human individuals but of societies and species, in terms of their relationship with the world around them by Richard Grove and Mark Elvin. Clearly some environmental historians argue from a materialist/structuralist perspective while others argue from much more of a cultural perspective. There is some disagreement about whether the natural world constitutes any kind of order or pattern that we can know and, if it does, whether that order can be apprehended by means of science or not. There is also debate on what is natural and what is not, whether indigenous people managed the whole environment or only some part of it, how much was wilderness and how much was mythical. There are divergent opinions over the extent to which nature influences human affairs, some taking the position of limited environmental determinism, others insisting that culture determines all. Worster believes in straddling both worlds by asserting that the cultural history of nature is as significant as the ecological history of culture. In considering how the field has developed historians have given a prime role to the workings of nature independent of human actions while at the same time continuing to place more of an emphasis on human interactions with the rural and nonarable environment. As Caroline Ford had argued many of the recent studies in environmental history stress the blurred aspect of the nature–culture divide. This article attempts to explore the various themes within colonial environmental history by examining the work of some of its practitioners.

Themes and Contexts in Colonial Environmental History

Environmental history, then, seeks to address the lacunae of the absence of nature in the study of history by developing new perspectives on the historian's enterprise. In histories of the empire, for example, this absence is particularly marked considering the fact, as Richard Grove argues, that the development of an environmental sensibility can be traced to the encounter of seventeenth and eighteenth century western Europeans, especially naturalists, medical officers and administrators with the startlingly unfamiliar environments of the tropics and with the damage done to these environments by them. Thus in Grove's work this unnatural history of the empire has been revised and the growing domain of environmental history has taken root in studies of the empire. Both Grove and Worster argued strongly for the need to think globally about environmental history. Donald Worster notes that one needs 'to take an all inclusive view to study the planet as a single integrated system that has been radically reorganized by a single, integrated economy, technology and culture … a point in history when people on every continent began to experience the same reality…and to satisfy their personal needs by drawing bon the most remote parts of the earth.' He notes that it is not easy to put precise dates on that era, which is still incomplete, but we should acknowledge as foundational the discovery of the western hemisphere, the invention of new communication and transportation technologies, and the appearance of worldwide markets. At its core, global environmental history must deal with capitalism as the pioneering, and still the most important, architect of that new integrated world economy.

As a discipline, Environmental history seems to have originated in Europe and America in part as an offshoot of the postwar wave of environmentalism which first made itself felt in the 1960s in Europe and America. Rachel Carson's *The Sea around us* (1952) and *Silent Spring* (1962) were a clarion call not just for new environmentalism but for environmental history. In 1967, there appeared another volume of quite extraordinary breadth and depth, Clarence Glacken's, *Traces on the Rhodian shore*. This book really marks a transition between historical geography and environmental history, although in some respects it is a unique study in the history of environmental and cultural change. The 1970s saw the growing strength of environmental history in the US with publications of books by authors such as Roderick Nash, Alfred Crosby and Donald Worster. By the 1980s and 1990s other books followed notably by Richard Grove, Carolyn Merchant, William Cronon, Donald Hughes and John McNeill. Most of the reviews of environmental history argue that the discipline was most firmly rooted in the US and had only a limited following elsewhere including

in Europe. However Grove's pioneering work has attempted to shift the discussion away from the American debates seeking instead to uncover the historiography of environmental history which he believed developed in the colonial tropics in the context of the Dutch, French, and English maritime empires. I shall discuss in this article specifically about Worster's and Grove's studies, which are a useful background to world environmental history and self consciously have carved out the field of environmental history and global environmental history. As Worster has recently noted, 'To think globally about environmental history means transcending national boundaries or local concerns that today bind all peoples and all ecosystems together and to understand how that happened and what the consequences have been.'

Worster's early work *Nature's Economy* traced the history of ecological ideas from Gilbert White, who wrote the *Natural History of Selbourne* in 1789 to Linnaeus whose essay 'The Oeconomy of Nature' in 1749 became the single most important summary of the world ecological point of view still in its infancy. Man and his ambitions in the natural economy are an integral part of the Linnaean model, and occupy a special place of dignity and honor.

> All these treasures of nature, so artfully contrived, so wonderfully propagated, so providentially supported through her three kingdoms, seem intended by the creator for the sake of man. Everything may be made subservient to his use; if not immediately, yet mediately, not so to that of other animals. By the help of reason man tames the fiercest animals, pursues and catches the swiftest, nay he is able to reach even those, which lie hidden in the bottom of the sea.

Worster then moves on to look at Thoreau and Charles Darwin. For Darwin the discovery of the Galápagos Islands introduced a counternarrative to the Arcadian vision of Gilbert White and Thoreau. In the Argentine pampas, especially along the La Plata river, the arrival of a European population with its domestic stock had worked disaster on the aboriginal order of nature. 'The countless herds of horses, cattle and sheep,' not only have altered the whole aspect of the vegetation, but they have banished the guanaco, deer and ostrich,' he wrote. Darwin had enough evidence to believe that the present order of animals was by no means the first to live on this land, and that death was as possible for a whole ecological system as it was for any of its members. Such a realizaion made its contribution to Darwin's awareness of the potent forces arrayed against the living. Extinction and conflict were far from the qualities of Arcadia. Wherever he turned he saw the 'universal signs of violence' Among Darwin's chief guides were Lyell's *Principles of Geology* published 1830 and Alexander Von Humboldt's *Personal Narrative* of 1807, a pioneering work, one might argue, in ecological biology. This is all very informative and interesting. However what Worster fails to point out is that perhaps most important among the 16 major books which Darwin carried with him on the Beagle was that by Alexander Beatson, an Indian army engineer who published *Tracts Relative to the Island of St Helena* in 1816. This book included a listing by William Roxburgh of the endemic plants of St Helena Island and comments on their rates of extinction. This became vital in the construction of Darwin's theory of natural selection.

So too, Worster in attempting to examine the origins of ecology, looks exclusively at the Anglo-American world, neglecting the development of understandings about man and nature in the tropical world. To understand this we have to look at the work of Richard Grove in the 1980s and 1990s and to his pioneering work in locating the emergence of a unique environmental awareness, not in the metropolitan center but in the colonial periphery from a very early period. Much of this awareness was based on the development of climatic anxieties. Indeed world environmental history is today increasingly concerned with the historical impact of global climatic anomalies, a major theme in Grove's work, which I now want to discuss.

In *Green Imperialism*, Grove argues that the emergence of a truly global environmental awareness was a very specific historical development. It depended on new empirical knowledge of the scale of the world and actual observations of human ability to change the natural environment on a global basis. Global environmental awareness was thus directly connected to a new capacity for people to travel large distances, transform the earth, and acquire knowledge about the environment. Two elements were necessary for this awareness to emerge: first, the institution of capital and shareholder-rich maritime trading companies backed by state legislation and assistance; and second, the settlement of previously uninhabited islands and continental peninsulas in the tropics and subtropics by colonial settlers and planters. The profit motives and mechanisms of these trading companies, especially of the East India Companies of Portugal, the Netherlands, Britain, and France, resulted in intensive cash-crop plantation activities on oceanic islands and the clearing of forests for agriculture, fuel wood and ship construction.

This process had already begun with Portuguese and Spanish settlement and plantation agriculture on the Azores, Canaries, and Madeira Islands during the fourteenth century, but the sheer scale of its impact was massively expanded as the European trading companies developed their routes to India, the East Indies, and the Caribbean. As early as the 1670s, the catastrophic consequences of their capital- and labor-intensive activities became clear as the early island colonies experienced drought due to the drying up of perennial streams, soil erosion, dust storms, and the disappearance of animal and plant species. These developments all made practical survival on oceanic islands difficult and encouraged wider questions about the sustainability of a confined settlement. Islands soon became symbolic of the explored world and encouraged ideas about limited resources and the need for conservation or sustainability. (The idea of regional environmental degradation or control was not new; indeed the word 'conservancy' was first adopted in Britain in the 14th century with relation to the control of whole river basins, such as that of the Thames river. Similarly, in the Venetian Republic, well-developed ideas existed about the control of deforestation in the hills in order to control erosion and silting downstream. These initiatives may have been early signs of responses by new, highly sophisticated maritime states to the first consequences of early merchant capitalism and trade that had a global reach. Indeed, even before the advent of large continental-based European empires in Asia, Africa, and the Americas, the scale of artificially caused environmental change

was already being transformed as European maritime countries started to exploit new kinds of natural resources on a global scale. Sugar and other crops essential to the new urban markets of Europe were cultivated on small islands, especially in the West Indies, Indian Ocean, and East Indies. After about 1400, fisheries extended to an oceanic scale as seals and whales were hunted from pole to pole. See Grove, *Green Imperialism, Colonial Expansion, Tropical Island Edens and the Origins of Environmentalism, 1600–1800*.)

The extension of what Immanuel Wallerstein has called the capitalist "world system" on a global scale between 1200 and 1788 had a critically important dimension in terms of resource exploitation. European merchants and companies found that they could exploit the trade goods, markets, and resources of almost every land, in what became an expanding commodity frontier, the "unending frontier" in the title of John Richards important recent book on the environmental history of the early modern world. However, capitalist accumulation and trade developed quite autonomously in South and East Asia, something Wallerstein was really unable to incorporate in his global theory. But in the centers of indigenous capital too, major transformations of the natural landscape took place. Some, like the deforestation of the Ganga basin, had already been long in progress, but they were quickly accelerated after 1400 as powerful mercantile empires developed. However, it was in the tropics that the destructive environmental impact of globalization made its most obvious impact to the observer. One of the first symptoms of the early phases of globalization was the marginalization, enslavement, and then extinction of small indigenous cultures, especially those of island peoples; the indigenes of the Canary Islands are a classic example. However, it was on uninhabited islands such as St. Helena and Mauritius that the full effects of highly capitalized plantations, forest clearance, and import of alien animals (especially pigs, goats, and rats) were first observed. The extinction of the dodo made a great impression on contemporary naturalists. The fact that oceanic islands were perceived as highly desirable 'Edenic' locations in long-running European cultural traditions served to emphasize the shock of their manifest and rapid degradation. Moreover, their degradation threatened their role as watering and supply stations for company ships. In these circumstances, the colonial governments of many small islands became environmentalist, if only to ensure their own survival and that of their agricultural settlers and slaves.

It is possible to argue that early environmental inquiry seems to have been driven by neo-Malthusian concerns about the possible demise of civilization as a result of the 'limits to growth.' A parallel concern is the discussion of the historical collapse of previous societies linked to apparent resource exhaustion and the failure of institutions to adapt to looming crises in time to prevent disaster. The classic discussion is the debate about Easter Island and the fate of its population doomed by deforestation and overfishing. Neo-Malthusian ghosts haunt the discussion of resource shortages and the failure of adaptation on the part of many societies for example in the recent work of Jared Diamond, *Collapse*. But it is worth noting that the theme of collapse also haunts European history, Environmental matters have been more fundamental to imperial politics than is usually understood. In Richard Grove's terms, colonial ecological interventions, especially in deforestation and subsequently in forest conservation, irrigation and soil 'protection,' exercised a far more profound influence over most people than the more conspicuous and dramatic aspects of colonial rule that have traditionally preoccupied historians. Over the period 1670 to 1950, very approximately, a pattern of ecological power relations emerged in which the expanding European states acquired a global reach over natural resources in terms of consumption and then too, in terms of political and ecological control. The reasons for this are not very complicated, but they need restating more frequently than has been the practice in most discussions of environment.

The argument presented by Grove about the expanding resource frontier of Europe has also been replicated by other environmental historians notably by Alfred Crosby in *Ecological Imperialism* where he argues that the process of imperial expansion, whether in terms of direct conquest in what Crosby calls the neo-Europes, or indirect disruptions as a consequences of trading patterns and military actions, fundamentally changed many ecological processes. The introduction of horses to the Americas, rabbits to Australia, or even the humble potato from the Americas to Europe, changed the environments of these places. This is not the first time such changes had altered the planet's ecology; the emergence of agriculture and the domestication of animals has meant that the Holocene, the geological period since the last glacial episode, has been one of anthropogenic-induced changes in most places, but the accelerating speed and scale of change in the last half millennium is what is most important. The ecological dimension of such imperialism is what needs much more attention than it has received until relatively recently. However, Crosby focuses exclusively on the white settler colonies and omits any discussion of the much extensive regions of the colonial tropics. In Grove's terms, the focus on the political and the administrative dimensions of empire have occluded the practical material impacts of colonization on people's lives and on land, animals, fish, forests and other facets of their ecological contexts. The environment has, in these terms, simply been taken for granted until recently when the ecological dimension of human history, minus the distractions of environmental determinism, is once again being worked into the picture.

Grove argues for example, that the Caribbean and its littoral, along with Bermuda, has been a very important area for working out the processes going on in world environmental history in the context of European economic expansion and globalization. Some of the first comprehensive forest-protection legislation on such colonies was introduced after 1620 in Bermuda and a little later in the Caribbean Leeward Islands. In Montserrat, the mountain forests of the island were protected from felling after 1702 by a rigid ordinance, with the knowledge that unrestricted logging caused soil erosion and flooding on lower grounds and in towns. The Caribbean islands, with their large settler and slave populations, came under sustained ecological pressure at an early date and, as on Mauritius and St. Helena, awareness quickly grew of the physical changes and extinctions brought about by commercial clearance. As early as 1616, measures had been taken to protect the indigenous edible sea birds in Bermuda. By the mid-eighteenth century, overfishing and major reductions in catches were taking place around many now

densely populated islands. Other legislation followed making conservation an integral part of colonial landscape control. Before the 1760s, the effects of colonial economic globalization were addressed on a piecemeal basis in order to protect local food, fuel, timber supplies, and what were already recognized as rare island species. However, in the mid-1760s, responses to deforestation in particular suddenly changed. This was due to the rapid spread of a theory first enunciated in France by Pierre Poivre that linked deforestation to rainfall and regional climate change. In the ensuing century, forest-reserve legislation responding to fears of deforestation-induced climate change slowly began to spread around the world, especially throughout the French, British, and Dutch empires. By the late 1830s and the 1840s, the reiteration of climatic environmentalism by Alexander von Humboldt and Jean-Baptiste Boussingault was being acted upon by environmentally minded scientists and officials working not just only on the islands but also on the large land masses of India, Southeast Asia, Southern Africa, and Australia, where the demands of European colonial empires were now bringing about deforestation at an unprecedented speed. It is to these specifics that we will now turn. The context is the growing discussion of the environment by scientists, travellers and administrators in the context of empire in the nineteenth century.

Early Colonial Scientists, Environmentalism and Environmental History

It is no accident that the earliest writers to comment specifically on rapid environmental change in the context of empires were scientists who were themselves often actors in the process of colonially stimulated environmental change. In a study, Grove and Damodaran have argued that the early pioneers of an environmental critique of the European and American empires depended on having an historical perception of rapid rates of ecological change, and access to evidence for rapid change. As early as the mid-seventeenth century we find that intellectuals and natural philosophers such as Richard Norwood and William Sayle in Bermuda, Thomas Tryon in Barbados and Edmond Halley and Isaac Pyke on St Helena were all already well aware of characteristically high rates of soil erosion and deforestation in the colonial tropics, and of the urgent need for conservationist intervention especially to protect forests and threatened species.

On French colonial Mauritius (the Isle de France) Pierre Poivre and Philibert Commerson framed pioneering forest conservation legislation designed specifically to prevent rainfall decline in the 1760s. In India William Roxburgh, Edward Balfour, Alexander Gibson and Hugh Cleghorn (all Scottish medical scientists) wrote alarmist narratives relating deforestation to the danger of climate change. Their distinctively modern environmentalist views owed a great deal to the precocious commentaries of Alexander von Humboldt in his *Personal Narrative* and in the *Cosmos*. East India Company scientists were also well aware of French experience in trying to prevent deforestation and rainfall change in Mauritius. William Roxburgh, together with Alexander Beatson on St Helena, went on further to observe the incidence of global drought events which we know today were globally teleconnected El Nino events. The rise of imperial networks of information thus enabled the emergence of a new global environmental awareness as well as the first accurate accounts of global change.

The writings of Edward Balfour and Hugh Cleghorn in the late 1840s in particular illustrate the extent of the permeation of a global environmental consciousness and could be said to constitute some of the first writings in world environmental history. In very similar fashion Baron Ferdinand Von Mueller in Australia, George Perkins Marsh and Franklin Benjamin Hough in the United States and John Croumbie Brown in South Africa all wrote multiple and voluminous texts that display formidable textual knowledge of evidence of global environmental change through time, which was used by all these men as material to warn of the dangers of future environmental profligacy and global ruin. It is significant that most of these men published their most important texts during the 1860s, a period which we could appropriately name the 'first environmental decade,' and which embodies a convergence of thinking about ecological change on a world scale that may well have been permeated in part by ontological and existential concerns elicited by the publication of Charles Darwin's *The Origin of Species* in 1859.

It was in the particular circumstances of environmental change at the colonial periphery that what we would now term 'environmentalism' first made itself felt and its colonial proponents were often in a position to make use of historical evidence for environmental change in government records and thus became *de facto* environmental historians. Victorian texts such as Strzelecki's *Physical Description of New South Wales*, Berthold von Ribbentrop's *Forestry in the British Empire*, Brown's *Hydrology of South Africa*, Cleghorn's *Forests and Gardens of South India* and Marsh's *Man and Nature* were not only vital to the onset of environmentalism; they were also meticulously documented works in environmental history. One preoccupation stands out in them above all. This was a growing interest in the potential human impact on climate change, in particular a fear that human activity, especially deforestation, might lead to global desiccation. This fear grew steadily in the wake of colonial expansion and fed into postcolonial fears about desertification articulated by international bodies and global nongovernmental organizations (NGOs). It also critically affected the early direction of environmental history. The preoccupation with climate probably owed much of its strength to notions linking climate, civilizational degeneration and racial types. Many of these notions had first been formally articulated by the Comte de Buffon.

Particularly after the 1860s, and even more after the great Indian famines of 1876- and 1899–1902, these connections encouraged and stimulated the idea that human history and environmental change might be firmly linked. The years immediately prior to 1900 saw a renewed interest developing in what were essentially millennial theories of global desiccation closely related to contemporary convictions about the 'inevitable' extinctions of both indigenous people and large tropical mammals such as the African Elephant. These theories were reinforced by climatic events. By the end of the nineteenth century it was evident that short-term vicissitudes of the African climate were of considerable economic importance. In East Africa the level of Lake Victoria rose suddenly in 1878 after unusually heavy rains. A few months later there was abundant

rain over the Blue Nile's catchment in Ethiopia and disastrous Nile floods followed in Egypt. In Southern Africa in 1862 and between 1881 and 1885 there were severe droughts. Alexander Knox in *The Climate of the Continent of Africa* (1911) pointed to what he saw as a decline in the rainfall of nineteenth century Senegal. The mean at St Louis in the 1830s had been about 600 mm; over the years 1892–1905 it was about 400 mm; in 1903 it was only 125 mm. Over the next decade rainfall was greater in Senegal and the rest of the Sudano-Sahelian but in 1913 there came a drought disaster, the terrible results of famine in Northern Nigeria that year being graphically described by Hastings in *Nigerian Days*. In Egypt the flow of the Nile was phenomenally low, and in Southern Africa there was widespread drought.

During the same period a number of American geographers started to pose a postglacial desiccation of the environments of central Asia and China based on the twin tenets that wet conditions characterized the glacial phases of the Pleistocene and that aridity had increased since the warming of the Pleistocene icesheets in the Holocene. Travellers in central Asia pointed to the occurrence of dry water courses and lakes and abandoned settlements as evidence of this desiccation and suggested that deteriorating environmental conditions had spurred successive nomadic invasions of their more civilized neighbors during periods of increased aridity. A major early exemplar of this group of scholar-travellers was an American, Ellsworth Huntington, a geographer and environmental determinist whose views were formed by his travels and intelligence activities in Central Asia. His first major work, *The Pulse of Asia* (1907) set an agenda for both desiccationism and environmental determinism. Both Huntington and Kropotkin (the latter in a landmark article published in *The Geographical Journal* in 1904) were critically influenced by the rise of contemporary anxieties in the tropics and a growing interest in climatic interpretations of history, boosted by the great Indian famines of the late nineteenth century. While at Harvard, Huntington had been a pupil of William Morris Davis at a time when the latter had been developing his theories of evolutionary landscape geomorphology. Huntington was, together with Andrew Douglass, one of the first practitioners of dendrochronology. They pioneered the use of this new dating technique on the Sequoia trees of California, whose immense age made a great impression on Huntington. But his formative experiences were in Turkey, where he taught at Euphrates College in 1897–1901 and in Turkestan. After Harvard, Huntington accompanied William Davis on the Pumpelly (1903) to Turkestan and then went with Barrett's (1905–6) expeditions to central Asia.

Colonial Environmental Ideas in the Twentieth Century

Throughout the 1920s and 1930s the experience of central Asia, which had exerted so much influence on Huntington, continued to exert an influence on the desiccationist school. The Chinese meteorologist Coching Chu, in 1926, summarized much of this in a paper on 'Climatic pulsations during historic times in China' Geographical periodicals and institutions were, as in the previous century, important as fora for the desiccation debate. In Africa this meant that the concerns of the 1920s now began to embrace some colonial territories that had not featured at all in the earlier environmental literature of the years before the Great War, but which were now the subject of considerable colonial interest and infrastructure investment. In the 1920s the Colonial Office was the main employer of British biologists, geologists and geographers. This was especially the case in Anglo-Egyptian Sudan, about which some of the first literature on desert-spreading or desertification now began to be written. A pioneer in this area was E.W. Bovill, who echoed Schwartz in South Africa in his 1921 paper on 'the encroachment of the Sahara on the Sudan.' His arguments were further followed up in an article entitled 'Sahara' in 1929 Bovill's articles were in turn taken much further by G.T. Renner in one of the first articles to paint Africa as a potentially famine-ridden continent, under the title 'A famine zone in Africa; the Sudan,' published in 1926. A widespread semiarid region meant that a disproportionately large alarmist South African literature developed on environmental matters. Prominent among these was the 1926 work by J.C. Smuts on *Holism and Evolution*, a work heavily influenced by the ecological theories of Arthur Tansley.

The emergence of a period of environmental alarmism in North America consequent on the prolonged 'dustbowl' droughts in the southern United States in the early 1930s supplemented the existing colonial panic over desert-spreading. This increased local anxieties and affected policy in some British and French colonies. Thus the New Deal conservationism of the United States was emulated particularly in the east and central African colonies. Soil erosion had already become a prominent issue in India during the period 1890–1925 and huge investments to control it were made, for example, in the Etawah region of the United Provinces of northern India. These efforts, like similar measures in West and South Africa, long predated the American 'Dustbowl' alarmism. In 1934 E.P. Stebbing, a very prominent Indian forester, and onetime Professor of Forestry at Edinburgh University, visited West Africa. Stebbing was the earliest historian of Indian forests and by the time of his visit has already published a three volume work on *The forests of India*, much of which detailed the history of environmental concern and early conservationism among the first surgeon-foresters of the East India Company medical service. His short visit to the French and British West African colonies, made significantly during the dry season, provoked him into writing a feverish warning on what he saw as the dangers of desert-spreading. The title of this essay 'The encroaching Sahara; the threat to the West African colonies' indicates that he had almost certainly read Bovill's similarly titled 1921 article on 'The encroachment of the Sahara.'

While Stebbing's somewhat hysterical warnings were downplayed by local colonial scientists, who had much greater experience of the causes, rates and seasonality of local desertification and erosion, his terminology was taken up with alacrity by governing circles in Paris and London. Stebbing's writings led directly to the founding of the Anglo-French Boundary Forest Commission. This commission, started in 1934, soon found that Stebbing's warnings were largely unjustified, and his analysis was decisively dismissed by B. Jones, a member of the commission, in an article published, as Stebbing's had been, in *The Geographical Journal*.

Notwithstanding, in 1937 Stebbing, now Professor of Forestry in Edinburgh, pursued his concerns in *Forestry in West Africa and the Sahara; a Study of Modern Conditions* and in an inflammatory work called *The Creeping Desert in the Sudan and Elsewhere in Africa*. He was not alone. In 1938 Francis Ratcliffe, fresh from his investigations of the causes of soil erosion in South Australia and Queensland, published *Flying Fox and Drifting Sand*, a savage indictment of the impact of extensive outback agriculture, and the first of a long line of apocalyptic books on the degradation of the Australian environment, many of which exercised a disproportionate influence on global environmental concerns far outside the antipodean context. Ratcliffe was familiar with the writings of Keith Hancock, later a prominent historian of the British Empire and Commonwealth, who in 1931 had published a virulent attack on profligate deforestation and land-clearing by settlers Hancock, a Quaker, was the biographer of Jan Smuts and was undoubtedly influenced by the 'holistic' views of the latter, not least in his role as an early environmental campaigner and environmental historian. Both men were profoundly interested in the historical impact of white colonial settlement on indigenous peoples and their environments at an imperial and world scale, Hancock ultimately writing on this theme in what was one of the earliest and most influential texts in Australian environmental history. In Hancock's mind the two world wars, Nazi death camps and the ecological crisis were linked catastrophes in a personal existential crisis which was partially expressed in his writing of environmental history.

In 1938 Gordon East published the first of many editions of *The Geography behind History*, a wide-ranging work which echoed many of the contemporary colonial anxieties albeit in a rather more detached and academic mode. Thus by 1938 the global environmentalist ball had truly started to roll, propelled largely by colonial exemplars and Stebbing's alarums, along with comparable simultaneous warnings from the United States which were soon being echoed by a cacophony of popular writers and journalists and above all by Graham Jacks and R.O. Whyte in their inflammatory, semi-racist and highly colored account book of 1939 entitled *The Rape of the Earth; a World Survey of Soil Erosion*. This book skillfully marshaled evidence for the kind of rapid environmental change which had been predicted by John Croumbie Brown and G.P. Marsh 70 years earlier, but referring to world-wide evidence of a detail which they had not been in a position to compile. The book set the scene for the postwar British (and French) colonial obsession with soil erosion and gullying in their postwar 'second colonial occupations,' as well as for the global desertification mania which developed in the 1970s in the wake of major droughts in West Africa. Conceivably, *The Rape of the Earth* owed some of its impact to underlying political anxieties in Britain, France, and the United States about the growing threat now posed by fascist regimes and above all by Nazi Germany both on the domestic and colonial front. Environmental fears were, conceivably, more palatable and thus easier to articulate and confront.

These agendas of colonial environmental thinking had taken over academic debate by the 1930s. Once the discipline of environmental history began to establish itself as a separately defined field of scholarship in the 1970s, these early ideas continued to fuel academic thinking. Grove's work seized the initiative for early environmental scholarship away from the American academy where practitioners believed it to be most deeply rooted to the colonial context of the tropics. Colonial environmental thinking he argued had an important impact on the emergence of the field.

More recently the story of colonial environmental history has developed in different ways. For example in South Asia, a shift toward an ecological questioning of conventional agrarian history developed among historians of South Asia in the early 1980s, and particularly in the minds of Richard Tucker and John Richards, the latter being a specialist in the monetary and agrarian history of Mughal India. Both men were convinced that ecological changes accompanying economic transition in the seventeenth to nineteenth centuries, while clearly large-scale, had never been properly quantified. Almost immediately Richards and Tucker, working in tandem, realized that their questioning could not be confined to South Asia but was equally requiring of answers researched in terms of global economic history in general, particularly with the advent of an era in which the connections between deforestation, carbon dioxide production, and global warming were becoming major popular anxieties in what Teresa Brennan was already calling the 'Age of Paranoia.' In two major edited works they therefore set out to review the global history of deforestation, especially in the tropics. However, in making the quick and logical intellectual leap from South Asia to world history, Richards and Tucker had ironically left the environmental history of South Asia itself largely undone with the exception of some very limited essays, regional studies, and essay collections. This is now being redressed in several recent studies on the environmental history of South Asia. In particular, the work of K. Sivaramakrishnan, Mahesh Rangarajan, Ravi Rajan, Rohan D'Souza, Vasant Saberwal, Dhirendra Dangwal. Not only have these works made a substantial presence in mainstream South Asian history writing but in terms of institutional advance as well. As D'souza notes 'these studies have propelled the South Asian experience to the centerstage of the field of environmental writing.' In some respects the early work in Indian environmental history resembled that in the Americas and Africa, positing an ecological golden age but more recent work is starting to challenge this notion.

Landscape Imagery and Colonial Environmental History

Another burgeoning trend in colonial environmental history is in landscape imagery and construction, which builds on the work in the 1980s and 1990s of Paul Carter, Simon Schama, Denis Cosgrove, and Steve Daniel. In the colonial context, the recent work of Graham Burnett, Matthew Edney, and David Arnold explores the culture of the tropics in implicitly rejecting the nature–culture dualism and blurring the boundaries. This is particularly true of David Arnold's work, *The Tropics and the Traveling Gaze* which examines the invention of the tropics from the eighteenth century. The eighteenth century was inter alia, an age of decorative gardens and of gardeners. One of England's more notable contributions to the neoclassical movement was the landscape garden in which the formal, regular, and ostentatious shapes of the baroque era were swept

away to be replaced with no-less cultivated, but naturalistic landscapes complemented by artificial lakes and classical ruins blended into the surroundings. The picturesque landscape of Capability Brown and Humphrey Repton are examples of these. But, as Addison noted:

> Why may not a whole estate be thrown into a kind of garden by frequent plantations, which may turn as much to the profit as the pleasure of the owner?

The search for more exotic decorative plants was intensified in the new era of exploration as evident from David Mackay's work (*In the Wake of Cook: Exploration, Science and Empire, 1780–1801*, 1985). Eventually, with the voyages of Bougainville and Cook, the Arcadian quest encompassed the south Pacific. The perception of the tropics that explorers and their artists carried with them was one of a naturalistic landscape with lush forests, steep cliff walls and waterfalls, instead of lakes and temples. A naive and noble savage formed part of the scenery as he sheltered in the romantic grotto from the elemental powers of nature. This romantic prospect frequently conflated the dramatic and the verdant landscape of Tahiti with the real economy of its inhabitants. Incidentally the same notions were transferred to the Indian Ocean islands, an apparently Utopian realm that has been little explored by scholars.

In his magnificent study, *The European Vision and the South Pacific* Bernard Smith has shown how exotic topography was elevated to high places reserved for ideal landscapes by the late eighteenth century, and underlines the significance of the fact that the term picturesque was being used to embrace the growing interest in the exotic. As a tradition conceived by Uvedale Price in 1794, the picturesque was perceived as an aesthetic category distinct from the Burkean notion of the sublime and the beautiful; it was distinguished by roughness, sudden variation and irregularity. While artists like William Hodges still sought to make the tropical landscape conform to the taste of their time by adapting it to the Italianate landscape there was an emerging perception in the late eighteenth century among philosophers and writers such as von Humboldt that the noblest landscapes were to be found in the tropics.

This burgeoning interest in landscape imagery and construction, some of which fed into environmental history, explicitly uses Saidian notions. For example in Burnett's work entitled *Masters of All They Surveyed*, on the exploration of British Guyana in the nineteenth century argues that colonial constructions of landmarks became not only 'places to see, but places that enhanced seeing and a technique for establishing a control over space.' He notes that, like the imperial picturesque of India which was used by British topographical surveyors both as a natural representation of the Indian landscape and as an invitation to British conquest, representations of the Guyanan landscape worked in similar ways. Explorers such as Richard Schomburgk, 'were aware of the power of their images and the processes involved in renaming, landmarking and resource assessment leading to the establishment of colonial boundaries and colonial order.' There are problems with using explicitly Saidian notions for understanding colonial engagements with the landscape. Bernard Smith's story is not one of simple appropriation, control and establishment of order through topographical art, but a story of a much more complex relationship between art and ideas. While between 1750 and 1850, a great deal of artistic talent was brought in to serve the biological and geographical sciences in the colonies a 'vigorous and aesthetically vibrant art movement was beginning to emerge in the context of the exploration of the new worlds.' While depiction of plants and animals within a typical landscape became increasingly common landscape painters were beginning to recognize the fact that the tropical world contained its own distinctive sceneries, which had a visual unity of its own. It was in the paintings of William Hodges that these ideas first began to challenge neoclassical values, but it was through the work of Alexander von Humboldt that this new artistic movement was celebrated, influencing such writers and artists as Gustav Carus and John Ruskin and in a circular sense, reinvigorating artistic and environmental sensibilities in Europe. Tim Bonyhady's (2000) book, entitled *The Colonial Earth*, a superb essay on the development of environmentalism in pre-1900 Australia, contains an excellent analysis of this development of independent colonial sensibilities, especially in connection with the work of the painter Eugene von Guerard.

Although what was termed 'nature's wild garden' by the Romantics had its effects on Sir Joseph Banks when he traveled with Cook, the sort of gardens in which he was interested by the 1780s was largely botanic ones. At Kew, Chelsea, and Edinburgh, men were trained in the empirical processes of observation, comparison and evaluation, and they began to gather comprehensive collections of native and exotic plants for microscopic examination. Drawing on the pioneer work in plant classification carried out by Linneaus, Britain despatched plant collectors to every corner of the globe, and these men sent home living plant specimens, seeds and dried samples to their patrons and employers. As early as the seventeenth century in Malabar, the botanist and traveler Van Reede moved further away from European precepts to a wholly Malayali classification of plants. Van Reede's reliance on local informants and his frequent contacts with Tiya and Ezhava 'low-caste' plant collectors helped to incorporate local Ezhavas botanical knowledge into the *Hortus Malabaricus Indicus*. This gigantic twelve-volume work which adhered religiously to local systems of plant classification and which rejected both Arabic classification and nomenclature and Brahminical systems, as well as European knowledge, all of which had proved virtually useless to the realities of plant identification and systematic knowledge as far as the Dutch were concerned. The extraordinary utility of the indigenous Malabar plant knowledge systems and systems of classification as compiled by Van Rheede and Itti Achuden, his main informant, were explicitly and remarkably recognized by Linnaeus. Indeed Linnaeus simply incorporated the Ezhava classification system wholesale, and without changes, when he came to describe the tropical plants of South India.

From the middle of the eighteenth century onwards the utilitarian strain in natural history became more pronounced as scientists and collectors interested themselves in the practical benefits which might be reaped from their endeavors. In many respects, it was thought, the world of nature had not endowed equally the peoples of the earth. Botanists in particular believed that their technical skills would enable them to rectify the imbalance by transferring certain natural productions from one region to another, and acclimatizing them in

the new environment. Periodic famines in India, for example, could be averted by planting Sago or other drought-resistant crops. English cotton mills might be supplied with raw cotton from improved seeds, while transplanted flax could provide cheap fibers. Above all, attempts were made to transplant spice and pepper species into new territories where they were previously unknown.

Such schemes had the mercantilist purpose of destroying the monopolies or predominance of rival nations and colonies and substituting that of another. Particular efforts were made by the French and British to break the Dutch spice monopolies in the East Indies. Botany and great power rivalry became intertwined as nations endeavored to guard their precious treasures. It was in this context that Captain Bligh's expedition took place. The story of Sir Joseph Banks, breadfruit and Tahiti is well known. Less well known is that it stimulated the advent of forest protection in the West Indies, just as spice transplantation to Mauritius had stimulated forest preservation there during the ancient regime.

However most of the studies of the culture of the landscape in the context of empire are about the colonization of space and how landscapes gain a moral and even redemptive significance. Some of it is a postmodern interpretation of landscape imagery and of the landscape as a cultural text. In doing so, writers such as Arnold uses a range of sources mainly Paul Carter, Michel Foucault and Bernard Cohn to see the environment as socially constructed and to suggest the idea that the 'tropics were invented as much as they were encountered' for example by the British in India in the nineteenth century. The idea of the tropics as 'warm, fecund, luxuriant, paradisical and pestilential' was central to the constitution of British colonial knowledge and was a critical ingredient in the larger colonizing process.

Thus we can see that environmental history allows for divergent opinions over the extent to which nature influences human affairs, some taking the position of limited environmental determinism, others insisting that culture determines all. While Crosby, Grove and Diamond fall into the former category, Arnold, Carter, Cosgrove and Daniels fall into the latter. Worster's point about straddling both worlds by asserting that the cultural history of nature is as significant as the ecological history of culture is a pertinent one and gives us new ways of understanding the history of European intervention in the tropical world. This article has attempted to explore the historiography of tropical environmental history which can now claim to be an important subfield within the discipline of environmental history. As John Mcneill has noted 'interest in environmental history depends, in large part, on anxiety about contemporary social problems.' (John Mcneill, 'Observations on the nature and culture of environmental history, *History and Theory*, 42, 2003.) With ecological concerns increasing environmental history and especially tropical environmental history is a burgeoning and innovative field of interest.

See also: Climate, History of; Economic History; Environment and Development; Environmental Archaeology; Environmental Conservation and Development: Critical Perspectives; Human Ecology.

Bibliography

Arnold, David, 2006. The Tropics and the Travelling Gaze. Ohio.

Brennan, Teresa, 2003. Globalisation and Its Terrors. London.

Bonyhady's, Tim, 2000. The Colonial Earth. Melbourne.

Bovill, E., 1921. The encroachment of the Sahara on the Sudan. Journal of the Royal African Society 20, 23–45.

Burnett, Graham, 2000. The Master of All They Surveyed, Exploration, Geography and a British Eldorado. Chicago.

Crumbie Brown, John, 1875. Hydrology of South Africa. London.

Carter, Paul, 1996. The Lie of the Land. London.

Chu, Coching, 1931. Climatic changes during historic time in China. Journal of the China Branch of the Royal Asiatic Society LXII, 32–40.

Chu, Coching, 1926. Climatic pulsations in historic times in China. The Geographical Review XVI (1), 274–282.

Cleghorn, H.F., 1861. The Forests and Gardens of South India. Edinburgh, Oliver and Boyd.

Corona, Gabriella (Ed.), 2008. What is global environmental history. 2, 228–249.

Cosgrove, D., Daniel, S., 1989. The Iconography of the Landscape, Essays on the Symbolic Representation, Design and Use of Past Environments. Cambridge.

Cronon, William, 2003. Changes in the Land, Indians, Colonists and the Ecology of New England. Hill and Wang.

Crosby, Alfred, 2004. Ecological Imperialism, the Biological Expansion of Europe, 900–1900. Cambridge.

Comte de Buffon, Leclerc, G.L., 1749–1804. Paris. Histoire naturelle: Generale et particuliere, 44 vols.

Darwin, Charles, 1859. On the Origin of Species by Means of Natural Selection; or the Preservation of Favoured Races in the Struggle for Life. London.

Davis, W.M., 1889. The rivers and valleys of Pennsylvania. National Geographic Magazine 1, 183–253.

Diamond, Jared, 2005. Collapse, How Societies Choose to Fail or Succeed. New York.

D'souza, Rohan, 2006. The Drowned and the Dammed, Colonial Capitalism and Flood Control in Eastern India. Delhi.

East, Gordon, 1938. The Geography behind History. T. Nelson Publishers, London.

Edney, Matthew, 1999. Mapping an Empire, Geographical Construction of British India. Chicago.

Elvin, Mark, 2004. The Retreat of the Elephants, an Environmental History of China. Yale.

Ford, Caroline, March 2007. Natures fortunes: new directions in the writing of European environmental history. The Journal of Modern History 79, 112–133.

Glacken, Clarence, 1990. Traces on the Rhodian Shore, Nature and Culture in Western Thought from Ancient Times to the End of the Eighteenth Century. Reprint, California.

Grove, Richard, 1995. Green Imperialism, Colonial Expansion, Tropical Island Edens and the Origins of Environmental Consciousness, 1600–1860. Cambridge.

Grove, R., Damodaran, V., Sangwan, S., 1998. Nature and the Orient; the Environmental History of South and Southeast Asia. Oxford.

Grove, Richard, 1998. Indigenous knowledge and the significance of south west India for Portugese and Dutch constructions of tropical nature. In: Grove, Richard, Damodaran, Vinita, Sangwan, Satpal (Eds.), Nature and the Orient, the Environmental History of South and South East Asia, Delhi, pp. 198–204.

Grove, R.H., 2005. Revolutionary weather; the climatic and economic crisis of 1788–1795 and the discovery of El Nino. In: Sherratt, T. (Ed.), A Change in the Weather; Climate and Culture in Australia. National Museum of Australia Press, Canberra, pp. 128–140.

Grove, R., Damodaran, V., 14 and 21 October 2006. Imperialism, intellectual networks and environmental change, origins and evolution of global environmental history, 1676–2000. Economic and Political Weekly, Parts 1 and 2, 41 (41), 4345–4356; 41 (42), 4497–4505.

Guha, R., Gadgil, M., 1992. This Fissured Earth; towards an Ecological History of India. Oxford, India.

Hancock, W.K., 1931. Australia. C. Scribner's Sons, New York.

Hancock, W.K., 1972. Discovering Monaro; a Study of Man's Impact on His Environment. Cambridge University Press, Cambridge.

Hancock, W.K., 1973. Today, Yesterday and Tomorrow; the 1973 Boyer Lectures. The Australian Broadcasting Commission, Sydney, p. 49.

Hastings, A.C.G., 1925. Nigerian Days. John Lane, London.

Hurst, H.E., 1952. The Nile. Constable, London, p. 304.

Hughes, Donald, 1975. Ecology in Ancient Civilisations. New Mexico Press.

Ellsworth Huntington. Explorations in Turkestan (1905); The Climatic Factor (1914); Civilization and Climate (1915); Climatic Changes (1922); Principles of Human Geography (1940); Mainsprings of Civilization (1945).

Huntington, Ellsworth, 1907. The Pulse of Asia, a Journey in Central Asia Illustrating the Geographic Basis of History. Houghton, Mifflin and Company, Boston.
A. von Humboldt, Personal Narrative of Travels to the Equinoctial Regions of the New Continent, 1799–1804 (H. Williams, Trans.), 6 vols., London, 1819; Cosmos (Mrs Sabine, trans.) 2 vols., London, 1848.
Jacks, G.V., Whyte, R.O., 1939. The Rape of the Earth; a World Survey of Soil Erosion. Faber and Faber Ltd, London.
Jones, B., 1938. Desiccation and the West African colonies. Geographical Journal XCI 401–423.
Knox, Alexander, 1911. The Climate of the Continent of Africa. Cambridge University Press, Cambridge, p. 158.
Kropotkin, Prince P., 1904. The desiccation of Eur-Asia. The Geographical Journal 23, 722–741.
Lefroy, J.H., 1877–79. Memorials of the Discovery and Early Settlement of the Bermudas or Somers Islands, 1515–1685. Comp. From the Colonial Records and Other Original Sources. Longmans, Green, and Co., London.
Marsh, J.P., 1864. Man and Nature or Physical Geography as Modified by Human Action. New York.
Mackay's work, David, 1780–1801, 1985. In the Wake of Cook: Exploration, Science and Empire, London.
Mcneill, John, 2001. Something New under the Sun: An Environmental History of the Twentieth Century World. Norton.
Mcneill, John, 2003. Observations on the nature and culture of environmental history. History and Theory 42, 5–43.
Merchant, Carolyn, 1990. The Death of Nature, Women, Ecology and the Scientific Revolution. Harper.
Nash, Roderick, 2001. Wilderness and the American Mind. Yale.
Joe Powell. Signposts to Tracts; Hancock and Environmental History. Unpublished TS. Paper to 1998 Symposium at the Australian National University, Canberra, 1–3 April 1998.
Rajan, Ravi, 2008. Modernising Nature, Forestry and Imperial Eecodevelopment. Oxford.
Rangarajan, Mahesh, 1999. Fencing the Forest. Delhi.
Ratcliffe, Francis, 1938. Flying Fox and Drifting Sand; the Adventures of a Biologist in Australia. R.M. McBride & Co., New York.
Ribbentrop, B., 1899. Forestry in British India. Calcutta.
Richards, John, 2003. The Unending Frontier; an Environmental History of the Early Modern World. Berkeley.
Renner, G.T., 1926. A famine zone in Africa; the Sudan. The Geographical Review 16, 583–596.
Saberwal, Vasant, 1999. Pastoral Politics Shepherds Bureaucrats and Conservation in the Western Himalaya. Delhi.
Schama, Simon, 1996. Landscape and Memory. Fontana.
Sivaramakrishnan, K., 1999. Modern Forests, State Making and Environmental Change in Eastern India. Delhi.
Smith, Bernard, 1960. European Vision of the South Pacific 1768–1850: A Study in the History of Art and Ideas. Clarendon Press, Oxford, England.
Stebbing, E.P., 1937. Forestry in West Africa and the Sahara; a Study of Modern Conditions. W. and R. Chambers, London.
Stebbing, E.P., 1922. The Forests of India, 3 vols. Edinburgh.
Stebbing, E.P., 1934. The encroaching Sahara: the threat to the West Africancolonies. The Geographical Journal 85, 506–524.
Stebbing, E.P., 1953. The Creeping Desert in the Sudan and Elsewhere in Africa: 15 to 13 Degrees Latitude. McCorquodale and Co., Khartoum.
Strzelecki, P.E., 1845. Physical Description of New South Wales and Van Diemens's Land. London.
Tryon, Thomas, 1684. Friendly Advice to the Gentlemen-Planters of the East and West Indies: In Three Parts by Philotheos Physiologus *alias*. Printed by Andrew Sowle, London.
osmas Tsokhas. Reflections of the life and work of W.K. Hancock. The Round Table 314, 213–226.
osmas Tsokhas. A search for transcendence; philosophical and religious dialogues in W.K. Hancock's biography of J.C. Smuts. The Round Table 358, 67–88.
Tucker, Richard P., Richards, John F. (Eds.), 1988. Global Deforestation in The Nineteenth Century World Economy, Durham, 1983; World Deforestation in the Twentieth Century, Durham.
Winiwarter, Verena, et al., 2004. Environmental history in Europe from 1994 to 2004: enthusiasm and consolidation. In: Environment and History, 10, pp. 501–530.
Worster, Donald, 1979. Dust Bowl: The Southern Plains in the 1930s. Oxford University Press, New York.
Worster, Donald, 1994. Natures Economy, a History of Ecological Ideas. Cambridge.

Environmental Justice in the United States

Robert D Bullard, Texas Southern University, Houston, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Despite significant improvements in environmental protection over the past several decades, millions of Americans continue to live in unsafe and unhealthy physical environments. Many economically impoverished and politically powerless individuals are exposed to greater health hazards in their homes, on the jobs, and in their neighborhoods when compared to their more affluent counterparts. A growing body of evidence reveals that people of color and low-income persons have borne greater environmental and health risks than society at large. Over the last two decades, grassroots activists in the United Staes have attempted to change the way government implements environmental, health, and civil rights laws. These groups have also organized, educated, and empowered themselves to make government and industry responsive to their needs.

Introduction

Numerous studies reveal that low-income persons and people of color have borne greater health and environmental risk burdens than the society at large. Environmental justice is defined as the fair treatment and meaningful involvement of all people regardless of race, color, national origin, or income with respect to the development, implementation, and enforcement of environmental laws, regulations, and policies. Fair treatment means that no group of people, including racial, ethnic, or socioeconomic groups should bear a disproportionate share of the negative environmental consequences resulting from industrial, municipal, and commercial operations or the execution of federal, state, local, and tribal programs, and policies. The grassroots environmental justice movement challenged the dominant environmental protection paradigm – a framework that manages, regulates, and distributes risks. On the other hand, the environmental justice paradigm embraces a holistic approach to formulating health policies and regulations, developing risk reduction strategies for multiple, cumulative and synergistic risks, ensuring public health, enhancing of public participation in environmental decision-making, promoting community empowerment, building infrastructure for achieving environmental justice and sustainable communities, ensuring interagency cooperation and coordination, developing innovative public/private partnerships and collaboratives, enhancing community-based pollution prevention strategies, ensuring community-based sustainable economic development, and developing geographically oriented community-wide programming.

Despite significant improvements in environmental protection over the four decades, millions of Americans continue to live in unsafe and unhealthy physical environments. Hardly a day passes without the media discovering some community or neighborhood fighting a landfill, incinerator, chemical plant, or some other polluting industry. This was not always the case. In the 1970s, the concept of environmental justice had not registered on the radar screens of environmental, civil rights, or social justice groups. Nevertheless, it should not be forgotten that Dr Martin Luther King, Jr. went to Memphis in 1968 on an environmental and economic justice mission for the striking black garbage workers. The strikers were demanding equal pay and better working condition. Of course, Dr King was assassinated before he could complete his mission.

Birth of a New Movement

The national environmental justice movement emerged around the issues of fairness, social equity, and environmental protection. This new movement embraces the principle that all communities are entitled to equal protection and enforcement of environmental, health, employment, housing, transportation, and civil rights laws and regulations that impact quality of life. The environmental justice framework rests on developing tools and strategies to eliminate unfair, unjust, and inequitable conditions and decisions. The framework also attempts to uncover the underlying assumptions that may contribute to and produce differential exposure and unequal protection. It brings to the surface the *ethical* and *political* questions of 'who gets what, when, why, and how much' (Bullard, 1996). The national environmental justice movement/framework has its origins in the United States of America. In America, low income and people of color communities are disproportionately affected by toxic waste facilities and dirty industries; as well as, internationally where hazardous wastes move from Organization for Economic Co-operation and Development (OECD) states to non-OECD states (Bullard and Johnson, 2000).

Early Struggles

Another landmark garbage dispute took place in Houston, Texas, a decade after the Memphis garbage strike. African-American homeowners in Houston began a bitter fight to keep the Whispering Pines sanitary landfill out of their suburban middle-income neighborhood. The residents formed the Northeast Community Action Group (NECAG). In 1979, NECAG and their attorney, Linda McKeever Bullard, filed a class action lawsuit to block the garbage facility from being built. The 1979 lawsuit, *Bean versus Southwestern Waste*

Management Corp., was the first of its kind to challenge the siting of a waste facility under civil rights law.

The landmark Houston case occurred 3 years before the environmental justice movement was catapulted into the national limelight in the rural and mostly African-American Warren County, North Carolina. The environmental justice movement has come a long ways since its humble beginning in Warren County where a hazardous-waste landfill ignited protests and over 500 arrests. This marked the first time Americans had been arrested protesting the siting of a waste facility.

The Warren County protests provided the impetus for US General Accounting Office to conduct an independent investigation. That study revealed that three out of four of the off-site, commercial hazardous waste landfills in Region 4 (which comprises eight states in the south) were located in predominantly African-American communities, although African-Americans made up only 20% of the region's population (US General Accounting Office, 1983).

After waiting more than two decades, victory finally came to the residents of predominately black Warren County. Since 1982, county residents lived with the legacy of a 142-acre toxic waste dump. Detoxification work began on the dump in June 2001 and the last clean-up work was slated to end the latter part of December 2003. State and federal sources spent $18 million to detoxify or neutralize contaminated soil stored at the Warren County polychlorinated biphenyl (PCB) landfill. A private contractor hired by the state dug up and burned 81 500 tons of oil-laced soil in a kiln that reached more than 800°F to remove the PCBs. The soil was put back in a football-size pit, recovered to form a mound, graded, and seeded with grass.

The protests also led the Commission for Racial Justice (1987) to produce *Toxic Waste and Race*, the first national study to correlate waste facility sites with demographic characteristics. Race was found to be the most potent variable in predicting where these facilities were located – more powerful than poverty, land values, and home ownership or occupancy status. In 2007, the United Church of Christ commissioned a study, 'Toxic Wastes and Race at Twenty 1987–2007,' that chronicles significant environmental justice milestones since 1987 and the report provide a collection of essays from environmental justice leaders on a range of issues (Bullard et al., 2007).

The People of Color Summit

The 1991 First National People of Color Environmental Leadership Summit was probably the most important single event in the young movement's history. Held in Washington, DC, the four-day Summit was attended by over 650 grassroots and national leaders from around the world. Delegates came from all 50 states including Alaska and Hawaii, Puerto Rico, Chile, Mexico, and as far away as the Marshall Islands.

The Summit broadened the environmental justice movement beyond its antitoxics focus and demonstrated that it is possible to build a multiracial grassroots movement around environmental and economic justice. On 27 September 1991, Summit delegates adopted 17 'Principles of Environmental Justice.' By June 1992, Spanish and Portuguese translations of the Principles were being used and circulated by NGOs and environmental justice groups at the Earth Summit and Global Forum in Rio de Janeiro.

In October 2002, the United Church of Christ and environmental justice leaders convened the Second National People of Color Environmental Leadership Summit (Summit II) in Washington, DC. Summit II organizers planned the four-day meeting for 500 participants. However, more than 1400 delegates from grassroots and community-based organizations, faith-based groups, organized labor, civil rights, youth, and academic institutions made their way to the nation's capital to participate in the historic gathering, a clear indication that the movement was alive and well. A special budget was carved out for Summit II youth and student programming. More than three-fourths of Summit II attendees were people of color from community-based organizations and networks.

Government Response

The federal Environmental Protection Agency (EPA) in the United States manages, regulates, and distributes risks. During its 43-year history, the EPA has not always recognized that many government and industry practices (whether intended or unintended) have adverse impact on poor people and people of color. In 1992, after meeting with community leaders, academicians, and civil rights leaders, the EPA (under the Bush Administration) admitted there was a problem, and established the Office of Environmental Equity. The name was changed to the Office of Environmental Justice under the Clinton Administration.

The agency produced *Environmental Equity: Reducing Risks for All Communities*, a comprehensive report that examines environmental justice concerns in low-income and people of color communities (US Environmental Protection Agency, 1992). The EPA also established a 25-member National Environmental Justice Advisory Council (NEJAC) under the Federal Advisory Committee Act. The NEJAC is comprised of stakeholders representing grassroots community groups, environmental groups, nongovernmental organizations, state, local, and tribal governments, academia, and industry.

Environmental Justice Health Symposium

In February 1994, seven federal agencies, including the Agency for Toxic Substances and Disease Registry (ATSDR), National Institute for Environmental Health Sciences, EPA, National Institute of Occupational Safety and Health, National Institutes of Health, Department of Energy, and Centers for Disease Control and Prevention (CDC) sponsored a national health symposium, 'Health and Research Needs to Ensure Environmental Justice.' The February conference was attended by more than 1000 scientists, researchers, health professionals, educators, government officials, and community stakeholders. Scholarships were provided to nearly 300 community leaders.

Environmental justice leaders offered their own environmental protection framework. This framework adopts a public health model of prevention (i.e., elimination of the threat before harm occurs) as the preferred strategy, shifts the burden of proof to polluters/dischargers who do harm, who discriminate, or who do not give equal protection to people of

color, low-income persons, and other 'protected' classes, allows disparate impact and statistical weight or an 'effect' test, as opposed to 'intent,' to infer discrimination, and redresses disproportionate impact through 'targeted' action and resources.

Executive Order 12898

In response to growing public concern and mounting scientific evidence, President Clinton on 11 February 1994 issued Executive Order 12898, 'Federal Actions to Address Environmental Justice in Minority Populations and Low-Income Populations.' The Order attempts to address environmental injustice within existing federal laws and regulations. It reinforces the Civil Rights Act of 1964, Title VI, which prohibits discriminatory practices in programs receiving federal funds. The Order also focuses the spotlight back on the 1969 National Environmental Policy Act, a law to ensure for all Americans a safe, healthful, productive, and aesthetically and culturally pleasing environment (Council on Environmental Quality, 1997).

11 February 2014 marked the twentieth anniversary of the historic Environmental Justice Executive Order 12898.

Although the Environmental Justice Executive Order was signed by President Clinton 20 years ago, it has never been fully implemented. After lying dormant for nearly a decade under the George W. Bush administration, the Environmental Justice Executive Order received some new life in 2010 under the Obama administration – with the reinvigoration of the Interagency Working Group that called for updating more than a dozen agencies' Environmental Justice strategic plans and Plan Environmental Justice 2014, a road map designed to help EPA integrate environmental justice into its programs, policies, and activities. The goals of the plan are to protect health in communities overburdened by pollution; empower communities to take action to improve their health and environment; and establish partnerships with local, state, tribal, and federal organizations to achieve healthy and sustainable communities.

Communities under Siege

A growing body of evidence reveals that people of color and low-income persons have borne greater environmental and health risks than the society at large in their neighborhoods, workplace, and playgrounds (Bryant, 1995; Bullard, 1999a,b,c, 2000; Stretesky and Hogan, 1998; Bullard and Wright, 2009; Bullard et al., 2011; Bullard and Wright, 2012; Adeola, 2012; Westra, 2013). An Institute of Medicine (1999) concluded that government, public health officials, and the medical and scientific communities need to place a higher value on the problems and concerns of environmental justice communities. The IOM study confirmed what most impacted communities have known for decades. People of color and low-income communities are (1) exposed to higher levels of pollution than the rest of the nation, and (2) experience certain diseases in greater numbers than the more affluent, white communities (Institute of Medicine, 1999).

The 2002 *Air of Injustice* study found dirty power plants are not randomly distributed across the US landscape. More than 78% of all African-Americans live within 30 miles of a power plant, the distance within which the maximum effects of the smokestack plumes are expected to occur. This is compared to 56% of whites and 39% of Hispanics (Black Leadership Forum et al., 2002).

In 2007, the *Toxic Wastes and Race at Twenty* study found people of color make up the majority (56%) of those living in neighborhoods within two miles of the nation's 413 commercial hazardous waste facilities (Bullard et al., 2007). People of color make up a much larger (over two-thirds) majority (69%) in neighborhoods with clustered facilities, and people of color in 2007 were more concentrated in areas with commercial hazardous sites than in 1987. Additionally, research on polluting facilities concluded that two-thirds of those living within a mile of such facilities are people of color, while two-thirds of those living more than 2½ miles away are white (Pastor et al., 2007).

Downey and Hawkins (2008) conducted a study on race, income, and environmental inequality in the United States in which they concluded African-Americans experience such a high air-pollution burden that black households with incomes of $50 000–$60 000 live in neighborhoods that are, on average, more polluted than neighborhoods of white households with incomes less than $10 000.

In tracking pollution from individual companies to specific communities, the corporate environmental justice performance scorecard shows the most polluted places tend to have significantly higher-than-average percentages of people of color (Ash et al., 2009). In examining the top 10 companies on the 'Toxic 100' list, people of color bear more than half of the human health impacts from the toxic air released by these industries.

Students of color are more likely to attend schools in heavily polluted areas than their white counterparts. While 44.4% of all white students in states attend schools located in the top 10% of the most polluted locations in the state, 81.5% of all African-American children and 62.1% of all Hispanic students attend schools in the most polluted zones (Mohai et al., 2011).

Unequal Protection

The *National Law Journal* uncovered glaring inequities in the way the federal EPA enforces its giant Superfund laws. White communities see faster action, better results, and stiffer penalties than communities where blacks, Hispanics, and other people of color live (Lavelle and Coyle, 1992). Clearly, communities that are located on the 'wrong side of the tracks' are at special risk from exposure to environmental hazards. Race has been found to be independent of class in the distribution of air pollution, contaminated fish consumption, location of municipal landfills and incinerators, abandoned toxic waste dumps, polluting industries, and lead poisoning in children (Commission for Racial Justice, 1987; Agency for Toxic Substances and Disease Registry, 1988; Pirkle et al., 1994).

Far too many people of color communities have become environmental 'sacrifice zones' where polluting industries expose residents who live on the fence line to dangerous emissions and releases into the air, water, and ground. African-Americans are 79% more likely than whites to live in neighborhoods where industrial pollution is suspected of posing the greatest health danger. In Houston, the nation's fourth largest

city, Hispanic and black residents who live near the polluting industrial facilities that line the Houston Ship Channel and near its booming Port pay the ultimate price with their health. As Houston expands its port, residents near the Houston Ship Channel fear the worst. Even before the Houston Port expansion, the mostly Hispanic and black East Houston neighborhoods bordering the port are already exposed to some of the worst air pollution in the country. A 2013 survey, conducted by the Healthy Port Communities Coalition, found that residents of five neighborhoods bordering the Houston Ship Channel suffer from higher rates of cancer and respiratory illnesses in the state.

Many poor and people of color communities are on the frontline of industrial accidents. And because of their close proximity to industrial corridors and transportation routes, they are exposed to elevated health risks from accidents, leaks, spills, explosions, and derailments. In 2004, more than half of the nation's 60 000 pressurized rail tank cars did not meet industry standards, and they raised questions about the safety of the rest of the fleet as well. Every day, deadly 'time bombs' haul their toxic loads through populated areas exposing millions of Americans to potential threats, both accidental and deliberate. Residents who live on the 'wrong side of the tracks' know their neighborhoods are prime locations for accidents waiting to happen.

The Harry Holt family wells and drinking water in Dickson, Tennessee, were poisoned with trichloroethylene (TCE) from a county-owned landfill. A battery of government tests were performed on duck ponds and wells where dogs were waiting to be euthanized. However, a black family's well that lies just 54 feet from the leaky Dickson County Landfill was not tested for 9 years, 1992 through 2001. It appears that government officials cared more about ducks and dogs than protecting the health and safety of law-abiding black homeowners and taxpayers. In 2003, the family sued the City and County of Dickson, the State of Tennessee and the company that dumped the TCE.

After more than 8 years of litigation, on 7 December 2011, a $5.6 million settlement agreement was finally worked out with the Dickson City and County governments on the NRDC and Holts' suit and a $1.74 million settlement to be paid to 11 Holt family members on the family's National Association for the Advancement of Colored People (NAACP) civil rights suit with monies from an October settlement reached with three companies that were defendants in the Holts' NRDC case. Harry Holt died of cancer in January 2007. His daughter, Sheila Holt Orsted is recovering from breast cancer. The county spent more than $3 million and the city almost $1.9 million on the lawsuits fighting the black family.

Even when the government has the facts about contamination and health impacts, there is no guarantee that it will act. For example, a federal survey 1999 and repeated in 2006 by the US ATSDR found dioxins in the Mossville, Louisiana, residents to be two to three times the average. Researchers have also discovered dangerous levels in residents' homes and on yards where children of Mossville play. Mossville is an unincorporated community, founded by African-Americans in the 1800s on the outskirts of Lake Charles, Louisiana, home to four vinyl production facilities. Vinyl chloride was documented to be present in the local air at levels 120 times higher than the ambient air standard. The pollution has become so unbearable that many residents have called for a community buyout and relocation. In July 2013, Sasol Lt., a South African energy giant, offered to voluntary buyout of homes in the Mossville area. The company plans a $14.5 billion complex near the community to turn natural gas into chemicals and fuels.

Childhood Lead Poisoning

Lead poisoning is the number one environmental health threat to children of color in the United States. Figures reported in the July 1994 *Journal of the American Medical Association* on the Third National Health and Nutrition Examination Survey (NHANES III) revealed that 1.7 million children (8.9% of children aged 1–5) are lead poisoned, defined as blood lead levels equal to or above 10 μg deciliter. The NHANES III data found African-American children to be lead poisoned at more than twice the rate of white children at every income level (Pirkle et al., 1994). Over 28.4% of all low-income African-American children were lead poisoned, compared to 9.8% of low-income white children. During 1976–91, the decrease in blood lead levels for African-American and Mexican-American children lagged far behind that of white children.

In California, a coalition of environmental, social justice, and civil libertarian groups joined forces to challenge the way the state carried out its lead screening of poor children. The Natural Resources Defense Council, the NACCP Legal Defense and Education Fund (LDF), the American Civil Liberties Union, and the Legal Aid Society of Alameda County, California won an out-of-court settlement worth $15 million–$20 million for a blood lead-testing program. The lawsuit, *Matthews vs Coye*, involved the failure of the state of California to conduct federally mandated testing for lead of some 557 000 poor children who receive Medicaid. This historic agreement triggered similar lawsuits in several other states that failed to live up to the federal mandates.

The Centers for Disease Control (2013) reported that about 535 000 children aged between 1 and 5 years in the United States are at risk for lead poisoning or about 1 in 38 children. The CDC reduced threshold for lead poisoning from 10 μg to 5 μg and found that between 1999 and 2010 2.6% of those children showed levels above the new threshold. And as was found in all previous reports, poor children, those living in dilapidated older houses, and African-American children are at higher risk for lead poisoning.

The Path of Least Resistance

The southern United States has become a 'sacrifice zone' for the rest of the nation's toxic waste. Toxic waste follows the path of least resistance. And in many instances, this translates into poor and people of color communities. In December 2008, a wall holding back 80 acres of sludge from the Tennessee Valley Authority (TVA), Kingston Fossil Plant broke spilling more than 500 million gallons of toxic coal ash over a dozen homes and up to 400 acres of the surrounding landscape, endangering aquatic life and the water supply for more than 25 000 residents. Six months after this tragedy in July 2009, a major

environmental injustice was perpetrated by EPA Region 4 approval of TVA's decision to ship 5.4 million cubic yards of toxic coal ash by railcar from the mostly white east Tennessee Roane County to a landfill located in the heart of the Alabama Black Belt, Perry County (69% African-American with more than 32% of its residents living in poverty) and to rural Taylor County, Georgia (41% of the population is African-American and more than 24% of residents live in poverty). TVA's toxic waste disposal decision is consistent with the theme of *Dumping in Dixie: Race, Class, and Environmental Quality* (Bullard, 2000), which illustrates that all communities are still not created equal. Shipping toxic waste to poor and people of color communities is also not news since an overwhelming share of waste facilities are located there.

Lax enforcement of environmental regulations has left the region's air, water, and land the most industry-befouled in the United States. Louisiana typifies this pattern. Nearly three-fourths of Louisiana's population, more than 3 million people, get their drinking water from underground aquifers. Dozens of aquifers are threatened by contamination from polluting industries. The Lower Mississippi River Industrial Corridor has over 125 companies that manufacture a range of products including fertilizers, gasoline, paints, and plastics. Environmentalists and local residents have dubbed this corridor as 'Cancer Alley' (Motavalli, 1998; Bullard, 2000).

Norco is located in St. Charles Parish, approximately 40 miles northwest of New Orleans. The town's name of Norco is derived from the New Orleans Refinery Company. The Old Diamond Community is more than 100 years old. As a result of Shell's Industrial expansion, the size of the Old Diamond Community has been shrinking.

The mostly African-American Old Diamond community in Norco is sandwiched between the Motiva and Shell refineries. The community is more than 100 years old and many of the residents can trace their ancestry to former slaves who purchased land from their former plantation owners. Pollution from the refineries has been a constant pain for the nearby residents who have been fighting for a buyout and relocation from the pollution. In 2002, after a decade of struggle, Concerned Citizens of Norco achieved an environmental justice victory. Shell agreed to a community buyout. And for her crusading environmental justice work, community leader Margie Richard, a lifelong Old Diamond community resident, won the Goldman Prize in 2004.

Corporate Welfare

Louisiana is also a leader in doling out corporate welfare to polluters. Industries routinely pollute the air, ground, and drinking water while being subsidized by tax breaks from the state. In the 1990s, the state wiped off the books $3.1 billion in property taxes to polluting companies. The state's top five worse polluters received $111 million dollars over the past decade (Barlett and Steele, 1998).

Subsidizing polluters is not only bad business, but it does not make environmental sense. A growing body of evidence reveals that "states with lower pollution levels and better environmental policies generally have more jobs, better socioeconomic conditions and are more attractive to new businesses" (Templet, 1995). Louisiana and other states could actually improve their general welfare by enacting and enforcing regulations to protect the environment.

In July 2013, the Southeast Louisiana Flood Protection Authority-East filed a lawsuit against dozens of energy giants asking the courts to force them to pay for decades of damage to fragile Louisiana coastal wetlands that help buffer the effects of hurricanes on the region (Schwartz, 2013). The lawsuit argues that energy companies such as BP and Exxon Mobil should be held responsible for fixing environmental damage they have caused to Louisiana wetlands – killing vegetation, eroding soil, and allowing salt water into freshwater areas. Louisiana Governor Jindal opposed the lawsuit.

Dumping on Native Americans

It should not be a surprise to anyone to discover that Native Americans have to contend with some of the worst pollution in the United States (Tomsho, 1990; Taliman, 1992). Native American nations are prime targets for waste trading. More than three dozen Indian reservations were targeted for landfills, incinerators, and other waste facilities in the early 1990s. Native American groups and their allies defeated most of these waste proposals.

Racism operates in energy production (mining of uranium) and disposal of wastes on Indian lands (Churchill and LaDuke, 1983). The practice has left many sovereign Indian nations without an economic infrastructure to address poverty, unemployment, inadequate education, and health care, and a host of other social problems. In 1999, residents on the Eastern Navajo Reservation in Church Rock and Crown Point, New Mexico formed Eastern Navajo Dine Against Uranium Mining (ENDAUM) and the New Mexico Environmental Law Center filed suit with the Nuclear Regulatory Commission to block a permit for uranium mining in Church Rock and Crown Point, New Mexico. ENDAUM has kept the fight going for over 16 years and Hydro Resources has yet to break ground. In December 2013, US Bankruptcy Judge Allan L. Gropper ruled that Anadarko and Kerr-McGee companies are liable for the contamination mess they have created on the Navajo Nation and set damages ranging between $5.1 and $14.1 billion (ICTMN Staff, 2013). And because of the abandoned uranium mines, tailings and contaminated waste sites that dot the reservation, many families have to travel 15–20 miles for water to drink, cook, bath in, and water their animals because their wells have been contaminated by uranium mining.

Transboundary Waste Trade

Hazardous waste generation and international movement of hazardous waste pose some important health, environmental, legal, and ethical dilemmas. Shipping hazardous wastes from rich communities to poor communities is a global problem. In the real world, all people, communities, and nations are not created equal. Unequal interests and power arrangements have allowed poisons of the rich to be offered as short-term remedies for poverty of the poor. This scenario plays out domestically (as in the United States where low income and people of color communities are disproportionately impacted by waste facilities and 'dirty' industries) and internationally – where

hazardous wastes move from OECD states to non-OECD states (Khoo and Rau, 2009).

The conditions surrounding the more than 2000 maquiladoras, assembly plants operated by American, Japanese, and other foreign countries, located along the 2000-mile US–Mexico border may further exacerbate the waste trade. The industrial plants use cheap Mexican labor to assemble imported components and raw material and then ship finished products back to the United States. Nearly a half million Mexican workers are employed in the maquiladoras.

All along the Lower Rio Grande River Valley maquiladoras dump their toxic wastes into the river, from which 95% of the region's residents get their drinking water (Hernandez, 1993). In the border cities of Brownsville, Texas and Matamoras, Mexico, the rate of anencephaly – babies born without brains – is four times the national average. Affected families have filed lawsuits against 88 of the area's 100 maquiladoras for exposing the community to xylene, a cleaning solvent that can cause brain hemorrhages, and lung and kidney damage.

The Mexican environmental regulatory agency is understaffed and not equipped adequately to address the problem (Barry and Simms, 1994), a situation that has not changed since.

As the United States tightens regulations on lead emissions, battery recycling operations have moved to Mexico. After 2008, new US limits on lead pollution have made domestic recycling costly. Lead emission standards in Mexico are one-tenth as stringent as US standards. The United States does not treat international battery shipments as hazardous waste as is common among developed nations. Between 2004 and 2011, the US exports of dead batteries to Mexico increased by anywhere from 449% to 525% (Secretariat of the Commission for Environmental Cooperation, 2012). In 2011, the dead batteries headed to Mexico would have filled 17 952 tractor-trailers. Mexico is rapidly becoming a 'pollution haven.'

Conclusion

The environmental justice movement emerged in response to environmental and social inequities, threats to public health, unequal protection, differential enforcement, and disparate treatment received by the poor and people of color. The movement redefined environmental protection as a basic right. It also emphasized pollution prevention, waste minimization, and cleaner production techniques as strategies to achieve environmental justice for all Americans without regard to race, color, national origin, or income.

The poisoning of African-Americans in Louisiana's 'Cancer Alley,' Native Americans on reservations, and Mexicans in the border towns all have their roots in a system characterized by economic exploitation, racial oppression, and devaluation of human life and the natural environment. Unequal political power arrangements have allowed poisons of the rich to be offered as short-term economic remedies for poverty and unemployment. Having industrial facilities in one's community does not automatically translate into jobs for nearby residents. Similarly, tax breaks and corporate welfare programs have produced few new jobs by polluting firms.

The environmental justice movement has established clear goals of eliminating unequal protection. This call for environmental and economic justice does not stop at the US borders but extends to communities and nations that are threatened by the export of hazardous wastes, toxic products, and 'dirty' industries. Environmental justice leaders are demanding that no community or nation, rich or poor, urban or suburban, black or white, should be allowed to become a 'sacrifice zone' or dumping grounds. They are also pressing governments to live up to their mandate of protecting public health and the environment.

See also: Environment and Development; Environment and Health; Environmental Health and Safety: Social Aspects; Environmental Movements; Environmental Risk; Neighborhood Revitalization and Community Development; Sustainable Development.

Bibliography

Adeola, F.O., 2012. Industrial Disasters, Toxic Waste and Community Impact: Health Effects and Environmental Justice Struggles Around the Globe. Lexington Books, Lanham, MD.

Agency for Toxic Substances and Disease Registry, 1988. The Nature and Extent of Lead Poisoning in Children in the United States: A Report to Congress. US Department of Health and Human Services, Atlanta, GA.

Ash, M., Boyce, J.K., Chang, G., Pastor, M., Scoggins, J., Tran, J.R., April 2009. Justice in the Air: Tracking Toxic Pollution from America's Industries and Companies to Our States, Cities, and Neighborhoods.

Barlett, D.L., Steele, J.B., November 23, 1998. Paying a price for polluters. Time, 72–80.

Barry, T., Simms, B., 1994. The Challenge of Cross Border Environmentalism: The US–Mexico Case, first ed. Resource Center Press, Albuquerque, NM.

Black Leadership Forum Clean the Air Georgia Coalition for the People's Agenda, and The Southern Organizing Committee for Economic and Social Justice, October 2002. Air of Injustice: African Americans & Power Plant Pollution.

Bryant, B., 1995. Environmental Justice: Issues, Policies, and Solutions. Island Press, Washington, DC.

Bullard, R.D. (Ed.), 1996. Unequal Protection: Environmental Justice and Communities of Color. Sierra Club Books, San Francisco.

Bullard, R.D., 1999a. Dismantling environmentalism in the USA. Local Environment 4, 5–19.

Bullard, R.D., 1999b. Building just, safe, and healthy communities. Tulane Environmental Journal 12, 373–404.

Bullard, R.D., Spring 1999c. Leveling the playing field through environmental justice. Vermont Law Review 23, 454–478.

Bullard, R.D., 2000. Dumping in Dixie: Race, Class and Environmental Quality, third ed. Westview Press, Boulder, CO.

Bullard, R.D., Johnson, G.S., 2000. Environmental justice: grassroots activism and its impact on public policy decision making. Journal of Social Issues 56, 555–578.

Bullard, R.D., Johnson, G.S., Torres, A.O., 2011. Environmental Health and Racial Equity in the United States: Building Environmentally Just, Sustainable, and Livable Communities. American Public Health Association, Washington, DC.

Bullard, R.D., Mohai, P., Saha, R., Wright, B., 2007. Toxic Wastes and Race at Twenty 1987–2007. A Report Prepared for the United Church of Christ Justice and Witness Ministries. United Church of Christ, Cleveland, OH.

Bullard, R.D., Wright, B., 2009. Race, Place and Environmental Justice after Hurricane Katrina: Struggles to Reclaim, Rebuild, and Revitalize New Orleans and the Gulf Coast. Westeview Press, Boulder, CO.

Bullard, R.D., Wright, B., 2012. The Wrong Complexion for Protection: How the Government Response to Disaster Endangers African American Communities. New York University Press, New York.

Centers for Disease Control and Prevention, April 5, 2013. Blood Lead Levels in Children Aged 1–5 Years, United States, 1999–2010. Morbidity and Mortality Weekly Report (MMWR).

Churchill, W., LaDuke, W., 1983. Native American: the political economy of radioactive colonialism. Insurgent Sociologist 13 (1), 51–63.

Commission for Racial Justice, 1987. Toxic Wastes and Race in the United States. Public Data Access, New York.

Council on Environmental Quality, 1997. Environmental Justice: Guidance under the National Environmental Policy Act. CEQ, Washington, DC.
Downey, L., Hawkins, B., 2008. Race, income and environmental inequality in the United States. Sociological Perspectives 51 (4), 759–781.
Hernandez, B.J., August 1993. Dirty Growth. The New Internationalist.
Institute of Medicine, 1999. Toward Environmental Justice: Research, Education, and Health Policy Needs. National Academy Press, Washington, DC.
Khoo, SU-Ming, Rau, Henricke, 2009. Movements, mobilities and the politics of hazardous waste. Environmental Politics 18 (6), 960–980.
Lavelle, M., Coyle, M., September 21, 1992. Unequal protection. The National Law Journal, 1–2.
Mohai, P., Kweon, B., Sangyun, L., Ard, K., 2011. Air pollution around schools is linked to poorer student health and academic performance. Health Affairs 30 (5), 852–862.
Motavalli, J., July/August 1998. Toxic targets: polluters that dump on communities of color are finally being brought to justice. E Magazine 9, 28–41.
Pastor, M., Sadd, J., Morello-Frosch, R., 2007. Still Toxic after All These Years: Air Quality and Environmental Justice in the San Francisco Bay Area.
Pirkle, J.L., Brody, D.J., Gunter, E.W., Kramer, R.A., Paschal, D.C., Glegal, K.M., Matte, T.D., 1994. The decline in blood lead levels in the United States: the national health and nutrition examination survey (NHANES III). Journal of the American Medical Association 272, 284–291.
ICTMN Staff, December 18, 2013. Navajo Nation Could Get $1 Billion in Damages for Uranium Mess. Indian Country Today.
Schwartz, John, July 24, 2013. Louisiana agency sues dozens of energy companies for damage to wetlands. The New York Times.
Secretariat of the Commission for Environmental Cooperation, 2012. Hazardous Trade? an Examination of US-Generated Lead-Acid Battery Exports and Secondary Lead Recycling in Mexico, the United States and Canada. Final Draft Report. Montreal, Canada.
Stretesky, P., Hogan, M.J., May 1998. Environmental justice: an analysis of superfund sites in Florida. Social Problems 45, 268–287.
Taliman, V., Fall 1992. Stuck holding the nation's nuclear waste. Race, Poverty & Environment Newsletter, 6–9.
Templet, P.H., Spring 1995. The Positive Relationship between Jobs, Environment and the Economy: An Empirical Analysis and Review. Spectrum, pp. 37–49.
Tomsho, R., November 29, 1990. Dumping grounds: Indian tribes contend with some of the worst of America's pollution. The Wall Street Journal, pp. A1, A6.
US General Accounting Office, 1983. Siting of Hazardous Waste Landfills and Their Correlation with Racial and Economic Status of Surrounding Communities. Government Printing Office, Washington, DC.
US Environmental Protection Agency, 1992. Environmental Equity: Reducing Risk for All Communities. EPA, Washington, DC.
US Environmental Protection Agency, 1998. Guidance for Incorporating Environmental Justice in EPA's NEPA Compliance Analysis. EPA, Washington, DC.
Westra, L., 2013. Environmental Justice and the Rights of Indigenous Peoples: International and Domestic Legal Perspectives. Earthscan Publisher.

Environmental Movements

Robert J Brulle, Drexel University, Philadelphia, PA, USA
Christopher Rootes, School of Social Policy, Sociology and Social Research, University of Kent, Canterbury, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Originating in nineteenth century concerns with the protection of unspoiled landscapes and natural resource conservation in the United Kingdom and the United States, environmentalism is now a global phenomenon, embracing a number of different perspectives. The development of different and multiple environmental movements worldwide is the outcome of contingent historical processes and events, the development of specific discourses, and the mobilization of material and cultural resources. There is no one 'environmental' movement. Rather, there are a number of different environmental movements that focus on a wide range of issues.

Introduction

Over the past century, concern over environmental degradation has manifested itself in various organized efforts across the globe. Environmental movements are networks of actors who advocate for a range of environmental concerns. There is not just one environmental movement (Doyle and MacGregor, 2014: 1). Rather, there are a number of partially overlapping networks that interact at local, regional, national, and international levels. These movements involve groups and organizations that range in size from small informal local groups that have no organizational affiliation, to large formalized international organizations, and even political parties, especially Green parties. The aims and orientations of these different groups vary widely, with diverse groups advocating the adoption of technological measures, various types of international accords, economic measures, or the wholesale transformation of capitalism. They also vary regarding the level of government policy they aim to influence, from local authorities and municipalities to national and international institutions, as well as specific corporate actors (Dietz and Garrelts, 2014).

Environmental movements are engaged in collective action motivated by their participants' shared identity or concern about environmental issues but they are not, however, identical with organizations or episodes of protest. It is only when organizations (and other, usually less formally organized actors) are networked and engaged in collective action, whether or not it involves protest, that an environmental movement exists (see Diani, 1995: 5; Rootes, 2004). Such linkages are not always readily visible. Where environmental movements are well established, they may focus on less visible forms of action such as lobbying and constructive engagement with governments and corporations rather than highly visible protest. Where environmentalism has become entrenched, there are many less publicly visible linkages among groups and organizations, and the full range of movement activities is less likely to be adequately reported in mass media. This has implications for scholars researching movements, with no single research strategy likely to yield a complete picture. Organizational studies neglect the less organized parts of the movement, and whilst studies of protest events (Rootes, 2003) may represent the most visible face of the movement, they cannot adequately portray less public – or simply less publicized – movement activities, including those that happen locally and out of sight of national mass media.

The Rise of Environmental Movements

The development of the different and multiple environmental movements across the globe is the outcome of contingent historical processes and events, the development of specific discourses, and the mobilization of material and cultural resources that results in a definable network of interaction. Over the past century, environmental movements have grown in range from a few isolated nature preservation groups in the United States, the United Kingdom, Germany, and other European countries into a worldwide phenomenon that encompasses an enormous range of issues. As industrialization and the exploitation of natural resources accelerated, their impacts on the environment became increasingly obvious and this motivated individuals to raise alarms about the consequences. In the 1960s, in a context of increasing aspirations of citizens to public participation, writers such as Rachel Carson cogently framed the issues of human actions that caused environmental degradation and thereby inspired individuals to join together in collective actions (Carson, 1962). These new cultural resources formed an important component in the development of environmentalism (Carmichael et al., 2012; Brulle and Benford, 2012). Motivated by these new perspectives and aided in the United States by foundation support, environmental movements rapidly expanded, starting in the 1960s in the United States and western Europe, and then spreading across the globe in subsequent decades. This, joined with the radical critique of capitalist industrialism and representative democracy associated with the New Left and various countercultural trends, created the public space for the development of new social movements, as well as furnishing their tactical repertoire. More radical environmental movement organizations (EMOs) that embraced nonviolent direct action arose and spread from the United States to western Europe. Friends of the Earth (FoE) and Greenpeace (Zelko, 2013) were started by environmentalists dismayed by the apparent timidity of more traditional environmental organizations such as the Sierra Club,

and in response to the demands of a new, self-confident, and participatory generation to participate directly in activism. The internationalism of the new EMOs and their skillful exploitation of mass media to excite public concern in order to put pressure on governments and corporate decision-makers were in tune with the times, and these were the fastest growing EMOs during the 1980s.

On both sides of the Atlantic, the history of environmentalism has been one of successive waves of critique, innovation, and incorporation. The rise of the new EMOs encouraged innovation in the agendas and tactics of older conservation organizations, and by the end of the 1980s networking among older and newer organizations and joint campaigns were increasingly common. By then, however, dissatisfaction with institutionalized reform environmentalism was stimulating the rise of groupings of radical ecologists and the environmental justice movement (EJM). Beginning in the 1970s in the United States and in parts of the global South, movements developed that were focused on environmental justice and the unequal environmental burden suffered by poor and disenfranchised peoples.

Many organizations formed in earlier waves of environmental concern have adapted to accommodate subsequent concerns and developments in ecological consciousness. In the United States, despite observers' doubts that established, 'wilderness obsessed' EMOs are capable of accommodating the concerns of the EJM, especially after 2000, the Sierra Club and Greenpeace took up aspects of the environmental justice agenda. In western Europe, the environmental justice frame has been increasingly adopted, especially by FoE, while established EMOs such as the World Wide Fund for Nature (WWF) and the national bird protection societies have developed more inclusive ecological perspectives and broadened the range of their campaigns to include issues of habitat and the welfare of human populations (Rootes, 2006). Despite noteworthy differences in perspectives, areas of concern, priorities, and tactics from one group or organization to another, there are network links of varying kinds and degrees of strength and intensity. It is the persistence of these that demonstrates the vitality of environmental movements (Rootes, 2004).

Values and Social Bases of Environmentalism

The rise of values and attitudes favorable to environmentalism is frequently explained in terms of 'post-materialism' (Inglehart, 1977). According to this thesis, post-material values prioritizing aesthetic, intellectual, and self-actualization needs are gradually displacing materialist values that place higher priority on economic and security needs, chiefly because younger generations raised in relative affluence and security are steadily replacing older generations who, during their formative years, more often experienced economic privation and the insecurities of war. The rise of environmentalism was counted among the consequences of this 'culture shift.'

Yet Dalton and Rohrschneider (2002) examined the link between individual endorsement of post-materialist values and membership in environmental organizations and found that the correlation was modest in wealthy nations and barely existed in poor ones. Post-materialism may not be as good a predictor of environmentalism as some observers assume because environmentalism combines a post-materialist aesthetic and principled concerns with environmental protection with essentially materialist concerns with safety and security, whether about air or water pollution, or the impacts of nuclear energy and climate change. Even global environmental concern, often considered to be a post-materialist gesture of concern by citizens of the rich North for the plight of the poor of the global South, might be portrayed as materialist since the effects of environmental degradation in the South will sooner or later be felt even in the North.

Post-materialism is also relatively weakly correlated with support for environmentalism because concerns for the environment are held both by different social groups: on the one hand, highly educated 'post-materialist' ecologists, who are more concerned about global environmental problems with relatively remote effects than they are fearful for their own security; and, on the other hand, by people, usually less well educated, who are more worried about the threats that environmental damage poses to their own material security. Post-materialism is a better predictor of environmental activism than of environmental concern because environmental activism, like most forms of political activism and unlike concern, is highly correlated with higher education. Similarly, the significant correlation that Dalton (2005) found between national affluence and membership in environmental organizations might exist simply because affluence is associated with membership in civic groups more generally.

Most research on the social backgrounds of environmental activists and members of national EMOs has concluded that they are relatively highly educated, more likely to be employed in teaching, creative, or caring professions, and very often the offspring of highly educated parents (Rootes, 1995). For this reason, environmentalism has sometimes been interpreted as the self-interested politics of a 'new class' comprised of people who are occupationally engaged with culture and the manipulation of symbols, and who are opposed or indifferent to the interests of those whose labor involves the manipulation of material things. However, environmental activists are not exclusively drawn from such backgrounds, and approval of EMOs and environmental activism, as well as pro-environment attitudes, is more widespread in most societies than the 'new class' thesis would suggest.

Grassroots environmental movements involve a broader cross section of society than major national EMOs. This is partly because less socially and economically privileged communities are more likely than the rich and the resourceful to be subjected to unwanted and environmentally damaging developments, but also because it is also easier to enter local political arenas than to participate effectively in national politics. Women play more prominent roles in grassroots mobilizations than in national EMOs, both because women are in general more involved in local communities, and also because they are more likely to be confident of acting in the local community than in the wider public sphere. Grassroots environmental activism is thus an important element of the mobilization of environmental movements: it makes the movement more broadly socially representative than it might otherwise be, and may also provide opportunities for social learning about environmental issues, acting as a school for

participation generally, an entry point for new activists and new issues, and even a source for the revitalization of national environmental organizations.

Culture and Forms of Action

Environmental movements across the globe involve a wide-ranging and complex network of interaction among a number of different organizations. They are composed of people and groups with many different viewpoints, which define their perception of the natural environment and their subsequent issue focus. Perspectives and priorities differ between and among environmentalists in the global North and the global South, as is apparent in the regional differences among the 74 national affiliates of FoE International (Doherty and Doyle, 2013). To understand such complex sets of interactions, sociologists have developed discursive frame analysis to capture and interpret different worldviews or 'discursive frames' within a social movement (Benford and Snow, 2000; Benford, 2011). A discursive frame is the set of cultural viewpoints that informs the practices of a community of social movement organizations, defines the goals and purposes of the organization, and provides guidance for action. In turn, this frame influences the organization's tactics, and methods of resource mobilization (Benford and Hunt, 1992; Brulle, 2000). It also defines a network of interaction among other similar organizations. The character of this network defines the scope and strength of the movement organization, relationships with other movement organizations and other institutions.

Within any social movement, there are generally multiple frames, which define distinct movement sectors or 'wings' that diverge in their definitions of problems, strategies, and methods of organization (Snow, 2004). This is certainly the case among the environmental movements. Theorists have identified a number of environmental networks based on their particular discursive frames, but these are highly contextual definitions that are applicable only within certain cultural and political boundaries. Organizational formalizations of ideological division are more likely in the formal party-political arena, where membership tends to be exclusive, than in movements, where organizations are more fluid, overlapping memberships are common, and the flexibility of the network structure may accommodate differences without much overt conflict. Moreover, common ground among organizations may mask greater diversity among the perspectives of rank-and-file activists (Wahlström et al., 2013).

Although the discursive frames adopted by environmentalists have consequences for the ways in which they campaign and the forms of organization they adopt (Brulle, 2000) but, as Dalton (1994) discovered, whether European EMOs had originally been committed to conservationism or ecologism made surprisingly little difference to their choices of strategies, tactics, and styles of action. The apparent convergence within the broad environmental movement sector was not simply a matter of the progressive institutionalization and incorporation of more radical organizations. FoE and Greenpeace may have started as activist groups and learned how to behave in ways that opened channels of communication to governments and/or powerful corporate actors, but some traditional conservationist organizations have become less conflict averse and more skilled in the use of mass media. EMOs' values do influence their strategy and tactics, but even more influential is the pattern of opportunities and constraints that derives from the structures of the national political systems within which those organizations operate (Dryzek et al., 2003).

Even where they embrace direct action, environmental movements almost always restrict themselves to nonviolent forms of action, but movement subcultures are not isolated from wider social and political contexts. Thus, violent environmental protests in Italy and Spain in the 1980s were part of the spillover from more general political conflicts – the political violence of the 1970s in Italy and militant Basque nationalism in Spain. Similarly, the wave of confrontational environmental protest in Britain in the 1990s was shaped by widespread and occasionally riotous protest against the Thatcher government's poll tax (Rootes, 2003). In general, however, national political cultures explain less of the variation in environmental activists' tactics than do political conjunctures.

The great majority of environmental organizations are either engaged in practical conservation work or focus upon informational and educational strategies or lobbying. Even for Greenpeace, the archetypical professionalized protest organization, protest is only a small part of its work. The small minority of groups that rely on protest are distinguished by the discourses they articulate. Those that employ the discursive frames of deep ecology and environmental justice are disproportionately responsible for the relatively small number of highly publicized environmental protests, especially in the United States. (Brulle, 2000: 246; Carmin and Balser, 2002).

Institutionalization

The numbers of members and supporters of preservation and conservation organizations and EMOs have grown since the 1970s. As a result, many of those organizations have become substantial operations with annual budgets in the millions, and employees numbering hundreds or thousands. As a result, they have increasingly become formal organizations, internally bureaucratized, and employing salaried professional employees rather than the enthusiasm of volunteers. Although some democratic organizations exist in the environmental movement, most lack formal participatory mechanisms and are effectively oligarchies governed by self-selecting boards (Brulle, 2000). These 'professional social movement' organizations (Zald and McCarthy, 1987) have been labeled 'astro turf' organizations (Cigler and Loomis, 1995) or 'protest businesses' (Jordan and Maloney, 1997), in which professional staff dominate and 'members' are merely financial supporters. Some, like Greenpeace, are unabashed about this; others preserve a semblance of democracy in principle but are in practice no less oligarchical (Rootes, 2009).

It used to be assumed that 'success' for a social movement meant its institutionalization, usually as a political party. Yet although Green parties emerged from the 1970s onward, and began by the 1980s to win seats in some European legislatures, they were more often products of antinuclear campaigns than direct representatives of environmental movements, and they are

not institutionalized vehicles of the environmental movement in the way that socialist parties were of the labor movement. Green parties have prospered, especially in northern Europe and Australasia, but their relations with environmental movements have been problematic, especially where Green parties have entered government, as they have in countries as diverse as Germany, Italy, France, Finland, the Czech Republic, and Ireland.

Where conditions have favored the development of institutionalized EMOs, they have become formal organizations embedded in networks of policy advocacy and implementation, and so have constituted a 'public interest lobby' (Diani and Donati, 1999) or an 'advocacy community' (Bosso, 2005). To protect their reputations in the eyes of more powerful actors in government and corporations, environmental organizations that have secured access on the basis of their expertise and politically nonpartisan status have generally been careful to keep some distance between themselves and others, such as Green parties, who have attempted to enter the political arena directly, and especially where Green parties have been only marginal players in formal politics.

In the United States, where political structures have discouraged the development of Green parties, charitable foundations have been a substantial source of funding for most major EMOs. However, most foundation funding goes to moderate organizations that use the traditional discourses of conservation, preservation, and liberal environmentalism. Concentrated on a relatively small number of large EMOs involved in political advocacy work (Brulle and Jenkins, 2005), charitable funding has steered the major environmental organizations into neoliberal environmental protection strategies that have become disconnected from the everyday concerns of their constituencies (Brulle, 2014). In Britain, foundation funding is less significant, but strongly favors organizations engaged in practical environmental protection and remediation rather than advocacy (Rootes, 2009).

Some claim that the institutionalization of EMOs, and their dependence upon foundation funding, has robbed them of the capacity and, even the will, to mobilize large numbers of people in collective action, thus depriving them of the 'veto power' – the ability to disrupt 'business as usual' – that social movement scholars have usually considered to be the ultimate basis of the power of social movements (van der Heijden, 1997). Interestingly, however, those organizations that claim to speak for grassroots environmental activists appear to be scarcely less oligarchic than others: most organizations associated with the EJM in the United States lack mechanisms of democratic accountability (Brulle and Essoka, 2005).

Elsewhere, the institutionalization of EMOs has not simply entailed the deradicalization of movements or a loss of shared identity. In Germany, a substantially institutionalized movement coexisted in the 1990s with the revival of highly confrontational, at times violent, antinuclear protest, as established EMOs gave assistance to local antinuclear groups (Rucht and Roose, 2003). In Britain during the 1990s, EMOs grew in numbers and size as well as access and influence, but reported environmental protest increased and became more, rather than less, confrontational (Rootes, 2003). In the United States, the rise of the climate movement, centered on the organization 350.org led by Bill McKibben, injected new activism and purpose into the environmental movement.

The rise of new, more radical groupings such as Earth First! can be traced to dissatisfaction with the apparent timidity of more established EMOs. Nevertheless, despite variations in their tactical repertoires, there is a measure of shared identity among environmental campaigners. Even the radical 'disorganizations' most committed to direct action have often been connected by networks of advice and support to more established organizations. The support that established EMOs have given to direct action groups may not have been very publicly visible, but the sense of collective identity among the constituent parts of the environmental movement grew as groups came to practice a division of labor and realized the advantages of cooperation, most notably around climate change. This stimulated the rise of new forms of direct action – such as the Climate Camps in Britain – and drew members and supporters of even conservation organizations into street demonstrations (Rootes, 2012).

These experiences show that it is possible for an environmental movement to retain many of the characteristics of an emergent movement whilst taking advantage of the opportunities presented by institutionalization. Institutionalization has not been a barrier to the mobilization of protest. Whilst some writers have referred to the 'self-limiting radicalism' of Green parties, it is the 'self-limiting institutionalization' of environmental movements that is more striking and that has more profound implications.

The survival of environmental movements and their resistance to the deradicalizing effects of institutionalization has distinguished them from the other 'new social movements,' perhaps because pressing environmental problems are a chronic condition of an industrialized world. Northern environmental movements are not universally anticapitalist but they have often, as in as campaigns against genetically modified organisms and neoliberal globalization, been influenced by the critical analyses of radically anticapitalist groups.

In the global South, because environmental campaigns are usually bound up with campaigns for human, democratic, and economic rights, they often do not take the form of specifically environmental movements, but they do campaign for global environmental justice and struggle against extractive industries, transboundary pollution and waste flows, free trade agreements, and the ideological and social systems that reinforce such practices, including militarization, racism, capitalism, patriarchy, and other social hierarchies. In other words, the global EJM merges concerns with ecological sustainability with a demand for social justice and human rights.

These movements view the global political economy as shifting risks and hazards from North to South, from rich nations to poor communities between and within nations, and from racially privileged elites to racially despised communities. Often radical in their approach, they challenge the sources of power in the global political economy – governments, transnational corporations, international financial institutions, and racist and classist ideologies – and seek to destabilize, disrupt, and open them up in the hope of creating a more just transnational political space.

In postcommunist central and eastern Europe and in east Asia, environmental movements often provided cover for prodemocracy activists (Lee and So, 1999). In sub-Saharan Africa, where states are weaker and affluent civil societies are scarcely

developed, environmental movements have been less vigorous (Mittelman, 1998) and, as in recently democratized central and eastern Europe, environmentalists' successes have often depended upon support from Northern/Western EMOs and human rights organizations. Dependence on foreign assistance may, however, displace local agendas and divert activists from efforts to mobilize local people. Perhaps for this reason, in challenging authoritarian states, environmental movements have had greatest impact where they mobilized local people and incorporated local cultural values and rituals, empowering local activists (Lee and So, 1999: 291).

Impacts

The direct impacts of social movements are always hard to distinguish from the effects of social and political changes external to the movement. The rise of environmental movements from the late 1960s coincided with the development of legislation and state agencies designed to protect the environment but, rather than the one leading to the other, both were products of developing environmental awareness amongst a larger public.

Environmental activists played major roles in preserving Californian redwood and Tasmanian forests and the Australian Great Barrier Reef, and Greenpeace's campaign against the dumping of waste at sea changed the policies of oil companies and perceptions of the sea as a limitless waste repository. Grassroots mobilization over issues of toxic pollution has forced states to take actions to remediate existing sites, and to institute preventive measures to forestall future incidents. Environmental movements have stimulated and mediated the concerns of the wider public, and reflected and amplified concerns that originated in scientific and policy communities. Sometimes they have succeeded in sensitizing mass publics and decision-makers to environmental issues that might not otherwise have been so salient.

However, one glaring failure of environmentalism has been their inability to mobilize meaningful action on climate change. After the failure of COP15 in 2009, much of the movement refocused on national and regional efforts. While this has produced some marginal results, it has yet to have any substantial impact upon the practices that contribute to the increasing urgency of the climate change issue, especially as, in deference to powerful economic interests, countries such as Canada and Australia retreat from previous commitments and developing countries continue to increase their greenhouse gas emissions. EMOs may have set the agenda of environmental reform and succeeded in framing the issues as matters of global collective responsibility but, as environmental issues have moved up political agendas, EMOs have struggled to meet expectations. Unable to participate effectively in high politics, they may be drawn back to the grassroots, but it is not clear whether this will produce a wider mobilization and more effective action on climate change.

Despite frequent references to 'the' environmental movement, and even to 'the global environmental movement,' there is so much variation among and within local and national environmental movements that a global environmental movement remains an aspiration rather than an accomplished fact. Obstacles to a global movement include the wide diversity of material conditions, cultures, languages, and political circumstances. National environmental movements have flourished under conditions of stable democracy in strong states, and the absence of a developed democratic global polity is a major obstacle to the development of a *global* environmental movement. However, if the globalization of industrial capitalism increases the scale of the assault on what remains of the pristine natural environment, the gradual extension of democratic politics encourages some hope that people may yet mobilize to limit environmental degradation.

Environmental movements are not monolithic structures but are composed of a variety of discursive communities, each of which has its own specific issue focus. Despite their diversity and temporary setbacks, environmental movements are likely to continue to be major players in politics across the globe.

See also: Environmental Archaeology; Environmental History; Environmental Justice in the United States; Environmental Sociology.

Bibliography

Benford, R.D., 2011. Framing global governance from below: discursive opportunities and challenges in the transnational arena. In: Corneliu, B., Markus, K. (Eds.), Arguing Global Governance. Routledge, London, pp. 67–84.

Benford, R.D., Scott, A.H, 1992. Dramaturgy and social movements: the social construction and communication of power. Sociological Inquiry 62, 36–55.

Benford, R.D., Snow, D.A., 2000. Framing processes and social movements: an overview and assessment. Annual Review of Sociology 26, 611–639.

Bosso, C.J., 2005. Environment Inc.: From Grassroots to Beltway. University Press of Kansas, Lawrence, KS.

Brulle, R.J., 2000. Agency, Democracy, and Nature: The U.S. Environmental Movement from a Critical Theory Perspective. MIT Press, Cambridge MA.

Brulle, R.J., 2014. Environmentalisms in the United States. In: Timothy, D., Sherilyn, M. (Eds.), Global Perspectives on Environmentalism. Praeger, New York, pp. 163–191.

Brulle, R.J., Essoka, J., 2005. Whose environmental justice? An analysis of the governance structures of environmental justice organizations in the US. In: Pellow, D.N., Brulle, R.J. (Eds.), Power, Justice and the Environment: A Critical Appraisal of the Environmental Justice Movement. MIT Press, Cambridge MA, pp. 205–218.

Brulle, R.J., Jenkins, J.C., 2005. Foundations and the environmental movement: priorities, strategies and impact. In: Faber, D., McCarthy, D. (Eds.), Foundations for Social Change: Critical Perspectives on Philanthropy and Popular Movements. Temple University Press, Philadelphia, PA, p. 74.

Brulle, R.J., Benford, R., 2012. "From Game Protection to Wildlife Management: Frame Shifts, Organizational Development, and Field Practices". Rural Sociology 77 (1), 62–88.

Carmichael, J., Jenkins, J.C, Robert, J.B, 2012. Building environmentalism: the founding of environmental movement organizations in the United States, 1900–2000. Sociological Quarterly, 53(2012), 422–453.

Carmin, J., Balser, D., 2002. Selecting repertoires of action in environmental movement organizations: an interpretive approach. Organization and Environment 15 (4).

Carson, R., 1962. Silent Spring. Houghton Mifflin, New York.

Cigler, A.J., Loomis, B.A., 1995. Interest Group Politics, fourth ed. Congressional Quarterly Press, Washington DC.

Dalton, R.J., 2005. The greening of the Globe? Cross-national levels of environmental group membership. Environmental Politics 14 (4), 441–459.

Dalton, R.J., Rohrschneider, R., 2002. Political action and the political context: a multilevel model of environmental activism. In: Fuchs, D., Roller, E., Wessels, B. (Eds.), Citizens and Democracy in East and West: Studies in Political Culture and Political Process. Westdeutscher Verlag, Opladen, pp. 333–350.

Dalton, R.J., 1994. The Green Rainbow: Environmental Groups in Western Europe. Yale University Press, New Haven CT and London.

Diani, M., Donati, P., 1999. Organizational change in Western European Environmental Groups: a framework for analysis. Environmental Politics 8 (1), 13–34.

Diani, M., 1995. Green Networks: A Structural Analysis of the Italian Environmental Movement. Edinburgh University Press.

Dietz, M., Garrelts, H., 2014. Routledge Handbook of the Climate Change Movement. Routledge, London.
Doherty, B.J., Doyle, T., 2013. Environmentalism, Resistance and Solidarity: The Politics of Friends of the Earth International. Palgrave Macmillan, Basingstoke.
Doyle, T., MacGregor, S. (Eds.), 2014. Environmental Movements around the World. ABC Clio/Praeger, Santa Barbara CA, Denver CO and Oxford.
Dryzek, J.A., et al., 2003. Green States and Social Movements. Oxford University Press.
Fox, S., 1981. The American Conservation Movement: John Muir and His Legacy. University of Wisconsin Press, Madison WI.
van der Heijden, H., 1997. Political opportunity structure and the institutionalisation of the environmental movement. Environmental Politics 6 (4), 25–50.
Hays, S.P., 1972. Conservation and the Gospel of Efficiency: The Progressive Conservation Movement, 1890–1920. Athenaeum, New York.
Hays, S.P., 1987. Beauty, Health, and Permanence: Environmental Politics in the United States, 1955–1988. Cambridge University Press, New York.
Inglehart, R., 1977. The Silent Revolution: Changing Values and Political Styles among Western Publics. Princeton University Press.
Jordan, G., Maloney, W., 1997. The Protest Business? Mobilizing Campaign Groups. Manchester University Press.
Lee, Y., So, A. (Eds.), 1999. Asia's Environmental Movements. M. E. Sharpe, Armonk NY and London.
Mittelman, J.H., 1998. Globalisation and environmental resistance politics. Third World Quarterly 19 (5), 847–872.
Nash, R., 1967. Wilderness and the American Mind. Yale University Press, New Haven CT.
Oelschlaeger, M., 1991. The Idea of Wilderness: From Prehistory to the Age of Ecology. Yale University Press, New Haven CT.
Rootes, C., 1995. A new class? The higher educated and the new politics. In: Maheu, L. (Ed.), Social Movements and Social Classes: The Future of Collective Action. Sage, London, pp. 220–235.
Rootes, C., 2004. Environmental movements. In: Snow, D.A., Soule, S.A., Kriesi, H. (Eds.), The Blackwell Companion to Social Movements. Blackwell, Oxford and Malden, MA, pp. 608–640.
Rootes, C., 2006. Facing south? British environmental movement organisations and the challenge of globalisation. Environmental Politics 15 (5), 768–786.
Rootes, C., 2009. Environmental NGOs and the environmental movement in England. In: Crowson, N., Hilton, M., McKay, J. (Eds.), NGOs in Contemporary Britain: Non-state Actors in Society and Politics since 1945. Palgrave Macmillan, Basingstoke, pp. 201–221.
Rootes, C., 2012. New issues, new forms of action? Climate change and environmental activism in Britain. In: van Deth, J., Maloney, W. (Eds.), New Participatory Dimensions in Civil Society: Professionalization and Individualized Collective Action. ECPR/Routledge, Abingdon, pp. 46–68.
Rootes, C. (Ed.), 2003. Environmental Protest in Western Europe. Oxford University Press.
Rucht, D., Roose, J., 2003. Germany. In: Rootes, C. (Ed.), pp. 80–108.
Snow, D.A., 2004. Framing processes, ideology, and discursive fields. In: David, A.S, Sarah, A.S., Hanspeter, K. (Eds.), The Blackwell Companion to Social Movements. Blackwell, Oxford, pp. 380–412.
Wahlström, M., Wennerhag, M., Rootes, C., 2013. Framing "the climate issue": patterns of participation and prognostic frames among climate summit protesters. Global Environmental Politics 13 (4), 101–122.
Zald, M.N., McCarthy, J.D., 1987. Social Movements in an Organizational Society. Transaction Books, New Brunswick, NJ.
Zelko, F., 2013. Make It a Green Peace! The Rise of a Countercultural Environmentalism. Oxford University Press.

Relevant Websites

http://dotearth.blogs.nytimes.com/ – DOT Earth Environmental Blog.
http://memory.loc.gov/ammem/amrvhtml/ – The Evolution of the Conservation Movement.
http://www.grist.org – Grist Magazine.
www.epa.gov – U.S. Environmental Protection Agency.

Environmental Planning

John C Keene, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

With the remarkable growth in the commitment to making the activities and programs of governmental, corporate, and nongovernmental organizations more 'sustainable,' there has been a corresponding expansion in environmental planning at all levels and in all segments of society. In fact, environmental planning and management of the impacts of these different organizations' activities on the natural environment constitute key components of any comprehensive strategy for sustainable development. Governments, corporations, and nongovernmental organizations such as universities have begun to reshape programs in all of the following policy areas so as to increase their contribution to greater global sustainability: managing urban growth, land development and building construction, use and management of water, increased energy efficiency in building use and operation, increased transportation efficiency in both mass transit and automobile usage, management and reduction of environmental pollution, taking into account environmental and social impacts as well as financial profitability in managing corporate policies, more efficient production and distribution of goods and services, and better management of solid waste recycling and disposal. Many organizations have voluntarily developed policies to make their operations climate neutral within 40 or 50 years.

Introduction

For millennia, humans have realized the importance of conserving important natural resources and minimizing the damage that their activities cause to the natural environment in which they live. They have followed broad strategies to protect environmental resources and to reduce or eliminate these impacts. They sought to shape land use patterns so as to protect people from natural hazards such as floods, to prevent inappropriate consumption of natural resources, and to control emissions or discharges directly. Environmental planning embraces those activities of government agencies and private organizations whose primary purpose is to manage the impacts of the human use on the natural environment, by taking thought for the future and choosing courses of action that promote appropriate goals.

In the first decade of the twenty-first century, a paradigm shift has occurred as the concepts of sustainable development have burrowed their way into more traditional ways of viewing man's impact on the natural environment. This shift has been fitful, with false starts and stops and uneven recognition and acceptance in various segments of society. It has been comprehensive because, often first motivated by concerns over such matters as global warming, loss of biodiversity, destruction of the stratospheric ozone layer, serious water and air pollution, depletion of major ocean species, it has come to address virtually all aspects of human behavior from energy conservation, to building construction and operation, waste disposal, transportation, production and consumption, conservation of natural resources, public health, social inclusion, and global poverty (see, e.g., The European Community's Principles for Sustainable Development, 2005).

Environmental planning has become a critically important component of the new paradigm of sustainable development. Because of the scope and complexity of the activities it addresses and of their far-reaching and variable impacts, the field of environmental planning defies easy definition. Furthermore, conceptions of the relationships between humanity and the environment change over time, so that the character, goals, methods, and indeed, the very ethical foundations of environmental planning are continuously evolving. At the present time, environmental planners carry out such varied responsibilities as the analysis of the natural environment that constitutes a component for comprehensive plans, assessment of and developing strategies for reducing the emission of greenhouse gases (GHGs) that are contributing to global warming, developing better methods of recycling, using, and disposing of liquid and solid wastes, the preparation of environmental impact statements for governmental programs, the drafting of plans for recreational areas, parks, and other types of open space, and the development of policies for the protection of prime farmland, wetlands, floodplains, and other natural resource areas as part of broader urban growth management policies and plans.

In the last decade, many universities, business corporations, cities, provinces, and nations have begun to formulate policies under the general rubric of climate action plans or sustainable development strategies. Although there is certainly no uniformity or even consensus about the nature and content of these actions, they appear to be converging on a set of policies that seek both to make the institutions, businesses, and governments more economically efficient and to promote the public interest by conserving energy, reducing the rate of climate change, and reducing the amount and type of environmental pollution that results from man's activities. The intriguing thing about many of these policies is that they have been undertaken voluntarily without any government mandate.

Furthermore, it is essential to understand that, to paraphrase the words of the National Research Council in its Report, *Toward Sustainable Agricultural Systems in the 21st Century* (*2010*), sustainable development "is a complex, dynamic and political concept that is inherently subjective in that different groups in society place different emphases on each of the its numerous goals … The pursuit of sustainability is not a matter of defining sustainable or unsustainable development, but rather, it is about assessing whether choices of the practices and systems would lead to more or less sustainable systems as

measured by society's goals. Improving sustainability will require identification of key metrics and indicators that can measure progress toward these goals, together with the monitoring and collecting of date and using adaptive management."

A Brief History of Environmental Planning

Although the term, environmental planning, is of relatively recent origin, mankind has undertaken the activities it now embraces for millennia. At least 5000 years ago, the earliest urban civilizations – the 'hydraulic' societies of Sumer, Egypt, and the Indus Valley civilization in present-day Pakistan – undertook major programs to harness natural forces. Large-scale irrigation projects provided the water needed to support the broad agricultural foundation on which the urban civilizations were built. Certainly, the Romans engaged in a substantial amount of what we would today include in environmental planning as they built extensive public works, such as the dams, reservoirs, aqueducts, and the water supply and sewerage systems that served their far-flung cities.

In historic Islamic cities from Morocco to Iran, abattoirs, tanneries, and textile dyeing operations were relegated to the outer, downstream, edges of the medina (the old city), far from the Grand Mosque, the central souk (market), and the principal residential areas. Across the Islamic world after the eighth century, gardens were associated with paradise, and many of the palaces contained elaborate gardens that gave their owners respite from the harshness of the climate. Superb examples such as the fourteenth-century Alhambra palace of the Nasrid caliph, Yusuf I, in Granada, Spain (Hoag, 1977), are creations that fall within the broad definition of environmental planning. Some represent the folly of ignorance of environmental imperatives, such as Fatehpur Sikri, Mughal Emperor Akbar's magnificent palace complex built near Agra in the second half of the sixteenth century. It was abandoned after less than 15 years of occupation because of an inadequate water supply system (Fatehpur Sikri).

In Europe, in the eleventh and twelfth centuries, after a long hiatus following the collapse of the Roman Empire in the fifth century, Italian cities adopted municipal codes that regulated both the disposal of wastes (forbidding, e.g., their discharge within the town) and the number and location of pollution-generating activities (Zupko and Laures, 1996). A growing interest in nature found expression in the Renaissance gardens of Florence, Rome, and Tivoli (McHarg, 1992). Here, in the sixteenth century, artists such as Bramante, Raphael, and Palladio imposed a simple geometry on the forms of nature. A century later, landscape gardeners such as André Le Notre developed formal Baroque gardens at Versailles and Vaux-le-Vicomte, that 'testified to the divinity of man and his supremacy over a base and subject nature' (McHarg, 1992). A strongly divergent tradition developed in England in the eighteenth century. Landscape gardeners such as Lancelot 'Capability' Brown and William Kent created the famed English landscape for their aristocratic clients, a setting for their manor houses that appeared natural but was as much an artifice of the human mind as the formal gardens of Paris and Rome (McHarg, 1992). Nature and its forms became something to be respected and idealized.

In the nineteenth century, interest in public health mounted, especially in the link between sewage discharges from rapidly growing cities, the pollution of their water supplies, and the outbreaks of typhoid and dysentery. Many major cities undertook the construction of public water supply and sewerage systems. In 1872, responding to a growing interest in the conservation of critically important natural areas, during the administration of President Grant, the United States created Yellowstone National Park, the first in what was to become an impressive system of parks. Frederick Law Olmstead designed a number of parks, the best known of which is Central Park in New York City.

The American city planning movement, which began in the early twentieth century, drew inspiration from the public health, housing, and City Beautiful movements, and from early German efforts to zone land according to its most appropriate land use. But it was concerned more with the management of new development by limiting densities and separating incompatible uses than it was with the moderation of human impacts on the natural environment and the conservation of natural areas. Environmental concerns took a backseat during World War I, the Roaring Twenties, the Depression, World War II, and the reconstruction and normalization years immediately thereafter. After World War II, Aldo Leopold articulated a new kind of ethic that called on mankind to serve as the steward of nature, not its sovereign (Leopold, 1949), and in 1962, Rachel Carson's *Silent Spring* fired the American imagination with her chronicle of the devastation that that country's profligate waste disposal, pesticide use, and weak environmental policies were wreaking on its environment (Carson, 1962). Ian McHarg argued eloquently and persuasively that plans for the development and redevelopment of rural and urban areas must be based on a comprehensive inventory of ecological data that present a complete picture of the geological, hydrological, pedological, limnological, physiographic, climatological, and biological systems in the area that is the subject of the plan (McHarg, 1969, 1992).

It was not until the 1960s and the early 1970s that a newly felt concern for the environment, and a deepening awareness of the damage that mankind was inflicting on it, resonated with something deep in people's souls. This found expression in the United States and across the world, in the enactment of a substantial body of environmental protection laws. More than in any preceding era, public and private decision makers have been called upon to take into account the environmental impacts of the various policies and infrastructural systems they propose and to develop appropriate planning strategies for dealing with them.

The recognition of the serious and growing impacts of man's activities on the environment and the development of the concept of sustainability as we now know it can be traced to the 1970s. The Club of Rome's *The Limits of Growth*, published in 1972, sounded a widely heard warning (Club of Rome, 1972). Also in 1972, the United Nations held a major conference on the Human Environment in Stockholm, and many view it to be the real font of the sustainability movement. The United Nations instituted its Environment Program in 1972, and the Cocoyoc Declaration in 1974 recognized that humanity was at a critical turning point. The Declaration stated: "Environmental degradation and the rising pressure on

resources raise the question of whether the 'outer limits' of the planet's physical integrity may not be at risk" (Friedman, 1992). Satisfying the basic needs of the world's poor was proclaimed as more important than simple growth maximization.

Most discussions of the concept of sustainability start with the definition found in the Report of the General Assembly of the United Nations' World Commission on Environment and Development (WCED) held in Stockholm in 1983. Its report, *Our Common Future* (often referred to as the Brundtland report and published in 1987) (Brundtland Report, 1987), articulated the following widely quoted definition of sustainability:

> Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs.

The Commission suggested a number of planning strategies for achieving sustainable development and recognized the importance of including social equity along with the more traditional environmental goals. It asserted that sustainability does not concern only the environment, it also must address economic inequalities and social inequities. Sustainable development rests on three pillars: protection of natural ecosystems, reduction of economic inequalities, and the creation of a more just world.

The 1992 United Nations Conference on Environment and Development (the Earth Summit) in Rio de Janeiro was a particularly important and influential international conference on environment. It adopted two major documents. The *Rio Declaration* recognized that social equity must coexist with economic development, and called on the polluting nations to contribute to cleaning up the environment. The 1992 Conference's *Agenda 21* was a lengthy document that set out a full set of goals and over 120 strategies for achieving them. Again, it recognized that "[H]humanity stands at a defining moment in its history. We are confronted with a perpetuation of disparities between and within nations, a worsening of poverty, hunger, ill health and illiteracy, and continuing deterioration of the ecosystems on which we depend for our well-being" (Sitarz, 1993). In the United States, in 1993, President Clinton created the President's Council on Sustainable Development and charged it with developing a national strategy for sustainable development in 1999, major components of which entailed environmental planning.

The 1996 Habitat II Conference in Istanbul continued that line of thinking, and confirmed that social equity and human resource development are essential components of a truly sustainable development process and a just world. Some 31 countries signed the Kyoto Protocol in December 1997, committing themselves to protecting the environment by bringing about a reduction in the emissions of GHGs, such as carbon dioxide, nitrous oxide, methane, and chlorofluorocarbons, by 5.2% from their 1990 levels by 2012 (Coogan, 2006). The European Union and its member states ratified it in 2002. The United States is not a party to the Protocol.

The European Community has been developing a comprehensive sustainable development policy since before the beginning of the twenty-first century. The Amsterdam Treaty of 1997 established sustainable development as a central objective of the Union. In 2001, the European Union adopted its first Sustainable Development Strategy at the Gothenburg Summit. In 2005, the European Council adopted a Draft Set of Guiding Principles for Sustainable Development (European Community's Principles for Sustainable Development, 2005), which recognized the international dimension of a commitment to sustainable development and, in June 2006, it adopted a renewed Sustainable Development Strategy. It recognizes that sustainability extends beyond purely environmental concerns to embrace social equity, reduction of poverty, respect for the needs of developing countries, employment policies, improved education, the rule of law, and social justice.

A Comprehensive Sustainability approach, such as is embodied in the European Sustainable Development Strategy, adopted in 2006 (European Union Sustainable Development Strategy, 2006), and supplemented in 2009, identified 11 key issues:

1. Establish Climate Change and Energy Policy.
2. Promote Sustainable Transportation Modes.
3. Promote Sustainable Production and Consumption.
4. Conservation and Management of Natural Resources.
5. Improve Public Health.
6. Facilitate Social Inclusion, Demography, and Migration.
7. Reduce Global Poverty and Face Sustainable Development Challenges.
8. Improve Education and Training.
9. Research and Development: Encourage research into sustainable development, economic growth, life cycle analysis, and risk analysis.
10. Develop Innovative Strategies to Move toward Sustainable Development.
11. Develop and Refine Indicators of Sustainability.

The 2009 Copenhagen talks on Green House Gas Reduction were inconclusive. The delegates from 190 nations who met in Cancun in December 2010 did not reach any broad agreement on steps to be taken to counteract climate change, but they did agree on a number of matters. For instance, developed industrial countries would contribute $30 billion over the next 2 years to help countries that are most vulnerable to the negative effects of climate change to reduce those effects. It is too soon to know what the effects of the Cancun Agreements will be, but a report by the United Nations Environment Program (United Nations Environment Program, November 2010) concluded, after a careful analysis of over 220 projections of GHG production, that the Copenhagen agreements would not be sufficient to meet the goal of limiting the increase in average global temperature to the 2 °C stabilization target.

In late 2011, the 17th Conference of the Parties met in Durban, South Africa, to address the next steps following the 1997 Kyoto Protocol of the United Nations Framework Convention on Climate Change. The final agreement committed the participating nations to work toward an agreement for the post-2020 era on steps to slow climate change. It launched the Green Climate Fund to support mitigation and adaptation in developing countries. It also clarified procedures for measuring, reporting, and verification of participating countries' climate efforts (Center for Climate and Energy Solutions Report).

In June 2012, over 50 000 people participated in the United Nations Conference on Sustainable Development (referred to colloquially as Rio +20) in Rio de Janeiro. While the conferees adopted a 283-paragraph agreement, 'The Future We Want' (The Future We Want, 2012), they made few if any new commitments to action. One-third of the agreement reaffirms earlier decisions and principles. Another portion confirms goals that should drive nations' activities, such as protecting the rights of indigenous peoples and promoting food security. The rest of the document stresses areas of heightened concern such as conservation of ocean biota, promoting sustainable production, and provision of potable water supplies. While the nations agreed only to develop a new set of sustainable development goals by 2013 (Rio +20 Summary), over 700 voluntary commitments were made by an assortment of nongovernmental organizations, governments, and private corporations that may result in more than $500 billion in expenditures promoting sustainable development (UN Rio +20 Conference Report). Among the most important of these private commitments were 'The Business Day,' which recognized the importance of the contributions of business corporations to sustainable development; 'the World Conference on Justice, Governance, and Law for Environmental Sustainability,' which recognized the central importance of the Rule of Law in achieving a sustainable society; 'the Global Initiative for Resource Efficient Cites,' which recognized the more than half of the people in the world live in cities that offer major opportunities for greater energy efficiency and more sustainable development; 'Global Insurance Principles,' which will promote Principles for Sustainable Insurance; and 'Sustainable Energy for All,' which will promote efforts to achieve sustainable energy by 2030 (Rio +20 Evaluation).

The Intellectual Context of Environmental Planning

As we have noted above, in the process of the shift to the new paradigm of sustainable development, environmental planning has been incorporated as a key element in many aspects of the broader efforts to achieve more sustainable development. Significantly, the locus of planning activities has broadened from primarily national governments to include provinces, states, corporate enterprises, and major institutional actors such as universities. Thus, environmental planners play important roles in:

1. Comprehensive growth management seeks to advance a broad spectrum of goals and objectives. It addresses the location, density, use, height and bulk, and phasing or timing of urban development and redevelopment, and the provision of various infrastructural systems. At the same time, it seeks to protect critically important natural resources, such as farmland, floodplains, marshes, and other areas of critical ecological concern. Comprehensive planning takes into account the underlying geology of the area, the impact of proposed development on groundwater, rivers, lakes, wetlands, and coastal waters, the impact of manufacturing and transportation activities on air quality, the effects of urbanization on the natural biota, the impacts of suburban development on the supply of prime agricultural land, and the effects of solid and liquid waste disposal on the natural environment. Environmental planning is a key component of comprehensive growth management.
2. Environmental pollution control policies are broadly concerned with air, water, and earth pollution, and such global problems as loss of biodiversity, global warming, and depletion of the stratospheric ozone level. Decades of contemporary efforts to protect the environment and reduce environmental pollution have confirmed the oneness of the mankind/natural environment system. Changes in the natural environment can cycle back and cause serious harm to mankind. Transportation, land development, and land use affect environmental quality, whether it be by covering over wetlands, changing the hydrological characteristics of watersheds so as to increase flooding and aggravate drought conditions, converting farmland to residential developments and shopping centers, or creating an automobile-dominated transportation system that generates a major proportion of dangerous environmental pollutants.
3. More efficient generation and use of energy leading to significant reductions in the use of nonrenewable energy sources and in the emissions of GHGs requires environmental planning in the areas of land development, materials conservation, building design, indoor environmental quality, transportation (including the design of more efficient automobiles, the increased use of renewable fuels, and the increase in the use of mass transit), the use of more efficient production systems, supply chain management, and life cycle analysis to reduce the lifetime impacts of different products on the natural environment, and the conservation and management of natural resources.

Research Issues and Major Categories of Environmental Planning

In the field of environmental planning, there are a number of areas of research that are attracting interest from academicians and practitioners. Especially since 2000, environmental planning, often primarily for the purpose of making operations more sustainable, occurs at the various levels of government, in a growing number of private institutions and nongovernmental organizations, and in business corporations.

Research on the Scientific Basis of Environmental Planning

Considerable work is being done in the natural sciences and ecology, generally, to strengthen the scientific underpinnings of an effective environmental planning effort. Some subjects range is fairly specific, such as determining a river's flood plains for differing flood recurrence intervals, or developing more cost-effective ways to remediate polluted industrial sites. Others are much broader, such as determining the extent to which the emission of GHGs from human activities is a substantial cause of global warming, the rise of sea levels, and the increasing violence of weather patterns. The growing emphasis on sustainable development has led to a substantial increase in the benchmarking of building energy consumption and the design of more efficient means of heating and cooling

buildings, the development of more efficient automobile engines that either use a renewable form of energy or are more efficient in their use of nonrenewable forms, the discovery of better ways to minimize raw material consumption in manufacturing processes, to recycle waste products, and to dispose of the remaining wastes in an ecologically protective manner. As nations come to recognize the true extent of global warming, they are exploring ways of planning how best to adapt to the environmental changes occasioned by the process, in the event that they are not able to deduce the rate of climate change significantly.

Environmental Planning by Governmental Agencies

Environmental planning embraces more than the various studies that environmental planners perform in the process of preparing a comprehensive plan or a more specific plan for making buildings more energy efficient or reducing the GHG emissions of transportation systems. It includes a careful assessment and selection of the most appropriate strategies of implementing the various goals of environmental planning and protection. Over the years, the Congress of the United States, for example, has created a virtual arsenal of methods for achieving environmental objectives (Plater et al., 2010). These are grouped in the taxonomy that follows.

A Taxonomy of Strategies for Implementing Environmental Plans

1. Strategies based on disclosure. Environmental assessment and impact statements covering proposed projects such as those mandated by the United States' National Environmental Policy Act of 1969 (NEPA), and numerous state counterparts, with respect to major governmental programs have a significant effect on the environment. This enables affected governmental agencies and citizen groups to find out about proposed actions and to assert their influence to modify their outcome. The technique has been extended to require business corporations to make disclosures of, for instance, the types of hazardous substances they are storing in their plants (required by the Emergency Planning and Community Right to Know Act).
2. Planning strategies. These include requirements for comprehensive planning, critical areas management such as floodplains or important marshes, facility siting laws for power plants and other possibly unwanted land uses, and resource management laws.
3. Strategies based on increased intergovernmental coordination, often accompanied by greater public participation in the decision-making process. For instance, the American Intergovernmental Coordination Act and NEPA call for collaboration among the various governmental agencies with responsibilities in a particular region.
4. Command and control strategies. These approaches have long been the workhorse of environmental planning and protection. A government establishes a regulation that restricts private activity, such as a floodplain zoning ordinance or a statute seeking to reduce air pollution by setting permissible emission levels. This regulation sets a standard with which the private actor must comply, and then applies sanctions to those who violate the standard. Command and control strategies may take a number of forms: (1) from the simple specification of certain environmental standards to harm-based strategies that set standards designed to protect the public health (such as the Clean Air Act's National Ambient Air Quality Standards for key pollutants), (2) technology-based strategies that require the installation of specified technology or technology that achieves a specified level of pollution reduction, (3) technology-forcing strategies, (4) roadblock strategies that prohibit certain activities such as the Endangered Species Act's prohibition of government programs that harm endangered species or destroy their habitat, (5) strategies for controlling market access, prohibiting certain kinds of actions, and requiring that firms internalize formerly externalized environmental costs, and, finally, (6) remedial liability strategies such as those of the Superfund Law that imposes strict, joint, and retroactive liability on any party that has contributed hazardous wastes to, or owned, a Superfund Site.
5. Strategies based on market-based incentives. Many American jurisdictions have experimented with techniques that seek to influence private actors' behavior by providing economic incentives or disincentives that induce citizens to change their behavior simply because it is more profitable. The 1990 amendments to the United States' Clean Air Act established a system for trading emissions allowance for oxides of sulfur (SO_x) and nitrogen (NO_x) that permit electric utilities to meet statutory requirements either by reducing their emissions by installing more effective air pollution control devices or by buying emission rights from other producers, whose emissions are well below their assigned level.
6. Tax expenditures. Many nations have created tax incentives that seek to increase after-tax income of individuals and companies that engage in conduct that is either environmentally protective or environmentally enhancing. Tax expenditures are variations from the normal structure of a particular tax designed to influence a business or individual's behavior by increasing its or his after-tax profits, such as investment tax credits for the installation of insulation, or more energy-efficient heating or cooling equipment.
7. Spending strategies. These strategies include grant-in-aid and loan programs, such as the United States $60 billion program for the construction of Publicly Operated Waste Water Treatment Plants under the Clean Water Act.
8. Demand management strategies. A number of programs seek to control demand for a particular product or facility, such as bottle return fees for soft drink bottles, high occupancy vehicle lanes, and parking limitations in congested or polluted areas.

Environmental Planning by Cities

Dozens, if not hundreds, of cities around the world are often undertaking voluntary programs to develop and implement strategies for reducing the emissions of GHGs and for making both municipal and private activities more sustainable, in an effort to reduce climate change and make their operations more efficient. In most cases these efforts started after 2000. They are best exemplified by the creation in 2005 of the C40 Cities Climate Leadership Group (C40). In 2011, it became fully

integrated with former President Bill Clinton's Climate Initiative Cities Program in the Carbon Disclosure Project, and received a major grant from the Bloomberg Philanthropies (C40 Cities). C40 Cities was created to reduce carbon emissions and to increase energy efficiency of large cities around the world. In 2011, its C40 Cities Climate Leadership Group counted 58 major cities as members from every continent, including Sydney, Tokyo, Beijing, Jakarta, Delhi, Cairo, Johannesburg, Moscow, Berlin, Paris, London, Bogotá, Buenos Aires, Mexico City, New York, Chicago, Los Angeles, and Toronto (C40 Cities, KPMG Report, 2011). The CDP 2011 Global Report on Cities was based on a questionnaire sent out to 64 cities to which 48 responded. It estimated that about 12% of the world's carbon emissions were produced in its cities' metropolitan areas and roughly one-eighth of the world's population lived within their limits (C40 Cities).

The CDP 2011 Global Report permits a number of generalizations. Many of the world's major cities are located in coastal areas and are particularly vulnerable to rising sea levels and more violent weather events that result from global warming. Thus, they have special incentives to begin to understand the nature and extent of their GHG emissions and to start finding ways to reduce emissions from both municipal and private sector and community facilities and operations. Their efforts have been focused in three major areas: (1) governance and planning, (2) GHG measurement and management, and (3) climate change risk assessment and management.

In many cases GHG emissions control was incorporated into regular comprehensive planning activities. Three typical environmental plans for dealing with climate change and a number of related environmental issues are those of New York City (New York, PlaNYC, 2009), Philadelphia (Philadelphia Greenworks, 2009), and London (London Sustainability, 2012). (This is the second edition of London's 2007 plan and it incorporates all of London's sustainability commitments.)

Most cities developed reporting systems for both municipal emissions and community emissions. They used a variety of methodologies to report these data, including spreadsheets and proprietary programs, that are designed specifically for this purpose.

These major cities are taking a hardheaded approach to risk assessment and management of the threats posed by global warming. As indicated above, many consider themselves at risk because of their low-lying coastal locations and are using a variety of different methods to project and assess these impacts on their physical structure and economies. Few of the cities have actually implemented their environmental plans for adapting to the anticipated risks because of the very novelty of the process of assessing climate change risks.

The CDP Report, *Cities 2011: A Global Report on C40 Cities*, reveals strategies and action programs that were still in their very early stages. Most have been undertaken since 2006. They show a need to (1) improve and standardize GHG measurement methodologies; (2) accelerate the development of data management software; (3) develop financial forecasting related to climate change investment; (4) create better tools for city-level risk assessment; (5) develop practical strategies and programs for reducing carbon emissions in both municipal and community sectors; and (6) perfect ways of monitoring progress in meeting programmatic goals (C40 Cities KPMG, CDP Report, 2011).

Many of the steps local governments can take are summarized in the United States Environmental Protection Agency's *Sustainable Design and Green Building Toolkit* (USEPA Toolkit, 2010). Buildings account for 40% of total energy consumption, 13% of water use, and 160 million tons of construction and demolition debris per year. More difficult to quantify but nevertheless significant are the effects of urban sprawl and its extensive land development patterns, Brownfield remediation, and loss of prime farm and forestland. The EPA's Sustainable Design Toolkit identifies five major aspects of land development and building construction that have impacts on the sustainability of communities. These categories have, in turn, a number of subcategories, each of which has its sustainable design objective, rationale, and overarching question. The report suggests strategies, tools, and techniques for engaging in environmental planning that will increase the level of sustainability for each of the categories and subcategories. The five categories are as follows:

1. Sustainable Sites and Responsible Land Development: This category addresses the development and minimization of the negative effects of land development and the impact of buildings on the natural environment.
2. Materials and Resource Conservation: This category examines ways of minimizing waste from construction sites and the use of green materials during construction.
3. Energy Conservation and Atmospheric Quality.
4. Water Efficiency, Conservation, and Management: This category addresses buildings' water consumption and ways of minimizing its impact on the environment.
5. Indoor Environmental Quality: This category addresses methods of improving indoor air quality.

Distributed and Often Voluntary Environmental Planning by Large Institutions and Business Corporations

Environmental Planning by Universities and Colleges

One of the remarkable developments during the first decade of the twenty-first century has been the decisions by hundreds of colleges and universities around the world, but especially in the United States, to undertake serious efforts to develop Climate Action Plans that will enable the institutions to reduce their GHG emissions. Most of these programs were adopted after 2000. In 2007, The American College & University Presidents' Climate Commitment organization (ACUPCC, 2007) prepared a document for signature by university and college presidents that committed the institution to agree to initiate two or more of seven actions designed to reduce GHG emissions within 2 years after their implementation start date. These actions included, for instance, (1) adopting a policy that all new campus construction would be built at least to the United States Green Building Council's Leadership in Energy and Environmental and Design (LEED) silver standard or its equivalent (USGBC LEED Program, 2012), (2) committing to begin purchasing at least 15% of the institution's energy consumption from renewable sources within 1 year of signing the commitment, and (3) adopting an energy-efficient appliance purchasing policy requiring purchase of ENERGY STAR

certified products in all areas for which such ratings exist (ACUPPC, 2007). By the end of 2011, 674 presidents had signed the commitment (many of whom headed multicampus universities such as the University of California) and 451 institutions had submitted Climate Action Plans. In other countries, many universities have also established significant sustainable development strategies, such as the universities of Oxford, Cambridge, and St Andrews.

In broad outline, a typical Climate Action Plan sets up a series of steps, such as those undertaken by the University of Pennsylvania in its Plan (the University of Pennsylvania CAP, 2009):

1. Undertake a complete inventory of campus emissions and emissions attributable to university activity, measured in metric tons of carbon dioxide equivalent (MTCDE).
2. Project the various types of emission ahead 40 years on the assumption that no actions were taken to reduce them (the 'business as usual' projection), including estimates of newly constructed floor space, and growing energy-consuming activities.
3. Break the total carbon dioxide emissions into different categories and develop strategies for reducing these emissions in each of these categories over a period of 40 years. Penn's environmental planners focused on six strategic areas:
   - Reduction of electric intensity.
   - Design and build new high-performance buildings by using the United States Green Building Council's LEED standards for new buildings and meeting the Energy Star 90 standard of the American Society of Heating, Refrigerating and Air-Conditioning Engineers.
   - Reduction of energy use of existing campus buildings. Improve the efficiency of existing utilities by reducing system losses, insulating steam lines, and repairing stem leaks.
   - Reduce air travel and encourage the use of mass transit and carpooling by students and employees.
   - Purchase Renewable Energy Credits. A Renewable Energy Credit represents and conveys the environmental and other nonpower attributes of 1 MW-h of renewable electricity generation.
4. Develop programs in each of these areas for reducing the amount of GHGs produced thereby. The University of Pennsylvania's planners projected 2050 total emissions under the 'business as usual' alternative to be about 600 000 MTCDE. By applying the various strategies outlined above, they estimate that, by FY2050, the MTCDE of each type would be reduced by the following percentages:
   - Reduce electricity intensity – 18%
   - High-performance new buildings – 16%
   - Renovate existing buildings – 27.8%
   - Recommission existing building systems – 17%
   - Improve efficiency of existing utilities – 3.5%
   - Reduce air travel and other emissions – 3.5%
   - Purchase Green Power RECs – 13%

This approach was based on concepts developed by Stephen Pascala and Rob Socolow and presented in an article in *Science* journal in 2004 (Pascala and Socolow, 2004). The University of Pennsylvania's Climate Action Program presents it graphically in Figure 1.

Almost all of the university climate action plans were adopted in the first decade of the twenty-first century. Even in the short time since then, many have evolved to focus on a broader range of environmental protection measures, so that they promote the broader goals of sustainability rather than just those of moving toward carbon neutrality. Typically, they address reduction of carbon emissions, improving energy efficiency, green building for new facilities, waste minimization and recycling, increased transportation efficiency, better use of

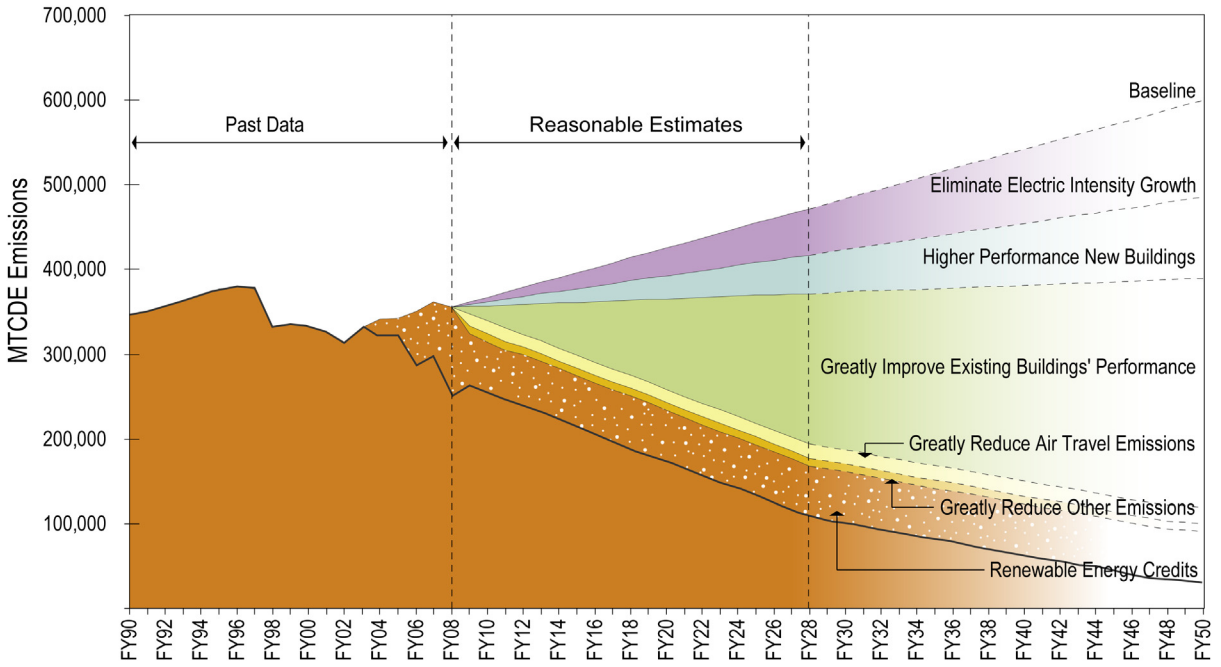

Figure 1 The University of Pennsylvania's 'Wedge Diagram' showing how each sector will contribute to reduction of CO_2 emissions.

water, refining food systems, and education in sustainable development policies and operation.

Environmental Planning by Business Corporations

Every business corporation is subject to the environmental protection laws of countries where it operates, such as those governing air pollution, water pollution, solid waste management and disposal, and use and storage of hazardous materials. Reviewing the corporate environmental planning that is involved in complying with those laws is beyond the scope of this article, other than as reviewed in Section Environmental Planning by Governmental Agencies, above.

Largely since 2000, there has been a notable increase in the extent to which private sector corporations have taken issues of sustainability into account in their corporate management and strategies. This growing interest has been evidenced in a number of ways (Feldman, 2009). Many companies have begun to speak of the 'triple bottom line,' which adds environmental and social measures of performance to the traditional financial ones. Others treat it under the rubric, 'corporate social responsibility (CSR),' and still others embrace sustainable business practices in many parts of their operations. Many corporations have created new management positions, such as Vice President for Sustainability or Vice President for CSR (Blackburn, 2007).

Corporations are adopting these policies for a variety of reasons. Most calculate that making their operations more sustainable will reduce energy, production, and marketing costs. Others are doing it because it is good public relations and they want to match shareholders rising expectations. In addition, market analysts are beginning to analyze and disclose corporations' environmental performance (Feldman, 2009; Makower, 2012), so that the investing public has greater access to information on corporate sustainability performance.

Conclusion

In the first few years of the twenty-first century, we have witnessed a remarkable rethinking of man's proper relationship with the natural environment, a process that will have profound implications on the full range of societal activities. For instance, the European Union's 2006 Sustainable Development Strategy addresses several pressing environmental issues, such as climate change, environmental pollution, loss of biodiversity, safe water supplies, energy conservation, building construction, and waste recycling and disposal, but it also addresses broader societal and economic issues such as public health, social inclusion and equity, global poverty, and underdevelopment. Thus, environmental planning is in the process of being reframed as part of more comprehensive strategies designed to achieve sustainability in the broadest sense of the word.

This process of rethinking societal goals is occurring in many different milieux. Most obviously, it is taking place at the ongoing series of international conferences that are addressing climate change and GHG emissions. It is taking place in national legislature after national legislature, and in regional and local governments. Most significant, perhaps, is that many large business corporations and substantial nongovernmental institutions such as universities are also becoming engaged in a broad spectrum of programs and policies that are designed to reduce the environmental impacts of these organizations' activities.

This process is by no means uniform and coordinated. It is fitful and uneven with advances here and retreats there. Because of its novelty, many organizations are only now beginning to face the political and economic implications of the policies and programs they have proclaimed. It is one thing, they are finding, to publish a climate action plan or a strategy for sustainable development. It is quite something else to find the resources and political will to actually implement the program or strategy. It is especially difficult in this time of recession, when income has declined and many needs demand attention, to move environmental, sustainability, and energy efficiency actions up on the list of budgetary priorities, no matter how persuasive are the arguments for protecting the environment and making the activities of the firm, the organization and society as a whole, more sustainable. Only time will tell whether the remarkable explosion in programs for environmental protection and sustainability will result in deep changes in the impacts of human activities on the natural environment.

Bibliography

Blackburn, William R, (Ed.), 2007. The Sustainability Handbook. The Environmental Law Institute, Washington, D.C.

Carson, Rachel, 1962. Silent Spring.

Club of Rome, 1972. The Limits of Growth.

Coogan, Douglas C, 2006. Corporate Governance and Climate Change. Making the Connection Ceres, Inc., New York.

Feldman, Ira Robert, 2009. Business and industry: transitioning to sustainability. In: Dernbach, John C. (Ed.), Agenda for a Sustainable America. Island Press, Washington, DC.

Friedman, John, 1992. Empowerment: the Politics of Alternative Development., Blackwell: Oxford

Hoag, J., 1977. Islamic Architecture. Rizzoli, New York.

Leopold, Aldo, 1949. A Sand County Almanac. New York.

McHarg, Ian, 1969 and 1992. Design with Nature. Natural History Press, John Wiley & Sons, Garden City, New York.

Pascala, S., Socolow, R., 2004. Stabilization Wedges: Solving the climate problem for the next 50 Years with Current Technologies. Science 305 (5686), 968–972.

Plater, Z., Abrams, R.H., Graham, R.L., 2010. Environmental Law and Policy: Nature, Law and Society, fourth ed. Aspen Publishers, Charlottesville, VA.

Sitarz, Daniel, 1993. Agenda 21: The Earth Summit Strategy to Save Our Planet.

The Brundtland Report, 1987. World Commission on Environment and Development (1987). Our Common Future. Oxford University Press, Oxford (The Brundtland Report).

The United Nations Conference on Sustainable Development, 2012. The Future We Want, http://www.uncsd2012.org/content/documents/727The%20Future%20We%20Want%2019%20June%201230pm.pdf.

The United States Green Building Council, 2012. LEEDS System http://www.usgbc.org/DisplayPage.aspx?CategoryID=19.

United States Environmental Protection Agency, 2010. Sustainable Design and Green Building Toolkit. U.S. Government Printing Office, Washington. DC.

Zupko, R., Laures, R., 1996. Straws in the Wind: Medieval Urban Environmental Law. Westview Press, New York.

List of Relevant Websites

ACUPCC, 2007. The American College & University Presidents Climate Commitment Organization – http://acupcc.org.

BRE Trust, The BRE Environmental Assessment Method – http://www.bream.org/page.jsp?id=66 and http://www.breeam.org/.
The Center for Climate and Energy Solutions report – http://www.c2es.org/blog/diringere/durban-how-big-a-deal.
Fatehpur Sikri – http://whc.unesco.org/en/list/255/.
The European Community's Principles for Sustainable Development, 2005. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2005:0658:FIN:EN:PDF; http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2005:0218:FIN:EN:PDF.
http://www.fatehpursikri.org/fatehpur-sikri-history.html.
University's Greenhouse Gas Policy – http://green.harvard.edu/greenhousegas.
European Community: Sustainable Development Strategy, 2006 – http://register.consilium.europa.eu/pdf/en/06/st10/st10917.en06.pdf.
C40 Cities – http://live.c40cities.org/about-us.
C40 Cities KPMG Report, 2011. http://live.c40cities.org/about, KPMG, CDP Cities 2011: A Global Report on Cities.
Makower, Joel, 2012. "The State of Green Business 2012," – http://www.greenbiz.com/research/report/2012/01/state-green-business-report-2012.
New York, PlanNYC 2011 – http://nytelecom.vo.llnwd.net/o15/agencies/planyc2030/pdf/planyc_2011_planyc_full_report.pdf.
London Sustainability Plan, 2012 – http://www.london2012.com/documents/locog-publications/london-2012-sustainability-plan.pdf.
U.S. E.P.A. Energy Star, 2012 – http://www.energystar.gov/.
U.S.G.B.C LEED Program – http://www.usgbc.org/DisplayPage.aspx?CategoryID=19 http://rs.acupcc.org/stats/tangible-actions/
The United Nations Environment Program, 2010. The Emission Gap Report: Are the Copenhagen Accord Pledges Sufficient to Limit Global Warming to 2° C or 1.5° C? Available from: www.unep.org/publications/ebooks/emissionsgapreport.
University of Cambridge Sustainability – http://www.admin.cam.ac.uk/offices/em/sustainability/environment/climate/.
University of Oxford Sustainability – http://www.admin.ox.ac.uk/estates/environment/.
University of St. Andrews Sustainability – http://www.st-andrews.ac.uk/about/Sustainability/.
Harvard University GHG Plan – http://green.harvard.edu/greenhousegas.
BRE Assessment – http://www.bream.org/page.jsp?id=66.
United States Environmental Protection Agency, 2012. Energy Star Program – http://www.energystar.gov/.
United States Environmental Protection Agency, Green Power Partnership – http://www.epa.gov/greenpower/gpmarket/rec.htm.
University of Pennsylvania's Climate Action Plan, 2009 – http://www.upenn.edu/sustainability/pdf/PENN-2009-Climate_Action_Plan.pdf.
University of Pennsylvania CAP Progress Report, 2011 – http://www.upenn.edu/sustainability/pdf/PENN_2011_Climate_Action_Plan_Progress_Report.pdf.
World Resources Institute, 2004. Greenhouse Gas Protocol – http://pdf.wri.org/ghg_protocol_2004.pdf.
Rio +20 Summary – http://www.wspenvironmental.com/newsroom/news-2/view/summary-outcomes-from-rio20-for-business-365.
U.N. Rio +20 Conference Report – http://www.uncsd2012.org/allcommitments.html.
Rio +20 Evaluation – http://www.cloudofcommitments.org.

Environmental Policy: Protection and Regulation

David Benson, Environment and Sustainability Institute (ESI), University of Exeter, Penryn, UK
Andrew Jordan, School of Environmental Sciences, University of East Anglia, Norwich, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A. Jordan, volume 7, pp. 4644–4651, © 2001, Elsevier Ltd.

Abstract

Environmental policy is primarily concerned with how to govern the relationship between humans and the natural environment in a mutually beneficial manner. Traditionally, it has been defined in terms of the problems it addressed, such as controlling pollution and waste flows and limiting habitat loss. However, policy makers have begun to appreciate that environmental policy will only be genuinely successful when it is sensitively integrated with other sectors and policy areas. Politically speaking, 'environmental policy integration' represents a new but potentially difficult stage in the continuing metamorphosis of environmental policy into a much wider policy supporting sustainable human development.

Environmental policy is essentially concerned with how best to govern the relationship between humans and the natural environment for their mutual benefit. This broad definition successfully captures the unbounded character of the environment as a distinct public policy field, touching as it does upon virtually *every* aspect of modern life. However, it lacks specificity: most areas of government policy affect the world 'out there.' A more precise definition is achieved by disaggregating the term into its component parts – i.e., 'environment' and 'policy' – although both are difficult to pin down.

Within politics, 'environment' remains a contested concept. Albert Einstein famously defined the environment as "everything that isn't me." Environmentalists would argue his interpretation fundamentally depersonalizes an inherently subjective, spiritual experience specific to every person. However, they would accept that his implied notion of 'everythingness' is useful as it powerfully confirms the complex and essentially unbounded nature of environmental policy, necessitating responses which are sensitively integrated with the main sectors of society. On this line of reasoning, *every* policy area has environmental implications. But the very 'everythingness' of the environment is why politicians often struggle to address the underlying roots of environmental damage through conventional sound-bite politics and ad hoc policy interventions. In solving environmental problems, politicians have to convince individuals to adapt their behavior for the good of society and future generations – never easy when benefits are diffuse and costs (or benefits foregone) are immediate. Finally, environmental issues raise difficult political, moral, and ethical questions which vote-chasing politicians instinctively avoid, thereby marginalizing environmental issues within mainstream politics.

If 'the environment' is amorphous then the term 'policy' is equally slippery. A dictionary definition is "a course of action or principle adopted or proposed by a government." However, some policies comprise a series of decisions, layered on top of one another. Moreover, policies may be formally adopted by national governments but the details of implementation are often determined by lower levels of government. Therefore, when a local pollution control officer negotiates with a polluting factory over a site license or emissions permit, he/she is arguably 'making' policy. Finally, policies also serve a symbolic purpose. Some are only ever statements of intent which governments have no real intention of enacting. Policy, therefore, should be thought of in terms of courses of action and inaction: as much an ongoing *process* of interaction as a discrete product or output of the political system, written down in legal texts.

Due to these and other conceptual ambiguities, the notion of 'environmental policy' is commonly defined in terms of the problems it addresses. Traditionally, environmental policies have been associated with controlling pollution, limiting natural habitat loss, managing waste, and reducing the impact of natural hazards (Connelly et al., 2012). These policies are normally implemented using regulatory tools although non-regulatory instruments such as taxes and voluntary agreements have become increasingly popular. However, environmental policy, in the broader sense described above, touches upon many more issues, such as food production, energy, transport, human health and safety, consumer choices, the survival of biological species, and even international security. In short, environmental policy addresses the complex, reciprocal links between modern society, the economy, and the environment, making it unique among policy areas (Ingram and Mann, 1983). The recognition that the increased pace, scale, and magnitude of contemporary human activity has pervasively affected the natural environment has been an important driver of environmental policy since the 1960s. Achieving a harmonious balance between society, the economy, and environmental systems is the main aim of *sustainable development*, the key concept for organizing environmental policy (Adger and Jordan, 2009).

The Main Characteristics of Environmental Policy

According to Albert Weale (1992: 5), the objective characteristics of a policy area have an important bearing upon the form of politics and policies that develop around it. This section explores the main characteristics of environmental policy.

Governing Public 'Goods' and 'Bads'

Environmental quality is a classic example of what economists term a nonexcludable or 'public' good. Thus the benefits of

environmental quality accrue to most if not all citizens (or, to put it another way, the costs of degradation are borne by society rather than just the polluter). This seemingly self-evident observation has important implications for the scope of environmental policy. First, large parts of the environment – the atmosphere and open oceans – are common property and nobody can be excluded from using them. While the benefits of resource use tend to be concentrated spatially, the costs or 'externalities' are typically dispersed across myriad individuals in the form of pollution (or economic 'bads'). Think of a smoking factory chimney pouring black soot onto surrounding houses. Here, the polluter has an economic incentive to continue consuming clean air knowing that the bads will be borne by the whole neighborhood. Second, while the financial cost of improving the environment often falls on one small section of society (e.g., the factory owner), the benefits are normally spread among myriad users. Polluters, who have a strong incentive to unite and fight to protect their pollution rights, have an advantage over those suffering the disbenefit. The latter are often too widely dispersed or individually suffer too little to mobilize effectively. Third, those fighting for the 'public' interest or the welfare of natural entities do not necessarily act automatically to achieve their individual or group self-interests. Mancur Olson (1965) argued that individuals have an incentive to 'free ride' by consuming environmental quality in the hope the cost will be borne by some other party or society as a whole. For example, polluters of a stream will continue to pollute hoping that a sufficient number of others desist from polluting it. However, the public good aspect of environmental problems also conditions attempts to abate environmental damage. Then, recreational users of our hypothetical river may well calculate that it is easier to free ride (i.e., wait for someone else to tackle the polluter and share in the benefits of clean water) than do something themselves. One reason why motorists continue to drive is that they cannot easily be excluded from the benefits (i.e., less air pollution and congestion) which are generated when some give up their vehicles and turn to public transport. This collective action problem is further developed in Garrett Hardin's (1968) famous parable of the 'tragedy of the commons.'

The Need for Regulation

Because environmental damage originates in otherwise socially legitimate activities like energy and food production, the state is often forced to police the consumption of public goods by limiting the damages that one section of society can impose upon others. Environmental policy is therefore inherently *regulatory* in nature, although regulation inevitably has distributive and redistributive consequences (Lowi, 1972). Lowi argued that whereas distributive policies (e.g., welfare programs) produce a close, dependence relationships between the state and interest groups, regulatory policies produce more conflictual politics as they force polluters to adopt costly abatement measures. However, not all interest groups enjoy the same level of influence or political power. Some routinely have more say than others in determining the content of policy by 'capturing' the regulatory process, although the influence of sectional groups is hotly disputed among political scientists (Dryzek and Dunleavy, 2009). Business groups do, however, enjoy a structural advantage over other groups – one which exacerbates the collective action problems suffered by environmental groups. According to Lindblom (1977), as *all* governments, whether Left or Right, need a healthy economy to provide jobs and welfare support, they have to adopt business-friendly measures without business having to take any observable action. Scientists too have their own independent policy objectives, and can be considered a kind of interest group. In some situations 'epistemic communities' (Haas, 1992) of scientific experts have decisively influenced policy by steering decision-makers.

Coping with Complexity and Uncertainty

Once human development exceeds the planetary capacity for self-repair, environmental assets can collapse and disappear, never to reappear. Admittedly, some environmental assets and functions are repairable or substitutable (individual trees can be replanted; water can be transported), but many others (a delicate wetland habitat for example) are not. Irreversibility complicates environmental policy making, differentiating it from more easily revisited areas such as tax or social affairs. Irreversibility implies that piecemeal policy responses should be substituted for ones that foresee problems and address them systematically. In practice, the environment often responds slowly to human interventions, and 'local' problems are sometimes only the observable tip of much more complicated and spatially dispersed processes. For instance, the pressure to pollute a local river with fertilizers may stem from a national policy to intensify agricultural production in order to compete in world markets. This raises the question of how decisions should be made at one level when consequences for other levels or sectors cannot be easily predicted. Should decision-makers err on the side of caution or proceed hoping that unfavorable consequences will not arise? This is one reason why the precautionary principle – the idea that policy makers should go 'beyond science' to address potential problems when the threat of harm is significant and potentially irreversible – has become important to policy debates in areas such as genetically modified food. But whenever environment impacts cannot be precisely determined, opportunities arise for vested interests to twist scientific ignorance to suit their own cause. Political systems also struggle to cope with genuine uncertainty; hence difficult environmental policy decisions are routinely handed to expert committees to deliberate upon, delaying still further the adoption of protective measures.

Governing through Time and Space

Although environmental quality is a classic example of a public good, the costs and benefits of policy do not necessarily accrue to the same citizens. Consequently equity and justice are central to most, if not all, environmental decisions. Those who suffer most from environmental damage are often the poorest and most vulnerable, while others who benefit have a vested interest in maintaining the status quo. The precise spatial and temporal distribution of costs and benefits produces alliances of self-interest which battle to determine

the scope and content of environmental policy. Constructing a bypass road around a congested village, for example, will greatly improve the welfare of its residents but at the cost of disruption to those in the path of the alternative route. Balancing these different needs becomes a central question of *politics* (or decision making) which is pervasive in all environmental policy making. Another problem facing policy makers is how to account for the well-being of future generations whose needs can only be anticipated. According to the most widely adopted interpretation of sustainable development, articulated in the UN Brundtland Report (WCED, 1987), policy makers should address the needs of current generations of humans without compromising the ability of future generations to meet theirs. Clearly, this raises thorny questions of justice and morality, to say nothing of due democratic process.

Governing across Political Borders

Unfortunately for policy makers, environmental problems show scant regard for political borders. Essentially, there are three types of international problems. The first arises when the legitimate activities of one state have international implications (e.g., global warming). Then there are global processes such as trade and international debt which have environmental implications. Finally, there is the problem of how to preserve resources such as rainforests and regional seas that reside in one part of the world but are valued by all. Olson's theory of collective action again illuminates the difficulties in achieving cooperation, this time between states rather than individuals. Regional arrangements, such as those pioneered in the European Union, provide the best evidence of successful international diplomacy (Jordan and Adelle, 2012). However, many disputes about responsibility for causing and remedying global environmental problems such as global warming hinge on a North-versus-South axis, though not exclusively so. In these circumstances, environmental policy relies upon regimes – i.e., international conventions and protocols – to build trust between participants. Many regimes have been negotiated since the 1970s (see Connelly et al., 2012). In some circumstances, they have been deliberately designed to offer 'selective incentives' (Olson, 1965) (i.e., technical and financial assistance) to incorporate reluctant countries, whose participation is nonetheless vital for success.

Integrating across Policy Areas

While the environment functions as an integrated whole, the bureaucracies managing it are typically fragmented in order to avoid administrative overlap and inefficiency. The mismatch between institutions and environmental problems has been an enduring problem for policy makers. The environment is probably unique in how it crosscuts traditional government sectors. In consequence, environmental policy making often produces furious conflicts between environmental ministries and more powerful parts of government representing finance, agriculture, transport, and so on. Accordingly, one of the great unmet challenges since the 1990s has been to *integrate* environmental concerns into these areas (Jordan and Lenschow, 2008). According to the UN Brundtland Report which set the terms of the international debate about sustainable development:

> Those responsible for managing natural resources and protecting the environment are institutionally separated from those responsible for managing the economy. The real world of interlocked economic and ecological systems will not change; the policies and institutions concerned must.
>
> WCED, 1987: p. 9

Often each department defends its own position while resisting a line that, while potentially beneficial to society as a whole, clashes with its 'own' interests. So even if the environment department wishes to push a greener line, it is likely to encounter fierce opposition from departments representing commercial interests (Jordan and Lenschow, 2008; Jordan and Schout, 2006).

Dealing with Nonhuman Entities

Environmental policy is probably the only policy domain which has the main goal of protecting nonhuman entities. Although most people support environmental protection the crucial question is to what extent should moral significance be ascribed to plants and animals, especially if it involves self-sacrifice and higher economic costs? There are essentially two ethical positions in an ongoing debate that underpins most environmental policy making: technocentric and ecocentric. Those who adopt a *technocentric* view regard nature as a source of resources to be exploited, whereas ecocentrists feel that the nonhuman world has interests and moral significance independent of human valuation – what economists term 'existence value.' While the technocentric view is hierarchical, manipulative, and managerial, the ecocentric view, by contrast, embraces community scale, natural rhythms, and a morality based on ecological principles. Business tends to adopt a more technocentric line, which conditions policy via the structural commitments outlined above. The 'green' case put by environmentalists is much more ecocentric. If anything, after decades of debate interests have converged on the center of this continuum of worldviews. Indeed sustainability – at least as defined by the Brundtland Report – has proven to be politically palatable precisely because it maintains that the world *can* grow its way out of ecological crisis.

Despite this convergence, contemporary environmental conflicts are still driven by the clash between worldviews. We see this in debates about genetically modified organisms and foods. The large chemical companies are largely technocentric, stressing the benefits of manipulating natural processes for human benefit. Their faith in the power of science to achieve a transition to modified crops in a predictable and controlled manner sets them apart from environmental groups who stress the fundamental uncertainties involved – which, if not respected, may result in irreparable damage to biodiversity. Informing this debate is the question of whether humans have the moral *right* to 'play God' with nature. The technocentric has few moral qualms about this, whereas the committed ecocentric does.

The Main Elements of Environmental Policy

A Brief History of Environmental Policy

As policy areas go, environmental policy is relatively youthful, dating back only as far as the early 1970s. Environmental provisions existed before that time but were relatively fragmented, aimed mainly at securing human health, and developed in a reactive and ad hoc manner. Environmental policy gained its impetus from the great upsurge of public interest in environmental matters during the late 1960s. Landmark books such as Rachel Carson's *Silent Spring* and *Blueprint For Survival* by the Ecologist magazine suggested a global crisis was imminent – a message which found receptive audiences in the US and Western Europe. The first evocative images transmitted from the space of Earth floating precariously in the inky blackness of deep space helped catalyze and intensify public concern. The unprecedented magnitude of human-induced environmental change throughout the twentieth century had also produced an objective reduction in environmental quality. Many national environmental ministries and the United Nations Environment Programme (UNEP) date from this period (Jänicke and Weidner, 1997). The policies adopted were mainly regulatory, specifying processes and emission standards. In those countries with less developed administrative systems, this early period saw the development of basic standards for air, water, waste, and noise pollution.

Another feature of this era was the growing internationalization of policy noted above. Early environmental regimes were often regionally based, such as UNEP's 1976 Barcelona Convention for protecting the Mediterranean (Connelly et al., 2012). More globally encompassing regimes then appeared in rapid succession to counter inter alia marine pollution, acid rain, ozone depletion, and hazardous waste shipments. During the 1990s, a new phase of international policy making emerged from agreements made at the United Nations Rio Earth Summit (1992). Most significantly, a UN Convention on Biological Diversity and a Framework Convention on Climate Change were agreed, the latter subsequently supported by the Kyoto Protocol (1997) aimed at limiting global climate emissions.

As sustainable development became the main discourse of environmental policy, the focus shifted from dealing with environmental damage – pollution, waste, and habitat loss – to addressing the underlying causes. The trend toward redesigning institutions and organizations to tackle problems at root is often labeled *ecological modernization* (Mol et al., 2009) – a political philosophy reflecting a technocentric view of sustainability. This involves moving from setting end of pipe standards to tackling social processes that generate emissions. The shift toward more diffuse and less interventionist policy tools in part reflects political pressure from business for more flexible and economically efficient alternatives to regulation. The 'new' tools in the armory of environmental policy now include (Wurzel et al., 2013):

- *Ecolabels*: These signal the environmental impact of products and processes to purchasers and investors. Eco-auditing and lifecycle assessments help to rate the eco-friendliness of products and processes by measuring their overall environmental impact.
- *Eco-tax reform*: Working with the grain of markets, taxes are designed to penalize economic 'bads' and, through hypothecation mechanisms, reward 'goods' such as labor. For example, in 1997, the UK adopted a tax on the waste deposited in landfills, the revenues from which are diverted to environmental projects.
- *Environmental assessment*: The aim here is to implement the principle of environmental policy integration. The US National Environmental Policy Act (1969) formally required an environmental assessment of individual development projects such as power stations. Although project level environmental impact assessment has since become accepted worldwide, more recent approaches to integration involve the 'strategic' assessment of higher level policies, plans, and programs.
- *Integrated product policies*: Using various devices to 'green' the production of goods from their design and production, through to final disposal. A good example is the EU's packaging waste directive which aims to reduce the overall quantity of waste material.
- *Strict liability regimes*: By making developers legally responsible for remedying the damage they cause, these encourage a more consistently preventative approach to environmental issues.

To these instruments, we might add resource-use policies that seek to transition economies toward greater sustainability. For example, the Japanese government promotes a '3R' policy of reduce, reuse, and recycle through multiple policy instruments, with an emphasis on avoiding environmental problems at source.

Climate change has emerged more forcefully as a policy issue at multiple levels since the late 1990s. A global 'leader' in this respect is the European Union, which has introduced a suite of innovative policies for climate mitigation, most notably an emissions trading system (Jordan and Adelle, 2012). Policy responses from national governments worldwide have been more mixed. Pioneering national legislative measures such as the UK Climate Change Act, that establishes legally binding annual emissions reduction targets, can be contrasted with the relative inaction of states such as the US where climate change has effectively become a policy nonissue at the federal level. This difference in national perceptions has slowed down the ongoing development of the Kyoto Protocol.

National Policy Styles

When studying these political responses to environmentalism, states and supranational organizations have evidently followed different paths reflecting their history and political mandate. The current allocation of responsibilities generally respects the principle of subsidiarity, whereby problems should be addressed at the lowest appropriate level. Global problems are addressed at the global level (i.e., the United Nations), regional problems at the regional level (e.g., the EU), and the rest at the national or subnational level. Environmental pressure groups and scientists have learned to work at all levels, removing environmental policy from the domain of purely state-guided action and placing it within a multilevel system of governance (see Jordan and Adelle, 2012).

In spite of internationalization, individual states continue to pursue different approaches, reflecting their individual geographical circumstances, political and constitutional makeup, cultural orientation, and economic conditions (Jänicke and Weidner, 1997). The interaction of these circumstances gives each country a distinct national 'policy style' (Richardson, 1982) which preexists its response to new political issues. Since the 1980s, Germany has adopted an anticipatory approach based on explicit standards and programs which are designed to 'force' the development of clean technologies. German approaches contrast starkly with the open, conflictual, and rule-bound nature of the US political system where the importance of the free market partly explains why the US has gone furthest in experimenting with economic instruments such as tradable permits.

The Policy Process

The policy process is normally conceptualized as sequential parts or stages. These are (1) problem emergence, (2) agenda setting, (3) consideration of policy options, (3) decision-making, (5) implementation, and (6) evaluation (Jordan and Adelle, 2012). According to this 'textbook' view of policy, environmental groups must first get a particular problem on the agenda for discussion and, if possible, consideration by policy makers. Policy makers then select the best course of action based on specialist advice, make the policy, then hand it to administrators for implementation. This stage-based view emphasizes that policy is a process involving many different parts of government. But in practice, policy issues are interconnected, policy makers fumble around for solutions in the context of great uncertainty and many internal and external constraints. Past practice often has a determining effect on how new issues are processed. To counter such uncertainty, increasing use is being made of scientific evidence to guide environmental policy development. Multiple policy 'tools' such as expert reports, consultation, and cost–benefit analysis are employed by governments to provide such decision support (Jordan et al., 2015).

A multitude of different models and theories have been developed to explain policy making (see Hill, 2009; Dryzek and Dunleavy, 2009). One strand is pluralist in its assumptions. *Pluralists* believe that political power is widely, although unevenly, spread throughout society. Although powerful groups exist in particular policy sectors, no single group is continuously capable of shaping the entire policy process. In policy terms, pluralists assume that agenda setting is open and competitive, with the government acting as an honest broker. Once adopted, though, policies must still be steered through the implementation process. Because of the competitiveness of the policy process, outcomes are unpredictable.

Neo-pluralists meanwhile argue that business occupies a 'privileged' position compared to other groups. Instruments like the mass media help to structure environmental politics by removing 'grand majority' issues concerning the fundamentals of the political order from the agenda, leaving citizens to debate residual 'secondary' concerns (Lindblom, 1977). Whereas pluralists assume that grievances are openly debated, neo-pluralists argue that they are organized out of politics by institutional rules and routines. Policy making occurs within small and stable groups of actors (or networks) clustered around particular government departments. Policy outcomes then generally reflect business preferences rather than those of environmentalists, who find it difficult to access decision-makers or are deterred from even trying.

Finally, *structuralists* believe that the state is under powerful structural pressure to nurture economic growth regardless of the environmental implications (Benton, 1996). On this view, most environmental controls are nothing more than a sham, introduced to pacify critics and keep the conflict between economic classes to manageable levels. For Marxist scholars, environmental problems are rooted in the unequal distribution of resources within society. There can be no lasting solution to such problems until the social system is structured more equally.

What Drives Environmental Policy and Politics?

Having defined the policy process and the main tools used, the obvious question is what actually drives it forward? Where do the inputs, or political issues, come from which create the agenda for environmental policy making? The answer is partly conditioned by which of the three models described above is used. These differences notwithstanding, ultimately the source of all green ideas is environmentalism, a broad social movement which emerged in the late 1960s reflecting a growing public demand for a better environment (Doyle and McEachern, 2007). Environmentalism, the roots of which stretch far back into the past, is reflected in a long-term shift in public opinion toward 'quality-of-life' issues, and an increasingly active and diverse range of 'green' pressure groups. Environmental concern has exhibited a broadly cyclical pattern over time, with particularly pronounced peaks in the late 1960s and 1980s. Closer scrutiny reveals that these short-term pulses coincided with periods of economic growth and social introspection.

Downs (1972) produced a simple model that captures these dynamics. He argues that environmental issues evolve slowly in a culture of scientific enquiry and suppression. This is usually a fairly secretive phase during which gloomy prognoses are carefully explored away from the glare of media concern. The second phase in his 'issue attention cycle' was one of media attention and anxiety. Since the 1960s, the environmental movement has benefited enormously from successive crises including Seveso, Three Mile Island, Chernobyl, and Fukushima. The third phase is the reactive process in which political institutions cope and adapt. During the fourth phase the costs of protection sink in, and demands grow for regulation to be dismantled to allow enterprise and economic growth. The so-called contrarians begin to question the accuracy of environmental claims (North, 1995). However, downgrading environmental protection succeeds only in propagating the next crisis, and so the cycle begins again.

Environmental policy has generally followed these oscillations. The early 1970s and the period 1989–92 were thus both marked by important institutional innovations. The respective focal points were international conferences at Stockholm (1972) and Rio (1992). In turn, these triggered public and private awareness, much greater financial investment, and

corporate planning. Significantly, however, political concern has not settled back to the original level, but has oscillated upward around a steadily rising curve. Ratchet-like, the laws and public institutions established to protect the environment during each upsurge, have helped fix environmental values, providing a new platform for subsequent policy making.

Conclusion

Environmental policy has reached a critical stage in its relatively short life. In the 1970s, governments across the world responded to the rise of environmental values by establishing agencies and laws to address specific issues as and when they arose. Policy makers regarded environmental problems as essentially technical in nature; policies were overwhelmingly reactive rather than proactive, and sectoral rather than integrated. But starting in the 1980s, the discourse surrounding environmental issues began to center on the concept of sustainable development (Adger and Jordan, 2009). Significantly, sustainable development is not simply economic advancement 'greened' by bolting on cleaner technologies and reducing waste, but a *process* of human development which simultaneously satisfies social, economic, and environmental goals, within and between different generations.

The emerging orthodoxy in the twenty-first century is that environmental thinking must be integrated into areas of policy traditionally regarded as 'non'-environmental such as agriculture, transport, and energy (Jordan and Lenschow, 2008). One important lesson learned in the last 40 years is that the fundamental driving forces of unsustainable behavior reside in these sectors. Unless these forces are addressed at their roots, environmental protection will always be reactive, piecemeal, and ad hoc. Politically speaking, the transition to environmental policy integration represents a new and potentially difficult stage in the continuing development of environmental policy. However, the political stakes are high because by greening other policy areas, environmental policy is in effect progressively undermining its own, hard-won independent existence. The irony is that if the environment is reflected fully in every aspect of policy making, there will be no need for an identifiable 'environmental' policy. Whether this outcome is ever likely to occur in practice, however, remains open to discussion.

See also: Climate Change: Economics; Sustainable Development: An Economic Perspective; Sustainable Development.

Bibliography

Adger, W.N., Jordan, A., 2009. Governing Sustainability. Cambridge University Press, Cambridge.
Benton, T. (Ed.), 1996. The Greening of Marxism. Guilford Press, London.
Connelly, J., Smith, G., Benson, D., Saunders, C., 2012. Politics and the Environment. Routledge, London.
Doyle, T., McEachern, D., 2007. Environment and Politics, third ed. Routledge, London.
Downs, A., 1972. Up and down with ecology: the issue attention cycle. Public Interest 28, 38–50.
Dryzek, J.S., Dunleavy, P., 2009. Theories of the Democratic State. Palgrave, Basingstoke.
Haas, P., 1992. Epistemic communities and international policy co-ordination. International Organization 46 (1), 1–35.
Hardin, G., 1968. The tragedy of the commons. Science 162 (3859), 1243–1248.
Hill, M., 2009. The Public Policy Process, fifth ed. Pearson, Harlow.
Ingram, H., Mann, D., 1983. Environmental Protection policy. In: Nagel, S. (Ed.), Encyclopaedia of Public Policy. Dekkar Publishers, New York.
Jänicke, M., Weidner, H. (Eds.), 1997. National Environmental Policies. Springer, Berlin.
Jordan, A.J., Schout, A., 2006. The Coordination of the European Union. Oxford University Press, Oxford.
Jordan, A.J., Lenschow, A. (Eds.), 2008. Innovation in Environmental Policy? Integrating the Environment for Sustainability. Edward Elgar, Cheltenham.
Jordan, A.J., Adelle, C. (Eds.), 2012. Environmental Policy in the EU, third ed. Routledge, London.
Jordan, A.J., Turnpenny, J., Benson, D., 2015. The tools of policy formulation: an introduction. In: Jordan, A.J., Turnpenny, J. (Eds.), The Role of Decision-support Tools in Policy Practice. Edward Elgar, Cheltenham.
Lindblom, C., 1977. Politics and Markets. Basic Books, New York.
Lowi, T., 1972. Four systems of policy, politics and choice. Public Administration Review 32, 298–310.
Mol, A.P.J., Sonnenfeld, D.A., Spaargaren, G. (Eds.), 2009. The Ecological Modernisation Reader. Routledge, London.
North, R., 1995. Life on a Modern Planet. Manchester University Press, Manchester.
Olson, M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.
Richardson, J. (Ed.), 1982. Policy Styles in Europe. Allen and Unwin, Europe.
WCED (World Commission on Environment and Development), 1987. Our Common Future. Oxford University Press, Oxford.
Weale, A., 1992. The New Politics of Pollution Control. Manchester University Press, Manchester.
Wurzel, R.K., Zito, A.R., Jordan, A.J., 2013. Environmental Governance in Europe. Edward Elgar, Cheltenham.

Environmental Risk

Ortwin Renn, Stuttgart Center for Interdisciplinary Risk and Innovation Studies at Stuttgart University (ZIRIUS), Stuttgart University, Stuttgart, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Liverman, volume 7, pp. 4655–4659, © 2001, Elsevier Ltd.

Abstract

It is important to distinguish between ecological risk assessment and risk evaluation. Assessment includes all systematic approaches to understand the consequences of human interventions into the natural ecosystem and to calculate the risks for potential damages. Risk evaluation includes the process of judging the acceptability of these interventions on the basis of a comparison between the expected benefits and risks. Both tasks require sound knowledge and ethical judgments. Judgments about acceptability are contingent on the priority of human values and social development goals. This is the reason why such judgments need a democratic legitimization based on an inclusive judgment and decision process incorporating representatives from science, politics, economics, and civil society.

Introduction

Environmental risks refer to the possibility or probability of negative impacts associated with human interventions into nature (Dobler et al., 2014: 2). Interventions are normally taken for resource depletion or using the natural environment as a sink for waste in all its forms. Risks stemming from these interventions include damages to human health, environmental quality, ecosystem functioning, changes in landscape, and loss of biodiversity.

Most interventions involving environmental risks to humans or nature are made for some reason: the actors who make such interventions want to secure goods or services to consumers, for example, to ensure long-term jobs and adequate incomes, to use natural resources for products and services, or to use nature for recycling waste materials from production and consumption that are no longer needed. None of these are done for reasons of brotherly love, but to maintain social interests. The list of human activities that pose risks onto others perpetrated for existential or economic reasons could be carried on into infinity. Human existence is bound to taking opportunities and risks (Radandt et al., 2008: 212).

With today's settlement densities and the predominantly industrial way of life, the human race is dependent on the technical remodeling of nature. Without doubt, human societies need interventions into the natural environment not only for survival, but also for the well-being of the innumerable people, goods, and services that reduce the stock of natural resources. With regard to the question of the acceptable risk level that human interventions in nature can cause, the question cannot be about 'whether' but – even better – about 'how much,' because it is an anthropological necessity to adapt and shape existing nature to human needs. For example, the philosopher Klaus Michael Meyer-Abich sees the situation as follows: " … we humans are not there to leave the world as though we had never been there. As with all other life forms, it is also part of our nature and our lives to bring about changes in the world. Of course, this does not legitimize the destructive ways of life that we have fallen into. But only when we basically approve of the changes in the world can we turn to the decisive question of which changes are appropriate for human existence and which are not" (Meyer-Abich, 1996).

It is therefore unavoidable to make a sensible judgment of the balance between necessary interventions into the environment and the risks posed by these interventions to human health and environmental quality. However, at the same time, it must be kept in mind that these interventions often inflict irreversible damage on the environment and destroy possible future-usage potentials for future generations. Above and beyond this, for the human race, nature is a cradle of social, cultural, aesthetic, and religious values, the infringement of which, in turn, has a major influence on people's well-being (Pidgeon et al., 2010: 95). On both sides of the equation, there are therefore important goods that have to be appreciated when interventions in nature occur. But what form should such an appreciation take?

If the pros and cons of the intervention in nature have to be weighed against each other, criteria are needed that can be used as yardsticks. Who can and may draw up such criteria, according to which standards should the interventions be assessed and how can the various evaluative options for action be compared with each other for each criterion?

Taking risks always involves two major components: an assessment of what we can expect from an intervention into the environment (be it the use of resources or the use of environments as a sink for our waste). This is the risk- and benefit-assessment side of the risk analysis. Next, we need to decide whether the assessed consequences are culturally or socially desirable. While the estimate of consequences broadly falls in the domain of scientific research and expertise, with uncertainties and ambiguities in particular having to be taken into account (IRGC, 2005; Renn et al., 2011), the question about the foundations for evaluating various options for action and about drawing up standards guiding action is a central function of ethics (Taylor, 1986). Both components, the assessment and the evaluation part of risk analysis will be explained in more detail in the sections below.

Generic Principles for Human Interventions into the Environment

Since contemporary society and the generations to come certainly use, or will use, natural resources, the question of legitimizing interventions into the natural environment is

mainly a question of how much of what nature needs to be conserved and what can be used for human purposes. Nature itself cannot show society what it is essential to conserve and how much nature can be traded for valuable commodities. Humans alone are responsible for a decision and the resulting conflicts between competing objectives. Appreciation and negotiation processes are therefore the core of the considerations about the ethical justification of rules for interventions. In this debate about the acceptability of environmental risk, it is advisable to make a distinction between inviolate principles and criteria, which allow trade-offs between them (Darwall, 2014: 14). Let us start with the principles: It follows on from the basic principle of conserving human life that all human interventions that threaten the ability of the human race as a whole, or a significant number of individuals alive today or in the future, to exist should be categorically prohibited. This refers to intervention threats to the systemic functions of the biosphere. Such threats are one of the guiding principles that must not be exceeded under any circumstances, even if this excess were to be associated with high benefits. In the language of ethics, this is a categorical principle; in the language of economics, a good that is not capable of being traded (Moldan et al., 2012: 7f.). The set of categorical prohibitions should, however, be used very sparingly because plausible trade-offs can be thought up for most principles, the partial exceeding of which appears intuitively. In the case of threats to existence, however, the categorical rejection of the behavior that leads to this is obvious.

But what does the adoption of categorical principles specifically mean for the political molding of environmental protection? In the past, a number of authors have tried to specify the minimum requirements for an ethically responsible moral system with respect to biosphere use. These so-called 'safe minimum standards' specify thresholds for the open-ended measurement scale of the consequences of human interventions that may not be exceeded even if there is a prospect of great benefits (Randall, 1988; Randall and Farmer, 1995). In order to be able to specify these thresholds in more detail, the breakdown into three levels proposed by the German Scientific Council for Global Environmental Change is helpful (WBGU, 2000: 325f.). These levels are:

- the global bio–geochemical cycles in which the biosphere is involved as one of the causes, modulator, or 'beneficiary';
- the diversity of ecosystems and landscapes that have key functions as bearers of diversity in the biosphere; and
- the genetic diversity and the species diversity that are both 'the modeling clay of evolution' and basic elements of ecosystem functions and dynamics.

When the first level is concerned, in which the functioning of the global ecosystem is at stake, categorical principles are obviously necessary and sensible, provided that no one wants to shake the primary principle of the permanent preservation of the human race. *Accordingly, all interventions in which important substance or energy cycles are significantly influenced at a global level and where globally effective negative impacts are to be expected are categorically prohibited*. Usually, no stringently causal evidence of the harmful nature of globally relevant information is needed; justified suspicion of such harmfulness should suffice.

On the second level, the protection of ecosystems and landscapes, it is much more difficult to draw up categorical rules. Initially, it is obvious that all interventions in landscapes in which the global functions mentioned on the first level are endangered must be avoided. Above and beyond this, it is wise from a precautionary point of view to maintain as much ecosystem diversity as possible in order to keep the degree of vulnerability to the unforeseen or even unforeseeable consequences of anthropogenic and nonanthropogenic interventions as low as possible. Even though it is difficult to derive findings for human behavior from observations of evolution, the empirically proven statement "he who places everything on one card, always loses in the long run" seems to demonstrate a universally valid insight into the functioning of systemically organized interactions (Renn, 2014: 502ff.). For this reason, the conservation of the natural diversity of ecosystems and landscape forms is a categorical principle, whereas the depth of intervention allowed should be specified on the basis of principles and standards capable of compensation.

The same can be said for the third level, genetic and species protection. Here too, initially the causal chain should be laid down: species conservation, landscape conservation, maintaining global functions. Wherever this chain is unbroken, a categorical order of conservation should apply. These species could be termed *primary key species*. This includes such species that are not only essential for the specific landscape type in which they occur, but also for the global cycles above and beyond this specific landscape type – thanks to their special position in the ecosystem. Probably, it will not be possible to organize all species under this functional contribution to the surrounding ecosystem, but we could also think of groups of species, for example, humus-forming bacteria. In second place, there are the species that characterize certain ecosystems or landscapes. Here they are referred to as *secondary key species*. They, too, are under special protection that is not necessarily under categorical reservations. Their function value, however, is worthy of special attention. Below these two types of species, there are the remaining species that perform ecosystem functions to a greater or lesser extent. Generally, here, too, as with the issue of ecosystem and landscape protection, the conservation of diversity as a strategy of 'reinsurance' against ignorance, global risks, and unforeseeable surprises is recommended (WBGU, 1999).

It remains to be said that from a systemic point of view, a categorical ban has to apply to all human interventions where global closed loops are demonstrably at risk. Above and beyond this, it makes sense to recognize the conservation of landscape variety (also of ecosystem diversity within landscapes) and of genetic variety and species diversity as basic principles, without being able to make categorical judgments about individual landscape or species types as a result.

Compensatory Principles and Standards

In order to evaluate partial infringements of compensatory principles or standards, which are referred to in the issue of environmental protection, we need rules for decision-making that facilitate the balancing process necessary to resolve compensatory conflicts. In the current debate about rules for

using the environment and nature, it is mainly teleological valuation methods that are proposed (Ott, 1993).

These methods are aimed at:

- estimating the possible consequences of various options for action at all dimensions relevant to potentially affected people,
- recording the infringements or fulfillments of these expected consequences in the light of the guiding standards and principles, and then
- aggregating them according to an internal key so that they can be weighed up in a balanced way.

On the positive side of the equation, there are the economic benefits of an intervention and the cultural values created by use, for example, in the form of income, subsistence (self-sufficiency), or an aesthetically attractive landscape (parks, ornamental gardens, etc.); on the negative side, there are the destruction of current- or future-usage potentials, the loss of unknown natural resources that may be needed in the future, and the violation of aesthetic, cultural, or religious attributes associated with the environment and nature.

There are, therefore, related categories on both sides of the equation: current versus possible uses in the future, development potentials of current uses versus option values for future use, shaping the environment by use versus impairments to the environment as a result of alternative use, etc. With the same or similar categories on the credit and debit side of the balance sheet, the decision is easy when there is one option that performs better or worse than all the other options for all categories. Although such a *dominant* (the best for all categories) or *subdominant option* (the worst for all categories) is rare in reality, there are examples of dominant or subdominant solutions.

Thus, for example, the overfelling of the forests of Kalimantan on the island of Borneo in Indonesia can be classed as a subdominant option, since the short-term benefit, even with extremely high discount rates, is in no proportion to the long-term losses of benefits associated with a barren area covered in *Imperata* grass. The recultivation of a barren area of this kind requires sums many times the income from the sale of the wood, including interest. Apparently there are no cultural, aesthetic, or religious reasons for conversion of primary or secondary woodland into grassland. This means that the option of deforestation should be classed as of less value than alternative options for all criteria, including economic and social criteria (Radandt et al., 2008: 231).

But if we disregard the dominant or subdominant solutions, an appreciation between options that violate or fulfill compensatory standards and principles depends on two preconditions: best possible knowledge of the consequences (what happens if I choose option A instead of option B?) and a transparent, consistent rationale for weighing up these consequences as part of a legitimate political decision process (are the foreseeable consequences of option A more desirable or bearable than the consequences of option B?) (Gregory et al., 2012: 17).

Knowledge

Adequate knowledge of the consequences is needed in order to reveal the systemic connections between resource use, ecosystem reactions to human interventions, and sociocultural context (Becker, 1999: 286). This requires interdisciplinary and transdisciplinary research and cooperation. The task of applied ecological research, for example, is to show the consequences of human intervention in the natural environment and how ecosystems are burdened by different interventions and practices. The economic approach provides a benefit-oriented valuation of natural and artificial resources within the context of production and consumption, as well as a valuation of transformation processes according to the criterion of efficiency. Cultural and social sciences examine the feedback effects between use, social development, and cultural self-perception. They illustrate the dynamic interactions between usage forms, sociocultural lifestyles, and control forms.

Interdisciplinary, problem-oriented, and system-related research contribute jointly to forming a basic stock of findings and insights about functional links in the relationship between human interventions and the environment, and also in developing constructive proposals as to how the basic question of an ethically justified use of the natural environment can be answered in agreement with the actors concerned.

Accordingly, in order to ensure sufficient environmental protection, scientific research, but especially transdisciplinary system research at the interface between natural sciences and social sciences is essential. Bringing together the results of interdisciplinary research, the policy-relevant choice of knowledge banks and balanced interpretation in an environment of complexity, uncertainty, and ambiguity are difficult tasks that primarily have to be performed by the science system itself (Klinke and Renn, 2012).

Valuation of Consequences

But knowledge alone does not suffice. In order to be able to act effectively and efficiently while observing ethical principles, it is necessary to shape the *appreciation process* between the various options for action *according to evaluative criteria* (Gowdy, 1999: 172ff.). To do this, first of all, it is necessary to identify the dimensions that should be used for a valuation. The discussion about the value dimensions to be used as a basis for valuation is one of the most popular subjects within environmental ethics. To apply these criteria in risk evaluation and to combine the knowledge aspects about expected consequences of different behavioral options with the ethical principles is the task of what we have called the risk governance (Renn, 2008: 47ff.).

The following section will present a risk governance framework for applying these principles into environmental policy making under risk. The main line of argument is that risk management requires an analytic-deliberative approach for dealing effectively and prudently with environmental risks.

A Risk Governance Approach to Environmental Risk Taking

Assessing potential consequences of human interventions and evaluating their desirability on the basis of subsequent knowledge and transparent valuation criteria are two of the central tasks of the risk governance process. Risk governance

deals with the identification, assessment, management, evaluation, and communication of risks in the context of plural values and distributed authority. It includes the totality of actors, rules, conventions, processes, and mechanisms and is concerned with how relevant risk information is collected, analyzed, and communicated, and how management decisions are taken (IRGC, 2005). Risk governance is both a descriptive (how are decisions made) as well as normative concept (how should decisions be made). In its application as a normative concept, it specifies the principles of good governance that include transparency, effectiveness and efficiency, accountability, strategic focus, sustainability, equity and fairness, respect for the rule of law, and the need for the chosen solution to be politically and legally feasible as well as ethically and publicly acceptable.

For dealing with complex environmental risks, the governance structure requires a set of resources associated with the prudent handling of complexity, uncertainty, and ambiguity. In 2005, the International Risk Governance Council (IRGC) suggested a process model for risk governance (IRGC, 2005; Klinke and Renn, 2012; Renn and Walker, 2008). This framework structures the risk governance process in four phases: pre-assessment, appraisal, characterization and evaluation, and risk management. Communication and stakeholder involvement were conceptualized as constant companions to all four phases of the risk governance cycle. Based on this framework and informed by many comments on the original framework, Klinke and Renn modified the original IRGC proposal (Klinke and Renn, 2012). The new framework consists of the steps: pre-estimation, interdisciplinary risk estimation, risk characterization, risk evaluation, and risk management. This is all related to the ability and capacity of risk governance institutions to use resources effectively (see **Figure 1**).

Appropriate resources include institutional and financial means as well as social capital (e.g., strong institutional mechanisms and configurations, transparent decision-making, allocation of decision-making authority, formal and informal networks that promote collective risk handling, education), technical resources (e.g., databases, computer soft- and hardware, etc.), and human resources (e.g., skills, knowledge, expertise, epistemic communities, etc.). Hence the adequate involvement of experts, stakeholders, and the public in the risk governance process is a crucial dimension to produce and convey adaptive and integrative capacity in risk governance institutions (Pelling et al., 2008).

Therefore, the plural values of a heterogeneous expert, stakeholder, and public preferences have to be incorporated in the risk governance process. But how can this be done given the wealth of competing values and preferences? Should we simply accept the results of opinion polls as the basis for making political decisions? Can we rely on risk perception results to judge the seriousness of pending risks? Or should we place all our faith in professional risk management?

If we turn to professional help to deal with plural-value input, economic theory might provide us an answer to this problem: If environmental goods are made individual and suitable for the market by means of property rights, the price that forms on the market ensures an appropriate valuation of the environmental good (Burgmann, 2005). Every user of this good can then weigh up whether he is willing to pay the price or would rather not use the good. With many environmental goods, however, this valuation has to be made by collective action, because the environmental good concerned is a collective or open access good. In this case, a process is needed that safeguards the valuation and justifies it to the collective. However, this valuation cannot be determined with the help of

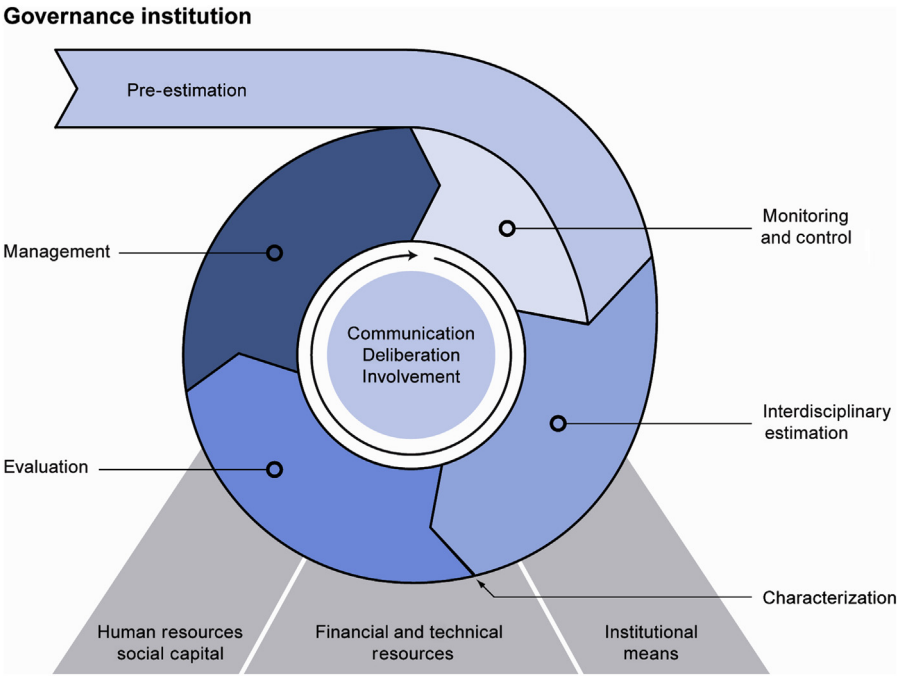

Figure 1 Adaptive and integrative risk governance model. Adapted from Klinke, A., Renn, O., 2012. Adaptive and integrative governance on risk and uncertainty. Journal of Risk Research 15 (3), 273–292.

survey results. Although surveys are needed to be able to estimate the breadth of preferences and people's willingness to pay, they are insufficient for a derivation of concrete decision-making criteria and yardsticks for evaluating the tolerability of risks to human health and the environment (Radandt et al., 2008: 235).

- First, the individual values are so widely scattered that there is little sense in finding an average value here.
- Second, the preferences expressed in surveys change much within a short time, whereas ethical valuations have to be valid for a long time.
- Third, preferences are frequently based on flawed knowledge or ad hoc assumptions both of which should not be decisive according to rational considerations.

What is needed, therefore, is a gradual process of assigning trade-offs in which existing empirical values are put into a coherent and logically consistent form. A report by the American Academy of Sciences on the subject of 'Understanding environmental risks' (US-National Research Council, 1996, 2008) comes to the conclusion that scientifically valid and ethically justified procedure for the collective valuation of options for risk handling can only be realized within the context of – what the authors coin – an analytic-deliberative process. *Analytic* means that the best scientific findings about the possible consequences and conditions of collective action are incorporated in the negotiations; *deliberative* means that rationally and ethically transparent criteria for making trade-offs are used and documented externally. Moreover, the authors consider that fair participation by all groups concerned is necessary to ensure that the different moral systems that can legitimately exist alongside each other should also be incorporated in the process.

A combination of analytic and deliberative processes can help explore the space of judgments and evaluations relating to complexity, uncertainty, and ambiguity (Renn, 2008: 284ff.). Most environmental agencies go through an environmental assessment process and provide opportunities for public review and comment. Many controversial environmental decisions become the focus of large analytical efforts, in which mathematical models are used to predict the environmental, economic, and health consequences of environmental management alternatives. Analysis should be seen as an indispensable complement to deliberative processes, regardless whether this analysis is sophisticated or not. Even simple questions need analytic input for making prudent decisions, especially in situations where there is controversy about the legitimacy of the human intervention into the environment.

Beyond the analytic input, the process of risk evaluation requires deliberation among and between different stakeholders. This second component 'deliberation' is even more difficult to put into practice than dealing with plural assessment results (Gregory et al., 2012: 24f.). Deliberation requires to (Webler, 1995):

- seek a consensus on the procedure that all participants want to employ in order to derive the final decision or compromise, such as voting, sorting of positions, consensual decision-making, or the involvement of a mediator or arbitrator;
- articulate and criticize factual claims on the basis of the 'state of the art' of scientific knowledge and other forms of problem-adequate knowledge (in the case of dissent, all relevant camps have the right to be represented);
- interpret factual evidence in accordance with the laws of formal logic and analytical reasoning;
- disclose their relevant values and preferences, thus avoiding hidden agendas and strategic game playing;
- process data, arguments, and evaluations in a structured format (e.g., a decision-analytic procedure) so that norms of procedural rationality are met and transparency can be created.

What needs to be deliberated? First, deliberative processes are needed to define the role and relevance of systematic and anecdotal knowledge for making far-reaching choices. Second, deliberation is needed to find the most appropriate way to deal with uncertainty in environmental decision-making and to set efficient and fair trade-offs between potential over- and underprotection. Third, deliberation needs to address the wider concerns of the affected groups and the public at large.

Why can one expect that deliberative processes are better suited to deal with environmental challenges than using expert judgment, political majority votes, or relying on public survey data (Renn, 2008: 306ff.)?

- Deliberation can produce common understanding of the issues or the problems based on the joint-learning experience of the participants with respect to systematic and anecdotal knowledge.
- Deliberation can produce a common understanding of each party's position and argumentation and thus assist in a mental reconstruction of each actor's argumentation. The main driver for gaining mutual understanding is empathy. The theory of communicative action provides further insights in how to mobilize empathy and how to use the mechanisms of empathy and normative reasoning to explore and generate common moral grounds (Webler, 1995).
- Deliberation can produce new options and novel solutions to a problem. This creative process can either be mobilized by finding win–win solutions or by discovering identical moral grounds on which new options can grow.
- Deliberation has the potential to show and document the full scope of ambiguity associated with environmental problems. Deliberation helps to make a society aware of the options, interpretations, and potential actions that are connected with the issue under investigation (Wynne, 1992). Each position within a deliberative discourse can only survive the crossfire of arguments and counterarguments, if it demonstrates internal consistency, compatibility with the legitimate range of knowledge claims, and correspondence with the widely accepted norms and values of society. Deliberation clarifies the problem, makes people aware of framing effects, and determines the limits of what could be called reasonable within the plurality of interpretations.
- Deliberations can also produce agreements. The minimal agreement may be a consensus about dissent. If all arguments are exchanged, participants know why they disagree.

They may not be convinced that the arguments of the other side are true or morally strong enough to change their own position; but they understand the reasons why the opponents came to their conclusion. In the end, the deliberative process produces several consistent and – in their own domain – optimized positions that can be offered as package options to legal decision-makers or the public. Once these options have been subjected to public discourse and debate, political bodies, such as agencies or parliaments can make the final selection in accordance with the legitimate rules and institutional arrangements such as majority vote or executive order. Final selections could also be performed by popular vote or referendum.

- Deliberation may result in consensus. Often deliberative processes are used synonymously with consensus-seeking activities (Coglianese, 1997). This is a major misunderstanding. Consensus is a possible outcome of deliberation, but not a mandatory requirement. If all participants find a new option that they all value more than the one that they preferred when entering the deliberation, a 'true' consensus is reached (Radandt et al., 2008: 245). It is clear that finding such a consensus is the exception rather than the rule. Consensus is either based on a win–win solution or a solution that serves the 'common good' and each participant's interests and values better than any other solution. Less stringent is the requirement of a tolerated consensus. Such a consensus rests on the recognition that the selected decision option might serve the 'common good' best, but on the expense of some interest violations or additional costs. In a tolerated consensus, some participants voluntarily accept personal or group-specific losses in exchange for providing benefits to all of society.

In summary, many desirable products and accomplishments are associated with deliberation (Chess et al., 1998). Depending on the structure of the discourse and the underlying rationale, deliberative processes can:

- enhance understanding;
- generate new options;
- decrease hostility and aggressive attitudes among the participants;
- explore new problem framing;
- enlighten legal policy-makers;
- produce competent, fair, and optimized solution packages; and
- facilitate consensus, tolerated consensus, and compromise.

In a situation of major value conflicts, the deliberation process may involve soliciting a diverse set of viewpoints, and judgments need to be made on what sources of information are viewed as responsible and reliable. Publication in scientific journals and peer review from scientists outside the government agency are the two most popular methods by which managers or organizers of deliberative processes try to limit what will be considered as acceptable evidence. Other methods are to reach a consensus among the participants up front, which expertise should be included in the deliberation or to appoint representatives of opposing science camps to explain their differences in public.

Conclusions

The objective of this article was to address and characterize the main components of assessing, evaluating, and managing environmental risks. Effective environmental risk governance needs a structure that assures the integration of technical expertise, regulatory requirements, and public values. These different inputs should be combined in such a fashion that they contribute to the best knowledge available in the field and to the plurality of values that govern the judgment of desirability for oneself and society as a whole. It does not make sense to replace technical expertise with vague public perceptions, nor is it justified to have the experts insert their own value judgments into what ought to be a democratic process.

An analytic-deliberative model of risk governance can provide assistance in problem structuring, in dealing with complex scientific issues and uncertainty, and in helping society to understand disagreements and ambiguity with respect to values and preferences. For supporting such an analytic-deliberative process, the IRGC risk framework can be used for environmental assessment and risk handling. It is designed to explore difficult issues, to focus debate and further analysis on the factors most important to the decision, and to provide for increased transparency and more effective exchange of information and opinions among the process participants.

The full power of the analytic-deliberative approach to risk governance lies in the combination of risk assessment through analytic methods and risk evaluation and management through deliberative procedures. Experienced analysts may be needed to guide the implementation of the analytical tools for aiding decisions. Skilled communicators and facilitators may be needed to design and conduct the deliberative processes as a means to provide a fair and comprehensive representation of societal values and preferences. Integrating analysis and deliberation can help to generate and sustain transparency and trust.

See also: Ecological Economics; Environmental Justice in the United States; Environmental Sociology; Participatory Action Research in Environmental and Ecological Studies.

Bibliography

Becker, E., 1999. Fostering transdisciplinary research into sustainability in an age of globalization: a short political epilogue. In: Becker, E., Jahn, T. (Eds.), Sustainability and the Social Sciences. Zen Books, London, pp. 284–289.

Burgmann, M., 2005. Risks and Decisions for Conservation and Environmental Management. Cambdrige University Press, Cambridge, US.

Chess, C., Dietz, T., Shannon, M., 1998. Who should deliberate when? Human Ecology Review 5 (1), 60–68.

Coglianese, C., 1997. Assessing consensus; the promise and performance of negotiated rule making. Duke Law Journal 46 (1), 1255–1333.

Darwall, S.L., 2014. Theories of ethics. In: Cohen, A.J., Heath Wellman, C. (Eds.), Contemporary Debates in Applied Ethics. Wiley Blackwell, Chichester, pp. 11–32.

Dobler, M., Lajili, K., Zéghal, D., 2014. Environmental performance, environmental risk and risk management. Business Strategy and the Environment 23 (1), 1–17.

Gregory, R., Failing, L., Harstone, M., Long, G., McDaniels, T., Ohlson, D., 2012. Structured Decision Making. A Practical Guide to Environmental Management Choices. Wiley Blackwell, Chichester.

Gowdy, J., 1999. Economic concepts of sustainability: relocating economic activity within society and environment. In: Becker, E., Jahn, T. (Eds.), Sustainability and the Social Sciences. Zen Books, London, pp. 162–181.

IRGC, International Risk Governance Council, 2005. Risk Governance: Towards an Integrative Approach, White Paper No 1. IRGC, Geneva.

Klinke, A., Renn, O., 2012. Adaptive and integrative governance on risk and uncertainty. Journal of Risk Research 15 (3), 273–292.

Meyer-Abich, K.M., 1996. Sustainable development? Wie nicht nur die Menschen eine 'dauerhafte' Entwicklung überdauern können. In: Clausthal, Forum (Ed.), Sustainable Development, vol. 1. TU Clausthal, pp. 10–21.

Moldan, B., Janoušková, S., Hák, T., 2012. How to understand and measure environmental sustainability: indicators and targets. Ecological Indicators 17, 4–13.

Ott, K., 1993. Ökologie und Ethik. Ein Versuch praktischer Philosophie. In: Ethik in Den Wissenschaften, vol. 4. Attempto, Tübingen.

Pelling, M., High, C., Dearing, J., Smith, D., 2008. Shadow spaces for social learning: a relational understanding of adaptive capacity to climate change within organisations. Environment and Planning A 40, 867–884.

Pidgeon, N., Simmons, P., Henwood, K., 2010. Risk, environment and technology. In: Taylor-Gooby, P., Zinn, J. (Eds.), Risk in Social Science. Oxford University Press, Oxford and London, pp. 94–116.

Radandt, S., Rantanen, J., Renn, O., 2008. Governance of occupational safety and health and environmental risks. In: Bischof, H.-J. (Ed.), Risk in Modern Society. Springer, Heidelberg and Berlin, pp. 127–258.

Randall, A., 1988. What mainstream economists have to say about the value of biodiversity? In: Wilson, E.O. (Ed.), Biodiversity. National Academy Press, Washington, DC, pp. 217–223.

Randall, A., Farmer, M.C., 1995. Benefits, costs, and the safe minimum standard of conservation. In: Bromley, D.W. (Ed.), The Handbook of Environmental Economics. New York and Oxford, Blackwell, pp. 26–44.

Renn, O., 2008. Risk Governance. Coping with Uncertainty in a Complex World. Earthscan, London.

Renn, O., 2014. Das Risikoparadox. Warum wir uns vor dem Falschen fürchten. Fischer, Frankfurt am Main.

Renn, O., Klinke, A., van Asselt, M., 2011. Coping with complexity, uncertainty and ambiguity in risk governance: a synthesis. AMBIO 40 (2), 231–246.

Renn, O., Walker, K., 2008. Lessons learned: a re-assessment of the IRGC frame-work on risk governance. In: Renn, O., Walker, K. (Eds.), The IRGC Risk Governance Framework: Concepts and Practice. Springer, Heidelberg and New York, pp. 331–367.

Taylor, P., 1986. Respect for Nature. A Theory of Environmental Ethics. Princeton University Press, Princeton.

US-National Research Council, Committee on Risk Characterization, 1996. Understanding Risk: Informing Decisions in a Democratic Society. National Academy Press, Washington, DC. http://books.nap.edu/catalog/5138.html.

US-National Research Council of the National Academies, 2008. Public Participation in Environmental Assessment and Decision Making. National Academies Press, Washington, DC.

WBGU – Wissenschaftlicher Beirat der Bundesregierung Globale Umweltveränderungen, 1999. Sondergutachten: Die Welt im Wandel: Umwelt und Ethik. Marburg, Metropolis.

WBGU – Wissenschaftlicher Beirat der Bundesregierung Globale Umweltveränderungen, 2000. Welt im Wandel: Erhaltung und nachhaltige Nutzung der Biosphäre. Springer, Berlin, Heidelberg and New York.

Webler, T., 1995. 'Right' discourse in citizen participation. An evaluative yardstick. In: Renn, O., Webler, T., Wiedemann, P. (Eds.), Fairness and Competence in Citizen Participation. Evaluating New Models for Environmental Discourse. Kluwer, Dordrecht and Boston, pp. 35–86.

Wynne, B., 1992. Risk and social learning: reification to engagement. In: Krimsky, S., Golding, D. (Eds.), Social Theories of Risk. Praeger, Westport, pp. 275–297.

Environmental Sciences

Chunglin Kwa, University of Amsterdam, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The environmental sciences are engaged in a remarkable effort of interdisciplinary cooperation and integration. Some long-running international scientific programs, notably the World Climate Research Programme and the International Geosphere-Biosphere Programme, play an important role therein. The origin of the environmental sciences is discussed, in particular physical geography and meteorology in the work of Alexander von Humboldt. The development of ecology during the nineteenth and twentieth centuries is traced through the work of Charles Darwin, and its historical bifurcation in evolutionary and systems approaches. The current focus on data is examined on the basis of the Long Term Ecological Research Network, and its precursor, the Grassland Biome study in the United States, 1968–76. The history of meteorology is discussed from the work of Vilhelm Bjerkness in Sweden to the development of the General Circulation Models, which form the basis of present studies of global warming and the 'human fingerprint.' Climate skepticism and geoengineering form two extreme but widely discussed attitudes toward global warming, in society and academia.

Introduction

The environmental sciences, meteorology, physical geography, ecology, and urban and regional planning are engaged in a remarkable effort of interdisciplinary cooperation and integration. The importance of climate change (and the related issue of biodiversity loss) is one driver of this process. The rise of the average global temperature will cause all kinds of interacting physical, biological, and social changes and they need to be addressed interdependently. Another driver is the political climate in which public expenditure on science is under pressure. The environmental sciences need to prove their continuing relevance for the public agenda and to show that they spend public funds wisely and efficiently. Orchestrating interdisciplinary integration of the environmental sciences is some worldwide research programs on the Earth system, under the auspices of the International Council for Science (ICSU). The World Climate Research Programme (WCRP) has been running since 1980, and the International Geosphere-Biosphere Programme (IGBP) became operational in 1988. At their start, the WCRP housed meteorology and other physical sciences, while the IGBP had ecology, atmospheric chemistry, and other biological, geological, and chemical sciences. Within the programs themselves, all kinds of disciplinary delineations were clearly visible. In recent decades, a remarkable effort in top-down planning has wiped away many of the dividing lines between disciplines, and it is expected that the WCRP and the IGBP will merge in the near future. The scientific results of the programs are increasingly evaluated on the basis of their proven relevance for climate assessments by the Intergovernmental Panel on Climate Change (IPCC) (ICSU/IGFA, 2009). There is also an increasing coordination with programs that investigate biodiversity (DIVERSITAS) and the social consequences of climatic change, the International Human Dimensions Programme. These programs also have ICSU as their sponsor. It is important to note that the ICSU is a body that speaks in the name of 'science,' not governments, while the scientists participating in the programs usually are university scientists and have a background in what used to be called 'fundamental science.' Insofar as the WCRP and the other programs do have a link with the world of policy, it is primarily through national funding agencies, such as the National Science Foundation in the United States, which have backgrounds in 'pure' science. However, increasingly, such programs have taken up socially oriented goals. The combined efforts of the program organizations and the funding agencies influence the 'little science' work modes at universities worldwide (Kwa, 2006).

Ecology

The Beginnings of Environmental Science in the Nineteenth Century

At the start of the twenty-first century, the environmental sciences, by achieving a fairly high degree of integration across disciplinary boundaries, have come full circle. Before the disciplines established themselves in the second half of the nineteenth century, several of them originated together through the work of just one individual, Alexander von Humboldt (1769–1859). Humboldt had received a university education in geology. When he was an Inspector of Mines, the inheritance of his family fortune enabled him to give up his position and to embark on a career in the service of the Prussian government. By then, he had been in contact with the naturalists on board Captain James Cook's scientific expeditions in the Pacific. In 1798, Humboldt began his own journey to Latin America. He climbed Mt Chimborazo, at the time considered the world's highest mountain. Upon his return in 1804, he became the most famous man in Europe after Napoleon. An important conceptual innovation of Alexander von Humboldt was the transformation of 'landscape' into a subject matter for science. Previously, the word landscape referred to a particular genre in painting, and Humboldt made good use of the painterly gaze, which confers unity on what it sees. Humboldt thus grasped a particular landscape, in effect, as an individual being, with a distinct 'physiognomy,' i.e., with its own visual characteristics and beauty (Kwa, 2005).

Foremost in the appearance of a landscape was its *vegetation*, another new concept, since it grouped plants according to the place where they lived rather than following their taxonomical order. Humboldt discovered that very different species occurred together, and that there were social relations among them, and he called these groupings plant associations.

The first modern textbook on ecology by the Danish botanist Eugenius Warming appeared in 1895, with the English translation appearing in 1909. Warming retained Humboldt's original precepts on plants in social associations, and worked them out in a precise manner, exposing the diversity of degrees and styles of interdependence that were found in nature. Josias Braun-Blanquet in Montpellier pursued this line of thought, and he and others in the so-called Zürich-Montpellier school produced a comprehensive system of Europe's plant associations. This vegetation science would play a major role in underpinning ecological support for nature conservation practices in many countries of continental Europe (Nicolson, 1989).

Darwin and Evolutionary Ecology

Charles Darwin's famous *Origin of Species*, first published in 1859, is as much a treatise on ecology as it is on evolutionary theory. Darwin owed much to Humboldt's holistic ecology, as he saw the various species of animals and plants forming an 'entangled' whole. But whereas Humboldt stressed harmony in nature, Darwin regarded nature much more as a creative force. Nature continually disturbs settled equilibria through the struggle of species scrambling for existence (Worster, 1985).

Darwin's work remained close to taxonomical practice in zoology and botany. He had worked for 8 years on the barnacle alone before publishing the *Origin of Species*, and he was in lifelong correspondence with Joseph Hooker, the leading taxonomist of his day, exchanging more than 1300 letters (Endersby, 2008). Hooker was the Director of the Royal Botanical Gardens at Kew, the center of an imperial network of collectors of plant species around the globe. The New York Botanical Garden was founded in 1891, just when ecology was developing into a field of study in the United States. The garden helped to facilitate the early development of ecology by providing training and research facilities (Kingsland, 2005). Although institutions devoted to taxonomy declined in importance during much of the twentieth century, the issue of biodiversity provided them with a new mission (Escobar and Paulson, 2005; Takacs, 1996).

Evolutionary ecologists stress interactions between species (populations). They do not regard the climate as the determinate factor in the makeup of animal communities, as ecosystem ecologists tend to do. Population dynamics has strong ties with applied ecology, including economic entomology (the study of pest insects in agriculture) and fishery ecology. Already by 1900, ecologists were working on biological pest control, i.e., the method of holding one insect species in check with another one. The publication of Rachel Carson's *Silent Spring* gave this research a boost (Perkins, 1982; Palladino, 1993).

Several approaches in population dynamics, in particular those of Robert May and C.S. Holling, have enlarged its theoretical basis in diverse ways. Both developed new mathematical techniques, which tackled the occurrence of more or less dramatic changes in ecological systems. In the work of Daniel Botkin, an evolutionary ecological approach is focusing on unique events in order to solve problems of nature conservation in local contexts, in contrast to the more general outlook of systems ecology. An evolutionary ecological style of work also is informing field experiments in ecology. Field experiments, such as the ongoing restoration of Lake Sempach in Switzerland, do not aim at controlling an environment, as in a laboratory experiment, but instead are oriented toward contingency (Schwartz and Krohn, 2011).

Ecosystems and Systems Ecology

Systems ecology is the mathematical, computer-implemented form of ecosystem ecology. The latter assumes that large entities such as landscapes are functionally integrated wholes and it attempts to uncover their structure and functioning.

The word 'ecosystem' was coined in 1935 by Arthur Tansley, a British ecologist. Tansley wanted to place the older associationist or community-oriented ecology on a mechanistic footing. This idea remained largely programmatic until the 1950s, when Eugene Odum pulled together some theoretical approaches and physical methodologies. Odum's synthesis launched ecosystem ecology. It incorporated an emphasis on 'productivity,' which was added to the analysis of food web relations first pioneered by Charles Elton in the United Kingdom. Odum made use of newly available physical methodologies, most importantly radioactive tracers. His approach was based on a recent discovery that several species accumulate radioactive isotopes, as they picked them up in minute quantities from the environment. Radioactive tracers permitted the elucidation of food webs both qualitatively (who eats what) and quantitatively. Odum also put the food web relations through a process of abstraction and conceived them as the circulation of matter and energy in the ecosystem. Through computer simulation of ecosystem models Odum's fellow ecosystem ecologists hoped to find the key to the rational exploitation and control of natural resources. They hoped to use the models to help them establish the conditions under which ecosystems preserve their stability and to identify the natural mechanisms that enable them to do so.

From 1968 to 1976, five very large systems ecology programs were implemented in the United States in the so-called biome studies of the International Biological Program (IBP), comprising large parts of the North American continent. The largest of these was the Grassland Biome Study at the Colorado State University. This project was headed by George Van Dyne, who by 1969 was principal investigator for 130 scientists and whose charismatic leadership went through a major crisis. These studies were pioneering in two ways: they embarked on large-scale data collection and they built computer models in an effort to simulate the entire biome. For collecting data, the Grassland Biome Study built a network; for instance, ecologists at a dozen land grant schools were contracted to clip and weigh grass at allotted pieces of land, in order to measure grass productivity. At the termination of the IBP, the comprehensive model was judged in the scientific community as both too overloaded with variables and too crude and static (Kwa, 1993; Aronova et al., 2010). Yet, ecological modeling was proven feasible, and this was true as well for the equally crude large-scale data gathering. When the National Science Foundation (NSF), in 1980, created the Long

Term Ecological Research Network (LTER), several former IBP sites were included (Kingsland, 2005; Coleman, 2010).

An outstanding feature of research at the LTER sites is the care devoted to storing data and making them available to the other sites and to researchers internationally. NSF's proactive data policy resulted in the designation of Network Data Managers centrally and for each of the sites. Metadata, the description of the data, become standardized and data are organized following 'ontologies,' a hierarchical form of data ordering developed in information science. However, several aspects of data management and data integration remain contested among scientists and difficult to implement (Baker and Bowker, 2007; Bowker, 2000).

Meteorology

Humboldt had invented isotherms, lines on a map that connect points with equal mean temperature. Together with isobars and a number of other graphic representations, such as of storm tracks, isotherms formed the basis of weather forecasting well into the twentieth century. The continuity suggested by the lines invited a hunt after ever more data, which no table with only 'point data' could have achieved (Edwards, 2010). But the skillful combination of various forms of data involved many rules of thumb.

Vilhelm Bjerknes and the Bergen School of Meteorology

Modern meteorology was founded by Vilhelm Bjerknes and his 'Bergen school' in Norway. He visualized the atmosphere as consisting of three-dimensional air masses moving around the globe, varying in temperature, density, and humidity. An important step was achieved by Bjerknes' son Jacob, who developed a model of the extratropical cyclone and how it moves. Bjerknes was able to exploit an existing military observational network for weather forecasting in western Norway, which enabled him to follow the movement of the cyclones. The model also brought him considerable popular support in the area. The major breakthrough occurred in 1919, when the concept of the 'polar front' was born. It would take several decades before the world's meteorologists were persuaded to 'see' a battle line, separating cold air masses from the north and warm air masses from the south. Bjerknes thought that this line stretched along the entire Atlantic, or indeed around the entire globe. Along the polar front, the cyclones, or low-pressure areas, were formed, which would hit the coasts of Europe with rain and storm. Bjerknes placed weather forecasting on a new footing, as the analysis of polar fronts for the purpose of forecasting relied on large networks for atmospheric data gathering (Friedman, 1989; Edwards, 2010).

The Bergen school's ideas received a decisive boost when Carl-Gustaf Rossby (1898–1967), a Swede who had worked with Bjerknes in Bergen, came to work at the US Weather Bureau during 1926–27. In the late 1930s, Rossby, by then at Massachusetts Institute of Technology, developed a method for calculating large-scale motion in the atmosphere, based on the movement of long waves (later to be called Rossby waves). He put the westward movement of the extratropical cyclones more rigorously into the context of the general circulation of the Earth's atmosphere, thus explaining semiquantitatively how the cyclones moved. On the basis of Rossby's work, precision in forecasting methods was greatly enhanced in succeeding years.

Numerical Prediction of the Weather with the Computer

John von Neumann, a mathematician, was one of the main architects of the modern computer. During World War II, he had been working on the atom bomb at Los Alamos. After the war, he initiated a project to build an 'all-purpose, automatic, high-speed electronic computing machine.' von Neumann also formulated the problems that computers were to solve, among them the modeling of the atmosphere. The reason why von Neumann chose meteorology may have been the military importance of weather forecasts. Jule Charney became von Neumann's collaborator in 1948. In early April 1950, the first successful numerical weather prediction was completed, on the basis of Charney's simple two-dimensional model of the atmosphere. Routine numerical weather prediction began by 1955.

The early models by Charney derived intellectually from the Bergen school. In the mid-1960s, however, these models were replaced in daily forecasting by models based on the so-called primitive equations, basic physical equations describing the atmosphere. As a result, the original Bergen 'fronts' moved to the background in numerical meteorology (Nebeker, 1995).

General Circulation Models

The US Weather Bureau created a General Circulation Research Section in 1955, renamed the Geophysical Fluid Dynamics Laboratory in 1963. General Circulation Models (GCMs) were to simulate the circulation of the Earth's entire atmosphere. Joseph Smagorinsky, its first Director, did not go for quick results. Instead, he believed that physical theory and numerical methods should be developed and corrected first, before any verisimilitude in the behavior of the empirically measured atmosphere could be established. During the 1960s, two more GCMs were developed, at the Lawrence Livermore National Laboratory and at the National Center for Atmospheric Research. These three models were the first of all the GCMs now in operation, estimated to be 18 if one counts only the models that link the atmosphere to the circulation of the oceans. There has been exchange of computer code between the models, and the models also are often compared with each other, leading to theoretical convergence. A related but different process is that the models are constantly being fine-tuned against all the known historical data sets of climatic variables. This is a two-way process, since the data are themselves constantly being recalibrated and reinterpreted according to modeling requirements. One obvious reason is that data sets were obtained through different measurement protocols and with different equipment. 'Raw data,' therefore, do not exist (Edwards, 2010).

Climate Skepticism

Global warming as a policy issue made its breakthrough in 1988, when James Hansen, the Director of the Goddard Institute for Space Studies, delivered a famous testimony before a Committee of the American Congress. Presidential Candidate George H. Bush promised 'to counter the Greenhouse effect by the White House effect.' A year later, physicist

William Nierenberg requested and received permission to present at the White House an unpublished report of the George C. Marshall Foundation (of which he was a cofounder), in which naturally fluctuating sun spots were identified as the sole cause of global warming and the link between global warming and human use of fossil fuels was denied (Oreskes and Conway, 2010). From then on, and in marked contrast with the European Union, the US government would refrain from policies to bring down carbon dioxide emissions in the atmosphere.

Frederick Seitz, another cofounder of the Marshall Institute, in a first-page article of the *Wall Street Journal* in 1996, accused the IPCC of fraud. The immediate target was a study by climate scientist Ben Santer, published in the IPCC's *Second Assessment Report* (1995). Santer had identified a 'human fingerprint' in the warming of the atmosphere, a particular aspect that could not be explained by the sun's activity but which could be explained by the raised level of carbon dioxide in the atmosphere. According to Seitz, the IPCC had not followed its own rules of peer-reviewed check before publishing Santer's study. The IPCC had in fact conducted such a check, although it has since sharpened its procedures. Santer's finding still stands, although at the time the damage to the IPCC and Santer personally was considerable.

From an overview of the interventions in the policy arena by Nierenberg, Seitz, and some fellow 'climate skeptics' such as Fred Singer, it has been found that their denial of a human influence on global warming is connected with a broader agenda. Enjoying considerable influence as elder statesmen of science, they have also made their mark by casting doubt on the deleterious effects of acid rain and the damaging effects of chlorofluorocarbons on the ozone layer, and they also denied the effects on health of passive smoking. The common denominator is that they opposed all governmental environmental regulation, in line with economic interests that oppose such regulations as antibusiness (Lahsen, 2005; Oreskes and Conway, 2010).

Nierenberg and Seitz were elite scientists, but in the cause of climate skepticism they were joined by a number of activists and citizen scientists. Steven McIntyre, a retired Canadian businessman, with an educational background in mathematics and economics, in 2005 launched 'Climate Audit,' a blog that scrutinizes climate data. McIntyre in particular attacked the validity of the reconstruction of the Earth's temperature over the last 1000 years by climatologist Michael Mann. Called 'the hockey stick graph,' the graphic reconstruction shows a steady but slow decline in temperature until 1900, and a steep rise thereafter. The controversy was long and bitter, and drew in many parties, including the US National Academy of Sciences and the US Congress. To date, the hockey stick has been confirmed by a number of subsequent temperature reconstructions (Edwards, 2010).

Geoengineering

A recent report issued by the UK Royal Society (2009) on geoengineering stands on the opposite end of the spectrum of scientific and societal reactions to the danger of global warming from that of the climate skeptics. Geoengineering as currently understood comprises two basic techniques for intervening in the global physics of the atmosphere: 'solar radiation management' and 'carbon dioxide capture.' The most serious method of solar radiation management is blasting aerosols in the atmosphere, in order to achieve a cooling effect similar to large volcanic eruptions. An example of carbon capture is fertilizing the ocean with iron, which would cause algal bloom and thereby capture carbon dioxide from the atmosphere. The Royal Society report asks "Why should appropriate geoengineering options not be added to the portfolio of options of dealing with climate change?" (Royal Society, 2009: p. 45). The report adds that "…geoengineering is likely to be technically feasible, and could substantially reduce the costs and risks of climate change" (p. 47).

A study by social scientists and philosophers, which was commissioned by the British Parliament, formulated a set of principles as part of a framework for the governance of geoengineering. One of these principles stated that geoengineering should be regulated as a public good and that there should be public participation in decision making (Rayner et al., 2010; Rayner, 2011). This has not prevented strong critiques of the high risks involved in geoengineering (Fleming, 2010; Kwa and van Hemert, 2011). At present, research on geoengineering techniques, such as computer simulations of their projected outcomes, is becoming part of the scientific planning process of the WCRP and the IGBP, along with research on the unintended consequences of geoengineering.

Conclusion

The examples of global warming research and geoengineering show the remarkable extent to which current environmental science has become politicized. It would be unreasonable to expect otherwise: the burden placed on societies, to design and implement mitigation and adaptation strategies for climate change, is not a light one. Since scientific evidence obtained in the laboratory and with the help of mathematical models cannot be simply translated into scientific evidence for public decision making, it is likely that there will be widely divergent societal assessments of risks and uncertainties (Jasanoff and Wynne, 1998; Hulme, 2009). Translation requires the involvement of scientists and politicians, and experts and lay citizens. Political scientist Mark B. Brown has argued that the assumption prevalent in policy advisory processes, that experts do not have interests and politicians and laypeople lack expertise, is wrong (Brown, 2009). Likewise, philosopher and anthropologist Bruno Latour wrote that societies lack the institutions that can 'take into account' the inclusion of new items and concerns into the political process regarding issues that involve the sciences (Latour, 2004). An exception mentioned by Latour is the IPCC, which is unique in that it is at the same time an intergovernmental and a scientific organization. This sympathetic assessment of the IPCC also implies that one IPCC is not enough to guide the development of the environmental sciences and their societal embedding.

See also: Climate Change and the Media; Climate Change; Darwin, Charles Robert (1809–82); Environmental Sociology; Risk, Sociology and Politics of; Science and Social Movements;

Science and Technology Studies: Experts and Expertise; Science, History of; Science, Sociology of; Scientific Controversies; von Humboldt, Alexander (1769–1859).

Bibliography

Aronova, E., Baker, K.S., Oreskes, N., 2010. Big science and big data in ecology: from the international geophysical year through the international biological program to the long term ecological research network (LTER), 1957–present. Historical Studies in the Natural Sciences 40, 183–224.

Baker, K.S., Bowker, G., 2007. Information ecology: open system environment for data, memories, and knowing. Journal of Intelligent Information Systems 29, 127–144.

Bowker, G., 2000. Biodiversity data diversity. Social Studies of Science 30, 643–684.

Brown, M.B., 2009. Science in Democracy: Expertise, Institutions, and Representation. The MIT Press, Cambridge, MA.

Coleman, D., 2010. Big Ecology: The Emergence of Ecosystem Science. University of California Press, Berkeley, CA.

Edwards, P.N., 2010. A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. MIT Press, Cambridge, MA.

Endersby, J., 2008. Imperial Nature: Joseph Hooker and the Practices of Victorian Science. University of Chicago Press, Chicago, IL.

Escobar, A., Paulson, S., 2005. The emergence of collective ethnic identities and alternative political ecologies in the Colombian Pacific rain forest. In: Paulson, S., Gezon, L.L. (Eds.), Political Ecology across Spaces, Scales and Social Groups. Rutgers University Press, New Brunswick, NJ, pp. 257–278.

Fleming, J.R., 2010. Fixing the Sky: The Checkered History Weather and Climate Control. Columbia University Press, New York.

Friedman, R.M., 1989. Appropriating the Weather. Cornell University Press, Ithaca, NY.

Hulme, M., 2009. Why We Disagree about Climate Change: Understanding Controversy, Inaction and Opportunity. Cambridge University Press, Cambridge.

ICSU/IGFA, 2009. Review of the International Geosphere-Biosphere Programme (IGBP): Report from an ICSU-IGFA Review Panel. ICSU, Paris.

Jasanoff, S., Wynne, B., 1998. Science and decisionmaking, chapter 1 of human choice and climate change. In: Rayner, S., Malone, E. (Eds.), The Societal Framework, vol. 1. Battelle Press, Columbus, OH, pp. 1–87.

Kingsland, S., 2005. The Evolution of American Ecology, 1890–2000. Johns Hopkins University Press, Baltimore, MD.

Kwa, C., 1993. Modeling the grasslands. Historical Studies in the Physical and Biological Sciences 24, 125–155.

Kwa, C., 2005. Alexander von Humboldt's invention of the natural landscape. The European Legacy 10, 149–162.

Kwa, C., 2006. Interdisciplinary research through informal science-policy interactions. Science & Public Policy 10, 457–467.

Kwa, C., van Hemert, M., 2011. Engineering the Planet: the issue of biodiversity in the framework of climate manipulation and climate governance. Quadern 76, 79–89.

Lahsen, M., 2005. Technocracy, democracy, and US climate politics: the need for demarcations. Science, Technology & Human Values 30, 137–169.

Latour, B., 2004. Politics of Nature: How to Bring the Sciences into Democracy. Harvard University Press, Cambridge, MA.

Nebeker, F., 1995. Calculating the Weather. Academic Press, New York.

Nicolson, M., 1989. National styles, divergent classifications. Knowledge and Society 8, 139–186.

Oreskes, N., Conway, E.M., 2010. Merchants of Doubt: How a Handful of Scientists Obscured the Truth from Tobacco Smoke to Global Warming. Bloomsbury Press, New York.

Palladino, P., 1993. On environmentalism: the origins of debates over policy for pest-control research in America, 1960–1975. In: Shortland, M. (Ed.), Science and Nature. British Society for the History of Science, Oxford, pp. 181–212.

Perkins, J.H., 1982. Insects, Experts and the Insecticide Crisis. Springer, New York.

Rayner, S., 2011. Climate Geoengineering Governance. Jahrbuch Ökologie (January). www.jahrbuch-oekologie.de/aktuelles_rayner.htm (accessed 06.09.13.).

Rayner, S., et al., 2010. Memorandum on Draft Principles for the Conduct of Geoengineering Research. House of Commons Science and Technology Committee enquiry into The Regulation of Geoengineering. HMSO, London.

Royal Society, 2009. Geoengineering the Climate: Science, Governance and Uncertainty. Royal Society, London.

Schwartz, A., Krohn, W., 2011. Experimenting with the concept of experiment: probing the epochal break. In: Nordmann, A., Radder, H., Schiemann, G. (Eds.), Science Transformed? Debating Claims of an Epochal Break. University of Pittsburgh Press, Pittsburgh, PA, pp. 119–134.

Takacs, D., 1996. The Idea of Biodiversity: Philosophies of Paradise. Johns Hopkins University Press, Baltimore, MD.

Worster, D., 1985. Nature's Economy: A History of Ecological Ideas. Cambridge University Press, Cambridge.

Environmental Sociology

Riley E Dunlap, Oklahoma State University, Stillwater, OK, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Environmental sociology, the study of societal-environmental interactions, developed four decades ago in the USA and has spread internationally and become institutionalized around the world. Key foci of the field include examining the social construction of environmental problems as well as analyzing the causes and impacts of, and solutions to, such problems. Major advances in methods and data availability are yielding a rapid increase in our understanding of driving forces and impacts, but there is great debate over potential solutions. The result is an intellectually vibrant field of study.

Environmental sociology emerged in the USA in the 1970s, sparked first by sociological interest in the growth of environmental awareness and activism and then by recognition of the societal relevance of the energy crisis and a growing number of ecological problems (Dunlap and Catton, 1979). It then spread to Europe, where citizen antinuclear protests and Green politics were major foci, and gradually throughout much of the rest of the world. In four decades, it has evolved from a small, marginal field to an increasingly mainstream area. For example, in the USA, many and perhaps a majority of sociology departments in colleges and universities now offer a course in the area, several graduate programs offer concentrations in it, and a growing number of departments are hiring specialists in environmental sociology. Concomitantly, various environmental sociology handbooks (e.g., Redclift and Woodgate, 2010; Gross and Heinrichs, 2010) and numerous textbooks, often in multiple editions (e.g., Bell, 2012; Gould and Lewis, 2014; Harper, 2012), are in print. Most importantly, environmental sociologists are producing a flood of journal articles, increasingly published in elite sociology journals (e.g., Jorgenson and Clark, 2012; Pais et al., 2014; Pellow and Brehm, 2013; York et al., 2003) and research monographs. Parallel trends are occurring in Europe, Asia, and to some degree in Latin America.

The growth and institutionalization of environmental sociology is also evident in terms of organizational developments. The American Sociological Association's Section on Environmental Sociology, founded in 1976 (and now labeled the Section on Environment and Technology), is thriving after modest growth in its first two decades (Dunlap and Catton, 1994), and similar research groups exist within the national sociological associations of several other nations (e.g., Canada and the UK). There is also an environmental sociology research network within the European Sociological Association, while an autonomous Japanese Association for Environmental Sociology has grown into the world's largest organizational body for the field. Most significantly, the International Sociological Association's Research Committee on Environment and Society (RC 24) has quickly grown into one of the largest and most active research committees in the ISA, and provides an excellent forum for the international exchange of ideas and perspectives among practitioners. In short, environmental sociology has become a widely recognized and institutionalized area within the discipline of sociology, and is flourishing not only within several nations but internationally as well.

The institutional and organizational growth of environmental sociology has been accompanied by its intellectual growth and vitality. As the rest of this article will show, the increasing availability of data on environmental problems and ecological conditions has spurred rapid progress in the methodological tools employed by environmental sociologists, while theoretical debates have become more prominent as well. Some of these developments reflect regional differences in methodological and theoretical approaches to environmental sociology across the world, with American and European perspectives being especially notable. I turn to these issues while also conveying the broad contours of the evolution of environmental sociology.

Early Emphases and Disciplinary Context

Modern sociological interest in the physical environment was largely a response to widespread societal awareness of environmental problems, and mobilization of support for environmental protection symbolized by celebration of the 1970 'Earth Day' in the USA and then the 1972 UN Conference on the Human Environment in Stockholm. Early sociological research on environmental topics involved analyses of public opinion toward environmental issues; environmental activism at both the individual and organizational levels; governmental agencies responsible for natural resource management and environmental protection; the roles of activists, media, scientists, and public officials in generating attention to environmental problems; and Green parties and politics (Yearley, 1991). This research applied perspectives from established sociological fields such as social psychology, social movements, political sociology, and organizational sociology to environmental topics, constituting a "sociology of environmental issues" (Dunlap and Catton, 1979).

The 1973–74 energy crisis highlighting the dependence of industrialized societies on fossil fuels, and increasing awareness of the seriousness of air and water pollution throughout the 1970s, ushered in a new strand of sociological research – examining how societies affect their environments and, in turn, are affected by changing environmental conditions such as pollution and resource scarcity. This concern with

 http://dx.doi.org/10.1016/B978-0-08-097086-8.91030-4

societal-environmental relationships reflected the emergence of a true 'environmental sociology,' and by the late 1970s, it was a small but vigorous field (Dunlap and Catton, 1979). Its focus on the relationships between modern societies and their environments represented a major departure from disciplinary norms, however, giving environmental sociology a critical stance relative to the larger discipline.

Sociology became a distinct discipline by emphasizing the social – as opposed to biological, geographical, and psychological – influences on social behavior. It developed during an era of general resource abundance, technological progress, and economic growth. As a result, sociology took root in a cultural worldview which assumed that sophisticated social organization (e.g., complex division of labor) and scientific and technological advances (e.g., nuclear energy) had freed industrial societies from resource and other environmental constraints. This assumption reinforced negative reactions to earlier excesses of 'environmental determinism' such as geographers' efforts to explain cultural differences via climatic variation. The result was that mid-twentieth-century sociology largely ignored the physical environment, and sociological references to 'the environment' typically meant the social context of the phenomena being investigated. More fundamentally, mainstream sociology was premised on a taken-for-granted 'human exemptionalism paradigm' that assumed modern, industrial societies had become exempt from ecological constraints (Catton and Dunlap, 1980).

Sociological analyses of the societal impacts of energy shortages and possibility of ecological 'limits to growth' constituted a significant disciplinary development in the 1970s, and clearly challenged the notion of human exemptionalism. This work was quickly supplemented by research on the social impacts of toxic contamination and other forms of pollution, as well as preliminary examinations of the societal factors generating environmental degradation. A particularly important topic in the USA was environmental justice, with a growing number of studies examining the inequitable distribution of environmental hazards across racial and socioeconomic strata as well as the growth of movements to combat the obvious injustices documented (Mohai et al., 2009). By the early 1990s, many environmental sociologists were ignoring disciplinary norms by analyzing the societal causes and impacts of environmental problems, albeit typically at a local or regional level, representing serious inquiry into societal-environmental interactions (Fischer-Kowalski and Haberl, 1993).

Major Trends and Emphases in the Field

As environmental sociology has evolved, its research emphases have formed around central aspects of environmental (now more typically termed 'ecological') problems: they are caused by human actions, they are treated as problems because of their negative impacts on humans and things that humans value (e.g., other species), solving them requires collective action, and all of the foregoing necessitate societal recognition that various conditions do in fact represent 'ecological problems.' The result is that the bulk of environmental sociology can be seen as focusing on the social construction of environmental problems or analyses of their causes, impacts, and potential solutions.

The Social Construction of Environmental Problems

As noted earlier, sociological interest in the environment was stimulated by societal attention to environmental problems, and much of early sociological work focused on analyzing the manner in which societal recognition developed and evolved. Such work involved studies of environmentalism (both of individual activists as well as environmental organizations), public opinion, and governmental responses (Dunlap and Catton, 1979). Eventually, attention was focused explicitly on the 'social construction' of environmental problems. Building upon such basic observations as levels of airborne particulates being ignored in one era and/or locale but seen as 'pollution' later on or in other places, a rich body of sociological literature has clarified how various environmental conditions become defined as 'problems'; the frequently contested nature of such problems; and the implications of competing interpretations of the sources, impacts, and solutions of the problems. Originally relying heavily on perspectives from social problems and the sociology of science (Yearley, 1991), over time such analyses have drawn on social movement theory, discourse analysis, and cultural sociology.

Numerous studies of claims, claims-makers and claims-making activities have yielded crucial insights into the social processes necessary to generate widespread acceptance of conditions – ranging from local toxic wastes to global climate change – as 'problematic.' Originally, the emphasis was on environmental activists as key claims-makers, but the importance of scientists has drawn increasing attention (Yearley, 2005). The vital role of the media both in publicizing and interpreting claims has also received attention. Such analyses have demonstrated that environmental problems do not simply emerge from objective conditions, but that their recognition is contingent upon issue entrepreneurs and advocacy coalitions (activists, scientists, and policy-makers) being successful in overcoming barriers to getting issues onto the public agenda, and then gaining widespread societal acceptance of their definition of the situation and calls for ameliorative action (Hannigan, 2006).

Early environmental problems like air and water pollution were readily perceptible, but newer problems such as toxic wastes, ozone depletion, and climate change depend on scientific measurements and interpretation for their discovery, analysis, and possible amelioration. The resulting heavy dependence of environmental advocates on scientific evidence has received scrutiny from sociologists. Analysts such as Yearley (2005) have emphasized that the environmental movement's reliance on science is a mixed blessing for several reasons: (1) demands for scientific proof can be used to stall action, particularly by unsympathetic politicians; (2) the probabilistic and tentative nature of scientific evidence falls short of the definitive answers laypeople and policy-makers seek; and (3) reliance on scientific claims makes environmental advocates vulnerable to counterclaims issued by 'skeptic scientists' supported by industry. This has been particularly true of anthropogenic climate change, where contrarian scientists – often backed by fossil fuels corporations and conservative think

tanks – have had considerable success in 'deproblematizing' global warming (McCright and Dunlap, 2010).

Constructivist approaches demonstrate that environmental problems do not simply emerge from changes in objective conditions, that scientific evidence is seldom sufficient for establishing conditions as problematic, and that the manner in which conditions are framed as problems requiring action is crucial. These are major insights, and represent essential sociological contributions. However, in the 1990s, some constructivists followed postmodern fads and 'deconstructed' not only environmental problems and controversies, but 'the environment' (or, more typically, 'nature') itself. Proclamations that "… there is no singular 'nature' as such, only a diversity of contested natures …." (Macnaghten and Urry, 1998: 1) were not uncommon, and posed a challenge to the predominant realist orientation of most environmental sociologists who took the 'reality' of problematic environmental conditions as given even while sensitive to the necessity of such conditions being defined and recognized as problems.

Constructivist excesses provoked a reaction from environmental sociologists of a realist bent, who argued that while one can deconstruct the concept of nature, an obvious human (and culturally bound) construction, this hardly challenges the existence of the global ecosystem and by implication various manifestations of ecosystem change (e.g., ozone depletion) construed as 'problems.' More specifically, a growing chorus of representatives of the realist camp argued that a strong constructivist approach, which ignores the validity of competing environmental claims, slips into relativism, has the potential of undermining environmental scientists and playing into the hands of their opponents, and precludes meaningful examination of societal-environmental relations seen as fundamental to environmental sociology and providing its distinctive focus on the material world (Dunlap and Catton, 1994; Murphy, 2002).

In response to such criticism, many constructivists emphasized that they were not denying the material 'reality' of environmental problems, as their postmodern rhetoric sometimes suggested, but were simply problematizing environmental claims and knowledge. In eschewing relativism in favor of 'mild' or 'contextual' constructivism, most constructivists have moved toward more common ground with their realist colleagues. The latter, in turn, have moved toward a 'critical realist' perspective that, although firmly grounded on acceptance of a reality independent of human understanding, recognizes that scientific (and other) knowledge is imperfect and evolving (Carolan, 2005). The result is that the 'realist–constructivist wars' of the 1990s have subsided; however, a remaining residue is visible in the stances adopted toward utilizing scientific evidence on environmental problems.

From Constructivism versus Realism to Agnosticism versus Pragmatism

The constructivist/realist divide in environmental sociology has evolved into a related but broader pair of stances termed 'agnosticism' and 'pragmatism' based on analysts' treatment of environmental phenomena in their research (Dunlap, 2010). Specifically, those adopting an agnostic stance (especially prevalent among but not confined to European scholars) toward environmental problems are reluctant to employ scientific indicators of these phenomena, preferring instead to problematize, contextualize, and sometimes deconstruct scientific evidence regarding issues such as anthropogenic climate change. For example, their analyses offer valuable insights into the complexities and contingencies of climate science, but because they are hesitant about granting ontological standing to the claims of climate scientists they avoid examining the causes and likely impacts of rising greenhouse gas (GHG) emissions (Yearley, 2009).

In contrast, the pragmatic approach, which is popular with (but not limited to) American scholars, employs the best available scientific data on a wide range of environmental conditions – carbon dioxide and other GHG emissions, deforestation, toxic wastes, and so on – in empirical analyses to investigate their relationships with an array of social phenomena. While many of these analyses of societal-environmental interactions involve rigorous statistical analyses of large, quantitative data sets, pragmatists also perform more qualitative, in-depth case studies that provide rich insights into the evolution of such interactions over long periods of time (Freudenburg et al., 2009).

The agnosticism about the state of environmental conditions that has roots in the postmodernist turn of the larger discipline, and the popularity of constructivist analyses within environmental sociology so prevalent in the 1990s, continues to have pervasive effects on sociological research and theorizing about societal-environmental relations. This is particularly true among European scholars, where there is far less work on the material dimensions of environmental problems (see Fischer-Kowalski and Haberl, 2007 for a notable exception). Instead, the agnostic camp tends to treat environmental matters as symbolic/ideational/cultural phenomena, to be analyzed via hermeneutic/interpretative approaches. Its efforts to incorporate the material world into sociological analyses are often limited to the discursive realm via talk about 'hybrids,' 'cyborgs,' and such (Dunlap, 2010).

So strong is this agnostic stance in Europe that even theorists holding a presumably more realist bent adopt agnostic positions toward the utility of scientific data for the study of environmental phenomena. For example, proponents of ecological modernization theory (EMT) (discussed below) tend to advocate skepticism toward rigorous empirical analyses of environmental conditions in favor of methods that are more qualitative and subjective. Leading ecological modernization theorists Mol and Spaargaren (see Mol et al., 2009), for example, express skepticism about the ability of empirical research to settle theoretical debates and largely reject the use of empirical methods and hypothesis testing in analyses of the environmental consequences of modernization. They further question the validity of using "natural science 'empirical facts'" and quantitative analyses to assess socially generated environmental problems (see quotes in York et al., 2010: 80).

Such a stance (not unique to Mol and Spaargaren) is at odds with the large and growing body of sociological research that relies heavily on measures of environmental conditions produced by natural scientists, government agencies, and sometimes environmental activists to analyze empirically diverse and often competing perspectives on the causes or driving forces of environmental degradation (e.g., York et al.,

2003), the impacts of degradation (Pais et al., 2014), and mechanisms of environmental reform (e.g., Roberts et al., 2004). While empirically oriented pragmatists within environmental sociology can benefit from the cautions raised by agnostics concerning the pitfalls of employing imperfect data on environmental conditions, the gulf between the two 'camps' concerning the utility of empirical evidence for testing theoretical perspectives and adjudicating debates between competing perspectives clearly poses a challenge for the field.

Causes or Driving Forces of Environmental Problems

From the outset, environmental sociologists have focused attention on the societal factors that produce environmental degradation (Dunlap and Catton, 1979: 252), although early work was hampered by a paucity of data on environmental conditions. Consequently, early work tended to examine arguments being put forth by natural scientists, particularly biological ecologists like Paul Ehrlich and Barry Commoner. Most influential in this regard was Schnaiberg's (1980) analysis of the weaknesses of emphasizing population growth, technological development, and materialistic consumption patterns as the key drivers of degradation. In contrast, Schnaiberg drew on political economy perspectives to develop a sophisticated alternative that stresses the inherent need for market-based firms to grow, to replace costly labor with advanced technologies, and to increase their use of resources and production of pollution in an ever-expanding production process. He argued that a powerful coalition of capital, state, and labor forms a 'growth coalition' that promotes and defends the resulting 'treadmill of production,' which yields endless growth (and escalating environmental degradation).

The treadmill model has been supplemented by a range of more explicitly Marxist analyses, including O'Connor's (1991) 'second contradiction of capitalism' notion, which argues that increasing the scale of production raises production costs and levels of pollution, as more expensive and difficult-to-obtain resources are continually substituted as those most accessible are depleted (1997). Like the treadmill, the second contradiction points to the unsustainability of endless economic growth. In a complementary fashion, Foster's use of Marx's concept of 'metabolic rift' clarifies the manner in which production and consumption in capitalist systems are inherently unecological (Foster et al., 2011). All three perspectives converge to provide a 'political economy' model that suggests that Capitalist-driven economic growth is the key factor producing environmental degradation worldwide, and the global free-market economy is the primary barrier to slowing such degradation and achieving ecological sustainability.

The resulting model, focused originally on the USA, has been extended to the international level with a marriage with World Systems Theory and its derivatives such as ecologically unequal exchange to create a global political economy model now widely used within environmental sociology (Jorgenson et al., 2014). With the growing availability of a wide range of data on resource use, pollution production and overall ecological impact at the level of nation-states, quantitatively sophisticated cross-national research has provided considerable insight into the driving forces of deforestation, GHG emissions, and ecological footprints (Jorgenson and Clark, 2012). A critical insight has been the degree to which the wealthy 'core' nations are able to protect their own environmental quality by using poorer, peripheral nations as both sources of natural resources and as sinks for pollution (Jorgenson and Clark, 2011). Such studies reinforce the pessimistic implications of the political economy model at the international level, suggesting that the global growth of a Capitalist-driven free-market economy is ecologically unsustainable as well as highly inequitable.

An alternative model of the driving forces of environmental degradation has evolved directly from the work of biologists such as Commoner and Ehrlich, particularly their famous 'IPAT' equation, which viewed (negative) environmental impacts as a function of population, affluence, and technology. Environmental sociologists have drawn on human ecology to develop a stochastic 'STIRPAT' model (technically, the STochastic Impacts by Regression on Population, Affluence and Technology model) allowing for predictions of a variety of environmental impacts, ranging from ecological footprints to GHG emissions to fertilizer and pesticide use (York et al., 2003). Such analyses reveal that population and affluence both have strong, positive effects on the ecological impacts of nations, and thus supplement rather than contradict findings from the political economy perspective by suggesting that growth of all types (whether in population, affluence, consumption) is incompatible with ecological sustainability.

As will be discussed below, both of the dominant strands of research on the driving forces of environmental degradation yield results that contrast with the more sanguine views offered by proponents of EMT. EMT not only fails to view Capitalism as inherently ecologically destructive, but suggests that a market economy coupled with technological innovation offers the best hope of solving environmental problems and moving toward a more sustainable future (Mol et al., 2009). EMT is examined in the section on 'solving environmental problems.'

Impacts of Environmental Problems: Inequitable Distribution and Environmental Justice

A key dimension of early work by environmental sociologists was a concern with the social impacts of environmental and resource problems, particularly the distributional aspects of negative impacts. Inequality has long been a central focus of the larger discipline, and it is thus no surprise that the inequitable impacts of energy shortages and the regressive impacts of policies to deal with them were major concerns in the 1970s (Dunlap and Catton, 1979: 263). With the discovery of toxic contamination in Love Canal and other communities, the distribution of exposure to environmental hazards became a central focus of environmental sociology. Numerous studies have documented the inequitable manner in which toxic wastes and other environmental hazards are distributed across racial and socioeconomic strata, with minorities and lower socioeconomic levels being disproportionately exposed to them (Mohai et al., 2009). Debates remain over whether race or class is the crucial factor, and increasingly sophisticated studies have helped clarify the complex interplay between these factors as well as the role of historical trajectories in influencing contemporary patterns of inequitable exposure (Pais et al., 2014).

Work in the realm of what is now termed environmental justice research has become methodologically rigorous, due to sophisticated analytical techniques and improved data availability, especially the use of geographic information systems (GIS) to overlay biophysical data with socioeconomic data across geographical units ranging from neighborhoods to census tracts. Such advances allow, for example, researchers to clearly identify urban 'heat islands' in which residents are exposed to dangerously high temperatures during heat waves and the socioeconomic characteristics of these neighborhoods – typically disproportionately populated with poor, racial minorities, elderly and isolated individuals (Harlan et al., 2013). Such technically strong work has advanced our knowledge of environmental 'injustices' immensely, and has been complemented by advances in theorizing the origins of environmental injustice and examinations of the emergence and international diffusion of social movements pursuing the goal environmental justice (Mohai et al., 2009).

At the other end of the geographical spectrum, a large and growing body of research and theorizing focuses on inequities in generating environmental degradation and experiencing the effects of such degradation throughout the global system. Intimately related to the previously noted work on driving forces, this scholarship has consistently documented inequities in the global economic system that benefit wealthy nations at the expense of poorer nations. Poor nations are pressured to engage in ecologically unequal exchanges with wealthy nations, providing their natural resources as inputs for use by multinational corporations (sometimes for materials produced in factories in their own nations, other times shipped abroad) in efforts to improve their own economies – but often with little success. At the same time, resource extraction generates environmental degradation as does the use of their territory as 'sinks' for international pollution, whether in the form of housing polluting factories that ship goods abroad or as actual dumping grounds for hazardous materials from wealthy nations (Jorgenson et al., 2014). Rising awareness of such processes is stimulating environmental justice campaigns around the world.

Beyond an explicit focus on equity, sociological work has documented the deleterious impacts of environmental problems created by human-created technological accidents and hazards. Unlike natural disasters such as floods, earthquakes, tornadoes, and hurricanes – which have been found to produce a therapeutic response in the affected communities, uniting the populace in helping victims and repairing damage – technologically induced disasters like release of toxic wastes tend to have a corrosive effect on community life. The frequent ambiguities inherent in documenting the sources and seriousness of the contamination often generate conflict among different sectors, especially those economically dependent on the perceived source and those most affected by the contamination. In-depth qualitative studies have documented the long-term erosion of community life and personal tolls on residents (Freudenburg, 1997).

Solving Environmental Problems

In the USA, environmental sociologists have paid less attention to possible solutions to environmental problems than to their causes and impacts, and early work on energy policies emphasized their regressive impacts rather than efficacy (Dunlap and Catton, 1979). Currently, various strands of relevant research, from a microlevel focus on developing 'behavioral wedges' to stimulate residential energy conservation (Dietz et al., 2009) to macrolevel analyses of the forces promoting and hampering the adoption of international environmental treaties (Roberts et al., 2004), are in progress. However, attention to solutions took a quantum leap forward when European environmental sociologists began analyzing and generalizing from instances of significant environmental reforms in some Northern European nations. The result was the emergence of EMT, a perspective that has enjoyed widespread visibility in Europe and, increasingly, in the rest of the world (Mol et al., 2009).

EMT called for a refocusing of the field, from analyzing the sources of environmental degradation to examining the sources and mechanisms of environmental reform. Building on industrial ecology models suggesting that production can grow with declining input of materials and resulting pollution, EMT quickly supplemented such claims of 'dematerializing' the economy via technological innovation with analyses of complementary societal developments. New forms of collaboration between industry, government, and civil society were seen as institutionalizing an 'ecological rationality' driven by science and technology that not only tempers environmentally harmful decision making, but stimulates a 'green capitalism' that marries the power of the market with the pursuit of environmental quality (Mol et al., 2009). EMT became popular in both academic and policy circles, in part because it not only sees Capitalism as friend rather than foe of environmental progress – a message that resonates well with powerful interests in an era of hegemonic free-market ideology – but embodies optimistic projections directly counter to the pessimistic views of both the political economy and human ecology perspectives. Such tendencies have led to it being portrayed as a very conservative approach to the possibility of achieving significant change in the current economic system (Foster, 2012).

Based on evidence from Northern European nations, EMT appeared to offer a welcome alternative to dominant perspectives in the field. Not surprisingly, however, its optimistic claims met with skepticism among American scholars, and critics soon pointed to the limitations of case studies of 'success stories' (whether individual factories or nations) and the focus on institutional outcomes (e.g., new laws and agencies) and behaviors (from establishment of corporate environmental officers to green consumerism). In contrast, a growing number of empirical studies examining actual environmental conditions such as GHG emissions and ecological footprints consistently find that economic growth produces higher levels of environmental degradation, especially at the global level (Jorgenson and Clark, 2011, 2012; York et al., 2003), and that improvements in environmental quality within some wealthy nations have come at the expense of their ecological loads being shifted to poor nations and the global commons (Jorgenson and Clark, 2011; York et al., 2010). Critics have further argued that in nations such as the USA, where free-market forces reign supreme, the sources of ecological rationality – environmental science, environmental pressure groups, and environmental agencies – are under constant

assault, threatening to create a process of ecological *demodernization*. The result is that not only does *unecological* modernization (i.e., growth as usual) continue, but ecological irrationality (like denial of anthropogenic climate change) is increasingly institutionalized (McCright and Dunlap, 2010).

These empirical tests and empirically based criticisms have produced something of an impasse, as leading proponents of EMT tend to dismiss the mounting number of studies disconfirming their optimistic projections. While not postmodern relativists, they have nonetheless adopted a very agnostic stance toward the use of scientific measures of environmental indicators and questioned the utility of empirical testing of theories as noted earlier (and documented in York et al., 2010). Unlike strong constructivists, who limit their analyses to the discursive realm, EMT proponents tend to limit theirs to the institutional and cultural realms and proclaim progress when regulations are adopted or green consumerism becomes more popular – without carefully assessing whether these developments produce discernible reductions in ecological impacts, especially at the global level (which is the bottom line when assessing the impacts of continued economic growth). Thus, to American 'realists,' EMT seems naively optimistic and lacking in empirical support, with its refusal to acknowledge troubling evidence of continuing ecological degradation representing a new version of human exemptionalism (Foster, 2012). Nonetheless, EMT may offer useful suggestions for developing nations that seek to lower their environmental impacts, such as China where it seems to be making inroads (Hong et al., 2014).

Complementing EMT is world polity theory (WPT), which posits the emergence of an evolving world culture that diffuses international norms concerning the duties of nation-states, including environmental protection. Emphasizing the role of both governmental and nongovernmental agents (international treaties, intergovernmental organizations, transnational environmental organizations, etc.), proponents of WPT argue that an international 'environmental regime' has been established that fosters environmental protection worldwide. Several empirical studies have documented the global diffusion of environmental protection mechanisms, including increased levels of activism from civil society sectors (especially environmental nongovernmental organizations) that put pressure on governments in less-developed nations to strengthen environmental protection efforts. The degree to which such activities can counteract the ecologically harmful unequal exchange relationships imposed on these nations by the global economic system and yield actual improvements in environmental quality still remains unclear (Jorgenson et al., 2014; Knight and Messer, 2012).

In addition, nations that were once innovators in environmental protection (especially the USA in the early 1970s), and thus major sources of stimulation for the establishment of an international environmental regime, are now reversing course as their governments are taken over by antiregulatory, free-market advocates. Watching the increasingly successful efforts of the American Right to dismantle environmental protection regulations and agencies in the USA (McCright and Dunlap, 2010), and the spread of similar efforts to other nations with neoliberal regimes such as Canada and Australia, makes this reversal a significant possibility – as witness the obstructive roles currently being played by these three nations in international negotiations on climate change policy.

Clearly there is need for additional sociological work on efforts to solve environmental degradation, especially to assess the likelihood that increasing ecological deterioration at the global level (e.g., rising GHG concentrations) can not only be slowed but eventually halted, particularly with a global economic system premised on endless growth. Thus far the empirical evidence on global trends does not match the optimistic projections of EMT and WPT (Jorgenson and Clark, 2011, 2012; Jorgenson et al., 2014; York et al., 2003, 2010).

Improved Data and Increasing Methodological Sophistication: An Illustration

At the beginning and throughout this article, reference has been made to the impact of the availability of improved data sets, especially on environmental conditions, and increasingly sophisticated methodological tools on the field of environmental sociology. For empirically inclined scholars, these trends are revolutionizing the field of environmental sociology and promise to continue to do so. Their impact is especially apparent in work on one of the oldest topics in the field, environmental attitudes and behaviors. Early work on public opinion evolved into efforts to carefully conceptualize "environmental concern" (Dunlap and Jones, 2000) and to develop theoretical models to predict and explain both the attitudinal and behavioral components of such concern. While European scholars often employed rational choice theory in both surveys and experimental studies of behaviors (Diekmann and Schmidt, 1998), Americans developed models of pro-environmental behavior building on theories of values and norm activation (Stern et al., 1999), while both contributed to clarifying key demographic predictors of environmental concern such as age, education, and political orientation.

The emergence in the 1990s of multinational data sets on citizens' views of environmental issues led to cross-national studies becoming popular. A debate evolved between Europeans working from a rational choice perspective viewing citizen concern as a result of affluence, and thus more prevalent in wealthy nations (Franzen and Meyer, 2012), and Americans viewing such concern as stimulated by awareness of environmental degradation, and thus widely distributed across nations and often higher in poor nations (Dunlap and York, 2008). The contrasting results are partially attributable to differences in the measures employed, with indicators of willingness to pay favored by proponents of the affluence perspective and a wider range of indicators used by proponents of the globalization perspective (Knight and Messer, 2012).

Increasingly, multilevel models examining the relative impact of national characteristics such as economic affluence and environmental degradation and individual characteristics such as values and income are being conducted. These studies provide rich insights into the interplay of personal characteristics and national contexts, showing, for example, that while *national* affluence does *not* lead to higher levels of citizen concern for the environment, within nations there may be (depending on the measure used) positive relationships

between postmaterial values and *individual* income and environmental concern (Fairbrother, 2013). More work along these lines, especially with improved data on environmental degradation within nations, should clarify the factors that enhance citizen support for environmental protection and how these vary with national and regional contexts.

Concluding Observations and Future Projections

The foregoing review has been limited to major emphases and controversies within the field of environmental sociology, particularly trends in research on the causes and impacts of environmental degradation and controversies regarding sociological work on potential solutions. It is worth highlighting three important issues: First, the increasing methodological sophistication of many practitioners and the ever-growing bodies of data on ecological conditions is yielding unprecedented advances in our understanding of the driving forces and impacts of ecological problems. Second, while the constructivist–realist divide has subsided, differing stances toward the utility of employing indicators of ecological conditions in empirical testing of theoretical perspectives are readily apparent, and hamper productive debate between agnostics and pragmatists. Third, and related to the second, the current major theoretical cleavage within the field is over the ability or inability of ecological modernization processes to halt environmental degradation, especially at the global level.

There are other aspects of environmental sociology worth noting, including the fact that over and above the just-noted differences, there is great diversity in methodological and theoretical approaches with strong bodies of work being produced by both qualitative and quantitative researchers, at the mirco- and macrolevel, and employing both constructivist and realist approaches. This diversity is enhanced by the geographical diffusion of the field, especially beyond North America and Europe (Dunlap, 2010). In Japan, for example, one can find scholars making heavy use of Western perspectives (Hasegawa, 2004) as well as developing indigenous ones (Torigoe, 2014). As environmental sociology becomes an ever more international field, one can expect this intellectual diversity to grow and produce beneficial effects.

There is also great variation in the degree to which environmental sociologists adopt multidisciplinary perspectives and engage in interdisciplinary projects. While some scholars focus on testing sociological theories such as WPT and do not venture outside of the discipline, others adopt research programs that inherently cross disciplinary boundaries (Fischer-Kowalski and Haberl, 2007). Given that the core focus of environmental sociology on societal-environmental interactions is now shared by a range of disciplines involved in 'coupled human and natural systems' research, and that many concerns of our field overlap with the growing area of 'sustainability science,' it is not surprising that environmental sociologists are actively engaged in these broader efforts (Liu et al., 2007).

In short, it seems appropriate to conclude that the field of environmental sociology, now in its fourth decade, has established itself as a respectable and intellectually vibrant field of study. The discovery of new ecological problems, and the recalcitrant nature of current ones, ensures that the field's subject matter will only grow in societal importance. One can only hope that as the field moves forward, it will be able to make more significant contributions to dealing with these problems, raising its profile beyond academia.

See also: Environment and Development; Environmental Attitudes and Behavior: Measurement; Environmental Justice in the United States; Environmental Movements; Industrial Ecology; Limits to Growth; Sustainability and Sustainability Science.

Bibliography

Bell, M.M., 2012. An Invitation to Environmental Sociology, fourth ed. Pine Forge Press, Thousand Oaks, CA.

Carolan, M.S., 2005. Society, biology and ecology: bringing nature back into sociology's disciplinary narrative through critical realism. Organization and Environment 18, 393–421.

Catton Jr., W.R., Dunlap, R.E., 1980. A new ecological paradigm for post-exuberant sociology. American Behavioral Scientist 24, 15–47.

Diekmann, A., Schmidt, P., 1998. Special section: testing rational choice models of behavior involving environmental issues. Rationality and Society 10, 77–78.

Dietz, T., Gardner, G.T., Gilligan, J., Stern, P.C., Vandenberg, M.P., 2009. The behavioral wedge: household actions can rapidly reduce U.S carbon emissions. Proceedings of the National Academy of Sciences 106, 18452–18456.

Dunlap, R.E., 2010. The maturation and diversification of environmental sociology: from constructivism and realism to agnosticism and pragmatism. In: Redclift, M., Woodgate, G. (Eds.), The International Handbook of Environmental Sociology, second ed. Edward Elgar, Cheltenham, UK, pp. 15–32.

Dunlap, R.E., Catton Jr., W.R., 1979. Environmental sociology. Annual Review of Sociology 5, 243–273.

Dunlap, R.E., Catton Jr., W.R., 1994. Struggling with human exemptionalism: the rise, decline and revitalization of environmental sociology. The American Sociologist 25, 5–30.

Dunlap, R.E., Jones, R.E., 2000. Environmental concern: conceptual and measurement issues. In: Dunlap, R.E., Michelson, W.H. (Eds.), The Handbook of Environmental Sociology. Greenwood Press, Westport, CT, pp. 482–524.

Dunlap, R.E., York, R., 2008. The globalization of environmental concern: an anomaly for postmaterialism. Sociological Quarterly 49, 529–563.

Fairbrother, 2013. Rich people, poor people, and environmental concern: evidence across nations and time. European Sociological Review 29, 910–922.

Fischer-Kowalski, M., Haberl, H., 1993. Metabolism and colonization: modes of production and the physical exchange between societies and nature. Innovation: The European Journal of Social Science Research 6, 415–442.

Fischer-Kowalski, M., Haberl, M. (Eds.), 2007. Socioecological Transitions and Global Change. Edward Elgar, Cheltenham, UK and Northampton, MA.

Foster, J.B., 2012. The planetary rift and the new human exemptionalism: a political-economic critique of ecological modernization theory. Organization & Environment 25, 211–237.

Foster, J.B., Clark, B., York, R., 2011. The Ecological Rift: Capitalism's War on the Earth. Monthly Review Press, New York.

Franzen, A., Meyer, R., 2012. Environmental attitudes in cross-national perspective: a multilevel analysis of the ISSP 1993 and 2000. European Sociological Review 26, 219–234.

Freudenburg, W.R., 1997. Contamination, corrosion and the social order: an overview. Current Sociology 45, 41–57.

Freudenburg, W.R., Gramling, R., Laska, S., Erikson, K.T., 2009. Catastrophe in the Making. Island Press, Washington, DC.

Gould, K.A., Lewis, T.L. (Eds.), 2014. Twenty Lessons in Environmental Sociology, second ed. Oxford University Press, New York.

Gross, M., Heinrichs, H., 2010. Environmental Sociology: European Perspectives and Interdisciplinary Challenges. Springer, Dordrecht.

Hannigan, J., 2006. Environmental Sociology: A Social Constructivist Perspective, second ed. Routledge, New York.

Harlan, S.H., Declet-Barreto, J.H., Stefanov, W.L., Petitti, D.B., 2013. Neighborhood effects on heat deaths: social and environmental predictors of vulnerability in Maricopa County, Arizona. Environmental Health Perspectives 121, 197–204.

Harper, C.L., 2012. Environment and Society, fifth ed. Pearson, Upper Saddle River, NJ.

Hasegawa, K., 2004. Constructing Civil Society in Japan: Voices of Environmental Movements. Trans Pacific Press, Melbourne, Australia.

Hong, D., Xiao, C., Lockie, S., 2014. China's economic growth and environmental protection: approaching a 'win-win' situation? In: Lockie, S., Sonnenfeld, D.A., Fisher, D.K. (Eds.), Routledge International Handbook of Social and Environmental Change. Routledge, London, pp. 45–57.

Jorgenson, A.K., Clark, B., 2011. Societies consuming nature: a panel study of the ecological footprints of nations, 1960–2003. Social Science Research 50, 226–244.

Jorgenson, A.K., Clark, B., 2012. Are the economy and the environment decoupling? A comparative international study, 1960–2005. American Journal of Sociology 118, 1–44.

Jorgenson, A.K., Dunlap, R.E., Clark, B., 2014. Ecology and environment. In: Sasaki, M., Zimmerman, E., Goldstone, J., Sanderson, S.A. (Eds.), Concise Encyclopedia of Comparative Sociology. Brill Academic Publishers, Leiden & Boston.

Knight, K.W., Messer, B.L., 2012. Environmental concern in cross-national perspective: the effects of affluence, environmental degradation, and world society. Social Science Quarterly 93, 521–537.

Liu, J., Dietz, T., Carpenter, S.R., et al., 2007. Complexity of coupled human and natural systems. Science 314, 1513–1516.

Macnaghten, P., Urry, J., 1998. Contested Natures. Sage, London and Thousand Oaks, CA.

McCright, A.M., Dunlap, R.E., 2010. Anti-reflexivity: the American conservative movement's success in undermining climate science and policy. Theory, Culture and Society 26, 100–133.

Mohai, P., Pellow, D., Roberts, J.T., 2009. Environmental justice. Annual Review of Environment and Resources 34, 405–430.

Mol, A.P.J., Sonnefeld, D., Spaargaren, G. (Eds.), 2009. The Ecological Modernization Reader: Environmental Reform in Theory and Practice. Routledge, New York.

Murphy, R., 2002. The internalization of autonomous nature into society. Sociological Review 50, 313–333.

O'Connor, J., 1991. On the two contradictions of capitalism. Capitalism. Nature, Socialism 2, 107–109.

Pais, J., Crowder, K., Downey, L., 2014. Unequal trajectories: racial and class differences in residential exposure to industrial hazard. Social Forces 92, 1189–1215.

Pellow, D.A., Brehm, H.N., 2013. An environmental sociology for the twenty-first century. Annual Review of Sociology 39, 229–250.

Redclift, M., Woodgate, G. (Eds.), 2010. The International Handbook of Environmental Sociology, second ed. Edward Elgar, Cheltenham, UK.

Roberts, J.T., Parks, B.C., Vasquez, A.A., 2004. Who ratifies environmental treaties and why? Institutionalism, structuralism and participation by 192 nations in 22 treaties. Global Environmental Politics 4, 22–64.

Schnaiberg, A., 1980. The Environment: From Surplus to Scarcity. Oxford University Press, New York.

Stern, P.C., Dietz, T., Abel, T., Guagnano, G.A., Kalof, L., 1999. A value-belief-norm theory of support for social movements: the case of environmentalism. Human Ecology Review 6, 81–97.

Torigoe, H., 2014. Life environmentalism: a model developed under environmental degradation. International Journal of Japanese Sociology 23, 21–31.

Yearley, S., 1991. The Green Case. Harper Collins, London.

Yearley, S., 2005. The sociology of the environment and nature. In: Calhoun, C., Rojek, C., Turner, B. (Eds.), The SAGE Handbook of Sociology. Sage, Thousand Oaks, CA and London, UK, pp. 314–326.

Yearley, S., 2009. Sociology and climate change after Kyoto: what roles for social science in understanding climate change? Current Sociology 57, 389–405.

York, R., Rosa, E.A., Dietz, T., 2003. Footprints on the earth: the environmental consequences of modernity. American Sociological Review 68, 279–300.

York, R., Rosa, E.A., Dietz, T., 2010. Ecological modernization theory: theoretical and empirical challenges. In: Redclift, M., Woodgate, G. (Eds.), International Handbook of Environmental Sociology, second ed. Edgar Elgar, Cheltenham, UK, pp. 77–90.

Environmental Stress and Health

John S Picou and Keith Nicholls, University of South Alabama, Mobile, AL, USA
Rainer Guski, Ruhr Universitat Bochum, Bochum, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R. Guski, volume 7, pp. 4667–4671, © 2001, Elsevier Ltd.

Abstract

This article examines the relationship of environmental stress to disturbances in behavior, mental health functioning, and public health for people impacted by a variety of ecosystem disruptions. The biophysical environment has enduring characteristics that are intimately linked to human social systems. Economic, social, cultural, and spiritual connections between the 'social' and 'biophysical' environments have been compellingly identified in the literature. Such exchange relationships between natural and human systems establish continuous dialectical interactions, which often results in changes to both. Extreme environmental stress in the form of catastrophic natural and technological disasters seriously undermines the human condition and quality of life in the modern world. We will briefly identify the context of contemporary risk and review the health consequences of extreme environmental stress.

Risk and Environmental Stress

The changing nature of risk in late modernity has been chronicled in the writings of Ulrich Beck (1992) and Kai Erikson (1994). Beck notes that 'manufactured' (human-generated) risks threaten the safety of the biophysical environment, while Erikson identifies the consequences of such risks for humans as a 'new species of trouble.' The relatively recent emergence of human-generated environmental risks and the extensive depletion of natural resources have produced stress and disruption to ecosystem services, social organizations, and the dynamic relationships between these entities (Kroll-Smith and Couch, 1993).

As defined by Beck (1992), modern environmental risks are characterized by new issues and concerns for human communities. First, these new environmental stressors are global, invisible, and often undetectable by human sensory perception. Second, they transcend generations; for example, massive amounts of radiation can impact future generations through an increased likelihood of birth defects. Third, such risks preclude causal attribution so that survivors can never be adequately compensated for damages (Beck, 1992). Downstream toxic contamination involves myriad sources of air, water, and soil pollution, while death, declines in public health, psychological pathology, and the loss of community are not compensable damages. This insidious 'collective trauma' invades the human condition on a permanent basis, separating the self from others, eroding trust and social relationships through the establishment of uncertainty and dread as social norms (Erikson, 1994).

Such anthropogenic environmental stress can alter traditional patterns of social organization by draining social capital and producing systemic corrosive communities (Freudenburg, 1997). This recognition of the potential untoward consequences of manufactured risks for both the natural and social environments has resulted in a reconsideration of the types of disasters that will characterize the twenty-first century. Traditionally viewed as 'natural' (meteorological and seismological events) or 'technological' (human-caused contamination), modern disasters now include a variety of hybrid 'natech' disasters, as well as acts of terrorism. Although terrorism is not a 'new' political event, the post-9/11 disaster landscape must include these purposeful, violent acts as human-caused destruction that involves the release of toxic chemicals and massive devastation of the built environment (Marshall et al., 2003). Furthermore, hurricanes, floods, and tsunamis will potentially impact chemical production and energy facilities, which, in turn, release toxins and radiation into the environment. Hurricane Katrina and the Fukushima tsunami are two obvious examples of these modern natech disasters (Picou, 2009a).

When we relate these catastrophic types of disasters to environmental stress, it becomes apparent that such events have enduring characteristics, which are beyond the control of residents of impacted communities. Although people may consider the toxic contamination produced by technological, terrorist, and natech disasters as harmful to their health and well-being, they can actually do very little to cleanup or reduce continuing radioactive and toxic emissions. The reaction by government and corporate authorities is often framed within a narrative that exposure poses little or no threat to ecosystem vitality or public health, and that the cleanup and containment of toxins, radiation, etc. has been successful. Such admonitions only exacerbate the fears and uncertainties of a distrusting public, resulting in long-term corrosion of the social fabric. Such secondary disasters stall timely ecological and community recovery from these extreme forms of environmental stress. In short, the consequences of modern technological, terrorist, and natech disasters become debated or contested in their aftermath. From official accounts of damages to protracted adversarial litigation, there is no consensus achieved on the final consequences of the event (Picou, 2009b).

A summary of these catastrophic types of disasters is presented in **Table 1**. Developed by Erikson (1994), this 2×2 property space identifies the cause of the disaster (natural or human) and type of destruction (toxic or nontoxic) as the defining characteristics of catastrophic environmental stress. Secondary disasters result from toxic contamination produced by both natural and human causes, and a continuing spiral of social impacts emerge (Gill, 2007). Indeed, human communities lose social capital, while residents experience

Table 1 Erikson's classification of disasters

	Cause	
Toxicity	*Human (technological)*	*Nature*
Nontoxic	Fires, dam collapses, airplane crashes, explosions	Hurricanes, floods, tornados, earthquakes
Toxic	Oil spills, toxic chemical spills, radiation leaks, toxic waste contamination, terrorism	Radon gas, natech scenarios – natural disasters that cause a technological disaster

Erikson, K.T., 1994. A New Species of Trouble: Explorations in Disasters, Trauma, and Community. W.W. Norton, New York.

deleterious medical and mental health consequences (Adeola and Picou, 2012, 2014).

The Context of Psychological Stress

Although there are many interpretations of how environmental risks and disasters impact people in the fields of medicine, sociology, and psychology, the Conservation of Resources (COR) stress model has been successfully used in research over the last 35 years (Hobfoll, 1989). As an explanatory framework, the COR model identifies resource loss, the threat of resource loss, and resource investment failure as primary determinants of psychological stress. This model considers four basic resources: (1) Objects (e.g., physical possessions, natural resources), (2) Conditions (e.g., good marriage; quality relationships), (3) Energies (e.g., money, knowledge), and (4) Personal characteristics (e.g., self-esteem, confidence). This comprehensive model predicts that psychological stress occurs when environmental risks and disasters alter values, produce extreme demands on individual and collective resources, overburden resource use, and produce strong mental damage (Hobfall, 1991). A variety of pathological psychological outcomes are predicted by the COR model, including anxiety, severe depression, posttraumatic stress disorder (PTSD), self-isolation, and anger. With technological, terrorist, and natech catastrophes comes acute uncertainty and dread on the part of survivors regarding the nature and extent of contamination and the future impacts of these forms of extreme environmental stress (Marshall and Picou, 2008).

Definitions of Environmental Stress

Although there are many different concepts of stress in the respective fields of medicine, psychology, and sociology, it is generally understood that stress is aversive for many people. Psychologists often employ the stress concept proposed by Lazarus (1966) and Lazarus and Launier (1978), which defines stress as a transaction between stimulus and response, i.e., the process that reflects the individual appraisal of a stimulus by persons who see an imbalance between environmental demands and their own response capabilities to cope adequately with the demands. A variety of mental health outcomes are associated with the recognition of loss or threat of loss to multiple types of resources conditioned by these stimulus/response transactions. These negative stimuli disrupt routine behavior and range from modest ambient stressors to cataclysmic events such as those noted above.

Characteristics of Environmental Stressors

Ambient stressors reflect continuous and intractable background characteristics of the physical environment (Campbell, 1983). Often unnoticed, the continuous hum of air conditioners, the permanent dust in an industrial area, and the faint hiss of the central heating system are common examples. People often adapt to ambient stressors, and they consider the costs of coping with such stressors to be more problematic than simply enduring them.

In addition, each type of environmental stressor can be categorized along several dimensions.

1. The major dimension is the degree of controllability over the environmental stressor (Campbell, 1983). Control may be either direct (e.g., switching off the neighbor's noisy lawn mover), indirect (e.g., closing the windows in order to reduce the noise penetration), social (e.g., calling authorities to stop the noise), or cognitive (e.g., knowing that the noise will stop in about an hour). Oftentimes, stress effects decrease with increased control, but control may have secondary stress effects (e.g., closing the windows will increase the temperature inside the house).
2. Related to cognitive control is the predictability of the stressor. High predictability of the occurrence and duration of an aversive agent may help people cope. For instance, residents in the vicinity of urban rail systems usually know the schedule of the trains and seldom complain about noise and vibration. Alternatively, knowing beforehand that a new airport will open in the neighborhood may cause greater stress effects than knowing that the old airport will close in the neighborhood (Hatfield et al., 1998).
3. Stressors may be more or less perceptually salient or identifiable. Some environmental agents are of rather low intensity and may either be not detected at all (e.g., low radioactive radiation), while others are only detected when intensity changes (e.g., the faint current of air conditioners). If an environmental agent is not perceptually salient, stress reactions may occur because people get informed about potential hazards. Such is the case for invisible and unknown toxic contamination.

Stressful life events also include incidents in the life cycle that require individual adaptive responses. Such events include changes in work, or residential environment, such as beginning a new job, moving to a new residential area, new construction work in the present residential area, or a perceptible change in the operating conditions of a nearby stressor. The event as such is usually short, but the behavioral consequences may extend through time.

Daily hassles are repeatedly occurring aversive events of ordinary life, such as arguments with colleagues, crowded classrooms, and traffic congestions on the route to work. Although they typically are more or less predictable, the individual has little means to avoid such hassles and their duration is rather short.

Variables Intervening between Stressors and Their Effects

Although varying degrees of environmental stress result in different social responses, it is also documented that the same stress exposure does not cause the same reaction in every person or in every situation. Situational, personal, and social variables determine whether an environmental agent becomes a severe problem for a person or a community.

There are variations in vulnerability to environmental stressors that characterize human populations (Donner and Rodriguez, 2008). The vulnerability model of environmental stress identifies socioeconomic status, gender, race, and age as examples of demographic vulnerability to extreme events (Thomas et al., 2013). On the other hand, the concept of resilience identifies the capacity of individuals, communities, and regions to resist, respond, absorb, and recover from extreme environmental stress. Resilient groups have the capacity to respond to environmental stress and recover from extreme disturbances in a timely fashion.

The Capital-Based Approach to Environmental Stress

Disaster researchers have identified five predominant forms of capital relevant to environmental stress: social, economic, physical, human, and natural (Portes, 1998; Aldrich, 2012). Social capital refers to the resources and abilities available to social groups through networks, trust, reciprocity, common values, and consensual rules (Bourdieu, 1986). Economic capital identifies the capacity to access financial resources, while physical capital refers to the scope and capacity of the built and modified environments. Natural capital refers to the abundance and availability of natural resources, while human capital refers to the coordination of all forms of capital to promote economic production. These forms of capital identify components for understanding vulnerability to extreme environmental stressors and the building of resilient responses to such events (Ritchie and Gill, 2007).

Case Studies of Catastrophic Environmental Stress

The following case studies briefly describe events that produced catastrophic environmental stress over the last 35 years.

Living in Contaminated Areas

Residential areas are sometimes built either on or in the vicinity of former industrial areas, and this can cause severe public health problems. One of the more infamous examples is Love Canal (Levine, 1982), a neighborhood in Niagara Falls, New York. This residential area was used by Hooker Chemical Company to bury 21 000 tons of assorted toxic waste. The emergence of chemicals in the air, water, and soil were detected in 1978, and health authorities warned the residents about the exposure risk. Over 900 families were evacuated. Despite attempts to cleanup this environmental tragedy, current residents have filed lawsuits and the physical and mental health impacts have had serious consequences for many residents (Levine, 1982). These negative outcomes include cancer, heart disease, anxiety, and severe depression.

A more recent example of residential contamination comes from the Wingate community in Fort Lauderdale, Florida. A municipal incinerator operated in this neighborhood for over 24 years and subsequent environmental testing revealed dioxin, mercury, arsenic, furans, and other carcinogenic compounds in the water and soil (Bevc et al., 2005). Research conducted 30 years after the closing of the incinerator found that physical symptoms of medical illness were related to exposure to chemicals, consumption of contaminated food, and mental health problems (Bevc et al., 2005). The human impacts of living in contaminated communities are serious for both the medical and mental well-being of residents and persist for extended periods of time (Edelstein, 2003).

Living with Nuclear Power Plants

Although the nuclear industry spends billions of dollars for public relations in the hope of changing public attitudes regarding the safety of nuclear power plants and nuclear waste, the general attitudes in Western societies remain skeptical. Such public responses are caused by the experience of technological catastrophes like Three Mile Island (TMI) in 1979 and Chernobyl in 1986. Seven years after the catastrophe at TMI, residents living near the damaged reactor reported more chronic stress symptoms and sleep disturbances than did control subjects (Davidson et al., 1987). The final health impacts of Chernobyl will never be known. However, increases in thyroid cancer have been documented in addition to those deaths reported in the first 4 months following the meltdown. Over 350 000 people were eventually evacuated from the highest radiation-contaminated areas. The psychological effects of these technological catastrophes on residents differ from those of natural hazards in many ways, e.g., they are more long lasting, and residents lose faith in public communications of authorities with respect to safety and control of health risks (Erikson, 1994).

The catastrophe at Chernobyl caused severe psychological changes in the European population, especially with respect to the evaluation of nuclear power, the probability of a nuclear accident, and the health risks for children (Drottz-Sjöberg and Sjöberg, 1990; Hüppe and Janke, 1994). Many social scientists believe that the inconsistent risk communication policies of public authorities has contributed to the stress symptoms of residents who were minimally exposed to direct nuclear radiation. At this time, the TMI-2 reactor is permanently out of operation and contains no radioactive materials or waste. The Chernobyl reactor is covered by a concrete sarcophagus, which will be replaced with a new, safer shelter in the future.

When Oil Invades Our Water

On 24 March 1989, the supertanker *Exxon Valdez* slammed into a well-marked reef in Prince William Sound, Alaska. Over 11 million gallons of crude oil were released into one of the most pristine marine ecosystems in North America, causing massive death and destruction of fish, birds, and a wide variety of marine mammal populations. Over 25 years of longitudinal research on ecological and social impacts clearly reveal severe consequences for both the natural and human environments. Four years after the *Exxon Valdez* oil spill, the

herring fishery collapsed, and to date this critical commercial and subsistence fishery has not recovered. Toxic impacts continue to be documented for sea otters, harlequin ducks, and many other species that are still being exposed to lingering oil in the spill area (Rice, 2009).

Research on the social and psychological impacts of the *Exxon Valdez* spill clearly reveal long-term economic, mental health, and community impacts. High levels of PTSD, depression, and community conflict persisted in Cordova, Alaska, for nearly 20 years after the spill (Picou, 2009b). As ecological impacts emerged over time and the protracted civil litigation continued until 2008, commercial fishers, Alaska Natives, and plaintiffs in the litigation were found to be at high risk for psychosocial damages (Picou et al., 2004; Gill et al., 2014).

On 20 April 2010, the Deepwater Horizon drilling platform, located 50 miles off the Louisiana Coast in the Gulf of Mexico, exploded, resulting in the release of approximately 200 million gallons of oil over a 3-month period. The immediate socioeconomic impacts for commercial and recreational fishing, tourism, and other coastal industries were devastating. Indeed, in the months after the explosion severe mental health impacts, similar to those of the *Exxon Valdez*, included uncertainty, PTSD, and suicides (Gill et al., 2011). The ecological consequences of this extreme form of environmental stress for the Gulf of Mexico will be monitored for many decades in the future. Furthermore, like the earlier *Exxon Valdez* spill, litigation will also continue well into the future, creating a context for prolonged disruption and stress for coastal residents.

Hybrid Forms of Environmental Stress

Natural disasters often have destructive consequences for chemical production systems, oil refineries, and local sources of toxic chemicals. This 'natural–technological synergy' results in technological failure, ecological contaminations, and toxic exposure for humans, creating a worst-case context for environmental stress. Air pollution, sediment contamination, and the movement of chemicals through the floodwaters of New Orleans from Hurricane Katrina created a classic natech disaster. In short, Hurricane Katrina was not only a massive meteorological event, but also a massive contamination event. The fact that Hurricane Katrina caused the release of 8 million gallons of oil is often overlooked by disaster researchers. At that time, this was the second largest oil spill in North American history, with only the 1989 *Exxon Valdez* releasing more oil (Picou, 2009a).

Earthquakes are another natural occurrence that can precipitate a natech disaster. Such was the case when a magnitude 9.0 earthquake struck off the coast of Japan on 11 March 2011. The quake generated a 50-foot tsunami which, in turn, damaged the Fukushima Daiichi nuclear power plant, resulting in a massive release of radioactive contamination. In addition to that released, a vast amount of radioactive water is being stored at the site. Cleanup and decontamination are expected to take decades. While there have been no deaths attributed directly to the radioactive release, hundreds died as a result of the ensuing evacuation and relocation of area populations. In addition, those exposed to radiation will experience severe stress and anxiety associated with the increased risk of cancer and birth defect for the rest of their lives and into the next generation.

Conclusions

Environmental stress must be framed in the context of modern risk. The global, uncontrollable, and invisible nature of contemporary risks produces extreme forms of environmental stress that threaten public health. The dual risks of natural and technological disasters will continue to challenge ecological resources and the social well-being of human populations throughout the twenty-first century.

See also: Ecology and Health; Globalization and Health; Injuries and Accidents: Psychosocial Aspects; Natural Disasters: Health-Related Aspects; Unemployment and Mental Health.

Bibliography

Adeola, F.O., Picou, J.S., 2012. Race, social capital, and the health impacts of Katrina: evidence from the Louisiana and Mississippi gulf coast. Human Ecology Review 19 (1), 10–24.

Adeola, F.O., Picou, J.S., 2014. Social capital and the mental health impacts of Hurricane Katrina: assessing long-term patterns of psychosocial distress. International Journal of Mass Emergencies and Disasters 32 (1), 121–156.

Aldrich, Daniel P., 2012. Building Resilience: Social Capital in Post-Disaster Recovery. The University of Chicago Press, Chicago.

Beck, U., 1992. Risk Society: Towards a New Modernity. Sage, London, UK.

Bevc, C.A., Marshall, B.K., Picou, J.S., 2005. Environmental justice and toxic exposure: toward a spatial model of physical health and psychological well-being. Social Science Research Journal. http://dx.doi.org/10.1016/j.ssreasaerch.2005.11.001.

Bourdieu, P., 1986. The forms of capital. In: Richardson, J. (Ed.), Handbook of Theory and Research for Sociology of Education. Greenwood Press, Westport, CT.

Campbell, J.M., 1983. Ambient stressors. Environment & Behavior 15, 355–380.

Davidson, L.M., Fleming, R., Baum, A., 1987. Chronic stress, catecholamines, and sleep disturbance at Three Mile Island. Journal of Human Stress 13, 75–83.

Donner, W., Rodriguez, H., 2008. Population composition, migration and inequality: the influence of demographic changes on disaster risk and vulnerability. Social Forces 87 (2), 1089–1114.

Drottz-Sjöberg, B.-M., Sjöberg, L., 1990. Risk perception and worries after the Chernobyl accident. Journal of Environmental Psychology 10, 135–149.

Edelstein, M., 2003. Contaminated Communities: Coping with Residential Toxic Exposure. Westview Press, Colorado.

Erikson, K.T., 1994. A New Species of Trouble: Explorations in Disasters, Trauma, and Community. W.W. Norton, New York.

Evans, G.W., Cohen, S., 1987. Environmental stress. In: Stokols, D., Altman, I. (Eds.), Handbook of Environmental Psychology, vol. 1. Wiley, New York, pp. 571–610.

Freudenburg, W.R., 1997. Contamination, corrosion and the social order: an overview. Current Sociology 45 (3), 19–40.

Gill, D.A., 2007. Secondary trauma or secondary disaster? Insights from Hurricane Katrina. Sociological Spectrum 27, 613–632.

Gill, D.A., Ritchie, L.A., Picou, J.S., 2014. Collaborative Research to Investigate and Document Social Impacts of High-Stakes Litigation Resolution in a Renewable Resource Community. Final report prepared for the National Science Foundation. Office of Polar Research, Washington, DC (ARC-0909497.).

Gill, D.A., Picou, J.S., Ritchie, L.A., 2011. The *Exxon Valdez* and BP oil spills: a comparison of initial social and psychological impacts. American Behavioral Scientist XX (X), 1–21. http://dx.doi.org/10.1177/0002764211408585.

Griefahn, B., Muzet, A., 1978. Noise induced sleep disturbances and their effects on health. Journal of Sound and Vibration 59, 99–106.

Hatfield, J., Job, R.F.S., Carter, N.L., Peploe, R., Taylor, R., Morell, S., 1998. Attitude-Mediated Reaction to Noise Influences Physiological Responses: Evidence Supporting Causality. Inter-Noise '98, Paper No. 440.

Hobfoll, S.E., 1989. Conservation of resources: a new attempt at conceptualizing stress. American Psychologist 44 (3), 513–524.

Hobfoll, S.E., 1991. Traumatic stress: a theory based on rapid loss of resources. Anxiety Research 4, 187–197.

Hüppe, M., Janke, W., 1994. The nuclear plant accident in Chernobyl experienced by men and women of different ages: empirical study in the years 1986–1991. Anxiety, Stress, and Coping 7, 339–355.

Kroll-Smith, J.S., Couch, S., 1993. Symbols, ecology & contamination: case studies in the ecological-symbolic approach to disaster. Research in Social Problems and Public Policy 5, 47–73.

Lazarus, R.S., 1966. Psychological Stress and the Coping Process. McGraw-Hill, New York.

Lazarus, R.S., Launier, R., 1978. Stress-related transactions between person and environment. In: Pervin, L., Lewis, M. (Eds.), Perspectives in Interactional Psychology. Plenum, New York, pp. 1–67.

Levine, A.G., 1982. Love Canal: Science, Politics, and People. Lexington Books, Lexington, MA.

Matthies, E., Höger, R., Guski, R., 2000. Living on polluted soil. Determinants of stress symptoms. Environment & Behavior 32, 270–286.

Marshall, B.K., Picou, J.S., Gill, D.A., 2003. Terrorism as disaster: selected commonalities and long-term recovery for 9/11 survivors. Research in Social Problems and Public Policy Journal 11, 73–96.

Marshall, B.K., Picou, J.S., 2008. Postnormal Science, precautionary principle, and worst cases: the challenge of twenty-first century catastrophes. Sociological Inquiry 78 (2), 230–247.

Miedema, H.M.E., Vos, H., 1999. Demographic and attitudinal factors that modify annoyance from transportation noise. Journal of the Acoustical Society of America 105, 3336–3344.

Picou, J.S., 2009a. Katrina as a natech disaster: toxic contamination and long-term risks for residents of New Orleans. Journal of Applied Social Science 4 (3), 39–55.

Picou, J.S., 2009b. When the solution becomes the problem: the impacts of adversarial litigation on survivors of the *Exxon Valdez* oil spill. University of St. Thomas Law Journal 7 (1), 68–88.

Picou, J.S., Marshall, B.K., Gill, D.A., 2004. Disaster, litigation and the corrosive community. Social Forces 82 (4), 1493–1522.

Portes, A., 1998. Social capital: its origins and applications in modern sociology. Annual Review of Sociology 24, 1–24.

Rice, S.D., 2009. Persistence, toxicity, and long-term environmental impact of the *Exxon Valdez* oil spill. University of St. Thomas Law Journal 7 (1), 55–67.

Ritchie, L.A., Gill, D.A., 2007. Social capital theory as an integrating framework for technological disaster research. Sociological Spectrum 27, 1–26.

Thomas, D.S.K., Philips, B.D., Lovekamp, W.E., 2013. In: Fothergill, A. (Ed.), Social Vulnerability to Disasters, second ed. CRC Press, Florida.

Winneke, G., Neuf, M., 1992. Psychological response to sensory stimulation by environmental stressors: trait or state? Applied Psychology: An International Review 41, 257–267.

Environment and Development

Chris J Barrow, College of Science, Swansea University, Swansea, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the evolution of attitudes, concepts and ethics concerning environment and development. It also defines 'environment' and 'development' and the evolution of environmentalism, the green movement and green politics. It further explores why environment and development interest took-off between the mid-1960s and 1974 and discusses global environmental change and the possibility of an environmental crisis. The coverage includes neo-Malthusians; Limits to Growth debate; sustainable development, the Precautionary Principle, the Polluter Pays Principle. The article also discusses changes in environment and development attitudes between 1972 and 1992 especially the growth of popular, business and media interest ("greening"). It examines growing worries about transboundary environment and development problems (e.g., global warming). It reviews achievements of the UN 1992 Rio 'Earth Summit,' including *Agenda 21*. It notes the growing awareness of environment- and development challenges have prompted development of environmental management.

Environment and Development

Since the 1950s people have become aware of problems like wildlife extinction and pollution. However, before the early-1970s few, except ecologists and other environmental scientists, were familiar with the terms 'environment' or 'development.' Today those terms are in wide usage and it is accepted that for development to be successful it must be environmentally sensitive (Barrow, 1995).

The Term 'Environment'

'Environment' may be defined as the total external conditions and influences affecting the life and fortunes of organisms, effectively 'the parameters of life.' It is shaped by living and nonliving elements that often interact in complex ways, making it a challenge to understand and predict future situations. The study of these interactions, 'ecology,' was founded in 1869 by Ernst Häckel, became established by the 1930s, and expanded from the 1970s. Ecology is often used loosely as a label for environmental concern.

Humans can adapt well to the environment they live in, and may be able to modify it to better suit their needs. But often they fail to improve or even sustain conditions, cause environmental degradation (local and even global) and then have to move away or suffer. A given environment may offer more opportunities than others but, ultimately, there are limits to human activity. If humans can satisfy their needs on a sustained basis without damaging the environment they may be said to have achieved 'sustainable development.' There are often a number of workable development strategies that could be pursued, some 'fit-in' with nature and at the other extreme some boldly change conditions. Wise development should also assess and prepare for environmental threats. Agriculture seeks to modify environments by inputs of energy (human labor or fuel-based mechanization), technology, and knowledge to sustain higher outputs of food or other crops.

People often talk of 'their local environment,' which is part of one global environment or biosphere. The biosphere can be subdivided into a vast number of more or less interacting 'ecosystems' (each a recognizably distinct assemblage of organisms adapted to particular environmental conditions). The ecosystem is the basic functional unit of ecology.

The Term 'Development'

Attempts to define 'development' reflect the current values of those making the definition, it is therefore difficult to agree one definition that is universally accepted and precise. Some see it as a learning process; others argue it is 'positive social transformation.' Another definition that is widely accepted is: *an ambiguous term for a multidimensional process involving material, social and organizational change, accelerated economic growth, the reduction of absolute poverty and inequality*. There is some overlap with *modernization*, so a group of people could develop yet remain unmodernized, or vice versa (Leys, 1996). Modernization is generally accepted to be a change toward a 'modern' economic, social and political system evolved in Western Europe and North America between the seventeenth century and present. Some argue that there has been a paradigm shift since the mid-1960s, from modern to *postmodern* development, involving less support for a mechanistic, reductionist and compartmentalized approach to the world's problems, in favor of a more *holistic*, multidisciplinary stance. (A researcher adopting a holistic approach seeks to study the whole, which may be more than the sum of the parts, so exploration of linkages and interactions is as important as studying component parts.)

Attitudes toward Environment and Development

Peoples' perception of development affects their exploitation of the environment. It is frequently argued that a faulty attitude toward development leads to environmental misuse and socioeconomic problems for humans (*see* Nature and Society in Geography).

Attitudes toward Environment and Development from Prehistoric Times to AD 1800

Before humans developed a sedentary lifestyle and their numbers increased beyond a few persons per square kilometer, environmental damage was largely temporary and limited. People moved on before pollution or erosion became too bad and, with the possible exception of some game animals, seldom overexploited resources. Although it has always been important to recognize natural threats and opportunities, the need to be concerned has grown as people have settled and multiplied, and as industry and technology have caused new demands and impacts.

People, with few alternatives to fall back upon, often evolved sustainable livelihood strategies and sound rules of environmental stewardship to ensure resources were not degraded. However, the poor may be forced to sacrifice sustainable livelihood and future opportunities to survive in the short term unless given aid and more affluent people may cause havoc through environmentally insensitivity.

In medieval Europe, the view was that the universe reflected Divine Perfection, consequently there was little attempt to think about development before the Renaissance. One of the first to consider the shaping secular human development was Thomas Moore, who published *Utopia* in AD 1516. By the mid-sixteenth century Europe's merchant class had grown in influence, possibly assisted by the spread of Calvinistic ethics (Weber, 1958). There was a growing, relatively liberal, capitalist economy and an increasing interest in science and rational study. From the mid-eighteenth century commerce, scientific reason, Industrial Revolution and urban migration had caused huge changes of attitude: a shift to the 'modern' outlook. This modern Western outlook is often accused of insensitivity toward the environment. However, there are environment and development problems in virtually all societies: in atheistic, Christian, Islamic, Buddhist and Hindu countries, regardless of whether the economy is socialist or free-enterprise capitalist.

Beyond the spreading cities, degraded farmlands and forests of Europe and eastern North America several islands began to suffer damage to biota and environment by the eighteenth century. So, by the 1760s efforts were made to try to protect forests on Tobago, Mauritius, and St. Helena, and some other islands. Even some oceanic fisheries, seal and whale stocks were under pressure by 1800. Naturalists in Europe and the Americas raised public interest in the environment (an early example being Gilbert White's 1798 study of nature) (for writings on the development of environmental attitudes see Wall, 1994). Nevertheless, if there were problems there was plenty of new land and resources to move on to.

Attitudes toward Environment and Development from AD 1800 to the 1940s

By AD 1800 the dominant attitude adopted in Western nations was that, whenever possible, the environment should be developed (exploited), and unspoiled areas were wasted or were even threatening. As the nineteenth century progressed the idea that humans could conquer environmental problems became established. Between 1830 and 1930 engineering and science made huge progress, but as pollution and population increased (see Ponting, 1991) and as scientists learnt more, it became clear that controlling the damage to environment would be difficult.

In Europe, Russia and North America growing cities, pollution and the clearance of forest and wilderness prompted a diverse group of intellectuals, romantics and 'protoanarchists' to question the way development was progressing (these included Thoreau, 1854; Kropotkin, 1899). Some saw the need to value nature for the inspiration and solace it could give.

The idea of safeguarding the environment against development existed in a localized ad hoc way before 1800, people protected certain plants or animals, establishing sacred groves, hunting parks, and enforcing taboos. However, by 1840 a growing number of naturalists, hunters, and foresters could see that what had once seemed limitless lands and resources were being degraded so they set up societies to promote larger-scale conservation and protection. National parks and reserves began to appear in the USA, Africa, India and other countries, and a number of forestry services were established. In the USA *preservationists* like John Muir worked to establish unspoilt wilderness reserves, and *conservationists* like Gifford Pinchot argued environmental protection could be combined with compatible resource exploitation (*see* Environmental Policy: Protection and Regulation). On both sides of the Atlantic nongovernmental organizations (NGOs) were established after 1830 to promote conservation and wildlife protection (like the Sierra Club – founded in 1892); some of these bodies played an important role after the Second World War in establishing powerful new environmental NGOs like the International Union for the Conservation of Nature and Natural Resources (IUCN).

Publication of Darwin's *The Origin of Species*… in 1859 helped established concepts like 'natural selection' and evolution, which encouraged some to argue that the human environment-development relationship was influenced or even controlled by physical conditions like climate. Race was also often seen as another determinant of development (*see* Environmental Determinism). These *environmental determinists* have been divided into the 'scientific' (e.g., Huntington, 1915), who tried to explore how nature might affect human fortunes, and those who promoted more 'crude and subjective' arguments. Supporters of *social Darwinism* argued for development to be shaped by competition and eugenics, rather than mutual aid and tolerance of the less developed (Pepper, 1984). Much discredited by the 1940s, it is probable that these concepts helped discourage social scientists from exploring environment and development issues until well into the 1970s (see Barrow, 2003: 12–31).

Serious drought and soil degradation in the US 'Dustbowl,' especially between 1932 and 1938, displaced many farmers. Novelist John Steinbeck drew on the misery caused by the disaster for his novel *The Grapes of Wrath* (1939), as did the folksinger Woody Guthrie; both were at first seen as subversives criticizing the American lifestyle (development dream) before they provoked public concern. The US Government had begun 'New Deal' measures to control regional land degradation by the mid-1930s, notably the foundation of the US Soil Erosion Service (in 1933) and its successor, the US Soil Conservation Service (in 1935). Just as the public in North America and Europe had started to see signs that there were environmental limits to development the Second World War and Cold War diverted attention and resources.

Attitudes toward Environment and Development after the 1940s

Between 1945 and the 1980s attention was focused on economic and industrial reconstruction, improving food production and the Cold War. Detonation of atomic bombs in 1945 signaled an ability to cause huge environmental damage. Major environment disasters took place: oil-spills, nuclear accidents, and other problems. The Cold War diverted attention and funds from environmental studies and development programs; but by triggering antinuclear weapons protest movements, it also helped to establish environmental politics in Europe. A few explored environment and development issues in the 1940s and 1950s (e.g., Teilhard de Chardin and Aldo Leopold), and some NGOs and public figures began to bring conservation and pollution issues to the attention of the public in North America and Europe. In 1949 the UN held one of the first postwar meetings, the Conservation Conference at Lake Success, New York State. In 1948 the IUCN, an NGO which has acted as a very influential network supporting conservation and environment and development issues was founded, now known as the Union for World Conservation, it continues to be important.

By the early-1960s attitudes toward environment and development had begun to change in Western nations, this is probably attributable to peoples' growing affluence and leisure time, development of the media, improved knowledge of the structure and function of nature (in part thanks to the exchange of information and research during the International Geophysical Year of 1957–1958 and the International Biological Program of 1964–1965), and diminishing colonial control. Rising levels of atomic weapons test fallout led by the 1960s to popular environmental fears. It has been suggested that the US Civil Rights Movement and 'hippie' culture encouraged people to ask questions about development and to value the environment more. Another significant development was that NGOs and groups of concerned individuals began to fight environmental issues in law courts. Interest in environmental matters 'took-off' in California, spreading to the rest of North America and Europe, reaching a peak between about 1966 and 1974. During this period what is variously described as 'environmentalism,' an 'environmentalist movement,' an 'environmental movement,' or an 'ecology movement' spread. A number of popular books helped provoke public interest, one of the most seminal and earliest being *Silent Spring* (Carson, 1962), which voiced concern about dichlorodiphenyltrichloroethane (DDT) pollution.

Today *environmentalism* is generally seen as a moral code or set of mediating values to guide human conduct. Environmentalism is often confused with *ecologism* which is more of a political ideology requiring its adherents to seek ethical and moral changes and to move away from giving priority to the welfare of humankind in any evaluation of development options. Environmentalism generally allows followers to put humans before nature and to seek managerial solutions rather than make far-reaching changes of development ethics and morals. Environmentalism encompasses a huge diversity of beliefs and attitudes, some pragmatic and willing to draw-upon technology (light green), others more romantic and suspicious of science. In general mid-1960s to mid-1970s environmentalism was dogmatic, rich in advocacy rather than workable proposals, and focused on the need for conservation, and the threats of overpopulation, pollution, corporate insensitivity and careless technical innovation (Farvar and Milton, 1972; Schumacher, 1973), it was also largely unpoliticized and restricted mainly to middle classes and intellectuals (*see* Political Ecology: Anthropological Perspectives). Many environmentalists, especially the *neo-Malthusians* (e.g., Ehrlich, 1972), warned of a coming 'crisis.' The feelings of 'crisis' were fanned when the USA and USSR space programs released pictures of Earth from space, prompting awareness that war and development took place within a finite, vulnerable system - 'Spaceship Earth.' At about this time the Club of Rome published *The Limits to Growth* (Meadows et al., 1972, 2004), which argued that in a finite world growth focused development was akin to cancer and zero-growth was needed. Amidst bitter debate this helped initiate a rethinking of development ethics and the nature of economics. Although it's roots lay in the 1930s the *Green Revolution* took-off in the 1960s, faltering and changing after 1973 (when oil-based inputs soared in cost). This was an initially clumsy attempt to use science and western development skills to raise agricultural production without clearing new land – i.e., intensification. Results were variable and after the 1990s the Revolution changed character as it learnt from mistakes and new knowledge, sustainability concerns and genetically modified organisms (GMOs) appeared (see Conway, 1997).

Warnings that the Earth faces an environmental and developmental crisis have increasingly been issued since the 1960s. The cause is identified as one or a combination of the following: population growth; cavalier use of nature; misapplication of technology; faulty development ethics. Environment and development problems can be expensive, socially damaging, or impossible to solve once they happen, it therefore makes sense to try to recognize and monitor crucial thresholds and act to prevent a crisis (a turning-point after which action is less effective or hopeless); yet there are risks in warning of threats and acting before there is proof unless a so-called 'win–win' approach can be found which gives benefit whatever the outcome. There are risks also that accepting there is a crisis might divert attention or promote undesirable 'crisis management,' rather than careful problem avoidance. Recognition of a crisis may be a mistaken overgeneralization and lead to failure to uncover causes.

In 1973 the UN Environmental Programme (UNEP) was established to promote environmental care worldwide. North America and Europe started to pass legislation in the late-1960s to develop more proactive planning which aims to reduce or avoid problems. One important examples of such legislation was the 1969 US National Environmental Protection Act (NEPA), among its achievements was the promotion of *environmental impact assessment* (EIA) procedures, and within 10 years it had been copied in most countries around the world. By 2010 EIA methods had improved and practitioners were policed by bodies like the International Association for Impact Assessment.

In spite of continuing environmentally damaging accidents, growing politicization of environmentalism, and the 1972 UN Conference on the Human Environment (held at Stockholm), popular interest dimmed between about 1974

and the late-1980s. The 1973–74 Organization of the Petroleum Exporting Countries (OPEC)-precipitated oil crisis and unsound international lending helped get many poor countries into debt by the 1980s, prompting increased exploitation of natural resources to earn foreign exchange which accelerated environmental damage also, as funds became scarce, poverty grew and the poor often resorted to environmentally damaging livelihood strategies. A response to debt was often structural adjustment which led to pressures on natural resources. Political changes in the USA by the late-1970s saw less government enthusiasm for expenditure on environmental protection, although there had been international agreement on a (nuclear weapons) Test Ban Treaty which reduced atmospheric fall-out levels.

Attitudes toward Environment and Development after the 1970s

At the time of what was effectively the first international ecology 'summit,' the 1972 UN Conference on the Human Environment at Stockholm, environmental matters were poorly reported, environmentalism was hardly politicized, few states had ministries or departments for the environment, and there was disinterest or even fear of calls for limits to growth and to protect nature. A number of representatives, especially those from developing countries, voiced the views that environmental care was a luxury their nations could not afford; that it was a ploy used by the rich to delay further development and subjugate or to exploit them; or, was hypocrisy coming from states which had damaged the environment to grow rich. When the 1992 UN Conference on Environment and Development met at Rio de Janeiro virtually all countries had ministries of environment; most delegates accepted environmental concern was a vital, integral part of development and not a luxury; the prevailing view was that rich and poor countries were interdependent parts of one global environment; and that practical and possibly painful measures must be taken to solve problems and promote more development in poor countries. In 20 years there had been a 'sea change.'

Certain publications helped achieve this shift, one was *Our Common Future*, the 'Brandt Report' (Independent Commission on International Development Issues, 1980), which spread the idea of developed and developing country interdependence eroding the view 'aid is charity.' Other influential publications included: the *World Conservation Strategy* (1980) and the 'Brundtland Report' (World Commission on Environment and Development, 1987). These publications rekindled public interest in environment and development issues; they also spread the concept of *sustainable development*. Sustainable development was more palatable than zero growth; it links aid, environment and development in a concept which, although difficult to give a precise and universal definition for, argues for stewardship to minimize environmental and social damage, efforts to ensure future options are not reduced and, if need be, aid to help achieve these aims.

During the 1980s the 'shorthand' *green* was increasingly applied to environmental issues. The meaning is vague, but in the main stands for environmentally friendly and environmentally sensitive, so 'greening' means 'improving environmental care,' and a 'green,' someone concerned for the environment (Adams, 2009). 'Greening' has marked the spread of environmental concern to a wider spectrum of Western (later other) societies, than in the 1970s. Interest spread to the public as a whole, and business, the legal profession, technology, the media, consumer protection bodies, marketing and advertising, economics and social studies, gender studies, tourism, theology and philosophy have 'gone green.' This is especially important with respect to economics and commerce, given the crucial role these have in development: an early green economics publication being *Blueprint for a Green Economy* (Pearce et al., 1989). By the late-1980s many bodies, including the World Bank and most aid agencies, had established environmental departments and green development guidelines.

Many interpret this greening as a new social movement: the *green movement*. Some see the green movement as a 'cultural attack' on the ills of modern society and economics, and it has been argued it is a manifestation of a postmodern and post-industrial cultural shift. The green movement (light green – deep green) is very diverse, but most would accept it embraces environmental sensitivity, social responsibility, grass-roots democracy, and nonviolence. However, greens may not be objective and even scientists may misjudge or apply received wisdom (see Lomborg, 2001) and sometimes business and institutions deliberately promote environmental falsehoods to look good ('greenwash').

Many greens scorn politics, or support whatever party they feel will best protect the environment; however, an aspect of greening has been the spread of environmental politics. Green politics first originated in New Zealand and West Germany with roots in antinuclear weapons protest, feminism, and the US Civil Rights Movement, early romantic environmentalists (like Henry Thoreau), and more recent writers on *social ecology*, environmental ethics (such as Arne Naess). Starting in Hamburg during the early-1970s, green candidates were elected to German State legislatures, and then the national Parliament. Early German activists include Petra Kelly and Rudolph Bahro. Only scatterings of greens have been elected to governments in Europe and other countries (virtually none in the USA), and if anything support has waned since the mid-1980s, however they triggered a greening (somewhat cynical) of established political parties. There is now scarcely a single country without green NGOs (increasingly networking globally), sympathetic newspaper, radio and television editors, and politicians interested in cultivating an environmentally sensitive image, and growing commercial interest in the environment.

Some recognize the threat to life and development of catastrophic natural events and advances in science. From the 1970s there has been more interest in terrestrial and extraterrestrial threats and the need to respect nature and its limits: in 1969 James Lovelock proposed the *Gaia hypothesis*, there are variants, but all argue that life on Earth has not simply adapted to physical conditions but is part of a complex interrelationship. If the hypothesis is accepted, the implication for development is that humans must 'fit in' and not upset mechanisms that maintain the environment or face catastrophic ecological change (*see* Gaia). In 2006 Lovelock stressed that Gaian limits had been exceeded but that avoidance of the catastrophic results was just possible. He argued technology would be vital in this and that cautious sustainable green

approaches might be too slow and so threatens survival (Lovelock, 2006).

Around 2000 many activists and institutions undertook stocktaking and some drew-up goals, notable among these were the Millennium Ecosystem Assessment - an international program launched in 2001 by the UN and completed in 2005 that aimed to provide information on environmental change and human well being. The findings indicated worrying damage to the biosphere that threatens future development. In 2000 the UN published the 'Millennium Development Goals 1990–2025.' By 2010 business and institutions were applying social responsibility and corporate environmental responsibility in their policymaking and the use of ecoauditing was expanding (e.g., ISO 14001 and the European Eco-Management and Audit System - EMAS).

Starting in the late-1980s and possibly catalyzed by the 1998 El Niño event (a periodic ocean/climatic fluctuation) and the 2004 Southeast (SE) Asian tsunami there has been a revived interest in environmental influences on development. This neoenvironmental determinism has drawn on archaeology, history and paleoecology: for example studies of climatic change impact on past civilizations (see Diamond, 2005). These studies might help reveal why adaptability is important and suggest ways people in the past deployed sustainable agriculture strategies that we might usefully reengineer today (such as using bio-char). By 2013 limited efforts were being made to develop early warning of meteorite, comet and asteroid strikes on Earth and of storm, volcanic and earthquake events.

At the time of writing it is apparent that 'development pressures' are damaging, or have destroyed, established livelihood strategies, leading to environmental damage and human misery. For example, the spread of sedentary agriculture has disrupted nonsedentary agriculturalists; economic policies may trigger land degradation; population growth or land development policies may promote settlement of unsuitable and vulnerable soils. Globalization has increased (once local activities are linked to and more and more dependant upon and at risk from distant manufacturing, research, transport, world trading patterns, influences, etc.). The world in the twenty-first century is more crowded and urbanized than ever before reducing resilience and making adaptation difficult. Environmental change and human actions might generate large numbers of ecorefugees, who on relocation may cause further environmental and socioeconomic problems (Black and Black, 1998). As traditional ways breakdown new ones must be found which improve and sustain human welfare without environmental damage. While the Cold War threat receded after 1985 there are new concerns: terrorism, rogue states and organizations acquiring nuclear, chemical, biological and computer warfare capability.

Governments, business, and NGOs increasingly adopt *environmental management* measures to try to ensure the *polluter pays* (i.e., whoever is responsible for environmental damage meets the cost of remedy) and minimize unwanted impacts, using precautionary techniques like environmental and social impact assessment, environmental accounting, and ecoauditing. Foreign aid is nowadays almost always conditional on the application of such measures to try and reduce or avoiding environment and development problems. Since the late-1980s concern about *transboundary environmental problems* has increased, these include acid deposition; stratospheric ozone depletion; chemical pollution; reduction of oceanic fish stocks, loss of biodiversity; and above all global warming. There has been progress in understanding the structure and function of the environment and in techniques for measuring and analyzing data-for example, satellite and drone remote sensing and geographical information systems (GIS) (for handling constantly updated data). The mobile phone is having a growing impact on how people perceive their environment, make their living and can react to environmental threats and degradation. At the Rio 'Earth Summit' in 1992, participating nations signed a Declaration on Environment and Development, Conventions on Biodiversity (protection) and Climate Change (control), and endorsed an environment and development action program: *Agenda 21* (Keating, 1993).

Since the 1990s fear has grown about environmental change, particularly global warming (greenhouse warming) and to a lesser extent sea-level rise, runaway changes to climate and ocean nitrification. Attempts to achieve international controls have initiated agreements like the Kyoto Protocol (started in 2005 and ongoing). Concern for climate change increasingly influences development efforts although some have ignored or resisted. There are increasingly controls on trade, technology, and taxation, aimed at reducing carbon dioxide and methane emissions that would have been unthinkable 30 years ago. Planning for global warming impacts is underway (e.g., Stern, 2009) and fear that present-day countries are less adaptable and more vulnerable to environmental change than were past peoples. Clearly, one development goal should be to seek better adaptability and resilience. Another twenty-first century challenge is from GMOs – microorganisms, plants, and animals that contain recombinant DNA and may have huge agricultural, healthcare or industrial potential. GMOs could raise food production, counter pollution and global change and have medical value, but if poorly managed they might threaten environment, human health and livelihoods. A treaty – the Cartagena Protocol on Biosafety was started in 2002 to try and improve assessment and management of GMO risks and potential.

Nations increasingly negotiate solutions for environment and development problems, but often back-off when it comes to spending money or foregoing profits. Governments and individuals are aware that consumption and technological advances can mean environmental damage and there are better monitoring and enforcement of controls but fear of an environmental crisis grows.

See also: Environmental Policy: Protection and Regulation; Sustainable Agriculture; Third World.

Bibliography

Adams, W.M., 2009. Green Development: Environment and Sustainability in a Developing World, third ed. Routledge, London.

Barrow, C.J., 1995. Developing the Environment: Problems and Management. Scientific & Technical. Longman, Harlow.

Barrow, C.J., 2003. Environmental Change and Human Development: Controlling Nature? Arnold, London.

Black, R., Black, D.R., 1998. Refugees, Environment and Development. Longman, Addison Wesley, London.
Carson, R., 1962. Silent Spring. Houghton Mifflin, Boston, MA.
Conway, G., 1997. The Doubly Green Revolution: Food for All in the 21st Century. Penguin, London.
Diamond, J., 2005. Collapse: How Societies Chose to Fail or Survive. Allen Lane, London.
Dresner, S., 2002. The Principles of Sustainable Development. Earthscan, London.
Ehrlich, P.R., 1972. The Population Bomb. Simon and Schuster, New York.
Farvar, M.T., Milton, J.P. (Eds.), 1972. The Careless Technology. Garden City Press, New York.
Houghton, J., 2004. Global Warming: The Complete Briefing, third ed. Cambridge University Press, Cambridge.
Huntington, E., 1915. Civilization and Climate. Yale University Press, New Haven, CT.
Independent Commission on International Development Issues, 1980. North-South: A Programme for Survival. Pan, London.
International Union for Conservation of Nature and Natural Resources, UNEP & WWF, 1980. World Conservation Strategy: Living Resources for Sustainable Development. IUCN, Gland, (Switzerland).
Keating, M., 1993. The Earth Summit's Agenda for Change: A Plain Language Version of 'Agenda 21'. Earthscan, London.
Kropotkin, P., 1899. Fields Factories and Workshops. Hutchinson, London (original in Russian); English edn. 1974 ed. by C. Ward.
Leys, C., 1996. Rise and Fall of Development Theory. Indiana University Press, Bloomington, IN; James Currey, Oxford.
Lomborg, B., 2001. The Sceptical Environmentalist: Measuring the Real State of the World. Cambridge University Press, Cambridge.
Lovelock, J., 2006. The Revenge of Gaia: Why the Earth Is Fighting Back – and How We Can Still Save Humanity. Allen Lane, London.
Meadows, D.H., Meadows, D.L., Randers, J., Behrens III, W.W., 1972. The Limits to Growth (A Report for the Club of Rome's Project on the Predicament of Mankind). Universe Books, New York.
Meadows, D.H., Randers, J., Meadows, D.L., 2004. The Limits to Growth: The 30-Year Update. Earthscan, London.
Pearce, D., Markandya, A., Bastier, E., 1989. Blueprint for a Green Economy. Earthscan, London.
Pepper, D., 1984. The Roots of Modern Environmentalism. Croom Helm, London.
Ponting, C., 1991. A Green History of the World. Sinclair-Stevenson, London.
Schumacher, F., 1973. Small Is Beautiful: A Study of Economics as if People Mattered. Briggs, London.
Stern, N., 2009. A Blueprint for a Safer Planet: How We Can Save the World and Create Prosperity. Bodley Head, London.
Thoreau, H.D., 1854. Walden; or, Life in the Woods. Ticknor and Fields, Boston (later published as Walden. Penguin, New York, 1983).
Wall, D., 1994. Green History: A Reader in Environmental Literature, Philosophy and Politics. Routledge, London.
Weber, M., 1958. The Protestant Ethic and the Spirit of Capitalism (English translation by T. Parsons-originally published in 1904 and 1905). Charles Scribner, New York.
Welford, R., 2000. Corporate Environmental Management: Towards Sustainable Development. Earthscan, London.
White, G., 1977[1798]. A Natural History of Selborne. Penguin, Harmondsworth.
World Commission on Environment and Development, 1987. Our Common Future (The Brundtland Report). Oxford University Press, Oxford.

Relevant Websites

http://www.greenbusinessnetwork.org.uk/ – Green Business Net.
http://www.iied.org – International Institute for Environment and Development (IIED) – influential NGO.
http://www.iisd.org – International Institute for Sustainable Development (IISD).
http://www.ipcc.ch – Intergovernmental Panel on Climate Change (IPCC).
http://jed.sagepub.com/ – Journal: The Journal of Environment and Development. Sage.
http://www.undp.org/content/undp/en/home/mdgoverview.html – The Millennium Development Goals.
http://www.unep.org/maweb/en/index.aspx – The Millennium Ecosystem Assessment.

Environment and Health

Brian King, The Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the relationships between human populations and the natural and physical environment. Because of the coupled relationships between human health and the environment, global institutions such as the World Health Organization have intervened to prevent infectious and noncommunicable diseases by attending to their social and ecological determinants. It is generally accepted that human health depends upon a sustainable environment that provides opportunities for healthy decision-making and effective disease management. Multiple scholarly disciplines attend to health and environment interactions, including the fields of environmental health, social epidemiology, medical geography, and political ecology. Additionally, activists and policy-makers have received international attention by showing how human health is affected through disproportionate exposure to positives and negatives within the surrounding environment. Whether this means proximity to facilities that produce toxins and pollutants or access to green space and bikeways that promote healthy lifestyles, these factors have contributed to the linking of health with concerns for social and environmental justice.

Human health is intricately connected to the environment. This recognition includes the natural environment and a range of biotic and abiotic factors, as well as the physical or human-built environment that facilitates the spread of infectious disease or the production of conditions that influence noninfectious, or noncommunicable, disease. As one example, malaria is transmitted by *Anopheles* mosquitoes that serve as the main vector for human populations. These mosquitoes are highly dependent upon the natural environment, specifically in terms of precipitation and temperature that can alter the environmental conditions that allow for their survival. Recent work has shown that daily fluctuations in temperature affect parasite infection, the rate of parasite development, and components of mosquito biology that combine to determine the intensity of malaria transmission (Paaijmans et al., 2010). Daily temperature changes have also been shown to play a role in dengue virus transmission by the *Aedes aegypti* mosquito (Lambrechts et al., 2011). Specifically, mosquito life span and susceptibility to dengue infection decreases when there are larger fluctuations in daily temperature as compared to only moderate daily temperature variations. The implication is that human and nonhuman species are directly impacted by the natural environment. Additionally, these coupled relationships are dynamic and can shift because of variations in temperature, precipitation, or ecosystem functioning.

Other infectious diseases with an unknown etiology are credited with being produced by the natural environment. The pathogen *Mycobacterium ulcerans* is the causal agent for Buruli ulcer disease, which is one of the most frequently occurring human mycobacterial diseases (Qi et al., 2013). Ongoing research is working to show how anthropogenic impacts on the natural environment, particularly land cover disturbance and climate change, could trigger human exposure to the pathogen. In addition to the ecological patterns, Buruli ulcer is also potentially involved with social systems such as livelihood production and settlement patterns. These three examples of malaria, dengue fever, and Buruli ulcer disease show that the transmission of infectious disease to human beings is related to nonhuman species, temperature, precipitation, and other drivers within the surrounding environment.

International agencies emphasize the role of the natural and physical environment in producing hazards such as those caused by unsafe drinking water or exposure to toxins and air pollutants. According to the World Health Organization, within developing countries, the primary environmentally caused diseases are diarrheal disease, lower respiratory infections, unintentional injuries and accidents, and malaria. Within wealthier and industrialized countries, cancer, cardiovascular disease, asthma, lower respiratory infections, and traffic injuries are major health hazards. According to the World Health Organization Public Health & Environment Global Strategy Overview (2011), environmental hazards are responsible for roughly a quarter of the total burden of global disease. In developing countries, the burden of environmental hazards is more strongly experienced by communicable diseases, such as malaria, dengue fever, and human immunodeficiency virus (HIV). In developed countries, environmental hazards are believed to have a larger impact upon noncommunicable diseases.

The central areas of emphasis for the World Health Organization in dealing with public health and the environment include improvements in water, sanitation, and hygiene to prevent waterborne diseases such as cholera. Indoor air pollution deriving from fuels and cook stoves, in addition to outdoor air pollution resulting from urban transportation and population growth, are also emphasized as requiring intervention and policy responses. Lastly, the World Health Organization notes the need for chemical regulation policies that reduce exposure from industrial and agricultural activities (World Health Organization, 2011). These types of environmental health dynamics attest to the role of unsustainable development practices in making certain populations vulnerable to illness.

The United Nations has recently emphasized the role of noncommunicable human diseases as the primary cause of death and disability in the world. Roughly two-thirds of annual

deaths are caused by cardiovascular diseases, cancers, chronic respiratory diseases, and diabetes. Because many of these can be prevented, public health experts emphasize the importance of reducing tobacco use and alcohol, unhealthy diets, and a lack of exercise. In terms of noncommunicable disease, there are a variety of behavioral and physical environmental conditions that support their expansion. There has been a growth of popularly written books that link the relationships between food production and consumption patterns with human well-being. Michael Pollan (2006), Eric Schlosser (2001), and Marian Nestle (2006) have received attention for suggesting that changes to food decision-making are needed to reduce obesity, dependence upon pesticides and oil, and support local economies and environmental systems. Central to this argument is that industrialized food production in the United States, particularly the dependence upon corn, has triggered infrastructure and cultural shifts that have repercussions for human health.

Environmental Health

A number of academic and policy fields study health and environment interactions with the intention of detailing effective interventions to ensure sustainable and viable possibilities for human well-being. Infectious diseases are connected to the natural environment and changes to global climate dynamics. The field of environmental health examines how human health is affected by the natural and physical environment. The disciplines of environmental epidemiology, toxicology, and exposure science are often recognized as central to environmental health. Environmental epidemiology studies the relationship between environmental exposures, such as chemicals, radiation, microbiological agents, and human health. Toxicology studies show how these exposures lead to specific health outcomes. Finally, exposure science examines human exposure to environmental contaminants by both identifying and quantifying exposures. In addition to these research fields, social science disciplines, such as social epidemiology, medical anthropology, and public health, have been influential in showing the specific connections between the natural environment and human health. The World Health Organization (2014) defines environmental health as the "physical, chemical, and biological factors external to a person, and all the related factors impacting behaviours. It encompasses the assessment and control of those environmental factors that can potentially affect health. It is targeted towards preventing disease and creating health-supportive environments. This definition excludes behaviour not related to environment, as well as behaviour related to the social and cultural environment, and genetics."

The field of epidemiology is also credited with making connections between human health and the surrounding environment. The work of Dr John Snow is widely recognized as advancing the science of germ theory through the identification of the cause of a cholera outbreak in Soho, England, in 1854. Snow mapped the incidence of the outbreak and interviewed local residents to identify the source as a water pump on Broad Street. The dismantling of the pump and subsequent cessation of the outbreak helped establish epidemiology as the foundation for public health while evidencing the relationships between the environment and human health.

Urbanization

The built environment is integral to human health. Within the United States, the expansion of the suburbs following World War II necessitated transportation to allow individuals to travel to employment and commercial centers. This transportation was largely the automobile with one consequence being an increasingly sedentary lifestyle that is credited with contributing to health challenges, such as obesity. Urbanization is also occurring on the global scale, with more than half of the world's population now residing in an urban area. The World Health Organization (2011) estimates that by 2050, 70% of the world's population will be living in towns and cities. While potentially increasing access to health care services and environmental benefits, in a public health bulletin the World Health Organization notes that urbanization could concentrate health risks while generating new hazards. Globally, road traffic injuries are the ninth leading cause of death, and most road traffic deaths occur in low- and middle-income countries. Urban air pollution kills around 1.2 million people globally each year, primarily due to cardiovascular and respiratory diseases. A significant proportion of urban air pollution is caused by motor vehicles, although industrial pollution, electricity generation, and in least developed countries, household fuel combustion are also major sources. Infectious disease, such as tuberculosis, can concentrate in large urban centers. In New York City, tuberculosis incidence is four times the national average. In the Democratic Republic of the Congo, 83% of people with tuberculosis live in cities. Lastly, the World Health Organization notes that urban environments are not always designed to support physical activity or promote healthy food consumption. Participation in physical activity is constrained by a variety of urban factors including overcrowding, high-volume traffic, heavy use of motorized transportation, poor air quality, and lack of safe public spaces and recreation facilities.

Another potential health effect with increasing urbanization is heat stress. Wilhelmi et al. (2013) assert that the urban heat island (UHI) effect is a leading cause of weather-related mortality in the United States and elsewhere. Different groups within the population are more vulnerable, specifically the young and elderly, and also those without access to air conditioning or other services that could mitigate the effects. It is anticipated that urban heat stress will intensify in the future with global climate change and an increasingly urbanized population.

While the global population increasingly urbanizes, there are still rural populations whose health and quality of life depend upon services that are not always available. In fact, one of the motivations for migration to urban environments is to seek out employment and social services that can positively influence human health. Within rural communities in the United States, there exist variations in proximity and ability to access health care centers, such as clinics or hospitals. One response has been the establishment of mobile clinics that extend the range of services for these populations.

Climate Change and Health

Anthropogenic climate change represents a significant challenge for the natural environment and human health. Variations in the amount, intensity, and duration of precipitation can alter the environmental conditions that allow certain disease vectors to survive. Additionally, temperature changes can shift the latitudes for ecosystems to support the movement of species into new areas. The Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (Smith et al., 2014) includes a chapter that addresses the links between global climate change and human health. The IPCC recognizes three pathways through which climate change affects health. The first pathway includes direct impacts that are tied to changes in the frequency of extreme weather including heat, drought, and heavy rain. The second pathway considers effects mediated through natural systems such as the role of disease vectors, water-borne diseases, and air pollution. As evidence of these effects, the report acknowledges that the spread of malaria in parts of Africa, or the West Nile Virus in the United States, demonstrates how infectious disease can accompany changing weather patterns. Lastly, the third pathway addresses effects heavily mediated by human systems, such as occupational impacts, undernutrition, and mental stress.

Climate change and health scenarios within the IPCC report suggest that there will be greater risk of injury, disease, and death due to more intense heat waves and fires. The report notes with high confidence that there will be increased risk of undernutrition stemming from reductions in food productivity and increased risks of food and waterborne disease and vector-borne diseases. The IPCC report also recognizes the social dynamics influenced by climate change that shape human health. Human populations could be disrupted due to rising sea levels, reduced access to environmental services, and agricultural decline. The report notes that until the mid-century climate change will most likely intensify existing health challenges in ways that affect the most vulnerable. The most vulnerable populations will likely have reduced work capacity and less productivity, which can cause a downward spiral into increased socioeconomic poverty. As such, the IPCC recommends addressing underlying social vulnerabilities to environmental change by investing in policies to increase access to clean water and sanitation, improve health care services including vaccination and child health services, reduce socioeconomic poverty, and improve the ability of populations to respond to disasters. The achievement of these goals will require substantial planning given that the IPCC recommends redesigning transport systems to promote active transport and reduce dependency upon motorized vehicles.

Political Economy and Environmental Health

Because health and environment interactions are mediated by social systems, the role of political, economic, cultural, and gendered dynamics in shaping health vulnerabilities and environmental sustainability have been addressed by a number of academic disciplines. The field of social epidemiology has demonstrated the social determinants of health and how these vary within settings. Medical anthropology has made a number of contributions on health and the built environment with one widely recognized element being how variations in access to disease and quality health care constitute structural violence against a population (Farmer, 1999). The field of medical geography has drawn upon disease ecology to examine how the natural environment, including features such as topography, vegetation, and climate, interact with vector and pathogen ecology to produce disease. These fields have been valuable in showing the role of social and ecological determinants in shaping human health and the ways that vulnerabilities vary across spatial and temporal scales.

Attending to the underlying conditions that contribute to the spread of infectious disease or facilitate noncommunicable conditions, scholarship has focused upon the structural and political economic dynamics that produce health inequalities. An early proponent of this approach was Meredith Turshen (1977: p. 48) who argued that disease ecology viewed economic and political processes to be irrelevant, and therefore suffered "from a failure to consider the relation of people to their environment in all its complexity." Building upon this with an examination of health and environment interactions in Tanzania, Turshen (1984) utilized a political ecology of disease approach to assert that historical and spatial systems were integral to human health and well-being. Specifically, she argued that the imposition of colonialism upon Tanzania transformed social and environmental systems in fundamental ways that would have lasting effects following the country's independence in 1962.

Within the social sciences, the research field of political ecology has documented the ways that political and economic systems shape environmental resource use and management. Political ecology has been identified as having five primary themes (Robbins, 2012), with one being the positioning of nonhuman actors into human systems to understand the resulting impacts for disease management. Political ecology has provided greater attention to environment and health interactions in recent years while drawing upon related scholarship from medical anthropology and health geography (King, 2010). Central to this wave of research is attending to the role of political and economic systems in producing disease vulnerabilities and differential health management capabilities. Disease discourses, including the meanings and practices that are inscribed within particular disease patterns, are also addressed by political ecology. Lastly, political ecologies of health also emphasize the coupled relationships between social and ecological systems in producing human health and well-being.

The reciprocal dynamics between social and ecological systems has been demonstrated with the HIV/AIDS epidemic. While research and policy have shown the ways in which the disease has impacted demographic patterns, national economies, and agricultural production, there are also ecological effects associated with the spread and management of HIV. Talman et al. (2013) identify a number of possible interactions between HIV and the natural environment, including the selling of livestock, land fragmentation, agricultural decline, and increased dependence upon natural resources as a coping mechanism. One of the emerging findings from current scholarship is the importance of food security in providing needed nutrition and vitamins to those infected with the HIV

virus. In work from central Mozambique, Kalofonos (2010) shows that adherence to antiretroviral therapy is tightly coupled with food needs and intracommunity tensions about who receives external assistance. While international public health and development organizations deserve recognition for increasing access to life-saving medications, the gamut of needs involved with HIV management indicate multifaceted interactions with the natural environment.

An additional way in which changes in the natural environment and human health are coupled is through an understanding of coproduction between humans and nonhuman species. Rather than focusing upon a unidirectional set of relationships, coproduction evaluates the ways in which nonhuman nature influences social systems. Scholarship has addressed how interactions with nonhuman species, such as a mosquito, results in adjustments by humans to manage disease outbreaks and public sentiment. In a case study addressing the management of *A. aegypti* and *Culex quinquefasciatus* mosquitoes in the state of Arizona, it has been shown that concerns about West Nile Virus have prompted different state responses ranging from surveillance and adulticide, GIS monitoring and larvacide, and education and public awareness (Robbins and Miller, 2013). As another example of this coproduction, Scott et al. (2012) examined the relationships between humans and pathogens, which they suggest are mutually conditioned due to continuous changes in exposures and institutional processes.

Health within Changing Environments

The post-World War II period marked the onset of emerging environmental concerns and legislation that were significant for coupling environmental health with human health. Particularly notable were the contributions from Rachel Carson who raised public awareness about the human and environmental dangers from pesticides in industrial agriculture. In *Silent Spring* Carson used her scientific training to argue that chemicals were becoming widely used for controlling disease and pests. She argued that the reliance upon these pesticides within the industrial agriculture, in addition to practices such as single crop farming and imported species, made human and nonhuman species vulnerable. Carson focused her attention upon dichlorodiphenyltrichloroethane (DDT) which had been converted from a preventative for malaria and typhus among soldiers to an agricultural insecticide following the conclusion of the war. Among the species directly impacted by DDT were ospreys and bald eagles, whose ability to reproduce was negatively affected due to the concentration of the pesticide in the food chain. When the bald eagle was adopted as the nation's symbol in 1782 there were 25 000–75 000 nesting pairs, however, by the early 1960s there were fewer than 450 pairs. The banning of DDT in 1972, coupled with the passage of the Endangered Species Act in 1973, is largely credited with the recovery of the bald eagle and other raptor populations.

In the wake of Carson's warnings and the first United States Earth Day in 1970, major pieces of environmental legislation were passed to protect the health and sustainability of the natural environment. These include the National Wilderness Act (1964), the Wild and Scenic Rivers and National Trails Act (1968), the Clean Air Act (1963), the Clean Water Act (1972), and the Safe Drinking Water Act (1974). The consequence has been an expansion of traditional environmental protection in the United States for particular landscapes, such as the creation of Yellowstone National Park in 1872, toward policy engagements that couple the possibilities for human health with the sustainability of the natural environment.

Environmental Justice

The relationships between the environment and human health are not equally experienced. The rise of case studies that show the disproportionate exposure to environmental hazards, such as proximity to facilities that produce toxins or pollutants, have helped establish links between health and social and environmental justice. One of the first incidents that brought public attention to environmental hazards was the Love Canal incident. Love Canal is a community in upstate New York that was built over 21 000 tons of toxic waste that had been dumped by Occidental Petroleum Corporation, which was previously Hooker Chemical Company. Historic rainfall caused the improperly stored containers to leach into the basements of 100 homes and a public school. This rainfall generated widespread concern because the landfill was believed to contain 82 different compounds of which 11 were suspected carcinogens at the time, including benzene. Declared a federal disaster area, residents organized to receive assistance in relocating, which was a process that unfolded for several years. In addition to raising public awareness as to the potential health hazards from unregulated exposure to toxins, Love Canal provided the catalyst for the establishment of Superfund Legislation that empowers the Environmental Protection Agency (EPA) to list and remediate toxic waste sites.

Following Love Canal, two cases helped advance the argument that disproportionate exposure to landfills and pollutants should be considered environmental racism. The announcement in 1978 that a landfill to store polychlorinated biphenyls (PCBs) was to be placed in Warren County, North Carolina, provoked widespread community protests against the decision. The Afton community, which was slated to be directly affected by the creation of the landfill, was situated in a region of high socioeconomic poverty and contained a largely African American population. Community residents organized nonviolent protests and attempted to use the court system to overturn the decision; however, the landfill opened in 1982. The second case was the Whispering Pines Sanitary Landfill which was intended to be placed in a largely African American neighborhood in Houston, Texas. In writing about the case, Robert Bullard (2005) noted that although African Americans made up 25% of the city's population at that time, the five city-owned landfills (100%) and six of the eight city-owned incinerators (75%) were in African American neighborhoods. Working with residents, Bullard and others at the Texas Southern University helped file a lawsuit to stop the landfill from opening. *Bean v. Southwestern Management Corp*. was the first lawsuit in the United States that charged environmental discrimination in the siting of a waste management facility. After several years in court it was ultimately unsuccessful, however, it helped set a precedent in using the Civil Rights Act to allege environmental racism.

The United States Environmental Protection Agency defines environmental justice as the "fair treatment and meaningful involvement of all people regardless of race, color, national origin, or income with respect to the development, implementation, and enforcement of environmental laws, regulations, and policies. EPA has this goal for all communities and persons across this Nation. It will be achieved when everyone enjoys the same degree of protection from environmental and health hazards and equal access to the decision-making process to have a healthy environment in which to live, learn, and work." Since the signing of Executive Order 12898 in 1994, the federal government is charged with making environmental justice part of its regular operations. The mainstreaming of environmental justice in U.S. federal policy has not been seamless, but it does present an opportunity for agencies to address inequities in exposure to environmental hazards that contribute to reduced opportunities and poor health.

Environmental justice considers the relationships between health and the environment and also the variations that exist within settings. In conclusion, environmental justice is not simply a function of disproportionate exposure to hazards but it also examines inequitable access to characteristics in the natural and physical environment that contribute positively to human health. The idea of a food desert has become common in scholarly and activist communities in showing how, due to planning and political economic processes, certain areas can lack access to full service grocery stores that offer healthy produce. Within urban environments there can be variations in access to green space that provides opportunities for recreation and contemplation. The recognition of these types of inequities has expanded traditional understandings of environmental justice to address the underlying factors that shape variations in exposure to both the positive and negative elements contributing to human health and well-being.

See also: Environment and Development; Environmental Education and Education for Sustainable Development; Environmental Justice in the United States; Environmental Movements; Environmental Policy: Protection and Regulation; Health Education and Health Promotion; Planning Issues and Sustainable Development; Social Ecology; Sustainable Development.

Bibliography

Bolton, A.T., Bolton, S., Walson, J.L., 2013. Interactions between HIV/AIDS and the environment: toward a syndemic framework. American Journal of Public Health 103 (2), 253–261.

Bullard, R.D., 2005. Neighborhoods 'zoned' for garbage. In: Bullard, R.D. (Ed.), The Quest for Environmental Justice: Human Rights and the Politics of Pollution. Sierra Club Books, San Francisco, pp. 43–61.

Carson, R., 1962. Silent Spring. Houghton Mifflin, New York.

Environmental Protection Agency, 2014. Environmental Justice. http://www.epa.gov/environmentaljustice/ (accessed 07.07.14.).

Farmer, P., 1999. Infections and Inequalities: The Modern Plagues. University of California Press, Berkeley, CA.

Kalofonos, I.A., 2010. "All I eat is ARVs": the paradox of AIDS treatment interventions in central Mozambique. Medical Anthropology Quarterly 24 (3), 363–380.

King, B., 2010. Political ecologies of health. Progress in Human Geography 34 (1), 38–55.

Lambrechts, L., Paaijmans, K.P., Fansiri, T., Carrington, L.B., Kramer, L.D., Thomas, M.B., Scott, T.W., 2011. Impact of daily temperature fluctuations on dengue virus transmission by *Aedes aegypti*. Proceedings of the National Academy of Sciences 108, 7460–7465.

Nestle, M., 2006. What to Eat. North Point Press, New York.

Paaijmans, K.P., Blanford, S., Bell, A.S., Blanford, J.I., Read, A.F., Thomas, M.B., 2010. Influence of climate on malaria transmission depends on daily temperature variation. Proceedings of the National Academy of Sciences 107, 15135–15139.

Pollan, M., 2006. The Omnivore's Dilemma: A Natural History of Four Meals. Penguin Press, New York.

Qi, J., Campbell, L.P., Van Ravensway, J., Finley, A.O., Merritt, R.W., Benbow, M.E., 2013. Buruli ulcer disease: the unknown environmental and social ecology of a bacterial pathogen. In: King, B., Crews, K.A. (Eds.), Ecologies and Politics of Health. Routledge, London, pp. 75–97.

Robbins, P., 2012. Political Ecology: A Critical Introduction, second ed. John Wiley & Sons Ltd, Malden, MA.

Robbins, P., Miller, J.C., 2013. The mosquito state: how technology, capital, and state practice mediate the ecologies of public health. In: King, B., Crews, K.A. (Eds.), Ecologies and Politics of Health. Routledge, London, pp. 196–216.

Schlosser, E., 2001. Fast Food Nation: The Dark Side of the All-American Meal. Houghton Mifflin, New York.

Scott, C., Robbins, P., Comrie, A., 2012. The mutual conditioning of humans and pathogens: implications for integrative geographical scholarship. Annals of the Association of American Geographers 102 (5), 1–9.

Smith, K.R., Woodward, A., et al., 2014. Human health: impacts, adaptation, and co-benefits. In: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, vol. I. Cambridge University Press, Cambridge, UK and NYC (Chapter 11).

Talman, A., Bolton, S., Walson, J.L., 2013. Interactions between HIV/AIDS and the environment: toward a syndemic framework. American Journal of Public Health 103 (2), 253–261.

Turshen, M., 1977. The political ecology of disease. Review of Radical Political Economics 9 (1), 45–60.

Turshen, M., 1984. The Political Ecology of Disease in Tanzania. Rutgers University Press, New Brunswick, NJ.

Wilhelmi, O., de Sherbinin, A., Hayden, M., 2013. In: King, B., Crews, K.A. (Eds.), Ecologies and Politics of Health. Routledge, London, pp. 219–238.

World Health Organization, 2011. WHO Public Health & Environment. Global Strategy Overview. http://www.who.int/phe/en/ (accessed 06.07.14.).

World Health Organization, 2014. Environmental Health. http://www.who.int/topics/environmental_health/en/ (accessed 10.06.14.).

Relevant Websites

http://www.cdc.gov/nceh/ – Centers for Disease Control and Prevention. National Center for Environmental Health (accessed on 08.07.14.).

http://www.eea.europa.eu/themes/human – European Environment Agency (accessed on 08.07.14.).

http://www.healthypeople.gov/2020/topicsobjectives2020/overview.aspx?topicid=12 – Healthy People.gov (accessed on 08.07.14.).

http://www.un.org/millenniumgoals/environ.shtml – United Nations, 2013. Millennium Development Goals (accessed on 01.07.14.).

http://www.who.int/topics/environmental_health/en/ – World Health Organization. Environmental Health (accessed on 10.06.14.).

Environments for Education

Barry J Fraser, Science and Mathematics Education Centre, Curtin University, Perth, WA, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by K. Marjoribanks, volume 7, pp. 4693–4697, © 2001, Elsevier Ltd.

Abstract

This article considers different educational environments (especially the classroom, school, and home), links between different environments, improving educational environments, and future directions, such as qualitative data-collection methods and personal forms of educational environment instruments.

Different Educational Environments

Psychosocial educational environment has been incorporated as several factors in Walberg's (1981) multifactor model of educational productivity. Based on an economic model of agricultural, industrial, and national productivity, this model holds that learning is a multiplicative, diminishing-returns function of student age, ability, and motivation; of quality and quantity of instruction; and of the four psychosocial environments – the home, the classroom, the peer group, and the mass media. Empirical probes of this model were made by Fraser et al. (1987) by carrying out extensive research syntheses involving the correlations of learning with the factors in the model. Also secondary analyses were conducted with National Assessment of Educational Achievement data (Walberg, 1986) and National Assessment of Educational Progress data (Fraser et al., 1986; Walberg et al., 1986). Classroom and school environments were found to be strong predictors of both achievement and attitudes even when numerous other factors in the productivity model were held constant. In the following subsections, consideration is given separately to the specific educational environments of the (1) classroom, (2) school, and (3) home.

Classroom Environment

Fraser's article on Classroom Climate (*see* Classroom Climate) in this *Encyclopedia*, together with a recent literature review of four decades of classroom-level environment research (Fraser, 2012), shows that a major thrust has been the development, validity, and use of various questionnaires. These publications review several lines of past classroom climate research, including evaluation of educational innovations, associations between student outcomes and classroom climate, and teachers' action–research attempts to improve their classroom climates.

School Environment

Earlier school environment instruments include Stern's (1970) College Characteristics Index (CCI) and Halpin and Croft's (1963) Organizational Climate Description Questionnaire (OCDQ). The Work Environment Scale (WES; Moos, 1981) was designed for use in any work milieu, but, to improve its face validity for use in schools, the word 'people' was changed to 'teachers,' 'supervisor' to 'senior staff,' and 'employee' to 'teacher' (Fisher and Fraser, 1983; Fraser et al., 1988). The WES was validated with 599 teachers in 34 primary and secondary schools in Tasmania (Docker et al., 1989).

The School-Level Environment Questionnaire (SLEQ) was designed especially to assess school teachers' perceptions of eight psychosocial dimensions of the environment of the school (Fisher and Fraser, 1991; Rentoul and Fraser, 1983). Two scales measure Moos's (1974) Relationship Dimensions (Student Support and Affiliation), one measures Personal Development Dimensions (Professional Interest), and five measure System Maintenance and System Change Dimensions (Staff Freedom, Participatory Decision Making, Innovation, Resource Adequacy, and Work Pressure). Using both exploratory and confirmatory factor analyses with data from 1106 teachers from 59 elementary schools in southwestern USA, Johnson and Stevens (2001) evolved a 35-item five-scale version of the SLEQ with sound psychometric properties. In South Africa, Aldridge et al. (2006) modified the SLEQ to make it suitable for use among teachers who were implementing outcomes-based education (OBE). With a sample of 403 teachers in 54 schools, they validated a questionnaire that combined modified versions of SLEQ scales with two new scales created by the researchers (Parental Involvement and Familiarity with OBE).

In Taiwan and based partly on the SLEQ, Huang and Fraser (2009) used the Science Teachers' School Environment Questionnaire (STSEQ) to assess the dimensions of Teacher–Student Relations, Collegiality, Principal Leadership, Professional Interest, Gender Equity, Staff Freedom, Innovation, Resources and Equipment, and Work Pressure among 300 female and 518 male science teachers from secondary schools. Sex differences in perceptions were reported for several scales even after controlling for teachers' background and school characteristics.

When Docker et al. (1989) used the WES with a sample of 599 teachers in investigating differences between the environments of various school types, reasonable similarity was found for preferred environment scales, but teachers' perceptions of their actual school environments varied markedly in that the climate in primary schools was more favorable than the environment of high schools for most scales. Similarly, when Fisher and Fraser (1991) used the SLEQ with 109 teachers in 10 schools, the most striking finding was that the climate in primary schools emerged as more favorable than the environment of high schools on most SLEQ scales.

In a study of the school-level environment of Catholic schools, Dorman et al. (1997) developed a school environment instrument that included modified versions of five SLEQ scales, but which added the two new scales of Empowerment (the extent to which teachers are empowered and encouraged to be involved in decision-making processes) and Mission Consensus (the extent to which consensus exists within the staff with regard to the overarching goals of the school). This instrument was used in studies of differences in the school environment of Catholic and government schools (Dorman and Fraser, 1996) and of associations between school environment and classroom environment (Dorman et al., 1997). For a sample of 208 science and religion teachers from 32 schools, Catholic school teachers saw their schools as more empowering and higher on Mission Consensus than government school teachers.

Home Environment

Generally it is agreed that school outcomes will be enhanced when home and school environments are in harmony with each other. However, relative to the voluminous research on school and classroom environments, much less research has considered the home environment, especially the joint and confounding influences of the school and the home on student outcomes. The famous Coleman et al. (1966) report drew attention to the way in which the contributions of the school and the home environments to variance in student outcomes are confounded. Kellaghan et al. (1993) clearly established the potency of positive home–school partnerships in improving student outcomes. When Allen and Fraser (2007) investigated both primary-school students' and their parents' perceptions of the classroom learning environment, it was found that both students and parents preferred a more positive learning environment, but that effect sizes for actual–preferred differences were much larger for parents than for students. At the kindergarten level, Robinson and Fraser (in press) reported that parents perceived more favorable actual environments than students, but students preferred more favorable environments than parents.

When schools are more involved with their communities and when teachers and parents see and talk more often with one another, they are more likely to know about one another's needs and are better able to work in tandem to promote the learning and welfare of students. Therefore, educational outcomes can often be improved if schools become more involved in the lives of their local communities (Berliner and Biddle, 1995). Parents have been found to have significant and positive influences on children's verbal ability, language, school-related knowledge and skills, and achievement in school subjects, especially when integrated into a network of community support (Kellaghan et al., 1993).

Links between Different Educational Environments

Although most individual studies of educational environments in the past have tended to focus on a single environment, there is potential in simultaneously considering the links between, and the joint influence of, two or more of the environments in Walberg's model of educational productivity. For example, Marjoribanks (1991, 1999) showed how the environments of the home and school interact to codetermine school achievement, and Moos (1991) illustrated links between school, home, and parents' work environments.

Several studies have investigated whether the nature of the school-level environment influences or transmits to what goes on in classrooms (i.e., the classroom-level environment). In South Africa, Aldridge et al. (2011) used a school environment instrument based on the SLEQ with 50 secondary-school science teachers from 50 different schools, together with a classroom environment questionnaire with the 2638 grade 8 students in the 50 classes of these 50 teachers. Overall, the school environment did not have a strong influence on what happens in classrooms. Other researchers who have investigated associations between school-level and classroom-level environments include Fraser and Rentoul (1982) and Fisher et al. (1995). When Dorman et al. (1997) administered a classroom environment instrument to 2211 students in 104 classes and a school environment instrument to the 208 teachers of these classes, only weak associations between classroom environment and school environment were found. Although school rhetoric often would suggest that the school ethos would be transmitted to the classroom level, it appears that classrooms are somewhat insulated from the school as a whole.

Fraser and Kahle (2007) simultaneously explored the influence of three of the environmental productivity factors in Walberg's model – namely, the school/class environment, the home environment, and the peer-group environment on students' cognitive and affective outcomes. Using secondary analysis of a large database from a Statewide Systemic Initiative (SSI) in the USA, the effects of these three types of environments on student outcomes were examined. Over 3 years, nearly 7000 students in 392 middle-school science and mathematics classes in 200 different schools responded to a questionnaire that assessed class, home, and peer environments, student attitudes, and achievement. Rasch analyses allowed comparison across student cohorts and across schools. Although all three environments accounted for statistically significant amounts of unique variance in student attitudes, only the class environment (defined in terms of the frequency of use of standards-based teaching practices) accounted for statistically significant amounts of unique variance in student achievement scores. Overall, this study reinforced the importance of extending research on school/classroom learning environments also to include the learning environments of the home and the peer group.

Improving Educational Environments

Although much research has been conducted on educational environments, less has been done to help teachers to improve the environments of their own classrooms or schools. However, Aldridge et al. (2012) have described how feedback information based on student or teacher perceptions can be employed as a basis for reflection upon, discussion of, and systematic attempts to improve classroom and school environments. Fraser and Fisher (1986) recommend five steps: (1)

assessment (students or teacher respond to both the preferred and actual forms of an environment questionnaire); (2) *feedback* (teachers are provided with feedback information derived from responses in the form of profiles representing the means of actual and preferred environment scores); (3) *reflection and discussion* (teachers engage in private reflection and informal discussion about the profiles in order to decide whether to attempt to change the environment in terms of some of the dimensions); (4) *intervention* (teachers introduce an intervention in an attempt to change the school or classroom environment); and (5) *reassessment* (the actual form of the scales is re-administered at the end of the intervention to see whether teachers/students are perceiving their school/classroom environments differently from before).

Fraser et al. (1988) used the WES as part of teacher development activities and reported a case study of a successful school change attempt in a primary school with a staff of 24 teachers. The SLEQ (Fisher and Fraser, 1991) was used in similar school improvement studies using the same basic strategy in a primary school with 15 teachers. After an intervention had been implemented for approximately 10 weeks, it was found that sizable changes had occurred in two of the targeted areas (of about two-thirds of a standard deviation and about half a standard deviation, respectively).

Future Directions

Qualitative Data-Collection Methods

Despite the considerable insights that have been generated through using educational-environment questionnaires that assess participants' perceptions, educators often feel the need to supplement these with qualitative and interpretive methods of data gathering. Furthermore, there are merits in moving beyond choosing between quantitative *or* qualitative methods, to combining quantitative *and* qualitative methods (Fraser and Tobin, 1991; Tobin and Fraser, 1998). In a mixed-methods study of learning environments in Taiwan and Australia, Aldridge et al. (1999) combined the use of a classroom environment questionnaire with classroom observations and interviews with students and teachers. The authors constructed narratives about what was going on in science classrooms in Taiwan and Australia, as well as identifying emergent themes. Overall, the qualitative information complemented the quantitative information and clarified patterns within the two countries and differences between them. Tobin et al. (1990) reported a study which focused on the goal of higher-level cognitive learning and which involved a team of six researchers intensively investigating the grade 10 science classes of two teachers over a 10-week period. Quantitative information from questionnaires assessing students' perceptions of classroom psychosocial environment was consistent with the observers' field records of the patterns of learning activities and engagement in each classroom.

Personal Forms of Educational Environment Instruments

Potentially there is a major problem with many existing educational environment instruments when they are used to identify differences between subgroups within an environment (e.g., males and females) or in the construction of case studies of individuals. The problem is that items are worded in such a way that they elicit an individual's perceptions of the environment as a whole, as distinct from that individual's perceptions of his/her own role within the environment. For example, items in the traditional class form might seek students' opinions about whether 'the work of the class is difficult.' In contrast, a personal form of the same items would seek opinions about whether 'I find the work of the class difficult.' Confounding could have arisen in past studies which employed the class form because, for example, males could find a class less difficult than females, yet males and females still could agree when asked for their opinions about the class as a whole. This distinction between personal and class forms is consistent with Stern et al.'s (1956) terms of 'private' beta press, the idiosyncratic view that each person has of the environment, and 'consensual' beta press, the shared view that members of a group hold of the environment.

When Fraser et al. (1995) developed and validated parallel class and personal forms of both an actual and preferred version of the Science Laboratory Environment Inventory, analyses confirmed that the personal form had a similar factor structure and comparable statistical characteristics to the class form. Also students' scores on the class form were found to be systematically more favorable than their scores on the personal form, perhaps suggesting that students have a more detached view of the environment as it applies to the class as a whole. As hypothesized, gender differences in perceptions were somewhat larger on the personal form than on the class form. Although a study of associations between student outcomes and their perceptions of the science laboratory environment revealed that the magnitudes of associations were comparable for class and personal forms, Fraser and McRobbie (1995) used commonality analyses to show that each form accounted for appreciable amounts of outcome variance which was independent of that explained by the other form. This finding justifies the decision to evolve separate class and personal forms because they appear to measure different, albeit overlapping, aspects of an educational environment.

See also: Classroom Climate; Educational Evaluation: Overview; Educational Productivity; Families as Educational Settings.

Bibliography

Aldridge, J.M., Fraser, B.J., Bell, L., Dorman, J.P., 2012. Using a new learning environment questionnaire for reflection in teacher action research. Journal of Science Teacher Education 23, 259–290.

Aldridge, J.M., Fraser, B.J., Huang, I.T.-C., 1999. Investigating classroom environments in Taiwan and Australia with multiple research methods. Journal of Educational Research 93, 48–62.

Aldridge, J.M., Fraser, B.J., Laugksch, R.C., 2011. Relationship between the school-level and classroom-level environment in secondary schools in South Africa. South African Journal of Education 31, 127–144.

Aldridge, J.M., Laugksch, R.C., Fraser, B.J., 2006. School level environment and outcomes-based education in South Africa. Learning Environments Research 9, 123–147.

Allen, D., Fraser, B.J., 2007. Parent and student perceptions of classroom learning environment and its association with student outcomes. Learning Environments Research 10, 67–82.

Berliner, D.C., Biddle, B.J. (Eds.), 1995. The Manufactured Crisis: Myths, Frauds, and the Attack on America's Public Schools. Longman, White Plains, NY.

Coleman, J.S., et al., 1966. Equality of Educational Opportunity. US Office of Education, Washington, DC.

Docker, J.G., Fraser, B.J., Fisher, D.L., 1989. Differences in the psychosocial work environment of different types of schools. Journal of Research in Childhood Education 4, 5–17.

Dorman, J.P., Fraser, B.J., 1996. Teachers' perceptions of school environment in Australian Catholic and government secondary schools. International Studies in Educational Administration 24 (1), 78–87.

Dorman, J.P., Fraser, B.J., McRobbie, C.J., 1997. Relationship between school-level and classroom-level environments in secondary schools. Journal of Educational Administration 35, 74–91.

Fisher, D.L., Fraser, B.J., 1983. Use of WES to assess science teachers' perceptions of school environment. European Journal of Science Education 5, 231–233.

Fisher, D.L., Fraser, B.J., 1991. School climate and teacher professional development. South Pacific Journal of Teacher Education 19 (1), 17–32.

Fisher, D.L., Grady, N., Fraser, B., 1995. Associations between school-level and classroom-level environment. International Studies in Educational Administration 23, 1–15.

Fraser, B.J., 1982. Development of short forms of several classroom environment scales. Journal of Educational Measurement 19, 221–227.

Fraser, B.J., 2012. Classroom learning environments: retrospect, context and prospect. In: Fraser, B.J., Tobin, K.G., McRobbie, C.J. (Eds.), Second International Handbook of Science Education. Springer, New York, pp. 1191–1239.

Fraser, B.J., Docker, J.G., Fisher, D.L., 1988. Assessing and improving school climate. Evaluation and Research in Education 2 (3), 109–122.

Fraser, B.J., Fisher, D.L., 1986. Using short forms of classroom climate instruments to assess and improve classroom psychosocial environment. Journal of Research in Science Teaching 5, 387–413.

Fraser, B.J., Giddings, G.J., McRobbie, C.J., 1995. Evolution and validation of a personal form of an instrument for assessing science laboratory classroom environments. Journal of Research in Science Teaching 32, 399–422.

Fraser, B.J., Kahle, J.B., 2007. Classroom, home and peer environment influences on student outcomes in science and mathematics: an analysis of systemic reform data. International Journal of Science Education 29, 1891–1909.

Fraser, B.J., McRobbie, C.J., 1995. Science laboratory classroom environments at schools and universities: a cross-national study. Educational Research and Evaluation 1, 289–317.

Fraser, B.J., Rentoul, A.J., 1982. Relationship between school-level and classroom-level environment. Alberta Journal of Educational Research 28, 212–225.

Fraser, B.J., Tobin, K., 1991. Combining qualitative and quantitative methods in classroom environment research. In: Fraser, B.J., Walberg, H.J. (Eds.), Educational Environments: Evaluation, Antecedents and Consequences. Pergamon, London, pp. 271–292.

Fraser, B.J., Walberg, H.J., Welch, W.W., Hattie, J.A., 1987. Syntheses of educational productivity research. International Journal of Educational Research 11, 145–252.

Fraser, B.J., Welch, W.W., Walberg, H.J., 1986. Using secondary analysis of national assessment data to identify predictors of junior high school students' outcomes. Alberta Journal of Educational Research 32, 37–50.

Halpin, A.W., Croft, D.B., 1963. Organizational Climate of Schools. Midwest Administration Center, University of Chicago, Chicago, IL.

Huang, S.-Y.L., Fraser, B.J., 2009. Science teachers' perceptions of the school environment: gender differences. Journal of Research in Science Teaching 46, 404–420.

Johnson, B., Stevens, J.J., 2001. Exploratory and confirmatory factor analysis of the School Level Environment Questionnaire (SLEQ). Learning Environments Research 4, 325–344.

Kellaghan, T., Sloan, K., Alvarez, B., Bloom, B., 1993. The Home Environment and School Learning: Promoting Parental Involvement in the Education of Children. Jossey-Bass, San Francisco.

Marjoribanks, K., 1991. Families, schools, and students' educational outcomes. In: Fraser, B.J., Walberg, H.J. (Eds.), Educational Environments: Evaluation, Antecedents and Consequences. Pergamon, London, pp. 75–91.

Marjoribanks, K., 1999. Environmental and individual influences on adolescents' aspirations: a moderation-mediation model. Learning Environments Research 2, 43–64.

Moos, R.H., 1974. The Social Climate Scales: An Overview. Consulting Psychologists Press, Palo Alto, CA.

Moos, R.H., 1981. Manual for Work Environment Scale. Consulting Psychologist Press, Palo Alto, CA.

Moos, R.H., 1991. Connections between school, work, and family settings. In: Fraser, B.J., Walberg, H.J. (Eds.), Educational Environments: Evaluation, Antecedents and Consequences. Pergamon Press, London, pp. 29–53.

Rentoul, A.J., Fraser, B.J., 1983. Development of a school-level environment questionnaire. Journal of Educational Administration 21, 21–39.

Robinson, E., Fraser, B.J., 2013. Kindergarten students' and parents' perceptions of science classroom environments: achievement and attitudes. Learning Environments Research 16, 151–167.

Stern, G.G., 1970. People in Context: Measuring Person-Environment Congruence in Education and Industry. Wiley, New York.

Stern, G.G., Stein, M.I., Bloom, B.S., 1956. Methods in Personality Assessment. Free Press, Glencoe, IL.

Tobin, K., Fraser, B.J., 1998. Qualitative and quantitative landscapes of classroom learning environments. In: Fraser, B.J., Tobin, K.G. (Eds.), The International Handbook of Science Education. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 623–640.

Tobin, K., Kahle, J.B., Fraser, B.J. (Eds.), 1990. Windows into Science Classes: Problems Associated with Higher-Level Cognitive Learning. Falmer Press, London.

Walberg, H.J., 1981. A psychological theory of educational productivity. In: Farley, F., Gordon, N.J. (Eds.), Psychology and Education: The State of the Union. McCutchan, Berkeley, CA, pp. 81–108.

Walberg, H.J., 1986. Synthesis of research on teaching. In: Wittrock, M.C. (Ed.), Handbook of Research on Teaching, third ed. American Educational Research Association, Washington, DC, pp. 214–229.

Walberg, H.J., Fraser, B.J., Welch, W.W., 1986. A test of a model of educational productivity among senior high school students. Journal of Educational Research 79, 133–139.

Epidemiology, Concepts and Methods of

Roselind Lieb, Division of Clinical Psychology and Epidemiology, University of Basel, Missionsstrasse, Basel, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides an overview of selected topics in the complex discipline of epidemiology. The aim is to focus the reader's attention to key points of epidemiology, such as definition and scope of epidemiology, criteria for causal inference, measures for morbidity and mortality, study designs, and measures of potential for prevention. Sample calculations help for a better understanding of the presented measures.

What Is Epidemiology?

Epidemiology is the scientific discipline that investigates the distribution of health phenomena (e.g., certain diseases, smoking, health behavior) in defined populations as well as factors that influence this distribution. Therefore, its premise addresses the search for predispositions and causal mechanisms for health phenomena in populations. Causal mechanisms may include a broad spectrum of biological, environmental, social, or individual factors. Epidemiology not only investigates specific factors, but also their sometimes highly complex interplay in the etiology of a certain disease. A prominent definition of epidemiology was given by Porta (2008), who defined epidemiology as "the study of the occurrence and distribution of health-related states or events in specified populations, including the study of determinants influencing such states, and the application of this knowledge to control the health problems" (Porta, 2008). In addition to the description of epidemiology given above, this definition additionally cites as an important component of the application of epidemiological knowledge in the broad field of public health.

The Scope of Epidemiology

If we want to describe the scope of epidemiology in more detail, several objectives can be addressed (see Gordis, 2009; Friis and Sellers, 2014; Tsuang et al., 2011):

1. The identification of the etiology or cause of health problems as well as the disease-relevant risk factors, that is, factors that increase a person's risk of developing the health problem. Once a factor has been shown to be a risk factor for a disease, the question of causality has to be addressed. Is the risk factor a causal risk factor, or, whenever the risk factor is modified, does this have an influence on the incidence of a disease (see terminology proposed by Kraemer et al., 1997)?
2. The investigation of the extent and burden of health problems in the community in terms of mortality and nonfatal consequences (e.g., disability associated with diseases, direct and indirect costs). This question is critical for decision making in health policy in order to improve, plan, organize, and provide health services.
3. The investigation of the natural history of diseases and health problems in terms of onset, course, and outcome.
4. The evaluation of the performance and effectiveness of existing health services. Do they cover needs in the community?

We can further divide the discipline of epidemiology into two major subdisciplines: descriptive epidemiology and analytic epidemiology (see Friis and Sellers, 2014). Descriptive epidemiology predominantly addresses the description of health phenomena in populations, their natural course, and the evaluation of health services. Analytic epidemiology deals with the extremely important question of risk factors and causality. Descriptive studies generally precede analytic studies, i.e., they identify the distribution of health problems, while analytic studies address the causes of the problems.

Measures of Morbidity

We have learned so far that the scope of epidemiology includes the description of health phenomena in populations. For this, several measures are used to describe the size of health problems. As basic measures, prevalence, incidence, mortality, and lethality will be defined in this section.

Prevalence

Prevalence is defined as the number of affected persons present in an at-risk population at a specific time or for a specific time period related to the total number of persons in that population at (for) the same time (period). This means that prevalence in fact describes a certain proportion or rate. The prevalence rate describes the proportion of the at-risk population that is affected by the disease of interest at the defined time (period).

For example, if we are interested in evaluating the prevalence rate of anxiety disorders in an at-risk population at a specific time (e.g., during the 12 months before an interview), we might use interviews to assess first the number of individuals affected by anxiety disorders defined according to the Diagnostic and Statistical Manual of Mental Disorders (DSM)-IV in that population. This number becomes the nominator for the prevalence rate. The denominator is the number of individuals in the at-risk population at this time. It

 http://dx.doi.org/10.1016/B978-0-08-097086-8.21028-3

is important to specify prevalence rates according to the time frame under consideration. For this, the term prevalence is further specified as point prevalence, period prevalence, and lifetime prevalence.

Point prevalence describes the prevalence rate of the disease of interest at a certain point in time (e.g., on a specific day).

Period prevalence describes the prevalence rate of the disease of interest in a chosen time period (e.g., within 12 months: a 12-month prevalence). It is important to point out that, for the determination of the period prevalence rate, it does not matter whether individuals developed the disease during this time period or whether they remitted or even died. All individuals affected by the disease during the time period count for the nominator.

Lifetime prevalence describes the proportion of individuals in the at-risk population that have been affected by the disease of interest at some point in their life up to the time of the assessment.

In general, prevalence rates are presented as a percentage of the at-risk population. Figure 1 presents a sample calculation for lifetime, 12-month prevalence, and point prevalence.

Incidence

Incidence is defined as the number of individuals in an at-risk population who for the first time in their life are affected during a specified time period related to the total number of persons in that population at the same time (period). Like prevalence, incidence describes a certain proportion or rate: the proportion of the at-risk population that for the first time is affected by the disease of interest during the defined time (period). For example, if we are interested in evaluating the incidence rate of anxiety disorders in an at-risk population during a specific time period (e.g., during the 12 months prior to the interview), we might first use interviews to assess the number of subjects who for the first time were affected by anxiety disorders (defined according to the DSM-V) in that population during this time period (i.e., they have never been affected before). This number becomes the nominator for the incidence rate. The denominator is the number of individuals in the at-risk population during this time period (e.g., all individuals who could have developed the disorder under consideration). Comparable to prevalence, incidence can be specified according to the time period under consideration. *Period incidence* describes the incidence rate for a certain time interval (e.g., 12-month incidence) while lifetime incidence (or cumulated lifetime incidence) describes the incidence rate for the whole life period up to the assessment (which then equals the lifetime prevalence rate). What is the difference between incidence and prevalence? While in the prevalence rate the nominator includes all individuals affected by the disorder of interest (irrespective of whether they were already affected before the time period under investigation), the nominator of the incidence rate includes only 'new' cases. Further, the denominator of the incidence rate includes only individuals that had not been previously affected by the disease, e.g., individuals with a history of the disease are excluded from the at-risk population. Figure 2 gives a sample calculation for 12-month incidence and lifetime incidence.

Mortality

Mortality expressed in quantitative terms (mortality rate) is the total number of deaths within a defined time period related to the number of individuals in the population within this period. Mortality is often specified for specific populations or specific

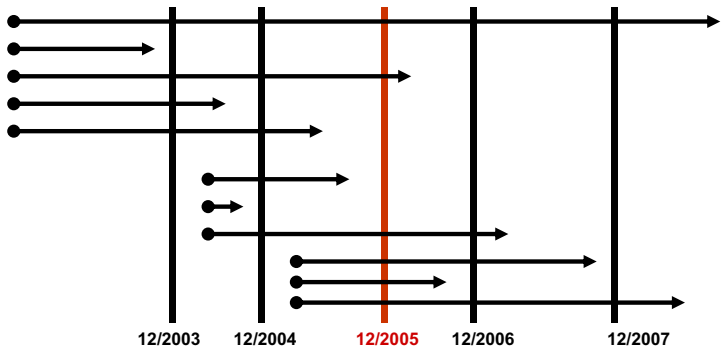

- **Determination of point prevalence in 12/2005 (number of cases based on population *n* = 100)**
  - **Point prevalence = 6/100 = 6%**
- **Determination of 12-month prevalence from 12/2004 to 12/2005 (number of cases based on population *n* = 100)**
  - **12-month prevalence = 8/100 = 8%**
- **Determination of lifetime prevalence (number of cases based on population *n* = 100)**
  - **Lifetime prevalence = 11/100 = 11%**

Figure 1 Sample calculation of point prevalence, 12-month prevalence, lifetime prevalence in a fictitious example.

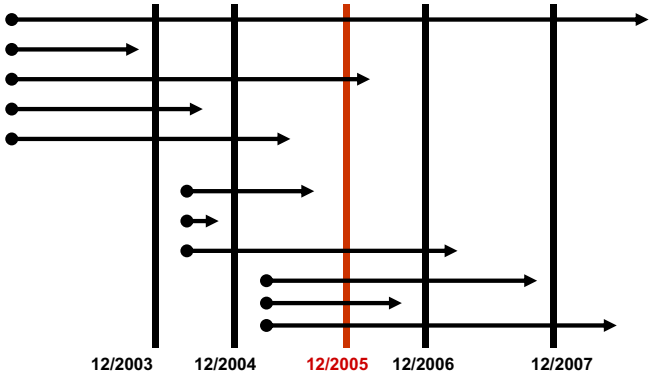

– **Determination of 12-month incidence from 12/2004 to 12/2005 (number of cases based on population)**
  - **12-month incidence: 3/92 (population *n* = 92; exclusion of 8 cases, that ever had the disease until 12/2004)**
– **Determination of lifetime incidence until 12/2005 (number of cases based on population *n* = 100)**
  - **Lifetime incidence: 11/100**

Figure 2 Sample calculation of incidence rates from different time periods in a fictitious example.

causes. Thus, the mortality rate for all causes refers to deaths from all causes in a specific time period, while cause-specific mortality rates refer to deaths from specific causes (e.g., from heart attacks) in a specific time period. Note that the denominator in both rates consists of the number of individuals in the population during this time period. Mortality also can be expressed for specific populations (e.g., men, women, infants up to 1 year, children up to 5 years). Thus, *infant mortality* (or the infant mortality rate) measures the risk of dying during the first year of life among all infants born alive.

Lethality

Lethality in contrast to mortality describes how many individuals diagnosed as having a certain disease die within a specific time period related to all individuals with the same diagnosis during this time period. Here, the denominator includes only individuals affected by the disease of interest and not the entire population (as in the mortality rate). For a sample calculation see Box 1.

Epidemiology, Risk Factor, and Causal Inference

As discussed at the beginning of this article, the search for causes of health problems is one of the major domains of epidemiology. Schwartz and Susser (2006a,b) identified the risk-factor paradigm as the basic paradigm in epidemiologists' search for the causes of diseases.

Using the risk-factor paradigm, epidemiologists search for factors (e.g., certain exposures) that increase the risk of a certain disease in individuals who have been exposed to these factors in comparison with individuals who have not been exposed. Risk means here the probability of developing a disease, but does not imply a deterministic association. Likewise, risk factors may or may not be sufficient or necessary for the development of the outcome under consideration. According to Kraemer et al. (1997), a risk factor can be defined as "a measurable characterization of each subject in a specified population that precedes the outcome of interest and which can be used to divide the population into 2 groups (the high-risk and the low-risk-group)" (Kraemer et al., 1997: p. 338). For a factor to be called a 'risk factor,' a statistically significant association between the factor and outcome under consideration must be shown. What is important is that a risk factor precedes the outcome and must show a significant relationship with the development of the outcome (e.g., with a health problem or disease). Next, it is important to distinguish between a risk factor and the cause of a disease. According to the concept of Kraemer et al. (1997), a risk factor must precede the outcome under consideration. This is also a critical component for the definition of cause, but not a sufficient one. The group around Kraemer suggests distinguishing between 'risk factor' and 'causal risk factor.' Here, a 'causal risk factor' meets all the criteria required for a 'risk factor,' but additionally, it must be shown that a manipulation of this factor changes the outcome. Note that here the term 'causal risk factor' and not the term 'cause' is used; this is because, according to Kraemer et al. (1997), the term 'cause' often inappropriately implies that the condition is a necessary and sufficient condition for the outcome of interest.

Box 1 Sample calculation of lethality

Assume 20 deaths among 100 individuals who have developed a heart disease throughout their life. Lethality rate here would be 20/100 = 20%.

Another approach goes back to the so called Hill criteria (1965). Hill (1965) suggested criteria that should be fulfilled before interpreting an established exposure–outcome association as a causal association (or, that the exposure causes the outcome). The Hill criteria are presented in **Table 1**.

These criteria were the starting point for the discussion and modification of criteria encouraging the interpretation that an exposure plays a causal role in an outcome (i.e., in a disease) (see Schwartz and Susser, 2006a,b; Rothman and Greenland, 2005). It is important in using this approach that causality cannot be tested empirically in establishing the causal role of a risk factor, but, rather, requires evidence and consistent results from different approaches and designs. The criteria give an orientation or guidelines for a conclusion about causality, but the conclusion itself depends on a judgment based on the existing evidence.

Epidemiological Study Designs

There are several study designs that can answer epidemiological questions. The design of a study determines on the one hand its limitations and threats to validity (inherent in every study design) and the precision and validity of its findings on the other. In the following section, the basic study designs in epidemiology will be presented. A discussion of the details of each study design and of their combinations is beyond the scope of this article. For this, the reader is referred to Susser et al. (2006), Gordis (2009), Friis and Sellers (2014), Rothman et al. (2008), and Twisk (2013).

Table 1 Criteria for evidence of causal association

Criterion	*Description*
Strength of association	The stronger the association, the more likely is a causal effect
Consistency of association	Similar associations have repeatedly been found by different people, in different places, circumstances, and times (replication)
Specificity of association	An exposure is associated only with one specific health outcome and not with others
Temporality of association	There is a clear temporal association between exposure (primary) and health outcome (secondary)
Biological gradient or dose–response	Higher exposure should be associated to a higher risk
(Biological) plausibility	Given the current (biological) knowledge the causation should be reasonable
Coherence	The observed association should not be inconsistent with or in contrary to existing knowledge
Experiment	An effect is more likely to be causal if it is based on the results of an experimental design
Analogy	An association between a comparable exposition or health outcome has already been documented

Source: Hill, S.B., 1965. The environment and disease: association or causation? Bulletin of the World Health Organization 83, 796–798.

As proposed by Friis and Sellers (1999), epidemiological study designs can be discussed on the basis of two factors: (1) Manipulation of the study factor (M), i.e., that the exposure of interest is controlled by the investigator, and (2) randomization of the study subjects (R), i.e., the random assignment of the study individuals to different exposure conditions. The involvement of these two factors result in three different study designs: experimental (M and R are involved), quasi-experimental (M but not R is involved), and observational studies (neither M nor R is involved).

Experimental and Quasi-Experimental Approaches

In an *experimental study*, the investigator both manipulates the study factor (e.g., certain exposures) and randomly assigns subjects to the exposure conditions. One common experimental design is the randomized controlled trial; here, the study subjects are randomly assigned to the different conditions of the study intervention (e.g., a new therapy vs treatment as usual) in order to evaluate and compare the effects of the interventions on a modification in health status (e.g., to evaluate whether a new therapy has better effects than treatment as usual). In a *quasi-experimental study*, the study factor is manipulated by the investigator but subjects are not randomly assigned to the study conditions.

The strengths of the experimental approach include their potential for high internal validity of the findings. Limitations clearly include its low external validity (e.g., generalizability of the findings to other subjects and settings). Likewise, the vast majority of assumed factors that may be causal in the development of a disease can never be examined experimentally for the whole development phase of the disease.

Observational Approach

Although experimental and quasi-experimental studies seem to be the best approaches to maximize internal validity, it is not possible to realize them for every question of interest (e.g., causal factors in the development of a disease). Accordingly, most epidemiological research regarding causal factors in disease development is related to observational studies. Observational studies (no randomization, no control of the exposure of interest) observe individuals in order to evaluate whether an association exists between a certain exposure and the health problem of interest. This study approach is clearly more vulnerable to the effect of bias than experimental approaches and therefore internal validity is more threatened. This problem, however, may at least be minimized by the choice of a good study design. Observational studies can be used to answer research questions of both descriptive and analytic epidemiology. Descriptive studies assess information that describes the health phenomenon of interest in a selected population. Analytic studies observe exposure and the health outcome of interest and test hypotheses about the association between these factors. Once associations have been found, the search for a causal relationship can go further (see above). Observational study designs that are most often used in epidemiology include cross-sectional

studies, case–control studies, and longitudinal (cohort) studies.

Cross-Sectional Study

The cross-sectional study (also called prevalence study) is characterized by a single observation. Information about present or past exposure and outcome of interest is collected for a defined population and at a specific point in time. This study design can provide information about the distribution of health phenomena in specific populations (e.g., prevalence rates in a certain country, men, women, specific age groups). Thus, it is a good design for the evaluation of the size of health problems in terms of prevalence rates. When evaluating the association between a certain exposure and an outcome of interest, the study population is normally divided into two groups: individuals who have been exposed and individuals who have not been exposed. For the same individual, health status is collected and the association between exposure and health outcome is evaluated. The major limitation of the cross-sectional design is that it is weak in detecting potential causal associations between exposure and outcome, because it is often not possible to determine a clear temporal association between exposure and outcome of interest. Another limitation includes the use of prevalence rates in the evaluation of the association. Cases of prevalence rates may not be representative of all individuals who develop the health problem of interest (they are only included in the incidence rates) and study findings therefore may not be generalizable. The strengths of this study design include a short study period and the inclusion of large representative populations (allowing maximization of external validity).

An example of a cross-sectional study is presented in Box 2.

Longitudinal (Cohort) Study

In contrast to the cross-sectional study design, the longitudinal study design is characterized by observing study individuals over a length of time. Longitudinal studies can be further grouped into prospective versus retrospective longitudinal studies. In a prospective longitudinal study design, an investigator observes a group of individuals (the study cohort) by following them forward over a certain period of time. Across the study period, incidence rates for the health outcome of interest are assessed in both groups. This can be done in one or several follow-up assessments. The temporal association between exposure and risk (incidence) for the outcome can then be determined. For a study sample see Box 3.

The retrospective longitudinal study design is characterized by the use of historical (that means previously collected) information about the exposure condition and the inclusion of incidence rates between exposure date and study date. It is important to note that the study design does not differ: in both prospective and retrospective longitudinal designs we compare incidence rates (risk) of exposed individuals with incidence rates of unexposed individuals.

The longitudinal study design is a powerful study design for the evaluation of risk factors for health problems. Using this design, the temporal relationship between exposure and outcome is given and all people at risk are observed. This clearly maximizes the internal validity of the findings. Nevertheless, potential confounding bias (see Box 4) is always inherent in this study type (since individuals are not randomly assigned to the different exposure conditions). Another limitation is that the design is not optimal for rare diseases and diseases that develop slowly over time. Rare and slowly developing diseases would demand an immense number of study participants plus a very long follow-up period. Both are

Box 2 Example of a cross-sectional study: The National Comorbidity Survey Replication Adolescent Supplement Study

The National Comorbidity Survey Replication Adolescent Supplement is a United States national representative epidemiological survey of 10 123 adolescents aged 13 to 18 years in the United States. For the study, all participants have been assessed (interviews, questions) once within one study period fielded in 2001–2004. The scope of the study was to provide estimates of prevalence, correlates, and patterns of service use for DSM-IV mental disorders among US adolescents. Mental disorders were assessed according to DSM-IV using a modified version of the World Health Organization Composite International Diagnostic Interview (see Merikangas et al., 2009).

Box 3 Example for a prospective longitudinal study: The Early Development Stages of Psychopathology Study

The prospective longitudinal Early Developmental Stages of Psychopathology Study assessed DSM-IV mental disorders and associated risk factors in a representative sample of 3021 adolescents and young adults aged 14–24 years at baseline (T0) in Munich, Germany. The baseline assessment took place at 1995. The study also included three follow-up surveys (T1, T2, and T3 in 1997, 1999, and 2005), a family history component (T0, T2, and T3), and parent surveys (T1 and T3) (see Lieb et al., 2000; Zimmermann et al., 2008).

Box 4 What is confounding?

Confounding is a central methodological issue in nonexperimental study designs. To explain what a confounding factor is, let us start with a simple example. Suppose that we observe a positive association between the exposure 'anxiety during adolescence' and the outcome 'depression in later life.' Perhaps we are now tempted to call this a 'true' association and to infer a causal relationship between adolescence anxiety and depression in later life. This 'true' association, however, could also be a 'false' one. Suppose that another variable, e.g., the experience of early traumatic events, is associated with both, the exposure (here: adolescence anxiety) and the outcome (here: depression in later life). Suppose further that if we would account for this variable, the association between exposure and outcome of interest would disappear or become smaller. In this scenario, all or part of the association could be explained by such a so-called confounding variable. It is obvious that in this scenario failure to account for confounding can lead to biased findings in terms of spurious associations. To avoid such spurious associations, epidemiologists always have to take into account potential confounding variables that may bias their findings.

theoretically possible, but not always practicable under given research conditions.

Case–Control Study

Another often used study design is the case–control study. This study design divides the participants into two groups at the beginning of the study: those who already have the disease (cases) and those who have not (yet) developed the disease of interest (controls). Then information regarding the exposure of interest is collected in both groups. Associations between exposure and disease status can then be evaluated. The advantage of case–control studies is that they can use small group sizes and are therefore cheaper to conduct than longitudinal studies. Limitations include potential confounding bias and the retrospective assessment of exposure, which may be affected by recall bias. These facts may threaten internal validity when evaluating temporal associations between exposure and the outcome of interest. Potential bias in the selection of cases and controls may additionally threaten external validity.

Measures of Association

The above-described study designs (case–control studies, longitudinal studies) allow the collection of information about exposure condition and the health outcome of interest. If we want to determine whether a certain exposure is associated with a certain outcome, we have to answer the question of whether individuals who are exposed have a higher risk of developing the outcome of interest than unexposed individuals. When the term risk is used as an equivalent for incidence rate, we want to evaluate whether exposed individuals show higher incidence rates than unexposed people. We see at first glance that cross-sectional and case–control studies are critical in determining such temporal associations, since they do not assess incidence rates. Measures of associations that can be calculated on the basis of incidence rates include *relative risk* (*RR*) and *odds ratio* (*OR*). OR additionally can be calculated only when prevalence rate are available.

Relative Risk (Risk Ratio)

The RR can be defined as the ratio of the risk of the health phenomenon (e.g., a disease) of interest among the exposed individuals to the risk among unexposed individuals. In epidemiology, risk is calculated by the incidence rate. In other words, RR can be calculated by comparing the probability of the health problem developing in exposed individuals with the probability in unexposed individuals. A sample calculation is shown in Box 5.

The RR quantifies the strength of the association between exposure and outcome and can be equal to one, greater than one, or less than one. These values can be interpreted as follows (see Gordis, 2009):

1. If RR = 1, risk in exposed individuals is equal to the risk in nonexposed individuals, i.e., no association.
2. If RR > 1, risk in exposed individuals is greater than the risk in unexposed individuals (positive association assuming statistical significance).
3. If RR < 1, risk in exposed individuals is lower than the risk in unexposed individuals (negative association assuming statistical significance).

As we have learned in the section on study designs, only longitudinal cohort studies allow for the direct assessment of incidence rates. RR can therefore not be calculated in cross-sectional studies and case–control studies.

Box 5 Sample calculation for the relative risk

Suppose that 30 of 100 individuals who have been exposed to an early traumatic event develop depression – this shows a risk (incidence rate) of 0.3. Fifteen of 100 individuals who have not been exposed to traumatic events develop depression. The risk to develop depression is (incidence rate) is 0.15. The RR is now calculated by the risk of the exposed group compared to the risk of the unexposed group (0.3/0.15 = 2.0). Hence, the risk of developing depression of the exposed group is twice as high in comparison to the risk of the unexposed group: RR = 2.0.

Odds Ratio

In contrast to the RR, the OR is defined as the ratio of two odds (instead of two risks). Taking the perspective of a longitudinal study, the OR is the ratio of the odds of the exposed individuals to the odds of the unexposed individuals. Here, the odds of the outcome are defined as the ratio of the probability that the outcome occurs to the probability the outcome does not occur. When using incidence rates from longitudinal studies the 'probability that the outcomes occurs' is equal to the incidence rate; 'the probability that the outcome does not occur' is then one minus the incidence rate. A sample calculation is presented in Box 6.

Like the RR, the OR is a quantitative measure of association. In longitudinal studies, both measures can be calculated. The advantage of the OR is that it can also be calculated in cross-sectional and case–control studies – that means that it can be determined as a measure of the association between exposure and outcome also in these study types. When we take, for example, the perspective of a case–control study, it is possible

Box 6 Sample calculation of odds ratio

Suppose that 30 of 100 individuals who have been exposed to a traumatic event develop depression. They have a probability of 0.3 to develop depression and a probability of 0.7 not to develop depression. Here, the odds is 0.3/0.7 = 0.42. Of 100 individuals who have not been exposed to a traumatic event, 15 develop depression. In this group, the probability to develop depression is 0.15 and the probability not to develop depression is 0.85. Hence, the odds is 0.15/0.85 = 0.18. The OR is calculated as the ratio of the odds of the exposed group to the odds of the unexposed group: 0.42/0.18, so the OR is 2.3. Thus, the odds of developing depression of an exposed group is 2.3 times higher than the odds of the unexposed group.

to calculate the odds of having the exposure in affected individuals (cases) related to the odds of having the exposure in unaffected individuals (controls). The problem is just the other way round. The odds of the exposure is calculated as the ratio of the probability the exposure occurs (or occurred) in cases to the probability the exposure occurs in controls. The OR is a symmetrical measure for association and can be determined as a measure for exposure–outcome association in studies with and without incidence rates.

Measures of Potential for Prevention

In the last section, measures of association between exposure and outcome were presented. In this section, we focus on a further question: if we have identified a causal exposure, and if we could eliminate this exposure, what effect would such a modification have for the group of exposed individuals? And what would the effect be for the whole population? Or, if we could eliminate the exposure, what proportion of the incidence in exposed individuals or in the population could be prevented? The two major measures that describe such proportions are presented in the following section. A more detailed discussion is provided by Gordis (2009) and Greenland et al. (2008).

Attributable Risk

The attributable risk (AR, also referred to as excess fraction) tells us what proportion of the incidence of the outcome of interest can be attributed to a certain exposure. It is important to note that we are focusing here on the group of exposed individuals. We determine how much of the incidence (risk) in this group could be prevented if the exposure were prevented. The AR as a difference can be calculated by subtracting the incidence in the nonexposed group (which describes the background risk) from the incidence in the exposed group. The AR as a proportion describes this difference in relation to the incidence in the exposed group; therefore, we calculate ((incidence in the exposed group) − (incidence in the nonexposed group))/(incidence in the exposed group). A sample calculation is presented in Box 7.

Population Attributable Risk

The population attributable risk (PAR, also referred to as population attributable fraction) also refers to an indicator of the potential for prevention, but on the population level. This measure describes the proportion of the incidence rate of a certain outcome in the population (including both exposed and unexposed individuals) that could be prevented if a causal factor for this outcome could be prevented. In other words, the PAR describes the AR for the population. There are several ways to calculate the PAR. The calculation that is comparable to that of the AR first subtracts the incidence in the nonexposed group from the incidence in the total population and then relates this proportion or fraction to the incidence in the total population. If the incidence in the total population is not known, it can be calculated as ((incidence in the exposed group)(proportion of those exposed in the total population)) + ((incidence in the nonexposed group)(proportion of nonexposed individuals in the total population)). A sample calculation is presented in Box 8.

Another way to calculate the PAR includes as its components the probability p that the risk factor occurs in the total population and the RR (exposed vs nonexposed individuals). With these components, the PAR can be calculated as $\mathrm{PAR} = (p \times (\mathrm{RR} - 1))/(p \times (\mathrm{RR} - 1) + 1)$.

Box 7 Sample calculation of attributable risk

Suppose the 12-month incidence of onset of depression in individuals who have been exposed to a traumatic event is 60% and in unexposed individuals it is 20%. AR can be calculated as following: (incidence in the exposed individuals – incidence in the unexposed individuals)/incidence in the exposed individuals. For this example, $(0.6 - 0.2)/0.6 = 0.66 = 66\%$ of the incidence in the exposed individuals is attributable to the exposition to a traumatic event.

Box 8 Sample calculation of population attributable risk

Suppose the incidence of onset of depression within 1 year is 10% in the total population. In individuals who have been exposed to an early traumatic event, the incidence is 15% and in unexposed individuals, it is 5%. PAR can be calculated as following: (incidence in the total population – incidence in the unexposed group)/incidence in the total population. For this example, $(0.1 - 0.05)/0.1 = 0.5 = 50\%$ of the incidence in the total population is attributable to the exposition to early traumatic event.

Conclusions

This article aimed to provide an overview over selected topics in the complex discipline of epidemiology. The aim was to focus the reader's attention to key points of epidemiology, such as definition and scope of epidemiology, criteria for causal inference, measures for morbidity and mortality, study designs, and measures of potential for prevention. The reader hopefully has gained some insight and understanding of key areas of epidemiology.

See also: Panic Disorder and Agoraphobia Across the Lifespan; Phobias Across the Lifespan; Social Phobia across the Lifespan.

Bibliography

Friis, R.H., Seller, T.A., 1999. Epidemiology for public health practice, second ed. Aspen Publishers, Gaithersburg, MD.

Friis, R.H., Sellers, T.A., 2014. Epidemiology for Public Health Practice, fifth ed. Jones and Bartlett Publishers, Sudburry, MA.

Gordis, L., 2009. Epidemiology, fourth ed. Elsevier/Saunders, Philadelphia.

Greenland, S., Rothman, K.J., Lash, T.L., 2008. Measures of effect and measures of association. In: Rothman, K.J., Greenland, S., Lash, T.L. (Eds.), Modern Epidemiology. Lippincott Williams & Wilkins, Philadelphia, pp. 51–70.

Hill, S.B., 1965[2005]. The environment and disease: association or causation? Bulletin of the World Health Organization 83, 796–798.

Kraemer, H.C., Kazdin, A.E., Offord, D.R., Kessler, R.C., Jensen, P.S., Kupfer, D.J., 1997. Coming to terms with the terms of risk. Archives of General Psychiatry 54, 337–343.

Kraemer, H.C., Stice, E., Kazdin, A., Offord, D., Kupfer, D., 2001. How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. American Journal of Psychiatry 158, 848–856.

Lieb, R., Isensee, B., von Sydow, K., Wittchen, H.-U., 2000. The early developmental stages of psychopathology study (EDSP). A methodological update. European Addiction Research 6, 170–182.

Merikangas, K.-R., Avenevoli, S., Costello, J., Koretz, D., Kessler, R.C., 2009. National comorbidity survey replication adolescent supplement (NCS-A): I. Background and measures. Journal of the Academic Child and Adolescent Psychiatry 48, 367–379.

Porta, M., 2008. A Dictionary of Epidemiology, fifth ed. Oxford University Press, New York.

Rothman, K.J., Greenland, S., 2005. Causation and causal inference in epidemiology. American Journal of Public Health 95 (Suppl. 1), 144–150.

Rothman, K.J., Greenland, S., Lash, T.L., 2008. Modern Epidemiology, third ed. Lippincott, Williams & Wilkins, Philadelphia.

Schwartz, S., Susser, E., 2006a. Detecting causes. In: Susser, E., Schwartz, S., Morabia, A., Bromet, E.J. (Eds.), Psychiatric Epidemiology: Searching for the Causes of Mental Disorders. Oxford University Press, Oxford, pp. 43–52.

Schwartz, S., Susser, E., 2006b. What is a cause? In: Susser, E., Schwartz, S., Morabia, A., Bromet, E.J. (Eds.), Psychiatric Epidemiology: Searching for the Causes of Mental Disorders. Oxford University Press, Oxford, pp. 33–42.

Susser, E., Schwartz, S., Morabia, A., Bromet, E.J., 2006. Psychiatric Epidemiology: Searching for the Causes of Mental Disorders. Oxford University Press, Oxford.

Tsuang, M.T., Tohen, M., Jones, P.B., 2011. Textbook in Psychiatric Epidemiology, third ed. Wiley-Blackwell, Chichester.

Twisk, J.W.R., 2013. Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide. Cambridge University Press, Cambridge.

Zimmermann, P., Brückl, T., Lieb, R., Nocon, A., Ising, M., Beesdo, K., Wittchen, H.-U., 2008. The interplay of familial depression liability and adverse events in predicting the first onset of depression during a 10-year follow-up. Biological Psychiatry 63, 406–414.

Epigenetic Inheritance

Eva Jablonka and Marion J Lamb, Tel Aviv University, Tel Aviv, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Not all of the information that is transmitted between generations of cells and organisms is encoded in DNA sequences. Information is also transmitted through epigenetic systems, which include the non-DNA parts of chromosomes, self-sustaining metabolic loops, self-propagating protein structures, and small RNA molecules. Some of the properties of epigenetic inheritance make it very different from genetic inheritance, particularly because new epigenetic variants can arise in response to the conditions of life. Consequently, understanding epigenetic inheritance is important for medicine, agriculture, public health, and environmental conservation. It also alters views about evolution.

Introduction

'Epigenetic inheritance' is a term used to describe the transmission from generation to generation of structural and functional variations that do not depend on differences in DNA base sequences. In its broadest sense, the term can also include the transmission between generations of nongenetic variations in behavior, but its use is usually restricted to transmission in cell lineages. The inheritance of induced epigenetic changes is part of normal development, but epigenetic variants can also be transmitted from one generation of individuals to the next. Understanding epigenetic inheritance is of practical importance for medicine and agriculture, and has theoretical implications for ideas about heredity and evolution.

Definitions and Historical Background

'Epigenetics,' 'epigenetic,' and 'epigenetic inheritance' are relatively new terms, derived from the much older term 'epigenesis.' According to the ancient and influential theory of epigenesis, development involves qualitative changes in organization, with the parts of an adult animal forming gradually and sequentially from an ovum that is amorphous and unorganized. In the rival theory of preformation, which was popular in the late seventeenth and eighteenth centuries, development was assumed to entail the growth or unfolding of parts that already exist in miniature in the egg (Pinto-Correia, 1997). Since the beginning of the nineteenth century, it has been accepted that development is through epigenesis, although a subtle form of preformation was introduced toward the end of the nineteenth century, when it was recognized that although developmental changes are qualitative, they are controlled by preformed determinants whose transmission and effects ensure hereditary continuity. For several decades after the rediscovery of Mendel's laws in 1900, the new science of genetics was concerned with the way these determinants, the genes, are transmitted from one generation to the next; their role in development received little attention until the second half of the twentieth century.

The introduction of the term 'epigenetics' by the British embryologist Conrad Waddington in the late 1930s (Waddington, 1975) reflected a growing interest in the relationship between genes and development, an interest that increased as the molecular nature of the gene was unraveled. Waddington coined the term epigenetics for the study of the causal mechanisms of development; the adjective 'epigenetic' is applied to the processes through which the genetic information in a fertilized egg is expressed and used in the production of the adult form, the visible phenotype. As the study of genetics and development progressed, it became clear that different cell types usually have identical genotypes, but express different sets of genes (one important exception proved to be the cells of the immune system, where epigenetic changes involve alterations in DNA sequences). Finding out which genes are expressed and what controls when and where they are expressed became the key to understanding the functional and structural specialization of cells and tissues during development. It also became clear that once a cell is committed to a particular form or function, that determined state is often transmitted to daughter cells, even though the stimulus that originally induced its development is no longer present. Thus, once a cell is committed to becoming a liver cell, it breeds true and by division gives rise to a lineage of liver cells; similarly, skin cells will give rise to skin cells, and kidney cells to kidney cells. This stable transmission of the alternative phenotypes of cells having the same genotype became known as 'cellular heredity,' 'cell heredity,' or 'epigenetic inheritance.'

Interest in the transmission of cellular states began to grow during the mid-1970s, as the molecular mechanisms and systems that underlie the control of gene expression started to be unraveled (Holliday, 1990). Gradually it was recognized that epigenetic information is transmitted not only between generations of cells in lineages within each individual, but also through the germline and gametes to the next generation of individuals. The concept of epigenetic inheritance was therefore extended and now has both narrow and broad meanings. The narrow meaning limits it to heredity in mitotically dividing cell lineages, whereas the broader meaning includes the transmission of epigenetic information between any reproducing entities – cells, individuals, and even groups of individuals. In this article, we first describe mechanisms of epigenetic inheritance in cell lineages within organisms, and then consider the evidence for and consequences of the transmission of epigenetic variations between

generations of individuals. More details of the historical development and scope of ideas about epigenetic inheritance can be found in Jablonka and Lamb (2002, 2011).

Mechanisms of Epigenetic Inheritance

The changes that occur during the determination and differentiation of both dividing cells (such as stem cells or germ cells) and nondividing cells (such as brain cells) are established and maintained by a variety of epigenetic mechanisms, some of which enable aspects of the cell's phenotype to be transmitted to descendent cells or organisms. We have called the mechanisms that underlie such transmission 'epigenetic inheritance systems' (EISs), and initially characterized three broad categories: steady-state systems, structural systems, and chromatin-marking systems (Jablonka and Lamb, 1995). A fourth EIS, based on the transmission of small regulatory RNAs, was recognized in the late 1990s (Jablonka and Lamb, 2005, 2010).

Steady-State Systems

Patterns of gene activity can persist in cell lineages through the operation of regulatory feedback loops. The simplest example is a gene that controls its own transcription through positive feedback. When the activity of such a gene is induced by an external stimulus, some of the gene's product binds to its own control region. This binding leads to the continued activity of the gene. In this way, gene activity that initially depended on the presence of an external inducer is maintained by self-regulation – by the continuous induction of the gene by its own product. If the concentration of the gene product is high enough and cell division is more or less equal, the gene's active state will be inherited, because the regulatory gene product will be transmitted to daughter cells as an automatic consequence of cell division. This type of self-regulation means that cells with the same DNA sequence can have different but stable patterns of cell-heritable activities, which either reflect their developmental history or are the result of stochastic switching between alternative states. Many examples of such self-perpetuating regulatory loops have been found in both animals and plants (Jablonka and Raz, 2009). In most cases, the systems are more complicated than the one just described because several genes and gene products are involved in the feedback loop, but the principle is the same.

Structural Inheritance Systems

Cells contain many complex three-dimensional structures whose assembly seems to depend on the presence of preexisting similar structures. In ways that are not fully understood, old structures are used as templates for the production of the new structures that are required for cell growth and multiplication. The best-known examples of this type of inheritance come from studies of single-celled organisms, such as the ciliated protozoan *Paramecium*. Genetically identical paramecia can show heritable differences in the patterns of cilia (tiny hairlike structures) on their cell surface, and experimentally altered patterns can be transmitted to daughter cells for hundreds of cell generations (Beisson, 2008). Although this type of inheritance is particularly conspicuous and stable in ciliated protozoa, structural inheritance is probably universal, since all cells rely on preexisting structures when they produce new ones. The growth of many cell membranes, for example, requires the presence of preexisting membrane (Cavalier-Smith, 2004). The templating properties of some aberrant proteins are believed to underlie the transmission of mad cow disease (bovine spongiform encephalopathy (BSE)) and several other neurodegenerative diseases of humans. In these diseases, misfolded proteins – prions – multiply by inducing function-affecting changes in the conformation of normal proteins. Not all prions are pathological, however. There are reasons for thinking that transmissible, self-propagating, prionlike proteins are involved in regular functions such as memory storage and innate immune responses (Soto, 2012). The best evidence for a functional role for prions comes from yeast, where switching between heritable prion and nonprion forms of certain proteins involved in DNA information processing enables cells to adapt rapidly to changed conditions. In a study of about 700 wild yeast strains, one-third were found to carry prions, many of which affected the phenotype in ways that in some conditions could be adaptive (Halfmann et al., 2012).

Chromatin-Marking Systems

Chromatin, the substance of chromosomes, is a combination of DNA, bound histone and nonhistone proteins, and RNA molecules. It is constantly being remodeled as small chemical appendages such as methyl groups are added to or removed from the nucleotides of DNA, as the amino acids in histones are modified by having chemical appendages added to or taken away from them, and as the various histone and nonhistone proteins and RNA molecules come and go from the complex. The different patterns of chromatin modifications are known as 'chromatin marks' and affect the way that DNA sequences are interpreted (Jablonka and Lamb, 2005, 2010). Epigenetic information that is contained in chromatin marks can be inherited by being carried from one cell generation to the next with replicated DNA. A good example of an inherited chromatin mark is a DNA methylation pattern. In many plants and animals, some of the cytosines in DNA are methylated. Although cytosine (C) is one of the four bases whose sequences encode genetic information, the presence of a methyl group (^{m}C) does not affect its coding properties. However, cytosine methylation within or around a gene often determines whether or not the gene is active, and how easy it is to activate: usually low levels of methylation in noncoding regions are associated with potential activity, while high levels are associated with inactivity. Changes in methylation patterns occur during growth and development, so the same DNA sequence can have several different methylation patterns, each reflecting a different functional state, and each is inherited stably through many cell divisions. One of the mechanisms through which methylation marks are transmitted to daughter cells depends on the way a cytosine in one strand of the DNA double helix is always paired with a guanine (G) in the other. Methylation usually occurs symmetrically on the two strands, most commonly in CG

doublets in which ^{m}C G is paired with G ^{m}C. Following DNA replication, in which the two parental DNA strands separate and are used as templates for building complementary daughter strands, the previously methylated sites remain methylated on the parental strand but are unmethylated on the newly synthesized daughter strand. However, special enzymes recognize the asymmetrical, half-methylated sites, and preferentially methylate the cytosines of the new strand (Holliday, 1990). Thus, because of the mirror symmetry of CG sites and the semiconservative nature of DNA replication, the specific pattern of methylation is reproduced at every cell division. In recent years, new types of DNA modification and additional cellular factors that are involved in methylation and demethylation have been discovered, and it has become clear that interactions between DNA methylation and the other heritable components of chromatin are important, so the overall picture of methylation marking is becoming far more complex (Law and Jacobsen, 2010). Nevertheless, DNA methylation remains the best understood chromatin-marking EIS.

Histone proteins, which form a large part of chromatin, are also important in chromatin marking. Which of the histone variants are present, the position of histones relative to DNA, and the way they are changed through acetylation, methylation, or other modifications of their amino acids all have roles in epigenetic regulation, and the organization and modifications of histones can to some extent be mitotically and sometimes meiotically inherited. Although the molecular mechanisms through which this occurs are far from clear, parental histone assemblies seem to be used as blueprints for restoring similar configurations in daughter cells (Margueron and Reiberg, 2010).

Different types of chromatin marks commonly interact, with the inheritance of one type being influenced by or even dependent on another (Bonasio et al., 2010). The type of histone and nonhistone proteins that are bound to DNA, the nature of the histone modifications, and the patterns of DNA methylation are all interdependent, so understanding the establishment, maintenance, and inheritance of epigenetic marks is extremely difficult.

Small-RNA Systems

Early in the twenty-first century, it became apparent that small RNAs (sRNAs) are central to the regulation of cellular activities and also take part in the transmission of epigenetic information (Ghildiyal and Zamore, 2009). RNA interference (RNAi) – gene silencing through sRNAs – has been found in almost all eukaryotes, and prokaryotes have analogous mechanisms. There are several types of RNAi, but in all of them precursor RNA molecules derived from external sources such as viruses or experimental manipulations, or from endogenous DNA sequences, are cut into sRNAs, usually 20–30 nucleotides long. After processing, these sRNAs are loaded onto protein complexes that bring about the silencing of nucleic acid sequences that are complementary to the sRNAs. The mechanisms through which sRNA–protein complexes silence gene activity include (1) binding to complementary messenger RNA (mRNA) sequences, thereby preventing their translation or causing their degradation; (2) targeting chromosome regions with similar DNA sequences, which they silence through DNA methylation or repressive histone modifications; and (3) excising homologous sequences from chromosomal DNA. Sometimes the products of sRNA-directed mRNA degradation can trigger the transcription of more precursor sRNAs, thus forming a self-sustaining loop. In whichever way sRNA-mediated silencing is achieved, it can be inherited, and there are several (not necessarily mutually exclusive) mechanisms through which this can occur. First, sRNAs can be replicated by an RNA-dependent RNA polymerase, and these replicated RNAs can then be distributed to daughter cells during cell division. Second, when sRNA–protein complexes guide chromatin modifications, the altered chromatin marks are inherited. Third, when sRNA–protein complexes lead to the excision of DNA sequences, the new, depleted DNA is inherited through conventional DNA replication. Fourth, some sRNAs can be moved horizontally from cell to cell.

Although for descriptive purposes we have distinguished four types of systems through which epigenetic information can be transmitted in cell lineages, each category includes multiple, sometimes disparate mechanisms. Moreover, the relative importance of the various mechanisms differs among organisms, and as the sRNA systems show clearly, the four inheritance systems are often interrelated and interdependent (Olovnikov et al., 2012).

Transgenerational Epigenetic Inheritance

Studies of the mechanisms of epigenetic inheritance were initially focused on cells in the cell lineages of multicellular organisms, but subsequently it was found that single-celled organisms such as bacteria and yeasts also transmit epigenetic variations to their descendants, and that multicellular organisms sometimes transmit epigenetic variations between generations (Jablonka and Lamb, 1995). Transgenerational transmission was recognized in two ways: either as genomic imprinting, in which a gene's expression is influenced by the sex of the parent who transmitted it, or as the long-term inheritance of acquired variations over several generations.

Genomic Imprinting

During the formation of sperm and eggs, chromatin is modified. The way that it is modified in the male and female germlines differs, so, at the beginning of embryonic development, the genes transmitted by the two parents are in different epigenetic states. If these differences persist during development and have phenotypic effects, the genes are said to be 'imprinted' (Ferguson-Smith, 2011). Imprints may involve differences in methylation patterns, histone marks, and sRNAs. The effect of genomic imprinting is that the expression of some genes depends on the sex of the parent from which they were inherited (parent-of-origin effects). Normally, imprints are erased and replaced in the germline, so when a gene that was inherited from one sex is transmitted to the next generation through the opposite sex, the imprints are reversed. Imprinting is widespread in both plants and animals and is known to be necessary for normal development. Several human diseases,

including Prader–Willi, Beckwith–Wiedemann, and Angelman syndromes, are associated with imprinting impairments.

Long-Term Epigenetic Inheritance

Epigenetic inheritance that lasts for more than a single generation has been found in all of the taxa in which it has been sought, from bacteria to humans. A literature search by Jablonka and Raz (2009) identified 102 well-documented cases of transgenerational epigenetic inheritance in 42 different species. This survey and subsequent reviews of epigenetic inheritance in plants (Richards, 2011) and mammals (Daxinger and Whitelaw, 2012) leave little doubt that epigenetic variants that can be transmitted through the germline are not uncommon, that several molecular mechanisms are involved in their transmission, and that the fidelity with which they are inherited depends on many different internal and external factors.

The most extensive evidence for epigenetic inheritance undoubtedly comes from work with plants, particularly with the small self-fertilizing thale cress, *Arabidopsis thaliana*, now the most common model organism used by botanical geneticists. Work summarized by Schmitz and Ecker (2012) shows that hundreds of genomic regions can be differentially methylated, and that many methylation marks are stably inherited between generations. The spontaneous rate of production of methylation variants is more than four orders of magnitude greater than that for gene mutations, and when methylation variants are induced, some of the new phenotypes are adaptive and have a high heritability.

In animals, one of the fascinating and important aspects of the evidence for epigenetic inheritance is that it shows that, contrary to the long-held beliefs of most biologists, induced or acquired changes in morphology, physiology, and behavior can be transmitted to future generations. For example, Remy (2010) found that olfactory imprints in the nematode worm *Caenorhabditis elegans* can be inherited. Studies of the same organism have shown that even a complex trait like lifespan can be altered by histone modifiers, and that the induced change – in this case, extended longevity – can be passed to several generations of descendants (Greer et al., 2011). Daxinger and Whitelaw (2012) have summarized data showing that even in mammals, some induced changes are transmitted to subsequent generations. This has been found in the mouse, where the expression of several metastable alleles can be heritably modified by environmental factors. *Agouti viable yellow* (A^{vy}), for instance, is an allele that has a noticeable effect on fur color, but also affects the incidence of diabetes and obesity. The allele's expression is variable, and this variation is correlated with methylation and histone marks in its control region. The chromatin marks and associated strength of gene expression are inherited. Feeding pregnant females with folic acid, betaine, or other methyl donors increases the expression of the A^{vy} allele in their offspring, and this increase is transmitted to subsequent generations, albeit not very stably.

There is substantial and worrying evidence, some of which is summarized by Jirtle and Skinner (2007), that physiological and behavioral variations resulting from exposure to environmental toxins or psychological stresses can be inherited. For example, four generations of male rats whose male ancestors were exposed as fetuses to the fungicide vinclozolin, which is a potent androgen suppressor, showed a greater incidence of testis abnormalities, infertility, kidney and prostate diseases, and immune system problems. They also had different adult responses to stress administered during adolescence (Crews et al., 2012). These changes were correlated with altered DNA methylation in the male germ line, and changes in gene activity in the brain. Psychological traumas are also known to affect future generations: mice that had been separated from their mothers for a few hours each day for the first 14 days of their lives later showed various behavioral and emotional abnormalities, which were transmitted by males to their offspring. The transmitted behaviors were associated with changes in germline and brain DNA methylation, and in gene expression (Franklin et al., 2010). In humans, starvation, psychological stress, and exposure to various toxins are all known to have long-term effects on physiology and behavior, and in some cases these effects have documented or suspected epigenetic underpinnings (Thayer and Kuzawa, 2011). Evidence for transgenerational transmission in humans is limited, but using surveys and historical data, Pembrey and his colleagues concluded that both smoking and food supply can have effects that are transmitted transgenerationally through the male line (Pembrey et al., 2006; *see also* Twin Studies and Epigenetics).

Genetic and Epigenetic Inheritance Compared

The properties of genetic and epigenetic inheritance in cell lineages overlap, but certain features are more characteristic of one than the other. Most differences stem from the fact that epigenetic systems are concurrently both transmission and response systems, whereas the genetic system is specialized more for transmitting information than for responding to internal or external cues. The response function of epigenetic systems means that, unlike most genetic variations, the occurrence of new heritable epigenetic variations is not random – variations are often induced by specific developmental signals or environmental challenges, and are targeted to particular regions of the genome. Sometimes, induced epigenetic variations are adaptive, although they need not be so. Whereas changes in the DNA sequence of a gene are rare, and reversion to the original sequence is even rarer, the response function of EISs means that, under some circumstances, especially in conditions of stress, the rate of change between epigenetic variants is very high, and coordinated changes may occur in the epigenetic marks of several genes. However, although some inherited epigenetic changes differ from typical genetic changes in being targeted, inducible, frequent, reversible, and adaptive, others resemble classical genetic mutations in that they are rare, random, nonadaptive, and unlikely to revert. Because changes in DNA can sometimes be targeted and involve many loci (Shapiro, 2011), no single criterion can be used to distinguish between the properties of the genetic and epigenetic systems, and molecular investigations are necessary to unequivocally identify the epigenetic or genetic nature of heritable variation.

Implications

Recognizing the ubiquity and developmental importance of cellular epigenetic inheritance has major implications for medicine, ecology, and related disciplines (Gilbert and Epel, 2009), and also for the way that we think about heredity and evolution (Jablonka and Lamb, 2010; Jablonka, 2013). It has forced the biological and medical communities to move away from the idea that heredity is determined solely by information carried in DNA sequences. In this article, we have focused on transmission in cell lineages and through the gametes, but as we indicated earlier, in its broad sense epigenetic inheritance includes routes of information transmission that bypass the germline. It therefore includes behavioral transmission of the type revealed by Weaver and his colleagues in laboratory rats (Weaver et al., 2004). These scientists found that the mothering style of rats – how much nursing, licking, and grooming they gave their pups – affected the way in which these youngsters responded to stress when adult, and the type of mothering that the adult females gave to their own pups. Rats raised by less nurturing mothers were more sensitive to stress than those raised by more caring mothers, and the female offspring of mothers who gave deficient care themselves became less caring mothers. The differences in mothering style and responses to stress were associated with epigenetic differences in chromatin marks and the expression of certain genes in the brain. It was possible to reverse the behavior of the timid adult offspring of the less caring mothers by using chemicals that changed their chromatin marks. However, although the poor mothering style and consequent fearful adult behavior were perpetuated in families, cross-fostering experiments showed that the epigenetic differences were not transmitted through the germline: foster mothers transmitted their style of care to their adopted pups as well as their own offspring. Therefore, although epigenetic marking was involved, transmission was through the mother rats' behavior, not through epigenetic marks carried in their germ cells.

Practical Implications

The importance of all types of heritable epigenetic variations in health and disease is today widely recognized. First, it is known that many cancers are associated with cell-heritable epigenetic modifications, such as alterations in methylation patterns and other aspects of chromatin organization. Second, chronic diseases such as some types of diabetes, coronary diseases, and mental disorders have been shown to have epigenetic correlates. Third, some hereditary diseases are caused by mutated (or epimutated) imprinted genes, which show peculiar but characteristic patterns of inheritance. Fourth, diseases such as BSE and kuru are caused by prions – transmissible, infectious, and abnormally folded proteins. Fifth, it is accepted that in order to cope with pathogenic microorganisms and other parasites, the effects of their adaptive epigenetic systems have to be countered. This knowledge is being incorporated into medical practice and epidemiological research programs, and new ways of recognizing, avoiding, and curing the diseases caused by heritable epigenetic changes are being developed.

Epigenetics has provided powerful tools for innovations in medicine and agriculture. Many of the early discoveries in the field were stimulated by attempts to solve problems associated with the low success rate of reproductive technologies such as cloning. Now that greater knowledge has made it easier to manipulate epigenetic marking, this technique is more successful. In addition, the possibility of replacing diseased or damaged tissues by using cells that have been epigenetically reprogrammed to become stem cells has made regenerative medicine a likely future growth area. The potential economic benefits of exploiting the sRNA systems were obvious as soon as the RNAi system was recognized, and private and public money has been pouring into this area of research. By introducing synthetic RNAs into cells, specific genes can be heritably silenced; not only is this a powerful tool for investigating gene functions, but also it provides a novel strategy for treating human diseases and for modifying commercial plants and animals.

The importance of epigenetic inheritance for ecological and environmental issues is acknowledged by the fact that a new discipline, ecological epigenetics, is beginning to emerge (Bossdorf et al., 2008; Ledón-Rettig et al., 2013). Most biologists now realize that the changes in ecological conditions that are taking place as a result of global warming will induce changes in physiology and behavior, some of which will be heritable and lead to new adaptations and new problems. The effects of chemical environmental pollutants are also likely to have enduring consequences. Not only do many of these chemicals persist in the environment for long periods, but also some induce epigenetic changes that are heritable, so even brief exposures can have long-term detrimental effects. For example, work with rats has shown that fetal exposure to low doses of some commonly used chemical pesticides, plastics, and hydrocarbon fuels can lead to epigenetic changes in the age of puberty, and alter gonadal functions in ways that reduce fertility. Each compound caused a distinct set of alterations in DNA methylation, and the induced epigenetic effects were inherited for three generations (Manikkam et al., 2012).

Although it is rarely recognized, epigenetic inheritance also has important implications for the social sciences. Thayer and Kuzawa (2011) are among those who have suggested that some of the persistent social and health disparities between different classes or races may be consequences not only of recent socioeconomic inequalities but also of the germline transmission of epigenetic effects resulting from the experiences of earlier generations. If so – and the work cited earlier in this article suggests it is likely to be so – it will be necessary to identify, compensate, and reverse any adverse inherited epigenetic effects, as well as deal with the current causes of inequality.

A Broader Concept of Inheritance and Evolution

When epigenetic inheritance and other non-DNA inheritance systems (such as the transmission of information through behavior or language; *see* Gene-Culture Coevolution) are included in heredity, it becomes a much richer and more general concept. No longer does heredity depend solely on variations in the sequences of nucleotides in DNA. Even at the cellular level of inheritance, the production and transmission

of variations involve several different types of molecules and several mechanisms. With some of these mechanisms, such as those underlying structural inheritance and steady-state systems, there is no real equivalent of the phenotype–genotype distinction, because the developmental reconstruction of the phenotype is an integral part of the transmission mechanism (*see* Genotype and Phenotype).

With all EISs, the inheritance of acquired (induced) variation is a common and regular occurrence, with the environment being both an inducer and a selector of epigenetic variations. Epigenetic inheritance can therefore be seen as a form of 'soft inheritance.' Ernst Mayr, one of the twentieth century's most influential evolutionary biologists, described soft inheritance as "the belief in a gradual change of the genetic material itself, either by use or disuse, or by some internal progressive tendencies, or through the direct effect of the environment" (Mayr, 1998, p. 15). Because he identified inheritance with genes and DNA, Mayr asserted, "It was perhaps the greatest contribution of the young science of genetics to show that soft inheritance does not exist" (Mayr, 1998, p. 17). Throughout the second half of the twentieth century, Mayr campaigned vigorously against the inclusion of soft inheritance in evolutionary thinking, trying to consolidate a view of Darwinism that was free of any Lamarckian overtones. For him and his many followers, the only heritable changes that were important in evolution were those that occurred in DNA. Consequently, there was a delay in recognizing the significance of epigenetic inheritance, which certainly reintroduces Lamarckian notions into evolutionary theorizing (Gissis and Jablonka, 2011). Today, because epigenetic inheritance and other types of nongenetic information transmission across generations are acknowledged to be widespread, the existence of 'soft' inheritance is no longer in doubt (Bonduriansky, 2012). Biologists are accepting a more pluralistic notion of heredity.

Theoretical models that incorporate an expanded concept of heredity are showing that epigenetic inheritance can have profound effects on the rate and direction of evolution. First, spontaneous epigenetic changes occur much more frequently than genetic changes, so adaptive responses to new challenges can be rapid. Second, epigenetic variants that are repeatedly induced by persistent environmental conditions can produce phenotypic change in a population without selection. Third, even variants that are not stably transmitted can be adaptive. For example, inheriting an epigenetic defense against a pathogen for just a few generations can be a good evolutionary strategy if the type of pathogen in the environment changes frequently, and evolutionary modeling has shown that the fidelity with which epigenetic variants are transmitted should evolve to match environmental fluctuations (Jablonka and Lamb, 1995). Fourth, as well as being a source of selectable variation in their own right, operating alongside the genetic system, EISs have a very significant influence on how the frequencies of genetic variants are changed by selection. If, for instance, an allele that previously had been heritably silent and therefore invisible to selection is heritably activated, even for a few generations, then the allele's frequency in the population may change because the epigenetic modification has made it visible. New theoretical frameworks are being developed that make it possible to analyze the interactions of the genetic and epigenetic systems, and these are showing that incorporating nongenetic inheritance drastically alters predictions about the course of evolutionary change that are based solely on genetic inheritance (Day and Bonduriansky, 2011; Geoghegan and Spencer, 2012). In addition to its strong effects at the population level, epigenetic inheritance is important in the process of speciation, and contributes to novel, large-scale, adaptive changes (Jablonka and Lamb, 1995, 2010).

The broadening of the concept of heredity, and recognition that natural selection acts on several different types of heritable variation, which began in the last decades of the twentieth century, has challenged the widely accepted gene-centered neo-Darwinian version of Darwinism (*see* Darwinism; Evolution, History of). A more comprehensive and more powerful Darwinian evolutionary theory is now being constructed, one that incorporates multiple and interacting inheritance systems.

See also: Behavioral Genetics: Psychological Perspectives; Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Darwinism; Evolution, History of; Familial Studies: Genetic Inferences; Gene–Culture Coevolution; Genotype and Phenotype; Lifespan Development, Theory of; Lifespan Development: Evolutionary Perspectives; Twin Studies and Epigenetics.

Bibliography

Beisson, J., 2008. Preformed cell structure and cell heredity. Prion 2, 1–8.

Bonasio, R., Tu, S., Reinberg, D., 2010. Molecular signals of epigenetic states. Science 330, 612–616.

Bonduriansky, R., 2012. Rethinking heredity, again. Trends in Ecology and Evolution 27, 330–336.

Bossdorf, O., Richards, C.L., Pigliucci, M., 2008. Epigenetics for ecologists. Ecology Letters 11, 106–115.

Cavalier-Smith, T., 2004. The membranome and membrane heredity in development and evolution. In: Hirt, R.P., Horner, D.S. (Eds.), Organelles, Genomes and Eukaryote Phylogeny: An Evolutionary Synthesis in the Age of Genomics. CRC Press, Boca Raton, FL, pp. 335–351.

Crews, D., Gillette, R., Scarpino, S.V., Manikkam, M., Savenkova, M.I., Skinner, M.K., 2012. Epigenetic transgenerational inheritance of altered stress responses. Proceedings of the National Academy of Sciences USA 109, 9143–9148.

Daxinger, L., Whitelaw, E., 2012. Understanding transgenerational epigenetic inheritance via the gametes in mammals. Nature Reviews Genetics 13, 153–162.

Day, T., Bonduriansky, R., 2011. A unified approach to the evolutionary consequences of genetic and nongenetic inheritance. American Naturalist 178, E18–E36.

Ferguson-Smith, A.C., 2011. Genomic imprinting: the emergence of an epigenetic paradigm. Nature Reviews Genetics 12, 565–575.

Franklin, T.B., Russig, H., Weiss, I.C., Gräff, J., Linder, N., Michalon, A., Vizi, S., Mansuy, I.M., 2010. Epigenetic transmission of the impact of early stress across generations. Biological Psychiatry 68, 408–415.

Geoghegan, J.L., Spencer, H.G., 2012. Population-epigenetic models of selection. Theoretical Population Biology 81, 232–242.

Ghildiyal, M., Zamore, P.D., 2009. Small silencing RNAs: an expanding universe. Nature Reviews Genetics 10, 94–108.

Gilbert, S.F., Epel, D., 2009. Ecological Developmental Biology: Integrating Epigenetics, Medicine, and Evolution. Sinauer, Sunderland, MA.

Gissis, S.B., Jablonka, E. (Eds.), 2011. Transformations of Lamarckism: From Subtle Fluids to Molecular Biology. MIT Press, Cambridge, MA.

Greer, E.L., Maures, T.J., Ucar, D., Hauswirth, A.G., Mancini, E., Lim, J.P., Benayoun, B.A., Shi, Y., Brunet, A., 2011. Transgenerational epigenetic inheritance of longevity in *Caenorhabditis elegans*. Nature 479, 365–371.

Halfmann, R., Jarosz, D.F., Jones, S.K., Chang, A., Lancaster, A.K., Lindquist, S., 2012. Prions are a common mechanism for phenotypic inheritance in wild yeasts. Nature 482, 363–368.

Holliday, R., 1990. Mechanisms for the control of gene activity during development. Biological Reviews of the Cambridge Philosophical Society 65, 431–471.

Jablonka, E., 2013. Epigenetic inheritance and plasticity: the responsive germline. Progress in Biophysics and Molecular Biology, in press.

Jablonka, E., Lamb, M.J., 1995. Epigenetic Inheritance and Evolution: The Lamarckian Dimension. Oxford University Press, Oxford.

Jablonka, E., Lamb, M.J., 2002. The changing concept of epigenetics. Annals of the New York Academy of Sciences 981, 82–96.

Jablonka, E., Lamb, M.J., 2005. Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life. MIT Press, Cambridge, MA.

Jablonka, E., Lamb, M.J., 2010. Transgenerational epigenetic inheritance. In: Pigliucci, M., Müller, G.B. (Eds.), Evolution – The Extended Synthesis. MIT Press, Cambridge, MA, pp. 137–174.

Jablonka, E., Lamb, M.J., 2011. Changing thought styles: the concept of soft inheritance in the 20th century. In: Egloff, R., Fehr, J. (Eds.), Vérité, Widerstand, Development: At Work with/Arbeiten mit/Travailler avec Ludwik Fleck, (Collegium Helveticum Heft 12). Collegium Helveticum, Zürich, pp. 119–156.

Jablonka, E., Raz, G., 2009. Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. Quarterly Review of Biology 84, 131–176.

Jirtle, R.L., Skinner, M.K., 2007. Environmental epigenomics and disease susceptibility. Nature Reviews Genetics 8, 253–262.

Law, J.A., Jacobsen, S.E., 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics 11, 204–220.

Ledón-Rettig, C.C., Richards, C.L., Martin, L.B., 2013. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Behavioral Ecology, 24, 311–324.

Manikkam, M., Guerrero-Bosagna, C., Tracey, M., Haque, M.M., Skinner, M.K., 2012. Transgenerational actions of environmental compounds on reproductive disease and identification of epigenetic biomarkers of ancestral exposures. PLoS One 7, e31901.

Margueron, R., Reiberg, D., 2010. Chromatin structure and the inheritance of epigenetic information. Nature Reviews Genetics 11, 285–296.

Mayr, E., 1998. Prologue: some thoughts on the history of the evolutionary synthesis. In: Mayr, E., Provine, W. (Eds.), The Evolutionary Synthesis: Perspectives on the Unification of Biology, second ed. Harvard University Press, Cambridge, MA, pp. 1–48.

Olovnikov, I., Aravin, A.A., Fejes Toth, K., 2012. Small RNA in the nucleus: the RNA-chromatin ping-pong. Current Opinion in Genetics and Development 22, 164–171.

Pembrey, M.E., Bygren, L.O., Kaati, G., Edvinsson, S., Northstone, K., Sjöström, M., Golding, J., the ALSPAC Study Team, 2006. Sex-specific, male line transgenerational responses in humans. European Journal of Human Genetics 14, 159–166.

Pinto-Correia, C., 1997. The Ovary of Eve: Egg and Sperm and Preformation. University of Chicago Press, Chicago, IL.

Remy, J.-J., 2010. Stable inheritance of an acquired behavior in *Caenorhabditis elegans*. Current Biology 20, R877–R878.

Richards, E.J., 2011. Natural epigenetic variation in plant species: a view from the field. Current Opinion in Plant Biology 14, 204–209.

Schmitz, R.J., Ecker, J.R., 2012. Epigenetic and epigenomic variation in *Arabidopsis thaliana*. Trends in Plant Science 17, 149–154.

Shapiro, J.A., 2011. Evolution: A View from the 21st Century. FT Press Science, Upper Saddle River, NJ.

Soto, C., 2012. Transmissible proteins: expanding the prion heresy. Cell 149, 968–977.

Thayer, Z.M., Kuzawa, C.W., 2011. Biological memories of past environments: epigenetic pathways to health disparities. Epigenetics 6, 798–803.

Waddington, C.H., 1975. The Evolution of an Evolutionist. Edinburgh University Press, Edinburgh.

Weaver, I.C.G., Cervoni, N., Champagne, F.A., D'Alessio, A.C., Sharma, S., Seckl, J.R., Dymov, S., Szyf, M., Meaney, M.J., 2004. Epigenetic programming by maternal behavior. Nature Neuroscience 7, 847–854.

Epigenetics

Jonathan D Turner and Sophie A Kirschner, Institute of Immunology, Centre de Recherche Public – Santé/Laboratoire National de Santé, Luxembourg, Grand Duchy of Luxembourg; and Graduate School of Psychobiology, University of Trier, Trier, Germany
Anne M Molitor and Konstantin Evdokimov, Institute of Immunology, Centre de Recherche Public – Santé/Laboratoire National de Santé, Luxembourg, Grand Duchy of Luxembourg
Claude P Muller, Institute of Immunology, Centre de Recherche Public – Santé/Laboratoire National de Santé, Luxembourg, Grand Duchy of Luxembourg; and Graduate School of Psychobiology, University of Trier, Trier, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The external environment plays a primordial role in determining gene expression, and protein production, which may in turn define a behavioral or physiological phenotype. This control occurs at both the transcriptional and translational levels and involves a wide range of mechanisms, providing a plethora of opportunities for altering the final level of functional proteins. These mechanisms all share the common feature that they do not alter the underlying genomic sequence, but they act as biological annotations to the genome. In this article we review these mechanisms, and how they are related to health. In particular the role of the three epigenetic mechanisms in controlling the glucocorticoid receptor, as well as the stress phenotype, is reviewed in detail.

The Three Principal Mechanisms of Epigenetic Regulation

DNA Methylation

Cytosine methylation is the principal epigenetic mechanism operating directly on the genomic DNA, and it is the only natural covalent DNA modification in vertebrates. Methylation occurs in mammals only on the 5′cytosine in a CpG pair, that is, on cytosines immediately followed by a guanine. Between 60 and 90% of genomic CpGs are thought to be methylated. CpG dinucleotides are not equally distributed throughout the genome, but are found in clusters, called CpG islands where they are significantly overrepresented. In the human genome there are about 29 000 CpG islands, covering about 1% of the genome. Dinucleotides within CpG islands have long been thought to be protected against methylation, while those throughout the remaining 99% of the genome are mostly methylated.

DNA methylation patterns are tissue specific. They are established during embryogenesis, and they are maintained during normal cell division (mitosis). However, during germ cell division (meiosis) both parental genomes are demethylated and during embryogenesis individual fetal methylation patterns are reestablished. This is particularly obvious in monozygotic twins born with virtually identical methylation patterns, that are different from those of their parents. Differences between the twins only develop after birth, and increase with age. In general, global methylation levels tend to decrease with age (Fraga et al., 2005). In mammalian cells, CpG methylation plays an important role in the regulation of tissue-specific gene transcription and silencing. It is also considered one of the principal interfaces between the environment and the host genome (Fraga et al., 2005; Szyf, 2011), and explains phenotypic variations in humans that are caused by the environment and have long-term programming effects.

Although methylation patterns may remain stable over long time periods, and global methylation levels from repetitive elements decrease at rates of only 0.06–0.12% per year (Bollati et al., 2009), DNA methylation is reversible either by enzymatic removal (He et al., 2011; Ito et al., 2011) or by the inhibition of maintenance of DNA methyltransferases such as DNMT1 during mitosis (**Figure 1**).

DNA 'Packaging' – Chromatin Structure and Modifications

In all eukaryotes the activity of genomic DNA is tightly regulated within the context of a nucleoprotein structure called chromatin. The fundamental unit of chromatin is the nucleosome. The nucleosome consists of an octamer of histones, including two copies of each core histone (e.g., H2A, H2B, H3, and H4) around which 146–147 base pairs of DNA are wrapped. Nucleosomes create a physical barrier to the accessibility of the underlying DNA. 'Open' or 'closed' conformations regulate fundamental DNA-template processes like transcription, DNA replication, DNA-damage repair, or DNA recombination. Chromatin can be dynamically remodeled by DNA methylation or covalent posttranslational histone modifications both of which provide a link between environmental conditions and the transcription of the genes in the remodeled area.

Posttranslational Modifications of Core Histones and Their Role in Transcriptional Regulation

The flexible N-terminal tails of the core histones protrude from the nucleosome body and are freely accessible to a growing list of histone-modifying enzymes that deposit (e.g., acetyltransferases, methyltransferases, kinases, ubiquitinases) or remove (e.g., deacetylases, demethylases, phosphatases, deubiquitinases) various covalent posttranslational modifications (PTMs). Histone PTMs are epigenetic marks playing an important role in defining the chromatin state that ultimately controls the transcriptional activity (Bannister and Kouzarides, 2011). Even though robust, they are considered

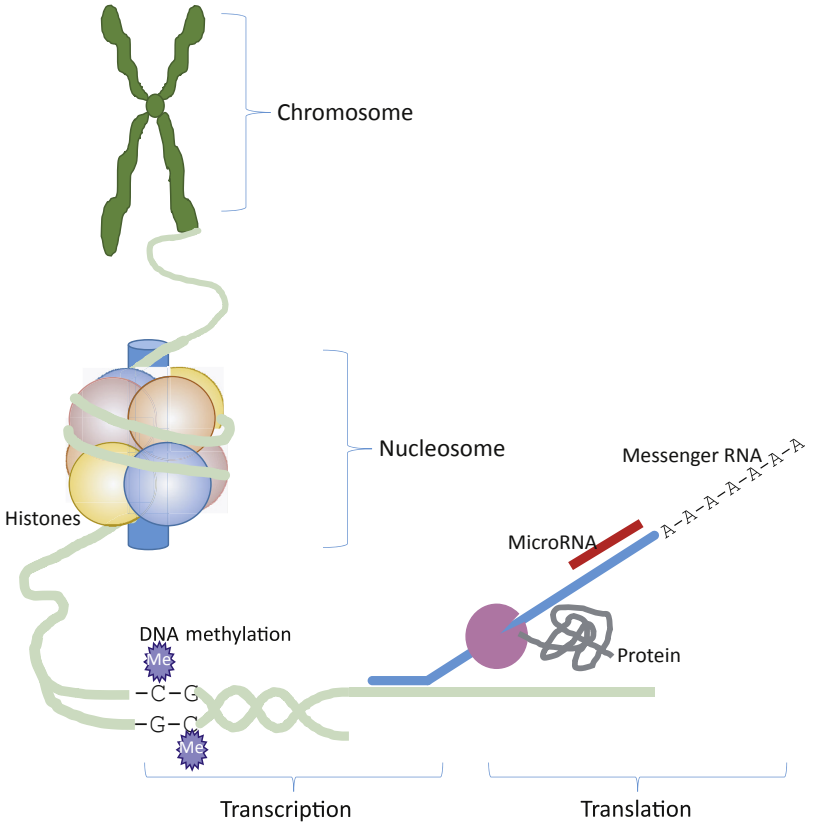

Figure 1 Schematic representation of the principal levels of epigenetic regulation. Genomic DNA is packaged in chromosomes consisting of nucleosomes with 146–147 basepairs of genomic DNA wrapped round them. The first level of epigenetic regulation is the modification of the histones that make the nucleosome. The second level of epigenetic regulation is DNA methylation, occurring on cytosines in CpG dinucleotides. After transcription, the mRNA is targeted as the third level of regulation by microRNAs. Targeting of mRNA regulates the production of protein from the mature messenger RNA.

to be more flexible than the CpG methylation marks. The link between PTMs and transcriptional activity depends essentially on the chemical nature of the covalent modification as well as the position and nature of the modified histone residue. In general, histone PTMs exert their effects via two main mechanisms: either they regulate biological processes by directly altering the overall chromatin structure or they influence the recruitment of secondary effectors that ultimately define the functional outcome of the PTMs. The combinatorial nature of the large number of available modifications and modification sites provides a 'histone code,' with specific modifications corresponding to either transcriptional repression or activation (summarized in Table 1; Figure 2).

Histone Acetylation

Histone acetylation directly influences the chromatin structure. The positive charge added by acetyl groups loosens the interactions between histones and the negatively charged DNA, resulting in a less tight packaging of the chromatin structure that renders the underlying DNA more accessible to the transcriptional machinery. Histone acetylation of lysine (K) residue side chains is performed by histone acetyltransferases and can be actively reversed by histone deacetylases (HDACs) (Yang and Seto, 2008). Histone acetylation is closely linked to DNA methylation. Cytosine methylation levels are interpreted by two families of methyl-CpG-binding proteins (Klose and Bird, 2006), the methyl-binding domain (MBD) family, and the Kaiso family. The interaction of these methylation interpreting

Table 1 The 'histone code' of common core histone modifications found in psychobiological paradigms and phenotypes

	Histone H3					*Histone H4*	*Histone H2B*
Residue	K4	K9	K27	K36	K79	K20	K5
Monomethyl			Activation		Activation	Repression	Activation
Dimethyl	Activation		Repression	Activation	Activation		
Trimethyl	Activation	Repression	Repression		Activation + Repression	Repression	Repression
Acetylation	Activation	Activation	Activation				Activation

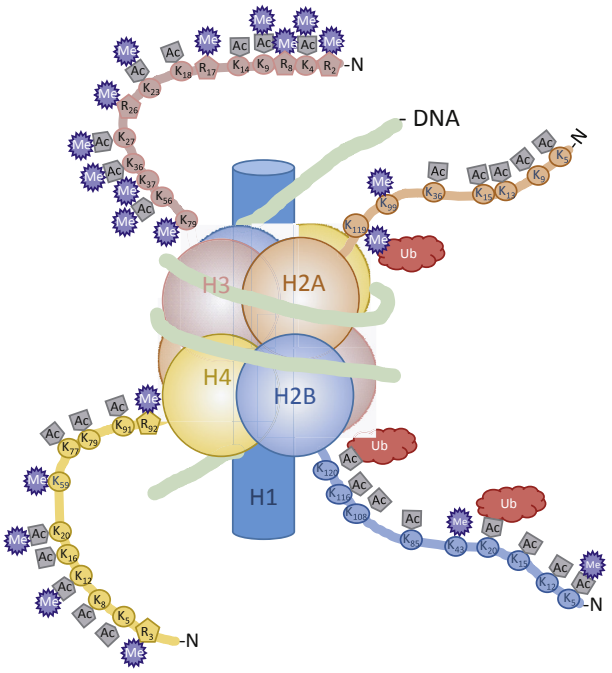

Figure 2 Schematic representation of the nucleosome, and histone tail modifications. The nucleosome consists of four core histones: H2A, H2B, H3, and H4. Epigenetic modifications occur on the histone tails. Modifications to the Lysine (K) and Arginine (R) residues shown are Ac acetylation, Me methylation, Ub ubiquitination. The linker histone, H1, has similar modifications that are not illustrated.

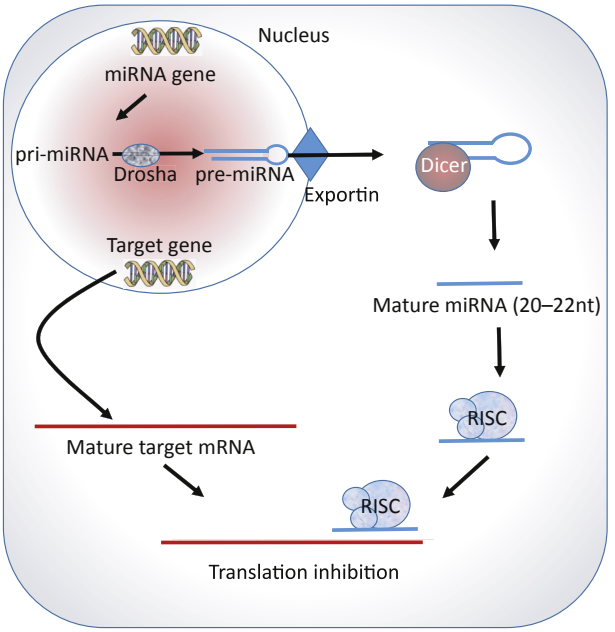

Figure 3 The principal of microRNA (miRNA) synthesis and actions. In the nucleus both the miRNA and the target are transcribed. Initially pri-miRNA is produced that is processed into pre-miRNA by Drosha and subsequently exported from the nucleus by Exportin. After processing by Dicer, the mature miRNA is incorporated into the RNA-induced silencing complex (RISC) and subsequently binds to the target mature mRNA.

proteins with HDAC complexes initiates the condensation of chromatin into transcriptionally closed structures (Prokhortchouk et al., 2001).

Histone Monoubiquitination

Reversible monoubiquitination (ub1) consists of the covalent addition of a single ubiquitin peptide (76 amino acids) by the sequential catalytic activity of an ubiquitin-activating enzyme (E1), an ubiquitin conjugation enzyme (E2), and an ubiquitin ligase (E3). The monoubiquitin peptide is about two-thirds of the size of a histone and thus constitutes an additional bulk that most probably leads to a profound change in the overall nucleosome conformation, and can be either a marker for high transcriptional activity for example, on H2B, or on the contrary for transcriptional repression for example, on H2A.

Histone Methylation

Histone methyltransferases deposit and histone demethylases remove methylation marks on the side chains of more than 20 different K or arginine (R) residues. Each K residue can be mono-, di-, or trimethylated and R residues, mono- or dimethylated further increasing complexity (Bannister and Kouzarides, 2011). In contrast to histone mono-ubiquitination and acetylation, histone methylations do not induce obvious changes in the chromatin structure, although they increase the hydrophobicity of the side chains potentially creating new binding surfaces for reader proteins, influencing transcriptional activity through altered recruitment of the transcriptional machinery (Bannister and Kouzarides, 2011).

MicroRNAs

MicroRNAs (miRNA) are short endogenous noncoding RNAs, which are involved in the posttranscriptional regulation of gene expression (Kim et al., 2009) by binding to the mRNA or targeted genes. miRNAs are usually transcribed, like normal coding genes, by RNA polymerase II, and go through a series of maturation steps (**Figure 3**), to make the final approximately 22 nucleotides long miRNA. This then forms the RNA-induced silencing complex (RISC) with Argonaute protein family (Kim et al., 2009), which in turn binds to mRNA by imperfect base pairing. Typically miRNAs anneal to 3′ untranslated regions of mRNAs. Binding of the RISC complex to mRNA inhibits mRNA translation, triggers mRNA deadenylation, degradation, and cleavage (Saraiya et al., 2013).

miRNAs Link the Environment to the Final Functional Protein Levels

miRNAs have been implicated in the translational control of many proteins, including key players in the hypothalamic–pituitary–adrenal (HPA) axis such as the glucocorticoid receptor (GR) (Vreugdenhil et al., 2009), as well as in many pathologies that have been linked to stress such as cardiovascular disease (Devaux et al., 2012), and depression (O'Connor et al., 2012). Stress itself has recently been associated with tissue-specific changes in intracellular miRNA levels in the frontal cortex but not in the hippocampus after restraint stress (Rinaldi et al., 2010), or after anxiety (Haramati et al., 2011). Similarly, after ACTH administration, mimicking HPA axis activation, Riester et al. observed a significant increase in a number of miRNA in the adrenal gland (Riester et al., 2012).

Genetics or Epigenetics?

Genetics versus Epigenetics

Along with the efforts of the International Human Genome Project to complete the first euchromatin sequence of the human genome (Hattori, 2005), a new generation of highly efficient microarray and sequencing-based techniques were developed. With a human reference genome and these powerful techniques new strategies such as genome-wide association studies (GWAS) became possible. In large-scale studies common genetic variants, called single nucleotide polymorphisms (SNPs) were associated with specific traits or phenotypes. The first successful attempt was published in 2005, when age-related macular degeneration was found to be strongly associated with mutations in the gene for complement factor H, encoding a protein involved in the regulation of inflammation (Edwards et al., 2005). However, GWAS showed its limitations when applied to more complex diseases depending both on genetic and environmental factors. The low power and resolution did not allow identifying multiple variants with modest effects and rare variants could only be detected in very large cohorts (Manolio et al., 2009). Moreover, to avoid unreliable or false-positive results, appropriate controls and arduous statistical analyses were essential. GWAS also made the assumption that genetic variants are the main reason for heritable variability in common diseases, which is disputable. For instance, genetic risk scores from a combination of the 24 most discriminating SNPs were statistically linked to coronary artery calcification; however, they explained only 2.5% of the variance observed in this disease (Van Setten et al., 2013). The importance of epigenetic parameters rather than classical genetic factors (such as SNPs or copy number variants) was highlighted in a recent longitudinal study by Baccarelli et al. (2010). The hazard ratio associated with lower global methylation levels and the incidence of cardiovascular disease was significantly higher than, for example, that of preexisting diabetes or smoking (4.0 vs 2.2 or 1.3 for ischemic heart disease; 5.7 vs 3.2 or 2.2 for stroke) and was independent of established risk factors. Although a direct comparison between epigenetic studies and GWAS is difficult, the rs1994016 locus has been shown to be associated with angiographic coronary artery disease with an odds ratio of 1.19 (1.13–1.24 95% CI). This is below the odds ratio of established risk factors such as smoking, and significantly below that of lower global DNA methylation levels (Reilly et al., 2011). There is also growing evidence that a large number of well-established cardiovascular disease risk factors operate through epigenetic pathways including smoking (Talikka et al., 2012), diabetes (Drong et al., 2012), hypertension (Xu et al., 2011), as well as the prenatal environment (Eriksson et al., 2011). Although cardiovascular diseases and psychopathologies may not share the same biological pathways and mechanisms, the latter are equally complex and depend both on genetic and environmental factors.

Epigenetic Modifications Can Be Inherited – DNA Methylation and miRNA

During the past decade it has become increasingly clear that nongenetic information can also be passed on to the next generations.

In 2009, Roth et al. used a rat model of infant maltreatment to investigate the epigenetic mechanisms involved in early-life adverse exposure. In rats with abusing and neglecting mothers expression of the *Bdnf* gene was consistently reduced in the prefrontal cortex where the promoter was also hypermethylated (Roth et al., 2009). Interestingly, the female offspring exposed to early-life maltreatment presented a similar poor maternal behavior and the same altered *Bdnf* DNA methylation pattern. When the offspring of abusive mothers were cross-fostered by normal mothers, methylation levels of the *Bdnf* promoter did not significantly change. This was the first evidence that DNA methylation induced by early-life maltreatment can be transmitted across generations. The exact mechanisms of epigenetic inheritance still remain unclear. In 2013, Rodgers et al. showed that paternal stress reprograms the HPA axis of the next generation regulation through alterations in sperm miRNA composition. Before breeding, male mice were exposed to 6 weeks of chronic stress, either during puberty or adulthood. This stress regimen perturbed the HPA axis of the males and this new phenotype was transmitted to the next generation along with the perturbed gene expression and DNA methylation levels (Rodgers et al., 2013). To elucidate transmission mechanisms, miRNA composition of the sperm of the fathers was investigated. Nine miRNAs were found significantly overexpressed, four of which targeted the de novo methylation enzyme Dnmt3a, and two proteins involved in the miRNA processing, Tnrc6b and Mtdh. These results are suggestive of a dynamic epigenetic transmission across generations. However, the authors could not demonstrate whether the paternal stress or the sperm miRNA composition was directly responsible for the epigenetic reprogramming of the offspring. Nevertheless, these different reports provide strong evidence for epigenetic mechanisms in the inheritance of phenotypes shaped by environmental exposure.

In humans there are so far only few examples of epigenetic inheritance. Fetal malnutrition has effects that persist into subsequent generations. Mothers who were born during the Dutch winter famine (1944–45) and who were exposed to severe malnutrition, had sequelae that were still visible in their own children (Painter et al., 2008). Mothers born with significantly reduced birth weight usually had children with normal birth weight, but of decreased length at birth, and increased ponderal index. Their children also had a 1.8 (95% CI 1.1–2.7) times greater risk of poor overall health in later life.

There is also circumstantial evidence for trans-generational programming of the HPA axis. Children of female Holocaust survivors that developed post-traumatic stress disorder (PTSD) had lower mean cortisol levels across a complete 24-h cycle, than those of survivors without PTSD, suggesting either an inherited programming, or a response to altered maternal behavior during a susceptible period in early life (Yehuda et al., 2007).

Epigenetics in Behavior and Psychology

Epigenetics of the Stress Response

External challenges trigger reactions both in the central nervous system and in peripheral organs. As one of the first reactions, stress activates the neuronal circuits of the HPA axis. Expression

patterns of genes in the limbic brain regions (e.g., prefrontal cortex, hippocampus, amygdala, nucleus accumbens) are modulated and behavioral phenotypes are adopted. In both developing tissues during early life and in fully differentiated adult cells, the transcriptional response to environmental signals is controlled at the epigenetic level (Meaney, 2007; Szyf, 2009). During and acute challenge the effects are transient and, in general, return to basal level once the stimuli disappeared to reinstate the initial homeostasis (McEwen, 2008; Rozanov, 2012), however both the reaction and initial homeostasis can be epigenetically programmed.

Epigenetic mechanisms have been implicated in social and maternal behavior (Weaver et al., 2004), learning and memory (Blaze and Roth, 2013), drug addiction (Vialou et al., 2013), and in several stress-related neurodevelopmental and psychopathological disorders (Kubota et al., 2012; McEwen et al., 2012). Stress response genes have been clearly shown to be susceptible to epigenetic methylation as highlighted in **Figure 4**. For instance GR expression in the hippocampus, essential in regulating the stress response, is susceptible to DNA methylation in both rodents and man (McGowan et al., 2009; Weaver et al., 2004; Turner et al., 2008). Methylation of both the rat and human GR inversely correlates with receptor expression levels (Cao-Lei et al., 2013; Witzmann et al., 2012), however there is some debate as to whether this regulation is due to methylation of individual CpGs (McGowan et al., 2009; Weaver et al., 2004), or wider promoter-wide regions (Cao-Lei et al., 2013; Witzmann et al., 2012).

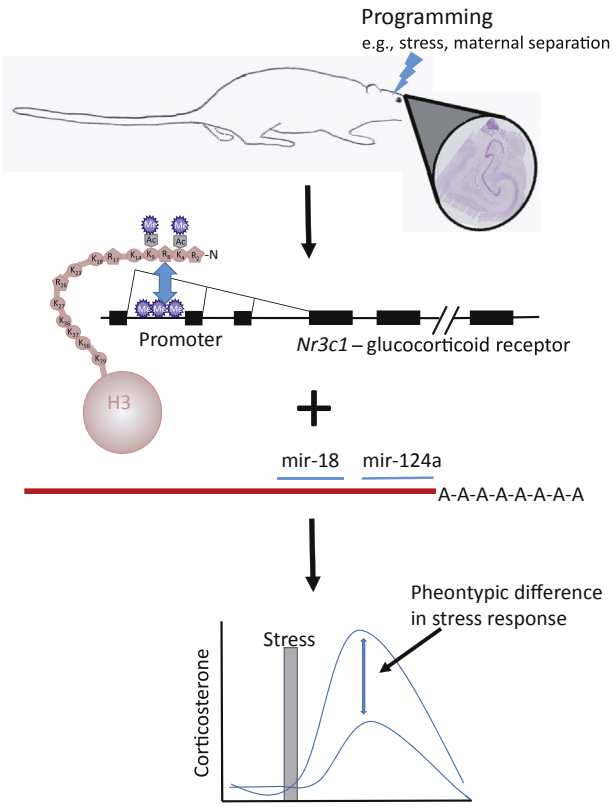

Figure 4 Schematic representation of the epigenetic regulation of the stress response and the glucocorticoid receptor. After exposure to a programming event such as stress or maternal separation, for example, (Weaver et al., 2004), the glucocorticoid receptor (*Nr3C1*) gene is programmed in the hippocampus, particularly at the histone H3 (Weaver et al., 2004; Weaver et al., 2007), the 1_7 promoter (Weaver et al., 2004), and by miRNAs-18 and 124a (Vreugdenhil et al., 2009). The functional effect of this programming is the phenotypic response to subsequent stress.

In depression, a mental disorder that has been linked to both stress and stress susceptibility, the activity of chromatin modifying enzymes was studied. The genome-wide levels of histone PTMs such as H3Ac, H4Ac, H3K4me3, H3K9me3, and H3K27me3 levels were clearly associated with stress responsiveness in diverse regions throughout the brain (Vialou et al., 2013). Conversely, systemic administration of HDAC inhibitors showed significant antidepressant-like effects in rodent models, but had no effect on unstressed mice (Gundersen and Blendy, 2009; Tsankova et al., 2007).

Current studies suggest highly complex mechanisms that are not limited to the already highly variable chromatin states. There is also evidence that the HPA axis responsiveness and hippocampal GR levels are modulated by miRNAs (Uchida et al., 2008; Vreugdenhil et al., 2009), underlying the stress hypersensitivity of Fischer 344 rats compared with the related hyposensitive Lewis strain (Lytle et al., 2007).

Dysregulation of the HPA axis and the stress response is also often observed in mental disorder. A role has also been proposed for DNA methylation in major depressive disorder (MDD; e.g., GR) (Alt et al., 2010; McGowan et al., 2009) and other mental disorders (Szyf, 2011). Several antidepressants alter global and gene-specific methylation levels in the rodent brain, as part of their therapeutic mechanisms (Milutinovic et al., 2007). Methylation of the serotonin transporter (5-HTT) promoter exacerbates behavioral stress reactivity in monkeys, and increases the risk of mood disorders (Kinnally et al., 2011).

Epigenetics and Behavior

Over the past 20 years there have been many reports of programming of biological systems and behavioral responses. Most of these were associated with defined susceptibility windows. Both in human and animal models maternal stress during gestation increases both anxiety- and depression-like behavior which persists during adulthood. Recent data suggest that epigenetic regulation of gene expression is involved (Sun et al., 2013).

A model in which the epigenetic regulation of behavior has been clearly shown is the Prairie vole (Microtus ochrogaster). These are the only socially monogamous rodents that form male–female bonds. Partner preferences are formed following male–female cohabitation and mating. Recently, Wang et al. facilitated this behavior in females, in absence of mating by inhibiting HDACs. This treatment increased histone acetylation in promoters of the oxytocin receptor (OTR and oxtr) and vasopressin V1a receptor (V1aR, avpr1a) in the nucleus accumbens of the brain producing epigenetic regulation patterns similar to those established by natural mating-induced partner preference. This was interpreted as direct evidence for epigenetic regulation of pair bonding.

Human data are somewhat limited. However, individuals with suicidal ideation or suicide completers have aberrantly downregulated *Bdnf* gene expression and protein levels. Here

Bdnf transcription was shown to be regulated by DNA methylation. The *Bdnf* promoter was significantly more methylated in these patients, linking epigenetic alterations in the brain with suicidal behavior (Keller et al., 2010; Kang et al., 2013).

Epigenetic and Exercise – DNA Methylation

Exercising is widely recognized today to prevent chronic diseases, to decrease the risk of mortality and to increase longevity. The underlying mechanisms are poorly understood but recent epigenetic studies may provide some clues. In 2013, Rönn et al. demonstrated that a 6-month exercise regimen changed the DNA methylation profiles in adipose tissues in healthy men with otherwise low levels of physical activity. Indeed, increases in DNA methylation levels at individual CpGs ranged from 0.2 to 10.9%, affecting adipocyte metabolism and fat storage (Rönn et al., 2013). Increased DNA methylation was shown in particular in genes implicated in obesity and type 2 diabetes. This suggests that common epigenetically controlled loci are involved in different metabolic diseases.

Exercising also showed promise in preventing drug addiction relapses in rodents. For example cocaine seeking was reduced by wheel running in abstinent addicted rats (Peterson et al., 2014). Although this was not tested the authors suspected an epigenetic mechanism, regulating *Bdnf* gene in the PFC region of the brain. A similar study by Sanchez et al. found that exercise may also reduce the relapse vulnerability for nicotine addiction in adolescent rats (Sanchez et al., 2013).

Are There Particular Sensitivity Windows?

Epigenetics plays an important role as an interface between genes and the environment. A wide range of environmental stressors modify the epigenome with dramatic long-term effects on the exposed individual and sometimes even on following generations. Type and time of exposure determine the outcome.

In utero

Epidemiological observations and animal studies demonstrated the importance of epigenetic changes during the developmental period. Environmental conditions, including maternal diet, uterine blood flow, and exposure to chemicals, can have severe consequences on offspring gene expression. Metals in the environment can inhibit specific methylation or demethylation enzymes either by direct binding or indirectly through reactive oxygen species actions. For instance Cadmium crosses the placental barrier and leaves distinct DNA methylation patterns in the newborn DNA. Analysis of DNA methylation associated with Cadmium exposure showed that methylation patterns in newborn white blood cells were disrupted in transcriptional regulation and apoptosis-related genes. The CpGs that were methylated were, importantly, within previously identified transcription factor-binding sites. Kundakovic et al. demonstrated that in utero exposure to environmentally relevant doses of Bisphenol A, a product extensively used in the production of plastics, induces lasting epigenetic disruptions in juvenile mice. The methylation levels of estrogen receptors and Dnmt (1 and 3A) were significantly altered in the cortex of males and the hypothalamus of females. Those epigenetic changes were associated with gender-specific social and anxiety-like behavior, especially sexually dimorphic phenotypes. Also the diet plays an important role in the epigenetics of mother–baby pairs. Epigenetically active enzymes depend on levels of cofactors and metabolites present in the intra- and extra-cellular environment. Some nutrients such as folate, choline, methionine, and vitamin B (folic acid metabolism) are essentials for the establishment and maintenance of DNA methylations both during development and throughout life. Deficiencies resulting from changes in the diet can have severe consequences on gene regulation. In rodents, in utero folate deficiency resulted in embryonic and postnatal growth retardation with long-term brain defects. Kerek et al. showed that early methyl donor deficiency (MDD) in rats may induce persistent brain defects by reducing signal transducer and activator of transcription 3 (Stat3) signaling (Kerek et al., 2013). MDD was also associated with an increased expression of miR-124, affecting Stat3 mRNA expression and activation levels. The authors concluded that impaired Stat3 levels in the brain, caused by miR-124 upregulation, reduce cell proliferation and increase apoptosis that may account for a definitive atrophy of specific brain regions.

The in-utero period is important for the phenotypical programming of the adult. As outlined above, there is growing evidence that even small or transient environmental changes such as maternal infections or stress during pregnancy are major risk factors for later neuropsychiatric disorders such as depression and anxiety (Enayati et al., 2012), schizophrenia (Brown and Derkits, 2010), and autism (Kinney et al., 2008). In-utero stress has recently been found to activate epigenetic mechanisms that could be linked to schizophrenia (Matrisciano et al., 2013). It was proposed that decreased GABAergic gene expression (e.g., glutamic acid decarboxylase 67 and reelin) concomitant with the overexpression of *DNMT* in GABAergic neurons can contribute to the neuropathology of psychotic disorders (Guidotti et al., 2011), and both were induced by prenatal restraint stress (Matrisciano et al., 2013).

Early Life

The principal rodent model of early life care is licking–grooming and arched-back nursing (LG-ABN). Low levels of LG-ABN were demonstrated to directly translate into increased DNA methylation of one particular CpG within the GR promoter, with profound and long lasting effects on the recipients stress response (Weaver, 2007). Enhanced methylation led to reduced GR expression, and cross-fostering of pups from high to low caring mothers demonstrated that the effect was due to the maternal behavior (Weaver et al., 2004, 2007). Maternal separation however did not produce similar differences in DNA methylation of the GR promoter (Daniels et al., 2009).

Translating such rodent data to humans is difficult. Nevertheless, early life adversity may provide some insight into the role of early life in epigenetic programming. For instance the care provided by orphanages has received significant attention, and even in the best institutions the care

cannot match the normal human parent–infant relation. Even in the best institutions employed caregivers working in shifts in charge of 20 children or more (Gunnar et al., 2009) with clear consequences for care quantity and quality (Smyke et al., 2007). A continually changing caregiver is the strongest stressor currently known for human (and nonhuman primate) infants (Johnson et al., 2002). It has also been shown that the changing care provider, in what would otherwise be considered a rich environment, significantly alters both social and emotional behavior for many years (Hodges and Tizard, 1989).

The psychological development of such individuals has been intensely followed in the early years postinstitutionalization including social–emotional behavior, attachment, and relationships (Smyke et al., 2002), peer interactions, and emotion regulation (Hodges and Tizard, 1989) as reviewed by Gunnar et al. (2000). More recently, there have been a large number of correlational neuroimaging studies performed on postinstitutionalized children, including for example PET, diffusion tensor imaging, and electroencephalogram. All of these studies suggest that exposure to institutions in early life induces neuronal mechanisms that underpin behavioral differences later in life. These have recently been extended by physiological studies such as the Minnesota International Adoption Project on postinstitutionalized children that have identified risk factors for cardiovascular diseases (Gunnar et al., 2009), disturbances in the HPA axis which increased with the duration of institutionalization, and epigenetic methylation.

Using a representative series of 27 000 of the 30 million CpGs covering 14 000 genes throughout the human genome, institutionalized children did not have a methylation pattern that was clearly distinct from the parent-raised control children. However, there was a series of 838 genes that showed small, but statistically significant, intergroup differences in methylation levels, that were predominantly (89%) increased after institutionalization (Naumova et al., 2011).

McGowan et al. identified a number of CpG dinucleotides in the GR promoter that were differentially methylated in suicide completers that had been subject to such abuse compared with either control nonabused nonsuicide victims, or to nonabused suicide completers (McGowan et al., 2009).

Adolescence

Development, particularly of certain brain regions, continues until early adulthood, meaning that epigenetic modifications, or programming are not restricted to the perinatal period and childhood (Fraga et al., 2005). For example, the HPA axis plasticity, together with persisting neuroendocrine changes, is maintained during adolescence and into adulthood with consequent phenotypic changes in the stress response (Schmidt et al., 2010). This plasticity is thought to be due to increased DNA methylation, associated with significantly reduced MR and GR levels in certain hippocampus regions (Schmidt et al., 2010), directly affecting the long-term stress response.

There is currently no human data available confirming the underlying epigenetic mechanisms, however studies addressing this such as TRAILS are underway (Bosch et al., 2012). Rodent data show that methylation of the *Gr* in HPA axis tissues, such as the pituitary and adrenal glands, is significantly increased by chronic stress during adolescence (Witzmann et al., 2012). Similarly Gr methylation in the maternal side of the placenta was sensitive to the external environment (Bromer et al., 2013). As such, it would be reasonable to assume that epigenetic mechanisms such as altered DNA methylation underlie the observations of the adolescent environment inducing long-term phenotypic alterations.

Conclusions

Epigenetics promises to improve our understanding of the mechanisms underlying our continued adaption to our environment throughout life. Results, both from animal models and humans, start to reveal specific epigenetic markers and modifications that result in stable phenotype-specific gene expression patterns and behavioral phenotypes. Mechanisms have been implicated in our response to stress or other environmentally challenges at every epigenetic level (chromatin, DNA methylation, and microRNAs).

The three principal epigenetic mechanisms are regulated throughout life with higher epigenetic susceptibility during early life or adolescence. So far tissue-specific epigenetic regulation has been studied almost exclusively on single, permanently regulated genes such as GR or BDNF. First genome-wide studies, however, strongly suggest that transcriptional and translational regulation of many other genes is sensitive to epigenetic regulation. Systematic genome-wide epigenetic studies using cutting-edge next generation sequencing will greatly accelerate our understanding of phenotypic adaptation to our environment.

This will also allow to further exploring epigenetic mechanisms as targets for drug treatment, beyond HDAC inhibitors as antidepressants. However, since HDACs and most other chromatin-modifying enzymes are ubiquitously expressed specific interventions will represent a considerable challenge.

See also: Behavioral Medicine; Culture as a Determinant of Mental Health; Environmental Stress and Health; Personality and Health; Personalized Medicine; Psychoneuroimmunology; Social Epigenetics; Stress, Coping and Health.

Bibliography

Alt, S.R., Turner, J.D., Klok, M.D., Meijer, O.C., Lakke, E.A., Derijk, R.H., Muller, C.P., 2010. Differential expression of glucocorticoid receptor transcripts in major depressive disorder is not epigenetically programmed. Psychoneuroendocrinology 35, 544–556.

Baccarelli, A., Tarantini, L., Wright, R.O., Bollati, V., Litonjua, A.A., Zanobetti, A., Sparrow, D., Vokonas, P., Schwartz, J., 2010. Repetitive element DNA methylation and circulating endothelial and inflammation markers in the VA normative aging study. Epigenetics 5.

Bannister, A.J., Kouzarides, T., 2011. Regulation of chromatin by histone modifications. Cell Research 21, 381–395.

Blaze, J., Roth, T.L., 2013. Epigenetic mechanisms in learning and memory. Wiley Interdisciplinary Reviews: Cognitive Science 4, 105–115.

Bollati, V., Schwartz, J., Wright, R., Litonjua, A., Tarantini, L., Suh, H., Sparrow, D., Vokonas, P., Baccarelli, A., 2009. Decline in genomic DNA methylation through aging in a cohort of elderly subjects. Mechanisms of Ageing and Development 130, 234–239.

Bosch, N.M., Riese, H., Reijneveld, S.A., Bakker, M.P., Verhulst, F.C., Ormel, J., Oldehinkel, A.J., 2012. Timing matters: long term effects of adversities from prenatal period up to adolescence on adolescents' cortisol stress response. The TRAILS study. Psychoneuroendocrinology 37, 1439–1447.

Bromer, C., Marsit, C.J., Armstrong, D.A., Padbury, J.F., Lester, B., 2013. Genetic and epigenetic variation of the glucocorticoid receptor (NR3C1) in placenta and infant neurobehavior. Developmental Psychobiology 55, 673–683.

Brown, A.S., Derkits, E.J., 2010. Prenatal infection and schizophrenia: a review of epidemiologic and translational studies. American Journal of Psychiatry 167, 261–280.

Cao-Lei, L., Suwansirikul, S., Jutavijittum, P., Meriaux, S.B., Turner, J.D., Muller, C.P., 2013. Glucocorticoid receptor gene expression and promoter CpG modifications throughout the human brain. Journal of Psychiatric Research 47, 1597–1607.

Daniels, W.M., Fairbairn, L.R., Van Tilburg, G., Mcevoy, C.R., Zigmond, M.J., Russell, V.A., Stein, D.J., 2009. Maternal separation alters nerve growth factor and corticosterone levels but not the DNA methylation status of the exon 1(7) glucocorticoid receptor promoter region. Metabolic Brain Disease 24, 615–627.

Devaux, Y., Vausort, M., Goretti, E., Nazarov, P.V., Azuaje, F., Gilson, G., Corsten, M.F., Schroen, B., Lair, M.L., Heymans, S., Wagner, D.R., 2012. Use of circulating microRNAs to diagnose acute myocardial infarction. Clinical Chemistry 58, 559–567.

Drong, A.W., Lindgren, C.M., Mccarthy, M.I., 2012. The genetic and epigenetic basis of type 2 diabetes and obesity. Clinical Pharmacology & Therapeutics 92, 707–715.

Edwards, A.O., Ritter 3rd, R., Abel, K.J., Manning, A., Panhuysen, C., Farrer, L.A., 2005. Complement factor H polymorphism and age-related macular degeneration. Science 308, 421–424.

Enayati, M., Solati, J., Hosseini, M.H., Shahi, H.R., Saki, G., Salari, A.A., 2012. Maternal infection during late pregnancy increases anxiety- and depression-like behaviors with increasing age in male offspring. Brain Research Bulletin 87, 295–302.

Eriksson, J.G., Kajantie, E., Thornburg, K.L., Osmond, C., Barker, D.J., 2011. Mother's body size and placental size predict coronary heart disease in men. European Heart Journal 32, 2297–2303.

Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suner, D., Cigudosa, J.C., Urioste, M., Benitez, J., Boix-Chornet, M., Sanchez-Aguilera, A., Ling, C., Carlsson, E., Poulsen, P., Vaag, A., Stephan, Z., Spector, T.D., Wu, Y.Z., Plass, C., Esteller, M., 2005. Epigenetic differences arise during the lifetime of monozygotic twins. Proceedings of the National Academy of Sciences of the United States of America 102, 10604–10609.

Guidotti, A., Auta, J., Chen, Y., Davis, J.M., Dong, E., Gavin, D.P., Grayson, D.R., Matrisciano, F., Pinna, G., Satta, R., Sharma, R.P., Tremolizzo, L., Tueting, P., 2011. Epigenetic GABAergic targets in schizophrenia and bipolar disorder. Neuropharmacology 60, 1007–1016.

Gundersen, B.B., Blendy, J.A., 2009. Effects of the histone deacetylase inhibitor sodium butyrate in models of depression and anxiety. Neuropharmacology 57, 67–74.

Gunnar, M.R., Bruce, J., Grotevant, H.D., 2000. International adoption of institutionally reared children: research and policy. Development and Psychopathology 12, 677–693.

Gunnar, M.R., Frenn, K., Wewerka, S.S., Van Ryzin, M.J., 2009. Moderate versus severe early life stress: associations with stress reactivity and regulation in 10-12-year-old children. Psychoneuroendocrinology 34, 62–75.

Haramati, S., Navon, I., Issler, O., Ezra-Nevo, G., Gil, S., Zwang, R., Hornstein, E., Chen, A., 2011. MicroRNA as repressors of stress-induced anxiety: the case of amygdalar miR-34. Journal of Neuroscience 31, 14191–14203.

Hattori, M., 2005. Finishing the euchromatic sequence of the human genome. Tanpakushitsu Kakusan Koso. Protein Nucleic Acid, Enzyme 50, 162–168.

He, Y.F., Li, B.Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., Sun, Y., Li, X., Dai, Q., Song, C.X., Zhang, K., He, C., Xu, G.L., 2011. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science 333, 1303–1307.

Hodges, J., Tizard, B., 1989. Social and family relationships of ex-institutional adolescents. Journal of Child Psychology and Psychiatry 30, 77–97.

Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C., Zhang, Y., 2011. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science 333, 1300–1303.

Johnson, J.D., O'connor, K.A., Deak, T., Spencer, R.L., Watkins, L.R., Maier, S.F., 2002. Prior stressor exposure primes the HPA axis. Psychoneuroendocrinology 27, 353–365.

Kang, H.-J., Kim, J.-M., Lee, J.-Y., Kim, S.-Y., Bae, K.-Y., Kim, S.-W., Shin, I.-S., Kim, H.-R., Shin, M.-G., Yoon, J.-S., 2013. BDNF promoter methylation and suicidal behavior in depressive patients. Journal of affective disorders 151, 679–685.

Keller, S., Sarchiapone, M., Zarrilli, F., Videtic, A., Ferraro, A., Carli, V., Sacchetti, S., Lembo, F., Angiolillo, A., Jovanovic, N., Pisanti, F., Tomaiuolo, R., Monticelli, A., Balazic, J., Roy, A., Marusic, A., Cocozza, S., Fusco, A., Bruni, C.B., Castaldo, G., Chiariotti, L., 2010. Increased BDNF promoter methylation in the Wernicke area of suicide subjects. Archives of General Psychiatry 67, 258–267.

Kerek, R., Geoffroy, A., Bison, A., Martin, N., Akchiche, N., Pourie, G., Helle, D., Gueant, J.L., Bossenmeyer-Pourie, C., Daval, J.L., 2013. Early methyl donor deficiency may induce persistent brain defects by reducing Stat3 signaling targeted by miR-124. Cell Death and Disease 4, e755.

Kim, V.N., Han, J., Siomi, M.C., 2009. Biogenesis of small RNAs in animals. Nature Reviews Molecular Cell Biology 10, 126–139.

Kinnally, E.L., Feinberg, C., Kim, D., Ferguson, K., Leibel, R., Coplan, J.D., John Mann, J., 2011. DNA methylation as a risk factor in the effects of early life stress. Brain Behavior and Immunity 25, 1548–1553.

Kinney, D.K., Munir, K.M., Crowley, D.J., Miller, A.M., 2008. Prenatal stress and risk for autism. Neuroscience & Biobehavioral Reviews 32, 1519–1532.

Klose, R.J., Bird, A.P., 2006. Genomic DNA methylation: the mark and its mediators. Trends in Biochemical Sciences 31, 89–97.

Kubota, T., Miyake, K., Hirasawa, T., 2012. Epigenetic understanding of gene-environment interactions in psychiatric disorders: a new concept of clinical genetics. Clinical Epigenetics 4.

Lytle, J.R., Yario, T.A., Steitz, J.A., 2007. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. Proceedings of the National Academy of Sciences of the United States of America 104, 9667–9672.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., Mccarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F.C., Mccarroll, S.A., Visscher, P.M., 2009. Finding the missing heritability of complex diseases. Nature 461, 747–753.

Matrisciano, F., Tueting, P., Dalal, I., Kadriu, B., Grayson, D.R., Davis, J.M., Nicoletti, F., Guidotti, A., 2013. Epigenetic modifications of GABAergic interneurons are associated with the schizophrenia-like phenotype induced by prenatal stress in mice. Neuropharmacology 68, 184–194.

McEwen, B.S., 2008. Central effects of stress hormones in health and disease: understanding the protective and damaging effects of stress and stress mediators. European Journal of Pharmacology 583, 174–185.

McEwen, B.S., Eiland, L., Hunter, R.G., Miller, M.M., 2012. Stress and anxiety: structural plasticity and epigenetic regulation as a consequence of stress. Neuropharmacology 62, 3–12.

McGowan, P.O., Sasaki, A., D'alessio, A.C., Dymov, S., Labonte, B., Szyf, M., Turecki, G., Meaney, M.J., 2009. Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse. Nature Neuroscience 12, 342–348.

Meaney, M.J., 2007. Environmental programming of phenotypic diversity in female reproductive strategies. Advances in Genetics 59, 173–215.

Milutinovic, S., D'alessio, A.C., Detich, N., Szyf, M., 2007. Valproate induces widespread epigenetic reprogramming which involves demethylation of specific genes. Carcinogenesis 28, 560–571.

Naumova, O.Y., Lee, M., Koposov, R., Szyf, M., Dozier, M., Grigorenko, E.L., 2011. Differential patterns of whole-genome DNA methylation in institutionalized children and children raised by their biological parents. Development and Psychopathology 1–13.

O'Connor, R.M., Dinan, T.G., Cryan, J.F., 2012. Little things on which happiness depends: microRNAs as novel therapeutic targets for the treatment of anxiety and depression. Molecular Psychiatry 17 (4), 359–376.

Painter, R.C., Osmond, C., Gluckman, P., Hanson, M., Phillips, D.I., Roseboom, T.J., 2008. Transgenerational effects of prenatal exposure to the Dutch famine on neonatal adiposity and health in later life. British Journal of Obstetrics and Gynaecology 115, 1243–1249.

Peterson, A.B., Abel, J.M., Lynch, W.J., 2014. Dose-dependent effects of wheel running on cocaine-seeking and prefrontal cortex Bdnf exon IV expression in rats. Psychopharmacology 231 (7), 1305–1314.

Prokhortchouk, A., Hendrich, B., Jorgensen, H., Ruzov, A., Wilm, M., Georgiev, G., Bird, A., Prokhortchouk, E., 2001. The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. Genes & Development 15, 1613–1618.

Reilly, M.P., Li, M., He, J., Ferguson, J.F., Stylianou, I.M., Mehta, N.N., Burnett, M.S., Devaney, J.M., Knouff, C.W., Thompson, J.R., Horne, B.D., Stewart, A.F., Assimes, T.L., Wild, P.S., Allayee, H., Nitschke, P.L., Patel, R.S., Myocardial Infarction Genetics, C., Wellcome Trust Case Control, C., Martinelli, N., Girelli, D., Quyyumi, A.A., Anderson, J.L., Erdmann, J., Hall, A.S., Schunkert, H., Quertermous, T., Blankenberg, S., Hazen, S.L., Roberts, R., Kathiresan, S., Samani, N.J., Epstein, S.E., Rader, D.J., 2011. Identification of ADAMTS7 as a novel locus for coronary atherosclerosis and association of ABO with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies. Lancet 377, 383–392.

Riester, A., Issler, O., Spyroglou, A., Rodrig, S.H., Chen, A., Beuschlein, F., 2012. ACTH-dependent regulation of microRNA as endogenous modulators of glucocorticoid receptor expression in the adrenal gland. Endocrinology 153, 212–222.

Rinaldi, A., Vincenti, S., De Vito, F., Bozzoni, I., Oliverio, A., Presutti, C., Fragapane, P., Mele, A., 2010. Stress induces region specific alterations in microRNAs expression in mice. Behavioural Brain Research 208, 265–269.

Rodgers, A.B., Morgan, C.P., Bronson, S.L., Revello, S., Bale, T.L., 2013. Paternal stress exposure alters sperm microRNA content and reprograms offspring HPA stress axis regulation. Journal of Neuroscience 33, 9003–9012.

Rönn, T., Volkov, P., Davegårdh, C., Dayeh, T., Hall, E., Olsson, A.H., Nilsson, E., Tornberg, A., Dekker Nitert, M., Eriksson, K.-F., Jones, H.A., Groop, L., Ling, C., 2013. A six months exercise intervention influences the genome-wide DNA methylation pattern in human adipose tissue. PLOS Genetics 9, e1003572.

Roth, T.L., Lubin, F.D., Funk, A.J., Sweatt, J.D., 2009. Lasting epigenetic influence of early-life adversity on the BDNF gene. Biological Psychiatry 65, 760–769.

Rozanov, V., 2012. Epigenetics: stress and behavior. Neurophysiology 1–19.

Sanchez, V., Moore, C.F., Brunzell, D.H., Lynch, W.J., 2013. Effect of wheel-running during abstinence on subsequent nicotine-seeking in rats. Psychopharmacology 227, 403–411.

Saraiya, A.A., Li, W., Wang, C.C., 2013. Transition of a microRNA from repressing to activating translation depending on the extent of base pairing with the target. PLoS One 8, e55672.

Schmidt, M.V., Scharf, S.H., Liebl, C., Harbich, D., Mayer, B., Holsboer, F., Muller, M.B., 2010. A novel chronic social stress paradigm in female mice. Hormones and Behavior 57, 415–420.

Smyke, A.T., Dumitrescu, A., Zeanah, C.H., 2002. Attachment disturbances in young children. I: the continuum of caretaking casualty. Journal of the American Academy of Child & Adolescent Psychiatry 41, 972–982.

Smyke, A.T., Koga, S.F., Johnson, D.E., Fox, N.A., Marshall, P.J., Nelson, C.A., Zeanah, C.H., Group, B.C., 2007. The caregiving context in institution-reared and family-reared infants and toddlers in Romania. Journal of Child Psychology and Psychiatry 48, 210–218.

Sun, H., Guan, L., Zhu, Z., Li, H., 2013. Reduced levels of NR1 and NR2A with depression-like behavior in different brain regions in prenatally stressed juvenile offspring. PLoS One 8, e81775.

Szyf, M., 2009. The early life environment and the epigenome. Biochimica et Biophysica Acta 1790, 878–885.

Szyf, M., 2011. The early life social environment and DNA methylation: DNA methylation mediating the long-term impact of social environments early in life. Epigenetics 6, 971–978.

Talikka, M., Sierro, N., Ivanov, N.V., Chaudhary, N., Peck, M.J., Hoeng, J., Coggins, C.R., Peitsch, M.C., 2012. Genomic impact of cigarette smoke, with application to three smoking-related diseases. Critical Reviews in Toxicology 42, 877–889.

Tsankova, N., Renthal, W., Kumar, A., Nestler, E.J., 2007. Epigenetic regulation in psychiatric disorders. Nature Reviews Neuroscience 8, 355–367.

Turner, J.D., Pelascini, L.P., Macedo, J.A., Muller, C.P., 2008. Highly individual methylation patterns of alternative glucocorticoid receptor promoters suggest individualized epigenetic regulatory mechanisms. Nucleic Acids Research 36, 7207–7218.

Uchida, S., Nishida, A., Hara, K., Kamemoto, T., Suetsugi, M., Fujimoto, M., Watanuki, T., Wakabayashi, Y., Otsuki, K., Mcewen, B.S., Watanabe, Y., 2008. Characterization of the vulnerability to repeated stress in Fischer 344 rats: possible involvement of microRNA-mediated down-regulation of the glucocorticoid receptor. European Journal of Neuroscience 27, 2250–2261.

Van Setten, J., Isgum, I., Smolonska, J., Ripke, S., De Jong, P.A., Oudkerk, M., De Koning, H., Lammers, J.W., Zanen, P., Groen, H.J., Marike Boezen, H., Postma, D.S., Wijmenga, C., Viergever, M.A., Mali, W.P., De Bakker, P.I., 2013. Genome-wide association study of coronary and aortic calcification implicates risk loci for coronary artery disease and myocardial infarction. Atherosclerosis 228, 400–405.

Vialou, V., Feng, J., Robison, A.J., Nestler, E.J., 2013. Epigenetic mechanisms of depression and antidepressant action. Annual Review of Pharmacology and Toxicology 53, 59–87.

Vreugdenhil, E., Verissimo, C.S., Mariman, R., Kamphorst, J.T., Barbosa, J.S., Zweers, T., Champagne, D.L., Schouten, T., Meijer, O.C., De Kloet, E.R., Fitzsimons, C.P., 2009. MicroRNA 18 and 124a down-regulate the glucocorticoid receptor: implications for glucocorticoid responsiveness in the brain. Endocrinology 150, 2220–2228.

Weaver, I.C., 2007. Epigenetic programming by maternal behavior and pharmacological intervention. Nature versus nurture: let's call the whole thing off. Epigenetics 2, 22–28.

Weaver, I.C., Cervoni, N., Champagne, F.A., D'alessio, A.C., Sharma, S., Seckl, J.R., Dymov, S., Szyf, M., Meaney, M.J., 2004. Epigenetic programming by maternal behavior. Nature Neuroscience 7, 847–854.

Weaver, I.C., D'alessio, A.C., Brown, S.E., Hellstrom, I.C., Dymov, S., Sharma, S., Szyf, M., Meaney, M.J., 2007. The transcription factor nerve growth factor-inducible protein a mediates epigenetic programming: altering epigenetic marks by immediate-early genes. Journal of Neuroscience 27, 1756–1768.

Witzmann, S.R., Turner, J.D., Meriaux, S.B., Meijer, O.C., mu ller, C.P., 2012. Epigenetic regulation of the glucocorticoid receptor promoter 1(7) in adult rats. Epigenetics 7, 1290–1301.

Xu, X.F., Cheng, F., Du, L.Z., 2011. Epigenetic regulation of pulmonary arterial hypertension. Hypertension research 34, 981–986.

Yang, X.-J., Seto, E., 2008. Lysine acetylation: codified crosstalk with other post-translational modifications. Molecular Cell 31, 449–461.

Yehuda, R., Teicher, M.H., Seckl, J.R., Grossman, R.A., Morris, A., Bierer, L.M., 2007. Parental posttraumatic stress disorder as a vulnerability factor for low cortisol trait in offspring of holocaust survivors. Archives of General Psychiatry 64, 1040–1048.

Epigenetics of Brain Plasticity and Behavior

Richelle Mychasiuk, Faculty of Medicine, Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Epigenetics refers to changes in gene expression/organism phenotype that occur without altering the DNA sequence. A delicate balance exists between the epigenome (the representative genome that includes all inhibitory and activating, covalent and chemical markers), and the environment that shapes development and long-term plasticity of the human brain, and thus behavioral output. Although epigenetic processes are essential for development and cellular differentiation, they also provide a conduit by which environmental influences can permanently alter neuronal structure and function. From an evolutionary perspective, this provides the organism with a mechanism for instant adaptation. Research is beginning to uncover the means by which epigenetic processes mediate brain development and plasticity and may provide us with the resources needed to restore dysfunctional learning and enhance neuroplasticity in the adult brain.

Introduction

Although cells of a multicellular organism are genetically homogeneous, they differ structurally and functionally from each other based upon the differential expression of genes. Many changes in gene expression arise during development but are steadily maintained throughout life. They are often heritable, transmitted from one generation to the next. Epigenetics refers to changes in gene expression or cellular phenotype that occur without changes to the DNA sequence. Epigenetic processes are essential for development and differentiation; they also transpire in mature cells in response to environmental influences. These epigenetic modifications provide organisms with a means of instant adaptation, whereby gene activity states can be altered by environmental factors and then propagated from one generation of cells to the next.

Brain development and plasticity are highly integrated processes that rely heavily on similar mechanisms of epigenetic programming. Although, often thought of as distinct events both processes use epigenetic factors to alter functioning of the brain. For example, synaptic pruning, a primary mechanism used by the brain during development, is also a key modulator of neuroplasticity. Although there are critical windows for brain development in which epigenetic programs are working at a heightened level, neuroplasticity is an ongoing process that requires steady modification to signaling cascades, gene expression, and protein production. The main purpose of this article is to describe the influences of epigenetics on brain development and plasticity. While this article will attempt to distinguish between the two processes, there is considerable overlap, and it may be difficult to differentiate where one mechanism begins and the other ends.

Neurodevelopment refers to the processes that generate, shape, and reshape the central nervous system (specifically the brain) from embryogenesis to the final years of life. Although brain development typically follows a predictable pattern that is generated by a genetic blueprint, experiences during this sensitive period have the ability to alter epigenetic patterns of many genes. Characterization of methylation patterns in the normal brain indicates that epigenetic alterations are dynamically regulated and region dependent. Temporal analysis demonstrates significant epigenetic activity in the prenatal period, followed by a marked slow decline after birth that continues with aging. This temporal arrangement is consistent with genome-wide demethylation at the point of conception followed by rapid re-methylation of the genome that is maintained throughout life. Several studies show that methylation of the brain's epigenome increases with age, with regional differences that coincide with maturational rates of the brain (see Ladd-Acosta et al., 2007; Siegmund et al., 2007). Importantly, these studies not only demonstrate the increased opportunity for epigenetic modification in the early years, but also emphasize the continuous nature of epigenetic programming and neuroplasticity.

Epigenetics permeates all aspects of development, especially in the brain. The mammalian brain is unique in this respect as it develops intimately with the environment, both prenatally and postnatally in the context of all experiences. These experience-induced changes produce stable gene-expression modifications that alter gene transcription throughout life and stimulate specific behavioral and psychological outcomes. The adaptive nature of the brain is produced by the dynamic relationship that exists between the epigenome, environment, and behavioral outcomes.

Similarly, a fundamental property of the brain is its propensity to change. Neuroplasticity refers to changes in neural pathways and synapses, which occur in response to changes in behavior, environment, and other neural processes. Neuroplasticity occurs at a variety of levels that range from behavioral (e.g., learning) to structural (e.g., cortical reorganization following injury). Common belief has changed from the previous position that the brain is a physiologically static organ, to the consensus that the brain is plastic only during critical windows, and finally to the current conviction that the brain remains plastic throughout life, but that the permeability to change waxes and wanes with maturational state and plasticity reserves. Although changes in the brain/neuroplasticity can be transient (i.e., fluctuations in neurotransmitters) or persistent (i.e., alterations to dendritic morphology or synaptic pruning), each modification requires epigenetic activity. For example, dendritic arborization can only occur following alterations in protein production that result from changes in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.55058-2

gene expression. Similarly, the neurogenesis associated with improved hippocampal functioning is dependent upon upregulation of specific growth factors, an epigenetic modification.

Mechanisms of Epigenetic Programming

The DNA sequence is only a foundation for understanding how the genetic program is read. A second layer of interpretation is required to fully appreciate the possibilities arising from the original sequence. There are 3 main mechanisms of epigenetic regulation used to transmit these genetic adaptations: RNA regulation, posttranslational modification to histone proteins, and the most commonly studied DNA methylation (see Figure 1).

RNA Regulation

Evidence indicates that RNAs direct many aspects of chromatin regulation. A sophisticated RNA regulatory network involves RNA editing and non-protein-coding RNAs that orchestrate an aspect of epigenetic programming. Non-coding RNAs (ncRNAs) dominate the transcriptional output of mammals and regulate many levels of gene expression during development. RNA-binding co-regulators are recruited to transcriptional units by gene-specific ncRNAs to modulate chromatin structure. ncRNAs may recruit transcriptional repressors or transcriptional activators functioning in both gene activation and gene silencing. microRNAs are posttranscriptional regulators that bind to the complementary sequence of mRNA in the 3′ untranslated region (UTR) usually resulting in gene silencing. RNA regulation is a continuous process involved in long-term memory formation, context-specific behavioral responses, and cellular development and differentiation (Mehler and Mattick, 2007).

Posttranslational Modification to Histone Proteins

Epigenetic programming via posttranslational modification of histone proteins occurs by altering the accessibility of the transcriptional machinery to the DNA. Posttranslational modification to histone proteins includes processes such as

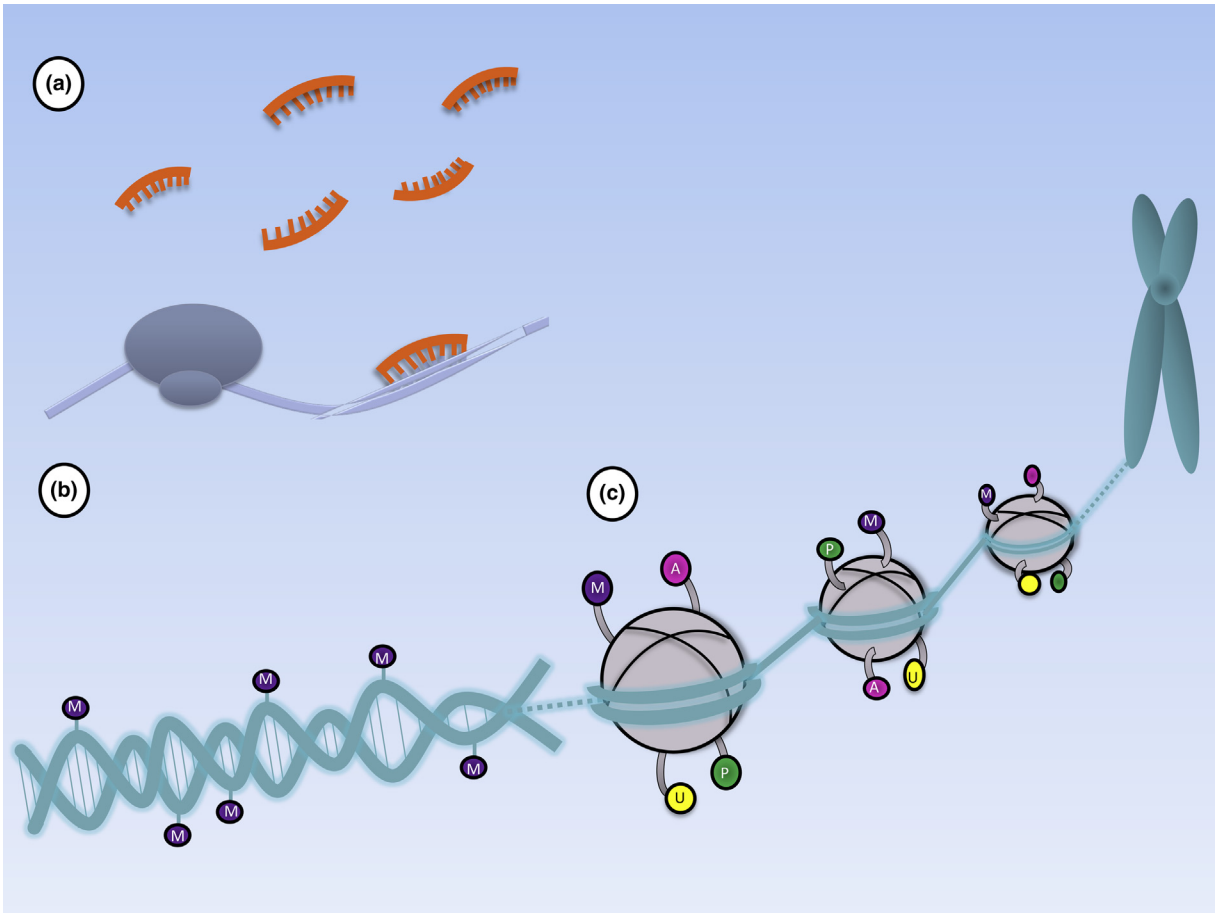

Figure 1 The cartoon demonstrates the three main levels of control utilized by the cell to regulate expression of specific genes and maintain epigenetic patterning. (a) Gene expression is regulated by microRNAs that bind to transcribed sequences of mRNA prior to protein translation. (b) DNA methylation is the most stable form of gene silencing, whereby methyl groups (M) are bound to cytosine residues in a gene's promoter region. (c) Histone modifications, including methylation (M), acetylation (A), ubiquination (U), and phosphorylation (P), are commonly used mechanisms that alter accessibility of transcriptional machinery to DNA sequences.

methylation, acetylation, and ubiquitination. Histone acetylation is the most commonly studied posttranslational modification. Acetylation is associated with decondensation of chromatin and increased transcriptional activity whereas histone deacetylation is linked to transcriptional repression (Fagiolini et al., 2009). The acetylation state is controlled by histone deacetylases (HDACs) and histone acetyltransferases (HATs) (Fagiolini et al., 2009). Histone methylation is generally associated with inactivation of transcription and condensation of the chromatin to a highly compact form.

DNA Methylation

DNA methylation is the primary contributor to the stable maintenance of particular gene expression states (Jaenisch and Bird, 2003). Gene silencing via DNA methylation occurs when a methyl group binds to the cytosine residue of a CpG island (Champagne, 2010). This addition attracts other methyl-binding proteins to the site; when these proteins are bound to the gene's promoter region, RNA polymerase cannot bind and the gene cannot be transcribed (Champagne, 2010). This process does not alter the DNA sequence. It merely limits access to the gene's promoter region. Switching between expression and silence is unlikely to occur by chance due to the presence of self-reinforcing mechanisms (Jaenisch and Bird, 2003). DNA methylation is tightly regulated; epigenomes are highly organized and surprisingly nonrandom with respect to methylation patterns (Borrelli et al., 2008). There are few methylation differences from neuron to neuron, or even neuron to neuron in two unrelated individuals.

Genomic Imprinting

With the exception of genomic imprinting, each parent contributes a single copy of all autosomal genes to their offspring at the moment of conception. Genomic imprinting is an epigenetic phenomenon independent of typical Mendelian inheritance, whereby specific experience-dependent changes can be propagated from one generation to the next. In genomic imprinting, parental genes are methylated in a parent-of-origin-specific manner. This heritable methylation pattern completely silences one parental allele, ensuring all of the genetic information is derived from the other parent. Therefore, genomic imprinting leads to selective expression of one of the two alleles, and has the potential to significantly influence brain function and behavior by altering paternal contributions to neurodevelopment and also long-term epigenetic programming (Wilkison et al., 2007).

Prenatal Epigenetic Influences on Brain Development and Plasticity

An organism's ability to survive is dependent upon its capacity to adapt to its environment. For mammals, this adaptive response begins prenatally, as the fetus adjusts to the intrauterine environment, with the assumption that the prenatal and postnatal environments are similar. The prenatal brain is exceedingly susceptible to environmental influence, as this is a sensitive period for brain development (Fagiolini et al., 2009). Gluckman and Hanson (2005) hypothesized that the epigenetic modifications experienced by the fetus in response to the intrauterine environment prepares the fetus for the particular environment it is likely to be exposed to later in life. These early-life responses, also known as fetal programming, occur rapidly in a single generation, to increase the offspring chances of survival. In response to environmental stimuli, the brain changes at the genetic, cellular, and structural levels. Throughout an organism's lifespan there is a continuous reciprocal relationship between environmental stimuli and epigenetic response. Protein synthesis is in a constant flux, resulting in a brain that demonstrates remarkable plasticity. However, evidence now indicates that if brain changes are made to compensate for prenatal disturbances, plasticity reserves may not be available later in life.

The prenatal period is a critical time for epigenetic processes, many of which are vital for development and differentiation, but others have been shown to influence growth, metabolism, and behavior. There is increasing evidence to show that maternal nutrition, physiology, and psychological states epigenetically influence the developing fetus (Champagne, 2010). Epidemiological studies of cohorts prenatally exposed to famine are associated with increased risk for schizophrenia, affective disorders, and other neurodevelopmental disorders. Many studies have linked schizophrenia to early-life adversity and differential epigenetic patterning, with postmortem analysis of brain tissue from schizophrenic individuals demonstrating aberrant epigenetic marking and altered CpG methylation of over 100 genes. Similarly, prenatal maternal stress has been associated with a plethora of negative long-term neurological outcomes. Studies demonstrate that gestational stress effects expression of hundreds of genes in the developing brain, including DNA Methyl Transferases (DNMTs), HDACs, and HATs. The implications of these findings are significant, as the epigenetic changes identified in hundreds of genes persist long after the initial insult, altering the plastic response to all future experiences.

Postnatal Epigenetic Influences on Brain Development and Plasticity

Epigenetics and Postnatal Brain Development

Although epigenetic mechanisms have been intrinsically implicated in mediating the heightened levels of neuroplasticity early in development, the ongoing plasticity demonstrated by the developing and mature brain also results from epigenetic processes. Synaptic networks and cortical tracts are extremely permeable to sensory stimulation and environmental influence during 'critical periods' of postnatal development. The heightened level of epigenetic activity that permits substantial learning and reorganization during these temporal windows has been attributed to the activation and subsequent inhibition of specific molecular pathways and signaling cascades (Fagiolini et al., 2009). Research demonstrates that HDAC, HAT, and DNMT activity is essential for neuroplasticity during these time periods and knocking out any of these genes disrupts cortical organization and function. In addition, the maturation of myelin (which often accompanies closure of critical windows) is associated with epigenetic activation of Nogo receptors, a network known to block neuroplasticity. It

appears that critical windows are associated with heightened epigenetic activity, but once windows close stable inhibitory epigenetic marks are added to the genome. Although these inhibitory marks are less responsive to environmental influences, the normal decrease in plasticity associated with the hyper-methylated aging brain can be reinstated with HDAC and DNMT inhibitors. HDAC and DNMT inhibitors have been administered via nutritional supplements and drug combinations, but further work is necessary to understand the mechanisms underlying the specificity and persistence of these alterations.

Epigenetics and Brain Plasticity

In addition to epigenetic regulation of postnatal brain development, epigenetics is also involved in learning and memory, and experience-dependent neural plasticity. Studies demonstrate that the adult neuronal response to environmental stimuli via epigenetic mechanisms and epigenetic regulation of a multitude of genes is necessary for plastic changes in the adult brain. Memory consolidation and long-term potentiation (LTP) are dynamically regulated by cascades of signaling events that are dependent upon changes in gene expression. In the hippocampus specifically, the formation of a memory following a contextual fear-learning paradigm is associated with histone acetylation, DNA methylation, and transcriptional activation of memory-enhancing genes such as Reelin and BDNF. Inhibition of DNMTs can actually block memory formation following contextual fear conditioning, whereas inhibition of HDACs enhanced formation of these memories. Furthermore, the activity of epigenetic programmers involved in learning and memory is time-specific and regionally dependent. Researchers have shown that the time period immediately following an electric shock is associated with altered methylation patterns in the hippocampus but that gene expression profiles in the neocortex preserve the memory days later (see Miller et al., 2008; Miller and Sweatt, 2007).

As it is now known that epigenetic programming is involved in normal learning and memory, reprogramming epigenetic modifiers can also restore dysfunctional learning processes. Despite prevailing evidence demonstrating the powerful effects of environmental enrichment (EE) on neuroplasticity, the underlying mechanisms have only recently been examined. EE has proven efficacious in reducing neurodegeneration and increasing neuroplasticity and long-term cognitive reserves. Recent evidence suggests that in the adult brain, EE modulates NMDA signaling pathways in the hippocampus, increases histone acetylation, and decreases global methylation patterns; each of which are epigenetic modifications associated with increased plasticity. Interestingly, EE can restore dysfunctional learning in cognitively impaired mice by increasing histone acetylation in the hippocampus and neocortex. Furthermore, this epigenetic modification is heritable and transmitted across generations; offspring of cognitively impaired mice exposed to EE show no deficit-forming memories related to contextual learning, even when cross-fostered to nonenriched mothers (Arai et al., 2009).

See also: Brain Development: Ontogenetic Neurobiology of; Genes and Behavior: Lessons from the Fruit Fly; Neural Plasticity; Prefrontal Cortex Development and Development of Cognitive Function.

Bibliography

Arai, J., Li, S., Hartley, D.M., Feig, L., 2009. Transgenerational rescue of a genetic defect in long-term potentiation and memory formation by juvenile enrichment. Journal of Neuroscience 29, 1496–1502.

Borrelli, E., Nestler, E., Allis, D., Sassone-Corsi, P., 2008. Decoding the epigenetic language of neuronal plasticity. Neuron 60, 961–974.

Champagne, F., 2010. Epigenetic influence of social experiences across the lifespan. Developmental Psychobiology 52, 299–311.

Fagiolini, M., Jensen, C., Champagne, F., 2009. Epigenetic influences on brain development and plasticity. Current Opinion In Neurobiology 19, 207–212.

Gluckman, P., Hanson, M., 2005. The Fetal Matrix: Evolution, Development and Disease. Cambridge University Press, New York.

Jaenisch, R., Bird, A., 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nature Genetics 33, 245–254.

Ladd-Acosta, C., Pevsner, J., Sabunciyan, S., Yolken, R., Webster, M., Dinkins, T., Callinan, P., Fan, J., Potash, J., Feinberg, A., 2007. DNA methylation signatures within the human brain. American Journal of Human Genetics 81, 1304–1315.

Mehler, M., Mattick, J.S., 2007. Noncoding RNAs and RNA editing in brain development, functional diversification, and neurological disease. Physiological Reviews 87, 799–823.

Miller, A.C., Sweatt, D., 2007. Covalent modification of DNA regulates memory formation. Neuron 53, 857–869.

Miller, C.A., Campbell, S.L., Sweatt, D., 2008. DNA methylation and histone acetylation work in concert to regulate memory formation and synaptic plasticity. Neurobiology of Learning and Memory 89, 599–603.

Siegmund, K., Connor, C., Campan, M., Long, T., Weisenberger, D., Biniskiewicz, D., Jaenisch, R., Laird, P., Akbarian, S., 2007. DNA methylation in the human cerebral cortex is dynamically regulated throughout the life span and involves differentiated neurons. PLoS ONE 2, e895.

Wilkison, L., Davies, W., Isles, A., 2007. Genomic imprinting effects on brain development and function. Nature Reviews Neuroscience 8, 832–843.

Epilepsy

G Campbell Teskey and Jordan S Farrell, Faculty of Medicine, Health Sciences Center, Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P.A. Schwartzkroin, volume 7, pp. 4710–4714, © 2001, Elsevier Ltd.

Abstract

Epilepsy is a common neurological disease that is characterized by an increased likelihood of experiencing seizures: events of pathological, hypersynchronous activity of neural networks. The development of epilepsy is influenced by complex interactions between genes and the environment, and is most common in young and old populations. About 60–70% of patients can control seizures with antiseizure drugs, which inhibit seizure development and propagation. The remaining individuals often require surgical resection or novel methods of controlling seizures. Current challenges include identifying therapeutic targets that prevent the development of epilepsy, understanding unexpected deaths, and addressing the causes of behavioral impairments and brain damage.

Epilepsy is one of the most common neurological diseases and is marked by the occurrence of periods of excessive neural synchrony (an electrographic seizure) that occur following innocuous stimuli that would not cause a seizure in a non-epileptic brain. Thus, the core concept of epilepsy is that the brain has a chronically low seizure threshold and is subjected to recurrent seizures. Epilepsy is a significant cause of morbidity and can be fatal in some instances. While our understanding of this disease has grown tremendously over the last century, much of its causes and consequences are still poorly understood. The purpose of this article is to provide a brief summary of current knowledge, identify knowledge gaps, and give insight into recent developments and potential treatments for epilepsy in the future.

Defining Seizures and Epilepsy

Given the heterogeneity of seizure types, there is often confusion when defining seizures and epilepsy. Recognizing this, the International League Against Epilepsy (ILAE) publishes reports that define and classify seizures and epilepsy. Currently, the ILAE defines an epileptic seizure as "a transient occurrence of signs and/or symptoms due to abnormal excessive and synchronous neuronal activity in the brain" (Fisher et al., 2005). Epilepsy is defined as "a chronic condition of the brain characterized by an enduring propensity to generate epileptic seizures, and by the neurobiological, cognitive, psychological, and social consequences of this condition. The definition of epilepsy requires the occurrence of at least one epileptic seizure" (Fisher et al., 2005). Definitions of epilepsy have been subject to change over the years, and may be modified again in the future, as knowledge in the field of epilepsy is constantly being updated.

Focal versus Generalized Seizures

Seizures are classified by how neural networks that participate in seizures are distributed. Focal seizures (previously referred to as partial) arise from a synchronous, hyperactive local region of the brain. The symptoms of a focal seizure depend on which neural networks participate in the seizure. Thus, focal seizures may have motor, sensory, autonomic, and/or psychic features. Focal seizures are further classified as to whether awareness is retained or altered (dyscognitive).

Generalized seizures may start as a focal seizure (as in focal epilepsy), but spread rapidly and bilaterally to distributed networks in both hemispheres. Alternatively, in primary generalized epilepsy, seizures begin in more widespread neural networks. Generalized seizures can further be classified as follows:

- tonic – clonic (formerly grand mal) – rigidity followed by rhythmic muscle contractions;
- absence (formerly petit mal) – brief periods of staring, blinking, or movements;
- clonic – rhythmic muscle contractions;
- tonic – Contraction of flexor and extensor muscles, rigid posture;
- atonic – Loss of muscle tone resulting in drop attacks; and
- myoclonic – Jerking of extremities symmetrically.

Etiology

While a typical mammalian brain can have a seizure induced through a variety of means including; electrical stimulation, lesions, convulsant drugs, and rapid withdrawal from chronic alcohol exposure, the causes of epilepsy are classified as genetic, structural-metabolic, or unknown. Genetic epilepsies (formerly termed idiopathic) occur as a result of mutations that directly promote both seizures and epilepsy. For example, mutations in genes that encode ion channels can contribute to abnormal excitability and stimulate a state of hypersynchrony. As the name suggests, structural/metabolic epilepsies result from brain abnormalities in structure and/or metabolism. For instance, epilepsies that develop from injuries to the brain (e.g., stroke, trauma) are classified as structural/metabolic. In the case, where there are no known structural/metabolic and genetic abnormalities, they are termed unknown. The number of unknown epilepsies continues to shrink with the

development of superior technologies that identify structural, metabolic, or genetic causes.

It should be noted that the boundaries between these epilepsies are not clear-cut. Genetics and changes in brain structure and metabolism likely interact to produce epilepsy. Genetic predispositions, for example, may make an individual more likely to develop epilepsy after a brain injury. Thus, the etiology of epilepsy should be understood as an interaction of genes and the environment.

Status Epilepticus

For the most part, seizures are self-terminating. In some cases, however, the driving force of the seizure does not abate, often due to a brain insult or drug withdrawal, and pharmacological intervention with a fast-acting benzodiazepine, barbiturate, general anesthesia, and/or ketamine is required to end the seizure. Status epilepticus is the term given to this life-threatening condition where seizures do not self-terminate. Mortality due to status epilepticus is much more common in adults than children (26 vs 3%, respectively) (DeLorenzo et al., 1996) and survivors often live with significant morbidity due to brain damage. Previously, 20 min of continuous seizure activity or bouts of seizure activity without regaining awareness was the accepted duration before pharmacological intervention was recommended. Today, however, 5 min of seizure activity is recommended because we have learned that most seizures terminate in less than a few minutes and seizures over 5 min in duration are unlikely to self-terminate (Lowenstein et al., 1999).

Sudden Unexpected Death in Epilepsy

Sudden unexpected death in epilepsy (SUDEP) refers to sudden death in patients with epilepsy where there is no identifiable cause of death (e.g., status epilepticus, drowning, falling), based on pathology and toxicology (Tomson et al., 2008). Most evidence suggests that SUDEP occurs during or, more often, after a seizure. Therefore, SUDEP is not entirely 'unexplained,' since recent seizures seem indirectly causal, but the direct cause of death remains unknown. Respiratory and cardiac failure may be common causes of death, but events leading up to this have yet to be characterized. Given that people with epilepsy die unexpectedly at a rate that is 24 times that of the general population (Ficker et al., 1998), SUDEP remains a large health challenge.

Epidemiology

Banerjee et al. (2009) has reviewed studies from around the world and concluded that the prevalence of epilepsy ranges from 0.2 to 4.1%. It is also noted that developed nations have lower prevalence and incidence of epilepsy as compared to developing nations. Some studies suggest that low-socioeconomic status, male gender, and some races have a higher prevalence of epilepsy, although these results are not consistent.

Seizures are most common in early development and senescence. In early life, the brain may be subjected to numerous factors that provoke seizures like hypoxia, inflammation, injury, and poor pre- and postnatal care. Febrile seizures are the most common type of seizures in infants and young children with up to 5% of children between 6 months and 5 years having had at least one febrile seizure (Reid et al., 2009). Although early life seizures have often been considered benign events, longer seizures are associated with an increased risk of developing epilepsy later in life. The aged brain is also subjected to events that increase the likelihood of seizures. Stroke is the most common cause of seizures in the elderly but epilepsy can also be caused by neurodegenerating diseases, dementia, tumors, head injury, and infection.

Diagnosis and Treatment

Diagnosing epilepsy no longer requires a physician to witness a seizure. In fact, epilepsy can often be diagnosed with an electroencephalogram (EEG) when a patient is not experiencing seizures. An EEG can detect abnormal changes in electrical potential across electrodes secured to a patient's scalp. In some epilepsies, however, interictal (between seizures) EEG can be completely normal and physicians will need to employ other diagnostic strategies. Patient or witness descriptions of seizure events (e.g., duration, loss of awareness, behavior, auras) will give a physician an idea of what type of epilepsy is most likely. Other techniques, such as Magnetic resonance imaging (MRI), can help diagnose epilepsy by identifying structural abnormalities associated with epilepsy (e.g., scarring). It is also important to conduct blood work to ensure seizures are not a consequence of a separate condition.

The first line of treatment for epilepsy is antiseizure drugs (ASDs), which inhibit seizure development and propagation. ASDs are often referred to as anticonvulsant and antiepileptic drugs, although these terms are inaccurate. ASDs have a diverse range of mechanisms to raise seizure thresholds. Common mechanisms include enhancing the action of the inhibitory neurotransmitter GABA (e.g., gabapentin), stabilizing the inactive state of sodium channels (e.g., carbamazepine, valproate), and inhibiting T-type calcium channels (e.g., ethosuximide). Most ASDs have broad mechanisms of action, many of which are unknown, that overlap with other ASDs. Most people with epilepsy will be able to work with their physician and find a drug and dosage that causes few side effects.

Unfortunately, 30–40% of people with epilepsy are unable to completely control seizures with ASDs and are termed refractory, pharmacoresistant, or intractable. Children and some adults with refractory epilepsy often benefit from the ketogenic diet: a high fat, low carbohydrate diet with anticonvulsant action. In adults, surgical resection of epileptogenic tissue is the most common treatment for intractable epilepsy and has a success rate of 70%. More recently, deep brain stimulation (DBS), a very successful treatment for Parkinson's disease, may be useful in refractory epilepsy patients. Bilateral stimulation of the anterior thalamus has been successful at reducing seizure frequency in many patients and DBS shows promise but more research is needed.

As is evident from the examples above, epilepsy treatments are focused on neural excitability. Even though new ASDs have

been developed, the percentage of patients with refractory epilepsy has not changed. In other words, new ASDs are still not treating intractable epilepsy, although they may have fewer side effects. Current research suggests that glia may be an appropriate pharmacological target (Devinsky et al., 2013), especially in intractable epilepsy. Astrocytes play a very important role in regulating the extracellular space, which can have a profound influence on seizure susceptibility. Astrocytic membrane proteins Aquaporin-4 and Kir4.1 regulate water and K^+ buffering, respectively, and are impaired following seizures. As a result of inadequate buffering, neural excitability is increased and may contribute to epileptogenesis. Given the importance of glia in regulating synaptic transmission, inflammation, blood flow, and neural excitability, more research is needed to identify potential therapeutic targets.

Consequences of Seizures

There are numerous structural and functional changes that occur after seizures, which give rise to specific behavioral impairments. To understand how the brain changes following seizures, it is important to understand what happens before, during, after, and between seizures.

Preictal Period

The preictal period occurs just before a seizure and is increasingly becoming a focus of epilepsy research. Identifying reliable markers that are expressed during the transition from preictal to ictal may allow patients to predict seizure onset and allow intervention. EEG changes in network activity can be observed minutes to hours before a seizure. There are also pronounced blood flow changes before seizures that may aid in identifying seizure onset (Baumgartner et al., 1998). Using functional MRI, Federico et al. (2005) found that a signal (Blood Oxygen Level Dependent) could be detected several minutes before seizure onset. Future treatments may employ implanted devices that measure these changes and intervene with electrical stimulation or pharmacological agents prior to a seizure.

Ictal Period

The intense neuronal firing during the ictal period is thought to be ultimately responsible for neuronal death, especially following prolonged seizures. The exact mechanism that leads to cell death has not been definitively determined but a large influx of calcium, which can lead to programmed cell death likely plays an important role. Following prolonged and severe seizures, as with status epilepticus, brain damage can be quite extensive.

Postictal Period

Following seizures, many genes are upregulated, including proinflammatory cytokines. These cytokines are thought to underlie the increased blood–brain barrier permeability that follows seizures as well as activation of astrocytes and microglia. Stains that select for dying neurons show that neuronal death peaks 12 h after a seizure and is thought to be due to the proapoptotic mechanisms activated during the ictal period.

Behaviorally, seizures often lead to deficits that are specific to brain regions involved in the seizure. In 1849, Dr Robert Bentley Todd first described a phenomenon that bears his name: Todd's paresis. It is frequently defined as motor paralysis contralateral to the seizure but may also affect a variety of sensory, motor, or cognitive processes, depending on where the seizure occurred. Furthermore, patients with medial temporal lobe epilepsy (TLE) and animal models of TLE have pronounced spatial memory deficits after seizures. These deficits last several minutes to hours and cannot be explained by flattening of the EEG. A few case reports have noted that Todd's Paresis is accompanied by reduced blood perfusion (Yarnell, 1975), but this has yet to be fully explored. There are data that show that this postictal period is associated with a pronounced ischemia/hypoxia that in turn appears to be responsible for the behavioral impairment (Teskey and Farrell, 2012).

Interictal Period and Comorbidities

With repeated seizures, changes in the brain occur that result in permanent behavioral disruption even when seizures are not occurring. These behaviors are often difficult to describe in patient populations because of the confounding effects of ASDs. Animal models have shown that seizures cause network reorganization and greater excitability, which disrupts the behavioral output of that brain region. With repeated neocortical seizures in rats, for example, movement representations on the motor cortex have lower thresholds, are much larger and have more multiple movements (Teskey et al., 2002). Behaviorally, these characteristics manifest as poor performance on skilled motor tasks. Thus, behaviors in the interictal period are specific to seizure networks and result in specific comorbidities (see Section Focal versus Generalized Seizures).

In patients with TLE, the most common pathology is hippocampal sclerosis, which is characterized by severe neuronal loss and an overabundance of glial cells, and is found specifically in the CA1 and subiculum of the hippocampus. It is important to note that MRI does not always detect hippocampal sclerosis. Recent developments in animal models suggest that glia may play a large role in epileptogenesis (see Section Epidemiology). Briefly, astrocytes create an extracellular environment that raises the excitability of neurons. Repeated seizures are also associated with production of inflammatory agents. Together, these signaling molecules enhance neuronal excitability. In sum, many glial-mediated changes occur in epileptic tissue and result in hyperexcitable tissue, which may explain interictal behaviors and the increased likelihood of experiencing seizures.

People with epilepsy also have neuropsychiatric disorders at higher rates than the general population. A 2011 survey of 172 959 people revealed that people who reported having epilepsy were significantly more likely to indicate that they had anxiety, depression, bipolar disorder, ADHD, sleep apnea, and movement disorders/tremors (Ottman et al., 2011). Depression is the most common comorbidity with 32.5% of

people with epilepsy also reporting depression. Interestingly, the relationship is bidirectional: depression increases the risk of epilepsy and epilepsy increases the risk on depression. While these comorbidities help our understanding of epilepsy, it is important to consider other factors that may increase the prevalence of comorbidities (e.g., antiepileptic drugs, psychosocial factors).

Conclusion

Our understanding of epilepsy has grown tremendously over the years. Epilepsy is influenced by interactions of genetic and environmental factors, which create brain changes that make seizures more likely. Epilepsy is more common in the young and old, partly due to seizure-promoting factors that are commonly experienced at these ages. New generations of ASDs may have decreased side effects for patients, but the amounts of refractory cases still remain to be 30–40%. Surgical resection of epileptic tissue is still the gold standard for intractable cases, but emerging technologies (e.g., DBS, seizure monitoring implants) could provide more treatment options. Furthermore, the role of glia in the CNS has been intensely studied in the last decade and potential implications in epilepsy are being explored. Given the role of glia in epileptogenesis, glial-targeted therapies could provide a new avenue of treatment – the development of epilepsy (i.e., not just the prevention of seizures). Ongoing challenges in the field of epilepsy are refractory cases, SUDEP, and understanding postictal and interictal phenomenon.

See also: Electroencephalography: Basic Principles and Applications; Kindling; Neural Plasticity.

Bibliography

Banerjee, P.N., Filippi, D., Hauser, A.W., 2009. The descriptive epidemiology of epilepsy – a review. Epilepsy Research 85 (1), 301–303.

Baumgartner, C., Serles, W., Leutmezer, F., Pataraia, E., Aull, S., Czeck, T., Pietrzyk, U., Relic, A., Podrek, I., 1998. Preictal SPECT in temporal lobe epilepsy: regional cerebral blood flow is increased prior to electroencephalography-seizure onset. Journal of Nuclear Medicine 39 (6), 978–982.

Devinsky, O., Vezzani, A., Najjar, S., De Lanerolle, N.C., Rogawski, M.A., 2013. Glia and epilepsy: excitability and inflammation. Trends in Neurosciences 36 (3), 174–184.

DeLorenzo, R.J., Hauser, W.A., Towne, A.R., Boggs, J.G., Pellock, J.M., Penberthy, L., Garnett, L., Fortner, C.A., Ko, D., 1996. A prospective, population based epidemiologic study of status epilepticus in Richmond, Virginia. Neurology 46, 1029–1035.

Federico, P., Abbott, D.F., Briellman, R.S., Harvey, A.S., Jackson, G.D., 2005. Functional MRI of the pre-ictal state. Brain 128 (8), 1811–1817.

Fisher, R.S., Boas, W.V.E., Blume, W., Elger, C., Genton, P., Lee, P., Engel, J., 2005. Epileptic seizures and epilepsy: definitions proposed by the international league against epilepsy (ILAE) and the international bureau for epilepsy (IBE). Epilepsia 46, 470–472.

Ficker, D.M., So, E.L., Annegers, J.F., O'Brien, P.C., Cascino, G.D., Belau, P.G., 1998. Population-based study of the incidence of sudden unexplained death in epilepsy. Neurology 51, 1270–1274.

Lowenstein, D.H., Bleck, T., Macdonald, R.L., 1999. It's time to revise the definition of status epilepticus. Epilepsia 40, 120–122.

Ottman, R., Lipton, R.B., Ettinger, A.B., Cramer, J.A., Reed, M.L., Morrison, A., Wan, G.J., 2011. Comorbidities of epilepsy: results from the epilepsy comorbidities and health (EPIC) survey. Epilepsia 52, 308–315.

Reid, A.Y., Galic, M.A., Teskey, G.C., Pittman, Q.J., 2009. Febrile seizures: current views and investigations. Canadian Journal of Neurological Sciences 36 (6), 679–686.

Teskey, G.C., Farrell, J.S., 2012. Brief seizures induce a severe ischemic/hypoxic episode. Epilepsia 53 (5), 104–105.

Teskey, G.C., Monfils, M.H., VandenBerg, P.M., Kleim, J.S., 2002. Motor map expansion following repeated cortical and limbic seizures is related to synaptic potentiation. Cerebral Cortex 12 (1), 98–105.

Tomson, T., Nashef, L., Ryvlin, P., 2008. Sudden unexpected death in epilepsy: current knowledge and future directions. Lancet Neurology 7 (11), 1021–1031.

Yarnell, P.R., 1975. Todd's paralysis: a cerebrovascular phenomenon? Stroke 6, 310–313.

Episodic Memory

Costanza Papagno, Università di Milano Bicocca, Milano, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Episodic memory is the memory of events in our own past and is a form of long-term, declarative memory. But how are episodic memories formed and how are they retrieved? The mechanisms and structures that determine whether an experience will be remembered or forgotten, the stages of memory processing, namely encoding, consolidation, and retrieval will be described, as well as the role of context in such processes.

Introduction

Information that is acquired in the course of experience, persists, and so can be consciously retrieved long after the experience is past is 'episodic' memory. However, episodic memory is not only memory of events in our personal past, but it also helps us to stay oriented in the present, and make plans for the future. This type of memory involves oneself as a participant and therefore it seems to be synonymous with autobiographical memory. Wheeler et al. (1997) have suggested that the type of awareness evident during retrieval primarily characterizes episodic memory whereas the content of the retrieved material (self-related) defines autobiographical memory. Episodic memory involves reexperiencing and being aware of the continuity of the experiencing self across time; autobiographical memory refers to the information that involves the rememberer but need not to entail the same subjective awareness (Gilboa, 2004). Autobiographical reexperiencing is only one aspect of autobiographical memory and seems unique to humans (more details about the distinction between episodic and autobiographical memory can be found in Conway, 2001).

The Canadian psychologist Endel Tulving in 1972 defined episodic memory as the conscious knowledge of temporally dated, spatially located, and personally experienced events or episodes. Episodic memory can be contrasted with semantic memory, our generic knowledge of the world. The relationship between the two remains controversial: one view suggests that semantic memory is simply the accumulation of many episodic memories for which the detailed contextual cue has disappeared, leaving only generic features (Squire, 1992), while Tulving (2002) postulated that episodic memory evolved out of semantic memory; many animals possess well-developed knowledge of the world and are capable of acquiring new information. Early humans were like these animals, but during evolution, episodic memory emerged. Its operations require, but go beyond, the semantic memory system.

Episodic memory has a context: when you recollect details about some episode, you are engaged in a sort of 'mental time travel,' and you are aware that the information you hold about it is bound to a particular experience. When events are retrieved, they are reexperienced in a quasiperceptual way, with conscious awareness. Every episodic memory involves a mental representation of the self as the agent or recipient of some action, or the experiencer of some state. Moreover, we have a conscious awareness that our reexperience of the past is different from our ordinary awareness of what is going on in our environment: the term 'autonoetic' awareness refers to the special kind of consciousness that allows us to be aware of the subjective time in which events happened. Consequently, episodic memory requires a subjectively sensed time.

Yet, the way episodic memory is tested is through learning lists of words (testing the to-be-learned material by means of free recall, cued recall, or recognition), stories recall or nonverbal material recall (such as immediate and delayed reproduction of spatial sequences or meaningless pictures): in such a way, one of the main features of episodic memory is missing, namely the temporospatial context. As Tulving (2002) pointed out, traditional laboratory experiments are almost invariably concerned with 'what,' but not with 'when' and 'where.' Therefore, episodic memory tests, for instance recalling lists of unrelated words, do not exactly deal with episodic memory. More recently, paradigms assessing also when or where a particular item was learned have been used.

Episodic memory processing involves three stages, interacting with each other: 'encoding,' 'consolidation,' and 'retrieval.' We experience an event and some aspects of this event get encoded. Encoding initiates a series of processes that take time – consolidation – and produces a memory trace. The stored memory trace can be retrieved. The original experience may involve the emotional arousal that was associated with the episode and this emotion may be remembered when the corresponding event is retrieved. Almost invariably, events are interpreted in a meaningful way through semantic knowledge. Such a complex process requires the activity of several brain structures and mechanisms.

Encoding

The term encoding indicates the various processes by which information is transformed into a memory representation. Episodic encoding requires the binding together of the various features of an event into an integrated memory representation. I make an experience, for example, I meet people during a party in a particular place one evening at a certain time of the year; I notice how they are dressed and look like, their voice and accent, the perfume, and perhaps I also think that they are boring or interesting people, therefore interpreting the experience and coloring it with an emotion. All of that produces a mental representation that records some or all of the aspects of this episode. The major feature of episodic representations is, therefore, that they associate together very different kinds of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.51032-0

information, including perceptual and semantic characteristics of objects, and how they are located in space and time relative both to each other and the self.

Several factors can affect encoding. One is the level at which we attend the experience, namely the amount of attention that we devoted to it, and the level at which we elaborate, interpret its meaning, relate it to other experiences and organize it.

In 1972, Craik and Lockhart developed a theory, the so-called 'level of processing' theory, which gave particular emphasis to the depth of processing. They suggested that the amount of information retained in long-term memory depends on how deeply it is processed during learning. For example, a given stimulus, say a word, can be processed in terms of its physical superficial features, or converted into its sound, or by evoking its meaning. Processing the word in terms of its meaning, requiring one to create a more durable memory trace, is deeper. In a later study, Craik and Tulving (1975) asked three groups of healthy participants to make judgments about viewed unrelated words: a group decided whether these words were written in lower or upper case, a second group had to make a rhyme judgment with a different word, and a third group had to decide whether a word fitted in a sentence. Afterward, the three groups were submitted to a recognition task, in which they had to categorize a list of words as either those that had already been presented or as new words. Participants who had to retrieve the appropriate meaning of the words (to decide whether they fitted in the sentence) obtained the best performance. Episodic memory would benefit, therefore, from this deep, semantic processing. However, patients with 'semantic dementia' who cannot meaningfully interpret episodes encode and recognize the perceptual features of experienced objects in a relatively normal way and they also can cope with spatiotemporal contexts (Papagno and Capitani, 2001). This suggests that adequate representations of perceptual features and spatiotemporal context can probably be made even without semantics.

The Craik and Tulving's experiment is an example of incidental learning, since participants were unaware that they would be tested for memory. There was not a purposeful intention to learn. This explains how we retain everyday experiences that we do not deliberately attempt to encode. However, attention can play a crucial role in encoding (see below).

The concept of processing depth is not particularly clear, and has been repeatedly criticized. It would be better explained in terms of richness and flexibility of the system. These two characteristics make semantic coding widely used in long-term memory: information that is encoded in terms of a rich and detailed description is likely to be more accessible than material that is processed in terms of simpler schemes. Generally, an effortful encoding leads to better recall. One is more likely to remember later the solving of a difficult problem than that of an easy one. A deeper elaboration of stimuli may result in distinctive records that are more discriminable from other encoded events and therefore can be more easily retrieved.

As mentioned, learning can be incidental or intentional. One reason for poor encoding is lack of attention. A failure to attend an event has as a consequence a reduced ability to retrieve it in the future, as it can be seen in eyewitness, when we can discover that many details of a scene have been completely ignored. Encoding is also reduced when we attend two tasks simultaneously. When we are performing a task, for example, cooking some elaborate meal without being an expert chef, and someone talks with us, we are more likely to forget what has been told, since we were involved in a distracting performance. Accordingly, the role of attention has been explored in laboratory, using dual task paradigms. In this type of task, the participants' attention is divided between encoding the information presented and performing a secondary task (for example, monitoring the position of an asterisk on a computer screen and press a button when it changes). This has detrimental effects on memory performance relative to conditions where full attention is paid to encoding the items (Craik et al., 1996).

A third aspect that helps encoding, beside level of processing and attention, is 'organization.' When participants are instructed to organize some material, and subsequently they are tested for recall, they retain this material just as well as participants who are instructed to learn it. In other words, there are major advantages in subsequent recall when participants are required to organize material. The term organizing generally refers to the grouping of items on the basis of shared semantic attributes (Gershberg and Shimamura, 1995), due to the familiarity that people already have with the to-be-learned material. Thus, a task requiring organization of material will overlap considerably with a deep encoding task since it requires processing of its meaning. Organizing has the additional requirement of manipulating material on the basis of its similarities or differences to other material.

A fourth variable affecting encoding is the 'distribution of practice in time.' The German psychologist Hermann Ebbinghaus (1850–1909) showed that it is better to distribute learning over two days relatively evenly rather than concentrate all in one day. He tested himself in learning meaningless consonant-vowel-consonant syllables (such as GUW, BEX), using both massive and distributed practice. At recall, he noticed that a suitable distribution of repetitions over time was more efficient than the massing of them in a single session, showing that distributed practice is better for encoding than massive practice. What is not clear, however, is how long the intertrial interval should be. This spacing effect can be related to several reasons: for example, attention can decrease when trials are massed together. When trials are spaced, the context can change enhancing a richer memory trace and additional retrieval routes to our memory. Indeed, the processing performed for the initial encounter and for repeated trials in massed practice is likely to be very similar: items are presented in the same context and in the same way. Distributed practice, on the other hand, fosters greater encoding variability: the context can change. A stimulus is more likely to be remembered when it is processed in different ways across trials. More recent research (Litman and Davachi, 2008) has suggested that the benefit of distributed learning is enhanced memory consolidation. More specifically, the time period of consolidation between distributed encoding session helps to stabilize memory traces. In addition, the reactivation of memory representations during the second encoding session may further facilitate consolidation.

The Context Effect

Context plays a crucial role. The effect of context at encoding depends on the contextual similarity between the target information and any subsequent interfering information.

When contextual information provides enough details to uniquely specify an event and distinguish it from other similar events stored in memory (Murnane et al., 1999), it reduces interference. This is evident in our everyday experience when retrieval of information does not immediately follow its encoding, but other information is acquired in the interval between encoding and retrieval of the same material. Indeed, learning different material in two different contexts confers a memory advantage over learning all the material in the same context. To sum up, learning different material in the same context impairs recall, due to interference, while the same context at encoding and retrieval enhances recall. This phenomenon, that is 'context dependency,' is well evident in a free recall experiment (Godden and Baddeley, 1975), in which divers learnt lists of words in two different environments: on the shore and underwater, and recalled the words in either the environment of original learning or in the alternative environment. Lists learnt underwater were best recalled underwater, and vice versa. Similarly, alcoholics may hide objects when drunk, forget where they have put it when sober, but remember it again in a subsequent binge; also lists of words learnt after having smoked marijuana were better recalled when these participants smoked again just before retrieval (Eich, 1980). Tulving (1983) has argued that the extent to which there is contextual similarity between the retrieval attempt and the initial encoding phase predicts the likelihood of remembering vs forgetting (the so-called encoding specificity hypothesis). Therefore, a crucial characteristic of episodic memory is that when an event is encoded, features of the physical environment are typically bound into the resulting episodic memory representation, providing an additional route to later recall. If these features will be in the environment at the time of retrieval, they will act as cues to memory and will help retrieval of further details of the event.

Neural Correlates of Encoding

The brain must not only encode the specific aspects of an event but also bind them in a manner that can later specify the spatiotemporal context in which the to-be-remembered event occurred. Converging evidence indicates that different regions of the medial temporal lobes may form representations of items, contexts, and item-context bindings and that areas in the dorsolateral prefrontal cortex are related to the processing of relationships between items. Indeed, the dorsolateral prefrontal cortex activity during encoding predicts subsequent memory specifically if participants are given encoding tasks that encourage relational processing or retrieval tests that are sensitive to memory for associations among items (Ranganath, 2010). Encoding requires in particular the activity of the left inferior prefrontal cortex, when deep encoding, that is semantic processing of verbal stimuli, is performed (Kapur et al., 1994). When visual learning is assessed, there is a greater engagement of the right frontal lobe (Brewer et al., 1998). This material-specific model of hemispheric asymmetry coexists with a process-specific model known as hemispheric encoding–retrieval asymmetry (HERA). According to the HERA, left prefrontal cortex is more involved than the right prefrontal cortex in episodic memory encoding, whereas the right prefrontal cortex is more involved than the left in episodic memory retrieval (Tulving et al., 1994).

The prefrontal activity also controls how attention is directed and coordinates posterior sensory cortices where components of an episode are represented. The prefrontal cortex seems also to represent the temporal features of an episode.

It appears therefore that the frontal lobes are heavily involved in episodic memory: an extensive discussion about the relation among episodic memory, autonoetic consciousness, and the frontal lobes can be found in Wheeler et al. (1997).

Consolidation

Memory changes with the passage of time after learning and after the initial registration of information, and a fraction of experienced episodes is permanently stored in the brain through consolidation, the maintenance of information over time. In sum, consolidation is the process that modifies the memory traces so that they become more stable over time: the memory is established in storage.

Memory consolidation is inferred from the existence of gradients of retrograde amnesia after damage to the brain, more specifically to parts of the medial temporal lobe: indeed, when the medial temporal lobe is damaged, old memories are in general better preserved than more recent ones. It is therefore assumed that damage interferes with some time-dependent process, which functions to make memory permanent. It is however not clear how long the period during which memory can be disrupted lasts before consolidation takes place. Following one view, early, fragile memory depends on the hippocampal system, located in the medial temporal lobe, while stable memory does not (Squire, 1992). Squire's 'standard model' of memory consolidation assumes that consolidation involves long-term interactions between hippocampal and extrahippocampal circuits. Originally, this model was intended to apply to explicit (declarative) forms of memory, including both semantic and episodic memory. The standard model supposes that the hippocampus plays a time-limited role in memory storage, either by first storing and then transferring a memory, or by acting as an index that binds extrahippocampal storage sites during a period when direct communication among these sites is inadequate to permit memory retrieval. These extrahippocampal sites then acquire the ability to interact such that the hippocampus is no longer required to retrieve memories. Nadel and Moscovitch (1997) challenged the assumption of a long-term consolidation processing strengthening neocortical neurons and resulting in the transfer of information initially stored in the hippocampal complex to the neocortex. They generated a new model called the 'multiple trace theory'; according to it, the hippocampal complex encodes all information that is attended or consciously apprehended and this information is sparsely encoded in a distributed ensemble of hippocampal neurons. Because the hippocampal complex encodes all information that is attended, each reactivation of a memory trace results in the creation of a newly encoded hippocampal trace, which again is sparse and distributed. By virtue of indexing a similar set of neocortical neurons encoding the features of the episode,

each trace shares some or all of the information about the initial episode. Therefore, episodic memory always depends upon the hippocampal complex and its provision of contextual information. As episodic memories age, they would either be forgotten or would have benefitted from the formation of multiple traces in the hippocampal complex and neocortex. Thus, older episodic memories would be associated with a greater number of traces. Retrieval would become easier as the number of traces, and the number of access routes to them, increase. Because memory traces are sparse and distributed, even minimal damage at any location in the hippocampal formation could affect acquisition, retention, and recovery of any single one of them. Newly acquired traces would be particularly vulnerable, but older memories, which are multiply represented, would be able to withstand the loss of more hippocampal complex tissue. As a result, partial damage to the hippocampal complex would affect remote episodic memory in relation to its age, producing a graded retrograde amnesia. Each time an episodic memory is retrieved a replicate trace is created within the hippocampal complex, spreading out the representation of that memory.

Retrieval

To remember a past event, information must be retrieved from episodic memory. Retrieval produces the subjective experience of consciously remembering the past. Episodic retrieval involves an interaction between a retrieval cue, self-generated or provided by the environment, and a memory trace, leading to the reconstruction of the aspects (some or all) of the past episode (Tulving, 1983). This interaction between the memory trace and the retrieval cue (and what follows) was termed 'ecphory' by Semon (Rugg and Wilding, 2000). Retrieval can take place through recall or recognition. Recall of episodic information is usually intentional and effortful, but it can occur unintentionally and automatically when a particular thought or perception triggers the recollection of an experience. Tulving argues that the same processes are involved in recall and recognition, although the particular retrieval cues are different, in recognition being copy cues, i.e., copies of the original stimuli. The memory trace, however, is not a static entity waiting for an appropriate cue. Rather, retrieved memories are the result of a complex interplay of stored information and retrieval cues.

Several factors influence whether an episodic retrieval attempt will be successful or not; one of such factors is the way in which the event was encoded, as mentioned above.

As reported, experiences such as remembering last birthday party or our first job interview represent unique events tied to a specific point in time and in space. The retrieval of this type of information is strongly influenced by the context within such episodes occurred, as the reported experiment on divers illustrates. However, a subsequent experiment on divers (Godden and Baddeley, 1980) using recognition instead of recall, found no significant context-dependency effect, and this is also confirmed with a wide range of drugs, including those used to anesthetize patients (Eich, 1980). This seems to suggest that recall and recognition are not the same processes, as Tulving proposed. It appears that the person's internal state helps to access the memory trace, but when the access is made easier by presenting an item for recognition, the initial search stage is not necessary.

In addition, results from cognitive, neuropsychological, and neuroimaging studies of human memory indicate that recognition memory performance reflects two distinct processes, referred as recollection and familiarity. Recollection is defined as the mental reinstatement of previous events and reflects the retrieval of qualitative information about a specific study episode, such as when or where an event took place, while familiarity reflects a more global measure of memory strength or stimulus recency without retrieval of any specific details about the study episode (Yonelinas et al., 2010). During recollection, unique details of a memory are recalled, which may include additional sensory information, such as the sounds paired with an event, or the emotions experienced during the initial encoding of it. In contrast, familiarity is a mental awareness, a subjective sense that an event has been experienced in the past, but the memory is lacking the unique details or mental reinstatement of the event that accompanies recollection. The difference between these two processes is easily demonstrated in everyday life by the example of meeting a person we recognize on the street. Sometimes we can specifically place where, or when, we had met her/him. This is a recollective process: we recognize an item on the basis of the retrieval of specific contextual details. In contrast, we sometimes strongly feel that we have met this person before, but cannot identify where or when we originally encountered her/him. The person is familiar, despite the fact that we cannot recollect unique details of the initial meeting. This is a familiarity process: a stimulus is matched against information in memory and if the stimulus and the stored information match or there is sufficient overlap, we find it familiar without having specific details. Several dual-process models have been developed to account for the observation that recognition memory judgments can be based on recollection or familiarity that have distinct modes of operation.

But how does a retrieval cue serve to bring back details about the past? Episodic memories are encoded by binding together the various features of a stimulus or event into an integrated representation, so an episodic memory consists in a conjunction of linked features. Any of these features is a potential route to memory and makes possible to have access to memories even when information is limited. When a retrieval cue – external or internal, namely provided by our own state – that corresponds to part of the encoded information finds a stored representation, other features bound to this representation are reactivated, reconstructing a whole from linked parts.

Neural Correlates of Retrieval

The distinction between familiarity and recollection has prompted a debate to define the role of the hippocampus and adjacent structures in recognition memory. Neuroimaging data have shown that in the hippocampus and the adjacent cortex, retrieval associated with recollection produces more activation than retrieval associated with familiarity (Eldridge et al., 2000), suggesting that this structure is more involved in recollection than in familiarity judgments or, alternatively, that a greater amount of information is retrieved during recollection

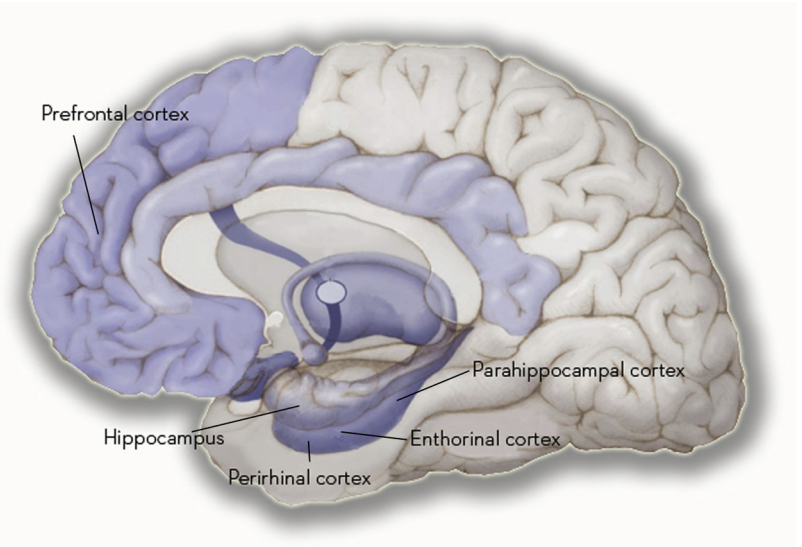

Figure 1 Prefrontal cortex and medial temporal lobe structures involved in episodic memory.

than familiarity judgments. In line with neuroimaging results, in patients with a damage to the hippocampal region, familiarity is as severely impaired as recollection (Manns et al., 2003). Apparently, both types of recognition memory depend on the integrity of structures within the medial temporal lobe, including the hippocampal region, but episodic recollection depends also on the frontal lobes (Shallice et al., 1994). Indeed, patients with frontal damage have difficulty in recollecting the details of earlier personal events. More specifically, the frontal lobes are important when we develop a strategy to retrieve, so that we select and represent the cues that can be used to reconstruct a memory. The frontal lobes also support mechanisms that resolve competition or interference between competing memories: more than one memory can be retrieved from a single cue and the memories compete to be fully retrieved. Interference during retrieval can produce forgetting and patients with frontal lesions are particularly prone to interference-based forgetting (Shimamura, 1995). Finally, the frontal lobes are important for evaluating or monitoring the quality and quantity of retrieved information (Figure 1).

There are, though, opposite views concerning the relative contribution of the hippocampus and the surrounding cortex to recollection and familiarity, since patients with hippocampal damage or dysfunction and a disproportionate deficit in recollection as compared to familiarity have been described (Vann et al., 2009). This is confirmed by neuroimaging studies (Ranganath et al., 2003), in which activity in the hippocampus and parahippocampal cortex during encoding was predictive of whether participants could subsequently recollect information about the context in which an item had been studied, but activity did not differentiate between missed items and items recognized on the basis of familiarity. In contrast, activity in the perirhinal cortex correlated with subsequent familiarity-based item recognition but was not sensitive to recollection. This suggests that the hippocampus and the perirhinal cortex subserve complementary encoding mechanisms that build representations that support recollection or familiarity, respectively.

Conclusion

Our episodic memory is capable of storing large amount of information and to make accurate retrieval. Retrieval depends on the type of encoding during the original experience, with context acting as a powerful recall cue. While retrieval can be relatively automatic, sometime recollection is similar to problem solving, since we have to follow our cues and check for the result of our search in order to reconstruct the whole episode. Future research, however, is still required to investigate several open issues, such as (1) the role of attention in deciding which aspects of an event will be treated as items and which as context, (2) the relative role of the hippocampus and adjacent cortex in storage as compared to retrieval, but also in familiarity, (3) the precise role of the posterior and frontal neocortices and the medial temporal lobe in encoding, storing, and retrieval of episodes, (4) how specific aspects of an event are processed and then bound together and inserted into a temporospatial context, and, finally, (5) the relation between semantic and episodic memory and how semantic memory contributes to episodic encoding.

See also: Amnesia: General; Autobiographical Memory; Declarative Memory, Neural Basis of; Hippocampus and Related Structures; Recognition; Retrieval; Semantic Memory.

Bibliography

Brewer, J.B., Zhao, Z., Desmond, J.E., Glover, G.H., Gabrieli, J.D., 1998. Making memories: brain activity that predicts how well visual experience will be remembered. Science 281, 1185–1187.

Conway, M., 2001. Sensory-perceptual episodic memory and its context: autobiographical memory. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 356, 1375–1384.
Craik, F.I.M., Govoni, R., Naveh-Benjamin, M., Anderson, N.D., 1996. The effects of divided attention on encoding and retrieval processes in human memory. Journal of Experimental Psychology: General 125, 159–180.
Craik, F.I.M., Lockhart, R.S., 1972. Levels of processing: a framework for memory research. Journal of Verbal Learning Verbal Behavior 11, 671–684.
Craik, F.I.M., Tulving, E., 1975. Depth of processing and the retention of words in episodic memory. Journal of Experimental Psychology: General 104, 268–294.
Eich, J.E., 1980. The cue-dependent nature of state-dependent retrieval. Memory and Cognition 8, 157–173.
Eldridge, L.L., Knowlton, B.J., Furmanski, C.S., Bookheimer, S.I., Engel, S.A., 2000. Remembering episodes: a selective role for the hippocampus during retrieval. Nature Neuroscience 3, 1068–1069.
Gershberg, F.B., Shimamura, A.P., 1995. Impaired use of organizational strategies in free recall following frontal lobe damage. Neuropsychologia 33, 1305–1333.
Gilboa, A., 2004. Autobiographical and episodic memory – one and the same? Evidence from prefrontal activation in neuroimaging studies. Neuropsychologia 42, 1336–1349.
Godden, D.R., Baddeley, A.D., 1975. Context-dependent memory in two natural environments: on land and underwater. British Journal of Psychology 66, 325–331.
Godden, D.R., Baddeley, A.D., 1980. When does context influence recognition memory? British Journal of Psychology 71, 99–104.
Kapur, S., Craik, F.I.M., Tulving, E., Wilson, A.A., Houle, S., Brown, G.M., 1994. Neuroanatomical correlates of encoding in episodic memory: levels of processing effect. Proceedings of the National Academy of Sciences USA 91, 2008–2011.
Litman, L., Davachi, L., 2008. Distributed learning enhances relational memory consolidation. Learning and Memory 15, 711–716.
Manns, J.R., Hopkins, R.O., Reed, J.M., Kitchner, E.G., Squire, L.R., 2003. Recognition memory and the human hippocampus. Neuron 37, 171–180.
Murnane, K., Phelps, M.P., Malmberg, K., 1999. Context-dependent recognition memory: the ICE theory. Journal of Experimental Psychology: General 128, 403–415.
Nadel, L., Moscovitch, M., 1997. Memory consolidation, retrograde amnesia and the hippocampal complex. Current Opinion in Neurobiology 7, 217–227.
Papagno, C., Capitani, E., 2001. Slowly progressive aphasia: a four-year follow-up study. Neuropsychologia 39, 678–686.
Ranganath, C., 2010. Binding items and contexts: the cognitive neuroscience of episodic memory. Current Directions in Psychological Sciences 19, 131–137.
Ranganath, C., Yonelinas, A.P., Cohen, M.X., Dy, C.J., Tom, S.M., D'Esposito, M., 2003. Dissociable correlates of recollection and familiarity within the medial temporal lobes. Neuropsychologia 42, 2–13.
Rugg, M.D., Wilding, E.L., 2000. Retrieval processing and episodic memory. Trends in Cognitive Sciences 4, 108–115.
Shallice, T., Fletcher, P., Frith, Chris, D., Grasby, P.M., Frackowiak, R.S.J., Dolan, R.J., 1994. Brain regions associated with acquisition and retrieval of verbal episodic memory. Nature 368, 633–635.
Shimamura, A.P., 1995. Memory and frontal lobe function. In: Gazzaniga, M.S. (Ed.), The Cognitive Neuroscience. The MIT Press, Cambridge, MA, pp. 803–813.
Squire, L.R., 1992. Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. Psychological Revue 99, 195–231.
Tulving, E., 1972. Episodic and semantic memory. In: Tulving, E., Donaldson, W. (Eds.), Organization of Memory. Academic Press, New York, pp. 381–403.
Tulving, E., 1983. Elements of Episodic Memory. Clarendon, Oxford.
Tulving, E., 2002. Episodic memory: from mind to brain. Annual Review of Psychology 53, 1–25.
Tulving, E., Kapur, S., Craik, F.I.M., Moscovitch, M., Houle, S., 1994. Hemispheric encoding/retrieval asymmetry in episodic memory: auditory sentence recognition. Proceeding National Academy of Sciences USA 91, 2012–2015.
Vann, S.D., Tsivilis, D., Denby, C.E., Quamme, J.R., Yonelinas, A.P., Aggleton, J.P., Montaldi, D., Mayes, A.R., 2009. Impaired recollection but spared familiarity in patients with extended hippocampal system damage revealed by 3 convergent methods. Proceedings of the National Academy of Sciences, USA 106, 5442–5447.
Wheeler, M.A., Stuss, D.T., Tulving, E., 1997. Towards a theory of episodic memory: the frontal lobes and autonoetic consciousness. Psychological Bulletin 121, 331–354.
Yonelinas, A.P., Aly, M., Wang, W.-C., Kohen, J.D., 2010. Recollection and familiarity: examining controversial assumptions and new directions. Hippocampus 20, 1178–1194.

Epistemic Communities and Collaborative Research

Morgan Meyer, Agro ParisTech, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses epistemic communities and collaborative research. The first part reviews the genealogy and the definition of the concept of epistemic community and, thereafter, the various developments and criticisms of the term. The second part looks at various forms of research collaborations between scientists, with stakeholders, and with laypersons. The two terms are abundant not only in capturing the ways in which actors work together in order to produce and publicize knowledge, but also to reveal how knowledge is entangled with politics, economics, and social issues.

Introduction

Today, decisions are increasingly based on expertise and scientific knowledge. When governments, for instance, take actions regarding health issues, telecommunications, the environment, or the economy, they usually call on experts to guide them and to provide a knowledge base for solving issues. But these experts are hardly ever just individual persons conveying their personal opinions. Experts are members of wider networks and communities, which are usually international and interdisciplinary in outlook. The Intergovernmental Panel on Climate Change or the Delors Committee, which wrote the blueprint for the European Economic and Monetary Union, are but two examples of such 'epistemic communities.' Epistemic communities play an increasingly important role in communicating knowledge, in defining international problems, and in shaping policies.

However, to clarify what we mean by communities, we need to focus not only on communities per se, but also on 'collaborative research.' The increase in collaborative research, and the shift from 'little science' to 'big science,' has been documented both in the history of science and through scientometric studies. (It is the science of measuring and analyzing science. See also the journal, *Scientometrics*, which has published numerous articles on collaboration.) In addition, there are now many cases in which we also observe research collaborations across and beyond communities (for example with industry). Such collaborations bring together not only researchers, but also stakeholders and even laypersons, and show that knowledge is frequently produced across institutional and cognitive boundaries.

Epistemic Communities: Genealogy and Definition of the Concept

The attention to epistemic communities has a relatively long tradition in the social sciences, predating the increasing prominence of the term since the 1990s. An early author concerned with both knowledge and the communal aspects of knowledge production and circulation is Ludwig Fleck via his notion of 'thought collectives' (1935). Fleck defined thought collectives as a "community of persons mutually exchanging ideas or maintaining intellectual interaction," and he used the term specifically to refer to scientists. In line with Fleck's view, Thomas Kuhn later coined the notion of 'paradigms' (Kuhn, 1962), arguing that any member of a scientific community needs to accept and work within a given paradigm. Paradigms – defined by Kuhn as "universally recognized scientific achievements that, for a time, provide model problems and solutions for a community of practitioners" – are most visibly revealed in the textbooks, lectures, and laboratory exercises of a scientific community. A paradigm defines both what counts as knowledge and who counts as a legitimate member of a knowledge community.

The first group of authors to use the term 'epistemic communities,' and develop it through the 1960s and 1970s, is Holzner and his coauthors (Holzner, 1968; Holzner and Salmon-Cox, 1977; Holzner and Marx, 1979). In their body of work, epistemic communities are conceptualized as communities for whom 'epistemic criteria' have primacy over other interests. What members of epistemic communities have in common is their faith in scientific methods to produce truth. While the influence of the work of Holzner and his coauthors has been relatively modest, the term 'epistemic community' was forcefully reintroduced in the early 1990s. Two authors from the global policy/relations tradition, Adler and Haas (Adler, 1992; Adler and Haas, 1992; Haas, 1992) paved the way for an increasing amount of studies on such communities. According to them, an epistemic community is "a network of professionals with recognised expertise and competence in a particular domain and an authoritative claim to policy-relevant knowledge" (Haas, 1992, p. 3). Epistemic communities, so the often-quoted definition reads, are defined as:

> (1) a shared set of normative and principled beliefs which provide a value-based rationale for the social action of community members; (2) shared causal beliefs, […] which then serve as the basis for elucidating the multiples linkages between possible policy actions and desired outcomes; (3) shared notions of validity – that is, intersubjective, internally defined criteria for weighting and validating knowledge in the domain of their expertise; and (4) a common policy enterprise – that is, a set of common practices associated with a set of problems to which their professional competence is directed, […].
>
> Haas, 1992, p. 3

This definition of epistemic communities builds on what both Fleck and Kuhn defined as 'thought collectives' and

'paradigms.' Haas and Adler's work on epistemic communities can thus be seen as a continuation of the tradition in the social sciences to provide a theoretical and analytical tool to understand collective knowledge production. However, the object of study has shifted: the key focus is no longer only on scientific communities alone, but on how experts influence policy decisions. An example discussed by Haas is the epistemic community that brought about policy change in the field of chlorofluorocarbons (CFCs): the Montreal Protocol which led to a phasing out of components containing CFCs, can arguably be attributed to the work and efforts of an international epistemic community.

Developments and Criticisms

Expanding Analytical Frameworks

Epistemic communities have initially been studied through their links with the world of policy-making: for example, effects of epistemic communities on, and implication with, waste management, environmental, or multinational policies (e.g., Verdun, 1999; Zito, 2001). Over the years, the framework for studies has expanded to include business enterprises (Håkanson, 2010; Gittelman, 2007), social movements (Dobusch and Quack, 2010), open source (Cohendet et al., 2001; Edwards, 2001), higher education (Choi, 2014), scientific disciplines (Ahlgren et al., 2013), and national/international security (Cross, 2011). The concept of epistemic communities is now used in a variety of academic fields: political science and international relations, economics, law, business studies, public administration, sociology, history, organization and management studies.

Epistemic communities are usually not limited to a specific location or site, but can consist of people in different countries and with various professional backgrounds. The usefulness of the notion is evident in the body of work that has drawn on the framework set out by Haas and Adler. By focusing their research on epistemic communities, scholars are thus able to provide an analysis that connects knowledge, collectivity, and policy. These interconnections are studied at various scales, times, and in different political and cultural contexts. Another advantage is that the various dynamics and effects of epistemic communities are accounted for not only dealing with knowledge creation and propagation but also concerned with the consequent decisions and actions, as the knowledge they produce and share usually leads to policy actions and institutional change.

Criticisms of the Concept of Epistemic Communities

Nonetheless, several points of criticisms have been raised about the definitions of epistemic communities.

- First, a too narrow focus on the knowledge and characteristics that epistemic communities share might overstate the formation of consensus and collective knowledge. Conversely, some definitions may not reflect the disagreements, tensions, and uncertainties that are evident among communities.
- Second, the assumed impact of epistemic communities and their ability in solving problems in the policy domain has often been taken for granted, and is sometimes not looked into in depth. It must be stressed that governments will hardly follow epistemic communities' claims if they diverge too much from their own interests. A precondition for the knowledge claims of epistemic communities is that policy actors are willing to listen to, accept, and act upon their knowledge.
- Third, the concept has above all been used to discuss scientists and scientific knowledge. But there are many other actors, such as diplomats or lawyers, who can form epistemic communities that are not necessarily based on scientific knowledge. Even though Adler and Haas explicitly state that the 'epistemic' in epistemic communities should not only be limited to scientific knowledge, most studies have nevertheless done so.
- Fourth, the formation and constitution of epistemic communities has not been analyzed much. Epistemic communities seldom just pop up. They usually form in response to a problem or controversy that allows them to be established, funded, and given the mission and legitimacy to produce and publish knowledge. In addition, the dissolution of epistemic communities needs to be addressed – whether it is due to their fragmentation, the fact that the policy issue they were meant to address was resolved, or due to other reasons.

The above points should be addressed – and have been discussed to some extent – in order to consolidate, broaden, and rethink the concept of epistemic communities. In particular, the emphasis on policy relevance and policy impact which has dominated the literature on epistemic communities should be reconciled with more anthropological and sociological readings of the term. Epistemic communities, in this view, provide a rich topic to examine: how communities produce and act with knowledge (i.e., how they communicate, distribute, accumulate, and store knowledge); how communities are constituted and stabilized; how they change over time and articulate their own histories and futures; how they differ in terms of their policy effects and in terms of the strength of their internal ties; and how they shape, demarcate, and negotiate the identities of their members and their collective identity more generally (see Meyer and Molyneux-Hodgson, 2010).

Collaborative Research

While collaborative research, unlike 'epistemic communities,' has not been extensively theorized and defined, examples of collaborative research nonetheless proliferate. Collaborative research can be understood along a continuum of different forms and formats. On the one side, there are research efforts that bring together scientists among the same scientific discipline. On the other side, there are research efforts that bring together scientists and laypeople which is where collaborative efforts cut across the boundaries of academic institutions. Between these two poles, we can observe collaborations involving researchers with various disciplinary backgrounds, for example, in interdisciplinary and multidisciplinary research endeavors, or involving researchers and other actors, e.g., medical practitioners, decision-makers, or industrial partners.

Collaborations between Scientists

Collaborations raise several important questions, including the trust and tensions between collaborators, the required efforts for coordinating and aligning work, and its effects on scientific productivity (see for instance, the special issue on scientific collaboration in the journal, *Social Studies of Science* (2005: 35 (5)); Davenport et al., 1998; and Shrum et al., 2007). One could argue, of course, that there is no such a thing as 'noncollaborative' research. Research is by its nature a collective and social practice. Whether doing or supervising a PhD, participating in a research project, writing and reviewing academic papers, authoring texts and reports, building and using equipment, or exchanging ideas and criticisms, researchers hardly work on their own when carrying out their activities. Having said this, it is necessary to stress the collaborative nature of knowledge production, particularly in scientific fields in which collaborative research is common practice (e.g., physics or biology) as opposed to fields where it is much rarer such as the humanities. One example, in the field of particle physics, is the European Organization for Nuclear Research (CERN) based in Geneva, Switzerland. Involving around 30 countries, several thousand employees, 600 universities and research facilities, the CERN is arguably the most iconic example of a research collaboration. Research papers from the CERN usually have several hundreds of authors. The current record of number of authors of a single research paper is just above 3000. Other large-scale collaborations include: the Human Genome Project, an international project completed in 2003 to sequence the human genome; the Human Brain Project, a project launched in 2013 to bring together the neurosciences, medicine, and computer science to better understand the human brain; and the numerous telescopes built in the desert of Atacama in Chile and financed by various countries since the 1970s.

A frequently shared assumption is that collaborations such as the Human Genome Project or the various experiments carried out at CERN are essentially unbiased and politically neutral. Especially their international setup and the large cohort of collaborating institutions and universities are seen as proof that science knowledge production is neutral. Yet, it is not only the complexity of experimental procedures, the interdisciplinary nature of scientific problems, and the high cost of some technical equipment that call for collaborative work. Numerous studies have shown that collaborations are inevitably shaped by wider historical, social, political, economic, and/or military issues (e.g., Parker et al., 2012; Knorr-Cetina, 1999). The Manhattan project, responsible for developing the atomic bomb during World War II, is an illustration of the entanglements of researchers and the military complex. The Human Genome Project raised economic questions (as it was rivaled by a similar project led by a private company) and called for the need to reflect on the ethical, legal, and social issues inherent in the project. Moreover, studies of the field of high-energy physics show that gender and power issues are frequent among science communities (Traweek, 1992). In general, debates and controversies about power, authorship, and ownership are frequent in large collaborations, as are disputes about the primacy and prestige about some scientific fields over others.

Collaborations with Laypersons

There are many examples of collaborations between amateurs and professional scientists: in the fields of natural history, astronomy, epidemiology, medicine, biology, environmental science, or informatics. In natural history, for instance, there is a long tradition of amateurs doing fieldwork and producing knowledge together with professionals (Star and Griesemer, 1989). These amateurs do often collaborate to produce maps, to feed databases, or to coorganize educational or museological activities. In the fields of medicine and epidemiology, there are also a number of examples: AIDS treatment activists have been involved in the production and evaluation of biomedical knowledge (Epstein, 1996), and there are patient associations that are actively engaged in scientific research and collaboration with professionals (Callon and Rabeharisoa, 2003). Thus, these actors challenge the assumption that only scientists or experts can engage in research practices and produce reliable knowledge. It shows that there are indeed areas in which the production of knowledge is based on collaborative work that spans across the boundaries of traditional knowledge communities, and that users, amateurs, or patients can play an active role. While this form of collaboration is sometimes called 'collaborative research,' there is a plethora of other terms used in the academic literature such as 'co-production,' 'participatory research,' or 'community-based research.'

The benefits of such collaborative research efforts are epistemic. Science can become more accessible and democratic by involving target groups or laypeople, and researchers have access to much bigger data sets and computer power. On the other hand, there are also nonepistemic driving forces. In the field of ecology and natural history, the ultimate aim is usually a better understanding and protection of the natural world. In epidemiology and medicine, laypeople often try to change research priorities and focus on health problems that are not funded and not visible in the public domain, and they sometimes go against established norms. As for the do-it-yourself biology movement or the hacker movement, they are providing alternative, experimental spaces accessible to various people and interests, they are developing workarounds around standardized technical equipment and protocols, and they are portraying themselves as an alternative to mainstream science. All in all, collaborative research involving laypeople seeks to address, in some form or another, ecological, health-related, or political issues that mainstream academic science does not or cannot deal with.

Collaborations with Stakeholders

As with laypeople, there are a number of examples of collaborative research efforts where researchers cooperate with various stakeholders, which include collaborations with decision-makers or collaborations with industries (such as the computer or the nuclear industry). Collaborations among researchers and medical practitioners, in particular, have become prominent in recent years. Translational medicine, a well-documented example, has been hailed as a new 'paradigm,' and one of the most important developments in medicine since the beginning of the twenty-first century. The essential idea of translational medicine is to create links

between basic research and clinical research, between researchers working in a laboratory and patients and doctors in a hospital (see Martin et al., 2008 and the special issue of the journal, *Tecnoscienzia* on the topic (2014: 5 (1))). The common slogan is to move 'from bench to bedside' because current links between laboratories and hospitals are deemed problematic. The links are seen as too time-consuming (it takes many years from the discovery of a new molecule or therapy to its use), too difficult (several disciplines, institutions, and professions need to collaborate and coordinate their efforts), and too expensive (the development of a molecule costs millions of euros and only about 5% of new molecules eventually become marketable products.) This is why political and academic actors are actively fostering translational medicine by establishing centers of translational research to help scientists collaborate with doctors and to translate knowledge from the laboratory into applicable innovations and products. An institutionalization and formalization of new kinds of collaborations is currently under way via the creation of new institutions, journals, networks, posts, and platforms dedicated to translational medicine.

This does not mean, however, that collaborations between researchers and hospitals now happen without any friction. Collaborations often reveal important discontinuities – which can be disciplinary, organizational, linguistic, cultural, ethical, and political – between biology and medicine. Translational medicine is sometimes described as being 'lost in translation' due to several barriers: the difficulty of moving from a model (animal, *in vitro*) to humans; the fact that diseases are not simple, but heterogeneous and complex objects; the presence of different ethical and practical norms; language and cultural barriers; the incompatibility of databases and samples; the lack of qualified people and of support. Research laboratories and hospitals are spaces with their own temporality and sociality, and they embody different regimes of knowledge production and knowledge use. The same counts for the case of partnerships between universities and industry, and between universities and policy-makers: the temporal, social, and cultural differences between these worlds might pose a challenge for collaborative research. In the case of university–industry linkages, issues about intellectual property, secrecy, patents, and academic freedom become particularly salient (a topic frequently covered in the journal, *Research Policy*). As for collaborations between universities and policy-makers, debates often focus on the applicability, usability, and relevance of academic knowledge and the position of experts (as opposed to scientists).

Concluding Remarks

In the social sciences, the term community has been used so ubiquitously that it almost went out of fashion. The term seems to have lost any precise meaning, and the amount of interdisciplinary and transdisciplinary research work has made the term 'discipline' seem old-fashioned. However, professionals and scientists still carry out research through (epistemic) communities, and despite collaborations across disciplines, this still has impacts on teaching, shaping identities, and organizing research work.

The success of the term epistemic community both as a theoretical framework, and as a tool for analyzing the relationships between knowledge and policy proves that the term community can be reinvigorated. Rather than doing away with the concept of community, the way forward is to raise new questions in order to examine the ways in which communities establish expertise, share beliefs, respond to policy issues, act on a global scale, and lead to change. As for the term collaborative research, it serves as a forceful reminder that research work is, by essence, a collaborative endeavor and that the variety of formats and practices of collaboration is the norm, and not the exception in science. And while the rise, diversity and challenges of collaborative research has now been relatively well documented, there still are exciting research avenues calling for further academic studies. The role of new collaborative tools such as crowdfunding, online file sharing systems, wikis, and communication technologies in shaping knowledge production are only beginning to be examined. Moreover, new opportunities have arisen concerning collaborations between the social sciences and the natural sciences. The benefits and drawbacks of being an embedded social scientist – for instance, in the fields of nanotechnology and synthetic biology – call for more academic work. (Embedded social scientists do not only analyze the work done by natural scientists but they closely work with them: they often share research funding, become integrated in interdisciplinary research teams, and are responsible to address the social and ethical issues of scientific research.) The recent Manifesto for Experimental Collaborations between Social and Natural Scientists (Balmer et al., 2012) proves that collaboration continues to be a challenging topic.

The two terms reviewed in this article are effective tools to not only capture the ways in which actors work together to produce and publicize knowledge, but also to reveal how knowledge is entangled with politics, economics, and social issues.

See also: Communities of Practice; Community Sociology; Cooperation: Sociological Aspects; Science, Public Engagement with; Science, Sociology of.

Bibliography

Adler, E., 1992. The emergence of cooperation: national epistemic communities and the international evolution of the idea of nuclear arms control. International Organization 46 (1), 101–145.

Adler, E., Haas, P.M., 1992. Conclusion: epistemic communities, world order, and the creation of a reflective research program. International Organization 46 (1), 367–390.

Ahlgren, P., Persson, O., Tijssen, R., 2013. Geographical distance in bibliometric relations within epistemic communities. Scientometrics 95 (2), 771–784.

Balmer, A., Bulpin, K., Calvert, J., Kearnes, M., Mackenzie, A., Marris, C., Martin, P., Molyneux-Hodgson, S., Schyfter, P., 2012. Towards a Manifesto for Experimental Collaborations between Social and Natural Scientists. Accessible at: http://experimentalcollaborations.wordpress.com.

Callon, M., Rabeharisoa, V., 2003. Research "in the wild" and the shaping of new social identities. Technology in Society 25 (2), 193–204.

Choi, S., 2014. Meritocratic epistemic communities: an alternative policy paradigm for higher education. Higher Education Quarterly (online first).

Cohendet, P., Creplet, F., Dupouet, O., 2001. Organisational innovation, communities of practice and epistemic communities: the case of Linux. In: Kirman, A., Zimmermann, J.-B. (Eds.), Economics with Heterogeneous Interacting Agents. Springer, Berlin, pp. 303–326.

Cross, M.A.K.D., 2011. Security Integration in Europe: How Knowledge-Based Networks Are Transforming the European Union. University of Michigan Press, Ann Arbor.

Davenport, S., Davies, J., Grimes, C., 1998. Collaborative research programmes: building trust from difference. Technovation 19 (1), 31–40.

Dobusch, L., Quack, S., 2010. Epistemic communities and social movements: transnational dynamics in the case of creative commons. In: Djelic, M.-L., Quack, S. (Eds.), Transnational Communities: Shaping Global Economic Governance. Cambridge University Press, Cambridge, pp. 226–251.

Edwards, K., 2001. Epistemic communities, situated learning and open source software development. Epistemic Cultures and the Practice of Interdisciplinarity (working paper).

Epstein, S., 1996. Impure Science: AIDS, Activism, and the Politics of Knowledge. University of California Press, Oakland.

Fleck, L., 1935. The Genesis of a Scientific Fact. Benno Schwabe & Co, Basel.

Gittelman, M., 2007. Does geography matter for science-based firms? Epistemic communities and the geography of research and patenting in biotechnology. Organization Science 18 (4), 724–741.

Haas, P.M., 1992. Epistemic communities and international policy coordination. International Organization 46 (1), 1–35.

Håkanson, L., 2010. The firm as an epistemic community: the knowledge-based view revisited. Industrial and Corporate Change 19 (6), 1801–1828.

Holzner, B., 1968. Reality Construction in Society. Schenkman, Cambridge.

Holzner, B., Salmon-Cox, L., 1977. Conceptions of research and development for education in the United States. Annals of the American Academy of Political and Social Science 434, 88–100.

Holzner, B., Marx, J., 1979. Knowledge Affiliation: The Knowledge System in Society. Allyn and Bacon, Boston.

Knorr-Cetina, K., 1999. Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press, Boston.

Kuhn, T., 1962. The Structure of Scientific Revolutions. Chicago University Press, Chicago.

Martin, P., Brown, N., Kraft, A., 2008. From bedside to bench? Communities of promise, translational research and the making of blood stem cells. Science as Culture 17 (1), 29–41.

Meyer, M., Molyneux-Hodgson, S., 2010. Introduction: the dynamics of epistemic communities. Sociological Research Online 15 (2).

Parker, J.N., Penders, B., Vermeulen, N. (Eds.), 2012. Collaboration in the New Life Sciences. Ashgate Publishing, Farnham.

Shrum, W., Genuth, J., Chompalov, I. (Eds.), 2007. Structures of Scientific Collaboration. MIT Press, Boston.

Star, S.L., Griesemer, J.R., 1989. Institutional ecology, translations and boundary objects: amateurs and professionals in Berkeley's museum of Vertebrate zoology, 1907–39. Social Studies of Science 19 (3), 387–420.

Traweek, S., 1992. Beamtimes and Lifetimes: The World of High Energy Physicists, second ed. Harvard University Press, Boston.

Verdun, A., 1999. The role of the Delors committee in the creation of EMU: an epistemic community? Journal of European Public Policy 6 (2), 308–328.

Zito, A.R., 2001. Epistemic communities, collective entrepreneurship and European integration. Journal of European Public Policy 8 (4), 585–603.

Epistemic Communities and Collaborative Research, History of

Kenneth Brad Wray, State University of New York at Oswego, Oswego, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Studies of epistemic communities and collaborative research in the social sciences have deepened the understanding of how science works, and more specifically how the social dimensions of scientific practice both enable and impede social scientists in realizing their epistemic goals. Two types of studies of epistemic communities are distinguished; general theories of epistemic communities aim to construct accounts of theoretical change applicable to all social scientific specialties, whereas historical studies emphasize the contingencies that affect specific social scientific disciplines, subfields, or institutions. Although collaborative research plays a less pronounced role in the social sciences than in the natural sciences, collaboration has enabled social scientists to investigate topics that would otherwise be difficult to investigate.

The Structure of Epistemic Communities

Research Communities

An epistemic community is a network of knowledge-specific experts who are able to generate, and verify knowledge claims. Epistemic communities exist in the social, as well as the natural sciences. Epistemic communities and their structure affect social scientists' prospects of realizing their epistemic goals. The practices surrounding collaborative research influence how research is done as well as what research is pursued.

The most important type of epistemic community in the social sciences is the specialty community, the collection of social scientists working in the same subject area. Those in a specialty community are tied together by research interests, standards, practices, shared background assumptions, and often conceptual frameworks. The structure and development of epistemic communities in the social sciences has been studied mainly by sociologists.

In *Structure of Scientific Revolutions*, Thomas Kuhn (1962/2012) suggests that in mature scientific fields the only people qualified to evaluate research contributions are other members of the research community. By the time a field has reached maturity, the research literature is quite technical, and those contributing to the literature assume a high level of mastery of the subject matter. Kuhn's theory of scientific change was developed in response to his study of the history of the natural sciences. But it does seem that research in the social sciences is developed and written principally for other specialists. Consequently, it should not be surprising if outsiders, lay people and even scientists working in other social scientific fields, are unable to comprehend the latest research in a field. Some, however, claim that the social sciences differ fundamentally from the natural sciences in this respect, given that the objects of study in the social sciences are people and institutions that are familiar to the lay person. But this is misleading. Social scientists have developed and continue to develop technical vocabularies for the phenomena they study. And even when social scientists use familiar terms, they often use these terms in very precise and technical ways that may not reflect the lay person's understanding (see Durkheim, 1951: pp. 41–42).

Communication Networks

Research communities are not only arbiters of research quality, determining who gets funding and what research gets published, they also provide an infrastructure through which scientists are able to advance their research goals, reaching out to other researchers. Diana Crane (1972) argues that epistemic communities are united by 'invisible colleges,' informal communication networks facilitating the fast and efficient transfer of information across the research community. Not all researchers, however, are connected with the invisible college in their field. Crane found that those who are not well connected tend to be less productive than their peers who are connected.

There is some evidence that the communication networks in research communities are held together by the collaborative networks (Newman, 2001: p. 407). M.E.J. Newman found that in a number of research communities, between 80 and 90% of all authors are connected to each other through collaborative relationships. Some of the chains connecting any two scientists in such a network can be quite lengthy, but Newman suggests that these relationships probably facilitate communication in science.

'Epistemic Communities' as a Technical Term in International Relations

It is worth noting that the term 'epistemic communities' has acquired a very specific meaning in the subfield of international relations. It has become a technical term in this context. Some political scientists believe that the term is useful for picking out a new type of player or agent in international policy debates (see Haas, 1992). In this context, epistemic communities are groups of experts consulted by governments and other parties involved in political debates. The members of an epistemic community are bound together by "a shared set of normative and principled beliefs … shared causal beliefs … shared notions of validity … and … a common policy enterprise" (Haas, 1992: p. 3). These groups have "recognized expertise and competence in a particular domain and an authoritative claim to policy-relevant knowledge within that domain" (Haas, 1992: p. 3). And, other groups that are involved in such

debates aim to secure a specific outcome that supports the interests of their own group; whereas, when epistemic communities are concerned, that policy will be directed by knowledge. The point of invoking these new 'agents' in analyses of international policy debates is to account for how international policy decisions are actually made. "The epistemic communities approach focuses on [the] process through which consensus is reached within a given domain of expertise and through which the consensus knowledge is diffused … by other actors" (Haas, 1992: p. 23). The traditional list of agents in international policy debates – nation states, politicians, bureaucrats, and bureaucratic entities – provides an incomplete picture. Political scientists employing the term 'epistemic communities' in their analyses represent a particular theoretical orientation in the subfield of international relations. Their influence, although noteworthy, is relatively isolated.

Epistemic Communities: Social Sciences and Natural Sciences Compared

Consensus

Much of the understanding of epistemic communities in the social sciences is a result of comparisons with epistemic communities in the natural sciences. Kuhn (1962/2012) argues that in the natural sciences, unlike the social sciences, researchers agree upon the fundamentals. That is, researchers working in a particular field in the natural sciences accept the same theoretical framework and methodologies. Epistemic communities in the natural science are thus unified to a great extent. In the social sciences, on the other hand, researchers often work from competing theoretical frameworks. Originally, Kuhn expressed this point in terms of paradigms. The researchers in subfields in the natural sciences share a paradigm, whereas researchers in the social sciences lack paradigms. Although this generalization may not apply across all areas of research in the social sciences, there does appear to be far less consensus about core knowledge in the social sciences than in the natural sciences, where core knowledge is understood to be the settled accepted beliefs in a field.

Textbooks

The broad consensus that characterizes epistemic communities in the natural sciences is reflected in the content of textbooks. Competing textbooks in the natural sciences do not present competing theories. In the natural sciences, young scientists-in-training are taught the accepted theories only. The social sciences are different. In sociology, for example, it is not uncommon for an introductory textbook to present competing theoretical frameworks: the structural–functional paradigm; the social-conflict paradigm; and the symbolic-interaction paradigm (see, for example, Macionis, 1997; see also Lavirière et al., 2006: p. 521). Indeed, the same phenomenon, crime, for example, may be examined through the lenses of all three sociological paradigms. Young sociologists-in-training thus learn from the beginning of their training that their field is divided with respect to theoretical orientation. Stephen Cole (1992) notes some important differences with respect to the sources cited in textbooks in the natural sciences and in textbooks in the social sciences. First, in the natural sciences, textbooks tend to cite older research. Sociology textbooks, on the other hand, predominantly cite recent research. Second, natural science textbooks focus on the core knowledge in a field, whereas textbooks in the social sciences focus more on the research frontier. The research frontier is where scientists are currently conducting research, and where investigations are unresolved. Cole suggests that this difference between textbooks in the natural sciences and textbooks in the social sciences may be due to the fact that the core knowledge is smaller in the social sciences than in the natural sciences (see Cole, 1992: p. 131). He also found that, although the natural sciences and the social sciences differ with respect to core knowledge, there is little difference at the research frontier. Natural scientists and social scientists are equally divided about issues at the research frontier. This is evident from the fact that in both the social sciences and the natural sciences there is little agreement about the value or worthiness of grant proposals (Cole, 1992: p. 123).

Publications

There are also significant differences between social scientists and natural scientists with respect to publication patterns. In a recent study of the various publications in journals indexed in the *Web of Science*, Lavirière et al. (2006) found that 72% of the publications of natural scientists are articles. The remaining publications are principally meeting abstracts, letters, notes, and review articles. In contrast, only about 59% of the publications by social scientists are articles. And, less than 0.1% of the publications of natural scientists are book reviews, whereas 27% of the publications by social scientists are book reviews. Lavirière et al. (2006) take this last difference to indicate the relative importance of books in the social sciences. Although the scholarly book is now obsolete in the natural sciences, books are still an integral part of scholarship in the social sciences.

Collaborative Research

Collaborative Research in the Natural and Social Sciences

Collaborative research is a vital part of epistemic communities and can be carried out in different ways. Collaborative research is ubiquitous in the natural sciences. In some fields, more than 85% of the publications are multiauthored (see Zuckerman and Merton, 1973). Knorr-Cetina (1999) has compared the publication and research practices in two scientific cultures where collaboration is common, the culture of high energy physics (HEP) and the culture of molecular biology. Knorr-Cetina found that the two cultures differ significantly in important ways. Research in HEP is conducted at large research institutions, like the European Center for Nuclear Research, and such research is frequently reported in articles authored by many scientists. A single article may be coauthored by 50 or more scientists (see Thagard, 1999: p. 169). Research in molecular biology, on the other hand, is done by smaller teams, and as a result articles have far fewer coauthors.

Table 1 Percentage of articles that are coauthored, by decade and area

Years	*Natural sciences (%)*	*Social sciences (%)*	*Humanities (%)*
1920–29	49	6	1
1930–39	56	11	2
1940–49	66	16	1
1950–59	83	32	2

Adapted from Zuckerman, H., Merton, R.K., 1973. Age, aging, and age structure in science. In: Merton, R.K. (Ed.), The Sociology of Science: Theoretical and Empirical Investigations, Edited and with an Introduction by N.W. Storer. University of Chicago Press, Chicago, pp. 497–559.

The research cultures in the various social sciences also differ from each other. Although collaboration has become more popular in the social sciences, the single-authored article is still the norm in many fields. Zuckerman and Merton (1973) provide data on the differences between the percentage of papers that are multiauthored in the natural sciences, the social sciences, and the humanities. These data are drawn from a sample of key journals in various fields (adapted from Zuckerman and Merton (1973): p. 547, Table 7) (see **Table 1**).

The data in **Table 1** take one to the end of the 1950s. Lavirière et al. (2006) provide more recent data on the prevalence of collaborative research. Their data are from the following Thomson Scientific databases: the *Science Citation Index*, the *Social Sciences Citation Index*, and the *Arts & Humanities Citation Index* (Lavirière et al., 2006: p. 523). See **Table 2**.

Researchers in the social sciences collaborate less than researchers in the natural sciences, but significantly more than researchers in the humanities.

International Collaboration

It is also worth noting that social scientists collaborate with researchers from other countries at half the rate that natural scientists do. About 20% of all the articles published in the natural sciences involve international collaboration, whereas only about 10% of all the articles published in the social sciences involve international collaboration (Lavirière et al., 2006: p. 526). Levels of international collaboration also vary from country to country. In China, South Korea, and Belgium, for example, social scientists are about four times more likely to collaborate with scholars from other nations than are social scientists in the United States (Lavirière et al., 2006: p. 527; see the Footnote on page 525). In a study of articles published between 1984 and 1994 in 13 key social science journals, James Endersby reports that about 54% of articles are coauthored (see Endersby, 1996: p. 382, Note 7).

Table 2 Percentage of articles that are coauthored, by decade and area

Year	*Natural sciences (%)*	*Social sciences (%)*	*Humanities (%)*
1980	76.3	28.4	5.4
1985	80.3	38.4	6.5
1990	84.6	43.6	6.9
1995	87.5	48.1	8.2
2000	89.3	54.1	8.1

Endersby also found that "the mean number of authors per article is 1.78" (p. 382, Note 7). Thus, very few articles in the social sciences are produced by large research teams. In fact, only 1 of the 13 journals had more than 3% of its published articles authored by teams of four or more, the *Journal of Personality and Social Psychology*. Indeed, collaboration is more common in social psychology than in other social sciences (see Endersby, 1996: Table 3 on page 383). There have been a few studies of collaborative research in specific disciplines in the social sciences. John Hudson (1996), for example, examines the changing trends in collaborative research in economics, drawing data from eight of the leading English language economics journals. Hudson looks at data from 1950 to 1993. Hudson found that between 1950 and 1965, "the *highest* proportion of multi-authored articles for any of the eight journals was 15.6 percent" (p. 154; emphasis added). In contrast, between 1974 and 1993 the *lowest* proportion of multiauthored articles for any of the eight journals was 35.5% (p. 154). Clearly, there has been a significant shift in the research culture in economics.

Reasons for Increasing Collaboration

Hudson (1996) believes that three factors contributed to the increase in collaborative research between 1950 and 1993. First, Hudson notes that there was significant "growth in the subject areas covered by economics" (p. 155). Second, he notes that economics became increasingly more technical. These two considerations force economists to work collaboratively. Unless they work collaboratively, there are many things they would be unable to study effectively, or at least unable to study effectively in a timely fashion. The third factor Hudson identifies as contributing to the increase in collaborative research in economics is the sheer growth of the profession. There are now significantly more economists, and thus significantly more researchers with whom one can collaborate (p. 155). Knowing that one can draw on the skills and conceptual and material resources of one's colleagues may lead one to think of research projects that would otherwise be inconceivable, or at least impractical. This third factor, rather than forcing collaboration on economists, provides an incentive to collaborate. Hudson notes that collaborative research in economics is more popular in quantitative studies than in qualitative studies. In fact, in *Economic History Review*, a journal that publishes primarily qualitative studies, fewer than 15% of the papers are multiauthored (1996: pp. 155–156). Importantly, Hudson recognizes that there may in fact be costs to this change in research practices in economics. For example, working as part of a team, individual economists may have to suppress their more radical and creative ideas, thus stifling the creativity of the field as a whole (1996: p. 157). Economics is not the only discipline in the social sciences in which collaboration has been studied. James Moody (2004) examines the changing trends in collaborative research in publications indexed in *Sociological Abstracts* between 1975 and 1999. He found that there was an increase in coauthorship from 31% in the period from 1975 to 1985, to 38% for the period from 1989 to 1999 (Moody, 2004: p. 221). He also found an increase in the number of authors involved in coauthored papers from an average of 2.4 authors in the

period from 1975 to 1985, to an average of 2.7 authors in the period from 1989 to 1999. Like Hudson, Moody found that in those subfields where quantitative studies are more popular than qualitative studies, collaboration is also more popular (2004: p. 214). Collaborative research is more popular in the sociology of medicine, the sociology of the family, and social psychology, areas that draw on "large datasets and [that require] cumbersome analyses that lend themselves to a substantive division of labor" (p. 223). In contrast, there is little collaboration in political sociology, the sociology of knowledge, and the sociology of science, where sociologists seldom work with large data sets (p. 223). In fact, Moody found that "those writing quantitative papers are over 5 times more likely to have coauthors than someone who has not written quantitative work" (p. 226). Moody also found that the collaborative practices in sociology give rise to a complex collaboration network. This network "is *not* held together by a small number of network stars," nor is the network "composed of distinct, separate clusters" (p. 228). Instead, Moody discovered that the network in sociology connects "people … from widely different specialties" (p. 228). Moody suggests that, as a consequence, sociology as a discipline has "permeable theoretical boundaries" (p. 228). Concepts developed in one specialty are readily adopted in neighboring specialties. Paul Thagard (1999) notes that scientists collaborate for many different reasons: (1) collaborative research can aid a scientist in teaching or mentoring scientists-in-training; (2) collaboration enables one to get a research project done more quickly; (3) collaboration enables scientists to collect together a set of skills and knowledge no single individual scientist has; and (4) collaborative research enables scientists to draw on conceptual and methodological resources in numerous fields (see Thagard, 1999: p. 170). These same reasons probably influence social scientists in deciding whether to become involved in collaborative research projects.

Interdisciplinary Collaborations

Often collaborative research projects involve social scientists in the same field, each bringing to the research project a different set of skills and knowledge. But sometimes social scientists from different fields collaborate. For example, economists and psychologists sometimes collaborate (Kahneman et al., 1990 is an influential example of such a collaboration). Sometimes social scientists are involved in collaborative research projects with natural scientists or medical researchers. These sorts of collaborations lead to special challenges including finding or creating a vocabulary that both collaborators can comprehend. Trostle and Sommerfeld (1996) provide a summary of the various projects and challenges at the intersection of medical anthropology and epidemiology. What is striking about this body of research is that the social scientists and the medical scientists involved are principally collaborating on applied projects, specifically, projects related to public health. The possibility of such collaborations has not been left to chance. In fact, by 1996 "at least 19 anthropologists [were] employed at the Centers for Disease Control and Prevention (CDC) … the central epidemiological research agency for the US federal government" (p. 255). Trostle and Sommerfeld believe that this body of research shows that the borders between disciplines are permeable (1996: p. 254). Importantly, some of this research helps scientists better understand key assumptions in their research. For example, "anthropologists working in epidemiological settings have published important critiques of racial classifications within public health statistics" (p. 259). Clearly, there are opportunities for social scientists to make similar contributions when they collaborate with scientists in other fields as well.

General Theories of Conceptual Innovation and Change in the Social Sciences

In addition to studying the structure of research communities, sociologists of science have developed theories of conceptual innovation and change in the social sciences. Randall Collins (1998) has developed a general theory of intellectual change in which he argues that "intellectual creativity is concentrated in chains of personal contacts, passing emotional energy and cultural capital from generation to generation" (p. 379). He also argues that "creativity moves by oppositions" (p. 379). Specifically, Collins claims that "new ideas unfold by negating the major points of rival positions in a shared topic" (p. 379). He also notes that there is a limit to how many competing schools of thought can exist at one time. His careful study of a number of examples of intellectual change drawn from history reveals that between three and six competing schools are sustainable at any one time in a field (p. 380). Collins intends his theory of intellectual change to apply to many different intellectual fields, and across cultures, and he provides evidence in support of his account from China, India, Japan, and Europe.

Andrew Abbott (2001) has developed an alternative account of conceptual change in the social sciences. Like Collins, Abbott believes there is an inner logic that leads to developments in the social sciences, a logic that is not found in the natural sciences. According to Abbott, new research areas and methodological approaches are created in the social sciences through a fractallike process. Sociology, for example, is often dominated by either a consensus paradigm or a conflict paradigm (pp. 17–18). A consensus paradigm takes consensus as the natural state of human groups, and the problem is to explain how conflicts arise. A conflict paradigm, on the other hand, assumes that conflict is the natural state of human groups, and the problem is to explain how consensus is achieved. According to Abbott, when the consensus paradigm is in ascendancy, inevitably some of the issues that the conflict paradigm treats adequately will be taken up by sociologists working in the consensus paradigm. Thus, epistemic communities are divided, with some researchers working in a thoroughgoing consensus paradigm, and others developing an alternative view that imports concepts and methods previously used by those working in a conflict paradigm. Social scientists are thus frequently reinventing old theories and concepts. But with each successive reinvention, social scientists are led to new insights.

Although their accounts of intellectual change differ in detail, both Collins and Abbott believe that a general account of intellectual change is possible. In this respect, their work is

continuous with traditional sociology of science, aspiring to develop a general theory.

Historical Studies of Epistemic Communities and Conceptual Change

Local and National Focus and Consumption

In contrast to the general theories mentioned above, historical studies of social sciences tend to emphasize unique contingencies that affect the development of disciplines and the institutions that support them. Such studies make no attempt to generalize from the particular cases they examine.

Historical studies of epistemic communities in the social sciences are an effective means to appreciate the contingency of the contemporary organization of the social sciences. Histories of specific institutions reveal even more contingency and variability. Michel Foucault's influential work reminds us of the complicated relationships between the social sciences and political power and social control (see, for example, Foucault, 1990). Social scientists are often called upon by others outside of the academic world for assistance, and this has shaped various epistemic communities as they respond to external pressures and incentives. The effective marketing of products, for example, draws on research in the psychology of perception. And the way social workers interact with the populations they serve is shaped by recent social scientific research. Immigrant women, for example, are treated differently as social scientists learn more about them (see Hacking, 1999, Chapters 4 and 5). Foucault suggests that because the knowledge generated by social scientists enables political states to have far-reaching control over their citizens' lives, social scientists can find themselves complicit in aiding political states in oppressing people. Clearly, this is not a problem unique to the social sciences. The workforce in the natural sciences has also been harnessed by those in power. Nuclear weapons, for example, were developed with the aid of physicists, and earlier developments in biological warfare were made possible with the assistance of chemists. Whether or not social scientists are as complicit as Foucault suggests, there is evidence that social scientists are more responsive to national needs than researchers in the natural sciences and are more inclined to publish for a national audience rather than an international audience. Like scholars in the humanities, social scientists "publish more often in their mother tongue and in journals with a more limited distribution" (Lavirière et al., 2006: p. 521). Thus, the research of social scientists is more often created for local consumption and is more often directed by local concerns than research in the natural sciences.

Methodologies

Epistemic communities are not immune to influence from the broader culture in which they work. For example, there is a long tradition in German-speaking countries of regarding the social sciences and the natural sciences as fundamentally different, distinguished by their different methodologies. The social sciences, unlike the natural sciences, are alleged to be interpretative enterprises and to employ hermeneutical methods, like the methods used in the humanities. Hermeneutical methods are concerned with getting at the meaning of human actions and institutions. A hand gesture can mean different things in different cultures. A particular gesture that might cause hostility in one culture may convey friendship in another culture. Max Weber, for example, holds this view (see Weber, 1977). In contrast, in the English-speaking world, researchers in the social sciences frequently aspire to emulate researchers in the natural sciences, employing similar methods, and having similar aims. In fact, this attitude is not unique to social scientists working in English-speaking countries. In his research on suicide, Durkheim looks for patterns in the statistical data across a range of European countries, much as many natural scientists look for patterns in the data they collect (see Durkheim, 1951).

Defining Disciplines

The situation is even more complex than this picture suggests. A single science can have different developmental trajectories in different cultures. Anthropology, for example, has developed in significantly different ways in the United States and the United Kingdom, largely a consequence of differences in the institutional organization of the various sciences in the two cultures. In the United States, anthropology is clearly identified with the social sciences, and archaeology is regarded as one of the four major subfields in anthropology. In contrast, in Europe, archaeology is more closely connected with the discipline of history (see Sharer and Ashmore, 2003: pp. 33, 60). Similarly, although the field of criminology is identified as part of sociology in the United States, in Europe it has affinities with the fields of medicine, psychiatry, and law (see Akers, 1992: p. 5).

Broader social factors also sometimes influence the development of a field in specific national settings. For example, in the United States, the field of criminology became less academic and more concerned with vocational training after "the 1967 President's Commission on Law Enforcement and the Administration of Justice" (Akers, 1992: p. 7). The Commission recommended "enhancing higher education for criminal justice personnel," which led to a great expansion of undergraduate programs in criminology (p. 7). Similar stories could be told about other disciplines (see Patterson, 2001: p. 4, for a discussion of how anthropologists in the United States came to study acculturation and assimilation as immigration was increasing).

Summary

Research on epistemic communities and collaboration in the social sciences has deepened the understanding of the challenges social scientists face when they investigate the social world. But there is not a widely accepted unified theory of how social scientists realize their epistemic goals. This is in part a function of the fact that different disciplines, fields, and subfields have different norms and practices, and thus generate and verify knowledge through different means. As collaboration becomes more common in the social sciences, it is likely that social scientists will continue to investigate the benefits and drawbacks of working collaboratively. Although

the pooling of resources afforded by collaboration enables social scientists to investigate things they might not otherwise be able to investigate, new challenges arise as social scientists need to develop the social skills to work in collaborative research settings.

See also: Academic Research and Employment: Recent Changes in Europe and the United States; Epistemic Communities and Collaborative Research; Evolutionary Epistemology; Sociology, Epistemology of.

Bibliography

Abbott, A., 2001. Chaos of Disciplines. University of Chicago Press, Chicago.

Akers, R.L., 1992. Linking sociology and its specialties: the case of criminology. Social Forces 71 (1), 1–16.

Cole, S., 1992. Making Science: Between Nature and Society. Harvard University Press, Cambridge, MA.

Collins, R., 1998. The Sociology of Philosophies: A Global Theory of Intellectual Change. Belknap Press, Cambridge, MA.

Crane, D., 1972. Invisible Colleges: Diffusion of Knowledge in Scientific Communities. University of Chicago Press, Chicago.

Durkheim, E., 1951. Suicide: A Study in Sociology (J.A. Spaulding, G. Simpson, Trans.). The Free Press, New York.

Endersby, J.W., 1996. Collaborative research in the social sciences: multiple authorship and publication credit. Social Science Quarterly 77 (2), 375–392.

Foucault, M., 1990. The History of Sexuality: A Introduction, vol. 1. Vintage Books, New York.

Haas, P.M., 1992. Introduction: epistemic communities and international policy coordination. International Organization 46 (1), 1–35.

Hacking, I., 1999. The Social Construction of What? Harvard University Press, Cambridge MA.

Hudson, J., 1996. Trends in multi-authored papers in economics. Journal of Economic Perspectives 10 (3), 153–158.

Kahneman, D., Knetsch, J.L., Thaler, R.H., 1990. Experimental tests of the endowment effect and the Coase Theorem. Journal of Political Economy 98 (6), 1325–1348.

Knorr-Cetina, K., 1999. Epistemic Communities: How the Sciences Make Knowledge. Harvard University Press, Cambridge, MA.

Kuhn, T.S., 1962/2012. Structure of Scientific Revolutions, fourth ed. University of Chicago Press, Chicago.

Lavirière, V., Gingras, Y., Archambault, É., 2006. Canadian collaboration networks: a comparative analysis of the natural sciences, social sciences and the humanities. Scientometrics 68 (3), 519–533.

Macionis, J.J., 1997. Sociology, sixth ed. Prentice Hall, Upper Saddle River, NJ.

Moody, J., 2004. The structure of a social science collaboration network: disciplinary cohesion from 1963 to 1999. American Sociological Review 69, 213–238.

Newman, M.E.J., 2001. The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences of the United States of America 98 (2), 404–409.

Patterson, T.C., 2001. A Social History of Anthropology in the United States. Berg, Oxford.

Sharer, R.J., Ashmore, W., 2003. Archaeology: Discovering Our Past, third ed. McGraw-Hill Higher Education, New York.

Thagard, P., 1999. How Scientists Explain Disease. Princeton University Press, Princeton.

Trostle, J.A., Sommerfeld, J., 1996. Medical anthropology and epidemiology. Annual Review of Anthropology 25, 253–274.

Weber, M., 1977. The Methodology of the Social Sciences (E.A. Shils, H.A. Finch, Trans.). The Free Press, New York.

Zuckerman, H., Merton, R.K., 1973. Age, aging, and age structure in science. In: Merton, R.K. (Ed.), The Sociology of Science: Theoretical and Empirical Investigations, Edited and with an Introduction by N.W. Storer. University of Chicago Press, Chicago, pp. 497–559.

Epistemic Cultures

Karin K Cetina, University of Chicago, Chicago, IL, USA
Werner Reichmann, University of Konstanz, Konstanz, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Epistemic cultures are cultures of creating and warranting knowledge; 'epistemic' refers to knowledge and truth-related goals and practices as central elements of scientific and professional fields. The notion epistemic culture challenges the idea of the epistemic unity of the sciences and suggests that the sciences are in fact differentiated into cultures of knowledge, which are characteristics of scientific fields or research areas, each reflecting a diverse array of practices and preferences coexisting under the blanket notion of science. This article traces the history of the concept, its various meanings, original research that supports the idea, and recent extensions to areas of practical relevance (such as learning and policy making), to the humanities, finance, and other fields. It shows how the notion has been used to explain the difficulties of interdisciplinary research. The article also discusses the uses of the idea of 'knowledge cultures' on a macrolevel to compare and assess attitudes toward nonknowledge and epistemic mentalities inherent in the knowledge-related institutions and policies of particular countries.

What Are Epistemic Cultures?

The term *epistemic cultures* grows out of the so-called laboratory studies of science and technology, a direction of research that focuses on the practice of science as distinct from the history of scientific ideas, the logic of scientific discovery, or the analysis of scientists' writings. The process of doing science tends to become black boxed once its outcomes are accepted as 'facts' or as technologies in use. In the early 1970s, lab studies emerged as a strategy to open up this process by watching scientists at work. The early laboratory studies attempted to work out how scientists accomplish and create their findings by focusing on a particular lab and field and by choosing an observational methodology (Knorr Cetina, 1981; Latour and Woolgar, 1979; Lynch, 1985). Some also focused on questions of organization and learning (Traweek, 1988). Second-generation studies targeting scientific practice often had broader goals and more diverse methodologies. The term epistemic cultures entered the vocabulary via Knorr Cetina's comparative study of two influential scientific fields, high-energy physics (HEP) and molecular biology (1991, 1999). The question that motivated a comparative approach was could the findings of early laboratory studies be generalized across scientific fields? Could one extrapolate from the kind of work conducted in the benchwork labs that had originally been studied, to other sciences: for instance, those that involved both huge detectors and equally huge collaborations? The notion 'epistemic cultures' is in a sense the answer to this question. It disrupts the idea of the epistemic unity of the sciences and suggests that the sciences are in fact differentiated into cultures of knowledge that are characteristic of scientific fields or research areas, each reflecting a diverse array of practices and preferences coexisting under the blanket notion of science.

Knorr Cetina defined epistemic cultures as cultures of creating and warranting knowledge (1999: p. 1); and as amalgams of arrangements of mechanisms and elements bound together by affinity, necessity, and historical coincidence which, in a given field, make up how we know what we know. The word 'epistemic' is significant here. It points to knowledge and truth-related goals and practices as the central elements of scientific fields. When we study the epistemic side of science we do not focus on laboratories as organizational units, on scientists as workers, or on scientific disciplines as professions – we focus on how what counts as knowledge and technology is accomplished in designated settings through specific strategies that generate, validate, and communicate scientific accomplishments. When you study epistemic cultures you are more likely to investigate how practitioners distinguish signal from noise, or move from *in vitro* experimentation to *in silico* simulation, or how they decide which figure to trust when experimental outcomes are uncertain, than on these practitioners' budgets, career moves, or conflicts with management. In a sense, you step through the looking glass – everything is the same, budgets, careers, and managers still matter, yet everything is different. The setting is now a new world defined by research in which much of what we focus on when we study organizations and professions gets bracketed and pushed into the background, and priorities shift to needs and problems that arise with objects and equipment in laboratory life. Experimental high-energy physicists somewhat dismissingly distinguish between 'the system' and 'the science' or 'the experiment.' The system was the world of administration and financing which made what they did possible, but that was sharply differentiated from what they were possessed by and cared for, namely the science. Practitioners themselves draw a line between the epistemic and the nonresearch side of science, which they often try to avoid and keep at bay. We follow practitioners when we study epistemic cultures – but, as they do, we also need to understand and watch 'the system,' and how it enables and constrains the science.

The notion 'epistemic' targets a particular level of analysis. But what does the notion of 'culture' target? A first answer is that it targets a level of practice, rather than the level of symbols and meanings, which students of culture usually tend to examine. Practices are not devoid of meaning; to discover practices, it is necessary to gain familiarity with the frames of meaning within which people enact their lives, and symbolic

rituals or writings are equally practices. But to capture how scientists create knowledge, one also needs a working familiarity with what they do in laboratories – the spaces where most of the activities of the natural sciences take place – and with the way meaning is nested in and arises from use. For example high-energy physicists approached their detector, a massive machine, via metaphors of a living being. This was not *dissociated* from inquiry, but formulated their ways of experiencing the machine *during* inquiry. The focus on practice moves the level of cultural analysis 'down' to the realm of material regularities, but without losing sight of symbolic regularities and the ways these are associated with the material.

Material regularities, of course, include more than human practice, and this gets us to a second answer to the question of 'culture.' The comparison between different sciences brought to the fore distinctive characteristics of each field which included a wide range of elements; for example, distinctive traditions of teamwork and publication, specific epistemic strategies, different meanings of the empirical, and distinctive notions of reality as it was dealt with by the science. Different fields appeared to employ not only different vocabularies of knowledge and target different objects of study, but they saw different things: reality was an interiorized system of reference and not shared across disciplines. Science was not as unitary as had been thought; not just in relation to method (the Vienna Circle of philosophers has proposed the methodological unity of science) but also with respect to how reality itself was accessed and constructed. Economists, for example, think of the economy in terms of 'fundamentals,' which on a macroeconomic level are factors like interest rates, GDP growth, the trade balance and deficit, inflation levels, and so on. These factors indicate an underlying reality – the state and health of an economy – which macroeconomists tend not to pursue further. Real-time economic transactions and their meanings for people – in other words, everyday reality – is not the baseline from which such fields start. Rather one starts with 'measures,' variables with causal powers which describe the being of an economy. Different epistemic cultures locate such fundamental descriptors differently, although they remain aware of the basic structures and elements we distinguish in everyday life.

What the notion of epistemic culture implies therefore is that the disunity of science potentially extends to all levels, including the ontological. In fact, it is hard to see how it could be otherwise. Cultural specificities arise, one assumes, when domains of social life become separated from one another – when they curl up upon themselves and become self-referential systems that orient more to internal and previous system states than to the outside environment. The cultural scripts that emerge may be 'policed' by participants who may see to it that activities are plausible, effective, displayable as rational; scripts may also grow together in the sense of becoming mutually adapted over time, forced by the need to accomplish things in practice. All this may make cultural domains at least somewhat internally coherent (though certainly not free of conflict) – in addition to being separated and bounded. The ATLAS detector in HEP, consisting of layers of 'barrel' and 'endcap' instruments and a beampipe, requires that all its parts work together without detrimental interference – no cooperation, no results! Science and expertise are obvious candidates for cultural divisions; they are pursued by specialists separated off from other specialists by long training periods, intense divisions of labor, distinctive technological tools, particular financing sources, and the need to come up with results and display them in public through publishing.

The notion of an epistemic culture takes up where this assessment leaves off. It brings into focus the content of knowledge-oriented lifeworlds as lived, taken for granted worlds, including specific constructions of the referent (the objects of knowledge), distinctive models of the epistemic subject (scientists), and particular structures of existence with which a field engages. It is particular to these lifeworlds that they include 'strange' (e.g., nonhuman) entities with which some form of coordination has to be achieved. The notion of a scientific laboratory as distinct from the related idea of an organization may illustrate this. Laboratories are organized. But unlike the notion of 'complex organization' they are not about the coordination of human groups. Where knowledge is concerned, the notion of organization tends to assume away the rich processes and relationships between experts, organizational formats, and epistemic objects, through the simple presupposition of knowledge workers' technical competence. While the ways in which object-centeredness is enacted in organized laboratory settings is open-ended, the term itself points to 'mixed' systems of care and attention developing around epistemic entities in which organizational governance crucially works through the manipulation of problem content, rather than solely through people and structures.

The second answer to the question of what the notion of an epistemic *culture* targets, therefore, is that it brings into view not only the level of practice, but also a nexus of lifeworlds and the machineries of knowing that develop within a specialty. Epistemic cultures tend to have rich and potentially complex internal environments, with warped geometries that result from their turning or curving in upon themselves and a tendency to impose and expand their own structures and concerns. But they can be torn by conflicts and disagreements and the less than peaceful coexistence of participants (human and nonhuman). Machineries of knowing may or may not include a tactics of conflict prevention and management. To make a collaboration of several thousand physicists (and numerous detector parts) work for the 20–30 years *longue durée* of a HEP experiment, numerous tactics that enhance cooperation were developed. Such tactics may be lacking in fields where individual scientists pair up with individual topics of research. In the second case, we are likely to find mechanisms in place that discourage cooperation and justify its opposite.

Three Illustrations for Epistemic Cultures – HEP vs Microbiology

To illustrate how different epistemic cultures shape the production of knowledge, consider some distinctions between HEP and microbiology. First, the two fields produce two kinds of knowledge. HEP processes and deals with ephemeral,

'phantasmatic' objects – subatomic particles – which are too small and too fast to be observed directly. What HEP 'sees' are the traces these particles create in detectors. HEP thus depends on massive arrangements of technical instruments (detectors and accelerators), which are placed seven stories underground. Since these instruments are built afresh for every new experiment, creating, calibrating, protecting, and understanding this equipment takes up much of the time of an experiment. Knorr Cetina shows that the results emerging from such experiments involve, and may almost entirely consist of, "negative knowledge" that is not "non-knowledge but knowledge of the limits of knowing, of the mistakes we make in trying to know, of the things that interfere with our knowing, of what we are not interested in and do not really want to know" (1999: p. 64). Microbiology, in contrast, deals with natural and quasi-natural objects, accesses and manipulates them directly, and despite the existence of large genomic centers, still mostly involves benchwork in small labs. Physics collaborations today involve approximately 2000 physicists, while benchwork groups may range in size between a few and three or four dozen scientists. Microbiology seeks positive knowledge rather than negative knowledge, often accomplishing it through the 'blind variation' of experimental conditions (1999: p. 108).

This study also compared symbolic classifications in the two fields. Although physicists have an elaborate technical terminology at their disposal, they frequently discussed machine and particle behavior in terms of a symbolic vocabulary, using metaphors that described the detector as if it had a physiology and was a living being. For example, detectors could age and have various ailments and illnesses. Detectors also seemed to be moral and social beings. They could act up and behave 'well' or 'badly,' and various detector parts could communicate with each other and 'agree' or 'disagree.' They could 'cooperate' or not (Knorr Cetina, 1999: pp. 118–119). Machines then were categorized in terms of a vocabulary of organisms. Molecular biology, on the other hand, which dealt with organisms and organism's parts, resorted to technical images. A single cell organism or a laboratory mouse was seen as a production system that could be standardized, controlled, optimized, and reproduced.

The two fields also exhibited markedly different forms of social organization, and understandings of how laboratories and experiments should be managed. HEP required large collaborations that worked together as 'posttraditional communities.' They used collective and communitarian principles that configured the collaboration, rather than the individual, as the epistemic subject (Knorr Cetina, 1999: pp. 166–171). The simplest illustration of this – and one of the mechanisms for accomplishing a collective order that rejected hierarchy – was that hundreds of authors, essentially all accredited members of the collaboration, were listed alphabetically on publications. In other words, epistemic agency was delegated to the experiment as a whole; for example, conference invitations were extended to the experiment rather than to individual scientists. Delegating epistemic agency to the collective unleashed something that was urgently needed in this area: cooperation. In contrast, microbiological laboratories were organized around a laboratory leader, the most powerful individual in a hierarchical structure with a fixed position (the last) on publications. In contrast to HEP, cooperation was always tenuous within the individualistic and hierarchical structures of molecular biology labs; it required trust and continual renegotiation of interest (Knorr Cetina, 1999: p. 171).

Preceding Notions of Knowledge, Culture, and the Disunity of Sciences

Several early analysts, writing from the viewpoint of philosophy or sociology of knowledge, developed ideas about the situatedness of thinking and knowledge production within a particular society or community. These authors also often suggested that truth might not be as universal as it seemed, but is relative to what shapes a particular style of thinking. These arguments put into question the notion of linear progress in science. For example, Ludwik Fleck (1979), a biologist and doctor writing in the 1930s, developed the notion of a thought collective and of thought styles. Fleck proposed that scientists were in fact locked into a particular style of thinking that dominated in a thought collective: a scientific specialty that developed its own language and cognitions which were not understandable by other thought collectives. Fleck has often been credited with anticipating and preceding Kuhn's (1962) notion of the incommensurability of scientific paradigms, and the idea that scientific revolutions rather than linear progress describe and explain change in science. Mannheim (1990), writing in the 1950s as a sociologist, was interested in how social conditions shape scientific thinking. He emphasized the 'Seinsgebundenheit' of knowledge, using a notion that anticipates feminist notions of knowledge as gendered and inescapably perspectival (e.g., Hartsock, 1998).

Cultural differences between disciplinary cultures were also noted and elaborated. Snow (1959) famously argued that there is a cultural gap between the humanities and the natural sciences; and Lepenies (1985) saw the social sciences as wedged between these two groups of disciplines. Kuhn's notion of a scientific paradigm – what a scientific community shared in terms of methodological preferences, theory, and experimental practice – paved the way for a cultural approach that was less general than Snow's and more grounded in the empirical details of the internal processes of science. Snow's notion of the two cultures gained currency again through the so-called Science Wars in the 1990s, when combatants from the natural sciences and from sociology of science debated the question of relativism and what it meant to say that knowledge was constructed. In an attempt to bridge the divide that had ensued, Collins and Labinger (2001) brought together social and natural scientists, and found broad agreement between them in their dialogue about the legitimacy and authority of science.

Snow's ideas revived themes that were part of the Methodenstreit (struggle over methods) in German universities in the 1890s, and that were echoed again later when Habermas (1971) distinguished between instrumental and communicative knowledge during the debates over positivism in England and Germany. He linked the differences to human interests rather than culture.

The notion of epistemic culture contrasts with these debates in several ways. First, it is an empirically based concept, approach, and theory that requires close attention to knowledge production at the site of its respective practices. Second, it is not pitched at the level of whole disciplines but at the level of research fields, and it does not sit well with general dichotomies that group sets of disciplines together based on how they are named rather than what they do. It is entirely possible, from an epistemic culture point of view, that the differences between some disciplines are not as substantial as the differences between *subfields* of a single discipline; or that a natural science field appears closer to a humanities area than to another physics specialty. HEP can again serve as an example. Given that it works with traces of particles that are always historical (in the sense that they have always already decayed when they are 'seen' and processed), and since the movements of these particles have to be elaborately 'reconstructed' (native term), HEP appears to have more in common with semiotics than with a benchwork science. Although it does not use the same methods of reconstruction, results are often painfully indirect – in the sense of requiring multiple interpretations and many intermediary steps, deriving results from traces for which there is no original. When one studies experimental HEP it helps to have knowledge of historical processes, of semiotics, and of closed system theory. The same cannot be said about the type of molecular biology that is based on scientists' embodied work, requiring scientists to develop 'golden hands' (to become skilled experimenters) and a 'feeling for the organism' they handle (Fox Keller, 1983).

The Career of the Concept of Epistemic Cultures

The notion of epistemic cultures was widely received in social studies of science and in the sociology of scientific knowledge, as well as in social theory and in other fields (e.g., Burri, 2005; Cazaux, 2003; Cutcliffe, 2001; Markovsky, 2000; Mogey, 2001; Pinch, 2000). It motivated further studies that we group together as follows. First, authors used the notion to investigate in detail the epistemic characteristics of further specialties and fields of knowledge inside and outside academia. Second, several authors elaborated the notion of epistemic cultures and developed it further. Third, the notion was used to explain the difficulties of interdisciplinary research. Fourth, research on organizations used the concept to understand and optimize complex organizations. Fifth, there is a heterogeneous category of studies that investigates features of knowledge areas that can be understood in terms of the notion of epistemic cultures.

Analyses of Research Fields Using the Notion of Epistemic Cultures

Several authors enriched the notion of epistemic cultures by analyzing a variety of scientific disciplines, and by characterizing the respective fields and their truth-related strategies and practices. Some authors also applied the notion to nonscientific contexts, such as in online platforms and politics. For example, Yonay and Breslau (2006) studied the epistemic culture of 'mainstream economics' (p. 349), asking how economists connect their formal models, which are based on unrealistic assumptions, with empirical research and data-driven inquiry. They argue that their epistemic culture enables them to mediate between the two sides. "Rather than involving scientific instruments or data-gathering procedures, this mediation is accomplished by vaguely defined but generally accepted conventions regarding the movement from reality to models" (p. 377). The models construct an easy to handle and manipulate version of reality, an epistemic strategy familiar to both social sciences and natural sciences.

Along similar lines, Müller (1996) applied the concept of epistemic cultures to the social sciences and their 'modeling dilemma,' using it to formulate differences between practices of generating and validating knowledge. Reuter (2013) also applied the idea to the social sciences, calling for more reflexive intercultural research that makes practitioners aware of their own epistemic strategies because they may become problematic when other cultures are investigated. Grégoire Mallard, in an article based on ethnographic research with Balzac scholars (2005), highlights the presence of diverse epistemic cultures in literary criticism, a humanities field, by tracing the evolution of Editions of Balzac's *La Comédie Humaine*, from a unified unadorned text in the 1930s to a searchable electronic format of different versions. He shows how the different schools of interpretation in Balzac criticism (traditional, semioticians, sociocritics) constructed these diverse editing technologies to influence the evolution of literary theories. Kuoppamäki (2002) analyzed architectural models resulting from particular sketching techniques, which are part of a common 'configurational epistemic culture.' Architectural models or sketches cannot be true or false and it would not be appropriate to use these categories in this case. Accordingly, we are not seeing a 'truth machinery' at work, as in a natural science, but configurational strategies: the epistemic culture connects the architectural knowledge of many with the configurational sketch of the individual. Cole (2013) investigated the epistemic culture of forensic sciences calling it 'forensic culture,' and cultural patterns were also identified in mathematics. In her study about strategies of proofing and validating mathematical results, Heintz (2000a,b) shows that mathematics experiences epistemic cultural variations over time. But mathematics also illustrates the difficulties of sociological inquiries, given its low degree of interpretative flexibility and the high degree of consensus about the validity of epistemic strategies.

Epistemic Cultures in Nonacademic Contexts

As Epstein (2008) argued, the concepts and ideas of science and technology studies potentially migrate to other sociological fields and they may become useful to the study of nonacademic areas, particularly in a knowledge society in which knowledge processes become differentiated and permeate most areas of practice. Epistemic cultures will therefore be found in expert settings outside of the scientific laboratory. For example, Wagner and Newell (2004) explain the diverse uses of standardized software packages in universities in terms of the differences between the epistemic culture of the lower ranked members of the administration, the higher university management (the level of the university president or rector and the vice-chancellor), and the level of researchers.

Jensen (2007) used the idea of epistemic cultures to challenge traditional theories of learning. Assuming that epistemic cultures 'spill over' from academia to professionals, she seeks to identify professional cultures in terms of how they engage in continuous learning. Jensen compared nurses, accountants, teachers, and computer engineers. All groups use the rhetoric of the knowledge society, saying that they are lifelong learners. But how they access new knowledge varies considerably between these groups. The learning process is an 'unfolding process' characterized by a 'back and forth' trial and error – and by the interference of emotions, such as boredom or frustration. Jensen shows that epistemic cultural characteristics do not spill into nonacademic professions automatically, and recommends that the process should be supported by learning aids and tools that are adapted to professional cultures.

The policy-making area has also been studied with respect to its knowledge characteristics. In her study of a political campaign, Heath (2012) investigates the relationship between the dominating epistemic culture of experts and the public policy that ensues, finding that when knowledge production strategies change, public policy also changes. Heath examined the impact of quantitative studies that correlated "the growth of welfare, the rise in unwed childbearing, and the poverty" of individuals (p. 385), and claimed that children born from unwed couples are more likely to end in poverty. These findings informed public debates and lead to regulation promoting marriage. Heath argues that "the political efforts at establishing an epistemic culture of marriage promotion in Oklahoma built on a general 'knowledge culture' concerned with the significance of heterosexual marriage for sustaining the social fabric of American society" (p. 392).

One of the most knowledge-intensive areas is the area of finance, home to various lines of traditional financial research but also to the evaluative epistemic cultures of rating agencies (e.g., McKenzie, 2011), to information knowledge processes characterized by an epistemic of consumption (Knorr Cetina, 2010), and to technologies that raise their own questions of culture-formation and transformation (Knorr Cetina and Preda, 2007). A final example for a nonacademic epistemic culture is the widely known online encyclopedia Wikipedia. Wray (2009) critically compares the norms, collaboration-styles, and knowledge validating practices of Wikipedia authors with those of scientists, arguing that Wikipedia encourages a 'puckish' culture, in contrast to science where seriousness dominates. Wikipedia includes (more or less harmful) jokes, defamations, and anonymous contributions written to undermine its mission (e.g., fake biographies of celebrities) – hence the notion 'puckish' – while there is little place for humor and jokes in science.

Further Developments of the Notion Epistemic Culture

The last sections focused on how authors in various scientific, professional, and expert contexts have used the notion of epistemic culture. This section offers three examples of studies that expand the notion theoretically and develop it further. For example, Böschen et al. (2006) argue that there are not only different cultures of producing and validating positive knowledge but also 'cultures of nonknowledge.' These manifest different degrees of awareness of and attention to nonknowledge; and different ways of anticipating the unexpected, of associating it with risk, and of dealing with unforeseen and unintended results. Some fields such as ecology cherish a naturalistic view of the world in order to avoid the unintentional elimination of unrecognized key details of a phenomenon. Giere (2002) criticizes Knorr Cetina's "most provocative idea" (p. 639) of the elimination of the epistemic subject (see Knorr Cetina, 1999: pp. 159–191) in experimental HEP. He argues that this part of the epistemic culture of HEP could also be understood in terms of notions of distributed cognition: of expertise distributed among the hundreds of members of a HEP experiment and physics instruments. The more individualistic sciences would then simply lack this feature.

Finally, Lamont provides a nuanced account of 'epistemological styles' (Lamont, 2009: p. 54) prevalent in different social sciences and leading to different criteria of evaluation when practitioners evaluate grant proposals, that reflect different disciplinary cultures. Lamont distinguishes four styles of identifying and defining excellence in proposal evaluations, which she links to differences in their consensus orientation: comprehensive, constructivist, positivistic, and utilitarian. In her account, philosophy is a 'problematic' discipline whose weakness and strength is its preference for discussion and argument rather than quick decisions. English literature, in contrast is a demographically threatened low consensus discipline with intellectual disagreements at its core. History, like economics, is much more consensual. It defines excellence on the basis of good 'craftsmanship' in archival work. Both history and economics appear to have no problems in defining and recognizing excellence (p. 103).

Epistemic Cultures and Questions of Interdisciplinary

One major area of research in which the notion of epistemic cultures has served as a descriptive and explanatory concept is that of interdisciplinary or multidisciplinary research. What happens when different epistemic cultures get in touch with each other? How do members of different epistemic cultures cooperate with each other, and do they even understand each other? Conversely speaking, does the existence of epistemic cultures explain and predict some of the difficulties of interdisciplinary research? While the concept of epistemic cultures "is useful for highlighting the specificities of the different scientific cultures and research practices […]", Albert et al. argued (2008: p. 2522) that "it remains somehow static as it does not provide insight on the power relationships between groups of scientists." Differing epistemic cultures are likely to devalue other epistemic cultures, especially when they do not have direct contact with them (2008).

In line with this argument, the notion of epistemic cultures has been used to highlight and explain conflicts between fields, which may be due to incompatible practices of knowledge production. Mørk et al. (2008) suggested that such conflicts can limit the productivity of knowledge creation and make it more path dependent. This means that once knowledge production takes off in a particular direction, it is difficult to roll it back to the beginning and start over in a different direction. Van House (2004) studied a digital platform that

provides environmental data for members of different epistemic cultures, including scientists, laymen, and bureaucratic organizations. She finds that the online communication makes the differences between epistemic machineries visible, but it also "facilitates the flow of information […] among epistemic cultures" (Van House, 2004: p. 11).

Other authors do not focus on conflicts between epistemic cultures but analyze the change and development that occurs when members of epistemic cultures meet and have to confront their differences. Kastenhofer (2007), for example, shows how biotechnology emerged from a process of convergence between several fields with quite different epistemic cultures. Kastenhofer argues that such processes of convergence can encourage cooperation between disciplines, integrate research groups and organizations, and assimilate epistemic cultures. Jacobs (2013: Chapters 1 and 4) argues more broadly that interdisciplinary research, when it occurs, leads over time to new syntheses that create new specialties with their own research strategies, journals, and institutional strongholds. In other words, they develop their own specialized epistemic culture. Other studies find that processes of 'convergence' may also produce uncertainties and inconsistent knowledge claims. For example, Torgersen (2009) argues that biology became 'interdisciplinary' with the development of genomics (see also Dean, 1979), and had to rethink criteria of evidence, experimental practices, object-classifications, and the role of hypotheses in research. He also argues that, in the end, a new epistemic culture emerged, but not without a lengthy process of specifying and reconfiguring previous cultural practices. Kertcher (2010), studying electronically mediated collaborations, identifies problems of innovation adoption having to do with differences between epistemic cultures participating in the collaborations, but he also shows ways through which participants managed to build bridges between disciplines while attempting to achieve interdisciplinary knowledge integration.

Interventionist Studies: Epistemic Cultures in Organizations

Another body of literature deploys the notion epistemic culture to intervene in, manage, or change social domains. The studies typically focus on knowledge-intensive organizations, arguing that the notion can serve as a tool in organizational design and professional development. Management theories pay attention to epistemic cultures because the notion helped to develop their "sense of the glue that keeps tightly knit groups of researchers together as 'epistemic communities' or communities that are defined by their shared tacit knowledge of process, context, history, and culture […]." The central management question here is: "How open are these coherent groups to cooperation and partnership with others?" (Gore, 2011: p. 99). The literature suggests that different epistemic cultures – defined as sets of shared knowledge-oriented practices – characterize not only scientific research, but also organizations and strategy (Kaplan, 2011: p. 320), and it recommends a practice-oriented approach to understand these cultures (e.g., Clegg et al., 2011: p. 117).

Dragonetti (2003) argues that management theory accepts knowledge to be an important resource for corporations. Still, it does not ask how knowledge is produced within corporate contexts. Based on an ethnographic study of several companies, she shows that the process of knowledge creation in organizations departs from the linear models on which knowledge management relies. Such models are like straightforward recipes, when in fact the real processes of knowledge creation appear much more chaotic and ambiguous. Dragonetti distinguished two types of knowledge: traditional knowledge ('know-how') and declarative knowledge ('know-what'). 'Know-what' was less codified and included social routines; while 'know-how' circumscribed all work-related routines and was strongly codified. Know-what and know-how involved different knowledge creation processes. For example, when know-how was created the pressure to achieve consensus was high, resulting in a more hierarchical social organization. Dragonetti calls upon management models to come up with an empirically based 'knowledge theory' to better control knowledge production within corporations.

Gore (2011) provides another example of the use of the epistemic culture notion in organizational research. He studied how universities' knowledge management can encourage cooperation between groups that share an epistemic culture (e.g., academic researchers) with relevant groups in industry. Following Gore, the notion epistemic culture is not only an analytic concept that explains differences between groups of researchers, but also a starting point for developing strategies that overcome these differences and prompt people to cooperate. Moisander and Stenfors (2009) make similar suggestions. In their empirical study they discuss the challenges of different epistemic cultures, different kinds of knowledge, and conflicting conceptions of the subject matter of knowledge (p. 229). The conflicts divide management science scholars and practitioners who may be required to use the strategic tools developed by research. Managers in 'postbureaucratic organizations' exhibit different epistemic cultures than 'modernist' operation research and management science scholars (p. 234). For Weissenberger-Eibl and Ebert (2010) epistemic cultures are an organizational characteristic that captures how knowledge is treated within the organization. Since management should consider all accepted ways of using, identifying, storing, distributing, and validating knowledge, organizational epistemic cultures are a central aspect of the management of organizations. Another study analyzes how PowerPoint presentations are engaged in strategy making in corporations and organizations (Kaplan, 2011). She argues that PowerPoint is not just effective or ineffective; it should rather be seen as part of 'strategy-in-the-making' (Moisander and Stenfors, 2009), and as such it shapes organizational knowledge.

Macroepistemic Cultures/Knowledge Cultures

The concept of epistemic cultures can also be used on a macrolevel to characterize the knowledge attitudes, institutional arrangements, and knowledge policies of different societies. Knorr Cetina (2007) proposed the notions of knowledge cultures and macroepistemic cultures for this reason, to

differentiate this type of analysis from that of the truth-related strategies and categories which are central to understanding research cultures. We tend to see knowledge as a component of society: of its economic, social, and political life. But we can also turn the argument around and consider social, political, and economic life as part and parcel of a particular knowledge culture. The distinction is not simply between the knowledge processes and their cultural context. It is rather a distinction between the interiorized description and conceptualization of knowledge processes opened up through empirical inquiry (the epistemic cultures) *and the description of society in knowledge terms*. For example, national legal frameworks and research expenditures not only enable, restrict, and contextualize research, they also exemplify the knowledge culture of a country.

Jasanoff (2005) provides an example for different macro-level knowledge cultures. Jasanoff (2005: p. 151) studied questions of assisted reproduction, stem cells, genetically modified food, and the status of an embryo (whether a 'lump of cells' is a person, or a fetus, or just a lump of cells) by investigating the national interpretations of the US, Great Britain, and Germany. As governments and the public interpret through debate and regulation what biotechnology promises or threatens, they come to quite different conclusions depending on whether the procedure is perceived to result in boundary crossing 'monsters,' is linked to a disorder, or is simply seen as giving rise to new hybrid forms. The resulting legal frameworks (Germany forbids gestational surrogate mothers, for instance) reflect attitudes toward new knowledge with wide ranging institutional, financial, and epistemic consequences, for example for the type of research viable in a country.

Another example of a knowledge-related category pertaining to the wider society is the notion of a risk society: the idea that we are increasingly confronted with risks and uncertainties emanating from the very technological, scientific, and other achievements of modernity. The risks are seen as ontologically threatening and cannot simply be eliminated by more expert knowledge (e.g., Beck, 1992). From the present perspective, the description of Western societies as risk societies expresses and elaborates a particular cultural viewpoint on risks: an 'epistementality' that may shape the way expert knowledge is embedded in legal frameworks, schemes of citizen-participation, and the like (Gibbons et al., 1994).

See also: Culture and Actor Network Theory; Culture and Institutional Logics; Culture, Production of: Prospects for the Twenty-First Century; Knowledge Society, History of; Networks and Meaning; Organizations and Culture; Science and Technology Studies, History of.

Bibliography

Albert, Mathieu, Laberge, Suzanne, Hodges, Brian D., Regehr, Glenn, Lingard, Lorelei, 2008. Biomedical scientists' perception of the social sciences in health research. Social Science & Medicine 66 (12), 2520–2531.

Beck, Ulrich, 1992. From industrial society to risk society. Theory, Culture and Society 9, 97–123.

Böschen, Stefan, et al., 2006. Scientific cultures of non-knowledge in the controversy over genetically modified organisms (GMO) – the cases of molecular biology and ecology. GAIA – Ecological Perspectives for Science and Society 15 (4), 294–301.

Burri, Regula, 2005. Book review: wissenskulturen – ein Vergleich naturwissenschaftlicher Wissensformen by karin knorr cetina. Kölner Zeitschrift für Soziologie und Sozialpsychologie 57 (2), 371–373.

Cazaux, Geertrui, 2003. Review essay: constructing nature: on salmon, mushrooms, and epistemic cultures. Journal of Contemporary Ethnography 32 (5), 626–636.

Clegg, Stewart, Carter, Chris, Kornberger, Martin, Schweitzer, Jochen, 2011. Strategy: Theory & Practice. SAGE, London.

Cole, Simon A., 2013. Forensic culture as epistemic culture: the sociology of forensic culture. Studies in History and Philosophy of Biological and Biomedical Studies 44 (1), 36–46.

Collins, Harry, Labinger, Jay A. (Eds.), 2001. The One Culture? The University of Chicago Press, Chicago & London.

Cutcliffe, Stephen, 2001. Book review: epistemic cultures: how the sciences make knowledge by Karin Knorr Cetina. Science, Technology, & Human Values 26 (3), 390–393.

Dean, John, 1979. Controversy over classification: a case study from the history of botany. In: Barnes, Barry, Shapin, Steven (Eds.), Natural Order: Historical Studies of Scientific Culture. SAGE, Beverly Hills & London, pp. 211–230.

Dragonetti, Nicola C., 2003. Organisations as Epistemic Cultures. INSEAD Working Paper.

Epstein, Steven, 2008. Culture and science/technology: rethinking knowledge, power, materiality, and nature. Annals of the American Academy of Political and Social Science 619, 165–182.

Fleck, Ludwik, 1979. Genesis and Development of a Scientific Fact. The University of Chicago Press, Chicago & London.

Fox Keller, E., 1983. A Feeling of the Organism: The Life and Work of Barbara McClintock. Freeman, San Francisco, CA.

Gibbons, Michael, Limoges, Camille, Nowotny, Helga, Schwartzman, Simon, Scott, Peter, Trow, Martin, 1994. The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies. Sage, London.

Giere, Ronald N., 2002. Distributed cognition in epistemic cultures. Philosophy of Science 69 (4), 637–644.

Gore, Tim, 2011. Epistemic communities in universities. Procedia – Social and Behavioral Sciences 10, 98–103.

Habermas, Jürgen, 1971. Knowledge and Human Interests. Beacon Press, Boston.

Hartsock, Nancy, 1998. The Feminist Standpoint Revisited and Other Essays. Westview Press, Boulder, CO.

Heath, Melanie, 2012. Making marriage promotion into public policy: the epistemic culture of a statewide initiative. Qualitative Sociology 35 (4), 385–406.

Heintz, Bettina, 2000a. Die Innenwelt der Mathematik. Zur Kultur und Praxis einer beweisenden Disziplin. Springer, Wien.

Heintz, Bettina, 2000b. In der Mathematik ist ein Streit mit Sicherheit zu entscheiden: Perspektiven einer Soziologie der Mathematik. Zeitschrift für Soziologie 29 (5), 339–360.

Jacobs, Jerry, 2013. In Defense of Disciplines. Interdisciplinarity and Specialization in the Research University. The University of Chicago Press, Chicago.

Jasanoff, Sheila, 2005. In the democracies of DNA: ontological uncertainty and political order in three states. New Genetics and Society 24 (2), 1139–1155.

Jensen, Karen, 2007. The desire to learn: an analysis of knowledge-seeking practices among professionals. Oxford Review of Education 33 (4), 489–502.

Knorr Cetina, Karin, 1991. Epistemic cultures: forms of reason in science. History of Political Economy 23 (1), 105–122.

Knorr Cetina, Karin D., 1981. The Manufacture of Knowledge. Pergamon Press, Oxford.

Knorr Cetina, Karin, 1999. Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press, Cambridge & London.

Knorr Cetina, Karin, 2007. Culture in global knowledge societies: knowledge, cultures, and epistemic cultures. Interdisciplinary Science Reviews 32 (4), 361–375.

Knorr Cetina, Karin D., 2010. The epistemics of information: a logic of knowledge consumption. Journal of Consumer Culture 10 (2), 1–31.

Knorr Cetina, Karin D., Preda, Alex, 2007. The temporalization of financial markets: from network markets to flow markets. Theory, Culture and Society 24 (7–8), 123–145.

Kaplan, Sarah, 2011. Strategy and powerpoint: an inquiry into the epistemic culture and machinery of strategy making. Organization Science 22 (2), 320–346.

Kastenhofer, Karen, 2007. Converging epistemic cultures? Innovation: the European Journal of Social Sciences 20 (4), 359–373.

Kertcher, Zack, 2010. Gaps and bridges in interdisciplinary knowledge integration. In: Anandarajan, Asokan, Anandarajan, Murugan (Eds.), e-Research Collaboration – Theory, Techniques and Challenges. Springer, Berlin & Heidelberg, pp. 49–64.

Kuhn, Thomas S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.

Kuoppamäki, Riitta, 2002. On Architectural Sketching. DATUTOP 23.

Lamont, Michèle, 2009. How Professors Think. Inside the Curious World of Academic Judgement. Cambridge University Press, Cambridge.

Latour, Bruno, Woolgar, Steve, 1979. Laboratory Life: The Construction of Scientific Facts. Princeton University Press, Princeton, NJ.

Lepenies, Wolf, 1985. Die drei Kulturen: Soziologie zwischen Literatur und Wissenschaft. Hanser, München.

Lynch, Michael, 1985. Art and Artifact in Laboratory Science: A Study of Shop Work and Shop Talk in a Research Laboratory. Routledge & Kegan Paul, London.

Mallard, Grégoire, 2005. Interpreters of the literary canon and their technical instruments: the case of Balzac criticism. American Sociological Review 70 (6), 992–1010.

Mannheim, Karl, 1990. Ideologie und Utopie. Vittorio Klostermann, Frankfurt am Main.

Markovsky, Barry, 2000. Book review: Epistemic cultures: how the sciences make knowledge by Karin Knorr Cetina. Contemporary Sociology 29 (3), 556–557.

McKenzie, Donald, 2011. The credit crisis as a problem in the sociology of knowledge. American Journal of Sociology 116 (6), 1778–1841.

Mogey, John, 2001. Review essay: Sociology of knowledge: trust, truth, and troth. Social Inquiry 71 (3), 394–397.

Moisander, Johanna, Stenfors, Sari, 2009. Exploring the edges of theory-practice gap: epistemic cultures in strategy-tool development and use. Organization 16 (2), 227–247.

Mørk, Bjørn Erik, Aanestad, Margunn, Hanseth, Ole, Grisot, Miria, 2008. Conflicting epistemic cultures and obstacles for learning across communities of practice. Knowledge and Process Management 15 (1), 12–23.

Müller, Karl H., 1996. Epistemic cultures in the social sciences. The modeling dilemma dissolved. In: Hegselmann, Rainer, Müller, Ulrich, Troitzsch, Klaus G. (Eds.), Modeling and Simulation in the Social Sciences from the Philosophy of Science Point of View. Klüwer Academic Publishers, Dordrecht, pp. 29–63.

Pinch, Trevor, 2000. Book review: Epistemic cultures: how the sciences make knowledge by Karin Knorr Cetina. American Journal of Sociology 105 (5), 1472–1474.

Reuter, Julia, 2013. Doing science – doing culture. Interkulturelles Forschen als Gegenstand Interkulturereller Forschung. Interculture Journal – Online-Zeitschrift für interkulutrelle Studien 12 (20), 11–23.

Snow, Charles P., 1959. The Two Cultures and the Scientific Revolution. Cambridge University Press, Cambridge.

Torgersen, Helge, 2009. Fuzzy genes: epistemic tensions in genomics. Science as Culture 18 (1), 65–87.

Traweek, Sharon, 1988. Bedtimes and Lifetimes: The World of High Energy Physicists. Harvard University Press, Cambridge, MA.

Van House, Nancy, 2004. Epistemic machineries of environmental online communication. In: Sharl, Arno (Ed.), Environmental Online Communication. Springer, Berlin, pp. 199–208.

Wagner, Erica L., Newell, Sue, 2004. 'Best' for whom?: the tension between 'best practice' ERP packages and diverse epistemic cultures in a university context. The Journal of Strategic Information Systems 13 (4), 305–328.

Weissenberger-Eibl, Marion, Ebert, Dominik, 2010. Communities of practice: So gestalten Sie die Wissenskultur in Organisationen. Zeitschrift Führung + Organisation 79 (6), 360–366.

Wray, K. Brad, 2009. The epistemic cultures of science and Wikipedia: a comparison. Episteme 6 (1), 38–51.

Yonay, Yuval, Breslau, Daniel, 2006. Marketing models: the culture of mathematical economics. Sociological Forum 21 (3), 345–386.

Epistemic Doubt During Adolescence

Robert B Ricco, California State University, San Bernardino, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Naïve beliefs about the nature of knowledge and knowing – epistemic beliefs – are present from childhood onward and undergo development through a relatively fixed sequence of levels from realism through relativism to rationalism. There appears to be more than one pass through this sequence, with the first pass occurring from middle childhood through late adolescence, and the second pass occurring from late adolescence through early adulthood. The period of relativism during adolescence is characterized by generic doubt about the possibility of establishing a valid, rational basis for one's beliefs, coupled with interesting and varied ways of coping with that doubt.

Epistemology is a subdiscipline of philosophy that is concerned with the nature, origin, limits, and validation of human knowledge (Hofer, 2002). Several decades of research within the fields of psychology and education have made it clear that lay adults, adolescents, and older children possess an interrelated set of beliefs about knowing and knowledge, which might be called a naïve or informal theory of epistemology (Hofer, 2004; Kitchener, 2002). Further, there is substantial agreement that given certain social and educational experiences, these naïve epistemologies undergo development in accordance with a relatively fixed sequence of levels. Several developmental models have been proposed (e.g., Baxter-Magolda, 1992; Chandler et al., 2002; King and Kitchener, 1994; Kuhn et al., 2000b; Perry, 1970). There is strong agreement across these models concerning the beginning and end points of development and considerable agreement about a key intermediary level. At the same time, the models differ substantially with respect to timing and the proposed age of onset for the various levels.

The first models of the development of naïve epistemologies focused on late adolescence through early adulthood and were derived from cross-sectional and longitudinal data regarding change in epistemic beliefs across the college years and beyond. Perry (1970), King and Kitchener (1994), and Baxter-Magolda (2002) generated similar models covering this age period. The models implied that there was little significant development concerning epistemic beliefs prior to the later stages of high school. In the initial level or stage of these models, variously described as an absolutist (Baxter-Magolda, 2002; Kuhn, 2000), prereflective (King and Kitchener, 2002), realist (Chandler, 1987), or dualist (Perry, 1970) level, knowledge is considered to be a direct copy of objects and events in the world. Fully objective and devoid of subjective elements, knowledge, in this view, derives from experts and other authority figures or from direct personal experience. Facts originating in this manner are self-evident, beyond dispute, absolutely correct, and certain. Within this position, there is no obvious need for justification or validation of knowledge claims above and beyond a simple, and allegedly transparent, process of assembling the facts and determining which position in a dispute is supported by them. If the uncertainty of knowledge claims or the indeterminacy of problems and controversies are at all acknowledged within this initial epistemic level, it is merely in a temporal sense. That is, observations can be in error, either because of faulty instrumentation or methodology, or because the observer has only limited access to the facts, but further research – individual or collective – will eventually resolve disputes in all such cases.

The epistemic realism of the first stage of development is progressively undermined by the gradual appreciation of the subjective, interpretive, and perspectival nature of all knowing, with the resultant understanding that certainty and determinacy are not generally attainable except within contrived or limited conditions. Knowing is increasingly viewed as influenced by prior knowledge, interests, goals, expectations, and other biases. Meaning is understood to be a construction by the knower in interaction with the known. Ultimately, within each developmental model, these insights into the constructive nature of the knowing process lead to the emergence of a full-blown or radical relativism that reflects a complete swing from the pure objectivism of the first developmental level to a pure subjectivism wherein there are no nonarbitrary beliefs. Within this multiplist (Kuhn et al., 2000b) or relativist (Chandler et al., 1990; Perry, 1970) level of epistemic development, there is essentially no distinction between fact and opinion, and all bases for determining what is true or right are considered arbitrary. Some researchers have gone so far as to claim that relativism constitutes a fifth, and distinctly adult, stage of Piagetian cognitive development – a postformal stage that supplants the logic of formal operations with a dialectical logic (Kramer et al., 1992).

Finally, at least in some individuals, relativism gives way to an evaluativist, rationalist level (Chandler et al., 2002; Kuhn et al., 2000b) – the end point of development – also described as commitment within relativism (Perry, 1970), reflective reasoning (King and Kitchener, 1994), or contextual knowing (Baxter-Magolda, 2002). In effect, the extreme tendencies toward objectivism and subjectivism, evident, respectively, in the two previous developmental periods, are now coordinated and balanced. The key advance is an appreciation that there are fallible but legitimate criteria for adjudicating between competing claims and that choice in the face of uncertainty and indeterminacy can still be rational. As in the relativist period, beliefs are understood to have a subjective and defeasible aspect and neither side in a controversy is considered to be right in any simple or

absolute sense. However, it is now recognized that one side may be legitimately superior to the other and that this is demonstrable through a kind of reasoning to the best explanation based on rational considerations of the available evidence.

Epistemic Doubt and Adolescent Development

Research guided by this developmental model initially led to conclusions that reflection upon, and revision of, naïve epistemic beliefs was more characteristic of early adult development than of the adolescent or preadolescent years (Kitchener, 2002). More recent findings, however, indicate that there is more than one pass through this sequence of levels and that epistemic development may be recursive (Chandler, 1987), with the same issues regarding the nature of knowing being addressed at multiple points in development as the individual encounters distinctly new academic or social contexts (Chandler et al., 2002). There may also be periods of regression to earlier positions. While data with college students does point to a relatively late-occurring progression through these developmental levels, it is equally clear that when measured in age-appropriate ways, the same progression is present somewhat earlier (Hallett et al., 2002). There is now ample evidence, during preadolescence, or even middle childhood, of relativistic thinking (Boyes and Chandler, 1992; Kuhn et al., 2000b; Mansfield and Clinchy, 2002), and an appreciation of the active role of the knower in the acquisition of knowledge (Carpendale and Chandler, 1996; Pillow and Henrichon, 1996; Reich et al., 1994; Smith et al., 2000). As Chandler (1987) has pointed out, adolescence is, very typically, a period of profound *epistemic doubt*, i.e., doubt about the possibility of establishing a valid basis for one's beliefs, coupled with interesting and varied ways of coping with that doubt. Resolving the tension between an emerging constructivist, interpretivist view of the mind and the increasingly disturbing implications of that view for strongly held beliefs in multiple domains of knowledge is one of the principle developmental tasks of adolescence and the focus of this encyclopedia article.

Systematic, generic doubt about the underpinnings of deeply held beliefs – as evident within the relativist or multiplist level – would not be possible without a view of the mind as active in the process of knowing, and a view of knowledge as the product of a dynamic, interpretive, constructive process – meaning as made, not found. It stands to reason, then, that the epistemic misgivings of adolescence must be based on a view of knowing that departs significantly from the absolutist and objectivist view, which characterizes the initial stage in adult models of epistemic development (Chandler, 1987; Chandler et al., 1990). While higher order insights into how knowledge is constructed by the knower may well be dependent on the explicit and formal treatment of theories, paradigms, and perspectives that is a unique part of a college curriculum, it is equally clear that a basic appreciation of the active role of the knower in the acquisition of knowledge predates early adulthood, and even adolescence, and emerges gradually across the middle childhood years (Carpendale and Chandler, 1996; Pillow and Henrichon, 1996; Reich et al., 1994; Smith et al., 2000). According to multiple accounts, it is at 6 or 7 years of age that children first show evidence of appreciating that when information about an event or object is ambiguous or incomplete, an individual's beliefs will be influenced by (i.e., reflect bias from) prior experience, expectations, and self-interest (Mills and Keil, 2005; Pillow, 1991; Pillow and Henrichon, 1996; Pillow and Mash, 1999). Further, children of this age are starting to appreciate, in rudimentary ways, how cognitive processes such as inference and planning can give rise to beliefs (Pillow, 1995).

Although some researchers have argued that a constructivist view of knowing may have even earlier origins (Perner, 1991), the evidence is questionable. For example, while 4-year-olds recognize that beliefs are representations of reality and can differ from that reality, as evident from their success on false-belief tasks, they generally view beliefs as originating in perception – not in cognitive processes of interpretation. Thus, older preschool children recognize that if an individual does not have perceptual access to an event, they will have inaccurate beliefs about that event. However, success with false belief is not sufficient to ensure an appreciation of how individuals with equal perceptual access can, nonetheless, arrive at different and equally legitimate interpretations of objects and events (Carpendale and Chandler, 1996). The emergence of the latter appreciation seems best located in the initial period of middle childhood.

The emergent constructivism of the early school years is extended as children begin to reflect upon and make comparisons across cognitive activities, identifying similarities and differences regarding the properties and function of these activities. Ten-year-olds tend to overlook several cognitive processes involved in the acquisition of knowledge when categorizing mental process verbs (Schwanenflugel et al., 1994). This indicates that while children of this age may consider the interpretive role of the mind in the manipulation of information once it is acquired, they do not fully appreciate the minds' equally important interpretive role in knowledge acquisition. Older children, on the other hand, do appreciate the importance, and potentially interpretive roles, of comprehension, selective attention, and recognition in knowledge acquisition. We also see a significant change from 8 to 11 years in the dimensions children employ when comparing and categorizing mental processes. By the end of this period, children are placing significant importance on a mental activity's likelihood of resulting (or not) in certain knowledge (Schwanenflugel et al., 1996, 1998). The growing focus on the certainty–uncertainty dimension is evident in the distinctions children begin to make during these years between deductive (certain) and inductive (uncertain) inference and between weak and strong forms of induction (Moshman, 2004). This new-found concern with degrees of certainty, and a growing awareness that knowledge claims can be uncertain and solutions to knowledge problems indeterminate, indicates that children are coming to recognize the possibility and legitimacy of alternative meanings in rational inquiry (Pillow and Pearson, 2009; Schwanenflugel et al., 1996).

As children approach and move through preadolescence, their emerging constructivist account of the mind leads to

a new kind of dualism that Chandler (1987; Chandler et al., 1990) label 'defended realism.' This represents an initial attempt to maintain the advantages of an absolutist or objectivist epistemology – namely, an assurance that some knowledge is certain and some controversies resolvable – while acknowledging mounting evidence of uncertainty regarding a variety of knowledge claims. In this phase, children implement a kind of bifurcation of experience. In certain domains, such as religion, where it becomes clear that disputes are not the result of limitations regarding access to objective fact, controversies are regarded as unresolvable and a matter of individual preference and opinion. Such domains come to be regarded as arenas for opinion, rather than fact, and for an appreciation of multiple perspectives, rather than truth (Chandler et al., 2002). Other domains come to be viewed as 'objective' and continue to be protected from significant epistemic doubt. However, as sources of bias in these self-declared objective domains become increasingly apparent, the basis for the split is progressively undermined.

The phase of defended realism coincides with the period of concrete operational thought (Inhelder and Piaget, 1964) and the limitations of concrete operations when it comes to generalization and abstraction mean that epistemic doubts, when they arise, tend to remain situational, and isolated to particular cases (Chandler et al., 1990, 2002). In addition, implications of the fallible nature of a particular belief for other beliefs to which it is related in semantic or episodic memory are not recognized. These cognitive limitations of concrete operations tend to prevent epistemic doubt from fully undermining the dualism of defended realism. With the advent of formal operations (Inhelder and Piaget, 1958), however, a deeper appreciation of the logical implications of constructivism emerges and a more generic and systematic form of epistemic doubt develops (Chandler et al., 2002). Because formal operations are operations upon operations – a meta level of operativity – doubt now becomes more principled, systematic, and potentially endemic to all acts of knowing. The dualism of defended realism is not likely to survive the onset and consolidation of formal operational thought. In addition, the new reflective capacity made possible by formal operations – particularly the capacity to reflect on the knowing process (Chandler et al., 1990; Kuhn and Franklin, 2006) – contributes to an appreciation that subject and object are split in inherently unbridgeable ways.

The emergence of a thoroughly systematic and generic epistemic doubt in the adolescent years, following closely upon the emergence of formal operations, is a significantly under-researched aspect of adolescent development. Nonetheless, available evidence suggests that the prevailing adolescent response to this pervasive doubt initially involves elements of both skepticism and dogmatism (Chandler, 1987), possibly intermixed in the same position. This apparent functional equivalence of seemingly opposite attitudes is important. Although diametrically opposed, skepticism and dogmatism are each an adaptation to epistemic doubt from within an operating assumption that rationality requires certainty and determinacy. Skepticism is a full-on acceptance of the logical implications of constructivism, while dogmatism is a 'leap of faith' to investing certain authority figures with 'better reason to know' (Chandler et al., 1990: p. 380) than the adolescent. Because rationality remains tied to certainty and determinacy in adolescent skepticism, the rejection of the possibility of certain knowledge and of fully resolvable disputes and controversies is also a rejection of the possibility of any rational basis for deliberation and choice. Consequently, this adaptive skepticism is associated with decision-making strategies of a distinctly noncognitive nature, including impulse (the avoidance of cognition), intuition, and conformity (Chandler, 1987). Adolescent dogmatism involves an identical assumption about the dependence of rationality on certainty and determinacy. However, in the case of dogmatism, there is an acceptance of various authority figures as embodiments or sources of the rationality that the doubting adolescent is unable to establish through his or her own cognitive activities. Both these adaptations – skepticism and dogmatism – could be characterized as surrendering control (Bendixen, 2002) of one's beliefs to either noncognitive mental processes (as in the case of skepticism) or external authorities (as in the case of dogmatism). Surrender is to be contrasted with taking ownership of one's beliefs and subjecting them to critique and evaluation. Within skepticism/dogmatism, beliefs are in no need of systematic justification because they are seen as being no less legitimate than the next person's beliefs and/or because they have backing of respected authorities.

Evidence suggests that these adaptations to generic epistemic doubt are extended to the process of identity formation as well. Adolescents in the first level of epistemic development – the absolutist or realist level – as well as adolescents in the somewhat more advanced phase of defended realism, typically score in the diffusion or foreclosure statuses of Marcia's ego-identity typology (Boyes and Chandler, 1992). These two statuses both involve a rejection of the idea that exploration is a necessary precondition for making a commitment to particular values, beliefs, or goals. From the standpoint of an epistemic realism or a defended realism, exploration is arguably irrelevant since it assumes that the evaluation of beliefs and values, for example, is complex and deeply subjective or personal – something that is not consistent with a realist position. Adolescents who have reached the second epistemic level – that of relativism or multiplism – and who also espouse a general skepticism, tend to be in either the moratorium status or the diffused status (Boyes and Chandler, 1992). These statuses share a common lack of commitment. Again, this is consistent with a radical relativism because the latter epistemology allows no basis for making commitments. Finally, there is a clear association between the identity achiever status (exploration leading to commitment) and having attained the rationalist or evaluativist epistemic level (Boyes and Chandler, 1992; Krettenauer, 2005). In particular, longitudinal research shows that the evaluativist position specifically contributes to identity achievement (Krettenauer, 2005).

For many adolescents, particularly if college bound, the skepticism–dogmatism adaptation to generic epistemic doubt gives way to a rationalist or evaluativist

epistemology; the third principle epistemic level, in which doubt is more effectively handled by a commitment to justifying and validating knowledge claims in the absence of certainty. The key to moving into a postskeptical rationalism would appear to be a full appreciation of the notion that there can be rational bases for deciding between alternative claims, even though certainty is impossible and alternative perspectives are irreducibly legitimate. What is the basis for this key insight leading the individual beyond relativism and the paralysis of generic epistemic doubt? One central factor appears to be the progressive development of the skills of rational inquiry and argumentation, something that is itself the result of metacognitive advances associated with formal operational thought, direct instruction in formal educational settings, and opportunities to engage in guided exploration, debate, and discussion (Kuhn and Franklin, 2006). As shown by Kuhn and her colleagues, an appreciation that rationality does not require certainty and determinacy develops as individuals become competent at inquiry and argumentation under conditions of uncertainty (Kuhn et al., 2000a, 2008). Such competence includes the capacity to represent one's beliefs as knowledge claims per se and to consider them independently of potential evidence that might confirm or refute such claims, the ability to generate predictions from knowledge claims and to test such predictions, an appreciation of what can be inferred from the outcome of such tests, and an ability to entertain alternative explanations and counterarguments. In the absence of such competence, alleged facts cannot be elevated to the status of testable claims and therefore, are indistinguishable from opinions, the sole stock and trade of a relativist epistemology.

Is Epistemic Development Recursive or Multidimensional?

As a concluding discussion, let us consider what has been noted above as a fundamentally paradoxical finding in this literature. How can we explain the seemingly contradictory findings that the same developmental sequence of epistemic levels has been found in the period extending from middle childhood through late adolescence and in the years from late adolescence through early adulthood? One possibility, as suggested previously, is that there is more than one pass through this developmental sequence (Chandler et al., 2002). One empirically supported proposal for how and why this might happen concerns the distinction between brute facts and institutional facts proposed by John Searle (1969). Brute facts refer to aspects of our knowledge that do not obviously depend in any essential manner on social conventions, rules, or agreements for their meaning. The structure of a cell, the distance of the Earth from the Sun, or the atomic weight of iron are examples of brute facts. Institutional facts, by contrast, clearly derive their meaning, although not their truth, from particular social conventions, rules, or agreements. An individual's nationality or marital status, the current rate of inflation, and the 1964 convention platform of the Democratic Party are institutional facts. The brute–institutional distinction is problematic in some respects, but it appears to reflect lay intuitions and sensibilities. Evidence suggests that development from the absolutist epistemic level through the evaluativist level occurs earlier for institutional fact domains such as the social sciences than it does for brute fact domains such as the physical sciences (Hallet et al., 2002). This may partly explain the recursive nature of epistemic development. While adolescents are wrestling with epistemic doubt largely, if not exclusively, in regard to institutional facts such as religious and political beliefs, the college years appear to involve the development of similar doubts with respect to brute or physical facts. College also involves a continuous undermining of the brute–institutional distinction, as more and more brute facts come to engender the same element of interpretation and indeterminacy that the individual has already come to recognize in institutional facts.

There is an alternative explanation for why we find the same sequence of epistemic levels at two different developmental periods. The appearance of recursion may be misleading and could be produced by changes in relatively independent epistemic dimensions, and in their interrelationships, from preadolescence through the college years. Some researchers have proposed an alternative to the unidimensional, stagelike models of epistemic development discussed above. Schommer (1998) and Hofer (2004), for example, have proposed multidimensional models. In this view, naïve epistemology could be thought of as a system of distinct but interrelated dimensions or factors (Schommer, 1998). Some dimensions may emerge earlier than others, interrelationships among dimensions may change systematically with development, and there may be substantial individual differences in the pathways or end points of development for these dimensions (Schommer-Aikins, 2004). Each of these possibilities could lead to one distinct set of advances occurring before late adolescence and another relatively distinct set occurring across the early adult years.

Although the multidimensional approach has generally been concerned with the study of individual differences in epistemic beliefs, rather than development per se, some findings within this approach suggest developmental trends that might, ultimately, explain the paradox of apparent recursion across developmental periods. Schommer-Aikins and colleagues (Schommer-Aikins et al., 2000, 2005) report that middle school students' epistemic beliefs appear to be multidimensional, but have a simpler structure than found among high school and college students. In addition, coherent beliefs about knowledge acquisition (e.g., whether learning is an automatic or an effortful process) appear to emerge in advance of beliefs about the nature of knowledge (e.g., whether knowledge is simple or complex, permanent or revisable). The multidimensional approach has also provided evidence that an uncritical reliance on experts, textbooks, and other external sources of knowledge is more prolonged than a similarly uncritical reliance on personal experience as a source of knowledge (Conley et al., 2004; Greene et al., 2008).

Conclusions

Naïve beliefs about the nature of knowledge and knowing – epistemic beliefs – are present from childhood onward and undergo development through a relatively fixed sequence of levels from realism through relativism to rationalism. Although there appears to be more than one pass through this sequence, the hallmarks of relativism as experienced during adolescence include a unique, and possibly prolonged, period of generic epistemic doubt during which the individual comes to question whether there can be any rational basis to deeply held beliefs and values. This doubt is manifest, somewhat paradoxically, in forms of skepticism and dogmatism and has significant implications for other areas of development such as identity formation. The road out of this radical relativism appears to require the key recognition that there are rational bases for holding particular beliefs and for preferring one side in a dispute despite the fact that uncertainty and indeterminacy are endemic to all forms of knowing.

See also: Defiant Behavior During Adolescence across Cultures; Epistemological Beliefs and Motivation; Identity and Identification, Social Psychology of; Identity in Childhood and Adolescence; Personality Changes During Adolescence Across Cultures; Self and Identity Development During Adolescence across Cultures; Social Competence During Adolescence Across Cultures; Social Identity in Sociology.

Bibliography

Baxter Magolda, M.B., 1992. Students' epistemologies and academic experiences: implications for pedagogy. Review of Higher Education 15, 265–287.

Baxter Magolda, M.B., 2002. Epistemological reflection: the evolution of epistemological assumptions from age 18 to 30. In: Hofer, B.K., Pintrich, P.R. (Eds.), Personal Epistemology: The Psychology of Beliefs about Knowledge and Knowing. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 89–102.

Bendixen, L.D., 2002. A process model of epistemic belief change. In: Hofer, B.K., Pintrich, P.R. (Eds.), Personal Epistemology: The Psychology of Beliefs about Knowledge and Knowing. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 191–208.

Boyes, M., Chandler, M.J., 1992. Cognitive development, epistemic doubt, and identity formation in adolescence. Journal of Youth and Adolescence 21 (3), 277–304.

Carpendale, J.I.M., Chandler, M.J., 1996. On the distinction between false belief understanding and the acquisition of an interpretive theory of mind. Child Development 66, 1686–1706.

Chandler, M.J., 1987. The Othello effect: essay on the emergence and eclipse of skeptical doubt. Human Development 30 (3), 137–159.

Chandler, M.J., Boyes, M., Ball, L., 1990. Relativism and stations of epistemic doubt. Journal of Experimental Child Psychology 50, 370–395.

Chandler, M.J., Hallett, D., Sokol, B., 2002. Competing claims about competing knowledge claims. In: Hofer, B.K., Pintrich, P.R. (Eds.), Personal Epistemology: The Psychology of Beliefs about Knowledge and Knowing. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 145–168.

Conley, A.M., Pintrich, P.R., Vekiri, I., Harrison, D., 2004. Changes in epistemological beliefs in elementary science students. Contemporary Educational Psychology 29, 186–204.

Greene, J.A., Azevedo, R., Torney-Purta, J., 2008. Modeling epistemic and ontological cognition: philosophical perspectives and methodological directions. Educational Psychologist 43 (3), 142–160.

Hallet, D., Chandler, M.J., Krettenaur, T., 2002. Disentangling the course of epistemic development: parsing knowledge by epistemic content. New Ideas in Psychology 20, 285–307.

Hofer, B.K., 2002. Personal epistemology as a psychological and educational construct: an introduction. In: Hofer, B.K., Pintrich, P.R. (Eds.), Personal Epistemology: The Psychology of Beliefs about Knowledge and Knowing. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 3–14.

Hofer, B.K., 2004. Epistemological understanding as a metacognitive process: thinking aloud during online searching. Educational Psychologist 39 (1), 43–55.

Inhelder, B., Piaget, J., 1958. The Growth of Logical Thinking from Childhood to Adolescence. Basic Books, New York.

Inhelder, B., Piaget, J., 1964. The Early Growth of Logic in the Child. W. W. Norton & Co, New York.

King, P.M., Kitchener, K.S., 1994. Developing Reflexive Judgment. Jossey-Bass, San Francisco, CA.

King, P.M., Kitchener, K.S., 2002. The reflective judgment model: twenty years of research on epistemic cognition. In: Hofer, B.K., Pintrich, P.R. (Eds.), Personal Epistemology: The Psychology of Beliefs about Knowledge and Knowing. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 37–62.

Kitchener, R.F., 2002. Folk epistemology: an introduction. New Ideas in Psychology 20, 89–105.

Kramer, D.A., 1983. Post-formal operations? A need for further conceptualization. Human Development 26, 91–105.

Krettenauer, T., 2005. The role of epistemic cognition in adolescent identity formation: further evidence. Journal of Youth and Adolescence 34 (3), 185–198.

Kuhn, D., 2000. Metacognitive development. Current Directions in Psychological Science 9 (5), 178–181.

Kuhn, D., Black, J., Keselman, A., Kaplan, D., 2000a. The development of cognitive skills to support inquiry learning. Cognition and Instruction 18 (4), 495–523.

Kuhn, D., Cheney, R., Weinstock, M., 2000b. The development of epistemological understanding. Cognitive Development 15, 309–328.

Kuhn, D., Franklin, S., 2006. The second decade: what develops (and how). Handbook of Child Psychology. In: Damon, W., Lerner, R.M., Kuhn, D., Siegler, R.S. (Eds.), Cognition, Perception, and Language, vol. 2. Wiley, Hoboken, NJ.

Kuhn, D., Iordanou, K., Pease, M., Wirkala, C., 2008. Beyond control of variables: what needs to develop to achieve skilled scientific thinking. Cognitive Development 23, 435–451.

Mansfield, A.F., Clinchy, B., 2002. Toward the integration of objectivity and subjectivity: epistemological development from 10 to 16. New Ideas in Psychology 20, 225–262.

Mills, C.M., Keil, F.C., 2005. The development of cynicism. Psychological Science 16, 385–390.

Moshman, D., 2004. From inference to reasoning: the construction of rationality. Thinking and Reasoning. 10 (2), 221–241.

Perner, J., 1991. Understanding the Representational Mind. MIT Press, Cambridge, MA.

Perry, W.G., 1970. Forms of Intellectual and Ethical Development in the College Years: A Scheme. Holt, Rinehart, & Winston, New York.

Pillow, B.H., 1991. Children's understanding of biased social cognition. Developmental Psychology 27, 539–551.

Pillow, B.H., 1995. Two trends in the development of conceptual perspective taking: an elaboration of the passive-active hypothesis. International Journal of Behavioral Development 18, 649–676.

Pillow, B.H., Henrichon, A.J., 1996. There's more to the picture than meets the eye: young children's difficulty understanding biased interpretation. Child Development 67, 803–819.

Pillow, B.H., Mash, C., 1999. Young children's understanding of interpretation, expectation, and direct perception as sources of false belief. British Journal of Developmental Psychology 17, 263–276.

Pillow, B.H., Pearson, R.M., 2009. Children and adult's evaluation of their own inductive inferences, deductive inferences, and guesses. Merrill-Palmer Quarterly 55, 135–156.

Reich, K.H., Oser, F.K., Valentin, P., 1994. Knowing why i now know better: children's and youth's explanations of their worldview changes. Journal of Research on Adolescence 4 (1), 151–173.

Schommer, M., 1998. The influence of age and education on epistemological beliefs. The British Journal of Educational Psychology 68, 551–562.

Schommer-Aikins, M., 2004. Explaining the epistemological belief system: introducing the embedded systemic model and coordinated research approach. Educational Psychologist 39 (1), 19–31.

Schommer-Aikins, M., Duell, O.K., Hutter, R., 2005. Epistemological beliefs, mathematical problem-solving beliefs, and academic performance of middle school students. The Elementary School Journal 105 (3), 289–303.

Schommer-Aikins, M., Mau, W.C., Brookhart, S., Hutter, R., 2000. Understanding middle students' beliefs about knowledge and learning using a multidimensional paradigm. Journal of Educational Research 94 (2), 120–127.

Schwanenflugel, P.J., Fabricius, W.V., Alexander, J.M., 1994. Developing theories of mind: understanding concepts and relations among mental activities. Child Development 65, 1546–1563.

Schwanenflugel, P.J., Fabricius, W.V., Noyes, C.R., 1996. Developing organization of mental verbs: evidence for the development of a constructivist theory of mind in middle childhood. Cognitive Development 11, 265–294.

Schwanenflugel, P.J., Henderson, R.L., Fabricius, W.V., 1998. Developing organization of mental verbs and theory of mind in middle childhood: evidence from extensions. Developmental Psychology 34 (3), 512–524.

Searle, J.R., 1969. Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press, London.

Smith, C.L., Maclin, D., Houghton, C., Hennessey, M.G., 2000. Sixth grade students' epistemologies of science: the impact of school science experiences on epistemological development. Cognition and Instruction 18 (3), 349–422.

Epistemological Beliefs and Motivation

Kerstin Oschatz, Hector Research Institute of Education Sciences and Psychology, Tübingen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

For the last three decades researchers have studied beliefs about knowledge and knowing and their impact on learning. These so-called epistemological beliefs form a complicated and intricate system of beliefs. They not only influence academic performance or information processing but also interact with constructs like achievement motivation. This article will give a brief introduction into the research concerning epistemological beliefs and provide an overview on studies concerning their relation to achievement motivation.

Introduction

Many factors determine how we learn and influence our subsequent performance. Confidence in our own abilities, the knowledge we already have about the learning task, our goals – these factors are some of the major influences widely discussed within literature. For the last three decades, research in education and psychology has taken another factor into account: Beliefs about knowledge and knowing, also known as epistemological beliefs.

These beliefs become salient in a variety of educational experiences: Whenever individuals encounter new information they have to understand, evaluate, or integrate into their existing knowledge, the beliefs they hold about the structure and stability of knowledge or about the justification for their knowing are activated. Individuals have to explicitly or implicitly answer the questions, how they do know what they know and how to choose what and whom to believe (Hofer, 2001). For these answers they tap into their beliefs about knowledge and knowing, which they have both formally and informally acquired. Epistemological beliefs have been linked to various aspects of learning, such as strategy use, reasoning, academic performance, information processing, and motivation. Personal epistemology consequently plays an essential part in learning, yet it has proven to be a rather complicated construct. This is apparent in the number of constructs defined and used to assess personal epistemology, such as epistemological reflection (Baxter Magolda, 2004), epistemic cognition (King and Kitchener, 1994), or epistemological understanding (Kuhn et al., 2000).

The diversity is owed to diverging ideas about which aspects to include within personal epistemology and which aspects to exclude as associated but distinct constructs. Research has called for holistic approaches to personal epistemology, taking further aspects of cognition into account. "The need for an embedded systemic model of epistemological beliefs, that is, a model that includes many other aspects of cognition and affect, comes from the assumption that epistemological beliefs do not function in a vacuum. Indeed, at any given moment, learners' thoughts, actions, or motivations represent the convergence of multiple systems" (Schommer-Aikins, 2004). Although research on such connections is still in its infancy, on motivational grounds auspicious studies exist, which promise that research on epistemological beliefs could shed a new light on the foundations of achievement motivation. The purpose of this article is to give a brief introduction into the research concerning epistemological beliefs and to provide an overview on studies concerning their relation to achievement motivation.

Perspectives on Epistemological Beliefs

The exploration of questions concerning the nature and justification of knowledge has originated in the philosophical discipline of epistemology. The beliefs and theories that individuals hold about knowledge and knowing are subsumed by the terms epistemological beliefs, epistemic beliefs, or personal epistemology. The different conceptualizations vary with regard to their dimensions and boundaries, context sensitivity, and relationship to a variety of cognitive learning outcomes as well as their normative implications in educational settings. The following sections provide a brief overview of the main approaches following essential questions regarding the existing models, the development, the universality, and specificity of epistemological beliefs as well as questions about measures for epistemological beliefs and normative implications of the construct.

How Are Epistemological Beliefs Modeled?

Perry (1970) was one of the first to study individuals' epistemological beliefs as part of an attempt to understand how students interpreted their pluralistic educational experiences. His theory of epistemological development became the basis of a vast variety of studies and approaches concerned with individuals' personal epistemologies. In Perry's model of intellectual and ethical development, personal epistemology is unidimensional and develops from a dualistic perspective on knowledge and the world (being either right or wrong) via the beginning recognition of diversity and uncertainty toward a focus on responsibility and commitment within contextual relativism. At the developmental endpoint the self is perceived as an active maker of meaning and is in need to choose and affirm its own commitments (Hofer and Pintrich, 1997). Perry's research has been continued by researchers conceptualizing personal epistemology from a structural and developmental perspective, exploring how individuals interpret their educational experiences, taking aspects of gender into account (Baxter Magolda, 2004).

Extending Perry's scheme to the domain of everyday reasoning King and Kitchener (1994) assessed individuals'

perspectives on knowledge and their conceptions of justification and argument by presenting individuals varying in age and educational background with different ill-structured problems. Their *reflective judgment model* consists of seven stages, organized in three levels of prereflective, quasi-reflective, and reflective thinking. Using interviews, Kuhn developed a model consisting of three distinct epistemological stages ranging from an absolutistic perspective, over a multiplistic toward an evaluativistic perspective (Kuhn et al., 2000). All three models view epistemological beliefs as a unidimensional construct, in which individuals move through the different stages as they develop.

A competing perspective has been introduced by Schommer (1990). She shifted the way researchers understand epistemological beliefs by conceptualizing the construct from a multidimensional perspective, proposing a system of beliefs, in which different epistemological beliefs emerge and develop at different rates. Along with this new asynchronous perspective she introduced a novelty into research, assessing individuals' beliefs through a questionnaire. In her *Epistemological Beliefs Questionnaire* (EBQ) she identified four to five dimensions of epistemological beliefs, each of which is viewed as a continuum, ranging from beliefs about: (1) the structure of knowledge (knowledge consists of isolated bits of information vs knowledge as highly interrelated concepts), (2) the stability of knowledge (knowledge is certain vs tentative and evolving), (3) the source of knowledge (knowledge is handed down by authority vs individually constructed), (4) the nature of knowledge acquisition (learning occurs quickly vs gradual), and (5) the ability to acquire knowledge (intelligence is fixed vs incremental; Schommer-Aikins, 2004). Despite methodological issues (Qian and Alvermann, 1995) as well as further questionnaire developments (Hofer, 2000), Schommer-Aikins' EBQ remains the primary paper-and-pencil assessment of personal epistemology. Most importantly, her approach to the study of epistemological beliefs provided researchers with measures to explore relations with different factors in the learning process. For a more detailed account of the different epistemological models, please read the comprehensive review of Hofer and Pintrich (2002).

How Do Epistemological Beliefs Develop?

Effects of age and education on the development of knowledge beliefs were one of the first factors explored (Perry, 1970). As Buehl (2003) states, it is difficult to differentiate the internal processes of maturation from the effect of educational experiences that individuals encounter during their development. Research suggests that age works as a significant predictor for changes in students' beliefs about learning, whereas education functions as a significant predictor of the sophistication of epistemological beliefs (Schommer, 1998). Although there is little research in younger populations like elementary school-or middle-school students, findings suggest that the structure and specifications of epistemological beliefs become more complex in populations of greater age (Qian and Alvermann, 1995).

How Are Epistemological Beliefs Related to Cognitive Learning Outcomes?

Most of the research on epistemological beliefs in the last 20 years has focused on the nature of students' knowledge beliefs and how these beliefs relate to a variety of cognitive learning outcomes. In the most prominent studies higher levels of comprehension, strategy use, and academic performance have been associated with what researchers view as more sophisticated beliefs about knowledge (Schommer, 1998; Kardash and Howell, 2000). Advanced epistemological beliefs have been linked to higher levels of information processing (Kardash and Scholes, 1996) and conceptual change (Qian and Alvermann, 1995).

How Specific Are Epistemological Beliefs?

An ongoing debate within research concerns the question whether epistemological beliefs have to be conceptualized as domain specific or domain general. In a host of studies, knowledge beliefs are assessed without being contextualized in a specific field or domain of knowledge (Schommer, 1990). Contrasting, there are researchers exploring the structure of epistemological beliefs within or between distinct academic domains by using domain-specific questionnaires developed to deliberately tab epistemological beliefs with regard to specific subjects like mathematics or history (Buehl et al., 2002). Another strand of research administers the same questionnaire to students of different domains (Jehng et al., 1993) or to the same students, asking them to answer the same questions with regard to two different domains (Hofer, 2000). According to these studies, college students majoring in 'soft' domains hold more sophisticated epistemological beliefs than students majoring in 'hard' fields. Research suggests that students possess general as well as domain-specific epistemological beliefs and that these beliefs are interrelated. Some researchers therefore suggest a multilevel conception of epistemological beliefs (Trautwein and Lüdtke, 2007) in which domain-specific beliefs could be conceptualized as an 'applied form' of general beliefs.

How Are Epistemological Beliefs Measured?

The quality of instruments used to measure epistemological beliefs is the center of a critical debate within research (Bromme, 2005; Trautwein and Lüdtke, 2007). One way of assessing epistemological beliefs is by using open-ended or half-structured interviews, like the *Reflective Judgment Interview* (RJI) developed by King and Kitchener (1994). Although research has found satisfactory internal consistency for the RJI (Wood, 1997), these measures are not suitable for large samples due to the extensive and complex coding involved in their usage. Paper and pencil measures based on the EBQ (Schommer, 1990) therefore form the basis of most research in psychological and educational contexts. Items are generally based on global and decontextualized statements like "Scientists can ultimately get to the truth" (Hofer, 2000). The validity of such global measures is questioned by a variety of researchers. They claim that it remains unclear, what frame of reference individuals choose when evaluating these global statements (Bromme, 2005).

In addition, epistemological beliefs influence learning and thinking in a variety of forms, as they can function as declarative or procedural knowledge – explicit and directly assessable for the individual – or as tacit or implicit knowledge. As

Buehl and Alexander (2006) point out, these different forms of knowledge beliefs coexist and interact in complex ways. Epistemological beliefs are therefore referred to as "socially shared intuitions about the nature of knowledge and learning" (Jehng et al., 1993) or 'informal knowledge' (Elby and Hammer, 2001). So far no reliable instruments have been developed to measure implicit dimensions of epistemological beliefs.

How to Deal with the Normative Implications of Epistemological Beliefs?

Research on epistemological beliefs is intertwined with normative perspectives. Within the unidimensional approaches, epistemological beliefs have been categorized in a continuum of ongoing development with higher stages of development pointing to a somewhat superior epistemological understanding. Within her multidimensional model Schommer-Aikins (Schommer, 1990) introduced a distinction between 'naïve' and 'sophisticated' stances with regard to each of the proposed belief factors. With regard to the cultural dependency of epistemological beliefs it is important to note that current models conceptualize the endpoint of epistemological development in a formative and abstract way of thinking, which has to be understood as a normative characteristic of Western cultures: The ideal of a stronger individualism in thinking, freeing thought from the dictation of authority is a symbol of an individualistic and Westernized socialization. In collectivistic cultures, in which the perspective on the self has interindividual implications, epistemological development might turn toward positions propagating an acceptance of consensus instead of a thinking that strives to be increasingly independent (Cobern and Loving, 2008).

Furthermore the exploration of the relation between epistemological beliefs and different factors in the learning process has led to normative assumptions. Higher levels of strategy use, comprehension, and academic performance have been associated with what researchers view as sophisticated epistemological beliefs (e.g., knowledge is tentative, complex, and derived from reason). As a consequence, a research strand has developed, that explores which forms of education might foster sophisticated epistemological beliefs. However, some research suggests that the connection is not always linear with students sometimes expressing less advanced epistemological beliefs as they gain knowledge in a domain (Trautwein and Lüdtke, 2007). According to Elby and Hammer (2001) researchers have failed "to distinguish between the correctness and productivity of an epistemological belief." A belief should be viewed as productive "if it generates behavior, attitudes, and habits that lead to 'progress' as defined by the given person or community" (Elby and Hammer, 2001, p. 555). Bromme et al. (2008) suggest that *sophistication* should be conceptualized as flexibility of epistemological judgment with regard to differences between contexts. Research should consequently account for the grade of adaptivity or advancement. Following this perspective, epistemological judgments are generated by the interplay of epistemological beliefs, ontological knowledge, and topic-related knowledge as well as the person's ability to account for the demands of a context. Bromme et al. (2008) suggest a new perspective on the adaptivity of epistemological beliefs introducing the idea of the 'division of cognitive labor': Individuals can display an adaptive epistemological understanding without having a fully elaborated understanding of the topic or its underlying ontological assumptions, if they can make out who could give reliable information on a topic and how one could distinguish between competing sources of knowledge claims.

The Relation of Epistemological Beliefs to Achievement Motivation

Over the last decade there have been suggestions that it is necessary to view personal epistemology in a broader context and to explore its relation to achievement motivation (Hofer and Pintrich, 1997). Achievement motivation refers to motivation in situations related to an individuals' competence and achievement-related behaviors (Nicholls, 1984). The relationship between personal epistemology and achievement motivation can be viewed from two angles: One line of research examines the influences of epistemological beliefs on constructs within achievement motivation and self-regulation (e.g., Bråten and Strømsø, 2005). The other line focuses on motivational states as mediators of the effect of epistemological beliefs on learning outcomes (e.g., Richter, 2010). For both perspectives only few empirically sound studies exist. The relationship of epistemological beliefs and achievement motivation uses different theoretical motivational frameworks: Following Buehl (2003), research will therefore be grouped into three sections dealing with the relations of epistemological beliefs and goal orientation, achievement values, and competency beliefs, respectively.

Epistemological Beliefs and Goal Orientation

Achievement goal orientation theory represents one of the most influential approaches to academic motivation in contemporary educational psychology and looks at the reasons or purposes students have for engaging in academic tasks. Two types of goal orientations have been differentiated within literature: mastery-goal and performance-goal orientation. Adopting mastery goals (focus on learning, understanding, mastering tasks, and personal improvement) has been related to more effort and persistence in academic work, the use of a variety of learning strategies that encourage deeper information processing as well as the promotion of interest. Research has broken performance goals down into performance-approach goals (focus on the demonstration of a high competence with the purpose to outperform others or attain favorable judgments of competence) and performance-avoidance goals (focus on the demonstration of a high competence with the purpose to avoid any demonstration of incompetence). While performance-approach goals have been linked to more positive consequences for motivation and achievement, avoidance goals lead to maladaptive outcomes like less intrinsic motivation and poorer achievement (Eccles and Wigfield, 2002; Bråten and Strømsø, 2004). Research distinguishes several potential antecedents of goals, ranging from competence-related beliefs (need for achievement, competence expectancy, fear of failure, trait test anxiety) and contextual variables (evaluation practices) to personal factors like gender or implicit theories of intelligence. According to Dweck and Leggett

(1988), mastery goals are more likely to be adopted by students who believe intelligence is malleable (incremental belief), whereas performance goals tend to be adopted by students who believe intelligence is fixed (entity belief).

Hofer and Pintrich (1997) were among the first to suggest that epistemological beliefs may function as implicit theories that can give rise to personal goals in learning. Buehl (2003) gives a more detailed account of this possible relation: Beliefs about abilities and the desire to engage in a learning task depend on the perception of that specific task. In academic contexts, tasks require the acquisition or manipulation of knowledge. Therefore beliefs about knowledge might influence task-perception and thereby influence goal orientation (Buehl, 2003).

This relation has been confirmed by several studies examining the relation of epistemological beliefs and goal adoption, indicating that students with advanced epistemological beliefs tend to adopt mastery goals. Because all available studies used measures based on Schommer's EBQ (Schommer, 1990), results can be grouped in relation to the five dimensions of personal epistemology she proposed.

Several studies suggest that more advanced beliefs in regard to the *structure of knowledge* (students perceive knowledge more as highly interrelated and less as isolated bits of information) relate to a tendency to adopt mastery goals, whereas students with less advanced beliefs more likely adopt performance goals (Buehl and Alexander, 2006; Murphy et al., 2002; Paulsen and Feldman, 1999).

Studies regarding the dimension *nature of knowledge acquisition* also suggest more advanced beliefs (learning occurs gradual and requires effort) being related to the adoption of mastery goals and less advanced beliefs (learning occurs quickly) being related to the adoption of performance-avoidance goals (Bråten and Strømsø, 2004, 2005, 2006; Buehl and Alexander, 2006; Qian and Burrus, 1996). According to Bråten and Strømsø (2004), who examined the long-term adoption of such goals, this predictive pattern was consistent over time. However, few studies report contradictory findings: Murphy et al. (2002) found beliefs about the need for effort to acquire knowledge – which is typically considered to be an adaptive and advanced perspective on knowledge acquisition – to be related to less adaptive goal orientations. Contradictory findings in respect to both described dimensions stem from a study with students in the South Pacific Region by Phan (2008). Less advanced beliefs regarding the *nature of knowledge acquisition* and the *structure of knowledge* were linked to a tendency to adopt mastery goals. Phan (2008) discusses cultural influences as a possible explanation for these unusual findings.

Regarding the dimensions *stability of knowledge* and *source of knowledge* similar patterns of results were shown as have been reported for the two dimensions described above. Several studies describe more advanced beliefs like perspectives of knowledge being tentative and evolving as well as beliefs of knowledge being constructed by the individual as related to mastery-goal orientation. Less advanced perspectives in regard to the *stability* and *source of knowledge* (knowledge is certain, knowledge is handed down by authority) were linked to the adoption of performance-goal orientations (Bråten and Strømsø, 2004, 2005, 2006; Chen and Pajares, 2010; Kizilgunes et al., 2009; Phan, 2008; Qian and Burrus, 1996). On top of these diverse findings, evidence from a study by Bråten and Strømsø (2005) shows that different dimensions of epistemological beliefs predicted goal orientation in separate student populations of different academic contexts (business administration vs student teachers). These findings indicate that the relationship between epistemological beliefs and goal orientations might be influenced by contextual factors and might therefore be more complicated than initially assumed.

Regarding the epistemic dimension *ability to acquire knowledge* only few studies exist. Their results are in line with existing research on the relation between implicit beliefs of intelligence and goal adoption, indicating that advanced beliefs (intelligence is incremental) are related to mastery-goal orientation (Phan, 2008; Qian and Burrus, 1996).

The majority of studies support the general notion of more advanced epistemological beliefs being related to mastery-goal orientations across different epistemological dimensions. Because most studies only examined correlational relations between dimensions of epistemological beliefs and goal orientation they allow no claims concerning the causal direction of the relationship. Furthermore single studies indicate that aspects of culture and the learning context might influence the relationship.

Epistemological Beliefs and Achievement Values

One of the factors seen to be influential for students' academic choices within educational motivation research is the value that students attach to specific tasks or subjects. Within expectancy-value theory (Eccles and Wigfield, 2002), students' subjective task value along with their competence beliefs is seen to drive students' achievement behavior. Eccles and Wigfield (2002) distinguish between four components of task value. (1) Attainment value is defined as the personal importance of doing well on a task linked to the possibility of confirming salient aspects of an individuals' self-schema. (2) Intrinsic value can be understood as the enjoyment an individual gets from engaging in an activity linking this value to subjective interest. (3) The contribution of a task to the achievement of future goals is defined as utility value. (4) Finally, cost captures all the negative aspects that come along with engaging in a task. Within the model, it is assumed that achievement values are partly influenced by task-specific beliefs (perception of competence regarding the task and task-difficulty). As epistemological beliefs play an important role within learning, researchers have suggested, that they might contribute to an individual's perception of a learning task, but only very few studies address this relationship. Paulsen and Feldman (1999) report less advanced beliefs regarding the *nature of knowledge acquisition* (learning occurs quick) as well as the *ability to acquire knowledge* (intelligence is fixed) to be negatively correlated with students' achievement values. In a study with college students Buehl (2003) investigated the interconnection of achievement value (as a single construct) and domain-specific epistemological beliefs in the subjects history and mathematics. Advanced beliefs regarding the *structure of knowledge* (knowledge is complex and highly interrelated) and the *source of knowledge* (knowledge has to be constructed and is derived from reason) were related to higher levels of achievement value. On the contrary less advanced beliefs on the *stability of knowledge* (knowledge is certain) in both subjects

were also linked to higher achievement value. Enman and Lupart (2000) studied factors underlying undergraduate students' selection of science majors measuring epistemological beliefs and achievement values. Please note that the objective in this study was to measure epistemological beliefs and achievement value to predict students' choice of science as a major. Therefore the relation of both constructs was not the focus of this study. Enman and Lupart (2000) report that less advanced beliefs on the *ability to acquire knowledge* (belief in fixed ability) predicted if students selected science as their favorite subject and choose science as their major. Following Buehl (2003) 'choice of favorite subject' can be understood as a proxy for achievement value. Thereby this study can be interpreted as support for a link between less advanced beliefs on the *ability to acquire knowledge* and higher levels of achievement value.

All in all there is not enough empirical evidence to allow clear statements about the relationship of epistemological beliefs and achievement values. The described studies indicate that the interconnection might be complicated, as they show a mixture of advanced and less advanced knowledge beliefs being connected to higher levels of achievement value.

Epistemological Beliefs and Competency Beliefs

Individuals' achievement choices are not only driven by the values that individuals attach to specific tasks, but also by their beliefs about their own competency. Eccles and Wigfield (2002) point out competency beliefs as directly related to the question "Can I do this task?". Individuals' beliefs about their own competency can relate to their assessments of their ability to organize and execute the courses of action required to produce specific given attainments (*self-efficacy*) as well as to their beliefs about their ability to successfully master certain tasks in the future (*expectancy for success*; Eccles and Wigfield, 2002). Individuals who have strong beliefs about their own competencies show higher achievement and greater motivation and persistence toward challenging tasks (Eccles and Wigfield, 2002).

Individuals' competency beliefs (such as *self-efficacy* and *expectancy for success*) are seen to be influenced by ability beliefs as well as mastery experiences. These two factors are especially interesting in relation to epistemological beliefs, as they are based on student's perception of the task demands. Students' epistemological beliefs can be understood as an aspect of students' perception of the learning task (Buehl, 2003). Therefore students' judgments about their capabilities in reference to a learning task could be partly based on their conceptions of knowledge and how it is gained. Within the formation of self-efficacy and expectancy beliefs, these conceptions of knowledge and experiences with knowledge acquisition might influence students' perceptions of their abilities.

Only little research exists on the relationship of epistemological and competency beliefs. Results of several studies show that more advanced beliefs about the *source of knowledge* (knowledge is constructed) were positively related to self-efficacy and expectancy for success (Bråten and Strømsø, 2005; Chen and Pajares, 2010; Buehl, 2003). Conflicting to these findings Kizilgunes et al. (2009) report more advanced perspectives like beliefs that knowledge comes from reasoning, thinking and experimenting to be connected to lower self-efficacy, whereas less advances beliefs (knowledge being handed down by authority) were related to higher levels of self-efficacy.

Regarding the epistemological dimension *stability of knowledge* similar contradictory findings exist: Whereas most studies report less advanced beliefs about the *stability of knowledge* as negatively related to self-efficacy (Bråten and Strømsø, 2005; Chen and Pajares, 2010) and advanced beliefs as related to higher levels of self-efficacy (Kizilgunes et al., 2009; Phan, 2008), the opposite relationship was reported by Buehl (2003). In her study on domain specific epistemological beliefs in the subjects mathematics and history, less advanced beliefs on *the stability of knowledge* in mathematics or history were related to greater confidence in competency.

Results of several studies indicate a negative correlation of self-efficacy with less advanced beliefs about the *structure of knowledge* (knowledge exist in isolated bits and pieces; Neber and Schommer-Aikins 2002; Paulsen and Feldman, 1999), whereas more advanced perspectives on the interrelatedness of knowledge were positively correlated with self-efficacy (Buehl and Alexander, 2006) and expectancy for success (Buehl, 2003). Here Phan (2008) reports contradictory findings and discusses influences of culture as a possible explanation for the results.

Regarding the epistemological dimension *nature of knowledge acquisition* studies suggest more advanced beliefs (learning is a gradual and effortful process) to be related to higher levels of self-efficacy (Bråten and Strømsø, 2005; Buehl and Alexander, 2006). Similar findings were reported for more advanced beliefs on the *ability to acquire knowledge* (intelligence is incremental) and their relation to higher levels of self-efficacy (Bråten and Strømsø, 2005; Phan, 2008; Paulsen and Feldman, 1999).

All in all knowledge about the relationship of the various dimensions of epistemological beliefs and competency beliefs is based on a thin empirical basis and shows an inconclusive pattern. In tendency results indicate that more advanced knowledge beliefs are related to higher levels of self-efficacy, but contradictory findings exist. These results lead to the notion that the relationship under study might be complicated and possibly influenced by further factors. Supporting evidence for this interpretation comes from Bråten and Strømsø (2005). In their study different dimensions of epistemological beliefs predicted self-efficacy in student populations from two different academic contexts: Less advanced beliefs on the *nature of knowledge acquisition* as well as on the *ability to acquire knowledge* negatively predicted self-efficacy in student teachers. Epistemological beliefs concerning the *source of knowledge* as well as beliefs concerning the *stability of knowledge* predicted self-efficacy in business administration students. This suggests that the relationship between epistemological beliefs and competency beliefs might be influenced by contextual factors.

Discussion

The existing research in the cross-section of personal epistemology and achievement motivations suggests possible connections between constructs. Findings also indicate that the relationship is intricate and complex: Although most

empirical evidence points toward a linear relationship between the constructs, not all studies support this claim.

Regarding the relationship of epistemological beliefs with goal orientation findings suggest that more advanced knowledge beliefs are related to more adaptive goal orientations like mastery or performance-approach goals. Due to the mainly correlational relations, no secure claims concerning the causal direction of the relationship can be drawn. Nor do these studies allow precluding the possibility of an external third variable connecting the two constructs. Despite these limitations present findings suggest that epistemological beliefs could be added to the list of antecedents of achievement goals, thereby broadening the understanding of goal formation. It is left to future research in this field to explore the interplay between the diverse constructs known to affect goal formation. Further research is also required to explore the differential effects of specific dimensions of epistemological beliefs on goal adoption, which may even be influenced by cultural factors (Bråten and Strømsø, 2005, 2006; Phan, 2008). In reaction to their results of different epistemological dimensions have changing importance for goal orientation of students of different academic domains, Bråten and Strømsø (2005) suggest that differences in instructional practices might foster importance of certain epistemological perspectives for individuals. These differences in the way students' epistemological beliefs operate across academic domains could have instructional implications, which should be explored in future research (Bråten and Strømsø, 2005).

For the relationship of epistemological beliefs and achievement values only few studies exist. Although present studies suggest that epistemological beliefs might be related to achievement value, there is more research needed to allow reliable conclusions about the nature and direction of that relationship.

Research on potential relations of competency beliefs and epistemological beliefs do not allow causal claims about the direction of the relationship. Chen and Pajares (2010) make careful interpretations of their findings, suggesting that beliefs about learning and ability may be antecedents to epistemological beliefs. This idea is in line with a hypothesis by Hofer and Pintrich (1997) assuming that beliefs about knowledge and knowing develop rather late within adolescence, whereas beliefs about ability and learning are formed relatively early in childhood. Future research will be needed to address the relationship and possible developmental implications between epistemological beliefs and competency beliefs, exploring if the relationship changes as a function of maturation or educational experience.

Conclusion

Epistemological beliefs are a complicated but central construct in respect to learning that has only been explored for a couple of decades. In this article, epistemological beliefs have been introduced as beliefs about the structure and stability of knowledge and about the justification for knowing, which are formally and informally acquired in the process of socialization and enculturalization. They become salient whenever individuals encounter new information they have to understand, evaluate, or integrate into their existing knowledge. One central aspect in regard to epistemological beliefs is the normative perspective embedded in the construct. The distinction of more 'sophisticated' versus 'naive' stances in regard to knowledge beliefs was clearly constructed from a viewpoint of western cultures, transferring the ideal of a stronger individualism in thinking into the classification of knowledge beliefs. Lately there has been a shift in looking at these normative implications with researchers suggesting focusing on the 'productivity' (behavior or attitudes, that lead to 'progress' as defined by the given person or community; Elby and Hammer; 2001) or 'adaptivity' (Bromme et al. 2008) of epistemological beliefs. Following the perspective formulated by Bromme et al. (2008), *sophistication* should be conceptualized as flexibility of epistemological judgment with regard to differences between contexts.

Effects of epistemological beliefs on a variety of cognitive learning outcomes such as strategy use, reasoning, academic performance, information processing and motivation have been studied in the last decade. Research from this last strand concerning the relationship of epistemological beliefs and achievement motivation has been the center of this article. As outlined in detail above, existing evidence in this field indicates a somewhat linear relationship between the constructs with more advanced knowledge beliefs being related to more adaptive motivational states. Not all studies support this claim though, suggesting the opposite connection of less advanced epistemological beliefs with more positive motivational states and thereby reinforcing questions concerning the 'productivity' or 'adaptivity' of epistemological beliefs in respect to learning.

In a more general approach to the question of possible relations between achievement motivation and personal epistemology, Bråten and Olaussen (2005) explored how certain distinct student motivation profiles drawn from measures of personal interest, mastery goal orientations, task-value beliefs, and self-efficacy are related to epistemological beliefs. Their results showed distinct motivational cluster groups with differences in regard to students' epistemological beliefs. Students with higher levels of motivation and self-efficacy showed a tendency to hold more advanced knowledge beliefs. Although no causal conclusion can be drawn from this study, Bråten and Olaussen (2005) interpret their results as support for the theoretical claim that epistemological beliefs may be important for academic motivation.

Despite the great need for further research, findings so far indicate that epistemological beliefs may offer educators an additional means to influence students' motivation and achievement. Based on their longitudinal study, Bråten and Olaussen (2005) suggest that the relationship between dimensions of personal epistemology and academic motivation is relatively stable over time. They therefore conclude that students, whose academic motivation rises over time, may be supported in this enhancement through more advanced epistemological beliefs. Taking up the discussion on normativity in reference to Elby and Hammer (2001), this could also require focusing on those epistemological beliefs, which have proven to be most 'productive' in supporting certain motivational components like achievement goals or task value.

Would this untested possibility be confirmed, positive motivational states could be supported through fostering of the 'right' epistemological beliefs. The downside of this claim is that studies exploring the conditions involved in epistemic change are still in their infancy. Researchers are only starting to explore central factors in the development of epistemological beliefs.

Bendixen and Rule (2004) discuss 'personal relevance' as a possible condition for change of epistemological beliefs, including topic interest, emotional involvement, and high self-efficacy as factors within this condition. This points toward the circular relationship between epistemological beliefs and motivational constructs already mentioned in the outline of this article. To shed light into this complicated and intricate system of constructs will be a major but promising challenge for future research in the field of motivation and epistemological beliefs.

See also: Affect-Regulation Motivation; Avoidance and Approach Motivation: A Brief History; Leisure Activities Choices among Adolescents; Mastery Learning; Motivation, Learning, and Instruction; School Achievement: Motivational Determinants and Processes; Schooling: Impact on Cognitive and Motivational Development; Self-Regulated Learning: Theories, Measures, and Outcomes; Self-Regulation in Adulthood; Test Anxiety and Academic Achievement.

Bibliography

Baxter Magolda, M.B., 2004. Evolution of a constructivist conceptualization of epistemological reflection. Educational Psychologist 39 (1), 31–42.

Bråten, I., Strømsø, H.I., 2006. Predicting achievement goals in two different academic context. A longitudinal study. Scandinavian Journal of Educational Research 50 (2), 127–148.

Bråten, I., Strømsø, H.I., 2005. The relationship between epistemological beliefs, implicit theories of intelligence, and self-regulated learning among Norwegian postsecondary students. British Journal of Educational Psychology 75 (4), 539–565.

Bråten, I., Strømsø, H.I., 2004. Epistemological beliefs and implicit theories of intelligence as predictors of achievement goals. Contemporary Educational Psychology 29, 371–388.

Bendixen, L.D., Rule, D.C., 2004. An integrative approach to personal epistemology: a guiding model. Educational Psychologist 39 (1), 69–80.

Buehl, M.M., 2003. At the Crossroads of Epistemology and Motivation: Modeling the Relations between Students' Domain-Specific Epistemological Beliefs, Achievement Motivation, and Task Performance (Unpublished Doctoral Dissertation). University of Maryland, College Park.

Buehl, M.M., Alexander, P.A., 2006. Examining the dual nature of epistemological beliefs. International Journal of Educational Research 45, 28–42.

Buehl, M.M., Alexander, P.A., Murphy, P.K., 2002. Beliefs about schooled knowledge: domain general or domain specific? Contemporary Educational Psychology 27, 415–449.

Bromme, R., 2005. Thinking and knowing about knowledge: a plea for and critical remarks on psychological research programs on epistemological beliefs. In: Hoffmann, M., Lenhard, J., Seeger, F. (Eds.), Activity and Sign – Grounding Mathematics Education. Springer, New York, pp. 191–201.

Bromme, R., Kienhues, D., Stahl, E., 2008. Knowledge and epistemological beliefs: an intimate but complicate relationship. In: Khine, M.S. (Ed.), Knowing, Knowledge, and Beliefs: Epistemological Studies across Diverse Cultures. Springer, New York, pp. 423–444.

Chan, K., Elliott, R.G., 2002. Exploratory study of Hong Kong teacher education students' epistemological beliefs: cultural perspectives and implications on beliefs research. Contemporary Educational Psychology 27, 392–414.

Chen, J.A., Pajares, F., 2010. Implicit theories of ability of grade 6 science students: relation to epistemological beliefs and academic motivation and achievement in science. Contemporary Educational Psychology 35, 75–87.

Cobern, W.W., Loving, C.C., 2008. An essay for educators: epistemological realism really is common sense. Science & Education 17, 425–447.

Dweck, C., Leggett, E.L., 1988. A social-cognitive approach to motivation and personality. Psychological Review 95 (2), 256–273.

Eccles, J.S., Wigfield, A., 2002. Motivational beliefs, values and goals. Annual Review of Psychology 53, 109–132.

Elby, A., Hammer, D., 2001. On the substance of a sophisticated epistemology. Science Education 85 (5), 554–567.

Enman, M., Lupart, J., 2000. Talented female students' resistance to science: an exploratory study of post-secondary achievement motivation, persistence, and epistemological characteristics. High Ability Studies 11, 161–178.

Hofer, B.K., 2001. Personal epistemology research: implications for learning and teaching. Journal of Educational Psychology Review 13 (4), 353–383.

Hofer, B.K., 2000. Dimensionality and disciplinary differences in personal epistemology. Contemporary Educational Psychology 25, 378–405.

Hofer, B.K., Pintrich, P.R. (Eds.), 2002. Personal Epistemology: The Psychology of Beliefs about Knowledge and Knowing. Lawrence Erlbaum, Mahwah, NJ.

Hofer, B.K., Pintrich, P.R., 1997. The development of epistemological theories: beliefs about knowledge and knowing and their relation to learning. Review of Educational Research 67 (1), 88–140.

Jehng, J.J., Johnson, S.D., Anderson, R.C., 1993. Schooling and students' epistemological beliefs about learning. Contemporary Educational Psychology 18, 23–35.

Kardash, C.M., Howell, K.L., 2000. Effects of epistemological beliefs and topic-specific beliefs on undergraduates' cognitive and strategic processing of dual-positional text. Journal Educational Psychology 92 (2), 524–535.

Kardash, C.M., Scholes, R.J., 1996. Effects of preexisting beliefs, epistemological beliefs, and need for cognition on interpretation of controversial issues. Journal of Educational Psychology 88 (2), 260–271.

King, P.M., Kitchener, K.S., 1994. Developing Reflective Judgment: Understanding and Promoting Intellectual Growth and Critical Thinking in Adolescents and Adults. Jossey-Bass, San Francisco.

Kizilgunes, B., Tekkaya, C., Sungur, S., 2009. Modeling the relations among students' epistemological beliefs, motivation, learning approach, and achievement. The Journal of Educational Research 102 (4), 243–255.

Kuhn, D., Cheney, R., Weinstock, M., 2000. The development of epistemological understanding. Cognitive Development 15, 309–328.

Muis, R.K., Bendixen, L.D., Haerle, F.C., 2006. Domain-generality and domain specificity in personal epistemology research: philosophical and empirical reflections in the development of a theoretical framework. Educational Psychological Review 18, 3–54.

Murphy, P.K., Buehl, M.M., Monoi, S., Long, J.F., 2002. Understanding the achievement of inner-city adolescents: the influence of epistemological beliefs and achievement goal orientation on academic achievement. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Neber, H., Schommer-Aikins, M., 2002. Self-regulated science learning with highly gifted students: the role of cognitive, motivational, epistemological, and environmental variables. High Ability Studies 13 (1), 59–74.

Nicholls, J.G., 1984. Achievement motivation: conceptions of ability, subjective experience, task choice, and performance. Psychology Review 91, 328–346.

Paulsen, M.B., Feldman, K.A., 1999. Epistemological beliefs and self-regulated learning. Journal of Staff, Program, & Organizational Development 16 (2), 83–91.

Perry, W.G., 1970. Forms of Intellectual and Ethical Development in the College Years: A Scheme. Holt, Rinehart and Winston, New York.

Phan, H.P., 2008. Multiple regression analysis of epistemological beliefs, learning approaches, and self-regulated learning. Electronic Journal of Research in Educational Psychology 6 (1), 157–184.

Qian, G., Alvermann, D., 1995. Role of epistemological beliefs and learned helplessness in secondary school students' learning science concepts from text. Journal of Educational Psychology 87, 282–292.

Qian, G., Burrus, B.M., 1996. The role of epistemological beliefs and motivational goal in ethnically diverse high school students' learning from science texts. In: Leu, D.J., Kinzer, C.K., Hinchman, K.A. (Eds.), Literacies for the 21st Century. National Reading Conference, Chicago, pp. 159–169.

Richter, T., 2010. Epistemological beliefs and epistemic strategies in self-regulated learning. Metacognition and Learning 5 (1), 47–65.

Schommer-Aikins, M., 2004. Explaining the epistemological belief system: introducing the embedded systemic model and coordinated research approach. Educational Psychologist 39 (1), 19–29.

Schommer, M., 1998. The role of adults' beliefs about knowledge in school, work, and everyday life. In: Smith, M.C., Pourchot, T. (Eds.), Adult Learning and Development. Lawrence Erlbaum, Mahwah, NJ, pp. 127–143.

Schommer, M., 1990. Effects of beliefs about the nature of knowledge on comprehension. Journal of Educational Psychology 82, 498–504.

Trautwein, U., Lüdtke, O., 2007. Predicting global and topic-specific certainty beliefs: domain-specificity and the role of the academic environment. British Journal of Educational Psychology 77, 904–934.

Wood, P.K., 1997. A secondary analysis of claims regarding the reflective judgment interview: internal consistency, sequentiality and intraindividual differences in ill-structured problem solving. In: Smart, J.C. (Ed.), Higher Education: Handbook of Theory and Research, vol. XII. Agathon, Edison, pp. 245–314.

Equality and Inequality: Legal Aspects

Susan Sturm, Columbia Law School, Columbia University, New York, NY, USA

Published by Elsevier Ltd.

Abstract

Law's role in advancing equality remains a focal point of attention and controversy in both theory and practice. Many of the recurring disagreements in case law, commentary, and politics can be traced to different theories of inequality, informed by differing conceptions of law's role in its remediation. These theories, in turn, reflect differences in the underlying assumptions about the causes of inequality, the continuing significance of identity and discrimination, the appropriateness of considering race and other social categories in remedying inequality, the appropriate balance between inequality and competing values, and the legitimate role of courts and the state more generally in addressing social problems. Courts in the United States and the European Union increasingly use equality theory to police the boundaries of efforts by other public and private actors to address persistent inequality, particularly racial inequality. This article lays out four major approaches to inequality and how they operate differently in judicial and nonjudicial venues: (1) formal inequality of treatment, (2) structural inequality, (3) outcome inequality, and (4) inequality of access. It includes discussion of novel equality theories that have emerged in legal arenas outside the courts and identifies theories designed to reduce the conflict between equality as defined by the courts and nonjudicial bodies pursuing policies aimed at reducing inequality.

Law's role in advancing equality remains a focal point of attention and controversy in both theory and practice. Explicit exclusion based on race or gender from employment, education, voting, and housing has become widely viewed as morally unacceptable and less common. Yet we live in a world of increasing inequality, one in which a widening income, wealth, health, and opportunity gap intersects with age-old inequalities along racial, gender, ethnic, disability, immigrant, and age lines (Sturm, 2014; Massey, 2007; Sharkey, 2013; Wilkinson and Pickett, 2009). That inequality persists in many significant domains – employment, housing, education, wealth, criminal justice – is uncontroversial. Nonetheless, the legal relevance and actionability of enduring and embedded inequality remains highly contested.

Many of the recurring disagreements in case law, commentary, and politics proceed from the clash between theories of equality, tied to differing conceptions of law's role in its remediation. These theories in turn reflect differences in the assumptions about the causes of inequality, the continuing significance of identity and discrimination, the appropriateness of considering race and other social categories in remedying inequality, the appropriate balance between inequality and competing values, and the proper role of courts specifically and the state more generally in addressing social problems.

The meaning of inequality and scope of remediation also vary depending on the locus of the body elaborating equality norms and remedies. Judicial, legislative, administrative, and nongovernmental bodies have differing conceptions of legitimacy, authority, and state power, and those differences affect the way they engage with issues of equality. In the US and European Union (EU) jurisprudence, courts applying equality principles use equality theory to police the boundaries of efforts by public and private actors to address persistent inequality through explicit and direct action requiring differential treatment, particularly, racial inequality. Most recently, commentators and policy makers have begun exploring frameworks that locate equality within a broader normative framework, decenter the role of the judiciary in advancing equality and embrace hybrid, polycentric strategies for addressing enduring inequality (De Burca, 2006; McCrudden, 1999; Fishkin, 2014; Sturm, 2006a, 2014). This article lays out four major approaches to inequality: (1) formal inequality of treatment, (2) structural inequality, (3) outcome inequality, and (4) inequality of access. It includes innovative theories for reducing inequality that have emerged in legal arenas outside the courts, and identifies theories designed to reduce the conflict between equality as defined by the courts and equality as defined in nonjudicial arenas pursuing policies aimed at reducing inequality. It also contrasts current US antidiscrimination law with more proactive legislative and administrative strategies developing in Europe and in nonjudicial arenas in the US.

Formal Inequality of Treatment

The baseline conception of inequality in the US and EU antidiscrimination law is embodied in principle of equal treatment. The law pursues equality by eliminating race, gender, or other suspect social categories as a legitimate basis of classifying people for the purpose of allocating resources or making decisions (Collins, 2003). Sometimes called the anticlassification principle (Fiss, 1971; Siegel, 2004), formal equality focuses on the state of mind of the decision maker; it targets intentional reliance on race, gender, or another protected category as the violation of the equality norm. It is process oriented, in the sense that it regards the taking account of a person's identity in the treatment of students, voters, employees, prospective homeowners, and people more generally, as both the harm and the wrongful conduct.

The hallmark narrative of formal equality has as its protagonist individuals who deliberately exclude or marginalize other individuals based on their race, gender, sexual orientation, or other protected classification. This model defines the wrongful conduct as intentional, unequal treatment based on race, gender, sexual orientation, age, or some other

class-based characteristic that is viewed as arbitrary. For inequality to count as legally relevant under this theory, it must be demonstrably caused by this form of deliberate, class-based exclusion.

The premise of this model is that race, gender, and other class characteristics are irrelevant to the underlying decision, and that taking class membership into account is thus arbitrary and unfair (Fiss, 1971; Fallon, 1984). The goal of the formal equality model is 'color blindness': decisions affecting people's status and opportunity should be made without regard to social categories such as race or gender. Race, gender, and other categories of status operate as fixed, universal, and static categories based on membership in a set, unchanging group identified by predetermined, and often physical, characteristics such as skin color or sex. The assumption is that group status is, by definition, irrelevant to decisions based on merit, that considering class status is by nature arbitrary, denigrating, and unfair, that decision makers can make decisions without noticing or taking account of racial and gender characteristics, and that equality can be achieved by a rule that prohibits decision makers from considering race or gender.

The formal equality norm rests on a principle of the dignity of the individual. For those who subscribe to formal equality as an umbrella conception, 'stigma' is the crucial element in analyzing racial classifications. As the US Supreme Court recently stated, "To be forced to live under a state-mandated racial label is inconsistent with the dignity of individuals in our society" (Parents Involved in Community Schools v. Seattle School District No. 1, 127 S. Ct. 2738 (2007)). The Supreme Court recently extended its concern with preserving individual dignity and avoiding stigmatization to the context of sexual orientation, invalidating as inconsistent with equality principles a law that "places same-sex couples in an unstable position of being in a second-tier marriage. The differentiation demeans the couple . . . and it humiliates tens of thousands of children now being raised by same-sex couples" (United States v. Windsor, 133 S. Ct. 2675 (2013)). Canadian, French, and the EU law similarly bases prohibitions against disparate treatment on dignitary values, often drawing directly from United States antidiscrimination laws of the 1970s (De Burca, 2012; Moreau, 2010). For example, the European Convention for the Protection of Human Rights and Fundamental Freedoms prohibits discrimination "on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status," and grounds this equal treatment mandate in the "independent value of respect for individual dignity" (Collins, 2003; Tremblay, 2012). Proponents of formal equality also identify racial polarization as a concern underlying color blindness jurisprudence (Siegel, 2011; Yoshino, 2011): "All state-imposed classifications that rearrange burdens and benefits on the basis of race are likely to be viewed with deep resentment by the individuals burdened. The denial to innocent persons of equal rights and opportunities may outrage those so deprived and therefore may be perceived as invidious" (Regents of the University of California v. Bakke, 438 U.S. 265, 294 n. 34 (opinion of J. Powell)). Formal inequality theory provides remedies to individuals who have been subjected to a process tainted by racial or gender considerations. It is decidedly ahistorical and individualistic in its focus, and compensatory in its explicit remedial approach. Subscribers to formal equality view race and gender as biological or immutable categories that also operate as "a fixed yet radically empty feature[s] of every person's identity" (Siegel, 2000: p. 91). According to this view of race, in particular, the category has no socially relevant content (Siegel, 2000).

Many who embrace a formal equality approach believe current racial inequality results from factors unrelated to racial or gender discrimination, or at least to factors that lie beyond the constitutionally defined boundaries of equality. The election of Barack Obama, a man of mixed Caucasian and African heritage, to the US presidency in 2008 has prompted some conservatives to suggest that the United States has moved to a postcivil rights society, in which race no longer operates in a manner that is actionable through the courts. "Post-racialism is a set of beliefs that coalesce to posit that racial discrimination is rare and aberrant behavior as evidenced by America's and Americans' pronounced racial progress" (Barnes et al., 2010: p. 968). One critical commentator summarized the view that the mechanisms of discrimination can be reduced to a simple paradigm – explicitly discriminatory policies and rogue individuals: "Discrimination is over. The market is bias-free. The law's job is to find the aberrant actor who just didn't get the memo" (Gertner, 2012: p. 111).

Commitment to race blindness is not limited to the US. For example, French law insists on race neutrality and "leaves no room for the possibility of justifying racial distinctions to remedy past racial harms or to promote diversity, as U.S. law does" (Suk, 2007: p. 13). Suk traces the color blindness norm in France in part to conceptions of equality and citizenship that regard social distinctions of any sort as a threat to republicanism (Suk, 2007).

Formal equality embraces the idea that decision makers such as employers and educational institutions may, and usually do, properly differentiate among people based on qualifications relevant to capabilities and performance. By taking social categories such as sex or race into account, decision makers unjustifiably depart from merit-based decision making. This violates basic notions of fairness because people deserve to be treated 'on the merits,' based on an individualized assessment of their capacities (Fallon, 1984; Gutmann, 1996; Suk, 2007). Exclusion based upon characteristics that are beyond the control of the affected individual are viewed as particularly morally problematic (Fiss, 1976; Brest, 1976). Decisions based on class membership thus offend dignitary values, in addition to undermining instrumental values of efficiency. Perhaps the most vehement expression of this view in the US can be found in Justice Scalia's concurring opinion invalidating voluntary affirmative action in government contracting based upon its inconsistency with the principle of color blindness: "To pursue the concept of racial entitlement—even for the most admirable and benign of purposes—is to reinforce and preserve for future mischief the way of thinking that produced race slavery, race privilege, and race hatred. In the eyes of government, we are just one race here. It is American" (Adarand Constructors vs. Pena 115 S. Ct. 2097 [1995]).

Those who embrace formal inequality may disagree about the degree of enduring inequality, but they share skepticism about the legitimacy and efficacy of efforts to address this form of inequality by taking race into account as a factor in allocating

benefits and opportunities to individuals (Rappaport, 2013; Suk, 2007; Browne, 1993; Rabkin, 1997; Sowell, 1984). Courts properly focus on remedying wrongful acts by perpetrators intending to exclude people based on their race, age, sex, or sexual orientation. Efforts to remedy inequality by making race a basis for individual advancement are deeply suspect under this approach. Justice Roberts restates his commitment to the color blindness principle in the context of higher education admissions: "The way to stop discrimination on the basis of race is to stop discriminating on the basis of race" (Parents Involved in Community Schools v. Seattle School District No. 1, 30). Formal equality principles increasingly operate to challenge the legality of efforts to remedy enduring inequality, particularly, in the context of affirmative action, voting rights, school desegregation, and other voluntary state efforts to reduce persistent, group-based inequality. (*see* Affirmative Action: Comparative Policies and Controversies; Affirmative Action: Empirical Work on Its Effectiveness).

Critiques of Formal Equality

The dominance of formal equality in American antidiscrimination law has produced a wide array of criticism. The most prevalent concern proceeds from the observation that color blindness is an inadequate response to continuing inequality. Most fundamentally, critics point out that a formal equality theory leaves intact a legacy of state-produced inequality and subordination. It ignores the role that group status continues to play in shaping the life chances, opportunities, and experiences of group members. It locks in the continued effects of past formal exclusion, imposing the consequences of that continued exclusion on the excluded groups themselves. By defining equality solely in terms of the conduct of the perpetrator, a formal equality approach normalizes conditions of marginalization and exclusion and the group dimensions of these material conditions. It does not begin to address the perspective of those who experience exclusion, marginalization, and unequal enjoyment of resources and status (Crenshaw, 1988; De Burca, 2012; Gotanda, 1991; powell, 2009). At the same time, "It conceals the distributive consequences of group-salient practices in a semantic code that defines instrumentally rational utterances as race-neutral" (Siegel, 2000: p. 101). "Acceding to a worldview on which racial inequity is primarily the product of present bad actors rather than largely a matter of historically embedded hierarchies fosters the misdescription of a central social problem and therefore helps make it less likely that the problem will be addressed through appropriate means" (Primus, 2003: p. 499).

On a conceptual level, David Strauss has pointed out the logical inconsistencies of color blindness as an antidiscrimination principle. To protect a designated group against treatment targeting that group, the state must pay attention to and indeed specially protect the interest of that group. Thus, color blindness is contradictory. Strauss also points out that the metaphor of color blindness really does not fit its use in the nondiscrimination context. Color blindness in the literal sense prevents an individual from seeing color; color blindness as an equality model prevents an individual from acting on what he or she sees. This approach presumes (incorrectly) that individuals can see race but keep it from influencing their perceptions, conduct, and decisions (Strauss, 1986).

Recent work in the social sciences challenges the assumption that conscious decision making adequately describes the current dynamics of exclusion and unequal treatment. Bias is produced not only as a result of motivation but also as a product of racial or gender schemata, multiracial group interactions, unequal access to networks, and patterns of conflict resolution or avoidance (Bagenstos, 2006; Dasgupta, 2004; De Burca, 2006; Krieger, 1995; Sturm, 2001). Bias or exclusion often results from interactive patterns of noninclusion, as much as from exclusion motivated by bias (Sharkey, 2013; Wilkins and Gulati, 1996). Members of nondominant groups may well be treated differently based on their race, notwithstanding the decision maker's lack of intention to do so.

Historians have also challenged the theory of formal equality for its ahistoricism. Color blindness ignores historically constituted patterns of group inequalities that are "socially pervasive (articulated across social domains), and socially persistent (articulated over time)" (Siegel, 2000: p. 82). Examining race and gender inequality from a historical standpoint reveals relations of group inequality embedded in the social organization of work, reproduction, and sexuality (Siegel, 2000). "In actuality, the United States has been plagued with … the consistent use of racial classifications to privilege (certain groups of) whites and disadvantage nonwhites" (Harris, 2000, 2009). The focus on classification itself – because it makes suspect race-conscious action taken to combat racism – justifies the maintenance of existing social and political relations (Harris, 2000). As Kenji Yoshino has noted, "an equal protection jurisprudence that turns formalistically on facial discrimination will, from an antisubordination perspective, get it exactly backward. On the one hand, this jurisprudence invalidates affirmative action programs seeking to aid historically subordinated groups. On the other hand, it upholds second-generation discrimination that continues to subordinate those groups" (Yoshino, 2011: p. 767). Similar critiques have been made of color blindness and equal treatment jurisprudence in Europe, there by yielding more far-reaching adoption of more proactive and structural approaches in the regulatory arena (Suk, 2007; De Burca, 2012).

Robert Post applies this structural account of the dynamics of bias to the law itself. Using what he calls a sociological account of the field, he contends that law does not and cannot cause employers to make judgments about qualifications in a way that is totally race and sex blind. Legal actors are themselves engaged in a social practice of shaping the norms and conventions by which sex and race are given social meaning. "Antidiscrimination law always begins and ends in history, which means that it must participate in the very practices that it seeks to alter and to regulate" (Post, 2000: p. 17). To pretend otherwise is to allow courts to rely on unarticulated assumptions and premises that legitimate some racial or gender meanings and challenge others.

A final critique relates to the threat to the court's legitimacy stemming from the gap between inequality as recognized by law and inequality as people experience it (Sturm, 2006b; Sturm, 2014). As that gap grows, the refusal of law to acknowledge persistent inequality undermines legitimacy of the law as an instrument of justice. This legitimacy deficit is

compounded when the court not only defines equality narrowly but proscribes the role of other branches of government or private entities in identifying and addressing more structural dimensions of persistent inequality.

Similar observations of persistent inequality left unaddressed by antidiscrimination strategies have been made in Europe. However, European and the US antidiscrimination law diverge in their response to these shared critiques of formal equality. Despite the fact that "core concepts and provisions of EU anti-discrimination law … were drawn from the United Kingdom's anti-discrimination law of the 1970s, … [which] were in turn strongly influenced by U.S. law at the time" these legal principles served as a predicate for dramatic expansion of European anti-discrimination laws through "new legislation and treaty provisions, new institutions such as the EU Fundamental Rights Agency, renewed civil society activism" (De Burca, 2012: p. 3). This difference has been traced to a shared European idea that "the provision and promotion of the human and social welfare of their peoples is a core aspect of the responsibility of the states" (De Burca, 2012: p. 7), in contrast with the US legal and political culture, which privileges the idea of individual freedom (De Burca, 2012). It is also possible that these differences stem in part from the fact that "the United States is several decades ahead of the European Union" in the evolution of its antidiscrimination law and policy so that "Europe has not yet experienced the powerful backlash that probably awaits an ambitious program" (De Burca, 2012: p. 10).

Structural Inequality of Treatment

A second approach focuses on structural, situational, and relational dynamics that produce and sustain inequality. Referred to as addressing second, third, and fourth generations of discrimination (Fredman, 2012; Sturm, 2001), this theory provides a framework that continues to account for inequality as a result of unintentional and embedded inequality as well as intentional biased treatment. It broadens the focus of analysis from formal and deliberate decisions to the social context and practices that produce decisions, and from individuals to groups and the contexts that shape their mobility, status, and circumstances.

Structural inequality analyses rest in part on goals similar to those of formal inequality theories: fair treatment and dignity. Differences in treatment between majority and minority group members that are arbitrary, in the sense that they are not justified by criteria of economic rationality, violate this principle of equality. This approach questions the fairness assumption underlying formal equality theory's meritocracy principle to the extent that it presumes that "qualifications are set that are relevant to the legitimate social purposes of the position in question" (Gutmann, 1996: p. 126). If differences in treatment are attributed to stereotyping, cognitive bias, power differentials, or differential access to the networks that structure advancement, they violate this fairness principle, even if the decision maker did not consciously intend that result. The harm is not only the stigma or loss of dignity that results from substituting group characteristics for individual assessment. It is also the difference in access that results from patterns of exclusion that emerge from social interaction or cognitive bias (Directorate General of Human Rights, 2004; Krieger, 1995; Sturm, 1998).

Structural inequality theorists go beyond the individualistic versions of these values to connect equality concerns to consider group-based issues of access and opportunity, and some connect equality values to more basic human concerns of diversity, equity, 'full participation,' 'belonging,' civic agency, and democratic legitimacy (powell, 2012; Sturm 2014). Gender mainstreaming policies in Europe embrace structural approaches to addressing inequality, and define equality positively to mean "an equal visibility, empowerment and participation of both sexes in all spheres of public and private life" in order to promote "the full participation of women and men in society" (Directorate General of Human Rights, 2004: p. 8; De Burca, 2006). They strive for the "(re)organization, improvement, development and evaluation of policy processes, so that a gender equality perspective is incorporated in all policies at all levels and at all stages, by the actors normally involved in policy making" (Directorate General of Human Rights, 2004: p. 12).

Structural inequality conceptions rest on a range of causation theories. One cluster continues to focus on the mind-set of decision makers, but shifts the focus from conscious and intentional bias to unconscious bias and stereotyping. People categorize along racial, gender, and ethnic lines, and these cognitive processes lead us to perceive and judge people differently, in ways that have a direct effect on employment, housing, admissions, and other decisions allocating opportunities and resources (Krieger, 1995; Bagestos, 2009; Sturm, 2001). Though seemingly inconsistent with strict formal equality, the US Supreme Court's opinion in *Hibbs* locates constitutional harm in 'firmly rooted' "stereotype-based beliefs about the allocation of family duties" that operate to the disadvantage of women in "situations in which work and family responsibilities conflict." Policies requiring employers to provide or adjust family leave have been justified as a means "to eliminate their reliance on and perpetuation of invalid stereotypes, and thereby dismantle persisting gender-based barriers to the hiring, retention, and promotion of women in the workplace" (*Hibbs*, 123 S. Ct. at 1981 n.10).

Structural inequality also can include a more institutionally embedded causal theory, focused on ways that inequality resides in the institutions and communities that structure people's lives (McCrudden, 1999; De Burca, 2006). These dynamics are multilevel, crossinstitutional, and nontechnical, embedded in cultures and routines, and often hard wired into policies and structures (Green, 2003; Sturm, 2001, 2014). They occur through cumulative advantages and disadvantages over a long period of time and produce isolation or marginalization from the experiences and relationships crucial to mobility and achievement. john powell has described the 'grammar of durable racial inequality' through the lens of racialization, which means "the set of practices, cultural norms, and institutional arrangements that are both reflective of and simultaneously help to create and maintain racialized outcomes in society" (powell, 2009: p. 785). Structural approaches combine attentiveness to process and outcomes, examining the mechanisms and barriers that

shape whether people of different races have "the wherewithal to succeed, thrive and contribute" and emphasize reflection, capacity building, and networks as strategies for improvement (Sturm et al., 2012; Sturm, 2006b).

This form of inequality is often difficult to identify at the level of individual decisions. For this reason, structural equality of treatment approaches often focus on patterns of unequal treatment affecting groups of similarly situated individuals. Unexplained disparities in treatment of similarly situated members of majority and minority groups are problematic under this approach (*Teamsters vs. United States*, 431 US 324 [1977]; De Burca, 2012; Lenhardt, 2011). Fault, in the sense of conscious intent to exclude based on irrelevant and suspect classifications, is not necessarily a criterion for discrimination.

Structural inequality theorists disagree about the appropriate remedial role of the judiciary, with European law showing more openness to structural definitions of equality at least in some areas (De Burca, 2012; de Witte, 2012). Some commentators in the US advocate expanding judicial doctrine to include unconscious bias and structural discrimination (Green, Krieger). Some question the viability of relying on the courts to define these equality norms. Some have advocated a shift to liberty principles or universalist goals as the appropriate strategy for tackling structural inequality and encouraging cross-racial collaboration (Yoshino, 2011). Others believe the courts retain a more limited but still-significant role in tackling structural inequality – one focused on "prompting—and creating occasions for—normatively motivated inquiry and remediation by non-legal actors in response to signals of problematic conditions or practices" (Sturm, 2006: p. 47; see also Lenhardt, 2011).

Some structural inequality theorists and policy makers have looked beyond the courts to other arenas that have more latitude and capacity to catalyze institutional and systems change, such as administrative agencies and hybrid public/private accountability systems, as locations for reducing structural inequality. For example, Olatunde Johnson has identified a series of federal 'equality directives' in the US that use administrative tools of conditioned spending, policy making, and oversight to promote substantive inclusion with respect to race, ethnicity, language, and disability. These measures require states to take proactive, front-end affirmative measures to address structural barriers to equality (Johnson, 2012). Susan Sturm has documented the role of public/private collaborations that embrace 'full participation' as a way to address structural inequality in particular communities and institutions through public policy aimed at institutional transformation (Sturm, 2014). john powell has put forward the role of law in redesigning policies and systems to connect underserved communities and their residents "to the opportunity structures that improve well-being and pave the path to success" (powell, 2012: p. 19).

In Europe, mainstreaming commissions and administrative bodies have advocated and in some countries embraced far-reaching structural approaches to advancing equality, intended to "change the structures in society which contribute to maintaining the unequal power balances between women and men" (Directorate General of Human Rights, 1998: p. 8). These initiatives locate gender and racial equality within broader principles of democracy and human rights (Directorate General of Human Rights, 2004; De Burca, 2012).

Critiques of Structural Inequality

Some critics question the practicability of this structural move through judicial intervention, given the courts' and politicians' reluctance to go beyond formal inequality as the normative underpinnings of judicial intervention (Bagenstos, 2006; Fredman, 2012). In the US, doctrinal limitations and deference to decision makers such as employers and government bodies, along with recent cut-backs on the availability of class actions, further limit the viability of judicial remedies for structural discrimination (Bagenstos, 2006; Gertner, 2012). Libertarian critics question whether courts should become enmeshed in addressing patterns that are not the responsibility of an individual and are pervasive. If cognitive bias is widespread, indeed, ubiquitous, is it appropriate to rely on the law to change embedded patterns of interaction? (Wax, 1999) Scholars concerned about realizing rights in practice note that this form of bias is embedded in cultures and organizational structures, and courts are limited in their capacity and willingness to disrupt these patterns.

Conservative critics have questioned the viability of a model of discrimination that, by its nature, can only be detected by looking at patterns involving groups over time (Browne, 1993; Rabkin, 1997). Some have also questioned the empirics; they are particularly concerned about extending laboratory-based findings of implicit bias to account for bias in the workplace (Tetlock and Mitchell, 2009; Wax, 1999). Critics from the left have also questioned discrimination theory, particularly unconscious bias, as too limited an explanation because it prompts superficial strategies that leave the underlying structures of durable inequality unchanged (Banks and Ford, 2009; Loury, 2000; Wilson, 1996). Courts themselves are subject to these biases, and are unlikely to find discrimination in the absence of evidence of animus (Gertner, 2012). Unconscious bias discourse "reinforces a misguided preoccupation with mental state, and perpetuates an obsession with antidiscrimination law, rather than policy reform, as a means of realizing racial justice goals … The best political approach, over the long term, is to straightforwardly define and defend policy goals, and then figure out how to achieve them" (Banks and Ford, 2009: p. 1059). Scholars and policy makers have questioned whether structural inequality alone provides a theory of change sufficient to move beyond critique and to construct the institutions and settings that reduce inequality and to enable people who are currently marginalized to thrive (Loury, 2000; Sturm, 2014).

Critics of noncoercive strategies for reducing structural inequality in both the US and Europe also question the robustness of their implementation and their capacity to address durable inequality in arenas that are most stratified and least likely to cooperate in process-oriented equality regimes (Fredman, 2012; Selmi, 2005). The duty to 'have due regard' for equality in Europe or to exercise reasonable care in US sexual harassment cases has been applied with sufficient

deference to public bodies to produce cosmetic rather than substantive change (Fredman, 2012; Krawiec, 2003). There is a fear that gender mainstreaming has been introduced "at the expense of specific measures which focus on long-term, systemic and structural inequalities faced by women in the labour market" (Shaw, 2005: p. 262).

Inequality of Outcome

This concept of inequality focuses on differences in the material conditions of members of particular groups. It is rooted in social reality and looks at the distribution of jobs, income, and status of members of groups that have previously been subjected to systematic exclusion and subordination (Fredman, 2012; Fiss, 1976; Fallon, 1984). The focus of this theory is the group, rather than the individual who has been excluded as a result of group status. Groups continue to matter because they define the opportunity structures affecting their members. For these theorists, group status must be directly addressed because it is the level at which inequality operates. For some, groups also have independent normative value, reflected in the patterns of cultural and political meaning that emerge among individuals who identify as a group (Gotanda, 1991). These theorists justify the focus on groups not only in instrumental terms, but also as an independent value.

The empirical premise of outcome inequality theory is that current inequality continues to be the result of systematic and legally sanctioned subordination. Legal intervention is justified to remedy persistent and arbitrary differences in resources, social status, and opportunity based on group membership. The current system for allocating opportunities continues to reflect biased criteria that systematically disadvantage nondominant groups. Thus, group-based remediation compensates for departures from merit-based decision making (Fiss, 1971; Harris and Narayan, 1994). The goal of this approach to equality is outcome oriented: disparities in group status should be reduced or eliminated. This conception of inequality has also been called the antisubordination principle: the conviction that it is wrong for the state to engage in practices that enforce the inferior social status of historically oppressed groups (Fiss, 1976; Siegel, 2004). Equal outcome theories have been recently advanced through a comparative perspective that contrasts the United States' aversion to quotas with the use of quotas as a way to advance democratic legitimacy in other countries, such as Brazil and France (Suk, 2012).

The normative justifications for the equal outcomes theory are either compensatory and remedial or grounded in theories of democratic legitimacy. Past group-based subordination justifies current group-based remediation. This theory rests on the premise that the group continues to have social meaning. The group continues to experience harm that is a consequence of the history of group-based exclusion, and members of the group experience the current manifestations of that harm. For this reason, matching harm to individual injury is thus both unnecessary and counterproductive (Crenshaw, 1988). Moreover, group-based remediation is necessary to overcome barriers to participation that are perpetuated through facially neutral processes and criteria (Siegel, 2000; Abrams, 1989). Reducing the continuing effects of past group-based discrimination has served as a justification for using affirmative action in employment, requiring preclearance of voting systems, and overturning restrictive voter registration requirements and gerrymandering, which have produced systematic underparticipation of protected groups in the political process (Sturm, 2001; Guinier, 1994).

Other justifications for the outcomes theory of discrimination are more pragmatic: bias is difficult to prove and process rules are easy to evade. Outcome measures serve as an adequate proxy for bias that continues to affect access to fair process and opportunity. If a process is fair and race or gender neutral, it will produce a result that more or less mirrors the composition of the pool (Suk, 2012). Outcome measures that strive for a workforce that mirrors the racial or gender composition of the pool simply accomplish directly what may be more difficult to achieve through reforming the decision-making process. These measures also provide an easy-to-implement remedy that responds to concerns about judicial capacity to detect and correct patterns of bias (Fiss, 1971; Suk, 2012).

Finally, some theorists link representation to theories of democratic legitimacy, drawing on the language relating to diversity to highlight the importance of enhancing parity in representation among groups that have been marginalized from public participation (Suk, 2012). Unlike the US, Europe has shown greater openness to a 'parity democracy' paradigm that uses quotas to increase participation of women in political and corporate decision making, based on an understanding of inequality as a problem of democracy and citizenship (Suk, 2012; Robio-Marin, 2012).

Critiques of Outcome Theories of Equality

Adherents of formal equality criticize outcome-based theories as inconsistent with the principle of race neutrality; they believe it offends dignitary and fairness values by allocating benefits based on an irrelevant and demeaning category. These approaches are seen as advancing quotas, which fail to recognize people as individuals and invite racial polarization and balkanization (Yoshino, 2011; Suk, 2012; Siegel, 2011). Some critics take the view that any approach that treats race or gender as a salient category, even for remediation, reinforces the continued significance of those categories and is thus discriminatory. They also question the appropriateness of remedial strategies that focus on groups. This skepticism proceeds in part from an increasing willingness to question the empirical assumption that current inequality is the result of past legal subordination. Since the 1970s, some scholars have developed alternative explanations for persistent inequality that emphasize characteristics of the racial or gender group, rather than patterns of exclusion or marginalization (Glazer and Moynihan, 1975; Sowell, 1984). Once the causal link between current economic or social inequality and legal exclusion or subordination is questioned, skepticism undercuts the perceived legitimacy of group-based remedies. Indeed, some critics of outcome theories of equality question whether racial or gender status continues to have any meaning at the level of the group. Some conservative critics thus suggest that group-based remediation is just another form of racial or gender-based decision making, and is thus illegitimate

(Clegg, 2013; Belz, 1991; Abram, 1986; Scalia, 2007; Roberts, 2007). Finally, they treat outcome-based equality theories as inconsistent with principles of merit-based decision making and thus contrary to both efficiency and dignitary values (Browne, 1993).

Outcome theories have also been criticized by progressives as an inadequate framework to define equality policy because they rest so heavily on pragmatic considerations and lack an adequate normative foundation. Outcome theorists typically view their strategies as contingent and temporary. They do not generally defend group-based redistribution as an end in and of itself, but rather as a means to an end. Nor do they reveal what it was about current practice that is objectionable, other than the result. The absence of a clearly articulated normative justification invites challenges to the legitimacy of any efforts to address racial or gender bias. Outcome approaches, if used as the exclusive theory justifying equality initiatives, may thus stunt the refinement, rearticulation, and reformation of the normative meaning of racial and gender participation and exclusion. They may also reinforce the perception that equality and efficiency are competing goals (Sturm, 2001; Sturm and Guinier, 1996).

Inequality of Access

Unequal access theories focus on inequalities that result from unfair criteria and structures of interaction. They target decision making that is formally fair but functionally biased in favor of members of a particular group. The central question is whether the structure, process, or criteria of decision making arbitrarily favors members of one group over another for reasons unrelated to the articulated goals underlying the decision (Sturm, 2001).

The most widely articulated conception of equal access theory accepts the moral legitimacy of differentiating among applicants based on meaningful differences in their qualifications to carry out the functions of a position. Reasonable disagreements exist, however, concerning the social purposes relevant to employment or admissions, the relative importance of multiple purposes, and the best way to assess, which applicants meet those qualifications. "Those members of society who now have the power to set qualifications may tend to value those qualifications that favor people similar to themselves" (Gutmann, 1996: p. 124). Moreover, these power differentials shape the structures for assessing individuals' qualifications and opportunities for advancement. These standards and structures are no more likely to produce valued social outcomes than others that do not exclude members of nondominant groups. As such, these exclusionary standards and practices are unfair, and thus discriminatory (Strauss, 1998).

The classic case of *Griggs vs. Duke Power Co.* (401 US 424 [1971]) lays out this theory through a fable: "Congress has now provided that tests or criteria for employment or promotion may not provide equality of opportunity merely in the sense of the fabled offer of the milk to the stork and the fox." Discrimination thus occurs when an employer relies upon conditions, criteria, or practices that favor members of a particular group, if that advantage is arbitrary in the sense that it does not advance the articulated goals of the employer.

This theory defines discrimination as conduct that deprives individuals or groups of meaningful access to educational or employment opportunities through the use of standards, criteria, or structures that fail to minimize the expression of bias, or that arbitrarily advantage one group over the other (Friedman, 1986). It builds on the practical insight that discrimination and inequality often result from arbitrary, unstructured, and unaccountable decision making. It uses outcome measures not to define the harm, but to signal the likelihood that an employment or admission process is biased, and thus to trigger scrutiny of the adequacy and fairness of the criteria and their application. The beneficiaries of this theory are both individuals, who are denied the opportunity to compete based on unbiased selection processes, and groups, which are systematically disadvantaged in their access to jobs or colleges.

Equal access goals have been part of the equality story in the US from the time of the original civil rights act, which at its introduction was defined to assure that, in President Kennedy's words, "public funds, to which all taxpayers of all races contribute, not be spent in any fashion which encourages, entrenches, subsidizes or results in racial discrimination" (Johnson, 2012: p. 26). Olatunde Johnson has documented the adoption 'equality directives' in variety of legal and policy settings, which require public and private actors to examine their practices to identify disparate impact and design inclusionary alternatives (Johnson, 2012).

Equal access theory has thus emerged as an integrative approach to discrimination that accepts the inevitability, if not the legitimacy, of merit-based distinctions but introduces structural and group notions of fairness to the question of how to define and implement merit-based decision making. Equal access theory has been overlooked in many interpretations of disparate impact's normative underpinnings. For example, Richard Primus offers two competing interpretations of disparate impact: reducing conscious or unconscious discrimination or increasing integration. He overlooks the equal access theory that most neatly matches the language, because it is really about advancing merit-based decision making and removing barriers that arbitrarily bias workplaces and higher educational opportunities in favor of particular groups.

The US and European conceptions of equal access seem to be moving in opposite directions. The US courts have been narrowing their definitions and approach to disparate impact, while the European Court of Human Rights has begun to "recognize the notion of indirect discrimination" (Besson, 2012; De Burca, 2012: p. 20).

Critiques of Equal Access Theory

Equal access theory has been called both over- and under-inclusive. It's potential for requiring radical restructuring of schools or workplaces prompts challenges based on the intrusiveness of such a theory into the autonomy of employers, admissions officers, and government officials. This criticism often proceeds from the view that courts should not engage in the process of institutional restructuring that they act illegitimately if they move beyond the role of dispute resolution and

compensation for discrete individual harms. The under-inclusiveness critique challenges the acceptance of merit or market criteria as a morally defensible basis for distinguishing between the opportunities of potential employees, college students, voters, etc. The weaker version of this critique, preferred by outcome theorists, argues that the continued pervasiveness of a caste system so infects the distribution of opportunities as to delegitimate a system of merit that produces racially identifiable results (Strauss, 1998). The stronger version of this critique challenges the moral legitimacy of differentiating access to social goods based on any arbitrarily distributed characteristics, including talents and motivation. "There is no more reason to permit the distribution of income and wealth to be settled by the distribution of natural assets than by historical and social fortune …. From a moral standpoint the two seem equally arbitrary" (Rawls, 1974: p. 74).

Judicial reluctance to embrace equal access in its robust form, focusing on discrete practices, has led some legal theorists and policy makers to question disparate impact as a viable theory for addressing persistent inequality (Bagenstos, 2006). For example, Michael Selmi concludes that disparate impact "has had a strikingly limited impact outside of the context of written employment tests and is, in fact, an extremely difficult theory on which to succeed." He contends that "whatever gains the disparate impact theory has produced likely could have been obtained through other means, particularly in large urban cities, and that the theory may have had the unintended effect of limiting our conception of intentional discrimination" (Selmi, 2006: p. 701). Many scholars, policy makers, and advocates are looking beyond the courts for arenas where barriers to mobility, access, and opportunity can be identified and challenged in a way that takes account of the mechanisms producing and reducing persistent inequality.

The differences in approach to equality summarized in this article recur in and help to explain deep fractures in judicial doctrine, legislative policy making, and public discourse about equality. Current debates often pose formal equality of treatment and equal outcome as the only two available conceptions of equality and force a choice between them. In the US, the complex dynamics of bias, along with the decreasing political and judicial viability of outcome and access theories of equality, have increased the attention to and importance of structural inequality, full participation, and unequal access theories of equality developed outside the courts. These theories are, however, the least normatively and empirically developed in the US. New governance approaches in Europe have opened up experimentation with positive theories of equality in the context of regional and supranational organizations, but these have yet to produce robust judicial enforcement or accountability systems. Emerging policy and scholarship emphasizes the decentering of courts as the primary elaborator of equality norms, the importance of nonjudicial venues for elaborating equality norms, and the emergence of equality strategies linked to broader normative goals. It remains to be seen whether courts and legal actors will respond dynamically to these new understandings of the dynamics and meaning of inequality and its relationship to law.

See also: Affirmative Action: Comparative Policies and Controversies; Affirmative Action: Empirical Work on Its Effectiveness; Discrimination and the Law; Equality: Philosophical Aspects; Injustice: Legal Aspects; Rights: Legal Aspects.

Bibliography

Abrams, K., 1989. Gender discrimination and the transformation of workplace norms. Vanderbilt Law Review 42, 1184–1248.

Aleinikoff, T.A., 1990. Case for race consciousness. Columbia Law Review 91, 1060–1125.

Andrews, M., 1999. The Political Economy of Hope and Fear: Capitalism and the Black Condition in America.

Bagenstos, S., 2006. The structural turn and the limits of antidiscrimination law. California Law Review 1, 13.

Banks, R.R., Ford, R.T., 2009. (How) does unconscious bias matter? Law, politics and racial inequality. Emory Law Journal 58, 1053.

Barnes, M.L., Chemerinsky, E., Jones, T., 2010. A post-race equal protection? The Georgetown Law Journal 98, 967–1004.

Barthlett, K.T., Harris, A.P., 1998. Gender and Law: Theory, Doctrine, Commentary, second ed. Aspen Law & Business, New York.

Belz, H., 1991. Equality Transformed: A Quarter-Century of Affirmative Action. New Brunswick.

Browne, K., 1993. Statistical proof of discrimination: beyond "damned lies". Washington Law Review 68, 477–558.

Besson, S., 2012. Evolutions in non-discrimination law within the ECHR and the ESC systems: it takes two to tango in the council of Europe. American Journal of Comparative Law 60, 147.

Brest, P., 1976. In defence of the antidiscrimination principle. Harvard Law Review 90, 1–54.

Clegg, R., 2013. Commentary on Fisher: better off than we were a year ago. SCOTUS blog. Available at. http://www.scotusblog.com/2013/06/commentary-on-fisher-better-off-than-we-were-a-year-ago/.

Collins, H., 2003. Discrimination, Equality and social inclusion. The Modern Law Review 66, 16–43. http://dx.doi.org/10.1111/1468-2230.6601002.

Crenshaw, K., 1988. Race, reform, and retrenchment: transformation and legitimation in antidiscrimination law. Harvard Law Review 101, 1331–1387.

Dasgupta, N., 2004. Implicit ingroup favoritism, outgroup favoritism, and their behavioral manifestations. Social Justice Research 17 (143), 157.

De Burca, G., 2006. EU race discrimination law: a hybrid model? In: De Burca, G., Scott, J. (Eds.), Law and New Governance in the EU and the US 97.

De Burca, G., 2012. The trajectories of European and American antidiscrimination laws. The American Journal of Comparative Law 60, 1–22.

Directorate General of Human Rights, 1998. Gender Mainstreaming: Conceptual Framework, Methodology and Presentation of Good Practices. Final Report of Activities of the Group of Specialists on Mainstreaming.

Directorate General of Human Rights, 2004. Gender Mainstreaming: Conceptual Framework, Methodology and Presentation of Good Practices. Final Report of Activities of the Group of Specialists on Mainstreaming.

Edsall, T.B., Edsall, M.D., 1992. Chain Reaction: The Impact of Race, Rights, and Taxes on American Politics. W.W. Norton & Co., New York.

Epstein, R., 1992. The Case against Employment Discrimination Law. Harvard University Press, Cambridge.

Fallon, R., 1984. Firefighters v. Stotts: conflicting models of racial justice. Supreme Court Review 1, 1–25.

Fishkin, J., 2014. Bottlenecks: A New Theory of Equal Employment Opportunity. Oxford University Press, USA.

Fiss, O., 1971. A theory of fair employment laws. University of Chicago Law Review 38, 235–314.

Fiss, O., 1976. Groups and the equal protection clause. Philosophy & Public Affairs 5, 107–177.

Fredman, S., 2012. Breaking the mold: equality as a proactive duty. The American Hournal of Comparative Law 60, 265.

Friedman, J., 1986. Redefining equality, discrimination, and affirmative action: the access principle. Texas Law Review 65, 41–99.

Gertner, N., 2012. Losers rules. Yale Law Journal On Line 122, 109. Available at: http://yalelawjournal.org/the-yale-law-journal-pocket-part/procedure/losers%E2%80%99-rules/.

Glazer, N., Moynihan, P. (Eds.), 1975. Ethnicity: Theory and Experience. Harvard University Press, Cambridge, MA.

Gotanda, N., 1991. A critique of 'Our constitution is color-blind'. Stanford Law Review 44, 1–68.

Green, T., 2003. Discrimination in workplace dynamics: toward a structural account of disparate impact theory. Harvard Civil Rights-Civil Liberties Law Review 38 (91), 145.

Griggs vs. Duke Power Co. (401 US 424 1971).

Guinier, L., 1994. The Tyranny of the Majority. The Free Press, New York.

Gutmann, A., 1996. Responding to racial injustice. In: Appiah, A., Gutmann, A. (Eds.), Color Conscious: The Political Morality of Race. Princeton, NJ, pp. 106–178.

Harris, A., 2000. Equality trouble: sameness and difference in twentieth century race law. California Law Review 88, 1923.

Harris, A., 2009. Theorizing class, gender, and the law: three approaches. Law & Contemporary. Problems 72 (37), 39.

Harris, L.C., Narayan, U., 1994. Affirmative action and the myth of preferential treatment: a transformative critique of the terms of the affirmative action debate. Harvard Blackletter Journal 11 (1), 4.

Johnson, O., 2012. Beyond the private attorney general: equality directors in american law. New York University, Law Review 87, 1384–1386.

Krawiec, K.D., 2003. Cosmetic compliance and the failure of negotiated governance. Washington University Law Review 81, 487.

Krieger, L.H., 1995. The content of our categories: a cognitive bias approach to discrimination and equal employment opportunity. Stanford Law Review 47, 1161–1248.

Lenhardt, R., 2011. Race audits. Hastings Law Journal 62, 1527.

Loury, G.C., 2000. The Anatomy of Racial Inequality. Harvard University Press, Cambridge, MA.

Massey, D., 2007. Categorically Unequal, the American Stratification System, p. 5.

McCrudden, C., 1999. Mainstreaming equality in the governance of Northern Ireland. Fordham International Law Journal 22, 1696.

Moreau, S., 2010. What is discrimination? Philosophy and Public Affairs 38 (2), 143–179.

Nev Dep't of Human Res. v. Hibbs, 538 U.S. 721, 123 S. Ct. 1972 (2003).

Nielson, L.B., Nelson, R., 2008. Handbook of Employment Discrimination Research: Rights and Realities. Springer.

Oliver, M.L., Shapiro, T., 1995. Black Wealth, White Wealth: A New Perspective on Racial Inequality. Routledge, New York.

Parents Involved in Community Schools v. Seattle School District No. 1, 127 S. Ct. 2738 (2007).

Post, R., 2000. Prejudicial appearances: the logic of American antidiscrimination law. California Law Review 88 (1), 17.

powell, j., 2009. Post-racialism or targeted universalism? Denver University Law Review 86 (1), 3–4.

powell, j., 2012. Poverty and race through a belongingness lens. PolicyMatters 1 (5).

Primus, R., 2003. Equal protection and disparate impact: round three. Harvard Law Review 117, 493.

Rabkin, J., 1997. Racial divisions and judicial obstructions. In: Devins, N. (Ed.), Redefining Equality. Oxford University Press, Oxford.

Rappaport, Michael B., April 3, 2013. Originalism and the Colorblind Constitution. San Diego Legal Studies Paper No. 13-115. Available at SSRN. http://ssrn.com/abstract=2244610. or. http://dx.doi.org/10.2139/ssrn.2244610.

Rawls, J., 1974. In: A Theory of Justice, pp. 74–75.

Robio-Marin, R., 2012. A new european parity-democracy sex equality model and why it won't fly in the United States. The American Journal of Comparative Law 60, 99.

Roberts, S., 2007. Parents Involved in Community Schools v. Seattle School District No. 1, 30.

Scalia, R., 2007. Parents Involved in Community Schools v. Seattle School District No. 1, 30. Roberts, dissenting/.

Selmi, M., 2005. Sex discrimination in the nineties, seventies style. Employee Rights and Employment Policy Journal 9 (1).

Selmi, M., 2006. Was the disparate impact theory a mistake? UCLA Law Review 53, 701.

Sharkey, P., 2013. Stuck in Place, pp. 1, 6.

Shaw, J., 2005. In: Holder (Ed.), Mainstreaming Equality and Diversity in European Union Law and Policy, p. 255.

Siegel, R., 2000. Discrimination in the eyes of the law: how 'color blindness' discourse disrupts and rationalizes social stratification. California Law Review 88, 77, 82, 101.

Siegel, R., 2004. Equality talk: antisubordination and anticlassification values in constitutional struggles over brown. Harvard Law Review 117, 1470, 1472–1473, 1486, 1500, 1530–1531.

Siegel, R., 2011. From colorblindness to Antibalkanization: an emerging ground of decision in race equality cases. Yale Law Journal 120, 1278.

Sowell, T., 1984. Civil Rights: Rhetoric or Reality? University of Chicago.

Strauss, D., 1986. The myth of colorblindness. Supreme Court Review 99.

Strauss, D., 1998. The illusory distinction between equality of opportunity and equality of result. In: Devins, N., Douglas, D.M. (Eds.), Redefining Equality.

Sturm, S., 1998. Race, gender, and the law in the twenty-first century workplace: some preliminary observations. University of Pennsylvania Journal of Labor and Employment Law 2, 639.

Sturm, S., 2001. Second generation employment discrimination: a structural approach. Columbia Law Review 101, 458–568.

Sturm, S., 2006a. The architecture of inclusion: advancing workplace equality in higher education. Harvard Journal of Law and Gender 29, 249.

Sturm, S., 2006b. Law's role in addressing complex discrimination. In: Nielson, L.B., Nelson, R.L. (Eds.), Handbook of Employment Discrimination Research, pp. 35–59.

Sturm, S., 2014. Reframing the civil rights narrative: from compliance to collective impact. In: Sarat, A. (Ed.), Civil Rights in the American Story. Cambrigde University Press, Cambridge, pp. 145–181.

Sturm, S., Guinier, L., 1996. The future of affirmative action: reclaiming the innovative ideal. California Law Review 84, 953–1036.

Sturm, S., Eatman, T., Saltmarsh, J., Bush, A., 2012. Full Participation: Full Participation: Building the Architecture for Diversity and Public Engagement in Higher Education (White Paper). Columbia University Law School: Center for Institutional and Social Change. Available at: www.fullparticipation.net.

Suk, J., 2007. Equal by comparison: unsettling assumptions of anti-discrimination law. American Journal of Comparative Law 55, 295–345, 325, 327.

Suk, J., 2012. From antidiscrimination to equality: stereotypes and the life cycle in the United States and Europe. American Journal of Comparative Law 60, 75.

Sunstein, C., 1994. The anticaste principle. Michigan Law Review 92, 2410.

Teamsters vs. United States, 431 US 324 1977.

Tetlock, P.E., Mitchell, G., 2009. Implicit bias and accountability systems: what must organizations do to prevent discrimination?. In: Staw, B.M., Brief, A. (Eds.), Research in organizational behavior, vol. 29. Elsevier, New York, pp. 3–38.

Tremblay, L., 2012. Promoting equality and combatting discrimination through affirmative action: the same challenge – questioning the canadian substantive equality paradigm. American Journal of Comparative Law 60, 181–204.

United States v. Windsor, 133 S. Ct. 2675 (2013).

Wax, A., 1999. Discrimination as accident. Indiana Law Journal 74, 1129.

Wilkins, D., Gulati, G.M., 1996. Why are there so few Black lawyers in corporate law firms? an institutional analysis. California Law Review 493, 493–614.

Wilkinson, R., Pickett, K., 2009. The Spirit Level: Why Greater Equality Makes Societies Stronger.

Williams, J., 1987. Eyes on the Prize: America's Civil Rights Years, 1954–1965. Penguin, New York.

Wilson, W.J., 1996. When Work Disappears: The World of the New Urban Poor. Vintage, New York, pp. 19–24.

Witte, B. de, 2012. New institutions for promoting equality in Europe: legal transfers, national bricolage and European governance. American Journal of Comparative Law 60, 49.

Yoshino, K., 2011. The new equal protection. Harvard Law Review 768, 802–803.

Equality of Opportunity

Meir Yaish, University of Haifa, Haifa, Israel

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A.F. Heath, volume 7, pp. 4722–4724, © 2001, Elsevier Ltd.

Abstract

The term *equality of opportunity* means different things to different scholars, policy makers, and popular audiences. This article explores the concept *equality of opportunity* with a focus on how it is used by students of stratification and social mobility, particularly in sociology. I suggest that a narrow definition of the concept *equality of opportunity* is the most fruitful one to adopt. Such a definition will lead students of inequality the opportunity to focus their analyses on inequality of results and conditions, rather than debating the conditions under which inequality can be legitimated. After all, equality of opportunity is partly determined by equality of conditions, whereas equality of results is much more visible and less elusive than equality of opportunity in its many interpretations.

The term *equality of opportunity* has philosophical, normative, political, ideological, and practical aspects, and it means different things to different scholars, policy makers, and popular audiences. At the philosophical normative end, the term is at the heart of a debate and discussion about just society. At the practical end, sociologists, and recently also economists, equate the term with social and economic mobility within a class or within other kinds of social structures. The two ends meet when students of social and economic mobility make inferences about *fairness* and *social justice* from studying social and economic mobility. There is neither an exclusive nor an agreed upon definition of the term social mobility. In fact, the term is elusive and difficult to operationalize and its critics argue that one of its main purposes is to mask and/or justify social and economic inequality of conditions. This article explores the term, or more accurately the concept, *equality of opportunity*, with a focus on how it is used by students of stratification and social mobility, particularly in sociology.

Equality of Opportunity and Stratification

Social stratification and mobility processes can be depicted as a race, with winners and losers obviously. Thus, social stratification and mobility processes are associated with inequality of results (of the race) and conditions. How one defines the concept *equality of opportunity* is important in assessing whether or not social inequality is fair and justifiable. A very wide array of definitions for the concept equality of opportunity exists in the literature, providing many – nearly all possible – justifications for the existence of inequality in society.

A thin definition of the concept *equality of opportunity* postulates that opportunities are equal as long as no overt discrimination on the grounds of race, ethnicity, gender, and the like prevents anyone from entering a race. This minimal definition is also known as formal equality of opportunity, and is concerned with equality of the access to available opportunity. Examples for the lack of equal access to opportunity used to be very common in the past. For example, for many years women did not have access to the political spheres and were denied access to higher education. This formal definition of the concept is apparent in the legislation.

Entering the race does not necessarily mean winning it, and thus the concept equality of opportunity can be extended to include also the rules governing the competition and how the winner is selected, that is, equality in the selection processes, once access is granted. At the formal level, an extension of the concept to include also selection processes would require, in the context of the labor market for example, the adaptation of race, age, gender, etc., blind procedures for job hiring and firing. Under such a policy, equality of opportunity does not necessarily mean that social background is not associated with the winners of the race. This is so simply because of the advantages and disadvantages that social background entails. Accordingly, a child of educated professional parents is better equipped to take advantage of the educational system than a child of uneducated working class parents. Defining equality of opportunity with reference to equal chances for access and advancement is akin to Sorokin's characterization of modern democratic societies, as societies in which "there are no judicial or religious obstacles to climbing or going down" (Sorokin, 1959: p. 153).

Equality of Opportunity and Mobility

A requirement that social background not be associated with the results of the race perhaps represents the broadest definition of the term equality of opportunity. This definition of the concept is very common in sociology, and also in economics, particularly among students of social stratification and mobility. Accordingly, intergenerational social and economic mobility is equated with equality of opportunity, such that the by now standard concept of *relative mobility* is taken as a measure of equality of opportunity.

Relative mobility can be explored by a transition matrix showing movements in, for example, a class hierarchy across generations. Relative rates are then treated as a set of (odds) ratios that measure the chances of having one origin rather than another for persons of one class relative to the comparable chances for persons in other classes. In the simplest case of

a 2 × 2 mobility table, the one odds ratio calculable would be given by the following equation:

$$\frac{f_{11}/f_{12}}{f_{21}/f_{22}}$$

where, f_{11} is the number of cases immobile in category 1, f_{12} the number mobile from category 1 origins to category 2 destinations, and so on. An odds ratio of 1 in this example implies complete statistical independence of the origin and the destination categories involved, or complete equality of opportunity. In mobility tables with more than two categories, it is possible, of course, to calculate more than one odds ratio, and other statistical techniques, such as log-linear analysis, are available to assess statistical independence.

Thus, if the odds of working class kids to end up later on in their lives in the upper class equals the odds of those born into the upper class to retain their privileged positions, then relative mobility – or *social fluidity* – is high and equality of opportunity exists within this society's class structure. This depiction of the term is best illustrated with references to the consequences of industrialization processes for stratification and mobility regimes. Accordingly, rapid economic development – a key element in the industrialization process – has resulted in the upgrading of the occupational structure of societies such that many new opportunities were created at the top of the occupational hierarchy. If these new opportunities are then taken equally by members of different social backgrounds, then equality of opportunity prevails. This interpretation equally applies to other social processes, most notably the educational attainment process. Here, the expansion of tertiary educational systems of most Western societies has created new opportunities at the top of the educational hierarchy. Equality of educational opportunity requires that these opportunities be shared by members of all social backgrounds equally.

The requirement that social background not be associated with the results of the competition represents a very wide definition of the term *equality of opportunity*, one that might require a compensatory mechanism for those underprivileged in society. A number of such mechanisms are discussed in the literature; the simplest one is a level playing field mechanism in which people are equalized before the race begins. Any inequality of results afterward is then justified. This mechanism, however, is at odds with normative ideas such as social justice and the notion of fairness. As Aristotle in his *Nicomachean Ethics* holds, if persons are not equal, they will not have nor be entitled to equal shares.

Equality of Opportunity and Competence

A more functional definition of the term equality of opportunity should therefore take into account people's competencies to realize the available opportunities (Marshall et al., 1997). That is, people with similar competencies should have equal chances to take on the available opportunities. In the educational attainment field this definition of the term equality of opportunity requires that members of various social backgrounds, but with similar ability, have equal chances for advancement in the educational hierarchy. In the social mobility field this definition normally requires that members of various social backgrounds, but with similar education, have equal access to the new opportunities that are created as a result of industrialization processes at the top of the occupational hierarchy.

In the sociological literature this idea is very clearly made by advocates of the so-called liberal thesis of industrialization advanced by American social scientists such as Talcott Parsons, Clark Kerr, and Daniel Bell. Accordingly, as well as changes in the structure of opportunities itself, industrialization alters the processes by which individuals are allocated to particular positions within the division of labor. Central to this argument is the shift away from *ascription* toward *achievement* as the principle criterion for social selection (Treiman, 1970). It is at this point that the theme of *meritocracy* comes to the fore. As Daniel Bell puts it: "by the logic of a meritocracy, these high-scoring individuals, no matter where they are in society, should be brought to the top in order to make the best use of their talent" (1974: p. 414). Thus, as merit comes to dominate social selection, the association between individuals' social origins and their destinations tends to weaken. As I have indicated earlier, when the association between social background and current social position is very weak, relative rates of mobility become equal and *equality of opportunity* is reached.

A major problem with this definition of the concept relates to the question: which competencies merit rewards? Put another way, can we justify unequal rewards on the basis of some heritable skills, IQ, or beauty? These questions can be linked to the issue of compensatory mechanisms earlier introduced, by focusing on those less privileged in society. Here we thus ask: the lack of which competencies should be compensated for?

The general consensus in the political philosophy literature is the idea that in striving for equality individuals should be held responsible for what happens to them (Roemer, 2002). That is, individuals should be held responsible for their preferences but not their resources (Dworkin, 1981). According to this line of reasoning, then, inequalities that arise from circumstances beyond our control should be considered illegitimate, while those that are the result of choice and preferences should be deemed legitimate (Phillips, 2004). But can we really distinguish between preferences and circumstances? Are not preferences also the results of one's position in the social hierarchy (circumstances)?

Examples for the intertwining of the two are numerous. For example, Hakim (2002) argues in her so-called preference theory that women select occupations that allow them to maintain the lifestyle they value. Most women, according to this argument, prefer a balanced work–family life and therefore select occupations that allow a better combination of the two roles. These 'preferences', however, do not necessarily reflect a choice free of constraint, as in the absence of public support many women are bound to adjust their occupational choices so that they meet their familial responsibilities. In a similar way, Breen and Goldthorpe (1997) explain the constancy of inequality in educational opportunity and, indeed, the existence of a stable class structure (cf Goldthorpe, 2000) in most modern societies by class-based educational choice. That is, members of the different social classes make different educational choices, which then result in the reproduction of the class structure. Hence, one can argue that ascriptive forces (circumstances) might often find ways of expressing

themselves as choice and preferences, as noted by Halsey (1977: pp. 173–186) in the context of social mobility in Britain.

Conclusion

To obviate these problems, I argue in favor of the narrowest possible definition of the concept *equality of opportunity*, one that requires the removal of all formal discrimination practices. Such a clear and lucid description of the term will lead students of inequality the opportunity to focus attention on studying inequality of results and conditions, rather than debating the conditions under which inequality can be legitimated. After all, equality of opportunity is partly determined by equality of conditions, while equality of results is much more visible and less elusive than equality of opportunity in its many interpretations.

See also: Income Inequality; Industrialization; Inequality Measurement; Inequality, Social; Meritocracy and Tokenism; Mobility, Social; Social Inequality and Schooling; Social Justice: Historical and Theoretical Considerations; Social Stratification.

Bibliography

Bell, D., 1974. The Coming of the Post-Industrial Society. A Venture in Social Forecasting. Heinemann, London.

Breen, R., Goldthorpe, J.H., 1997. Explaining educational differentials - towards a formal rational action theory. Rationality and Society 9 (3), 275–305.

Dworkin, R., 1981. What is equality? Part 2: equality of resources. Philosophy and Public Affairs 10, 283–345.

Goldthorpe, J.H., 2000. On Sociology. Oxford University Press, Oxford.

Hakim, C., 2002. Lifestyle preferences as determinants of women's differentiated labor market careers. Work & Occupations 29 (4), 428–459.

Halsey, A.H., 1977. Towards meritocracy? The case of Britain. In: Karabel, J., Halsey, A.H. (Eds.), Power and Ideology in Education. Oxford University Press, New York, pp. 173–186.

Marshall, G., Swift, A., Robert, S., 1997. Against the Odds? Social Class and Social Justice in Industrial Societies. Clarendon Press, Oxford.

Phillips, A., 2004. Defending equality of outcome. Journal of Political Philosophy 12, 1–19.

Roemer, J.E., 2002. Equality of opportunity: a progress report. Social Choice and Welfare 19 (2), 455–472.

Sorokin, P.A., 1959. Social Mobility. The Free Press, Glencoe.

Treiman, D.J., 1970. Industrialization and social stratification. In: Laumann, E.O. (Ed.), Social Stratification: Research and Theory for the 1970s. Bobbs-Merrill, Indianapolis, pp. 207–234.

Equality: Philosophical Aspects

Emmanuel Picavet, Université Paris-1 Panthéon-Sorbonne, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.J. Arneson, volume 7, pp. 4724–4729, © 2001, Elsevier Ltd.

Abstract

Equality is an enduring source of challenges for social practice, institutions, and analysis. Being no replica of identity in features of reality, it calls for normative judgment, hence its importance in moral and political thought. The close association of normative and descriptive elements typically brings about controversies about the right and relevant ways to secure the equal treatment of persons in their social life. Equality in status, equal opportunities, and equal rights are major concerns in this respect.

Equality is an object of inquiry in moral and political philosophy, as an ideal which pertains to the status of human beings as such (in a cosmopolitan perspective) and also to the treatment of individuals as citizens of a political entity and to the proceedings of social life. Although the equality ideal is usually vindicated, endorsed, or studied as a *de jure* requirement or an ideal to be rationally pursued, it must be considered in close association with the observation of empirical facts about the meaning attached to equal treatment in various cultural settings. The kind of equality which must be enforced so that reciprocity prevails is thus known to be highly dependent on the cultural characteristics of social life. While inequalities are plural, they jointly impact interpersonal relationships and social life, and they are under the common yoke of social institutions (Nadeau, 2007: chap. 2). The correlative social nature of equality aspirations suggests that equality is to be framed as a system of adequate social relationships (as in the variety of democratic equality defended by Anderson, 1999; see also Scheffler, 2003) rather than identified with patterns of distribution (or redistribution) of goods and bads. The complexity of moral and political thought about equality reflects the problems involved in the articulation of distributive operations and underlying concerns about equal treatment, equality in relationships, and equal dignity.

Equality as a Moral and Political Principle: The Variety of Philosophical Concerns about Equality

Ancient philosophers gave an expression to the ideal of the equality of all men, notwithstanding the contrast with their own political environment. In the Stoa, Zeno's *Republic* gave the example of a completely uniform *polis*, comprising only fellow-citizens and local residents, with only one way of life and order (Plutarch, *On the Fortune of Alexander*, 329). Ancient Greece provided analyses of equal treatment in social and political proceedings. In the Aristotelian treatment of monetary exchange, it is assumed that by means of convention, money makes different things tolerably commensurable and establishes the kind of equality that is necessary for exchange to take place (*Nicomachean Ethics*, V, 8). In the *Magna Moralia* (I, xxxiii, 28), it is explained that if a man is content to take less than his due, given that equality would have been a man's just due, injustice is only apparent; such a man is in fact exchanging part of his due for honor, praise, and the like. Inequality in friendship was also an object of scrutiny; those who are superior in wealth, for example, do not expect to love, but to be loved (they consider that they repay those who love them with the things in which they have a superiority); this is an object of ambition but being loved confers no benefit (*Magna Moralia*, II, xi, 34–40).

In Aristotle's *Nicomachean Ethics* (V, 2), the notion of equality plays an essential part in the definition of the just man; along with obeying the law, the just man is respectful of 'equality,' in contrast with the unjust man's taking more than his part of the good things, or trying to get less than his part of the bad things. In this classic treatment of justice, equality has the meaning of respecting a correct standard of behavior with respect to bad or good things, which makes it clear that philosophy should address the basic problem of setting up an interpersonal standard for the definition of such conduct. Aristotelian theory further developed the notion of proportionate equality, which made it possible to establish a direct connection with justice; in the *Magna Moralia*, it is thus considered proportionate that "one who has large possessions should pay a large amount in taxes, whilst he who possesses little pays little; and likewise that one who has toiled little also receives little. The ratio of the labours should equal the ratio of the receipts, and the ratios of labour to receipt equal one another" (*Magna Moralia*, I, xxxiii, 9, trad. G. Cyril Armstrong).

Equality had been studied in ancient times in connection with various moral subjects, ranging from having an equal temper in the face of adversity, or the importance of avoiding a degrading anger-inspired contest with an inferior (Seneca, *De Ira*, II, xxxiv, 1), to renouncing the conquest of advantage by violent means in favor of the kind of equality established by justice (Cicero, *De Officiis*, xix, 63). But equality as a moral ideal to be pursued through the fundamental rights and prerogatives attached to citizenship is by and large a modern notion. Equality among all human beings is a moral ideal which finds institutional expression, for example in the principles of political constitutions or international charters, or in officially proclaimed moral values (such as "Liberty, Equality, Fraternity" in republican France) or in fundamental legal arrangements. This can be illustrated by the emphasis of Condorcet on equality as the foundation of a republican constitution, which

helps preserve freedom and dignity, in accordance with nature, reason and justice ("*Exposition des principes et des motifs du plan de Constitution*," 1792, in *Œuvres*, vol. 12).

In some cases, the equality principle is taken to refer to the enjoyment of goods or property, or the command of revenue. But other politically and economically significant inequalities have a more procedural character, as they originate in the operations of social interaction: inequalities in information, freedom (Carter, 1999: chap. 3), recognized merit or effort, recognized talent or aptitude, level of independence, social influence, or friendly support. Kolm's theory of macro-justice sets out to give a firm grounding to the association between equal freedom and fairness in income redistribution, starting from a procedural view of equal social liberty for individuals who face restrictions on their choices (in particular, budget limits – see Kolm, 2005).

Significant Inequalities and the Ideal of Equality

In the political and social processes which convert empirical differences (which might partly originate in natural causes) into problematic political, social, or economic 'inequalities,' there is an important role for personal aspirations, goals, and values. Inequalities that matter are inequalities with respect to those human goals and values we should not ignore when we consider the sources of good or meaningful life in the human species. "Giving appropriate weight to interests for the sake of well-conceived equality might invite us not to give equal weight to all interests (however futile), in order to give more weight to those interests which deserve to be satisfied, or whose owners are entitled to have them satisfied" (D. Locke, 1981: p. 533). Such reasons support the centrality given, in ethical or political discourse, to a number of essential social dimensions of inequalities (which are usually not mutually independent from an empirical point of view): inequalities in property and income, educational inequality (unequal access to formal education, unequal opportunities in higher education, unequal quality of educational services), health-services inequality, job-market inequality, housing inequalities (access to housing facilities, quality of housing), inequality with respect to the availability of cultural goods and services, unequal quality of the environment, and so on. Due to the characteristics of particular societies in specific historical periods, such inequalities might have considerable overlap, leading to massive inequalities between social groups, communities, casts, or social classes.

In the eyes of most egalitarians, there is a deep connection between legal equality (the equal status granted by commonly recognized norms and institutions) and the achievement of a substantial form of equality, such as equal access to goods or services, or the command of equal revenue. Notwithstanding the practical difficulties and the fact that we are unlikely to achieve perfection in this respect, collective action should then be oriented toward the achievement of equal benefits for all from organized social life. Here, the means might have complicated connections with the ends. In many cases, public decision-making mechanisms will turn out to be necessary to make this kind of progress toward equality, but such mechanisms might prove deeply problematic from the point of view of equality itself. Indeed, they call for the institution of power differentials among persons, which can generate worrying inequalities.

Other, idiosyncratic practical difficulties should not be overlooked: inequalities are usually deeply rooted in social and economic life, in culture and traditions, in historical trends, and in geographical data. As a result, the effectiveness of collective action in the fight against those forms of inequality which are generally recognized as problematic is itself problematic, and an object of scrutiny. Furthermore, the practical obstacles might devaluate the currency of egalitarian speech, if the latter is excessively ambitious while we lack effective means of action. Concentration on specifically important dimensions of inequality, such as health care and education, is quite useful for pragmatic purposes. In addition to keeping the problems tractable, this strategy is usually rooted in needs which can be described as basic or essential. It is thus also amenable to plausible interpersonal comparisons and consensus.

Nevertheless, real egalitarians understandably want to go well beyond the development of widely accessible quality services in health care and education. Equality along more numerous dimensions of objective welfare can be vindicated with plausibility and, indeed, promoting equality in welfare lies at the heart of a variety of egalitarianism (see Arneson, 1989; Guibet Lafaye, 2006: chap. 2). With lesser intrinsic plausibility in predominantly free-market economies, equality in wealth or revenue is still a living ideal in the heart of hard-line egalitarians; for all its lack of plausibility when separately considered, this objective is influential as a limiting case when it comes to arguing for or against reforms that impact the prospects of greater or lesser material inequality, or inequality in access to goods and services generally speaking.

Concern for an encompassing view of equality might lead to the acceptance of inequality in treatment by state authorities, if such unequal treatment is framed as a way to remedy deep-seated social inequalities. This is the root of the 'affirmative action' strategies, which use differential opportunities to be appointed to significant positions in the public sector, or differential public support for educational opportunities, as tools to remedy pervasive and culturally inherited social prejudices and inequalities. Such concerns have led to short-term, transitory social strategies.

It must be acknowledged that we need go beyond the traditional precepts of equal treatment according to the law in order to secure significant results in terms of substantial equality, even though the enjoyment of an equal legal status with respect to basic rights, obligations, and liberties (irrespective of sex, religion, ethnic background, marital status, etc.) is obviously a problem in several countries nowadays. The resulting initiatives might appear problematic in the perspective of those doctrines which promote a purely process-based view of social justice (which is supposed to be rooted in equal respect for persons and their property, after the pattern described by Kolm, 2005; see also, for thorough analyses, Gamel and Lubrano, 2011). The problem is further connected with the frequent suspicion that redistributive or corrective social and economic action is, of necessity, dependent on particular value judgments, in such a way that individuals arbitrarily benefit or suffer from the opinions of others, in violation of

the moral equality of all human beings. This tension in the requirements of equality is, to a large extent, responsible for persistent conceptual and practical difficulties in the fight against inequalities. In order to effectively promote equality in the enjoyment of the goods or advantages gained through social life and common institutions, we must go beyond the sort of laws which formally institutionalize the protection of persons and their property, once a framework for the enforcement of a basic equal liberty for all is secured (let's say, in the spirit of John Rawls's first "principle of justice," formulated and explained in *A Theory of Justice*). To go beyond, we need laws that are political instruments in the fight about these inequalities which are found to be important by the rulers, with respect to the enjoyment of various advantages.

Many important issues in the field of inequality are mediated by individual rights. This raises specific problems because equality in terms of rights has its own importance, even though this form of equality is not automatically conducive to equality on all scores. As shown by J. Habermas (1992: chap. IX, sec. II), there is a dialectical link between right-based equal freedom and the unequal opportunities given to individuals to take advantage of their equal freedom (given the empirical social conditions that are partly the result of rights-exercising); this dialectic, Habermas argues, is central to the development of law in the contemporary world. Historically, the compatibility of first-generation (eighteenth-century) fundamental human rights with worrying social and economic inequalities (such as educational or health-related inequalities in particular) has been a major reason to try to develop a more encompassing concept of human rights, in order to take the social and economic dimensions of life into account.

Unemployment, no less than unrewarding activities, threatens the equal ability of citizens to value the kind of life they can choose to live according to their own conception of the good (Guillarme, 1999: chap. 5); other challenges are posed by segregated social roles and by the kind of identity-threatening social disrespect that is given prominence in Honneth (1992), which motivate effort to integrate egalitarian redistributive views and recognition concerns (Phillips, 1999; Fraser and Honneth, 2003). All in all, the connection between workplace activity (or redundancy), social respect, and social rights stands as a major challenge for equality in our times, in the face of evolving, often problematic forms of workplace and housing inclusion or exclusion. Promoting equality in individual functionings and capabilities (as classically defined by Sen, 1992) is a plausible strategy for addressing and interrelating a number of core insights and concerns in the field, even though the anthropological and sociological underpinnings of the capabilities approach are quite specific.

Natural Equality and the Social Treatment of Individuals

Equality is an ideal with respect to both representations of the nature of human persons and conceptions about the fair treatment of persons by other persons as well as collective agents. It is usually hypothesized that the second perspective is determined by the first one. In particular, holding the view that human beings are equal beings is widely thought to be a sufficient condition for holding that human beings should be treated equally. At the level of fundamental moral dispositions, this is convincing indeed: equality of status implies that, when human beings are simply considered as human beings, there is no reason to believe that they should be treated unequally in their mutual relationships, and unequal claims would be unjustified in the first place.

Such ideas have had a profound influence in the history of mankind. In particular, the gradual realization that men and women are equal, as well as human beings with different cultural background and skin colors, has been a crucial source of inspiration for progress in the equal treatment of men and women, and similarly in the case of equal treatment for people with different cultural backgrounds (consequences must then be drawn for the colonial experience) and in the case of equal treatment for people with different skin colors (with consequences for the abolition of slavery and apartheid). In detailed social, administrative, or political problems, the basic claims associated with the fundamental status of human being are not usually considered in a direct manner. Even when equal treatment is an option, it is generally not automatically considered the best one, owing to the association of the relevant individual claims with the particulars of joint work and mutual advantage, personal capacities and the public goods, individual talents and overall economic efficiency, and the combination of personal traits and social life.

Most of the claims about justice are compatible with serious cases of inequality which can often plausibly be interpreted as examples of the unequal treatment of individuals in (or by) society. More often than not, such inequalities are actually required by the selected views about justice. This raises a natural question: if 'justice' requires a departure from strict equality, how can it count as an expression of justice in the first place? Does it not stand to be criticized on account of the violation of equality? A standard answer consists in saying that ideals of justice are always ideals of equality or equal treatment in definite respects which are considered central or most important for the questions at hand; this might involve inequality in other respects, in such a way that fairness or equity is safe. This answer accounts for the diversity in opinions or doctrines about justice when it comes to deriving their implications in terms of inequality or equality (Sen, 1992).

For example, theories of justice which focus on equality in the enjoyment of personal rights (such as property rights) are, in principle, compatible with wide inequalities in terms of material welfare or financial resources. As a matter of fact, respect for free, consent-based interactions between adult persons normally commits one to accept unequal outcomes, unless these are being actively corrected so that they conform to equality ideals in all essential respects. If the latter kind of correction is taken to be inappropriate, then one should be prepared to welcome inequality. Other theories, which articulate ideals of equal treatment in connection with welfare or resources, might accommodate unequal protection of persons and goods, or the denial of due process in judicial activity. For example, this awkward property might originate in the fact that progress toward economic equality is taken to involve revolutionary episodes and class conflict.

To be sure, inequality raises problems of identification: how can we check that two situations are 'equal'? Which criteria can we use to assess the presence of inequality (the kind of inequality we are interested in) in particular cases? Furthermore, given a general description of the kind of inequality that is of interest, judgment in particular cases presupposes the availability of criteria that enable us to process the pieces of information that are included in the data. Normatively speaking, the main problems of inequality have to do with the connection between inequality and social justice (or fairness), as emphasized in the socialist and liberal political traditions. This accounts for the usual negative value judgment about inequality, even though inequality discourse is seemingly descriptive. As noted by Hilary Putnam (2002), description and value judgments are intimately connected in the ways to address inequalities. Consider the interest we take in the interpretation of empirical measures or descriptions of social and economic inequalities in given societies; this kind of interpretation is deeply influenced by the selection of morally or politically relevant aspects of the distribution and by the choice of the underlying characteristics that are the object of description or measurement. The interplay of norm-based value judgments and empirical statements is remarkable here, as it occasionally conditions the recognition of empirical knowledge as relevant, faithful to its purpose, or even adequate. Indeed, a failure to develop morally or politically relevant measures of inequality – measures that enable one to assess the kind of inequality that should be attenuated or suppressed – is likely to be interpreted as a failure to describe (or measure) inequality in the first place.

To be sure, active fight against inequality by political or legal means is not always appraised positively (e.g., it was repeatedly and notoriously criticized by Friedrich von Hayek). In addition, divisions occur about the seriousness or the relevance of inequalities. For example, in the long-standing controversies in contemporary Greece about differentiated tax rules, there is no homogeneity of judgment about the problematic nature of the Church's advantages; for some, this amounts to a serious departure from the standard of equality; for others, this is an inconsequential form of differentiated treatment, or an inequality that should be regarded as negligible if we focus on the associated advantages, for example, with respect to the preservation of the nation's cultural heritage. Actively fighting inequality is sometimes criticized on account of the alleged conflict with everyone's free access to the benefits of one's own effort – an argument which is regularly defeated in various contexts by appealing to the fact of widespread social interdependence in education, in the division of labor, and in the social background of fruitful activity.

The kind of equality which is of interest for normative social science and for moral and political philosophy seems to be amenable to description as a variety of equal treatment in an organized social process. Normative discourse about inequality is typically not a discourse about natural differences (or differences with a very strong natural grounding), such as differences in physical strength, quick-mindedness, and so forth, even though the common descriptions of such interpersonal differences are often based on comparisons which provide a common ground with inequality speech: people have a 'more or less' robust constitution, they are 'more or less' quick-minded, and so on. When differentiations of this kind are taken into account, the underlying questions have to do with the neutralization of their influence in the processes of social life, possibly by means of compensation schemes. Relevant questions might also have to do with the collective use of natural assets. How are we collectively to benefit – in an efficient and fair way – from a distribution of natural assets? A typical, more detailed question is, for example: how should we organize social institutions so that the unequal distribution of individual dispositions is put to use in a satisfying manner (both efficient and equitable), given our differences in needs?

Equality, Democracy, and Contractarian Methodology

Human beings are not naturally unequal between themselves to the extent they are in their social life. Moreover, the development of social inequalities might make it difficult for some individuals to give their consent to social arrangements that threaten comparatively advantageous positions in these inequalities, especially if their own situation is less favorable in such alternative patterns of the social order. From a rationalist perspective, however, a moral judgment about the social order should *a priori* commit ourselves neither to the defense of an antecedent, inherited order, nor to the *defense* of personal interests in a special way. On the contrary, it should equip us with a more detached perspective – the kind of perspective needed when we raise questions about the justice or fairness of social institutions. These are some of the reasons which give credibility to John Rawls's statement according to which "despite the fact that the contract doctrine seems at first to be a somewhat special conception, particularly in its treatment of inequalities, it may still express the principles of justice which stand in the background and control the weights expressed in our everyday judgments" (Rawls, 1967: *Collected Papers*, p. 153).

Contractarianism (the family of moral, ethical, and economic doctrines which place emphasis on possible unanimous agreements on basic social arrangements after the pattern of a 'social contract' in order to evaluate, justify, or criticize such arrangements) invites us to judge social inequalities by the standards that would be selected by free people in a state of equality, as proposed by the core models of social contract theories. Philosophical analysis should equip us with a benchmark 'equal' situation which can offer a basis for normative reasoning; such is the general meaning of an original position's design in contractarian theory. It should be observed, though, that the theoretical functions of this design are plural. Following Hobbes and Locke, contractarian theorists refer to a virtual original position which is a state of (at least approximate) equality. This accounts for the defects or positive reasons why people should be content with institutional arrangements of some sort (e.g., the existence of sovereign authority, as classically justified by Hobbes in his *Leviathan* in the successive 1651 and 1668 versions, or the existence of judicial arbitration, as explained in Locke's *Two Treatises of Government*, 1689). Rational choice and self-interest are then instrumental in identifying the reasons why the described state should be viewed as imperfect and transient.

Rousseau's assumptions about the initial equality of men in their natural condition also perform this kind of reason-providing function. Both in the *Social Contract* (1762) and in the sketch of a correct collective arrangement in the second *Discourse (Discours sur l'origine et les fondements de l'inégalité parmi les hommes, 1755)*, the natural equality of men provides a standard to appreciate both the poor condition of men as a result of the progress of civilized life (which involves property, power and wealth differentials, and so on), and the kind of freedom that is regained by a true social contract, given the widespread interdependence which is a result of social life. But the partly narrative, partly predictive social theory which is developed in the second *Discourse* enables contractarian theory to play the additional role of explaining how poorly reflective and badly informed unanimous choices, starting from an intermediate state of inequality among men, might account for the social enforcement of unequal institutions and, subsequently, massive inequality in modern societies.

In spite of the egalitarian features of 'states of nature' or original positions, the conclusion of a contractarian argument is not necessarily egalitarian. At the procedural level, however, equality is required. The ideal of equality is expressed through the methodological requirements which make it possible to reach conclusions in an impartial way. In Kant's use of the social contract as a criterion of justice, the condition of a possible universal assent of all citizens to public action yields an impartial test for revealing the unjust nature of given acts of government. Along different routes, this line of inquiry was pursued in Harsanyi (1953, 1955), who proposed a renewal of contractarian methodology (starting from rational-choice premises and leading to neo-utilitarian statements), and in Rawls's *Theory of Justice* (1971).

The conclusions of Rawls's theory of social justice are compatible with marked inequalities in the panels of 'primary goods,' that is to say, goods which are involved in all avenues of life, no matter the personal choices or conceptions of the good life. Such goods are secured to individuals by the basic features of a social 'structure' and their allocation is the proper object of analysis for the proposed economic and political justice. The ideal of equality is expressed through the structure of the overall treatment of justice, which culminates in the statement of the two 'principles of justice' with their companion priority rules. The equal allocation of a maximally defined set of basic individual freedoms is firmly granted by the first principle. 'Fair equality of opportunity' in the Rawlsian sense states that, given equal talent and similar ambition, the prospects of reaching social positions (with their associated prerogatives and advantages) are the same, and the positions are open for all in the first place. Beyond this requirement, the second principle, which makes use of the 'maximin' logic (an ordering principle which invites us to look first at the situation of the worst off when ranking possible distributive patterns, then at the situation of those who occupy the other positions), further secures that all individuals are, in some sense, equally treated as ends in themselves: each of them should be able to view society as best organized for his or her own endowment in all-purpose primary goods.

Following Rawls (1971), equality is deeply connected with the procedural analysis of justice. The model of a choice under a 'veil of ignorance' in an 'original position' is rooted in a quest for equality and impartiality in the deepest layer of the theory, although is leads to conclusions which are questionable from an egalitarian point of view. ('Veil of ignorance' is an argumentative strategy which focuses on a thought experiment through which a number of empirical (social and personal) characteristics of individuals are put aside, so that their influence on the outcome of choice problems is nullified.) Following the proposed methodology, all members of society are supposed to have a say in the collective articulation of principles of justice for the regulation of the basic structure of their society. But the kind of judgment which must be reached is moral in character; it should help us answer questions about the justice of social arrangements, and these cannot be reduced to strategic or tactical political questions, nor to economic efficiency matters. In the continuity of Locke, Rousseau, and Kant, Rawls develops a vision of equality according to which all human beings should be treated equally in connection with the basic social structure of society, and this finds an expression in a choice under a 'veil of ignorance,' which blocks the influence of contingent factors such as age, sex, and social origin. This methodology nullifies the influence of contingent factors which might impair impartiality in the choice of principles for the whole of society.

Contractarian methodology yields benchmarks for assessing the legitimacy of power, and the equality assumptions which are typical of this methodology are congenial to the moral underpinnings of democratic theory. This convergence allows adequate contractarian justifications of power to be recognized as valuable contributions to the legitimacy of democratic power. From a practical point of view, however, investing specific people with special powers or prerogatives is a major source of inequality. Obedience to authority, even when its justification is clearly associated with equality-based normative reasoning and judgments, is a source of concern with respect to equality.

Controversies about Possibly Legitimate Sources of Inequality

Neutralizing power asymmetries among individuals, even legitimate inequalities, is central to the quest for proximity that is associated with contemporary methods of participative deliberation. Nevertheless, participatory democracy on a local scale draws attention to a less-than-general concept of shared will. The resulting dialectical tension with what is referred to as the common good or the general will at the level of the political community at large is in turn a source of worries for the equality of citizens. If obedience to political authority is to be understood as obeying to one's own prescribed law after the pattern of Rousseau's ideal social contract, a key requirement of the kind of equality which is being pursued in rule-making is that we only refer to one kind of united wills. Giving weight to separate interests, even in the form of shared interests for the concerned parties, might be problematic for the effective pursuit of equality among citizens. There is a philosophical problem in this respect, because inclusion of various groups of concerned persons in political debates, including local groups of stakeholders,

might turn out to be crucial to giving each individual a voice which can be heard in an effective way, thus promoting equality in the form of a remedy to power asymmetries; it might also be crucial to helping individuals in the elaboration of their own values or interests in a reflective way, thus giving moral significance to equal involvement in debates and decision-making.

Acting out of respect for equality might be compatible with a differentiated treatment of persons whenever personal effort or contribution is itself differentiated. Otherwise – so the argument goes – effort or personal involvement (or merit) would not receive equal consideration; think of the shareholder model of just retribution, classically used by Leibniz (1702), for the purpose of clarifying the common notion of justice. For example, lack of punishment for offenders would not amount to a contribution to equal treatment but should rather be viewed as a violation of the requirement of equality.

In contemporary philosophy, this line of reasoning is still influential in the doctrines which identify a difference between tolerable inequalities and other inequalities, which should come to an end. The difference created by the display of personal merit offers a popular criterion for the differentiation between the two kinds of inequalities. People should not be held responsible for the differences between them that testify to the role of causes which are beyond their full control, such as the genetically determined features of their bodily constitution, or the early stages of the education they received, and even their talent and (dis)abilities. As they do not control such things, they should not be held responsible for their consequences in their own life; this has been described as a development of the idea of equality of opportunity, which is central to liberal and democratic thought (Freeman, 2007: p. 111). This opens the way to the critical reflection on collective attitudes toward social and economic inequalities, if we insist that these attitudes should testify to the commitment to suppress the contingent influence of accidental causes, thus enabling us to realize a significant kind of equal treatment in our social life (Dworkin, 1981; Cohen, 1989; Arneson, 2001; Fleurbaey, 2008; Spitz, 2008).

At the same time, this raises questions about the connection between equality and a fair treatment of individuals in their social life. According to some, the kind of respect for individuals that should be displayed by society must have its departure point in the serious consideration of individuals as they are. Their nature is not just the nature of rational persons: they are also constituted as individuals by the contingencies of life, through their own particular experience as living beings. Therefore there would be nothing partial or unfair about letting them being treated in a differentiated manner due to their contrasted ability to benefit from their own personal features (gifts or handicaps, social origin, age, etc.) and social involvement (educational connections, friendships, family links, etc.). On the contrary, if fairness is supposed to commit us to nullify the impact of contingent factors which are beyond the control of individuals, active intervention is required by the imperative of equal treatment. Along the same lines, it can even be argued that fairness ideals commit us to the active support of 'fair' inequalities (those economic inequalities which result from contrasted workplace effort, educational initiative, or active social involvement in collectively useful directions).

See also: Contractarianism; Human Rights and Social Work; Rawls, John (1921–2002); Rousseau, Jean-Jacques (1712–78).

Bibliography

Anderson, E., 1999. What is the point of equality? Ethics 109 (2), 287–337.

Arneson, R.J., 1989. Equality and equal opportunity for welfare. Philosophical Studies 56 (1), 77–93.

Arneson, R.J., 2001. Luck and equality. Proceedings of the Aristotelian Society 75, 73–90. Suppl.

Carter, I., 1999. A Measure of Freedom. Oxford University Press, Oxford.

Cohen, G.A., 1989. On the currency of egalitarian justice. Ethics 99, 906–944.

Condorcet, J. A-N. C. de, 1849. In: Arago, F., O'Connor, Mme (Eds.), Oeuvres. F. Didot, Paris (F. Frommann Verlag, Stuttgart and Bad Cannstatt, 1968).

Dworkin, R., 1981. What is equality? part I: equality of welfare; part II: equality of resources. Philosophy and Public Affairs 10, 185–246, 283–345.

Fleurbaey, M., 2008. Fairness, Responsibility, and Welfare. Oxford University Press, Oxford.

Fraser, N., Honneth, A., 2003. Redistribution of Recognition? A Political-Philosophical Exchange. Verso, London and New York.

Freeman, S., 2007. Justice and the Social Contract. Oxford University Press, Oxford.

Gamel, C., Lubrano, C. (Eds.), 2011. On Kolm's Theory of Macrojustice. Springer, Berlin and Heidelberg.

Guibet Lafaye, C., 2006. Justice sociale et éthique individuelle. Presses de l'université Laval, Québec.

Guillarme, B., 1999. Rawls et l'égalité démocratique. Presses universitaires de France, Paris.

Habermas, J., 1992. Faktizität und Geltung. Suhrkamp, Frankfurt (Engl. transl. by W. Rehg, 1998. Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy. MIT Press, Cambridge, Mass.).

Harsanyi, J.C., 1953. Cardinal utility in welfare economics and the theory of risk-taking. Journal of Political Economy 61, 434–435.

Harsanyi, J.C., 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. Journal of Political Economy 63, 309–321.

Honneth, A., 1992. Kampf um Anerkennung. Suhrkamp Verlag, Frankfurt (Engl. transl.: The Struggle for Recognition: The Moral Grammar of Social Conflicts. Polity Press, Cambridge, 1995).

Kolm, S.-C., 2005. Macrojustice, the Political Economy of Fairness. Cambridge University Press, Cambridge.

Leibniz, G.W., 1702/1994. Méditation sur la notion commune de justice. In: Sève, R. (Ed.), Leibniz: le droit de la raison. Vrin, Paris, pp. 107–136.

Locke, D., 1981. The principle of equal interests. The Philosophical Review XC (4), 531–559.

Nadeau, C., 2007. Justice et démocratie. Les Presses de l'université de Montréal, Montréal.

Phillips, A., 1999. Which Equalities Matter? Polity Press, Cambridge.

Putnam, H., 2002. The Collapse of the Fact-Value Dichotomy and Other Essays. Harvard University Press, Cambridge, Mass.

Rawls, J., 1967. Distributive justice. In: Laslett, P., Runciman, W.G. (Eds.), Philosophy, Politics, and Society. Blackwell, Oxford, pp. 130–153. Reprinted as chap. 7. In: Freeman, S. (Ed.), Collected Papers. Cambridge, MA and London, 1999.

Rawls, J., 1971/1999. A Theory of Justice. Harvard University Press, Cambridge, MA.

Scheffler, S., 2003. What is egalitarianism. Philosophy and Public Affairs 31 (1), 5–39.

Sen, A.K., 1992. Inequality Reexamined. Cambridge University Press, Cambridge, MA.

Spitz, J.-F., 2008. Abolir le hasard? Responsabilité individuelle et justice sociale. Vrin, Paris.

Equilibrium: History of the Concept

Till Düppe, Université du Québec à Montréal (UQAM), Montréal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Equilibrium analysis largely identifies the scope of theory in today's economic research. This prominent place has several historical origins. Before the notion was first used, equilibrium arguments were already applied in early modern writings on trade. Not before the late eighteenth century, the concept became part of the vocabulary of political economists. It was used as an analogical and, moreover, normative concept referring to the natural order of the economy. At the end of the nineteenth century, the concept gained analytical meaning, specifically because of the difference between static and dynamic equilibrium. Between the 1940s and the 1960s, the concept grew into a metatheoretical notion as the corner stone of axiomatic general equilibrium analysis, which in turn would become the bedrock of current mathematical modeling. Equally in game theory, the concept is the main analytical tool, though not limited to competitive interactions in markets. It allows formalizing institutional environments and strategic interaction on an individual and group level. In nearly all macroeconomic modeling, general equilibrium theory continues to be the prevailing reference point.

Equilibrium analysis largely identifies the scope of theory in today's economic research. Since there are many economic theories, the concept refers only vaguely to specific beliefs or ideas about economic reality. Instead, it provides cohesion to the economic way of modeling in the social sciences, and thus corresponds with a 'style of reasoning' typical for the discipline of economics (Hacking, 1992).

The prominent place that the concept occupies in economics today has several origins. Before the notion was first used, equilibrium arguments were already applied in early modern writings on trade. Not before the late eighteenth century, the concept became part of the vocabulary of political economists. It was used as an analogical and, moreover, normative concept referring to the natural order of the economy. At the end of the nineteenth century, the concept gained analytical meaning, specifically because of the difference between static and dynamic equilibrium. Between the 1940s and the 1960s, the concept grew into a metatheoretical notion as the corner stone of axiomatic general equilibrium analysis, which in turn would become the bedrock of current mathematical modeling. Equally in game theory, the concept is the main analytical tool, though not limited to competitive interactions in markets. It allows formalizing institutional environments and strategic interaction on an individual and group level. In nearly all macroeconomic modeling, general equilibrium theory continues to be the prevailing reference point despite its limited use for empirical analysis.

Since the concept functioned differently at each point of time, one better views its history as a sequence of transformations of different, not necessarily reducible scientific cultures in economics. The history of the concept is nevertheless significant insofar as in current economic research as well as in economic policy more recent and at the same time older meanings of the concept are used interchangeably.

Early Forms of Equilibrium Arguments

Before the concept was used explicitly, equilibrium arguments can be traced back to early modern economic writings of the seventeenth century. Since the form of argument is fairly simple, one could without doubt find yet earlier uses. The discursive situation from which these arguments emerged was early modern writings on trade, mostly pamphlets written in the context of the council of trade. Equilibrium arguments provided a rationale for fixing quantitative measures such as price limits, tax levels, import quotas, etc. Previous arguments that used the Aristotelian and medieval concept of the just price, lacked such a rationale. To give but one example: William Petty, when speaking of the limit between a reasonable rate of interest and illegitimate usury, argued that the interest rate should be as high as the rent of land, for *if not*, financial investment would soon outweigh agricultural production (Petty, 1899: 48). Similar arguments regarding excess supply and demand can be found in the writings of Sir Dudley North or John Locke (Letwin, 1963: 198; Tieben, 2012). Equilibrium arguments are a form of hypothetical reasoning that replaces normative arguments about economic measures such as price ceilings or tax levels. They require a notion of interdependence of two or more economic variables, the change of one of which induces the change of the other and vice versa. The causal mechanism of this endogenous change, however, must not be fully specified. Hence, equilibrium arguments were present in economic writings before the concept was explicitly used in terms of a specific ontology of the nature of economic order.

Not all arguments involving interdependence, however, are equilibrium arguments, as, for example, Francoise Quesnay's *tableau économique* that traces the causal effects of growth in one sector, agriculture, onto all others (Quesnay, 1766). Even though equilibrium arguments do not require full causal explanations, they equally should be kept apart from identity arguments that were also frequently used in seventeenth- and eighteenth-century economic writings. The so-called Say's law might be the best-known example that states that (aggregate) supply and demand cannot diverge because they are mutually determined. Identities cannot be unequal and thus serve as principles from which equilibrium arguments can be constructed. An example of a very influential early equilibrium argument that shows the difference between identity,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03081-6

equilibrium, and causality is Hume's so-called price specie flow (1752: 82ff.). Hume argued that a trade surplus cannot be maintained in the long run, as mercantilist writers assumed, because the inflow of gold would increase domestic prices that in turn leads to less export and more import until the balance of payments is restored. The identity of price level and money supply is assumed, the equilibrium argument is hypothetical, and the actual causal mechanism of the economic agents who respond to price differentials is unspecified.

Analogical Uses of the Concept of Equilibrium

In order to possess explanatory value, equilibrium arguments must include an ontological dimension beyond hypothetical reasoning: the nature of the causal mechanism that brings about endogenous change must be specified. Classical political economists established this explanatory value of the concept by means of the analogy between economic and natural order. Elements of the market economy, say, different parties of trade, relate to one another like elements in other domains where the belief in an ordered system was already established. This analogical usage of the term was decisive for establishing the belief in an orderly sphere of the economy as an object of systematic inquiry.

The very first appearance of the concept in the history of social thought is difficult to identify because the concept has been already established in other than social writings where it described mechanical movements, interstellar forces, anatomic proportions, which often had, typical for eighteenth-century authors, also creationist connotations. The French writer Pierre de Boisguilbert (1646–1714), to whom Spiegel assigns first usage, used the term when comparing economic and divine order (Spiegel, 1991: 614). For him, equilibrium refers to a state of opulence. He also used the analogy with mechanics according to the original Latin root of the word – 'equal balance' – in order to explain the price mechanism:

> It is like a balance suspended in equilibrium because the weight is equal in each of the two scales. A hundred pounds of any material whatever cannot receive an increase of even two pounds in one of the two scales without destroying the equilibrium. This, then, is precisely what takes place with regard to the price of grain. An increase or a diminution of twenty sacks at the market or staple house above or below the ordinary supply of five hundred sacks offered every week, destroys the balance, and makes it fall immediately on one side.
>
> Boisguilbert, 1966[1707]: 859

After such occasional uses in the early eighteenth century, the concept became part of the vocabulary of political economists during the Scottish enlightenment of the late eighteenth century. While the concept was also used by James Steuart in 1769 (see Milgate, 2008), the *locus classicus* of the analogical concept of equilibrium is Adam Smith when speaking of the 'gravitational force' exerted by the natural price (Smith, 1961 [1776]: 60ff.). The natural price, according to Smith, reflects (long-run) production costs while the market price varies according to short-term deviations of demand or supply. These deviations are 'forced' back to the natural price by means of rivalry between either consumers or producers.

> The natural price … is, as it were, the central price, to which the prices of all commodities are continually gravitating. Different accidents may sometimes keep them suspended a good deal above it, and sometimes force them down even somewhat below it. But whatever may be the obstacles which hinder them from settling in this center of repose and continuance, they are constantly tending towards it.
>
> Smith, 1961[1776]: 65

'Gravitational forces' of the market refer to the effects of changes of demand (that is degrees of wants for conveniences and of needs for necessities) given the limitations of production (that are determined by local natural resources, technology, capital maintenance, and labor costs). Such are the economic 'forces' to be brought into balance by the price mechanism that imposes costs and benefits on decisions of producers and consumers. Being in rivalry with one another, they react to price differentials.

Next to physical or celestial analogies, anatomic analogies are also frequently used in economic writings of the eighteenth century. After William Harvey's discovery of blood circulation, economic writers commonly described trade in anatomical terms. Money flows through the economy like blood through veins. Thomas Hobbes was one of the first, John Pollexfen, Jacob Vanderlint, Philip Cantillon, and others followed in calling the public treasure the heart of the state and money its blood that makes its organs reach out to other states – e.g., by paying armies (see Walter, 2011: 40–47). This anatomical analogy held both ways to the extent that the notion of the economy was habitually used in medical treatises when one spoke of the economy of the body and its natural equilibrium. Diseases were defined in terms of the divergence from this equilibrium, such that equilibrium took on older meanings of moderation and lack of excess.

These analogical uses helped establishing the field of political economy as an inquiry into an orderly object of study. It established the belief that "certain systematic or persistent forces, regular in their operation, are at work in the economic system," which is the most general feature of the concept of equilibrium according to Milgate (2008). The analogy was put forward against the belief that economic phenomena are manifestations of the blind forces of power and thus are not intelligible or predictable in systematic terms. This contrast with the concept of power will be present throughout further transformations of the concept.

The analogical use also established the very prominent normative meaning of a state of harmony of interests in a free-market society. Smith's natural price implied harmony between the three classes of labor, capitalists, and landowners. His analogy nourished the idea that once left alone, the economy spontaneously forms its inner order such that market interventions seeking power over the economy disturb this natural order, and thus the harmony of interests. Free competition allows the natural forces of the economy to bring about an orderly, harmonious, and thus desirable state of the economy. Even though Smith also propagated many government responsibilities, he thought of free trade as the system of natural liberty: complete resource mobility and free entry and exit in all markets. His system of natural liberty referred to institutions that induce rivalry between economic agents that are necessary for the 'gravitational forces' of the market to be at work (Smith, 1961[1776]: 687).

Classical political economists hardly ever felt obliged to spell out the analogy of the concept of equilibrium to such extent that the nature of economic forces could be empirically tested or even analytically determined. In fact, if used solely in an analogical fashion without the hypothetical reasoning referred to in the previous section, classical political economists added little to the normative arguments of natural law philosophers. In turn, the normative meaning of the concept of equilibrium, until today, has never been fully argued for without analogical elements (see Düppe, 2011).

The fact that the equilibrium concept rather imported familiarity into economic discourse than provided explanations of the market mechanism was still true when late-nineteenth century neoclassical writers used the concept, such as William Stanley Jevons. His analogy was that of the equilibrium of prices and that of a pendulum:

> We can no more know nor measure gravity in its own nature than we can measure a feeling; but, just as we measure gravity by its effects in the motion of a pendulum, so we may estimate the equality or inequality of feelings by the decisions of the human mind. The will is our pendulum, and its oscillations are minutely registered in the price lists of the markets.
>
> Jevons, 1970[1871]: 11

Already before Jevons, however, notably with John Stuart Mill, skepticism regarding the empirical merits of the analogy emerged. Mill emphasized that economic laws hold only under very restrictive conditions that can, in contrast to natural laws, never be fully observed in reality (because of the absence of controlled experiments). Due to noneconomic interfering causes only tendencies toward states of equilibrium can be observed. The classical laws of the market are true only 'in the abstract,' that is, "they are only true under certain suppositions, in which none but general causes – causes common to the whole class of cases under consideration – are taken into account" (Mill, 1844: 144–145). In this fashion, Mill opened the way to the analytical notion of the concept of equilibrium, specifically by pointing to the significant fact that initial conditions of the equilibrium, in the concrete, may change 'on the way,' and that hence the static idea of a market equilibrium must be supplemented with a theory of dynamic change (Mill, 1848: 421ff.). Thus, Mill introduced a distinction that was to remain present in all equilibrium analysis: the difference between static and dynamic equilibrium analysis.

The Analytical Concept of Equilibrium

It was only in the last decades of the nineteenth century, specifically after the work of Léon Walras (1834–1910) that equilibrium became a theoretical concept that could be analyzed in its own terms without reference to other domains of research. Decisive was the idea of representing the economy as a system of equations in Walras's *Elements of Pure Economics* (2003[1874]). Similar to nonconceptual uses of the word when one speaks of a balance of payment to be in equilibrium, equilibrium as the solution of a system of equations refers to the state of market clearance where aggregate demand is equal to aggregate supply. Equations are separately defined for demand and supply in all goods and factor markets that each depend on the price of the good of this market and all other goods and factor prices being either complements or substitutes. Equilibrium refers to a set of prices that solves all equations simultaneously.

Even if Walras's inspiration was a textbook by Louis Poinsot on the mechanical equilibrium of a solar system (Ingrao and Israel, 1990: 88), and even if later in his life he defended his approach by an analogy between economic and mechanical concepts (scarcity and utility relate to one another like force and energy, see Mirowski and Cook, 1990), Walras did not believe that economic and natural phenomena are reducible to one another. He moved beyond analogical reasoning insofar he put emphasis on the common mathematical structure while recognizing the ontological differences between these domains. Accordingly, Walras distinguished, similar to Mill, between pure, applied, and social economics, while only in the context of pure economics, an equilibrium could be properly defined. Little ontological commitment was transmitted by Walras's analytical concept of equilibrium. Instead, he opened the door to mathematical modeling, specifically to linear algebra.

The aim of Walras's system was the representation of the interdependence of all markets. Agents all over the economy that appear to be unrelated are connected by the price system. Equilibrium thus contrasts to fragmentation of the economy, in which case one could say very little about a market economy as a whole. This interest in capturing interdependence is often contrasted with Alfred Marshall's (1920[1890]) analysis of markets: the latter representing partial, the former general, equilibrium theory. Marshall sought to understand the mechanism of single markets, *ceteris paribus* all effects on other markets. What allowed Marshall to do so was a distinction between causal factors being relevant for different periods of time (1890: Book V). He distinguished, first, the market period in which goods had been already produced, supply is thus fixed, such that prices depend solely on demand; second, the short-run period in which production can be expanded altering some production factors, but not all due to prime costs; third, in the long-run all production technologies can change, while demand is stable. Consequently, as to whether or not one considers being in equilibrium depends on the period of time one considers. Such was Marshall's take on the problem of dynamics – the problem that the endogenous changes induced by a disequilibrium changes the initial conditions of equilibrium, such as income and technology.

Walras's instead, in his framework of general equilibrium analysis, bypassed the problem of dynamic change by speaking of an 'auctioneer' who identifies the set of equilibrium prices by 'crying out' random prices before any transaction is taken (Jaffe, 1967). This comparison with the stock market was clearly not an identification of the causes that bring prices into equilibrium, as Marshall would have it, but the possibility of an algorithm that describes the change of prices within his system of equations. The Walrasian auctioneer, as an analytical creation, allowed for equilibrium theory to develop theoretical problems of its own prior to its applications. When Walras considered actual price changes, he drew back and referred once again to a natural analogy, a lake agitated by the wind, saying that initial conditions are in continuous change due to the changes caused by a disequilibrium (Walras, 2003[1874]: 380).

Nature, no longer a source of the belief in an orderly domain, changes faster than the market can adapt.

Next to the static and dynamic equilibrium, another analytical difference that became prominent from contrasting Marshall and Walras was that of aggregation and disaggregation: Can the conditions of market equilibrium be traced down to the individual level? Marshall assumed what is nowadays called a representative agent such that even the marginal utility of income was supposed to be the same across all income levels. Walras equally started from aggregate market demand and supply functions, even if his set of equations could in principle be spelled out in terms of disaggregated individual functions – or so he and other neoclassical economists assumed for a long time until the opposite was proven in the 1970s, as we see below.

In the hands of Walras, a significant transformation of the concept took place regarding its normative welfare dimension. Perfect competition for Walras means that prices are independent of demand and supply functions, i.e., they are 'given' to economic agents. Gone was the close link between rivalry and free-market institutions such that it was no longer obvious in which institutional context equilibrium theory would apply. This was indeed not the task of pure economics according to Walras. As a result, the normative meaning of the concept of equilibrium was no longer the harmony of interests and the natural system of liberty, but the absence of waste and the efficient use of resources as the formal implication of market clearance. Neoclassical welfare economics is in fact less an argument for specific economic institutions or policies, but the development of criteria according to which various institutions and policies, interventionist or not, could be evaluated.

Equilibrium and Time

The distinction between static and dynamic equilibrium remained the driving force of innovations involving the concept of equilibrium in the first half of the twentieth century. An important complication of the problem of dynamics was, rather than income effects, the uncertainty of long-term investment decisions that has short-term effects. An important consequence of this uncertainty is that there is no reason to trust that profit rates are unique across all markets. And so the concept of equilibrium was in the background of a range of institutional questions regarding monetary and business cycles theories that were put forth by Austrian economists. Böhm-Bawerk, for example, spoke of an intertemporal equilibrium considering both long-term and short-term plans simultaneously in a trade off between current and future consumption (Böhm-Bawerk, 1959[1899]). In further course, including uncertainty in equilibrium analysis, most notably after John Maynard Keynes, generated a whole bunch of new analytical problems that were largely unknown to political economists, and, in further course, only vaguely linked with traditional price theory. What came to be labeled as macroeconomics gained shape in a certain distance to equilibrium analysis.

It was not before John Hicks (1904–89), in response to the work of Lindahl and Hayek, and in trying to reconcile Keynesian and Walrasian theories, that a different interpretation of equilibrium was introduced to solve problems of time and uncertainty: the temporary equilibrium (Hicks, 1939). Hicks specified the process by which an equilibrium comes about in *sequential* terms. He spoke of a market period called a 'week' defined as the period in-between transactions that allow for carrying out production plans. Equilibrium was thus no longer defined in terms of long-run forces and short-run disturbances. Adaptations and decisions were sequentially separated. Insofar as at each transaction day a new equilibrium is achieved Hicks bypassed problems of the coordination of expectations. At each period of time there are different market clearing prices.

Also the normative dimension of the equilibrium concept was heavily influenced by these new problems of time, as is evident in Oskar Lange's idea of market socialism (1938). Lange reinterpreted Walras's general equilibrium theory as an economic theory of socialism with public ownership of production factors. The central planning bureau sets prices like a Walrasian auctioneer, observes excess demand and supply via inventory data, and reaches equilibrium by adapting factor prices accordingly. Against this use of equilibrium analysis, Friedrich Hayek put forth a concept of competition that is opposed to any notion of equilibrium: competitive institutions are preferable as they result in the most adaptive utilization of local knowledge. It is only in a continuously changing environment, that is, in disequilibrium, that the price mechanism shows to be superior to a bureaucratic system (Hayek, 1945).

Axiomatic Equilibrium Analysis

Between the 1940s and the 1960s the concept of equilibrium became the central means by which mathematical modeling was advanced, and conceptual and other literary economic theory were replaced. The concept of equilibrium became the symbol of the culture of economics being different from other social sciences (Düppe and Weintraub, 2014). The joint of this development was the application of the axiomatic approach in mathematics to general equilibrium theory. The early contributions to this development were mathematical existence proofs written by John von Neumann, John Nash, Kenneth Arrow, Gérard Debreu, and Lionel McKenzie (using various versions of Brouwer's fixed-point theorem, see Weintraub, 1983; Debreu, 1987). This launched the period of several decades of advancing mathematical tools in order to determine axioms that are necessary to prove, next to existence, uniqueness, stability, and the optimality of an equilibrium (Ingrao and Israel, 1990). By the 1970s, definitions of equilibrium conditions (the axioms) became a necessary requirement for all economic theory.

The axiomatization of equilibrium theory was, similar to mathematics at the same time, coined by an 'image conflict' (Dalmedico, 2001). In the spirit of John von Neumann, several mathematical economists understood the axiomatic approach as a clarification of the behavioral and institutional elements of equilibrium theory: producers and technologies, consumers and commodities. Axioms are seen as basic ontological properties that must be true for an equilibrium in a competitive economy to possibly exist, to be empirically determined (uniqueness), to be stable, and to have desirable welfare properties (optimality). The early existence proofs were

considered a preliminary exercise that would generate new solutions to traditional questions (Arrow, 1968). The success of the axiomatic method would depend heavily on the plausibility of its axioms as being descriptive of economic reality. Specifically the rationality of optimizing agents emerged as the pivotal condition of equilibrium.

According to another image of the axiomatization of equilibrium theory, however, the mathematical structure of the theory is only loosely related to its possible interpretations. Innovations of new mathematical tools can then enlarge, limit, or create new forms of explanations on the level of economic theory. Several mathematical economists, many of which were associated with Gérard Debreu, focused on the development of new mathematical tools such as measure theory and nonstandard analysis, among others. The guiding epistemic virtue was mathematical generality (which does not necessarily correlate with greater economic applicability): topology being more general than, for example, differential equations. As questions of stability, among others, required tools that are lower in this hierarchy than those necessary for existence proofs, the latter played a greater role than the former. The pivotal axioms that emerged from this axiomatic approach were properties such as convexity, monotonicity, differentiability, continuity rather than the behavioral notion of rationality. Moreover, equilibrium analysis was not related to competitive institutions, a concept that does, for example, not appear in Debreu's *Theory of Value* (1959). The axiomatic method results in the separation of mathematical and economic reasoning that has indefinite and uncontrolled effects on the latter.

This image conflict was hardly ever confronted such that there was never a consensus among practitioners in economic theory what the exact status of axioms is: hypothesis, basic ontological property, or auxiliary construct for creating a plurality of contextual theories (Düppe, 2012). Accordingly, the concept of equilibrium, until today, plays a double role in economic theory: it refers to the analytical solvability of a model, and at the same time to the solution concept of actual economic problems by modeling the consequences for an entire system. After the diffusion of general equilibrium models in the entire discipline during the 1970s, the concept drove a wedge between the analysis of theoretical structures and the elaboration of economic interpretations.

Within the inner circles of mathematical economists, however, the research program of axiomatic general equilibrium theory showed rather negative results: stability could not be specified, uniqueness of an equilibrium required ad-hoc assumptions that were considered rather strong, and at the top of it, in 1974, Hugo Sonnenschein, Rolf Mantel, and Debreu showed that excess demand functions lack structure even if one assumes well-behaved, i.e., rational individual preferences. The theory imposes little structure on empirical data. Even simple questions of comparative statics (such as the effects of tax cut on demand and supply) that were already by the 1960s and still are standard in economic education could no longer be considered rigorous. One has to impose additional structure in order to generate empirically definite results (such as constant elasticity of utility functions, a specific income distribution, and the like).

By the mid-1970s, the hopes of early post-World War II years for the future of general equilibrium theory were largely disappointed. Mathematical economists such as Frank Hahn, however, considered these negative results most valuable insofar as they allow for a falsification of the beliefs that were traditionally linked with the notion of equilibrium, above all its normative aspects (Hahn, 1973). In terms of the present article, axiomatic analysis of equilibrium theory helped correcting the beliefs that the analogical concept of equilibrium has established.

Game Theoretical Concepts of Equilibrium

After John von Neumann's and Oskar Morgenstern's seminal book, *The Theory of Games and Economic Behavior* (1944), also game theory, with a certain delay, transformed the concept of equilibrium. One research question that has been pursuit during the 1960s was specifically formative, that of 'core convergence.' The core is an equilibrium concept going back to Edgeworth's theory of exchange: it includes all allocations, for which no subset of individuals would be better off, if they separated from the other individuals (in the case of two individuals, the core excludes those allocations, which leaves individuals worse than alone). The question of core convergence was whether or not, once the amount of individuals increase, the set of core allocations shrinks toward the competitive price solution. Core convergence thus opened a path to integrate game theory in the tradition of economic theory enlarging the scope of equilibrium theory from competition to strategic behavior. In fact, competitive behavior appeared to be a special case of noncooperative games – 'oceanic games,' as Milnor and Shapley called them (1961). Core convergence helped in "rescuing" the standards of mathematical modeling that were established in general equilibrium theory where they led into the rather negative results mentioned above (Rizvi, 1994).

Historically speaking, this revival of game theory in the 1960s bore aspects of irony. The 'core' referred back to another, slightly more general concept in cooperative game theory, the 'stable set.' Cooperative game theory allows for some coalitions to be binding even if some individuals would be better off without partaking in this coalition. This was the central equilibrium concept in von Neumann's and Morgenstern's book (1944). Mostly because of the interpretations of the Austrian trained economist Morgenstern, game theory was initially presented in contrast to competitive equilibrium as the latter excluded the exertion of strategy and power. This opposition was forgotten once competition was considered to be analytically reducible to game theory.

The most dominant equilibrium concept in game theory, however, is the noncooperative solution concept named after John Nash: no player can gain by changing behavior unilaterally taking into account the behavior of other players. The existence proof of such equilibrium was an inspiration also for Arrow and Debreu to proof existence in general equilibrium theory (Düppe, 2012). Further refinements of the Nash equilibrium became central to all game theory, causing an entire industry of research elaborating information constraints imposed on all or a subset of players: different kind of games (asymmetric games, sequential games, mixed strategies, evolutionary games where players act according to (genetic) traits), different market forms (for

example, Cournot equilibrium in a duopolistic market, or Lindahl equilibrium as a solution to public good problems), but also different types of auctions. Game theory, much in contrast to general equilibrium theory, is thus considered descriptive of behavioral content and institutional contexts. None of its institutional applications, however, replace the concept of Nash equilibrium.

Game theory also proved to be crucial for the import of new empirical methods into economics, such as cognitive theories and experimental methods. The results, however, concern the behavioral assumptions, but not the concept of equilibrium itself. It is still difficult to imagine a theory identifiable as economic theory without reference to the concept of equilibrium.

Equilibrium Analysis in Today's Macroeconomics

Despite the fact that competitive general equilibrium analysis, as shown above, has turned out to imply little for empirical research, it continues to be the sheer exclusive reference point, in combination with econometric methods, for all macroeconomics theories. General equilibrium theory has been omnipresent ever since the work of Robert Lucas, Thomas J. Sargent, Edward C. Prescott, and among others (Snowdon and Vane, 2005). It is used as the analytical solution to so-called dynamic stochastic general equilibrium models. These models use the general equilibrium framework subject to stochastic shocks in order to explain historical time series data given calibrated exogenous variables. The analytical problems stated in axiomatic equilibrium analysis are bypassed by imposing structure on economic agents (usually Cobb–Douglas functions, which brings back the representative agent of Marshall's partial equilibrium analysis).

After the 2008 financial crisis, dynamic stochastic general equilibrium (DSGE) models and with it the imperative of equilibrium analysis was put under heavy attack (Colander, 2010). Even though in newly emerging fields such as in econophysics the concept of equilibrium appears to be expendable, since agent-based simulation techniques bypass analytical solvability of mathematical modeling, the great mass of macroeconomic theory is far from incorporating these tools and moving away from equilibrium analysis.

See also: Axiomatic Theories; Cooperative Game Theory; Discipline Formation in the Social Sciences; Economics, History of; Game Theory: Noncooperative Games; Game Theory; Liberalism: Political Doctrine and Impact on Social Science; Political Economy, History of; Social Mechanism; Welfare Economics.

Bibliography

Arrow, Kenneth J., 1968. Economic equilibrium. In: International Encyclopedia of the Social Sciences, vol. 2. Harvard University Press, Cambridge, MA. Reprinted in The Collected Papers of Kenneth J. Arrow.

Böhm-Bawerk, Eugen, 1959[1899]. Capital and Interest. Reprinted. Libertarian Press, South Holland, IL.

Boisguilbert, Pierre de, 1966[1707]. Traité de la nature, culture, commerce et intérêt des grains. In: Pierre de Boisguilbert ou la Naissance de l'Économie Politique, 2 vols. Institut National D'Études Démographiques, Paris, pp. 827–878.

Colander, David, 2010. The economics profession, the financial crisis, and method. Journal of Economic Methodology 17 (4), 419–427.

Dalmedico, Amy D., 2001. An image conflict in mathematics after 1945. In: Bottazzini, U., Dalmedico, A.D. (Eds.), Changing Images in Mathematics: From the French Revolution to the New Millenium. Routledge, London and New York, pp. 223–253.

Debreu, Gérard, 1986. Theoretic models: mathematical form and economic content. Econometrica 54 (6), 1259–1270.

Debreu, Gérard, 1987. Existence of general equilibrium. In: Durlauf, Steven N., Blume, Lawrence E. (Eds.), The New Palgrave Dictionary of Economics.

Debreu, Gerard, 1959. Theory of Value: An Axiomatic Analysis of Economic Equilibrium. Yale University Press, New Haven, CT.

Düppe, Till, 2011. How economic methodology became a separate science. Journal of Economic Methodology 18 (2), 163–176.

Düppe, Till, 2012. Arrow and Debreu de-homogenized. Journal of the History of Economic Thought 34 (4), 491–514.

Düppe, Till, Weintraub, E. Roy, 2014. Finding Equilibrium: Arrow, Debreu, McKenzie and the Problem of Scientific Credit. Princeton University Press, Princeton.

Hacking, Ian, 1992. 'Style' for historians and philosophers. Studies in History and Philosophy of Science Part A 23 (1), 1–20.

Hahn, Frank, 1973. On the Notion of Equilibrium in Economics. Cambridge University Press, Cambridge.

Hayek, Friedrich A., 1945. The use of knowledge in society. American Economic Review 35 (4), 519–530.

Hicks, John R., 1939. Value and Capital. Clarendon Press, Oxford.

Hume, David, 1752. Political Discourses. Oxford University.

Ingrao, Bruna, Israel, Georgio, 1990. The Invisible Hand: Economic Equilibrium in the History of Science. MIT Press, Cambridge, MA.

Jaffe, William, 1967. Walras theory of tatonnement: a critique of recent interpretations. Journal of Political Economy 75 (1), 1–19.

Jevons, William S., 1970[1871]. Theory of Political Economy. Edited from the second ed. (1879) by R.D.C. Black. Penguin, Harmondsworth.

Lange, Oskar, Taylor, Fred, 1964[1938]. On the Economic Theory of Socialism. McGraw Hill, New York.

Letwin, William, 1963. The Origins of Scientific Economics: English Economic Thought 1660–1776. Methuen, London.

Marshall, Alfred, 1920[1890]. Principles of Economics, eighth ed. Macmillan, London.

Milgate, Murray, 2008. The development of the concept of equilibrium. In: Durlauf, Steven N., Blume, Lawrence E. (Eds.), The New Palgrave Dictionary of Economics, second ed. Palgrave Macmillan.

Mill, John Stuart, 1844. Essays on Some Unsettled Questions of Political Economy, 1874, reprinted, second ed. Augustus M. Kelley, New York.

Mill, John Stuart, 1848. Principles of Political Economy, 1871 (reprinted 1909), sixth ed. Longmans, Green & Company, London.

Milnor, John W., Shapley, Lloyd S., 1961. Value of Large Games, II: Oceanic Games. RAND Memorandum.

Mirowski, Philip, Cook, Pamela, 1990. Walras's economics and mechanics: translation, commentary, context. In: Samuels, Warren (Ed.), Economics as Discourse. Kluwer, pp. 189–215.

Neumann, John von, Morgenstern, Oskar, 1944. Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ.

Petty, William, 1899. In: Hull, Charles H. (Ed.), The Economic Writings of Sir William Petty: Together with the Observations upon the Bills of Mortality (More Probably by Captain John Graunt). Cambridge University Press, Cambridge.

Quesnay, François, 1766. Analyse de la formule arithmétique du Tableau économique de la distribution annuelle d'une nation agricole. Journal de l'Agriculture, du Commerce et des Finances 5 (3), 11–41.

Rizvi, Abu Turab S., 1994. Game theory to the Rescue? Contributions to Political Economy 13, 1–28.

Smith, Adam, 1961[1776]. In: Cannan, E. (Ed.), An Inquiry into the Nature and Causes of the Wealth of Nations. Methuen, London.

Snowdon, Brian, Vane, Howard R., 2005. Modern Macroeconomics: Its Origins, Development and Current State. Edward Elgar.

Spiegel, Henry W., 1991. The Growth of Economic Thought, third ed. Duke University Press, Durham.

Tieben, Bert, 2012. The Concept of Equilibrium in Different Economic Traditions: An Historical Investigation. Edward Elgar, Cheltenham, UK.

Walras, Léon, 2003[1874]. Elements of Pure Economics: Or the Theory of Social Wealth. Routledge, London.

Walter, Ryan, 2011. A Critical History of the Economy: On the Birth of the National and International Economies. Abingdon, Routledge, Oxon.

Weintraub, E. Roy, 1983. The existence of a competitive equilibrium: 1930–1954. Journal of Economic Literature 21 (1), 1–39.

Equity and Education

Salma Ismail, Faculty of Humanities, School of Education, University of Cape Town, Cape Town, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Equity and education are long-standing goals of many societies and there are diverse perspectives on ways to provide equal opportunities and resources for all. The literature reviewed reveals wide differences in educational outcomes both within and between countries and suggests that education policy and institutions on their own cannot solve structured inequalities. It is also argued that in a time of economic liberalization equity in education in many countries is more difficult to attain and that inequalities can be deepened. Research indicates that social mobility is more difficult to achieve for students from lower income households who continue to receive poor quality education and therefore social hierarchies and opportunity structures remain intact. The literature suggests that public policy can make a difference and cites Korea, Malaysia, and China as examples of countries, which have performed strongly in international tests. Researchers argue that what is required is collaboration between all involved to fulfill the promises that education can empower all citizens to participate in a globalized world.

Introduction

Equity is usually defined as fairness, impartiality, and justice and is related to equal opportunity (Sen, 2009). The world's progress toward these goals is usually measured by comparisons across countries and involves many disciplines. A review of the literature suggests that in many contexts equity inspired goals are not always realized and where they have been attained political and economic changes can reverse equity gains. The key reasons cited for not achieving equity are that not everyone has equal resources and structural poverty excludes those who are poor or cultural and religious practices are discriminatory against gender, age, diversity, or disability.

Education is seen as having positive effects on society by providing skills and promoting social stability, inclusion, and in creating modern democracies (Stromquist, 2005). It is also signaled as a major factor, which would make an impact on reaching equity goals and decreasing poverty.

Equity in education is usually linked to equal access to formal education opportunities and resources. The idea often assumes economic growth and social transformation. Equality of opportunity is generally considered to maximize the total social good, increases social gains, and can lead to diffusion of power to individuals. Education is seen as determining opportunity of outcomes, livelihood, and economic growth. The lack of education is often perceived as the most pervasive cause of inequality and underdevelopment. The immense technological changes that are synonymous with globalization have compressed time and space with the use of the Internet and electronic media. Alongside these are innovations in education technology which have given rise to new measures of equity and inequalities in education such as technological access and the digital divide. The new technologies have also renewed a focus on the significance of lifelong learning (LL) for adults.

Others argue that there are also prejudices against those who are not considered intelligent and this leads to variability so everyone does not receive the same education. In a similar vein, analysts and theoreticians conclude that the quality of education varies for different income levels with those at the lower end receiving an inferior education and thereby the levels of inequality remain static or increase unevenly.

The Gini coefficient a common measurement of income inequality within countries captures the ratio between the most and least wealthy in a country with zero equating to zero inequality and one indicates complete inequity. This gap between rich and poor has increased in most countries in spite of more spending on education. A comparison of the Gini coefficient between countries show a wide range with one of the most equal countries such as Denmark scoring 0.248 and one of the most unequal South Africa scoring 0.65 http://data.worldbank.org/indicator/SLPOV.GINI. The implications for equity in education is that those with less income are less likely to receive a good quality education or be able to access tertiary education to enable them to find better employment opportunities.

Similarly, the human development index which measures longevity, health, and knowledge track a higher score for the North than the South. The human poverty index which measures life expectancy, adult literacy, water quality, and underweight children confirms the above data and indicates that Central Africa is the worst off region, while Japan has the best score www.wacra.org/world%20Poverty%20MAPPDF. This has led social theorists to conclude that education is but one force among many to distribute resources (Stromquist, 2005).

Education has evolved from the pursuit to attain knowledge for itself to include to lead, train, and nourish, as well as to see the improvement of societies and nation-states. On an individual level the outcomes are measured by the attainment of knowledge, skills, and an understanding of what is learned and an increase in cognition. The learner's identity and character may undergo change and the learner may become more socially responsive and also be desirable that the learner's ambitions are expanded and curiosity sparked and he or she aspires to seek and to create more knowledge, to be critical, for personal or economic advancement or pleasure, for social purposes, justice, and to excel in educational pursuits.

On a societal and country level the outcomes are measured by comparisons done by international agencies, such as the World Bank, The Organisation for Economic Cooperation and Development (OECD) by using the Programme for International Student Assessment (PISA) and the International Association for the Evaluation of Educational Achievement. Some of the measurements used to determine the quality, effectiveness, and impact of education are, the improvement in the quality of life and longevity, adult literacy, knowledge held in a society, returns on education, numbers, and progress of girls in education, how well students are prepared for life and the global economy, the income gap, as well as social and political stability http://www.OECD-ilibrary.org/education/education-indicators–in-focus_22267077.

Education in a formal sense means an institutionalized system wherein one enters a program of learning which is certified. It includes preprimary, primary, secondary, and tertiary pursuits which involved academic or vocational knowledge and skills and is strongly linked to employment opportunities, social mobility, and transformation.

The state is the main provider of education at a primary and secondary school levels and in most countries subsidizes tertiary education. However, in a rapidly changing global context and with the financial instability since 2008, many governments have cut spending on higher education. In response, there has been a growth in distance education and privatized education provided by business and independent education institutions. Distance education provided by universities is a means to secure funds and to reach a global audience. These measures have meant that education for profit has gained in ascendancy in some countries and threaten to reinforce social hierarchies and opportunity structures.

Another economic factor, which has impacted on equity and education in the early twenty-first century is a downturn in the global economy. This has led to high rates of youth unemployment which subsequently impacts on access and success in education with implications for economic growth and social and political stability (Scarpetta et al., 2010). Researchers point to other challenges in achieving equity in education linked to technology and new forms of social media which have resulted in changes in the global culture especially among youth. These new methods of communication bring to the fore the need for new and complex understandings in dealing with tensions between local and global cultures that defy attempts at homogenization within and across education systems.

There is a radical viewpoint (Morley, 2012; Thompson, 2000) which suggests that education does not necessarily increase social mobility but can reproduce inequalities by providing opportunities and resources for those already established in the society. Breault and Lack (2009) argue that factors such as the family capital (incomes and educational level of parents) represent an essential barrier of educational possibilities of graduates of public high schools and also that there is a considerable inequality even of educational intentions for lower income pupils as issues of equity in these schools are steeped in deficit notions of fixing the problem. They further argue that children from families with higher social positions receive full secondary education and then go on to study in universities, while children from families with low family capital are limited by the general secondary education and then study in technical colleges. These views are supported by empirical research conducted by UNESCO and the OCED (2010) PISA study done in 2009 shows that many 15-year-olds who performed poorly in PISA tests are from socioeconomically disadvantaged backgrounds and that poor home backgrounds reinforce low-quality performances.

Nonformal and informal education or learning usually means noncredit learning, which takes place outside of formal institutions and can occur in the everyday. This kind of education is usually led by nongovernmental or civil society organizations, trade unions or occurs in social movement activities. Internships, apprenticeships, or on-the-job training are other forms of nonformal learning and can play a role toward accreditation. In many parts of the world like in India, or in some parts of Africa many people learn in the informal system as they do not have the resources or cultural practices forbid girls to attend school (Rogers, 1992, 2005).

Nonformal programs offering literacy programs have experienced uneven successes. The successes in Cuba and in the previously socialist and communist countries have not been replicated in most of the developing countries (OECD – World Inequality Database on Education). Women and youth in rural areas in developing countries still constitute the bulk of the target group for nonformal education (Rogers, 1992, 2005).

The impact of nonformal education is not easily measured and learning in social movements is usually measured by social change, how power is redistributed, and the growth of democracy (Rogers, 2005). These are significant measures of equality in new democracies such as in South Africa which gained a new democratic government in 1994 and in Tunisia and Egypt where, in 2012, social movements unseated the ruling powers. This phenomenon is now widely known as the Arab Spring. Govender and Chinna Rao (2012) point out that the victories in these countries over the past few years may not necessarily have replaced unequal relations or changed societies fundamentally.

Interventions to Achieve Equity in Education

The most significant intervention to achieve equity through education on a global scale has been to provide free and compulsory primary and secondary schooling for children or what is often referred to as basic education. The provision of basic education on a mass scale is seen as a significant factor in increasing people's life chances. A quote from UNESCO's International Education statistics of the global population of primary school age states that at present 655 million children are in school and 61 million of children are out of school. UNESCO estimates that this number has declined but of this half is in Africa (www.unesco-MDG.org).

According to UNESCO progress to reach the Millennium Development Goals (MDG) and the goals set by the Education for All movement, one of which is to have universal education for all by 2015 is hampered by the continuing wars in Africa, ethnic prejudices, migration, HIV/AIDS, and tuberculosis, as well as malaria. Other factors that contribute to not reaching the MDGs are that in some communities in particular in rural areas girls are often excluded from formal secular schooling in

spite of great efforts being made by governments and women's movements to bring awareness of the importance of schooling and also to offer single sex schooling. The increase in the statistics for girls in schools in urban areas can be attributed to modernization which has meant an increase in urbanization and therefore the necessity to be independent and education is seen as one means to achieve self-sufficiency (www.World-Inequality-Database-on-Education). The increase of rural education is another feature of a country's path to development and in some countries, for example, in China it is hoped that improved rural education will slow down the rural population's movement to the city. In a study to measure the quality of rural schooling in Malaysia conducted by Othman and Muijis (2012) they found that generally teachers perceived no differences between urban and rural schools in educational quality.

There has also been an increase in religious schools attributed to religious fundamentalism and an increase in home schooling as parents become disaffected and prejudiced against multiculturalism, and the academic standards of public schooling.

Another feature of globalization has been the continuous migrations of populations all over the world, as well as an increase in the number of economic and political refugees. An OCED study of 2006 concludes that 175 million people live outside their home country and in OCED countries there has been an increase of 5% in the immigrant population between 2000 and 2009. New forms of discrimination against immigrant populations or economic or political refugees are on the increase as these populations are excluded from participating in everyday activities in the host country. In some countries, where there is high unemployment migrant populations often become targets of abuse. Governments, which abide by the United Nations Refugee policies, have allowed migrants and refugees access to education as it is acknowledged by them as a basic human right. This policy is also seen as contributing to political stability; it allows for the attainment of citizenship, for refugees to make a positive contribution to social cohesion, and contributes to human development. There is very little research to monitor progress in this area and no proven statistics to indicate that these attributes have been achieved. The 2009 PISA study indicated that immigrant populations are a significant community in schools who may not receive an equitable education. The PISA study showed that generally immigrant populations underperform in school. Some of the factors identified to explain underperformance are language barriers, poor learning opportunities, and public policies that do not take into account the diversity of immigrant populations (OECD, 2010–PISA study 2009).

Another significant intervention has been to provide access to excluded groups of people by participation to be inclusive of gender, religion, age, race, indigenous groups, or marginalized groups. Access into higher education of new constituencies is seen as a new form of democratization because it widens participation and is a form of social justice and contributes to economic development and social harmony (Thomas and Tight, 2011; Thompson, 2000).

Alternative admissions policies known as Recognition of Prior Learning (RPL) or Assessment of Prior Education and Learning (APEL) for mature adults were much publicized in the 1990s in higher education. These policies are associated with LL and redress and allow for alternative access for mature adults who do not have the required formal qualifications to use prior learning to gain access to an education program or to gain credit toward certification (Thomas and Tight, 2011; Thompson, 2000).

Widening access has not always meant that previously excluded groups have progressed seamlessly or have successfully completed their education, or excelled in their educational attainments. This has led to much research in the field of how to assist excluded groups from remaining in the education system and also to excel, as well as to allow them to become part of an academic community as staff and students (Thomas and Tight, 2011; Thompson, 2000).

A number of recommendations from the research have led to various reforms in education some of which are to create inclusive admissions policies and practices while making education provision available at convenient times and places for working adults. Opening up access to poor students has required financial assistance and scholarships, as well as measures to support academic development and in some cases the duration of the academic program has been extended. These reforms have to confront critiques of lowering academic standards in particular where admission requirements are lowered. The emphasis in most elite universities is on creating a meritocracy and maintaining an elite group (Featherman et al., 2010; Thomas and Tight, 2011).

In developing countries which have entered the global economy and have suffered racial oppression, colonialism, and imperialism like South Africa, India, and Brazil, and lag behind in professional skills there have been huge investments in education in particular in higher education. There have been concerted attempts to produce a skilled and professional work force. The interventions in these countries to end the inequalities have been to enhance schooling opportunities, which include more liberal admissions criteria, and a financial aid scheme financed by universities and the governments or a loan system to enter both public and private institutions (Akoojee and Nkomo, 2011; Featherman et al., 2010; Motala, 2001).

However, many students in higher education are academically underprepared because of an uneven schooling system therefore they are encouraged to attend extended degree programs. Because of the low numbers of skilled black people or professionals in South Africa in this population category there are various forms of affirmative action such as setting targets or quotas for certain programs such as medicine or careers such as engineering as these are disciplines and careers which were not offered to Black African students in apartheid South Africa (Featherman et al., 2010).

In Brazil, private higher education absorbs 70% because the public system is viewed as inefficient. This means that those who cannot afford the fees in private institution receive a low-quality education. In India, the impediments to access are poverty, the caste system, and gender discrimination impeded access to education. India has seen an increase in higher education and is now the second largest higher education sector in the world. The Indian state subsidizes higher education by almost 90% of costs and this has led to a huge expansion in higher education. However, state education is not always of same quality as in private institutions therefore there

has been a call for meaningful education for all (Akoojee and Nkomo, 2011).

Measures worldwide for schooling in certain countries have focused on improving equity by improving the quality of teaching which includes teacher training, encouraging progressive teaching methods rather than more didactic models that emphasize technical rationality. Other interventions for teachers have ranged from incentives in particular for rural teachers, apprenticeships, learnerships, workshops, in-house training, and professional courses. Alongside these innovations are calls for varied forms of assessment strategies such as including oral and written forms of assessments and to scaffold assessments, as well as to allow for cumulative and formative assessment models. Further measures to improve education performance and learning opportunities have ranged from increased physical resources, such as providing basic infrastructure for water and electricity, classroom facilities, improved instructional materials, calls for greater parent and community involvement, and incentives for teachers and principals (Othman and Muijs, 2012).

There is also a need for tutor assistance. In particular to adult students who often have to balance conflicting demands of work, family, and studies and may need more personal attention. Critical education theorists argue that a transformation in demography and different access routes such as RPL and APEL also require curriculum and pedagogy changes, as well as reforms in institutional structures and governance. They support the call for a critical pedagogy which supports social and institutional change as a means to improve equity (Thomson and Tight, 2011).

Widening access through use of technology commonly known as e-learning is also changing the way that teaching is practiced, how research is produced, and how knowledge is shared among teachers and learners and with the public. The uses of education technologies are changing pedagogic relationships and influence new theories of teaching and learning. Some argue that e-learning shifts more responsibility onto the learner however most are in favor and conclude that open access to resources via the Internet has empowered many people who otherwise would not be able to access knowledge. E-learning has increased social aspirations to an extent whereby societal expectations need to be taken into account in what is researched and how the knowledge is shared.

Other debates on this issue are that education should have a social justice approach and that such an approach should rely on dialectical, inquiry based, and interpretive strategies, which value critical reflection. Some social justice approaches argue that education should see the social engagement of universities and schools in community and social projects. Parents and communities should participate in governance, in curriculum, and pedagogy so that marginalized communities can be more empowered and express their own agency.

There are policy makers, educationist, and communities who wish to see changes in leadership and management and argue that democratic governance requires greater accountability and transparency from the leadership, as well as democratization of knowledge. There are also countercalls for education to be linked to social entrepreneurship in particular where there is a limited possibility of employment in the formal economy and more reliance on the informal economy (Breault and Lack, 2009).

However, research has shown that most of the interventions have limited impact and the success of these reforms is uneven. Even though education has progressed for many there are persistent exclusions based on a clear intersection of gender, class, race, and other forms of disadvantage. Therefore, theorists from multiple disciplines argue that education cannot overcome deep-rooted structural inequities and more serious attention needs to be paid to build a social movement that combine critical voices of the university, school leaders and staff, students, and the community to create a vision of the education that is a dynamic environment which includes love, energy, and determination to social justice and the attainment of equity (adapted from Breault and Lack, 2009: 165).

Measurements of Equity in Education

Education is viewed as a strong determinant of career, life chances, employment opportunities, life expectancy, and can contribute to social change and national development. However, employment opportunities are affected by labor market conditions and the dictates of demand and supply. Governments and industry are calling for a stronger link between education and the labor market and demands of the economy. But there is not sufficient evidence to prove that markets can determine education needs or the other way around.

OECD's equity indicators (i.e., schooling, earning, and employment) for the OECD countries measure how the global crisis has affected employment. These indicators illustrated that schooling and an upper secondary qualification is the norm in most societies for participation in the labor market and for entry and success at tertiary level, as well as integration into society and to participate in a globalized economy. OECD figures indicate that during the financial crisis people with a higher education qualification were more likely to be employed and have higher earnings compared to less educated people.

However, figures from 2004 to 2008 show some contradictions and indicate that during the economic crisis an upper secondary qualification among 20–24-year olds was no longer a guarantee to providing employment or decreasing poverty. In 2011, youth unemployment in the age range 15–24 years in the United Kingdom was at 21.9%. In OECD countries, youth unemployment was more than double the unemployment rate in the general population (17%) with an increase in Spain of 51%, Switzerland 7%, and in SA the figure was 52.5% http://www.oced-ilbrary.org/education.

Other forms of measurements dominated human capital theory or social efficiency are concerned with returns on investment. These approaches measure the relationship between inputs and outputs, the number of graduates in terms of the investment made, whether education achieves positive outcomes in the labor market, and a social efficiency perspective which argues that public expenditure on education must yield results anything less is wasteful expenditure. Educationists who favor an equity perspective oppose this view and argue that this discourse has led to tensions between equity and

efficiency with profound social consequences as poor performing schools face closure.

Stromquist (2005) and Othman and Muijs (2012) have argued for more qualitative studies which are attentive to history and culture to measure equity and efficiency with a concern for redistribution of educational opportunity and to improve the learning experiences of pupils. In their (Othman and Muijs, 2012) study in Malaysia to measure perceived differences in rural and urban schools they focused on processes rather than products and looked at the influence of education resources, school climate, parent involvement, and leadership (Othman and Muijs, 2012).

Equity Scorecards were used in a study in Ghana and Tanzania to measure responses to the MDG which is set to halve poverty by 2015 (Morley, 2012). The scorecards measured who was admitted, who gained access to which course of study, retention, completion, and levels of achievement. The data revealed that many poor students both men and women benefitted from widening participation policies but that academic choices are gendered and classed as both poorer and women students were underrepresented in Science and Technology programs and this suggests reproduction rather than transformation of social structures (Morley, 2012: 359). Similar results from Northern countries such as Australia where indigenous people have entered higher education from the 1970s also attributed low success rates to indigenous people having to fit into culture that is alien and one that does not empower marginalized cultures.

Further measures besides completion rates which ensure that there is accountability for public funds and investment in education is measured by the effectiveness of teaching and learning in education at a national level. Other measures are the number of students who successfully complete formal education and have the required graduate attributes which are often defined as changes in student lives and understanding of the world. Qualities which make them better and more responsible citizens, or better human beings, or social responsiveness and measures of redress are also measures of equity in education. Recently, the World Bank has also reported on the effects of corruption in government on achieving equity goals.

Some critical discourses argue that in a knowledge economy policy makers need to assert social values and work against pervasive politics of individualism and institutional competition in publication and research. They argue for more internal and international collaboration between researchers based on solidarities rather than contracts or rankings. They also favor publishing on the Internet in open access networks so as to make knowledge available to a wider audience and readership. In addition that publishing open source material should not be subject to copyright but publications be made available for public education and to encourage wider debate and discussion as this can stimulate common interests, solutions to issues and problems (Thompson, 2000: 182–183).

Conclusion

The notion of equity in education is important and has moved people in the past and will continue to move people in the future (Adaptation of Sen's use of social justice 2009: 402).

Education policies, interventions, and measurements point to a range of ideological positions which implies that the discussions are not neutral. However, there is substantial agreement among different positions that poverty structures access into education, success, and future mobility. Therefore, it is argued that what is required is the redistribution of resources as education on its own cannot deliver equality or erase all the inequities and cannot deal with all the social problems. The research indicates that the achievements are uneven and that social structures are reproduced, that the influences of global capitalism impacts on countries and people differently across the world and has sometimes deepened inequities.

The task to achieve equity through education is multilayered and will require the cooperation of governments, multinationals, social movements, academics, researchers, policy makers, and world institutions to seek innovative ways to achieve good quality education for all at the different levels of education so as to achieve the MDG for 2015. The alliances should also agree to shape a world responsive to social justice and solidarity (Stromquist, 2005; Govender and Chinna Rao, 2012).

See also: Affordable Private Schools in Developing Countries; Educational Effectiveness, The Field of; Educational Philosophy; Educational Sociology; School Choice; School Effectiveness Research; Socialization and Education: Theoretical Perspectives.

Bibliography

Akoojee, S., Nkomo, M., 2011. Widening equity and retaining efficiency: considerations from the IBSA (India, Brazil and South Africa) southern coalface. International Journal of Educational Development 31, 118–125.

Breault, Rick A., Lack, B., 2009. Equity and empowerment in PDS work: a review of the literature (1999 to 2006). Equity and Excellence in Education 42 (2), 152–168.

Equality in Higher Education Statistical Report, 2009. Equality Challenge Unit. ECU 2010, London.

Featherman, D.L., Hall, M., Krislov, M., 2010. The Next 25 Years – Affirmative Action in Higher Education in the United States and South Africa. University of Michigan Press, USA.

Govender, J., Rao, Y.C., 2012. Pamphlet for Workshop on Seeking Common Ground for Change. Hosted by Congress of Swiss Sociological Association. University of Berne.

Morley, L., 2012. Researching absences and silences in higher education: data for democratization. Higher Education Research and Development 31 (3), 353–368.

Motala, S., 2001. Quality and indicators of quality in South African education: a critical appraisal. International Journal of Educational Development 21, 61–78.

OECD, 2010. PISA 2009 Results: Overcoming Social Background – Equity in Learning Opportunities and Outcomes, vol. II. http://dx.doi.org/10.1787/9789264091504-en (accessed 21.01.13.).

Othman, M., Muijs, D., 2012. Educational quality differences in a middle-income country: the urban-rural gap in Malaysian primary schools. School Effectiveness and School Improvement. An International Journal of Research, Policy and Practice. http://www.tandfonline.com/doi/abs/10.1080/09243453.2012.691425.

Reynolds, D., Chapman, C., Kelly, A., Muijs, D., Sammons, P., 2012. Educational effectiveness: the development of the discipline, the critiques, the defence, and the present debate. Effective Education 3 (2). http://www.tandfonline.com/doi/abs/10.1080/19415532.2011.6861683(2).

Rogers, A., 1992. Adults Learning for Development. Cassell Education: University of Michigan.

Rogers, A., 2005. Non-formal Education: Flexible Schooling or Participatory Education. CERC Series 15. Springer, USA.

Sen, A., 2009. The Idea of Justice. Penguin Books, England.

Scarpetta, S., Sonnet, A., Manfredi, T., 2010. Rising Youth Unemployment during the Crisis: How to Prevent Negative Long-term Consequences on a Generation?. OECD Social, Employment and Migration Working Papers, No. 106 OECD Publishing. http://dx.doi.org/10.1787/5kmh79zb2mmv-en.

Stromquist, N.P., Spring 2005. Comparative and international education: a journey towards equality and equity. Harvard Education Review 75 (1), 89–111.

Stewart, F., 1996. Globalisation and education. International Journal of Education Development 16 (4), 327–333.

Thompson, J. (Ed.), 2000. Stretching the Academy – The Politics and Practice of Widening Participation in Higher Education. NIACE, England.

Thomas, L., Tight, M., 2011. Institutional Transformation to Engage a Diverse Student Body. In: A Book Series of International Perspectives on Higher Education Research, vol. 6. Emerald Book, UK.

Relevant Websites

www.oecd-ilbrary.org/education-indicators in focus (accessed 14.11.12.) – OECD iLibrary, 2012. Education Indicators in Focus.

http://www.oced-ilbrary.org/education/how-well-are-countries-educating-young-people (accessed 14.01.13.) – OECD iLibrary Educating Young People.

www.oecd.org/els/socialp (accessed 22.01.13.) – OECD, 2011. An Overview of Growing Inequalities in OECD Countries.

www.education-inequality.org (accessed 22.01.13.) – OECD – World Inequality Database on Education.

www.unesco.education-statistics.org (accessed 21.12.12.) – UNESCO Education Statistics.

http://www.unesco.org/new/en/education/themes/leading-the-international-agenda/efareport/ – UNESCO Education for All, 2012. Monitoring Report.

http://data.worldbank.org/indicator/SLPOV.GINI (accessed 14.01.13.) – World Bank Poverty Indicators by Geni-coeffecient.

www.wacra.org/world%20Poverty%20MAPPDF (accessed 14.01.13.) – World Bank-Human Development Index.

Equity in Law

André-Jean Arnaud, Droit et Société, FMSH, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Equity is usually defined as fairness, impartiality, justness. Consequently, equity in law is said to exist when justice administered between parties appears as based upon natural reason or ethical judgment – whether or not it is in conformity with written (state) law. Indeed, equity can be considered either as a concept immanent to the law or as a set of principles outside the law, the function of which is to allow the judge to give soft interpretations in concrete cases in order to reach not too unmerciful solutions. Hence, equity is working at large in legal systems where the judge plays an important role, like common law or other systems where state law has a reduced role, while the official use of it is very insignificant in other legal systems, such as positivist continental laws. Equity in law can be examined under two main headings. Equity in law has always been considered (1) *ethically* as an opportunity to refer to equity against positive (state) law, and (2) *politically* as the way to open law to necessities borne from the economic developments.

Equity as a Common Expedient

In this sense, equity is positioned between law and ethics. In the Western legal tradition, equity was even culturally perceived as the only foundation of law. Equity and justice go hand in hand in Western legal philosophy. But the Roman root *aequus* was translated by two words that acquired two very distinct meanings through the ages: *equity*, on the one hand, and *equality*, on the other. The distinction led to opposing political theories on law and justice.

The Roman law defined legal work as skillfulness in settling what is good and fair : *Ars boni et aequi* (Celsius and Ulp, *Dig*., I, 1, 1, pr.). Among traditional legal principles, we find also the rule according to which equity must be expected in everything but above all in law (Paul, *Dig*., 50, 17, 90). As Pretorian law, a great part of everyday Roman legal regulation was progressively built upon equity in order to complete positive law. The Roman distinction between *ius honorarium* and *ius civile* has been seen by some authors as a forerunner of what would later exist in England as a distinction between common law and equity: a "body of rules existing alongside the original civil law, founded on distinct principles, and claiming incidentally to supersede civil law in virtue of a superior sanctity inherent in these principles" (Maine, 1861: p. 25).

Equity is used to replace legal rules drawn from statute or common law when these rules are considered too inadequate or unyielding for the expectation of justice in a specific case. Hence, equity was always considered as a common expedient against legal (state) justice where this latter appeared too distant from ethics, morals, or customs. This route to justice was historically employed to such a degree that the gambit was embodied in many legal traditions, re-emerging from time to time according to need.

In opposition to legal unfairness, equity is regarded as either a refuge (equity prompts legal rules to law givers) or a recourse (equity is the alternative source of a redress or remedy to those provided by established law).

From ancient times, equity was considered to be an efficient means to establish justice in law. Without equity, law would be unjust ('*Summum ius, summa injuria*'). Within legal anthropology, Henri Sumner Maine showed in his work *Ancient Law* that equity is considered an instrument to bring harmony to progressive societies. According to this author, equity is one of the three agencies by which primitive law has been adapted to the requirements of a developing society and an improving morality.

Equity was also considered in Western legal tradition as a complement or a corrective to positive law. There are a number of aphorisms on this topic, such as *Hæc æquitas suggerit, etsi deficiamur jure* or *Æquitas nihil aliud est quam jus quam lex scripto prætermisit*; that is, equity must be followed where there is no legal norm to reach a solution.

But, when the man in the street talks of 'equity,' he is thinking of ideal justice that may even be contrary to the law. Here lies that faith in the fairness of a claim to some natural law existing above positive laws or statutes. Equity focuses on the imperfections of law and can even be preferred to a legal norm where it provides a useful corrective to legal institutional justice by improving an equitable outcome: *Placuit in omnibus rebus præcipuam esse justitiæ æquitatisque quam stricti juris rationem*. Equity is better than a solution drawn from a strict interpretation of law. Such a theory, elaborated from Aristotle onward, runs through the Western legal tradition through Roman law, subsequent Canon Law (Grossi, 1998) and right through continental legal regimes before the French Revolution.

That law can be brought into harmony with society through equity is yet currently demonstrated by legal sociologists, whose research illustrates specific circumstances in which societies make appeals to alternative procedures where equity plays an important role. This is particularly clear when it is not ingrained within a culture to instinctively call on the law for help as the main way to resolve conflicts, such as in traditional Chinese or Japanese cultures. Moreover, equity is often grasped as an opportunity for a workable alternative to state law, where it is not envisaged by a written legal norm. In the last decades of the twentieth century, for instance, equity was cited in alternative movements by both Italian and Brazilian judges (*Alternative use of Law* and *Alternative Justice*, in Arnaud, 1993).

Equity as a Means of Adapting Legal Regulation to Political Prerequisite for the Development of an Economy

All through the building of the occidental legal tradition, while laws were considered as means to protect communities and ensure their reproduction following a specific pattern, equity also appeared as the way to open up law to necessities born from the development of trade. Here, equity appears (1) as an earlier technique to create a society for merchants, and (2) in the last decades of the twentieth century, as a means to implement a global legal order.

Equity as a Foundation of Trading Societies

At the end of the Middle Ages, the opposing pressures of the wish of kings to concentrate absolute powers in their hands, the opening up of everyday life to trade and merchants, as well as the emergence of a *bourgeoisie*, led the various national laws to adapt themselves to the development of landed and propertied interests. Western laws gradually expanded into opposing systems: the common law and equity tradition on the one hand and continental legal positivism on the other.

After centuries of legal rationalization, equity was definitively taken up from the discourse of ethics and located as one of the foundations of law, appealing to natural law and reason, even as against positive law and statutes (Hugo Grotius, Samuel Pufendorf, Christian Wolff, and their disciples in Northern Europe; Jean Domat in France).

As to the continental Romano-canonic tradition, equity was quickly made subservient to state law. Despite the influence of Domat's work on French jurisprudence, equity was forbidden in France by the *Ordonnance civile* of 1667. The French Revolution of 1789 was partially directed against the discretionary power of judges. A project to reintroduce equity as a tool in judicial interpretation in the French Civil Code was drastically cut down by Bonaparte. Finally, the judge was considered merely to be the mouthpiece of state law. It is true that we can find counterexamples in nineteenth-century France, such as judicial decisions by the *Bon Juge Magnaud* who used equity to avoid enforcing the law against the poor. But this use of equity was a rarity, and almost all such decisions were reversed by the court of appeal.

We can also identify some slight remains of equity in civil law, even in the French Civil Code (art. 565, 1135; 1854). The strongest examples are, however, drawn from the Swiss Civil Code, articles 1 and 4: "The law must be applied in all cases which come within the letter or the spirit of any of its provisions. Where no provision applies, the judge shall decide according to the existing customary law, and, failing which, according to the rules which he would lay down if he had himself to act as legislator. Herein he must be guided by approved legal doctrine and case-law – where the law expressly leaves a point to the discretion of the judge, or directs him to take circumstances into consideration, or to consider whether a ground alleged is material, he must base his decision on principles of justice and equity" (see Lloyd and Freeman, 1985: p. 1231).

At the same time, in the common law countries, legal developments went in the opposite direction. There, indeed, equity refers to that part of the law of England that is derived not from the custom of the realm nor the enactments of Parliament but from the decisions of the old Court of Chancery, consolidated in 1873, along with the other superior courts in the supreme court of judicature. Such equity pertains to a right recognized by a court of equity through a specific procedure. Here, we are faced with procedural equity and not 'natural justice.' Jurisdiction in equity is one of the three clearly marked divisions into which the jurisdiction of the court falls, together with a miscellaneous jurisdiction and a common law jurisdiction. The equitable jurisdiction of the court originated in the power that the chancellor assumed in the Middle Ages to modify a strict and formal legal system on moral grounds.

There are a number of principles used as guidelines when it is necessary to make appeal to this notion of equity: equity acts *in personam*; equity acts on the conscience; equity will not suffer a wrong to be without a remedy; equity follows the law; equity looks to the intent rather than the form; equity imputes an intent to fulfill an obligation; equitable remedies are discretionary; delay defeats equities; he who comes into equity, must come with clean hands; he who seeks equity must do equity; equity takes into account the balance of convenience; where there are equal equities, the law prevails and the first in time prevails; equity, like nature, does nothing in vain; equity never wants a trustee; equity aids the vigilant; equality is equity; equity will not assist a volunteer; equity will not permit a statute to be a cloak for fraud.

In England, because the remedies that the common law courts could give were early limited by the forms of writs available, men turned to the chancellor for equity. The two systems came to be administered by wholly independent tribunals whose decisions were frequently in conflict. The early jurisdiction of the Court of Chancery was in its nature extremely vague. From the beginning of the seventeenth century, it is possible to discern certain general principles in the exercise of the court's jurisdiction. The early seventeenth century was the principal turning point in the history of the court, for the common lawyers were seeking to prohibit the Chancery from interfering with common law rights. King James I resolved the conflict between the two jurisdictions in favor of the Chancery. From this time onward, although the conflict did not die down until the end of the century, the court gradually developed jurisdiction coordinate with that of the common law courts and on settled principles. In the eighteenth and nineteenth centuries, it came to follow precedent as strictly as they did.

The principal distinction between the two systems was that, whereas the rules of common law were supposed to have been established since time immemorial, those of equity could in almost every case be traced to the chancellor who invented them. In the nineteenth century the jurisdiction of Chancery, now clearly defined, was frequently classified according to its relationship to the common law as being exclusive, concurrent or auxiliary. By the middle of the twentieth century, the Chancery division consisted of the Lord Chancellor and judges divided into groups that were respectively assigned matters within the peculiar jurisdiction of the division.

In the United States, equity was developed with relative consistency during the nineteenth century. The expansion of the equity system was subject to statutory grants bestowed to

each state, which constitutes a specific feature of American legal history. During the Civil War, the equitable jurisdiction was generally attributed to courts of double jurisdiction. Since 1848, many states excluded the distinction between law and equity in favor of one 'civil action.' In the federal courts there was a complete fusion of law and equity by the adoption, in 1938, of the federal rules of civil procedure, promulgated by the US Supreme Court under authority of an act of Congress.

Equity as a Referent in the Framework of Contemporary Political Theory

Recently, the use of equity was proposed by neoliberal legal philosophers as a way to put into common practice some fundamental legal principles. Hence, equity has become a referent in the framework of globalization, re-emerging in many new areas for legal work, for example, private, regional, and international judicial assemblies, and directly in relation to global governance.

The importance of the recourse to equity in contemporary neoliberal legal and political philosophy is due to (1) the predominance of the Anglo-American legal tradition in international legal relations, (2) the special focus given to the above-mentioned principle according to which 'equality is equity,' and finally (3) the claim for a minimal state.

After much controversy surrounding the Aristotelian idea of 'corrective justice,' equity was reintroduced as the legal principle of equal treatment of people in the allocation of legal benefits and obligations. John Rawls designates as '*fair equality of opportunity*' the equity of opportunities that cannot be reduced to a merely hypothetical right (Rawls, 1971; para.14 and 46). Here arises a paradox: the more the state intervenes in establishing equality of opportunities, the more it stands apart from the ideal of contractualism that characterizes the minimal state. Consequently, there could be no coherent basis for imposing common global lawful and unique interpretations (for example, of human rights). This means that it would not be possible to universally expect the same behavior through all institutions, whether national, local, or international.

Popper's claim for the importance of tradition, a tradition that offers the moral structure of a society, is grounded in the idea that herein lies the social meaning of justice and equity. On the one hand, tradition reveals the degree of moral feeling that each society attains; on the other hand, should there be conflict between opposite interests, tradition is the moral framework that allows us to find fair or equitable compromises (Popper, 1963: Chap. 17, para. 3). Institutions alone give a society its legal structure appropriate to this moral background.

According to Friedrich Hayek, a free society – what he calls a 'great society,' similar to Popper's 'open society' – is a spontaneous order formed by individuals each pursuing their own interest. There is no common hierarchy of values; instead, a legal system that protects each individual's ability to pursue his/her own values provides the rules of behavior that discipline society and produce peace. Chief elements of such a legal system are the right to own property and the right to exchange one's entitlements on any mutually agreeable terms; the particular spontaneous order that arises when individuals exchange their entitlements is the 'market.' In Hayek's view, the market plays an invaluable role in harmonizing people's disparate plans: changes in the prices of goods give signals to entrepreneurs to produce more or less of different products, even though these entrepreneurs may be entirely ignorant of the causes of these price changes. Free, unplanned societies are thus characterized by greater order and less waste of resources than can ever exist in a planned society, which relies on a degree of explicit knowledge among the planners that can never be attained.

In the game of the market, each participant has equal formal rights. The results are unequal, but this is not unfair, Hayek explains; indeed, the possibility of such inequality (for example, between entrepreneurs who guess consumer demand correctly or incorrectly, or between risk takers whose risks do or do not materialize) is what allows the market to harmonize individuals' plans and to use resources rationally. Success or failure must depend only upon each person's action and cannot flow from the knowledge and power held by public authorities.

For Hayek, the 'great society' is not merely a national order; the benefits of the impersonal market order apply all the more as the great society is extended to the international level through globalization. Thus, inequality of results is also a characteristic of the global order, but equity is realized provided that everybody is treated equally according to these abstract rules of fair behavior.

States from time to time wish to introduce social justice through legal regulation. But equality cannot be implemented beyond equal treatment before the law. To the extent concepts of social justice seek to accomplish anything more, Hayek writes (Hayek, 1973), they require external planning, destroy the spontaneous feature of the market, prevent the great society from developing, and are thus incompatible with liberty. Moreover, redistribution between groups in the name of social justice produces conflict between groups. In other words, equality of result – even in the name of charity – or intervening to equalize opportunity, rather than equal treatment by law, leads to arbitrary outcomes. Concrete human circumstances may evoke compassion and lead to requests to modify the abstract rules. But, Hayek argues, humankind requires an impersonal global order to thrive, and meddling with the order to produce what he calls 'the mirage of social justice' reveals a lack of intellectual and moral maturity. Authoritarian intervention generates disorder and cannot be equitable. To speak of concrete common goals is specific to small closed societies in which people were protected by community at the price of their freedom. In contrast, a great society of free people can exist only if everybody submits to 'equal' abstract rules.

In this context, it is easy to understand that the state must play a minimal role. The 'entitlement theory' (Nozick, 1974) based upon a discussion of Hart's 'principle of fairness' (Hart, 1961) challenges the claims for state action aimed at implementing equality among people.

Where the state is striking space, the judge becomes a masterpiece in the framework of the legal structure. The judge enjoys extended powers in such a society. He has to correct permanently the disturbances within an order that is spontaneous, not artificial, and that tells people what they have to do.

Such a judge must decide in equity in the sense that he has to decide if people's behavior is appropriate to what can normally be expected in comparison with the common conduct of the members of the same group.

The law of the judges is a translation of what people have learned from their actions. Judges' function is to maintain a permanent and efficient order in people's behavior. The judge contributes to improving a given order, an order that must not be upset by changes within relationships between citizens.

Such contemporary theoretical views on equity are particularly significant. Globalization is promoting neoliberal views on law, justice, and equity. Indeed, equity has already become a common referent for regional and international courts of justice. This leads in turn to many ambiguities, for instance in the implementation of the European Union's *Equal Treatment Directive* against sex discrimination. This last rule is based on a general feeling of equity but supposes some affirmative action through community (state) intervention (see the *European Law Journal*, 1998; Posner, 1993: pp. 404–405). The same ambiguity can be found in the report by the Commission on Global Governance entitled *Our Global Neighborhood* (Commission on Global Governance, 1995). In this report, written at the request of the United Nations, we read that "Justice and equity are essential human values. Respect for them is indispensable for peace and progress… Where large number of citizens are treated unfairly or denied their due… discontent is inevitable and conflict likely." However, 'a concern for equity is not tantamount to an insistence on equality.' Therefore, equity 'does call for deliberate efforts to reduce gross inequalities, to deal with factors that cause or perpetuate them, and to promote a fairer sharing of resources' (Commission on Global Governance, 1995: p. 51). This is good as political principle, but what practical legal rule can flow from such abstract, inconsistent, and equivocal discourse?

Nevertheless, and despite the lack of general agreement on its relationship with equality, equity appears as a newly contemporary and significant notion that promises great possibilities for redesigning law making, the administration of justice, and the process of judicial decision making.

Recent Uses of Equity: A Contrast between Legal Tools and Social Claims

When speaking of equity today, one of the main problems is the confrontation between theories, ideas, wishes, and utopia about what must be equity, on the one hand, and the reality of trade, economic, and financial aspirations, on the other hand. In other words, modern usage of this term includes a conception of fairness and equality of treatment for people in disparate circumstances. Toward that end, one can identify claims about *social equity* (that there should be fairness in the elaboration and implementation of public policy, as well as in the equitable distribution of public services), *gender equity* (that, regardless of gender, people should be treated equally as human beings or as rights-bearing citizens), *race equity* (that, regardless of race, people should be treated equally as human beings or as rights-bearing citizens), and *pay equity* (that employees should receive comparably equal pay for comparably equal work).

The problems tied to gender, race, and cultures pertain to social equity, but more especially to *cultural equity*. This expression encompasses family background, child welfare, education, role models and mentors, experiences of gender and race-related issues, choice of field and career, professional and personal life goals. Teaching diversity and educating citizens in a multicultural society is becoming a global obsession, concretized down through the implementation of more and more national legal norms of affirmative action. Many colleges and universities consider the term *social equity* as synonymous with *social equality*. However, some argue that this is mere pretention to equality, and that it necessarily creates other forms of inequality (Messerli, 2012).

The recent uses of equity have complicated the meaning of this term in the legal and social universe. While some uses may appear to have appropriated the term, the notion of equity, through the development of claims to social equity, seems to be increasingly implicated both in social claims and in calls to redesign public policies and law making.

See also: Civil Law; Common Law; Equality and Inequality: Legal Aspects; Equality: Philosophical Aspects; Globalization: Legal Aspects; Justice and Law; Justice: Philosophical Aspects; Legal Systems, Classification of; Lex Mercatoria; Natural Law; von Hayek, Friedrich A. (1899–1992).

Bibliography

Aequitas et Bona fides, 1955. Festgabe Simonius. Helbing & Lichtenhahn, Basle, Switzerland.

Équité. In: Arnaud, A.-J. (Ed.), Dictionnaire encyclopédique de Théorie et de Sociologie du droit, second ed. LGDJ, Paris.

Arnaud, A.-J., 1998. De l'Équité des marchands à l'équité du marché. In: Entre modernité et mondialisation. Cinq leçons d'histoire de la philosophie du droit et de l'État. LGDJ, Paris.

Arnaud, A.-J. (Ed.), 2010. Dictionnaire de la globalisation – droit, science politique, sciences sociales. LGDJ-Lextenso Ed, Paris.

Bullard, R.D., Johnson, G.S., Torres, A.O., 2011. Environmental Health and Racial Equity: Building Environmentally Just, Sustainable, and Livable Communities. Westview Press, Boulder, CO.

Bürgi, W., 1939. Ursprung und Bedeutung der Begriffe "Treu und Glauben" und "Billigkeit" im schweizerischen Zivitrecht. Stämpfli, Berne, Switzerland.

Castán Tobeñas, J., 1954. La Formulación judicial del derecho – Jurisprudencia y arbitrio de equitad, second ed. Instituto Editorial Reus, Madrid.

Commission on Global Governance, 1995. The Global Neighbourhood. The Report of the Commission on Global Governance. Oxford University Press, Oxford, UK.

De Page, H., 1931. A propos du gouvernement des juges. L'équité en face du droit. EMPEA, Sirey-Bruylant, Paris. www.empea.net/.

Einolf, K.W., 2007. Building a socially responsible equity portfolio using data envelopment analysis. Philosophica 80, 71–103.

Esser, J., 1963. Wandlungen von Billigkeit und Billigkeitsrechtsprechung im modernen Privatrecht. In: Summum ius summum injuria. Mitgliedern der Tübingen Juristenfakultät. Mohr, Tübingen, Germany, pp. 22–40.

European Law Journal, 1998. Special Issue 4/2 on Sex Equality.

Frederickson, H.G., 1990. Administration and social equity. Public Administration Review 50 (2), 228–237.

Grossi, P., 1998. "Aequitas canonica," Quaderni Fiorentini per la storia del pensiero giuridico moderno, vol. 27. Giuffrè, Milan.

Hart, H.L.A., 1961. The Concept of Law. Clarendon Press, Oxford, UK.

Hayek, F.A., 1973. Law, Legislation and Liberty. Routledge & Kegan Paul, London.

Lloyd of Hampstead, Lord, Freeman, M.D.A., 1985. Lloyd's Introduction to Jurisprudence, fifth ed. Stevens & Sons, London.

Maine, H.J.S., 1861. Ancient Law: Its Connection with the Early; History of Society and Its Relation to Modern Ideas. J. Murray, London.

Manaï, D., 1985. Le juge entre la loi et l'équité. Essai sur le pouvoir d'appréciation du juge en droit suisse. Payot, Lausanne, Suisse.

Marini, C.M., 1959. Il Giudizio di equità nel processo civile. Premesse teoriche, Padova, Italy.

Messerli, J., 2012. Should affirmative action policies, which give preferential treatment based on minority status, be eliminated? http://www.balancedpolitics.org/affirmative_action.htm (updated 01.07.12.).

Metzger, J.T., 1996. The theory and practice of equity planning: an annotated bibliography. Journal of Planning Literature 11, 112–126.

Mommsen, T. (Ed.), 1954. Corpus Juris Civilis, Digestum, vol. 1. Weidemann, Berlin.

Newman, R.A. (Ed.), 1973. Equity in the World's Legal Systems: A Comparative Study. Bruylant, Brussels, Belgium.

Nozick, R., 1974. Anarchy, State and Utopia. Basic Books, New York.

Popper, K.R., 1963. Conjectures and Refutations. Routledge and Kegan Paul, London.

Posner, R.A., 1990/1993. The Problems of Jurisprudence, first ed. Harvard University Press, Cambridge, MA.

Rabello, A.M. (Ed.), 1997. Aequitas and Equity: Equity in Civil Law and Mixed Jurisdictions. Harry and Michael Sacher Institute for Legislative Research and Comparative Law, The Hebrew University of Jerusalem, Israel. Race Equity Project, www.ejrc.cau.edu/raceequitysmartgrowth.htm and also http://equity.lsnc.net/.

Rawls, J., 1971. A Theory of Justice. Belknap Press of the Harvard University Press, Cambridge, MA.

Rayman, P., 2001. Beyond the Bottom Line: The Search for Dignity at Work. Palgrave, Oxford.

Ridder, H., 1950. Aequitas und equity. Archiv für Rechts- und Sozialphilosophie 39, 181–200.

Treuhaft, S., Glover Blackwell, A., Pastor, M., 2011. California Tomorrow: Equity is the Superior Growth Model. http://www.policylink.org.

Zaki, M.S., 1990. Équité. Archives de Philosophie du Droit, 87–118.

Equivalence and Transfer Problems in Cross-Cultural Research

Amina Abubakar, Tilburg University, Tilburg, The Netherlands; and Utrecht University, Utrecht, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Problems in transferring tests from one context to another arise from construct-, method-, or item- level bias. Using examples from research in different parts of the world, the author highlights and discusses specific problems in transferring tests across cultures. A systematic approach to making adaptations to tests and procedures to ensure their suitability in a new cultural context is presented. The approach has a strong focus on community involvement and the use of mixed method approaches during the test adaptation process.

Introduction

There are a large number of psychological tests and scales that are adequately standardized and normed for use by developmental psychologists; most of these are commercially available and widely distributed. However, a significant number of these tests have been developed and standardized for use in a very limited demographic and geographical setting i.e., largely developed and standardized among white middle class North American and European children and families. Consequently, most of these tests may not be culturally appropriate for use by other groups that were not included in the development and standardization procedures, unless their appropriateness has been shown. When developing a test, most test developers will, in their choices of items, tasks, stimuli, and administration procedures, be informed by what is happening in their immediate social, economic, and cultural contexts (Gopaul-McNicol and Armour-Thomas, 2002). Taking this scenario into consideration, the implication is that the cultural context of a significant proportion of the world's population did not contribute to the conceptualization of both the test contents and the administration procedures. Most researchers and clinicians in emerging and developing parts of the world find themselves lacking the necessary tools and measures to assess child development and outcomes.

Cultural influences may lead to several forms of bias in cross-cultural assessments that do not carefully consider the impact of test transference into a new culture (Van de Vijver and Tanzer, 2004). Three forms of bias can be distinguished: construct, method, and item bias (Van de Vijver and Tanzer, 2004). The first one is *construct bias*, which occurs when an instrument only partially samples the domains that define a construct (Van de Vijver and Hambleton, 1996; Van de Vijver and Tanzer, 2004). For instance, as the development of the modern intelligence test was based purely on Western definitions of scholastic achievements, it has a strong emphasis on aspects such as reasoning, memory, and acquired knowledge. However, in non-Western countries such as those in Africa, there is evidence that the conceptualization of intelligence has a very strong social aspect (Mpofu, 2001; Grigorenko et al., 2001). Consequently, the use of non-Western measures may not adequately measure intelligence in African contexts.

The second bias consists of *method bias*, which refers to problems that arise due to administration procedures, instrument and sample characteristics. Most assessment measures have laid down procedures and conditions for administration. It is usually assumed that all children understand the norms within the assessment context and would respond accordingly (Gopaul-McNicol and Armour-Thomas, 2002). However, it has been observed that in several instances, the test administration procedures may be culturally inappropriate (Foxcroft, 2002). Included in this are the examples of unfamiliarity with involving themselves in a one-to-one dyadic interaction with adults, unfamiliarity with directly responding to questions asked by an adult, and/or balancing between speed and accuracy. For example, in rural Kenya, children spend a significant amount of time with siblings; requesting such children to fully engage in a developmental assessment task in a one-to-one assessment with an unfamiliar adult led to significant problems in task engagement (Abubakar et al., 2007). However, introducing the test in the presence of a sibling and allowing the children to settle in gradually, enhanced their test engagement. Similarly, in a Singaporean study, it was noted that most children were too shy to express themselves in the presence of an adult; consequently, it was difficult to evaluate their expressive language.

Instrument characteristics such as stimulus and task familiarity have been observed to influence test-taking behavior and performance across different cultural contexts. Malda et al. (2010b), in a study in South Africa comparing children of Setswana and Afrikaans origin, observed that children from both cultural groups performed significantly better in tasks that involved items, vocabulary, and experiences drawn directly from their cultural settings. In an assessment of children from different cultural groups in South East Asia (Hmong, Laotians, Vietnamese, and Cambodians), several aspects of the Denver Developmental Test had to be modified due to their unfamiliarity with the test takers (Miller et al., 1984). For instance, language items that had plurals had to be excluded because none of the languages used by the children taking part in the study used plural forms. Additionally, they observed that while children of Cambodian origin could easily engage in 'pat-a-cake' play as part of the assessment, the children from the other cultural groups were unfamiliar with this game, hence could not participate in it.

The last form of bias is *item bias*. This may arise due to differential functioning of items in different contexts (Van de Vijver and Tanzer, 2004). For example, in a study in West

Africa, digit span was found to have significant bias against the Wolof-speaking children, because among the Wolof people, digits are counted to the base of five, while among the Mandika people, digits are counted to the base of 10. In this case, the Wolof-speaking children were disadvantaged because they had to remember much longer digits after the initial five digits (Jukes and Grigorenko, 2010). Moreover, in developmental psychology, item bias may arise due to differences in both speed and patterns of acquisition of developmental milestones. For instance, Vierhaus et al. (2011) observed that in the assessment of developmental outcomes among Cameroonian children using the Bayley Scales of Infant Development (BSID) scores, the performance of the Cameroonian children diverted from the expected patterns. While unable to perform some of the 'simple' motor tasks, they were nevertheless able to perform 'more advanced' tasks. Consequently, strictly following the discontinuation rule of the BSID may potentially underestimate the performance levels of Cameroonian children.

Dealing with Problems of Transference: Adaptation and Assembly

There are two options for dealing with problems of transference across cultural contexts: *adaptation* or *assembly* (Malda and Van de Vijver, 2005). Adaptation involves a systematic evaluation of all aspects of a measurement instrument, and modifying it where needed in order to make it more suitable to the context. Assembly involves the development of a new measurement instrument. These approaches are expensive and time-consuming. However, the availability of an instrument that has been developed and rigorously evaluated on-site significantly enhances the validity of the data collected (Van de Vijver, 2002). Whether one adapts or assembles a new tool will largely be informed by salient aspects of the study design. For instance:

1. Target sample: As previously discussed, sample characteristics largely influence test-taking. For instance, working in a context where there are no standardized tests, the needs of the rural and urban populations may vary. With globalization and exposure to westernized education, urban children may be more able to cope with tests from other contexts. However, this cannot be assumed; formal evaluations of the suitability of the measures need to be carried out before being used.
2. Target construct: Some constructs have been observed to have much easier transferability between contexts as compared with others. In our work in rural Kenya, we have observed that measures of motor development are relatively easy to adapt to the rural African context. In adapting motor development measures, our main challenge was ensuring that test material and administration procedures were appropriate to this context (Abubakar, 2008). However, concepts such as parenting have been observed to be much more 'culturally influenced,' making it more difficult to come up with adequate procedures (Abubakar, 2008).

Test adaptations have to address all sources of bias previously discussed, to ensure that the adapted test is appropriate for the context. Malda et al. (2008) have proposed a framework of the types of adaptations that need to be carried out in order to be able to present the appropriate results. The slightly modified version of the framework on areas that need adaptation is presented in Table 1.

Moreover, whether one has decided to adapt or assemble the literature, cross-cultural psychology advocates the use of systematic approaches in carrying out this exercise. Various guidelines exist on how to carry out adaptations. In the next section, the author presents a four-step test adaptation approach that is based on field work done in rural Kenya, and cross-cultural work done elsewhere (Abubakar, 2008; Holding et al., 2008). This approach involves four phases, namely, construct definition, item development, scale development, and test evaluation. In the next section, each of these steps will be discussed.

Construct Clarification: A Mixed Method Procedure

In this first stage, the aim is to clearly define the construct to be measured, as well as to identify behavior indicators of the construct of interest. During this stage, one seeks to understand the full behavior repertoire that needs to be assessed, adequate approaches for assessing the behavior, and the best way to access this information. The use of a mixed method approach allows for the elicitation of both *etic* and *emic* aspects of the construct to be studied.

Etic aspects refer to aspects of the construct that can be accessed from a context of universalism, i.e., aspects of the content, items, stimulus, and administration procedures that do not require changes or modifications. To attend to the etic aspects of the study, one is required to carry out a systematic

Table 1 A summary of the types of adaptations that may need to be carried out

Type of adaptation	*Explanation of the adaptation*
Construct-driven	These are adaptations aimed at including items or modifying items to ensure that the construct of interest in the target population is fully captured.
Theory-driven	These are adaptations carried out to take into consideration the theoretical underpinning of the tests. For example, ensuring that all digits in the digit span test are adequate.
Linguistic-driven	In this category, adaptations are carried out to accommodate differences in the language structure of the original test.
Culture-driven	These adaptations are made to take into account the norms and standards that are culturally defined.
Sample-driven	These are adaptations made to ensure the tests are suitable for the sample being studied. Included here are modifications to administration procedures to cater for literary levels and modifications to the test materials, to ensure that the study sample is familiar with the stimulus being presented.

review of existing literature in order to provide a critical evaluation of existing measures of the construct of interest, and help identify a suitable instrument and items to be used. The review process could result in various kinds of choices for the test developer. First, one could identify a single instrument whose characteristics make it the best possible choice; this test would then be selected for adaptation. For instance, in looking for a measure of intelligence for use in India, Malda et al. (2010), used the Kaufman Assessment Battery for Children, version II, 2004 (KABC-II) as the basis for their work. They then set out to make adaptations to ensure that the content and the stimulus were appropriate for the context. A second scenario is that a variety of tests are used, and items are 'borrowed' from different scales to form the basis of the new scale. In this case, one assembles a new measure based on existing scales. A third scenario is where one reviews the literature, does not use any existing items, but uses the existing theoretical information to design a tool.

Emic aspects refer to culture-specific aspects of the construct being studied, or culture-specific ways in which a construct understudy needs to be evaluated. The emic approach would involve quantitative data collection techniques such as focus group discussion, observation, interviews, and item rating with the members of local communities. In carrying out emic aspects of test development, two groups of the local community form an interesting target group. First, the local experts (by this the author means members of the community who are literate and possess expertise in the relevant areas): for instance, for a developmental psychology test development process, experts such as teachers, nurses, and local psychologists may be a valuable source of information. They provide the middle ground on information between what is in the literature and the community-level information. Second, the target community needs to be involved in this process; they provide the culture-specific aspects of the study. Moreover, by involving them, one is able to access information that tells them how best to access the data they need. Observing the natural setting in which people operate on a day-to-day basis allows for the formulation of the necessary procedures and skills. Methods for eliciting community understanding include Focus Group Discussions (FGDs), individual in-depth interviews, and direct observation. The role of the target community is diverse and may include definition of the construct in the local community, providing face validity of the item/tasks, and evaluating the cultural appropriateness of the items/tasks. For instance, in a study in Kilifi, consulting with the local population contributed to enhanced item wording of a translated version of the 10 Questions Questionnaire. In this study, though, they had used the forward- and back-translation approach for translating the measure; in the pilot stage, unacceptably high rates of 'impairments' in three specific areas were reported. In a series of focus group discussions with the mothers, it was indicated that the item wording in the local language had ambiguity. For example, the question on hearing impairments could mean the child either had a hearing problem or that the child is inattentive. In another example, use of local consultations has been dependent on access to local idioms and labels e.g., there is no Shona word for depression, but the term '*kufungisisa*' which means 'thinking too much' overlaps with depression (Patel et al., 1997).

Item Development: Translation and Formulation

The aim of this step is to prepare a list of potentially acceptable items, in clear and unambiguous language, through the integration of the information collected in Step 1. During this stage, items identified from the literature need to be translated adequately, while ideas identified from the community need to be adequately formulated into appropriate items so that they can be used in a scale. In the next stage, we look at issues related to translations and adaptations.

Item Translations

Various approaches to test translations have been proposed. One of the most popular approaches is the translation, back-translation approach. However, for several decades this approach has been criticized for being inadequate in providing an adequate translation of the original item. Brislin noted that the translation, back-translation approach has various problems. For example, the quality of the translation can be compromised when one has poor translators who stick to the grammar of the original language. This approach will make a back-translation easy; however, the scale will be inappropriate for the target. Furthermore, a back-translator may make sense of a poor translation and produce an adequate back-translation, thus covering up for the translation errors. We have previously used a process that combines the various translation processes to maximize the quality of the translated items. The steps in the evaluative process can be summarized as (1) Production of at least two forward-translations by two independent native speakers; the target of this stage is conceptual translations as opposed to literal translations. In ensuring a good translation, one has to select the most appropriate translators. Some of the important characteristics of good translators include being bilingual, being familiar in the cultural context, as well as being familiar with the construct one would like to assess. (2) Next, a back-translation of each of these translations is carried out. (3) A harmonization step, in which a panel of experts is formed to discuss and agree on the different versions. The panel of experts usually includes those involved in forward- and back-translations alongside the other experts; important here is the participation of a linguistic expert. During the harmonization process, the translations are examined in the light of the original text, and disagreements discussed until a consensus is reached. An adequate translation produces items that are conceptually equivalent, easy to read, natural sounding in the target language, and easily administered to the target population.

Formulation of the Appropriate Items

Item wording needs careful consideration. The simpler an item, the more likely it is to function well, especially within populations which have low literacy or who are unfamiliar with test-taking. Various studies in developmental and personality psychology indicate that items wording may impact on the validity and factorial structure of the measures. This may arise if the item is too complex (thus raising ambiguity), or if the items use negation (leading to careless scoring or results in response style in scoring e.g., extremity).

Scale Development: Refining the Scale through Pretesting and Piloting

At the scale development stage, activities include the full development of the scale, the development of scoring, and administration procedures. Training of those who will administer the tests ensures that the developed procedures are adequately implemented. Scale development involves a decision on how items are presented to the participants. This is done to take into consideration the characteristics of the sample and the local contexts. For instance, in the adaptation of measures of postpartum depression such as the Edinburgh Postnatal Depression Scale (Cox et al., 1987) for use in resource constrained settings, it was observed that mothers who were illiterate found it difficult to comprehend the measures presented in the four-statement formula. To deal with this problem, the statements were reformulated, and presented one after the other. The modification of the administration procedures led to enhanced comprehension, and validity of the adapted scale. Moreover, in developing scales one has to take into consideration the ability level of those who will administer the scales.

Prepilot and Cognitive Interviews

Having developed the scale, one has to evaluate test contents, administration, and scoring procedures in order to ensure their adequacy. At this point, test materials are modified as required. Various steps can be taken to ensure that the content is of good quality. One such step is cognitive interviews during the prepilot stage. These involve holding in-depth interviews with participants to elicit information related to how the participants understand the items, process the information, and respond to it. Some of the information one would attempt to obtain from this process is to what extent participants understand the questions in the way they were intended to be understood, and whether the items contain phrases or expressions which are either difficult, rarely used, or socially inappropriate. Cognitive interviews can be carried out either concurrently while the respondent goes through a questionnaire, or retrospectively as part of a debriefing discussion.

Test Evaluation: Psychometric and Nonpsychometric Approaches

In evaluating the quality of the adapted tests, there are standard psychometric and nonpsychometric approaches to be used.

Psychometric Evaluation

Following adaptation, an evaluation of the quality of the developed test is very important. The extant literature has sufficient guidelines on the psychometric approaches to be considered (Hambleton et al., 2005). The evaluation of validity and reliability of the tests is very important. Additionally, test norms for the newly developed or adapted tests need to be considered. If the scores of the developed tests are to be used for group comparisons, ensuring equivalence of the developed test scores is an important aspect of the developed procedures.

Nonpsychometric Approaches

Nonpsychometric information can be collected from participants, test administrators, or simply during the assessment session (Hambleton et al., 2005).

Observational Notes during the Tests

Data collected during the test session have the potential to provide significant information on the suitability of the adapted testing. Some of the relevant information here includes time taken to administer the test, and the reaction of the participants to test procedures, instructions and methods.

Feedback from the Test Administrators

Test administrators can provide valuable feedback on the quality of the adapted test based on the ease with which they were able to engage the participants, the ease with which they could administer the tasks, and the general quality of the interactions. If some tasks induced a negative reaction or uneasiness from either the test administrator or interviewer, a review of those tasks may be required to make them more acceptable.

Feedback from Participants

When working in communities where people are unfamiliar with the concept of psychological testing, it is helpful to get a participant to provide information on their experience in participating in psychological assessments. During such sessions, the participants can point out content, methods, and procedures they may have found difficult to fully engage in. This would provide the test developer with information on how to further improve the test, or alternatively information that can explain the patterns of results from the psychometric analysis.

Documentation

An important aspect of test development is to keep a good record of all test development and adaptations procedures alongside the psychometric properties of the newly developed test (Hambleton et al., 2005). The documentation of the adaptation process is important in providing evidence that the test content, procedures, and scoring are appropriate for all cultural groups, and that the context of the test is being applied. Documentation allows for interested parties to be able to trace the adaptations made. One needs to keep a record of different steps in the process, including the initial forward version, recommendations made by the panel of experts, the back-translation, problems faced during the pretest period, all modifications made, and the final version of the questionnaire used. It is also necessary to describe the samples used in this process (i.e., the composition of the expert panel and the pretest respondent samples). For the latter, the number of individuals as well as their basic characteristics should be described, as appropriate.

Conclusion

Transferring tests from one context to another presents significant problems that can threaten the validity of data. Systematic approaches to test development and adaptations

can contribute to the development of tests and procedures that are both culturally appropriate and psychometrically sound.

See also: Cross-Cultural Research Methods in Sociology.

Bibliography

Abubakar, A., 2008. Infant-toddler Development in a Multiple Risk Environment. Ridderprint, Ridderkerk.

Abubakar, A., van de Vijver, F., Mithwani, S., Obiero, E., Lewa, N., Kenga, K., Holding, P.A., 2007. Assessing developmental outcomes in Kenyan children following prophylaxis for seizures in cerebral malaria. Journal of Health Psychology 12, 413–430.

Cox, J.L., Holden, J.M., Sagovsky, R., 1987. Edinburgh Postnatal Depression Scale (EPDS). The British Journal of Psychiatry 150.

Foxcroft, C.D., 2002. Ethical issues related to psychological testing in Africa: what I have learned (so far). Online readings in psychology and culture. Available from: http://www.wwu.edu/∼culture.

Gopaul-McNicol, S., Armour-Thomas, E., 2002. Assessment and Culture. Academic Press, New York.

Grigorenko, E.L., Geissler, P.W., Prince, R., Okatcha, F., Nokes, C., Kenny, D.A., Sternberg, R.J., 2001. The organization of Luo conceptions of intelligence: a study of implicit theories in a Kenyan village. International Journal of Behavioural Development 25, 367–378.

Hambleton, K.R., Merenda, P.F., Spielberger, C.D., 2005. Adapting Educational and Psychological Tests for Cross-cultural Assessment. Lawrence Erlbaum Associates, Mahweh, NJ.

Holding, P., Abubakar, A., Wekulo-Kitsao, P., 2008. A systematic approach to test and questionnaire adaptations in an African context. In: Paper presented at the 3mc, Berlin, Germany.

Jukes, M.C.H., Grigorenko, E.L., 2010. Assessment of cognitive abilities in multiethnic countries: the case of the Wolof and Mandinka in the Gambia. British Journal of Educational Psychology 80 (1), 77–97.

Malda, M., van de Vijver, F.J.R., 2005. Assessing cognition in nutrition intervention trials across cultures. In: Vorster, H.H., Blaauw, R., Dhansay, M.A., Kuzwayo, P.M.N., Moeng, T.L., Wentzel-Viljoen, E. (Eds.), Annals of Nutrition and Metabolism. Karger, Basel.

Malda, M., Van de Vijver, F.J.R., Srinivasan, K., Transler, C., Sukumar, P., Rao, K., 2008. Adapting a cognitive test for a different culture: an illustration of qualitative procedures. Psychology Science Quarterly 50, 451–468.

Malda, M., van de Vijver, F.J.R., Srinivasan, K., Transler, C., Sukumar, P., 2010a. Traveling with cognitive tests: testing the validity of a KABC-II adaptation in India. Assessment 17 (1), 107–115. http://dx.doi.org/10.1177/1073191109341445.

Malda, M., van de Vijver, F.J.R., Temane, Q.M., 2010b. Rugby versus soccer in South Africa: content familiarity contributes to cross-cultural differences in cognitive test scores. Intelligence 38 (6), 582–595. http://dx.doi.org/10.1016/j.intell.2010.07.004.

Miller, V., Onotera, R.T., Deinard, A.S., 1984. Denver developmental screening test: cultural variations in Southeast Asian children. Journal of Pediatric 104, 481–492.

Mpofu, E., 2001. Indigenization of the psychology of human intelligence in sub-Saharan Africa. In: Lonner, W.J., Dinnel, D.L., Hayes, S.A., Sattler, D.N. (Eds.), Online Readings in Psychology and Culture Unit 5, (Chapter 2). Center for Cross-Cultural Research, Western Washington University, Bellingham, Washington USA. http://www.ac.wwu.edu/∼culture/index-cc.htm.

Patel, V., Simunyu, E., Gwanzura, F., Lewis, G., Mann, A., 1997. The Shona Symptom Questionnaire: the development of an indigenous measure of common mental disorders in Harare. Acta Psychiatrica Scandinavica 95 (6), 469–475.

Van de Vijver, F.J.R., 2002. Cross-cultural assessment: value for money. Applied Psychology: An International Review 51 (4), 545–566.

Van de Vijver, F.J.R., Hambleton, R.K., 1996. Translating tests: some practical guidelines. European Psychologist 1, 89–99.

Van de Vijver, F.J.R., Tanzer, N.K., 2004. Bias and equivalence in cross-cultural assessment: an overview. European Review of Applied Psychology 54, 119–135.

Vierhaus, M., Lohaus, A., Kolling, T., Teubert, M., Keller, H., Fassbender, I., Schwarzer, G., 2011. The development of 3- to 9-month-old infants in two cultural contexts: Bayley longitudinal results for Cameroonian and German infants. European Journal of Developmental Psychology 8, 349–366.

Erikson, Erik Homburger (1902–94)

James E Marcia, Simon Fraser University, Burnaby, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Erik Erikson is an ego psychoanalytic theorist who developed a psychohistorical approach to human development and formulated an epigenetic sequence of ego developmental stages encompassing the human life span from infancy through old age. His theory provides a linkage between intrapersonal Freudian theory and social–contextual approaches to personality development. He emphasizes strengths accrued by the individual in the context of social institutions or rituals keyed to individual physiological development.

Erik Erikson (1902–94) emigrated from Germany to the United States in 1933. While working at institutions on both the east and west coasts of that country, he formulated his psychohistorical approach, which, in contrast to most other psychoanalytic schools, emphasizes the impact of social factors on personality development. Most importantly he provides a schedule for ego development extending from infancy through old age. His life span developmental schema consists of eight age-specific 'crises' in ego growth. Each of these has a somatic, social, and psychological component. The successful resolution of each crisis is important for optimal movement through the succeeding stages. The fifth psychosocial stage, late adolescence, is characterized by the crisis of 'ego identity vs identity confusion', the concept with which Erikson's name is most closely linked. Some of his best-known books are *Childhood and Society*, *Young Man Luther*, *Gandhi's Truth*, and *Identity: Youth and Crisis*. Although criticized by both 'left' and 'right' wing psychoanalytic thinkers and by some feminists, Erikson's theory remains one of the most influential, and researched, psychoanalytic theories of development. His work appears in almost all textbooks on human development as well as personality theory.

Erik Erikson was born on 21 June 1902 in Frankfurt, Germany, to a Danish Jewish mother, Karla Salomonsen. He was eventually given his stepfather's name (Homburger) and later christened himself, Erikson. He grew up in Germany, completed his gymnasium in Karlsruhe and in 1921 enrolled in art school. However, he soon left this and, after a *Wanderschaft* period, he secured his first professional position in the late 1920s as a teacher, together with Peter Blos, in Vienna's Heitzing School founded by Eva Rosenfeld, Dorothy Burlingham, and Anna Freud. The latter became Erikson's analyst and supervisor in play therapy with children. Erikson married a Canadian dancer, Joan Serson, in 1930. She was to become a significant influence upon, and contributor to, his life's work. In 1933, now the father of two sons, Kai and Jon, he left his incompleted studies at the University of Vienna, was voted a full member of the Vienna Psychoanalytic Society and emigrated to the United States with his young family. He established a private practice in Boston and was made a member of the Boston Psychoanalytic Society despite that body's parochial policy of limiting membership to physicians. During the next 5 years, Erikson worked at the Judge Baker Center in Boston, the Murray Clinic at Harvard, secured a position at Yale with the Gesell clinic and Institute of Human Relations, and began an important professional relationship with the anthropologist, Margaret Mead. In 1936 he became a father for the third time with the birth of his daughter, Sue. His fourth child, Neill, was born much later in 1944.

In 1938, Erikson, whose surname until this time had been his stepfather's (Homburger), changed his name to Erikson, became a US citizen and moved to California to take a position at Macfarlane's Institute of Child Welfare in Berkeley and established a substantial private practice in San Francisco in 1942–43. While working with returning veterans at the Mount Zion Rehabilitation clinic, he began a friendship with the Jungian analyst Joseph Wheelwright, who broadened Erikson's Freudian perspective. In 1949 he became a professor at Berkeley, in spite of lacking the usual academic credentials, but soon left after refusing to sign the McCarthy era loyalty oath – considered by some a 'Here I stand' gesture that may have prefigured his forthcoming book on Martin Luther.

In 1950 he presented his and Joan's outline of the human life cycle to a White House Conference on Children and Youth and saw his most influential book, *Childhood and Society*, published. He returned to the eastern US in 1951 to join the Austen Riggs Center in Stockbridge, Massachusetts, as clinician and author. There he established important professional and personal relationships with ego psychoanalysts Robert Knight and David Rapaport. At this time, he completed his second major book, *Young Man Luther* (1958). He published the monograph *Identity and the Life Cycle* in *Psychological Issues* (1959) expanding both on the life cycle outline and on ego identity, a concept, together with 'identity crisis' that was to become associated with his name in the mind of the public. The preface to this article by David Rapaport provides a valuable introduction to ego psychology in general and locates Erikson within that tradition.

With help from David Riesman and Robert White, Erikson moved from Stockbridge to Boston and Harvard in 1960 and became a professor, but outside of any particular academic department. In the midst of, and to some extent with reference to, the student unrest of the 1960s, Erikson began work on his second psychobiography, *Gandhi's Truth* (1969) for which he received both the Pulitzer Prize and the National Book Award. In addition, *Insight and Responsibility* (1964), *Identity: Youth and Crisis* (1968), and a revision of *Childhood and Society* (1963/1985) were also completed during this period. These books

reflected a growing shift in Erikson's thinking from the development of the individual, to the social context of individuals' development, to the moral and ethical implications of different forms of social organization for humankind. In this transition, the work of his friend, Paul Tillich, was influential. Clearly, Erikson had begun to move quite a distance from his orthodox psychoanalytic initiation by Anna Freud.

By the beginning of the 1970s, Erikson had become one of the most respected and honored thinkers within the psychoanalytic developmental area, described by one reviewer as 'the most influential of living psychoanalysts'. He delivered two sets of honorary lectures: the Godkin at the John F. Kennedy School for Government at Harvard in 1972, and the Jefferson for the National Endowment for the Humanities in Washington, DC, in 1973. These eventuated in two more books, *Toys and Reasons* (1977) and *Dimensions of a New Identity* (1974). Another book, *Life History and the Historical Moment* (1975), occasioned significant criticism from both feminists and New Left sources who found Erikson too politically and psychologically conservative.

Erikson's most important final writings were on the last life cycle stage of Integrity. These were *Reflections on Dr Borg's Life Cycle*, a commentary on Bergman's film, 'Wild Strawberries', *The Life Cycle Completed: A Review* (1982), and *Vital Involvement in Old Age* (1986) written with Joan Erikson and Helen Kivnick. A synthesis of Erikson's unpublished conference papers and manuscripts has been published by Hoare (2002). Erikson continued to do clinical work until 1990. He died on 12 May 1994, in Harwich, Massachusetts.

Erikson lived through the world-shaping events of the twentieth century: World War I, the Great Depression, World War II, and the United States' social–political Coming of Age in Vietnam. The national and ideological strands that the Erikson gathered up and wove into his psychosocial developmental theory were numerous: Danish, German, North American, Jewish, Protestant, classical psychoanalytic, ego psychoanalytic, existential, and even Jungian. His literary/theoretical excursions into a form of biography that he labeled 'psychohistory' further widened his intellectual scope to include Luther in Reformation Germany, Gandhi in preindependence India, Hitler in pre-World War II Germany, Gorky in Bolshevik Russia, and Jesus Christ in biblical Palestine. Although only one of a number of psychoanalytic thinkers who left Europe for America in the 1930s, only Erikson integrated and transformed these disruptive experiences into a clinically, theoretically, and empirically useful outline of the stages of normal human development. Two significant biographies on Erikson's life and work have been published: Coles (1970) and especially Friedman (1999).

In the beginning of his work, Erikson took a 'configurational' approach to children's play therapy, wherein he considered the spatial arrangement of items chosen by the child to be as important as the items themselves or the stories linking them. He carried this configurational approach over into his psychosocial developmental chart that portrays the human life cycle as consisting of eight central developmental crises, each having its more or less specific age-related time of ascendance. The first three stages: trust vs mistrust (infancy), autonomy vs shame, doubt (early childhood), and initiative vs guilt (play age) emerge from the Freudian psychosexual stages: oral, anal, and phallic. Erikson's stages of ego growth in childhood extend beyond their foundational Freudian erogenous body zones to include the development of associated modes of behavior (e.g., active and passive incorporation at infancy). A further extension of development from body zones and their behavioral modes is his description of specific stages of ego growth. These stages reflect the individual's senses of oneself and of the world and are shaped by the child-rearing practices of particular parents embedded within their particular culture.

The description of life cycle stages after childhood constitutes one of Erikson's unique contributions to psychoanalytic developmental theory. He continues his outline of a sequential ego development to extend beyond childhood through the adult years: Industry vs inferiority (school age), identity vs identity confusion (adolescence), intimacy vs isolation (young adulthood), generativity vs stagnation (middle adulthood), and integrity vs despair (old age). All of these ego developmental stages have their basis in physiological changes and all occur within the context of relevant social institutional 'rituals' (e.g., basic trust at infancy within 'motherhood'; industry at school age within 'schools' or other cultural child education practices, etc.). It is this grounding in bodily changes and their psychological significance for psychosocial growth that provides the basis for the theory's cross-cultural applicability. Finally, he also noted a schedule of associated 'virtues', which emerge from the resolution of each stage: basic trust > 'hope', autonomy > 'will', initiative > 'purpose', industry > 'competence', identity > 'fidelity', intimacy > 'fidelity', generativity > 'care', and integrity > 'wisdom' (**Figure 1**).

The emergence of each of the eight stages in this sequence is governed by an epigenetic principle comprising a built-in developmental progression of individual physical and psychological change occurring within a social context of age-related 'average expectable' demands and provisions. Because Erikson located personality development firmly within the social context, he is to be credited with initiating psychosocial developmental theory. Optimal development in childhood consists of a balance between the opposite poles described above, tilted in favor of the more positive alternative. Beginning in adolescence, resolution of a developmental crisis incorporates both poles into a synthesis. For example, identity development necessarily includes an element of diffusion or confusion; generativity formation incorporates some stagnation; and there is no integrity without some accompanying despair.

The primarily positive resolution of each stage is considered to be crucial for the resolution of the next stage. However, consistent with Erikson's configurational approach, there is an aspect to the design of his developmental chart that allows for remediation of inadequately resolved issues or precocious resolution of 'future' developmental issues. Each stage is assumed to arise during every chronological period, although the nature of that stage is shaped by the dominant issue of the period. For example, a basic trust vs mistrust crisis reoccurring at intimacy will take a form configured by the adult nature of relationships typical of early adulthood. Because they arise again at each succeeding chronological period, previously unresolved age-specific issues can be remediated through the provision of 'better-than-average expectable' conditions (e.g., counseling, psychotherapy, or enriched milieux). Similarly,

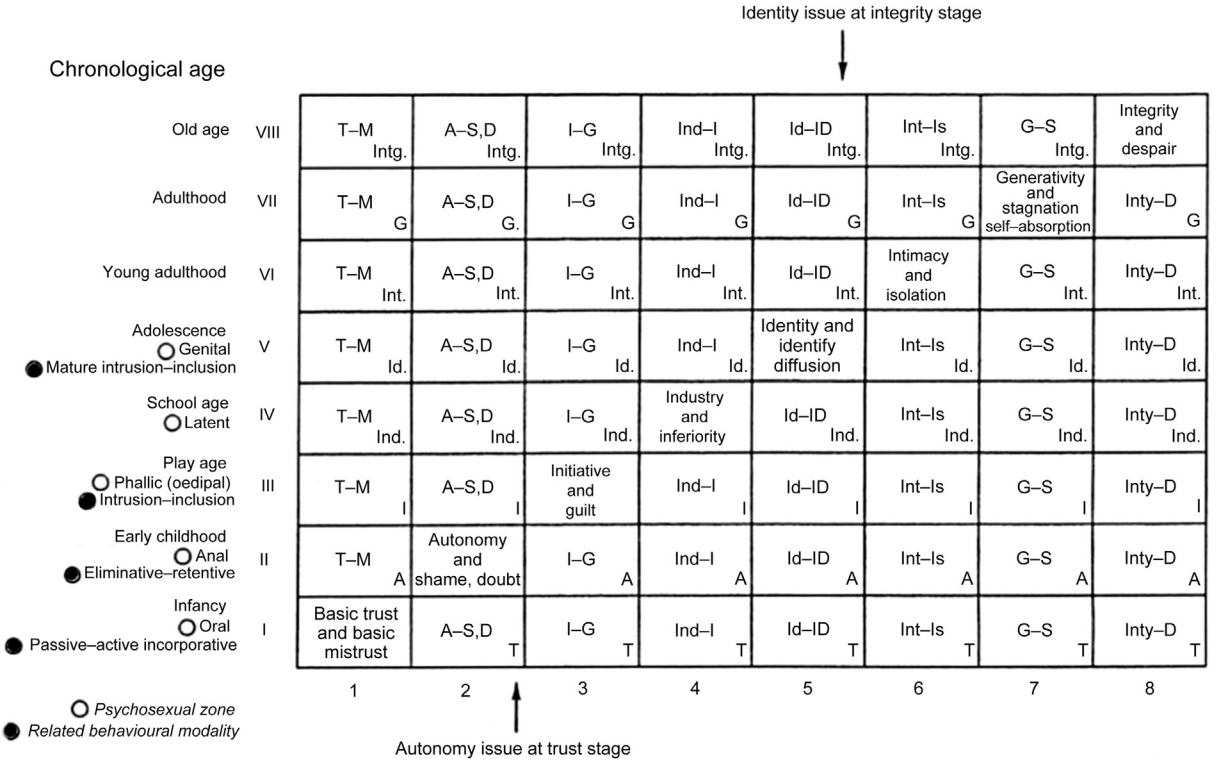

Figure 1 Psychosocial developmental diagram. Note that Erikson's original 'vs' has been replaced with 'and'. This reflects his thinking that stages are not a matter of 'either–or' resolution but involve a blend of the polar qualities, optimal development being a predominance of the positive pole. In addition, all squares in the chart are filled in to indicate that every stage occurs at every other stage.

because forerunners of succeeding stages are present at every stage, the theory suggests the possibility of precocious resolution of stages not yet encountered at the suggested chronological age.

In the broadest view, Erikson proposes that there are crucial periods ('crises') of development characterizing the human life cycle for which societies have developed institutions that, more or less successfully, furnish a matrix within which these developmental crises are resolved. Ideally, societal rewards and demands mesh with individual needs and abilities at each of these stages. Whereas Freud furnished a view of personality from the 'inside out', looking to the individual's unconscious in order to understand cultural products, Erikson, while accepting the essential Freudian dynamic principles, added a perspective from the 'outside in', beginning at the level of the whole individual and describing a particular culture's impact on human development. Also, in contrast to Freud, Erikson saw ego growth as continuous throughout the life cycle – from infancy through old age.

Erikson has been criticized from a number of quarters. Orthodox psychoanalytic theorists (e.g., Jones, Eissler) saw Erikson's emphasis on social factors and the ego in the development of personality as paying insufficient attention both to the causative role of early childhood experiences and to the primacy of unconscious factors in determining later adult personality. In short, they found him insufficiently 'Freudian'. Early feminist critics (e.g., Weisstein, Millett) took Erikson to task for his assumption of a unique developmental trajectory for women based upon their biological differences from men – in other words, too 'Freudian'. Later feminists (e.g., Gilligan) criticized him for paying insufficient attention to women's differences from men in terms of their capacities for care and relationship. Object relations theorists (e.g., Fairbairn, Winnicott) and self-theorists (Kohut) felt that Erikson was somewhat superficial in underemphasizing the importance of early mother–child interactions and overemphasizing the importance of social factors in personality development. By contrast, some historians and sociologists (e.g., Kushner, Weinstein, and Platt) have found Erikson's theory too little concerned with underlying causative social factors and questioned the universality of his eight-stage model. Social critics from the left (e.g., Fromm, Kovel) thought Erikson too conservative in his apparent acceptance of the social status quo. And, finally, social scientists (e.g., J. McV. Hunt) found Erikson's writing too imprecise to allow for empirical research. Erikson, himself, was critical of attempts to validate his theory with experimental methods, preferring more clinical and social approaches. The abundance of research that has emerged in the past 30 years largely vitiates the criticism of nontestability of the theory.

Erikson's primary contribution, emerging from the smoke of all the controversy noted above, has been a comprehensive overview of the human life cycle that incorporates somatic, psychological, and social elements. The influence of his psychosocial approach has been pervasive throughout ego psychoanalytic theory, developmental psychological research, and educational counseling, although there is no 'Eriksonian'

school of psychotherapy. Almost every psychology textbook on personality theory and developmental psychology of children, adolescents, and adults has a section devoted to Erikson. Although his descriptions of early developmental stages usually are covered, it is the identity issue at adolescence with which he is most closely associated, followed by his descriptions of adult life cycle stages. Even political theory, especially the 'politics of identity', has been influenced by his approach. Erikson has amended Freudian theory to make it more understandable, 'experience near', growth focused, researchable, and socially useful. Although criticized by classical Freudians, Erikson's extension of psychoanalytic theory into the realm of normal as well as optimal human development, may have saved that theory from its excessive interiority, pessimism, and obscurantism. In addition, because of its essentially interpersonal nature, Erikson's theory has provided a valuable link between drive-based classical and ego psychoanalytic theory and relationship-based object relations theory. From a social–political perspective, the psychosocial developmental outline proposed by Erikson is sufficiently clear and putatively universal that it can provide a guide for social programs that focus on establishing optimal conditions for human development across the life cycle.

To date, research has established validity for the last five Eriksonian stages: industry (Kowaz and Marcia, 1991); identity (Marcia, 1966; Marcia et al., 1993); intimacy (Orlofsky et al., 1973); generativity (Bradley and Marcia, 1998; McAdams and de St. Aubin, 1992); and integrity (Hearn et al., 2012; Walasky et al., 1983–1984). Kroger (2007) has authored a textbook on adult development. As well, two handbooks on identity theory and research have been published: (Marcia, et al., 1993) and (Schwartz, et al., 2012). Extension of Eriksonian theory into political psychology has been undertaken by Hoover (1997). A graduate school and treatment center based upon Erikson's theory, the Erikson Institute, has been established in Chicago. As well, an international society (Society for Research in Identity Development) and a related research journal (*Identity: An International Journal of Theory and Research*), both having their impetus in Erikson's work, have been in existence since 1998. Research on this theory continues into the present, much of it having to do with applications to child and adult development.

See also: Ethnic Identity, Psychology of; Infancy and Childhood: Emotional Development; Psychohistory; Self and Identity Development During Adolescence across Cultures.

Bibliography

Bradley, C., Marcia, J.E., 1998. Generativity-Stagnation: a five category model. Journal of Personality 66 (1), 39–64.

Coles, R., 1970. Erik H. Erikson: The Growth of His Work, first ed. Atlantic Monthly Press, Boston, MA.

Erikson, E.H., 1958. Young Man Luther: A Study in Psychoanalysis and History, first ed. Norton, New York.

Erikson, E.H., 1959. Identity and the Life Cycle: Selected Papers. Psychological Issues 1. International Universities Press, New York.

Erikson, E.H., 1963/1985. Childhood and Society, second ed. Norton, New York.

Erikson, E.H., 1964. Insight and Responsibility: Lectures on the Ethical Implications of Psychoanalytic Insight, first ed. Norton, New York.

Erikson, E.H., 1968. Identity, Youth and Crisis, first ed. Norton, New York.

Erikson, E.H., 1969. Gandhi's Truth: On the Origins of Militant Nonviolence, first ed. Norton, New York.

Erikson, E.H., 1974. Dimensions of a New Identity: The 1973 Jefferson Lectures in the Humanities, first ed. Norton, New York.

Erikson, E.H., 1975. Life History and the Historical Moment, first ed. Norton, New York.

Erikson, E.H., 1976. Reflections on Dr Borg's life cycle. Daedalus 105, 1–28.

Erikson, E.H., 1977. Toys and Reasons: Stages in the Ritualization of Experience, first ed. Norton, New York.

Erikson, E.H., 1982. The Life Cycle Completed: A Review. Norton, New York.

Erikson, E.H., Erikson, J.M., Kivnick, H.Q., 1986. Vital Involvement in Old Age. Norton, New York.

Friedman, L.J., 1999. Identity's Architect: A Biography of Erik H. Erikson. Scribner, New York.

Hearn, S., Saulnier, G., Strayer, J., Glenham, M., Koopman, R., Marcia, J.E., 2012. Between integrity and despair: toward construct validation of Erikson's eighth stage. Journal of Adult Development 19 (1), 1–20.

Hoare, C.H., 2002. Erikson on Development in Adulthood: New Insights from the Unpublished Papers. Oxford University Press, New York.

Hoover, K., 1997. The Power of Identity: Politics in a New Key. Chatham House, Chatham, NJ.

Kowaz, A., Marcia, J.E., 1991. Development and validation of a measure of Eriksonian industry. Journal of Personality and Social Psychology 60, 390–397.

Kroger, J., 2007. Identity Development: Adolescence through Adulthood, second ed. Sage, Thousand Oaks, CA.

Marcia, J.E., 1966. Development and validation of ego identity status. Journal of Personality and Social Psychology 3, 551–558.

Marcia, J.E., Waterman, A.S., Matteson, D.R., Archer, S.A., Orlofsky, J.S., 1993. Ego Identity: A Handbook for Psychosocial Research. Springer-Verlag, New York.

McAdams, D., de St. Aubin, E., 1992. A theory of generativity and its assessment through self-report, behavioural acts, and narrative themes in autobiography. Journal of Personality and Social Psychology 62, 1003–1015.

Orlofsky, J., Marcia, J.E., Lesser, I., 1973. Ego identity status and intimacy versus isolation crisis in young adulthood. Journal of Personality and Social Psychology 3, 415–423.

Schlein, S. (Ed.), 1987. A Way of Looking at Things: Selected Papers from 1930–1980, first ed. Norton, New York.

Schwartz, S.J., Luyckx, K., Vignoles, V.L. (Eds.), 2012. Handbook of Identity Theory and Research, vols. I and II. Springer, New York.

Walasky, M., Whitbourne, S.K., Nehrke, M., 1983–1984. Construction and validation of an ego integrity status interview. International Journal of Aging and Human Development 18 (1), 61–72.

Erikson's Theory of Psychosocial Development

Li-fang Zhang, The University of Hong Kong, Hong Kong

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article demonstrates that contrary to some people's belief that Erikson is passé, Erikson is very much in fashion. It does so by providing an up-to-date, panoramic picture of the scholarship founded on Erik Erikson's theory of psychosocial development. It begins with a succinct introduction of the principal tenets of Erikson's theory and those of the most widely studied construct rooted in Erikson's theory – that of identity status proposed by James Marcia. Following this, are a critical analysis of the major critiques on Erikson's theory and a synthetic review of the chief efforts to improve the theory. The third part highlights empirical evidence supporting the validity of the Eriksonian stages and that of identity statuses. Finally, taking a firm stand on the recurrent, long-standing controversial issues over Erikson's theory, the article draws three conclusions – that Erikson's theory is highly relevant today and perhaps will be for a long time into the future; that it concerns both self and society; and that it applies to both genders. It is further contended that despite its limitations, the Eriksonian literature has practical implications for education and beyond.

Introduction

One of the major achievements in the history of psychology has been Erik Erikson's (1950, 1968) theory of psychosocial development. This article is divided into four parts. First, Erikson's theory and Marcia's (1966) concept of identity status are described. Second, principal critiques and the evolution of Erikson's theory are highlighted. Third, empirical findings rooted in the theory are reviewed. In the final part, some conclusions are drawn.

Theory of Psychosocial Development and Concept of Identity Status

Erikson first presented his theory of psychosocial development in his book, *Childhood and Society* (1950). Drawing heavily on the biological principle of epigenesis, Erikson articulated a sequence of stages during which ego development occurs, with each stage being related to all the others. Erikson's theory was profoundly influenced by Freud's (1894) psychoanalytic approach. However, as Hoare (2005) put it, as "the second stage psychoanalytic theorist" (p. 19), Erikson shifted Freudian thought "upward in consciousness, outward to the social world, and forward throughout the complete life span" (Hoare, 2005).

Central to Erikson's theory is the concept of identity. For Erikson, identity was multidimensional, to be viewed through a wide-angle lens. According to Erikson, an individual develops his or her identity through a sequence of eight stages that define the life cycle: (1) trust versus mistrust; (2) autonomy versus shame and doubt; (3) initiative versus guilt; (4) industry versus inferiority; (5) identity versus role confusion; (6) intimacy versus isolation; (7) generativity versus stagnation; and (8) ego integrity versus despair. Each stage is marked by a psychosocial crisis that is the product of the soma (body), the ego (psyche), and society. At each stage, the individual confronts new challenges and tries to overcome them, and each stage builds upon the successful completion of earlier stages. However, this does not mean that the pattern of individuals' psychosocial development is linear. Rather, it is cyclical in that the challenges of stages not successfully completed may be expected to reappear as problems at a later stage in life.

Over the years, Erikson's theory has engendered much theorization and empirical work, the most notable theorization being that of James Marcia (1966). Upon thoroughly examining the various common issues facing youths at the time, Marcia (1966) conceived the theoretical construct of 'identity status.' In particular, he elaborated Erikson's original adolescence stage of 'identity versus identity confusion.' According to Marcia, the criteria for attaining a mature identity are based on two essential variables: crisis/exploration and commitment. Crisis/exploration refers to the process during which adolescents actively examine developmental opportunities and identity issues, question goals and beliefs defined by their parents, and begin to search for personally appropriate alternatives in respect to occupation, goals, values, and beliefs. Commitment refers to the extent to which an individual is personally committed to self-selected aspirations, beliefs, goals, values, and an occupation. Based on the absence or the presence of crisis/exploration and commitment, Marcia characterized four identity statuses: identity achievement, moratorium, foreclosure, and diffusion (Marcia, 1966).

Major Critiques and Proposed Modifications

Like any other theory, Erikson's has had its share of criticism. This part first presents criticisms of Erikson's theory as a framework for understanding human development in the contemporary era. It then introduces major efforts to modify Erikson's original theory of psychosocial development.

Criticisms of Erikson's Theory of Psychosocial Development

Throughout the years, critiques of Erikson's theory have continued to appear from various perspectives, including those

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23200-5

of radical behaviorists, postmodernists, human developmentalists, and feminist writers. However, the criticisms seemed to be have been made in a piecemeal fashion, with a lack of dialogue among the critics. Largely due to this lack of communication, some scholars noted that there appears to be a growing consensus that Erikson is passé (Côté, 2002a). In response to the apparent loss of interest in Erikson's theory, Sorell and Montgomery (2001) posed what they saw as the three central questions concerning Erikson's theory. First, does the historical and sociocultural context in which Erikson's theory was constructed weaken its contemporary usefulness? Second, can a theory that ties individual development to a biological ground plan – hence placing strong emphasis on uniformity of experience – offer guidance for investigations into identity formed in the contemporary era, which is characterized by fast changing and multicultural contexts? Third, is the theory inherently androcentric and therefore unable to serve as a useful framework for understanding females', or even males', psychosocial development?

Concerning the first question, Sorell and Montgomery (2001) argued that Erikson's theory is a very useful guiding framework in contemporary settings, especially for identity research. The authors further contended that Erikson's theory distinguishes itself from all other developmental theories in that its construction was based upon a wide range of personal experiences and observations drawn from a variety of cultural contexts.

Regarding the theory's linking individual development to a biological ground plan, Sorell and Montgomery (2001) pointed out two problems. First, although Erikson extensively elaborated the link between biological maturation and psychosocial development from infancy through adolescence, he made no mention of the relationship between the two beyond adolescence. Second, Sorell and Montgomery (2001) noted that some had criticized Erikson for his 'anatomy is destiny' viewpoint concerning psychosocial development for males and females and the role that reproductive anatomy plays in Erikson's concept of self and identity. However, Sorell and Montgomery (2001) disputed this common criticism, stating that Erikson integrated the notion of biological ground plan into his theory of psychosocial development in a manner that is far more sophisticated than the 'anatomy is destiny' interpretation offered by earlier critiques.

As regards the third issue, Sorell and Montgomery (2001) believed that Erikson's theory is essentially androcentric because "it is organized by a definition of identity development as the construction of an individuated and separated self" (p. 113). They pointed out that only the first three Eriksonian stages have to do with caring relationships, thus supporting the viewpoint of earlier feminist writers (e.g., Archer, 1992; Gilligan, 1982) that Erikson's theory largely concerns males' identity development.

To stimulate scholarship on Erikson's theory, Côté (2002a) invited commentaries on Sorell and Montgomery's (2001) article. While they were either in agreement with or silent about the first two issues raised by Sorell and Montgomery (2001), the commentators largely diverged from Sorell and Montgomery on the third issue – the question of whether or not Erikson's theory is so androcentric that it is unable to describe females' psychosocial development. For example, Côté (2002b) convincingly argued that a balanced view of the selves can easily be seen in Erikson's notion of generativity, maintaining that Erikson had made it clear that the development of the stable sense of self is a logically necessary precursor to the establishment of stable relational bonds with others. Côté (2002b) further contended that Erikson's theory addresses the notion of care far beyond the first three stages. From an Eriksonian perspective, Côté concluded, the necessary condition for adulthood is a sense of active and caring responsibility for others, not only within one's own generation, but also for future generations.

Côté (2002b) was not alone in his view that Erikson did emphasize the relational self. Levine (2002) insisted that Erikson "always knew and, in fact, wrote about at great length" (p. 273) the dialectic of independence–interdependence. Kroger (2002) articulated her divergence from Sorell and Montgomery's (2001) view that Erikson's theory was androcentric at two levels. At a general level, quoting widely from Erikson's (1968) "*Identity, Youth and Crisis*," Kroger noted that Erikson's definitions of identity "continually emphasize the role of the individual and his or her interpersonal context as intricately intertwined in the development of ego identity" (p. 260). At a more specific level, focusing on Erikson's analysis of the notion of generativity, Kroger (2002) successfully made the case that generativity – the ability to care for one's offspring, one's work, one's community, and one's future generations – should be a task that is highly relevant for both genders.

Modifications to Erikson's Theory

Since the late 1980s, much effort has gone into modifying Erikson's theory of psychosocial development. One line of work has been the elaboration of Erikson's publications concerning adulthood. For example, based on Erikson's unpublished materials, Carol Hoare (2002) wrote a monograph, "*Erikson on development in adulthood: New insights from unpublished papers*," in which she provided an exhaustive account of Erikson's thoughts on what it means to be an adult. In particular, she lucidly interpreted Erikson's conceptions of such notions as insightful adult, wise adult, and mature adult. Further based on the published and unpublished writings of Erikson and on her interviews with Erikson's widow (Joan Erikson), Hoare (2009) wrote a paper in which she explored the meaning of spirituality in adult identity development, following the trajectory framed by Erikson – that of the progression from morality to ideology, to aesthetics, and finally to mature ethics.

While also focusing on an adult stage (generativity versus stagnation), Slater's (2003) work distinguished itself by adding new concepts to Erikson's (1968) epigenetic diagram, thus increasing the breadth of the central crisis of generativity versus stagnation. These new concepts of psychosocial conflicts include (1) inclusivity versus exclusivity, (2) pride versus embarrassment, (3) responsibility versus ambivalence, (4) career productivity versus inadequacy, (5) parenthood versus self-absorption, (6) being needed versus alienation, and (7) honesty versus denial.

Another major approach to modifying Erikson's work has centered on addressing the concern over the lack of association between biological change and psychosocial change beyond adolescence, and on proposing a ninth Eriksonian stage. In response to the criticism that Erikson's epigenetic theory did not describe the relationship between biological and psychosocial change beyond adolescence, Joan Erikson (Erikson and Erikson, 1997) expanded Erik Erikson's book published in 1982, elaborating at length both the biological changes and psychosocial changes concerning the 'integrity versus despair' stage and the implications of increased life expectancy in many contemporary societies. Furthermore, Joan Erikson added a ninth stage to Erikson's original scheme, focusing on the physical needs of the aging body and what they mean for integrity.

Joan Erikson's idea of proposing a ninth stage was reinforced by Forrest and Côté's (2002) discussion of what they called a subdivision of Erikson's eighth stage of life – "The mortal stage of late life." Forrest and Côté maintained that mortality presents "the optimal task of Realization (in many positive coping and existential senses) versus Denial (or other nonrecognition or emotional paralysis) or Fear (or apprehension or even terror)" (p. 329).

Finally, retaining the eight psychosocial crises in Erikson's (1959) model, Capps (2004) conceptualized each crisis and its resolution within the context of a decade. According to this reclassification, six of the eight stages occur in the adult years, thus reorienting the life cycle scheme more toward adulthood from early childhood. Moreover, Capps proposed a ninth and a tenth decade, with the ninth decade's central developmental issue being, again, 'basic trust versus basic mistrust,' and the tenth's repeating the life stage of 'autonomy versus shame and doubt.'

Empirical Evidence

Erikson's 1951 study heralded a line of research on the various Eriksonian developmental stages, while Marcia's (1966) empirical work marked the beginning of research on the notion of identity status (Eriksonian stage 5). Both lines of research have yielded findings that have not only lent strong support to Erikson's theory, but also helped to address some of the theoretical concerns previously discussed.

In this part, Erikson's theory of psychosocial development is examined in light of research evidence. It is divided into five sections. The first provides a brief survey of the research scope and methodology. The second reviews major work on testing the validity of the Eriksonian stages. The third highlights studies on gender-relevant concepts. The fourth introduces studies that examine the relationship between the Eriksonian stages and other variables. The final section underscores research on identity.

Scope and Methodology of Research

With the exception of some studies of young children (e.g., Archer, 1989a; Erikson, 1951; Wang and Viney, 1996), the majority of the studies have been conducted among college students and adults, with some specifically concerning the elderly population (e.g., Brown and Lowis, 2003; Hughes and Peake, 2002; Hannah et al., 1996; Zauszniewski and Martin, 1999). Furthermore, these studies have been undertaken in different parts of the world, with most of them having been conducted in North America.

Concerning research methodology, like many other areas in psychology, research on psychosocial development has been largely cross sectional. Moreover, studies investigating the Eriksonian stages have tended to employ self-report inventories that elicit responses to be scored on Likert scales, while those examining identity statuses have been much more likely to use such qualitative approaches as semistructured and autobiographical interviews, observations, and analyses of life history. Finally, between 1951, when Erikson conducted his first empirical study, and the first decade of the twenty-first century, when Luyckx and his colleagues (Luyckx et al., 2008) constructed the Dimensions of Identity Scale, more than two dozen inventories assessing the Eriksonian stages and the Marcian identity statuses were created (see the Appendix for a list of the major inventories identified in the literature).

Validating the Eriksonian Stages

Studies focusing on validating the Eriksonian stages fall roughly into six research themes. These include (1) validating the factor structures of inventories, (2) testing developmental differences based on age and attributes associated with adulthood, (3) investigating psychosocial development variations across times in different cohorts, (4) examining psychosocial differences in cross-cultural contexts, (5) studying the developmental trajectory for each of Erikson's eight original stages, and (6) validating the eighth and ninth stages.

Validating Inventories' Factor Structures

Over the years, factor analysis of Eriksonian-stage inventories has been carried out constantly. One such inventory is Constantinople's (1969) Inventory of Psychosocial Development (IPD) designed to assess the first six of Erikson's stages. For example, Varghese (1980) tested the IPD among 252 graduate students in the United States. Factor analysis of the scores from the 12 (six positive and six negative) Eriksonian subscales yielded two significant components. The first factor, named actualization, was composed of six variables that represented the positive aspects of personality: trust, autonomy, initiative, industry, identity, and intimacy. The second factor, termed retrogression, comprised six variables that symbolized the negative aspects of personality: mistrust, self-doubt, guilt, inferiority, diffusion, and isolation.

Another inventory is Rosenthal et al.'s (1981) Erikson Psychosocial Stage Inventory (EPSI). Factor analysis of the EPSI has also lent strong support to Erikson's theory. For example, Gray et al. (1986) tested the EPSI among 534 university freshmen and sophomores in the United States. Although the seven factors (e.g., initiative, industry, identity, friendship, dating, goal clarity, and self-confidence) obtained from the analysis did not precisely correspond to the relevant Eriksonian stages, each of the factors made substantive sense in supporting Erikson's observations on late adolescent development.

Several other popular inventories have also been submitted to factor analysis. These include, among others, the Ego Identity Scale (Rasmussen, 1964), the Ego Stage Development (Caillet, 1980), and the Inventory of Psychosocial Balance (Domino and Affonso, 1990). With very few exceptions (e.g., Caillet and Michael, 1983), studies examining the factor structures of the inventories suggest that the Eriksonian stages can be assessed by reliable and valid inventories.

Finally, it is worth mentioning that in terms of examining Erikson's construct of psychosocial development through factor analysis of particular instruments, some researchers have used multiple inventories, looking for a latent variable known as the general factor of psychosocial development (GFPD). For example, Dunkel et al. (2012) conducted a series of three studies among ethnically diverse university students in the United States. Results of correlational and confirmatory factor analyses in all three studies supported the hypothesized GFPD.

Age and Attributes Associated with Adulthood

One of the assumptions underlying Erikson's theory is that an individual's psychosocial development progresses with age. It is further assumed that people with higher levels of psychosocial development would display more attributes associated with adulthood. Such assumptions have been proved to be true by quite a few studies. For example, early in 1981, when Rosenthal et al. (1981) first constructed the EPSI, they tested their inventory among 622 adolescents (aged 9 and 11 years) from nine metropolitan high schools in Australia – after having pilot tested their inventory with 106 secondary school students (aged 9 and 10 years). The researchers concluded that, as expected, older students scored higher in the positive direction on each of the six scales assessed.

In developing their Modified Erikson Psychosocial Stage Inventory (MEPSI), based on that of Rosenthal et al. (1981), Darling-Fisher and Leidy (1988) found positive relationships of psychosocial development to age and attributes associated with adulthood. The researchers found that among 168 participants (aged between 19 and 86 years), age was positively related to generativity and ego integrity levels, and that those who reported exercising regularly scored significantly higher in the positive direction of the psychosocial attributes (also Brown and Lowis, 2003).

Psychosocial Development across Times

Support for Erikson's psychosocial development theory has also been demonstrated by the results from carefully designed cross-sectional studies. For example, Zuschlag and Whitbourne (1994) analyzed data from three generations of college students. Following up on Constantinople's (1969) classic study, the investigators administered Constantinople's (1969) IPD to two cohorts of students attending the same university in 1977 and 1988. The findings suggested that, for all three times of testing (i.e., 1969, 1977, and 1988) and for most of the Eriksonian stages, college seniors were more advanced in their psychosocial development than their younger counterparts.

Cross-Cultural Studies

Although very few studies have made direct comparisons between/among cultural groups, some scholars have pointed to a certain degree of universality concerning psychosocial development. For example, Wang and Viney (1996) compared the psychosocial development of school-aged Australian and Chinese students, using the Content Analysis Scales of Psychosocial Maturity (Viney and Tych, 1985). It was concluded that, although some differences in the individual scale scores were found between the two cultural groups, a general pattern of similarity was evident (also Zhang, 2010).

The Unique Developmental Trajectory of Each Eriksonian Stage

Two related basic tenets in Erikson's theory are (1) each stage normally reaches ascendancy at a particular time in the lifespan, and (2) despite each stage having an overall trajectory describing the average individual's progression through the stages, the interaction between individuals and environments could bring about individual differences in these developmental trajectories. These theoretical arguments have been supported by empirical evidence. For example, Sneed et al. (2006) conducted a 34-year longitudinal study with 175 adults in the United States, examining three of Constantinople's (1969) IPD scales: trust, identity, and ego integrity. As expected, each of the three stages showed a distinct developmental trajectory: the trust stage followed a positive linear trajectory; the identity stage followed a curvilinear trajectory that leveled off in middle adulthood; and the integrity stage followed a curvilinear trajectory with an increasing trend in middle adulthood.

Validating the Eighth and Ninth Stages

A number of studies have focused on the eighth and ninth stages – the extreme later life stages (e.g., Hannah et al., 1996; Hughes and Peake, 2002; Njus and Johnson, 2008). Although research on these extreme later life stages, especially research involving the ninth stage, is limited, the findings of this research support the validity of the ninth stage. For instance, Brown and Lowis (2003) examined the eighth and ninth stages among 32 women in their 80s and 90s and 32 women in their 60s in the United Kingdom. Results suggested a significant relationship of age to stage 9 but not to stage 8. Moreover, a statistically significant difference was found between mean scores of the two age groups for stage 9 but not for stage 8.

Gender-Relevant Differences in the Eriksonian Stages

Although one of the long-standing controversies over Erikson's theory concerns its usefulness in explaining females' psychosocial development, studies investigating this topic have largely reported gender differences as a by-product and of secondary importance. A number of such studies (e.g., Mellor, 1989; Zhang, 2013; Zuschlag and Whitbourne, 1994) found that women generally scored significantly higher on most of the Eriksonian stages, although the opposite finding has been reported occasionally (e.g., Rosenthal et al., 1981).

While studies focusing on gender-relevant differences are scarce, the findings of these studies are informative. For example, Mellor (1989) investigated the relationship between the first six Eriksonian stages and relational self-definitions

among 388 students in public schools in the United States. Results indicated that on three of the six stages (trust, initiative, and intimacy), males with separate self-definitions scored higher than did females with like self-definitions. Meanwhile, females with connected self-definitions scored significantly higher on the same three Eriksonian stages than did males with connected self-definitions. These findings, as noted by Mellor (1989), are in line with Gilligan's notion of gender-based self–other relationships.

Waterman and Whitbourne (1982) examined the relationship between sex-role orientation as measured by the Bem Sex Role Inventory (Bem, 1974) and psychosocial development as measured by Constantinople's (1969) IPD among university students and adults. It was concluded that the findings generally supported the view that the simultaneous endorsement of both masculine and feminine qualities was more conducive to adaptive capacity. The significance of this study lies in its telling finding that not only biological sex, but also sex-role orientation, matters in psychosocial development.

The Eriksonian Stages and Their Correlates

Since the late 1960s, the Eriksonian stages have been tested among different populations. Chief among these are senior adults and college students, particularly the latter.

Research on Senior Adults

Among senior adults, the Eriksonian stages have been examined against some positive attributes that are theoretically expected to be displayed by psychosocially well-adjusted senior adults and against a wide range of indicators of late-life maladjustment. Examples of these positive attributes include purpose in life, self esteem, and self-realization (e.g., Hannah et al., 1996); learned resourcefulness (Zauszniewski and Martin, 1999); and spiritual well-being (Hughes and Peake, 2002). Examples of maladjustment indicators are death anxiety (Fishman, 1992), depression (Hughes and Peake, 2002), and sleep disturbance (Wagner et al., 1983). As a whole, this literature provides strong support for the construct validity of the Eriksonian stages in that better psychosocial development is significantly associated with adaptive attributes, and lower levels of psychosocial development are significantly related to maladaptive attributes.

For example, Zauszniewski and Martin (1999) tested the relationship between the first eight Eriksonian stages as measured by the MEPSI (Darling-Fisher and Leidy, 1988) and learned resourcefulness as evaluated by the Self-Control Schedule (Rosenbaum, 1980). Research participants were 60 senior adults, aged between 65 and 90 years. Results showed that 37% of the variance in learned resourcefulness was accounted for by trust, autonomy, initiative, industry, and identity, whereas the contributions of intimacy, generativity, and integrity were not statistically significant. Such results led the researchers to conclude that the mastery of the developmental tasks through adolescence may play a critical role in determining how resourceful senior adults are in their daily activities. In another study, Hughes and Peake (2002) found support for their prediction that a better resolution of the conflicts in stages 8 and 9 would be related to a healthy sense of spiritual well-being.

A better resolution of the Eriksonian stages has also been proved to be significantly negatively associated with attributes indicative of maladjustment among senior adults. For instance, Wagner et al. (1983) examined the relationship of the level of resolution of the conflict between integrity and despair to sleep behaviors among 122 institutionalized and noninstitutionalized senior adults (aged between 60 and 89 years). They found that effective resolution of this conflict tended to reduce sleep disturbances. In another study, Fishman (1992) found that higher levels of ego integrity among 115 senior adults (aged between 65 and 93 years) were significantly correlated with lower levels of death anxiety.

Research on College Students

A broad spectrum of variables has been examined as the correlates of the Eriksonian stages among the college student population. These variables can be crudely classified into three types: cognition-oriented, affect-oriented, and behavior-oriented. A good number of studies have investigated the relationships between the Eriksonian stages and cognition-oriented variables. In the United States, for instance, Njus and Johnson (2008) tested their hypothesis that psychosocial development would be positively related to the need for cognition (NFC) – a social-cognitive individual-difference variable defined by Cacioppo et al. (1996) as the desire to engage in effortful thinking. The NFC scale and the subscale of Rosental et al.'s (1981) EPSI were administered to 200 incoming college students and to about half of those students again 15 months later. At both times, students who scored higher on the NFC scale also scored higher on the EPSI identity subscale. Among both Hong Kong and mainland Chinese university students, Zhang and her collaborator (Zhang, 2010; Zhang and He, 2011) found that creativity-generating and relationally oriented intellectual styles positively contributed to psychosocial development, whereas norm-favoring and self-oriented intellectual styles did so negatively. In Germany, Busch and Hofer (2012) found that high self-regulatory capacities benefited the development of intimacy among college students.

The Eriksonian stages have been tested against a range of affect-oriented variables, including academic motivation, self-esteem, purpose in life, internal locus of control, ego strength, anxiety, depression, and the Big Five personality traits. In Canada, Faye and Sharpe (2008) concluded that psychosocial development (as represented by identity and intimacy) played a critical role in facilitating academic motivation in college. Zhang (2013) concluded that among Chinese college students, the successful resolution of particular Eriksonian crises reversely predicted neuroticism, but positively predicted extraversion, openness, agreeableness, and conscientiousness.

Finally, the Eriksonian stages have also been investigated against behavior-oriented variables, including health behaviors (Schwartz et al., 2010; Wagner et al., 1983), career-decision behaviors (Cohen et al., 1995), and interpersonal behaviors (Ávila et al., 2012). For example, using multivariate Poisson regression analysis, Schwartz et al. (2010) found that higher scores on several measures of the Eriksonian stages were

negatively related to such health-risk behaviors as binge drinking, illicit drug use, sexual risk behaviors, and risky driving among college students. Among 423 college students, Cohen and his research associates (1995) found that higher degrees of successful identity resolution were related to higher levels of career decisiveness. Ávila concluded from two waves of data collection that identity development was predicted by romantic attachment, but not by parental attachment.

Research on Identity Statuses

Much more concerted effort has been devoted to studying identity statuses than to research on the Eriksonian stages. Since the mid-1960s, an extensive body of research validating the construct of identity statuses has emerged. To date, several insightful reviews of empirical work on identity statuses have been published. This literature indicates that results from the existing work can indeed offer answers to some of the major theoretical questions centered on the nature of the construct of psychosocial development in general and on the nature of the development of identity statuses in particular. Such issues concern the process, domain, timing, pattern/direction, and stability of identity development, as well as the identification of correlates of identity statuses. In all the studies concerning the aforementioned issues, gender differences in identity development remain predominant.

The process of identity development refers to the particular identity status typical of an individual's approach to arriving at one's self-definition. The key concern here is whether or not one is disposed toward exercising sophisticated decision making, as in the case of a moratorium identity status individual or an identity achiever. The domain of identity development refers to the particular content areas in which one's self-definition is formed, including family, ethnicity, ideology, sex-role orientation, religious beliefs, and vocation. The timing of identity development concerns the specific time in one's life when identity activities are taking place in different domains. The key question regarding timing is, do different domains become more salient for different genders as a function of different points within varying contexts? The pattern/direction of identity development concerns the trajectory of each of the four identity statuses. The basic developmental assumption is that all individuals go through identity diffusion, foreclosure, moratorium, and identity achievement – in that order. Finally, a related issue is the stability (or changeability) of identity statuses. The principal question here is, is it possible for individuals to regress from a more advanced identity status to an identity status that is indicative of lack of identity maturity?

Several independent reviews of the literature (e.g., Archer, 1989a; Cramer, 2000) have all concluded that adolescents and young adults generally demonstrate increasingly more sophisticated identity activities with increasing age. They have further concluded that research strongly supports Marcia's (1980) hypothesis that identity is mainly formed between the ages of 18 and 22 years. Moreover, few gender differences exist in the process, timing, and pattern/direction of identity development. The domains in which gender differences tend to be found are those of sexuality and family roles, which are likely to hold greater salience for women than for men. Further with respect to developmental patterns, the fundamental developmental hypothesis of the identity status model assumes a decrease in diffusion and foreclosure and an increase in achievement during the course of development. However, comprehensive reviews of the literature (Archer, 1989a; Kroger et al., 2010) have consistently pointed out that, although developmental shifts are largely progressive, regressions do occur.

Kroger et al. (2010) examined the stability (or change) of developmental patterns of identity status during adolescence and young adulthood through meta-analysis of 124 studies published between 1966 and 2005. Results from 11 longitudinal studies showed that the mean proportion of adolescents making progressive identity status changes was 36%, compared with 15% who made regressive changes and 49% who remained stable. Cross-sectional studies indicated that the mean proportion of moratorium individuals rose steadily to age 19 years and declined thereafter, while the mean proportion of identity achievers rose between late adolescence and young adulthood; foreclosure and diffusion statuses declined over the high school years, but fluctuated throughout late adolescence and young adulthood.

Together, these reviews suggest that, although there is a general tendency for individuals' identity statuses to progress from diffusion to foreclosure, to moratorium, and finally to achieved identity, the trajectory of identity development is nonlinear. The exploration-commitment processes underlying the identity statuses are lifelong.

Finally, identifying the correlates of the choice of developmental pathways has been a research interest of scholars for a long time. As early as 1958, Hartmann conjectured that those who attain more adaptive identity statuses would tend to function better in educational, personal, and social domains. Such an insightful supposition has long been empirically supported. More than three decades ago, Marcia's (1980) extensive review of the identity literature showed that identity achievement (and moratorium to a lesser extent) tended to be associated with psychological variables commonly perceived to be desirable, whereas identity diffusion (and foreclosure to a lesser degree) tended to be associated with psychological variables normally deemed to be undesirable. Throughout the years, reviews of a similar nature focusing on different variables have reached the same conclusion – that identity achievement carries the most adaptive value, whereas identity diffusion is the most detrimental to related aspects of psychological functioning, learning, and development, with moratorium being the second most beneficial and foreclosure being the second most damaging (e.g., Hoegh and Bourgeois, 2002).

Conclusions, Limitations, and Implications

This article has demonstrated that there has been much theorizing and empirical testing concerning Erikson's construct of psychosocial development and its companion notion of identity status conceived by Marcia. After reading

what has been presented, one could naturally reach the conclusion that Erikson is not passé; on the contrary, Erikson is very much in fashion. More specifically, the literature reviewed should justify a firm response to each of three fundamental concerns raised by Sorell and Montgomery (2001). First, despite the historical and sociocultural context in which Erikson's theory was constructed, the theory is highly relevant today and perhaps will be for a long time into the future. Second, the popular criticism that the relationship between biological maturation and psychosocial development is merely elaborated from infancy through adolescence is not well grounded, the reason being that Erikson's theory has been shown to take a strong interactionist view of the entire human life span. From the very start, Erikson (1959) maintained that self and society must operate jointly to foster psychosocial development. Finally, although gender differences do exist in aspects of psychosocial development, this does not detract from the fact that Erikson's theory serves both genders. At the very least, empirical findings from studies of both Erikson's stages and Marcia's identity statuses so far have shown more evidence of similarities across genders than of differences.

Limitations

Notwithstanding the above, it should be recognized that scholarship concerning Erikson's theory is far from adequate relative to the value of the theory. A number of limitations exist in the current literature. Most obviously, compared with research on Marcia's identity statuses, research on Erikson's stages is far from being sufficient and systematic. The second limitation concerns the narrow range of research populations covered in almost all lines of investigation. For example, the existing research on the correlates of Erikson's stages among senior adults has, thus far, been conducted exclusively in the United States. Third, the existing research suffers from a severe lack of methodological rigor, most apparently manifested in the lack of adequate control for confounding factors that either moderate or mediate the variables under investigation, the lack of longitudinal investigations, and the lack of intervention studies. A final limitation is that very few studies reported effect sizes of their findings. As much as possible, future empirical research should aim to overcome the aforementioned and other limitations in the existing research. To substantiate the various theoretical claims, comprehensive, systematic, and well-controlled experimental longitudinal investigations must be undertaken in a wide range of populations internationally. Likewise, sound theorizing concerning the work of Erikson and Marcia is highly contingent upon an intimate internal dialogue among scholars.

Practical Implications

Regardless of the said limitations, the existing literature has practical implications for education and beyond. The theoretical argument that psychosocial development results from the interaction between individuals and their environments has long been supported by abundant empirical evidence. Given this, it is reasonable to believe that psychosocial development can be taught. If so, fostering positive psychosocial development either through direct interventions or by developing the attributes of the correlates of the psychosocial development construct (or identity statuses) would make very good sense, be that in academic or in nonacademic settings.

Appendix: Key Measures of the Eriksonian Stages and Identity Statuses

Year	*Author*	*Inventory title*
1964	Rasmussen	Ego Identity Scale
1966	Marcia	Identity Status Interview
1966	Marcia	Ego Identity – Incomplete Sentences Blank
1969	Constantinople	Eriksonian Inventory of Psychosocial Development
1970	Simmons	The Identity Achievement Scale
1980	Caillet	Ego Stage Development Inventory
1981	Rosenthal, Gurney, and Moore	Erikson Psychosocial Stage Inventory
1984	Côte	Ego Identity Scale
1984	Grotevant and Adams	The extended version of the Objective Measure of Ego Identity Status
1984, 1985	Viney and Tych	The 16 content analysis scales of psychosocial maturity
1986	Oschse and Plug	Ego Development Scale
1988	Darling-Fisher and Leidy	The modified version of Erikson Psychosocial Stage Inventory
1988	Hawley	Measures of psychosocial development
1989	Adams, Bennion, and Huh	Objective Measure of Ego-Identity Status-2
1990	Domino and Affonso	Inventory of Psychosocial Balance
1992	McAdams and de St. Aubin	The Loyola Generativity Scale
1997	Côte	Identity Stage Resolution Index
1997	Lowis and Raubenheimer	Ego Integrity Scale
1998	Adams	Objective Measure of Ego-Identity Status-2 Revised
1998	White, Wampler, and Winn	Identity Styles Inventory
2007	Schwartz	Identity Confusion Inventory
2008	Luyckx et al.	Dimensions of Identity Development Scale

See also: Friendship During Adolescence and Cultural Variations; Friendship During Childhood and Cultural

Variations; Personality Development: Systems Theories; Piaget's Theory of Cognitive Development; Vygotsky's Sociocultural Theory.

Bibliography

Archer, S.L., 1989a. The status of identity: reflections on the need for intervention. Journal of Adolescence 12 (4), 345–359.

Archer, S.L., 1989b. Gender differences in identity development: issues of process, domain and timing. Journal of Adolescence 12 (2), 117–138.

Archer, S.L., 1992. A feminist's approach to identity research. In: Gullotta, G.R., Montemayor, R. (Eds.), Adolescent Identity Formation: Advances in Adolescent Development, vol. 4. Sage, Newbury Park, CA, pp. 25–49.

Ávila, M., Cabral, J., Matos, P.M., 2012. Identity in university students: the role of parental and romantic attachment. Journal of Adolescence 35, 133–142.

Bem, S.L., 1974. The measurement of psychological androgyny. Journal of Consulting and Clinical Psychology 42, 155–162.

Brown, C., Lowis, M.J., 2003. Psychosocial development in the elderly: an investigation into Erikson's ninth stage. Journal of Aging Studies 17 (4), 415–426.

Busch, H., Hofer, J., 2012. Self-regulation and milestones of adult development: intimacy and generativity. Developmental Psychology 48 (1), 282–293.

Cacioppo, J.T., Petty, R.E., Feinstein, J., Jarvis, W., 1996. Dispositional differences in cognitive motivation: the life and times of people varying in need for cognition. Psychological Bulletin 119, 157–253.

Caillet, K.C., 1980. Ego Stage Development Inventory. Unpublished Self-report Inventory. California State University, Long Beach, CA.

Caillet, K.C., Michael, W.B., 1983. The construct validity of three self-report instruments hypothesized to measure the degree of resolution for each of the first six stage crises in Erikson's developmental theory of personality. Educational and Psychological Measurement 43, 197–209.

Capps, D., 2004. The decades of life: relocating Erikson's stages. Pastoral Psychology 53 (1), 3–32.

Cohen, C.R., Chartrand, J.M., Jowdy, D.P., 1995. Relationships between career indecision subtypes and ego identity development. Journal of Counseling Psychology 42 (4), 440–447.

Constantinople, A., 1969. An Eriksonian measure of personality development in college students. Developmental Psychology 1, 357–372.

Côté, J.E., 2002a. Commentary: focus on Sorell and Montgomery's feminist perspectives on Erikson's theory: their relevance for contemporary identity development research. Identity: An International Journal of Theory and Research 2 (3), 255–256.

Côté, J.E., 2002b. Commentary on "Feminist perspectives on Erikson's theory: their relevance for contemporary identity development research". Identity: An International Journal of Theory and Research 2 (3), 277–280.

Cramer, P., 2000. Development of identity: gender makes a difference. Journal of Research in Personality 34, 42–72.

Darling-Fisher, C.S., Leidy, N.K., 1988. Measuring Eriksonian development in adult: the modified Erikson psychosocial stage inventory. Psychological Reports 62, 747–754.

Domino, G., Affonso, D.D., 1990. A personality measure of Erikson's life stages: the inventory of psychosocial balance. Journal of Personality Assessment 54 (3 and 4), 576–588.

Dunkel, C.S., Kim, J.K., Papini, D.R., 2012. The general factor of psychosocial development and its relation to the general factor of personality and life history strategy. Personality and Individual Differences 52, 202–206.

Erikson, E.H., 1950. Childhood and Society. W.W. Norton & Company, New York, NY.

Erikson, E.H., 1951. Sex differences in the play configurations of preadolescents. American Journal of Orthopsychiatry XXI, 667–692.

Erikson, E.H., 1959. Identity and the Life Cycle. Psychological Issues Monograph, vol. I (1). International Universities Press, New York, NY.

Erikson, E.H., 1968. Identity: Youth and Crisis. W.W. Norton & Company, New York.

Erikson, E.H., 1982. The Life Cycle Completed: A Review. W.W. Norton & Company, New York.

Erikson, E.H., Erikson, J.M., 1997. The Life Cycle Completed: Extended Version. W.W. Norton & Company, New York.

Faye, C., Sharpe, D., 2008. Academic motivation in university: the role of basic psychological needs and identity formation. Canadian Journal of Behavioral Science 40 (4), 189–199.

Fishman, S., 1992. Relationships among an older adult's life review, ego integrity, and death anxiety. International Psychogeriatrics 4, 267–277.

Forrest, D.V., Côté, L.J., 2002. The mortal stage of late life. Journal of the American Academy of Psychoanalysis 30 (3), 329–340.

Freud, S., 1894. The neuro-psychoses of defence. In: Strachey, J. (Ed. and Trans.), The Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 3. Hogarth Press, London, pp. 43–68.

Gilligan, C., 1982. In a Different Voice: Psychological Theory and Women's Development. Harvard University Press, Cambridge, MA.

Gray, M.M., Ispa, J.M., Thornburg, K.R., 1986. Erikson psychosocial stage inventory: a factor analysis. Educational and Psychological Measurement 46, 979–983.

Hannah, M.T., Domino, G., Figueredo, A.J., Hendrickson, R., 1996. The prediction of ego integrity in older persons. Educational and Psychological Measurement 56 (6), 930–950.

Hannush, M.J., 2006. Book review: Hoare's "Erikson on development in adulthood: new insights from unpublished papers". Journal of Phenomenological Psychology 37 (1), 115–120.

Hartmann, H., 1958. Ego Psychology and the Problem of Adaptation. International Universities Press, New York (Original work published in 1923).

Hawley, G.A., 1988. Measures of Psychosocial Development: Professional Manual. Psychological Assessment Resources, Inc., Lutz, FL.

Hoare, C.H., 2002. Erikson on Development in Adulthood: New Insights from Unpublished Papers. Oxford University Press, New York.

Hoare, C.H., 2005. Erikson's general and adult developmental revisions of Freudian thought: "Outward, forward, upward". Journal of Adult Development 12 (1), 19–31.

Hoare, C.H., 2009. Identity and spiritual development in the papers of Erik Erikson. Identity: An International Journal of Theory and Research 9, 183–200.

Hoegh, D.G., Bourgeois, M.J., 2002. Prelude and postlude to the self: correlates of achieved identity. Youth and Society 33 (4), 573–594.

Hughes, D.E., Peake, T.H., 2002. Investigating the value of spiritual well-being and psychosocial development in mitigating senior adulthood depression. Activities, Adaptation and Aging 26 (3), 15–35.

Kroger, J., 2002. Commentary on "Feminist perspectives on Erikson's theory: their relevance for contemporary identity development research". Identity: An International Journal of Theory and Research 2 (3), 257–266.

Kroger, J., Martinussen, M., Marcia, J.E., 2010. Identity status change during adolescence and young adulthood: a meta-analysis. Journal of Adolescence 33 (5), 683–698.

Levine, C., 2002. Women, men, and persons: a response to "Feminist perspectives on Erikson's theory: their relevance for contemporary identity development research". Identity: An International Journal of Theory and Research 2 (3), 271–276.

Luyckx, K., Schwartz, S.J., Berzonsky, M.D., Soenens, B., Vansteenkiste, M., Smits, I., et al., 2008. Capturing ruminative exploration: extending the four-dimensional model of identity formation in late adolescence. Journal of Research in Personality 42, 58–82.

Marcia, J.E., 1966. Development and validation of ego identity status. Journal of Personality and Social Psychology 3, 551–558.

Marcia, J.E., 1980. Identity in adolescence. In: Adelson, J. (Ed.), Handbook of Adolescent Psychology. Wiley, New York, pp. 159–187.

Mellor, S., 1989. Gender differences in identity formation as a function of self-other relationships. Journal of Youth and Adolescence 18 (4), 361–375.

Njus, D., Johnson, D.R., 2008. Need for cognition as a predictor of psychosocial identity development. The Journal of Psychology 142, 645–655.

Rasmussen, J.E., 1964. Relationship of ego identity to psychosocial effectiveness. Psychological Reports 15, 815–825.

Rosenbaum, M., 1980. A schedule for assessing self-controlling behaviors: preliminary findings. Behavior Therapy 11, 109–121.

Rosenthal, D.A., Gurney, R.M., Moore, S.M., 1981. From trust to intimacy: a new inventory for examining Erikson's stages of psychosocial development. Journal of Youth and Adolescence 10 (6), 525–537.

Schwartz, S.J., 2001. The evolution of Eriksonian and Neo-Eriksonian identity theory and research: a review and integration. Identity: An International Journal of Theory and Research 1 (1), 7–58.

Schwartz, S.J., Forthun, L.F., Ravert, R.D., Zamboanga, B.L., Umaña-Taylor, A.J., et al., 2010. Identity consolidation and health risk behaviors in college students. American Journal of Health Behavior 34 (2), 214–224.

Schwartz, S.J., Zamboanga, B.L., Wang, W., Olthuis, J.V., 2009. Measuring identity from an Eriksonian perspective: two sides of the same coin? Journal of Personality Assessment 91 (2), 143–154.

Slater, C.L., 2003. Generativity versus stagnations: an elaboration of Erikson's adult stage of human development. Journal of Adult Development 10 (1), 53–94.

Sneed, J.R., Whitbourne, S.K., Culang, M.E., 2006. Trust, identity, and ego integrity: modeling Erikson's core stages over 34 years. Journal of Adult Development 13, 148–157.

Sorell, G.T., Montgomery, M.J., 2001. Feminist perspectives on Erikson's theory: their relevance for contemporary identity development research. Identity: An International Journal of Theory and Research 1 (2), 97–128.

Varghese, R., 1980. Factor structure of an Eriksonian measure of personality. Psychological Reports 46, 475–478.

Viney, L.L., Tych, A., 1985. Content analysis scales measuring psychosocial maturity in the elderly. Journal of Personality and Assessment 49, 311–317.

Wagner, K.D., Lorion, R.P., Shipley, E.T., 1983. Insomnia and psychosocial crisis: two studies of Erikson's developmental theory. Journal of Consulting and Clinical Psychology 51, 595–603.

Wang, W., Viney, L.L., 1996. A cross-cultural comparison of Eriksonian psychosocial development. School Psychology International 17, 33–48.

Waterman, A.S., Whitbourne, S.K., 1982. Androgyny and psychosocial development among college students and adults. Journal of Personality 50 (2), 121–133.

Zauszniewski, J.A., Martin, M.H., 1999. Developmental task achievement and learned resourcefulness in healthy older adults. Archives of Psychiatric Nursing XIII (1), 41–47.

Zhang, L.F., 2010. Further investigating thinking styles and psychosocial development in the Chinese higher education context. Learning and Individual Differences 20, 593–603.

Zhang, L.F., 2013. Psychosocial development and the big five personality traits among Chinese university students. Learning and Individual Differences 23, 163–167.

Zhang, L.F., He, Y.F., 2011. Thinking styles and the Eriksonian stages. Journal of Adult Development 18, 8–17.

Zuschlag, M.K., Whitbourne, S.K., 1994. Psychological development in three generations of college students. Journal of Youth and Adolescence 23 (5), 567–577.

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 8

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser–Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 8

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 8 *xlv*

VOLUME 8

Erotic Capital 1
Catherine Hakim

Errors in Statistical Assumptions, Effects of 8
Mark J Schervish

Essentialism 13
Nick Haslam and Maria Abou-Abdallah

Estimation: Point and Interval 16
Roger L Berger

Ethical Codes, Professional: Business Codes 21
Josef Wieland

Ethical Consumers 25
Veronika A Andorfer

Ethical Dilemmas, Research, and Access to Treatment 31
Dirceu B Greco

Ethical Guidelines and Codes of Conduct in Social and Behavioral Research 37
Thomas L Van Valey

Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute 43
Jean E McEwen, Joy T Boyer, Nicole C Lockhart, David J Kaufman, and Lawrence C Brody

Ethical Practices: Institutional Oversight and Enforcement 48
Sophia F Dziegielewski

Ethical Questions in Social and Behavioral Sciences, History of 53
Andrea Ploder

Ethical Turn in International Development 60
Eleanor Fisher

Ethics and Psychiatry 63
Allen R Dyer and Eindra K Khin

Ethics and Values 71
Peter Railton and Alex Silk

Ethics for Biomedical Research Involving Humans: International Guidelines 78
John R Williams

Ethnic Cleansing, History of 83
Liliana Riga

Ethnic Conflict, Geography of 91
Frederick W Boal

Ethnic Conflicts 95
John Coakley

Ethnic Identity and Ethnicity in Archaeology 102
Margarita Díaz-Andreu

Ethnic Identity and Work 106
Amelie F Constant

Ethnic Identity, Psychology of 113
Karmela Liebkind, Tuuli A Mähönen, Sirkku Varjonen, and Inga Jasinskaja-Lahti

Ethnic Intermarriage 118
Delia Furtado

Ethnic, Racial, and Nationalist Movements 123
Susan Olzak

Ethnicity and Educational Achievement 130
Fang Gao and Gerard A Postiglione

Ethnicity and Ethnic Groups: Historical Aspects 136
Mathias Bös

Ethnicity and Migration in Europe 142
Karen Phalet, Gülseli Baysu, and Kaat Van Acker

Ethnicity: Anthropological Aspects 148
Richard Jenkins

Ethnicity, Race, Oppression, and Social Work: The Canadian Case 152
Jennifer Clarke, Gordon Pon, Akua Benjamin, and Annette Bailey

Ethnicity, Sociology of 157
Eliezer Ben-Rafael and Yistzhak Sternberg

Ethnoarchaeology 162
Margaret Beck

Ethnocentrism 166
RA LeVine

Ethnocentrism and Prejudice: History of the Concepts 168
Boris Bizumic

Ethnogenesis in Anthropology 175
Michiyuki Matsuda

Ethnography 178
Antonius CGM Robben and Jeffrey A Sluka

Ethnography in Applied Social Research 184
David M Fetterman

Ethnohistory 192
Pauline T Strong

Ethnology 198
Gisela Welz

Ethnomethodology, General 203
Steven E Clayman

Etymology 207
Steven N Dworkin

Eudaemonism 212
Heidrun Friese

Eugenics as a Basis of Population Policy 218
Garland E Allen and Marius Turda

Eugenics as an International Movement 224
Garland E Allen

Eugenics in Asia 233
Frank Dikötter

Eugenics in Europe 237
Richard S Fogarty

Eugenics in the Americas 246
Patience A Schell

Eugenics, History of 253
Diane B Paul and Marius Turda

European Religion from the Atlantic to the Black Sea 258
Grace Davie

European Union Law 264
Jürgen Basedow

Europe, Archaeology of 269
Peter F Biehl

Europe: Sociocultural Overview 273
Christian Giordano

Euthanasia 280
Dieter Birnbacher

Evangelical, Revivalist, and Pentecostal Religions 285
Mark Hutchinson

Evans-Pritchard, Sir Edward E. (1902–73) 289
Wendy James

Event History Analysis 295
Richard J Cook and Jerald F Lawless

Event History Analysis: Applications 305
Hans-Peter Blossfeld

Event History Analysis in Continuous Time 310
Per K Andersen and Niels Keiding

Event Sequence Analysis: Methods 320
Andrew Abbott

Everyday Life, Anthropology of 323
Orvar Löfgren

Evidence-Based Medicine 327
Matthias Egger

Evidence-Based Practice 332
Jeremy Steglitz, Jennifer L Warnick, Sara A Hoffman, Winter Johnston, and Bonnie Spring

Evidence-Based Psychiatry 339
Megan J Hobbs

Evidence-Based Psychotherapy 344
Andreas Beelmann and Nina Heinrichs

Evidentials in Linguistics 351
Zlatka Guentchéva

Evolution and Language: Cultural Transmission 357
Bill Thompson and Kenny Smith

Evolution and Language: Overview 364
William Croft

Evolution and Language: Phylogenetic Analyses 370
Simon Greenhill

Evolution: Diffusion of Innovations 378
Everett M Rogers

Evolution, History of 382
Peter J Bowler

Evolution, Natural and Social: Philosophical Aspects 388
Brian Skyrms

Evolutionary Approach to Education Science 392
David C Geary

Evolutionary Approaches to Human Behavior: Philosophical Aspects 399
Ben Fraser and Kim Sterelny

Evolutionary Demography: A Darwinian Renaissance in Demography 406
Rebecca Sear

Evolutionary Demography: A Synthesis of Two Population Sciences 413
Daniel A Levitis

Evolutionary Developmental Psychology 420
Karin Machluf and David F Bjorklund

Evolutionary Economics 430
Christian Cordes

Evolutionary Epistemology 437
David L Hull

Evolutionary Group Dynamics 441
Tatsuya Kameda, Mark Van Vugt, and Robert Scott Tindale

Evolutionary Medicine 448
Alexandra Alvergne

Evolutionary Social Psychology 460
Douglas T Kenrick

Evolutionary Sociology 463
Rosemary L Hopcroft

Evolutionary Theory, Structure of 469
Elisabeth Lloyd

Exchange in Anthropology 476
Gustav Peebles

Exchange: Social 482
Karen S Cook

Executive Functions during Childhood, Development of 489
Sanja Šimleša and Maja Cepanec

Executive Government 497
Wolfgang C Müller

Expectancy-Value-Cost Model of Motivation 503
Kenneth E Barron and Chris S Hulleman

Expectations, Economics of 510
George W Evans and Seppo Honkapohja

Experiential Psychotherapy 517
Leslie S Greenberg

Experimental Design: Bayesian Designs 521
Merlise A Clyde

Experimental Design: Compliance 527
Alix I Gitelman

Experimental Design: Large-Scale Social Experimentation 531
Carl N Morris

Experimental Design: Overview 535
Angela M Dean

Experimental Design: Randomization and Social Experiments 541
Robert F Boruch and Benjamin Brumley

Experimental Economics 546
Gigi Foster

Experimental Laboratories: Psychological 552
Daniel Greenberg

Experimental Laboratories: Social and Economic 556
Richard Gonzalez

Experimental Methods in Survey Research in Demography 559
Duncan Thomas and Elizabeth Frankenberg

Experimentation in Psychology, History of 566
GCG Trudy Dehue

Experimenter and Subject Artifacts: Methodology 571
Ralph L Rosnow and David B Strohmetz

Experiment in Science and Technology Studies 576
Peter Dear

Expert Testimony 582
Margaret B Kovera and Jacqueline L Austin

Expertise, Anthropological Approaches to 588
Dominic Boyer

Explanation: Conceptions in the Social Sciences 593
Hans Westmeyer

Explanatory Style and Health 598
Athanassios Douzenis and Dionysis Seretis

Exploratory Data Analysis: Univariate Methods 600
David C Hoaglin

Exposure Therapies and Stress Inoculation: A Brief Overview 605
David Phillips and Dean Lauterbach

Extended Schools 612
Alan Dyson

External Validity 617
Laura C Leviton

Extraversion 623
Caroline Fielden, Lisa E. Kim, and Carolyn MacCann

Eye Movement, Control of: Oculomotor Control 628
Bryan Kolb

Eye Movements in Reading 631
Keith Rayner

Eyeblink Classical Conditioning 635
Kevin L Brown and John H Freeman

Eyewitness Testimony 642
Robert A Nash, Maciej Hanczakowski, and Giuliana Mazzoni

Face Perception Development During Early Childhood 650
Gudrun Schwarzer and Claudia Freitag

Face Recognition Models: Computational Approaches 658
Zhen Lei and Stan Z Li

Face Recognition, Psychological and Neural Aspects 663
Aina Puce

Facial Emotion Expression, Individual Differences in 667
Andrea Hildebrandt, Sally Olderbak, and Oliver Wilhelm

Facial Perception 676
Oliver Wilhelm, Andrea Hildebrandt, and Werner Sommer

Factionalism 683
Jeremy Boissevain

Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis 686
Ralph O Mueller and Gregory R Hancock

Factor Analysis and Latent Structure Analysis: Overview 691
David J Bartholomew

Factor Analysis and Latent Structure: IRT and Rasch Models 698
Brian W Junker

Factor Analysis and Latent Variable Models in Personality Psychology 703
Ralf Schulze, Anna-Lena Hilger, and Philipp M Engelberg

False Memories 709
Rebecca M Nichols, Daniel Bogart, and Elizabeth F Loftus

Familial Studies: Genetic Inferences 715
Vincent P Diego, Jack W Kent, Jr, and John Blangero

Families and Households, Formal Demography of 725
Frans Willekens and Evert van Imhoff

Families as Educational Settings 731
Beth Kurtz-Costes

Family and Kinship, History of 738
Tamara K Hareven

Family and Psychiatric Illness: Family Therapy 745
Judith Lask

Family and Schooling 751
Thomas Kellaghan

Family and Work 758
Marisa Young

Family as Institution 766
Linda J Waite

Family Business 772
Guido Corbetta

Family Law in the United States 779
Martha A Fineman and June Carbone

Family Planning: Program Effects 786
Grant Miller and Kimberly S Babiarz

Family Planning Programs: Feminist Perspectives 794
Carmen Barroso

Family Policy 799
Mary Daly

Family Size Preferences 805
Elizabeth Thomson

Family: The Anthropology of the Concept and Its History 809
Jean La Fontaine

Family Theory: Competing Perspectives in Social Demography 815
Susan L Brown

Family Theory: Economics of Childbearing 821
John F Ermisch

Family Theory: Economics of Intergenerational Relations 826
Alessandro Cigno

Family Theory: Economics of Marriage and Divorce 829
Yoram Weiss

Fashion, Sociology of 835
Sellerberg Ann-Mari and Aspers Patrik

Fatherhood 840
Sandra Hofferth and Frances Goldscheider

Fathers and Fathering: Social Work Perspectives 844
Jonathan Scourfield

Fear Conditioning: Overview 849
Herta Flor and Niels Birbaumer

Fear of Crime 854
Jodi Lane

Fear-Potentiated Startle 860
Alfons O Hamm

Fear: Psychological and Neural Aspects 868
Blake A Pellman and Jeansok J Kim

Federalism 875
Arthur Burris

Female Genital Mutilation 878
Aud Talle

Female Power at Work 883
David J Maume and Leah Ruppanner

Female Quotas in Parliament: Do They Make a Difference? 888
Mona L Krook

Femininities 893
Elroi J Windsor

Feminism in Anthropology 898
Louise Lamphere

Feminisms and Acculturation around the Globe 905
Akosua Adomako Ampofo, Edwin Asa Adjei, and Maame Kyerewaa Brobbey

Feminist Criminology and Gender Studies 912
Susan F Sharp

Feminist Legal Theory 918
Genevieve R Painter

Feminist Political Theory and Political Science 926
Joni Lovenduski

Feminist Psychotherapy 929
Carolyn Zerbe Enns

Feminist Research and Social Work 935
Ruth Phillips

Feminist Theory: Radical Lesbian 942
Wendy L Lee

This page intentionally left blank

CONTRIBUTORS TO VOLUME 8

Andrew Abbott
University of Chicago, Chicago, IL, USA

Maria Abou-Abdallah
University of Melbourne, Melbourne, VIC, Australia

Garland E. Allen
Washington University in St. Louis, St Louis, MO, USA

Alexandra Alvergne
University of Oxford, Oxford, UK

Akosua Adomako Ampofo
University of Ghana, Legon, Ghana

Per K. Andersen
University of Copenhagen, Copenhagen, Denmark

Veronika A. Andorfer
Göttingen, Germany

Sellerberg Ann-Mari
Lund University, Lund, Sweden

Edwin Asa Adjei
University of Ghana, Legon, Ghana

Jacqueline L. Austin
John Jay College of Criminal Justice, CUNY, New York, NY, USA

Kimberly S. Babiarz
Stanford University, Stanford, CA, USA

Annette Bailey
Ryerson University, Toronto, ON, Canada

Kenneth E. Barron
James Madison University, Harrisonburg, VA, USA

Carmen Barroso
International Planned Parenthood Federation/Western Hemisphere Region, New York, NY, USA

David J. Bartholomew
London School of Economics, London, UK

Jürgen Basedow
Max-Planck-Institute for Comparative and International Private Law, Hamburg, Germany

Gülseli Baysu
Kadir Has University, İstanbul, Turkey

Margaret Beck
University of Iowa, Iowa City, IA, USA

Andreas Beelmann
Friedrich-Schiller-University Jena, Jena, Germany

Eliezer Ben-Rafael
Tel-Aviv University, Ramat Hasharon, Israel

Akua Benjamin
Ryerson University, Toronto, ON, Canada

Roger L. Berger[†]
Arizona State University, Phoenix, AZ, USA

Peter F. Biehl
SUNY Buffalo, Buffalo, NY, USA

Niels Birbaumer
University of Tübingen, Tübingen, Germany

Dieter Birnbacher
Institut für Philosophie, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany

Boris Bizumic
ANU College of Medicine, Biology and Environment, The Australian National University, Canberra, ACT, Australia

David F. Bjorklund
Florida Atlantic University, Boca Raton, FL, USA

John Blangero
Texas Biomedical Research Institute, San Antonio, TX, USA

Hans-Peter Blossfeld
University of Bielefeld, Bielefeld, Germany

[†] Deceased.

Frederick W. Boal
Queen's University Belfast, Belfast, Northern Ireland, UK

Daniel Bogart
University of California, Irvine, CA, USA

Jeremy Boissevain
University of Amsterdam, Amsterdam, The Netherlands

Robert F. Boruch
Graduate School of Education, University of Pennsylvania, Philadelphia, PA, USA

Mathias Bös
Leibniz Universität Hannover, Hannover, Germany

Peter J. Bowler
Queen's University Belfast, Belfast, UK

Dominic Boyer
Rice University, Houston, TX, USA

Joy T. Boyer
National Human Genome Research Institute, Bethesda, MD, USA

Lawrence C. Brody
National Human Genome Research Institute, Bethesda, MD, USA

Kevin L. Brown
The University of Iowa, Iowa City, IA, USA

Susan L. Brown
Bowling Green State University, Bowling Green, OH, USA

Benjamin Brumley
University of Pennsylvania, Philadelphia, PA, USA

Arthur Burris
California State University, Hayward, CA, USA

June Carbone
University of Minnesota Law School, Minneapolis, MN, USA

Maja Cepanec
University of Zagreb, Zagreb, Croatia

Alessandro Cigno
University of Florence, Firenze, Italy

Jennifer Clarke
Ryerson University, Toronto, ON, Canada

Steven E. Clayman
University of California, Los Angeles, CA, USA

Merlise A. Clyde
Duke University, Durham, NC, UK

John Coakley
Queen's University, Belfast, Northern Ireland

Amelie F. Constant
George Washington University, Washington, DC, USA; Temple University, Philadelphia, PA, USA; and IZA – Institute for the Study of Labor, Bonn, Germany

Karen S. Cook
Stanford University, Stanford, CA, USA

Richard J. Cook
University of Waterloo, Waterloo, ON, Canada

Guido Corbetta
Bocconi University, Milan, Italy

Christian Cordes
University of Bremen, Bremen, Germany

William Croft
University of New Mexico, Albuquerque, NM, USA

Mary Daly
University of Oxford, Oxford, UK

Grace Davie
University of Exeter, Exeter, UK

Angela M. Dean
The Ohio State University, Columbus, OH, USA

Peter Dear
Cornell University, Ithaca, NY, USA

G.C.G. Trudy Dehue
University of Groningen, Groningen, The Netherlands

Margarita Díaz-Andreu
ICREA-Universitat de Barcelona, Barcelona, Spain

Vincent P. Diego
Texas Biomedical Research Institute, San Antonio, TX, USA

Frank Dikötter
The University of Hong Kong, Hong Kong, People's Republic of China

Athanassios Douzenis
Attikon University Hospital, Athens University Medical School, Athens, Greece

Steven N. Dworkin
University of Michigan, Ann Arbor, MI, USA

Allen R. Dyer
The George Washington University, Washington, DC, USA

Alan Dyson
University of Manchester, Manchester, England

Sophia F. Dziegielewski
College of Health and Public Affairs, University of Central Florida (UCF), Orlando, FL, USA; and Merritt Island, FL, USA

Matthias Egger
University of Bern, Bern, Switzerland

Philipp M. Engelberg
Bergische Universität Wuppertal, Wuppertal, Germany

Carolyn Zerbe Enns
Cornell College, Mount Vernon, IA, USA

John F. Ermisch
University of Oxford, Oxford, UK

George W. Evans
University of Oregon, Eugene, OR, USA; and University of St. Andrews, Fife, UK

David M. Fetterman
San Jose, CA, USA

Caroline Fielden
University of Sydney, Sydney, NSW, Australia

Martha A. Fineman
Emory University — School of Law, Atlanta, GA, USA

Eleanor Fisher
University of Reading, Reading, UK

Herta Flor
Heidelberg University, Mannheim, Germany

Richard S. Fogarty
University at Albany, SUNY, Albany, NY, USA

Gigi Foster
University of New South Wales, Sydney, NSW, Australia

Elizabeth Frankenberg
Duke University, Durham, NC, USA

Ben Fraser
Australian National University, Canberra, ACT, Australia

John H. Freeman
The University of Iowa, Iowa City, IA, USA

Claudia Freitag
University of Giessen, Giessen, Germany

Heidrun Friese
Intercultural Communication, Chemnitz, Germany

Delia Furtado
University of Connecticut, Storrs, CT, USA

Fang Gao
Hong Kong Baptist University, Hong Kong, People's Republic of China

David C. Geary
University of Missouri at Columbia, Columbia, MO, USA

Christian Giordano
Université de Fribourg, Fribourg, Switzerland

Alix I. Gitelman
Oregon State University, Corvallis, OR, USA

Frances Goldscheider
University of Maryland, College Park, MD, USA

Richard Gonzalez
University of Michigan, Ann Arbor, MI, USA

Dirceu B. Greco
Federal University of Minas Gerais, Belo Horizonte, Brazil

Daniel Greenberg
College of Charleston, Charleston, SC, USA

Leslie S. Greenberg
York University, Toronto, ON, Canada

Simon Greenhill
Australian National University, Canberra, ACT, Australia

Zlatka Guentchéva
LACITO — CNRS (Centre National de la Recherche Scientifique), Paris, France

Catherine Hakim
Civitas, London, UK

Alfons O. Hamm
University of Greifswald, Greifswald, Germany

Gregory R. Hancock
University of Maryland, College Park, MD, USA

Maciej Hanczakowski
Cardiff University, Cardiff, UK

Tamara K. Hareven
University of Delaware, Newark, DE, USA

Nick Haslam
University of Melbourne, Melbourne, VIC, Australia

Nina Heinrichs
University of Braunschweig, Braunschweig, Germany

Andrea Hildebrandt
Ernst-Moritz-Arndt University Greifswald, Greifswald, Germany

Anna-Lena Hilger
Bergische Universität Wuppertal, Wuppertal, Germany

David C. Hoaglin
Sudbury, MA, USA

Megan J. Hobbs
Clinical Research Unit for Anxiety and Depression, School of Psychiatry, University of New South Wales, Sydney, NSW, Australia

Sandra Hofferth
University of Maryland, College Park, MD, USA

Sara A. Hoffman
Northwestern University, Chicago, IL, USA

Seppo Honkapohja
Bank of Finland, Helsinki, Finland

Rosemary L. Hopcroft
University of North Carolina-Charlotte, Charlotte, NC, USA

David L. Hull[†]
Northwestern University, Evanston, IL, USA

Chris S. Hulleman
University of Virginia, Charlottesville, VA, USA

Mark Hutchinson
Southern Cross College, Chester Hill, NSW, Australia

Wendy James
Oxford University, Oxford, UK

Inga Jasinskaja-Lahti
University of Helsinki, Helsinki, Finland

Richard Jenkins
University of Sheffield, Sheffield, UK

Winter Johnston
Northwestern University, Chicago, IL, USA

Brian W. Junker
Carnegie Mellon University, Pittsburgh, PA, USA

Tatsuya Kameda
Hokkaido University, Hokkaido, Japan

David J. Kaufman
National Human Genome Research Institute, Bethesda, MD, USA

Niels Keiding
University of Copenhagen, Copenhagen, Denmark

Thomas Kellaghan
St Patrick's College, Dublin, Ireland

Douglas T. Kenrick
Arizona State University, Tempe, AZ, USA

Jack W. Kent
Texas Biomedical Research Institute, San Antonio, TX, USA

Eindra K. Khin
The George Washington University, Washington, DC, USA

Jeansok J. Kim
University of Washington, Seattle, WA, USA

Lisa E. Kim
University of Sydney, Sydney, NSW, Australia

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada

Margaret B. Kovera
John Jay College of Criminal Justice, CUNY, New York, NY, USA

Mona L. Krook
Rutgers University, New Brunswick, NJ, USA

Beth Kurtz-Costes
University of North Carolina, Chapel Hill, NC, USA

Maame Kyerewaa Brobbey
University of Ghana, Legon, Ghana

Jean La Fontaine
University of London, London, UK

Louise Lamphere
University of New Mexico, Albuquerque, NM, USA

Jodi Lane
University of Florida, Gainesville, FL, USA

Judith Lask
Kings College London, London, UK; and United Kingdom Council for Psychotherapy, London, UK

Dean Lauterbach
Eastern Michigan University, Ypsilanti, MI, USA

Jerald F. Lawless
University of Waterloo, Waterloo, ON, Canada

Wendy L. Lee
Bloomsburg University of Pennsylvania, Bloomsburg, PA, USA

[†] Deceased.

Zhen Lei
Chinese Academy of Sciences, Beijing, China

R.A. LeVine
Harvard University, Cambridge, MA, USA

Daniel A. Levitis
Max-Planck Odense Center on the Biodemography of Aging, University of Southern Denmark, Odense, Denmark

Laura C. Leviton
The Robert Wood Johnson Foundation, Princeton, NJ, USA

Stan Z. Li
Chinese Academy of Sciences, Beijing, China

Karmela Liebkind
University of Helsinki, Helsinki, Finland

Elisabeth Lloyd
Indiana University, Bloomington, IN, USA

Nicole C. Lockhart
National Human Genome Research Institute, Bethesda, MD, USA

Orvar Löfgren
University of Lund, Lund, Sweden

Elizabeth F. Loftus
University of California, Irvine, CA, USA

Joni Lovenduski
Birkbeck College, London, UK

Carolyn MacCann
University of Sydney, Sydney, NSW, Australia

Karin Machluf
Florida Atlantic University, Boca Raton, FL, USA

Tuuli A. Mähönen
University of Helsinki, Helsinki, Finland

Michiyuki Matsuda
Kyoto University, Kyoto, Japan

David J. Maume
University of Cincinnati, Cincinnati, OH, USA

Giuliana Mazzoni
University of Hull, Hull, UK

Jean E. McEwen
National Human Genome Research Institute, Bethesda, MD, USA

Grant Miller
Stanford University, Stanford, CA, USA

Carl N. Morris
Harvard University, Cambridge, MA, USA

Ralph O. Mueller
University of Hartford, West Hartford, CT, USA

Wolfgang C. Müller
University of Vienna, Vienna, Austria

Robert A. Nash
University of Surrey, Guildford, UK

Rebecca M. Nichols
University of California, Irvine, CA, USA

Sally Olderbak
Ulm University, Ulm, Germany

Susan Olzak
Stanford University, Stanford, CA, USA

Genevieve R. Painter
University of California, Berkeley, CA, USA

Aspers Patrik
Uppsala University, Uppsala, Sweden

Diane B. Paul
Harvard University, Cambridge, MA, USA

Gustav Peebles
The New School, New York, NY, USA

Blake A. Pellman
University of Washington, Seattle, WA, USA

Karen Phalet
Center for Social and Cultural Psychology, Leuven, Belgium

David Phillips
Eastern Michigan University, Ypsilanti, MI, USA

Ruth Phillips
The University of Sydney, Sydney, NSW, Australia

Andrea Ploder
University of Graz, Graz, Austria

Gordon Pon
Ryerson University, Toronto, ON, Canada

Gerard A. Postiglione
The University of Hong Kong, Hong Kong, People's Republic of China

Aina Puce
Indiana University, Bloomington, IN, USA

Peter Railton
University of Michigan, Ann Arbor, MI, USA

Keith Rayner
University of California, San Diego, La Jolla, CA, USA

Liliana Riga
University of Edinburgh, Edinburgh, Scotland, UK

Antonius C.G.M. Robben
Utrecht University, Utrecht, The Netherlands

Everett M. Rogers†
University of New Mexico, Albuquerque, NM, USA

Ralph L. Rosnow
Temple University, Philadelphia, PA, USA

Leah Ruppanner
University of Melbourne, Parkville, VIC, Australia

Patience A. Schell
School of Language and Literature, King's College, University of Aberdeen, Aberdeen, Scotland, UK

Mark J. Schervish
Carnegie Mellon University, Pittsburgh, PA, USA

Ralf Schulze
Bergische Universität Wuppertal, Wuppertal, Germany

Gudrun Schwarzer
University of Giessen, Giessen, Germany

Jonathan Scourfield
Cardiff University, Cardiff, Wales, UK

Rebecca Sear
London School of Hygiene and Tropical Medicine, London, UK

Dionysis Seretis
Attikon University Hospital, Athens University Medical School, Athens, Greece

Susan F. Sharp
University of Oklahoma, Norman, OK, USA

Alex Silk
University of Birmingham, Birmingham, UK

Sanja Šimleša
University of Zagreb, Zagreb, Croatia

Brian Skyrms
University of California, Irvine, CA, USA; and Stanford University, Stanford, CA, USA

Jeffrey A. Sluka
Massey University, Palmerston North, New Zealand

Kenny Smith
The University of Edinburgh, Edinburgh, UK

Werner Sommer
Humboldt-Universität zu Berlin, Berlin, Germany

Bonnie Spring
Northwestern University, Chicago, IL, USA

Jeremy Steglitz
Northwestern University, Chicago, IL, USA

Kim Sterelny
Australian National University, Canberra, ACT, Australia

Yistzhak Sternberg
Beit Berl Academic College and Open University of Israel, Tel-Aviv, Israel

David B. Strohmetz
Monmouth University, West Long Branch, NJ, USA

Pauline T. Strong
The University of Texas at Austin, Austin, TX, USA

Aud Talle
University of Oslo, Oslo, Norway

Duncan Thomas
Duke University, Durham, NC, USA

Bill Thompson
The University of Edinburgh, Edinburgh, UK

Elizabeth Thomson
Stockholm University, Stockholm, Sweden; and University of Wisconsin—Madison, Madison, WI, USA

Robert Scott Tindale
Loyola University Chicago, Chicago, IL, USA

Marius Turda
Oxford Brookes University, Oxford, UK

Kaat Van Acker
Max Planck Institute for the Study of Religious and Ethnic Diversity, Göttingen, Germany

Evert van Imhoff†

Thomas L. Van Valey
Western Michigan University, Kalamazoo, MI, USA

Mark Van Vugt
VU University Amsterdam, Amsterdam, The Netherlands

Sirkku Varjonen
University of Helsinki, Helsinki, Finland

† Deceased.

Linda J. Waite
University of Chicago, Chicago, IL, USA

Jennifer L. Warnick
Northwestern University, Chicago, IL, USA

Yoram Weiss
University of Tel-Aviv, Tel-Aviv, Israel

Gisela Welz
Goethe University, Frankfurt, Germany

Hans Westmeyer
Free University of Berlin, Berlin, Germany

Josef Wieland
Zeppelin University, Friedrichshafen, Germany

Oliver Wilhelm
University Ulm, Ulm, Germany

Frans Willekens
Max Planck Institute for Demographic Research, Rostock, Germany; and Netherlands Interdisciplinary Demographic Institute (NIDI), The Hague, The Netherlands

John R. Williams
Ottawa, ON, Canada

Elroi J. Windsor
Salem College, Winston-Salem, NC, USA

Marisa Young
McMaster University, Hamilton, ON, Canada

This page intentionally left blank

Erotic Capital

Catherine Hakim, Civitas, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Erotic capital is a combination of physical and social attractiveness. It is the fourth personal asset, alongside economic, social, and human capital. Erotic capital delivers social benefits throughout life, illustrated by gains of 15% extra votes for attractive candidates in elections. It has economic value, delivering earnings mark-ups of between 10 and 20% on average (17% for men and 12% for women). Attractive people gain more friends, lovers, spouses, and sexual experience. They are more persuasive, can sell more goods, services, and policies, are more likely to be perceived as honest and competent, and are treated accordingly. Social, economic, and technological changes are driving the rising social salience and economic value of erotic capital in the twenty-first century.

Erotic capital is a new concept in social science, a label for the personal asset of physical and social attractiveness that has increasing importance in the twenty-first century, and that offers substantial social benefits and economic value to those who choose to develop and exploit it.

Bourdieu's well-known classification of economic, social, and cultural (or human) capital developed for analyses of twentieth-century society is now expanded to include erotic capital as the fourth personal asset in the twenty-first century. Like human/cultural capital and social capital, erotic capital can be exchanged for money and vice versa. Unlike cultural/human and social capital, erotic capital is more randomly distributed by genetic good fortune rather than predicted by parental wealth. Beautiful children can be born into the poorest families, giving them opportunities for upward social mobility. Wealthy parents may have ugly offspring. This socially unpredictable and uncontrollable character of physical attractiveness gives it a 'dangerous' maverick character that has led to it being belittled as superficial and worthless in some cultures. But as Webster and Driskell (1983) recognized in an early paper, beauty confers status, so sociological theory cannot ignore it. New research shows that the economic value of erotic capital is now equal to the value of human capital in affluent societies – so this is not a trivial factor.

By the first decade of the twenty-first century, several writers working independently on sexuality recognized, around the same time, that physical attractiveness and sex appeal have serious value in sexual markets. Seeking a more theoretical term for sex appeal, they adopted various labels: sexual capital (Martin and George, 2006), erotic capital (Green, 2008), and erotic power (Hakim, 2012). These scholars were studying sexual markets, both heterosexual and gay, so the term *sexual capital* is more appropriate for this narrow context. The term *erotic capital* was adopted by Hakim, again working independently, to denote a broader combination of physical and social attractiveness that has value in all spheres of public and private life, including the labor market and politics. Hakim presented the most fully developed concept and theory of erotic capital, synthesized the widest range of research evidence, and discussed the problems of measurement and research methodology for this personal asset (Hakim, 2010, 2011). It is her theory that informs this review essay and that has been most widely taken up by other social scientists.

Defining Erotic Capital

Erotic capital is multifaceted. Beauty is clearly a central element, although there are cultural and temporal variations in ideas about what constitutes beauty. In Papua New Guinea, it is men who paint their faces with dazzlingly colorful patterns, but in the West it is the women who paint their faces. Some African societies admire women with large voluptuous bodies. In Western Europe, fashion models can be tall and stick-thin to the point of being anorexic. In previous centuries, women with small eyes and tiny rosebud mouths were considered delicately beautiful. The modern emphasis on photogenic features means that men and women with large eyes and mouths and 'sculptured' faces are prioritized. Studies show that symmetry, an even skin tone, and conventionality contribute to attractiveness. It is, in part, an achieved characteristic, as illustrated by the French concept of the *belle laide* or *beau laid*, an ugly woman or man who becomes attractive through presentational skills. Great beauty is always in short supply and is therefore universally valorized.

A second element is sexual attractiveness, which can be quite separate from classic beauty. To some extent, beauty is about facial attractiveness, while sex appeal is about a sexy body. However, sex appeal can also stem from personality and style, expressions of femininity or masculinity, and a characteristic of social interaction. Beauty tends to be static, hence is easily captured in a photo. Sexual attractiveness is about the way someone moves, talks, and behaves, so it can only be captured on film. Sex appeal is often linked to youth and can fade rapidly with age. The variation in personal tastes also means there is no universal ideal. Some men prefer women who are small and petite, while others prefer tall and elegant women. Some women prefer men with well-developed muscles and strong athletic bodies, while others prefer a slender, elegant appearance. Despite this variation in tastes, sex appeal is in short supply and is therefore valorized.

A third element of erotic capital is social: grace, charm, social skills in interaction, the ability to make people like you, feel at ease and happy, want to know you and, perhaps, desire you. Flirtatious skills can be learnt, but again are not universal. Some people in positions of power ooze charm and charisma; others have none at all. Some men and women are skilled at discreet flirtation in all contexts; others are incapable.

A fourth element is liveliness, a mixture of physical fitness, energy, and good humor. People who have a lot of life in them can be hugely attractive to others, as illustrated by people who are 'the life and soul of the party.' In most cultures, liveliness is displayed in dancing skills or sporting activities.

The fifth element concerns techniques of social presentation: style of dress, face-painting, tattoos, perfume, jewelry or other adornments, hairstyles, and the various accessories that people carry or wear to announce their social status and style to the world. Heads of state dress for public functions to emphasize their power and authority. Ordinary people going to a party or other social event dress to make themselves attractive as well as to announce their social status and wealth to any strangers they meet or to make style statements. The relative emphasis on sexy attire or social status statements depends on the venue and event. People who are skilled at social presentation appropriate to the social event are more attractive than people who look like homeless vagrants.

The sixth element is sexuality itself: sexual competence, erotic imagination, playfulness, and everything else that makes for a sexually satisfying partner. Whether or not someone is a good lover is known only to their partners, and national sex surveys provide no information on levels of sexual competence. It seems reasonable to assume that sexual skill is not a universal attribute, even among adults, and extreme competence is a minority asset. This last factor usually applies only in private, intimate relationships, whereas the first five exert their influence in all social interaction, visibly or invisibly, at the conscious or (more often) subconscious level.

For men as well as women, all six elements contribute to someone's erotic capital, although one or two elements may dominate. The value of erotic capital can also depend on someone's occupation, which may or may not valorize it. Erotic capital is generally of most value in contexts where there is an audience, where there is a need to persuade or sell, and where people are on public display in some way. The social and economic value of erotic capital is highlighted in occupations in the entertainment and hospitality industries, but is also significant in professional and managerial occupations. Erotic capital helps to sell products, services, ideas, and policies.

Erotic capital is thus a combination of aesthetic, visual, physical, social, and sexual attractiveness to other members of the society, and especially to members of the opposite sex, in all social contexts. In some cultures, fertility would be an additional element of women's greater erotic capital. Erotic capital is partly genetic but is not fixed at birth. Almost every element requires skills that can be learned and developed. The most obvious example is people who change the color of their hair to stand out and be more attractive, for example, by changing from brown to blonde. An ugly person can become attractive through the use of flattering hairstyles, cosmetics, physical fitness, flattering colors and styles of clothes, cosmetic surgery, and developing charming manners. Women generally have more of it than men, even in cultures where fertility is not an integral element, and they deploy it more actively.

Studies show that physical and social attractiveness are too closely interlinked to be disentangled, even though most studies try to tease out the impact of each of the six elements separately. The most notable finding from an ingenious study in Argentina was that attractiveness had the same impact on social interaction whether employers met workers face-to-face, saw only a photo of the worker, or never saw anyone at all and only spoke to people on the phone. As usual, attractiveness raised assessments of ability, success, and earnings by between 13 and 17% (Mobius and Rosenblat, 2006; Hakim, 2011: pp. 170–171).

The Increasing Importance of Erotic Capital in the Twenty-First Century

Erotic capital is of growing importance in the twenty-first century, and in knowledge economies, for three reasons.

First, the workplace has changed, so social skills and charm are crucial now in many jobs in modern economies. The shift from manual jobs in manufacturing industries to white-collar and service sector work makes erotic capital more important because good social relationships, and smooth social interaction, are crucial in the workplace. For example, even in academia, where the focus is on specialist knowledge, professors who are attractive are judged more competent by their students (Hamermesh, 2011).

Second, as people and societies become more affluent, they spend more on luxuries, including beauty, courtesy, charming good manners, nice clothes, pleasant service, and good looks. People who offer erotic capital in addition to any other talents will thus be more successful.

Third, we are all more visible now due to digital photography and the Internet. Even private individuals become exposed through their photos on Facebook and other social networking sites, and images of celebrities and politicians proliferate in the media. A good appearance matters more now, both in private and public life.

Currently, the emphasis is on educational qualifications as the route to success in life, especially among academics. However, recent research shows that physical and social attractiveness is *equally important* to qualifications for success in life, on average. For some individuals, such as models and film stars, it can be much *more important* and the source of great wealth.

The increasing emphasis on visual culture, and on erotic capital, in developed economies has led some commentators to criticize what they regard as the sexualization of society. This is most common in the Anglo-Saxon Puritan cultures of North America and Britain, where the value of erotic capital (and related luxuries and entertainments, such as sumptuous clothing, music, and dancing) has been denied, historically. Some feminist writers are now extending this Puritan tradition of denigrating appearance as trivial, superficial, and worthless (Jeffreys, 2005; Rhode, 2010). Almost invariably these critical writings assume that the new emphasis on attractiveness affects women exclusively, or primarily. For example, they point to the fact that women spend money (or are 'forced' to spend money) on cosmetics. This overlooks the fact that men's rising investment in erotic capital takes different forms – such as committing to regular exercise in order to maintain fitness and build an athletic body (Thompson and Cafri, 2007), using nutritional supplements to enhance their appearance and stamina, using products to

avoid male-pattern baldness and maintain a healthy head of hair, spending money on extensive body tattoos, and even cosmetic surgery – as well as keeping slim, having good haircuts, and buying flattering clothing, just like women.

Social anthropological evidence, for example, from African tribes, shows that men and women have always invested time, money, and effort in looking good. In modern societies, there is simply a greater choice of styles and more diversity than in societies where self-presentation was dictated by tribal affiliation, sex, marital status, age, and social roles.

The Distribution of Erotic Capital

Only a few North American surveys provide information on national distributions of attractiveness. Most of them use the five-point scale shown in Table 1. On this scale, the majority of men and women are graded 'average' in looks, between half and two-thirds. A quarter to one-third are graded above average for their age, including tiny minorities of exceptionally handsome people. Around 1 in 10 are judged below average for their age. In these surveys, and most other studies, women are assessed as better-looking than men, probably because they expend more effort on grooming.

The distribution observed in North America may be idiosyncratic rather than universal. For example, in Shanghai, Chinese people are reluctant to grade anyone as below average (Hamermesh et al., 2002). So almost all women are assessed as average (2/3), above average (29%), or exceptionally beautiful (6%).

The Economic Value of Erotic Capital

Hatfield and Sprecher (1986) produced the first research synthesis on the diverse social and economic benefits of attractiveness in everyday life. The studies they review attest to the impact of erotic capital in all social interaction, but most are relatively small scale, reflecting the dispersed and exploratory nature of the field in the twentieth century. Hakim (2010, 2011) and Hamermesh (2011) present systematic reviews of the recent research evidence on the advantages of attractiveness. Hamermesh reviews economic studies in a textbook designed to explain how economists measure discrimination. Hakim synthesizes research in sociology, social psychology, economics, politics, social anthropology, and history to demonstrate the social and economic value of erotic capital in all contexts and all cultures.

From the American studies he reviews (most of them his own) Hamermesh concludes (2011: pp. 45, 50) that the average economic gain for an above-average attractive person compared to an unattractive person is 15%, all else equal, but the beauty premium is slightly higher for men at 17% than for women at 12%, on average.

Hakim concludes that someone with high erotic capital gains an average financial mark-up of between 10 and 20%, as compared with someone with below-average appearance. While this may seem a small addition, she notes that this can add up to a substantial cumulative gain over a lifetime, given that any small marginal advantage can deliver a large gain in 'winner-take-all' competitions for jobs or business opportunities, and in other situations where there are numerous candidates but only one winner. With the impact of globalization, and the trend toward larger multinational companies, the financial rewards for top jobs have increased, and so has the competition for them. As a result, the importance and potential benefits of any marginal advantage in personal assets, including erotic capital, are clearly increasing in the higher echelons of professional and management jobs. This may explain why cosmetic surgery companies report a steady increase in male customers in the twenty-first century, even during recessions.

In addition, exceptionally attractive men and women can become multimillionaires by exploiting their erotic capital alone. The most obvious examples are photographic models such as Kate Moss, Elle Macpherson, and Gisele Bundchen. Many film stars also benefit substantially from investing in their appearance and from fronting advertisements for luxury goods, including watches. One of the most beautiful film stars, Elizabeth Taylor, is said to have earned more from promoting her celebrity perfumes than from her lifetime earnings as an actor. Similarly, the footballer David Beckham is said to earn more, globally, from appearing in advertisements than from his main career as an athlete.

Quantitative and qualitative studies, case studies, and laboratory experiments are consistent in showing a slightly higher average return to erotic capital for men than for women. Typically, this is in earnings or adult incomes attained, but studies often show a similar sex differential for promotion and other rewards in the labor market. The most startling example

Table 1 Distribution of looks in the United States and Canada, 1970s

	American studies				*Canadian study*	
	1971		*1977*		*1977–81*	
	M	*F*	*M*	*F*	*M*	*F*
1. Strikingly beautiful or handsome	3	3	1	2	2	2
2. Above average (good looking)	24	28	27	30	32	32
3. Average for age	60	52	60	52	58	57
4. Below average for age (quite plain)	11	15	11	14	7	8
5. Homely	2	2	1	2	1	1
$N = 100\%$	864	1194	959	539	3804	5464

Source: Table 2 in Hamermesh, D., Biddle, J., 1994. Beauty and the labor market. American Economic Review 84, 1174–1194.

of this sex differential is found in the incomes of Hollywood film stars: top male stars can earn double the annual incomes of top female stars, according to regular *Forbes* estimates.

As yet, explanations for the sex differential in returns to erotic capital are purely speculative. One possibility is that this is yet another reflection of women's failure to exploit their advantages, their failure to negotiate and bargain for a better deal, in both private and public life (Hakim, 2011: pp. 200–205). Another is that it is a straight reflection of discrimination against women. This is the explanation adopted by Rhode (2010) and Hamermesh (2011); both assume that the returns to attractiveness are entirely unjustified and discriminatory. The third possibility is that the differential is restricted to Puritan Western cultures, where the value of beauty and sexuality have always been denied, as Hakim pointed out (Hakim, 2011: pp. 63–86). Research indicates that women's erotic capital is more highly rewarded outside the Puritan Anglo-Saxon countries where the bulk of recent studies have been carried out. For example in Shanghai, the economic returns to female attractiveness are higher than in the United States. They are at least equal to the returns for men and much higher for very attractive women (Hamermesh et al., 2002; Hamermesh, 2011: pp. 138–139). There is consistent evidence from many other sources that most cultures value attractiveness more positively than in America and Britain (Hakim, 2011: pp. 135–162).

Some studies have sought to compare the returns to attractiveness with the returns to educational qualifications (as a proxy for human capital more broadly). The best studies show that attractiveness, qualifications, and self-confidence deliver equal gains in income once the impact of intelligence and ability has been separated out (Judge et al., 2009; Hakim, 2011: p. 170). It appears that many studies greatly overstate the impact of educational qualifications on incomes because they do not separate out the influence of IQ, which is time consuming to measure. These results also show that the economic value of erotic capital is not due simply to any consequential gain in self-confidence, as this has a separate, equal impact of its own on incomes.

The returns to erotic capital appear to be larger in the higher echelons of the occupational ladder, where performance assessments can become subjective. In any event, attractiveness offers additional benefits to the highly qualified. Studies of the graduates of top law schools, and MBA graduates, found that attractive men and women earn 10–12% more than the unattractive ones, especially after several years on the job, when their talents and performance have been proven. The largest corporations tend to have better-looking CEOs than smaller companies. More generally, employing beautiful people has been shown to raise companies' sales and profits (Hamermesh, 2011).

The Social and Political Value of Erotic Capital

The first decade of the twenty-first century was the 'lift-off' period for concerted social science attention to the social, sexual, political, and economic benefits of attractiveness, with many new studies in all disciplines. Several reviews of the evidence were published, some of them focusing on female beauty and sex appeal, apparently geared toward the new gender studies courses. In line with the Western world's emphasis on facial beauty, and the research opportunities opened up by digital photography, many studies had an initial focus on facial beauty and body shape, but they usually went wider to tap erotic capital more broadly (Marwick, 2004; Hakim, 2011; Hamermesh, 2011). However, most reviews focus on adult lives, ignoring social processes during childhood.

The benefits of attractiveness are seen from birth onward. Beautiful babies and children attract more attention, assistance, social interaction, tolerance, and gifts – from complete strangers as well as friends and family. As a result, attractive children develop their cognitive and social skills faster, just as tall children benefit from being treated as older than their age. They are given more opportunities and have a smoother path in life.

These benefits continue in adult life. Attractive people are perceived as more intelligent, competent, and honest, and treated accordingly. They are given more support, cooperation, and assistance of every kind, by complete strangers as well as teachers, colleagues, acquaintances, and friends. For example, men and women who are attractive, well-groomed, and well-dressed are less likely to be arrested and less likely to be found guilty – unless the offense involved exploiting their good looks. Attractive men and women get loans more easily, win bigger contributions to charity, are more readily welcomed into social groups, and attract more good-looking and/or wealthy spouses.

As Aristotle once noted, beauty is the best letter of introduction. Maupassant's 1885 novel *Bel Ami* (*Beautiful Friend*) describes how a poor but handsome young man exploits his erotic capital for dramatic upward social mobility in nineteenth-century France, finally marrying into a wealthy family. In China today, handsome young men and women set out similar aspirations on dating websites. Yet young women who do the same in the Western world can be denigrated as 'gold-diggers' rather than recognized as exploiting their talents in exactly the same way as do men. It appears that there is an unstated double standard when it comes to exploiting erotic capital: men can do it; women should not do it at men's expense.

Attractive, charming, and charismatic men and women stand out from a young age, are remembered positively, so find it easier to develop their social as well as human capital. They develop more positive personalities and become more likable and acquire more friends and contacts. The association between intelligence and attractiveness is small at 0.10, but it is unmistakable. It may be a recent development.

The social benefits of being attractive explain why such people are more persuasive and can sell more goods, services, ideas, and policies. Research shows that people prefer to interact with, or buy from, an attractive person, even when they are aware it makes no difference to the transaction. Beauty is a good in itself and is a desirable luxury. All these results apply equally to men and women and are not affected by the sex of the assessor (Hakim, 2011).

Political studies provide the best alternative to economics for measuring the social impact, or value, of erotic capital. Instead of measuring additional income or earnings, political

scientists measure the additional votes won by above-average attractive people compared with those who are below average in appearance (Berggren et al., 2010; Hakim, 2011).

Studies show that attractive candidates win 15% extra votes on average. This advantage applies equally to male and female candidates. A good appearance has the strongest impact on votes for right-wing political parties and for candidates who are newcomers to the political scene, those who have no established track record. These patterns are repeated in nonpolitical elections, for example, voting for candidates onto committees or other roles in professional organizations and representative bodies (Hamermesh, 2011). Young people especially benefit from the extra goodwill accorded to attractive newcomers.

Overall, studies of the social and economic benefits of erotic capital are highly consistent. There is a +15% average benefit; the benefit is greatest for relative strangers, for whom there is little other information; and the benefits are generally larger for men than for women, despite the fact that women tend to score higher on erotic capital. The social benefits are substantial enough to justify the economic returns to attractiveness on grounds of improved productivity.

The Value of Erotic Capital in Intimate Relationships

It is well established that attractive men and women have a smoother path in dating, mating and marriage markets, and tend to marry younger. Attractive women tend to marry spouses with greater economic, social, and human capital (Feingold, 1990; Baumeister and Vohs, 2004; Todd et al., 2007; Hamermesh, 2011).

It is notable that national sex surveys and studies of sexuality carried out in the twentieth century do not even consider the impact of attractiveness and sex appeal on sexual activity, let alone attempt to measure it. The key exception was the first ever national survey of sexuality carried out in Sweden in 1967 by a sociologist, Hans Zetterberg, who was aware of the secret status ranking of sex appeal and sought to measure it and study its impact. Unfortunately, he published no survey results on this topic.

However, Zetterberg's 1967 Swedish survey inspired an entire research program in Finland, including a series of sex surveys in Finland and other Baltic states. Here, they did collect data on sex appeal and publish results on its impact on sexual behavior (Kontula, 2009: pp. 121–123). They confirmed the results of other studies by social psychologists and sociologists showing that attractive people have an earlier sexual debut, acquire wider experience, have more affairs, and (in Europe at least) have more partners than men and women who lack sex appeal. They also express greater satisfaction with their sex lives.

Studies of gay communities in North America show that sexual capital is of even greater importance than in heterosexual relationships and meet markets. High standards of sex appeal and physical fitness are underlined in gay meeting places such as bathhouses where men parade naked, and in bars and clubs where the right look, the right style of body and clothes can be crucial to acceptance, social interaction, and desire. Men who score low on sexual capital, due to age or lack of fitness may be ostracized, have reduced chances of sexual hook-ups, can have little or no control over the choice of partner, and may feel obliged to accept less safe sexual practices imposed by the more attractive partner. In situations where there is little or no scope for the exchange of money and sexual capital, physical attractiveness may become the dominant factor in relationships and sexual status rankings.

Studies of men and women who sell sexual services and erotic entertainment (such as lapdancing and stripping) invariably show a substantial mark-up in prices and incomes for the most attractive entertainers. These studies confirm that facial beauty alone is less important than the complete package of erotic capital, including liveliness and social skills. Young beautiful lapdancers who are glacial in manner make far less money than less beautiful women who have more positive personalities and good social skills. The high fees achieved by some men and women who sell sexual services (Hakim, 2011) demonstrate the very high full economic value of erotic capital in the round (all six elements).

Measuring Erotic Capital and Research Methods

As yet, no one has tried to measure erotic capital in the round. Studies typically focus on particular elements, such as facial beauty, body shape, or dress style, but inevitably capture the wider impact of social skills as well as physical looks, as most scholars recognize (Hakim, 2011).

Nationally representative surveys using face-to-face interviews have asked interviewers to grade respondents' appearance on the five-point scale shown in **Table 1**. Other studies employ three to five judges, male and female, of diverse ages, to grade the attractiveness of photos or videos in order to produce an average score for each image, before using the images in laboratory experiments or other studies.

Analyses show an astonishing degree of agreement and consistency in rankings, irrespective of the sex of the grader, respondent, or image. There is consistency in rankings even across cultures and races (Hakim, 2011).

People who are scored above average when young tend to be seen the same way decades later, indicating consistency across the life cycle as well. Validity and reliability of assessments do not appear to be a problem. Most people can distinguish between their own personal tastes and impersonal general assessments of someone's attractiveness within their society. So beauty may be in the eye of the beholder in private lives, but most people can recognize an Aphrodite or Adonis as exceptional within their society. Even people who prefer blondes can see that Elizabeth Taylor and Pierce Brosnan were exceptionally attractive in their prime.

Social psychologists have yet to develop a reliable measure of social skills or social intelligence (as they label it), but methodological work is slowly bearing fruit. Similarly, although Hochschild's (1983/2003) concept of emotional labor has proven popular, there is no agreed method for measuring it properly. However, in everyday life, we have no difficulty distinguishing people who are likable and charming from those who are clumsy, rude, or abrasive.

The advent of digital photography opened up a whole new research field for analyzing facial beauty, body shape, and styles of movement using IT to manipulate photos and videos to test what specific features of people are perceived as attractive or unattractive, such as the ideal hip-to-waist ratio. This new field of research is yielding important new discoveries (such as the importance of symmetrical features), with more to come.

Computer-based laboratory experiments use IT to manipulate the appearance of photos or videos of confederates to allow researchers to measure how each precise alteration affects responses, in terms of cooperation, generosity, and liking, for example, in variants of the Prisoner's Dilemma game.

Some studies have hired exceptionally handsome and beautiful actors to play a role in laboratory experiments that mimic natural social settings, to assess whether and how their good looks alters the results obtained from studies using invisible or ordinary-looking actors.

In sum, researchers are being extremely inventive in devising new types of study that measure the overall attractiveness of individuals; identify the key elements of attractiveness; or test for the impact of attractiveness on the responses, attitudes, perceptions, and social behavior of viewers. The latest studies have started to assess the benefits of attractiveness *relative* to other assets such as human capital or intelligence (Judge et al., 2009).

Studies normally differentiate between attractiveness (erotic capital), height, and body weight (using body mass index (BMI) as the standard measure). It is well established that there are separate effects of being tall (rather than short), being slim or normal weight (versus obese), and being attractive (versus ugly). These three aspects of appearance all affect social interaction and perceptions, sometimes with different effects for men and women. For example, obesity in women is judged more harshly than in men. Tallness is an advantage for men and women, but not beyond a certain point, when it becomes anomalous or dysfunctional. The research summarized here concerns the effects of attractiveness, excluding the effects of BMI or height. It is notable that the legal cases discussed by Rhode (2010) concern obesity (and the overweight) more often than attractiveness. Like tallness, obesity quickly becomes dysfunctional, for example, when extremely tall or fat people cannot fit into car or airline seats. Ugliness is rarely dysfunctional in this way, even if it limits access to certain occupations.

Directions for Future Research

Research on the social, political, economic, and sexual effects of attractiveness is a new and growing field, with contributions from sociology, social psychology, economics, political science, historians, and IT specialists. Lawyers have also become involved, typically assuming that attractiveness is worthless and prompts unlawful discrimination (Rhode, 2010). The only theoretical perspective that encompasses all disciplines, methods, and social contexts is Hakim's (2010, 2011) theory of erotic capital. Further methodological developments and discoveries can be expected in the next two to three decades. One continuing barrier is the cross-disciplinary character of the research field, with scholars in one discipline failing to pick up and incorporate relevant work being done in another discipline. There is, as usual, a fairly large divide between experimental computer-based and laboratory studies, and research on real-life settings and actors, such as studies of the careers of law school and MBA graduates. Contrary to the usual pattern, real-life studies often reveal larger (long-term) impacts of erotic capital than laboratory experiments. This is another indication that erotic capital in the round has wider effects than any single element.

There is growing recognition of the importance of erotic capital in many fields of activity where appearance may initially seem irrelevant, such as sport (Guttman, 1996) and politics (Berggren et al., 2010), as well as the more obvious fields of advertising (Reichert, 2003) and sexual markets (Feingold, 1990; Martin and George, 2006; Hakim, 2011, 2012). There is also increasing attention to the role of physical attractiveness in men's lives, especially in gay sexual communities (Thompson and Cafri, 2007; Green, 2008), providing a welcome balance to the excessive focus on women. We can anticipate many more case studies of the role of erotic capital in particular fields and social contexts; its role in dating, sexual and marriage markets; and more research to test for the *relative* importance of social, human, and erotic capital alongside cognitive ability (IQ), a factor which is so often overlooked in social science research.

Research on Internet dating and speed-dating events, and Facebook-based analyses, provide a fruitful combination of laboratory-style controls and real-life behaviors (Todd et al., 2007). Some of the most interesting future research is likely to come from these studies, which often combine disciplinary perspectives and skills to best effect.

See also: Beauty and the Labor Market; Charisma, Social Aspects of; Obesity and Labor Market Performance; Obesity: Cultural and Biological Factors; Sexual Motivation; Sexuality, Theories of; Wisdom, Psychology of.

Bibliography

Baumeister, R.F., Vohs, K.D., 2004. Sexual economics. Personality and Social Psychology Review 8, 339–363.

Berggren, N., Jordahl, H., Poutvaara, P., 2010. The looks of a winner: beauty and electoral success. Journal of Public Economics 94, 8–15.

Feingold, A., 1990. Gender differences in effects of physical attractiveness on romantic attraction: a comparison across five research paradigms. Journal of Personality and Social Psychology 59, 981–993.

Green, A.I., 2008. The social organisation of desire: the sexual fields approach. Sociological Theory 26, 25–50.

Guttman, A., 1996. The Erotic in Sports. Columbia University Press, New York.

Hakim, C., 2010. Erotic capital. European Sociological Review 26 (5), 499–518.

Hakim, C., 2011. Erotic Capital: The Power of Attraction in the Boardroom and the Bedroom. Basic Books, New York. Published in Britain as Honey Money: Why Attractiveness is the Key to Success. Penguin, London.

Hakim, C., 2012. The New Rules: Internet Dating, Playfairs and Erotic Power. Gibson Square, London.

Hamermesh, D.S., 2011. Beauty Pays. Princeton University Press, Princeton, NJ.

Hamermesh, D.S., Xin, Meng, Junsen, Zhang, 2002. Dress for success – does primping pay? Labour Economics 9, 361–373.

Hatfield, E., Sprecher, S., 1986. Mirror, Mirror: The Importance of Looks in Everyday Life. State University of New York Press, New York.

Hochschild, A.R., 1983/2003. The Managed Heart. University of California Press, Berkeley, CA.
Jeffreys, S., 2005. Beauty and Misogyny: Harmful Cultural Practices in the West. Routledge, London and New York.
Judge, T.A., Hurst, C., Simon, L.S., 2009. Does it pay to be smart, attractive, or confident (or all three)? Relationships among general mental ability, physical attractiveness, core self-evaluations, and income. Journal of Applied Psychology 94 (3), 742–755.
Kontula, O., 2009. Between Sexual Desire and Reality: The Evolution of Sex in Finland (M. Makinen, Trans.). Population Research Institute, Helsinki.
Laumann, E.O., Gagnon, J.H., Michael, R.T., Michaels, S., 1994. The Social Organisation of Sexuality: Sexual Practices in the United States. University of Chicago, Chicago, IL.
Martin, J.L., George, M., 2006. Theories of sexual stratification: toward an analytics of the sexual field and a theory of sexual capital. Sociological Theory 24, 107–132.
Marwick, A., 2004. A History of Human Beauty. Hambledon, London.
Maupassant, G., 1885/1990. Bel Ami. Pocket, Paris.
Mobius, M.M., Rosenblat, T.S., 2006. Why beauty matters. American Economic Review 96, 222–235.
Reichert, T., 2003. The Erotic History of Advertising. Prometheus Books, Amherst, NY.
Rhode, D.L., 2010. The Beauty Bias: The Injustice of Appearance in Life and Law. Oxford University Press, New York.
Thompson, J.K., Cafri, G. (Eds.), 2007. The Muscular Ideal. American Psychological Association, Washington, DC.
Todd, P.M., Penke, L., Fasolo, B., Lenton, A.P., 2007. Different cognitive processes underlie human mate choices and mate preferences. Proceedings of the National Academy of Sciences of the United States of America 104 (38), 15011–15016.
Webster, M., Driskell, J.E., 1983. Beauty as status. American Journal of Sociology 89, 140–165.
Zetterberg, H.L., 1969/2002. Sexual Life in Sweden, translated with a new Introduction by Graham Fennell. Transaction Publishers, New Brunswick, NJ.

Errors in Statistical Assumptions, Effects of

Mark J Schervish, Carnegie Mellon University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 7, pp. 4739–4744, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

This article discusses the assumptions that underlie statistical procedures and how violations of these assumptions can affect the properties of a procedure. Both theoretical and simulation methods are discussed for assessing the sensitivity of properties of statistical procedures to violations of their assumptions. The example of the *t*-test is considered in detail.

Each statistical procedure rests upon a collection of assumptions for its validity. For example, many statistical procedures assume that a data set consists of a collection of independent observations. A large number of these same procedures also assume that the observations all have the same distribution. Each of these assumptions, as well as any others that a procedure assumes, may or may not be reasonable in a given application. If one or more of the assumptions justifying a procedure fail to hold, the statistical properties of the procedure may be different than they would be if the assumptions hold. This article discusses ways to assess the sensitivity of a statistical procedure to violations of its assumptions.

Introduction

The development of almost every statistical procedure is accompanied by a list of properties possessed by the procedure under some specified statistical assumptions. For example, the *t*-test is a uniformly most powerful unbiased level α test if one assumes that the data $X_1, \ldots, X_n$ comprise a sample of independent random variables with common normal distribution. Of course, one can perform a *t*-test with data $X_1, \ldots, X_n$ regardless of whether one makes some or all of the assumptions. One cannot expect the associated properties to hold, however, if one does not believe the assumptions. If one has reason to believe (or fear) that one or more of the assumptions used to justify a statistical procedure fail to hold, then one should be concerned with how sensitive the properties are to the assumptions. For example, suppose that the observations $X_1, \ldots, X_n$ have some distribution other than normal. How much does that affect the probability of type I error when I use the level α *t*-test? How much does it affect the power of the test at important alternatives? This article will not attempt to answer all of the questions of this sort. Rather, this article describes general ways to assess the sensitivity of statistical procedures to specific violations of their assumptions. An excellent discussion of the role of assumptions in statistical models was initiated by Box (1980).

Related to the question of sensitivity is the question of diagnosis. That is, can one tell whether or not the data suggest that certain assumptions are indeed violated? There are several statistical methods that have been developed to diagnose violations of assumptions of the linear regression model. Some of these are discussed in Linear Hypothesis: Regression (Graphics).

For a given statistical procedure, one first needs to be aware of what assumptions justify the procedure. Then one needs to decide which aspects of the procedure and which assumptions the sensitivity of which one wishes to assess. One needs to decide which violations of those assumptions one is willing to contemplate and finally how one wishes to measure the sensitivity. Each of these issues is examined in a detailed study of one statistical procedure in the section 'The *t*-test'.

Theoretical results concerning sensitivity can be difficult to derive, even for the most elementary types of violations. Hotelling (1961) and Efron (1969) analyze the effects of some specific types of violations on the distribution of the *t* statistic. Box (1953) discusses the effects of violations of assumptions on tests concerning variances.

The *t*-test

Since the *t*-test is such a popular statistical procedure, let us first consider violations of its assumptions. The assumptions for the validity of the *t*-test are that the data comprise a sample of independent random variables, all with the same normal distribution. Specifically, there are three key assumptions:

1. the observations are independent;
2. the observations all have the same distribution; and
3. each observation has a normal distribution.

There are many ways in which these assumptions can be violated. For example, the observations might be dependent. (*See* Time Series: General for discussion of one type of model of dependence between observations.) The observations might have different distributions. Alternatively, some or all the observations might have distributions other than normal. Finally, it is possible that more than one assumption is violated. For example, each observation could have its own distribution, none of which is normal, and they could be dependent.

The most crucial properties of a test procedure are its size and power. These can be measured from the distribution of the test statistic given a set of assumptions about the data. For example, consider a test of the null hypothesis that the mean μ of a sample of random variables is 0. If one were interested in measuring the sensitivity to the assumption that the data are normal, and if one contemplates the possibility that a linear function $a(X-b)$ of each observation X has a Cauchy distribution, then one must reinterpret μ to be something other than the mean (e.g., the median, b). The power of the test can then be written as a function of $a\mu$.

Theoretical Results

Clearly, it is not reasonable to expect to find theoretical results that handle every possible departure from the assumptions. For some violations, it is possible to determine analytically how sensitive a procedure will be. For example, suppose that the observations $X_1,\ldots, X_n$ have normal distribution with mean μ and common variance σ^2, but that they have common correlation ρ. And suppose that one wishes to perform a t-test of the null hypothesis that the mean of the distribution is μ_0. The t statistic has the form

$$T = \frac{\overline{X} - \mu_0}{\sqrt{S^2/n}} \qquad [1]$$

where $\overline{X}$ is the sample average and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \qquad [2]$$

Under the usual assumptions for the t-test, if $\mu = \mu_0$, then T has a t distribution with $n-1$ degrees of freedom. In the case of common correlation, it is possible to show that the random variable $U = T/\sqrt{1 + n\rho/(1-\rho)}$ has t distribution with $n-1$ degrees of freedom. If ρ is greater than 0, the distribution of T will be more spread out than the nominal t distribution if $\mu = \mu_0$. This will have the effect of making the size of the t-test larger than its nominal value. For example, if $\rho = 0.05$ and $n = 10$ and one uses a test with nominal size 0.05; the actual size is 0.10. If the null hypothesis is false and $\mu \neq \mu_0$, then U has noncentral t distribution with $n-1$ degrees of freedom and noncentrality parameter $\mu\sqrt{n/[1+(n-1)\rho]}$. The effect of this on the power of the test is not so obvious, but it is not as dramatic as the effect on the size.

For other violations, one can attempt to reason qualitatively what the effect will be. For example, suppose that the observations $X_1, \ldots, X_n$ come from a distribution with heavier tails than a normal distribution (i.e., observations far from the median occur with high probability). And suppose that one wishes to perform a t-test of the null hypothesis that the median of the distribution is μ_0. Some theoretical results concerning this case have been derived by Benjamini (1983). A heavy-tailed distribution will produce a value $\overline{X}$ with more variability than a normal distribution will. That is, we would expect to see large values of $\left|\overline{X} - \mu_0\right|$ with higher probability from a heavy-tailed distribution than from a normal distribution. But the effect of heavy tails on $\sqrt{S^2/n}$ should be even greater than the effect on $\overline{X}$ because the squaring gives the large deviations $X_i - \overline{X}$ much more prominence. Hence, one expects $|T|$ to be smaller if the data have a distribution with heavy tails than if the data have normal distribution. Not all violations lend themselves to qualitative analysis of this sort, however. And, as Benjamini (1983) points out, an intuitive analysis like this might only hold in special cases, as will be noted later in this article.

When one wishes to focus on a specific collection of possible violations, then one can perform simulations to assess both qualitative and quantitative sensitivity. Simulation has advantages and disadvantages when compared to theoretical analysis. A simulation can be designed to assess the sensitivity of any statistical procedure to any specific collection of violations of its assumptions. However, each simulation analysis assesses only the specific violations for which it was designed and does not tell us how sensitive the procedure is to other, albeit similar, violations. Of course, simulations are generally quick and easy to perform, so one could perform many of them to cover a large number of specific violations. The remainder of this section illustrates the sorts of simulations that one might perform in order to assess the sensitivity of various aspects of t-test to some of the more easily understood violations of assumptions.

Violations of Normality

Suppose that one thinks that the observations are normal with mean μ and variance σ^2, and that one wishes to test the null hypothesis that $\mu = 0$. Suppose that most interest lies in a test with level 0.05 and with the power of the test when $\mu = \sigma$. However, suppose that one is willing to contemplate the possibility that the data come from a distribution that is not normal. If that distribution does not have a mean or does not have a variance, then one will need to reinterpret μ and/or σ. For distributions without a mean, let μ stand for the median. For all distributions, let σ stand for 0.741 times the interquartile range (which happens to be the standard deviation when the data have a normal distribution). Now, one can simulate samples of size n from every distribution that one wishes to consider and compute the probability that $|T| > c$, where c is the 0.975 quantile of the t distribution with $n-1$ degrees of freedom. For distributions with mean (or median) 0, this probability will be the size of the test. For distributions with mean $\mu = \sigma$, this probability will be the power of interest. Table 1 reports the results of several simulations. Each row of the table represents two million simulated samples of size $n = 10$ or $n = 20$ from a distribution that is not normal. The simulated data from the t distribution have three degrees of freedom. The simulated data from the gamma distribution have parameter 2. The σ in the last column is 0.741 times the interquartile range. The 'normal' rows give the theoretical values under the normal distribution assumption. For each sample the statistic t was computed. The table reports the

Table 1 Simulated sizes and powers of the t-test under violations of normality

		Prob. of $\lvert T\rvert > c$	
Distribution	*Sample size (n)*	$\mu = 0$	$\mu = \sigma$
Laplace	10	0.042	0.593
	20	0.046	0.862
Cauchy	10	0.019	0.295
	20	0.020	0.325
t (3 degrees of freedom)	10	0.040	0.619
	20	0.043	0.848
Uniform	10	0.054	0.971
	20	0.052	1.0
Exponential	10	0.099	0.786
	20	0.081	0.997
Gamma (2)	10	0.077	0.838
	20	0.066	0.999
Normal	10	0.05	0.803
	20	0.05	0.989

estimated probability that $|T| > c$ for c equal to the 0.975 quantile of the t distribution with $n-1$ degrees of freedom.

A quick look at the rows of Table 1 corresponding to heavy-tailed distributions (Laplace, Cauchy, t) indicates what we argued earlier. That is, the probability of $|T| > c$ is smaller for these heavy-tailed distributions than for the normal. However, as is common with simulations, one must be careful not to generalize too far. If one had considered level 0.20 tests instead of level 0.05 tests, the results would have been slightly different. For example, with the Laplace distribution, the estimated level is slightly higher than 0.2 with samples of both sizes 10 and 20. This is consistent with the theoretical results of Benjamini (1983).

Violations of Common Distribution

For the assumption that the observations have the same distribution, there are many possible violations. The data might split into two (or a few) groups, each of which has a common distribution. Alternatively, each observation could have its own distribution. Or there might be just one (or a few) observation(s) with distribution different from the others. All of these are possible even if one believes that each of the individual distributions has a normal distribution.

Table 2 gives the results of several simulations. As before, each simulation is based on two million simulated samples of size $n = 10$ or $n = 20$. Each sample consists of n normal random variables with a common mean μ but with either 10% or 50% having a larger standard deviation than the rest. The larger standard deviation is either 2, 5, or 10 times as large as the smaller one. The value of σ in the last column is

$$\sigma = \sqrt{(1-p)\sigma_1^2 + p\sigma_2^2} \qquad [3]$$

where σ_1 is the smaller standard deviation, σ_2 is the larger standard deviation, and p is the proportion of the sample with the larger standard deviation. The 'ratio = 1' rows give the theoretical values under the assumption that the data have a common normal distribution. The table reports the estimated probability that $|T| > c$ for c equal to the 0.975 quantile of the t distribution with $n-1$ degrees of freedom. Note how the size of the test can be significantly lowered by the inclusions of 10% of the observations with a standard deviation 5 or 10 times as large as the rest. The power of the test at $\mu = \sigma$ is not much affected by this type of violation. However, the value of σ/σ_1 increases as the ratio of the two standard deviations increases. If one had calculated the power at a fixed alternative, there would have been a dramatic decrease in power as the ratio of standard deviations increased and as the proportion with the larger standard deviation increased.

Table 2 Simulated sizes and powers of the t-test under violations of common distribution

Ratio of standard deviations	*Sample size*	*Proportion with larger standard deviation*	*Prob. of $\|T\| > c$*	
			$\mu = 0$	$\mu = \sigma$
2	10	0.1	0.047	0.809
	10	0.5	0.046	0.804
	20	0.1	0.048	0.984
	20	0.5	0.048	0.985
5	10	0.1	0.030	0.831
	10	0.5	0.037	0.800
	20	0.1	0.033	0.959
	20	0.5	0.046	0.982
10	10	0.1	0.017	0.829
	10	0.5	0.033	0.799
	20	0.1	0.015	0.948
	20	0.5	0.045	0.981
1	10	0	0.05	0.803
	20	0	0.05	0.989

Violations of Independence

Violations of the independence assumption are a bit more difficult to think about. There are simple violations such as equal correlation that is not 0. Indeed, this case is the one that lent itself to an analytic solution earlier in this article. Another type of violation of independence is time-series dependence. This means that the amount of dependence between observations depends on the order in which they were observed. Many data sets are collected without recording when the observations were made. This would make it impossible to perform a traditional time series analysis. However, we can still assess the effect of time series dependence on the distribution of a statistic like t, even if one did not record the observation times. For example, suppose that observations arrive in the order $x_1, \ldots, x_n$, regardless of whether or not one knows this to be true. One might consider the possibility that $x_i = \rho x_{i-1} + \varepsilon_i$ for $i = 2, \ldots, n$. Here $\rho \in (-1, 1)$ and $\varepsilon_2, \ldots, \varepsilon_n$ are assumed to be mutually independent with variance $(1-\rho^2)\ \mathrm{Var}(x_1)$. This type of dependence is called autoregressive of order 1. (For more detail on autoregressive dependence, *see* Time Series: ARIMA Methods.)

Table 3 presents the results of several simulations. Each simulated sample has autoregressive dependence of order 1 with a different correlation. Each number in the table is based on 2 000 000 simulated samples. The rows with '$\rho = 0$' give the theoretical values under the assumption that the data are independent. The table reports the estimated probability that $|T| > c$ for c equal to the 0.975 quantile of the t distribution

Table 3 Simulated sizes and powers of the t-test under violations of independence

Sample size (n)	*Correlation (ρ)*	*Prob. of $\|T\| > c$*	
		$\mu = 0$	$\mu = \sigma$
10	0	0.05	0.803
	0.1	0.072	0.792
	0.3	0.138	0.773
	0.5	0.245	0.759
	0.7	0.412	0.761
	0.9	0.672	0.824
20	0	0.05	0.989
	0.1	0.074	0.982
	0.3	0.145	0.962
	0.5	0.256	0.928
	0.7	0.422	0.883
	0.9	0.688	0.864

with $n-1$ degrees of freedom. Even correlation as small as 0.1 has a noticeable effect on the size of the test.

Linear Regression

The assumptions which justify the standard inferences in linear regression are a bit more complicated than those which justify the t-test. In turn, the types of violations of these assumptions are also more complicated. To be specific, consider simple linear regression. The data comprise a set of pairs $(Y_1, x_1),\ldots,(Y_n, x_n)$, and the numbers $x_1,\ldots, x_n$ are considered fixed, known constants. Alternatively, the remainder of the probabilistic statements are to be understood conditional on these values. The assumptions are that

1. the random variables $Y_1,\ldots, Y_n$ are independent,
2. every Y_i has a normal distribution,
3. Var (Y_i) is the same for all i, and
4. there exist numbers β_0 and β_1 such that $E(Y_i) = \beta_0 + \beta_1 x_i$.

This section will briefly discuss violations of these assumptions, but not in as much detail as it did for the one-sample t-test. Failures of independence can come in the same forms that they did for the one-sample t-test. However, there are additional varieties available as well. For example, one might consider the possibility that there is correlation between Y_i and Y_j that depends on how far apart x_i and x_j are. Each Y_i can have a distribution that depends on x_i through more than just its mean and/or variance. Indeed, there are standard methods in regression analysis for diagnosing and dealing with cases in which it appears that Var (Y_i) depends on x_i. The dependence of $E(Y_i)$ and x_i can be very general. It might seem reasonable in some problems to assume that $E(Y_i)$ is a monotone function of x_i, but not always. For more discussion of regression methodology, including diagnosing violations of assumptions, *see* Linear Hypothesis: Regression (Basics) and Linear Hypothesis: Regression (Graphics).

Sensitivity of Bayesian Procedures

Bayesian procedures (*see* Bayesian Statistics), being statistical procedures, are also based upon probabilistic assumptions, and are thereby sensitive to violations of their assumptions. Typically, a Bayesian procedure is based on the same types of assumptions as any other procedure plus additional assumptions about the distributions of any parameters (such as μ and σ^2 in the t-test) that might appear in the statistical model. These latter assumptions are generally part of what is called the prior distribution of the parameters. Sensitivity of Bayesian procedures to violations of all of these assumptions can be assessed in a manner similar to what was done earlier in this article. However, those who use Bayesian procedures to represent subjective beliefs conditional on the observed data would be more interested in assessing sensitivity in a different manner.

For example, suppose that one wants an interval estimate of a parameter μ, and one models data $X_1,\ldots, X_n$ as independent normal random variables with mean μ and variance σ^2. One would also specify a prior distribution for (μ, σ^2). After observing the data values $X_1 = x_1,\ldots, X_n = x_n$, one would then compute the appropriate interval estimate (a, b). Then one could ask what interval would have been computed if one had modeled the data as having a different distribution. For example, suppose $(X_i - \mu)/(0.6745\sigma)$ are modeled as independent with Cauchy distribution. (The factor of 0.6745 makes the interquartile range of X_i the same as it would have been if the data were normal, namely 1.349σ.) Under this Cauchy assumption, a different interval (a', b') would be computed using the same observed data values $x_1,\ldots, x_n$. The change from the interval (a, b) to the interval (a', b') measures the sensitivity of the process of forming an interval estimate to this change of the assumption about the distribution of the data. This type of sensitivity of the process of forming an inference differs somewhat from the type of sensitivity of a particular procedure that was described earlier. Sensitivity of process is not a uniquely Bayesian concern, however, and it can also be measured for changes to assumptions in non-Bayesian inference.

There have been some successful attempts to take considerations of sensitivity to assumptions into account explicitly in the formulation of Bayesian models. Model uncertainty (see Draper, 1995) and Bayesian model averaging (Hoeting et al., 1999) attempt to develop methodology for expressing directly in the modeling process the belief that more than one out of several competing statistical models might be appropriate for the data, and it may not be clear which one is best. For example, one might not be able to decide whether to model data as normal random variables or as Cauchy random variables. The methods of model uncertainty and model averaging provide ways to pursue both models (as well as others that might seem plausible) without having to pretend as if we really believed firmly in precisely one of them.

Finally, there have been some methods developed to deal with specific common violations of assumptions. The assumption that all observations in a sample have the same distribution has received attention particularly in order to be able to accommodate the occasional outlying observation. One can model data as coming from a mixture of two or more distributions. One of the distributions is the main component on which most interest is focused. The others represent the types of data that might arise when something goes wrong with the data collection procedure. Each observation can be modeled as coming from one of the component distributions with a probability associated with each component. Indeed, using powerful simulation methods (*see* Markov Chain Monte Carlo Methods) one can even compute, for each observation, the conditional probability that it arose from each of the components given the observed data.

Conclusion

This article has discussed the need for carefully acknowledging the probabilistic assumptions made to justify a statistical procedure and some methods for assessing the sensitivity of the procedure to those assumptions. If one believes that there are violations of the assumptions that might reasonably arise in practice, and if the procedure is overly sensitive to violations of those assumptions, one might wish to select an alternative procedure.

There are two popular approaches to choosing alternatives to standard statistical procedures when one fears that assumptions are likely to be violated. One is to use robust procedures, and the other is to use nonparametric procedures. Robust procedures are designed to be less sensitive to violations of specific assumptions without sacrificing too much of the good performance of standard procedures when the standard assumptions hold. Nonparametric procedures are chosen so that their properties can be verified under fewer assumptions. For more details on these types of procedures *see* Robustness.

See also: Linear Hypothesis: Regression (Basics); Linear Hypothesis: Regression (Graphics); Robustness; Time Series: ARIMA Methods; Time Series: General.

Bibliography

Benjamini, Y., 1983. Is the *t* test really conservative when the parent distribution is long-tailed? Journal of the American Statistical Association 78, 645–654.

Box, G.E.P., 1953. Non-normality and tests on variances. Biometrika 40, 318–335.

Box, G.E.P., 1980. Sampling and Bayes' inference in scientific modeling and robustness (with discussion). Journal of the Royal Statistical Society, (Series A) 143, 383–430.

Draper, D., 1995. Assessment and propagation of model uncertainty (with discussion). Journal of the Royal Statistical Society, (Series B) 57, 45–97.

Efron, B., 1969. Student's *t*-test under symmetry conditions. Journal of the American Statistical Association 64, 1278–1302.

Erceg-Hurn, David M., Mirosevich, Vikki M., 2008. Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. American Psychologist 63 (7), 591.

Hampel, Frank R., Ronchetti, Elvezio M., Rousseeuw, Peter J., Stahel, Werner A., 1986. Robust Statistics: The Approach Based on Influence Functions. John Wiley & Sons, New York.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C., 1999. Bayesian model averaging: a tutorial (with discussion). Statistical Science 14, 382–417.

Hotelling, H., 1961. The behavior of some standard statistical tests under nonstandard conditions. In: Neyman, J. (Ed.), Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, CA, pp. 319–359.

Zuur, Alain F., Ieno, Elena N., Elphick, Chris S., 2010. A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution 1 (1), 3–14.

Essentialism

Nick Haslam and Maria Abou-Abdallah, University of Melbourne, Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Psychological essentialism is the belief that an entity or category has an underlying essence that determines its identity. Essentialist intuitions are common in laypeople's thinking about biological kinds and some social categories. In particular, essentialist beliefs about social groups have been shown to have a variety of problematic implications, contributing to social stereotyping, prejudice, and discrimination. This article reviews concepts and empirical studies on essentialism.

What Is Essentialism?

The concept of 'essentialism' originated in philosophy, where it usually refers to the claim that all instances of a certain category – particularly those categories known as 'natural kinds' – share an underlying essence that determines their identity. For instance, the essence of the element *gold* might be its atomic structure and the essence of the species *tiger* might be its DNA structure. In both cases, the proposed essence held in common by all instances of the element or species is causally responsible for their observable features but is not directly observable itself.

Essentialism has been a contentious idea and most philosophers question whether categories do in fact have essences. However, whether or not essentialism is philosophically justified, everyday people appear to hold essentialist intuitions about a variety of categories. This lay intuition, dubbed 'psychological essentialism' (Medin and Ortony, 1989), has been the focus of a variety of research programs within cognitive, developmental, and social psychology (Gelman, 2003). Cognitive psychologists have examined how essentialist thinking enables category learning by leading people to expect deep-seated similarities among the instances they encounter. Developmental psychologists have documented essentialist thinking about biological kinds among preschool children and explored some of the factors that promote it. Social psychologists have explored the implications of essentialist thinking about social groups, and in particular its role in stereotyping and prejudice.

This article will emphasize the social psychological study of essentialist thinking. First we examine theoretical approaches that have animated this work, before turning to a review of several key research findings.

Essentialism in Social Psychology

The first reference to essentialism within social psychology lies in Gordon Allport's classic text, *The Nature of Prejudice* (1954). Allport proposed that a belief in category essences is a fundamental cognitive ingredient of the rigid style of thinking that underpins prejudice. This idea disappeared for almost 40 years before Rothbart and Taylor (1992) argued that people commonly apply essentialist thinking to social categories, such as races and genders. They identified two key aspects of this form of thinking: it led people to see categories as unalterable and 'inductively potent' (i.e., a rich source of information about a person who belongs to the category). Essentializing a social group leads people to see it as fixed and identity-defining. Belonging to such a group, therefore becomes an unchanging and unchangeable part of who they are as a person. According to Rothbart and Taylor, essentialist thinking is destructive because it accentuates perceived differences between social groups and makes existing social arrangements seem inevitable. It also leads people to draw conclusions about people's underlying natures on the basis of their superficial differences and to overlook the role of social, historical, and cultural forces in the shaping of individuals.

Following Rothbart and Taylor's work, several distinct theoretical approaches to the study of essentialist thinking have arisen. One approach, linked to anthropologist Lawrence Hirschfeld (1996), holds that essentialist thinking is one 'mode of construal' that is available to people to make sense of the kinds of people they encounter. Children come into the world cognitively prepared to see essence-based groupings such as 'races.' Another approach, proposed by anthropologist Francisco Gil-White (2001), argues that essentialist thinking is not simply an abstract mode of construal, but a view of categories based on our 'folk biology.' On this account, people construe biological species as if they have distinct essences, and sometimes they may apply this way of thinking to human groupings by analogy. Just as members of a species seek one another out as mates (endogamy) and produce other members as progeny (descent), so do members of ethnic groups. Thus, it is unsurprising that children come to perceive ethnic groups as species-like natural kinds.

Three other major approaches to the study of psychological essentialism have come out of social psychology rather than anthropology. Vincent Yzerbyt and colleagues (e.g., Yzerbyt et al., 1997) have linked essentialist thinking to stereotyping and proposed that it has a system-justifying function. Essentialist beliefs about social categories such as races and genders serve as explanations for existing social stereotypes (e.g., the claim that women are innately maternal or intrinsically unsuited for certain occupations on account of deep-seated genetic or neural factors). Because these essences are understood to be innate and fixed, they make existing social arrangements, such as inequalities, seem inevitable and ordained by nature.

A second psychological account of essentialism is based on the work of Nick Haslam (e.g., Haslam et al., 2000). He has argued that essentialist beliefs about social categories have multiple facets, which form two distinct dimensions. On the one hand, some groups are seen as natural kinds, which involves beliefs that they have crisp, either–or category boundaries (discreteness), a biological basis (naturalness), are invariant across time and cultures (historical stability), and membership of the group is fixed (immutability). On the other hand, some groups are reified or seen as 'entitative.' These groups are believed to be inductively potent (informativeness), homogeneous (uniformity), and to have deep-seated (inherence) and defining features (necessity). Essentialist beliefs are therefore not unitary and social groups can be essentialized in different ways.

An influential third approach to the study of psychological essentialism, albeit one that does not employ the term, examines 'lay theories' about human attributes (Hong et al., 2001). These theories refer to beliefs about the malleability versus fixity of these attributes: for example, people who believe that intelligence can be modified are 'incremental theorists,' whereas those who believe it is fixed are 'entity theorists.' Holding an entity theory is equivalent to having an essentialist belief in the immutability of an attribute, and the many deleterious psychological consequences of entity theories – for motivation, social judgment, and stereotyping – speak to the problematic nature of essentialist thinking.

Empirical Investigations

Numerous psychological studies have now examined the implications, correlates, and determinants of essentialist thinking. Some of these have addressed general questions. For example, it has been shown that those who tend to essentialize differences between people also tend to endorse the accuracy of social stereotypes (Bastian and Haslam, 2006), indicating that essentialism is associated with stereotyping as Yzerbyt and others had proposed. Similarly, it has been shown that essentialist thinking is promoted by the use of generic statements about groups (e.g., "men are aggressive": Leslie and Gelman, 2012). However, most social psychological research on essentialist thinking has studied it in relation to particular social categories. Key findings of this work for the primary social categories of interest will be reviewed below.

Race, Ethnicity, and Culture

Several studies have examined whether individual differences in essentialist beliefs about race or ethnicity are associated with prejudiced attitudes. Generally speaking, this work has found correlations between essentialist thinking and measures of racism, particularly when essentialist beliefs involving genetic factors have been examined (i.e., "genetic essentialism": Dar-Nimrod and Heine, 2011). Relatedly, researchers have found that essentialist thinking among members of a host culture is associated with a less welcoming attitude toward immigrants, and that essentialist thinking among immigrants may also correlate with reluctance to integrate with the host culture. Essentialist thinking also has implications for people's ability to negotiate bicultural identities. One study (Chao et al., 2007) found that Asian-Americans who held racially essentialist beliefs were more troubled when asked about their bicultural identity and were less able to perform a task that required them to switch between their Asian and American identities. Even so, essentialist thinking may have benefits for members of ethnic minorities, such as immigrant communities, by enabling them to maintain a sense of identity, assert a political voice, and resist demands to assimilate (Verkuyten, 2003).

Gender

Gender categories are commonly essentialized even among young children, often in ways that accentuate perceived gender differences. For example, adults are more likely to judge one of their own attributes as unchanging when told it is associated with gender rather than when told it is not, and those who hold biological explanations of gender differences see those differences as more pronounced. There is also evidence that individuals who endorse genetic determinism and biological explanations show higher levels of sexist attitudes (Keller, 2005). This relationship between gender essentialism and gender stereotype endorsement extends to self-stereotyping. For example, women who believe that gender differences are biologically based are more likely to endorse stereotypically feminine traits (Coleman and Hong, 2008), and fathers who held more essentialist views about gender differences provided less direct care for their young children (Gaunt, 2006). In another important study, women performed worse on a math test when told that women tended to perform worse on it than men (i.e., 'stereotype threat'), but only when they had been given a biological essentialist explanation of gender differences in mathematical aptitude (Dar-Nimrod and Heine, 2006).

Sexual Orientation

Essentialist thinking about sexual orientation is interesting in part because it seems to have a somewhat complex association with prejudice. While endorsement of essentialist beliefs about race and gender tend to be associated with racism and sexism, respectively, some essentialist beliefs about homosexuality appear to be related to the absence rather than the presence of antigay attitudes. For example, Haslam and Levy (2006) found that antigay attitudes were associated with the essentialist belief that homosexuality is a discrete category, but were negatively associated with the essentialist beliefs that homosexuality is immutable and historically stable. Importantly, these belief dimensions predicted antigay attitudes independently of established predictors of prejudice such as authoritarianism and political conservatism. More religious people held more negative attitudes because, in part, they believed homosexuality to be mutable, and men held more negative attitudes than women because they believed it to be more discrete.

Mental Disorder

Psychologists have recently begun to study the implications of essentialist beliefs about mental disorders. Such beliefs – that disorders are natural kind with sharp boundaries and biological determinants – roughly corresponds to the biomedical or disease model of mental disorders. Research has shown that many people do in fact hold essentialist beliefs about at least some mental disorders (Ahn et al., 2006) and that these beliefs are linked to prejudice toward those who suffer from them (psychiatric stigma). A growing body of evidence shows that people who endorse more biomedical explanations for mental disorders tend to have less favorable attitudes toward sufferers. Research by Phelan (2005) indicates how 'genetic essentialist' beliefs about mental disorders might promote rather than reduce stigma, and that genetic attributions increase the perceived seriousness and persistence of mental disorders, as well as the perceived risk that the condition would affect the family members of those afflicted.

Personality

Some research beliefs about personality attributes has focused on the extent to which personality traits are seen *as* essences. Although all traits are to some extent essentialized, where the trait concept implies some degree of fixity and underlying causal influence on behavior, some dispositions may be more essentialized than others. Much of this work has examined the development of children's beliefs about the innateness and stability of traits. Giles (2003), for example, shows that children tend to hold essentialist beliefs about aggressiveness, where children who exhibit it are seen as likely to remain that way, and their aggressiveness is attributed to intrinsic rather than environmental factors. Among adults, the traits that are highly essentialized tend to be those that are seen as desirable, prevalent, and emotion-related. As deep-seated essences, these traits are seen as particularly central to a person's identity and to 'human nature' (Haslam et al., 2004).

Conclusions

Psychological essentialism is an intriguing concept that has generated multiple lines of theory and research across several of psychology's subdisciplines. Although at one level, they reflect abstract 'ontological' intuitions about the nature of social reality, at another they have important implications for everyday social perception and intergroup relations. People who tend to essentialize the social world tend to believe in the reality of social stereotypes, form rapid category-based judgments about other people and, with some exceptions, hold more negative attitudes toward traditionally stigmatized and disadvantaged groups. When they hold essentialist beliefs about their own groups, they tend to be constrained by in-group stereotypes and to have more difficulty in navigating the complexities of their identities. Overall, essentialist thinking tends to produce an image of the social world as composed of distinct identities and to deepen the divides between them. For this reason, it is an important topic of study for social and developmental psychologists.

See also: Acculturation; Concept Formation in Childhood: Cross-Cultural Perspectives; Group Processes, Social Psychology of; Heterosexism and Homophobia; Identity and Identification, Social Psychology of; Intergroup Relations; Mentally Ill, Public Attitudes towards; Racism: Social Psychological Perspectives; Reification: History of the Concept; Sexism; Social Categorization; Social Psychology; Stereotypes in Social Psychology; Xenophobia: Social Psychological Aspects.

Bibliography

Ahn, W., Flanagan, E., Marsh, J.K., Sanislow, C., 2006. Belief about essences and the reality of mental disorders. Psychological Science 17, 759–766.
Allport, G.W., 1954. The Nature of Prejudice. Addison-Wesley, Reading, MA.
Bastian, B., Haslam, N., 2006. Psychological essentialism and stereotype endorsement. Journal of Experimental Social Psychology 42, 228–235.
Chao, M., Chen, J., Roisman, G.I., Hong, Y.Y., 2007. Essentializing race: implications for bicultural individuals' cognition and physiological reactivity. Psychological Science 18, 341–348.
Coleman, J.M., Hong, Y.Y., 2008. Beyond nature and nurture: the influence of lay gender theories on self-stereotyping. Self and Identity 7, 34–53.
Dar-Nimrod, I., Heine, S.J., 2006. Exposure to scientific theories affects women's math performance. Science 314, 435.
Dar-Nimrod, I., Heine, S.J., 2011. Genetic essentialism: on the deceptive determinism of DNA. Psychological Bulletin 137, 800–818.
Gaunt, R., 2006. Biological essentialism, gender ideologies, and role attitudes: what determines parents' involvement in child care. Sex Roles 55, 523–533.
Gelman, S.A., 2003. The Essential Child: Origins of Essentialism in Everyday Thought. Oxford University Press, Oxford, UK.
Giles, J.W., 2003. Children's essentialist beliefs about aggression. Developmental Review 23, 413–443.
Gil-White, F., 2001. Are ethnic groups biological species to the human brain? Essentialism in human cognition of some social groups. Current Anthropology 42, 515–554.
Haslam, N., Levy, S., 2006. Essentialist beliefs about homosexuality: structure and implications for prejudice. Personality and Social Psychology Bulletin 32, 471–485.
Haslam, N., Bastian, B., Bissett, M., 2004. Essentialist beliefs about personality and their implications. Personality and Social Psychology Bulletin 30, 1661–1673.
Haslam, N., Rothschild, L., Ernst, D., 2000. Essentialist beliefs about social categories. British Journal of Social Psychology 39, 113–127.
Hirschfeld, L.A., 1996. Race in the Making: Cognition, Culture, and the Child's Construction of Human Kinds. MIT Press, Cambridge, MA.
Hong, Y., Levy, S.R., Chiu, C., 2001. The contribution of the lay theories approach to the study of groups. Personality and Social Psychology Review 5, 97–105.
Keller, J., 2005. In genes we trust: the biological component of psychological essentialism and its relationship to mechanisms of motivated social cognition. Journal of Personality and Social Psychology 88, 686–702.
Leslie, S.J., Gelman, S.A., 2012. Quantified statements are recalled as generics: evidence from preschool children and adults. Cognitive Psychology 64, 186–214.
Medin, D., Ortony, A., 1989. Psychological essentialism. In: Vosniadou, S., Ortony, A. (Eds.), Similarity and Analogical Reasoning. Cambridge University Press, New York, pp. 179–195.
Phelan, J.C., 2005. Geneticization of deviant behavior and consequences for stigma: the case of mental illness. Journal of Health and Social Behavior 46, 307–322.
Rothbart, M., Taylor, M., 1992. Category labels and social reality: do we view social categories as natural kinds? In: Semin, G.R., Fiedler, K. (Eds.), Language and Social Cognition. Sage, London, UK, pp. 11–36.
Verkuyten, M., 2003. Discourses about ethnic group (de-)essentialism: oppressive and progressive aspects. British Journal of Social Psychology 42, 371–391.
Yzerbyt, V., Rocher, S., Schadron, G., 1997. Stereotypes as explanations: a subjective essentialistic view of group perception. In: Spears, R., Oakes, P.J., Ellemers, N., Haslam, S.A. (Eds.), The Social Psychology of Stereotyping and Group Life. Blackwell, Cambridge, UK, pp. 20–50.

Estimation: Point and Interval

Roger L Berger[†], Arizona State University, Phoenix, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Casella[†], R.L. Berger, volume 10, pp. 4744–4749, © 2001, Elsevier Ltd.

Abstract

In parametric statistical inference, knowledge about a population parameter yields knowledge about the entire population. Thus, methods of estimating population parameters are cornerstones to statistical analysis. Point estimators provide a single value as an estimate of a parameter. Set estimators provide a set of possible values. Set estimators quantify uncertainty about the parameter through the size of the set and the probability of the set covering the parameter. Various methods for deriving point and set estimators and various methods for evaluating estimators are discussed.

Introduction

When sampling from a population described by a density or mass function $f(x|\theta)$, knowledge of θ yields knowledge of the entire population. Hence, it is natural to seek a method of finding a good estimator of the point θ, that is, a good *point estimator*. However, a point estimator alone is not enough for a complete inference, as a measure of uncertainty is also needed. For that, we use a *set estimator* in which the inference is the statement that $\theta \in C$ where $C \subset \Theta$, Θ is the parameter space (set of all possible values of θ), and $C = C(\boldsymbol{x})$ is a set determined by the value of the data, $\boldsymbol{X} = \boldsymbol{x}$, observed. If θ is real-valued, then we usually prefer the set estimate C to be an interval. Our uncertainty is quantified by the size of the interval and its probability of covering the parameter.

Point Estimators

In many cases there will be an obvious or natural candidate for a point estimator of a particular parameter. For example, the sample mean is a natural candidate for a point estimator of the population mean. However, when we leave a simple case like this, intuition may desert us, so it is useful to have some techniques that will give reasonable candidates for consideration. Those that have stood the test of time include these.

Method of Moments

The method of moments (MOM) is, perhaps, the oldest method of finding point estimators, dating back at least to Karl Pearson in the late 1800s. One of the strengths of MOM estimators is that they are usually simple to use and almost always yield some sort of estimate. In many cases, unfortunately, this method yields estimators that may be improved upon.

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sample from a population with density or mass function $f(\boldsymbol{x}|\theta_1, \ldots, \theta_k)$. MOM estimators are found by equating the first k sample moments to the corresponding k population moments. That is, we define the sample moments by $m_j = \sum_{i=1}^{n} X_i^j/n$ and the population moments by $\mu_j(\theta_1, \ldots, \theta_k) = E_{\boldsymbol{\theta}}\boldsymbol{X}^j$ for $j = 1, \ldots, k$. Then we set $m_j = \mu_j(\theta_1, \ldots, \theta_k)$ and solve for $\theta_1, \ldots, \theta_k$. This solution is the MOM estimator of $\theta_1, \ldots, \theta_k$.

Maximum Likelihood

For a sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from $f(\boldsymbol{x}|\theta_1, \ldots, \theta_k)$, the likelihood function is defined by

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k)$$

The values of $\theta_1, \ldots, \theta_k$ that maximize this function are those parameter values for which the observed sample $\boldsymbol{x}$ is most likely and are called the maximum likelihood estimators (MLEs). If the likelihood function is differentiable (in θ_i), the MLEs can often be found by solving

$$\frac{\partial}{\partial\theta_i}\log L(\boldsymbol{\theta}|\boldsymbol{x}) = 0, \quad i = 1, \ldots, k$$

where the vector with coordinates $\frac{\partial}{\partial\theta_i}\log L(\boldsymbol{\theta}|\boldsymbol{x})$ is called the *score function* (see Schervish, 1995: Section 2.3).

Example: If $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a sample from a Bernoulli (p) population, the likelihood function is

$$L(p|\boldsymbol{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

and differentiating $\log L(p|\boldsymbol{x})$ and setting the result equal to zero gives the MLE $\widehat{p} = \sum_{i=1}^{n} x_i/n = \bar{x}$, the sample mean. This is also the MOM estimator.

If instead we have a sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from a binomial (k, p) population, where p is known and k is unknown, the likelihood function is

$$L(k|\boldsymbol{x}, p) = \prod_{i=1}^{n} \binom{k}{x_i} p^{x_i}(1-p)^{k-x_i}$$

and the MLE must be found by the numerical maximization. The MOM gives the closed form solution

$$\widehat{k} = \frac{\bar{x}^2}{\bar{x} - \sum_i (x_i - \bar{x})^2/n},$$

which can take on negative values. This illustrates a shortcoming of the MOM, one not shared by the MLE. Another, perhaps more serious shortcoming of the MOM estimator is that it may not be based on a sufficient statistic (*see* Statistical Sufficiency), which means it could be inefficient in not using all of the available information in a sample. In contrast, both MLEs and Bayes estimators (*see* Bayesian Statistics) are based on sufficient statistics.

[†] Deceased.

Bayes

In the Bayesian paradigm a random sample $X_1, \ldots, X_n$ is drawn from a population indexed by θ, where uncertainty about θ can be described by a probability distribution (called the *prior distribution*). After the sample is taken, the prior distribution is updated with this sample information. The updated prior is called the *posterior distribution*.

If we denote the prior distribution by $\pi(\theta)$ and the sampling distribution by $f(\boldsymbol{x}|\theta)$, then the posterior distribution, the conditional distribution of θ given the sample, $\boldsymbol{x}$, is

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)},$$

where $m(x) = \int f(x|\theta)\pi(\theta)\mathrm{d}\theta$ is the marginal distribution of $\boldsymbol{x}$.

Example: Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sample from a Bernoulli (p) population. Then $Y = \sum_i X_i$ is binomial (n, p). If p has a beta (α, β) prior distribution, that is

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1},$$

the posterior distribution of p given y is

$$\begin{aligned}\pi(p|y) &= \frac{f(y|p)\pi(p)}{m(y)} \\ &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1},\end{aligned}$$

which is a beta distribution with parameters $y+\alpha$ and $n-y+\beta$. The posterior mean, a common Bayes estimator of p, is

$$\widehat{p}_B = \frac{y+\alpha}{\alpha+\beta+n}.$$

Evaluating Point Estimators

There are many methods of deriving point estimators (robust methods, least squares, estimating equations, invariance), but the three in Section Point Estimators are among the most popular. No matter what method is used to derive a point estimator, it is important to evaluate the estimator using some performance criterion.

One way of evaluating the performance of a point estimator W of a real-valued parameter θ is through its *mean squared error* (MSE), defined by

$$\mathrm{MSE}_\theta W = E_\theta(W-\theta)^2 = \mathrm{Var}_\theta W + (E_\theta W - \theta)^2.$$

Defining the bias of a point estimator by $\mathrm{Bias}_\theta W = \mathrm{E}_\theta W - \theta$ yields $\mathrm{MSE}_\theta W = \mathrm{Var}_\theta W + (\mathrm{Bias}_\theta W)^2$. An estimator whose bias is identically (in θ) equal to zero is called *unbiased*.

For an unbiased estimator we have $\mathrm{MSE}_\theta W = \mathrm{Var}_\theta W$; if an estimator is unbiased, its MSE is equal to its variance. If $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a sample from a population with mean θ and variance σ^2, the sample mean is an unbiased estimator of θ $(E_\theta \overline{X} = \theta)$, and

$$\mathrm{MSE}_\theta \overline{X} = \mathrm{Var}_\theta \overline{X} = \frac{\sigma^2}{n}.$$

Controlling bias does not guarantee that MSE is minimized. In particular, it is sometimes the case that a trade-off occurs between variance and bias. For example, in sampling from a normal population with variance σ^2, the usual unbiased estimator of σ^2, $S^2 = \sum_{i=1}^n (X_i - \overline{X})^2/(n-1)$, has $\mathrm{MSE}_\sigma S^2 = 2\sigma^4/(n-1)$. An alternative estimator of σ^2 is the MLE $\widehat{\sigma}^2 = \sum_{i=1}^n (X_i - \overline{X})^2/n = (n-1)S^2/n$. This is a biased estimator of σ^2 with

$$\mathrm{MSE}_\sigma \widehat{\sigma}^2 = \frac{2n-1}{n^2}\sigma^4 < \frac{2}{n-1}\sigma^4 = \mathrm{MSE}_\sigma S^2,$$

showing that $\widehat{\sigma}^2$ has smaller MSE than S^2. Thus, by trading off variance for bias, the MSE is improved.

Measuring performance by the squared difference between the estimator and a parameter is a special case of a function called a *loss function*. The study of the performance and optimality of estimators evaluated through loss functions is a branch of decision theory. In addition to MSE, based on squared error loss, another popular loss function is absolute error loss, $L(\theta, W) = |W - \theta|$. Both of these loss functions increase as the distance between θ and W increases, with minimum value $L(\theta, \theta) = 0$. That is, the loss is minimum if the estimator is correct.

In a decision theoretic analysis an estimator is evaluated through its *risk function*. For an estimator W of θ, the risk function is $R(\theta, W) = \mathrm{E}_\theta L(\theta, W)$; the risk function is the average loss. If the loss is squared error, the risk function is the MSE.

Using squared error loss, the risk function (MSE) of the binomial Bayes estimator of p is

$$\begin{aligned}\mathrm{MSE}_p \widehat{p}_B &= \mathrm{Var}_p \widehat{p}_B + \left(\mathrm{Bias}_p \widehat{p}_B\right)^2 \\ &= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2.\end{aligned}$$

In the absence of good prior information about p we might choose α and β to make the risk function of $\widehat{p}_B$ constant (called an *equalizer rule*). The solution is $\alpha = \beta = \sqrt{n/4}$, yielding

$$\mathrm{MSE}_p \widehat{p}_B = \frac{n}{4(n+\sqrt{n})^2}.$$

We can also use a Bayesian approach to the problem of loss function optimality, where we would use the prior distribution to compute an average risk, $\int_\Theta R(\theta, W)\pi(\theta)\mathrm{d}\theta$, known as the *Bayes risk*. We then find the estimator that yields the smallest value of the Bayes risk. Such an estimator is called the Bayes rule with respect to a prior π.

To find the Bayes decision rule for a given prior π we write the Bayes risk as

$$\int_\Theta R(\theta, W)\pi(\theta)\mathrm{d}\theta = \int_x \left[\int_\Theta L(\theta, W(x))\pi(\theta|x)\mathrm{d}\theta\right] m(x)\mathrm{d}x,$$

where the quantity in square brackets is the expected value of the loss function with respect to the posterior distribution, called the *posterior expected loss*. It is a function only of $\boldsymbol{x}$ and not a function of θ. Thus, for each $\boldsymbol{x}$, if we choose the estimate $W(\boldsymbol{x})$ to minimize the posterior expected loss, we will minimize the Bayes risk, and $W(\boldsymbol{x})$ is the Bayes rule.

For squared error loss, the posterior expected loss is minimized by the mean of the posterior distribution. For absolute error loss, the posterior expected loss is minimized by the median of the posterior distribution. If $\boldsymbol{X} = (X_1, \ldots, X_n)$

is a sample from a normal population, with mean θ and variance σ^2 ($n(\theta, \sigma^2)$), and the prior on θ is $n(\mu, \tau^2)$, the posterior mean is

$$E(\theta|x) = \frac{\tau^2}{\tau^2 + (\sigma^2/n)}\bar{x} + \frac{\sigma^2/n}{\tau^2 + (\sigma^2/n)}\mu.$$

Because the posterior distribution is normal, it is symmetric and the posterior mean is the Bayes estimator for both squared error and absolute error loss. The posterior mean in our binomial/beta example, $\widehat{p}_B = (y + \alpha)/(\alpha + \beta + n)$, is the Bayes estimator for squared error loss.

We can loosely group evaluation criteria into large sample (asymptotic) methods and small sample methods. Our calculations using MSE and risk functions illustrate small sample methods. In large samples, MLEs typically perform very well, being asymptotically normal and efficient, that is, attaining the smallest possible variance. Other types of estimators that are derived in a similar manner (for example, M-estimators – *see* Robustness in Statistics) also share good asymptotic properties. For a detailed discussion see Casella and Berger (2001), Lehmann (1999), Stuart et al. (1999), or Lehmann and Casella (1998: Chapter 6).

Interval Estimation

Reporting a point estimator of a parameter θ only provides part of the story. The story becomes more complete if an assessment of the error of estimation is also reported. Informally, this can be accomplished by giving an estimated standard error of the estimator, and, more formally, this becomes the reporting of an interval estimate. If $\boldsymbol{X} = \boldsymbol{x}$ is observed, an *interval estimate* of a parameter θ is a pair of functions, $L(\boldsymbol{x})$ and $U(\boldsymbol{x})$, for which the inference $\theta \in (L(x), U(x))$ is made. The *coverage probability* of the random interval $(L(\boldsymbol{X}), U(\boldsymbol{X}))$ is the probability that $(L(\boldsymbol{X}), U(\boldsymbol{X}))$ covers the true parameter, θ, and is denoted by $P_\theta(L(\boldsymbol{X}) < \theta < U(\boldsymbol{X}))$.

By definition the coverage probability depends on the unknown θ, so it cannot be reported. What is typically reported is the *confidence coefficient*, the infimum of the coverage probabilities, $\inf_\theta P_\theta(L(X) < \theta < U(X))$.

If $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a sample from a population with mean θ and variance σ^2, a common interval estimator for θ is

$$\bar{x} - 2\frac{s}{\sqrt{n}} < \theta < \bar{x} + 2\frac{s}{\sqrt{n}},$$

where $\bar{x}$ is the sample mean and s is the sample standard deviation. The validity of this interval can be justified from the Central Limit Theorem, because

$$\frac{\overline{X} - \theta}{S/\sqrt{n}} \to n(0, 1),$$

the standard normal distribution. We then see that the coverage probability (and confidence coefficient) of this interval is approximately 95%.

The above interval is a large sample interval, because its justification is based on an asymptotic property. There are many methods for constructing interval estimators that are valid in small samples, including these.

Inverting a Test Statistic

There is a correspondence between acceptance regions of hypothesis tests (*see* Hypothesis Testing in Statistics) and confidence sets, summarized in the following theorem.

Theorem 1: For each $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a level α test of H_0: $\theta = \theta_0$. For each sample point $\boldsymbol{x}$, define a set $C(\boldsymbol{x})$ in the parameter space by

$$C(x) = \{\theta_0 : x \in A(\theta_0)\}.$$

Then the random set $C(\boldsymbol{X})$ is a confidence set with confidence coefficient $1 - \alpha$. Conversely, let $C(\boldsymbol{X})$ be a $1 - \alpha$ confidence set. For each $\theta_0 \in \Theta$, define

$$A(\theta_0) = \{x : \theta_0 \in C(x)\}.$$

Then $A(\theta_0)$ is the acceptance region of a level α test of H_0: $\theta = \theta_0$.

Example: If $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a sample from an $n(\theta, \sigma^2)$ population, with σ^2 known, the level α test of H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$ will accept the null hypothesis, if $|\bar{x} - \theta_0|/(\sigma/\sqrt{n}) < z_{\alpha/2}$. This inequality can be equivalently written as

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \theta_0 < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

This interval of θ values, for which the null hypothesis is accepted at level α, is a $1 - \alpha$ confidence interval for θ.

Pivotal Intervals

An elegant method of constructing set estimators is the use of pivotal quantities (Barnard, 1949). A random variable, $Q(\boldsymbol{X}, \theta) = Q(X_1, \ldots, X_n, \theta)$, is a pivotal quantity (or pivot), if the distribution of $Q(\boldsymbol{X}, \theta)$ is independent of all parameters. If we find a set C such that $P(Q(\boldsymbol{X}, \theta) \in C) = 1 - \alpha$, then the set $C(\boldsymbol{X}) = \{\theta: Q(\boldsymbol{X}, \theta) \in C\}$ has coverage probability $1 - \alpha$.

In *location* and *scale* cases, once we calculate the sample mean $\overline{X}$ and the sample standard deviation S, examples of pivots are:

Form of pdf	*Type of pdf*	*Pivot*
$f(x - \theta)$	Location	$\overline{X} - \theta$
$f(x/\sigma)/\sigma$	Scale	S/σ
$\dfrac{f((x - \theta)/\sigma)}{\sigma}$	Location–scale	$\dfrac{\overline{X} - \theta}{S}$

In general, differences are pivotal for location problems, while ratios (or products) are pivotal for scale problems. *see* Fiducial and Structural Statistical Inference.

Example: Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sample from an exponential (λ) population, and let $T = \sum_{i=1}^n X_i$. T has a gamma (n, λ) distribution. This gamma pdf is

$$f(t|\lambda) = \Gamma(n)\left(\frac{t}{\lambda}\right)^{n-1} e^{-t/\lambda}\frac{1}{\lambda}.$$

In this gamma pdf t and λ appear together as t/λ, and, in fact, the gamma (n, λ) pdf is a scale family in λ. If $Q(T, \lambda) = 2T/\lambda$, then $Q(T, \lambda)$ has a gamma $(n, 2)$ (also called χ^2_{2n}) distribution,

which does not depend on λ. Hence, the quantity $Q(T, \lambda)$ is a pivot, and a $1-\alpha$ pivotal interval is

$$\frac{2T}{\chi^2_{2n,\alpha/2}} < \lambda < \frac{2T}{\chi^2_{2n,1-\alpha/2}},$$

where $P(\chi^2_{2n} > \chi^2_{2n,\alpha}) = \alpha$.

Bayesian Intervals

If $\pi(\theta|\boldsymbol{x})$ is the posterior distribution of θ given $\boldsymbol{X} = \boldsymbol{x}$, then for any set $A \subset \Theta$ the posterior probability of A is

$$P(\theta \in A|x) = \int_A \pi(\theta|x)\mathrm{d}\theta.$$

If $A = A(\boldsymbol{x})$ is chosen so that this posterior probability is $1-\alpha$, then A is called a $1-\alpha$ *credible set* for θ.

The interpretation of the Bayes interval estimator is different from the classical intervals. In the classical approach to assert 95% coverage is to assert that in 95% of repeated experiments the realized intervals will cover the true parameter. In the Bayesian approach a 95% coverage means that the probability is 95% that the parameter is in the realized interval. In the classical approach the randomness comes from the repetition of experiments, while in the Bayesian approach the randomness comes from uncertainty about the value of the parameter (summarized in the prior and posterior distributions).

Example: Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sample from an Poisson (λ) population, and assume that λ has a gamma (a, b) prior, where a is an integer. The posterior distribution of λ is

$$\pi(\lambda|x) \text{ is gamma}\left(a + \sum_{i=1}^{n} x_i, \frac{b}{nb+1}\right).$$

Thus the posterior distribution of $[2(nb+1)/b]\lambda$ is $\chi^2_{2(a+\sum_i x_i)}$, and a $1-\alpha$ Bayes credible interval for λ is

$$\frac{\chi^2_{2(a+\sum_i x_i),1-\alpha/2}}{2(nb+1)/b} < \lambda < \frac{\chi^2_{2(a+\sum_i x_i),\alpha/2}}{2(nb+1)/b}.$$

We can form a Bayes credible set by using the *highest posterior density* (HPD) region of the parameter space, also. That is, choose c so that the set $C(x) = \{\lambda: \pi(\lambda|x) > c\}$ satisfies

$$1-\alpha = P(\lambda \in C(x)|x).$$

Such a construction is optimal in the sense of giving the shortest interval for a given $1-\alpha$, although if the posterior is multimodal the set may not be an interval.

Other Intervals

We have presented two-sided parametric confidence intervals that are constructed to cover a parameter. Other types of intervals include one-sided intervals, distribution-free intervals, prediction intervals, and tolerance intervals.

One-sided intervals are those in which only one endpoint is estimated, such as $\theta \in (L(X), \infty)$. Distribution-free intervals are intervals whose probability guarantee holds with few (or no) assumptions on the underlying population. The other two interval definitions, together with the usual confidence interval, provide us with a hierarchy of inferences, each more stringent than the previous.

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sample from a population with cdf $F(\boldsymbol{x}|\theta)$. If $C(\boldsymbol{x}) = (L(\boldsymbol{x}), U(\boldsymbol{x}))$ is an interval, for a specified value $1-\alpha$ it is a:

1. *confidence* interval if, for all θ, $P_\theta(L(X) < \theta < U(X)) \geq 1-\alpha$;
2. *prediction* interval if, for all θ, $P_\theta(L(X) < X_{n+1} < U(X)) \geq 1-\alpha$, where X_{n+1} is an independent, new observation;
3. *tolerance* interval if, for all θ and for a specified value p, $P_\theta(F(U(X)|\theta) - F(L(X)|\theta) \geq p) \geq 1-\alpha$.

So a confidence interval covers a parameter, a prediction interval covers a new observation, and a tolerance interval covers a proportion of the population. Each gives a different inference with the appropriate one being dictated by the problem at hand. Vardeman (1992) discusses the importance of each of these intervals.

Conclusions

Point estimation is one of the cornerstones of statistical analysis and the basic element on which many inferences are based. Inferences using point estimators gain statistical validity when they are accompanied by an interval estimate, providing an assessment of the uncertainty. We have mainly discussed parametric point and interval estimation, where we assume that the underlying model is correct. Such an assumption can be questioned, and considerations of nonparametric or robust alternatives can address this (*see* Robustness). For more on these subjects see, for example, Boos and Stefanski (2013), Hettmansperger and McKean (1998), or Staudte and Sheather (1990). Full treatments of parametric point and interval estimation can be found in Casella and Berger (2001), Stuart et al. (1999), or Schervish (1995).

Acknowledgment

This article is a revised version of an article on the same topic in the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, which was coauthored with the late George Casella, University of Florida.

See also: Bayesian Statistics; Likelihood in Statistics; Quantile Regression; Statistical Identification and Estimability; Statistical Sufficiency.

Bibliography

Barnard, G.A., 1949. Statistical inference (with discussion). Journal of the Royal Statistical Society, Series B 11, 115–139.

Boos, D.D., Stefanski, L.A., 2013. Essential Statistical Inference, Theory and Methods. Springer, New York.

Casella, G., Berger, R.L., 2001. Statistical Inference, second ed. Wordsworth/Brooks Cole, Pacific Grove, CA.

Hettmansperger, T.P., McKean, J.W., 1998. Robust Nonparametric Statistical Methods. Wiley, New York.

Lehmann, E.L., 1999. Elements of Large-Sample Theory. Springer-Verlag, New York.
Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation, second ed. Springer-Verlag, New York.
Schervish, M.J., 1995. Theory of Statistics. Springer-Verlag, New York.
Staudte, R.G., Sheather, S.J., 1990. Robust Estimation and Testing. John Wiley, New York.
Stuart, A., Ord, J.K., Arnold, S., 1999. Kendall's Advanced Theory of Statistics, Vol. 2A, Classical Inference and the Linear Model, sixth ed. Wiley, London.
Vardeman, S.B., 1992. What about the other intervals? The American Statistician 46, 193–197.This page intentionally left blank

Ethical Codes, Professional: Business Codes

Josef Wieland, Zeppelin University, Friedrichshafen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Business codes are a functional equivalent to national or internal deficits of regulation by way of directing companies to self-binding and self-control. They are the foundation of collective and individual perception and provide shared mental models for cooperation, and they determine and define the identity of those corporations that bind themselves to standards of behavior. They show a corporation's responsibility and can assure economic benefit as well as reduce the risk of losing economic and social reputation. These effects can be reached only by actually maintaining the standards of behavior within the business routine of the company. For that purpose, they have to be implemented and integrated into the operative processes of the organization and its management. One of the most important instruments in this respect is values management systems.

Business codes are formal or informal standards of behavior that give structure and direction to economic action and behavior. They codify the values of a corporation, an industry, or a whole sector, and thereby create identity and shape one's perception. Conceived of as self-binding acts, business codes enable and provide constraints to collective actions within the economy. In Western economies, as well as in the emergence of a global economy, they play an important role as governance structures because of, and insofar as they make possible, the expansion of economical and societal cooperation.

History and Actuality

Voluntary standards of behavior in the economy as a means of self-binding and constraint can be found in early history. Their emergence is connected to the guilds of craftsmen and merchants in the Middle Ages, where they regulated behavior toward the other members of the guilds, customers, coworkers, and employees. In the beginning, they were informal and derived from tradition, common rites, and the customs belonging to a trade. With the expansion of long-distance trading, these standards were laid down and were adopted by traveling merchants. Thus, they became increasingly formal, which meant that they became enforceable by third parties. These early business codes served to ensure contracts concerning future delivery and payment of goods and were enforceable by guild courts and the threat of ostracism (Mitchell, 1969).

To the extent that national and international business laws were established and enforceable by government bodies during the nineteenth century, private standards of behavior based on the threat of ostracism and loss of reputation gradually started to lose their importance. But in the 1970s, a renaissance of private standards of behavior emerged from two main sources. The first was the internationalization and globalization of markets and production networks, which were and still are unaccompanied by developments of political, juridical, and cultural institutions and their enforceability. Such business codes include maintaining human rights as well as social and legal standards in cooperative relationships, for example, with customers, suppliers, and partners, and they define common perspectives in areas where different values standards as well as cross-cultural values conflicts hinder the establishment of transcultural teams (cf Wieland, 2010a).

Business codes are an element of an emerging normative order of globally accepted norms of good behavior (Dunning, 2003). In developing global standards, new players like the BRICS states (i.e., Brazil, Russia, India, China, and South Africa) are very likely to play an increasingly decisive role. Therefore, business codes must reflect cultural diversity and at the same time transcultural values. Real globalization of the economy and society, however, still is accompanied by the twofold problem of institutional and organizational deficit. In contrast to international standards, which are decided on by individual states or intergovernmental organizations and subsequently enforced throughout the world, the design and establishment of global and globally accepted standards and rules is still in a development process. This is due, not least, to a lack of internationally recognized and motivated organizations capable of promoting such a development process for globally accepted norms to regulate fair individual and organizational behavior. The Human Guidelines on Business and Human Rights, the Organization for Economic Cooperation and Development (OECD) Guidelines for Multinational Enterprises, the United Nations Global Compact, or the International Labor Organization 'core norms' are exemplary and valuable examples of the emergence of a global normative order. In this context, multistakeholder dialogues as standard setters are of utmost importance (Wieland, 2011).

The second source was the organizational decentralization of corporations, which led to an increase in 'gray areas' and decreasing possibilities of utilizing direct monitoring and control mechanisms. Here, business codes have to define a common set of values to give the company identity and a behavioral orientation. To reach this aim, it needs to form vision, mission, and value statements, compliance standards, and rules for dealing with the company's stakeholders (customers, employees, suppliers, environment) (Weaver, 1993). Such standards of behavior can be enforced by sanctions and incentives within the company. With regard to

customers and suppliers, the companies' reputation and brand names are important enforcing mechanisms.

Minimum Standards, Problem Standards, and Corporate Standards

Business codes are characterized by self-binding possibilities of private enforcement and by the identity of the actors. Self-binding by voluntary constraints of action, certainty of expectation in contracts by enforcement possibilities (sanctions, incentives, reputation), and the creation of corporate identity by means of codified and realized values in business routine make standards of behavior a functional equivalent to national regulations and laws. Whenever national or supernational regulations and their enforcement do not exist or are not feasible, regardless of the identity of an actor, private standards of behavior become a dominating issue of economic and societal governance. Therefore, they are known as 'soft laws' (Fatouros, 1994). By creating certainty of expectation for economic action in societies that are characterized by division of labor, business codes directly influence the state of welfare in any given society. With these reliable and trustworthy rules, more cooperation and more transactions are made possible than without them.

The distinction between minimum standards, problem standards, and corporate standards enables us to group modern business codes along with the kinds of cooperation and transactions that are steered by them.

Minimum Standards

Minimum standards are instruments serving the governance of global business transactions – for example, Guidelines for Multinational Enterprises (OECD, update 2011) and Corporate Governance Guidelines (OECD, 1991/1998/2004), a Code of Conduct on Transnational Corporations (United Nations, 1994), and ISO 26.000 Social Responsibility (SR) (ISO 2010), to name just a few. They define minimum requirements for such areas as working conditions, involvement of unions, safety at work, consumer protection, competition, technology transfer, investment in developing countries, corruption, and human rights.

Minimum standards emerge from bargaining processes between national governments, national and international organizations of employers and unions, and any nongovernmental organizations (NGOs) involved, which are directed by a recognized international organization (Fatouros, 1994). Their status is precarious. They are voluntary but not arbitrary. These agreements are obligatory for the signing parties, although they cannot be enforced by law. Whether it is possible to implement them therefore depends on the voluntary efforts of the contractors as well as on whether it is possible to take action against rule breakers or to damage their reputation in a cost-generating way. Their strength lies in the fact that they provide standards that possess some chance of realization, although there is no suitable international law available or accessible. Their weak points are the difficulty of implementing and controlling them. Nevertheless, they are a significant component of economic globalization because they are a functional equivalent to the still underdeveloped international business law. Hence, they create a certain formal and political certainty within the world economy that is a crucial demand for worldwide prosperity.

Problem Standards

Problem standards define proposals of behavior that help to solve certain problems and that go far beyond the rules of the game asserted through the minimum standards. One example from the area of ecological sustainability is the 'responsible care' process of the chemical industry's international associations and companies. From the area of social responsibility, the efforts to abolish child labor and forced labor according to the Council on Economic Priorities' standard Social Accountability 8000 may be cited. In the field of sustainability, Rainforest Alliance Certified is a pertinent example. Problem standards are highly ambitious compared with the potentialities and mentality of many companies. Hence, they are accepted by only a few companies, mostly by those that have elaborated them, usually in cooperation with NGOs. This may happen in associations or chambers of commerce or in loose assemblies of those companies for whose business actions the problem in question is instantly relevant. Their strength lies in their model character and in the fact that they make practical experiences concerning these standards possible. Their weakness lies in what is mostly very difficult control and in the possible competitive disadvantages for those companies that do not succeed in overcoming the costs of implementing and keeping the standards through prices.

Corporate Standards

Corporate standards codify the constitutive and action-directing values of a corporation. As codes of conduct or codes of ethics, they are components of corporate governance, that is, of the leading, steering, and controlling mechanisms of an economic organization. They also are expressions of the distinct philosophy and policy of the corporation. The values of a corporation can be classified as performance values (e.g., profit, quality), cooperative values (e.g., loyalty, team spirit), communication values (e.g., respect, transparency), and moral values (e.g., integrity, fairness) (Wieland, 2000). A firm gains identity and distinctiveness as soon as a specific values mix drawn from these four classes of values endows it with a particular profile. Particularity here means that it fits best with the promises of the company, its tradition and history, and the aspired direction of future developments. As a rule, these business codes provide regulations for coping with suppliers and customers, shareholders, and employees, as well as rules of law and of environment protection. They are part of an overall values management system by which the business codes are operated and pursued in corporate practice. Normally, this is executed by means of internal and external communication, by training, by ethics officers, and by being incorporated in labor contracts and other policies and procedures, and the system is controlled by monitoring audits (Weaver et al., 1999; Wieland, 2010b).

Values management systems thus are a fundamental presupposition to solve all the potential conflicts of interest in everyday business (Davis and Stark, 2001).

The systems themselves can be divided into law-driven and values-driven ones (Paine, 1994). Although the first are compliance programs that focus on the observation of the rules of law and are organized by the legal or compliance department of a corporation, the latter as values management systems aim at a comprehensive development of the corporate culture specifically concerning relations with the totality of stakeholders. Codes of ethics therefore never make any statements about existing matters of fact, and they never stand in for the moral consensus of all parties that are interested in the company. It is just the other way around. They define the values of an organization against all stakeholders and declare them conditional for further interaction. They express the normative side of the firm. As conditions of cooperation with and within an organization, these values are shaped like commitments of performance. The company commits itself to these values and promises to continually work on the fulfillment of its commitment. This makes it obligatory for all stakeholders also to bind themselves to the values. In the last decade of the twentieth century, about 80% of the big US companies had adopted such codes (Center for Business Ethics, 1992; Stevens, 2008; Kaptein, 2008). In the second decade of the twenty-first century, a code of ethics or code of conduct for big and multinational companies has become a standard in Europe.

Apart from the reasons for the renaissance of business ethics mentioned earlier, another important factor is legal developments. Because of the introduction in the United States and the United Kingdom of organizational liability in cases of deviant behavior of employees, incentive systems have been installed that demonstrate the due diligence and make it financially worthwhile to document before the court the implementation and communication of a code of ethics, so that in case of an individual breach of law, the organization cannot be held responsible for it (U.S. Sentencing Commission, 1995). Legal developments similar to this can be observed in Europe (U.K. Bribery Act, 2011). In this context, business codes aim to achieve a higher level of transparency and thereby enhance the means of control in the corporation. The strength of standards of behavior in corporations is that they are highly specific and integrate the involved actors in maintaining them by the implementation of incentives. They cause problems, however, for any form of control and by the fact that they have to be introduced into the operative actions (policies and procedures) of the corporation in a relevant way.

Functions and Effect

Although all the three kinds of business codes refer to different types of cooperation and transactions, they are identical in regard to their functions and effectiveness. These can be differentiated as follows:

1. Business codes are a functional equivalent to national or internal deficits of regulation by way of directing companies to self-binding and self-control.
2. Business codes determine and define the identity of those corporations that bind themselves to standards of behavior. These corporations are distinguished from other organizations in the public arena of realization and communication.
3. Business codes are the basis, as well as a criterion of selection, for corporate decisions. Hence, they define the range of liability a company claims beyond existing law.
4. Business codes are the foundation of collective and individual perception and provide shared mental models for cooperation.
5. Business codes internally and externally indicate the behavioral preferences of an organization and at the same time communicate a challenge to all its stakeholders to bind themselves to them.
6. Business codes provide certainty of expectation in business relations and make possible the building up of mutual trust by creating a check against informational uncertainty and opportunistic behavior.
7. Their prevalence, on the one hand, relies on economic benefits that arise from the acceptance of standards of behavior. On the other hand, one runs the risk of losing economic and social reputation in any case of violation of the legal and moral obligations to which they are committed.
8. The aspects mentioned – self-binding, identity, responsibility, shared mental models, certainty of expectation, building up trust, and reputation – constitute a system of mutual effects that lead to an expansion of economic and societal cooperation and to accompanying gains from trade.
9. These effects can be reached only by actually maintaining the standards of behavior within the business routine of the company. For that purpose, they have to be implemented and integrated into the operative processes of the organization and its management. One of the most important instruments in this respect is values management systems.

See also: Economics and Ethics; Ethics and Values; Ethics for Biomedical Research Involving Humans: International Guidelines; Kantian Ethics and Politics; Virtue Ethics; Weber, Max (1864–1920); Wisdom, Psychology of.

Bibliography

Center for Business Ethics at Bentley College, 1992. Installing ethical values in large corporations. Journal of Business Ethics 11, 863–867.

Davis, M., Stark, A., 2001. Conflict of Interest in the Profession. Oxford University Press, Oxford.

Dunning, John H., 2003. Making Globalization Good: The Moral Challenges of Global Capitalism. Oxford University Press, Oxford.

Fatouros, A. (Ed.), 1994. Transnational Corporations: The International Legal Framework. Routledge, London.

Kaptein, M., 2008. The Living Code: Embedding Ethics into the Corporate DNA. Remouf Publishing Company.

Mitchell, W., 1969. An Essay on the Early History of Law Merchant. Burt Franklin Press, New York.

Paine, L.S., 1994. Managing for organizational integrity. Harvard Business Review 72, 106–117.

Stevens, B., 2008. Corporate ethical codes: effective instruments for influencing behavior. Journal of Business Ethics 78 (4), 601–609.

United States Sentencing Commission, 1995. Corporate Crime in America: Strengthening the 'Good Citizen' Corporation. United States Sentencing Commission, Washington, DC.

Weaver, G.R., 1993. Corporate codes of ethics: purpose, process, and content issues. Business and Society 32 (1), 44–58.

Weaver, G.R., Trevino, L.K., Cochran, Ph.L., 1999. Corporate ethics programs as control systems: influences of executive commitment and environmental factors. Academy of Management Journal 42 (1), 41–57.

Wieland, J., 2000. The ethics of governance. Business Ethics Quarterly 11 (1), 73–87.

Wieland, J., 2010a. Global economic ethic as transcultural management. In: Küng, H., Leisinger, K.M., Wieland, J. (Eds.), Manifesto Global Economic Ethic. Consequences and Challenges for Global Businesses. Deutscher Taschenbuch Verlag (dtv), München, pp. 208–222 (German/English).

Wieland, J., 2010b. How can we manage values? Practical guidelines. In: Küng, H., Leisinger, K.M., Wieland, J. (Eds.), Manifesto Global Economic Ethic. Consequences and Challenges for Global Businesses. Deutscher Taschenbuch Verlag (dtv), München, pp. 223–264 (German/English).

Wieland, J., 2011. How geopolitical change alters social and environmental standards the ISO 26000 SR case. In: UN Global Compact International Yearbook. Macondo Media Group, Münster.

Zadek, S., Pruzan, P., Evans, R. (Eds.), 1997. Building Corporate Accountability. Earthscan, London.

Ethical Consumers

Veronika A Andorfer, Göttingen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ethical consumers are not only concerned with the price and quality of products and services, they also consider the consequences of their consumption practices. Researchers and practitioners alternatively describe these consumers as 'green,' 'sustainable,' 'responsible,' 'socially conscious,' 'conscientious,' 'critical,' 'active,' 'political,' or 'moral.' At the beginning of the 1980s, scholars and professionals used the concept almost exclusively to describe green consumers being mindful of the environmental impact of their day-to-day consumption patterns. Over time, however, it has come to denote a much broader and growing field of contemporary consumer concerns.

Ethical consumers are not only concerned with the price and quality of products and services, they also consider the consequences of their consumption practices. Researchers and practitioners alternatively describe these consumers as 'green,' 'sustainable,' 'responsible,' 'socially conscious,' 'conscientious,' 'critical,' 'active,' 'political,' or 'moral.' At the beginning of the 1980s, scholars and professionals used the concept almost exclusively to describe green consumers being mindful of the environmental impact of their day-to-day consumption patterns. Over time, however, it has come to denote a much broader and growing field of contemporary consumer concerns. Typically, the ethical dimension involves environmental, political, social, or religious aspects when consumers purchase, use, modify, and discard goods and services (Harrison et al., 2007). Cases of ethical consumers are usually attributed to the industrialized markets of North America and Europe.

Standard examples of ethical consumption include buying organic, regional/local, or fair trade produce; boycotting commodities manufactured under dubious working conditions and supporting sweat-free companies; avoiding products tested on animals; investing in firms that operate in a socially responsible manner; recycling waste; compensating CO_2 emissions from ground and air travel; preferring a vegetarian or vegan diet; voluntarily abstaining from purchasing consumer goods; and observing buy nothing days.

As diverse as the manifestations of ethical consumers and their consumption practices in everyday life are the scholarly contributions on the topic. Researchers from wide-ranging disciplines such as agricultural sciences, anthropology, business and marketing, consumer studies, economics, environmental science, philosophy, geography, political science, (social) psychology, sociology, and sustainability studies have explored the individual characteristics and motivations of ethical consumers and the structural determinants of their consumption practices.

Characteristics of Ethical Consumers

Who Are They?

Profiling the ethical consumer is difficult. Researchers find only few sociodemographic variables that consistently link the social bases of ethical consumers (Devinney et al., 2010; Papaoikonomou et al., 2011; Peattie, 2010). Empirical evidence suggests that ethical consumers tend to be older, female, well educated, and have higher-than-average incomes (e.g., Cotte and Trudel, 2009; Papaoikonomou et al., 2011 Tallontire et al., 2001). Researchers have not been successful in identifying ethical consumers on the basis of their personality and personal motivations. Ethical consumers' practices do, however, seem to be guided by religious beliefs and values of universalism, solidarity, and altruism (e.g., Doran, 2009; Doran and Natale, 2011; Peattie, 2010).

What Do They Do?

The diverse profile of ethical consumers is matched by the diversity of practices they pursue in everyday life to demonstrate their principles. To classify ethical consumers' practices, scholars commonly differentiate 'buycotts' (Friedman, 1996) from 'boycotts.' Boycotting – also called negative buying – denotes consumers' refusal to buy products and services that harm the environment, communities, animals, or workers. It may also involve consumers refusing to buy products from a specific company to punish it for its unethical conduct (e.g., North American and European consumers boycotting Nestlé products from 1977 to the present because of the company's aggressive marketing strategy to promote breast milk substitutes in developing countries). Buycotting – or positive buying – involves consumers deliberately choosing ethical products and services over conventional alternatives. In doing so, consumers reward a company for its ethical policies.

In addition to the dichotomy of boycotts and buycotts, scholars identify a variety of other types of ethical consumption practices. 'Consumer action' commonly includes aspects such as lobbying policy makers, writing petitions, and initiating campaigns (e.g., Tallontire et al., 2001: p. 7). Harrison et al. (2007: p. 3) identify three additional types: (1) 'ethical screening/rating of producers' such as publishing ethical consumer magazines to inform consumers about ethical/unethical cooperate behavior; (2) 'relationship purchasing,' for instance, community-supported agriculture; and (3) 'anti-consumerism or sustainable consumerism' understood as consumers trying to avoid unsustainable products entirely. In academic discussions, this latter type is sometimes specified as

'voluntary simplicity/ethical simplifiers' meaning consumers voluntarily cutting down on their total consumption and choosing to live a simpler life to promote their ethical objectives. By contrast, 'sustainable consumer habits' summarizes selected eco-friendly consumer practices such as recycling waste or using water-saving house appliances (Papaoikonomou et al., 2011: p. 213).

A different classification scheme categorizes ethical consumption practices according to ethical consumers' primary areas of concern. These fall broadly into the categories, 'people,' 'politics,' and 'environment' (Yates, 2010: p. 196). Examples include issues such as consumerist lifestyles; free trade in tropical commodities; abusive labor practices; local economy; repressive regimes; environmental sustainability; biodiversity, nature, and endangered species; genetically modified organisms; and animal welfare (Starr, 2009).

Ethical consumers' day-to-day actions are manifold and the list of potential paths of ethical consumption grows continuously. As such, the terms 'ethical consumption' and 'ethical consumers' seem to be incoherent umbrella terms for a wide variety of phenomena in contemporary consumer society rather than a clearly defined set of consumers (Lewis and Potter, 2011). Except for the classical boycott–buycott dichotomy, scholars and practitioners have not yet established a commonly accepted typology of ethical consumption behaviors.

Where Are They?

Scholars and practitioners picture ethical consumers almost exclusively in the industrialized societies of North America, Western Europe, and Scandinavia. However, their geographical scope widens as new middle class consumers arise in the emerging economies of Eastern Europe, Asia, and the global South.

Cases of ethical consumers have mostly been studied in the United States, Canada, the UK, Western and Northern Europe, and to a lesser extent in Australia and the New Zealand (Andorfer and Liebe, 2012; Cotte and Trudel, 2009; Peattie, 2010). Academic work usually focuses on single countries. Few scholars have conducted cross-national research projects.

General population surveys and opinion polls suggest high levels of self-reported ethical consumption behavior in North America and Western Europe. In the first wave of the European Social Survey in 2002–03, for instance, around 40–50% of respondents from Scandinavia, Switzerland, Germany, the UK, and France indicated that they purchased products for political, ethical, or environmental reasons in the 12 months prior to the survey. Around 30% of Swedish, Swiss, British, and French respondents report to have boycotted products. Only 10–15% of respondents from Southern and Eastern European countries gave these answers (Koos, 2012; Yates, 2010). A 2010 survey in the United States found that around 50% of respondents indicated that they deliberately buy eco-friendly products (PEW Research Center, 2010).

These data need to be viewed with caution for two reasons: First, the high levels of self-reported buycotting behavior found in most general population surveys usually exceed the market shares of ethical products and services by far. For example, the Food and Agriculture Organization of the United Nations (FAO, 2009) reports that fair trade coffee – the oldest and most widely available fair trade product – accounts for less than 1% of the total European coffee market. In addition, market shares vary considerably between countries. For example, fair trade coffee accounted for 20% in the UK, 7% in France, 1% in Germany (Krier, 2008); U.S. market shares were at 3% in 2006 (Fair Trade Federation, 2008). Second, large-scale population surveys or market data might not always be appropriate to catch the complex actions and motivations of ethical consumers. These are generally omnibus studies (covering multiple topics), with just one or two items directly related to ethical consumer behavior. Thus, such research may be too insensitive to capture the 'ordinary' ethical consumers who (habitually) act in environmentally or socially sustainable ways and the 'ethical simplifiers' who pursue a lifestyle outside capitalist market societies.

Scholars have only begun to study ethical consumers in nonaffluent consumer societies as new middle classes with growing income levels and educational background gain momentum in emerging markets. Thus, studies on the propensity and motivations of ethical consumers in these countries remain scarce. Research on ethical consumption in Brazil and Chile (Ariztía et al., 2014), China (Gomersall and Wang, 2011), Poland (Koszewska, 2011), South Africa (Tustin and Jongh, 2008), and Turkey (Kirezli and Kuşcu, 2012), for example, suggests evolving segments of aspiring ethical consumers.

Historical Context

Lewis and Potter (2011) suggest that the rise and mainstreaming of ethical consumers in affluent societies at the turn of the twenty-first century represents a shift toward a "growing politicization of life and lifestyle practices" (p. 5) in mass consumer culture. However, ethical consumers are not exclusively of contemporary cultural concern. Historically, the two classical types of ethical consumer practices – boycotts and buycotts – date at least back to the eighteenth and the nineteenth centuries (Lewis and Potter, 2011: p. 6); and consumer boycotts focusing on economic objectives such as lower prices and higher quality products can be found as early as the fourteenth century (Klein et al., 2004). Prominent historical examples of social conflicts linked to consumer politics include the boycott of English goods by the Boston Tea Party in 1773 and the buycott of the White Label Campaign in 1899, when US women's consumer leagues rewarded department stores for their good labor conditions with a label (Lewis and Potter, 2011: p. 6).

Lang and Gabriel's (2007) analysis of the historical roots of ethical consumer movements identifies four waves: (1) the cooperative movement of English working-class people to meet their needs for high-quality and affordable food in 1844; (2) The value-for-money movement in the United States at the beginning of the twentieth century, which aimed at informing and educating consumers to get their money's worth; (3) 'Naderism,' named for Ralph Nader, the U.S. lawyer and consumer activist who dedicated his career to informing consumers of and empowering consumers against the alleged manipulation and systematic disinformation of consumers by businesses; and (4) Consumer concerns about the detrimental

impacts of affluent lifestyles on the environment appeared in the 1970s and 1980s. Forming a social movement in the 1990s, an increasing number of consumers expressed their disapproval of companies' environmentally harmful activities and refused to buy their products. A classic example of a consumer boycott motivated by perceived environmental abuse is the boycott of Shell gas stations in Germany in 1995 triggered by the company's plans to dispose the Brent Spar oil platform into the North Atlantic.

Since the 2000s, a broader ethical turn has linked ethical consumer practices to popular critiques of globalization and global social justice, overconsumption, materialism, and unsustainable lifestyles (Lewis and Potter, 2011: p. 7). Scholars discuss several reasons for the rise of ethical consumers (e.g., Harrison et al., 2007; Lewis and Potter, 2011; Newholm and Shaw, 2007; Stehr, 2008): increasing media coverage on the impact and risks of affluent lifestyles, state withdrawal from traditional areas of civic responsibilities with a subsequent upsurge of NGOs and new forms of civic participation, growing affluence and higher levels of education, long-term shifts from materialist values of basic needs to postmaterialist values of self-fulfillment and care for others, and expressing personal concerns and issues of self-image through consumption in today's consumer culture.

Approaches to Understanding Ethical Consumption and Ethical Consumers

The rise of ethical consumers and their practices in the everyday is matched by an increasing interest of scholars especially in the humanities and the social sciences. Since the beginning of the 2000s, researchers have published more and more academic work on the topic (Cotte and Trudel, 2009; Papaoikonomou et al., 2011). Key research questions cover a spectrum ranging from aspects of individualized ethical consumer decision-making, on one end, to structural and collective determinants of consumer movements, on the other end.

The philosophical basis of ethical consumption is a key research area within consumer ethics. Focusing on ethical consumption and ethical consumers in terms of personal moral conduct, scholars apply different philosophical theories to evaluate ethical consumption. Moreover, they discuss the meaning of 'ethical' and explore its dimensions of morality.

Scholars in business and marketing have focused on identifying and categorizing the ethical consumer. These empirical efforts began with profiling ethical consumers and establishing market segments based on sociodemographic variables (gender, income, education, etc.) and personality characteristics. However, as noted above, these findings proved to be unstable. Scholars hence moved on to the question of which factors determine individual ethical consumer practices.

Drawing on the concepts of attitudes, social psychologists have applied modified versions of the attitude–intention–behavior framework (mostly Ajzen's theory of planned behavior; TPB; Ajzen, 1991) to model ethical consumer practices. In most models of TPB, ethical consumption behavior is determined by the intention to perform this behavior, which in turn is influenced by corresponding attitudes and a subjective norm. Perceived behavioral control influences both, intentions and behavior (see Armitage and Connor, 2001 for a general meta-analytic review). Extended models of TPB include, for example, ethical obligation and self-identity as additional determinants (e.g., Shaw et al., 2000). Influenced by Schwartz's processual norm activation model (Schwartz, 1977) norms are again identified as key determinants of ethical consumption. Activated in a four-step decision-making process, it is the personal norm of feeling morally obliged to consume in an ethical manner that is to exert a positive effect on ethical consumption. It is assumed that in order to convert the moral obligation into actual behavior, the personal norm needs to be activated or formed – either as awareness of need (i.e., individual recognition that something needs to be done with regard to the situation at hand) or awareness of responsibility (i.e., individuals' recognition that they are responsible for doing something; see, for example, Guagnano, 2001 for an empirical application to recycled paper products).

Especially in research on green consumption, much scholarly work and discussion has concentrated on the effect of environmental concern as a general attitude toward pro-environmental consumption practices and whether its effect decreases as behavioral costs increase for consumers, the so-called low-cost hypothesis (Diekmann and Preisendörfer, 2003). Empirical evidence of the low-cost hypothesis is mixed, and the link between environmental concern and behavior is often found to be weak.

Economists who have examined the market shares of ethical products find that few consumers put their money where their mouth is, i.e., few actually go out of their way to implement their ethical concerns in the marketplace. Scholars and practitioners have discussed the gap between what consumers say (in surveys) and what they do (in the marketplace), known as the 'attitude–behavior gap' or 'ethical purchasing gap.' Some scholars argue that only improved theories will better account for the attitude–behavior link (see, for example, Carrington et al., 2010; Newholm and Shaw, 2007: p. 257f. with citations therein). Others argue that only better methods than conventional surveys will diminish the ethical purchasing gap (for more details see Section on Methodological Issues below).

Economists' standard approach to understanding consumer behavior focuses on consumers' willingness to pay (WTP) a premium for the ethical features of products. Concentrating primarily on product price and budget restriction, such studies assume that consumers choose the product alternative that gives them the highest expected utility per cost unit. Consumers reveal their preference in choosing the more expensive ethical product, and consequently, it is assumed that these consumers gain additional utility from the ethical attributes. Estimating consumers' WTP, researchers apply various forms of stated preference methods (e.g., contingent valuation, discrete choice experiments) and field experiments. Empirical evidence shows that consumers are indeed willing to pay an ethical premium, for example, for organic and fair trade products (see, for example, Rotaris and Danielis, 2011 and Papaoikonomou et al., 2011 and citations therein).

Identity theory in sociology informs a number of scholars of consumer behavior in analyzing ethical consumption. Drawing on the work of Bauman (1988) and Giddens (1991), who see consumption as an identity project central to the individual expression of self-meaning and the personal selection of

self-image in affluent societies of late modernity, studies find that ethical consumption can be considered as a means to express moral concerns and to construct a (consumer) identity as moral person.

Social scientists have also explored ethical consumption behavior in the context of conspicuous consumption, class, and distinction (Lewis and Potter, 2011: 13). Connected to the arguments of Veblen (1899) and Bourdieu (1984), that consumption practices can serve as a means to prove, retain, and reinforce social distinction and power structures in society, scholars discuss ethical consumer practices as elite projects of social distinction. Indeed, empirical evidence suggests that consumers with higher social status or higher class positions are more prone to ethical consumption (e.g., Johnston et al., 2011).

Human and cultural geographers have examined ethical consumption in the context of space and place (Lewis and Potter, 2011: pp. 13–16). Studies in this vein have focused on questions of commodity chains starting at the producer side of ethical consumption, caring at a distance, food miles, and, more recently, on the networks and organizations shaping ethical consumption (Clarke, 2008).

Criticizing the individualistic bias and the predominant model of ethical consumers as rational actors, scholars in geography and other disciplines argue that ethical consumption is primarily determined by macrolevel factors of networks, organizations, political regulations, and market structures that provide the infrastructure of ethical consumption, and less by factors on the level of individual ethical consumers.

Moving toward the other end of the spectrum, it is questions of civic engagement and collective action that point to the collective and contextual aspects of ethical consumption. Key issues in studies on political consumption and political consumerism evolve around consumption as voting and consumer sovereignty, consumer-citizenship and empowerment (Gabriel and Lang, 2006: Chapter 10), and individualized collective action (Micheletti, 2010). This perspective links ethical consumer practices to individuals' political concerns and serve as a means for political change beyond traditional approaches of political participation – potentially at the expense of (national) government responsibility or crowding out other means of political change that might be more effective. However, empirical evidence for these assertions is inconclusive. Normative concerns have also been raised that only the wealthy can afford to vote on ethical issues in the marketplace (Littler, 2011: p. 34). From a social movement perspective, (boycotting) campaigns of consumers or NGOs highlight the social dilemma of engaging individuals in the (costly) provision of public goods such as human rights, biodiversity, and animal welfare.

Methodological Issues

Scholars draw on a diverse set of methods to empirically investigate hypotheses about ethical consumers and their consumption practices. However, cross-sectional surveys predominate across disciplines as the sources of data that are used to address these hypotheses and gain insights into consumers' ethical purchase behavior, attitudes, and motivations. Cross-sectional surveys (of consumers or a general population) are subject to well-known methodological constraints. For example, simple item measures such as "When I go shopping, I prefer organic products" are typically found in these surveys. The drawback of such simple item measures is that respondents tend to overreport behavior they believe to be socially desirable. Some researchers apply qualitative approaches for data collection, such as in-depth interviews, focus group discussions, and content analysis, to reconstruct the social context of ethical consumers and their consumption habits. But social desirability bias can occur with in-depth interviews and focus group discussions as well.

Some researchers argue for study designs that provide respondents with incentives to uncover their 'true' attitudes, motivations, or behavior. Commonly, researchers recommend discrete choice experiments and contingent valuation methods as incentive compatible alternatives to standard population surveys. Likewise, lab and field experimental designs can be used to overcome biases in other data collection methods in order to study consumers' ethical purchase decisions (e.g., Cotte and Trudel, 2009; Devinney et al., 2010).

Critical Remarks

Along with the mainstreaming of day-to-day ethical consumers and the growing body of interdisciplinary research, scholars and practitioners alike have raised objections with how the field is conceptualized theoretically and practically.

Often, fault is found with the category of the ethical consumer itself. Just as the related term 'ethical consumption,' it is considered to be too inconsistent, too fragmented, and too ambiguous to be meaningful (e.g., Devinney et al., 2010; Littler, 2011). Many studies analyze only a small selection of ethical consumption practices; often it is the scholars who set the focus of research. Hence, researchers run the risk of missing the complex picture of consumers' ethical consumption strategies (Papaoikonomou et al., 2011: p. 213). In daily life, ethical consumer practices might contradict each other. For instance, European or North American consumers saving on food miles (i.e., avoiding food that is transported long distances from producers to consumers; Paxton, 1994) and at the same time buying fair trade products from faraway places (Littler, 2011: 28).

A second common objection raises the question of effectiveness: Can consumers change markets to be ethical? While proponents answer in the affirmative, critics consider ethical consumers to have only marginal influence in society. Some skeptics regard ethical goods as niche products unfit for the mass market, as a means for consumers to ease their guilty conscience, and as a means for companies to earn more profits by whitewashing or greenwashing. Devinney et al. (2010: p. 36), for example, find the image of the ethical consumer greatly exaggerated by activists. They exhort consumers to understand that only producers can change the moral conduct of markets.

A third objection points to the notion of ordinary consumers' habits being 'unethical' when contrasted with those of their ethical counterparts. Critics stress that many consumers frame their ordinary consumption in moral terms; an aspect, which is often neglected in popular and academic discourse on ethical consumption. From this perspective, focusing on ethical consumption solely in terms of highly visible consumer practices turns a blind eye on the less spectacular forms of

day-to-day consumption (Gronow and Warde, 2001; Lewis and Potter, 2011: p. 13; Newholm and Shaw, 2007: p. 260).

Avenues for Future Research

To advance the knowledge and understanding of ethical consumers and their practices, scholars need to address several gaps in research. First, future scholarly work would profit from a commonly accepted definition and typology of ethical consumers and their corresponding behavior. Second, an integrated theoretical framework of ethical consumer behavior would make it easier to study yet unresolved research topics (e.g., Newholm and Shaw, 2007; Papaoikonomou et al., 2011; Peattie, 2010). Future scholarship needs a solid theoretical basis to address the following questions: How does one become an ethical consumer and which factors shape this socialization process in different social groups? To what extent are ethical consumers embedded in their specific cultural and local context and are there cross-cultural similarities and differences? How are individual ethical consumers linked to collective projects of ethical consumption in families, households, and communities?

Third, acknowledging the plethora of ethical concerns and consumers' manifold motivations for their ethical practices, the need to draw on a variety of methods arises. Studying the same ethical consumption practice with different methodological approaches would allow more detailed empirical and theoretical analyses. At the same time, this strategy provides researchers with the opportunity to assess the extent to which their choice of methods influences the study of ethical consumers. Having the above mentioned research questions in mind, future scholarly work would also profit from more large-scale comparative and longitudinal studies to analyze the cultural context and socialization processes of ethical consumers. These need to be complemented with more in-depth qualitative studies to better understand the microlevel processes of ethical consumer practices.

Permission/Credit Line

This article is based on material originally published as 'Ethical Consumption.' In: The Wiley-Blackwell Encyclopedia of Consumption and Consumer Studies, edited by Daniel Cook and J Michael Ryan. John Wiley and Sons Ltd., Oxford, 2015.

See also: Alternative Food Movements; Conspicuous consumption and distinction, history of; Consumer Psychology; Consumption, Geography of; Consumption, History of; Consumption, Sociology of; Environmental Movements; Pro-Environmental Behavior; Sustainable Lifestyles; Sustainable Practices; Sustainable Production and Consumption.

Bibliography

Ajzen, I., 1991. The theory of planned behavior. Organizational Behavior and Human Decision Processes 50, 179–211.

Andorfer, V.A., Liebe, U., 2012. Research on fair trade consumption—a review. Journal of Business Ethics 106, 415–435.

Ariztía, T., Kleine, D., Brightwell, M.d.G.S., Agloni, N., Afonso, R., Bartholo, R., 2014. Ethical consumption in Brazil and Chile: institutional contexts and development trajectories. Journal of Cleaner Production 63, 84–92.

Armitage, C.J., Connor, M., 2001. Efficacy of the theory of planned behaviour: a meta-analytic review. British Journal of Social Psychology 40, 471–499.

Bauman, Z., 1988. Freedom. Open University Press, Milton Keynes.

Bourdieu, P., 1984. Distinction. A Social Critique of the Judgement of Taste. Routledge and Kegan Paul, London.

Carrington, M.J., Neville, B.A., Whitwell, G.J., 2010. Why ethical consumers don't walk their talk. Towards a framework for understanding the Gap between the ethical Purchase intentions and actual buying behaviour of ethically minded consumers. Journal of Business Ethics 97, 139–158.

Clarke, N., 2008. From ethical consumerism to political consumption. Geography Compass 2, 1870–1884.

Cotte, J., Trudel, R., 2009. Socially Conscious Consumerism: A Systematic Review. Network for Business Sustainability Knowledge Project Series.

Devinney, T.M., Auger, P., Eckhardt, G.M., 2010. The Myth of the Ethical Consumer. Cambridge University Press, Cambridge.

Diekmann, A., Preisendörfer, P., 2003. Green and greenback. The behavioral effects of environmental attitudes in low-cost and high-cost situations. Rationality and Society 15, 441–472.

Doran, C.J., 2009. The role of personal values in fair trade consumption. Journal of Business Ethics 84, 549–563.

Doran, C.J., Natale, S.M., 2011. ἐμπάθεια (Empatheia) and Caritas. The role of religion in fair trade consumption. Journal of Business Ethics 98, 1–15.

Fair Trade Federation, 2008. Interim Report on Fair Trade Trends. Wilmington, USA. http://www.fairtradefederation.org/ht/d/sp/a/GetDocumentAction/i/6944 (retrieved 09.03.11.).

FAO, 2009. The Market for Organic and Fair-trade Coffee. FAO, Rome. http://www.fao.org/fileadmin/templates/organicexports/docs/Market_Organic_FT_Coffee.pdf (retrieved 19.01.11.).

Friedman, M., 1996. A positive approach to organized consumer action. The "Buycott" as an alternative to the boycott. Journal of Consumer Policy 19, 439–451.

Gabriel, Y., Lang, T., 2006. The Unmanageable Consumer. Sage, London.

Giddens, A., 1991. Modernity and Self-identity. Polity, Cambridge.

Gomersall, K., Wang, M.Y., 2011. Expansion of Fairtrade products in Chinese market. Journal of Sustainable Development 5, 23–32.

Gronow, J., Warde, A. (Eds.), 2001. Ordinary Consumption. Routledge, London.

Guagnano, G.A., 2001. Altruism and market-like behavior. An analysis of willingness to pay for recycled paper products. Population and Environment 22, 425–438.

Harrison, R., Newholm, T., Shaw, D., 2007. Introduction. In: Harrison, R., Newholm, T., Shaw, D. (Eds.), The Ethical Consumer. Sage, London, pp. 1–8.

Johnston, J., Szabo, M., Rodney, A., 2011. Good food, good people: understanding the cultural repertoire of ethical eating. Journal of Consumer Culture 11, 293–318.

Kirezli, Ö., Kuşcu, Z.K., 2012. Exploring fair trade attitude and fair trade behavior of Turkish consumers. In: 8th International Strategic Management Conference, vol. 58, pp. 1316–1325.

Klein, J., Smith, N., John, A., 2004. Why we boycott. Consumer motivations for boycott participation. Journal of Marketing 68, 92–109.

Koos, S., 2012. What drives political consumption in Europe? A multi-level analysis on individual characteristics, opportunity structures and globalization. Acta Sociologica 55, 37–57.

Koszewska, M., 2011. The Ecological and ethical consumption development prospects in Poland compared with the western European countries. Comparative Economic Research 14, 101–123.

Krier, J.-M., 2008. Fair Trade 2007. New Facts and Figures from an Ongoing Success Story. A Report on Fair Trade in 33 Consumer Countries. Dutch Association of Worldshops, Culemborg.

Lang, T., Gabriel, Y., 2007. A brief history of consumer activism. In: Harrison, R., Newholm, T., Shaw, D. (Eds.), The Ethical Consumer. Sage, London, pp. 39–53.

Lewis, T., Potter, E., 2011. Introducing ethical consumption. In: Lewis, T., Potter, E. (Eds.), Ethical Consumption. A Critical Introduction. Routledge, New York, pp. 3–23.

Littler, J., 2011. What's wrong with ethical consumption? In: Lewis, T., Potter, E. (Eds.), Ethical Consumption. A Critical Introduction. Routledge, New York, pp. 27–40.

Micheletti, M., 2010. Political Virtue and Shopping. Individuals, Consumerism, and Collective Action. Palgrave Macmillan, New York.

Newholm, T., Shaw, D., 2007. Studying the ethical consumer. A review of research. Journal of Consumer Behaviour 6, 253–270.

Papaoikonomou, E., Ryan, G., Valverde, M., 2011. Mapping ethical consumer behavior: integrating the empirical research and identifying future directions. Ethics & Behavior 21, 197–221.

Paxton, A., 1994. The Food Miles Report: The Dangers of Long-distance Food Transport. SAFE Alliance, London, UK. http://www.sustainweb.org/publications/?id=191 (retrieved 01.08.12.).

Peattie, K., 2010. Green consumption. Behavior and norms. Annual Review of Environment and Resources 35, 195–228.

PEW Research Center, 2010. Millennials. A Portrait of Generation Next. Washington. http://www.pewsocialtrends.org/files/2010/10/millennials-confident-connected-open-to-change.pdf (retrieved 01.03.10.).

Rotaris, L., Danielis, R., 2011. Willingness to pay for fair trade coffee: a conjoint analysis experiment with Italian consumers. Journal of Agriculture & Food Industrial Organization 9 (1), 1–22, Article 6.

Schwartz, S.H., 1977. Normative influences on altruism. Advances in Experimental Social Psychology 10, 221–279.

Shaw, D., Shiu, E., Clarke, I., 2000. The contribution of ethical obligation and self-identity to the theory of planned behaviour. An exploration of ethical consumers. Journal of Marketing Management 16, 879–894.

Starr, M.A., 2009. The social economics of ethical consumption: theoretical considerations and empirical evidence. The Journal of Socio-Economics 38, 916–925.

Stehr, N., 2008. Moral Markets. How Knowledge and Affluence Change Consumers and Products. Paradigm Publishers, Boulder.

Tallontire, A., Rentsendorj, E., Blowfield, M., 2001. Ethical Consumers and Ethical Trade. Policy Series 12. Natural Resources Institute, Chatham, UK.

Tustin, D.H., Jongh, D. de, 2008. Ethical consumerism as a key consideration for future brand strategy development in South Africa. Southern African Business Review 12, 24–49.

Veblen, T., 1899. The Theory of the Leisure Class. Macmillan, New York.

Yates, L.S., 2010. Critical consumption. European Societies 13, 191–217.

Relevant Website

www.ethicalconsumer.org – Ethical Consumer: The Alternative Consumer Organisation.

Ethical Dilemmas, Research, and Access to Treatment

Dirceu B Greco, Federal University of Minas Gerais, Belo Horizonte, Brazil

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Rights and controversies of access to ethically sound clinical trials and right to universal access to established and new products of research are discussed. It is proposed the development of a comprehensive set of ethical requirements, with participation of government officers, international representative agencies, investigators, activists and civil society with an agenda to define where relevant clinical trials are to be performed, to adequately inform the prospective volunteers so that they can autonomously decide to participate and to ascertain posttrial access to developed products. Dilemmas in research ethics, including but not limited to access to care and treatment and the use of placebo will be dealt with and may then be settled.

There must be an internationally accorded decision on universal access to health care which will help guarantee the inclusion of the most prevalent illnesses, including orphan or neglected diseases, such as tuberculosis, malaria, leprosy, leishmaniasis, schistosomiasis, trypanosomiasis, better qualifying the original Millennium Development Goals. Universal access to health care implies the intensification in the fight to eliminate poverty, and as individuals/countries emancipate themselves it will help confirm that the status quo of inequality can be defeated. Dilemmas related to apparent scarce resources should be confronted with other funds, such as military investments and financial transaction taxes.

Introduction and Definitions

Universal access to quality care and treatment in a public health setting is an achievable human right. The ethics of conducting clinical trials are part of this. If clinical trials are conducted ethically has to be considered in relation to, the day-to-day difficulties of research involving vulnerable populations and countries.

The history of trials involving human subjects is blemished by many unethical experiments and the situation is not different in relation to access to decent health care throughout the world. It is fair to say that the interruption of the Tuskegee Syphilis Study (1932–72) was due to public outrage after news reports in 1972 in both the Washington Post and The New York Times. In this experiment there were "issues of race and racism, vulnerability of particular communities or populations, power, deception, exploitation, and questions about consent, appropriate study design and research hypotheses, distinctions between research and treatment, and standard of care" (Parker et al.; Fairchild and Bayer, 1999).

Other unethical trials were conducted in the US (Beecher, 1966; Hornblum, 1997) in the post-Second World War period and also outside the US – the Guatemala sexually transmitted diseases inoculation project is another hallmark of an unethical trial – (Reverby, 2011) after the issuing of Nuremberg Code (1947) and the then recently released Declaration of Helsinki – DoH (1964) (WMA).

The risks of unethical trials, probably not so deplorable as the ones mentioned above, were in part related to the exponential expansion in the last few decades of clinical trials originated in a developed but usually performed in developing countries (Macklin, 2004; Lavery et al., 2007).

The establishment of phase III efficacy trials, with thousands of volunteers and the globalization of the pharmaceutical industry facilitated the performance of these trials outside industry's country of origin. Add to that the easiness of recruiting participants in vulnerable environments, with possible laxer ethical requirements, with high burdens of disease and less demanding individuals, who many times see this participation as the only means of accessing medical care. On the other hand, more stringent rules in developed countries, where the participants' rights are at least formally protected, are part of the same picture, precluding the approval of some trials there.

Intellectual Background, Schools of Thought and Leading Advocates, Current and Emerging Directions

The necessity to protect research participants gave rise to the Nuremberg Code in 1947 (Shuster et al., 1997) and in the US to the Belmont report in 1979, the latter issued a few years after public disclosure of the Tuskegee study. In the Belmont Report three principles were included -autonomy, beneficence and justice, which according to the authors express the general values underlying rules in the common morality.

In 1979, Beauchamp and Childress (2001) helped consolidate a new phase for bioethics with the development of a theory of 'applied ethics,' known as the Principlism theory, which included the three principles of the Belmont report: autonomy, beneficence and justice, adding the forth principle of postmaleficence.

Principlism ethics has been mentioned in many documents, with critics arguing that it may not cover the whole spectrum of moral decisions. A thorough discussion of principlism is in Pierre Malia's Critical Overview of principles (2013), in Clouser and Gert (1990) and in Holms manuscripts (1996). Malia concurs with the great influence of Beauchamp and Childress in the development of bioethics, but argues that the principlism approach, "starting from American common-morality, leads to an underdevelopment of beneficence and justice, and that the methods offered for specification and balancing of principles are inadequate."

On the same token a Latin-American movement calls for 'an intervention bioethics' (Garrafa and Porto, 2003). The authors take in account what they understand as a "growing lack of political analysis of moral conflicts and of human indignation. The indiscriminate utilisation of the bioethics justification as a neutral methodological tool softens and even cancels out the seriousness of several problems, even those that might result in the most profound social distortions." Based on these premises, and on the concept that equity means 'treating unevenly the unequal,' they "propose a hard bioethics or intervention bioethics, in defence of the historical insights and rights of economically and socially excluded populations that are separated from the international developmental process."

Another bioethical approach, especially aimed at the destitute also originated in Latin America. Schram and Braz propose a Bioethics of Protection. According to the authors, this came in the scope of applied ethics in Latin America, related initially to the existing moral conflicts in Public Health in developing countries, combining the dialogue of consecrated bioethical tools and the context of the praxis of the Latin-American stakeholders. "To the traditional descriptive and normative functions of applied ethics, bioethics of protection proposes an additional function consisting in the 'protection' with priority of individuals and harmed groups, in order to permit them to develop, if possible, their capabilities, necessary to their empowerment, i.e., necessary to a decent life in a conflagrated world, and, possibly, searching for rational and/or reasonable solutions" (Schramm and Braz, 2008).

It becomes evident that the definition, principles, application and reason d'etre of bioethics is complex as the world we live in – multicultural, multilingual, with obscene disparities and violent.

Issues and Expected Changes

Here is one important dilemma- is it possible to keep the same exacting ethical standards independently of the origin, ethnicity or economic power of the volunteers?

My clear answer is Yes – origin, economic power, ethnicity must not be used to treat individuals with less rigorous ethical values, as we are dealing with human beings born with the same universal rights (Universal Declaration of Human Rights, 1948).

One of the risks of doing otherwise is the setting of double standards in clinical trials. And as access to care and common goods are not the same in different settings, not only in different countries but also within countries, performance of clinical trials in the developing and even with less well to do individuals in the developed world, carries the risk of cutting corners in ethical requirements and this actually has happened more often than not. Furthermore, unethical conduct is not limited to private pharmaceutical companies. Ethically fraught studies were also performed by foreign and local health agencies with the willing participation of local and international investigators (Reverby, 2011; Greco, 2013; Presidential Commission for the Study of Bioethics Issues, 2011).

It is many times mentioned that it is impossible to keep the same standards in poor environments and if research is not performed there these populations will be deprived of the progress of science. This is not a true dilemma as many of the trials performed in developing countries were not useful at all for their population. One example is the set of trials that occurred in the early 1980s to ascertain the efficacy of a vaccine against hepatitis B. They took place in both the developed and in a few developing world countries and showed that the vaccine was very effective. When commercially released its cost was not within reach of most countries. Even in 2000, the vaccine's availability in public health in Africa was limited to Botswana, Gambia, Mauritius, Seychelles, South Africa, Swaziland and Zimbabwe. Reasons listed for delayed introduction included high vaccine cost, weak infrastructure and low financing priority among donors (Arevshatian et al., 2007). Thus the question is not of performing health research anywhere for the sake of science but of doing relevant trials where volunteers can be adequately protected and when there is assurance of access to the research products to all that may need them – access in the sense of universal availability in the public health system.

There is another intertwined question: Health research is a broad term and search for knowledge is a human characteristic but the main objective of health research must of course be what above mentioned, that is provision of access to the efficacious products of research wherever needed.

If a vulnerable population/community/country has no condition for participating in a clinical research, the first perceived dilemma of precluding them of access to progress of science, is also counter argued with: first, that access to the products is more important than where the research is performed; second, if the specific research is of extreme value for this population and should be there performed, research conditions must be established locally prior to beginning of the trial.

Thus, inasmuch as clinical research trials with new medications or vaccines are undoubtedly necessary, they can be carried out only where independent ethical requirements and due evaluation is available, where risks of exploitation and double standards can be circumvented. Research must be designed and only implemented when the populations in which it is carried out stand to benefit from the results (CIOMS, 2002). Even with perceived and sometimes real urgency to search for better preventive methods, and more effective medications and vaccines, what is urgently needed is to have them available to all and not only to the trial participants. That will be dealt with in the treatment priorities dilemmas.

Thus the apparent dilemma of not performing a trial in certain circumstances is not an ethical dilemma as upon completion of the trial, if the product is shown to be effective, there should be international pressure to make it available (and affordable) wherever needed.

With this background two international documents, from UNAIDS and from WHO, listed circumstances/conditions in which biomedical research trials should not be performed. They include (UNAIDS, 2007; WHO, 2010):

- when the capacity to conduct independent and adequate scientific and ethical review does not exist;
- where voluntary participation and freely decided consent cannot be obtained;
- when conditions affecting potential vulnerability or exploitation may be so severe that the risk outweighs the benefit of conducting the trial in that population;

- when agreements have not been reached among all research stakeholders on access to medical care and treatment;
- when an agreement has not been reached on responsibilities and plans to make trial products (drugs, other treatments, or preventive measures) that prove to be safe and effective, available to communities and countries where they have been tested, at an affordable price.

Autonomy Dilemma

It has been argued that individual autonomy in trials can be limited for the sake of scientific progress. Yet often the relevance of the pursued knowledge is not well defined, or the expected results will not be relevant for this population. Moreover, many times trials are performed among vulnerable (social and/or economically) individuals/communities. Individual well being has to be prioritized against the value for the community (Rawls, 1971). To take one example: the AZT placebo-controlled trial among HIV infected women in Asia, Africa, and Central America in the 1990s. Participant's vulnerability was clear as they came from a poor environment, had difficulty in understanding the informed consent terms, had limited access to adequate health care, making them prone to participate in a trial just to receive medical care. In 1997, a strong editorial by Marcia Angell (1997, 2000) and a paper by Lurie and Wolfe (1997) in the same issue of the New England Journal of Medicine criticized the ethics of these studies performed outside the US. A shorter scheme of AZT, eliminating its intravenous use was tested not against the gold standard ACTG 076 (MMWR, 1995) but against placebo instead. Angell questioned if these trials were not reminiscent of the Tuskegee study. This editorial and article brought about fierce discussions that ensued strong pressure from US investigators and agencies to change the 1996 version of the World Medical Association (WMA) Declaration of Helsinki (DoH), at that time the most internationally respected set of ethical guidelines on human research. Their intent was to modify two articles, exactly those dealing with access to best proven medical care to all trial participants and with restrictions on placebo use when an efficacious treatment exists. Pressures were overwhelming and WMA backed off adding two notes of clarification, making it easier to dodge requirements of treating participants with respect and rights irrespectively of their origin or economic power, setting grounds for double standards in clinical research. The item on Placebo use, that in the 1996 version was not permitted when an efficacious control was available, has been changed and the 2013 version confirmed the possibility of its use even in certain circumstances where an efficacious control exists. This is an example of how guidelines may not be in the best interest of developing countries and to counteract this situation, countries have autonomy to be more stringent in protecting clinical trials participants. This was the case in Brazil: the National Research Ethics Commission (CONEP) in its Resolution 466/2012 (Conselho Nacional de Saúde (CNS) Brasil, 2012), which regulates all Brazilian human research states on the use of placebo: "Benefits, risks, burdens and effectiveness of a new method should be tested against those of the best current prophylactic, diagnostic, and therapeutic methods. This does not exclude the use of inert placebo in studies where no proven diagnostic or therapeutic method exist." On the same token the Brazilian Federal Medical Council (Conselho Federal de Medicina (CFM) Brasil, 2010) in the 2010 version of the National Code of Medical Ethics considered unethical for any physician to participate in a research project where placebo is used in the control group when there are efficacious and effective drugs for the disease being tested. And more recently the Latin American and Caribbean Medical Confederation in their XVI General Assembly convened in 2013 in Pachuca, México issued the *Declaracion de Pachuca* (Confederacion Medica latinoamericana y del Caribe, 2013), where it unanimously rejected any human medical investigation utilizing placebo when there is an efficacious drug for the disease being studied. They also propose that governments should neither authorize nor finance drugs that used placebo for their evaluation when a proven control existed.

Safeguarding against Exploitation

There are procedural ways to partially curtail vulnerability issues that may occur in clinical trials. For example, establishment of an internationally approved and enforceable research ethics document, with guidelines based on human rights aimed at universal access to decent health care, plus the availability of publicly disclosed research proposals and results. With this in place studies such as Tuskegee, Guatemala, the AZT trials for infected mothers performed in the late 1980s and many other not as (in)famous that occurred in many countries will never be permitted again.

It will of course be more difficult to overcome the hurdles to achieve the much-needed fundamental changes. Who and where are (and should be) the real stakeholders? The real stakeholders are all individuals who can use the results of good science. Of course this may sound vague but it means that is possible to involve individuals, communities, public health officers and the research team in decisions related to relevant clinical trials so that efficacious products are made available in the public health settings.

In this direction the following points, modified from UNAIDS 2004 Guidance document (UNAIDS, 2004) should be considered:

- All stakeholders including individuals, communities, public health agencies and officers, and the research team, both from developed and developing countries, should collaborate during all stages of the study, from the development of the protocol to the publication and application of the results in public health to all who need them. The response to the AIDS epidemic may be used as a good example of this broad involvement.
- Decisions about access to the products which result from the studies should be made based upon the principles of justice, meaning that those who have to cope with the heaviest burden should receive the appropriate benefits. Participants must have posttrial access to drugs, vaccines, interventions, prevention strategies and any other benefits resulting from the study.
- The involvement of international representative agencies (such as UN/WHO) will be a crucial asset to universalize access.

- Financing of studies should include funds for strengthening local capacities, not only those related to the specific aims of the project, but also local infra-structure, institutions and host country capacity to conduct research projects.
- It is extremely important to deepen the discussion on the informed consent process. Such consent must be obtained in ways that respect the dignity and cultural values of the volunteers.
- Inclusion of vulnerable individuals shall warrant special justification and appropriate protection and should occur only when the aim of the project is to their own benefit.

Social Determinants, Power Relations, and Emancipation

To confront disparity, exploitation and poverty there is no place for the so-called *empowerment*, which supposedly means that someone is going to transfer power to someone else, does not hold true. Empowerment is supposed to 'give' capacity to individuals in relation to their needs and expectations. Unfortunately, this is a figure of rhetoric as power is never given, very seldom shared. Most of times it means a top down way of marginally providing a little to those in need, to calm down their claims. These vulnerable individuals/countries will praise the 'donors' for helping them get what is nothing more than their rights. This only perpetuates dependency, in a new sort of colonialism – Give the rings, save the fingers and maintain the status quo of disparity and vulnerability.

What is here proposed is *emancipation* in the sense proposed by the Brazilian educator Paulo Freire. Writing on education for freedom, Freire held that human liberation will neither occur accidentally nor as a concession, but it will be a conquest of human praxis, demanding constant fighting to have it achieved. In Freire's method there is a moment when the oppressed take the nitiative to fight and emancipate themselves from the oppressors (Freire, 2005). Thus, emancipation goes hand-in-hand with access to education, citizenship, human rights and the confrontation of disparity.

Ethics of Universal Access to the Products of Research and to Adequate Health Care (Universal Health Access)

Access to the Products of Research – Political Decision + International Participation

Even with a well-established research ethics framework, access to the developed products, that is, making them available and affordable is getting increasingly complex. Despite possible political and budgetary commitment, costs of new, more complex and patent-protected medications are straining health available resources worldwide. Thus there is an urgent need for decreasing these costs and this may happen, for instance, through international collaboration in procurement. Examples include the Pan American Health Organization (PAHO) Strategic fund, UNITAID and Clinton Health Access Initiative (CHAI).

PAHO's strategic fund is a cooperation mechanism to improve access to and supply management of strategic public health supplies. As a procurement mechanism, it allows participating members to utilize a common fund for payment of authorized purchases of essential public health commodities. The Fund is capitalized through the allocation of the 3% procurement service charge. As more countries participate, the purchasing power of the Fund is expected to increase resulting in lower product prices (PAHO).

UNITAID was launched in 2006 at the opening of the 61st Session of the General Assembly of the United Nations by Brazil, France, Chile, Norway and the United Kingdom It uses innovative financing to increase access to treatments and diagnostics for HIV/AIDS, malaria and tuberculosis in low-income countries. A novel way of financing UNITAID was established through a small levy on airline tickets in several countries. As of the end of 2011, this levy has collected US $1.06 billion – 66% of the US $1.6 billion raised by UNITAID.

The CHAI was established in 2002 addressing the HIV/AIDS crisis in the developing world and aiming at strengthening local health systems. CHAI "works to improve markets for lifesaving medicines and diagnostics, lower the costs of treatments, and expand access to life-saving technologies" (CHAI).

These mechanisms should be associated with national or regional political decision to implement local production of generics, WHO pre-qualified, at fair prices to counteract the ever increasing costs of patent medicines. There are legal tools for issuing compulsory licensing of drugs to protect public health, in accordance with TRIPS allowed flexibilities (Declaration on the TRIPS agreement and public health; Greco and Simao). It must be added that a substantive part of the costly research on several drugs were and still are occurring in universities and research institutes financed mainly by public funds in the US and Europe, which must also be brought to the global discussion on patents, transfer of technology and price setting.

Access to Adequate Health Care (Universal Health Access)

Universal access to good quality healthcare is *sine qua non* for achieving real distribution of research results. To reverse the current situation, intense changes in the world order will be necessary to attain fair distribution of resources, equity and to decrease vulnerability of all involved. The dilemma here perceived is that the status quo of access and funding for health is scarce and it is not possible to change that. One counterargument used by many concerned with access to adequate health care, is that funding for all disease-related expenses should be weighed against other uses of public monies and holds that funding for health should be increased by decreasing the budget for military purposes.

To help discuss the apparent dilemmas in relation to access to funding for health care, the AIDS epidemic is a very special example. Much has been written about the enormous economic burden of the AIDS epidemic, including direct (medication, out-patient care, hospices, hospitalization), indirect (research, education, prevention, loss of GNP) and social costs for the individual, their communities and countries (Mann, 1987). It must be remembered that in most of the third world, HIV infection is added to a list of diseases already endemic (e.g., schistosomiasis, Chagas, leishmania, hepatitis, malaria, leprosy). It is also true that even in developed countries access to decent and affordable health care is not universal (Stillman, 2013).

In 2013, Piot and Quinn (2013) and Brandt, (2013) underlined the important role of the AIDS epidemic in shaping global health, even with all the suffering the epidemic

imposed to millions of individuals. Pressure from society and also from the disease itself may be strong enough to modify these 'fixed budgets.' To better illustrate the skewed distribution of available funds the 2013 report of the Stockholm International Peace Research Institute (SIPRI) estimated that world military spending in 2012 was $1.75 trillion, of which $682 billion were spent by the US, $166 billion by China and $90.7 billion by Russia. There was increased spending in Asia, Eastern Europe, the Middle East, North Africa and Latin America. Military expenditures were several hundred times WHO's annual budget of $395 billion for the 2012–13 period for all their programs. It is also several hundred times higher than the annual budget of the US Centers for Disease Control (CDC) of over $10.0 billion, and the Gates Foundation, which in 2012 donated $1.3 billion to the Global Fund to Fight AIDS, Tuberculosis and Malaria (Chelala, 2013). To change this distribution is of course difficult and complex but worth pursuing. More recently a movement was set in motion aiming at finding resources from other available sources by establishing a Financial Transaction Tax (FTT) for Global Health. It is estimated that the 2008 economic crisis has put an extra 200 million people, which had nothing to do with the reasons for the crisis, living on less than $2 a day in some of the world's poorest countries (Oxfam, 2010). On the other hand, the financial sector was and has been bailing out and is continuing to make huge profits. This FTT could make banks, hedge funds and other parts of the financial sector pay their fair share to cover for what was their responsibility and provide long term financing for global health, to help attain the Millennium Development Goals (MDGs) (2000). An FTT even with an average tax of just 0.05% on transactions (such as bond and share sales) could raise as much as $409 billion a year (Leading group on innovative financing to fund development, 2010).

Conclusions/Perspectives

Unfortunately, health disparities will not be solved just by standards and guidelines that regulate research and researchers; not even exclusively by treating everyone as equals in studies involving human beings.

Justice and thus, equal access to treatment and care, will be served when individuals, communities and countries emancipate themselves to fight for their rights. In his *History of the Peloponnesian War*, Thucydites (411 BC) wrote that Justice would prevail when those who are not subjected to injustice are as indignant as those who are. I dare counter with the following: Justice will only prevail when those affected by injustice emancipate themselves to fight for their rights. Making sure that equity is respected in clinical research is of course a step toward reversing the current injustice in the allocation of resources for health research and can contribute to the emancipation of volunteers, researchers and society, to fight for their rights as citizens.

Perspectives for Ethics in Research

A concerted action of governments, international representative agencies, investigators, activists and civil society will be crucial to develop together an agenda to define where relevant clinical trials are to be performed, how to prepare the sites for these trials, to adequately inform the prospective volunteers so that they can autonomously decide to participate and to ascertain posttrial access to the developed products.

A comprehensive and harmonized set of ethical requirements must be developed, ideally under the umbrella of WHO and UNESCO. This requirements could use established international documents, such as Universal Declaration on Bioethics and Human Rights (UNESCO, 2005) and the CIOMS International ethical guidelines for biomedical research involving human subjects (CIOMS, 2002). These world representative institutions could convene a multinational, multidisciplinary working group, including civil society to elaborate and open up for discussion a draft document of ethical requirements based on human rights and equity. Dilemmas in research ethics, including but not limited to access to care and treatment and the use of placebo will be dealt with and may then be settled. This document should be broad enough to encompass a road map for universal access to the developed efficacious products. This will be a daunting but worthwhile task to really and truly convince the world that access to health is not a commercial commodity but a human right, that progress in research and development must be available to all, and must be shared throughout the world.

Perspectives for Universal Access to Health Care

It must be noted the recent decision at the United Nations 67th General Assembly on Universal Health Coverage (GA December 2012). A consensus text was adopted encouraging member states to plan and pursue transition of national health care systems toward universal coverage (UN 67th General Assembly, 2012). This is in accordance with the International Covenant on Economic, Social and Cultural Rights directives on rights to the highest attainable standard of health (UN ECOSOC, 2000). Examples of public health access, such as the universal public health systems of Brazil, Canada and UK, confirm that it is attainable and feasible in different settings.

As already mentioned, AIDS can and should be used as a global health example with a multipronged and multidisciplinary involvement of all stakeholders, including health authorities/governments, sponsors, scientists, and civil society (Brandt, 2013; Piot and Quinn, 2013). The decision on universal access to health care will help guarantee the inclusion of other (orphan or neglected) diseases with much less clout but as significant and disseminated as HIV/AIDS, including, but not limited to, tuberculosis, malaria, leprosy, leishmaniasis, schistosomiasis, trypanosomiasis, better qualifying the original Millennium Development Goals.

More importantly universal access to health care implies the intensification in the fight to eliminate poverty, and as individuals/countries emancipate themselves it will help confirm that the status quo of inequality can be defeated.

See also: Bioethics; Common Rule: US Regulatory Protections for Human Research Participants; Ethics for Biomedical

Research Involving Humans: International Guidelines; Fieldwork: Ethical Aspects; Informed Consent of Research Subjects; Intellectual Property Rights: Ethical Aspects; Objectivity of Research: Ethical Aspects; Placebo-Controlled Trials, Ethics of; Research Funding: Ethical Aspects.

Bibliography

Angell, M., 18 September 1997. The ethics of clinical research in the third world. New England Journal of Medicine 337, 847 (Editorial).

Angell, M., 2000. Investigator's responsibilities for human subjects in developing countries. New England Journal of Medicine 342, 967–969.

Arevshatian, L., Clements, C.J., Lwanga, S.K., Misore, A.O., Ndumbe, P., Seward, J.F., Taylor, P., 2007. An evaluation of infant immunization in Africa: is a transformation in progress? Bulletin of the World Health Organization 85 (6). http://dx.doi.org/10.2471/BLT.06.031526. Accessed at: www.who.int/bulletin/volumes/85/6/06–031526/en/.

Beauchamp, T., Childress, J., 2001. Principles of Biomedical Ethics. Oxford University Press, Oxford.

Beecher, H.K., 1966. Ethics and clinical research. New England Journal of Medicine 274, 1354–1360.

Belmont Report – Ethical Principles and Guidelines for the Protection of Human Subjects of Research, 18 April 1979. The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.

Brandt, A.M., 2013. How AIDS invented global health. New England Journal of Medicine 368 (23), 2149–2152.

Chelala, C., 2013. Distorted Priorities: Military Spending vs. People's Health. Accessed at: www.commondreams.org/view/2013/04/30-0.

CIOMS International Ethical Guidelines for Biomedical Research Involving Human Subjects, 2002. Council for International Organizations of Medical Sciences, Geneva. http://www.cioms.ch/publications/layout_guide2002.pdf.

Clinton Health Access Initiative, accessed at: www.clintonfoundation.org/our-work/clinton-health-access-initiative.

Clouser, K.D., Gert, B., 1990. A critique of principlism. Journal of Medicine and Philosophy 15 (2), 219–236.

Confederación Medica Latinoamericana y del Caribe, 2013. Declaration de Pachuca, www.confemel.com/asambleas/xviaao/declaracion_helsinki.pdf.

Conselho Federal de Medicina (CFM), 2010. Accessed at: www.portalmedico.org.br/novocodigo/integra.asp CFM – codigo de ética médica – artigo 106.

Conselho Nacional de Saúde (CNS), Brasil, 2012. Accessed at: conselho.saude.gov.br/resolucoes/2012/ppReso466.pdf.

Declaration on the TRIPS agreement and public health. DOHA WTO Ministerial, 2001. Accessed at: www.wto.org.

Fairchild, A.L., Bayer, R., 1999. Uses and abuses of Tuskegee. Science 284, 919–921.

Freire, Paulo, 2005. Pedagogy of the Oppressed (Myra Bergman Ramos, Trans.). The Continuum International Publishing Group Inc., New York, ISBN 0-8264-1276-9, 183 pp.

Garrafa, V., Porto, D., 2003. Intevention bioethics: a proposal for peripheral countries in a context of power and injustice. Bioethics 17 (5–6), 399–416.

Greco, D.B., 2013. Emancipação na luta pela equidade em pesquisas com seres humanos. Revista Bioética 21 (1), 20–31.

Greco, D.B., Simão, M., July 2007. Brazilian policy of universal access to AIDS treatment – sustainability challenges and perspectives. Acquired Immune Deficiency Syndrome 21 (Suppl. 4), S37–S45.

Holms, S., 1996. Not just autonomy – the principles of American biomedical ethics. Journal of Medical Ethics 21 (6), 332–338. http://dx.doi.org/10.1136/jme.21.6.332.

Hornblum, A.M., 1997. They were cheap and available: prisoners as research subjects in twentieth century America. British Medical Journal 315, 1437–1441.

Lavery, J.V., Grady, C., Wahl, E.R., Emanuel, E.J., 2007. Ethical Issues in International Biomedical Research a Case Book. Oxford Press, Oxford.

Leading group on innovative financing to fund development, 2010. Globalizing Solidarity: The Case for Financial Levies. www.healthpovertyaction.org/policy-and-resources. www.robinhoodtax.org.

Lurie, P., Wolfe, S., 1997. Unethical trials of interventions to reduce perinatal transmission of the human immunodeficiency virus in developing countries. New England Journal of Medicine 337 (12), 853–856.

Macklin, R., 2004. Double Standards in Medical Research in Developing Countries. Cambridge, ISBN 0521 833884.

Malia, P., 2013. Critical Overview of Principles in the Nature of the Doctor-Patient Relationship Health Care Principles through the Phenomenology of Relationships with Patients, vol. 2. SpringerBriefs in Ethics, ISBN 978-94-007-4939-9. VI, 86 pp.

Mann, J., 1987. AIDS. World Health Forum 8, 361–370.

Marx, K., Engels, F., 1848. Manifesto of the Communist Party.

U.S. Public health service recommendations for human immunodeficiency virus counseling and voluntary testing for pregnant women. Morbidity and Mortality Weekly Report 44 (RR-7), 7 July 1995, 1–15.

Oxfam, Development Finance International, 2010. The Impact of the Global Financial Crisis on the Budgets of Low-Income Countries.

Pan American Health Organization (PAHO), accessed at: www.paho.org.

Parker, L.S., Alvarez, H.K., Thomas, S. Ethics and Public Health Curriculum Module. The Legacy of the Tuskegee Syphilis Study, accessed at: https://medicine.mercer.edu/mu-medicine/ethics/second/cases/upload/Parker-Tuskegee.pdf.

Piot, P., Quinn, T.C., 2013. Global health: response to the AIDS pandemic – a global health model. New England Journal of Medicine 368 (23), 2220–2225.

Presidential Commission for the Study of Bioethical Issues, September 2011. "Ethically Impossible" STD Research in Guatemala from 1946 to 1948. Washington, DC.

Rawls, John, 1971. A Theory of Justice. Belknap Press of Harvard University Press, ISBN 0-674-00078-1 (revised 1999).

Reverby, Susan M., January 2011. "Normal exposure" and inoculation: a PHS "Tuskegee" doctor in Guatemala, 1946–48. Journal of Policy History Special Issue on Human Subjects 23 (1), 6–28.

Schramm, F.R., Braz, M., 2008. Bioethics of protection: a proposal for the moral problems of developing countries? International Journal of Bioethics 19 (1), 73–86.

Shuster, E., Markman, J.R., Markman, M., 1997. Fifty years later: the significance of the nuremberg code. New England Journal of Medicine 337, 1436–1440.

Stillman, M., 2013. Dead walking men. New England Journal of Medicine 369 (20), 1880–1881.

Thucydites, 411 BC, 1954. The History of the Peloponesian Wars. Penguin Classics, London.

UNAIDS, 2007. Ethical Considerations on HIV Prevention Trials. Revised in 2012: www.unaids.org/en/media/unaids/contentassets/documents/unaidspublication/2012/jc1399_ethical_considerations_en.pdf.

UNAIDS/WHO, 2004. Guidance on Ethics and Equitable Access to HIV Treatment and Care, ISBN 9241 592583.

UN Economic and Social Council, Committee on Economic, Social, and Cultural Rights, 22nd Session, 25 April–12 May 2000. General Comment No. 14: The Right to the Highest Attainable Standard of Health http://www.un.org/documents/ecosoc/docs/2001/e2001-22.pdf.

UNESCO, 2005. Universal Declaration on Bioethics and Human Rights. http://portal.unesco.org/en/ev.php-URL_ID=31058&URL_DO= DO_TOPIC &URL_SECTION=201.html.

UN 67th General Assembly, 2012. www.un.org/news/press/docs/2012/ga11326.doc.

UNITAID, accessed at: www.unitaid.eu/.

UN Millenium Development Goals, 2000. http://www.un.org/millenniumgoals/.

UN Universal Declaration of Human Rights, 1948. Accessed at: www.un.org/en/documents/udhr, Disponível em: http://bioethics.gov/ cms/sites/default/files/PCSBI-IRP_Research-Across-Borders_0.pdf.

WHO, 2010. Ethical Guidance of Tuberculosis Prevention, Care and Treatment – whqlibdoc.who.Int/Publications/2010/9789241500531_eng.Pdf. WHO Guidance on Ethics of Tuberculosis Prevention, Care and Control.

World Medical Association Declaration of Helsinki, www.wma.net.

Ethical Guidelines and Codes of Conduct in Social and Behavioral Research

Thomas L Van Valey, Western Michigan University, Kalamazoo, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article briefly reviews the situation regarding research ethics in Australia, Canada, Great Britain, and the United States. It notes some of the areas where regulations referring to social/behavioral research are similar, as well as areas where they differ. It then turns to the codes of ethics of various professional organizations, first describing them and then comparing and contrasting them. Finally, it points to the future of social/behavioral research and the emerging trend toward broader notions of research integrity and the responsible conduct of research.

Introduction

The guidelines and codes of conduct that exist today have essentially been created in response to major incidents, mostly in the area of medical/health research with human subjects that became highly publicized. They can certainly be traced directly back to the Nuremberg Code (1949) and the Declaration of Helsinki (1962). In the United States, the Belmont Report (1979) built upon these documents, and in turn, became the basis for what has come to be known as the Common Rule (1981), which applies wholly or in part to most federal agencies. In addition, many academic institutions accept the Common Rule as the basis for their own institutional policies, whether they receive federally funded research or not.

While the original focus of most regulations was on medical/health research and the protection of human subjects from physical harm, other well-publicized incidents resulted in social and behavioral research being included (Coleman et al., 2005). Nevertheless, although they apply to the social and behavioral sciences, most of the current regulations and guidelines dealing with human subjects continue to focus on issues most relevant to medical/health research. For this reason, some tensions exist in the regulatory context. These tensions tend to revolve around methodological issues (e.g., qualitative and other approaches seldom used in the health/medical areas) and/or substantive issues (e.g., topics that often produce controversial viewpoints, such as crime, poverty, and human sexuality).

In recent years, emphasis has begun to extend beyond the protection of human subjects (and animals) in research to the broader notions of research integrity and the responsible conduct of research (RCR). While it includes the protection of human subjects, RCR also covers such things as conflicts of interest, data management, authorship, and safety issues. With some exceptions, this article will deal primarily with the guidelines and codes of conduct relevant to the use of human subjects in research, and less with other dimensions of professional activity.

Regulatory Contexts

Australia

In Australia, the *National Statement on Ethical Conduct in Human Research* (2007) and the *Australian Code for the Responsible Conduct of Research* (2007 – referred to as the Research Code) are the primary sources for regulations and guidelines. Australia has established Human Research Ethics Committees that review proposed research projects. The Australians have identified a set of four principles upon which reviews are based:

1. Integrity
2. Respect for persons
3. Beneficence
4. Justice

The Research Code provides a framework for responsible research practice and lays out guidelines in data and record management, publication of findings, authorship, conflicts of interest, and other areas (http://www.nhmrc.gov.au/book/national-statement-ethical-conduct-human-research).

Canada

In Canada, the fundamental regulations and guidelines that apply to all research – including health, humanities, natural and social sciences, and engineering – is the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*. The first edition was produced in 1998 and was revised in 2010. For the social and behavioral sciences, this is primarily the purview of the Social Sciences and Humanities Research Council of Canada, the major funding source. In addition to the Tri-Council Policy, Canadian government-funded researchers and their institutions must also follow other policies pertaining to research integrity, peer review, and conflicts of interest. Numerous other research organizations in Canada have adopted the Tri-Council Policy as the basis for their own operations. (The full text of the statement is available at http://www.ethics.gc.ca/pdf/eng/tcps2/TCPS_2_FINAL_Web.pdf)

The three core principles underlying the Canadian Tri-Council Policy are:

1. Respect for persons
2. Concern for welfare
3. Justice

Research Ethics Boards (REBs) are the entities charged with the actual reviews of proposed research. The REBs are created and maintained by the institutions that have jurisdiction over human research.

The United Kingdom

In the United Kingdom, the responsible regulatory bodies are the National Research Ethics Service and the Economic and Social Research Council. For the social and behavioral sciences, the *Framework for Research Ethics* provides the basic principles and guidelines. It was first published in 2005 and revised in 2010 and again in 2012 (it is available at http://www.esrc.ac.uk/_images/Framework-for-Research-Ethics_tcm8-4586.pdf). The *Framework for Research Ethics* explicitly identifies six principles that should be addressed in any review of proposed research:

1. Research should be designed, reviewed, and undertaken to ensure integrity, quality, and transparency.
2. Research staff and participants must normally be informed fully about the purpose, methods, and intended possible uses of the research, what their participation in the research entails, and what risks, if any, are involved.
3. The confidentiality of information supplied by research participants and the anonymity of respondents must be respected.
4. Research participants must take part voluntarily, free from any coercion.
5. Harm to research participants and researchers must be avoided in all instances.
6. The independence of research must be clear, and any conflicts of interest or partiality must be explicit.

In the United Kingdom, the reviews of research are carried out by Research Ethics Committees, according to procedures established by the organizations that sponsor the research.

The United States

In the United States, the Common Rule (originally established in 1981) is the basic set of regulations and guidelines, primarily aimed at the treatment of human subjects in research. As such, it was one of the first sets of regulations and guidelines established, following public and political outcries over such incidents as the Tuskegee syphilis study (1932–72), and human radiation experiments (1944–74), along with the Milgram experiments on obedience (published in 1963), and the Zimbardo prison experiment (which took place in 1971). The National Research Act of 1974 created the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research which was charged with developing guidelines for the protection of human subjects in research. That commission produced the Belmont Report (1979), which identified three fundamental ethical principles:

1. Respect for persons
2. Beneficence
3. Justice

These principles continue to serve as the basis for the ethical reviews of research in the United States (http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html).

It was also as a result of the National Research Act of 1974 that Institutional Review Boards were mandated, initially for those institutions with federally funded research, but now more generally. Although there have been a number of subsequent bioethics commissions, the 1974 commission that produced the Belmont Report is generally regarded as being the most influential. Today, the Office for Human Research Protection, within the Department of Health and Human Services, has the primary responsibility for enforcing the Common Rule (The Common Rule (45 Code of Federal Regulations 46) is available at http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html).

In addition to the protection of human subjects, the United States also supports a separate agency – the Office of Research Integrity (ORI) – for the detection, investigation, and prevention of research misconduct. ORI was created, in 1992, as a result of congressional hearings on cases of research misconduct that occurred in the late 1970s and early 1980s. According to the federal regulations, research misconduct is defined as "fabrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results…." While fabrication is the use of made-up data or results, falsification is defined as the changing or omitting data or research processes. Plagiarism is defined as the wrongful appropriation of another author's research or ideas. However, research misconduct does not include honest errors and differences of opinion (http://ori.dhhs.gov/federal-research-misconduct-policy).

Since 2000, the role of ORI has expanded beyond the detection and investigation of research misconduct within the Department of Health and Human Services. Today, it is also actively engaged in a program of research and education designed to promote the RCR and research integrity.

Comparison of Regulatory Contexts

These various regulations dealing with the protection of human subjects differ primarily in the details of the approaches they use to address the public responsibility that governments have toward the protection of their citizens. For example, they use different names to refer to their review committees, and the composition and specific activities of those committees vary from country to country. They also differ somewhat in the extent to which their regulations go beyond the protection of human subjects to the RCR. For example, Australia has a separate document relating to RCR, and elements of RCR are present in the Canadian and United Kingdom documents (e.g., confidentiality, conflicts of interest). In contrast, in the United States, this function falls under separate regulations, and is under the auspices of an entirely separate agency.

The regulatory contexts for social/behavioral research are similar in that they all have origins in Nuremberg and Helsinki. Moreover, there is an extraordinary degree to which the government regulations referring to social/behavioral research are similar. All of them are based on sets of core principles, most of which are either identical to or resonate with the Helsinki Declaration or the Belmont Report (e.g., Justice, and Respect for Persons in the United States, Canada, and Australia). All underscore the need to avoid doing harm to the people who are asked to participate in research, and to get their voluntary consent to participate. Also, all call for the creation of entities (i.e., ethics review committees) to review research proposals that involve human subjects. In addition, in each country, these

review committees are also highly decentralized, being located in each institution that supports research activities.

In sum, the regulatory contexts of all four governments reviewed, and indeed, much of the European Union, are remarkably consistent in their basic form and structure. It is in the specificity of definitions and operations where they differ. Thus, researchers seeking to carry out their research cross-nationally must use extreme care to make sure that they do not inadvertently violate regulations that are different from the ones to which they are accustomed.

Major American Social and Behavioral Professional Associations

American Psychological Association

The document, *Ethical Standards of Psychologists*, was first published in 1953. Since then, it has been amended 11 times, most recently in 2010, and is now titled *The Ethical Principles of Psychologists and Code of Conduct* (http://www.apa.org/ethics/code/index.aspx). It has an introduction, a preamble, a set of general principles, and a number of ethical standards. The document explicitly states that the preamble and general principles are aspirational in nature. The five general principles are as follows:

1. Beneficence and nonmaleficence
2. Fidelity and responsibility
3. Integrity
4. Justice
5. Respect for people's rights and dignity

In comparison, the ethical standards are to be considered by psychologists as enforceable rules of conduct. There are 10 Ethical Standards, each with multiple subsections. They cover a wide range of professional topics, including privacy and confidentiality, record keeping and fees, research and publication, education and training, assessment, and therapy.

The American Psychological Association (APA) maintains an Ethics Office. "The APA Ethics Office has five primary roles: to adjudicate ethics complaints; to provide ethics education; to assist the association in developing policies related to the Ethics Code; to promote the presence of, and interest in, ethics throughout the APA membership and governance process; and to offer consultation on ethics-related questions" (Behnke, 2007). As part of its responsibilities, the Ethics Office cooperates with the Ethics Committee in making judgments on complaints. It also supports members and the association in dealing with ethical issues and provides consultation. Lastly, the Ethics Committee offers workshops and seminars on ethical standards (http://www.apa.org/ethics/).

The Ethics Committee is given the responsibility for recommending rules and procedures for its operation, as well as executing investigations of unethical conduct by members of the APA. While the Committee may dismiss an allegation of unethical conduct, it recommends other actions (e.g., censure, expulsion) to the APA Board of Directors. There are two types of investigations typically carried out by the committee: (1) show cause proceedings – where formal action has previously been taken by another body (e.g., a felony conviction); and (2) complaints alleging violations of the ethics code that may be brought either by individuals or by the Ethics Committee itself.

With regard to complaints of unethical conduct, a preliminary evaluation is first undertaken to determine if the respondent (i.e., the subject of the complaint) is a member of the APA (which is required), and if the complaint has been submitted correctly. If a case is opened, the Director of the Ethics Office issues a charge letter to the respondent and conducts an investigation. The respondent is able to comment on all information to be considered by the committee. The Ethics Committee reviews all of the materials and decides how the case will be resolved. The Committee can dismiss the complaint, recommend reprimand or censure, recommend that the respondent resign from the APA, or recommend expulsion. In certain instances, the respondent has the right of appeal, which may include an in-person proceeding before a formal hearing committee (http://www.apa.org/ethics/code/committee.aspx).

American Sociological Association

The *Code of Ethics* of the American Sociological Association (ASA) was initially approved by its members in 1971. In the early years, as it was in the nation at large, ethical behavior in the conduct of research was the focal issue. Over time, a much wider range of ethical issues were recognized as relevant to a Code addressing the professional conduct of sociologists (The full text of the ASA Code of Ethics and the policies and procedures for the Committee on Professional Ethics (COPE) are available on the ASA web site at http://www.asanet.org/about/ethics.cfm).

The development of the code began with the first ad hoc committee on professional ethics, which was created in 1961. However, there was opposition within the membership to the establishment of a formal ethics code, and the project was shelved (Rhoades, 1981). Another ASA committee was appointed in 1967, and it prepared a draft that focused largely on the conduct of research. That initial version of the Code was approved by the ASA membership in 1969, but it was not operational until 1971 when the ASA had completed the task of establishing the procedures for the enforcement of the Code. Since 1971, the Code has been revised four times. The major changes increased attention to the ethical situations presented by the employment of sociologists in government as well as in for-profit and nonprofit organizations, and a section was added on fair employment practices. Revisions to the COPE policies and procedures focused on ensuring the confidentiality of the complaint process and underscored the degree to which mediation should be initially employed in the resolution of complaints.

The current version of the Code was approved by the ASA membership in 1997. It includes an introduction, a preamble, and five general ethical principles:

1. Professional competence
2. Integrity
3. Professional and scientific responsibility
4. Respect for people's rights, dignity, and diversity
5. Social responsibility

These principles of ethical professional conduct are expressly aspirational. They underlie the 20 specific ethical standards (many of which contain multiple subsections) that

are the enforceable standards of ethical conduct for sociologists. The enforcement of the Code through COPE applies only to current members of the Association. In becoming a member, individuals explicitly agree to adhere to the Code's ethical standards and can thus be held accountable. If an ASA member's conduct is determined to have been unethical under the Code, the sanctions available to COPE include: a private reprimand (that may include an action to rectify a wrong); a public reprimand; denial of privileges; and termination of membership for a given period of time (Kennedy, 1996).

At the ASA, the full-time Executive Officer of the Association is the liaison to COPE (although this duty may be delegated). COPE procedures explicitly call for the Executive Officer to attempt to resolve complaints before initiating a formal investigation. In recent years, the ASA has typically received an average of about 20 inquiries annually. Most involve interpretations of the Code or allegations concerning individuals who are not members of the Association. The few remaining are mostly about issues that the Executive Officer or the chair of COPE can respond to informally, and mediate without a full investigation by the committee. During a recent 5-year period, only one official complaint proceeded through full investigation and decision, and this frequency is not unusual. (The material describing the ASA Code and its enforcement are largely drawn from Van Valey and Hillsman, 2007.)

American Anthropological Association

Statement on Ethics: Principles of Professional Responsibility – the ethics code of the American Anthropological Association (AAA) was first published in 1971. Subsequent revisions were published in 1998, 2009, and most recently, in 2012. The document contains a preamble and a set of seven core principles:

1. Do no harm.
2. Be open and honest regarding your work.
3. Obtain informed consent and necessary permissions.
4. Weigh competing ethical obligations due collaborators and affected parties.
5. Make your results accessible.
6. Protect and preserve your records.
7. Maintain respectful and ethical professional relationships.

Each of the principles is accompanied by a brief discussion and link to other resources. The AAA preamble states, "…these principles are intended to foster discussion, guide anthropologists in making responsible decisions, and educate." Consequently, the AAA does not have a mechanism for enforcing its code of ethics (http://www.aaanet.org/profdev/ethics/upload/Statement-on-Ethics-Principles-of-Professional-Responsibility.pdf).

International Sociological Association

The International Sociological Association's (ISA) Code of Ethics includes a preamble and four sets of specific ethical standards to which all members of ISA must adhere (http://www.isa-sociology.org/about/isa_code_of_ethics.htm).

The code was approved in 2001, and the four standards are:

1. Sociology as a field of scientific study and practice
2. Research procedures
3. Publication and communication of data
4. Extra-scientific use of research results

Each of these standards has several subsections. The specific goals of the Code of Ethics are to protect individuals and groups that are in involved in sociological research, and to provide a framework for the research of ISA members (http://www.isa-sociology.org/about/isa_code_of_ethics.htm).

Other Academic and Professional Organizations

Many other academic and professional organizations in the social sciences also have codes of ethics. Of those, most appear to have used either the code of the APA or the code of the ASA as a model (e.g., the Association for Applied and Clinical Sociology). For example, among the regional sociological associations in the United States, the North Central Sociological Association has adopted a version of the ASA code produced by its Committee on the Profession, Freedom, and Responsibility. Others, rather than writing or adapting their own, have taken the position that their members will simply abide by the ASA code (e.g., the Pacific Sociological Association, and the Midwest Sociological Association). Some of the regional sociological associations (e.g., the Southern Sociological Society, the Eastern Sociological Society) do not have explicit codes of ethics. However, they do have committees in place that consider issues of an ethical nature. Nevertheless, with the exception of the ASA itself, none of the other sociological associations have active enforcement procedures in place.

Comparison of Professional Codes of Ethics

It is first apparent that most American professional associations in the social and behavioral sciences do have a code of ethics or some functional equivalent. Indeed, many of them are quite similar in form and even in content. It is also clear that the codes of ethics maintained by these professional associations are far less concerned with the treatment of human subjects than the governmental regulations. While they do adopt (or at least adapt) ethical principles similar to those expressed in the Belmont Report, there is clearly more attention paid to a much wider range of professional roles and behaviors. In each case, there are standards (or statements, as in the AAA) that deal directly with research activities, such as informed consent, respect for people's rights, and the research record. However, there are many more that cover other interpersonal contexts such as exploitation, harassment, and conflicts of interest, as well as professional activities such as teaching, authorship, publication, plagiarism, and clinical or applied work, both in and out of the workplace. This, of course, is to be expected, since the nature of professional associations is to focus on the full range of professional activities and responsibilities carried out by their members.

It is notable that the APA and the ASA are the only major professional social/behavioral science associations with an enforcement process that allows members and others to submit a charge of unethical behavior about a member. In both cases,

it is gratifying that the process is relatively infrequently used. However, also in both cases, there is substantial infrastructure support required. In each case, an organizational officer and a number of members are required to deal with the inquiries and allegations that take place. While most of the support is in the form of the time expended by these people, there are also record keeping, communications, and logistic needs as well.

The governmental process most comparable to code enforcement by professional associations is the detection and investigation of research misconduct by ORI. Yet, the vast majority of infractions reported by the professional associations deal with matters other than research misconduct, while research misconduct was the primary purpose for the creation of the government agency.

The Responsible Conduct of Research

In 2010, at the 2nd World Conference on Research Integrity held in Singapore, a statement on research integrity was established. As an attempt to establish common principles and responsibilities that extend across all research, a set of 4 fundamental principles and 14 responsibilities were agreed upon – *The Singapore Statement*. As its Preamble states, "While there can be and are national and disciplinary differences in the way research is organized and conducted, there are also principles and professional responsibilities that are fundamental to the integrity of research wherever it is undertaken." (The full text and related materials are available at http://www.singaporestatement.org/statement.html.)

The Singapore Statement identifies four fundamental principles:

1. Honesty in all aspects of research.
2. Accountability in the conduct of research.
3. Professional courtesy and fairness in working with others.
4. Good stewardship of research on behalf of others.

In addition, it lays out 14 responsibilities that are necessary to the integrity of research, regardless of where it takes place:

1. Integrity – insure the trustworthiness of research.
2. Adherence to regulations – be aware of and adhere to regulations and policies.
3. Research methods – use of appropriate methods, critical analysis of data, fully and objectively report findings.
4. Research records – clear and accurate, to allow replication and verification.
5. Research findings – share data and findings.
6. Authorship – include only those who meet applicable criteria.
7. Publication acknowledgment – acknowledges others who make contributions but who do not merit coauthorship.
8. Peer review – be fair, prompt, and vigorous, while respecting confidentiality.
9. Conflict of interest – disclose financial and other conflicts that could compromise the trustworthiness of the work.
10. Public communication – limit public communications to areas of expertise.
11. Reporting irresponsible research practices – notify appropriate authorities.
12. Responding to irresponsible research practices – develop procedures for responding to allegations, as well as protecting those who report them.
13. Research environments – encourage integrity through education, policies, and reasonable standards for advancement.
14. Societal considerations – weigh societal benefits against risks.

While *The Singapore Statement* clearly goes well beyond the ethical treatment of human subjects in research, it does identify the direction that ethical guidelines and codes of conduct are moving, not only for the social and behavioral sciences, but also for research in general. The 3rd World Conference on Research Integrity was held in Montreal in May 2013, and as a result issued *The Montreal Statement on Research Integrity in Cross-Boundary Research Collaboration*s (http://wcri2013.org/Montreal_Statement_e.shtml). This document focuses on issues central to cross-national and cross-disciplinary research, and moves the conversations regarding research ethics and the RCR even more broadly than *The Singapore Statement*. The 4th World Conference on Research Integrity is slated to be held in Rio de Janeiro in 2015.

The Future of Research Ethics

The governmental regulations and professional codes of ethics described above are only examples of the kinds of regulations that currently exist around the world – everywhere that research takes place. Given the early events that generated so much governmental concern about research ethics, it is hardly surprising that the treatment of human subjects became the object of so many regulations and oversight. Moreover, there have been an unfortunate number of recent breaches that reaffirm the need for vigilance. After all, the support and success of research – in all fields – ultimately depends on the public trust, and those early events, as well as others later on, challenge that foundation of trust.

The Singapore Statement and *The Montreal Statement* are solid indicators of what appears to be a trend away from the focus primarily on the treatment of human subjects and toward the wider range of issues implicit in research ethics. There are also indications of this trend in some of the governmental regulations (e.g., ORI) and disciplinary statements. Concerns over confidentiality, authorship, publication, conflicts of interest, and other issues do appear. However, at present, they still appear to be largely secondary to the need to protect human subjects from harm. Nevertheless, it does appear that this trend toward research ethics more broadly construed will continue and grow.

See also: Associations, Scholarly; Common Rule: US Regulatory Protections for Human Research Participants; Ethical Dilemmas, Research, and Access to Treatment; Ethical Practices: Institutional Oversight and Enforcement; Fieldwork: Ethical Aspects; Informed Consent of Research Subjects; Privacy of Individuals in Social Research: Confidentiality; Problem Solving and Reasoning: Case-based; Research Funding: Ethical Aspects; Research Publication: Ethical Aspects; Social Science Infrastructure: North America (Research and Teaching); Social Science Infrastructure: North America (Resourcing and Funding); Social Science

Infrastructure: North America (Social Science Academies and Related Organizations); Social Science Professions and Professionalization; Surveys and Polling: Ethical Aspects.

Bibliography

Behnke, S., September 2007. The work of the APA ethics office: frequent calls we receive. Monitor on Psychology 38 (8), 56.

Coleman, C.H., Menikoff, J.A., Goldner, J.A., Neveloff Dubler, N., 2005. The Ethics and Regulation of Research with Human Subjects. Matthew Bender & Co., Newark.

Haney, C., Banks, C., Zimbardo, P., 1973. Interpersonal dynamics in a simulated prison. International Journal of Criminology and Penology 1, 69–97.

Kennedy, J., 1996. Hot off the press: revised draft of the code of ethics. Footnotes 24 (6), 12. American Sociological Association, Washington, DC.

Milgram, S., 1963. Behavioral study of obedience. Journal of Abnormal and Social Psychology 67 (4), 371–378.

Rhoades, L., 1981. A History of the American Sociological Association: 1905–1980. American Sociological Association, Washington, DC, p. 55.

Van Valey, T., Hillsman, S., 2007. The code of ethics of the American sociological association. Bulletin of the Swiss Society of Sociology 132, 11–15.

Relevant Websites

http://www.apa.org/ethics/code/index.aspx – American Psychological Association's Ethical Principles of Psychologists and Code of Conduct.

http://www.asanet.org/about/ethics.cfm – American Sociological Association's Code of Ethics.

http://ethics.aaanet.org/category/statement/ – AAA Ethics Blog. Statement on Ethics.

http://www.ethics.gc.ca/pdf/eng/tcps2/TCPS_2_FINAL_Web.pdf – The Interagency Advisory Panel on Research Ethics (PRE). *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* (TCPS).

http://www.esrc.ac.uk/about-esrc/information/research-ethics.aspx – Economic and Social Research Council. Framework for Research Ethics.

http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html – US Department of Heath & Human Services. The Belmont Report.

http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html – US Department of Heath & Human Services. Code of Federal Regulations.

http://www.isa-sociology.org/about/isa_code_of_ethics.htm – International Sociological Association's (ISA) Code of Ethics.

http://www.nhmrc.gov.au/book/national-statement-ethical-conduct-human-research – National Health and Medical Research Council. National Statement on Ethical Conduct in Human Research (2007) (Updated March 2014).

http://ori.dhhs.gov/federal-research-misconduct-policy – The Office of Research Integrity. Federal Research Misconduct Policy.

http://www.singaporestatement.org/statement.html – Singapore Statement on Research Integrity.

http://wcri2013.org/doc-pdf/MontrealStatement.pdf – The 3rd World Conference on Research Integrity. Montreal Statement on Research Integrity in Cross-Boundary Research Collaborations.

Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute

Jean E McEwen, Joy T Boyer, Nicole C Lockhart, and David J Kaufman, National Human Genome Research Institute, Bethesda, MD, USA
Lawrence C Brody, National Human Genome Research Institute, Bethesda, MD, USA

Abstract

The Human Genome Project was a large coordinated effort designed to produce a highly detailed and dynamic picture of our genetic material. It was clear from the beginning of the Project that the results would raise a host of ethical, legal, and social issues. To address these questions, the National Human Genome Research Institute created the Ethical, Legal, and Social Implications (ELSI) Program, with the charge to fund basic and applied research on the ethical, legal, and social implications of genetic and genomic research for individuals, families, and communities. Nearly 25 years later, 'ELSI' has become a recognized term for research that deals with the social, legal, and ethical issues arising from advances in biomedical science.

Introduction

The Ethical, Legal, and Social Implications (ELSI) Program is a research program at the National Human Genome Research Institute (NHGRI), one of 27 institutes and centers of the US National Institutes of Health (NIH). In the mid-1980s there were a number of discussions about the need to have a detailed map of the human genome. These discussions were formalized in 1990 as the Human Genome Project (HGP) – a project funded jointly by the US Department of Energy and the NIH. The leaders of the HGP were aware that producing highly detailed information about human genomes might lead to discrimination based on an individual's genetic makeup. Also in the background were various discriminatory activities that had earlier carried out in the name of the eugenics movement. To ensure that the social implications of the science were addressed in step with the technical advances, it was decided that a portion of the funding for the HGP would support research on the ethical, legal, and social implications of the research.

Thus, the formal launch of the HGP in 1990 included the creation of the ELSI Program – a program designed to anticipate and address these issues. In the ensuing 24 years, 'ELSI' has become a recognized shorthand for this type of work even beyond human genome research. 'ELSI' is now often used as an adjective when describing other research programs and is also sometimes used as a noun when discussing specific programs, elements of infrastructure, and components of a research program.

The ELSI program received additional recognition as a defined entity via a 1993 US Congressional mandate that decreed that NHGRI shall allocate 5% of its annual research budget to support research into the ethical, legal, and social implications of genetics and genomics research (National Institutes of Health Revitalization Act of 1993). The rationale for the Program is to facilitate rigorous consideration of the implications of advances in genetics and genomics research at the same time that the underlying science is being conducted.

The ELSI Program is currently the only dedicated extramural bioethics research program at the NIH and is the largest funder in the world of research focused on ethical, legal, and social issues in genetics and genomics. The Program's budget has grown from $1.57 million USD in fiscal year 1990 to approximately $18 million USD in 2014. Since its inception, the Program has awarded more than $320 million USD in research support and has funded over 480 projects; collectively, these have resulted in thousands of scientific publications (McEwen et al., 2014).

The Program, which is administratively situated within NHGRI's Division of Genomics and Society (http://www.genome.gov/27550080), is shaped by ongoing advice from the National Advisory Council for Human Genome Research (NACHGR) (http://www.genome.gov/10000905). This body includes experts in genomic science and technology and genetic and genomic medicine as well as experts in emerging ELSI issues. The Program also obtains regular input from the Genomics and Society Working Group, a subgroup of the NACHGR established to make recommendations on short- and long-range planning and priority setting (http://www.genome.gov/27551917).

The present-day ELSI Program is fundamentally a research and training program. Although the studies that the Program supports often help to inform the development of policies related to genomic science and of education and community outreach activities, the direct support of policy development activities is beyond the Program's purview. In the past, the Program supported direct funding for education activities aimed at professionals or the public, but funding for education projects was discontinued in 2004 based on growing recognition that the Program was neither financially nor organizationally equipped to contribute meaningfully to the enhancement of 'genomic literacy' by producing educational materials. However, research projects designed to investigate the best practices to enhance and extend the knowledge of genomics still fall within the portfolio. In addition, the Program supports a variety of meetings, workshops, and policy conferences, some of which include educational components. The Program also supports many research training activities, and has been especially committed to

training aimed at increasing the diversity of the investigators who conduct ELSI research.

Overview of the ELSI Program Portfolio

The current research priorities of the ELSI Program fall into four broad categories: psychosocial and ethical issues in genomics research; psychosocial and ethical issues in genomic medicine; legal and public policy issues; and broader societal issues (Green et al., 2011). In 2014, roughly 40% of the current budget funded studies of genomics research issues, 33% to studies of issues in genomic medicine, 11% to legal and policy studies, and 16% to studies of broader societal issues (McEwen et al., 2014). These percentages are not fixed and the Program has been designed to evolve as genomic science and genomic medicine advance.

The Program's research portfolio incorporates work by investigators from a broad range of disciplines, including (among others) genetics and genomics, clinical medicine, bioethics, the social sciences (e.g., psychology, sociology, anthropology, political science, and communication science), history, philosophy, literature, law, economics, health services, and public policy. Many of the individual grants are themselves highly multidisciplinary. Funded projects employ a wide range of empirical (both quantitative and qualitative) and nonempirical methodologies. These range from experimental and quasi-experimental trials, surveys, focus groups, and structured and semistructured interviews, to ethnographic, legal, philosophical, archival, and oral history research. The ELSI Program aims to maintain a balance of grants that utilize empirical and nonempirical methods, and nearly half the grants in the Program's portfolio employ multiple methods.

Projects funded by the Program may be undertaken either as stand-alone efforts or embedded in much larger multidisciplinary projects with multiple objectives, ranging from technology improvement to clinical trials. In the latter types of projects, the ELSI implications of the science can be investigated in parallel and in real time along with the primary objectives of the underlying study.

Funding Mechanisms and Mechanics

The ELSI Program solicits both investigator-initiated and program-initiated research. Investigator-initiated research applications are solicited through general Program Announcements that contain broad statements of ongoing areas of programmatic interest.

Program-initiated studies – studies of particular high-priority topics that are periodically identified as requiring immediate or more focused attention – are solicited through targeted Requests for Applications or Requests for Proposals. Past examples of these Program-initiated projects include calls for research into such topics as the ethical and legal implications of cancer genetics predisposition testing, the societal implications of genetic variation research, and the issues raised by the return of genomic research and test results and incidental findings.

A currently active example of a Program-initiated project is the Centers of Excellence in ELSI Research (CEER) Program (http://www.genome.gov/1504773). The purpose of the CEER Program is to support the creation and maintenance of the infrastructure necessary to foster highly transdisciplinary academic research, facilitate the translation of such research into health, research, and public policies and practices, and train the next generation of investigators in the field. In its current form, this program consumes roughly one-third of the ELSI budget.

Although the ratio of investigator-initiated to Program-initiated research fluctuates from year to year, historically, most of the research funded by the Program has been investigator-initiated, reflecting the Program's desire to promote the intellectual independence of its supported investigators. All grants submitted to the Program undergo a peer review process that utilizes the standard review criteria applied to all NIH applications.

Program Priorities – Breadth versus Depth

In setting program priorities, the ELSI Program strives to maintain a balance between breadth and depth in its research portfolio, while taking into account the broad range of stakeholder interests involved. The need for careful priority-setting has been especially pronounced since 2010 because of the shift that is occurring in the field of genomics from an emphasis on basic research (research aimed at understanding the structure and biology of genomes and the biology of disease) toward an emphasis on research with identified human participants and immediate clinical applications (Green et al., 2011). The latter type of research raises an expanded array of ethical issues, which the ELSI Program is increasingly called upon to address.

Pressures on the Program to spread its limited research dollars across a wider range of topic areas began to occur just as the budget for NIH (and thus NHGRI) became more restricted, which has heightened priority-setting challenges for the Program. For this reason, the Program must continually reassess the extent to which its resources should be allocated to the support of certain categories of research which, in a less-constrained budgetary climate, would likely have been supported, such as studies that are 'disease-specific' as distinct from those aimed at the development of models with potentially broader applicability. Similar questions have begun to arise about the extent to which the Program's resources should be devoted to research on issues arising from uses of genomic information in nonmedical contexts (i.e., studies addressing topics such as potential uses of genomic data in insurance, employment, education, criminal justice, family law, and the military) or research that has a non-US or explicitly international focus.

Integration of ELSI Research with Genomics Research and Policy Activities

Interactions between the ELSI and genomics research communities and between the ELSI research and policy activities have always been a feature of the ELSI Program, but such

interactions have grown more frequent since the early 2000s. On the research side, as earlier noted, the CEER Program has as an explicit goal the building of transdisciplinary research teams that include both ELSI and genomics investigators. Several large community resource projects coordinated by NHGRI also provide opportunities for interactions between genomics investigators and those with expertise in the ethical, legal, and social issues those projects raise. Opportunities for collaborations have come about through 'stand-alone' RFAs issued to support ELSI research conducted in parallel with corresponding genomics initiatives, through the establishment of Working Groups for large initiatives to provide guidance on particular issues, and more recently, through the direct 'embedding' of ELSI research within genomics research protocols.

On the policy front, many investigators supported by the ELSI Program participate actively with various policy bodies. At the national level, these include the Secretary's Advisory Committee on Genetics, Health and Society, the Secretary's Advisory Committee on Human Research Protections, the Discretionary Advisory Committee on Heritable Disorders in Newborns and Children, the Presidential Commission for the Study of Bioethical Issues, and the policy and bioethics committees of most of the major professional societies related to genetics.

Increasing integration between the ELSI research enterprise and the genomics research and genomics policy enterprises carries with it a number of significant benefits. For example, in the research realm, directly embedding ELSI research into active, ongoing genomics research provides a 'natural laboratory' for conducting such studies, sometimes generating more reliable and more practically relevant findings than those arising from hypothetical studies. Closer collaborations between ELSI investigators and genomics researchers have also helped ELSI investigators become more conversant with genomic science and genomic investigators become more conversant with some of the methods used in other disciplines (including normative methods) – thus enhancing mutual respect and understanding. On the policy side, increasing integration between ELSI investigators and those charged with the development of policy has tended to increase the policy relevance of the research. This increases the amount and quality of evidence on which to base policies related to genomics whether in the sphere of research, clinical medicine, or society more broadly, and enhances the likelihood that these policies will be solidly evidence based.

On the other hand, increased integration between the ELSI research enterprise and the genomics research and policy enterprises has occasionally come with a downside, as some risk exists that ELSI investigators will come to be seen as mere 'facilitators' of genetic and genomic research, or as tools of others' particular policy agendas. Indeed, concerns have occasionally been expressed that encouraging too close an integration between the enterprises runs the risk of a subtle erosion in the legitimacy of ELSI investigators' standing as independent and autonomous academic investigators. In fact, however, a large portion of the ELSI budget is spent on investigator-initiated grants, helping to ensure the continuation of a healthy independent ELSI research agenda.

The challenges involved in promoting enhanced integration among ELSI research, genomics research, and policy development activities, while simultaneously maintaining the objectivity, intellectual independence, and integrity of ELSI investigators, have been well recognized since the inception of the ELSI Program. To a large extent, these challenges arise from the simple fact that the Program, while charged with supporting critical investigations into the ethical, legal, and social implications of genomics research, is organizationally, and indeed physically, situated within the very agency that supports the underlying science.

Impact of the ELSI Program

Assessing the impact of the ELSI Program is challenging for several reasons. First, many of the studies funded by the ELSI Program involve basic research (e.g., the clarification of terminology and concepts or the analysis of moral frameworks, values, or cultural constructs) – a kind of research that does not always have a direct impact on policy or practice but which does help to provide a foundation on which more applied studies can be built. Second, not all of the issues addressed by ELSI research have direct relevance to specific policy enactments (at least under a narrow definition of 'policy'). Third, the trajectory between the dissemination of ELSI research findings and the establishment of research, clinical, or broader societal policies related to genomics tends to be nonlinear. Finally, as ELSI research is often aimed at anticipating where the science and technology will be in the future, the program may fund research into the implications of genomic methods, techniques, and applications that fail to materialize.

Indeed, the impact of ELSI research has often come less from the direct 'translation' of published study findings into a formal embodiment of policy than from ELSI investigators – operating independently as scholars, beyond the immediate scope of their NIH-funded activities – serving directly on particular commissions or other policymaking bodies or providing expert testimony or other forms of expert analysis to those groups (Burke et al., 2014). The most consequential impact of ELSI research may, in fact, come about in even more subtle ways, such as through the contributions the studies have made to incremental changes in the cultural milieu in which genomics research is conducted, genomic medicine is implemented, and genomic information is incorporated into decision-making in various areas of society more broadly.

In the realm of genomics research, for example, ELSI research on issues related to privacy and the potential identifiability of genomic data (Kaufman et al., 2009; McGuire, 2008) has led to changes in the way investigators draft, and the way institutional review boards review, consent forms and consent processes for genomic studies. Such research has also influenced the ongoing development of NIH-wide policies on genomic data sharing and proposed revisions to US federal regulations regarding the conduct of research.

Research on the ethical issues relating to the use of stored genetic samples (Clayton et al., 1995; Fullerton et al., 2010) has influenced the development of policies and practices for biobanks and biorepositories and the development of innovative governance mechanisms for these entities. Research on community engagement, community consultation, and group interests in genetic variation research, and on the ethical and

social considerations involved in the labeling of populations (Foster et al., 1999; Sankar and Cho, 2002) has led to the development of novel protocols to engage and consult with the socially defined communities approached for participation in several major international research projects and to the adoption of more precise and ethically sensitive nomenclature for the populations studied. Ongoing studies of novel approaches to community-based participatory research (James et al., 2012; Kelley et al., 2012) are serving as a model for genomics research conducted with diverse populations and for emerging 'Citizen Science' initiatives.

The impact of research supported by the ELSI Program can also be observed in the evolution of policies related to the use of genomic information in the clinic. A long line of studies of the ethical and social issues in various settings, including population-based screening (Anderson et al., 2005; Clayton et al., 1996), carrier testing (Callanan et al., 1999; Faden et al., 1994), prenatal testing (Botkin, 1998; Robertson, 2003), newborn screening (Botkin, 2005; Waisbren et al., 2003), and predictive testing of both children and adults for both early and late-onset conditions (Burke et al., 1997a,b; Tercyak et al., 2007), and (most recently) whole exome and whole genome sequencing have contributed to the development of a number of Points to Consider documents, recommendations, guideline statements issued by professional organizations, and occasionally to enactments by other policymaking bodies. Studies of the attitudes of research participants and patients about the return of results and incidental findings (Kaufman et al., 2009) and studies of the experiences of customers of direct-to-consumer genetic test marketing companies with the receipt of genetic test results (Bloss et al., 2011; Bollinger et al., 2013) are contributing to researchers' and clinicians' awareness of what people want and expect, how they perceive risk, and how complex genomic information can be communicated in ways that are easier for people to comprehend.

In the broader societal domain, normative and legal analyses relating to genetic privacy and the risk of genetic discrimination supported by the Program, especially in its earliest years (Hudson et al., 1995; Rothenberg et al., 1997), helped create momentum that led, sequentially, to the issuance by the Equal Employment Opportunity Commission of guidance suggesting that discrimination based on genetic predisposition is prohibited by the Americans with Disabilities Act, to an Executive Order protecting federal employees from genetic discrimination in the workplace, to the inclusion in the Health Insurance Portability and Accountability Act of a provision prohibiting group health insurers from excluding individuals from group coverage based on genetic predisposition, and eventually, to passage of the US Genetic Information Nondiscrimination Act, which prohibits genetic discrimination in most areas of health insurance and employment. Research on ethical, legal, and social issues related to uses of DNA in the courtroom (Noble and Moulton, 2006; Scherr, 2013) has helped to sensitize judges, lawyers, civil liberties advocates, and members of the law enforcement community to the issues involved in the collection and potential use of DNA samples from people brought into the criminal justice system.

Research on the ethical and social implications of emerging behavioral genetics findings (Parens et al., 2006; Phelan, 2005) has led gradually to greater media awareness of the need for nuanced discussions of this issue and has heightened awareness among judges, and others involved in the legal system of the complexities involved in using behavioral genetics evidence in courtrooms and other societal structures. In addition, legal and economic analyses regarding the effects of gene patents and other types of intellectual property protection on genomics research and on access to genetic and genomic tests (Henry et al., 2002; Rai and Sampat, 2012) have provided data that have helped to inform policy development in this area.

Conclusion

In the first 24 years of the ELSI Program, the research it funded often focused on the theoretical possibilities of genomic science. Today, the sequence of the human genome has been completed and genomic technologies are improving at a rapid pace. This setting makes it even more critical to carefully examine and explore the implications of these scientific advances. Ignoring the social, legal, and ethical implications of this work could lead to a loss of return on scientific investment if the products of these advances are not understood or accepted. Supporting a research agenda that engages the public and various professional communities that are likely to be impacted by the work can help to ensure that the scientific and technical advances will not be lost in translation.

See also: Biobanking: Ethical Issues; Bioethics in the Post-genomic Era; Bioethics: Genetics and Genomics; Bioethics: Philosophical Aspects; Citizen Science; Community Genetics; Direct-to-Consumer Personal Genetic Testing; Eugenics as a Basis of Population Policy; Eugenics as an International Movement; Eugenics in Europe; Eugenics in the Americas; Eugenics, History of; Forensic Genetic Databases: Ethical and Social Dimensions; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetics and Forensics; Genetics and Social Justice; Genetics and Society; Genetics: Legal Aspects; Human Genome Diversity Project: History; Human Genome Project: History and Assessment.

Bibliography

Anderson, R.T., Press, N., Tucker, D.C., Snively, B.M., Wenzel, L., et al., 2005. Patient acceptability of genotypic testing for hemochromatosis in primary care. Genetics in Medicine 7 (8), 557–563.

Bloss, C.S., Schork, N.J., Topol, E.J., 2011. Effect of direct-to-consumer genomewide profiling to assess disease risk. The New England Journal of Medicine 364 (6), 524–534.

Bollinger, J.M., Green, R.C., Kaufman, D., 2013. Attitudes about regulation among direct-to-consumer genetic testing customers. Genetic Testing and Molecular Biomarkers 17 (5), 424–428.

Botkin, J.R., 1998. Ethical issues and practical problems in preimplantation genetic diagnosis. The Journal of Law, Medicine & Ethics 26 (1), 17–28, 3.

Botkin, J.R., 2005. Research for newborn screening: developing a national framework. Pediatrics 116 (4), 862–871.

Burke, W., Daly, M., Garber, J., Botkin, J., Kahn, M.J., et al., 1997a. Recommendations for follow-up care of individuals with an inherited predisposition to cancer. II. BRCA1 and BRCA2. Cancer genetics studies consortium. JAMA 277 (12), 997–1003.

Burke, W., Petersen, G., Lynch, P., Botkin, J., Daly, M., et al., 1997b. Recommendations for follow-up care of individuals with an inherited predisposition to cancer. I. Hereditary nonpolyposis colon cancer. Cancer genetics studies consortium. JAMA 277 (11), 915–919.

Burke, W., Appelbaum, P., Dame, L., Marshall, P., Press, N., Pyeritz, R., Sharp, R., Juengst, E., 2014. The translational potential of research on the ethical, legal, and social implications of genomics. Genetics in Medicine.

Callanan, N.P., Cheuvront, B.J., Sorenson, J.R., 1999. CF carrier testing in a high risk population: anxiety, risk perceptions, and reproductive plans of carrier by "non-carrier" couples. Genetics in Medicine 1 (7), 323–327.

Clayton, E.W., Hannig, V.L., Pfotenhauer, J.P., Parker, R.A., Campbell 3rd, P.W., Phillips 3rd, J.A., 1996. Lack of interest by nonpregnant couples in population-based cystic fibrosis carrier screening. The American Journal of Human Genetics 58 (3), 617–627.

Clayton, E.W., Steinberg, K.K., Khoury, M.J., Thomson, E., Andrews, L., et al., 1995. Informed consent for genetic research on stored tissue samples. Journal of the American Medical Association 274 (22), 1786–1792.

Faden, R.R., Tambor, E.S., Chase, G.A., Geller, G., Hofman, K.J., Holtzman, N.A., 1994. Attitudes of physicians and genetics professionals toward cystic fibrosis carrier screening. American Journal of Medical Genetics 50 (1), 1–11.

Foster, M.W., Sharp, R.R., Freeman, W.L., Chino, M., Bernsten, D., Carter, T.H., 1999. The role of community review in evaluating the risks of human genetic variation research. The American Journal of Human Genetics 64 (6), 1719–1727.

Fullerton, S.M., Anderson, N.R., Guzauskas, G., Freeman, D., Fryer-Edwards, K., 2010. Meeting the governance challenges of next-generation biorepository research. Science Translational Medicine 2 (15), 15cm3.

Green, E.D., Guyer, M.S., NHGRI, 2011. Charting a course for genomic medicine from base pairs to bedside. Nature 470, 204–213.

Henry, M.R., Cho, M.K., Weaver, M.A., Merz, J.F., 2002. Genetics. DNA patenting and licensing. Science 297 (5585), 1279.

Hudson, K.L., Rothenberg, K.H., Andrews, L.B., Kahn, M.J., Collins, F.S., 1995. Genetic discrimination and health insurance: an urgent need for reform. Science 270 (5235), 391–393.

James, R., Starks, H., Segrest, V.A., Burke, W., 2012. From leaky pipeline to irrigation system: minority education through the lens of community-based participatory research. Progress in Ccommunity Health Partnerships: Research, Education, and Action 6 (4), 471–479.

Kaufman, D.J., Murphy-Bollinger, J., Scott, J., Hudson, K.L., 2009. Public opinion about the importance of privacy in biobank research. The American Journal of Human Genetics 85 (5), 643–654.

Kelley, M., Edwards, K., Starks, H., Fullerton, S.M., James, R., et al., 2012. Values in translation: how asking the right questions can move translational science toward greater health impact. Clinical and Translational Science 5 (6), 445–451.

McEwen, J.E., Boyer, J.T., Sun, K.Y., Rothenberg, K.H., Lockhart, N.C., Guyer, M.S., 2014. The ethical, legal, and social implications (ELSI) program of the national human genome research institute: reflections on an ongoing experiment. Annual Review of Genomics and Human Genetics 15 (4), 1–4.

McGuire, A.L., 2008. Identifiability of DNA data: the need for consistent federal policy. The American Journal of Bioethics: AJOB 8 (10), 75–76.

Noble, A.A., Moulton, B.W. (Eds.), 2006. DNA fingerprinting and civil liberties. The Journal of Law, Medicine & Ethics 34 (2), 171–475.

Parens, W., Chapman, A.R., Press, N., 2006. Wrestling with Behavioral Genetics: Science, Ethics, and Public Conversation, first ed. The Johns Hopkins Univesrity Press, Baltimore.

Phelan, J.C., 2005. Geneticization of deviant behavior and consequences for stigma: the case of mental illness. Journal of Health and Social Behavior 46 (4), 307–322.

Rai, A.K., Sampat, B.N., 2012. Accountability in patenting of federally funded research. Nature Biotechnology 30 (10), 953–956.

Robertson, J.A., 2003. Extending preimplantation genetic diagnosis: the ethical debate. Ethical issues in new uses of preimplantation genetic diagnosis. Human Reproduction 18 (3), 465–471.

Rothenberg, K., Fuller, B., Rothstein, M., Duster, T., Ellis Kahn, M.J., et al., 1997. Genetic information and the workplace: legislative approaches and policy changes. Science 275, 1755–1757.

Sankar, P., Cho, M.K., 2002. Toward a new vocabulary of human genetic variation. Science 298 (5597), 1337–1338.

Scherr, A.E., 2013. Genetic privacy & the fourth amendment: unregulated surreptitious DNA harvesting. Georgia Law Review Winter 47 (2), 445.

Tercyak, K.P., Peshkin, B.N., Demarco, T.A., Patenaude, A.F., Schneider, K.A., et al., 2007. Information needs of mothers regarding communicating BRCA1/2 cancer genetic test results to their children. Genetic Testing 11 (3), 249–255.

103rd US Congress. 1st Session, 1993. In: National Institutes of Health Revitalization Act of 1993. Title XV, Subtitle C, Sec. 1521, Subpart 4, Secs. 485B(a)(6) and 485B(b)(1).

Waisbren, S.E., Albers, S., Amato, S., Ampola, M., Brewster, T.G., et al., 2003. Effect of expanded newborn screening for biochemical genetic disorders on child outcomes and parental stress. JAMA 290 (19), 2564–2572.

Relevant Websites

http://www.genome.gov/15014773 – Centers of Excellence in ELSI Research (CEER).
http://www.genome.gov/27550080 – Division of Genomics and Society.
http://www.genome.gov/10000905 – National Advisory Council for Human Genome Research.
http://www.genome.gov/27551917 – NHGRI Genomics and Society Working Group.

Ethical Practices: Institutional Oversight and Enforcement

Sophia F Dziegielewski, College of Health and Public Affairs, University of Central Florida (UCF), Orlando, FL, USA; and Merritt Island, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.J. Levine, volume 7, pp. 4770–4774, © 2001, Elsevier Ltd.

Abstract

This article provides an overview of the history, development and regulations, composition and performance, hopes and hindrances related to the committees in the United States charged with the protection of human subjects in research studies. Often referred to as Ethical Review Committees, Ethical Review Boards, or the Institutional Review Board, the roles of these risk-management committees are explored, with an emphasis on contemporary issues and concerns as well as the power they have in requesting modifications to scientific protocol and in making the ultimate decisions for approval.

Lessons from the Past and Development of Regulation

Most researchers are bound to a professional code of ethics and operate within ethical guidelines. Simply defined, in clinical practice ethics are rules to help professionals distinguish between what is right and wrong. When ethical practice becomes research, however, the focus switches slightly to focusing on minimizing the risks of those who participate while conducting scientifically valid research. The fluid nature of ethical conduct, by definition, is a continuous process (Fujii, 2012). Ethical practice, with its inherent decision-making principles, needs constant updating and reexamination to ensure that environmental and technological advancements are included. Therefore, there is rarely a simple response or check-box answer that will ensure that all human subjects are protected. Ethical decision-making is a complex task, and it can also be somewhat elusive.

The rights and welfare of human subjects have not always been considered; at times individuals have been used with no concern for their autonomy or safety. Many professionals are familiar with the Nazi medical experiments of the 1930s and 1940s where brutal mutilative surgery was performed in the name of science. Following World War II (1946–49), tribunals were held in the Palace of Justice in Nuremberg, Germany to review the travesties that were reported. These tribunals recommended a set of research ethics termed 'The Nuremberg Code.' The Nuremberg Code of 1947 outlined 10 principles related to ethical conduct. This Code was the first to emphasize the importance of protecting research participants by means of informed consent with the absence of coercion and properly designed experimentation that minimizes risk and maximizes benefits. In the 1960s, the issue of human participant rights was further highlighted by the Declaration of Helsinki formulated by the World Medical Association in Helsinki, Finland. This document went beyond the Nuremberg Code by focusing primarily on ethical principles related directly to medical research and is often considered an overarching precursor to the guidelines practiced today in the United States. The Declaration of Helsinki provided the basis for an international standard that recognized the importance of biomedical research involving human subjects and that mandated that each research protocol be reviewed by a specially appointed Ethical Review Committee (ERC) (World Medical Association, 2000: Article 13).

The horrors uncovered in the Nuremberg tribunals, were brought closer to home when the American people were alerted to the Tuskegee Syphilis Study. This shocking violation of human dignity still ripples through the hearts and minds of a previously trusting public. This study first began in 1932 when the U.S. Public Health Service (USPHS) started a 40-year clinical study of the effects of untreated syphilis on black males. Over 600 low-income study subjects were promised free health care by the government, along with meals and burial rights. When a cure for the disease was discovered in 1940s (penicillin), the participants were not told and some were actually prevented from receiving treatment. In May 1997, President Clinton issued a formal apology to the survivors of the Tuskegee study and their families. In addition, a second study, from 1946 to 1948, also supported by the USPHS, in collaboration with Guatemalan health authorities, deliberately infected prisoners with sexually transmitted diseases such as syphilis. The Guatemalan prisoners were provided treatment, however, this was all done without their informed consent. After the discovery and exposure of the Guatemala study in 2010, the President of the United States issued an apology to the Guatemalan people in general because those individuals who were actually harmed could not be found (NIH, Fact Sheet, n.d.).

Given the misuse of research subjects in the past it became clear that a set of guidelines was essential if research ethics were to be put into practice. The National Research Act was signed into law by the U.S. Congress on 12 July 1974. This Act authorized the creation of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. This commission identified basic ethical principles and in 1978 released a document entitled *The Belmont Report*. *The Belmont Report* identified three major principles: (1) respect for persons; (2) beneficence and 'do no harm'; and (3) justice (The Belmont Report, 1979).

For researchers in the United States, this effort culminated with what is usually referred to as the 'Common Rule.' This regulation is advanced by the Office of Human Research Protections (OHRP) which is responsible for safeguarding the rights, welfare, and well-being of research subjects in the

conduct of federal research. To ensure compliance OHRP, supported by the U.S. Department of Health and Human Services (HHS), oversees biomedical and behavioral science research institutions ensuring compliance with HHS regulations as described in Title 45, Part 46, Code of Federal Regulations (45 CFR 46). The Common Rule defines research as a systematic investigation designed to develop or contribute to generalizable knowledge. This includes research, testing, and evaluation on human subjects. The Common Rule further defines human subjects as any living individual where data about the individuals are obtained either through intervention or interaction or includes private, identifiable data [45 CFR 46.102(f)]. "Thus, according to the Common Rule, as currently written, research involving material from deceased individuals (e.g., autopsy material) or the use of specimens that are completely anonymous (i.e., a link to subject identity does not exist), would not be subject to the Common Rule, although state and local regulations and policies may apply" (Bledsoe and Grizzle, 2013: p. 324).

When looking specifically at nonbiomedical research, effective and efficient application of the Common Rule requires:

1. an Institutional Review Board (IRB) that has sufficient expertise in social and behavioral research;
2. an IRB that understands and utilizes the flexibility in the Common Rule; and
3. investigators who understand the potential for social and psychological risk in their research.

The Health Insurance Portability and Accountability Act

The importance of those in clinical practice as well as researchers gaining access to a participant's health-related information is coupled with a need to protect the privacy and ensure the consent of all research participants. This need to control a patient's identifiable information provided the impetus for the creation of the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule in 1996, (Health Information Privacy, 2014). Based on previous incidents of researcher misconduct and the public fear that private information would be used or misused without proper informed consent, guidelines were put in place to control access to both health and mental health-related information. Designed to protect the consumer, this Act identified and clearly defined what constitutes a participant's Protected Health Information (PHI). At times PHI may need to be released and this Privacy Rule allows for the release of such disclosures for treatment or other purposes with the expectation that certain protections be put into place. In practice today, HIPAA has carried over into research activity and when PHI is collected, a detailed plan for outlining what will be collected and analyzed, how it will be retrieved, who will have access, how it will be securely stored, and when it will be destroyed will be a necessary part of the participant's awareness requiring consent. The researcher must explain and justify his or her reasoning when requesting a waiver of participant, such as when the study is minimal risk, the waiver will not adversely affect the rights and welfare of the subjects, and the research would not be practicable if consent were required.

Ethical Oversight in the United States: The Formulation of the IRB

The Nuremberg Code and the original Declaration of Helsinki were helpful in formulating the basis for ethical conduct of research, but made no direct statement as to the importance of formulating an ethical review panel. Historically, the expectation for ethical research was placed on the investigator with little, if any, system of checks and balances. In the United States, the first federal document requiring committee review was issued on 17 November 1953. This document and its guidelines applied only to research conducted at the newly opened Clinical Center at the National Institutes of Health (NIH) (Lipsett et al., 1979). The first mention of committee review from an international perspective was in Tokyo with a revision of the Declaration of Helsinki (World Medical Association, 1975).

Learning from the mistakes of the past, all research-savvy institutions in the United States that conduct research involving human subjects have access to an *IRB*. In some cases, the IRB may be viewed as an independent ethics committee or referred to as an *Ethical Review Board (ERB) or Research Ethics Committees (REC)*. In countries other than the United States, the term REC is most common, whereas in the United States the most common term is IRB. In Canada, the term Research Ethics Board (REB) has been used with similar authority to approve or disapprove research proposals (Medical Research Council of Canada, 1987). In the United States, one significant pronouncement that can be seen as a forerunner for today's IRB came on 8 February 1966, from the Surgeon General of the USPHS. The Surgeon General issued the first federal policy statement requiring research institutions to establish the committees which subsequently became known as RECs and ultimately IRBs (Curran, 1970). Focusing on USPHS grant funded research, the grantee's institution, in cooperation with the investigator and his or her institutional associates, were expected to be responsible for the oversight of projects involving human subjects.

The conception of IRB oversight rests in three important concepts. (1) These boards serve as the official designated guardians of the rights and welfare of the participants that serve in research studies under their purview. As central units, it is clear that these boards have the responsibility and are empowered to review and monitor the research activities and ultimately have the ability to approve or deny the research protocol being submitted. (2) As the scope of research extends far beyond traditional expectations, an updated systematic review is needed that takes into account participant awareness, consent, and safety, but also balances federal regulation with ethical expectations in research conduct. Lastly (3), the IRB is expected to ensure that the research is conducted in a transparent and scientifically feasible way while evaluating whether it is ethically sound, thereby protecting the rights of all human subjects involved.

Defining vulnerable subjects can be problematic. In its simplest definition 'vulnerable' includes individuals who, in

health care or research are at the greatest risk of being harmed or have greater susceptibility to being harmed (Hurst, 2008). Protecting human subjects who are most vulnerable rests at the heart of ethical conduct. This type of protection requires additional attention to any structural or systematic bias and introduces a *subject surrogate* that serves as an expert in the area. For example, related to the unique and vulnerable role of prisoners, DHHS regulations require that an REC/IRB member regularly review research that involves this vulnerable category of subjects. This board representative is specifically assigned to review any research related to this population. This reviewer scrutinizes these protocols to ensure that the rights of vulnerable populations have been addressed. Other vulnerable groups include children, pregnant women, and those that are considered mentally incompetent.

Ensuring Ethical Conduct: The CFR

Today, in the United States it is clear to all that human participant research must be conducted in compliance with federal regulations. The Division of Education and Development provides guidance to individuals and institutions conducting HHS-supported human subjects' research. The Division of Policy and Assurances administers the *Federal-wide Assurance* (FWA) of compliance and registration of IRBs. When an IRB is formed a FWA is completed with the HHS, and, therefore, must meet all the HHS regulations. All risk-management reviews need to be conducted with subsequent approval required by law. All research-related activities carried out under the institution's auspices that are supported by federal funds are subject to regulation by the Food and Drug Administration. Often referred to as 'checking the box,' in addition to the legal requirements for federally funded research, many institutions also voluntarily extend the requirement for IRB review to all research involving human subjects.

Once an IRB is established (as published in the Federal Register on 15 January 2009), it must be registered. IRB registration falls under (74 FR 2358), 21 CFR Part 56, IRBs, and amended with regard to IRB registration (21 CFR 56.106). HHS monitors and updates this Internet-based registration system for federally funded research. In addition, an assurance document is generated between the institution and HHS that the institution will comply with federal guidelines for the protection of human subjects and adhere to regulations [45 CFR part 46] with the assurance mechanism described in 45 CFR 46.103.

Understanding how membership of this committee is established is an important first step in ensuring that it can complete its important tasks. IRB membership must include at least five appointed members. It is expected that every effort should be made to create a culturally and professionally diverse committee taking into account race, gender, and cultural backgrounds, as well as affiliation. There is no regulation that says that members cannot be compensated. In addition, at least one of the members should be from the scientist area and one must be a nonscientist. There should also be at least one appointed member that has no connection with the institution apart from committee membership, although absence of this member can still constitute a quorum for decision-making. There are many professionals that can serve on the committee and depending on the types of research reviewed, composition can vary. It is important to note that IRB members do not need to cover every discipline or research area; when needed the IRB can and should obtain information and guidance from outside consultants as they make their research determinations.

Establishment of the IRB: Basics for Structure and Functioning

For many institutions the IRB is located in the facility and remains an integral part of its research activities. For others, the activities of the IRB may be outside of the institution, often referred to as *central review boards*. Regardless, the oversight of the studies conducted by either affiliated or nonaffiliated researchers is their purview.

The purpose of the IRB, ERB, or the ERC is clear in that it is designed to protect the rights of human subjects that participate in research-related activities. The goal of these boards is to ensure that research involving human subjects is designed to conform to relevant ethical and legal standards while safeguarding the rights and welfare of individual research subjects, ensuring informed consent, and minimizing risks while maximizing any anticipated benefits. Originally, these committees were created to provide oversight for biomedical research; however, today their domain has been extended to also cover research in the social and behavioral sciences. To complete this task, these boards are expected to perform a type of risk–benefit analysis that ensures all research conducted with its oversight will be performed ethically and legally within expected compliance standards. When research is conducted in the United States, every precaution needs to be taken to ensure that regulations designed to address the needs of participants are followed.

Membership of the IRB should contain an interdisciplinary group of professionals made up of diverse professionals such as doctors, nurses, lawyers, scientists (including those skilled in the biomedical area and behavioral scientists such as psychologists and social workers) as well as clergy that bring professional expertise to the review process. These professionals under the leadership of a Chair utilize a jury model that incorporates a risk-management strategy capable of confirming scientific merit as well as ethical conduct always reflecting common sense principles of reasonable expectations.

Avoiding Conflicts of Interest

Although incredible gains in knowledge and human understanding and betterment may result from research findings, how they were obtained and what was done to human subjects to achieve them remains a core issue. Secondary gain and whether the researcher may personally benefit from the research study requires careful consideration.

To help avoid the potential for 'conflict of interest,' all studies, especially those that are deemed to be more than minimal risk require the diverse membership of a board of

scientists and nonscientists that oversee the process. In addition, it is expected that at each meeting – prior to starting protocol reviews and the risk-management discussions – that members should identify any conflicts of interest. Once identified, the member should absent him or herself from discussions and abstain from voting on any relevant protocols. The Chair of the IRB is responsible for ensuring this is done at each meeting and if the Chair has a conflict, a substitute Chair for the proceedings should be appointed.

Mission Creep and Criticisms of the IRB

IRB leadership and participation is not for the squeamish and it should come as no surprise that this type of work rests heavily on all involved. Historically, complaints about the IRB overstepping its bounds are nothing new. Concerns related to the development of ethical guidelines, codes, or sets of procedures concerning the 'conduct of research' prevailed in the medical research community (Curran, 1970). Furthermore, in the 1970s biomedical scientists and social behavioral scientists alike remained harshly critical of the role of the IRB and its risk-assessment practices (Levine, 1988). The social behavioral researchers felt strongly they were far less likely to harm human subjects, when compared to a biomedical scientist, and they questioned whether they should be held to the same rules. The behavioral scientists believed that because they did not complete medical procedures or gather tissue from participants and only spoke with them that equivalent regulations were an unconstitutional constraint on their freedom of speech (Levine, 1988).

Similar to earlier objections, concerns are still voiced regarding many areas of research oversight, including community-based participatory research. These complaints from investigators related to oversight of their research activities are not uncommon and at times the assumption is they fall upon deaf ears (Wolf, 2010). At the root of this controversy is whether the REB/REC/IRB really has the skill and expertise to approve or disapprove the scientific design of research protocols (Levine, 1988). After all, who knows better the scientific design than those who created it? This requires that REB/REC/IRBs identify what constitutes a strong and comprehensive design with the intention of providing oversight and awareness of what also meets ethical requirements designed to protect human subjects. Moreover, they argue, the IRBs obligation to determine that risks to subjects are reasonable in relation to anticipated benefits necessarily relies on a prior determination that the scientific design is adequate, for if it is not, there will be no benefits and any risk must be considered unreasonable.

Just by virtue of the power vested in their review, guidance, monitoring, and approval power it is clear why some researchers are wary of the IRB and the potential for *mission creep*. Mission creep defined simply is when the power vested in the IRB extends its bounds and goes beyond what is covered in the federal regulations. This type of extension, in its worst case scenario, is conflict driven and not directly related to the protection of human subjects. It therefore may be directly related to institutional legal concerns and protecting the university or in pushing the self-interest of board members or others in institutional leadership positions. It is clear that there is often a fine line between ensuring a subject's rights, institutional liability, and balancing what might be termed overarching yet limited federal guidelines – all of which rarely provide direct guidance for every situation. For those serving on the IRB, as well as the scientific investigators, should not this always be a test for weight and balances? Should federal law be adhered to the letter even when situations are not fully addressed by the general guidelines provided? Yet, who monitors the IRB and how is their work overseen?

Although no comprehensive answer is provided, steps along the corrective path include: Continuing education for board, investigators, and the community; and, regulation of IRBs by an independent body such as Association for the Accreditation of Human Research Protection Programs. This independent accreditation body can work to protect the rights of human subjects and foster clear guidelines for institutions, self-assessment, peer-review of all policy and procedures, and education.

When continuation for RECs and IRBs are required, this type of training needs to be updated regularly and reflective of changing technologies in research methods as well as public policy initiatives (Wolf, 2010). Most IRBs already have such policies and procedures in place but clear budgeting and allowances for such updating are vital. This includes budgeting for attendance at conferences where the sharing of the latest information and exchange of ideas with those familiar with IRB requirements and activities is essential. For those on the board, required training should take place at each meeting as well as through a systematized fashion on a cycle of renewal. In addition, there are certificates and additional training that can make board members aware of how to best apply the federal guidelines and any subsequent changes that occur. In turn, the scientific investigators should also receive updated training. To have investigators aware of the latest rules makes it a win–win for all in terms of review, modifications, and subsequent approvals. Lastly, the human subject, now generally referred to as the participant in the study, clearly can benefit from learning more about informed consent and what to or not to expect in any research study, not just the one he or she is invited to participate. An educated public helps to avoid confusion and can empower participants to ask the questions they have not thought to ask before. This requires that continuing education remain rich with updated policies and procedures essential for the success of any IRB.

Conclusions: Pulling It All Together

After the lessons learned in the past, some researchers continue to question the IRBs role in the protection of human subjects in clinical and behavioral science research. With the gains made through regulation, some professionals continue to wonder whether there are still gaps that occur in study conception and implementation (Slade et al., 2012).

This is complicated by federal regulations that are written in vague language in contrast to the clearer direction provided for protecting subjects' rights. A board composed primarily of professionals may be more scientifically curious than those with less training in the lay public. Such a board could be more

likely to vote in favor of proposals that may be considered of higher risk based on the potential for the gain of scientific knowledge. Lastly, as is true in any group process, a gathering of the minds can lead to *group think* where members of the group are influenced by a charismatic leader.

Gaining approval of the IRB has never been more persuasive when increasingly academic settings require faculty to conduct research and publish the results to advance the profession. Similar, to publish these research findings and contribute to the literature, academic, and rigorous journals will not publish the information gathered if it has not had prior IRB review and approval. This makes escape from IRB oversight almost impossible if one wishes to contribute scientific information to the literature. Pressure on IRBs to meet unrealistic deadlines and expectations that go beyond federal regulations can lead to 'mission creep,' a concept that only slows the process and discourages a professional collegial and supportive relationship between the researcher and the board (Saleem and Khalid, 2011).

Furthermore, are we really drawing a line in the sand when we try to separate out research from clinical practice? Can one really be distinct from the other? If so, then applying different rules makes sense. If not, trying to apply different rules and oversight of certain aspects of practice by calling it research may mislabel and subsequently misrepresent clinical inquiry (Anderson, 2010).

When the potential for mission creep is mixed with a blurred definition of what is clinical practice and what constitutes research, the lines between the two heighten the existence of the other. The more blurred the definitions, the greater the mission creep and the more complicated decisions become. Contemporary considerations and global connections make the situation even more complicated. When scientific investigations that are funded by U.S. enterprises such as pharmaceutical companies extend into developing countries, the need for control with limited red tape becomes even more evident (Saleem and Khalid, 2011). How does one truly stop the seductive benefits of secondary gain filled with researcher bias? Also, with changing technology problems continue to evolve related to ensuring consent to storing data in the cloud. As scientific and technological advances continue, the usual conflicts related to the varied professional interpretations of what constitutes ethical conduct designed to protect human subjects will make disagreements and differences in opinion unavoidable. Research regulations and the work of the IRB committee remains at the heart of risk-management review. These activities, however, will continue to fall short if this responsibility is not assumed deep in the core activities of the principal investigator and his or her research colleagues. In closing, this leads to one primary assumption, IRBs can help provide risk-management assessment, but the true ethical imperative to safeguard the participant must always rest with the individual investigator and his or her colleagues.

Few would disagree that the use of human subjects in research has led to significant scientific and medical advancements and equally few would question the need to protect research subjects from the potential for abuse (Bledsoe and Grizzle, 2013). In fact the rights and welfare of human subjects should always be considered first because it is the only way to ensure participant autonomy and safety.

See also: Common Rule: US Regulatory Protections for Human Research Participants; Ethical Dilemmas, Research, and Access to Treatment; Ethics for Biomedical Research Involving Humans: International Guidelines; Objectivity of Research: Ethical Aspects; Privacy of Individuals in Social Research: Confidentiality; Research Ethics, Cross-Cultural Dimensions of; Research Ethics: Research.

Bibliography

Anderson, J.A., 2010. Clinical research in context: reexamining the distinction between research and practice. Journal of Medicine and Philosophy 35, 46–63.

Bledsoe, M.J., Grizzle, W.E., 2013. Use of human specimens in research: the evolving United States regulatory, policy and scientific landscape. Diagnostic Histopathology 19 (9), 322–330.

Curran, W.J., 1970. Government regulation of the use of human subjects in medical research: the approaches of two federal agencies. In: Freund, P.A. (Ed.), Experimentation with Human Subjects. George Braziller, New York, pp. 402–454.

Fujii, L.A., 2012. Research ethics 101: dilemmas and responsibilities. Political Science and Politics 45 (4), 717–723.

Health Information Privacy, Updated 2014. U.S. Department of Health and Human Services. From: www.hhs.gov/ocr/privacy/ (retrieved 30.03.14.).

Hurst, S.A., 2008. Vulnerability in research and health care: describing the elephant in the room. Bioethics 22 (4), 191–202.

Levine, R.J., 1988. Ethics and Regulation of Clinical Research, second ed. Yale University Press, New Haven, CT.

Lipsett, M.B., Fletcher, J.C., Secundy, M., 1979. Research review at NIH. Hastings Center Report 9 (1), 18–21.

Medical Research Council of Canada, 1987. Guidelines on Research Involving Human Subjects. Medical Research Council of Canada, Ottawa, Canada.

NIH: Fact Sheet on 1946–1948 U.S. Public Health Service Sexually Transmitted Diseases (STD) Inoculation Study. HHS.Gov: US Department of Health and Human Services. From: http://www.hhs.gov/1946inoculationstudy/factsheet.html (retrieved 29.03.14.).

Saleem, T., Khalid, U., 2011. Institutional review boards- A mixed blessing. International Archives of Medicine 4 (19). Retrieved: http://www.intarchmed.com/content/4/1/19.

Slade, K.L., Carreau, N.A., Heald, P., 2012. Ethics of clinical trials in dermatology. Clinics in Dermatology 30, 226–230.

The Belmont Report, 18 April 1979. U.S. Department of Health and Human Services. From. http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html (retrieved 29.03.14.).

Wolf, L.E., 2010. The research ethics committee is not the enemy: oversight of community-based participatory research. Journal of Empirical Research Ethics: An International Journal 5 (4), 77–86.

World Medical Association, October 1975. Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects (Amended), 29th WMA General Assembly, Tokyo, Japan.

World Medical Association, October 2000. Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects (Amended), 52nd WMA General Assembly, Edinburgh, Scotland. http://www.wma.net/e/policy/17c.pdf.

Relevant Websites

http://www.ahrq.gov/ – Agency for Healthcare Research Quality U.S. Department of Health and Human Services.

From HHS.gov, retrieved from: www.hhs.gov/ohrp/about/index.html – Office of Human Research Protections About OHRP.

http://www.hhs.gov/ocr/privacy/hipaa/understanding/special/mhguidance.html – HIPAA Privacy Rule and Sharing Information Related to Mental Health.

http://www.hhs.gov/ocr/privacy/hipaa/understanding/special/mhguidance.html – HIPAA Mental Health Privacy Rules.

http://www1.va.gov/oro/ – VA Health Care: Office of Research Oversight.

Ethical Questions in Social and Behavioral Sciences, History of

Andrea Ploder, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A systematical discussion of Ethical Questions in the Social and Behavioral Sciences (SBS) started to unfold in the United States during the 1940s and 1950s and came to a provisional climax in the late 1960s and early 1970s. During this time, several ethical codes were released and collected volumes on the matter published. The debate brought about considerable shifts in perspective in the professional communities. Throughout the 1970s and 1980s, it slowly flashed over from the United States to Canada and Europe. Today, ethical codes and committees have become scientific 'normality' in most of the SBS.

Ethical Questions in the Social and Behavioral Sciences (SBS) are manifold. They touch on a broad variety of topics and are discussed with reference to a wide range of principles – from respect for human dignity and informed consent to the misuse of expertise and the upholding of scientific standards. Today, ethical codes and committees have become scientific 'normality' in most of the SBS. They appear in supervising agreements, dissertations, research funding forms, etc. From a historical perspective, the institutionalization of ethical concerns is a rather recent phenomenon. Ethical codes have developed with different rapidity, shape, and binding force among the disciplines and national communities.

This entry addresses the history of research ethics in the SBS from different angles. It starts with a sketch of early developments in other disciplines and professional contexts – the 'prehistory' of research ethics in the SBS and continues with early debates and institutionalization processes in psychology, sociology, and anthropology. The next section discusses some prominent cases in order to show the dynamics of the development from within the debates that shaped them. The last section addresses further factors that need to be taken into account in order to understand the history of research ethics in its complexity.

Prehistory

Within the SBS, ethical questions were no big topic before 1940. There was hardly any debate in leading journals, and academic associations did not release official codes of conduct. Some debates from the late nineteenth and early twentieth centuries, which would nowadays be framed as ethical, were framed as epistemological problems at that time.

Outside the SBS, the debate on research ethics has much a longer tradition. The most prominent arena was the field of medicine and medical research, where the first known codification dates back to the fifth century BC – the Hippocratic Oath. In 1830, John William Wilcock published his *Laws Relating to the Medical Profession* and in 1912, the *American Medical Association* released *The Principles of Medical Ethics* (for further detais see Lock, 1995, p. 515). The *Nuremberg Code*, released in 1947 subsequent to one of the Nuremberg trials, stated ethical principles for humane experimentation and medical research. It focused on informed consent, non-coercion, and the benefit of the research participants. Moreover, a lot of professional codes of ethics were released during the 1940s, most of them by nonacademic professional organizations in the United States (see Hobbs, 1948).

Early Debates and Institutionalization

Debates on ethical issues came up at different times and with different intensities throughout the various national SBS communities and disciplines. Starting prominently in the United States, they spread to other countries and international associations. **Table 1** compares professional codes of ethics in psychology, sociology, and anthropology in different national and international contexts by the year of their first release. It shows remarkable asynchronicity between the United States, Australia, and Europe, and also between the disciplines.

In economics, debates on ethical questions came up during the 1970s as well, one very early proponent being Kenneth J. Arrow (1973). He argued that – under certain circumstances – "ethical codes can contribute to economic efficiency" (p. 318). Still, Arrow and other economists concerned with ethical questions were and still are focused on *business* ethics – not on ethical standards for economic *research*. In recent years however, national and international economic associations started to formulate disclosure policies for their journals. In July 2012, the American Economic Association released a disclosure policy for all of its journals. The same goes for the Royal Economic Society (Great Britain) and its Economic Journal (2012), as well as the Canadian Economic Association with its journal CJE (2014). Neither of these associations have released a code of ethics for its members so far. According to its current office, the European Economic Association (EEA) set up an Ethics Committee in August 2013 which reported back to the EEA Council in August 2014. The Royal Economic Society – to give an example – requires a disclosure of "any financial or other interests or connections that might raise any question of bias in the work reported or the conclusions" for every submission to the *Economic Journal* (RES, 2012).

As becomes clear from the table above, early debates on ethical guidelines for SBS researchers were focussed on the psychological, sociological, and anthropological

Table 1 Codes of Ethics in the Social and Behavioral Sciences: years of first release

	Psychology	*Sociology*	*Anthropology*
International Associations	1976 Resolution Concerning Professional Ethics in Psychology (The International Union of Psychological Science)	2001 Code of Ethics (International Sociological Association)	——— (International Union of Anthropological and Ethnological Sciences)
European Associations	1995 Carta Ethica or Charter of Professional Ethics for Psychologists (European Federation of Psychologists' Associations)	——— (European Sociological Association (ESA)). (According to its current office, the ESA plans to release a Code of Ethics by 2014.)	——— (European Association of Social Anthropologists)
United States of America	1953 Ethical Standards of Psychologists: A summary of ethical principles (American Psychological Association) 1973 Ethical Principles and Guidelines in the Conduct of Research With Human Participants (American Psychological Association)	1971 Code of Ethics (American Sociological Association)	1971 Code of Ethics (American Anthropological Association (AAA)) 1973 Statements on Ethics. Principles of Professional Responsibility (AAA)
Canada	1986 Code of Ethics for Psychologists (Canadian Psychological Association (CPA).) (According to its current office, the CPA used the APA Code from the 1950s until the release of their own Code of Ethics. The first steps toward a CPA Code were taken around 1970.)	1994 Statement of Professional Ethics (Canadian Sociological Association)	1994 Statement of Professional Ethics (Canadian Sociology & Anthropology Association)
	1998 Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (Canadian Institute for Health Research, Natural Science and Engineering Research Council, Social Science and Humanities Research Council)		
Australia	1949 Code of Professional Conduct (since 1997: Code of Ethics) (Australian Psychological Society)	1998 Ethical Guidelines (The Australian Sociological Association (TASA).) (According to the current TASA office, the need for a code of ethics was first raised in 1971. A code was drafted in the 1970s but not adopted until 1998.)	1985 Code of Ethics (Australian Anthropological Society (AAS).) (In 1981, the AAS formed an Ethics Committee to draw up a Code of Ethics. Peter Sutton drafted a Code (see Sutton, 1986) which was revised and finally adopted in 1985. For further historical details, see AAS, 2012.)
Great Britain	2009 Code of Ethics and Conduct (British Psychological Society)	2002 Statement of Ethical Practice (British Sociological Association)	1999 Ethical Guidelines for Good Research Practice (Association of Social Anthropologists of the UK and Commonwealth)
Germany	1998 Ethische Richtlinien (Deutsche Gesellschaft für Psychologie)	1992 Ethik-Kodex (Deutsche Gesellschaft für Soziologie)	2009 'Frankfurter Erklärung' zur Ethik in der Ethnologie (Deutsche Gesellschaft für Völkerkunde)

communities in the United States. The Australian, Canadian, and European codes refer to them considerably in style and content. In the following, these debates will be examined in a bit more detail.

Psychology

Within the SBS, the field of psychology in the US was the first discipline to deal extensively with ethical questions. One reason for this early engagement might be the strong academic networks between medicine and psychology during the first half of the twentieth century. Moreover, many psychologists were involved in the treatment of World War II soldiers. They were asked to determine the draft eligibility of prospective soldiers or care for their mental health after their return from the field. Ethical dilemmas occurred quite frequently (Fisher, 2012, p. 3).

In 1940, a Committee *on Scientific and Professional Ethics* recommended the election of a Standing Committee and the release of a professional code to the *American Psychological Association* (APA). In retrospect, U.S. psychologist Nicholas Hobbs (1948) notes five motives for the interest of the community in a professional code of ethics: (1) "a growing sense of professional unity," (2) the need "to ensure public welfare, (3) to promote sound relationships with other professions, (4) to reduce intra-group misunderstandings, and (5) to promote professional standing of the group as a whole" (p. 80).

In 1947, the APA appointed the first *Standing Committee on Ethical Standards for Psychology* (chaired by Edward Tolman). It drafted a code of conduct and circulated it for broad discussion within the professional community (see Hobbs, 1948) before its final release in 1953. The procedure (proposed by Nicholas Hobbs (1948), a member of the committee) was highly participatory and drew from psychological insights to group dynamics and motives for ethical behavior. The goal was to "develop a technique which will be effective in modifying human behavior" (Hobbs, 1948, p. 82) and help to ethically imporove the professional conduct of psychologists. The code was based on 1000 reports of APA members describing and discussing ethically relevant situations and experiences in their professional life. From these reports, the Committee distilled a long and case-oriented code draft and played it back to the members for further revision. In 1959, it was revised and compressed, got more comprehensive, and then consisted of 18 principles and a preamble.

After a long period of minor revisions and adaptations to new fields of research and new methods, the APA Code was substantially revised in 1992. Since that time, the code distinguishes *aspirational principles* from 180 *specific ethical standards*. The principles are very general and meant to represent the foundational values of the profession, while the standards are meant to be distinct enough to serve as a basis for enforceable decisions by the APA and other Ethical Boards. In 1996, the APA Ethics Committee appointed an Ethics Code Task Force (chaired by Celia B. Fisher) in order to prepare another major revision of the code (Fisher, 2012, p. 6). The revised version was again subjected to a major participatory process and released in 2002. Apart from some new provisions concerning methodical innovations, a wider range of addressees, and some considerations related to the Internet, its major achievement was a higher level of concreteness, detail, and enforceability.

Sociology

After some scattered remarks and claims for regulation in earlier years, U.S. sociologists engaged in a broad discussion of ethical questions from the late 1950s on. The main impulse came from a number of controversial studies, extensively debated in journals and at conferences. These debates are an interesting document for a process which prepares the release of ethical codes in the late 1960s and early 1970s and some of them will be considered in more detail below.

In 1961, the American Sociological Association (ASA) appointed a Committee on Professional Ethics ('COPE,' chaired by Robert C. Angell) in order to discuss ethical issues systematically. Due to fundamental controversies, a new Committee was appointed in 1967 (chaired by Edgar A. Schuler), with the mission to formulate a statement on ethics for sociological research. After several revisions, the ASA Code of Ethics, a Council Preamble, and a Committee Preamble were approved by the members in 1971. The Committee Preamble formulated general statements on the identity of sociology as a profession and the nature of its Code of Ethics. It considered the fact that a lot of sociologists did not agree on the necessity of the Code of Ethics (e.g., Becker, 1964) and explicitly compared the development of this Code to the development of Common Law: based on cases from which principles are drawn slowly and carefully (Schuler, 1969: 317). The Code itself was rather short and vague, focusing (very generally) on the subjects' right to privacy, confidentiality, and protection from personal harm; on the researchers' duty of objectivity and integrity; and on the declaration of sources of financial support. Apart from a rather general claim in the Council's Preamble not to "manipulate persons to serve their quest for truth" (p. 318), the question of disguised observation – a highly controversial topic at that time – was avoided. The code got substantially revised in 1984 and 1997. The 1997 revision was preceded by 3 years of discussion within the COPE and aimed at a clarification of style, improvement of enforcement tools, and accordance with the new requirements of granting agencies. At the 2014 ASA meeting in San Francisco, the current Executive Officer suggested to revisit the Code of Ethics again. A committee has been formed.

Anthropology

The debate on research ethics in anthropology also started within the U.S. community. In contrast to psychology and sociology, it was not primarily focused on the relationship between researcher and research participants but on sponsoring relationships. Due to its long-term character, anthropological and ethnological fieldwork often requires the approval and sponsoring of state or private authorities. Sometimes these authorities have specific interests in the area and tie their support to a scientific pursuit of these interests. Intervention may reach from the adjustment of research questions to an obligation to pass on politically sensitive research findings to the authorities and/or hold them back from public and

scientific community. In an early reaction to this problem, the Council of the American Anthropological Association (AAA) released a one-page Resolution on the Freedom of Publication in 1948. It dealt with the question of academic freedom in situations of external research funding and claimed freedom of interpretation and publication for anthropological research results. During the 1960s, the debate expanded significantly. The discipline got involved in several scandals around controversial political projects including anthropological research. The most prominent case is related to *Project Camelot* and will be discussed in further detail below.

In 1965, the AAA *Executive Board* (EB) sent a report to the *Council of Fellows* (COF) concerning 'secret research' and the allegedly unethical involvement of U.S. anthropology with state politics. The Council asked the Board to investigate accusations in detail and the Board founded a *Committee on Research Problems and Ethics*. Chair Ralph L. Beals presented a report (the so-called *Beal Report*) at the AAA plenary session in 1966 which led to the release of the first AAA code of ethics by the COF – the *Statement on Problems of Anthropological Research and Ethics* in 1967. One year later, an *Interim Committee on Ethics* was installed. It recommended the election of a *Standing Committee on Ethics* (COE) and presented a draft for an ethical code. Both suggestions were controversially discussed: a lot of AAA members considered them to be unnecessary or itself unethical. Nevertheless, the first official COE was elected in 1970 and released a code of ethics called *Principles of Professional Responsibility*, as well as a text on *Role and Function of the Committee on Ethics* in 1971. In the same year, another case of dubious connections between research and secret governmental plans came up – the 'Thailand case.' It provoked severe tensions between the COE and the EB, led to the foundation of an ad hoc committee (chaired by Margaret Mead) and divided the Association into two fractions. The *Mead Report* disburdened researchers in Thailand, stating that no violation of the 1967 statement had taken place. Still, the discussion remained controversial and emotional until the end of the Vietnam War in 1973.

Afterward, the emphasis of the COE shifted toward a micro-level, focusing on questions of informed consent, anonymization, plagiarism, and so on (Hill, 1987). The last revisions of the code took place in 1998, 2009, and 2012. Compared to the Codes in psychology and sociology, the AAA code is still rather short (app. 6 pages) and principle-oriented, and the matters for its enforcement are very restricted.

Leading Cases

The history of ethical codes in the SBS covers only a small – yet important – part of the history of ethical questions. Ethical debates do not start with codifications. Typically, they start with concerns, complaints, and convictions of scientists and laymen about alleged ethical wrongdoings. Therefore, an important part of the history of ethical questions is a *history of cases* which leads to discussions about ethical problems within a profession. The relevance of 'leading cases' – especially for the debate within the United States – is underlined by the plan to compliment the ASA Code of Ethics with a casebook in the late 1990s. In the following, some of the most prominent cases will be discussed. The choice does obviously neither claim for completeness nor for representativeness but it helps to understand the structure and style of discussions in the early stage of the debate.

The 1950s and 1960s were a high time for ethical debates on cases in the SBS. Apart from those discussed here, one might mention Alfred C. Kinsey on *Sexual Behavior in the Human Male/the Human Female* (1948/1953), Festinger/Riecken/Schachter: *When Prophecy Fails* (1956), Laud Humphreys *Tearoom Trade* (1970), Philip Zimbardos *Stanford Prison Experiment* (1971), and many others. Many of them provoked controversies in sociological, psychological, and/or anthropological journals and some of them still serve as reference points for arguments and reevaluations of research ethical questions. Altogether, these cases document a process of growing sensitivity toward ethical questions in the SBS, resulting in the release of ethical codes in the late 1960s and 1970s. This process has at least three interesting characteristics. First, a shift in ethical convictions: the idea that ethical considerations are subjective in nature and therefore no legitimate criteria for scientific work, shifted toward a notion of professional ethics, binding on other than subjective grounds. Second, a step back from the conviction that scientific progress justifies all methods of investigation. Increasingly, ethical qualities reached the status of an important criterion for good research. And third, growing differentiation: the conviction formed that there is more than one kind of ethically relevant research relationship and more than one type of ethical responsibility. Different kinds of secrecy, grades of anonymization, kinds of informed consent, etc., started to be distinguished (see e.g. Roth, 1962).

The Springdale Case (1958)

The Springdale case became one of the leading cases for questions of anonymization. It refers to a community study by Arthur J. Vidich and Joseph Bensman (*Small Towns in Mass Societies*, 1958), reporting on a fictitious town in rich ethnographic detail. The material stems from fieldwork in Ithaca, where Vidich was engaged in another research project. He considered his notes and other material from the project to be valuable and decided to write a book on the structure and organization of the community. He tried to anonymize town and community members, but the true identity of 'Springdale' was easily revealed.

The book received broad attention, and many colleagues expressed concern about the ethical implications of the study. William Foote Whyte (1958) launched a debate on the case in *Human Organization*, resulting in a big discussion on professional ethics. Apart from failed anonymization, Vidich and Bensman were attacked for disguised investigations and for using material from the project without permission from the project leaders.

In their first response, Vidich and Bensman refused to concede any ethical problems in their research and claimed to have acted for the sake of science in every respect. They argued for the impossibility of further anonymization without changing the story (and its sociological punch line) completely. They underlined the importance of the study for the field of community studies and claimed that everything they said about the town and its members was common sense in

Springdale. Among others, Whyte (1958) pointed toward the difference between orally passed on local knowledge and written accounts with the pledge of scientific evidence. The second edition of the book (Vidich and Bensman, 1968) contains supplementary materials and a discussion of methodological and ethical questions. In his 2009 autobiography, Vidich also reports about the circumstances and faith of the study at length (Chapter 8).

Alcoholics Anonymous (1960)

John Lofland's and Robert Lejeune's study on *Alcoholics Anonymous* (AA) became a leading case on problems of secret observation. In 1959, the two sociologists held a research seminar at Columbia University and sent some of their graduate students to AA gatherings for participant observation. The observers concealed their identity as researchers and pretended to be newcomers to the group. Their dressing and behavior was planned and rehearsed in detail before they entered the field. After each meeting, the students took field notes and discussed their observations with the teachers. The use of a hidden recorder was considered, but dropped for reasons of feasibility. In 1960, Lofland and Lejeune sent a report of their main findings and some methodological reflections to the journal *Social Problems* (Lofland and Lejeune, 1960).

In the next issue, Fred Davis (1961) accused the authors of "premediated deception" (p. 364), lack of informed consent, and "tacit disrespect shown for AA's values" (p. 364). Moreover, he took the case as an example of a general moral dilemma of sociological research: the exploitation of an 'out-of-research-role' in order to elicit information for research purposes. Altogether, he considered the approach of Lofland and Lejeune to be a violation of the "collective conscience of the community, if not that of the profession," altogether "less than human (sic), and hence unworthy of a discipline which, whatever else it represents itself as, also calls itself by that name" (p. 365). Lofland's reply to this critique took a rather radical stand against Davis' allegations: "our obligation to science," Lofland wrote, "is the objective, full and unbiased execution and reporting of observation." Ethical considerations were therefore "simply not part of the legitimate criteria for judging research. Therefore, *as* scientists and professionals, we have no reason to be concerned with Mr. Davis' opinions of our personal moral standards." (p. 366) In this article, Lofland disputed the merits of professional research ethics altogether. Since the generalization of individual moral convictions can lead to absurd laws of conduct, he argued, moral dilemmas of researchers were completely up to the individual and "should not be solved by the intervention of the profession at large or other professionals speaking personally" (p. 367). (Some years later, Lofland changed his opinion on this matter.)

Lofland's reaction provoked a broad discussion within the U.S. sociological community. In 1962, Julius A. Roth tried to structure the debate in a letter to the editor of *Social Problems*. He distinguished different layers of consideration (size of the observed group, intimacy of pre-research relationship, level of organization and power of the observed group, systematic character of research, etc.) and concurring moral intuitions (outcome- versus principle-oriented) concerning secret observation. Interestingly, Roth did not call for a code of ethics. On the contrary he stated: "We are more likely to develop a useful and satisfying working ethic by analyzing the research process of the sociologist himself than by drawing up written codes of ethics which merely perpetuate current moral biases and restrict rather than aid further ethical development" (p. 284).

In 1965, Kai T. Erikson took up on the topic with another text in *Social Problems*. He called for a distinction between 'personal morality' and 'professional ethics' and interpreted the latter as a contract between professionals. This contract implies responsibilities in relationships of the researcher and (1) his subjects, (2) his colleagues, (3) his students (who are asked to do the work), and (4) his data. Disguised observation, Erikson concluded, implies problems in all of these dimensions. Moreover, Erikson suggested to distinguish between the concealment of one's identity as a researcher and one's research goals and its deliberate misrepresentation, the latter being unethical in any case. Erikson refers to other cases (like Springdale) as well and suggests to take first steps toward a "general posture on the question of disguised participant observation" (p. 259).

The Milgram Experiment (1961)

The Milgram Experiment is one of the most famous experiments in social psychology and still serves as a reference case for informed consent and the risk to inflict moral and psychological harm to research participants. It started in 1961, during the trial on Adolf Eichmann. With his experiments, U.S. psychologist Stanley Milgram wanted to find out what makes people obey to an authority, if they are asked to act against their personal moral convictions. He tried to test how much pain people were prepared to inflict to another person in mere situational obedience to an authority. He deceived the participants about the real research question and asked them to play the part of a teacher who tries to improve his/her student's results by inflicting electroshocks for each mistake. In fact, the alleged 'students' were part of the research team and no electroshocks got ever displayed. The results were unexpected: 65% of the participants were prepared to administer a 450 V shock, although the alleged students claimed to stop the experiment and displayed severe pain. (For details on the experiment, see the chapter on *Obedience*.)

Despite the huge interest in the results by colleagues and public (Milgram's findings are still considered as a landmark in the field of obedience studies), it was strongly contested for the ethical implications of the research setting. The most prominent and insistent critic was Diana Baumrind, a psychologist in Berkeley. After the 1963 publication of some of Milgram's findings in the *Journal of Abnormal and Social Psychology*, Baumrind wrote a comment in the *American Psychologist* (1964), addressing four major ethical problems of the experimental setting: (1) The participants did not know which kind of study they were engaging in; (2) The participants were subjected to extreme stress (heavy sweating, trembling, stuttering, and nervous laughter were reported). They might have suffered psychological harm and an uneasiness to trust authorities as a result of their participation in the experiment; (3) The participants were manipulated, discomforted, and embarrassed during the experiment. Thereby they were (4) denied their dignity and their right not to be

used as a mere purpose for the researchers' (secret) goals. According to Baumrind, all of these problems can – if not addressed properly – lead to a decline in public image of psychology and is therefore ethically problematic toward the profession and colleagues. She calls for full information about the possible risks and systematic care for eventual harms of participants in the aftermath.

In his response, Milgram (1964) claimed that stress as well as indignation of the research participants were neither intended nor accepted but came completely unexpected. To refrain from the risk of unexpected reactions of research participants, Milgram argues, would mean to refrain from scientific inquiry. He also presented the results from a questionnaire among the participants, indicating that 84% of the participants were retrospectively glad to have participated in the experiment and only 1.3% said that they were either sorry or very sorry to have participated. Up to one year after the experiment, participants were provided with results, appreciation for their effort, and the opportunity to talk about their reactions. Milgram's reply does not directly address Baumrind's concern about the concealment of the research question.

Project Camelot (1964–65)

Project Camelot was the code for a social science research project initiated by the U.S. Army in 1964. Conceived by a Special Operation Research Office in the American University of Washington D.C. and Research and Development Unit of the U.S. Army, the management was confided to Rex Hopper, a scholar with expertise on revolutionary movements in Latin America. The goal of the project was to study the cultural, economic, and political conditions of internal war by means of systems analysis and with an international team of social science researchers. According to one of its first documents (a letter of invitation to participate, mailed to a selected number of international researchers), Project Camelot was meant "to determine the feasibility of developing a general social systems model which would make it possible to predict and influence politically significant aspects of social change in the developing nations of the world" (1964, quoted from Horowitz, 1967, p. 4f.). Due to several miscommunications, breaches of secrecy, and diplomatic entanglements the project got stopped before it really started. Nevertheless, the Camelot affair caused big discussion within academia, the military, the public, and the state officials of several countries (especially Chile and the United States). Some of the involved anthropologists got suspected of espionage for the U.S. Government or unwittingly supporting political interventions against local populations. More generally the question came up, if and under which circumstances state funding of and/or political interests in research are likely to restrict scientific freedom to an unbearable extent. The case was discussed in U.S. Congress hearings and received broad attention by the media as well as professional academic organizations. For details on the course of events, see Horowitz, 1967.

After its termination, the project lost public interest but remained a hot topic among social scientists for many years. Freedom of scientific inquiry in a situation of governmental funding and military supervision, responsibility for research data, and political implications of research were extensively debated with regard to Project Camelot. Until today, it as a reference case for ethical dilemmas of state-funded research. It significantly influenced the release of the first AAA Ethical Code in 1965 and was also quoted in connection with the ASA Code of Ethics of 1969.

Other Relevant Factors

So far, this entry focused on the foundation of ethical codes and some leading cases in psychology, sociology, and anthropology. Of course, this is only part of the story. In order to give an idea of the complexity of the development as a whole, this last section addresses some additional factors that are important for the history of ethical questions in the SBS.

The history of research ethics in a discipline is closely linked to the *history of its research methods*. Experimental research designs are more likely to trigger problems of informed consent and the risk of psychological, physical, and moral harm, ethnographies often touch on problems of disguised observation and anonymization, and expensive methods requiring external funding (like extensive fieldwork or large-scale experiments) are likely to get in trouble with questions of tortious influence and conflict of interests. New methods go along with new ethical problems, and the import of methods from other disciplines comes together with the ethical problems related to it.

The history of research ethics is also strongly related to the *history of the disciplines* and their academic organization. Ethical questions became more and more urgent with the rising number of researchers, the increase of public research funding, and the growing public interest in the results of this research.

The history of research ethics is also linked to the *history of research funding*. The development of ethical codes of conduct historically coincides (at least in the United States) with the development of large-scale research funding programs. Apart from ethical problems linked to funded research (see the discussion above), ethical nonobjection also became one of the leading criteria for research funding institutions. Today, many research funding agencies require a certificate of harmlessness by a professional ethics committee. Accordingly, universities are required to install such committees and release a code that guides its work.

The history of research ethics is also connected to the *history of academic journals*, where most of the debates on ethical questions took and still take place. A platform for frequent professional exchange enables quick, intense, and transparent debates. For the development of collective ethical convictions, standards, and codes, this kind of discussion is crucial. An analysis of the most prominent cases shows that the debate often follows a three-step pattern: (1) breach of an unsaid ethical principle in a particular research case, (2) accusation and discussion of the case in professional journals and/or in public, and (3) enacting or modification of a professional code of ethics.

Finally, the history of research ethics is closely connected to the *history of scientific trust*. The fast-growing development of research ethics in the SBS took place during a time of public disappointment with the merits of science after World War II. The public as well as scholars started to look more closely at the

ethical quality of academic research. In the late 1960s, student revolts as well as the upcoming science studies supported this critical approach and further decomposed the holy grail of science. This situation was remarkably different from the early twentieth century, when most people (scholars as well as nonacademics) held a strong belief in the powers of scientific reasoning (for a similar argument, see Lock, 1995, p. 519).

Problems of research ethics are always located in a *broad field of norms and interests*, not all of them being ethical in nature (see Filstead, 1970, p. 266). The history of research ethics is always embedded in the history of political, professional, and institutional norms and interests.

Ethical convictions change over time. What was considered to be ethically neutral at one time can be considered to be ethically unbearable at another; minor ethical concerns can significantly grow in importance. The cases discussed above have shown that quite clearly. Therefore, the history of research ethics is not only a history of codification or changing moral sentiments of academics but also a history of moral change on a broader scale.

Conclusion

Ethical questions in the SBS started to unfold in the United States during the 1940s and 1950s and came to a provisional climax in the late 1960s and early 1970s. During this time, several ethical codes were released and books on the matter published (e.g., Horowitz, 1967; Filstead, 1970). The debate brought about considerable shifts in perspective in the professional communities. In the beginning, the debate was highly controversial and emotional, revealing both the significance of ethical questions for the community and the wide range of opinions on how they can be solved. In some of the early case discussions, scholars argued against ethical considerations with reference to the 'moral neutrality of research.' With the release of ethical codes and installation of ethical boards in the leading professional associations and universities, this view got more and more contested.

Throughout the 1970s and 1980s, the debates on research ethics slowly flashed over from the United States to Canada and Europe. Some scholarly associations used the codes of their U.S. colleagues before releasing their own. In Europe, the first ethical codes were released during the early 1990s. The 1990s was also a high time for research ethics in the United States, where a lot of ethical codes were considerably reviewed.

See also: Anthropological Research, Ethics of; Applied Social Research, History of; Ethical Guidelines and Codes of Conduct in Social and Behavioral Research; Privacy of Individuals in Social Research: Confidentiality; Research Funding: Ethical Aspects; Research Publication: Ethical Aspects ; Social Psychology: Research Methods.

Bibliography

American Economic Association, 2012. Disclosure Policy. http://www.aeaweb.org/aea_journals/AEA_Disclosure_Policy.pdf (23.03.14.).

American Psychological Association, 1949. Developing a code of ethics for psychologists. A first report of progress. American Psychologist 4 (1), 17.

Australian Anthropological Society, 2012. Code of Ethics. http://www.aas.asn.au/docs/AAS_Code_of_Ethics.pdf (23.03.14.).

Arrow, K.J., 1973. Social responsibility and economic efficiency. Public Policy 21, 303–318.

Baumrind, D., 1964. Some thoughts on ethics of research: after reading Milgram's "Behavioral Study of Obedience." American Psychologist 19 (6), 421–423.

Becker, H.S., 1964. Against the code of ethics (for the ASA). American Sociological Review 29 (3), 409–410.

Davis, F., 1961. Comment on 'initial interactions of newcomers in alcoholics anonymous'. Social Problems 8, 364–365.

Filstead, W.J. (Ed.), 1970. Qualitative Methodology. Firsthand Involvement with the Social World. Markham, Chicago.

Fisher, C.B., 2012. Decoding the Ethics Code. A Practical Guide for Psychologists. Sage, Thousand Oaks, CA.

Gingrich, P., Wien, F., 2007. Ethics of sociological practice: comments from the Canadian Sociological Association. In: Swiss Sociological Association: Ethik in der soziologischen Praxis. Seismo, Zürich, pp. 19–20.

Hill, J., 1987. The committee on ethics: past, present, and future. In: Cassell, J., Jacobs, S. (Eds.), Handbook on Ethical Issues in Anthropology. American Anthropological Association, Washington, DC., pp. 11–19.

Hobbs, N., 1948. The development of a code of ethical standards for psychology. American Psychologist 3 (3), 80–84.

Horowitz, I.L., 1967. The Rise and Fall of Project Camelot. Studies in the Relationship between Social Science and Practical Politics. M.I.T. Press, Cambridge, London.

Lock, S., 1995. Research ethics – a brief historical review to 1965. Journal of Internal Medicine 238, 513–520.

Lofland, J.F., Lejeune, R.A., 1960. Initial interactions of newcomers in alcoholics anonymous: a field experiment. Social Problems 8, 102–111.

Lowe, G.E., 1966. The camelot affair. Bulletin of Atomic Scientists Tome 22 (5), 44–48.

Milgram, S., 1963. Behavioral study of obedience. The Journal of Abnormal and Social Psychology 67, 371–378.

Milgram, S., 1964. Issues in the study of obedience: a reply to Baumrind. American Psychologist 19 (11), 848–852.

Royal Economic Society, 2012. The Economic Journal Ethics Policy. Publication Ethics and Conflicts of Interest. http://www.res.org.uk/view/ethicspolicyEconomic.html (23.03.14.).

Roth, Julius A., 1962. Comments on 'Secret observation'. Social Problems 9 (3), 283–284.

Sieber, J.E., Tolich, M.B., 2013. Planning Ethically Responsible Research. A Guide for Students and Internal Review Boards. Sage, Newbury Park.

Schuler, E.A., 1969. Toward a code of ethics for sociologists: a historical note. American Sociologist 3, 316–318.

Sutton, P., 1986. Draft code of ethics for professional anthropologists in Australia. Australian Anthropological Society Newsletter 29, 17–23.

Vidich, A.J., 2009. With a Critical Eye. An Intellectual and His Times. Newfound Press, Knoxville.

Vidich, A.J., Bensman, J., 1958. Small Towns in Mass Society. Class, Power, and Religion in a Rural Community. Princeton University Press, Princeton.

Vidich, A.J., Bensman, J., 1968. Small Towns in Mass Society. Class, Power, and Religion in a Rural Community. Princeton University Press, Princeton.

Whyte, W.F., 1958. Editorial: Freedom and Responsibility in Research. The Springdale Case. Human Organisation 17 (2), 1–2.

Ethical Turn in International Development

Eleanor Fisher, University of Reading, Reading, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The ethical turn in international development relates to a tendency to question the deleterious impact of international development action on local populations and environment. As a consequence, new courses of action are proposed in order to generate socially just and environmentally sustainable global change. This tendency is most prominent in relation to the development impacts of globalization of production and trade but also appears across a wide range of development sectors.

International development relates to change in living standards by developing countries through the elimination of poverty and the creation of institutional environments conducive to material improvement. It is a broad field encompassing topics as wide ranging as economics, politics, social transformation, overseas development assistance, health, infrastructure, education, governance, humanitarian relief, environmental change, human rights, and gender equality. Ethical concerns enter into all of these aspects. However an 'ethical turn' emerges most strongly out of the globalization of production and trade and is linked to demand by citizen-consumers to know where and how products and services originate, the conditions under which they have been produced and traded, and the sustainability of these processes.

The ethical turn does not refer to coherent shared politics, values, or practices, but points to a tendency within international development, which emphasizes consideration of peoples' quality of life in terms of healthy, secure, and sustainable living and work environments, and of the impact of development processes on people and the planet. Thus, the ethical turn has been consolidated through diverse international campaigns over labor rights, the mineral trade and conflict, food quality, animal welfare, health concerns, natural resource depletion, and environmental sustainability. Increasingly, citizen-consumers demand that ethical standards are upheld by companies and organizations producing and selling products and services from developing countries; they also demand that governments and international organizations demonstrate leadership in setting ethical development agendas and in ensuring ethical development practice.

Globalization and Market Integration

Stiglitz (2006) conceives globalization as encompassing the increased integration of people and countries due to a change in information communication technologies, reduction in transportation costs, and the removal of cross-border barriers to the flow of goods, services, and capital. One aspect of globalization is an increasingly interdependent global market, as described by Kingsbury (2012). Another aspect of globalization, linked also to neoliberal capitalism, is a transformation in established orthodoxies on the relationship between the state, market, and society. The changing power of supra- and substate authorities, and the influence of nongovernmental organizations and companies mean a multiplicity of social actors and power regimes come together beyond the nation-state (Grande and Pauly, 2005). These processes are illustrated by Fisher (2012) in a case study of Wales, the world's first Fair Trade Nation. In the Welsh example international development practice is driven by government's close relationship to civil society organizations and businesses promoting ethical trade.

In a global era driven by neoliberal capitalism, in which the centrality of the market is championed, the power of transnational corporations comes to the fore. This corporate power has potential for adverse social and environmental effects, as companies seek to source products and services at the lowest cost without considering broader externalities. Impoverished people may be paid a pittance for their labor and may work in conditions where negative impact on life well-being can be profound (Brown, 2000). Linked to this are questions over business standards in developing countries and who holds multinational corporations to account for their actions.

Associated with changing relationships between state, market, and society is the rise of organizations that are explicitly in favor of globalization and play contradictory roles in the ethical turn in international development – both by acting as standard bearers for human development and human rights and, conversely, through stimulating unintended social protest in response to their actions. These organizations include the World Trade Organization (WTO), the International Monetary Fund (IMF), the World Bank, the Organization for Economic Cooperation and Development (OECD), and the United Nations (UN). The WTO, IMF, OECD, and World Bank focus on aspects of economic development and the UN and its agencies on global security through the political, humanitarian, and social aspects of global development (Kingbury, 2012).

The Governance of Ethics through Voluntary Regulation

Voluntary regulation has emerged as an important mechanism for introducing ethical standards into aspects of international development. As Brunsson and Jacobsson (2000) describe, standards are rules that identify the course of action that those who adopt them should follow. The state has historically been the main body that promotes standards for regulation, often but not always backed by legislation. However, as global governance processes have become more complex, and in the context of an ideological emphasis on the neoliberal market as opposed to state control, voluntary regulation has become increasingly prominent across a wide range of development sectors. Standards associated with voluntary regulation perform

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10153-9

two key functions: product differentiation and risk management. They are part of a wider 'tripartite standards regime,' which includes standard setting, certification, and accreditation, and reflects evolving governance processes in which state-based systems have shifted to networks of regulation that incorporate state, corporate, and nongovernmental actors (Loconto and Busch, 2010). Against this background an important issue is how local development actors, interests, and values shape the effect of standards on processes of development transformation relevant to ethical issues (Viteri and Arce, 2013).

Both transnational social movement organizations and corporate actors promote the use of voluntary standards as mechanisms for ensuring that ethical issues are taken into account in international development in response to concerns by citizen-consumers about safety, quality, ethics, and production conditions (SC of the SKASC, 2012). This has led to a burgeoning of standards, certification, and product labels, as illustrated by information held by the International Trade Centre on over 100 voluntary standards (www.standardsmap.org).

Global development organizations, such as UN agencies, also play important roles in leadership for global standards. This leadership is reflected in legally binding agreements, such as some labor standards promoted by the International Labor Organization (e.g., on child labor) and the Minemata Convention on Mercury, a global treaty to protect human health and the environment from the adverse effects of mercury, whose adoption has been led by United Nations' Environment Programme through the Global Mercury Partnership and Intergovernmental Negotiation Committee. Standards also relate to nonmandatory guidance, such as the OECD's 2011 Due Diligence Guidance for Responsible Supply Chains of Minerals from Conflict-Affected and High-Risk Areas.

Corporate Social Responsibility

Corporate social responsibility has been adopted by business as an approach to international development that brings ethical issues to the fore. It focuses on consideration of the social and environmental impacts of business activity and how they can be mitigated. Blowfield (2005) argues that the incorporation of corporate social responsibility as an approach to international development raises crucial questions about who benefits and in what ways; in particular, is it used as a vehicle for certain interest groups to redefine the meaning of international development success? In examining this question and the degree to which business focuses on instrumental outcomes in development as opposed to intrinsic gains, Edward and Tallontire (2009) outline a model to examine businesses' role in development. To do this, the model presents two continuums: one focuses on the business–society relationship and the other on concepts of development (Tallontire, and Nelson 2013). Strength of this model is that it provides a means to observe and analyze the changing role of business in development, rather than positing ideal positions for or against. Examples of empirical analyses of the relationship between corporate social responsibility and development and relevant ethical issues are provided in an edited collection by Barrientos and Dolan (2006), examining how ethical sourcing has entered mainstream food retailing and whether it can address problems facing workers and producers in the global food system.

Social Movements and Public Campaigns

Nongovernmental organizations, trade unions, and the many different actors that are part of civil society have become increasingly important in international advocacy over ethical issues in developing countries, influencing policies, and affecting the contexts in which they are made. Micheletti et al. (2006) suggest that the global market in particular becomes a site for ethics and citizen action. They propose the idea of political consumption in which the market becomes a site for politics, expressing political values beyond traditional markers of value – price, taste, and quality. This can be seen in historically prominent international campaigns, such as the movement against apartheid in South Africa between 1959 and 1994, student protests in the 1970s over the multinational corporation Nestle's sale of formula milk to mothers in the developing world, and violent demonstrations with diverse motivation in reaction to meetings of the WTO and Group of Eight (G8) leaders of the world's most powerful countries. For Kriesberg (1997), these and similar campaigns reflect the rise of transnational social movement organizations, made possible by democratization, global integration, and the convergence and diffusion of values.

Social movements promoting ethical causes in international development draw complex linkages between a wide range of development issues and actions that can be taken by government, business, and society. Two contrasting examples illustrate these processes.

The first example comes from the African mining sector, which in the twenty-first century is a crucial driver for the 'mineralization' of economies and societies across the African continent (Bryceson et al., 2014). Emphasis on ethical issues in African mining has been stimulated by high-profile international campaigns on 'conflict' or 'blood' diamonds and minerals – where profits from the mineral trade fuel violent conflict – and by corporations worried about reputational damage as result of the negative social and environmental impacts of mining. These issues have brought traceability and accountability to the fore in the mineral trade. Focusing on artisanal gold production for fair trade markets, Fisher and Childs (2014) provide a case study on how voluntary regulation on ethical standards has entered into mineral production and trade in sub-Saharan Africa. They illustrate how a large number of ethical standards have been developed to certify the absence of conflict, good labor conditions, mitigation of environmental impacts, and development benefits for artisanal producers. These standards combine voluntary initiatives, e.g., the Fairtrade and Fairmined Gold Standard, with mandatory legislation, e.g., the United States Dodd–Frank Wall Street Reform and Protection Act, and nonmandatory guidance by international organizations, e.g., UN Guidelines on Due Diligence.

An important question is whether ethical standards, such as the Fairtrade and Fairmined Gold Standard, create new opportunities for market-led development transformation or whether they reinforce established patterns of power and

inequality with potentially detrimental and unintended consequences for artisanal producers or workers in large-scale mining operations. Fisher and Childs propose that Fairtrade has the potential to benefit artisanal gold producers, but suggest that there are significant impediments in practice to this occurring in ways that challenge established power and market relations to the benefit of the most marginalized in mining communities. See also Lompo and Trani (2013), who examine corporate social responsibility and development in the oil sector, presenting a case from Nigeria.

A second example of linkages between ethical issues, development, and actions taken by social movement organizations is from the Latin American health sector. As countries develop, governments face the problem of what policy response to make to noncommunicable diseases. Among these diseases are the consequences of poisoning from reliance on use of dangerous pesticides in food crop production and subsequent deleterious effects on human health. Sherwood et al. (2013) examine pesticide use, noncommunicable diseases, and social action on the ethics of development in Ecuador. They outline three significant policy developments in modern food supply: making food accessible to the whole population, the arrival of energy-compact food products, and industrial food production – all have been associated with agricultural productivity reliant on high pesticide use. One unintended consequence relates to rising levels of obesity due to the high sugar and fat content of the food products. Challenging the use of pesticides and arguing for the need for healthier food consumption are social movement organizations, which propose alternatives to the dominant food regime. Sherwood et al. (2013) argue that the sheer violence of the pesticide technology has captured the public imagination, generating challenges to seemingly all-powerful and impervious food ideologies and business interests. An outcome of the ethical countermovement was the introduction of legislative measures to eliminate highly toxic pesticides.

Examples from the African mining sector and Latin American health sector contrast markedly; nonetheless both highlight how international development issues generate ethical concerns and how social actors from civil society, government, and business respond to these concerns in ways that seek to challenge established practices and power relations in the favor of local people.

Conclusion

The ethical turn in international development relates to a tendency to question the deleterious impact of international development action on local populations and environment. As a consequence, new courses of action are proposed in order to generate socially just and environmentally sustainable global change. This tendency is most prominent in relation to the globalization of production and trade, but also appears across a wide range of development sectors. In taking action on ethical issues, the changing dynamics between civil society, state, and market become apparent. The power of transnational corporations comes to the fore, but so does the ability of transnational social movement organizations to challenge the status quo in terms of power hierarchies and accepted values, often through using the market as a site to express the politics of consumption but also by lobbying governments and global institutions.

See also: Culture and Economic Development; Development and the State; Ecocultural Theories of Development; Environment and Development; Global Development: Markets and Wellbeing; Globalization and Health; Globalization and Transnational Corporations; Globalization and World Culture; Globalization, Anthropology of; Globalization: Welfare and Social Work; Migration and Development; Neoliberalism; Planning Issues and Sustainable Development; Sustainable Agriculture; Sustainable Development: An Economic Perspective.

Bibliography

Barrientos, S., Dolan, C. (Eds.), 2006. Ethical Sourcing in the Global Food System. Earthscan, London & Sterling, VA.

Blowfield, M., 2005. Corporate social responsibility: reinventing the meaning of development? International Affairs 81 (3), 515–524.

Brown, D.K., 2000. International Trade and Core Labour Standards: A Survey of the Recent Literature, OECD Labour Market and Social Policy Occasional Papers 43. OECD Publishing. http://ideas.repec.org/p/tuf/tuftec/0005.html (accessed 22.06.13.).

Brunsson, N., Jacobsson, B., 2000. The contemporary expansion of standardization. In: Brunsson, N., Jacobsson, B., Associates (Eds.), A World of Standards. Oxford University Press, Oxford.

Bryceson, D.F., Fisher, E., Jønsson, J.B., Mwaipopo, R., 2014. Mining and Social Transformation in Africa: Mineralizing and Democratizing Trends in Artisanal Production. Routledge, London, New York.

Edward, P., Tallontire, A., 2009. Business and development – towards re-politicisation. Journal for International Development 21 (6), 819–833.

Fisher, E., 2012. The "Fair trade nation": market-oriented development in devolved European regions. Human Organization 71 (3), 255–267.

Fisher, E., Childs, J., 2014. An ethical turn in African mining: voluntary regulation through fair trade. In: Bryceson, D., Fisher, E., Jønsson, J.B., Mwaipopo, R. (Eds.), Mining and Social Transformation in Africa: Mineralizing and Democratising Trends. Routledge, London, New York.

Grande, E., Pauly, L. (Eds.), 2005. Complex Sovereignty: Reconstituting Political Authority in the Twenty-first Century. Toronto University Press, Toronto.

Kingsbury, D., 2012[2008]. Globalization and Development, in International Development: Issues and Challenges, second ed. Palgrave Macmillan, Basingstoke. pp. 112–137.

Kriesberg, L., 1997. Social movements and global transformation. In: Smith, J., Chatfield, C., Pagnucco, R. (Eds.), Transnational Social Movements and Global Politics: Solidarity beyond the State. Syracuse University Press, Syracuse, New York.

Loconto, A., Busch, L., 2010. Standards, techno-economic networks, and playing fields: performing the global market economy. Review of International Political Economy 17 (3), 507–536.

Lompo, K., Trani, J.-F., 2013. Does corporate social responsibility contribute to human development in developing countries? Evidence from Nigeria. Journal of Human Development and Capabilities 14 (2), 241–265.

Micheletti, M., Follesdal, A., Stolle, D. (Eds.), 2006. Politics, Products and Markets: Exploring Political Consumerism Past & Present. Transaction Publishers, New Brunswick, New Jersey.

Sherwood, S., Arce, A., Berti, P., Borja, R., Oyarzun, P., Bekkering, E., 2013. Tackling the new materialities: modern food and counter-movements in ecuador. Food Policy 41, 1-10.

Steering Committee of the State-of-Knowledge Assessment of Standards and Certification. 2012. Toward Sustainability: The Roles and Limitations of Certification. RESOLVE, Inc., Washington, DC. Available at: http://www.resolv.org/site-assessment/files/2012/06/Report-Only.pdf (accessed 31.07.12.).

Stiglitz, J., 2006. Making Globalization Work: The Next Steps to Global Justice. Allen Lane, London.

Tallontire, A.M., Nelson, V.J., 2013. Fair trade narratives and political dynamics, Social Enterprise Journal 9, 28–52.

Viteri, L., Arce, A., 2013. The negotiation of quality standards: a social interactionist approach to fruit and vegetable distribution in Argentina. International Journal of Sociology of Agriculture and Food 20 (1).

Ethics and Psychiatry

Allen R Dyer and Eindra K Khin, The George Washington University, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Psychiatry is a profession defined and guided by well-established norms dating back to the Hippocratic Oath and carefully delineated by the standards of modern biomedical ethics and professionalism. Key principles, such as beneficence (confidentiality), autonomy (informed consent), nonmaleficence, and justice provide bases for clinical judgment in specific contexts. Commercialism in health care introduces potential conflicts for the professional, whose primary allegiances are to the patient and those served. Several new challenges are faced by psychiatrists and other professionals in the twenty-first century. These include telemedicine, electronic health records, and social networks/social media.

Ethical codes and traditions are important in defining the norms of professional conduct and even what it means to be a profession. Insofar as ethics deals with intellectual attempts to understand right and wrong (reason) as well as the affective impact of struggling with moral issues (emotion), psychiatric thinking offers an important perspective on ethics. This article deals with both ethics in psychiatry and psychiatry's contributions to ethics. Although changes in society and especially the recent advent of managed care have led to challenges to professional identity, the Hippocratic tradition continues to provide a perspective of enduring values that calls into question many of the assumptions of modern culture and transcends the pressures of political expediency.

Ethics concerns what is right and what is wrong. It is often used synonymously with morality, but sometimes a distinction is made between ethics, the systematic approach to understanding right and wrong, and morality, the forces which govern right conduct. Ethics is central to human experience, and plays a vital role in the profession of psychiatry. In a time of cultural change, the appeal to ethics can help to resolve conflicts of both civic life and self-understanding. Often ethical principles are used in an attempt to find a common ground or a 'higher ground' for resolving disputes or to find a basis for acting in a dilemma.

Ethical codes and traditions are important in defining the norms of professional conduct and even what it means to be a profession. Insofar as ethics deals with intellectual attempts to understand right and wrong (reason) as well as the affective impact of struggling with moral issues (emotion), psychiatric thinking offers an important perspective on ethics. This article deals with both ethics in psychiatry and psychiatry's contributions to ethics.

Types of Ethical Theories

There is in ethics, a constant tension between rule-based (deontological) approaches and approaches that appeal to principle (teleological). Deontological ethical theories maintain that there are rules of action with moral validity independent of the consequences that require adherence, such as keeping a promise, maintaining justice, or following a commandment of God or the state. Teleological ethical theories are based on the principle of utility (utilitarianism) or beneficence, which proposes that the ultimate moral standard is the general happiness of all people concerned, or the greatest net balance of good over evil ('the greatest good for the greatest number'). Situation ethics represents another form of ethical reasoning, called the consequentialist approach, in which an action is judged to be morally right or wrong by assessing the consequences of the action. Much of clinical medicine is consequentialist in that the goodness of actions is judged by their outcomes.

Ethics of Virtue and Character

Both teleological and deontological approaches to ethics have one serious shortcoming. They depend on the activities of the rational mind of a motivated moral agent to apprehend right and wrong. Moral philosophers and psychiatrists alike appreciate that not everyone is motivated to be moral; that even good people sometimes do bad things; and that people are capable of rationalizing self-interest (Hauerwas, 1977, 1981; Pincoffs, 1971, 1980).

A delineation of obligations, however essential, is not the whole picture of ethics. Virtue, character, and integrity are every bit as important in understanding right conduct. For the professions especially, integrity in concrete situations is the hallmark of virtuous life.

A Brief History of Medical Ethics

The ethical traditions of the medical profession originate in the fourth century BC with the Oath of Hippocrates. This oath, sworn by healer–craftsmen, has come to symbolize the ethical ideals of the medical profession. In the nineteenth century, first the British Medical Association, and subsequently the American Medical Association (AMA) and professional organizations in other countries began to adopt formal codes of ethics modeled on the principles of the Hippocratic Oath.

By the mid-1960s, the values underlying the doctor–patient relationship began to change. The most notable factor was the emergence of new technologies, which began to change the way we thought about medicine and even what it meant to be human: genetic engineering, organ transplantation, advanced life support, safe abortion and birth control, and medical and

surgical treatments based on clinical research. No longer was medical care primarily palliative. No longer could it be assumed that the patient would necessarily want what the doctor might offer (Dyer, 1997a,b).

Also in this era, several social movements began to change the way society thought about assumed relationships, with an increased emphasis on the right of the autonomous person to self-determination. Informed consent and autonomy came to be the keys to medical ethics and medical decision-making. Medical ethics in this era became bioethics, as it came to be realized that the decisions in medicine were decisions in which society had a stake.

By the 1970s, economics increasingly became a part of thinking about ethics. Allocation of resources became a concern, and the question was often asked, "Could health costs be contained without rationing care?" The principle of justice eventually superseded beneficence and autonomy in ethical discourse, with an emphasis on fair distribution of medical resources.

The Ethics of Advertising

Most physicians are not inclined to seek patients though advertising. Health care organizations, however, may make their appeal through advertising, especially in a market-driven health care economy. The 1957 version of the AMA *Principles* prohibited 'solicitation' of patients, by which was meant 'obtaining patients by deception,' or making false claims (AMA, 1957). After the Supreme Court decision Federal Trade Commission (FTC) versus AMA (Greenhouse, 1982), professional organizations such as the AMA can no longer formally constrain physicians from advertising.

Advertising has two basic goals: (1) dissemination of information and (2) product differentiation. For physicians, providing information about services offered and fees charged is consistent with patient benefit, but attempts to differentiate the product are suspect. Licensing, accrediting, and credentialing organizations assume the role of maintaining standards. Beyond the activities of individual practitioners, the larger ethical question for advertising is truthfulness. Notable is the advertising of unhealthy products such as tobacco products and alcohol. Much advertising is antithetical to professional goals of patient benefit (Dyer, 1995, 1997a,b).

The Place of Ethics in the Definition of a Profession

What differentiates a profession from a trade? What makes medicine a profession? The emphasis on the commerce of technology in medicine in recent years may obscure the centrality of an ethical attitude toward the patient, which has traditionally been considered to be the defining feature of a professional life.

A profession may be defined by (1) its knowledge, technology, and expertise, or (2) its ethics or values. The habits of modern thought might lead us to believe that this is an either–or choice. Clearly for psychiatry, as for most of medicine, allopathic technologies such as drugs and procedures have become important and perhaps even central in some people's minds. The recent primary care thrust has reopened consideration of a more holistic approach to healing, including talking. Ethics is more fundamental to professional definition than technology. Technology is useful as long as it serves ethical ends but not as an end in itself. Knowledge, technology, and expertise are not commodities to be bartered in the marketplace, but skills which may be used to ethical ends under the purview of professional values.

The Hippocratic Oath and the Hippocratic Tradition

The Hippocratic Oath is a remarkable document, not so much because it answers the ethical questions posed by modern medicine, but because it frames those questions. It articulates principles, notably the principle of patient benefit, which is often helpful to physicians in sorting out where their allegiances lie. Taken together with its corollary, the principle of nonmaleficence (*Primum non nocere*, first do no harm), the principle of beneficence puts the patient at the center of ethical decision-making. The Hippocratic Oath provides an ethical perspective that calls into questions many of the assumptions of modern culture and transcends the pressures of political expediency.

The Contemporary Relevance of the Hippocratic Oath

A reading of the Hippocratic Oath reveals many phrases, which responsible physicians might not want to follow literally. Medicine was delineated as a higher calling, a covenant with the gods. The Oath provided a relationship with a like-minded family of individuals who shared certain values and self-understanding. It defined the physician's life in a beneficial relationship to those served. Its paternalism was questioned when society emphasized individualistic autonomy, but the Oath's emphasis on justice foreshadowed a larger concern for community interests.

The Hippocratic tradition has at various times and places served as an ethical reference external to immediate expediency, by which physicians may gauge their behavior and to which they can appeal (or by which they may be judged by their colleagues). For example, the World Psychiatric Association censured the practices of psychiatrists in the former Soviet Union and other Eastern European countries for the detention of political dissidents. Such mislabeling with psychiatric diagnoses is called the political abuse of psychiatry, for the profession is used for ends other than patient benefit (Bloch, 1991; Bloch and Reddaway, 1977). Such extreme abuses make a case, which is not always appreciated, for professions being autonomous organizations self-regulated by a code of ethics rather than serving the interests of society, which may be politically conflicted. The Hippocratic Oath is an ever-present reminder that acting in the interest of someone other than the patient creates a conflict, a conflict which is often experienced internally by the physician.

Honor in Medicine

The Hippocratic Oath speaks of honor, holiness, and purity. The good name and respect of both the practitioner and the profession are to be earned through consistent application of

the ethical principles, and the code provides guidance. Good conduct becomes 'characteristic' of the professional, and a code of conduct organizes one's thinking about good conduct both symbolically and in very practical matters.

The Fiduciary Principle

The doctor–patient relationship is spoken of sacredly as a fiduciary relationship, a relationship based on trust. That trust must be earned, at the least by the hard work required to acquire the skills necessary to apply the art. But it must also be earned by the consistent application of attentive response to the patient's needs, what we call 'responsibility.' Unlike law or business, where the trustee may act for the client, in medicine the physician must act with the patient and with the patient's consent, implying a partnership. That action may be paternalistic, but if that concern verges on control, the modern patient may well lose trust and confidence.

Hippocratic Oath for Psychiatrists

Maurice Levine (1972) articulated a Hippocratic Oath for psychiatrists with some unique insights about the values of this profession. His code articulates the importance of self-knowledge or at least self-awareness and constant self-scrutiny. He stresses the need to recognize in ourselves the feelings that patients can stir up, particularly seductive patients, and the importance of getting consultation (or supervision) not just in training but whenever it might be necessary. This is a useful reminder at times when many treatments are briefer and more biological, because many of the complaints of unethical conduct received by the American Psychiatric Association (APA) come from patients whose doctors did not sufficiently deal with their own feelings and acted out in treatment situations. Self-reflection is an important ethical tradition, not only in psychiatric and psychoanalytic circles, but throughout Western culture dating back to the Socratic admonition to 'Know thyself.'

Contemporary Codes of Ethics (AMA and APA)

When the AMA was founded in 1847, it adopted a code based largely on the Hippocratic principles of beneficence and honor. This code (*The Principles of Medical Ethics*) has been revised every few decades (most recently in 1957 and 1980). When the APA adopted its first code in 1973, it decided to use the AMA Principles *With Annotations Especially Applicable to Psychiatry*. The principles should be read teleologically, not deontologically as a list of rules. The requirement of honorable behavior of the physician goes to the physician's character and virtue and cannot be reduced to a list of rules. The physician must understand the principles in terms of a higher calling and act in accord with the dictates of conscience.

The APA (1993) may impose four possible sanctions for misconduct by a psychiatrist: admonishment (informal warning), reprimand (formal censure), suspension (not to exceed 5 years), and expulsion. In this sense, this code is minimalist but essential, serving as a tool for professional discipline. A psychiatrist should know what is specified in the code because he or she might be subject to sanctions. The modern physician should also understand the principles that underlie the code in order to be able to act in situations that have not been explicitly spelled out. The following discussions illustrate the constant tension between rule and principle and the importance of the physician striving for virtuous action.

Sexual Misconduct and Boundary Violations

Sexual contact with a patient is unethical. This is one of the least ambiguous sections of the ethical code. It is a tradition that goes back to the Oath of Hippocrates, which speaks of such conduct as 'mischief.' It is important for psychiatrists because the intimacy of the treatment activates strong feelings and fantasies in the doctor–patient relationship, the discussion of which may be essential to healing. In psychiatry especially, the nature of such feelings the patient may have for the doctor (transferences, i.e., derived from significant relationships in the past and activated in the treatment) and feelings the doctor may have for the patient (countertransferences) receive close scrutiny. They are no less important for other physicians, other therapists, or other professionals. It is sometimes argued that the reason sexual contact is proscribed is because of the power differential, an argument which applies equally to employers and employees, supervisors and supervisees, and teachers and students. Most basically, the importance of trust in the therapeutic relationship requires forbearance. As a rule, sexual contact with a patient is proscribed.

Those more legalistically inclined might wonder if erotic feelings might be acted on if the professional relationship were terminated. Transferences endure over time, so the interests of the patient–client could never be served by crossing this boundary. The most recent version of the APA *Annotations Applicable to Psychiatry* spells this out: "Sexual activity with a current or former patient is unethical." Other examples of boundary violations, which rest on similar considerations, are business relations with a patient, using the professional relationship for other contacts, or profiting from information gained from the therapeutic relationship.

The Boundary between Ethics and Law

Ethics and the law share a similar concern for right and wrong. But the ethical requirements of professionals do not coincide identically with the requirements of all citizens. The Oath of Hippocrates held physicians to a standard that was not required of all citizens. The AMA *Principles* and APA *Annotations* recognize this tension and expect physicians to 'respect the law' and 'seek changes in those requirements, which are contrary to the best interests of the patient.' Civil disobedience against unjust laws is a respected ethical tradition. A physician is a citizen whose professional role requires adherence to principles such as patient benefit above all.

Confidentiality

Confidentiality, understood as secrecy, in the doctor–patient relationship is one of the most fundamental principles of professional ethics. Confidentiality has its roots in the

confidence or trust that is placed in the physician and on which the treatment depends. It has become increasingly difficult to safeguard confidentiality. Insurance companies, managed care organizations, other providers, courts, and sometimes families claim a right to know what is going on in patients' treatments. The possibility of misuse of information from insurance data banks is a large issue. Confidentiality is one of the most fragile tenets in the contemporary era.

At times, principles may conflict. For example, the *Tarasoff* decision obligates a psychotherapist who has knowledge that a patient may intend harm to a third party to notify that third party of the potential risk. Mandatory reporting of suspected child abuse is another example in which the interest of the state in protecting children legally supersedes the physician's duty to maintain secrecy. In a larger context, the erosion of confidentiality threatens to undermine the possibility of a kind of therapy that requires developing openness to another person confident that what is disclosed is only for the patient's benefit.

The 'False Memory' Controversy

Much conflict has developed around the issue of therapists directing therapy to recover memories of childhood sexual abuse, which may not have occurred. In particular, there have been instances reported in which therapists used suggestive methods such as hypnosis in order to recover memories of past sexual abuse that were believed to be repressed. Alleged perpetrators of such abuse have been prosecuted and even jailed. Therapists' zeal about these issues may lead to unethical behavior, even with good intentions, with destructive results for their patients and families (Chodoff, 1999).

The Double Agent Problem

Double agents work both sides. There are situations in which the physician works for someone other than the patient, and there may be a conflict between the patient and the physician's employer resulting in a divided allegiance and compromised confidentiality. The classic case of this is the physician working for the military, trying to get soldiers back into combat. This is ethically analogous to treating the psychotic prisoner incompetent to face execution. The student health psychiatrist doing an evaluation for the dean or the psychiatrist doing a pre-arraignment examination have loyalties that go beyond the patient's best interest and which require at the least disclosure of the purpose of the examination and how the information will be used. Increasingly physicians working for corporations, managed care organizations, and government hospitals have interests, which radically transform the nature of the relationship between doctor and patient.

Fees, Billing, and Reimbursement

The financial arrangements for medical service strike at the heart of the ethical issues in how medicine is valued. Inevitably there must be some sort of exchange for the services rendered, and inescapably there must be some sort of feelings about the monetary transaction. In the dyadic doctor–patient relationship that is a very direct transaction, though it may be buffered by insurance companies, billing and collection processes. Patient benefit and fee-for-service have traditionally gone hand in hand, and the fee (alongside honor and duty, of course) has helped focus the physician's attention on whose benefit should receive attention. One of the general criticisms often leveled against traditional (Hippocratic) medical ethics is that it does not deal with the physician's larger obligations to society and particularly, to the distribution of medical services.

One financial arrangement in particular that sometimes questioned is the practice of charging for missed sessions. The APA *Annotations* has specifically addressed this and has said, "It is ethical for the psychiatrist to make a charge for a missed appointment when this falls within the terms of the specific contractual agreement with the patient" (Sect. 2, paragraph 6). The idea of patients having responsibility for the treatment and ultimately for their health and well-being is not something usually considered in medical ethics, but it goes beyond the idea of medicine as technology that the doctor delivers and stresses the patient's investment and commitment to the healing process.

Another financial arrangement that is addressed throughout the AMA *Opinions* and APA *Annotations* is fee splitting. Physicians should receive remuneration for the work they do, and they should obtain patients by referral based on the quality of their work and the reputation they earn in the eyes of satisfied patients and other physicians. Any incentive such as splitting the fee with a referring physician is considered unethical. It offers a temptation to place monetary gain above the best interest of the patient.

Informed Consent

An issue that has been receiving increasing attention recently is the importance of preserving patient autonomy and control in the context of medical decision-making and participation in research. Recent debate has occurred around the issue of whether patients with psychiatric illnesses are competent to give consent in such situations. The danger in such debate is that psychiatric patients can be further stigmatized with policies that are overprotective and confuse psychopathology with a generalized lack of competence. Efforts to educate patients regarding the nature of medical decisions and research projects should protect their rights by helping them to make decisions for themselves (Appelbaum, 1998).

Ethics and Health Care Financing

The current revolution in health care financing challenges many traditional assumptions of medical ethics. The most fundamental change is the departure from a dyadic doctor–patient relationship. Doctors in the new environment are being transformed into 'providers,' patients into 'consumers,' and any number of 'third parties' claim an interest in what goes on in that relationship.

In the economic model of marketplace transactions, much more is at stake than how the payment is to be allotted. The metaphor of health care as an industry radically transforms the healing process to physiological interventions that take place in a delimited period of time. Physicians become technicians, and

patients become recipients of technology, with both likely to be strangers to one another. Medicine is transformed from a human service into a commodity.

In the interest of efficiency or economy, these transformations might seem warranted, but the practices we are currently witnessing in the name of managed care must themselves be subjected to ethical scrutiny. Certainly much of what is going on is not ethically justifiable, nor does it serve the ends it was purported to serve. The transformation of medicine from a profession to a trade by the FTC was justified as a cost-saving measure on the grounds that increasing consumer choice (through advertising) would lower costs. The value placed on autonomy set the stage for a transformation in which consumers have very little choice about the health care they receive. Health care has become an investment opportunity where vast amounts of money have been siphoned away from service delivery into the pockets of executives and shareholders of megacorporations. The market solution has been a nonsolution to this point in that it has only succeeded in lowering costs without adequately addressing quality of service or the distribution and allocation of services.

The ethical physician is inevitably in tension with a system designed to limit or deny the care given to his or her patients. While recognizing that resources are finite, the ethical physician must struggle to do the best possible for individual patients without compromising integrity.

Novel Ethical Challenges in Twenty-First Century

Over the last couple of decades, we have witnessed a remarkable wave of information and communication technology advances that has fundamentally transformed the landscape of the field of medicine. In this new age of technology, it has become imperative that health care providers and administrators are equipped with the most up-to-date knowledge in this field to be able to deliver best possible care to the patients (Tang and Helmeste, 2000). As exciting as it is to have these novel digital, it is also vital to keep in mind that there is no substitute for prudent clinical judgment. While technology itself is neither ethical nor unethical, it is the intent and manner in which the technology is implemented that impacts the question of appropriate utilization. In many cases, lack of clear professional guidelines and legal regulations leads to concerning ethical issues (Fleming et al., 2009).

Telepsychiatry

Until recently, psychiatry has been a particular branch of medicine that did not rely heavily on high technology. However, the nature of psychiatry, which mainly utilizes audiovisual information lends itself particularly well to the practice of telemedicine. Second, in the last century, mental health services have undergone a major reform shifting away from institutionalized care to care in the community. This has produced a substantial strain on the system as the task of recruiting and retaining sufficient qualified providers has proved to be a challenge (Norman, 2006). Telepsychiatry has been viewed by many as the answer to this barrier to access and at present is important for countries with significant rural or dispersed populations such as Canada, Australia, and Russia (Tang and Helmeste, 2000). The cost effectiveness of telepsychiatry has been less clear. Some research shows that it can be less expensive for patients, reducing travel time, travel costs, and time off work. Other considerations, such as take-up rate and usage of the service, will factor into cost effectiveness as well. It is also important to keep in mind that the cost analysis equation of telepsychiatry is constantly changing with the varying cost of equipment and reliability of the technology (Norman, 2006). The ethical conundrum of telepsychiatry realizes the uniquely positive impact that it can have on patients, providers, and clinical outcomes, as well as the potential for harm and abuse that may ensue (Fleming et al., 2009).

Licensure and Credentialing Requirements: In the United States, there is a requirement that physicians hold a full, unrestricted license in all states in which they practice. An exception exists in most states, but not all, that allows physicians licensed in that state to consult with licensed physicians from other states and sometimes other countries. Parallel to the uncertainty surrounding licensure, there is also a question of whether a telepsychiatry consultant is required to be credentialed both at the base institution and at the remote site (Hyler and Gangure, 2004).

The question of licensure and credentialing requirements is often cited as a major barrier to the development of telemedicine networks and services, including telepsychiatry services. For many professionals, acquiring and maintaining multiple licenses is a significant professional and financial burden in complying with multiple state, region, or country regulations. Regardless, given that individual areas in the world differ in their definitions of practice, it is important that practitioners adhere to their area's regulations to ensure legally and ethically sound practice (Miller et al., 2005).

Informed Consent, Privacy, and Confidentiality: The principle of respect for patient autonomy dictates that physicians do their due diligence to facilitate the informed consent process by enhancing patients' understanding of their care and encouraging them to be active participants. As part of the informed consent process for telepsychiatry, patients should be advised on how their health information will be accessed, stored, transmitted, transferred, or protected. Given the novelty of telepsychiatry modality, a preliminary discussion with the patient regarding expectations concerning the sessions and treatment needs to take place before the consultation to establish a clear, explicit understanding of risks and benefits as well as shared responsibilities (Hyler and Gangure, 2004).

For centuries, Hippocratic Oath has tasked physicians to safeguard patients' privacy and confidentiality. This task has never seemed more daunting in this millennium as physicians' control over patient information has become diminished with the implementation of such practices as telepsychiatry and electronic medical/health record systems. Experts have warned that electronic transmission of patient information has increased the opportunity for infringement of the privacy and confidentiality of personal psychiatric information. In the case of telepsychiatry, this amplified threat for security breach is twofold: technological risks that originate from software of computer systems and risks directly posed by unauthorized

personnel (Hyler and Gangure, 2004). To address the first concern, risk management strategy would involve utilizing identity authentication (unique user IDs and passwords), secure networks with remote access controls, and encryption programs as well as adhering to the existing, regionally relevant legal and professional guidelines (Miller et al., 2005). The other unique privacy concern in telepsychiatry arises from possibility of outsiders to witness the session without being seen on the screen. This may range from presence of staff at the site of consultation or nonclinical technicians at the remote site of a technical service provider. Care must be taken to employ necessary precaution measures to protect against this type of potential security breaches as well (Hyler and Gangure, 2004; Norman, 2006).

Impact on Doctor–Patient Relationship: As health care technology permeates clinical practice more and more, we may be at risk of becoming technology centric rather than patient centric. Telehealth should function as an adjunct to, not a replacement of, the traditional physical encounter as many believe that physical proximity is still of paramount importance in securing the reciprocal connections integral to a healing and trusting relationship (Fleming et al., 2009). Given that telepsychiatry interactions are purely based on limited audiovisual cues of varying quality, physicians as clinicians no longer have the benefit of all their senses in uncovering and appreciating the richness and complexity of the patient (van Wynsberghe and Gastmans, 2009). Fundamental clinical cues such as eye contact, posture, facial expressions, body positioning, voice quality and tone, and hesitations, all of which can be relevant and pivotal in assessing the patient's emotional state, may be compromised or missed altogether due to limited audiovisual quality of telepsychiatry practice. Conversely, this setting can also limit the therapist to perform supportive gestures during videoconferencing sessions, which may be perceived as lack of empathy by the patient (Norman, 2006). This potential for depersonalization may present as a major barrier in relating to patients as whole persons and in turn has the potential to impede physicians' ability to provide the best care possible (van Wynsberghe and Gastmans, 2009).

Electronic Health Records

National electronic health record (EHR) initiatives are in progress in many countries around the world including Australia, Canada, Denmark, England, Estonia, France, Scotland, and the United States (Spriggs et al., 2012; Sutrop, 2011). Anticipated benefits from these nationwide initiatives include improvement in quality and continuity of care by improving efficiency, decreasing medical errors, reducing cost, reducing health disparities, improving the health of the public, and advancing education and research. Along with these benefits, there are also risks associated with the development, expansion, and maintenance of medical and health information in electronic form (Kopala and Mitchell, 2011). Ethical concerns surrounding informed consent, privacy, and confidentiality in EHRs are similar to those with telepsychiatry as discussed above. Some EHRs allow for the creation of sensitive notes that can only be viewed by certain groups of clinicians while others notify users when they are about to access sensitive information (Nielsen et al., 2013).

Proper Documentation and Data Accuracy: Clinical information pertaining to psychiatric care tends to contain more sensitive data than most other domains of medicine. Psychiatrists are encouraged to avoid unnecessary details and substitute general language as appropriate for specific details (Nielsen et al., 2013).

Despite the high hopes for EHRs to reduce medical errors, the very nature of EHRs raises concerns for the quality, accuracy, and reliability of data in the electronic records. Given the longitudinal and multiple-user nature of these records, they can contain extensive amounts of data that may not be current or relevant in a given clinical setting. Improper use of operational functions such as 'cut and paste' may result in an inaccurate representation of the patient's current condition and treatment. Worse, some EHRs employ autopopulate features that may introduce out-of-date, irrelevant, incomplete, or inaccurate information into the records. Concerns for incomplete data due to loss or destruction of data during data transfer have been noted as well (Kopala and Mitchell, 2011).

Communication and Collaboration: The doctor–patient relationship of present time differs greatly from past relationship when a physician was the sole provider of health care. Nowadays, a patient's health care is provided by multiple physicians of multiple specialties, usually in integrated delivery systems or at least coordinated in some way. Physicians today also share health care responsibilities with other nonphysician providers as well as support staff (Rothstein, 2010). One of the appeals of EHRs is the enhanced ease of information sharing across clinicians and service sectors. However, collaborative care requires clinicians to be mindful about effective communication that reflects a balance of restraint and release of information (Nielsen et al., 2013). Experts have called for different ways of disclosing health information only on a 'need-to-know' basis, limiting based on the role of the health care provider, date of the prior encounters, type of health care provided, diagnosis, or procedure (Rothstein, 2010). Authorization for release of information must specify what information will be disclosed, the purpose of the disclosure, the recipient of the information, and the date on which the consent expires. Some argue that reasonable effort must be made to limit the release of protected health information to the minimum necessary to fulfill the purpose of the request for information (Nielsen et al., 2013). The opponents of this stance warn that this approach may lead to making health care decisions based on incomplete information (Rothstein, 2010).

Social Networking/Media

With the emergence of social networking sites and media sharing sites over the last decade, such sites have become widely employed all over the world as a framework for self-expression as well as a means of communication. By enabling individuals and communities to gather, communicate, share, and collaborate, the benefits associated with these sites are numerous, ranging from rapid dissemination of useful information, increased interaction with friends and family, and development of new relationships, to facilitation of social and political change. However, they have also made it

increasingly challenging for physicians to keep their personal and professional lives separated. This is particularly relevant for the new generation of clinicians entering the health care field with ingrained communication habits, which are fundamentally different than their predecessors' (von Muhlen and Ohno-Machado, 2012). This paradigm shift has in turn created a unique set of ethical concerns.

Potential for Breaching Patient's Confidentiality: It is not unusual for users of social networking sites to employ this forum as a medium to vent about trials and tribulations of their everyday lives, including work-related issues. Physicians are no exception to this trend. Given this widespread practice, it is important to understand that one may still inadvertently reveal the identity of a patient without using any protected health information. A seemingly harmless anecdotal post about an interesting patient without personal identifying information may still be recognizable to a colleague with contextual knowledge about the case. Therefore, it is best not to post any information about patients on the social networking sites, no matter how deidentified it is (Ginory et al., 2012).

Some may argue that exceptions already exist where the public use of deidentified or disguised information about a patient is allowed. Such exceptions are generally granted for educational purposes such as presentation in scientific papers. Blatantly self-serving nature of posting information about patients as a status update is what distinguishes such unjustifiable uses from legitimate exceptions (Gabbard et al., 2011).

Potential for Publicizing Unprofessional Conduct: In addition to the aforementioned danger of breaching patients' privacy, other unprofessional conduct may involve postings including inappropriate language, negative comments about jobs or coworkers, derogatory comments about patients, depictions of substance use, and sexually suggestive/explicit material (Baer and Schwartz, 2011; Bosslet, 2011).

While physicians engaging in conduct unbecoming of their profession is not a new concern, physicians' unprofessional conduct in the digital era is accompanied by a novel set of implications. To begin with, the anonymity, albeit incorrectly perceived, and detachment from real-time social cues inherent in online forums can serve as the seductress for the users and lead them down the path of disinhibition to the level that may be atypical in real-life situations. Physicians are not immune to this temptation and may be more prone to lapses of professionalism in the setting of social networking (Baer and Schwartz, 2011).

One crucial point to note about the social networking sites, whose very nature relies on dissemination of information through 'second-degree' friends, is that it facilitates the spread of information faster than any other media. Therefore, online information available through social networking sites is not only easily accessible and widely available but also very durable. Once any unprofessional conduct has been displayed in cyberspace, it is very difficult, if not impossible, to erase the digital footprint. Even if one deletes the original unfavorable post, it will continue to exist in cyberspace unless everyone with whom it has been shared delete it as well (Bosslet, 2011). It is also imperative to note that questionable online content, be they words, pictures, or videos, has the potential not only to reach an unintended audience but also to be taken out of context (Baer and Schwartz, 2011).

Since public display of physician's lack of common sense and sensibility can be potentially detrimental to personal integrity, relationships with patients and colleagues, and employment prospects, physicians should be mindful that traditional expectations regarding the conduct of the medical profession still apply in this nontraditional context. In the face of this newly emerging pitfall, physicians should be vigilant about privacy settings on social networking sites and educate themselves about the potential effects of technological and information-sharing advances on professionalism (Baer and Schwartz, 2011).

Potential for Boundary Transgressions: 'Friend requests' on social networking sites bring yet another challenge to maintaining a proper boundary in doctor–patient relationship. Physicians may worry about the potential negative impact on the therapeutic alliance, if they were to decline a friend request from a patient. On the other hand, if it were to be accepted, each party would be allowed full access to the other's profile, which would most likely be more revealing than what is generally disclosed in a treatment relationship, especially from the physician's side. Depending on the type of information revealed, this may result in a boundary crossing or worse, a boundary violation (Gabbard et al., 2011; Ginory et al., 2012). In some parallel way, a friend request may be conceptualized as a gift offer from a patient to a therapist, where conscious and unconscious motivations and expectations associated with the gift should be further explored (Ginory et al., 2012).

Before becoming 'friends' with the patient on a social networking site, the physician must ask himself or herself how an online connection would be directly related to patient care. Given the concern for forming dual relationships, which in most cases do not have a therapeutic value and in some cases can be straight up harmful to the patient, the physician should always strive to establish and maintain appropriate boundaries (Ginory et al., 2012).

Safeguard Measures

Technological advances in the health care field have been adopted at different paces in different areas of the world, based on existing infrastructure and resources. Physicians should keep abreast with their regional, governmental, and institutional guidelines to ensure that proper administrative, physical, and technical safeguards are employed in their practice when using these digital tools for patient care. When indicated in cases of electronic medical records, appropriate deidentification methods should be utilized to assure patients' privacy and confidentiality. In the event that a security breach occurs involving personal identifiable information, specific steps must be taken to address the incident and mitigate potential harmful effects. Technical expertise should be sought to guarantee the use of these digital tools in a productive and safe manner. As with any traditional clinical encounter, helping patients understand the procedures and policies governing the care, custody, control, and disclosure of their health information using these new digital interfaces, as well as risks and benefits associated, is a crucial part of the informed consent process

(Klein, 2011). When in doubt, physicians would do well to remember one of the cornerstones of professionalism, respect for patient autonomy, from which springs obligations to ensure informed consent, privacy, and confidentiality (Fleming et al., 2009).

Conclusion

Looking back to the longstanding traditions of medical ethics, we are now in a position to better appreciate the place of the Hippocratic Oath in contemporary medicine. Though it is sometimes criticized for being anachronistic, its true value is apparent. The Hippocratic tradition in medicine provides a perspective of enduring values by which the shortcomings of modern economic expediency may be judged.

See also: Culture as a Determinant of Mental Health; Ethical Dilemmas, Research, and Access to Treatment; Ethics for Biomedical Research Involving Humans: International Guidelines; Suicide.

Bibliography

American Medical Association, 1957. Opinions and Reports of the Judicial Council. AMA, Chicago.

American Medical Association, 1997. Code of Medical Ethics (150th Anniversary Edition); Current Opinions with Annotations of the Council on Ethical and Judicial Affairs. AMA, Chicago.

American Psychiatric Association, 1993. The Principles of Medical Ethics with Annotations Especially Applicable to Psychiatry. American Psychiatric Press, Washington DC.

Appelbaum, P.S., 1998. Missing the boat: competence and consent in psychiatric research. American Journal of Psychiatry 155, 1486–1488.

Baer, W., Schwartz, A.C., 2011. Teaching professionalism in the digital age on the psychiatric consultation-liaison service. Psychosomatics 52, 303–309.

Bloch, S., 1991. The political misuse of psychiatry in the Soviet Union. In: Bloch, S., Chodoff, P. (Eds.), Psychiatric Ethics, second ed. Oxford University Press, Oxford, UK, pp. 493–516. New York and Melbourne (Australia).

Bloch, S., 1999. Misuse and abuse of psychiatry. an overview. In: Blosch, S., Chodoff, P., Green, S.A. (Eds.), Psychiatric Ethics, third ed. Oxford University Press, Oxford, UK, pp. 493–516.

Bloch, S., Reddaway, P., 1977. Psychiatric Terror: How Soviet Psychiatry Is Used to Suppress Dissent. Basic Books, New York.

Bosslet, G.T., 2011. Commentary: the good, the bad, and the ugly of social media. Academic Emergency Medicine 18, 1221–1222.

Chodoff, P., 1999. Misuse and abuse of psychiatry: an overview. In: Bloch, S., Chodoff, P., Green, S.A. (Eds.), Psychiatric Ethics, third ed. Oxford University Press, Oxford, UK, pp. 49–66.

Dyer, A.R., 1988. Ethics and Psychiatry: Toward Professional Definition. American Psychiatric Press, Washington, DC.

Dyer, A.R., 1995. Advertising. In: Reich, W. (Ed.), Encyclopedia of Bioethics, second ed. Macmillan, New York, pp. 76–79.

Dyer, A.R., 1997a. Ethics, advertising, and assisted reproduction: the goals and methods of advertising. Women's Health Issues 7, 143–148.

Dyer, A.R., 1997b. Ethics of human genetic intervention: a postmodern perspective. Experimental Neurology 144, 168–172.

Fleming, D.A., Edison, K.E., Pak, H., 2009. Telehealth ethics. Telemedicine and e-Health 15, 797–803.

Gabbard, G.O., Kassaw, K.A., Perez-Garcia, G., 2011. Professional boundaries in the era of the Internet. Academic Psychiatry 35, 168–174.

Ginory, A., Sabatier, L.M., Eth, S., 2012. Addressing therapeutic boundaries in social networking. Psychiatry 75, 40–48.

Greenhouse, J., 24 March, 1982. Justices uphold right of doctors to solicit trade. The New York Times 10.

Hauerwas, S., 1977. Truthfulness and Tragedy: Further Investigations into Christian Ethics. University of Notre Dame Press, Notre Dame, IN.

Hauerwas, S., 1981. Community and Character. University of Notre Dame Press, Notre Dame, IN.

Hyler, S.E., Gangure, D.P., 2004. Legal and ethical challenges in telepsychiatry. Journal of Psychiatric Practice 10, 272–276.

Klein, C.A., 2011. Cloudy confidentiality: clinical and legal implications of cloud computing in health care. Journal of American Academy of Psychiatry and the Law 39, 571–578.

Kopala, B., Mitchell, M.E., 2011. Use of digital health records raises ethics concerns. JONA'S Healthcare Law, Ethics, and Regulation 13, 84–89.

Levine, M., 1972. Psychiatry & Ethics. G Braziller, New York.

Miller, T.W., Burton, D.C., Hill, K., Luftman, G., Veltkemp, L.J., Swope, M., 2005. Telepsychiatry: critical dimensions for forensic services. Journal of American Academy of Psychiatry and the Law 33, 539–546.

von Muhlen, M., Ohno-Machado, L., 2012. Reviewing social media use by clinicians. Journal of the American Medical Informatics Association 19, 777–781.

Nielsen, B.A., Baum, R.A., Soares, N.S., 2013. Navigating ethical issues with electronic health records in developmental-behavioral pediatric practice. Journal of Developmental & Behavioral Pediatrics 34, 45–51.

Norman, S., 2006. The use of telemedicine in psychiatry. Journal of Psychiatric and Mental Health Nursing 13, 771–777.

Pincoffs, E., 1971. Quandary ethics. Mind 75, 552–571.

Pincoffs, E., 1980. Virtue, the quality of life and punishment. The Monist 63, 23–27.

Rothstein, M.A., 2010. The Hippocratic bargain and health information technology. Journal of Law, Medicine & Ethics 38, 7–13.

Spriggs, M., Arnold, M.V., Pearce, C.M., Fry, C., 2012. Ethical questions must be considered for electronic health records. Journal of Medical Ethics 38, 535–539.

Sutrop, M., 2011. Changing ethical frameworks: from individual rights to the common good? Cambridge Quarterly of Healthcare Ethics 20, 533–545.

Tang, S., Helmeste, D., 2000. Digital psychiatry. Psychiatry and Clinical Neurosciences 54, 1–10.

van Wynsberghe, A., Gastmans, C., 2009. Telepsychiatry and the meaning of in-person contact: a preliminary ethical appraisal. Medicine, Health Care and Philosophy 12, 469–476.

Ethics and Values

Peter Railton, University of Michigan, Ann Arbor, MI, USA
Alex Silk, University of Birmingham, Birmingham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the twentieth century, moral philosophy and the theory of value came to be explicitly subdivided into two elements: metatheory, which concerns questions about meaning, justification, metaphysics, and psychology; and normative theory, which concerns more substantive questions about what is good, right, or valuable. In earlier centuries, these two kinds of question were seldom distinguished. More recently, dissatisfaction with any sharp separation of these areas of inquiry has grown: how are we to judge the adequacy of any account of meaning without asking whether it does some justice to widespread substantive convictions? The distinction nonetheless will make the presentation below more manageable. The focus of this article will almost exclusively be on the so-called 'analytic,' Anglo-American philosophy and its claimed antecedents, although fundamental work on ethics and the theory of value has of course been done in other traditions.

Ethics

Metaethics

There is no agreed-upon distinction between the terms 'ethics' and 'morality,' and they are used in philosophy more or less interchangeably (although Williams, 1985 associates 'morality' especially with duty-based systems). Both terms show, in their origin, a relation to social customs – 'ethos' and 'mores' – and both have a descriptive as well as a normative sense. In the former sense, we can ask "What was the Greek warrior's ethic?" or "How has morality evolved in Western Europe over the past two hundred years?" In the latter sense, we can ask "What is the ethical thing to do when one accidentally learns of a deep confidence?" or "Is it moral for us to carry out animal experimentation?" (*see* Conventions and Norms: Philosophical Aspects).

The descriptive aspect seems fairly straightforward, at least once one has noticed that there are wider and narrower uses (see Gibbard, 1990). In the wide sense, a system of morality is virtually any code of conduct that concerns how to live, regardless of its particular content or social import; a warrior ethic might favor unstinting cruelty toward one's enemies; a personal morality might glorify selfishness or esthetic accomplishment over everything else, even if society finds this abhorrent. In the narrow sense, morality is taken to involve norms of conduct with a distinctive purport, sometimes characterized in terms of a moral point of view, as contrasted with, say, an esthetic or prudential point of view (see Baier, 1958). Morality, narrowly understood, is said to involve some or all of: impartiality, concern for the well-being of others, treating like cases alike, willingness to cooperate, a notion of personal agency and responsibility, associated feelings of impersonal anger, guilt, or deservingness. In the wide sense, we can ask of virtually anyone or any society, "What is his (or their) morality?"; in the narrow sense, we can ask, "Does this individual (or this society) have a morality, and, if so, when did it emerge?" Much of the interest of philosophers and social scientists in individual moral development or in the historical development of morality concerns morality in the narrow sense. There is an active debate over whether various ancient civilizations recognized morality, as distinct from enlightened prudence or religious requirement (see Annas, 1993). Some social psychologists have suggested that there are three main types of morality: an ethic of autonomy, dominant in the West, based primarily on concepts like those of individual freedom, rights, fairness, justice; an ethic of community, prominent in much of East Asia, based primarily on concepts like those of duty, loyalty, hierarchy, interdependence; and an ethic of divinity, well-elaborated in India, based primarily on concepts like those of sanctity, the natural or sacred order, sin, pollution (Shweder et al., 1997; see also Fiske, 1991; Rozin et al., 1999).

An important moment came when Prichard (1912) convinced virtually everyone that showing certain principles of conduct to be advantageous would not suffice to show they had distinctively moral standing. Recent interest in the possible evolutionary origins of morality also typically concerns morality in the narrow sense: 'debunking' views tend to emphasize that the gene- and kin-oriented mechanisms of natural selection are not impartial; 'vindicating' views tend to emphasize that these mechanisms could have equipped us with psychological resources – empathy, cooperativeness, concern for fairness – that can be mobilized culturally on behalf of genuinely moral practices (e.g., see Thompson, 1995; Katz, 2000). Some debunking views argue that the fact that evolutionary forces have tremendously influenced the contents of our moral judgments and evaluative attitudes gives us reason to deny that there are moral or normative truths that hold independently of the evaluative attitudes of the agents to whom they apply (see Joyce, 2006; Kitcher, 2006; Street, 2006). Research into the neural bases for our moral intuitions has led some to discount the trustworthiness of certain of these intuitions (see Greene, 2003; Singer, 2005). 'Experimental philosophers' have used scenario-based survey instruments to raise similar questions and to challenge philosophers' grasp of 'common-sense morality' (Knobe and Nichols, 2008). The normative import of these arguments, however, is strongly debated (see Copp, 2008; Berker, 2009; Enoch, 2010).

Philosophical analysis of moral language in the normative sense also has focused on the narrow use of the term, although this has not eliminated controversy. Central to the disputes has

been the questions of whether the primary semantic function of moral sentences is to describe or represent the world, the question of 'cognitivism.' A paradigmatic factual sentence like 'Alice is smoking' conventionally represents the world as being a certain way, as being such that Alice has the property of smoking. It distinguishes among ways the world might be. Cognitivists maintain that moral sentences like 'Smoking is wrong' conventionally perform the same function. Noncognitivists disagree. They maintain that moral sentences conventionally distinguish among ways moral systems might be. Expressivists contend that the nature of moral beliefs is, in the first instance, to be explained in terms of their psychological properties – say, as states of norm-acceptance or planning states – rather than in terms of their intentional properties as states that represent the world as being a certain way. They maintain that the primary semantic function of moral sentences is to express desire-like attitudes or commands intended to influence behavior (see Ayer, 1936; Stevenson, 1937, 1944; Hare, 1952; Blackburn, 1984, 1998; Gibbard, 1990, 2003; Schroeder, 2008), although some expressivists allow that moral sentences conventionally express ordinary descriptive beliefs as well (see Schroeder, 2009 for an overview). Although many have been skeptical about whether expressivists can give an adequate compositional semantics for moral language (the 'Frege–Geach problem'), the influence of the expressivist strategy can be seen in areas outside metaethics; expressivist semantics have also been given for epistemic modals and adjectives like 'might' and 'probably,' for attitude verbs like 'knows,' for predicates of personal taste like 'tasty,' among others.

'Noncognitivism' is in some ways an unfortunate label, since moral discourse clearly involves a great deal of cognition and conceptual content, according to even those who claim its primary function is to express attitudes. For an attitude to count as distinctively moral, it must represent itself as responsive to certain features rather than others (those characterized by a moral point of view), and as regulative for certain emotions (e.g., guilt or resentment). Recently, so-called 'minimalist' theories of truth have gained currency, according to which the predicates 'true' or 'false' can appropriately be applied to sentences of a discourse so long as they enter seamlessly into familiar logical relations, patterns of inference, etc. (Horwich, 1990). Because moral sentences unquestionably do this – 'Stealing is wrong', unlike an imperative like 'Don't you dare steal!' can be the premise of a normal logical inference and embed under any expression that takes a clausal argument – there has been a tendency to concede on all sides of the debate that moral sentences can be said to be 'truth-apt' in the minimal sense. Contemporary noncognitivists also often accept that there are moral beliefs, properties, and facts. This might or might not suffice to earn the label 'realism' about moral properties (see Blackburn, 1993). If it does, some have worried that expressivists – or at least those who wish to endorse certain characteristic realist claims, given an expressivist interpretation – will not be able to distinguish themselves from traditional realists (see Gibbard, 2003; Dreier, 2004; Scanlon, 2003, 2009).

A burning issue is: even if utterances of moral sentences express beliefs, do moral judgments have an 'internal' or conceptual connection with motivation, such that making a moral judgment necessarily involves an element of endorsement not found in ordinary factual judgments? One can perfectly appropriately assert or believe that such-and-such rock is granite, for example, without expressing or having any favorable or unfavorable attitude toward that state of affairs. Similarly, if one cannot sincerely assert or believe that such-and-such situation is immoral while having no attitude of disapproval, then there remains a fundamental difference between a moral and a 'merely descriptive' discourse.

Defenders of such a difference must insist that this expressive function belongs either to the pragmatics of moral discourse or to its semantics. Expressivists and subjectivists pursue the latter path. Expressivists claim that utterances of moral sentences conventionally express motivational attitudes, and that in order to count as having a certain moral belief one must have certain emotional capacities and dispositions, general and particular. A simple subjectivism holds that utterances of moral sentences report rather than express that one (or one's society or group) approves or disapproves. Moral beliefs, on this view, are beliefs about one's attitudes or norms (or those of one's society or group). It may seem that expressivism and subjectivism result in a kind of moral relativism, according to which an individual's moral obligations essentially depend on what norms or principles she (or her society or group) accepts. However, expressivism and subjectivism, in the senses described above, are semantic positions about what moral sentences mean. As such, they are compatible with a variety of substantive theses about the universality of moral truths and the nature of moral properties. This is to their advantage; although relativist views have enjoyed a certain popularity, they typically account poorly for moral disagreement and often verge on incoherence, since they may include a nonrelativistic moral recommendation of tolerance in response to moral difference (but see Harman and Thomson, 1996 for a sophisticated debate on relativism).

Subjectivists about moral properties claim that moral properties are constituted in part in terms of psychological properties of agents. Contemporary subjectivists often take a sort of objectivist approach, thinking of moral properties neither as independent features of the world, nor as constituted wholly by psychological properties, but as a pairing of features of situations and characteristic (or ideal-typical) human responses to those features – rather in the way that dangerousness might be seen as a pairing of risk and fear – and, therefore, as response dependent (see Railton, 1986a, 1998a). Some theorists have been inspired in part by an analogy between moral properties and secondary qualities, such as color (McDowell, 1985). Color judgments show a high degree of objectivity and intersubjective agreement, so much so that they are often used in scientific observation and social coordination (e.g., 'red' for 'stop'). Yet, what it is to be a certain color may nonetheless essentially depend on kinds of subjective response, and those without a certain human-typical perceptual apparatus may be incapable of seeing color. Moral judgments might be similar. Those without a certain human-typical motivational apparatus might be incapable of moral experience, but since virtually all of us have (or could develop) these motivations, moral judgments can also be highly objective and intersubjective. In the moral case, the response – like fear – to the relevant

feature of the world would have a motivational aspect, perhaps arising from moral training in the virtues rather than sheer human physiology. For someone equipped with moral character, then, moral experience and judgment are necessarily accompanied by a motivating response. Does this sort of subjectivism lead to relativism – different patterns of moral training (say) yield different, yet locally apt, moral judgments? Not necessarily, since we can take our moral discourse, like our actual color discourse, to be anchored to our actual responses (by a semantic device known as 'rigidification'; see Railton, 1998b; Wiggins, 1998). The difficulty now is that those in different moral traditions may simply turn out not to have a common moral vocabulary with us.

Nonsubjectivists typically reject the conceptual or internal connection between moral judgment and motivation, and claim that moral judgments can be made without linguistic impropriety even by someone for whom they are normally motivationally inert. Such 'externalist' approaches recognize that in our ordinary life we presuppose certain typical motivations – so that it will be somewhat misleading to utter 'This situation is immoral' if one does not intend thereby to indicate or express any negative attitude toward it. Expressions of belief, and of moral evaluations, play certain roles in our social lives owing to various prevailing or presupposed background conditions, but these do not all become part of the conventional meaning of these words. We need not be driven to surprising conclusions like relativism or rigidification by the so-called 'motivation problem,' the problem of capturing the intimate connection between moral judgment and motivation. Externalists are divided into naturalists, who think the properties attributed in moral discourse are within the ambit of ordinary empirical knowledge, and nonnaturalists, who think of moral properties as in some sense irreducible or *sui generis*.

The first need is to explain how morality can be so controversial, if moral questions are in principle resolvable by matters of natural fact; the second need is to explain how moral properties can be known or referred to if they are not causally related to our experience, or if they are causally related to our experience, how irreducibly moral properties could be so related (see Scanlon, 2003, 2009; Huemer, 2005; Shafer-Landau, 2005; Parfit, 2006, 2011a; Wedgwood, 2007). Some naturalists outright identify moral concepts with certain naturalistic concepts; others give a broadly functional characterization of the role or meaning or moral terms – the truisms about morality and its social role – and leave conceptually 'open' the question of which particular acts, practices, and so on play the relevant role (for some varieties, see Brandt, 1979; Sturgeon, 1985; Railton, 1986b; Boyd, 1988; Brink, 1989; Lewis, 1989; Smith, 1994; Jackson and Pettit, 1996; Jackson, 1998; Schroeder, 2007). Of course, one possibility is that nothing can be found that will play all or most of the roles ascribed to moral properties, resulting in an error theory of moral discourse rather like atheism about religion (Mackie, 1977). Nonnaturalists argue that morality is not alone in showing the need to give an account of how we are able to talk about or have knowledge of nonempirical phenomena – mathematics poses similar problems (see Ross, 1930; Field, 1989; Wiggins, 1998; Enoch, 2010).

G.E. Moore, a nonnaturalist externalist, famously diagnosed a 'naturalistic fallacy' in moral philosophy, but the import of this claim has been widely misunderstood (Moore, 1903). He did not doubt, and few in the twentieth century have doubted, that moral properties supervene, in some sense, on natural properties – roughly, those properties accessible to empirical experience and investigation. It is, according to this view, a mistaken picture of value to think that two worlds could be identical in all empirical features yet different in their evaluative properties – e.g., one good and the other bad. In this sense, a kind of transition from 'is' to 'ought' is possible: given the way in which moral properties supervene on natural properties, we can learn whether acts are good or bad by learning of their empirical features. Moore's point is that moral concepts are conceptually distinct from the natural concepts or criteria that we might use to apply them – we do not contradict ourselves by asking whether happiness is good in itself, even if, as a matter of fact, it is (Frankena, 1939; Gibbard, 2003).

Oscar Wilde has Mrs Cheveley say, after being reprimanded by Gertrude Chiltern, "Do you know, Gertrude, I don't mind your talking morality a bit. Morality is simply the attitude we adopt toward people whom we personally dislike" (*An Ideal Husband*, Act II). Such a 'debunking' depends on a contrast between the normative purport of morality and the normative status it actually possesses or deserves. If '*x* is wrong' simply meant "I personally don't like *x*," there would be no presumption that it would be bothersome to be told that what one is doing is wrong. Morality might simply dress up personal dislikes in judgmental clothing. But giving an account of the meaning of moral terms – a metatheory – involves saying just what this dressing up consists of. (A collection of related essays and a more extensive survey is to be found in Darwall et al., 1997.)

Normative Ethics

It is difficult to imagine that what is right and wrong, good and bad, and the like have nothing to do with likes and dislikes, or broadly, with what matters to people or to sentient beings generally. Thus, there does appear to be something profoundly subjective about what is moral. Yet, ethics strives to develop objective forms of evaluation on this subjective base. We can see normative moral theories as seeking to develop a synthesis of subjective and objective elements that: (1) presents a credible account of moral rightness and value, one which fits reasonably well with our firmest moral judgments on reflection; (2) is psychologically and socially realistic ('ought implies can,' – a condition that is growing in importance in philosophy as we learn more about cognitive and motivational psychology and social dynamics); (3) involves no suspect metaphysical or epistemological assumptions (see Parfit, 1984, 2011a; Sinnott-Armstrong and Timmons, 1996; Campbell and Hunter, 2000; Huemer, 2005; Shafer-Landau, 2005; Wedgwood, 2007; Scanlon, 2009); and (4) enables us to locate a compelling basis for moral evaluation that would explain how those judgments could credibly be justified, criticized, and improved – since we are all aware of moral judgments that once seemed to many people obviously right or self-evident (such as the permissibility of slavery or the

moral irrelevance of animals), that we have subsequently come to reject. A normative moral theory should illuminate the underlying basis or rationale for the correct moral judgments (if such can be found), even if it calls for some normative rethinking on our part. This set of concerns is sometimes called seeking wide reflective equilibrium (Rawls, 1971; *see* Reflective Equilibrium).

We find three major living traditions in normative moral theory, one claiming as its spiritual ancestor Aristotle, the second David Hume, and the third Immanuel Kant.

On Aristotelian accounts, the fundamental question of ethics is "What is the good life?," and the answer is thought to take the form: the virtuous life (for some Aristotelian approaches, see Aristotle, 1989; Foot, 1978, 2001; MacIntyre, 1981; McDowell, 1985; Williams, 1985; Hursthouse, 1999). Virtues are robust traits of character, drawing on normal human faculties and motivations but shaped by proper moral training that enable an individual to see what is best or required, and act appropriately, in a wide range of circumstances. Virtue thus involves cognitive, affective, and conative elements. Moreover, virtues are thought to be mutually reinforcing ('the unity of the virtues'), and to yield a life that is itself good for the virtuous person to lead. Some have argued that virtue ethics presupposes a faulty picture of human psychology, and that situationism in social psychology shows that there are no character traits, hence no virtues that human beings could develop (Harman, 1999; Doris, 2002). But virtue ethicists have responded by engaging more with empirical psychology and emphasizing how robust and multitrack are the dispositions with which virtues are identified (Sreenivasan, 2002; Prinz, 2009; Annas, 2011). Typically, one cannot formulate strict principles of action that regulate the evaluations, deliberations, and actions of the virtuous agent: moral conduct essentially involves the exercise of judgment (shaped by the 'moral experience' of the virtuous agent), yet this need not preclude fairly objective recommendations about specific situations – much as we think that persons of sensitivity, discernment, and knowledge tend to converge in esthetic judgment even if esthetic evaluation cannot be captured in terms of strict rules. Although virtue theories take agents and their motives as the primary locus of moral evaluation, the view need not simply be individualistic. Aristotle, for example, thought the virtuous life to be possible only in the right sort of social and political setting, and only if one plays one's appropriate role in that setting (*see* Virtue Ethics).

Humean accounts, although they may allow for the possibility of virtue, typically do not assign it so fundamental a role. They seek to understand morality as fundamentally a scheme of social cooperation and regulation that enables individuals and groups to achieve sought-after ends that they could not attain individually, or on the strength of native instinct. In Hume's own view (see Hume, 1888, a related view is found in Smith, 1976), we are natively equipped with many of the sentiments necessary for such cooperation and regulation – empathy (what Hume called 'sympathy'), an intrinsic concern for kith and kin, a willingness to defer to custom, and so on – but others within this tradition (e.g., Bentham) suppose humans to be egoistic hedonists. Each strand solves the 'motivational problem' for morality in a distinctive way – Hume by the reshaping of native sentiment toward greater impartiality and willingness to cooperate through social conventions and 'artificial' in addition to 'natural' virtues, Bentham by the use of socially backed 'sanctions' for changing the incentives individuals face. Both of these strands recognize the need for a common, objective standard for moral assessment: 'the general interest.'

An influential form of this general approach is utilitarianism, which has taken a vast number of particular forms. Essential to it are two features: (1) a notion of a person's good or well-being that can be assessed prior to or independently of judging moral worthiness ('the good is prior to the right') – a theory of intrinsic nonmoral value (which need not be hedonistic, but could be perfectionist or involve the satisfaction of informed desires, personal or impersonal); and (2) a way of evaluating acts or practices across individual lives and over time by comparing tendencies to promote intrinsic nonmoral value, impartially assessed (which characteristically is thought of as maximizing, but need not be). Such value-based theories are often called consequentialist or teleological (goal-oriented or end-oriented; *see* Consequentialism Including Utilitarianism). Morality partakes of the subjective – it is essentially concerned with human interests and responses – because intrinsic goodness is grounded in what does, or could, matter to people (or to sentient beings generally) for its own sake. Yet morality is objective, because it seeks to ground distinctively moral judgments on an unbiased and impartial assessment of how acts and practices affect the long-term realization of intrinsic values in our lives.

On Kantian accounts, there is an attempt to locate a ground of rightness that is not in this way wholly dependent on such tendencies to promote nonmoral good. In some sense, the end does not entirely justify the means. Certain constraints on right action, prerogatives for individual choice, and duties to self and others, exist that are not functions solely of actual or expected consequences. Indeed, there is on Kantian views some resistance to the idea that we can characterize 'intrinsic nonmoral good': happiness derived from using another inappropriately as a means, for example, torturing an innocent, is not, they claim, a good at all. Normative theories that thus claim that the right is prior to the good – that facts about moral rightness ground, or are more fundamental than, facts about intrinsic value – are sometimes called deontological (duty based). Although Kant's own view is a bit hard to categorize at its foundation, views of Kantian inspiration have become the dominant and best developed form of philosophical deontology. (Other forms include Lockean natural rights theories, e.g., Nozick, 1974; *see* Libertarianism; and rational intuitionism, e.g., Ross, 1930. For a discussion, see Scheffler, 1994.) Humean moral theorists are able to give a rationale for encouraging non-consequentialist moral norms or motivations by a 'second-order' consequentialist argument: they are necessary to avoid individually or collectively self-defeating behavior (see Parfit, 1984; Railton, 1984). But how are we to give a rational justification of moral actions and duties, or constraints, prerogatives, and special obligations, if we are willing to fly directly in the face of consequences for what we most value? The Kantian approaches this question from the standpoint of a responsible agent asking "What have I reason to do?" (*see* Practical Reasoning (Philosophical Aspects)). From this standpoint, we recognize that desires and goals do not automatically give

reasons to act, since they are themselves up for scrutiny – and not merely because some are stronger or more basic than the others. Perhaps a desire for revenge, however powerful or however much a part of human nature, ought to be unequivocally fought, or rather weighed in an individual or social 'utility function.' We know from experience that we are free enough from the control even of 'natural' desires and interests to do this.

If one claims that one's act is rationally permitted, one in effect licenses the principle on which it is based as legitimate for anyone in similar circumstances. Like ideal legislators, we thus authorize ourselves to act by 'making law,' subjecting ourselves to the very same law even when we would prefer otherwise; the imperative is categorical. This universalizability condition on our wills brings an element of objectivity and generality to our decisions, based not on actual or expected consequences, but on a notion of ourselves as being autonomous (literally, 'self-law-giving'; *see* Autonomy, Philosophy of), seeking the moral law (see Kant, 1956; also *see* Kantian Ethics and Politics). Some contemporary Kantians prefer to understand this in terms of a hypothetical social contract: we show that our action is permitted or required by rules that we would all agree to (or find reasonable) in a situation of initial freedom and equality. Some acts or practices that would have good consequences – promoting social peace by treating individuals paternalistically in a secretive way, or permitting inequalities that raise total utility but do not tend to benefit all members of society – might thereby be ruled out (Rawls, 1971; Scanlon, 1998; for a related view, expressed in terms of the conditions for nondistorted social communication, see Habermas, 1990; also *see* Contractarianism). Other contemporaries inspired by the notion of autonomy have seen in this doctrine a radical affirmation of individual human freedom. Jean-Paul Sartre, for example, gave a distinctively existentialist account of 'choosing for all mankind,' holding that we can lapse into bad faith if we attempt to escape this absolute responsibility for our acts by pointing to things 'beyond our control,' such as 'regrettable circumstances,' 'natural urges,' or even 'moral principle' itself (Sartre, 1946).

One might see these three traditions as each emphasizing a different, yet important root of morality: the notion of the moral life as the best or fittest sort of life for humans (Aristotle); the notion of the moral life as one of benevolence and freedom from prejudice (Hume); and the notion of the moral life as one of respect for one's own freedom and rationality as well as that of others (Kant). Can all these elements be combined into one normative theory? No overarching synthesis has been found, and any view that seeks to be at all systematic cannot, so to speak, guarantee that it is optimal in three dimensions at once. Kant, aware there could be no guarantee, nonetheless took it to be an essential condition of the rationality of moral conduct that autonomous self-legislation would in the end yield both good consequences overall and a good life for humans – happiness and virtue would be united as the 'highest good,' if not in this life, then in the afterlife – but later Kantians have largely given up this doctrine of 'rational faith' and argued that Kant's account can stand even without postulating that these conditions are met. In his landmark *The Methods of Ethics*, Henry Sidgwick argued that "the intuitional method rigorously applied yields as its final result the doctrine of pure universalistic Hedonism" (1907: pp. 406–7). In a similar spirit, Derek Parfit has defended the provocative thesis that the best versions of Kantianism, consequentialism, and contractualism agree in their moral evaluations of all possible cases; they are "climbing the same mountain on different sides": the principles whose universal acceptance everyone could rationally will are identical to the principles whose universal acceptance would lead to the best results, which are identical to the principles no one could reasonably reject (Parfit, 2011a,b). This leaves open, among other things, the question of which normative factors are most fundamental. Even then, many are not convinced (see the commentaries in Parfit, 2011a).

An alternative response to this sense of multiple roots of morality is to claim that no systematic account of normative ethics should be attempted. Different circumstances, roles, personal histories, and so on make different sorts of moral consideration more or less appropriate, and this is an ineluctable fact of moral life. A relation with one's spouse or child is seldom best regulated by 'the general interest' or 'universalizable maxims,' a legislator cannot generally rely on personal virtue to tell her how to shape a complex piece of environmental law, and in our relations with animals we cannot regulate our conduct as if we were operating in a reciprocal realm of rationality. Experience, wisdom, and caring, not theory, enable us to act and deliberate well in a complex moral world (see Larmore, 1987). This view was given some impetus by, and finds continuing support in, a growing literature on moral development, where individuals who score highly on traditional tests of moral-theoretical sophistication may do poorly in practice in comparison with those capable of greater personal caring, empathy, and adaptiveness (see Gilligan, 1982; Damasio, 1993). Defenders of normative theory insist that systematic ambition is an essential antidote to arbitrariness or narrow-mindedness. Normative theories need not be understood as offering rule-like standards for acts or for moral reasoning; such theories must of course be combined with suitable motivation and sensitivity. Moral particularists (Dancy, 1993, 2004; Hooker and Little, 2000) and some feminists (see Card, 1991; Walker, 2007) counter that the systematizing impulse can itself become a form of moral insensitivity and arbitrariness. Disagreement remains concerning how best to understand the relation between normative theory and ethical reasoning and practice.

This is probably a healthy dialectic in normative ethics. It is important to emphasize, however, that resistance to normative moral theorizing need not be combined with a resistance to metaethical theorizing: moral particularism is often defended on the ground of an account of the meaning of moral concepts or on the ground of moral justification. Thus, on one influential view (Ross, 1930), morality consists of many mutually independent prima facie duties and permissions, the bearing of which in particular situations is not governed by general rules. Ross grounded his view in a nonnaturalist intuitionism in metaethics, and numerous contemporary defenses of Ross-like normative theories or other forms of particularism draw upon the idea that moral concepts are response dependent (see Dancy, 1993) or that fundamental values are plural and incommensurable (see Williams, 1981; Stocker, 1990; Chang, 1997; Thomson, 1997).

Theory of Value

In the first half of the twentieth century, theory of value (axiology) was a flourishing, central area of philosophy, represented by such major figures as G.E. Moore (1903), John Dewey (1929), C.I. Lewis (1962), and Ralph Barton Perry (1954). It has since gone so thoroughly out of style that it is difficult for a contemporary philosopher to recognize the themes or structure of a book like Perry's *Realms of Value*, with its chapters on *Motor-Affective Psychology*, *Gradation, Integration, and Mutation*, *The Science of Science*, and *The History of History*. This may partly be the result of a familiar tendency within disciplines to subdivide into specialities: ethics, esthetics, rational choice, law, and epistemology – all of these are domains of evaluation, yet they seldom are united together under a common rubric. Yet, like specialization generally, it involves some loss of perspective. The value-theoretic aspects of some subdisciplines within philosophy became, at least until fairly recently, relatively invisible (epistemology, for example) or radically de-emphasized ('philosophical aesthetics'). This has also given the mistaken impression that the so-called 'fact/value' distinction is disciplinary – ethics on the one side, science on the other. But value theorists would be quick to point out that values (but not simply moral values) as well as facts are fundamentally involved in all areas of human practice and inquiry.

There are excellent reasons for focusing on value as such. There is independent interest in investigating the nature and structure of value. Some questions about objectivity, cognitivism, and realism are familiar from metaethics, but other questions, such as the distinctions among the various forms of value (intrinsic, extrinsic, inherent, personal, impersonal, and so on), the possibility of commensurability or interpersonal comparison among values, and the relation of value to valuing attitudes or affect, merit sustained scrutiny. Fortunately, recent years have seen a growing revival of philosophical research into the nature of value (Anderson, 1993; Hurka, 1996; Railton, 1998b; Broome, 1999; Zimmerman, 2001; Rønnow-Rasmussen and Zimmerman, 2005; Rønnow-Rasmussen, 2011; *see* Value Pluralism).

Moreover, that moral philosophy in the twentieth century has been overwhelmingly preoccupied with questions of rightness and obligation (deontic or directive categories), rather than questions of value (axiological or evaluative categories), has given us a lopsided picture of morality and of the 'motivation problem' for ethics. Perhaps an 'ought' judgment – better, a 'must' judgment – expresses some motivation on the speaker's part. But judgments that a way of life would be valuable seem to highlight more the dependence of normative facts on the actual or potential motivation of those to whom they are applied. Further, evaluative judgments in general seem more directly responsive to facts of psychology, sociology, and history than are directive judgments (see Williams, 1985; Wiggins, 1998). The emphasis on deontic judgment has also tended to divorce philosophical ethics from value-involving subjects where duty plays a less conspicuous role. Kant and Hume, for example, saw quasi-esthetic evaluative judgment as at the foundation of morality.

Finally, a certain lost opportunity for philosophy for mutual enlightenment between 'analytic' and 'continental' philosophy occurred during the twentieth century. Moore (1903) observed that the philosopher whose outlook seemed closest to his own was Brentano (1902). Theorists within the phenomenological tradition in philosophy and psychology, such as Köhler (1938) and Heidegger (1962), explored our experience of value in life in ways quite comprehensible from the standpoint of, say, that of Dewey (*see* Phenomenology: Philosophical Aspects). Frankfurt School (see Habermas, 1990) criticisms of an overly narrow construal of value can be read alongside a growing body of Anglo-American literature critical of current notions of value in philosophy and economics, especially in rational choice theory and cost–benefit analysis (see MacLean, 1986; Anderson, 1993; also *see* Economics and Ethics).

See also: Consequentialism (Including Utilitarianism); Conventions and Norms: Philosophical Aspects; Dewey, John (1859–1952); Evolution, Natural and Social: Philosophical Aspects; Hume, David (1711–76); Kant, Immanuel (1724–1804); Kantian Ethics and Politics; Kohlberg, Lawrence (1927–87); Law and Morality: A Continental European Perspective; Law and Morality: An Analytical Perspective; Moral Reasoning in Psychology; Norms; Relativism: Philosophical Aspects; Social Science, The Idea of; Values, Social Psychology of; Virtue Ethics.

Bibliography

Anderson, E., 1993. Value in Ethics and Economics. Harvard University Press, Cambridge.
Annas, J., 1993. The Morality of Happiness. Oxford University Press, New York.
Annas, J., 2011. Intelligent Virtue. Oxford University Press, New York.
Aristotle, 1989. The Nichomachean Ethics (W.D. Ross, Trans.), rev. ed. Oxford University Press, Oxford.
Ayer, A., 1936. Language, Truth, and Logic. Dover, New York.
Baier, K., 1958. The Moral Point of View. Cornell University Press, Ithaca.
Berker, S., 2009. The normative insignificance of neuroscience. Philosophy and Public Affairs 37, 293–329.
Blackburn, S., 1984. Spreading the Word. Clarendon Press, Oxford.
Blackburn, S., 1993. Essays in Quasi-realism. Oxford University Press, New York.
Blackburn, S., 1998. Ruling Passions. Clarendon Press, Oxford.
Boyd, R., 1988. How to be a moral realist. In: Sayre-McCord, G. (Ed.), Essays on Moral Realism. Cornell University Press, Ithaca, pp. 181–228.
Brandt, R.B., 1979. A Theory of the Good and the Right. Oxford University Press, New York.
Brentano, F.C., 1902. The Origin of Knowledge of Right and Wrong (Cecil Hague, Trans.). Constable.
Brink, D.O., 1989. Moral Realism and the Foundations of Ethics. Cambridge University Press, Cambridge.
Broome, J., 1999. Ethics Out of Economics. Cambridge University Press, Cambridge.
Campbell, R., Hunter, B. (Eds.), 2000. Moral Epistemology Naturalized. University of Calgary Press, Calgary.
Card, C. (Ed.), 1991. Feminist Ethics. University Press of Kansas, Lawrence, KS.
Chang, R. (Ed.), 1997. Incommensurability, Incomparability, and Practical Reason. Harvard University Press, Cambridge.
Copp, D., 2008. Darwinian skepticism about moral realism. Philosophical Issues 18 (1), 186–206.
Damasio, A.R., 1993. Descartes' Error: Emotion, Reason, and the Human Brain. G.P. Putnam, New York.
Dancy, J., 1993. Moral Reasons. Blackwell, Cambridge.
Dancy, J., 2004. Ethics without Principles. Clarendon Press, Oxford.
Darwall, S., Gibbard, A., Railton, P. (Eds.), 1997. Moral Discourse and Practice: Some Philosophical Approaches. Oxford University Press, New York.
Dewey, J., 1929. The Quest for Certainty. Minton Balch, New York.
Doris, J.M., 2002. Lack of Character: Personality and Moral Behavior. Cambridge University Press, Cambridge.

Dreier, J., 2004. Meta-ethics and the problem of creeping minimalism. Philosophical Perspectives 18, 23–44.
Enoch, D., 2010. The epistemological challenge to metanormative realism: how best to understand it, and how to cope with it. Philosophical Studies 148 (3), 413–438.
Field, H., 1989. Realism, Mathematics, and Modality. Blackwell, New York.
Fiske, A.P., 1991. Structures of Social Life: The Four Elementary Forms of Human Relations. Free Press, New York.
Foot, P., 1978. Virtues and Vices. Blackwell, Oxford.
Foot, P., 2001. Natural Goodness. Clarendon Press, Oxford.
Frankena, W., 1939. The naturalistic fallacy. Mind 48, 464–477.
Gibbard, A., 1990. Wise Choices, Apt Feelings. Harvard University Press, Cambridge.
Gibbard, A., 2003. Thinking How to Live. Harvard University Press, Cambridge.
Gilligan, C., 1982. In a Different Voice: Psychological Theory and Women's Development. Harvard University Press, Cambridge.
Greene, J.D., 2003. The secret joke of Kant's soul. In: Sinnott-Armstrong, W. (Ed.), Moral Psychology, The Neuroscience of Morality: Emotion, Brain Disorders, and Development, vol. 3. MIT Press, Cambridge, pp. 35–79.
Habermas, J., 1990. Moral Consciousness and Communicative Action (Shierry Weber Nicholsen & Christian Lenhardt, Trans.). MIT Press, Cambridge.
Hare, R., 1952. The Language of Morals. Oxford University Press, Oxford.
Harman, G., 1999. Moral philosophy meets social psychology: virtue ethics and the fundamental attribution error. Proceedings of the Aristotelian Society 99, 315–331.
Harman, G., Thomson, J.J., 1996. Moral Relativism and Moral Objectivity. Blackwell, Cambridge.
Heidegger, M., 1962. Being and Time. John Macuarrie & Edward Robinson. Harper, New York.
Hooker, B., Little, M. (Eds.), 2000. Moral Particularism. Oxford University Press, Oxford.
Horwich, P., 1990. Truth, second ed. Blackwell, Oxford.
Huemer, M., 2005. Ethical Intuitionism. Palgrave MacMillan, New York.
Hume, D., 1888. A Treatise of Human Nature. L.A. Selby-Bigge, second ed. Oxford University Press, Oxford.
Hurka, T., 1996. Perfectionism. Oxford University Press, New York.
Hursthouse, R., 1999. On Virtue Ethics. Oxford University Press, Oxford.
Jackson, F., 1998. From Metaphysics to Ethics: A Defense of Conceptual Analysis. Oxford University Press, Oxford.
Jackson, F., Pettit, P., 1996. Moral functionalism and moral motivation. Philosophical Quarterly 46, 82–86.
Joyce, R., 2006. The Evolution of Morality. MIT Press, Cambridge.
Kant, I., 1956. Groundwork of the Metaphysics of Morals (H.J. Paton, Trans.), third ed. Harper & Row, New York.
Katz, L.D. (Ed.), 2000. Evolutionary Origins of Morality: Cross-disciplinary Perspectives. Imprint Academic, Bowling Green, OH.
Kitcher, P., 2006. Between fragile altruism and morality: evolution and the emergence of normative guidance. In: Boniolo, G., De Anna, G. (Eds.), Evolutionary Ethics and Contemporary Biology. Cambridge University Press, Cambridge, pp. 159–177.
Knobe, J., Nichols, S. (Eds.), 2008. Experimental Philosophy. Oxford University Press, New York.
Köhler, 1938. The Place of Value in a World of Facts. Liveright, New York.
Larmore, C.E., 1987. Patterns of Moral Complexity. Cambridge University Press, Cambridge.
Lewis, C.I., 1962. An Analysis of Knowledge and Valuation. Open Court, La Salle.
Lewis, D., 1989. Dispositional theories of value. Proceedings of the Aristotelian Society 63 (Suppl), 113–137.
MacIntyre, A., 1981. After Virtue. University of Notre Dame Press, Notre Dame.
Mackie, J., 1977. Ethics: Inventing Right and Wrong. Penguin, Harmondsworth.
MacLean, D., 1986. Values at Risk. Rowman & Allanheld, Totowa, NJ.
McDowell, J., 1985. Values and secondary qualities. In: Honderich, T. (Ed.), Morality and Objectivity. Routledge & Kegan Paul, London, pp. 110–129.
Moore, G., 1903. Principia Ethica. Cambridge University Press, New York.
Nozick, R., 1974. Anarchy, State, and Utopia. Basic Books, New York.
Parfit, D., 1984. Reasons and Persons. Oxford University Press, Oxford.
Parfit, D., 2006. Normativity. In: Shafer-Landau, R. (Ed.), Oxford Studies in Metaethics, vol. 1. Clarendon Press, Oxford, pp. 325–380.
Parfit, D., 2011a. On What Matters, vol. II. Oxford University Press, Oxford.
Parfit, D., 2011b. On What Matters, vol. I. Oxford University Press, Oxford.
Perry, R.B., 1954. Realms of Value: A Critique of Human Civilization. Harvard University Press, Cambridge.
Prichard, H., 1912. Does moral philosophy rest on a mistake? Mind 21, 21–37.
Prinz, J., 2009. The normativity challenge: cultural psychology provides the real threat to virtue ethics. Journal of Ethics 13, 117–144.
Railton, P., 1984. Alienation, consequentialism, and the demands of morality. Philosophy and Public Affairs 13 (2), 134–171.
Railton, P., 1986a. Facts and values. Philosophical Topics 14 (3), 5–31.
Railton, P., 1986b. Moral realism. Philosophical Review 95, 163–207.
Railton, P., 1998a. Aesthetic value, moral value, and the ambitions of naturalism. In: Levinson, J. (Ed.), Aesthetics and Ethics: Essays at the Intersection. Cambridge University Press, Cambridge, pp. 59–105.
Railton, P., 1998b. Red, bitter, good. European Review of Philosophy 3, 67–84.
Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge.
Rønnow-Rasmussen, T., 2011. Personal Value. Oxford University Press, Oxford.
Rønnow-Rasmussen, T., Zimmerman, M.J. (Eds.), 2005. Recent Work on Intrinsic Value. Springer, Dordrecht.
Ross, W., 1930. The Right and the Good. Oxford University Press, Oxford.
Rozin, P., Lowery, L., Imada, S., Haidt, J., 1999. The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral ethics (community, autonomy, divinity). Journal of Personality and Social Psychology 76, 574–586.
Sartre, J.-P., 1946. L'existentialism Est Un Humanisme. Nage, Paris.
Scanlon, T., 1998. What We Owe to Each Other. Harvard University Press, Cambridge.
Scanlon, T., 2003. Metaphysics and morals. Proceedings and Addresses of the American Philosophical Association 77 (2), 7–22.
Scanlon, T., 2009. Being Realistic about Reasons: The John Locke Lectures 2009. Oxford.
Scheffler, S., 1994. The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions, second ed. Oxford University Press, Oxford.
Schroeder, M., 2007. Slaves of the Passions. Oxford University Press, Oxford.
Schroeder, M., 2008. Being for: Evaluating the Semantic Program of Expressivism. Oxford University Press, Oxford.
Schroeder, M., 2009. Hybrid expressivism: virtues and vices. Ethics 119, 257–309.
Shafer-Landau, R., 2005. Moral Realism: A Defense. Oxford University Press, Oxford.
Shweder, R.A., Much, N.C., Mahapatra, M., Park, L., 1997. The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. In: Brandt, A., Rozin, P. (Eds.), Morality and Health. Routledge, New York, pp. 119–169.
Sidgwick, H., 1907. The Methods of Ethics, seventh ed. Hackett, Indianapolis.
Singer, P., 2005. Ethics and intuitions. The Journal of Ethics 9, 331–352.
Sinnott-Armstrong, W., Timmons, M. (Eds.), 1996. Moral Knowledge?. Oxford University Press, New York.
Smith, A., 1759/1976. The Theory of Moral Sentiments. Oxford University Press, New York.
Smith, M., 1994. The Moral Problem. Blackwell, Oxford.
Sreenivasan, G., 2002. Errors about errors: virtue theory and trait attribution. Mind 111, 47–68.
Stevenson, C., 1937. The emotive meaning of ethical terms. Mind 46, 14–31.
Stevenson, C., 1944. Ethics and Language. Yale University Press, New Haven.
Stocker, M., 1990. Plural and Conflicting Values. Oxford University Press, New York.
Street, S., 2006. A darwinian dilemma for realist theories of value. Philosophical Studies 127, 109–166.
Sturgeon, N.L., 1985. Moral explanations. In: Copp, D., Zimmerman, D. (Eds.), Morality, Reason, and Truth. Rowman & Allanheld, Totowa, NJ, pp. 49–78.
Thompson, P. (Ed.), 1995. Issues in Evolutionary Ethics. State University of New York Press, Albany.
Thomson, J.J., 1997. The right and the good. Journal of Philosophy 94, 273–298.
Walker, M.U., 2007. Moral Understandings: A Feminist Study in Ethics, second ed. Oxford University Press, New York.
Wedgwood, R., 2007. The Nature of Normativity. Oxford University Press, Oxford.
Wiggins, D., 1998. Needs, Values, Truth: Essays in the Philosophy of Value, third ed. Clarendon Press, Oxford.
Williams, B., 1981. Moral Luck. Cambridge University Press, Cambridge.
Williams, B., 1985. Ethics and the Limits of Philosophy. Harvard University Press, Cambridge.
Zimmerman, M.J., 2001. The Nature of Intrinsic Value. Rowman & Littlefield, Lanham, MD.

Ethics for Biomedical Research Involving Humans: International Guidelines

John R Williams, Ottawa, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.J. Levine, volume 7, pp. 4788–4792, © 2001, Elsevier Ltd.

Abstract

In response to the atrocities committed by German medical researchers before and during World War II, the judges at the Nuremberg doctors' trial formulated the first set of international ethics guidelines for biomedical research on humans, the Nuremberg Code. Subsequently, numerous other guidelines have been developed and revised, beginning with the World Medical Association's Declaration of Helsinki and followed by the Council for International Organizations of Medical Sciences' International Ethical Guidelines for Biomedical Research Involving Human Subjects and the International Council on Harmonization's Guideline for Good Clinical Practice, to mention only the most influential. This article compares these four documents on how they protect participants in biomedical research.

Prior to World War II there were no international ethics guidelines for biomedical research involving human subjects. Since then, numerous sets of guidelines have been developed and promulgated, beginning with the Nuremberg Code (1947) and followed by the World Medical Association's (WMA) Declaration of Helsinki (1964, revised nine times, most recently in 2013), the International Ethical Guidelines for Biomedical Research Involving Human Subjects developed by the Council for International Organizations of Medical Sciences (CIOMS) in collaboration with the World Health Organization (WHO) (1982, revised in 1993 and 2002), and the International Council on Harmonization's (ICH) Guideline for Good Clinical Practice (1996), to mention only the most influential. These will be referred to respectively as 'Nuremberg,' 'DoH,' 'CIOMS,' and 'ICH.'

At the close of World War II, in Nuremberg, Germany, certain leaders of the Nazi regime were put on trial for having committed 'crimes against humanity.' During the course of one of these trials, the so-called doctors' trial, the first international statement of research ethics, the Nuremberg Code, was written. Subsequent international guidelines were designed to improve on or to interpret the efforts of their predecessors. WMA began its project in 1953 because it recognized a need for professional guidelines designed by physicians for physicians (as opposed to the Nuremberg Code, which was formed by jurists in a legal trial). The result of this project was the DoH, which was adopted at the World Medical Assembly in Helsinki, Finland in 1964. CIOMS introduced its Guidelines in 1982 by stating that they were designed to provide guidance for the correct application of the principles of the DoH, particularly when the research is initiated by researchers and sponsors in technologically developed countries and carried out in developing countries. Subsequent revisions of the CIOMS Guidelines in 1993 and 2002 departed from some of the DoH provisions and they are no longer considered simply as a commentary on the DoH. At the time of this writing, CIOMS Guidelines are once again being revised. Like CIOMS, ICH explicitly references the DoH:

> Good Clinical Practice (GCP) is an international ethical and scientific quality standard for designing, conducting, recording and reporting trials that involve the participation of human subjects. Compliance with this standard provides public assurance that the rights, safety and well-being of trial subjects are protected, consistent with the principles that have their origin in the Declaration of Helsinki, and that the clinical trial data are credible.
>
> Guideline E6: p. 1

These four documents vary considerably in length and scope. Nuremberg contains just 504 words in 10 paragraphs. The latest (2013) version of the DoH has 2240 words in 37 paragraphs. CIOMS 2002 is 65 pages long, consisting of a statement of general ethical principles, a preamble and 21 guidelines, each with an extensive commentary, plus a five-page list of items to be included in a research protocol. ICH is 53 pages long; however, it covers many other aspects of research besides ethics, the treatment of which is relatively brief. The titles or subtitles of the documents reflect differences in their scope. Nuremberg has no subtitle but the text makes clear that it applies to all types of experiments on humans. The original (1964) DoH was subtitled 'Recommendations Guiding Doctors in Clinical Research.' In the 1975 version 'clinical research' was changed to 'biomedical research involving human subjects.' In the 2000 and all subsequent versions, 'biomedical' was replaced by 'medical.' All three versions of the CIOMS Guidelines use the term 'biomedical research.' ICH focuses exclusively on clinical trials.

Ethical Universalism or Pluralism?

Discussions of international guidelines for research ethics generally reflect a tension between ethical universalism and cultural pluralism. Universalists hold that ethical principles are similar, if not identical, in all human societies and in all historical periods. It is as if ethical principles exist in human nature and are there to be discovered. The reason that they appear to change from time to time is that our ability to discern and understand them changes (one hopes for the better). The pluralistic position, by contrast, is that ethics are invented (not discovered) in the course of conversations conducted within particular societies; the results will necessarily reflect the histories and traditions of the cultures in which the conversations are conducted (Levine, 1996).

The authors of Nuremberg and the DoH attempted to create universal documents. ICH is addressed to the United States, European Union, and Japan but is referenced in many other countries. By contrast, CIOMS, which although it recognizes some ethical principles as 'transcendent,' also acknowledges the validity of some degree of cultural pluralism. The CIOMS writers aspire to 'global applicability' of a document that provides adequate ethical guidance for all known contemporary societies with the understanding that revisions will be required from time to time as ethics change in response to changing circumstances and understandings (Levine, 1996).

International statements of research ethics have no legal authority except to the extent that parts or all of them are incorporated in the laws of nations or other political entities having the authority to enact and enforce laws. This point notwithstanding, these documents have important moral status in that those who wish to criticize the ethics of various research activities generally rely on one or more of these documents to ground their judgments.

The Content of the Ethical Guidelines

Three ethical principles are widely recognized as basic or fundamental in the ethics of biomedical research and practice. These principles, which were first articulated in their most widely used form by the U.S. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (National Commission), are 'respect for persons,' 'beneficence,' and 'justice.' The first explicit reference to these principles in the international statements is in the 1993 version of the CIOMS Guidelines. However, the National Commission found in Nuremberg and the DoH implicit statements of respect for persons and beneficence (National Commission, 1978; Levine, 1988); concern for justice is also implied in the introduction to Nuremberg. Whether or not these principles are universally valid is the topic of a controversy beyond the scope of this article (Levine, 1996; Macklin, 1992).

The purpose of the ethical guidelines is to provide guidance on how research involving human subjects should be carried out so that it is responsive to the requirements of the basic ethical principles. Because they are considered relevant to virtually all fields of human behavior, the principles themselves are stated at such a level of abstraction that they do not provide clear direction on what behaviors are required or forbidden. Ethical norms (or rules) are incorporated in ethical guidelines to clarify what behaviors are prescribed or proscribed by the document. Ethical norms are either substantive or procedural. A substantive norm specifies behaviors that are required (or forbidden) because they are morally right (or wrong). Some procedural norms specify procedures that should be performed to determine the most appropriate behavior when the behavior has not been specified by a substantive norm. Other procedural norms are designed to provide documentary evidence that research has been conducted in accord with the relevant substantive norms (Levine, 1988).

An illustration. The principle of respect for persons requires that human persons must never be treated only as means to the ends of others. The substantive norm that requires informed consent is a specification of one way in which this principle is to be made operational in the conduct of research involving human subjects. The procedural norm that requires review and approval by a research ethics committee affords a method for determining what specific bits of information must be divulged to prospective subjects in a particular research protocol. Another procedural norm that requires the signing of a consent form provides documentary evidence that the behavior required by the substantive norm has been accomplished.

Ethical codes written by ethical universalists tend to have a high proportion of substantive norms. Those that espouse cultural pluralism tend to have a larger proportion of procedural norms. Substantive norms predominate in Nuremberg and the DoH while CIOMS and ICH contain a relatively large number of procedural norms along with abundant commentary on how the norms should be interpreted in various specific circumstances.

All of the substantive norms articulated in the leading international research ethics guidelines are variations on and elaborations of seven general normative statements (Levine, 1988). The first three of these are most closely related to the principle of beneficence which requires that those who perform research should maximize possible benefits and minimize possible harms and, further, that they should not deliberately injure any human being.

1. There should be good research design. This norm is linked inextricably to the purpose of doing research – to contribute to the development of generalizeable new knowledge. If the research is not designed adequately, there is little or no chance that it will make the desired contribution to the development of knowledge. As stated in CIOMS (2002: p. 23), "scientifically invalid research is unethical in that it exposes research subjects to risks without possible benefit."
2. There should be competent investigators. One purpose of this norm is identical to that of the first; competent investigators are much more likely than incompetents to succeed in producing beneficial new knowledge. In addition, there must be at least one member of the team of researchers who has the clinical competence to monitor the health of the subjects with a view toward prompt intervention as indicated to preserve the health of the subject by, for example, ending the individual's participation in the research or by administration of an antidote to some toxic substance. As stated in the DoH (paragraph 12): "Medical research involving human subjects must be conducted only by individuals with the appropriate ethics and scientific education, training and qualifications. Research on patients or healthy volunteers requires the supervision of a competent and appropriately qualified physician or other health care professional."
3. There should be a favorable (or reasonable) balance of risks and benefits. This is a judgment that must be made by the investigator and affirmed by the research ethics committee before the prospective subject is presented with an invitation to participate. The risks of interventions and procedures that hold out the prospect of direct benefit to the subject are justified much the same as they are in the conduct of medical practice – in terms of the anticipated benefit to the subject. The risks and benefits of interventions and

procedures that do not hold out the prospect of direct benefit for the individual subject (nonbeneficial procedures) must be justified in terms of the anticipated benefit to society. Nuremberg (Principle 6) requires that: "The degree of risk to be taken should never exceed that determined by the humanitarian importance of the problem to be solved." Plans to employ nonbeneficial procedures that present more than minimal risk in vulnerable populations require special justification. If in the course of conducting the research, the balance of risks and benefits ceases to be favorable, the investigator is required to intervene appropriately. Nuremberg requires: "During the course of the experiment the scientist … must be prepared to terminate the experiment … if he has probable cause to believe … that a continuation … is likely to result in injury, disability or death" (Principle 10). The DoH does not require separate analysis of the risks and benefits of individual components of research as described above. Rather, it calls for the evaluation of the entire research protocol as either beneficial or nonbeneficial: "Physicians who combine medical research with medical care should involve their patients in research only to the extent that this is justified by its potential preventive, diagnostic or therapeutic value and if the physician has good reason to believe that participation in the research study will not adversely affect the health of the patients who serve as research subjects" (paragraph 14).

The next two of the general substantive norms are most related to the principle of respect for persons:

4. There should be informed consent. The first principle of Nuremberg set forth the definition of informed consent that, with variations, has been incorporated in all subsequent guidelines and regulations:

> The voluntary consent of the human subject is absolutely essential. This means that the person involved should have legal capacity to give consent; should be so situated as to be able to exercise free power of choice, without the intervention of any element of force, fraud, deceit, duress, overreaching, or other ulterior form of constraint or coercion; and should have sufficient knowledge and comprehension of the elements of the subject matter involved as to enable him to make an understanding and enlightened decision. This latter element requires that before the acceptance of an affirmative decision by the experimental subject there should be made known to him the nature, duration, and purpose of the experiment; the method and means by which it is to be conducted; all inconveniences and hazards reasonably to be expected; and the effects upon his health or person which may possibly come from his participation in the experiment.

In this way, the prospective subject is invited to accept the researcher's goals as if they were his or her own at least for the duration of his or her participation. The result is that the subject is not treated as *merely* a means to the researcher's goals.

5. There should be respect for privacy and maintenance of confidentiality. 'Privacy' and 'confidentiality' are often incorrectly used as synonyms. 'Privacy' refers to the "freedom of the individual to pick and choose for himself the time and circumstances under which, and most importantly, the extent to which his attitudes, beliefs, behavior and opinions are to be shared with or withheld from others" (Kelman, 1977). 'Confidentiality' refers to a mode of management of private information; if a subject shares private information with (confides in) an investigator, the investigator is expected to refrain from sharing this information with others without the subject's authorization or some other justification. CIOMS requires that: "The investigator must establish secure safeguards of the confidentiality of subjects" research data. Subjects should be told the limits, legal or other, to the investigators' ability to safeguard confidentiality and the possible consequences of breaches of confidentiality (Guideline 18).

The final two general substantive norms are most closely related to the principle of justice which requires fairness in the distribution of both the burdens and benefits of research.

6. Selection of subjects should be equitable. From the 1940s through the mid-1980s, the prevailing attitude among commentators and policy makers in the field of research ethics was that participation in research was dangerous to the subjects and that researchers were inclined to exploit the socially and economically less powerful individuals who were most likely to serve as research subjects. For this reason, the ethical guidelines and regulations of that time focused on the protection of the prospective subject from bearing an unfair share of the burdens of research. The criterion for determining whether an individual (or a class of individuals) was in need of 'special protections' was 'vulnerability.' In the 1970s, vulnerability tended to be equated with 'limited capacity to give informed consent' and special procedural and substantive norms were developed to protect the rights and welfare of children, prisoners, the institutionalized mentally infirm, and fetuses (Levine, 1998). The category 'vulnerable' has been expanded subsequently to include others including "medical and nursing students, subordinate hospital and laboratory personnel, employees of pharmaceutical companies, and members of the armed forces or police" (CIOMS, 2002: p. 64), and patients with incurable diseases, persons in nursing homes, unemployed or impoverished persons, patients in emergency situations, ethnic minority groups, homeless persons, nomads, refugees, and those incapable of giving consent (ICH E6: p. 8).

Since the mid-1980s, research participation increasingly has been recognized as beneficial to both individuals and the class of persons they represent. This change began as a response to the AIDS activists who in 1986 pointed out that participation in a placebo-controlled clinical trial of AZT was not to be considered burdensome; it was, they argued, beneficial in that it was the only way people could have access to the only drug that anyone believed might be directed at the cause of their disease. Since then, ethical guidelines and regulations have begun to interpret the requirement for equitable selection of subjects as requiring also that vulnerable persons be assured of receiving a fair share of the benefits of participation in research. The 2013 DoH states: "Medical research with a vulnerable group is only justified if the research is responsive to the health needs or priorities of this group and the research cannot be

carried out in a non-vulnerable group. In addition, this group should stand to benefit from the knowledge, practices or interventions that result from the research" (paragraph 20).

7. There should be compensation for research-induced injury. CIOMS requires that "Research subjects who suffer physical injury as a result of their participation are entitled to free medical treatment for such injury and to such financial or other assistance as would compensate them equitably for any resultant impairment, disability or handicap" (2002: Guideline 19).

Exploitation of Subjects in Developing Countries

Industrial sponsors and investigators from technologically developed countries often conduct research designed to evaluate their therapeutic innovations in developing countries. Some commentators (e.g., Macklin, 2004) claim that this is done to lower the costs of the research and to take advantage of the less complex and sophisticated regulatory systems typical of developing countries. This is considered unjust in that these innovations, once proved safe and effective, are generally not available to the residents of developing countries owing to economic factors.

With the aim of preventing such exploitation CIOMS requires special justification for conducting research in developing countries (e.g., the research must be responsive to the health needs and priorities of the population or community in which it is to be carried out) and special obligations for the sponsors (e.g., they must engage in capacity-building in the host country to enable it to conduct independent ethical and scientific review). The 2002 version of these Guidelines requires that any product developed or knowledge generated in such research should be made "reasonably available for the benefit of that population or community" (Guideline 10).

The DoH requires that a new intervention must be "tested against those of the best current interventions(s), except in [limited] circumstances" (paragraph 33). This standard, which is similar to CIOMS Guideline 11 (although CIOMS uses the term 'established effective intervention'), is designed to prevent exploitation of all patients, including those residing in developing countries. However, some clinical trial experts argue that it may also have the undesired effect of obstructing the development for use in developing countries of inexpensive alternatives to expensive therapies employed in developed countries. The latter are generally the 'best current interventions.' An alternative which would not have this disadvantage is 'the highest attainable and sustainable' standard (Levine, 2000). 'Highest attainable' means that one should provide the best one can offer under the circumstances of the research rather than settle for the standard prevailing in the host country (often no therapy at all). The upper limit of what should be provided is defined by 'sustainability.' It would be unethical to develop a therapy that could not be made available to residents of the host country (sustained) after the research has been completed. Such research would not have been responsive to the health needs and priorities of the host country.

Informatics and Biobanking

Whereas the Nuremberg Code was written once and for all and therefore reflects the state of research at the time, the DoH and CIOMS have been revised periodically to take account of advances in research as well as other social, political, economic, and ethical developments. The ICH could likewise benefit from updating but the organization has given no indication that this will happen.

With the rapid growth of computing power from the 1970s onward, great advances have been made in medical informatics. The proliferation and increasing sophistication of databases and biobanks have enabled researchers in epidemiology, genetics and many other fields to work far more efficiently than before. At the same time these research tools have engendered new ethical issues and sharpened old ones such as privacy and confidentiality. Recent revisions of the DoH have incorporated some of these concerns while maintaining its status as primarily a statement of principles. The first paragraph of the 2000 and subsequent versions of the DoH specifies that medical research involving human subjects includes research on identifiable human material and data. A new paragraph in the 2008 version stated: "For medical research using identifiable human material or data, physicians must normally seek consent for the collection, analysis, storage and/or reuse. There may be situations where consent would be impossible or impractical to obtain for such research or would pose a threat to the validity of the research. In such situations the research may be done only after consideration and approval of a research ethics committee" (paragraph 25). This paragraph was only slightly modified in the 2013 DoH.

CIOMS 2002 contain just one brief reference to health databases and biobanks, in the commentary on Guideline 18, 'Safeguarding Confidentiality,' the section on "Issues of Confidentiality in Genetics Research": "When biological samples are not fully anonymized and when it is anticipated that there may be valid clinical or research reasons for linking the results of genetic tests to research subjects, the investigator in seeking informed consent should assure prospective subjects that their identity will be protected by secure coding of their samples (encryption) and by restricted access to the database, and explain to them this process" (p. 77).

These very brief mentions of such important aspects of contemporary biomedical research are partly justifiable by the fact that CIOMS, WMA, and ICH have other documents that treat these subjects more fully: the CIOMS *International Ethical Guidelines on Epidemiological Studies* (1991, revised in 2009), the WMA's *Declaration on Ethical Considerations Regarding Health Databases* (2002), and ICH's *Choice of Control Group and Related Issues in Clinical Trials E10* (2000). The advantage of separate documents is that the subjects can be treated more fully than they could in the DoH, CIOMS, and ICH E6 Guidelines; the disadvantage is that they are not nearly as well known and therefore less influential.

Other Guidelines

Although the four documents described above are arguably the most important international statements on research

ethics, they are by no means the only ones. The following also deserve consideration, especially in countries where they have mandatory or recommended status:

- Council of Europe: Additional Protocol to the Convention on Human Rights and Biomedicine concerning Biomedical Research (2004),
- UNESCO: Universal Declaration on Bioethics and Human Rights (2005),
- WHO: Standards and Operational Guidance for Ethics Review of Health-Related Research with Human Participants (2011),
- UNAIDS/WHO: Ethical Considerations in Biomedical HIV Prevention Trials (2007, 2012),
- European Union: Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on Clinical Trials on Medicinal Products for Human Use, and Repealing Directive 2001/20/EC.

See also: Biobanking: Ethical Issues; Bioethics in the Post-genomic Era; Bioethics: Genetics and Genomics; Bioethics; Ethical Dilemmas, Research, and Access to Treatment; Genetics and Indigenous Communities: Ethical Issues; Informed Consent of Research Subjects; Objectivity of Research: Ethical Aspects; Placebo-Controlled Trials, Ethics of; Research Ethics, Cross-Cultural Dimensions of; Research Funding: Ethical Aspects; Research Publication: Ethical Aspects; Stem Cell Research and Cloning.

Bibliography

Bankowski, Z., Levine, R.J., 1993. Ethics and Research on Human Subjects. international Guidelines. Proceedings of the XXVIth CIOMS conference, CIOMS Round Table Conference. WHO, Geneva.

Carlson, R.V., Boyd, K.M., Webb, D.J., 2004. The revision of the Declaration of Helsinki: past, present and future. British Journal of Clinical Pharmacology 57, 695–713.

Council for International Organizations of Medical Sciences, 2002. International Ethical Guidelines for Biomedical Research Involving Human Subjects. CIOMS, Geneva.

European Union: Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on Clinical Trials on Medicinal Products for Human Use, and Repealing Directive 2001/20/EC. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2014.158.01.0001.01.ENG.

International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), 1996. Guideline for Good Clinical Practice E6(R1). www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6_R1/Step4/E6_R1__Guideline.pdf.

ICH, 2000. Choice of Control Group and Related Issues in Clinical Trials E10. www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E10/Step4/E10_Guideline.pdf.

Kelman, H.C., 1977. Privacy and research with human beings. Journal of Social Issues 33, 169–199.

Levine, R.J., 1988. Ethics and Regulation of Clinical Research, second ed. Urban & Schwarzenberg, Baltimore, MD.

Levine, R.J., 1996. International codes and guidelines for research ethics: a critical appraisal. In: Vanderpool, H.Y. (Ed.), The Ethics of Research Involving Human Subjects: Facing the 21st Century. University Publishing Group, Frederick, MD, pp. 235–259.

Levine, R.J., 2000. Global issues in clinical trials. In: Osterweis, M., Holmes, D.E. (Eds.), Global Dimensions of Domestic Health Issues. Association of Academic Health Centers, Washington, DC, pp. 119–130.

Macklin, R., 1992. Universality of the Nuremberg code. In: Annas, G.J., Grodin, M.A. (Eds.), The Nazi Doctors and the Nuremberg Code: Human Rights in Human Experimentation. Oxford University Press, New York, pp. 240–276.

Macklin, R., 2004. Double Standards in Medical Research in Developing Countries. Cambridge University Press, Cambridge, UK.

Schmidt, U., Frewer, A. (Eds.), 2007. History and Theory of Human Experimentation: The Declaration of Helsinki and Modern Medical Ethics. Franz Steiner Verlag, Stuttgart.

The Nuremberg Code, 1947. www.hhs.gov/ohrp/archive/nurcode.html.

United States National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. DHEW Publication No. (OS) 78-0012, Washington, DC. www.fda.gov/ohrms/dockets/ac/05/briefing/2005-4178b_09_02_Belmont%20Report.pdf.

World Medical Association Declaration of Helsinki Ethical Principles for Medical Research Involving Human Subjects. Adopted by the 18th WMA general assembly, Helsinki, Finland, June 1964 and Amended by the 29th WMA General Assembly, Tokyo, Japan, October 1975; 35th WMA General Assembly, Venice, Italy, October 1983; 41st WMA General Assembly, Hong Kong, September 1989; 48th WMA General Assembly, Somerset West, Republic of South Africa, October 1996; 52nd WMA General Assembly, Edinburgh, UK, October 2000; 53rd WMA General Assembly, Washington, DC, October 2002 (Note of Clarification added); 55th WMA General Assembly, Tokyo, Japan, October 2004 (Note of Clarification added); 59th WMA General Assembly, Seoul, Republic of Korea, October 2008; and 64th WMA General Assembly, Fortaleza, Brazil, October 2013. www.wma.net.

Ethnic Cleansing, History of

Liliana Riga, University of Edinburgh, Edinburgh, Scotland, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term ethnic cleansing emerged during the 1990s' wars of Yugoslav secession, but it has since been used to describe and theorize a range of ethnic conflicts, including forms of genocide and mass atrocity. Drawing on classical and modern historical cases, ethnic cleansing may include colonial or frontier cleansing of indigenous populations, forcible population transfers, and mass expulsions following partitions of postwar settlements, ideological 'classicides,' and elements of genocide. Theorizations have taken a number of analytical lenses, giving rise to issues such as democratic or wartime cleansing, humanitarian interventions and postconflict reconciliation, and ethnic cleansing involving resource or environmental conflicts.

From its original instantiation in the 1990s' wars of Yugoslav secession, scholars have extended the term ethnic cleansing to descriptively cover a wide range of empirical cases and to conceptually capture a distinctive category or type of ethnic conflict and nationalist violence. After outlining legal and scholarly definitions of ethnic cleansing, this article situates understandings of ethnic cleansing as a social phenomenon in wider historical contexts. It then examines some of the most influential theoretical perspectives or conceptual frameworks used to understand ethnic cleansing, and closes by considering these approaches in light of case studies and issues recently raised as the study of ethnic cleansing takes new directions.

Definitions and Properties

First popularly used in the 1990s during the Yugoslav wars (in Serbo-Croatian, *etničko čišćenje*, or, more precisely, *čišćenje terena*, cleansing the terrain), a key characteristic feature or property of ethnic cleansing – central to both legal and scholarly definitions – is that it involves the removal or mass expulsion of peoples from a territory. Based on the Yugoslav experience, in 1994 the United Nations Commission of Experts formally defined ethnic cleansing as a "rendering [of] an area ethnically homogenous by using force or intimidation to remove from a given area persons from another ethnic or religious group"; it is a "purposeful policy designed by one ethnic or religious group to remove by violent or terror-inspiring means the civilian population of another ethnic or religious group from certain geographic areas" (UN Commission of Experts Report, 1994: 33). Although legally recognized by the UN as a 'form of genocide,' it is nevertheless distinguishable from genocide both by intentionality and by a 'killing requirement': genocide requires the *intent* to destroy a group *as such* (often implying a quantitative threshold), while ethnic cleansing does not require proof of this intent but simply the removal (not necessarily killing) of a group from a territory. In international law, therefore, unlike genocide, ethnic cleansing does not formally require international intervention (Article II of the United Nations' Convention on the Prevention and Punishment of the Crime of Genocide includes the intent 'to destroy, in whole or in part, a national, ethnical, racial or religious group, as such').

Scholarly conceptualizations have similarly focused on the removal of a people from a territory by force or intimidation as the constitutive feature of ethnic cleansing. Sociologist Michael Mann (m: 11) defines 'murderous ethnic cleansing' as "the removal of members of one such [ethnic] group of another such group from a locality they define as their own," and a nation can be such a group when it claims collective political rights to a given territory. The historian Norman Naimark (2001: 3) problematizes the distinction between genocide and ethnic cleansing, arguing that ethnic cleansers intend "to remove a people and often all traces of them from a concrete territory…to get rid of the 'alien' nationality, ethnic, or religious group and to seize control of the territory they had formerly inhabited"; but whereas it is similar at one end to forced deportations, at the other it 'bleeds into genocide.'

Likewise, Gellately and Kiernan (2003: 19–20) conceptualize ethnic cleansing as only partially distinct from either ethnic conflict or genocide: because it involves the "'purification' of a territory, not necessarily of a population," the (non) violent deportation of an ethnic group from the territory can precede and/or accompany genocide, making it a separate and distinguishable event. Political scientist Brendan O'Leary et al. (2001: 32) conceives of "ethnic expulsion as a right-peopling strategy," distinct from those strategies that move borders: it is "the intended, direct or indirect, forcible movement by state officials, or sanctioned paramilitaries, of the whole or part of a community from its current homeland, usually beyond the sovereign border of a state." Geographers Toal and Dahlman (2011: 5) define it more sharply as a territorializing military tactic, "as much about seizing and consolidating territory as it is about identity…[it] involves the ethnicization of space," whereby a "landscape is wiped clean and available for re-inscription as an ethnically homogeneous homeland."

Importantly, then, ethnic cleansing has been most commonly placed on both empirical and conceptual continua of practices of organized or unsystematic violence. Although there is no precise international crime under the term 'ethnic cleansing,' it is deemed 'contrary to international law,' and its practices or instrumentalities – forcible deportation, torture, arbitrary detention, rape and sexual assault, or extra-judicial executions – are considered war crimes, crimes against humanity, or otherwise fall under the terms of the Genocide Convention (UN Commission of Experts Report, 1994: Part III,

Section B). Moreover, these practices have been extended beyond simply the physical removal of a people, to also include the removal or erasure of evidence of cultural identities and practices. So ethnic cleansing commonly involves the destruction or desecration of religious symbols and buildings, cultural artifacts, and other traces of a groups' cultural existence. While ethnic cleansing can legally blur into – or become a subset of – genocide, for most scholars and legal experts it still retains valuable empirical elasticity and theoretical utility for analyzing a range of historical and contemporary practices that do not, by themselves, constitute genocide (see Blecher, 2005; Hayden, 1996) (*see* Ethnic Conflicts).

Ethnic Cleansing in Historical Contexts

Since the early 1990s, therefore, scholarly work across the social sciences has sought to stretch ethnic cleansing's empirical elasticity by extending its conceptual insights to a wide range of practices, and to an even wider variety of historical and contemporary contexts. The broadest and most influential organizing historical framework for understanding ethnic cleansing problematizes its essential 'modernity.' A number of scholars view ethnic cleansing as an empirically modern phenomenon entwined with key sociopolitical and ideological processes associated with modernity, such as the construction of citizenship, nationalism, democratization, or the institutional rise of the nation state (see, among others, Mann 2005; O'Leary et al., 2001; Naimark, 2001; Bauman, 1989). Yet other scholars locate instances of ethnic cleansing in antiquity or in pre-modern contexts, especially the ancient empires (see, for instance, Kiernan, 2007; Bell-Fialkoff, 1996).

These latter scholars argue that the practices of removing, killing, or expelling peoples in premodern historical contexts invoked religious, political, or other ideologies, and that they effectively possessed the essential features of modern ethnic cleansing, albeit recognizing that its force and systematic intensity have increased with the political capacities of the modern state. The most commonly cited examples of the cleansing of 'undesirable populations' are the atrocities, displacements, and massacres that characterized Babylonian, Greek, and Assyrian societies (compare dissenting discussions, Kiernan, 2007: Ch. 1; Mann, 2005: Ch. 2). Kiernan (2007), for instance, argues that 'cults of antiquity' and animosities toward non-agriculturalists characterized classical massacres; and Bell-Fialkoff (1996) suggests that over the course of 3000 years, cleansing was variously used to maintain control over populations, including religious minorities in early modern history. Moreover, Goody (2002) argues that religion was often the basis of cultural identifications and ascriptions in the early stages of the formation of modern nation-states, underscoring its centrality to 'cleansing.'

Ethnic cleansing 'modernists,' by contrast, argue that the core constitutive property of ethnic cleansing is its drive to homogenize territories, something historically located in the modern nationalist political claim to be ruled by co-ethnics, in a modern notion of 'we the people,' and in the ideologies and instruments associated with the rise of the modern state (*see* War and Nationalism; War and Democracy; Nation-State and War). While these scholars work from different epistemological perspectives, a belief in ethnic cleansing's essentially modern imperative is shared: Mann (2005), for instance, argues that it is the 'dark side of democracy,' Hinton (2002: Ch. 1) that it is the 'dark side of modernity,' and for Bauman (1989), the Holocaust itself was constitutively implicated in modernity. On this view, the cluster of underlying sociopolitical forces relevant to understanding ethnic cleansing – most especially democracy, welfare states, and citizenship – are organized around ethnic or national criteria of inclusion and exclusion, so ethnic cleansing is characterized by its political exclusivism.

The distinction within this organizing historical framework between 'modernists' and 'pre-modernists,' therefore, often turns less on evaluations of ethnic cleansing's practices, tactics, or instrumentalities than it does on assessments of the modernity of its ideological justifications and commitments – including whether or not group ascriptions are viewed as 'primordial' or as cultural constructs of political modernity.

Yet whether seen as premodern or modern, a loose typology of historical cases of ethnic cleansing has emerged in the scholarly literature. One such set of cases involves 'frontier ethnic cleansing,' or those colonial policies and settler practices that have historically targeted indigenous populations (cf Kiernan, 2007: Part II; Mann, 2005: Ch. 4; Anderson, 2005; Hinton, 2002: Part I). Tsarist Russia's settler policies between 1855 and 1866 toward groups in its Caucasian borderlands involved the colonial cleansing of between 500 000–900 000 Muslims, who experienced pressured and coerced emigration as a result of Tsarist land, settlement, and agricultural policies (Breyfogle, 2003). Similarly, US President Andrew Jackson's Indian Removal Act of 1830 and the forcible relocation and removal of Cherokee, Choctaw, Seminole, and other Native American nations, known as the Trail of Tears, has been considered a form of ethnic cleansing (Carson, 2008; Kiernan, 2007: Ch. 8). The related conquest of Texas in the nineteenth century involved a kind of ethnic frontier warfare, whereby settlers challenged native populations for hunting lands and ultimately 'cleansed' them from their territories (Anderson, 2005). German colonists' killing of an estimated 80% of Namibia's Herero population in German South West Africa between 1904 and 1908 (by some accounts prefiguring Nazi ideological commitments) is also considered frontier or colonial ethnic cleansing (Sarkin, 2011; Madley, 2004). These historical cases of ethnic cleansing – sometimes blurring into the genocide or the 'ethnocide' of particular indigenous populations – are constituted less by specific practices or tactics of 'cleansing,' and more by its ideological or colonialist underpinnings, whether civilizational, expansionary, or settler.

A related set of cases links postcolonial instances of ethnic cleansing to specific colonial identity regimes or to the aftereffects of colonial practices of local rule and social control. Mamdani's (2001) study of Rwanda, for example, argues that the logic of the 1994 killings in Rwanda (and earlier cycles of Hutu/Tutsi killings and reprisals) can only be understood against the historical context of Belgian and British colonial racializations of social ascriptions and hierarchies. And although the character and impact of the Belgian colonial state's identity regimes in neighboring Burundi differed from those of Rwanda, they similarly reshaped Tutsi and Hutu relations administratively, ascriptively, and in terms of landholding (Uvin, 2002; Lemarchand, 1994: Ch. 3).

Another set of historical cases that have been characterized as ethnic cleansing are those that involved mass expulsion or forcible population transfers, most often associated with the dissolution of empires and the creation of nation-states in East Central Europe and the 'demographic engineering' that these processes entailed. Ther and Siljak (2001: 1–2), for instance, conceptualize this form of ethnic cleansing as sanctioned or nonsanctioned "repressive actions" that involve the "involuntary transfer of entire ethnic groups from one country to another," especially deportations (the coercive transfer of minorities to other regions within a country) and expulsions (the forced removal of minorities from a country). To make the newly drawn political boundaries coincide with ethnolinguistic boundaries, between 1944 and 1948, 31 million people – Germans, Romanians, Bulgarians, Poles, Ukrainians, Czechs, Slovaks, Greeks, Serbs, Croats, and Macedonians among others – were forcibly transferred or expelled in internationally sanctioned treaties or by nonsanctioned resettlement policies (cf Hayden, 1996; Ther and Siljak, 2001; Várdy and Tooley, 2003: Part II; Snyder, 2003). In this context, the forcible resettlement of ethnic Germans from Czechoslovakia is sometimes referred to as 'liberal ethnic cleansing' because it was sanctioned by treaty at Potsdam in 1945.

But these population transfers had been preceded by the purposeful expulsion of other minorities in East Central Europe, Russia, the Balkans, and the territories of the Ottoman Empire, during and after the Balkan Wars of 1912–13 and during and after WWI, to consolidate demographic claims to territories: two million people were forcibly expelled or deported by the 1923 Treaty of Lausanne in sanctioned population exchanges that redefined the borders of the states of Turkey, Greece, and Bulgaria that emerged from the Ottoman Empire (see the contributions in the 2008 Special Issue of the *Journal of Genocide Research* 10(1)). These internationally sanctioned population exchanges formalized expulsions and sanctioned certain property claims. In Tsarist Russia during WWI, the army forcibly deported approximately 800 000 Jewish and German minorities away from the war front in what were considered 'ethnic expulsions' (Lohr, 2003). Most famously, in 1915, millions died as the Ottoman Empire deported its entire Armenian population from the northeast frontier in what moved beyond ethnic cleansing to genocide.

This set of historical cases in particular highlights certain definitional issues: just as the continuum between ethnic violence at one end and genocide at the other is typically taken to contain ethnic cleansing within it, so too the issue arises as to when fear- or intimidation-induced emigration becomes forced migration or coercive displacement and transfer, and, relatedly, whether legally sanctioned population transfers or exchanges should be similarly considered forced relocation or 'repatriation.' The consensus on this has shifted, as the practices of ethnic cleansing have been enveloped within a human rights framework in the post-WWII period (see discussion below). So while in the early- and mid-twentieth century, international efforts considered legally sanctioned population movements to be viable alternatives to minority rights or protections, today they are generally viewed as unacceptable 'liberal' forms of cleansing.

A related set of historical cases further underscores the difficulty in capturing what, precisely, is considered ethnic cleansing. Some scholars view violent state-directed or spontaneous large-scale population movements and massacres occasioned by state partitions to be instances of ethnic cleansing. The partition of Pakistan and India in 1947, for instance, and the resulting large-scale massacres that occurred in the Punjab were, like those in Rwanda, cycles of retribution – not state sponsored, but not unorganized either: between 200 000–360 000 Hindus and Muslims were killed as up to 12 million crossed the new borders of the partition (Brass, 2003; Hansen, 2003).

Finally, a distinct kind of ethnic cleansing involves what Mann (2005) terms 'classicide,' or mass expulsions, killings, and deportations whose targets are class or political enemies, and which may or may not be culturally marked. Such cases include the Soviet era deportations of Chechen-Ingush and Crimean Tatars (Naimark, 2001: Ch. 3), the class enemy-based murderous ethnic cleansing in Cambodia (see Mann's 2005: Ch. 11 account), or Stalin era ethnic deportations (Martin, 1998). In the context of Soviet ethnic deportations (including ethnically based agricultural resettlement), Martin (1998), for instance, argues that ethnic cleansing occupies a middle position between genocide and nonviolent pressured ethnic emigration – in this case occasioned not by Russian nationalism, but by Soviet xenophobia. These cases of classicide also blur into a set of cases considered 'politicide': recent Ethiopian law has defined the forcible removal of political groups in the country as ethnic cleansing, and the extension of this kind of legislation to other countries is expected (Tiba, 2007). Naimark (2011), too, argues that the politically defined killing and forcible resettlement of kulaks and Ukrainian peasants under Stalin also constituted forms of ethnic cleansing/genocide; and Gratz (2011) argues that the specific targeting of political, religious, and cultural elites in the violence in Bosnia similarly constituted ethnic cleansing as 'eliticide.'

Theorizing Ethnic Cleansing

These various historicizations have given rise to several overlapping explanatory or analytical frameworks for theorizing ethnic cleansing. Individual scholars often draw on one or more of these theoretical frameworks. The most common analytical framework examines ethnic cleansing through ethnic, nationalist, cultural, or other identity lenses. Seen as one potential result of the political claim to be ruled by co-ethnics, for instance, Mann's (2005) influential explanatory framework situates murderous ethnic cleansing along a continuum of violent or nonviolent practices or policies; these policies are also arrayed along a separate axis that involves the elimination, accommodation, or containment of cultural difference. While sometimes treating the relevant social groups as essentialized or reified categories, Mann's case studies (Yugoslavia, Armenia, the Nazi Holocaust, Cambodia, Rwanda) suggest that the intent to cleanse, and the degrees of violence associated with it, are variously dependent on state factionalization (often precipitated by war), ethnonationalist 'core constituencies,' and the intersectionality of ethnic differences with other social distinctions, most notably class or region. His analytical schema ranges, therefore, from nonviolent multiculturalism, to pogroms, to violent displacement, to genocide (subsuming ethnic cleansing). So 'murderous ethnic

cleansing' can involve not only forced displacement, but also biological sterilization and genocide. O'Leary's et al. (2001) explanatory framework likewise situates ethnic cleansing (e.g., expulsion and genocide) as a strategy for *eliminating difference*, distinguishing it from those strategies used for *managing difference* (e.g., arbitration or consociation). Hinton (2002) conceptualizes ethnic cleansing (and genocide) as a process that essentializes identity differences – a defining dimension of modernity – and then seeks to erase these differences through a homogenizing nationalism.

A key feature of these theoretical contributions is that ethnic cleansing is moored around nationalism or culture, and the key issue is how and why cultural differences are eliminated, controlled, or displaced from a given territory. Political frameworks – nation-state formation, the dissolutions of empires, democratization – are important processes whereby the cultural and the political or territorial are made congruent. So, for instance, modern ethnic cleansing is theorized as a form of organic or integral nationalism, which fuses the *demos* of democracy with the *ethnos* of nationalist homogeneity – and it centrally does so to control a given territory (Mann, 2005: 3–4). The idea of 'unmixing peoples,' and its associated forms of population transfer, similarly conceives of ethnic cleansing mainly through a nationalist or cultural lens (Blecher, 2005; Brubaker, 1995). Snyder's (2003: 197) conceptualization of Polish–Ukrainian ethnic cleansing in 1943, for example, as the "violent removal of populations in the pursuit of nationally pure space," conceptualizes the essence of ethnic cleansing as being primarily about cultural difference. These nationalism-based analytical frameworks, then, conceptualize ethnic cleansing as a cultural or identity problem of politics.

A second, related analytical axis for understanding ethnic cleansing focuses attention more specifically on distinctions in ethnic cleansing's motivating or organizing ideologies. The principle beliefs, rationales, and justifications given for the removal or killing of an outgroup can range from nationalist, religious, or sectarian beliefs, to racialist, secular political, and imperial or civilizational ideologies (cf Snyder, 2012; Kiernan, 2007; Straus, 2006; Mann, 2005; Lohr, 2003; Weitz, 2003). Viewed through the prism of its ideological commitments, these scholars explore the symbolic, political, or discursive language used, the intellectual origins of ideological underpinnings, the social carriers of those beliefs, and the self-ascribed beliefs of those involved in implementing or carrying out ethnic cleansing. In short, this theoretical framework seeks to retrieve the ideological or ideational underpinnings of those states, groups, and constituencies possessing, or involved in producing, organized exclusivist ideologies, and in particular it highlights the key social and political dynamics associated with 'othering': stigmatization, dehumanization, and racialization (Hagan and Rymond-Richmond, 2008).

A third explanatory framework for understanding ethnic cleansing seeks to better understand its perpetrators, its victims, and especially the critical role of the state and its various agents as instruments of ethnic cleansing. Here, careful consideration is given to those involved in the actual practices of ethnic cleansing: who are its proponents, instigators, or perpetrators? How involved are 'ordinary' citizens? How involved are organized societal groups such as paramilitaries, political elites, police, or armed gangs? Importantly, attention is given to explaining both individual-level motivations and the large-scale involvement of citizens (cf Straus, 2012; Mann, 2005; Mamdani, 2001; Staub, 1989). Scholars working within this broad conceptual framework examine not only the dynamics of reluctant or purposeful masses in the participation of ethnic cleansing, and in particular the use or deployment of norms of coercion (e.g., Bhavnani, 2006), but they also examine why some refuse to participate in ethnic cleansing or genocide (e.g., Dumitru and Johnson, 2011). Scholars also focus on explaining the violent dynamics and escalatory scale of ethnic cleansing, and the mediating macrolevel forces and microlevel motivations of individuals and midlevel 'constituencies' (Mann, 2005); on examining the relationships between leaders and followers, bystanders and aggressors, their social psychologies, the microdynamics of identity self-ascriptions and moral choices in the violent contexts of ethnic cleansing (Monroe, 2012; Snyder, 2010; Midlarsky, 2005); and on intentionality, and the relationship between beliefs and behaviors, particularly in contexts where killings are individualized, personal, or intimate (Straus, 2012, 2006; Verwimp, 2005; Waller, 2002). Straus' (2006) wider account of the ethnic cleansing/genocide in Rwanda, for example, focuses on the local dynamics of the killings in Rwanda by relying on in-depth interviews with imprisoned *génocidaires*; Dumitru and Johnson (2011) link macrotheories of ethnic war to microlevel explanations of ethnic killing by examining the underlying qualities of the ethnic relationships that deeply contextualize the ethnic cleansing in the first instance; whereas Uvin (1998) anchored an analysis around the salience and ubiquity of underlying structural violence in Rwandan society, something that generated a pervasive sense of fear, humiliation, and frustration – anticipating violence.

A fourth common organizing framework for theorizing or understanding ethnic cleansing is that of human rights, war crimes, and international law. Here scholars explore the entwined relationships among ethnic cleansing, genocide, and conceptions of human rights violations, crimes against humanity, and mass atrocity (Hagan and Rymond-Richmand, 2008; Hagan et al., 2006; Gellately and Kiernan, 2003; Hinton, 2002). Viewed within a human rights or international legal perspective, the empirical focus is often on assessing evidence or determining a set of practices that constitute breaches of human rights. The practices of ethnic cleansing, in other words, are the central focus, and they are read through the conceptual lens of international human rights law. So, for instance, unlawful detention, extrajudicial killings, rape and sexual assault, or targeted destruction of property are legal categories through which the tactics used for the mass rendering of an area ethnically homogenous, or to effect population displacement by intimidation or force, may involve laying siege to cities and indiscriminately shelling civilian populations, executing non-combatants, the use of concentration or detention camps of torture, execution, rape, and inhumane treatment, and razing villages to intimidate or uproot populations. This empirical overlap between the tactics or practices involved in genocide, ethnic cleansing, and war crimes or mass atrocity has also led to a re-framing of ethnic cleansing as a particular subset of human rights violations, that is, ethnic cleansing as a type of 'hard' human rights violation (Hagan and Rymond-Richmond, 2008; Hagan et al., 2006).

Comparisons, Cases, and Issues

While there are few analytically synthetic works on ethnic cleansing (cf Mann, 2005; Weitz, 2003), there are a number of comparative case studies organized around cases selected for historical, thematic, or geographical reasons (cf Snyder, 2012; Naimark, 2011, 2001; Carmichael, 2009; Gellately and Kiernan, 2003; Hinton, 2002). There are many more individual case studies that focus on either bounded instances of ethnic cleansing, or on mass killing and displacement amounting to some combination of ethnic cleansing and genocide. The most commonly explored cases are the violent dissolution of the former Yugoslavia, the Nazi Holocaust, Rwanda, East Timor, Cambodia, Guatemala, and Darfur among others (*see* Holocaust, The). Scholarly works on these most notable cases usually involve questions of how and why the ethnic violence occurred, what could have been done to prevent it, and what the policy 'lessons learned' might be.

But ethnic cleansing as both an empirical descriptor and as a conceptual construct is evolving, and more recent developments around the world are increasingly coming under scrutiny as instances of ethnic cleansing, in addition to revisiting historical instances of ethnic violence through more contemporary lenses, for instance. As a result, a number of analytical issues are emerging as suggestive not only of the development of new typologies of ethnic cleansing, but also of the directions that the history of ethnic cleansing is beginning to take.

1. The most cited feature of ethnic cleansing in the many individual case and comparative studies is its common occurrence under cover of war, civil conflict, or in the context of violent state partitions or secessions (*see* Civil Wars). In addition to some of the cases noted earlier, this includes recent studies of cycles of sectarian cleansings between Shi'a and Sunni in Iraq (Kenyon Lischer, 2008); the forced evictions of Tamil minorities in northwest Sri Lanka in 2008–09 following the end of a civil war; the ethnic cleansing characteristic of the conflicts in Africa's Great Lakes (Lemarchand, 2009); and the continuing coerced displacement of the Nuba in South Kordofan in southern Sudan in the wake of the secession of Southern Sudan in 2011. This is not new, of course; Mann (2005), Shaw (2007, 2003), Naimark (2001), Lohr (2003), Straus (2006), and Snyder (2012) among others have theorized along these lines with respect to earlier historical cases. However, as the empirical relationships among war, civil conflict, and ethnic cleansing are better understood, ethnic cleansing is increasingly theorized with less abstraction and greater local, empirical contextualization, embedding moments of ethnic cleansing within local politics – and pointing the way to the emergence of new empirical typologies around this type of ethnic cleansing (cf Kasara, 2010; de Waal, 2005a).
2. A related general consensus is that ethnic cleansing is often associated with processes of democratization: democratic legitimacy is used as a 'we the people,' exclusivist ideology (Mann, 2005); or ethnic elites play 'ethnic cards' for purposes of social mobilization, or more specifically because of the political stakes involved in the ethnic gerrymandering of electoral districts (e.g., Kasara, 2010; Malik, 2000). Kasara (2010), for instance, shows how arson attacks in particular neighborhoods in Nairobi and other cities in Kenya around the country's 2007 elections were forms of targeted ethnic cleansing, designed to segregate neighborhoods for electoral purposes. Similarly, the recent ethnic displacement and killing of Muslim Rohingyas in Buddhist majority Myanmar's current fragile and tenuous democratic liberalization involves pogroms and the burning of homes and villages along its western coast.
3. A third issue is the pervasiveness of rape and sexual assault in contexts of ethnic cleansing (Meger, 2011; Allen, 1996). Either spontaneous and unsystematic, or organized and systematic, rape, and in some cases forced sterilization, are usually theorized in the context of war more generally, or by conceptualizing women as the biological carriers of 'the nation,' and rape and sexual assault as the violent inscription of cultural identities on women's bodies (Mostov, 2000). A related issue receiving more recent attention is that of children born of ethnic cleansing rape and its related postethnic cleansing traumas (Carpenter, 2010) (*see* War, Political Violence, and Effects on Children).
4. Scholarly work on ethnic cleansing is also seeking more explicit policy prescription: how ethnic cleansing can be averted, or how and what kind of interventions can prevent escalation from lower levels of ethnic violence (Robinson, 2010; Mann, 2005: Ch. 17; O'Leary et al., 2005). Indeed, the 2005 UN World Summit Document (A/RES/60/1, 2009 UN Report) regarding the Responsibility to Protect (R2P) outlines states' responsibilities (and those of the international community) to protect populations against genocide, war crimes, ethnic cleansing, and their incitement. So Robinson (2010), for instance, shows how international armed intervention averted ethnic cleansing in East Timor in 1999; and O'Leary et al. (2005) argue that various forms of power-sharing or consociational arrangements can mitigate both the causes and effects of sectarian cleansing (*see* Conflict Mediation).
5. Another key issue highlighted in recent cases of ethnic cleansing has been its link to underlying resource and environmental conflicts – both those contexts characterized as resource-rich with tensions around the politicized extraction of resources, and those contexts that are resource-scarce (see especially the 2012 Special Issue of the *Journal of Peace Research* 49 (1)). Understanding ethnic cleansing in these contexts involves, for instance, exploring the roles of 'moral geographies' and the ethnically based claims to land of indigenous cultures. This is most notably examined in those conflicts along the African Sahel Belt, where resource-based ethnic cleansing occurs in contexts of stark poverty, fragile livelihoods, and vulnerable environments as it combines with ethnically marked access or control of social and political resources, and weak or infrastructurally fragile states. These have been highlighted as key factors underlying the ethnic cleansing Darfur and in the eastern Horn of Africa, for instance (de Waal, 2005b; Tubiana, 2007; UNEP, 2007). Locating an important element of Darfur's ethnic cleansing in the ways in which desertification and drought in Western Sudan placed severe environmental pressures on both pastoral and agricultural

livelihoods among Arab and non-Arab Darfuris, scholars point to the subsequent politicization of these underlying tensions, and link local and national politics to the independent effects of resource scarcity.

6. A sixth emergent issue in studies of ethnic cleansing may be broadly described as a cluster of related postethnic cleansing issues: reconstruction, return, and transitional, restorative, or postconflict justice and reconciliation. The practices of transitional and restorative justice, international and regional tribunals for reconciliation and accountability, and attention to the roles of memory and postconflict history teaching have all gained greater attention as focus turns to *postethnic cleansing* needs and realities (Straus and Waldorf, 2011; Hinton and O'Neill, 2009; Staub, 2006; Snyder and Vinjamuri, 2003/2004; Hinton, 2002: Part 4; Borneman, 2002; Minow, 1998). These include better understandings of ethnic displacement and refugee crises (cf Zetter, 2007; Peteet, 2005) and of postconflict environmental issues (e.g., UNEP, 2007). Moreover, following the Yugoslav conflicts, a 2005 UN resolution was adopted to accept the 'Pinhiero' Principles on Housing and Property Restitution for Refugees and Displaced Persons (outlined in UN Pinhiero Report, 2002–2005). It defined the rights and obligations of states toward those displaced by conflict, and specified that refugees and displaced populations, many occasioned by ethnic cleansing, have the specific rights to return to their pre-war homes (not just to their pre-war countries) or they are otherwise entitled to restitution of their homes and properties. This is also seen as a particularly important element in restorative justice and in building sustainable peace (Leckie and Huggins, 2011), and it has given rise to greater attention being paid to the modalities and possibilities for the reversal of ethnic cleansing, most exhaustively examined in the context of Bosnia-Herzegovina (Toal and Dahlman, 2011; Williams, 2005).
7. Another direction taken to better understand the modalities of ethnic cleansing is methodological: the methodologies involved in identifying ethnic cleansing (and genocide), measuring its social and environmental effects, and finding more sophisticated ways of assessing or quantifying its statistical parameters (deaths, disappearance, or displacement figures, for instance) are being extended to include the use of new technologies. In particular, this involves a greater diffusion of the use of satellite imagery/GIS to examine levels of housing and building destruction; overturned earth through infrared imagery to assess the existence of mass graves; and spatial geomodeling to determine incidents and locations of ethnic violence-related events, such as arson-induced displacement, road destruction to prevent population returns, burned crop fields, and the construction of detention centers or tent camps of the displaced (Hagan et al., 2006; Tabeau and Bijak, 2005). A contextualization of ethnic cleansing within a war crimes framework, combined with the adoption of these methodologies, has prompted sociologists, for instance, to critically re-assess ethnic cleansing as a 'complex humanitarian emergency'; this requires a better grasp of disease, health, and environmental effects, and innovative analytic techniques to better calculate mortality and displacement rates to take the full measure of the violence involved (e.g., Hagan et al., 2006). This scholarly work at once parallels and complements that recently undertaken by a number of human rights organizations and (inter) governmental agencies, such as the Harvard Humanitarian Initiative and the United Nation's 'Operation Satellite and Applications Program' (cf http://www.unitar.org/unosat/; http://hhi.harvard.edu/). These efforts embed the contemporary unfolding of ethnic cleansing (and genocide) into environmental and other humanitarian emergencies frameworks and link ethnic cleansing's methods and tactics to wider projects involving 'crisis mapping' through satellite imagery analysis (*see* Geographic Information Systems and Remote Sensing).
8. Finally, an important dimension in understanding the dynamics of ethnic cleansing has recently involved embedding it in geopolitical neighborhoods, border regions, and regional conflicts, with particular attention paid to the migration of the conflict through the large number of either cross-border refugees or internally displaced persons. This has given rise to a more sustained consideration of the workings of the short- and long-term volatile dynamics of displaced minorities in geopolitically vulnerable conflict neighborhoods and to how cycles of ethnic cleansing can occur. The displacement of Tutsis and Hutus across Rwanda, Burundi, and the Democratic Republic of Congo (Lemarchand, 2009; Prunier, 2008), that of Darfuris in Sudan and Chad (Behrends, 2007), and that of Ethiopians in Somalia, Eritrea, and Eastern Africa (see brief account by Calhoun, 1999) have all prompted greater theoretical sensitivity to how the ethnic cleansing of minority populations – and especially to their lives in resettlement or refugee camps – can carry conflicts across borders, and entwine them with those of other, equally vulnerable political contexts or disaffected populations. Indeed, the politicizing effects of displacement are particularly acute where displaced populations live in large refugee settlements that have a kind of semipermanent status, for example, the new identities and politics emerging in the camps of Somali refugees in Kenya or Darfuri refugees in Chad.

See also: Civil Wars; Conflict Mediation; Ethnic Conflict, Geography of; Ethnic Conflicts; Genocide and War; Holocaust, The; Military, War, and Politics; Nation-State and War; Remote Sensing; Territorial Conflicts; War and Democracy; War and Nationalism; War, Sociology of; Warfare in History.

Bibliography

Allen, Beverly, 1996. Rape Warfare: The Hidden Genocide in Bosnia-Herzegovina and Croatia. University of Minnesota Press.

Anderson, G.C., 2005. The Conquest of Texas: Ethnic Cleansing in the Promised Land. University of Oklahoma Press, 1820–1875.

Bauman, Zygmunt, 1989. Modernity and the Holocaust. Cornell University Press.

Bell-Fialkoff, Andrew, 1996. Ethnic Cleansing. St. Martin's Press.

Behrends, Andrea, 2007. The Darfur conflict and the Chad/Sudan border: regional context and local reconfigurations. Sociologus 57 (1), 99–131.

Bhavnani, Ravia, 2006. Ethnic norms and interethnic violence: accounting for mass participation in the Rwandan genocide. Journal of Peace Research 43 (6), 651–669.

Blecher, Robert, 2005. Citizens without sovereignty: transfer and ethnic cleansing in Israel. Comparative Studies in Society and History 47 (4), 725–754.
Borneman, John, 2002. Reconciliation after ethnic cleansing: listening, retribution, affiliation. Public Culture 14 (2), 281–304.
Brass, Paul, 2003. The partition of India and retributive genocide in the Punjab, 1946–47: means, methods, and purposes. Journal of Genocide Research 5 (1), 71–101.
Breyfogle, Nicholas, 2003. Heretics and Colonizers: Forging Russia's Empire in the South Caucasus. Cornell University Press.
Brubaker, Rogers, 1995. Aftermaths of empire and the unmixing of peoples: historical and comparative perspectives. Ethnic and Racial Studies 18 (2), 189–218.
Calhoun, Craig, 1999. Ethiopia's ethnic cleansing. Dissent 47–50.
Carmichael, Cathie, 2009. Genocide before the Holocaust. Yale University Press.
Carpenter, R. Charli, 2010. Forgetting Children Born of War: Setting the Human Rights Agenda in Bosnia and beyond. Columbia University Press.
Carson, James Taylor, 2008. Obituary of nations': ethnic cleansing, memory, and the origins of the old South. Southern Cultures 14 (4), 6–31.
Dumitru, Diana, Johnson, Carter, 2011. Constructing interethnic conflict and cooperation: why some people harmed Jews and others helped them during the holocaust in Romania. World Politics 63 (1), 1–42.
Gellately, Kiernan, 2003. The study of mass murder and genocide. In: Gellately, Kiernan (Eds.), The Spector of Genocide: Mass Murder in Historical Perspective. Cambridge University Press.
Goody, Jack, 2002. Bitter icons and ethnic cleansing. History and Anthropology 13 (1), 1–12.
Gratz, Dennis, 2011. Eliticide in Bosnia and Herzegovina and its impact on the contemporary understanding of the crime of genocide. Nationalities Papers 39 (3), 409–424.
Hagan, John, Rymond-Richmond, Wenona, 2008. The collective dynamics of racial dehumanization and genocidal victimization in Darfur. American Sociological Review 73, 875–902.
Hagan, John, Schoenfeld, Heather, Palloni, Alberto, 2006. The science of human rights, war crimes, and humanitarian emergencies. Annual Review of Sociology 32, 329–349.
Hansen, A.B., 2003. The Punjab 1937–47–a case of genocide? In: Jensen, S.L.B. (Ed.), Genocide: Cases, Comparisons and Contemporary Debates. Danish Centre for Holocaust and Genocide Studies, pp. 22–83.
Hayden, Robert, 1996. Schindler's fate: genocide, ethnic cleansing, and population transfers. Slavic Review 55, 727–748.
Hinton, Alexander Laban (Ed.), 2002. Annihilating Difference: The Anthropology of Genocide. University of California Press.
Hinton, Alexander Laban, O'Neill, Kevin, 2009. Genocide: Truth, Memory and Representation. Duke University Press.
Journal of Genocide Research, 2008. Special Issue: Late Ottoman Genocides 10 (1).
Journal of Peace Research, 2012. Special Issue: Climate Change and Conflict 49 (1).
Kasara, Kimuli, 2010. Local segregation and inter-ethnic violence. In: Paper Presented to the American Political Science Association, September 2–5, 2010.
Kiernan, Ben, 2007. Blood and Soil: A World History of Genocide and Extermination from Sparta to Darfur. Yale University Press.
Kenyon Lischer, Sarah, 2008. Security and displacement in Iraq: responding to a forced migration crisis. International Security 33 (2), 95–119.
Leckie, Scott, Huggins, Chriss (Eds.), 2011. Conflict and Housing, Land, and Property Rights: A Handbook on Issues, Frameworks, and Solutions. Cambridge University Press.
Lemarchand, René, 1994. Burundi: Ethnic Conflict and Genocide. Cambridge University Press.
Lemarchand, René, 2009. The Dynamics of Violence in Central Africa. University of Pennsylvania.
Lohr, Eric, 2003. Nationalizing the Russian Empire: The Campaigns against Enemy Aliens during World War I. Harvard University Press.
Madley, Benjamin, 2004. Patterns of frontier genocide: 1803–1910: the aboriginal Tasmanians, the Yuki of California, and the Herero of Namibia. Journal of Genocide Research 6 (2), 167–192.
Malik, John, 2000. The Dayton agreement and elections in Bosnia: entrenching ethnic cleansing through democracy. Stanford Journal of International Law 36, 303–355.
Mamdani, Mahmood, 2001. When Victims Become Killers: Colonialism, Nativism, and the Genocide in Rwanda. Princeton University Press.
Mann, Michael, 2005. The Dark Side of Democracy: Explaining Ethnic Cleansing. Cambridge University Press.
Martin, Terry, 1998. The origins of Soviet ethnic cleansing. The Journal of Modern History 70 (4), 813–861.
Meger, Sarah, 2011. Rape in contemporary warfare: the role of globalization in wartime sexual violence. African Conflict and Peacebuilding Review 1 (1), 100–132.
Midlarsky, Manus, 2005. The Killing Trap: Genocide in the Twentieth Century. Cambridge University Press.
Minow, Martha, 1998. Between Vengeance and Forgiveness: Facing History after Genocide and Mass Violence. Beacon Press.
Monroe, Kristen, 2012. Ethics in an Age of Terror and Genocide: Identity and Moral Choice. Princeton University Press.
Mostov, Julie, 2000. Sexing the nation/desexing the body: the politics of national identity in the former Yugoslavia. In: Mayer, T. (Ed.), Gender Ironies of Nationalism. Routledge, pp. 89–112.
Naimark, Norman, 2011. Stalin's Genocides. Princeton University Press.
Naimark, Norman, 2001. Fires of Hatred: Ethnic Cleansing in Twentieth Century Europe. Harvard University Press.
O'Leary, Brendan, McGarry, John, Salih, Khaled (Eds.), 2005. The Future of Kurdistan in Iraq. University of Pennsylvania.
O'Leary, Brendon, Lustick, Ian, Callaghy, Thomas, 2001. Right-sizing the State: The Politics of Moving Borders. Oxford University Press.
Peteet, Julie, 2005. Landscapes of Hope and Despair: Palestinian Refugee Camps. University of Pennsylvania Press.
Prunier, Gérard, 2008. Africa's World War: Congo, the Rwandan Genocide, and the Making of a Continental Catastrophe. Oxford University Press.
Robinson, Geoffrey, 2010. If You Leave Us Here, We Will Die": How Genocide Was Stopped in East Timor. Princeton University Press.
Sarkin, Jeremy, 2011. Germany's Genocide of the Herero: Kaiser Wilhelm II, His General, His Settlers, His Soldiers. James Curry.
Shaw, Martin, 2007. What Is Genocide. Polity.
Shaw, Martin, 2003. War and Genocide: Organized Killing in Modern Society. Polity.
Snyder, Jack, Vinjamuri, Leslie, 2003/2004. Trials and errors: principle and pragmatism in strategies of international justice. International Security 28 (3), 5–44.
Snyder, Timothy, 2012. Bloodlands: Europe between Hitler and Stalin. Basic Books.
Snyder, Timothy, 2003. The causes of Ukrainian–Polish ethnic cleansing, 1943. Past and Present 179, 197–234.
Staub, Ervin, 2006. Reconciliation after genocide, mass killing or intractable conflict: understanding the roots of violence, psychological recovery, and steps toward a general theory. Political Psychology 27 (6), 867–894.
Staub, Ervin, 1989. The Roots of Evil: The Origins of Genocide and Other Group Violence. Cambridge University Press.
Straus, Scott, 2012. Retreating from the brink: theorizing mass violence and the dynamics of restraint. Perspective on Politics 10 (2), 343–362.
Straus, Scott, Waldorf, Lars, 2011. Remaking Rwanda: State-building and Human Rights after Mass Violence. University of Wisconsin Press.
Straus, Scott, 2006. The Order of Genocide: Race, Power, and War in Rwanda. Cornell University Press.
Tabeau, Ewa, Bijak, Jabuk, 2005. War-related deaths in the 1992–1995 armed conflicts in Bosnia and Herzegovina: a critique of previous estimates and recent results. European Journal of Population 21, 187–215.
Tiba, Firew Kebede, 2007. The Mengistu genocide trial in Ethiopia. Journal of International Criminal Justice 5 (2), 513–528.
Ther, Philip, Siljak, Ana, 2001. Redrawing Nations: Ethnic Cleansing in East-Central Europe, 1944–1948. Rowan and Littlefield.
Toal, Gerard, Dahlman, Carl, 2011. Bosnia Remade: Ethnic Cleansing and its Reversal. Oxford University Press.
Tubiana, J., 2007. Darfur: a war for land? In: Waal, Alex de (Ed.), War in Darfur and the Search for Peace. Global Equity Initiative, Harvard University, (Chapter 3).
United Nations Environment Programme, 2007. Sudan: Post-Conflict Environmental Assessment. UNEP, Geneva. http://postconflict.unep.ch/(2006.
UN Pinheiro Report, 2002–2005. 'Economic, Social and Cultural Rights: The Return of Refugees' or Displaced Persons' Property', Commission on Human Rights. UN Economic and Social Council (E/CN.4/Sub.2/2002/17).
UN Report, 2009. Implementing the Responsibility to Protect. http://www.un.org/en/preventgenocide/adviser/responsibility.shtml.
Uvin, Peter, 2002. On counting, categorizing and violence in Burundi and Rwanda. In: Kertzer, David, Arel, Dominique (Eds.), Census and Identity: The Politics of Race, Ethnicity, and Language in National Censuses. Cambridge University Press.
Uvin, Peter, 1998. Aiding Violence: The Development Enterprise in Rwanda. Kumarian Press.
Várdy, Stevan Béla, Hunt Tooley, T. (Eds.), 2003. Ethnic Cleansing in 20th Century Europe. Social Science Monographs, Columbia University Press.

Verwimp, Philip, 2005. An economic profile of peasant perpetrators of genocide: micro-level evidence from Rwanda. Journal of Development Economics 77 (2), 297–323.

de Waal, Alexander, 2005a. Who are the Darfurians? Arab and African identities, violence and external engagement. African Affairs 104 (415), 181–205.

de Waal, Alexander, 2005b. Famine that Kills. Oxford University Press, Darfur, Sudan.

Waller, James, 2002. Becoming Evil: How Ordinary People Commit Genocide and Mass Murder. Oxford University Press.

Weitz, Eric, 2003. A Century of Genocide: Utopias of Race and Nation. Princeton University Press.

Williams, Rhodri, 2005. Post-conflict property restitution and refugee return in Bosnia and Herzegovina: Implications for international standard-setting and practice. International Law and Politics 37, 441–553.

Zetter, Roger, 2007. More labels, fewer refugees: remaking the refugee label in an era of globalization. Journal of Refugee Studies 20 (2), 172–192.

Ethnic Conflict, Geography of

Frederick W Boal, Queen's University Belfast, Belfast, Northern Ireland, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Much ethnic conflict is territorially based. From a geographical perspective, ethnic conflict can be viewed at three scales – the interstate, the intrastate, and the predominantly micro-urban level. However, conflicts at these scales are intimately connected and interact with each other. The outcomes of conflict can produce secession, segregation, or at least some degree of separation of the groups concerned. Again, this can be viewed across a range of scale levels. A number of territorially based solutions or at least means of regulating ethnic conflict can be delineated – territoriality, dominance, and mutuality.

Ethnic conflict is a worldwide phenomenon. Much of it is territorially based, entailing disputes over the control of space. It occurs over a wide range of spatial scales, from the interstate level to the urban neighborhood. Violence takes place at the extreme, though a significant amount of ethnic conflict is characterized by nonviolent behavior. Causal factors can range from material welfare concerns to identity issues. A range of territorially based solutions or at least attempts at conflict regulation have been put forth.

Spatial Scales

It is possible, from a geographical perspective, to classify ethnic conflicts into three categories – the interstate, the intrastate, and the microscale or intra-urban. However, it must be stressed that there are powerful linkages between the various scales. Events at the interstate scale can reverberate down into individual states and may even have consequences at the level of urban neighborhoods. Likewise, conflicts at the urban microscale can impact at the larger spatial scales, even as far as triggering ill-feeling or worse, in the 'international' arena. No matter what the scale, spatial concentration appears to be very significant in providing a possible base for the generation of conflict. Monica Duffy Toft (2003) and others have suggested that there are two mechanisms at work here: motivation driven and opportunity driven. The first derives from the fact that groups with a high degree of spatial concentration are more likely to display a strong attachment to their territory and are, therefore, more likely to fight for it. Indeed, in some cases, the space may have almost become sacralized. The second mechanism requires a relationship between a group's concentration and its ability to develop and sustain collective organization for violence. Work by Nils Weidman (2009) indicates that the second – opportunity – mechanism is the stronger, though it can be argued that all this is a bit of a chicken-and-egg situation (chicken and egg both playing important roles), with loyalty/attachment to territory and the organizational benefits of spatial concentration forming part of an overriding, inseparable territorial nexus. In terms of cause and effect, it is also worth noting that while spatial concentration may be a possible generator of conflict, such concentration can actually be the product of conflict in the first instance.

Ethnic Conflict at the Interstate Scale

The geographical distributions of many ethnic groups do not conform neatly to existing state boundaries (the Basques and the Kurds provide clear examples of this). The misfit between state and ethnic group boundaries is stressed by Niall Ferguson (Ferguson, 2006: lv) when he notes that "in certain parts of the world there was [and is] an exceptional mismatch between ethnic identities and political structures." He points, in particular, to the ethnic map of eastern and central Europe between World Wars I and II. This mismatch between state and ethnic group boundaries may generate claims. As Donald Horowitz (1985: 281) has put it "if irredentism is conceived as a movement to retrieve ethnic kinsmen and their territory across borders, the common disjunction of group boundaries and territorial boundaries offers scope for irredentas aplenty." Horowitz further notes that a decision to attempt to forcibly retrieve ethnic group members across a border is, in the main, a government decision. This contrasts with secession, which is an ethnic group's decision to break away from the state that it currently finds itself penned into. Secession may involve an attempt to form a separate state, or it may mean breaking away from one state and joining another.

Stanley Lieberson (1972) has stressed that an ethnic group has the possibility of reducing or eliminating any disadvantage it may suffer, by opting for political separation. As he points out such a strategy is not available to other disadvantaged groups in a stratified society (for instance those disadvantaged on grounds of gender, age, or economic position). Thus, as Lieberson outlines it, the most fundamental difference between ethnic and other forms of stratification lies in the fact that the former is nearly always the basis for the internal disintegration of the existing boundaries of the state. Since the objective is most likely to be the formation of a new state, the best way to describe such secessionist groups is to refer to them as 'ethno-nationalists.' Such attempts at the disruption of existing states will not be greeted favorably by existing governments. Thus secession is likely to be pursued and to be resisted by the use of violence.

Interstate ethnic conflict also manifests itself through the roles of 'external national homelands' (Brubacker, 1996) and ethnic diasporas. With the former a state may feel an obligation toward its fellow ethnics in one or more other states. This obligation may translate into attempts to influence those other

states' policies toward the homeland state's co-ethnics, or to the offering of immigration and citizenship privileges for returning members of the ethnic diaspora. Involvement, however, may extend to irredentist claims. Ethnic diasporas, on the other hand, may express concerns about circumstances in their original (real or mythical) home countries. In this case elements in the diaspora communities will encourage their 'own' government to adopt policies that are perceived to be advantageous to fellow ethnics in the 'old country' or ethnic homeland. Irish Americans vis-à-vis Ireland, Jewish Americans vis-à-vis Israel, and Arab Americans vis-à-vis Palestine are clear instances of this. Indeed, some of these groups may well adopt stances that are more extreme than those taken by their fellow ethnics 'back home,' or as Samuel P. Huntington (1997) has put it, they may be 'more Catholic than the Pope.' Huntington also observes that diaspora communities become particularly active in the age of electronic communication: "the commitments of diasporas are reinvigorated and sometimes polarized by constant contact with their former homes." (1997: 274).

Ethnic conflict at the interstate level is also evident in what Samuel Huntington labels 'fault-line wars.' While these wars erupt at the interfaces between 'civilizations' (labeled by Huntington 'Western,' 'Islamic,' 'African,' and so on), they frequently involve localized ethnic groups, acting, as it were, as the standard bearers of their respective global-scale groupings. Of course, some of these fault-line wars occur at the intrastate level. In either case they are struggles for control of people and of territory.

Population transfers play a significant role in many interstate ethnic conflicts. If borders cannot be adjusted, people may be 'adjusted' instead. Sometimes the ethnic territories are 'purified' by agreement, as with the transfer of Greeks and Turks in the early 1920s. Frequently, however, transfers are achieved by expulsion or – the ultimate savagery – by genocide. Finally, it can be noted that some transfers are associated with what John McGarry (1998) calls 'demographic engineering' on the part of states. Here 'agents' may be moved in, being population allotted special roles on behalf of the state concerned; on the other hand some groups perceived as 'enemies' may be moved out, enemies being defined as groups whose present locations pose problems for the authorities and an obstacle to their goals (these goals generally being one of territorial control). As with all types of population transfer, these movements can occur both within and between states.

Ethnic Conflict at the Intrastate Scale

Gurr and Harff (1994), in their book on ethnic conflict in world politics, place ethnic groups in four categories – ethnonationalists, indigenous peoples, communal contenders, and ethnoclasses. Ethnonationalists are relatively large and regionally concentrated ethnic groups, which live within the boundaries of a single state or straddle several adjacent ones. They are likely to be seeking a greater degree of autonomy or even independent statehood. Indigenous people may also be seeking some degree of autonomy, but are particularly concerned about the discrimination and exploitation experienced at the hands of the more technologically advanced people who, by and large, control them. Communal contenders, unlike ethnonationalists, do not seek separation or secession; rather, they seek to share power in the governance of the state they reside in.

As with communal contenders, ethnoclasses seek equality. Unlike contenders, however, they are usually spatially dispersed, rarely having a well-defined territorial base (for example, where ethnic entrepreneurs operate in specialized, spatially dispersed economic niches). Consequently, in their case, strategies aimed at secession from the existing state or at seeking some degree of autonomy within it are not relevant.

Thus, while some ethnic group conflicts impinge at the interstate level, many are contained within one state and do not spill over 'international' boundaries. The roles played by dominant groups in these situations are crucial – they set the context for the other ethnic groups present. What Hennayake (1992) calls 'majority ethnonationalism' can generate a reaction in a situation he refers to as 'interactive ethnonationalism.' In a similar vein, Brubacker (1996) writes of 'nationalizing states' where dominant elites promote the language, culture, demographic position, economic flourishing or political hegemony of their own group, creating an environment disadvantageous to the cultural aspirations and material interests of the other ethnic groups present within the state. Indeed, here issues of inclusion and exclusion are encountered, seen at their most general in the distinction between 'civic' and 'ethnic' nationalism. With the former, the nation is ethnically inclusive of all those who subscribe to the nation's political creed. With the latter, the claim is made that an individual's deepest attachments are inherited, not chosen (Ignatieff, 1993). National identity is defined by ethnic identity – *ein Volk: ein Staat*. However, it must also be stated that the existence on the ground of a clear distinction between 'ethnic' and 'civic' has been questioned by the research of Hjerm (1998) and others.

The maintenance of political and social cohesion in multiethnic states will be possible under three contrasting circumstances: first, where we find civic nationalism to be dominant, second, where one ethnic group is dominant, and, third, where an external power exerts hegemony. This latter situation has been particularly prevalent where imperial powers have operated – from the British, French, German, Austro-Hungarian, and Russian empires of the nineteenth and early twentieth centuries to what Michael Ignatieff (1993) refers to as the Soviet and American joint *imperium* after World War II. However, with the collapse of all but one of these, the lid of the pressure cooker has been lifted and apparently solid states have fragmented (the Soviet Union and Yugoslavia being the most notable instances). This has led to the (re-)emergence of previously suppressed or slumbering ethnic groups. Such a reawakening has threatened the fabrics of many states, generating movements that call for varying degrees of separation (from secession to regional autonomy).

Ethnic struggles at the intrastate level, from time to time, have been characterized both by what may be called 'census wars' and by 'symbolic strife.' With reference to the former, Donald Horowitz (1985) has coined the phrase 'winning the census' – here there are concerns about existing (ethnic) demographic balances and future trends. As he puts it (p. 194) "numbers are an indicator of whose country it is." Census watching becomes an obsession; arguments occur regarding the accuracy of census enumerations, and in some cases ethnic

numeral balances are so sensitive that the census is abandoned (as happened, for instance, in Lebanon; see discussion below on demography).

Matters of symbolism also take on a huge importance in states with internal ethnic conflict. Raymond Breton (1984) argues that the symbolic order is a key component in constructing and maintaining a collective entity. A dominant ethnic group will attempt to impose its symbolic order on the state – flags, official language, anthems, and so on. Other ethnic groups may resist this process, with the consequence that many symbols become contested and contribute to disunity. Ethnonationally contested spaces, in particular, are notable for their symbolic dissensus. Of course, symbols may well help create unity within a particular ethnic group, but in terms of inter-ethnic relations they divide.

Within-state ethnic conflict may arise between indigenous or well-rooted populations, on the one hand, and recent, less well-rooted immigrants on the other. If the immigrant flow is overwhelming (numerically or technologically) the indigenous ethnic group(s) will find themselves marginalized (as was the case in North America). However many immigrant flows have to accommodate themselves to the preexisting social and political environments. In this case, conflict will be over equality of treatment – the immigrant will not expect (or presumably wish) to take over the state that receives them.

Ethnic Conflict at the Microscale

Ethnic conflict gains its greatest intensity at the small scale – here individuals and small groups interface with each other as part of the daily round. Rural, relatively low population density environments can experience intense conflict (as well as day-to-day coexistence), but it is usually in the urban, big city environments that ethnic antagonisms can achieve a peculiarly focused, 'molecular' nastiness.

In the urban context, foreign origin migrants attempt to establish themselves as do those from within the state, some of which may also have their origins in ethnically differentiated regions. While conflict is not necessarily an inevitable outcome, it is of frequent occurrence, as the migrants compete with each other and with the receiving society for scarce resources of housing, employment, educational opportunities, and so on. In these circumstances many ethnic newcomers experience a degree of residential segregation, part based on the wishes of the immigrant and migrant ethnics themselves, part based on the somewhat hostile response of those who see themselves as the 'hosts.'

From a bottom-up perspective, ethnic groups may be seen to cluster together residentially for reasons of physical defense, from a wish to distance themselves from the embarrassment of contact with ethnocentric others, and, indeed from a wish to have a base for organizing politically and, thereby, gaining a say in the decision making in the wider society through judicious use of the electoral system. In the more extreme conflict environments, the segregated residential cluster can provide a base for organizing physical attack against a wider society that is seen to be oppressive. This is where the urban guerrilla is most likely to thrive. Finally segregation can provide a context for cultural preservation, one where the ethnic culture can be tended and, indeed, transmitted to future generations. Some of this clustering is for very positive reasons, but conflict can also be a key generating factor (Boal, 1987).

From the viewpoint of the encompassing society, segregation also has its merits – it provides a means of containing 'alien' populations, it insulates host culture from what may be perceived as undesirable influences, it provides a means of manipulating and minimizing the electoral impact of the immigrant ethnics and, at the extreme, it sets up the ethnic clusters as readily definable targets for attack. Beyond the urban scale, Nurit Kliot (1986) refers to intermingled ethnic clusters in Lebanon as 'hostage situations.' The same situation can apply at the urban scale.

Much ethnic conflict at the urban scale is relatively low key, entailing competition for access to the resources of everyday life and frictions over cultural matters. However, some cities become cockpits of ethnonational strife. In this situation, local conflicts and competition become embroiled in issues of sovereignty, and are thereby greatly intensified. Meron Benvenisti (see Bollens, 2000) has labeled such cities 'polarized,' while Joël Kotek (1999) has applied the term 'frontier.' According to him, frontier cities are above all disputed places because they are subject to contradictory and opposing sovereignty claims. For Kotek, a frontier city is "a territory for two dreams."

The urban encapsulation of ethnonational conflict may be seen as the city merely mirroring wider conflicts over sovereignty, secession, irredentist claims, and desires for regional autonomy. More accurately, cities displaying ethnonational encapsulation are, in reality, key players in the conflicts in their own right. Examples such as Jerusalem, Belfast, Brussels, Montreal, Nicosia, and Sarajevo make the point. Indeed some of these cities are the epicenters of the wider conflict and are the focus of particularly intense ethnonational disputation. Jerusalem stands out in this regard.

Demographic Balance

A factor with the potential to trigger ethnic conflict, operating at all geographic scales, is the influence of demographic balance. At the state level, demographic balance between competing ethnic groups can be very significant in terms of power within the political system. Shifts in demographic balance can be seen as zero-sum situations – one group's gain is the other group's loss. Such shifts in demographic balance are generated by differential birth rates (the ethnic conflict is won or lost in bed; at one time in Québec this was referred to as *la ravanche des berceaux*), differential death rates, or by differential migration flows (voluntary or forced), or by the combined operation of two or more of these processes.

At the regional and urban scale levels, demographic balance can also be very significant, where one group may use its majority position to impose its cultural norms and to shift resource distribution to its advantage (housing, employment, educational resourcing, and municipal services) Slack and Doyon (2001: 158) encapsulate all this in their discussion of Bosnia, claiming that "when ethnonationalism becomes a political force, demography matters." They go on to note that "the threat of decline of one group at the expense of

another further aggravates the pressure for interethnic conflict."

To repeat: the significance of demography needs always to be judged within a series of spatial frameworks – state, region, city, or neighborhood.

Conflict Regulation and Resolution

Solutions to ethnic conflicts can take a number of forms – territorial approaches, dominance approaches, and mutuality approaches. Territorial solutions entail actions that increase the ethnic homogeneity of specified territorial spaces. The most extreme approaches involve the removal – by genocide or by forced population transfer – of one or more ethnic groups, producing what might be called 'purified' spaces. More acceptably, territorial approaches can involve the creation of separate states for each ethnonational group (partition or secession), or for cantonal or federal arrangements within the one state. Of course processes of integration/assimilation can also remove differences within a given state. If these occur by consent, then interethnic relations are likely to be positive; if they occur in a coercive manner the consequences will be alienation and the consequent storing up of future conflict.

Dominance is where hegemonic control is exerted by the most powerful group present in a multiethnic environment. It is unlikely to provide a long-term solution to interethnic conflict – rather it suppresses it, leading to a situation where one is likely to find reluctantly acquiescent ethnic minorities at best and considerable alienation at worst.

Mutuality is where ethnic groups in conflict reach a point where they are prepared to enter into a process of mutual recognition and acceptance. Consociational/power sharing arrangements within the state will be the outcome. In this situation we would find what has earlier been referred to as civic nationalism. Space is shared in a process of mutual accommodation (see discussions in O'Leary and McGarry, 1995; in Boal, 1999). Indeed, as suggested by Staub (2013), even in severe situations such as genocide, measures can be undertaken to achieve reconciliation.

See also: Borders, Anthropology of; Censuses: History and Methods; Cities: Internal Structure; Diaspora; Ethnic Conflicts; Ethnic, Racial, and Nationalist Movements; Ethnicity and Ethnic Groups: Historical Aspects; Ghetto; Identity and Identification; Immigration: Political Aspects; Minorities; Peace Processes; Residential Segregation in the United States; Territorial Conflicts; Urban Conflicts.

Bibliography

Boal, Frederick W., 1987. Segregation. In: Pacione, Michael (Ed.), Social Geography: Progress and Prospect. Croom Helm, London, pp. 90–128.

Boal, Frederick W., 1999. Seeking the common ground. Geopolitics 4, 239–261.

Bollens, Scott A., 2000. On Narrow Ground: Urban Policy and Ethnic Conflict in Jerusalem and Belfast. State University of New York Press, Albany, NY.

Breton, Raymond, 1984. The production and allocation of symbolic resources: an analysis of the linguistic and ethno-cultural fields in Canada. Canadian Review of Sociology and Anthropology 21, 123–144.

Brubacker, Rogers, 1996. Nationalism Reframed. Cambridge University Press, Cambridge, UK.

Connor, Walker, 1994. Ethnonationalism: The Quest for Understanding. Princeton University Press, Princeton, NJ.

Ferguson, Niall, 2006. The War of the World: History's Age of Hatred. Allen Lane, London.

Gurr, Ted. R., Harff, Barbara, 1994. Ethnic Conflict in World Politics. Westview Press, Boulder, CO.

Hennayake, Shantha K., 1992. Interactive ethnonationalism: an alternative explanation of minority ethnonationalism. Political Geography 11, 526–549.

Hjerm, Mikael, 1998. National identities, national pride and zenophobia: a comparison of four western countries. Acta Sociologica 41 (4), 335–347.

Horowitz, Donald, 1985. Ethnic Groups in Conflict. University of California Press, Berkeley, CA.

Huntington, Samuel. P., 1997. The Clash of Civilizations and the Remaking of World Order. Simon and Schuster, London.

Ignatieff, Michael, 1993. Blood and Belonging: Journeys into the New Nationalism. BBC Books, London.

Kaplan, David H., 1992. Nationalism at a micro-scale: educational segregation in montreal. Political Geography 11, 259–282.

Kliot, Nurit, 1986. Lebanon: a geography of hostages. Political Geography Quarterly 5, 199–220.

Kotek, Joël, 1999. Divided cities in the European cultural context. Progress in Planning 52, 227–237.

Lieberson, Stanley, 1972. Stratification and ethnic groups. In: Richmond, Anthony H. (Ed.), Readings in Race and Ethnic Relations. Pergamon, Oxford, UK, pp. 199–209.

McGarry, John, 1998. Demographic engineering: the state-directed movement of ethnic groups as a technique of conflict regulation. Ethnic and Racial Studies 21, 613–638.

O'Leary, Brendan, McGarry, John, 1995. Regulating nations and ethnic communities. In: Breton, Albert, Galleotti, Gianluige, Salmon, Pierre, Wintrobe, Ronald (Eds.), Nationalism and Nationality. Cambridge University Press, Cambridge, UK, pp. 245–289.

Staub, Ervin, 2013. Building a peaceful society. American Psychologist 68 (7), 576–589.

Slack, J. Andrew, Doyon, Roy R., 2001. Population dynamics and susceptibility for ethnic conflict: the case of Bosnia and Herzegovina. Journal of Peace Research 38, 139–161.

Toft, Monica D., 2003. The Geography of Ethnic Violence. Princeton University Press, Princeton, NJ.

Waterman, Stanley, 1987. Partitioned states. Political Geography Quarterly 6, 151–170.

Weidman, Nils B., 2009. Geography as motivation and opportunity. Journal of Conflict Resolution 53, 526–543.

Ethnic Conflicts

John Coakley, Queen's University, Belfast, Northern Ireland

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.M. Williams, volume 7, pp. 4806–4810, © 2001, Elsevier Ltd.

Abstract

The literature on ethnic conflict has mushroomed in the past three decades, but, despite a large volume of analytical output, important questions remain unanswered. There is no consensus on how this term is to be defined, nor is there full agreement on the identification and classification of cases. Nevertheless, there is substantial overlap between major cross-national data collections, and these have facilitated research into the causes and consequences of ethnic conflict. The associated qualitative literature has helped to advance both academic research in the area and policy relevant exploration of paths toward ethnic conflict resolution.

The study of conflicts labeled 'ethnic' has emerged only relatively recently as a major focus of research, though the phenomenon to which it now refers may well have been analyzed earlier using different terminology. During the two decades 1970–89, which were marked by many instances of domestic strife that would today be described precisely as 'ethnic conflict', only nine academic articles bearing this expression in their title are recorded in the Scopus bibliographical database, out of more than 600 000 English-language articles in the area of the humanities and social sciences. During the 1990s, however, this exploded to 94 articles, and in the first decade of the twenty-first century it expanded further to 164 articles (out of a total of 1.8 million English-language articles in the humanities and social sciences). More broadly based bibliographical tools, such as Google Scholar and library catalogs, report a rather greater number of items, but here, too, the proportionate upsurge in interest in the 1990s is obvious.

Any effort to generalize about this burgeoning literature needs to start by noting that the expression 'ethnic conflict' itself gives rise to a range of potential challenges. These include problems of definition (in that each of the two words in the expression lacks an unambiguous meaning), difficulties of description (since even listing and outlining the parameters of cases of ethnic conflict take us into disputed terrain), issues of explanation (because differing theoretical perspectives lead to varying conclusions as to how ethnic conflict is to be accounted for), and disagreements over outcomes (as conflicting empirical data and theoretical expectations lead to incompatible conclusions about the manner in which ethnic conflict is managed). This article considers these matters in turn.

Definition

Efforts to define 'ethnic group', and related terms such as 'nation', have had a long and unhappy history. Walker Connor (1978) summarized the dilemma in the title of a much-cited article, 'a nation is a nation, is a state, is an ethnic group, is a …', and the definitional problem has been addressed over the years in the social science literature (for example, McKay and Lewins, 1978; Brown, 1993: 4–5; Chandra, 2006). Authoritative international bodies have tried to bring intellectual order to bear, as in the case of UNESCO's glossary in the area of 'ethnic questions' (UNESCO, 1977) and a parallel initiative by the Research Committee on Conceptual and Terminological Analysis of the International Social Science Council (Riggs, 1985). Individual scholars have also sought to tackle the question by compiling comprehensive lists (for example, Spira, 1999).

The need for an agreed terminology is acute not just in scholarly analysis but also in public policy making. The legacy of efforts to clarify language in this area in the former Soviet Union, with its complex ethnically mixed population, is still to be seen in certain communist-run states such as China. Soviet nationalities policy, developed in the early twentieth century and adopted in much of the communist world, required the formal identification of ethnic groups, since they were seen as entitled to institutional recognition. Soviet terminology thus came to reflect a hierarchy of organizational forms, ranging from the most developed collectivity, the 'nation', through 'nationality', 'ethnic group', and 'ethnographic group' (Connor, 1984: 217–239; Fedoseyev et al., 1977: 17–50).

Eventually the challenge of identifying ethnic groups struck western policy makers too – they needed data on ethnic divisions with a view to designing specific policy initiatives (for example, in promoting measures to reduce inequality). By the early twenty-first century many states that had apparently been blind to ethnic difference in the past now began to invite residents to respond in the population census to questions about their ethnic (and sometimes 'racial') background. For example, the US census of 2010 asked two such questions: 'Is person X of Hispanic, Latino, or Spanish origin?' and 'What is person X's race?' (Humes et al., 2011: 1). The English and Welsh census of 2011 asked two similar questions: 'What is your ethnic group?' and 'How would you describe your national identity?' (ONS, 2012: 12–13). In Canada, the corresponding question in 2011 was 'What were the ethnic or cultural origins of this person's ancestors?' (Statistics Canada, 2013: 5). These questions may have been confusing, but in each case the range of precoded response categories made the task of respondents easier, even if it meant forcing them into boxes defined by census takers. Survey research analysis has engaged in a similar struggle to measure ethnic identity, but without reaching a consensus on a single, preferred method to

tackle this extremely challenging question (Coakley, 2007: 581–583).

Efforts to resolve the terminological problem by inventing entirely new words have been less than successful. For example, Anthony Smith (1971: 187–191) adopted the French term 'ethnie' as an alternative, and Pierre van den Berghe (1981: 22) used a similar noun, 'ethny'. But these terms failed to be widely adopted. Arguably the only case where new terminology made an impression on the lexicon of ethnic studies was Walker Connor's (1994) popularization of the terms 'ethnonational' and 'ethnonationalism'; but these have rather more specific meanings than 'ethnic'. To facilitate further discussion here, it may be appropriate simply to adopt a stipulative approach, and to define an ethnic group as 'a large collectivity whose members are linked by certain cultural characteristics – including sense of sharing a common past – which they and others see as defining a social boundary between members and nonmembers of the group' (Coakley, 2012: 11). This of course leaves certain loose ends (it is not clear what 'large' means, or how 'cultural' characteristics may be measured); but it helps to narrow the parameters of what is at stake.

Because of these difficulties of terminology, determining when a conflict is 'ethnic' is far from straightforward. Of course, ethnic identity and conflict are symbiotically related; it has for long been acknowledged that social conflict feeds on differentiation, but also shapes it (see Coser, 1956). When, though, is a confrontation between two or more ethnic groups actually a 'conflict'? Does a purely latent 'conflict' merit that description? For example, outside observers and some insiders may perceive an apparently objective, fundamental conflict of interest between two ethnic groups, as when two such groups coexist in circumstances of deep social stratification. But conflict of interest would not necessarily be widely interpreted as real 'conflict', and a particular point along a scale must be used to define the moment where tension and disagreement spill over into conflict (for an overview of approaches toward explaining this transition, see Opp, 2009). Sidestepping the issue of intensity, one useful definition sees ethnic conflict as referring to 'situations in which combatants take recourse to the systematic use of violence for strategic purposes and in which at least one combatant defines itself primarily in relation to a distinctive ethnic identity' (Cordell and Wolff, 2011: 4). This, of course, still leaves the cutoff point undefined, and researchers are by no means unanimous in responding to this, as will be seen in the next section.

Description and Classification

One of the clearest illustrations of the debilitating consequences of the terminological confusion discussed in the last section is the set of difficulties that encounter the researcher who seeks to identify and classify cases of ethnic conflict. Ethnic conflicts are of course a subset of political conflicts more generally, so it is appropriate as our starting point to look at the major cross-national data collections that exist in this area, and then to assess their value as instruments for the compilation of exhaustive lists of cases of ethnic conflict.

One of the earliest attempts to offer a systematic, quantitative approach to intrastate (and, indeed, interstate) conflicts was the path-breaking study by David Singer and Melvin Small (1972) of the 'correlates of war' (COW). This evolved subsequently into a major database of intrastate wars and of participants in these, spanning the period 1816–2007 (Sarkees and Wayman, 2010). Two other early American projects with an overlapping agenda supplemented and in certain respects competed with this. One, the 'Polity' project, focused on political system characteristics and variables accounting for stability and change (Eckstein and Gurr, 1975). Its current version (Polity IV) extends over the period 1800–2012 (Marshall et al., 2013). The second had a more specific concern with minorities: the 'Minorities at Risk' project, which concentrates on minorities that have achieved a minimum degree of political mobilization (see Gurr et al., 1993; MAR, 2014). While of themselves these two data sets do not code conflict, they have been used as a basis for other collections which do – for example, in the case of the work of Fearon and Laitin (2003), or the 'major episodes of political violence' (MEPV) data set developed by Monty Marshall (2013).

This body of data has been reinforced by parallel programs in Europe. Two examples may be cited. First, as part of a strong Scandinavian tradition of peace research, the Uppsala Conflict Data Program (UCDP) and the International Peace Research Institute, Oslo (PRIO) have produced – among other related data sets – successive editions of an 'armed conflict' data set spanning the years since 1945 (Gleditsch et al., 2002; Themnér and Wallensteen, 2013). A second widely used data set has emerged from the Heidelberg Institute for International Conflict Research (HIIK), which produced a regularly updated 'Kosimo' ('conflict simulation model') data set (Pfetsch and Rohloff, 2000), though in its reconfigured version it is not currently publicly available. These by no means exhaust the range of such data sets, but are sufficiently diverse to illustrate the issues of measurement that arise.

Of course, many data collections sidestep difficulties of measurement by relying mainly on qualitative reporting: the *Encyclopedia of Modern Ethnic Conflicts* (Rudolph, 2003), for instance, the *Encyclopedia of Ethnic Relations* (Levinson, 1994), and the various editions of *Mondes rebelles* (Balencie and de La Grange, 2005). But most of the programs already mentioned typically present a large volume of qualitative descriptive material alongside the quantitative indicators. Some are associated with the COW data set (for example, Sarkees and Wayman, 2010). MAR (2014) maintains a set of valuable online reports on the minorities included in its quantitative data set. The Polity IV Project (2014) produces a set of regularly updated country reports. Qualitative reports on armed conflicts have appeared annually since 1993 in PRIO's *Journal of Peace Research* (normally in the penultimate issue), parallel to an online *Conflict Encyclopedia* maintained by UCDP (2014); and HIIK (2014) has published a detailed annual *Conflict Barometer* since 1992 (in German in the early years, but in English from 2002 onwards).

These collections, however, immediately highlight significant problems of definition and classification: as well as varying interpretations of conflict and of cutoff points for inclusion, different data sets classify conflicts in different ways as regards their substantive basis (such as the extent to which they are 'ethnic') and their levels of intensity. If we select four of

the data sets discussed above, a brief description of their approach to definition and classification will highlight this point. They define their primary elements of analysis as follows:

Correlates of War (COW): Intrastate wars are wars that are fought within state borders between a government and non-government forces (civil war), or at least two non-government forces (intercommunal war) … in order to be classified as a civil war, the central government should be actively involved in military action with effective resistance for both sides, and there should be at least 1000 battle related deaths during the civil war.
(see Sarkees and Wayman, 2010)

Major Episodes of Political Violence (MEPV): Major episodes of political violence involve at least 500 directly-related fatalities and reach a level of intensity in which political violence is both systematic and sustained (a base rate of 100 directly-related deaths per annum).
(Marshall, 2013: 7)

Political conflicts (HIIK – Kosimo): Conflict is defined as the clashing of opposing interests or positional differences around national values and issues (independence, self-determination, borders and territory, access to or distribution of domestic or international power). A conflict has to be of some duration and is carried out between at least two parties (states, groups of states, organizations or organized groups) that are determined to pursue their interests and win their case. At least one party is the organized state.
(Pfetsch and Rohloff, 2000: 386)

Armed conflicts (UCDP–PRIO): An armed conflict is defined … as a contested incompatibility that concerns government or territory or both where the use of armed force between two parties results in at least 25 battle-related deaths. Of these two parties, at least one is the government of a state.
(Gleditsch et al., 2002: 618–619)

The associated data sets, and their classification systems, are summarized in **Table 1**. It will be seen that the four data sets have different criteria for classifying conflict types. Two (MEPV and Kosimo) recognize a specifically 'ethnic' category, and the Kosimo data set handles the problem of complex cases by using a multiple coding system that allows conflicts to be simultaneously classed as falling into more than one category. In the case of COW, the 'intercommunal war' category overlaps very substantially with that of ethnic conflict, but the addition of a second category ('civil war over local issues') captures other ethnic-type cases without being excessively inclusive. In the case of the UCDP–PRIO data set, the two-dimensional classification system (incompatibility type and conflict type) seems promising at first, in that restricting cases to internal conflicts directed at territory seems plausible; but the 'incompatibility' dimension is of limited value in identifying an ethnic category, since most of the internal conflicts which were classified as being directed at 'government' rather than 'territory' clearly had an ethnic component.

The data sets vary also in their approach to measuring conflict intensity – a recurring problem in the analysis of civil wars (Sambanis, 2004: 815). The Kosimo data set is the most sensitive to conflicts of all kinds: it equates these with clashing interests, even if no deaths result. For example, the smouldering but latent Flemish-Walloon conflict in Belgium (whose high point as regards violence was reached in early 1968, when Flemish students rioted and threw 'Molotov cocktails' in Louvain) is included only in this data set; indeed, its sponsoring body, HIIK, monitored no fewer than 58 conflicts in Europe alone in 2012, though most were at a low level of intensity (HIIK, 2013: 4). For the UCDP–PRIO data set, a threshold of 25 battle deaths is stipulated; this rises to 500 in the MEPV data set, and to 1000 in COW.

Table 2 presents a comparison of the four data sets in respect of their reporting of what might be defined as ethnic conflicts, broken down by continent. To enhance comparability, a standard start date (1945) is taken, thus eliminating most of the cases in the COW project, which begins in 1815. The closing date ranges from 1999 to 2012. It is possible to identify categories in these data sets that approximate ethnic conflicts (though not necessarily using this label in all cases). The much lower numbers recorded by COW and MEPV derive from another inconsistency in approach across the data sets. The 'episodes' that they identify are inclusively defined: a long-running conflict with shifting participation by actors still counts as just one conflict. In the case of UCDP–PRIO, by contrast, coding is highly sensitive to any change in participants, and as coded here such a change results in the identification of a new conflict; but, in addition, it is likely that some cases have been included here which are not ethnic conflicts at all.

These data sets, then, have some value in generating a checklist of potential cases of ethnic conflict, and it is relatively easy to whittle that down to plausible cases. It is true, as Eck (2008: 39) argues, that these different but overlapping definitions result in potentially incompatible and confusing lists of cases, since this diversity 'illuminates not only the theoretical disagreements regarding what conflict is, but has produced confusion in interpreting the statistical COW'. But she regards this position as preferable to the homogeneity that would emerge from use of a single definition in circumstances where there is no consensus about what conflict is – a healthy warning against naïve attempts to believe that a list of such conflicts on which all would agree might ever be drawn up.

Explanation

Why do ethnic conflicts break out? This straightforward but virtually unanswerable question breaks itself down into at least three subordinate questions. First, how do population groups come to see themselves as ethnically related – in other words, what explains the transition from a social category sharing certain linguistic, religious, cultural, or phenotypical features to a self-conscious ethnic group? Second, in what circumstances does an ethnic group become politically mobilized, making demands on the state that set it in conflict with other groups? Third, what causes a politically mobilized ethnic group to abandon conventional political processes and move to the use of physical force in advancing its objectives?

There is substantial agreement on the broad parameters of an answer to the first question, about the manner in which self-conscious ethnic groups come into existence. The appearance of such groups may be explained in part by the impact of differential social and economic dislocation on

Table 1 Overview of four conflict data sets

Substantive types	*Levels of intensity*	*Cases included*
Correlates of War, version 4.1, 2011		
(1) civil war for central control; (2) civil war over local issues; (3) regional internal war; (4) intercommunal war	Duration in days, and number of battle deaths	Intrastate wars, 1816–2007; N = 442
Major Episodes of Political Violence (MEPV), 2012 ed.		
(1) International; (2) international independence; (3) civil; (4) ethnic; each except the second is divided into two subcategories, 'violence' and 'war'	Intensity of either violence or war along a 10-point scale from low to high	Country – years, 1946–2012; 329 episodes; N = 9057
Political conflicts (HIIK–Kosimo), 1999 ed.		
Conflict over (1) territory, borders, sea borders; (2) decolonization, national independence; (3) ethnic, religious or regional autonomy; (4) ideology, system; (5) internal power; (6) international power; (7) resources; (8) other; multiple categorization possible	Four levels: (1) latent conflict – completely nonviolent, open verbal claims; (2) crisis – nonviolent, more intensive claims; (3) severe crisis – sporadic irregular use of force, blockade, threat; (4) war – organized continuing use of force	Political conflicts, 1945–99; N = 641
Armed conflicts (UCDP–PRIO), version 4-2013		
Two incompatibility types: (1) directed at the government or political system; (2) directed at the territory, as in demands for autonomy or secession; further distinction between four conflict types: (1) interstate; (2) extrastate (including colonial and imperial wars); (3) internationalized internal armed conflict; (4) internal armed conflict	(1) Minor (at least 25 battle deaths per year, but fewer than 1000 overall); (2) intermediate (25–999 per year, but more than 1000 overall); (3) major (at least 1000 per year)	Conflict – years, 1946–2012; 253 conflicts; N = 2098

HIIK, Heidelberg Institute for International Conflict Research; UCDP, Uppsala Conflict Data Program; PRIO, International Peace Research Institute, Oslo.
Source: Data sets from COW, 2014. COW Wars v. 4.0, 1816–2007 (Dataset). www.correlatesofwar.org (accessed 15.01.14.); MEPV, 2014. Major Episodes of Political Violence, 1946–2012. www.systemicpeace.org (accessed 15.01.14.); HIIK, 2011. Kosimo Dataset. www.hiik.de (accessed 01.02.11.); UCDP, 2014. Uppsala Conflict Data Program. www.pcr.uu.se/research/ucdp/ (accessed 15.01.14.).

Table 2 Distribution of cases of ethnic conflict by continent and data set, 1945–2012

Continent	*COW 2007*	*MEPV 2012*	*HIIK–Kosimo 1999*	*UCDP–PRIO 2012*
Africa	35	48	36	65
America, Latin, and Caribbean	1	3	3	19
America, Northern	0	0	1	1
Asia	48	45	62	71
Europe	17	13	25	29
Oceania	1	2	4	1
Total	102	111	131	186

Note: Continents follow UN definitions. Cases included are those classified as either civil war over local issues or intercommunal war (COW), those specifically defined as ethnic (MEPV), those where the conflict type has a component labeled ethnic, religious or regional autonomy, even if this is not the dominant type (HIIK–Kosimo), and internal armed conflicts, whether or not with an international component, including some nonethnic conflicts (UCDP–PRIO). COW, correlates of war; MEPV, major episodes of political violence; HIIK, Heidelberg Institute for International Conflict Research; UCDP, Uppsala Conflict Data Program; PRIO, International Peace Research Institute, Oslo.
Source: Data sets from COW, 2014. COW Wars v. 4.0, 1816–2007 (Dataset). www.correlatesofwar.org (accessed 15.01.14.); MEPV, 2014. Major Episodes of Political Violence, 1946–2012. www.systemicpeace.org (accessed 15.01.14.); HIIK, 2011. Kosimo Dataset. www.hiik.de (accessed 01.02.11.); UCDP, 2014. Uppsala Conflict Data Program. www.pcr.uu.se/research/ucdp/ (accessed 15.01.14.).

groups defined by distinctive cultural or other features, and by population movement processes or geopolitical changes (such as military conquest) that bring culturally differentiated groups into close contact with each other. Sudden, substantial movement from the countryside into culturally alien towns, or migration across countries, may in certain circumstances promote the emergence of a collective identity that distinguishes the 'new' group from the earlier settled population – a population which, in turn, may itself find a new sense of collective identity in the face of 'alien' newcomers. Imperial expansion and territorial boundary change may bring similar consequences. While scholars are divided as to the process by which this identity transformation takes place, and the factors that govern it, few, if any, now subscribe to the view that this is a 'primordial' process, driven by innate characteristics within the coexisting groups (Coakley, 2013).

The second question attracts less consensus: how does a previously dormant ethnic group become politically mobilized, making specific and determined policy demands on the state? Structural features (such as growing socioeconomic inequalities) may play a role, but perception of such features is critical. It is normally social, cultural, or political leaders who highlight issues of these kinds; indeed, elites play a central role in this stage of the mobilization process, whether by highlighting the 'injustices' under which subordinate groups labor, or by pointing to the 'threat' to the position of dominant groups that other ethnic groups represent. For new immigrant minorities, marked by varying levels of integration in the host society from generation to generation, from country to country, and from city to city, immediate barriers to social advancement may color attitudes toward the state, and in certain configurations lead to forceful dissent, whether in the form of rioting or by more structured means. Ethnic mobilization may take the form of a new, urban middle class moving into a leadership role, but in other

circumstances religious elites are decisive actors, and dominant groups can rely on the military, the bureaucracy and the aristocracy (groups unevenly represented from case to case) as important resources.

Finally, in what circumstances does a transition take place from electoral dissent or peaceful protest to violent rebellion or intergroup warfare? To start with, there may be circumstances where peaceful or 'constitutional' routes simply do not exist: subordinate groups have no access to legal channels for advancing their demands, either because the state is an authoritarian one, or because these rights are withheld from them, either explicitly or through the use of surrogate exclusionary criteria. Alternatively, even if the conventional electoral or conflict-resolving devices of a liberal democracy are available, one group may feel that it is consigned to permanent minority status, so that armed rebellion is the only alternative to being permanently outvoted in a state whose legitimacy is leaders question. Conversely, the leaders of dominant groups may see violent repression of dissident minorities as the only appropriate response in defending their interests (which they typically represent as the interests of the state). Alongside rational choice-type explanations for the outbreak of ethnic conflict, social psychological theories (focusing on matters of collective identity and security) have much to contribute; and the role of external intervention also needs to be borne in mind (Cordell and Wolff, 2010: 25–43).

Outcomes

How do ethnic conflicts evolve? In what circumstances do they simply bubble on, whether marked by low- or high-intensity violence? How are they sometimes brought to an effective end? Often, the outcome is a consequence of the uneven nature of the conflict between the two sides: the conflict in its overtly violent form is brought to a conclusion by the military victory of one side rather than through a negotiated deal (Licklider, 1995: 684). The defeat of the Tamil Tigers in Sri Lanka in early 2009, signaling the decisive victory of the other side, is an instance. Yet there are more positive examples, where contending parties have managed to hammer out a deal that persists at least in the short and medium term, as in South Africa and Lebanon.

In the case of such negotiated settlements, the central dilemma has been expressed pithily by Roy Licklider (1995: 681): 'how do groups of people who have been killing one another with considerable enthusiasm and success come together to form a common government?'. The issue is not just agreement but also durability: peace settlements after prolonged ethnic conflicts may be particularly fragile. There is a strong tendency for civil wars to reignite even after a settlement, with half of all negotiated agreements breaking down within 5 years (Licklider, 1995: 685). Comparative research points to the general factors that matter: the intensity and character of violence, the capacity of domestic political actors to manage this, and the effectiveness of external forces in intervening (Doyle and Sambanis, 2000). One comprehensive overview of the literature identified a set of background circumstances that are likely to have an impact, in one direction or another, on the survival prospects of a peace settlement (Hartzell et al., 2001: 188–190):

- *Historical period* (agreements reached after the end of the Cold War are more likely to persist, since the actors are less likely to be proxies for rival great powers)
- *State tradition* (agreements within states with a democratic tradition are more likely to last, since elites have more extended experience of political accommodation)
- *Character of underlying cleavage* (agreements between groups divided by ideological considerations are more likely to survive than those between contending ethnic groups, since it is easier to establish a middle ground)
- *Intensity of conflict* (agreements at the end of a low-intensity conflict are more likely to survive, since if conflict intensity has been high the level of continuing mutual suspicion is likely to be greater)
- *Duration of conflict* (agreements that conclude a long conflict are more likely to endure than those which conclude a short conflict, since protracted war is more likely to highlight the attractions of the political path).

The same authors add that the content of the agreement itself also matters. First, agreements that provide for territorial autonomy are more likely to succeed, since the most politically sensitive issues can be defused by allowing each group to adopt its favored formula in its own area, a stable balance between the groups can be established, and additional resources can be made available at substate level. Second, agreements that allow for third-party enforcement are more likely to endure, since they serve to reduce the security concerns of the parties to the conflict and provide a mechanism for intervention at critical moments (Hartzell et al., 2001: 191–193).

These points take us close to the ultimate political design of a society in which ethnic conflict has been substantially resolved, regardless of the means by which this has been brought about. To what extent may we see predictable patterns here? Some analysts cast the net very wide, identifying a range of endpoints for ethnic conflict, from elimination through genocide to accommodation through formal institutional recognition of difference (for example, Horowitz, 1990; Coakley, 1992; McGarry and O'Leary, 1993; Schneckener and Senghaas, 2003: 166–167). For others, the only 'solutions' that merit the description are those where a formula for stable government has been reached and generally accepted (for example, Zartman, 1990; Smooha and Hanf, 1992; Cohen, 1997; Peleg, 2007: 78–83).

These authors collectively produce a long list of approaches, but two dimensions seem to underlie these: the degree to which they rest on a generous regime of individual rights, and the extent to which they acknowledge groups as important components in society. The broad outcomes may be summarized by considering four combinations of positions in respect of these dimensions (Coakley, 2012: 219–239).

- Elimination: Neither individual rights nor collective existence are acknowledged; subordinate ethnic groups are physically removed in a process of 'ethnic cleansing', as in the genocide of most of Europe's Jewish population during the Second World War, or the expulsion of the German populations of Poland and Czechoslovakia after the war.

- Exclusion: Individual rights are not acknowledged, but there is some recognition of group identity. Minorities are, however, excluded from mainstream political life and may be given restricted forms of autonomy, as in the 'homeland' policy of South Africa under apartheid.
- Integration: Full individual rights are recognized, but the existence of subordinate ethnic groups is denied. All of the population may share fully in the government of the state, but must do so by accepting the norms and values of the dominant culture. This may mean the assimilation of minorities, as in the traditional model of the Jacobin state in France, or the 'melting pot' metaphor in the USA.
- Accommodation: Full individual rights are conceded, but group rights are also recognized. This may result in such strategies as power sharing government (or consociation), territorial autonomy, or nonterritorial autonomy (contemporary Belgium illustrates all three approaches).

Conclusion

The difficulties of definition that have impeded the analysis of ethnic conflict are reflected in the discussion above, and in the nature of the quantitative data sets that have been described. Valuable though such data sets are, it is almost impossible to avoid also introducing qualitative considerations, and the rich supplementary descriptive text material presented as additional documentation for these data sets is of central importance. In the first place, it allows us to whittle down the number of cases by excluding those which clearly do not have a significant ethnic dimension (though there remain some marginal cases). Second, it draws attention to cases that have been excluded from the quantitative data sets. The very extensive associated analytical literature that has developed in the past three decades provides the materials for substantial progress in exploring the causes and consequences of ethnic conflict, and in identifying possibly pathways toward settlement. While theoretical disagreements will continue, there is at least scope for limited progress within particular theoretical frameworks in arriving at a balanced understanding of a phenomenon whose implications for societal harmony are so potentially devastating.

See also: Civil Wars; Data Bases and Statistical Systems: International Comparative Research; Ethnic Conflict, Geography of; Ethnic, Racial, and Nationalist Movements; Ethnicity and Ethnic Groups: Historical Aspects; Nationalism: General; Peace Processes; Peace; Racial Relations; Social Movements and Political Violence; Social Protest; Violence: Public; War, Sociology of.

Bibliography

Balencie, J.-M., de La Grange, A. (Eds.), 2005. Les nouveaux mondes rebelles, fourth ed. Éditions Michalon, Paris.

Brown, M.E., 1993. Causes and implications of ethnic conflict. In: Brown, M.E. (Ed.), Ethnic Conflict and International Security. Princeton University Press, Princeton, NJ, pp. 3–26.

van den Berghe, P.L., 1981. The Ethnic Phenomenon. Elsevier, New York.

Chandra, C., 2006. What is ethnic identity and does it matter? Annual Review of Political Science 9, 397–424.

Coakley, J., 1992. The resolution of ethnic conflict: towards a typology. International Political Science Review 13 (4), 343–358.

Coakley, J., 2007. National identity in Northern Ireland: stability or change? Nations and Nationalism 13 (4), 573–597.

Coakley, J., 2012. Nationalism, Ethnicity and the State: Making and Breaking Nations. Sage, London.

Coakley, J., 2013. Primordialism and the study of nationalism. In: Williams, M.H. (Ed.), The Multicultural Dilemma: Migration, Ethnic Politics and State Intermediation. Routledge, London, pp. 153–168.

Cohen, F.S., 1997. Proportional versus majoritarian ethnic conflict management in democracies. Comparative Political Studies 30 (5), 607–630.

Connor, W., 1978. A nation is a nation, is a state, is an ethnic group, is a…. Ethnic and Racial Studies 1 (4), 377–400.

Connor, W., 1984. The National Question in Marxist-Leninist Theory and Strategy. Princeton University Press, Princeton, NJ.

Connor, W., 1994. Ethnonationalism: The Quest for Understanding. Princeton University Press, Princeton, NJ.

Cordell, K., Wolff, S., 2010. Ethnic Conflict: Causes – Consequences – Responses. Polity, Cambridge.

Cordell, K., Wolff, S., 2011. The study of ethnic conflict: an introduction. In: Cordell, C., Wolff, S. (Eds.), Routledge Handbook of Ethnic Conflict. Routledge, London, pp. 1–12.

Coser, L., 1956. The Functions of Social Conflict. Routledge and Kegan Paul, London.

COW, 2014. COW Wars v. 4.0, 1816–2007 (Dataset). www.correlatesofwar.org (accessed 15.01.14.).

Doyle, M., Sambanis, N., 2000. International peacebuilding: a theoretical and quantitative analysis. American Political Science Review 94 (4), 779–801.

Eck, K., 2008. An overview and typology of conflict data: the advantages of data diversity. In: Kauffmann, M. (Ed.), Building and Using Datasets on Armed Conflicts. IOS Press, Amsterdam, pp. 29–40.

Eckstein, H., Gurr, T.R., 1975. Patterns of Authority: A Structural Basis for Political Inquiry. John Wiley & Sons, New York.

Fearon, J.D., Laitin, D.D., 2003. Ethnicity, insurgency, and civil war. American Political Science Review 97 (1), 75–90.

Fedoseyev, P.N., et al., 1977. Leninism and the National Question. Progress, Moscow.

Gleditsch, N.P., Wallensteen, P., Eriksson, M., Sollenberg, N., Strand, H., 2002. Armed conflict 1946–2001: a new dataset. Journal of Peace Research 39 (5), 615–637.

Gurr, T.R., Harff, B., Marshall, M.G., Scarritt, J.R., 1993. Minorities at Risk: A Global View of Ethnopolitical Conflicts. United States Institute of Peace, Washington, DC.

Hartzell, C., Hoddie, M., Rothchild, D., 2001. Stabilising the peace after civil war: an investigation of some key variables. International Organisation 55 (1), 183–208.

HIIK, 2011. Kosimo Dataset. www.hiik.de (accessed 01.02.11.).

HIIK, 2013. Conflict Barometer 2012. Heidelberg Institute for International Conflict Research, Heidelberg.

HIIK, 2014. Heidelberg Institute for International Conflict Research. www.hiik.de/en/ (accessed 15.01.14.).

Horowitz, D.L., 1990. Ethnic conflict management for policymakers. In: Montville, J.V. (Ed.), Conflict and Peacemaking in Multiethnic Societies. Lexington Books, Lexington, MA, pp. 115–130.

Humes, K.R., Jones, N.A., Ramirez, R.R., 2011. Overview of Race and Hispanic Origin: 2010 [2010 Census Briefs]. www.census.gov/prod/www/decennial.html (accessed 15.01.14.).

Levinson, D., 1994. Ethnic Relations: A Cross-cultural Encyclopedia. ABC Clio, Santa Barbara, CA.

Licklider, R., 1995. The consequences of negotiated settlements in civil wars, 1945–1993. American Political Science Review 89 (3), 681–690.

MAR, 2014. Minorities at Risk. www.cidcm.umd.edu/mar/ (accessed 15.01.14.).

Marshall, M.G., 2013. Major Episodes of Political Violence (MEPV) 1946–2012. www.systemicpeace.org (accessed 15.01.14.).

Marshall, M.G., Gurr, T.R., Jaggers, K., 2013. Polity IV Project: Political Regime Characteristics and Transitions, 1800–2012: Dataset Users' Manual. Center for Systemic Peace, George Mason University, Fairfax, VA.

McGarry, J., O'Leary, B., 1993. Introduction: the macro-political regulation of ethnic conflict. In: McGarry, J., O'Leary, B. (Eds.), The Politics of Ethnic Conflict Regulation: Case Studies of Protracted Ethnic Conflict. Routledge, London, pp. 1–40.

McKay, J., Lewins, F., 1978. Ethnicity and the ethnic group: a conceptual analysis and reformulation. Ethnic and Racial Studies 1 (4), 412–427.

MEPV, 2014. Major Episodes of Political Violence, 1946–2012. www.systemicpeace.org (accessed 15.01.14.).

ONS, 2012. Ethnicity and National Identity in England and Wales 2011 [2011 Census, Key Statistics for Local Authorities in England and Wales]. www.ons.gov.uk/ons/rel/census/2011-census (accessed 15.01.14.).

Opp, K.-D., 2009. Theories of Political Protest and Social Movements: A Multidisciplinary Introduction, Critique, and Synthesis. Routledge, Abingdon.

Peleg, I., 2007. Democratising the Hegemonic State: Political Transformation in the Age of Identity. Cambridge University Press, Cambridge, MA.

Pfetsch, F., Rohloff, C., 2000. KOSIMO: a databank on political conflict. Journal of Peace Research 37 (3), 379–389.

Polity IV Project, 2014. Polity IV Project: Political Regime Characteristics and Transitions, 1800–2012. www.systemicpeace.org/polity/polity4.htm (accessed 15.01.14.).

Riggs, F. (Ed.), 1985. Ethnicity: Intercocta Glossary: Concepts and Terms Used in Ethnicity Research (Pilot Edition). Committee on Conceptual and Terminological Analysis of the International Social Science Council, Honolulu, HI.

Rudolph Jr., J.R. (Ed.), 2003. Encyclopedia of Modern Ethnic Conflicts. Greenwood Press, Westport, CT.

Sambanis, N., 2004. What is a civil war? Conceptual and empirical complexities of an operational definition. Journal of Conflict Resolution 48 (6), 814–858.

Sarkees, M.R., Wayman, F.W., 2010. Resort to War: A Data Guide to Inter-state, Extra-state, Intra-state, and Non-state Wars, 1816–2007. CQ Press, Washington, DC.

Schneckener, U., Senghaas, D., 2003. In quest of peaceful coexistence – strategies in regulating ethnic conflict. In: Daftary, F., Troebst, S. (Eds.), Radical Ethnic Movements in Contemporary Europe. Berghahn, Oxford, pp. 165–200.

Singer, J.D., Small, M., 1972. The Wages of War, 1816–1965: A Statistical Handbook. John Wiley, New York.

Smith, A.D., 1971. Theories of Nationalism. Duckworth, London.

Smooha, S., Hanf, T., 1992. The diverse modes of conflict-regulation in deeply divided societies. International Journal of Comparative Sociology 33 (1–2), 26–47.

Spira, T. (Ed.), 1999. Nationalism and Ethnicity Terminologies: An Encyclopedic Dictionary and Research Guide. Academic International Press, Gulf Breeze, FL.

Statistics Canada, 2013. Ethnic Origin Reference Guide: National Household Survey, 2011. www5.statcan.gc.ca/ (accessed 15.01.14.).

Themnér, L., Wallensteen, P., 2013. Armed conflict, 1946–2012. Journal of Peace Research 50 (4), 509–521.

UCDP, 2014. Uppsala Conflict Data Program. www.pcr.uu.se/research/ucdp/ (accessed 15.01.14.).

UNESCO, 1977. UNESCO Thesaurus. UNESCO, Paris [current ed. available on CD-rom].

Zartman, I.W., 1990. Negotiations and prenegotiations in ethnic conflict: the beginning, the middle, and the ends. In: Montville, J.V. (Ed.), Conflict and Peacemaking in Multiethnic Societies. Lexington Books, Lexington, MA, pp. 511–533.

Ethnic Identity and Ethnicity in Archaeology

Margarita Díaz-Andreu, ICREA-Universitat de Barcelona, Barcelona, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, ethnicity is defined as that aspect of a person's self-conceptualization and his or her conceptualization by other individuals that results from identification with one or more broader groups, on the basis of perceived cultural differentiation and belief in a common descent. This definition understands ethnicity as multidimensional, situational, and fluid and makes its link with material culture much more complex than traditionally perceived. Within the last decade, the term 'community' has been introduced into archaeology to refer to a social collective that shares a geographical space and many have agreed that it is a useful concept. For archaeologists, talking about communities instead of ethnic groups circumvents the problem of assigning an ethnic identity to groups for which only material culture remains.

Many may think that the study of ethnicity is a recent topic in archaeology, but this is only partially true. Archaeologists have always been involved in the identification of ethnic groups in the past, but have defined these groups using different terms. In the early years, archaeologists called ethnic groups 'peoples' and 'nations,' terms that were replaced with 'archaeological cultures' from the start of the twentieth century (Díaz-Andreu, 1996). It is only since the 1980s that the growing influence of anthropology (*see* Ethnicity: Anthropological Aspects) on archaeology has persuaded the latter to begin to talk about ethnicity and ethnic groups.

Ethnicity will be defined here as that aspect of a person's self-conceptualization and his or her conceptualization by other individuals that results from identification with one or more broader groups, on the basis of perceived cultural differentiation and belief in a common descent (Díaz-Andreu, 1998: p. 199, see also Jones, 1997: p. 1). However, not all scholars would agree with this definition. It does not focus on the ethnic group in itself as something objective and describable, as many archaeologists would like. Far from it, the definition provided here stresses perception and, although it does not intend to deny commonalities among the members of the same group, including race (*see* Race: Ethnicity, and Health), it does not consider them a fixed set of cultural rules agreed by all members of the ethnic group.

The above definition considers ethnic identity (*see* Identity and Identification) as multidimensional, because it argues that people can identify with one or more broader groups. It proposes, therefore, that multiple ethnic affiliations usually coexist and overlap in the same person. Each individual is (or is potentially) associated with various groups that could be classed within the definition of ethnicity. An example from Europe would be a Mancunian, who may feel themselves to be a Mancunian, English, British, European, or a world citizen. They can even be of Pakistani descent. However, because of our political education, if asked for their ethnic identity, this person would provide a single answer, selecting only one ethnic identity. This reaction has been expected from each citizen since the end of the eighteenth century. However, this might not necessarily be what people feel and what people felt in the past or in non-Western societies (Metcalf, 2010).

Two more aspects may be added to the proposed definition of ethnicity. The first is that ethnic identity can be described as situational. Each of these complementary and enriching ethnic identities can come into play individually in various situations, depending on where the people are, what the situation is and what political ideology they hold. Each individual is active regarding his or her ethnic identities and the daily negotiations will entail that one or several of them are made manifest. Second, ethnic affiliations are fluid. They are and have been in a continuous process of flux, as is reflected by the appearance and disappearance of group names throughout the several centuries referred to in written sources.

Within the last decade, the term 'community' has been introduced into archaeology to refer to a social collective that shares a geographical space (Canuto and Yaeger, 2012; Mac Sweeney, 2011). Communities are seen as groups of individuals held together by living together in a settlement, and also by practices of affiliation: shared ways of acting and behaving, a common understanding of the social meaning of objects, agreement on routines of consumption, exchange, and the curation of material culture and the physical setting in which practices should take place. Some authors have greeted the arrival of this concept as an acknowledgment that not all group identity is necessarily bound by ethnic ties. However, one could also see the use of this concept as a way of circumventing the problem of assigning an ethnic identity to groups for which only material culture remains. It would be possible to argue that community identity may or may not have an ethnic component. Faced with a lack of written information that can inform about the feelings and identity belief of the individuals in a group, archaeologists are on safer ground talking about community rather than ethnic identity.

Of the two theoretical approaches currently predominant in the study of ethnicity, the primordialist and the instrumentalist (Jones, 1997), the conceptualization of ethnicity provided above fits in with the latter. The primordialist perspective considers ethnicity as something inherent, acquired by birth. Thus, the instinctive attachment of the individual to his or her territory and culture is emphasized and the ethnic group is regarded as the natural universal unit of human organization. The instrumentalist perspective, however, views ethnicity as a fluid and situational group identity. It is through practice that individuals perceive and negotiate their ethnic allegiances. This perspective stresses the social aspect

 http://dx.doi.org/10.1016/B978-0-08-097086-8.13032-6

of ethnicity as a regulatory device for social behavior and social organization within each ethnic group and between different ethnic groups. Archaeology has traditionally perceived ethnic identity from a primordialist stance and it is only in the last two decades that the instrumentalist viewpoint has made its appearance.

Archaeological discussion of ethnicity has a long record. Throughout much of archaeology's history, material culture has been understood in a very straightforward manner as a basis for inferring ethnicity. The presence of specific types of material culture and the absence of others have been considered as diagnostics to define groups. For example, the characteristic prehistoric incised beaker pot was seen to indicate the presence of the Beaker culture in an area and, similarly, Samian pottery attested the presence of Romans. This way of looking at ethnicity has been dominant from the early days of archaeology, although it was only clarified theoretically at the turn of the nineteenth century and in the early twentieth century by approaching it from the cultural historical point of view. Culture history first appeared in Austro-German anthropology (Zwernemann, 1983). It was adopted in archaeology a decade later, mainly through the work of the German archaeologist Gustaf Kossinna (Trigger, 1980: p. 44). The preeminence of German archaeology at this time favored the spread of the theory elsewhere. However, anthropology has long worked with a concept of 'culture' (Jenkins, 1997), whereas archaeology has only very recently been rethinking its use of 'culture' and beginning to explicitly discuss, problematize, and theorize ethnicity.

Within the cultural historical perspective, ethnic groups were called cultures (or civilizations in the case of non-primitive societies). The first definition of culture was proposed by Childe. As he put it, "we find certain types of remains—pots, implements, ornaments, burial rites, house forms—constantly recurring together. Such a complex of regularly associated traits we shall term a 'cultural group' or just a 'culture'" (Childe, 1929: pp. v–vi). In Childe and other culture historians' views of culture/ethnicity, there was no place for fluidity and negotiation, nor for perception. One could say that, in practice, archaeologists considered peoples almost as a by-product of objects. For example, changes in material culture were often interpreted as a result of invasions, i.e., as the substitution of one ethnic group by another. The following statement is typical of the archaeology of the period: "These two groups of hoards [in the British Late Bronze Age] may be taken to indicate first, definite invaders … and second, their fusion with the inhabitants and the absorption, modification, or discarding of their exotic types" (Kendrick and Hawkes, 1932: p. 135)

Despite the emphasis on invasions and migrations, individuals (who were presumably involved) were perceived as being passive spectators of the more important process. Hence Childe, when discussing the archaeology along the River Danube, could say, "whether these Moravian peasants actually came themselves up the Danube is immaterial; their culture did" (Childe, 1927: p. 85).

Culture history archaeologists viewed cultures or ethnic groups as organic beings, as living personalities. Their features were conceptualized like those of an individual. A culture was born, developed, and eventually declined, before moving on to another cycle of emergence, rise, and decay. The impact of these ideas is evident in the quasi-obsession with origins at that time, exemplified by such titles as *The Dawn of European Civilization* (Childe, 1925) and *The Beginnings of Egyptian Civilization* (Brunton, 1929).

In many ways, processual archaeology or New Archaeology continued some of the trends seen in culture history into the 1960s and 1970s (Jones, 1997: pp. 107–109; *see* Theory in Archaeology). The aim of processual archaeology was to develop methods for understanding the formation of the archaeological record and to refine approaches for inferring past behavior from observations of that record. This led to a growing interest in, for example, site formation processes (*see* Geoarchaeology), ethnoarcheology (*see* Ethnoarchaeology) and the use of analogy. There was thus an emphasis on the use of a rigorous hypotheticodeductive method (involving the testing of models and the use of inferential generalizations about the past). Within this perspective, culture was seen as a system of interrelated components (Binford, 1968: pp. 16–23). It was, however, to be criticized later, in the same way as functionalist anthropology (*see* Ethnicity: Anthropological Aspects), for its rather fixed views of human behavior.

As distinct from archaeology, the subjectivist or emic perspective for the study of ethnicity was increasingly being accepted in anthropology, at least from the 1970s (Jones, 1997). The general recognition of the role of perception in the definition of identity, and in particular of ethnicity, is partly a consequence of the influential work of Fredrik Barth. Barth (1969) radically changed the study of ethnicity with his proposal that ethnicity was something that could not be objectified, but that was defined by diagnostic socially relevant factors for membership and the (successful) ascription of the members to their ethnic group. Although he pointed out that all ethnic groups have their own overt signals (diacritical features that people look for and exhibit to show identity), together with basic value orientations (the standards of morality and excellence by which performance is judged), he considered that the critical features defining ethnicity were self-ascription and ascription by others, and that there was "no simple one-to-one relationship between ethnic units and cultural similarities and differences. The features that are taken into account are not the sum of 'objective' differences, but only those which the actors themselves regard as significant" (Barth, 1969: pp. 13–14). Barth's proposal ruled out the possibility of there being an objective and objectified way to designate ethnicity and use material culture as a critical factor in the study of ethnicity, and instead perception took on an essential role. Barth's ideas were later refined by other authors. Among them the work by Ronald Cohen should be emphasized. Cohen (1978) went even further than Barth when he stressed the situational and fluid nature of ethnicity, the influence of which was only possible as far as it was a meaningful element in social interaction. Although neither Barth nor Cohen, nor any of the other authors, arrived at the contention expressed above – that ethnic identity is multidimensional, and it is likely that multiple ethnic identifications coexist in the same person – Cohen (1978) almost reached this conclusion. Some of his examples clearly pointed in that direction. He even stated that "the same person can be categorized according to different criteria of

relevance in different situations" (Cohen, 1978: p. 388), but instead of conceding that more than one of these characterizations could be ethnic, he continued to consider ethnicity as monolithic.

Barth and Cohen were not dealing with prehistoric peoples, and never considered the issue of whether ethnicity was an appropriate concept to be applied to the past. There have been some claims that the concept is not applicable to precapitalist societies. Eriksen (1993: p. 80), for example, argued that outside Europe, ethnicity appeared only after peoples' contact with capitalism. There is some truth in this view, as some new ethnic groups were indeed created by colonizers (Cohen, 1978: p. 383; Gulliver, 1969: p. 15), or even by anthropologists (Cohen, 1978: pp. 380–383). However, colonizers seem to have created these new ethnic groups on the basis of preexisting ones, amalgamating, separating, or reorganizing them. This, rather than denying ethnicity's existence, demonstrates how ethnicity changes throughout time.

Colonizers may not have been successful in all their new creations, but in some they undoubtedly were. In the case of anthropologists (*see* Ethnicity: Anthropological Aspects), Cohen acknowledges that some of them had to provide a name for a 'tribe' (a concept that Cohen sees as equivalent to ethnic unit) "even when the group faded imperceptibly into other more or less similar named groups and was broken up into named subgroupings that had strong we/they feelings dividing them" (Cohen, 1978: p. 381). Following the definition of ethnicity provided above, one should interpret these anthropological categorizations on the one hand as reflecting the multiple levels of ethnic affiliation among these groups. On the other, they should be understood as an example of how anthropologists themselves used their own sense of a 'we/they' division in order to classify 'the other' in their own (nationalistic) language. Their nationalistic perspective would lead them to deny all groupings as ethnic and hence to the need to decide which of them was the only one – the proto-ethno-nation – that should be considered as their object of study. As indicated above, nationalism has left a definite mark on the way ethnicity has been understood, mainly with respect to its one-dimensional characterization.

In archaeology, the first major break with the processual paradigm did not occur until the early 1980s, much later than in anthropology (*see* Ethnicity: Anthropological Aspects), to which it was clearly related. The seminal book in this process was Ian Hodder's *Symbols in Action* (Hodder, 1982b), which was, not coincidentally, an ethnoarchaeological study (*see* Ethnoarchaeology). Hodder depicted the manipulation of material culture by several groups in East Africa as active in creating social relationships, and stressed the need to find ways of interpreting the meaning(s) of material culture. Wider discussions of this topic were, however, well underway by then, as shown by many of the contributions to *Symbolic and Structural Archaeology* (Hodder, 1982a). The major aim of this body of research was the study of living cultures "in order to shed some light on the analysis and interpretation of cultures in prehistoric archaeology"(Hodder, 1982a: p. 1). Interestingly, Hodder's ethnoarchaeological work originated from an overtly processual hypothesis, which was to test the relationship between resources, interaction, and cultural similarity (1982a: p. 8), and it was assumed that cultural similarity would 'reflect' degrees of interaction (Hodder, 1982a: p. 9). However, observations made during Hodder's ethnographic fieldwork showed that while the use and distribution of certain items such as drinking cups, wooden eating bowls and stools followed the expected patterns, those of other items, such as spears and calabashes, did not. Hodder concluded that "the extent to which cultural similarity relates, for example, to interaction, depends on the strategies and intentions of the interacting groups and how they use, manipulate and negotiate material symbols as part of those strategies" (Hodder, 1982a: p. 185). These interacting groups were ethnic (tribal), but other subgroups within those societies, such as younger men and women, were also seen to follow their own strategies in their relations with the dominant older men.

Despite Hodder's important contribution on ethnicity, for several years archaeology remained reluctant to analyze this issue of ethnic identity. In 1989, Shennan published a much-cited article in this area, but in it he referred mainly to culture instead of ethnicity (Shennan, 1989). This was not the case in the Polish journal, *Archaeologia Polona*, which in 1991 dedicated an entire issue to ethnicity and archaeology. The articles published here, together with others that have subsequently appeared (e.g., Díaz-Andreu, 1998; Dietler, 1994; Funari, 1999), show that the debate was becoming more intense. This process was further fostered by the 1997 publication of Siân Jones's book, *The Archaeology of Ethnicity*, which provides archaeologists with an excellent overview of the state of the debate. In the last few years, many archaeologists have continued discussing ethnicity in archaeology (Bawden and Reycraft, 2009; Curta, 2005; Di Hu, 2013; Levy, 2008; Lucy, 2005; Stojanowski, 2010; Weik, 2009).

See also: Ethnic Identity, Psychology of; Ethnicity and Ethnic Groups: Historical Aspects; Ethnicity: Anthropological Aspects; Ethnoarchaeology; Social Identity in Social Psychology.

Bibliography

Barth, F., 1969. Introduction. In: Barth, F. (Ed.), Ethnic Groups and Boundaries; the Social Organization of Culture Difference. Universitetsforlaget, Oslo, pp. 9–38.

Bawden, G., Reycraft, R., 2009. Exploration of punctuated equilibrium and culture change in the archaeology of Andean ethnogenesis. In: Marcus, J., William, P.R. (Eds.), Andean Civilization: A Tribute to Michael E. Moseley. Cotsen Institute of Archaeology, University of California, Los Angeles, pp. 195–210.

Binford, L.R., 1968. Archaeological perspectives. In: Binford, S.R., Binford, L.R. (Eds.), New Perspectives in Archaeology, Atherton Chicago. Aldine, New York, pp. 1–32.

Brunton, G., 1929. The beginnings of Egyptian civilization. Antiquity 3, 456–467.

Canuto, M.A., Yaeger, J., 2012. Introducing an archaeology of communities. In: Nichols, D.L., Pool, C.A. (Eds.), Oxford Handbook of Mesoamerican Archaeology. Oxford University Press, Oxford, New York.

Childe, V.G., 1925. The Dawn of European Civilization. Kegan Paul, London.

Childe, V.G., 1927. The Danube thoroughfare and the beginnings of civilization in Europe. Antiquity 1, 79–91.

Childe, V.G., 1929. The Danube in Prehistory. Clarendon Press, Oxford.

Cohen, R., 1978. Ethnicity, problem and focus in anthropology. Annual Review of Anthropology 7, 379–403.

Curta, F. (Ed.), 2005. Borders, Barriers and Ethnogenesis: Frontiers in Late Antiquity and the Middle Ages. Brepols, Turnhout.

Díaz-Andreu, M., 1996. Constructing identities through culture. In: Graves-Brown, P., Jones, S., Gamble, C. (Eds.), Cultural Identity and Archaeology. Routledge, London, pp. 48–61.

Díaz-Andreu, M., 1998. Ethnicity and Iberians. The archaeological crossroads between perception and material culture. European Journal of Archaeology 2, 199–218.

Dietler, M., 1994. Our ancestors the Gauls. Archaeology, ethnic nationalism and the manipulation of Celtic identity in modern Europe. American Anthropologist 96, 584–605.

Eriksen, T.H., 1993. Ethnicity and Nationalism. Anthropological Perspectives. Pluto, London.

Funari, P.P.A., 1999. Etnicidad, identidad y cultura material. Un estudio del Cimarrón Palmares, Brasil, Siglo XVII. In: Zarankin, A., Acuto, F.A. (Eds.), Sed non Satiata. Teoría Social en la Arqueología Latinoamericana Contemporánea. Ediciones del Tridente, Buenos Aires, pp. 77–96.

Gulliver, P.H., 1969. Introduction. In: Gulliver, P.H. (Ed.), Tradition and Transition in East Africa. Studies of the Tribal Element in the Modern Era. Routledge & Kegan Paul, London, pp. 5–38.

Hodder, I. (Ed.), 1982a. Symbolic and Structural Archaeology, New Directions in Archaeology. Cambridge University Press, Cambridge, UK.

Hodder, I., 1982b. Symbols in Action. Ethnoarchaeological Studies of Material Culture. Cambridge University Press, Cambridge, UK.

Hu, Di, 2013. Approaches to the archaeology of ethnogenesis: past and emergent perspectives. Journal of Archaeological Research 21, 371–402.

Jenkins, R., 1997. Rethinking Ethnicity. Arguments and Explorations. Sage Publications, London.

Jones, S., 1997. The Archaeology of Ethnicity. Routledge, London.

Kendrick, T.D., Hawkes, C.F.C., 1932. Archaeology in England and Wales. Methuen, London.

Levy, T., 2008. Ethnic identity in Biblical Edom, Israel and Midiam. Some insights from mortuary contexts in the lowlands of Edom. In: Schloen, D. (Ed.), Exploring the Longue Durée. Essays in Honour of Lawrence E. Stager. Eisenbrauns, Winona Lake, pp. 251–261.

Lucy, S., 2005. Ethnic and cultural identities. In: Díaz-Andreu, M., Lucy, S., Babic, S., Edwards, D. (Eds.), The Archaeology of Identity. Approaches to Gender, Age, Status, Ethnicity and Religion. Routledge, London, pp. 86–109.

Mac Sweeney, N., 2011. Community Identity and Archaeology. Dynamic Communities at Aphrodisias and Beycesultan. University of Michigan Press, Ann Arbor.

Metcalf, P., 2010. The Life of the Longhouse. An Archaeology of Ethnicity. Cambridge University Press, Cambridge.

Shennan, S., 1989. Introduction: archaeological approaches to cultural identity. In: Shennan, S. (Ed.), Archaeological Approaches to Cultural Identity. Unwin Hyman, London, pp. 1–32.

Stojanowski, C., 2010. Bioarchaeology of Ethnogenesis in the Colonial Southeast. University of Florida Press, Gainesville.

Trigger, B.G., 1980. Gordon Childe. Revolutions in Archaeology. Thames and Hudson, London.

Weik, T., 2009. The role of ethnogenesis and organization in the development of African-Native American settlements. An African Seminole model. International Journal of Historical Archaeology 13, 206–238.

Zwernemann, J., 1983. Culture History and African Anthropology. A Century of Research in Germany and Austria. In: Uppsala Studies in Cultural Anthropology, vol. 6. University of Uppsala, Uppsala, Sweden.

Ethnic Identity and Work

Amelie F Constant, George Washington University, Washington, DC, USA; Temple University, Philadelphia, PA, USA; and IZA – Institute for the Study of Labor, Bonn, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Immigrants do not fare as well as natives in economic terms; even after including many controls, an unexplained part remains. The ethnic identity entered the field of labor and migration economics in an effort to better explain the economic outcomes of immigrants, their behavior, and their often perceived as irrational and suboptimal choices. Quantifying ethnic identity is a major issue; even more challenging is to measure its impact on economic outcomes such as the probability to work or the earnings of immigrants. The thin but burgeoning theoretical and empirical literature shows that ethnic identity has a significant impact on the economic behavior of immigrants.

Introduction

The Greek and Latin etymology of the words ethnic and identity shows that these concepts exist among humans for at least 2000 years. Ethnic is derived from the Greek word *ethnos* that means nation, people, caste, and tribe. It is akin to the Greek word *ēthos* that refers to the spirit, customs, beliefs, and aspirations of a group. The word identity comes from the Latin *identitas* and *idem* and denotes the concept of sameness and oneness. Together they define the notion of a person's individuality and self-recognition, as a person is part of and shares the culture of a larger group. Together, they emphasize boundaries and juxtaposition to the otherness. While reference groups become essential in ethnic identity formation and adaptation, bloodlines are not so relevant.

Social scientists have long understood the importance of ethnic identity as it shapes people's lives individually and within a society and a country. Social psychologists, sociologists, anthropologists, historians and economic sociologists, and political scientists have been studying ethnic identity for decades in social, sociopolitical, psychological, and educational contexts. However, the construct entered the realm of economics and became the subject of in-depth studies only recently. Quantifying the concept of ethnic identity is a major issue. The nonavailability of data, endogeneity issues, selection biases, and inadequate theoretical frameworks have been some of the hurdles economists face.

Studying ethnic identity and its impact on the labor market is of outmost importance; it can offer key insights to the economic outcomes of immigrants in the host country, and suggestions to an effective migration policy that caters to needs of immigrants and encourages their integration. Other social phenomena such as ethnic diversity, multiculturism, ethnic mobilization, diaspora effects, ethnic uprising, minority issues, intercultural dialogue, national identity, etc., make more sense after we understand ethnic identity.

Ethnic identity as a concept, its meaning, and use in economics is presented in Part 1. The theoretical considerations and empirical advancements in measuring ethnic identity are offered in Part 2. How attached people are to their ethnic group and see themselves compared to the majority identity can influence their labor market decisions. Part 3 elaborates how these aspects of ethnic identity relates to choices of participating in the labor force, working in the primary labor market, using social networks to look for a job, employment level (prestige), as well as wage. Finally the conclusion summarizes.

Ethnic Identity as a Concept

The definition of ethnic identity is crucial when one wants to build a theory and provide an analytical framework to evaluate the impact of ethnic identity on economic phenomena. There is no specific definition of ethnic identity in economics. Researchers use common definitions of the words – as they are defined in dictionaries and in other sciences – and often adapt them to their specific study.

Identity is interconnected to the perception individuals have about themselves, their being and their role in society; it is their distinguishing character that includes both the physical appearance and the sense of belonging or not to a group. It is what makes individuals unique and different from others and it is self-defined. It denotes how strongly an individual is attached, relates, or feels a sense of belonging to a particular ethnic group or culture. It signifies a sense of continuity of belonging to a group that shares certain elements such as cultural practices, language, interactions with society and ethnic peers, as well as preferences for same-ethnic friends or colleagues. The ethnic identity of individuals describes their social, psychological, and consumer behavior (Constant et al., 2006b). In contrast to ethnicity, or ancestry, which is a permanent characteristic and externally identifiable, ethnic identity can continuously morph throughout a person's life. It can change or develop due to age, relationships, experience with others, discrimination, religious influence, time, and migration, among a number of other factors. But since it can change and requires self-assessment, it is challenging to capture and measure it in a way to allow for comparisons or evaluations.

As sociologists have shown, ethnic or racial identity can alter when conflicts arise; they distinguish the concepts of situational ethnicity and opposing identity. In the US, segregated neighborhoods can create the structural conditions for the development of 'an oppositional culture that devalues work, schooling, and marriage' and impedes success in the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.94005-4

larger economy (Massey and Denton, 1993). If ethnic identity varies from that of most people in a country, it can also be defined as the minority, or 'oppositional,' identity, which simply indicates that it is outside the majority, or dominant, identity in society. Battu et al. (2007) find that nonwhites choose to adopt 'oppositional' identities; some individuals may identify with the dominant culture and others may reject it, even if it implies adverse labor market outcomes. More recently, Mason and Matella (2014) using self-identification as white as a measure of acculturation and the fraction of all hate crimes directed at Muslims as a measure of stigmatization show that after 11 September in the US, there was a statistically significant and substantively large decrease in the probability that Arab and Islamic Americans will self-identify as white.

When individuals live in their home country are part of the majority (There are some exceptions such as the Jews, who are a minority in other countries and become a majority when they migrate to Israel.) and their ethnic identity is not in conflict or contrast with others. However, when people move to a new country, they immediately feel the clash of cultures. The contrast of cultures, languages, mores, and customs poses a polylemma. An immigrant ponders whether to keep their identity and remain true to their origins, to shed all ancestral attachments and adopt the new culture, or to find some combination of both. Ethnic identity applies to natives as well. By being in contact with immigrants and their culture, natives can also alter their own identity. This additional complication of moving identities makes the measurement of ethnic identity all the more difficult. The ethnic identity of immigrants – the attachment to the culture, language, traditions, etc., of the home country – can be further challenged, changed, or redefined based on their experiences with additional years in the host society, particularly if they face discrimination or acceptance. How people perceive and interpret their surroundings result in a personal and individual adaptation (Phinney et al., 2001). Ethnic identity may or may not influence an individual's other identities such as national identity – how attached a person feels to his/her country. But it has been found that religion is a decisive factor for ethnic identity evolution (Constant et al., 2009).

The way natives receive immigrants, their attitudes, and perceptions about immigrants can trigger different ethnic identity reactions among immigrants. If immigrants switch identities because natives treat them differently, this may very well result in different economic integration patterns. In a theoretical and empirical study, Manning and Roy (2010) examine how people see others and how they see themselves. They focus on the cultural assimilation of immigrants in the UK, the British identity and rights and responsibilities in societies. They find that while almost all UK-born immigrants identify as British, and the rest feels more British with additional stay in the UK, the white UK-born population does not perceive these immigrants as British. Interestingly, they are worried that Pakistanis who feel British are causing problems but are not worried about Italians who do not feel British and cause problems.

How to Measure Ethnic Identity

Ethnic identity becomes all the more salient in economics. Akerlof and Kranton (2000) helped increase discourse about the role of identity in economics. A strand of literature is developing that places identity, behavior, and personality traits in the heart of labor markets and the performance of individuals. The general goal is to explain economic labor market integration and unexplained wage differentials whether examining sex, racial, minority, or immigrant-native differences.

Theoretical Advancements

Once the concepts of one's identity, personality, and well-being entered the utility function of the individual agent and became part of modern economics, researchers showed that these concepts significantly affect economic behavior and can even explain 'irrational' and suboptimal choices and outcomes. Akerlof and Kranton (2010) demonstrated that when facing economic decisions, including those in the labor market, people choose different paths, even when given the same incentives; the missing explaining link is identity. The ethnic identity and its significance in economic behavior triggered new research in labor and migration economics. How a person sees himself or herself compared to others and fitting into a larger group (i.e., the ethnic community or society at large) can lead to nonrational economic outcomes (Constant and Zimmermann, 2011).

Akerlof and Kranton (2000) offer a novel theoretical framework of the utility maximization function and a more realistic economic rigor to neoclassical economics. They conceive a person's self-identification as powerful motivation for behavior, implying that achieving an 'ideal self' and being comfortable with one's identity is bound to affect one's utility function. In this case, utility increases; otherwise, utility decreases. Accordingly, even rational individuals may reach nonoptimal solutions if they cogitate about their identity.

Incorporating a broad class of beliefs including identity, Bénabou and Tirole (2011) developed a model that can explain labor market integration and wage differentials. The authors also examine self-serving beliefs linked to pride, dignity, or wishful thinking that arise endogenously. For example, some individuals may have the human capital and the drive to integrate and flourish in the labor market, but they may not succeed because their behavioral norms are in the way and their self-identity images are muddled.

To explain the evolution of behavioral ethnic codes that people follow in order to maintain social acceptance, Kuran (1998) created a theory of reputational cascades; they can explain why similar societies may exhibit different levels of ethnic activity. The theory incorporates the speed of acting ethnic, chosen under social pressures, which in turn the individuals themselves create and sustain; interdependencies among individual incentives are strong and crucially affect personal choices. Darity et al. (2006) developed a long-term theory of racial (or ethnic) identification formation. In this evolutionary game theory model, there allow for multiple equilibria; one, in which everybody follows an individualist identity strategy, another in which everybody pursues a racialist (or ethnic) identity strategy, and yet another equilibrium as a mixture of both. Accordingly, race or ethnicity may be more or less significant for both market and nonmarket social interactions. The foundation of this theory is the positive impact of racial identity on the productivity of

social interactions, which also explicates the persistence of racial or ethnic advantages in market economies.

The Ethnosizer

As mentioned, one barrier to more thorough economic studies was measuring ethnic identity in order for comparison. A breakthrough came with the 'ethnosizer,' a measure of the intensity of a person's ethnic identity (Constant et al., 2006b). The authors conceived and defined ethnic identity as the balance between the commitment to or self-identification with the culture and society of the country of origin and the host country. With the ethnosizer, they offer a complete concept of ethnic identity and a new approach for empirical modeling.

The ethnosizer can be one-dimensional, two-dimensional, or multidimensional to capture the dynamic and multidimensional character of ethnic identity. The authors allow the ethnosizer to differ among immigrants of the same origin, and/or be comparable among immigrants of different ethnic backgrounds. The one-dimensional ethnosizer in which ethnic identities toward the home and host country are mutually exclusive, it is very much related to the assimilation literature. It assumes that every time a person becomes more assimilated toward the host country he/she has to become less attached toward the home country. This can be visualized by drawing the Cartesian plane and focusing on the upper right quadrant, in which the horizontal axis measures commitment to the home country and the vertical axis measures commitment to the host country. The line that connects the two axes (the hypotenuse in Figure 1) illustrates how immigrants alter their ethnic identity; it is the one-dimensional ethnosizer.

Constant et al. (2006b, 2009), (These states mirror Phinney et al.'s (2001) four acculturation strategies in the psychology field, built off of previous acculturation theory (Berry, 1980).) created the two-dimensional ethnosizer that allows for more realistic possibilities of the ethnic identity of immigrants such as the simultaneous commitment to both countries and other permutations. The ethnosizer is based on five elements that best capture the essence of ethnic identity: language, culture, societal interaction, history of migration, and ethnic self-identification. (Zimmermann et al. (2007a) compared the ethnosizer to the ethnic self-identification measured directly from the German Socio-Economic Panel (GSOEP) and showed that the ethnosizer is a better measure of identity.) As in the one-dimensional ethnosizer, the horizontal axis measures the commitment to the home country and culture and the vertical axis commitment to the host country and culture with respect to the five elements. In this case, immigrants are allowed to traverse through the space defined by the box that the axes can form. This box identifies four distinct states of ethnic identity, differentiated by the strength of commitments to each country. (1) Assimilation: a prominent identification with the host culture and society, a firm conformity to the norms, values, and codes of conduct, and a weak identification with the origin. (2) Integration: a simultaneous dedication to and identification with the ancestral and the host society. This is a state of biculturalism. (3) Marginalization: a pronounced detachment from either the dominant culture or the culture of origin. (4) Separation: a strong commitment to the culture of origin even after years of emigration, along with weak involvement in the host culture and country realities. A migrant can undergo a more intricate journey through the four states. An

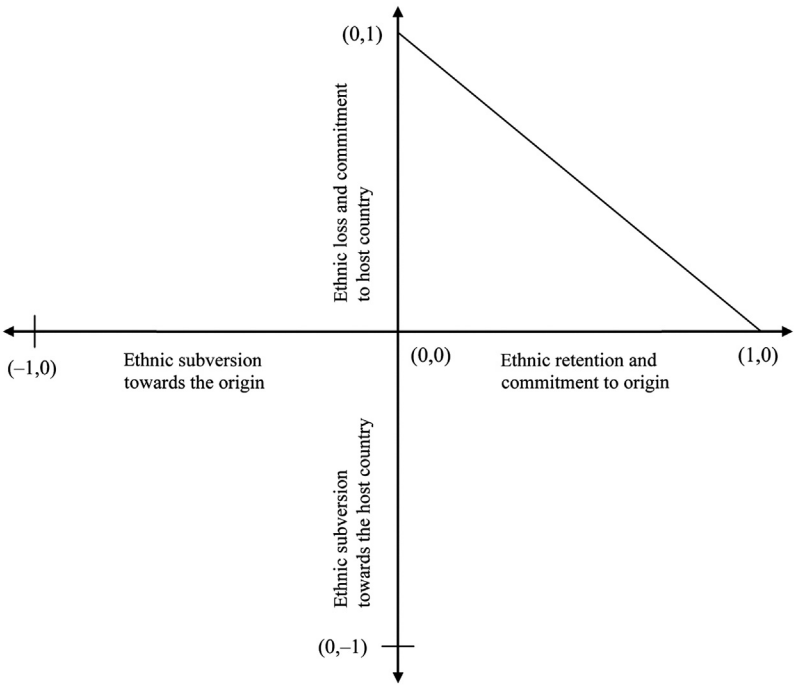

Figure 1 Complete illustration of ethnic identity; retention, relinquishment, and subversion. Reproduced from Constant, A.F., Zimmermann, K.F., 2008. Measuring ethnic identity and its impact on economic behavior. Journal of the European Economic Association 6 (2–3), 424–433.

interesting aspect of the ethnosizer is that in what state individuals end up varies among immigrants even from the same country.

Figure 1 illustrates how Constant and Zimmermann (2008) expanded the theoretical possibilities of the ethnosizer by allowing the commitments to the countries to become negative, thus taking into consideration the multidimensionality of ethnic identity. Accordingly, ethnic identity includes the extreme cases of 'subvert,' negative, or undermining manifestation of ethnic identity. Using the same setting of the Cartesian coordinates, a migrant can move from the upper right quadrant to the lower right or the upper left or the lower left. Moving to the right along the positive part of the horizontal axis expresses ethnic retention and increasing commitment to the country of origin. Point (1,0) shows maximum commitment to and self-identification with the culture of origin and zero commitment to the host country. It is possible that immigrants move beyond point (1,0), as they identify more fanatically with the country of origin. For example, this is the case of radical individuals practicing more extreme views than even comparable coethnics who stay in the country of origin. The ethnic identity can also manifest along the negative part of the horizontal vector indicating that immigrants turn against their own heritage and culture.

Moving up from (0,0) to (0,1) on the positive vector of the vertical axis, indicates increasing identification with the host country to the point of reaching maximum assimilation. The ethnic identity of immigrants can move beyond point (0,1); such is the case of overzealous migrants, who identify with the host country even more than comparable natives. Another possibility is that immigrants actively resent the host culture and develop a subvert self-identification toward it. This is illustrated by the movement on the negative vector of the vertical axis. The lower right quadrant characterizes individuals who keep the ancestral identity, but turn against the host country. In the lower right quadrant, individuals are turning against both countries.

Epstein and Heizler (2015) provide a theoretical framework that sheds light on the ethnosizer literature. In their theoretical model of utility maximization, they incorporate ethnic identity and add a time dimension. They show that two identical migrants arriving at two different periods will have a different the ethnic identification, which will increase over time.

Empirical Advancements

Measuring and quantifying ethnic identity is a difficult task. This intrigues economists, yet generally speaking the field of economics has been unhurried to undertake thorough research on ethnic identity, mainly due to endogeneity issues and sample biases (Constant and Zimmermann, 2013). Do individuals accept work opportunities and make decisions because of their ethnic identity, or do people's identities define their work opportunities? These questions reveal the endogeneity problem. Empirically, endogeneity and sample bias are additionally affected by the fact that respondents are self-evaluating and self-reporting their ethnic attachment making it difficult to prove causality.

Another dimension of the difficulty is lack of data that contain appropriate information about ethnic self-identification and attachment to the home or host country. There are some exceptions such as the Ethnic Diversity Survey in Canada, the British Household Panel, or the GSOEP. Lastly, it is difficult to effectively and quantitatively distinguish between discrimination and an individual's own ethnic choices or preferences; this holds true for many tastes or preferences, as well as for ethnic identity influencing one's employment choices.

Utilizing the GSOEP waves 2000–02, and combining information of the five elements that compose the ethnic identity, Constant et al. (2006b, 2009) created the two-dimensional ethnosizer. The measure ranges from zero, indicating full commitment to the host country, to one indicating full commitment to the country of origin. The authors find that men are closer to the native German ethnicity than women due to greater language use and cultural aspects attachment but the gender difference is not significant. They also find that young migrants are integrated and assimilated the most; immigrants with higher education or college degrees (from the home country) integrate, but do not assimilate. Regarding the ethnicities of immigrants in Germany, the authors show that ex-Yugoslavs assimilate more and separate less than Turks, but they also marginalize more. While Greeks, Spaniards, and Italians were no different than Turks, other ethnicities integrated and assimilated more. The results from the ethnosizer are mostly influenced by premigration characteristics; making this measure exogenous to the economic and social processes that immigrants experience in Germany (Zimmermann, 2007). (Hochman and Davidov (2014) used the GSOEP and found that it is German language proficiency that has a strong impact on the national identification of immigrants and not the other way around.)

Georgiadis and Manning (2013) studied the correlates of national identity in Great Britain and showed that these people who feel they are well treated and tolerated with their values are more likely to identify with the country and have feelings of belonging to society. Also, individuals who live around their likes have a stronger feeling that they too belong to the country.

Below, I focus on labor market outcomes, specifically individual decisions to work, how many hours to work, the use of informal ethnic networks to look for work, as well as employment and wages.

Ethnic Identity and the Labor Market

Some empirical economists embarked upon the quest to improve human capital models and gain a greater understanding of the behavioral determinants of occupational and monetary success by utilizing ethnic identity. The question is whether ethnic identity matters and is useful to immigrants and natives. The rationale is that immigrants are mostly useful in the host country when they possess different talents and skills than natives, and when they bring with them characteristics that are relatively scarce. This improves labor market efficiency. When a population is homogeneous, there is always the risk of lost creativity. There are, however, costs and benefits associated with this cultural capital embodied in immigrants. When

immigrants and natives are complements to each other, both can profit as well as the economy and society can benefit from creativity, dynamism, and greater prosperity. Most of the empirical studies and findings are based on cross-sectional evidence and on a few countries.

When nonwhites in the UK identify with their social environment, their culture of origin, and when social networks assist them in finding jobs – they end up with totally different choices than other otherwise identical individuals (Battu et al., 2007). On the other hand, Battu and Zenou (2010) show that social environment influences the individuals' identity choice. Those nonwhite British who develop and manifest oppositional and extreme identities are penalized in the labor market, facing a 6–7% lower probability of being employed.

Linking identities, conflicting behavioral norms, and job attributes, Russo and van Hooft (2009) find that individuals who experience conflicting norms in the labor market choose job characteristics that can reduce the degree of conflict such as favorable working hours and good relationships with colleagues and managers. While men favor a career when a conflict between career and leisure arises, women do not. Bisin et al. (2011) find that in European countries, immigrants face more difficulties to enter the labor market than natives do. Across all levels of education, those with strong ethnic identities experience a penalty in the labor market. The authors suggest that more flexible labor markets can ease these difficulties for immigrants, but not those with strong ethnic identities. In contrast, in Sweden, income and employment differences for foreign-born men depend on the strength of the *majority* identity, irrespective of ethnic identity (Nekby and Rödin, 2007).

Constant et al. (2006a) studied the effect of ethnic identification on the probability of working in Germany by sex. The probability of immigrant men working is not affected when they are integrated. The probability of working for immigrant women, on the other hand, is significantly affected if they are integrated. Constant and Zimmermann (2009) extend the ethnosizer framework to model labor force participation and earnings disaggregated by sex. They find that the ethnic identity of immigrants in Germany is a strong determinant of their labor force participation. Interesting gender dynamics show that immigrants (both men and women) who are separated and marginalized are less likely to work than those who are assimilated and demonstrate a strong commitment to the German society. In addition, women who were integrated tended to work more than those who were assimilated, but there was no significant difference between integrated men and assimilated women. Curiously, they find that ethnic identity does not significantly affect the earnings of men and women workers. They explain this finding through the selection process in the labor market. In other words, ethnic identity can make a difference in the probability to work, but once immigrants have a job, their ethnic identity does not affect their earnings.

Applying the ethnosizer on different data in Germany – the evaluation data set on the unemployed – Constant et al. (2011) demonstrated that those unemployed immigrants who are separated experience a relatively slow reintegration into the labor market.

Job Search and Working in Ethnic Enclaves or the Primary Market

Identity within an enclave or ethnic group shapes an individual's decisions regarding work. Being part of an informal network can have positive and negative effects on employment and wages. An individual may have access to information and job opportunities within their ethnic group thus more likely to find a job through their network. Benefits could include work that aligns with ethnic practices or religious beliefs. Repercussions may be that the job offers a wage below her human capital capabilities or is outside of the primary labor market. Further, the individual may be stuck in the enclave and never be able to access the entire labor market possibilities. Ethnic identity can influence the decision to enter the primary labor market (labor force participation). More opportunities, along with pressure, encouragement, or willingness to economically assimilate may drive an individual to and work with the majority population, as opposed to working in the enclave. It is also possible that an individual combines both markets such as a medical doctor who works in the primary market, but because of his/her ethnic advantage in language for example, can also have coethnic patients and an expanded clientele.

Pendakur and Pendakur (2005) find that some individuals in Canada desire to work among coethnics and thus are willing to 'pay' for it through accepting traditionally classified low-quality jobs but that are within the ethnic enclave. Regarding job search, the authors find almost 40% found their current job through an informal network, including family and friends. A study on ethnic identity and employment in the US finds that ethnic minorities prefer to have same-race friends despite the economic gains of having white (in this case the majority group) friends, who open social networks in the mainstream labor market (Battu et al., 2007). Choosing to conform to the 'white's norm' depends on peer pressure, offered wages, and the impact on the minority unemployment rate (Battu et al., 2007). This delicate balance between outsiders' perceptions versus those of an individual is demonstrated again in Canada through examining 'visible minorities'; those with strong ethnic identity work jobs with lower occupational prestige than visible minorities without deep ethnic identity (Pendakur and Pendakur, 2005).

Therefore depending on the strength of individuals' ethnic identity, behavioral decisions will be based on a trade-off between expected social and economic outcomes. It seems likely that a person who strongly identifies with her ethnicity could favor social networks to find a job or work in the ethnic economy, despite it possibly being a suboptimal economic decision, since she gains social reassurance and benefits which she may personally value more than additional employment prestige or salary.

An individual's attachment to his ethnic group can affect the decision to work, the (un)employment rate, choosing to work in the primary or ethnic economy, the level (prestige or quality) of his job, and if employment aligns with his human capital. A study on ethnic minorities in England finds that nonwhites adopting an oppositional identity (i.e., strong ethnic identity) experience higher unemployment and low probabilities of finding a job compared to "other non-whites who choose to adopt the white's norm" (Battu and Zenou, 2010; similarly to

Mason's (2004) research showing wage differences depending on ethnic identity). The authors also show that there is a lower probability of being employed (6–7% lower) for those with 'extreme identities.' This is a smaller effect than domestic educational qualifications, interethnic marriage, and home and car-ownership, yet a larger effect than language fluency and ethnic enclaves (Battu and Zenou, 2010; Mason, 2004).

As stated, there are benefits to working in the ethnic economy, but it has been found that participation in it also lowers social activities in the overarching society, creating a social cost (Fong and Ooka, 2002). It is possible that working in the ethnic economy also has economic costs, as it may not efficiently utilize human capital or may offer a lower wage than the primary labor market. For example, an individual may work a job involving the production or sale of ethnic goods, perhaps in an ethnic restaurant, market, or clothing store because she has the necessary language and ethnic know how despite also being overeducated for the position. On the other hand, she may willingly accept, or even prefer, this job since it may offer social benefits or lowers her economic risk.

The trade-off between economic gains and 'social reassurance' exemplifies why seemingly economically irrational behavior may actually be rational for the individual (Akerlof and Kranton, 2000; Pendakur and Pendakur, 2005: p. 2). It also demonstrates the reality that traditional economic measurements for employment level or prestige – such as earnings – are no longer always the appropriate tools to evaluate labor market outcomes. Attempting to include personal utility stemming from ethnic community payoffs seems to be integral for economists to include in analyses.

Wages

Ethnic identity may have direct effects on an individual's wage, which could be surprising if one had assumed wage differences stemmed from how *others* perceive the individual. But Mason (2004) shows that Mexican- and Cuban-Americans were able to increase their income and hourly wages by 'acculturating into a non-Hispanic white racial identity' (although it was not enough to overcome penalties for those with a dark complexion). Acculturation is similar to ethnic identity, as seen above, in the sense that an individual decides how attached she is to a culture or ethnic group. If a person self-reports a low ethnic identity (weak attachment) and identifies more with the majority identity, it can result in higher economic gains than peers with a high ethnic identity (Mason, 2004), particularly if she interacts with the majority and has access to the primary labor market.

Studying the effect of ethnic identification on income, Zimmermann (2007) applied the two-dimensional ethnosizer and showed that integrated men and women experience a dramatic earnings growth; the increase in the earnings of women is higher. Full separation and full marginalization led to a decrease in labor earnings for both male and female migrants. Using the ethnosizer and Greek data, Drydakis (2012) found that assimilation and integration significantly increase the wages of immigrants; separation and marginalization decrease them. More recently, Gorinas (2014) extended the *ethnosizer* by developing *the modernization index* to measure openness to majority norms. He showed that while immigrant employment in general is almost unaffected by the ethnic identity of immigrants in Denmark, immigrant women benefit from significantly better employment outcomes.

Endogeneity Issues

Experiments offer a way to circumvent the endogeneity issue. Following social psychology methods to introduce an exogenous variation in identity, Benjamin et al. (2010), could identify the marginal effect of a particular social category by priming subjects as respondents – temporarily – lean toward the norms associated with the salient category. That is, the authors prime identities with unknown norms and can infer what those norms are through the behavioral response to the prime.

Conclusion

How a person's multifaceted identities mix, conflict, or reinforce each other has been and continues to be an interesting research topic across the social sciences and an important point for public policy, including immigration policy. Varying levels of ethnic identity have social benefits to enrich a country's overall society and culture, as well as economic benefits since those with strong attachments to an ethnic group can possess different human capital and skills than the majority identity. At issue is whether immigrants should assimilate, or integrate in the host country. Likewise economic migrants are willing to move abroad, but how willing or likely are they to adopt the host country's culture and societal norms? Migrants become assimilated, integrated, marginalized, or separated into society (the four states of the ethnosizer), which can directly impact their labor market outcomes.

Competing directly with natives and no longer having any culture-specific human capital as an additional qualification is not advantageous for immigrants. At the same time, preserving one's ethnic identity with dignity and pride is important, and integration can be enriching and rewarding for immigrants.

Research thus far has shown a strong impact of ethnic identity on the labor market. Some individuals can avoid the labor market penalties through relinquishing their ethnic identity and showing greater attachment to the majority, or dominant, identity (i.e., stronger assimilation); although those with certain factors – such as dark complexion – may be unable to do so. Economic gains may rise with increased interaction with the majority population since these connections can open new opportunities in the primary labor market and individuals may experience less discrimination. These interactions help to explain why second-generation immigrants show greater economic assimilation.

Of those self-assessed of having an ethnic identity in Canada, about 40% found their current job through an informal social network, which demonstrates the usefulness of extracting information from the ethnic group and benefiting from labor market outcomes. For those working, jobs may be in the ethnic economy rather than in the primary labor market. These positions carry a trade-off between economic gains and

social benefits (within the ethnic group) that is slowly being understood and captured in data.

Decisions to work, pay taxes, and employment levels and prestige all influence the nation's productivity, efficiency, and output. Hence immigrants' ethnic identity is a public policy topic. It is improbable that a country would aim to marginalize people based on ethnic identity, likewise it is unusual that individuals are separated from both home and host cultures. If a nation's objective is to have assimilated migrants, then giving preference to young migrants with degrees is a strong approach to immigration policy (Constant et al., 2006a,b). If a country decides that national identity is more important, they can create policy to encourage assimilation and for group to feel accepted, which makes it more likely that national identity will be strong (Phinney et al., 2001).

The two-dimensional ethnosizer depends on premigration characteristics, and is exogenous to economic activity. Postmigration variables play practically no role in explaining the *de facto* exogeneity of the ethnosizer variables found in the literature.

See also: Assimilation of Immigrants; Discrimination: Racial; Ethnic Identity, Psychology of; Ethnicity and Educational Achievement; Ethnicity and Ethnic Groups: Historical Aspects; Ethnicity and Migration in Europe; Ethnicity, Sociology of; Ethnicity: Anthropological Aspects; Immigration: Social Psychological Aspects; International Migration; Migration and Social Capital; Migration, Economics of; Migration, Theory of; Migration: Cultural Aspects; Prejudice and Discrimination.

Bibliography

Akerlof, G.A., Kranton, R.E., 2000. Economics and identity. Quarterly Journal of Economics 115, 715–753.

Akerlof, G.A., Kranton, R.E., 2010. Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being. Princeton University Press, Princeton, NJ.

Battu, H., McDonald, M., Zenou, Y., 2010. Oppositional identities and the labor market. Journal of Population Economics 20, 643–667.

Battu, H., Zenou, Y., 2010. Oppositional identities and employment for ethnic minorities: evidence from England. Economic Journal 120, F52–F71.

Benjamin, D.J., Choi, J.J., Strickland, A.J., 2010. Social identity and preferences. American Economic Review 100, 1913–1928.

Bénabou, R., Tirole, J., 2011. Identity, morals, and taboos: beliefs as assets. Quarterly Journal of Economics 126 (2), 805–855.

Berry, J., 1980. Acculturation as varieties of adaptation. In: Padilla, A.M. (Ed.), Acculturation: Theory, Models and Some New Findings. Westview, Boulder, Colorado, pp. 9–25.

Bisin, A., Patacchini, E., Verdier, T., Zenou, Y., 2011. Ethnic identity and labour market outcomes of immigrants in Europe. Economic Policy 26, 57–92.

Constant, A.F., Gataullina, L., Zimmermann, K.F., 2006a. Gender, Ethnic Identity and Work. IZA Discussion Paper No. 2420. Institute for the Study of Labor, Bonn, Germany.

Constant, A.F., Gataullina, L., Zimmermann, K.F., 2006b. Ethnosizing Immigrants. IZA Discussion Paper No. 2040. Institute for the Study of Labor, Bonn, Germany.

Constant, A.F., Gataullina, L., Zimmermann, K.F., 2009. Ethnosizing immigrants. Journal of Economic Behavior and Organization 69, 274–287.

Constant, A.F., Kahanec, M., Rinne, U., Zimmermann, K.F., 2011. Ethnicity, job search and labor market reintegration of the unemployed. International Journal of Manpower 32 (7), 753–776.

Constant, A.F., Zimmermann, K.F., 2011. Work and Money: Payoffs by Ethnic Identity and Gender. Research in Labor Economics 29, 3–30.

Constant, A.F., Zimmermann, K.F., 2008. Measuring ethnic identity and its impact on economic behavior. Journal of the European Economic Association 6 (2–3), 424–433.

Constant, A.F., Zimmermann, K.F., 2011. Migration, ethnicity and economic integration. In: Jovanovic, M.N. (Ed.), International Handbook of Economic Integration. Edward Elgar Publishing, Cheltenham, UK, pp. 145–168.

Constant, A.F., Zimmermann, K.F., 2013. Immigrants, ethnic identities and the nation-state. In: Constant, A.F., Zimmermann, K.F. (Eds.), International Handbook on the Economics of Migration. Edward Elgar Publishing, Cheltenham, UK, pp. 259–275.

Darity, W.A., Mason, P.L., Stewart, J.B., 2006. The economics of identity: the origin and persistence of racial identity norms. Journal of Economic Behavior and Organization 60 (3), 283–305.

Drydakis, N., 2012. Ethnic identity and immigrants' wages in Greece. International Journal of Intercultural Relations 36, 389–402.

Epstein, G.S., Heizler (Cohen), O., 2015. Ethnosizing immigrants: a theoretical framework. IZA Journal of Migration Mimeo 4, 8.

Fong, E., Ooka, E., 2002. The social consequences of participating in the ethnic economy. International Migration Review 36, 125–146.

Georgiadis, A., Manning, A., 2013. One nation under a groove? Understanding national identity. Journal of Economic Behavior & Organization 93, 166–185.

Gorinas, C., 2014. Ethnic identity, majority norms, and the native–immigrant employment gap. Journal of Population Economics 27, 225–250.

Hochman, O., Davidov, E., 2014. Relations between second-language proficiency and national identification: the case of immigrants in Germany. European Sociological Review 30 (3), 344–359.

Kuran, T., 1998. Ethnic norms and their transformation through reputational cascades. Journal of Legal Studies 27 (2), 623–659.

Massey, D.S., Denton, N., 1993. American Apartheid: Segregation and the Making of the Underclass. Harvard University Press.

Mason, P.L., 2004. Annual income, hourly wages, and identity among Mexican Americans and other Latinos. Industrial Relations 43, 817–834.

Mason, P.L., Matella, A., 2014. Stigmatization and racial selection after September 11, 2001: self-identity among Arab and Islamic Americans. IZA Journal of Migration 3 (20), 1–21.

Manning, A., Roy, S., 2010. Culture clash or culture club: national identity in Britain. Economic Journal 120 (542), F72–F100.

Nekby, L., Rödin, M., 2007. Acculturation Identity and Labor Market Outcomes. IZA Discussion Paper No. 2826. Institute for the Study of Labor, Bonn, Germany.

Pendakur, K., Pendakur, R., 2005. Ethnic Identity and the Labour Market: Research on Immigration and Integration in the Metropolis. Working Paper Series No. 05-10.

Phinney, J.S., Horenczyk, G., Liebkind, K., Vedder, P., 2001. Ethnic identity, immigration, and well-being: an interactional perspective. Journal of Social Issues 57, 493–510.

Russo, G., van Hooft, E., 2011. Identities, conflicting behavioural norms and the importance of job attributes. Journal of Economic Psychology 32 (1), 103–119.

Zimmermann, L., Zimmermann, K.F., Constant, A.F., 2007a. Ethnic self-identification of first-generation immigrants. International Migration Review 41, 769–781.

Zimmermann, L., Gataullina, L., Constant, A., Zimmermann, K.F., 2007b. Human capital and ethnic self-identification of migrants. Economics Letters 98/3, 235–239.

Zimmermann, K.F., 2007. Migrant ethnic identity: concept and policy implications. Ekonomia 10 (1), 1–17.

Ethnic Identity, Psychology of

Karmela Liebkind, Tuuli A Mähönen, Sirkku Varjonen, and Inga Jasinskaja-Lahti, University of Helsinki, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.S. Phinney, volume 7, pp. 4821–4824, © 2001, Elsevier Ltd.

Abstract

This article provides an overview of the definitions, development, and key correlates of ethnic identity. Ethnic and multiple identities are discussed primarily from the perspective of social and acculturation psychology, which traditionally employ a sociocognitive approach. In addition, a discursive approach to these phenomena is presented and discussed. Ethnicity is a matter of subjective belief in common ancestry but individuals use different criteria for defining their ethnic identity to construct it situationally. In this article, different theories, methodologies, and single studies on ethnic identity and intergroup relations are linked to individual, interactive, and societal levels of analysis.

Defining Ethnic Identity

Ethnic identity will here be discussed primarily from the perspective of social, acculturation, and discursive psychology. Before that, two major concepts are discussed, ethnicity and identity, as they have been given numerous different definitions within the social sciences.

The term 'ethnicity' has its roots primarily in anthropology and ethnology. Ethnicity refers to a sense of belonging to a particular (assumed) ancestry and origin (Liebkind, 2006). An ethnic group is thought to exist whenever the belief in common descent is used to bind people together to some degree. This sense of origin is often accomplished by defining ethnicity in terms of metaphors of kinship: ethnicity is family writ large. What makes an ethnic group specific is this genealogical dimension. While ethnic group is sometimes used to refer only to those in less powerful positions, ethnicity refers to everybody; all people are members of (at least) one ethnic group, although ethnic group membership has more often been studied in minority than in majority groups. The majority/minority distinction has also been criticized, not least because it tends to confound relative group size with status or power. However, within social psychology, this distinction is widely used, acknowledging that numerical asymmetries often, although not necessarily, coincide with status or power asymmetries.

Even though identity has the reputation of being one of the slipperiest concepts in the social scientist's lexicon, it can be fruitful in conceptualizing ways in which people conceive themselves and are characterized by others. Identity is all about the intricacies, dilemmas, contradictions, and imperatives of the relationship between individuals and their social environment (Verkuyten, 2005). People's identities can be seen to have two components, personal identity and social identity, and the latter derives from the recognition of and value attached to membership in various groups (Tajfel, 1981). Research on self-awareness shows that people can distinguish personal from social identity and that the momentary dominance of one or the other depends on the situation and shapes behavior.

Ethnicity is often seen as an essential part of identity, but the salience of ethnicity varies situationally as well as during the lifetime of an individual. Ethnic identity refers to an individual's sense of self in terms of membership in a particular ethnic group. Although ethnic identity is sometimes used to refer simply to one's self-label or group affiliation, it is generally seen as embracing various aspects, such as self-identification, feelings of belongingness and commitment to a group, a sense of shared values, and attitudes toward one's own ethnic group (Liebkind, 2006).

Ethnic identity differs from other social identities because ethnicity is a matter of subjective belief in common ancestry. The notion of where one came from provides people with an important sense of an enduring identity and feelings of connectedness (Verkuyten, 2005). Different criteria such as physical similarities, cultural characteristics, language, religion, and historical events and myths can all play a role in the definition and justification of a common origin. However, ethnic identity is not necessarily tied to culture: a sense of ethnic identity can remain strong, although cultural changes take place (Verkuyten, 2005).

Subjective beliefs in common origin, descent, and history are socially constructed and therefore always subject to reinterpretations, adjustments, and change. As all social identities depend crucially on acknowledgement and acceptance by others, also ethnic identity is sustainable to the extent that it is expressed and affirmed in social interaction (Verkuyten, 2005). However, although ethnic identities are malleable, they are not complete fabrications. A person's ethnicity is ascribed in the sense that one cannot choose the ethnic group into which one is born, but it is achieved to the extent that the meaning it acquires for one's total identity can be a matter of choice. The fact is that the ascribed aspect of ethnicity can be played down to the extent that ethnic identity is made equal to other social identities, but transmitted as it is in primary socialization, ethnicity is often a relatively pervasive part of identity (Liebkind, 2006).

Levels of Analysis in and Different Perspectives on the Study of Ethnic Identity

Three levels of analysis have typically been distinguished in the study of ethnic identity: individual, interactive, and societal. The first level involves intraindividual (psychological)

processes and personal characteristics, and the focus can be, for example, on pride in and identification with one's ethnic group. The second level is the level of interaction and it refers to the dynamics of everyday contacts in different situations. The emphasis is here on the emergence and maintenance of ethnic identity in situated interaction and ethnic identity is examined in terms of an ongoing process of social definitions and negotiations. The first and second levels are the ones most often studied by social psychologists. The third level of analysis is the societal level and it is made up of political, ideological, cultural, and economic features. Here, ethnic identity is investigated, for example, in relation to legislation, dominant discourses, and the power and status differences between ethnic groups. Social psychologists frequently take also this level into account when studying ethnic identity, then often focusing specifically on intergroup relations and how they are affected by prevailing intergroup attitudes in society. The level of interaction mediates between the other two; it is in interactions that societal relations, beliefs, norms, and values are reproduced, changed, or challenged, and it is in interactions that a sense of self and ethnic identity is formed. Through these interactions, the institutionalized practices, group statuses, and labels affect the way people define and position themselves. Simultaneously, people can challenge and respond to these social structures and expectations, and actively struggle to establish and affirm their own sense of ethnic identity in interactions (Verkuyten, 2005). The emergent and changing qualities of ethnic identity thus coexist with people's relatively stable sense of their self.

There are many different theoretical and methodological approaches to the study of ethnic identity. One approach can be described as sociocognitive, and it informs us about the cognitive, evaluative, and emotional dimensions of ethnic identity (e.g., Social Identity Theory, Tajfel, 1981). Within acculturation psychology, ethnic identity is perceived to be that aspect of acculturation, which concerns the subjective sense of belonging to a group or culture, while acculturation is a broader construct, encompassing a wide range of behaviors, attitudes, and values that change with contact between cultures (Phinney et al., 2001). A third perspective is that of discursive psychology and it contributes to the understanding of identities as contextually and interactionally constructed and negotiated phenomena (e.g., Verkuyten and de Wolf, 2002). These general perspectives complement each other and offer together a broad understanding of ethnic identity. Next, the development of ethnic identity will be briefly discussed, followed by some examples of research on ethnic identity and intergroup relations within the acculturative and other sociocognitive approaches and, finally, discursive approaches.

The Development of Ethnic Identity

Traditionally, ethnic identity development has been treated as part of a broader process of linear development. In this perspective, an ethnic identity is developed and modified as children become aware of other groups and of the cultural/ethnic differences between themselves and others, and attempt to understand the meaning of their ethnicity within the larger setting.

According to the stage model of ethnic identity, developed in the United States by Jean Phinney (1989), ethnic identity begins in a rudimentary form in childhood and remains largely unexamined; that is, children have not thought through for themselves the meaning of their ethnic group membership. Ethnic identity becomes especially central during adolescence when ethnic minority adolescents must negotiate the identity formation process in terms of interpersonal relationships and occupation as well as contend with their minority group status within the majority culture and define how this aspect of their identity informs who they are and who they will become. The optimal outcome of the ethnic identity formation process during childhood and adolescence is the achievement of a secure and confident sense of one's ethnic group memberships. Through the joint processes of exploration and commitment, individuals are expected to move in adolescence from ethnic identity diffusion (lack of a clear identity) to either foreclosure (a commitment without exploration) or moratorium (a period of exploration) and to ethnic identity achievement, involving a firm commitment to one's ethnic group(s) based on an exploration that has led to a clear understanding of ethnicity. By adulthood, most people have acquired a relatively stable and secure sense of themselves as ethnic group members, that is, an achieved ethnic identity (Phinney and Ong, 2007). Although ethnic identity formation has been mainly considered as a major developmental task of adolescence, the process may continue in cycles that involve further exploration or rethinking of the meaning of one's ethnicity.

The development of ethnic identity does not happen in a social vacuum; social identities, intergroup attitudes, and interactions influence each other reciprocally. Social identity processes have implications for intergroup relations because one has a need for a positive self-concept and, as a consequence, there will be bias in social comparisons with other groups to look for ways in which one's own group can be favorably distinguished from other groups (Tajfel, 1981). This group-serving tendency – intergroup bias – can take the form of favoring one's in-group and/or derogating out-groups and sometimes the strength of identification with one's in-group correlates positively with in-group favoritism and/or out-group derogation (e.g., Hodson et al., 2003). However, others assume a negative relationship between in-group identification and in-group favoritism and/or out-group derogation; research on ethnic identity development in adolescents and young adults suggests that strong ethnic identities constitute markers of maturity, consideration of intergroup differences, and are ultimately associated with less intergroup bias (e.g., Wittig, 2008). Thus, a strong ethnic identification among minority members may in some contexts imply positive and in other negative attitudes toward the national out-group.

These mixed results may be due to substantial differences in the ways ethnic identity is measured. Besides the cognitive, emotional, evaluative, and behavioral dimensions of identification measured in various combinations, Roccas et al. (2008) distinguish between four conceptually distinct modes of social identification: importance (how much I view the group as part of who I am), commitment (how much I want to benefit the group), superiority (how much I view my group as superior to other groups), and deference (how much I honor and

submit to the group's norms, symbols, and leaders), with all of them having different antecedents and consequences. Another reason for the mixed results may be that ethnic identities are sometimes reduced to a fixed and singular ethnic identity. Yet it is especially the development of complex and flexible identities (including both national and one or more ethnic components), positive intergroup attitudes, and positive intergroup interactions that are the essence of the development of adaptive social identities among adolescents.

The Complex Structure of Multiple Identities

The development of and changes in ethnic identity represent parts of an individual's larger identity project. Especially for ethnic minority individuals, multiple categories are readily available. For them, ethnic identity and identity as a member of the dominant society (national identity) can be thought of as two dimensions of group identity that vary independently, resulting in varying degrees in multiple identities and in- and out-group differentiation (Liebkind, 2006). In acculturation research, an immigrant's ethnic and national identities are thought of as two dimensions of group identity that vary independently; that is, each identity can be either secure or undeveloped. An individual who retains a strong ethnic identity while also identifying with the new society is considered to have an integrated (or bicultural) identity. One who has a strong ethnic identity but does not identify with the new culture has a separated identity, while one who gives up an ethnic identity and identifies only with the new culture has an assimilated identity. The individual who identifies with neither has a marginalized identity. Actual identity categories and the strength of identification with them depend on a number of factors, including characteristics of immigrant groups and of the places where they have settled (Liebkind, 2006; Phinney et al., 2001). A person can also identify simultaneously with multiple minority and/or majority groups and thus show various degrees of social identity complexity or adopt different modes of identity representation at different times, either during different periods of life or under different conditions or emotional states (Roccas and Brewer, 2002).

The multiply identified individuals face two challenges: how to balance the importance of different group memberships, and how to integrate them (Phinney and Alipuria, 1996). Although having multiple important identities may increase the likelihood of conflict between them, multiple ethnic and national identities may also coexist successfully (e.g., Phinney et al., 2001; Verkuyten, 2005). However, interference between multiple central identities are likely if the cultures associated with them differ; Martinovic and Verkuyten (2012) found that the negative relationship between religious and national identification was strong among Turkish Muslims in Germany and the Netherlands who perceived incompatibility between Western and Islamic values and ways of life.

When different social identities do not converge, there are different ways in which the individual may structure the representations of the in-groups to reconcile the competing identities (Roccas and Brewer, 2002). Even when considering only one ethnic identity at the time together with the national identity, the content of the resulting combined identity may include varying proportions of those two components, and these components may be either parallel or hierarchical; immigrants may perceive their own ethnic group and the national out-group as subgroups of a common national in-group (e.g., Verkuyten, 2005). Multiple identification is, however, possible only if the superordinate identity is sufficiently inclusive and facilitates successful intergroup interaction. Also, multiply identified individuals can claim membership in two or more groups only to the extent that they are accepted by others as members of those groups. One major problem is that members of majority and minority groups may have different preferences for the content of the superordinate identity, that is, what that identity and the relations between the subgroups within it should look like. Although both majorities and minorities may prefer identity integration as a goal, majority members may mean a more assimilative identity and minority members may mean a more separate identity with the term (e.g., Verkuyten, 2005).

Ethnic Identity, Well-Being, and the Consequences of Perceived Discrimination

According to a meta-analysis among ethnic minorities in North America (Smith and Silva, 2011), ethnic identification is more strongly related to positive than to compromised well-being. In addition, bicultural identities are both typical and beneficial for psychological well-being (e.g., Berry et al., 2006). Moreover, secure and positive bicultural identity is found to be associated not only with better well-being but also with more positive attitudes toward out-groups (Phinney and Alipuria, 1996).

However, the relationship between ethnic identity and adaptation and well-being is likely to be moderated by a number of additional factors related to social context, such as national policies, local circumstances within the community, and personal relationships. A consistent and robust finding among adults and adolescents alike is that perceived discrimination has direct and strong impact on the psychological well-being and health of immigrants. Perceived discrimination also affects the ways minority group members identify with both minority and majority groups.

However, individuals react differently to the same circumstances; while some may downplay or reject their own ethnic identity, others may assert their pride in their cultural group as a way of dealing with negative attitudes. The different identity reactions available for ethnic minority members to cope with the experience of being discriminated against include (1) seeking to assimilate into the majority by increasing national identification, (2) developing a so-called reactive ethnicity, or (3) reducing national identification. The Rejection-Identification Model (Branscombe et al., 1999) argues, and research has attested, that increasing group identification in response to perceived discrimination can reduce the detrimental effects of perceived discrimination on psychological well-being as the disadvantaged group provides a sense of acceptance and belonging. In some studies, however,

the relationship between perceived discrimination, ethnic identity, and well-being has not been obtained (e.g., Jasinskaja-Lahti et al., 2009). Different factors, such as the lack of cohesion in the immigrant community or initially high identification may prevent immigrants from increasing their identification with the minority group. Instead, ethnic minority individuals may react to perceived discrimination by withdrawing themselves from the superordinate national group, followed by negative out-group attitudes. Thus, as suggested by the Rejection-Dis-Identification Model (Jasinskaja-Lahti et al., 2009), perceived discrimination may in some contexts prevent minority members from developing a sense of belonging to a national superordinate group. However, causal chains are reciprocal; the level of national identification influences changes in ethnic identity as a reaction to perceived discrimination. Fuller-Rowell et al. (2013) found in their longitudinal study that perceived discrimination in the first year of college decreased ethnic identity commitment across the college years among participants with a weaker national identity, but increased ethnic identity commitment among participants with a stronger national identity. Higher levels of national identity were also associated with greater increases in ethnic identity commitment over time.

However, no single model can fully and universally account for the complex interplay between ethnic identity, perceived discrimination, and well-being. In some studies, high in-group identifiers have been found to be harmed more by perceived discrimination, in others high in-group identification has reduced the psychological costs of discrimination and often it has no effect at all (Pascoe and Richman, 2009).

Negotiating Ethnic Identity

Examining ethnic identities on the level of interactions means studying the actual social processes in which people define and negotiate their complex identities in a specific context. When ethnic identities are studied within discursive psychology, the focus is on what people do with their talk and how they do it, for example, in how people make sense of their identities and how groups are discursively formed through talk (e.g., Verkuyten and de Wolf, 2002). Within this approach, ethnic identity is thus conceptualized as a discursive construct and analyzed contextually in terms of discursive action. One can, for instance, examine how social categories are constructed and used in social interaction and with what consequences. For example, classifications used to categorize various immigrant groups can have implications for how they are treated (Goodman and Speer, 2007).

Discursive studies on ethnicity have focused not only on majority group members' constructions of group distinctions and identities, but also on ethnic minority identities. Verkuyten and de Wolf (2002) showed how Chinese people living in the Netherlands used a variety of discursive and cultural resources when constructing and accounting for their ethnic identities in three different ways: The first way called 'being' was ontological in that a biological model was often used in this discourse, where ethnicity was determined by ancestry. In the second discourse, characterized by emotional motivations and called 'feeling', ethnicity was based on the person's own feelings of belonging to the ethnic group. This feeling, in turn, was explained as a result of early socialization. The third discourse was all about 'doing,' that is, the focus was on ways of acting that are typical for this particular ethnic group, for example, language and specific behaviors. Other examples of discursive research on ethnic minorities include Malhi et al. (2009) who found that South Asian-Canadian women preferred hybrid, multicultural identities over monocultural identities when describing themselves, and Ali and Sonn (2010) who explored how Cypriot Turkish people in Australia constructed multihyphenated identities by using different discourses, that is, those of modern Muslims, of language, of phenotype and ancestry, and of generation. Ali and Sonn (2010) concluded that these discourses highlighted both the current sociopolitical context and collective history and constructing hyphenated identities functioned as an act of self-determination and a way of resisting assimilation.

Constructions of ethnic identities are always situated by nature, relevant to a particular time and place. Varjonen et al. (2013) examined the way a group of ethnic Finnish migrants constructed their ethnic identities before and after migrating from Russia to Finland and found that the main interpretative repertoires used when constructing identities both in Russia and in Finland were the biological repertoire, the repertoire of socialization and the repertoire of intergroup relations. Although the process of ethnic migration is officially based on criteria of ethnic Finnish ancestry, Finland turned out to be the very place where the participants stated that being Finnish was impossible. Before migration, participants mostly presented themselves as Finns, whereas after migration Finnish identity was explicitly problematized and a larger variety of self-labels was used.

In Conclusion

Whether one adopts a cognitive or discursive approach, successful integration of multiple identities can be seen as providing means for the individual to construct for himself or herself both insider and outsider positions in relation to more than one ethnic group. Ethnicity can be an important part of anyone's identity, although research on ethnic identity has focused mainly on minority members. This is due to increasing international migration, which has prompted a need for identity negotiations concerning cultural maintenance vs assimilation and minority members' possibilities to combine ethnic and national identities. Clearly, minority groups have less power than majorities to influence the outcome of such negotiations. However, the way a superordinate national identity is defined within a society affects the sense of belonging of minority and majority members alike.

As noted above, ethnic identities are relational, situational, and constantly negotiated in flexible ways. Furthermore, constructing ethnic identity in social interaction does not exclude a relatively stable sense of oneself as a member of one or more ethnic groups, developed over the life span. National majority groups sometimes use essentialist – and therefore exclusionist – criteria for defining the content and boundaries of national identity, depicting ethnic identity as a rather unchangeable

phenomenon. However, also ethnic minority groups sometimes use essentialist arguments about their own 'authentic' identity in order to resist assimilation and to express a wish to preserve their heritage identity. Thus, essentialist arguments – or any other arguments members of minority and majority groups may use in defining their ethnic identities – are neither 'good' nor 'bad' as such. Which arguments people will use in defining themselves in interaction depends on where and with whom they are, and who they want to be in that particular situation. The authors therefore emphasize the deeply social nature of ethnic identity. In addition, as multiple identities are both increasingly typical and often beneficial for psychological well-being, there is no reason to reduce ethnic identity to a fixed and singular entity but to encourage the development of complex and flexible identities that include both national and one or more ethnic components.

See also: Acculturation; Critical Psychology; Culture and the Self: Implications for Psychological Theory; Ethnicity and Migration in Europe; Identity and Identification, Social Psychology of; Immigration: Social Psychological Aspects; Intergroup Relations; Levels of Analysis in Social Psychology; Racism: Social Psychological Perspectives; Social Constructionism; Social Identity in Social Psychology; Xenophobia: Social Psychological Aspects.

Bibliography

Ali, L., Sonn, C.C., 2010. Constructing identity as a second-generation Cypriot Turkish in Australia: the multi-hyphenated other. Culture & Psychology 16 (3), 416–436.

Berry, J., Phinney, J., Sam, D., Vedder, P. (Eds.), 2006. Immigrant Youth in Cultural Transition: Acculturation, Identity, and Adaptation across National Contexts. Erlbaum, Mahwah.

Branscombe, N.R., Schmitt, M.T., Harvey, R.D., 1999. Perceiving pervasive discrimination among African Americans: implications for group identification and well-being. Journal of Personality and Social Psychology 77, 135–149.

Brown, R., 2010. Prejudice: Its Social Psychology, second ed. Wiley-Blackwell, Malden.

Fuller-Rowell, T.E., Ong, A.D., Phinney, J.S., 2013. National identity and perceived discrimination predict changes in ethnic identity commitment: evidence from a longitudinal study of Latino college students. Applied Psychology 62, 406–426.

Goodman, S., Speer, S.A., 2007. Category use in the construction of asylum seekers. Critical Discourse Studies 4 (2), 165–185.

Hodson, G., Dovidio, J.F., Esses, V.M., 2003. Ingroup identification as a moderator of positive-negative asymmetry in social discrimination. European Journal of Social Psychology 33, 215–233.

Jasinskaja-Lahti, I., Liebkind, K., Solheim, E., 2009. To identify or not to identify? National disidentification as an alternative reaction to perceived ethnic discrimination. Applied Psychology: An International Review 58, 105–128.

Liebkind, K., 2006. Ethnic identity in acculturation. In: Sam, D.L., Berry, J.W. (Eds.), The Cambridge Handbook of Acculturation. Cambridge University Press, Cambridge, pp. 78–96.

Malhi, R.L., Boon, S.D., Rogers, T.B., 2009. 'Being Canadian' and 'being Indian': subject positions and discourses used in South Asian–Canadian women's talk about ethnic identity. Culture & Psychology 15 (2), 255–283.

Martinovic, B., Verkuyten, M., 2012. Host national and religious identification among Turkish Muslims in Western Europe: the role of ingroup norms, perceived discrimination and value incompatibility. European Journal of Social Psychology 42, 893–903. http://dx.doi.org/10.1002/ejsp.1900.

Mähönen, T.A., Jasinskaja-Lahti, I., 2012. Anticipated and perceived intergroup relations as predictors of cultural and national identities: a longitudinal study on Ingrian-Finnish return migrants. European Psychologist 17, 120–130.

Pascoe, E.A., Richman, L.S., 2009. Perceived discrimination and health: a meta-analytic review. Psychological Bulletin 135, 531–554.

Phinney, J., 1989. Stages of ethnic identity development in minority group adolescents. Journal of Early Adolescence 9, 34–49.

Phinney, J.S., Alipuria, L.L., 1996. At the interface of culture: multiethnic/multiracial high school and college students. Journal of Social Psychology 136, 139–158.

Phinney, J.S., Horenczyk, G., Liebkind, K., Vedder, P., 2001. Ethnic identity, immigration, and well-being: an interactional perspective. Journal of Social Issues 57, 493–510.

Phinney, J.S., Ong, A.D., 2007. Conceptualization and measurement of ethnic identity: current status and future directions. Journal of Counseling Psychology 54, 271–281.

Roccas, S., Brewer, M.B., 2002. Social identity complexity. Personality and Social Psychology Review 6, 88–106.

Roccas, S., Sagiv, L., Schwartz, S., Halevy, N., Eidelson, R., 2008. Toward a unifying model of identification with groups: integrating theoretical perspectives. Personality and Social Psychology Review 12, 280–306.

Sala, E., Dandy, J., Rapley, M., 2010. 'Real Italians and wogs': the discursive construction of Italian identity among first generation Italian immigrants in Western Australia. Journal of Community and Applied Social Psychology 20 (2), 110–124.

Smith, T.B., Silva, L., 2011. Ethnic identity and personal well-being of people of color: a meta-analysis. Journal of Counseling Psychology 58, 42–60.

Tajfel, H., 1981. Human Groups and Social Categories. Cambridge University Press, Cambridge, England.

Varjonen, S., Arnold, L., Jasinskaja-Lahti, I., 2013. 'We're Finns here, and Russians there': a longitudinal study on ethnic identity construction in the context of ethnic migration. Discourse & Society 24, 110–124.

Verkuyten, M., 2005. The Social Psychology of Ethnic Identity. Psychology Press, London.

Verkuyten, M., de Wolf, A., 2002. Being, feeling and doing: discourses and ethnic self-definitions among minority group members. Culture & Psychology 8, 371–399.

Wittig, M.A., 2008. A mutual acculturation model for understanding and undermining prejudice among adolescents. In: Levy, S.R., Killen, M. (Eds.), Intergroup Attitudes and Relations in Childhood through Adulthood. Oxford University Press, Oxford, UK, pp. 220–233.

Ethnic Intermarriage

Delia Furtado, University of Connecticut, Storrs, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ethnic intermarriage is often used as measure of social distance between ethnic groups. This article presents statistics on the prevalence of intermarriage and reviews theories of marriage which might be helpful when interpreting these statistics. After a review of the empirical literature on the determinants of interethnic marriage, the article ends with a discussion of its consequences both for those directly involved and for society as a whole.

Ethnic intermarriage occurs when people from two different ethnic groups marry each other. While ethnicity is always used to describe people with similar cultural attributes (such as language, cuisine, and dress), because it is socially defined, the notion of ethnicity can differ across societies. This article will focus on the determinants and consequences of ethnic intermarriage in the US, but the issues discussed are applicable in many contexts.

Significance of Ethnic Intermarriage

Intermarriage rates are often used to measure social distances between groups. These marriages typically only occur in large numbers after a minority group has adopted the cultural patterns of the mainstream. Not only are they reflective of assimilation and ethnic acceptance among those in the marriage, but also they generate interethnic commingling among family members, friends, and broader social networks which in turn results in further assimilation and acceptance. In contrast, marriage within ethnicity, or ethnic endogamy, simplifies the transmission of ethnic traits to new generations and in this way perpetuates the distinctions between ethnic groups.

Given the view of interethnic marriage as the final stage of assimilation (Gordon, 1964), analyses of marriage patterns can offer insights into whether today's immigrants are integrating as rapidly as those in the past, whether some groups assimilate faster than others, and to which segments of the host society immigrants and their descendants assimilate.

Prevalence of Intermarriage

According to a recent Pew Research Center report (Wang, 2012), 6.8% of all new marriages in the US in 1980 involved people of different races or ethnicities while the corresponding figure in 2010 was 15. The report defines out-marriage as a marriage between people from different racial groups or a marriage between a Hispanic and non-Hispanic regardless of race. Among the 2010 newlyweds, whites had the lowest rate of out-marriage (9%) followed by blacks (17%), Hispanics (26%), and Asians (28%). There were no meaningful gender differences in intermarriage tendencies for whites and Hispanics, but black males were significantly more likely to marry outside of their race than black females and Asian females were more likely to out-marry than Asian males (Wang, 2012).

While statistics based on this very broad race–ethnicity-based conception of intermarriage are certainly informative, they can mask interesting patterns that surface when using country of origin or ancestry to measure ethnicity. **Table 1** shows endogamy rates by nativity and year for the 20 most populous ancestries in the US in 1980. Statistics are computed using the 1980 Census and 2008–10 American Community Surveys. A marriage is considered endogamous if the first self-reported ancestries of the spouses match. As can be seen in the table, the foreign born have significantly higher endogamy rates than the native born. This pattern is consistent with the notion that immigrants are less assimilated and so are likely to have less in common with potential spouses of different ancestries. However, since many of the foreign born may have

Table 1 Endogamy rates by ancestry, year, and nativity

	1980		*2008–10*	
	Native born	*Foreign born*	*Native born*	*Foreign born*
Danish	7.8	30.1	5.6	24.0
Dutch	18.3	46.6	15.4	26.4
English	53.1	44.8	36.7	30.9
French	21.5	35.4	16.0	23.1
German	40.7	47.3	36.9	33.8
Greek	28.8	79.2	17.2	66.5
Irish	30.5	48.3	26.0	35.9
Italian	39.6	76.6	26.0	55.6
Norwegian	21.7	39.6	16.4	19.4
Portuguese	30.6	82.9	17.4	68.7
Swedish	12.6	36.6	8.5	19.0
Czechoslovakian	20.8	45.1	9.9	31.6
Hungarian	11.3	51.4	6.6	42.8
Polish	30.0	59.8	16.6	70.5
Russian	40.3	57.0	24.0	63.1
Mexican	73.2	84.2	61.8	86.8
Puerto Rican	52.6	76.2	38.8	65.1
Spanish	37.5	59.3	24.7	53.3
Chinese	57.9	87.2	41.3	84.7
Japanese	74.2	50.1	42.3	45.4

Notes: Endogamy rates computed using data from the Integrated Public Use Microdata 5% state sample of the 1980 Census along with the 2008–10 American Community Surveys.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31097-2

arrived in the US already married, this result might also be explained by their not having the opportunity to marry outside of their ethnicity. The table also shows that, except for the foreign born in three ethnic groups, endogamy rates were lower in 2008–10 than they were in 1980, a trend consistent with recent findings of increasing acceptance of intermarriage (Wang, 2012). Perhaps most noteworthy in the table is the considerable degree of heterogeneity in endogamy rates across ancestries. In order to properly interpret these relationships, it is useful to consider what drives marriage decisions.

Theories of Marriage

Social exchange theory was the first framework developed for understanding intermarriage decisions (Davis, 1941; Merton, 1941). It implies that if whites face higher social costs of interracial, black–white marriage, as compared to blacks, then the white spouse must be compensated for such marriages to actually occur. This compensation might often come in the form of higher socioeconomic status of the black spouse.

Empirical analyses of social exchange yield mixed results. Because taboos against interracial marriage have diminished greatly through the years (Wang, 2012), some of the discrepancies across studies are due to differences in when the study was conducted. However, even when using the same data source, researchers arrive at divergent conclusions because of differences in empirical approach (Gullickson and Fu, 2010; Kalmijn, 2010; Rosenfeld, 2005, 2010). Regardless of methodology, the degree of empirical support for social exchange theory differs across racial groups (Fryer, 2007; Qian, 1999).

The theory of marriage developed by Gary Becker (1973) predicts exchange on some characteristics but matching on others. He conceptualized households as small firms producing 'commodities' such as children, health, companionship, food, and clean clothes. Since these commodities typically require money and time to produce, the model implies exchanges between high-wage men who specialize in the labor market and low-wage women who specialize in domestic skills. The theory also implies matching on characteristics when these matches would make the couple more efficient at producing commodities. For example, spouses with the same ethnic background can more efficiently 'produce' children with ethnic traits.

David Lam's theory of marriage (1988) focuses on couples' joint consumption, as opposed to production, of household goods. The theory predicts that optimal matches are made based on similar demands for household public goods. Because many of the goods shared among family members are ethnicity-based (ethnic traits in children as well as ethnic meals and vacations to the home land), then it is optimal for individuals with similar demands for these goods to marry. Given the emergence of labor-saving household technologies such as microwaves and dishwashers as well as a functioning service industry that enables families to outsource many household activities, today's couples seem to form family based more on consumption complementarities of the type proposed by Lam than on production complementarities proposed by Becker (Stevenson and Wolfers, 2007). Nevertheless, all three marriage models either directly imply or are consistent with high rates of ethnic endogamy and can provide insights into the types of people that are most likely to intermarry.

Determinants of Ethnic Intermarriage

As there is no clear way to measure ethnicity in a US context, empirical studies of ethnic endogamy use several different measures. Most research is limited to the variables available in the large population-representative surveys (Census and Current Population Survey (CPS)): race, Hispanic ethnicity, ancestry, country of birth, and parents' countries of birth. More detailed country of origin information for people that have been in the US for several generations is also available for some Asians and Hispanics. While the variety of ways ethnicity can be defined can make detailed comparisons between studies difficult, the distinctions are not generally very important given the correlations between categories. For example, immigrants born in Japan or second-generation immigrants with Japanese parents typically list a Japanese ancestry. That said, to err in the side of caution, the discussion below focuses on the drivers of intermarriage that tend to be robust across studies.

The determinants of ethnic intermarriage can be separated into three main categories: ethnic preferences, preferences for characteristics which are correlated with ethnicity, and marriage market characteristics.

Ethnic Preferences

Both the Becker and Lam theories of marriage predict that people who are more connected to their ethnic backgrounds are more likely to marry within their ethnicities. By exchange theory, those who are less integrated may need to be more highly compensated before agreeing to out-marry. It should not be surprising, therefore, that immigrants who arrive in the US at an older age or have been in the country for less time are more likely to choose a same-ethnicity spouse (Qian et al., 2012; Chiswick and Houseworth, 2011). Similarly, immigrants who are less fluent in English are more likely to marry endogamously (Bleakley and Chin, 2010).

Ethnic preferences also play a role for second- and later generation immigrants. While the native born are typically fluent in English, their marriage decisions often depend on whether English is their mother tongue (Stevens and Swicegood, 1987). Regardless of language ability, US-born children of immigrants can be sensitive to their parents' preferences for their spouses. This is consistent with the evidence that ethnic endogamy is more prevalent among those marrying young (Chiswick and Houseworth, 2011) and those from countries with traditions of marrying young (Kalmijn and van Tubergen, 2010) but less prevalent among those marrying for the second and third times (Chiswick and Houseworth, 2011).

In addition to early life experiences, acceptance of different cultures might result from events and experiences occurring later on in life. Participation in the Armed Forces may broaden people's worldview as a result of being stationed overseas or sharing barracks with Americans of different ethnicities within

the US (Fryer, 2007). Military experience is indeed positively correlated with out-marriage (Furtado and Theodoropoulos, 2011; Chiswick and Houseworth, 2011).

Education may also give rise to a more worldly view which might make people more comfortable sharing a home with someone with different cultural attributes. While the highly educated tend to be less prejudiced toward ethnic and racial minorities (Hello et al., 2002), researchers have found conflicting evidence on the relationship between education and the strength of ethnic identity (Duncan and Trejo, 2011; Feliciano, 2009). After accounting for other mechanisms through which schooling affects marriage, researchers using data from recent years provide evidence that education decreases preferences for ethnic endogamy (Furtado and Theodoropoulos, 2011; Kalmijn, 2012), but Furtado (2012) does not find any evidence of this relationship using 1970 data.

Preferences for Related Characteristics

High rates of ethnic endogamy may not be driven by ethnic preferences per se, but by preferences for other characteristics that are correlated with ethnicity. For example, given the large role played by family in the rituals and practices of many religions, it may not be surprising that religious homogamy rates are quite high (Sherkat, 2004). Preferences for a spouse with the same religion can ultimately result in high rates of ethnic endogamy given the strong relationship between ethnicity and religion. It may be challenging, for example, for a Hindu spouse-searcher to find a non-Indian Hindu.

It is difficult to analyze the role of religion in interethnic marriage decisions because of the lack of information on religion in most nationally representative surveys in the US. However, the percentage of the country of origin that is, Christian has been found to decrease endogamy among childhood arriving immigrants and the native-born children of immigrants in the US (Kalmijn and van Tubergen, 2010). Using data from Sweden, Dribe and Lundh (2011) provide evidence that several cultural factors, including religion, play a role in explaining endogamy rates.

While religious homogamy is generally decreasing (Sherkat, 2004), assortative matching on education has increased since the 1960s (see Schwartz and Mare, 2005, for a discussion of potential explanations). Given the importance of matching on education and the variation in average schooling levels across ethnic groups, individuals with education levels that are typical within their ethnic group may end up marrying endogamously simply because it is relatively easy to meet a coethnic with a similar level education. On the other hand, ethnics with education levels that deviate from the norm within their ethnicity are likely to find it more difficult to find a same-ethnicity and same-education spouse and so compromises must be made.

Several studies have found that the closer a person's own education is to his ethnic group's average, the more likely he is to marry endogamously (Chiswick and Houseworth, 2011; Kalmijn and van Tubergen, 2010). Furtado (2012) and Furtado and Theodoropoulos (2011) show that an increase in schooling leads to a decrease in the probability of marriage within ethnicity for people in low-education groups but an increase for those in education ethnicities. Consistent with this evidence, Kalmijn (2012) shows that education increases out-marriage most in low-education groups and least in high-education groups. Interestingly, he never finds that education increases endogamy, not even in the most highly educated groups. He hypothesizes that, as discussed above, education tends to decrease in-group preferences so that while the highly educated members of high-education groups have access to same-education, same-ethnicity spouses, they may not value ethnic endogamy to the same degree as those with less education.

Schooling may also affect marriage patterns via the preferences of marriage-market participants who do not identify with any particular ethnic background. For example, it may be that, as predicted by exchange theory, people are only willing to marry ethnic minorities, especially those with strong ethnic attachments, if they are compensated in the form of a spouse's higher education (Gullickson and Fu, 2010). It may also be that ultimate marriage patterns are driven by the majority population's perceptions of ethnic minorities, perceptions that are likely to be influenced by average schooling levels of the ethnic group. Kalmijn (2012) finds that, conditional on a person's own education, belonging to a group with a higher average education is associated with an increase in the likelihood of out-marriage. He interprets this result as evidence of the importance of natives' perceptions of a person's ethnic group.

Marriage Market Characteristics

Besides preferences, intermarriage patterns are driven by opportunities to meet people from different ethnic groups. It is certainly easier for a person belonging to a large ethnic group to find an attractive match within ethnicity. Indeed, whites, the majority race in the US, have higher racial endogamy rates than any other group despite the fact that when racial minorities out-marry they typically marry whites (Wang, 2012). Opportunities to meet same-ethnicity mates are also driven by residential segregation patterns. The decline in Hispanic intermarriage in the 1990s can be explained partially by increases in Hispanic residential segregation (Lichter et al., 2007).

In order to take into account both the size of the ethnic group and residential segregation, studies typically look at the impact of the proportion of a person's neighborhood or city that is of the same ethnicity. Studies consistently find a strong association between this variable and the probability of in-marriage (Chiswick and Houseworth, 2011; Kalmijn and van Tubergen, 2010), but caution is required when interpreting this result since people with stronger preferences for endogamy are likely to choose to live around coethnics. In addition, residential segregation can reinforce group solidarity and ethnic identity (Lichter et al., 2007). Regardless of what exactly drives the relationship between ethnic group size and endogamy, it is important to keep in mind the distinction between opportunity and preferences when interpreting changing endogamy rates over time or comparing endogamy rates across ethnic groups.

Another marriage market characteristic that drives intermarriage patterns is sex ratios. Males in groups with larger male to female ratios in the US are likely to find it more difficult to

meet suitable same-ethnicity spouses whereas the opposite would be true for females in these groups. Sex ratios have been found to predict interethnic marriage for both first- and second-generation immigrants (Angrist, 2002; Chiswick and Houseworth, 2011), but this relationship is not as robust as that of between ethnic group size and endogamy as some studies find no effect (Kalmijn and van Tubergen, 2010).

Variation in the availability of same-ethnicity potential spouses may also arise from changes in the flow of new immigrants to the US. A steady immigrant inflow from Mexico in the past century provides many potential same-ethnicity partners for Mexicans of all generations (Lichter et al., 2007). New immigrants also help sustain institutions, such as churches and newspapers, which facilitate social interactions within ethnicity. Stevens and Swicegood (1987) find that belonging to an ethnic group with many non-English speakers, either because of a large number of immigrants in the group or because second- and later generations continue to speak their origin language, increases endogamy rates even among people whose first language is English. They attribute this result to the role of language in shaping socialization experiences and ethnic identification regardless of one's own preferred language.

Consequences of Intermarriage

Interethnic marriage patterns are often used as measures of assimilation not only because more assimilated groups are more likely to intermarry but also because intermarriage results in further integration. While it is true that those with better English-speaking skills are more likely to marry outside of ethnicity, sharing a household with a native English speaker is likely to further improve a person's English proficiency. In a similar way, immigrants married to natives are likely to learn more about US customs and traditions.

Intermarriage affects those directly involved in the marriage as well as the communities surrounding these marriages. Children of intermarried couples are less likely to identify with an ancestry (Duncan and Trejo, 2011). Moreover, an intercultural marriage typically brings together two sets of families, friends, and acquaintances that most likely have different cultural attributes. This ethnic commingling breaks down barriers between ethnic groups.

Several studies have found a positive relationship between marriage to a native and wages of immigrants in several different countries (Meng and Gregory, 2005, for Australia; Meng and Meurs, 2009, for France; Kantarevic, 2004, for the US, Çelikaksoy, 2007, for Denmark). Since the natives that immigrants marry are most often from different ethnicities (Furtado and Theodoropoulos, 2010), we can think about these marriages as interethnic. Furtado and Theodoropoulos (2010) find that immigrants arriving in the US before age 19 are more likely to be employed if they are married to a native than if they are married to another immigrant.

The problem with interpreting these relationships is that it is unclear whether marriage decisions directly affect worker productivity and job matching or, as discussed in the previous section, whether productive immigrants with better networks are more likely to marry natives. Studies that have taken steps to correct for this by focusing on people compelled to out-marry as a result of marriage market conditions instead of personal preferences have generally continued to find positive effects of marrying a native (Meng and Gregory, 2005; Furtado and Theodoropoulos, 2010). Furtado and Theodoropoulos (2010) examine the mechanisms through which marriage to a native increases employment rates of immigrants. They find little to no evidence that the returns to marrying a native are driven by undocumented immigrants' gaining the right to work legally in the US but provide several pieces of evidence suggesting that connections to native social networks play an important role for all childhood immigrants.

Intermarriage patterns also have important implications for the interpretation of the most commonly used measures of intergenerational assimilation in the US. While the CPS asks questions about parents' countries of birth, 1970 was the last time the Census asked for this information. No national survey asks for grandparents' countries of birth. Thus, when attempting to measure intergenerational assimilation beyond the second generation, researchers typically rely on self-reports of ancestry or ethnic background. This may result in biased estimates of intergenerational mobility, especially among descendants of mixed marriages, if the likelihood of identifying with a particular ancestry is associated with a person's education and labor market success. Duncan and Trejo (2011) present evidence that high-skilled Mexicans are less likely, relative to low-skilled Mexicans, to marry other Mexicans and that children of these mixed marriages are less likely to identify as Mexican. These two facts, in conjunction with intergenerational transmission of skill, imply that Mexicans in third and higher generations will be more likely to self-report Mexican ancestry if they are low skilled. This is exactly what they find in the data suggesting that the commonly constructed estimates of intergenerational assimilation for Mexicans are too low.

Conclusion

As argued throughout this article, interethnic marriage rates can be interpreted as a measure of social distance between groups. Not only is intermarriage more likely when groups are more integrated into mainstream society, but marriage further integrates ethnic groups. In fact, intermarriage has been referred to as the spoon that stirs the ethnic melting pot (Lichter et al., 2007).

Recent statistics suggest that intermarriage rates, although increasing, are still quite low. Perhaps even more interesting is the variation across ethnic groups in the likelihood of intermarriage. While it may be tempting to interpret changes over time and differences across groups solely in terms of preferences, the literature suggests that many structural factors drive endogamy rates. Regardless of the causes of intermarriage, a union of two people with different ethnic backgrounds begets many social interactions between the couple's social circles resulting in further breakdowns of ethnic barriers. Children of interethnic marriages tend to have weaker ethnic identities and are themselves more likely to marry outside of ethnicity. Thus, further analysis of the causes and consequences of ethnic intermarriage is certainly warranted given its role as

a barometer of current levels of assimilation and ethnic relations as well as a forecaster of future levels of immigrant integration.

See also: Ancestors, Anthropology of; Assimilation of Immigrants; Assortative Mating in the Marriage Market; Ethnic Conflict, Geography of; Ethnic Identity and Ethnicity in Archaeology; Ethnic Identity, Psychology of; Ethnicity and Ethnic Groups: Historical Aspects; Immigration Policy; Immigration, Citizenship, and Integration: Social Work Connections; Immigration: Political Aspects; Marriage; Migration: Cultural Aspects.

Bibliography

Angrist, J., 2002. How do sex ratios affect marriage and labor markets? Evidence from America's second generation. The Quarterly Journal of Economics 117, 997–1038.

Becker, G.S., 1973. A theory of marriage: part I. Journal of Political Economy 81, 813–846.

Bleakley, H., Chin, A., 2010. Age at arrival, English proficiency, and social assimilation among U.S. immigrants. American Economic Journal: Applied Economics 2, 165–192.

Chiswick, B., Houseworth, C., 2011. Ethnic intermarriage among immigrants: human capital and assortative mating. Review of Economics of the Household 9, 149–180.

Çelikaksoy, A., 2007. A wage premium or penalty: an analysis of endogamous marriage effects among the children of immigrants. Danish Journal of Economics (Nationaløkonomisk Tidsskrift) 145, 288–311.

Davis, K., 1941. Intermarriage in caste societies. American Anthropologist 43, 376–395.

Dribe, M., Lundh, C., 2011. Cultural dissimilarity and intermarriage: a longitudinal study of immigrants in Sweden 1990–2005. International Migration Review 45, 297–324.

Duncan, B., Trejo, S.J., 2011. Intermarriage and the intergenerational transmission of ethnic identity and human capital for Mexican Americans. Journal of Labor Economics 29, 195–227.

Feliciano, C., 2009. Education and ethnic identity formation among children of Latin American and Caribbean immigrants. Sociological Perspectives 52, 135–158.

Furtado, D., 2012. Human capital and interethnic marriage decisions. Economic Inquiry 50, 82–93.

Furtado, D., Theodoropoulos, N., 2010. Why does intermarriage increase immigrant employment? The role of networks. B.E. Journal of Economic Analysis & Policy 10. Article 101.

Furtado, D., Theodoropoulos, N., 2011. Interethnic marriage: a choice between ethnic and educational similarities. Journal of Population Economics 24, 1257–1279.

Fryer, R.G., 2007. Guess who's been coming to dinner? Trends in interracial marriage over the 20th century. Journal of Economic Perspectives 21, 71–90.

Gordon, M.M., 1964. Assimilation in American Life. Oxford University Press, New York.

Gullickson, A., 2006. Education and black/white interracial marriage. Demography 43, 673–689.

Gullickson, A., Fu, V.K., 2010. Comment: an endorsement of exchange theory in mate selection. American Journal of Sociology 115, 1243–1251.

Hello, E., Scheepers, P., Gijsberts, M., 2002. Education and ethnic prejudice in Europe: explanations for cross-national variances in the educational effect on ethnic prejudice. Scandinavian Journal of Educational Research 46, 5–24.

Kalmijn, M., 1998. Intermarriage and homogamy: causes, patterns, trends. Annual Review of Sociology 24, 395–421.

Kalmijn, M., 2010. Educational inequality, homogamy and status exchange in black-white intermarriage: a comment on Rosenfeld's critique. American Journal of Sociology 115, 1252–1263.

Kalmijn, M., 2012. The educational gradient in intermarriage: a comparative analysis of immigrant groups in the United States. Social Forces 91, 453–476.

Kalmijn, M., van Tubergen, F., 2010. A comparative perspective on intermarriage: explaining differences in marriage choices among national origin groups in the United States. Demography 47, 459–479.

Kantarevic, J., 2004. Interethnic Marriages and Economic Assimilation of Immigrants. IZA Discussion Papers 1142. Institute for the Study of Labor (IZA).

Lam, D., 1988. Marriage markets and assortative mating with household public goods: theoretical results and empirical implications. Journal of Human Resources 23, 462–487.

Lichter, D.T., Brown, J.B., Qian, Z., Carmalt, J.M., 2007. Marital assimilation among Hispanics? Evidence of declining cultural and economic incorporation? Social Science Quarterly 88, 745–765.

Meng, X., Gregory, R.G., 2005. Intermarriage and the economic assimilation of immigrants. Journal of Labor Economics 23, 135–176.

Meng, X., Meurs, D., 2009. Intermarriage, language, and economic assimilation process: a case study of France. International Journal of Manpower 30, 127–144.

Merton, R.K., 1941. Intermarriage and social structure: fact and theory. Psychiatry 4, 361–374.

Qian, Z., 1999. Who intermarries? Education, nativity, region, and interracial marriage, 1980 and 1990. Journal of Comparative Family Studies 30, 579–597.

Qian, Z., Glick, J., Baston, C., 2012. Crossing racial, ethnic, and nativity boundaries: ethnic differences in marriage and cohabitation. Demography 49, 651–675.

Rosenfeld, M.J., 2005. A critique of exchange theory in mate selection. American Journal of Sociology 110, 1284–1325.

Rosenfeld, M.J., 2010. Still weak support for status-caste exchange: a reply to critics. American Journal of Sociology 115, 1264–1276.

Sherkat, D.E., 2004. Religious intermarriage in the United States: trends, patterns, and predictors. Social Science Research 33, 606–625.

Schwartz, C.R., Mare, R.D., 2005. Trends in educational assortative marriage from 1940 to 2003. Demography 42, 621–646.

Stevenson, B., Wolfers, J., 2007. Marriage and divorce: changes and their driving forces. Journal of Economic Perspectives 21 (2), 27–52.

Stevens, G., Swicegood, G., 1987. The linguistic context of ethnic endogamy. American Sociological Review 52, 73–82.

Wang, W., 2012. The Rise of Intermarriage: Rates, Characteristics Vary by Race and Gender. Pew Research Center Social and Demographic Trends, Washington, DC.

Ethnic, Racial, and Nationalist Movements

Susan Olzak, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ethnic, racial, and nationalist social movements organize around distinct ethnic identities, but they differ with respect to the scope of their claims, activities, and goals. Ethnic nationalist movements can be distinguished empirically from ethnic/racial movements by the presence of demands claiming legitimate rights to sovereignty and/or authority to administer a specific territory. As a result, nationalist movements are likely to come into conflict with existing state authorities and the international system. Theories explaining the rise, persistence, and decline of these movements vary with respect their emphasis on cognitive, rational, or instrumentalist motivations for such movements, and to the degree these movements come in conflict with state authorities.

Introduction and Definitions

Though they are commonly analyzed separately, ethnic, racial, and nationalist (E/R/N) movements voice strikingly similar claims of sovereignty rights and ethnic integrity. Race is conventionally defined in terms of putative biological or inherited features, while ethnicity tends to refer to cultural and acquired traits such as language, religion, and custom. Various combinations of claims, goals, tactics, and organizational forms distinguish different types of E/R/N movements. This section first establishes some working definitions for ethnic and nationalist social movements, lists various forms of E/R/N movements, and then briefly reviews several relevant research traditions.

Social movements generate collective action advocating fundamental changes in the political or economic arrangements in a society. Social movements typically involve sustained activity over time and place (whereas collective action may be fleeting). Most scholars also find that adherents of a social movement tend to support a coherent ideology that defines its core identity.

The defining features of ethnic and racial social movements (E/R) are that claims are made based upon particular identity or boundary, defined by the presence of racial or ethnic markers (Barth, 1969). These markers typically include skin pigmentation, ancestry, language, and a history of discrimination, conquest, or other shared experience. Recent research tends to subsume racial and ethnic movements under the label of ethnicity. This is due to the fact that distinct biological characteristics do not easily allow observers to distinguish among categories that are fluid, fuzzy, and permeable. For simplicity (and to avoid invoking unscientific assumptions about the genetic basis of race), most researchers prefer the more generic label of ethnicity (Brubaker, 2004).

Successful challenges from E/R/N movements depend upon levels of group solidarity and mobilization capacity. In order to examine the effect of these factors on movement outcomes, it is necessary to distinguish a movement's mobilization capacity apart from ethnic solidarity. Solidarity is characterized as the conscious identification (and loyalty) with a particular race or ethnic population, measured by attitudes, institutional involvement (or organizational participation), and monitoring capacity. Mobilization is the capacity to harness resources (including solidarity, organizations, and material resources) in an effort to reach some collective goal. The US civil rights movement is a prototypical E/R social movement. As a template for collective action, the civil rights movement generated many subsequent movements (and countermovements), including the antiapartheid movement in South Africa, rights movements for Muslims in Western Europe, and many other settings (Andrews, 2013).

While race and ethnic movements make reference to a shared identity based upon cultural markers, phenotype, outward expressions of loyalty, celebrations, violation of minority rights, or in- and out-group labels, nationalism is a social movement making a territorial claim. According to Hechter (2000), nationalist movements articulate claims for self-determination and authority over a specified territory and espouse grievances based upon the fact that these demands are not now being satisfied. Ethnic nationalism also includes references to shared ethnic identity as a basis for these demands.

The pursuit of sovereignty rights typically also provokes conflict (and perhaps also warfare) with existing regimes. Such conflict can remain quiescent for long periods of time, erupting suddenly into full-blown armed guerrilla warfare, or civil war, depending on regime strength, outside support, primary export commodities, internal mobilization of resources, and reaction by state authorities to nationalist movements (Fearon and Laitin, 2003, 2005).

However, nationalist movements do not always organize around race or ethnic identity. Instead, they may rest upon a group's geographical concentration, political jurisdiction and leadership, and/or legacy claims to legitimate authority. Some scholars refer to this as 'civic nationalism,' in which national identity crosscuts ethnic identities. Civic nationalism is often characterized as more inclusive, as in multiculturalism (but see Brubaker, 2004).

Brubaker (1996: pp. 6–12) categorizes nationalism as one of the three types of collective action mobilized by national minorities, nationalizing states, and external national homelands. This tripartite definition has the advantage of analyzing nationalism as a confrontation with an existing

regime, empire, colonial power, or host nation. Put differently, ethnicity becomes transformed into nationalism when it makes specific historical claims and attempts to administer the group as a political community.

Social movement perspectives add the insight that E/R/N movements can be further distinguished by their relative duration, target, tactics, violence, and audience. These distinctions yield eight broad categories: (1) autonomy movements that demand special rights over a particular territory; (2) civil rights protests that demand expansion of a group's civil and economic rights or demand an end to discrimination; (3) antagonist movements directed against specific ethnic targets, including collective attacks ranging from genocide, ethnic cleansing, and mob violence, to symbolic threats; (4) state-strengthening nationalism, which attempts to unify diverse cultures (state-building nationalism) or merge politically divided territories into one state (unification nationalism); (5) secession movements claiming rights of withdrawal from formal state authority (to join or form a newly independent state); (6) separatist movements demand their own state authority which involves replacing existing state authority and administrative powers; (7) irredentist movements are those that claim rights over a territory currently administered by another state authority; and (8) genocide or ethnic cleansing, which is an extreme form of violent social movements against a target population which aims to either annihilate or remove an entire population.

While these definitions clarify some important distinctions, the application of these terms often becomes tricky when conducting research on E/R/N movements because race and ethnic boundaries are porous, dynamic, and flexible. From different political vantage points and at different times, the same movement may be seen as engaging in senseless violence or as a nationalist liberation movement. Moreover, movements adopt new goals and tactics to adjust to shifts in the political environment. Such changes by movements may engender fears that a movement's original authenticity has become compromised or coopted. While these transformative qualities of social movements undoubtedly create difficulties for researchers, there is a hidden advantage as well. Taking these shifts into account might reveal new information about how group identity becomes transformed into social movements.

Core Research Questions

Questions about the origins, persistence, and success of movements drive the majority of all social science research efforts in this area. First, how does ethnic, racial, religious, territorial, or national identity become transformed into active social movements? This question underscores the importance of maintaining a distinction between cultural or symbolic expressions of identity (e.g., ethnic self-help organizations, immigrant festivals, head scarves) from social movements that express claims for expanded rights to some set of authorities, or violence directed against specific groups. Accordingly, scholars ask under what conditions will specific boundaries (e.g., language vs. skin pigmentation) come to be politically activated. In other words, it is important to frame the salience of ethnicity in social movements as something to be explained.

Second, are the factors explaining the emergence the same as those that encourage the persistence of E/R/N movements? Until recently, social scientists have focused mainly on the internal characteristics of states to explain E/R/N movements. Thus poverty, rough terrain, imposition of direct rule, warlord corruption, or some other structural feature of the political system or economy triggers have been identified as factors raising levels of insurgency (Fearon and Laitin, 2003). Others have drawn attention to the advantages of viewing social movements as a function of global and transnational mechanisms that encourage ethnic insurgency (Olzak, 2006, 2011). Still others emphasize the importance of the existence of ethnic grievances driven by ethnic exclusionary practices directed against less powerful minorities (Esteban et al., 2012; Wimmer, 2002).

A third question asks, are ethnic movements truly novel, or are they simply political movements that once, earlier in history, adopted other forms? Without undertaking a long historical perspective, answering this question is difficult. However, it seems plausible that social movements are now more likely to be couched in distinctly ethnic terms, as a function of self-determination norms and UN declarations on minority rights. Instead, for most scholars, evidence that these movements are modern rests on emergence of a shared identity of a 'people' with boundaries beyond a parochial village or kinship group (Anderson, 1991; Brubaker et al., 2006). The irony is that many 'modern' nationalist movements (e.g., Islamic nationalist movements) have invoked themes demanding return to the past.

A fourth orienting question concerns the shift in the scope of activity: What are the mechanisms that cause E/R/N movements to expand their scope from local concerns to encompass national goals? Brubaker (1996) describes 'nationness' as an institutional process that begins to crystallize with state-expansion. Hechter's (2000) theory asserts that nationalist movements reacted to the implementation of direct rule by an external colonial or imperialist power. Global perspectives suggest that the density of international connections among states may be reducing the significance of social movements bounded by national borders (Olzak, 2006).

Another set of issues raise questions about the nature of the relationship between the existence of ethnic divisions and the likelihood of internal civil war. For instance, they ask, under what conditions does the activation of ethnic boundaries promote armed conflict or separatist insurgencies? Several lines of research have suggested that there is a strong link between ethnic cleavages and violence, in which group differences mobilize and sustain the capacity for groups to incite civil wars. For instance, Sambanis (2001) finds that civil wars based upon ethnic and/or religious identities are more likely to erupt in countries with high levels of ethnic heterogeneity and low levels of political democracy (but see Fearon and Laitin, 2003).

Scholars have also explored the reverse causal argument. In this view, economic (or political) instability results from prior conflict, or that economic decline follows the public's anticipation of civil unrest. In an attempt to sort out the causal ordering of economic effects on ethnic wars, current empirical evidence supports the notion that economic decline raises rates

of internal civil war, rather than the reverse, but that political instability and state strength may be endogenous to the process of ethnic and nationalist mobilization (see Olzak, 2004).

While the evidence connecting ethnic diversity to the onset of armed conflict has been equivocal, ethnic diversity may prolong the duration of civil wars. Fearon and Laitin (2003) find that the duration of violent civil conflict increases when there are a small number of large ethnic groups, when there are conflicts over land use, and when rebels have access to external (or contraband) resources.

Leading Theoretical Perspectives

Perspectives offering explanations of the emergence, growth, and decay of E/R/N social movements emphasize one or more processes of changing economic, political, or social conditions. Each tradition has generated a number of important empirical studies, which are linked together by common theoretical concepts and mechanisms.

Colonialism and Ethnicity

The legacy of colonialism provides a number of instructive lessons for understanding the emergence and timing of nationalist movements. Territorial boundaries drawn during periods of colonialist rule (especially in Africa and the Middle East) provide examples of how colonialism encouraged ethnic cleavages that subsequently form the basis of independence movements based upon ethnic nationalism. During periods of state-formation, outcomes depend upon complicated negotiations between opponents, nation-builders, and external participants. As norms of self-determination gained momentum internationally, colonialist regimes became perceived as wielding illegitimate power over the indigenous population. As a result, nationalist movements of independence often adopted an ethnic character, often in contradistinction to colonialist racial/ethnic identities.

E/R/N movements are fundamentally embedded in (often contradictory) legends and myths about various group identities and actions that have shaped their histories. Language, religion, immigration, and migration histories all play a role in building the defining characteristics of a region. However, periods of nation-building that coincide with independence movements apparently play a central role in determining the nature of identity of an imagined 'nation.'

State- and Nation-Building Processes

Theorists once assumed that the processes of state-building (obtaining legitimate authority and administrative rights to govern a territory) and nation-building (building a sense of shared identity in a population) should be analyzed as an evolutionary set of stages, such assumptions seem naive today. Evidently the process of creating a legitimate nation with an accepted system of authority and leaders is better conceptualized as a dynamic set of negotiated meanings (Brubaker, 1996). National identities and ethnic communities are constantly being reconstructed and boundary lines redrawn.

Anderson (1991) provided a useful starting place for understanding why nationalist and ethnic movements aim to reconcile the lack of correspondence between state boundaries and national identity. Hechter (2000), Fearon and Laitin (2003), and many others have emphasized the fact that few (if any) nation-states are homogeneous entities; not only do states sometimes encompass many nations (as in the notion of multiculturalism) but national movements can exist without espousing goals of statehood. If a 'nation' is demarcated by a self-identified boundary, then one nation may be dispersed across multiple state boundaries (as in the concept of a Kurdish nation). Such movements may eventually express goals of creating a new state organized around a common national identity. However host states will nearly always resist these movements, often with force. This implies that, even if they are only temporarily successful, ethnic movements can undermine attempts at state-building that assume that state and nationality boundaries are coterminous.

The literature on state-formation has suggested that ethnic movements are most likely to turn violent during early periods of state-formation, when issues of power and authority remain unresolved (Hechter, 2000). E/R/N movements are likely to confront newly independent (or newly democratic) regimes that have not completely won the hearts and minds of all inhabitants. Outcomes depend upon a complicated power balance among opponents, nation-builders, or external states, each of whom may have conflicting interests.

Initial periods of state- and nation-building also play a central role in determining the nature of the identity of an imagined 'nation' (Anderson, 1991). In this view, nations are 'birth marked' by prior conflicts – marked by religion, region, ethnicity, anticolonialism, or conquest – that prevailed during a formative historical period. Violence is especially likely when state-building efforts activate resistance movements based upon these historic cleavages.

Economic and Instrumental Approaches

This tradition rests on the assumption that ethnic movements are better analyzed in terms of politically and/or economically motivated demands rather than instigated by primordial ethnic identities (Blattman and Miguel, 2010). One prominent version of this line of argument suggests that ethnic social movements are based upon shifts in economic inequalities among ethnic groups. Other instrumentalist positions emphasize that ethnic movements are led by social movement leaders who use ethnic identities to gain leverage over competitors in contests for political power.

Competition Theories of Race and Ethnic Relations

Competition theories of race/ethnic movements emerged to explain the puzzle prompted by the empirical failure of modernization theory to account for the apparent rise in E/R/N movements in both developing and developed countries. Competition theories of race and ethnic relations suggest that development encourages competition to the extent that declines in ethnic and racial disadvantage occur. This in turn triggers efforts by dominant groups to exclude competitors, using both legal and illegal methods against minorities

(Olzak, 1992; Olzak and Shanahan, 2003). Lynching, ethnic cleansing, and exclusionary laws are prominent examples of strategies used against minorities (Olzak, 2013).

Competition theory has been supported by empirical studies of Chinese laborers, United States labor movement, contemporary South Africa, analyses of postindustrial racial conflict in the United States, analyses of race and ethnic conflict in cities, and of lynching in rural areas in nineteenth-century United States, and the former republics of the Ukraine, Latvia, Estonia, and Lithuania (Beissinger, 2002; Olzak, 2004).

Rational-Choice and Game-Theoretic Approaches

Rational-choice theorists state that modern ethnic movements occur with regularity because ethnic groups lower monitoring costs and increase benefits attached to ethnic mobilization, allowing ethnic groups to overcome the free-rider problems attached to other types of mobilization efforts (Hechter, 2000). The US civil rights movement has also been analyzed within a rational-choice framework that emphasizes how mobilization that increases disruptions and concession costs to elites has proven decisive to successful movement outcomes (Luders, 2010).

Because ethnic groups are able to form dense social networks easily, costs of monitoring commitment are minimized, which fosters ethnic mobilization. Building on rational-choice models, Fearon and Laitin (1996) have linked the strategic aspects of ethnic identity to violence, as elites build on existing ethnic loyalties (see also Petersen, 2002). Such loyalties can prove fatal to group members. Moreover, the presence of genocidal norms (defined as a threat of sanctions to in-group members who decline participation in ethnic mayhem) increases the scale of ethnic violence. This perspective offers an explanation for one persistent and counterintuitive finding in the literature: Despite a history of intergroup cooperation, tolerance, intermarriage, and trust among different groups interacting within a region, the intensity of ethnic cleansing and violence can remain high, due to the presence of genocidal norms.

Similarly, theorists have extended these ideas to consider the implications of game theory for ethnic mobilization, including outbreak of ethnic war (Fearon and Laitin, 1996). While armed ethnic rebellions tend to last longer than nonethnic ones (Fearon, 2004), a variety of ethnic and cultural characteristics have few systematic effects on the onset or duration of civil wars in general (Fearon and Laitin, 2003).

Political/Institutional Approaches

Political perspectives emphasize the role of shifts in political constraints and opportunity structures that influence the trajectory of E/R social movements. These theories emphasize institutional arrangements, court rulings and reforms, and regional concentration of ethnic populations as viable political instruments leading to mobilization. Two studies from India illustrate these points. For instance, Chandra (2004) argues that political systems based upon ethnic patronage systems can inadvertently provide the foundation for permanent hostilities. Alternatively, Varshney (2002) finds that when business, civic, and voluntary associations integrate and/or crosscut ethnic lines, ethnic confrontations are significantly less likely to erupt (see also Horowitz, 2001).

In cross-national studies of the influence of political structures, some scholars have emphasized the centrifugal force of ethnic political parties, which maintains ethnic loyalties through institutional arrangements and patronage based on ethnic loyalties. Such forces produce fierce loyalties when language, religion, or some other marker also distinguishes a population that is geographically concentrated (Fearon and Laitin, 2003). Other scholars have emphasized proximate causes or triggering mechanisms, such as political changes in authority, collapse of colonial authorities or empires, or transition to market economies, which presumably increases ethnic inequality and promotes ethnic competition (Chua, 2003).

Political structures can be arrayed along a continuum that indicates the extent to which political parties become ethnic vehicles for contesting power. One view of ethnic party systems ('consociationalism') holds that these arrangements reduce ethnic tensions by channeling extrainstitutional mobilization and energy within institutionalized party contests. Alternatively, ethnic parties may instead actively encourage ethnic mobilization because they render ethnic boundaries more salient. Wilkinson (2004) finds empirical support for this argument, reporting that states with a high degree of institutionalization of ethnic parties in states in India produces significantly higher rates of ethnic violence and hostility (see also Varshney, 2002).

Weak States, Regime Instability, and Ethnic Mobilization

States often respond to challenges from nationalist movements by alternating between strategies of repression and concession (Hechter, 2000). For example, the evidence suggests that states shift from strategies of repression to accommodation, depending on both the virulence of dissident protest behavior and state capacity to repress these challenges. However, others have suggested the intriguing hypothesis that it is the vacillation of states itself that incites nationalist violence, signaling a weakness in the state's internal capacity to act (Olzak, 2004).

The potential for ethnic separatism also influences the intensity of collective violence in a country and this effect is stronger in states with weakened political institutions. Evidence from civil wars in Bosnia and Kosovo provide another example of how regime instability shapes opportunities for E/R/N movements. At the same time, policies that involve ethnic resettlement programs often concentrate ethnic populations and create new networks that provide new recruits for mobilizing ethnic violence, as examples from the West Bank in Israel, or the Kurds in Germany, suggest. Thus transitions to democracy may mobilize ethnic movements by offering new political advantages to ethnic groups that were more easily submerged in repressive regimes.

The imposition of external political authority on ethnic minorities compared to imposition of structures of indirect authority has important consequences for ethnic and nationalist movements. Hechter (2000) has argued that the seeds of nationalist movements are embedded in specific political structural arrangements in which a colonial or federated authority cedes formal authority to local leaders. Under such conditions, local elites are delegated political power and authority by centralized authorities, yet the power of local

elites is fundamentally based upon regional identities and loyalties. When central authority is weakened or challenged (by external events such as war, famine, or economic crises), or when central authority is withdrawn (as in the case of the Soviet Union), local elites can mobilize on the basis of regional/ethnic identity.

Institutional Forms of Ethnic Exclusion

Wimmer et al. (2009) emphasize an argument that suggests that the form E/R/N movements will take depends upon a regime's formal institutional arrangements that exclude or privilege members of ethnic populations. Exclusion can take the form of formal disenfranchisement (as in the case of South Africa under apartheid), or more discriminatory patterns of ethnic favoritism, clientelism, and other forms (Wimmer, 2002). Research consistent with this perspective suggests that ethnic conflict peaks under conditions in which exclusionary practices, demographic polarization, and ethnic inequality interact (Esteban et al., 2012; Blattman and Miguel, 2010).

Brubaker (2011) reflects these themes in his work on ethnonationalism in former Soviet states. While some have argued that ethnic tensions and nativist movements have arisen as Soviet power and authority (and resources) related to Cold War alliances declined in importance, Brubaker (2011) argues that the federated system of regional and ethnically defined republics in the Soviet Union created the structural basis for ultimate disintegration of the Soviet Union.

Cultural and Cognitive Perspectives

Recent theoretical analyses emphasize both the cultural and cognitive components of social movements, suggesting that group identity is both an important mobilizing strategy and a consequence of mobilization. In particular, movements invoke one or more cultural themes of nationalism, rights of self-determination, expansion of human rights, and basic rights of sovereignty (Brubaker, 2004). Another line of argument within this cognitive tradition emphasizes the causal role of emotion in producing and escalating ethnic violence. In this view fear, resentment, and revenge play a key role in determining the conditions under which widespread ethnic cleansing and rioting occur (Horowitz, 2001; Petersen, 2002). Scholars from this tradition emphasize the importance of historical contingencies that seek to identify the particular historical outcomes that are shaped by broad cultural factors as well as idiosyncratic twists of fate.

An advantage of this view is that ethnic identity can be analyzed as the result (rather than the cause) of collective action. This insight opens up the possibility for studying how social mechanisms of contact, conflict, borrowing, and other forms of interaction cause new ethnic or racial categories to emerge. Interesting questions about the collective nature of conflict arise. For example, as group violence and revenge escalates on either side to a conflict, small-scale or individual skirmishes became redefined as events requiring a collective response. In analyzing forces escalating group conflict, scholars from the cognitive tradition emphasize the emergent properties of both identities and conflict.

Research guided by this tradition tends to focus on ethnic social movements that did occur, which limits the explanatory power of such studies. Another drawback of a purely cognitive perspective is that it is difficult to determine the causal ordering of emergent group identity and ethnic mobilization.

Globalization and the Internationalization of Ethnic/Nationalist Movements

Economic and political crises that once affected only local areas now have repercussions in vastly different and formerly unconnected regions and states. Since the advent of modern media, civil wars, terrorist acts, and acts by ethnic social movements have produced reactions across national borders. Taking an international perspective helps clarify how economic interdependence within states has reinvigorated ethnic politics.

Transnational Ethnic Social Movements

Transnational ethnic social movements are social movements that span multiple national borders, target forces of global integration, or are social movements that concern global-level issues (e.g., global environmental concerns). Regional associations such as the European Union (EU), Organization of the Petroleum Exporting Countries, North Atlantic Treaty Organization, and other supranational organizations promote interstate migration and decrease reliance of regions within states on the military and economic power of the nation-state. Multistate organizations also provide an audience for insurgent groups demanding new sovereignty rights (Koopmans and Statham, 2000). In this view, an increasingly dense network of international economic relations, exemplified by multinational corporations, growing trade and foreign investment, and supranational economic associations, will continue to produce more large-scale ethnic movements.

International Governmental Associations

One (perhaps unanticipated) consequence of the integration of the EU is that ethnic tensions have risen rather dramatically (Koopmans and Statham, 2000). Furthermore, as political and economic barriers have declined, labor (and capital) flow move more freely across states. However, one potential consequence of increasing flows of foreign workers across borders has been that new right-wing parties across Western Europe have mobilized sentiment against foreign workers. To the extent that the integration of the European Union has restructured local politics within European countries, the opportunity has arisen for ethnic politics on both sides of the immigration question. As a consequence, antiforeigner sentiment, nationalist political parties, and attacks on foreigners also appear to be rising in most Western European countries, especially in Germany, France, the Netherlands, and England (Koopmans and Olzak, 2004).

This perspective suggests that individual states will become less powerful in negotiations when confronting nonstate actors and/or transnational movements than in the past. As state economies and politics become more integrated, international

associations and events occurring outside state boundaries will become increasingly salient. It seems likely that as integration of the world's states (politically, diplomatically, and economically) proceeds, ethnic groups become less constrained by their own state authorities. The growing predominance of an integrated set of states ironically decreases the ability of any one state to direct all of its internal affairs. Moreover, highly integrated nation-states cannot simply repress, jail, or torture the ethnic challengers, without risking international condemnation, sanctions, and boycotts. Furthermore, neighboring countries may directly or indirectly finance campaigns of instability, using political refugees or exiles as mercenary soldiers. There is growing evidence on transnational environmental and human rights movements that supports these contentions (Olzak, 2006).

Diffusion of ideologies, resources, and personnel has accelerated these trends. Ethnic social movements occurring in neighboring countries have powerful diffusion properties, destabilizing or threatening nearby regimes. Sambanis (2001) argues that elite factions (or warlords) offering military and financial support from neighboring countries have played crucial roles in prolonging ethnic wars in Africa and Central Asia in recent years. Although it is difficult to study (because many of the transactions are clandestine and sources of data are unreliable), corruption feeds upon an increased flow of arms, mercenaries, and illegal drugs. It is likely that networks of local warlords also fuel ethnic wars (without state or international sanctions) in neighboring countries (Fearon and Laitin, 2003; Fearon, 2004).

Globalization and Ethnic Conflict

A global strategy offers arguments about forces of globalization that produce inequality, competition, and mobilization. Olzak (2006) holds that integration of a world economic and political system has encouraged ethnic fragmentation within states. It does so by increasing access of formerly disadvantaged groups to political resources, creating new political opportunities for mobilizing, and increasing levels of economic inequality in peripheral countries, which increases the potential for competition and conflict among groups within these states. As the world's states have become more directly linked through communication and media channels, information about inequality and claims for redress of this inequality has increased sharply. Thus, global forces of integration tend to crystallize and empower local level cleavages, increasing solidarity and heighten the capacity to mobilize movements challenging state authority (Olzak, 2011).

Global perspectives suggest another way that an ideology supporting human rights has accelerated the acceptance of an ideology supporting minority rights, especially ethnic ones. In this view, as the worldwide human rights movement gained momentum, claims for national sovereignty, group rights, and freedom became intertwined. In this view, as nation-states became linked together in networks of military and economic associations, national political boundaries weaken and political regimes become vulnerable to international and external challenges. The same forces that encouraged the diffusion of nationalism as an ideology also affect ethnic movements within and between state boundaries.

Military interdependence constitutes an obvious way that international relations affect conflicts within countries. Although such strategies are not new, superpowers arm and train ethnic and subnational groups in order to stabilize or in some cases destabilize regimes. The cases of rebellions financed and supported by transnational forces (on both sides of the struggles) in civil wars in Afghanistan, Nicaragua, Vietnam, Sudan, Iraq, and many other settings illustrate this point. If arguments suggesting a link between international networks and E/R/N movements are correct, then an increase in transnational processes could potentially weaken the viability of the nation-state as a main organizing strategy for territorial authority and control.

Pan-Ethnic/Religious Movements

Reactive movements are based upon pan-ethnic (or ethno-religious) claims that span national borders. Such movements often espouse ideologies that contrast with Western models of the so-called liberal state. In the contemporary era, some of these movements have been characterized as loosely connected by antimodernist claims that have a basis in both ethnic identity and religious fundamentalism. These movements espouse a form of identity politics (such as the wearing of head scarves), but they diverge from typical nationalist movements because they rarely express specific claims to govern specific territories. Whether or not such movements will continue to gain momentum remains an open question.

Evaluation of Various Perspectives

Many scholars analyze different types of E/R/N movements separately, by historical period, regional groupings, specific goals, or tactics. While it has become popular for scholars to claim that E/R/N movements are produced by a constellation of historically contingent factors, this strategy has hampered our ability to develop powerful explanations that identify some general mechanisms of social change. Yet other social scientists have demonstrated that there are substantial theoretical payoffs attached to analyzing the similarities across E/R/N movements. By paying attention to the commonalities among forms of ethnic and nationalist movements, we stand to gain more leverage over questions about how protest escalates and diffuses, or how spontaneous protests become transformed into sustained (or violent) social movements that challenge existing authority structures.

Approaches that seek to emphasize the continuities and discontinuities among social movements and their emergent forms allow cumulative and testable theories to be constructed and evaluated. In contrast, if multiple types of ethnic mobilization (from civil rights movements to ethnic civil wars) are analyzed separately by country, time period, and movement goals, it becomes impossible to know when to stop creating new categories and crafting unique explanations to cover each new occurrence of a nationalist event or ethnic campaign. Truly comparative work that seeks to build theories that can be falsified empirically holds far more promise than does a strategy that views each

movement as a unique and separate category. In the context of complex dynamics of transnational social movements, activists and organizations engage in an activity that engages and activates ethnic identity within nations. Such reactive movements can have reverberations to kindred or disaporic groups beyond a single nation. Understanding the commonalities among these forms in the context of a global system seems especially relevant.

Assessing the relative importance of various causal factors explaining ethnic movements over time is also hampered by the fact that ethnic and racial boundaries (and labels) often change over time. Answers to questions about the nature and trajectory of ethnic movements lie in conducting careful empirical analyses and comparisons of different kinds of events – ethnic, civil rights, national, religious, civil wars, and autonomy social movements of various kinds that share some (but not all) root causes.

Recent trends have taken these dynamics into account by emphasizing the emergence of a more densely connected global system. The globalization of social movements has led current research on ethnic and national movements away from a sole emphasis on internal features of states and toward the international context of collective actions. Clearly the internationalization of a world economy and political integration of organizational, diplomatic, and trade linkages have prompted us to reconsider previous assumptions that rest on stable characteristics of states. Research reviewed here depicts social movements that have produced strikingly similar social movements that share similar forms, goals, tactics, and ideologies. Thus theories that focus solely on the internal bases of discontent now seem short sighted.

A resulting network of economic and political ties based upon ethnicity potentially unites populations in different states but divides them against their own host states. As a consequence, ethnic mobilization at the global level provides fertile ground for new types of movements based upon national, ethnic, and other cultural identities. Such movements can be simultaneously local and international in scope. Because of this flexibility, E/R/N are able to shift strategies quickly and adapt to changing political and economic conditions. By turning to explanations firmly based on theories of international connections and processes, we may be better able to understand the emergence of new forms of nationalism.

See also: Civil Rights Movements; Collective Behavior, Sociology of; Ethnicity, Sociology of; Nationalism, Sociology of; Political Sociology; Racial Relations.

Bibliography

Anderson, B., 1991. Imagined Communities. Verso, London, UK.

Andrews, K., 2013. Civil rights movement (United States). In: della Porta, D., Klandermans, B., McAdam, D. (Eds.), The Wiley-Blackwell Encyclopedia of Social and Political Movements (accessed 13.04.13) http://onlinelibrary.wiley.com/store/10.1002/9780470674871.wbespm451/asset/wbespm451.pdf?v=1&t=hfif78ue&s=bc3636815932b8d6a770f76c9f5edf3e99a95fe9.

Barth, F., 1969. Ethnic Groups and Boundaries. Sage, Los Angeles, CA.

Beissinger, M., 2002. Nationalist Mobilization and the Collapse of the Soviet State. Cambridge University Press, Cambridge.

Blattman, C., Miguel, E., 2010. Civil war. Journal of Economic Literature 48, 3–57.

Brubaker, R., 1996. Nationalism Refrained: Nationhood and National Question in the New Europe. Cambridge University Press, Cambridge, UK.

Brubaker, R., 2004. Ethnicity without Groups. Harvard University Press, Cambridge, MA.

Brubaker, R., 2011. Nationalizing states revisited: projects and processes of nationalization in post-Soviet states. Ethnic and Racial Studies 17, 1785–1814.

Brubaker, R., Feischmidt, M., Fox, J., Grancea, L., 2006. Nationalist Politics and Everyday Ethnicity in a Transylvanian Town. Princeton University Press, Princeton.

Chandra, K., 2004. Why Ethnic Parties Succeed. Cambridge University Press, Cambridge.

Chua, A., 2003. World on Fire: How Exporting Free Market Democracy Breeds Ethnic Hatred and Global Instability. Doubleday, New York.

Esteban, J., Mayoral, L., Ray, D., 2012. Ethnicity and conflict: theory and facts. Science 336, 858–865.

Fearon, J., 2004. Why do some civil wars last so much longer than others? Journal of Peace Research 41, 275–302.

Fearon, J., Laitin, D., 1996. Explaining interethnic cooperation. American Political Science Review 90, 715–735.

Fearon, J., Laitin, D., 2003. Ethnicity, insurgency, and war. American Political Science Review 97, 75–90.

Fearon, J., Laitin, D., 2005. Primary commodity exports and civil war. Journal of Conflict Resolution 49, 483–507.

Hechter, M., 2000. Containing Nationalism. Oxford University Press, New York.

Horowitz, D.L., 2001. The Deadly Ethnic Riot. University of California Press, Berkeley.

Koopmans, R., Olzak, S., 2004. Discursive opportunities and the evolution of right-wing violence in Germany. American Journal of Sociology 110, 198–230.

Koopmans, R., Statham, P., 2000. Challenging Immigration and Ethnic Relations Politics. Oxford University Press, Oxford.

Luders, J., 2010. The Civil Rights Movement and the Logic of Social Change. Cambridge University Press, New York.

Olzak, S., 1992. The Dynamics of Ethnic Competition and Conflict. Stanford University Press, Stanford.

Olzak, S., 2004. Ethnic and nationalist social movements. In: Snow, D., Soule, S.A., Kriesi, H. (Eds.), The Blackwell Companion to Social Movements. Oxford University Press, Oxford, pp. 666–693.

Olzak, S., 2006. The Global Dynamics of Race and Ethnic Mobilization. Stanford University Press, Stanford.

Olzak, S., 2011. Does globalization breed ethnic discontent? Journal of Conflict Resolution 55, 3–32.

Olzak, S., 2013. Competition theory of ethnic/racial conflict and protest. In: della Porta, D., Klandermans, B., McAdam, D. (Eds.), The Wiley-Blackwell Encyclopedia of Social and Political Movements (accessed 14.04.13) http://onlinelibrary.wiley.com/store/10.1002/9780470674871.wbespm043/asset/wbespm043.pdf?v=1&t=hfiewvfu&s=cb2ca449f98009db63e1d6d84bb4293687fba0f2.

Olzak, S., Shanahan, S., 2003. Racial policy and racial conflict in the urban United States, 1869–1924. Social Forces 82, 481–517.

Petersen, R., 2002. Understanding Ethnic Violence. Cambridge University Press, Cambridge, UK.

Sambanis, N., 2001. Do ethnic and nonethnic civil wars have the same causes? Journal of Conflict Resolution 45, 259–282.

Varshney, A., 2002. Ethnic Conflict & Civic Life: Hindus and Muslims in India. Yale University Press, New Haven.

Wilkinson, R., 2004. Electoral Competition and Ethnic Violence in India. Cambridge University Press, Cambridge.

Wimmer, A., 2002. Nationalist Exclusion and Ethnic Conflict. Cambridge University Press, Cambridge.

Wimmer, A., Cederman, L.E., Min, B., 2009. Ethnic politics and armed conflict: a configurational analysis of a new global data set. American Sociological Review 74, 316–337.

Ethnicity and Educational Achievement

Fang Gao, Hong Kong Baptist University, Hong Kong, People's Republic of China
Gerard A Postiglione, The University of Hong Kong, Hong Kong, People's Republic of China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides an overview of the causal relationship between ethnicity and educational achievement. It sets educational achievement as a process in which culture is transmitted differentially and results in inequalities across ethnic and racial groups. This article summarizes key theoretical paradigms including genetic intelligence, cultural characteristics within a particular ethnic or racial community and social forces and structural attributes in students' school and classroom experiences, and most significant empirical research in each paradigm.

Introduction

In contrast to race referred to groups of people who have perceived differences and similarities in biological traits deemed by society to be socially significant, meaning that people treat other people differently because of them, ethnicity refers to shared cultural practices, perspectives, and distinctions that set apart one group of people from another. That is, ethnicity is a shared cultural heritage, whether real or imagined. The most common characteristics that discern various ethnic groups are ancestry, a sense of history, language, religion, ritual, and forms of dress. Ethnic differences are not inherited, but learned. Yet, in day-to-day life race and ethnicity are often viewed as inseparable because the social construction of a racial category, for example, Asian or Black, carries with it a host of cultural meanings. We choose ethnicity as a focus because perceived racial groups more often than not have a shared cultural heritage, whether real or imagined. More importantly, this article addresses educational achievement as a process in which culture is transmitted differentially and results in inequalities across ethnic and racial groups.

School education is neither socioculturally nor sociopolitically neutral. Through institutional arrangements and pedagogical practices, certain ethnic identity and its underlying values, beliefs, and characteristics are represented within education spaces, yet the cultural dimensions of other ethnic groups may be marginalized by education as 'minority' culture. As Pierre Bourdieu (1986, 1991) points out, in the concept of cultural capital, that the culture of the dominant group always possesses the legitimate and valued cultural capital in education sector, and thus prevents other cultures from obtaining any legitimacy in daily school life. Such process of cultural socialization in education, ultimately, yields the system of societal stratification beyond linguistic and cultural differences, implying that those coming from other cultural backgrounds will be sidelined in the struggle for academic achievement and later, social status.

Inequality on the basis of ethnicity occurs in both the developed and the developing world and has become a matter of international urgency. In any pluralistic society, scores of students from some racial/ethnic minority groups are not on a par with either those of mainstream students or those of students from other racial/ethnic minority groups. To ameliorate the educational problems is thus to ensure equitable access of underprivileged students to all levels of education (Coleman et al., 1966; Postiglione, 1999). Various explanations have been drawn on to account for ethnic students' educational achievement and attainment and most implicate the role of ethnicity. This article thus identifies important theoretical paradigms that explain ethnicity-driven discrepancy in achievement and summarizes most significant empirical evidence in each paradigm. It introduces its readers to three major theoretical standpoints that represent different schools of thoughts in the field of research including: genetic, cultural, and social/structural. However, the readers are expected to bear in mind that ethnic origin is not the solely contributory force to academic gap; and other factors made of race, socioeconomic status (SES) (*see* Socioeconomic Status and Academic Achievement), gender (*see* School Achievement as Moderated by Sociodemographic Status: The US Case), and ability, also cause discrepancy in achievement. This article, albeit aware of the significance of intersectionality, subscribes to the role of ethnicity in inequality of education.

Theoretical Explanations for Achievement Gap

Among the well-known theoretical models, genetic and cultural perspectives demonstrate the effects of home or community background as a set of 'inputs' that children bring with them to school on their academic outcomes, yet the structural explanatory model attributes race, ethnicity, and cultural factors as key social forces that predict educational processes, practices, and policies, which in turn produce variations in educational outcomes. The emerging theories are consistent in their affirmation of the significance of ethnicity being a necessary condition for academic achievement, but are divergent in their stances on the ways ethnicity impacts upon achievement. Of significance for the divergence is whether inherited and cultural characteristics directly influence students' learning and achievement in school, or innate traits guide educational policies, processes, and practices, the nature of which determines the ultimate learning outcomes.

Paradigm 1: Genetic Deficit Theory

Genetic deficit theory, also called 'biological determinism', associates gap in achievement with racial/ethnic groups'

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92140-8

differences in cognitive abilities that arise from innate biological differences. This approach places race and ethnicity as static and relevant studies adopt intelligence quotient tests (IQ tests), and posit racial/ethnic patterns of genetic variation as a main determinant of one's intellectual capabilities over environmental and cultural attributes. Genetic deficit theory or biological determinism was a prominent explanatory paradigm in the 1960s, particularly in the US. Its advocates include Arthur Jensen, Richard Herrnstein, and Charles Murray. In 1969, Arthur Jensen, American educational psychologist, published an article entitled 'How much can we boost IQ and scholastic achievement?' in which Jensen advocates that IQ test is a scientific measure of intelligence, and 80% of a person's intelligence is solely dependent on racially heritable genes. He accordingly concludes that compensatory education of the US for less intelligent African-American children is ineffective and apparently has failed. Jensen's claim is supported by Hans Eysenck, a psychologist born in Germany and spending his professional career in the UK. Eysenck asserts that genetics is more determined by one person's cognitive ability than by environment. In his book of 1973 'The inequality of man', Eysenck writes: "The whole course of development of a child's intellectual capabilities is largely laid down genetically, and even extreme environment changes … have little power to alter this development" (pp. 111–112). Eysenck further suggests that Italian, Spanish, Portuguese, and Greek immigrants in the US are intellectually inferior and are in relevance to the populations in their countries of origin in Europe. The frame of reference has been utilized by the supporters of the 'New Right' in the US, who deliberately claim that government expenditure on education is unlikely to solve the problem of underachievement among genetically deficit student population. Today, the genetic paradigm continues to germinate within the academy as exemplified by Richard Herrnstein and Charles Murray's 1994 book 'The bell curve'. Using empirical statistical analysis, Herrnstein and Murray posit that the differences between 40 and 80% in intelligence between individuals are caused by genetic factors; 60% of the variation in IQ test scores between Black and White Americans accounts for differences in inherited intelligence between Blacks and Whites. They thereof come to the conclusion that racially inherited differences in IQ are a major factor predicting why some ethnic groups cannot excel in school.

Genetic deficit perspective, however, has attracted criticisms of not only its theoretical rationale, but its research methodology. Soon after the publication of 'The bell curve', the American Psychological Association's Board of Scientific Affairs responded to the increasing controversy around it in the statement that there is no certain support for racial differences and genetic hypothesis. In the edited volume 'The bell curve debate' (Jacoby and Glauberman, 1995), that involves 81 authors' writing comments on the book 'The bell curve', one of the authors Leon Kamin contends by using the separated twins as an example that the separated twins are always reared in similar environments, which demonstrates much more the observed similarities in IQ test scores. In addition to the contention that genetics cannot solely account for difference in intelligence, critics as well question the validity of intelligence based on IQ tests. For instance, Stephen Gould (1994) criticizes the four underlying assumptions made by Herrnstein and Murray ((1) Intelligence must be reducible to a single number. (2) Intelligence must be capable of rank ordering people in a linear order. (3) Intelligence must be primarily genetically based. (4) Intelligence must be essentially immutable) and argues that the assumptions are false and thus the entire argument based on them is false.

Paradigm 2: Culture-Relevant Explanations

Cultural Deficit Theory

Aroused by the thrust to eradicate poverty in the US and by the rejection of geneticists' argumentation, cultural deficit theory, also termed 'culture of poverty' or 'culture of deprivation', emerged during the early 1900s, and was prominent in the 1960s. This theory replaces nature with nurture as the main cause for school failure. Its theorists assume that some racial/ethnic groups do not experience a cognitively stimulating family environment and thereof their cultural attributes or practices are inferior and intellectually impoverished (e.g., limited outlooks toward the future, a lack of work ethic, low intellectual abilities, and other deficiencies such as early initiation to sex among children, female-headed households, or a limited interest in education, as identified by Stanley Eitzen and Maxine Baca Zinn in 2006), thus preventing them from achieving in school (e.g., Bereiter and Engelmann, 1966; Deutsch, 1967; Hess and Shipman, 1965) and resulting in culturally based cross-generational poverty (Lewis, 1959).

Cultural deficit theory, in spite of a highly ethnocentric view (e.g., Baratz and Baratz, 1970; Valentine, 1968), receives much attention in the community of education and it helps place the responsibility for school failure to minority children and their families (Ryan, 1976; Sleeter, 1993). For example, it is widely claimed that cultures or religions of some minorities such as Latinos in the US, Afro-Caribbeans in the UK, or Tibetans in China are inferior and in turn restrict the groups' possibilities of success in education. The explanation is compatible with the view that White working-class students' relatively low achievement arises from the cultural pathology of working-class people (e.g., Reay, 2001). As a frame of reference, cultural deficit model has its far-reaching significance in terms of the formulation in education policy and instructional practice that low-income populations and in specific racial/ethnic minority students at a disadvantage are required to overcoming cultural deficits caused by their families and communities in order to achieve in education. In the US, one of the representative policy initiations is the head start, a program launched in 1965 by the US Department of Health and Human Services, in order to alter the cultural attributes of low-income ethnic minority groups by the provision of comprehensive education, health, nutrition, and parent involvement services to low-income families and their children (Zigler and Styfco, 1996).

Cultural deficit paradigm flourished in the 1960s and today it continues to be a significant explanation for some racial/ethnic groups' underperformance in school and their subsequent limited mobility in society. For instance, in response to Herrnstein and Murray's work in 'The bell curve', Jencks and Phillips (1998) constitute the heredity-environment dichotomy and suggest that Black students' failure in

education is attributed to their family culture other than schools.

Cultural Difference Theory

In the late 1960s, cultural deficit theory was challenged by sociolinguistics, majorly by William Labov who criticizes ethnocentric cultural biases and argues that there is cultural difference in communication style between teachers and their minority students and "nonstandard dialects are highly structured systems…" (1969: p. 31). Rather than blaming either teachers or students, cultural difference or cultural compatibility/incompatibility model asserts that whether students excel or not in school reflects the cultural characteristics of the students, either compatible or incompatible with the dominant culture of schools; a lack of compatibility plays a vital role in school failure (e.g., Au, 1980; Davidson, 1992; Delgado-Gaitan, 1987; Erickson, 1987; Heath, 1983; Philips, 1972; Trueba, 1988; Vogt et al., 1987). Fredrick Erickson is a main advocate of cultural difference theory. Erickson pays attention to the 'micro' elements of classroom and points out that cultural difference theory provides 'a way of seeing classroom troubles as inadvertent misunderstanding – teachers and students playing into each other's cultural blind spots' (1987: p. 336). The home/school 'discontinuities'/difference model proceeds from the anthropological notion that minority students, families, and communities contain rich cultural and cognitive resources, termed by Moll et al. (1992) 'funds of knowledge'. The reason for the underachievement of the students is that mainstream schooling based on the prevailing culture espouses different values from those in minority culture.

Empirical research documents the variations in learning styles, expectations, languages, and interpersonal communication patterns between home and school, in particular between teachers and students as a source of limited success for ethnic minority students. For instance, Au and Mason's (1981) experimental study on native Hawaiian first graders illustrates a causal linkage between the cultural communication patterns of teaching and academic achievement. This study also supports the assertion made by Barnhardt and Heath that culturally responsive pedagogy, that is, accommodating to students' cultural norms, can not only reduce culture shock encountered by students, but affirm symbolically the culture of ethnic minority in school. Then cultural adaptation, in turn, will beget higher school performance. Other studies from the psychological view document the variations in learning styles and contend that racially and ethnically diverse students' learning results could be enhanced if their schooling experiences are able to resonate with their cognitive or learning styles (Boykin, 1986; Hale-Benson, 1986; Ramírez III and Castañeda, 1974; Shade, 1982).

Bourdieu's concept of capital that refers to the incompatibility of culture between home and school is particularly relevant to the cultural difference argumentation (Bourdieu, 1986; Bourdieu and Passeron, 1977; Lareau and Horvat, 1999; De Graaf, 1986). Cultural capital hypothesis suggests that privileged parents occupy a great deal of cultural resources that ensure their children's mastery of academic contents. Bourdieu views culture not as the traditional sociological term, but as a capital resource capable of generating 'profits' (Lareau and Weininger, 2003). He argues in an attempt to explain differences in achievement in France during the 1960s that the education system plays the role in reproducing and legitimizing the culture of the dominant classes (highbrow aesthetic culture including fine art and classical music) and in turn puts racial/ethnic minority students in disadvantage to navigate through the culturally biased academic knowledge and assessment. Among the empirical studies on the effects of cultural capital on educational achievement, De Graaf's (De Graaf, 1986; De Graaf et al., 2000) fieldwork presents the causal relationship between parental cultural capital and children's achievement. Using the Dutch educational system as an example for study, De Graaf highlights that the Dutch educational system values cultural elements, especially in terms of reading behavior that remains an important role to play in children's educational outcomes. De Graaf breaks the 'classical operationalization' of cultural capital, yet explains parents' cultural capital in terms of both the dominant highbrow cultural dispositions and linguistic and cognitive skills as human capital transmitted over generation. According to De Graaf, linguistic and cognitive skills, in specific reading abilities acquired at home contribute to children's human capital and in turn account for their success in education. Bourdieu in his writing of 'The forms of capital' (1986) develops types of capital, that include cultural capital, economic capital, social capital, and symbolic capital. In addition to cultural capital acquired in the home via exposure to a given set of family cultural practices, research evidence points out that social capital, namely, social resources offered by parents, tied to their high educational and occupational status positions, have important implications for educational outcomes. James Coleman's large-scale study of the influential factors for academic achievement articulates that there is a strong correlation between achievement and family background and environment (Coleman et al., 1966).

The cultural difference model implies that how the curriculum and the school environment act on such perceived differences is of importance with a view to decreasing academic gap. As a consequence, a call for multiculturalism in education appears alongside the 'ethnic revitalization' movements in many nation states during the 1960s and 1970s (including the civil rights movement in the US) with the goal of affirming the social, cultural, and historical experiences of students from diverse cultural backgrounds. Although multicultural education causes the warning that the celebration of diversity may undermine national unity and cohesion (Huntington, 2004; Schlesinger, 1998), multiculturalists such as James A. Banks argue that the European-American culture of schools invalidates the cultural experiences of students from non-European backgrounds and thus an infusion of multicultural education can be part of the solution to improving the educational achievement of students from diverse cultural backgrounds (Banks, 1999). One endeavor is to develop culturally relevant/responsive pedagogy that gives attention to the reflection of teaching toward a diversity of cultural practices (e.g., Ladson-Billings, 1994).

Cultural Ecological Model

Beyond the purely culture-centered explanations, Ogbu's cultural ecological model transforms the discourse toward a macro level of sociosystemic analysis in the investigation of school success or failure. As an America-based scholar, John Ogbu (1978, 1987, 1998) criticizes sociolinguistics-based cultural difference perspective and in his comparative research, he ties different minority groups' academic achievements in

the US not to school itself, but to the ways that they enter into the American society and receive schooling. He distinguishes between voluntary and involuntary minorities. Voluntary minorities such as Asian Americans are defined as immigrants who voluntarily come to the US in search of a better life. Involuntary minorities such as African-Americans or Native Americans as 'castelike' minority groups are those who have been incorporated into the US through slavery or conquest. Ogbu's cultural ecological explanation connects individual student's experience of schooling with the situation of the individual at the school and community level and in the wider society and its politics and economics. Empirical research evidence (e.g., Gibson, 1987) has shown that voluntary minorities are more likely to succeed educationally because they hold the meritocratic view and observe education as a means for success and upward mobility. On the contrary, involuntary minorities underachieve in schooling, which is seen as opposite to their culture and identities and do not believe the possibility of social mobility. It is argued that while involuntary families and community members hold negative beliefs toward school achievement, their children are less likely to engage in meaningful learning activities.

Cultural ecological model receives criticism because of its heavy emphasis on socioeconomic determinism without any room given to human agency. The success of Asian-Americans as voluntary immigrants in education is widely cited as evidence that involuntary minority students cannot acquire positive outcomes of education despite any special provision of bilingual or multicultural education. In addition, the model also fails to capture the variation within groups. Carla O'Connor's (1999) research questions perceptions of opportunity can predict academic engagement and she points out that it is dangerous to overgeneralize some populations of students and the heterogeneity or multiplicity within each group should be presented in the research of achievement disparity. For instance, even though involuntary minority groups, at large, do not indicate high performance in school, exceptions can be found and thus raise serious doubts about the adequacy of cultural ecological explanation. Moreover, according to this model, Asian-Americans have been seen as a monolithic group with shared educational and socioeconomic success. It ignores the fact that there is increasing disparity among Asian-Americans in educational and economic accomplishments (Gao, 2010; Kibria, 2002). Hence it may be dangerous to excessively emphasize cultural ecological explanation of educational achievement, in which school organization and teaching process actually play important roles.

Paradigm 3: Structural Inequality in Education

The structural/institutional paradigm of explanation for educational achievement emphasizes that race, ethnicity, and cultural factors within a particular family/community do not predict whether students learn and achieve in school, but as key social forces predict educational processes, practices, and policies that produce variations in educational outcomes. Whether educational institution reproduces or diminishes social inequality on the basis of race and ethnicity has been the most important research by sociologists of education. Proponents of the theoretical model shed light upon school funding, resources, curriculum and pedagogy, teacher expectation of students, tracking and ability grouping, and other school-level variables and contend that structural inequality of education may provide advantages of some racial/ethnic groups and marginalize or produce disadvantages for other groups. For instance, American sociologist Eric Hanushek criticizes Coleman's claim that differences in school resources are not indicators of student achievement, and schools cannot overcome differences in students' family background and SES. Hanushek (Hanushek, 1971; Rivkin et al., 2005) questions Coleman's examination on school quality by funding and teacher certification and teaching experience, instead, he argues that quality of teachers should be evaluated on the basis of student learning. Thereof Hanushek asserts that there are great differences across schools, which are not examined in Coleman Report, and schools certainly play an important role in student achievement and differences in teachers are extraordinarily significant in terms of achievement in school.

Structural explanations highlight the importance of equality of educational opportunity for the establishment of a meritocratic society where academic achievement and educational attainment are only relied upon students' academic merits rather than their ascribed racial/ethnic background. For scholars with the paradigm, to remedy or shorten the achievement gap is to critically examine the education system in order to counter any institutional racism (*see* Racism, Sociology of) and discrimination. Critical race theory (*see* Critical Race Theory) and critical pedagogy are actively promoted in order to make institutional injustice visible and challenge dominant power structures (e.g., Bell, 1976; Giroux, 1981). Relevant empirical studies range widely but mainly deal with three areas: school segregation, tracking, and teacher expectation (see Cook, 1983; Oakes, 1990; Rosenthal and Jacobson, 1968), and have important implications for educational policy and practice.

Unequal Access to School and Desegregation Policy

Although Coleman Report shows that school characteristics explain less of the variation in student achievement in comparison to family background, it evidently attests to a causal linkage between participation in predominately race-segregated schools and African-American students' lower test scores in the US. The report and other studies negate the possibility of segregated Black schools to provide the same quality of education as White schools and argue that access to White schools can ensure equality of educational opportunity for Black students. The research findings contribute to the implementation of desegregation policy and subsequently the change of student demography by bussing students in order to overcome the effects of residential segregation on public school demographics. Though desegregation policy in the US fell into dispute (see Clotfelter, 2004; Cook et al., 1984; Schofield, 1995), it has led to desegregation policy initiation in other regions of the world. For example, in Hong Kong new primary/secondary school admission policy takes racial/ethnic integration as a meaningful means to providing minority students with equal educational opportunities and ultimately to embrace them in the mainstream society (Gao, 2011).

Tracking System and Within-Class Ability Grouping

Empirical studies in the field of tracking and ability grouping examine a common educational practice, that is, assigning

students to different tracks and within-class ability groups, and explore the effects of tracking and ability grouping on academic achievement (Oakes, 1990). According to relevant research evidence, track assignments are based on both academic ability and by nonacademic factors because of the fact that racial/ethnic minority and low-income students are disproportionately placed in lower academic tracks. Researchers assert that higher tracks often offer a higher quality of teaching and instruction, and high-track students have opportunities to access to superior curriculum in a class environment, which is conducive to learning, and thus create unequal opportunities for students to learn. Statistical data have indicated the influence of tracking/ability grouping on academic achievement while controlling ability and prior achievement (e.g., Slavin, 1987). It is noted that students in higher tracks learn more and faster than those in lower tracks and the gap between high-track students and low-track students increases over the course of schooling. Hence, educational institution by its practice of tracking, does not ameliorate inequality of education, but maintain or even exacerbate educational inequality along the line of race or ethnicity. As a highly contradictory issue, tracking system has been criticized by some scholars under the consideration of equality of educational opportunity, but supported by others arguing its facilitation for teaching and instruction. At the current times, educational policy and practice mainly focus on how to modify tracking system in order to diminish its inequality while retaining its pedagogical benefits (Hallinan, 2001).

Teacher Expectation of Students and Labeling

Social-psychological processes of teacher expectation and labeling (*see* Labeling Theory, History of) account for race/ethnicity-based inequality of educational experiences in school. A body of research studies examines teacher expectation and its effects on students' behavior and academic achievement (Rist, 1970; Rosenthal and Jacobson, 1968). These studies report that teachers may have different expectations for students' performance and racial/ethnic minority students disproportionately receive the lowest teacher expectations. For example, Rosenthal and Jacobson (1968) in their study note that a teacher has lower expectation of children of 'darker skin color', or children whose mothers are 'on welfare', than of children of 'an upper-middle-class community'. Gillborn (1990) in his study of the educational experiences of Afro-Caribbean and South Asian children also points out that different racist stereotypes underlie White teachers' assumptions and expectations of Afro-Caribbean and South Asian students. In contrast to 'troublesome' Afro-Caribbean students, Asian students such as in Wright's (1992) research are always labeled by White teachers as students who have language problems and thus are excluded from class discussion. Teacher expectations ultimately create a self-fulfilling prophecy and greatly impact on student learning and achievement. The rationale behind is that students who are believed by teachers to have high ability tend to have more frequent and positive teacher-student and peer interactions and receive more attention from teachers that are conducive to learning. Intervention research and educational programs such as status expectations research are undertaken in order to raise racial/ethnic minority students' status in order for them to achieve more equal classroom participation with greater self-esteem.

Conclusion

This article outlines educational inequality in relevance to ethnicity. In the previous half century, a central concern in the sociology of education has been the relationship between educational inequality and societal stratification. Two traditional views in the field – structural functionalism and conflict theory – clarify this relationship in different ways. For the structural functionalists (e.g., Durkheim, 1969), the society is meritocratic and fair competition for society's resources is necessary for upward mobility. Equal opportunities to compete should be available to all persons of a society regardless of their social origin or other ascribed backgrounds, and inequality in the educational outcomes is seen as beneficial to the function of society. In comparison, Marx's (see Bowles and Gintis, 1976) conflict theory does not see fair competition as viable because ascribed backgrounds of individuals disadvantage certain people, and turn to the distribution of educational resources and rewards to address inequality. Conflict theorists question the ability of the education system to eradicate inequality and argue that a fairer and more efficient system is needed. It is thus important to note that the concept of ethnicity is not static so that the issue of education and ethnicity is relatively complicated and any explanation for achievement gap has to be proceeded with caution and appropriation to certain societal contexts, where either functionalism or conflict view may play a more important role. Given the ongoing challenges regarding educational stratification, communities, families, and schools are required to collaborate with each other so as to provide sufficient resources and supportive learning environments, in which ethnic minority students' academic needs can be met. The debate on the influence of ethnicity on student's educational outcomes will continue into the twenty-first century under the promotion of No Child left Behind.

See also: Critical Race Theory; Critical Theory: Contemporary; Labeling Theory, History of; Race and Academic Motivation; Race and Racism in the Twenty-First Century; Racism, Sociology of; School Achievement as Moderated by Sociodemographic Status: The US Case; Socioeconomic Status and Academic Achievement.

Bibliography

Au, K.H., 1980. Participation structures in a reading lesson with Hawaiian children: analysis of a culturally appropriate instructional event. Anthropology and Education Quarterly 11 (2), 91–115.

Au, K.H., Mason, J.M., 1981. Social organizational factors in learning to read: the balance of rights hypothesis. Reading Research Quarterly 17 (1), 115–152.

Banks, J.A., 1999. An Introduction to Multicultural Education. Allyn and Bacon, Needham Heights, MA.

Baratz, S.S., Baratz, J.C., 1970. Early childhood intervention: the social science base of institutional racism. Harvard Educational Review 40, 29–50.

Bell, D., 1976. Serving two masters: integration ideals and client interests in school desegregation litigation. The Yale Law Journal 85 (4), 470–516.

Bereiter, C., Engelmann, S., 1966. Teaching Disadvantaged Children in the Preschool. Prentice-Hall, Engelwood Cliffs, NJ.
Bourdieu, P., 1986. The forms of capital. In: Richardson, J.G. (Ed.), Handbook of Theory and Research for the Sociology of Education. Greenwood Press, New York, pp. 241–258.
Bourdieu, P., 1991. Language and Symbolic Power. Polity Press, Cambridge.
Bourdieu, P., Passeron, J.C., 1977. Reproduction in Education, Society and Culture. Sage, London.
Bowles, S., Gintis, H.A., 1976. Schooling in Capitalist America: Educational Reform and the Contradictions of Economic Life. Basic Books, New York.
Boykin, A.W., 1986. The triple quandary and the schooling of Afro-American children. In: Neisser, U. (Ed.), The School Achievement of Minority Children: New Perspectives. Erlbaum, Hillsdale, NJ, pp. 57–92.
Clotfelter, C.T., 2004. After "Brown": The Rise and Retreat of School Desegregation. Princeton University Press, Princeton, NJ.
Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J.M., Mood, A.M., Weinfeld, F.D., York, R.L., 1966. Equality of Educational Opportunity. U.S. Government Printing Office, Washington, DC.
Cook, T., 1983. What Have Black Children Gained Academically from School Integration? Examination of Meta-analytic Evidence. National Institute of Education, Washington, DC.
Cook, T., Armor, D., Crain, R., Miller, N., Stephan, W., Walberg, H., Wortman, P., 1984. School Desegregation and Black Achievement. National Institute of Education, U.S. Department of Education, Washington, DC.
Davidson, K.L., 1992. A comparison of native American and White students' cognitive strengths as measured by the Kaufman assessment battery for children. Roeper Review 14 (3), 111–115.
De Graaf, N.D., de Graaf, P.M., Kraaykamp, G., 2000. Parental cultural capital and educational attainment in the Netherlands: a refinement of the cultural capital perspective. Sociology of Education 73 (2), 92–111.
De Graaf, P.M., 1986. The impact of financial and cultural resources on educational attainment in the Netherlands. Sociology of Education 59, 237–246.
Delgado-Gaitan, C., 1987. Literacy for Empowerment: The Role of Parents in Children's Education. Falmer Press, London.
Deutsch, M., 1967. The Disadvantaged Child. Basic Books, New York.
Durkheim, E., 1969. The Evolution of Social Thought. Routledge and Kegan Paul, London.
Erickson, F., 1987. Transformation and school success: the politics and culture of educational achievement. Anthropology & Education Quarterly 18 (4), 335–356.
Eysenck, H.J., 1973. The Inequality of Man. Temple Smith, London.
Gao, F., 2010. Being a Model Minority: Schooling Experiences of Ethnic Koreans in China. Lexington Books, Lanham, MD.
Gao, F., 2011. Linguistic capital continuity and change in educational language policies for South Asians in Hong Kong primary schools. Current Issues in Language Planning 12 (2), 251–263.
Gibson, M.A., 1987. The school performance of immigrant minorities: a comparative view. Anthropology & Education Quarterly 18 (4), 262–275.
Gillborn, D., 1990. 'Race', Ethnicity and Education: Teaching and Learning in Multi-Ethnic Schools. Unwin Hyman, London.
Giroux, H.A., 1981. Ideology Culture & the Process of Schooling. The Falmer Press, London.
Gould, S.J., 28 November 1994. Curveball. The New Yorker, pp. 139–149.
Hale-Benson, J., 1986. Black Children: Their Roots, Culture, and Learning Styles, second ed. Johns Hopkins University Press, Baltimore, MD.
Hallinan, Maureen T., 2001. Equality in education. In: Levinson, D., Cookson, P., Sadovnik, A. (Eds.), Education and Sociology: An Encyclopedia. Routledge Falmer, New York, pp. 241–246.
Hanushek, E., 1971. Teacher characteristics and gains in student achievement: estimation using micro data. The American Economic Review 61 (2), 280–288.
Heath, S.B., 1983. Ways with Words: Language, Life, and Work in Communities and Classrooms. Cambridge University Press, New York.
Herrnstein, R.J., Murray, C., 1994. The Bell Curve: Intelligence and Class Structure in American Life. Free Press, New York.
Hess, R.D., Shipman, V.C., 1965. Early experience and the socialization of cognitive modes in children. Child Development 36 (4), 869–886.
Huntington, S.P., 2004. Who Are We? The Challenges to America's National Identity. Simon & Schuster, New York.
Jacoby, R., Glauberman, N., 1995. The Bell Curve Debate: History, Documents, Opinions. Random house, New York.
Jencks, C., Phillips, M., 1998. The Black-White Test Score Gap. Brookings Institution Press, Washington, DC.
Jensen, A.R., 1969. How much can we boost IQ and scholastic achievement? Harvard Educational Review 39 (1), 1–123.
Kamin, L.J., 1995. Lies, damned lies and statistics. In: Jacoby, R., Glauberman, N. (Eds.), The Bell Curve: History, Documents, Opinions. Random House, New York, pp. 81–105.
Kibria, N., 2002. Becoming Asian American: Second-Generation Chinese and Korean American Identities. The Johns Hopkins University Press, Baltimore, MD.
Labov, M., 1969. The Study of Nonstandard English. National Council of Teachers of English, Washington, DC.
Ladson-Billings, G., 1994. The Dreamkeepers: Successful Teachers of African-American Children. Jossey-Bass, San Francisco, CA.
Lareau, A., Horvat, E.M., 1999. Moments of social inclusion and exclusion race, class, and cultural capital in family-school relationships. Sociology of Education 72, 37–53.
Lareau, A., Weininger, E.B., 2003. Cultural capital in educational research: a critical assessment. Theory and Society 32, 567–606.
Lewis, O., 1959. Five Families: Mexican Case Studies in the Culture of Poverty. Basic Books, New York.
Moll, L.C., Amanti, C., Neff, D., González, N., 1992. Funds of knowledge for teaching: using a qualitative approach to connect homes and classrooms. Theory Into Practice 31 (2), 132–141.
Oakes, J., 1990. Multiplying Inequalities: The Effects of Race, Social Class, and Tracking on Opportunities to Learn Mathematics and Science. Rand Corporation, National Science Foundation.
O'Connor, C., 1999. Race, class, and gender in America: narratives of opportunity among low-income African American youths. Sociology of Education 72, 137–157.
Ogbu, J.U., 1978. Minority Education and Caste: The American System in Cross-Cultural Perspective. Academic Press, New York.
Ogbu, J.U., 1987. Variability in minority school performance: a problem in search of an explanation. Anthropology and Education Quarterly 18 (4), 312–334.
Ogbu, J.U., 1998. Voluntary and involuntary minorities: a cultural ecological theory of school performance with some implications for education. Anthropology and Education Quarterly 29 (2), 155–188.
Philips, S.U., 1972. Participant structures and communicative competence: warm springs children in community and classroom. In: Cazden, C.B., John, V.P., Hymes, D. (Eds.), Functions of Language in the Classroom. Teachers College Press, New York.
Postiglione, G.A., 1999. China's National Minority Education: Culture, Schooling and Development. Falmer Press, New York.
Ramírez III, M., Castañeda, A., 1974. Cultural Democracy, Bicognitive Development, and Education. Academic Press, New York.
Reay, D., 2001. Finding or losing yourself? Working-class relationships to education. Journal of Education Policy 16 (4), 333–346.
Rist, R., 1970. Student social class and teacher expectation: the self-fulfilling prophesy of ghetto education. Harvard Education Review 40 (3), 411–454.
Rivkin, S.G., Hanushek, E.A., Kain, J.F., 2005. Teachers, schools, and academic achievement. Econometrica 73 (2), 417–458.
Rosenthal, R., Jacobson, L., 1968. Pygmalion in the Classroom. Holt, Rinehart & Winston, New York.
Ryan, W., 1976. Blaming the Victim. Random House, New York.
Schlesinger Jr., A.M., 1998. The Disuniting of America: Reflections on a Multicultural Society. Norton, New York.
Schofield, J.W., 1995. Review of research on school desegregation's impact on elementary and secondary school students. In: Banks, J.A., Banks, C.A.M.G. (Eds.), Handbook of Research on Multicultural Education. MacMillan, New York, pp. 597–617.
Shade, B., 1982. Afro-American cognitive style: a variable in school success? Review of Educational Research 52 (2), 219–244.
Slavin, R., 1987. Ability grouping and student achievement in elementary schools: a best evidence synthesis. Review of Educational Research 57 (3), 293–336.
Sleeter, C., 1993. How white teachers construct race. In: McCarthy, C., Crichlow, W. (Eds.), Race, Identity and Representation in Education. Routledge, New York, pp. 157–171.
Trueba, H., April 1988. Meaning from Text: Cultural Knowledge Required for Second Language Literacy. Paper Presented at the Annual Meeting of American Educational Research Association. New Orleans, LA.
Valentine, C.A., 1968. Culture and Poverty. University of Chicago Press, Chicago, IL.
Vogt, L.A., Jordan, C., Tharp, R.G., 1987. Explaining school failure, producing school success: two cases. Anthropology and Education Quarterly 18, 276–286.
Wright, C., 1992. Race Relations in the Primary School. David Fulton Publishers, London.
Zigler, E., Styfco, S.J., 1996. Reshaping early childhood intervention to be a more effective weapon against poverty. In: Darby, M.R. (Ed.), Reducing Poverty in America: Views and Approaches. Sage, Thousand Oaks, pp. 310–333.
Zinn, M.B., Eitzen, D.S., 2006. Diversity in Families, eighth ed. Allyn and Bacon, Boston, MA.

Ethnicity and Ethnic Groups: Historical Aspects

Mathias Bös, Leibniz Universität Hannover, Hannover, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term ethnicity is used for a broad spectrum of groups in history. Ethnic community building is used to foster solidarity in macro-group structures, to legitimize social inequalities by descent, and to fulfill individuals' need for belonging. In the social sciences there is a growing consensus that ethnicity is socially constructed and historically contingent, but that many people perceive it as an important, unchangeable part of their identity.

Is 'Ethnicity' Ancient or Modern?

Some argue that ethnicity is as old as human kind. Hittites, Luvians, Minoans, Canaanites, Amorites, Elamites – these are just a few examples of ethnic groups that Anthony D. Smith lists as important in the political world of the ancient Near East 2200–1700 BC in his book *The Ethnic Origins of Nations* (1986). In this perspective ethnic groups have been a significant factor in the rise and fall of political powers in the history of Europe and the Near East for thousands of years.

There is, however, a totally different view that sees ethnicity as a typically modern phenomenon. This argument is expressed most forcefully by Nathan Glazer and Daniel P. Moynihan, in their seminal reader *Ethnicity: Theory and Experience*.

> We are suggesting that a new word reflects a new reality and a new usage reflects a change in that reality. The new word is 'ethnicity', and the new usage is the steady expansion of the term 'ethnic group' from minority and marginal subgroups at the edges of society – groups expected to assimilate, to disappear, to continue as survivals, exotic or troublesome – to major elements of a society. … [T]here is something new afoot in the world, and…we may label it 'ethnicity.'
> (Glazer and Moynihan, 1975: p. 5)

In one view ethnicity is ancient; in the other it is modern. This contradiction might be due to the fact that both views focus on different kinds of ethnicity. Smith sees ethnic groups as geopolitical groups. He observes the long history of often violent ethnic conflicts associated with group memberships and forms of belonging. Glazer and Moynihan are talking about the group formation processes in times of migration, urbanization, and globalization. For them ethnicity is something that emerges within the political system of modern multicultural societies. Reflecting on the historical aspects of ethnicity Randall Collins argues:

> Ethnicity is an intrinsically messy topic because the historical processes that produce it are intrinsically messy. Our analytical problems stem from the fact that ethnicity is always a distorted concept, an attempt to impose a pure category on a social reality that is not at all pure.
> (Collins, 1999: p. 78)

Brief History of the Vocabulary of Ethnicity

The adjective 'ethnic' came into use in English during the fifteenth century, denoting characteristics of groups that are not considered Christians or Jews. At that time the term ethnic was roughly equivalent to words like heathen or pagan. In the nineteenth century the adjective ethnic was increasingly used in its modern meaning as 'pertaining to or having common racial, cultural, religious, or linguistic characteristics, esp. designating a racial or other group within a larger system' (*The Oxford English Dictionary* 1989). The word ethnic derives from the Greek term **ethnos** and the corresponding adjective **ethnikos**. The word denoted any nonspecific collective, like warriors or bees. Later it was mostly used in respect to foreigners and collectives outside the city-state, sometimes with a derogatory connotation; it roughly translates into what in English today is called a nation.

The term 'nation' derives from the Latin word **natio**, denoting a group of common descent quite similar to breed or stock. In English, nation was used in respect to foreign groups, as in 'Indian nations' during the times of colonial America. But it meant other groups of a kind as well, for example, lovers or school boys. Increasingly the meaning shifted from an emphasis of common descent or kind to processes associated with political groups. Today one of the main differences between ethnic group and nation lies in the fact that the latter implies much more strongly the idea of political self-determination and the demotic element of group action than the former. The meaning of nation moves closer to that of a demos, a term which in the social sciences refers to the political aspect of a people. The Greek term **demos** was used for the common (free) people in a city-state. In the nineteenth century the word demos was occasionally used in English to refer to groups which act together to influence their political destiny, like in the 'Celtic Demos'.

The Greek world had a quite complex structure of ethnic groups such as Dorians, Ionians, Aeolians, and Achaeans (Hall, 2002) and many Greek authors like Herodotus described people in a way we would call ethnic today. Descriptions of ethnic groups are often found in Judeo-Christian European historical texts, but the social scientific concept 'ethnic' is modern. It was only at the beginning of the nineteenth century that the term ethnic became increasingly popular in the social sciences, where 'ethnological studies' set out to examine

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62015-9

'peoples' with different customs, religions, or appearances (Gabbert, 2006). From the end of the nineteenth century onward, the adjective ethnic was also used to describe minorities within a larger political unit. Within North American and Western European societies, immigrant groups and minorities were often called 'ethnic communities'.

English lacks a noun corresponding to the adjective ethnic, like the French 'ethnie', so the expression ethnic group or ethnic community is used in the social sciences to refer to a group of people with the property of being ethnic. The term 'ethnic community' was commonly used for immigrant groups or minorities until the middle of the twentieth century (Ware, 1937[1934]). The word 'ethnicity' is even younger. Having come into use at the beginning of the 1940s (Sollors, 1996), ethnicity refers to the characteristic of a person or a group. This stands in contrast to the term race that can be used to refer to a group and to the characteristic of belonging to a group social defined by biological markers.

The term 'race' is used in the same context as ethnic groups, denoting geopolitical macro-group structures or societal minorities in everyday life and in the social sciences. Like ethnic group, race is a comparatively new concept. Its etymological origins are in Spanish in the times of the Spanish inquisition. **Raza** basically meant a common stock of humans or animals, or simply a status group like nobles. During the nineteenth century race acquired a narrower meaning, referring to a group that shares a set of biological characteristics, as in 'human race'. Biological characteristics were also used to form subdivisions in human kind; one of the first taxonomies by J.F. Blumenbach (1865[1795]) used skin color and conformation of the head to characterize five races: American, Caucasian, Ethiopian, Malay, and Mongolian. Racial typologies were used to define groups of human beings assuming genetically based biological differences as congruent with differences in culture and character between human groups. Genetically based biological differences refer to the assumption that differences of races are based in genomic inheritance and emphasize the biological aspect of descent. The word genetically stems from the Greek **genos** for family, house, lineage, or species. The idea of race as a hierarchy of genetically determined human cultural groups lies at the very hard of racism (Hirschfeld, 1938). One of the reasons for the use of the term ethnic group in the social sciences was the attempt to substitute the strong biological meanings of race with a term that emphasizes the more flexible nature of differences between groups of assumed common ancestry (Fenton, 2010).

Whether the terms race and ethnicity refer to the same social phenomenon is open to discussion, but typical phrases like 'race and ethnic studies', 'race and ethnic relations', or 'racial and ethnic disparities' somehow imply that they belong to a common set of social facts. Race and ethnicity are today considered as forming an important part of identity. Identity refers to the sameness of a person or group with itself over time and to characteristics that differentiate a group or person from another. In the social sciences, the term identity came into broader use at the same time as the word ethnicity. The two terms share the same ambivalence because they denote the need for belonging and autonomy (Gleason, 1996[1983]).

The word ethnic identity refers on the one hand to that what makes a person or a group a stable unit in time with characteristics deeply rooted in the biography of the individual or the history of the group, something that is difficult or even impossible to be changed. A person from the German state of Bavaria may consider himself or herself Bavarian in respect to food, clothing, or German dialect and considers these preferences as authentic and important. On the other hand ethnic identity refers to that which makes a person or group special or different from others. Traveling through Europe, this very same Bavarian can self-describe as a German when talking to a Frenchmen; while discussing the political situation with a New Yorker, the Bavarian can position himself or herself as an European. In this respect ethnic identity is something highly situational, which depends on the relations to other groups and persons in a given situation, something that can change often.

Approaches to Ethnic Groups and Ethnicity: A Sample

Since the nineteenth century, the social sciences have developed numerous approaches to the study of ethnic groups and ethnicity. The following sample of perspectives is no attempt to give a full overview. It focuses on those approaches – largely from sociology, political science, and cultural anthropology – which have either made a historically important contribution to the social scientific study of these concepts and/or which are still important to understanding ethnicity as a social phenomenon today.

For the evolution of a social scientific perspective on ethnic groups and ethnicity three scholars were particularly important, although they never or only rarely used the term ethnic group: Sumner, DuBois, and Weber.

William Graham Sumner (1840–1910) held the first professorship in sociology at Yale. His work *Folkways* (1907 [1906]) analyzes what in everyday language is often meant when one talks about ethnic groups. Sumner emphasizes the common traditions, a common way of life, a shared culture, and a group's myth of common ancestry as an important basis for group solidarity. In his book he synthesizes the huge corpus of ethnographic literature of his time and compares a wide range of groups, especially groups that he calls 'primitive people' around the world, depicting their differences in customs. To describe hostile sentiments in the in-group toward out-groups he coined the word ethnocentrism as the root for patriotism and chauvinism, and a main cause of intergroup conflict, even wars. Sumner was convinced that these ethnocentric preferences were something learned as a child, but he held a quite social Darwinistic view when it comes to the power structure between different groups.

William Edward Burghardt DuBois (1868–1963) was a sociologist, historian, and civil rights activist who devoted his whole scholarly oeuvre to the analysis of Africans and African-Americans as groups of common heritage and ancestry (DuBois, 1940). W.E.B. DuBois' sociology of ethnic or racial groups as macro-group structures is important for the history of the social sciences because he analyzed the complex structural causes of group solidarity. Summarizing his large opus under this respect, ethnic or racial group solidarity can be caused on three levels: (1) Ethnicity or race forms an important part of an individual's social identity. Here group membership is socially defined by common ancestry, common historical destiny, and

specific cultural expressions. (2) Ethnic or racial groups constitute a structure of social inequality. DuBois emphasizes the often coercive character of those groups and their institutional settings like family, neighborhood, church, work, and associations. (3) Ethnic or racial groups are a part of the macro-group structure in every society that is constituted as a nation-state and the world system. Here ethnic or racial groups are a part of the global power structure and are a crucial factor in processes of colonization, migration, and nation building (Bös, 2008).

The first full-fledged definition of an ethnic group was developed by the sociologist and political economist Max Weber (1864–1920):

> We shall call 'ethnic groups' those human groups that entertain a subjective belief in their common descent because of similarities of external habitus or of customs or both, or because of memories of colonization and migration; this belief must be important for the propagation of group formation; conversely, it does not matter whether or not an objective blood relationship exists. Ethnic membership (Gemeinsamkeit) differs from the kinship group precisely by being a (believed-in) membership, not a group defined by actual social action, like the latter. In our sense, ethnic membership does not constitute a group; it only facilitates group formation of any kind, particularly in the political sphere. On the other hand, it is primarily the political community, no matter how artificially organized, that inspires the belief in common ethnicity. This belief tends to persist even after the disintegration of the political community, unless drastic differences in the custom, physical type, or, above all, language exist among its members …
>
> (Weber, 2007[1920]: p. 301)

For Weber the membership in an ethnic group is based on the subjective belief in group membership regardless of any 'objective blood relationship'. Furthermore he argues that subjective beliefs in common descend follows from factors like similarities in customs or historical memories of colonization and not the other way around. Historically, ethnicity is not the cause of cultural differences but a subjective interpretation of group relations based on existing differences. Technically speaking in the language of modern social sciences, being ethnic is the dependent not the independent variable. Or using Marx's classical distinction between class 'in itself' and class 'for itself', there is only an ethnic group for itself and nothing like an ethnic group in itself. There is no such thing as an ethnic position in the social structure independent from the perception of an actor, being a member of an ethnic group is always a conscious and active process. Instead, ethnic groups are caused by social structures and they promote social structures. The propagation of group formation usually goes hand in hand with ethnic closure, that is, the use of ethnicity by a group to monopolize life chances (Bös, 2004).

Besides a few pages on ethnic groups in 'Economy and Society', Weber did not use the concept of ethnic groups in his social analysis. But many authors implicitly or explicitly refer to his pathbreaking conceptionalization. He also straightforwardly rejected the biological notion of race as an explanation of historical processes in the social sciences. Like ethnicity, Weber saw race as a cultural construction; in fact, he conceptualized race as a subtype of ethnic groups characterized by socially constructed physical markers.

The interactionistic approach to ethnic groups, developed by US-American sociologists like Robert Ezra Park (1864–1944) and William Isaac Thomas (1863–1947), was influenced by Weber and, more strongly, by the work of the German sociologist Georg Simmel (1858–1918). For Park and Thomas society is a huge web of interactions between individuals and between groups. Epitomized by Simmels' famous 'web of group affiliations', individuals are defined by their memberships in a whole set of different groups, these multiple memberships are the structural condition to become an autonomous individual.

The definition of a situation by the actor himself or herself is the key to understanding his or her action; as W.I. Thomas famously formulated, when men define a situation as real, it yields real consequences. What it means to define a situation in reference to ethnic group membership can be analyzed, for example, in letters from immigrants to friends and relatives in their country of origin (Thomas and Znaniecki, 1958[1918]). For immigrants, ethnicity is something newly created in the immigration situation of the host country. The social structure of the host country plays an important role in this type of ethnogenesis. Societies develop social forms like the 'marginal man' (Park, 1928) or the 'stranger' (Simmel, 1983[1908]) as stable social meaning structures that govern the internal logic of a definition of the situation. But besides this radical actor orientation, especially Park tended to develop sweeping historical generalizations of intercultural processes like the race relation cycle: contact, conflict/competition, accommodation, and assimilation. He considered these stages to be universal; all racial or ethnic groups go through this cycle when they come into contact with a majority.

At the beginning of the last century it was quite common for social scientists to see world history as the consequence of a power struggle between human races, where lower races were doomed to fail. The battle against this form of scientific racism was primarily fought by cultural anthropologists like Franz Boas (1858–1942). His two main arguments against scientific racism were statistical and historical. Statistically, the distribution of characteristics like intelligence between ethnic or racial groups may show a small difference in means between groups, but they nevertheless overlap extremely so that a highly intelligent individual of one racial group has a high statistical chance to be more intelligent than the average person of another group. This means that a small statistical difference in the means of two distributions cannot legitimize the different treatment of these groups, for example, in the education system. Furthermore Boas argued that even if there are statistical differences, for example in bodily features between groups, these differences tend to equalize within a short time span, if these groups live in the same social and environmental conditions.

The other main line of argumentation advanced by cultural anthropologists was historical. Over time, all ethnic groups and races have made important technological and social contributions to the evolution of mankind. This historical insight undermines the assumption of the evolutionary superiority of one group over the other. At the end of the interwar period, the key assumptions of scientific racism were proven wrong: there are no genetically based cultural differences between groups that are socially meaningful; the alleged evolutionary-based

hierarchy between races is only a myth, and 'race mixture' is no indicator for the decline of a racial group.

The falsification of racist ideas led social scientists to abandon the term race from their theoretical vocabulary. This point was eloquently made in Ashley Montagu's book *Man's Most Dangerous Myth: The Fallacy of Race* (1997[1942]). Later on (1951) Montagu was one of the main authors of the UNESCO statement against the use of the word race in politics or science. This position was also seen as a lesson learned from the role of the ideology of race in the hideous crimes of Nazi Germany. Whereas cultural anthropologists adopted this stance, especially Anglo-Saxon sociology kept using the term race as a scientific concept.

The best expression of the American sociological perspective on ethnicity after World War II is surely Warner and Srole's book *The Social Systems of American Ethnic Groups*, which was published in the Yankee City Series:

> The future of American ethnic groups seems to be limited; it is likely that they will be quickly absorbed. When this happens one of the great epochs of American history will have ended and another, that of race, will begin. Paradoxically, the force of American equalitarianism, which attempts to make all men American and alike, and the force of our class order, which creates differences among ethnic peoples, have combined to dissolve our ethnic groups. Until now these same forces have not been successful in solving the problem of race.
>
> (Warner and Srole, 1945: pp. 295–296)

In this book Warner formulated the American relational model of the terms 'race' and 'ethnicity'. For him, the important difference is that race refers to socially recognized bodily features, with ethnicity forming a kind of cultural subdivision of races. This can be called Model I: societies are divided into races that may be further divided into ethnic groups. The alternative, Model II, depicts societies as divided into ethnic groups, only some of which are 'racialized'. This is the Weberian model in which 'ethnicity' is the general term, and in some cases one sees biological features used as markers for ethnic groups.

Milton Gordon's pathbreaking study *Assimilation in American Life* (1964) formulated two central arguments: ethnic group formation within society is closely connected to economic factors and assimilation is always a complex, multidimensional process. The economic status of a group is often the prime factor that determines its lifestyle and its attitudes. Since the congruency between ethnic groups and classes can be observed quite often, Gordon coined the term ethclasses. Even more importantly he conceptualized the multidimensionality of ethnic group relations. He formulated seven dimensions: (1) Acculturation: the adoption of cultural features like language or customs. (2) Structural assimilation: the participation in institutions of the host society, especially in the economy. (3) Marital assimilation: intermarriage and 'mixed families' as important process of assimilation in the private sphere. (4) Identificational assimilation: the minority identifies with the values and aims of the surrounding society. (5) Attitude reception assimilation: the reduction or absence of prejudice. (6) Behavior reception assimilation: the reduction or absence of discrimination. (7) Civic assimilation: the inclusion into the system of political power. These dimensions are not necessarily synchronous; people can be integrated, for example, on an identificational dimension, but disintegrated in a structural dimension. There can be a largely acculturated and structurally integrated ethnic group, but still a high level of discrimination.

A significant refinement in understanding ethnic group processes was Barth's (1969) concept of 'ethnic markers'. The main structural problem of an ethnic group is border maintenance: For every ethnic group it is important to know who belongs to the ethnic group and who does not. Cultural markers – including language, hairstyles, or food preferences – can serve as indicators of membership, as ethnic markers. This kind of ethnic differentiation is largely independent from other cultural patterns of the ethnic group. More often than not, cultural patterns are quite similar between ethnic groups. As already stated in Weber's definition, the propagation of ethnic community building can be based on many other factors than cultural differences. One case in point is the power conflict between different groups in former Yugoslavia and the increasing ethnicization of groups like Serbs and Bosnians.

Given its reflexive nature, ethnicity is always a cultural phenomenon, but ethnicity is not the same as culture. Culture encompasses the meaning structure and the praxis of human beings in their material and nonmaterial aspects. In the ideal typical case of a totally isolated human group, culture might be congruent with the borders of a (protoethnic) group, but in real life, owing to the diffusion of things, ideas or common traditions, culture is never limited to a single group. Ethnic groups, especially if they live close together, mostly share a huge amount of cultural features, ethnic markers often amount to that what Freud once called the 'narcissism of minor differences'. However, from the perspective of an actor, especially in highly emotionalized conflict situations, all acts and things can be ethnicized. For the ethnic mind everything becomes ethnic.

The fact that ethnic groups are primarily defined by ethnic markers and not by their 'cultural content' is one of the reasons that ethnic group formations are so flexible over time. The concept of symbolic ethnicity (Gans, 1999[1979]) elaborates exactly on that point. In studying the Jewish community in the USA, Herbert Gans observed that some ethnic cultural features (e.g., how to observe the Sabbath) changed to fit in the way of life of the majority. These changes do not weaken, but strengthen the emotional attachment to the community. In such a case of symbolic ethnicity, ethnic group formations fit structurally into society and serve as an important identity resource.

The old preoccupation of ethnic research with huge comparisons was revitalized again in the 1980s and 1990s, for example, in Daniel Horowitz's *Ethnic Groups in Conflict* (1985). The book compares deep ethnic conflicts in states in Asia, Africa, and the Caribbean that became independent after World War II. Here the focus of analysis shifts away from the assumed nature of an ethnic group to ethnic relations. Horowitz shows how the social form of ethnic conflict gets its own, often violent dynamic when a conflict system is ethnicized. In postcolonial situations elites or interest groups often promote the idea of ethnicity in order to pursue their interests. These processes were analyzed by Brass (1991) in his book *Ethnicity and Nationalism: Theory and Comparison* that focuses on cases in India and the

former Soviet Union. He emphasizes the fluid nature of ethnic macro-group structures and political situations, where elites struggle over resources and for political control.

In nation building processes in general an integral part of the ethnicization process is the invention of traditions (Hobsbawm and Ranger, 1983). Like ethnic groups, nation-states reconstruct their history, describing their contemporary political unit as the logical product of a series of historical events: in the case of France, for example, this constructed narrative starts with the Gauls 1000 BC through the 'golden age' under Charlemagne onto the French revolution and Napoleon. Only very few inhabitants of modern France would find it especially appealing to live under the material and cultural conditions like the Gauls, nevertheless many may consider themselves as a part of a common historical collective.

Especially literature that discusses the relation between ethnic groups and power structures, apply rational choice theory to explain ethnic mobilization (Hechter et al., 1982). This seems counterintuitive since ethnicity epitomizes a social phenomenon that is dominated by emotions and, sometimes, irrational behavior. Nevertheless one can make the argument that ethnic behavior is the result of strategic action, through which people pursue their own interests. This line of thinking connects to the idea, already put forward by Weber, that the propagation of ethnic community building is usually associated with an attempt to monopolize life chances.

An important part of the cultural turn in the social sciences in the 1980s was the increasing interest in the history of ethnicity, be it in social and cultural anthropology (Chapmann et al., 1989) or in sociology, as in the sociology of literature (Sollors, 1986) or in race and ethnic relations (Waters, 1990). All of these approaches try to describe and conceptionalize the ambivalences of ethnicity; its flexible yet rigid nature, its coercive yet emancipating character. The introduction of concepts such as diaspora and trans-nationalism further advanced the theoretical debate about race and ethnicity. These concepts, sometimes framed in terms of globalization, advanced the interpretation of the history of group formation outside the nation-state (Gilroy, 1993; Kivisto, 2001).

One important recent contribution to the theorization of the historical aspects of ethnicity is Jan Nederveen Pieterse's book *Ethnicities and Global Multiculture: Pants for an Octopus* (2007). Today ethnic community building processes result from the hybridization of Asian, African, American, and European cultures. According to Nederveen Pieterse, ethnicization can have different structures: domination, enclosure, competition, and optional. He even argues that in a time of optional ethnic identities within global multiculture the change of the social form 'ethnicity' is so deep that we can speak of postethnicity.

Conclusion

In the plethora of voices in the social sciences ethnicity is something old, something new, or something we are nearly over with, but the concept nevertheless serves as a focal point for new creative research and for intriguing theoretical developments.

The emphasis in this article is mainly on English-speaking Western European and North American approaches to ethnic groups and ethnicity. Like most subfields of the social sciences, research on ethnic groups has distinct national histories in respect to empirical research and theorizing. Especially for the USA, it is fair to say that the history of social scientific research on ancestry-based groups is more a history of race than of ethnicity. This fascination with race directly reflects the focus of societal discourses: a fixation that Gilroy (2000), among others, deems increasingly problematic.

In the social sciences there is surely a growing consensus about ethnicity: although perceived by many people as an important, unchangeable part of their identity, scholars see ethnicity as socially constructed and historically contingent. As long as solidarity in macro-groups is produced through tradition, as long as social inequalities are legitimized by descent, and people feel the need for belonging, researchers will not run out of material to which the concepts of ethnic groups and ethnicity can be applied.

See also: Ethnic Cleansing, History of; Ethnicity, Sociology of; Ethnicity: Anthropological Aspects; Eugenics, History of; Race Identity; Racism, History of; Simmel, Georg (1858–1918); Weber, Max (1864–1920).

Bibliography

Barth, F., 1969. Ethnic Groups and Boundaries. The Social Organization of Culture Difference (Results of a Symposium Held at the University of Bergen, 23rd to 26th February 1967). Universitetsforlaget, Allen & Unwin, Bergen, London.

Bös, M., 2004. Reconceptualizing modes of belonging: advancements in the sociology of ethnicity and multiculturalism. In: Genov, N. (Ed.), Advances in Sociological Knowledge over Half a Century. VS-Verlag, Wiesbaden, pp. 221–244.

Bös, M., 2008. Ethnizität. In: Baur, N., Korte, H., Löw, M., Schroer, M. (Eds.), Handbuch Soziologie. VS-Verlag, Wiesbaden, pp. 55–76.

Blumenbach, J.F., 1865[1795]. On the Natural Variety of Mankind. In: Bendyshe, T. (Ed.), The Anthropological Treatises of Johann Friedrich, third ed. Longman, London, pp. 65–144.

Blumenbach, Longman et al., London, pp. 145–276.

Chapmann, M., McDonald, M., Tonkin, E., 1989. Introduction – history and social anthropology. In: Tonkin, E., McDonald, M., Chapmann, M. (Eds.), History and Ethnicitya. Rotledge, New York.

Collins, R., 1999. Chapter 3: Balkanization or Americanization: a geopolitical theory of ethnic change. In: Collins, R. (Ed.), Macrohistory: Essays in the Sociology of the Long Run. Stanford University Press, Stanford, CA, pp. 70–109.

DuBois, W.E.B., 1940. Dusk of Dawn: An Essay toward an Autobiography of a Race Concept. Harcourt Brace, NY.

Fenton, S., 2010. Ethnicity. Polity Press, Cambridge.

Gabbert, W., 2006. Concepts of ethnicity. Latin American and Caribbean Ethnic Studies 1 (1), 85–103.

Gans, H.J., 1999[1979]. Symbolic ethnicity. In: Gans, H.J. (Ed.), Making Sense of America: Sociological Analyses and Essays. Rowman & Littlefield, Lanham, MD, pp. 167–202.

Gilroy, P., 1993. The Black Atlantic: Modernity and Double Consciousness. Harvard University Press, Cambridge, MA.

Gilroy, P., 2000. Against Race: Imaging Political Culture beyond the Color Line. Harvard University Press, Cambridge, MA.

Glazer, N., Moynihan, D.P. (Eds.), 1975. Ethnicity: Theory and Experience. Harvard University Press, Cambridge, MA.

Gleason, P., 1996[1983]. Identifying identity: a semantic history. In: Sollors, W. (Ed.), Theories of Ethnicity. New York University Press, New York, pp. 460–488.

Gordon, M.M., 1964. Assimilation in American Life: The Role of Race, Religion, and National Origins. Oxford University Press, New York.

Hall, J.M., 2002. Hellenicity: between Ethnicity and Culture. University of Chicago Press, Chicago, IL.
Hechter, M., Friedmann, D., Appelbaum, M., 1982. A theory of ethnic collective action. International Migration Review 16 (1), 412–434.
Hirschfeld, M., 1938. Racism. Victor Gollancz Ltd, London.
Hobsbawm, E.J., Ranger, T.O., 1983. The Invention of Tradition. Cambridge (Cambridgeshire). Cambridge University Press, New York.
Horowitz, D.L., 1985. Ethnic Groups in Conflict. University of California Press, Berkeley, CA.
Kivisto, P., 2001. Theorizing transnational immigration: a critical review of current efforts. Ethnic & Racial Studies 24 (4), 549–577.
Montagu, A., 1997[1942]. Man's Most Dangerous Myth: The Fallacy of Race. Columbia University Press, New York.
Nederveen, P.J., 2007. Ethnicities and Global Multiculture: Pants for an Octupus. Rowman & Littlefield Publisher, Inc, Lanham, MD.
Park, R.E., 1928. Human migration and the marginal man. American Journal of Sociology 33 (6), 881–893.
Simmel, G., 1983[1908]. Soziologie – Untersuchungen über die Formen der Vergesellschaftung. Dunker & Humblot, Berlin.
Smith, A.D., 1986. The Ethnic Origins of Nations. Blackwell Publishers, Oxford.
Sollors, W., 1986. Beyond Ethnicity: Consent and Descent in American Culture. Oxford University Press, New York.
Sollors, W., 1996. Foreword: theories of American ethnicity. In: Sollors, W. (Ed.), Theories of Ethnicity: A Classical Reader. New York University Press, New York, pp. x–xliv.
Sumner, W.G., 1907[1906]. Folkways: A Study of the Sociological Importance of Usages, Manners, Customs, Mores, and Morals. Ginn, Boston, MA.
The Oxford English Dictionary, second ed., 1989. Earlier Version First Published in New English Dictionary, 1891 ed., Online Version March 2012. http://www.oed.com/view/Entry/64786 (accessed 31.03.12.).
Thomas, W.I., Znaniecki, F., 1958[1918]. The Polish Peasant in Europe and America. Dover Publications, New York.
Ware, C.F., 1937[1934]. Ethnic communities. In: Seligman, E.R.A. (Ed.), Encyclopaedia of the Social Sciences. The Macmillan Company, New York, pp. 607–613.
Warner, W.L., Srole, L., 1945. The Social Systems of American Ethnic Groups. Yale University Press; H. Milford Oxford University Press, New Haven, London.
Waters, M.C., 1990. Ethnic Options: Choosing Identities in America. University of California Press, Berkeley, CA.
Weber, M., 2007[1920]. The origin of common ethnic beliefs: the language and cult community. In: Kalberg, S. (Ed.), Max Weber: Readings and Commentary on Modernity. Blackwell Publishing, Malden, MA, pp. 299–304.

Ethnicity and Migration in Europe

Karen Phalet, Center for Social and Cultural Psychology, Leuven, Belgium
Gülseli Baysu, Kadir Has University, İstanbul, Turkey
Kaat Van Acker, Max Planck Institute for the Study of Religious and Ethnic Diversity, Göttingen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article approaches migration and ethnic diversity issues from the social psychology of intergroup relations. From an intergroup-relations perspective, minority and majority experiences of ethnic diversity are interdependent; and they are jointly shaped by the prevailing diversity climate in a particular society. Accordingly, the article consists of three main sections. The first section sketches different patterns and policies of migration and ethnic diversity at the societal level. The second section discusses majority attitudes in European societies. The third and last section shifts focus to minority experiences of acculturation against the backdrop of generally restrictive diversity climates in Europe.

Immigration and ethnic minority issues pose major social, political, and intellectual challenges to contemporary Europe. European societies are coping with rapidly diversifying immigration streams. New arrivals include asylum seekers, refugees, sojourners, and documented or undocumented immigrants from all over the world. Additionally, growing numbers of European students and highly qualified professionals are living and working outside their countries of origin within the European Union. While the center and south of Europe are more recent destinations, in the northwest of Europe the second generation of post-1965 immigrant workers and (post)colonial immigrants, is coming of age. Meanwhile, new immigration into the northwest of Europe continues. Across Europe, most countries have extended civil rights to immigrant minorities, and some have developed targeted policies to include immigrant minorities. Yet, immigrant–host relations in European societies are generally unequal and often tense. Especially, Muslim minorities are facing widespread prejudice and discrimination (Heath et al., 2008).

Against this background, social psychological research in Europe has addressed the key question when migration and ethnic diversity are divisive issues; and when they are positive forces in the lives of individuals, groups, and society at large (Chryssochoou, 2004). This article reviews research on ethnic diversity among immigrant minorities in Northwestern Europe. It does not cover other sources of ethnic diversity in an enlarged European context, such as the presence of nonimmigrant or 'indigenous' minorities with either national (e.g., ethnic Hungarians outside state borders) or ethnic status (e.g., Roma and Jewish minorities in Hungary). For the purpose of this article, immigrant minorities include the first generation of immigrants proper and their native-born offspring, the so-called second generation. Host majorities refer to a nonimmigrant reference group of native-origin citizens in the same city, school, or country. The terms majority and minority culture or identity will be used to distinguish the dominant national cultures and identities of the receiving countries from the heritage cultures and the distinct ethnic or religious identities of immigrant minorities. Multiculturalism refers to the acceptance and valuing of ethnic diversity and to inclusive diversity policies that accommodate minority culture and identity.

With its great variety of national and ethnic cultures and institutional arrangements, Europe constitutes a natural laboratory for studying the different ways in which societies organize immigrant–host relations. Varying diversity patterns and policies across Europe give rise to different social and psychological realities for both immigrant minorities and host majorities. Parallel and largely separate research lines in social psychology have studied (1) sources of ethnic prejudice and ways to reduce prejudice among host majorities, and (2) acculturation processes and their consequences for the success and well-being of immigrant minorities. Combining both research lines, this article takes an integrative approach of diversity research from the social psychology of intergroup relations (Brown and Zagefka, 2011). From an intergroup-relations perspective, minority and majority experiences of ethnic diversity are mutually interdependent; and they are jointly shaped by the prevailing diversity climate in a particular society (Phalet and Kosic, 2006). Exemplary findings will be organized theoretically around a multilevel and multigroup approach of migration and ethnic diversity. A multilevel approach connects psychological experiences of ethnic diversity in intergroup encounters (at the microlevel) to the ways in which societies organize and represent diversity and intergroup relations (at the macrolevel). In addition, a multigroup approach distinguishes between, most often asymmetrical, minority and majority group perspectives on ethnic diversity and intergroup relations.

This article consists of three sections. The first section sketches different patterns and policies of migration and ethnic diversity at the societal level, and discusses their implications for immigrant–host relations. The second section draws on prejudice research to discuss generally restrictive public attitudes in European societies, and how they relate to majority experiences and representations of ethnic diversity. The last section shifts focus to minority experiences of acculturation against the backdrop of less welcoming diversity climates and majority attitudes in Europe.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.24040-3

Diversity Patterns and Policies in Europe

Ethnic Diversity and Inequality

In Western Europe, issues of ethnic diversity have been mostly concerned with the incorporation of immigrants from outside Europe and their descendants. In addition, European enlargement and the uneven burden of the current economic crisis within the European Union, have contributed to increasing numbers of so-called free movers in Europe, including many students and professionals. Different origins, motives, and destinations of immigrants have implications for ethnic diversity patterns in European societies. Across Europe, children of immigrant workers from less developed countries or regions are facing significant socioeconomic disadvantages and widespread ethnic prejudice and discrimination. For instance, the first generation of Turkish and Moroccan 'guest workers' were recruited in the 1960s and 1970s mainly from less developed rural areas within Turkey and Morocco to be employed in unskilled industrial jobs in Europe. From the 1980s onward, family reunion and family formation enabled their permanent settlement and the establishment of major immigrant communities. In the same period, the post-industrial transition of European economies was pushing them out of a shrinking industrial sector into massive unemployment and economic inactivity. The first generation has passed their inferior group position at the bottom of the stratification heap on to the next generation, as evident from persistent educational disadvantage, high poverty rates, residential segregation in underprivileged urban areas, high risks of unemployment and inactivity, and restricted access to stable and well-paid jobs. In spite of antidiscrimination directives in Europe and legislation in most Western European societies, there is hard evidence of ethnic discrimination, for instance, in the labor market (Heath et al., 2008). Finally, the second generation has become a prime target of public hostility against Muslim communities in European societies (Voas and Fleischmann, 2012).

Migration and Integration Policies

In historical immigration countries, such as the United States and Canada, immigrants and their offspring are an indispensable part of the historical process of nation formation. In contrast, the nations of 'the old Europe' face the challenge of redefining and opening up their historically rooted national identities and institutions to immigrants from diverse ethnic and religious backgrounds (Phalet and Kosic, 2006). Most states in the northwest of Europe have gradually and selectively extended the civil and social rights that used to be the exclusive entitlements of the dominant national group to immigrants. Moreover, most have developed targeted social policies that converge on some variant of immigrant integration as a middle ground between multiculturalism and assimilation. In South and Central European countries where mass migration is more recent, diversity policies are less developed and more piecemeal. Under the umbrella of 'immigrant integration,' different countries are in fact reproducing and simultaneously reinventing their particular histories of nation building (Phalet and Kosic, 2006). Accordingly, countries differ in the (restricted) access of immigrants and their children to legal residence and national membership, in their protection from direct and indirect discrimination as well as in their social policies to promote equal opportunities in schools and in the workplace, and full participation in civil society and in politics (MIPEX, 2010). Overall, integration policies in the northwest of Europe range from assimilation, when minorities are required to give up their ethnic ties and traditions to become full members of the receiving society (e.g., in France), to some variants of multiculturalism, when multiple ethnic and national cultures and identities are accepted and valued (e.g., in Sweden, and until recently in the Netherlands).

An illustrative example is the case of national differences in the institutional incorporation of Islam in Europe. As distinct from a North American conception of freedom of religion as 'equal liberty' for all, European principles of 'religious tolerance' are premised on privileged relations between states and established churches; and they do not usually require full recognition of all religions (Fleischmann and Phalet, 2012). Looking beyond this common restriction, European countries are extending different histories of state–church relations to incorporate their Muslim citizens. For example, the Netherlands has publicly recognized and accommodated Muslim minorities by including Islam into a national history of religious pluralism. Contrarily, Islam has been denied formal recognition in Germany, where state support is restricted to historically established Christian churches. Also, France has been reluctant to accommodate Islam in accordance with a national tradition of secularism, or strict separation of church and state. These distinct institutional pathways entail different social realities for Muslim minorities across Europe. For example, only where Islam has been granted public recognition (e.g., in the Netherlands, not in Germany), are Muslim minorities able to combine sustained religious involvement with their social integration into mainstream society (Fleischmann and Phalet, 2012).

To conclude, migration histories and related patterns and policies of ethnic and religious diversity vary across Europe, yet they are clearly distinct from those in historical immigration countries such as Canada. In the European migration context, ethnic boundaries most often coincide with persistent social inequalities; and diversity policies usually fall short of multicultural accommodation.

Majority Perspectives on Ethnic Diversity: Attitudes and Intergroup Relations

Majority Attitudes and Intergroup Threat

In contrast to prevailing multiculturalism in historical immigration countries like Canada, representations of national culture and identity in European societies are more ethnically rooted and bounded, and hence less open to ethnic diversity (Verkuyten and Martinovic, 2012). In many European countries, issues of migration and ethnic diversity are highly contested in public debates and in the political arena; and inclusive integration policies are challenged by the rising tide of anti-immigrant parties and popular sentiments (Phalet and Kosic, 2006). From an intergroup-relations approach to prejudice, people will develop negative attitudes toward

those out-groups that are seen to threaten important in-group values or entitlements (Stephan and Stephan, 2000). Thus, when threat representations are induced by media images associating immigrants with crime or violence, or when threat is chronically salient, for instance, in the aftermath of terrorist attacks or in times of economic crisis or political turmoil, intergroup attitudes become more negative (Meeussen et al., 2012). Across Europe, many majority members see immigrant minorities as a threat to 'our' economic interests and cultural values (ESS, 2002). Such perceptions of threat reliably predict hostile attitudes toward minority members as well as restrictive acculturation expectations (Brown and Zagefka, 2011). Accordingly, most majority members demand that immigrants adopt the dominant culture and way of life; and they disapprove the desire of many immigrants to maintain their ethnic cultures and identities (ESS, 2002).

Particularly, public opinion surveys reveal a general trend toward more hostile majority attitudes toward Muslim values and ways of life (Voas and Fleischmann, 2012). Moreover, majority members tend to perceive Islamic and European values, and ways of life as mutually exclusive or incompatible (Chryssochoou and Lyons, 2011; Verkuyten and Martinovic, 2012). This is evident, for instance, from majority perceptions of Muslim family and political values as violating egalitarian gender role values and democratic rights and liberties in European societies. More generally, restrictive acculturation expectations among majority members are premised on an acculturation model that defines minority and majority cultures and identities as conflicting commitments (Van Acker and Vanbeselaere, 2012). For instance, Belgian majority members perceived the preference of Turkish immigrants to maintain their minority culture as incompatible with embracing the Belgian culture (Chryssochoou and Lyons, 2011). Perceived incompatibility suggests that Berry's 'integration' model of acculturation (Berry, 2001), which combines dual orientations toward majority and minority cultures and identities, may not always travel well to European migration contexts.

Media Representations and Contact Experiences

Feelings of threat and negative attitudes are not restricted to disadvantaged and typically most prejudiced segments of the majority population (ESS, 2002). Nor are they consensual or inherent in the dominant group position of majority members, as evident from ongoing public debates and deep political divisions over diversity issues within majority groups (Meeussen et al., 2012). Looking beyond classic prejudice research, therefore, social–psychological research looks for explanations of majority attitudes in the way intergroup relations with immigrant minorities are represented in society at large and experienced in daily contact situations.

First, restrictive majority attitudes have been related to prevailing media and political representations of immigrant minorities as threatening majority culture, welfare, and safety (Van Acker, 2012). Thus, content analysis of Flemish-Belgian newspapers, Internet forums, and party platforms revealed a widespread 'threat representation,' where minority members are seen to threaten majority interests and values. Threat representations associate immigrant minorities with negative stereotypes of terrorists or fortune seekers and thus, attribute responsibility for social problems to minority actors ('blaming the victim'). The findings suggest that, by diverting responsibility away from the majority group, threat representations in media and politics serve to legitimize restrictive diversity attitudes and policies. In contrast, a less frequent 'discrimination' representation sees majority members as discriminating or disrespecting immigrant minorities. The latter representation attributes responsibility for social problems to majority actors or policies. In an experimental follow-up study, Belgian participants were randomly assigned to reading realistic media messages either on 'intergroup threat' or on 'discrimination' (versus a control condition) and then reported their acculturation attitudes. Generally, restrictive acculturation expectations in the control condition were replicated or reinforced in the threat condition and they were significantly attenuated in the discrimination condition. Participants were most reluctant to support minority culture maintenance after reading a threatening media message, and most supportive after reading about discrimination.

Another example comes from a study about a real-life event, the 'Van Gogh murder' in the Netherlands. Van Gogh was a prominent Dutch film producer, who was assassinated by a Moroccan-Dutch Muslim allegedly due to his film criticizing the way women are treated in Islam. A study of intergroup threat in Dutch media following this incident (Meeussen et al., 2012) showed that the way intergroup events are framed in mass media can reinforce or attenuate negative majority attitudes. Specifically, media coverage of dissenting minority voices, e.g., condemning the violent act of a fellow Muslim as violating Islamic norms and values, effectively buffered the generalization of threat feelings to the (Muslim) minority group as a whole. To conclude, prevailing threat representations can be effectively challenged by exposing majority audiences to alternate perspectives or minority voices.

Finally, threat representations at the societal level have repercussions for daily experiences of intergroup contact with minority members. An extensive research tradition on intergroup contact reveals positive contact with minority members as a powerful tool to reduce prejudice (Pettigrew and Tropp, 2006). Against the background of less welcoming diversity climates and majority attitudes in Europe, however, the real-life contact experiences of majority members are often restricted and not always positive. More attention should be paid to the impediments of structural inequalities and legitimizing ideologies in real-life interactions between majority and minority group members. For instance, a diary study in ethnically mixed secondary schools in Belgium shows that pupils who endorse a threat representation of majority–minority relations, actually experience more threat and less positive effect in their daily interactions with minority members (Van Acker, 2012). This finding suggests that threat representations at the societal level are sustained by self-reinforcing cycles of threat through negative contact experiences. In contrast, endorsement of an alternative representation of discrimination predicted more positive effect and perspective taking during daily interactions.

Minority Perspectives on Ethnic Diversity: Acculturation and Adjustment

Minority Acculturation: Culture Maintenance and Religious Identity

In this last section, the authors turn to minority perspectives. Generally, restrictive majority attitudes impose pragmatic constraints on the acculturation options that are open to minority members (the viability question); and they affect the trade-off between psychological costs and rewards attached to bicultural integration (the adaptivity question). Berry's 'integration' type of acculturation attitudes refer to the preference to combine both minority and majority cultures and identities, in contrast with monocultural preferences, either for the minority culture ('separation') or for the majority culture ('assimilation'). Cross-cultural research with immigrant minorities shows that the second generation most often continues to attach high importance to the minority culture. In addition, they also report some measure of belonging to the majority culture, albeit levels of national identification among minorities are lower on average than among the host majority (Berry et al., 2006). While there is considerable variation in the self-relevance of the majority culture and identity, the psychological reality of most minority members and of the second generation in particular, is thus bicultural. More precisely, the majority culture becomes psychologically significant against the background of a continued and strong attachment to the minority culture (Verkuyten and Martinovic, 2012). This is best illustrated by a major cross-cultural study of the acculturation experiences of minority youth (Berry et al., 2006), which includes historical immigration countries as well as West, South, and Central European countries. The comparative findings confirmed that bicultural integration was indeed most preferred overall; and that integration predicted equally or more positive adjustment outcomes for minority youth relative to separation or assimilation.

Minority acculturation and adjustment vary considerably with the warmth of welcome in intergroup relations, however. To take into account different intergroup climates, recent research on acculturation and intergroup relations develops a multigroup perspective that combines majority and minority acculturation attitudes in particular intergroup settings. This research yields converging evidence of mutual influence between ('perceived') majority and minority acculturation attitudes; and relates diverging minority and majority attitudes (e.g., between majority Dutch preference for assimilation and minority Turkish preference for separation) to increased threat feelings and more negative intergroup attitudes (Brown and Zagefka, 2011). In European migration contexts, minority integration is often at odds with majority demands of assimilation. In these contexts, therefore, minority and majority cultures were more often experienced as conflicting; and separation was sometimes preferred over integration (for instance, by Turkish minority youth). Taken together, the evidence qualifies the psychological viability of integration, especially in more recent and less diverse European migration contexts.

One example of minority culture maintenance despite restrictive majority attitudes and discrimination is the importance of religion among second-generation Muslims in Europe. Sustained religiosity is well documented for the Turkish and Moroccan second generation in several European countries (Güngör et al., 2011). Religious continuity was related to the religious socialization of the second generation within close-knit immigrant families and communities. Immigrant parents purposefully and successfully instilled Islamic religious attachments, practices, and beliefs in their children. In parallel, coethnic peer networks further reinforced religiosity. Religious socialization in childhood predicted religiosity later in life through orientating children toward heritage culture maintenance. In line with an integration type of acculturation, religious socialization was unrelated to mainstream culture adoption.

Depending on different diversity climates in European societies, religiosity may also take on a reactive edge, however, opposing religious maintenance to mainstream culture adoption (Voas and Fleischmann, 2012). Cross-national comparisons provide evidence for more strict or reactive forms of religiosity in acculturation contexts that are antithetical to multiculturalism (for instance, in Germany (Fleischmann and Phalet, 2012)). Similarly, and in line with reactive religiosity, some evidence associates personal experiences of discrimination with increased religious identification and more strict forms of religious practice among second-generation Muslims. At the same time, there is also evidence of integration, combining sustained religiosity with majority culture contact and adoption. For instance, the religious attachment of second-generation Turkish-Belgian Muslims was compatible with positive attitudes toward adoption of the Belgian majority culture (Güngör et al., 2011). To conclude, the evidence on the psychological viability of strong religiosity in bicultural Muslims is mixed and likely to vary with the prevailing diversity climate.

Minority Adjustment: Bicultural Integration and School Performance

Turning to the adaptivity question, the adaptive value of minority acculturation attitudes, and of integration in particular, depends crucially on the acceptance of diversity by host majorities (Brown and Zagefka, 2011). To document the interplay of majority attitudes with minority acculturation and adjustment, the authors discuss in some more detail recent research on minority acculturation and school success. Schools are intergroup contexts that can make distinctive minority identities salient; and that attach more or less value to minority cultures and identities. When the school context explicitly or implicitly conveys the message that one's minority identity is valued and accepted, the school becomes an 'identity safe' learning environment for minority students. On the contrary, when the school disregards or devalues minority identities, this amounts to an 'identity threatening' environment. From the perspective of minority students, the school composition (e.g., ethnic segregation), the quality of intergroup contact (e.g., discrimination experiences), as well as salient negative stereotypes about the academic competence of minority groups are possible sources of identity threat (Baysu et al., 2013). In a field experiment in ethnically diverse Belgian schools, experimentally induced stereotype threat increased

test anxiety and decreased minority performance on a verbal analogies test, but only among participants who had experienced discrimination in school (Baysu, 2011). The interaction of chronic and situational forms of identity threat on minority performance suggests that early experiences of discrimination may have long-lasting consequences by making minority children vulnerable to stereotype threat in the classroom.

Integration often (although not always) predicts better psychological adjustment for minority members, including school success (Berry et al., 2006). In identity threatening contexts, however, when the majority rejects the double membership claims of minorities, integration may have less positive consequences for minority well-being and success. Accordingly, Turkish minority members who were high on integration were found to be more vulnerable in discriminatory school environments (Baysu et al., 2011). Thus, integrated Turkish students who reported more personal experiences of discrimination in school were more likely to leave school early and less likely to enter higher education than their assimilated or separated peers. In contrast, integrated Turkish pupils who had little or no experiences of discrimination had better chances of school success than their separated or assimilated peers. An experimental follow-up study similarly assessed the acculturation attitudes of Turkish and Moroccan minority members in Belgian schools (Baysu, 2011). Rather than using self-reported discrimination experiences, however, this study experimentally induced identity threat. Immediately preceding a verbal ability test, the experimental condition activated negative stereotypes of minority students as lacking in Dutch language mastery. Again, integrated participants showed worse performance, lower state self-esteem, and more negative feelings of frustration and anxiety under threat. In contrast, in the absence of threat, they were doing much better and they outperformed their peers with separation or assimilation attitudes. Finally, a study in British schools found longitudinal evidence of the psychological costs of integration in an achievement context (Brown et al., 2013). Asian minority children with an integration strategy of acculturation, while reporting higher levels of self-esteem and peer acceptance, also evidenced more emotional symptoms of social anxiety in school than children with either separation or assimilation strategies.

To conclude, the findings suggest that integration is a two-edged sword (see also Chryssochoou and Lyons, 2011; Simon et al., 2013; Wiley and Deaux, 2011), enabling sustained school engagement and better performance in the absence of identity threat, yet yielding most negative outcomes under identity threat. This is especially relevant for a better understanding of minority acculturation and adjustment in European migration contexts, which are generally less welcoming and less sympathetic to multiculturalism. In the words of a second-generation Turkish-Belgian youngster: "However hard we try, as long as Flemish society does not accept us and always, whatever we do or say, underlines that 'They are Turks, they are Muslims anyways', we cannot fully integrate. We will always be called, and defined as, Turk and Muslim. …Still, I live here. I work here. I have everything here. Turkey is my motherland. Belgium is my fatherland. I should be equally part of both."

Conclusion

This article reviewed selected research from both minority and majority perspectives on diversity. Minority and majority attitudes and experiences are situated against the background of a less welcoming diversity climate in today's Europe. Taken together, social–psychological findings suggest that ethnic diversity can be a divisive or a positive force; and that much depends on the way societies organize and represent intergroup relations between immigrants and hosts. In spite of persistent ethnic inequalities in most countries, diversity patterns vary widely across Europe. Similarly, diversity policies usually fall short of multicultural accommodation, yet they can be more sympathetic – or antithetical – to multiculturalism. The resulting diversity climate at the societal level shapes the diversity attitudes and experiences of both majority and minority members. Thus, common threat representations of immigrant minorities in media and politics predicted more restrictive acculturation attitudes and less positive daily contact experiences in majority members, suggesting self-reinforcing cycles of threat engendering more threat. Importantly, alternate views and minority voices effectively can break this cycle of threat, for instance, through shifting group perspectives or responsibility attributions. Similarly, on the minority side, European research findings qualify the psychological viability and adaptive value of bicultural integration, or the combination of minority and majority cultures and identities, in less welcoming environments. For instance, evidence of so-called reactive religiosity suggests that Muslims who experience personal discrimination perceive their religious identity as incompatible with the majority culture. Not only the psychological viability but also the adaptive consequences of integration depend crucially on majority attitudes toward diversity. For instance, bicultural integration made minority students most successful in so-called identity-safe school environments, where diversity is accepted and valued. In contrast, in an identity-threatening environment or situation, for instance, in the presence of stereotype threat or when experiencing discrimination, students with an integration strategy of acculturation were least successful. Taken together, these new social–psychological findings lay out an emerging European research agenda for the social psychology of migration and ethnic diversity. They combine parallel and hitherto largely separate research streams on majority prejudice and minority acculturation. To articulate an integrative intergroup-relations perspective on ethnic diversity, the studies the authors reviewed develop a multigroup and multilevel approach of minority and majority experiences in different societal contexts. To this end, they cover the full range of cross-nationally comparative, longitudinal, and experimental research strategies in social psychology. While diversity issues are commonly associated with images of the threatening immigrant or Muslim, social psychology is well equipped to articulate different experiences and representations of diversity as a challenge for Europe.

See also: Acculturation; Ethnic Identity, Psychology of; Identity and Identification, Social Psychology of; Immigration: Social

Psychological Aspects; Intergroup Relations; Racism: Social Psychological Perspectives; Stereotypes in Social Psychology; Xenophobia: Social Psychological Aspects.

Bibliography

Baysu, G., 2011. An Intergroup Perspective on School Success: Turkish & Moroccan Minorities in Western Europe (Ph.D. thesis). University of Leuven, Leuven.

Baysu, G., Phalet, K., Brown, R., 2011. Dual identity as a two-edged sword: identity threat and minority school performance. Social Psychology Quarterly 74 (2), 121–143.

Baysu, G., Phalet, K., Brown, R., 2013. Relative group size and minority school success: the role of intergroup friendship and discrimination experiences. British Journal of Social Psychology. http://dx.doi.org/10.1111/bjso.12035.

Berry, J.W., 2001. A psychology of immigration. Journal of Social Issues 57 (3), 615–631.

Berry, J.W., Phinney, S.P., Sam, D.L., Vedder, P. (Eds.), 2006. Immigrant Youth in Cultural Transition. Acculturation, Identity and Adaptation across National Contexts. Lawrence Erlbaum, Mahwah, NJ, London.

Brown, R., Baysu, G., Nigbur, D., Rutland, A., Watters, C., Cameron, L., Hossain, R., LeTouze, D., Landau, A., 2013. Acculturation attitudes and social adjustment in ethnic minority British children: a longitudinal study. Personality and Social Psychology Bulletin. http://dx.doi.org/10.1177/0146167213500149.

Brown, R., Zagefka, H., 2011. The dynamics of acculturation: an intergroup perspective. Advances in Experimental Social Psychology 44, 129–184.

Chryssochoou, X., 2004. Cultural Diversity: Its Social Psychology. Blackwell, Oxford.

Chryssochoou, X., Lyons, E., 2011. Perceptions of (in)compatibility between identities and participation in the national polity of people belonging to ethnic minorities. In: Azzi, A.E., Chryssochoou, X., Klandermans, B., Simon, B. (Eds.), Identity and Participation in Culturally Diverse Societies. A Multidisciplinary Perspective. Wiley-Blackwell, Oxford, UK, pp. 69–88. http://dx.doi.org/10.1002/9781444328158.ch4.

ESS (European Social Survey), 2002. Downloadable from: www.europeansocialsurvey.org/.

Fleischmann, F., Phalet, K., 2012. Integration and religiosity among the Turkish second generation in Europe. A comparative analysis across four capital cities. Ethnic and Racial Studies 35 (2), 320–341.

Güngör, D., Fleischmann, F., Phalet, K., 2011. Religious identification, beliefs, and practices among Turkish-Belgian and Moroccan-Belgian Muslims: intergenerational continuity and acculturative change. Journal of Cross-Cultural Psychology 42 (8), 1356–1374.

Heath, A., Rothon, C., Kilpi, E., 2008. The second generation in Western Europe: education, unemployment and occupational attainment. Annual Review of Sociology 34, 211–235.

Meeussen, L., Phalet, K., Meeus, J., Van Acker, K., Montreuil, A., Bourhis, R., 2012. "They are all the same": low perceived typicality and outgroup disapproval as buffers of intergroup threat in mass media. International Journal of Intercultural Relations. http://dx.doi.org/10.1016/j.ijintrel.2012.05.002.

Migration Integration Policy Index (MIPEX), 2010. Overall Rankings. http://www.integrationindex.eu/topics/2638.html.

Pettigrew, T., Tropp, L., 2006. A meta-analytic test of intergroup contact theory. Journal of Personality and Social Psychology 90 (5), 751–783.

Phalet, K., Kosic, A., 2006. Acculturation in European societies. In: Sam, D.L., Berry, J.W. (Eds.), The Cambridge Handbook of Acculturation Psychology. Cambridge University Press, Cambridge, pp. 331–348.

Simon, B., Reichert, F., Grabow, O., 2013. When dual identity becomes a liability: identity and political radicalism among migrants. Psychological Science 24 (3), 251–257.

Stephan, W.G., Stephan, C.W., 2000. An integrated threat theory of prejudice. In: Oskamp, S. (Ed.), Reducing Prejudice and Discrimination. Erlbaum, Mahwah, NJ, pp. 23–46.

Van Acker, K., 2012. Europe's Real and Present Threat? How Representations of Intergroup Relations Shape Attitudes towards Muslim Minorities (Ph.D. thesis). University of Leuven.

Van Acker, K., Vanbeselaere, N., 2012. Heritage culture maintenance precludes host culture adoption and vice versa: Flemings' expectations of Turks' acculturation behavior. Group Processes & Intergroup Relations 15, 133–145.

Verkuyten, M., Martinovic, B., 2012. Immigrants' national identification: meanings, determinants and consequences. Social Issues and Policy Review 6, 82–112.

Voas, D., Fleischmann, F., 2012. Islam moves West. Religious change in the first and second generations. Annual Review of Sociology 38, 525–545.

Wiley, S., Deaux, K., 2011. The bicultural identity performance of immigrants. In: Azzi, A.E., Chryssochoou, X., Klandermans, B., Simon, B. (Eds.), Identity and Participation in Culturally Diverse Societies: A Multidisciplinary Perspective. Wiley-Blackwell, Chichester-Oxford, pp. 49–68.

Ethnicity: Anthropological Aspects

Richard Jenkins, University of Sheffield, Sheffield, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article opens with a brief discussion of the historical centrality to anthropology of a particular model of collective identity – the small-scale, territorially defined, corporate group – and a related assumption of bounded cultural differentiation. Reflecting a gradual move away from perspectives rooted in evolutionism and colonialism, the anthropological model of collectivity has changed, from the tribe to the ethnic group or community. Ethnicity in one sense or another remains the dominant current framework, and its key themes are membership and boundary. There are two major anthropological positions: the transactionalist or situationalist, following Barth (with an intellectual pedigree stretching back to Weber), and the primordialist, usually attributed to Geertz (and deriving from the European Romantic tradition). The disagreement between the two points of view rests at least in part on a misreading of both Barth and Geertz. It is possible to identify a basic anthropological model of ethnicity that accommodates understandings of ethnicity as both negotiable and situational, and potentially deeply rooted and powerful in its consequences.

The ancient Greek word *ethnos*, referring to human collectivities living and acting together, is typically translated today as 'people' or 'nation.' Over the last 100 or more years, the concepts of ethnicity and ethnic group that are derived from it have passed into everyday discourse. They are central to the politics of group differentiation and advantage in the culturally diverse democracies of Europe and North America. With notions of 'race' in disrepute since 1945, ethnicity became a rallying cry in the bloody reordering of the post–Cold War world: 'ethnic cleansing' stands shoulder to shoulder with earlier euphemisms such as 'racial hygiene' and 'the final solution.' In the markets of the affluent West and North, to call a commodity 'ethnic' is to imply the value of 'authenticity.' In the social sciences, ethnicity – including related topics such as racism, nationalism, and diversity – is a long-established speciality.

Corporate Groups and Cultural Differences

One of anthropology's earliest ambitions as a comparative scientific enterprise was to explain and understand human diversity. This aspiration remains central to the social and cultural anthropological imagination – and to the limited, but significant, unity of purpose that divergent national and intellectual anthropological traditions have succeeded in maintaining – ever since. The different approaches to this quest that have characterized the discipline's twentieth-century history have generally been rooted in two interrelated anthropological axioms. On the one hand, there is an understanding of human social life as based in membership in definite, bounded *corporate groups*. On the other, the basic unity of humankind is understood as crisscrossed by patterns of *cultural differentiation*.

These fundamental concepts emphasize similarity (corporate groups) and difference (cultural differentiation), the two basic dimensions of all processes of social identification. In groups we are both together (as us) and apart (from them). Although the two notions, and their implication in each other, remain constant, how they are defined, and the relationship between them, has changed since the nineteenth century. There has been a gradual shift from documenting the social structure and customs of 'tribes' to attempting to understand the complex dynamics of 'ethnic groups.'

During the colonial and immediately postcolonial eras, anthropology's theoretical preoccupation with corporate groups was manifest in the assumption that 'primitive peoples' were characteristically organized into tribal groups. In one move, the concept of 'the tribe' distanced tribal society from civilized society, in both everyday and analytical discourse, and provided the anthropologist with a theoretical model of the nature of 'uncivilized' social organization that could organize ethnographic data and function as a framework for cross-cultural comparison. At the same time as the difference between 'them' and 'us' was established, the basic similarity between different sorts of 'them' was proclaimed. This 'tribe-group' model was not, however, an explicit analytical framework; it was, rather, implicit in most ethnographic studies, neither examined nor questioned.

The Anthropology of Ethnicity

One of the earliest anthropologists to doubt the usefulness of the notion of the tribe was Edmund Leach, who argued, in his study of highland Burma (1954), that tribes, as discrete bounded entities, were essentially models developed by outsiders, for their own purposes, not locally meaningful principles of everyday social organization. Locals might talk about themselves *as if* there were clear-cut collective identities and entities, but everyday interaction and organization revealed a more complex pattern of overlap, variation, and fluidity.

By the 1960s, the notion of 'the tribe' was gradually being replaced in British social anthropology by the perhaps less colonial-sounding 'ethnic group.' However, the underlying assumptions about corporate groups and cultural difference had probably not changed much. Within American anthropology, the increasing use of an ethnicity model, over the same

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12062-8

period, can also be understood as part of a long-term, incremental shift in the discipline's axiomatic meta-theoretical model of difference: from 'race,' to 'culture,' to 'ethnicity' (Wolf, 1994).

Ethnic Groups and Boundaries

The paradigm shift toward a social constructionist model of ethnic groups followed the publication of the Norwegian symposium edited by Barth, *Ethnic Groups and Boundaries* (1969). The subtitle of this collection of essays – "The Social Organization of Culture Difference" – tells us a great deal about both anthropological understandings of ethnicity (cultural differentiation) and Barth's take on it (as social organization). In the introduction, he argued against the presumption that the social world was made up of 'distinct named groups.' Ethnicity, the boundaries of ethnic groups, and indeed their ontological status as collectivities, could not be accepted uncritically as fixed aspects of the social reality in question. Instead, Barth insisted that ethnic identity – and its production and reproduction in routine social interaction – must be treated as essentially problematic features of social life, as emergent properties of everyday interaction. The ethnographer must examine the practices and processes whereby ethnicity and ethnic boundaries are socially constructed.

Barth and his colleagues shifted the anthropological emphasis from the static evocation of tribal identity as a feature of social structure to a recognition of ethnic identity as a dynamic aspect of social organization: ethnic identification is something that people can and do manipulate, depending on the needs and possibilities of the context. People are not 'at the mercy' of their ethnicity, nor do they do things simply because they are ethnically 'X' or 'Y.' Ethnicity is not an imperative. In acknowledging this, *Ethnic Groups and Boundaries* marked a shift from an interest in the 'uncivilized' peoples of the colonies and ex-colonies to a more equitable concern with the complex heterogeneity of all societies.

Barth's emergent alternative framework had several key features. First, the analysis of ethnicity starts from the definition of the situation held by social actors. Second, the focus of attention then becomes how ethnic boundaries are maintained or changed in the structured interaction between 'us' and 'them' that takes place across boundaries. Third, the ethnicity of actors is not necessarily fixed; it is situationally defined. Fourth, ethnic identity depends on ascription, by members of the ethnic group in question and by outsiders with whom they interact. Fifth, ethnicity is not a matter of 'real' cultural differentiation; differences are in the eye of the beholder; the 'cultural stuff' that had been hitherto believed to determine group identification is somewhat irrelevant. Finally, ecological issues are influential in producing and reproducing ethnic identity: economic competition for scarce resources plays an important role in the generation of ethnicity.

Ethnicity as Socially Constructed

Barth's social constructivist model of ethnicity can be traced back to the ideas of sociologists such as Max Weber, Everett Hughes, and Erving Goffman. Although *Ethnic Groups and Boundaries* was the model's most systematic, extended, and influential statement, similar arguments can also be found in the work of contemporary North American anthropologists such as Moerman (1965) and Cohen (1969). It has dominated discussions of ethnicity within European and North American anthropology ever since.

Handelman's development of this paradigm (1977) acknowledged the importance of boundaries, and the situational negotiability of ethnic membership. However, he argued that Barth overemphasized the centrality of social interaction in establishing boundaries, at the expense of acknowledging the role of culture. For Handelman, the cultural content of ethnicity is an important aspect of its social organization; a dichotomy between the cultural and the social is misleading. He argued further that ethnicity is incorporated – made socially significant – in differing degrees of intensity and territoriality, on which depends its salience in individual experience. Moving from 'the casual to the corporate,' Handelman distinguished the *ethnic set*, the *ethnic category*, the *ethnic network*, the *ethnic association*, and the *ethnic community*. Ethnicity can, for example, organize everyday life without ethnic *groups* featuring locally as meaningful social forms, an insight that Brubaker has since developed (2004).

This anthropological model of ethnicity has become a normal science paradigm to the degree that it is somewhat taken for granted. It has many virtues. In stressing flexible social process and the practices of actors, it encouraged a move away from the various strands of determinism of the 1970s and early 1980s. The emphasis upon social construction discourages the intrusion of biological conceptions of ethnicity or race into social analysis. Finally, the importance that is attached to the views and self-perceptions of social actors themselves undermines the taken-for-granted ethnocentrism of much social science. These are important gains, and in all of these respects the approach anticipated some of the valuable aspects of postmodernism and poststructuralism.

However, the approach has suffered at the hands of some of those who have most wholeheartedly embraced it. In stressing concepts such as 'group,' 'category,' and 'boundary,' and processes of 'maintenance,' Barth – despite his intentions to the contrary – contributed to the further reification of the ethnic group as an enduring corporate entity. In particular, the spatial and physical overtones of the notion of the 'boundary' allow the superficial or the careless to celebrate the situational flexibility of ethnicity without taking on board the more difficult questions about the nature of collective social forms in which Barth is interested. In their enthusiasm for processes of boundary maintenance – negotiation and strategizing – many of those whose analyses owe much, if not everything, to Barth have taken for granted what he did not, that is, the 'group' that is the object of the research.

Three other, related, criticisms are worth mentioning. First, although the locations where Leach, Barth, and others did their fieldwork – South-East Asia and the Middle East, for example – do allow for flexibility and a degree of choice with respect to identification, the same is not true everywhere. In many other contexts, such as Northern Ireland or Kosovo, there may be much less room for maneuver. Second, many anthropologists, in their celebration of the positivities of everyday life, still appear reluctant to accord full significance to power imbalances; this, however, is changing (Gledhill, 2000). The third

criticism concerns the relative lack of theoretical attention that has been given to the relationships between ethnicity, on the one hand, and other principles of social identification, particularly class and stratification, on the other.

None of these problems, however, are inescapable theoretical weaknesses of the general anthropological approach. Recently, a solution to some of them has been put forward in the application of the distinction between *groups* and *categories* to the study of inter-ethnic relations (Jenkins, 2008). It is important to distinguish between two analytically distinct processes of ascription: *group identification* and *social categorization*. The first occurs inside the ethnic boundary (group members identify themselves), the second outside and across it (a collectivity and its members are categorized by nonmembers). Social categorization, in particular, is intimately bound up with power relations. It reflects the capacity of one group successfully to impose its categories of ascription upon another set of people, and the resources that the categorized collectivity can draw upon to resist that imposition. Acknowledging the significance of this distinction places relationships of domination and subordination on the theoretical center-stage.

Ethnicity as Personally Meaningful

If ethnic identity is a flexible matter, dependent on situation, it is legitimate to ask what it *really* means to the individuals concerned. This question has generated discussion about the nature of ethnic identity, and is a perennial textbook topic. Is ethnicity a fundamental, *primordial* aspect of human existence and self-consciousness, unchanging and unchangeable in the demands it makes upon individuals and the bonds it creates between the individual and the group? In other words, is ethnicity an aspect of human nature? Or is it, to whatever extent, defined *situationally*, strategically or tactically manipulable, and capable of change at both the individual and collective levels? Is it, indeed, wholly socially constructed, as the post-Barth conventional wisdom might suggest?

This debate takes its place alongside philosophical controversies about human nature, the human condition, and the capacity of humans to intervene in their own lives, to determine or to be determined. But the 'primordial model' also has other historical roots, in the Romantic reaction to Enlightenment rationalism. Finding some of its more respectable voices in Herder and Hegel, primordialism naturalizes ethnic groups and justifies chauvinistic ethnic sentiments, providing an intellectual charter for murderous ethnic chauvinism, nationalism, and genocide.

Academically speaking, primordialism can be identified in debates about the modernity and nature of nationalism (e.g., Smith, 1994). The anthropologist who is most often identified with a primordial model of ethnicity is Geertz (1973: pp. 255–310). Writing originally in 1963, Geertz was concerned to understand the obstacle that 'primordial attachments' – such as tribalism, parochialism, or communalism – posed to the development of the modern political sentiment of citizenship in emergent postcolonial 'new states.' This was an important recognition that, whatever else it is about, ethnicity is also about affect, and the sometime power of culture to move people.

So, on the one hand, we have negotiable identities, while on the other, there are primordial attachments. The apparent conflict between the 'primordialist' and 'situationalist' perspectives is, however, not as clear-cut as it seems. It is hard to resist the conclusion that the debate has generated more heat than light. The protagonists are certainly usually misrepresented, sometimes to a remarkable degree.

To look at what Geertz actually said, he explicitly recognized the role of culture in defining the primordial 'givens,' and that primordial attachments are historically and situationally variable in their power to move people. He is also clear that what matters is that ties of blood, language, and culture are *seen* by actors to be ineffable and obligatory; that they are *seen* as natural. He is concerned with the terms in which attachments are understood and mobilized locally; with what people *believe*. Finally, Geertz further argues that these putative 'primordial attachments,' far from being necessarily ancient, may actually be stimulated and quickened by the political modernization of nation-building.

Barth, for his part, has never neglected the power and stability of ethnic identifications. As much concerned with the persistence of ethnic boundaries as with their flux, what he said was that under certain, not uncommon, circumstances ethnic change *can* happen, not that it *must*. Furthermore, his point of view is often presented fossilized in its 1969 incarnation. Since then he has explored the importance of ongoing, and historically relatively stable, 'streams of tradition' or 'universes of discourse' within the constraints of which ethnic identities are produced and reproduced in practice. Most recently, he has acknowledged the need to add micro and macro levels of analysis to the intermediate level upon which his original arguments concentrated. This means looking at what ethnicity means for individuals.

There are other reasons for becoming impatient with this debate. In the evocative words of one teaching text, it offers a contrast between 'ethnicity in the heart' and 'ethnicity in the head' (Banks, 1996: pp. 185–187). This alerts us to the need to acknowledge affect and emotion in our considerations of ethnicity. But there is no necessary contradiction between instrumental manipulation, on the one hand, and sentiment on the other. They may, in fact, go hand in hand. They may also, of course, conflict; but so may different sentiments or opposing instrumental goals.

There is, however, good reason to reject totally any strongly primordialist view. Too much ethnographic evidence exists of the fluidity and flux of ethnic identification, and of the differing degrees to which ethnicity organizes social life in different settings, for any other position to be sensible. However, we cannot deny the longevity and stubbornness, under particular circumstances, of ethnic attachments. But to acknowledge this point does not necessitate either embracing a primordial model or abandoning the idea of social construction. It is not stretching the point to regard ethnic differentiation – the social construction of 'us' and 'them,' marked in cultural terms – as a ubiquitous feature of human sociability, and hence of all human societies. Arguments about whether or not ethnicity is 'situational' *or* 'primordial' confuse the ubiquity of a social phenomenon such as ethnicity with implacable presocial powers of causation. To suggest that ethnicity is ever-present, as one of the 'givens' of human social life, is not to endorse primordialism.

Within anthropology, the primordial intellectual position is unconvincing largely because it is something of a straw man. Debates about 'instrumentalism' and 'primordialism' exaggerate the differences between, for example, Geertz and Barth, and do neither justice. More broadly, it is tempting to suggest that the notion of the 'primordial' should be banished from the social science lexicon altogether. On the other hand, however, it would be irresponsible to forget the debate altogether. Crude primordialism is essentially a common-sense view, with enormous power in the world. It still occasionally makes an academic showing in the guise of sociobiology – now calling itself 'evolutionary psychology' – and discussions about territoriality. These views cannot be simply ignored, and an anthropological perspective, rooted in social constructionist assumptions, has much to offer as a critique of the naturalization of chauvinism and ethnic conflict.

A Basic Anthropological Model of Ethnicity

Despite the criticisms and debates summarized here – and in unacknowledged harmony with post-1980s discourses about 'difference,' globalization, trans-nationality, and so on – a 'basic anthropological model of ethnicity' remains in place as disciplinary conventional wisdom. In its foundational emphasis upon social construction, its recognition of the fluidity as well as the solidity of identity, and its refusal to biologize or naturalize what are sociocultural phenomena, it remains the cornerstone of *any* anthropological approach to the topic. It can be summarized thus:

- Ethnicity is about cultural differentiation (bearing in mind that similarity and difference are inextricably connected).
- Although ethnicity is centrally concerned with culture – shared universes of meaning – it is also rooted in, and to a considerable extent the outcome of, social interaction.
- Ethnicity is no more fixed or unchanging than the culture of which it is a component or the social situations in which it is produced and reproduced.
- Ethnicity as a social identity is collective and individual, externalized in social interaction and internalized in personal self-identification.

However, that there is a *relatively* consensual model, combined with the anthropological enthusiasm for detailed ethnography rather than systematic theory, may lead to some things being taken for granted. In particular, there is a perpetual need to struggle against the tendency to reify culture and ethnicity. Although we talk about them in these terms, neither culture nor ethnicity is 'something' that people 'have' or to which they 'belong.' They are, rather, complex repertoires that people experience, use, learn, and 'do' in their daily lives, within which they construct an ongoing sense of themselves and an understanding of their fellows.

The reification of ethnicity may also lead to its being seen as typically – or even only – an attribute of the Other. We need to remind ourselves all the time that each of *us* participates in an ethnicity – perhaps more than one – just like *them*, just like the Other or 'the ethnic minorities.' *National* identity may, however, be easier to recognize in ourselves than ethnicity. Nationalism and the construction of national identity are explicit projects of the state (and, if nothing else, we have passports). In the 'modern world,' the contours and contents of national identity, and the contexts of its uses and justifications, may be more visible than ethnicity. It is therefore no surprise that the anthropology of nationalism is a growth area (Eriksen, 2010), or that the developing field of 'border studies' – focusing inter alia on the processual and cultural character of national boundaries – owes much to the anthropological model of ethnicity (Wilson and Donnan, 2012).

Although anthropologists may not have to remind themselves that national identity, as a variation on the theme of ethnicity, is socially constructed, there is a need to keep this idea firmly in the public eye. The timeliness of Wolf's warning (1994) that 'race,' 'culture,' and 'people' are 'perilous ideas,' makes the social constructionist, anthropological model of ethnicity as relevant now as it was in the 1960s. Anthropology's comparative global reach and ethnographic research approach, its emphasis upon culture and social construction, and its capacity to see individual trees as well as the collective wood, relativizes notions about ethnicity and undermines the naturalization of ethnic or national identity. That remains a vital contribution to our understanding of the modern world.

See also: Afrocentric Approaches to Social Work; Constructivism and Constructionism: Methodology; Differentiation: Social; Ethnic Identity and Ethnicity in Archaeology; Ethnicity and Ethnic Groups: Historical Aspects; Ethnicity and Migration in Europe; Ethnicity, Sociology of; Identity in Anthropology; Integration and Multiculturalism: Focus on Western Europe; Language and Ethnicity; Nationalism, Sociology of; Nationalism: General; Race and Racism in the Twenty-First Century; Racism, History of; Racism, Sociology of; Religion: Nationalism and Identity; Social Constructionism; Social Identity in Sociology; Tribe.

Bibliography

Banks, M., 1996. Anthropological Constructions of Ethnicity: An Introductory Guide. Routledge, London.
Barth, F. (Ed.), 1969. Ethnic Groups and Boundaries: The Social Organization of Culture Difference. Universitetsforlaget, Oslo.
Brubaker, R., 2004. Ethnicity without Groups. Harvard University Press, Cambridge, Mass.
Cohen, Y.A., 1969. Social boundary systems. Current Anthropology 10, 379–403.
Eriksen, T.H., 2010. Ethnicity and Nationalism: Anthropological Perspectives, third ed. Pluto, London.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Gledhill, J., 2000. Power and Its Disguises, second ed. Pluto, London.
Handelman, D., 1977. The organization of ethnicity. Ethnic Groups 1, 187–200.
Jenkins, R., 2008. Rethinking Ethnicity: Arguments and Explorations, second ed. Sage, London.
Leach, E.R., 1954. Political Systems of Highland Burma: A Study of Kachin Social Structure. Athlone, London.
Moerman, M., 1965. Ethnic identification in a complex civilization: who are the Lue? American Anthropologist 67, 1215–1230.
Smith, A.D., 1994. The problem of national identity: ancient, medieval and modern? Ethnic and Racial Studies 17, 375–399.
Wilson, T.M., Donnan, H. (Eds.), 2012. A Companion to Border Studies. Wiley-Blackwell, Oxford.
Wolf, E.R., 1994. Perilous ideas: race, culture, people. Current Anthropology 35, 1–12.

Ethnicity, Race, Oppression, and Social Work: The Canadian Case

Jennifer Clarke, Gordon Pon, and Akua Benjamin, Ryerson University, Toronto, ON, Canada
Annette Bailey, Ryerson University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In 1987, the Canadian Association of Schools of Social Work adopted policies and accreditation standards that reflect the profession's commitment to address issues of race, ethnicity, and cultural diversity in its programs and curricula. Task force recommendations to emphasize antiracism in schools of social work were contested and resisted. Since then, various shifts in perspectives have emerged and adopted in social work to varying degrees. Despite efforts to advance antiracism, and more specifically anti-Black racism and anticolonialism in social work education, anti-oppression has been more palatable to mainstream social workers. To advance the profession, these perspectives must be understood and addressed.

Introduction

Canada as a white settler society has a strong aversion to speaking about racism. This is despite the existence of the *Canadian Multiculturalism Act* of 1988 (Haque, 2012), which aims to recognize cultural and racial diversity and acknowledge the freedom of all Canadians to preserve, enhance, and share their cultural heritage. The Act was purposed to promote full and equitable participation of individuals and communities in building the Canadian nation state (Henry and Tator, 2010; Thobani, 2009). The Act influenced Canadian discourses of race and racism (St. Denis, 2011). However, the extent of its influence is not yet fully realized because there remains a deeply engrained aversion to speaking about racism in Canadian institutions (Henry and Tator, 2010; Simpson et al., 2011; St. Denis, 2011).

The profession of social work is not immune to avoiding discussions about racism in Canada. Most schools of social work in Canada do not address issues of race and ethnicity very well, and only few require mandatory courses on antiracism in their curricula (Henry and Tator, 2010). Henry and Tator (2010) contend that most schools of social work do not prepare students well to work with racialized and Aboriginal populations. Lena Dominelli has called for a social work praxis that addresses racism at the individual, systemic, and structural level (Dominelli, 1988). Yet, many schools of social work continue to embrace social order perspectives such as systems and ecological theories that seek to understand individuals in relation to their environments (Mullaly, 1997). These approaches do not address systemic and structural forms of oppressions including racism and its intersecting correlates (Carniol, 2010). Rather, they emphasize assimilationist perspectives, which merely reify whiteness (Razack, 2002).

In the late 1970s and 1980s, social work scholars such as Maurice Moreau, Bob Mullaly, and Ben Carniol advanced structural social work, which seeks to 'redefine problems in their material and collective contexts,' and transform intersecting oppressions such as racism, classism, ableism, sexism, and gender inequality (Moreau, 1979: p. 17). Groundbreaking publications including Ben Carniol's *Case Critical: The Dilemma of Social Work in Canada* (1987) and Bob Mullaly's *Structural Social Work* (1997) hold that addressing individual problems and aspects of social differences are inseparable from larger systemic and structural factors. Thus structural social work responds to the limitations of social order perspectives such as systems and ecological theories (Mullaly, 1997). More recently in Canada, anti-oppression perspectives in social work have grown in popularity.

This article argues that anti-oppression discourses often silence antiracism. We assert that a commitment to preparing social work students to effectively meet the needs of ethnic, cultural, racialized, and Indigenous populations require deliberate attention to critical race feminism and anticolonialism frameworks that challenge the liberalism of multiculturalism and intentionally address systemic and structural barriers affecting these populations. Toward this end, we assert that schools of social work must address issues of colonialism and racism in their programs and curricula, including anti-Black racism.

A Historical Perspective of Race and Social Work Education

At the Canadian Association of Schools of Social Work (CASSW) general assembly in 1987, members voted in favor of adopting policies and accreditation standards that "reflect social work education's commitment to prepare students to work effectively with diverse ethnic, cultural, and racial populations" (CASSW, 1991: p. 3). Subsequently the CASSW Board of Directors unanimously passed a motion to refer the matter to the Educational Policy Committee for advancement and integration. Hence, the Task Force on Multicultural and Multiracial Issues in Social Work Education was launched. By May 1991, the task force membership was established and their actions materialized in the publication of *Social Work Education at the Crossroad: The Challenge of Diversity*. Based on interviews with students, faculty, field, and community representatives, this document heralded the need for greater emphasis on antiracism in schools of social work. Its recommendations covered areas of organizational structure and policies, content, students, faculty and instructors, field work, and community. In 1992, according to the CASSW, all schools of social work were to revise their education policies to reflect the changes in accreditation standards.

Since the 1991 CASSW document, Canadian schools of social work have undergone various changes in perspectives in response to cultural and racial diversity. However, many schools responded with resistance (Henry and Tator, 2010). Some of the resistance to antiracism took the form of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28107-5

accusations that its proponents were creating a hierarchy of oppressions (Barnoff and Moffatt, 2007; McLaughlin, 2005), which placed racism as the most salient or pressing oppression. Given the severe discomfort and volatility that was generated by discussions of antiracism in social work, it was not unsurprising that a shift emerged in Canada from antiracism to anti-oppression (McLaughlin, 2005; Pon et al., 2011). To understand this dynamics, we next examine efforts to advance antiracism in Canadian social work.

Efforts to Advance Antiracism in Social Work

In the 1990s, from African Canadian grassroots activism of teachers, community leaders, and academics, antiracism education emerged in Canada (Wood, 2000). Antiracism education would influence disciplines such as social work. Antiracism education challenged the liberalism of multiculturalism that emphasizes song, food, and festivals, while never having to name and examine whiteness (Razack, 2002). Antiracism uses race as the primary lens through which to understand and respond to intersecting and interlocking systems of oppressions such as racism, sexism, classism, heterosexism, and ableism (Dei, 2008). Using race as the point of entry is an explicit political stance to counter the pervasive denial of systemic and structural racism in Canada (Dei, 2008). Centering race also challenges "…benign conversations about culture and identity" and the management of diversity, which reproduces whiteness (Jeffery, 2005: p. 418).

The need for antiracism in social work is particularly pressing given Thobani's (2009) assertion that the rise of the postwar Keynesian welfare state in Canada was founded on principles of white supremacy that functioned to exalt white Canadian nationals as benevolent, kind, and caring, while simultaneously constructing racialized and Aboriginal people in Canada as constituting threats to the nation. Constructed as threats to nationalism, racialized immigrant and Aboriginal populations were deemed to be in need of surveillance and containment by social services workers (Thobani, 2009). A sustained consequence of this is the exclusion of racialized and Aboriginal people from equal and equitable access to social benefits. This inequity concerns social work, since its mission is to prioritize social justice especially for marginalized and vulnerable populations.

Integrating antiracism into social work education helps to prepare social work students to be change agents in practice. This is vital because current practice reflects an exclusion of racialized immigrants and Aboriginal people from employment in social services and as professors of social work. Henry and Tator (2010) noted that it is only in the 1990s that mainstream human services agencies in Canada began hiring from ethno-racial minority communities. For example, 94% of all child welfare workers in Canada are White (Fallon et al., 2003). Besides, racialized social workers face barriers such as lack of promotion opportunities, nonrecognition of foreign credentials, and devaluation of their knowledge and skills (Clarke, 2010, 2012; Gosine and Pon, 2010; Henry and Tator, 2010). Of equal concern is the exclusion of Aboriginal and racialized people from tenure stream positions in schools of social work in Canada (CASSW, 1991). These exclusions of individuals have led racialized immigrant communities in Canada to form their own social services organizations for the sake of survival (Thomas, 2006).

Organizations such as the Association of Black Social Workers (Thomas and Hamilton-Hinch, 2006), Black Action Defense Committee (Williams, 2005), and Urban Alliance on Race Relations (Urban Alliance on Race Relations, 2002) influenced the societal context of antiracism and challenged systemic racism in Canadian institutions. Despite their actions in challenging racial injustice, there is still resistance to antiracism discourse and a reliance on multicultural discourse in social work education and practice. However, as Lee and Bhuyan (2013) argue, "multicultural discourse in social work relies on a deficit model of culture, which according to social work historian Yoosun Park (2005), constructs culture as difference that must be managed" (p. 103). Further, multicultural social work promotes the discourse of cultural competency, which silences race and racism, depoliticizes culture as neutral and devoid of power relations, and fails to analyze the role of whiteness in social work (Pon, 2009).

The Silencing of Antiracism by Anti-Oppression Perspectives

In light of the criticisms leveled against antiracism, there has been a shift away from this perspective toward anti-oppression in social work. In the new millennium anti-oppression has emerged as a widely accepted approach to social work in Canada (Mullaly, 2002; Baines, 2010). McLaughlin (2005) asserts that proponents of antiracism such as Lena Dominelli (1988, 2002) shifted from antiracism to anti-oppression to avoid the contestation and appearance of a hierarchy of oppressions or ranking of oppressions (McLaughlin, 2005). Anti-oppression, unlike antiracism, does not use race as the point of entry or lens through which to view interlocking systems of oppression. Rather it purports to view all oppressions simultaneously and equally.

Anti-oppression aims to eradicate all forms of oppression. It is defined as "a form of social work practice which addresses social divisions and structural inequalities in the social work that is done with clients (users) or workers" (Dominelli, 2002: p. 36). Anti-oppression acknowledges worker power and expert knowledge and requires critical self-reflection to ensure an egalitarian client-centered practice (Healy, 2005). It is eclectic in its borrowing of critical perspectives including antiracism, feminism, Marxism, structuralism, postmodernism, and post-structuralism.

Several scholars have raised concern that anti-oppression discourse avoids debates about racism, while still seeming to be progressive (Pon et al., 2011; Razack, 2002; Williams, 1999). Others like Healy (2005), Briskman (2007) and Walter, Taylor and Habibis (2011) argue that the "focus on multiple forms of oppression dilutes and dissipates the significance of race in everyday and institutional interactions" Walter, Taylor and Habibis (2011, p. 12). For example, in a document entitled *Anti-Oppression Framework for Child Welfare in Ontario* (Wong and Yee, 2010), which is endorsed by all of Ontario's child welfare agencies, the word racism occurs only once in the body of the text, the word 'black'

appears only once, and colonialism, is never mentioned. This example calls attention to the way in which anti-oppression discourse silences and can easily avoid issues of racism and colonialism, while purporting to be responsive to diversity.

Toward a Critical Race Theory and Anticolonialism

Emergent calls to address anti-oppression limitations have pointed to critical race theory (Razack and Jeffery, 2002), critical race feminism (Clarke, 2012; Pon et al., 2011; Razack et al., 2010), anticolonialism (Pon et al., 2011; Sinclair et al., 2009), and decolonization (Baskin, 2010; Lavallée, 2007; Lawrence and Dua, 2005), especially as these theories address the fundamental stakes of white supremacy, imperialism, racism, settler colonialism, and interlocking oppressions. The extant literature on the ongoing colonialism and racism experienced by Aboriginal and racialized peoples in Canada is extensive (Baskin, 2010; Clarke et al., 2012; Dei, 2008; James, 1996; Lawrence and Dua, 2005). Despite this scholarship, there is relatively little discussion of racism and colonization in social work education. To address this gap, three Aboriginal professors Raven Sinclair et al. (2009) recently published a book titled *Wicihitowin: Aboriginal Approaches to Social Work*, which is a required reading for upper level students in the social work program at Ryerson University. The text provides social work students with knowledge of Indigenous worldviews, research and practice, the history of colonization, and processes of decolonization in Canada. While Indigenous content is incorporated into the social work curricula at Ryerson and such core courses are offered to students, these remain on the margin in many social work programs in Canada.

On the other hand, some schools have developed specific Indigenous programs that are grounded in Indigenous epistemologies (Sinclair et al., 2009). Although these programs are accountable to the Indigenous community and have Elders physically present in classrooms, they operate within the constraints of the CASSW, with national standards that marginalize Aboriginal ways of knowing (Bruyere, 2007). This problem is compounded by the relatively few Aboriginal (2.1%) and racialized (17%) faculty in the academy, which limits scholarship, research, and teaching (Canadian Association of University Teachers, 2010). Social work's attempt to bring *other ways of knowing* into its curricula is critical in order to "…challenge Eurocentric discourses" and "…subvert dominant thinking that re-inscribes colonial and colonizing relations" (Dei and Kempf, 2006: p. 3). Calls to address the limitations of antiracism and anti-oppression must also address the de-linking of antiracism and anti-colonialism in much of social work education, which helps to fuel discursive and political tensions between migrants, settlers, racialized, and Aboriginal peoples in Canada (Lawrence and Dua, 2005). Antiracism theory has been criticized for excluding an Indigenous perspective, and not reflecting on Indigeneity and ongoing colonialism in the settler society (Lawrence and Dua, 2005). As Lawrence and Dua (2005) contend, much of "Aboriginal activism against settler domination takes place without people of color as allies" (p. 120). As a profession, social work has not been able to effectively advance Indigenous decolonization or address racism and other forms of oppression, including the marginalization of Blacks. Thus, there is a need for social work education to focus on anti-Black racism.

Where to From Here? Anti-Black Racism in Social Work Education and Practice

The social work profession has been slow to address racism, colonialism, and white supremacy in Canada. Antiracism has helped to advance critical perspectives for understanding racism in Canada and the impact on racialized people (Dei, 2008; Henry and Tator, 2010; James, 2003). However, the antiracism movements to combat racism against Blacks in many Canadian institutions have been challenging. This has led to calls for the development of a framework that critically examines the specificities of the Black community with institutional racism, as evidenced by their overrepresentation in the criminal justice system and child welfare system; racial profiling in policing; disproportionate and extreme poverty; high levels of unemployment; and restricted access to housing and health care (African Canadian Legal Clinic, 2012). Social work's response to anti-Black racism has been minimal.

The term anti-Black racism was coined in the 1990s by Black leaders, grassroots activists, community members, and academics to reflect a particular type of racism that is directed at members of the Black community in Toronto and their resistance to this discrimination (Benjamin, 2003). Anti-Black racism began with a focus on racism in the criminal justice system, following a series of police shootings of Black men in Toronto and elsewhere in Canada, and the lack of restitution for these killings (Benjamin, 2003). The concept of anti-Black racism emerged out of internal political and ideological struggles and the external and systemic conditions of racism in Canadian society. It has drawn upon anticolonial, anti-imperialist, Marxist, Socialist, Feminist, Black Nationalist, and other political ideologies and stances, to understand and address the experiences of Black people (or African Canadians (Benjamin, 2003; Warner, 2006). As such, anti-Black racism constitutes dialectic between racism directed at Blacks and their resistance to this discrimination.

The concept of anti-Black racism, however, did not gain much traction until the Provincial government under the New Democratic Party (NDP) hired Stephen Lewis to conduct an investigation into what mainstream media termed the Yonge Street "riot" of 1992. This uprising occurred when members of Toronto's Black community and other social justice activists took to the street to resist police violence and harassment, and the killings of young Black men and women in Toronto and Halifax, and the Rodney King beating by White police officers in the United States (Benjamin, 2003; Warner, 2006). The Black Action Defense Committee (BADC) of Toronto, a grassroots organization that was formed in 1988 after the shooting of Lester Donaldson, led a coalition of Black community groups in a community-wide struggle against the shootings (Benjamin, 2002). Lewis met with members of the Black community and named anti-Black racism as a particular and severe experience of racism by African Canadians, particularly youth. Lewis' use of the term anti-Black racism validated the work of Black

community members who had been naming and resisting this form of discrimination for decades.

Anti-Black racism borrows from Gayatri Spivak's (1990) notion of strategic essentialism. This concept is a political and social act of uniting under a common identity such as Black in order to take effective action against social oppressions (Benjamin, 2003). Some argue that anti-Black racism affords Blacks 'special treatment' and supports a hierarchy of oppression (i.e., racism toward Black is worse than racism toward other racialized and oppressed groups). This argument misses two salient points. First, anti-Black racism recognizes that there are different forms of racism directed toward different groups and populations in Canadian society (e.g., slavery against Blacks; cultural genocide of Aboriginal peoples through the 1876 Indian Act; residential school system and their historic struggles against assimilation and for sovereignty; the internment of Japanese Canadians during the Second World War; and the indenture of Chinese settlers) (Henry and Tator, 2010). These specific experiences, struggles, and consequences demand differing ways to address the particularities of each group. This means attention is given to the specific struggles of Blacks. Secondly, anti-Black racism stands in solidarity with groups and communities struggling against their specific form of racism and other oppressions. Today, there is little usage of the concept or knowledge of anti-Black racism in social work.

The dialectic of anti-Black racism has much to offer social service agencies and the profession of social work, particularly insights into community organizing. Three key principles for community organizing have been advanced by BADC. The first principle is that the learning of past Black struggles is important to addressing systemic racism presently. Such learnings can come from Pan-Africanism, Black Radicalism, antiracism, Black cultural nationalist organizing, anti-imperialism, anticolonialism, postcolonial Black feminism, Black Queer, and other Black struggles (Benjamin, 2003). The political struggles of anti-Black racism can be contentious but is important in resisting Whiteness and working for inclusive and respectful principles (Benjamin, 2003). The second principle is that antiessentialist principles and practices must be maintained throughout when organizing from an anti-Black racism perspective. This principle means that there is no way of phenotypically or culturally being identified as Black. Black is intersected and interlocked with different social identities (e.g., Black is intersected with social identities of class, age, cultural backgrounds, sexual identities, and disabilities). This means that anti-Black racism must attend to the diverse voices and perspectives, and the differing political positions of the organizers in a respectful and inclusive manner (Benjamin, 2003).

A third principle concerns a deep ongoing commitment to continuous leadership of giving time, energy, and other resources as a requirement in organizing from an anti-Black racism perspective (Benjamin, 2003). The commitment to leadership in organizing must begin with a historic understanding that when Black leaders and organizations have attempted to make systemic change, their efforts have often been seen as untenable and unacceptable by those in power. Thus many community leaders and organizations have come under attack by the state (Benjamin, 2002, 2003; Williams, 2005). Leaders must therefore be skilled in strategy and tactic to challenge attacks by those in power who attempt to redirect or refuse to engage with their demands for systemic change in racist policies and practices. Benjamin (2003) argues that "naming racism and anti-Black racism is, consequently and critically, a form of resistance to White, capitalist hegemony and domination. To not name racism is often to give ground to those hegemonic forces that mask racism…under the guise of seemingly liberal and democratic notions of equality" (p. 80). Anti-Black racism offers a framework for theorizing about power and privilege that enables whiteness and white supremacy to be exposed (Dei, 2000). Anti-Black goes beyond race to tackle and work toward the development of a just and equitable society.

Today, implementing the concept of anti-Black racism in social work is difficult, as racism has been exacerbated by neoliberal policies, globalization, and the phenomena of color blindness in the Obama era of purported postraciality. Nevertheless, putting forward the concept of anti-Black racism is critical in moving the profession toward addressing racism, colonialism, and white supremacy in Canadian society. The profession must address such questions as: Where are the programs and courses on race? If they exist, are they addressing questions of anti-Black racism? How many Black faculty members are being hired in schools of social work? How many faculty members have scholarship that includes anti-Black racism? How many Black students are registered in social work programs? Do schools recognize that due to racism in their lives, Black students may need additional support and mentoring, or to speak with a faculty member who understands their issues? Addressing these questions is vital in moving anti-Black racism theory and action forward.

Conclusion

The silencing of antiracism by anti-oppression reinforces Canada's aversion to speaking about racism and colonialism, and consequently social work scholars have proposed instead anti-Black racism and anticolonialism. The rich tradition of Aboriginal and Black Canadian resistance to racism and colonialism offers social work education and practice a critical pedagogy of survival and social transformation. There is in Canada a wealth of knowledge and experience among Aboriginal and racialized elders, community leaders, activists, social workers, students, and scholars that is untapped in social work. Our histories however have often been erased from official classroom textbooks. Anti-Black racism and anticolonialism aim to propel social work to take action and reveal these hidden histories and ways of knowing that promise to fulfill the 1987 Canadian social work education commitment to "…prepare students to work effectively with diverse ethnic, cultural, and racial populations" (CASSW, 1991: p. 3).

See also: Afrocentric Approaches to Social Work; Charity and Philanthropy: Overview; Critical Reflexivity and Social Work Practice; Critical Social Work Practice; Culturally Appropriate Interventions in Social Work; Essentialism; Gender and Academic Motivation; Human Rights and Social Work;

Migration, Ethnicity, Aging and Social Work Practice; Neoliberalism; Oppression; Prejudice and Discrimination; Race and Academic Motivation; Radical Social Work; Reflexivity; Social Action and Social Justice; Social Work Education; Whiteness and White Supremacy.

Bibliography

African Canadian Legal Clinic, July 2012. Canada's Forgotten Children: Written Submissions to the Committee on the Rights of the Child on the Third and Fourth Reports of Canada. African Canadian Legal Clinic, Toronto.

Angod, L., 2006. From post-colonial to anti-colonial politics: differences, knowledge and R.v.R.D.S. In: Dei, G.J.S., Kempf, A. (Eds.), Anti-Colonialism and Education: The Politics of Resistance. Sense Publishers, Netherlands.

Baines, D., 2010. Doing Anti-oppressive Practice, second ed. Fernwood, Halifax.

Barnoff, L., Moffatt, K., 2007. Contradictory tensions in anti-oppressive practice in feminist social services. Affilia: Journal of Women and Social Work 22, 56–70.

Baskin, C., 2010. Strong Helpers' Teachings: The Value of Indigenous Teachings in the Helping Professions. Canadian Scholars' Press, Toronto.

Benjamin, A., 2003. The Dialectics of Anti-Black Racism. Doctoral Dissertation. OISE/UT, Toronto.

Benjamin, A., 2002. The social and legal banishment of anti-racism: a black perspective. In: Chan, W., Mirchandani, K. (Eds.), Crimes of Colour: Racialization and the Criminal Justice System in Canada. Broadview Press, Toronto, pp. 177–190.

Briskman, L., 2007. Social Work with Indigenous Communities. The Federation Press, Leichhardt, NSW.

Bruyere, G., 2007. Making circles: renewing first nations ways of helping. In: Coates, J., Graham, J.R., Swartzentruber, B., Ouellette, B. (Eds.), Spirituality and Social Work: Selected Canadian Readings. Toronto Scholars Press, Toronto, pp. 259–272.

Canadian Association of Schools of Social Work, May 1991. Social Work Education at the Crossroads: The Challenge of Diversity. The Report of the Task Force on Multicultural and Multiracial Issues in Social Work Education. Canadian Association of Schools of Social Work, Ottawa.

Canadian Association of Schools of Social Work, March 2000. Canadian Association of Schools of Social Work: Anti-racist Training and Materials Project.

Carniol, B., 1987. Case Critical: The Dilemma of Social Work in Canada. Between the Lines Press, Toronto.

Carniol, B., 2010. Case Critical: Social Services and Social Justice in Canada, sixth ed. Between the Lines Press, Toronto.

Canadian Association of University Teachers, January 2010. The Changing Academy? A Portrait of Canada's University teachers. Education Review 12 (1). Retrieved from http://www.caut.ca/docs/education-review/the-changing-academy-a-portrait-of-canada-rsquo-s-university-teachers-%28jan-2010%29.pdf?sfvrsn=14.

Clarke, J., 2010. The challenges of child welfare involvement for Afro-Caribbean Canadian families in Toronto. Children and Youth Services Review 33 (2), 274–283.

Clarke, J., 2012. Beyond child protection: Afro-Caribbean service users of child welfare. Journal of Progressive Human Services 23 (3), 223–257.

Clarke, J., Aiello, O., Chau, K., Atcha, Z., Rashidi, M., Amarel, S., 2012. Uprooting social work education. LEARNing Landscapes 6 (1), 81–105.

Dei, G., 2000. Contesting the Future: Anti-Racism and Canadian Diversity. In: Nanacoo, S. (Ed.), 21st Century Canadian Diversity. Canadian Educators, Toronto, pp. 295–319.

Dei, G.S., 2008. Racists Beware: Uncovering Racial Politics in Contemporary Society. Sense Publishers, New York.

Dei, G.J.S., Kempf, A. (Eds.), 2006. Anti-colonialism and Education: The Politics of Resistance. Sense Publishers, The Netherlands.

Dominelli, L., 2002. Anti-oppressive Social Work: Theory and Practice. Palgrave Macmillan, Houndsmills Basingstoke, Hampshire (England).

Dominelli, L., 1988. Anti-racist Social Work. Macmillan, London.

Fallon, B., MacLaurin, B., Trocmé, N., Felstiner, C., 2003. A national profile of child protection workers. In: Kufeldt, K., McKenzie, B. (Eds.), Child Welfare: Connecting Research, Policy, and Practice. Wilfrid Laurier University Press, Waterloo, pp. 41–52.

Gosine, K., Pon, G., 2010. On the front lines: the voices and experiences of racialized child welfare workers in Toronto, Canada. Journal of Progressive Human Services 22 (2), 135–159.

Haque, E., 2012. Multiculturalism within a Bilingual Framework: Language, Race, and Belonging in Canada. University of Toronto Press, Toronto.

Healy, K., 2005. Social Work Theories in Context: Creating Frameworks for Practice. Palgrave MacMillan, Basing-Stoke.

Henry, F., Tator, C., 2010. The Colour of Democracy. Racism in Canadian Society. Nelson Education, Toronto.

James, C.E., 1996. Perspectives on Racism in the Human Services Sector. University of Toronto Press, Toronto.

James, C.E., 2003. Seeing ourselves: Exploring ethnicity, race and culture, 3rd Ed. Thompson Educational, Toronto.

Jeffery, D., 2005. What good is anti-racist social work if you can't master it'?: exploring a paradox in anti-racist social work education. Race, Ethnicity and Education 8 (4), 409–425.

Lavallée, L., 2007. Healing through the medicine wheel, pimatisiwin. A Journal of Aboriginal and Indigenous Community Health 5 (1).

Lawrence, B., Dua, E., 2005. Decolonizing antiracism. Social Justice 32 (4), 120–143.

Lee, E., Bhuyan, R., 2013. Negotiating within whiteness in cross-cultural clinical encounters. The Social Service Review 87 (1), 98–130.

McLaughlin, K., 2005. From ridicule to institutionalization: anti-oppression, the state and social work. Critical Social Policy 25 (3), 283–305.

Moreau, M., 1979. A structural approach to social work practice. Journal of Social Work Education 5 (1), 78–94.

Mullaly, B., 1997. Structural Social Work. McClelland & Stewart, Toronto.

Mullaly, B., 2002. Challenging Oppression. Oxford University Press, Don Mills.

Park, Yoosun, 2005. Culture as deficit: A critical discourse analysis of the concept of culture in contemporary social work discourse. Journal of Sociology and Social Welfare 32 (3), 11–34.

Pon, G., Gosine, K., Phillips, D., 2011. Immediate response: addressing anti-native and anti-black racism in child welfare. International Journal of Child, Youth and Family Studies 3 and 4, 385–409.

Pon, G., 2009. Cultural competency as new racism: an ontology of forgetting. Journal of Progressive Human Services 20 (1), 59–71.

Razack, N., 2002. Transforming the Field. Fernwood, Halifax.

Razack, N., Jeffery, D., 2002. Critical race discourse and tenets of social work. Canadian Social Work Review 17 (2), 257–271.

Razack, S., Smith, M., Thobani, S., 2010. Introduction. In: Razack, S., Smith, M., Thobani, S. (Eds.), States of Race: Critical Race Feminism for the 21st Century. Canadian Scholars Press, Toronto, pp. 1–19.

Simpson, S.J., James, C.E., Mack, J., 2011. Multiculturalism, colonialism, and racialization: conceptual starting points. The Review of Education, Pedagogy, and Cultural Studies 33, 285–305.

Sinclair, R., Hart, M.A., Bruyere, G., 2009. Wicihitowin: Aboriginal Social Work in Canada. Fernwood, Halifax.

Spivak, G., 1990. The Post-colonial Critic: Interviews, Strategies, Dialogue. Routledge, New York.

St. Denis, V., 2011. Silencing aboriginal curricular content and perspectives through multiculturalism: "there are other children here". The Review of Education, Pedagogy and Cultural Studies 33, 306–317.

Thobani, S., 2009. Exalted Subjects. University of Toronto Press, Toronto.

Thomas, W.B., 2006. Fighting for Change: Black Social Workers in Nova Scotia. Pottersfield Press, Lawrence town Beach.

Thomas, B.W., Hamilton-Hinch, B., 2006. Four journeys – one vision: ABSW comes to Halifax. In: Bernard Thomas, Wanda (Ed.), Fighting for Change: Black Social Workers in Nova Scotia. Pottersfield Press, Lawrence town Beach, pp. 18–26.

Urban Alliance on Race Relations, 2002. Saving Lives: Alternative to the Use of Lethal Force by Police. Urban Alliance on Race Relations, Toronto.

Warner, R., 2006. Theoretical Framework for the Racism. In: Lloyd, Bethan (Ed.), Violence and Health Project: A Working Paper. http://rvh.socialwork.dal.ca/resources.html. From the World Wide Web.

Walter, M., Taylor, S., Habibis, D., 2011. How white is social work in Australia? Australian Social Work 64 (1), 6–19.

Williams, C., 2005. To unnerve and detect: policing black activists in Toronto. In: Visano, L.A. (Ed.), Law and Criminal Justice: A Critical Inquiry. APF Press, Toronto, pp. 95–113.

Williams, C., 1999. Connecting anti-racist and anti-oppressive theory and practice. Retrenchment or reappraisal? British Journal of Social Work 29 (2), 211–230.

Wong, H., Yee, J.Y., 2010. Anti-oppression Framework for Child Welfare in Ontario. Ontario Association for Children's Aid Societies, Toronto.

Wood, M., 2000. Black education in Canada: past, present and future. In: Goldstein, T., Selby, D. (Eds.), Weaving Connections: Educating for Peace, Social and Environmental Justice. Sumach Press, Toronto, pp. 99–119.

Ethnicity, Sociology of

Eliezer Ben-Rafael, Tel-Aviv University, Ramat Hasharon, Israel
Yistzhak Sternberg, Beit Berl Academic College and Open University of Israel, Tel-Aviv, Israel

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E. Ben-Rafael, Volume 7, pp. 4838–4842, © 2001, Elsevier Ltd.

Abstract

In contrast to the expectations of many researchers of modernity, ethnicity has become, today more than ever, a major aspect of the social reality. This article considers, defines, and distinguishes, in present-day terms, the notions of ethnic identity, identification, and identifying. It further delves into the sway of ethnicity in contemporary transformational processes as a factor in the multiculturalization of societies and diasporas.

Introduction

Ethnicity has constituted, in various forms, a characteristic aspect of many historical societies. The word derives from *ethnikos* and means 'heathen' in Greek (Spencer, 2006: 45); it was a conspicuous phenomenon in Athens, Rome, and Alexandria and was also a factor in the formation of the Indian caste system and the *millet* in Muslim societies. Ethnicity, as a rule, contributed to the establishing of segregation among segments of the population.

The rise of modernity, and with it the spread of new universalistic values and the construction of nation-states upholding concepts of nation and citizenship, led many scholars to believe that ethnicity was now doomed to lose its social significance. Social dynamics, however, would gainsay those expectations. Today, most observers cannot but acquiesce that ethnicity has in fact become a major driving force in contemporary social allegiances and transformations.

This article defines ethnicity in present-day terms, and elicits, in a sociological vein, its principal aspects which are considered here under the notions of collective identity, ethnic identification, and identifying. This phenomenon is linked to major developments in contemporary reality, analyzed in the final section.

Defining Ethnicity

Ethnicity is a notion that refers to social entities sharing real or putative ascriptive features like a common origin or cultural–linguistic legacy which assumedly command special collective commitment, as well as their retention and transmission (see Jenkins, 2007). While this definition is widely accepted – with diverse nuances – by researchers, divergences of views and understandings also divide them in certain respects, especially regarding the designation of factors accounting for the formation of ethnic bonds, or their dislocation. More particularly, an ongoing confrontation sets in conflict the primordialist and the circumstantialist. An early formulation of the 'primordialist' view was made by Shils (1957: 42) who argued that primordialism exists when "… a certain ineffable significance [is] attributed to the tie of blood." Geertz (1973: 259–260) added that the strength of such primordial bonds, and the types they illustrate

> …differ from person to person, from society to society, and from time to time. But for virtually every person, in every society, at all times, some attachments seem to flow more from a sense of natural – some would say spiritual – affinity than from social interaction.

It is in this vein that Bayar (2009) acknowledges that primordialism offers a convincing hypothesis in light of the persistence of ethnicity-related phenomena over long periods of time, in some groups. Support for this thesis is also offered by Cavalli-Sforza et al. (1994), who found that ethnic groups may have statistically significant distinctive genetic features.

The 'circumstantialist' or 'constructivist' perspective, in contrast, is a reminder that individuals have multiple identities. These allegiances are constructed and reconstructed over time in different contexts. Changing individual and social interests (Ratcliffe, 2010; Stone and Dennis, 2003) account satisfactorily for this dynamism. One of the first formulations of this approach by Fredrik Barth (1969: 6) stated that

> Ethnicity is a matter of social organization above and beyond questions of empirical cultural differences…[it is a matter of] self-ascription and ascription by others in interaction.

Andreas Wimmer (2008) elaborates a taxonomy of strategies for ethnic boundary construction and reconstruction. He points out the practice of redrawing collective boundaries by including new members, excluding marginal segments, or challenging the hierarchical ordering of social categories.

Another debate relates to the identities that should be included in the general notion of ethnic identity. Jenkins (2007) distinguishes racial, national, regional, local, and communal identities but argues at the same time that all these identities are simply conceptual variations of the more general type of collective identity, consisting of ethnicity. Not everyone, however, endorses this kind of outlook. Van den Bergh (1967: 9) drew a distinction between ethnicity as 'socially defined on the basis of cultural criteria' and race, which is 'socially defined on the basis of physical criteria.'

Eriksen (1993) adds that an ethnic group proper never completely overlaps the nation.

For the purpose of conceptual clarity, one may propose here that ethnicity designates this general kind of collective identity of a group – which may itself be subdivided in different manners – whose members are aware that they share a real or putative common ancestry. As such, they are engaged, in Giddens' (1991) words, in a reflexive project of identity building. Like any kind of collective-identity building, this project assumes that: (1) in one way or another, individuals feel committed to people whom they see as fellow-members of their group; (2) they perceive that group as conveying some cultural singularity; and (3) they tend, to some extent and in certain respects, to distance themselves from *others* whom they consider nonmembers (Ben-Rafael, 2001). This notion of ethnic identity which focuses on the awareness of individuals, describes an essentially subjective phenomenon and, as such, can never be seen as acquired permanently. It may involve dilemmas alongside unambiguous assertions; it may be phrased in different (possibly even contradictory) terms, diachronically or synchronically, among different circles pertaining to the same group.

These definitions are not opposed to constructionist interpretations that explore contextual aspects to explain identity elaborations. On the other hand, neither do these definitions exclude the primordialist principle. They recognize primordialism, not as independent of human will, but as an attribute that individuals perceive as given, and as conveying essential significance for them. Because of its subjective nature, however, that self-perceived singularity may also be formulated in different terms, as constructivists maintain, according to personalities, milieus, or circumstances.

Identification

Another dimension of ethnicity which is often confused with that notion of identity, but points in a different direction, consists of the issue of ethnic identification. Indeed, ethnic identification does not refer to the tenets of the collective identity but rather to the question of how much importance individuals ascribe to that identity. Like the formulations of identity itself, identification may vary greatly among fellow-ethnics and these variations are manifest in daily life, social norms, and involvement in collective goals. The highest degree of ethnic identification is illustrated by individuals who see their ethnic identity as their principal, if not only, collective identity; the lowest degree of identification is exemplified by individuals who completely disregard their ethnicity and view it just as an incidental 'origin.'

Though distinct dimensions of the ethnic condition, identity, and identification are never completely independent from each other. The weakening of ethnic identification may lead individuals to reformulate their ethnic identity in less precise terms. On the other hand, where religion is a component of a group's singularity, it often cements identification on the basis of transcendental convictions. In the words of Hervieu-Léger (2009: 447), faith presupposes submission to the religious authority which warrants the permanence of truth. Religions tend to explicitly convey a notion of exclusiveness over veracity (Sternberg, 2010). As assessed by Huntington (1993: 27) religion 'discriminates sharply' between people: A person may be half-French and half-Arab but it is much more difficult to conceive of being half-Catholic and half-Muslim. Hence, the religious revival witnessed in parts of the world today (Martin, 2010), particularly especially in its fundamentalist variants (see Almond et al., 2003), strengthens ethnoreligious cohesion as well as the potential for conflict between ethnic groups and other groups or individuals (Schnapper, 2005).

Other circumstances may also have an impact on identification. As Gold (2007) has shown, ethnic identification is strong in societies where populations are clearly differentiated by unequal participation in resources. This may lead to an ethnic division of labor in society, or even to split labor markets (Bonacich, 1972). In this respect, Light and Gold (2000) speak of an ethnic economy where a certain ethnic group takes control of given professional branches and staff most positions – from command echelons to rank-and-file. Such situations result in growing ethnic identification and retention of ethnic markers, motivated by economic objectives (Lichter, 2007).

The political regime, the configuration of power in society, and opportunities for building a constituency are other factors likely to influence ethnic identification. Since democracy is grounded in competition for support between parties and leaders, groups of many kinds – like local communities, professional categories, and class-oriented lobbies – may be able to build up constituencies and caucuses. This includes ethnic groups articulating identity politics (Calhoun, 1994; Bernstein, 2005; Saha, 2007). Political entrepreneurs belonging to ethnic entities may exploit available opportunities, and the more successfully they gather public support, the greater their contribution to the recognition of their *folk* and the legitimization of their constituency as a political actor. This kind of legitimization is in fact the imposition, bottom–up, of multiculturalism as societal reality (Ben-Rafael, 2010).

That legitimization is not, however, always unwelcomed by the dominant culture, which may even encourage it. By dominant culture, we refer to the beliefs, values, and symbols sustained by society's leading political forces and which express how those forces interpret the historical legacy and long-range interests of the setting. Among other issues, this dominant culture involves basic attitudes toward social–cultural heterogeneity as features of the social order. It may encourage groups to abandon their particularism on behalf of prevailing norms and values, or to retain their particularism jointly with the endorsement of the dominant culture's hard-core principles and symbols, at the least. In Germany, for instance, the dominant culture has long emphasized the notion of *Volk* (people) and its anchorage in history and language (Brubaker, 1992). Until recently, people of non-German origin, even those living in the country for decades, continued to be seen as outsiders. In the United States of today, contrastingly, the 'salad-bowl' metaphor (Hollinger, 1995) does not preclude groups from inserting themselves in society and becoming full members, while retaining some particularistic markers. In France, 'republicanism' defines the nation as a 'community of citizens' (Schnapper, 2005) and expects new groups to fully assimilate.

All in all, the group's socioeconomic condition – whether underprivileged or better-off, the nature of the dominant culture's approach to the group – whether segregative or assimilationist, and the group's own aspirations regarding insertion in society – whether moved by the drive to fully assimilate into the mainstream or to retain markers of distinctiveness, contribute in various ways to the emergence or retention of ethnicity. Hence, socioeconomically underprivileged ethnics, an exclusionist attitude of the dominant culture vis-à-vis the group, and a strong dedication of ethnics to their singular markers should, together, contribute to the group's confinement to the social margins, and its forming a kind of *enclave*. On the other hand, socioeconomic mobility of the group's members, an inclusionist dominant culture, and a flexible and open form of collective self-perception among the people concerned, in fact advance acculturation, and beyond it, tendencies to *assimilation*. Midway, one finds intermediary cases illustrating various possibilities. A socially mobile group would therefore be likely to acculturate eagerly to the mainstream culture, though the extent to which it tends to assimilate is still influenced by the openness of the dominant culture. Where a discrepancy exists between the group's general deprivation and the dominant culture's openness, one might expect mobile elements within the group to leave it and try to assimilate on their own. This, however, is contingent on the group itself generally aspiring to fusion with the mainstream. If it does not, one would rather expect that some of the mobile elements remain faithful to their group and resist the temptation of assimilation (see Ben-Rafael, 2001 for detailed hypotheses).

In all these, we should recall that the background of many contemporary cases of ethnicity consists of the experience of migration (Banton, 2008). Most of the time, immigrants who reach a target society demonstrate an irresistible tendency to grow more and more similar to mainstream society (Brubaker, 2001). However, the ethnic boundaries do not necessarily become totally blurred, even when former migrants acquire the dominant language and norms (Alba and Nee, 2003, 2007). On this point, researchers have suggested the concept of 'segmented' assimilation to indicate that upward mobility, far-reaching acculturation, and partial assimilation of a large part of a given ethnic group may occur through persistent combined biculturalism (Vermeulen, 2010).

It should also be noted that cultural convergence does not necessarily entail the disappearance or weakening of ethnic identification – especially when insertion does not take place in a highly favorable context. Different groups of immigrants coexisting in the same environment easily become hostile to each other in the context of difficulties in obtaining access to social resources. On the other hand, immigrants may also arouse among long-established citizens anti-immigrant feelings and nativism among people who consider themselves 'true' nationals and feel socially, economically, or culturally threatened by the newcomers. Sometimes, these 'nativists' turn to violent actions as a means of intimidation, or to direct national politics in order to impose restrictive measures against immigration (see Sternberg, 2009; Schrag, 2010; Lucassen, 2005). One thinks here of Le Penism in France and similar phenomena in Germany, Austria, and the Netherlands. However, such exclusionist and hostile manifestations tend rather to strengthen immigrants' ethnic identification.

Indentifying and Being Identified

This latter aspect brings out the additional issue of how *others* see ethnics and identify them. This identifying may be appreciative or neutral, but it is quite often rather negative and loaded with prejudice and stereotypes. These *others* are ethnics' fellow-countrymen in general, who may themselves respond to a wide diversity of labeling. Hence, at the societal level, this identifying repeats itself with each particular group and demarcates the multiple lines of the multicultural segmentation of society.

Identifying is thus a distinct aspect of the discourse about ethnicity which, however, is often confused with the notions of identity and identification. Identity, we saw, refers to the tenets of a group's collective identity, while identification investigates the importance that individuals grant that identity. Identifying, for its part, designates how people or institutions refer to individuals as members of a given group: It may be based on errors of discernment and eventually influence the individuals concerned in one way or another (ultimately by effectively identifying themselves as ethnics). However, identifying per se does assess the identity, let alone the identification of people in ethnic terms: when someone perceives a person as a 'Turk' or a 'Jew' this does not in itself make that person a Turk or a Jew. This does not belie the fact that being identified by a given ethnic label always has consequences for the individual concerned – especially when the identifier possesses power and has hostile intentions.

In actual fact, identifying actors in precise ethnic terms is by no means an easy task in present-day reality. Ethnic groups, indeed, are by no means impermeable to the influence of their environment and over time, volens nolens, attenuate their cultural particularism by converging toward the mainstream culture. In this, they come to exemplify what scholars designate as *cultural hybridization*, a notion that describes a given culture's acquisition of behavior patterns and values that a different one adheres to. The result, according to Jan Nederveen Pieterse (2000), consists of cultural alterations and changes that lead in turn to greater fluidity of collectives' social boundaries. Hence, collective boundaries have nowadays come to resemble dotted lines rather than clearly traced traits. In such circumstances, people quite easily find opportunities for slipping away from their group. Once outside the group, they may join other groups – which often happens through marriage – or locate themselves in the nonethnic category where the only feeling of attachment and commitment is toward the dominant culture and the national society as a whole. Multiculturalism, indeed, consists of a condition where sociocultural groups coexist and pressurize society in various directions but where more than a few people – possibly even a majority – do not perceive themselves as ethnics, and see in the nation or the national society their only meaningful attachment.

This does not signify, however, that individuals who see themselves as nonethnics may still be wrongly identified as ethnics by those around them (Kukutai and Didham, 2012), and here lies a major loophole for ethnicity statistics (Simon and Piché, 2012). As a rule, some studies use

observable 'objective' criteria like names, skin color, or place of residence in order to identify ethnic categories which in fact do not necessarily overlap the 'subjective' criteria referring to identity and identification. This conceptual clarification is especially important at a time when ethnicity is exercising an enormous impact on social reality, as it does nowadays.

Present-Day Impacts: Diaspora and Transnationality

Many cases of present-day ethnicity display another major characteristic that adds to the phenomenon's importance as a factor driving society toward new horizons. We refer, in the context of present-day globalization, to advances in communication and transportation technology and the new configurations of international relations, enabling ethnic groups dispersed worldwide to stay interconnected enduringly and on a daily basis with fellow-ethnics settled elsewhere as well as – where relevant – with their original homeland. Diaspora has always designated a commitment that goes beyond its literally significance in Greek, i.e., *dispersion* (διασπορά) (Dufoix, 2008). This commitment means that dispersed people stemming from the same cultural or territorial origin constitute a *transnational entity* (see Glick Schiller et al., 1995; Cohen, 1997). This entity, whose boundaries cut across state borders, is in many cases structured by institutions of its own which contribute to diasporans' retention of their collective identity and identification. As shown by Robin Cohen (2009), in some cases, the diasporans' identity is cemented by powerful myths and forms a *solid*-diaspora model, while in others it exemplifies a more *liquid* model influenced by the local culture and marked by less sacred icons (Waldinger, 2011).

The formation of transnational entities is most of the time the outcome of intercountry and intercontinent migration. Other possibilities or particular forms of migration are also illustrated by recent history. Hence, Rogers Brubaker (2009) speaks of the 'accidental diaspora' that develops as a result of the demise of political entities and border changes – illustrated by the Russian population living in the Baltics who became ethnic minorities in newly independent states. Another manifestation of transnationality consists of migrants from different groups who find themselves close – socially, culturally, or linguistically – to each other, and relatively remote from the mainstream in the host society. Such groups may tend to amalgamate into a kind of *pan-diaspora*. Latin-Americans of different origins who become *Hispanics* in the United States are an example, as are *Arab Muslims* of diverse origins in Europe. This does not rule out situations in which Mexicans or Cubans, on the one hand, and Moroccans or Algerians on the other, may concomitantly continue displaying their specific identities.

Still another pattern is the *returning diaspora* consisting of people of a given origin who, for whatever reasons, decide to *come home* and resettle in their or their ancestors' original homeland. In the meantime, they have absorbed another culture, have experienced the transformation of their legacy, and adopted new visions of the world. Good examples are Brazilian Japanese (Tsuda, 2003) resettling in Japan, or Ethnic Germans from Ukraine and Russia who resettle in Germany (Münz, 2002). Such groups may become an ethnic group in their new/old country by inversion of their collective identities: What was their ethnic identity now becomes their national identity, and what was their national identity, their ethnic identity.

One of the factors that may still encourage diasporans to consider going *back home* consists of the special relations their *original homeland's* government and other various organizations often retain with their nationals abroad. The first motive of these institutions, as well as of the basic positive response by diasporans to these efforts, is probably a general feeling of national solidarity. A second and more concrete motive appears to be the fact that governments of original *homelands* may find it beneficial to nurture strong relations with communities abroad when these have become significant political actors and could support their interests (Laguerre, 2006). On the other hand, diasporans who wish to achieve public power in their society of residence may also find the backing of homelanders to be beneficial. These games belong to identity politics which strengthens groups' ethnic identification but which, by the same token, fosters society's fragmentation as well (Ben-Rafael, 2010).

Moreover, groups of diasporans do not only contribute to the multiculturalization and fragmentation of their societies but also to the multiculturalization of their own diasporas. While diasporans' retention of ethnicity – even in the form of biculturalism – erodes the cohesion of their societies of residence, the prints left by the latter on the diverse communities attached to the same diaspora tend to transform the latter into a disparate entity.

All in all, ethnicity is a major factor in the transformation of contemporary societies at the local level and a lever for multicultural transnationalism globally. It contributes to the advent of a new societal and global (dis)order (Ben-Rafael, 2010; Ben-Rafael and Sternberg, 2009). Contrary to past predictions, contemporary ethnicity has by no means disappeared. On the contrary, in a diversity of forms and interpretations, it plays a role of its own in transforming the social world.

See also: Acculturation; Collective Identity; Diaspora; Multiculturalism.

Bibliography

Alba, R., Nee, V., 2003. Remaking the American Mainstream: Assimilation and Contemporary Immigration. Harvard University Press, Cambridge, MA.

Alba, R., Nee, V., 2007. Assimilation. In: Ritzer, J. (Ed.), Blackwell Encyclopedia of Sociology. Blackwell, Malden, MA.

Almond, G.A., Appleby, R.S., Sivan, E., 2003. Strong Religion: The Rise of Fundamentalisms around the World. The University of Chicago Press, Chicago.

Banton, M., 2008. The sociology of ethnic relations. Ethnic and Racial Studies 31 (7), 1267–1285.

Barth, F. (Ed.), 1969. Ethnic Groups and Boundaries: The Social Organization of Culture Difference. Little, Brown and Co, Boston, MA.

Bayar, M., 2009. Reconsidering primordialism: an alternative approach to the study of ethnicity. Ethnic and Racial Studies 32 (9), 1639–1657.

Ben-Rafael, E., 2001. Ethnicity, sociology of. In: International Encyclopedia of the Social and Behavioral Sciences, vol. 7. Elsevier, London, pp. 4838–4842.

Ben-Rafael, E., 2010. Diaspora, Sociopedia. ISA.

Ben-Rafael, E., Sternberg, Y., 2009. Epilogue: Chaos and gestalt. In: Ben-Rafael, E., Sternberg, Y. (Eds.), Transnationalism: Diasporas and the Advent of a New (Dis) order. Brill, Leiden and Boston, pp. 687–692.

Bernstein, M., 2005. Identity politics. Annual Review of Sociology 31 (1), 47–74.

Bonacich, E., 1972. A theory of ethnic antagonism: the split labor market. American Sociological Review 37 (5), 547–559.
Brubaker, R., 1992. Citizenship and Nationhood in France and Germany. Harvard University Press, Cambridge, MA.
Brubaker, R., 2001. The return of assimilation? Changing perspectives on immigration and its sequels in France, Germany, and the United States. Ethnic and Racial Studies 24 (4), 531–548.
Brubaker, R., 2009. Accidental diasporas and external 'homelands' in Central and Eastern Europe: past and present. In: Ben-Rafael, E., Sternberg, Y. (Eds.), Transnationalism: Diasporas and the Advent of a New (Dis)order. Brill, Leiden and Boston, pp. 461–482.
Calhoun, C. (Ed.), 1994. Social Theory and the Politics of Identity. Blackwell, Oxford, UK.
Cavalli-Sforza, L.L., Menozzi, P., Piazza, A., 1994. The History and Geography of Human Genes. Princeton University Press, Princeton, NJ.
Cohen, R., 1997. Global Diasporas. An Introduction. UCL Press, London.
Cohen, R., 2009. Solid, ductile and liquid: changing notions of homeland and home in diaspora studies. In: Ben-Rafael, E., Sternberg, Y. (Eds.), Transnationalism: Diasporas and the Advent of a New (Dis)order. Brill, Leiden and Boston, pp. 117–133.
Dufoix, S., 2008. Diasporas. University of California Press, Berkeley.
Eriksen, T.H., 1993. Ethnicity and Nationalism. Pluto Press, London.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Giddens, A., 1991. Modernity and Self-Identity: Self and Society in the Late Modern Age. Polity Press, Cambridge, MA.
Glick Schiller, N., Basch, L., Blanc-Szanton, C., 1995. From immigrant to transmigrant: theorizing transnational migration. Anthropological Quarterly 68 (1), 43–68.
Gold, S.J., 2007. Race and ethnic consciousness. In: Ritzer, J. (Ed.), Blackwell Encyclopedia of Sociology. Blackwell, Malden, MA.
Hervieu-Léger, D., 2009. Roman Catholicism and the challenge of globalization. In: Ben-Rafael, E., Sternberg, Y. (Eds.), Transnationalism: Diasporas and the Advent of a New (Dis)order. Brill, Leiden and Boston, pp. 445–459.
Hollinger, D.A., 1995. Postethnic America: Beyond Multiculturalism. Basic Books, New York.
Huntington, S.P., 1993. The clash of civilizations? Foreign Affairs 72 (3), 22–49.
Jenkins, R., 2007. Ethnicity. In: Ritzer, J. (Ed.), Blackwell Encyclopedia of Sociology. Blackwell, Malden, MA.
Kukutai, T., Didham, R., 2012. Re-making the majority? Ethnic New Zealanders in the 2006 census. Ethnic and Racial Studies 35 (8), 1427–1447.
Laguerre, M.S., 2006. Diaspora, Politics and Globalization. Palgrave Macmillan Press, New York.
Lichter, M., 2007. Ethnic and racial division of labor. In: Ritzer, J. (Ed.), Blackwell Encyclopedia of Sociology. Blackwell, Malden, MA.
Light, I., Gold, S., 2000. Ethnic Economics. Academic Press, San Diego, CA.
Lucassen, L., 2005. The Immigrant Threat: The Integration of Old and New Migrants in Western Europe since 1850. University of Illinois Press, Urbana and Chicago.
Martin, D., 2010. Pentecostalism: a Christian revival sweeping the developing world. In: Ben-Rafael, E., Sternberg, Y. (Eds.), World Religions and Multiculturalism: A Relational Dialectic. Brill, Leiden and Boston, pp. 93–118.
Münz, R., 2002. Ethnos or Demos? Migration and citizenship in Germany. In: Levy, D., Weiss, Y. (Eds.), Challenging Ethnic Citizenship: German and Israeli Perspectives on Immigration. Berghahn Books, New York and Oxford, pp. 15–35.
Pieterse, J.N., 2000. Globalization as hybridization. In: Lechner, F.J., Boli, J. (Eds.), The Globalization Reader. Blackwell, Oxford, pp. 99–105.
Ratcliffe, P., 2010. Ethnic Group, Sociopedia. ISA.
Saha, S.C. (Ed.), 2007. The Politics of Ethnicity and National Identity. Peter Lang, New York.
Schnapper, D., 2005. Ethnic revival and religious revival in 'providential democracies'. In: Ben-Rafael, E., Sternberg, Y. (Eds.), Comparing Modernities. Brill, Leiden and Boston, pp. 205–221.
Schrag, P., 2010. Not Fit for Our Society: Immigration and Nativism in America. University of California Press, Berkeley.
Shils, E., 1957. Primordial, personal, sacred and civil ties: some particular observations on the relationships of sociological research and theory. The British Journal of Sociology 8 (2), 130–145.
Simon, P., Piché, V., 2012. Accounting for ethnic and racial diversity: the challenge of enumeration. Ethnic and Racial Studies 35 (8), 1357–1366.
Spencer, S., 2006. Race and Ethnicity: Culture, Identity and Representation. Routledge, London and New York.
Sternberg, Y., 2009. Contemporary immigration in comparative perspective. In: Ben-Rafael, E., Sternberg, Y. (Eds.), Transnationalism: Diasporas and the Advent of a New (Dis)order. Brill, Leiden and Boston.
Sternberg, Y., 2010. Religion, territory and multiculturalism. In: Ben-Rafael, E., Sternberg, Y. (Eds.), World Religions and Multiculturalism: A Relational Dialectic. Brill, Leiden and Boston, pp. 353–365.
Stone, J., Dennis, R. (Eds.), 2003. Race and Ethnicity: Comparative and Theoretical Approaches. Blackwell, Oxford, UK.
Tsuda, T., 2003. Strangers in the Ethnic Homeland: Japanese Brazilian Return Migration in Transnational Perspective. Columbia University Press, New York.
van den Bergh, P.L., 1967. Race and Racism: A Comparative Perspective. John Wiley, New York.
Vermeulen, H., 2010. Segmented assimilation and cross-national comparative research on the integration of immigrants and their children. Ethnic and Racial Studies 33 (7), 1214–1230.
Waldinger, R., 2011. Immigrant Transnationalism, Sociopedia. ISA.
Wimmer, A., 2008. Elementary strategies of ethnic boundary making. Ethnic and Racial Studies 31 (6), 1025–1055.

Ethnoarchaeology

Margaret Beck, University of Iowa, Iowa City, IA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W.A. Longacre, volume 7, pp. 4842–4846, © 2001, Elsevier Ltd.

Abstract

Ethnoarchaeology is an important kind of middle-range research, consisting of ethnographic fieldwork undertaken by archaeologists. Ethnoarchaeologists study the material traces of ongoing human behavior so that archaeologists can better infer past human behavior from material patterns. Ethnoarchaeological fieldwork has been conducted all over the world, in settings ranging from hunting-gathering camps to state-level societies, and addresses a wide variety of material remains and artifact classes. Work focusing on refuse disposal and midden formation is highlighted in this article to illustrate ethnoarchaeology's contributions and relevance for the archaeological record.

Archaeologists study human societies, as do other anthropologists, but they study them primarily through their material culture (Renfrew and Bahn, 2012). The public may think of archaeologists such as those studying Maya culture, painstakingly toiling in Honduras to collect sixth-century artifacts from a site like Copán. These scholars do not observe Classic Maya society directly, but instead see the end products of behavior, such as artifacts, features, and spatial patterns.

How do archaeologists make sense of the archaeological record to learn about the people who created it? They must study the relationships between human behavior and material culture (Reid et al., 1974, 1975), specifically developing links or correlates between dynamic events and activities and their static material traces (Schiffer, 1975). This study of cause and effect has been labeled 'middle-range research' (Binford, 1978). Ethnoarchaeology is an important kind of middle-range research, consisting of ethnographic fieldwork undertaken by archaeologists. Ethnoarchaeologists study the material traces of ongoing human behavior so that archaeologists can better infer past human behavior from material patterns.

Other disciplines offer information on the material traces of ongoing human behavior too, including sociocultural anthropology, history, sociology, and geography. Data originally collected for nonarchaeological applications may be extremely useful to archaeologists, but is often too vague on crucial points of interest. An archaeological perspective is important for understanding "the relationships of material culture to culture as a whole" (David and Kramer, 2001: 2). Other labels for ethnoarchaeology include action archaeology, living archaeology, or archaeological ethnography.

Applying ethnoarchaeological data to archaeological cases (and the use of middle-range research in general) requires the use of cross-cultural generalizations, albeit with defined limits (or boundary conditions) to their applications (David and Kramer, 2001). One early example is Ascher's (1961) claim that analogies can be drawn between cultures that adapted to similar environments in similar ways. In this respect, ethnoarchaeology differs from the direct-historical approach, which applies knowledge of a group's current or recent culture to the study of its prehistory and depends upon cultural continuity over time. Although Jesse Walter Fewkes was the first to use the term 'ethnoarchaeologist' in 1900, Fewkes meant an archaeologist using the direct-historical approach and not an ethnoarchaeologist as currently defined (David and Kramer, 2001).

Ethnoarchaeology among hunter-gatherers may have had the biggest impact on archaeological interpretation, at least as measured in citation rates in *American Antiquity* articles (Skibo, 2009). Two recent applications of hunter-gatherer ethnoarchaeology include the use of data on site structure to interpret an open-air Paleoindian site (Bamforth et al., 2005) and use of data on long-distance logistical mobility to interpret middle Holocene sites in the Great Lakes region (Lovis et al., 2005). But the field of ethnoarchaeology is quite broad, conducted all over the world in settings ranging from hunting-gathering camps to state-level societies (see overview by David and Kramer, 2001). Topics include a wide variety of material remains and artifact classes, such as architecture, faunal bone, botanical remains, basketry, flaked stone, ceramics, glass, and metals. Activities and cultural processes have also been studied, such as the use of space within buildings, sites and settlements, and broader landscapes; site abandonment; and the development and transition of decoration and style.

The Development of Modern Ethnoarchaeology

In their book *Ethnoarchaeology in Action*, David and Kramer (2001) divide the history of ethnoarchaeology into three periods – Initial, New Ethnoarchaeology, and Recent – and I follow this framework and their characterization of these periods. The Initial period is mostly explorative and descriptive. David and Kramer mark the beginning of modern ethnoarchaeology with Maxine Kleindienst and Patty Jo Watson's (1956) paper *Action archaeology: the archaeological inventory of a living community*, in which Kleindienst and Watson argue that archaeologists rather than cultural anthropologists need to conduct ethnoarchaeological fieldwork. Other landmark studies in this period include Raymond Thompson's (1958) *Modern Yucatecan Maya Pottery Making* and George Foster's (1960) article *Life expectancy of utilitarian pottery in Tzintzuntzan, Michoacan, Mexico*, both of which explicitly consider archaeological application of their data.

The orientation of ethnoarchaeology shifted with the development of the New Archaeology or processual archaeology (e.g., Binford, 1962; *see* Theory in Archaeology; Archaeology: Philosophy and Science), producing the New Ethnoarchaeology. Processual archaeologists often use ethnoarchaeology as a source of testable hypotheses within an explicitly scientific framework. A related school of thought, behavioral archaeology, promotes using ethnoarchaeology to develop 'laws' that relate variables in behavior and material culture (Schiffer, 1978).

With its focus on environmental adaptation, the New Archaeology directed ethnoarchaeology toward studies of landscape and resource use. Many classic ethnoarchaeological studies of foragers were conducted during this period in the 1960s and 1970s, including work with the Ngatatjara Aborigines in Australia (e.g., Gould, 1980), the !Kung San in Southern Africa (e.g., Yellen, 1977), the Nunamiut in Alaska (e.g., Binford, 1978), and the Alyawara of central Australia (e.g., O'Connell and Hawkes, 1984). Influential studies of agricultural villagers from this period include research by Patty Jo Watson and Carol Kramer in Iran (e.g., Kramer, 1982; Watson, 1979) and William Longacre in the Philippines (e.g., Longacre, 1974).

David and Kramer (2001) start the Recent period with the publication of Ian Hodder's (1982) book *Symbols in Action*, in which Hodder argues that artifacts are not only passive end products of behavior, but also active participants in culture. This period is characterized not only by the continuation of New Ethnoarchaeology research agendas and projects, but also by the growing influence of post-processualism, which downplays function and adaptation and emphasizes the role of human choice and agency in culture. The sociologist Pierre Bourdieu and his practice theory became influential for Hodder and for many subsequent archaeologists.

Throughout the Recent period and up to the present, we see the rewards of three decades or more of modern ethnoarchaeology. They include published results from multiyear research projects such as work with the Hadza in Tanzania (e.g., O'Connell et al., 1988), the 1970s fieldwork of the Coxoh Ethnoarchaeology Project in the Maya highlands (e.g., Hayden and Cannon, 1984), the 1970s and later fieldwork of the Kalinga Ethnoarchaeological Project (e.g., Longacre and Skibo, 1994; Stark and Skibo, 2007), and the 1980s and later fieldwork of the Mandara Archaeological Project in Cameroon and Nigeria (e.g., David et al., 1988). They also include volumes that collect ethnoarchaeological data from multiple projects and regions, along with other kinds of data, to illuminate a particular topic. Examples are single-authored monographs such as Dean Arnold's (1985) *Ceramic Theory and Cultural Process*, Richard Blanton's (1993) *Houses and Households: A Comparative Study*, and Robert Kelly's (1995) *The Foraging Spectrum*, and edited volumes such as *Ceramic Ethnoarchaeology* (Longacre, 1991), *The Interpretation of Archaeological Spatial Patterning* (Kroll and Price, 1991), *Domestic Architecture and the Use of Space* (Kent, 1991), *The Ethnoarchaeology of Refuse Disposal* (Staski and Sutro, 1991), *From Bones to Behavior* (Hudson, 1993), and *Ethnozooarchaeology* (Albarella and Trentacoste, 2011). These synthetic works make use of global "cross-cultural comparisons as a uniformitarian strategy for learning in anthropology," as advocated by Lewis Binford (2001: 5) throughout his career.

Not all archaeologists believe ethnoarchaeology can produce scientific law-like generalizations about the relationships between human behavior and material culture, or that such generalizations can be applied broadly enough to be meaningful or useful (e.g., Hegmon, 2000). The long-term impact of classic ethnoarchaeological works is being reassessed by a recent series in the journal *Ethnoarchaeology: Journal of Archaeological, Ethnographic, and Experimental Studies*, which features retrospective essays by prominent ethnoarchaeologists coupled with an outside review. The profiled ethnoarchaeologists and works include not only Dean Arnold and *Ceramic Theory and Cultural Process* (Arnold, 1985, 2011; Harry, 2011), as cited above, but also Patty Jo Watson and *Archaeological Ethnography in Western Iran* (Kamp, 2010; Watson, 1979, 2010) and Richard Gould and *Living Archaeology*, which addresses Aboriginal lifeways in Australia (Gould, 1980, 2011; Holdaway and Allen, 2011). The area-focused works by Watson (1979) and Gould (1980) provide not only analogies for interpreting specific archaeological patterns, but also broaden the debate about the applicability of cross-cultural patterns. In spite of vigorous argument by both authors that their findings have cross-cultural relevance, they are most relied upon by archaeologists within their respective regions. Both works also underscore the complexity of the interactions between people and objects, and include information of broader anthropological interest with perhaps no clear archaeological application. In contrast, Arnold (1985) compiles data from multiple sources to create explicitly cross-cultural models, some of which (such as the distance to resources model) have been very widely applied. All of these prominent ethnoarchaeologists and many others have helped to bring a greater understanding of anthropological context into archaeological work, with or without the production of specific, widely used correlates.

Ethnoarchaeology is still a vibrant, productive discipline, as shown by publications in the last decade, with work continuing among both foraging and agricultural societies (e.g., Arthur, 2006; Jones, 2009; Lupo, 2006; Politis, 2007; Searcy, 2011). In addition to *Ethnoarchaeology*, which began publication in 2009, ethnoarchaeological work also frequently appears in other outlets including the journals *American Antiquity*, *Journal of Archaeological Method and Theory*, *Journal of Archaeological Science*, and in particular *Journal of Anthropological Archaeology*.

Ethnoarchaeology at Work: The Study of Refuse Deposits

To illustrate a deliberate cross-cultural approach to the application of ethnoarchaeological results, I turn to examples of ethnoarchaeology's contributions in the area of refuse or midden deposits. Refuse deposits are vitally important to the archaeologist, because most of the archaeological records consist of intentionally discarded objects. To make valid archaeological interpretations from them, we need to understand how items reach these deposits, how to determine which households or activity areas supply refuse to a particular deposit, and how the sample of material culture here represents material culture in use. If we come from industrialized state-level societies, the refuse disposal systems

of the societies we study are often very different from our own. Ethnoarchaeology is a valuable strategy to bridge the gap in our understanding, helping us to see the general principles of refuse disposal in different settings and to model how these deposits formed in archaeological cases (see papers in Staski and Sutro, 1991).

People handle refuse in different ways depending upon the density of settlement and the length of occupation (Schiffer, 1987; Wilson, 1994). Primary refuse, discarded where the activity took place, is more likely during brief occupations with outside activities. Secondary refuse is moved away from the use location for disposal, in increasingly formal ways with increasing population density. Formal, repeatedly used secondary refuse deposits are known as middens (Needham and Spence, 1997). Modern U.S. residents may rarely, if ever, see their middens on steroids – landfills – in which their discarded materials rest, and determining where a specific household's trash goes is a real challenge (Royte, 2006). In smaller communities and less formalized refuse disposal systems, including those earlier in U.S. history (Crane, 2000), the catchment areas of middens are smaller and physically closer.

Several communities studied ethnoarchaeologically – including the Maya in Chiapas, Mexico (Deal, 1985, 1998; Hayden and Cannon, 1983), households in the Sierra de los Tuxtlas of Veracruz, Mexico (Arnold, 1990), and the Shipibo-Conibo in the Peruvian Amazon (DeBoer and Lathrap, 1979) – have household midden accumulations immediately around the perimeter of the house clearing or house lot. Eventually, as documented in Chiapas, these deposits (known as 'toft' deposits) might be hauled away to ravines or neighborhood dumps (Hayden and Cannon, 1983: 146, Table 1).

A different midden accumulation pattern was seen in a small rural village in the mountains of northern Luzon, the Philippines, where there is limited availability of flat surfaces for building increased settlement density. Working as part of the Kalinga Ethnoarchaeological Project, Beck and Hill (Beck, 2006; Beck and Hill, 2004) documented 32 active middens in a community with 70 households. Even though household trash was not always sent to tofts or household middens next to the house lot, the catchment area of most middens could be reconstructed. Most households sent their daily household waste to only one midden, although some used two or three. Even when a household used more than one midden, however, it sent most of its waste to only one. The middens themselves were divided into groups based on how many households were in the catchment area. Household middens were used by only one household (receiving trash from, on average, 6 m away). Local middens were used by 2–5 households (receiving trash from, on average, 11 m away), and communal middens were used by six or more households (receiving trash from, on average, 35 m away). Household and local midden users travel less distance to discard material because they tend to live around the village exterior or next to an empty house lot, and therefore have the space for a midden by the house. Communal midden users live in the more densely occupied areas of the village and, with houses rather than empty space on all or most sides, must use a midden located in some other portion of the village. All the communal middens are located along the central axis of the village.

In marked contrast to modern U.S. discard behavior (Rathje and Murphy, 1992; Rathje et al., 1992), edible food for humans and materials that can be reused or recycled do not make it into ethnoarchaeological middens (Beck, 2006). Materials with any potential, whatsoever, are kept instead in 'provisional discard' (Deal, 1985: 259) near houses. Adult humans therefore do not scavenge in middens, but animals do, and dogs in particular take a heavy toll on discarded faunal remains. Children play in middens as well, breaking up objects that entered the midden in larger pieces and moving items around (see Hammond and Hammond, 1981). If trash deposits start to interfere with village needs (for example, if they become too large or too smelly, or if they sit on land now desired for another purpose), they will be moved or cleaned up by burning. This disturbance reduces the average ceramic sherd size significantly.

Ethnoarchaeological studies of modern garbage make important research contributions not only to archaeology, but to social science more broadly. The Garbage Project was founded at the University of Arizona in 1971 by faculty member William Rathje (Rathje and Murphy, 1992; Rathje et al., 1992). A Harvard-trained archaeologist who had previously studied the Classic Maya, Rathje used archaeological techniques to investigate modern household trash. This was a very productive way to examine modern household behavior and address social problems such as food waste and rapid landfill accumulation. The accumulation of nonbiodegradable trash is increasingly a problem in areas with limited space for disposal (see Jones et al., 2012 for a case on the small Fijian island of Nayau).

Garbage Project data were so illuminating, in part, because they proved to be a necessary complement to other research techniques. There were marked differences in data from the 'front door' (interviews) versus 'back door' (garbage). Garbage, as a nonreactive measure of behavior, "is the evidence of what people *did*, not what they *think* they did, what they think they should have done, or what they think the interviewer thinks they should have done" (Harrison et al., 1975: 13). For example, food consumption and food waste patterns were very different in the garbage versus survey data, with vast underreporting of food waste and alcohol consumption and vast overreporting of vegetable consumption. The Garbage Project illustrates that informants may not provide complete or accurate, and that material culture is another important line of information.

Ethnoarchaeology, as we now know it, has been part of our discipline for over 50 years. In that time, it has increased our understanding of the role of material culture in human lives – often (but not always) in ways that help us interpret material patterning in the archaeological record. Many archaeologists study societies very different from their own and do so indirectly because they can only observe the results of activities, not the activities themselves. Ethnoarchaeology cannot directly show us the people we seek to understand, but it can help us infer what they did, by providing better knowledge of comparable economies, crafts, and activities as well as site formation principles.

See also: Archaeology, Theory in; Ceramics in Archaeology; Ethnography; Fieldwork in Social and Cultural Anthropology; Household Archaeology; Hunter-Gatherer Societies, Archaeology of; Hunting and Gathering Societies: Anthropology; Materiality and Culture; Zooarchaeology.

Bibliography

Albarella, U., Trentacoste, A. (Eds.), 2011. Ethnozooarchaeology. Oxbow Press, Oxford.

Arnold, D.E., 1985. Ceramic Theory and Cultural Process. Cambridge University Press, Cambridge.

Arnold, D.E., 2011. Ceramic theory and cultural process after 25 years. Ethnoarchaeology 3, 63–98.

Arnold, P.J., 1990. The organization of refuse disposal and ceramic production within contemporary Mexican houselots. American Anthropologist 92, 915–932.

Arthur, J.W., 2006. Living with Pottery: Ethnoarchaeology among the Gamo of Southwest Ethiopia. University of Utah Press, Salt Lake City.

Ascher, R., 1961. Experimental archeology. American Anthropologist 63, 793–816.

Bamforth, D.B., Becker, M., Hudson, J., 2005. Intrasite spatial analysis, ethnoarchaeology, and Paleoindian land-use on the Great Plains: the Allen site. American Antiquity 70, 561–580.

Beck, M.E., 2006. Midden ceramic assemblage formation: an ethnoarchaeological case study from Kalinga, Philippines. American Antiquity 71, 27–51.

Beck, M.E., Hill Jr., M.E., 2004. Rubbish, relatives, and residence: the family use of middens. Journal of Archaeological Method and Theory 11, 297–333.

Binford, L.R., 1962. Archaeology as anthropology. American Antiquity 28, 217–225.

Binford, L.R., 1978. Nunamiut Ethnoarchaeology. Academic Press, New York.

Binford, L.R., 2001. Constructing Frames of Reference: An Analytical Method for Archaeological Theory Building Using Ethnographic and Environmental Data Sets. University of California Press, Berkeley.

Blanton, R.E., 1993. Houses and Households: A Comparative Study. Springer-Verlag, New York.

Crane, B.D., 2000. Filth, garbage, and rubbish: refuse disposal, sanitary reform, and nineteenth century yard deposits in Washington, DC. Historical Archaeology 34, 20–38.

David, N., Kramer, C., 2001. Ethnoarchaeology in Action. Cambridge University Press, Cambridge.

David, N., Sterner, J., Gavua, K.B., 1988. Why pots are decorated. Current Anthropology 29, 365–389.

Deal, M., 1985. Household pottery disposal in the Maya Highlands: an ethnoarchaeological interpretation. Journal of Anthropological Archaeology 4, 243–291.

Deal, M., 1998. Pottery Ethnoarchaeology in the Central Maya Highlands. University of Utah Press, Salt Lake City.

DeBoer, W.R., Lathrap, D., 1979. The making and breaking of Shipibo-Conibo ceramics. In: Kramer, C. (Ed.), Ethnoarchaeology: Implications of Ethnography for Archaeology. Columbia University Press, New York, pp. 102–138.

Foster, G., 1960. Life-expectancy of utilitarian pottery in Tzintzuntzan, Michoacan, Mexico. American Antiquity 25, 606–609.

Gould, R.A., 1980. Living Archaeology. Cambridge University Press, Cambridge.

Gould, R.A., 2011. From living archaeology to disaster archaeology. Ethnoarchaeology 3, 187–202.

Hammond, G., Hammond, N., 1981. Child's play: a distorting factor in archaeological distribution. American Antiquity 46, 634–636.

Harrison, G.G., Rathje, W.L., Hughes, W.W., 1975. Food waste behavior in an urban population. Journal of Nutrition Education 7, 13–16.

Harry, K.G., 2011. Building ceramic theory: a twenty-five year retrospective on Dean Arnold's work. Ethnoarchaeology 3, 99–104.

Hayden, B., Cannon, A., 1983. Where the garbage goes: refuse disposal in the Maya Highlands. Journal of Anthropological Anthropology 2, 117–163.

Hayden, B., Cannon, A., 1984. The Structure of Material Systems: Ethnoarchaeology in the Maya Highlands. In: SAA Papers, vol. 3. Society for American Archaeology, Washington, DC.

Hegmon, M., 2000. Advances in ceramic ethnoarchaeology. Journal of Archaeological Method and Theory 7, 129–137.

Holdaway, S.J., Allen, H., 2011. A retrospective review of Richard A. Gould's living archaeology. Ethnoarchaeology 3 (2), 203–220.

Hudson, J. (Ed.), 1993. From Bones to Behavior: Ethnoarchaeological and Experimental Contributions to the Interpretation of Faunal Remains. Southern Illinois University, Carbondale. Center for Archaeological Investigations Occasional Paper Series Issue 21.

Jones, S., 2009. Food and Gender in Fiji: Ethnoarchaeological Explorations. Lexington Books, Lanham, MD.

Jones, S., McCown, A., Messersmith, M., Andrews, C., Cormier, L., 2012. Talking trash. Ethnoarchaeology 4, 147–184.

Kamp, K., 2010. The impact of Patty Jo Watson's ethnography. Ethnoarchaeology 2, 248–254.

Kelly, R.L., 1995. The Foraging Spectrum: Diversity in Hunter-Gatherer Lifeways. Smithsonian Institution Press, Washington, DC.

Kent, S. (Ed.), 1991. Domestic Architecture and the Use of Space: An Interdisciplinary Cross-cultural Study. New Directions in Archaeology Series. Cambridge University Press, Cambridge.

Kleindienst, M.R., Watson, P.J., 1956. Action archaeology: the archaeological inventory of a living community. Anthropology Tomorrow 5, 75–78.

Kramer, C., 1979. Ethnoarchaeology: Implications of Ethnography for Archaeology. Columbia University Press, New York.

Kramer, C., 1982. Village Ethnoarchaeology: Rural Iran in Archaeological Perspective. Studies in Archaeology. Academic Press, New York.

Kroll, E.M., Price, T.D. (Eds.), 1991. The Interpretation of Archaeological Spatial Patterning. Plenum Press, New York.

Longacre, W.A., 1974. Kalinga pottery-making: the evolution of a research design. In: Leaf, M.J. (Ed.), Frontiers of Anthropology. D. Van Nostrand, New York, pp. 51–67.

Longacre, W.A. (Ed.), 1991. Ceramic Ethnoarchaeology. The University of Arizona Press, Tucson.

Longacre, W.A., Skibo, J.M. (Eds.), 1994. Kalinga Ethnoarchaeology: Expanding Archaeological Method and Theory. Smithsonian Institution Press, Washington, DC.

Lovis, W.A., Donahue, R.E., Holman, M.B., 2005. Long-distance logistic mobility as an organizing principle among holocene hunter-gatherers: a great lakes middle holocene settlement system. American Antiquity 70, 669–693.

Lupo, K.D., 2006. What explains the carcass field processing and transport decisions of contemporary hunter-gatherers? Measures of economic anatomy and zooarchaeological skeletal part representation. Journal of Archaeological Method and Theory 13, 19–66.

Needham, S., Spence, T., 1997. Refuse and the formation of middens. Antiquity 71, 77–90.

O'Connell, J., Hawkes, K., Blurton Jones, N., 1988. Hadza hunting, butchering, and bone trans- port and their archaeological implications. Journal of Anthropological Research 44, 113–161.

O'Connell, J., Hawkes, K., 1984. Food choice and foraging sites among the alywara. Journal of Anthropological Research 40, 504–535.

Politis, G.G., 2007. Nukak: Ethnoarchaeology of an Amazonian People. Left Coast Press, Walnut Creek, CA.

Rathje, W.L., Hughes, W.W., Wilson, D.C., Tani, M.K., Archer, G.H., Hunt, R.G., Jones, T.W., 1992. The archaeology of contemporary landfills. American Antiquity 57, 437–447.

Rathje, W.L., Murphy, C., 1992. Rubbish! The Archaeology of Garbage: What Our Garbage Tells Us about Ourselves. HarperCollins, New York.

Reid, J.J., Rathje, W.L., Schiffer, M.B., 1974. Expanding archaeology. American Antiquity 39, 125–126.

Reid, J.J., Schiffer, M.B., Rathje, W.L., 1975. Behavioral archaeology: four strategies. American Anthropologist 77, 864–869.

Renfrew, C., Bahn, P., 2012. Archaeology: Theories, Methods, and Practice. Thames and Hudson, New York.

Royte, E., 2006. Garbage Land: On the Secret Trail of Trash. Little, Brown, New York.

Schiffer, M.B., 1975. Archaeology as behavioral science. American Anthropologist 77, 836–848.

Schiffer, M.B., 1978. Methodological issues in ethnoarchaeology. In: Gould, R.A. (Ed.), Explorations in Ethnoarchaeology. University of New Mexico Press, Albuquerque, pp. 229–247.

Schiffer, M.B., 1987. Formation Processes of the Archaeological Record. University of New Mexico Press, Albuquerque.

Searcy, M.T., 2011. The Life-giving Stone: Ethnoarchaeology of Maya Metates. The University of Arizona Press, Tucson.

Skibo, J.M., 2009. Archaeological theory and snake-oil peddling: the role of ethnoarchaeology in archaeology. Ethnoarchaeology 1, 27–56.

Stark, M.T., Skibo, J.M., 2007. A history of the Kalinga Ethnoarchaeological Project. In: Skibo, J.M., Graves, M.W., Stark, M.T. (Eds.), Archaeological Anthropology: Perspectives on Method and Theory. The University of Arizona Press, Tucson, pp. 93–110.

Staski, E., Sutro, L.D. (Eds.), 1991. The Ethnoarchaeology of Refuse Disposal. Arizona State University Anthropological Research Papers No. 42, Tempe.

Thompson, R.H., April 1958. Modern Yucatecan Maya Pottery Making: Memoirs of the Society for American Archaeology, vol. 23, No. 4, Part 2.

Watson, P.J., 1979. Archaeological Ethnography in Western Iran. Columbia University Press, New York.

Watson, P.J., 2010. Archaeological ethnography in Western Iran, 1959–1960. Ethnoarchaeology 2, 235–247.

Wilson, D.C., 1994. Identification and assessment of secondary refuse aggregates. Journal of Archaeological Method and Theory 1, 41–68.

Yellen, J.E., 1977. Archaeological Approaches to the Present: Models for Reconstructing the Past. Studies in Archaeology. Academic Press, New York.

Ethnocentrism

RA LeVine, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term ethnocentrism passed from social science theory into common English usage during the twentieth century. In present usage it means the belief that one's culture is superior to others the use of a frame of reference derived from one's own culture to judge the arrtibutes of another culture, often in disparaging terms. The term was apparently invented in 1906 by the sociologist William Graham Sumner, who conceptualized it as a means of promoting solidarity within what he called 'ingroups' and antagonism toward 'outgroups.' He claimed ethnocentrism in this sociological as well as cultural sense to be a universal tendency of intergroup relations among humans. After 1950 Sumner's theory came under criticism from sociologists, who argued that individuals often belong to multiple groups and often admire outgroups, and from anthropologists, who described parts of the world in which ethnic identities, and thus alliances and antagonisms, were unstable historically. As Sumner's sociological formulation and claims of universality lost credibility, the vlaue of the ethnocentrism concept as a descriptor of recurrent attitudes gave it a permanent home in the English language.

Ethnocentrism is a term that passed from the social sciences into broader English discourse during the twentieth century. In current usage, it means culturally biased judgment, i.e., applying the frame of reference provided by one's culture to an object, action, person, or group of a different culture. The term as introduced by William Graham Sumner in his seminal treatise *Folkways* (Sumner, 1906) was explicitly sociological and included biased judgment as one of several features in a theory linking group solidarity with intergroup antagonism as a universal characteristic of human societies. Sumner's theory influenced research on ethnic relations and nationalism during the first half of the twentieth century. Later, the critique of Sumner by Merton and Rossi (1957) gave attention to the concept of ethnocentrism as problematic in social research, and LeVine and Campbell (1972) used the term as a rubric under which to review theories of inter- group relations and attitudes and to construct a framework for comparative research on the subject. By the end of the century, however, interest in ethnocentrism as a sociological concept had waned, even as the adjective 'ethnocentric' was commonly used as a descriptor or epithet in discussions of intercultural attitudes.

Sumner's brief but dense and multifaceted discussion of ethnocentrism in *Folkways* remains a primary source of theory on this subject. Although Sumner acknowledged previous work by the Austrian sociologist Ludwig Gumplowicz and Coser (1956) has cited Marx and other earlier theorists for certain aspects of ethnocentrism, Sumner himself may have coined the term ethnocentrism (as well as 'ingroup' and 'outgroup') and was responsible for endowing it with wide-ranging theoretical implications for the social sciences:

> The conception of 'primitive society' which we ought to form is that of small groups scattered over a territory. … A group of groups may have some relation to each other … which draws them together and differentiates them from others. Thus a differentiation arises between ourselves, the we-group, or in-group, and everybody else, or the others-group, out-groups. The insiders in a we-group are in a relation of peace, order, law, government and industry, to each other. Their relation to all outsiders, or others-groups, is one of war and plunder, except so far as agreements have modified it … .
>
> .. *Sentiments in the in-group and towards the out-group.* The relation of comradeship and peace in the we-group and that of hostility and war towards others-groups are correlative to each other. The exigencies of war with outsiders are what makes peace inside, lest internal discord should weaken the we-group for war. These exigencies also make government and law in the in-group, in order to prevent quarrels and enforce discipline. Thus war and peace have reacted on each other and developed each other, one within the group, the other in the intergroup relation. … Sentiments are produced to correspond. Loyalty to the group, sacrifice for it, hatred and contempt for outsiders, brotherhood within, warlikeness without – all group together, common products of the same situation. These relations and sentiments constitute a social philosophy. It is sanctified by connection with religion.
>
> … Virtue consists in killing, plundering, and enslaving outsiders.
>
> ... *Ethnocentrism* is the technical name for this view of things in which one's own group is the center of everything, and all others are scaled and rated with reference to it. Folkways correspond to it to cover both the inner and the outer relation. Each group nourishes its own pride and vanity, boasts itself superior, exalts its own divinities, and look with contempt on outsiders. Each group thinks its own folkways the only right ones, and if it observes that other groups have other folkways, these excite its scorn. Opprobrious epithets are derived from these differences. 'Pig-eater', 'cow-eater', 'uncircumcised', 'jabberers', are epithets of contempt and abomination. The Tupis called the Portuguese by a derisive epithet descriptive of birds which have feathers around their feet, on account of trousers. . For our present purpose the most important fact is that ethnocentrism leads a people to exaggerate and intensify everything in their own folkways which is peculiar and which differentiates them from others. It therefore strengthens the folkways (Sumner, 1906: pp. 12–13).

This passage shows both the scope of Sumner's discussion – covering attitudes and stereotypes as well as social relationships, organizational features and warfare – and his commitment to the theory that ingroup-outgroup dynamics generate internal solidarity and cultural coherence as well as hostility and warfare. Sumner claimed these dynamics and the concomitant attitudes to be universal in human societies, although he was aware that not all peoples were warlike. Social scientists before and after Sumner discovered many of the same phenomena in their own

research, but his influential passage became a point of reference for subsequent social theory and research on war and peace, ethnic relationships, and intergroup conflicts and attitudes.

Ethnocentrism as a process simultaneously promoting ingroup solidarity and outgroup hostility has been described at different levels, in many places and historical periods, not only by sociologists and anthropologists but also by political scientists, historians, and journalists. It is as familiar now, particularly in ethnic politics, as it was in Sumner's time. But Sumner's theoretical claims have been criticized and largely rejected by sociologists and anthropologists. Merton and Rossi (1957), in formulating the sociological theory of reference groups, pointed out that individuals often belong to, or wish to belong to, more than one group, making the ingroup-outgroup boundary ambiguous and the outgroup often an object of admiration rather than contempt. They argued that the ethnocentrism syndrome as Sumner formulated it should be seen as a special case of human intergroup relations rather than universal, as he claimed.

Anthropologists have been similarly critical. Describing many parts of the world where ethnic groups do not have bounded territories and individuals and groups change ethnic affiliation, anthropologists cast serious doubt on generalizations presuming that human groups have clear boundaries and stable membership (Leach, 1954; Barth, 1956, 1969; Goody, 1956; Moerman, 1965; LeVine and Campbell, 1972: pp. 81–113). Furthermore, the threat-dependent ethnic and national identities that Sumner assumed were products of a natural human tendency have increasingly been seen by anthropologists as deliberately constructed by leaders to gain adherents and seize power. Finally, it now seems likely that the intense nationalism and threatening military competition of the Western world in 1906 predisposed Sumner and his readers to accept ethnocentrism as a universal fact of human social life in advance of the comparative evidence. With better evidence and a different experience of international relations, we are less inclined to accept his theory.

Thus the sociology of Sumner's ethnocentrism syndrome, viewed in the light of recent evidence and newer perspectives, seems a dated fallacy of early twentieth century social science. The central problem has been that without clear and stable boundaries between ingroup and outgroup, the question of whether ingroup solidarity and outgroup hostility are consistently interdependent cannot be unequivocally answered. Group boundaries, if they ever were clear, have become more and more blurred as international migration and communication have intensified. Sumner's concept, as he generalized it, is not an accurate guide to the real world, but the question arises, if Sumner's formulation of the ethnocentrism syndrome has no merit as social theory, why does it seem so familiar as ethnographic description? How it is that, despite globalization, the contemporary world produces so many examples of conflict between ethnic groups differing in language, religion, and other customary practices, with mutual disparagement and threat-dependent solidarity escalating to intergroup violence? Answering this question through social research requires the investigation of the ethnocen-trism syndrome as a contingent historical phenomenon, examining in detail how the ideological and organizational patterns to which Sumner called attention have emerged in particular cases and how long they have lasted.

The term ethnocentrism remains in use to refer to culturally biased judgment, the act of applying the frame of reference from one's own culture to someone or something belonging to another culture. The *Dictionary of Anthropology* (Barfield, 1997: p. 55) defines ethnocentrism as 'the belief that one's own culture is superior to others, which is often accompanied by a tendency to make invidious comparisons. In a weaker form, ethnocentrism is the tendency to look at other cultures through the filter of one's own cultural presuppositions.' As in Sumner's examples in the passage quoted, ethnocentric judgments are usually derisive epithets, in which outgroup practices that deviate from ingroup norms form the basis for an insulting characterization: 'pig-eaters,' 'uncircum-cised,' 'jabberers.' This can reflect a naive form of ethnocentrism, called phenomenal absolutism by Segall et al. (1966), in which a person unreflectively takes his own culture's values as objective reality and automatically uses them as the context in which he judges less familiar objects and events. Alternatively, it can reflect a judgment made with full knowledge of cultural variability, but with an equally or even more scornful attitude. An ethnocentric judgment can be treated as a cognitive phenomenon, but it has emotional component as well. In its cultural form it is usually called an ethnic stereotype. Such stereotypes were intensively studied in American social psychology before and after World War II, up to the mid- 1950s (Allport, 1954). Since then, there has been relatively little research on ethnic stereotypes, but the concepts of ethnocentrism and stereotypes are still widely used in the mass media, in programs of training in intercultural communication, and in public discussions of racial prejudice and ethnic relations. The teaching of anthropology, through publications as well as in the classroom, is often justified by its role in reducing the ethnocentric attitudes ofWestern populations.

See also: Ethnic Conflict, Geography of; Ethnic Conflicts; Ethnic Identity, Psychology of; Groups, Sociology of; Nationalism, Sociology of; Nationalism: General; Prejudice in Society: Psychological Perspectives; Xenophobia: Social Psychological Aspects.

Bibliography

Allport, G.W., 1954. The Nature of Prejudice. Addison-Wesley, Cambridge, MA.
Barfield, T. (Ed.), 1997. The Dictionary of Anthropology. Blackwell, Oxford, UK.
Barth, F., 1956. Ecologic relationships of groups in Swat, North Pakistan. American Anthropologist 58, 1079–1089.
Barth, F. (Ed.), 1969. Ethnic Groups and Boundaries: The Social Organization of Culture Difference. Little, Brown, Boston, MA.
Coser, L.A., 1956. The Functions of Social Conflict. Free Press, Glencoe, IL.
Goody, J.R., 1956. Social Organization of the LoWiili. Her Majesty's Stationery Office, London.
Leach, E.R., 1954. Political Systems of Highland Burma. Bell, London.
LeVine, R.A., Campbell, D.T., 1972. Ethnocentrism: Theories of Conflict, Ethnic Attitudes and Group Behavior. Wiley, New York.
Merton, R.K., Rossi, A., 1957. Contributions to the theory of reference group behavior. In: Merton, R.K. (Ed.), Social Theory and Social Structure. Free Press, Glencoe, IL.
Moerman, M., 1965. Ethnic identification in a complex civilization: Who are the Lue? American Anthropologist 67, 1215–1230.
Segall, M.H., Campbell, D.T., Herskovits, M.J., 1966. The Influence of Culture on Visual Perception. Bobbs-Merrill, Indianapolis, IN.
Sumner, W.G., 1906. Folkways. Ginn, New York.

Ethnocentrism and Prejudice: History of the Concepts

Boris Bizumic, ANU College of Medicine, Biology and Environment, The Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ethnocentrism and prejudice are two distinct, though closely related, concepts. Since the early twentieth century, social scientists have tended to see them as fundamental social scientific concepts. This article presents a brief history of how social scientists conceptualized ethnocentrism and prejudice, and discusses their study with a focus on four prominent explanations that guided empirical research: (1) evolutionary, (2) threat and conflict, (3) self-aggrandizement, and (4) socialization and normative.

As phenomena, both ethnocentrism and prejudice have been observed and discussed since Ancient times. They became, however, fundamental social scientific concepts in the early twentieth century. Social scientists have widely studied both concepts, but they often meant different things to different researchers. This article will, therefore, first focus on the brief history of the two concepts and their varied, and often confusing, usages across disciplines. This will be followed with a history of explanations of ethnocentrism and prejudice, with a focus on four broad categories of explanations, which have been applied to both phenomena and which have guided empirical research. These theoretical approaches are (1) evolutionary, (2) threat and conflict, (3) self-aggrandizement, and (4) socialization and normative.

A Brief Conceptual History

It is very difficult to define ethnocentrism and prejudice. Conceptual analysis has rarely been applied to the two concepts. Neither concept has become clearer over time, and there is no universal agreement among social scientists on what ethnocentrism and prejudice mean. As a result, any discussion of these phenomena and their explanations is hindered due to vague, problematic, and inconsistent conceptualizations.

Ethnocentrism

The term 'ethnocentrism' combines the original Greek word 'ethnos' with the word 'center,' and its literal translation is that a particular 'ethnos' (ethnic group, nation, or people) is in the center. The concept has existed since at least the 1870s, and it was probably Ludwig Gumplowicz (born 9 March 1838; died 19/20 August 1909) who coined the concept (Gumplowicz, 1879, 1881), and used it in his books and papers written in German and Polish. William G. Sumner (born 30 October 1840; died 12 April 1910) later adopted and popularized the concept (Sumner, 1906). Given that Sumner did not reference Gumplowicz's work on ethnocentrism, later researchers (e.g., Adorno et al., 1950; LeVine and Campbell, 1972) have incorrectly credited Sumner with inventing the concept in 1906 (for the early conceptual history, see Bizumic, 2014).

Gumplowicz saw ethnocentrism as a similar concept to geocentrism and anthropocentrism, but with a focus on one's own ethnic group or people. He viewed it as a belief that one's own ethnic group is more important, superior, and better than any other group. Sumner similarly saw ethnocentrism as a "view of things in which one's own group is the center of everything, and all others are scaled and rated with reference to it" (Sumner, 1906: p. 13). He also included feelings of ingroup devotion and cohesion as well as rejection of outgroups, group superiority, and exploitative intergroup relations in his explicit conceptualizations of ethnocentrism (Sumner, 1906, 1911). Sumner, however, also assumed that ethnocentrism would almost always translate into outgroup negativity, and that ingroup positivity and outgroup negativity are highly intercorrelated.

Later social scientists based their conceptualizations of ethnocentrism on Sumner's writings, and given the complexity and breadth of his discussions of ethnocentrism, they tended to see it in various ways. Certain researchers have focused on intragroup factors, others on intergroup factors, and still others on outgroup factors. Occasionally, and influenced by the work of Theodor W. Adorno and colleagues (Adorno et al., 1950), ethnocentrism has been seen primarily as hostility or prejudice against a variety of outgroups. The influential book on theories of ethnocentrism by Robert LeVine and David Campbell focused on Sumner's conceptualizations and identified 23 facets in ethnocentrism, which they conceived as potentially a single dimension consisting of ingroup positivity and outgroup negativity (LeVine and Campbell, 1972). Empirical research, however, suggested that ingroup positivity (i.e., ingroup love) and outgroup negativity (i.e., outgroup hate) are often unrelated phenomena and not necessarily mutually interdependent (Brewer, 1999). Indeed, more recent cross-cultural measurement of ethnocentrism has suggested that there are two kinds of ethnocentrism: (1) intragroup ethnocentrism, which includes a sense of strong group cohesion and devotion to one's own ethnic group, and (2) intergroup ethnocentrism, which includes preference for ethnic ingroups over outgroups, a belief in ethnic superiority, a desire to maintain ethnic purity, and approval of exploitation of ethnic outgroups if it is in the interest of one's own ethnic group (Bizumic and Duckitt, 2012). These two kinds of ethnocentrism express a strong belief in the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03153-6

importance of one's own ethnic group and cover most historical definitions of ethnocentrism. These two kinds of ethnocentrism were shown to be empirically and conceptually distinct from mere ingroup positivity and outgroup negativity or prejudice.

Prejudice

The concept of prejudice itself is a much older concept than that of ethnocentrism. It comes from the Latin word 'praejudicium,' originally meaning 'prior judgment,' but over time acquiring the notion of generally unfavorable predetermined judgments. Prejudice has been a legal term for centuries, but isolated writings by social scientists on prejudice appeared in the nineteenth century. The concept, however, became popular in the social sciences only in the twentieth century when prejudice became a significant social and, by extension, social scientific problem. In fact, until the 1920s, social scientists did not consider prejudice as an important concept for research, and generally justified its existence, at least when the targets of prejudice were non-Western ethnic groups (Duckitt, 2010). This changed due to numerous societal reasons, and social scientists started to produce considerable work on prejudice, which they began seeing as a generally unjustified negative attitude toward a group of people, including every individual member.

Social scientists have broadly conceptualized prejudice as a negative attitudinal construct (Allport, 1954; Banton, 1998), although they have often focused on different kinds of negative attitudes, which at times overlapped with ethnocentrism (e.g., Adorno et al., 1950). Prejudice has often been seen as a negative attitude, based on an inaccurate and rigid generalization, toward a particular group or its group members (e.g., Allport, 1954). Many social scientists have argued that all kinds of prejudice are inaccurate, rigid, and unjustifiable. Others, however, have focused on a broader conceptualization. Indeed, the most important aspect of prejudice, on which most social scientists appear to agree, is that it includes a negative affect toward an outgroup, and this is reflected in both conceptual and empirical work. For example, Walter Stephan and Cookie Stephan (2000) saw and measured prejudice primarily as negative affect toward outgroups, involving both emotional reactions (e.g., hatred, disdain) and evaluations (e.g., disliking, disapproval) as well as the absence of positive emotions and evaluations. Social scientists have historically also distinguished between specific prejudice (negative attitude toward a specific group) and generalized prejudice (negative attitude that is generalized across many outgroups), which they often saw to be an aspect of ethnocentrism (Adorno et al., 1950). As mentioned above, however, research has shown that ethnocentric attitudes are conceptually and empirically distinct from prejudice (Bizumic and Duckitt, 2012; Brewer, 1999).

Ethnocentrism, Prejudice, and Social Groups

Since the beginning of the social scientific research into ethnocentrism and prejudice, most work has focused on ethnocultural and racial groups – even though both were at times related to any kind of social group. For example, despite the existence of 'ethno' in ethnocentrism, the concept was at times used to refer to diverse kinds of groups, including gender groups. Nevertheless, most social scientists see ethnocentrism as a characteristic of large-scale ethnocultural groups – though at times they also apply it to multiethnic nation states (e.g., US ethnocentrism) or even whole continents (e.g., European ethnocentrism). By definition, the concept of prejudice can be applied to any kind of social group, and social scientists at times have used it to refer to negative attitudes toward other species. Social scientists have, however, mainly used the concept to refer to prejudice against ethnocultural and racial groups, even though at times it was applied to women, religious groups, homosexuals, and other vulnerable or minority groups. Historically, both concepts have been predominantly studied among ethnic majority group members, whereas ethnic minorities and other vulnerable groups were often seen as recipients and targets of prejudiced attitudes, and their levels of ethnocentrism and prejudice have been largely unexplored.

A History of Explanations of Ethnocentrism and Prejudice

Social scientists have attempted to explain ethnocentrism and prejudice since the beginning of the social sciences, but the popularity and usefulness of various explanations have waxed and waned over time as well as across disciplines and across cultures. The influence of different theoretical explanations during particular periods has not been caused by the production of better theories or more convincing data in favor of certain explanations over others, but by a variety of external societal and political factors, which oriented social scientists toward different questions (Duckitt, 2010). In general, historical explanations have focused on the role of: (1) evolution, (2) threat and conflict, (3) self-aggrandizement, and (4) socialization and norms in the development of ethnocentrism and prejudice. These broad kinds of explanations have guided most of the social scientific research into ethnocentrism and prejudice.

Evolutionary Explanations

Evolutionary explanations of ethnocentrism and outgroup hostility are as old as the social sciences. In fact, the social sciences developed in part due to the Darwinian revolution in the nineteenth century when social phenomena, as extensions of natural phenomena, became legitimate scientific topics. It is not surprising, therefore, that many early theorists have attempted to explain ethnocentrism using an evolutionary framework. There is a wide range of evolutionary explanations, from Darwin's early theory and social Darwinist explanations to the more recent sociobiological and evolutionary psychological explanations. Ethnocentrism was explicitly related by Gumplowicz (1881) to anthropocentrism (giving humans the central place in the world), which evolutionary theory heavily criticized.

A common theme of evolutionary explanations is that humans are biologically hardwired to engage in both ethnocentrism and prejudice, and that ethnocentrism is evolutionarily useful because it enhances both individual and group

survival. Ethnocentrism and prejudice are, therefore, conceived as natural- and biology-based phenomena. Charles Darwin (born 12 February 1809; died 19 April 1882) wrote that in animals, "sympathy is directed solely toward the members of the same community, and therefore toward known, and more or less beloved members, but not to all the individuals of the same species" (Darwin, 1875: p. 106). Accordingly, humans, like other animals, also tend to engage in preferences toward members of their own community over others. Darwin also thought that intergroup competition makes people more cooperative with other members of their own group and that cooperative groups are more likely to thrive than noncooperative groups. Herbert Spencer (born 27 April 1820; died 8 December 1903), a social Darwinist, proposed two moral codes that exist universally in intragroup and intergroup relations (Spencer, 1892). He argued that one moral code approved of hostility toward those outside the group, whereas another code postulated that one should be highly cooperative with those inside one's own group. Spencer assumed that these exist among both industrial and nonindustrial societies. Sumner (1906), also a social Darwinist, elaborated upon the survival value of ethnocentrism, and assumed that outgroup hostility is a natural and normal correlate of ethnocentrism. Arthur Keith (1948), furthermore, argued that group survival is possible only if ingroup cooperation and intergroup competition are balanced, and that neither too much cooperation nor too much competition is good for groups. Although diverse, early explanations saw ethnocentrism and outgroup hostility as biology based and having survival value. Evolutionary theorists, however, frequently went beyond available data to claim that certain social groups are biologically superior and others are biologically inferior, and that it is right for the superior groups to rule over the inferior groups.

Although popular at the beginning of the twentieth century, evolutionary explanations became less influential from the 1930s until the 1970s. Anthropologists, sociologists, and psychologists began to produce strong evidence that brought into question many of the central assumptions of crude social Darwinism, such as the belief that there are inherent biological differences between different ethnic and racial groups. Anthropologists, in particular those who followed Franz Boas's (born 9 July 1858; died 22 December 1942) approach, have also advanced ideas of cultural relativism, which argued that cultures can only be understood from within and that cross-cultural comparisons (especially those advocating cultural superiority) are not useful and valid approaches for understanding and explaining human social behavior. Boas (1940) successfully questioned many crude biological and evolutionary applications to human groups and any inherent biological differences between members of different ethnic and racial groups. He also showed that many differences between ethnic and racial groups are better explained with environmental rather than biological factors. Finally, unrestrained ethnocentrism and prejudice, based on a naive evolutionary theorizing of social Darwinists, showed its extremely dark side during the 1930s and 1940s among the Nazis and fascists. Following the Second World War, evolutionary explanations of social phenomena, such as ethnocentrism and prejudice, have almost completely retreated.

Nevertheless, evolutionary explanations regained certain popularity again in the 1970s, and since then have been increasing in influence within academic circles. For example, social scientists used sociobiological theory to explain ethnocentrism. This theory proposes the concept of 'inclusive fitness,' which assumes that ingroup preferences are useful because genes could be passed to next generations not only from parents to their offspring but also through kin. Pierre van den Berghe (1978), for example, argued that both human and nonhuman animals prefer their kin over nonkin due to evolutionary reasons, and that altruism and self-sacrifice aimed at nonkin are not useful biologically. He assumed that humans tend to extend their kinship to the whole ethnic group, which is a unique kind of large-scale social group due to the existence (or, at least, a belief in the existence) of biological relatedness. This extension, therefore, tends to promote people's ethnocentric sacrifice and preference on behalf of their own group.

The past several decades have seen more evolutionary work on ethnocentrism and prejudice, including a focus on the role of explicit biological influences (e.g., genes and hormones), investigating ingroup preferences in other primates, and using simulation methodologies to test evolutionary underpinnings of ethnocentrism. The deterministic and crude evolutionary explanations characteristic of early social Darwinism have been supplemented with more sophisticated explanations. These explanations, however, often exaggerate the inevitability of ethnocentrism (though not necessarily outgroup prejudice) and minimize the role of social, cultural, and political factors.

Threat and Conflict Explanations

Many explanations of ethnocentrism and prejudice have focused on different kinds of threats and conflicts. These have ranged from realistic (e.g., economic) threats between groups to symbolic threats (e.g., threats stemming from perceiving different value or belief systems) and intrapsychic threats (e.g., psychodynamic threats and conflicts). A common theme in these theories is that ethnocentrism and prejudice are defensive reactions of individuals or groups against threats and conflicts.

Realistic conflict theory assumes that ethnocentric intragroup and intergroup attitudes as well as outgroup hostility are caused by realistic threats to the group and its welfare. Early sociologists were already concerned with the role of these threats in intergroup relations. For example, building on early realistic conflict theories of Gumplowicz and others, Sumner (1906) argued that ethnocentrism and outgroup hostility are functional in defending the ingroup against outgroup threats and advancing ingroup interests while in conflict with other groups. The work on realistic conflict theory and realistic threats became prominent in sociology and social psychology in the 1950s and 1960s. The theory was then elaborated upon and summarized by LeVine and Campbell (1972). This theory focuses on realistic threats to a group as opposed to unrealistic (i.e., psychodynamic, intrapsychic) threats and postulates that ethnocentrism and outgroup hostility are useful in helping groups defend against threat. It proposes that groups in conflict and under threat are more likely to defend themselves if they are ethnocentric and prejudiced. If, however, a group in conflict is accepting and considerate of outgroups

(i.e., nonethnocentric and nonprejudiced), it may risk being exploited or even destroyed.

Psychodynamic theorists, influenced by Sigmund Freud (born 6 May 1856; died 23 September 1939), however, assumed that ethnocentrism and prejudice are caused by intrapsychic threats and conflicts. Accordingly, these theorists have invoked a variety of defense mechanisms, which are thought to help individuals protect against these threats. This approach became particularly popular during the 1930s and 1940s because at that time, prejudice was being increasingly seen as irrational and unjustifiable (Duckitt, 2010). The most widely used defense mechanism to explain both ethnocentrism and prejudice was displacement. Freud (1930/1961), for example, argued that group life and ingroup cohesion are possible because people are able to displace their natural aggression and hostility onto outgroups.

The most influential psychodynamic theory of ethnocentrism and prejudice is authoritarian personality theory (Adorno et al., 1950). This theory posits that overly punitive and strict parenting causes an authoritarian personality, which is prone to ethnocentrism and generalized prejudice. Due to parental punitiveness and dependence on them, authoritarian people experience conflicting feelings, such as fear, love, and hostility. To cope with these ambivalent feelings, they employ defense mechanisms, such as repression, externalization, and displacement, which result in idealization of one's own group and prejudice against ethnic and other minority groups. For example, these authors wrote "the prejudiced subject's ambivalence toward his parents, with a repression and externalization of the negative side of this ambivalence, may be a factor in determining his strongly polarized attitudes, such as his uncritical acceptance of the ingroup and violent rejection of the outgroup" (Adorno et al., 1950: p. 482). Many later studies have supported the claim that authoritarianism is related to ethnocentrism and prejudice, but failed to support the psychodynamic explanation.

In contrast to other psychodynamic explanations, terror management theory has become influential in the 1990s and the 2000s. This theory builds on the work by Ernest Becker, and argues that the intrapsychic threat of mortality drives ethnocentric and prejudiced attitudes (Solomon et al., 1991). According to the theory, people are threatened by the awareness of eventual death and their cultural value systems serve as buffers against this existential anxiety. Any threat to one's own cultural system, therefore, makes people affirm their own cultural system and reject the systems of other groups. Indeed, research appears to indicate that ethnocentrism and prejudice may increase when people are reminded of their own mortality.

Psychologists and sociologists have proposed a wide range of other threats and conflicts to account for ethnocentrism and prejudice. For example, Milton Rokeach (1960) developed a belief congruence theory in the 1950s and 1960s. The theory argues that perceiving different beliefs and values in outgroup members, which conflict with one's own beliefs and values, might be sufficient to create group preferences and discrimination. According to the theory, people prefer those who have similar beliefs and values to themselves, and it is not ethnicity that causes ethnic preferences, but the perceived differences on important specific issues between those categorized as ethnic ingroup and outgroup members. Ethnic categorization is, therefore, important only as a symbol for "complexes of beliefs which to one degree or another are seen to be similar to and different from our own" (Rokeach, 1960: p. 391).

A later attempt was made to integrate different kinds of threats into broader theoretical frameworks. For example, integrated threat theory (Stephan and Stephan, 2000) proposed that in addition to realistic threats (i.e., threats to the general group welfare) and symbolic threats (i.e., threats to the values and beliefs of group members), there are two other threats as causes of prejudice: intergroup anxiety (i.e., irritation and fear caused by interpersonal interactions with members from another group) and negative stereotype threats (i.e., a negative perception of another group). A large body of research has indeed shown that both prejudice and ingroup, ethnocentric preferences increase as a result of these different threats (Riek et al., 2006).

Threat and conflict have been shown to have potent influences on ethnocentrism and prejudice. A variety of methodologies and approaches have shown that both ethnocentrism and prejudice tend to increase as a result of various threats. Nevertheless, social scientists have discovered that a certain amount of ethnocentrism may exist even in the absence of realistic conflicts and threats (Tajfel and Turner, 1979). Accordingly, it is still not clear whether threat and conflict could fully explain, or just exacerbate, already existing levels of ethnocentrism and prejudice. There are also endless examples from history of strong and threatening ethnic groups fully assimilating weaker ethnic groups. Threat may, therefore, at times create less, and not more, ethnocentrism and prejudice.

Self-Aggrandizement Explanations

A number of social scientists have explained ethnocentrism and prejudice as phenomena that make people feel good about themselves because they can increase people's sense of self-importance and esteem. Another group of explanations has assumed that these phenomena could be used to justify inequalities between different kinds of groups. These explanations refer to various self-aggrandizing functions of ethnocentrism and prejudice.

The potential to use ethnocentrism and prejudice to increase a person's self-esteem can be seen in Gumplowicz's description of ethnocentrism as a self-congratulatory delusion that ethnic groups employ, as well as in the psychodynamic links of narcissism with intergroup relations and prejudice (Freud, 1930/1961). Accordingly, ethnocentrism and prejudice might be easily seen as constructs that can make group members think and feel good about themselves. This explanation became particularly popular in the 1970s with the development of social identity theory by Henri Tajfel and John Turner (1979), which proposed that the need for positive self-esteem underlies ethnocentrism. Social identity research showed that ingroup bias in the lab (which Tajfel and Turner saw as an analog concept to ethnocentrism in the outside world) may exist even in artificially arranged minimal group situations. These situations are characterized by the absence of interaction between group members, the lack of knowledge about who ingroup or outgroup members are, as well as the absence of individual self-interest to discriminate. The only thing that people know in such situations is that

they are members of one group and not another. Still, numerous studies have shown that in these situations, people discriminate in favor of their ingroup over outgroups.

Social identity theory argued that when people categorize themselves into ingroups and other people into outgroups, they experience their self at a different level. As a result of a need for positive self-esteem (also called positive group distinctiveness), they want to feel good about their groups and they want their groups to be superior to other groups on relevant dimensions, and this need for positive self-esteem drives ethnocentric preferences. The theory has had a profound influence on various approaches that use categorization processes into ingroups and outgroups to explain intergroup relations and prejudice. Nevertheless, much subsequent work within this theoretical tradition has resulted in a variety of approaches that have tended to focus less on the motivational role of self-esteem and more on universal categorization processes, which in interaction with normative and socialization processes may reduce or increase ethnocentrism and prejudice.

A related group of self-aggrandizing approaches to ethnocentrism and prejudice has origins in Marxist and elite theories. Ethnocentrism and prejudice not only are useful in making people feel good about themselves but also may be useful in helping one group (usually the dominant group within a single society) exploit and rule over other groups. This approach, accordingly, claims that ethnocentrism and prejudice can be used by groups in order to exploit and dominate other groups for the sake of their own individual or group self-aggrandizement.

This approach came into widespread acceptance during 1960s and 1970s when numerous social movements of the 1960s led to an increase in the popularity of explanations based on Marxist theory (Banton, 1998). These explanations have focused on large-scale exploitative tendencies of dominant groups, which tend to use ethnocentrism and prejudice in order to promote their own selfish interests without consideration for the well-being of other groups. For example, Oliver C. Cox (1948) argued that prejudice is helpful in the dominance and exploitation of outgroups because it makes these outgroups appear inferior and deserving of exploitation. Cox, therefore, saw prejudice as an instrument of capitalism that helps the dominant group exploit subordinate groups, and he argued that relations between people of different racial groups represent relations between different social classes. Another example of this approach could be seen in the work of Teun A. van Dijk (1993), who used elite theory to argue that elites, including political, corporate, and academic elites, construct ethnocentric and prejudiced attitudes in various ways, and impose them on the rest of the population – including both ethnic minority and majority group members. His research using discourse analysis indeed suggested that ethnocentric and prejudiced attitudes are to a large extent constantly enforced by the elites (van Dijk, 1993).

A related theoretical framework, social dominance theory (Sidanius and Pratto, 1999), has focused on causes and consequences of social dominance orientation, which refers to individual differences in acceptance of hierarchical intergroup relations. This individual orientation appears to predispose people to support a wide variety of ideologies and attitudes, including ethnocentrism and prejudice, which give a differential value to social groups. Research indeed suggests that people's individual differences in their motivation for dominance and self-enhancement, together with individual differences in the motivation for security, appear to be the two main driving forces behind individual differences in ethnocentrism and prejudice (Duckitt, 2001).

Socialization and Normative Explanations

Social scientists have very frequently used socialization and norms to explain both ethnocentrism and prejudice. These explanations postulate that ethnocentrism and prejudice develop and persist in humans due to socialization in or conformity with ethnocentric and prejudiced social norms. The process of socialization usually involves internalization of social norms and other social influences.

Socialization and normative explanations gained prominence in the second half of the twentieth century. They were partly a reaction to individual difference explanations, such as authoritarian personality theory (Adorno et al., 1950). In an influential critique of personality explanations of prejudice, Thomas Pettigrew (1958) argued that normative influences are more important for the development of prejudice than internal psychological factors, such as authoritarian personality. For example, he showed that although Whites in the United States and South Africa tended to have similar scores on authoritarianism, South African Whites were more likely to show negative attitudes toward Blacks. Similarly, people in the South and North of the United States tended to have similar scores on authoritarianism, but the Southerners were more prejudiced against Blacks. Pettigrew, therefore, argued that prejudiced norms guide people's expressions of prejudice, and that they are much more important than individual differences. Similarly, Ralph Minard's (1952) research in the United States showed that Black and White miners conformed to dominant nonethnocentric norms of interaction and cooperation while at work, and to dominant ethnocentric norms of separation while they were in their communities. Studies such as these showed that socialization and social norms appear to have powerful influences on ethnocentrism and prejudice, and many subsequent approaches to reducing these phenomena focused on changing social norms.

Socialization and normative approaches have also been employed to interact with basic cognitive processes, such as social categorization, to explain ethnocentrism and prejudice. Social categorization approaches have, in general, postulated that categorization into groups is a basic precondition for ethnocentrism and prejudice (Allport, 1954). Although influenced by Allport's (1954) work, these theories have become particularly influential explanations of ethnocentrism and prejudice following the development of social identity theory in the late 1970s (Tajfel and Turner, 1979). They tended to show that how people categorize their social world has important implications for the development of ethnocentrism and prejudice. People tend to learn about social categories to which they belong and do not belong, the values associated with these social categories, and preferences for the social categories to which they belong over others. In fact,

ethnocentrism and prejudice are unlikely to exist if people failed to identify with certain social categories and disidentify with others. Given the general and universal processes of categorization, social categorization researchers have attempted to investigate what social conditions would foster or reduce prejudice and ethnocentrism. For example, social categorization researchers have argued that providing people with cross-cutting group categories might enable them to engage in less prejudice and ethnocentrism (Brewer, 1999).

Another group of explanations that focuses on socialization and norms of two different kinds of prejudice became popular in the 1980s and continues to capture the interest of social scientists until the current day (Crandall and Eshleman, 2005). These explanations propose two different kinds of prejudice. They suggest that as a result of different norms and social learning, people in Western societies have been developing different kinds of prejudiced attitudes. For example, due to egalitarian, nonethnocentric, and antiprejudiced norms, which are characteristic of many Western societies, new kinds of prejudice have been developing, which are replacing the traditional, blatant forms of prejudice. Although social scientists have used various terms for these kinds of prejudice (e.g., aversive, modern, symbolic, implicit, and subtle kinds of prejudice), what they have in common is that they are more subtle and sophisticated, but still predispose people to intergroup discrimination. This research has also suggested that people often may not even be aware that they are ethnocentric or prejudiced, but still have ethnocentric and prejudiced attitudes and act on them.

Despite a wide variety of socialization and normative theories, they do appear to share the same underlying emphasis on the role of external factors in shaping ethnocentrism and prejudice, and are therefore top–down theories, even though they may often interact with the more fundamental social cognitive processes. These theories assume that both ethnocentrism and prejudice could be transcended when socialization and social norms propagate humanitarian, nonprejudiced, and tolerant attitudes. Whether this is fully possible is an important question because socialization into any cultural world view appears to involve ethnocentrism, given that all cultures, at least at an implicit level, assume that one's own cultural norms, rules, and ethical systems are right and correct, and by extension, somewhat better than others.

Summary and Concluding Remarks

Social scientists from various disciplines have studied ethnocentrism and prejudice, and defined them in diverse ways, with ethnocentrism being primarily seen as a strong sense of intragroup and intergroup importance of one's own group, and prejudice as a negative, often unjustifiably negative, attitude toward an outgroup. Although sometimes used interchangeably, ethnocentrism and prejudice have been historically perceived as distinct, and research does support their conceptual and empirical distinction. Both have at times been used to refer to many kinds of groups, but most historical usages refer to large-scale ethnocultural groups in ethnocentrism and large-scale ethnocultural and racial groups in prejudice – although in the last several decades, there has been an increasing amount of research into prejudice against women, homosexuals, and other vulnerable groups.

Social scientists have approached ethnocentrism and prejudice from many angles, and these explanations could be categorized as those invoking: (1) evolutionary factors, (2) threat and conflict, (3) self-aggrandizement, and (4) socialization and norms. There is a wide variety of explanations within each of the categories, and some transcend even these broad categorizations. Nevertheless, it appears that social scientists have consistently drawn on these four broad kinds of explanations while trying to account for both ethnocentrism and prejudice.

Ethnocentrism and prejudice have appeared useful to researchers from different backgrounds and interests, and these concepts have been shown to be very fruitful and engaging. Unfortunately, conceptual clarity has not been of primary interest to the social scientists who have studied ethnocentrism and prejudice, and they have often failed to agree on right conceptualizations. Consequently, any summary of these wide areas of research needs to grapple with many conceptual issues and vague operationalizations. Both concepts have also been shown to be very difficult to change and eliminate, and their persistence still puzzles social scientists, especially those who reject evolutionary explanations. Any fruitful integrated approach to ethnocentrism and prejudice must, therefore, be able to address the four different kinds of explanations, which all have substantial bodies of supporting research, and shows how they independently, or in interaction with each other, produce ethnocentrism and prejudice.

See also: Attitude: History of Concept; Psychology, History of (Twentieth Century); Sociology, History of.

Bibliography

Adorno, T.W., et al., 1950. The Authoritarian Personality. Harper, New York.
Allport, G.W., 1954. The Nature of Prejudice. Addison-Wesley, Reading, MA.
Banton, M., 1998. Racial Theories. Cambridge University Press, Cambridge.
Bizumic, B., 2014. Who coined the concept of ethnocentrism? A brief report. Journal of Social and Political Psychology 2, 3–10.
Bizumic, B., Duckitt, J., 2012. What is and is not ethnocentrism? A conceptual analysis and political implications. Political Psychology 33, 887–909.
Boas, F., 1940. Race, Language and Culture. Macmillan, New York.
Brewer, M.B., 1999. The psychology of prejudice: Ingroup love and outgroup hate? Journal of Social Issues 55, 429–444.
Cox, O.C., 1948. Caste, Class, and Race: A Study in Social Dynamics. Doubleday, New York.
Crandall, C.S., Eshleman, A., 2005. The justification-suppression model of prejudice: an approach to the history of prejudice research. In: Crandall, C.S., Eshleman, A. (Eds.), Social Psychology of Prejudice: Historical and Contemporary Issues. Lewinian Press, Lawrence, KS, pp. 237–267.
Darwin, C., 1875. The Descent of Man: And Selection in Relation to Sex, second ed. D. Appleton and Company, New York.
Duckitt, J., 2001. A dual process cognitive-motivational theory of ideology and prejudice. In: Zanna, M.P. (Ed.), Advances in Experimental Social Psychology. Academic Press, San Diego, CA, pp. 41–113.
Duckitt, J.H., 2010. Historical overview. In: Dovidio, J.F., et al. (Eds.), The SAGE Handbook of Prejudice, Stereotyping and Discrimination. Sage, London, pp. 29–44.
Freud, S., 1961. Civilization and its discontents. In: Starchey, J. (Ed.), The Standard Edition of the Complete Psychological Works of Sigmund Freud. Hogarth Press, London, pp. 57–145 (Original work published 1930).

Gumplowicz, L., 1879. Das Recht der Nationalität und Sprachen in Oesterreich-Ungarn (The Right of Nationality and Languages in Austria-Hungary). Wagner'schen Universitäts-Buchhandlung, Innsbruck.
Gumplowicz, L., 1881. Rechtsstaat und Socialismus (Legal State and Socialism). Wagner'schen Universitäts-Buchhandlung, Innsbruck.
Keith, A., 1948. A New Theory of Human Evolution. Watts & Co, London, UK.
LeVine, R.A., Campbell, D.T., 1972. Ethnocentrism: Theories of Conflict, Ethnic Attitudes, and Group Behavior. John Wiley & Sons, New York.
Minard, R.D., 1952. Race relationships in the pocahontas coal field. Journal of Social Issues 8, 29–44.
Pettigrew, T.F., 1958. Personality and sociocultural factors in intergroup attitudes: a cross-national comparison. Journal of Conflict Resolution 2, 29–42.
Riek, B.M., Mania, E.W., Gaertner, S.L., 2006. Intergroup threat and outgroup attitudes: a meta-analytic review. Personality and Social Psychology Review 10, 336–353.
Rokeach, M., 1960. The Open and Closed Mind: Investigations into the Nature of Belief Systems and Personality Systems. Basic Books, New York.
Sidanius, J., Pratto, F., 1999. Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression. Cambridge University Press, New York.
Solomon, S., Greenberg, J., Pyszczynski, T., 1991. A terror management theory of social behavior: the psychological functions of self-esteem and cultural worldviews. In: Advances in Experimental Social Psychology. Academic Press, San Diego, CA, pp. 93–159.
Spencer, H., 1892. The Principles of Ethics. Williams and Norgate, London.
Stephan, W.G., Stephan, C.W., 2000. An integrated threat theory of prejudice. In: Oskamp, S. (Ed.), Reducing Prejudice and Discrimination. Lawrence Erlbaum, Hillsdale, NJ, pp. 23–46.
Sumner, W.G., 1906. Folkways: A Study of the Sociological Importance of Usages, Manners, Customs, Mores, and Morals. Ginn and Company, Boston, MA.
Sumner, W.G., 1911. War and Other Essays. Yale University Press, Freeport.
Tajfel, H., Turner, J.C., 1979. An integrative theory of intergroup conflict. In: Worchel, S., Austin, W.G. (Eds.), The Social Psychology of Intergroup Relations. Brooks-Cole, Monterey, CA, pp. 33–47.
van den Berghe, P.L., 1978. Race and ethnicity: a sociobiological perspective. Ethnic and Racial Studies 1, 401–411.
van Dijk, T.A., 1993. Elite Discourse and Racism. Sage, Newbury Park, CA.

Relevant Websites

http://www.tolerance.org – The Authors of the Website Describe It as, "A Place for Educators to Find Thought-provoking News, Conversation and Support for Those Who Care about Diversity, Equal Opportunity and Respect for Differences in Schools."

https://www.archive.org – Includes Many Early Historical Books into Ethnocentrism and Prejudice, Such as the Work by Herbert Spencer, Ludwig Gumplowicz, William G. Sumner, and Arthur Keith.

http://ethics.sandiego.edu/Applied/race/ – Numerous Internet Resources Related to Race, Ethnicity and Multiculturalism.

https://implicit.harvard.edu – On-line Tests that Investigate Ethnic, Racial, and Other Group Preferences and Prejudice at the Unconscious Level.

http://www.understandingprejudice.org – Various Resources on Prejudice, Intolerance, Ethnocentrism.

Ethnogenesis in Anthropology

Michiyuki Matsuda, Kyoto University, Kyoto, Japan

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 7, pp. 4854–4857, © 2001, Elsevier Ltd.

Abstract

Ethnogenesis is a process by which a distinct ethnic group emerges with a common sociocultural consciousness of belonging. The concept, coming on the anthropological studies in the 1960s, has been regenerated during the 1990s with new implications. It is used to focus on the interaction between global and local histories, especially on local initiatives to deal with the impact of globalization. Ethnogenesis therefore refers to cultural and political struggles by indigenous and enslaved populations to create enduring identities at times of often violently imposed radical change resulting from European domination. In the contemporary world system, they have begun to assert their sociocultural and politico-economic rights by reclaiming 'natural' sentiment of ethnicity. Ethnogenesis is thus becoming a key concept to diagnose and understand society in the twenty-first century.

Ethnogenesis is a process by which a distinct ethnic group emerges with a common sociocultural consciousness of belonging. In the 1960s Singer (1962) analyzed the process among African-Americans and Sturtevant (1971) subsequently introduced this term into the anthropological study of native Americans. Since then it has been widely utilized for examining the consciousness of belonging among various populations.

Although Sturtevant did no more than point out how ethnogenesis could be used to analyze the historical establishment of group distinctiveness, the concept has been regenerated during the 1990s with new implications. It is used to focus on the interaction between global and local histories, especially on local initiatives to deal with the impact of globalization. Ethnogenesis therefore refers to cultural and political struggles by indigenous and enslaved populations to create enduring identities at times of often violently imposed radical change resulting from European domination – such as demographic collapse, forced relocation, or enslavement. The idea of ethnogenesis refutes the established view of societies in the Third World as isolated, static, and ahistorical. Instead it pays attention to their creativity and ways of making history.

Theoretical Background

Currently the focus on issues of ethnicity has shifted from study of particular ethnic groups to looking at the dynamic process of ethnogenesis. This was made possible by the continued development of new theoretical understandings of ethnicity. Following Barth's contribution (1969), the subjective and interactive aspects of ethnicity began to gain attention. Despite the pre-eminence accorded the objective and primordial aspects of ethnicity by earlier perspectives, Barth did not regard ethnic identity as constituting an *a priori* reality, but as being an interactive process of self-definition and classification by others.

Actors categorize themselves and others in relation to how they interact with one another. It is when they use an ethnic identity to categorize themselves that Barth sees an ethnic group emerging. He rejected the prevailing idea that geographical and cultural isolation or separation is essential to the preservation of ethnic groups and the maintenance of cultural diversity. Instead he insisted that ethnic differentiation forms an overall social system; what is to be studied, consequently, is not the internal relations of the individual ethnic group, but the way the different ethnic groups establish boundaries. This perspective led to the view that ethnicity involved a formative process.

Another shift in the 1980s was the emergence of new perspectives that sought to deconstruct ethnicity as an essential entity. Deconstructionism in the social sciences exercised its influence to show how seemingly essential categories such as sex, generation, and race are far from being natural and *a priori*, but are an invention of modern society. Hobsbawm and Ranger showed that what had been taken to be traditional elements in symbols, customs, myths, and identity were actually invented as part of a dynamic process in the course of modern nation-building. According to this view, ethnic identity is also a modern invention that has been shaped in a political and historical way within the frame of the nation-state. This deconstructionist perspective, together with a subjective active view of ethnicity, provided a basis for the social sciences to theorize the phenomenon of ethnogenesis.

Ethnogenesis and Ethnocide

The concept of ethnogenesis has developed from studies of societies in the New World. Here the indigenous population was conquered and oppressed in the centuries after 1492. In addition, an enormous number of people with different cultural-linguistic backgrounds were transported from Africa and set to work as slaves. This is a process of ethnocide.

Ethnocide means the extermination of a culture. It involves the systematic destruction of cultural identities, ecosystems, political authorities, and religious rituals and belief systems. Common forms of ethnocide included the forced conversion to the religion of the colonizers, the banning of the use of native languages and cultural practices, and the forced relocation from the homeland to reservations.

Although it differs from genocide, which means the physical destruction of a whole people, the distinction is not always clear, and the two phenomena may occur together. Following the arrival of Columbus in the New World in 1492, the Arawak and Carib peoples he encountered, whose population is estimated to have numbered around 10 million, were almost entirely wiped out within 100 years. Virtually no indigenous people now remain in the Caribbean, although some were able to flee to the American continent where they merged with other survivors to generate new cultures.

In the century after the discovery, the indigenous population of South America, estimated to have been around 40 million, declined to about 440 000. In Australia in the nineteenth century the situation was much the same, although there were not as many people to be killed. The people of Tasmania, with an estimated population of 23 000 in 1803, were exterminated before the end of the century. Other people, on the verge of cultural extinction, have attempted to establish distinct cultural identities in order to reorganize their societies and secure their existence. This is a process of ethnogenesis. There is therefore a direct relationship between ethnocide and ethnogenesis. Where ethnocide takes place, one will usually also find ethnogenesis.

In the present day, peoples who were enslaved or marginalized under colonial domination are able to revive their cultural identities and reclaim their right to existence. This has gained momentum with the setting-up of statutory and voluntary organisations in the international context.

Mechanisms of Ethnogenesis

Minority populations which were marginalized in the world politico-economic system needed a shared consciousness of belonging. This required definitions of group boundaries, and processes of cultural creation, revitalization, and appropriation.

The creation of a religious ideology and practice was one means of integrating different groups into one unit. The enslaved African populations of the Caribbean, for instance, had originally come from a variety of sociocultural backgrounds with different languages, religions, value systems, and social structures. Despite these differences, they invented rituals, oaths, and a common set of religious values. This bricolage of different African elements became marked with the emergence of a Maroon identity in Jamaica in the middle of the seventeenth century.

In 1655, Britain took Jamaica from Spain by force. African slaves took advantage of the resulting turmoil to escape from the plantations into the forests and mountains. These deserters comprised a medley of ethnic origins, from Senegambia and the Gold Coast to the Congo and Angola, and even to the east coast of Africa. They were able to establish their own political authority and military alliances in the seventeenth and eighteenth centuries. Such polities created and mobilized common religious ideologies and practices based on rituals of spirit possession, which produced a shared consciousness and generated a common identity.

Ethnic identity may also be developed through a revitalization of precolonial symbols and ritual practices, as was the case with the Seminoles of Northern Florida. In the eighteenth century, when British, French, and Spanish forces were competing in the region, a distinct Seminole ethnicity emerged from the variety of Muskogean and other peoples with different cultural-linguistic backgrounds, who had colonized this region. They established new chiefdoms which closely followed the traditional Muskogee patterns. Among Muskogean peoples, who had previously been only loosely related with each other, there was the powerful division of Kawita, which overpowered the others and standardized and transformed a variety of politico-religious cultures under its hegemony. The early Seminole settlements still acknowledged Kawita's suzerainty. This meant giving a central role to rituals of the mythical creation of sacred fire, originally controlled by Kawita's traditional chiefs whose authority was thereby revitalized in the newly created Seminole societies.

Ethnogenesis may also take place through appropriation. From the late fifteenth century, indigenous peoples in the Third World were in most cases conquered by overwhelmingly superior European powers and forced to accept the religion, language, and way of living of the colonizers. In doing so, however, they often created a new ethnic identity by the way they transformed and adapted the ritual symbols of the Europeans. For instance, the Kapon and Makushi of Venezuela and Guyana emerged when the Christian saints and deities brought by the European colonizers were reinterpreted according to the traditional cosmology of indigenous peoples who used these ritual symbols to forge new cultural identities.

Ethnogenesis in the Twentieth Century

Although the process of ethnogenesis can be clearly seen in small-scale societies in the Third World following colonization, it has also developed in the process of modern nation-building. The ideology of the nation-state expressly aims to transform different ethnies into one nation with a single national culture. The principle is often different from the practice. The ruling group, in most cases the majority group but sometimes a numerical minority having dominant power, has generally monopolized the political-economic interests and imposed its culture on the others in the name of the national culture. In response, marginalized peoples have attempted to generate a new ethnic identity in order to claim their sociocultural and politico-economic rights within the frame of the nation-state. Modern ethnogenesis may be classified in three main ways.

The first type is integration. Those peoples who originally had formed separate and independent polities without any consciousness of belonging together were given the same name and treated as one group by European explorers, missionaries, anthropologists, and administrators. They accepted this name and formed a new identity in accordance with the imposed boundary, in order to resist or negotiate with the colonial regime, especially in modern nationalist movements that sought political independence. In the process, a new ethnicity emerged. This type of ethnogenesis can be seen among the Aborigines in Australia, the Inuit in the Arctic regions and the San in the Kalahari, where the central government has adopted policies of multiculturalism to secure minority group

rights – the government needs fixed sociocultural units in order to assist political autonomy and cultural preservation.

The second type is amalgamation, which occurred frequently in African societies at the time of independence. In Kenya after World War II, for example, an active African nationalism was organized by two major ethnic groups, the Kikuyu and the Luo. Political elites from smaller ethnic groups united under the hegemony of these two groups and gradually formed new cultural identities. Thus the Luyia emerged in the 1940s to bring together 17 Bantu-speaking peoples, among them the Maragoli, Bukusu, and the Wanga, who lived in what was then the North Kavirondo district of western Kenya. The word Luyia, meaning the communal space in which people sit around the fire, was appropriated for the the idea of mobilizing a consciousness of brotherhood among those from western Kenya who studied at Makerere University in Kampala.

The same sort of amalgamation took place among the 10 or so Eastern Cushitic-speaking ethnic groups in the Rift Valley. This ethnogenesis was promoted by those who studied at the prestigious Alliance high school in Nairobi in the 1940s. They invented the name of a new ethnic group, the Kalenjin – literal meaning 'I'm telling you' (in the plural sense) – and it started to spread among influential members of the home communities in the 1950s. Aidan Southall has termed this type of ethnogenesis 'supertribalization.'

The third type is separation. During the period of colonial domination, some small indigenous ethnic groups were subsumed into bigger neighboring groups or were put together in one large unit, either voluntarily or forcibly. Later they have attempted to revive their former cultural identity and secede from the constituted group, to create an independent ethnic group for themselves. The Suba, mainly living in the small islands of Lake Victoria, had been Luo-ized for a century, and their language was beginning to become extinct. But after the multiparty system was reintroduced into the Kenyan political scene in 1991, which caused the nation-state to become more ethnicized, the Suba political elite began to differentiate Suba from Luo and promoted the Suba culture and language. In so doing they are again forming themselves into a distinct ethnic group.

In the contemporary world system indigenous peoples and ethnic minorities who were once colonized and marginalized have begun to assert their socio-cultural and politico-economic rights. Because refashioned ethnic identities play a large part in such assertions, ethnogenesis is becoming a key concept in understanding society in the twenty-first century.

Bibliography

Barth, F. (Ed.), 1969. Ethnic Groups and Boundaries. Universitetsforlaget, Oslo, Norway.

Henry, M., 1962. Studies in Siberian Ethnogenesis. University of Toronto Press, Canada.

Hill, D. (Ed.), 1996. History, Power, and Identity: Ethnogenesis in the Americas. University of Iowa Press, Iowa City, IA.

Hobsbawm, E., Ranger, T.O. (Eds.), 1992. The Invention of Tradition. Cambridge University Press.

Matsuda, M., 1998. Urbanisation from Below. Kyoto University Press, Kyoto, Japan.

Roosens, E., 1989. Creating Ethnicity: The Process of Ethnogenesis. Sage.

Singer, L., 1962. Ethnogenesis and Negro-Americans today. Social Research 29, 419–432.

Smith, A., 1986. Ethic Origins of Nations. Blackwell, Oxford, UK.

Sollors, W., 1989. The Invention of Ethnicity. Oxford University Press.

Southall, A., 1970. The illusion of tribe. In: Gutkind, P.C.W. (Ed.), The Passing of Tribal Man in Africa. Brill, Leiden.

Sturtevant, W., 1971. Creek into Seminole. In: Leacock, E., Lurie, N. (Eds.), North American Indians in Historical Perspective.

Ethnography

Antonius CGM Robben, Utrecht University, Utrecht, The Netherlands
Jeffrey A Sluka, Massey University, Palmerston North, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ethnography is the investigation and description of cultures and societies through fieldwork. Ethnographers conduct empirical research by immersing in local settings, establishing good rapport, and using qualitative methods such as participant observation and open interviews. This intersubjective construction of ethnographic knowledge requires close attention to forms of textual representation. The twentieth-century standard of long-term, face-to-face fieldwork has in the twenty-first century been broadened by multisited fieldwork and computer-mediated research.

The term ethnography refers to the first-hand study of people, cultures, and subjects in local settings, and to their description and analysis in written texts. Ethnography is as much the practice of investigation as the reporting of empirical findings. The investigator is on the one hand the research instrument that accesses reality through personal observations, experiences, and interactions, and on the other hand the author of ethnographic knowledge whose texts interpret the study object through different forms of argumentation and writing. The history of ethnography reflects its dual meaning as research and representation. Most key shifts and innovations analyzed here were pioneered by sociocultural anthropology, which continues to embrace ethnography with its experience-near data collection as the heart of the discipline. Sociology has made several important contributions, notably in urban studies, but has developed a growing preference for quantitative methods.

The historical antecedents of ethnography date back to classical antiquity in the writings of Herodotus and Tacitus about the Egyptians, Scythians, and Germanic tribes. They were travelers, like their medieval successors Marco Polo and the Muslim contemporary Ibn Battuta, who spent short periods at multiple locations and gathered knowledge through personal observations and local conversations (Launay, 2010). They developed a distinct genre of travel writing that reflected a sense of wonder at humankind's sociocultural variation, and added creative imaginations to contrast supposedly 'superior' civilizations with 'inferior' and 'primitive' cultures.

The rise of science and humanism during the Renaissance encouraged a more systematic study of non-Western cultures, an approach that had been predated in the fourteenth century by the Muslim historian Ibn Khaldun. The comparison of peoples and cultures called for scholarly reflection on universal classifications and proper research methods. The manual about data collection by French philosopher Joseph-Marie Degérando, written in 1800, was formative for ethnography. It served as the basis for the 1840 manual of the French Ethnological Society and the *Notes and Queries on Anthropology* of 1841 whose multiple editions were required reading for generations of British anthropologists (Urry, 1972).

Mid-nineteenth-century ethnographic studies by social scientists such as Morgan, Spencer, Tyler, and Frazer emphasized the classification of data according to evolutionary principles. For example, Lewis Henry Morgan (1985) discerned in his 1877 book *Ancient Society* stages of cultural evolution by comparing human societies in terms of subsistence, government, language, family types, religion, property, house life, and architecture. The sociocultural evolutionists were guided by a systematic data collection that distinguished their scientific reporting from contemporary travel accounts. Detailed ethnographic descriptions were presented according to general classes, categories, types, and taxonomies that facilitated cross-cultural comparison.

By the late-nineteenth century, ethnography underwent a major change when the quality of ethnographic data, the comparative method, and cultural evolution theory became questioned. Alfred Haddon's expedition to the Torres Straits islands presaged this change. First-hand empirical research was conducted by a multidisciplinary team of scientists that surveyed the archipelago as one cultural area, rejected speculative cross-cultural comparisons, and collected local classifications in the native language, although hardly observing their manifestation in people's everyday practices. Haddon introduced the term fieldwork into anthropology, and promulgated the merging of the armchair anthropologist with the field researcher into the ethnographer (Stocking, 1996). Franz Boas and Bronislaw Malinowski came to personify this ideal at the beginning of the twentieth century through their epistemological critiques of, respectively, anthropology's evolutionism and ethnographic methodology.

Founding Fathers of Ethnography: Boas and Malinowski

The German-American Franz Boas revolutionized ethnography in North America by emphasizing in-depth data collection and inductive analysis, while rejecting cultural evolution theory and the comparative method. He criticized the classifications of evolutionists as arbitrary, Eurocentric, and deduced from unsubstantiated laws and principles. Cultural phenomena that appeared similar were categorized erroneously into one class, even though their functions and meanings differed. The display of material culture in ethnographic museums demonstrated this misguided approach. Nineteenth-century museums decontextualized artifacts, such as bows and arrows, musical instruments, baskets, and clothing, by grouping them in separate display cases. Exhibitions were organized along the alleged

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12065-3

evolutionary development of material culture. Instead, Boas argued that different artifacts of one society should be grouped together as expressions of one cultural whole.

The understanding of cultures as integrated wholes, instead of manifestations of evolutionary stages, radically changed ethnographic practice and description. Data collection was no longer guided by preestablished classifications but by a holistic approach sensitive to the meanings, patterns, and internal consistency of sociocultural phenomena and traits within one culture. Still, Boas' writings demonstrate the vestiges of earlier styles of ethnography that clustered data in terms of analytical categories such as material culture, mythology, social organization, the arts, and rituals. Boas did not possess a thorough knowledge of local languages, and worked in partnership with a key informant who was taught how to present and translate local knowledge. These narratives were then reworked into native texts that erased their dialogical production and generated authoritative accounts (Briggs and Bauman, 1999). The lasting importance of Franz Boas to ethnography consists of his emphasis on the detailed description of cultures as integrated wholes, and his fundamental critique of the comparative method.

The Polish-British Bronislaw Malinowski (1984) carried ethnography forward by combining Boas' concern for rigorous data collection with experience-rich descriptions, and adding the importance of language fluency and the study of worldviews. Malinowski became the great synthesizer of ethnography, both as field-worker and author, and defined ethnography in terms of three objectives that required distinct methods and types of textual notation. One, the 'cultural anatomy' or 'firm skeleton of tribal life' must be delineated in a holistic manner as patterns and regularities. People act according to patterns unknown to them, so the field-worker must derive them from their behavior through observation and inductive inference. The field-worker records these regularities through systematic, quantitative documentation. Two, the 'flesh and blood' or 'imponderabilia of actual life' must be described by observing people's ordinary behavior and participating in their activities. The ethnographer's attention is not on the regularities of social life but on their individual enactment – including the field-worker's – and the subtle deviations from the patterns. These daily observations are recorded in an ethnographic diary. Finally, ethnography aims to study people's 'spirit,' 'subjective states,' or 'typical ways of thinking and feeling.' The ethnographer is not interested in people's psychology but in their verbalization of worldviews, opinions, and ideas as members of society. Speech is written down verbatim in the local tongue to constitute a corpus of native texts, very much like Boas had proposed.

Malinowski interpreted fieldwork as a dialectic of empathy and detachment, and turned participant observation into ethnography's key research method. Detachment enabled the delineation of rules and regularities embedded in thought and behavior through detailed observation. Empathy allowed the ethnographer to assume an insider's perspective through active participation in the local setting. The ethnographer's ultimate goal was 'to grasp the native's point of view' by drawing on the systematic documentation, the diary, and the native texts. Although other anthropologists had written ethnographies based on fieldwork, Malinowski most fully developed the research strategy under the influence of his intellectual mentor Rivers and his method of intensive study.

The author Malinowski (1984) made an effort to establish his ethnographic authority in his 1922 *Argonauts of the Western Pacific*. Occasionally, writing in the first pronoun and in the present tense, he described his lengthy residence among the Trobriand islanders, his isolation from 'white men' (Europeans), and his participation in ceremonies and rituals. Unlike Boas' neutral texts composed along distinct categories such as material culture and rituals, Malinowski's three ethnographies fuse his personal experiences with those of the Trobrianders to evoke the life cycles of ritual interisland exchanges, human reproduction, and agriculture. His studies combine a scientific stance, an experiential approach, and a narrative style that evokes lived experience (Payne, 1981).

Malinowski set the standard for ethnographic fieldwork in sociocultural anthropology with research in a local setting, conducted in the local language, data collection through participant observation, and inductive rather than deductive analysis. Similar developments were occurring in the United States, not only among the students of Boas but also especially among sociologists at the University of Chicago.

The Chicago School of Urban Sociology

During the 1920s and the 1930s, sociologists at the University of Chicago, influenced primarily by the work of American social psychologist George Herbert Mead, developed new forms of mostly urban-based ethnographic research, which became known as the 'Chicago School.' Led by Robert Park, Earnest Burgess, Herbert Blumer, William Thomas, Albion Small, Charles Horton Cooley, Florian Znaniecki, and Louis Wirth, the Chicago School represented the first major attempt to conduct systematic ethnographic fieldwork – particularly community studies – in urban environments beginning with Chicago but then spreading to other cities and to small town life and rural settings such as Jim Crow segregation in the south. Credited with institutionalizing, if not establishing, sociology as a modern science, Chicago School sociologists produced hundreds of monographs about topics such as Jewish ghettos and other ethnic quarters, taxi-dance halls, professional thieves, hobos, gangs, juvenile delinquents, labor strikers, and salesladies. They encouraged their students to do fieldwork, and reconceptualized their methodology around participant observation, which they referred to as a 'hands on' or 'getting one's hands dirty' method and which they opposed to 'armchair philosophy' based on library research.

Their research combined emerging theoretical approaches – mainly symbolic interactionism – with qualitative and quantitative methods of field research. Defining the city as a 'sociological laboratory,' many researchers lived in the community they studied for an extended period of time. They analyzed the everyday life and symbolic interactions characteristic of specific groups and communities, focused on social structures and environmental factors rather than genetic or individual personality characteristics, believed that the 'natural environment' a community inhabits is the main factor in shaping human behavior, and viewed the city as a microcosm: "In these great cities, where all the passions, all the energies of mankind are

released, we are in a position to investigate the process of civilization, as it were, under a microscope" (Park, 1928: p. 890).

The Chicago School ethnographers employed an 'ecological' approach (now called social ecology) to their studies, producing spot maps of cities identifying the location of specific behaviors such as alcoholism, homicide, suicide, psychoses, and poverty based on official data including census reports, housing and welfare records, and crime statistics, relating these data spatially to different neighborhoods. They employed multiple methods of triangulation and developed techniques of field research including interviewing, diaries, self-reported life histories, and case analyses to provide subjective balance to this quantitative and spatial analysis. They also identified and distinguished between 'patterns' (structures) and 'processes' (actions or behavior) as the constituting forces in society. They were eager to develop tools to not only research and understand society but also change or improve it, particularly by directing urban planning and social intervention agencies.

Chicago School ethnographers eschewed complex or abstract theoretical language and their approach was systematically influenced by Mead's ideas concerning the self, the other, social interaction, and the human capacity to be rational and take the role of the other. As Deegan observes, one of the fundamental theoretical foundations of Chicago School ethnography was Mead's understanding that "the scientific model of observation, data collection and interpretation is fundamentally a human project. Sociologists can learn to take the role of others because this is how all humans learn to become part of society" (2008: p. 19).

The New Ethnography

In the mid-1950s, an American school of thought emerged that called itself 'ethnoscience' or 'the New Ethnography.' Falling within the Boasian interest in cultural knowledge, the New Ethnography tried to develop a more rigorous ethnographic method than the rather intuitive, experiential approach of the Malinowskian tradition. Replication and reliability were the tests of validity. If ethnographers could reproduce the findings of previous field-workers, and could conduct themselves in culturally acceptable ways by using the ethnographic knowledge, then the research was considered valid.

Drawing on linguistic techniques such as componential analysis, ethnoscientists set out to isolate cultural classifications through structured interviews. The relations among contrasting conceptual categories were used to arrive at the cognitive model that determined people's thoughts and actions. The ethnographer should determine the boundaries of a cultural domain, such as a kinship system or a color system, and describe its internal structure or folk classification. A useful method to determine cultural classifications is through contrast sets. Informants are asked to contrast one kin term or color term with another. The answers are organized into taxonomies of mutually exclusive components that share only one culturally relevant feature, such as 'aunt' and 'uncle,' but not 'canoe' and 'uncle' (Sturtevant, 1964).

This research agenda posed the challenge of how to describe cognitive models. Recourse was taken again to linguistics, and in particular to Kenneth Pike's adaptation of vernacular and systematic descriptions of language to cultural phenomena. Pike coined the terms 'emic' and 'etic' in reference to the difference between phonemics and phonetics. The emic approach reproduces local terminologies, and constructs the cultural grammar underlying social behavior. The etic approach consists of the culture-free anthropological lexicon that allows the translation of local terms into generalizable knowledge. For example, 'aunt' is an emic term meaningful only in Anglo-American kinship terminology, while 'mother's sister' and 'father's sister' are the corresponding etic kinship terms. The emic approach should not be confused with Malinowski's native's point of view, which refers to people's experience of the world, not their cognition. The New Ethnography regarded the emic approach as an ethnographic method to map the tacit cultural logics, cognitive principles, and folk taxonomies that organize people's emotions, thoughts, and behavior (D'Andrade, 1995).

Contrary to its name, the New Ethnography hardly produced ethnographies but mostly articles on kinship and legal terminologies, color categories, and plant and animal taxonomies. A highlight of ethnoscience is the study *Basic Color Terms* by Berlin and Kay (1969). The authors compare the color lexicons of 98 languages and conclude that some societies distinguish as few as two colors but that there are 11 basic color categories: white, black, red, green, yellow, blue, brown, purple, pink, orange, and gray. Furthermore, all languages have at least black and white, and may then add additional colors along either one of two fixed temporal orders. Finally, the combined temporal order is evolutionary: color vocabularies with few color terms correlate with simple societies and simple technologies. Such decontextualized, evolutionary comparison seemed to defy the Boasian critique but demonstrated how decades of thorough *in situ* fieldwork had raised the quality of ethnographic data.

Criticism of the New Ethnography was raised with regard to its privileging of language over practice, the reduction of culture to cognition, the emphasis on the human mind, the neglect of intracultural variation, and the lack of attention to meaning. Ethnographic attention shifted during the 1970s from an emphasis on cultural classifications to symbolic meanings and local interpretations.

Thick Description and Reflexive Ethnography

The semiotic conceptualization of culture expounded by symbolic and interpretive anthropology revolutionized ethnography during the 1970s. Following the intellectual footsteps of German sociologist Max Weber, American anthropologist Clifford Geertz (1973) conceptualized culture as a web of meanings spun by the people themselves. Ethnography moved from its empiricist tendency toward hermeneutic interpretation. Fieldwork became like reading an old, partially faded manuscript with all the accompanying misreadings and mistranslations. Drawing on French philosopher Paul Ricoeur (1971), Geertz resembled cultures to documents as systems of meaning embodied in symbolic forms because of four similarities between cultural practices and texts: they are both meaningful; have observable objectifying consequences; supersede the original conditions,

relevance, and authorship of their production; and address different audiences.

Geertz defined ethnography as thick description. Thick description tried to unravel the conceptual structures and frames of local interpretation by studying people's conflicting social discourses in order to arrive at cultural systems of meaning. These systems are not emic analyses or descriptions of the actor's point of view but anthropological constructions of the actor's interpretations of his or her actions, emotions, beliefs, and experiences. The test of validity is ultimately the ethnography's imaginative quality. Textual representations take the form of detailed analyses of specific cultural phenomena as interpretations of larger sociocultural realities. Geertz's (1973) interpretation of cockfights in Bali is a typical study. He describes these events as social gatherings at which men gamble and play putting their prestige, masculinity, and social statuses at risk. The cockfight reveals the multiple fixed social hierarchies of Balinese society without ever changing them. It is therefore a metasocial commentary, a text about Balinese experience informing people about the lives they live.

The attention of interpretive anthropology to local meaning prompted the critique that the intersubjective construction of field data remained untouched. Reflexive anthropology arose in the 1980s to examine the narrative genres, forms, styles, structure, and rhetoric of ethnographic texts (Clifford and Marcus, 1986). This approach was strongly influenced by the practice theory of French sociologist Pierre Bourdieu, the study of discourses of power by French social philosopher Michel Foucault, and the dialogic theory of Russian philosopher Mikhail Bakhtin. Ethnographic research became interpreted as a practice of domination because of the unequal power relation between the authoritative fieldworker and the local informant.

Reflexive ethnography implies the study of the fieldworker's interpretive presuppositions and ethnographic authority, the interaction with research participants, the intercultural construction of ethnographic knowledge, and the narrative style of research findings. Ethnographers became conscious of how knowledge was produced in research and reporting, as exemplified by Paul Rabinow's (1977) analysis of his fieldwork in a Moroccan town. Intercultural translation by key informants required that they first understand their own culture before they could describe it to a foreign researcher who lacked any experiential knowledge of the culture under study. Moroccan interlocutors who were situated at the social periphery proved to be excellent interpreters because they could see across social boundaries.

In an era in which researched 'others' increasingly 'spoke back' to what was written about them, and with a heightened awareness of the relationship between power and the construction of knowledge, ethnographers sought to increase the voice of the 'other' through more active involvement of research participants. This was achieved by extensive use of direct quotes, coauthorship, reciprocity, collaboration, and partnership. Vincent Crapanzano (1980) addressed the intersubjective construction of cultural meaning in an exegesis of his research relation with the illiterate Moroccan tilemaker Tuhami. He reproduced the dialogic quality of fieldwork by presenting a literal transcription of distinct voices, questions and answers of Tuhami and himself, intermingled with anthropological commentary. Crapanzano chose an experimental writing style to convey this field encounter; other fieldworkers took an additional step toward ethnographic fiction and poetry.

Narrative Ethnography, Testimonio, and Ethnographic Fiction

The 1980s and 1990s saw the emergence of new forms of ethnographic writing, including narrative ethnography, *testimonio*, and ethnographic fiction. *Testimonio* or testimonial narrative was based on first-person accounts of situations involving violence, marginalization, or struggle for survival to bear witness to and denounce human rights abuses. Mostly emerging from Latin America, narratives such as the autobiographical *I, Rigoberta Menchú: An Indian Woman in Guatemala* (Menchú and Burgos-Debray, 1984) were intended to give a voice to the voiceless. *Testimonio* refers to a narrative in book form, told in the first person by a narrator who is a direct participant or witness of the events they recount, and is usually presented as a 'life story.' First and foremost, *testimonio* was intended to be a tactic of empowerment and survival in which the subaltern brought the existence of oppressive and repressive conditions to the world's attention in order to achieve social and political changes. It represented a form of experimental ethnography based on a new approach to fieldwork and a new relationship between the ethnographer and the 'other' (Beverly, 2004).

One of the most interesting developments in experimental forms of ethnographic writing has been the emergence of ethnographic fiction combining art and science. Ethnography and fiction have been juxtaposed in a number of theoretical areas, such as literary and postmodern interests in ethnography as a form of literature. Although ethnographic fiction has been criticized as inappropriate for consideration as ethnography, there has been a long tradition dating back to the nineteenth century of ethnographers writing fiction informed by fieldwork experiences (Narayan, 1999). There have been two main forms of juxtaposing ethnography with fiction, namely, fictional ethnography and ethnographic fiction, both of which aim to convey ethnographic realities. Fictional ethnography is written by trained ethnographers who write fictionalized accounts presented as 'real' ethnographies, whereas ethnographic fiction is usually written as novels or stories by writers who lack formal ethnographic training. Examples of the first include Elenore Smith Bowen (1964), Carlos Castaneda (1968), and Florinda Donner (1982); the second includes Susanna Kaysen (1990) and Mischa Berlinski (2008).

Both Castaneda's and Donner's books claim to be 'real' ethnographies by 'real' ethnographers, but have been revealed to essentially be ethnographically informed fiction or literature. Goldschmidt characterizes Castaneda's first books as "both ethnography and allegory," noting that it "demonstrates the essential skill of good ethnography – the capacity to enter into an alien world" (1968: pp. vii–viii). While Castaneda maintained that his experiences were real, his critics concluded that his five Don Juan books were a hoax (de Mille, 1980). Donner's book was also presented as a 'real' ethnographic account, but it too was revealed as fiction based on real ethnographic research.

Kaysen and Berlinski are not ethnographers. Their anthropological novels are based on extensive research and achieve what Bowen, Castaneda, and Donner are acclaimed for, conveying both ethnographic knowledge about 'other' peoples and cultures and evoking the complexities of the ethnographic encounter through fictionalized accounts.

While there is debate about whether this 'creative nonfiction' strays too far from fact into the realm of fiction, and about the ethics of injecting fictional inventions within works labeled 'ethnography,' Pratt (1986) and Narayan (1999) argue that ethnography has much to gain from writing outside established discursive traditions because the crossings between ethnography and fiction have been mutually enriching for both genres.

Engaged and Covert Ethnography

The first decade of this century has been marked by the contradictory emergence of 'engaged' and 'militarized' ethnography. The trends described above solidified into a politically engaged ethnography, which seeks to critically attract a wider readership and contribute to public debate about the urgent crises facing humankind today. Since its inception, social science was intended as an applied science. As a product of the European Enlightenment, it was based on the conviction that reason could and should be applied for the reduction of suffering and improvement of the human condition. In a contemporary reaffirmation of this vision, the core ethos and eidos of engaged ethnography is its ethical imperative to not only 'avoid harm' but to intentionally work to 'do some good' where it is needed. Engaged ethnography sees ethnographers as witnesses, instead of either dispassionate observers or political activists, who connect their readers to the world's trouble spots and sensitize them to the suffering and struggles of disprivileged human beings. It seeks to generate public discussion, influence opinion, and critically engage politicians and policymakers with the explicit goal of fostering genuine social change (Sanford and Angel-Ajani, 2006; Sluka and Robben, 2012).

Unfortunately, that all is not well with the epistemology of ethnography was revealed by the reemergence of covert mission-related counterinsurgency research by social scientists deployed by the US Army, which created renewed debate about the ethics, practice, and praxis of ethnographic fieldwork today. The goal of counterinsurgency research is population control and is viewed by the military as an element of 'combat power' – that is, as a weapon. Anthropologists first confronted the issue of militarized ethnography during the Vietnam War. This controversy led directly to the adoption of the first code of anthropological ethics in 1971, which effectively banned such research for over 30 years. However, in 2006, the US Army initiated a new experimental counterinsurgency program called the Human Terrain System, which began to embed social scientists with combat brigades in Iraq and Afghanistan to gather ethnographic intelligence. This generated renewed controversy among anthropologists, most of whom view it as fundamentally unethical, inherently harmful to those studied, and an attempt to 'weaponize' or 'recolonize' ethnography as an applied science of state control.

Militarized ethnography is a deceptive and unacceptable form of covert ethnographic research because legitimate consent cannot be given at the end of the barrel of a gun; the people studied are subjected to the research rather than willing participants in it; they are not informed of the real aims and agenda; confidentiality is not guaranteed; the data obtained are likely to be used to directly control, exploit, or harm them; and the findings are officially secret and will never be published or made available to them. Involvement of ethnographers in counterinsurgency research is not a form of engaged ethnography, as its practitioners pretend, but rather its antithesis because it contradicts the fundamental intention of engaged ethnography to function as a form of empowerment and resistance to imperialism, domination, and exploitation.

Conclusion

The accelerating globalization in the 1990s affected the Boasian-Malinowskian canon of fieldwork deeply. The idea of cultures as integrated wholes and their study through a personal immersion in one setting for the duration of 1 or 2 years became untenable. Research of hybrid cultures, global interconnections, translocal subjects, and new media demanded other ethnographic approaches. Multisited ethnography has been one strategy to examine the movements, paths, and connections among people, goods, and knowledge in different locations (Marcus, 1995; Hannerz, 2004). Field-workers make short trips to a patchwork of sites, collaborate with other field-workers, and skype with research participants. Netnography has emerged as the latest development in ethnography. Computer-mediated research of online communities and social media is combined with ethnographic fieldwork (Kozinets, 2010). Online interviews, Web text analyses, and Internet crowd surfing have not displaced participant observation and face-to-face interviews but are becoming important additional research methods. Ethnography thus continues to be a valuable strategy of research and representation that adapts to changing empirical circumstances.

See also: Anthropological Research, Ethics of; Anthropological Writing; Anthropology and History; Anthropology: Overview; Cognitive Anthropology; Community Studies: Anthropological; Comparative Method in Anthropology; Culture Shock; Ethical Practices: Institutional Oversight and Enforcement; Ethnocentrism; Fieldwork in Social and Cultural Anthropology; Fieldwork: Ethical Aspects; Informed Consent of Research Subjects; Interpretation in Anthropology; Militaries and Militarization, Anthropology of; Multisited Field Studies; Objectivity of Research: Ethical Aspects; Privacy of Individuals in Social Research: Confidentiality; Reflexivity in Anthropology; Research Ethics, Cross-Cultural Dimensions of; War: Anthropological Aspects, Historical Development of.

Bibliography

Berlin, B., Kay, P., 1969. Basic Color Terms: Their Universality and Evolution. University of California Press, Berkeley.

Berlinski, M., 2008. Fieldwork: A Novel. Picador, New York.

Beverly, J., 2004. Testimonio: On the Politics of Truth. University of Minnesota Press, Minneapolis.
Bowen, E.S. (pseudonym for Laura Bohannan), 1964. A Return to Laughter: An Anthropological Novel. Natural History Press, Garden City.
Briggs, C., Bauman, R., 1999. "The foundation of all future researches": Franz Boas, George Hunt, native American texts, and the construction of modernity. American Quarterly 51, 479–552.
Castaneda, C., 1968. The Teachings of Don Juan: A Yaqui Way of Knowledge. Washington Square Press, New York.
Clifford, J., Marcus, G.E. (Eds.), 1986. Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley.
Crapanzano, V., 1980. Tuhami: Portrait of a Moroccan. The University of Chicago Press, Chicago.
D'Andrade, R., 1995. The Development of Cognitive Anthropology. Cambridge University Press, Cambridge.
de Mille, R. (Ed.), 1980. The Don Juan Papers: Further Castaneda Controversies. Ross-Erikson, Santa Barbara.
Deegan, M.J., 2008. The Chicago school of ethnography. In: Atkinson, P., Coffey, A., Delamont, S., Lofland, J., Lofland, L. (Eds.), Handbook of Ethnography. Sage, London, pp. 11–25.
Donner, F., 1982. Shabono: A Visit to a Remote and Magical World in the South American Rainforest. Triad, London.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Goldschmidt, W., 1968. Foreword. In: Castaneda, C. (Ed.), The Teachings of Don Juan. University of California Press, Berkeley, pp. vii–viii.
Hannerz, U., 2004. Foreign News: Exploring the World of Foreign Correspondents. The University of Chicago Press, Chicago.
Kaysen, S., 1990. Far Afield. Vintage Books, New York.
Kozinets, R.V., 2010. Netnography: Doing Ethnographic Research Online. Sage, London.
Launay, R. (Ed.), 2010. Foundations of Anthropological Theory: From Classical Antiquity to Early Modern Europe. Wiley-Blackwell, Malden.
Malinowski, B., 1984[1922]. Argonauts of the Western Pacific. Waveland Press, Prospect Heights.
Marcus, G.E., 1995. Ethnography in/of the world system: the emergence of multi-sited ethnography. Annual Review of Anthropology 24, 95–117.
Menchú, R., Burgos-Debray, E., 1984. I, Rogoberta Menchú: An Indian Woman in Guatemala. Verso, London.
Morgan, L.H., 1985[1877]. Ancient Society. University of Arizona Press, Tucson.
Narayan, K., 1999. Ethnography and fiction: where is the border? Anthropology and Humanism 24, 132–147.
Park, R.E., 1928. Human migration and the marginal man. American Journal of Sociology 33, 881–893.
Payne, H.C., 1981. Malinowski's style. Proceedings of the American Philosophical Society 125, 416–440.
Pratt, M.L., 1986. Fieldwork in common places. In: Clifford, J., Marcus, G.E. (Eds.), Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley, pp. 27–50.
Rabinow, P., 1977. Reflections on Fieldwork in Morocco. University of California Press, Berkeley.
Ricoeur, P., 1971. The model of the text: meaningful action considered as a text. Social Research 38, 529–562.
Sanford, V., Angel-Ajani, A. (Eds.), 2006. Engaged Observer: Anthropology, Advocacy, and Activism. Rutgers University Press, New Brunswick.
Sluka, J.A., Robben, A.C.G.M., 2012. Fieldwork in cultural anthropology: an introduction. In: Robben, A.C.G.M., Sluka, J.A. (Eds.), Ethnographic Fieldwork: An Anthropological Reader, second ed. Wiley-Blackwell, Malden, pp. 1–47.
Stocking, G.W., 1996. After Tylor: British Social Anthropology, 1888–1951. The Athlone Press, London.
Sturtevant, W.C., 1964. Studies in ethnoscience. American Anthropologist 66, 99–131.
Urry, J., 1972. Notes and Queries on Anthropology and the Development of Field Methods in British Anthropology, 1870–1920. Proceedings of the Royal Anthropological Institute of Great Britain and Ireland. pp. 45–57.

Relevant Websites

http://savageminds.org/ – Savage Minds.
http://ethnographymatters.net – Ethnography Matters.
http://en.wikipedia.org/wiki/Ethnography – Ethnography.
http://cead.org.nz – Contemporary Ethnography Across the Disciplines.
Photoethnography.com.
http://architectonictokyo.com/Architectonic_Tokyo__An_Anthropology_of_the_Moving_Earth.html – Architectonic Tokyo.
http://www.aaanet.org/resources – Anthropology Resources on the Internet.
http://www.topsite.com/best/ethnography; http://eth.sagepub.com/reports/most-cited – 10 Best Ethnography Websites.
http://www.loc.gov/folklife/other.html – Ethnographic Studies Internet Resource Page.
Atlasti.com – ATLAS.ti: The Qualitative Data Analysis & Research Software.
http://www-bcf.usc.edu/~genzuk/Ethnographic_Research.html – Center for Multilingual, Multicultural Research; University of Southern California. *A Synthesis of Ethnograhic Research*.

Ethnography in Applied Social Research

David M Fetterman, San Jose, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ethnography is the art and science of describing a group or culture. The ethnographer writes about the routine, daily lives of people. Fieldwork is the heart of the ethnographic research design. In the field, basic anthropological concepts, data collection methods and techniques, and analysis are the fundamental elements of 'doing ethnography.' Selection and use of various pieces of equipment – including the human instrument – facilitate the work. This process becomes product through analysis at various stages in ethnographic work – in field notes, memoranda, and interim reports, but most dramatically in the published report, article, or book.

Ethnography is the art and science of describing a group or culture. The description may be of a small tribal group in some exotic land or of a classroom in middle-class suburbia. The ethnographer writes about the routine, daily lives of people. The more predictable patterns of human thought and behavior are the focus of inquiry.

Ethnographers are noted for their ability to keep an open mind about the groups or cultures they are studying. However, this quality does not imply any lack of rigor. The ethnographer enters the field with an open mind, not an empty head. Before asking the first question in the field, the ethnographer begins with a problem, a theory or model, a research design, specific data collection techniques, tools for analysis, and a specific writing style. The ethnographer also begins with biases and preconceived notions about how people behave and what they think – as do researchers in every field. Indeed, the choice of what problem, geographic area, or people to study is in itself biased. Biases serve both positive and negative functions. Controlled, biases can focus and limit the research effort. Uncontrolled, they can undermine the quality of ethnographic research. To mitigate the negative effects of bias, the ethnographer must first make specific biases explicit. A series of additional quality controls, such as triangulation, contextualization, and a nonjudgmental orientation, place a check on the negative influence of bias.

An open mind also allows the ethnographer to explore rich, untapped sources of data not mapped out in the research design. The ethnographic study allows multiple interpretations of reality and alternative interpretations of data throughout the study. The ethnographer is interested in understanding and describing a social and cultural scene from the emic, or insider's perspective. The ethnographer is both storyteller and scientist; the closer the reader of ethnography comes to understanding the native's point of view, the better the story and the better the science.

Overview

Ethnographic work begins when the ethnographer selects a problem or topic and a theory or model to guide the study. The ethnographer simultaneously chooses whether to follow a basic or applied research approach to delineate and shape the effort. The research design then provides a basic set of instructions about what to do and where to go during the study. Fieldwork is the heart of the ethnographic research design. In the field, basic anthropological concepts, data collection methods and techniques, and analysis are the fundamental elements of 'doing ethnography.' Selection and use of various pieces of equipment – including the human instrument – facilitate the work. This process becomes product through analysis at various stages in ethnographic work – in field notes, memoranda, and interim reports, but most dramatically in the published report, article, or book.

This article provides a summary of the concepts, methods and techniques, equipment, analysis, writing, and ethics involved in ethnographic research. This approach highlights the utility of planning and organization in ethnographic work. The more organized the ethnographer, the easier his or her task of making sense of the mountains of data collected in the field. Sifting through notepads filled with illegible scrawl, listening to hours of digital voice recordings, labeling and organizing piles of digital photos, and cross-referencing data stored on hard drives or clouds are much less daunting to the ethnographer who has taken an organized, carefully planned approach.

The reality, however, is that ethnographic work is not always orderly. It involves serendipity, creativity, being in the right place at the right or wrong time, a lot of hard work, and old-fashioned luck. Thus, although this discussion proceeds within the confines of an orderly structure, I have made a concerted effort to ensure that it conveys as well the unplanned, sometimes chaotic, and always intriguing character of ethnographic research.

While in most research analysis follows data collection, in ethnographic research analysis and data collection begin simultaneously. An ethnographer is a human instrument and must discriminate among different types of data and analyze the relative worth of one path over another at every turn in fieldwork, well before any formalized analysis takes place. Clearly, ethnographic research involves all different levels of analysis. Analysis is an ongoing responsibility and joy from the first moment an ethnographer envisions a new project to the final stages of writing and reporting the findings.

Concepts

A few of the most important concepts that guide ethnographers in their fieldwork include culture, a holistic perspective, contextualization, emic perspective and multiple realities, etic perspective, and a nonjudgmental orientation.

Culture

Culture is the broadest ethnographic concept. Definitions of culture typically espouse either a materialist or an ideational perspective. The classic materialist interpretation of culture focuses on behavior. In this view, culture is the sum of a social group's observable patterns of behavior, customs, and way of life (see Harris, 1968: 16). The most popular ideational definition of culture is the cognitive definition. According to the cognitive approach, culture comprises the ideas, beliefs, and knowledge that characterize a particular group of people. This second definition specifically excludes behavior. Obviously, ethnographers need to know about both cultural behavior and cultural knowledge to describe a culture or subculture adequately. Although neither definition is sufficient, each offers the ethnographer a starting point and a perspective from which to approach the group under study.

Many anthropologists consider cultural interpretation ethnography's primary contribution. Cultural interpretation involves the researcher's ability to describe what he or she has heard and seen within the framework of the social group's view of reality. A classic example of the interpretive contribution involves the wink and the blink. A mechanical difference between the two may not be evident. However, the cultural context of each movement, the relationship between individuals that each act suggests, and the contexts surrounding the two help define and differentiate these two significantly different behaviors. Anyone who has ever mistaken a blink for a wink is fully aware of the significance of cultural interpretation for avoiding misunderstanding (see Fetterman, 1982: 24; Geertz, 1973: 6; Wolcott, 1980: 57, 59).

Holistic Perspective and Contextualization

Ethnographers assume a holistic outlook in research to gain a comprehensive and complete picture of a social group. Ethnographers attempt to describe as much as possible about a culture or a social group. This description might include the group's history, religion, politics, economy, and environment. No study can capture an entire culture or group. The holistic orientation forces the fieldworker to see beyond an immediate cultural scene or event in a classroom, hospital room, city street, or plush offices in Washington, D.C., New York, or Chicago. Each scene exists within a multilayered and interrelated context.

Contextualizing data involves placing observations into a larger perspective. For example, in one of my studies of an alternative high school for dropouts, it was important to describe the inner-city environment in which the schools were located – an impoverished neighborhood in which pimping, prostitution, arson for hire, rape, and murder were commonplace. This helped policy makers understand the power of certain elements in the community to distract students from their schoolwork. This description also provided some insight into the often lucrative alternatives with which the school competed in attracting and retaining students. Contextualization helped to provide a more accurate characterization of the school's degree of difficulty and helped to prevent a common error – blaming the victim (see also Fetterman, 1987a).

Emic and Etic Perspectives

The emic perspective – the insider's or native's perspective of reality – is at the heart of most ethnographic research. The insider's perception of reality is instrumental to understanding and accurately describing situations and behaviors. Native perceptions may not conform to an 'objective' reality, but they help the fieldworker understand why members of the social group do what they do. In contrast to a priori assumptions about how systems work from a simple, linear, logical perspective – which might be completely off target – ethnography typically takes a phenomenologically oriented research approach. An emic perspective compels the recognition and acceptance of multiple realities. Documenting multiple perspectives of reality in a given study is crucial to an understanding of why people think and act in the different ways they do. Differing perceptions of reality can be useful clues to individuals' religious, economic, or political status and can help a researcher understand maladaptive behavior patterns.

An etic perspective is an external, social scientific perspective on reality. Some ethnographers are interested only in describing the emic view, without placing their data in an etic or scientific perspective. They stand at the ideational and phenomenological end of the ethnographic spectrum. Other ethnographers prefer to rely on etically derived data first, and consider emically derived data secondary in their analysis. They stand at the materialist and positivist philosophical end of the ethnographic spectrum. Most ethnographers simply see emic and etic orientations as markers along a continuum of styles or different levels of analysis. Most ethnographers start collecting data from the emic perspective, then try to make sense of what they have collected in terms of both the native's view and their own scientific analysis. Just as thorough fieldwork requires an insightful and sensitive cultural interpretation combined with rigorous data-collection techniques, so good ethnography requires both emic and etic perspectives.

Nonjudgmental Orientation

A nonjudgmental orientation requires the ethnographer to suspend personal valuation of any given cultural practice. Maintaining a nonjudgmental orientation is similar to suspending disbelief while watching a movie or play, or reading a book – one accepts what may be an obviously illogical or unbelievable set of circumstances in order to allow the author to unravel a riveting story.

Concepts of culture, a holistic perspective, contextualization, emic perspective and multiple realities, etic perspective, and a nonjudgmental orientation place a check on our observations. They help the fieldworker see differences that may invalidate pat theories or hypotheses about observed events in the field. In some cases, these differences are systematic patterned activities for a broad spectrum of the community,

compelling the fieldworker to readjust the research focus; to throw away dated and inappropriate theories, models, hypotheses, and assumptions; and to modify the vision of the finished puzzle. In other cases, the differences are idiosyncratic but useful in underscoring another, dominant pattern – the exception that proves the rule. In most cases, however, such differences are instructive about a level or dimension of the community that had not received sufficient consideration (see Lavenda and Schultz, 2012 for additional discussion about concepts).

Methods and Techniques

The ethnographer is a human instrument. Ethnographic methods and techniques help to guide the ethnographer through the wilderness of personal observation and to identify and classify accurately the bewildering variety of events and actions that form a social situation.

Fieldwork

Fieldwork is the hallmark of research for both sociologists and anthropologists. The method is essentially the same for both types of researchers – working with people for long periods of time in their natural setting. The ethnographer conducts research in the native environment to see people and their behavior given all the real-world incentives and constraints. This naturalist approach avoids the artificial response typical of controlled or laboratory conditions. Understanding the world – or some small fragment of it – requires studying it in all its wonder and complexity. The task is in many ways more difficult than laboratory study, but it can also be more rewarding (see Mcall, 2006; Blommaert and Jie, 2011).

The fieldworker uses a variety of methods and techniques to ensure the integrity of the data. These methods and techniques objectify and standardize the researcher's perceptions. Of course, the ethnographer must adapt each one of the methods and techniques discussed below to the local environment. Resource constraints and deadlines may also limit the length of time for data gathering in the field – exploring, cross-checking, and recording information.

Selection, Sampling, and Entry

Research questions shape the selection of a place and a people or program to study. The ideal site for investigation of the research problem is not always accessible. In that event, the researcher accepts and notes the limitations of the study from the onset. Ideally, the focus of the investigation shifts to match the site under study.

The next step is to decide how to sample members of the target population. Most ethnographers use the big-net approach conducive to participant observation – mixing and mingling with everyone they can at first. As the study progresses, the focus narrows to specific portions of the population under study. The big-net approach ensures a wide-angle view of events before the microscopic study of specific interactions begins.

Ethnographers typically use informal strategies to begin fieldwork, such as starting wherever they can slip a foot in the door. (An introduction by a member is the ethnographer's best ticket into the community.) The most common technique is judgmental sampling; that is, ethnographers rely on their judgment to select the most appropriate members of the subculture or unit, based on the research question. Some experienced ethnographers use a rigorous randomized strategy to begin work – particularly when they already know a great deal about the culture or unit they are studying. However, using a highly structured randomized strategy without a basic understanding of the people under study may cause the researcher to narrow the focus prematurely, thus eliminating perhaps the very people or subjects relevant to the study.

Participant Observation

Participant observation characterizes most ethnographic research and is crucial to effective fieldwork. Participant observation combines participation in the lives of the people under study with maintenance of a professional distance that allows adequate observation and recording of data.

Participant observation is immersion in a culture. Ideally, the ethnographer lives and works in the community for 6 months to a year or more, learning the language and seeing patterns of behavior over time. Long-term residence helps the researcher internalize the basic beliefs, fears, hopes, and expectations of the people under study. The simple, ritualistic behaviors of going to the market or to the well for water teach how people use their time and space, how they determine what is precious, sacred, and profane. The process may seem unsystematic; in the beginning, it is somewhat uncontrolled and haphazard. However, even in the early stages of fieldwork the ethnographer searches out experiences and events as they come to his or her attention. Participant observation sets the stage for more refined techniques – including projective techniques and questionnaires – and becomes more refined itself as the fieldworker understands more and more about the culture. Ideas and behaviors that were only a blur to the ethnographer on entering the community take on a sharper focus. Participant observation can also help clarify the results of more refined instruments by providing a baseline of meaning and a way to reenter the field to explore the context for those (often unexpected) results.

In applied settings, participant observation is often noncontinuous, spread out over an extended time. Often the contract research budgets or time schedules of applied research do not allow long periods of study – continuous or noncontinuous. In these situations, the researcher can apply ethnographic techniques to the study, but cannot conduct ethnography.

Interviewing

The interview is the ethnographer's most important data-gathering technique. Interviews explain and put into a larger context what the ethnographer sees and experiences. General interview types include structured, semistructured, informal, and retrospective interviews.

Formally structured and semistructured interviews are verbal approximations of a questionnaire with explicit research goals. These interviews generally serve comparative and representative purposes – comparing responses and putting them in the context of common group beliefs and themes. A structured or semistructured interview is most valuable when the fieldworker comprehends the fundamentals of a community from the 'insider's' perspective. At this point, questions are more likely to conform to the native's perception of reality than to the researcher's.

Informal interviews are the most common form in ethnographic work. They seem to be casual conversations, but where structured interviews have an explicit agenda, informal interviews have a specific but implicit research agenda. The researcher uses informal approaches to discover the categories of meaning in a culture. Informal interviews are useful throughout an ethnographic study for discovering what people think and how one person's perceptions compare with another's. Such comparisons help the fieldworker to identify shared values in the community – values that inform behavior. Informal interviews are also useful for establishing and maintaining healthy rapport.

The ethnographer uses retrospective interviews to reconstruct the past, asking informants to recall personal historical information; retrospective interviews can be structured, semistructured, or informal. This type of interview does not elicit the most accurate data. People forget or filter past events. In some cases, retrospective interviews are the only way to gather information about the past. In situations where the ethnographer already has an accurate understanding of the historical facts, retrospective interviews provide useful information about individuals.

Key Actor or Informant Interviewing

Some people are more articulate and culturally sensitive than others. These individuals make excellent key actors or informants. In a social group under study, this individual is one of many actors, though they may not be a central or even an indispensable community member. Yet this individual becomes a key actor in the theater of ethnographic research and plays a pivotal role, linking the fieldworker and the community.

Key actors can provide detailed historical data, knowledge about contemporary interpersonal relationships (including conflicts), and a wealth of information about the nuances of everyday life. Although the ethnographer tries to speak with as many people as possible, time is always a factor. Therefore, anthropologists have traditionally relied most heavily on one or two individuals in a given group.

Key actors require careful selection. They are rarely perfect representatives of the group. However, they are usually members of the mainstream – otherwise, they would not have access to up-to-date cultural information. Key actors may be cultural brokers, straddling two cultures. This position may give them a special vantage point and objectivity about their culture. They may also be informal or formal leaders in the community. Key actors come from all walks of life and all socioeconomic and age groups.

The key actor and ethnographer must share a bond of trust. Respect on both sides is earned slowly. The ethnographer must take time to search out and spend time with these articulate individuals. The fieldworker learns to depend on the key actor's information – particularly as cross-checks with other sources prove it to be accurate and revealing. Ethnographers establish long-term relationships with key actors who continually provide reliable and insightful information. Key actors can be extremely effective and efficient sources of data and analysis.

At the same time, the ethnographer must judge the key actor's information cautiously. Overreliance on a key actor can be dangerous. Every study requires multiple sources. In addition, the fieldworker must take care to ensure that key actors do not simply provide answers they think the fieldworker wants to hear. The ethnographer can check answers rather easily, but must stay on guard against such distortion and contamination. Another, subtler problem occurs when a key actor begins to adopt the ethnographer's theoretical and conceptual framework. The key actor may inadvertently begin to describe the culture in terms of this a priori construct, undermining the fieldwork and distorting the emic or insider's perspective (for further discussion of the role of key informants, see Dobbert, 1982; Ellen, 1984; Freilick, 1970; Goetz and LeCompte, 1984; Pelto, 1970; Spradley, 1979; Taylor and Bogdan, 1984).

Questionnaires

Structured interviews are close approximations of questionnaires. Questionnaires represent perhaps the most formal and rigid form of exchange in the interviewing spectrum – the logical extension of an increasingly structured interview. However, questionnaires are qualitatively different from interviews because of the distance between the researcher and the respondent. Interviews have an interactive nature that questionnaires lack. In filling out a questionnaire, the respondent completes the researcher's form without any verbal exchange or clarification. Knowing whether the researcher and the respondent are on the same wavelength, sharing common assumptions and understandings about the questions, is difficult – perhaps impossible.

Misinterpretations and misrepresentations are common with questionnaires. Many people present idealized images of themselves on questionnaires, answering as they think they should to conform to a certain image. The researcher has no control over this type of response and no interpersonal cues to guide the interpretation of responses. Other problems include bias in the questions and poor return rates.

Despite these caveats, questionnaires are an excellent way for fieldworkers to tackle questions dealing with representativeness. They are the only realistic way of taking the pulses of hundreds or thousands of people. Anthropologists usually develop questionnaires to explore scientific concerns after they have a good grasp of how the larger pieces of the puzzle fit together. The questionnaire is a product of the ethnographer's knowledge about the system, and the researcher can adapt it to a specific topic or set of concerns. Ethnographers also use existing questionnaires to test hypotheses about specific conceptions and behaviors. However, the ethnographer must establish the relevance of a particular questionnaire to the target culture or subculture before administering it.

Online surveys and questionnaires provide an efficient way to document the views of large groups in a short period of time.

The questions are posted on the web, including yes/no, all that apply, open-ended, and five-point Likert scale questions. Respondents are notified about the location of the survey on the web (with a specific URL), enter their responses, and submit their survey online. The results are automatically calculated. The responses are often visually represented in a bar chart or similar graphic display as soon as the data are entered. This saves the ethnographer from the initial mailing costs, time-consuming and expensive postal reminders, and the expense of data entry concerning all the submitted surveys. Ethnographers can help computer-phobic respondents or those who do not have access to computer or the Internet complete the survey and enter their responses in the same online data base if necessary.

Projective Techniques

Projective techniques supplement and enhance fieldwork, they do not replace it. These techniques are employed by the ethnographer to elicit cultural and often psychological information from group members. Typically, the ethnographer holds an item up and asks the participant what it is. The researcher may have an idea about what the item represents, but that idea is less important than the participant's perception. The participant's responses usually reveal individual needs, fears, inclinations, and general worldview.

Project techniques, however revealing, rarely stand alone. The researcher needs to set these techniques in a larger research context to understand the elicited response completely. Projective techniques can elicit cues that can lead to further inquiry or can be one of several sources of information to support an ongoing hypothesis. Only the ethnographer's imagination limits the number of possible projective techniques. However, the fieldworker should use only those tests that can be relevant to the local group and the study.

Unobtrusive Measures

Most ethnographic methods are interactive: They involve dealing with people. A variety of other measures, however, does not require human interaction and can supplement interactive methods of data collection and analysis. These methods require only that the ethnographer keep eyes and ears open. Unobtrusive measures allow the ethnographer to draw social and cultural inferences from physical evidence (see Webb et al., 1966).

Outcroppings

Outcropping is a geological term referring to a portion of the bedrock that is visible on the surface – in other words, something that sticks out. Outcroppings in inner-city ethnographic research include skyscrapers, burned-out buildings, graffiti, the smell of urine on city streets, yards littered with garbage, a Rolls-Royce, or a syringe in the schoolyard. The researcher can quickly estimate the relative wealth or poverty of an area from these outcroppings. Initial inferences are possible without any human interaction. However, such cues by themselves can be misleading. A house with all the modern conveniences and luxuries imaginable can signal wealth, or can signal financial overextension verging on bankruptcy. The researcher must place each outcropping in a larger context. A broken syringe can have several meanings, depending on whether it lies on the floor of a doctor's office or in an elementary schoolyard late at night. On the walls of an inner-city school, the absence of graffiti is as important as its presence.

Changes in a physical setting over time can also be revealing. For example, an increase in the number of burned-out and empty buildings on a block indicates a decaying neighborhood. Conversely, an increase in the number of remodeled and revitalized houses may be indicative of gentrification, in which wealthy investors take over the neighborhood. The fieldworker must assess this abundant information with care, but should not ignore it or take it for granted.

Written and Electronic Information

In literate societies, written documents provide one of the most valuable and timesaving forms of data collection. In studies of office life, I have found past reports, memoranda, and personnel and payroll records invaluable. Mission statements and annual reports provide the organization's purpose or stated purpose and indicate the image the organization wishes to present to the outside world. Budgets tell a great deal about organizational values. Electronic mail is often less inhibited than general correspondence and thus quite revealing about office interrelationships, turf, and various power struggles. Proper use of this type of information can save the ethnographer's years of work.

Equipment

Notepads, computers, tape recorders, cameras – all the tools of ethnography are merely extensions of the human instrument, which aid to memory and vision. Yet these useful devices can facilitate the ethnographic mission by capturing the rich detail and flavor of the ethnographic experience and then helping to organize and analyze these data. Ethnographic equipment ranges from simple paper and pen to laptops and cloud storage, from digital recorders to digital cameras. The proper equipment can make the ethnographer's sojourn in an alien or unfamiliar culture more pleasant, safe, productive, and rewarding.

Analysis

Analysis is one of the most engaging features of ethnography. It begins at the moment a fieldworker selects a problem to study and ends with the last word in the report or ethnography. Ethnography involves many levels of analysis. Some are simple and informal; others require some statistical sophistication. Ethnographic analysis is iterative, building on ideas throughout the study. Analyzing data in the field enables the ethnographer to know precisely which methods to use next, as well as when and how to use them. Through analysis, the ethnographer tests hypotheses and perceptions to construct an accurate conceptual framework about what is happening in the social group under study. Analysis in ethnography is as much a test of the ethnographer as it is a test of the data.

Thinking

First and foremost, analysis is a test of the ethnographer's ability to think – to process information in a meaningful and useful manner. The ethnographer confronts a vast array of complex information and needs to make some sense of it all – piece by piece. The initial stage in analysis involves simple perception. However, even perception is selective. The ethnographer selects and isolates pieces of information from all the data in the field. The ethnographer's personal or idiosyncratic approach, together with an assortment of academic theories and models, focuses and limits the scope of inquiry. However, the field presents a vast amount of material, and in understanding day-to-day human interaction, elementary thinking skills such as those required for processing information are as important as ethnographic concepts and methods.

The ethnographer employs many useful techniques to make sense of the forests of data, from triangulation to the use of statistical software packages. All these techniques, however, require critical thinking skills – notably, the ability to synthesize and evaluate information – and a large dose of common sense.

Triangulation

Triangulation is basic in ethnographic research. It is at the heart of ethnographic validity, testing one source of information against another to strip away alternative explanations and prove a hypothesis. Typically, the ethnographer compares information sources to test the quality of the information (and the person sharing it), to understand more completely the part an actor plays in the social drama, and ultimately to put the whole situation into perspective (see Webb et al., 1966, for a detailed discussion of triangulation).

Patterns

Ethnographers look for patterns of thought and behavior. Patterns are a form of ethnographic reliability. Ethnographers see patterns of thought and action repeated in various situations and among various players. Looking for patterns is a form of analysis. The ethnographer begins with a mass of undifferentiated ideas and behavior, and then collects pieces of information, comparing, contrasting, and sorting gross categories and minutiae until a discernible thought or behavior becomes identifiable. Next the ethnographer must listen and observe, and then compare his or her observations with this poorly defined model. Exceptions to the rule emerge, variations on a theme are detectable. These variants help to circumscribe the activity and clarify its meaning. The process requires further sifting and sorting to make a match between categories. The theme or ritualistic activity finally emerges, consisting of a collection of such matches between the model (abstracted from reality) and the ongoing observed reality.

Any cultural group's patterns of thought and behavior are interwoven strands. As soon as the ethnographer finishes analyzing and identifying one pattern, another pattern emerges for analysis and identification. The fieldworker can then compare the two patterns. In practice, the ethnographer works simultaneously on many patterns. The level of understanding increases geometrically as the ethnographer moves up the conceptual ladder – mixing and matching patterns and building theory from the ground up (see Glaser and Strauss, 1967, for a discussion of grounded theory).

The observer can make preliminary inferences about the entire social system by analyzing the behavior that is subsumed within the pattern, as well as the patterns themselves. Ethnographers acquire a deeper understanding of and appreciation for a culture as they weave each part of the ornate human tapestry together, by observing and analyzing the patterns of everyday life.

Key Events

Key or focal events that the fieldworker can use to analyze an entire culture occur in every social group – Geertz (1973) eloquently used the cockfight to understand and portray Balinese life. Key events come in all shapes and sizes. Some tell more about a culture than others, but all provide a focus of analysis (see also Geertz, 1957).

Key events, like digital photographs or videos, concretely convey a wealth of information. Some images are clear representations of social activity; others provide a tremendous amount of embedded meaning. Once the event is recorded, the ethnographer can enlarge or reduce any portion of the picture. A rudimentary knowledge of the social situation will enable the ethnographer to infer a great deal from key events. In many cases, the event is a metaphor for a way of life or a specific social value. Key events provide lenses through which to view a culture.

Key events are extraordinarily useful for analysis. Not only do they help the fieldworker understand a social group, but the fieldworker in turn also can use them to explain the culture to others. The key event thus becomes a metaphor for the culture. Key events also illustrate how participation, observation, and analysis are inextricably bound together during fieldwork.

Content Analysis

Ethnographers analyze written and electronic data in much the same way they analyze observed behavior. They triangulate information within documents to test for internal consistency. They attempt to discover patterns within the text and seek key events recorded and memorialized in print.

Ethnographers may subject internal documents to special scrutiny to determine whether they are internally consistent with program philosophy. Reviews may also reveal significant patterns. It is often possible for the ethnographer to infer the significance of a concept from its frequency and context in the text (see Graneheim and Lundman, 2004; Neurendorf, 2001; Roberts, 1997; Stemler, 2001).

Statistics and Software

Ethnographers use nonparametric statistics more often than parametric statistics because they typically work with small samples. Parametric statistics require large samples for statistical significance. The use of nonparametric statistics is also more consistent with the needs and concerns of most anthropologists. An anthropologist typically works with nominal and

ordinal scales. Nominal scales consist of discrete categories, such as sex and religion. Ordinal scales also provide discrete categories as well as a range of variation within each category – for example, reform, conservative, and orthodox variants within the category of Judaism. Ordinal scales do not determine the degree of difference between subcategories. The Guttman (1944) scale also known as cumulative scaling or scalogram (Trochim, 2006a) analysis is one example of an ordinal scale that is useful in ethnographic research.

The chi-square test and the Fisher's exact probability test are popular nonparametric statistical tools in anthropology. However, all statistical formulas require that certain assumptions be met before the formulas may be applied to any situation. A disregard for these variables in the statistical equation is as dangerous as neglect of comparable assumptions in the human equation in conducting ethnographic fieldwork. Both errors result in distorted and misleading efforts at worst, and waste valuable time at best.

Ethnographers use parametric statistics when they have large samples and limited time and resources to conduct all the interviews. Survey and questionnaire work often requires sophisticated statistical tests of significance. T-tests are used to determine whether the means of two groups are statistically significant (Trochim, 2006b). Analysis of covariance design in regression analysis is another common test used when sample size permits in ethnography. Ethnographers also use the results of parametric statistics to test certain hypotheses, cross-check their own observations, and generally provide additional insight (see Fetterman, 2010, for discussion of problems with statistics).

There are many statistical software packages used to facilitate analysis ranging from Excel for descriptive statistics to SPSS. In addition, there are a number of software packages designed to facilitate qualitative data analysis, such as ATLAS.ti, MAXQDA, HyperResearch, and QDA.

Writing

Ethnography requires good writing skills at every stage of the enterprise. Research proposals, field notes, memoranda, interim reports, final reports, articles, and books are the tangible products of ethnographic work. The ethnographer can share these written works with participants to verify their accuracy and with colleagues for review and consideration. Ethnography offers many intangibles, through the media of participation and verbal communication. However, written products, unlike transitory conversations and interactions, withstand the test of time.

Writing good field notes is very different from writing a solid and illuminating ethnography or ethnographically informed report. Note taking is the rawest kind of writing. The note taker typically has an audience of one. Thus, although clarity, concision, and completeness are vital in note taking, style is not a primary consideration (see Emerson et al., 1995).

Writing for an audience, however, means writing to that audience. Reports for academics, government bureaucrats, private and public industry officials, medical professionals, and various educational program sponsors require different formats, languages, and levels of abstraction. The brevity and emphasis on findings in a report written for a program-level audience might raise some academics' eyebrows and cause them to question the project's intellectual effort. Similarly, a refereed scholarly publication would frustrate program personnel, who would likely feel that the researcher is wasting their time with irrelevant concerns, time that they need to take care of business. In essence, both parties feel that the researcher is simply not in touch with their reality. These two audiences are both interested in the fieldwork and the researcher's conclusions, but have different needs and concerns. Good ethnographic work can usually produce information that is relevant to both parties. (See Fetterman, 1987b, for discussion of the ethnographer as rhetorician. See also Yin, 1994, for discussion of differing audiences in the presentation of a case study.)

Writing is part of the analysis process as well as a means of communication (see also Hammersley and Atkinson, 1983). Writing clarifies thinking. In sitting down to put thoughts on paper, an individual must organize those thoughts and sort out specific ideas and relationships. Writing often reveals gaps in knowledge. If the researcher is still in the field when he or she discovers those gaps, the researcher needs to conduct additional interviews and observations of specific settings. If the researcher has left the field, field notes, e-mails, videoconference chats, and telephone calls must suffice (Fetterman, 1996). Embryonic ideas often come to maturity during writing, as the ethnographer crystallizes months of thought on a particular topic. From conception–as a twinkle in the ethnographer's eye – to delivery in the final report, an ethnographic study progresses through written stages. (For additional discussions of ethnographic writing, see Fetterman, 2010; Madison, 2005; Wolcott, 1990. See Van Maanen, 2011 for some of the rhetorical and narrative devices used in ethnographic work, including realist, confessional, and impressionist tales.)

Ethics

Ethnographers do not work in a vacuum, they work with people. They often pry into people's innermost secrets, sacred rites, achievements, and failures. In pursuing this personal science, ethnographers subscribe to a code of ethics that preserves the participants' rights, facilitates communication in the field, and leaves the door open for further research.

This code specifies first and foremost that the ethnographer do no harm to the people or the community under study. In seeking a logical path through the cultural wilds, the ethnographer is careful not to trample the feelings of natives or desecrate what the culture calls sacred. This respect for social environment ensures not only the rights of the people, but also the integrity of the data and a productive, enduring relationship between the people and the researcher. Professionalism and a delicate step demonstrate the ethnographer's deep respect, admiration, and appreciation for the people's way of life. Noninvasive ethnography is not only good ethics, but is also good science (see American Anthropological Association, 2012; Fetterman, 1983; Rynkiewich and Spradley, 1976; Weaver, 1973). Basic underlying ethical standards include the securing of permission (to protect individual privacy), honesty, trust (both implicit and explicit), and reciprocity.

Conclusion

This article highlighted the basic elements of ethnography. It included a discussion about concepts, methods and techniques, equipment, analysis, writing, and ethics. Step by step, this article provided a path through the complex terrain of ethnographic work.

See also: Anthropology, History of; Applied Social Research, History of; Applied Sociology; Cultural Sociology, History of; Habitus: History of a Concept; Interviewing, History of; Mixed Methods; Qualitative Research; Sociology, History of; Unobtrusive Methods, History of.

Bibliography

American Anthropological Association, 2012. Statement on Ethics: Principles of Professional Responsibility. American Anthropological Association, Arlington, VA. Available at: http://www.aaanet.org/profdev/ethics/upload/Statement-on-Ethics-Principles-of-Professional-Responsibility.pdf.

Blommaert, J., Jie, D., 2011. Ethnographic Fieldwork: A Beginner's Guide. Multilingual Matters, Great Bristol.

Denzin, N.K., 1989. Interpretive Interactionism. Sage, Newbury Park, CA.

Denzin, N.K., 2003. Performance Ethnography: Critical Pedagogy and the Politics of Culture. Sage, Thousand Oaks, CA.

Dobbert, M.L., 1982. Ethnographic Research: Theory and Application for Modern Schools and Societies. Praeger, New York.

Ellen, R.F., 1984. Ethnographic Research: A Guide to General Conduct. Academic Press, New York.

Emerson, R., Fretz, R., Shaw, L., 1995. Writing Ethnographic Fieldnotes. University of Chicago Press, Chicago, IL.

Fetterman, D.M., 1982. Ethnography in educational research: the dynamics of diffusion. Educational Researcher 11 (3), 17–29.

Fetterman, D.M., 1983. Guilty knowledge, dirty hands, and other ethical dilemmas: the hazards of contract research. Human Organization 42, 214–224.

Fetterman, D.M., 1987a. Ethnographic educational evaluation. In: Spindler, G.D. (Ed.), Interpretive Ethnography of Education: At Home and Abroad. Lawrence Erlbaum, Hillsdale, NJ.

Fetterman, D.M., November 18–22, 1987b. Multiple Audiences Reflect Multiple Realities. Invited presentation at the 86th Annual Meeting of the American Anthropological Association, Chicago.

Fetterman, D.M., 1996. Videoconferencing on-line: enhancing communication over the internet. Educational Researcher 25 (4), 23–26.

Fetterman, D.M., 2010. Ethnography: Step by Step. Sage, Newbury Park, CA.

Freilick, M. (Ed.), 1970. Marginal Natives: Anthropologists at Work. Harper & Row, New York.

Geertz, C., 1957. Ritual and social change: a Javanese example. American Anthropologist 59, 32–54.

Geertz, C., 1973. The Interpretation of Cultures: Selected Essays. Basic Books, New York.

Glaser, B.G., Strauss, A.L., 1967. The Discovery of Grounded Theory: Strategies for Qualitative Research. Aldine, Chicago.

Goetz, J.P., LeCompte, M.D., 1984. Ethnography and Qualitative Design in Educational Research. Academic Press, New York.

Graneheim, U.H., Lundman, B., 2004. Qualitative content analysis in nursing research: concepts, procedures, and measurements of trustworthiness. Nurse Education Today 24 (2), 105–112.

Guttman, L., 1944. A basis for scaling qualitative data. American Sociological Review 9, 139–150.

Hammersley, M., Atkinson, P., 1983. Ethnography: Principles in Practice. Tavistock, London.

Harris, M., 1968. The Rise of Anthropological Theory. Thomas Y. Crowell, New York.

Lavenda, R., Schultz, E., 2012. Core Concepts in Cultural Anthropology. McGraw-Hill, New York.

Madison, S.D., 2005. Critical Ethnography: Method, Ethics, and Performance. Sage, Thousand Oaks, CA.

Mcall, G.J., 2006. The fieldwork tradition. In: Hobbs, D., Wright, R. (Eds.), The Sage Handbook of Fieldwork. Sage, Thousand Oaks, CA.

Neurendorf, K., 2001. The Content Analysis Cookbook. http://academic.csuohio.edu/kneuendorf/content/resources/car.htm.

Pelto, P.J., 1970. Anthropological Research: The Structure of Inquiry. Harper & Row, New York.

Roberts, C., 1997. Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts. Lawrence Erlbaum, New Jersey.

Rynkiewich, M.A., Spradley, J.P., 1976. Ethics and Anthropology: Dilemmas in Fieldwork. John Wiley, New York.

Spradley, J.P., 1979. The Ethnographic Interview. Holt, Rinehart & Winston, New York.

Stemler, S., 2001. An overview of content analysis. Practical Assessment, Research, & Evaluation 7 (17). http://pareonline.net/getvn.asp?v=7&n=17.

Taylor, S.J., Bogdan, R.C., 1984. Introduction to Qualitative Research Methods: The Search for Meanings. John Wiley, New York.

Trochim, W., 2006a. Guttman Scale. Research Methods Knowledge Base. http://www.socialresearchmethods.net/kb/scalgutt.htm.

Trochim, W., 2006b. T-test. Research Methods Knowledge Base. http://www.socialresearchmethods.net/kb/stat_t.htm.

Van Maanen, J., 2011. Tales of the Field: On Writing Ethnography. University of Chicago Press, Chicago.

Weaver, T., 1973. To See Ourselves: Anthropology and Modern Social Issues. Scott, Foresman, Glenview, IL.

Webb, E.J., Campbell, D.T., Schwartz, R.D., Sechrest, L., 1966. Unobtrusive Measures: Nonreactive Research in the Social Sciences. Rand McNally, Chicago.

Wolcott, H.F., 1990. Writing up Qualitative Research. Sage, Newbury Park, CA.

Yin, R.K., 1994. Case Study Research: Design and Methods, second ed. Sage, Thousand Oaks, CA.

Ethnohistory

Pauline T Strong, The University of Texas at Austin, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ethnohistory is an interdisciplinary approach to indigenous, colonial, and postcolonial culture and history. Combining the approaches of history, cultural anthropology, and archaeology, ethnohistory has most often focused on the cultures and histories of the indigenous peoples of settler societies in the Americas, Australia, New Zealand, the Pacific, and South Asia. Ethnohistory encompasses both particularistic and comparative scholarship and embodies productive tensions among historical, anthropological, and indigenous perspectives on cultural and historical processes.

Introduction

Ethnohistory is an interdisciplinary approach to indigenous, colonial, and postcolonial culture and history that developed as a coherent scholarly field in the United States in the 1950s. Combining the approaches of history, cultural anthropology, and archaeology, ethnohistory centers on reconstructing the history of non-European peoples, including their experiences of colonization and resistance. Ethnohistory began as an applied field, as historians and anthropologists collaborated on US tribal land claims cases, and it soon became institutionalized in an organization now known as the American Society for Ethnohistory (ASE). Ethnohistory expanded unevenly into other regional contexts, especially in South Asia and the Pacific, but the Americas has remained its center of gravity. Although there are programs in ethnohistory in university, museum, and governmental contexts, most ethnohistorians are trained primarily as anthropologists, historians, or archaeologists, such that ethnohistory remains an interdisciplinary pursuit rather than a discipline in itself.

There are strong connections between ethnohistory and other forms of interdisciplinary inquiry such as historical anthropology, historical archaeology, social history, oral history, subaltern studies, colonial studies, and indigenous studies, but ethnohistory is a fairly coherent scholarly formation that remains true to its particularistic and empirical roots. This article surveys the history of the field, some of its contributions to knowledge, and some of the ways in which ethnohistory has integrated and revealed tensions among historical, anthropological, and indigenous perspectives on cultural and historical processes. Such tensions have led to a productive reflexivity among some ethnohistorians.

The Development of Ethnohistory

Following the enactment in the United States of the Indian Claims Commission Act of 1946, anthropologists and historians were asked to serve as expert witnesses in court cases adjudicating tribal claims to territory. As legal historian Christian McMillen (2007) shows in *Making Indian Law: The Hualapai Land Case and the Birth of Ethnohistory*, this development was influenced by a landmark Supreme Court decision, *United States v. Santa Fe Pacific Railroad Company* (1941), that defined the basis of indigenous land claims as evidence of occupancy from 'time immemorial.' Conducting research for both sides (Indian tribes on the one hand and the Justice Department on the other), expert witnesses relied on colonial documents, oral histories, and ethnographic research to delineate indigenous concepts of rights to territory and resources; indigenous patterns of land use; indigenous and colonial understandings of treaty rights; the history of appropriation of land and resources from indigenous communities; and cultural continuities and discontinuities among indigenous populations. After 1978, when the Indian Claims Commission issued its final report, ethnohistorians continued to serve as expert witnesses in other contexts, including tribal recognition cases.

By 1978, however, ethnohistorians were conducting research that went well beyond the testimonial. The pursuit of ethnohistory consolidated a previously marginalized focus on Indian history within the field of history; in anthropology it brought a historical lens to culture change within indigenous cultures. The diachronic emphasis of ethnohistorical research provided a corrective to the synchronic nature of traditional ethnographic research, in which indigenous people were largely treated as static and timeless. Ethnography's focus on the 'ethnographic present' developed as a corrective to the speculative history of nineteenth-century historical evolutionism, but erroneously treated indigenous people as if they were 'without history,' in anthropologist Eric Wolf's (1982) memorable phrase. The neglect of history among anthropologists should not be overemphasized; however, the work of James Mooney on the Ghost Dance, published in the late nineteenth century, is an important exception to the rule. In the early twentieth century, John Swanton, Frank Speck, William Duncan Strong, Julian Steward, and William Fenton used historical methods in reconstructing indigenous culture; the latter became important in the development of ethnohistory. The acculturation studies carried out by Robert Redfield, Melville Herskovits, Robert Redfield, Oscar Lewis, and others in the 1930s also constituted a move toward understanding cultural change. In Britain, Evans-Pritchard's account of structural change among the Bedouins of Cyrenaica launched an interest in structural change and cultural conceptions of time that moved the discipline away from the synchronicity of structural functionalism.

Research on the politics, economics, ideologies, and social relations of what Georges Balandier called the 'colonial situation' (Cohn, 1980) became central to the development of

ethnohistory as a field. Ethnohistorians have productively analyzed the political economy of colonial empires, the impact of epidemics on indigenous populations, the nature of imperial bureaucracies, indigenous strategies of survival and resistance, and Christian evangelization and indigenous responses, among other topics. Studies of the gendered nature of the colonial situation, and especially changes in gender roles under colonialism and Christian evangelization, led to the development of feminist ethnohistory. Anthropologist Eleanor Leacock conducted important early work in this area, researching Innu (Naskapi) hunting territories and the impact of the fur trade on gender roles. In *Women and Colonization* (Etienne and Leacock, 1980), Leacock offered an historical materialist explanation of gender inequality that maintained that the allied forces of capitalism, colonialism, and Christianity led to the degradation of women's status among Native Americans, though not without resistance. Other studies inspired by Leacock have shown that the impact of capitalism and Christianity on gender roles has been extremely complex, varying according to indigenous gender relations as well as the manner in which indigenous economies articulated with the capitalist economy.

More recently what might be called the ethnohistory of sexuality has emerged in the work of scholars such as Ramón Gutierrez, Will Roscoe, and Ann Stoler interested in the dynamics of race, class, and sexuality in colonial systems of inequality. But this scholarship is associated less with ethnohistory than with an efflorescence of work in historical anthropology that emerged in the 1980s and 1990s. Historical anthropology, more a subfield of anthropology than an interdisciplinary field, is mainly outside the scope of this article (but *see* Anthropology and History).

The ASE, founded in 1954, awards the Erminie Wheeler-Voegelin Book Award annually to the author of the best book in the field. This award and the companion Robert F. Heizer Article Award offer an excellent way to keep abreast of developments in ethnohistory. The Society publishes the journal *Ethnohistory*; other venues in which ethnohistorical works frequently appear are *Comparative Studies in Society and History* and *History and Anthropology*. The ASE presidential address is published each year in *Ethnohistory* and often addresses current theoretical or methodological issues in the field.

Contributions to Knowledge

Ethnohistorians have compiled numerous collections of primary documents, including documents in indigenous languages, and cultural histories of particular indigenous peoples. These are important for providing fine-grained diachronic perspectives on indigenous experiences and perspectives. Of those ethnohistories that go beyond the particular to establish or exemplify significant strands of ethnohistorical scholarship, we will only be able to consider a representative sample.

One of the most influential ethnohistorical monographs ever published is anthropologist Anthony Wallace's *The Death and Rebirth of the Seneca*. In this ethnohistory Wallace draws on both archival and ethnographic research to reveal the development of the Handsome Lake religion among the Seneca Indians in the eighteenth century, which he analyzes as an example of a revitalization movement. This book has been a model for subsequent analyses of cultural revitalization, including *Reassessing Revitalization Movements: Perspectives from North America and the Pacific Islands*, edited by Michael Harkin. This collection looks comparatively at Pacific cargo cults, the Ghost Dance movement on the US Plains, and contemporary social movements, showing the comparative power and contemporary significance of Wallace's approach to cultural revitalization. Ethnohistorians have also looked at revitalization in the context of indigenous schools, cultural centers, language programs, and economic development efforts.

In the area of political economy the work of environmental historian Richard White has been particularly significant. Beginning with a monograph on the colonial production of political and economic dependency among the Choctaws, Pawnees, and Navajos, White has developed an approach to ethnohistory that focuses on changing modes of production, social relationships, and relationships to the environment. He and other scholars have chronicled the impact of the fur trade, farming, mining, and other European interventions on indigenous ecologies and economies, analyzing the forms of dependency and innovation that these interventions have engendered. Another approach to political economy is featured in *Captives and Cousins: Slavery, Kinship and Community in the Southwest Borderlands*, by historian James Brooks. This book analyzes the significance of indigenous forms of captivity, servitude, and adoption as forms of intercultural exchange and the significant changes to these patterns that occurred under colonialism.

Richard White is also known for the important concept of the 'middle ground.' First developed in *The Middle Ground: Indians, Empires, and Republics in the Great Lakes Region*, the concept refers to common patterns of social relationships and meaning that develop in situations of culture contact. Structural anthropologist Marshall Sahlins offers a somewhat similar concept, 'structures of the conjuncture,' in *Historical Metaphors and Mythical Realities: Structure in the Early History of the Sandwich Islands Kingdom*. While ethnohistorians trained in history, like White, are more likely to employ the more informal framework of 'the middle ground,' the approaches of Sahlins and White both offer ways of conceptualizing the hybrid social and cultural forms that develop in the course of intercultural encounters. Sahlins's earlier work on hunting and gathering economies has also been influential on ethnohistories like Robert Brightman's *Grateful Prey: Rock Cree Human-Animal Relationships*, that focus on indigenous economies and cultural logics. Like many anthropologists, Brightman is concerned with delineating the culturally specific concepts of the person that form a central part of indigenous cultural beliefs and practices.

Relations between indigenous peoples and the colonizing state have been a recurring concern of ethnohistory. Bringing the time depth of archaeology to the question, Bruce Trigger, Samuel Wilson, and others have shown that societies with higher degrees of political and economic integration tended to adapt more successfully to European conquest (Rogers and Wilson, 1993). In an extensive set of ethnohistorical works on India, anthropologist Bernard Cohn developed the concept of 'rule by records,' which has inspired numerous other important studies of colonial rule as a form of knowledge and power. Anthropologist Nicholas Dirks' *Castes of*

Mind: Colonialism and the Making of Modern India, for example, historicizes the notion of 'caste,' analyzing it neither as an essential cultural attribute nor as a colonial 'invention,' but as a complex product of indigenous culture, colonialism, and resistance.

One of the most productive arenas of ethnohistorical research is the study of colonial missions and indigenous responses. In *Ambivalent Conquests: Maya and Spaniard in Yucatan*, historian Inga Clendinnen considers missionary work and resistance among the Mayas in the Yucatan peninsula in the sixteenth century. Several influential accounts of conversion and resistance have been offered by anthropologists John and Jean Comaroff, including the latter's *Body of Power, Spirit of Resistance: The Culture and History of an African People*, which analyzes the Church of Zion among the Tshidi as a form of indigenous agency. Among the many studies of religious conversion, resistance, and syncretism in North America, anthropologist Edward Spicer's *Cycles of Conquest: The Impact of Spain, Mexico, and the United States on the Indians of the Southwest* stands out for its scope, while the many works of historian James Axtell are important for contrasting the conversion strategies of the French, English, and Iroquois. Anthropologist Sergei Kan has expanded the range of ethnohistorical comparison in his accounts of Tlingit religion and the impact of the Russian Orthodox Church.

Ethnohistorians have offered important contributions to the understanding of 'ethnogenesis,' or the formation of cultural groups through historical processes of culture contact and mixture. Studies of the Métis of Canada, the Seminoles of the US Southeast, the Garifuna of the Caribbean, and the Yoruba of Africa, among others, have employed this concept productively. *History, Power, and Identity: Ethnogenesis in the Americas*, a collection edited by anthropologist Thomas Hill, considers the development of new cultural identities among indigenous and African American groups, tying the emergence of new identities to global processes of domination, resistance, and exchange.

The linguistic anthropologist Michael Silverstein (1997) has called for, and demonstrated, an ethnohistory of communication, specifically, an ethnohistory of 'languages of encounter.' This ethnohistory concerns the reciprocal transformation of linguistic forms as well as the linguistic communities in which they are embedded. While Silverstein focuses on North America, his argument is relevant anywhere a 'contact community' has formed in which certain kinds of communication (pidginized forms used in trade or diplomacy, for example) occur among people who otherwise belong to distinct language communities. He emphasizes that the 'languages of encounter' used in trade or evangelization (Chinook Jargon, for example) differ significantly from the creole languages that developed on plantations. A monograph exemplifying the ethnohistory of communication, which is also an 'ethno-ethnohistorical' inquiry into indigenous conceptions of history, is *Understanding Tolowa Histories: Western Hegemonies and Native American Responses*, by linguistic anthropologist James Collins. Among other things, this book looks at the relationship between discourses of place and the expropriation of land.

Fortunately, ethnohistory has not been immune to the critique of ethnographic authority and the move to reflexive and collaborative scholarship that is characteristic of anthropology as a whole. An example of the new collaborative ethnohistory is *History is in the Land: Multivocal Tribal Traditions in Arizona's San Pedro Valley*, by archeologists Thomas Ferguson and John Colwell-Chanthaphonh. This ethnohistory is polyphonic, directly incorporating perspectives from representatives of the Tohono-O'odham, Hopi, Zuni, and San Carlos Apache tribal peoples. Reflexive scholarship in ethnohistory often concerns methodological issues, as discussed below, but there is also an important body of reflexive works on the representation of indigenous peoples. These include a critical examination of the role of history and anthropology in producing and circulating such representations. In *Land Filled With Flies: A Political Economy of the Kalahari*, anthropologist Edwin Wilmsen examines the ideological construction of the San (Bushmen) of southwestern Africa as 'primitive' rather than as a displaced and excluded underclass. Similarly, historian Jean O'Brien's *Firsting and Lasting: Writing Indians Out of Existence in New England* analyzes how local histories contributed to the discursive extinction of Indian peoples in New England, an extinction made familiar in the 'last of his tribe' trope. Coming full circle to the beginnings of ethnohistory, anthropologist Jack Campisi's *Tribe On Trial* reflects on his own experience as an expert witness in the Mashpee tribal recognition case, showing the problematic nature of the concept of 'tribe.'

Methodological Issues

A recurrent theme in ethnohistory has been the tension between the methodologies of anthropology and those of history, tensions aptly captured in historian Jennifer Brown's metaphor of 'strange bedfellows, kindred spirits.' Historians have tended to rely on the critical analysis of colonial documents, while anthropologists have put more credence in oral history, sometimes also 'upstreaming,' as William Fenton called it, from contemporary cultural practices to engage in informed speculation about the past. Anthropologists have also emphasized comparative analysis and indigenous conceptions of history, leading Raymond Fogelson to advocate for 'ethno-ethnohistory.'

As Fogelson's ironic reduplication suggests, there has been some tension over the name of the field itself. To what does the 'ethno' in 'ethnohistory' refer? For some scholars it refers substantively to indigenous and other ethnic groups, making ethnohistory the history of these groups. For others, it has methodological significance and refers to an ethnological or ethnographic approach to history. For still others, using an analogy to terms such as 'ethnobotany' or 'ethnoastronomy,' the name of the field refers to ethnic and indigenous groups' conceptions of their own pasts. All of these endeavors have taken place under the rubric of 'ethnohistory,' sometimes in the very same works. In light of these difficulties some scholars, especially those conducting research outside of the Americas, prefer to avoid the term altogether, referring instead to 'historical anthropology' or 'anthropological history.' Yet the term 'ethnohistory' persists, likely because a scholarly tradition and community has grown up around it and because it privileges one discipline less than the alternative terms.

Ethnohistorians typically rely on multiple forms of data, all of which must be utilized with considerable care. These include the following:

- Colonial and other institutional documents, including travel journals, missionary records, administrative records, judicial records, treaty records, memoirs, policy statements, and published narratives. In interpreting these primary documents the ethnohistorian must sort out (as much as possible) official categories, prejudices, and misapprehensions from indigenous beliefs and practices, putting the documents within the larger contexts of colonialism, capitalist expansion, evangelization, and indigenous and nationalist social movements. Like any historian, the ethnohistorian must also be aware of the selective and contingent nature of archiving (Galloway, 2006).
- Written records in indigenous languages. These are particularly significant for literate cultures, but may exist for oral cultures in translated form. They also may exist in written form for indigenous peoples like the Cherokee who developed a written form of their language after contact or those, like the Lakota, who learned to read and write in their native language from missionaries. Regarding his work with Euro-American and Lakota documents, Raymond DeMallie writes, "in a fundamental sense they represent conflicting realities, rooted in radically different epistemologies. The challenge of ethnohistory is to bring these two types of historical data together to construct a fuller picture of the past" (1993: 516).
- Archaeological evidence, which may offer a significantly longer diachronic dimension than historical and ethnographic evidence. Daniel Rogers and Samuel Wilson describe ethnohistorical and archaeological sources as 'complementary investigative routes' into postcontact cultural change (1993: 7), with archaeology particularly valuable for information on demography and geography, settlement and land use patterns, travel routes, and ethnic relationships.
- Collections, including maps, illustrations, photographs, and artifacts. Much ethnohistorical research occurs in museum collections, and these must be understood as subject to the same processes of selection, appropriation, misinterpretation, reinterpretation, and loss as historical archives (Galloway, 2006).
- Oral traditions, including oral histories, genealogies, folk tales, and place names. Like institutional documents, oral traditions must be contextualized within contemporary social structures and projects. As Africanist Jan Vansina has emphasized, oral traditions represent the viewpoints of particularly situated actors and may shift over time.
- Ethnographic research, including systematic participant observation, aimed at finding traces of, or attitudes toward, the past in the present. For some, ethnographic fieldwork is what differentiates the ethnohistorian from the conventional historian; for others, it is a more general 'anthropological orientation' toward culture (Cohn, 1968). In either case, the results of ethnographic research must always be understood as situated within a particular moment and particular ethnographic encounter. Ethnographic research produces its own archive, and ethnohistorians typically rely on the ethnographic research of previous generations as well as their own. A fine example of an ethnohistorian working through an ethnographic archive are Jennifer Brown's ethnohistorical works on the Berens River Ojibwa, using materials from A.I. Hallowell's fieldwork there in the 1930s.

Given this range of data, and the different forms of training that historians, anthropologists, and archeologists bring to the their research, it is not surprising that methodological issues arise frequently in the field of ethnohistory. Among these issues are:

- The problem of what Samuel Wilson calls 'mixed epistemologies,' with history, cultural anthropology, and archaeology each relying on different modes of explanation and validation. Wilson contrasts a neoevolutionary approach focused on long-term and large-scale processes with historical approaches more concerned with small-scale, shorter term processes. He compares this to the contrast Braudel draws between process-oriented history (the *longue durée*) and event-oriented history. This difference in scale poses "the challenge of integrating macroprocesses and microprocesses of culture change into a coherent analysis" (Rogers and Wilson, 1993: 21; 23).
- The contingency and partiality (in both senses) of documentary sources. Patricia Galloway (2006) notes the significance of material not noticed or not available to European observers; material misinterpreted by European observers; and material that has been lost. This situation requires a methodology of suspicion and convergence, in which documentary sources are viewed as compromised and explanations may be strengthened by the convergence of several lines of evidence, including archaeological, linguistic, and geographic evidence.
- The role of oral history, tribal history, and/or living memory as a supplement or corrective to written documents, which largely represent the perspective of colonial institutions. Jan Vansina (1985) has offered a methodology for the interpretation of oral history, including a way of distinguishing among ecological time, genealogical time, the sociological (recurring) calendar, and the time of extraordinary events. Anthropologists such as Julie Cruikshank have emphasized the importance of indigenous orientations to time, such as prophecy, that contrast sharply with linear time.
- The ability of ethnohistory to deal adequately with plural interpretations, including tensions or contradictions between (and within) archaeological, documentary, and oral histories. Historian Fred Hoxie has argued eloquently for an ethnohistory that aims "to describe community lives in their own terms," producing "stories that allow for an open vision—one that is coherent but attends to several layers of meaning and many co-existing interpretations" (1997: 606; 610). Anthropologist Raymond Fogelson agrees with this vision, while enumerating the difficulties posed to the ethnohistorian by "bi- or multicultural frames of reference …, by different modes of discourse, by documentation that cannot always be limited to written manuscripts, and by recognition of different conceptions of reality" (1989: 141).

- The role of theoretical concepts in the field, for example Michel Foucault's notion of bodily practice (as employed by Jean Comaroff, for instance), the structuralism of Marshall Sahlins, or Eric Wolf's use of world systems theory. Many ethnohistorians of a historical bent prefer highly particularistic analyses and express suspicion of the comparative analysis and theoretical framing characteristic of many of the ethnohistories produced by anthropologists. Fred Hoxie (1997), for example, counterposes the pluralistic ethnohistory he advocates to Eric Wolf's attempt, in *Europe and the People Without History*, to locate tribal histories within a global system. Many anthropologists, on the other hand, have lauded and tried to emulate Wolf's systematic vision.

Anthropologist Bernard Cohn (1980) offered a wise mediation between what he calls 'anthropologyland' and 'historyland,' noting that history and anthropology have a common subject matter, 'otherness'; a common project, translation; and a common interest in studying social transformation. He called for an 'anthropological history' in which the colonial situation is a privileged, unified analytic field for the study of the construction and representation of culture – indigenous culture, colonial culture, and postcolonial culture alike. Such an endeavor, he insisted, requires a working experience of both the field and the archive. This endeavor also needs to be highly reflexive, he insisted, and attuned to the ways in which current scholarly concepts are often remnants of colonial forms of rule. These include such basic concepts as 'culture,' 'race,' 'tribe,' and 'ethnic group.' Even the division between history and anthropology is a remnant of colonial notions of the timeless nature of 'primitive' people.

Emerging Directions in Ethnohistorical Research

In a provocative review of 'The State of Ethnohistory' Shepard Krech asked, "should ethnohistory cease to exist?" (1991: 346). Given the convergence or interpenetration of anthropological and historical methods since 1980 – in which scholars in both disciplines have been influenced by the Annales School of social history, world systems theory, the interpretive turn, and practice theory – is the field a legitimate one? Or does ethnohistory reproduce a colonialist practice in which 'history' refers to knowledge of the past in Western, literate societies, while 'ethnohistory' refers to all who are outside of that category? Was Eric Wolf correct in arguing that "The more ethnohistory we know, the more clearly 'their' history and 'our' history emerge as part of the same history" (19)?

The importance of this question has only increased in the decades since 1991 as the field of indigenous studies has grown in significance. Some argue that the conditions that gave rise to the interdisciplinary formation of ethnohistory have changed enough that it is no longer a viable field. Others, such as Jennifer Brown (1991), argue that ethnohistory has brought historians, anthropologists, archeologists, and indigenous people together in a fruitful enterprise that will always be changing but should not be scuttled.

Krech's position can be reversed, as Patricia Galloway (2006) does in *Practicing Ethnohistory: Mining Archives, Hearing Testimony, Constructing Narrative*. "If we are to go on practicing 'ethnohistory' at all," she writes, "and if by that we mean trying to write something we call history on the basis of incomplete and biased testimonies that we only partially understand, then we need to recognize that there is really no history that is *not* ethnohistory for a species whose daily personal communication is a clash of idiolects" (27). Perhaps the same argument can be made for anthropology: is there really no anthropology that is not ethnohistory for a species that lives in time?

The future of ethnohistory most likely lies between these two extremes. Ethnohistory will probably neither disappear nor subsume all of history, not in the foreseeable future anyway. It will likely continue to develop its strengths: being a 'middle ground,' as Brown (1991) puts it, between history and anthropology, between archive and field, between colonialist and indigenous perspectives. It will be strongest when practiced by scholars with experience in both the archive and the field, and when indigenous and nonindigenous scholars are collaborating side by side. It will be strongest when practiced by scholars with expertise in indigenous languages. It will be strongest when it is broadly comparative, and when local practices and meanings are tied to global processes and power relations. It will be strongest when multiple and sometimes conflicting, worldviews and perspectives are taken into account. And it will be strongest when it proceeds in a reflexive manner, subjecting its own practices to critique and revision. Such scholarship may not be called ethnohistory in the future, but it will reject the artificial, colonialist distinction between culture, prehistory, and history that is institutionalized in the academy's distinction between the fields of anthropology, archaeology, and history.

See also: Anthropology and History; Colonialism, Anthropology of; Colonization and Colonialism, History of; Historical Archaeology; Subaltern History.

Bibliography

Brightman, R., 1993. Grateful Prey: Rock Cree Human-animal Relationships. University of California Press, Berkeley, CA.

Brooks, J.F., 2002. Captives and Cousins: Slavery, Kinship and Community in the Southwest Borderlands. University of North Carolina Press, Chapel Hill, NC.

Brown, Jennifer S.H., 1991. Ethnohistorians: strange bedfellows, kindred spirits. Ethnohistory 38 (2), 113–123.

Campisi, J., 1991. The Mashpee Indians: Tribe on Trial. Syracuse University Press, Syracuse, NY.

Clendinnen, I., 2003. Ambivalent Conquests: Maya and Spaniard in Yucatan, 1517–1570, second ed. Cambridge University Press, Cambridge.

Cohn, B.S., 1968. Ethnohistory. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences, vol. 6. The Macmillan Company and The Free Press, New York, pp. 441–448.

Cohn, B.S., 1980. History and anthropology: the state of play. Comparative Studies in Society and History 22 (2), 198–221.

Collins, J., 1998. Understanding Tolowa Histories: Western Hegemonies and Native American Responses. Routledge, New York.

Comaroff, J., 1985. Body of Power, Spirits of Resistance: The Culture and History of a South African People. University of Chicago Press, Chicago, IL.

DeMallie, R.J., 1993. "These have no ears": narrative and the ethnohistorical method. Ethnohistory 40 (4), 515–538.

Dirks, N.B., 2001. Castes of Mind: Colonialism and the Making of Modern India. Princeton University Press, Princeton, NJ.

Etienne, M., Leacock, E., 1980. Women and Colonization: Anthropological Perspectives. Praeger Publishers, New York.
Ferguson, T.J., Colwell-Chanthaphonh, J.S., 2006. History Is in the Land: Multivocal Tribal Traditions in Arizona's San Pedro Valley. University of Arizona Press, Tucson, AZ.
Fogelson, R.D., 1988. The ethnohistory of events and Nonevents. Ethnohistory 36 (2), 133–147.
Galloway, P., 2006. Practicing Ethnohistory: Mining Archives, Hearing Testimony, Constructing Narratives. University of Nebraska Press, Lincoln, NE.
Harkin, M.E. (Ed.), 2004. Reassessing Revitalization Movements: Perspectives from North America and the Pacific Islands. University of Nebraska Press, Norman, OK.
Hill, J.E. (Ed.), 1996. History, Power, and Identity: Ethnogenesis in the Americas, 1492–1992. University of Iowa Press, Iowa City, IA.
Hoxie, F., 1997. Ethnohistory for a tribal world. Ethnohistory 44 (4), 595–615.
Krech III, S., 1999. The state of ethnohistory. Annual Review of Anthropology 20, 345–375.
McMillen, C., 2007. Making Indian Law: The Hualapai Land Case and the Birth of Ethnohistory. Yale University Press, New Haven, CT.
O'Brien, J., 2010. Firsting and Lasting: Writing Indians Out of Existence in New England. University of Minnesota Press, Minneapolis, MN.
Rogers, J.D., Wilson, S.M. (Eds.), 1993. Ethnohistory and Archaeology: Approaches to Postcontact Change in the Americas. Plenum Press, New York and London.
Sahlins, M., 1981. Historical Metaphors and Mythical Realities: Structure and Early History of the Sandwich Islands Kingdom. University of Michigan Press, Ann Arbor, MI.
Silverstein, M., 1997. Encountering language and languages of encounter in North american ethnohistory. Journal of Linguistic Anthropology 6 (2), 126–144.
Spicer, E., 1962. Cycles of Conquest: The Impact of Spain, Mexico, and the United States on the Indians of the Southwest, 1533–1960. University of Arizona Press, Tucson, AZ.
Vansina, J., 1985. Oral Tradition as History. University of Wisconsin Press, Madison, WI.
Wallace, A.F.C., 1969. The Death and Rebirth of the Seneca. Alfred A. Knopf, New York.
White, R., 1991. The Middle Ground: Indians, Empires, and Republics in the Great Lakes Region, 1650–1815. Cambridge University Press, New York.
Wilmsen, E.N., 1989. A Political Economy of the Kalahari. University of Chicago Press, Chicago, IL.
Wolf, E., 1982. Europe and the People without History. University of California Press, Berkeley, CA.

Relevant Websites

http://www.ethnohistory.org/ – American Society for Ethnohistory.
http://cssh.lsa.umich.edu/ – Comparative Studies in Society and History.
http://www.naisa.org/ – Native American and Indigenous Studies Association.

Ethnology

Gisela Welz, Goethe University, Frankfurt, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the historiography of the social sciences, ethnology is often taken to represent an early stage in the development of the anthropological disciplines. In the course of the twentieth century, ethnology was eclipsed by Social Cultural Anthropology. In many European countries, however, the term ethnology is used persistently. Here, a plethora of national ethnologies had evolved in the nineteenth century. Even today, this is reflected in the coexistence of divergent research programs and academic specializations called Ethnology. Addressing the historical transformation of ethnology allows us to reflect on the fact that scholarship in anthropology is far from homogeneous across national boundaries.

Ethnology started in the eighteenth century as a systematic attempt to acquire and compare information about the cultures of peoples and nations throughout the world. The first section of this chapter is dedicated to an understanding of Ethnology as a precursor of Cultural and Social Anthropology. However, transformations in the anthropological disciplines during the twentieth century led to a redefinition of ethnology as a specific approach within anthropology that is dedicated to cross-culturally comparative analyses. The second section of the chapter takes up this issue and discusses the historically contingent relationships among ethnology, ethnography, and anthropology. In many European countries, where national ethnologies evolved in the nineteenth century, there was a divergent development, which is briefly touched on in the third part of the chapter. The concluding section of the chapter addresses the role of ethnology within the context of global anthropologies in a postcolonial world.

Ethnology as the Formative Period of Anthropology

In the historiography of the social sciences, ethnology is taken to represent an early stage in the development of the anthropological disciplines. Academic circles at the German universities of Goettingen and Halle first used the term 'ethnology' in the late-eighteenth century. Coined as a derivation of the Greek word 'ethnos,' meaning 'a people,' the term 'ethnology' in its most general meaning indicates a scholarly interest in how aggregations of human beings are distinct from each other in terms of material culture, language, religion, moral ideas, or social institutions. Early developments in ethnology also included speculative theories about assumed interrelationships among the differences in cultural and biological groups. In the nineteenth century, ethnology had emerged as a line of inquiry into the culture history of humankind, initially supported and brought to public attention by the founding of learned societies and museums dedicated to ethnology in both Europe and the United States. This in turn prepared the ground for the inclusion of ethnology as an academic discipline in the curricula of universities, first in England, then in Germany and the United States.

In 1888, German-born natural scientist Franz Boas, who during the following decades would initiate and establish cultural anthropology as an academic discipline in the United States, lectured on 'The Aims of Ethnology' before an audience of well-educated German immigrants who were members of a civic society in New York City. He identified himself as an ethnologist and defined ethnology's agenda as follows: "The task of ethnology is the study of the total range of phenomena of social life. Language, customs, migrations, bodily characteristics are subjects of our studies. Thus its very first and most immediate object is the study of the history of mankind; not that of civilized nations alone, but that of the whole of mankind, from its earliest traces found in the deposits of the ice age, up to modern times" (Boas, 1948, 628).

Coming to ethnology in the late-nineteenth century, Boas found himself at the cusp of an important paradigm shift that changed the face of this still young discipline. Until this point, ethnology had been driven by the impetus to discover the hidden laws of human cultural development. Scholars hoped to do this by aggregating data about as many cultures as possible – both past and present – and by tabulating and classifying them. Findings were supplied by archaeology, itself an emerging discipline, or came from travel accounts, mission reports, and increasingly also from more sophisticated ethnographic surveys of indigenous groups in Africa, Asia, and the Americas. Most of the nineteenth century ethnologists attributed the manifestly observable cultural diversity of humankind to what they considered to be different levels of cultural development: The more 'primitive' groups were classified as being far down on the lowest rungs of the steep ladder of ascent to civilization. At the same time, these 'savages' personified, in the gaze of their more progressed contemporaries, the earliest evolutionary stages of humanity. Therefore, this approach, personified by, among others, Edward B. Tylor in England (1832–1917), Lewis Henry Morgan (1818–81) in America, and to some extent by Adolf Bastian (1826–1905) in Germany, was labeled – not by its protagonists, but by later opponents and historiographers – 'evolutionism.'

As increasingly more exact and detailed information about non-European cultures was made available, many evolutionist generalizations became untenable. At the turn of to the twentieth century and already anticipated by Boas in the lecture cited above, a new theory took over. It shifted the emphasis away from determining universal rules of cultural development; instead, evidence of migration and of contact between

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12066-5

groups and the resulting exchanges and transfers of cultural elements were then taken to be the key to the understanding of human history and cultural diversity. The underlying assumption was that human culture emerged from a single or a small number of origins, and the ensuing dispersal of cultural elements came about through migration and diffusion – hence the label of 'diffusionism' that was attributed to this new school. It had its strongest and methodologically innovative proponents among museum ethnologists such as Fritz Graebner (1877–1934) in Germany, who used similarities in the shapes and makeups of pottery vessels, weaponry, ornaments, and other artifacts as the bases to map geographical areas where cultures were assumed to have a shared origin. This enabled German-speaking ethnologists, among them Leo Frobenius (1873–1939) and Pater Wilhelm Schmidt (1868–1954), at the beginning of the twentieth century, to lay out a worldwide system of so-called Kulturkreise, areas of shared cultural traits. With their evolutionist predecessors, representatives of diffusionism shared a tendency toward sweeping generalizations and a disregard for the cultural meanings associated with the artifacts under scrutiny. They were not interested in the social integration of groups or in the cultural context of the customs they used as the bases for their theoretical speculations; rather, they treated them as isolates, much as if they were specimens of material culture in a collection's glass case. Also, while insisting that their endeavors were historical in nature, they established a frozen chronology of past cultures rather than dating events and gaining insights into historical processes.

Franz Boas (1858–1942), who had been trained by Bastian in Germany and was appointed a professor at Columbia University in New York City in 1899, rejected not only evolutionism – which had an American stronghold in the Smithsonian Institution's Bureau of American Ethnology in Washington DC – but German diffusionism as well. He founded a new approach, 'the historical method,' which could perhaps best be described as a moderate type of diffusionist ethnology. He and his students, however, shied away from projective theory and, indeed, from generalizations on a grand scale. They built up an enormous database from fieldwork among contemporary North American indigenous populations; their concept of culture areas paid careful attention to proven historical relationships among native American groups and also integrated environmental factors. With their emphasis on empirical research, the Boasians, students of Boas who became university professors in their own right at the most prestigious American universities, ended the era of speculative ethnology. In Great Britain, evolutionism proved to be more tenuous and was not superseded by diffusionist ideas until the 1920s.

Anthropology, Ethnology, Ethnography

In the late nineteenth and well into the twentieth century, anthropology became increasingly prevalent as an umbrella term to refer to a broad domain of scholarly interests in the biological, cognitive, and cultural dispositions of the human species, including physical anthropology, linguistics, and archaeology along with ethnology. In the course of these developments, ethnology was integrated as a subfield of the growing concern of anthropology. In the United States, the term ethnology continued to be used, but its meaning contracted to denote anthropological approaches that compared data concerning a number of groups or societies cross-culturally. Cross-cultural comparison had already been an indispensable tool for the historical reconstruction of cultural evolution as well as of cultural diffusion in the nineteenth century, both in the United States and in academic institutions in Europe, but it was not until the advent of the functionalist school that cross-cultural comparison had emerged as the principal methodology of British social anthropology. Relinquishing the generalizing thrust of earlier social anthropology, later approaches used comparison to inquire into the interplay of variable and invariable factors shaping social relations in a given culture by contrasting it with others like it, often in a controlled sample of regionally equivalent cultural groupings. Meanwhile, in the United States, the statistical analysis of nonmaterial cultural data had been further refined. The Human Relations Area Files at Yale University, a project that, since 1937, had abstracted and systematized the information on cultures around the world from ethnographic reports, developed into a huge cross-cultural data repository to be used for quantitative analyses and surveys.

By the end of World War II, the meaning of 'ethnology' had contracted to define but one of the three scholarly practices that form the anthropological enterprise. Ethnography was understood as data gathering in a single society, usually in a spatially and temporally bounded situation, such as when the fieldworker spends a year as a participant-observer in a local community and later writes up the findings in a stand-alone, case-study-type monograph (also called ethnography). Ethnology, by contrast, utilized the information collected in a number of ethnographies in order to engage in systematic cross-cultural comparisons between two or more societies. Ethnology in this sense was considered both distinct from and dependent on ethnography as the primary data gathering by fieldwork conducted in individual societies, but it also included, at times, information garnered from historical research. In a report on the status of anthropology in higher education submitted to UNESCO in 1954, French anthropologist Claude Lévi-Strauss defined ethnography, ethnology, and anthropology as three consecutive stages in the production of anthropological knowledge, each building up on the results of the other. Ethnology, to his mind, effected a synthesis by the systematic comparison of ethnographic data across cultures, but only anthropology, as the third and final stage, aspired to theory building with a generalizing impetus, attempting to come to universal statements about the nature of humanity (Lévi-Strauss, 1975).

In cultural anthropology in 1960s America, new developments predicated a significant shift away from comparative work. Most notably, Clifford Geertz's championing of 'thick description' suggested "not to generalize across cases but to generalize within them" (Geertz, 1973, 26) by in-depth immersion in one culture and ethnographic representation of one individual case study. During the 1980s, criticism of cross-cultural comparativism became increasingly vocal. In the second half of the twentieth century, many cultural anthropologists avoided cross-cultural perspectives altogether.

Instead, in-depth ethnographic accounts of particular cultural settings were considered the best reflection of anthropology's critical abilities within the concert of the humanities. In a contradictory movement, ethnography also was cast into doubt by a movement of textual critics and postcolonial historiographers of anthropology, but it emerged as the signature knowledge practice of cultural and social anthropology by the late-twentieth century.

Two important scholarly journals carry the designation 'ethnology' in their names: Ethnology, initiated by George Murdock in 1962, and the American Ethnologist, founded in 1972 at the suggestion of the American Anthropological Association 'to meet the growing demand to publish specifically socio-cultural anthropological articles' and also serving as the journal of the much older American Ethnological Society (founded in 1842). Indicative of the state of ethnology in English-language anthropology, these journals today present themselves to audiences and contributors as important venues for ethnographically produced findings in cultural and social anthropology. Ethnology is often taken today as an older and somewhat outdated synonym of cultural anthropology with which it shares the designation of study of present-day and recent cultures. Therefore, although ethnography as a mode of producing anthropological knowledge became unexpectedly central, especially in the United States and in those parts of the international arena where American-derived cultural anthropology is hegemonic, the term ethnology faded into the background. There were attempts to reassess comparative methodologies in anthropology, and new approaches were introduced to strengthen what is considered the inherently critical potential of comparison (see Gingrich and Fox, 2002), but 'ethnology' is no longer universally employed as a term to denote a specific approach in the production of anthropological knowledge. This may also have to do with the etymological connotations of the term 'ethnology' itself, because it contains the notion of 'ethnos,' of a bounded and homogeneous peoplehood as the basic unit of social life and of anthropological research. Many of the new research concerns of anthropology – migration and mobility, media and computer-mediated communication, statehood and supranational governance, commodities and consumption, science and technology – today entail a turning away from the idea of 'ethnos.' As a consequence of these transformations, anthropologists have abandoned established assumptions about how culture relates to territory. The primordialist notion of ethnos and the conventional culture concept of anthropology intimately bound up with it have been widely criticized, deconstructed, and abandoned.

The Multiplicity of Ethnologies in Europe

Today more than ever, Tomas Gerholm's and Ulf Hannerz's observation holds: "between different national and continental traditions as well as over time, there has been little consensus over the meaning of 'anthropology', 'ethnology' and 'ethnography'"(1982: 21). In Europe, the disciplinary landscape of anthropology is far from uniform. The names of academic disciplines, their boundaries, and their designated domains of scholarly inquiry differ from one country to the next, sometimes even among individual universities and research institutions. A closer look, however, reveals some remarkably consistent patterns. A split emerged in the nineteenth century, resulting in two separate anthropological disciplines coexisting in academic teaching and research.

On the one hand, national ethnologies emerged in individual countries, which "contributed to the cultural homogenization process of each nation" (Frykman, 2012, 578), embodied in the "the study of folklore and folk music, customs and costumes, housing and handicraft as they existed in peasant society" (Gerholm and Hannerz, 1982, 22). On the other hand, academic interest in the cultures of so-called primitive peoples and premodern peasant societies coalesced in the formation of research programs that during the twentieth century grew into a more general, often comparison-oriented anthropology. The names given to these two branches of the anthropological enterprise differed according to national languages from country to country. Because the bifurcation appears to have been articulated most strongly in the German-speaking countries, however, it has been labeled the 'Volkskunde/Voelkerkunde split' (Gerholm and Hannerz, 1982), even in English-language histories of the disciplines. Traditionally, the national ethnologies grouped under the Volkskunde heading were concerned with the historical investigation of and the cultural construction of a national heritage. Conversely, those disciplines belonging to the Voelkerkunde category focused on non-European 'primitive' peoples. The latter scholarly line of inquiry became dominant in societies marked by colonial expansionism, whereas the former interest led to the establishment of folklore and folklife studies dedicated to the philology and history of national and regional cultures in most of Eastern, Central, and Northern Europe. Although both "developed within the same traditions of cultural theory, and the early pioneers read much of the same classics, their position in academia came to be very different. European ethnology and folklore was defined as belonging to the humanities with links to history, literature, art history and language" (Loefgren, 2008, 119), whereas the comparative study of extra-European cultures was, more often than not, considered to be a social science. Since the 1960s, both sections of the anthropological enterprise in continental Europe have undergone marked modernization, in some countries leading to a rapprochement of the disciplines. The designation of European Ethnology had originally been invented in 1955 at a conference in Arnhem, during one of the first postwar encounters of scholars from various European countries. Those who championed this designation wanted it to be strongly inclusive of the differing orientations of national ethnologies and folklore studies by emphasizing a common European perspective. The 1964 founding of the International Society of Ethnology and Folklore initially emphasized differences rather than commonalities. Through the years it has proven to function as a Europe-wide organization of scholars who do research in Europe within the context of national ethnologies and their modernized successor disciplines. The Europe-wide journal 'Ethnologia Europaea' also grew from the concern with European ethnology that was most pronounced in the Scandinavian countries. There, ethnology originally was a museum-based research program concerned with the material culture of non-European, small-scale societies. However, it had transformed

itself into an internationally acclaimed social anthropology by the 1970s. The label ethnology was then appropriated by the successors of national folklore studies and prefixed with 'European.' "As European ethnologists in the 1960s oriented their research towards modern industrial society rather than the vanishing peasant traditions, the basic concept of folk culture was exchanged for the study of the everyday and 'the ordinary people'"(Loefgren, 2008, 126). The popular culture, cultural creativity, and consumption as a symbolically expressive practice took the place of tradition, customs, and heritage on the disciplinary agenda.

In the German-speaking countries of Europe, the national ethnologies had, after World War II, initially rejected the new label of European Ethnology, continuing to identify themselves instead as Volkskunde. During the modernization process of the disciplines in the 1970s, the term ethnology was adopted by Voelkerkunde (Haller, 2012), while many university departments and programs previously called Volkskunde acquired new names, such as cultural studies, empirical cultural research, European ethnology, and cultural anthropology. Their name changes signified a shift toward the problem-oriented study of contemporary German, Swiss, or Austrian society informed by new developments in social thought, especially the critical theory of the Frankfurt School, but also with advances in social research (Bausinger, 1986). In addition, early British cultural studies of the Birmingham school proved a strong inspiration for the new focus on everyday life, consumption, and expressive cultures (Warneken, 2006). In Germany, the striving for a new rationale for the discipline had been animated by the need to come clear of any associations with the Nazi past that had haunted the discipline since the end of World War II. The rejection of the old Volkskunde designation signaled a theoretical paradigm shift as much it did a political turning point, leaving behind the politically conservative, often at least latently nationalist, research agenda (Bendix, 1997).

In many other European countries, even those not characterized by a marked split between ethnologists working abroad and those working in their own societies, an anthropological turn could be observed by the 1980s, aligning academic programs and research orientations more closely with Anglo-American cultural and social anthropology. An exception is France, where 'ethnologie' continues to be a term valid for both the intra- and extra-European research endeavors conducted by anthropologists (see Segalen, 2008). There, since the 1970s, the Ethnologie Contemporaine de la France represented a new orientation of anthropological inquiry into French society with a strong ethnographic flavor. Its focus on urban and rural communities is not unlike the research interest that a group of social anthropologists and sociologists in the United Kingdom started to pursue under the heading of Anthropology of Britain, generating community studies that were acclaimed as important case studies of social change in twentieth-century England, Scotland, and Wales (see Macdonald et al., 2005).

In Eastern European countries, before the transition of 1989, the division of labor among disciplines had been organized according to the Marxian classification of academic endeavors. The discipline was often called ethnography, rather than ethnology, and was considered an auxiliary to history. Ethnography in the socialist countries, however, had sometimes included scholars working both within and outside of Europe. Yet the flavor of ethnology in many countries was strongly nationalist, its research celebratory of working-class and rural folk cultures. Even after the post-1989 transformations, the "traditional canon of Volkskunde lingered on longer in East and Central Europe" (Kiliánová, 2012, 104) than it had in Germany and the Scandinavian countries, leading the Polish anthropologist Michal Buchowski to demand, even in 2012, that "the myth of nationalist ethnology in (post)socialist Central and East Europe must be dispelled" (Buchowski, 2012, 68). In many Eastern European countries, however, the homegrown national ethnologies and ethnographic research programs were resolutely 'anthropologized' (Rihtman-Augustin and Capo Zmegac, 2004), adopting an orientation toward American cultural anthropology and/or the British-derived social anthropology.

Today, social anthropologists claim to be the most visible and numerous group of anthropologists in Europe. In 1989, the European Association of Social Anthropologists (EASA) was conceived by its founders as a counterbalance to the American hegemony in anthropology. It has evolved into one of the main arenas for young scholars working in the interface of ethnology, anthropology, and sociology, many of them focusing on contemporary European societies rather than on non-European field sites.

Many of the contributors to EASA's events and publication series come from southern Europe. There, new departments of anthropology have emerged, and they combine research interests in their home countries with the most advanced theoretical developments in international anthropology. In Spain, Portugal, and Greece, nationalist ethnologies and folklore studies were successfully superceded and discarded in favor of a social-science-oriented anthropology that tackles problems of contemporary Mediterranean societies, such as urban growth, social change, and immigration. Southern European anthropologists have also entered into a productive dialogue with an older Europeanist specialization that had developed in American cultural anthropology and British social anthropology. In the 1970s, Anglo-American social and cultural anthropologists had already conducted community studies in Mediterranean countries (Boissevain, 1994).

A Future for Ethnology?

Historically, ethnology is a Europe-derived academic project. In many European countries, ethnology not only continues to be widely used as a disciplinary designation, but also the term has been adopted, and even reinvented, to suggest new directions taken in the anthropological investigation of contemporary European societies. European ethnologies today, however, constitute a minor academic tradition within the broader arena of global anthropologies. The majority of their publications are not in English but in their respective countries' languages. As a consequence, they have less visibility than those of their colleagues who participate in English-language dominated, transnational discourse. Meanwhile, most markedly in the United States but also in other Anglo-Saxon countries, ethnology has receded and been eclipsed by anthropology. The impact of this development extends into other countries whose

academic institutions employ anthropologists who have acquired their academic training and university degrees in North American or British universities or at institutions of learning influenced by them. Yet the beginning of the twenty-first century also sees anthropological studies in new nations on the rise, such as Brazil, Mexico, and India, for some of whom the ethnological tradition of inquiry into the cultures of indigenous groups is a problematic legacy of the colonial past but one that they have long since left behind and instead, have successfully developed their own anthropological institutions, theoretical discourses, and research agendas (Lins Ribeiro and Escobar, 2006). These postcolonial anthropologies of the global south, for the most part, have no stake in ethnology or in its history (Argyrou, 2002). Inquiring into ethnology's scope in the twenty-first century allows us to reflect on the divergent disciplinary developments that make up the academic landscape of global anthropologies. Anthropology, like any other scholarly discipline, is not a universal endeavor but a historically contingent project infused with hegemonic power and national interests. It is also able to generate innovative energy and political critique.

See also: Anthropology and History; Anthropology at Home; Center–Periphery Relationships; Europe: Sociocultural Overview; Folklore; Global South.

Bibliography

Argyrou, Vassos, 2002. Anthropology and the Will to Meaning. A Postcolonial Critique. Pluto Press, London.

Bausinger, Hermann, 1986. Volkskultur in der technischen Welt. Campus-Verlag, Frankfurt/New York. Translated as: Bausinger, Hermann, 1990. Folk Culture in a World of Technology (Folklore Studies in Translation). University of Indiana Press, Bloomington.

Bendix, Regina, 1997. In Search of Authenticity. The Formation of Folklore Studies. University of Wisconsin Press, Madison.

Boas, Franz, 1948. Race, Language and Culture. Macmillan, New York.

Boissevain, Jeremy, 1994. Towards an anthropology of European communities. In: Goddard, Victoria A., Llobera, Josep R., Shore, Cris (Eds.), The Anthropology of Europe. Identities and Boundaries in Conflict. Berg, Oxford/Washington, DC, pp. 41–56.

Buchowski, Michal, 2012. Anthropology in postsocialist Europe. In: Kockel, Ullrich, Nic Craith, Máiréd, Frykman, Jonas (Eds.), A Companion to the Anthropology of Europe. Wiley-Blackwell, Malden, Oxford, Chichester, pp. 68–87.

Frykman, Jonas, 2012. A tale of two disciplines: European ethnology and the anthropology of Europe. In: Kockel, Ullrich, Nic Craith, Máiréd, Frykman, Jonas (Eds.), A Companion to the Anthropology of Europe. Wiley-Blackwell, Malden, Oxford, Chichester, pp. 572–589.

Geertz, Clifford, 1973. The Interpretation of Cultures. Selected Essays. Basic Books, New York.

Gerholm, Tomas, Hannerz, Ulf, 1982. Introduction: the shaping of national anthropologies. Ethnos 47 (I–II), 5–35.

Gingrich, A., Fox, R.G. (Eds.), 2002. Anthropology, by Comparison. Routledge, London/New York.

Haller, Dieter, 2012. Die Suche nach dem Fremden. Geschichte der Ethnologie in der Bundesrepublik. Campus-Verlag, Frankfurt.

Kiliánová, Gabriela, 2012. Mitteleuropean ethnology in transition. In: Kockel, Ullrich, Nic Craith, Máiréd, Frykman, Jonas (Eds.), A Companion to the Anthropology of Europe. Wiley-Blackwell, Malden, Oxford, Chichester, pp. 103–121.

Lévi-Strauss, Claude, 1975. Anthropologie. Diogène 90, 3–30.

Lins Ribeiro, Gustavo, Escobar, Arturo (Eds.), 2006. World Anthropologies: Disciplinary Transformations within Systems of Power. Berg, New York.

Loefgren, Orvar, 2008. When small is beautiful? The transformations of Swedish ethnology. In: Nic Craith, Máiréad, Kockel, Ullrich, Reinhard, Johler (Eds.), Everyday Culture in Europe: Approaches and Methodologies. Ashgate, Aldershot, pp. 119–132.

Macdonald, Sharon, Edwards, Jeanette, Savage, Mike (Eds.), November, 2005. Community, continuity, and change in the study of Britain. A Festschrift for Ronnie Frankenberg. Sociological Review 25 (4).

Rihtman-Augustin, Dunja, Capo Zmegac, Jasna, 2004. Ethnology, Myth, and Politics: Anthropologizing Croation Ethnology. Ashgate, Aldershot.

Segalen, Martine (Ed.), 2008. L'Europe et ses Ethnologies. Ethnologie Française 38 (4).

Warneken, Bernd Juergen, 2006. Die Ethnographie popularer Kulturen. Eine Einfuehrung. Boehlau, Wien, Koeln, Weimar.

Relevant Websites

http://www.aesonline.org/ – American Ethnological Society. American Ethnologist (journal).

http://ethnology.pitt.edu/ojs/index.php/Ethnology/about – Ethnology. An International Journal of Cultural and Social Anthropology.

www.easaonline.org/about/founmemb.shtml – European Association of Social Anthropologists.

http://www.siefhome.org/ – Société Internationale d'Ethnologie et Folklore.

Ethnomethodology, General

Steven E Clayman, University of California, Los Angeles, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 7, pp. 4865–4870, © 2001, Elsevier Ltd.

Abstract

Ethnomethodology is a mode of inquiry devoted to studying the practical methods of common sense reasoning used by members of society in the conduct of everyday life. It was developed by Harold Garfinkel in an effort to address certain fundamental problems posed by Talcott Parsons' theory of action. Parsons' motivational approach to the problem of order (emphasizing internalized values) implicates an analytically prior cognitive problem of order involving the process by which concrete actions are produced and rendered intelligible in relation to their circumstances. Garfinkel's classic studies were designed to expose the methods of common sense reasoning that actors employ in this process. Contemporary research initiatives within ethnomethodology have revitalized a wide range of social science subfields, including the study of language and social interaction, the inner workings of bureaucratic and people-processing institutions, and the construction of formal scientific knowledge.

Ethnomethodology is a mode of inquiry devoted to the study of the practical methods of common sense reasoning used by members of society in the conduct of everyday life. The significance of this seemingly mundane subject matter resides in the fact that practical reasoning is what enables societal members to make sense of the circumstances in which they find themselves, find ways of acting in relation to those circumstances, and thereby contribute to the production and maintenance of an intelligible social world. Ethnomethodology, as the study of such reasoning practices, is thus concerned with the very foundations of social order.

It should be clear from this brief definition that ethnomethodology is not primarily a theory of social life, nor is it a methodology for the study thereof. It is in the first instance a discipline of inquiry devoted to a distinctive order of phenomena, one that is sometimes understood to be orthogonal to the concerns of mainstream social science. However, these phenomena have important theoretical, methodological, and substantive ramifications, and in the course of elucidating them ethnomethodology has had a major impact across a range of social science disciplines. Although it originated within sociology, ethnomethodology's sphere of influence extends to anthropology, cognitive science, communication, linguistics, psychology, and the philosophy of the social sciences.

Although a few studies appeared in the late 1950s and early 1960s, the nascent approach crystallized and attracted widespread attention in 1967 with the publication of Harold Garfinkel's *Studies in Ethnomethodology*. This was a time of great intellectual ferment in the social sciences, and ethnomethodology contributed in no small way to the revolutions that challenged prevailing theoretical perspectives and methodological approaches. In the domain of sociological theory, ethnomethodology helped end the dominance of the structural functionalist paradigm associated with Talcott Parsons, and led to a major reconceptualization of the theory of action and its relationship to social structure. It was also an important impetus behind the broad-based social constructionist movement that revolutionized theorizing about subjects ranging from deviance to gender. In methodology, ethnomethodological studies challenged the scientistic pretensions of positivistic research methods, and fostered greater sensitivity and self-reflection among methodologists of various stripes. Finally, ethnomethodology inspired numerous research initiatives that revitalized a wide range of social science subfields, including the study of language and social interaction, the inner workings of bureaucratic and people-processing institutions, and the construction of formal scientific knowledge.

Intellectual Origins

The origins of ethnomethodology can be traced back to seminal research conducted in the late 1940s by Harold Garfinkel in the course of his graduate work in sociology at Harvard University. Garfinkel was a student of Talcott Parsons and he took Parsons' voluntaristic theory of action as a basic frame of reference, but he was also deeply influenced by the phenomenological writings of Alfred Schütz and the teachings of Aaron Gurwitsch. (Sometime later, the ordinary language philosophy of Ludwig Wittgenstein would become another source of inspiration.) The phenomenological enterprise directed Garfinkel's attention to certain fundamental pretheoretical problems posed by Parsons' theory of action, problems which were not adequately addressed within the Parsonian framework. What would later come to be known as 'ethnomethodology' is in large part the result of Garfinkel's sustained effort to confront these problems by empirical means (*see* Parsons, Talcott (1902–79)).

Parsons had sought to develop a theory of social action that would solve the Hobbesian problem of order while retaining purposive human agency within its framework. At the same time, he wished to avoid the limitations and pitfalls of strictly utilitarian thinking, which cannot account for the ends toward which action is oriented, and which treats intrinsic rationality as the sole standard governing the selection of means. Parsons' solution centered around the role of moral values in social action – he proposed that, through socialization, such values become internalized and incorporated into the actor's personality, where they guide both the selection of ends of

action and the normative means by which they are sought. Insofar as such values are also institutionalized within a society and are shared by societal members, patterned activity and social cohesion will be the natural result.

Parsons' theory places primary emphasis on the motivational wellsprings of action, and in so doing it has relatively little to say about how concrete actions are managed and coordinated in real time. Thus, as Heritage (1987: 228) has observed, it is less a theory of action per se than a theory of the dispositions that give rise to action. Correspondingly, Parsons' theory fails to give serious consideration to the knowledge and forms of reasoning that actors themselves bring to bear in the course of producing and recognizing actual conduct. This insensitivity to actors' *endogenous* perspectives is exacerbated by Parsons' decision to define rational action in terms of an *exogenous* standard based on consistency with the application of scientific/logico-empirical methods. Any deviation from this standard is regarded as the embodiment of ignorance and error, in which case actors' own explanations of their actions may be dismissed in favor of a 'scientific' explanation cast in the language of norms and values.

Garfinkel recognized that Parsons' motivational solution to the Hobbesian problem of order implicates an analytically prior *cognitive* problem of order, one that cannot be resolved without due consideration to the common sense knowledge and practical reasoning employed by actors themselves. How do actors analyze their circumstances, determine which features are relevant for present purposes, and grasp what those features 'mean?' How do they know what forms of behavior will be recognized as normatively appropriate under those circumstances? How in other words is the fundamental intelligibility of action and circumstance accomplished? It is questions like these that Garfinkel sought to answer.

Garfinkel's Classic Studies

Although Garfinkel's central analytic questions were inspired by phenomenology, his way of addressing them was distinctively his own. He eschewed highly interpretive methods involving introspection or freewheeling *verstehen* in favor of a rigorously empirical approach based on close observation of social behavior. Since the methods of reasoning of interest are those embodied in social activities, they are publicly available to both lay members of society and professional analysts and are, at least in principle, eminently observable.

Analyzing such methods of reasoning-in-action, however, is intrinsically difficult because they are so deeply taken for granted. As an omnipresent resource for the management of social life, practical reasoning is not ordinarily an object of conscious reflection in its own right. Its very familiarity renders it elusive and resistant to analysis.

Garfinkel's ingenious solution to this problem entailed seeking out extraordinary situations in which the sense-making process is foregrounded, exaggerated, and hence rendered conspicuous. For the most part, these situations were engineered by Garfinkel and his associates through a series of experiments and quasi-experimental demonstrations. The situations often involved social actors encountering anomalous events – typically, particular social actions – that are incongruous with a pre-existing or default 'definition of the situation.' In one such experiment, subjects were engaged in a game of tic-tac-toe by experimenters who would erase and reposition the subjects' mark before placing their own mark on the grid. In another experiment, subjects were told to ask yes/no questions of a counselor but, unbeknownst to them, the 'answers' were given according to a random schedule. Elsewhere the focus shifted to incongruities that arise naturalistically rather than experimentally, and to the practices of social actors rather than observers – an example is the study of 'Agnes,' a person whose masculine genitalia and biography were at odds with her claim to female status. Across these examples, incongruity and its management served to throw processes of practical reasoning into sharp relief.

The results revealed that interactants would strive to render the situation as coherent and intelligible, invoking a wide range of background assumptions, contextual knowledge, and other elements of common sense – often in the form of natural language accounts – to this end. Three patterns of response were recurrent. Subjects would either (1) 'normalize' the anomalous event by reinterpreting it as consistent with the prior definition of the situation, (2) 'demonize' the event by treating it as a motivated and morally suspect departure from normality, or (3) reconstitute the environing situation so as to make it congruent with what is ostensibly taking place.

These outcomes demonstrate forcefully that interactants make sense of a given action by considering not only the intrinsic properties of the action itself, but also its relationship to the broader social environment in which it is embedded. The meaning and import of any particular action is thus thoroughly context-dependent. This would be a trivial matter if contexts were autonomous and stable – 'the' context of an action could then serve as an archimedian vantage point from which to disambiguate that action. Contexts, however, do not have such transcendent properties – the relevant context at any given moment is derived primarily from the actions through which it is composed, and even a single incongruous action can dramatically alter the context that is understood to be in play. Practical reasoning is thus a dynamic and reflexive process whereby the sense of particular actions and environing social situations are ongoingly adjusted and reconciled with one another. By implication, the experiential reality and temporal persistence of any social situation – whether a game of tic-tac-toe, a course of counseling advice, or a person's sexual status – rests upon a foundation of such reasoning practices.

Theoretical Implications

Garfinkel's findings amount to a major reconceptualization of the fundamental locus of social order. Most social science theories view social life as organized by structural entities (e.g., social institutions, cultural symbol systems, structures of race, class, and gender, etc.) that stand outside of the flow of events in everyday life and exert a more or less determining influence on the course of those events. Such theories embody what might be thought of as a *top-down* conception of social order. Garfinkel, in contrast, offers a thoroughly *bottom-up* conception, and this theoretical inversion is a natural consequence of his decision to treat social order as a cognitive rather than

a Hobbesian problem – not a problem of how conflict is avoided and solidarity maintained, but a problem of how the social world, whatever its character, becomes intelligible and accountable to its members. From this vantage point, every orderly feature of social life is an ongoing contingent accomplishment, the result of members' concerted effort to make those features recognizable to one another and the basis for subsequent action.

This theoretical inversion has further implications for the theory of action, and in particular for the role of social norms (as well as rules, conventions, etc.) in the conduct of action (Heritage, 1984, 1987; Wilson, 1971). In the traditional Parsonian view, norms regulate action by specifying what courses of action are appropriate under given circumstances. This view presupposes that situations, norms, and actions are independent entities, with each situation standing outside of the actions contained within it, and predefined norms constraining those actions to unfold along situationally appropriate lines. However, Garfinkel's findings about the nature of practical reasoning suggest that this viewpoint is fundamentally misguided. Far from being independent entities, situations, actions, and norms stand in a co-constitutive or reflexive relationship to one another. Correspondingly, actors are knowledgeable agents at the very center of this process, with the capacity to alter or transform the 'definition of the situation' through their actions, and to decide upon the sense and applicability of the norms deemed relevant to that situation. Thus, the experiential reality of social norms, like every other organized feature of social life, rests upon a foundation of practical reasoning whereby action is produced and rendered intelligible in normative terms. This does not mean that norms are inconsequential for social organization, but their primary significance is constitutive rather than regulative – norms play a crucial role as a resource for imputing meaning and motivation to situated behavior (*see* Action, Theories of Social; Norms).

Contemporary Research Initiatives

Following Garfinkel's classic studies, ethnomethodological research developed in a number of fruitful directions. While some researchers continued to elucidate the generic properties of practical reasoning (e.g., Pollner, 1987), most shifted toward examining such reasoning as it is applied in various specialized domains of social life. At the same time, the predominant research methods employed by ethnomethodologists underwent a corresponding shift away from quasi-experimental designs and toward more naturalistic methods involving the direct observation of ordinary conduct. The resulting corpus of studies resists easy summary, but three prominent lines of work will be distinguished.

People-Processing Institutions

Ethnomethodologists first explored various organizational environments involved in people processing: schools, public welfare offices, police departments, and so on (e.g., Cicourel, 1968; Sudnow, 1965; Wieder, 1974). These studies repeatedly documented the inadequacy of codified rules, formal procedures, and informal norms as explanations of organizational conduct. Rules and allied phenomena fail to capture the elaborate judgmental work necessary to implement the rules in concrete circumstances, and more generally to perform competently the tasks intrinsic to each setting. These previously unexamined professional competencies and discretionary judgements became the focus of close scrutiny and sustained analysis, shedding new light on the complex inner workings of people-processing institutions.

The findings have important implications for our understanding of organizational decision-making and its products. The institutions under examination routinely generate official designations of persons (e.g., 'criminal,' 'juvenile delinquent,' 'qualified welfare recipient,' etc.) and their actions (e.g., 'burglary,' 'assault,' etc.). Such designations do not result from workers mechanically applying clearly defined criteria to each case, but involve various ad hoc considerations guided by common sense knowledge of what outcome would be normal and reasonable under the circumstances. Insofar as the resulting designations form the basis for calculations of official statistics (on rates of crime, mortality, poverty, etc.), these studies also cast doubt on the validity of official statistics and their value for social scientific research (*see* Bureaucracy, Sociology of).

Science and Technology

Ethnomethodologists have also ventured into the laboratory to explore the highly technical competencies that underlie the creation of scientific and mathematical knowledge, and the production and use of technological artifacts. This line of research is most closely associated with the later work of Garfinkel himself and his immediate associates (e.g., Livingston, 1986; Lynch, 1993). It is reminiscent of Edmund Husserl's later writings on the European sciences in that a primary objective is to elucidate the unexplicated foundations of scientific knowledge. However, where Husserl's enterprise was essentially philosophical in character and retained a phenomenological concern with transcendental consciousness, the ethnomethodological approach proceeds empirically by examining publicly available details of situated 'workbench' practices.

Much research in this vein is concerned with the genesis and reproduction of technical discoveries ranging from physical objects to mathematical theorems. Such discoveries are arrived at, accountably validated, and made intersubjectively available through complex courses of practical reasoning and embodied activity that receive scant attention in scientific texts. This may seem surprising, given that scientists are supposed to document their own methods so as to permit replication, but scientists' methodological descriptions – like all abstracted accounts of situated action – are necessarily incomplete. Ethnomethodologists thus seek to recover the mundane praxiological foundations of discovering work, and in so doing they re-specify discovered objects as locally produced and naturally accountable achievements. Furthermore, the very concept of discovery is transformed by this research. If physical and mathematical objects are unavoidably intertwined with situated practices, then the discovery of any such object is at the same time the discovery of a substrate of practices through which that object

may be accountably produced and reproduced within concrete situations.

Ethnomethodological studies of science and technology hold the promise of yielding findings that represent recognized contributions to the various technical disciplines being studied. Indeed, the practical value of ethnomethodology is increasingly recognized in computer science, software engineering, and human–computer interaction (Suchman, 1987; Button, 1993).

Talk and Social Interaction

Perhaps the most widespread contemporary variant of ethnomethodology is what has come to be known as *conversation analysis*. This burgeoning field was developed by Harvey Sacks, originally Garfinkel's student and colleague, in collaboration with Emanuel Schegloff and Gail Jefferson (e.g., Sacks, 1992; Atkinson and Heritage, 1984).

Conversation analysis (henceforth CA) involves the study of practical reasoning as it is put to use in the conduct of spoken interaction. The domain of interaction – what Erving Goffman referred to as *the interaction order* – is more general than any of the specialized institutional domains investigated by ethnomethodologists, for interaction lies at the heart of virtually all of these institutions and extends as well to informal encounters between persons. Moreover, in the history of the human species, interaction developed long before other societal institutions came into being, and is in this respect 'the primordial site of sociality' (Schegloff, 1988).

CA differs from other lines of ethnomethodological research not only in substance but in methodology. Conversation analysts rely exclusively on audio- and video-recordings of interactional data, and transcripts that capture the details of interaction as it actually occurs. Such data have numerous advantages – they can be examined repeatedly, analyzed at an unprecedented level of detail, and reproduced in published works so that readers can independently assess the validity of analytic claims.

The resulting research enterprise has generated an impressive array of interlocking and cumulative findings on a wide range of subjects. These include the organization of turn taking, action sequences, lexical choice, the relationship between talk and nonvocal activities, and the collaborative management of various interactional activities (e.g., giving advice, delivering good and bad news, telling troubles, etc.). More recently, researchers have applied the analytic resources of CA to various phenomena that intersect with, and can be informed by, the study of talk-in-interaction. These include how talk is organized in various institutional settings, and how it serves as a medium for the accomplishment of occupational tasks such as medical examinations, classroom lessons, journalistic interviews, trial examinations, and so on (e.g., Boden and Zimemrman, 1991; Drew and Heritage, 1992). Researchers have also begun to explore how the study of talk-in-interaction can illuminate linguistic phenomena such as grammar (Ochs et al., 1996), as well as medical disorders such as aphasia that manifest themselves at the level of speech (Goodwin, 1995; Heeschen and Schegloff, 1999).

See also: Conversation Analysis: Sociological; Parsons, Talcott (1902–79); Science and Technology Studies, Ethnomethodology of.

Bibliography

Atkinson, J.M., Heritage, J. (Eds.), 1984. Structures of Social Action: Studies in Conversation Analysis. Cambridge University Press, Cambridge, UK.
Boden, D., Zimmerman, D.H. (Eds.), 1991. Talk and Social Structure. Polity Press, Cambridge, UK.
Button, G. (Ed.), 1993. Technology in Working Order: Studies of Work, Interaction and Technology. Routledge, London.
Cicourel, A.V., 1968. The Social Organization of Juvenile Justice. Wiley, New York.
Drew, P., Heritage, J. (Eds.), 1992. Talk at Work. Cambridge University Press, Cambridge, UK.
Garfinkel, H., 1967. Studies in Ethnomethodology. Prentice-Hall, Englewood Cliffs, NJ.
Goodwin, C., 1995. Co-constructing meaning in conversations with an aphasiac man. Research on Language and Social Interaction 28, 233–260.
Heeschen, C., Schegloff, E.A., 1999. Agrammatism, adaptation theory, and conversation analysis: on the role of the so-called telegraphic style in talk-in-interaction. Aphasiology 13, 365–405.
Heritage, J., 1984. Garfinkel and Ethnomethodology. Polity Press, Cambridge, UK.
Heritage, J., 1987. Ethnomethodology. In: Giddens, A., Turner, J. (Eds.), Social Theory Today. Polity Press, Cambridge, UK, pp. 224–272.
Livingston, E., 1986. Ethnomethodological Foundations of Mathematics. Routledge, London.
Lynch, M., 1993. Scientific Practice and Ordinary Action: Ethnomethodology and Social Studies of Science. Cambridge University Press, Cambridge, UK.
Ochs, E., Schegloff, E.A., Thompson, S.A., 1996. Interaction and Grammar. Cambridge University Press, Cambridge, UK.
Pollner, M., 1987. Mundane Reason: Reality in Everyday and Sociological Discourse. Cambridge University Press, Cambridge, UK.
Sacks, H., 1992. Lectures on Conversation. Blackwell, Oxford, UK.
Schegloff, E.A., 1988. Goffman and the analysis of conversation. In: Drew, P., Wootton, A. (Eds.), Erving Goffman: Exploring the Interaction Order. Polity Press, Cambridge, UK, pp. 89–135.
Suchman, L.A., 1987. Plans and Situated Actions: The Problem of Human Machine Communication. Cambridge University Press, Cambridge, UK.
Sudnow, D., 1965. Normal crimes. Social Problems 12, 255–276.
Wieder, D.L., 1974. Language and Social Reality. Mouton, The Hague.
Wilson, T.P., 1971. Normative and interpretive paradigms in sociology. In: Douglas, J. (Ed.), Understanding Everyday Life. Routledge, London, pp. 57–79.

Etymology

Steven N Dworkin, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a complete revision of the previous edition article by W.P. Lehmann, volume 7, pp. 4874–4877, © 2001, Elsevier Ltd.

Abstract

This article begins by defining etymology as both the identification of word origins and the preparation of histories of individual words or word families. It next offers a brief overview of the history of this linguistic subdiscipline. The article then discusses such topics as the relationship of etymology to other branches of historical linguistics, the diffusion of the findings of etymological research (especially in etymological dictionaries), theoretical and methodological questions, and the insights provided by etymological reconstruction of the lexicon of undocumented language states into the cultures and social structures of past historical periods.

Definition and Scope of Etymology

Etymology can be concisely defined as the search for and identification of the origin of a given word in a specific language or of a given series of historically related words in a language family. It is the etymologist's responsibility to uncover and describe the morphophonological form and the semantic range of the base that forms the starting point for the evolution over time and space of the lexical item under investigation. For many specialists in this field, etymology has come to designate the preparation of the complete history or biography of a word. Identification of the word's immediate formal and semantic origin is now but the first step in this task, which includes accounting for all facets of the chosen word's formal and semantic history, the derivatives to which it has given rise, its rivalry with other referentially similar lexical items, its spread into other languages as a loanword (if appropriate) and, when relevant, the circumstances leading to its demise or slide into obsolescence. This approach to etymology seems feasible only for languages with a long and abundant recorded history. Etymology is consequently an integral part of the broader discipline of diachronic lexicology.

In general terms, a given word can (1) form part of the language's inherited lexicon, (2) represent a borrowing from a language with which the language under study has come into contact, or (3) result from the various (often language-specific) internal processes of lexical creation available to speakers, such as prefixal and suffixal derivation, infixation, compounding, sound symbolism, and lexical blends. The identification of the elements that fall into each of these three broad categories in a specific language forms part of the brief of etymology, the sole linguistic discipline that is exclusively diachronic in scope. Most etymological research devoted to individual languages or language families concentrates on the history of the inherited lexicon, and, to a lesser extent, on loanwords. Derivatives have traditionally received far less attention. Yet, the history of each apparent derivative deserves individual study in order to differentiate authentic derivatives created within the language at issue from those which, genetically, are actually inherited from the lexicon of the ancestor language, or are borrowings from yet another language (e.g., the many Spanish derivatives in – *aje* which, for the most part, are borrowings from neighboring Gallo-Romance and Catalan). Etymology is one branch of historical linguistics that needs to make use of the findings of many of the other fields, such as phonology, inflectional and derivational morphology, syntax, semantics, and lexical typology, as well as relevant extralinguistic knowledge garnered from the study of political, social, cultural, and literary history, and textual criticism. Unless the analyst is dealing with the lexicon of an isolate language (such as Basque), research into the history of the lexicon of an individual language needs to be comparative in nature, with the presentation of cognates from other members of the relevant language family. Often, evidence from a related language may tip the scales in favor of one of several competing etymological hypotheses, or even point the way to a new solution. Even when dealing with an isolate language, comparative evidence from attested regional varieties of that language may be useful in etymological reconstruction.

Etymology in the Past

The origin of words has attracted the attention of scholars and thinkers interested in human language for over two millennia. Early philosophers thought that the original meaning of a word reflected its true essence, a belief reflected in the Greek label *etymologia* (a compound of *etymos* 'true' and *logos* 'word'). Such etymologizing plays a significant role in various religious traditions. The Bible and the subsequent Jewish and Christian traditions of biblical exegesis contain numerous passages in which the text seeks to identify the original essential meaning of a word, especially of personal and place names. A similar approach to the relationship between linguistic signs and the realities that they name is found in the Chinese *Shih ming* (*c*.CE 200). Old French *ethimologie* and Old Italian *etimologia* are documented in thirteenth-century texts. English *etymology*, rarely found before the sixteenth century, is a borrowing from Old French. With the awakening of scholarly interest in vernacular languages during the Renaissance, there began to appear in the sixteenth and seventeenth centuries, etymological dictionaries of numerous European languages. Many of the etymologies proposed in these prescientific works are fanciful, often based on fortuitous formal and semantic resemblances.

In the western scholarly tradition, etymology became a truly scholarly discipline only with the establishment of the comparative method and the founding of Indo-European (IE) comparative linguistics in the early nineteenth century with its emphasis on the regularity of sound change, a basic underlying principle of etymological research and reconstruction. This period witnessed the publication of the first great etymological dictionaries of IE languages (e.g., Pott, 1833–36; Diez, 1853; Fick, 1871; Kluge, 1883). As a scholarly discipline, etymology forms part of the European linguistic tradition. The major etymological dictionaries of numerous non-European languages and language families were prepared over the course of the last century by European and American experts.

Etymology and Other Linguistic Fields

Etymology finds itself in a symbiotic relationship with some of the other disciplines of historical linguistics. In 1875, the American pioneer William Dwight Whitney (1827–94) declared (echoing thoughts voiced earlier in the century by Jakob Grimm): "The whole process of linguistic research begins in and depends upon etymology" (quoted in Malkiel, 1993: 20). The sound correspondences that form the backbone of the study of a language's formal evolution are extrapolated from a large quantity of accepted and verified etymologies. An equally reliable etymological database is needed to study the patterns of semantic evolution. Yet, a deep knowledge of formal and semantic evolutionary patterns identified on the basis of well-established etymologies is essential for the reconstruction or identification of the etymon of the many words in a language whose origins are unclear, controversial, or completely unknown. Many of the nineteenth-century pioneers of etymological research authored both historical grammars and, separately, etymological dictionaries of the same language or language family, for example, Jakob Grimm (1822–1837, 1854–1960 (the latter work was carried out in collaboration with his brother Wilhelm Grimm)) for German and Friedrich Diez (1836–1844, 1853) and Wilhelm Meyer-Lübke (1890–1902, 1911–1920) for Romance. In all three cases, publication of the historical grammar preceded the corresponding etymological dictionary. The close relationship between etymology and other fields of historical linguistic research constitutes one of the main themes running through the etymological writings of the Romanist Yakov Malkiel (see, for example, the essays collected in Malkiel, 1968).

Etymology has to distinguish between the immediate and the distant origin of the word under study. Etymological studies of the lexicon of a specific language tend to work backward from the modern language. Historians of each language or language family must delimit, in light of the available linguistic and extralinguistic data, how far back in time to trace the word's origin. For example, etymological studies of the inherited lexicon of a Romance language traditionally seek to identify the appropriate (documented) Latin base of a Romance word, without discussing the more distant Italic or IE origin of the relevant Latin form (although there are exceptions, such as the Indo-Europeanist Giacomo Devoto's *Avviamento alla etimologia italiana*, which identifies, when relevant, the IE root of the etymon). In the case of the numerous borrowings from other languages that entered each Romance language through contact situations, the etymologist will identify the etymon that entered from the immediate source language, without elaborating on its origins in the donor language. To give but one example, the historian of the Spanish lexicon will identify Spanish *azul* 'blue' as a borrowing from Arabic (its immediate origin) but will not discuss the circumstances of this word's entry into Arabic from Persian. On the other hand, the author of an etymological dictionary of Gothic can choose to identify either the proto-Germanic or the proto-Indo-European root of the lexical item under examination.

Etymology and Reconstruction

On the basis of cognate sets, the oldest available data, and the application of the comparative method, etymological dictionaries for most languages or language families strive to reconstruct undocumented roots as the appropriate starting points. Practitioners of Romance etymology are blessed (or cursed?) by having at their disposal the rich documentation afforded by written Latin, reflecting the different written registers of the spoken language (labeled variously as Spoken Latin, Vulgar Latin, and Proto-Romance) underlying the Romance languages. Traditional Romance etymology has contented itself with presenting as the etymon the form of the base as it appears in written Latin, with asterisked hypothetical bases invoked as a last resort in the absence of a written Latin correlate. Often, the standardized orthography of the traditional Latin form masks the phonetic reality of the etymon in the spoken language, which can be recovered through application to the Romance data of the comparative method. This is the underlying principle of the *Dictionnaire Étymologique Roman* project codirected by Eva Buchi and Wolfgang Schweickard, in which the Proto-Romance starting point is presented in phonemic transcription (although the Romance descendants are offered in standard orthography). Each entry ends with explicit reference to the relevant form of the base in written Latin. I shall offer one example of this procedure. The Romance data permit the reconstruction at the level of Proto-Romance of the bases */askUlt-a-*/eskult-a-/, which correspond to ausculty */eskult-a-/are 'to listen' of written Latin. This use of comparative reconstruction in Romance etymology has been the subject of a lively debate among specialists (see, for example, Varvaro, 2011a,b and, in response Buchi and Schweickard, 2011a,b).

The attempt to identify the etymologies for a language whose genetic affiliation is uncertain poses different problems. Well-established cognate sets play an important role in the reconstruction of etymological bases. In the case of very large language families such as Sino-Tibetan Burman or Austronesian, whose genetic subgroupings are highly controversial, the etymologist must take care in weighing the role of possible cognates in an etymological reconstruction, and must distinguish between formal and semantic resemblances that reflect actual genetic affiliation versus those that may result from processes of language contact.

Over the years etymologists have refined their techniques for reconstructing the form of the relevant bases underlying the words or word families under study. The semantic side of

the (reconstructed) base has not been a primary focus of etymological research. Emile Benveniste's seminal study (1954) lays out many of the difficult and complex issues involved in semantic reconstruction. Typical is the declaration made in Orel (2003: xii): "The semantic reconstruction of the proto-Germanic words was not even attempted: too many complications and arguments would have followed the decision to ascribe meanings to Proto-Germanic words." Only recently have some specialists attempted to apply the principles of the comparative method to the reconstruction of the semantic range of the proposed etymological base. It would be reasonable to assume that meanings shared by all or many of the descendants of a given base (reconstructed or documented) go back to the semantic range of that item in the protolanguage. Meanings found only in a small number of geographically scattered cognate forms may reflect inheritances from the protolanguage, or, alternatively, might result from independent but parallel semantic innovations brought about by the working of such cognitive processes as metaphor and metonymy. It is certainly realistic to posit polysemy at the level of the etymon (Rankin, 2003). One of the innovative features of the aforecited *Dictionnaire Étymologique Roman* is its meticulous reconstruction and systematic ordering in each entry of the semantics of the posited proto-Romance base. The reconstructions are based on the attested semantic range of each Romance cognate rather than on the meanings of the written Latin equivalents.

Etymological Dictionaries

Etymological dictionaries of entire language families can adapt a prospective perspective, taking as the starting point for each entry the documented or, more often, reconstructed lexicon of the source language, and tracing the evolution of each base into the individual descendant languages. This approach is also feasible for the study of the vocabulary of individual languages. In addition to the pan-Romance *Romanisches Etymologisches Wörterbuch* (1911, 1920_3), the multivolumed *Französisches Etymologishes Wörterbuch* (von Wartburg et al., 1922–2002) and the ongoing *Lessico Etimologico Italiano* (Pfister, 1979) are arranged alphabetically by etymon. A similar structure is found in the innovative and methodologically controversial *Dictionnaire Étymologique Roman*, whose entries are arranged according to the reconstructed proto-Romance base (presented in phonemic transcription), rather than by the written Latin equivalent. This dictionary has limited its scope in its first stage of elaboration to approximately 500 proto-Romance bases that have survived into all, or most of, the Romance languages. However, this prospectively arranged system poses serious problems when deciding in which entry to place the many lexical items in the target language whose origins are far from transparent. The team responsible for the continuation of the *Französisches Etymologishes Wörterbuch* has had to prepare several fascicules, arranged by semantic categories, devoted to words of unknown origin. In dictionaries arranged according to the etymon, indexes of the words studied are essential for the nonspecialist reader interested in the origin of a specific word.

Etymological information can be transmitted in several forms. The layperson seeking to know the origin of a given word in a specific language will often find a concise identification of the etymology in the relevant entry in a monolingual reference dictionary. For example, most dictionaries of, say, French will tell the reader of the entry for *père* 'father' that the word comes from the Latin *pater*, but will not explain the formal or semantic details of its evolution. With regard to the origin and history of words whose etymology is unknown or disputed, specialists usually pass on their findings or hypotheses to their colleagues in the form of brief notes or more elaborated articles in scholarly journals, usually read only by specialists. Far less frequent today are monographs or even books devoted to the detailed study of the history of individual words or of entire word families. The culmination of etymological research on the lexicon of a language or language family is the etymological dictionary where each entry discusses, insofar as is feasible, the given word's formal and semantic history, as well as offering, when appropriate, a scholarly justification for the proposed etymology and, if appropriate, a critical review of competing explanatory hypotheses. Indeed, it is often an etymological dictionary whose entries make available to the wider scholarly community the findings first reported in detailed and specialized etymological articles and monographs.

Today, most major languages or language families have at least one etymological dictionary devoted to their lexicon. An excellent and thorough critical overview (with bibliography of primary and secondary sources) of such dictionaries appears in Eva Buchi's contribution to the forthcoming *Oxford Handbook of Lexicography*. Malkiel's (1976) typology of etymological dictionaries retains its value. He adduced eight criteria for classifying such dictionaries: (1) time depth (period to which the etymologies are traced back), (2) direction of analysis (prospection or retrospection), (3) range (languages dealt with), (4) grand strategy (structural division of the dictionary), (5) entry structuring (linear presentation of the chosen features), (6) breadth (information given in the front- and back matter vs within the individual entries), (7) scope (general lexicon vs parts of it, e.g., borrowings), and (8) character (author's purpose and level of tone).The entries in most such dictionaries are arranged alphabetically by the word whose origin it seeks to elucidate or by the chosen etymon, rather than by semantic categories or lexical families.

Etymology, Theory, and Methodology

For the most part, etymologists prefer to devote their efforts to the identification of etyma and the preparation of word histories/lexical biographies rather than to discussions of methodology and theory. Although Jules Gilliéron originally coined his famous dictum 'Each word has its own history' as a response to the Neogrammarian view of the rigidity of the so-called sound laws, his words could equally apply to lexical history, including etymology. Each etymological solution is unique. Consequently, etymology is a linguistic subdiscipline that does not readily lend itself to rigid methodological systematization or vociferous and polemical theoretical debate. There can be no unified theory of etymological analysis, a situation that may account for the current estrangement between etymology and general historical linguistics. Some renowned etymologists (e.g., Vittorio

Bertoldi have gone so far as to describe etymology as an 'art' (cf the title of his introductory manual L'arte della etimologia (1952)).

Etymology no longer occupies a central place in the discourse of historical linguistics. Etymological method today seldom receives lengthy attention in manuals and handbooks of historical linguistics and is rarely afforded separate book-length treatment. Notable recent exceptions in the English-speaking world are the overviews of etymology by Anatoly Lieberman (2005) and Philip Durkin (2009) aimed at the educated nonspecialist reader. At the beginning of the twentieth century, Romance etymologists debated the relative weight to be accorded to formal versus semantic criteria in resolving complicated etymological cruxes. In retrospect, this was a pointless discussion, as the etymologist must weigh all the relevant linguistic (and, if appropriate, nonlinguistic) evidence. There exists no algorithm for evaluating each case. Etymology is language specific, and most etymologists specialize in the history of one language or one language family. It is next to impossible to identify etymological universals. Can we even speak of 'general etymology' in the same way as 'general phonetics' or 'general semantics?' A comparative etymological analysis of the lexicon of individual languages or language families undertaken by a team of specialists may identify certain semantic fields or grammatical categories that possibly contain a high number of elements of unclear origin, or may reveal certain cross-linguistic patterns in lexical history. In the Romance languages, the origin of many of the words corresponding to 'boy, child' is controversial: Fr. *garçon*, Sp. *muchacho*, *niño*, *chico*, *mozo*, Ptg. *rapaz*, It. *ragazzo*, and Rum. *băiat*. Although outside the Romance domain, one may include here the English *boy*, originally 'servant.'

Etymology, Loanwords, and History

All words evolve in a historical, cultural, and social context. For many etymologists (including the author of this article), the opportunity to study lexical history in such a framework is one of the joys of humanistically slanted diachronic lexical research. The link between diachronic lexical studies and material reality can be seen in the *Wörter und Sachen* school of dialect geography that flourished among European linguists in the first half of the twentieth century. Such studies focused, from a historical perspective, on how a language or group of languages expressed a given reality of material civilization (e.g., animal-, plant-, bird-names, farm implements, tools, etc.). The identification and analysis of loanwords resulting from language contact (at the level of both oral and written language) require close examination of the relevant extralinguistic circumstances, including accurate dating and philological analysis of the first documentation of the relevant forms in both the source and recipient language. A loanword can exist in the recipient language for a long time before it surfaces in a text, the only source of documentation for past stages of the language. The historical study of loanwords requires the etymologist to study the integration and diffusion of the neologism through the speech community, processes that may take a long time after its introduction into the host language (for an example of the application of this approach to Spanish, see Dworkin, 2012). A recent cross-linguistic study of loanwords in about 40 languages (Haspelmath and Uri Tadmor, 2009) has demonstrated that lexical items belonging to certain semantic fields (e.g., emotions and values, motion, kinship, spatial relations, body parts, and sense perception), semantic areas that are universal to all human societies, are rarely borrowed. In contrast, fields such as clothing and grooming, religion and belief, the house, law, agriculture and vegetation, social and political relations, food and drink, domains subject to a high degree of intercultural influence, show higher rates of borrowing. These considerations may guide etymologists in deciding in doubtful cases whether it is reasonable to propose that a specific word may be a borrowing. The results of etymological research form the foundation of a subfield known as linguistic paleontology, whose findings allow historians to analyze (more often than not, reconstructed) lexical data in order to recover details or formulate hypotheses concerning early (often prehistoric) societies, their religious beliefs, their kinship systems, the structure of the local economy, etc. (cf the attempts to reconstruct diverse facets of proto-Indo-European society, or the attempts to use specific items of the reconstructed lexicon to locate the homeland of speakers of proto-Indo-European).

See also: Comparative Method in Linguistics; Indo-European Languages; Internal Reconstruction in Linguistics; Lexical Semantics; Lexicology and Lexicography.

Bibliography

Benveniste, E., 1954. Problèmes sémantiques de la reconstruction. In: Benveniste, E. (Ed.), Problèmes de linguistique générale, Word, vol. 10, pp. 131–144. Martinet, A., Weinreich, U. (Eds.), Linguistics today. Linguistic Circle of New York, New York. Reprinted in Benveniste, E., 1966. Problèmes de linguistique générale. Gallimard, Paris, pp. 289–307.

Bertoldi, V., 1952. L'arte dell'etimologia. Liguori, Naples.

Buchi, E. Etymological dictionaries. In: Durkin, P. (Ed.), The Oxford Handbook of Lexicography. Oxford University Press, Oxford, forthcoming.

Buchi, E., Schweickard, W., 2011a. Sept malentendus dans la perception du DERom par Alberto Vàrvaro. Revue de linguistique romane 75, 305–312.

Buchi, E., Schweickard, W., 2011b. Ce qui oppose deux conceptions de l'étymologie romane. Réponse à Alberto Vàrvaro et contribution à un débat méthodologique en cours. Revue de linguistique romane 75, 628–635.

Devoto, G., 1966. Avviamento alla etimologia italiana. F. le Monnier, Florence.

Diez, F., 1836–1844. Grammatik der romanischen Sprachen, 3 vols. Weber, Bonn.

Diez, F., 1853. Etymologisches Wörterbuch der romanischen Sprachen, fifth ed. Adolph Marcus, Bonn. 1887.

Durkin, P., 2009. The Oxford Guide to Etymology. Oxford University Press, Oxford.

Dworkin, S.N., 2012. The History of the Spanish Lexicon: A Linguistic Perspective. Oxford University Press, Oxford.

Fick, R., 1871. Vergleichendes Wörterbuch der indogermanischen Sprachen. Vandehoeck and Ruprecht, Göttinge.

Grimm, J., 1822–1837. Deutsche Grammatik. Bei Deitrich, Göttingen.

Grimm, J., Grimm, W., 1854. Deutsches Wörterbuch. S. Hitzel, Leipzig.

Haspelmath, M., Uri, T., 2009. Loanwords in the World's Languages: A Comparative Handbook. Mouton de Gruyter, Berlin.

Kluge, F., 1883. Etymologisches Wörterbuch der deutschen Sprache. K.J. Trübner, Strassburg. Numerous revised editions and reprints.

Lieberman, A., 2005. Word Origins and How We Know Them. Etymology for Everyone. Oxford University Press, Oxford.

Malkiel, Y., 1976. Etymological Dictionaries: A Tentative Typology. University of Chicago Press, Chicago and London.

Malkiel, Y., 1968. Essays on Linguistic Themes. Blackwell, Oxford.
Malkiel, Y., 1993. Etymology. Cambridge University Press, Cambridge.
Meyer-Lübke, W., 1896–1902. Grammatik der romanischen Sprachen, 4 vols. Fues, Leipzig. French Translation (1890–1906). Grammaire des langues romanes. 4 vols. Welter, Paris.
Meyer-Lübke, W., 1911–1920. Romanisches Etymologisches Wörterbuch, third ed. Winter, Heidelberg. 1935.
Orel, V., 2003. A Handbook of Germanic Etymology. Brill, Leiden and Boston.
Pfister, M., 1979. Lessico etimologico italiano. Dr. Ludwig Reichert, Wiesbaden.
Pott, A., 1833–36[1859–1876]. Etymologische Forschungen auf dem Gebiete der Indogermanischen Sprachen. Meyer, Lemgo.
Rankin, R., 2003. The comparative method. In: Joseph, B., Janda, R. (Eds.), The Handbook of Historical Linguistics. Blackwell, Oxford, pp. 183–212.
Varvaro, A., 2011a. Il DERom: un nuovo REW? Revue de linguistique romane 75, 297–304.
Varvaro, A., 2011b. La 'rupture épistemologique' del DERom: Ancora sul metodo della etimologia romanza. Revue de linguistique romane 75, 623–627.
von Wartburg, W., et al. (Eds.), 1922–2002. Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes, 25 vols. Klopp/Winter/Teubner/Zbinden, Bonn; Heidelberg; Leipzig-Berlin; and Basel.

Eudaemonism

Heidrun Friese, Intercultural Communication, Chemnitz, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Eudaimonia and concepts of the good life and human flourishing have been of central concern for philosophy and not least with reference to Aristotelian thought. Tied to the Self, the relations to others and the just political order, it has been one of the key concepts of ethical-political thinking during the seventeenth and eighteenth centuries and – despite the eviction of *eudaimonia* in Kantian reasoning and the categorical imperative – has been revived in various strands of utilitarianism, critical thought, moral philosophy, virtue ethics, and (feminist) ethics of care.

Eudaimonia (lit. 'a good demon,' *Glück*, *prospérité*, *bonheur*, happiness, human flourishing, good life) has been one of the central concepts of ethics and philosophy. Happiness or better: the good life was, for once, connected to virtue; it was conceptualized in subjective terms, with reference to the self, to the (restraining of) passions and affects, to self-realization and the realization of personal objectives, or of one's overall life. Second, happiness has been cast in relation to the other, as friendship, love, recognition, as ethically founded intersubjectivity. Third, it has been connected to the just political order. The notion can be traced from its 'classical' conceptions in Aristotle, the Epicurean and stoic traditions, to its reelaboration in the philosophy in early modernity to Kant, who radically evicted the connection of happiness with virtue. Despite Kant's subjectivation of happiness, the notion returned persistently into the political realm – as 'redemption,' 'the principle of hope,' as 'emancipation,' 'undamaged life' (Frankfurt School), and, in the version of existentialist philosophy, as 'authenticity,' up to recent attempts of philosophy to found happiness in a just order, as, for example, in John Rawls, Martha Nussbaum, and Martin Seel, in virtue ethics and moral philosophy (Elisabeth Anscombe, Bernard Williams, Alasdair MacIntyre) or (feminist) ethics of care (Carol Gilligan, Virginia Held, Nel Noddings).

Following up on the ancient concepts of *eudaimonia*, the notion has been given a great variety of different meanings. Its semantic range reaches from 'the favor of fate, the advantageous lot, the happy circumstance,' all referring to accidental, external circumstances, to action and achieved goals making up the 'good life.' Across this breadth, thus, the term includes that which is outside of the reach of human beings, i.e., destiny (as in the ancient figures of *Tyche* or *Fortuna*, the goddess of fate, who blindly distributes her offerings), as well as a durable state that can be influenced and, in principle, achieved by virtuous human action. Precisely, the determination of such action, guided by the pursuit of happiness, has long been the basis of ethics and moral philosophy and of their debates about action guided by virtue, which then in turn was to determine and safeguard happiness and the good life (Caillé et al., 2001).

Eudaimonia is thus one of the central concepts of ethics and moral philosophy. It has been cast in a variety of ways and points into a number of different directions. Happiness was, for once, conceptualized in subjective terms, with reference to the self (*see* Self: Philosophical Aspects), to the passions and affects, to the realization of one's objectives, or of one's overall life. Second, it can be cast in relation to the other, as friendship, love, recognition, care, as ethically founded intersubjectivity. And finally, happiness and the good life have been connected to the community and its just political order. Even though this relation found its strongest expression in ancient philosophy, it has by far not disappeared from current discourse. In contrast, it may possibly even be claimed that the emphasis on justice as a precondition for happiness and the 'good life' presents, just as much as the reference to the self and to the other, an inescapable point of concern of modern (moral) philosophy. Additionally, conceptualizations of happiness show a temporal tension between fulfillment across all of the life span, on the one hand, and fulfillment in the sudden and epiphanic moment, on the other. Further, it is marked by a transcendental religious tension that stretches from revelation, redemption, and salvation to joy. It carries with it a political–social–ethical tension between an individualistic view of personal self-realization and the conception of a good and just political order. Finally, a tension with regard to a theory of action is unavoidable, a tension that situates the singular human being between the capability to act autonomously and fate, destiny, and luck (Nussbaum, 2001). In a variety of ways, these tensions can be traced back to the heritage of Greek reason.

The 'Classical' Tradition

In the Nicomachean Ethics (Nagel, 1980; Destrée, 2003), engaging with Plato's conception, Aristotle links the concept of *eudaimonia* to the three human life forms (Wolf, 1996): to pleasure, from which most human beings gain their view about what happiness is; to the service of the state; and to life dedicated to philosophy. Because reason is a distinctively human capacity, neither profane enjoyment nor politics and the art of governing, but *theorein* is the highest in the hierarchy of human activities, in which *eudaimonia* can be realized. "Happiness is activity (*energeia*) in accordance with virtue" (X, 7) and "the highest good reachable for human beings presents itself as the activity of the soul in the sense of its essential prowess (*areté*)" (Aristotle, 1934: I, 6, see Heidegger, 1992). The activity of *theorein* is aiming at the cognition of that which is eternal. Being independent of changes in the external circumstances it is the most persistent among human activities. Is full of pleasure – on the one hand, because it is continuous and stretches across an entire life; on the other, because it participates in the eternal

 http://dx.doi.org/10.1016/B978-0-08-097086-8.26027-3

that it gets to recognize. Furthermore, to be active in the sense of *theorein* is a purpose for itself, the final telos, that which is pursued for its own sake, since it does not produce anything that lies outside of itself; it is self-sufficient. For Aristotle, happiness is an eminently social and political matter. It is tied to the polis, a just political community and is as well not accomplished without friendship, without the relations to others. Aristotle, thus, linked happiness to human essence in need to be actualized, to the relation to others in friendship, and to institutionally supported forms of living and acting together (Garver, 2011). With the disruption of the order of the polis as well as of Hellenic philosophy, in contrast, for the Stoa and for Epicurean hedonism (and as well for skepticism), *eudaimonia* becomes a concept that refers to purposes that reside in that which is available to human beings. *Eudaimonia* is, then, the striving for whatever is available and, vice versa, it asks for the diminution of demands and the avoidance of pain. Despite all differences between these ancient Greek schools of thought, they also show commonalities, such as in particular in the connection of virtue and happiness. Ancient Greek philosophies, as Kant stated, follow the same method in rejecting any validity to the assertion that virtue and happiness could be two separate elements of the highest good. They "consequently sought the unity of the principle in accordance with the rule of identity," even if they differ in the choice of the fundamental concept. "The Epicurean said: to be conscious of one's maxim leading to happiness is virtue; the Stoic said: to be conscious of one's virtue is happiness. For the first *prudence* was equivalent to morality; for the second, who chose a higher designation for virtue, *morality* alone was true wisdom" (Kant, 1993a: A 200, pp. 239–240/book II, chapter 2, p. 229).

Eudaimonia in the Philosophy of Early Modernity

Conceptions of happiness relating individuals to society changed profoundly during the period known as early modernity; these changes were part of broader reproblematizations in the moral and social philosophy of the time. Political thinking of this period introduces happiness as a component and as a means of politics. Happiness becomes one of the key concepts of ethical-political thinking during the seventeenth and eighteenth centuries, a thinking that at the same time reorders the relations between the concepts for affects and those for morality. No longer the affective drives of the human beings are the object under consideration, but rather human actions and their consequences. Affects as such are considered neither as good nor evil, it is rather individual selfishness and self-love (*amour-propre*) that moves into the center of concern (Mauzi, 1969). This focus generates the new problématique of moral philosophy: If it is the case that human beings pursue their happiness on the basis of their drives and affects, then it is the task of moral philosophy to determine how – despite that – a good social order can exist and be maintained.

The new emphasis on the consequences of individual action went along with a reconceptualization of the political. From now on, it was precisely their striving for happiness that made human beings inclined to subject themselves to the common order. This reasoning emerges in classical form in Thomas Hobbes's (1588–1679) *Leviathan*. Hobbes first determines 'felicity' as the attempt to completely accomplish the objectives of human desires. Since, however, all human beings hope to reach this goal, the one potentially turns into a danger for the other. To avoid such a situation of *bellum omnium contra omnes*, of 'war of every man against every man,' human beings assemble to form a state that itself embodies no longer any conception of happiness or of the good life, but that guards to prevent any situation from arising in which the various strivings for happiness collide.

At the end of this period, during which happiness remains the focal point of theoretical constellations, an attempt is made to indeed erect a polity that is committed to support its members' striving for happiness. The American Declaration of Independence even declares the 'pursuit of happiness' to be a human right (Jefferson, 1776). A historically novel compromise is here elaborated between a classic substantive and an individualistic instrumental conception of the state: the state respects the autonomy of the individual and the plurality of their views, but does not entirely renounce the proposition of a general guiding idea about the 'good life.' The individuals' striving for happiness – to the protection of which the state has been created, exactly in the sense of the theories of delegation – in its sum produces the common good. This common good, however, is now in a particular way mixed with elements of a Protestant ethic, in which inner-worldly happiness, for example, as economic success, gains divine approval. This unique privileging of the pursuit of happiness as a human right, which cannot be found in European declarations, is rooted in a particular relation between the state and the citizens. It is nevertheless close to European conceptions in another respect, namely in understanding the citizen predominantly as an economic subject, a *bourgeois* in contrast to a *citoyen*, whose pursuit of property needs to be protected by the state, a protection that is considered as being the primary task of the state.

The utilitarianism of the eighteenth and nineteenth centuries similarly still places happiness in the center of moral and social philosophy (*see* Utilitarian Social Thought, History of). The ethics of utilitarianism considers individual – as well as general – utility as the purpose of action and sets thus 'good' synonymous with 'useful.' Jeremy Bentham (1748–1832) calls for 'the greatest happiness of the greatest number' as the collective goal; by promoting the well-being of all, the well-being of the individual is promoted as well. Morals and legislation are that art to regulate human actions such that to bring forth 'the greatest possible quantity of happiness' (*An Introduction to the Principles of Morals and Legislation* 1789). Drawing on Bentham, John Stuart Mill (1806–73) considers the pursuit of the greatest possible happiness of all as the supreme goal as well (*Utilitarianism* 1861). On the one hand, utilitarianism maintained a conception according to which the pursuit of happiness is the most important motive of human action, and it also kept the link between happiness and politics. What was added were amendments that had become historically unavoidable, namely the view that every human being determines their happiness in their own way and strives for it, and that everybody has the same right to do so. On the other hand, however, an *aporia* between the common good and the diversity of the human pursuits of happiness had been created. This *aporia* was to be dealt with by assuming the comparability of the various individualized and

subjectivized pursuits. Measurability and quantification were the tools to make the diverse compatible, since the state's task to pursue the greatest happiness of the greatest number demanded the possibility of adding up the manifoldness of forms of individual happiness. Quantification of happiness is the attempt to maintain a measure for collective happiness by at the same time making only a minimum of substantive assumptions about happiness as such. Utilitarianism had inevitably to fail when radically subjectivized conceptions of happiness were to be compared to each other, measured and sublated in an overall concept of utility.

The Eviction of Eudaimonia from Philosophy: Immanuel Kant

In continental moral and political philosophy, the most systematic attempt to reorder the relation between action and happiness was undertaken by Immanuel Kant – it was at the same time an attempt with far-reaching consequences for later philosophy. Moral philosophy had tried to demonstrate that there is no, and could not possibly be any, contradiction between ethical demands on the one side, and the pursuit of happiness (or of utility or interest) on the other. Thus, it stated that virtuous action promotes happiness. Kant radically breaks this connection of virtue and happiness.

Happiness remains for Kant as well a concept that has a basic influence on human action: "Happiness is 'the satisfaction of all our desires'; *extensive*, in regard to their multiplicity; *intensive*, in regard to their degree; and *protensive*, in regard to their duration" (1974a: II, B 834, p. 677 [p. 518]), "*happiness* is the state of a rational being in the world in the whole of whose existence *everything goes according to his wish and will*, and rests, therefore, on the harmony of nature with his whole end as well as with the essential determining ground of his will" (1993a: A 224–225). This concept, however, is not suitable as a foundation of moral philosophy, according to the following reasoning.

First, happiness can hardly be the true purpose of nature, since in such a case, human action would not be guided by practical reason, steering the will, but subjected to the instincts (1993b: BA 5, p. 20). Second, the concept of happiness cannot be determined unequivocally, since it rests on experience, which is never closed, on the one hand (1993b: BA 46, p. 47, 1993a: A 46, p. 133), and since, on the other hand, the 'idea of happiness' requires an absolute whole, a present and future maximum of well-being. The contradiction between the principled openness of experience and the idea of the accomplishment of happiness cannot be sublated within time. Furthermore, the conception of a reachable maximum of happiness is impossible for two reasons. No sum of happiness can ever be added up, since only similar perceptions can be added up; feeling, however, is highly variable according to the manifoldness of being touched in the very intricate state of life. In addition, the nature of the human being can never halt and be satisfied in possession and enjoyment. The inclinations change, they grow, and they always leave an even larger emptiness behind them than the one hoped to fill. (The dynamics of needs and their limitless character (*apeiron*) was already of concern for Hellenic philosophy.) The fact that each human being may have a specific view of happiness, and which furthermore may undergo change in the course of life, puts a further limit to any general determination of happiness. Happiness is "a subjective basic feeling of enjoyment of dissatisfaction" (1993a: A 46, p. 133); a subjective feeling, however, cannot be the foundation of general rules for action. As a consequence, thus, it is not possible to demand the agreement of the pursuit of happiness with moral requirements – which precisely was the former assumption: that namely the pursuit of happiness realizes ethical requirements since happiness resides only in virtuous action.

Such subjectivization separates happiness from virtue and ethics and subordinates it to the categorical imperative. "Act in such a way that the maxim of your will could always serve as the principle of a general law." It is no longer the pursuit of happiness that makes an action moral, but only the respect of the moral law and the obedience to duty, since it is only in them that human freedom is anchored. The categorical imperative relates to "the idea of freedom" (1993b: BA 102–103, p. 84), since free will is nothing but "autonomy, that is, the property of the will to be its own law" (1993b: BA 97–98, p. 81). Happiness is accidental and not at our disposal, but freedom is effective "independent of external causes that determine it" (1993b: BA 97–98, p. 81, BA 99–100, pp. 82–83). A human being does not act morally because he or she hopes for happiness – "If *eudaemonia* (the principle of happiness) instead of *eleutheromania* (the principle of freedom for inner legislation) is set as the foundation, the consequence will be euthanasia (the soft death) of all morality" (1997: AIX, p. 506) – but because he or she follows the principle of freedom of inner legislation. This entails that morality has been robbed of the foundation on which it could ground itself, that is, the hope for the remuneration for moral action that would arrive at some point. And by implication the pursuit of happiness loses the justification that was derived from its link to morality. Simmel (1989: p. 253) called this the "mutual independence of ethicality and happiness." 'If our ethical action,' Simmel wrote, "were only a detour on the path to happiness, then the point of our freedom would show itself again interwoven with the dependence on the powers of being, without whose favors no happiness could be complete."

The expulsion of the concept of happiness from moral philosophy has its counterpart in the eviction of happiness out of the realm of the political. For Kant, freedom has taken the place of happiness as the goal of politics – a concept driven further by romanticism, although setting out from the non-*eudaimonistic* ways of reasoning in transcendental philosophy. Kant's categorical imperative expels the concept of *eudaimonia* from some strands of philosophical discussion.

Redemption, Happiness, and Political Order

The emerging social sciences have started out from this rupture and have largely accepted it. In his theory of religiosity and morality, Durkheim reinterprets happiness and identifies it in religious experience, in moments of 'collective effervescence,' of self-transcendence and loss of self. Happiness is seen as the experience of being-one-with-oneself or of being submerged in the world, as a "feeling of strengthening" (Durkheim, 1981: p.

291), as the subjective 'oceanic feeling,' which Freud (1994) considers without any comprehension. Simmel, too, considers spiritual salvation (*Seelenheil*) as a state in which complete harmony with the world has been achieved, a state that shows the full liveliness and inner touchability of the perceived, the immediate being (Simmel, 1995). "Happiness, as a relation of transcendence, thus has two fundamental forms. It exists as the experience of either *oneness* or *otherness*, as the dissolution of all difference, or as the apprehension of, and unconditional surrender to, absolute difference" (Ferguson, 1992: pp. 11–12). Such experience articulates social integration and, again, connects the individual with the other and with a 'holy' collectivity, the ideal community of justice.

Within the main line of critical theorizing, that is, in Theodor W. Adorno and Max Horkheimer, the subjectivization of happiness is further radicalized. The foundational character of the aporias that result from the dualism of individual and society is rejected. Instead, the joint attack of liberalism (*bourgeois* ideology) and capitalism is held responsible for the destruction of the link between freedom and happiness.

Happiness is privatized in 'administered society' and under the reign of 'culture industry,' and it becomes the expression of societally determined false consciousness. Due to the horror of the circumstances, individual happiness becomes nothing but the dreadful arrangement of a petty-bourgeois idyll with its swollen sentiment. The *bourgeois* promise of happiness arising from property and possession becomes under conditions of inhumanity a curse. The promise of the commodity turns into pale appearance; in the totality of objectivizable socioeconomic relations, the connection between the individual, happiness, and society is interrupted and the promise of happiness only admitted as aesthetic experience (Adorno). But, within the same critical tradition, and by appropriating critically Freud's reflections on the pleasure principle and the reality principle, Herbert Marcuse (1955/1966) aims at developing a social theory that elaborates on a conception of possible happiness and permits the critique of both, the societal circumstances as well as the empirically existing conceptions of happiness. "The concept (of happiness) denotes a more-than-private, more-than-subjective condition; happiness is not in the mere feeling of satisfaction but in the reality of freedom and satisfaction" (Marcuse 1955/1966: pp. 103–104). The concept of happiness has to be liberated from bourgeois conformism and relativism; even the singular human beings cannot be 'judge of their own happiness' under these circumstances. It is precisely the task of critical analysis to show the 'possibility of a happier real constitution of humanity' beyond the existing 'false happiness.'

In another line of critical thinking, Walter Benjamin turns political time, directed toward happiness, into the key element of his philosophy of history. "The image of happiness … is tainted by time through and through … In other words, the imagination of happiness resonates inalienably with the one of redemption" (1991a: p. 693). Happiness is here not in the (past) event, but in its possibility, to which a 'weak messianic power' is given; it is recognizable only in its being missed historically (1991a: p. 694). "The order of the profane has to erect itself with the support of the idea of happiness" (1991b: p. 203) – and in the *Angelus Novus*, Benjamin finds the image for history in which the past, that which is complete (misfortune and misery), and the future, that which has remained incomplete (happiness), get reverted and in which misfortune and misery turn into something incomplete hoping for redemption (1991c: pp. 588–589). Ernst Bloch, too, reinstates the old connection between the individual, passion, and society; happiness becomes again an inalienable part of human being. "Because hope's grounding in the human drive for happiness is too difficult to be destroyed; and too evidently has it always been the engine of history" (1978: I, p. 515). Happiness, thus, even though pushed into the private realm since Kant, persistently returns into the political realm – as 'redemption,' 'the principle of hope,' as 'emancipation,' 'undamaged life,' and, in the version of existentialist philosophy, as 'authenticity.'

During the past few years, happiness has not only become the object of dubious empirical quantitative research on levels of satisfaction, but also again a theme of social philosophy, partly by referring to classical liberalism, partly also by returning to ancient traditions. Under terms such as justice, the good life, and self-realization, the question about concepts of the good life in modernity and the societal conditions under which happiness can unfold is again raised (Thomä, 2003; Thomä et al., 2011). The attempt is, thus, renewed to locate happiness not only in an 'accomplished' relation of the individual to her- or himself, but also in relation to others and to a just social order.

Hannah Arendt already held a view of political happiness, of 'public happiness,' that could not be equated with the obligation of the state to care for the welfare of its citizens (such a view of *bonheur*, in contrast, is, in her view, coresponsible for the loss of freedom in the course of the French Revolution). In the American Revolution, however, the connection between 'public happiness and political freedom,' which is an 'inalienable part of the structure of a republican community,' could be maintained at least for some period (Arendt, 1994: p. 178). The new connection between happiness and freedom occurs in the sign of plurality and a public sphere as guardians against both dangers; that namely substantive conceptions of happiness could suffocate freedom or that abstract and atomized freedom empties out any conception of happiness at all.

Drawing on the classical contract theories by John Locke and Jean-Jacques Rousseau and situating himself in relation to Immanuel Kant, John Rawls develops a 'theory of justice' (1971) for modern, pluralistic, and democratically constituted societies. For Rawls's theory, the 'right' enjoys primacy over the 'good,' since value-pluralistic societies will be able to agree only over a minimum set of foundations for what is right, but not over what is good. Rawls, thus, renounces any attempt to formulate a substantive theory of the good life. The starting point of his considerations is the hypothetical construction (the 'intuitive thought') of a society as a fair system of cooperation between free and equal persons. Since these considerations provide the general political framework for the 'good life as justice and fairness,' a human being can for Rawls be called "happy … during those periods when he is successfully carrying through a rational plan and he is with reason confident that his efforts will come to fruition" (Rawls, 1971: p. 550). Rawls ties his theory of justice to a critique of utilitarianism. For him the well-being of people does not depend on the degree to which individual preferences are realized. It is

important only that people are provided with the basic means to the same amount. As Martha Nussbaum, however, points out, Rawls did not consider what people do with these goods, to what they are enabled, and whether the space of enablements that thus opens up is not marked by striking inequalities.

Nussbaum (1990, 1993) herself resorts to Aristotle to sketch those political foundations that are to enable all people to lead a good life. She does not advocate any liberal relativism, but rather a universalism, based on an anthropological conception, that refers to needs and capacities which all people across all divides of culture, class, and sex have in common. Such basic needs constitute moral demands; they are an inalienable part of ethics and moral philosophy. At the same time, Nussbaum's perspective aims at pushing back the primacy of reason in moral philosophy, since moral principles had hitherto mostly been anchored in human reason and, as feminist critique remarked, had been considered predominantly an exclusive male capacity (Gilligan et al., 1989; Held, 2005; Noddings, 1984). Her ethics of virtue attempts to counteract the insufficient regard for the particularity of situations in moral theories that are based on general principles. In addition, it aims at integrating emotional dimensions into ethics. To define the concept of virtue more precisely, and not merely to understand it by recourse to rules, a 'superior instance' needs to be found; for Nussbaum this is a theory of the good. Starting out from the 'basic structure of the human form of life' (mortality, corporeality, cognitive abilities, practical reason, sociality, etc.), she develops an 'open list' of the basic capacities of human beings and of the good. She herself characterizes the list as 'vague,' which means that it is open for concretization in each specific cultural context, and then puts specific demands toward politics. The objective of measures taken by the state is that they enable the citizens to choose a good form of leading their lives, to develop their capacities, and to live in accordance with practical reason and with their social ties.

In his attempt at clarifying the relations between a moral life and a good life, Seel (1995) similarly defends a "universalism of the good," which is not linked to particular forms of society. This universalism, however, is no essentialism that claims the existence of values independent of human volition or raises anthropological facts to the status of values or norms. Seel develops a position that he calls "reflexive *eudaimonism*" and asks what it entails, from the viewpoint of different people, "to sensibly lead a succeeding life" (p. 10), defining a 'good life' formally on the basis of moral respect, social recognition, and political legitimacy. The incarnation of good human life is a "life, open to the world and self-determined, the moral and legal protection of which includes a respect for all individuals who are able to have a (however determined) good life" (p. 11).

'Happiness,' 'the good life' are, thus, always also tied to sociocultural circumstances and the social *imaginaire* (Friese, 2011). It remains the task of philosophy as well as of the social sciences to think the – political – conditions under which a good life and happiness stay open as a possibility. Happiness as the relation to oneself, to others, and to the community thus, belongs to reflection, to experience, and to the dream (Mauzi, 1969).

See also: Arendt, Hannah (1906–75); Critical Theory; Frankfurt School: Institute for Social Research; Justice: Philosophical Aspects; Kant, Immanuel (1724–1804); Kantian Ethics and Politics; Morality 'East' and 'West': Cultural Concerns; Philosophy, Psychiatry, and Psychology; Self: History of the Concept; Subjective Wellbeing, Psychology of; Utilitarian Social Thought, History of; Virtue Ethics.

Bibliography

Arendt, Hannah, 1963/1994. On Revolution. Penguin, Harmondsworth, UK.
Aristotle, 1934. The Nicomachean Ethics. Harvard University Press, Cambridge, MA.
Benjamin, Walter, 1991a. Über den Begriff der Geschichte. In: Tiedemann, R., Schweppenhäuser, H. (Eds.), Gesammelte Schriften, vol. I2. Suhrkamp, Frankfurt am Main, Germany, pp. 691–704.
Benjamin, Walter, 1991b. Theologisch-politisches fragment. In: Tiedemann, R., Schweppenhäuser, H. (Eds.), Gesammelte Schriften, vol. I2. Suhrkamp, Frankfurt am Main, Germany, pp. 203–204.
Benjamin, Walter, 1991c. Das Passagen-Werk. In: Tiedemann, R., Schweppenhäuser, H. (Eds.), Gesammelte Schriften, vol. V. Suhrkamp, Frankfurt am Main, Germany, pp. 1–2.
Bloch, Ernst, 1978. Das Prinzip Hoffnung, vols. 1–2. Suhrkamp, Frankfurt am Main, Germany.
Caillé, Alain, Lazzeri, Christian, Senellart, Michel (Eds.), 2001. Histoire raisonnée de la philosophie morale et politique. Le bonheur et l'utile. La Découverte, Paris.
Destrée, Pierre (Ed.), 2003. Aristote. Bonheur et vertus. Puf, Paris.
Durkheim, Emile, 1912/1981. Les formes élémentaires de la vie réligieuse. Alcan, Paris.
Ferguson, H., 1992. Religious Transformation in Western Society: The End of Happiness. Routledge, London.
Freud, Sigmund, 1994. Das Unbehagen in der Kultur. Fischer, Frankfurt am Main, Germany.
Friese, Heidrun, 2011. Geglücktes Leben. Europäischer Universitätsverlag, Bochum/Berlin, Germany.
Garver, Eugene, 2011. Aristotle's Politics: Living Well and Living Together. Chicago University Press, Chicago.
Gilligan, Carol, Ward, Janie, McLean Taylor, Jill, Bardige, Betty (Eds.), 1989. Mapping the Moral Domain: A Contribution of Women's Thinking to Psychological Theory and Education. Harvard University Press, Cambridge, MA.
Heidegger, Martin, 1992. Platon: Sophistes. Gesamtausgabe, vol. 19. Klostermann, Frankfurt am Main, Germany [1997. Plato's Sophist. (R. Rojcewicz, A. Schuwer, Trans.). Indiana University Press, Bloomington, IN].
Held, Virginia, 2005. The Ethics of Care: Personal, Political, and Global. Oxford University Press, Oxford.
Kant, Immanuel, 1781/1974a. Kritik der reinen Vernunft. In: Weischedel, W. (Ed.), Werkausgabe, vols. III/IV. Suhrkamp, Frankfurt am Main, Germany [Critique of Pure Reason, Politis, V. (Ed. and Trans.) based on Meiklejohn 1993. Everyman, London].
Kant, Immanuel, 1785/1993b. Grundlegung zur Metaphysik der Sitten. In: Weischedel, W. (Ed.), Werkausgabe. Suhrkamp, Frankfurt am Main, Germany, pp. 11–102.
Kant, Immanuel, 1788/1993a. Kritik der praktischen Vernunft. In: Weischedel, W. (Ed.), Werkausgabe, vol. VII. Suhrkamp, Frankfurt am Main, Germany, pp. 107–302.
Kant, Immanuel, 1790/1974b. Kritik der Urteilskraft. In: Weischedel, W. (Ed.), Werkausgabe, vol. X. Suhrkamp, Frankfurt am Main, Germany.
Kant, Immanuel, 1798/1997. Die Metaphysik der Sitten. In: Weischedel, W. (Ed.), Werkausgabe, vol. VIII. Suhrkamp, Frankfurt am Main, Germany.
Marcuse, Herbert, 1955/1966. Eros and Civilization. Beacon Press, Boston.
Mauzi, Robert, 1969. L'idée du bonheur dans la littérature et la pensée françaises au XVIIIe siècle. Colin, Paris.
Nagel, Thomas, 1980. Aristotle on eudaimonia. In: Rorty, A.O. (Ed.), Essays on Aristotle's Ethics. University of California Press, Berkeley, pp. 7–14.
Noddings, Nel, 1984. Caring: A Feminine Approach to Ethics and Moral Education. University of California Press, Berkeley, CA.
Nussbaum, Martha C., 1990. Aristotelian social democracy. In: Douglas, R.B., Mara, G., Richardson, H. (Eds.), Liberalism and the Good. Routledge, New York, pp. 203–252.

Nussbaum, Martha C., 1993. Non-relative virtues: an Aristotelian approach. In: Nussbaum, M., Sen, A. (Eds.), The Quality of Life. Clarendon Press, Oxford, UK.
Nussbaum, Martha C., 2001. The Fragility of Goodness. Luck and Ethics in Greek Tragedy and Philosophy (Revised Edition). Cambridge University Press, Cambridge.
Rawls, John, 1971. A Theory of Justice. Belknap Press of Harvard University Press, Harvard, MA.
Seel, Martin, 1995. Versuch über die Form des Glücks. Studien zur Ethik. Suhrkamp, Frankfurt am Main, Germany.
Simmel, Georg, 1989. Die Lehre Kants von Pflicht und Glück. In: Rammstedt, O. (Ed.), Gesamtausgabe, vol. 7. Suhrkamp, Frankfurt am Main, Germany, pp. 247–254.
Simmel, Georg, 1995. Die Religion. In: Rammstedt, O. (Ed.), Gesamtausgabe, vol. 40. Suhrkamp, Frankfurt am Main, Germany, pp. 39–118.
Thomä, Dieter, 2003. Vom Glück in der Moderne. Suhrkamp, Frankfurt am Main, Germany.
Thomä, Dieter, Henning, Christoph, Mitscherlich-Schönherr, Olivia (Eds.), 2011. Glück. Ein Interdisziplinäres Handbuch. Metzler, Stuttgart, Germany.
Wolf, Ursula, 1996. Die Suche nach dem guten Leben. Platons Frühdialoge. Rowohlt, Reinbek, Germany.

Relevant Websites

http://www.encyclopedia.com/topic/ethics.aspx#5-1O101:Ethics-full – Entry 'Ethics.'
http://www.iep.utm.edu/virtue – The Internet Encyclopedia of Philosophy.
unsdsn.org/wp-content/…/2014/…/WorldHappinessReport2013 – World Happiness Report.

Eugenics as a Basis of Population Policy

Garland E Allen, Washington University in St. Louis, St. Louis, MO, USA
Marius Turda, Oxford Brookes University, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G.E. Allen, pp. 4889–4896, © 2001, Elsevier Ltd.

Abstract

This article considers the eugenic basis of population policy by reviewing national and international developments during the twentieth century. Examples from the history of eugenics are invoked in order to contextualize this topic both diachronically and conceptually.

Population policy refers to control of human population by legislation or other official means. It involves setting goals for such population parameters as total size, the maintenance of demographic stability, and birth and mortality rates. Population policy includes both privately funded programs, such as the contraception campaign promoted in India in the 1950s by the Rockefeller Foundation, and publicly funded programs such as those promoted throughout the Third World by the United States Agency for International Development (USAID). There are also state-controlled population policies, such as the natalist programs in Ceausescu's Romania, the sterilization of Roma women in Czechoslovakia and, more recently, of indigenous women in Fujimori's Peru as well as China's current restrictive family planning. Often, there are religious aspects involved in certain population policies such as the Catholic Church's traditional opposition to contraception and abortion. Finally, there are social and economic factors that shape population policy, most notably voluntary sterilization in Brazil.

In the twentieth century, population policy has often been associated with eugenics. The term 'Eugenics,' derived from the Greek *ευγενής*, or *eugenes*, and was first coined by English statistician Francis Galton (1822–1911) in his book *Inquiries into Human Faculty and Its Development* to refer to one born 'good in stock, hereditarily endowed with noble qualities' (Galton, 1883). As an intellectual and social movement, eugenics came to mean, in the words of one of its strongest American supporters Charles B. Davenport (1866–1944), "the science of human improvement by better breeding" (Davenport, 1910). For both Galton and Davenport, better breeding implied improving the biological quality of the human species by the scientific principles of heredity. Eugenicists wanted to develop policies that dealt not only with the quality of the population but also its growth. In this sense they sought to increase the reproductive rate of that segment of the population deemed to be biologically (meaning genetically) well-endowed (positive eugenics), and limit or prevent the reproduction of that segment of the population deemed to be biologically inferior or degenerate (negative eugenics).

Eugenics was essentially based on biological, or genetic determinism (the idea that genes determine more of what a person is like, both in personality and in physical traits, than environment; eugenics took the side of 'nature' in the long-standing nature–nurture debate). The largely ignored model of heredity Gregor Mendel (1882–1884), the Moravian monk, had published as early as 1866 – subsequently known as the system of Mendelian inheritance – challenged the theory of blending inheritance (namely that an offspring's makeup constituted a blend of his or her parents' traits). Mendel's research on the nature of inheritance proved that an offspring's genetic structure was not simply a synthesis of that of the parents, but that particular genetic traits could be either dominant or recessive, and that these were passed on in accordance with statistical laws.

By the time of World War I the Mendelian principles of heredity were being applied with great success to the breeding of many varieties of animals and plants; it also revolutionized theories of human genetics and evolution. It was in this atmosphere that eugenics flourished. It was hailed as a scientific solution to recurrent social and biological problems. If alcoholism, criminality, pauperism, and feeblemindedness were all determined by genes, then the way to eliminate these problems would be to prevent the reproduction of those individuals thought to carry those genes. While eugenicists were concerned with the inheritance of physical abnormalities such as hemophilia, color blindness, brachydactyly (shortened digits), and debilitating diseases such as Huntington's chorea (all recognized to be genetic prior to 1910), they were ultimately more interested in human behavioral and personality traits – including mental ability, mental illness, 'moral degeneracy,' and criminality, to list only a few – all of which they claimed were genetically determined to a significant degree.

Encouraged by the reception to Galton's ideas in Britain, the Eugenics Education Society was established in 1907 in London. Other countries followed this example and in Scandinavia, for instance, a Swedish Society for Race Hygiene was formed in Stockholm in 1909. American eugenicists had established the Station for Experimental Evolution at Cold Spring Harbor as early as 1904, followed by the Eugenics Record Office in 1910. The French Eugenics Society was established in 1912. In Central Europe, Czech, Austrian and Hungarian societies were established in 1913 and 1914. Also in 1914, an Italian Eugenics Committee convened at the Roman Society of Anthropology. By then, the German Society for Racial Hygiene and the German League for National Regeneration and Heredity, both established in 1905, were intensely involved in disseminating eugenics to the general public, both at home and abroad, as illustrated by the Hygiene Exhibition organized in Dresden in 1911. However, it was in the United States, Britain, the Scandinavian countries, and Germany where eugenics as an

intellectual and social movement reached its greatest strides and from eugenicists' point of view, greatest ideological and political results (Adams, 1990).

Eugenic Research: Its Aims and Problems

Following Galton's research into the genealogy of illustrious British individuals, the family pedigree chart became the main analytical means for displaying and analyzing data on the inheritance of one or another trait. Information was collected from among a group of relatives, and it was displayed as a family pedigree chart, emphasizing that traits ran in families and thus how heredity influenced many behavioral, personality, and mental traits. The American eugenicist Charles Davenport claimed, for example, that thalassophilia, or 'love of the sea,' was a sex-linked Mendelian recessive appearing in the families of prominent US naval officers (Davenport, 1919). That the condition must be sex-linked was clear to Davenport, since in pedigree after pedigree only males in the various families he studied became naval officers. Such simplistic Mendelian arguments were then extended from individuals to large groups as a way of explaining the differences between racial, ethnic, and national groups.

Because of their concern with the apparent rise in 'feeblemindedness,' eugenicists embraced the nascent psychometric movement (intelligence testing) starting as early as 1912. In that year, Henry H. Goddard (1866–1957) administered a version of the French Binet–Simon test to immigrants arriving at New York's Ellis Island (Zenderland, 1998). Although the Binet–Simon test was intended to measure only an individual's mental functioning at a given point in time, Goddard and a host of psychometricians in the United States and Europe considered that it also measured innate, or genetically determined intelligence. Goddard coined the term 'feeblemindedness' (which he claimed was genetic) to refer to those people who scored below a given level (70) on his tests, and claimed that it was transmitted from generation to generation. It was also Goddard who carried out the famous study of the Kallikak family, showing what was presumed to be the hereditary effects of feeblemindedness and degeneracy (Goddard, 1912). For American eugenicists, the new mental tests seemed as a precise, quantitative tool for measuring an otherwise elusive, but fundamental human trait. Eugenically inclined psychometricians felt that the scientifically trained expert could at last solve with rational and scientific methods the social problems that muddle-headed politicians and social workers had been unable to solve during the preceding century.

A major problem with eugenic research was that it quickly fell behind the work being done by geneticists in laboratories throughout the world. By the 1920s many geneticists had come to recognize that genes (collectively, known as the organism's genotype) interact with one another (a phenomenon known as epistasis) and with the environment to produce any given trait (the organism's phenotype). The interaction of genes with one another and environmental factors (for example, temperature, drugs, diet) was found to alter the gene's expression in profound ways, yielding a variety of alternative phenotypes from the same genotype. Even the position of a gene along the length of its chromosome turned out to affect the way that gene is expressed. This malleability, or plasticity, of the genotype–phenotype relationship meant that even knowing the genotype of the offspring at conception did not provide a certain means of predicting the phenotype. While they adopted much of the exciting new work in genetics in the 1920s and 1930s, it is apparent that many eugenicists did not really understand it fully, or take to heart its implications for their social programs.

Eugenics and the Control of Reproduction

One of the aims eugenicists across the world shared was the desire to allegedly improve the population's quality by controlling its reproduction. Eugenic sterilization is one of methods used toward achieving this goal. In the United States, Harry H. Laughlin (1880–1943) was particularly active in lobbying for the passage of a number of state eugenic sterilization laws in the 1920s and 1930s. Indeed, Laughlin drew up a 'Model Sterilization Law' that served as a prototype from which each state could derive its own modified version (Laughlin, 1922). Although the earliest such laws had been passed before World War I, the majority was established during the interwar period (Reilly, 1992). Eugenic sterilization was aimed specifically at those individuals in mental or penal institutions who, from family pedigree analysis, were considered likely to give birth to 'socially defective' offspring. Sterilization could be ordered only after a patient had been examined by a eugenics committee, usually composed of a lawyer or family member representing the individual, a judge, and a doctor or other eugenic 'expert' (Reilly, 1992). Between 1907 (when the first such law was put into effect in Indiana) and 1941, over 38 000 eugenic sterilizations were performed in the United States, with California leading all other states (Reilly, 1992). By the early 1960s, the number had risen to some 60 000. The most famous of the sterilization cases in the United States was that of Carrie Buck in Virginia. When Virginia passed its eugenic sterilization law in 1924, a test case was arranged the next year to determine if the law was constitutional. Buck vs. Bell was tried in 1925 in the Virginia Circuit Court (Lombardo, 1985). When the lower court ruled in favor of the law, an appeal was sent to the Supreme Court of the United States, where Justice Oliver Wendell Holmes wrote the majority report. In upholding the lower court ruling, Holmes made his oft-repeated assertion: "Three generations of imbeciles are enough."

In Europe, the debate on eugenic sterilization focused on two aspects: what social and medical categories would be subjected to sterilization, and what legislative forms such policies would require, namely, whether they would be voluntary or compulsory? In 1928, the Canton of Vaud in Switzerland authorized the sterilization of those suffering from hereditary mental diseases and those deemed feebleminded. A year later, the Danish Sterilization Law was introduced. Sterilization was contingent upon the medical diagnosis provided by a physician and upon the consent of the individual, the family or the legal guardian. A Committee for Legalizing Eugenic Sterilization was formed in Britain in 1929, and a bill for voluntary sterilization of mental defectives was introduced to Parliament in 1931. The campaign was, however,

unsuccessful as British eugenicists were divided over which preventive methods was most efficient, sterilization or institutional segregation. Amidst the growing acceptance of negative eugenics amongst Christian states in Europe and beyond, Pope Pius XI's issued the Encyclical on Christian Marriage, *Casti Connubii* in 1930. The encyclical castigated the prevention of 'unworthy' life advocated by eugenicists both as an expression of excessive secularization and of the state's interference in the individual and family's private sphere. The opposition of the Catholic Church notwithstanding, eugenic sterilization remained a popular method of population control during the 1930s: it was also adopted by the Baltic states and was discussed in Eastern Europe in countries like Hungary, Romania, and Poland (Turda and Weindling, 2007).

It was, however, in Nazi Germany after 1933 that eugenics served as the main basis of population policies (Weiss, 1990). The Law for the Prevention of Progeny with Hereditary Diseases was officially announced on 14 July 1933, and became effective on 1 January 1934. In theory, it stipulated that "anyone with hereditary diseases may be rendered sterile by surgical means, when, according to medical experience, it is highly probable that the offspring of such person will suffer from severe inherited mental or bodily disorders." In practice, however, several medical categories were outlined, namely: (1) hereditary feeblemindedness; (2) schizophrenia; (3) manic-depressive insanity; (4) hereditary epilepsy; (5) Huntington's chorea; (6) blindness; (7) deafness; and (8) severe hereditary malformation (Gütt et al., 1934). Chronic alcoholics were also subjected to the law. Hereditary Health Courts were established in 1934 to supervise the implementation of sterilization.

Other countries, such as Sweden, emulated the German sterilization model and introduced compulsory sterilization of persons suffering from insanity, feeblemindedness, and other mental disorders in 1934. Similar legislation followed in Norway (1934), Finland (1935), and Estonia (1936), allowing for compulsory sterilizations on medical, social, and eugenic grounds. In other countries, like France, for example, the Nazi model of compulsory sterilization was viewed with suspicion due to its totalitarian and racial aims. Akin to some Italian and Romanian eugenicists, French eugenicists also distinguished between a 'Latin eugenics' understood to focus more on improving the social environment and education, and an 'Anglo-Saxon eugenics' seemingly preoccupied with negative prevention, selective breeding, and racial protectionism.

The empowerment of the state to enact population and eugenic policies was not restricted to the United States, Nazi Germany, or Fascist Italy. Demand for a eugenically predisposed state was vast and extensive, coloring the outlook of eugenicists of various political orientations, such as anarchists in Spain, socialists in Britain, feminists and Social Democrats in Germany and Sweden, and health educators in Turkey. A number of Fabians, including luminaries such as G.B. Shaw (1856–1950) and H.G. Wells (1866–1946) argued that an effective and comprehensive eugenic program could best be pursued in a socialist state. This was also an idea entertained by the American eugenicist Herman J. Muller (1890–1967), who believed that Stalin would be the architect of a new humanity, one in which eugenic rationality would reign supreme.

Eugenics as Population Policy

A qualitative population policy, able to accompany quantitative measures, was vital for the success of eugenics; moreover, this task was a national duty for both the state and the eugenicists, especially during the interwar period. It was World War I that brought population policy (known as *Bevölkerungspolitik* in the German-speaking countries) and eugenics together. Examples include the German and the Austrian Societies for Population Policy and the Hungarian Society of Racial Hygiene and Population Policy. In the wake of the dramatic demographic changes brought about by the war, population experts reviewed the relationship between marriage and illegitimate children, the importance of nutrition and the distribution of arable land and unemployment. All these guidelines and suggestions for practical measures in the field of eugenics and population policy were supposed to provide Europe's much-needed demographic growth. In many respects, Italy's fascist welfare program served as the model. The National Organisation for the Protection of Motherhood and Infancy, established in 1925, emphasized positive incentives to increase the quantity and quality of the population. Fascist demographers and eugenicists attempted to unify maternal and child welfare with rural development schemes, offering in fact a nationwide program of social and biological engineering.

These programs gradually took over during the 1920s and 1930s, placing natalist ideas of demographic growth and large families at the center of eugenic thinking. In each European country, eugenic population policies assumed various overtones according to national circumstances. In the United States, the biologist Raymond Pearl (1879–1940) claimed that eugenicists had focused too much attention on the quality of offspring from particular (usually 'degenerate') families, and had lost sight of the equally important issue of the rapid birth rate of the biologically 'degenerate' at home and abroad. In a series of scientific and popular articles he began to articulate the ideology of population control. While critical of much of the current eugenic propaganda, Pearl did not disavow basic eugenic principles. He simply emphasized the importance of quantitative as well as qualitative considerations. In Central and Eastern Europe, for instance, it was the state that was particularly dedicated to population policies, such as those of resettlement, meant to strengthen the racial character of those regions perceived to be either depopulated or under threat by other ethnic groups.

By the year 1940, eugenic population policies became the norm in Europe, as exemplified by the Bulgarian Law for the Protection of the Nation, the National Fund for the Protection of Family and Population in Hungary, the Law for the Protection of the Race adopted by the regime of Widkun Quilsing in Norway, or the French Law for the Protection of Mothers and Children. Undeniably, these laws advocated the same political protectionism of the dominant racial group and the exclusion of ethnic minorities, especially Jews, but they also, like in the Bulgarian and French cases, introduced sanitary screening and mandatory premarital examination, aiming at a general biotypological investigation of the nation. Facing World War II, the eugenic rhetoric of the 1940s intensified in its racial tone. Eugenicists in Hungary and in France, for instance, profited from this ideologically favorable environment to establish racial institutes, like the Hungarian Institute for National

Biology (established in 1940) and the French Foundation for the Study of Human Problems (created in 1941).

Maintaining the nation's demographic potential thus became of prime political importance alongside instruments for eliminating the 'dysgenic groups,' be they defined socially or ethnically. By investing the state with the powers for protection of the family and the nation, a new eugenic ethos was devised, whereby ideas of social and biological improvement were combined in innovative ways. On the one hand, social hygiene, public health, and family welfare were based on the broad concept of eugenics; on the other, political reform and economic development became coterminous with the preservation of the race.

In the aftermath of World War II, the United States emerged as a contender for global power, inheriting the mantle of international economic and political involvement that had previously been carried by the British and, to a lesser extent, the French. In 1952, President Dwight D. Eisenhower formed a blue-ribbon panel, headed by David Rockefeller of the Chase National (later Chase-Manhattan) Bank, to evaluate the future prospects for US industrial and commercial expansion around the world. In the same year, the Rockefeller Foundation organized a conference in Williamsburg, Virginia, under the direction of John D. Rockefeller III, to study what Pearl had highlighted as the 'population problem.'

The blue-ribbon panel concluded that the single greatest obstacle to US industrial expansion (meaning access to cheap raw materials, labor, and ultimately markets) was the rapid population growth occurring in most Third World countries. Such population increases not only required the retention of more of countries' raw materials (including agricultural products) at home, but also fueled local revolutionary movements aimed at eliminating foreign control. Not surprisingly, the Williamsburg conference called for the formation of an organization to look into methods for controlling population growth. Out of this meeting came the proposal for establishing the Population Council, with Rockefeller funds and John D. III himself as President. Frederick Osborn (1889–1981), still Secretary of the American Eugenics Society at that time, soon took over as Council President, running both organizations out of his office in Rockefeller Center in New York.

One of the first major projects of the Council was the funding of investigations of various forms of birth control, including the intrauterine device (IUD), and eventually, to the first birth-control pill (Watkins, 1998). Initial trials in Puerto Rico beginning in April 1956, led to field studies in other Third World countries such as India and Pakistan. One of the major incentives used to promote these studies was the economic benefits that would accrue to families that had fewer children. That such pilot population control movements encountered considerable resistance in the countries where they were introduced, suggested that local people saw the matter quite differently. The programs provided a lesson in the cultural and social components of reproductive life. For many in agricultural, Third World countries, large families were the only way to insure some manner of economic survival. Children worked on farms at an early age, and thus more than earned their 'keep,' while large families often provided the only form of old-age insurance in countries that lacked programs for the elderly. This lesson is still one to be learned by leading political and many environmental and conservation groups, who continue to advocate population control as the major agenda for Third World countries. Today, for example, the USAID, a branch of the State Department, devotes a significant portion of its budget to its Population Office. Programs emerging from USAID support population control programs in over 20 Third World countries.

Criticism of Eugenics, 1925–50

Almost from the beginning, many of the basic premises of eugenics were brought under critical scrutiny by scientists and the general public alike. Among the first and most important critics from the biological side was Thomas Hunt Morgan (1866–1945), the *Drosophila* geneticist at Columbia, in 1933 the first recipient of the Nobel Prize for work done in genetics, and one of the earliest members of the Committee on Animal Breeding of the American Breeders' Association (a Committee that had taken on itself organizing the first eugenic activities in the United States). In a personal letter sent to his close friend Davenport in 1915, Morgan resigned from the Committee, explaining that the claims being made on behalf of eugenics were exaggerated beyond what any scientific evidence could support (Allen, 1978). Although Morgan did not criticize eugenics in print at that time, from 1925 on, starting with his book *Evolution and Genetics* (Morgan, 1925) he publicly chastised eugenicists for lumping many mental/behavioral conditions together under a rubric like 'feeblemindedness' and treating them as if it had a single underlying cause. Especially critical, he noted, was the fact that personality and social traits were the most malleable of our characteristics, and hence the most difficult in which to tease apart the relative contributions of heredity and environment:

> The important point is that mental traits in man are those that are most often the product of the environment which obscures to a large extent their inheritance, or at least makes very difficult their study. (Morgan, 1925)

The pedigree charts on which Mendelian analyses were based, were also strongly criticized by Morgan and others as giving the impression of tracing genetically determined transmission when, of course, they equally well could be interpreted as showing culturally determined effects:

> The pedigrees that have been published showing a long history of social misconduct, crime, alcoholism, debauchery and venereal disease are open to the same criticism from a genetic point of view; for it is obvious that these groups of individuals have lived under demoralizing social conditions that might swamp a family of average persons. It is not surprising that, once begun from whatever cause, the effects may be to a large extent communicated [socially] rather than inherited [genetically]. (Morgan, 1925)

In other words, pedigree charts do not separate 'nature' from 'nurture.'

Other geneticists followed suit. Raymond Pearl also voiced public criticism of eugenicists' claims, especially those

regarding the supposed biological inferiority of various racial and ethnic groups (Pearl, 1927). Along with his colleague Herbert S. Jennings (1868–1947) at Johns Hopkins University, Pearl felt that propagandists like Laughlin, Madison Grant, and others made claims that went far beyond any reasonable evidence (Barkan, 1992). Jennings devoted the better part of the ensuing decade (after 1924) to lecturing all over the United States about the dangers of oversimplified version of eugenics. Muller, a student of Morgan's, and subsequently himself to become a Nobel laureate in genetics, also delivered a searing attack on 'old style' eugenics at the Third International Eugenics Congress in New York City in 1932. Muller, who still harbored strong eugenic beliefs, argued that until the economic and social environment could be equalized it would be impossible to know how much of any individual's 'social inadequacy' was due to heredity and how much to environment (Muller, 1933). Much of feeblemindedness might be due to poor maternal diet during pregnancy or during the early years of childhood, or to maternal drug and alcohol use; although these effects might appear at a superficial glance to be genetic, they had nothing to do with defective germplasm. English mathematician and sometime eugenics critic Lancelot Hogben (1895–1975) made one of the clearest statements at that point of time about the oversimplified concept of genetics that informed much of the eugenics movement: "No statement about a genetic difference has any scientific meaning unless it includes or implies a *specification of the environment* in which it manifests itself in a particular manner" (Hogben, 1933: p. 14).

In 1934, a special blue-ribbon committee of the British government published the Brock Report (named for its chairman, Laurence B. Brock, head of the Joint Committee on Mental Deficiency), outlining difficulties in determining the relative contribution of genetics to mental deficiency. Among others, the Committee included Ronald A. Fisher, himself a staunch eugenicist but also one of the world's foremost statisticians and population geneticists. The Brock Report agreed with the criticisms of eugenics put forward a few years earlier by Lionel Penrose, who had demonstrated that the category of 'feeblemindedness' was ill-defined; was used carelessly and indiscriminately to cover a host of mental problems that were caused in many cases by disease and social deprivation; and for the most part little was known of the genetic cause of mental disability (Kevles, 1985).

An even more comprehensive and devastating critique of the whole range of eugenic claims was published by the American Neurological Association in the mid-1930s, just a few years after the appearance of the Brock Report (Committee of the American Neurological Association, 1936). Work for the book was funded by the Carnegie Foundation through the New York Academy of Sciences. Like the Brock Report, the findings of the Committee found that by the mid-1930s abundant evidence threw into doubt such well-accepted eugenic beliefs as: (1) The vast number of mental and moral/personality traits are genetically determined, (2) Environment (upbringing) plays a relatively minor role in the development of mental and personality traits, and (3) Sterilization is the most effective means of reducing the frequency of socially inadequate behavior in future generations.

The Committee of the American Neurological Association supported many of the general conclusions of the Brock Report. Citing the work of Morgan, Jennings, and others in the United States, and that of J.B.S. Haldane in England, the authors argued that we do not know 'enough about the mental diseases and the effects of the social structure to manipulate in a successful way the procreation of the race' (Committee of the American Neurological Association, 1936). The level of genetic knowledge was too slight to formulate any social policies or carry out any specific manipulations guaranteed to alter the genetic basis of future generations. More specifically, it was found that eugenicists consistently downplayed the role of the environment in the development of mental and behavioral problems, even when it seemed obvious that social and economic conditions must be playing a highly significant role. Herbert S. Jennings made the same point, emphasizing that 'environment' included the internal, cellular, as well as external environment in affecting how genes were expressed. Even if a person is clearly shown to have a genetic defect, as in hereditary thyroid deficiencies, Jennings argued, the condition can be treated with dietary and/or drug supplements, which can restore normal function. A whole host of human anatomical and physiological conditions that were clearly genetic did not need to be debilitating or limiting to the individual, according to Jennings, because in many cases we possess the technological means to restore, or at least compensate for, the genetic defect. Genetics, according to Jennings, was not destiny (Jennings, 1927). For eugenicists, however, it was, and this remained one of the underlying assumptions of hereditarian thinking from the past to the present.

On the eve of World War II (literally 1 week before), another strong criticism of eugenics emerged as the 'Geneticists' Manifesto,' a document signed in early August 1939, by a group of geneticists assembled in Edinburgh at the VIIth International Congress of Genetics. Among the major initiators were J.B.S. Haldane (1892–1964) and Lancelot Hogben, and H.J. Muller. The Manifesto put forth Muller's argument that from a genetics point of view, as long as vast disparities in wealth and access to resources continued to exist in western societies, no claims could be made about the relative contribution of heredity over environment to human mental capacity and social behavior (Chase, 1977). They condemned racialist thinking, especially the attempt to rank races with respect to one another on the basis of genetic worth. The geneticists who initiated and signed the Manifesto were among the most well known and respected workers in the field during the twentieth century.

Despite these severe criticisms, eugenics continued to thrive in the Scandinavian countries and especially in Nazi Germany (between 1933 and 1945). After passage of the sterilization laws in 1934, the Nazi government enacted the infamous Nüremberg Laws in 1935, prohibiting marriage between Aryans and Jews and other 'non-Nordic' groups. Doctors were required to report to the Ministry of Health cases of what they deemed to be hereditary diseases among their patients, and after the large-scale deportation to concentration camps, became involved in a variety of genetic experiments on twins. All in all, German population experts carried out eugenic principles further than in any other country of Western Europe

or North America, notwithstanding the condemnation from other European eugenicists.

In the period after World War II overt criticisms of eugenics as the basis for population policy became more widespread. Revelation of Nazi atrocities committed in the name of eugenics played an important role in this change of attitude. In addition, hard-line eugenicists such as Davenport, Laughlin, Grant, and others on the American scene were retired or deceased by the end of the war, and were replaced by what historian Daniel J. Kevles has dubbed moderate, or reform eugenicists.' Although it might be questioned how really 'reformed' these eugenicists were, it is clear they were more knowledgeable about the nuances of what traits genes might and might not control, less overtly racist, and more modest in their claims for what eugenics might accomplish. Partly because of the criticism, it was difficult for the field of human genetics to get a fresh start after the war, and many geneticists shied away from dealing with human heredity (Kevles, 1985: 193ff; Ludmerer, 1972). Textbooks and other writings on human genetics were careful to distance themselves from old style eugenic thinking (Ludmerer, 1972). Laurence H. Snyder's *Human Genetics* (1951), for example, stated clearly that "[w]e should hesitate at accepting without careful analysis the claims and recommendations that are present in a large portion of the eugenics literature." An understanding of the historical development of the eugenics movement, Snyder went on to say, would suggest, "why professionally trained geneticists, on the whole, have given relatively little support to eugenic theories and programs" (quoted in Ludmerer, 1972). Anthropologist and prolific popular writer Ashley Montague echoed the same feeling in his 1959 *Human Heredity* when he wrote that much unfortunate public policy – by which he meant immigration restriction and state sterilization laws – had been based on "the ill-considered, prejudiced, and unscientific judgments of the late Harry Laughlin … Madison Grant … and Charles Benedict Davenport" (Montagu, 1959: pp. 25–26). And, in 1960, the American Medical Association came out with a report, 'A Reappraisal of Eugenic Sterilization Laws,' that called into question virtually all the assumptions of the old eugenic movement. In the decisive debates over the excesses and disastrous effects of eugenic dreams of perfectibility (particularly since the 1950s), the problematic connotations of population policy for the individuals and communities involved are ever-present. Understanding the eugenic basis for population policy is, therefore, essential in both its historical and comparative perspective.

See also: Bioethics: Philosophical Aspects; Darwinism; Eugenics as an International Movement; Eugenics in Asia; Eugenics in Europe; Eugenics, History of; Genetics and Society; Genetics: The New Genetics; Genomics, Ethical Issues In; Nazi Law; New Genetics and Race; Race: Genetic Aspects; Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective.

Bibliography

Adams, M. (Ed.), 1990. The Wellborn Science: Eugenics in Germany, France, Brazil and Russia. Oxford University Press, New York.

Allen, G.E., 1978. Thomas Hunt Morgan: The Man and His Science. Princeton University Press, Princeton, NJ.

Barkan, E., 1992. The Retreat of Scientific Racism: Changing Concepts of Race in Britain and the United States between the World Wars. Cambridge University Press, Cambridge, New York.

Chase, A., 1977. The Legacy of Malthus. Alfred A. Knopf, New York.

Committee of the American Neurological Association, 1936. Eugenical Sterilization: A Reorientation of the Problem. Macmillan, New York.

Davenport, C., 1910. Eugenics. Henry Holt, New York.

Davenport, C., 1919. Naval Officers: Their Heredity and Development. Carnegie Institution of Washington, Washington, DC.

Galton, F., 1883. Inquiries into Human Faculty and Its Development. Macmillan, London.

Goddard, H.H., 1912. The Kallikak Family: A Study in the Heredity of Feeble–Mindedness. Macmillan, New York.

Gütt, A., Rüdin, E., Ruttke, F., 1934. Gezetz zur Verhütung erbkranken Nachwuchses vom 14. Juli 1933 mit Auszug aus dem Gesetz gegen gefährliche Gewohnheitsverbrecher und über Massregeln der Sicherung und Besserung vom 24. Nov. 1933. J. F. Lehmann, Munich.

Hogben, L., 1933. Nature and Nurture. George Allen, London.

Jennings, H.S., 1927. Public health progress and race progress – are they incompatible? Science 66, 45–49.

Kevles, D.J., 1985. In the Name of Eugenics. Alfred A. Knopf, New York.

Laughlin, H.H., 1922. Eugenical Sterilization in the United States. Pscyhopatic Laboratory of the Municipal Court of Chicago, Chicago.

Lombardo, P.A., 1985. Three generations no imbeciles: new light on Buck v. Bell. New York University Law Review 60, 30–62.

Ludmerer, K., 1972. Genetics and American Society: A Historical Appraisal. Johns Hopkins University Press, Baltimore, MD.

Montagu, A., 1959. Human Heredity. World Publishing, New York.

Morgan, T.H., 1925. Evolution and Genetics. Princeton University Press, Princeton, NJ.

Muller, H.J., 1933. The dominance of economics over eugenics. Scientific Monthly 37, 40–47.

Pearl, R., 1927. The biology of superiority. American Mercury 1, 257–266.

Reilly, P., 1992. The Surgical Solution: A History of Involuntary Sterilization in the United States. Johns Hopkins University Press, Baltimore, MD.

Turda, M., Weindling, P.J. (Eds.), 2007. Blood and Homeland: Eugenics and Racial Nationalism in Central and Southeast Europe, 1900–1940. CEU Press, Budapest, Hungary.

Watkins, S.W., 1998. On The Pill: A Social History of Oral Contraceptives, 1950–1970. The John Hopkins University Press, Baltimore, MD.

Weiss, S., 1990. The race hygiene movement in Germany, 1904–45. In: Adams, M. (Ed.), The Wellborn Science, Eugenics in Germany, France, Brazil and Russia. Oxford University Press, New York.

Zenderland, L., 1998. Measuring Minds. Cambridge University Press, Cambridge, New York.

Eugenics as an International Movement

Garland E Allen, Washington University in St. Louis, St Louis, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Eugenics, the study of the social forces under human control that work to insure the best possible genetic future, was first coined as a term by Francis Galton in 1883. It became a widespread movement throughout Europe (England, France, Germany, and Scandinavia), North and South America (the United States, Canada, Cuba, and Brazil), Russia and the former Soviet Union, and Asia (China and Japan). Most supporters of eugenics were upper middle-class professionals and wealthy elites who saw the high birth rate of lower socioeconomic groups as a threat to social order and the progress of 'civilization.' Eugenicists in most countries sought to influence legislation on such reproductive issues as birth control and sterilization, and immigration restriction. Eugenics movements differed from country to country in terms of the theories of heredity on which they were based (Mendelian and Lamarckian) and local cultural influences such as religion and family structure.

The term eugenics, derived from the Greek ευγενεσ (eugenes) was first coined by English mathematician and geographer Francis Galton (1822–1911). In his book *Inquiries into Human Faculty and Its Development* eugenics refers to "one born 'good' in stock, hereditarily endowed with noble qualities" (Galton, 1883: p. 24). Eugenics was considered to be the human counterpart of scientific animal and plant breeding.

In his book, Galton lays out all the dimensions that came to characterize eugenics as an ideology and social/political movement during the first half of the twentieth century. These were as follows: (1) adherence to the methods of selective breeding as applied to the improvement of the human species; (2) a strong belief in the power of heredity in determining physical, physiological, and mental (including personality) traits; (3) an inherent ethnocentrism and racism that included belief in the inferiority of some races or groups and superiority of others (a view extended to ethnic groups and social classes as well); and (4) belief in the power of science, rationally employed, to solve social problems, including ones so seemingly intractable as pauperism, crime, violence, urban decay, prostitution, alcoholism, and various forms of mental disease, including manic depression and 'feeblemindedness' (retardation).

It took some time after Galton first propounded the idea for eugenics organizations and activities to emerge around the world, but by the time of World War I (1914), eugenics had become a truly international movement. Thus, the standard historical view of eugenics as a largely British, American, and German phenomenon is inaccurate. Eugenics developed in a variety of national contexts with a wide range of ideological and political programs (Adams, 1990b: pp. 4–6, 1990d: pp. 217–231). The earliest eugenics movements were founded in Germany in 1904, in Britain in 1907, and in the United States in 1908–10. Other eugenics movements appeared around the world shortly thereafter: in Western Europe between 1910 and 1920: France, Spain, Hungary, the Scandinavian countries (including Finland), Russia, Latin America (Cuba, Brazil, and Mexico), Canada, Africa (particularly South Africa), and Asia (China and Japan). While eugenics in its modern form – that is, viewed as a branch of genetics – clearly originated in the West, its spread around the globe gave it an international flavor that from the start retained a common core of scientific and political beliefs while developing local variations. Eugenicists themselves were consciously internationalists, arguing that like all sciences, eugenics knew no national boundaries. Recently, various scholars have applied what is called 'social movement theory' to understanding the common global processes (economic, social, scientific, and cultural transfer) that helped to foster the rise of eugenics and other such movements simultaneously in a number of different countries and national contexts (Barrett and Kurzman, 2004). Despite this international profile, however, it was in Western countries such as the United States and, of course Germany under the Nazi regime (1933–45), that eugenics as an intellectual and social movement reached its greatest strides, and from eugenicists' point of view, enjoyed its greatest ideological and political success.

The international character of eugenics is reflected in the three major international congresses, at which eugenics enthusiasts gathered literally from around the world. The first was held in London in 1912, while the second and third were held in New York (the third congress was held at the American Museum of Natural History) just following the sixth International Congress of Genetics, in 1932. Papers and photographs of exhibits from the second and third congresses fill several impressive volumes.

Although eugenics is often thought of as a politically conservative movement, globally its supporters came from widely divergent political backgrounds. Some, like Charles B. Davenport, Harry H. Laughlin, and Madison Grant in the United States, and Eugen Fischer in Germany, were reactionary, while others, such as Britain's J. B. S. Haldane, the United States' H. J. Muller, or Russia's Aleksandr Serebrovsby, were influential left-wing eugenicists who held socialist or even communist views (Paul, 1984: p. 567). Left-wing eugenicists were not necessarily egalitarians, however, and few envisioned a society in which all people were truly 'equal.' Their model was a meritocracy based on inherited biological worth. It was the reliance on a rational, scientific approach to matters of reproduction and family planning that attracted many, including some liberals, to the eugenic ideal (Kevles, 1986).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.82001-2

Background: Degeneration Theories and the Rise of Eugenics (1890–1914)

In both the West and the East, the late nineteenth century saw, for different reasons, a widespread belief that the human species was in various phases of deterioration and decline (Paul, 1995). Much of this pervasive pessimism came from observations in the social and political arenas. British eugenicists were particularly concerned with the high fecundity and presumed inherited mental degeneracy of the urban working class, particularly those labeled as 'paupers' or (in Britain) the 'residuum.' By contrast, American eugenicists were more concerned with the number of 'feebleminded' who appeared to be filling the prisons and insane asylums in the wake of urbanization and industrialization after 1865 (the US Civil War) and with the genetic deficiencies of the 'new immigrants' around the turn of the century. In Germany, various racial and ethnic groups were thought to harbor an undue proportion of mentally ill, psychotic, psychopathic, and psychiatric members. In France and Brazil, where ideas of the inheritance of acquired characteristics held more sway than in most other countries, public health measures, rather than reliance on selective breeding, were seen as the most effective means to combat degeneracy. In Asia, general theories of degeneration spread with the encroaching influence of Western imperialism from the 1860s onward. The inability of the Asian populations toward off imperial penetration or intra-Asian rivalries was seen as partly a result of scientific and technological backwardness and of physical and cultural weakness in the face of the Western onslaught.

Eugenics and Scientism

Eugenics throughout the world was viewed as the application of rational and 'modern' scientific principles to the solution of age-old social problems (Pickens, 1968). Eugenicists and many of their supporters had a high regard for science and technology, symbols in the Western countries of the triumph of enlightenment and rationality, and of modern capitalist control over, and exploitation of, nature for human gain. The goal of science was to predict and control, an ideology extended to society at large. The period in which eugenics developed was globally one of great economic, social, and political instability: two major world wars; the dissolution of the great empires of Europe in the wake of rebellions abroad; rapid industrialization in the West, especially the growth of heavy industry; and urbanization, all of which contributed to the rise of such problems as alcoholism, prostitution, crime, mental illness, and social unrest. In the major Western countries, organized labor gained great strength, leading to increased and bloody conflicts in the period 1890–1940. Communism made major gains during the period, with the Bolshevik Revolution in Russia, the organization of Workers' Councils that took over several cities in Germany after the Armistice in 1918, and with the expansion of such groups in the United States as the International Workers of the World (the IWW, or 'Wobblies') who carried out a successful general strike in Seattle and other areas of the northwest United States in 1919.

Outside of Europe, similar, sometimes related, historical events led to similar responses. Latin America underwent convulsive economic and social change in the wake of the decline of Spanish and rise of North American colonialism. In Asia, the struggle in China and Japan toward 'modernization,' inter-imperialist rivalries culminating in the Sino–Japanese War of 1894–95, and the Russo–Japanese war of 1905, all stemmed from growing expansionism related to trade, and control of raw materials and markets between East and West and within rising industrial powers, like Japan, in Asia itself. As in the West, science and technology in Asia were associated with modernism and economic advancement.

As a supposed new 'science,' eugenics meshed well with modernist ideology. Eugenics was based on what was thought to be the cutting-edge research in a new field (genetics): it was mathematical, experimental, rigorous, predictive, and purported to apply the principles of science to the solution of previously intractable social problems. Worldwide, it provided an incentive for academically trained practitioners (geneticists) to join the new managerial class that was beginning to apply highly trained expertise to everything from organizing factory production lines, offices, and schoolrooms, to 'scientific' agriculture and not least, human reproduction. Applied to society, eugenics became part of the ideology of 'social engineering' that was replacing the older notion of *laissez faire* associated with social Darwinism. It was not enough to let nature take its course: scientifically trained social engineers had to step in and direct that change along rational lines. Eugenically, this change involved the guidance of human reproduction to ensure that only those with good heredity passed on their genes to future generations. Eugenics was viewed as the 'control of human evolution.' Eugenicists argued that to control heredity at the source was to be efficient – to prevent problems before they occurred, rather than let them happen and then deal with the consequences. In various forms – either as proposed breeding and sterilization programs or maternal and other public health proposals, the efficiency argument dominated much of eugenic social rhetoric. In the United States, this collage of views and ideas was swept up into a somewhat heterogeneous movement known as progressivism. Progressives advocated the use of science to solve social problems, and embraced the efficiency movement of Frederick Winslow Taylor (1856–1915), which had been developed in the United States initially in an industrial context. Similar strains and arguments appeared in different forms in eugenics movements around the world.

Since other articles discuss various specific aspects of eugenics movements around the world (*see* Eugenics in Europe; Eugenics as the Basis of Population Policy; Eugenics, History of; Eugenics in Asia), this article will focus on the general nature and extent of international eugenics activities, comparing the similarities and differences among the various national movements.

British and American Eugenics, 1880–1940

British and American eugenics shared a number of intellectual and social concerns, as well as personal contacts (Kevles, 1986). The major importer of eugenics to the United States, Charles B. Davenport (1866–1944), spent part of the year from 1899 to 1900 with Galton and his protégé, Karl Pearson (1857–1935),

at University College, London, learning statistics and biometrics and gaining initiation into Galton's 'secular religion of eugenics.' American eugenics thus grew directly out of the vision that Galton, and later Pearson, developed. At the same time, American eugenics developed its own particular twists, especially its early acceptance of particulate, discontinuous, and Mendelian concepts of heredity, a relationship not shared by their British counterparts until the later 1920s.

Mendelism notwithstanding, American and British eugenicists shared the basic assumption that such traits as alcoholism, criminality, physical and mental 'defects,' socioeconomic position, and intelligence, were all biologically inherited, with the worse forms concentrated in the lower classes and among certain races. Horrified by the significantly higher birthrate of the lower classes, eugenicists in the United States and Britain argued that the observed decline in productivity and 'national efficiency' was due to the increasing number of less productive social defectives that were being born and were having to be cared for by the state. Eugenics was a 'scientific' approach to improving social efficiency. In Britain, Pearson's largely theoretical orientation to eugenics led him to warn eugenicists to avoid politics. Partly as a result, British eugenicists had much less legislative success than their counterparts in the United States (Kevles, 1986).

In several ways, the American movement followed a different trajectory from the movement in Britain. First, American eugenicists became strong advocates of psychometrics, the burgeoning new school of I.Q. testing, which took it for granted such tests measured an innate, unchanging mental function. While British eugenicists also assumed that mental ability was inherited, they did not employ the I.Q. (or other similar) tests to the same extent as in the United States. Second, leaders of American eugenics were from the start anxious to play a role in the public arena – in what today would be called formation of public policy. Eugenicists such as Harry H. Laughlin (1880–1943), Superintendent of the Eugenics Research Office at Cold Spring Harbor, NY, USA, translated eugenic principles into social action (Ludmerer, 1972). Eugenicists lobbied for passage of the Johnson–Reed Act (Immigration Restriction Act) in the United Sates Congress in 1924, while Laughlin designed a 'Model Sterilization Law' and eugenicists lobbied for the passage of eugenical sterilization laws in a number of states during the 1920s and 1930s. The first eugenical sterilization laws were passed in the United States (the first were passed in the state of Indiana, 1907), although by the mid-1930s, such laws were in place in a number of Western countries (Norway, Sweden, Finland, Germany, and Canada). Between 1907 and 1965, more than 60 000 people in the United States – mostly in state institutions for the feeble-minded, psychiatric hospitals, and prisons – were sterilized. Like their British counterparts, American eugenicists were highly motivated by the ideas of national efficiency and rational planning (Vecoli, 1960; Allen, 1986: pp. 255–260).

Eugenics on the Continent: Germany, Scandinavia, and France

On the continent, eugenics in Germany and Scandinavia shared many features in common, while France took its own, unique direction. The German and Scandinavian movements were based strongly on what has been termed notions of 'hard heredity' (i.e., that most mental and moral attributes are rigidly determined by genes at conception), and employed basic concepts of Mendelian genetics. While the ideology of Nordic and Aryan superiority was present in both movements, it represented only one strand, and not always the most prominent strand in eugenic thought prior to the ascendancy of the National Socialists in Germany in 1933 (Weiss, 1990). Ideologues of the German movement, such as Fritz Lenz (1887–1976) and Eugen Fischer (1874–1967), were racial purists, but during the Weimar years kept a 'subdued' profile on overtly racial issues. Because racial purity was a cornerstone of the Nazi regime, however, eugenics after 1933 became openly racist and pro-Nordic (Weiss, 2010).

A number of common ideological convictions characterized German eugenics from its pre-1900 days, right down through the 'final solution': (1) The conviction that *laissez-faire* capitalism was at heart dysgenic because it allowed individuals to reproduce at will, with no concern for the future well-being of society. (2) The belief that Germany's population was degenerating, as evidenced by the debacle of World War I, increases in labor unrest, and social problems such as alcoholism, criminality, prostitution, mental illness, and retardation. (3) A strong belief in the value of a managed society based on scientific expertise.

At the same time, several features distinguish German eugenics from its counterparts elsewhere: (1) More German eugenicists were drawn from the medical profession than in either England or the United States (Proctor, 1988). (2) During the Weimar Republic and later the Nazi regime, eugenics received far more government funding than in most other countries. German geneticists and eugenicists all held posts on important government advisory councils dealing with matters of eugenics and national health (Weindling, 1985, 1989: pp. 308–309). (3) A major feature of German eugenics was the subsequent role it came to play in the National Socialist government and the Holocaust, between 1933 and 1945.

Some historians have argued, cogently, that eugenics became so prominent and virulent in Germany because of the humiliation of defeat in World War I and the exacerbating terms of the Treaty of Versailles. Among the factors leading to military defeat, eugenicists argued, was the nation's national degeneracy, the influences of racial impurity, declining birthrates among the intelligentsia, and economically prosperous, and the anti-eugenic policies associated with socialist egalitarianism (Meineke, 1963: pp. 30–34; Mosse, 1964: pp. 237–240). Worsening economic conditions due to reparations payments and loss of the mineral-rich Ruhr Valley, gave, among average Germans, added weight to arguments for improving national efficiency by eliminating what the Nazis termed 'useless eaters' or 'lives not worth living' (Weindling, 1989: pp. 305–310). Under national socialism, eugenics eventually became the scientific rationale for euthanasia and genocide.

The Scandinavian countries, including Finland, had well-entrenched eugenics movements by the 1920s, those in Norway and Sweden being the largest and most politically successful. Scandinavian eugenics showed much of the character of the other Western movements: an emphasis on efficiency, commitment to legislative action, and a strong

pro-Aryan and Nordic stance (Broberg and Roll-Hansen, 1996). In the period between 1910 and 1925, Scandinavian eugenics was characterized by a strong belief in the superiority of the Nordic races, and thus displayed a nationalist and racist undertone. These took the form of social legislation, particularly the passage of eugenical sterilization laws (Denmark's law was passed in 1919, Norway's in 1934, and Sweden and Finland's in 1935). The rates of sterilization varied considerably: between 1935 and the 1960s, Sweden sterilized over 62 000 people – an astonishing figure for a country whose total population at the time was less than 7 million (Roll-Hanson, 1996: p. 263). Eugenics achieved its greatest political success in Sweden, although the most vocal international voice from Scandinavia was that of Jon Alfred Mjøen (1860–1939), a Norwegian pharmacist turned eugenic activist. Although shunned by many scientists in his own country for his racist views, Mjøen was viewed as a major theoretician by eugenicists in Sweden and Denmark. Nonetheless, Scandinavian eugenics as a whole never adopted the virulent Nordic racism found among Nazi elements in Germany. As advanced welfare states, the Scandinavian countries also couched their eugenics proposals in terms of 'national efficiency.' Hit hard by the same economic problems faced by other countries during the interwar period, and especially after 1929, the Scandinavian countries were seeking to reduce state expenditures wherever possible, and the supposedly hereditarily defective became an even more likely target.

Eugenics in France

The French, as other Europeans, harbored a growing concern about what was perceived as the degeneration of society, in France's case associated with a declining birthrate and loss of military hegemony on the continent after the Franco-Prussian war of 1870. French eugenics also embraced as strongly as elsewhere the idea of rational management by scientifically trained experts, especially in areas of social policy. Like other countries during the same period, France was beset with a variety of economic, social, and political problems that called for sweeping changes in both public attitudes and government policy (Thomson, 1960: pp. 173–186). Before, but especially after, World War I, class struggle accelerated in such publicly visible forms as large-scale strikes (Tilly et al., 1975: pp. 71–76). Like Germany, eugenics in France was strongly supported by members of the anthropological, medical, and public health professions. French eugenics differed from that in England, the United States, and Germany, however, in three very important respects: (1) The presence of a strong Enlightenment tradition of radical social theory and practice that laid greater emphasis than elsewhere on environmental, as opposed to biological, determinants of human personality and behavior. (2) Concomitant with their emphasis on the environment, French biologists maintained a strong commitment to neo-Lamarckian concepts of heredity, that is, the direct inheritance of environmentally induced variations. Consequently, eugenicists in France did not emphasize Mendelian genetics and Darwinian natural selection. (3) The strong influence of the Catholic Church, especially with regard to matters of reproduction, caused French eugenicists for the most part to avoid advocating measures such as contraception, sterilization, or abortion to deal with actual or potential defectives (especially after the Papal Bull, *Casti connubi* in 1930, which condemned eugenical sterilization). The effect of the church position was to widen an already existing division within the eugenicists' ranks: between liberal and conservative and between pro- and antisterilization forces. Thus, French eugenicists never attained the unity necessary to lobby effectively for specific eugenic measures at the legislative levels.

Illustrative of French eugenics is the work of Just Sicard de Plauzoles (b. 1872), a doctor, a reformer, and Secretary of the Venereal Disease Commission of the Ministry of Health. He was the author of two highly influential books, *Tuberculosis* (1900) and *Motherhood and the National Defense against Depopulation* (1909). As a public health official, Sicard joined the French Eugenics Society in 1922 under whose auspices he gave a popular course of lectures every year on 'social hygiene' at the Sorbonne, aimed at medical, health, and government officials. Sicard's brand of eugenics was cast in the terms of preventive medicine and public health, analogous to programs for public sanitation. Like many eugenicists, he was a devotee of the industrial efficiency movement, carrying his analogy to the point of referring to human beings as "work-producing machines" and eugenics as the technique for "procreating, perfecting and utilizing man" as "human capital" (Schneider, 1990: p. 89). In this vein, he also stressed the importance of quality over quantity: increasing the birth rate of superior human beings, like high-quality machines, would ultimately be more cost-effective. The working classes, he felt, were showing signs of degeneracy due to their higher birth rate combined with defective heredity. His eugenics was thus much more class based than race based. In addition to his influence through lecture course and writings, Sicard was also a legislative activist, helping to sponsor a bill in the Chamber of Deputies in 1932 calling for the repeal of a 1920 law that forbade the sale or advocacy of contraceptive devices (the bill did not pass).

Thus, while French eugenics shared with the British, German, and Scandinavian movements a desire to improve the overall biological quality of the population, it differed in several important respects: (1) The French eugenicists' theory of heredity – non-Mendelian and neo-Lamarckian, or 'soft heredity' – clearly set it apart from most of the other countries. (2) The emphasis on soft heredity thus led the French to focus more attention on such environmental factors as health care in general and maternal health care in particular, sanitation, nutrition, and public health programs. (3) The strong presence of the Catholic Church stymied any strong eugenic legislation such as compulsory sterilization laws; under these circumstances, environmental approaches were more logical and practical. As a coherent movement, then, eugenics in France never attained the high degree of organization and political effectiveness that it enjoyed in the United States or, later, Nazi Germany.

Eugenics in Russia

The development of eugenics in Russia occurred in a turbulent period in that country's history, framed by the onset of World War I (1914), the Bolshevik Revolution (1917), and the

subsequent attempt to fashion a new society based on distinctly different ideologies and social practices. As in other countries, eugenics made its entrance into Russian society in the first decade of the twentieth century along with Mendelian genetics, both of which were brought from Europe by Russian biologists studying abroad, such as Nikolai Kol'tsov (1872–1940) and Iurii Filipchenko (1882–1930). The eugenics movement took shape starting around 1920 when Filipchenko and other committed Bolsheviks saw eugenics as the perfect fit to postrevolutionary Soviet ideology: it was the application of a rational and materialist theory to a vision for a new kind of human society. Although vigorous at first, the eugenics movement as such in the Union of Soviet Socialist Republics was rather short-lived, flourishing only from 1920 through the early 1930s, when it came under increasing attack by the Communist Party as racist and elitist and was eventually disbanded (Graham, 1977: p. 1145; Adams, 1989: pp. 881–882).

Although Russian biologists had begun to develop a strong Mendelian research program by 1914, the advent of the war, the revolution of 1917, and the ensuing civil war retarded further development of most scientific work until after 1920. The decade between 1920 and 1930, in which Russian genetics and eugenics expanded rapidly, can be divided conveniently into two periods: 1920–25 and 1926–30. The eugenics of the first period was modeled on movements in the West, particularly Germany, with which Russian scientists traditionally had close ties. In the second period, Soviet eugenicists attempted to create their own socialist eugenics based on the Marxist principles of historical materialism and a scientific approach to society (Graham, 1977: p. 1145).

In the aftermath of the revolution, Mendelian genetics began to play an important role in Soviet agriculture. Resources were severely limited and the famines of 1920 and 1921 impressed on Lenin and other leaders the importance of the scientific breeding for increasing food production in the future. As in other countries, eugenics gained prominence by coat tailing on the expansion of Mendelian genetics. From there, it seemed only natural to extend genetic principles to the scientific breeding of humans as well.

Concrete form was given to Soviet eugenics in 1921 when its two most important organizations were formed: the Russian Eugenics Society and the Bureau of Eugenics of the Soviet Academy of Sciences (Graham, 1977: p. 1146). The Eugenics Society, composed of both biologists and laypersons was given a small state subsidy, and was oriented toward political activism. The Bureau, composed of professional geneticists, was fully subsidized by the state, and was more oriented toward research and scholarly publication. Workers at the Bureau were particularly concerned to insulate the new science of eugenics from entanglement in political matters. Members of the Society, by contrast, were eager to inject genetic ideas into the social and political arena and saw no conflict between eugenical principles and socialism (Graham, 1977: p. 1146). This approach commanded particular attention in a Marxist state, where a conscious policy of wedding theory and practice was considered the cornerstone of the new politics.

Like their counterparts in other Western countries, many of the younger members of the Russian Eugenics Society were alarmed about the dysgenic effects of war and declining birth rate among the intellectual and educated classes. This topic received much attention in the work of M. V. Volotskoi (1893–1944) in a 1923 book, *On the Sexual Sterilization of the Hereditarily Defective* (Adams, 1990: p. 175). Based on a survey of the state sterilization laws in the United States, Volotskoi argued for large-scale sterilizations of the unfit to improve the genetic character of the Russian people. Volotskoi's ideas met with considerable opposition, however, on moral as well as scientific and political grounds. Since birth rate had fallen disastrously in recent decades, proposals for birth control or sterilization were viewed as dysgenic. Thus, negative eugenic programs never developed a significant following in the Soviet Union.

A more viable option for pursuit of eugenic goals was the widespread belief in neo-Lamarckian inheritance. Like biologists and eugenicists in France and Germany, many in the Soviet Union adhered to some form of neo-Lamarckism. Volotskoi, for example, argued that neo-Lamarckism alone allowed for the rapid improvement in human biological qualities that eugenics called for. If the gene was as fixed an entity as classical Mendelian genetics suggested, there was little chance of creating the new Soviet human being by eugenical means. Through significant changes in the social and economic environment, however, Volotskoi saw the opportunity to achieve eugenic ends in a manner consonant with a Marxist worldview.

However, Mendelian geneticists such as Filipchenko and Serebrovsby pointed out that the social environment could affect the germplasm in a negative as well as positive way, so that the proletariat and peasantry, subjected for so many past generations to poor environment and nutrition would be likely to have inferior germplasm, a conclusion distasteful to most Marxists (Adams, 1990c: p. 177). Thus by the mid-1920s, Mendelian genetics and its associated eugenic ideology were already headed on a collision course with the neo-Lamarckians that was ultimately to have disastrous consequences for Russian biology in general, and genetics in particular. It is the debates around this issue that characterize the second period in the development of Soviet eugenics (1925–30).

In response to the Communist Party's call for implementation of the first Five Year Plan in 1928, Aleksandr Serebrovsky (1892–1948) along with the young medical geneticist Solomon G. Levit (1894–1938) proposed that if the Soviet Union were cleansed of all its hereditary defectives, the Five Year Plan could be achieved in two and a half years (Adams, 1990c: p. 180). Serebrovsky made his plan concrete by proposing the use of methods of artificial insemination recently developed for animal breeding to ensure the most eugenical matings unencumbered by the problem of forced marriage or intercourse. It was such proposals that brought eugenics, and its associated Mendelian genetics, under further scrutiny. An article published in 1925 in one of the major journals of Marxist theory launched a pointed attack on eugenics claiming that eugenicists emphasized biological determinants of human behavior to the total neglect of social and economic factors (Graham, 1977: p. 1150). Sounding a neo-Lamarckian refrain, the article suggested that attention to the human environment would go a lot farther toward improving human mental and physical health than eugenics.

At the same time, the pressure to produce rapid results for agricultural improvement produced a schism between the more theoretical geneticists and government and party leaders. Mendelian genetics seemed to yield few immediate results, and its association with eugenics made it appear to be an example of bourgeois elitism and racism (Joravsky, 1970: p. 30; Levins and Lewontin, 1985: p. 164; Adams, 1990c: p. 183). In response, however, Mendelian geneticists such as Kol'tsov and Serebrovsky argued that the best long-range results would be obtained by allowing geneticists to continue working on simple model organisms like *Drosophila* rather than the much more complex and experimentally difficult agricultural organisms like cows and chickens. Many Soviet geneticists, like their counterparts elsewhere, had adopted a technocratic approach, working in their specialized laboratories rather than in the fields and breeding farms (Adams, 1990c: p. 182). Government and Party officials increasingly saw these attitudes as bourgeois and elitist. Serebrovsky was attacked in 1930 for being overly technocratic and biologistic by treating human beings as resources like coal and gas (Adams, 1990c: p. 184). Largely as a result of these developments, in 1930 the Communist Party launched an attack on Mendelian geneticists as bourgeois experts and their associated eugenics claims were officially outlawed. A number of major Soviet geneticists/eugenicists were demoted or lost their positions and the Russian Eugenics Society was disbanded in that year. By 1931, the Soviet eugenics movement as such had ceased to exist in any organized form.

Although Mendelian genetics itself was beginning to come under attack in the early 1930s by neo-Lamarckian agronomist Trofim D. Lysenko (1898–1976) and his followers, it remained a viable research program well into the 1930s in such institutions as the Maxim Gorky Research Institute of Medical Genetics (Adams et al., 2005: pp. 251–252). However, the association of Mendelian genetics with eugenics along with its opposition to neo-Lamarckism led ultimately to the claim that geneticists had imported Fascist ideas into Soviet science, and to the arrest (and in some cases execution) of its leading proponents (Adams, 1989: p. 883). By the late 1940s, Lysenko's hereditarian claims became official Soviet policy and Mendelian genetics (and geneticists) banned. It would take Soviet genetics at least two more decades to recover.

Eugenics in Latin America

The existence of significant eugenics movements in non-European and non-Anglo-Saxon contexts has been ignored by historians of science until quite recently (Adams, 1990b; Stepan, 1991). Eugenics movements in Latin America rose to prominence between the world wars and commanded considerable attention within the worldwide eugenics community. Interest in eugenics stimulated some of the first genetics courses in Latin American universities, and was often an important factor in child welfare, maternal health, infectious disease, and immigration legislation in the various countries (Stepan, 1990: p. 110). Eugenics was also discussed vigorously at various Latin American scientific and public health conferences, as well as being the focus of two specific Pan-American conferences of its own (Cuba in 1927 and Argentina in 1934). A special eugenics society, the *Fédération International Latine des Sociétés d'Eugénique*, was founded in 1936 under the leadership of Corrado Gini, a well-known Italian eugenicist and president of the Italian Society of Genetics and Eugenics. The *Fédération* held one international congress, in Paris in 1937. Moreover, leaders of eugenics movements in other countries expressed keen interest in the developments in Latin America. Not only Gini in Italy, but also Davenport and Laughlin in the United States offered considerable encouragement to fledgling societies in Cuba and Mexico.

One of the most important differences between Latin American and European eugenics was the political economy of neocolonial 'underdevelopment' in which most Latin American countries existed in the first decades of the twentieth century. Although just beginning to experience industrialization and urbanization after World War I, Brazil, for example, was still largely an agricultural country with an impoverished rural population that was illiterate and dominated by the Catholic Church. Much of the land was held as plantations by local family oligarchies and foreign investors who produced more for export than for local consumption. The country was ruled by a tiny elite largely educated abroad. Biology, such as existed, was housed only in medical schools and agricultural stations. Historian Nancy Stepan has pointed to the highly mixed racial composition of most South American countries as a factor differentiating the Brazilian eugenics movement from that in Europe or the United States (Stepan, 1990: p. 11).

Four specific factors helped to make eugenics attractive in the period after 1920 in countries such as Brazil (Stepan, 1990: pp. 112–114): (1) A rising new nationalism and sense of independence gained from the war experience. The collapse of Germany, whose science and culture had been so much admired in Brazil and other Latin American countries before the war, generated a new sense of independence about what Latin Americans could accomplish on their own. (2) There was increased recognition of 'the social question,' that is, the appalling poverty and poor health of the working and peasant classes. After 1915, the combination of poverty, massive migrations, and immigration from abroad (particularly Italy) produced a radicalization of politics. A large-scale strike in 1917 in São Paolo, Brazil, was met with brutal repression by the government and industry leaders. Under these circumstances, eugenics began to play the same role as it had in the United States and Britain: a means for breeding out the biological defectives who were claimed to be the cause of such disruptions. (3) Growing enthusiasm for science, and particularly biological science, as a way toward economic and social progress, especially in the area of public health, had been imported into Latin America from Europe and the United States. As in both France and Germany, eugenics in Latin America was easily incorporated by an expanding medical profession eager to assert its own power and expertise in generating new or reformed public health policy. (4) Finally, the widespread racial intermixing (between people of Indian, African, and European descents) that characterized most Latin American countries had two very different effects in Latin America: on the one hand, as in Europe, it raised the specter of race suicide and impending degeneracy; on the other, large numbers of South Americans had mixed ancestry, thus making

racial classification so fluid that the sharp distinctions made in European countries and the United States never developed with the same intensity.

With the exception of Cuba, most early eugenics programs in Latin America were derivative of those in Europe, particularly France (Stepan, 1990: p. 114). This meant that they often adopted version of neo-Lamarckian rather than Mendelian concepts of heredity, and for a variety of local reasons were also much more concerned with improvements in sanitation, nutrition, and public health. Indeed, most of the members of local eugenics organizations were drawn from the medical or other related professions.

Because it had one of the largest and most well-developed eugenics movements, Brazil has become one of the best studied examples of eugenics in Latin America. As a dependent colony even after liberation in the earlier part of the nineteenth century, Brazil was ruled by an elite of mostly Portuguese ancestry that shared the general European racial views regarding indigenous peoples or those with mixed ancestry in the New World. And with the abolition of slavery in 1888, many of the ruling elites feared the degenerative effects of racial miscegenation. At the same time, the colonial powers needed to deal with issues of tropical disease that affected their administrators and military personnel, as well as local populations.

Thus, in the earlier phase of Brazilian (and other Latin American) eugenics movements, the focus was on environmental improvements which, under a general neo-Lamarckian framework, was expected to yield long-term hereditary, as well as short-term health and medical benefits. This does not mean, however, that traditional ideas about biological heredity were absent. For example, at one of the early organizing meetings of what was to become the São Paulo Eugenics Society in 1917, the medical doctor Renato Ferraz Kehl (1889–1974) opened with a description of Galton's 'new science of eugenics,' and proceeded to discuss several policy proposals, including the desirability of prenuptial physical and mental exams and a proposed revision to current Brazilian law allowing consanguineous marriages.

After starting from this more anthropological and public health perspective, by the 1940s a new generation of eugenicists had adopted Mendelian genetics and began to apply it to formulate Latin American eugenics policies. Greatly stimulated by the visit of American geneticist Theodosius Dobzhansky (1900–75) to Brazil in 1943, eugenicists in Latin America, unlike their US or German counterparts, used genetics to argue that miscegenation, like crossing different strains of animals and plants, could produce superior, not inferior or 'mongrel,' lines. The Mendelian eugenicists thus came into conflict with the neo-Lamarckian eugenicists, creating a division within the ranks of the movement as a whole.

Mendelian geneticists and eugenicists were particularly prominent in agricultural schools, and like their counterparts elsewhere were particularly concerned with issues of inbreeding and outbreeding and their potential effects on the viability of hybrid crosses. Mendelian critics of neo-Lamarckism, and its effects on the progress of eugenics, included two highly influential scientists: the anthropologist Edgar Roquette-Pinto (1884–1954), Director of the National Museum of Anthropology, and the biologist Octavio Dominguez (n.d.) at the Agricultural School of Piracicaba (near São Paulo). Roquette-Pinto had enormous prestige in scientific circles and his promotion of Mendelian theory was particularly important in developing that side of Brazilian eugenics. He claimed in 1929 at the First Brazilian Congress of Eugenics that what he considered the 'confusion' between eugenics and sanitation/public heath would be cleared up by the time of the second congress, and eugenics could then begin to make great strides in Brazil. Dominguez was a major writer of textbooks that promoted an orthodox view of Mendelism, much of it derived from Galton, Pearson, Davenport, Castle, and other Anglo-American writers. He was also a member of both the American Genetical Association (originally American Breeders Association) and the Eugenics Society of London. Although not free of views of human hierarchical worth, on the whole Mendelian geneticists and eugenicists in Brazil tended to be far less overtly racist than their counterparts in Britain and the United States. In the face of largely foreign-dominated economic, political, and social expansion, however, racism, particularly in the form of and marginalization of the continent's black and Indian populations, remained one facet of Brazilian and other Latin American eugenics movements.

The Latin American experience illustrates the quite different path that eugenics could take in an environment far removed from the centers of scientific advancement, one in which scientific ideas were derived largely from outside. The strong orientation of most Latin American eugenics to personal and public hygiene, as well as to nutrition and medical care, clearly reflects the existence of widespread poverty, malnutrition, and disease (hookworm, malaria, and yellow fever) among the vast and largely rural populations. With economic and health problems of such magnitude, issues of race and race mixture were less central among Latin American eugenicists than among their counterparts in the industrialized world.

Eugenics in Asia

Scholars studying the history of modern Asian science have pointed out that eugenics had a role to play in many of those societies. Moreover, it was not merely Western eugenics imported and imposed on Asian society *carte blanche*. In both China and Japan, eugenics fitted well with movements for the increased promotion of science and technology associated with conscious attempts at what was called 'modernization.' Mendelian genetics was introduced to both China and Japan in the early decades of the twentieth century by Western scientists who traveled to Asia, as well as by Asian scientists who came to work in laboratories in the West. Unlike their Western counterparts, eugenicists in both China and Japan relied less on a formal, organized set of genetic theories, promoting eugenics primarily as a 'modern' way of dealing with traditional problems. In addition, there appears to have been no strong, organized neo-Lamarckian tradition in China or Japan as existed in the major European movements, so when eugenics was introduced along with Mendelian genetics, it did not encounter the same sort of theoretical resistance as elsewhere. However, in China after the communist revolution in 1949, Lysenkoist views became an important ideological foundation for the study of heredity, and thus a conflict between Mendelian and

neo-Lamarckian views became part of the scientific debates well into the 1950s and 1960s (Schneider, 1988: pp. 28–29).

Chinese eugenics, when it was introduced in the 1920s and 1930s, was perhaps unique among all nations in that few if any well-trained scientists, including biologists, were participants. Rather, it was a movement of intellectuals and journalists (Schneider, 2003: p. 132). The most active eugenicist with any scientific background in this period was Pan Guangdam (1898–1967), who had studied biology, but never worked as an active biologist. He promoted a kind of 'positive eugenics,' which appeared to raise fewer moral issues than negative eugenics (Schneider, 2003: p. 139). For Chinese intellectuals and 'progressives,' eugenics became a way of grounding the calls for social and economic reform in science, a view akin to the Western principle of rational, social control. While not relying on Lamarckian notions of inheritance of acquired characteristics, during the 1920s and 1930s Chinese eugenicists emphasized the importance of education, public health, and other social measures, along with limiting family size. There was relatively little in the way of statistical or genetic research carried out by eugenicists in Republican China.

As elsewhere, eugenics in China was particularly supported by the rising professional middle classes, whose ideas were in various ways an amalgam of imported Western views and traditional Chinese notions of race and class. Despite the lack of explicit genetic input, as historian Frank Dikotter has argued, eugenic and racial thinking permeated much of Chinese culture from the latter part of the nineteenth century through the first half of the twentieth century (Dikotter, 1992).

After the advent of the communist regime in 1949, traditional eugenics was attacked as racist and fascist as it had been in the Soviet Union, and a few earlier proponents of eugenics recanted their views (Schneider, 2003: p. 139). By the 1970s, however, population growth and calls for population control policies were increasing. This culminated in 1978 with the One Child Law, which raised considerable concern both inside and outside of China. Whether the policy has strictly eugenical implications is highly debated, but it is clearly one of the most drastic reproductive control policies developed by any country in the past or present.

In Japan, during the early decades of the twentieth century, eugenics was linked more closely than in China with both Mendelian genetics as well as movements for birth control, sex education, and public health. In this context, conditions such as venereal disease, tuberculosis, alcoholism, and leprosy were all considered inherited and eugenicists argued they should fall under the same regulations as those governing public health matters. The difference is due in large part to the much closer connection between Western (particularly German and American) biologists and those in Japan. Here too, as in the West, eugenics was seen as a scientific means of combating social and moral degeneracy. In 1935, Japanese eugenicists were successful in getting the Imperial Diet to pass a Racial Eugenics Protection bill, which favored the sterilization of hereditarily mental defectives. A sterilization law based on the German 1933 law (in turn based on model laws in the United States) was drafted in 1936, but a fully comprehensive eugenics law was not adopted until 1940.

Ironically, although eugenics was on the wane in the West after World War II, it became more prominent in Japan during that same period. A full-scale Eugenical Protection Law was passed in 1948, allowing for involuntary sterilization of those with undesirable, supposedly inherited traits. By 1996, over 16 000 such operations had been carried out. This law was very much influenced by the experience of US geneticists, particularly those involved with the Atomic Bomb Casualty Commission, a joint effort between Allied (mostly United States) and Japanese scientists to investigate the medical and genetic effects of high-energy radiation from the atomic blasts in Hiroshima and Nagasaki. Some of these geneticists had, or had once had, eugenical sympathies. In turn, many major Japanese geneticists had been educated in the United States in the 1920s and 1930s when eugenics was in its heyday. Eugenics in Japan therefore bore a stronger resemblance to its Western heritage than did eugenics in China.

Conclusion

Although the term eugenics was used worldwide, eugenic movements in various countries were shaped by local economic, social, political, and cultural factors, and thus differed from one another in substantial ways. It has become clear that, as a movement, eugenics was not monolithic. Although tied to rigidly hereditarian Mendelian genetics in the United States, Scandinavia, and somewhat later in England and Germany, eugenics was more aligned with neo-Lamarckian views of inheritance in France, Latin America, and the Soviet Union.

In their orientation toward various social issues and proposed solutions, eugenicists also varied considerably from country to country. In some, such as the United States, Scandinavia, and Germany, concern with racial issues played a prominent role – indeed the Nazis claimed the 'Third Reich' was based on the concept of the 'racial state.' In these countries, concern with the ill effects of race crossing was widespread, while in countries such as Brazil, race mixture was considered by some eugenicists actually to be a source of biological improvement (appealing to the concept of hybrid vigor). For British eugenicists, the supposed degeneration apparent among the lower classes was the counterpart of the supposed inferiority of racial and ethnic groups in the less developed countries. In all three cases, it was the economically dispossessed who formed the focal point of eugenicists' fears, and the targets of their programs.

Eugenics movements also differed widely in the social policies on which they focused. In those countries such as France, Brazil, and to some extent the Soviet Union, where a more neo-Lamarckian view prevailed and where the medical profession was more actively involved in eugenics, proposals tended to promote environmental reforms, particularly with regard to public health. In those countries such as the United States and Britain, and especially Germany after 1933, where more strictly Mendelian views were predominant, segregation and/or sterilization of supposed 'defectives,' as well as immigration restriction, were more commonly proposed solutions.

It should be pointed out that the scientific basis for eugenical claims – that complex behavioral, personality, and mental traits are inherited in a 'hard-wired' fashion, and particularly that they follow Mendelian patterns of inheritance – has never found strong empirical support. Most eugenicists had a simplistic understanding of Mendelian

genetics, and persistently ignored environmental inputs into determining social traits. Various geneticists in the West, such as Thomas Hunt Morgan (1866–1945) and Herbert Spencer Jennings (1868–1947) strongly criticized the genetic claims on which eugenics was based, but their public critiques came late in the movement (starting in the mid-1920s) and appear to have done little to stop it.

Underlying all the differences, eugenics movements worldwide shared a concern for efficiency, the ideology of scientific management, the role of middle-class experts in establishing social and political policy, and a strong faith in science. Underlying all this, however, in almost all eugenics movements, was a powerful undercurrent of apprehension about the supposed degenerating biological quality of the human species. Eugenics programs, from California to the Baltic Sea, were aimed at increasing the biological worth of the human species by applying the latest genetic and medical knowledge. That such knowledge was often incomplete or flawed ultimately brought the eugenic tree to bear a bitter fruit.

See also: Eugenics as a Basis of Population Policy; Eugenics in Asia; Eugenics in Europe; Eugenics, History of; Genomics, Ethical Issues In.

Bibliography

Adams, M.B., 1989. The politics of human heredity in the USSR, 1920–1940. Genome 31, 879–884.

Adams, M.B. (Ed.), 1990a. The Wellborn Science. Eugenics in Germany, France, Brazil, and Russia. Oxford University Press, New York.

Adams, M.B., 1990b. Introduction. In: Adams, M.B. (Ed.), The Wellborn Science. Eugenics in Germany, France, Brazil, and Russia. Oxford University Press, New York, pp. 3–7.

Adams, M.B., 1990c. Eugenics in Russia, 1900–1940. In: Adams, M.B. (Ed.), The Wellborn Science: Eugenics in German, France, Brazil and Russia. Oxford University Press, New York, pp. 153–216.

Adams, M.B., 1990d. Towards a comparative history of eugenics. In: Adams, M.B. (Ed.), The Wellborn Science: Eugenics in German, France, Brazil and Russia. Oxford University Press, New York, pp. 217–231.

Adams, M.B., Allen, G.E., Weiss, S.F., 2005. Human heredity and politics: a comparative institutional study of the Eugenics Record Office at Cold Spring Harbor (United States), the Kaiser Wilhelm Institute for Anthropology, Human Heredity and Eugenics (Germany), and the Maxim Gorky Medical Genetics Institute (USSR). Osiris 2, 232–262 (Second Series).

Allen, G.E., 1986. The eugenics record office at Cold Spring Harbor, 1910–1940: an essay in institutional history. Osiris 2, 225–264 (Second Series).

Barrett, D., Kurzman, C., 2004. Globalizing social movement theory: the case of eugenics. Theory and Society 33, 487–527.

Broberg, G., Roll-Hansen, N. (Eds.), 1996. Eugenics and the Welfare State: Sterilization Policy in Denmark, Sweden, Norway, and Finland. Michican State University Press, East Lansing, MI.

Dikotter, F., 1992. The Discourse on Race in Modern China. Hurst Publishers, London, UK.

Galton, F., 1883. Inquiries into Human Faculty and Its Development. Macmillan, London.

Graham, L.R., 1977. Science and values: the eugenics movement in Germany and Russia in the 1920s. American Historical Review 82, 1133–1164.

Joravsky, D., 1970. The Lysenko Affair. Harvard University Press, Cambridge, MA.

Kevles, D.J., 1986. In the Name of Eugenics. Alfred A. Knopf, New York.

Levins, R., Lewontin, R., 1985. The Problem of Lysenkoism. In: Levins, R., Lewontin, R. (Eds.), The Dialectical Biologist. Harvard University Press, Cambridge, MA, pp. 163–196.

Ludmerer, K., 1972. Genetics and American Society: A Historical Appraisal. Johns Hopkins University Press, Baltimore.

Meineke, F., 1963. The German Catastrophe. Beacon Press, Boston.

Mosse, G.L., 1964. The Crisis of German Ideology. Intellectual Origins of the Third Reich. Grosset and Dunlap, New York.

Paul, D., 1984. Eugenics and the left. Journal of the History of Ideas 45, 567–590.

Paul, D., 1995. Controlling Human Heredity: 1865–Present. Humanities Press International, Atlantic Highlands, NJ.

Pickens, D., 1968. Eugenics and the Progressives. Vanderbilt University Press, Nashville, TN.

Proctor, R., 1988. Racial Hygiene: Medicine under the Nazis. Harvard University Press, Cambridge, MA.

Roll-Hansen, N., 1996. Norwegian eugenics: sterilization as social reform. In: Broberg, G., Roll-Hansen, N. (Eds.), Eugenics and the Welfare State: Sterilization Policy in Denmark, Sweden, Norway and Finland. Michigan State University Press, East Lansing, MI, pp. 151–194.

Schneider, L.A., 1988. Genetics in Republican China. In: Bowers, J.Z., Hess, J.W., Sivin, N. (Eds.), Science and Medicine in Twentieth-century China: Research and Education. Center for Chinese Studies, University of Michigan, Ann Arbor, MI.

Schneider, L.A., 2003. Biology and Revolution in Twentieth-century China. Rowman and Littlefield, London, UK.

Schneider, W.H., 1990. The eugenics movement in France, 1890–1940. In: Adams, M.B. (Ed.), The Wellborn Science. Eugenics in Germany, France, Brazil, and Russia. Oxford University Press, New York, pp. 109–169.

Stepan, N., 1990. Eugenics in Brazil, 1917–1940. In: Adams, M.B. (Ed.), The Wellborn Science. Eugenics in Germany, France, Brazil, and Russia. Oxford University Press, New York, pp. 110–152.

Stepan, N., 1991. The Hour of Eugenics. Race, Gender, and Nation in Latin America. Cornell University Press, Ithaca, NY.

Thomson, D., 1960. Democracy in France: The Third and Fourth Republics. Oxford University Press, Oxford, UK.

Tilly, C., Tilly, L., Tilly, P., 1975. The Rebellious Century, 1830–1930. Harvard University Press, Cambridge, MA.

Vecoli, R., 1960. Sterilization: a progressive measure? Wisconsin Magazine of History 43, 190–202.

Weindling, P., 1985. Weimar eugenics: the Kaiser Wilhelm Institute for anthropology, human heredity, and eugenics in social context. Annals of Science 42, 303–318.

Weindling, P., 1989. Health, Race and German Politics between National Unification and Nazism, 1870–1945. Cambridge University Press, Cambridge, UK.

Weiss, S.F., 1990. The race hygiene movement in Germany, 1904–1945. In: Adams, M.B. (Ed.), The Wellborn Science. Eugenics in Germany, France, Brazil, and Russia. Oxford University Press, New York.

Weiss, S.F., 2010. The Nazi Symbiosis: Human Genetics and Politics in the Third Reich. University of Chicago Press, Chicago.

Eugenics in Asia

Frank Dikötter, The University of Hong Kong, Hong Kong, People's Republic of China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Eugenics not only spread in Europe between the two World Wars, but also developed a large following in Asia, in particular, in China and Japan. Eugenics was promoted by social reformers who believed that the reproductive rights of individuals should be subordinated to the rights of an abstract racial collectivity. These debates culminated in several eugenics laws in Japan after World War II, while the People's Republic of China passed eugenic legislation in 1995.

Eugenics, a term coined in 1883 by the English scientist Francis Galton, had its heyday during the first half of the twentieth century. Its proponents claimed that breeding principles such as assortative mating and artificial selection could improve the quality of human offspring. 'Positive eugenics' would ensure that individuals with above-average abilities would breed at a higher rate than ordinary people. 'Negative eugenics' would restrict the reproduction of inferior people, while those having subnormal abilities would have to be physically prevented from perpetuating their 'infirmities.' Mainline eugenicists assumed that intellectual capacity and behavioral traits were inherited and could not be significantly enhanced by education. Supporters included not only scientists, in particular geneticists, but also social reformers, journalists, and other members of the emerging liberal professions. Until recently, however, it was often assumed that eugenics only appealed to extreme right-wing thinkers or organizations in Germany or the United States, and the multifarious dimensions and extraordinary appeal of eugenics to individuals of very different social backgrounds, political convictions, and national affiliations was ignored.

Far from being a politically conservative and scientifically spurious set of beliefs that remained confined to the Nazi era, eugenics belonged to the political vocabulary of many modernizing forces between the two World Wars. It was not only supported by scientific societies, pressure groups, and political institutions in such different countries as France, Brazil, and Sweden, but it also developed a large following in Asia, in particular in China and Japan.

China

Eugenics in Asia is sometimes dismissed as a derivative manifestation of a more 'authentic' discourse in 'the West,' a misleading interpretation that can only impoverish our understanding of the complexities of cultural and intellectual encounters between population groups around the world. When the history of eugenics is explored in a variety of cultural, social, and political contexts, it resists any reductive explanation. As in European cultures, popular hereditarian notions – encapsulated in such sayings as 'like begets like' – helped the emergence and acceptance of eugenic ideas in Asia. Well before Darwin's time, experts in China had attained a sophisticated level of knowledge about selection and breeding, particularly in botany and in goldfish. On the other hand, many medical publications discussed in great detail ways of improving the physical quality of a line of descent while also highlighting the medical dangers to future offspring of careless reproduction. Patrilineal culture, in which reproduction was represented as a potentially dangerous process that should be carefully regulated in order to enhance the possibilities of engendering healthy male descendants, permeated traditional medical literature in imperial China. The individual, seen in relationship to the lineage, was held to be responsible not only for his or her own reproductive behavior, but also for the health of future offspring.

Interest in eugenics, however, only appeared in China in the wake of the reform movement of the 1890s, which openly sought to challenge imperial institutions and orthodox ideology. In search of wealth and power after the country's disastrous defeat against Japan in 1894–95, reform-minded scholars used Darwinian theories – inspired by Herbert Spencer and T. H. Huxley – to propose a radically new political vision based on faith in progress. Slogans such as 'survival of the fittest' and 'struggle for existence' became popular in China, a country increasingly represented as a 'race' that ought to be strengthened in its confrontation with imperialist countries. Public health, individual hygiene, even sexual behavior were targeted by reformers such as Kang Youwei and Liang Qichao as areas in need of state regulation in order to ensure the 'survival of the race' in a context of international competition. Only after the collapse of the empire in 1911 and its substitution by a republic, however, did eugenics develop a much wider following. Eugenics in republican China (1911–49), as in other countries during the interwar period, was not so much a clear set of scientific principles, as a 'modern' way of talking about social problems in biologizing terms: politicians with mutually incompatible beliefs and scientists with opposed interests could all selectively appropriate eugenics to portray society as an organic body which had to be guided by biological laws. Eugenics gave scientific authority to social fears and moral panics and provided an overarching rationale for welfare policies in China. Powered by the prestige of science, it allowed modernizing elites in China to represent their prescriptive claims about social order as objective statements irrevocably grounded in the laws of nature. Eugenics promoted a biologizing vision of society in which the reproductive rights of individuals were subordinated to the rights of an abstract organic collectivity.

Eugenic discourse had different meanings to different social groups in different contexts: it was appropriated and used in a variety of ways to accommodate radically divergent purposes and interests, being flexible enough to unite individuals or constituencies with contradictory agendas. One of the major dividing lines within eugenic discourse was that between a hard approach to heredity, based on genetics in Britain and Germany, and a soft approach combining an emphasis on the environment with hereditarian explanations.

The soft approach to eugenics was common in republican China as well as in other countries as diverse as Brazil and France. Professional groups in China campaigned for better public education in social hygiene and sexual health in the name of race improvement. Moreover, unlike researchers in countries that pursued a hard Mendelian approach, supporters of eugenics in China did not produce significant biological research or statistical studies. An emphasis on the virtues of education accounts for the relative absence in republican China of formal institutions, official organs, or professional organizations centered around the promotion of eugenic ideals. Rather than emanating from a solid organizational foundation, eugenic discourse was supported by a variety of voices in the social field: it attracted popular journalists, social reformers, medical writers, sex educators, university professors, and political ideologues, all attempting to promote medical knowledge and reproductive health for the sake of a more eugenic nation.

Eugenics was actively supported by members of the professional classes, in particular intellectuals, university teachers, sociologists, and biologists, and cut across most political positions, from the fascist core of the Guomindang to the communist theories of Chen Duxiu. Cheap textbooks on heredity and genetics explained to the public the dangers of racial degeneration. Primers, self-study manuals, cheap pamphlets, and 'ABC' introductions to mainline eugenics were published throughout the 1920s. Hereditary principles, endowing sex with a biological responsibility for future generations, were thought to further highlight the need for social discipline. In the 1930s, eugenic discourses became increasingly common in medical circles, 'degeneration' and 'racial hygiene' being the catchwords of the day. There is little evidence, however, to show that the widespread interest in eugenic ideas was translated into actual practice before the advent of a socialist regime, although some isolated cases, such as the execution of lepers by the police in Guangdong province in the 1930s, point to a more radical approach in questions of public health during the Nanjing decade.

The Second World War, followed by the civil war between the Communist Party and the Guomindang from 1945 to 1949, was the main reason why eugenic ideas did not achieve institutional form. The Committee for the Study of Population Policies, organized by the Ministry of Social Affairs in 1941, was the first official attempt to approach eugenics systematically. It recommended the segregation of physically and mentally handicapped persons from the normal population for what was called 'cultural advancement and racial rejuvenation.' Because people were thought to be unequally endowed, the report advocated a differential birth rate: 'Thus viewed,' said the report, "some individuals may have children, others not." The committee – whose members included Chen Da and Pan Guangdan – also encouraged the use of sterilization for the racial rejuvenation of the country and recommended that "(s)teps should be taken to segregate persons of hereditary defects, physical or mental, from the normal population" (Dikötter, 1998).

After the communist victory in 1949, the class bias of eugenics was systematically condemned, while genetics was denounced as 'bourgeois science.' As in the Soviet Union, Lysenko's rejection of Mendelian genetics and T. H. Morgan's chromosome theory became dominant in China in the 1950s and 1960s for ideological reasons. This trend was only reversed after the accession of Deng Xiaoping to power in 1978. Medical experts and population specialists were put into powerful positions as official policies aimed at the limitation of births were initiated by the government from the early 1980s onward. Besides the one-child family program, by which the government only in exceptional circumstances allows parents to have more than one or occasionally two children, an important component of these policies has been the improvement of the quality of newborn babies.

The one-child family policy has effectively prepared the terrain for a better acceptance of eugenic legislation among large sectors of the urban population: many families that are allowed to have one child only are keen to avoid 'defective' births themselves and also tend to favor boys. A prolific medical discourse has responded to some people's concern for healthy offspring, a virtuous child, or even a genetically improved line of descent. Numerous pamphlets of scientific vulgarization thus dispense advice on the art of engendering a prodigy child, suggesting that the eugenist vision of the government is shared to a great extent by a population that is anxious to avoid what are called 'inferior births.'

As the result of such consensus, local regulations have been passed virtually unopposed in a number of provinces and cities since the late 1980s. On 25 November 1988, the Standing Committee of the National People's Congress of Gansu Province passed the country's first law prohibiting 'mentally retarded people' from having children. According to the law, those whose condition was either inherited or a consequence of marriage between close relatives were not allowed to have children; none were allowed to get married unless they had undergone sterilization surgery; those who had married before the promulgation of the law were also to undergo sterilization surgery; and pregnant women suffering from mental retardation were required to have their pregnancies terminated. Individuals who violated the law by allowing 'mentally retarded people' to have children would be punished by both administrative and economic means. Liaoning Province enacted a similar law in February 1990, while Zhejiang Province passed a law on the preservation of 'eugenic health' in June 1992 to combat the increasing incidence of 'cretins,' who were estimated to number more than 300 000. The Henan eugenics law of 1992 requires that if one partner in a married couple suffers from a 'chronic mental disorder' such as schizophrenia or manic depression, that partner shall be sterilized. The implementation of such sterilization laws at the provincial level is advanced by a system of incentives and disincentives. People who refuse to be sterilized or anyone who impedes the sterilization of another may be subject to penalties. For example, Henan's eugenics law

requires prompt reporting of those diagnosed as 'needing to be sterilized' who refuse to accept the operation. Differential fines for noncompliance with 'eugenic measures' are imposed for rural and urban residents.

Under pressure from a variety of lobbies, mainly family planning experts and geneticists, the provincial laws finally culminated in the People's Republic of China passing eugenic legislation at the national level at the beginning of 1995. The law aims to prevent 'inferior births' from becoming a burden on the state and society. Renamed the Maternal and Infant Health Law after protests against a preliminary draft was entitled Eugenics Law (*Youshengfa*), it supports the systematic 'implementation of premarital medical check-ups' in order to detect whether one partner in a couple suffers from a 'serious hereditary,' venereal, or reproductive disorder, a 'relevant mental disorder', or a 'legal contagious disease': it asserts that those 'deemed unsuitable for reproduction' should be urged to undergo sterilization or abortion, or to remain celibate, in order to prevent 'inferior births.' During pregnancy, medical testing will be compulsory and can be followed by termination if the fetus is found to have a serious disorder. The requirement to undergo a premarital health check was abolished in 2003, although in some provinces doctors continue to have the power to prevent people judged to be 'unfit' from marrying.

In the absence of any detailed studies by social scientists and population experts, concrete information on the implementation of eugenic laws is still missing. Foreign observers, however, have emphasize that the implementation of eugenic legislation not only undermines the rights of the person vis-à-vis the state, but it is based on controversial, if not antiquated, theories of human heredity (O'Brien, 1996; Lancet, 1995; Nature, 1994).

Japan

In Japan, eugenics developed gradually after modernizing intellectuals encountered evolutionary theories in the works of Charles Darwin and Herbert Spencer in the 1880s. In the wake of the Meiji Revolution (1868), science and technology were perceived to be the key to national defense in an era of intense imperialism. As in many other countries, racial improvement became a popular notion among leading intellectuals who viewed the world in terms of a 'struggle for existence' among competing 'races.' Belief that individual reproductive decisions should be controlled by the state in the name of collective health flourished during the first half of the twentieth century. In the 1910s, works on eugenics by Umino Kōtoku, Nagai Hisomu, and Saitō Shigesaburō appeared, closely related to an increased interest in birth control, sex education, and public health.

Specialized journals were also established in the 1920s, both the *Yūzenikkusu* (Eugenics) and the *Yūsei Undō* (Eugenics Movement) being crucial in spreading eugenic discourse to a wider audience. As in China, belief in the power of heredity led many followers of the eugenics movement to view venereal disease, tuberculosis, leprosy, and alcoholism as heritable conditions that should be strictly administered by the state in order to prevent the increase of 'inferior offspring.' Eugenics became part of such widely discussed themes as evolution, degeneration, civilization, and modernity, and touched on a wide variety of emerging fields such as maternity, psychiatry, criminology, public health, and sex education (Sumiko and Bartholomew, 1998; Suzuki, 1975).

In Japan, too, the orthodox Mendelian roots of eugenics were often disregarded in favor of a more flexible approach to eugenics that emphasized belief in the transmission of acquired characteristics. Contrary to China, however, eugenic ideas provided a rationale for welfare policies with deep institutional and legal ramifications. A first attempt to implement eugenic legislation took the form of an appeal to the Imperial Diet in the early 1920s (Matsubara, 1998). Proposed by the New Women's Federation, the petition sought to legally restrict the marriage of individuals with venereal diseases as a means of strengthening the nation. Despite support from leading intellectuals as well as members of the Imperial Diet, the petition was not adopted. Only in 1934–35 was a Racial Eugenics Protection Bill introduced by the Imperial Diet, which proposed the sterilization of persons with mental and physical diseases deemed to be hereditary, including serious criminals who were claimed to transmit their 'bad traits' to their descendants. The Japanese Association of Racial Hygiene, on the other hand, campaigned for a eugenics law limiting sterilization to patients with hereditary diseases. Headed by Nagai Hisomu, the group drafted a sterilization bill in 1936 based on the German Sterilization Law of 1933. Only in 1940, however, was a National Eugenic Bill passed, its enactment leading to heated controversy among government representatives and members of the medical professions. Objections came both from nationalists defending the need of the state to increase its population and from geneticists questioning the scientific validity of eugenics. While the bill came into effect a year later, the government was compelled to prevent the enforcement of eugenic sterilization: by the end of the Second World War, only 454 operations were performed under the eugenics law.

After Japan's defeat in the Second World War, questions of 'racial survival' surfaced again in the context of occupation by the Allied Powers (Matsubara, 1998). As abortion was legalized, a broad consensus emerged within medical and political circles around the need to prevent not only 'inferior descendants,' but also the reproduction of patients deemed to have 'infectious diseases' such as leprosy and nonhereditary mental disorders in the name of 'eugenic protection.' Supported by conservative advocates as well as socialist politicians, the Eugenics Protection Law (*Yūsei hogohō*) of 1948 gave legitimacy to eugenic sterilization without the consent of the patients considered. The law remained the foundation for eugenic sterilizations throughout the second half of the twentieth century: involuntary operations reached over 16 500 cases by 1996, while a further 1500 leprosy sufferers segregated in sanatoriums had been sterilized with consent by 1992. In 1996, the law was finally revised and renamed the Maternal Protection Law, while the eugenic clauses on sterilization and abortion were deleted. Even after these legal changes, however, the government refused to critically examine the practice of compulsory sterilization on eugenic grounds, despite severe criticisms from a variety of civil rights groups (Norgren, 1998).

Conclusion

The legacy of eugenics is still highly visible today in many countries, including less researched parts of the world outside 'the West.' As the examples of China and Japan demonstrate, eugenic ideas elaborated during the first half of the twentieth century continue to carry scientific and legal credibility. While crude eugenic programs sponsored by the state may have become less frequent today, eugenic ideas are prominent in contemporary debates about new medical techniques for isolating and manipulating genes. During the heyday of eugenics, assumptions about 'race,' gender, and class pervaded genetic research: with the revolution in human genetics since the end of the twentieth century, social and political factors continue to exert formidable influences on scientific research. Eugenic laws have been passed at the provincial and national levels in China, a country that has curbed the reproductive freedoms and civil rights of individuals on behalf of the 'public good.' Even in democratic countries like Japan, however, marginalized people may be treated in a discriminatory way, as social prejudice and economic interest influence the genetic information made available to families, employers, hospitals, insurance companies, or welfare systems. As much as 'ethnicity' is a concept that is sometimes used as a thinly disguised substitute for the outdated notion of 'race,' discredited eugenic ideas have thrived on rather than suffered from the revolution in human genetics in some parts of the world: East Asia is no exception.

Bibliography

China's misconception of eugenics. Nature 367, 1994, 1–2.

Dikötter, F., 1998. Imperfect Conceptions: Medical Knowledge, Birth Defects and Eugenics in China. Columbia University Press, New York.

Matsubara, Y., 1998. The enactment of Japan's sterilization laws in the 1940s: a prelude to postwar eugenic policy. Historia Scientiarum 8 (2), 187–201.

Norgren, T., 1998. Abortion before birth control: the interest group politics behind postwar Japanese reproductive policy. Journal of Japanese Studies 24, 59–94.

O'Brien, C., 1996. China urged to delay 'eugenics' law. Nature 383, 204.

Sumiko, O., Bartholomew, J.R., 1998. Eugenics in Japan: some ironies of modernity, 1883–1945. Science in Context 11 (3–4), 545–565.

Suzuki, Z., 1975. Geneticists and the eugenics movement in Japan. Japanese Studies in the History of Science 14, 157–164.

Western eyes on China's eugenic law. The Lancet 346, 1995, 131.

Eugenics in Europe

Richard S Fogarty, University at Albany, SUNY, Albany, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article surveys the history of eugenics in Europe, noting its origins in the growing comprehension of the power of science to understand the human body and in the increasing ambitions of the state and the social and scientific elites to exert control over the social body. The article then surveys major events and personalities of the eugenics movement in various European countries over the course of the nineteenth and twentieth centuries, placing the movement firmly in its political, social, and cultural contexts.

Although eugenics was not exclusive to Europe, it is there that it originated and found some of its most fully-elaborated theories and practices. Precursors to this so-called science of improving human populations through genetics and theories of heredity had existed in the West since antiquity, with Plato's advocacy of judicious breeding of the citizens of his ideal *Republic* constituting one of the earliest and most famous examples. What might be called a proto-eugenic discourse emerged during the seventeenth and eighteenth centuries, as ideas on human perfectibility, theories of generation and inheritance, and considerations of demography and national strength all became growing concerns with intellectuals, scientists, and the ruling elites. The nineteenth century saw accelerating attention to these matters, as the simultaneous development of powerful new scientific discourses and emergent centralized and bureaucratized nation-states gave powerful impetus to ambitions and attempts to manipulate human health for the supposed betterment of large populations. The eugenics movement, both in its intellectual and scientific forms and its manifestations in state policies, reached its apogee in the first half of the twentieth century. The frenzy of racial utopian thinking that marked the Second World War and the Holocaust seriously undermined faith in eugenics, as did other postwar developments in science and politics. By the end of the twentieth century, eugenics was so thoroughly discredited that its name disappeared from acceptable mainstream discourse. Yet its legacies have not disappeared and continued advances in the ability to manipulate human life at the genetic level mean that the problems raised by eugenics are far from vanishing for good.

An Englishman, Francis Galton, a cousin of Charles Darwin, coined the term eugenics in his 1883 book *Inquiries into Human Faculty and Its Development*, deriving it from the Greek word *eugenes*, literally meaning 'well-born.' Galton, who had been interested in questions of heredity since his work *Hereditary Genius: An Inquiry into Its Laws and Consequences* of 1869, described eugenics as meaning "good in stock, hereditarily endowed with noble qualities." He and those who came after him described eugenics as a science, relying on the latest insights into the development of life provided by Darwin's revolutionary theory of evolution and an increasing understanding of genetics. The perceived social benefits of the science of eugenics, that of applying ideas of heredity in practice to improve the health of the social body, provided eugenicists with a wide audience and in some cases official government support. But this very practicability subjected eugenics to even more effective social and cultural pressures than is usually the case in the formation of bodies of scientific knowledge (even if social and cultural forces are never absent from the development of science). Steven Jay Gould has pointed out that certain disciplines are particularly subject to external social pressures because they have a low ratio of data to social impact – they tell more about social change than anything else, and they formulate their original questions "in such a restricted way that any legitimate answer can only validate a social preference" (Gould, 1996: pp. 54–55). The social preference of eugenicists was always very clear: the prevalence of certain elements of society or sections of the population over others, whether defined by racial, ethnic, or class superiority. Some individuals and groups were more 'fit' than others and deserved to propagate their kind, while other 'less fit' individuals and groups should face restrictions or prohibitions on their procreative activities. In extreme cases, eugenicists contested and even sought to prevent the continued existence of the 'less fit.'

Cultural and political factors and social preferences decisively influenced the format of eugenic research and activity in a variety of national contexts. Different social concerns over matters of race and class differentiated these separate national eugenics movements, even as some commonalities and international scientific networks linked them in guiding research imperatives and focusing eugenic policy. The desire to bring growing expertise to bear on public policy also played a role, because medicine was becoming increasingly professionalized over the late-nineteenth and early twentieth centuries. It is no accident that leading eugenicists were most often crusading medical men, not laboratory geneticists. In short, eugenics is a prime example of what Tzvetan Todorov calls "scientism … the use of science to establish an ideology" (Todorov, 1993: p. 114), through the increase of the persuasive power of medico-scientific discourse on social problems during this period. The idiosyncratic science of eugenics provides an explicit example of how closely the scientific endeavor is intertwined with immediate social, political, and cultural imperatives. Eugenics was an important force in nearly all industrialized societies and certainly in all of Europe by the early twentieth century. In each case, the particular needs and

desires of the prevailing national political and social regime seeking to classify, control, maximize, and marshal its population and human capital, determined its eugenic policies to a large extent.

Origins

The Scientific Revolution of the seventeenth century had profound effects not only upon the European intellectual climate but also on Western political development. By the eighteenth and nineteenth centuries, scientific knowledge was increasing in volume, scope, and sophistication at a remarkable rate. Many new fields of inquiry rapidly developed and institutionalization and professionalization soon followed with dramatic growth. At the same time, scientists were busy forging links with the emerging centralized and bureaucratic states of Europe. Those in power soon realized that a partnership between science and politics could offer a mutually beneficient relationship, and politicians and legislators began to lean heavily upon scientific experts, methods, and data in formulating social policies. Effective social policy was certainly an urgent priority in this age of industrialization and centralization. Modernity brought with it deep tensions that constantly threatened the tenuous hold that the new bureaucracies maintained upon both the peoples over whom they ruled and the seemingly impersonal accelerating forces of history. The new sciences appeared to offer a coherent method for resolving these tensions within the social order.

The French thinker and historian Michel Foucault designated "what brought life and its mechanisms into the realm of explicit calculations and made knowledge-power an agent of transformation of human life" as "biopower"(Foucault, 1990: p. 143). This marked the emergence, simultaneous with the advent of modernity, of myriad new technologies designed to exert control over and channel the energies of individual bodies in the ostensible interests of the larger social body. The state, or groups representing the interests of the state, wielded "a certain mode of detailed political investment of the body," and a ubiquitous "new microphysics of power," acting as disciplinary authorities, closely monitoring people's health and movements (Foucault, 1977: p. 139). This attention to individual bodies was predicated on a view of the body as a machine, which could be manipulated and fine-tuned to maximize its utility and output. The goal of these controlling apparatuses in focusing on particular human beings was to maximize the utility of populations as a whole. This "biopolitics of the population" sought to promote the health and well being of society, which is more than just the sum of its parts. Society itself, in this line of reasoning, is analogous to a living organism. It is subject to diseases and cures and could be healthy or ill, strong or weak, normal, or pathological.

New technologies of power made the extension of biopower possible during the nineteenth and twentieth centuries. The ruling elites and the states they served, both from within and outside formal political bodies (professionals such as scientists and doctors, scientific societies, expert advisors, and technocrats) were above all concerned to divide the populations in which they lived and over whom they had control into 'normal' and 'pathological,' or at least to place individual bodies in a hierarchy of value. These were conscious and unconscious strategies of power, to be sure, but they were also responses to the political, economic, and class tensions called into being by modernity. In this sense, eugenics became a perfect example of both the powers and the anxieties unleashed by modernity. Not only did the emergence of mass culture, industrialization, and the many other novel problems and conditions of modern life generate new concerns about the social body, but these same conditions made imperative the development of new methods for controlling and organizing individual bodies. Political and social authorities began to concern themselves with the examination, classification, diagnosis, organization, and treatment of people, both as single, distinct entities and as aggregate populations.

The connection between individual bodies and the aggregate population which they constituted was never quite so important as it became during the rapid economic and demographic expansion of the late eighteenth century. In 1798, Thomas Malthus published his thoughts on the relationship between the two in his *An Essay on the Principle of Population*. In it, he postulated that human beings' biological reproductive drives were inexorably propelling them to multiply beyond the limits of their available sustenance and that rapid technological advances only postponed the inevitable overpopulation of the planet. Malthus reversed the conventional wisdom of the time concerning healthy bodies. According to his argument, healthy individual bodies did not contribute to the overall well-being of society. Rather, virile and well-fed people reproduced at a dangerously higher rate. This thinking was in line with an elite Victorian tradition of suspicion and fear of the awesome and seemingly autonomous power of certain segments of the population. By the middle of the nineteenth century, educated and upper class Europeans would come to see the unchecked proliferation of members of the laboring class as a threat to (their) civilization.

Yet inter-racial worries laid the groundwork for these intra-racial concerns and decisively shaped eugenic anxieties and discourse. Scientific language and methods already informed nineteenth-century thinking about race when Charles Darwin's *On the Origin of Species* appeared in 1859 and provided racial theorists with a new and powerful way to conceptualize racial divisions. Writers such as Arthur de Gobineau had always been adamant about placing the various human races in a qualitative hierarchy with the white race at the top. He justified fears of miscegenation with empirical data demonstrating the inferiority of both nonwhites and persons of mixed race. But now racial theorists combined Darwin's theories of evolution and natural selection with their observations about the relative levels of 'civilization' achieved by the various races, maintaining that whites represented the most advanced stage of human evolution. They cited the technological and artistic achievements and near worldwide hegemony of Europe as conclusive proof of white superiority and even argued that the white race constituted a different, more highly developed species of human being. This flew in the face of Darwin's skepticism about truly 'specific' differences among different groups of human beings. In his 1871 *The Descent of Man*, Darwin cited cross-fertility and the 'fusion' of humans of different races who lived in close proximity and who chose to intermingle and

breed across racial lines, as conclusive evidence that the various races among humans did not constitute different species. Still, racialist thinkers ignored Darwin's conclusions on race while retaining his evolutionary theories in order to invest their opinions with the impressive mantle of the new science.

So underlying eugenics was a peculiarly scientific form of social Darwinism. Eugenicists advocated methods and policies based on Herbert Spencer's "survival of the fittest," and rather than natural selection, eugenicists sought to promote the concept that artificial selection would safeguard and improve the health of the social body. They justified these artificial interventions into the process of evolution by citing the artificially dysgenic conditions of modern civilization. Galton himself vaunted these interventions as "more merciful and not less effective" than natural selection, which relied upon "excessive production and wholesale destruction," while eugenics sought to bring "no more individuals into the world than can be properly cared for and those only of the best stock." But why was the need for eugenics so increasingly urgent during precisely this period in European history? Ultimately, eugenicists transmuted concerns over racial competition and interracial contamination into fears of intra-race decay, known by the French term *dégénérescence*, or degeneration. Darwin's theory of evolution, combined with ideas about human heredity, led many to believe that a 'devolution' of the human species was possible. Writers like Max Nordau fulminated against the corruption of modern culture that they saw around them and determined that the origin of this degeneration must be found in human heritability. Nordau built upon the work that originally introduced the term, B.A. Morel's 1857 *Traité des dégénérescences*, and defined the condition as "a morbid deviation from an original type," which was transmissible and would spread more widely and increase in intensity with the propagation of the original carrier's descendants. The conditions of modern life, centered in the frenetic, unhygienic environments of ever larger cities promoted this decay, which then developed further in the unhealthful altered circumstances of natural heredity. Nordau's special contribution, aside from popularizing the notion in his international best-selling *Degeneration* of 1892, was to diagnose the condition in the new medico-scientific discourse that resonated so loudly in *fin-de-siècle* Europe. With such a powerful and compelling analytical tool, he and others were able to attribute biological causes to perceived social ills.

Although Nordau found degeneration everywhere, even in the ranks of the artistic and literary elite, discourse about degeneration came to focus primarily on the unfortunate poor and the demonstrably 'feeble-minded,' mentally and emotionally disabled persons. The intensifying pressures of modernity in Europe over the nineteenth century led to political struggles that thrust 'the masses' decisively into the political arena. Spectacular examples during the upheavals of 1848 and the 1870 Paris Commune frightened the newly empowered bourgeoisie and provoked intellectuals to view the large pool of wage-laboring poor as a biological as well as a political threat. Thus, the social and political elites could characterize social problems in the comfortingly objective terms of science. A medical diagnosis of inherent and heritable degeneracy conveniently and compellingly explained the impersonal and inscrutable forces of history. Concerns about degeneration proliferated, and penetrated into works of popular literature, the social sciences, and political debates, as well as the medical and biological sciences.

Discourses about degeneration, a social anxiety articulated as a medico-biological concept, predictably took on different characteristics in different national contexts. In Great Britain, the term resonated distinctly with class tensions. For example, psychiatrist Henry Maudsley focused on the rampant mental deficiency among the working poor in urban industrial centers. The squalor and inbreeding of the slums in which they lived, he argued, rendered the poor pathologically degenerate, and the unchecked twin processes of degeneracy and procreation threatened to engulf British society. The widely publicized poor condition of recruits from the working class serving in the Boer War (1899–1902) convinced many of the urgency of stopping the apparent decline of the nation's biological stock. In France, the birthplace of the theory of *dégénérescence*, concerns mixed with anxieties over the nation's declining birthrate, especially in comparison with a more fecund and recently united Germany (the process of German unity itself crystalizing in the humiliation of France in the 1870 Franco-Prussian War). Fear of population decline put the focus on specific socio-medical problems, such as tuberculosis, alcoholism, and pre- and post-natal health, the amelioration of which could stem the decline and even regenerate the nation. In Italy, the most salient evidence of degeneration seemed to be intractable criminality. Unrest and crime in this recently and uneasily unified nation-state led legendary criminologist Cesare Lombroso to search for ways to identify hereditary criminals by their physical and mental characteristics. He and men like Enrico Ferri recommended the segregation of 'criminal types,' preventing their reproduction, thus checking the effects and spread of biological-social decline. In the Scandinavian countries of Demark, Norway, and Sweden, rapid modernization and its attendant problems, declining birthrates, and emigration seemed to suggest an anthropological approach to degeneration, in this region where the discipline was particularly well developed. Racial hygiene movements organized to address the issue, and Denmark's 'Anthropological Committee' eventually became the leading advocate in that country for eugenic solutions to the problem of degeneration.

Degeneration was the problem for which eugenics was the solution – the rationale was that the medical problem of degeneration called for a medical cure. Eugenicists, often doctors or men with a background in medicine or public hygiene, identified degeneration as the root of all modern social and political problems. Eugenics offered a utopian way out of the dead end into which modernity seemed to be driving human civilization and its highest expression in Europe's advanced industrialized societies. Ironically, and perhaps ingeniously, modernity could cure itself, since eugenics represented to its adherents the most modern, cutting edge science of the day, built upon the solid foundations of Darwinian biology and pioneering work in the understanding of human genetics. The problems of modernity, once attributed to biology and inheritance, could be made subject to scientific study and treatment. But it was clear to many eugenicists that it was going to be necessary to enlist the coercive power of the state in order to enact specific eugenic measures.

Eugenics in Formation, 1904–33

If the origins of eugenic thinking lie far back in the Western tradition (in dreams of perfecting individual human bodies by intervening in the procreative process, thereby perfecting the social body as a whole) and if the nineteenth century saw the crystallization of the component elements of eugenic thinking (social Darwinism, racism, and theories of degeneration) through scientific advances in understanding human biology, then the early twentieth century was an era of institutionalization and the articulation of eugenics as a body of both thought and social policy. Galton, writing in 1908 about the science and movement he had founded, succinctly described the twin imperatives of eugenics: firstly, "to check the birth-rate of the Unfit, instead of allowing them to come into being, though doomed in large numbers to perish prematurely"; secondly, to improve "the race by furthering the productivity of the Fit by early marriages and healthful rearing of their children." These two forms became known as 'negative' and 'positive' eugenics and they combined in different ways in different national contexts, depending upon the social and cultural forces driving interest in eugenics and the policies that flowed from them. In some places negative eugenics held sway, while in others positive eugenics prevailed. In others still, the two coincided, overlapped, or even mixed. Eugenic thinking was never uncontested, to be sure, but in virtually all European societies during this period the idea prevailed that social and civilizational health was largely reducible to heredity, and that humans could intervene effectively in the hereditary process by controlling the reproductive behavior of certain elements of society.

Positive eugenics, the less well-known and seemingly more benign of the two, involved the encouragement of healthy, eugenically sound marriages, the improvement of pre- and post-natal care, campaigns against diseases such as alcoholism and tuberculosis, and a host of other programs. Proponents designed these measures to maintain the health of the fit, so that they would reproduce healthy and numerous offspring, and to improve the health of the less fit members of society, who would then pass on to their descendants their improved condition (or at least would not pass on degenerative diseases). Negative eugenics, on the other hand, involved the prohibition of dysgenic unions, where one or both of the partners was allegedly biologically unfit. Attempts to prevent undesirables from propagating their kind usually took one or more of four separate forms: marriage restrictions, immigration restrictions, sexual segregation, and sterilization. Clearly, the state would be required to play a preponderant role in implementing these policies since they all required coercive authority, legal regimes, broad control over large segments of society, and restrictions on individual freedom.

The state would have to provide the legal and most of the administrative infrastructure for the execution of eugenic policies. Nonetheless, the status of eugenics as a broad-based social and scientific movement meant that individuals and interest groups outside the state would play a very large role in originating and elaborating eugenic policies at the level of ideas, pressuring for their implementation via government action, and staffing the bureaucracies that carried them out. Social reformers of all stripes advocated for eugenic policies, but the real impetus for the movement came from medical biologists, physicians, and academics in the social sciences. This mix of advocates, along with the generalized social anxieties of the bourgeois elites faced with an accelerating modernity, gave eugenics its distinctive quality of being at once a social movement and a science. The ways that social and cultural imperatives decisively influenced the direction of the science of eugenics and eugenic policies is clear from the different forms that they took in the various countries of Europe during the first half of the twentieth century.

Although rooted in the nineteenth century, eugenics as a movement began and rapidly spread at the very beginning of the twentieth century. Books, concepts, organizations, and even legislation laying the foundations for the movement appeared earlier, but the year 1904 saw the constitution in London of the Eugenics Record Office, the first institution to deploy the term by name and take the science explicitly as its guiding principle. The United States would emerge nearly simultaneously as one of the most important sites of eugenic thinking and policy, with its own Eugenics Record Office, at Cold Spring Harbor in New York, opening in 1910 (the British institution had changed its name in 1907 to the Francis Galton Laboratory for the Study of National Eugenics). Over the following decades, eugenics would also develop as an important force in societies across Asia, Latin America, and beyond, but the movement originated and was anchored earliest and most firmly in Europe.

Great Britain, of course, as the home of both Darwin and Galton and the site of the original Eugenics Record Office, blazed the path that other nations would follow, though each would eventually pursue a eugenics that best suited their social, cultural, and political circumstances. In the United States the eugenics movement under the influence of men like Charles Davenport displayed a strong Mendelian bent and a concern with genetics, and under the influence of the nation's distinctive history and culture, displayed a clear preoccupation with race and immigration. Eugenicists in Britain viewed class divisions as the most pressing problem to which eugenics might offer a solution. British science not only resisted Mendelianism, but because the British class system was notoriously rigid, the problem of degeneration appeared in this context as a diseased urban blight brought on by rapid industrialization and the need for cheap labor, spreading out into the 'healthy' population. Race and ethnic diversity, at this stage of British history and outside of Ireland, did not raise critical problems domestically. British eugenicists, in short, worked in the tradition of Galton, who stressed the heredity of both physical and social characteristics and warned of the consequences of a more fertile but less intelligent underclass. Class anxiety about degeneration and solutions by means of eugenics later manifested themselves as warnings about internal, class-based social and hereditary decay, in the writings and pronouncements of writers like H.G. Wells and Havelock Ellis.

But men of science were influential in the British eugenics movement. Karl Pearson was the most famous, a mathematician and biometrician who taught statistics at University College London, a protégé of Galton and head of the Galton Eugenics Laboratory for almost three decades. He also held the first Galton Chair of Eugenics at University College London, which Galton had endowed upon his death in 1911. Britain was an international center of eugenics research and activity,

seeing the foundation of the Eugenics Education Society in 1907 (becoming simply the Eugenics Society in 1926), with Galton as honorary president. From 1911 to 1928, Charles Darwin's son, Major Leonard Darwin served as the Society's President. The First International Congress of Eugenics in London in 1912 paid tribute to Britain's importance to the international movement. Pearson, who was not a member of the Eugenics Society, founded the important international journal, *Annals of Eugenics*, in 1926.

Still, despite all of this intellectual activity, enacting eugenic policies proved difficult in Britain, much more so than in the United States. An 1895 Connecticut law prohibited marriages of epileptics, imbeciles, and the feebleminded, while Indiana's 1907 Compulsory Sterilization Law pioneered legal sterilization, targeting confirmed criminals, idiots, imbeciles, and rapists. The two laws together established the United States as the most fertile ground, along with Germany eventually, for extreme negative eugenics measures. In Great Britain, in contrast, even lawmakers sympathetic to eugenic ideas found restrictive measures unpalatable, and politicians on the side of labor were concerned with the effect of negative measures on the working poor. Thinkers of a socialist bent, such as George Bernard Shaw, still accepted the widespread discourse on eugenics, but advocated more positive means. In Shaw's case, this involved removing rigid class distinctions and the social restrictions they imposed to allow for better, more eugenic pairings.

France saw an important eugenics movement arise during this very same period, but eugenics took a very particular form in that country. Even more than in Britain, French scientists, thinkers, and legislators resisted negative measures and, in fact, provided one of the most congenial homes in the world to positive eugenics. The particular social and political problems that characterized French discourse about degeneration – fear of a declining birthrate and international marginalization – along with patriotic sentiment, led French eugenicists to subscribe to the theories on heredity propounded by their countryman Jean-Baptiste Lamarck, rather than those of the Englishman Darwin. Lamarckian biology led them to believe that physical changes wrought by the environment on an individual organism during its lifetime were heritable by its offspring. This meant that positive eugenic measures implemented to improve the health of individual members of society would improve the biological capital from which future generations would draw. In a strictly Galtonian or Mendelian view of heredity, however, positive measures could improve the health of the already fit and encourage them to produce more numerous fit offspring, but the unfit carried, fatally and permanently, the seeds of future decay and had to be stopped with negative measures from procreating. The French eugenic movement's positive approach aligned it with the pronatalist and social hygiene movements. Even advances in understanding genetics – Mendelianism and Weismann's theory of the continuity of the 'germ plasm' – failed to alter the approach of most French eugenicists. The movement, highly influenced by the social and political imperatives to maintain and increase the quantity of the population, rejected negative measures that would prevent procreation. It sought instead to improve the quality of the population, which according to Lamarckian principles would, in turn, increase the quantity.

Physician Adolphe Pinard, a leading and typical figure in French eugenics, built his career and fame on public advocacy of pre- and post-natal health. This science of healthy babies, known as *puériculture*, stressed the importance of parental health not just during pregnancy, but even at the moment of conception. Parents, after all, were in a position to pass on their acquired and maintained health to their children. Enthusiasm brought back by the French delegation to the First International Eugenics Congress in London in 1912 led to the founding of the French Eugenics Society at the end of that year. Pinard was a founding member and vice president, and the organization published a monthly journal, *Eugénique*. Physicians constituted more than half the group's founding members. The Society's agenda included government action to manage sexual activity, marriage, birth and life expectancies via *puériculture*, social hygiene, and eugenics.

Suspension of activities during World War I slowed the group's momentum and caused it to lose membership. The 1924 creation of the National Office of Social Hygiene incorporated a part of the Society's agenda and in 1926 *Eugénique* ceased publication. Yet the Great War's appalling demographic impact inspired French eugenicists to reemphasize social hygiene and positive eugenics in order to rejuvenate a French population facing greater and more specific threats than a vague, generalized 'degeneration.' Now, more than ever, the production of more and healthier children was critically urgent and if the Society was organizationally weak, other institutions and actors were now more open to eugenic ideas. Public discussions of eugenics proceeded, even if they did not rejuvenate the Society. In 1926 Pinard, now a member of the Chamber of Deputies, proposed legislation to require all French males to submit to a premarital medical examination. But years of debate held up the bill's progress. It would not be until later in the 1930s and during World War II under the Vichy regime that interest in eugenics (even negative eugenic proposals unthinkable in earlier periods dominated by fears of depopulation) had concrete results.

Eugenic discourse in Germany differed significantly from that of its French neighbor. Germans did not fear a declining birthrate since their nation enjoyed a demographic boom in the late nineteenth and early twentieth centuries. Thus, German eugenicists, instead of molding their scientific views around cultural fears over numerical population decline, were subject to social Darwinist and racist cultural anxieties, talking freely about the survival of the fittest and about restricting the reproduction of certain members of society. Nowhere is the impact of cultural imperatives, on eugenics as well as on scientific discourse in general, more clear than in the contrast between the French and German 'styles' of eugenics.

Despite this contrast, both movements concerned themselves deeply with what they both called hygiene. But while eugenics in France and elsewhere often had close links with social hygiene movements, the German eugenics movement was rooted in *Rassenhygiene*, or racial hygiene, betraying an underlying concern with safeguarding both racial health and purity. Alfred Ploetz, who coined the term, evinced a keen interest in manipulating heredity to improve the health of the German population, or 'race.' In 1904 he founded the eugenicist journal *Archiv für Rassen- und Gesellschaftsbiologie*, and in the following year, the Society for Racial Hygiene

(*Gesellschaft für Rassenhygiene*). Although some German eugenicists were skeptical of restrictive, negative measures, worrying like their French counterparts over the size of their population as well as its quality, there was still a marked tendency, heightened by the intense attention to race, to think in terms of preventing the propagation of the unfit while promoting the fecundity of the fit. Also strongly present in German eugenics was a technocratic and managerial logic, best exemplified by physician Wilhelm Schallmayer, who strove to improve national efficiency through racial hygiene. Although he rejected the ardent racism of many other German eugenicists, he advocated a rational management of population via cost-benefit analyses to determine the future social cost of dysgenic reproduction, along with government action to prevent or reduce this dysgenic behavior.

But the logic of extending state power to increase 'national efficiency,' through the rational management of the population as a resource, involved dividing individuals into categories of valuable and less valuable, or valueless. Once the Nazis later created an environment where racial criteria almost exclusively guided these determinations, the logic of German eugenics played out to its murderous end. What Foucault called the "biopolitics of the population," the regulation of individual bodies through medico-scientific discourses in the supposed interests of the larger social body, was evident in German eugenics from an early stage, long before the Nazis made the approach notorious. German eugenicists, even those eschewing racism like Schallmayer, saw it as their duty, and the state's duty, to watch over and supervise, to maintain and regulate the biology and reproductive behavior of each individual as part of the species, or at minimum of the larger national community.

The three Great Powers of Britain, France, and Germany were certainly not the only homes of eugenic activity in the early twentieth century. Russia, Austria-Hungary, and Italy, themselves sometimes also ranked as Great Powers, all hosted important eugenics movements. Eugenics in Russia before the Bolshevik Revolution of 1917 rejected the emphasis on class and race found in other national movements, emphasizing the role of the environment in shaping heredity, much as in France. Geneticists such as Konstantinovich Kol'tsov and Yuri Alexandrovich Filipchenko were the key figures in the movement by the 1920s, and a move toward Mendelianism and a 'socialist' eugenics characterized their efforts. But Lamarckianism was more in line with official Soviet ideology, negative measures never found wide acceptance, and social hygiene was of continuing importance to the movement. Eugenics in the Austro-Hungarian empire before its dissolution in 1918 was closely aligned with the German *Rassenhygiene* movement in a 'Greater German' approach to questions of race and heredity, while public health policies in the Austro-Hungarian empire indicated that leaders in Vienna sought to enhance imperial control and interests via intervention in the health of individuals and the larger social body. In Italy, although Lombroso's ideas about hereditary criminality were internationally influential, the movement as a whole hewed to an environmental view that rejected negative eugenics. Moreover, a strong Catholic tradition, which showed itself hostile to eugenics in many countries, precluded strict biological determinism. Italian biology as a whole tended not to stress racial purity and to accept variation and diversity as healthy.

Switzerland and the Scandinavian nations were also active sites of eugenic activity. Under the influence of eugenicists, the Swiss canton of Vaud, in 1928, became the first European polity to enact a compulsory sterilization law aimed at the mentally ill. One of the earliest European societies devoted to discussing and promoting eugenic ideas was Denmark's *Den Antropologiske Komité* (Anthropological Committee), founded in 1904, which later became a private eugenics society. All three Scandinavian countries of Denmark, Sweden, and Norway, as well as the Nordic states of Finland and Iceland and the Baltic states of Latvia and Estonia, introduced eugenics-inspired sterilization laws. Almost all of these, however, came into force during the 1930s. If the period from a decade before the First World War through the late 1920s saw the flourishing of eugenic thinking and activity in scientific journals, research programs, and societies, it was the following decade, (which also saw the rise to power in Germany of Adolf Hitler) and the years of the Second World War that constituted the height of eugenic policymaking and enforcement by European states.

Eugenics in Power, 1933–45

The story of eugenics in power during the 1930s and 1940s is more quickly told than that of the movement during the period of its formation. This is not because eugenics activity tapered off – on the contrary it grew in intensity and concrete effects – but because the story is, in many cases, a straightforward one of policymaking and state action. This is most obviously true in the case of Germany, under the explicitly racist and eugenicist Nazi regime from 1933. But the repeated crises of the 1930s and the outbreak of war in 1939 stoked desires and opened up opportunities in many countries to act to achieve eugenicist ends, whether it was manifested as an urge to protect populations from the alleged perils of hereditary decay or as an impulse to realize utopian visions of racial and ethnic purity and human perfection.

Germany is the most important and notorious case, and if an account of the origins and early stages of eugenics must begin with Great Britain, the story in the 1930s and 1940s must start with Hitler's ascent to power in 1933. The first and most general effect of the Nazi regime on German eugenics was to weld together the concern of some eugenicists with welfare measures designed to improve public and health and hygiene in the interest of the future of the population, and the obsession of the Nazis and not a small number of eugenicists with race. Only a few months after coming to power, Hitler promulgated a sterilization law aimed at those with supposedly hereditary diseases such as feeblemindedness and other mental or physical impairments. The eugenicist logic underlying the measure was clear from its name, Law for the Prevention of Hereditarily Diseased Offspring (*Gesetz zur Verhütung erbkranken Nachwuchses*), and eugenicists indeed took a central role in its framing. Ernst Rüdin was a key figure, as he had been in the movement since he helped Ploetz found the Society for Race Hygiene in 1905. Eventually, some 375 000 people were legally sterilized, virtually all without their consent and some without even knowing they had undergone the procedure. The eugenic logic of sterilization

combined with one of the Nazis' most ardent racial obsessions in 1937, when 385 children born of German mothers and French colonial soldiers during the post-WWI occupation of the Rhineland were sterilized.

Nazi measures to segregate racially undesirable people from the ethnically pure German population, in order to prevent the pollution of the gene pool through uncontrolled reproduction, displayed the same blend of eugenic logic and strategies combined with Nazi racial logic. The 1935 Nuremburg Laws that forbade sexual relations between Germans and non-Germans (Jews, primarily, but also other so-called racial inferiors) took inspiration from the work on interracial mixing by Eugen Fischer, who in 1927 had become the first head of the Kaiser Wilhelm Institute for Anthropology, Human Heredity, and Eugenics in Berlin. Rector of the University of Berlin from 1933, Fischer remained closely tied to the Nazis and their racial policies, and in 1935 upon promulgation of the Nuremberg Laws, he publicly praised them and thanked Hitler for having presented to eugenicists such an opportunity to "make their research useful to the general public." Not all German eugenicists, before or during the Nazi period, shared Hitler's lethal racism, but many did. Fischer was typical for his collaborative and even participatory stance in relation to Nazi policies.

Notoriously, these would lead to euthanasia and mass murder. The T4 Program, in which doctors and technicians murdered some 70 000 mentally ill and disabled Germans between 1939 and 1941, was the precursor to even larger administrative, technological, and industrialized slaughter during the war. The links between eugenicists and this program, and the later plan to murder millions of Jews and other so-called 'subhumans' in order to eliminate the dangers they posed to the biological health of a greater German and Aryan empire, are not entirely straightforward. Personnel involved in the T4 Program, drawing directly on their previous experiences, helped develop the methods and infrastructure of industrialized killing in the death camps of the Eastern Front, once it became clear that mass murder by shooting and starvation did not work quickly or efficiently enough. Many in the eugenics movement denied that they or their ideas were instrumental in these events. Nonetheless, a number of physicians and scientists, clearly influenced by eugenic ideas and related genetic research, were indeed involved and saw their careers flourish under the Nazis and even in the camps themselves (Dr. Josef Mengele's murderous experiments in Auschwitz being perhaps the most extreme and notorious example). In any case, eugenic policies and these extreme and grotesque crimes partook of the same logic: a desire to divide human beings into 'hereditarily healthy' (and thus desirable) and 'hereditarily pathological' (and thus undesirable) groups. Eugenicists and committed Nazis (when they were not one and the same, which they often were) both sought, sometimes by different means, to promote the existence of the first group, and to isolate, diminish, and eliminate the second.

In France, the 1930s saw a shift among some eugenicists toward accepting a greater degree of restrictive intervention to assure the quality of the population, with quantity relegated as a lesser priority. Adolphe Pinard's efforts in the 1920s aimed at mandating a premarital examination for all French males to assess their genetic suitability for marriage and procreation. The proposal stalled in the legislature for years, but it demonstrated a growing openness to negative eugenic measures. As the 1930s wore on, physicians and others prominent in the eugenics movement advocated immigration restrictions to protect France's biological capital, and some even promoted birth control despite continuing concerns over falling birthrates and their effect on France's international position in an increasingly dangerous Europe. Charles Richet, president of the French Eugenics Society immediately after the First World War, exemplified this new attitude. His *La sélection humaine* (1919) accepted Lamarckian-style environmental factors as having powerful long term effects on species, but also counseled shorter term restrictive measures to improve the race based on the iron laws of heredity. His proposals included the removal and quarantine of the sick, prohibitions on interracial marriage, "the elimination of the mentally deficient" (through neglect, not euthanasia), and the prohibition of marriage for the 'abnormal.' The most internationally famous of French eugenicists, Alexis Carrel, winner of the Nobel Prize in Physiology or Medicine for 1912, spent the bulk of his career in the United States at the Rockefeller Institute for Medical Research. His international best-seller *Man, the Unknown* (1935), contained both positive and negative eugenic ideas, but he was very clear about how, in his view, the unfit burdened society. In one famous passage he suggested "humanely and economically" disposing of criminals and the criminally insane, "in small euthanasic institutions supplied with proper gases."

It is not clear that this one statement in a long and wide-ranging book is enough to place Carrel firmly in the camp of German eugenicists and physicians deeply compromised by their participation in Nazi euthanasia and genocide programs. It is true that extreme negative eugenics remained relatively marginal in France throughout the 1930s, but Carrel's views and career nonetheless reflect well the trajectory of French eugenics after 1940. After Paris fell to the Germans, Carrel met with the head of the collaborationist Vichy government, Marshal Pétain, and agreed to head a new *Fondation française pour l'étude des problèmes humaines*, founded in November 1941. In the occupied portion of France, the German authorities provided the impetus for indigenous French racism and anti-Semitism to be expressed in eugenic terms. The anti-Semitism of the period cannot be reduced to eugenics, but eugenics often provided support and a vocabulary to buttress even the wildest anti-Jewish propaganda and the most sinister anti-Jewish measures. For example, Louis Darquier de Pellepoix, head of the notorious *Commissariat général aux questions juives*, a center for racist and anti-Semitic propaganda, defended the roundup and deportation of Jews as a 'public hygiene' measure. For its part, Vichy France instituted the most enduring eugenic legacy, the final passage, in 1942, of the long-stalled law instituting a premarital exam, which remains in force to this day. Its sponsors proclaimed it as "the first time [that] a eugenic measure appears in French legislation." Even here, though, the largely positive French style of eugenics was in evidence, since the mandate for the premarital exam was embedded in a series of other measures promoting social hygiene and maternal health. Moreover, though the exam was mandatory, it placed no restrictions on the right to marry nor did it require the parties to divulge the results, even if they revealed a serious infectious disease. It was in this sense merely informative, and

sought, in the words of the law's authors, "only to confront the future couple with their conscience and their responsibility." Thus, even under Vichy, French eugenics avoided the hereditary and racial utopianism of extreme negative eugenics, although that vision triumphed on the other side of the Rhine and spread its malign influence throughout much of Europe.

In Europe broadly, the 1930s and 1940s did see an increase in the ability of eugenicists to influence public policy. Great Britain, the original home of eugenics, was largely an exception. The Eugenics Society remained active throughout the period and concerned itself with race mixing, birth control, and even sterilization, but the political and social environment remained inconducive to the implementation of explicitly eugenic measures both inside and outside the context of the state. A number of other European nations adopted sterilization legislation during the 1930s. But this would prove to be the apogee of eugenics in Europe, even in Germany. It was in part because of the triumph of eugenic thinking, married to racism and totalitarian government, in Germany that eugenics saw a slow but sure decline after the end of the Second World War.

Eugenics in Decline

Events in Germany and German-occupied Europe discredited attempts at genetic social engineering, though eugenics really disappeared only slowly and incompletely over the second half of the twentieth century. The 'Doctors' Trial' at Nuremburg in 1946–47 exposed many of the worst medical abuses under the Nazis, and the close collaboration of physicians and scientists in the implementation of the Nazis' racist policies, attempts at biological engineering, and murder. Nevertheless, prosecution for sterilization was difficult because of its technical legality even before the war began, and because the defense could easily point to similar, and similarly ambitious and widespread, eugenic policies and practices in countries like the United States. Some doctors and scientists associated with eugenics and Nazi policies continued to pursue their work and even held important positions in political and academic institutions for decades after the end of the war.

Eugenic thinking and practice also persisted beyond 1945 in many other places in Europe. Vaud, the only Swiss canton to institute legal sterilization, in 1928, continued the practice for decades, officially repealing its legal basis only in 1985. In France, the prestigious *Institut national d'études démographiques* (INED) formed in 1945 from elements of Alexis Carrel's Vichy-sponsored foundation and with some of its personnel. In the immediate postwar period, the population geneticist Jean Sutter (1910–70), at INED, continued to promote eugenic solutions to traditional French problems associated with public health and population. In Great Britain, the Galton Chair of Eugenics at University College London remained a preeminent position in the academic study of genetics, the Eugenics Society continued its work, and eugenic ideas exerted some influence over medical curricula and even social policy. Italian eugenicists reframed their field of interest as 'genetic demography,' continuing to work largely in the same vein as before the war. Sterilizations continued apace in Scandinavia through the 1970s. In Greece, interest in eugenics actually increased after 1945.

Yet these circumstances should not obscure the long-term decline of the movement. As the full extent of Nazi atrocities in the name of biological engineering sank in, reinforced by continuing revelations, eugenics, 'racial hygiene,' and associated projects suffered by association. This was one reason that eugenicists increasingly rebranded their work: genetic demography in Italy, euthenics in France, social or population biology and a host of other terms, some of them certainly legitimate academic or scientific fields in their own right, elsewhere. And eugenics societies and institutes either faded away or changed their names and even research foci. The German Society for Racial Hygiene ceased to exist after the war, the Swedish Institute for Race Biology became a Department of Medical Genetics at Uppsala University in 1959, and the Francis Galton Laboratory for the Study of National Eugenics renamed itself the Galton Laboratory of the Department of Human Genetics and Biometry at University College London in 1963. The British Eugenics Society waited until 1989 to change its name to the Galton Institute, but change it did. Other organizations outside of Europe – in Canada, Hong Kong, Australia, and the United States – changed their names between 1950 and 1973, dropping references to race, hygiene, and eugenics. In particular, the term eugenics became and remains, largely taboo. The major exception to this was the Hellenic Eugenics Society, founded in Greece in 1954. But even formerly eugenic policies shifted in focus, if not in notorious effect. Sterilizations continued into the 1970s in Scandinavia, it is true, and under the same laws inaugurated by advocates of eugenics. But operations decreasingly occurred because of eugenic concerns or motivations, and instead took individual medical problems or voluntary birth control as their primary rationale. By 1977, Denmark, Norway, and Sweden had all repealed eugenic legislation, replacing it with laws allowing for voluntary sterilization. The symbolic end to academic eugenics had occurred much earlier in its original home of Great Britain, with the ascension of Lionel S. Penrose to the Galton Chair in eugenics. He was, in fact, skeptical about the claims of eugenics to the status of science and was openly critical of some of its specific claims. In 1954, he changed the name of the journal founded by Karl Pearson, *Annals of Eugenics*, to *Annals of Human Genetics*, and changed the name of his position to the Galton Professorship of Human Genetics.

It is important to note that the horrors perpetrated by the Nazis and others were not the only factors contributing to the decline of eugenics. The end had come earlier in the Soviet Union, for example. Stalin's hostility to the claims of eugenics as incompatible with socialism and his own visions for a transformed society brought the curtain down on eugenics decisively after he rose to absolute power in 1929. Human genetics in general suffered from association with 'fascist views' in the 1930s, and from Trofim Lysenko's quixotic and deeply destructive crusade against Mendelianism through the 1950s. The Soviet takeover of Eastern Europe after 1945 spelled the end of eugenics there as well. Other factors eroding the authority of eugenics, at least in its most explicit forms, included the discrediting of race as an idea in science and medicine, as well as in anthropological understandings of human characteristics and behavior; a growing awareness of racial injustice propelled by anticolonial and civil rights movements; and increasing skepticism among broad swaths of

world public opinion about absolutist scientific claims and about the potential for harm that often accompanies scientific and technological advances. One must also not underestimate the effect of increasingly sophisticated scientific understandings of the workings of heredity and genetics. In this view, as scientists like Lionel Pearson discovered, the claims of eugenics appear crude and simplistic – not just wrong, but harmfully so.

Legacies, Echoes, and the Future of Eugenics

At the beginning of the twenty-first century, eugenics seems to many to have disappeared from view. In some ways, it is the science that dare not speak its name, in that eugenic concerns lie only in the background of some modern research in the fields of biology, genetics, demography, and public health. Yet this is most often a distant heritage and the personnel and research agendas deeply compromised by close association with eugenics have by now largely passed from the scene. Eugenics now manifests itself as a term of abuse, a way to make sometimes exaggerated (though sometimes not so exaggerated) criticisms of others' social views and political beliefs. Thus the government of Niger condemned a French author's denunciation of that country's high birthrate and extreme poverty as a 'crime,' and his invocation of sterilization as a solution to this global problem, as 'fascist,' declaring that his comments could not be more racist and they praise eugenics. This reaction demonstrates the reverberating legacy of German and other eugenicists' fateful linkage of racism, government action, and social and population policy.

The legacy of eugenics is also present in contemporary Europe because its victims have not disappeared from the scene. This is all the more true since eugenic practices such as sterilization continued up through the very recent past. Individuals sterilized against their will, and often even without their knowledge, have mobilized to force the governments that perpetrated these acts to recognize the injustice and, in some cases, pay reparations. Some efforts have been unsuccessful, as in Switzerland. Sweden has paid damages to victims, but the practice of sterilization, arguably against one's will, has apparently continued. In 2013, 142 transsexuals brought suit against the Swedish government seeking compensation for having been forced to accept mandatory sterilization in order to undergo sex change operations between 1972 and 2012.

Eugenics also often emerges in present-day discussions of prenatal diagnosis, abortion, and individual and medical decisions about reproduction. For many, the question now is not one of states seeking to extend biopower through control over the reproductive activity of individuals and thus over the larger social body, but of individual parents and doctors making choices about reproduction that modern medical science now makes possible. Beyond prenatal diagnosis, some observers see the specter of eugenics in the growing power science exerts in fertility treatment, genetic manipulation, and so-called 'designer babies.' What is still at issue is the core problem that led scientists, doctors, and others in the past to think in terms of race betterment and eugenics, that is, the impulse to make value judgments about life at the most basic biological level, to arrange life hierarchically into categories of greater or lesser degrees of worthiness, and to act accordingly.

See also: Darwinism; Eugenics as an International Movement; Eugenics, History of; Evolution, History of; Race: Genetic Aspects.

Bibliography

Adams, M.B., 1990. The Wellborn Science: Eugenics in Germany, France, Brazil, and Russia. Oxford University Press, New York.
Bashford, A., Levine, P., 2010. The Oxford Handbook of the History of Eugenics. Oxford University Press, New York.
Bowler, P.J., 1993. Biology and Social Thought, 1850–1914. Office for History of Science and Technology, Berkeley.
Burleigh, M., Wippermann, W., 1991. The Racial State: Germany, 1933–1945. Cambridge University Press, Cambridge, UK.
Carol, A., 1995. Histoire de l'eugénisme en France: les médecins et la procréation XIXe-XXe siècle. Éditions du Seuil, Paris.
Foucault, Michel, 1977. Discipline and Punish: The Birth of the Prison. Vintage, New York.
Foucault, Michel, 1990. The History of Sexuality, Volume I: An Introduction. Vintage, New York.
Gould, Stephen J., 1996. The Mismeasure of Man. W.W. Norton, New York.
Kevles, D.J., 1985. In the Name of Eugenics: Genetics and the Uses of Human Heredity. Harvard University Press, Cambridge, MA.
Nye, R., 1984. Crime, Madness, and Politics in Modern France: The Medical Concept of National Decline. Princeton University Press, Princeton, NJ.
Paul, D.B., 1995. Controlling Human Heredity, 1865 to the Present. Humanities Press, New Jersey.
Pick, D., 1989. Faces of Degeneration: A European Disorder, c. 1848-c. 1918. Cambridge University Press, New York.
Proctor, R.N., 1988. Racial Hygiene: Medicine under the Nazis. Harvard University Press, Cambridge, MA.
Schneider, W.H., 1990. Quality and Quantity: The Quest for Biological Regeneration in Twentieth-century, France. Cambridge University Press, Cambridge, UK.
Todorov, Tzvetan, 1993. On Human Diversity: Nationalism, Racism, and Exoticism in French Thought. Harvard University Press, Cambridge, MA.
Weindling, P., 1989. Health, Race, and German Politics from National Unification to Nazism, 1870–1945. Cambridge University Press, Cambridge.
Weiss, S.F., 1987. Race, Hygiene, and National Efficiency: The Eugenics of Wilhelm Schallmayer. University of California Press, Berkeley, CA.

Eugenics in the Americas

Patience A Schell, School of Language and Literature, King's College, University of Aberdeen, Aberdeen, Scotland, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses eugenics in the Americas – a scientific ideology that profoundly has affected the history of this region – and identifies common themes and varied examples for comparison. These examples demonstrate the influence of the 'well-born science' on medicine, psychiatry, education, and public policy. This article includes a broad overview of eugenics in South and North America, pointing out areas of differences and commonalities. The chapter also discusses examples of formal eugenics organizations, in their national and pan-American context and the utility of eugenics for its practitioners. The article finishes discussing examples of eugenic practice through immigration policy, *puériculture* programs, and sterilization.

Introduction

From Argentina to Alberta, eugenics as a science and an approach for understanding human populations profoundly has affected the history of the Americas. Despite the great diversity of geography, historical experiences, and human populations that characterizes these two continents, the Americas in the late nineteenth and into the twentieth centuries shared a history of colonialism, unease about indigenous, former enslaved and mixed-race populations, attempts to manage immigration, and social problems associated with growing urban populations. The spread of eugenics, despite its diverse manifestations, responded to this situation. Discussion of eugenics, in terms of both policy and practice, in South and North America offers evidence of the diversity and adaptability of this movement to 'breed better humans.' Although it is impossible to present a complete overview of eugenics in the Americas because of space constraints, this article will approach eugenics through the identification of common themes and present varied examples to offer comparative insights that also indicate the elasticity of the concept. These wide-ranging examples also demonstrate the influence of the 'well-born science' on medicine, psychiatry, education, and public policy.

This article will start with a broad overview of eugenics in South and North America, pointing out areas of differences and commonalities. Next, this article discusses examples of formal eugenics organizations, in their national and pan-American context, and the utility of eugenics for its practitioners. The article ends by looking at examples of eugenic practices through immigration policy, *puériculture* programs, and sterilization.

An Overview of Eugenics in the Americas

In Latin America, eugenics was fed by and grafted onto concerns emerging in the nineteenth century about the racial composition of these new nations, in which 'problem' populations, including indigenous, African-ancestry or mixed-race people, were considered a hindrance to national modernization projects. These concerns about the population mix were compounded with fears that a tropical climate in warm countries, such as Brazil, was inimical to progress and contributed to racial degeneration. Nineteenth-century elites despaired of building modern, productive, civilized nations from such diverse human populations. To improve their populations, countries like Argentina, Mexico, Peru, and Brazil attempted to encourage, with differing degrees of success, European (ideally northern) immigration; public health and education programs responded to these perceptions of a problematic population mix (Hochman et al., 2010, p. 495; Orbegoso, 2012, p. 232). Yet by the early twentieth century, elites also had begun to believe that making a modern nation required some incorporation of these populations. In Chile, for example, questions about national identity emerging from warfare, reflections that were prompted by a centenary of independence, and fears that European immigration undermined the national character, fostered the concept of the 'Chilean race,' which symbolically incorporated the poor, mixed-race, and indigenous populations into the new imagined national stock (Subercaseaux, 2007, pp. 31–32, 63). Across the region, a wave of cultural nationalism shifted depictions of Latin America's non-White populations from being a hindrance to progress to being an essential component of a unique national identity; nonetheless, elites continued to anticipate that African and indigenous elements of the population eventually would be 'whitened' away (Chomsky, 2000, pp. 420–422). The former Mexican minister of education, José Vasconcelos, in *La raza cósmica* (1925) (*The Cosmic Race*), lauded the *mestizo* (originally meaning mixed indigenous and Spanish) as the result of an aesthetic, spiritual eugenics that combined the best aspects of Europeans, indigenous people, Asians, and Africans. Nonetheless, his vision was based on the assumption that the European contribution redeemed this 'cosmic race' (Stern, 2000, pp. 61–62).

Attitudes toward the poor and ethnically diverse populations had begun to shift from the later nineteenth century. Increased urbanization and nascent industrialization, however, brought such issues as contagious illnesses, like tuberculosis and syphilis, alcoholism, infant mortality, labor activism, and unhygienic housing to the attention of the public and medical professionals. Lumped together, these problems were referred to as the 'social question.' Eugenic influences were frequently evident in public health responses to this social question. Nancy Leys Stepan has argued that the version of eugenics that flourished in Latin America, influencing state policies, national debates, public health, and the sciences, was influenced heavily

by neoLamarckian ideas of the inheritance of acquired characteristics (1991, p. 76). Within a neoLamarckian approach, alcoholism, for example, could be seen as a eugenic issue because, although it resulted from social problems, social action to address it would have positive eugenic results. Keeping 'nature' and 'nurture' together united eugenics, public health, and sanitation programs, as any intervention in aspects of 'nurture', over time, could benefit the 'nature' of the national population (Stepan, 1990, pp. 119–122; see also Stepan, 1991, pp. 67–84). Focus on 'nurture' in Latin America meant that puériculture, the scientific cultivation of the child, and homiculture, which turned the puériculture approach to the study of adult life cycles, were both popular manifestations of eugenics evidenced in public, childhood, and maternal health programs (Schell, 2010, pp. 478, 485). But the later shift to Mendelian genetics did not undermine the case for eugenics. Recent research on Brazil has indicated that proponents of Mendelian genetics also supported eugenics, even as they distinguished between problems due to genetic inheritance and social conditions (Hochman et al., 2010, p. 494).

Part of the malleability of eugenics, and its resilience as a concept, was the elasticity and context specificity of defining who was 'unfit' (see Levine and Bashford, 2010, p. 7). Even as ideas of 'better birth,' which underpinned eugenics can be found throughout the Americas, the expression of that sentiment and the populations of 'undesirables' against whom it was directed was highly particular. Bearing this elasticity in mind, Latin American countries offer interesting eugenics case studies, as their populations (usually largely composed of indigenous, African-ancestry, or mixed-race peoples) were considered 'dysgenic,' that is, they reproduced 'undesirable' traits, in Western Europe and the United States. Thus, focus on eugenic policies in their various guides direct us to those sectors of the population that prompted anxiety for the state, the medical professions, educationalists, and journalists, among others.

In Canada and the United States, eugenics was one expression of the new public health movement, stemming from a belief that understandings of heredity could improve the health and welfare of the population at large. Part of this public health movement allowed the discussion of previously taboo topics, such as birth control and venereal disease, under the guise of social reform (McLaren, 1990, pp. 28, 68). Eugenics also built on fears, known as 'race suicide,' that the Anglo-Saxon population was not reproducing quickly enough when compared with other groups, such as immigrants and non-White populations (Roberts, 1997, 60–61; McLaren, 1990, pp. 44–45). Eugenics was adopted readily in Canada and the United States due to widespread notions of scientific racism, which held that biologically existing 'races' had innate, inheritable characteristics that determined capacities. In the United States, theories of scientific racism also justified segregationist policies toward the African-American population (Roberts, 1997, 61). Moreover, as in Latin America, in Canada and the United States, the Aboriginal–indigenous peoples were viewed through a colonial mentality that held that this population represented an obstacle for society's advancement, doubted the intellectual capacities of these peoples, and judged them unfit for modernity (Strange and Stephen, 2010, p. 534). Thus, in these countries, eugenic discourse and practice permeated medicine, public health, immigration policy, and popular culture. Ian Robert Dowbiggin argued that, in Canada and the United States, in the early twentieth century, the profession of psychiatry was so infused with eugenic beliefs that the task of the historian is to "determine the different *shades* of eugenicism" (1997, p. xi; italics in original). That task is appropriate for the Americas at large.

Institutions, Organizations, and Practitioners

One important site for the creation of eugenics knowledge in the United States, which then fed into circulation throughout the Americas, evolved from a research project funded by the Carnegie Institute, in 1904, into the Eugenic Record Office, at Cold Spring Harbor, Long Island. The project was run by University of Chicago professor and biologist Charles B. Davenport, whose team researched the genetic history of extended families believed to be carriers of hereditary abnormalities, evidenced in "laziness, mental retardation, and immoral habits, as well as high fecundity." On the basis of his team's work, Davenport concluded that heredity not only determined instances of diseases, such as hemophilia, but also alcoholism, insanity, and criminality. The Eugenic Record Office trained experts (including 250 field workers) for and supplied research to the growing eugenics movement as well as disseminated findings in such publications as the *Eugenical News* (Roberts, 1997, pp. 61–62. Quotation p. 62).

But eugenics in the United States moved well beyond the Atlantic seaboard (Stern, 2005b). The 1915 Panama Pacific Exhibition, held in San Francisco, offers an interesting example of eugenics beyond the US East Coast, but also as a growing popular science. Simultaneous to the Panama Pacific Exhibition, San Francisco hosted the National Conference on Race Betterment; eugenicists wanted to take advantage of the potential audiences. The Pan-Pacific exhibition's organizers were delighted, as the two events both sought to promote civilization and progress, and they designated one week of the Pan-Pacific Exhibition as 'Race Betterment Week.' The 'Race Betterment Exhibit' at the Pan-Pacific Exhibition sought to prove declining standards of racial health nationally and became one of the most popular displays. For these eugenicists and supporters, 'race betterment' depended in large part on ensuring that middle-class white women took up their social responsibility as women to be mothers who reproduced healthy citizens within marriage. This cementing of selected women's primary social responsibility as home and family came at a time, the 1910s, when new opportunities for women in education, work, and leisure threatened to distract them from their reproductive duties (Kline, 2001, pp. 14–19, 29). Concerns about women and sexuality were fundamental themes to eugenics in the Americas at large.

According to Stepan, the start of organized eugenics in Latin America came a few years later, with the founding, in Brazil in 1918, of the São Paulo Eugenics Society, whose 140 members were almost all physicians. The organization's arguments that linked long-standing public health concerns, such as tuberculosis, alcoholism, and venereal diseases, to the need to make Brazil a eugenic nation, gained positive coverage in the local

and regional press. Although this society lasted only a few years, eugenics approaches and discourse permeated public health discussions and publications (Stepan, 1991, pp. 47–50). In Mexico, a normalization of eugenics was evident at the 1921 Mexican Congress of the Child, which debated maternal health issues and population 'whitening,' as well as narrowly voting in favor of forced sterilization of criminals. By the end of the decade, the Mexican Society of Puériculture, founded in 1929, included a eugenics section that discussed heredity, sex education, and birth control. Interestingly, the founding members of the Mexican Eugenics Society (Mexico City, 1931) were from the Puériculture Society's eugenics wing. The new organization, which also debated sex education, birth control, eugenic procreation, and marital health, drew its membership from high echelons of politics, public health, the judiciary, and criminologists, and its influences were evident in state campaigns against venereal disease and alcoholism, as well as in the 1935 legal requirement for a prenuptial certificate (Schell, 2010, p. 485; Stepan, 1991, pp. 55–58; Suárez y López Guazo, 2000, pp. 5, 7, 11, 13).

In Argentina, formal eugenics emerged after physician Víctor Delfino attended London's First International Congress of Eugenics. Delfino founded the Argentina Eugenics Society in 1918, which made immigration control one of its principal campaigns amid concerns about the impact of immigrants on national culture. By 1932, in a swing to the right, the Argentine Association of Biotypology, Eugenics, and Social Medicine was founded. Close cultural links between Argentina and Italy, a result of that worryingly large-scale Italian immigration, helped account for the inclusion of 'biotypology,' which was an Italian science that posited that "human populations could be divided into distinct types with their own characteristic illnesses and psychological makeup." Understanding biotypology would allow a state to use its human resources most effectively. The association offered training in biotypology's methodologies and diagnostics, including running a clinic for determining biotype and offering appropriate interventions (Stepan, 1991, pp. 58–61. Quotation p. 60).

Although eugenics policies and practice undoubtedly had national manifestations, eugenicists in the Americas formed part of a pan-American movement, most readily evident through international meetings. For example, when the first pan-American Conference on Eugenics and Homiculture was held in Havana in 1927, US eugenicist Charles Davenport, along with his ally leading Cuban eugenicist Domingo F. Ramos Delgado, lobbied delegates to approve a pan-American 'Code of Eugenics and Homiculture,' which mandated classification of all residents of the Americas into categories of 'good,' 'bad,' and 'doubtful' to determine who could marry, have children, and migrate. For Davenport, the code addressed concerns about the racial composition of the US's neighbors, as these potential migrants threatened eugenics efforts within the United States. Other delegates, however, argued against the implication that Latin American populations were dysgenic, and the code was defeated in favor of a homiculture approach to eugenics, which endorsed public health and education measures (Guy, 1998, p. 272; Schell, 2010, pp. 478–479; Stepan, 1991, pp. 178–182). At that same conference, delegates from Latin America voted that each state had the right to control immigration in ways that were appropriate to its own ethnic mix and that would not undermine national sovereignty (Stepan, 1990, p. 134).

By the time Buenos Aires hosted the second pan-American eugenics conference in 1934, the international context had changed, especially with the National Socialist rise to power in Germany (1933). Although the eugenics movement had grown in Latin America, it was situated firmly within preventive medicine, social welfare, and only voluntary measures offering advice on eugenic reproduction. The 1930 papal encyclical *Casti Connubii* (*On Christian Marriage*), condemning eugenics and sterilization, also diminished the influence of 'negative' eugenics, such as sterilization, in Latin America, as well as French-Canada, while strengthening Catholic opposition at that end of the eugenics spectrum (Klausen and Bashford, 2010, p. 105; Stepan, 1990, p. 135; Strange and Stephen, 2010, p. 526). Stepan concluded that the firm marginalization of the US version of negative eugenics at the Buenos Aires conference ended the pan-American eugenic movement (1991, pp. 61, 183–186). But beneath this formal support for 'positive' eugenics in Latin America, historians also have found evidence of negative eugenics practice, discussed later in this article. In the United States and Canada, in contrast, negative eugenics was supported by both policy and practice.

To understand how eugenics came to have such a pervasive influence, Dowbiggin argued that, in the case of psychiatry in Canada and the United States, professional reasons have to be placed at the forefront. As the practice of psychiatrists was largely limited to state asylums and public health organizations, where their work could seem like being glorified guards of incurable patients, eugenics offered a reassuring explanation for psychiatrists' failure to 'cure' institutionalized patients and additionally gave them an excuse to move beyond the asylum (1997, pp. x, 27, 50). McLaren concurred, arguing that for medical doctors in Canada, advances in medical science in the late-nineteenth century, including 'germ theory' (the discovery in this period that microorganisms cause some infectious diseases), improved the status of physicians, who claimed unique knowledge for diagnosing illness and recognizing threats to public health. Eugenicists' arguments that often social and health problems were hereditary, not environmental, further strengthened doctors' professional authority. Meanwhile, the state bolstered the status of medical professionals, turning to these experts to understand and resolve problems associated with industrialization, foreign immigration, urbanization, and low fertility of the Anglo-Saxon population (1990, p. 29). Stepan found parallel transformations in the medical profession in Latin America. The number of doctors increased, their training became more professional and formalized, and they worked in a context in which there were growing expectations that medical science could design successful interventions to stop the endemic diseases decimating populations. In Brazil, for instance, successful public health campaigns against yellow fever, smallpox, and the bubonic plague earned acclaim for the Oswaldo Cruz Institute, which eventually became federally funded, and hero status for its eponymous director (1991, pp. 42–43). These medical heroes contributed new ways to resolve problems associated with segments of the population considered dangerous or threatening.

Targeting the 'Unfit'

Throughout the Americas, gender became one of the major axes around which eugenics rotated: Gender norms, male and female social roles, ideas about normal and abnormal sexuality, and reproduction all have gender at their core. Women's behavior and their procreative choices were of particular concern for eugenicists. Yet the eugenics movements also provided new professional opportunities for middle-class women. Women became both objects for eugenic intervention, and practitioners of eugenics (Kline, 2001, p. 5; see also Stern, 2010).

Multiple examples indicate how eugenics frequently targeted women, particularly because of their role in childbirth and childrearing. In Chile, a 1911 publication by A. Moraga Porras (*Influencia de la cultura física en la formación del carácter. Educación física de la mujer*, or The Influence of a Physical Fitness Culture on Character Formation. The Physical Education of Women), argued that women's physical health mattered more than their intellect because of their procreative role and their responsibility to provide the nation with healthy children who would improve the 'Chilean race' (Subercaseaux, 2007, p. 57). Likewise, in Argentina, the majority of eugenic medical studies were oriented toward women, as there was concern among demographers that population decline was due to women's failure of maternal feeling or eugenic unfitness (Eraso, 2007).

In both Canada and the United States, as will be discussed, concern about the 'inappropriate' sexual behavior of young working-class women, who then were defined as 'morally' unfit, made them vulnerable to eugenic interventions. For instance, in Canada morally unfit women could be committed to the Ontario Provincial Asylum for Idiots, despite the lack of formal eugenic policies as a framework for doing so. A 1924 amendment to the Juvenile Delinquents Act had given the Canadian courts authority to take action to police morality, and young women, Aboriginal or Anglo-Celtic working class, were most at risk of state policing (Strange and Stephen, 2010, pp. 527–528).

In Canada, eugenics also singled out the 'feeble-minded' and immigrants as dangerous for the eugenic nation, giving the state responsibility to act in the national interest. Helen MacMurchy, who had been the first women to earn a medical degree from the University of Toronto, was appointed in 1913 to be Canada's first 'inspector of the feeble-minded.' She blamed the country's 'epidemic' of feeblemindedness on alcoholism, crime, infectious diseases, philanthropy, and immigrants (Strange and Stephen, 2010, pp. 526–527. Quotation p. 527). Eugenic concerns also influenced Canadian immigration policy; a 1906 change to immigration law prohibited the entry of potential immigrants who were feeble-minded or who were morally dubious criminals. This amendment additionally gave the Department of Immigration the authority to deport immigrants about whom fears arose retrospectively. If one medical doctor raised concerns about mental illness, potential mental illness, or moral failings, these concerns constituted grounds for deportation. Mental fitness was not the only grounds for restricting immigration; starting from 1885, immigration legislation included race as a criterion on which entry could be refused. A 1923 federal law explicitly prohibited all Chinese immigrants (Strange and Stephen, 2010, p. 529). Nancy Ordover's argument that, in the United States, anxiety about Mexican immigration resulted from a "dual image of immigrants as both containing foreign bodies, in this instance disease [such as bubonic plague], and constituting foreign bodies within the national entity" (2003, p. 37; italics in original) was also the case in Canada. But controlling immigration was a common eugenic measure in other parts of the Americas: In the late 1930s, the Brazilian constituent assembly voted to restrict immigration by setting racial quotas and tests of fitness for entry; one of the goals of this legislation was to protect the eugenic processes already at work in Brazil (Stepan, 1990, p. 143).

Positive and Negative Eugenics in Latin America

Although immigration controls could limit external dysgenic influences, 'puériculture,' a form of eugenics widespread in Latin America, which inextricably linked mother and child, would help improve the national stock. Puériculture also built on gender stereotypes in which women were given a central role as the mothers of families who made the foundations of a strong state. In Argentina, members of the Association of Biotypology presented a plan to create a state organization for the welfare of mothers and infants, passed by congress in 1936, which resulted in the National Department of Maternity and Infancy. The new department had oversight of eugenic intervention before marriage, before conception, and in pre- and postpartum care, and it responded to concerns about high rates of infant mortality and 'weakening' diseases among the adult population (Eraso, 2007, pp. 793–794, 802–803). In Brazil, as well, where slavery had only been abolished in 1888, eugenicists promoted puériculture and wise marriage choice for young women (Stepan, 1990, pp. 123–124). Mexico, too, embraced puériculture. The Department of Public Health's School Hygiene Service implemented many of the resolutions from the First Congress of the Child, held in 1921, including training women and girls in puériculture, building playgrounds, and sending health visitors to supervise young women (pregnant or not) in Mexico City's poor neighborhoods. Later programs included tracking the progress of local mothers and children, as well as teaching families about basic hygiene. These programs fed into the Mexican Society of Puériculture, mentioned earlier, which was founded in 1929, and the origin of the Mexican Eugenics Society (Stern, 1999, pp. 373–374).

Although eugenics policies in Latin America generally fell into the category of 'positive' action, historians have shown how 'negative' eugenics also influenced practice. In Argentina, although permanent sterilization was banned in principle, the procedure could be carried out legally in cases in which there was a danger to the mother. In practice, this caveat allowed for permanent eugenic sterilizations of women during the 1930s, when a more conservative national mood contributed to greater concerns about population fitness. Interpreting this exception broadly, doctors carried out sterilizations in cases in which pregnancy was believed to complicate existing conditions, such as diabetes or tuberculosis. But there were other cases in which sterilizations were carried out based on less-evident health risks. For instance, in 1928, a leading

obstetrician sterilized a woman who had leprosy, using eugenic reasons as one of his justifications. On the basis of her research, Yolanda Eraso concluded that eugenic sterilizations, presented as procedures on medical grounds, were more practiced than has been recognized and the legal framework offered doctors considerable discretion in performing the operation. At the 1940 Fourth Argentine Congress of Obstetrics and Gynecology, Uruguayan participants were more open about practicing sterilization, which was sanctioned by the Uruguayan Commission of Hygiene (2007, pp. 796, 813–819, 822).

In neighboring Brazil, even as elites feared that the population was too small and too diseased for the national space, psychiatrists, criminologists, and experts in legal medicine discussed sterilization, along with birth control and abortion. One account indicates that the director of the National Insane Asylum had sterilized women patients suffering from 'sexual' illnesses. Largely, however, the conservative Brazilian medical establishment was against sterilization on any grounds (Stepan, 1990, pp. 124–125). Meanwhile, in North America, Veracruz State passed Mexico's only eugenic sterilization law in 1932, but there is no evidence it was ever implemented (Schell, 2010, pp. 485–486). Puerto Rico, a US colony, in contrast, served as an island laboratory for birth control testing, funded by the US federal government, and for widespread sterilization amid eugenic concerns about 'overpopulation,' which had no base in empirical data. By the mid-1960s, an estimated 34% of women of childbearing age had been sterilized. How voluntary these sterilizations were remains subject of debate, but alternative nonpermanent forms of birth control were not available, limiting the choices open to women (Schell, 2010, pp. 481–484; see Briggs, 2002; Ramírez de Arellano and Seipp, 1983).

Further north, however, both the United States and Canada practiced eugenic sterilizations widely and over decades. In Canada, during the 1920s and 1930s, calls for the sterilization of the dysgenic became a nationwide campaign, with the support of leading politicians and business representatives and the Canadian National Committee for Mental Hygiene. There were some critics of the growing influence of eugenics. In 1922, an economist at the University of Alberta, H.S. Patton, complained that the popularity of eugenics was a result of the growing power of "professional experts who sought to control scientifically every aspect of life." Nonetheless, at the University of Manitoba, in the 1930s, V.M. Jackson, author of *Rat Ranching in Manitoba*, taught eugenics problems in his zoology courses. Science took control of many people's private lives when the Western provinces of British Columbia and Alberta passed sterilization laws. In the case of Alberta, the legislation built on farmers' support gained through the logic of stockbreeding. A board was created to review the cases of those about to be discharged from asylums, authorizing sterilization as a preventive measure in cases in which a disability might be passed on to children; and, from 1937, that sterilization could be without the patient's consent. Strange and Stephen compared Alberta's local political culture to the US Deep South, noting there was no oversight of the board; the head of the board (1928–1965) authorized more than 3200 sterilizations, about 60% of which were actually undertaken. During the 1950s and 1960s, Alberta's sterilization rates were higher than California's rates, discussed later, and the statue was not repealed until 1972. Strange and Stephen concluded that, although these sterilization laws were not overtly designed to target particular groups, determined by race, gender, or age, in practice young women of Aboriginal background were most likely to be sterilized, partially due to racism and anxiety about young women's sexual and reproductive morality. In conclusion, Strange and Stephen have expanded on Alexandra Stern's argument that Western eugenics in the United States has particular attributes, finding that the same holds true for Canada (Strange and Stephen, 2010, pp. 530–534; Stern, 2005b; see also McLaren, 1990, pp. 26–27, 95–106. Quotation p. 27).

In the field of eugenic sterilization, however, the United States was such a pioneer that its influence has been traced to later German policies (Thomson, 2010, p. 120). Indiana passed the first US law of sterilization in 1907 prompted by studies of allegedly defective families, and in the context of the public health focus of the Progressive Era. Indiana's law mandated the sterilization of criminals, 'idiots,' 'imbeciles,' and some rapists (Dowbiggin, 1997, p. 77). Two years later, California became the third state to legalize forced sterilization, in asylums and prisons; the legislation was expanded in 1913 and 1917, adding protection for surgeons and emphasizing a eugenic rationale for the procedure. California defined sterilization as a public health measure, undertaken for the state's collective good. By 1921, 80% of sterilizations nationally had taken place in California, in part because, as Stern argued, the state had a broad eugenics movement with supporters from education, medicine, government, and the agricultural sector. In 1927, the Supreme Court sanctioned California's use of eugenic sterilization, in *Buck v. Bell*, drawing on states' rights to vaccinate under public health guises as the legal precedent. With the public health endorsement of the Supreme Court, by 1932, 27 states had passed sterilization laws and 3900 procedures had been performed (2005a, pp. 1129–1130).

In the subsequent decades, California continued to top the nation's sterilization tables, even in per capita comparisons, and the procedure's supporters argued that the practice must be extended beyond the state's institutions. In these state institutions, such as mental hospitals, sterilizations were disproportionately carried out on immigrants and African and Mexican-Americans. Anxiety about female sexuality and gender norms further characterized the procedure; many of the sterilized female patients in their late teens and early 20s had been institutionalized because they were 'promiscuous' or 'nymphomaniacal' for having a child while unmarried (Stern, 2005a, pp.1130–1131).

Men were sterilized, too, but the explanations for sterilizations were gender specific, again indicating how important gender is to understanding eugenics. Wendy Kline argued that, at California's Sonoma State Home for the Feebleminded, from 1918 to 1944 between 56 and 62% of women were sterilized, which was a slightly higher rate than men (she does not give the exact percentage rate for men). Moreover, the reasons given for the sterilizations differed by gender: The women were sterilized to protect society from their allegedly threatening sexual activity, while the men were sterilized for the benefit of their own health. One study, from 1926, estimated that 25% of the young women sterilized had been sent to the Sonoma State Home expressly for the procedure, often because of their expressions of sexuality; they were released after only a few months (2001, pp. 51–54).

Sterilizations in California dropped in the 1950s, as legal changes tightened administrative requirements for physicians and protection for patients increased; nonetheless, well into the 1970s, the procedure continued to be performed at state institutions. Using federal support for family planning and birth control, and also due to a new public awareness of family planning issues, in the late 1960s, California turned its attention to the reproduction of working-class Mexican-American women. Stern argued that in a context in which family planning models endorsed two or three children, increasing resources for birth control, and a lack of standardized consent for sterilization, situations arose in which women had little protection against coercion (2005a, pp. 1132–1133). At Los Angeles County General Hospital, obstetric residents were expected to meet quotas for tubal ligations and, in a two-year period from July 1968, 'elective' tubal ligations rose by 470%. In *Madrigal v. Quilligan*, 10 sterilized women, all Spanish speakers who immigrated to the United States in their teens, worked for low wages, and had been sterilized after difficult births ending in a caesarean, took out a class action lawsuit against the hospital. The lead defendant, obstetrician Edward James Quilligan, had been heard to express eugenic views about Mexican women burdening society by having too many babies. Although the plaintiffs lost, the hospital subsequently improved its compliance with federal guidelines regarding consent in sterilizations (Stern, 2005a, pp. 1134–1135). Stern concluded that in the 70 years its sterilization law was in force, more than 20 000 nonconsensual sterilizations in state-run medical institutions had been performed in California (2005a, p. 1128).

Yet as Roberts argued, the practice of nonconsensual sterilization was not limited to California or to procedures undertaken through eugenics laws: Under the auspices of expanding health care provision, African-American women in the US South underwent nonconsensual hospital sterilizations; one activist estimated that 60% of the African-American women of Sunflower County, Mississippi, had been sterilized postpartum without consent. When the family of two teenagers sterilized without consent in 1973, fourteen-year-old Minnie Lee and twelve-year-old Mary Alice Relf, turned to the Southern Poverty Law Center for redress, the resulting class action lawsuit concluded that federally funded programs had sterilized between 100 000 and 150 000 poor women and girls annually: almost half of them were African-American. Sterilization programs aimed at the Native American population reached an estimated 25% of women; between 1973 and 1976 more than 3000 nonconsensual sterilizations were performed. Roberts concluded that, "eugenic sterilization enforced social judgements cloaked in scientific terms" (1997, pp. 70, 90, 93–95; quotation 70).

Sterilization was not the only way in which the 'unfit' could be prevented from 'harming' the population. Martin S. Pernick has found widespread public support for, and interest in, eugenic euthanasia practiced on newborns, such as a girl born in 1917 who was microcephalic and did not have a complete skull. Her case responded to surgeon Harry Haiselden's publicity campaign about his decisions to deny treatment to 'defective' newborns in the first decades of the twentieth century. To make his case, Haiselden lectured publicly, wrote articles for the *Chicago American*, brought reporters to see the infants and interview their mothers, and collaborated on a fictionalized film based on one case of infant euthanasia called *The Black Stork*. His public stance on the issue brought his actions and those of the parents under legal scrutiny, but despite there being grounds of neglect on which to prosecute, neither Haiselden nor any of the parents were ever charged. By the 1920s, public support for euthanasia had dimmed and supporters of eugenics distanced themselves from infanticide, but the connection reemerged in the 1930s, as some US eugenicists supported German eugenic practices including euthanasia (1996, pp. 4–5, 8, 31, 160–167).

Conclusion

In this brief overview of eugenics in the Americas, despite evident diversity, there are common contextual features and manifestations of eugenics. In all cases, American countries were dealing with postcolonial situations, including ethnically diverse populations that included some combination of African and indigenous peoples. Depending on national demographics, responses to this diverse population differed, but always emerged from a concern, especially in the early twentieth century, that these populations were an impediment to progress and needed to be altered, improved, or eliminated. Immigration was a key issue, as well, both as a eugenic salvation, assuming that the correct people would settle, and potentially as another threat to national health. This overview also made clear how eugenics was a profoundly gendered project, in which women's maternal role and sexuality were polemical areas of state and medical intervention. Through eugenic discourses, the state and medical professionals took it upon themselves to determine what normal sexuality was for women, which women would be encouraged to reproduce or be sterilized, and how to help the right women reproduce a strong nation. Yet although race and gender were important components to eugenics, eugenics singled out 'dysgenic' populations in other ways. The 'feeble-minded,' the alcoholic, the syphilitic, and the poor also were subjected to eugenic interventions. Thus, study of eugenics in the Americas offers evidence for the durability and adaptability of a concept that influenced social and health policy and practice for much of the twentieth century. In Latin America, eugenics appealed in part because it offered a scientific method to foster a 'biologically consolidated nation.' Latin American elites were fearful that their national potential was undermined by lack of unity in culture, purpose, and population, after four centuries of racial mixing (Stepan, 1999, p. 106). Yet eugenics also grafted onto older notions about racial hierarchy, couched as scientific racism in the nineteenth century, appropriate gender behavior, the politics of power, and national identity throughout the Americas. The history of eugenics in the Americas reminds us of the ways in which science can and has served the interests of perpetuating power and powerlessness. The pervasiveness and adaptability of eugenics also serves as a reminder that the history of eugenics remains highly relevant to the present.

See also: Catholicism; Christianity in Latin America; Darwinism; Eugenics as a Basis of Population Policy; Eugenics as an

International Movement; Eugenics in Asia; Eugenics in Europe; Eugenics, History of; Familial Studies: Genetic Inferences; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Screening for Disease; Race: Genetic Aspects; Science and Social Movements; Scientific Controversies.

Bibliography

Briggs, L., 2002. Reproducing Empire: Race, Sex, Science and U.S. Imperialism in Puerto Rico. University of California Press, Berkeley, Los Angeles, and London.

Chomsky, A., 2000. "Barbados or Canada?" Race, immigration, and nation in early-twentieth-century Cuba. Hispanic American Historical Review 80, 415–462.

Dowbiggin, I.R., 1997. Keeping America Sane: Psychiatry and Eugenics in the United States and Canada, 1880–1940. Cornell University Press, Ithaca and London.

Eraso, Y., 2007. Biotypology, endocrinology, and sterilization: the practice of eugenics in the treatment of Argentinian women during the 1930s. Bulletin of the History of Medicine 81, 793–822.

García González, A., Alvarez Peláez, R., 1999. En busca de la raza perfecta: Eugenesia e higiene en Cuba (1898–1958). Consejo Superior de Investigaciones Científicas, Madrid.

García González, A., Alvarez Peláez, R., 2007. Las trampas de poder: Sanidad, eugenesia y migración. Cuba y Estados Unidos (1900–1940). Consejo Superior de Investigaciones Científicas, Madrid.

Guy, D.J., 1998. The Pan American child Congresses, 1916 to 1942: Pan Americanism, child reform, and the welfare state in Latin America. Journal of Family History 23, 272–291.

Hochman, G., Lima, N.T., Maio, M.C., 2010. The path of eugenics in Brazil: dilemmas of miscegenation. In: Bashford, A., Levine, P. (Eds.), The Oxford Handbook of the History of Eugenics. Oxford University Press, New York and Oxford.

Klausen, S., Bashford, A., 2010. Fertility control: eugenics, neo-Malthusianism, and feminism. In: Bashford, A., Levine, P. (Eds.), The Oxford Handbook of the History of Eugenics. Oxford University Press, New York and Oxford.

Kline, W., 2001. Building a Better Race: Gender, Sexuality, and Eugenics from the Turn of the Century to the Baby Boom. University of California Press, Berkeley, Los Angeles and London.

Levine, P., Bashford, A., 2010. Introduction: eugenics and the modern world. In: Bashford, A., Levine, P. (Eds.), The Oxford Handbook of the History of Eugenics. Oxford University Press, New York and Oxford.

McLaren, A., 1990. Our Own Master Race: Eugenics in Canada, 1885–1945. Oxford University Press, Toronto, Oxford and New York.

Orbegoso, A., 2012. Eugenesia, tests mentales y degeneración racial en el Perú. Revista de psicología 14, 230–243.

Ordover, N., 2003. American Eugenics: Race, Queer Anatomy, and the Science of Nationalism. University of Minnesota, Minneapolis.

Pernick, M.S., 1996. The Black Stork: Eugenics and the Death of "Defective" Babies in American Medicine and Motion Pictures since 1915. Oxford University Press, New York and Oxford.

Ramírez de Arellano, A.B., Seipp, C., 1983. Colonialism, Catholicism, Contraception: A History of Birth Control in Puerto Rico. University of North Carolina Press, Chapel Hill, NC.

Roberts, D., 1997. Killing the Black Body: Race, Reproduction, and the Meaning of Liberty. Vintage Books, New York.

Schell, P.A., 2010. Eugenics policy and practice in Cuba, Puerto Rico, and Mexico. In: Bashford, A., Levine, P. (Eds.), The Oxford Handbook of the History of Eugenics. Oxford University Press, New York and Oxford, pp. 477–492.

Strange, C., Stephen, J.A., 2010. Eugenics in Canada: a checkered history, 1850s–1990s. In: Bashford, A., Levine, P. (Eds.), The Oxford Handbook of the History of Eugenics. Oxford University Press, New York and Oxford, pp. 523–538.

Stepan, N.L., 1990. Eugenics in Brazil, 1917–1940. In: Adams, M.B. (Ed.), The Wellborn Science: Eugenics in Germany, France, Brazil, and Russia. Oxford University Press, New York and Oxford, pp. 110–152.

Stepan, N.L., 1991. "The Hour of Eugenics:" Race, Gender, and Nation in Latin America. Cornell University Press, Ithaca and London.

Stern, A.M., 1999. Responsible mothers and normal children: eugenics, nationalism, and welfare in post-revolutionary Mexico, 1920–1940. Journal of Historical Sociology 12, 369–397.

Stern, A.M., 2000. Mestizofilia, biotipología y eugenesia en el México pos-revolucionario: Hacia una historia de la ciencia y el estado, 1920–1960. Relaciones 81, 59–91.

Stern, A.M., 2005a. Sterilized in the name of public health: race, immigration, and reproductive control in modern California. American Journal of Public Health 95, 1128–1138.

Stern, A.M., 2005b. Eugenic Nation: Faults and Frontiers of Better Breeding in Modern America. University of California Press, Berkeley, Los Angeles, and London.

Stern, A.M., 2010. Gender and sexuality: a global tour and compass. In: Bashford, A., Levine, P. (Eds.), The Oxford Handbook of the History of Eugenics. Oxford University Press, New York and Oxford.

Suárez y López Guazo, L., 2000. Medicina y mejoramiento racial: la eugenesia en México. Boletín mexicano de historia y filosofía de la medicina 3, 4–15.

Suárez y López Guazo, L., 2005. Eugenesia y racismo en México. Universidad Nacional Autónoma de México, México, DF.

Subercaseaux, B., 2007. Raza y nación. El caso de Chile. A contra corriente: Una revista de historia social y literatura de América Latina 5, 29–63.

Thomson, M., 2010. Disability, psychiatry, and eugenics. In: Bashford, A., Levine, P. (Eds.), The Oxford Handbook of the History of Eugenics. Oxford University Press, New York and Oxford.

Trent, J.W., 1994. Inventing the Feeble Mind: A History of Mental Retardation in the United States. University of California Press, Berkeley, Los Angeles, and London.

Eugenics, History of

Diane B Paul, Harvard University, Cambridge, MA, USA
Marius Turda, Oxford Brookes University, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D. Paul, volume 7, pp. 4896–4901, © 2001, Elsevier Ltd.

Abstract

This article provides an overview of the history of eugenics, from the Greek philosophers to current debates on genetic screening. The historical sources of eugenics are explored next to its modern scientific foundations such as Darwinism and Mendelism. Moreover, various national contexts are highlighted in an attempt to illustrate how diverse eugenics was, both geographically and thematically.

Eugenics is the theory and practice of improving human heredity, typically through control of reproduction. Although the concept is very old, stretching back at least to the ancient Greeks, the modern movement dates to the nineteenth century when Francis Galton, a British polymath and first cousin of Charles Darwin, proposed the term 'eugenics' from the Greek *eugenes*, for 'well-born.' He defined it broadly as "the science of improving stock, which is by no means confined to questions of judicious mating, but which, especially in the case of man, takes cognisance of all influences that tend in however remote degree to give the more suitable races or strains of blood a better chance of prevailing speedily over the less suitable than they otherwise would have had" (Galton, 1883: p. 17). Galton also distinguished between 'positive' eugenics, which promotes breeding by those possessed of superior hereditary qualities, and 'negative' eugenics, which deters breeding by those with physical, mental, or moral defects.

Eugenics before Galton

Abandonment or outright killing of deformed infants was practiced in the ancient world, although both the rationale for and the extent of infanticide are matters of controversy (Roper, 1913; Riddle, 1992). Infanticide of defective infants was common in Sparta and was approved by both Plato and Aristotle; indeed, the latter thought it should be mandated by law. In the *Republic*, Plato describes a utopian state in which the family is replaced by a community of wives and children, and in which reproduction was regulated by the state and its trusted guardians (Cornford, 1941). Plato's eugenic thinking, which was further elaborated in the *Laws*, further influenced the seventeenth-century Dominican monk, Tomasso Campanella. His *City of the Sun* (1623) describes a utopia where children are bred for the good of the commonwealth, and whose inhabitants are scornful of those who neglect the breeding of their own kind. Galton himself wrote a utopian eugenic novel under the title *The Eugenic College of Kantsaywhere* (now digitized and available at http://www.ucl.ac.uk/library/special-coll/ksw.shtml). In addition to philosophy and utopianism, examples of eugenic thinking can also be found in medical literature, including the poem entitled *Callipaedia* written in 1655 by Claude Quillet and the *Essai sur la manière de perfectionner l'espèce humaine (Essay on How to Perfect the Human Species)* published in 1756 by Charles Auguste Vandermonde, and in the writings of nineteenth-century phrenologists and physiognomists, who were fascinated with the hereditary character of disease and mental traits and the implications for reproduction.

From Galton to the 1930s

Charles Darwin and Evolutionary Thought

The publication of Darwin's *Origin of Species* (1859) was central to the emergence of modern eugenics. Darwin argued that evolution occurred through a process of fierce competition in which organisms with advantageous traits would tend to survive and reproduce their kind, while the ill-adapted would leave few if any progeny. He and his contemporaries viewed this process as a progressive force that tended toward the constant physical and mental improvement of organic beings. But many Victorians worried that natural selection was no longer at work since modern medicine, charity, and other humanitarian measures prevented the culling of the physically and mentally weak. Moreover, it seemed as though the worst elements of society were breeding at a much faster rate than the best. Darwin expressed a common middle-class anxiety about the 'differential birthrate' when he remarked in *The Descent of Man* (1871) that if various checks "do not prevent the reckless, the vicious, and the otherwise inferior members of society from increasing at a quicker rate than the better class of men, the nation will retrograde, as has occurred too often in the history of the world. We must remember that progress is no invariable rule." Darwin himself did not advocate any practical measures to counter the alleged degeneration of modern society. But in 1869, his cousin Galton published the first modern eugenics manifesto, *Hereditary Genius*, prompting Darwin to exclaim: "I do not think I ever in all my life read anything more interesting and original" (Galton, 1908: p. 290).

Galton's Argument

Like many Victorian scientists, Galton was interested in a wide range of subjects, from geography to phrenology, and he was also an avid inventor. But his central passion was eugenics. Believing that heredity and not environmental influences could

explain the racial, cultural, and social differences between populations, Galton applied quantitative measurements and analysis of variants in the physical dimensions of the human body to explain social class differentials in British society. Galton aimed to prove that human mental, moral, and temperamental traits, and not just physical ones, were transmitted from parents to offspring. These traits included a gregarious personality; a truthful or courageous character; a liking for alcohol, drugs, or gambling; and a tendency to pauperism, crimes of violence, or fraud. He was particularly concerned to prove the heritability of traits making for social success.

His method was simple. Consulting biographical dictionaries, he established that men who had achieved eminence in various scientific fields were much more likely than others to have had fathers who were themselves eminent; in short, that achievement ran in families. Of course Galton knew that individuals also inherit social advantages, and he conceded that this might explain success in certain fields. No admirer of the landed aristocracy, which still dominated parliament, the army, and civil service, Galton noted that "it is notorious that neither the Legislature nor the army afford, in their highest ranks, an open arena to the ablest intellects" (Galton, 1865: p. 161). But in the 'open fields' of science, literature, and the law, individuals without talent would surely fail, no matter how distinguished their families, while conversely, the talented would succeed, no matter how unfavorable their background.

Unfortunately, it seemed that those with the most talent were the least reproductively prolific. Galton argued that there was an urgent need to reverse this trend. In his view, this could be achieved by convincing those of 'good' stock to marry, preferably with each other, and to have many children. His ideas for inducing the fittest to reproduce included celebrating their marriages in Westminster Abbey, providing state 'dowries' to encourage early marriage, subsidized housing settlements where gifted couples could raise large families, as well as payments to women graduating from college who married before age 26 and on the birth of each child.

Changing Views of Heredity

Galton believed that the hereditary material was transmitted intact from one generation to another. However, his contemporaries – including Darwin – were 'Lamarckians,' who assumed that characteristics acquired by organisms during their lifetime (primarily through the use and disuse of particular organs) could be transmitted to their offspring. The Lamarckian or 'soft' view of heredity blurs the distinction between nature and nurture, since it assumes that inherited traits can be suppressed or redirected by changing the conditions of life. Thus, even if social problems such as criminality, promiscuity, or pauperism resulted from poor heredity, the solution could reside in improved nutrition, education, or other social reforms. But if heredity is impervious to environmental changes, the only way to eliminate the problems it caused would be through selective breeding.

Developments in the burgeoning field of genetics after 1900 were to the development of eugenics what the seminal theory of natural selection was to the development of biology in the nineteenth century. The germ-plasm theory proposed by the German cytologist August Weismann (1834–1914), which rejected the inheritance of acquired characteristics, and the hereditary model proposed by the Moravian monk Gregor Mendel (1822–1884), later known as the system of Mendelian inheritance, changed the course of eugenics. In the new view of heredity, no matter how hard individual parents worked to improve their minds or bodies, or whatever social reforms might be instituted, no child would be born healthier, happier, or more virtuous or intelligent as a result.

Eugenics Becomes Popular

Galton's work initially received a lukewarm reception, especially in the popular press. But after the turn of the century, eugenics caught on. Between 1890 and 1930, eugenics movements of some form developed in over 30 countries, encompassing Central and South as well as North America, Southeast Asia as well as Europe (Adams, 1990; Turda and Weindling, 2007; Bashford and Levine, 2010). By the time the First International Eugenics Congress convened in London in 1912, eugenic ideals had been embraced by more than 400 participants from most European countries and the United States. In Germany, a variant of eugenics called racial hygiene emerged at the end of the nineteenth century, particularly in the writings of Alfred Ploetz. Most German eugenicists, following Ploetz, shifted the focus of eugenics from a Galtonian preoccupation with the individual's inherited qualities to those of the broader national community. Racial hygiene, according to Ploetz, was exclusively concerned with the hereditary qualities of the population, and its aims were twofold: (1) to both increase and further those hereditarily 'superior' individuals and (2) to decrease – if elimination was not possible – those considered racially undesirable (Turda, 2010).

Eugenic societies drew most of their members from the professional middle classes: doctors, anthropologists, lawyers, teachers, social workers, ministers, and so on. To some extent eugenics reflected middle-class anxieties about social and political position. Numerous demographic studies seemed to demonstrate that workers and agricultural laborers had more children than the middle and upper-middle classes, with those living in the worst conditions producing the largest families (actual patterns of fertility are a matter of dispute; see Szreter, 1996). Another concern was the ostensibly rapid increase in the number of 'feebleminded.' In 1910, the American psychologist Henry H. Goddard proposed a three-tiered grading system for mental deficiency. Those with a mental age of one or two he termed 'idiots,' those with a mental age of three to seven 'imbeciles,' and those with a mental age between nine and twelve 'morons' (Zenderland, 1999). Members of the 'high-grade feebleminded' or 'moron' class posed the greatest threat to society because they could not easily be differentiated from normal persons. Unlike idiots and imbeciles, they were often physically fit and attractive, and thus had no trouble either finding mates or producing children. But lacking foresight and self-restraint, and unable to distinguish right from wrong, they easily fell into lives of pauperism, alcoholism, sexual promiscuity, and crime. Exhibits at museums, expositions, and state fairs described the disastrous consequences of allowing reproduction by these defectives to go unchecked.

On the basis of pedigree studies demonstrating the passage of traits in several generations of the same family (such as Goddard's study of the Kallikak family), feeblemindedness was judged to be heritable. According to the best science of the day, feeblemindedness resulted from a single recessive gene. It thus initially seemed that the trait could be eliminated if its possessors were prevented from breeding. By the 1920s, that optimistic assessment was shown to be wrong, since most of the offending genes would be hidden in apparently normal carriers. Policies such as eugenic segregation or sterilization, which targeted only those affected, would thus work much more slowly than previously assumed. But most eugenicists reasoned that even a small improvement of the population was better than none (Paul and Spencer, 1995).

Modern scholarship has challenged the once conventional identification of eugenics with political and social conservatism. As diverse ideologically as it was dispersed geographically, eugenics was embraced by a broad array of professional and political elites. In China, eugenics flourished during the Republican Era, where it was invoked by modernizers in their efforts to counter tradition (Dikötter, 1998). In post-1945 Greece, eugenics was invoked as modern technology of health and social planning. Fabian and other socialist eugenicists turned the concern that the worst were outbreeding the best into a case for social and economic change, arguing that only in a society that abolished economic inequality would it be possible to distinguish genetic value from environmental good fortune (see Kevles, 1995; Paul, 1998). These Left eugenicists criticized the racial and class bias that informed most eugenics movements, arguing that the real problem was the high reproductive rates among the least prudent and foresighted of every class.

Although many eugenicists were socially conservative, sex reformers invoked eugenics in support of access to birth control, legalization of divorce and abortion, and radical changes in gender roles, including expansion of educational and employment opportunities for women (who were often drawn to the eugenics cause; for example, nearly half the members of the British Eugenics Education Society were female). Birth-control advocates argued that access to contraception would allow poor women, of inferior heredity, to voluntarily limit their births. According to the manifesto of the American Birth Control League, authored by Margaret Sanger, "funds that should be used to raise the standard of our civilization are diverted to maintenance of those who should never have been born" (Franks, 2005; Engelman, 2011). (Opponents countered that contraception would be used disproportionately by the middle class.) Even 'free love' or abolition of marriage was promoted on the grounds that marriage was often based on class or other genetically irrelevant factors (Paul, 1995; Hansen, 1996).

Eugenic Measures

Until recently, studies of eugenics focused on Britain and the United States. Since the 1980s, however, the focus has greatly broadened to include Canada, Latin America, Germany, Russia, France, Spain, Scandinavia, Australia, China, and Japan and a number of Central and Southeastern European countries. Eugenics took quite different forms in different places. In some Latin American countries as well as in Italy, Spain, and France, where Lamarckianism was rather strong and Catholicism the dominant religion, concerns with biological degeneration did not result in the introduction of sterilization laws. In France, for instance, the puèriculture (prenatal- and child-care), movement, spearheaded by Adolphe Pinard, stressed the value of breastfeeding and of improved diet and more rest for pregnant women, as well as a premarital examination law to prevent dysgenic unions among those individuals with hereditary defects. In Italy, Nicola Pende proposed the science of *biotypology* as a form of eugenic improvement based on identifying the constitutional types within the nation. Instead of sterilization and segregation, biotypologists promoted medical research (particularly endocrinology), maternal treatment, child welfare, and educational campaigns. In these countries, environmental and hereditary improvement were generally thought to be mutually reinforcing, rather than exclusionary (on France, see Schneider, 1990; on Latin America, see Stepan, 1991; on Italy, see Cassata, 2011; on Argentina, see Gomez Di Vincenzo, 2013).

Furthermore, in Central and Southeastern Europe, eugenicists valued the peasantry as an embodiment of racial fertility and national strength. They praised the rural communities while denigrating the city and urban spaces as degenerate and damaging to the racial health of the nation. Moreover, in some national cases, such as Romania, the size of their ethnic minority populations prompted the state to promote eugenic policies on behalf of the ethnic majority (Turda and Weindling, 2007; Promitzer et al., 2011). Even in North America and the Protestant countries of Europe, eugenics encompassed a wide range of activities. In Britain, moreover, where the modern movement originated, there was no effort to pass coercive legislation (which would have been politically impractical, in part due to the strength of the Labor Party, which viewed eugenics as hostile to workers). Following Galton's lead, the Eugenics Education Society emphasized the need to convince their compatriots of the value of controlling reproduction. Thus, eugenics consisted primarily of education and propaganda, such as the Pauper Pedigree Project, which compiled family trees of typical degenerate families (Mazumdar, 1992).

Coercive eugenic measures were, however, adopted in many countries. Beginning in 1907, when Indiana authorized the compulsory sterilization of 'criminals, idiots, rapists, and imbeciles,' sterilization laws were passed in more than 30 American states, two Canadian provinces, all the Scandinavian countries, Estonia, Germany, and Japan. In the United States, opposition from the Catholic Church, concern that the procedure (vasectomy in men, tubal ligation in women) would promote sexual promiscuity and venereal disease, and especially legal challenges to the constitutionality of state laws, originally limited the number of sterilizations (Reilly, 1991). To address the legal issues, Harry Laughlin, Superintendent of the Eugenics Record Office at Cold Spring Harbor, New York (the center of the organized eugenics movement in the United States), drafted a model law. Its use by the state of Virginia was tested in the 1927 case of *Buck v. Bell*. In his famous decision in that case, Justice Oliver Wendell Holmes wrote: "It is better for all the world if, instead of waiting to execute degenerate offspring for their crime, or to let them

starve for their imbecility, society can prevent those who are manifestly unfit from continuing their kind…Three generations of imbeciles is enough." It is significant that Holmes was a progressive jurist, and that his decision in *Buck v. Bell* was approved by seven other members of the Supreme Court, including his distinguished colleague Louis D. Brandeis, famous for his defense of right to privacy (Lombardo, 2008).

In the aftermath of the case, a dozen states revised existing laws or enacted new ones (with more than half the 60 000 legal sterilizations performed in California). In the United States, concerns about deterioration had been heightened by the results of the mental testing program initiated by the army during World War I. Psychologists tested more than 1.75 million recruits and concluded that nearly half the white draft was feebleminded, with blacks and recent immigrants scoring worst. The testing program thus exacerbated anxiety about the quality of immigrants from Southern and Eastern Europe, whose numbers increased dramatically beginning in the 1880s. The newcomers from Russia, Poland, Italy, and the Balkans were resented for their cultural and linguistic differences, support for big-city political machines, and willingness to work for low wages. Eugenicists added that they were also of poor stock, and that mixing with older stock Americans would produce biological degeneration. In the first decade of the century, Theodore Roosevelt had warned repeatedly that old-stock Americans were losing a desperate 'war of the cradle' with lesser breeds, a theme that recurred frequently in both academic and popular literature (Lovett, 2007). In 1924, President Calvin Coolidge signed into law the Immigration Restriction Act, which restricted entrants from any European country to 2% of the foreign-born of the same national origin recorded in the 1890 census, that is, before the massive influx of new immigrants.

American developments were noted in Germany, where a draft law permitting sterilization with the consent of the person involved or his or her guardian had been drafted at the end of the Weimar Republic. In *Mein Kampf*, Adolf Hitler had characterized National Socialism as 'applied biology.' The Nazi government quickly passed a compulsory law, which extended to a wide range of ostensibly genetic conditions. The law was often praised by non-German commentators for its procedural safeguards (Kopp, 1936). Approximately 400 000 Germans were eventually sterilized under the Nazi regime (Bock, 1986: pp. 230–246). In 1939, the Nazis instituted a euthanasia program aimed at eliminating the 'useless eaters' in mental hospitals.

Neither the sterilization nor euthanasia policies were racially oriented. But soon after the Nazis came to power, they enacted the first of the Nuremberg laws, which forbade Germans to marry Jews and others of 'inferior blood.' Other actions targeting Jews, Gypsies, the offspring of German mothers and black French soldiers, and other reviled groups followed, culminating in the Holocaust. Events in Germany were watched closely in Scandinavia, where many eugenicists were critical of Nazi racism but impressed by the sterilization law. Denmark, Sweden, Norway, Finland, and Iceland passed laws of their own during the 1930s (Broberg and Roll-Hansen, 1996).

Support for sterilization reflected fears about degeneration that were shared by elites elsewhere in Europe and in America. It also reflected the Social Democratic emphasis on communal responsibilities, as well as the view that a state that has taken on the responsibility of caring for its citizens must limit the numbers of those who will burden it. With its objectifying scientific gaze, eugenics contributed to this ideal of a healthy national community, one in which those individuals and ethnic groups who were deemed dangerous to the nation were relegated to institutions and marginal social spaces, traditionally reserved for the elderly and the infirm (Turda, 2013).

Opposition to compulsory sterilization did not necessarily imply the rejection of eugenics per se, as clearly evidenced by the British or Italian cases. In fact, most critics took for granted that mentally deficient individuals should be persuaded or prevented from breeding. The issue was how best to achieve this end. Skeptics variously charged that the laws promoted promiscuity, were ineffective, were based on unfounded assumptions about the heredity of mental defect or criminality, and were inevitably biased in their application. A common concern was that sterilization would provide a cheap substitute for the more desirable policies of segregation, training, and community supervision. Even the Catholic Church – the most powerful foe of sterilization – was not opposed to all forms of eugenics (much less committed to the view that procreation is a private matter), and Catholicism's traditional approaches to contraception, marriage, and family life were sometimes viewed as compatible with eugenic teachings (Muckermann, 1933). In the 1930 encyclical *Casti Connubii*, Pope Pius XI castigated eugenicists' desire to prevent 'unworthy' life as an expression of excessive secularization and of the state's interference in the individual and family's private sphere. But while sterilization was condemned as an illegitimate usurping by the state of power over the bodies of its subjects, eugenics was "accepted, provided lawful and upright methods are employed within the proper limits."

Eugenics after World War II

Although it is often said that eugenics was discredited by its association with Nazi barbarities, the reality is far more complex. In the aftermath of World War II, eugenics was increasingly criticized, but eugenic practices and rhetoric often lingered and for some, the word 'eugenics' retained a positive valence. Most eugenics societies in the West survived the war, including the American Eugenics Society and the British Eugenics Society. In Japan, the Eugenic Protection Law was introduced in 1948 (replacing the 1940 National Eugenic Law), while in Greece the Hellenic Eugenics Society was established only in 1952. In many jurisdictions, including several US states, eugenic sterilizations continued after the war. Indeed, in most Scandinavian countries, they peaked in the late 1940s. Racially oriented sterilizations were performed on Romani women under the Communist regime in Czechoslovakia and between 1996 and 2000 on rural indigenous women in Fujimori's Peru.

In the intellectual discourse of the 1950s and 1960s, claims that eugenics had been discredited *tout court* were challenged by those who believed that the core Galtonian aim of human genetic improvement could and should be defended. In the United States, it was common, especially among scientists, to distinguish a good eugenics – one respectful of individual

liberty, disease-focused, scientifically sophisticated, and uninflected by race or class – from the bad eugenics of the past. However, this effort ultimately failed, and favorable references to eugenics became increasingly rare.

Recently, several philosophers have attempted to rehabilitate both the core ideal and the term, arguing, as had many postwar scientists, that eugenics per se is not objectionable – that racism, class-bias, social conservatism, and violations of individual liberty, while prominent aspects of the history of eugenics, were not constitutive of it (Wikler, 1998; see also Kitcher, 1997; Agar, 2004; Glover, 2006). But in most academic and public discourse, to label a practice as eugenics became *ipso facto* to condemn it. As a result, efforts to oppose a 'good' to a 'bad' eugenics have generally been replaced by efforts to demarcate what *is* from what *is not* eugenics.

See also: Darwinism; Eugenics as a Basis of Population Policy; Eugenics as an International Movement; Eugenics in Asia; Eugenics in Europe; Genetics: The New Genetics; New Genetics and Race.

Bibliography

Adams, M.B. (Ed.), 1990. The Well-Born Science: Eugenics in Germany, France, Brazil, and Russia. Oxford University Press, New York.
Agar, N., 2004. Liberal eugenics. In: Defense of Human Enhancement. Wiley-Blackwell, Hoboken, NJ.
Bashford, A., Levine, P. (Eds.), 2010. The Oxford Handbook of the History of Eugenics. Oxford University Press, New York.
Bock, G., 1986. Zwangsterilization im Nationalsozialismus. Westdeutscher, Opladen, Germany.
Broberg, G., Roll-Hansen, N. (Eds.), 1996. Eugenics and the Welfare State: Sterilization Policy in Denmark, Sweden, Norway, and Finland. Michigan State University Press, East Lansing, MI.
Cassata, F., 2011. Building the New Man: Eugenics, Racial Sciences and Genetics in Twentieth-Century Italy. CEU Press, Budapest, NY.
Cornford, F.M., 1941. The Republic of Plato. Oxford University Press, London.
Dikötter, F., 1998. Imperfect Conceptions: Medical Knowledge, Birth Defects and Eugenics in China. Columbia University Press, New York.
Engelman, P.C., 2011. A History of the Birth Control Movement in America. Praeger, Westport, CT.
Franks, A., 2005. Margaret Sanger's Eugenic Legacy: The Control of Female Fertility. McFarland, Jefferson, NC.
Galton, F., 1865. Hereditary talent and character. Macmillan's Magazine 12, 157–166, 318–327.
Galton, F., 1883. Inquiry into Human Faculty and Its Development. Macmillan, London.
Galton, F., 1908. Memories of My Life, second ed. Methuen, London.
Glover, J., 2006. Choosing Children: Genes, Disability, and Design. Oxford University Press, Oxford, UK.
Gomez Di Vincenzo, J.A., 2013. "Biotipificar al soberano." Biotipología, psicotecnia, educación y prescripción de roles sociales en Argentina (1930–1943). Rhesis, Buenos Aires.
Hansen, B.F., 1996. Something rotten in the state of Denmark: eugenics and the ascent of the welfare state. In: Broberg, G., Roll-Hansen, N. (Eds.), Eugenics and the Welfare State: Sterilization Policy in Denmark, Sweden, Norway, and Finland. Michigan State University Press, East Lansing, MI, pp. 9–76.
Kevles, D.J., 1995. In the Name of Eugenics: Genetics and the Uses of Human Heredity. Harvard University Press, Cambridge, MA.
Kitcher, P., 1997. The Lives to Come: The Genetic Revolution and Human Possibilities. Free Press, New York.
Kopp, M.E., 1936. Legal and medical aspects of eugenic sterilization in Germany. American Sociological Review 1, 761–770.
Lombardo, P., 2008. Three Generations, No Imbeciles: Eugenics, the Supreme Court, and Buck v. Bell. Johns Hopkins University Press, Baltimore, MD.
Lovett, L.L., 2007. Conceiving the Future: Pronatalism, Reproduction, and the Family in the United States, 1890–1938. University of North Carolina Press, Chapel Hill, NC.
Mazumdar, P., 1992. Eugenics, Human Genetics, and Human Failings: The Eugenics Society, Its Sources and Its Critics in Britain. Routledge, London.
Muckermann, H., 1933. Eugenik und Katholizismus. Alfred Metzner, Berlin.
Paul, D.B., 1995. Controlling Human Heredity: 1865 to the Present. Prometheus, Amherst, NY.
Paul, D.B., Spencer, H.G., 1995. The hidden science of eugenics. Nature 374, 302–304.
Paul, D.P., 1998. The Politics of Heredity: Essays on Eugenics, Biomedicine, and the Nature–Nurture Debate. SUNY Press, Albany, NY.
Promitzer, C., Trubeta, S., Turda, M. (Eds.), 2011. Health, Hygiene and Eugenics in Southeastern Europe to 1945. CEU Press, Budapest, NY.
Reilly, P.R., 1991. The Surgical Solution: A History of Involuntary Sterilization in the United States. Johns Hopkins University Press, Baltimore, MD.
Riddle, J.M., 1992. Contraception and Abortion from the Ancient World to the Renaissance. Harvard University Press, Cambridge, MA.
Roper, A.G., 1913. Ancient Eugenics. Oxford University Press, London, UK.
Schneider, W., 1990. Quality and Quantity: The Quest for Biological Regeneration in Twentieth-Century France. Cambridge University Press, Cambridge, MA.
Stepan, N.L., 1991. The Hour of Eugenics: Race, Gender, and Nation in Latin America. Cornell University Press, Ithaca, NY.
Szreter, S., 1996. Fertility, Class and Gender in Britain 1860–1940. Cambridge University Press, Cambridge, MA.
Turda, M., 2010. Modernism and Eugenics. Palgrave Macmillan, Basingstoke, UK.
Turda, M., 2013. Crafting a healthy nation: European eugenics in its European context. In: Turda, M. (Ed.), Crafting Humans: From Genesis to Genetics and Beyond. V&R Unipress, Göttingen, pp. 109–126.
Turda, M., Weindling, P. (Eds.), 2007. Blood and Homeland: Eugenics and Racial Nationalism in Central and Southeast Europe, 1900–1940. CEU Press, Budapest, NY.
Wikler, D., 1998. Eugenic values. Science in Context 11, 455–470.
Zenderland, L., 1999. Measuring Minds: Henry Herbert Goddard and the Origins of American Intelligence Testing. Cambridge University Press, New York.

European Religion from the Atlantic to the Black Sea

Grace Davie, University of Exeter, Exeter, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The first part of this article considers the relevance of the religious factor to definitions of Europe; it also outlines the major confessional divisions within the continent. The second section introduces a range of factors that must be taken into account in order to understand fully the nature of religion in twenty-first-century Europe. The final section deals with the bewildering variety of church–state relationships in this part of the world; it also contains additional material on Russia.

Introduction

The concept of Europe – whether considered geographically or culturally – is not self-evident. While the Western, coastal edge is clear enough, the distinctiveness of Europe within the Eurasian land mass is less so. The first part of this article considers the relevance of the religious factor to the definitional question; it also outlines the major confessional divisions within the continent. The second section introduces the very varied factors that must be taken into account in order to understand fully the nature of religion in twenty-first-century Europe. The final section deals with the bewildering variety of church–state relationships in this part of the world; it also contains additional material on Russia.

The Religious Boundaries of Europe

The complexities involved in the concept of Europe are addressed in Davies (1996). In the early modern period, the idea of Europe gradually disentangled itself from the concept of 'Christendom,' a process in which the religious factor played a decisive role. Given that the nations of Europe had endured centuries of religious war, they were disinclined to remind themselves of what they were supposed to have in common; the notion of Europe provided a more neutral alternative. A longer historical perspective reveals, however, that Christendom had itself been divided long before the early modern period: between an Orthodox East (centered first in Constantinople and then in Moscow) and a Catholic West (centered in Rome), a dispute that came to a head for a mixture of theological and political reasons in the eleventh century.

A central question regarding Europe's self-understanding is connected to this division: should this or should it not include Russia, an Orthodox and Slavophile state that extends to the Bering Sea? This article will include Russia, but will at the same time emphasize the distinctiveness of the Orthodox parts of the continent. Such distinctiveness takes the form of a markedly different cultural heritage. The nations associated with Western Christianity have, for example, experienced the Renaissance, the Reformation, the scientific revolution, the Enlightenment, and the Romantic revival, and take these pivotal movements for granted. Countries further east have been through an equally formative but entirely different historical evolution. For them, the ebb and flow of Islam have been of greater significance. Indeed, it is possible to argue that the concept of Europe finally gives way to something else at the point where Orthodoxy gives way to Islam, on a line that has fluctuated over centuries and continues, from time to time, to play a crucial and not always peaceful role in the establishment of identities. The 1990s disputes between Russia and Chechnya or between Armenia and Azerbaijan are obvious illustrations. The continuing controversies surrounding Turkey's role in Europe – and, more specifically, regarding Turkey's membership in the European Union – are equally pertinent.

Given its long-term cultural significance, the boundary between Catholic and Orthodox Christianity is more important for understanding Europe's religious life than the much later division between communist and noncommunist Europe. With this in mind, it is useful to introduce the terms 'Eastern' and 'Central' Europe when dealing with the former communist countries. The countries of Eastern Europe (European Russia, Romania, Bulgaria, Serbia, and Macedonia) belong to the Orthodox tradition, whereas the countries of central Europe (the Baltic States, Poland, the former East Germany, the Czech Republic, Slovakia, Hungary, Slovenia, and Croatia) developed within Western Catholicism. It is only in the postwar period that they have been separated from the economic and cultural traditions of the West. The enlargements of the European Union in 2004 and 2007 are easier to understand against this background. The 2004 enlargement was almost exactly coterminous with the Western tradition. The 2007 enlargement extended beyond this. The position of Greece is in some ways anomalous; for a large part of the postwar period, it was the only Orthodox country within the Union. In a much longer time scale, however, the contribution of Greek culture to the understanding of what it means to be European in the Western sense is hard to overestimate. Greek rationalism is a formative factor in European identity.

The border between Eastern and Central Europe is, however, as problematic as any other, for on it are found the Greek Catholic or Uniate churches – those, in other words, which have been received under the jurisdiction of Rome but retain their own ritual, practice, and canon law. In the disputed identities of this part of Europe, moreover, the Uniate churches are of considerable importance (notably in the Ukraine and Belarus). Further south and in a rather different way, Bosnia-Herzegovina has been uncomfortably squeezed between Orthodox Serbia and Catholic Croatia, a situation rendered all

 http://dx.doi.org/10.1016/B978-0-08-097086-8.84009-X

the more complex by the presence of a sizable Muslim population dating from the time of the Ottoman occupation. The 1990s conflagration in the Balkans cannot be understood without taking the religious factor into account.

The subsequent divisions of the West into areas or nations that are primarily Catholic or Protestant (or combinations of the two) are inseparable historically from the emergence of the nation state as the dominant form of European political life. The processes by which such divisions occurred are highly complex, involving economic, social, political, as well as religious issues. Ongoing debates among historians deal with which of these factors were the most important historical determinants and how they fit together. What remains indisputable, however, is the unprecedented upheaval in the ordering of Christian society in the sixteenth century, an upheaval that included the emergence of separate political entities or nation states, some of which expressed their independence from papal interference in the form of a state church.

Part of the ambiguity lies in the understanding of the term 'Reformation.' Does this imply innovation and the breaking of new ground? Or does it involve a return to and rediscovery of primitive excellence? Were those who endorsed the theological changes taking place at this time looking primarily for radical change or for conservative independence? Motives were mixed. The more conservative interpretation, however, was bound to appeal to those political rulers anxious to establish independence from external authority, but with a careful eye on stability within. Both were possible within the Lutheran concept of a 'godly prince.' Sometimes the prince had jurisdiction over a whole kingdom or kingdoms. Such was the case in the Nordic countries, where Lutheranism became embodied in their state churches. Elsewhere the process was far more local and concerned relatively small patches of land. The German case exemplifies the latter, leading to patterns that are not only extant but also influential some 400–500 years later.

The Reformation took different forms in different places. In addition to Lutheranism, parts of Europe – notably the Swiss, the Dutch, the Scots, some Germans, some Hungarians and Czechs, and a small but significant minority of French people – were attracted first by Huldrych Zwingli, but then by John Calvin, toward a more rigorous version of Protestantism. Calvinism was both more radical and more restrained: radical in the sense of a new kind of theology based on the doctrines of predestination and redemption, but restrained in terms of its stringent moral codes. The effect of this particular combination on the subsequent economic development of Europe has provided inexhaustible material for an ongoing debate among historians and sociologists alike.

Broadly speaking, Western Europe divided itself into a Catholic South (Spain, Portugal, Italy, and France, but including Austria, Belgium, and Ireland), a Protestant North (the Nordic countries and Scotland), and a range of 'mixed' countries in between (England and Wales, Northern Ireland, the Netherlands, Germany, and Switzerland). Central Europe exemplifies similar categories, though the geography is rather more complicated. Lithuania, Poland, Slovakia, Slovenia, and Croatia are firmly Catholic; Estonia and (most of) Latvia are Lutheran, and relate closely to their Nordic neighbors (a commonality strongly emphasized as the Baltic republics regained their political independence in the early 1990s); and Hungary and the Czech Republic, finally, are rather more mixed (primarily Catholic but with significant Protestant minorities). In other words, boundaries gradually emerged all over Europe dividing one nation from another, one region from another, and one kind of Christianity from another. Boundaries, moreover, imply dominance as well as difference. Majorities and minorities were, and still are, created depending on the precise location of the line in question.

One final point concludes this discussion: the growing secularity of parts of Europe, noting that the secular (just like the religious) takes different forms in different places. Most striking of all are the low levels of religiousness discovered in Estonia, the former East Germany, and the Czech Republic. It would be unwise, however, to attribute this decline solely to the influence of communism. It derives from a complex mix of long-term historical trends and more recent political events. Ecclesiological factors also play their part. It is clear that the Protestant churches of Central Europe were less able to withstand the pressures of the Soviet bloc than their Catholic neighbors – a fact very largely accounted for by the institutional power of the Catholic Church. Such a statement should not detract from the courage of individual resistors inspired by Protestant ideals.

Patterns of Religion in Modern Europe: Factors to Take into Account

This section draws from a longer article (Davie, 2006) to set out six very different factors that – in contrasting ways – contribute to a better understanding of the place of religion in modern Europe. The crucial point to grasp is that these factors push and pull in different directions. Additional material can be found in Byrnes and Katzenstein (2006), Davie (2000), Jenkins (2009), and Willaime (2004).

The six factors are (1) the role of the historic churches in shaping European culture; (2) an awareness that these churches still have a place at particular moments in the lives of modern Europeans, even though they are no longer able to discipline the beliefs and behavior of the great majority of the population; (3) an observable change in the churchgoing constituencies of the continent, which operate increasingly on a model of choice, rather than one of obligation or duty; (4) the arrival into Europe of groups of people from many different parts of the world, with very different religious aspirations from those seen in the host societies; (5) the reactions of Europe's secular elites to the increasing salience of religion in public as well as private life; and (6) a growing realization that the patterns of religious life in modern Europe should be considered an 'exceptional case' in global terms – they are not a global prototype.

Cultural Heritage

The starting point concerns the undisputed role of the historic churches in shaping European culture, bearing in mind that other factors (notably Greek rationalism and Roman organization) also contributed. One example will suffice: the Christian tradition has had an irreversible effect on the shaping of time and space in this part of the world. Both week and year follow the Christian cycle, even if the major festivals of

Christianity are beginning to lose their resonance for large sections of the population. The same is true of space. Wherever you look in Europe, there is a predominance of Christian churches, some of which retain huge symbolic value. This is not to deny that in some parts of Europe (notably the larger cities), the skyline is becoming an indicator of growing religious diversity. Europe is evolving, but the legacies of the past remain deeply embedded in both the physical and cultural environment.

Historical Legacies: The Old Model

Physical and cultural presence is one thing; a 'hands-on' role in the everyday lives of European people is quite another. Commentators of all kinds agree that, with very few exceptions, the latter is no longer a realistic aspiration for the historic churches of Europe. That does not mean, however, that these institutions have entirely lost their significance as markers of religious identity. These ambiguities will be explored in two ways: first, through the notion of 'believing without belonging'; and, second, through the concept of 'vicarious religion.'

One of the most striking features of religious life in contemporary Europe is the mismatch between different measurements of religiousness. There exists, first of all, a set of indicators that measure firm commitments to (1) institutional life and (2) credal statements of religion (in this case, Christianity). These indicators, moreover, are closely related to each other insofar as institutional commitment both reflects and confirms religious belief in its orthodox forms. The believing Christian attends church to express his or her belief and to receive affirmation that this is the right thing to do. Conversely, repeated exposure to the institution and its teaching necessarily disciplines belief.

No observer of the current religious scene disputes these facts – that is, that these dimensions of European religion are both interrelated and in serious decline. There is, however, considerable debate about the consequences of this situation. The complex relationship between belief (in a wider sense) and practice is central to this discussion, for it is clear that a manifest reduction in the 'hard' indicators of religious life has not, in the short term at least, had a similar effect on rather less rigorous dimensions of religiousness. For the time being, the latter remain relatively strong (the data are clear on this point). It is precisely this state of affairs that is captured by the phrase 'believing without belonging,' the popularity of which in both pastoral and sociological accounts of religious life in modern Europe indicating its perceived accuracy.

The connections between emerging patterns of belief and the institutional churches are, however, complex. Not only do the latter continue to exist, they quite clearly exert an influence on many aspects of individual and collective lives in Europe. The notion of vicarious religion is helpful in this context. By 'vicarious,' I mean the notion of religion performed by an active minority but on behalf of a much larger number, who implicitly at least not only understand but also approve of what the minority is doing (Davie, 2007). The first half of the definition is relatively straightforward and reflects the everyday meaning of the term – that is, to do something on behalf of someone else (hence the word 'vicar'). The second half is more controversial and is best explored by means of examples. Religion can operate vicariously in a wide variety of ways. Churches and church leaders perform ritual on behalf of others, notably the occasional offices – and if these services are denied or inappropriately handled, this causes offense and criticism. At the same time, church leaders and churchgoers embody moral codes on behalf of others, even when those codes have been abandoned by large sections of the populations that they serve. Churches, finally, can offer space for the vicarious debate of unresolved issues in modern societies. They are, for example, expected to contribute to – though not dominate – the moral issues of our time, be they bioethical, socioeconomic, or geopolitical.

From Obligation to Consumption: An Emerging Market

That said, Europe's diminishing but still significant churchgoers – those who sustain the tradition on behalf of the people described in the previous section – are beginning to behave differently. Here an observable change is clearly taking place, best summarized as a shift from a culture of obligation or duty to a culture of consumption or choice. What was once simply imposed or inherited becomes instead a matter of personal choice. People go to a church (or another religious organization) because they want to, maybe for a short period or maybe for longer, but there is no obligation on the part of Europeans either to attend a place of worship in the first place or to maintain a connection that no longer meets a need.

As such, this pattern is entirely compatible with vicariousness: the churches need to be there in order that people may attend them if they so choose. The nature of the relationship, however, gradually alters – a shift that is discernible in both practice and belief, not to mention the connections between them. There is, for example, an easily documentable change in the patterns of confirmation in the Church of England (confirmation is the Christian rite of passage denoting full or adult membership in the Church). The overall number of confirmations has dropped dramatically in recent decades, evidence once again of institutional decline. In England, though not yet in the Nordic countries, confirmation is no longer a teenage rite of passage, but a relatively rare event undertaken as a matter of personal choice by people of all ages. As a result, there is a very marked rise in the proportion of adult confirmations among the candidates overall – by no means enough, however, to offset the fall among teenagers. Baptism is following a similar trend. No longer the norm in England, it is becoming increasingly a matter of choice on the part of churchgoing parents on behalf of their children. It is, therefore, more likely to signify membership of a voluntary community than allegiance to a national identity. For all these reasons, the rite itself is changing in nature. It is important to remember, however, that patterns of religion in Europe vary markedly from place to place. In a number of countries (notably the Nordic countries, Poland, and parts of Southern Europe), a very high proportion of children will still be baptized as infants into the majority (national) church.

The situation can be summarized as follows: a new model of religiousness is gradually, if unevenly, beginning to assert itself across the continent. And if the form of religion captured by the term 'vicarious religion' is in effect a public utility, the new model is more like a market – in which there are gainers

and losers (see Davie, 2006). Both models, however, will continue to exist for the foreseeable future: each fills the gaps left by the other. Broadly speaking, the public utility caters for the passively religious among Europeans (the majority), and the market for the more active minority. In this sense, the market lies over the public utility rather than replacing it.

New Arrivals

The fourth factor concerns the growing number of new arrivals in almost all European societies. There have been two stages in this process. The first was closely linked to the urgent need for labor in the expanding economies of postwar Europe – notably in Britain, France, West Germany, and the Netherlands. The second wave of immigration occurred in the 1990s and included, in addition to the places listed above, the Nordic countries, Ireland, and the countries of Mediterranean Europe (Greece, Italy, Spain, and Portugal) – bearing in mind that many of these places, until relatively recently, have been countries of emigration rather than immigration. Unsurprisingly, the situation changed again following the 2008 financial crash.

Both waves of immigration were economically motivated; the implications for the religious life of Europe were nonetheless considerable – bearing in mind that they vary from place to place, depending on both host society and new arrivals. Britain and France offer an instructive comparison. In Britain, immigration has been much more varied than in France, both in terms of provenance and in terms of faith communities. It is important to remember that a significant proportion of British immigrants come from the West Indies and West Africa and is Christian in origin; these people are, moreover, far more formed in their faith than most British people. Britain is also a country where ethnicity and religion crisscross each other in a bewildering variety of ways (only Sikhs and Jews claim ethno-religious identities). In France, conversely, immigration has been largely from the Maghreb; as a result, France has one of the largest Muslim communities in Europe (circa 5 million) – an almost entirely Arab population. Rightly or wrongly, 'Arab' and 'Muslim' have become interchangeable terms in popular parlance in France.

Beneath these differences, however, lies a common factor: the growing presence of other faith communities in general and of the Muslim population in particular is challenging some deeply held European assumptions (Nielsen et al., 2009). The notion that faith is a private matter and should, therefore, be proscribed from public life – notably from the state and from the education system – is widespread in Europe. Conversely, many of those who are currently arriving in this part of the world have markedly different convictions, and offer – simply by their presence – a challenge to the European way of doing things. As a result, European societies have been obliged to reopen debates regarding the place of religion in public as well as private life: hence the heated controversies about the wearing of the veil in the school system, about the rights or wrongs of publishing material that one faith community in particular finds offensive, and about the location of 'non-European' religious buildings. There have been moments when a lack of mutual comprehension, together with an unwillingness to compromise on many of these issues, have led alarmingly fast to dangerous confrontations, both in Europe and beyond.

Secular Reactions

Such episodes raise a further point: the extent to which the secular elites of Europe use these events in order to articulate alternatives – ideological, constitutional, and institutional – to religion. It is important to remember, however, that such elites, just like their religious counterparts, vary noticeably from place to place. Once again, the contrast between Britain and France is helpful, noting that the process of secularization has taken place differently in each case (Martin, 1978).

What in Britain, and indeed in most of Northern Europe, occurred gradually (starting with a declericalization of the churches from within at the time of the Reformation) became in France a delayed and much more ideological clash between a hegemonic, heavily clerical church and a much more militant secular state. This became known as 'la guerre des deux Frances' (the war of two Frances), which dominated French political life well into the twentieth century. Its legacies still remain in the form of a self-consciously secular elite and a lingering suspicion concerning religion of all kinds – the more so when it threatens the public sphere. The fact that these threats are no longer Catholic but Muslim does not alter the underlying reaction. The quintessentially French concept of *laïcité* is deployed both to describe and to justify this approach (*laïcité* is a French term that implies the absence of religion from the public sphere). In Britain, something rather different has happened: overlapping elites (both religious and secular) demonstrate a greater propensity to work together in order to encourage mutual respect between different world faiths. Their goals, moreover, are distinctive: they are more likely than their French counterparts to look for a solution to a particular problem, than to establish a principle (such as *laïcité*) and apply this strictly to all cases.

That said, Britain now plays host to some aggressively secular voices. Resurgent religion brings with it resurgent atheism, a fact that is increasingly noticeable. The group that has become known as the New Atheists may not be all that numerous, but they have been active in public debate. The New Atheists claim that religion should not simply be tolerated but should be countered, criticized, and exposed by rational argument. Leading members include Richard Dawkins, Daniel Dennett, Sam Harris, and the late Christopher Hitchens. It is important to remember, however, that New Atheists stand out even among secularists. The reason is straightforward: unbelief is as diverse as belief and can be placed along a continuum ranging from the mildly agnostic to the ideologically committed. Unbelief, moreover, is unevenly spread across the continent. It is noticeably more common in Northern Europe than the South, and further East it is colored by the legacies of communism. The study of the secular (both in Europe and elsewhere) is increasingly emerging as a subject in its own right. Central to this development is a growing awareness that the religious and the secular are intimately linked; they constitute two sides of the same coin.

Europe as an Exceptional Case

So far, we have looked at Europe from within. Davie (2002) takes a further step, reversing the 'normal' question: instead of asking what Europe is in terms of its religious existence, it asks what Europe is not. It is not a vibrant religious market such as

that found in the United States; it is not a part of the world where Christianity is growing exponentially, very often in Pentecostal forms, as is the case in the Southern Hemisphere (Latin America, Sub-Saharan Africa, and the Pacific Rim); it is not a part of the world dominated by other faiths, but is increasingly penetrated by these; and it is not, for the most part, subject to the violence often associated with religion and religious difference in other parts of the globe – the more so if religion becomes entangled in political conflict. Hence the inevitable, if for some disturbing, conclusion: the patterns of religion in modern Europe, notably its relative secularity, might be an exceptional case in global terms.

Such a statement goes straight to the heart of an urgent and as-yet-unresolved question: is secularization intrinsic or extrinsic to the modernization process? In other words, is Europe secular because it is modern (or at least more modern than other parts of the world), or is it secular because it is European and has developed along a particular pathway unlikely to be repeated elsewhere? The latter seems the more likely case.

Church–State Relationships

A broad-brush profile of religion in Europe provides the background for a more detailed discussion of church and state relationships; these in turn constitute the parameters within which religious life on a national level takes place. Once again, it is helpful to draw on the framework established in Martin's seminal work on the development of secularization (Martin, 1978). Following Martin, Europe is a unity by virtue of having possessed one Caesar and one God (the legacy of Rome); it is a diversity because of the existence of nations. The patterns of European religion derive from the ongoing tensions and partnerships between religion and the search for national integrity and identity. These tensions and partnerships have resulted in a wide variety of church–state relationships within the continent as a whole (Madeley and Enyedi, 2003).

In Western Europe, such relationships range from the state churches of Northern Europe, through the pillarized (politico-denominationally differentiated) cultures of the Netherlands and Belgium, to the considerable diversity contained within the United Kingdom and Ireland (including a Calvinist national church in Scotland, an Anglican establishment in England, assorted nonconformity in Wales, and a technically secular but heavily Catholic state in Ireland). Germany is a biconfessional state in which both the Lutheran and Catholic churches remain important public bodies. Latin Europe includes the privileged Catholic churches of Italy, Spain, and Portugal and the rather different situation in France (the only European nation with a rigorously secular school system in addition to the state itself). In this part of Europe, and especially in France, anti-clericalism has played a historically important role.

In Eastern and Central Europe, it is necessary to start with the autocephalous churches of Orthodox Europe, all of which have been closely aligned with political power over centuries ('autocephalous' means self-governing, that is, churches whose leaders do not report to a higher-ranking authority such as the Pope). In Central Europe, there are examples (notably Poland) of the Catholic Church becoming the effective carrier of an alternative memory during the communist period; elsewhere – in Hungary and Czechoslovakia – the churches were more compromised. For reasons already explained in this article, the Lutheran churches in, for example, Estonia and the former East Germany all but collapsed altogether. Throughout the region, post-communism has brought a new set of challenges in terms of church–state relationships and the associated privileges of either party. Initial euphoria following 1989, often centered on the national church, has in many cases given way to acrimony as disputes over power and property prove difficult to resolve. The collapse of Yugoslavia is a special case; in this part of Europe, national and confessional differences reinforce each other. Both have been sharpened by recent conflict.

Church, State, and Religion in Russia

Russia has been predominantly Orthodox for more than 1000 years, an anniversary celebrated in 1988 – a date that marked a turning point in the fortunes of the Russian Orthodox Church. After seven decades of official atheism, government-sponsored events took place to mark the millennium, and a number of churches and monasteries were allowed to reopen. Several factors are important in this connection. The first is to appreciate the long-term and deep-seated connections between church and state in Russia. Russian Orthodoxy is a prime example of an autocephalous church; founded in 988, it acquired its own Patriarchate in 1589. Its influence was immense, especially when the tsar was weak. At the end of the seventeenth century, under Peter the Great, the balance of power shifted. The Church was subjugated to imperial power, but its links with the state persisted. Both were victims of the Revolution in 1917. For the next 70 years, the Church became the target of Soviet repression. The level of persecution varied over the period, but was at times intense – under communism, the Russian Orthodox Church experienced huge losses of personnel, land, and property. It is worth noting that the destruction of the Church (indeed, of religion more generally) began two generations earlier in the Soviet Union than it did in Eastern or Central Europe. Indeed, parts of the latter experienced a formative period of democracy in the interwar period.

The situation following 1989 should be seen in this light. Given the extent of its persecution under Soviet rule, Orthodox Christianity returned to Russian society astonishingly fast, symbolized in the rebuilding of the Cathedral of Christ the Savior in Moscow. Increasingly, however, the Orthodox Church has become a focal point for both nationalists and conservatives, so much so that a close alliance between the government and the Russian Orthodox Church has become a defining characteristic of Vladimir Putin's tenure. The wheel has turned full circle (Marsh, 2011). An important consequence of this situation concerns the ambiguous status of other faith communities, including the Orthodox churches that place themselves under the Ecumenical Patriarchate in Constantinople rather than Moscow. The Orthodox Church in Estonia is a case in point. Relationships with the Catholic Church remain tense, and a number of Protestant groups have encountered bureaucratic restrictions.

The Muslim population in Russia is relatively young and is expected to grow. It is largely an ethnic religion and is concentrated in the nationalities of the North Caucasus and among the Tatars and Bashkirs of the Volga basin. The situation

concerning Jews continues to evolve. Historically, the Russian Empire hosted the largest population of Jews in the world. Persecution and pogroms led to mass emigration in the late nineteenth and early twentieth centuries, the majority going to the United States. Following 1989, there has been a further exodus – large numbers have left for Israel, the United States, and Germany. The recent renaissance of Judaism in modern Germany is very largely due to this relocation.

Concluding Remarks

The Europe that is emerging as the twenty-first century develops is a rapidly changing place. From a religious point of view, one of the most significant evolutions is undoubtedly the increasing representation of faiths other than Christian. Analytical concepts must evolve accordingly. European religion is giving way to the 'religions of Europe,' a continent that now houses a significant representation of Muslims, Sikhs, Hindus, and Buddhists in addition to the Jewish communities that have played such a crucial role in Europe's recent history. Bearing this in mind, it is clear that the tensions between unity and diversity articulated by Martin more than 30 years ago remain central, though in forms that are peculiar to late modern rather than early modern European society.

It is clear, moreover, that two rather different things are happening at once. On the one hand, the process of secularization continues, though more so in some parts of Europe than in others; on the other hand, religion has returned – somewhat unexpectedly – to the public sphere. The combination is hard to handle. At precisely the moment when they are most needed, Europeans are losing the vocabulary, tools, and concepts that they require in order to have an informed conversation about faith. The result all too often is an ill-informed and ill-mannered public debate about issues of extreme importance to the democratic future of the continent.

See also: Catholicism; Orthodox Churches; Religion: Nationalism and Identity; Secularization.

Bibliography

Byrnes, T., Katzenstein, P. (Eds.), 2006. Religion in an Expanding Europe. Cambridge University Press, Cambridge.

Davie, G., 2000. Religion in Modern Europe: A Memory Mutates. Oxford University Press, Oxford.

Davie, G., 2002. Europe: The Exceptional Case. Parameters of Faith in the Modern World. Darton, Longman and Todd, London.

Davie, G., 2006. Religion in Europe in the 21st century: the factors to take into account. European Journal of Sociology 47, 271–296.

Davie, G., 2007. Vicarious religion: a methodological challenge. In: Ammerman, N. (Ed.), Everyday Religion: Observing Modern Religious Lives. Oxford University Press, New York, pp. 21–37.

Davies, N., 1996. Europe: A History. Oxford University Press, Oxford.

Jenkins, P., 2009. God's Continent: Christianity, Islam and Europe's Religious Crisis. Oxford University Press, New York.

Madeley, J., Enyedi, Z. (Eds.), 2003. Church and State in Contemporary Europe: The Chimera of Neutrality. Frank Cass, London.

Martin, D., 1978. A General Theory of Secularization. Blackwell, Oxford.

Marsh, C., 2011. Religion and the State in Russia and China: Suppression, Survival, and Revival. Continuum, London.

Nielsen, J., Akgönül, S., Alibašić, A., Maréchal, B., Moe, C. (Eds.), 2009, Yearbook of Muslims in Europe, vol. 1. Brill, Leiden This is an annual publication.

Willaime, J.-P., 2004. Europe et religions: les enjeux du XXIe siècle. Fayard, Paris.

Relevant Websites

http://www.europeanvaluesstudy.eu/ (a useful statistical source) – The European Values Study.

http://www.issp.org/, especially 1991, 1998 and 2008 (a useful statistical source) – The International Social Survey Programme.

http://www.pewforum.org/Muslim/Mapping-the-Global-Muslim-Population(14).aspx – Mapping the Global Muslim Population (Europe overview).

http://pewresearch.org/pubs/1731/muslim-networks-movements-western-europe – Muslim Networks and Movements in Western Europe.

European Union Law

Jürgen Basedow, Max-Planck-Institute for Comparative and International Private Law, Hamburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The European Union is the result of a unique attempt at international integration. While other integration movements have always relied on a continuous political affirmation of integration by the participating states, the European Union is characterized by a transfer of sovereign rights of legislation, execution, and adjudication to central institutions. At the same time, the Treaty of Rome has introduced the legal framework of a common market, providing for individual rights and obligations of citizens in the whole Union. This article sets out the basic legal structures of the Union and the Common Market, the types of legal instruments and their essential effects, and the efforts undertaken to harmonize national legal systems.

The European Union was established by the Treaty on European Union (TEU), signed by the members of the European Community (EC) at Maastricht, Netherlands, on 7 February 1992 (Official Journal of the EC – OJ – 1992 C 224/1). It initially related to three areas commonly called the pillars of the Union: the economic communities (see below), the common foreign and securities policy, and cooperation in the fields of justice and home affairs. They were consolidated by the Treaty of Lisbon of 13 December 2007 (OJ, 2007 C 306). Following the Treaty of Lisbon, the foundational law of the Union can be found in three instruments: the TEU, the Treaty on the Functioning of the European Union (TFEU), and the Charter of Fundamental Rights (consolidated versions in OJ, 2010 C 83).

Historically, the Union is rooted in three communities – which were often thought to be one community, but which were in fact three communities – with different focuses: the EC which was established as the European Economic Community by the Treaty of Rome on 25 March 1957 (United Nations Treaty Series – UNTS vol. 298, p. 11); the European Coal and Steel Community (ECSC) established by the Treaty of Paris on 18 April 1951 (UNTS vol. 261: p. 140); and the European Atomic Energy Community (EURATOM), established by a separate Treaty of Rome on 25 March 1957 (UNTS vol. 298: p. 167). These communities, which formed the first pillar of what became the European Union, were invested with supranational powers of legislation, execution, and adjudication, which they received by transfer from the member states; these powers were exercised by the organs of the communities. Alongside this first pillar, the second and third pillars of the European Union were constituted by policies and forms of intergovernmental cooperation in the fields of 'common foreign and security policy' and 'cooperation in the fields of justice and home affairs,' renamed 'Police and Judicial Cooperation in Criminal Matters' by the Treaty of Amsterdam on 2 October 1997 (OJ, 1997 C 340/1). While the three communities were set up by six founding member states (Belgium, France, Germany, Italy, Luxembourg, and the Netherlands), later accessions (Austria, Denmark, Finland, Great Britain, Greece, Ireland, Portugal, Spain, and Sweden) made the Community grow to 15 member states; a further extension of the Union toward 12 states in central and eastern Europe was put into effect in 2004 (Cyprus, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia, and Slovenia) and 2007 (Romania and Bulgaria). Croatia became the 28th member state in 2013. The success of this integration is almost entirely due to the transfer of sovereign powers to the three original communities and in particular to the EC. The remainder of this article will therefore deal only with that part of the law of the Union which can be traced back to a transfer of sovereign powers, i.e., with Community Law in the traditional sense. The terms 'Union' and 'Community' will be employed interchangeably.

A Union Based upon the Rule of Law

As the Court of Justice of the European Union has repeatedly held, the European Union/Community is a "Community based on the rule of law, in as much as neither its Member States nor its institutions can avoid a review of whether the measures adopted by them are in conformity with the basic constitutional charter, the Treaty" (Case C-2/88 Imm. – *Zwartveld* – (1990) European Court Reports (ECR) I – 3365 para. 16). In comparison with previous models for the peaceful cooperation and integration of states, the Community approach is outright revolutionary. While political considerations and diplomacy still have a great impact on the settlement of conflicts between member states, national sovereignty is no longer unlimited. States have to respect the law of the Union, lest violations are ascertained by the Court of Justice. Access to the Court is not limited to authorities of member states and the institutions of the Union; legal proceedings may also be commenced by private parties in some cases and generally by national courts of law, which request a preliminary ruling concerning the interpretation or the validity of acts of the Union (see art. 267 TFEU). The enforcement of the law of the Union is thus indirectly entrusted, beyond national governments, to the great number of national judges. While a change of government policy often threatens the performance of international obligations which states have assumed under traditional treaties, the law of the Union is much more stable due to its decentralized enforcement. An agreement among some, or even all, governments of member states not to comply with a specific legislation of the Union would fail sooner or later, unless it was embodied in an amendment of the basic Treaty. Treaty

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86030-4

amendments, in turn, have to be ratified by the parliaments of all the member states (i.e., the Union has no supranational powers here). The repercussions for national policies are apparent: member states are compelled to conduct their policies within the institutions of the Union. The pursuit of a separate policy outside that framework is not entirely excluded but will often meet with opposition from Community Law and the agencies called upon to enforce it. Within the institutions of the Union it is, however, almost impossible for a state to continuously pursue a separate policy. Thus, the rule of law established by the Treaty of Rome has narrowed the political margin of discretion left to the member states, and has channeled their policies toward integration.

Institutions

Under art. 13 TEU, the tasks of the Community are carried out by seven institutions: the European Parliament, the European Council of the Heads of State or Governments, the Council (of ministers responsible for the various branches of government), the Commission, the Court of Justice, the European Central Bank, and the Court of Auditors.

In the legislative process of the Union, only the European Parliament, the Council, and the Commission are involved. While the traditional model of Western constitutions provides for a separation of legislative and executive powers between Parliament and Government, the respective roles of the three European Institutions are more complicated.

The 'Commission' may originally have been conceived as a kind of European executive to implement the integration program of the Treaty. But this task implies wide margins of political discretion, concerning both the priorities and the content of the measures to be adopted. Moreover, the Commission is the only institution which is entitled under the Treaty to initiate legislative proceedings. It is therefore a political heavyweight in the integration process. The governments of the member states each have the right to nominate one member to the Commission, who is, however, completely independent in the performance of his or her duties (see arts. 245 s. TFEU).

It is up to the Council to represent the interests of member states within the Union. The Council consists of a representative of each member state. While each representative may veto a decision where unanimity is required, the votes of the Council members are weighted in accordance with the size of the member state, where the Treaty requires a decision taken by qualified majority (see art. 16 para. 4 TEU and art. 238 TFEU). The Treaty of Lisbon has introduced the so-called enhanced cooperation, whereby a minimum of nine member states may establish areas of advanced integration or cooperation, and this area of cooperation must be open for subsequent access by other member states; enhanced cooperation requires prior permission granted by an act of the Union (see art. 20 TEU and arts. 326 s. TFEU).

The European Parliament initially was named the Assembly and had very little weight in the legislative process, its rights being limited to mere consultation. Since then, various amendments to the Treaty have improved the European Parliament's legal status and thereby increased its political weight. At present, its members are elected by direct universal suffrage (art. 14 TEU); however, much greater relative weight is granted to the votes stemming from small member states like Luxembourg or Malta than to those originating in large member states like Germany. While the rights of the Parliament are still limited to a mere consultation in some areas such as immigration policy (cf art. 77 para. 3 TFEU), most legislative acts now have to be approved by both the Council and the Parliament in the so-called ordinary legislative procedure under art. 294 TFEU. The Parliament still lacks the power to initiate legislation by its own volition.

Legislative Powers of the Union

The Union has no inherent powers, it can only act within the limits of the powers conferred upon it by the member states in the Treaty (cf art. 5 para. 2 TEU). Nevertheless, the effect of that limitation is mitigated by three factors: first, the Court of Justice interprets the empowering provisions of the Treaty in a wide manner; second, the Treaty contains provisions like art. 352 TFEU that confer upon the Union an open-ended power; and third, by the doctrine of implied powers, the Court of Justice has extended the Union's powers even beyond the limits imposed by the Treaty to international commitments insofar as they help to protect the uniform application of the legislation of the Union vis-à-vis third states (Case 22/70 – ERTA – (1971) ECR 263 para. 28).

The exercise of the Union's competence is further limited by the principles of subsidiarity and proportionality (see art. 5 para. 3 and 4 TEU). The legal significance of the principle of subsidiarity is rather limited as compared with its sweeping use in political discussions. It does not relate to the objectives of a legislative action of the Union, which have to be established in conformity with the Treaty by the political institutions. Subsidiarity pertains only to the question of whether such objectives can be better achieved by the Community than by the member states – the so-called comparative efficiency test. It is highly questionable whether the Court of Justice is in a better position than the political institutions of the Community to decide upon that issue. This may be different in the case of the principle of proportionality, which prohibits the Community from taking action that goes beyond what is necessary to achieve the objectives of the Treaty. In the context of the basic freedoms (see Section Principles of the Internal Market), the Court of Justice has, in fact, frequently assessed the proportionality of national restrictions imposed by member states. In the future, better protection of the principles of subsidiarity and proportionality may be effected by the Protocol Two annexed to the Treaty of Lisbon, which grants certain procedural rights to the national parliaments of the member states.

Sources of Union Law

It is common to distinguish the primary and secondary or derivative law of the Union. The primary law of the Community consists of the founding treaties indicated in the introductory paragraph of this article and of some additional international conventions. Moreover, the general principles

common to the laws of the member states are sometimes classified as part of the primary law. These general principles are explicitly referred to in art. 340 para. 2 TFEU in the context of the noncontractual liability of the Community for damage caused by its institutions or by its servants; however, the Court of Justice considers the general principles as a source of law even outside that field and has recourse to them in other matters, too. However, if the distinction between primary and secondary sources of law indicates a ranking, it would appear that only certain general principles such as fundamental rights can actually count among the primary sources of Union law. Whatever the status of the general principles may be, they allow for the flexibility that every legal system needs in order to adjust to new and unforeseen facts and situations.

Art. 288 TFEU lists five types of *secondary sources* of the law of the Union: regulations, directives, decisions, recommendations, and opinions. While recommendations and opinions are not binding, most Union acts are adopted either in the form of regulations or directives. Regulations are directly applicable in all member states, which means that the Union can impose obligations on and grant rights to the citizens of the member states without the intervention of national legislatures. In contrast, a directive is binding only upon the specific member state to which it is addressed, and only to effect the result to be achieved; in order to be applicable to the citizens of the Community, it requires implementation, which can be carried out by the member states in accordance with the forms and methods of their legal system (see art. 288 para. 3 TFEU). It soon turned out that member states were sometimes slow or reluctant to implement directives. In a long line of cases, the Court of Justice has, therefore, extended the scope of directives. National laws, even if enacted prior to a directive, have to be interpreted so as to conform with it (Case 14/83 – *von Colson* – (1984) ECR, 1891 para. 26); insofar as the directive is sufficiently precise and deals with 'vertical' relations involving the defaulting member state, it may even be applied directly by the national courts and authorities (Case 8/81 – *Becker* – (1982) ECR 53 para. 24 s.). While such a direct effect is not admitted in 'horizontal' relations (Case C-91/92 – *Faccini Dori* – (1994) ECR I-3325 para. 22 seq.), the private party suffering loss from the nonimplementation of the directive may invoke the defaulting member state's government liability (see Section Ensuring Compliance with Community Law). Although the case law has eliminated some of the differences between the directive and the regulation, conspicuous differences remain. The regulation creates uniform law whose appearance is the same in all member states, while the directive provides for a harmony of solutions at best; but even after that harmonization, it is still difficult to assess the precise state of the law in other member states in the area at issue.

It is generally acknowledged, and has been affirmed on various occasions by the Court of Justice, that the law of the Union is a separate legal system distinct from both international law and the national laws of the member states. Ever since the leading case *van Gend & Loos* was decided in 1963, the Court has also stressed the need for the uniform application of Union law in all member states (case 26/62, (1963) ECR 1). Two consequences follow from this premise: Union law has to be interpreted autonomously, i.e., without reference to definitions established in national law; moreover, as recognized in *van Gend & Loos*, the law of the Union must be given direct effect and precedence over national law, insofar as its provisions are sufficiently clear and unambiguous. In the course of time the conditions for direct effect were watered down. At present, direct effect is the norm rather than the exception; it applies to many Treaty provisions, to regulations, and even to many directives (see above). By the doctrine of direct effect the Court of Justice has obliged the national courts to enforce Community law; a duty, which has been accepted over time by the national judiciaries.

Ensuring Compliance with Community Law

The institutions of both the Union and the member states must comply with the law of the Union. Acts adopted by the institutions of the Union are subject to legal review (cf arts. 263 and 265 TFEU). In the case of a violation of its laws committed by the Union or by its servants, the Union may incur liability under art. 340 para. 2 TFEU. With regard to the member states, the Treaty ensures their compliance with Union law in a much less comprehensive way. Under arts. 258 and 259 TFEU, the Commission and member states are entitled to bring a violation before the Court of Justice. In practice, only the Commission has made use of this right. The Treaty does not provide for a direct right of action of individuals who complain of a violation of Union law committed by a member state, nor does it establish the liability of member states for such violations. Both gaps have been filled by judge-made law. By the aforementioned doctrine of direct effect, the Court has enabled and instructed the national courts to disregard national provisions that violate the law of the Union. Private parties may therefore invoke the law of the Union in national courts, which can ask the Court of Justice for a preliminary ruling under art. 267 TFEU if they have doubts about the interpretation of the relevant provisions of Union law. In addition the Court of Justice has established the member state's liability for violations of Union law. Insofar as the respective member state has manifestly and gravely disregarded the legal limits of its political discretion, it has to compensate private individuals who have suffered loss as a consequence of that violation. This concerns, in particular, cases of a continuous breach of the obligation to implement a directive (cases C-46/93 and 48/93 – *Brasserie du Pêcheur v. Germany and R. v. Secretary for Transport, ex parte Factortame* – (1996) ECR I – 29; cases C-187 et al. – *Dillenkofer v. Germany* – (1996) ECR I – 4845).

Principles of the Internal Market

Among the various policies of the Union listed in art. 3 TEU, the establishment of the Common Market – later renamed the Internal Market – has always been of central significance. With regard to other policies, the Treaty either limits the activities of the Union to the 'support' of efforts made by the member states (e.g., with regard to health protection, education, industry, or culture, cf art. 6 TFEU), contains vague indications as to the direction of a Community policy (economic and social cohesion, research, transeuropean networks, consumer protection, cf art. 4 TFEU) or empowers the Community only to conduct

a policy in a given area without indicating any specific targets (agriculture, transport, environment, energy). In contrast to these policies, the establishment of a Common Market or Internal Market is not only an unequivocal objective stated in art. 3 para. 3 TEU, the basic principles of such a market are also spelled out in clear and operative rules throughout the Treaty. Thus, art. 26 para. 2 TFEU defines the Internal Market as "an area without internal frontiers in which the free movement of goods, persons, services and capital is ensured." These basic freedoms are guaranteed in further provisions of the Treaty which the Court of Justice has interpreted as producing direct effect. National provisions which subject certain markets and activities to regulations in the interest of consumers, of safety, of the environment, etc., are therefore regarded as restrictions of the basic Community freedoms and will be approved only if mandatory requirements of the public interest so command (case 120/78 – *Cassis de Dijon* – (1979) ECR 649 para. 8, 9). By elaborating in a long line of cases, both the basic rule of free movement and the exception in favor of national restrictions, the Court of Justice effectively paved the way for the deregulation of many markets in Europe in the 1980s and 1990s. While political and economic discussions on deregulation were often deadlocked, a single judgment of the Court of Justice would announce a new deal for further negotiations. The case law being inspired by a strong attitude against all kinds of discrimination on grounds of nationality (cf art. 18 TFEU), the Court has effectively opened national markets to competitors from other member states.

The basic freedoms are directed against restrictions imposed by states. Their implementation extends the scope of private liberties to engage in certain activities but they do not compel individuals and companies to actually do so. Market actors may also fear the intensification of competition and therefore try to maintain by private agreement the barriers to market access, which were previously ordered by national legislation. The risk of such anticompetitive agreements and practices had been foreseen by the founders of the Community, who therefore explicitly provided for the prohibition of anticompetitive agreements and of abuses of market dominance in arts. 101 and 102 TFEU. The enforcement of competition law, which is entrusted to the Commission and to national competition authorities, must be regarded as a complement of the basic freedoms: while the latter purport to open up markets by removing national restrictions, the former is meant to protect that competition against private restrictions.

While the basic freedoms have been implemented to a large extent, the opening up of internal markets is far from being a reality in many sectors. This is due to a great number of remaining and sometimes hidden obstacles. Thus, while nationality requirements for certain jobs have been abolished all over Europe, it is still lawful for member states to require a particular type of qualification, which, in the absence of a harmonization in the field of education, can hardly be acquired by foreigners in other member states; consequently, these jobs are still closed for foreign applicants. Another impediment to the opening up of internal markets is in the diversities relating to the legal framework of transactions, in particular in civil and commercial law; they deter many small traders from making use of their basic freedoms. While the Union is empowered to approximate the laws of the member states for the purposes of the internal market (cf art. 114 TFEU), it makes slow progress in that area.

The need for coordination of national economic policies was already considered a corollary of the common market in the early days of the European Economic Community (see art. 119 s. TFEU). It became even more conspicuous when the majority of the member states agreed upon monetary union in the Treaty of Maastricht and introduced the Euro as the common currency (art. 127 s. TFEU), thereby abandoning the role of exchange rate fluctuations as an arbitrator between the divergent economic policy orientations of member states. Tensions between the existence of the common currency administered by the European Central Bank and obvious deficits of the alignment of national economic policies have led to serious crises of the balance of payments and public debt in some member states and continue to be a major challenge to the stability of the Union.

Fundamental Rights

The EEC Treaty of 1958 did not contain a list of unalterable rights. While some fundamental rights of that type were guaranteed by single provisions scattered over the Treaty, such as prohibitions of discrimination or the basic freedoms, their scope was limited. It was therefore possible that secondary legislation could infringe private rights that were protected as fundamental rights granted by national constitutions. With respect to such situations, the German Constitutional Court indicated that it would subject the secondary legislation of the Union to the basic rights granted by the German constitution, so long as the Community had not developed a charter of basic rights of its own. A constitutional dispute between the Community and some member states appeared inevitable. In response, the European Court of Justice began to acknowledge the fundamental rights flowing from the constitutional traditions common to the member states and from the European Convention on Human Rights (UNTS vol. 213, p. 221) as an integral part of the general principles of law protected by the Court (case 11/70 – *Internationale Handelsgesellschaft* – (1970) ECR 1125 at 1134). It was only in December 2007 that the Treaty of Lisbon conferred binding effect on the Charter of Fundamental Rights of the European Union (cf art. 6 TEU). The charter can be expected to provide a basis for further judicial integration of the Union.

See also: Constitutional Courts; Globalization: Legal Aspects; International Law and Treaties.

Bibliography

Cartou, L., Gruber, A., Clergerie, J.-L., Rambaud, P., 2006. L'union Européenne, sixth ed. Dalloz, Paris.

Grabitz, E., Hilf, M. Das Recht der Europäischen Union, loose leaf ed., vol. 3. Beck, Münche.

von der Groeben, H., Schwarze, J., 2003–2004. Kommentar zum Vertrag über die Europäische Union und zur Gründung der Europäischen Gemeinschaft, sixth ed., vol. 4. Nomos, Baden-Baden.

Hartley, T., 2010. The Foundations of European Community Law, seventh ed. Oxford University Press, Oxford.

Kapteyn, P., VerLoren van Themaat, P., 1995. Inleiding tot het recht van de Europese Gemeenschappen, fifth ed. Kluwer, Deventer.

Lenaerts, K., van Nuffel, P., 2011. Europees Recht. Maklu, Antwerpen, 2011. European Union Law, third ed. Sweet & Maxwell, Thomson Reuters, London (in English).

Mégret, J., Waelbroeck, M., Dony, M. Commentaire Mégret, several vols. Bruylant, Bruxelles.

Mengozzi, P., 1997 (Trattato di diritto commerciale e di diritto pubblico dell'economia, diretto da Galgano, F.). Il diritto comunitario e dell'Unione europea, vol. 15. CEDAM, Padova.

Oppermann, T., Classen, C.D., Nettesheim, M., 2011. Europarecht, fifth ed. Beck, München.

Tesauro, G., 2010. Diritto dell'Unione europea, sixth ed. Cedam, Padova.

Weatherill, S., Beaumount, P., 1999. EU Law, third ed. Penguin Books, London.

Wyatt, D., Dashwood, A., 2011. Wyatt and Dashwood's European Union Law, sixth ed. Sweet & Maxwell, Oxford.

Europe, Archaeology of

Peter F Biehl, SUNY Buffalo, Buffalo, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 7, pp. 4912–4917, © 2001, Elsevier Ltd, with revisions made by C. Bonsall.

Abstract

European archaeology is the study of the material culture left by past inhabitants of the European continent over a time range extending back at least one and a half million years. Our predecessors were as mixed ethnically as they were culturally and made their homes in Europe's diverse microregions and microclimatic zones. They left a rich legacy of material traces and it is from these that archaeologists can reconstruct the complex patterns of their lives. The study of European archaeology had its beginnings during the Renaissance, when the classical world was rediscovered. But it was not until the end of the nineteenth century, when improved methods of classification, excavation, and relative chronology emerged, that archaeology became an academic discipline. By the twentieth century, European archaeology's major tenets had been established as had conflicting schools of archaeological thought. When the Iron Curtain fell in 1989, European archaeology experienced a new rupture. Terrain that was once impenetrable and colleagues and research that for decades had been unreachable suddenly became available. This exchange has both incited rich dialogue and stirred stimulating controversy. Spectacular recent finds such as the Iceman from the Italian Alps and technological advances such as the ability to analyze prehistoric DNA are revolutionizing archaeology revealing new insights into the past.

This article is divided into five parts, each of which outlines the major chronological and cultural shifts in European pre- and protohistory. Also included are major finds that have significantly altered or enriched knowledge of the past.

Europe's Origins – The Paleolithic and the Mesolithic

The early stages of human settlement in Europe took place against a background of severe fluctuations in climate and sea level that marked the Pleistocene 'Ice Age,' changes that probably impacted profoundly on demographic patterns and subsistence. The first European hominins may have belonged to the species known as *Homo erectus*, which originated in Africa but had dispersed into Southwest Asia 1.8 mya. Finds of simple stone tools indicate the presence of humans in southern Europe by about 1.5 mya. Fragmentary hominin fossils, thought to be over 1 million years old, have been found at the sites of Orce and Atapuerca in Spain and assigned to the species *Homo antecessor*. A new hominin species, *Homo heidelbergensis*, associated with more advanced stone tool technologies (both ultimately of African origin) was present in Europe 0.6 mya. This early period in European prehistory is known as the Lower Paleolithic. It was succeeded by the Middle Paleolithic, beginning around 300 000 years ago and marked by more advanced ('prepared core') stone flaking technologies. The emergence of prepared core technologies broadly coincided with the appearance of a new hominin species in Europe, *Homo neanderthalensis*. Unlike their predecessors, the Neanderthals buried their dead. Some burials appear to have been accompanied by grave goods, indicating that these hominins had some sort of belief system with notions of an afterlife. The Neanderthals hunted and scavenged, but they also ate fish, birds, and wild plant foods.

The Middle Paleolithic ended when the Neanderthals were rapidly replaced by the first modern humans (*Homo sapiens*) who had originated in Africa over 150 000 years ago and later dispersed into Asia and eventually Europe. *Homo sapiens*, whose physical appearance and cognitive abilities were similar to ours, were present in several parts of Europe 40 000 years ago. Their expansion into Europe marked the beginning of the Upper Paleolithic stage and was perhaps facilitated by a rapid warming of climate around 45 000 years ago. During this time, there were important advances in technology and a wider range of materials (including bone, antler, ivory, and shell) was used for manufacturing tools, weapons, and body ornaments. Cave paintings first appear in the Upper Paleolithic. The paintings mainly depict animals; however, abstract signs and patterns are also common. The sites of Altamira in northern Spain and Lascaux, Chauvet, and Cosquer in southern France offer particularly vivid examples. Remarkable are the carved female figurines depicted with exaggerated breasts, stomachs, buttocks, and thighs. They have been found all over Europe; among the best-known examples are those from Willendorf in Austria and Dolní Věstonice in the Czech Republic.

The Mesolithic marks a continuation of the hunter-gatherer way of life into the Holocene, when interglacial conditions returned to Europe once more. The time range of the Mesolithic varies from region to region; it began around 10 000 BC and came to an end with the arrival of farming, which took several millennia to spread across Europe. With the rapid warming of climate at the end of the Pleistocene, the tundra and steppe that had covered much of Europe and were home to vast herds of wild horse, reindeer, and bison were replaced by forests that supported much smaller numbers of animals like red deer and wild boar, which in turn supported fewer people. It is hardly surprising, therefore, that many Mesolithic communities relied on fishing for their subsistence, with the largest concentrations of sites found along sea coasts and the banks of large lakes and rivers. The dietary emphasis on aquatic resources and the existence of quite large cemeteries like those at Skateholm in southern Sweden, Zvejnieki on Lake Lubāns in Latvia, and sites along the Danube and Dniepr rivers in southeastern Europe, suggest that Mesolithic fishing communities often lived in permanent or semipermanent settlements. But Mesolithic people were not just fishers; they also hunted land game (with

bows and arrows, spears, and traps) and gathered wild plants for food, medicinal, and other purposes. Most groups also kept dogs. In some regions, including much of Scandinavia and the eastern Baltic, where the growing season for crops and pasture is short, a fishing-hunting-gathering economy continued long after farming was introduced to other parts of Europe.

The Emergence of Farming and Social Complexity: The Neolithic and the Copper Age

The Neolithic is often identified by the appearance of domestic animals and crops. This huge change revolutionized European history. It is accompanied by an increase in the size and complexity of settlements as well as the use of pottery and polished stone tools, and in some regions, the erection of elaborate monuments. During the Neolithic, people began to view the world in a whole new way. Farming revolutionized ways of life and subsistence. The human diet shifted from an almost total dependence on wild animals to a diet based primarily on cereals and domesticated animals. Sedentary life had demographic consequences and as populations increased, society became more complex and stratified.

Early Neolithic communities came to Europe from the Near East and Anatolia around 7000 BC, and first settled in the southeastern part of the continent in Greece and the Balkans. As in Anatolia, subsistence was based on emmer, wheat, sheep, goats, pigs, and cattle. Farmers inhabited open settlements and lived in freestanding, rectangular houses. Some settlements were inhabited for long periods and the constant building and rebuilding led to the emergence of large settlement mounds (tells). From these southeastern settlements, Neolithic communities dispersed across Europe. The movement went in two general directions: to the north and northwest and to the west along the Mediterranean coast. The process of colonization varied tremendously, as did the adoption of domesticates. In southern Scandinavia and the British Isles, for example, it occurred around 4000 BC, while in central Germany, it occurred about 1300 years earlier. The western route of Neolithic colonization led along the Mediterranean coast to southern Iberia and southern France, both of which were colonized around 5500 BC. These routes have been identified closely with the spread of Impressed Ware Pottery. From the Mediterranean, domestication and farming moved northward along the Atlantic Coast into the Parisian Basin. Here, the southern route met the northern route. This route, which followed the Danube through central Europe, is associated with Linear Pottery and showed advances in the use of longhouses, incised decorated pottery, and the shoe-last adze for clearing forests. By 5300 BC, the Neolithization of continental Europe was complete. The most significant pottery emerged from the Chasseen culture in France, the Michelsberg culture between the Rhine delta and the Alps, and the Trichterbecher culture in northern and eastern Europe. Along the Atlantic Coast, from Portugal in the south, the Orkney Isles and Scandinavia in the north, and Poland in the east, large megaliths in the form of standing stones, stone circles, henges, and cursus monuments have been found. During the fifth millennium BC, copper metallurgy was introduced into southeastern Europe, specifically in the Balkans and the Carpathian Basin. This period, which lasted from 4500 to 2500 BC, is hence known as the Copper Age.

Economic changes such as the introduction of the wheel and plow, wool production, and the increased use of animals for specialized tasks occurred. Together, this is known as the Secondary Product Revolution. At this time, more emphasis was placed on creating cemeteries than on erecting domestic structures. Cemeteries were a means of establishing group identity and instilling a sense of belonging. A spectacular example of this is in Varna on the Black Sea coast of Bulgaria. Here, individual graves featured sophisticated gold and copper work. By 3000 BC, the explosion of new innovations ruptured traditional lifestyles across the continent and change occurred rapidly. As wealth became portable, emphasis shifted from places to people and from collective to personal possessions. A symbol of this change is Corded Ware pottery, which is decorated with impressed cords. Such pottery has been found in male graves located under circular mounds across northern and central Europe.

In addition to the pottery, men were found in single graves accompanied by stone battle-axes. This has both been interpreted as a sign of the importance of fighting and drinking and of individualization within a stratified society. A second type of pottery, handless drinking-cups known as Bell-Beakers, appeared concurrently and slightly later in the expansive area from Scotland to Sicily and Spain to central Europe. Graves were flat graves and in some regions also individual round mounds and, along with Bell-Beakers, contained daggers and archery equipment such as triangular barbed-flint arrowheads and wrist guards of fine stone. The vast distribution of the Bell-Beakers across temperate Europe and the Mediterranean indicates an increasingly mobile lifestyle that arrived just as traditional social structures disintegrated.

The Emergence of Elites: The Bronze Age

During the second millennium BC, bronze became the universal medium of ordinary objects and also of prestige. Horses and wheeled vehicles, textiles, amber necklaces, and metal dress-pins for clipping wool clothing emerged as visible signs of wealth. Trade in bronze expanded and bronze became popular for manufacturing weapons, ornaments, and tools. Most importantly, metallurgy shifted as copper became regularly alloyed and tin began to replace arsenic. Although copper is relatively rare as a naturally occurring material, tin is even rarer and the creation and circulation of bronze objects depended on regular supplies of both. This led to the rise of leading centers of metalwork, each of which made technical advances in bronze casting that were widely diffused and adopted. As techniques of war changed, so did the types of weapons used. New bronze weapons and designs spread across the continent.

In central Europe, in a region known as the Carpathian Basin, where the Únětice culture flourished, large hoards of bronze objects and magnificent burial mounds filled with gold and bronze objects have been uncovered. The influence of these communities stretched from Denmark to Italy to Greece, and to some degree, the British Isles. In the western Mediterranean, similar developments around 2200 BC spawned the establishment of new centers along the coast of the Iberian Peninsula. Here, hilltop settlements containing rectangular

houses and streets were surrounded by stone walls. These settlements lasted throughout the second millennium and are called El Agrar. The proximity of urban cultures in the eastern Mediterranean also provided alternative centers of trade and exposure to novel eating and drinking habits, clothing, furniture, means of transport, prestige ornaments, and weaponry.

Demand for bronze by the populations of the North Pontic steppes spurred trade and soon communication and exchange networks were established. North of the Caspian Sea, new grave architecture that included timber-built underground chambers known as Timber Grave culture was found. Eastward, the closely related Andronovo culture perfected the use of horses and created a horse-driven chariot complete with a bridle-bit with spiked cheek pieces made of antler. At the timber grave cemetery of Sintashta in the southern Urals, shaft graves similar to those of Mycenae were found containing wheeled chariots and horses.

When the palatial societies of the eastern Mediterranean collapsed after 1300 BC, dramatic changes swept across Europe. This period is known as the late Bronze Age and is characterized by hill-forts, flat cemeteries, scrap hoards, and votive finds from rivers. The most striking sign of the changes is the almost-universal shift from burying the dead in graves to cremating them and putting their ashes in pits or funerary urns. The period lasts until *c.*800 BC and has been named the 'Urnfield Period.' It is a transitory time, a phase in between the Bronze Age and the Iron Age, in which economic, social, military, and religious changes that are of pivotal importance in the formation of historic Europe occurred. For the first time, swords, spearheads, armor, safety pins, and certain ornaments from central and northern Europe appeared in the Mediterranean. Hilltop forts and stockades on lower ground became popular around 1100 BC and lasted until *c.*550 BC. At Biskupin, near Bydgoszcz in Poland, a fortified settlement island with 13 parallel rows of houses divided by roads and encircled by a rampart were excavated. The layout resonates communal order and indicates the presence of a coercive power. Throughout the period, exchange networks allowed goods to travel over vast distances and for the first time salt from the Austrian Alps and other places emerged as a valuable trade commodity. Hallstatt, the richest cemetery of the period, sits on top of a salt mine. Spiritual life also took on great importance and the bird became a beloved symbol. Figurines in the shape of birds have been found, as have numerous pots, cauldrons, and cult wagons with bird applications. Numerous rock carvings with engravings of humans, animals, boats, and a variety of symbols were found in Sweden (Tanum), the Alps at Val Camonica, north of Brescia, and on Monte Bego near Bordighera. The abundance of swords, daggers, spearheads, arrowheads, and armor found indicates the importance of war and the extensive choices available to the warrior.

From Chieftain to 'State' Organization in Celtic Europe: The Iron Age

By 600 BC, iron became the standard material for tools and weapons, while bronze was used for high-quality art products. The explosion of trading ports around the Mediterranean provided a strong incentive for European development. Mediterranean goods were sought avidly and the most valuable ones were acquired by the wealthy as proof of their exalted status. The valleys of the Rhone and Saône provided easy access to the region from Burgundy to southern Germany. This west Hallstatt zone was pivotal as it spanned the upper courses of the Seine, Saône, Rhine, and Danube rivers and directly linked the continent north of the Alps. Greek-led trade intensified through the colony of Massalia in the sixth century and the elite in this northern zone controlled trade of exotic Mediterranean products, including prestige equipment for wine-drinking rituals of the Greek symposium. Large valuables, such as the human-sized bronze crater from Vix, were imported and were exchanged for local products and slaves. The native elite maintained full control of these exchanges and likely held lavish feasts and gift exchanges among important chieftains, including the famous chief from Hochdorf. At graves, social prestige was demonstrated by the inclusion of costly sets of wine-drinking equipment and other exotic gear with the dead. Funerary rituals took place in large-scale religious structures surrounding burial mounds. At Glauberg in Hessen, a 1-km-long ceremonial path leading to a grave from the early La Tène period was recently discovered. Hilltops also emerged as prime locations for defended residences for the nobility. At Mount Lassois in Burgundy and at the Heuneburg in Germany, huge quantities of Mediterranean luxury goods were found. The Heuneburg was surrounded by a mud-brick wall built on dry stone footings with forward-projecting rectangular bastions. This technique, common enough in the Greek world, was totally alien to temperate Europe, and is a striking example of how avidly the European elite adopted Mediterranean achievements, even when they were impractical for the climate. In the east, in Slovenia, a second group of less-wealthy communities known as the east Hallstatt zone arose from the eighth to the fifth centuries. They were characterized by fortified settlements such as Magdalenska Gora, Vače, and Stična and large cemeteries with aristocratic graves furnished with sheet bronze body armor and decorated situlae.

The rise of a new warrior elite and the drying up of the market for luxury goods in the early fifth century destabilized the east and west Hallstatt systems and by the middle of the fifth century caused their demise. A new, northern elite arose and relied on native craftsmen to produce luxury goods. Numerous late fifth-century elite tombs have been found containing locally made masterpieces of Celtic-style art. Schools of creative metalworkers emerged during this period and their work led to the creation of the first truly Pan-European art style, known as Celtic or La Tène art. A rapid population increase, evident in the growth in the number of cemeteries, ensued and shook the social equilibrium. Over time, the system collapsed and resulted in a widespread migration. Late fifth- and early fourth century cemeteries, which extend south into the Po Valley and east along the Danube into Hungary, trace the path of this migration. Some early groups also crossed the great Hungarian plains and settled in Transylvania. In the Marne-Moselle region, elite burials continued at Somme-Bionne, Reinheim, and Waldalgesheim. But overall, the number of cemeteries decreased, indicating migration. In the tombs, young men were usually found accompanied by spears, swords, and shields. Wealthier men had bronze helmets and some possessed two-wheeled chariots. It is these warriors, called 'Gauls' or 'Celts,' who, in tens of thousands, made their way across Europe seeking new lands to

conquer and settlements to plunder. The Celtic migrations between 450 and 200 BC can be traced archaeologically across Europe through the La Tène culture, which is characterized by art depicting mobility, upheaval, and turbulence.

The northern limit of La Tène culture lay at the mouth of the Rhine and across southern Germany and northern Italy to the Carpathians. Further east, the Celts were well established in the lower Danube region. In the early fourth century, they began moving eastward into Transylvania and in the third century they pushed into southern Thrace and Greece. In the northeast, Celtic groups crossed the Prut and Dnestr, exploiting the power vacuum left by the decline of the Scythians. One of the few surviving pieces of silver made during the disrupted Celtic period is the Gundestrup cauldron found in Jutland. The puzzle of this elaborate cauldron encapsulates the complexity of these migrations. The techniques used to make it are typically Augustian Thracian. Some of the figures on it such as a warrior wearing a bird-crested helmet look Celtic, while other motifs are reminiscent of steppe shamanism. The cauldron was probably carried north as booty by raiding Germans.

Center and Periphery: Barbarian Europe and the Roman Empire

For Rome, the Rhine River marked an ethnic divide between the Celts to the south and the Germans to the north. In reality, tribes on both sides of the river had integrated to such a degree that few ethnic differences existed. Wealth derived from constant access to particularly fertile land and did not build up in the hands of any single group. Status could, however, be acquired by acts of bravery and leadership in raids. Power at this time lay with the elite and was measured by entourage and the ability to bestow patronage. The elites formed the tribal council. Caesar overran Gaul fairly easily because the Gauls were a settled people with a well-developed agricultural base and tribes centrally located on permanent defended settlements known as oppida. Oppida were the first urban centers in temperate Europe and have been found from France to Serbia. At Manching in southern Germany, the settlement was densely built up with regularly laid out timber buildings aligned along straight, ordered streets. A wide range of craft skills were practiced and coins were minted. Commodities such as wheel-turned pottery, glass beads, bracelets, and ironwork of every kind were done on an industrial scale. In oppida like Bibracte, cargoes of wine amphorae were exchanged with the local aristocracy for slaves and other commodities. This stability made Gaulish society soft and vulnerable. Germans were the opposite. Mobility and warfare were central to their lives and they had few southern luxuries. Since they had no oppida to protect, they could simply retreat and disappear. The Roman Empire desperately needed a continuous supply of raw materials and manpower. As its frontiers continued to expand, valuable metals and other commodities arrived. Rome's greatest dependence lay in its need for slaves to keep the Roman system working. The scale of the slave trade was enormous. Judging from the distribution maps of artifacts in Free Germania, Romans traded coins extensively, but also traded sets of metal vessels used for wine-drinking rituals, buckets, situlae, and jugs. In this zone, a number of rich burial sites known as Lübsow were found. The graves contained extraordinary luxury goods, indicating that the dead were members of the elite, who, by virtue of their rank, were able to acquire a set of the most costly Roman items. Such local elite served as middle men between the Roman world and the Baltic region. This system of socio-economic zones continued until the middle of the second century AD, when population expansion, social instability, and folk movements in central Europe built up and pressured tribes along the frontier. They then threatened the stability of the Empire. The territory occupied by the warrior burials coincides with the region from which the Burgundians and Vandals are thought to have emerged. The first recorded indications of the growing problem came in AD 162 when a frontier tribe, the Chatti, attempted to migrate south into Roman territory. A few years later, the Longobardi and Marcomanni crossed the Danube into Transdanubian Hungary. This was followed by even more massive movements by other tribes. These events, and the Marcomannic War fought by Rome between AD 166 and 180, mark the beginning of a wave of migrations which within three centuries destroyed the Empire and laid the groundwork for the formation of medieval Europe.

See also: Archaeology and Cultural Memory; Archaeology, Politics of; Classical Archaeology; Conflict and War, Archaeology of: Behavior and Social Organization; Ethnic Identity and Ethnicity in Archaeology; Migrations, Colonizations, and Diasporas in Archaeology; Prehistoric Art; Slavery: Comparative Aspects; States and Civilizations, Archaeology of; Tribe.

Bibliography

Barker, G. (Ed.), 1999. Companion Encyclopedia of Archaeology, 2 vols. Routledge, London and New York.

Bailey, G., Spikins, P. (Eds.), 2008. Mesolithic Europe. Cambridge University Press, Cambridge, UK.

Barker, G., 2009. The Agricultural Revolution in Prehistory: Why Did Foragers Become Farmers? Oxford University Press, Oxford, UK.

Bogucki, P., Crabtree, P. (Eds.), 2003. Ancient Europe 8000 B.C. to A.D. 1000: Encyclopedia of the Barbarian World, 2 vols. Scribner, New York, USA.

Clottes, J., 1996. The Cave Beneath the Sea: Paleolithic Images at Cosquer. Abrams, New York.

Die Gletschermumie aus der Kupferzeit: Neue Forschungsergebnisse zum Mann aus dem Eis 1999. Schriften des Südtiroler Archäologiemuseums 1. Folio-Verlag, Bozen.

Gamble, C., 1999. The Palaeolithic Societies of Europe, second ed. Cambridge University Press, Cambridge, UK.

Hänsel, B. (Ed.), 1998. Man and Environment in Bronze Age. The Bronze Age: The First Golden Age of Europe. Oetker-Voges Verlag, Kiel, Germany.

Herrmann, F.-R., Frey, O.-H., 1996. Die Keltenfürsten vom Glauberg: Ein frühkeltischer Fürstengrabhügel am Hang des Glauberges bei Glauburg-Glauberg, Wetteraukreis. Archäologische Gesellschaft in Hessen, Landesamt für Denkmalpflege, Wiesbaden, Germany.

Hoops, J., 2001. Reallexikon der Germanischen Altertumskunde, 17 vols. Walter De Gruyter, Berlin and New York.

Hansel, B. (Ed.), 1998. Man and Environment in the European Bronze Age. Oetker-Voges Verlag, Kiel, Germany.

Kristiansen, K., 1998. Europe before History. Cambridge University Press, Cambridge, UK.

Lichardus, J., Lichardus-Itten, M., 1985. La protohistoire de l'Europe. Le Neolithique et le Chalcolithique entre la Mediterranee et la Mer Baltique. Nouvelle Clio Ibis, Paris.

Müller-Karpe, H., 1998. Grundzüge früher Menschheitsgeschichte, 5 vols. Konrad Theiss Verlag, Stuttgart, Germany.

Scarre, C. (Ed.), 2013. The Human Past: World Prehistory and the Development of Human Societies, third ed. Thames and Hudson, London, UK.

Europe: Sociocultural Overview

Christian Giordano, Université de Fribourg, Fribourg, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Most concepts of Europe as a unitary imagined community are characterized by a bipolar scheme in which Europe is opposed to 'anti-Europe.' However, it is misleading to think of Europe as a united civilization or as a sum of cultural areas. Europe is considered as a system of structurally and culturally diverse historical regions.

The article also deals with the relevance of European nation-states in the past and in the framework of the present globalization processes. It shows how Europe experiences a difficult, tense, and contradictory transition from a Europe of the national communities to a postnational Europe.

The Term 'Europe'

In ancient Greece, the term 'Europe' evoked legendary or divine characters from the mythological universe rather than a geographic site.

The notion of Europe took on a more political and sociocultural connotation during the Carolingian period (seventh to eighth century). According to a medieval anonymous chronicle, the soldiers led by Charles Martel, who fought victoriously against the Arabs at Poitiers (732), were 'European.' At this time, the notion of Europe was assimilated with the Frankish Kingdom, and the idea of Europe as a political-spiritual entity developed along with the expansion of Arab-Muslim settlements north of the Strait of Gibraltar.

One of the recurrent themes of various subsequent concepts of Europe as an 'imagined community' (Anderson, 1983) can be noticed *in statu nascendi* from the latter observation. It is characterized almost without exception by a scheme in which the notion of Europe appears together with another contrasting term. In most cases, it is a dichotomy or bipolarity, rather than a comparison between what Europe is reckoned to be and what is supposedly its opposite. This 'European self-awareness,' as Denis de Rougemont (1990) calls it, becomes even more tangible at the time of geographical discoveries and the first colonial conquests.

From an anthropological point of view, Europe discovers itself by discovering and conquering the rest of the world. Between Europe and anti-Europe, which from time to time could be the Arab-Muslim culture or Asia, Africa, or the Orient, there is a reflective relationship in which the former identifies its basic traits through a presumed diametrical difference with the latter. Balthazar Gracian (1601–58) coined the famous phrase "Europe is the admirable face of the world." Following the same train of thought, the *Grand Dictionnaire historique de Louis de Moreri* (published in 1674) illustrates the European traits that have allowed Europeans to overpower the people of other continents. It is quite surprising to read statements in this publication that seem to foretell the stereotypes of German Völkerpsychologie as well as American studies on modal personality and national character. Similar arguments were expressed more briefly and clearly by Carl von Linné (1707–78); the Swedish naturalist describes Homo Europaeus as subtle and creative while Asians are melancholic and miserly and Africans are sly and lazy.

During the *ère des philosophes*, almost all of the eighteenth century's great thinkers – Montesquieu, Voltaire, Jean-Jacques Rousseau, Ferdinando Galiani, Melchior-Gaspar de Jovellanos, Antoine de Condorcet, Edward Gibbon, Edmund Burke – pondered over the 'European peculiarity.' Their dissertations share a common denominator: a scheme, usually bipolar, in which at least one other term of comparison appears besides Europe. Thus the main question for most of these authors is why European nations are the only ones that have reached such levels of civilization, while the other societies are still savage, primitive, or barbarian. Leaving aside the answers, almost all eighteenth-century texts on Europe draw either implicit or explicit parallels and hierarchies between 'we' and 'they,' between Europe and anti-Europe.

During the nineteenth century, the notion of Europe had an increasingly hegemonic connotation. The century was characterized by a bold optimism that turned out to be an awful self-deception. In historical terms, European elites had surprisingly underestimated the changes that had been going on beyond their continent. For example, they had overlooked the dynamic and intensive socioeconomic and cultural relations involving various Asian societies on the Indian Ocean's rim well before the arrival of European colonial empires; the formation of pre-Columbian empires in Mexico and Peru; the Japanese development model spurred by the emperor Mutsuhito, better known as Meiji Tenno (1852–1912), who had opened his country to foreign commerce; and even the American 'great transformation.' Throughout this century and to some extent the next one, Europeans would obstinately emphasize their superiority, as Eric Wolf has aptly highlighted, by stubbornly disowning the history of others (Wolf, 1982).

It is no surprise that the term 'Europe' was being increasingly used as a synonym with 'civilization' and 'progress' (Schmitt, 1974). From a European point of view, the Old Continent was the only one that could establish a 'civilizing movement' to 'break through the darkness that enfolds whole populations' as Leopold II (1835–1909), King of the Belgians, stated in a solemn speech in 1876. For this typical representative of the time, European colonial expansion in the heart of the anti-Europe of that moment, namely Africa, was 'a crusade worthy of this century of progress.'

Nineteenth-century certainties were radically curtailed by the First World War, which the German philosopher Karl Löwith (1897–1973) keenly called 'the first European civil war' (Löwith, 1990); by the critical analysis of the nationalistic and totalitarian frenzy between the two World Wars; by Second World War traumas; and finally by the melancholic awareness of crumbling colonial empires. Europe as a notion became less triumphalist and much more defensive. After 1945, Europe was perceived as suffering from a chronic economic, cultural, and above all political and military fragility. This intrinsic deficiency could be checked only by a strong synergy between all the continent's components sharing the same fate. In this respect, the French sociologist Edgar Morin (1987) has fittingly spoken about Europe as a *Schicksalsgemeinschaft*, a term first used by the Austro-Marxist theorist Otto Bauer (1881–1938). From a sociological point of view, the European fragility concept implies a menace coming from external hostile forces. Each community, especially if it is a *Schicksalsgemeinschaft*, is strengthened or crystallized if under pressure from an actual or presumed enemy. The part of Europe that after the Second World War was not under Soviet control always reverted to the anti-Europe theme to emphasize who the menace was. From 1945 until 1989, while Europe was establishing its first joint institutions, the leading anti-Europe menace was the Soviet and communist sphere.

The fall of the Berlin Wall (1989) and the subsequent disintegration of the opposing blocs coincided with the birth of a new European institution, the European Union (EU). A fragile and 'menaced' Europe was still a common perception in the Old Continent despite this new highly specialized and official legal apparatus. Behind this defensive attitude, some authoritative, although overly severe, 'Euro-skeptic' critics detected the danger of the creation of a European fortress that wishes only to defend an acquired economic standing threatened by supposed external menaces. At the end of the twentieth century, the bipolarity between Europe and anti-Europe still existed. Now anti-Europe is mainly the planet's boundless south, with its migratory waves toward the Old Continent, its high demographic rates, and political–religious conflicts.

What Europe Is Not: Some Misleading and Dangerous Ideas

During the twentieth century, notions of Europe with powerful ideological implications have occasionally been spread; yet they can hardly endure a historical, sociological, and anthropological analysis. The idea of European civilization as a monolithic entity was most popular and politically relevant. European civilization was seen as the bulwark of Christianity. This concept of Europe was an ideological backbone of antidemocratic political movements. Though admittedly in conflict with Christian doctrine and ecclesiastic institutions in many aspects, German Nazism and Italian Fascism envisioned themselves as the upholders of European civilization and strenuous paladins of Christianity, particularly against Soviet Bolshevism, which was stigmatized as Asian and hence alien to Europe. Other régimes with similar ideological backgrounds, such as authoritarian and dictatorial governments and movements in Central and East European countries as well as Francoism in Spain and Salazarism in Portugal, resorted to analogous ideas.

Europe as a united civilization embodied in the Christian Occident is an idea that can still be found in present-day European extreme rightists, such as the French Front National and neo-Nazi and neofascist fringes in Germany, Austria, and Italy. These movements resort to similar concepts to assert their hostility toward immigration, especially from Islamic and/or non-European countries. It is a dangerous idea because it implies the pursuit of cultural, denominational, and maybe even ethnic homogeneity based upon the 'exclusion strategy' and the 'purity myth,' which caused so many of Europe's tragedies during the twentieth century. This vision of Europe is misleading since it is far-removed from the continent's past and present social realities.

Europe: A System of Historical Regions

The German philosopher Ernst Bloch (1885–1977) advises us not to turn a highly differentiated social entity like Europe into a sum of 'cultural areas' because this would turn "History … into a circus [with] three or more [separate] rings" (Bloch, 1985). Europe must be considered a system of strictly (inter)dependent, yet structurally diverse historical regions. Hungarian historian Jenö Szücs (1983) was the first one to apply a strictly scientific method to the problem of Europe's internal frontiers, and thereby to the question of its historical divisions, and consequently to Europe's sociostructural differences.

Szücs stresses that from the early Middle Ages, Europe began to differentiate socioeconomically, giving rise to two different poles: *Europa occidentalis* and *Europa orientalis*. The boundary would be an ideal line running from the Elbe River (a socioeconomic boundary between western and eastern Germany) to the Leitha River (for many centuries a pervious frontier between Austria and Hungary). A second economic and sociostructural division, which includes all societies involved in the Great Schism (1054), came to pass between the eleventh and fourteenth centuries. Thus, there are three historical regions with very diverse socioeconomic, political, and cultural characteristics. The first includes the territory from the Atlantic to the Elbe–Leitha ideal line, corresponding more or less with Carolingian Europe, today believed to be the quintessence of Western Europe. The central historical region east of the Elbe–Leitha line but west of the eastern and southern borders of the original kingdoms of Hungary and Poland (Grand Duchy of Lithuania included), approximately overlaps what is known today as mid-Eastern Europe, also called *Mitteleuropa*. Everything east of the Baltic–Black Sea line or south of the Carpathians is Eastern Europe proper.

Szücs' arguments are relevant for the social sciences because they reveal how processes dating back to a remote past have created lasting economic, sociostructural, and cultural differences in Europe; however, his hypothesis of at least three historical regions must be broadened and further examined. For this purpose Immanuel Wallerstein's analysis of Europe incorporated in a 'world-system,' yet divided into a 'core' and 'peripheries' (Wallerstein, 1974), is well suited. A new system to organize and manage the economy emerged and spread

between 1450 and 1640. Max Weber labeled this new system rational capitalism. According to Wallerstein (1974), as well as to French historian Fernand Braudel (1979) among others, expansionist movements are inherent to this new model of economic and social behavior. The capitalist economy's worldwide expansion implies new relationships between economic partners, that is, societies, states, regions, towns, and so forth. These new relationships entail a new international division of labor that will permeate social structures and cultural models of all societies involved.

The international division of labor within the world-system is also based on a systematic separation between core and peripheries. Europe at first, and then the whole world, came to be divided into a number of regions. Some of these underwent a historical acceleration, quickly modernizing most of their socioeconomic structure, while others suffered centuries of socioeconomic stagnation if not growing impoverishment. The following classification highlights the roles of Europe's historical regions within the new international division of labor.

Northwestern Europe

The first core of the world-system developed in this region. Steady decline of the feudal system, early establishment of an extended net of self-governing towns (*freie Reichstädte*, *comuni*, etc.), striking capital accumulation followed by important technological changes culminating in the industrial revolution, rise of strong territorial states with more culturally homogeneous populations, colonial expansion outside Europe: these are only some of the most evident processes marking the socioeconomic and cultural structure of this historical region. From the onset, it included southern England, France, Flanders, Holland, and the westernmost part of Germany. Northern Spain and Northern Italy, other areas of Germany, and most of Switzerland, Austria, Bohemia, and Scandinavia, joined the core in subsequent periods.

Mediterranean Europe

This historical region, for centuries the continent's socioeconomic and cultural hub, was turned into a vast periphery by the new international division of labor. It provided the core with foodstuffs and raw materials for the growing textile industry. From a socioeconomic point of view, the area comprising Portugal, Spain, and Italy was ruled by a system of wheat monoculture and/or stock farming with an absentee owner aristocracy. The vast properties were administrated by large tenants who would sublease to smaller ones, who in turn would sublease to the smallest peasants. This long renting chain led to a devastating fragmentation. Too many people without a productive activity lived exclusively off the land and backs of those who actually tilled it.

Even though the banking system, rational capitalism's pillar, was developed in Renaissance Italy, and the two Iberian states were the first ones to rely on imperial colonies, Mediterranean Europe's structure was rooted in agrarian societies and economies from the sixteenth century up to most of the twentieth century. The rising bourgeoisie of this historical region did not play a specific role, nor did it have its own specific economic behavior. Its ideal was to imitate the aristocracy. This matched the economic and social behavior of categories of tenants such as the *señoritos* in Andalusia, La Mancha, and Extremadura, or the *gabelloti* and *galantuomini* in Sicily and southern Italy, who were inclined to a passive existence on the principle of 'dignified idleness.'

Central–Eastern Europe

This historical region, with few exceptions, lies between the two ideal boundaries traced by Szücs. It is Europe's second vast periphery and has been one of the core's major suppliers of foodstuffs. The typical socioeconomic scenery of Europe's second periphery is in western and eastern Prussia, the Baltic countries, the Hungarian plain, and the Romanian Danube plain, together with Poland's so-called eastern territories (most of which were lost after the Second World War). In these boundless areas were the landed estates of Central–Eastern Europe's aristocracy: important Polish and Hungarian noble families, Baltic overlords of Germanic ancestry, and Romanian boyars. The institution of serfdom was the rule in these estates nearly up to the end of the nineteenth century. In fact, rather than being abolished, serfdom relationships were revived in what historians call the 'second serfdom.' Great landowners introduced new farming methods in order to increase crops, mainly cereals, causing harsher labor and life circumstances for the peasants. Central–Eastern Europe's second serfdom was also accompanied by the formation of new groups of great landowners with a strong corporate spirit and staunch political and socioeconomic ambitions, such as the Prussian *Junker*, Polish *Szlachta*, and Hungarian gentry.

Southeastern Europe

This historical region mainly includes the Balkans (except Romania). The Ottoman Empire that dominated it for almost 500 years became a part of the world-system only during the eighteenth century, that is, during its manifest decadence. However, even under the Ottoman domination the Balkans was a poor periphery on the precarious border between the Ottoman and Austro-Hungarian empires. Therefore, the circumstances of this historical region should not be compared to other more flourishing socioeconomic areas of the Ottoman Empire, such as parts of Syria or Egypt. To make matters worse, in the Balkans there was a degeneration of the patrimonial system based on 'tax-farming.' The Balkans' fate within the world-system was to become a 'periphery of the periphery.'

Eastern Europe

Europe's third historical region, according to Szücs' classification, comprises mainly Russian territory west of the Urals with the exception of the areas south of the Carpathians. During its first centuries, this enormous country was an 'external area' as regards the world-system, with a distinctive socioeconomic system. A communitarian land property regime and a peculiar patrimonialism by which an imperial official 'arrogated' property administrated on behalf of the czar were the socioeconomic structure's cornerstone. Historians compared the situation engendered by such a structural peculiarity, which lasted up to the Russian Revolution (1917), to a 'bog of misery'

(Riasanovsky, 1984). However, Eastern Europe's gradual integration in the world-system, beginning in the eighteenth century, remained incomplete. Russia continued to be a European external area and never became a real periphery or a core-state as other historical regions already described.

Historical Regions and Present Socioeconomic Gaps

The establishment of the world-system and the growth of a new international division of labor, which have molded the European core and peripheries, are the primary components of an epochal historical change whose socioeconomic and cultural consequences are visible to this day.

The gaps that appeared between Europe's historical regions during the formation of the world-system are still plain to see. The North–South gap mirrors past historical sociostructural and economic differences between the core and the Mediterranean periphery. To date, even the integration of these two historical regions within the EU has not eliminated the gap. The enlargement of the EU to the east is a problematic issue due to West–East discrepancies (and not only to 50 years of socialism) that embody old boundaries between the European nucleus and the central–eastern periphery. The Balkan situation, which is characterized by a feeble industrialization and an incomplete modernization, is due to the fact that for centuries this area was a periphery of the periphery. Finally, present-day Russia, with its distinct development model torn between capitalistic models and (pre)Socialist communitarist revivals, is adhering to its typical tradition of lingering on the borders of the world-system. The core-periphery gaps are detectable not only interregionally but also within the states. In fact, states that straddle historical regions have strongly dualistic societies, economies, and cultures. For example, Italy is divided into north and south; Spain has discrepancies between Catalonia and Spanish Basque Country on the one hand and the rest of the country on the other; in former Czechoslovakia there were differences between Bohemia and Slovakia; in former Yugoslavia, Slovenia and Croatia were more connected to the core while Serbia, Bosnia, and Macedonia belonged to the southeast periphery.

At the end of the twentieth century and in the early twenty-first century, the structural gap between core and peripheries was held to have been overcome thanks to the creation of the EU, to its various socioeconomic reallocative mechanisms, and especially to the introduction of a single currency. The current government debt crisis of the Mediterranean area countries (Portugal, Italy, Greece, and Spain), the severe economic problems in East and Southeast Europe's former communist countries (Latvia, Lithuania, Hungary, Slovakia, Romania, and Bulgaria), and the hegemonic aims and stances of some of Europe's northwestern nations (Germany in particular) are evidence that asymmetrical and dependency relations between core and peripheries are definitely not over.

Perceived Gaps: Core and Peripheral Identities

Core–periphery gaps go beyond structural data. They are also 'lived,' and therefore perceived and constructed realities. In fact, in the core there is a feeling of belonging dominated by the concept of techno-economic and social progress, of human perfectibility, and, based on the values of enlightened rationalism, of secularized and individualized societies, the spirit of capitalism, and so forth. Max Weber would have described such identities as 'disenchanted' ones. Maybe the most fitting description of this type of identity is precisely the one supplied by Max Weber in his model on Occidental uniqueness (Weber, 1986), which, despite what Max Weber intended to prove, was used to validate European superiority and the core's hegemonic aims on the peripheries.

In European peripheries instead, especially among the intellectual elites, there is an awareness of the marginality if not the backwardness of one's country or its historical region. However, there is also a strong belief that one's nation or people has outstanding excellence and virtues. Peripheral elites have been coping with the discrepancy between their nation's perceived socioeconomic inferiority and its presumed spiritual and moral superiority even at the brink of the formation of their respective nation-states. European peripheries' ruling-class interpretation of this apparent contradiction has been remarkably ambivalent as regards to relationships with the core as a socioeconomic and cultural complex. Quoting Reinhard Bendix's terminology, the core is looked upon as a 'reference society,' with whom there is both an 'identification' and a 'distinction' relationship (Bendix, 1978). An emulation of the core is encouraged, while dissociation is maintained as well. Over the past 150 years, almost every European periphery has undergone major modernization processes emulating the core, which affected all aspects of society. However, these processes were accompanied by archaistic countertrends (agrarian populist movements, nationalisms, etc.) that instead glorified what was deemed as the past, tradition, culture, and identity of one's group. A modernization drive and a nativistic revival were both at work.

A modernization drive basically means a Europeanization, particularly of politics (establishment of parliamentary institutions and legal systems), economy (industrialization), culture (more theaters, art museums, avant-garde arts, etc.), and even everyday life (new habits, conventions, and ephemeral fads from the capitals of 'reference societies' – Paris, Vienna, or Berlin). Nativistic revival means stressing one's distinguishing difference from the core and revitalizing imagined social institutions, guardians, and bearers of primordial virtues, such as the village, rural community, medieval municipality, or guilds.

Europe and the Sociocultural Order of Nation-States

Since the nineteenth century, capitalist economy's expansion in world-system Europe has been accompanied by a proliferation of nation-states that were established (except in the United Kingdom) through a combination or a separate use of two main models first conceived in France and Germany (Brubaker, 1992). Notwithstanding some differences, both concepts are based on the doctrine by which each nation is entitled to its own territorial state and each state must include only one nation.

Because of this ideal, the last two centuries have been marked by repeated efforts to make single national territories more and more ethnically and culturally homogeneous. Already from the twelfth to the seventeenth centuries, Western

Europe was overrun by several homogenization waves in the shape of religious conflicts and persecutions. Even in the more culturally homogeneous nation-states of Western Europe such as France, however, assimilation processes continued to the end of the nineteenth century (Weber, 1976). In Central and Eastern Europe instead, where the doctrine was applied only after the downfall of the multicultural empires, the processes of ethnocultural recomposition aimed at 'ethnic purity' have been implemented only during the nineteenth and twentieth centuries through a series of boundary revisions, forced assimilations, expulsions, aimed and planned immigrations, deportations, ethnic wars, and genocide.

Four main periods can be identified in the various processes of ethnic separation that over the last two centuries concerned the European nation-states, especially in the central and eastern parts of the continent. The first period occurred predominantly in the Balkans, immediately after the creation of the first nation-states in the nineteenth century. Vast population groups of Turkish origin or just of Islamic faith were forced to leave the region. As administrators of the Ottoman Empire, some did indeed represent the hated occupiers, but members of social strata that had little in common with the ruling class were also expelled. During the great 'Crisis in the Orient,' which led to the bloody Russian–Turkish war, between 1875 and 1878, a million and a half people were repatriated.

The second virulent phase took place between 1913 and 1925. It was characterized by the forced transfer of whole minority groups, yet it was internationally recognized and guaranteed. In diplomatic parlance, it was euphemistically termed 'population exchange.' Substantial groups of Albanians from Kosovo and western Macedonia were transferred to Turkey after the Balkan Wars (1913) mainly because of their religion. Particularly after the creation of the kingdom of Yugoslavia, the Islamic populations were replaced with Serbian, Montenegrin, Croatian, and Slovenian people. The 'population exchange' between Greece and Turkey was even more dramatic. The Treaty of Lausanne in 1923 ratified a series of reciprocal expulsions and hasty repatriations caused by the Greek military catastrophe during the campaign in Asia Minor. After the defeat, Greece was overrun by refugees from the coasts of western Anatolia, plus the Greeks from the Black Sea area and the Caucasus fleeing from the new Bolshevik régime's repressions. A country with 4.5 million inhabitants faced the arrival of 1.3 million refugees. At the same time, the 'population exchange' provided for the departure of Turks and Albanians.

The third phase of ethnic homogenization includes the decade between 1940 and 1950, marked by the Nazi policy of annihilation, transfer, and expulsion of whole ethnic groups and by Stalinist deportations and purges. Along with the holocaust of the so-called 'transnational minorities,' that is, the Jews and Roma, there were massive population movements in Central and Eastern Europe. Eleven and a half million Germans were expelled from the Ostgebiete (eastern territories) while three million Poles, particularly from the regions absorbed by the Soviet Union after the Second World War, were settled in Silesia and in the southern portion of eastern Prussia. Thus Poland became an almost monoethnic country.

The treaties between Czechoslovakia and Hungary, and between Hungary and Yugoslavia, which also provided for reciprocal population exchanges, date back to the same period. Finally, Stalin consolidated his conquests in the western part of the Soviet Union through a policy of planned mobility. This involved the deportation of populations considered accomplices of the enemy and therefore traitors of the great patriotic war (Latvians, Lithuanians, Estonians, etc.) to Siberia or central Asia. It also involved replacing them with more 'reliable' immigrants, mainly of Slavic origin.

The fourth virulent phase of ethnic homogenization is the wave of ethnic separations that has devastated Central and Eastern Europe over the past 25 years. It can be traced back to Socialist Bulgaria with its approach to minority problems. In the second half of the 1980s, the Bulgarian solution was to expel and/or forcibly assimilate ethnic Turks. The phase continued with the disintegration of the three countries born after the First World War through a multiethnic and multinational logic: Yugoslavia, the Soviet Union, and Czechoslovakia. The resulting new and old states are all based on the nation-state principle. This latest wave of separations, which aims at the homogenization of territories with highly differentiated ethnic structures and claimed by one or more nations, continued in the 1990s in Bosnia and Kosovo, and may not yet have come to an end, particularly in the Balkans and former Soviet Europe.

Postnational Europe in the Making?

Over the final 50 years of the twentieth century, there were major population movements in Europe; this is the most outstanding characteristic of this period. These migrations for economic (job seekers) or political reasons, namely humanitarian (asylum seekers), have affected Western Europe more than Central–Eastern Europe. Until 1989, Soviet hegemony in this area allowed only a limited and often planned mobility, even though immigrant laborers were not totally lacking (Vietnamese in Bulgaria, former Czechoslovakia, German Democratic Republic, and the USSR as well). Within the EU and its previous community institutions, a gradual, albeit slow, frontier abatement among member states eased transnational mobility. From the end of the 1950s, unprecedented numbers of laborers from the Mediterranean and Southeast peripheries (Italy, Spain, Portugal, Greece, and the former Yugoslavia) reached the core (France, Germany, Austria, Switzerland, Benelux, the United Kingdom, and Scandinavia). Here they found jobs in industry and in fields that are not sought after by the local population because they are considered low-status activities. From the 1970s, these immigrants were joined by an increasing number of laborers from the Mediterranean area bordering on the Old Continent (Turkey and Maghreb). Although, so far some countries, such as Germany and Switzerland, challenge this designation, northwestern European countries became immigration societies in this period, notwithstanding strict measures introduced in the early 1970s regulating foreigners' permanent entry.

The circumstances in Euro-Mediterranean former emigration societies (Italy, Spain, Portugal, and Greece), which in the meantime have become immigration countries, are slightly different from Northwestern Europe's situation. Migration is mainly from so-called Third World countries (sub-Saharan

Africa, Southeast Asia, China, the Indian subcontinent, Latin America, and so forth) and from the continent's most underdeveloped areas (Albania). From the end of the 1980s, however, what became known as mass unemployment worsened in all of Western Europe. Therefore, recruitment of foreign workers fell drastically while political and humanitarian immigration increased.

At the same time, large numbers of migrants reached the core from non-European regions and from former Socialist bloc countries after the fall of the Berlin Wall (1989) with the subsequent opening of frontiers. Facing a difficult sociopolitical transformation and an unstable economy, Central–Eastern Europe is not the preferred destination for these migrants, who regard it only as a stopover on their way to Western Europe. Therefore, former 'Real Socialism' countries become a temporary buffer zone.

These migratory circumstances have deeply changed the present Europe's setup, calling in question the sociocultural order created by nation-states based on the ideal of a monoethnic territory. The growing presence of diasporas, transnational groups, creolized communities, and deterritorialized collectivities in Europe's social and cultural setting, plus minorities' emphatic claims and demands (linguistic, religious, and ethnic), have fired discussions on nation-state adequacy and the need for postnational organizations with new types of transnational, flexible, or multiple citizenships, as Ferry (1992), Pierré-Caps (1995), Kleger (1997), Ong (1999), Linklater (2002), or Sassen (2002) foresee. However, such projects might be taken into serious consideration only as the twenty-first century progresses.

For the time being, globalization in Europe has led neither to a neoliberal homogenization as suggested by Fukuyama (1992) nor to a McDonaldization of society as conceived by Ritzer (1996, 1998). Europe has instead experienced a *glocalization*, as postulated by Robertson (1995). More to the point is the resurgence of political regionalisms. Current versions of political regionalisms, contrary to those of the 1970s, can no longer be likened to ideology or opinion movement rebellions if not mass uprisings. Instead, they involve arrangements and compromises with the national states' governments. Hence, present-day regionalisms are not so much *counter*- as *co*-phenomena. By now, the term 'regionalism' also has its place in the common parlance of politicians and bureaucrats. We need only mention phrases such as 'the Europe of regions' or terms as 'devolution,' 'federalism,' and 'autonomism.'

Present-day regionalisms in Europe are political orientations seeking accommodations or compromises with globalization through negotiations. They are grounded in the appeal to local identitary belongings in an attempt to adapt to radical sociostructural changes triggered and/or imposed by the globalization process. Furthermore, they are social strategies aiming to overcome globalization-induced difficulties and problems via the mobilization of local societies whose members think and/or believe that they belong to the group because of specific shared qualities. Finally, regionalism is a response in line with national states' gradual cession of sovereignty to social, political, and economic institutions of global governance.

Admittedly, between the gradual cession of national states' sovereignty and regionalistic claims, there is now a dialectic relation that generates new political setups through constant accommodations, in some cases even based on a lasting, though often shaky, power sharing, both locally and nationally. Not to mention that nowadays in Spain or Romania, Italy or Bulgaria, governing is a difficult task without the support of parties advocating regionalism.

In the European context, we can distinguish between ethnic regionalism and cosmopolitan regionalism. By ethnic regionalism we mean a sociopolitical movement striving to build an imagined collectivity whose belief in commonality is based on ethnic criteria regarded as qualities unique to one's own group, such as origin, ancestry, history, traditions, culture, religion, language, and last but not least, territory.

At first glance, cosmopolitan regionalism by its very nature should be considered postethnic (in line with the definition of this term by David Hollinger, 1995), transboundary, and transnational. In principle, this type of regionalism:

- should make boundaries between groups involved in these types of regionalism more flexible, if not indeed variable,
- should allow the individuals involved to have plural belongings,
- should accept hybrid identities for these same individuals.

Cosmopolitan regionalisms, at least in principle, have a 'social imaginary' that certainly lies outside that of the national state; however, as far as their currently inadequate political realization is concerned, for the time being we can only be skeptical.

As an example, consider the Euroregions, which by juridical definition are transboundary and thus transnational. These were meant to be institutions in charge of carrying out cosmopolitan and postethnic projects at a local level. In most cases, however, despite the proliferation of these projects and some instances of zeal and initial enthusiasm that led to belief and hope in the formation of bottom-up participation movements, the top-down aspect was strikingly prevalent. This demoted the Euroregions to a question of bureaucracy and technocracy for the most part, ultimately leading local societies to back out of the projects.

In the main, realized regionalisms are ethnoregionalisms grounded in the 'social imaginary' stemming from the national state, and moreover from an ethnic-based one. Ultimately, as the emblematic cases of Catalonia, Galicia, Spanish Basque Country, Corsica, Flanders, Walloon Provinces, Scotland, and 'Padania' show, until now the ethnoregionalisms resulting from globalization have been unable to express political models as alternatives to those of the classic national state (Calhoun, 2007).

Bibliography

Anderson, B., 1983. Imagined Communities. Reflections on the Origin and Spread of Nationalism. Verso Editions and NLB, London.

Bendix, R., 1978. Kings or People. Power and the Mandate to Rule. University of California Press, Berkeley, Los Angeles, London.

Bloch, E., 1985. Tübinger Einleitung in die Philosophie. Suhrkamp, Frankfurt/Main.

Braudel, F., 1979. Civilisation matérielle, économie et capitalisme, XVe-XVIIe siècle. Le temps du monde, vol. 3. Librairie Armand Colin, Paris.

Brubaker, R., 1992. Citizenship and Nationhood in France and Germany. Havard University Press, Cambridge, MA, London.

Calhoun, Craig, 2007. Nations Matter, Culture, History and the Cosmopolitan Dream. Routledge, London.
Ferry, J.M., 1992. Pertinence du postnational. In: Lenoble, J., Dewandre, N. (Eds.), L'Europe au soir du siècle. Identité et démocratie. Editions Esprit, Paris, pp. 39–57.
Fukuyama, F., 1992. The End of History and the Last Man. Macmillan Free Press, New York.
Hollinger, D., 1995. Postethnic America: Beyond Multiculturalism. Basic Books, New York.
Kleger, H. (Ed.), 1997. Transnationale Staatsbürgerschaft. Campus, Frankfurt/Main.
Löwith, K., 1990. Der Mensch inmitten der Geschichte. Philosophische Bilanz des 20. Jahrhunderts. Verlag J.B. Metzler, Stuttgart.
Linklater, A., 2002. Cosmopolitan citizenship. In: Isin, E.F., Turner, B.S. (Eds.), Handbook of Citizenship Studies. Sage, London, pp. 317–330.
Morin, E., 1987. Penser L'Europe. Gallimard, Paris.
Ong, Aihwa, 1999. Flexible Citizenship: The Cultural Logic of Transnationality. Duke University Press, Durham NC.
Pierré-Caps, S., 1995. La multination. L'avenir des minorités en Europe centrale et orientale. Editions Odile Jacob, Paris.
Riasanovsky, N.V., 1984. A History of Russia. Oxford University Press, Oxford.
Ritzer, G., 1996. The McDonaldization of Society: An Investigation into the Changing Character of Social Life. Pine Forge Press, Thousand Oaks, London, New Delhi (Revised Edition).
Ritzer, G., 1998. The McDonaldization Thesis: Explorations and Extensions. Pine Forge Press, Thousand Oaks, London, New Delhi.
Robertson, R., 1995. Glocalization: time-space and homogeneity–heterogeneity. In: Featherstone, M. (Ed.), Global Modernities. Sage, London, pp. 25–44.
Rougemont, D de, 1990. Vingt-huit siècles d'Europe. La conscience européenne à, travers les textes – d'Hésiode à nos jours. Christian de Bartillat, Editeur, Paris.
Sassen, S., 2002. Towards post-national and denationalized citizenship. In: Isin, E.F., Turner, B.S. (Eds.), Handbook of Citizenship Studies. Sage, London, pp. 277–291.
Schmitt, C., 1974. Der Nomos der Erde im Völkerrecht des Jus Publicum Europaeum. Duncker & Humblot, Berlin.
Szücs, J., 1983. Vázlat Európa Három Történeti Régiójáról (The Three Historical Regions of Europe). Magvetö Könyvkiadó, Budapest.
Wallerstein, I., 1974. The Modern World-System. Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century. Academic Press, New York, San Francisco, London.
Weber, E., 1976. Peasants into Frenchmen: The Modernization of Rural France, 1800–1914. Stanford University Press, Stanford.
Weber, M., 1986. Gesammelte Aufsätze zur Religionssoziologie. Mohr & Siebeck, Tübingen.
Wolf, E.R., 1982. Europe and the People without History. University of California Press, Berkeley.

Euthanasia

Dieter Birnbacher, Institut für Philosophie, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Euthanasia' is derived from Greek words meaning 'dying well' or 'good death.' First used by Greek poets from about 400 BC onward, the term was adopted by Greek and Roman philosophers and historians to denote a death that comes quickly and without great and prolonged suffering. Today, the term is primarily used for interventions by others to ease death. The article addresses both questions in the ethics of euthanasia and questions concerning the evolving research on the practice of euthanasia, mainly in European countries. In order to structure the discussion, it distinguishes five meanings of 'euthanasia': (1) caring for the dying by relieving suffering psychologically and medically, and by spiritual assistance; (2) letting die suffering patients by not artificially prolonging life; (3) actively hastening death to shorten suffering by assisting patient suicide, by voluntary or nonvoluntary active euthanasia; (4) withholding life-sustaining treatment from a patient in a persistent state of unconsciousness; and (5) ending the lives of children and adults seen as a burden to society or to their families.

For many reasons, during the last 20 years, euthanasia has increasingly become a subject of social scientific research. Analogously to public opinion about sexual morality, public opinion about the morality of medical end-of-life practices seems to undergo rapid changes, partly in consequence of the continuing prolongation of the human life span by medical progress and social improvements, partly in consequence of the decreasing influence of religious traditions on social attitudes and legislation, at least in the industrialized world. Living wills are increasingly used by patients to provisionally exercise their constitutional rights to self-determination in case they lose their capacity to give valid consent to medical treatment in the future, and they are increasingly backed up by institutional safeguards (cf Council of Europe, 2004). Living wills are increasingly made the subject of explicit legislation, and more and more hospitals include living wills of patients in their conditions of admission. Furthermore, social science research is urgently needed in order to assess the far-reaching empirical claims made by the parties on both sides of the political controversy: the claim, on the one hand, of right-to-die societies that the personal preferences concerning the time and mode of dying continue to be insufficiently respected by legislators and medical societies, and the claim, on the side of conservatives, that a more liberal legislation creates a slippery slope to morally inacceptable misuse and abuse. Psychological and social research in the end-of-life context, however, faces particular difficulties and ethical challenges. Anecdotal evidence suggests that physicians tend to be secretive about their treatment of end-of-life patients and that official statistics tend to be unreliable, particularly in countries with laws prohibiting even less controversial forms of euthanasia such as termination of treatment. Furthermore, research in end-of-life contexts is hampered by severe ethical restrictions. Patients at the end of life are generally highly vulnerable and research involving them may be felt as a problematic intrusion.

The article addresses both questions in the ethics of euthanasia and questions concerning the evolving research on the practice of euthanasia, mainly in European countries. In order to structure the discussion, it distinguishes five meanings of 'euthanasia': (1) caring for the dying by relieving suffering psychologically and medically, and by spiritual assistance; (2) letting die suffering patients by not artificially prolonging life; (3) actively hastening death to shorten suffering by assisting patient suicide, by voluntary or nonvoluntary active euthanasia; (4) withholding life-sustaining treatment from a patient in a persistent state of unconsciousness; and (5) ending the lives of children and adults seen as a burden to society or to their families.

Introduction

In the original Greek, 'euthanasia' meant 'good dying' (*eu thanatos*). First used by Greek poets from about 400 BC onward, the term was adopted by Greek and Roman philosophers and historians to denote a death that comes quickly and without great and prolonged suffering. The term 'euthanasia' was not used for the intervention, medical, or otherwise, to ease a patient's death or to end his or her suffering. Such intervention cannot, however, have been uncommon in antiquity, since assisting a patient's suicide was explicitly ruled out in the so-called Hippocratic oath. Whoever swore this oath (probably the members of a small Pythagorean sect of physicians) committed him- or herself to "give no deadly medicine to anyone if asked, nor suggest any such counsel."

In modern times, the meaning of the concept has considerably changed. Though its core meaning of 'good dying' was preserved, it no longer refers to the quality (however defined) of the process of dying as such, but to the cooperation of someone else (the physician, the priest, or others) in making death easier for the dying. Francis Bacon, in a passage of *De dignitate et augmentis scientiarum* (1623), was the first to distinguish two kinds of assistance to which, according to him, a dying patient has a claim, *euthanasia interior*, the "spiritual preparation of the soul" for death and *euthanasia exterior*, "whereby the dying may pass more easily and quietly out of life" (Bacon, 1857: p. 387). Bacon's point was that the purely spiritual preparation provided by the priest was not enough but should be supplemented by medical means to ease the dying process.

By this token, Bacon suggested the introduction of what is now called palliative care, the systematic effort to relieve, by medical means, pain, and other symptoms causing suffering to the dying. Euthanasia in this sense, Bacon insisted, was not only a natural part of the duties of the physician but also a requirement of plain humanity. Though psychological and spiritual assistance in dying continues to be recognized as one of the central elements in the care for the dying (often, however, neglected in practice), the modern concept of euthanasia includes the 'easing of death' only in so far as this is effected by the application or nonapplication of medical means.

The term 'euthanasia' is not commonly applied to procedures by which the death of a patient is eased by any action or omission the patient carries out without the direct or indirect involvement of a second party, such as suicide in a nonfinal state of a progressive disease or the refusal of nutrition and hydration with the aim of hastening death (Chabot, 2012).

Five Types of Euthanasia

The principal meanings of the term 'euthanasia' in ethical and legal discussion correspond to the various forms in which medical assistance in dying can be given:

1. Relieving suffering by palliative care, e.g., by giving pain-killing drugs and suppressing symptoms such as vomiting, unrest, or feelings of suffocation. Palliative care for the dying is rarely referred to as 'euthanasia' in the absence of the more typical practices (2) and (3). However, the term 'indirect euthanasia' is generally applied to cases in which medical interventions aimed at relieving suffering (such as administering drugs, sedating a patient with narcotics, or operating on a patient to make the dying process more tolerable) hasten a patient's death.
2. Withholding life-sustaining medical treatment (either by not initiating treatment or by terminating it at some point) where this implies that the patient dies at an earlier date than he or she would if treatment were continued and where the intention is wholly or primarily to reduce the patient's suffering. This form of euthanasia is commonly referred to as 'passive euthanasia,' irrespective of whether withholding or withdrawing life-sustaining treatment is explicitly requested by the patient. The term 'passive' is often taken to suggest that this variant of euthanasia is legitimate (in contrast to 'active' forms of euthanasia), which, however, is a misunderstanding. 'Passive' is in itself value-neutral and refers to nothing but the fact that the death of the patient is primarily caused by his or her bad state of health and not by any kind of intervention.
3. Actively hastening death to shorten suffering, either by killing the patient ('active euthanasia'), by assisting the patient to kill himself ('assisted suicide'), or by putting the patient in a prolonged state of unconsciousness (e.g., by giving him or her a strong dose of narcotics) ('palliative sedation'), with the intention to withdraw life-sustaining treatment while the patient is in this state.

Interest in euthanasia of these various types has grown in proportion to two historical developments: the growth of the resources of modern medicine and its increased capacity to maintain life functions in states of terminal disease; and the growing interest of patients to decide themselves about the way they want to die, especially where physicians tend to preserve life in situations of imminent death at the cost of severe suffering.

Political impulses toward making euthanasia a patient right continue to come primarily from the euthanasia and 'right-to-die' societies that have developed all over the world (the British 'Voluntary Euthanasia Legalization Society' was founded in 1935, the 'Euthanasia Society of America' in 1938). They vary, however, in the forms of euthanasia they want to see legalized by law and professional codes. All societies advocate the extension and improvement of psychological and palliative care (including, for the most part, palliative sedation) for the dying as well as termination of treatment where this is explicitly requested by the patient (cf the manifesto of the World Federation of Right-to-Die Societies, 2013). The differences between them concern the extent to which they advocate voluntary active euthanasia (as well as the related forms of assisted suicide). While for the Hospice movement (Yennurajalingam and Bruera, 2011) it has become partly a matter of principle to reject active euthanasia (and in many cases also physician-assisted suicide), most euthanasia societies demand the legalization both of (physician-)assisted suicide and active euthanasia under what might be called the "paradigmatic euthanasia conditions" (Birnbacher, 2004): (1) the patient is in an irreversible state of disease or impairment of functioning, (2) the patient suffers intolerably, (3) the patient explicitly wishes to die, (4) this wish is not only momentary but persistent. Concerning (1), some societies restrict the definition of disease to clear cases of somatic disease such as end-stage cancer, whereas others, such as the Swiss society 'Exit,' include also psychiatric diseases such as major depression and schizophrenia. A number of societies, however, favor assisted suicide over voluntary active euthanasia as a more unambiguous expression of the patient's autonomous will. At present, an officially approved and legal practice of active euthanasia exists in the Netherlands, Belgium, and Luxembourg, an officially approved and legal practice of physician-assisted suicide in Switzerland and in the states of Oregon, Washington, Montana and Vermont, USA.

The common element of the variants (1) to (3) of euthanasia is that the primary and ultimate aim of giving euthanasia is the relief of suffering, either from physical pain or from other inconveniences of the patient's condition. Shortening the life of the patient is either intended as a means to that end (active euthanasia, passive euthanasia, assisted suicide) or accepted as a side effect (indirect euthanasia). This common element is lacking in practices (4) and (5).

4. Withholding life-sustaining treatment from a patient in a persistent state of unconsciousness (as in the so-called *coma vigile*) with the resulting (and mostly intended) death of the patient. In cases of this sort, the intention cannot be to relieve the patient's suffering because an unconscious patient cannot experience suffering. (There is, however, an unclarity on this point. In some cases, a residual form of consciousness, the so-called minimal conscious state can be diagnosed from the patient's behavior and brain scans.)

Euthanasia in these cases is generally justified either by the patient's presumed will not be kept alive in a persistent state of unconsciousness (as documented, e.g., by an advance directive), by the insistence of the family distressed by a treatment they often see as degrading and contrary to human dignity, or, in advanced stages, by considerations of medical futility (studies have shown that the probability of an 'awakening' from a so-called 'vegetative state' is about 1% after 1 year). Relevant jurisdiction in the US and the UK shows a tendency to allow the withdrawal of life-sustaining treatment after several years of 'persistent vegetative state' given the approval of the family.

5. Ending the lives of adults and children seen as a burden to society or to parents. In German-speaking countries, the German equivalent to the term 'euthanasia' inevitably evokes the National Socialist practice of killing children with malformations and the inmates of psychiatric hospitals and is normally replaced, in reference to practices (1)–(4), by the term *Sterbehilfe* (aid in dying). Moreover, memories of the NS-'Euthanasie' explain at least part of the violent resistance, especially in German-speaking countries, to philosophical justifications of infanticide (such as by Peter Singer and Helga Kuhse) (Kuhse and Singer, 1985; Singer, 2011: Chapter 6). The motives behind the killing, with the cooperation of their physicians, of 70 000 patients of psychiatric institutions in Germany from 1940 to 1941 were partly economic (one criterion of selection was whether patients were able to work), partly an ideology of racial purity. After stopping the active killing in 1941, a further 90 000 patients were exterminated by 'softer' methods such as malnutrition (Benzenhöfer, 2009: Chapter VII, 3). Calling these killings 'euthanasia' was a deliberate euphemism, as was the term *lebensunwert* (unworthy of living) used by the authors Binding and Hoche in a seminal book as early as 1920 and suggesting a poor subjective quality of life (the contrary of 'a life worth living') but in fact expressing an objective judgment about the usefulness of a life to society. Ideas of a selective killing of 'useless' lives have their roots in the Social Darwinism of the late nineteenth century, though there are important earlier philosophical models. A rigid eugenic policy involving compulsory abortion and exposure of 'weak' children (probably on the Spartan model) were part of Plato's model state in the *Republic*. A curious blending of euthanasia in the sense of relieving suffering and euthanasia in the sense of getting rid of people seen as a burden occurs in Thomas Morus' *Utopia* (1516), where priests and state officials persuade the terminally ill to kill themselves rather than to 'survive their own death.'

Ethical Problems

While euthanasia in sense (1) is almost universally accepted and euthanasia in sense (5) almost universally rejected by medical ethics and medical law, ethical and legal discussion of euthanasia in each of the other meanings is highly controversial both on the national and the international level. Moreover, the discussion often shows a particular emotional quality. There are at least four controversial issues:

1. How to resolve the conflict between the *principle of patient autonomy* and the *best interest principle* in euthanasia situations. According to the principle of patient autonomy, medical treatment (except in emergency situations) must be based on the informed consent of the patient. If the patient is unconscious or otherwise unable to give valid consent, physicians are required to take into account the presumed consent of the patient, as indicated by written or oral advance directives, by the views expressed by the patient about how he or she would like to be treated in relevant situations and what else can be inferred from the patient's beliefs and worldview. At the same time, the physician is also bound by the best interest principle, which says that he or she should further the well-being of the patient in the best possible way. This is generally taken to entail both prolonging the life of the patient as well as an obligation to relieve suffering even without an explicit request. Generally, this principle is taken to justify shortening the life of a patient with terminal illness wherever there is intense suffering even in the absence of a capacity to give valid consent, provided that the intention of the physician is exclusively or primarily directed at relieving suffering.

Problems arise whenever the two principles come into conflict, as when a patient refuses an operation that, in all probability, would save his or her life at an acceptable level of subjective quality of life, whereas following his or her refusal would lead to instant death. There is considerable consensus among medical ethicists that a medical paternalism (i.e., treating the patient against his or her will, see Häyry, 1991) is objectionable even in cases of markedly 'unreasonable' patient wishes except when there is evidence that the patient is not sufficiently informed about his or her prospects, that his or her will is unstable, distorted by neurosis, psychosis, or excessive affect, or is alterable by additional information and argument. Though, on the best interest principle, the physician is called upon to persuade the patient to revise his or her decision, it is widely agreed that the physician should not resort to biased information, persuasion or even deception to change the patient's will, since this would undermine trust and infringe patient autonomy no less than compulsory treatment.

2. Whether actively discontinuing treatment (e.g., by turning the switch of a respirator) is normatively equivalent to strict inaction (e.g., by 'freezing' oxygen pressure at a level that is insufficient for the patient to survive). Ethicists and lawyers predominantly take both modes of action to be normatively equivalent. They see no morally significant difference between 'active' and 'passive' discontinuing nor between not starting treatment and terminating treatment. (Discontinuing treatment with the result or the intention of a patient's death in a euthanasia situation generally counts as 'passive' even if involving positive action because the state of the patient is merely restored to the condition in which it was before treatment.) There is, however, a considerable psychological difference. As a rule, physicians and nurses find it easier to let a patient die 'passively' by strict inaction. The reason is that, in this case, causality is attributed primarily to factors other than one's own behavior. This is markedly different whenever the death of

the patient follows more or less immediately on an active termination of treatment.

3. Whether active killing should be allowed to doctors (or to others) in euthanasia situations. Voluntary active euthanasia, i.e., actively killing a patient to end his or her suffering on the patient's explicit wish is an attractive option because of substantial doubts whether all states of suffering in the last phase of a patient's life (e.g., in some forms of end-stage cancer) can be effectively relieved by discontinuing treatment and/or palliative measures. Furthermore, many of the ethical arguments against voluntary active euthanasia on the level of principles suffer from the fact that they attribute much more normative weight to the distinction between activity (killing) and passivity (letting die) than is thought acceptable, and indeed desirable, in other medical contexts. In these contexts, the physician usually has not only a duty not to harm but also a duty to treat. In particular, the question arises how the acceptance of indirect euthanasia (which also involves active intervention) by most legislatures can be made consistent with the rejection of voluntary active euthanasia. The crucial difference seems to be that death is accepted as a side effect in the former case, whereas it is not held to be acceptable as a means to reducing suffering in the latter case. This difference has been regarded as highly relevant by the so-called doctrine of 'double effect' rooted in Catholic moral theology, which puts much ethical weight on intentions and less on consequences. This principle survives, as it were, in the legal treatment of euthanasia, whereas it is not generally accepted by most legal systems in other contexts.

More compelling arguments against legalizing voluntary active euthanasia are pragmatic considerations of potential misuse and abuse. Establishing a practice of active euthanasia risks that the autonomy of the patient is overruled by the wishes of relatives, the standards of the medical profession, and the interests of society. Due to a program of continuous survey research about actual euthanasia practice in the Netherlands, it was found out that in the Netherlands that along with a publicly approved practice of voluntary euthanasia (2.2% of all deaths in 2001, 2.8% in 2010), there has been an undeclared practice of nonvoluntary euthanasia that deliberately disregarded the procedural rules laid down by the Royal Dutch Medical Association. Though far from proving that a practice of voluntary euthanasia necessarily has a practice of nonvoluntary euthanasia in its wake, the opponents of the legalization of active euthanasia have a point in quoting the "Dutch experience" as a warning (see, e.g., Thomasma and Marshall, 1998). It is true, a majority of medical ethicists think that palliative sedation and physician-assisted suicide are, wherever possible, preferable as alternatives to active euthanasia, the former because it is interpreted as a purely palliative measure and the latter because it goes further than active euthanasia in asserting the autonomy of the patient (Quill et al., 1997; Orentlicher, 1998). However, since palliative sedation is frequently applied together with the withdrawal of artificial nutrition and hydration (e.g., in the Netherlands, cf Centraal Bureau, 2003: p. 67), it often goes together with the intention to hasten the death of the patient no less than the more controversial forms of euthanasia. On the other hand, in the Netherlands active euthanasia is often preferred over assisted suicide because of physical weakness or incapacity of patients (Onwuteaka-Philipsen et al., 2003: p. 4).

Social Science Research on Euthanasia

Social science studies about the practice of passive and active euthanasia have begun to be undertaken only recently, partly after the model of the Dutch Remmelink report, which was done in 1990–91 at the request of the Dutch government in preparation of euthanasia legislation. This nationwide study included interviews with 405 physicians (carried out by 30 experienced physicians and five other academics) and a questionnaire study with the physicians of a sample of 7000 recently deceased persons. The main result of this study was a surprisingly high incidence of euthanasia in the Netherlands prior to regulation. Fifty-four percent of the physicians interviewed confirmed that they had practiced active euthanasia or assisted suicide on the request of patients and a further 34% said that they could conceive of situations in which they would be prepared to do so (van der Maas et al., 1991: p. 669). Twenty-seven percent said that at some time they had ended a patient's life without his or her explicit request and a further 32% said that they would be prepared to do so under appropriate circumstances. The study was repeated every 5 years showing that the situation in the Netherlands has not substantially changed up to now (van der Maas et al., 1996; Onwuteaka-Philipsen et al., 2003).

Substantially lower figures were found in an Australian and a British questionnaire study in which 29% (12%) of all doctors who returned a completed questionnaire confirmed that they had taken active steps to end a patient's life. A large proportion of the British respondents (46%) was reported to say that they would consider taking active steps to bring about death if it was legal to do so (Kuhse and Singer, 1988: p. 623, Ward and Tate, 1994: p. 1332). A metastudy of all relevant surveys of the period 1990–2010 showed, however, that this estimate was exaggerated and that the value is no larger than about 25% (cf McCormack et al., 2012: p. 26).

Legislatures and medical associations continue to differ greatly in the practices they legalize, tolerate, or prohibit (Heide et al., 2003; Birnbacher and Dahl, 2008). A striking example is that of the neighboring and culturally highly similar countries the Netherlands and Germany. Differently from the Netherlands, in Germany active euthanasia is still largely a political taboo in spite of a high approval (about 70%) by the general population. The only questionnaire study done so far showed that the proportion of physicians practicing active euthanasia was maximally 5% (Kirschner and Elkeles, 1998: p. 251). In contrast to voluntary active euthanasia, physician-assisted suicide is not a criminal offense in Germany but declared unethical and correspondingly sanctioned by the medical associations. According to a recent survey, however, more than one-third of German physicians can think of conditions under which they would consider assisting a patient suicide. The situation is, again, different in Switzerland where voluntary active euthanasia is legally prohibited but where assisting suicide is legally prohibited only in cases of self-serving motives

and is generously tolerated by the medical associations and by state authorities. The resulting contrast in legislation between Switzerland and neighboring countries has led to the new phenomenon of 'euthanasia tourism,' especially with the establishment of organizations in Switzerland that, differently from older Swiss organizations such as Exit, offer their services to patients from abroad.

See also: Bioethics; Death and Dying, Psychology of; Death and Dying, Sociology of; Death, Anthropology of; Medical Profession.

Bibliography

Bacon, F., 1857. Of the dignity and advancement of learning, books II–IV. In: Spedding, J. (Ed.), The Works of Francis Bacon. Longman, London.

Benzenhöfer, U., 2009. Der gute Tod? Geschichte der Euthanasie und Sterbehilfe. Vandenhoeck & Ruprecht, Göttingen, Germany.

Birnbacher, D., 2004. Terminale Sedierung, Sterbehilfe und kausale Rollen. Ethik in der Medizin 16 (2004), 358–368.

Birnbacher, D., Dahl, E. (Eds.), 2008. Giving Death a Helping Hand. Physician-Assisted Suicide and Public Policy. An International Perspective. Springer, Dordrecht, The Netherlands.

Centraal Bureau voor de Statistiek, 2003. Het levenseinde in de medische praktijk. Resultaten sterfgevallenonderzoek 2001. Voorburg/Heerlen, The Netherlands.

Chabot, B.E., 2012. Hastening death through voluntary cessation of eating and drinking. In: Youngner, S.J., Kimsma, G.K. (Eds.), Physician-Assisted Death in Perspective. Assessing the Dutch Experience. Cambridge University Press, Cambridge, UK.

Council of Europe, 2004. Euthanasia, Vol. II: National and European Perspectives.

Häyry, H., 1991. The Limits of Medical Paternalism. Routledge, London, UK.

Hastings Center, 1987. Guidelines on Termination of Life-Sustaining Treatment and the Care of the Dying. Indiana University Press, Bloomington, IN.

Heide, A., et al., 2003. End-of-life decision-making in six European countries. A descriptive study. Lancet 362, 345–350.

Kirschner, R., Elkeles, T., 1998. Ärztliche Handlungsmuster und Einstellungen zur Sterbehilfe in Deutschland. Gesundheitswesen 60, 247–253.

Kuhse, H., Singer, P., 1985. Should the Baby Live? The Problem of Handicapped Infants. Oxford University Press, Oxford, UK.

Kuhse, H., Singer, P., 1988. Doctors' practices and attitudes regarding voluntary euthanasia. Medical Journal of Australia 148, 623–627.

van der Maas, P.J., van Delden, J., Pijnenborg, L., Looman, C.W.N., 1991. Euthanasia and other medical decisions concerning the end of life. Lancet 338, 669–674.

van der Maas, P.J., et al., 1996. Euthanasia, physician-assisted suicide, and other medical practices involving the end of life in the Netherlands, 1990–1995. New England Journal of Medicine 335, 1699–1705.

McCormack, R., Clifford, M., Conroy, M., 2012. Attitudes of UK doctors towards euthanasia and physician-assisted suicide: a systematic literature review. Palliative Medicine 26, 23–33.

Onwuteaka-Philipsen, B.D., et al., 2003. Euthanasia and other end-of-life decisions in the Netherlands in 1990, 1995, and 2001. Lancet 364, 1–5.

Orentlicher, D., 1998. The Supreme Court and terminal sedation. An ethically inferior alternative to physician-assisted suicide. In: Battin, M.P., Rhodes, R., Silvers, A. (Eds.), Physician-Assisted Suicide. Expanding the Debate. Oxford University Press, New York.

Quill, T.E., Lo, B., Brock, D.W., 1997. Palliative options of last resort. Journal of the American Medical Association 278, 2099–2104.

Singer, P., 2011. Practical Ethics, third ed. Cambridge University Press, Cambridge, UK.

Thomasma, D.C., Marshall, P.A. (Eds.), 1998. Asking to Die: Inside the Dutch Debate about Euthanasia. Kluwer, Dordrecht, The Netherlands.

Ward, B.J., Tate, P.A., 1994. Attitudes among NHS doctors to requests for euthanasia. British Medical Journal 308, 1332–1334.

World Federation of Right-to-Die Societies, 2013. Manifesto. http://www.worldrtd.net/manifesto.

Yennurajalingam, S., Bruera, E. (Eds.), 2011. Oxford American Handbook of Hospice and Palliative Medicine. Oxford University Press, New York.

Evangelical, Revivalist, and Pentecostal Religions

Mark Hutchinson, Southern Cross College, Chester Hill, NSW, Australia

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 1752–1756, © 2001, Elsevier Ltd.

Abstract

This article points out the connections between three of the great themes of modernity – the interplay between Protestant Christianity and the universalizing tendencies of modern culture and modernist revivalism. Not only did the Reformation result from and encourage the rise of capitalism, it provided cohering worldview ideologies and organizational principles which contributed to modern science, European colonialism, and expansion by migration, revival, missions, the formation of imperial professionalism, and a global vision which relativized national identities. It is of continuing interest to sociologists not just because of the remarkable global spread of Pentecostalism, but because of its implications for secularization theory.

While 'globalization' as a concept arose as a fiscal planning tool in the 1950s, its extension as an explanatory concept for the sort of territory previously covered by such frameworks as 'world systems theory' dates from the 1970s. At the time of writing, Roland Robertson and Peter Beyer are the key writers in the area as it relates to religion. Indeed, Robertson's pioneering framework arises out of the conundrum which religion presents for standard modernization theory. The rise of Islamic and other forms of religious fundamentalism denies the linear rationalist assumptions of modernization theory and such offspring as secularization theory, which suggest that regional difference and religious conviction should be passing away in the face of rational technical systems and the homogenization of culture. Robertson has since then further articulated the theory through the concept of 'globalization,' which allows for simultaneous global homogenization of culture and local reassertions of difference. Robertson's theory that cultural forces move faster than political and economic forces has received considerable support from historians and analysts of Christian missions, which have noted that the indigenization of Christianity has historically been faster and deeper in Asia and Africa than the movement of formal missionary and colonial structures. In South America, by contrast, there is a body of opinion which states that the combination of state apparatus and church extension actually retarded the extension of Christianity. It is clear, then, that globalization theory is vital for the understanding of the expansion of Christianity generally, just as the study of the expansion of Christianity is vital for understanding the processes of globalization.

While a slow starter to get into missions, evangelical Christianity began to articulate an increasingly sophisticated global vision from the late 1600s, with the rise of the Pietist movement. As Lewis has noted, Pietism restored a sense of personal call to Protestant orthodoxy which, through agencies such as the University of Halle and Moravian missions, expressed itself in a growing internationalism. The correspondence of Zinzendorf, the migration of Moravians, and missionary outreach to Russia, Greenland, and India, created networks and examples of successful missions which informed the expanding networks of new settler societies and expanding capitalism. It was on a trip to the New World that John Wesley, for instance, first ran into Moravians, who, through the influence of Peter Böhler, provided the synthesizing principle to Wesleyan eclectic theology and practice which in turn made it a vibrant missionary force. It was likewise among Scots-Irish and Dutch settlers in the New World that religious revival produced the combination of Protestant theology and religious experience which has become known as 'Evangelical Christianity.' Definitions differ, usually with personal affiliation, but a standard definition of Evangelicalism offered by David Bebbington locates the movement in a quadrilateral of theology and praxis involving activism, cross-centeredness, Bible-centeredness, and conversionism. The fact that nothing like the French Revolution occurred in England helped legitimize the movement as a source of social and cultural stability. The inner energies of the movement, combining a universalizing crucicentrism and biblicism, with conversion-oriented activism, meant that the tens of thousands of converts to the Evangelical Awakening on both sides of the Atlantic provided a vast pool of willing support to the evangelical program to change globalizing British society. The leaders of the movement – in particular William Wilberforce, Lord Shaftesbury, the Clapham sect, and Selina, Countess of Huntingdon – were able to draw upon these networks of support in a dynamic program of social activism over the century 1750–1850. This activism arguably provided important impulse force to the processes of globalization, just as globalizing Western economies provided important resources and opportunities for evangelical expansion and change.

Protestants from Luther's time were faced with an expanding globe in tension with an increasingly outmoded ecclesiology. While, during Luther's own lifetime, Diaz rounded the Cape of Good Hope (1488), Columbus was driven west in pursuit of a millennial vision to land on the shores of America (1492), and other parts of the world were opened to Western influence by da Gama, Cabot, Cortes, Magellan, and Pizarro, Protestants were combining with increasingly nationalistic ruling classes (the nobility in Germany, bourgeoisie in Switzerland and The Netherlands) to reinstitute the 'Christendom' ideal within the national churches of rising nation states. While Catholic empires ruled the seas, Protestant missions were restricted to Europe. While the theology became increasingly arid and formalized, the impact of these churches was to reinforce the nexus between nationalism and capitalism.

Capitalism expanded during the period through royal chartered companies, combining national expansion and capitalist enterprise. Halle Pietism specifically used this link to expand into Russia, just as Puritans used the chartered company to establish themselves in Massachusetts Bay. The CMS first fought with and later used the East India Company in Asia, while the Claphamites moved into West Africa on the tails of colonization companies. While Weber's 'spirit of Protestantism' has been revised and attacked in some quarters, it is nonetheless a useful heuristic when viewing the interaction of shipping, chartered, and colonization companies on the one hand, and missionary agencies on the other.

Protestantism provided an 'inner fascination with the world,' and evangelicalism adopted the same fascination and extended it. There has been considerable work done, for instance, on the link between enlightenment and Evangelicalism, indicating that Evangelicalism, through its commercial and missionary emphases, was a major carrier of rationalizing Western concepts of time and space into the rest of the world during the late nineteenth and early twentieth centuries. Not only did Wesley proceed on the assumption of an 'inner enlightenment,' using both the vocabulary and technologies of enlightenment to further the evangelical cause (he was, for instance, an assiduous publisher), but Puritan and evangelical input into the sciences from the sixteenth to the nineteenth centuries was immense. Clergy were not only involved in agencies such as the Royal Society, but were key suppliers of information to enlightenment projects such as the social sciences. (Malcolm Prentis, for example, has demonstrated the link between clerical collection and the construction of an international culture of research in anthropology.)

The logical place to look for the globalizing influence of evangelicalism is through the agency of missions. The marginality of missions at the beginning of the nineteenth century demonstrates the close association of evangelical expansion with Western expansion. By 1800, North America was subject to European settlement, but only on the eastern half and the western rim of the continent, and though there was no real challenge to Hispanic hegemony in South America, Christianization had not really pierced deeply below the surface of the indigenous and emerging plantation cultures. The impact of missions in the Indic, wider Asian, and Muslim worlds was negligible. Obviously, then, developments in the nineteenth century were critical for the development of global Christianity in our own period. Europe at the time was weak and divided religiously. The industrial revolution encouraged both concentration of the sort of wealth which could capture foreign markets and provided the scientific curiosity and the speed of travel to take advantage of new opportunities. While most European expansion in the eighteenth century had been mercantilist in nature – with small colonies of isolated Europeans trading with Asian nations in sealed cantons such as Shanghai, Macau, and Hong Kong – the nineteenth century saw a shift toward extensive settlement and active domination of local politics. At least in part, this was because it was now technically possible, with advances in communication and military hardware, to dominate other less advanced cultures. So Britain absorbed India, Burma, and Ceylon into a new Empire, France annexed Indo-China, and The Netherlands finalized its annexation of Indonesia. Other European powers founded colonies, such as the German colonies in PNG, and annexed parts of Africa, such as Togoland, Southwestern Africa, and the like. At least for the British, Stanley argues that there was a consistent imperial impulse to British society, but no plan of conquest, most acquisitions having been gained in order to protect trading rights. Formal control was never exercised when the much cheaper form of informal control could be used, and it was the increasing need for formal control, driven by the intensified competition of European powers for global resources, which drove Britain to acquire and then defend India as the keystone in its defense of the East.

In terms of the other global religious cultures, evangelicalism also faced weakened opposition. One advantage of the time was that Islam was in retreat – in 1821 Greece was retaken, and from the 1830s European powers began dividing up Islamic parts of Northern Africa such as Morocco, Algeria, and Tunisia. Roman Catholicism at the beginning of the nineteenth century was under siege from the forces of modernization. With the eclipse of Portugal and Spain as imperial powers, increasingly the push for missions in the Catholic world came from France, Belgium, and (significantly for Australia) Ireland. While the Orthodox churches did not pursue much in the way of missions, despite the spread of the Russian Empire during this period, the Protestants, energized by the Evangelical revival, certainly did, both within and outside the English speaking world. Stanley suggests that, while there were agencies such as the Society for the Propagation of the Gospel with a missionary element to them, before the Evangelical awakening there was no consistent acceptance of the burden of foreign missions by either Anglican or Dissenting Protestants. Europe was the same. In 1810, for instance, a revival known as the Reveil spread out of Geneva and affected the French Protestant world deeply, leading to considerable missionary endeavor overseas. Norwegian revivals, most notably commenced by Hans Nielsen Hauge, a lay reader preaching in a way which was not legalized until 20 years after his death, energized Scandinavian missions to North America and later to South America. German missionaries, in the tradition of Halle and the Moravians, traveled all over the world. The interaction of these revival movements and missions can be seen in their common adherence to interdenominational agencies such as the Evangelical Alliance, which was founded in 1847, and held conferences in France, Belgium, Switzerland, Germany, Canada, Sweden, and the United States.

The model provided by the Alliance of a voluntary society was to become the classic form for Protestant missions. Many were formed at the prompting of chaplains and clergy traveling with army, trading company or political representatives. The first of them was the Baptist Missionary Society (1792), inspired by William Carey, who in turn had been inspired to mission by the accounts of the travels of such explorers as Captain James Cook in the Pacific. It was followed by the London Missionary Society (1795), the British and Foreign Bible Society (1804), the Wesleyan Methodist Missionary Society (1813), and many others. Because the Church of Scotland refused the requests of its Evangelical party to form their own missionary agency, most Scots siphoned off into other agencies, such as the LMS. In 1810, preceded by many small bodies dedicated to evangelizing First Nations' peoples,

the first American foreign agency was founded in the American Board of Commissioners for Foreign Missions.

The extension of evangelical missions had a number of major interactions with the growth of global society. The first was that, with increased need for personnel, the missions started their own Bible training institutions. These in turn legitimized the long-standing 'old dissenting' academies, supported religious emancipation and social reform, and acted to break down the exclusivism of the established church-linked universities. European states, at slower or faster rates, responded to this broadening of the religious franchise (among other secularizing tendencies) by stepping back from the official sanctioning of one form of religious truth or another. By the end of the century, most European states had become generally 'Christian' countries without confessional attachments, and many, in their colonies (particularly in Canada and Australia) were actively encouraging liberal democratic states. Not only did missions change the education of the increasingly vital professional class at home, but in exporting the Western educational model, also exported Western professional values into the nation states which began to arise in Asia, Africa, and the Pacific from the late nineteenth century. In exporting values such as professional disinterestedness, the importance of technical expertise, and basic human rights, a common culture was established which facilitated not only the spread of global society, but also (through preservation and re-skilling of Fourth World peoples) the means for local opposition to global homogenization. The specter of evangelical African bishops nay-saying advanced liberal proposals at a Lambeth conference, or Aboriginal evangelicals controlling the balance of power in the national assemblies of the United churches of Canada and Australia, have demonstrated the interactivity between evangelical missions and mediation of global culture.

The establishment of comity arrangements between evangelical missions in foreign countries highlights the second major interaction. The movement of religious traditions outside their national boundaries produced a relativization of their content and form, creating movements toward ecumenism based on 'common gospel' assumptions, diminishing the importance of liturgical and cultural tradition, and homogenizing the culture of international Christianity. This was imported back into the West as de-denominalization, and into Third World countries as 'alliance Christianity,' enabling Christianity to be described as a unitary object in contact with other world religions. This sense of relativization of form sparked concerns about relativization of truth, concerns expressed most markedly in the fundamentalism/modernism debates of the 1920s onwards.

Fundamentalism self-consciously identifies itself as a countermodernist movement, particularly in terms of theological modernism. Evangelicalism maintains something of this thrust, at the end of the twentieth century increasingly identifying itself as a counter-postmodernist movement. On another level, however, these debates can be seen as localized resistance to homogenizing forces at the global level, resistance seen most markedly in the general opposition of national evangelicals to agencies such as the World Council of Churches. In some cases, this meant the localization of evangelicalism into extreme nationalist forms. In most cases, however, especially after the rise of the National Association of Evangelicals in the USA and the Billy Graham Organization, the centrality of missions and a self-identity partly defined by contention for the faith caused evangelicals after World War II to take the same globalizing path that the WCC had already taken. This is reflected in the formation of agencies such as the Lausanne Committee, the World Evangelical Fellowship, and worldwide denominational fellowships such as the World Alliance of Reformed Churches. Particularly after World War II, these agencies provided the basis for expansion of English and, with increasing hegemony, American evangelicalism onto the world stage. By the 1990s, it was clear that denominations had become almost irrelevant to the self-definition of most evangelicals, and that localized megachurches were more important for agenda setting.

Likewise, evangelical missions had shifted from being denominationally and agency based to increasing integration with global movements in migration, occupation, and tourism. Short-term mission, and mission originating from countries in the 'South,' contribute nearly as much to the total Christian missionary workforce as do traditional, First World, agency based missions, and is likely soon to outstrip it.

Within these tendencies in evangelicalism in the nineteenth century arose the issue of 'revivalism.' Revivalism is a technical accommodation between the experience of revival, as formalized during and after the First Great Awakening by Calvinists such as Jonathan Edwards, and the utilitarianism implicit in early nineteenth century postenlightenment thought. The emphasis of Edwards and Wesley in defending the use of 'means' for the furtherance of religious revival were codified and extended during frontier revivals in New York State and the expanding American frontier. Not least among the codifiers was Charles Grandison Finney, whose *Lectures on Revival* became the great 'how to' book of the nineteenth century. Revivalism as it expanded from the USA to England, and then around the world (by way of church and migrant diasporas) developed an array of techniques to maintain the association between Christian mass evangelism and spiritual response. It became the tool of choice for many evangelicals as Western societies seemed to be slipping away from their commitments to national religious frameworks, a set of ideas crystallized in the premillennial theology of archetypal evangelistic movements such as the Brethren, Baptists, and Methodists. As a technical accommodation to utilitarianism, revivalism was a natural client and promoter of globalization, which was very largely fueled by the technical innovation of high modernist capitalism. So revivalism quickly spread by the development of Methodist itineracy into the big campaign evangelism of D.L. Moody, Alexander Somerville, Gypsy Smith, and the like. Such itineracy became global, as shipping, then automobiles, and then airplanes, brought other parts of the globe within reach. As communications technologies also improved – through telegraphed newspaper reports, international postage, radioed accounts, and finally television coverage (culminating in the global campaign by satellite of Billy Graham in 1996) – evangelicalism spread with revivalism to all parts of the globe. It is important to note in this regard the high degree of evangelical participation in modern technological enterprises, from Back to the Bible Radio, to the massive media empire of the Universal Church of the Kingdom of God in Brazil. This has been seen as a paradox by some commentators – it is less obviously so when revivalism's accommodation to

technological utilitarianism is taken into account. Though revivalism is not limited to evangelicalism (Catholic revivalism, for instance, or the adoption of revivalistic techniques by Hindu and Buddhist promoters in South and South-east Asia), and not all evangelicalism is revivalistic, it is clear that the two are blood relations and natural partners in many cultures around the world. At the same time, the fact that revivalism is a distillation of 'revival tradition' into a technological form which can thereby cross cultural boundaries makes it eminently transportable. For this reason, Edith Blumhofer has called the major global form of modern revivalism, pentecostalism, 'world evangelicalism,' because of its facility with communications technologies. The ability of pentecostalism to 'cast' itself as a communal form of thought through the abstracting processes of print media and CITs makes it a prime globalizing force. On the one hand, this makes for global expansion of evangelicalism through its revivalist and Pentecostal forms. On the other hand, the same process homogenizes its distinctives and causes it to adapt new local distinctives as it moves across cultures. The lack of a single center (unlike either Islam or Catholicism) is thus a great facilitator of growth, but is also the prime cause of internal dissension and fragmentation as various local evangelical cultures seek to adopt or invent normative centers which enable local communities and protect them from the effects of fragmentation. Neo-fundamentalism, Calvinist renewal movements in the Southern Baptist Convention, the Toronto and Brownsville revivals – each of these may be seen as the creation of new centers, which feed from the energies of globalization, and attempt to create islands of order amidst the chaos of global society.

In particular, evangelicalism as it stands at the beginning of the twenty-first century is clearly largely a two-thirds world phenomenon. As Piggin, Reed, and others have shown, revival is the inculturation of Christianity into local cultures. It is not too much to say, therefore, that the theology of revival, particularly as evidenced in Pentecostal revivalism since the 1950s, is a mainstream legitimating structure for the indigenization of Protestant evangelicalism. This is particularly evident in revivalistic outbreaks such as the East Africa Revival in the 1930s, the blossoming of Pentecostalism in Brazil and later in Argentina, and the Aboriginal revival commencing at Galiwin'ku in Australia's Northern Territory from 1978. In each case, evangelicalism struck new roots and found new accommodations with existing religious cultures which have in turn been exported into the global religious network.

While the East Africa revival had comparatively little experience of charismatic manifestations, Brazil and the Aboriginal revival have been highly charismatic in their combination of local and general characteristics. As Freston notes, the massive expansion of evangelicalism in that showcase of evangelical inculturation, Brazil, is likewise largely Pentecostal in nature. As he noted in 1998, a "survey of evangelical institutions in Greater Rio de Janeiro discovered that, of the 52 largest denominations, 37 were of Brazilian origin, virtually all Pentecostal. While only 61% of all evangelical churches were Pentecostal, 91% of those founded in the previous 3 years were" (Freston, 1998: p. 74). Of the 450 million 'evangelicals' and 750 million 'Great Commission Christians' estimated by David Barrett to exist in 1990, some 350 million could be counted within the Pentecostal or charismatic camp. It is clear that at the end of the twentieth century, evangelicalism's expanding edge was clearly Pentecostal/charismatic in nature, leaving the national traditions of the First World heartlands significant challenges of reinvention in their struggle to survive the disappearance of their defining ethnic and national boundaries.

See also: Globalization and World Culture; Globalization, Anthropology of; Historical Thought and Historiography: Christian Tradition; Religion, Sociology of; Religion: Evolution and Development.

Bibliography

Bebbington, D.W., 1992. Evangelicalism in Modern Britain: A History from the 1730s to the 1980s. Baker Book House, Grand Rapids, MI.

Beyer, P., 1994. Religion and Globalization. Sage Publications, Thousand Oaks, CA.

Blumhofer, E.L., 1993. Restoring the Faith: The Assemblies of God, Pentecostalism, and American Culture. University of Illinois Press, Urbana, IL.

Blumhofer, E.L., Spittler, R.P., Wacker, G.A. (Eds.), 1999. Pentecostal Currents in American Protestantism. University of Illinois Press, Urbana, IL.

Carwardine, R., 1978. Transatlantic Revivalism: Popular Evangelicalism in Britain and America, 1790–1865. Greenwood Press, Westport, CT.

Featherstone, M., Lash, S., Robertson, R. (Eds.), 1995. Global Modernities. Sage Publications, Thousand Oaks, CA.

Freston, P., 1998. Latin American perspectives. In: Hutchinson, M., Kalu, O. (Eds.), A Global Faith: Essays on Evangelicalism and Globalization. CSAC, Sydney, pp. 69–88.

Marsden, G.M., 1980. Fundamentalism and American Culture: The Shaping of Twentieth Century Evangelicalism, 1870–1925. Oxford University Press, New York.

McGrath, A., 1995. Evangelicalism and the Future of Christianity. InterVarsity Press, Downers Grove, IL.

Noll, M.A., Bebbington, D.W., Rawlyk, G.A. (Eds.), 1994. Evangelicalism: Comparative Studies of Popular Protestantism in North America, the British Isles, and beyond, 1700–1990. Oxford University Press, New York.

Poewe, K. (Ed.), 1994. Charismatic Christianity as a Global Culture. University of South Carolina Press, Columbia, SC.

Pollak-Eltz, A., Salas, Y. (Eds.), 1998. El pentecostalismo en América Latina entre tradición y globalización. Ediciones Abya-Yala, Quito, Ecuador.

Robertson, R., 1992. Globalization: Social Theory and Global Culture. Sage, London.

Robertson, R., Garrett, W.R. (Eds.), 1991. Religion and Global Order. Paragon House, New York.

Smith, T.L., 1980. Revivalism and Social Reform: American Protestantism on the Eve of the Civil War. Johns Hopkins University Press, Baltimore, MD.

Synan, V., 1997. The Holiness-Pentecostal Tradition: Charismatic Movements in the Twentieth Century, second ed. W. B. Eerdmans, Grand Rapids, MI.

Thomas, G.M., 1989. Revivalism and Cultural Change: Christianity, Nation Building, and the Market in Nineteenth-Century United States. University of Chicago Press, Chicago.

Evans-Pritchard, Sir Edward E. (1902–73)

Wendy James, Oxford University, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Edward Evans-Pritchard was a leading scholar of the twentieth century, holding the Chair in Social Anthropology at Oxford from 1946 to 70. Originally trained in history, Evans-Pritchard became famous for his outstanding ethnographic works on the Azande and Nuer peoples of the Anglo-Egyptian Sudan. In raising classic questions about the nature of reason, political order, and religion across the gulf then imagined between 'civilized' and 'primitive' societies, these studies brought a new sophistication into anthropological theory. He inspired a generation of followers whose work helped spread an international recognition for social anthropology, as an academic subject also relevant to practical affairs.

Background, Education, and Career Outline

Edward Evans-Pritchard was born at Crowborough in Sussex on 21 September 1902, the son of an Anglican clergyman. He attended Winchester College, a leading independent school in the British public school tradition, from which he entered Exeter College at Oxford University in 1921. His family background was less conventional, however, than this account might indicate, as his father was from Caernarvon and spoke Welsh, while his mother's family came from Liverpool. The history syllabus he followed as an undergraduate was itself on the conventional side, even for the period, but Evans-Pritchard then sought a more independent academic path and went on to study anthropology at the London School of Economics and Social Science. Through Seligman he embarked on several periods of field research in the Anglo-Egyptian Sudan from 1926 to 40, and in the early years of the Second World War was able to fashion opportunities for anthropological work on the Ethiopian border, in Kenya and in Libya, in conjunction with his military service. It was during the period from the late 1930s to 1940s that other key events in his life took place. In 1939 he married Ioma Heaton-Nicholls, daughter of the South African high commissioner in London; and in the early 1940s he was received into the Roman Catholic Church. On returning to Britain at the end of the War, Evans-Pritchard was appointed in 1945 to a Readership in Cambridge; and then in 1946, to the Chair of Social Anthropology in Oxford. Settling here, he and his wife raised their lively family of five children; while he came to play a key role in British academic life. He became president of the Royal Anthropological Institute (1949–51), a founder (1946) and later first life president of the Association of Social Anthropologists of the Commonwealth, and a fellow of the British Academy (1956). He retired in 1970, and his outstanding life's contribution was recognized with a knighthood in 1971. He died at his home in Oxford, on 11 September 1973.

Early Influences

During his time as an undergraduate at Oxford, Evans-Pritchard found his obligatory studies of mainstream history, mainly about kings and battles, somewhat constricting. However, Oxford already had the fine ethnographic collections of the Pitt Rivers Museum, and a tradition of related lectures, initiated by Edward Tylor in the 1880s. There was also provision for research and teaching in physical anthropology. Exeter College was already well established as a focus for anthropological activities in a broad sense, as a result of the enthusiasms of R.R. Marett, a practicing archaeologist with a strong interest in early forms of religion and society. He had been appointed as university Reader in Social Anthropology in 1910, but even by the 1930s had only developed an embryonic 'department'. However, in 1928, Marett became rector of Exeter College, and was able to maintain a lively stream of visiting speakers. Even as a student, Evans-Pritchard would have had the chance to meet many archaeologists and anthropologists there. It is extremely likely that he also attended the very popular lectures given in Oxford by the young philosopher–archaeologist R.G. Collingwood, ranging from topics in the archaeology of Roman Britain to questions linking academic philosophy with 'real life' experience. Evans-Pritchard was a little later to review Collingwood's *Roman Britain*, warmly endorsing what we would call today its 'sociocultural' approach to the everyday life and practices of the ordinary folk.

It was after Evans-Pritchard had embarked on his researches in the Sudan that he first had his own experience of teaching; this was at King Fuad University in Cairo from 1932 to 34, and during this time he published a series of critical, theoretical papers, which foreshadowed some of his mature work. In 1935 he was appointed to a temporary research lectureship in African Sociology back in Oxford, to work with Marett from time to time but also to use Oxford as a base for continuing field research. As Marett's expected retirement date approached (1936) there was much discussion in Oxford about the future of anthropology. A special interest was taken in this matter by All Souls College, which actively supported the establishment of a Chair in Social Anthropology, rather than simply 'Anthropology', on the grounds that enlightened governance required knowledge of the peoples of the empire, and would benefit from scholarly attention to questions of their culture, welfare, and interests (Davis, 2007). After a good deal of internal politicking, the chair was duly established in 1936, in conjunction with a fellowship at All Souls. The first appointee was A.R. Radcliffe-Brown, well known along with Malinowski for his pioneering the importance of first-hand fieldwork for

anthropology; he was also already known for his vision of the systematic nature of the (somewhat legalistic) principles governing the forms of social structure in small-scale societies. He could not take up the chair immediately but arrived in 1937 and began to pursue the cause of 'social' anthropology with a will. His efforts were inevitably set back by the outbreak of the War.

Evans-Pritchard's Field Research: The Analytic Ethnographer

Drawn by the promise of working with experienced field anthropologists, Evans-Pritchard moved on in 1924 from Oxford to the London School of Economics and Social Science. C.G. Seligman was already teaching there, and Bronislaw Malinowski arrived the same year. Seligman, with his wife Brenda, had been engaged in anthropological work for the government of the Anglo-Egyptian Sudan since 1910. They were beginning to seek others to continue their project, and Evans-Pritchard seized his opportunity. After completing a required survey in the southern Blue Nile region in 1926, he was allowed a free hand in embarking on his own project to study the Azande kingdoms of the equatorial forests bordering the Congo basin in the southwest. This field research, together with a series of visits between 1930 and 1936 to the very different region of the upper Nile flood plains inhabited by the Nuer, whose reputation for being a 'difficult' and elusive people far outweighed anything else known about them, laid the foundation for Evans-Pritchard's later career and his shaping of the discipline of social anthropology. A steady output of ethnographic articles mainly on the Azande and the Nuer also appeared through the 1930s, and indeed in the case of the Azande continued to appear throughout his life (a bibliography is provided in Beidelman, 1974). But it was the two remarkable monographs appearing at the end of the decade that confirmed Evans-Pritchard's outstanding contribution to the development of thematic, analytical anthropology. They were *Witchcraft, Oracles, and Magic among the Azande* (1937) and *The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People* (1940), both published by the Clarendon Press in Oxford, who successfully pleaded the need to obtain a subsidy in each case from the Sudan Government.

The Azande analysis (1937) immediately struck chords; Collingwood was one of the first to refer to it in print in his *Principles of Art* (1938). After the War, the book became widely influential in the debate opening up within philosophy over the socially and historically placed foundations of what could be understood as knowledge. It also resonated with the more flexible approaches being taken by historians toward the social context of 'witchcraft' phenomena in the European and American past. Because of its pace and well-constructed storyline, the book easily moved into teaching syllabuses and has always been popular with students. An abridged edition appeared in 1976, mainly to serve teaching courses, and has become a standard landmark in interdisciplinary discussions about the nature of knowledge, proof, belief, and evidence. The Azande and their doggedly 'rational' and clever way of monitoring and justifying the operation of their oracles (still quite surprising to many readers) have helped to define anthropology's contribution to such discussions. They have certainly entered almost every textbook on the subject published since the War. Ironically, however, the appeal of the Azande has more recently converged with new and easy forms of cultural relativism, which lead students to ask "Why should not the Azande believe in witchcraft, if it suits their way of life, and who am I to criticize them?" This is not Evans-Pritchard's line; on the contrary, he shows how most ordinary Azande themselves are not at all comfortable with the presence of witchcraft. It is those in powerful and privileged positions who focus on the idea, as the source of suffering and death, and turn their accusations on the relatively helpless. Sometimes cruel kings and princes ultimately control the oracles. Even at the household level only senior men have access to them, while women constitute a large proportion of the accused. Evans-Pritchard's analysis is explicitly set in a political and historical context that offers a more nuanced reading than that of simple relativism, and ultimately, it can be argued that this extra depth is what gives the text its lasting qualities.

The Nuer, as a people, and *The Nuer* (1940) as Evans-Pritchard's first book on them, next entered the life and language of academic anthropology. This was the first time that 'œcology' – now simply rendered as 'ecology' – had explicitly been selected for analysis in a field monograph. It was also the first analytical field study of the political principles behind patterns of hostility and alliance between apparently unrelated groups of mobile herders, here cast in a strong idiom of moral obligation invoked through the situational rhetoric of paternal descent crosscut by marriage alliances and maternal kinship links. Also for the first time, structured conceptions of space, and of time, were shown to have a social dimension and to constitute an overall framework of shared understandings within which political action was cast and could be 'read' by an observer. The book offered quite a different vision of 'social structure' from the empiricism of Radcliffe-Brown, and, significantly, has since been interpreted as a forerunner of the abstract 'structuralism' associated especially with Lévi-Strauss; Louis Dumont (in his introduction to the French translation, 1975) even claimed that it had not been properly appreciated by English-speaking anthropologists. *The Nuer* nevertheless served as model and inspiration to a wide range of field studies of 'segmentary' political systems both in Africa and elsewhere. With its successors *Kinship and Marriage among the Nuer* (1951a, written as a companion text and explicitly focusing on kinship rather than politics) and *Nuer Religion* (1956) it provoked a considerable industry of 'reanalysis', criticism, and comment. The point was often made that Evans-Pritchard's work was too closely associated with the interests of the (then) Sudan Government and with the perspectives of British colonial and imperial authorities generally, that this historical context was missing, and that his work and those ethnographies inspired by it were still in line with the 'structural functionalism' of Radcliffe-Brown. The secondary debates flourished in American and British anthropology for some decades until fresh field research began to alter the position of 'the Nuer' within anthropology by asking questions about them and their social world from a different standpoint than that of Evans-Pritchard's founding texts, taking into account aspects of social interconnection across the wider region of the southern Sudan and the longer term as well as postcolonial social history

of North East Africa as a whole and its ongoing struggles (for example, Johnson, 1994; Hutchinson, 1996). *The Nuer*, along with the collection on *African Political Systems*, coedited with Meyer Fortes and appearing in the same year, nevertheless marked the birth of the subdiscipline of field-based political anthropology and a new dialog with the growing subject of political science.

The final volume in Evans-Pritchard's trilogy devoted to the Nuer also helped create a subdiscipline, in this case the ethnographic study of religion, and also established a new baseline for interdisciplinary conversations extending as far as theology. *Nuer Religion* (1956) was written quite provocatively as a challenge to academic theology, requiring its specialists as well as a wider readership in religious studies to take tribal systems of belief and practice seriously. Evans-Pritchard's presentation of Nuer trust in *kwoth*, the Spirit of the above, has the schematic character of a creed, but at the same time offers a glimpse of a lived and shared reality. Nuer religion is certainly presented for us in terms which Evans-Pritchard has drawn, deliberately, from the traditions of Judaism and the Christian church: we are confronted not only with a 'high god' in the Nuer world, but with sin, with prayer, with souls, priests, and prophets. He portrays for us in a distinctly biblical tone, an elemental confrontation between God and Man, demanding recognition across the gulf between the historical religions and their sacred texts on the one hand and the worlds they often denigrate on the other as erring paganism or blind custom. But the analysis is not merely of belief in or representation of the divine. It focuses equally on practice; for example, upon the ritual of sacrifice in its context of bringing formerly feuding groups together, in reconciliation with each other and with the various forces of spirit. The work is presented essentially as social, rather than theological, analysis. It emphasizes how diversity of perception and representation reflects the relativity of points of view within Nuer society as established in his first book, and it draws the line (explicitly) at entering the subjectivity of religious experience. A turn toward this very subjectivity, nevertheless, has been at the heart of more recent work in the anthropology of religion, opening up avenues to a more direct exploration of experience and faith within the broad extent of the major world religions.

While Evans-Pritchard's best-known ethnographic work was carried out in the southern Sudan in the late 1920s and 1930s, he also found opportunities to undertake field research in several other countries after the outbreak of the War. He was commissioned in the Sudan Defence Force in 1940, and fought alongside locally recruited militia from the Anuak, whose people straddled the border with Ethiopia. He drew attention in early publications to the context of kingship among the Anuak, and – on the basis of passing acquaintance and historical sources rather than dedicated fieldwork – he was to later to publish what became a classic account of the related 'divine kingship' among the Shilluk (the Frazer Lecture, 1948, in Evans-Pritchard, 1962). In 1942 Evans-Pritchard was posted to Libya, where he served in the British Military Administration following the liberation of that country from Italian occupation. His investigations of the Bedouin tribes of the interior bore fruit in his book the *Sanusi of Cyrenaica* (1949), an explicitly historical treatment of some of the ideas about segmentary social relations which he had worked out for the Nuer. By the time the War was over, Evans-Pritchard had achieved a greater ethnographic range than any of his British predecessors. He had also found innovative ways of presenting his ethnography in relation to problems of a philosophical or sociopolitical kind. The combination of these qualities was the basis on which, after the War, he was able to relaunch and develop his discipline.

The Rejuvenation of Anthropology in Oxford

Back in Oxford, with his appointment as Professor of Social Anthropology in 1946, Evans-Pritchard was able initially to build on Radcliffe-Brown's work in defining and promoting 'social anthropology' and acknowledging its affiliation to the French school of Durkheim and the journal *L'Année Sociologique*. He was no more able, however, than his predecessor to establish an undergraduate degree in anthropology as such, because of the scale of sibling rivalry between interests of museum studies, biology, anatomy, sociology, and so on. So Evans-Pritchard focused on the development of 'social anthropology' as a field specifically for graduate coursework and doctoral theses almost always based on field research. He fostered a thriving community of scholars and researchers, drawn from people with a wide variety of nationalities and academic backgrounds. This diversity stimulated important new approaches and helped achieve an international profile for 'British social anthropology' in the scholarly tradition of the humanities. He facilitated studies of places and peoples in the Indian subcontinent, the Middle East, and the Mediterranean, thus opening the way for anthropology to extend its comparative scope from the 'primitive' or 'preliterate' to include studies of the social forms and cultural life of communities embedded within world's major civilizations. The Institute of Social Anthropology at Oxford trained many young anthropologists who left for distinguished careers elsewhere, including several universities in the UK, in Europe, in North America, and in the growing number of Universities in the newly independent countries of Asia and Africa (Rivière, 2007: Epilogue). Even back in the Sudan, a department of Social Anthropology & Sociology was founded in 1958 in the University of Khartoum, with Ian Cunnison as its first Professor, soon to be joined by several other anthropologists also trained in Evans-Pritchard's Institute – including Talal Asad, Lewis Hill, Ahmed al-Shahi, Peter Harries-Jones and myself. Fruitful links with anthropologists based in North America continued throughout Evans-Pritchard's years in the Oxford Chair; there were not only many American visiting scholars, but he himself spent a period in 1950 as a visiting professor at Chicago, and in 1957 as a fellow at the Center for Advanced Study in Behavioral Sciences, Stanford, California.

Although Evans-Pritchard did not embark on new field research after the War, he continued to draw on his field materials in a range of publications through his life. He also sought to define, and to some extent redefine, the historical sources of key ideas and methods in modern social anthropology. He lost no time in promoting a series of translations of essays from the French journal *L'Année Sociologique*, by Durkheim, Mauss, Hertz, Hubert, and their colleagues. Some of these translations have become established classics

worldwide – for example, Marcel Mauss's essay on *The Gift* (first translation 1954), which remains a foundation text in the teaching of anthropology, especially economic anthropology. The French connection continued – Oxford gave a lively reception to the ideas of Claude Lévi-Strauss, and indeed a warm welcome to the lecture he delivered on the occasion of his award of a honorary degree, in the early 1960s.

Evans-Pritchard was also engaged, throughout his post-War career, in a long series of comparative, thematic, and historical studies, which, by looking back to authors of the French and Scottish Enlightenment traditions, helped shape the emerging profile of the discipline of social anthropology as he understood it. It remains true that his concrete ethnographic analyses contain and exemplify his real genius, and to appreciate his 'theories' or 'methods' we need, indeed, to engage with these. But in his later writings he made explicit his life-long sympathy for the historian's perspective and his ambition to establish social anthropology as 'a kind of historiography', a comparative study of the forms of social life, which did not caricature those forms as mere empirical data but recognized their moral character and their participation in the same historical world as that of the observer (1951b, 1962). He specifically evoked R.G. Collingwood's vision of what 'history' could be (Collingwood, 1946) in sketching his own ambitions for social anthropology, and also followed Collingwood's example in drawing up a new academic lineage for his disciplinary field as he saw it. The genealogy he sketched out for social anthropology no doubt owed something to his admiration for the work of Isaiah Berlin, whom he knew well through their common membership of All Souls College. From a collection of little-known and unpublished lectures and writings by Evans-Pritchard edited and brought out posthumously by André Singer (1981), we can see, for example, the continuing lines of sympathetic thought on the nature of human social life identified on the French side from Montesquieu (1689–1775) through Condorcet (1743–94) to Comte (1798–1857 – acknowledged father of sociology) and Lévy-Bruhl – who enters our own era of twentieth century anthropological debate and overlaps with the school of Durkheim. On the Scottish side we start with Lord Kames (1696–1782), and continue with Ferguson, Millar, McLennan, Robertson Smith, and so again to the lively debates of the late nineteenth and early twentieth centuries. Evans-Pritchard took the writings of all these, along with those of writers already acknowledged as anthropology's ancestors (such as Tylor, Frazer, etc.) seriously and pinpointed some of their real insights. As a minor but charming example: he notes that in the late eighteenth century Condorcet 'supported all the liberal programs of his time', and was a believer in the perfectibility of man. All peoples about whom we know, suggested Condorcet, "fall somewhere between our present degree of civilization and what we are told about savage tribes". His work amounted to a history of human thought (Evans-Pritchard, 1981: p. 35).

It has sometimes been suggested that Evans-Pritchard's style of anthropology could be glossed as 'the translation of culture', and indeed this was adopted as the title of one of the main volumes of essays offered to him on his retirement in 1970 (Beidelman, 1971). However, as a gloss on his overall repositioning of anthropology this phrase does not really fit the bill – deriving more from late twentieth century ideas of 'culture' difference than from the disciplined comparison of social forms toward which Evans-Pritchard strove. Mauss's essay on *The Gift*, for example, is clearly about more than the 'translation of culture'. Evans-Pritchard was certainly concerned with translation, but essentially as a craftsman in handling the subtleties of language as such, as for example in his elegant paper on *sanza*, the way in which Azande employ proverbs in double-talk (1962), or in his treatments of poetry and oral literature. From his early training in history, Evans-Pritchard remained always sensitive to language, not so much as a part of 'culture', but as integral to the independent intellectual, political, and aesthetic agency, which he saw as key to the life of the peoples he studied. A project he undertook with colleagues in the 1960s was the launching of the Oxford Library of African Literature, a series that flourished over the following two decades, and had considerable influence in developing anthropology's openness to art, performance, narrative, and the multiple uses of both written and oral textual material. Several distinguished African authors were included in the series, which had significant impact beyond academia in the narrow sense.

In a very few years from his 1946 appointment, Evans-Pritchard had become something of a living legend, a 'household name' in anthropology. His works not only helped to define his subject to successive generations of students, both within anthropology and in sister disciplines. He and his colleagues also pursued a campaign for its public recognition, through radio talks (1951b), addresses to bodies such as the Fawcett Society, and reviews in journals such as the *Times Literary Supplement* and *The Listener*.

In achieving the particular success of social anthropology in Oxford, Evans-Pritchard initially had to create some distance between his approach and that of competing claims over resources, recognized posts, and approved degrees from three other interested parties – at least in Oxford – from the museum side, where ethnology was at first taught quite independently; the physical anthropology side, which had its own posts and courses; and from Sociology, which posed a new threat to the Chair of Social Anthropology, campaigning actively to take it over. The volume edited by Rivière to mark the centenary of Oxford's anthropology (2007) provides a new set of perspectives on the whole period. With hindsight one can admire the success of Evans-Pritchard in defending the 'new' social anthropology of his day, particularly in forging a clear identity in respect of sociology, which in its new post-War vigor was seen as a richer cousin and tremendous rival (and indeed still is, especially in claims to public funding). Here, his emphasis lay on anthropology's qualitative appreciation of *relational structures and processes* in social life. He pressed for us to make the widest possible use of worldwide ethnography, in part as a way of looking critically at modern-day Western social life. Anthropologists found themselves trying to explain, in a way that was not common in sociology at that time, the principles behind the uses of language in the making of persons and in meaningful interaction; as well those behind the choreography of kinship and marriage, the beliefs and rituals of religion, or even the likely hidden agendas behind the workings of authority, power, and political game-playing outside the modern nation-state.

Legacy

The lasting significance of Evans-Pritchard's lifework lies partly in its successful move away from simple observer's empiricism and a narrowly based definition of the subject matter of anthropology, toward an engagement with key debates in history, politics, literary criticism, philosophy, theology, and religious studies (for evaluations, see, for example, Douglas, 1980; Gellner, 1981, Barnes, 1987; James, 2007). At the same time, these subjects have come increasingly to recognize the value of empirical inquiry into the lives, opinions, memories, and practices of people living 'in the world', while the feminist critique and new fields such as gender studies, cultural, and media studies have adopted many ideas from anthropology. They have also helped to give a new lease of life to the idea of 'ethnography' as a reflexive engagement with the world.

Anthropology has moved definitively from the 'tribal periphery' toward the comparative study of the major civilizations, including the practice of the world religions; it is now tackling modern economic ideas and international development, urban life, and global processes. A generation of anthropologists born long after Evans-Pritchard's era have naturally sought to refashion and refresh the subject in the fast-changing world from the late twentieth into the present century, and various lines of criticism have been directed toward his work. Along with his contemporaries, he has been taken to task for not keeping more of an intellectual distance from the colonial situation in which anthropology had such a privileged position. He has been criticized for being too impressionistic in his ethnographic portrayals, for not adopting a more differentiated view of the peoples he was studying, and not attending more carefully to the specifics of their lives. He has been accused of using a patronizing tone and perpetuating character stereotypes. On the question of religion, he has been accused of presenting Nuer beliefs as too biblical and the people as too pious. He has also had his fair share of accusations over 'male bias'. Details of his Nuer ethnography (though rarely of his accounts of the Azande) have been combed over and reanalyzed in order to support various more 'rigorous' accounts based on modern social or psychological theories. Evans-Pritchard himself rarely responded directly to such criticisms. From the mid-1960s on, in common with others whose lives had become involved in the Sudan in one way or another, he was shocked and angry about the deepening civil war in the country and its effect on people he remembered, and who remembered him. Perhaps partly because of this he had no taste for picking over his earlier writings. He developed various enthusiasms for both radical and conservative causes, and seemed to enjoy his growing reputation for eccentricity (see Lienhardt's memoir, 1974). Age added more of a waspish tone to his famous wit and camaraderie, while the debates over his work were left to others.

With hindsight, and the turn of another century, it can be seen that Evans-Pritchard's work remains a source of inspiration despite the broadside against 'functionalism' and imperial collusion. He was 'accused' of working for the Sudan government, but at the time he was aware of the contradictions of his position and engaged in debate with the colonial authorities (Johnson, 1982). He put very vividly those paradoxes of multiple belonging or 'positionality', which the modern student ponders, focusing in a still helpful way on the relativity of 'self and other' in the making of personal and group identities, as we might rephrase the relevance of segmentary theory today. He placed 'dialogue' both between the Azande themselves and between himself and his informants firmly at the heart of the analysis of knowledge and belief that he carried out among them. He also respected, explicitly, the privacy and inner religious consciousness of the Nuer whose world became in so many other ways a public property through his work. It is clearly true that in many respects Evans-Pritchard failed to take 'colonialism' into account in his analyses, or to allow for historical change; but then again, looking back, he is foremost among those who worked to reorient anthropology toward history. His treatment of field research was never as an end in itself; field research was only one of many methods, including many of those scholarly methods standard to history and the social sciences, to be drawn into anthropological analysis.

Moreover, Evans-Pritchard's anthropology was broader, in practice, than many imagine, including, for example, the sensitive use of informants' own accounts about relations between men and women among the Azande – long before gender or sexuality had become an academic topic (1974). Nor was he confined to work in 'social' anthropology alone; he made quite extensive collections of ethnographic objects, musical recordings, and photographs, most now deposited in the Pitt Rivers Museum in Oxford, and available online. In Britain, social anthropology has sought to reestablish debates with the study of material culture, art, and music on the museum side and also with medical and biological anthropology. Evans-Pritchard's work remains a source of questions relevant to these new debates. By the late twentieth century, Evans-Pritchard's vision of the translatability of language, culture, and experience had come to be widely accepted as a starting point for anthropology, together with a holistic view of 'sociality' as an essentially shared human trait. Today we might claim that the 'social anthropology', which Evans-Pritchard promoted as distinct from the museum studies, or biological and evolutionary studies of his day, is now able to engage at a sensitive level with advanced modern work in these areas, and help foster new kinds of integration across the anthropological field. The structure of research and teaching in his own former Institute has since 1990 brought together a range of different specialist graduate courses within the overall institution of the 'School of Anthropology and Museum Studies', between which there is much exchange and collaboration. There are also interdisciplinary undergraduate degrees in which the legacy of 'social anthropology' is prominent and makes a constructive contribution to conversations with the new biological sciences of humanity and today's lively work of the archaeologists.

Evans-Pritchard was celebrated with several volumes in his honor during his lifetime, and with others dedicated to his memory. A research fellowship in the name of his wife, Ioma Evans-Pritchard, is established at St. Anne's College in Oxford, and All Souls College holds an annual series of Evans-Pritchard lectures primarily designed for young scholars to present their research on aspects of the anthropology, archaeology, or history of the major regions in which Evans-Pritchard himself worked: Africa, the Middle East, and the Mediterranean.

See also: Anthropology, History of; Anthropology: Overview; Comte, Auguste (1798–1857); Durkheim, Emile (1858–1917); Functionalism, History of; Historical Thought and Historiography: Sub-Saharan Africa; Imperialism; Mauss, Marcel (1872–1950); Political Anthropology; Religion, History of; Social Structure; Witchcraft.

Bibliography

Barnes, John A., 1987. Edward Evan Evans-Pritchard. Proceedings of the British Academy 73, 447–489.

Beidelman, Tom O. (Ed.), 1971. The Translation of Culture: Essays to E.E. Evans-Pritchard. Tavistock, London.

Beidelman, Tom O. (Ed.), 1974. Evans-Pritchard: A Bibliography. Tavistock, London.

Collingwood, Robin G., 1938. The Principles of Art. Clarendon Press, Oxford, UK.

Collingwood, Robin G., 1946. The Idea of History. Clarendon Press, Oxford, UK.

Davis, John, 2007. In: Rivière (Ed.), How All Souls Got Its Anthropologist, pp. 62–82.

Douglas, Mary, 1980. Evans-Pritchard. Harvester Press, Fontana, Brighton, UK.

Dumont, C. Louis, 1975. Preface to the French edition of the Nuer. In: Beattie, J.H.M., Lienhardt, R.G. (Eds.), Studies in Social Anthropology. Clarendon Press, Oxford, UK.

Evans-Pritchard, Edward E., 1937. Witchcraft, Oracles, and Magic among the Azande. Clarendon Press, Oxford, UK. Abridged edition issued in 1976.

Evans-Pritchard, Edward E., 1940. The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People. Clarendon Press, Oxford, UK.

Evans-Pritchard, Edward E., 1949. The Sanusi of Cyrenaica. Clarendon Press, Oxford, UK.

Evans-Pritchard, Edward E., 1951a. Kinship and Marriage among the Nuer. Clarendon Press, Oxford, UK.

Evans-Pritchard, Edward E., 1951b. Social Anthropology. Cohen & West, London.

Evans-Pritchard, Edward E., 1956. Nuer Religion. Clarendon Press, Oxford, UK.

Evans-Pritchard, Edward E., 1962. Essays in Social Anthropology. Faber & Faber, London. Chapter 1, 9.

Evans-Pritchard, Edward E., 1974. Man and Woman Among the Azande. Faber and Faber, London.

Evans-Pritchard, Edward E., 1981. A History of Anthropological Thought. In: Singer, André (Ed.), Introduction by Ernest Gellner. Faber & Faber, London, pp. xiii–xxxvi.

Fortes, Meyer, Evans-Pritchard, Edward E. (Eds.), 1940. African Political Systems. Oxford University Press, London, UK.

Hutchinson, Sharon E., 1996. Nuer Dilemmas: Coping with Money, War, and the State. University of California Press, Berkeley, CA and Los Angeles.

James, Wendy, 2007. 'A Feeling for Form and Pattern, and a Touch of Genius': E-P's Vision and the Institute 1946–70. In: Rivière (Ed.), A History of Oxford Anthropology. Berghahn, Oxford, pp. 98–118.

Johnson, Douglas H., 1982. Evans-Pritchard, the Nuer, and the Sudan political service. African Affairs 81, 231–246.

Johnson, Douglas H., 1994. Nuer Prophets: A History of Prophecy from the Upper Nile in the Nineteenth and Twentieth Centuries. Clarendon Press, Oxford.

Lienhardt, Ronald G., 1974. Evans-Pritchard: a personal view. Man n.s. 9, 301–303.

Mauss, Marcel, 1954. The Gift: Forms and Functions of Exchange in Archaic Societies (I. Cunnison, Trans.). Cohen & West, London.

Rivière, Peter G. (Ed.), 2007. A History of Oxford Anthropology. Berghahn Books, New York & Oxford.

Event History Analysis

Richard J Cook and Jerald F Lawless, University of Waterloo, Waterloo, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Methods for event history analysis are well suited to the study of dynamic phenomena in social and behavioral research. In such cases, stochastic models for times to events, recurrent events, or multistate transitions are often used to provide insight into the factors influencing event occurrence or changes in behavior. Intensity functions play a central role in model formulation and accommodate the study of time trends, fixed and time-varying covariates, and the study of local dependence between two or more processes of interest. Complications due to unexplained heterogeneity between individuals, incomplete observation, and other factors are discussed.

Introduction to Event History Analysis

An Overview

Event history analysis deals with the modeling and analysis of events and related phenomena that occur within individuals, populations, or systems over time. For simplicity, we consider settings in which individuals are at risk of different types of events, but the units observed could equally well be population subgroups or systems.

Methods for event history analysis have a diverse range of application within the behavioral and social sciences, involving topics such as fertility, education, employment, cohabitation and marriage, voting, addiction, and utilization of social services. Broad objectives of analysis include (1) understanding and characterizing processes governing event occurrence in individuals, (2) characterizing variability across individuals and relating it to individual characteristics, interventions, or other factors, and (3) assessing the relationship of time-varying factors or processes to the events of interest. For example, in studies of women's fertility, we might want to relate events such as pregnancy to their age, educational status, employment history, and history of cohabitation and marriage.

Subareas that fall within the domain of event history analysis follow, along with two illustrative settings in which event history methods are useful.

Survival analysis, which deals with times to a specific event, or equivalent durations of a specific condition (Kalbfleisch and Prentice, 2002; Lawless, 2003).

Recurrent event analysis, which deals with events that can occur repeatedly over time (Cook and Lawless, 2007).

Analysis of multiple events, which is concerned with the simultaneous analysis of different types of events (Cook and Lawless, 2007; Wei et al., 1989).

Transitional multistate analysis, which deals with settings where individuals occupy and move between specified states over time (Andersen and Keiding, 2002).

Example 1: Recidivism Among Criminal Offenders

Factors associated with recidivism, that is, repeated criminal acts or convictions, are of interest in connection with crime prevention and rehabilitation of offenders. This is often studied using techniques of survival or event history analysis, since charges and convictions are potentially repeatable events. For example, analysis of recidivism among juvenile offenders would consider demographic factors such as age, gender, and family, along with previous experience with the justice system (e.g., numbers of prior convictions, types of penalties imposed). This would typically involve multivariate event history methods, since there are different types of offenses and many possible outcomes to a criminal proceeding. Techniques for survival analysis are often used to fit relevant models; for example, Skardhamer and Telle (2012) consider the effects of postrelease employment on times to recidivism in Norway.

Example 2: Childbearing and Employment Among Women

Many questions are of interest concerning patterns of employment among women following childbirth. For example, are they likely to return to work part-time, or to delay reentry to their previous field of employment? Are they likely to move to self-employment? Such issues can be addressed using multistate models with different states of employment. The effects of various factors on transitions between states and on state sojourn durations can be considered: factors of interest include education, previous employment history and income, marital status, spousal employment, and previous births.

We next introduce some notation for event history models and then consider some particular types of analysis.

Notation and Key Functions

The types of events mentioned in the previous section can all be considered within the broad statistical framework of multivariate counting processes (Andersen et al., 1993). In particular, suppose J types of events are of interest and that events for a generic individual occur randomly over time. Let $N_j(t)$ denote the number of type j events ($j = 1, \ldots, J$) that occur over the time interval $(0, t)$. The stochastic process $\{N_1(s), \ldots, N_J(s), 0 < s\}$ is referred to as a multivariate counting process and the history of event occurrence up to time t^- is denoted $H(t) = \{N_1(s), \ldots, N_J(s), 0 < s < t\}$ and represents the data on events up to that time. This framework covers single, recurrent, or multiple events; multistate processes are also covered since it is possible to view different types of state-to-state transitions as different

types of events. In the multistate settings we often use additional notation: we assume there is a given set of K states $\{1, \ldots, K\}$ and let $Z(t)$ denote the state occupied at time t.

Let $\Delta N_j(t) = N_j((t+\Delta t)^-) - N_j(t^-)$ denote the number of events of type j over the time interval $[t, t+\Delta t)$. The intensity function for a type j event is defined as

$$\lambda_j(t|H(t)) = \lim_{\Delta t \downarrow 0} \frac{P\left(\Delta N_j(t) = 1 \middle| H(t)\right)}{\Delta t} \qquad [1]$$

(Andersen et al., 1993; Cook and Lawless, 2007). Under mild conditions that exclude the occurrence of more than one event at any given time and the occurrence of arbitrarily large numbers of events in a finite window of time, intensity functions fully specify the multivariate event process and hence offer a useful framework for modeling. The key challenge in event history analysis is often to understand the nature of these intensities, and to approximate them with specific mathematical forms. In most situations event occurrence may be affected by fixed or time-varying covariates, denoted $x(s)$, and in such cases we include their values up to time t in the history and let $H(t) = \{N_1(s), \ldots, N_J(s), 0 < s < t; x(s), 0 < s \leq t\}$.

The intensity functions [1] are typically the focus of attention when we want to understand the dynamics of event occurrence and features such as temporal effects or relationships between different events. In some contexts, however, we may wish to make simple summaries of data and comparisons between different types of individuals on the basis of these. In this case, it is appealing to consider marginal features of event processes. Two such features are:

(a) mean functions, $\mu(t) = E\{N(t)\}$ or corresponding rate functions, $\rho(t) = \mathrm{d}\mu(t)/\mathrm{d}t$ for recurrent events; and
(b) prevalence functions, $P_k(t) = P(Z(t) = k)$ for multistate models.

Note that in the case of a single survival time variable T, we can consider failure (death) as the sole type of event, in which case $\mu(t)$ in (a) is the distribution function for T. Equivalently, if we view the survival setting as a model with the two states: alive (1) and dead (2), then the distribution function corresponds to the prevalence function $P_2(t)$ in (b). Estimation of marginal characteristics requires fewer modeling assumptions than are required for intensity-based analyses, and marginal methods are widely used in randomized studies where simple comparisons between different groups are desired. Challenges in using marginal characteristics can arise, however, as we mention in Section Further Remarks.

Books on stochastic processes give many types of models for recurrent events (e.g., Daley and Vere-Jones, 2007; Grandell, 1997) and multistate processes (e.g., Cox and Miller, 1977; Ross, 2006). Markov processes, for which intensities [1] are functions only of the elapsed time t since the initiation of the process, and semi-Markov (or Markov renewal) processes, for which they are functions of the elapsed time since the most recent event, are mathematically tractable and widely studied. They are useful in many settings but models that allow other types of dependency on the process history are often needed. We turn now to the analysis of specific types of event history data, beginning with recurrent events. We focus on model fitting and related inference; discussions of model checking can be found in the cited references.

Analysis of Recurrent Events

Consider a recurrent event with counting process $\{N(s), 0 \leq s\}$ and intensity function $\lambda(t|H(t))$. We remark that the time origin $t = 0$ is part of the model, and in some settings there may be appealing alternative choices. In any case, suppose an individual is observed over a specified time interval (t_0, τ) and let $Y(u) = I(t_0 < u \leq \tau)$, where $I(\cdot)$ is an indicator function such that $I(A) = 1$, if A is true and equals zero otherwise. The probability density for the occurrence of n events over (t_0, τ) at times $t_1 < \cdots < t_n$ is then (Cook and Lawless, 2007; Sec. 2.1),

$$\prod_{\ell=1}^{n} \lambda(t_\ell|H(t_{\ell-1})) \cdot \exp\left(-\int_0^\infty Y(u)\lambda(u|H(u))\mathrm{d}u\right), \qquad [2]$$

which can be rewritten as

$$\prod_{\ell=1}^{n} \left\{ \lambda(t_\ell|H(t_{\ell-1})) \exp\left(-\int_{t_{\ell-1}}^{t_\ell} \lambda(u|H(u))du\right)\right\} \cdot \exp\left(-\int_{t_n}^{\tau} \lambda(u|H(u))du\right).$$

We note that the first n terms in the product above have the same form as a likelihood contribution from a survival model with 'hazard function' $\lambda(t|H(t))$ based on failure times that are left-truncated at $t_{\ell-1}$, $\ell = 1, 2, \ldots, n$ respectively; the final term corresponds to a time which is left-truncated at t_n and right-censored at τ. As a consequence of this connection, software for survival analysis can be exploited to fit many recurrent event models. In particular, if $x(t)$ is a $p \times 1$ vector of fixed covariates, time-varying covariates, or aspects of previous event history, then modulated Markov models of the Cox (1972) form are widely applicable. These specify the intensity as

$$\lambda(t|H(t)) = \lambda_0(t)\exp(x'(t)\beta), \qquad [3]$$

where β is a $p \times 1$ vector of regression coefficients. Software for fitting Cox models, such as the coxph function in R (R Development Core Team, 2011) or S-Plus (TIBCO Spotfire S+®, 2008) or PROC PHREG in SAS (SAS Institute Inc., 2008), can be used to conduct analysis; Therneau and Grambsch (2000, Chapter 8) and Cook and Lawless (2007, Chapter 3) provide details and illustrations. Cook and Lawless (2007, Chapter 4) also show how modulated semi-Markov models, for which $\lambda_0(t)$ in [3] is replaced by $\lambda_0(t - t_{N(t^-)})$, can similarly be handled, where $t - t_{N(t^-)}$ is the time elapsed since the most recent event. In each case the likelihood is assumed to be a product of terms of the form [2] across independent individuals, and the data frame on which analysis is based includes one line for each period that each individual is at risk of an event; individuals may therefore contribute several lines to the data frame (Therneau and Grambsch, 2000).

Several additional points deserve mention. First, likelihood methods remain valid when the start (t_0) and stop (τ) times of observation for each individual are random and depend on prior (but not future) process history (Aalen and Husebye, 1991). Second, the process history $H(t_0)$ prior to the start of follow-up is needed in order to specify the intensity function for some models. Third, it is assumed here that full observation of event occurrence times and covariates is possible; cases

where some information may be missing require special treatment (e.g., Lawless and Zhan, 1998; Cook and Lawless, 2007, Chapter 7; Sun and Zhao, 2013).

The model [3] can also be used to provide analysis of mean and rate functions, as defined in the Introduction. Suppose that only fixed covariates x are considered in [3] and that the form of $\lambda_0(t)$ is left unspecified. It is then known as the Andersen–Gill model (Andersen and Gill, 1982) and the rate and mean functions for an individual with covariate vector x are, respectively,

$$\rho(t|x) = \lambda_0(t)\exp(x'\beta), \quad \mu(t|x) = \Lambda_0(t)\exp(x'\beta), \qquad [4]$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u)\mathrm{d}u$. These are log-linear models, which are widely used for the analysis of event counts over prespecified time intervals (McCullagh and Nelder, 1989). With a sample of m independent individuals in which the counting process $\{N_i(s), 0 < s\}$ for individual i is observed over (t_{i0}, τ_i), we let $\mathrm{d}N_i(s) = 1$ if an event occurs at time s and 0 otherwise, $Y_i(s) = I(t_{i0} < u \le \tau_i)$, and the covariate vector be x_i. Then, the estimate of the cumulative baseline intensity is

$$\widehat{\Lambda}_0(t) = \int_0^t \mathrm{d}\widehat{\Lambda}_0(s) = \sum_{s_r \in D(t)} \mathrm{d}\widehat{\Lambda}_0(s_r) \qquad [5]$$

where $D(t)$ is the set of unique event times in the sample at or before t,

$$\mathrm{d}\widehat{\Lambda}_0(s) = \frac{\sum_{i=1}^m Y_i(s)\mathrm{d}N_i(s)}{\sum_{i=1}^m Y_i(s)\exp\left(x_i'\widehat{\beta}\right)}, \qquad [6]$$

and $\widehat{\beta}$ is the estimate obtained from Cox model software, as discussed above. The estimates for β and $\Lambda_0(t)$ obtained using software for Cox regression are robust provided the multiplicative form in [4] is valid, and coxph and PROC PHREG have options that yield robust variance estimates to ensure valid inference when the Andersen–Gill (Poisson) model does not hold, but [4] does; see Cook and Lawless (2007; Sec. 3.6 and 7.2). In the absence of covariates (i.e., in a one-sample problem), the denominator in [6] simplifies to $\sum_{i=1}^m Y_i(s)$ and the resulting estimator in [5] is called the Nelson–Aalen estimator of the common mean function (Cook and Lawless, 2007). More generally, $\widehat{\Lambda}_0(t)\exp(x'\widehat{\beta})$ is an estimate of the mean function in [4]. In some situations, there is more heterogeneity (variability) in event occurrence across individuals than is accounted for by observed covariates. Sometimes this is due to inadequate incorporation of previous event history in models for the intensity, but it may also be due to unobserved factors that affect event occurrence. For the latter case, it can sometimes be helpful to introduce latent variables to represent unmeasured characteristics and use them to accommodate unexplained variation between individuals. For example, if u_i is a random effect associated with individual i, then the conditional intensity specification

$$\lambda_i(t|H_i(t), u_i) = u_i\lambda_0(t)\exp\left(x_i'(t)\beta\right) \qquad [7]$$

is often used, with the u_i assumed to be independent and identically distributed with some distribution $G(u)$. Such models are often termed 'frailty' models and can be fitted with the R/S-Plus coxph function; see Therneau and Grambsch (2000, Chapter 9) or Cook and Lawless (2007; Sec. 3.6) for details. It should be noted that all random effect models that can be fitted with existing software assume random effects for a given individual are fixed over time; this is a strong assumption.

Heterogeneity and dependence on past events cannot be clearly distinguished in many contexts. For example, suppose that u_i in [7] has a gamma distribution with $E(u) = 1$ and $var(u) = \phi$. The 'observable' intensity obtained by averaging over u_i is then (Cook and Lawless, 2007; Sec. 2.2.3)

$$\lambda(t|H_i(t)) = \left\{\frac{1 + \phi N_i(t^-)}{1 + \phi\Lambda_0(t)\exp\left(x_i'\beta\right)}\right\}\lambda_0(t)\exp\left(x_i'\beta\right) \qquad [8]$$

where for simplicity we restrict attention to the case of fixed covariates. It is apparent from [8] that the occurrence of an event leads to an instantaneous rise in the risk of subsequent event occurrence over the baseline rate $\lambda_0(t)$, and that this decreases as $t - t_{i,\,Ni(t^-)}$ increases. Note that [7] is mathematically equivalent to [8] plus the gamma assumption for u_i, and the only way we can know whether [8] arises from heterogeneity or from a direct effect of the number of previous events is through external information regarding the process (Cook and Lawless, 2013).

Multitype Recurrent Events

When recurrent events are classified into types, as discussed in Section Introduction to Event History Analysis, questions may be directed at marginal type-specific processes or at the nature of association between event types. Intensity models based on [1] and joint data on all types of events give likelihoods of the form

$$L \propto \prod_{j=1}^{J}\left\{\left[\prod_{t_{jr}\in D_j}\lambda_j\left(t_{jr}|H\left(t_{jr}\right)\right)\right]\exp\left(-\int_0^\infty Y(u)\lambda_j(u|H(u))du\right)\right\}, \qquad [9]$$

where $D_j = \{t_{jr}, r = 1, 2, \ldots, N_j(\tau)\}$ is the set of times of type j events occurring over (t_0, τ), $j = 1, 2, \ldots, J$. Provided the intensities for the different events do not share any parameters, the J different contributions to [9] can be maximized separately.

It is important to note that the likelihood [9] is based on intensities for each type of event conditioning on the prior history of all event types; using it is different than conducting independent analyses of each type. If $J = 2$, for example, intensities of the form

$$\lambda_j(t|H(t)) = \lambda_{j0}(t)\exp\left(x'(t)\beta_j + N_{3-j}(t)\theta_j\right) \quad j = 1, 2 \qquad [10]$$

may be specified, in which case $\exp(\theta_j)$ quantifies how the risk of a type j event changes with the occurrence of a type $3 - j$ event; if $\theta_1 = \theta_2 = 0$, then the two event processes are independent. An appealing feature of this formulation is that it allows one to explore asymmetric forms of so-called local dependence (Aalen et al., 2008) since we may have, for example, $\theta_1 = 0$ but $\theta_2 > 0$. In that case the occurrence of a type 2 event does not alter the risk of type 1 event, but type 1 events increase the instantaneous risk of a type 2 event.

When interest lies in examining covariate effects on separate event types, marginal models are often useful and are formed

without conditioning on the history of the other event types. That is, we consider separate models

$$\lambda_j^m\left(t\middle|H_j(t)\right) = \lim_{\Delta t\downarrow 0}\frac{P\left(\Delta N_j(t) = 1\middle|H_j(t)\right)}{\Delta t} \qquad [11]$$

where $H_j(t) = \{N_j(s), 0 < s < t; x(s), 0 < s \leq t\}$, and estimate the parameters by maximizing [9] with [11] in place of $\lambda_j(t|H(t))$. Simultaneous inferences about different types can be valid in some settings, provided robust covariance estimates are used (Cook and Lawless, 2007; Lin et al., 2000).

Another common approach for addressing dependence between event types is to assume independence between processes conditional on a shared random effect. In this case it is common to specify models similar to [7], of the form

$$\lambda_j(t|H(t),u) = \lim_{\Delta t\downarrow 0}\frac{P\left(\Delta N_j(t) = 1\middle|H(t),u\right)}{\Delta t} = u\lambda_j^m\left(t\middle|H_j(t)\right),$$

where u is a latent variable with $E(u) = 1$ and $var(u) = \phi$. If covariates are fixed, the processes are Poisson given the random effect (i.e., $\lambda_j^m(t|H_j(t)) = \rho_j(t|x)$) and if, in addition, u is gamma distributed, then in the case $J = 2$ the observed intensity for type j events becomes

$$\lambda_j(t|H(t)) = \left\{\frac{1+\phi\left[N_1\left(t^-\right)+N_2\left(t^-\right)\right]}{1+\phi[\mu_1(t|x)+\mu_2(t|x)]}\right\}\rho_j(t|x), \qquad [12]$$

where $\mu_j(t|x) = \int_0^t \rho_j(u|x)\mathrm{d}u$ is the mean function and $\rho_j(t|x)$ is the rate function for type j events, $j = 1,2$. This model extends [8] to allow the risk of type 2 events to depend on the cumulative number of type 1 events and since $\phi > 0$, it is geared toward accommodating positive associations. A limitation, however, is that this model may not adequately distinguish the effect of an event of the same type and of a complementary type. We also remark that the proportional increase in risk resulting from event occurrence in [10] is constant until the occurrence of the next event, whereas the increase in [12] is transitory.

A more general model that allows both positive and negative associations involves the specification $\lambda_j(t|H(t), u_j) = u_j\rho_j(t|x)$, where $u = (u_1,u_2)'$ follows a bivariate distribution with $E(u_j) = 1$, $var(u_j) = \phi_j$, $j = 1, 2$, and $cov(u_1,u_2) = \phi_{12}$. If $\mu(s,t) = \mu(t) - \mu(s)$, a consequence of this formulation is that $\mathrm{var}\{N_j(s,t)|x\} = \mu_j(s,t) + \mu_j^2(s,t)\phi_j$, $cov\{N_j(s_1, t_1), N_j(s_2, t_2)|x\} = \phi_j\mu_j(s_1, t_1)\,\mu_j(s_2, t_2)$, $j = 1, 2$, and $cov\{N_1(s_1, t_1), N_2(s_2, t_2)|x\} = \phi_{12}\mu_1(s_1, t_1)\,\mu_2(s_2, t_2)$, where $s_1 < t_1 < s_2 < t_2$. Thus, the separate parameters ϕ_1 and ϕ_2 accommodate heterogeneity and history dependence for the marginal processes, and a third parameter ϕ_{12} governs the association between processes. Much attention has been given to bivariate log-normal random effects distributions, but multivariate distributions based on copulas provide a richer class of models and enable the specification of different families of marginal densities for random effects (Cook et al., 2010).

Multistate Models

Dynamic processes involving distinct stages can often be conveniently represented by multistate models. As mentioned in the Introduction, such models require specification of the state space, allowable transitions, and the form of the component intensity functions. If $Z(t)$ represents which of K distinct states is occupied at time t, then $\{Z(s), 0 \leq s\}$ denotes the stochastic process. The history of the multistate process and a $p \times 1$ vector $x(t)$ of external (exogenous) covariates is denoted $H(t) = \{Z(s), 0 \leq s < t; x(s), 0 \leq s \leq t\}$. Then if $Y_k(t) = I(Z(t^-) = k)$ indicates an individual is at risk of a transition out of state k at time t,

$$\lambda_{k\ell}(t|H(t)) = \lim_{\Delta t\downarrow 0}\frac{P\left(Z\left(t+\Delta t^-\right) = \ell\middle|Y_k(t) = 1, H(t)\right)}{\Delta t}, \ell \neq k, \qquad [13]$$

denotes the intensities of $k \to \ell$ transitions. Modulated Markov models have intensities of the form

$$\lambda_{k\ell}(t|H(t)) = Y_k(t)\lambda_{k\ell}(t)\exp(x'(t)\beta_{k\ell}), \qquad [14]$$

where $\beta_{k\ell}$ is a $p \times 1$ vector of regression coefficients. The $Y_k(t)$ term in [14] is present to reflect the implicit conditioning on $Z(t^-) = k$ and that $k \to \ell$ transitions are otherwise impossible at time $t > 0$. Such models are useful when the dynamics of the process are most naturally modeled using the time since the origin of the process; examples include processes affected by aging or deterioration over time.

Modulated semi-Markov processes have intensities of the form

$$\lambda_{k\ell}(t|H(t)) = Y_k(t)h_{k\ell}(B_k(t))\exp(x'(t)\alpha_{k\ell}) \quad \ell \neq k \qquad [15]$$

where $B_k(t)$ denotes the time since entry to state k if $Y_k(t) = 1$ and $h_{k\ell}(\cdot)$ is a baseline function governing the cause-specific hazard of the sojourn time in state k ending because of a $k \to \ell$ transition. The parameters $\alpha_{k\ell}$ in [15] reflect the effect of the covariates on this transition but these have different meaning than the coefficients in [14].

If $x(t)$ is broadened to allow internal (endogenous) covariates capturing key aspects of the process history, [14] and [15] give more general intensity models. For example, if in [14] one element of $x(t)$ is $B_k(t)$, then the associated coefficient represents a duration dependence in the intensity. The simplest such model with $p = 1$ has the form $\lambda_{k\ell}(t|H(t)) = Y_k(t)\ \lambda_{k\ell}(t)\ \exp(B_k(t)\ \beta_{k\ell})$ and if $\beta_{k\ell} \neq 0$, the instantaneous risk of a $k \to \ell$ transition depends on both the time since the onset of the process and the time since state k was most recently entered. A model [15] with $p = 1$ and $x(t)$ a specified function of time has the same property with a different primary timescale. There are many options for specifying the form of the intensity functions, and intensities for different types of transitions within a given multistate process may warrant distinct forms of dependence on time.

Andersen et al. (1993) show that for multistate processes the likelihood function based on an observed set of transitions for an individual is $L \propto \prod_{k\neq\ell} L_{k\ell}$, where if $D_{k\ell}$ is the set of observed times of $k \to \ell$ transitions and $\overline{Y}_k(u) = Y(u)Y_k(u)$,

$$L_{k\ell} = \left[\prod_{t_r\in D_{k\ell}}\lambda_{k\ell}(t_r|H(t_r))\right]\exp\left(-\int_0^\infty \overline{Y}_k(u)\lambda_{k\ell}(u|H(u))du\right). \qquad [16]$$

Like [2], this has the form of a likelihood arising in survival analysis and many multistate models can be fitted with survival analysis software; see Therneau and Grambsch (2000, Sec. 8.6) and Cook and Lawless (2007; Sec. 6.7) for illustrations.

Likelihood-based methods facilitate the study of intensity-based (i.e., conditional) models for transitions and the durations of sojourns in specific states. Marginal process features are also of frequent interest and, as with recurrent events, simple estimates can be obtained for some types of models. With Markov models and fixed covariates, in particular, if $d\mathbb{A}(t|x)$ is a $K \times K$ matrix with entries $\lambda_{k\ell}(t|x)\,dt$ in the off-diagonal and $-\Sigma_{\ell \neq k}\lambda_{k\ell}(t|x)dt$ on the diagonal, and if $\mathbb{I}$ is a $K \times K$ identity matrix, then product integration (Andersen et al., 1993) gives $K \times K$ transition probability matrices for $0 \leq s \leq t$ as

$$\mathbb{P}(s,t|x) = \prod_{(s,t]}\{\mathbb{I} + d\mathbb{A}(t|x)\} \qquad [17]$$

with (k,ℓ) entry $[\mathbb{P}(s,\ t|x)]_{k\ell} = P(Z(t) = \ell|Z(s) = k, x)$. This is an important and useful result for several reasons. First, estimates of the parameters governing the transition intensities obtained by maximizing [16] can be inserted into [17] to obtain an estimate of $\mathbb{P}(s,t|x)$. Second, $\mathbb{P}(s,t|x)$ can be used to calculate estimates of marginal features such as the mean time in a given state, say ℓ, over $(0,\ \tau)$ through

$$E\left\{\int_0^\tau I(Z(u) = \ell)\mathrm{d}u\middle|x\right\} = \sum_{k=1}^{K}\int_0^\tau P(Z(u) = \ell|Z(0) = k, x)P(Z(0) = k|x)\mathrm{d}u.$$

Third, estimates of [17] can be useful in cost–benefit analyses by associating costs per unit time with each state and extending the preceding expression. In addition, such estimates are robust to departures from the Markov assumption; see Aalen et al. (2001) and Cook et al. (2009). Finally, in settings where transition times are not observed because individuals are only seen at periodic inspection times (see Section A Study in Childhood Smoking Prevention), likelihoods can be constructed based on elements in [17].

Some Illustrative Analyses

Trend in Hospital Readmission among Individuals with Affective Disorder

Kessing et al. (2004) report on a study of recurrent hospitalizations involving individuals with affective disorder. A main question of interest was whether the risk of readmission tends to increase with the number of prior admissions. The expected number of hospitalizations as a function of time since disease onset (defined as the first admission to hospital due to affective disorder) is also of interest. Here we consider data on 10 523 individuals having a discharge from a first hospitalization between 1 January 1994 and 31 December 1999. Follow-up of an individual was censored upon the diagnosis of schizophrenia, an organic disorder, or death. A total of 6498 readmissions were observed with the number of readmissions

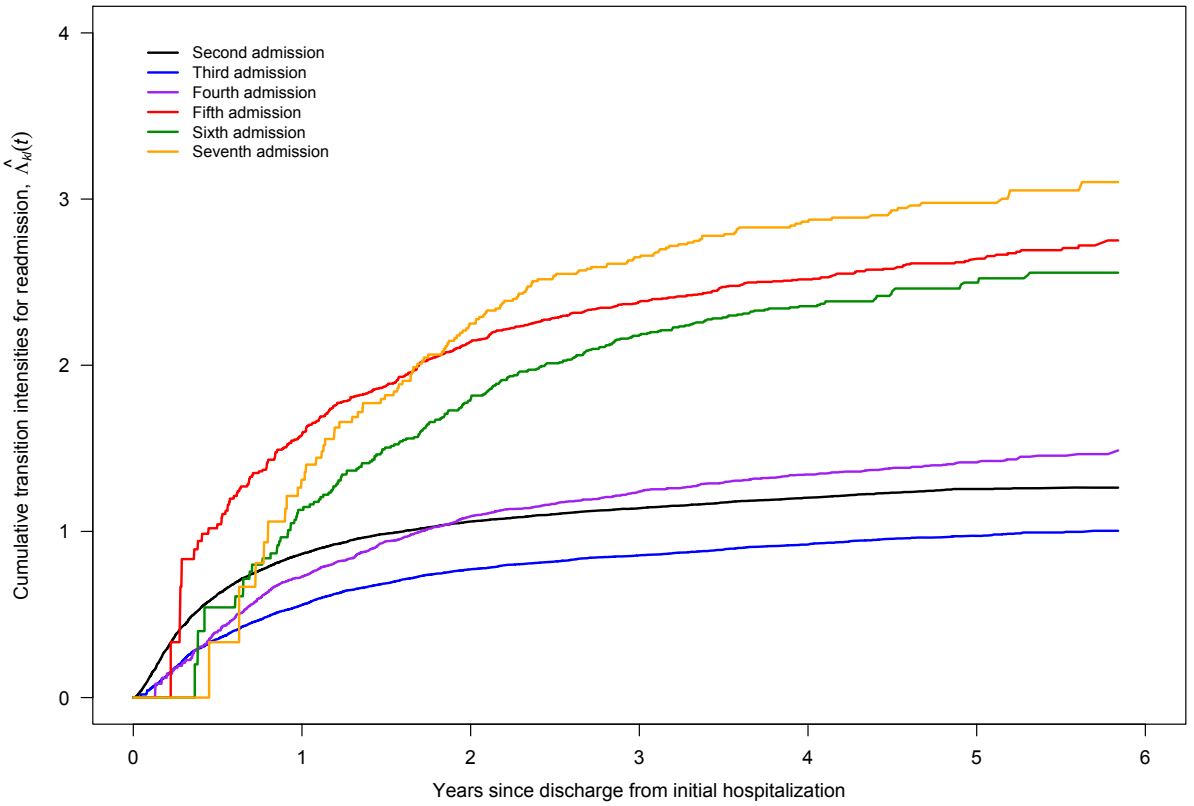

Figure 1 Nonparametric Nelson–Aalen estimates of the cumulative intensities for readmission in the study of affective disorder, stratified by the number of prior admissions.

ranging from 0 to 89. Figure 1 gives plots of nonparametric Nelson–Aalen estimates of the cumulative Markov transition intensities for consecutive hospitalizations (see [14] and [6]), with no adjustment for covariates. Note that there is a suggestion of a trend toward increased risk of hospitalization with increasing number of prior hospitalizations; this is inferred from the increasingly steeper slopes of the cumulative transition intensities. The same feature is seen in analyses based on semi-Markov models for the readmission intensities based on [15]. The left panel of Figure 2 shows estimated baseline cumulative intensity functions in a semi-Markov model involving covariates. As discussed in the section on recurrent events, it is difficult to distinguish trends or dependence on prior events, from heterogeneity. To illustrate this point, the right panel of Figure 2 shows the corresponding estimates based on a mixed semi-Markov model with individual-level random effects; the apparent trend of increasing readmission rates is now greatly attenuated. The much more pronounced trend seen in Figure 1 and the left panel of Figure 2, thus, could be a consequence of one or more of [1] a time trend, [2] residual effects from previous hospitalizations, or [3] heterogeneity of subjects due to unobserved factors.

A Study in Childhood Smoking Prevention

To illustrate a multistate analysis based on intermittently observed subjects, we consider here a randomized youth smoking prevention study designed to evaluate the effectiveness of an experimental health promotion program on smoking behavior in children (Cameron et al., 1999). At the beginning of grade 6, 4:1 randomization was carried out with the majority of children receiving experimental intervention. Annual assessments of smoking behavior were carried out late in the year in grades 7–12. Children can be classified at any given time into one of $K = 3$ states defined as never smoked (state 1), smoker (state 2) or 'quitter' (state 3). These states are represented in Figure 3 where the arrows indicate possible transitions. Once an individual has smoked, he or she cannot subsequently qualify as 'never smoked,' so state 1 is transient, but individuals can quit and subsequently resume smoking, so $3 \rightarrow 2$ transitions are allowed.

Table 1 gives the raw counts of transitions between states over successive annual assessments. Note that while there are individuals who were in state 1 at one assessment and state 3 on the next, they are presumed to have passed through state 2 at least once and these transitions could have been made at any time between the assessments. The R package *msm* (Jackson, 2011) can be used to fit time-homogenous Markov models to data from the intervention arm and control arm separately to give insight into the dynamic aspects of the smoking process. Here we let

$$\mathbb{I} + \mathrm{d}\mathbb{A}(t) = \begin{bmatrix} 1-\lambda_{12}(t)\mathrm{d}t & \lambda_{12}(t)\mathrm{d}t & 0 \\ 0 & 1-\lambda_{23}(t)\mathrm{d}t & \lambda_{23}(t)\mathrm{d}t \\ 0 & \lambda_{32}(t)\mathrm{d}t & 1-\lambda_{32}(t)\mathrm{d}t \end{bmatrix}$$

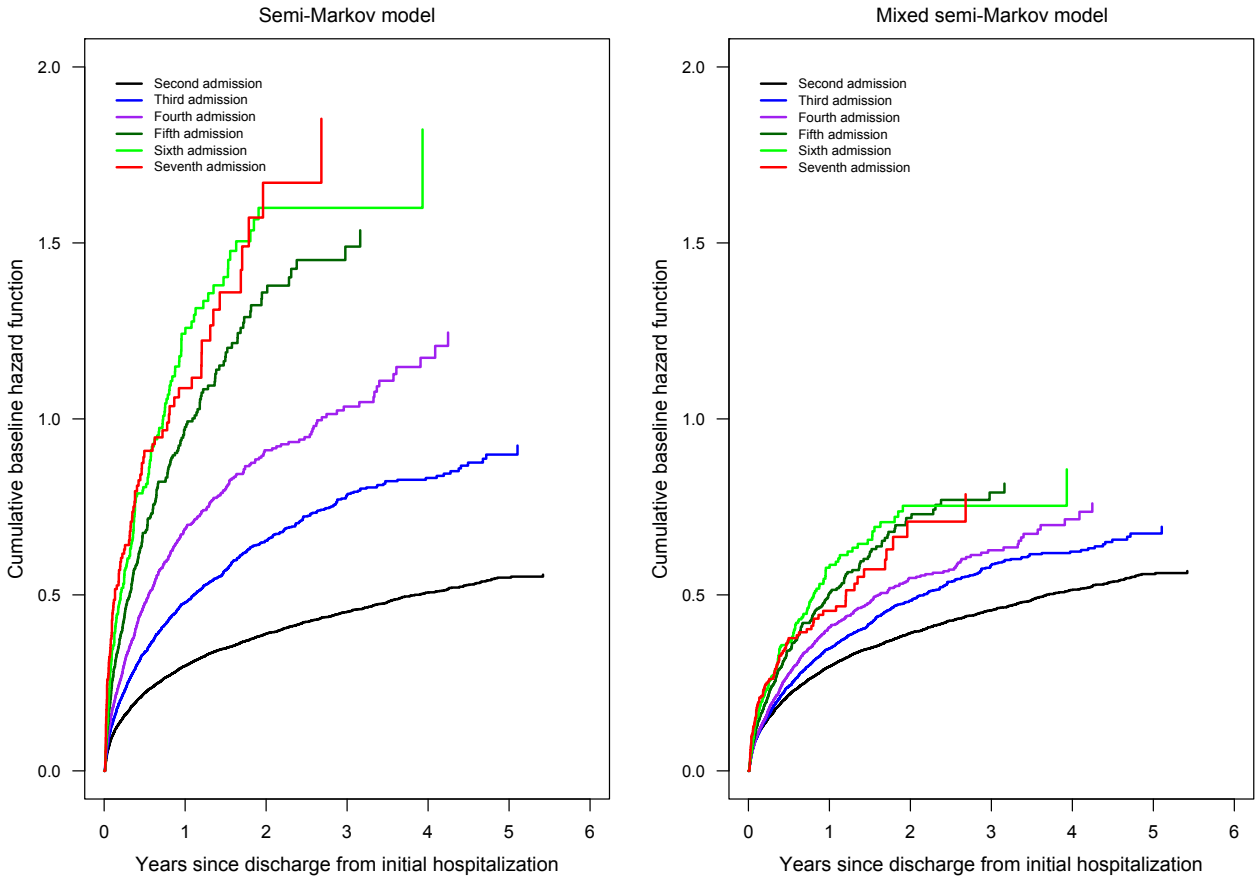

Figure 2 Nonparametric estimates of cumulative baseline hazards from semi-Markov and mixed semi-Markov models for hospitalizations.

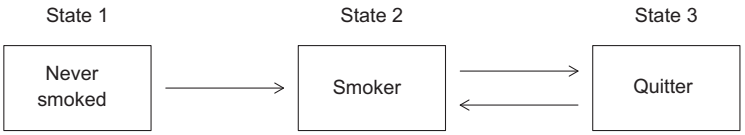

Figure 3 Multistate diagram for dynamic modeling of smoking behavior for analysis of the study by Cameron, R., Brown, K.S., Best, J.A., Pelkman, C.L., Madill, C.L., Manske, S.R., Payne, M.E., 1999. Effectiveness of a social influences smoking prevention program as a function of provider type, training method, and school risk. American Journal of Public Health 89, 1827–1831.

Table 1 Number of transitions between states for the smoking behavior data

		Control			*Intervention*		
To		*1*	*2*	*3*	*1*	*2*	*3*
From	1	2339	325	179	8658	1217	678
	2	0	910	216	0	3071	705
	3	0	266	372	0	907	1330

Cameron, R., Brown, K.S., Best, J.A., Pelkman, C.L., Madill, C.L., Manske, S.R., Payne, M.E., 1999. Effectiveness of a social influences smoking prevention program as a function of provider type, training method, and school risk. American Journal of Public Health 89, 1827–1831.

and obtain $\mathbb{P}(s, t)$ from [17]. If individual i is observed at assessment times $a_{i0}, a_{i1}, \ldots, a_{i,Ri}$, the likelihood for a sample of m independent individuals is

$$\prod_{i=1}^{m}\left\{\prod_{r=1}^{R_i} P\left(Z_i(a_{ir}) = z_{ir} \middle| Z_i\left(a_{i,r-1}\right) = z_{i,r-1}\right)\right\}.$$

If $\pi_k = P(Z(a_0) = k)$ governs the initial distribution, where $a_{i0} = a_0$ denotes the common baseline assessment in grade 6, a natural estimate $\widehat{\pi}_k$ is based simply on the empirical frequencies. One can then obtain

$$\widehat{P}(Z(t) = \ell) = \sum_{k=1}^{4} \widehat{P}(Z(t) = \ell | Z(a_0) = k) \cdot \widehat{\pi}_k. \qquad [18]$$

as an estimate of the prevalence for state ℓ as a function of time. These estimates of marginal state occupancy probabilities can be compared to crude estimates of the proportion of individuals in each state at each assessment time. Figure 4 shows quite good agreement between crude estimates and model-based estimates, providing some empirical validation of the Markov assumption. A time-homogeneous Markov model and a piecewise constant Markov model were fitted, where the latter adopted constant transition intensities over the elementary grades and different constant intensities during high school. The agreement with the crude prevalence estimates was noticeably better for the piecewise constant model and a likelihood ratio test of the time homogeneous assumption against this alternative leads to rejection of the null hypothesis (Control:LRS $= 112.69$, $P(\chi^2_3 > 112.69) < 0.0001$; Intervention:LRS $= 398.56$, $P(\chi^2_3 > 398.56) < 0.0001$).

Proportional intensity models of the form $\lambda_{k\ell}(t|H(t))| = \lambda_{k\ell} \exp(x'(t)\,\beta)$ were also fit to assess the effect of the intervention program and the effect of a social models risk score, which takes the value 1 for children with no parents or friends that smoke, 2 if one parent or friend smokes, and 3 if two or more parents or friends smoke. The social models risk score varies with time and was updated at each assessment when available. The results of fitting a multiplicative intensity model with the treatment indicators (Model 1) and a model with treatment and the time-dependent social models risk (Model 2) are given in Table 2.

There is no evidence of an intervention effect on the transition rates for either model. There is strong evidence, however, of the impact of the social models risk on smoking behavior. Students with a risk score of 2 have a 1.65 higher risk of starting to smoke, while children with a score of 3 have a 2.82 times higher risk. Strong reverse effects are seen on the quit rate. Interestingly, there is insignificant evidence of an effect of social models risk on resumption of smoking.

Further Remarks

Event history models are an invaluable tool in studying life history processes. When full information is available about events or states experienced by individuals in a study, survival analysis software can be employed to fit certain models. In particular, software for the Cox model allows a wide range of semiparametric models to be fitted, including those given here in [3], [7], [14], and [15]. In studies where individuals are seen intermittently, with only the states occupied at those observation times being available, analysis is more difficult. Analysis in this case is usually based on Markov models; see Cook and Lawless (2007) and for software, Jackson (2011). We have commented on the usefulness of models for marginal process features such as rate, mean, and prevalence functions. Estimates based on Markov models for recurrent events and multistate processes have certain robustness properties, provided that the length or frequency of follow-up for individuals is independent of their event processes. When this independence condition is violated, inverse probability weighting methods can be employed to adjust for event history-dependent follow-up in some cases (Cook et al., 2009). A second challenge for marginal analyses is that when processes involve 'terminal' events or states such as death, the specification of marginal event rates and mean functions is more complicated. For discussion and methodology associated with these issues, see Cook et al. (2009).

Understanding of causal relationships between time-varying factors and important lifetime events is explicitly or implicitly the goal in many fields of study. Methods for event history analysis offer a powerful framework for studying dynamic processes and causal factors that influence process dynamics. There are different schools of causal inference but given the focus on modeling phenomena as they unfold over time, event history methods are best suited to answering questions within

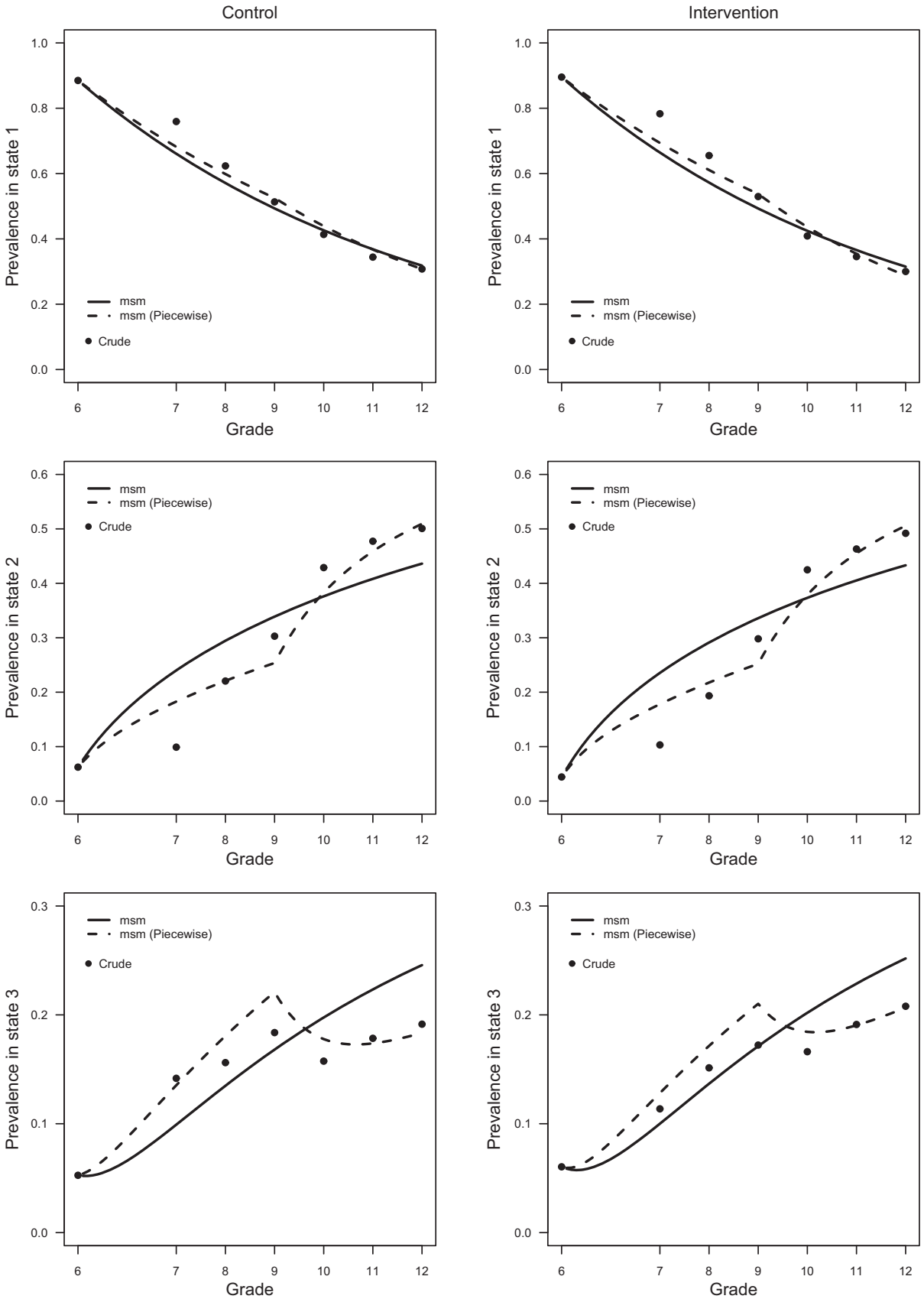

Figure 4 Estimated prevalences in states 1, 2, and 3 for individuals in the control (left) and intervention (right) groups based on a time-homogeneous and a two-piece Markov model fitted using the *msm* function (Jackson, C.H., 2011. Multi-state models for panel data: the msm package for R. Journal of Statistical Software 38, 1–28) and [18], with the corresponding crude estimates represented by dots; the spacing on the x-axis reflects the fact that grade 6 assessments were early in the academic year and each subsequent annual assessment took place at the end of the academic year.

Table 2 Parameter estimates from two Markov regression models for smoking behavior involving intervention and social models risk (SMR); relative risk (RR) is the exponential of the regression coefficient

			Model 1			Model 2		
Transition		Effect	RR	95% CI	p	RR	95% CI	p
1 → 2	Intervention (x_1)	Treatment vs Control	1.02	(0.92,1.12)	0.7091	1.01	(0.92,1.12)	0.7859
	SMR ($x_2(t)$)	2 vs 1				1.65	(1.51,1.80)	<0.0001
	SMR (x_3(t))	3 vs 1				2.82	(2.51,3.17)	<0.0001
2 → 3	Intervention (x_1)	Treatment vs Control	1.02	(0.88,1.18)	0.7663	0.98	(0.84,1.14)	0.7767
	SMR ($x_2(t)$)	2 vs 1				0.73	(0.63,0.86)	0.0002
	SMR ($x_3(t)$)	3 vs 1				0.34	(0.29,0.40)	<0.0001
3 → 2	Intervention (x_1)	Treatment vs Control	0.99	(0.83,1.17)	0.9001	0.97	(0.81,1.15)	0.7308
	SMR ($x_2(t)$)	2 vs 1				0.98	(0.81,1.19)	0.8445
	SMR ($x_3(t)$)	3 vs 1				0.89	(0.73,1.08)	0.2349

the Granger–Schweder framework (Granger, 1969; Schweder, 1970); see Chapter 9 of Aalen et al. (2008) for an excellent discussion. The remarks in Section Multitype Recurrent Events on studies of local dependence through the use of intensity-based models are a brief allusion to this approach.

We have emphasized the use of intensity-based models with multiplicative covariate effects, but as in standard survival analysis, covariates may be studied on the basis of time-transform models (Cheng et al., 1995), additive models (Aalen et al., 2001), or hybrid models with additive and multiplicative effects such as the Cox–Aalen model (Martinussen and Scheike, 2006). The timereg package in R facilitates the fitting of quite general models having time-dependent covariates with time-varying effects, and this facility proves particularly helpful in problems involving causal inferences (see Aalen et al., 2008, Chapter 9).

In some situations individuals or processes may be grouped into clusters and in that case it may be important to account for intracluster association. This can be done in ways analogous to those in Section Multitype Recurrent Events, either by the introduction of cluster-specific random effects (Cook et al., 2004; Sutradhar and Cook, 2008) or by the development of robust variance estimation for marginal methods. For robust methods involving time to event data see Lee et al. (1992), for recurrent events see Cook and Lawless (2007; Sec. 8.5 and 8.9) and references therein, and for multistate models see Lee and Kim (1998). Finally, in some settings the selection or follow-up of individuals in a study may be related to their event histories. Methods for adjusting for selection effects are then needed; see Cook and Lawless (2007, Chapter 7) and Cook and Lawless (2014; Sec. 6).

Acknowledgments

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada to RJC (Grant No. 101093) and JFL (Grant No. 103167) and an operating grant from the Canadian Institutes for Health Research to RJC (Grant No. 105099). Richard Cook is Canada Research Chair in Statistical Methods for Health Research. Data on hospitalizations in affective disorder were kindly provided by Professor Lars Vedel Kessing and Professor Per Kragh Andersen who have used the data in prior publications on recurrent events. We thank Professors Steve Brown and Roy Cameron for permission to use the data from the Waterloo Smoking Prevention Project, which was funded by NHLBI (U.S.) and NHRDP (Canada). Assistance with the statistical computing from Ker-Ai Lee is also gratefully acknowledged.

See also: Data Bases and Statistical Systems: Health; Event History Analysis: Applications; Generalized Linear Mixed Models; Hierarchical Models: Random and Fixed Effects; Longitudinal Causal Inference; Longitudinal Data: Event–History Analysis in Discrete Time; Markov Models and Social Analysis; Semiparametric Models; Stochastic Models; Survival Analysis: Introduction; Survival Analysis: Overview.

Bibliography

Aalen, O.O., Borgan, Ø., Fekjær, H., 2001. Covariate adjustment of event histories estimated from Markov chains: the additive approach. Biometrics 57, 993–1001.

Aalen, O.O., Borgan, Ø., Gjessing, H.K., 2008. Survival and Event History Analysis: A Process Point of View. Springer, New York.

Aalen, O.O., Husebye, E., 1991. Statistical analysis of repeated events forming renewal processes. Statistics in Medicine 10, 1227–1240.

Andersen, P.K., Borgan, O., Gill, R.D., Keiding, N., 1993. Statistical Methods Based on Counting Processes. Springer, New York.

Andersen, P.K., Gill, R.D., 1982. Cox's regression model for counting processes: a large sample study. The Annals of Statistics 10, 1100–1120.

Andersen, P.K., Keiding, N., 2002. Multi-state models for event history analysis. Statistical Methods in Medical Research 11, 91–115.

Cameron, R., Brown, K.S., Best, J.A., Pelkman, C.L., Madill, C.L., Manske, S.R., Payne, M.E., 1999. Effectiveness of a social influences smoking prevention program as a function of provider type, training method, and school risk. American Journal of Public Health 89, 1827–1831.

Cheng, S.C., Wei, L.J., Ying, Z., 1995. Analysis of transformation models with censored data. Biometrika 82, 835–845.

Cook, R.J., Lawless, J.F., 2007. Statistical Analysis of Recurrent Events. Springer, New York.

Cook, R.J., Lawless, J.F., 2013. Concepts and tests for trend in recurrent event processes. Journal of Iranian Statistical Society 12, 35–69.

Cook, R.J., Lawless, J.F., 2014. Statistical issues in modeling chronic disease in cohort studies. Statistics in Biosciences 6, 127–161.

Cook, R.J., Lawless, J.F., Lakhal-Chaieb, L., Lee, K.-A., 2009. Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: application to skeletal complications in cancer metastatic to bone. Journal of the American Statistical Association 104, 60–75.

Cook, R.J., Lawless, J.F., Lee, K.-A., 2010. A copula-based mixed Poisson model for bivariate recurrent events under event-dependent censoring. Statistics in Medicine 29, 694–707.

Cook, R.J., Yi, G.Y., Lee, K.-A., Gladman, D.D., 2004. A conditional Markov model for clustered progressive multistate processes under incomplete observation. Biometrics 60, 436–443.

Cox, D.D.R., Miller, H.D., 1977. The Theory of Stochastic Processes, vol. 134. Chapman and Hall, London.

Cox, D.R., 1972. Regression models and life tables (with discussion). Journal of the Royal Statistical Society. Series B (Methodological) 34, 187–220.

Daley, D.J., Vere-Jones, D., 2007. An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure, vol. 2. Springer, New York.

Grandell, J., 1997. Mixed Poisson Processes. Chapman and Hall, London.

Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society 37, 424–438.

Jackson, C.H., 2011. Multi-state models for panel data: the msm package for R. Journal of Statistical Software 38, 1–28.

Kalbfleisch, J.D., Prentice, R.L., 2002. The Statistical Analysis of Failure Time Data, second ed. John Wiley and Sons, New York.

Kessing, L.V., Hansen, M.G., Andersen, P.K., 2004. Course of illness in depressive and bipolar disorders. British Journal of Psychiatry 185, 372–377.

Lawless, J.F., 2003. Statistical Models and Methods for Lifetime Data, second ed. John Wiley and Sons, Inc., Hoboken, NJ.

Lawless, J.F., Zhan, M., 1998. Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. Canadian Journal of Statistics 26, 549–565.

Lee, E.W., Kim, M.Y., 1998. The analysis of correlated panel data using a continuous-time Markov model. Biometrics 54, 1638–1644.

Lee, E.W., Wei, L.J., Amato, D.A., Leurgans, S., 1992. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In: Klein, J.P., Goel, P.k. (Eds.), Survival Analysis: State of the Art. Springer, Netherlands, pp. 237–247.

Lin, D.Y., Wei, L.J., Yang, I., Ying, Z., 2000. Semiparametric regression for the mean and rate functions of recurrent events. Journal of the Royal Statistical Society. Series B (Methodological) 62, 711–730.

Martinussen, T., Scheike, T.H., 2006. Dynamic Regression Models for Survival Data. Springer, New York.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman and Hall, London.

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Austria, Vienna, ISBN 3-900051-07-0. http://www.R-project.org/.

Ross, S.M., 2006. Introduction to Probability Models. Academic Press, Orlando, FL.

SAS Institute Inc., 2008. SAS/STAT® 9.2 User's Guide. SAS Institute Inc., Cray, NC.

Schweder, T., 1970. Composable Markov processes. Journal of Applied Probability 7, 400–410.

Skardhamer, T., Telle, K., 2012. Post-release employment and recidivism in Norway. Journal of Quantitative Criminology 28, 629–649.

Sun, J., Zhao, X., 2013. Statistical Analysis of Panel Count Data. Springer, New York.

Sutradhar, R., Cook, R.J., 2008. Analysis of interval-censored data from clustered multistate processes: application to joint damage in psoriatic arthritis. Journal of the Royal Statistical Society. Series C (Applied Statistics) 57, 553–566.

Therneau, T.M., Grambsch, P.M., 2000. Modeling Survival Data: Extending the Cox Model. Springer, New York.

TIBCO Spotfire S+®, 2008. 8.1 Guide to Statistics Volume 2. TIBCO Software Inc., Palo Alto, CA.

Wei, L.-J., Lin, D.Y., Weissfeld, L., 1989. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. Journal of the American Statistical Association 84, 1065–1073.

Event History Analysis: Applications

Hans-Peter Blossfeld, University of Bielefeld, Bielefeld, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 7, pp. 4941–4946, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Event history analyses investigate collections of individuals or units moving between different observed states. This article discusses applications of this type of analysis. It first distinguishes different ways of collecting data, contrasting event-oriented observations. Then, it details basic terminology and the central concept of the transitional rate. In conclusion, it discusses the inclusion of time-varying covariates in transition rate models and describes three successful applications of event history analysis in different substantive contexts.

'Event history analysis' is the study of processes that are characterized in the following general way: (1) there is a collection of units (which may be individuals, organizations, societies, etc.), each moving among a finite (usually small) number of states; (2) these changes (or events) may occur at any point in time (i.e., they are not restricted to predetermined points in time); and (3) there are time-constant and/or time-varying factors influencing the timing of events. In this article, various examples of event history applications are presented, the advantages and problems of different observation plans are discussed, the terminology and central concepts of event history analysis (e.g., transition rate or censoring) are demonstrated, and practical complications of statistical models (e.g., exponential, piecewise constant, Cox, Gompertz–Makeham, Weibull, Sickle, Log-logistic, Log-normal, or Gamma model) are addressed.

'Event history analysis' studies a collection of units, each moving among a finite (usually small) number of states. An example is individuals who are moving from unemployment to employment. In this article, various examples of event history applications are presented, the concepts and advantages of event history analysis are demonstrated, and practical complications are addressed.

Examples of Continuous-Time, Discrete-State Processes

Event history analysis is the study of processes that are characterized in the following general way: (1) there is a collection of units (which may be individuals, organizations, societies, etc.), each moving among a finite (usually small) number of states; (2) these changes (or events) may occur at any point in time (i.e., they are not restricted to predetermined points in time); and (3) there are time-constant and/or time-varying factors influencing the timing of events. Examples are workers who move between unemployment and employment; men and women who enter into consensual unions or marriages; companies that are founded or closed down; governments that break down; people who are mobile between different regions or nation-states; consumers who switch from one brand to another; prisoners who are released and commit another crime; people who show certain types of verbal and nonverbal behaviors in social interactions; students who drop out of school; incidences of racial and ethnic confrontation, protest, riot, and attack; people who show signs of psychoses or neuroses; patients who switch between the states 'healthy' and 'diseased,' and so on. In event history analysis, the central characteristics of the underlying stochastic process are mirrored in the specific way theoretical and mathematical models are built, data are collected, and the estimation and evaluation of models is done.

Different Observation Plans

In event history analysis, the special importance of broader research design issues has been stressed. In particular, different observation plans have been used to collect information on the continuous-time, discrete-state substantive process. These various schemes produce different types of data that constrain the statistical analysis in different ways: cross-sectional data, event count data, event sequence data, panel data, and event history data.

The cross-sectional observation design is still the most common form of data collection in many fields. A cross-sectional sample is only a 'snapshot' taken on the continuous-time, discrete-state substantive process (e.g., the jobs of people at the time of the interview). Event history analysis has shown that one must be very cautious in drawing inferences about explanatory variables on the basis of such data, because, implicitly or explicitly, social researchers have to assume that the process under study is in some kind of equilibrium. Equilibrium means that the state probabilities are fairly trendless and the relationships among the variables are quite stable over time. Therefore, an equilibrium of the process requires that the inflows to and the outflows from each of the discrete states be equal over time to a large extent. This is a strong assumption that is not very often justified in the social and behavioral sciences (see Blossfeld and Rohwer, 1995).

A comparatively rare type of data on changes of the process are event count data (Andersen et al., 1993). They record the number of different types of events for each unit in an interval (e.g., the number of racial and ethnic confrontations, protests, riots, or attacks in a period of 10 years). Event sequence data provide even more information. They record the sequence of

specific states occupied by each unit over time (e.g., *see* Event Sequence Analysis: Methods). The temporal data most often available to the scientist are panel data. They provide information on the same units of analysis at a series of discrete points in time. In other words, there is only information on the states of the units at predetermined survey points, and the course of the process between the survey points remains unknown. Panel data certainly contain more information than cross-sectional data, but involve well-known distortions created by the method itself (e.g., panel bias, attrition of sample). In addition, causal inferences based on panel approaches are much more complicated than has been generally acknowledged (Blossfeld and Rohwer, 1995). With regard to the continuous-time, discrete-state substantive process, panel analysis is particularly sensitive to the length of the time intervals between the waves relative to the speed of the process. They can be too short, so that too few state transitions will be observed; or too long, so that it is difficult to establish a time order between the events (Allison, 2014).

The event-oriented observation design records the complete sequence of states occupied by each unit and the timing of changes among these states. For example, an event history of job mobility consists of more or less detailed information about each of the jobs and the exact beginning and ending dates of each job. Thus, this is the most complete information one can get on the continuous-time, discrete-state substantive process. Such event history data are often collected retrospectively with life history studies. Life history studies are normally cheaper than panel studies and have the advantage that they code the data into one framework of codes and meaning. But retrospective studies also suffer from several limitations (see Blossfeld and Rohwer, 1995). In particular, data concerning motivational, attitudinal, cognitive, or affective states are difficult (or even impossible) to collect retrospectively because the respondents can hardly recall the timing of changes in these states accurately. Also, the retrospective collection of behavioral data has a high potential for bias because of its strong reliance on autobiographic memory. To reduce these problems of data collection, modern panel studies (e.g., the Panel Study of Income Dynamics (PSID) in the United States, the British Household and Panel Survey (BHPS) in the UK, or the Sozio-oekonomisches Panel (SOEP) in Germany) use a mixed design, which provides traditional panel data and retrospectively collects event history data for the period before the first panel wave and between the successive panel waves.

Event History Analysis

Event history analysis, originally developed as independent applications of mathematical probability theory in demography (e.g., the classical life table analysis and the product-limit estimator of Kaplan and Meier (1958)), reliability engineering, and biostatistics (e.g., the path-breaking semiparametric regression model for survival data of David Cox (1972)) has gained increasing importance in the social and behavioral sciences (see Tuma and Hannan, 1984; Blossfeld et al., 1989; Lancaster, 1990; Andersen et al., 1993; Blossfeld and Rohwer, 1995; Broström, 2012; Blossfeld et al., 2007) since the early 1980s. Due to the development and application of these methods in various disciplines, the terminology is normally not easily accessible to the user. Central definitions are therefore given first.

Basic Terminology

Event history analysis studies transitions across a set of discrete states, including the length of time intervals between entry to and exit from specific states. The basic analytical framework is a state space and a time axis. The choice of the time axis or clock (e.g., age, experience, marriage duration, etc.) used in the analysis must be based on theoretical considerations and affects the statistical model. Dependent on practical and theoretical reasons, there are event history methods using a discrete (Yamaguchi, 1991; Vermunt, 1997) or continuous time axis (Tuma and Hannan, 1984; Blossfeld et al., 1989; Courgeau and Lelièvre, 1992; Blossfeld and Rohwer, 1995). Discrete-time event history models are, however, special cases of continuous-time models and are therefore not further discussed here. An episode, spell, waiting time, or duration – terms that are used interchangeably in the literature – is the time span a unit of analysis (e.g., an individual) spends in a specific state. The states are discrete and usually small in number. The definition of a set of possible states, called the state space, is also dependent on substantive considerations. Thus, a careful, theoretically driven choice of the time axis and design of state space are crucial, because they are often serious sources of misspecification.

The most restricted event history model is based on a process with only a single episode and two states (an origin state j and a destination state k). An example may be the duration of first marriage until the end of the marriage, for whatever reason (separation, divorce, death). In the single episode case, each unit of analysis that entered into the origin state is represented by one episode. Event history models are called multistate models, if more than one destination state exists. Models for the special case with a single origin state but two or more destination states are also called models with competing events or risks. For example, a housewife might become 'unemployed' (meaning entering into the state 'looking for work'), or start being 'full-time' or 'part-time employed.' If more than one event is possible (i.e., if there are repeated events or transitions over the observation period), the term multiepisode model is used. For example, an employment career normally consists of a series of job shifts.

Censoring

Observations of event histories are often censored (Keiding, 2014). Censoring occurs when the information about the duration in the origin state is incompletely recorded. An episode is fully censored on the left, if starting and ending times of a spell are located before the beginning of an observation period (e.g., before the first panel wave), and it is partially censored on the left, if the length of time a unit has already spent in the origin state is unknown. This is typically the case in panel studies, if individuals' job episodes at the time of a first panel wave is known but no further information about the history of the job is collected. Left censoring is normally a difficult problem because it is not possible to take the effects

of the unknown episodes into account. It is only without problems, if the assumption of a Markov process is justified (i.e., if the transition rates do not depend on the duration in the origin state).

The usual kind of censoring, however, is right censoring. In this case, the end of the episode is not observed but the observation of the episode is terminated at an arbitrary point in time. This type of censoring typically occurs in life course studies at the time of the retrospective interviews or in panel studies at the time of the last panel wave. Because the timing of the end of the observation window is normally determined independently from the substantive process under study, this type of right censoring is unproblematic and can easily be handled with event history methods (Kalbfleisch and Prentice, 1980 – type I censoring). Finally, episodes might be completely censored on the right. In other words, entry into and exit from the duration occurs after the observation period. This type of censoring normally happens in retrospective life history studies in which individuals of various birth cohorts can only be observed over very different spans of life. To avoid sample selection bias, such models have to take into account variables controlling for the selection; for example, by including birth cohort dummy variables (see Yamaguchi, 1991).

The Transition Rate

The central concept of event history analysis is the transition rate. Because of the various origins of event history analysis in the different disciplines, the transition rate is also called the hazard rate, intensity rate, failure rate, transition intensity, risk function, or mortality rate:

$$r(t) = \lim_{t' \to t} pr(t \leq T < t' | T \geq t)/(t' - t) \qquad [1]$$

The transition rate provides a local, time-related description of how the process evolves over time. It can be interpreted as the propensity (or intensity) to change from an origin state to a destination state, at time t. But one should note that this propensity is defined in relation to a risk set ($T \geq t$) at t, i.e., the set of units that still can experience the event because they have not yet had the event before t.

Statistical Models

The central idea in event history analysis is to make the transition rate, which describes a process evolving in time, dependent on time (t) and on a set of covariates, x:

$$r(t) = g(t, x) \qquad [2]$$

The causal interpretation (*see* Explanation: Conceptions in the Social Sciences) of the transition rate requires that we take the temporal order in which the processes evolve very seriously. In other words, at any given point in time, t, the transition rate $r(t)$ can be made dependent on conditions that happened to occur in the past (i.e., before t), but not on what is the case at t or in the future, after t.

There are several possibilities to specify the functional relationship g (see Blossfeld and Rohwer, 1995). For example, the exponential model, which normally serves as a baseline model, assumes that the transition rate can vary with different constellations of covariates x, but that the rates are time-constant. The piecewise constant exponential model allows the transition rate to vary across fixed time periods with period-specific effects of covariates. There are also different parametric models, which are based on specific shapes of the time-dependence (e.g., the Gompertz-Makeham, Weibull, Sickle, Log-logistic, Log-normal, or Gamma model). If the time-shape of the baseline hazard rate is unspecified and only possible effects of covariates x are estimated, the model is called a semi-parametric or partial likelihood model (Cox, 1972). Finally, there are also models of unobserved heterogeneity in which the transition rate can be made dependent on the observed covariates x, the duration t, and a stochastic error term.

Parallel and Interdependent Processes

The most important scientific progress permitted by event history analysis is based on the opportunity to include explicitly measured time-varying covariates in transition rate models (see Blossfeld and Rohwer, 1995). These covariates can change their values over process time in an analysis. Time-varying covariates can be qualitative or quantitative, and may stay constant for finite periods of time or change continuously. From a substantive point of view, time-varying covariates can be conceptualized as observations of the sample path of parallel processes. These processes can operate at different levels. In sociology, for example, (1) there can be parallel processes at the level of the individual's different domains of life (e.g., one may ask how upward and downward moves in an individual's job career influence his/her family trajectory); (2) there may be parallel processes at the level of some few individuals interacting with each other (e.g., one might study the effect of the career of the husband on his wife's labor force participation); (3) there may be parallel processes at the intermediate level (e.g., one can analyze how organizational growth influences career advancement or changing household structure determines women's labor force participation); (4) there may be parallel processes at the macro level (e.g., one may be interested in the effect of changes in the business cycle on family formation or career advancement); and (5) there may be any combination of such processes of type (1) to (4). For example, in the study of life-course, cohort and period effects, time-dependent covariates at different levels (see below) must be included simultaneously (multilevel analysis).

In dealing with such systems of parallel processes, the issue of reverse causation is often addressed in the methodological literature (see, e.g., Kalbfleisch and Prentice, 1980; Tuma and Hannan, 1984; Blossfeld et al., 1989; Yamaguchi, 1991; Courgeau and Lelièvre, 1992). Reverse causation refers to the (direct or indirect) effect of the dependent process on the independent covariate process(es). Reverse causation is often seen as a problem because the effect of a time-dependent covariate on the transition rate is confounded with a feedback effect of the dependent process on the values of the time-dependent covariate. However, Blossfeld and Rohwer (1995) have developed a causal approach to the analysis of interdependent processes that also works in the case of interdependence, too. For example, if two interdependent processes, Y_t^A and Y_t^B, are given, a change in Y_t^A at any (specific) point in time

t' may be modeled as being depend on the history of both processes up to, but not including t'. Or stated in another way: what happens with Y_t^A at any point in time t' is conditionally independent of what happens with Y_t^B at t', conditional on the history of the joint process $Y_t = (Y_t^A, Y_t^B)$ up to, but not including, t ('principle of conditional independence'). Of course, the same reasoning can be applied if one focuses on Y_t^B instead of Y_t^A as the 'dependent variable.' Beginning with a transition rate model for the joint process, $Y_t = (Y_t^A, Y_t^B)$, and assuming the principle of conditional independence, the likelihood for this model can then be factorized into a product of the likelihoods for two separate models: a transition rate model for Y_t^A, which is dependent on Y_t^B as a time-dependent covariate, and a transition rate model for Y_t^B, which is dependent on Y_t^A as a time-dependent covariate. From a technical point of view, there is therefore no need to distinguish between defined, ancillary, and internal covariates (see, e.g., Kalbfleisch and Prentice, 1980), because all of these time-varying covariate types can be treated in the estimation procedure. Estimating the effects of time-varying processes on the transition rate can easily be achieved by applying the method of episode splitting (see Blossfeld and Rohwer, 1995).

Unobserved Heterogeneity

Unfortunately, researchers are not always able to include all the important covariates into an event history analysis. These unobserved differences between subpopulations can lead to apparent time-dependence at the population level and additional identification problems. There exists a fairly large literature on misspecified models in general. In particular, there have been several proposals to deal with unobserved heterogeneity in transition rate models (the so-called frailty models) (see, e.g., Blossfeld and Rohwer, 1995).

Illustrative Applications

The following three applications illustrate the technical strength of event history models in different substantive contexts.

A Multilevel Analysis of Life-Course, Cohort, and Period Effects in Career Mobility

Any standard statistical analysis using age, year of birth (as a cohort measure), and chronological year of measurement (as a period measure) leads to an identification problem (age = year of measurement − year of birth). The technical solution to that problem often is to assume that at least one of these measures is constant over time or to apply arbitrary statistical constraints. Blossfeld (1986) in an event history study of career mobility of German men approached this problem from a more substantive point of view. He suggested to rely more on explicit measures for the theoretically important macro effects of cohort and period and used 14 time series from official statistics indicating the long-term development of the labor market structure in Germany. Based on factor analysis, he first reduced the 14 highly correlated time series to two orthogonal macro factors, which represented the 'monotonic trend in the level of modernization' and the 'cyclic changes in labor market conditions,' and then matched these macro data to individual event career data to conduct a multilevel analysis. To represent the changing conditions under which cohorts enter the labor market, Blossfeld used the factor scores of these two factors at the year men entered the labor force. To introduce period as a time-dependent covariate, he used the method of episode splitting. In accordance with this method, the original job episodes were split into subepisodes every year so that the factor scores of the two macro factors could be updated in each job episode for all employed men every year. Modeling cohort and period effects this way not only provides more direct and valid measures of the explanatory concepts, but also results in the disappearance of the problem of nonestimable parameters above. It was, therefore, not necessary to impose (in substantive terms normally unjustifiable) constraints on the parameters to identify life-course, cohort, and period effects.

Parallel Processes at the Level of the Individual: Women's Rising Educational and Career Investments, and Changes in the Process of Family Formation

Using event history models, Blossfeld and Huinink assessed the hypothesis of the 'new home economics' that women's rising educational and career investments will lead to a decline in marriage and motherhood (see Blossfeld and Rohwer, 1995: 143). Because the accumulation of educational and career resources is a lifetime process, they analyzed it as a dynamic process over the life course. To model the changes in the accumulation of qualifications over the life course, Blossfeld and Huinink updated a time-varying covariate 'level of education' at the age when a woman obtained a particular educational rank in the hierarchy. They also generated a time-dependent dummy variable indicating whether or not a woman is attending the educational system at a specific age. Finally, for the dynamic modeling of job-specific investments in human capital over the life course, Blossfeld and Huinink have developed a specific time-dependent labor force experience measure, which they called 'career resources.' The time-dependent increments and final levels of this measure were designed to reflect the 'quality of each job' and the respective 'employment durations.' The time-dependent measure also took into account that career resources are declining when women interrupted their employment career. Blossfeld and Huinink applied the method of episode splitting to include these various time-dependent dimensions of education and career together with additional control variables in the rate equation. The results of their longitudinal analysis were surprising: contrary to the expectations of the economic approach to the family, women's entry into marriage obviously was independent of women's level of education and their career resources. Equally surprisingly, the effect of women's level of education on the rate of their first birth was even positive. An explanation for the positive effect on the timing of first birth is easily demonstrated with an event history model: women who remain in school longer and attain high qualifications are first shifting their fertility decisions to higher ages, and then, they are subject to pressure not only from the

potential increase in medical problems connected with having children late, but also from societal age norms. Thus the higher the educational attainment level, the stronger the catching-up effect. This application demonstrates that the standard empirical support for the economic theory of the family claimed on the basis of cross-sectional data might be misleading. Cross-sectional data do not permit a differentiation between the time-dependent effect of accumulation of human capital over the life course and the time-dependent effect of participation in the educational system.

A Causal Approach to Interdependent Systems: The Effects of Pregnancy on Entry into Marriage for Couples Living in Consensual Unions

Finally, the application of a causal approach to interdependent systems is demonstrated by Blossfeld and Rohwer in an analysis of the effect of pregnancy on entry into marriage in Germany (see Blossfeld and Rohwer, 1995: 157). This study shows that the effect of a fertility event on entry into marriage is highly time dependent: It starts to rise at some time shortly after conception (thus there is some time lag involved between cause and effect), increases during pregnancy to a maximum at around 7–8 months after conception, and then decreases again reaching a level even below the starting level at around 6 months after the birth of the child. Thus, the dynamic interdependence of fertility and nuptiality mainly occurs in a very specific phase of the life course and the unfolding effect of fertility on nuptiality is highly dynamic over time. Of course, standard statistical applications based on cross-sectional observation can be hardly used to analyze such situations in a meaningful way. However, in event history analysis, there is an easy way to approach that problem. One can use a series of time-dependent dummy variables each indicating a month since the occurrence of conception (the causal event). They show that the effects of the time-dependent dummy variable at first increase, reach a maximum, and then decrease again. Thus, starting with conception, there is an increasing normative pressure to avoid an illegitimate birth that increases the marriage rate, particularly for people who are already 'ready for marriage.' But with an increasing number of marriages, the composition of the group of still unmarried couples shifts toward couples being 'less ready for marriage' or being 'not ready for marriage,' which, of course, decreases the fertility effect again. This is an important result not only in substantive terms; the methodological lesson for event history analysis is also revealing. One should be very careful in modeling qualitative parallel processes with only one dummy variable (e.g., in the current application using only the time of conception, or the time of birth for the switch of the time-dependent covariate) because the effect itself might be highly time-dependent. For example, Blossfeld and Rohwer show that the fertility effect disappears completely, when the time-dependent covariate is switched at the time of birth because then a period with a low-marriage rate up to the time of discovery of conception and a period with a high-marriage rate during pregnancy is aggregated and compared with a period of low rate after the birth of the child.

See also: Event History Analysis; Longitudinal Data: Event–History Analysis in Discrete Time.

Bibliography

Allison, P.D., 2014. Event History and Survival Analysis: Regression for Longitudinal Event Data. Series: Quantitative Applications in the Social Sciences. SAGE Publications, Thousand Oaks, CA.

Andersen, P.K., Borgan, Oslash, Gill, R.D., Keiding, N., 1993. Statistical Models Based on Counting Processes. Springer-Verlag, New York.

Broström, G., 2012. Event History Analysis with R. CRC Press, New York.

Blossfeld, H.-P., 1986. Career opportunities in the Federal Republic of Germany: a dynamic approach to the study of life-course, cohort, and period effects. European Sociological Review 2, 208–225.

Blossfeld, H.-P., Hamerle, A., Mayer, K.U., 1989. Event History Analysis. Erlbaum, Hillsdale, NJ.

Blossfeld, H.-P., Rohwer, G., 1995. Techniques of Event History Modeling. New Approaches to Causal Analysis. Erlbaum, Mahwah, NJ.

Blossfeld, H.-P., Golsch, K., Rohwer, G., 2007. Event History Analysis with Stata. Lawrence Erlbaum, New York.

Courgeau, D., Lelièvre, E., 1992. Event History Analysis in Demography. Clarendon, Oxford.

Cox, D.R., 1972. Regression models and life-tables. Journal of the Royal Statistics Society 34, 187–220.

Kalbfleisch, J.D., Prentice, R.L., 1980. The Statistical Analysis of Failure Time Data. Wiley, New York.

Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53, 457–481.

Keiding, N., 2014. Event history analysis. Annual Review of Statistics and its Application 1, 333–360.

Lancaster, T., 1990. The Econometric Analysis of Transition Data. Cambridge University Press, Cambridge, UK.

Tuma, N.B., Hannan, M.T., 1984. Social Dynamics. Models and Methods. Academic Press, Orlando, FL.

Vermunt, J.K., 1997. Log-Linear Models for Event Histories. Thousands Oaks, Newbury Park, CA.

Yamaguchi, K., 1991. Event History Analysis. Sage, Newbury Park, CA.

Event History Analysis in Continuous Time

Per K Andersen and Niels Keiding, University of Copenhagen, Copenhagen, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Event history data are obtained by observing individuals over time, focusing on times of occurrence of certain events and the types of event occurring. A review is given of event history analysis (in continuous time) based on multistate models. Examples of such models include the two-state model for survival data and the competing risks and disability models. The likelihood function for a multistate model is presented using the theory of counting processes with special emphasis on models with piecewise constant transition intensities, nonparametric models, and models for the transition intensities including covariates along the lines of the semiparametric Cox proportional hazards regression model for survival data. Direct models for various marginal features of event history data, such as state occupation probabilities, are also discussed. Finally, the influence of observational patterns on the inference is discussed.

Introduction

Event history analysis deals with data obtained by observing individuals over time focusing on events occurring for the individuals. Thus, typical outcome data consist of *times of occurrence* of events and of the *types of events* which occurred. Frequently, an event may be considered as a *transition* from one state to another and, therefore, *multistate models* will often provide a relevant modeling framework for event history data. Multistate models are discussed from several points of view in the books by Andersen et al. (1993), Kalbfleisch and Prentice (2002), Martinussen and Scheike (2006), Aalen et al. (2008), Blossfeld and Rohwer (1995), and Courgeau and Leliévre (1992), see also survey papers by Andersen and Keiding (2002), Andersen and Pohar Perme (2008), and Keiding (2014).

Survival Data

The simplest multistate model is the two-state model for survival data with one initial state 0: alive and one final state 1: dead, see **Figure 1**. The observation for a given individual will here in the simplest form consist of a random variable, say T, representing the time from a given origin (time 0) to the occurrence of the event death. The distribution of T may be characterized by the probability distribution function $F(t) = \text{Prob}(T \leq t)$ or, equivalently, by the survival distribution function $S(t) = 1 - F(t) = \text{Prob}(T > t)$. It is seen that $S(t)$ and $F(t)$, respectively, correspond to the probabilities of being in state 0 or 1 at time t. If every individual is assumed to be in state 0 at time 0 then $F(t)$ is also the *transition probability* from state 0 to state 1 for the time interval from 0 to t. In continuous time the distribution of T may also be characterized by the *hazard rate function*.

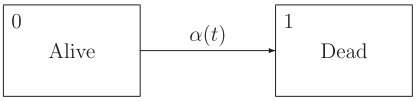

Figure 1 The two-state model for survival data.

$$\alpha(t) = -\mathrm{d}\log S(t)/\mathrm{d}t = \lim_{\Delta t \to 0} \frac{\text{Prob}(T \leq t + \Delta t | T \geq t)}{\Delta t}$$

that is

$$S(t) = \exp\left(-\int_0^t \alpha(u)\mathrm{d}u\right)$$

Thus, $\alpha(\cdot)$ is the *transition intensity* from state 0 to state 1, that is, the instantaneous probability per time unit of going from state 0 to state 1. The intensity of going from state 1 to state 0 is 0 and state 1 is then said to be *absorbing*, defined as a state from which further transitions cannot occur. A *transient* state is a state which is not absorbing and in this simple two-state model state 0 is transient.

In general, event history analysis deals with inference for transition intensities and transition or state occupation probabilities in multistate models. This includes estimation and hypothesis tests for these quantities and analysis of regression models where these quantities are related to (possibly time-dependent) explanatory variables observed for the individuals under study. Most frequently, multistate models are defined by their transition intensities from which transition or state occupation probabilities may be derived by plug-in or by micro-simulation. However, sometimes the main *target of inference* for event history analysis is different and, for example, direct, *marginal* models for probabilities may be studied. Examples are provided below.

A typical feature of event history analysis is the inability to observe complete event histories, for example, by the end of the observation period all individuals under study may not have reached an absorbing state. In survival analysis this would correspond to individuals still being alive by the end of the study and this kind of *incomplete observation* is known as *right-censoring*. Furthermore, all individuals may not have been observed from the same time origin. This kind of incomplete observation where individuals are only observed conditionally on not having reached an absorbing state by the time of initiation of the study is known as *left-truncation*. Restricting attention to right-censoring a crucial problem is whether the available incomplete data enables one to make valid inference

on parameters in the multistate model for the complete data. The condition for this is known as *independent* right-censoring and the interpretation is that a sample observed after independent right-censoring is representative for the population without censoring. This means that individuals who are censored should have neither lower nor higher risk of future events than individuals who are not censored (see, e.g., Andersen et al., 1993: Chapter III; Kalbfleisch and Prentice, 2002: p. 194).

Right-censoring may be caused by a number of different mechanisms. Suppose, for example, that one is interested in the disease-specific mortality for patients suffering from a given disease like skin cancer and followed from their times of diagnosis (time 0). Then right-censoring may be caused by

1. a patient still being alive by the time of closure of the study,
2. a patient being lost to follow-up,
3. a patient dying from causes other than skin cancer.

In case (1) there are usually no reasons to believe that censoring should not be independent since the time of closure of a study can most often be assumed independent of the process observed. In case (2) one may be suspicious that patients lost to follow-up could be individuals who generally feel well (or the opposite) and, therefore, do not want to take part in a study any more and this could lead to *dependent* censoring. Situation (3) is more tricky since here the population without censoring is quite hypothetical: it is hardly reasonable to think of a population of skin cancer patients for whom it is impossible to die from other causes. This fact has consequences for how to choose an appropriate model in this situation and a relevant model to consider could be a *competing risks model*, see Figure 2.

The problem is that even though it may be possible to make valid inference for the single cause-specific hazard $\alpha_1(t)$ of dying from skin cancer by using methods for survival data and treating deaths from other causes as right-censored observations then, to evaluate the survival probability for skin cancer patients, the intensity of dying from other causes is also needed. In Section Multistate Models we return to the relationship between the two cause-specific hazards and the transition probabilities in Figure 2. The conclusion is that a *partial model specification* where the hazard, $\alpha_2(t)$, of dying from other causes is not modeled may suffice for some, but not all, purposes and a multistate model like the one depicted in Figure 2 can be a sensible model to study in spite of the fact that focus may be on a single transition. Partial model specification is further discussed in Section Statistical Model Specification.

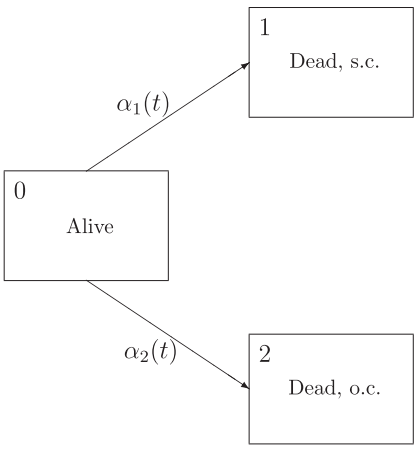

Figure 2 Competing risks model for skin cancer (s.c.) mortality (o.c., other causes).

Multistate Models

In this section, we will present a number of different multistate models including the two-state model for survival data and the competing risks model introduced in Section Survival Data. We will begin with a few definitions. Some of the definitions are formulated loosely but they may all be made rigorous within the framework of so-called *marked point processes* (Andersen et al., 1993: Chapter III; Arjas and Haara, 1984).

A multistate process is a stochastic process $(X(t), t \in \mathcal{T})$ with a finite state space $\mathcal{S} = \{1, \ldots, p\}$ and with right-continuous sample paths: $X(t+) = X(t)$. Here, $\mathcal{T} = [0, \tau]$ or $[0, \tau]$ with $\tau \leq +\infty$. The process has *initial distribution* $\pi_h(0) = \text{Prob}(X(0) = h)$, $h \in \mathcal{S}$. A multistate process $X(\cdot)$ generates a *history* $\mathcal{X}_t$ consisting of the observation of the process in the interval $[0, t]$. Relative to this history we may define *transition probabilities* by

$$P_{hj}(s,t) = \text{Prob}(X(t) = j | X(s) = h, \chi_{s-})$$

for $h, j \in \mathcal{S}$, $s, t \in \mathcal{T}$, $s \leq t$, and for $j \neq h$ *transition intensities* by the derivatives

$$\alpha_{hj}(t) = \lim_{\Delta t \to 0} P_{hj}(t, t + \Delta t)/\Delta t$$

which we shall assume exist. Some transition intensities may be 0 for all t. Graphically, multistate models may be illustrated using diagrams like Figure 1 or 2 with boxes representing the states and with arrows between the states representing the possible transitions, that is, the nonzero transition intensities (Hoem, 1976; Andersen et al., 1993: Section I.3). A state $h \in S$ is *absorbing* if for all $t \in \mathcal{T}$, $j \in \mathcal{S}$, $j \neq h$, $\alpha_{hj}(t) = 0$, otherwise h is *transient*. The *state occupation probabilities* $\pi_h(t) = \text{Prob}(X(t) = h)$ are given by

$$\pi_h(t) = \sum_{j \in \mathcal{S}} \pi_j(0) P_{jh}(0,t)$$

Notice that the $P_{hj}(\cdot,\cdot)$, the $\pi_h(\cdot)$, and the $\alpha_{hj}(\cdot)$ all depend both on the probability measure Prob and on the history. The process is *Markovian* if transition intensities $\alpha_{hj}(t)$ only depend on the history via the state $h = X(t)$ occupied at time t. Sometimes one is interested in considering an extended history which also includes observed *covariates*. If only time-fixed covariates Z are studied then the observed history is $\mathcal{F}_t = \mathcal{X}_t \vee \mathcal{Z}_0$ whereas time-dependent covariates $Z(t)$ give rise to an extended history of the form $\mathcal{F}_t = \mathcal{X}_t \vee \mathcal{Z}_t$ where in both cases $\mathcal{Z}_t$ is the history generated by the covariates in $[0, t]$. We shall here focus on the *purely endogenous* case where $\mathcal{Z}_t \subset \mathcal{X}_t \vee \mathcal{Z}_0$, that is, the covariates are either all time-fixed or the random development of the time-dependent covariates is fully specified by the history of the process itself (see, however, Section Statistical Model Specification).

The Two-State Model for Survival Data

This model, illustrated in Figure 1, has $p = 2$ states and only one possible transition from state 0 to state 1. The corresponding transition intensity $\alpha_{01}(t)$ is given by the hazard rate function $\alpha(t)$, while $\alpha_{10}(t) = 0$ for all t, that is, state 1 is absorbing. The initial distribution is degenerate in 0: $\pi_0(0) = 1$ and the process is Markovian. Covariates may be entered into the model using a regression model for $\alpha(\cdot)$.

The Competing Risks Model

This model has one transient state 0: alive and a number, k, of absorbing states, state h, $h = 1,\ldots, k$ corresponding to death from cause h. Thus, there are $p = k + 1$ states. The model is illustrated for $k = 2$ in Figure 2. The transition intensities $\alpha_{0h}(t)$ for $h = 1,\ldots, k$ are given by the *cause-specific hazard* functions.

$$\alpha_h(t) = \lim_{\Delta t \to 0} \text{Prob}(\text{Dead from cause } h \text{ by } t + \Delta t | T \geq t)/\Delta t$$

where T is the survival time. The initial distribution is degenerate in 0, the only transient state of the model, that is, $\alpha_{hj}(t) = 0$ for all $h \neq 0$ and all $j \neq h$. The transition probabilities are the *overall survival probability*

$$P_{00}(0,t) = S(t) = \text{Prob}(T > t)$$

$$= \exp\left(- \int_0^t \sum_{h=1}^{k} \alpha_h(u)\, \mathrm{d}u \right)$$

and the cumulative incidences

$$P_{0h}(0,t) = \int_0^t S(u-)\alpha_h(u)du, h = 1, \ldots, k$$

Like the simple two-state model ($k = 1$) the competing risks model is Markovian and covariates may be included via regression models for the cause-specific hazards.

The Illness–Death Model

This model is illustrated in Figure 3, which by now should be almost self-explanatory. Often the time t is the age of the individual, and usually individuals will be assumed to be in state 0 at $t = 0$. However, individuals will not always be observed from $t = 0$ as shall be further discussed in Section Statistical Model Specification below regarding observational patterns. The mortality $\alpha_{12}(t)$ of the diseased (the *lethality*) may sometimes depend on duration d since entry to state 1 in addition to the dependence on age t. (Notice that, despite the notation, $\alpha_{12}(t)$ then depends on the *random* time of the most recent transition into 1.) If $\alpha_{12}(t)$ does not depend on d the process is Markovian, otherwise it is a semi-Markov process, which is clearly purely endogenous.

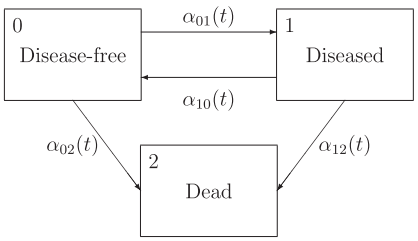

Figure 3 The illness–death or disability model.

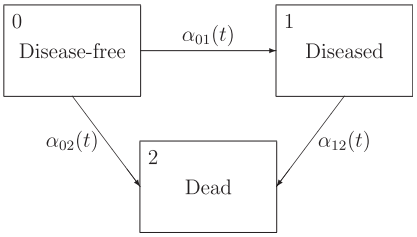

Figure 4 Unidirectional illness–death model.

In Figure 3 we have indicated the possibility of reversibility: the transition back from state 1 to 0 is possible. It will turn out that the simple unidirectional model in Figure 4 is rather easier to analyze statistically. Thus, the transition probabilities in this model have simple explicit expressions.

$$P_{00}(s,t) = \exp\left(- \int_s^t (\alpha_{02}(u) + \alpha_{01}(u))\mathrm{d}u \right)$$

and (in the Markovian case)

$$P_{01}(s,t) = \int_s^t P_{00}(s,u-)\alpha_{01}(u)P_{11}(u,t)\mathrm{d}u$$

where

$$P_{11}(s,t) = \exp\left(- \int_s^t \alpha_{12}(u)\mathrm{d}u \right)$$

For this reason, and because of the desire to have a more detailed description, the reversible illness–death model is sometimes replaced by a model with one state for each new 'disease occurrence.' An example from nuptiality is indicated in Figure 5, where the possible transition from each state to death is omitted from the figure. (Only finitely many states are assumed!)

Interaction between Life–history Events

This Markov model, illustrated in Figure 6 describes the joint behavior of two life events A and B; if $\alpha_{0B} = \alpha_{A,AB}$ but $\alpha_{0A} \neq \alpha_{B,AB}$, A is called *locally dependent* on B but B is not locally dependent on A. The temporal order of events allows for this *asymmetric* concept of dependence which yields more

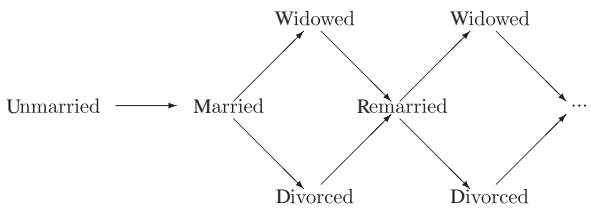

Figure 5 A model for marriage and dissolution.

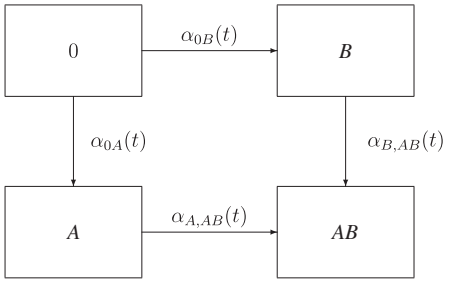

Figure 6 Interaction between life history events.

information for drawing causal inference than the standard symmetric association concepts from cross-sectional studies. Similar duration dependence as in the illness–death process might be added. A model of this type was discussed by Aalen et al. (1980) for a study of interaction between menopause and a certain chronic skin disease (see also Blossfeld and Rohwer, 1995; Courgeau and Leliévre, 1992).

Bone Marrow Transplantation

A model combining most of the above features has been studied in detail (e.g., Klein et al., 1993; Keiding et al., 2001) as describing some of the possible states of a leukemia patient following bone marrow transplantation, see **Figure 7**. Patients have been given various kinds of therapy to temporarily keep the disease down, they are said to be in *remission*. In our context these patients are followed since bone marrow transplantation ($t = 0$), initially considered in state 0. Two different types of complications are considered: acute graft-vs-host disease (A), chronic graft-vs-host disease (C), and a special state AC is defined for those patients acquiring both A and C. Patients are followed until relapse of the leukemia (R) or death (D) while still in remission. Relapsed patients are not followed further in this context though, of course, a mortality rate among relapsed could be of interest in some settings. If all transition rates depend only on time t since transplantation, we have again a Markov process, but various kinds of duration dependence (semi-Markov process models) may also be relevant.

Counting Process Representation, Likelihood

Assume that multistate processes $X_i(t)$ such as described in Section Multistate Models are observed over intervals $[0, \tau_i]$ for individuals $i = 1,\ldots, n$. In this section, we will assume that τ_i is a fixed (i.e., nonrandom) time of termination of observation for individual i. Random right-censoring (Section Survival Data) is treated in Section Observational Patterns. Since $X_i(t)$ is constant between transitions, it is equivalent to record $X_i(0)$ and the *counting processes*.

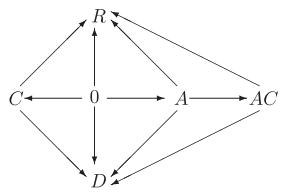

Figure 7 A model for events following bone marrow transplantation: acute and chronic graft-vs-host disease, relapse, and death.

$$N^i_{hj}(t) = \#\,(\text{direct transitions } h \to j \text{ in } [0,t] \text{ for } i$$

described by the times T^{ik}_{hj} of these transitions, where

$$0 < T^{i1}_{hj} < \ldots < T^{iN^i_{hj}(\tau_i)}_{hj} \le \tau_i$$

Let

$$N_{hj}(t) = \sum_{i=1}^{n} N^i_{hj}(t)$$

It will turn out to be useful to also introduce $Y^i_h(t) = I\{X_i(t-) = h\}$ and

$$Y_h(t) = \#\,(\text{individuals at risk in state } h \text{ at time } t-)$$
$$= \sum_{i=1}^{n} Y^i_h(t)$$

Note that, for $t > \tau_i$, $N^i_{hj}(t) = N^i_{hj}(\tau_i)$ and $Y^i_h(t) = 0$; these can thus be considered to be defined on $(0, \infty)$.

For individual i denote the initial distribution $(\pi^i_h(0))$, the density of time-fixed covariates $f(Z_i)$, and the transition intensities $\alpha^i_{hj}(t)$, then the likelihood is

$$\prod_{i=1}^{n} f(Z_i)\pi_{X_i(0)} \prod_{h \ne j} \prod_{k=1}^{N^i_{hj}(\tau_i)} \alpha^i_{hj}\left(T^{ik}_{hj}\right) \exp\left(-\int_0^{\tau_i} \alpha^i_{hj}(t) Y^i_h(t)\,\mathrm{d}t\right)$$

It is very common to condition on Z_i and on the initial values $X_i(0)$ (the distribution of which may often be degenerate anyway), and consequently omit the factors $f(Z_i)\pi_{X_i(0)}$ from the likelihood. We shall do so without further comment in the sequel.

Recall from above that the notation $\alpha^i_{hj}(t)$ represents possible dependence of the transition intensity on the whole history $\mathcal{X}^i_t$ of the process. Thus, $\alpha^i_{hj}(t)$ may well contain covariates and other random elements, as already exemplified.

Statistical Model Specification

As discussed above the basic parameters in event history analysis are the *transition intensities* from which a full probabilistic description of the processes is possible. Thus, assessing their dependence on covariates, possibly using various stratifications, will often be an important first step of the analysis. However, frequently there are other targets of inference and additional functionals are useful, particularly various types of transition probabilities obtained by integrating certain functions of the transition intensities. Sometimes, it is therefore desirable to attempt to model such *marginal* features of the event history data directly. A final activity may be *prediction*, both as an illustration of the dynamics and for concrete practical purposes.

Markov Processes

The most important class of models is the (continuous time) *Markov process* $X(t)$ on a finite state space $\mathcal{S} = \{1,\ldots, p\}$

where the dependence of $\alpha^i_{hj}(t)$ on the history $\mathcal{X}_t$ introduced at the beginning of Section Multistate Models is only via the current state h of $X(t)$ (and possibly via time-fixed covariates). Statistical models are usually obtained by specifying the class of transition intensities $(\alpha^i_{hj}(t))$ for each individual i.

Constant and Piecewise Constant Transition Intensities

The simplest class of models is obtained by keeping the transition rates constant: $\alpha^i_{hj}(t) = \alpha^i_{hj}$. Piecewise constant intensities

$$\alpha^i_{hj}(t) = \alpha^{i(k)}_{hj}, t_{k-1} < t \leq t_k, t_0 = 0$$

form the next step up, and both of these choices are of widespread use, particularly in large studies in econometrics, epidemiology, sociology, and demography.

For ease of notation focus on the case with constant intensities and assume first that all individuals have the same transition intensities, $\alpha^i_{hj} = \alpha_{hj}$. The likelihood then simplifies to

$$\prod_{h \neq j} \alpha^{N_{hj}(\tau)} e^{-\alpha_{hj} S_h}$$

where $\tau = \max_i \tau_i$ and

$$S_h = \sum_i \int_0^{\tau_i} Y^j_h(t) \mathrm{d}t$$

Since this likelihood looks like one resulting from observing $N_{hj}(\tau)$ events in a Poisson process with intensity α_{hj} observed over the interval $(0, S_h)$, except that here S_h is random, this likelihood, and even the model, is often associated with Poisson's name. Maximum likelihood estimation is elementary, yielding

$$\widehat{\alpha}_{hj} = N_{hj}(\tau)/S_h$$

the classical *occurrence/exposure rate*. Asymptotic inference may be obtained from the information

$$-D^2 \log L = N_{hj}(\tau)/\alpha^2_{hj}$$

$$\mathrm{Var}(\widehat{\alpha}_{hj}) \sim \alpha^2_{hj}/N_{hj} \sim N_{hj}(\tau)/S^2_h$$

However, these variance expressions build on a correctly specified model and for that reason it is often desirable to apply *robust variances* based on the so-called *sandwich formula*, see, e.g., Keiding (2014).

Inclusion of time-fixed covariates is conveniently achieved log-linearly.

$$\alpha^i_{hj} = \alpha_{0hj} e^{\beta'_{hj} Z_i}$$

The Poisson likelihood easily accommodates this in what is often termed Poisson regression models, which may be treated using standard generalized linear model software.

Transition probabilities for the constant and piecewise constant Markov process models are explicit functions of the transition intensities (e.g., Chiang, 1968), allowing direct plug-in maximum likelihood estimation, as well as calculation of standard error estimates via the delta-method.

Other Parametric Models for the Transition Intensities

Although the piecewise constant model is often sufficient to describe the dependence of intensities on time, other possibilities exist. Certain mathematical functions of time may generate the model, such as the Gompertz-Makeham model for mortality.

$$\alpha(t) = \alpha + \beta\gamma^t$$

but except for mortality studies in actuarial and some demographic contexts such parametric models are used rather little. One reason for this may be the powerful development of methodology for nonparametric statistical inference, where $\alpha_{hj}(t)$ is left unspecified.

Freely Varying (Nonparametric) Transition Intensities

Assume first that the transition intensities are the same for all individuals but that they are allowed to vary freely with time: $\alpha^i_{hj}(t) = \alpha_{hj}(t)$. Statistical inference is then conveniently phrased in terms of the counting process approach pioneered by Aalen (1978) (see Andersen et al., 1993, for a detailed exposition). Estimators (which may be given a nonparametric maximum likelihood interpretation) of the integrated intensities

$$A_{hj}(t) = \int_0^t \alpha_{hj}(u) \mathrm{d}u$$

are obtained as the *Nelson–Aalen estimators*

$$\widehat{A}_{hj}(t) = \int_0^t \frac{J_h(u)}{Y_h(u)} \mathrm{d}N_{hj}(u) = \sum_i \sum_{k:0<T^{ik}_{hj} \leq t} \frac{1}{Y_h\left(T^{ik}_{hj}\right)}$$

$J_h(u) = I\{Y_h(u) > 0\}$, with variance estimators

$$\widehat{\sigma}^2\left(\widehat{A}_{hj}(t)\right) = \int_0^t \frac{J_h(u)}{Y_h(u)^2} \mathrm{d}N_{hj}(u) = \sum_i \sum_{k:0<T^{ik}_{hj} \leq t} \frac{1}{Y_h\left(T^{ik}_{hj}\right)^2}$$

An elaborate mathematical theory based on stochastic integrals and martingales is available to study exact and asymptotic properties of these estimators.

When estimates are desired of the transition intensities $\alpha_{hj}(t)$ themselves rather than their integrals, smoothing techniques are necessary.

An important feature of the nonparametric approach is its elegant generalization (due to Aalen and Johansen, 1978), to estimating *transition probabilities*. The basic tool is the (matrix) product-integral. Let I be the identity matrix and G a matrix-valued function. The corresponding *product-integral* is defined as

$$\prod_0^t (I + G(\mathrm{d}s)) = \lim_{\max|t_v - t_{v-1}| \to 0} \prod (I + G(t_v) - G(t_{v-1}))$$

where $0 = t_0 < t_1 < \ldots < t_n = t$ is a partition of $[0, t]$. In particular, if G is continuous and scalar

$$\prod_0^t (1 + G(\mathrm{d}s)) = e^{-G(t) - G(0)}$$

and if G is a scalar step function

$$\prod_0^t (1 + G(\mathrm{d}s)) = \prod_{k=1}^K \left(1 + \Delta G\left(t_{(K)}\right)\right)$$

where $t_{(0)} = 0$ and $0 < t_{(1)} < \ldots < t_{(K)} < t$ are the jump times of G and

$$\Delta G\left(t_{(k)}\right) = G\left(t_{(k)}\right) - G\left(t_{(k-1)}\right)$$

So far, transition intensities $\alpha_{hj}(\bullet)$ were only defined for $j \neq h$. Define now

$$\alpha_{hh}(t) = -\sum_{j \neq h} \alpha_{hj}(t)$$

and the intensity matrix function

$$A(t) = \left(\left(\int_0^t \alpha_{hj}(u)\mathrm{d}u\right)\right)$$

Then the matrix $P(s, t) = ((P_{hj}(s, t)))$ of transition probabilities

$$P_{hj}(s, t) = \mathrm{Prob}(X_i(t) = j | X_i(s) = h)$$

is given by

$$P(s, t) = \prod_s^t (I + A(\mathrm{d}u))$$

The *Aalen–Johansen estimator* of $P(s, t)$ is obtained by plugging the matrix of Nelson–Aalen estimators $(\widehat{A}_{hj}(t))$ into this formula.

$$\widehat{P}(s, t) = \prod_s^t (I + \widehat{A}(\mathrm{d}u))$$

For the simple two-state model for survival data, $\widehat{P}_{00}(0, t)$ reduces to the classical Kaplan and Meier (1958) estimator $\widehat{S}(t) = \prod_{T_i \leq t}(1 - \mathrm{d}N_{01}(T_i)/Y_0(T_i))$ of the survival function $S(t)$.

As documented in detail by Andersen et al. (1993: Section IV.4), there is a well-developed theory, again based on stochastic integrals and martingales, about the asymptotic properties of the Aalen–Johansen estimator.

Semi- and Nonparametric Regression Markov Models

For Markov models with several states, there will often be too little empirical basis for estimating freely varying transition intensities between all states, so that more parsimonious models are required. One possibility (demonstrated in detail by Klein et al., 1993; Keiding et al., 2001; for the bone marrow transplant example above) is to postulate that transition intensities *into* a state are proportional: $\alpha_{hj}(t) = \alpha_{0j}(t)e^{\beta_h}$ for all (relevant) h. The statistical analysis of this model is equivalent to a Cox regression model with time-dependent covariates indicating the departure state h.

The semiparametric Cox model (proposed by Cox (1972a,b) for survival data and for renewal processes, respectively) may also be used for including covariates not depending on time, this preserving the Markov property. The consequent generalization of the Aalen–Johansen estimator was studied by Andersen et al. (1991b, 1993: p. 512ff.) and implemented for the bone marrow transplant problems just mentioned. An alternative, nonparametric *additive* hazard regression model is due to Aalen (1980) with general discussion by Martinussen and Scheike (2006) and Aalen et al. (2008).

Beyond Markov Processes

The most important deviations from the Markov property in practice are various kinds of duration dependence, where transition intensities depend on other time origins than $t = 0$, typically the time at entry to the present state. There are two main approaches to handling these.

As long as transition intensities depend only on one time origin each (e.g., all intensities depend only on duration in the present state), a model for the multistate process may be obtained by combining independent submodels for each transition intensity. These may, in turn, be modeled as constant or piecewise constant or by non- or semiparametric models, and as long as there is a unidirectional flow in the model, transition probabilities are still straightforward explicit functionals which may be estimated by plugging in the intensity estimates. Variance calculations may, however, become less direct.

More elaborate models will include several time origins (such as both age, disease duration, duration since marriage dissolution), often in piecewise constant intensity models or semiparametric regression models such as the Cox model. An example concerning mortality of diabetics with or without nephropathy was studied by Andersen (1988), see also the consequences for Danish life insurance premiums (Ramlau-Hansen et al., 1987).

Nonparametric estimation of transition probabilities in the non-Markov unidirectional illness–death model (**Figure 4**) was studied by Meira-Machado et al. (2006) while, more generally, Datta and Satten (2001) showed that *state occupation probabilities* $\pi_h(t)$ may be estimated using the Aalen–Johansen estimator without a Markov assumption. Also, without a Markov assumption, transition probabilities may be estimated from a specification of all transition intensities using *micro-simulation* (e.g., Mitton et al., 2000).

The heterogeneity between individuals cannot always be accounted for by *observed* covariates, and at least some residual heterogeneity will have to be modeled as random variation. An important class of models here is the multiplicative *frailty models*, obtained by postulating random (usually transition-specific) *frailties* W^i_{hj}, and conditional transition intensities.

$$\alpha^i_{hj}\left(t | W^i_{hj}\right) = \alpha_{0hj}(t) W^i_{hj} e^{\beta' Z_i}$$

(e.g., Hougaard, 2000). By integrating over the (multivariate) frailty distribution, observable features of the multistate model may be derived.

Hypothetical Calculations in Multistate Models

As mentioned earlier, there is often considerable interest in studying the consequences of the estimated transition intensities by calculating summary measures such as transition probabilities. When a full model has been estimated, this can be done not only of the model observed in this world but, also, the consequences of an assumed (or fitted) multistate model may be further illustrated by calculating transition probabilities

in hypothetical models obtained by changing some of the parameters. An elaborate example of this was given by Keiding et al. (2001) for the bone marrow transplantation context, but similar calculations have been performed in the competing risks model ever since the first discussion by Bernoulli (1760) of the effect on population mortality of removing smallpox through vaccination. The interpretational justification of such calculations was discussed heatedly, particularly in the 1970s (see Gail, 1975; Prentice et al., 1978; Kalbfleisch and Prentice, 2002; Chapter 8), see Andersen and Keiding (2012) for a recent contribution to that discussion.

Partial Model Specification

We have so far assumed that the multistate model was completely specified through statistical modeling of all transition intensities. In a series of papers (Pepe, 1991; Pepe et al., 1991; Pepe and Mori, 1993; Couper and Pepe, 1997), Pepe and coworkers have developed estimates of certain functionals in multistate models without assuming a full probability structure. One example is the *prevalence* of a transient condition indicated by state c defined as

$$P_{0c}(0,t)/\sum_{j\in\mathcal{S}_T} P_{0j}(0,t)$$

where $\mathcal{S}_T$ is the set of transient states and 0 is a fixed initial state. The idea is to estimate numerator and denominator separately by simple linear combinations of Kaplan–Meier estimates. Easily applicable variance estimates are then available and in one application (Klein et al., 2000) it was shown that the precision of the Pepe approach was close to the more elaborate (and restrictive) complete Markov model.

Regression models for certain marginal features of a multistate process have been studied both in specific cases and in a more general framework. Thus, the model by Fine and Gray (1999) provides a direct link between a cumulative incidence function in a *competing risks* model and (time-fixed) covariates, and Lawless and Nadeau (1995) (see also Lin et al., 2000) studied a similar model for the mean number of events in a *recurrent events* model. General approaches to regression models for state occupation probabilities in a multistate model (however, with much emphasis on the simple competing risks model) were provided by the 'direct binomial regression models' by Scheike and Zhang (2007) and Scheike et al. (2008) and by the use of 'pseudoobservations' (e.g., Andersen et al., 2003; Klein and Andersen, 2005). Since these approaches are not likelihood based, estimated variances for regression coefficients are obtained using the *sandwich formula* for the relevant *estimating equations*.

Another example of partial model specification occurs when the model contains time-dependent covariates which are not purely endogenous. In fact, for time-fixed covariates we just (cf Section Counting Process Representation, Likelihood) conditioned on their observed values without specifying their distribution, $f(Z_i)$, but for time-dependent covariates such a conditioning is more tricky. Formally, the likelihood will contain factors for the stochastic development of $Z_i(t)$ given the history $\mathcal{F}_{t-} = \mathcal{X}_{t-} \vee \mathcal{Z}_{t-}$ and the likelihood described in Section Counting Process Representation, Likelihood, is not the full likelihood but only a partial likelihood for the parameters for the transition intensities $\alpha^i_{hj}(t)$. This means that inference for the transition *intensities* may be based on this partial likelihood whereas the transition *probabilities* typically will depend also on the parameters in the model for $Z_i(t)$. Thus, if the model contains time-dependent covariates which are not purely endogenous then transition probabilities cannot be estimated using only a partial model specification. One approach is to set up a *joint model* for the multistate process $X_i(t)$ and the time-dependent covariates $Z_i(t)$. When $Z_i(t)$ only takes a finite number of values this joint model could, again, be a multistate model where $Z_i(t)$ is now endogenous (e.g., Andersen, 1986; Andersen et al., 1991b; Cortese and Andersen, 2010). Examples of more general joint models were presented by Wulfsohn and Tsiatis (1997) and Henderson et al. (2000), see also the review by Tsiatis and Davidian (2004).

A simpler approach for this problem which may be particularly useful when *dynamic prediction* is an issue is to use *landmarking* (e.g., van Houwelingen, 2007; see also van Houwelingen and Putter, 2012). Here, a series of (e.g., Cox regression) models is set up, each at a prechosen landmark time, and each model then conditions on the current values of possibly time-dependent covariates.

Time-Dependent Confounding: Beyond the Range of Event History Analysis

A main purpose of including covariates in event history models is to correct effect estimates for confounding, and the necessary care to take in this connection is well known as long as only confounders known at baseline are considered, the *time-fixed* covariates. When considering 'controlling for' *time-varying* or *time-dependent* covariates, new interpretational difficulties arise, and these are not directly addressed by the technical statistical considerations around likelihood construction mentioned in Section Partial Model Specification. Pioneered by Robins (1986) in connection with the century-old attempts to obtain a conceptual framework for the healthy worker effects, the core problem may be illustrated by **Figure 8**. Here X_t and Y_t are covariate and response at time t. Since Y_t is a confounder for the association between X_t and $Y_{t+}1$, we should control for (i.e., condition on) Y_t, but because Y_t is also intermediate between X_{t-1} and Y_{t+1}, we should not dilute the effect between X_{t-1} and Y_{t+1} by conditioning on Y_t. Using the counterfactual framework for causal analysis, Robins (1986) defined causal effects in this context and explained (see in particular Robins et al., 1992) that Cox models with time-dependent covariates cannot estimate this causal effect. Robins's work has generated a large literature about methods for handling time-dependent confounding, see the excellent tutorial by Daniel et al. (2013) for a survey. These methods

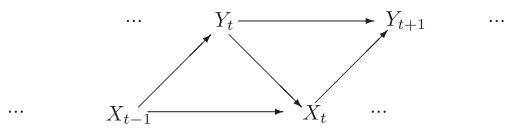

Figure 8 Illustration of time-dependent confounding.

go beyond the event history analysis framework discussed in this article.

Observational Patterns

As emphasized in the introduction, event history models are rarely observed completely. Some patterns of incomplete observation are more easily handled than others, and this section aims at introducing some of the more important classes.

Delayed Entry and Right-Censoring

Many of the statistical methods introduced in this article depend directly or indirectly on the likelihood representation of Section Counting Process Representation, Likelihood. Two patterns of incomplete observations are particularly easily tractable, because they lead to only minor modification of the likelihood: *Delayed entry* (or *left-truncation)* where individual i enters at some time V_i and *right-censoring* where nothing is known about i after some time U_i. Both V_i and U_i may be random although only dependent on the previous history of the process (see, e.g., Andersen et al., 1993, for precise specification of this and further discussion). The reason for the particular tractability of these mechanisms is that the at risk indicator $Y_h^i(t) = I\{X_i(t-) = h\}$ just needs to be amended to

$$Y_h^i(t) = I\{X_i(t-) = h, V_i < t \leq U_i\}$$

and the likelihood-based statistical methods then all go through.

The situation is more complex when modeling marginal features directly. Thus, in such cases simple inference will typically rely on an assumption that right-censoring is completely independent of the multistate process and otherwise the way in which right-censoring depends on covariates needs to be modeled explicitly (e.g., Fine and Gray, 1999; Scheike and Zhang, 2007). Also, left-truncation becomes much more tricky for marginal models (Geskus, 2011).

Interval Censoring

An important class of incomplete observational patterns consists in the times of some (but often not all) transitions not being known exactly but only up to an interval, e.g., between visits to a clinic or between censuses. (The *number* of transitions is usually assumed to be known.) A classical approach in demography (Hoem, 1969) is to approximate the exposure.

$$S_h^i = \int_{V_i}^{U_i \wedge \tau_i} Y_h^j(t)\mathrm{d}t$$

another to impute values in the observation interval for the time at risk. Systematic studies of nonparametric maximum likelihood estimation under interval censoring in the unidirectional illness-death model are by Frydman (1992, 1995), Frydman et al. (2013), Joly and Commenges, 1999 and Joly et al. (2012). Kay (1986) and Andersen et al. (1991a) exemplified interval-censored observation in the reversible illness-death model, using piecewise constant intensity models. Sun (2006) summarized methods for analysis of interval censored survival data.

Conditioning in Multistate Models

Many observational patterns in event history analysis may be described by conditional distributions in simpler models, which often describe 'direct' observations that are practically unobtainable. A prime example is the left-truncation just described, another is right-truncation with widespread use in studies of AIDS patients whose development is often observed conditional on having contracted the disease before the study entry. For general discussions see Kalbfleisch and Lawless (1989), Keiding and Gill (1990), and Gross and Huber-Carol (1992). A more elaborate application of such retrospective observational plans obtained by conditioning in an underlying prospective Markov process model was documented by Aalen (1980) for the four-state interaction of life history events example described above (**Figure 6**).

An important sampling frame for the illness–death model without recovery is the *prevalent cohort study* where a cross-sectional sample of diseased is taken at a fixed calendar time. Keiding (1992) (see also Lund, 2000) discussed the conditions for correct inference on the mortality $\alpha_{12}(t)$ based on follow-up of the diseased, studied under left-truncation, and compared with inference based on the *length-biased* durations which include the retrospective time from disease onset, as well as to the *forward recurrence time* from sampling to death, assuming stationarity.

A comprehensive substantive discussion of biases in prevalent cohorts was given by Brookmeyer (1998). Retrospective estimation of incidence based on the disease onset information of the survivors and independent lethality information was exemplified by Keiding et al. (1989), see also Ogata et al. (2000).

Sometimes interval censoring is extreme: in a cross-sectional study it is for all individuals only known whether or not an event has happened at age of sampling. Methodology for analyzing such *current status* data was surveyed by Sun (2006), cf Keiding (1991, 2006). *Current duration* data (backward recurrence times) may sometimes be used to estimate mortality from the age distribution of those living at a cross-section, assuming stationarity. Keiding et al. (2002, 2012) and Scheike and Keiding (2006) surveyed application of this approach to time-to-pregnancy studies, and Slama et al. (2006, 2012) applied the technique to a French telephone survey of time to pregnancy.

Concluding Remarks

We have attempted to provide an exposition of problems and solutions in event history analysis in continuous time. The probability models are continuous-time nonhomogeneous Markov processes and generalizations, the main statistical techniques are log-linear intensity models as well as non- and semiparametric counting process-based models. It is emphasized that considerable care (and consequent methodological innovation) is motivated by the many indirect observational patterns.

See also: Event History Analysis: Applications; Event History Analysis; Life Course and Event History Analysis; Life Table; Longitudinal Data: Event–History Analysis in Discrete Time; Markov Chain Monte Carlo Methods; Multistate Transition Models in Demography; Survival Analysis: Introduction.

Bibliography

Aalen, O.O., 1978. Nonparametric inference for a family of counting processes. Annals of Statistics 6, 701–726.

Aalen, O.O., 1980. A model for non-parametric regression analysis of counting processes. Springer Lecture Notes in Statistics 2, 1–25.

Aalen, O.O., Borgan, Ø., Keiding, N., Thormann, J., 1980. Interaction between life history events: nonparametric analysis for prospective and retrospective data in the presence of censoring. Scandinavian Journal of Statistics 7, 161–171.

Aalen, O.O., Borgan, Ø., Gjessing, H.K., 2008. Survival and Event History Analysis. Springer-Verlag, New York.

Aalen, O.O., Johansen, S., 1978. An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. Scandinavian Journal of Statistics 5, 141–150.

Andersen, P.K., 1986. Time-dependent covariates and Markov processes. In: Moolgavkar, S.H., Prentice, R.L. (Eds.), Modern Statistical Methods in Chronic Disease Epidemiology. Wiley, New York, pp. 82–103.

Andersen, P.K., 1988. Multistate models in survival analysis: a study of nephropathy and mortality in diabetes. Statistics in Medicine 7, 661–670.

Andersen, P.K., Borgan, Ø., Gill, R.D., Keiding, N., 1993. Statistical Models Based on Counting Processes. Springer, New York.

Andersen, P.K., Hansen, L.S., Keiding, N., 1991a. Assessing the influence of reversible disease indicators on survival. Statistics in Medicine 10, 1061–1067.

Andersen, P.K., Hansen, L.S., Keiding, N., 1991b. Non- and semiparametric estimation of transition probabilities from censored observations of a non-homogeneous Markov process. Scandinavian Journal of Statistics 18, 153–167.

Andersen, P.K., Keiding, N., 2002. Multi-state models for event history analysis. Statistical Methods in Medical Research 11, 91–115.

Andersen, P.K., Keiding, N., 2012. Interpretability and importance of functionals in competing risks and multi-state models. Statistics in Medicine 31, 1074–1088.

Andersen, P.K., Klein, J.P., Rosthøj, S., 2003. Generalized linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika 90, 15–27.

Andersen, P.K., Pohar Perme, M., 2008. Inference for outcome probabilities in multi-state models. Lifetime Data Analysis 14, 405–431.

Arjas, E., Haara, P., 1984. A marked point process approach to censored failure data with complicated covariates. Scandinavian Journal of Statistics 11, 193–209.

Bernoulli, D., 1760. Essai d'une nouvelle analyse de la mortalité causé par la petite vérole, et des avantages de l'inoculation pour le prévenir. Historie avec le Mémoires, Académie Royal des Sciences 1–45.

Blossfeld, H., Rohwer, G., 1995. Techniques of Event History Modeling. Erlbaum, Mahwah, NJ.

Brookmeyer, R., 1998. Biased sampling of cohorts in epidemiology. In: Armitage, P., Colton, T. (Eds.), Encyclopedia of Biostatistics. Wiley, New York, pp. 338–350.

Chiang, C.L., 1968. Introduction to Stochastic Processes in Biostatistics. Wiley, New York.

Cortese, G., Andersen, P.K., 2010. Competing risks and time-dependent covariates. Biometrical Journal 52, 138–158.

Couper, D., Pepe, M.S., 1997. Modelling prevalence of a chronic condition: chronic graft-versus-host disease after bone marrow transplantation. Statistics in Medicine 16, 1551–1571.

Courgeau, D., Leliévre, E., 1992. Event History Analysis in Demography. Clarendon Press, Oxford, UK.

Cox, D.R., 1972a. Regression models and life tables. Journal of the Royal Statistical Society B 34, 187–220.

Cox, D.R., 1972b. The statistical analysis of dependencies in point processes. In: Lewis, P.A.W. (Ed.), Stochastic Point Processes. Wiley, New York, pp. 55–66.

Daniel, R.M., Cousens, S.N., De Stavola, B.L., Kenward, M.G., Sterne, J.A.C., 2013. Methods for dealing with time-dependent confounding. Statistics in Medicine 32, 1584–1618.

Datta, S., Satten, G.A., 2001. Validity of the Aalen-Johansen estimators of stage occupation probablities and Nelson-Aalen estimators of integrated transition hazards for non-Markov models. Statistics and Prob Letters 55, 403–411.

Fine, J.P., Gray, R.J., 1999. A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association 94, 496–509.

Frydman, H., 1992. A non-parametric estimation procedure for a periodically observed three-state Markov process, with application to AIDS. Journal of the Royal Statistical Society Series B 54, 853–866.

Frydman, H., 1995. Semiparametric estimation in a three-state duration dependent Markov model from interval-censored observations with application to AIDS data. Biometrics 51, 502–511.

Frydman, H., Gerds, T., Groen, R., Keiding, N., 2013. Nonparametic estimation in an "illness-death" model when all transition times are interval-censored. Biometrical Journal 55, 823–843.

Gail, M., 1975. A review and critique of some models used in competing risk analysis. Biometrics 31, 209–222.

Geskus, R.B., 2011. Cause-specific cumulative incidence estimation and the Fine and Gray model under both left truncation and right censoring. Biometrics 67, 39–49.

Gross, S.T., Huber-Carol, C., 1992. Regression models for right truncated left censored survival data. Scandinavian Journal of Statistics 19, 193–213.

Henderson, R., Diggle, P., Dobson, A., 2000. Joint modeling of longitudinal measurements and event time data. Bioststistics 1, 465–480.

Hoem, J.M., 1969. Fertility rates and reproduction rates in a probabilistic setting. Biométrie-Praximétrié 10, 38–66.

Hoem, J.M., 1976. The statistical theory of demographic rates. A review of current developments. Scandinavian Journal of Statistics 3, 169–185.

Hougaard, P., 2000. Analysis of Multivariate Survival Data. Springer-Verlag, New York.

van Houwelingen, H.C., 2007. Dynamic prediction by landmarking in event history analysis. Scandinavian Journal of Statistics 34, 70–85.

van Houwelingen, H.C., Putter, H., 2012. Dynamic Prediction in Clinical Survival Analysis. CRC/Chapman and Hall, Boca Raton.

Joly, P., Commenges, D., 1999. A penalized likelihood approach for a progressive three-state model with censored and truncated data: application to AIDS. Biometrics 55, 887–890.

Joly, P., Gerds, T.A., Qvist, V., Commenges, D., Keiding, N., 2012. Estimating survival of dental fillings based on interval censored data and multi-state models. Statistics in Medicine 31, 1139–1149.

Kalbfleisch, J.D., Lawless, J.F., 1989. Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS. Journal of the American Statistical Association 84, 360–372.

Kalbfleisch, J.D., Prentice, R.L., 2002. The Statistical Analysis of Failure Time Data, second ed. Wiley, New York.

Kaplan, E.L., Meier, P., 1958. Non-parametric estimation from incomplete observations. Journal of the American Statistical Association 53, 457–481.

Kay, R., 1986. A Markov model for analyzing cancer markers and disease states in survival studies. Biometrics 42, 855–865.

Keiding, N., 1991. Age-specific incidence and prevalence: a statistical perspective. Journal of the Royal Statistical Society A 154, 371–412.

Keiding, N., 1992. Independent delayed entry. In: Klein, J.P., Goel, P.K. (Eds.), Survival Analysis: State of the Art. Kluwer, Dordrecht, pp. 309–326.

Keiding, N., 2006. Event history analysis and the cross-section. Statistics in Medicine 25, 2343–2364.

Keiding, N., 2014. Event history analysis. Annual Review of Statistics and its Applications 1, 333–360.

Keiding, N., Gill, R.D., 1990. Random truncation models and Markov processes. Annals of Statistics 18, 582–602.

Keiding, N., Hansen, O.H.H., Sørensen, D.N., Slama, R., 2012. The current duration approach to estimating time to pregnancy (with discussion). Scandinavian Journal of Statististics 39, 185–213.

Keiding, N., Holst, C., Green, A., 1989. Retrospective estimation of diabetes incidence from information in a current prevalent population and historical mortality. American Journal of Epidemiology 130, 588–600.

Keiding, N., Klein, J.P., Horowitz, M.M., 2001. Multistate models and outcome prediction in bone marrow transplantation. Statistics in Medicine 12, 1871–1885.

Keiding, N., Kvist, K., Hartvig, H., Tvede, M., Juul, S., 2002. Estimating time to pregnancy from current durations in a cross-sectional sample. Biostatistics 3, 565–578.

Klein, J.P., Andersen, P.K., 2005. Regression modeling of competing risks data based on pseudo-values of the cumulative incidence function. Biometrics 61, 223–229.

Klein, J.P., Keiding, N., Copelan, E.A., 1993. Plotting summary predictions in multistate survival models: probabilities of relapse and death in remission for bone marrow transplantation patients. Statistics in Medicine 12, 2315–2332.

Klein, J.P., Keiding, N., Shu, Y., Szydlo, R.M., Goldman, J.M., 2000. Summary curves for patients transplanted for chronic myeloid leukemia salvaged by a donor lymphocyte infusion: the current leukemia free survival curve. British Journal of Haematology 109, 148152.

Lawless, J.F., Nadeau, J.C., 1995. Nonparametric estimation of cumulative mean functions for recurrent events. Technometrics 37, 158–168.

Lin, D.Y., Wei, L.J., Yang, I., Ying, Z., 2000. Semiparametric regression for the mean and rate functions of recurrent events. Journal of the Royal Statistical Society, series B 62, 711–730.

Lund, J., 2000. Sampling bias in population studies – how to use the Lexis diagram. Scandinavian Journal of Statistics 27, 589–604.

Martinussen, T., Scheike, T.H., 2006. Dynamic Regression Models for Survival Data. Springer-Verlag, New York.

Meira-Machado, L., Una-Alvarez, J., Cadarso-Suarez, C., 2006. Nonparametric estimation of transition probabilities in a non-Markov illness-death model. Lifetime Data Analysis 12, 325–344.

Mitton, L., Sutherland, H., Week, M. (Eds.), 2000. Microsimulation Modelling for Policy Analysis. Challenges and Innovations. Cambridge University Press.

Ogata, Y., Katsura, K., Keiding, N., Holst, C., Green, A., 2000. Empirical Bayes age-period-cohort analysis of retrospective incidence data. Scandinavian Journal of Statistics 27, 415–432.

Pepe, M.S., 1991. Inference for events with dependent risks in multiple end-point studies. Journal of the American Statistical Association 86, 770–778.

Pepe, M.S., Longton, G., Thornquist, M., 1991. A qualifier Q for the survival function to describe the prevalence of a transient condition. Statistics in Medicine 10, 413–421.

Pepe, M.S., Mori, M., 1993. Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? Statistics in Medicine 12, 737–751.

Prentice, R.L., Kalbfleisch, J.D., Peterson, A.V., Flournoy, N., Farewell, V.T., Breslow, N., 1978. The analysis of failure time data in the presence of competing risks. Biometrics 34, 541–554.

Ramlau-Hansen, H., Jespersen, N.C.B., Andersen, P.K., Borch-Johnsen, K., Deckert, T., 1987. Life insurance for insulin-dependent diabetics. Scandinavian Actuarial Journal 19–36.

Robins, J.M., 1986. A new approach to causal inference in mortality studies with a sustained exposure period application to control of the healthy worker survivor effect. Mathematical Modelling 7, 1393–1512.

Robins, J.M., Blevins, D., Ritter, G., Wulfsohn, M., 1992. G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. Epidemiology 3, 319–336.

Scheike, T., Keiding, N., 2006. Design and analysis of time to pregnancy. Statistical Methods in Medical Research 15, 127–140.

Scheike, T.H., Zhang, M.-J., 2007. Direct modeling of regression effects for transition probabilities in multistate models. Scandinavian Journal of Statistics 34, 17–32.

Scheike, T.H., Zhang, M.-J., Gerds, T.A., 2008. Predicting cumulative incidence probability by direct binomial regression. Biometrika 95, 205–220.

Slama, R., Ducot, B., Carstensen, L., Lorente, C., de La Rochebrochard, E., Leridon, H., Keiding, N., Bouyer, J., 2006. Feasibility of the current duration approach to study human fecundity. Epidemiology 17, 440–449.

Slama, R., Hansen, O.K.H., Ducot, B., Bohet, A., Sorensen, D., Giorgis Allemand, L., Eijkemans, M.J.C., Rosetta, L., Thalabard, J.C., Keiding, N., Bouyer, J., 2012. Estimation of the frequency of involuntary infertility on a nation-wide basis. Human Reproduction 27, 1489–1498.

Sun, J., 2006. The Statistical Analysis of Interval-Censored Failure Time Data. Springer-Verlag, New York.

Tsiatis, A.A., Davidian, M., 2004. Joint modeling of longitudinal and time-to-event data: an overview. Statistica Sinica 14, 809–834.

Wulfsohn, M.S., Tsiatis, A.A., 1997. A joint model for survival and longitudinal data measured with error. Biometrics 53, 330–339.

Event Sequence Analysis: Methods

Andrew Abbott, University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 7, pp. 4966–4969, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

This article introduces the development of the method of event sequence analysis (SA) in social science. It begins by locating it in the tradition of exploratory data analysis. It then lays out various existing sequence analytic strategies that vary in the amount of narrative detail they retain as well as some methodological issues with SA. It concludes with a discussion of the potential applications of SA and its promise for adoption within social science following from algorithmic developments in biology and computer science.

Event sequence analysis (SA) methods are methods seeking patterns that recur across a number of sequences. They are applied to data on social sequences (occupational careers or life courses, for example) (e.g., Kogan, 2007; Malo and Muñoz-Bullón, 2003) as well as to historical sequences (of policy adoption, for example) and cultural sequences (different versions of a particular story, for example). The most common techniques for sequence analysis are optimal matching techniques, so-called because they operate by finding distances between pairs of sequences on the basis of an 'optimal match' of the two sequences – a minimal series of operations that transform one sequence into the other. In computer science, the optimal matching problem is usually called the string alignment problem. In social sequence analysis, the matrix of pairwise distances between sequences is used in any standard two-way analysis scheme like scaling or cluster analysis to produce categorizations or dimensionalizations of a sequence space. Sequence analysis has been used in a variety of empirical areas with varying success. It is a descriptive rather than causal technique and will typically be used in combination with causal methods rather than to their exclusion.

SA methods are methods seeking patterns that recur across a number of sequences. They are applied to data on social sequences like occupational careers or life courses, as well as to historical sequences such as might emerge in policy adoption or in the development of organizational structures and to cultural sequences like the various orders of elements in different versions of a widespread ritual or story. The most common techniques for seeking these patterns are standard sequence comparison algorithms widely available in biological software.

Sequence Analysis and Its Origins

Sequence analytic methods are part of a general turn toward the use of exploratory techniques in social science. As such they stand with techniques like cluster analysis, multidimensional scaling, and factor analysis. All of these methods seek radical simplification of the data space in the hope that analyzable patterns can be found. They contrast with the many causal techniques that take high dimensionality of the data space for granted and analyze that space by seeking to find the impact of a set of 'independent' dimensions on one 'dependent' dimension of the data.

This turn to exploratory techniques has a number of sources. One of these has been the stockpiling of immense data sets analyzable in many diverse ways. Another has been the more or less exploratory use of causal methods enabled by commodification; where hundreds of regressions precede those actually reported, the idea of classical hypothesis testing becomes questionable. Exploratory analytic techniques have generally been more common in applied social science rather than in academic work, having their widest use in market research. Within the academy, they have seen wider use in psychology than in any other social or behavioral science. While not unknown in political science and sociology (see, e.g., the various works of E.O. Laumann, such as Laumann and Knoke, 1987), they are virtually absent from economics.

The turn to sequence analysis in particular, however, proceeded from a critique of causal conceptions articulated by Abbott (1983, 1988, 1997) and Abell (1987). Both writers criticized what they called the 'variables' tradition for assuming that the social world is best imagined in terms of independent units of analysis whose properties 'cause' one another rather than in terms of chains of actions and events. Abell's argument was the more abstract, being rooted in a theory of social action based on rational choice. Abbott's was more empirical, being driven by the problem of comparing histories across social units. Both writers sought methods that would identify and compare similar chains of events across varieties of social sequence data.

Sequence analytic strategies vary in the amount of narrative detail they retain. Essentially, the more narrative detail retained the less powerful the modes of generalization or pattern search available. Strongest in retaining detail is David Heise's (1989) event structure analysis (ESA), which is essentially an interactive program for establishing the essential narrative dependencies in a complex story. The analyst inputs events and answers questions posed by the program about which events depend on which others and how. The program then produces a minimal representation of these dependencies. While it is very useful for establishing rigorous narrative patterns, ESA has not seen widespread use as a technique for generalizing about sequence patterns (for an example, see Griffin, 1993).

Abell's (1993) method for narrative analysis retains less empirical detail. It begins with a 'narrative table.' The table is essentially a matrix of actors across actions, implicitly like the Heise pattern, but ignoring the specific content of events to focus on the network 'shape' of the connections between events. Narratives are similar if they can be reduced homomorphically to the same network structure. Abell's method has not seen substantial use, perhaps because his notion of similar narrative structure is not seen by empirical analysts as of immediate utility, despite its theoretical elegance. (But see the comments by various writers on Abell's techniques in *Journal of Mathematical Sociology* 18, Numbers 2 and 3.)

Abbott's approach simplifies sequence data far more than do those of Abell or Heise, insisting on unilinear sequences of simple events. These sequences are then compared using the widely available pattern-matching algorithms that have emerged in computer science and biology since the 1970s (for a review, see Abbott and Tsay, 2000; Durbin, 1998; Von Heijne, 2012; Brzinsky-Fay and Kohler, 2010). As might be expected given its simplicity, Abbott's method has seen the widest use of SA methods. It is variously referred to as sequence analysis, optimal matching analysis, or optimal alignment, the latter two being names for the standard algorithms used to measure resemblance between sequences.

Abbott's method (hereafter abbreviated OM for its most common name, optimal matching) presupposes a set of 'sequence data,' that is, a set of strings (sequences), each one of which is a vector of elements representing a succession of events in some discrete order. In career data, these might be sequences of jobs held in each month over some period. In cultural analysis, they might be orders of gestures in rituals. Events can recur or return an indefinite number of times, and there is no necessary constraint of equal length on the sequences.

Simply described, Abbott's method consists in using dynamic programming algorithms to define distances between all possible pairs of sequences in a data set. These distances are then analyzed by clustering or scaling to produce some typology of sequence types. This typology may be used as a simple description or as a dependent or independent variable in more standard causal techniques. Thus stated, Abbott's formulation of sequence analysis for the social sciences is equivalent to what is known in computer science as the multiple string alignment problem, on which there is an extensive literature (see Gusfield, 1997: Chapter 14).

Algorithms for multiple string alignment are typically based on simple algebra establishing 'Levenshtein distance' between two strings, the minimum number of edit operations (insertions, deletions, replacements) necessary to transform one string into the other. Although in some applications more complex operations have been investigated (see papers in Sankoff and Kruskal, 1983), the string comparison literature has focused on the simplest three operations.

Finding the minimal edit distance between two sequences is a matter of inspection unless varying weights are placed on varying operations. Typical weighting schemes involve a fixed insertion and deletion weight coupled with replacement weights that vary in terms of what replaces what. Under such weighting schemes, the minimal edit distance problem must be solved by a simple dynamic recursion usually attributed to the classic paper of Needleman and Wunsch (1970).

Sequence analysts may also be interested in the question of whether a set of sequence data has some common subpattern, i.e., whether there is a characteristic subsequence that may appear in many or most sequences. This problem, known technically as the multiple local alignment problem, has been solved by a Gibbs sampling algorithm due to Lawrence et al. (1993). Abbott has applied it to the analysis of rhetorical subsequences in sociological articles (Abbott and Barman, 1997).

Methodological Issues

Two kinds of methodological issues have arisen in the area of SA methods. In the main, these concern OM methods, as these are the most widely applied. The first such issues are those proper to the method itself, the second those concerning its utility. Like all methods, OM has its own internal assumptions and difficulties. Issues of data coding have vexed many users, although there is some indication that the methods, like most exploratory methods, are less vulnerable to the vagaries of coding than methods taking the data as given. One coding issue in particular has seemed worrisome, that of coding the underlying temporality. Some have used regular time intervals; others have collapsed continuous episodes. Another important issue has been the setting of weights of replacement and insertion/deletion. A wide variety of schemes have been used, but relatively few analysts have studied the impact of varying schemes. It is clear that all these areas need more experimentation than has presently been done. This is also true in terms of validation. Most typologies and clusterings in the OM literature have been accepted as is, without further validation. A few papers have used Monte Carlo validation methods.

The other major issues with OM methods in particular, as with SA methods more generally, concern their utility. There are two versions of this concern, one small scale, the other more general. The small-scale issue involves what SA methods can really add to event history or other methods investigating hazard rates for particular kinds of events. These are the standard 'causal' methodologies that are applied to sequence data. In areas where such methods are well developed (examination of careers, labor force involvement, and marital status), SA methods seem to standard analysts to have little to add. In other areas, however, particularly in the analysis of sequential cultural structures like stories or rituals, event history methods themselves have no utility. It is not yet clear whether SA methods can provide standard-method career analysts with results that will surprise them. (Like any paradigm, event history methods to some extent enforce criteria for results that favor their own types of results.) All the same, the eventual success of SA methods will depend on finding regularities that event history methods cannot find. A clear discussion (and example) of such regularities appears in Stovel et al. (1996).

The more general question about SA's utility seems to have a clearer answer. There is little question that the broad project of employing pattern-matching techniques in social science data is a sound one. Partisans of SA methods entered the area largely to carry out a theoretical agenda that began as an attack on standard methodologies. But in turning to the computer

science and biological literature, they found one of the most rapidly advancing areas of modern scientific methodology. Seeking patterns directly, rather than through hypothesized models, is not only good science, but also may be one of the only means of addressing the monumental data sets of the future.

Future Developments

To date, most applications of SA have involved a fairly limited substantive range. The most common application has been to career data. Some have worked with particular job histories, others with broader occupational classes. Still others have combined family data into these patterns. A number have worked with the 'careers' of larger social organizations – professions, counties, and nations. Still other applications have involved patterns of activity – at travel agencies, libraries, decision-making groups, and the like. A final group of studies has considered cultural artifacts – stories, dances, academic articles, and so on. As these examples show, the method has been typically applied where standardized patterns are strongly expected.

Most early SA was done using fairly simple computer packages adapted for the purpose. At the beginning of the twenty-first century standard biological packages are being adapted for social use (e.g., CLUSTAL by scholars at the University of Strasbourg). Applications of SA can be expected to grow more common once the actual algorithm (not just the weighting scheme) can be tailored to the type of sequence regularity expected in the data (e.g., Gauthier et al., 2010).

The extraordinary flowering of pattern-matching methods in biology and computer science will undoubtedly lead to greater application of these methods within social analysis. They are already widely used in market research, of course, and therefore constitute a central part of what is by far the most extensive and comprehensive form of social scientific analysis at present. To be sure, the causal paradigm against which SA methods were developed remains strong within the empirical social sciences. But in the past, widespread availability of new methods has led to wholesale adoption even in the face of earlier paradigms. Indeed, the present causal methods spread precisely in this way. Given the rise of pattern-matching as a general strategy of scientific inquiry, SA and related methods will almost certainly spread. They will also constitute the entering wedge of a much broader set of descriptive approaches to social reality that can be expected to transform social inquiry in the next decades.

See also: Dimensionality of Tests: Methodology; Event History Analysis: Applications; Event History Analysis: Applications; Growth Curve Analysis; Interrupted Time Series Quasi-Experimental Designs; Longitudinal Data: Event–History Analysis in Discrete Time; Statistical Pattern Recognition.

Bibliography

Abbott, A., 1983. Sequences of social events. Historical Methods 16, 129–147.
Abbott, A., 1988. Transcending general linear reality. Sociological Theory 6, 169–186.
Abbott, A., 1997. Of time and space. Social Forces 75, 1149–1182.
Abbott, A., Barman, E., 1997. Sequence comparison via alignment and Gibbs sampling. Sociological Methodology 27, 47–87.
Abbott, A., Tsay, A., 2000. Sequence analysis and optimal matching methods in sociology. Sociological Methods and Research 29, 3–33.
Abell, P., 1987. The Syntax of Social Life Oxford. Oxford University Press, Oxford, UK.
Abell, P., 1993. Some aspects of narrative method. Journal of Mathematical Sociology 18 (2–3), 93–134.
Brzinsky-Fay, C., Kohler, U., 2010. New developments in sequence analysis. Sociological Methods & Research 38, 359–364.
Durbin, R. (Ed.), 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, New York.
Griffin, L.J., 1993. Narrative, event structure analysis, and causal interpretation in sociology. American Journal of Sociology 98, 1094–1133.
Gusfield, D., 1997. Algorithms on Strings, Trees, and Sequences. Cambridge University Press, Cambridge, UK.
Gauthier, J.A., Widmer, E.D., Bucher, P., Notredame, C., 2010. Multichannel sequence analysis applied to social science data. Sociological Methodology 40, 1–38.
Heise, D., 1989. Modeling event structures. Journal of Mathematical Sociology 14, 139–169.
Kogan, I., 2007. A study of immigrants' employment careers in West Germany using the sequence analysis technique. Social Science Research 36, 491–511.
Laumann, E.O., Knoke, D., 1987. The Organizational State. University of Wisconsin Press, Madison, WI.
Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., Wooton, J.C., 1993. Detecting subtle sequence signals. Science 262, 208–214.
Malo, M.A., Muñoz-Bullón, F., 2003. Employment status mobility from a lifecycle perspective: a sequence analysis of work-histories in the BHPS. Demographic Research 9, 471–494.
Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48, 443–453.
Sankoff, D., Kruskal, J.B., 1983. Time Warps, String Edits, and Macromolecules. Addison-Wesley, Reading, MA.
Stovel, K., Savage, M., Bearman, P., 1996. Ascription in achievement. American Journal of Sociology 102, 358–399.
Von Heijne, G. (Ed.), 2012. Sequence Analysis in Molecular Biology: Treasure Trove or Trivial Pursuit. Elsevier, Amsterdam.

Everyday Life, Anthropology of

Orvar Löfgren, University of Lund, Lund, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of everyday life is an interdisciplinary field that stands for a special interest in the commonplace and the seemingly trivial routines and activities of daily life as well as a focus on ordinary people as creative actors rather than passive consumers or objects of domination. Like few other similar concepts, the study of everyday life has been used not only as a research perspective or analytical tool but also as a research ideology. The study of everyday life developed in the postwar period as an alternative to, or critique of, mainstream traditions in the social sciences, often with a focus on anthropological, ethnographic, and qualitative approaches.

The Study of Everyday Life

From the 1960s onward, the concept of 'the study of everyday life' became an increasingly popular notion in many social sciences, such as anthropology, history, and sociology, as well as in the expanding field of cultural studies.

It has come to signal a special interest in the commonplace and the seemingly trivial routines and activities of daily life as well as a focus on 'ordinary people' as creative actors rather than passive consumers or objects of domination. The different ways in which this seemingly simple concept has been put to work illustrate changing concerns in social and cultural research, as well as conflicts between different research paradigms. Like few other similar concepts, the study of everyday life has been used not only as a research perspective or analytical tool but also as a research ideology. The study of everyday life often developed as an alternative to, or critique of, mainstream traditions.

The use of the concept as an alternative approach means that it is constantly being constructed in opposition to something else. It can be organized as the study of the life and thoughts of ordinary or common people against that of elites, intellectuals, or experts. It can be defined as the routines, habits, and phenomena that tend to be neglected because their 'everydayness' makes them invisible or self-evident, unlike colorful rituals, celebrations, or other activities with high visibility or prestige. Here, it is 'the everyday' as contrasted to 'the Sunday best' in social life, working with polarities like mundane/sacred, informal/formal, repetitive/unique, and trivial/special. Everyday life can further be seen as the arena of expressions and meanings threatened by invasion from above, by institutions like the state, the market, or just 'the system.' This perspective favors a view of everyday life as resistance against domination. The focus on resistance is often linked to a view of everyday life as a site of cultural creativity. Finally, 'the study of everyday life' often defines a research orientation rather than an empirical field. It is then seen as a certain style of doing research, usually with a focus on qualitative and ethnographic methods – exploring the hidden worlds of daily life. The argument is often that by focusing on seemingly banal phenomena new and sometimes surprising analytical insights into major social or cultural issues may be produced.

The study of everyday life has emerged as an arena of intense cross-disciplinary dialogues, but it is also an interest that has developed in a constant interchange with new public and political ideas about the importance of everyday life (see Gullestad, 1991) (*see* Common Sense, Anthropology of).

The Making of a Concept

The development of everyday life as a research topic is thus a complex one. It has entered the social sciences along many paths and in this process been charged with very different theoretical and ideological meanings. There are two nineteenth-century traditions that can be seen as starting points for the twentieth-century interest in the study of everyday life. The first of these was the works of those cultural historians who in the late-nineteenth century turned to the everyday life of 'ordinary people' to create an alternative to the classic history of kings and elites. The second tradition was the interest among early social researchers and reformers in exploring the lives of urban poor or rural peasants, starting with Henry Mayhew's classic work *London Labor and the London Poor* of 1861.

Sociological Approaches

It was this latter interest in questions of social welfare that influenced the interwar ethnographic studies of the so-called Chicago School of American Sociology. An early example is Nels Anderson's book *The Hobo: The Sociology of Homeless Men* of 1923.

This interest in urban everyday life of the Chicago School was developed further by the new generations of postwar American sociologists. Behind the new label 'the sociology of everyday life' very different theoretical traditions and research perspectives were found. One was the symbolic interactionism of Erving Goffman's (1959) pioneering work *The Presentation of Self in Everyday Life*, with its qualitative, ethnographical approach to the seemingly trivial or invisible routines of the everyday as it surfaced in patterns of social interaction. For early phenomenological contributions to this sociology of everyday life, see, for example, Schutz (1973) and Douglas (1971). The third new approach was ethno-methodology (see Garfinkel, 1984), mainly devoted to understanding the construction of

common-sense knowledge in everyday life. Today, the sociology of everyday life usually stands for a qualitative and ethnographic approach (see Glassner and Hertz, 1999).

In France, the sociology of the everyday developed along very different lines. Here, the studies of popular mentalities in the Annales School of social history and especially the works of Fernand Braudel created an early analytical interest in the role of everyday life for understanding society. The sociologist and rather unorthodox Marxist Henri Lefebvre started to explore the concept of the everyday already in the interwar period and published a number of books on the topic between 1947 and 1991 (e.g., Lefebvre, 1971). His interest in everyday life was connected with the rapid modernization France underwent especially after the Second World War with the creation of a new consumer-based and standardized everyday life. He used the term 'everydayness' (*le quotidien*) to describe what he saw as the banality and conformity of modern everyday life and the ways in which it was colonized by modern capitalism. But Lefebvre's approach was not only centered around questions of alienation and domination. His treatment of the problems of 'everydayness' also stressed the ways in which social scientists underestimated the role of the everyday, not only as a site of domination but also as an arena with potentials for transcendence and change – the peak experiences or special moments embedded in it (see Shields, 1999).

Lefebvre's discussions of everyday life had a strong philosophical profile, which may be one of the reasons his work did not come to inspire empirical studies in the same way as the work of another French sociologist, Michel de Certeau (1984), whose book was very influential in shaping the approaches to the study of everyday life in many disciplines during the 1980s and 1990s. Lefebvre and de Certeau both shared an interest in the uses of space in everyday life, but the similarities end there. Unlike Lefebvre, de Certeau focused on everyday life as an arena of cultural creativity and resistance, and above all on consumption as creative cultural production.

Anthropological Approaches

In anthropology, the actual concept 'study of everyday life' was not at first used very frequently, in spite of the fact that anthropology often was seen as the science of the everyday with its ethnographic and qualitative explorations of life in different settings over the world. One reason for the lack of an explicit use of the concept of everyday life was that it was seen as a cultural category belonging to the modern Western world. In non-Western societies, this polarity between the everyday and its counterparts, however defined, did not seem very relevant. In the broad anthropological approach to the study of culture, there was no need for a subcategory labeled everyday culture.

In many ways, the interdisciplinary interest in studies of everyday life among historians and sociologists from the 1960s and onward meant an anthropologization of both approaches and theoretical frameworks. It was not only the ethnographic methods and the metaphors of 'discovering' and 'exploring' everyday life as a hidden but also the exotic culture that accounted for this anthropologization. The anthropological connection also had to do with studies of everyday life emphasizing a cultural perspective on society. It was in the habits, rituals, routines, and traditions of daily life that 'culture' was embedded, as invisible to the actors as 'water to the fish,' to quote a popular metaphor from the 1970s.

A much more explicit use of the category of the everyday was found in the anthropological branch of European ethnology, which from the 1960s onward redefined itself as a discipline in which the study of everyday life in European settings, past and present, was central. This transformation may serve as an example of how the concept came to be given strong symbolic and ideological charges (see Frykman and Löfgren, 1996).

As European ethnologists in the 1960s oriented their research toward modern industrial society rather than vanishing peasant traditions, the basic concept of folk culture was exchanged for the study of the everyday and 'the ordinary people.' These new concepts not only signaled a break with earlier generations but also worked as a boundary marker against neighboring disciplines. Ethnologists studied everyday life with a qualitative perspective unlike sociologists and historians; the concept of everyday life was used to show that the ethnological interest in society was not concerned with formal institutions and macrostructures. In reality, the concept was thus not about the contrast between everyday and non-everyday phenomena, but a way of stressing a research interest focusing on neglected groups, activities, and spheres.

Everyday Life and Cultural Creativity

Another research tradition in which the notion of everyday life became very important was the emerging field of cultural studies. As in European ethnology, this interest in everyday life during the 1980s and 1990s was linked closely to the new concept of cultural creativity.

Like the notion of everyday life, the concept of creativity functioned as a counterargument in cultural research. Groups and settings that at first glance seemed characterized by passivity, simply reproducing dominant patterns, were in the new studies coming out as culture builders and bricoleurs: busy reworking, elaborating, and thickening their everyday lives with new meanings and routines. This approach is very striking in de Certeau's work as well as in many of the studies made by the so-called Birmingham School of cultural studies. A good example is Willis' (1990) book *Common Culture: Symbolic Work at Play in the Everyday Cultures of the Young*. Here everyday creativity becomes not only a counterargument against the presupposition that creativity is an elite resource, a cultural capital of the gifted, of artists, writers, intellectuals, and so forth, but also against the notion that modern mass consumption is a passifying, homogenizing force in people's lives.

When consumption was redefined as 'cultural production' or 'symbolic production,' the creativity of everyday life moved in to occupy a central position in this production process. This emphasis on cultural creativity as resistance must be seen as a reaction against an earlier discourse on the seduction of mindless consumers: locked up in the iron cage held together by market forces. Creativity became, in some ways, the weapon of the weak – a positive strategy of resistance. More rarely, the concept was used to describe strategies of dominance and oppression; other terms were used for the creative and imaginative ways of control and manipulation.

The pairing of cultural creativity and everyday life can thus be seen as an attempt to redress the balance, to develop a more actor-oriented approach in studies of mass consumption. In this process, however, the 'when, where, how, why, and who' of cultural creativity in everyday life was narrowed down and produced certain paved roads: a tendency to over-research some aspects and arenas and overlook others (see Löfgren, 1997).

This slanting of the two concepts was not a very conscious process. It started with the focus on consumption as creativity, a creativity much focused on symbolic production and expressive forms in what Featherstone (1991) has called the aesthetization of everyday life. This very selective emphasis was made more narrow by a marked concentration on specific social settings, above all youth cultures with their constant experimentation with identities and reworking of mass media and mass consumption styles, images, and commodities.

With a selective focus like this, cultural creativity became a very positive concept. It was seen as enriching, elaborating, and 'thickening' everyday life, a process through which people make their daily existence colorful, unique, specific, distinct, and above all positive. The perspective usually was reserved for those defined as the underdogs of the modern world: consumers, workers, women, teenagers, and colonial and postcolonial subjects, and this is where the concept is linked closely to ideas of counter-hegemony, to the tactics of resistance in the world of mass consumption or in postcolonial processes of globalization.

Problems and Potentials

Just like everyday life, creativity became a concept mainly for the age of late modernity. The academic interest in the two concepts mirrors a public and political concern too – especially in Northern European and North American settings – which had its roots in the countercultures of the 1960s and 1970s and the interests in alternative lifestyles as well as the democratization of cultural expression. During this period, the idea of the importance of everyday life became a powerful political symbol, condensing and unifying a number of sometimes conflicting values (see Gullestad, 1991). Everyday life was 'where the action was,' the location of 'real or authentic life.' By studying and engaging this arena, silent groups and interests could be given a voice.

Such political and popular interests influenced academic research. Some activities and people easily came to be seen as more everyday than others. Another danger of this perspective was that everyday life was too easily seen as something 'down there,' in a way that seldom reflected upon the problem of what the opposite of the everyday was supposed to be. It was forgotten that everyday life existed not only in retirement homes or on the streets but also in the corridors of power, that there were everyday practices not only in immigrant suburbs but also in scholarly research projects. Researchers who sometimes said that they wanted to 'get out into everyday reality' missed the point that what is interesting about the everyday is that it is ever-present.

Taking such ambiguities and possible ideological overtones into account, research interest in the study of the everyday has on the whole been important both in revitalizing and in redirecting scholarly attention to the role of the seemingly trivial activities and ideas embedded in the commonplace. During the last decades of the twentieth century, the study of everyday life constituted a very productive cross-disciplinary field, developing innovative research strategies for destabilizing or problematizing the taken-for-grantedness of much daily life.

The study of the everyday has also been a strategy to avoid paved roads in research and develop back-door entrances to major issues. Two examples may illustrate such strategies. The first is taken from the French sociologist Kaufmann's ethnographies of daily life, his study of the conflicting values of modern marriages analyzed through the routines of doing the laundry in a sample of French families (Kaufmann, 1998). Through extensive interviews with wives and husbands about their laundry habits, Kaufmann throws new light on the constant negotiations typical of modern marriages, negotiations about divisions of labor, definitions of love and individual space, as well as changing perceptions of the moral economy of the family.

Similar examples of such a use of back-door entrances to major issues can be found in those studies of nationalism, which do not take ideology or official rhetoric of major institutions as their starting point, but want to understand the power of 'the national' through the modes in which it is anchored and materialized in everyday life. Such studies, for example, Billig (1995), illustrate the powerful ways in which the nation-state has been made visible and tangible in a wealth of everyday situations and settings, and how citizens have been touched by the state.

Analyzing Undercurrents

During the early 2000s, there was a renewed interest in studies of the everyday, often as an interdisciplinary dialogue between anthropologists and sociologists. One reason for this was the wish to develop new theoretical and methodological tools for analyzing the nondiscursive dimensions of social life as well as the interaction with the material world. A good example of this focus is the renewed interest in mundane routines and habits as an often-unconscious undercurrent in everyday life (for an overview, see Shove et al., 2009). Studying everyday routines is methodologically challenging, as they often rest in the body rather than in the conscious mind, naturalized into reflexes. How are routines learned and how are they internalized into implicit knowledge? In what kinds of situations do they surface? What happens when they are contested? Questions like these came to engage many scholars (see the discussion in Ehn and Löfgren, 2010: 71ff.)

Another reason behind this new interest in routines has to do with the fact that their seemingly insignificance or invisibility often hides issues of power, freedom, and control. By looking ethnographically closer at routines, it becomes possible to destabilize the notion that they are about 'just going through the same movements.' The repetitious nature of routines often hides important micro-changes that eventually may transform them into something else, as Tom O'Dell (2006) has pointed out.

Over the last years, the classic discussions of routines and habituation in the works of scholars like Pierre Bourdieu (1977), Michel de Certeau (1984), and Norbert Elias (1939/1978) may be contrasted with some more recent

studies such as those by Ben Highmore (2010), Elizabeth Shove et al. (2007), and Sarah Pink (2012), who have developed a more nuanced understanding of how routines work in everyday life.

Studies like these work as a reminder that the seemingly insignificant may be a good starting point for exploring the workings of society. Studies of everyday life have produced techniques and methods for understanding how a lot of cultural energy may be condensed or crystallized in a trivial situation, an unperceived object, or a prosaic routine. They have also led to a much greater reflexivity about the manners in which the study of the everyday is conditioned by the researcher's own embeddedness in his or her own everyday practices.

See also: Collective Beliefs, Sociological Explanation of; Collective Memory, Anthropology of; Ethnology; Family: The Anthropology of the Concept and Its History; Folklore; Hegemony: Anthropological Aspects; Home, Anthropology of; Play, Anthropology of; Symbolic Interaction: Methodology; Tradition, Anthropology of.

Bibliography

Billig, M., 1995. Banal Nationalism. Sage, London.
Bourdieu, P., 1977. Outline of a Theory of Practice. Cambridge University Press, Cambridge.
de Certeau, M., 1984. The Practice of Everyday Life. University of California Press, Berkeley, CA.
Douglas, J. (Ed.), 1971. Understanding Everyday Life. Routledge and Kegan Paul, London.
Ehn, B., Löfgren, O., 2010. Doing Nothing? An Ethnography of Waiting, Daydreaming and Other Non-events. University of California Press, Berkeley, CA.
Elias, N., 1939/1978. The Civilizing Process. Blackwell, Oxford.
Featherstone, M., 1991. Consumer Culture and Postmodernism. Sage, London.
Frykman, J., Löfgren, O. (Eds.), 1996. Force of Habit: Exploring Everyday Culture. Lund University Press, Lund, Sweden.
Garfinkel, H., 1984. Studies in Ethnomethodology. Polity Press, Oxford, UK.
Glassner, B., Hertz, R. (Eds.), 1999. Qualitative Sociology as Everyday Life. Sage, Thousand Oaks, CA.
Goffman, E., 1959. The Presentation of Self in Everyday Life. Doubleday, Garden City, NY.
Gullestad, M., 1991. The transformation of the notion of everyday life. American Ethnologist 18 (3), 480–499.
Highmore, B., 2010. Ordinary Lives: Studies in the Everyday. Routledge, London.
Kaufmann, J.-C., 1998. Dirty Linen. Couples as Seen Through Their Laundry. Middlesex University Press, Middlesex, UK.
Lefebvre, H., 1971. Everyday Life in the Modern World. Harper & Row, New York.
Löfgren, O., 1997. Scenes from a troubled marriage: Swedish ethnology and material culture studies. Journal of Material Culture 2 (1), 95–113.
O'Dell, T., 2006. Cultural backdraft. In: Löfgren, Orvar, Wilk, Richard (Eds.), Off the Edge. Experiments in Cultural Analayis. Museum Tusculanum Press, Copenhagen, pp. 113–118. (Also published as Ethnologia Europaea 35, pp. 1–2.)
Pink, S., 2012. Situating Everyday Life: Practices and Places. Sage, London.
Schutz, A., 1973. Structures of the Life-World. Northwestern University Press, Evanston, IL.
Shields, R., 1999. Lefebvre, Love and Struggle: Spatial Dialectics. Routledge, London.
Shove, E., Watson, M., Ingram, J., Hand, M., 2007. The Design of Everyday Life. Berg, Oxford.
Shove, E., Trentmann, F., Wilk, R., 2009. Time, Consumption and Everyday Life. Berg, Oxford.
Willis, P., 1990. Common Culture. Westview Press, Boulder, CO.

Evidence-Based Medicine

Matthias Egger, University of Bern, Bern, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Evidence-based medicine (EbM) was developed in the 1990s and was quickly adopted worldwide. EbM is defined as the integration of current best evidence from patient-centered clinical research, clinical experience, and patient values and preferences in making decisions about the care of individuals. Evidence-based medicine proceeds in four steps: (1) the formulation of a precise clinical question, (2) the identification of the best available evidence, (3) the critical appraisal of the evidence, and (4) its application in clinical practice. Systematic reviews of relevant studies and their statistical synthesis in meta-analysis are important resources for evidence-based health care. Evidence-based medicine has been described as "one of modern medicine's greatest intellectual achievements" but some resent the limitation of the freedom of doctors to treat a patient in the way they believe is best.

Introduction

Evidence-based medicine (EbM) was first proposed as 'a new approach to teaching the practice of medicine' in 1992 (Evidence-Based Medicine Working Group, 1992), but its roots extend back to the 1950s and earlier. The application of epidemiological thinking and methods to study medical practice and health services from the 1950s onwards, pioneered by Alvan Feinstein (1925–2001) and Thomas C. Chalmers (1917–95), which later became known as *Clinical Epidemiology* (Sackett, 1991), the rise of randomized controlled trials to evaluate medical interventions, and the introduction of systematic reviews and meta-analyses in the 1990s were all instrumental. The 1972 book by Archibald Cochrane (1909–88) on *Effectiveness and Efficiency – Random Reflections on Health Services* (Cochrane, 1972) inspired Iain Chalmers and others to set up the *Cochrane Collaboration*, which assists providers, policy makers, and patients by preparing and disseminating systematic reviews of health care interventions. Richard Smith and Drummond Rennie, who were involved as editors of medical journals, have recently created an oral history of EbM on video, with perspectives from individuals who played pivotal roles in its development (Smith and Rennie, 2014).

EbM stands for a practice of medicine that is based both on expert opinion and clinical skills and up-to-date data from relevant patient-oriented research. Proficiency in how to interpret the results from such clinical studies and how to use them for making decisions in the context of the care of an individual patient are as important as background knowledge on pathophysiological mechanisms. David Sackett, who is widely regarded as the 'father of EbM,' and colleagues defined EbM as,

> the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research.
>
> Sackett et al., 1996

How accurate is exercise electrocardiography in the diagnosis of coronary heart disease? What is the life expectancy of HIV-infected individuals in the era of potent antiretroviral combination therapy? How efficacious is an inhibitor of the BCR-ABL tyrosine kinase in older men with chronic myeloid leukemia? EbM critically evaluates diagnostic and screening tests, prognostic information, and therapeutic interventions to answer these types of questions. To be useful, such research must be based on patient-centered outcomes, i.e., outcomes that people care about, such as survival, physical, and cognitive function, symptoms, and health-related quality of life. The available evidence is then systematically reviewed and, if appropriate, statistically combined in meta-analyses (Egger et al., 2001). The body of data is finally graded in terms of its validity and used in guidelines to inform clinical practice (Guyatt et al., 2008).

Clinical expertise and skills are essential in many areas of clinical practice, for example, when making a diagnosis after careful history taking and clinical examination of the patient, or by integrating the patient's preferences in making decisions about his or her care, based on the experience gained over years of working with patients with similar symptoms and conditions (Straus, 2011). Values and preferences vary widely between patients. They may depend on patients' age and other sociodemographic characteristics, their prior experience with treatments or health outcomes, but also on the methods used to assess preferences (MacLean et al., 2012). **Figure 1** summarizes the three key elements contributing to EbM.

Why Do We Need EbM?

You might ask, haven't doctors always practiced EbM? Perhaps surprisingly, the answer is no: before the advent of EbM the use of critically appraised results from patient-oriented research was less common and less systematic. The opinion of experienced clinicians, based on their authority and status, was often uncritically accepted, even in the absence of good data from clinical research supporting these views. Physicians practiced medicine primarily based on what was taught at medical school, their experience and local habits, leading to striking variations.

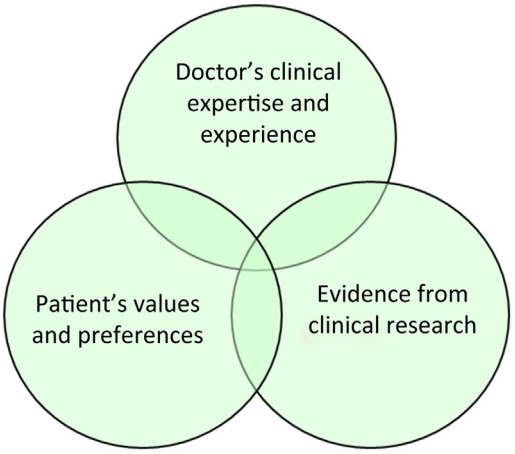

Figure 1 Key components of EbM.

Terms like 'expert-based,' 'opinion-based,' or 'eminence-based' medicine are sometimes used to describe this situation. Much emphasis was given to pathophysiological reasoning, data on potential mechanisms and surrogate endpoints.

Surrogate endpoints are measures of intermediate effects of a treatment that correlate with a patient-centered outcome; for example, cardiovascular risk factors such as blood pressure or cholesterol levels, which are associated with cardiovascular disease. The short-term change in a risk factor observed with a drug is then interpreted as if the long-term health benefits corresponding to this change are guaranteed to follow (Fleming and DeMets, 1996). The widespread use in the 1980s of flecanaide, a type-1 antiarrythmic drug, to prevent sudden cardiac deaths is one of many examples where this was erroneously assumed. Sudden deaths are often caused by disturbances of the heart beat (arrhythmia) or cardiac arrest. The risk is higher in people who previously had a heart attack. Many such patients were therefore given flecanaide or similar drugs prophylactically, particularly in the United States. The drugs were effective in suppressing ventricular premature beats (the surrogate endpoint), but a large randomized controlled trial, the Cardiac Arrhythmia Suppression Trial (CAST), showed that total mortality was higher with flecainide than with placebo (Echt et al., 1991).

In his book entitled *Deadly Medicine*, the journalist Thomas Moore argued that "tens of thousands of heart patients died in America's worst drug disaster" (Moore, 1995). Although the accuracy of Moore's figures is disputed, it is clear that many patients were exposed to an increased risk of death before the CAST trial stopped this practice. Other examples of failed surrogate endpoints include tolbutamide and phenformin which lower blood glucose but were associated with a higher risk of death, or fluoride which enhances bone mineral density, but caused more fractures overall (Fleming and DeMets, 1996).

Key Steps

EbM proceeds in four key steps from formulating a clinical problem to solving it based on the best available evidence. These steps are discussed in the following sections.

Formulate Precise Clinical Questions

The formulation of a precise question is key to clarifying the nature of the problem and to avoid getting loss in a huge amount of literature later on. It is helpful to develop the question in terms of the patient group concerned, the intervention (typically drugs or other medical interventions), the comparison (standard care, or no intervention), and the (patient-centered) outcome. The mnemonic *PICO* is often used in this context: *Patient, Intervention, Comparison and Outcome*. For example, an elderly woman who suffered a myocardial infarction 4 years ago reports breathlessness and swollen ankles, i.e., typical signs of heart failure. She has been taking a nonsteroidal anti-inflammatory drug (NSAID) for the past month because of knee pain. A relevant question in this situation might be as follows: In elderly women with a history of coronary heart disease who developed signs of heart failure on NSAIDs, is it sufficient to stop the anti-inflammatory drug? Should drugs against heart failure, for example diuretics, be given immediately? Are diuretics sufficient or should they be combined with other drugs?

Identify the Best Available Evidence

Once the question has been defined the search for relevant literature can start in order to identify the relevant evidence. This is an iterative process which involves the definition of appropriate search terms in electronic databases such as *PubMed* or *Embase*, based on the components of PICO. For example, let's assume our question is whether insulin pumps improve the control of blood glucose and the quality of life in children with type I diabetes, in comparison with daily insulin injections. The search terms will relate to study populations of children with type I diabetes, different types of insulin pumps as the intervention of interest compared to conventional care with daily insulin injections, and outcomes levels of hemoglobin A1c, hypoglycemic episodes and quality of life. We use Boolean operators to combine search terms in the databases: OR to capture several terms for one concept (for example 'insulin infusion system' OR 'insulin pump' OR 'artificial pancreas') and AND to combine the different searches for the components of PICO.

It is of course important to focus on the best possible studies, but which are the best studies? Several systems have been developed to classify studies according to their vulnerability to bias. **Figure 2** shows the *hierarchy of evidence* proposed by the US Agency for Health Care Policy (US Department of Health and Human Services, Public Health Service, Agency Health Care Policy and Research, 1992) to establish causal effects of medical interventions. Among different study designs, randomized controlled trials (RCTs) provide the highest level of evidence: in large trials randomization will ensure that the groups to be compared are similar, thus preventing any distortion of results due to an imbalanced distribution of prognostic factors. Studies that compare patients that received a drug with individuals who received no treatment or another drug are at high risk of bias. People receiving a drug are generally sicker than those who do not, and there are good reasons why one drug and not another is prescribed; reasons that are often associated with the severity and prognosis of the disease. In well-designed cohort studies, case–control studies or

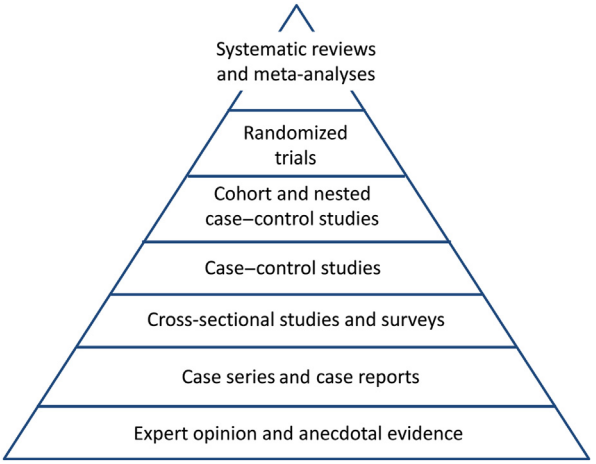

Figure 2 Levels of evidence.

before–after studies such biases can be reduced, but only randomization can guarantee that prognostic factors are balanced between groups. Several RCTs are required to document the consistency of effects in different patient groups and health care settings. They are ideally summarized in systematic reviews and, where appropriate, statistically combined in meta-analyses. Search terms denoting study designs (for example, 'randomized controlled trial' or 'meta-analysis' as publication types in the PubMed database) are therefore often combined with the PICO-related search described above.

Critically Appraise the Evidence

Although we may have focused on the study designs that minimize bias in our search of the literature, we still need to critically appraise the relevant studies one by one in terms of internal and external validity. Internal validity describes the extent to which the results are valid for the patients included, i.e., not distorted by bias or confounding. External validity relates to the extent to which the results can be applied to different patient groups and settings. For example, can the findings from a study done in a University tertiary care clinic be applied to a small district hospital, or to primary care?

Appraising internal validity is a matter of understanding the strengths and limitations of the design, conduct and analysis of a study. Internal validity is a prerequisite for external validity: a flawed study cannot be externally valid. The often incomplete and intransparent reporting of studies is a problem when assessing the internal and external validity of studies. Important methodological details and details on the study population and intervention are often missing from the scholarly articles. In recent years this situation has been addressed by the development of reporting guidelines for different study types. A network to Enhance the QUAlity and Transparency Of health Research (EQUATOR) has built a comprehensive searchable library of reporting guidelines, which is freely available on the web (The EQUATOR Network, 2014).

Bias is a systematic deviation of a study's result from a true value. Typically, bias is introduced during the design or implementation of a study and cannot be remedied later. Bias arises from flawed information or flawed patient selection so that associations are found that do not exist in reality. Confounding produces relations that are real, but do not reflect causal relationships because of some other factor that is not accounted for. Bias and confounding need to be distinguished from random error, a deviation from a true value caused by statistical fluctuations in the data (Vandenbroucke et al., 2007).

Many possible sources of bias have been described and a variety of terms are used. Two simple categories are helpful: information bias and selection bias (Vandenbroucke et al., 2007). Information bias occurs when systematic differences in the completeness or the accuracy of data distort results. Selection bias is introduced if the groups that are being compared in a study are not representative of those the results will be applied to. For example, in clinical trials attrition due to loss of participants during the study may introduce selection bias: those who stay in the study often differ from those who drop out, for example with respect to severity of disease, or side effects of drugs. The data available at the end of the study may thus have become biased. Confounding means confusion of effects. A study might seem to show either an association or no association between an exposure and the risk of a disease. In reality, the seeming association or lack of association is due to another factor that increases the risk of the disease but that is also associated with the exposure. The other factor is called the confounding factor or confounder. Confounding thus gives a wrong assessment of the potential 'causal' association of an exposure. For example, a study might find that yellow fingers increase the risk of developing lung cancer. The confounding factor here is of course heavy smoking, which makes your fingers go yellow and also increases your risk of lung cancer (Smith and Phillips, 1992). Proper randomization of patients to comparison groups in clinical trials prevents confounding. In other studies confounders can be taken into account in the analysis, but results might still be distorted by residual confounding (the confounding that remains after attempts to control for it).

The Cochrane Collaboration developed a helpful 'Risk of Bias' tool to assess the methodological quality of RCTs (Higgins et al., 2011). Useful checklists for other types of studies are available from the Critical Appraisal Skills Program in the United Kingdom (Critical Appraisal Skills Program (CASP), 2014).

Apply the Evidence in Clinical Practice

As Archibald Cochrane observed in 1972, "Between measurements based on RCTs and benefit … in the community there is a gulf which has much been under-estimated" (Cochrane, 1972). Unlike the assessment of the internal validity of a study, there are no accepted guidelines on how the external validity, generalizability, or applicability of studies should be assessed. Assessing external validity is a matter of judgment, based on a careful assessment of the setting of trials (for example, the level of care or the type of health care system), the selection of patients (the eligibility and exclusion criteria), the characteristics of the patients included in the study (for example, the severity of disease), the difference between the trial protocol and routine practice (for example, the timing and dose of the drug), and other factors (Rothwell, 2005). The consistency or inconsistency of results across different studies is

an important consideration, which supports or questions the generalizability of results to routine clinical practice.

Clinical decisions can then be made, taking into account the values or preferences of patients. The critically appraised body of evidence also forms the basis for developing guidelines and recommendations. The GRADE (Grading of Recommendations Assessment, Development and Evaluation) system is often used for this purpose (Guyatt et al., 2008). GRADE is a system for rating the quality of the evidence and the strength of recommendations in guidelines, which is increasingly adopted by organizations worldwide, including the World Health Organization.

Resources and Tools for Evidence-Based Medicine

Systematic Reviews and Meta-analyses

Systematic reviews and meta-analyses are an important source of evidence for EbM. Systematic reviews are 'studies of studies.' Like any study of patients or healthy volunteers, they are done using a systematic approach to minimize bias and random error, which is described in a study protocol. The protocol details the eligibility criteria for studies to be included and describes the search strategy to identify potentially eligible studies, and their critical appraisal. Systematic reviews may, or may not, include meta-analysis, a statistical combination of results from several studies to produce a single estimate of the effect of an intervention. Systematic reviews allow for a more objective appraisal of the evidence than traditional, narrative reviews and might help to resolve uncertainty and identify areas where further studies are needed. Meta-analysis will enhance the precision of estimates of intervention effects, leading to reduced probability of false negative results, and potentially to a timelier introduction of effective interventions. Methods for analyzing the evidence that is available on the benefits and harms for several drugs or an entire class of drugs (for example, NSAIDs) have also become available. Such *network meta-analyses* simultaneously analyze direct evidence from trials comparing two drugs (for example, A vs C, or B vs C), and indirect evidence (A vs B) and are increasingly used in *comparative effectiveness research*, which aims to provide evidence on the benefits and harms of different treatment options (Cipriani et al., 2013).

The Cochrane Library

The *Cochrane Library*, the main product of the Cochrane Collaboration, is a prominent source of systematic reviews of health care interventions ('Cochrane reviews'). At the time of writing, the database included over 5000 systematic reviews as well as a large collection of randomized controlled trials. The logo of the Cochrane Collaboration (**Figure 3**) shows a meta-analysis of seven RCTs of a short, inexpensive course of a corticosteroid given to women about to give birth too early. Each horizontal line represents the results of one trial (the shorter the line, the more certain the result); and the diamond represents their combined results. The position of the diamond to the left of the vertical line indicates that the treatment studied is beneficial. Indeed, this treatment reduces the odds of the babies of these women dying from the complications of immaturity by 30–50%. As no systematic review and meta-analysis of these trials had been published until 1989, most obstetricians had not realized that the treatment was so effective, and many premature babies have probably suffered and died unnecessarily (The Cochrane Collaboration, 2014).

Figure 3 The logo of the Cochrane Collaboration.

The Campbell Collaboration, named after Donald Campbell (1916–96), an American psychologist and thinker, was set up to prepare systematic reviews of high-quality research on effective methods and interventions in the fields of social welfare and social work, education and learning, and crime and delinquency (The Campbell Collaboration, 2014).

Other Resources

Numerous resources on evidence-based health care are available, including books, databases offering online access to medical evidence, journals dedicated to EbM, and resources for patients. Applications customized for use on mobile devices are also increasingly available. For example, the *National Institute for Health and Care Excellence* (NICE) in the United Kingdom provides primary care practitioners with summaries of the current evidence base and practical guidance in respect of over 300 conditions relevant to primary care. Produced by the BMJ Group, *Best Practice* summarizes the evidence from relevant systematic reviews, guidelines on the diagnosis, prognosis, treatment, and prevention for a large number of clinical specialties, ranging from psychiatry to cardiac surgery. The *Physiotherapy Evidence Database* (PEDro) provides access to over 27 000 RCTs, systematic reviews, and clinical practice guidelines in physiotherapy. The website of the Cochrane Collaboration provides an excellent 'webliography' of EbM resources (The Cochrane Collaboration, 2014).

Critique, Limitations, and Debates

EbM has recently been described as "one of modern medicine's greatest intellectual achievements" (Smith and Rennie, 2014), but its reception was mixed in the 1990s and critical views continue to be voiced. Sackett and colleagues addressed some of points raised in an editorial with the title "Evidence based

medicine: what it is and what isn't" (Sackett et al., 1996): the pronounced variations observed in clinical practice demonstrated that not 'everyone is already doing it' and the experience from frontline medicine documented the feasibility of EbM. Furthermore, Sackett and colleagues argued that EbM is not 'cookbook' medicine because it integrates evidence from patient-centered research with expertise, experience, and patient choice, and that EbM is not restricted to evidence from RCTs but considers other studies to answer relevant questions, for example, cohort studies on prognosis (Sackett et al., 1996).

In his oral history of EbM, Iain Chalmers expressed anger that in medical school he was "not given the tools to find out what worked" in medical practice. Many of his generation share this anger: the role of EbM in medical education cannot be overestimated. However, medicine is not an exact science and neither is EbM: assessing the quality of the evidence and applying the 'best evidence' from clinical research to an individual patient is to some extent subjective, but this is not always fully appreciated by EbM enthusiasts. Ironically, the properties of some of the tools promoted for EbM have not been well evaluated, for example, the GRADE system, and others are known to produce misleading results in some situations, for example, meta-analysis (Egger and Smith, 1995). Some resent that the freedom to treat a patient in the way the doctor believes is best has been limited by EbM, and argue that "patients are individuals, and need individual treatment" (Hampton, 2002). This view will likely gain momentum with the advent of personalized medicine, where treatments are tailored to the individual's genomic makeup. The potential of EbM will strongly depend on the health service and the values and attitudes of the public, but EbM is often discussed in isolation. For example, in the United States or Switzerland fee for service payments can create strong incentives to adopt interventions that are well reimbursed, even if the evidence base is poor. Finally, sometimes you meet colleagues, perhaps on guidelines committees, who appear to practice EbM as a religion and are at risk of losing sight of common sense. This sentiment was expressed in an entertaining article on the efficacy of parachutes to prevent 'death related to gravitational challenge,' which ended with the recommendation that the most radical protagonists of EbM should participate in a double blind, randomized, placebo controlled, crossover trial of the parachute (Smith and Pell, 2003). While such a trial should probably be avoided, it is clear that much has been achieved and much remains to be done to further develop the principles and practice of EbM and to better understand its potentials and limitations.

See also: Functioning, Disability and Health, International Classification of; Health Care, Rationing of; Health Interventions, Community-based; Therapeutic Effectiveness Assessment: Approaches in Contemporary Research.

Bibliography

The Campbell Collaboration [WWW Document], 2014. URL: http://www.campbellcollaboration.org/ (accessed 21.05.14.).

The Cochrane Collaboration [WWW Document], 2014. URL: http://www.cochrane.org/ (accessed 20.05.14.).

Critical Appraisal Skills Programme (CASP) [WWW Document], 2014. Crit. Apprais. Ski. Programme CASP. URL: http://www.casp-uk.net/ (accessed 19.05.14.).

Cipriani, A., Higgins, J.P.T., Geddes, J.R., Salanti, G., 2013. Conceptual and technical challenges in network meta-analysis. Annals of Internal Medicine 159, 130–137. http://dx.doi.org/10.7326/0003-4819-159-2-201307160-00008.

Cochrane, A.L., 1972. Effectiveness and Efficiency: Random Reflections on Health Services. Nuffield Provincial Hospitals Trust, London.

Evidence-Based Medicine Working Group, 1992. Evidence-based medicine. A new approach to teaching the practice of medicine. Journal of the American Medical Association 268, 2420–2425.

Echt, D.S., Liebson, P.R., Mitchell, L.B., Peters, R.W., Obias-Manno, D., Barker, A.H., Arensberg, D., Baker, A., Friedman, L., Greene, H.L., 1991. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The cardiac arrhythmia suppression trial. New England Journal of Medicine 324, 781–788. http://dx.doi.org/10.1056/NEJM199103213241201.

Egger, M., Smith, G.D., 1995. Misleading meta-analysis. British Medical Journal 311, 753–754.

The EQUATOR Network [WWW Document], 2014. URL: http://www.equator-network.org/ (accessed 20.05.14.).

Egger, M., Smith, G.D., Altman, D.G., 2001. Systematic Reviews in Health Care Meta-analysis in Context. BMJ Books, London.

Fleming, T.R., DeMets, D.L., 1996. Surrogate end points in clinical trials: are we being misled? Annals of Internal Medicine 125, 605–613. http://dx.doi.org/10.7326/0003-4819-125-7-199610010-00011.

Guyatt, G.H., Oxman, A.D., Vist, G.E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., Schünemann, H.J., GRADE Working Group, 2008. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. British Medical Journal 336, 924–926. http://dx.doi.org/10.1136/bmj.39489.470347.AD.

Hampton, J.R., 2002. Evidence-based medicine, opinion-based medicine, and real-world medicine. Perspectives in Biology and Medicine 45, 549–568.

Higgins, J.P.T., Altman, D.G., Gøtzsche, P.C., Jüni, P., Moher, D., Oxman, A.D., Savovic, J., Schulz, K.F., Weeks, L., Sterne, J.A.C., Cochrane Bias Methods Group, Cochrane Statistical Methods Group, 2011. The cochrane collaboration's tool for assessing risk of bias in randomised trials. British Medical Journal 343, d5928.

MacLean, S., Mulla, S., Akl, E.A., Jankowski, M., Vandvik, P.O., Ebrahim, S., McLeod, S., Bhatnagar, N., Guyatt, G.H., American College of Chest Physicians, 2012. Patient values and preferences in decision making for antithrombotic therapy: a systematic review: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest 141, e1S–23S. http://dx.doi.org/10.1378/chest.11-2290.

Moore, T.J., 1995. Deadly Medicine: Why Tens of Thousands of Heart Patients Died in America's Worst Drug Disaster. Simon & Schuster, New York.

Rothwell, P.M., 2005. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 365, 82–93. http://dx.doi.org/10.1016/S0140-6736(04)17670-8.

Sackett, D.L. (Ed.), 1991. Clinical Epidemiology: A Basic Science for Clinical Medicine, second ed. Little, Brown, Boston.

Sackett, D.L., Rosenberg, W.M., Gray, J.A., Haynes, R.B., Richardson, W.S., 1996. Evidence based medicine: what it is and what it isn't. British Medical Journal 312, 71–72.

Smith, G.C.S., Pell, J.P., 2003. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. British Medical Journal 327, 1459–1461. http://dx.doi.org/10.1136/bmj.327.7429.1459.

Smith, G.D., Phillips, A.N., 1992. Confounding in epidemiological studies: why "independent" effects may not be all they seem. British Medical Journal 305, 757–759.

Smith, R., Rennie, D., 2014. Evidence based medicine – an oral history. British Medical Journal 348, g371. http://dx.doi.org/10.1136/bmj.g371.

Straus, S.E., 2011. Evidence-Based Medicine: How to Practice and Teach It. Elsevier Churchill Livingstone, Edinburgh.

US Department of Health and Human Services, Public Health Service, Agency Health Care Policy and Research, 1992. Acute Pain Management: Operative or Medical Procedures and Trauma. Agency for Health Care Policy and Research Publications, Rockville, MD (AHCPR Pub 92–0038.). Agency for Health Care Policy and Research Publications, Rockville, MD.

Vandenbroucke, J.P., von Elm, E., Altman, D.G., Gøtzsche, P.C., Mulrow, C.D., Pocock, S.J., Poole, C., Schlesselman, J.J., Egger, M., STROBE Initiative, 2007. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. PLoS Medicine 4, e297. http://dx.doi.org/10.1371/journal.pmed.0040297.

Evidence-Based Practice

Jeremy Steglitz, Jennifer L Warnick, Sara A Hoffman, Winter Johnston, and Bonnie Spring, Northwestern University, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Evidence-based practice (EBP) is an approach that aims to improve the process through which high-quality scientific research evidence can be obtained and translated into the best practical decisions to improve health. The interprofessional model of EBP emphasizes shared decision-making within the context of the most important advances of the various health professions. The model depicts three data streams that are integrated in the decision-making process: evidence, resources, and patient characteristics. Health professionals can play several different roles in the EBP process, including primary researchers, systematic reviewers, and clinicians. Carrying out the EBP process involves five steps, including Ask, Acquire, Appraise, Apply, and Analyze and Adjust. A new generation of research designs, such as the Sequential Multiphased Adaptive Randomized Trial, has been put forward to develop treatment algorithms that optimally capture the Apply, Analyze and Adjust steps of the EBP process.

Definition

Evidence-based practice (EBP) is an approach that aims to improve the process through which high-quality scientific research evidence can be obtained and translated into the best practical decisions to improve health. Research findings derived from the systematic collection of data through observation and experiment, as well as the formulation of questions and testing of hypotheses comprise the evidence supporting practice. EBP harmonizes the standards used to conduct, report, evaluate, and distribute research results so as to increase their application to practice and policy. EBP also involves the use of conscientious and explicit decision-making that integrates consideration of the best available research evidence, client characteristics (including preferences), and resources. Best available research is defined as contextually relevant and best in quality, according to consensually accepted scientific standards for different types of questions. Practical decisions relevant to EBP often involve the selection of an assessment or intervention. While professionals practicing evidence-based medicine (EBM) often need to choose among treatments involving drugs or devices, those practicing evidence-based behavioral medicine (EBBM) usually make selections among nondrug and nondevice behavioral or psychosocial interventions.

History of EBP

The origins of EBP are usually dated to 1910, when the American Medical Association and the Carnegie Foundation commissioned Abraham Flexner, a research scholar at the Carnegie Foundation for the Advancement of Teaching, to survey American and Canadian medical schools. The Flexner Report, as it was called, represented a major effort to reform medical education by placing it on a scientific foundation. Flexner surveyed 155 medical schools and severely criticized the training offered by many of them. His findings revealed that most medical schools offered lax clinical training, a curriculum not based on science, and a motivation that promoted profit rather than public service (Flexner, 1910). The Flexner Report established an educational quality standard that many of the existing medical schools could not meet. Therefore, more than half of all medical schools closed by 1935 (Beck, 2004). This report is widely regarded as the start of the EBM movement.

A second main catalyst for the EBM movement came from Archibald Cochrane, a British epidemiologist who aimed to establish a rational, systematic basis for determining what treatments should be covered by health care (Cochrane, 1972). Cochrane argued that because resources for health care are inevitably limited, it is essential that scarce dollars be allocated only for procedures of demonstrated worth. He argued that randomized controlled trials (RCTs) offer the most unbiased, reliable method to evaluate the effectiveness of treatments, warranting their placement at the top of a hierarchy of evidence. Accordingly, findings from high-quality RCTs are given greater credence than those from observational studies, case studies, and expert opinion when determining whether a treatment is effective. Followers of Cochrane's work subsequently established the Cochrane Collaboration (www.cochrane.org), a worldwide network that tracks, critically appraises, and synthesizes results of RCTs, publishing their findings online.

In the 1990s, a group of clinical epidemiologists working at Canada's McMaster University under the direction of David Sackett and Gordon Guyatt spearheaded the third initiative that catalyzed the EBM movement. This group's mission was to close the research-to-practice gap by encouraging physicians to engage in lifelong learning about new research findings (Sackett and Rosenberg, 1995a,b). The McMaster group was motivated by evidence that health professionals primarily implement treatment practices learned during training but neglect new and often more efficacious treatments that emerged subsequently (Isaacs and Fitzgerald, 1999). To change this habit, the McMaster group developed methods for health professionals to find, assess, and apply research results. However, the group encountered resistance from health professionals who believed that exclusive practice based upon

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10540-9

research findings devalued clinical expertise and experience (Haynes et al., 1996). To overcome this perceived slight against practicing clinicians, Guyatt et al. (1992) renamed the approach 'evidence-based medicine' in place of 'scientific medicine.' In this newest model, EBM was presented as an approach that tied together and utilized research, patient characteristics, and expertise to formulate best treatment practice, as opposed to relying solely on research findings (Haynes et al., 1996; Sackett et al., 1996).

EBM to EBBM

To evaluate interventions besides drugs or devices, health professionals in the behavioral sciences also needed a standard to evaluate behavioral treatments. The first entity to take on this task of conceptualizing EBBM was the Society of Behavioral Medicine's EBBM Committee. Established in 2000 with support from the National Institutes of Health (NIH) Office of Behavioral and Social Science Research (OBSSR) under Acting Director, Peter Kaufmann, the first EBBM Committee, first chaired by Karina Davidson, defined its scope to include behavioral interventions that prevent disease, promote health and adherence to treatment, or change biological determinants of behavioral conditions (Davidson et al., 2003). Initially, this committee familiarized behavioral medicine researchers with the Consolidated Standards of Reporting Trials (CONSORT) guidelines that encourage comprehensive, transparent reporting of RCTs in medical journals (Schulz et al., 2010). This effort was one of many that led behavioral science journals to adopt the CONSORT guidelines for publishing clinical trials. Among the first behavioral science journals to adopt CONSORT were *Annals of Behavioral Medicine*, *Health Psychology*, *International Journal of Behavioral Medicine*, and the *Journal of Consulting and Clinical Psychology*. The EBBM committee also addressed other weaknesses in the quality of behavioral clinical trials, especially with regard to the analytic approach. Numerous behavioral treatment trials were found to have analyzed data only from those who completed the final assessment in a clinical trial or who experienced a full dose of treatment (Pagoto et al., 2009; Spring et al., 2007). Since then, use of the intent-to-treat policy has increasingly become normative in behavioral science, such that data from all randomized participants are included in study analyses according to the condition to which they were assigned.

EBBM Evolves to EBBP, and Then to EBP

By 2006, the US health-care crisis was in full swing. With it came the need for a better, more integrated system of care that addressed mental as well as physical health for the sick and prevention for the well. The only way to accomplish all of this was by the coordinated efforts of an interprofessional team. It became clear that the EBBM approach needed to be upgraded to include all health professionals. Thus, OBSSR sponsored the Council on Evidence-Based Behavioral Practice (EBBP), chaired by Bonnie Spring, and its scientific and clinician advisory boards. The composition of the Council and the Boards was determinedly interprofessional, combining representatives from medicine, nursing, psychology, social work, public health, and information sciences (www.ebbp.org). The Council's first task was to formulate a conceptual model that could accommodate the diverse historic traditions as well as the individual- and population-level behavioral interventions that different health professions implement.

Initially, the conceptual model for EBM emphasized only a single parameter: research (Sackett et al., 1996). Later, the EBM model expanded to include other considerations such as clinical experience and specific patient needs. The EBM definition stated that: "evidence-based medicine requires the integration of the best research evidence with clinical expertise and the patient's unique values and circumstances" (Strauss et al., 2005). EBBM expanded upon EBM by adding nondrug, nondevice treatments. EBBP went one step further by consolidating across different disciplinary frameworks for EBP. The goal was to develop a conceptual model that could be shared by a more diverse interprofessional health-care team, whose members all require core competency in EBP (Greiner and Knebel, 2003). This shared EBP model supports jointly held vocabulary, foundational assumptions, and practice principles that unite the team of professionals in medicine, nursing, psychology, social work, public policy, and information sciences. A unified EBP eliminates the need to have separate models for different disciplines (Satterfield et al., 2009; Spring and Hitchcock, 2009).

Interprofessional Model of EBP

The interprofessional model of EBP (**Figure 1**) emphasizes shared decision-making within the context of the most important advances of the various health professions, including those mentioned above. The model depicts three data streams that are integrated in the decision-making process: evidence, resources, and patient characteristics. The interprofessional EBP model is grounded in an ecological framework that emphasizes the importance of considering environmental and organizational spheres when conceptualizing the problem and designing a course of treatment.

Best Research Evidence

Evidence refers to research findings from the systematic collection of data through observation and experiment grounded in

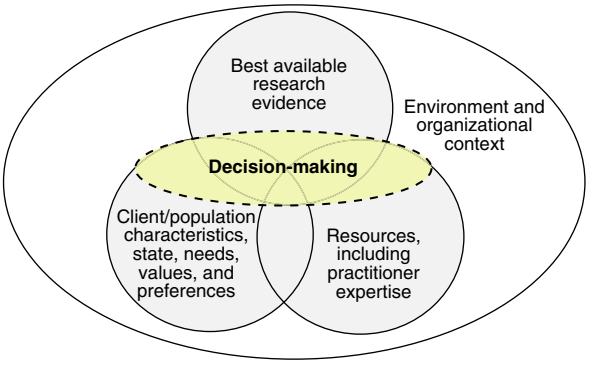

Figure 1 Interprofessional model of EBP.

the formulation of questions and hypothesis testing. That, which is deemed best research evidence is contingent on the particular question that needs to be addressed (Sackett and Wennberg, 1997). A question concerning etiology or prognosis, for example, is optimally answered through a longitudinal cohort research study design. On the other hand, a question about efficacy and effectiveness of treatments benefits from an RCT that is less susceptible to bias and error. Treatment-based questions can be addressed particularly well with the systematic review, which synthesizes the findings from multiple RCTs (Oxford Center for Evidence-Based Medicine, 2001). Recent interest in personalized care has generated interest in contextualized research evidence that is uniquely applicable to a particular patient and practice context (Weaver et al., 2005; Westfall et al., 2007). This has led to a resurgence of interest in the single-case experimental design (Dallery et al., 2013), which some representations of the evidence hierarchy place at the top of the evidence pyramid (Figure 2).

Resources

Resources refer to the skills and infrastructure that are required to provide EBPs. Resources needed to deliver treatments include physical, technological, financial, and personnel assets (e.g., office space, technological support, insurance reimbursement, and expert health professionals trained in an evidence-based treatment). Additional resources may include institutional endorsement by higher administration and agreement from other system components to make a treatment available.

The interprofessional EBP model breaks down clinician expertise into four categories of skill: assessment skills, EBP process skills, communication and collaboration skills, and engagement and intervention skills.

1. *Assessment skills* refer to the appraisal of patient characteristics, problems, values and expectations, and environmental factors.
2. *EBP process skills* are defined by competency in carrying out the steps of the EBP process: ask well-formulated questions, acquire best available research evidence, appraise quality and relevance of evidence, apply evidence, analyze change, and adjust treatment accordingly.

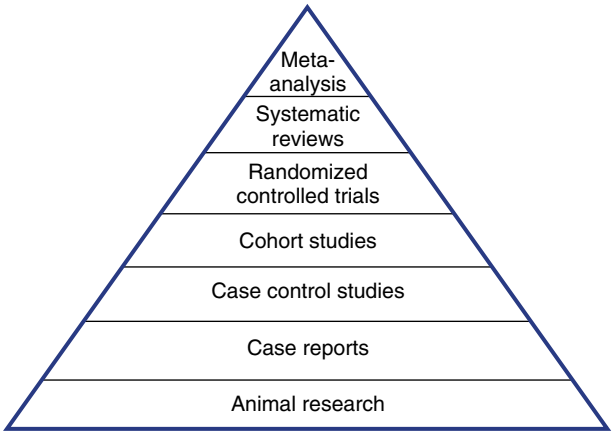

Figure 2 Evidence hierarchy.

3. *Communication and collaboration skills* involve the capacity to convey information clearly and appropriately. Further, they include the ability to listen, observe, adjust, and negotiate in order to achieve an agreed-upon treatment plan.
4. *Engagement and intervention skills* entail proficiency at motivating interest, constructive involvement, and positive change from individuals, groups, organizations, communities, and other entities affected by health decisions.

A recent development in EBP is resource-sensitive practice guidelines that review evidence for appropriate practice recommendations given an available level of resources (cf Fried and Krabshuis, 2008). Resource-sensitive guidelines enable decision-makers to appraise the level of intervention intensity that optimizes the available degree of accessible infrastructure, human capital, and financial resources.

Patient Characteristics

Patient characteristics are a key set of contextualizing factors. They include individual attributes such as state and trait variation in condition, needs, history of treatment response, values, and preferences that all influence whether a treatment is well matched to a particular patient. When deciding whether available research evidence is relevant to a given individual, health professionals need to assess the comparability between patient and study population. Tailoring surface aspects of the treatment can enhance its acceptability to the patient, so long as modifications do not stray so far from fidelity to core treatment elements that treatment loses its effectiveness (National Cancer Institute, 2006).

Patient preferences are a singular type of contextualizing variable. Although patient preferences are a particularly vital part of shared decision-making, they are the least developed aspect of the EBP model. Shared decision-making is grounded in the empowering of patients to self-manage their health and health care. Two preconditions for shared decision-making exist. The first is departure from a paternalistic model of care in which the clinician makes decisions on the patient's behalf. The second is adoption of a culturally informed model of care, whereby health professionals assist patients in clarifying their own values and treatment preferences.

Health Professionals' Roles in EBP

Health professionals can play several different roles in the EBP process (Figure 3).

First, they can be primary researchers who directly contribute to forming the evidence base. Primary researchers not only develop new treatments, but also design, conduct, analyze, and report research that evaluates the efficacy and effectiveness of interventions. Ideally, they will conduct RCTs to evaluate whether a treatment works. If, however, time and resources are insufficient to conduct an RCT, primary researchers may use alternative designs such as an intermittent time series.

A second role that health professionals can play is that of systematic reviewers, whereby they act as evidence synthesizers. They aggregate primary research to analyze and interpret synthesized findings that can be accessed and used efficiently

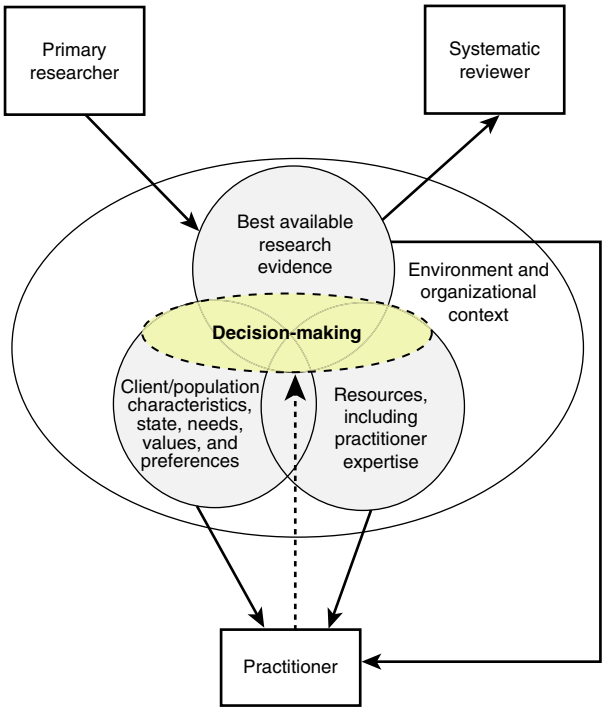

Figure 3 Roles of health professionals in the EBP model.

by health professionals. The role of the systematic reviewer is particularly critical within the EBP framework due both to the rapid proliferation of the scientific literature and health professionals' limited time to remain comprehensively informed of new research. The EBP system is made possible because systematic reviewers collect and analyze the full body of new and old studies that address clinically relevant questions. Systematic reviewers then disseminate their findings to health professionals in the form of succinct summaries that offer EBP recommendations for practicing clinicians.

Systematic review methodology includes a series of steps, including the formulation of a structured PICOT question that specifies the target population (P), candidate intervention (I) to be evaluated, comparator (C) intervention, patient outcome (O) of interest, and over what time frame (T) the outcome is to be assessed. After formulating the PICOT question, the systematic reviewer develops a comprehensive and unbiased protocol whose objective is to identify research that addresses the question. Once the relevant studies have been acquired, a decision about whether to include them in the review can be made relative to protocol entry and exclusion criteria. Included studies are then critically appraised for methodological quality, and their data extracted and synthesized to reach an answer to the question at hand. Synthesis is sometimes performed quantitatively, if the included interventions and study designs are sufficiently homogeneous; alternatively, synthesis can be performed qualitatively. Increasingly, systematic reviews constitute a requisite basis for EBP guidelines and health policies.

The third role that health professionals can play is that of the clinician. Assuming one of the most complex and challenging roles in EBP, the clinician extracts and uses data from each of the three EBP circles. Unlike the primary researcher or systematic reviewer, the clinician interacts directly with the two circles of the EBP model that concern patient characteristics and resource considerations. Additionally, health professionals are research consumers in that they access research evidence and assess its quality and relevance for the patient and context at hand. Secondary, synthesized, critically preappraised evidence sources, such as systematic reviews or EBP guidelines on www.guidelines.gov are intentionally designed to give the busy health professional a way to efficiently find the best research-tested answer to the most commonly asked practice questions. In some circumstances, however, available systematic reviews and treatment guidelines may not provide an answer, requiring the clinician to search the primary literature to identify relevant research. To enhance the simplicity of the clinician's complex job, the five-step EBP process delineates a recommended series of steps that health professionals can follow to address each of the three circles of the EBP model.

The Five Steps of EBP

Carrying out the EBP process involves five steps (Figure 4):

Step 1: Ask patient-oriented, well-formulated questions about the health status and contexts of individuals, communities, or populations.

Step 2: Acquire the best available evidence to answer the questions.

Step 3: Appraise the evidence critically for validity and applicability to the problem at hand.

Step 4: Apply the evidence by engaging in collaborative health decision-making with the affected individual(s) and/or group(s). Implement the health practice. Appropriate decision-making integrates the context, values, and preferences of the individual, community, or population. It also integrates available resources, including professional expertise.

Step 5: Analyze the new health practice and adjust practice accordingly. Evaluate implications for future decision-making, disseminate the results, and identify new informational needs.

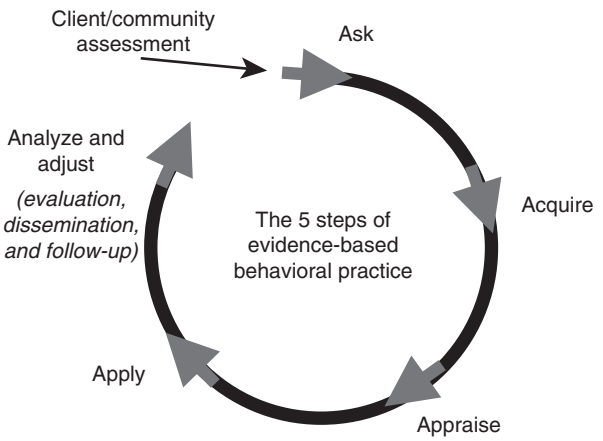

Figure 4 The 5As approach of EBP.

Competencies

Ask

EBP health professionals ask important, practice-relevant questions. They know how to translate their information needs into well-formulated, answerable questions. Further, they differentiate among various types of practical questions, including assessment, intervention, prognosis, harm, cost-effectiveness, and seek the best type(s) of evidence to answer each kind of question.

Acquire

EBP health professionals answer their questions by efficiently and effectively searching for the best available evidence. Specifically, EBP health professionals understand how to seek answers to their questions by accessing clinical guidelines and systematic reviews of research on health procedures. They know the difference between primary and secondary (synthesized) research evidence, and can translate questions into efficient search plans. They can use available technology and information systems to stay up-to-date regarding research relevant to their questions.

Appraise

EBP health professionals critically appraise evidence based on its quality and applicability to the specific population and circumstances at hand. When evaluating research on interventions, it is important to consider both internal and external validity. Internal validity represents the extent to which research was designed and conducted in a way that enables change to be causally attributed to the intervention as opposed to extraneous variables. External validity reflects whether characteristics of the research population or intervention context can be generalized to the current population, interventionist, or circumstances. Applicability refers to the clinician's judgment vis-à-vis the fit of the evidence with the circumstances.

In terms of specific competencies, EBP health professionals know the strengths and weaknesses of different kinds of research evidence for answering different kinds of health questions. They can evaluate the quality and strength of primary research evidence based on study design and execution. They understand how to synthesize research evidence and how to evaluate the quality and strength of evidence in systematic reviews and practice guidelines. They can identify gaps in evidence that suggest future research. Finally, they evaluate the applicability of the evidence for a particular individual, community, or population.

Apply

Finding the best available research evidence is one thing. Applying it is more complex in EBP, because it requires shared decision-making between health professionals and those affected by an intervention. The aim of the shared decision-making is to arrive at an action plan that balances the applicable evidence, the resources available to implement the best practice, and how the values and preferences of those affected influence the acceptability of the practice.

Analyze and Adjust

EBP health professionals participate in continuous quality improvement. After initiating an intervention that aggregate research suggests is evidence based, the clinician analyzes change and adjusts intervention accordingly. As such, the adaptation of interventions to changing individuals in changing contexts over time is at the core of EBP's transition between the Apply and the Analyze and Adjust. In other words, one initially applies the best 'one size fits all' treatment based on systematic review or aggregate evidence. Subsequently, the adaptation of treatment becomes individualized based on the person's own unique response to the sequence of offered treatments.

Methodological and Practical Challenges

Despite the past decades' great strides toward acceptance and full implementation of EBP, continued *barriers* remain. Rigorously designed research studies are expensive and take time to complete. Often, research findings are simply lacking or insufficient to provide a basis for policy and practice decisions. Gaps in the existing evidence base are especially noteworthy in areas of nondrug treatment interventions and preventive care (Maciosek et al., 2006; Moyer et al., 2005). The U.S. Preventive Services Task Force (USPSTF) was created by the U.S. Department of Health and Human Services in 1984 to address this challenge. *The Guide to Clinical Preventive Services*, which appraises systematic reviews of research graded on its evidence quality was created to allow governing bodies to identify effective evidence-based preventive services (Woolf and Atkins, 2001). The evidence reviewed in the current *Guide* provides the basis for all preventive services mandated by the Affordable Care Act to be covered by insurance without copays. Even so, the evidence about many preventive care practices earns a grade of 'i' for insufficient evidence. As such, the state of the science too often fails to accommodate the demands of policy-makers and health professionals.

Some have gone so far as to challenge the utility and relevance of RCTs as the gold standard research design for evaluating intervention effectiveness. Beyond the expense and duration of RCTs, other criticisms allege an overemphasis on internal validity (freedom from bias) over external validity (sample representativeness, generalizability) (Altman et al., 2001). Other challenges are that RCT designs are sometimes not feasible to implement when, for example, policy-makers or communities decline to accept potential random assignment to a control condition.

Another ongoing tension is that some clinicians chafe under the perceived restrictions that EBP imposes on professional autonomy. While they may appreciate research-tested practices as useful tools, they regard treatment as something of an art form and prefer being given the creative license to try out and develop novel treatment tactics. Additionally, health professionals may lack training in EBP. Although many training programs train students in the use of specific evidence-based treatments, few educate students on the EBP process and steps (Gambrill, 2007a). Discussions are just beginning regarding the best timing and configuration for EBP in the training curriculum (Jenson, 2007) and optimal teaching techniques (Gambrill, 2007b; Sackett et al., 2000).

Emerging Directions

The methods and standards of EBP are now given in health care. Establishing and disseminating an evidence base for behavioral treatments alongside medical treatments is necessary to make them accessible to patients. Providing access is, in turn, essential to achieve patient-centered care because a majority of patients express a preference for behavioral over medical treatments (e.g., Raue and Schulberg, 2007). Systematic reviews and treatment guidelines that include behavioral treatments are needed, particularly because most health insurers now require these EBP components as prerequisites for treatments to become covered practices.

Currently, most evidence-based behavioral treatments resemble evidence-based medical treatments in taking the form of a single, best 'one size fits all' intervention. However, few clinicians believe that one-size treatment fits all. More often clinicians practice by applying a sequence of interventions that they adapt over time depending upon the patient's response. A common protocol is to begin with a best modal practice, evaluate response, and then adapt or change treatment in the case of insufficient response or nonresponse. The sequence of decision rules about how to titrate and adjust treatment contingent upon patient response has been called a treatment algorithm or operational guideline. The algorithm represents a strategy that links together particular treatments that have research support, a strategy about what treatment should be tried first, and a set of tactics or decision rules about what treatment to try next in the case of treatment failure. Like a specific treatment, an algorithm that links together a sequence of research-tested interventions can itself be tested and proven by research to be an efficacious strategy, as compared to a fixed treatment or an alternative treatment algorithm. An example of adaptive treatments that showed efficacy in clinical trials is the Action to Control Cardiovascular Risk in Diabetes trial (e.g., Sullivan et al., 2013). Moreover, once validated, such operational treatment guidelines can be programmed into electronic health records or mobile devices to provide guidance and decision support to clinicians about how to choose the next step in treatment. Approaches like the operational guideline capture both the Apply and the Analyze and Adjust steps of the EBP process. A new generation of research designs, such as the Sequential MultiPhased Adaptive Randomized Trial (SMART) has been put forward to develop such treatment algorithms (e.g., Almirall et al., 2012). We expect these new adaptive interventions to become the EBP wave of the future.

See also: Applied Social Research, History of; Behavioral Economics, History of; Health Education and Health Promotion; Implementation Science; Policy Analysis; Prevention Research; Science and Politics: Value Neutrality.

Bibliography

Almirall, D., Compton, S.N., Gunlicks-Stoessel, M., Duan, N., Murphy, S.A., 2012. Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. Statistics in Medicine 31 (17), 1887–1902.

Altman, D.G., Shulz, K.F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., et al., 2001. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Annals of Internal Medicine 134, 663–694.

Beck, A.H., 2004. The Flexner report and the standardization of American medical education. Journal of the American Medical Association 291 (17), 2139–2140.

Cochrane, A., 1972. Effectiveness and Efficiency: Random Reflections on Health Services. Royal Society of Medicine Press, London.

Dallery, J., Cassidy, R.N., Raiff, B.R., 2013. Single-case experimental designs to evaluate novel technology-based health interventions. Journal of Medical Internet Research 15 (2), e22. http://www.jmir.org/2013/2/e22/.

Davidson, K.W., Goldstein, M., Kaplan, R.M., Kaufmann, P.G., Knatterud, G.L., Orleans, C.T., et al., 2003. Evidence-based behavioral medicine: what it is and how do we achieve it? Annals of Behavioral Medicine 26 (3), 161–171.

Davidson, K., Trudeau, K., 2004. Evidence-based behavioral medicine. In: Anderson, N. (Ed.), Encyclopedia of Health and Behavior. SAGE Publications, Inc., Thousand Oaks, CA, pp. 320–324. http://dx.doi.org/10.4135/9781412952576.n94.

Flexner, A., 1910. Medical Education in the United States and Canada: A Report to the Carnegie Foundation for the Advancement of Teaching (Bulletin No. 4). The Carnegie Foundation for the Advancement of Teaching, New York City, pp. 346, OCLC 9795002.

Fried, M., Krabshuis, J., 2008. Can "cascades" make guidelines global? Journal of Evaluation in Clinical Practice 14 (5), 874–879.

Gambrill, E., 2007a. To be or not to be: will five-step be used by clinicians? Research on social work practice 17, 428–434. A Review of Norcross, J.C., Beutler, L.E., Levant, R.F. (Eds.), Evidenced-based practices in mental health: Debate and dialogue on the fundamental questions. American Psychological Association, Washington, DC.

Gambrill, E., 2007b. Transparency as the route to evidenced-informed professional education. Research on Social Work Practice 17, 553–560.

Greiner, A.C., Knebel, E. (Eds.), 2003. Health Professions Education: A Bridge to Quality. National Academy Press, Washington, DC.

Guyatt, G., Cairns, J., Churchhill, D., Cook, D., Haynes, B., Hirsch, J., et al., 1992. Evidence-based medicine. A new approach to teaching the practice of medicine. Journal of the American Medical Association 268, 2420–2425 ('Evidence-based Medicine Working Group').

Haynes, R.B., Sackett, D.L., Gray, J.M., Cook, D.J., Guyatt, G.H., 1996. Transferring evidence from research into practice: 1. The role of clinical care research evidence in clinical decisions. ACP Journal Club 125 (3), A14–A16.

Isaacs, D., Fitzgerald, D., 1999. Seven alternatives to evidence based medicine. British Medical Journal 319 (7225), 1618.

Jenson, J.M., 2007. Evidence-based practice and the reform of social work education: a response to Gambrill and to Howard and Allen-Meares. Research on Social Work Practice 17, 569–573.

Maciosek, M.V., Coffield, A.B., Edwards, N.M., Flottemesch, T.J., Goodman, M.J., Solberg, L.I., 2006. Priorities among effective clinical preventive services: results of a systematic review and analysis. American Journal of Preventive Medicine 31 (1), 52–61.

Moyer, V.A., Klein, J.D., Ockene, J.K., Teutsch, S.M., Johnson, M.S., Allan, J.D., 2005. Childhood obesity working group, US preventive services task force. Screening for overweight in children and adolescents: where is the evidence? A commentary by the childhood obesity working group of the US preventive services task force. Pediatrics 116 (1), 235–238.

National Cancer Institute, 2006. Using What Works: Adapting Evidence-Based Programs to Fit Your Needs, Module 3, NIH Publication No. 06-5874. Retrieved from: http://cancercontrol.cancer.gov/use_what_works/start.htm.

Pagoto, S., Kozak, A.T., John, P., Bodenlos, J., Hedeker, D., Spring, B., 2009. Intention-to-treat analyses in behavioral medicine randomized clinical trials: the impact of CONSORT. International Journal of Behavioral Medicine 16 (4), 316–322.

Raue, P., Schulberg, H.C., 2007. Psychotherapy and Patient Preferences for the Treatment of Major Depression in Primary Care. Trends in Depression Research. Hauppauge, New York.

Sackett, D.L., Straus, S.E., Richardson, W.S., Rosenberg, W., Haynes, R.B., 2000. Evidence-based Medicine: How to Practice and Teach EBM, second ed. Churchill Livingstone, New York.

Sackett, D.L., Rosenberg, W.M., 1995a. On the need for evidence-based medicine. Health Economics 4, 249–254.

Sackett, D.L., Rosenberg, W.M., 1995b. The need for evidence-based medicine. Journal of the Royal Society of Medicine 88, 620–624.

Sackett, D.L., Rosenberg, W.M., Gray, J.A., Haynes, R.B., Richardson, W., 1996. Evidence-based medicine: what it is and what it isn't. British Medical Journal 312, 71–72.

Sackett, D.L., Wennberg, J.E., 1997. Choosing the best research design for each question. British Medical Journal 315, 1636.

Satterfield, J.M., Spring, B., Brownson, R.C., Mullen, E.J., Newhouse, R.P., Walker, B.B., 2009. Toward a transdisciplinary model of evidence-based practice. The Milkbank Quarterly 87 (2), 368–390.

Schulz, K.F., Altman, D.G., Moher, D., 2010. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Annals of Internal Medicine 154 (11), 1–11.

Spring, B., Hitchcock, K., 2009. Evidence-based practice in psychology. In: Weinter, I.B., Craighead, W.E. (Eds.), Corsini's Encyclopedia of Psychology, fourth ed. Wiley, New York, pp. 603–607.

Spring, B., Nelville, K., 2010. Evidence-based practice in clinical psychology. In: Barlow, D. (Ed.), Oxford Handbook of Clinical Psychology. Oxford University Press, New York, pp. 128–149.

Spring, B., Pagoto, S., Knatterud, G., Kozak, A., Hedeker, D., 2007. Examination of the analytic quality of behavioral health randomized clinical trials. Journal of Clinical Psychology 63, 53–71.

Strauss, S.E., Richardson, W.S., Glasziou, P., Haynes, R.B., 2005. Evidence-Based Medicine: How to Practice and Teach EBM, third ed. Elsvier, New York.

Sullivan, M.D., Katon, W.J., Lovato, M.E., et al., 2013. Association of depression with accelerated cognitive decline among patients with type 2 diabetes in the ACCORD-MIND trial. JAMA Psychiatry 70 (10), 1041–1047.

Weaver, C.A., Warren, J.J., Delaney, C., et al., 2005. Bedside, classroom and bench: collaborative strategies to generate evidence-based knowledge for nursing practice. International Journal of Medical Informatics 74 (11–12), 989–999.

Woolf, S.H., Atkins, D., 2001. The evolving role of prevention in health care contributions of the U.S. Preventive Services Task Force. American Journal of Preventive Medicine 20 (3S), 13–20. http://dx.doi.org/10.1016/S0749-3797%2801%2900262-8.

Westfall, J.M., Mold, J., Fagan, L., 2007. Practice-based research-"Blue Highways" on the NIH roadmap. Journal of the American Medical Association 297 (4), 403–406.

Relevant Websites

www.ebbp.org – Evidence-Based Behavioral Practice (EBBP) project.
www.cochrane.org – The Cochrane Collaboration.
www.guidelines.gov – National Guideline Clearinghouse.

Evidence-Based Psychiatry

Megan J Hobbs, Clinical Research Unit for Anxiety and Depression, School of Psychiatry, University of New South Wales, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Andrews, volume 18, pp. 12287–12292, © 2001, Elsevier Ltd.

Abstract

Efficacious, effective, and efficient psychiatric treatments are needed to reduce the burden of mental disorders. Evidence-based psychiatry is the process that facilitates the translation of research from the bench to the bedside. The 'evidence' in evidence-based psychiatry is graded according to a hierarchy. Randomized controlled trial(s) are considered to be at the apex of this hierarchy. Most randomized controlled trials use categorical disorders as their primary outcome variable but categorical disorders have limited validity. Pragmatic approaches to improving the 'evidence' in evidence-based psychiatry by accounting for the nature of mental disorders are, therefore, needed and are described herein.

One in two adults will experience a mental disorder in their lives and most of these adults will experience more than one mental disorder (Teesson et al., 2009). As well as being common in the community, mental disorders are highly burdensome (Murray et al., 2013). With the purpose of averting this burden, some governments spend billions of dollars annually on mental health services (Australian Institute of Health and Welfare, 2012). The development and implementation of efficacious, effective, and efficient psychiatric treatments are therefore important public health issues.

Researchers are responsible for establishing treatment efficacy. This is achieved by developing and refining treatment protocols and evaluating treatment outcomes in relation to a null or an active comparator. Mental health practitioners are responsible for providing efficacious treatments to their patients and when mental health professionals use these treatments and achieve comparable outcomes to efficacious studies, these treatments are said to be effective. Governments have limited funds and it is, therefore, ideal that the most efficient treatments are funded, that is, treatments that produce the greatest health benefits for the least government dollars should be funded. This process of translating empirical research from the bench to the bedside is the primary goal of evidence-based psychiatry. This goal is achieved when mental health practitioners make "conscientious, explicit and judicious use of the current best evidence in making decisions about the care of individual patients" (Sackett et al., 1996).

The 'evidence' in evidence-based psychiatry is a graduated construct, the degree of which can be gauged by mental health professionals in relation to the following hierarchy (Geddes and Harrison, 1997; **Table 1**).

When evaluating the degree or quality of evidence in favor of a respective treatment, mental health professionals should prefer data from higher ranked methods over data from lower ranked methods. Data from lower ranked methods can, however, inform clinical decision making when data from higher order methods are unavailable (Geddes and Harrison, 1997). The 'evidence' in evidence-based psychiatry is therefore the sum of the quantitative data from empirical evaluations of respective treatments preferably from the synthesis of multiple randomized controlled trials (Möller, 2009). Psychiatry as a science has much to benefit from employing evidence-based principles. Evidence-based psychiatry holds promise of better health outcomes for patients, legitimizes the treatment protocols that mental health professionals choose, and that governments fund on the basis of quantifiable data that can be disproven using scientific method rather than personal opinion or political agendas.

Despite the potential benefits of evidence-based psychiatry for patients, mental health professionals, and governments, evidence-based psychiatry has been criticized on a number of grounds. Most critiques concern the treatment of randomized controlled trial(s) as the 'best' evidence for preferring particular psychiatric interventions. Experts have suggested, for instance, that it is difficult to generalize the results of randomized controlled trials to the complexities of psychiatric practice, that meta-analyses of multiple trials can bias results, that evidence-based psychiatry does not reflect clinical decision making accurately and that relying heavily on quantitative data minimizes the importance of professional expertise in the diagnosis and treatment of psychiatric disorders (Gupta, 2007; Maier, 2006).

The most evidence-based criticism of evidence-based psychiatry concerns the validity of psychiatric disorders and the use of these disorders as the primary outcome measures in randomized controlled trials. The international psychiatric classifications – the *Diagnostic and Statistical Manual for Mental Disorders-5* (DSM) and the *International Classification of Diseases-10* (ICD) – primarily use a categorical approach to defining

Table 1 Hierarchy of evidence in evidence-based psychiatry

Evidence from a meta-analysis of randomized controlled trials
Evidence from at least one randomized controlled trials
Evidence from at least one controlled study without randomization
Evidence from at least one other type of quasiexperimental study
Evidence from nonexperimental descriptive studies, such as comparative studies, correlation studies and case-control studies
Evidence from expert committee reports or opinions and/or clinical experience of respected authorities

Reproduced from Geddes, J.R., Harrison, P.J., 1997. Closing the gap between research and practice. British Journal of Psychiatry 171, 220–225.

mental disorders. Lists of symptoms and symptom thresholds (e.g., the number and/or combination of symptoms) are used to distinguish between categorically distinct cases of specific mental disorders (e.g., a patient has disorder X but not disorder Y) and to distinguish between cases and noncases of disorder (e.g., a patient has disorder X or they do not). As a result of this categorical approach to classification, randomized controlled trials tend to focus on whether the treatment under trial significantly reduces the number of patients who meet diagnostic criteria for a disorder. This approach to measuring clinical change assumes that there is a group of patients who each meet criteria for the same disorder that these patients experience the disorder in the same way and will therefore respond to treatment in a uniform manner. There are robust data that show that these assumptions are incorrect. The DSM and ICD disorder classifications do not identify mutually exclusive groups of patients. Many of the specifically defined mental disorders share common risk factors and clinical correlates. Moreover, mental disorders are not experienced uniformly. People experience mental disorders to different degrees and in different ways. The inaccurate assumptions that form the basis of most randomized controlled trials therefore threaten the validity of relying on randomized controlled trial data pursuant to the evidence-based psychiatry principles.

Although the validity of psychiatric disorders is a real methodological limitation of most randomized controlled trials and consequently, a limitation of evidence-based psychiatry, the alternative to evidence-based psychiatry is clinical decision making and a mental health system that are uninformed by empirical data. The latter has a significant likelihood of resulting in poor treatment outcomes for patients, bringing the field of psychiatry into disrepute and a mental health system funded based solely on political agendas. This alternative to evidence-based psychiatry is unacceptable. Pragmatic approaches for improving the design of randomized controlled trials in light of the nature of mental disorders are therefore needed. Herein the risk factors and clinical correlates shared by many of the specifically defined mental disorders and the different ways that people experience mental disorders are reviewed. Following on from this discussion, practical ways of accounting for the nature of mental disorders in the design of randomized controlled trials are detailed. These methods are presented as tangible ways of improving the 'evidence' in evidence-based psychiatry.

Many Mental Disorders Share Risk Factors and Clinical Correlates

The notion that the international psychiatric classifications carve nature at its joints and identify discrete disease entities that respond differentially to specific treatments is false (Kendell and Jablensky, 2003). The common mental disorders (e.g., anxiety, depression, conduct, and substance-related disorders) cooccur more often than would be expected by chance (Teesson et al., 2009). This pattern of cooccurrence suggests that these disorders are different manifestations of common underlying liabilities. Indeed, internalizing and externalizing liabilities explain the development and maintenance (Kessler et al., 2011), and the phenotypic and genotypic covariation of the anxiety and unipolar mood disorders, and the substance-related and antisocial disorders, respectively (Andrews et al., 2009; Goldberg et al., 2009; Kendler et al., 2003; Krueger and Markon, 2006). This internalizing–externalizing structure of common mental illness has been replicated across age groups (Fergusson et al., 2011; Lahey et al., 2008; Laurens et al., 2012), ethnicities (Eaton et al., 2012), and countries (Krueger et al., 2003; Slade and Watson, 2006). The internalizing–externalizing model of the occurrence of the common mental disorders has also been extended to explain the occurrence of psychotic-like experiences with a third distinct but correlated factor explaining the cooccurrence of psychotic-like experiences in adults (Wright et al., 2013) and in children (Laurens et al., 2012).

People Experience Mental Illness to Different Degrees

Sadness and anxiety are normal reactions to situational triggers. Although everyone experiences sadness and anxiety, not everyone will develop an affective or an anxiety disorder during their lives (Kessler et al., 2005). Similarly, all antisocial adults have life-long histories of antisocial behavior but not all adolescents who have conduct problems grow into antisocial adults; most do not (Moffitt, 1993). Indeed, more than half of children report that they experience psychotic-like experiences but only a small proportion of adults will develop a psychotic-like disorder (Laurens et al., 2007; Morgan et al., 2011). As well as there being groups of people who do and do not develop threshold mental disorder(s) during their lives, people who do experience threshold mental disorders tend to oscillate between experiencing subthreshold symptoms and their threshold disorder (e.g., Bienvenu et al., 1998). These data suggest that people experience mental illness to varying degrees and vary in their propensity to develop active mental illness.

Structural research supports the conclusion that people experience mental health problems to different degrees. Item response modeling is a latent variable technique that has provided unique insights into the dimensionality of specific mental disorders as well as the relationship between the symptoms that are used to define the disorder and the severity of the disorder. Item response modeling achieves this by describing the relationship between the probability of reporting psychiatric symptoms, diagnostic severity, and the psychometric characteristics of symptoms (Reckase, 2009). Specifically, this probabilistic relationship is described using a 'person parameter' (e.g., theta) that indexes the severity of the respective person being asked about their psychiatric experiences. In addition to the person parameter, 'item parameters' are used to describe the relationship between diagnostic latent severity, symptom characteristics, and the probability of endorsing the symptom. These are a *discrimination parameter* (i.e., an *a*-parameter) that reflects the ability of the symptom to distinguish between continuous degrees of the underlying disorder factor and a *difficulty parameter* (i.e., *b*-parameter) that describes the degree of severity that equates to a 50% likelihood of endorsing the symptom. (Note that the number of *b*-parameters used in item response modeling is equal to the number of categories used to measure the presence of the symptom minus one. The meaning of the difficulties parameters will vary as a function of the number of response

categories.) A key example of how item-response modeling has been used to refine clinical phenotypes comes from the substance-use literature. In DSM-IV, substance-related disorders were classified using separate abuse and dependence categories. The validity of continuing to delineate between substance abuse and dependence in DSM-5 was tenuous. Contrary to the biaxial DSM-IV classification, replicated data from clinical and epidemiologic samples showed that substance abuse and dependence reflected one latent construct. Item-response modeling studies that examined how the DSM-IV abuse and dependence criteria indexed the latent construct of substance misuse also showed that the abuse criteria did not always index milder substance misuse than the dependence criteria (Langenbucher et al., 2004; Lynskey and Agrawal, 2007; Teesson et al., 2002). As a result of these data, the DSM-IV abuse and dependence criteria were coalesced into single diagnostic categories for each respective substance type in DSM-5.

People Experience Mental Disorders in Different Ways

In contrast to the homogeneous assumptions of most randomized controlled trials, epidemiological and clinical data show that people who experience mental disorders do so in heterogeneous ways. This means that two patients who meet the diagnostic criteria for the same disorder can experience the disorder in very different ways. *Latent class analyses* are one of the latent modeling techniques that have provided insights into the different ways that individuals experience mental health problems. Latent class analyses are 'person-centered' analyses that identify discrete groups of people based on the endorsement of observed variables such as psychiatric symptoms (Lazarsfeld and Henry, 1968). Latent class analyses do not make a priori assumptions about the number of different ways that people experience psychiatric symptoms or how specific people experience psychiatric symptoms. Instead, latent class analyses identify the number of groups and the group membership of people based on how and which symptoms they report (Lazarsfeld and Henry, 1968). Researchers have used latent class analyses to examine, for instance, the different ways that individuals experience depression. These studies have shown that individuals can experience 'severe depression' and tend to experience all of the DSM-IV depressive symptoms. Individuals can also experience 'psychosomatic depression' (e.g., changes in appetite, sleep disturbances, psychosomatic complaints, and impaired concentration) or 'cognitive-emotional' depression, which is characterized by feelings of worthlessness and excessive guilt, difficulty concentrating, and suicidal thoughts (Carragher et al., 2009; Sullivan et al., 1998). Indeed, there are clinically meaningful subtypes of other mental disorders such as panic disorder (Goodwin et al., 2002; Kircanski et al., 2009; Roberson-Nay and Kendler, 2011) and social phobia (Kessler et al., 1998).

Improving the 'Evidence' in Evidence-Based Psychiatry

Although multiple disorders share risk factors and clinical correlates and people experience specific mental disorders to varying degrees and in different ways, evidence-based psychiatry should not be disbanded by mental health professionals or the governments that fund health services. Instead, there are quantitative methods that are available that could be employed in randomized controlled trials in ways that would account for the nature of mental disorders. These methods, if employed in randomized controlled trials, would mitigate the criticism of evidence-based psychiatry that psychiatric disorders have limited validity and by association so do randomized controlled trials that use disorders as their primary outcome measure.

Generalized mixture models provide one way of improving the 'evidence' in evidence-based medicine. Generalized mixture models can be used to examine the trajectories of change of single or multiple observable variables such as psychiatric symptoms and can map the trajectories of continuous and categorical observable variables (Muthén et al., 2002). This means that generalized mixture models avoid the homogeneous designs that are typically employed in randomized controlled trials and can be used to examine how patients' presenting symptoms influence their treatment response by providing:

- a way to identify groups of patients whose symptom profiles differ and
- a way to examine whether these discrete groups differ in their trajectory and/or rate of treatment response.

For example, generalized mixture models could account for the nature of mental disorders by examining how the comorbid symptoms of multiple disorders influence treatment response and thereby account for the fungible boundaries between categorical diagnoses. Moreover, generalized mixture models could be used to examine how individual differences in psychiatric experiences influence treatment response. Randomized controlled trials that employ generalized mixture models would therefore avoid the homogeneous assumptions that have been argued to be a limitation of evidence-based psychiatry.

A limited number of randomized controlled trials have used generalized mixture modeling to examine the relationship between clinical characteristics and treatment response. For example, generalized mixture models of depression have shown that some depressed patients improve with treatment, some patients get more depressed with treatment, and for some patients treatment has no effect (Cuijpers et al., 2005; Gunn et al., 2013; Sunderland et al., 2012). Patients with different degrees of depression also experience change at different times during treatment (Lutz et al., 2009) and respond to different treatments, for instance, combination therapy is more effective than monotherapies for some depressed patients (Stulz et al., 2010). Depressed patients who respond differently to treatment also differ in terms of their sociodemographics and clinical correlates including age of onset, comorbidity, and familiality of psychopathology (Cuijpers et al., 2005; Gunn et al., 2013). There has been sparse use of generalized mixture models in anxiety and substance-use trials. Although generalized mixture models are only now making traction in the clinical treatment literature, the aforementioned studies demonstrate that the limited validity of categorical psychiatric disorders does not need to be a limitation of evidence-based psychiatry.

Conclusions

The process of evidence-based psychiatry can benefit patients, mental health practitioners, and the governments that fund mental health services by promoting the use of efficacious, effective and/or efficient treatments. Like any method, evidenced-based psychiatry has its limitations. Randomized controlled trials fall at the apex of the evidence hierarchy. Yet despite the limited validity of categorical psychiatric disorders, randomized controlled trials use these disorders as their primary outcome variable. This practice does not reflect the nature of mental disorders accurately. Pragmatic approaches such as item response, latent class, and generalized mixture modeling are gaining traction in the epidemiological and clinical literature. These methods can be used both to identify epidemiologically meaningful input variables for randomized clinical trials as well as identifying how individual differences in mental health problems influence trajectories of clinical change. Incorporating these latent variable methods in future randomized controlled trials would improve the 'evidence' in evidence-based psychiatry by accounting for the nature of mental illness.

See also: Alcohol-Related Disorders; Antidepressant Drugs; Anxiety and Anxiety Disorders; Depression; Diagnostic Processes in Clinical Psychology and Psychotherapy; Epidemiology, Concepts and Methods of; Evidence-Based Psychotherapy; Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis; Factor Analysis and Latent Variable Models in Personality Psychology; Mental Disorders, Epidemiology of; Neuroticism; Nosology in Psychiatry; Panic Disorder; Personality Disorders; Personality Theory and Psychopathology; Personality, Trait Models of; Psychiatric Care, Economics of; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychological Treatments: Randomized Controlled Trials.

Bibliography

Andrews, G., Goldberg, D.P., Krueger, R.F., Carpenter, W.T., Hyman, S.E., Sachdev, P.S., Pine, D.S., 2009. Exploring the feasibility of a meta-structure for DSM-V and ICD-11: could it improve utility and validity? Psychological Medicine 39, 1993–2000.

Australian Institute of Health and Welfare, 2012. Mental Health Services – in Brief 2012. Australian Institute of Health and Welfare, Canberra.

Bienvenu, O.J., Nestadt, G., Eaton, W.W., 1998. Characterizing generalized anxiety: temporal and symptomatic thresholds. Journal of Nervous Mental Disorders 186, 51–56.

Carragher, N., Adamson, G., Bunting, B., McCann, S., 2009. Subtypes of depression in a nationally representative sample. Journal of Affective Disorders 113, 88–99.

Cuijpers, P., van Lier, P.A.C., van Straten, A., Donker, M., 2005. Examining differential effects of psychological treatment of depressive disorder: an application of trajectory analyses. Journal of Affective Disorders 89, 137–146.

Eaton, N.R., Keyes, K.M., Krueger, R.F., Noordhof, A., Skodol, A.E., Markon, K.E., Grant, B.F., Hasin, D.S., 2012. Ethnicity and psychiatric comorbidity in a national sample: evidence for latent comorbidity factor invariance and connections with disorder prevalence. Social Psychiatry and Psychiatric Epidemiology 48, 701–710.

Fergusson, D.M., Boden, J.M., Horwood, L.J., 2011. Structural models of the comorbidity of internalizing disorders and substance use disorders in a longitudinal birth cohort. Social Psychiatry and Psychiatric Epidemiology 46, 933–942.

Geddes, J.R., Harrison, P.J., 1997. Closing the gap between research and practice. British Journal of Psychiatry 171, 220–225.

Goldberg, D.P., Andrews, G., Krueger, R., Hobbs, M.J., 2009. Emotional disorders: cluster 4 of the proposed meta-structure for DSM-V. Psychological Medicine 39, 2043–2059.

Goodwin, R.D., Hamilton, S.P., Milne, B.J., Pine, D.S., 2002. Generalizability and correlates of clinically derived panic subtypes in the population. Depression and Anxiety 15, 69–74.

Gunn, J., Elliott, P., Densley, K., Middleton, A., Ambresin, G., Dowrick, C., Herrman, H., Hegarty, K., Gilchrist, G., Griffiths, F., 2013. A trajectory-based approach to understand the factors associated with persistent depressive symptoms in primary care. Journal of Affective Disorders 148, 338–346.

Gupta, M., 2007. Does evidence-based medicine apply to psychiatry? Theoretical Medicine and Bioethics 28, 103–120.

Kendell, R.E., Jablensky, A., 2003. Distinguishing the validity and utility of psychiatric diagnoses. American Journal of Psychiatry 160, 4–12.

Kendler, K.S., Prescott, C.A., Myers, J.M., Neale, M.C., 2003. The structure of genetic and environmental risk factors for common psychiatric and substance use disorders in men and women. Archives of General Psychiatry 60, 929–937.

Kessler, R.C., Berglund, P., Demler, O.V., Jin, R., Merikangas, K.R., Walters, E.E., 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity replication. Archives of General Psychiatry 62, 593–602.

Kessler, R.C., Cox, A.D., Green, J.G., Ormel, J., McLaughlin, K.A., Merikangas, K.R., Petukhova, M., Pine, D.S., Russo, L.J., Swendsen, J., Wittchen, H.-U., Zaslavsky, A.M., 2011. The effects of latent variables in the development of comorbidity among common mental disorders. Depression and Anxiety 28, 29–39.

Kessler, R.C., Stein, M.B., Berglund, P., 1998. Social phobia subtypes in the National Comorbidity Survey. American Journal of Psychiatry 155, 613–619.

Kircanski, K., Craske, M.G., Epstein, A.M., Wittchen, H.-U., 2009. Subtypes of panic attacks: a critical review of the empirical literature. Depression and Anxiety 26, 878–887.

Krueger, R.F., Chentsova-Dutton, Y.E., Markon, K.E., Goldberg, D.P., Ormel, J., 2003. A cross-cultural study of the structure of comorbidity among common psychopathological syndromes in the general health care setting. Journal of Abnormal Psychology 107, 216–227.

Krueger, R.F., Markon, K.E., 2006. Reinterpreting comorbidity: a model-based approach to understanding and classifying psychopathology. Annual Review of Clinical Psychology 2, 111–133.

Lahey, B.B., Rathouz, P.J., van Hulle, C., Urbano, R.C., Krueger, R.F., Applegate, B., Garriock, H.A., Chapman, D.A., Waldman, I.D., 2008. Testing structural models of DSM-IV symptoms of common forms of child and adolescent psychopathology. Journal of Abnormal Child Psychology 36, 187–206.

Langenbucher, J.W., Labouvie, E., Martin, C.S., Sanjuan, P.M., Bavly, L., Kirisci, L., 2004. An application of item response theory analysis to alcohol, cannabis, and cocaine criteria in DSM-IV. Journal of Abnormal Psychology 113, 72–80.

Laurens, K.R., Hobbs, M.J., Sunderland, M., Green, M.J., Mould, G.L., 2012. Psychotic-like experiences in a community sample of 8,000 children aged 9 to 11 years: an item response theory analysis. Psychological Medicine 42, 1495–1506.

Laurens, K.R., Hodgins, S., Maughan, B., Murray, R.M., Rutter, M.L., Taylor, E.A., 2007. Community screening for psychotic-like experiences and other putative antecedents of schizophrenia in children aged 9-12 years. Schizophrenia Research 90, 130–146.

Lazarsfeld, P.F., Henry, N.W., 1968. Latent Structure Analyses. Plenum Press, New York.

Lutz, W., Stulz, N., Kock, K., 2009. Patterns of early change and their relationship to outcome and follow-up among patients with major depressive disorders. Journal of Affective Disorders 118, 60–68.

Lynskey, M.T., Agrawal, A., 2007. Psychometric properties of DSM assessment of illicit drug abuse and dependence: results from the national epidemiologic survey on alcohol and related conditions (NESARC). Psychological Medicine 37, 1345–1355.

Maier, T., 2006. Evidence-based psychiatry: understanding the limitations of a method. Journal of Evaluation in Clinical Practice 12, 325–329.

Moffitt, T.E., 1993. Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy. Psychological Review 100, 674–701.

Möller, H.-J., 2009. Is evidence sufficient for evidence-based medicine? European Archives of Psychiatry and Clinical Neuroscience 259, S167–S172.

Morgan, V.A., Waterreus, A., Jablensky, A., Mackinnon, A., McGrath, J.J., Carr, V., Bush, R., Castle, D., Cohen, M., Harvey, C., Galletly, C., Stain, H.J., Neil, A., McGorry, P., Hocking, B., Shah, S., Saw, S., 2011. People Living with Psychotic Illness 2010. Commonwealth Government of Australia, Canberra, Australia.

Murray, C.J.L., Vos, T., Lozano, R., Naghavi, M., Flaxman, A.D., Michaud, C., Ezzati, M., Shibuya, K., Salomon, J.A., Abdalla, S., Aboyans, V., Abraham, J., et al., 2013. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the global Burden of disease study 2010. Lancet 380, 2197–2223.

Muthén, B.O., Hendricks Brown, C., Masyn, K., HKhoo, S.-T., Yang, C.-C., Wang, C.-P., Kellam, S.G., Carlin, J.B., Liao, J., 2002. General growth mixture modeling for randomized preventative interventions. Biostatistics 3, 459–475.

Reckase, M.D., 2009. Multidimensional Item Response Theory. Springer, New York.

Roberson-Nay, R., Kendler, K.S., 2011. Panic disorder and its subtypes: a comprehensive analysis of panic symptom heterogeneity using epidemiological and treatment seeking sampled. Psychological Medicine 41, 2411–2421.

Sackett, D.L., Rosenberg, W.M.C., Gray, J.A.M., Hayes, R.B., Richardson, W.S., 1996. Evidence based medicine: what it is and what it isn't. British Medical Journal 312, 71–72.

Slade, T., Watson, D., 2006. The structure of common DSM-IV and ICD-10 mental disorder in the Australian general population. Psychological Medicine 36, 1593–1600.

Stulz, N., Thase, M.E., Klein, D.N., Manber, R., Crits-Christoph, P., 2010. Differential effects of treatments for chronic depression: a latent growth model reanalysis. Journal of Consulting and Clinical Psychology 78, 409–419.

Sullivan, P.F., Kessler, R.C., Kendler, K.S., 1998. Latent class analysis of lifetime depressive symptoms in the national comorbidity survey. American Journal of Psychiatry 155, 1398–1406.

Sunderland, M., Wong, N., Hilvert-Bruce, Z., Andrews, G., 2012. Investigating trajectories of change in psychological distress amongst patients with depression and generalised anxiety disorder treated with internet behavioural therapy. Behaviour Research and Therapy 50, 374–380.

Teesson, M., Lynskey, M.T., Manor, B., Baillie, A., 2002. The structure of DSM-IV cannabis use disorders in the community. Drug and Alcohol Dependence 68, 255–262.

Teesson, M., Slade, T., Mills, K., 2009. Comorbidity in Australia: findings of the 2007 national survey of mental health and wellbeing. Australian and New Zealand Journal of Psychiatry 43, 606–614.

Wright, A., Krueger, R.F., Hobbs, M.J., Markon, K.E., Eaton, N.R., Slade, T., 2013. The structure of psychopathology: toward an expanded quantitative empirical model. Journal of Abnormal Psychology 122, 281–294.

Evidence-Based Psychotherapy

Andreas Beelmann, Friedrich-Schiller-University Jena, Jena, Germany
Nina Heinrichs, University of Braunschweig, Braunschweig, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents the conceptual and methodological foundations of evidence-based psychotherapies (EBPs). Starting with a definition and guidelines for EBP, it gives a short overview of the historical development of psychotherapy research and EBP. It describes different types of psychotherapy research, such as efficacy, effectiveness, efficiency, and process research, and discusses several methodological issues concerning research design, selection of samples, and outcome criteria, as well as the interpretation of research results. The article closes with a short overview of EBP within the field of treatment of depression and anxiety.

Introduction: What Is Evidence-Based Psychotherapy?

The term evidence-based psychotherapy (EBP) refers to empirically tested psychotherapeutic treatments and strategies. Hence, EBP requires above all empirical data and knowledge on *outcomes* based on informative and meaningful studies of clinically representative samples. However, recent psychotherapy research also includes more complex studies such as analyses of treatment aptitude (e.g., which treatment is most effective for whom and under what circumstances), specifications of processes of change (how and why clients improve and what are the most important predictors of positive outcomes), and calculations of the efficiency of psychotherapy (cost–benefit ratios).

In recent years, the movement toward EBP can be traced back to the initiatives of a Task Force set up by the Clinical Section of the American Psychological Association (see Chambless and Hollon, 1998) that developed specific guidelines for identifying empirically supported therapies (EST). According to these guidelines, a therapy can be assumed to be empirically based if it meets a set of methodological and substantial criteria. For example, well-established treatments require (1) at least two between-group design studies (i.e., at least one treatment and one control group that are compared before and after the intervention) conducted by two independent research groups; (2) studies using a treatment manual, specified clinical groups, and reliable and valid outcome criteria; and (3) results that demonstrate superiority to medication, placebo control, or alternative treatments *or* equivalence to already established treatment. As an alternative to experimental between-group studies, these results can also be gained from a larger series of single-case experimental designs (the Task Force group suggested a minimum of 9).

Evidence in support of psychotherapy has increased dramatically over the last 30 years (Lambert and Ogles, 2004). However, as to be expected, the entire 'shift' toward EBP and EST guidelines has been very controversial with concerns ranging from what is the best way to collect and analyze 'the evidence' (see the section Selected Methodological and Conceptual Issues) to the necessary criteria for attaining an evidence-based status. This has led to alternative and extended guidelines and lists of EBP (Chambless and Ollendick, 2001). These lists have subsequently been updated and revised to cover new aspects such as evidence-based therapist–patient relationship factors (e.g., Norcross, 2002).

A Short History of EBP and Psychotherapy Research

The movement toward EBP started with a programmatic article published by Hans-Jürgen Eysenck in 1952 stating that the empirical studies on psychotherapy available at that time indicated that the effectiveness of psychotherapy was not superior to spontaneous remission. This general criticism led to a widespread promotion of psychotherapy research over the following decades. The hundreds of controlled studies that followed resulted in an extensive body of outcome research. Initial systematic reviews, especially the first meta-analysis by Smith and Glass (1977), Smith et al., 1980 have largely confirmed the general effectiveness of psychotherapeutic treatments. However, even in these first systematic reviews, the authors already discussed a more differentiated view because, for example, not all approaches and methods seemed having had the same degree of empirical support.

As a result, the question that therapeutic approach works best is still being evaluated and discussed intensively up to today (Wampold, 2013). However, the stage marked by the competition between rival therapeutic approaches has been supplemented by treatment aptitude research focusing on which treatment is most appropriate for which kind of problem. In addition, it has been argued that it is more important to understand the changes during the therapeutic process and which main factors within this change lead to successful psychotherapies (Orlinski et al., 2004).

Today, hundreds of systematic reviews and meta-analyses summarizing results on different approaches (e.g., cognitive behavior therapy), on different mental health problems (e.g., depression), and on different therapeutic factors (e.g., therapist–client relationship) are available (see Lambert and Ogles, 2004; Orlinski et al., 2004). The most recent developments in psychotherapy research, however, are attempts to identify more general factors or therapeutic principles instead of validating specific therapy procedures or methods. Nonetheless, despite this movement toward a general psychotherapy, there is still an extensive discussion on the relative effectiveness of different

 http://dx.doi.org/10.1016/B978-0-08-097086-8.21081-7

therapeutic approaches. To some extent, these discussions may arise from the different professional identities that are still common in psychotherapy research and practice (e.g., researchers/therapists with psychodynamic vs behavioral orientations). These identities lead to different interests and may blur the perception, analysis, and interpretation of empirical results (Gaudiano and Miller, 2013).

Types of Psychotherapy Research

Consumer Satisfaction

Consumer satisfaction studies come from the tradition in market research by asking clients how satisfied they are with a product or service. These are mostly based on surveys using large representative samples. This perspective on psychotherapy was long ignored, probably because client satisfaction with therapy was viewed as a less valid criterion. However, this perspective on outcome research has also confirmed the general efficacy of psychotherapy. One prominent example is Seligman's (1995) study that provided additional evidence that participants of psychotherapy benefit from a broad array of treatments. At the same time, one should be aware of important methodological limitations to this approach such as the use of posttherapy retrospective ratings as a measure of improvement as well as biased selection within samples.

Efficacy/Effectiveness Research

The most important question for EBP concerns the outcomes of psychotherapeutic treatments. Outcome research can be differentiated into *efficacy research* (i.e., outcome research under well-controlled settings with randomized trials) and *effectiveness research* (i.e., outcome research in clinically representative settings). The most important issues in these types of research are the study design, the threats to internal and external validity, and the selection of outcome criteria (see the section Selected Methodological and Conceptual Issues). Randomized controlled trials (RCTs) are regarded as the 'gold standard' of efficacy research, whereas quasi-experiments are typically used in effectiveness research. There are several differences between these two strategies, for example, homogeneous versus heterogeneous clinical samples or randomization versus statistical control as strategies for controlling between group differences. However, although RCT and efficacy research is viewed as a necessary strategy for EBP, effectiveness research is viewed as a complementary step for the transfer of EBP into evidence-based practice.

Efficiency Research

Because financial costs and resources have become an important issue in health care systems, the question of costs in relation to benefit are also important in psychotherapy and EBP. Efficiency research can be divided into cost–benefit and cost-effectiveness analyses (Yates, 1995). Cost–benefit analyses compare the monetary cost of psychotherapy in relation to its monetary benefits. Monetary costs are usually taken to be the cost of the therapy itself, the loss of working time for the client, the cost for qualification and training of therapists, materials expenses such as rental costs for rooms, and so forth. Monetary benefits are, for example, lower cost for medical treatments, fewer days out of work, more taxes paid, and the like. However, because effects of psychotherapy are usually hard to calculate in monetary terms (e.g., what is the monetary benefit of a better marital relationship?), cost-effectiveness analyses are replacing monetary value as a way of measuring the benefits of a particular outcome (e.g., how much money is necessary to cure someone from depression with a particular treatment?). Although the efficiency of psychotherapy has been evaluated much less than its efficacy, existing results usually reveal that psychotherapy is also successful in monetary terms (see, e.g., Layard et al., 2007).

Process Research

Psychotherapy process research aims to specify relevant change processes during treatment and identify those within-treatment factors that allow the best prediction of successful outcomes. In general, the specification of relevant change processes should lead to a better fit between psychotherapeutic treatments and client characteristics (treatment aptitude studies) and a better understanding of general factors and mechanisms of change (Orlinski et al., 2004). Psychotherapy process research is diverse and covers various issues and levels of analysis. For example, process characteristics can range from microlevel variables such as concrete therapist–client interaction to characteristics of therapeutic sessions and specific phases (e.g., how well a sustainable therapeutic bond between therapist and client has been established) up to the analysis of biographical change processes over years. In addition, process research can be conducted in several ways, for example, by identifying characteristic patterns of change during therapy, outcomes moderators (e.g., results for specific subgroups), and outcome mediators (e.g., quality of therapist–client relationship in relation to outcomes). In general, a number of process characteristics have been identified as having a significant effect on psychotherapy outcomes. For example, a number of studies have found that variables such as a positive therapeutic bond or alliance are the most influential characteristics of a successful psychotherapy (Orlinski et al., 2004).

Meta-Analysis

Because thousands of efficacy and effectiveness studies have investigated outcomes of psychotherapy over the past decades, systematic efforts have been made to summarize these results via meta-analysis. The term meta-analysis refers to systematic methods for aggregating empirical results of different primary studies on the same topic, and it is a widely used approach in psychotherapy research (Lambert and Ogles, 2004; Lipsey and Wilson, 1993). Typically, a meta-analysis starts with an exact definition of the eligibility criteria for the inclusion of a study in a research review (e.g., what kind of intervention has been evaluated, which clients have been treated, etc.) followed by an intensive literature search based on these selection criteria. After systematically coding the characteristics of the included studies, unified effect size parameters are calculated for each

study. This metric allows the calculation of mean results across a large body of studies and the identification of which study characteristics systematically relate to the results.

At the beginning of the meta-analytic era, this approach was used for broadband reviews of psychotherapy outcomes. For example, Smith and Glass (1977) found a mean effect size of $d = 0.85$ across 475 studies, indicating that the average psychotherapy client is superior to 80% of the controls receiving no or only placebo treatment. More recent psychotherapy meta-analyses have addressed specific questions, such as specific treatments, specific disorders, specific therapist factors, or even combinations of these factors (e.g., Butler et al., 2006; Cabral and Smith, 2011; Cuijpers et al., 2011; Tefikow et al., 2013). Meta-analysis has also been used to integrate comparative treatment research in which different therapeutic approaches are compared directly (e.g., Shapiro and Shapiro, 1982; Tolin, 2010). In sum, meta-analytic research has revealed that psychotherapy is generally effective over a great variety of mental health disorders; that cognitive behavioral therapy is the best investigated approach; and, finally, that although a lot of within-therapy factors influence psychotherapy outcomes, the therapeutic alliance between therapist and client is the fundamental base for therapeutic change.

Research on Transfer into Evidence-Based Practice

In addition to gathering evidence on the empirical foundation of psychotherapy, the evidence-based movement and the development of treatment guidelines aim to promote evidence-based practice. However, recent research has shown that – as in other fields of intervention – transferring EBP into practice is not an easy task. As a result, a macroperspective on the dissemination of EBP within the health services and the road to evidence-based practice is becoming an important field of research (see Gaudiano and Miller, 2013). This entails implementation control and at least a basic documentation of the implementation of individual therapies (e.g., within systematic quality assurance), as well as research on the disseminability, availability, acceptability, and utilization of EBP within the mental health services.

Selected Methodological and Conceptual Issues

Research Designs

Research designs are strategic plans for examining causal effects, that is, plans regarding how to confirm proposed effects of psychotherapy in a valid way. RCTs are viewed as the 'gold standard' in EBP. Within RCTs, clients are assigned randomly to experimental groups in studies typically comparing one or more treatment condition with one (or more) control condition. Relevant outcome criteria are then used to assess these groups at least before and after the treatment.

There are different types of nontreated controls such as placebo, attention, low-intensity, or no-treatment conditions. Placebo or attention controls are applied to control unspecific effects of psychotherapy (i.e., effects due simply to applying some kind of treatment). However, the rationale of placebo and attention groups is discussed in a controversial manner, because it is argued that these 'unspecific' effects are a genuine part of psychotherapeutic treatments. Real no-treatment conditions are often questioned for ethical reasons (i.e., no treatment in the case of psychological disorders). Therefore, studies often apply waiting list controls (i.e., clients who will receive therapy after the posttest assessment). Apart from the discussion on control conditions, the assessment of long-term effects (e.g., follow-up assessments after the termination of treatment) is regarded as crucial for EBP.

Because randomization is not always possible, at least in clinical settings, design alternatives such as quasi-experiments and single case studies are frequently used to evaluate psychotherapeutic treatments. Within quasi-experiments, assignment to groups (treatment vs controls) is not based on randomization. In fact, effectiveness studies within real clinical settings often try to collect data on widely comparable units (clients from Clinic A vs Clinic B) when examining the outcomes of psychotherapy. Of course, these strategies have disadvantages due mainly to limited comparability and subsequent threats to validity such as selection bias (see below). However, it is argued that they have a higher external validity and are more appropriate for generalization and the transfer to practical settings. Nonetheless, depending on the special field of investigation, these designs really suffer from having limited value in claiming causal inference.

Another problem in psychotherapy research is collecting sufficiently large and homogeneous clinical samples, because some disorders are relatively rare. One solution in these cases is experimental single-case designs (see Kazdin, 2011). These study individual cases more intensively and under special conditions in order to make valid causal inferences. One of the most frequently used single-case designs is the multiple baseline design (MBD). MBD uses continuous assessments prior to (baseline), within, and after psychotherapeutic treatments. The design varies according to whether there are multiple cases (more than one), multiple behaviors (treatment effects are tested on at least two different behaviors), or multiple settings (treatments in different situational conditions). The core principle of this design is that these multiple units are not tested at the same time but successively in order to see whether a possible effect appears repeatedly. This repeated testing should strengthen the causal inference.

Problems in Recruitment and Selection into the Treatment

One of the most challenging tasks within empirical clinical studies is to collect comparable treatment and control groups that are matched in terms of sociodemographic variables, type and severity of clinical disorder, motivation for therapy, and so forth. Randomization prior to therapy with RCTs should guarantee that clients are initially comparable. However, premature termination of treatment and attrition in experimental studies are still possible even in RCTs, and these aspects need to be considered when analyzing the results. For example, highly motivated clients usually have a higher probability of being assigned to the treatment conditions because they will terminate their participation in a study if randomly assigned to the control group (or, vice versa, less motivated clients will terminate treatment prematurely). These selections

into treatment can cause severe threats to validity, because motivation for change at the beginning of the treatment normally correlates significantly with the treatment outcome, and this usually leads to an overestimation of treatment effects. Quasi-experiments (in which complete comparability is not expected) offer some strategies for controlling or enhancing comparability (also used in RCTs when randomization failed). For example, one can search for comparable units on the individual level (i.e., two individuals share a set of same characteristics and then one of these experimental 'twins' is assigned to the condition at random). With respect to premature termination and attrition, researchers should at least control whether terminators differ from completers on relevant characteristics (e.g., sociodemographic variables, pretest level of outcome criteria). A more specific technique is intention-to-treat analysis (ITT). This compares all clients in the treatment group (regardless of whether they terminate treatment) with subjects in the control condition. Using this strategy, premature termination usually leads to lower estimations of treatment effects because terminators are expected to have weaker outcomes then completers. However, ITT works on the premise that terminators do not leave the entire study (at least pre–posttest data should be available). Finally, comparability can also be achieved by statistical procedures. One of the simplest strategies is analysis of covariance in which treatment outcomes are corrected for significant covariates (e.g., initial values of the outcome criteria). Modern statistical procedures use techniques such as propensity score matching to attain comparability among comparison groups. Estimated propensity scores are used to select comparable clients for the treatment and control condition or correct results according to these propensity scores. The propensity scores are usually calculated on the basis of empirically derived predictors that lead to selection into treatment and as a result, they control associated threats to validity (e.g., selection bias).

Type of Outcome Criteria

Selecting reliable and valid outcome criteria is an important step in a psychotherapy study. This includes not only methodological (e.g., psychometric quality of assessment tools) but also substantial (e.g., what will be measured?) issues. Substantial issues relate to the goals of treatment and imply the operationalization of outcome criteria derived from these goals. In general, psychotherapy has three global aims: (1) to cure from a disorder or at least relieve symptoms or problems; (2) to promote client development (e.g., personality change, adaptive functioning); and (3) to promote the client's general quality of life. Curing a disorder or relief from symptoms is superordinate to both others goals, because a decrease in problems will naturally lead to an increase in a person's development and quality of life. Specifying outcome criteria to measure these goals basically has to address the following four questions:

- What should be assessed?
  The first step is to determine the *content of outcome criteria*. For example, what should be assessed when treating depression? In general, three types of measures are applied: (1) ratings of client satisfaction with the treatment and subjective ratings of the outcome; (2) clinical measures assessing the symptoms, disorders, and important individual problems; and (3) additional assessments of important psychosocial aspects and further outcomes. Satisfaction with the treatment and direct subjective ratings of effects are usually applied within consumer satisfaction studies (see above). Clinical measures include the assessment of the clinical status through clinical interviews using approved clinical diagnostic systems (ICD-10, DSM-V), an assessment of clinical symptoms by symptom list (e.g., the Symptom Checklist-90-R), and problem-related assessment tools (e.g., rating tools such as the Beck Depression Inventory). In addition, it is considered useful to evaluate psychotherapy by individualizing treatment goals according to each individual client and assessing the expected and real changes using, for example, goal attainment scaling (GAS). GAS is an assessment tool that reviews defined individual treatment goals by using a uniform rating of expected versus real improvements. Finally, most psychotherapy studies apply additional aspects of the outcome measurement, such as self-control, self-efficacy, quality of life, and personality assessments. Independent from this selection, outcome criteria should also consider different types of human cognitions, emotions, and behavior (i.e., multitrait assessment).
- What kind of assessment tool?
  The selection of the assessment method depends upon the psychometric properties (e.g., standardization, reliability, validity) of the method as well as on its practicability and past utilization. It is advisable to use at least two different methods (i.e., multimethod assessment), although the cost of collecting data and the availability of informants have to be taken into account (e.g., collecting observational data is normally time consuming). Especially in psychotherapy treatment research, instruments should be standardized for clinical populations. However, the application of standardized tests is not without its problems, because these have usually been constructed to assess status and not change. Therefore, it is necessary to select instruments with a demonstrated sensitivity to psychological and clinical change.
- Who is the informant?
  To some extent, the effects of psychotherapy also depend on the informant, that is, on who delivers the data on possible changes. There are several options here: the client, the therapist, independent experts, or relevant others (e.g., partner, parents in case of child psychotherapy). The most important informant is the client, although multi-informant assessment is usually preferable.
- When to collect the data?
  Outcome criteria can be assessed in three different ways: one single measurement after the termination of the therapy (direct change measurement); pre–postcomparisons, preferably with long-term follow-ups; and, finally, multiple measurements at baseline, within, and after therapy. Multiple measurements are naturally expensive, but deliver detailed and reliable information on therapeutic processes. In addition, repeated follow-ups are considered to be necessary for evaluating long-term effects.

Statistical Analysis and Interpretation of Results

The statistical analysis of outcomes depends mainly on the time and frequency of assessment points. In general, there are three different statistical indicators for therapeutic improvements: First, improvement rates can be calculated based mainly on clinical diagnoses of clients (e.g., proportions of improved clients compared to nontreated groups). Second, pre–posttest data are usually confirmed by *statistical* and additionally by *clinical significance* (Ogles et al., 2001). Clinical significance requires comparisons between a treated clinical group and normative scores (e.g., from nonclinical groups). Clinical significance is assumed when treated clinical groups score as high as members of a nonclinical group. Finally, change can be expressed as standardized *effect sizes* as applied mostly within meta-analyses (Ellis, 2010). Most common in intervention research is the *d* index (standardized differences between a treated and a nontreated group at posttest or follow-up expressed in standard deviation units) and the odds ratio (odds for improvement within the treated vs nontreated group).

Conflict of Interest

One important aspect of psychotherapy research concerns the independence of results. Meta-analytic data have revealed that psychotherapy studies carried out by program developers or proponents of a specific treatment approach deliver higher effects than studies that are not influenced by some kind of personal or financial conflict of interest. These influences emerge when, for example, program developers are not willing to publish negative or zero results on 'their' treatment (the so-called publication bias) or report only significant results in their publications. To counter such sources of bias, the aforementioned treatment guidelines insist on at least one independent replication study with comparable results before an EBP can be assumed.

Evolution of EBP

Different types of research have indicated that the evolution of EBP cannot be achieved by one single study. Therefore, the development of EBP is normally a process passing through

Table 1 Empirically supported psychotherapeutic interventions for anxiety and depression

Disorder	*Treatment*	*Empirical support*	*Conclusion*
Depression	Behavior therapy	Strong	Efficacious
	Cognitive therapy	Strong	Efficacious
	Cognitive behavioral analysis system of psychotherapy	Strong	Efficacious
	Interpersonal therapy	Strong	Efficacious
	Nondirective supportive therapy	Strong	Efficacious
	Acceptance and commitment therapy	Modest	Probably efficacious
	Behavioral couple therapy	Modest	Probably efficacious
	Emotion-focused therapy	Modest	Probably efficacious
	Short-term psychodynamic therapy	Modest	Probably efficacious
Generalized anxiety disorder	Cognitive and behavioral therapies	Strong	Efficacious
Obsessive-compulsive disorder	Behavior therapy (exposure and response prevention, ERP)	Strong	Efficacious
	Cognitive therapy	Strong	*Without ERP only probably efficacious*
	Acceptance and commitment therapy	Modest	Probably efficacious
	Eye movement desensitization and reprocessing	Modest	Probably efficacious
	Stress management	Little	Likely not efficacious
	Psychodynamic therapy	Little	Likely not efficacious
Panic disorder	Cognitive behavior therapy	Strong	Efficacious
	Applied relaxation	Modest	Probably efficacious
	Psychodynamic treatment	Modest	Probably efficacious
Post-traumatic stress disorder	Prolonged exposure	Strong	Efficacious
	Present-centered therapy	Strong	Efficacious
	Cognitive processing therapy	Strong	Efficacious
	Stress inoculation therapy	Modest	Probably efficacious
	Eye movement desensitization and reprocessing	Strong	Efficacious
	Psychological debriefing	Little	Likely not efficacious Potentially harmful
Social phobia	Cognitive and behavioral therapies	Strong	Efficacious
Specific phobias	Exposure therapies	Strong	Efficacious

Based upon the American Psychological Association, 2008; modified by updated results from several recent meta-analysis articles (e.g., Cuijpers, P., Andersson, G., Donker, T., van Straten, A., 2011. Psychological treatment of depression: results of a series of meta-analyses. Nordic Journal of Psychiatry 65, 354–364; Hollon, S.D., Ponniah, K., 2010. A review of empirically supported psychological therapies for mood disorders in adults. Depression and Anxiety 27, 891–932).

successive steps. At the beginning, pilot studies with group samples or a single-case design is used to test the general acceptability and feasibility of a novel psychotherapeutic method. Subsequently, RCTs – ideally from different research groups – are needed to confirm the efficacy of this new approach. These studies should be followed by preferably diverse effectiveness studies to test the application in practice. These studies should be supplemented by efficiency and process-outcome studies to determine cost–benefit ratios and provide information on relevant therapeutic processes. Finally, studies on dissemination should be conducted to investigate the transfer to mental health services and evidence-based practice.

Evidence-Based Psychotherapies

After about 60 years of intensive research on psychotherapy outcomes, there is now an extensive body of research showing that EBP can help patients to maintain mental health and increase their quality of life (APA, 2013). However, as the body of evidence for psychotherapy approaches is so vast, a complete illustration of EBP would go beyond the scope of this article (see Chambless and Ollendick, 2001; Chambless and Klonsky, 2013; or relevant web sites, e.g., http://www.psychologicaltreatments.org and http://www.evidencebasedpsychotherapies.org, for reviews). As an example, **Table 1** presents an overview of EBP in the field of depression and anxiety as two very prevalent mental disorders.

The most promising approach across all types of depression and anxiety disorders is cognitive and behavior therapies (CBTs). CBT encompasses various approaches and has a large body of evidence that it successfully treats these disorders. A recent meta-analysis of nonrandomized effectiveness studies on outpatient CBT for anxiety disorders (Hans and Hiller, 2013) has also demonstrated that outpatient individual and group CBT is very effective in routine practice settings and has a high level of acceptability among clients. Other therapeutic approaches and methods have more specific potentials. Examples are interpersonal psychotherapy on depression or eye moment desensitization and reprocessing on post-traumatic stress disorder.

Conclusion

Taken together, there is substantial evidence that psychotherapy is efficacious. This is true for a number of different approaches, although not all demonstrate similarly strong evidence across all disorders. However, when focusing on the relative efficacy of these treatment approaches, recent reviews have shown that it is difficult to confirm superiority of any one approach over another. When only studies directly comparing two treatments are considered (which were both intended to be used in therapeutic settings, i.e., bona fide psychotherapies), many yield effect sizes around zero (e.g., Benish et al., 2008; Wampold et al., 2011). This implies a lack of superiority of one evidence-based treatment (classified either as strong or modest in comparison to a control group; see **Table 1**) over another (Benish et al., 2008) and also illustrates that the relative efficacy of evidence-based therapies (EBTs) versus treatment-as-usual (TAU) may depend on the specific TAU conditions (Wampold et al., 2011). For example, EBT is relatively more effective when TAU is defined as minimal mental health services. However, it does seem less efficacious when compared to psychotherapy services delivered in routine care. Another example is a recent meta-analysis comparing the efficacy of seven psychotherapeutic interventions for patients with depression (Barth et al., 2013). Again, each of the psychotherapeutic interventions was deemed superior to a waiting list control condition (with moderate-to-large effect sizes), whereas comparisons between these interventions yielded at best small effect sizes.

These results have encouraged researchers to focus on general principles of change rather than on comparisons between therapy procedures or methods. For example, Goldfried (2013) has recently suggested five common change principles on the basis of empirical evidence: (1) facilitating expectations that therapy can be helpful; (2) establishing an optimal therapeutic alliance; (3) facilitating the patient's awareness of those factors within themselves, others, and their environment that contribute to their problems; (4) facilitating corrective experiences; and, finally, (5) encouraging continued reality testing. These principles are supposed to reflect general change processes across different treatment procedures. Indeed, deriving principles of change from the large body of evidence available for psychotherapies maybe more helpful in developing evidence-based practice than comparing outcomes for different treatment procedures.

See also: Anxiety and Anxiety Disorders; Depression; Post-Traumatic Stress Disorder; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychological Treatments: Randomized Controlled Trials; Psychotherapy Process Research.

Bibliography

American Psychological Association, 2013. Recognition of psychotherapy effectiveness. Psychotherapy 50, 102–109.

Barth, J., et al., 2013. Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. PLoS Medicine 10, 1–17.

Benish, S.G., Imel, Z.E., Wampold, B.E., 2008. The relative efficacy of bona fide psychotherapies for treating post-traumatic stress disorder: a meta-analysis of direct comparisons. Clinical Psychology Review 28, 746–758.

Butler, A.C., Chapman, J.E., Forman, E.M., Beck, A.T., 2006. The empirical status of cognitive-behavioral therapy: a review of meta-analyses. Clinical Psychology Review 26, 17–31.

Cabral, R.R., Smith, T.B., 2011. Racial/ethnic matching of clients and therapists in mental health services: a meta-analytic review of preferences, perceptions, and outcomes. Journal of Counseling Psychology 58, 537–554.

Chambless, D.L., Hollon, S.D., 1998. Defining empirically supported therapies. Journal of Consulting and Clinical Psychology 66, 7–18.

Chambless, D.L., Klonsky, E.D., 2013. Compendium of empirically supported treatments. In: Koocher, G.P., Norcross, J.C., Greene, B.A. (Eds.), Psychologists' Desk Reference, third ed. Oxford University Press, New York, pp. 159–166.

Chambless, D.L., Ollendick, T.H., 2001. Empirically supported psychological interventions: controversies and evidence. Annual Review of Psychology 52, 685–716.

Cuijpers, P., Andersson, G., Donker, T., van Straten, A., 2011. Psychological treatment of depression: results of a series of meta-analyses. Nordic Journal of Psychiatry 65, 354–364.

Ellis, P.D., 2010. The Essential Guide to Effect Sizes. Statistical Power, Meta-analysis, and the Interpretation of Research Results. Cambridge University Press, Cambridge.

Gaudiano, B.A., Miller, I.W., 2013. The evidence-based practice of psychotherapy: facing the challenges that lie ahead. Clinical Psychology Review 33, 813–824.
Goldfried, M.R., 2013. What should we expect from psychotherapy? Clinical Psychology Review 33, 862–869.
Hans, E., Hiller, W., 2013. A meta-analysis of nonrandomized effectiveness studies on outpatient cognitive behavioral therapy for adult anxiety disorders. Clinical Psychology Review 33, 954–964.
Hollon, S.D., Ponniah, K., 2010. A review of empirically supported psychological therapies for mood disorders in adults. Depression and Anxiety 27, 891–932.
Kazdin, A.E., 2011. Single-Case Research Designs: Methods for Clinical and Applied Settings, second ed. Oxford University Press, New York.
Lambert, M.J., Ogles, B.M., 2004. The efficacy and effectiveness of psychotherapy. In: Lambert, M.J. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change. Wiley, New York, pp. 139–193.
Layard, R., Clark, D., Knapp, M., Mayraz, G., 2007. Cost-benefit analysis of psychological therapy. National Institute Economic Review 202, 90–98.
Lipsey, M.W., Wilson, D.B., 1993. The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. American Psychologist 48, 1181–1209.
Norcross, J.C. (Ed.), 2002. Psychotherapy Relationships That Work: Therapist Contributions and Responsiveness to Patients. Oxford University Press, New York.
Ogles, B.M., Lunnen, K.M., Bonesteel, K., 2001. Clinical significance: history, application, and current practice. Clinical Psychology Review 21, 421–446.
Orlinski, D.E., Rønnestad, M.H., Willutzki, U., 2004. Fifty years of psychotherapy process-outcome research: continuity and change. In: Lambert, M.J. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change. Wiley, New York, pp. 307–389.
Seligman, M.E.P., 1995. The effectiveness of psychotherapy: the consumer report study. American Psychologist 50, 965–974.
Shapiro, D.A., Shapiro, D., 1982. Meta-analysis of comparative therapy outcome studies. A replication and refinement. Psychological Bulletin 122, 203–215.
Smith, M.L., Glass, G.V., 1977. Meta-analysis of psychotherapy outcome studies. American Psychologist 32, 752–760.
Smith, M.L., Glass, G.V., Miller, T.I., 1980. The Benefits of Psychotherapy. John Hopkins University Press, Baltimore.
Tefikow, S., et al., 2013. Efficacy of hypnosis in adults undergoing surgery or medical procedures: a meta-analysis of randomized controlled trials. Clinical Psychology Review 33, 623–636.
Tolin, D.F., 2010. Is cognitive-behavioral therapy more effective than other therapies? A meta-analytic review. Clinical Psychology Review 30, 710–720.
Wampold, B.E., et al., 2011. Evidence-based treatments for depression and anxiety versus treatment-as-usual: a meta-analysis of direct comparisons. Clinical Psychology Review 31, 1304–1312.
Wampold, B.E., 2013. The good, the bad, and the ugly: a 50-year perspective on the outcome problem. Psychotherapy 50, 16–24.
Yates, B.T., 1995. Cost-effectiveness analysis, cost-benefit analysis, and beyond: evolving models for the scientist-manager-practitioner. Clinical Psychology Review 2, 385–398.

Relevant Websites

http://www.evidencebasedpsychotherapies.org – Psychotherapy randomized controlled and comparative trials (database on psychotherapy RCTs for meta-analysis).
http://www.psychologicaltreatments.org – Society of Clinical Psychology. Website on Research-Supported Psychological Treatments.

Evidentials in Linguistics

Zlatka Guentchéva, LACITO – CNRS (Centre National de la Recherche Scientifique), Paris, France

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by Z. Guentchéva, volume 7, pp. 4972–4975, © 2001, Elsevier Ltd.

Abstract

This article presents an overview of the notions of evidentials and evidentiality. It briefly outlines the history of evidentials and the various types of evidentiality established. It looks at these notions' semantics and summarizes some of the more influential publications on the matter. It further looks at interaction between evidentials and tense/aspects morphemes, and discusses some theoretical issues. It mentions some sources for the emergence of evidentials and recent controversy over grammatical versus lexical approaches.

The label evidential originated in descriptions of verbal affixes in American Indian languages where evidence for certain statements must be specified (e.g., dreams, sounds, hearsay, inferences, sightings, cf Boas, 1911, 1947). It occurs once in Boas' posthumously published *Kwakiutl grammar* (1947, p. 237) but the notion is retained in several of his studies on verbal categories, and is generally discussed under the heading 'source of information' with the gloss 'evidently (as is shown by evidence)' (Boas, 1947, p. 245; 1911, p. 496; see Jacobsen (1986, p. 4–7) and Aikhenvald (2004, pp. 11–17) for further references of grammaticalized information sources). However, it was Jakobson (1957/1971, p. 135) who first introduced the term in a broad theory of generic verbal categories as "a tentative label for the verbal category" seen as an interaction between three events: "a narrated event (E^n), a speech event (E^s), and a narrated speech event (E^{ns}), namely the alleged source of information about the narrated event. The speaker reports an event on the basis of someone else's report (quotative, i.e. hearsay evidence), of a dream (revelative evidence), of a guess (presumptive evidence), or of his own previous experience (memory evidence)." This is illustrated by Bulgarian examples contrasting with 'direct narration' ($E^{ns} = E^s$) versus 'indirect narration' ($E^{ns} \neq E^s$): *zaminala* 'it [the steamer] is claimed to have sailed' versus *zamina* 'I bear witness; it [the steamer] sailed.'

Referring to grammaticalized information sources, evidentials were first related to a closed class set of obligatory verbal morphemes denoting the source of subjective knowledge or the source of information (for an historical overview see Jacobsen, 1986; de Haan, 1999, 2001; Comrie, 2000; Aikhenvald, 2004; Plungian, 2010). The notion appears to have become well-established by the mid-1960s (Jacobsen, 1986: 6) but it seems to have been used rather infrequently by Americanists until the publication of the collective volume by Chafe and Nichols (1986). It was progressively extended to a repertoire of grammatical devices coding the different kinds of information sources on which statements are based, and used in descriptions of typologically different languages. In this 'broad sense,' evidential markers encode not only the type of information source but also the speaker's attitude toward the reliability of the information (based on hearsay or inference), as well as the degree of probability underlying the truth value of the utterance (Chafe and Nichols, 1986, p. vii, Mithun, 1986) or its epistemological assessment independently of its grammatical status (Chafe, 1986). As a result, the relations and boundaries between epistemic modality and evidentiality has been often debated: evidentiality has been categorized as part of modality or speaker beliefs (Palmer, 1986); from a typological point of view, Van der Auwera and Plungian (1998, pp. 85–86) show various forms where the two notions cluster, more precisely in the inferential domain; the diachronic relationship between the two notions has been questioned by Bybee et al. (1994, pp. 195–205); de Haan (2005) has noted that there is a connection between evidentiality and deixis. Data from different languages show that evidentiality does not relate to the truth or falsity the speaker's statement (Guentchéva, 1996, p. 11; Lazard: 2000, p. 212; Aikhenvald, 2003, 2004, p. 5, Joseph, 2003, p. 309).

The notion of evidentiality became mainstream mostly in linguistics after the publication of the collective volume by Chafe and Nichols (1986), as well as papers published by Barnes (1984) and Malone (1988) on Tuyuca, or Willett's first typological study (1988), and especially with Aikhenvald's monograph (2004) dealing exclusively with the category of evidentiality in a cross-linguistic perspective. Numerous articles dealing with various meanings of evidentials, and other collective volumes have been published such as Guentchéva (1996), Johanson and Utas (2000), Aikhenvald and Dixon (2003), Guentchéva and Landaburu (2007), and Diewald and Smirnova (2010), as well as the special issues of the *Journal of Pragmatics* 33 (Dendale and Tasmowski (eds), 2001) and of *Functions of Language* 16 (1) (2009).

In formal linguistics, Cinque (1999) was the first to include evidentiality as a separate category distinct from modality in his work on the cartography of the syntax structure. In formal semantics, several studies on evidentials in Cuzco Quechua have been carried out, for example, Faller (2002). Interaction between evidentiality and culture has also been investigated in anthropological linguistics (Basso, 2008).

Earlier descriptions of what today would be called evidential systems go back to al-Kāšǧārī's analysis of the

di-past/*miş*-past opposition in *Dīwān Luǧāt at-Turk* (Compendium of the Turkic Dialects) in the eleventh century (Friedman, 1986) and other European studies at the end of the nineteenth century and in the beginning of the twentieth century Haarmann's study (1970) of languages in the Euro-Asian area where 'indirect experience' is analyzed as a separate category and contrasted with neutral verbal forms, rather than as a subtype of a general category expressing information sources, as well other European studies seem to have had little influence on Anglo-American linguistics.

Linguistic terminology is not homogeneous and various terms have been adopted in different grammatical traditions: *indirect* (Haarmann, 1970) or *indirective* for Turkic languages (Johanson, 2000), *nonconfirmative* for Balkan Slavic (Friedman, 1986, 1999), *validational* or *verificational* in Quechua (Weber, 1986), *data-source* in the Jaqi languages (Hardman, 1986), *mediative* (Lazard, 1996; Guentchéva, 1996b), *presumptive* in Romanian, *oposredstvenost* distinguishing *renarrative* and *conclusive* in Bulgarian, *kosvennaja zasvidel'stvannost'* ('indirect evidence') or *zaglasnost* (lit. 'behind one's back') in Russian, etc.

Over the last few decades, the study of evidentiality has given rise to increasingly abundant research. Even though most scholars recognize the notions 'information source,' 'evidence,' 'type of evidence,' 'source of evidence,' or 'mode of knowing' as basic governing principles of the category of evidentiality, it must be observed that these notions are interpreted in different ways and the choice between these notions has clear consequences for analysis of the phenomenon with regard to the diversity of languages. Aikhenvald (2003, pp. 18–20, 2003, pp. 147–151, 2008) and Lazard (2000, p. 214, 2001) advocate a strict grammatical definition of evidentiality and a clear-cut distinction between grammatical and lexical expressions of evidentiality (such as the English adverbs *reportedly*, *allegedly*, and *evidently*, particles referring to verbal report or inference, and parentheticals). On the other hand, scholars such as Squartini (2007), Wiemer (2007), and Boye and Harder (2009) aim to establish functional grounds for cross-linguistic analysis and a unified framework for both lexical and grammatical devices.

It has often been noted that all languages have lexical, syntactic, or other means of referring to information sources, but not all languages have an evidential grammatical category. Grammatical evidentiality "is understood as stating the existence of a source of evidence for some information; that includes stating that there is some evidence, and also specifying what type of evidence there is." (Aikhenvald, 2003, p. 1). Aikhenvald (2004) has brought together data on over 500 languages with obligatory evidential morphemes and reports that about one quarter of the world's languages has some type of grammatical evidentiality. She lists six semantic parameters used in languages with grammatical evidentiality, five being the maximum within a given grammatical system: *visual* (evidence acquired through seeing), *nonvisual sensory* (evidence through hearing, smell, and taste, and sometimes touch), *inference* (based on visible or tangible evidence or results), *assumption* (evidence other than visible results: logical reasoning, assumption, or general knowledge), *reported* (reported information with no reference to who reported it), and *quotative* (information with an overt reference to the quoted source). This can be illustrated with the following example from Tariana, an Arawak language spoken in northwest Amazonia (Aikhenvald, 2004, pp. 2–3):

1. a. Juse irida di-manika-*ka*
   José football 3person.Masc.Sg-play-RECENT.PAST.VISUAL
   'José played football (we saw it)'

   b. Juse irida di-manika-*mahka*
   José football 3person.Masc.Sg-play-RECENT.PAST.NON-VISUAL
   'José played football (we heard it)'

   c. Juse irida di-manika-*nihka*
   Jose' football 3person.Masc.Sg-play-RECENT.PAST.INFERRED
   'José played football (we infer it from visual evidence)'

   d. Juse irida di-manika-*sika*
   José football 3person.Masc.Sg-play-RECENT.PAST.ASSUMED
   'José played football (we infer it from general knowledge)'

   e. Juse irida di-manika-*pidaka*
   football 3person.Masc.Sg-play-RECENT.PAST.REPORTED
   'José played football (we were told)'

Numerous languages, especially within the East Asia-New World complex, encode in their verbal system specific forms for expressing the way in which the information uttered by the speaker has been acquired. However, evidential systems often interact simultaneously with other linguistic categories such as tense and aspect, person, number, and gender. Like Tariana, Tuyuca (Eastern Tucanoan language, spoken in Brazil and Columbia), often presented as having a 'classical evidential system,' distinguishes five evidentials. To say 'he played soccer,' the speaker has to choose the appropriate suffix (Barnes, 1984, p. 256):

2. a. díiga apé-*wi* (I saw him play)

   b. díiga apé-*ti* (I heard the game and him, but I did not see it or him)

   c. díiga apé-*ɣi* (I have seen evidence that he played: his distinctive shoe print)

   d. díiga apé-*ɣigi* (I obtained the information from someone else)

   e. díiga apé -*hĩɣi* (It is reasonable to assume that he did)

As shown in (2), Tuyuca distinguishes: *visual* in (2a), *auditory* in (2b), and *inferential* in (2c), used to present information based on observed traces of an earlier event; *reported* in (2d) and *assumed* in (2e), based on general knowledge about someone's habits. In fact, Tuyuca is a language with very complex interactions between evidentiality and other categories, namely person and tense as illustrated in the following compound construction with an auxiliary verb

where the visual evidential *-wi* interacts with the 'resultative' morpheme *-ri*:

3. Wesé sóe-ri-gɨ nĩi-wi.
   Field burn-RES-Masc.Sg AUX-VIS.3Sg.Masc.PAST
   Gloss: I saw his field and it had been burned
   'He burned his field.' (Barnes, 1984: 259)

It is clear that (3) describes a resultative state due to an earlier event which the speaker did not witness. From a diachronic point of view, the visual suffix could be a relic from an earlier language state (Malone 1988, p. 132); moreover, examples as (3) show that visual interpretation appears to have resulted from the reanalysis of a tense morpheme.

In Wintu, a Pomoan language, evidentiality correlates with aspectual morphemes. According to Schlichter (1986, p. 54), Wintu has an evidential deictic system establishing a strict grammatical distinction between direct 'visual experience' (unmarked) and, indirect experience, that is "an alternative to the temporal orientation of "Western" civilization." It has four evidentials:

4. a. heket wira waca:-bi-**nt^{h}e**ˑ-*m*
   Someone come cry-Imperfective-NON-VISUAL SENSORIAL-DUB=3 person subject
   'Someone is coming crying (I hear)'

   b. Č'epkal neˑ**l** baˑ-bi-**nt^{h}i**-da h
   bad we eat-Imperfective-SENSE-we
   'We have been eating bad things (I sense)'

   c. Minel kir-**ke**ˑ-m
   die asp-hearsay–dub=3 person subject
   'He has died. (I am told)'

   d. Ničay ʔewin suke-**re**ˑ
   Nephew here stand-INFERENTIAL
   'My nephew must have been here (see tracks)'

   e. ʔlmtop nuquˑʔ**el**
   berries ripe-EXPECTATIONAL
   'The berries must be ripe (it is that time of the year)'

The nonvisual *-nt^{h}er* occurs most often in the third person, as in (4a) and involves 'a fact' known to the speaker through a sense other than vision, or by any kind of intellectual experience or 'sixth sense.' It interacts closely with person, agentivity, and aspect. When in first person (where the subject marker *-da* is required) as in (4b), it involves the speaker's experiential role. Suffixes *-keˑ* and *-reˑ* indicate, respectively, information known through hearsay (4c), and information based on inference (4d). Finally, the suffix *-ʔel* in (4e) is usually used to specify the speaker's 'experience with similar situations, regular patterns, or repeated circumstances common in human life.' It thus has an 'experiential' meaning, though it can also function as a hearsay marker in myths. In fact, (4d) and (4e) denote two types of reasoning (Peirce, 1974, vol. II, p. 636; vol. V, p. 171). In (4d), the speaker indicates, with *-reˑ*, that an earlier plausible event can be inferred. The first step of the reasoning process is the presence of the observed traces, that is, 'q'; the second step is the appeal to memorized general knowledge ('if p, then q'); and the third step is that the claim 'p' is plausible i.e., the event has occurred; abduction, made on the basis of the observed traces 'q' and the knowledge of general laws (Desclés, 2000, p. 106). The utterance is therefore a resultative state of the event denoted by 'p.' The use of *-ʔel* is based on generalization: the speaker indicates that his statement is inferred from experience (at this time of the year the berries should already be ripe): 'the existence of phenomena such as we have observed in cases which are similar'; in fact, *-ʔel* encodes inductive reasoning (in the sense of Peirce, 1974, vol. II, p. 636).

In a number of languages, like in Wintu, there are more than one way of marking inference. For instance, Kashaya (Oswald, 1986) has two sets of inferential markers implying two types of reasoning: *-qă*, called the Inferential I (which regularly changes to *-q^h*), and *-bi*, called inferential II (which can co-occur with evidential morphemes):

5. sinam-**q^h**
   drown-INFR.I
   'He must have drowned'

6. duʔk'u-**bí**-na cohtocch-**q^h**
   finish-INFER.II-SS leave-INFER.I
   'He must have finished and left (the work is done and he is no longer here)'

The reanalysis of these two inferentials could show that they are the result of two types of reasoning: *-qă*, used on the basis of the observed traces, implies that the event is reconstructed through abductive reasoning, while *-bi* is not only a distributional variant of *-qă* but also the result of reasoning leading to a probabilistic conclusion.

Evidentiality is also well attested in Tibetan languages where special suffixes indicate that the situation has been seen (*-song*), is common knowledge or has been heard from someone else (*-pare'*), or inferred (*-sha'*) as in the following example (Tournadre, 1996):

7. mātrokungkar-la sāyom-ki' khābgpa Mangpo lo-*'song/ -pare'/-sha'*
   Metrokungkar-LOC earthquake-ERG house many destroy-AUX
   'In Metrokungkar, the earthquake destroyed many houses'

As in other languages, in Tibetan the reportive morpheme *-sa* may co-occur with any of these verb forms, to indicate the source information of the speaker's informant. The system is even more complex, since these markers interact closely with both the person and the semantics of lexical verbs.

Reportive (also referred to as 'reported,' 'hearsay,' and 'renarrative', and also somewhat misleadingly called 'quotative') is one of the more widespread type of markers, and pertains to information obtained through someone else or hearsay, as shown in the following example from Andoke,

an Amazonian language spoken in Colombia (Landaburu, 2007, pp. 36–37):

8. a. ha-ʌtʌ púkə̄ o-ĩ-ti-ʌ b-aya
   your-son canoe this-i.-make-ACCORD3 ENGAG.-i.31
   (Fr. 'C'est ton fils qui fait un canoë')
   'It is your son who is making a canoe'

   b. Ha-ʌtʌ púkə̄ *o-ĩ-ti-ʌ* b-aya-**há**
   your-son canoe this-i.-make-ACCORD3 ENGAG.-i.31-ha
   (Fr. 'C'est ton fils qui fait un canoë, on me l'a dit.'
   '[I am told] it is your son who is making a canoe'

Some languages have more elaborated reportives. In Andoke, for instance, there are different reportive markers depending on discourse functions: for historical facts the speaker uses *-pé*, if he was present and *-hápé*, if the narrated events are old; for myths, a very, experienced shaman uses the suffix *ñé'*, while a shaman who has less authority may also use *-há ñé'* (Landaburu, 2007, p. 40).

In many languages, reportives coexist with indirect speech (which usually involves overt expression of *verba dicendi*, and deictic transposition which 'shifts' person, tense, and mood). In Maricopa (Gordon, 1986, p. 86), there is a special morpho-syntactic device with a complex interaction of aspect, tense, and evidentiality. Cliticized to the invariant verb form *'-ish* or *ish* "say," *-'a* in (10a), it conveys information acquired by hearing in an independent clause, while *-k* in (10b) introduces 'a complement whose verb is not marked with switch reference suffixes' and whose interpretation is that of indirect speech:

10. a. Bonnie-sh chuy-k-'ish-'a
   Bonnie-Sj marry-**k**-say+sh-hearsay
   '(They said, I hear tell) Bonnie got married'

   b. *Bonnie-sh* *chuy-k uu'ish-k*
   Bonnie-Sj marry-k say+Pl-Asp
   'They say Bonnie got married'

There is a fundamental difference between the evidential markers of American Indian or Tibeto-Burman languages and those of languages spoken in South Eastern Europe and Middle Eastern Asia (such as Albanian, Bulgarian, Macedonian, Iranian, and Turkish) which have forms derived from perfects and resultatives and generally identified as indirect evidentials. These are areas where analyzing these verb forms is often highly complex. This raises the question of whether these languages have the category of evidentiality. For instance, in Turkish, the inflectional marker *–mIş* originates in perfect forms and is the grammatical means of referring to hearsay, inference, or perceived information on the basis of traces or result; it contrasts paradigmatically with the unmarked past *-dI*: in the opposition *-dI/-mIş gel-di* 'has come/came' vs *gel-mIş* ('has obviously come/obviously came'), *-dI* past exhibits "always neutral uses in cases where the speaker considers the evidential distinction inessential" (Johanson 2003, p. 275). Although specification of the information source is not crucial in Turkish, Johanson (2000, 2003, p. 374) explicitly states that "the notion of indirectivity is in accordance with the cross-linguistic definition of evidentiality" and treats it as an evidential category. In Aikhenvald's classification, Turkish has a binary system ('with two choices') of the subtype nonfirsthand versus everything else.

The situation in Balkan Slavic is more complex because new paradigms of aspect-tense forms developed out of the old perfect. While Macedonian never uses the third person auxiliary in aspect-tense forms based on the *l*-participle (Friedman, 1986, 2000), the auxiliary *săm* 'be' can be used or omitted in the third person in Bulgarian, as shown in the following examples: *Toj se prestoril* (renarrative aorist) *na umrjal* 'He pretended to be dead (I am told)' and *Toj se **e** prestoril* (aorist 'conclusivus') *na umrjal* 'He has pretended to be dead (I infer)'; *Toj se prestruval* (renarrative imperfect) *na umrjal* 'He was pretending to be dead (I am told)' and *Toj se **e** prestruval* (imperfect 'conclusivus') *na umrjal* 'He must have pretended to be dead (I infer).' Although the precise function of the auxiliary's omission or presence continues to be a matter of some controversy, Bulgarian cannot be reduced to a Turkish-type binary system: it does not fit into the evidential subtypes 'firsthand and nonfirsthand,' nor in 'nonfirsthand and everything else.' The auxiliary's omission or presence typically involves the speaker: the statement is made on the basis of hearsay, rumor, report, folk tales, and narrating history (auxiliary omission) or on the basis of reasoning leading to formulating a plausible explanation of tangible evidence and can be used for reconstructing facts that are not observed (auxiliary presence).

The basic issue with regard to evidentiality is to find criteria for characterizing the concept of evidential. Anderson (1986) was the first to point out that not all forms that have been called 'evidential' share a common set of 'evidential functions.' He proposes four parameters to identify a true evidential as a specific grammatical phenomenon (Anderson, 1986, pp. 274–275): (a) Evidentials show the kind of justification for a factual claim which is available to the person making that claim […]; (b) Evidentials are not themselves the main predication of the clause, but are rather a specification added to a factual claim ABOUT SOMETHING ELSE; (c) Evidentials have the indication of evidence […] as their primary meaning, not only as a pragmatic inference; (d) Morphologically, evidentials are inflections, clitics, or other free syntactic elements (not compounds or derivational forms). Commenting on Anderson's four criteria, Aikhenvald (2003, p. 24) and Boye and Harder (2009) criticize some of them, but not for the same reasons: Aikhenvald because 'the distinction between inflectional and derivational categories is not clear-cut' (cf (d)), and Boye and Harder (2009) because they adopt a functional-cognitive approach and give arguments "based on diachronic data exemplifying one of the paths of development that gives rise to evidential elements" (p. 17).

Evidentials can have various origins (Aikhenvald, 2004, pp. 271–302); they may come from the grammaticalization of lexical categories, verbs, deictic and locative markers, or other word classes.

Reportives are very often derived from a verb of 'saying' and are used in dialogue as well as in tales and myths. In Lezgian (North Caucasian), the reportive suffix *-lda* is derived from the construction *luhuda* 'one says' (Haspelmath, 1993, p. 148). In Tibetan, the reportive *-sa*, is derived from the verb *zer* 'say.' In Makah, the suffix *wa:**t*** is a reconstructed passive

form of the verb *wa·* 'say;' (9b) is opposed to (9a) and expresses an indirect message (Jacobsen, 1986, p. 17):

9. a. *xu.bitdib$^{?}$**u***
   'He was snoring'

   b. *xu.bitdibit**wa.d***
   'He was snoring (I was told)'

Perception verbs are a source of visual and non-visual evidentials. In Maricopa, a Yuman language (Vauprés region, Colombia, and Brazil), the visual suffix *'yuu-* comes from the verb *yuu-k* 'see' (Gordon, 1986, p. 85). In Wintu, *-nt^{h}er* is a passive form of the perception verb *mut-* 'hear, feel, sense, and perceive' (Schlishter, 1986, p. 49). From a diachronic point of view, the Tuyuca visual is a 'relic auxiliary verb' from 'seem' or 'be perceived' (Malone 1988, p. 132). In Tariana, the visual *-nuka*, *-naka* is related to the first person singular form of the verb *-ka* 'see,' while *-mha* comes from the verb *-huma* 'hear and feel' (Aikhenvald, 2004, p. 274). The source for nonvisuals may be verbs meaning 'seem, be perceived, and feel,' for example, in East Tucanoan languages (Aikhenvald, 2003).

Verbs referring to existence and location as well as demonstrative pronouns can participate in developing evidentials. In Wintu, the experiential evidential *?el* is probably "a morpheme meaning something like 'it exists'" (Schlichter, 1986, p. 54).

Over the last decade, evidentiality has been defined as a functional-conceptual substance domain (Boye and Harder, 2009), a conceptual domain (Wiemer, 2010), and a semantic-functional domain (Diewald and Smirnova, 2010). As a result, evidentiality has been extended to 'languages traditionally considered as unaffected' by it (Squartini, 2008) and includes a range of devices in order to establish the functional grounds for a cross-linguistic analysis: adverbial expressions such as *reportedly, supposedly*, or *allegedly* in English; mental verbs such as *penser* think,' *trouver* 'think, judge,' *avoir l'impression* 'have the impression' in French (Dendale and Bogaert, 2007) or Eng. *think, suppose, find, claim.*); parentheticals such as Fr. French *dit-on* or *paraît-il);* particles referring to a verbal report (Wiemer, 2010); Fr. *selon*, Eng. *according to*, Russ. *Soglasno*, Bulg. *už, kaj, spored*, Georg. *mixedvit, tanaxmad, cnobit*, and so on. Given the broad variety of linguistic phenomena, it is difficult to define unambiguously the notion of evidentiality. To some extent, however, there is a real problem here: if '[e]vidential meanings range from lexical to grammatical functions' (Diewald and Smirnova, 2010, p. 1), the question is: 'how wide this range can be?'

Recently, (ad)mirative meanings have received a lot of attention and are the subject of growing interest. Such meanings arise when the speaker is confronted with an unexpected discovery and expresses the contrast between the expected situation and what actually happens ('I did not know that S (but now I do)'). The term admirative was first introduced by Dozon (1878) in Albanian (for the history of the term 'admirative,' see Friedman (2012) and references cited there). Admirative usages expressing surprise at a newly discovered fact are well known in Balkan Slavic, Turkish, and other Balkan languages (see also Gerdžikov, 1984 for discussion on their theoretical status in Bulgarian). The notion of mirativity, introduced by DeLancey (1997), and its status as a grammatical category has recently been discussed in the journal *Linguistic Typology* (2012/13).

See also: Inferences in Discourse, Psychology of; Knowledge, Anthropology of; Tense, Aspect, Mood and Evidentiality, Linguistics of.

Bibliography

Aikhenvald, A.Y., 2003. Evidentiality in typological perspective. In: Aikhenvald, A.Y., Dixon, R.M.W. (Eds.), Studies in Evidentiality. Benjamins, Amsterdam, Philadelphia, pp. 1–32.

Aikhenvald, A.Y., 2004. Evidentiality. Oxford University Press, Oxford.

Aikhenvald, A.Y., 2008. Information source and evidentiality: what can we conclude? In: Squartini, M. (Ed.), Evidentiality between Lexicon and Grammar (=Rivista di Linguistica 19- (1)), pp. 209–227.

Aikhenvald, A.Y., Dixon, R.M.W. (Eds.), 2003. Studies in Evidentiality. Benjamins, Amsterdam, Philadelphia.

Anderson, L., 1986. Evidentials, paths of change, and mental maps: typologically regular asymmetries. In: Chafe, W., Nichols, J. (Eds.), Evidentiality: The Linguistic Coding of Epistemology, Advances in Discourse Processes. Ablex, Norwood, NJ, pp. 273–312.

van der Auwera, J., Plungian, V.A., 1998. Modality's semantic map. Linguistic Typology 2, 79–124.

Barnes, J., 1984. Evidentials in the Tuyuca verb. International Journal of American Linguistics 50, 255–271.

Basso, E., 2008. Epistemic deixix in Kalapalo. Pragmatics 18 (2), 215–250.

Boas, F., 1911. Kwakiutl. In: Franz, Boas (Ed.), Handbook of American Indian Languages, Part 1, Smithsonian Institution. Bureau of American Ethnology Bulletin, 40, pp. 423–557.

Boas, F., 1911. Kwakiutl grammar, with the glossary of the suffixes. Transactions of the American Philosophical Society 37, 201–377.

Boye, K., Harder, P., 2009. Evidentiality: linguistic categories and grammaticalization. Functions of Language 16 (1), 9–43.

Bybee, J., Perkins, R., Pagliuca, W., 1994. The Evolution of Grammar: Tense, Aspect and Modality in the Languages of the World. University of Chicago Press, Chicago.

Chafe, W., 1986. Evidentiality in English conversation and academic writing. In: Chafe, W., Nichols, J. (Eds.), Evidentiality: The Linguistic Coding of Epistemology, Advances in Discourse Processes. Ablex, Norwood, NJ, pp. 261–272.

Chafe, W., Nichols, J. (Eds.), 1986. Evidentiality: The Linguistic Coding of Epistemology. Ablex, Norwood, NJ.

Cinque, G., 1999. Adverbs and Functional Heads: A Cross-Linguistic Perspective. Oxford University Press, Oxford.

Comrie, B., 2000. Evidentials: semantics and history. In: Johanson, L., Utas, B. (Eds.), Evidentials: Turkish, Iranian, and Neighbouring Languages. Mouton de Gruyter, Berlin, New York, pp. 1–12.

DeLancey, S., 1997. Mirativity: The grammatical marking of unexpected information. Linguistic Typology 1, pp. 33–52.

Dendale, P., Tasmowski, L. (Eds.), 2001. On Evidentiality, Journal of Pragmatics 33.3, 339–348.

Dendale, P., Van Bogaert, J., 2007. A semantic description of French lexical evidential markers and the classification of evidentials. In: Squartini, M. (Ed.), Evidentiality between Lexicon and Grammar (=Rivista di Linguistica 19- (1)), pp. 65–89.

Desclés, J.-P., 2000. Abduction and non-observability – some examples from language science and the cognitive science. In: Agazzi, E., Pauri, M. (Eds.), The Reality of the Unobservable. Observability, Unobservable and Their Impact on the Issue of Scientific Realism. Kluwer Academic Publishers, Dordrecht, pp. 87–122.

Desclés, J.-P., Guentchéva, Z., 2000. Enonciateur, locuteur, médiateur. In: Monod-Becquelin, A., Erikson, P. (Eds.), Les Rituels du dialogue: Promenades ethnolinguistiques en terres amérindiennes, Société d'ethnologie (Recherches thématiques 6), pp. 79–112. Nanterre.

Diewald, G., Smirnova, E., 2010. Linguistic Realization of Evidentiality in European Languages (Empirical Approaches to Language Typology 49). De Gruyter Mouton, Berlin.

Dozon, A., 1878. Manuel de la langue chkipe ou albanaise. Leroux, Paris.

Faller, M.T., 2002. Semantics and Pragmatics of Evidentials in Cuzco Quechua. Dissertation. Stanford University, Stanford.

Friedman, V., 1986. Evidentiality in the balkans: macedonian, bulgarian, and albanian. In: Chafe, W., Nichols, J. (Eds.), Evidentiality: The Linguistic Coding of Epistemology, Advances in Discourse Processes. Ablex, Norwood, NJ, pp. 168–187.

Friedman, V., 1999. Evidentiality in the Balkans. In: Hinrichs, U. (Ed.), Handbuch der Südosteuropa-Linguistik. Harrassowitz, Wiesbaden, pp. 519–544.

Friedman, V., 2000. Confirmartive/nonconfirmative in Balkan Slavic, Balkan Romance, and Albanian with additional observations on Turkish, Romani, Georgian, and Lak, Macedonian, Bulgarian, and Albanian. In: Johanson, L., Utas, B. (Eds.), Evidentials: Turkish, Iranian, and Neighbouring Languages. Mouton de Gruyter, Berlin, New York, pp. 329–366.

Friedman, V., 2012. Perphaps mirativity is phlogiston, but admirativity is perfect: on Balkan evidential strategy. Linguistic Typology 3, 505–527.

Gerdžikov, G., 1984. Преизказването на глаголното действие в българския езикѣ наука и изкуство. София.

Gordon, L., 1986. The development of evidentials in Maricopa. In: Chafe, W., Nichols, J. (Eds.), Evidentiality: The Linguistic Coding of Epistemology, Advances in Discourse Processes. Ablex, Norwood, NJ, pp. 75–88.

Guentchéva, Z. (Ed.), 1996a. L'énonciation Médiatiseée. Peeters, Louvain, Belgium.

Guentchéva, Z., 1996b. Le médiatif en bulgare. In: Guentchéva, Z. (Ed.), L'énonciation Médiatiseée. Peeters, Louvain, Belgium, pp. 47–70.

Guentchéva, Z., Landaburu, J. (Eds.), 2007. L'Enonciation médiatisée II: Le traitement épistémologique de l'information: Illustrations amérindiennes et caucasiennes. Peeters, Paris.

Haarmann, H., 1970. Das Indirekte Erlebnisform Als Grammatische Kategorie. Eine Eurasische Isoglosse. Harrassowitz, Wiesbaden.

Hardman, M.J., 1986. Data-source marking in the Jaqi languages. In: Chafe, W., Nichols, J. (Eds.), Evidentiality: The Linguistic Coding of Epistemology, Advances in Discourse Processes. Ablex, Norwood, NJ, pp. 113–136.

Haspelmath, M., 1993. A Grammar of Lezgian. Mouton de Gruyter, Berlin.

de Haan, F., 1999. Evidentiality and epistemic modality: setting boundaries. Southwest Journal of Linguistics 18, 38–101.

de Haan, F., 2001. The relation between modality and evidentiality. In: Müller, R., Reis, M. (Eds.), Modalität und Modalverben im Deutschen. Linguistische Berichte, Sonderheft 9. H. Buske, Hamburg.

de Haan, F., 2005. Encoding speaker perspective: evidentials. Studies in Language Companion Series. In: Frajzyngier, A. (Ed.), Linguistic Diversity and Language Theories. Benjamins, Amsterdam, pp. 379–397.

Jacobsen, W.H., 1986. The heterogeneity of evidentials in Makah. In: Chafe, W., Nichols, J. (Eds.), Evidentiality: The Linguistic Coding of Epistemology, Advances in Discourse Processes. Ablex, Norwood, NJ, pp. 3–23.

Johanson, L., 2000. Turkic indirectives. In: Johanson, L., Utas, B. (Eds.), Evidentials: Turkish, Iranian, and Neighbouring Languages. Mouton de Gruyter, Berlin, New York, pp. 61–87.

Johanson, L., 2003. Evidentiality in Turkic. In: Aikhenvald, A.Y., Dixon, R.M.W. (Eds.), Studies in Evidentiality. John Benjamins, Ampsterdam, pp. 273–291.

Johanson, L., Utas, B., 2000. Evidentials: Turkish, Iranian, and Neighbouring Languages. Mouton de Gruyter, Berlin, New York.

Joseph, B., 2003. Evidentials. Sommation, questions, prospects. In: Aikhenvald, A.Y., Dixon, R.M.W. (Eds.), Studies in Evidentiality. Benjamins, Amsterdam, Philadelphia, pp. 307–327.

Landaburu, J., 2007. La modalisation du savoir en langue andoke (Amazonie colombienne). In: Guentchéva, Z., Landaburu, J. (Eds.), L'énonciation médiatisée II. Le traitement épistémologique de l'information: illustrations amérindiennes et caucasiennes. Peeters, Louvain, Belgium, pp. 23–27.

Lazard, G., 1996. In: Guentchéva, Z. (Ed.), Le médiatif en persan, pp. 21–46.

Lazard, G., 2000. Le médiatif : considérations théoriques et application à l'iranien. In: Johanson, L., Utas, Bo (Eds.), Evidentials: Turkish, Iranian, and Neighbouring Languages. Mouton de Gruyter, Berlin, New York, pp. 209–228.

Lazard, G., 2001. On the grammaticalization of evidentiality. Journal of Pragmatics 33, 359–367.

Mithun, M., 1986. Evidential diachrony in northern Iroquoian. In: Chafe, W., Nichols, J. (Eds.), Evidentiality: The Linguistic Coding of Epistemology, Advances in Discourse Processes. Ablex, Norwood, NJ, pp. 89–112.

Malone, T., 1988. The origin and development of Tuyuca evidential. International Journal of American Linguistics 54, 119–140.

Oswald, R., 1986. The evidential system of kashaya. In: Chafe, W., Nichols, J. (Eds.), Evidentiality: The Linguistic Coding of Epistemology, Advances in Discourse Processes. Ablex, Norwood, NJ, pp. 29–45.

Palmer, R.F., 1986. Mood and Modality. Cambridge University Press, Cambridge, UK.

Peirce, C.S., 1974. Collected Papers, vols. I–VI. Belknap Press, Cambridge, MA.

Plungian, V.A., 2001. The place of evidentiality within the universal grammatical space. Journal of Pragmatics 33, 349–358.

Plungian, V.A., 2010. Types of verbal evidentiality marking: an overview. In: Diewald, G., Smirnova, E. (Eds.), Linguistic Realization of Evidentiality in European Languages (Empirical Approaches to Language Typology 49). De Gruyter Mouton, Berlin, pp. 15–58.

Schlichter, A., 1986. The origins and deictic nature of Wintu evidentials. In: Chafe, W., Nichols, J. (Eds.), Evidentiality: The Linguistic Coding of Epistemology, Advances in Discourse Processes. Ablex, Norwood, NJ, pp. 46–60.

Squartini, M. (Ed.), 2007. Evidentiality between lexicon and grammar (=Rivista di Linguistica 19- (1)).

Tournadre, N., 1996. Comparaison des systèmes médiatifs de quatre dialectes tibétains (tibétain central, ladakhi, dzongkha et amdo). In: Guentchéva, Z. (Ed.), L'énonciation Médiatiseée. Peeters, Louvain, Belgium, pp. 195–213.

Weber, D.J., 1986. Information, perspective, profile, and patterns in Quechua. In: Chafe, W., Nichols, J. (Eds.), Evidentiality: The Linguistic Coding of Epistemology, Advances in Discourse Processes. Ablex, Norwood, NJ, pp. 137–155.

Wiemer, B., 2007. Lexical markers of evidentiality in Lithuanian. In: Squartini, M. (Ed.), Evidentiality between Lexicon and Grammar. Rivista di Linguistica 19 (1), pp. 173–208.

Wiemer, M., 2010. Hearsay in European languages: toward an integrative account of grammatical and lexical marking. In: Diewald, G., Smirnova, E. (Eds.), Linguistic Realization of Evidentiality in European Languages, Empirical Approaches to Language Typology, 49. De Gruyter Mouton, Berlin, pp. 59–129.

Willett, T., 1988. Across-linguistic survey of the grammatization of evidentiality. Studies in Language 12, 51–97.

Evolution and Language: Cultural Transmission

Bill Thompson and Kenny Smith, The University of Edinburgh, Edinburgh, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Language exhibits a variety of structural properties that require evolutionary explanation. Many of these properties may result from processes of cultural evolution in populations of language learners. We review several key examples of how cultural transmission can lead to the emergence of linguistic structure, and outline the implications for evolutionary accounts of the origins of language.

Human language exhibits two remarkable and apparently contradictory properties. On the one hand, language is unique in the natural world, it is universal among our species, it displays striking complexity, and it has, by facilitating the development of complex social and technological systems, enabled humans to spread to all corners of the planet, and even out into space. From this perspective, language looks like a biological adaptation, which evolved in humans (and only humans) as a result of the fitness payoff associated with communication and information sharing. On the other hand, language is acquired incrementally through learning, it exhibits tremendous cross-cultural variation, and this variation poses no problem for the acquirer who could, as far as we can tell, master with equal confidence any existing language just by dint of exposure from birth. These properties might equally lead us to believe that language is a cultural trait, transmitted through social learning.

We can begin to reconcile these intuitions by acknowledging that language is the product of an interaction between two evolutionary systems: biological and cultural evolution. Recently, several independent findings have lent support to the view, which we favor here, that many of the defining characteristics of language can be explained largely as products of cultural evolution, with biological evolution having less to say about the specific structural properties of language. Debate has been reignited on whether there are truly any universal features of language, or whether these universals (typically taken as indicative of a species-wide 'language instinct') may be far fewer than has previously been supposed (Evans and Levinson, 2009). New techniques for reconstructing language histories (*see* Evolution and Language: Phylogenetic Analyses) suggest that, while large groups of languages do share features, these shared bundles of properties tend to pattern like products of shared cultural history, rather than biologically dictated behavioral universals (Dunn et al., 2011). In parallel, the biological basis for language has been questioned on evolutionary grounds (Chater et al., 2009), and on the basis of new evidence for plasticity in the functions of brain regions (Bedny et al., 2011). This shift in thinking has drawn attention to the need for a comprehensive understanding of how cultural process might give rise to the many observable properties of language, and, in turn, how these cultural processes might interact with biological evolution to shape the evolution of language and its cognitive underpinnings. A considerable body of research, involving (among others) linguists, biologists, psychologists, cognitive scientists, and computer scientists, is now dedicated to exploring the workings of the cultural evolution of language. Among the most basic questions underpinning this research agenda is that of the *problem of linkage*: namely, how do properties of individual learners give rise to population-level phenomena that we see in language? In simple terms: what kind of mechanism would allow individual language learners and users to unconsciously coordinate and produce a richly structured system of conventions shared throughout an entire population?

Cultural transmission is just such a mechanism: languages are transmitted by iterated learning, which is one type of cultural transmission, whereby learners acquire a behavior by observing similar behaviors in others who acquired those behaviors in the same way (Kirby et al., 2008). In the following two sections, we will present evidence, obtained through computational, mathematical, and experimental models of iterated learning, that processes of learning and use can lead to the emergence of linguistic systems that exhibit some of the fundamental features of human language. In the final section, we will turn again to biology, and discuss how this improved understanding of the potential role of cultural transmission and cultural evolution in shaping language might inform our understanding of how the human capacity for language evolved.

Language Transmission and the Evolution of Linguistic Structure

Hockett (1961) famously identified a set of design features thought to characterize the fundamental properties of language. Among these is the fact that language is *traditionally transmitted*, or passed from generation to generation through the mechanisms of vertical and oblique cultural transmission. (The distinction between vertical and oblique transmission rests on whether transmission is from parents to offspring (vertical), or from nonparental individuals (oblique). For simplicity, we will collapse this distinction, and refer to intergenerational transmission as vertical.) Vertical cultural transmission is analogous to genetic transmission in biology, and describes the transfer across generations of heritable traits: analogies between the cultural evolution of language and the biological evolution of species have been debated

since Darwin. While the study of language change and phylogenetic relationships between languages has been one of the longest studied subfields of linguistics (Grimm, 1819), only recently has a research program emerged to explore whether processes of cultural evolution arising from traditional transmission might be responsible for some of the other design features of language that Hockett identified. Perhaps the most influential methodology assumed in this framework is the use of computer simulations and mathematical models. Similar techniques have a long history of productive application in analogous topics of biology, and have been successfully modified for application to general models of cultural transmission in foundational research into social learning.

The evolution of language is a special case of cultural evolution due to several special properties that characterize its form and its transmission. In large part, these special properties stem from the fact that language is acquired through a procedure of inductive inference far more demanding than that required by other instances of social learning. Models of iterated learning explore the consequences of this kind of cultural transmission. Central to this approach is the concept of *learnability* (Brighton et al., 2005a,b). In principle, languages can vary with respect to the fidelity with which they are transmitted: some types of language might be more learnable than others. As a result of that variation, through cultural transmission some types of language may come to be more common than others, due to differential retention of different language types (or, more likely, different features within a single language). Computational and mathematical modeling allows the researcher to formally define the properties of such a system, and to study its behavior systematically. This approach has shown how the cultural transmission of language can explain the existence of several of the structural properties fundamental to language. One such structural property is *compositionality*. A communication system is compositional if, for a given complex signal, the meaning of that signal is a function of the meanings associated with the component parts of the signal, and of some set of rules that determine how those component parts can be combined. This is true of all human languages: put simply, in language, the meanings of a sentence are composed from the meanings of the morphemes, words, or phrases that make up that sentence. These component parts are put together in accordance with systematic rules that govern the patterns of legal combination and the meanings of those combinations. Taken together, these properties form the basis of syntactic structure in language. Compositionality is a universal, definitional feature of human language, but it is not present (as far as we currently know) to any great extent in the communication system of any other animal (The waggle dance of bees (von Frisch, 1974) and the alarm-calling system of Campbell's monkeys (Zuberbuhler, 2002) show that some minimal, tightly constrained form of compositionality does occur in animal communication, but there is a clear difference between these systems and the massively productive use of compositionality in human language.). If any property of language were to seem straightforwardly and plausibly attributable to a biologically evolved faculty of language in humans, it would be compositionality, since it is compositionality that gives language its open-ended expressive power. However, models of iterated learning have shown that there exists a preferable alternative explanation for this linguistic universal: compositionality can emerge as a result of cultural transmission, without this specific property of language needing to be hardwired into language learners.

This explanation relies on an age-old philosophical observation concerning the nature of learned behavior, traditionally treated as evidence for the necessity of innate knowledge, known in contemporary terms as *the poverty of the stimulus*. The poverty of the stimulus describes the idea that learners of a language, among other challenges they face, are exposed to only a subset of the language they are attempting to learn. While this argument has taken various forms, many controversial, there is at its core an incontestable truth: language is open ended and productive, encoding in grammar the capacity to generate a potentially infinite number of legal sentences, yet the learner's exposure is necessarily finite, not least because of a finite lifespan, but also because language acquisition is largely complete in the first decade of that life. In other words, you cannot see a language in its entirety before you learn it, because all languages are infinitely large. While this transmission *bottleneck* is sometimes presented as a logical argument demonstrating the importance of innate knowledge in language-learning, the cultural evolution perspective treats the poverty of the stimulus as a *selective pressure* facing the evolving language: the poverty of the stimulus presents a pressure for languages to be reconstructable on the basis of incomplete data. Languages, or parts of languages, which evidence their full structure in a greater proportion of their utterances will be more faithfully transmitted than languages whose utterances are less representative of the whole. One way a language can achieve this is by employing structural regularity in the mappings between meanings and signals.

Consider three possible kinds of language that provide different ways of mapping meanings to signals. A maximally *holistic* language employs a unique, undecomposable signal for every possible meaning. A maximally *underspecified* language would employ the same signal for every possible meaning. An ideal compositional language, unlike a fully underspecified language, provides a distinct signal for every meaning: however, unlike a holistic language, those signals are constructed via a rule-governed, compositional process of the sort described in this article, where the structure of the meaning is reflected in the signal. Figure 1 contrasts examples of these different kinds of mappings. What challenges would these systems pose to a learner?

To acquire the full holistic system, a learner must observe *all* meaning–signal mappings: no generalizations can be drawn from exposure to a subset, since by definition the holistic language offers no generalizations about how meaning and signal are related. Such a language would therefore be unlearnable from a subset: while you might be able to reproduce the (infinitely small) fraction of the language you witnessed during your lifetime, you would not be able to produce or interpret utterances you had not already encountered. In contrast, to acquire a fully underspecified system, the learner need only observe one pairing of meaning and signal: from a single observation, the learner can make accurate generalizations about the rest of the system. Acquiring a fully underspecified language in the face of a bottleneck on transmission is therefore

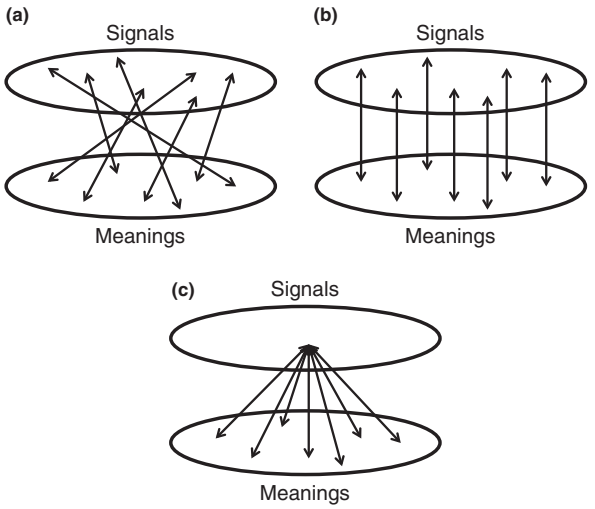

Figure 1 Three ways to map between meanings and signals. (a) Holistic mapping: little correlation between the forms of meaning and signals. Neighboring meanings do not necessarily map to neighboring signals; (b) a compositional mapping between meanings and signals. Here, there is a correlation between the form of a meaning and the form of a signal, such that exposure to a subset of these mappings will allow the learner to induce a strong generalization; (c) a maximally underspecified mapping: all meanings map to the same signal.

easy: to put it another way, underspecification is adaptive in the face of a transmission bottleneck. Various computational simulations of the iterated learning process show that, purely as a result of cultural transmission, underspecified languages will evolve: the most stable, learnable, transmissible system is a fully underspecified system (Brighton et al., 2005a,b). All that is required for this to happen is that learners are able to make inductive leaps: to go beyond that which is exemplified in linguistic experience and draw piecemeal generalizations not directly evidenced in their data and not previously present in the linguistic community. In its most general conceptualization, this amounts to a bias toward *compressed representations*: a preference for simple grammars, exploiting those generalizations that the data afford. Iterated learning with a bottleneck inherently favors generalization, but can actually lead to generalization only if learners are capable of drawing such generalizations, or compressing information. These conditions alone lead to the cultural evolution of maximally learnable, maximally underspecified languages: generalizations (of the form "several meanings can be expressed using the same signal") emerge and spread, until a single generalization ("*all* meanings are expressed using *this* signal") takes hold.

However, language is not maximally underspecified; it is compositional. From the perspective of the learner, compositional languages lie somewhere between these two extremes. Unlike fully underspecified languages, they require a substantial body of data to learn: you need to see enough data to figure out the rules and regularities that relate signal to meaning. However, unlike holistic languages, you do not need to see the full system provided you get enough data to work out the compositional system for producing and interpreting signals, you can reconstruct the whole language, or produce or interpret signals you never actually encountered during learning. Like maximally underspecified languages, compositional languages require learners who are capable of making the generalizations that data from a compositional language suggest. However, if compositional languages are to evolve through transmission, an additional bias must be in place: the evolution of compositional language relies on the presence of some bias against many-to-one meaning–signal mappings. That is, alongside the pressure for learnability, there must exist in the system some pressure against underspecification. A learning bias against many-to-one mappings (i.e., an antiambiguity bias) is a simple way to implement this pressure in simulations, and this kind of bias is empirically motivated by psycholinguistic research (e.g., see Brighton et al., 2005a,b for review; also, this article further discusses the use of language for communication as another pressure against underspecification). In summary, vertical cultural transmission in populations of learners who are biased to form compressed representations of their experiences, and to avoid underspecification, leads eventually, in the absence of any biological change, to the evolution of languages governed by (rudimentary) syntactic structure, where before there was none. On the grounds of parsimony alone, an explanation of this kind seems preferable to assuming that compositionality is a reflection of a hardwired constraint in learners.

Current theoretical work in this tradition is grounded on two major agendas: to explore whether other fundamental properties of language can also be explained as products of cultural evolution, and to search for general principles of language evolution through cultural transmission. Pursuing the former agenda, comparable accounts, similarly based on biased cultural transmission, have been offered for various other aspects of linguistic structure: most research so far has been concerned with the origins of syntactic phenomena, such as embedding and recursion (Batali, 2002; Kirby, 2002), constituent-order patterns (Kirby, 1998), and morphosyntactic regularity and irregularity (Kirby, 2001), although some researchers in the iterated learning model (ILM) tradition have also studied the cultural evolution of phonological (Boer, 2001) and semantic systems (how sound systems or word meanings evolve) (Perfors and Navarro, 2011).

Research focused on the more general principles of the cultural evolution of language aims to abstract away from specific instances and define the foundations of cultural transmission through inductive inference. Current work is mostly situated in the rational tradition and based on the principles of Bayesian learning (Griffiths and Kalish, 2007). Bayesian models of iterated learning have various merits: principally, this approach facilitates a high degree of abstraction, allows precise treatment of the role of bias in individual learning, and draws useful connections between models of language evolution and computational techniques heavily studied in the Artificial Intelligence and Machine Learning literature. A major objective of Bayesian ILMs is to establish the abstract link between properties of individual learners and population-level phenomena: do individual learning biases, such as those discussed in this article, straightforwardly determine the kinds of languages we see in human populations? If the answer to this question is yes, then typological universals, or in fact behavioral universals of many kinds, can act as a reliable map of the structure of

the mind. If the answer is no, and cultural transmission can amplify or wash out the effects of learning biases, such that the relationship between biases of individuals and universals is complex, then it may be a mistake to infer cognitive universals from typological generalizations.

As a first approximation, there is a consensus (supported by classic mathematical analyses of cultural evolution) that cultural evolution amplifies individual biases, introducing opacity in the relationship between properties of individuals and properties of cultural systems: put simply, just because we see a strong skew in the distribution of some cultural trait (say, a linguistic trait) across human populations, or even an apparently universal trait, we cannot conclude that, that reflects a strong or absolute constraint in individuals: it could be a consequence of a far weaker bias in individuals, amplified over many generations of transmission (Boyd and Richerson, 1985; Kirby et al., 2007). A significant finding to emerge from Bayesian analysis of Iterated learning is that under certain conditions, this general tendency may not hold: for a particular class of psychologically plausible learning algorithms, Iterated learning converges to a state entirely determined by the biases of individual learners (Griffiths and Kalish, 2007). This result demonstrates that relatively minor differences in our assumptions about the kind of learning mechanisms that learners employ can lead to dramatically different evolutionary hypotheses. It is not yet clear which pattern of results better characterizes the transmission and evolution of language. Much current work is motivated by this problem: aiming to determine systematically which conditions give rise to which pattern of linkage, by exploring alternative models of individual learning and by manipulating aspects of the transmission process such as population structure and transmission dynamics.

Alongside this approach has emerged another new methodology, intended to bypass the problems involved with making assumptions about the nature of learning biases: language evolution experiments with human participants. Language evolution experiments in the laboratory combine artificial language-learning techniques (Gomez and Gerken, 2000) with diffusion chain experimental techniques (Mesoudi and Whiten, 2008; Whiten and Mesoudi, 2008) to replicate and manipulate the process of iterated learning. Artificial language-learning experiments involve training participants on an experimenter-constructed miniature language and subsequently testing their knowledge of that language, often by eliciting productions. A diffusion chain is a simple laboratory model of cultural transmission, where the productions of one participant constitute the training input to the next participant, creating a chain of learners through which information is diffused. Language evolution experiments in the laboratory are in their early stages, although some fundamental findings have already emerged, and some important parallels between models and experiments are worth highlighting.

One early series of experiments replicates closely the simulation models of the emergence of compositionality discussed in this article (Kirby et al., 2008). In these experiments, participants are trained on pairs of associations between orthographic strings (written words of an artificial language) and sets of meanings, or objects, whose features vary along several dimensions (color, shape, and motion: e.g., a spinning red square or a bouncing blue triangle). The first participant in each diffusion chain is trained on mappings generated at random: each object is paired with a random word, so no regularities exist in the mappings (i.e., the language is holistic). After a short period of testing, the participant's knowledge of the language is tested by eliciting productions, whereby the participant is prompted to produce a string for an object. Some of these test trials will involve objects encountered during training paired with a label, but the test set also includes novel objects: participants may be prompted to produce labels for objects they have never seen before. In other words, as in the models, there is a bottleneck on transmission, forcing learners to attempt to generalize from the parts of the language they encountered during training. These productions then form the data with which the next participant is trained. This procedure is repeated, and the properties of the artificial language are tracked down the chain. As in the theoretical models, the pressure introduced by the bottleneck (and also participants' fallible memories; Cornish, 2010) creates an advantage for generalizable languages that can be more comprehensively reconstructed from exposure to a subset: the more a learner is able to make generalizations, the less likely it will be that novel items during testing will disrupt the transmission of the language to new learners. Experimental findings confirm that under these conditions, a culturally transmitted language evolves to become highly learnable: after only a few generations, the artificial language is largely underspecified,

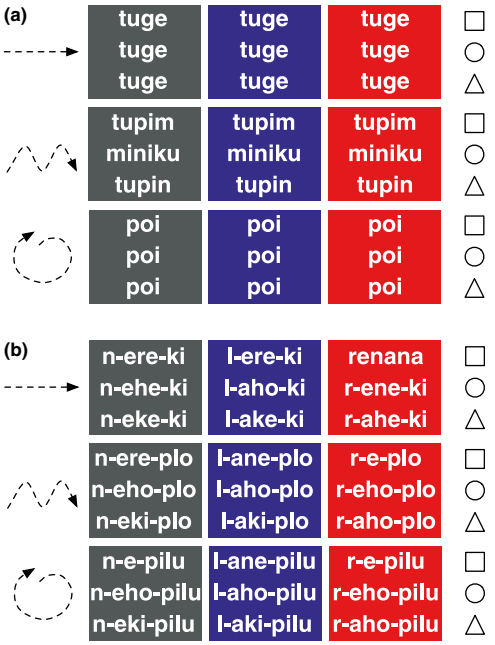

Figure 2 Experimentally evolved artificial languages. (a) An evolved language with a systematically underspecified mapping from meanings to signals; (b) an evolved compositional language. Reproduced from Kirby, S., Cornish, H., Smith, K., 2008. Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. Proceedings of the National Academy of Sciences of the United States of America 105 (31), 10681–10686.

employing a small repertoire of homonyms to cover a much larger space of possible meanings (see Figure 2(a)).

In a subsequent experiment, the data were manipulated as they were passed from participant to participant, such that any ambiguous signals in a participant's productions were removed by the experimenter before those productions were fed to the next learner in the chain. The motivation for this interference is twofold: first, equivalent theoretical models predict that to prevent an evolving language from collapsing to underspecification, some bias against homonymy is necessary, as discussed in this article. Since, at least in this task, no such bias appeared to be expressed by participants (or, rather, any such bias is very weak given the contracted training period and the strong requirement for generalization introduced by the transmission bottleneck), the experimental manipulation imposes an equivalent influence. Second, this manipulation can be taken as a proxy for the effects of language *use*. Presumably, in the real world (but not in this simple experiment), the requirement of language users that their language can be used to communicate imposes a strong pressure for expressivity in language: maximally underspecified languages are maximally useless for communication. Again, as predicted by the models, under this regime, the artificial languages develop a compositional structure (see Figure 2(b)): competing pressures for learnability and expressivity conspire to produce systematically structured compositional languages. The pressure for learnability is evidently due in large part to systemic properties of vertical cultural transmission. The pressure inhibiting underspecification, however, may be driven by biases of learners (as assumed in some models; e.g., Smith, 2004), or it may have its roots in another process: intragenerational, or *horizontal*, cultural transmission.

Language Use, Negotiation, and the Formation of Conventions

Broadly construed, language is a fundamentally social phenomenon: we use it to inform, amuse, impress, coerce, or cajole other people. Unsurprisingly, some of language's key features have received explanations that appeal to dynamic social interaction, in a manner comparable to the literature reviewed in this article explaining compositionality in terms of language-learning and transmission. A well-studied example, which is general enough to have implications for many aspects of language structure, is the emergence of *convention*. Communication relies on convention at all levels of language structure: in order to be communicatively useful, linguistic conventions must be shared by an entire linguistic community. How do conventions, particularly linguistic conventions such as a shared name for an object or a shared grammatical device for signaling some more abstract concept, arise in a population? This is a question about the nature of cultural transmission, since the processes by which individuals influence each other as they interact are questions about culture.

As with vertical cultural transmission, computational simulations and mathematical analyses of conventional alignment have produced useful insights. In principle, shared conventions can emerge from nothing and propagate rapidly (i.e., within a single generation) (Steels, 2011). That is, a population of learners who share no system of conventions can easily negotiate and maintain such a system culturally. The speed and accuracy with which populations converge, and the stability of an emerging system, depend crucially on the particular cognitive abilities assumed of learners. Simple capacities for memory and invention are sufficient to allow populations to maintain shared conventions, although these end up containing much interindividual variation, and are therefore communicatively suboptimal. Slightly more complex assumptions about how learners deal with competing conventions, known as strategies for *lateral inhibition* (equivalent to the biases against many-to-one mappings discussed in the Language Transmission and the Evolution of Linguistic Structure section), are necessary for optimally functioning conventional systems to emerge (Steels, 2011). Nevertheless, convention negotiation as a phenomenon may well underpin many aspects of linguistic structure: topics studied in this framework range from the lexicon, to syntactic constructions, to phonemic inventories, and to even more fine-grained systems such as case marking and paradigms for marking tense. We now have a clear understanding of how this kind of agreement can emerge culturally in populations, supported by evidence from agent-based computational simulations, mathematical analyses, and even robotic experiments (Steels, 2011).

Again, alongside theoretical and computational analyses, there exists a body of complementary experimental research on human participants. One strand of research in this area explores the emergence of another fundamental design feature of language: arbitrariness. At least to a first approximation, linguistic conventions are arbitrary. This is clearest in the case of words, which do not seem to resemble the things (objects, actions, or ideas) to which they refer: for instance, the English word 'tree' shares no obvious properties with the plants to which it refers (it is not tall, or green, and it does not sound like a tree creaking in the wind or irresistibly evoke the smell of crushed pine needles). We can imagine a language that is nonarbitrary: indeed, there might be advantages to such languages. For instance, a system of *iconic* signs (e.g., words or gestures that sound or look like the thing they refer to) would allow naïve users to straightforwardly understand and even produce novel signs.

A recent series of psychological experiments shows how conventionalized signs can emerge, and how from iconic beginnings those signs gradually become less and less iconic as a result of their transmission and use (Fay et al., 2010; Garrod et al., 2007). In the classic version of these experiments (Garrod et al., 2007), pairs of participants engage in a graphical communication task similar to the popular game Pictionary: they use drawings (i.e., not their preexisting symbolic language) to communicate about concepts or objects taken from a prespecified list, playing many such games repeatedly with a single partner. Initially, the communicative drawings tend to be highly iconic (e.g., if attempting to convey 'computer monitor,' you might do your best to produce a good drawing of a computer monitor, base unit, keyboard, etc.: see Figure 3). However, when pairs interact repeatedly (i.e., they try to communicate the same concepts several times), the emergent signs tend to become abstract and

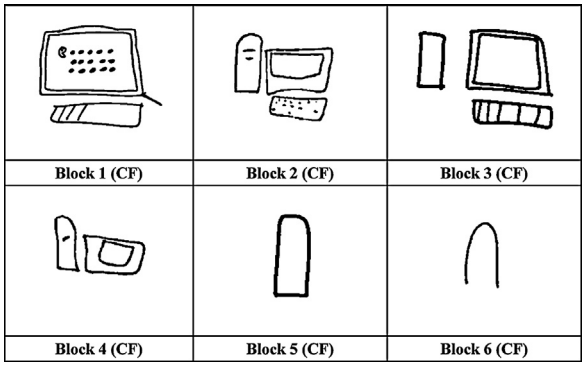

Figure 3 The evolution of a signaling system. Depictions begin highly iconic and over interactions approach arbitrariness. Reproduced from Garrod, S., Fay, N., Lee, J., Oberlander, J., MacLeod, T., 2007. Foundations of representation: where might graphical symbol systems come from? Cognitive Science 31 (6), 961–987.

arbitrary: rather than drawing the computer monitor, you might produce just an inverted U shape (see Figure 3) and be confident your partner knows what you mean. Because pairs have a shared cultural history and memory, their signing systems need not iconically depict their referents: each member of the pair knows the evolutionary history of the sign (e.g., that the inverted U resembles the base unit that you drew when you first produced an iconic drawing for 'computer monitor'), and so the signs themselves are free to become less iconic and more (seemingly) arbitrary. Iconicity in the system gradually deteriorates, in favor of simpler, more efficient conventions that nevertheless pick out specific objects or concepts. Onlookers, who try to identify the depicted object using these abstract signs without being exposed to the shared interaction history, are at chance, although a follow-up study (Caldwell and Smith, 2012) shows that witnessing these signs being used communicatively is sufficient to allow naïve individuals to adopt them as purely arbitrary signaling conventions, without access to the historical process by which those particular signs became established. Interaction and transmission are the keys: a simple process of social negotiation and cultural evolution can explain how an initially iconic signaling system can quickly become one of arbitrary, symbolic conventions. Similar phenomena are documented by fieldwork exploring emerging sign languages, although this too is a young enterprise (Sandler et al., 2005).

Coevolution: A Brief Word

The examples discussed so far are not exhaustive – other aspects of language structure have received explanations that draw heavily on cultural processes – and of course much remains to be done: we are far from offering a complete explanation of language in cultural terms. However, it is clear that there exist viable cultural explanations for the evolution of many of the most fundamental design features of language, including compositional syntax, word-order generalizations, morphological regularities, recursion, phonological regularities, shared expressive conventions, and arbitrary reference. The findings of this research enterprise raise profound questions about the role of biological evolution. If many important features of language could have arisen through cultural processes, what is left for biology to explain?

While it is important to draw a marked distinction between the biological evolution of the language faculty and the cultural evolution of languages themselves, it is a mistake to treat biological and cultural processes as independent: neither acts in isolation, and both feed into each other, acting together to create a coevolutionary system. To be clear, we cannot entertain cultural explanations for language without appealing to the mechanisms of learning and the inductive biases that underpin them (although the link between the bias and the details of the linguistic structure may be obscure). And bias is a topic for biology, where we isolate the need for cognitive bias, we identify a contribution of biology, and of the selective forces that maintain useful traits. Similarly, at least from some perspectives (Scott-Phillips et al., 2011; but see Laland et al., 2011 for a response), any appeal to biological adaptation in fact subsumes the cultural approach: natural selection pervasively exploits environmental regularities that play a part in boosting an organism's fitness. In principle, cultural processes are just one more environmental regularity that biological evolution can commandeer: where cultural processes do the work, we should expect this to be reflected in biology.

At present, very little is known about how these two processes interact to shape language. Preliminary research (Smith, 2004; Smith and Kirby, 2008) suggests that these interactions are nontrivial: opaque mappings between individual biases and population-level cultural phenomena may significantly constrain the kinds of adaptations that can arise through natural selection to support language and other cultural systems. If these indications are accurate, theories of linguistic nativism (a class of theoretical perspectives that hypothesize innate knowledge of linguistic structure) may have to be reconsidered on evolutionary grounds: rather than supporting the existence of a language faculty built around rich innate knowledge of linguistic structure, evolutionary reasoning might instead suggest biological underpinnings of language that are composed of the preadaptations necessary for the processes of cultural evolution.

Summary

Here, we have outlined the perspective that cultural evolution is a major force driving the emergence of structure in human language. Cultural evolution results from various kinds of cultural transmission: learning and use are two forms of transmission in particular that influence the nature of language. Many aspects of linguistic structure have received explanations that rely heavily on these processes, and many more remain to be explored in this way. This perspective is supported by evidence from computational, mathematical, theoretical, and experimental sources, which continue to inform our understanding of the origins and evolution of language, and to raise important questions concerning the evolutionary processes that shape human behavior.

See also: Evolution and Language: Phylogenetic Analyses.

Bibliography

Batali, J., 2002. The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In: Briscoe, E. (Ed.), Linguistic Evolution through Language Acquisition: Formal and Computational Models. Cambridge University Press, Cambridge, pp. 111–172.

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E., Saxe, R., 2011. Language processing in the occipital cortex of congenitally blind adults. Proceedings of the National Academy of Sciences of the United States of America 108 (11), 4429.

Boer, B., 2001. The Origins of Vowel Systems. Studies in the Evolution of Language. Oxford University Press, Oxford.

Boyd, R., Richerson, P.J., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago, IL.

Brighton, H., Kirby, S., Smith, K., 2005a. Cultural selection for learnability: three principles underlying the view that language adapts to be learnable. In: Tallerman, Maggie (Ed.), Language Origins: Perspectives on Evolution. Oxford University Press, Oxford, pp. 291–309.

Brighton, H., Smith, K., Kirby, S., 2005b. Language as an evolutionary system. Physics of Life Reviews 2 (3), 177–226.

Caldwell, C.A., Smith, K., 2012. Cultural evolution and perpetuation of arbitrary communicative conventions in experimental microsocieties. PLoS One 7 (8), e43807.

Chater, N., Reali, F., Christiansen, M.H., 2009. Restrictions on biological adaptation in language evolution. Proceedings of the National Academy of Sciences of the United States of America 27, 1015–1020.

Cornish, H., 2010. Investigating how cultural transmission leads to the appearance of design without a designer in human communication systems. Interaction Studies 11 (1), 112–137.

Dunn, M., Greenhill, S.J., Levinson, S.C., Gray, R.D., 2011. Evolved structure of language shows lineage-specific trends in word-order universals. Nature 473 (7345), 79–82.

Evans, N., Levinson, S.C., 2009. The myth of language universals: language diversity and its importance for cognitive science. Behavioral and Brain Sciences 32 (5), 429–448.

Fay, N., Garrod, S., Roberts, L., Swoboda, N., 2010. The interactive evolution of human communication systems. Cognitive Science 34 (3), 351–386.

Garrod, S., Fay, N., Lee, J., Oberlander, J., MacLeod, T., 2007. Foundations of representation: where might graphical symbol systems come from? Cognitive Science 31 (6), 961–987.

Gomez, R.L., Gerken, L., 2000. Infant artificial language learning and language acquisition. Trends in Cognitive Sciences 4 (5), 178–186.

Griffiths, T.L., Kalish, M.L., 2007. Language evolution by iterated learning with Bayesian agents. Cognitive Science 31, 441–480.

Grimm, J., 1819. Deutsche Grammatik, second ed. Gottingen.

Hockett, C.F., 1961. Linguistic elements and their relations. Language 37 (1), 29–53.

Kirby, S., 1998. Language Evolution without Natural Selection: From Vocabulary to Syntax in a Population of Learners. Edinburgh Occasional Paper in Linguistics 98-1. University of Edinburgh, Edinburgh.

Kirby, S., 2001. Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. IEEE Transactions on Evolutionary Computation 5 (2), 102–110.

Kirby, S., 2002. Learning, bottlenecks and the evolution of recursive syntax. In: Briscoe, E. (Ed.), Linguistic Evolution through Language Acquisition: Formal and Computational Models. Cambridge University Press, Cambridge, pp. 173–203.

Kirby, S., Cornish, H., Smith, K., 2008. Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. Proceedings of the National Academy of Sciences of the United States of America 105 (31), 10681–10686.

Kirby, S., Dowman, M., Griffiths, T.L., 2007. Innateness and culture in the evolution of language. Proceedings of the National Academy of Sciences of the United States of America 104, 5241–5245.

Laland, K.N., Sterelny, K., Odling-Smee, J., Hoppitt, W., Uller, T., 2011. Cause and effect in biology revisited: is Mayr's proximate-ultimate dichotomy still useful? Science 334 (6062), 1512–1516.

Mesoudi, A., Whiten, A., 2008. The multiple roles of cultural transmission experiments in understanding human cultural evolution. Philosophical Transactions of the Royal Society of London B: Biological Sciences 363, 3489–3501.

Perfors, A., Navarro, D.J., 2011. Language evolution is shaped by the structure of the world: an iterated learning analysis. In: Carlson, L., Hoelscher, C. (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society, pp. 477–482.

Sandler, W., Meir, I., Padden, C., Aronoff, M., 2005. The emergence of grammar: systematic structure in a new language. Proceedings of the National Academy of Sciences of the United States of America 102 (7), 2661–2665.

Scott-Phillips, T.C., Dickins, T.E., West, S.A., 2011. Evolutionary theory and the ultimate-proximate distinction in the human behavioral sciences. Perspectives on Psychological Science 6 (1), 38–47.

Smith, K., 2004. The evolution of vocabulary. Journal of Theoretical Biology 228 (1), 127–142.

Smith, K., Kirby, S., 2008. Cultural evolution: implications for understanding the human language faculty and its evolution. Philosophical Transactions of the Royal Society B: Biological Sciences 363, 3591–3603.

Steels, L., 2011. Modeling the cultural evolution of language. Physics of Life Reviews.

von Frisch, K., 1974. Decoding the language of the bee. Science 185, 663–668.

Whiten, A., Mesoudi, A., 2008. Establishing an experimental science of culture: animal social diffusion experiments. Philosophical Transactions of the Royal Society of London B: Biological Sciences 363, 3477–3488.

Zuberbuhler, K., 2002. A syntactic rule in forest monkey communication. Animal Behaviour 63 (2), 293–299.

Evolution and Language: Overview

William Croft, University of New Mexico, Albuquerque, NM, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.H. Greenberg, volume 7, pp. 4975–4982, © 2001, Elsevier Ltd.

Abstract

The close relationship between biological evolution and language was noted by Darwin himself in an oft-quoted passage from *The Descent of Man:* "the formation of different languages and of different species, and the proofs that both have been developed through a gradual process, are curiously parallel" (Darwin, 1882). In fact, the development of evolutionary theory in biology was inspired in part by the advances in historical linguistics in the early nineteenth century. In the twentieth century, evolutionary theory did not have much influence on linguistics, in part due to the advent of structuralism and the focus on synchronic linguistic analysis. In the past two decades, there has been a considerable increase in interest in the relationship between language and evolution. The interest has been manifested in three areas of recent research. The first is the evolutionary origin of the human language capacity, a topic that was actively avoided in linguistics for a century. The second is the employment of techniques of phylogeny reconstruction from biology in the analysis of genetic families of languages. Finally, there is the application of theories of evolutionary processes to language change, based on the hypothesis that, as Darwin wrote, the two are strikingly parallel.

The Evolutionary Origin of Language

The evolutionary origin of language is not technically an example of parallelism between biological evolution and language. It simply is an instance of biological evolution, specifically the evolution of a behavioral capacity in one particular species. This topic was avoided in linguistics for over a century because of the absence of empirical evidence about any stage of language prior to modern human language.

All modern human languages are general-purpose communication systems of a similar order of complexity; there is no reason to believe that any contemporary human languages represent an earlier stage in the evolutionary development of the language capacity. Written records of language are very recent – no more than 5000 years old – and they make clear that languages of that age are no different from contemporary human languages (*see* Writing, Evolution of). Languages otherwise do not leave material evidence, so the archaeological record leaves no direct record of the evolution of the language capacity.

Evidence from hominin skeletal remains suggests that there have been anatomical changes that indicate the increased ability of our ancestors to produce a wide range of vocalizations, and to control our breathing in such a way to produce linguistic sounds. It is generally assumed that the ability to produce a rich variety of vocalizations implies the existence of language at the time of these anatomical changes. However, it could be that such vocalizations had other functions than a general-purpose communication system, such as music and ceremony, or that the sort of language that such vocalizations were used for was quite different in its structure and communicative range than modern human language. Likewise, the great expansion of brain capacity in the development of humankind is an indicator of the emergence of the cognitive capacity for modern human language. But lacking more specific details of neural structure in hominins, this change cannot tell us more about the evolutionary origin of the human language capacity.

The taboo against studying the evolutionary origin of the language capacity has now fallen. A number of different approaches have been taken to tackling the question. All of them are necessarily indirect. Two of the commonest approaches are computer-based models simulating a society of interacting agents, to explore under what circumstances language structures might emerge (e.g., Kirby, 1999); and comparative ethology, comparing the communicative, symbolic, and interactional capacities of our nearest relatives, the primates, and other species that possess vocal communicative systems, such as various species of birds and cetaceans (e.g., Tomasello, 2008). Child language acquisition is also used as a reference point for language origins (Tomasello, 2008). Comparison to child language acquisition must be treated with caution because children learn language by being exposed to a modern human language, which will influence their course of development. Also, children largely already have the cognitive and social capacities of modern humans, so they do not represent the sort of hominins in which the modern human language capacity evolved. More recently, experiments using artificial language learning have come to be used in the search for language origins (*see* Evolution and Language: Cultural Transmission). Again, the subjects of artificial language learning experiments already have the cognitive and social capacities of modern human beings, and they also have knowledge of a modern human language at the point that they engage in the experiments. This latter point also applies to proposals about the origins of language based on more naturalistic examples of language 'creation,' such as pidgins, creoles, and certain sign languages.

Many earlier studies focused on the structure of the linguistic signal. In comparative ethology, the question was whether or not certain structures of human language utterances were unique, or could be found in the communication systems of other species. For example, it was claimed that vervet

monkey alarm calls were symbolic, in that they distinguish three different kinds of predators, and that bee dances and some bird songs involve a combinatoric syntax not unlike that of human languages. In the modeling approaches, the agents would produce strings of units and through simple rules of interpretation and feedback from such interactions, stable meanings of the units (i.e., symbols) and regular syntactic combinations were shown to emerge over several generations of interactions.

More recent studies have taken a broader perspective on the evolution of the human language capacity. Language structure does not evolve for its own sake, but only as part of a larger system of social interaction among human beings. Even the function of communication, often taken to be the raison d'être for language, actually serves the ultimate goal of achieving some joint action between persons. These joint actions form the basis of human culture and society. In other words, the evolution of language is part of the evolution of human social interaction and human culture (Tomasello, 2008).

This broader perspective implies, for example, that the computational modeling of the emergence of language will have limited success with disembodied agents simply reacting to whether a string of units refers to some object or not. In one modeling project from this perspective, the agents are robots that have perceptual and motor abilities, and the modeling includes visual processing and the establishment of joint attention on the part of the robots toward a potential referent for any utterance that is produced (Steels, 2011). Of course, this type of modeling is very ambitious, and requires solving problems in visual and auditory processing, motor activity, and social interaction as well as language structure processing and cognitive categorization and combination. In other words, one must model human beings in their full perceptual, motor, cognitive, and social abilities in order to model the emergence of language.

In the case of comparative ethological approaches, the broader perspective implies looking at the social behavioral capacities of primates and other species, including joint attention, reasoning about other individual's beliefs ('theory of mind'), coordination of actions, and the ability to engage in cooperative activity, as well as cognitive abilities of categorization, causal reasoning, and representing knowledge of displaced experience (i.e., not the 'here and now') that have been traditionally associated with the evolution of the language capacity. All of these social behaviors are preconditions for the evolution of modern human languages. Comparative studies suggest that many other species have the simple cognitive capacities listed above, but few if any have the social cognitive capacities required for language, and not to the degree that modern humans do.

Language embedded in its social interactional context is part of a very complex suite of human behavioral capacities. It is highly unlikely that this complex suite evolved suddenly. Thus, the transition to modern human language in the hominid lineage was probably a gradual one, with a number of intermediate stages of 'pre-language.' Unfortunately, there is still a great gap between animal communication systems known to us and modern human language, so just what those intermediate stages are is largely a matter of speculation. One common proposal is that there was a stage in which utterances consisted of only one symbolic unit (a one-word stage). Some argue that a multiword stage developed by combining these single units in a single utterance, while others argue that a multiword stage developed by reanalyzing the single symbolic unit (a complex vocalization) into meaningful parts. Other proposals are based on the functions of the linguistic units. An earlier stage may involve gestures that are indexical, establishing joint attention in order to coordinate joint actions; names for individuals and symbolic gestures that denote things or actions emerge in a later stage. Or an earlier stage may involve more direct interactions such as commands or questions; utterances whose sole purpose is to share information, especially about displaced experience, develop in a later stage. Such intermediate stages of pre-language only make sense in the context of plausible intermediate stages of social interaction between what is observed in other primates (or other species) and what occurs between human beings in contemporary societies.

The Use of Biological Methods for Language Classification (Phylogenies)

The most obvious respect in which biological evolution and language are parallel is that languages form lineages in much the same way as species do (*see* Evolution and Language: Phylogenetic Analyses). Languages have parents and daughters, and form trees of ancestors and descendants; these trees are called phylogenies in biology (*see* Comparative Method in Evolutionary Studies). In both domains, the ancestors are no longer present, because they have either become extinct or the ancestral species or language has split up into the contemporary species or languages. In only a few cases is there a material record of ancestral forms – fossil individuals of ancient species and written records of ancient languages. As a consequence, evidence for the phylogeny of both species and languages is indirect, namely, the traits of the currently existing species or languages. Comparative work among languages uncovers genetic trees, just as comparative work among species uncovers biological taxa. In some cases, historical linguists have arrived at conclusions about the nature of comparative historical analysis before biologists have. For example, an important principle of comparative historical linguistics is that a trait that is an innovation among daughter languages is more important evidence that the daughter languages form a subgroup than retention of a shared trait from an ancestral language. This principle was later independently adopted in the cladistic approach to constructing phylogenies (Hull, 1988).

Thus, both the basic fact – branching lineages – and the basic method – comparison of present-day members – are the same for biology and language. For this reason, recent work has increased in the application of phylogeny reconstruction methods from biology to language data (*see* Evolution and Language: Phylogenetic Analyses). For the most part, it has been biologists using these methods on language data, rather than linguists adopting these methods for historical linguistics. The basic reason for this fact is that the biological methods are essentially quantitative (and increasingly complex), while the methods of traditional historical linguists are essentially qualitative.

In comparative historical linguistics, the traits used to classify languages into genetic families (i.e., phylogenies) are words, that is, combinations of form and meaning. Words are particularly good traits to use for phylogenies. Words generally evolve independently of each other, so a large number of words represents a large set of independent pieces of evidence for a phylogeny. (In some cases, words are derivationally related, such as *long* and *length*, and so are not entirely independent.) The pairing of form and meaning is largely arbitrary, so other factors are unlikely to lead to convergence across languages. (There are cases of onomatopoeia, such as *meow*, but they are few and are avoided by historical linguists.) For this reason, historical linguists avoid schematic grammatical traits, such as subject-verb-object word order and the presence of grammatical gender. These traits are limited in their possibilities (much more so than the possible pairings of forms and meanings), and their presence is often due to universals of grammatical structure, of the sort studied in linguistic typology (*see* Linguistic Typology), rather than common ancestry of the languages that contain them. In biological terms, words are weakly filtered by selection but schematic grammatical patterns are not. Individual words form lineages themselves, but these word lineages bundle together, especially in basic vocabulary. This is what allows us to say that a language is the daughter of another language: most of its linguistic material comes from a single source. Of course, some words come from a language other than its parent; these are borrowings. Borrowings are more common in nonbasic vocabulary, although they occur in basic vocabulary as well. To the extent that borrowings can be identified and excluded from analysis, then the phylogeny of languages is tree like: a single parent and multiple daughters.

Most applications of biological methods in phylogeny reconstruction to language data begin with a set of words that are assumed to be cognate, that is, inherited from a common ancestor language (called a protolanguage). That is, for a set of languages, the data consists of whether or not some or all of the languages share a cognate for that meaning, for all the meanings in the word list. The phylogeny reconstruction algorithms use computational statistical methods to construct a phylogeny for the languages in question, given the distribution of cognate forms across the languages. These methods have increased in sophistication in recent years, aided by increases in computational power (e.g., Bouckaert et al., 2012; *see* Comparative Method in Evolutionary Studies; Evolution and Language: Phylogenetic Analyses).

The phylogeny reconstruction methods provide a more precisely quantifiable judgment weighing the evidence from the complex distribution of a large number of cognates. They have been applied to universally accepted, chronologically relatively shallow families such as Austronesian (in Oceania) and Bantu (in Africa) in order to clarify the subgroupings of those language families and provide support for hypotheses about the prehistoric migrations of the speakers of those languages. Such methods have their limits, however. For even a relatively small number of languages, the number of possible trees that can be constructed is far larger than the number that can be tested using current computing power. Hence, the phylogeny reconstruction algorithm can only test a small subset of the possible trees against the data provided by the list of cognate forms. (Much of the recent effort in refining phylogeny reconstruction algorithms is in how to explore the search space of possible trees.) One consequence of this fact is that a phylogeny reconstruction algorithm does not produce a single 'best' family tree; it produces a probability distribution of greater or less likely trees for the data.

Against this concern is the fact that even crude methods can find the major subgroups for the language families to which they are applied, such as the major subgroups of Indo-European. That is, the phylogenetic signal is very strong. But that signal is strong enough that visual inspection of cognate lists reveals those aspects of the phylogeny, and employment of quantitative methods is unnecessary. When it comes to resolving phylogenetic questions which are still a matter of debate among historical linguists, then the ability to find the most likely trees becomes an important issue in applying quantitative methods to historical linguistic problems.

Any list of cognates for a set of languages is going to give conflicting information about the phylogeny of those languages. Languages lose cognate forms; the original cognate changes meaning, or is lost entirely, and its form is replaced by another, unrelated form. Hence a missing cognate is not necessarily evidence that the language does not belong to the family in question. Deciding whether a set of cognate forms is a common innovation, justifying placing those languages into a subgroup, or a common retention, whose survival or loss is an accident, is not an automatic process. It requires balancing evidence from other cognate distributions to decide which cognates define a subgroup and which ones do not. The phylogeny reconstruction algorithms encode a particular way to make these decisions, based on their utility in biological evolution. The algorithm may or may not have the same utility in language change (this factor has not been discussed in the relevant literature).

The application of biological phylogeny reconstruction algorithms to language classification is actually a hybrid method. The data to which the algorithm is applied is already processed by a linguist: a linguist has decided which words are cognates and which ones are not. One of the chief problems in determining cognates is whether words that are similar in form and meaning across languages are borrowings or not. Some putative cognates may turn out to be borrowings. Some phylogeny reconstruction programs produce networks rather than trees. In most if not all cases in language history, ancestry vs contact are clearly distinct; for example, English is a Germanic language, albeit with many words borrowed via contact with French. Historical linguists look for patterns in the distribution of similar forms in order to differentiate common ancestry from borrowing. For example, similar forms across the bulk of basic vocabulary is indicative of common ancestry, whereas similar forms concentrated in particular semantic domains of nonbasic vocabulary (e.g., animal husbandry, religious ceremony) is indicative of contact; and a high degree of similarity of forms to the forms in just one language suggest that the latter language is the source of borrowing. Judgments based on differentiating types of vocabulary, or patterns of distribution across sets of languages, are not yet incorporated into phylogeny reconstruction algorithms.

What is of most interest to linguists (and nonlinguists) is the establishment of chronologically deeper language families than those widely accepted among linguists today. The

comparative historical linguistic method starts from languages that are accepted to form a group; phylogeny reconstruction algorithms do the same, by starting with cognates for accepted families. Deeper language families have been proposed, but are controversial among linguists. In principle, it is possible to create algorithms that would identify form–meaning similarities among a complete set of words across a large set of languages not accepted to be related; look for correspondences between the sounds occurring in equivalent positions in the words, which is strong evidence for cognacy; use distributional patterns of form–meaning similarities to differentiate likely cognates from likely borrowings; and use all of this information to postulate likely family trees. It is only very recently that the first efforts to devise such algorithms are being made.

Biologists interested in prehistoric human population history have compared phylogenies reconstructed from biological data to language family trees, and have observed parallels between biological phylogenies of humans (and the migration patterns they imply) and language family trees. The combination of biological and linguistic phylogenies has been used to clarify the prehistoric movements of Austronesian and Bantu peoples, where the language families are widely accepted; and the migration of Native Americans into the New World, where the language families that resemble the biological phylogenies remain controversial (Reich et al., 2012). The parallels between biological and language phylogenies do not necessarily match in all details, because human populations may give up their heritage language and shift to another language. Nevertheless, considering both biological and language families together gives us a more refined picture of human prehistory.

The Parallel between Biological and Linguistic Evolution

The fact that both languages and species form lineages, and numerous other parallels between biological evolution and language change, has led many linguists to discuss analogies between the two domains. Although observing analogies may provide inspiration for theories of language change in linguistics, they do not provide a principled basis for constructing an evolutionary theory of language change. In particular, when one finds a disanalogy between biological evolution and language change, it is unclear whether the disanalogy is due to simply a difference between the two empirical domains, or the disanalogy points to profound differences between biological evolution and language change. In fact, domain-general theories of evolutionary units and processes have been proposed. These theories are intended to subsume both biological evolution and some form of cultural evolution (*see* Cultural Evolution: Overview). There is major debate over the appropriateness of these theories; the following sections will briefly summarize the major issues.

Most biologists consider the fundamental principles of Darwin's evolutionary theory (Darwin, 1859) to continue to hold (*see* Darwinism). These principles are descent with modification and selection. Descent with modification leads to lineages, and occurs through a process that is now generally called replication. Indeed, evolution can be thought of as a model of change by replication, as opposed to a model of inherent change. Replication also occurs in cultural transmission, including language change (Mesoudi et al., 2004; Pagel, 2009). In both biological and cultural evolution, there is debate over what is the replicator, that is, what it is that replicates. In neo-Darwinian evolutionary theory, the paradigm case of the replicator is the gene, and we will use this example for initial illustration.

The fundamental properties of the replication process, most of which are explicated by Darwin, can be grouped into two categories (Dawkins, 1976). The first category has to do with how replicators form lineages. A replicated unit can itself be replicated; the replication process can proceed indefinitely. And some replicators must survive long enough to be replicated themselves (these properties are also called longevity and fecundity). The second category has to do with inheritance, namely what makes the replication process nonrandom. Replicated units possess much of the structure of the original unit. The replication process is mostly faithful. But replication can produce heritable variation: that is, the replication process introduces variation, and that variation is cumulative over generations of replicators.

Among theorists of cultural evolution, three sorts of entities have been proposed as replicators: artifacts, behaviors, and concepts, that is, mental entities. An issue with all of these replicator types is that their replication must be indirect. Artifacts and behaviors are reproduced by human agents, who use their cultural knowledge (i.e., mental conceptual structures) to replicate them. For concepts to be replicated from one person to another ('learning'), they must be publicly displayed in terms of human actions (behavior) or products (artifacts). In addition, concepts are not directly observable and there is evidence that the neural structures and patterns of activation for the same concept may differ from person to person.

In language change, the debate has been traditionally framed in terms of whether language change takes place through imperfect learning by children, or via language use. In evolutionary terms, this is a debate over whether the replicator is a grammar (that is, a speaker's knowledge about his or her language) or linguistic structures in utterances (such structures are replicated every time we speak). Both theories have been debated since the emergence of historical linguistics in the nineteenth century, and the debate continues.

There is considerable empirical evidence that imperfect learning by children cannot be the source of language change (Croft, 2000). Children's errors in acquisition (that is, their linguistic innovations) do not match the processes observed in language change. Children are very good at learning the language spoken around them even though they are not provided with direct negative evidence about the structure of their language (specialists in child language acquisition call this the 'no negative evidence problem'). Finally, children's language is not emulated by other social groups in language change; instead, children expend their effort in emulating other social groups, in particular their older peers when they become adolescents.

By contrast, the sort of variation that is found in the phonetic realization of sounds, and in the verbalization of experience into words and constructions, does match observed

language changes (Croft, 2010). An important element of the usage-based model of language change is that knowledge of language includes knowledge of language variation, and that language variation is pervasive in language. Hence language change is actually a process by which change occurs in the distribution of linguistic variants across phonetic space (for sounds) or conceptual space (for words or constructions). While adults are not as flexible as children in changing their patterns of language use, there is some flexibility, which is greater for some linguistic units (e.g., new words) and less for others (accent); and adults vary in their receptiveness to novel linguistic variants.

Cultural replicators have been called memes (Dawkins, 1976). However, memetics has combined the basic properties of replicators and replication with a 'selfish gene/selfish meme' theory of selection. Selection represents the other fundamental principle of Darwinian evolutionary theory. Selection is a process by which some replicators are successfully propagated in a population while other replicators go extinct (in the extreme case). Selection is a separate process from replication. Most evolutionary biologists recognize that evolution is a two-step process: the generation of variation via replication and selection of some variants over others (also called differential replication).

In the 'selfish gene/meme' approach, the replicator (gene or meme) functions as both the locus of replication and of selection. In biology, the organism is merely a vehicle to facilitate the gene's successful replication, and by analogy, in cultural change, a human agent is merely a means for concepts to differentially replicate themselves. This approach has been controversial in both biology and theories of cultural evolution including language change (but see Ritt, 2004).

Most biologists consider the selfish gene theory to miss the important roles that entities other than the gene play in biological evolution. Analogously, the selfish meme theory misses the important roles that entities other than the meme, in particular human agents, play in cultural evolution. Another generalized theory of evolution, the General Analysis of Selection, provides a richer, though still highly abstract, theory of evolutionary processes (Hull, 1988). The General Analysis of Selection includes the replication process, but analyzes selection as a process of environmental interaction that causes differential replication to take place. Environmental interaction defines two more roles, the environment and the interactor. In the neo-Darwinian theory of biological evolution, the canonical interactor is the organism: an organism interacts with its environment, and natural selection brings about the differential replication of its genes depending on the survival and reproductive success or failure of the organism.

The General Analysis of Selection can also be applied to language (Croft, 2000). The interactor is the human agent, namely the speaker. The speaker's environment includes the experience being communicated, the speakers with whom he or she converses, and the social structures that bind those speakers together in a speech community. In that context of language use, speakers make choices as to which linguistic variants to replicate. The variants that get replicated more frequently are propagated through a speech community, and in the end a historical linguist observes that a language change has taken place.

The General Analysis of Selection provides a framework for an evolutionary model of the processes that have been observed in sociolinguistics. Sociolinguistics observes variation in language use in a speech community. (Sociolinguists generally do not consider the origins of that variation; it has been argued to result from language production, as discussed earlier.) In particular, some of the variation in language use is socially structured: that is, it correlates with social parameters such as gender, age, and socioeconomic status. The association of variants with particular social parameters appears to drive the propagation of certain variants in the speech community, that is, selection of those variants by the speakers in the community.

Another element of a generalized evolutionary theory that plays a central role in cultural evolution and language change is population thinking. The term 'population thinking' refers to the definition of species that emerged by the neo-Darwinian synthesis in the early twentieth century. The population definition replaces the essentialist definition of species that preceded evolutionary theory. In the essentialist approach to species, a species is defined as an ahistorical type with certain essential properties. The essentialist approach is problematic for actual biological species for several reasons. There is a high degree of variation among individuals in a single species; 'essential' traits may be lacking in many individuals. Conversely, two species may overlap substantially in their traits, so 'essential' traits would not distinguish species (and in fact, there are many disputes about how many species there are in particular genera). Finally, the 'essential' traits of a species change over time as species evolve.

The population approach avoids all of these problems. In the population approach, a species is defined not by an essential trait of individuals, but an interactional trait among the members, namely interbreeding (in sexual organisms) and its converse, reproductive isolation between populations. This interactional trait is also the way in which populations form lineages: interbreeding leads to reproduction and lack of interbreeding eventually splits a population. Recently, different definitions of species have been proposed. However, these definitions appear to be the consequence of the fact that populations diverge slowly when interbreeding stops between the populations, and the recent species definitions differ as to what point diverging populations are said to constitute distinct species.

Languages suffer from the same problems with essentialist definitions as biological species. Languages are internally highly variable, so that one cannot provide an 'essential' set of words and/or grammatical rules that would define a language. Two languages may share large amounts of lexicon and grammar, and there are disputes in many cases as to whether there are two languages or two dialects of a single language. Finally, languages change, so 'essential' features of languages may change.

The general solution is to offer a social definition of a language, since a structural (i.e., essentialist) definition is unsatisfactory. More precisely, a population definition of a speech community will avoid the problems of defining a language, since a language is produced by members of the speech community (Croft, 2000; Mufwene, 2001). A population definition of a speech community is a set of speakers that converse with each other. Since what the members of the speech community produce in conversation is the language of that community, it accommodates

variation in the language. When a speech community splits, the languages they speak gradually diverge as the speakers of the two communities stop interacting. At first, the languages are structurally quite similar, but over time they will come to differ; linguists may disagree about the point at which they would say there are now two distinct languages.

Language populations, that is, speech communities, are never completely communicatively isolated from their neighbors. The result is language contact. Although this is sometimes held up as a disanalogy between language change and biological evolution, there is in fact much hybridization in the biological world outside of animals, and many instances of introgression that leads to the transfer of genes from one population to another through incomplete reproductive isolation. Language contact leads to sometimes large-scale transfer of linguistic structures or features from one speech community to another, particularly in the creation of new speech communities through the immigration of multiple populations in colonial situations (new dialect formation and creole formation; Trudgill, 2004; Mufwene, 2001, 2008). As in the case of the selection of certain linguistic variants in a single speech community, selection of linguistic variants from the members of different speech communities that come together in a colonial situation is influenced by social factors that operate in the circumstances of the colonization (the dominance of some groups and enslavement of others, the demographic proportions of speakers, and so on).

The employment of an evolutionary framework for understanding language change allows for the development of mathematical models of mechanisms and processes of language change. Such models are agent-based models, in which entities that represent speakers (as human agents) interact with each other and produce linguistic structures, constrained by a range of competing factors. The process of replicating these linguistic structures feeds back into future linguistic behavior of the agents in a cumulative fashion, leading to evolutionary change in the language produced by the agents/speakers. In other words, these mathematical models represent language as a complex adaptive system (Ellis and Larsen-Freeman, 2009).

The relationship between evolution and language is sufficiently parallel that both can be subsumed under a generalized model of evolutionary processes, and specific mathematical models can be applied to both. Although many linguists, particularly historical linguists, have observed the parallels between evolution and language, a reassessment of the nature of theories and models of language based on an evolutionary perspective is a relatively recent phenomenon in linguistics. The level of impact of evolutionary thinking on linguistic theorizing remains to be seen.

See also: Comparative Method in Evolutionary Studies; Cultural Evolution: Overview; Darwinism; Evolution and Language: Cultural Transmission; Evolution and Language: Phylogenetic Analyses.

Bibliography

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S.J., Alekseyenko, A.V., Drummond, A.J., Gray, R.D., Suchard, M.A., Atkinson, Q.D., 2012. Mapping the origins and expansion of the Indo-European language family. Science 337, 957–960.

Bybee, J.L., 2010. Language. Usage and Cognitive. Cambridge University Press, Cambridge.

Croft, W., 2000. Explaining Language Change: An Evolutionary Approach. Longman, Harlow, Essex.

Croft, W., 2010. The origins of grammaticalization in the verbalization of experience. Linguistics 48, 1–48.

Darwin, C.R., 1859. On the Origin of Species, first ed. John Murray, London. Available at: http://darwin-online.org.uk/contents.html (accessed 30.07.12).

Darwin, C.R., 1882. The Desent of Man, and Selection in Relation to Sex, second ed. John Murray, London. Available at: http://darwin-online.org.uk/contents.html (accessed 30.07.12).

Dawkins, R., 1976. The Selfish Gene. Oxford University Press, New York.

Ellis, N.C., Larsen-Freeman, D. (Eds.), 2009. Language as a Complex Adaptive System. (Special Supplement to Language Learning 59 (4).)

Hull, D.L., 1988. Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science. University of Chicago Press, Chicago, IL.

Kirby, S., 1999. Function, Selection and Innateness: The Emergence of Language Universals. Oxford University Press, Oxford.

Mesoudi, A., Whiten, A., Laland, K.N., 2004. Is human cultural evolution Darwinian? Evidence reviewed from the perspective of the origin of species. Evolution 58, 1–11.

Mufwene, S., 2001. The Ecology of Language Evolution. Cambridge University Press, Cambridge.

Mufwene, S., 2008. Language Evolution: Contact, Competition and Change. Continuum, London.

Pagel, M., 2009. Human language as a culturally transmitted replicator. Nature Reviews Genetics 10, 405–415.

Reich, D., et al., (63 authors), 2012. Reconstructing native American population history. Nature 488, 370–374.

Ritt, N., 2004. Selfish Sounds and Linguistic Evolution: A Darwinian Approach to Language Change. Cambridge University Press, Cambridge.

Steels, L., 2011. Modeling the cultural evolution of language. Physics of Life Reviews 8, 339–356.

Tomasello, M., 2008. Origins of Human Communication. MIT Press, Cambridge, MA.

Trudgill, P., 2004. New-dialect Formation: The Inveitability of Colonial Englishes. Edinburgh University Press, Edinburgh.

Relevant Websites

http://www.evolutionarylinguistics.org – Evolutionary Linguistics.
http://www.eva.mpg.de – Max Planck Institute for Evolutionary Anthropology.
http://www.evolang.org – Evolution of Language International Conferences.

Evolution and Language: Phylogenetic Analyses

Simon Greenhill, Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Language phylogenies are a potentially powerful way to answer questions about how languages and cultures evolve. Recently, phylogenetic methods have been applied to a range of questions about the evolution of human languages and cultures. This article reviews the historical background of these approaches and provides a detailed methodological overview. Three different applications of phylogenetic methods are discussed: how language phylogenies can be used to test population dispersal hypotheses, to investigate processes in language evolution, and to infer patterns in cultural evolution. The article discusses briefly some controversies over the use of these methods before closing with some future prospects.

Language as the Archives of History

There are more than 7000 languages spoken in the world today (Lewis, 2009). These languages can be divided into more than 200 major language families, each containing related languages that have diversified from a common ancestor. For example, the ~12 000 languages of the Austronesian language family originated with a common ancestor – known as Proto-Austronesian – around 5500 years ago (Blust, 1999; Pawley, 2002). The relationships between these languages trace how they and their speakers diverged and expanded over time. Therefore, a family tree or *phylogeny* of these languages holds many clues for investigating the past.

Recently, there has been a series of studies that have applied computational phylogenetic methods from evolutionary biology to languages. These *language phylogenies* have been used to test some prominent and controversial hypotheses in linguistics and anthropology. One subset of studies has tested population dispersal hypotheses including the spread of Indo-European speakers through Europe (Bouckaert et al., 2012; Gray and Atkinson, 2003), the peopling of the Pacific by Austronesian language speakers (Gray et al., 2009), and population expansions in sub-Saharan Africa (Holden, 2002), the Middle East (Kitchen et al., 2009), Japan (Lee and Hasegawa, 2011), and Australia (Bowern, 2010, 2012).

Another set of studies has used language phylogenies to quantify processes in language evolution. These studies have searched for deep linkages between the languages of Melanesia (Dunn et al., 2005), investigated the causes of variation in rates of language evolution (Atkinson et al., 2008; Greenhill et al., 2010; Pagel et al., 2007), and tested how language structures coevolve (Dunn et al., 2011).

Yet other studies have used language phylogenies as a backbone for testing hypotheses about human cultural evolution. These studies include inferring kinship systems (Jordan, 2011; Jordan et al., 2009), how farming and patriliny coevolved (Holden and Mace, 2005), and how political complexity arose in the Pacific (Currie et al., 2010).

Historical Background

Evolutionary biology and historical linguistics share many curious parallels (see Atkinson and Gray, 2005). To understand their respective study material, both fields need to make historical inferences: When did a species or language arise? Where did it come from and what relationships does it have with others? How has the species or language changed over time?

The tree of life is a commonly used concept in evolutionary biology to represent these historical relationships. Since at least the time of Lamarck (1809), trees have been used to describe the history of species and their relationships. It was Darwin, however, who planted this tree at the center of evolutionary theory: the only figure in the *Origin of Species* (Darwin, 1859) is the tree of life.

Darwin was late. Linguists had already been using trees to describe language relationships for some time (Schlegel, 1808; Schleicher, 1863). Trees quickly became commonplace in both evolutionary biology and historical linguistics as descriptions of history. Both fields started to develop quantitative methods for building family trees and making inferences from them. Historical linguistics invented *lexicostatistics* and *glottochronology*. Evolutionary biology started to develop *numerical taxonomy* and *phylogenetics*. Both fields encountered teething problems with these methods. However, here the curious parallels between evolutionary biology and historical linguistics begin to break down. Historical linguistics largely rejected quantitative methods while evolutionary biology developed and improved the methods into a large set of tools known as *phylogenetic methods*.

Historical Linguistics and the Rise and Fall of Lexicostatistics

In a series of papers, Morris Swadesh proposed that the best way to infer language relationships was to count the number of shared similarities on a word list of basic vocabulary. These similarities were not just 'look-alikes' but were *cognate* words. Cognates are words in two languages that have descended from a common ancestor. For example, the words for 'taboo' in Maori (*tapu*), Tahitian (*tapu*), and Hawaiian (*kapu*) are cognates. The words for 'three' are also cognate: *toru* in Tahitian and Maori and *kolu* in Hawaiian. Historical linguistics identifies these cognates by finding systematic sound correspondences between multiple words. Here, *k* in Hawaiian systematically corresponds with *t* in Tahitian and Maori in many words. These systematic sound correspondences diagnose specific phonological innovations that occurred during those languages'

 http://dx.doi.org/10.1016/B978-0-08-097086-8.81035-1

evolutionary history. In this case, the ancestor of these languages – Proto-Polynesian – probably had forms like *tolu* and *tapu* and changed all *t* to *k* after it diverged from Proto-Polynesian (Greenhill and Clark, 2011).

Swadesh argued that the number of cognates shared on a word list measured similarity. The word list Swadesh used contained a varying number of items of *basic vocabulary* that were selected as universal and easily identifiable concepts that could be matched across languages. These items included words for body parts (e.g., hand, head), personal pronouns (e.g., I, we), colors (e.g., black, green), lower numerals, and highly recurrent verbs (e.g., to walk, to swim).

According to Swadesh, the more cognates two languages shared in their basic vocabulary, the more closely related they were. Successively grouping the languages with the highest similarity would therefore build a family tree of those languages. This approach became known as *lexicostatistics*.

Swadesh extended lexicostatistics into *glottochronology* – a system for inferring the ages of language divergences. Glottochronology assumed that languages lost their shared cognates at a constant rate over time. Given a count of cognate similarities between two languages and a rate at which cognates were lost, one could infer divergence times with the following equation:

$$t = \frac{\log C}{2 \log r}$$

where C is the number of shared cognates and t is the amount of time. The value for r was the *retention rate*, i.e., the rate at which cognates were retained. Swadesh initially chose a value of 85% every 1000 years for r, based on the differences documented between Old and Modern English (Swadesh, 1950). A later study by Lees (1953) refined this rate to 81% every 1000 years based on a 200-item word list from 13 historically attested languages. Therefore, according to glottochronology, if two languages shared 81% of their basic vocabulary after 1000 years, they would share 66% after 2000 years and 43% after 3000 years, and so forth.

Swadesh (1950) first applied this method to the Salishan languages spoken in the Pacific Northwest. Lexicostatistics and glottochronology were rapidly applied to languages around the world (e.g., Elbert, 1953; Hymes, 1960; Lees, 1953; Swadesh, 1955, 1950). The results were of great interest and discussed by eminent anthropologists (e.g., Kroeber, 1955) and cited as revolutionizing prehistory (e.g., Murdock, 1964).

However, this popularity did not last due to some serious methodological issues. The supposedly universal word lists were not applicable to all languages ('ice' and 'snow,' e.g., are not common in the Pacific), nor did the simple pairwise count of similarity distinguish between items that were innovated in a lineage or were simply retained from some common ancestor. But the key criticism was rate variation. When rates of change varied between languages, there was a high risk that both the subgroups and the dates were wrong. The most damaging study by Bergsland and Vogt (1962) pointed out that Old Norse and Icelandic diverged around 1000 years ago. Glottochronology estimated their divergence at less than 200 years. In fact, rather than a very tightly constrained retention rate of around 81% cognate loss every 1000 years, Bergsland and Vogt found that the rate varied by around 15–20%.

While some practitioners of lexicostatistics worked to improve the methods (e.g., Embleton, 1981; Sankoff, 1973), they had lost the war. By this time, historical linguistics had largely rejected quantitative methods and lexicostatistics became a textbook example of how not to do linguistics. This rejection has been so complete that even 50 years later linguists are still cautioned that 'linguists do not do dates' (McMahon and McMahon, 2006).

Evolutionary Biology and the Phylogenetic Revolution

Meanwhile, in evolutionary biology, taxonomists had developed tools for building and dating trees that mirrored those found in lexicostatistics. A method known as unweighted pair-group method using arithmetic averages (Sokal and Michener, 1958), or UPGMA, built phylogenetic trees from pairwise genetic similarity just as lexicostatistics did with linguistic similarity. UPGMA and similar methods were extended to estimate divergence times using the assumption of a 'molecular clock' (Zuckerkandl and Pauling, 1965). Kimura (1968), for example, estimated the average time for one base pair change in the genome to be 1.8 years.

Just like linguists, biologists were quick to notice that there were major problems with this method. As with lexicostatistics, these 'numerical taxonomic' methods were undone by variation in rates of genetic change. Variation in rates could lead to markedly incorrect estimates of both the relationships between species (Felsenstein, 1978) and the estimated divergence times (Kirsch, 1969).

Unlike linguists, however, biologists actively sought to improve their methods to handle these problems. First, the problems with methods that calculated pairwise similarity (like UPGMA or lexicostatistics) led to the development of methods that built trees from the raw data itself rather than a summary statistic (Cavalli-Sforza and Edwards, 1967). Second, methods were developed to explicitly account for rate variation in molecular evolution (e.g., Penny et al., 2001; Yang, 1993). Third, the dating of divergence events has become much more robust with a range of new methods that incorporate known historical information and that 'smooth' or 'relax the clock' to account for variation in rates of change over time (e.g., Drummond et al., 2006; Sanderson, 1997). Finally, there has been substantial theoretical development in 'tree thinking' (O'Hara, 1997, 1988) as biologists have realized that trees are not merely diagrams of evolution, but provide ways of testing hypotheses about evolving systems.

Modern Phylogenetics

These methodological and theoretical advances have led to the development of a major field of *phylogenetics* (see also Comparative Method in Evolutionary Studies). The birth of this field was not easy or quiet – debates about phylogenetics were notoriously vicious (Hull, 1988). However, today these phylogenetic methods provide some of the most powerful ways of testing evolutionary hypotheses and have come to dominate evolutionary biology (Huelsenbeck et al., 2001). Phylogenies have become so important in biology that

around 15 new trees are published every day (Rokas, 2006), and being able to read phylogenies is seen as important to biologists as reading a map is to geographers (Gregory, 2008).

The ubiquity of phylogenetic methods in biology have meant that they have cross-fertilized neighboring disciplines including cultural evolution (e.g., Gray et al., 2007), the evolution of manuscripts (e.g., Howe et al., 2001), and musicology (e.g., Rzeszutek et al., 2011). One of the most successful cross-fertilizations has been the wave of studies applying phylogenetic methods to languages.

Bayesian Phylogenetic Methods

Currently, the approach of choice for language phylogenies is a set of tools known as Bayesian phylogenetic methods (Huelsenbeck et al., 2001). Bayesian phylogenetic methods treat the data as fixed observations and aim to fit a model of language change to the data, estimate the model parameters from these data, and use this to infer the best fitting trees using a likelihood framework. There are therefore four major components of a phylogenetic analysis: the data, the model, calculating the likelihood, and the tree search.

Data

The first component of the analysis are the data. Phylogenetic analyses need large amounts of good-quality data to resolve language subgroupings. Cognates from basic vocabulary are often used because they are relatively stable over time and resistant to borrowing (Greenhill et al., 2008). More importantly, however, there are large-scale databases of basic vocabulary and cognate data available for analysis (e.g., Greenhill et al., 2008; Greenhill and Clark, 2011). These data are usually encoded as a binary matrix showing presence or absence in each language (see Table 1).

Cognate data are not the only type of data that can be used – studies have also used information about language structures and typology (Dunn et al., 2005; Greenhill et al., 2010). The only restriction on the data is that there needs to be some way of modeling how they evolve over time.

Model

The next component is a model of how these data could change on a tree. Cognates can either be gained or lost. A cognate gain would occur when a language changes to have the same cognate set for an item such as when a new cognate is 'born' in an ancestral protolanguage. A loss of a cognate would occur when a language changes out of a cognate group when a new lexical innovation arises.

A very simple model of cognate change might allow cognates to be gained or lost at the same rate. This rate is known as a *parameter* of the model. Each model parameter can either be fixed to a known value or estimated from the data. This model with one parameter is quite unrealistic – perhaps cognates are gained at a different rate to being lost. We might expect that it is much harder to gain a new cognate set than it is to lose one. The model could therefore specify two rate parameters – one of cognate gains and one of cognate losses. Yet more complex models would allow the rates of gains and losses to vary over time, e.g., a *covarion* model (Penny et al., 2001), or to enforce the idea that losses are much more numerous than gains by forcing cognates to only arise once, e.g., a *stochastic dollo* model (Alekseyenko et al., 2008; Nicholls and Gray, 2006). All these simple models are often used in reconstructing language phylogenies.

More complex models with more parameters can better model how languages change. If the model is too simple – i.e., *underparameterized* – then one runs the risk of missing important aspects of language change and getting the wrong answer. However, as the model becomes more complex, substantially more data are needed to estimate the values of the parameters effectively (Burnham and Anderson, 1998). If the model is too complex for the available data, then it is *overparameterized* and estimates of the parameter values are likely to be incorrect. There is therefore a balance between an adequately complex model and the amount of data. In practice, most studies fit a range of models to the data and use a model comparison – such as a Bayes Factor (Kass and Raftery, 1995) – to choose the best fitting model.

Given a model of cognate change and some data, one can calculate the fit of those data onto any given tree. To do this, a numeric value called the *likelihood* is calculated, which is simply the probability of the data given, the tree, and the model. The likelihood is the sum of all possible state assignments to the internal nodes. As such it mimics the process of subgrouping languages or species based on shared innovations and not common retained traits. Because the likelihood is the outcome of repeatedly multiplying probabilities, it becomes very small and it is therefore common to report *log likelihood* scores instead.

Table 1 Cognates in some items of basic vocabulary in five Central Pacific languages

	Taboo	*Blood*	*To suck*	*Taboo*	*Blood*	*To suck*[3]	*To suck*[4]
Fijian	tabu[1]	drā	sucu-ma	1	0	0	0
Maori	tapu[1]	toto[2]	ngote[3]	1	1	1	0
Tahitian	tapu[1]	toto[2]	'ote[3]	1	1	1	0
Marquesan	tapu[1]	toto[2]	omo[4]	1	1	0	1
Hawaiian	kapu[1]	koko[2]	omo[4]	1	1	0	1

The left side of the table shows three items of basic vocabulary. Different cognate sets are marked with superscript numerals. Note that 'to suck' has two cognate sets. The right side of the table shows the binary presence/absence coding of these cognate sets. Noncognate items (Fijian 'drā' and 'sucu-ma') are ignored.

Finding the Most Probable Trees

The likelihood gives a way to score different trees according to how well they explain the data under the chosen model of cognate change. However, finding the best trees is hard – there are incredibly many possible permutations. Once there are more than 50 languages in the analysis, there are more possible trees than atoms in the universe.

To find the highly probable trees from this space of possible trees, a Bayesian inferential approach that combines the likelihood with prior knowledge of trees and parameter values is used to give a *posterior probability* distribution of trees. The posterior distribution contains the trees that fit the data well

(i.e., have high likelihoods) given the data, the model of cognate change, and the *priors*. Priors are the initial starting values used for the model parameters. Often priors are 'flat,' considering all possible values as equally probable, but if there is strong external evidence for what these parameters might be, then this can be explicitly taken into account (Huelsenbeck et al., 2001).

Calculating the posterior distribution is hard, requiring the integration of all model parameters over all possible branch-length combinations on every tree. However, a method known as *Markov-Chain Monte Carlo* (MCMC) can sample trees from the posterior probability distribution. To start with, MCMC randomly generates a tree. The initial values for each parameter are sampled from the prior distributions. This original tree is then randomly permuted in some way – by moving a branch on the tree or by modifying one of the model parameters. This permutation gives a new tree that is different from the previous tree by only one change. By calculating the likelihood on this new tree, it can be compared to the old tree to see if it explains the data better. If the new tree fits the data better, then one keeps this tree. If the new tree is worse, then one reverts to the old tree.

This process is repeated many millions of times, sampling the tree and the model parameters at each generation. To start with, the initial trees are poor estimates of the real tree and have incorrect parameter values – this is known as the *burn-in* period. However, because the method preferentially selects the better tree at each generation, the trees and parameters converge toward good estimates of their real values. After running the chain for many generations, the analysis essentially samples the best trees from the space of possible trees to give the posterior probability distribution.

Applications of Language Phylogenies

Testing Population Dispersal Hypotheses

A major use of language phylogenies has been to test hypotheses about how human populations expanded and dispersed across the globe (see also Cultural Evolution: Phylogeny versus Reticulation). One of the greatest human dispersals was the settlement of the Pacific. During the Holocene, a new culture arose in the Pacific and spread through the region from Taiwan, into Island Southeast Asia, on to Oceania, eventually reaching as far afield as Madagascar, Hawaii, Rapanui (Easter Island), and New Zealand. This journey was made all the more impressive by being undertaken without the help of modern sailing equipment. As these people settled this region, their languages diversified into the vast Austronesian language family. These Austronesian languages now comprise one of the largest language families in the world containing between 1000 and 1200 languages (Blust, 2009).

There are two competing theories about where these people came from and how they spread through the Pacific (Gray et al., 2009). The first theory, the *pulse-pause* scenario, proposes that the Austronesians originated in Taiwan around 5500 years ago and expanded through the Pacific in a series of expansion pulses and settlement pauses (Blust, 1999; Diamond and Bellwood, 2003; Pawley, 2002). According to this scenario, after the initial settlement of Taiwan there was a settlement pause of around 1000 years followed by a major expansion pulse into the Philippines, through Island Southeast Asia, along the coast of New Guinea, before reaching Tonga, Samoa, and Fiji around 2800–3000 years ago (Burley et al., 2012). Here, a second settlement pause occurred, coinciding with the development of the pre-Polynesian society (Kirch and Green, 2001). Around 1000 years ago, a second expansion pulse occurred as these peoples – now Polynesian – expanded into Eastern Polynesia, eventually reaching Hawaii, New Zealand, and Rapanui around 800 years ago (Wilmshurst et al., 2011).

In contrast, the second theory of Pacific settlement, the *slow boat* scenario, proposes that the Austronesians originated in Wallacea – the area around Sulawesi and the Moluccas (Oppenheimer and Richards, 2001). The slow boat scenario claims that the Austronesians expanded out of Wallacea around 13 000–15 000 years ago following rises in sea level after the submerging of the Sunda shelf at the end of the last ice age. According to this scenario, the Austronesians expanded in two prongs – the first headed north through the Philippines and into Taiwan, while the second traveled east along the New Guinea coast into Polynesia.

Gray et al. (2009) used language phylogenies to test between these two scenarios. Their dataset was more than 34 000 cognate sets from 400 Austronesian languages (Greenhill et al., 2008). The results from Gray et al. (2009) showed overwhelming support for the pulse-pause scenario. First, the Austronesian trees were rooted in the Formosan languages of Taiwan – not in Wallacea – and showed a single chainlike pattern rather than a two-pronged expansion.

Second, the two scenarios propose quite different timings for the Austronesian expansion. The pulse-pause predicts an origin around 5500 years while the slow boat predicts an age between 13 000 and 15 000 years. To test this prediction, Gray et al. (2009) used a nonparametric rate–smoothing approach to convert the branches on the trees (here measuring amount of lexical change) into time. This method smooths the observed rates of change across the trees in the posterior distribution while accounting for historical calibrations (Sanderson, 1997).

The historical calibrations were based on known historical information about parts of the tree, e.g., the development of the large Proto-Oceanic subgroup 3200–3600 years ago (Lynch et al., 2002), the settlement of the Reefs and Santa Cruz islands between 3000 and 3150 years ago (Green et al., 2008), and the historical attestation of the Chamic Subgroup between 1800 and 2500 years ago (Thurgood, 1999). Dating the trees in the posterior distribution showed an average age of 5230 years (95% highest posterior density interval = 4750–5800 years ago). This finding again supports the pulse-pause scenario instead of the slow boat.

Strikingly, the trees revealed that the pattern of expansion pulses and pauses visible in the trees matched those predicted by the pulse-pause scenario. If language diversification were linked to the population expansions, then expansion pulses will leave little time for linguistic changes to accumulate. In pulse regions, the trees will have small, short branches. In contrast, the settlement pauses will show long branches due to the increased amount of lexical changes accumulated. The Austronesian trees showed precisely this pattern – long branches matching the settlement pauses and short branches

matching the expansion pulses. In sum, the study by Gray et al. (2009) shows the great promise of phylogenetic methods for testing population dispersal hypotheses in terms of timing, pattern, and sequence.

Quantifying Processes of Language Change

Understanding the tempo of language change is critical for many questions in historical linguistics. What aspects of languages are stable over time and what types are rapidly changing? This information is important for resolving language relationships – if we want to identify deep language relationships, then slowly changing features should be our data of choice. On the other hand, more rapidly changing features can provide finer-grained resolution at shorter time scales.

A series of phylogenetic studies has investigated the tempo of change in lexical items. Pagel and Meade (2006) constructed trees of 87 Indo-European languages using Bayesian phylogenetic methods. They then calculated the instantaneous rates of change (i.e., the rate at any point on the phylogeny) for items of basic vocabulary using a statistical model of word evolution. This rate was converted into a *half-life*, i.e., the amount of time needed for a word to lose 50% of its cognates. The average half-life for each word in both families was 2343 years and a median of 1449 years. The slowest words in the Indo-European languages were 'two,' 'three,' 'five,' and 'who' with an estimated half-life of 19 600 years.

What might cause these variations in rates of change? One factor thought to influence the rate of change is frequency of use. Pagel et al. (2007) extended their previous study by comparing the rates of change in lexical items to the frequency those items were used in four large corpora of English, Spanish, Russian, and Greek languages. The results show that the rate of lexical change in the lexicon is predicted by the frequency at which words are used in modern-day languages. That is, the more a word is used in speech, the less likely it is to change. This frequency of use effect is strong, explaining about 50% of the variation in the rates of change in these items of lexicon.

Using phylogenetic methods to investigate rate variation also allows us to explore broader questions about the patterns of change. A long-standing debate in linguistics concerns how far back one can trace prehistory. The accepted rule of thumb suggests that the tempo of language change is so fast that any historical signal is lost after 10 000 years (Ringe, 1995). However, if there are features that are more stable, one might be able to trace language relationships at far deeper levels (Gray, 2005). Some authors have argued that certain structural features, like word order or pronoun systems, are very stable over time and can identify language relationships dating back tens of thousands of years (Nichols, 1992).

To test the hypothesis about structural features, Greenhill et al. (2010) used a phylogenetic approach to infer the rates of change in Austronesian and Indo-European languages for 200 items of basic vocabulary and 138 structural/typological features. The results showed that, on average, the lexical items changed slightly faster than the structural features in the Austronesian languages but there was no significant difference in the Indo-European languages. More importantly, many of the features proposed to be highly stable were not – showing marked variation between the two language families. For example, the subset of structural features related to nominal syntax has some of the highest rates of change in Austronesian, but is among the slowest in Indo-European. There was no significant relationship in rates of change between Austronesian and Indo-European. These results showed that structural features were not changing significantly slower than the lexicon and the rates were not consistent between families. This study highlights the problems with broad assumptions of stability between different languages and suggests that phylogenetic methods might provide a better way of moving these debates forward by directly testing these implicit assumptions.

Investigating Cultural Evolution

Language phylogenies provide a way of testing hypotheses about cultural evolution (see Cultural Evolution: Overview). A major problem for statistical tests on cultures is that cultures share similarity due to descent and are therefore not statistically independent. This is known as Galton's problem (Naroll, 1961). One of the great benefits of phylogenies is that they directly estimate the similarity due to descent, allowing this to be controlled for. This benefit has led to the development of a series of *phylogenetic comparative methods* for testing comparative hypotheses (Felsenstein, 1985). Unfortunately, large-scale cross-cultural databases are lacking (Greenhill et al., 2008). However, since languages trace the history of cultures well they can work as a proxy for culture history. When coupled with phylogenetic comparative methods they provide a powerful tool for investigating cultural evolution (Mace and Pagel, 1994).

For example, one study used the Austronesian language phylogenies described above to test how political complexity evolved (Currie et al., 2010). Anthropologists have classified societies into different stages of political complexity. These stages range from small *acephalous* societies without any permanent leadership, to *simple chiefdoms* with one level of leadership beyond the local community, to *complex chiefdoms* with two levels of leadership, and then to *states* with more than two levels. There has been much debate about how these differences in political complexity arose. Do societies move through these levels to increase in complexity (e.g., band to tribe to chiefdom to state)? Does this sequence always have to be followed, or can societies change between these levels freely?

To investigate these questions Currie et al. (2010) used the Austronesian language phylogenies discussed earlier (Gray et al., 2009). They first collected information about political complexity for the Austronesian societies speaking the languages used in the language trees. These data classified each society into one of four stages (acephalous, simple chiefdom, complex chiefdom, and state). The society stages were then mapped onto the Austronesian phylogenies from Gray et al. (2009). Currie et al. then fitted different models of how political complexity could evolve onto these trees and data. For example, one model specified that only sequential increases in complexity were possible. Another model specified that only sequential increases and decreases were

possible. Yet other models allowed major jumps or catastrophic declines in complexity. Calculating the fit of these models onto the trees and the data showed which model explained these data the best.

Currie et al.'s results showed that the best fitting model only allowed sequential increases and decreases in complexity. That is, the political complexity in these Austronesian societies tended to increase or decrease only in small steps between stages. There were no big jumps in complexity (e.g., from acephalous societies to states). The next best fitting model was one that allowed catastrophic decreases suggesting that while major increases in complexity might not happen, collapses do.

Controversies

Language phylogenies have often been controversial. Following the rejection of lexicostatistics, historical linguistics has had a 'tradition of hostility toward probabilistic modelling' (Sankoff, 1973, p. 96). Phylogenetic methods are often criticized for simply being 'lexicostatistics redux,' with all the fatal flaws inherent within lexicostatistics. However, phylogenetics developed in response to many of the same problems faced by lexicostatistics and has moved far beyond lexicostatistics in power and robustness (Greenhill and Gray, 2009).

The most substantive criticism of language phylogenies is that languages do not necessarily evolve like species do (Bateman et al., 1990). The main component of this argument is that languages can borrow cognates, typological features, or grammatical structures from other languages. English, for example, has borrowed around 60% of its lexicon from French and Latin (Embleton, 1986). The critics argue that high levels of borrowing cause phylogenetic methods to give inaccurate or invalid results.

This criticism has lead to a highly polarized debate between those who think that trees can represent linguistic and cultural evolution well and those who think that phylogenetic methods should be rejected (see Greenhill et al., 2009). The truth is somewhere in between. First of all, an important component of phylogenetic analysis is choice of data. The data of choice for most language phylogenies to date – basic vocabulary – are resistant to being borrowed. While English has borrowed up to 60% of its total lexicon from French and Latin, only about 6% of its basic vocabulary is from these languages (Embleton, 1986). Choosing the right data is half the battle.

Borrowing is not an all-or-nothing prospect – the key question is what levels of borrowing are problematic. To answer this question, Greenhill et al. (2009) simulated borrowing of words on language trees and then attempted to reconstruct their histories. The results showed that Bayesian phylogenetic methods were robust to even quite high levels of borrowing – 15–20% of the basic vocabulary every 1000 years. Therefore, we would expect these methods to work quite well even when borrowing is reasonably common.

When borrowing is substantial, alternative approaches do exist. The languages of Polynesia and Vanuatu, for example, are notorious for high levels of borrowing and complex dialect chains. In biology, phylogenetic network methods like NeighborNet or Split Decomposition (e.g., Huson and Bryant, 2006) have been developed to handle cases where there are multiple histories – such as when plants hybridize or when bacteria exchange genetic material between different species. In linguistics, these networks have been used to visualize complex and conflicting patterns of descent for some time (e.g., Bryant et al., 2005; Greenhill et al., 2010). Until now, phylogenetic networks were primarily visualization tools. Recently, however, these methods have been applied to languages to estimate the amount of lexical borrowing in Indo-European languages (Nelson-Sathi et al., 2011), and to quantify the amount of conflicting signal in various linguistic and cultural datasets (Gray et al., 2010). Perhaps the best way to analyze the complex descent patterns found in places like Vanuatu and Polynesia is to use these recently developed phylogenetic network methods rather than phylogenetic trees (Gray et al., 2010). These methods are still in their infancy but show great promise for helping investigate complex histories.

Conclusions and Future Prospects

To date, language phylogenetic work has been largely driven from outside historical linguistics but shows great potential for complementing the historical linguistic tool set. The future development of this field hinges on moving beyond the wholesale importation of methods from evolutionary biology. Major advances will come from *transforming* these methods to be more appropriate for languages, for example, through better models of sound change (e.g., Bouchard-Côté et al., 2009) or better ways for handling complex borrowing and diversification patterns (Gray et al., 2010).

The three applications presented above hint at the promise that a phylogenetic approach to languages has. In terms of testing population dispersal hypotheses, these methods can help us locate language families – and cultures – in both space and time. With well-dated trees of languages and cultures, one will be able to gain a much better understanding of prehistory. In terms of the processes in language evolution, phylogenetic methods can infer and quantify the large-scale broad patterns that have shaped the languages spoken today. As for investigating cultural evolution, language phylogenies give a robust set of tools for making inferences about how human societies have evolved. When combined with good linguistic and cultural data, language phylogenies provide a very powerful set of tools for answering questions about human prehistory and how languages evolve. These methods raise the possibility of an exciting new field of computational historical linguistics that can bring linguistics into the forefront of the human sciences.

See also: Comparative Method in Evolutionary Studies; Cultural Evolution: Overview; Cultural Evolution: Phylogeny versus Reticulation; Cultural Evolution: Theory and Models; Darwinism; Evolution and Language: Cultural Transmission; Evolution and Language: Overview; Evolution, History of.

Bibliography

Alekseyenko, A.V., Lee, C.J., Suchard, M.A., 2008. Wagner and Dollo: A stochastic duet by composing two parsimonious solos. Systematic Biology 57, 772–784.

Atkinson, Q.D., Gray, R.D., 2005. Curious parallels and curious connections–phylogenetic thinking in biology and historical linguistics. Systematic Biology 54, 513–526.

Atkinson, Q.D., Meade, A., Venditti, C., Greenhill, S.J., Pagel, M., 2008. Languages evolve in punctuational bursts. Science 319, 588.

Bateman, R., Grady, R., Funk, V.A., Mooi, R., Kress, W.J., Cannell, P., 1990. Speaking of forked tongues: the feasibility of reconciling human phylogeny and the history of language. Current Anthropology 31, 1–24.

Bergsland, K., Vogt, H., 1962. On the validity of glottochronology. Current Anthropology 3, 115–153.

Blust, R., 1999. Subgrouping, circularity and extinction: some issues in Austronesian comparative linguistics, Selected papers from the Eighth International Conference on Austronesian Linguistics. Symposium Series of the Institute of Linguistics, Academia Sinica, Taipei, Taiwan.

Blust, R.A., 2009. The Austronesian Languages. Pacific Linguistics, Canberra.

Bouchard-Côté, A., Griffiths, T.L., Klein, D., 2009. Improved reconstruction of proto-language word forms. In: Proceedings of Human Language Technologies. Association for Computational Linguistics, Boulder, Colorado, pp. 65–73.

Bouckaert, R.R., Lemey, P., Dunn, M., Greenhill, S.J., Alekseyenko, A.V., Drummond, A.J., Gray, R.D., Suchard, M.A., Atkinson, Q.D., 2012. Mapping the origins and expansion of the Indo-European language family. Science 337, 957–960.

Bowern, C., 2010. Historical linguistics in Australia: trees, networks and their implications. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 365, 3845–3854.

Bowern, C., 2012. The riddle of Tasmanian languages. Proceedings of the Royal Society B: Biological Sciences 279, 4590–4595.

Bryant, D., Filimon, F., Gray, R.D., 2005. Untangling our past: languages, trees, splits and networks. In: Mace, R., Holden, C.J., Shennan, S.J. (Eds.), The Evolution of Cultural Diversity: A Phylogenetic Approach. UCL Press, London, pp. 67–84.

Burley, D., Weisler, M.I., Zhao, J.-X., 2012. High precision U/Th dating of first Polynesian settlement. PLoS ONE 7 48769.

Burnham, K.P., Anderson, D.R., 1998. Model Selection and Inference: A Practical Information-Theoretic Approach. Springer, New York.

Cavalli-Sforza, L.L., Edwards, A.W., 1967. Phylogenetic analysis. Models and estimation procedures. American Journal of Human Genetics 19, 233–257.

Currie, T.E., Greenhill, S.J., Gray, R.D., Hasegawa, T., Mace, R., 2010. Rise and fall of political complexity in Island South-East Asia and the Pacific. Nature 467, 801–804.

Darwin, C., 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. Murray, London.

Diamond, J.M., Bellwood, P., 2003. Farmers and their languages: the first expansions. Science 300, 597–603.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLOS Biology 4, e88.

Dunn, M., Greenhill, S.J., Levinson, S.C., Gray, R.D., 2011. Evolved structure of language shows lineage-specific trends in word-order universals. Nature 473, 79–82.

Dunn, M., Terrill, A., Reesink, G., Foley, R.A., Levinson, S.C., 2005. Structural phylogenetics and the reconstruction of ancient language history. Science 309, 2072–2075.

Elbert, S.H., 1953. Internal relationships of the Polynesian languages and dialects. Southwestern Journal of Anthropology 9, 147–173.

Embleton, S., 1981. Lexicostatistical tree reconstruction incorporating borrowing. Toronto Working Papers in Linguistics.

Embleton, S., 1986. Statistics in Historical Linguistics. Studienverlag Brockmeyer, Bochum.

Felsenstein, J., 1978. The number of evolutionary trees. Systematic Zoology 27, 27–33.

Felsenstein, J., 1985. Phylogenies and the comparative method. The American Naturalist 125, 1–15.

Gray, R.D., 2005. Pushing the time barrier in the quest for language roots. Science 309, 2007–2008.

Gray, R.D., Atkinson, Q.D., 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426, 435–439.

Gray, R.D., Bryant, D., Greenhill, S.J., 2010. On the shape and fabric of human history. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 365, 3923–3933.

Gray, R.D., Drummond, A.J., Greenhill, S.J., 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. Science 323, 479–483.

Gray, R.D., Greenhill, S.J., Ross, R.M., 2007. The pleasures and perils of darwinizing culture (with phylogenies). Biological Theory 2, 360–375.

Green, R.C., Jones, M., Sheppard, P., 2008. The reconstructed environment and absolute dating of SE-SZ-8 Lapita site on Nendö, Santa Cruz, Solomon Islands. Archaeology in Oceania 43, 49–61.

Greenhill, S.J., Atkinson, Q.D., Meade, A., Gray, R.D., 2010. The shape and tempo of language evolution. Proceedings of the Royal Society B: Biological Sciences 277, 2443–2450.

Greenhill, S.J., Blust, R., Gray, R.D., 2008. The Austronesian basic vocabulary database: from bioinformatics to lexomics. Evolutionary Bioinformatics 4, 271–283.

Greenhill, S.J., Clark, R., 2011. POLLEX-online: the polynesian lexicon project online. Oceanic Linguistics 50, 551–559.

Greenhill, S.J., Currie, T.E., Gray, R.D., 2009. Does horizontal transmission invalidate cultural phylogenies? Proceedings of the Royal Society of London, Series B, Biological Sciences 276, 2299–2306.

Greenhill, S.J., Gray, R.D., 2009. Austronesian language phylogenies: myths and misconceptions about Bayesian computational methods. In: Adelaar, K.A., Pawley, A. (Eds.), Austronesian Historical Linguistics and Culture History: A Festschrift for Robert Blust. Pacific Linguistics, Canberra, pp. 375–397.

Gregory, T.R., 2008. Understanding evolutionary trees. Evolution: Education and Outreach 1, 121–137.

Holden, C.J., 2002. Bantu language trees reflect the spread of farming across Sub-Saharan Africa: a maximum-parsimony analysis. Proceedings of the Royal Society of London, Series B, Biological Sciences 269, 793–799.

Holden, C.J., Mace, R., 2005. 'The cow is the enemy of matriliny': using phylogenetic methods to investigate cultural evolution in Africa. In: Mace, R., Holden, C.J., Shennan, S.J. (Eds.), The Evolution of Cultural Diversity: A Phylogenetic Approach. UCL Press, London, pp. 217–234.

Howe, C.J., Barbrook, A.C., Spencer, M., Robinson, P., Bordalejo, B., Mooney, L.R., 2001. Manuscript evolution. Trends in Genetics 17, 147–152.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R., Bollback, J.P., 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294, 2310–2314.

Hull, D.L., 1988. Science as a Process. University of Chicago Press, Chicago, IL.

Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23, 254–267.

Hymes, D., 1960. Lexicostatistics so far. Current Anthropology 1, 3–44.

Jordan, F.M., 2011. A phylogenetic analysis of the evolution of Austronesian sibling terminologies. Human Biology 83, 297–321.

Jordan, F.M., Gray, R.D., Greenhill, S.J., Mace, R., 2009. Matrilocal residence is ancestral in Austronesian societies. Proceedings of the Royal Society B: Biological Sciences 276, 1957–1964.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.

Kimura, M., 1968. Evolutionary rate at the molecular level. Nature 217, 624–626.

Kirch, P.V., Green, R.C., 2001. Hawaiki, Ancestral Polynesia: An Essay in Historical Anthropology. Cambridge University Press, Cambridge.

Kirsch, J.A.W., 1969. Serological data and phylogenetic inference: the problem of rates of change. Systematic Zoology 18, 296–311.

Kitchen, A., Ehret, C., Assefa, S., Mulligan, C.J., 2009. Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. Proceedings of the Royal Society B: Biological Sciences 270, 2703–2710.

Kroeber, A.L., 1955. Linguistic time depth results so far and their meaning. International Journal of American Linguistics 21, 91–104.

Lamarck, J.B., 1809. Philosophie zoologique, ou Exposition des considérations relatives à l'histoire naturelle des animaux. Dentu, Paris.

Lee, S., Hasegawa, T., 2011. Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. Proceedings. Biological Sciences/The Royal Society 278, 3662–3669.

Lees, R.B., 1953. The basis of glottochronology. Language 29, 113–127.

Lewis, P.M. (Ed.), 2009. Ethnologue: Languages of the World. SIL International, Dallas, Texas.

Lynch, J., Ross, M.D., Crowley, T., 2002. The Oceanic Languages. Curzon Press, Richmond.

Mace, R., Pagel, M., 1994. The comparative method in anthropology. Current Anthropology 35, 549–564.

McMahon, A., McMahon, R., 2006. Why linguists don't do dates: evidence from Indo-European and Australian languages. In: Forster, P., Renfrew, C. (Eds.), Phylogenetic Methods and the Prehistory of Languages. McDonald Institute for Archaeological Research, Cambridge, pp. 153–160.

Murdock, G.P., 1964. Genetic classification of the Austronesian languages: a key to Oceanic culture history. Ethnology 3, 117–216.

Naroll, R., 1961. Two Solutions to Galton's problem. Philosophy of Science 28, 15–39.

Nelson-Sathi, S., List, J.-M., Geisler, H., Fangerau, H., Gray, R.D., Martin, W., Dagan, T., 2011. Networks uncover hidden lexical borrowing in Indo-European language evolution. Proceedings of the Royal Society B: Biological Sciences 278, 1794–1803.

Nicholls, G.K., Gray, R.D., 2006. Dated ancestral trees from binary trait data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70, 545–566.

Nichols, J., 1992. Linguistic Diversity in Space and Time. University of Chicago Press, Chicago, IL.

Oppenheimer, S.J., Richards, M.B., 2001. Fast trains, slow boats, and the ancestry of the Polynesian islanders. Science Progress 84, 157–181.

O'Hara, R.J., 1988. Homage to Clio, or, toward an historical philosophy for evolutionary biology. Systematic Biology 37, 142–155.

O'Hara, R.J., 1997. Population thinking and tree thinking in systematics. Zoologica Scripta 26, 323–330.

Pagel, M., Atkinson, Q.D., Meade, A., 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. Nature 449, 717–720.

Pagel, M., Meade, A., 2006. Estimating rates of lexical replacement on phylogenetic trees of languages. In: Forster, P., Renfrew, C. (Eds.), Phylogenetic Methods and the Prehistory of Languages. McDonald Institute for Archaeological Research, Cambridge, pp. 173–182.

Pawley, A., 2002. The Austronesian dispersal: languages, technologies and people. In: Bellwood, P., Renfrew, C. (Eds.), Examining the Farming/language Dispersal Hypothesis. McDonald Institute for Archaeological Research, Cambridge, pp. 251–273.

Penny, D., McComish, B.J., Charleston, M.A., Hendy, M.D., 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. Journal of Molecular Evolution 53, 711–723.

Ringe, D., 1995. "Nostratic" and the factor of chance. Diachronica 12, 55–74.

Rokas, A., 2006. Genomics and the tree of life. Science 313, 1897–1899.

Rzeszutek, T., Savage, P.E., Brown, S., 2011. The structure of cross-cultural musical diversity. Proceedings of the Royal Society B: Biological Sciences 279, 1606–1612.

Sanderson, M.J., 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Molecular Biology and Evolution 14, 1218–1231.

Sankoff, D., 1973. Mathematical developments in lexicostatistic theory. Current Trends in Linguistics 11, 93–113.

Schlegel, F., 1808. Ueber die Sprache und Weisheit der Indier: Ein beitrag zur Begründung der Alterthumskunde. Mohr & Zimmer, Heidelberg.

Schleicher, A., 1863. Die Darwinsche theorie und die sprachwissenschaft. Hermann Bohlau, Weimar.

Sokal, R.R., Michener, C.D., 1958. A statistical method for evaluating systematic relationships. The University of Kansas Science Bulletin 38, 1409–1438.

Swadesh, M., 1950. Salish internal relationships. International Journal of American Linguistics 16, 157–167.

Swadesh, M., 1955. Towards greater accuracy in lexicostatistic dating. International Journal of American Linguistics 21, 121–137.

Thurgood, G., 1999. From Ancient Cham to Modern Dialects. Oceanic Linguistics Special Publications, Hawaii.

Wilmshurst, J.M., Hunt, T.L., Lipo, C.P., Anderson, A.J., 2011. High-precision radiocarbon dating shows recent and rapid initial human colonization of East Polynesia. Proceedings of the National Academy of Sciences of the United States of America 108, 1815–1820.

Yang, Z., 1993. Maximum-Likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular Biology and Evolution 10, 1396–1401.

Zuckerkandl, E., Pauling, L., 1965. History of evolutionary molecules as documents. Journal of Theoretical Biology 8, 357–366.

Relevant Website

http://language.psy.auckland.ac.nz/austronesian/ – Austronesian Basic Vocabulary Database.

Evolution: Diffusion of Innovations

Everett M Rogers,[†] University of New Mexico, Albuquerque, NM, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 7, pp. 4982–4986, © 2001, Elsevier Ltd.

Abstract

This article discusses the model of the diffusion of innovations and describes this model's key variables. Diffusion is the process by which an innovation is communicated through certain channels over time among the members of a social system. An innovation is an idea, practice, or object that is perceived as new by an individual or other unit of adoption.

This article discusses the model of the diffusion of innovations and describe this model's key variables. First, however, a brief history of diffusion research is presented. The paradigm for diffusion research can be traced to the rural sociology research tradition, which began in the 1940s. Rural sociology is a subfield of sociology that focuses on the social problems of rural life. One rural sociology study in particular influenced the methodology, theoretical framework, and interpretations of later students in the rural sociology tradition, and in other diffusion research traditions. Ryan and Gross (1943) investigated the diffusion of hybrid seed corn among Iowa farmers. Hybrid seed, the result of agricultural research, was made available to Iowa farmers in 1928. The hybrid vigor of the new seed increased corn yields on Iowa farms, hybrid corn varieties withstood drought more effectively than the open-pollinated seed they replaced, and hybrid corn was better suited to harvesting by mechanical corn pickers. By 1941, about 13 years after its first release, the innovation was adopted by almost 100% of Iowa farmers. Ryan and Gross studied the relatively rapid diffusion of hybrid corn in order to obtain lessons learned that might be applied to the diffusion of other farm innovations. However, the intellectual influence of the hybrid corn study reached far beyond the study of agricultural innovations, and outside of the rural sociology tradition of diffusion research. Since the 1960s, the diffusion model has been applied in a wide variety of disciplines such as education, public health, communication, marketing, geography, sociology, and economics (Pelto, 1973; Rosen, 2000; Van den Ven et al., 1989; Zaltman et al., 1973). Diffusion studies in these various disciplines have ranged from the rapid diffusion of the Internet to the nondiffusion of the Dvorak keyboard (in typewriters and computers). The innovations that have been studied in diffusion research have mainly been technological innovations.

Diffusion is the process by which (1) an innovation (2) is communicated through certain channels (3) over time (4) among the members of a social system (Rogers, 1995). Diffusion is a special type of communication concerned with the spread of messages that are perceived as new ideas, and necessarily represents a high degree of uncertainty to the individual. The four main elements in the diffusion of new ideas are (1) the innovation, (2) communication channels, (3) time, and the social system.

[†] Deceased

The Innovation

An innovation is an idea, practice, or object that is perceived as new by an individual or other unit of adoption. The characteristics of an innovation, as perceived by the members of a social system, determine its rate of adoption. Some innovations diffuse relatively slower, and other innovations diffuse faster. Why do certain innovations spread more quickly than others? The characteristics which determine an innovation's rate of adoption are: (1) relative advantage, (2) compatibility, (3) complexity, (4) trialability, and observability.

Relative Advantage

This is the degree to which an innovation is perceived as better than the idea it supersedes. The degree of relative advantage may be measured in economic terms, but social prestige, convenience, and satisfaction are also important factors. It does not matter so much if an innovation has a great deal of objective advantage. What does matter is whether an individual perceives the innovation as advantageous. The greater the perceived relative advantage of an innovation, the more rapid its rate of adoption will be.

Compatibility

This is the degree to which an innovation is perceived as being consistent with the existing values, past experiences, and needs of potential adopters. An idea that is incompatible with the values and norms of a social system will not be adopted as rapidly as an innovation that is compatible. The adoption of an incompatible innovation often requires the prior adoption of a new value system, which is a relatively slow process.

Complexity

This is the degree to which an innovation is perceived as difficult to understand and use. Some innovations are readily understood by most members of a social system; others are more complicated and will be adopted more slowly. New ideas that are simpler to understand are adopted more rapidly than innovations that require the adopter to develop new skills and understandings.

Trialability

This is the degree to which an innovation may be experimented with on a limited basis. New ideas that can be tried on the

installment plan generally will be adopted more quickly than innovations that are not divisible. An innovation that is trialable represents less uncertainty to the individual who is considering it for adoption, who can learn by doing.

Observability

This is the degree to which the results of an innovation are visible to others. The easier it is for individuals to see the results of an innovation, the more likely they are to adopt it. Such visibility stimulates peer discussion of a new idea, as friends and neighbors of an adopter often request innovation-evaluation information about it.

Communication Channels

The second main element in the diffusion of new ideas is the communication channel. Communication is the process by which participants create and share information with one another in order to reach a mutual understanding. A communication channel is the means by which messages get from one individual to another. Mass media channels are more effective in creating knowledge of innovations, whereas interpersonal channels are more effective in forming and changing attitudes toward a new idea, and thus in influencing the decision to adopt or reject a new idea. Most individuals evaluate an innovation, not on the basis of scientific research by experts, but through the subjective evaluations of near-peers who have adopted the innovation. So the process is essentially social in nature, being driven by individuals talking to others about the new idea, as they give meaning to the innovation.

Time

The third main element in the diffusion of new ideas is time. The time dimension is involved in diffusion in three ways.

Innovation-Decision Process

First, time is involved in the innovation-decision process. The innovation-decision process is the mental process through which an individual (or other decision-making unit) passes from first knowledge of an innovation to forming an attitude toward the innovation, to a decision to adopt or reject, to implementation of the new idea, and to confirmation of this decision. An individual seeks information at various stages in the innovation-decision process in order to decrease uncertainty about an innovation's expected consequences.

Innovativeness

The second way in which time is involved in diffusion is in the innovativeness of an individual or other unit of adoption. Innovativeness is the degree to which an individual or other unit of adoption is relatively earlier in adopting new ideas than other members of a social system. There are five adopter categories, or classifications of the members of a social system on the basis on their innovativeness: (1) innovators, (2) early adopters, (3) early majority, (4) late majority, and (5) laggards.

Innovators

These are the first 2.5% of the individuals in a system to adopt an innovation. Venturesomeness is almost an obsession with innovators. This interest in new ideas leads them out of a local circle of peer networks and into more cosmopolite social relationships. Communication patterns and friendships among a clique of innovators are common, even though the geographical distance between the innovators may be considerable. Being an innovator has several prerequisites. Control of substantial financial resources is helpful to absorb the possible loss from an unprofitable innovation. The ability to understand and apply complex technical knowledge is also needed. The innovator must be able to cope with a high degree of uncertainty about an innovation at the time of adoption. While an innovator may not be respected by the other members of a social system, the innovator plays an important role in the diffusion process: that of launching the new idea in the system by importing the innovation from outside of the system's boundaries. Thus, the innovator plays a gatekeeping role in the flow of new ideas into a system.

Early adopters

These are the next 13.5% of the individuals in a system to adopt an innovation. Early adopters are a more integrated part of the local system than are innovators. Whereas innovators are cosmopolites, early adopters are localites. This adopter category, more than any other, has the greatest degree of opinion leadership in most systems. Potential adopters look to early adopters for advice and information about the innovation. This adopter category is generally sought by change agents as a local missionary for speeding the diffusion process. Because early adopters are not too far ahead of the average individual in innovativeness, they serve as a role-model for many other members of a social system. The early adopter is respected by his or her peers, and is the embodiment of successful, discrete use of new ideas. The early adopter knows that to continue to earn this esteem of colleagues and to maintain a central position in the communication networks of the system, they must make judicious innovation-decisions. The early adopter decreases uncertainty about a new idea by adopting it, and then conveying a subjective evaluation of the innovation to near-peers through interpersonal networks.

Early majority

This is the next 34% of the individuals in a system to adopt an innovation. The early majority adopt new ideas just before the average member of a system. The early majority interact frequently with their peers, but seldom hold positions of opinion leadership in a system. The early majority's unique position between the very early and the relatively late to adopt makes them an important link in the diffusion process. They provide interconnectedness in the system's interpersonal networks. The early majority are one of the two most numerous adopter categories, making up one-third of the members of a system. The early majority may deliberate for some time before completely adopting a new idea. 'Be not the first by which the new is tried, nor the last to lay the old aside,' fits the thinking of the early majority. They follow with deliberate willingness in innovations, but seldom lead.

Late majority

This is the next 34% of the individuals in a system to adopt an innovation. The late majority adopt new ideas just after the average member of a system. Like the early majority, the late majority make up one-third of the members of a system. Adoption may be the result of increasing network pressures from peers. Innovations are approached with a skeptical and cautious air, and the late majority do not adopt until most others in their system have done so. The weight of system norms must definitely favor an innovation before the late majority are convinced. The pressure of peers is necessary to motivate adoption. Their relatively scarce resources mean that most of the uncertainty about a new idea must be removed before the late majority feel that it is safe to adopt.

Laggards

These are the last 16% of the individuals in a system to adopt an innovation. They possess almost no opinion leadership. Laggards are the most localite in their outlook of all adopter categories; many are near isolates in the social networks of their system. The point of reference for the laggard is the past. Decisions are often made in terms of what has been done previously. Laggards tend to be suspicious of innovations and change agents. Resistance to innovations on the part of laggards may be entirely rational from the laggard's viewpoint, as their resources are limited and they must be certain that a new idea will not fail before they can adopt.

Rate of Adoption

The third dimension in which time is involved in diffusion is in rate of adoption. The rate of adoption is the relative speed with which an innovation is adopted by members of a social system. The rate of adoption usually is measured as the number of members of the system that adopt the innovation in a given time period. As shown previously, an innovation's rate of adoption is influenced by the five perceived attributes of an innovation.

The Social System

The fourth main element in the diffusion of new ideas is the social system. A social system is defined as a set of interrelated units that are engaged in joint problem-solving to accomplish a common goal. The members or units of a social system may be individuals, informal groups, organizations, and/or subsystems. The social system constitutes a boundary within which an innovation diffuses. How the system's social structure affects diffusion has been studied. A second area of research involved how norms affect diffusion. Norms are the established behavior patterns for the members of a social system. A third area of research has had to do with opinion leadership, the degree to which an individual is able to influence informally other individuals' attitudes or overt behavior in a desired way with relative frequency. A change agent is an individual who attempts to influence clients' innovation-decisions' in a direction that is deemed desirable by a change agency.

A fourth type of research involves the types of innovation-decisions (whether individual adoption decisions or organizational decisions, and whether they are made by an authority or by consensus). Another type ofresearch, conducted mainly by anthropologists, analyzed the consequences of innovation.

A crucial concept in understanding the nature of the diffusion process is the critical mass, which occurs at the point at which enough individuals have adopted an innovation that the innovation's further rate of adoption becomes self-sustaining. The concept of the critical mass implies that outreach activities should be concentrated on getting the use of the innovation to the point of critical mass. These efforts should be focused on the early adopters, the 13.5% of the individuals in the system to adopt an innovation after the innovators have introduced the new idea into the system. Early adopters are often opinion leaders, and serve as role-models for many other members of the social system. Early adopters are instrumental in getting an innovation to the point of critical mass, and hence, in the successful diffusion of an innovation.

Opinion Leaders

The two-step flow model of mass communication suggests that communication messages flow from a source, via mass media channels, to opinion leaders, who in turn pass them on to followers (Rogers, 1995). This two-step flow model of mass communication can be utilized to diffuse an innovation. The model focuses attention on the inter-media interface between mass media channels and interpersonal communication channels. Mass media channels are primarily creators of awareness-knowledge of innovations, while interpersonal networks are more important in persuading individuals to adopt or reject innovations.

The two-step flow model highlights the importance of opinion leaders in diffusing innovations, and especially in reaching the critical mass. Opinion leaders informally influence other individuals' attitudes or overt behavior in a desired way with relative frequency, so they are critical to the successful diffusion of innovations.

New Communication Technologies

We live in the age of a communication revolution. New and developing communication technologies, and, ultimately, equipment and systems, are influencing the communication industry and American society greatly. Take the Internet, for example. The Internet is a network of interconnected computer systems. By using the Internet as a data highway, an individual can communicate with colleagues and friends at great distance very cheaply, and gain direct access to information in other computer systems. Therefore, diffusion of the Internet has been exceedingly rapid. From 1990 to 2001, the number of individuals adopting the Internet in North America rose from approximately 5 million to 200 million. A huge percentage of the population is interconnected via the Internet. The critical mass in the diffusion of the Internet occurred in about 1989 or 1990, after 20 years of very slow increase in the rate of adoption of computer networks.

Most schools and local libraries are connected to the Internet, and can provide Internet access to individuals who otherwise might not own, or have access to, a computer. The potential for

using computer-mediated communication for outreach activities is enormous, and the nature of the diffusion process may be speeding up and becoming less spatially restricted.

See also: Diffusion in Sociological Analysis; Diffusion: Anthropological Aspects; Diffusion: Geographical Aspects; International Communication: History.

Bibliography

Pelto, J., 1973. The Snowmobile Revolution: Technology and Social Change in the Arctic. Cummings, Menlo Park, CA.

Rogers, M., 1995. Diffusion of Innovations, fourth ed. Free Press, New York.

Rosen, E., 2000. The Anatomy of Buzz: How to Create Word of Mouth Marketing. Doubleday/Currency, New York.

Ryan, B., Gross, C., 1943. The diffusion of hybrid seed corn in two Iowa communities. Rural Sociology 8, 15–24.

Van de Ven, H., Angle, H.L., Poole, M. (Eds.), 1989. Research on the Management of Innovation: The Minnesota Studies. Ballinger/Harper & Row, New York.

Zaltman, G., Duncan, R., Holbeck, J., 1973. Innovations in Organizations. Wiley, New York.

Evolution, History of

Peter J Bowler, Queen's University Belfast, Belfast, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article first outlines the various models of evolution that have been proposed, including the Darwinian theory and the most popular alternatives. It then outlines the early development of ideas about how life could have developed, including the Lamarckian theory of the inheritance of acquired characteristics. The development and publication of Darwin's theory is described, along with the controversies following its publication. Efforts to promote alternative theories in the late nineteenth century are described and the implications of the various theories for ideas about human origins and social evolution noted. The article concludes with the emergence of modern Darwinism and later debates such as those arising from sociobiology and evolutionary developmental biology.

The term 'evolution' is widely used to denote the development through time of societies, cultures, and more especially of living species. It is often contrasted with the view that these entities were divinely created as we see them today, and is routinely (but incorrectly according to modern biological theory) associated with the idea of progress. This article outlines the various models of evolution that have been suggested to account for the development of life and social organization, and then shows how the theories were formulated and popularized. Particular attention is paid to the work of Charles Darwin, whose theory of biological evolution by natural selection is now seen as the most influential expression of the basic idea of natural development. However, non-Darwinian evolutionary ideas also played a role in biology and were perceived to have implications for social evolution.

Models of Evolution

The term 'evolution' is derived from the Latin *evolutio*, denoting the unrolling of a scroll. In the eighteenth century it was applied to the growth of the embryo, then often described as the mere expansion of a preformed miniature. This application was retained up to the nineteenth century, by which time it was clear that the development of the embryo consisted of the progressive appearance of more complex structures. The philosopher and sociologist Herbert Spencer generalized the term by applying it to any process of natural progressive development, thereby creating the still common but inaccurate belief that all evolution must be progressive. In particular, Spencer applied the term to the development of societies and of life on earth.

Charles Darwin did not often use the term when describing his theory of the origin of species, but by the end of the nineteenth century, this had become the most common version of 'evolution.' However, Darwin's theory did not imply an inevitable progression toward complexity, and much controversy has surrounded the association between evolution and progress. Darwin's theory of natural selection now dominates biology and is being extended to many other domains; it is but one of a number of mechanisms that have been suggested to explain evolution, each of which has its own associated implications.

The Ladder of Progress

Perhaps the most popular model of evolution, and one that reveals the link with embryology, is of a linear hierarchy of stages of development from inception to maturity. Such a model simply assumes that all change is progressive and that there is only one direction of development. When applied to societies or species, the model can be adapted by supposing that many lines of development each ascend the same hierarchy but at different rates, so that the slow developers illustrate the ancestral stages already passed through by those at the top of the scale.

Evolutionary Trees

Although the ladder model is still popularly applied to the evolution of life on earth, and was once accepted in the social sciences, biologists and sociologists now realize that there is no single goal toward which evolution is developing – neither the human race nor Western civilization can be treated as the most mature phase of development. Evolution is better represented as a tree or a bush with many branches, each of which is constantly subdividing. The end product of one branch cannot be seen as the goal toward which all the others are ascending, and it is impossible for the end product of one branch to illustrate exactly the ancestral form of another branch. The living apes are not immature humans, because even if the common ancestor from which humans and apes have evolved would be classified as an ape, it would no longer be identical to any of the living ape species. If there is progress on the branching tree model, it has to be defined in more sophisticated terms so that there are many different ways to become more complex.

Darwin argued that branching evolution explains how we are able to classify species into groups within groups (species, genera, families, etc.). Two species descended from a common ancestor will still share many characters derived from that ancestor, and will share some characters with even more distant relatives. Darwin and his modern followers assume that each new character is developed only once, so all species exhibiting the character are descended from a common ancestor that also had the character. This assumption will be invalidated if there are cases where two lines of evolution evolve the same character independently (convergence or homoplasy). It will also be

 http://dx.doi.org/10.1016/B978-0-08-097086-8.81005-3

invalidated if there are even small-scale examples of ladderlike evolution, where several branches are constrained to evolve along the same pattern of development (parallelism).

Adaptation

The branching model had already been applied to the evolution of languages before Darwin and his followers used it to transform biology. Darwin's key insight was that there was no natural mechanism that could force evolution along a single predetermined path. Evolution is driven not by progress but by the necessity for each species to adapt to its environment in an ever-changing world. Branching occurs because a single population can sometimes become divided by a geographical barrier, after which each isolated population will develop independently in response to the different adaptive pressures it encounters. Progress to a more complex structure is at best a by-product of adaptation and is never inevitable in any one branch; indeed, many species are the products of degenerative evolution. In the animal kingdom, two branches of evolution can seldom if ever rejoin via hybridization (although this does happen in plants). Here, there is a clear contrast with social and linguistic evolution, where mergings and borrowings are frequent.

Darwinism and Lamarckism

Two mechanisms of adaptive evolution have been suggested: natural selection (Darwinism) and the inheritance of acquired characters (Lamarckism). Darwin assumed that the species consists of a population of organisms that exhibit some degree of individual variation and that these variations are inherited – we now explain them as genetic differences produced by mutations. He argued that if the population is exposed to a new environment, some variants will by chance be able to cope better with the new conditions; they will survive and reproduce, transmitting their favorable characters to the next generation. Any that are less fitted to the environment will be unlikely to breed and may even die, so their characters are eliminated from the population. Over many generations, this process of natural selection (Spencer called it the 'survival of the fittest') will change the species to adapt it to the environment.

Some time before Darwin, J.B. Lamarck proposed that the inheritance of characters acquired during an organism's lifetime could accumulate to give adaptive transmutation. An acquired character is produced by the organism's behavior, which in the wild is usually a response to the environment. Thus, the ancestral giraffes, striving to reach the leaves of trees as an alternative source of food, were able to lengthen their necks, just as a weightlifter can acquire better developed muscles. If such characters are inherited, the next generation of giraffes would be born with longer necks and would then stretch them further. This mechanism was once popular in biology, and corresponds quite closely to how humans transmit new ideas to their descendants. However, where teaching and learning provide a natural mechanism for ideas to be transmitted, modern genetics has shown that there is no way in which acquired biological characters can be inherited. Lamarckism has been banished from biology: The giraffe got its long neck because those who by chance inherited the genes producing that trait reproduced faster than those who carried the genes for shorter necks. The genetic differences are the product of mutations that are random as far as the requirements of the species are concerned.

Nonadaptive Evolution

There are also several mechanisms of biological evolution that suppose that not all developments are adaptive. In the nineteenth century, many naturalists believed that evolution was the unfolding of a predetermined pattern, perhaps analogous to or controlled by the forces of embryological development. Such ideas gave rise to the theory of recapitulation, which supposes that the evolutionary history of the species is traced out by the development of the embryo (or 'ontogeny recapitulates phylogeny'). This in turn suggests a model of evolution more in tune with the notion of a hierarchical ladder of development. The possibility that saltations (evolutionary jumps or macromutations) might have found new populations was once taken seriously even by geneticists. The theory of orthogenesis supposed that the species was somehow programmed to vary continuously in a particular direction, forcing groups of species to evolve in parallel. Such ideas are now largely discounted in biology, although the possibility that there might be constraints on the range of possible variation has recently been taken more seriously by the advocates of evolutionary developmental biology or 'evo-devo.' In the social sciences, the question of whether all cultural and social differences serve a practical purpose has been much debated, and this debate continues.

Pre-Darwinian Evolutionism

Darwin's was not the first theory of evolution, although the publication of his *Origin of Species* in 1859 triggered the conversion of most scientists to evolutionism. Several earlier versions of the idea were proposed, and although rejected by conservative thinkers and most scientists, they did much to prepare the way for the reception of Darwinism (Bowler, 1989c).

Creation and Design

The early Church Fathers did not take the Genesis story of creation in the Bible literally, and it was the Protestant theologians of the seventeenth century who established the worldview that survives in modern 'young earth' creationism. Archbishop James Ussher published his much ridiculed calculation that the earth was created in 4004 BC. In such a climate of opinion, a form of natural theology flourished in which it could be argued that God had designed and created all the modern species just as we see them today, each with the structures and instincts appropriate for its intended habitat. John Ray and many other naturalists stressed the adaptation of species to their environment, and argued that similarities could be used to classify species because they expressed the rational order of God's plan of creation – a view adopted by Carolus Linnaeus, who founded the modern system of biological classification (Greene, 1959). Others exploited the ancient concept of the 'chain of being,' in which the plan of creation was a linear hierarchy, with the human species at the top of the animal series. This was promoted without any sense of progression

through time, although the influence of the chain was retained in some of the early theories of progressive evolution.

Evolution in the Enlightenment

The more radical thinkers of the eighteenth-century Enlightenment recognized that the Earth was much older than Ussher's estimate and that the planet had changed considerably in the course of its history. Some explored the possibility that life had been produced by spontaneous generation from nonliving matter, and began to doubt that species remained unchanged through time. The French naturalist Georges Leclerc, Comte de Buffon, was especially influential in this respect, as were the materialist philosophers Denis Diderot and Baron d'Holbach (Roger, 1998).

Toward the end of the eighteenth century, Erasmus Darwin (grandfather of Charles) published a theory of transmutation in his *Zoonomia*, arguing that all living things were derived from a simple origin and had become more complex through their striving to adapt to the environment. These views were echoed in a more sophisticated form in J.B. Lamarck's *Zoological Philosophy* of 1809. Lamarck accepted that there was a basic progressive force driving living things to ascend the chain of being over many generations, but he accepted that no linear pattern is visible in the present animal and plant kingdoms, explaining the divergence as a consequence of life's constant need to adapt to an ever-changing environment. He invoked the inheritance of acquired characters (described above) to explain this process, and his name has been associated with that mechanism ever since (Corsi, 1988).

The Fossil Record

Historians used to think that Lamarck was marginalized by his great rival, Georges Cuvier, who pioneered the study of fossil bones to reconstruct the history of life. Cuvier showed that there was a succession of populations in the course of the earth's history but insisted that species disappear abruptly from the record through the action of geological catastrophes. He rejected transmutation but did not openly endorse the idea of divine creation – although the theory of a succession of creations was popularized by Cuvier's British followers, including William Buckland. Although they had extended the Earth's history, Cuvier and Buckland were convinced that the human race was a recent creation. Their work thus endorsed a view of human history in which there was no time for social evolution, many scholars still believing that language and the arts of civilization were taught to the earliest humans by their Creator.

Later historical work has shown that although the theory of successive creations was endorsed by the scientific establishment, more radical thinkers in both France and Britain were attracted to Lamarck's theory (Desmond, 1989). Darwin, who developed his theory in the 1830s, concealed it for fear of being associated with this radical movement. Historians also now recognize that there was an alternative non-Darwinian theory of evolution emerging in Germany among the followers of J.F. Blumenbach. This stressed the model of preordained development and invoked the parallel with embryological development. The situation in Britain changed significantly following the anonymous publication of Robert Chambers' *Vestiges of the Natural History of Creation* in 1844 (Secord, 2000). Chambers argued that an ideology of progress was just what the middle class needed to support its demands for reform, and provided a scientific foundation for this ideology based on the idea of progressive evolution. Although he presented evolution as the unfolding of a divine plan, his book's open recognition that humans are transformed animals, and that the mind is a product of the expansion of the brain, caused an outcry from conservative thinkers. Over the following decade, however, *Vestiges* popularized the idea of progressive evolution, creating a space in which Darwin's more scientific theory could have an impact.

The Development of Darwin's Theory

Charles Darwin was converted to evolutionism by his discoveries on the voyage of HMS *Beagle* (1831–36) and returned to England determined to understand how the process worked. He soon discovered the mechanism of natural selection and continued to work on the various aspects of his theory in relative secrecy. Much controversy has surrounded Darwin's discovery, because in addition to the many scientific factors influencing his thinking, he acknowledged the impact of Thomas Malthus' principle of population, a product of the ideology of laissez-faire economics.

Scientific Influences

While studying at Cambridge, Darwin was attracted to William Paley's Natural Theology of 1802, which restated the argument that species are adapted to their environment through the wisdom and benevolence of the Creator. On the voyage of the *Beagle*, his observations of South American geology converted him to Charles Lyell's uniformitarian position, which accounted for all geological change in terms of the action of observable forces over vast periods of time. Biogeographical studies, especially those of the birds of the Galapagos Islands, forced him to rethink the idea of Divine Creation, because he saw that individual islands had each acquired a distinct species of immigrant families, such as the finches. On his return to England, he soon accepted a complete evolutionary philosophy and began to search for a mechanism (Browne, 1995).

Darwin saw that evolution must be a branching process and that the ultimate cause was the adaptation of isolated populations to different environments. He studied animal breeding and realized that breeders produce new varieties by selecting out individual animals with the required character and breeding only from these. Artificial selection draws on the natural and largely random variability of the population and allows only individuals with a selected character to breed, thereby enhancing that character in future generations. Darwin was soon looking for a natural equivalent of this process.

Malthus and Population

A vital step in Darwin's thinking came when he read Malthus' *Essay on Population*, which argued that populations breed so rapidly that they outstrip their food supply. Drawing on hints in Malthus, Darwin realized that population pressure must generate a struggle for existence in which many individuals die.

He was now able to argue that any better adapted variant would be more successful in the struggle and would be more likely to survive and reproduce, while poorly adapted variants would die, thus establishing the theory of natural selection described in Section Darwinism and Lamarckism. This gave Darwin the basis on which to erect a comprehensive account of evolution. The role played by Malthus' ideologically loaded principle has, however, led to frequent charges that Darwin's theory is itself a projection of laissez-faire values on to nature, with the consequence that the theory would be used to justify the competitive ethos of Victorian capitalism (Young, 1985). Much controversy also surrounds the degree of Darwin's commitment to the idea of progress. Although the selection theory undermines the ladder model of progress, historians now accept that Darwin's thinking was influenced by the assumption that competition is the driving force of a more general form of progress. However, it is clear that Darwin's commitment to branching evolution went far beyond the conventional social values later linked to his theory, while his application of the population principle undermined Malthus' own ideology, which was an attempt to discredit the optimism of social progressionists.

Darwin let only a few close contacts know of his theory, while continuing to work on exploring its scientific implications. He came to appreciate that, even in a stable environment, specialization would allow natural selection to continue the process of divergence by which species separate from close relatives. Eventually, he began to write up his ideas for publication but was interrupted in 1858 by the arrival of a paper by Alfred Russel Wallace describing a similar (although not identical) theory. The joint Darwin–Wallace papers were published in that year, while Darwin rushed to complete the *Origin of Species* (Browne, 2002).

Darwinism in Biology

The *Origin of Species* sparked an intense debate that soon led to the conversion of most scientists and educated people to evolutionism (Hull, 1973; Ruse, 1979). Despite initial opposition by conservative thinkers, the belief that the human race was the outcome of a progressive trend built into nature was accepted. In science, too, creationism was replaced by the expectation that all living species are the products of natural processes, and here, too, there was a widespread assumption that such processes guaranteed progress, even if superimposed on a branching model of development. The success of Darwinism needs to be assessed with care because the popularity of this developmental model of evolution meant that the more radical implications of Darwin's thinking were evaded (Bowler, 1988). Far from welcoming the theory of natural selection, most nineteenth-century scientists, including some of Darwin's leading supporters, were lukewarm about it. Opposition grew toward the end of the century, with many alternative mechanisms of evolution being tried out.

The History of Life

Much of the first generation of evolutionists' efforts focused on reconstructing the course of the history of life on earth from anatomical, embryological, biogeographical, and fossil evidence (Bowler, 1996). Darwin himself treated this project with caution, but under the influence of the German biologist Ernst Haeckel, a whole generation of evolutionists strove to complete it. They had some successes, but the problems were much greater than they had anticipated and the project was sidelined by new developments in biology at the beginning of the twentieth century. Darwin had stressed the imperfection of the fossil record, but in a few important areas new discoveries helped to show how the major steps in evolution had taken place. Elsewhere, anatomical and embryological evidence was used to reconstruct evolutionary relationships. Darwin's theory implied that it would be unlikely for the same character to develop independently in different lines of evolution, so similarity of underlying structure could be taken as evidence of common descent. Some non-Darwinian theories, however, postulated predetermined trends that might generate similar structures independently. In embryology, Darwin's assumption that early stages of development would illustrate affinities was overtaken by Haeckel's recapitulation theory, according to which earlier stages could actually illustrate ancestral adult forms. This way of thinking was a product of the progressionist way of thought, presupposing a privileged line of development toward maturity. Only in biogeography did it become clear that evolution was a highly irregular process, depending on the hazards of migration in a world where geological forces were constantly remolding the topography.

Natural Selection and the Alternatives

Historians have focused on the debates over natural selection, but this is more a product of our modern preoccupation with that mechanism than of its significance in late nineteenth-century biology. One much discussed theme centers on Darwin's concept of heredity, which was quite unlike the model of unit character inheritance promoted by Mendelian genetics. It has even been argued that, without the concept of non-blending unit characters, natural selection was fatally flawed. It is true that Darwin was much concerned by Fleeming Jenkins' 1867 review, which argued that if male and female parental characters blend in the offspring, the advantages of a well-favored 'sport' or mutation would be diluted rapidly. However, as A.R. Wallace pointed out, this objection is irrelevant if there is a continuous range of variation in the population, which is the case for most characters (Gayon, 1998).

In fact, many of the objections to natural selection were intended only to show that evolution must be a more purposeful process than any mechanism based on random variation would allow. One major alternative that now became popular was the Lamarckian theory of the inheritance of acquired characters, in which the animals' own behavior directs their variation. Some objections to natural selection were aimed at the whole idea of adaptive evolution and were intended to show that some internally driven force directed the organisms' variation. These were very much a product of the rival developmental tradition that had become popular earlier in the century, especially in Germany. The theory of orthogenesis supposed that variation was controlled by internal developmental forces that generated linear evolutionary trends, while the theory of saltations assumed that macromutations could somehow find new breeding populations (Bowler, 1983).

Human Origins and Social Evolution

Many religious thinkers objected to the idea that the human soul was the product of evolution from an animal ancestry. Much of this initial opposition was overcome by stressing the purposeful nature of evolution, allowing the appearance of humanity to be presented as the culmination of a divine plan (Moore, 1979). Darwin's theory of evolution threatened this assumption, and in his *Descent of Man* of 1871, he argued that humans had evolved from African apes, which had stood upright as an adaptation to a new lifestyle on the open plains. Most late nineteenth-century accounts of human evolution, however, stressed the expansion of the brain as the main driving force of the process. At the same time, the idea of progressive evolution was applied to the history of human cultures and societies. Some evolutionists, including Spencer, stressed competition as the driving force of progress, leading to the charge that they were promoting a 'social Darwinism' designed to legitimize capitalist values. While Darwin's theory was certainly associated with this movement, it was by no means the only scientific justification offered.

Cultural Evolutionism

Coincidentally with the Darwinian revolution, archaeologists undermined the belief that the human race was a recent creation, and exposed a vast period of prehistory in which our ancestors had used only stone tools. It was this initiative rather than any influence from Darwinism that led anthropologists such as Edward B. Tylor to equate modern 'savages' (i.e., people with relatively unsophisticated technology) with the ancestral stages through which civilized humans had passed in prehistoric times on their march toward a more mature culture. Cultural evolutionism was a product of the nineteenth-century developmental viewpoint, based on the ladder model of progress (Bowler, 1989a). Tylor himself did not accept that humans had evolved from apes, but other cultural evolutionists, especially John Lubbock, linked the two modes of evolution and argued that 'savages' were biologically as well as culturally primitive – living examples of the earlier stages in human biological and mental evolution. Herbert Spencer's model of social evolution, while stressing ostensibly the divergent nature of evolution, still presented 'lower' races as surviving primitives incapable of matching the mental powers of the Anglo-Saxons.

Social Darwinism

Spencer is most often remembered, however, as the social philosopher who linked the idea of progress to the ideology of free enterprise individualism. As an exponent of laissez-faire, and the coiner of the term 'survival of the fittest,' it has been easy for critics to portray Spencer's enthusiasm for struggle as the motor of progress as a direct application of the Darwinian selection theory to society, hence the image of him as the founder of 'Social Darwinism.' Spencer worried about the elimination of the unfit, but in biology he defended Lamarckism, and his enthusiasm for struggle came more from a conviction that competition would stimulate individual self-improvement (which in a Lamarckian scheme could be passed on to future generations). Darwin's theory is thus only one of the scientific factors that stimulated the ideology of 'progress through struggle' – remembered preferentially because natural selection has now become the only mechanism by which that metaphor can be put into practice in biology (Bannister, 1979). In the later nineteenth century, there was more emphasis on the struggle between nations and races, which bypasses the central feature of natural selection within populations.

Modern Darwinism

By the end of the nineteenth century, the attempt to reconstruct the courses of biological and social evolution had run out of steam. The social sciences abandoned the evolutionary model and any suggestion that human behavior is predetermined by biological factors. Biologists began to study the actual processes of heredity and variation, leading to the emergence of Mendelian genetics. The geneticists destroyed Lamarckism, and although at first hostile to the selection theory, they were eventually convinced that genetic mutation was the source of the random variation that is the raw material of natural selection. By the mid-twentieth century, the 'Modern Synthesis' of Darwinism and genetics had come to dominate biology but remained controversial elsewhere.

The Modern Darwinian Synthesis

In the last decade of the nineteenth century, the statistician Karl Pearson developed a mathematical model of natural selection acting on the variations in a wild population (Gayon, 1998). Many biologists were now exploring the idea that new species were produced by saltations or macromutations, and it was partly from this tradition that the 'rediscovery' of Gregor Mendel's laws of particulate inheritance emerged. Ignored since the 1860s, these laws postulated the existence of unit characters transmitted unchanged from one generation to the next. In 1900, Mendel's laws were proclaimed as the basis for a new science of heredity, which William Bateson soon called 'genetics.' However, Pearson was already hostile to Bateson's saltationism, while Bateson rejected Pearson's commitment to continuity and adaptation. Genetics was thus perceived as a new alternative to Darwinism, although it destroyed the credibility of Lamarckism by suggesting that there is no plausible way in which acquired characters can affect the gene (Bowler, 1989b).

T.H. Morgan's studies of genetic mutations showed that most macromutations are fatal, while small mutations simply add to the genetic variation in the population. Soon, Morgan conceded that natural selection would affect the reproductive success of the various genes. In the 1920s and 1930s, R.A. Fisher, J.B.S. Haldane, and Sewall Wright brought together the statistical study of populations with the new genetics to create a science of population genetics based on the genetic theory of natural selection (Provine, 1971; Mayr, 1982). Naturalists such as Julian Huxley and Ernst Mayr now recognized that the adaptation of local populations to their environment was the most likely explanation of how species originate. The new selection theory was integrated with these wider developments to create what Huxley called the 'Modern Synthesis.'

Later Developments

The Darwinian synthesis has dominated biology since the mid-twentieth century, although there have been numerous developments within it and some efforts to limit its applicability. Some modern thinkers now argue that the concept of natural selection can be applied to almost all areas where change is observed, sweeping away the whole idea of a universe that is the product of divine purpose. Conversely, the emergence of evolutionary developmental biology (evo-devo) has renewed interest in the role of embryological factors in the creation of new characters and has thereby reinstated some topics that were central to the non-Darwinian programs of the nineteenth century (Amundson, 2005).

Perhaps the most controversial applications of modern biology are efforts to apply the theory of natural selection to human beings, especially in the area of social behavior. The science of sociobiology has been successful in explaining the social behavior of animals in terms of genetics and reproductive success, via the concept of the 'selfish gene' (Dawkins, 1976). However, E.O. Wilson's efforts to apply sociobiology to human behavior by suggesting that we are driven by instincts implanted by natural selection have been greeted with much suspicion by social scientists and have also been dismissed as a revived 'social Darwinism' by the political left (Segerstrale, 2000). In more recent decades, sociobiology has been replaced by evolutionary psychology, which postulates that human behavior is shaped by inbuilt behavioral constraints evolved originally to deal with the environment in which the earliest modern humans evolved (Barkow et al., 1992). This approach depicts the mind as a collection of more or less independent modules controlling behavior, rather than as a generalized information processor. It is opposed by rival methodologies including human behavioral ecology and a new form of cultural evolutionism, both of which focus on the malleability of behavior and our ability to respond to current environmental and social pressures. The question of how much evolutionary biology can tell us about operations of the human mind thus remains controversial.

See also: Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Darwin, Charles Robert (1809–82); Darwinism; Eugenics as a Basis of Population Policy; Eugenics, History of; Evolutionary Social Psychology; Malthus, Thomas Robert (1766–1834).

Bibliography

Amundson, R., 2005. The Changing Role of the Embryo in Evolutionary Thought: The Roots of Evo-Devo. Cambridge University Press, Cambridge.

Bannister, R.C., 1979. Social Darwinism: Science and Myth in Anglo-American Social Thought. Temple University Press, Philadelphia, PA.

Barkow, J.H., Cosmides, L., Tooby, J. (Eds.), 1992. The Adapted Mind. Oxford University Press, Oxford.

Bowler, P.J., 1983. The Eclipse of Darwinism: Anti-Darwinian Evolution Theories in the Decades around 1900. Johns Hopkins University Press, Baltimore, MD.

Bowler, P.J., 1988. The Non-Darwinian Revolution: Reinterpreting a Historical Myth. Johns Hopkins University Press, Baltimore, MD.

Bowler, P.J., 1989a. The Invention of Progress: The Victorian and the Past. Basil Blackwell, Oxford, UK.

Bowler, P.J., 1989b. The Mendelian Revolution: The Emergence of Hereditarian Concepts in Modern Science and Society. Athlone, London.

Bowler, P.J., 1989c. Evolution: The History of an Idea. University of California Press, Berkeley, CA.

Bowler, P.J., 1996. Life's Splendid Drama: Evolutionary Biology and the Reconstruction of Life's Ancestry, 1860–1940. University of Chicago Press, Chicago.

Browne, J., 1995. Charles Darwin: Voyaging. Jonathan Cape, London.

Browne, J., 2002. Charles Darwin: The Power of Place. Jonathan Cape, London.

Corsi, P., 1988. The Age of Lamarck: Evolutionary Theory in France, 1790–1830. University of California Press, Berkeley, CA.

Dawkins, R., 1976. The Selfish Gene. Oxford University Press, New York.

Desmond, A., 1989. The Politics of Evolution: Morphology, Medicine and Reform in Radical London. University of Chicago Press, Chicago.

Gayon, J., 1998. Darwinism's Struggle for Survival; Heredity and the Hypothesis of Natural Selection. Cambridge University Press, Cambridge, UK.

Greene, J.C., 1959. The Death of Adam: Evolution and Its Impact on Western Thought. Iowa State University Press, Ames, IA.

Hull, D.L., 1973. Darwin and His Critics: The Reception of Darwin's Theory by the Scientific Community. Harvard University Press, Cambridge, MA.

Mayr, E., 1982. The Growth of Biological Thought. Harvard University Press, Cambridge, MA.

Moore, J.R., 1979. The Post-Darwinian Controversies: A Study of the Protestant Struggle to Come to Terms with Darwin in Britain and America, 1870–1900. Cambridge University Press, Cambridge, UK.

Provine, W.B., 1971. The Origin of Theoretical Population Genetics. University of Chicago Press, Chicago.

Roger, J., 1998. The Life Sciences in Eighteenth-century French Thought. Stanford University Press, Stanford, CA.

Ruse, M., 1979. The Darwinian Revolution. University of Chicago Press, Chicago.

Secord, J., 2000. Victorian Sensation: The Extraordinary Publication, Reception and Secret Authorship of Vestiges of the Natural History of Creation. University of Chicago Press, Chicago.

Segerstrale, U., 2000. Defenders of the Truth: The Battle for Science in the Sociobiology Debate and Beyond. Oxford University Press, Oxford, UK.

Young, R.M., 1985. Darwin's Metaphor: Nature's Place in Victorian Culture. Cambridge University Press, Cambridge, UK.

Evolution, Natural and Social: Philosophical Aspects

Brian Skyrms, University of California, Irvine, CA, USA; and Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The development of theories of biological evolution and theories of social or cultural evolution is briefly surveyed. The question of empirical content is kept in the foreground. Contemporary theories of social evolution are at an earlier stage of development than contemporary theories of biological evolution. There is no counterpart to molecular genetics. However, contemporary theories of cultural evolution are not content free. They do generate some testable models.

In its most minimal sense, 'evolution' simply means gradual change by some natural process. In this sense, belief that species of plants and animals have evolved is at least as old as Aristotle's contrary belief in the fixity of the species. The theory of evolution acquires more content as the nature of the natural process involved is specified.

Darwin (1859) provided the broad outlines of the modern theory. There are traits that are inherited. There is some process that produces natural variation in these traits. The traits may affect the ability of the organism to reproduce, and thus the average number of individuals bearing the traits in the next generation. Therefore, those traits that enhance reproductive success increase in frequency in the population, and those that lead to below average reproductive success decrease in frequency. The three essential factors in Darwin's account are (1) natural variation, (2) differential reproduction, and (3) inheritance. The questions of the existence, extent, and nature of these factors are empirical scientific questions.

Could these processes, as they naturally occur, account for large changes over time? Darwin argued that they could by calling attention to the large changes that had come about from the selective breeding of plants and animals. Selective breeding is just a particularly severe form of differential reproduction, imposed by the breeder. Darwin emphasized this argument by calling his theory a theory of natural selection:

> I have called this principle, by which each slight variation, if useful, is preserved, by the term Natural Selection, in order to mark its relation to man's power of selection. But the expression often used by Mr. Herbert Spencer, of the Survival of the Fittest, is more accurate, and is sometimes equally convenient. We have seen that man by selection can certainly produce great results, and can adapt organic beings to his own uses, through the accumulation of slight but useful variations, given to him by the hand of Nature. But Natural Selection, as we shall hereafter see, is a power incessantly ready for action, and is as immeasurably superior to man's feeble efforts, as the works of Nature are to those of Art. (Darwin 1859, Origin of the Species, 6th edn., Chap. III)

The status of Spencer's description has been a matter of controversy. On superficial reflection on the slogan, survival of the fittest, more than one commentator has asked: "If fitness is fitness to survive, is not the theory of survival of the fittest simply a tautology?" We can give a preliminary examination of this idea in a highly idealized setting. Let us take Darwinian fitness just to mean expected number of offspring. Let us suppose that the population is effectively infinite, so that the expected number of offspring for a type can be taken to equal the average number of offspring. Let us suppose that reproduction is asexual and transmission of a type is perfect. If we had one generation every time unit, these assumptions specify a dynamics of population proportions called the discrete time replicator dynamics. Taking generations close together, or taking overlapping generations, it is possible to idealize to a system of differential equations, which represent the continuous time replicator dynamics (Hofbauer and Sigmund, 1988; Taylor and Jonker, 1978). This is an idealized model of differential reproduction. In this idealized model, one can ask whether it can be proved that the fittest type will go to fixation in the population.

To get any kind of affirmative answer, we must make more assumptions. Suppose that the fitness of a type is fixed. It does not depend on the frequency of types in the population. The environment is considered constant. Suppose that all the types of interest are initially present in the population and that one of them has the greatest Darwinian fitness. Then it is possible to prove that the dynamics will converge to a state of fixation of the fittest strategy, a state where its proportion of the population is 100%. In this sense, survival of the fittest can be mathematically demonstrated.

The demonstration depends not only on the conception of Darwinian fitness but also essentially on a battery of idealizations and assumptions. If the Darwinian fitness of a type is not fixed, but dependent on the frequency of other types in the population, then, even with all the other idealizations in place, the dynamics may not converge to any state at all. (It may cycle or even exhibit chaotic behavior.) If there is a fitter type not present in the population, then differential reproduction provides no way for it to get in. The model can be modified in various ways to allow for mutation, which may introduce the type in question into the population, but then the theorem is no longer true. If the population is finite, then only probabilistic conclusions can be reached. The fittest type may go extinct simply by bad luck. Moreover, in a finite population mutation cannot keep every possible type in play, but rather must be modeled as a chance process that introduces some new types from a very large space of possible types. With the slightest move toward realism the tautology goes away.

In the first half of the twentieth century, Darwinian theory was enriched by Mendelian genetics. Gregor Mendel, an Augustinian monk, had discovered basic principles of genetics

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63025-8

by a series of experiments in plant breeding carried out in the garden of the monastery (Mendel, 1866).

Mendelian genetics introduces a new source of variation in sexually reproducing species, but this variation has its limits. Can recombination of fixed genetic units account for the origin of new species and the whole sweep of Darwinian evolution? Darwinians thought not, and at the beginning of the twentieth century Darwinism and Mendelism were enemies. Only when mutation is included in the picture, so that more fundamental sorts of variation are possible on a longer timescale, is it evident that Mendelism and Darwinism can be thought of as reinforcing each other. This 'modern synthesis' of evolution and genetics was created in the first half of the twentieth century (Dobzhanski, 1937; Fisher, 1930; Haldane, 1932; Wright, 1968, 1969, 1977, 1978).

Mendelian genetics also enriches the account of differential reproduction. We now need to consider reproduction at more than one level. Individuals reproduce but so do genes. Fitness on one level is connected to that on the other, but in a way that can be complicated. The case of heterozygote superiority (as in the case of the sickle cell trait) provides another counterexample to the tautology claim discussed earlier. Heterozygotes cannot take over the population, simply because of the mechanics of genetic reproduction. Frequency-dependent selection reemerges at the level of the gene pool.

Mendelian genetics is a mathematical model (or a set of mathematical models). In 1953, Watson and Crick opened the door to understanding the physical basis of genetics at the molecular level. The dynamics of genetic replication, recombination, mutation, and expression can be studied at this level. The study of molecular genetics has provided many surprises such as the action of retroviruses that can transcribe their own information into the DNA of the host cell, the fact that bacteria can exchange genetic information, and the existence of plasmids – rings of DNA in bacteria that exist and replicate independently of the bacterial chromosome. With the theory of largely adaptively neutral mutations (Kimura, 1968, 1983), molecular genetics has also provided the basis of a molecular evolutionary clock, which can be used to reconstruct evolutionary history.

The history of the theory of biological evolution has been – and will continue to be – a story of the gradual amplification of Darwin's insights by empirical research. The three essential factors in Darwin's account – natural variation, differential reproduction, and inheritance – remain at the center of the theory, but in a form that Darwin could not have imagined.

By social evolution, we mean not the evolution of sociality, which may be one aspect of biological evolution, but rather a social process in some way analogous to the process of biological evolution. What is the analogy to fix upon? Different thinkers have had different aspects of biological evolution in mind (and sometimes, different conceptions of the nature of the biological process). In its most minimal sense, social and cultural evolution can just be thought of as social and cultural change. So interpreted, the theory of social evolution would be such a truism as to lack all interest. Are there deeper and more informative analogies that are worth exploring?

Some thinkers have thought of ideas that can spread through, or disappear from, a population as analogous to genes (Boyd and Richerson, 1985; Campbell, 1965; Cavalli-Sforza and Feldman, 1981; Lumsden and Wilson, 1981). These memes (Dawkins, 1976) can recombine in various ways. (A comprehensive discussion of this topic would have to range from nineteenth century associationalist psychology to modern genetic programming.) They can replicate themselves as they spread from mind to mind. They can be thought of as having a sort of Darwinian fitness in terms of the replicator process.

Others have focused on adaptive dynamics as the appropriate analogy. Darwinian evolutionary models fall, more or less, within a broad class of adaptive models, which also includes various kinds of learning models. In psychological theories of reinforcement learning, it is assumed that there is some 'reward' that is reinforcing, and the theory models the dynamics of behavior under various learning schedules. (The determination of what sort of thing is reinforcing is empirical, but extrinsic to the theory.)

We might then model various dynamical processes of social learning or adaptation instead of trying to model social evolution so as to have a close resemblance to biological evolution.

One common way for individuals to adapt in a social setting is to imitate others whom they observe to be successful according to some generally shared criterion of success. It is a remarkable fact that the replicator dynamics, introduced by Taylor and Jonker (1978) as a mathematical model of differential reproduction, has been derived from a number of different models of learning by imitation in a large population (Binmore et al., 1995; Björnerstedt and Weibull, 1996; Schlag, 1998). The replicator dynamics also has deep connections with classical psychological models of positive reinforcement learning (Borgers and Sarin, 1997; Beggs, 2005).

The replicator dynamics is a member of the broader class of adaptive dynamics according to which all behaviors with greater than average success increase their population proportion, which is a subset of the still broader class according to which the most successful behavior increases its population proportion. These broader classes of adaptive dynamics may also come into play in theories of social evolution (Cressman, 2003; Hofbauer and Sigmund, 2003; Samuelson, 1997; Sandholm, 2010; Weibull, 1995).

Populations may not be large, and encounters may not be random but rather structured. Learning by imitation in such settings may lead to different results than it would lead to in large random mixing populations, just as biological evolution may be affected by population structure and population size (Alexander and Skyrms, 1999; Anderlini and Ianni, 1997; Axelrod and Hamilton, 1981; Epstein, 1998; Frank, 1998; Grim, 1996; Hamilton, 1964; Lindgren and Nordahl, 1994; Nowak and May 1993; Pollack, 1989; Sober and Wilson, 1998; Skyrms, 2003). It is of particular interest that population structure can favor evolution or learning of cooperative behavior in interactions where it would not survive in large random mixing populations.

Classical reinforcement learning is the kind of learning that demands the least thought. A more sophisticated learner might use some kind of inductive reasoning to form beliefs about the aspects of the environment that, together with his/her actions, influence payoffs and then act optimally according to these beliefs. If the relevant aspects of the environment include the actions to be taken by other agents, then the reasoning involved may be complicated, depending on the strategic sophistication

of the agents and the character of the payoffs. (I am trying to predict what she will do. Is she trying to predict what I will do? If so what will she predict and what will she do and what is my best response? But might she be imagining that I am thinking this way? And so forth.) Here there is a whole spectrum of learning models, whose appropriateness depends on the inductive and strategic sophistication of the agents being modeled (Camerer and Ho, 1998; Fudenberg and Levine, 1998; Young, 2004).

If small probabilistic perturbations are added to deterministic dynamics of learning or evolution, then a new kind of equilibrium concept is appropriate, that of a stochastically stable equilibrium. In some cases, the nature of the underlying deterministic dynamics does not matter, and the stochastically stable equilibrium is determined by the payoff structure (Foster and Young, 1990; Kandori et al., 1993; Young, 1993, 2001). A small population with local interaction may reach stochastically stable equilibrium much faster than a large population with random interactions (Ellison, 1993). However, in the context of adaptive social networks, the usual stochastic stability results may be reversed (Skyrms and Pemantle, 2000).

The stage of development of the theory of social evolution is comparable to the earlier stages of the theory of biological evolution. There is no analog of molecular biology in view for social evolution. Rather, there are various mathematical models of learning that can be put to empirical test in special domains of application (Argiento et al., 2009; Blume et al., 1998; Camerer and Ho, 1998, 1999; Cavalli-Sforza et al., 1982; Friedman, 1996; Hewlett and Cavalli-Sforza, 1986; Roth and Erev, 1995; Skyrms, 1996, 2003, 2010; Suppes and Atkinson, 1960; van Huyck et al., 1995).

The processes of biological evolution and social evolution can interact (Boyd and Richerson, 1985; Boyd et al., 2010; Cavalli-Sforza and Feldman, 1981; Crawford, 1989; Richerson and Boyd, 2005). The biological evolution of humans as a social species sets the stage upon which social evolution is carried out. Social structure can feed back to biological evolution by imposing or modifying selective pressures. Basic cultural norms, for instance, the prevailing system of agriculture in a society, can be influenced by the biological evolution of other species. As both Darwin and Mendel observed, they can impose selective pressures on domesticated plants and animals. When cultural changes allow humans to survive in new or modified environments, selective pressures on humans themselves can be modified, as emphasized by Baldwin (1902). As humans continue to transform the planet, the biological effects of human cultural evolution become more and more extensive.

See also: Conventions and Norms: Philosophical Aspects; Cultural Evolution: Overview; Darwin, Charles Robert (1809–82); Darwinism: Social; Darwinism; Evolution, History of; Evolutionary Theory, Structure of; Game Theory: Noncooperative Games; Prisoner's Dilemma, One Shot and Iterated; Sociobiology: Overview.

Bibliography

Alexander, J.M., Skyrms, B., 1999. Bargaining with neighbors: is justice contagious? Journal of Philosophy 96, 588–598.

Anderlini, L., Ianni, A., 1997. Learning on a torus. In: Bicchieri, C., Jeffrey, R., Skyrms, B. (Eds.), The Dynamics of Norms. Cambridge University Press, New York.

Argiento, R., Pemantle, R., Skyrms, B., Volkov, S., 2009. Learning to signal: analysis of a micro level reinforcement model. Stochastic Processes and Their Applications 119, 373–390.

Axelrod, R., Hamilton, W.D., 1981. The evolution of cooperation. Science 211, 1390–1396.

Baldwin, J.M., 1902. Development and Evolution: Including Psychophysical Evolution, Evolution by Orthoplasty, and the Theory of Genetic Modes. MacMillan, New York.

Beggs, A., 2005. On the convergence of reinforcement learning. Journal of Economic Theory 122, 1–36.

Binmore, K.J., Gale, J., Samuelson, L., 1995. Learning to be imperfect: the ultimatum game. Games and Economic Behavior 8, 56–90.

Björnerstedt, J., Weibull, J., 1996. Nash equilibrium and evolution by imitation. In: Arrow, K., Columbotto, E., Parlman, M., Schmidt, C. (Eds.), The Rational Foundations of Economic Behavior. St. Martin's Press, New York, pp. 155–171.

Borgers, T., Sarin, R., 1997. Learning through reinforcement and the replicator dynamics. Journal of Economic Theory 77, 1–14.

Blume, A., DeJong, D.V., Kim, Y.G., Sprinkle, G., 1998. Experimental evidence on the evolution of meaning in messages in sender–receiver games. American Economic Review 88, 1323–1340.

Boyd, R., Richerson, P.J., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago.

Boyd, R., Richerson, P., Henrich, J., 2010. Gene-culture co-evolution in the age of genomics. Proceedings of the National Academy of Sciences (USA) 107, 8985–8992.

Camerer, C., Ho, T., 1998. Experience weighted attraction learning in coordination games: probability rules, heterogeneity and time-variation. Journal of Mathematical Psychology 42, 305–326.

Camerer, C., Ho, T., 1999. Experience-weighted attraction learning in normal form games. Econometrica 67, 827–874.

Campbell, D., 1965. Variation and selective retention in sociocultural evolution. In: Barringer, H.R., Blanksten, G., Mack, R.M. (Eds.), Social Change in Developing Areas: A Reinterpretation of Evolutionary Theories. Schenkman, Cambridge, MA.

Cavalli-Sforza, L.L., Feldman, M.W., 1981. Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press, Princeton, NJ.

Cavalli-Sforza, L.L., Feldman, M.W., Chen, K.H., Dornbush, S., 1982. Theory and observation in cultural transmission. Science 218, 19–27.

Crawford, V.P., 1989. Learning and mixed strategy equilibria in evolutionary games. Journal of Theoretical Biology 140, 537–550.

Cressman, R., 2003. Evolutionary Dynamics and Extensive Form Games. MIT Press, Cambridge, MA.

Darwin, C., 1859. On the Origin of the Species by Means of Natural Selection. J Murray, London.

Dawkins, R., 1976. The Selfish Gene. Oxford University Press, Oxford, UK.

Dobzhanski, T., 1937. Genetics and the Origin of Species. Columbia University Press, New York.

Ellison, G., 1993. Learning, local interaction and coordination. Econometrica 61, 1047–1071.

Epstein, J., 1998. Zones of cooperation in demographic prisoner's dilemma. Complexity 4, 36–48.

Fisher, R.A., 1930. The Genetical Theory of Natural Selection. Clarendon Press, Oxford, UK.

Foster, D., Young, P., 1990. Stochastic evolutionary game dynamics. Theoretical Population Biology 38, 219–232.

Frank, S.A., 1998. Foundations of Social Evolution. Princeton University Press, Princeton, NJ.

Friedman, D., 1996. Equilibrium in evolutionary games: some experimental results. Economic Journal 106, 1–25.

Fudenberg, D., Levine, D.K., 1998. The Theory of Learning in Games. MIT Press, Cambridge, MA.

Grim, P., 1996. Spatialization and greater generosity in the stochastic prisoner's dilemma. Biosystems 37, 3–17.

Haldane, J., 1932. The Causes of Evolution. Longmans, Green and Co, New York.

Hamilton, W.D., 1964. The genetical evolution of social behaviour. Journal of Theoretical Biology 7, 1–52.

Hewlett, B.S., Cavali-Sforza, L.L., 1986. Cultural transmission among Aka pygmies. American Anthropologist 88, 922–934.

Hofbauer, J., Sigmund, K., 1988. The Theory of Evolution and Dynamical Systems. 1998, second ed. Cambridge University Press, Cambridge, UK.

Hofbauer, J., Sigmund, K., 2003. Evolutionary game dynamics. Journal of the American Mathematical Society 40, 479–519.

Kandori, M., Malaith, G.J., Rob, R., 1993. Learning, mutation, and long run equilibria in games. Econometrica 61, 29–56.

Kimura, M., 1968. Evolutionary rate at the molecular level. Nature 217, 624–626.
Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK.
Lindgren, K., Nordahl, M.G., 1994. Evolutionary dynamics in spatial games. Physica D 75, 292–309.
Lumsden, C.J., Wilson, E.O., 1981. Genes, Mind and Culture. Harvard University Press, Cambridge, MA.
Mendel, G., 1866. Versuche über Pflanzenhybriden. Verhandlungen des Naturforscher-Vereins in Brünn. Tr. as [Experiments in Plant Hybridisation] 1960. Harvard University Press, Cambridge, MA.
Nowak, M., May, R., 1993. The spatial dimensions of evolution. International Journal of Bifurcation and Chaos 3, 35–78.
Pollack, G., 1989. Evolutionary stability of a viscous lattice. Social Networks 11, 175–212.
Richerson, P., Boyd, R., 2005. Not by Genes Alone: How Culture Transformed Human Evolution. University of Chicago Press, Chicago, IL.
Roth, A.E., Erev, I., 1995. Learning in extensive form games: experimental data and simple dynamic models in the intermediate term. Games and Economic Behavior 8, 164–212.
Samuelson, L., 1997. Evolutionary Games and Equilibrium Selection. MIT Press, Cambridge, MA.
Sandholm, W., 2010. Population Games and Evolutionary Dynamics. MIT Press, Cambridge, MA.
Schlag, K.H., 1998. Why imitate and if so, how? A boundedly rational approach to many multi-armed bandits. Journal of Economic Theory 78, 130–156.
Skyrms, B., 1996. Evolution of the Social Contract. Cambridge University Press, Cambridge, UK.
Skyrms, B., 2003. The Stag Hunt and the Evolution of Social Structure. Cambridge University Press, Cambridge, UK.
Skyrms, B., 2010. Signals: Evolution, Learning and Information. Oxford University Press, Oxford.
Skyrms, B., Pemantle, R., 2000. A dynamic model of social network formation. Proceeding of the National Academy of Sciences (USA) 97, 9340–9346.
Sober, E., Wilson, D.S., 1998. Unto Others: The Evolution and Psychology of Unselfish Behavior. Harvard University Press, Cambridge, MA.
Suppes, P., Atkinson, R.C., 1960. Markov Learning Models for Multiperson Interactions. Stanford University Press, Stanford, CA.
Taylor, P.D., Jonker, L.B., 1978. Evolutionarily stable strategies and game dynamics. Mathematical Biosciences 40, 145–156.
van Huyck, J., Battalio, C., Mathur, S., Ortmann, A., van Huyck, P.P., 1995. On the origin of convention: evidence from symmetric bargaining games. International Journal of Game Theory 24, 187–212.
Wright, S., 1968, 1969, 1977, 1978. In: Evolution and the Genetics of Populations, vols 4. University of Chicago Press, Chicago, IL.
Weibull, J.W., 1995. Evolutionary Game Theory. MIT Press, Cambridge, MA.
Young, H.P., 1993. The evolution of conventions. Econometrica 61, 57–84.
Young, H.P., 2001. Individual Strategy and Social Structure: An Evolutionary Theory of Institutions. Princeton University Press, Princeton, NJ.
Young, H.P., 2004. Strategic Learning and Its Limits. Oxford University Press, Oxford.

Evolutionary Approach to Education Science

David C Geary, University of Missouri at Columbia, Columbia, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Evolution has resulted in a system of evolved cognitive abilities that coalesce around the domains of folk psychology, biology, and physics. These cognitive domains direct attention toward and process information about other people, the flora and fauna in the local ecology, and the physical environment. Among the associated competencies are language, the ability to interpret facial expressions, knowledge of the behavior of familiar animals, and the ability to navigate in the physical world. The basic architecture of these competencies is universal and adapted to the nuances of local conditions as children play, interact with other people, and explore the environment. In modern societies, people must also acquire a host of cognitive competencies that have not evolved, including reading, writing, and basic mathematics. The learning of these nonevolved competencies requires adapting evolved cognitive systems for tasks for which they are not designed. The associated activities (e.g., direct instruction) appear to be very different from the activities (e.g., play) that foster the adaptation of evolved abilities to local conditions. Schooling thus involves the society-wide organization of children's activities, such that they learn competencies that would not otherwise emerge.

During the past decade, evolutionary analyses of human behavior, social relationships, and cognition have become increasingly common within psychology, albeit not without continued resistance from some social scientists. The goals here are to consider how theory and research in evolutionary psychology can be used to understand children's cognitive and academic development and to explore related educational issues, a subfield now known as evolutionary educational psychology (Geary, 2012a). This approach is illustrated with discussion of how evolved cognitive abilities, specifically language, can be modified through schooling to create evolutionarily novel ones, specifically reading, and with discussion of the attendant implications for understanding children's achievement motivation. Before these issues can be fully appreciated, an overview of the evolutionary approach to cognition and development is needed.

Evolution and Cognition

The basic assumption of evolutionary psychologists is that the human brain and cognitive abilities evolved to attend to, process, and guide behavioral response to evolutionarily significant information that coalesces around folk psychology, folk biology, and folk physics (see Geary, 2005). These biologically primary or core domains are predicted to be modular to the extent that they are biased to process specific forms of information; and they are predicted to be plastic to the extent that sensitivity to variation in the corresponding information patterns provides an advantage. A bias to orient to the human face and to automatically process key pieces of facial information (e.g., eyes) is an example of a modular primary ability (Kanwisher et al., 1997) that in this case is a component of folk psychology. If the ability to discriminate one person from the next affords a social advantage over those who cannot make these discriminations, then this modular system will evolve some degree of plasticity, but within constraints.

A taxonomy of folk modules is presented in **Figure 1** (Geary, 2005). The function of these folk systems is to focus behavior on attempts to achieve access to and control of the social, biological, and physical resources that tended to enhance survival or reproductive prospects during human evolution. Achieving control is not an easy task, of course, and if it occurs it is typically without conscious awareness of the corresponding evolved function. The development of friendships is supported by folk-psychological competencies, but does not appear to be guided by a motivation to control the behavior of these people, or at least there is often no explicit desire to do so. Nevertheless, having these relationships provides the social support that is correlated with physical and psychological health and in some contexts mortality risks (e.g., Taylor et al., 2000). These friendships are social resources that can enhance survival (e.g., support during times of social conflict) and reproductive prospects (e.g., a romantic partner) under the types of conditions found in traditional societies and presumably throughout human evolution.

Folk Psychology

The folk-psychological modules represent systems that process information related to the self, other individuals, and group dynamics, respectively. The first set includes awareness of the self as a social being and awareness of one's relationships with other people. Self-awareness is related to the ability to mentally project the self backward in time to recall and relive episodes that are of personal importance and to project oneself forward in time to create a self-centered mental simulation of potential future states that become goals to achieve (Suddendorf and Corballis, 1997). The individual level modules process the forms of information that guide one-on-one social dynamics and foster one-on-one social relationships (Bugental, 2000). The group-level modules enable individuals to break their social world into categories of people, including kin and members of favored in-groups and disfavored out-groups. People also have the unique ability to form in-groups on the basis of ideology, such as nation.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92010-5

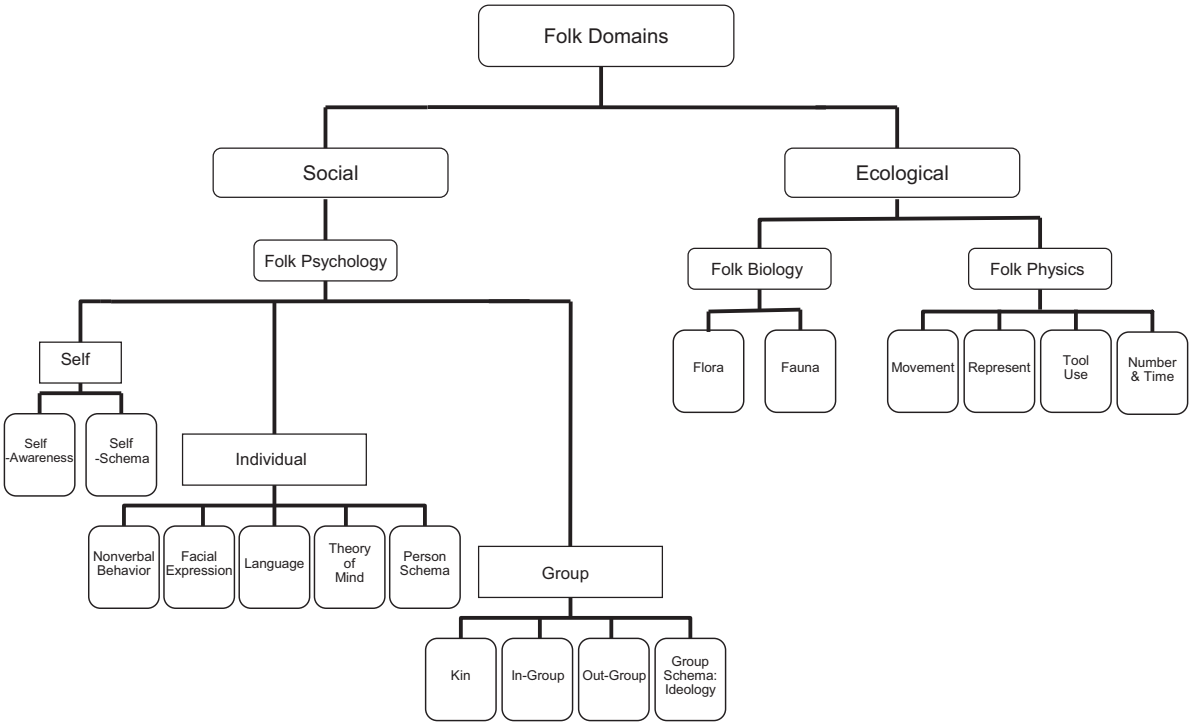

Figure 1 System of evolved modular systems in the domains of folk psychology, folk biology, and folk physics. Adapted from Geary, D.C., 2005. Origin of Mind: Evolution of Brain, Cognition, and General Intelligence. American Psychological Association, Washington, DC, p. 129.

Folk Biology and Physics

The folk-biological modules orient attention toward important features of the biological world, such as movement patterns of potential prey, and support the ability to develop taxonomies of other species and knowledge systems about the behavior, growth, and 'essence' (something akin to their 'spirit') of these species that in traditional societies support hunting and horticulture (Atran, 1998). The folk-physical systems support navigation, the formation of mental representations of physical features of the ecology, the construction of tools, and an implicit understanding of magnitude, number, and time. Some of these competencies, especially the ability to navigate, are similar to those found in other species and thus are not uniquely human (Tomasello and Call, 1997). The ability to construct and use tools, in contrast, far exceeds the competencies found in chimpanzees (*Pan troglodytes*) and other species (Povinelli, 2000).

Heuristics and Attributional Biases

Gigerenzer and colleagues' (Gigerenzer et al., 1999) heuristics represent 'rule of thumb' patterns of behavior that are evoked based on the implicit (below conscious awareness) processing of key ecological or social cues. For instance, in many species of bird, females are attentive and sexually responsive to the bright plumage of males of the same species without an explicit understanding that this coloration signals genetic health or the ability to provision offspring (Hill and McGraw, 2006). Females and their offspring thus benefit from mating with these males, resulting in the evolution of a rule of thumb ('mate with colorful males') heuristic for these species of bird; Simon (1956) referred to these implicit relations as bounded rationality, as illustrated in Figure 2. Many human biases (e.g., preference for symmetric faces in mates) reflect similar implicit heuristics. For humans,

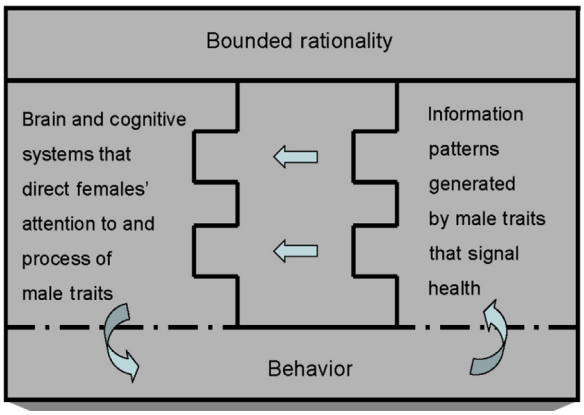

Figure 2 Bounded rationality is a coupling of ecological and social information patterns that occurred regularly during the species' evolutionary history and complementary brain and cognitive systems that have evolved to direct the organisms' attention to and process this information. The coupling can result in automatic behavioral responses that act on the ecological or social information, although people can override these biases and respond based on conscious evaluation of the information. Adapted from Geary, D.C., 2005. Origin of Mind: Evolution of Brain, Cognition, and General Intelligence. American Psychological Association, Washington, DC, p. 173.

heuristics can also include explicit inferential and attributional biases that are integral features of folk knowledge. The attributional biases that favor members of the in-group and derogate members of out-groups are well known (Stephan, 1985) and facilitate coalitional competition (Horowitz, 2001) Kahneman and Tversky's (1982) simulation heuristic involves use of the person schema to mentally simulate and make judgments about how the individual might react in one situation or another. The essence associated with folk biology allows people to make inferences (e.g., during the act of hunting) about the behavior of members of familiar species, as well as about the likely behavior of less familiar but related species (Atran, 1998). Attributions about causality in the physical world have also been studied. Children and adults have, as an example, natural, naive conceptions about motion and other physical phenomena (Clement, 1982).

Evolution and Cognitive Development

Over evolutionary time, the human developmental period has increased from about 11–13 years to 18–20 years (Bogin, 1999). The lengthening of this period comes with a profound risk – death before the age of reproduction – and thus would only evolve if there were benefits that outweighed this risk. Studies of other primates with comparatively long periods of development suggest that an important adaptive benefit is the accompanying ability to refine the physical, social, and cognitive competencies that support survival and reproduction. A long developmental period is found in all social mammals and the length of this period increases with increases in the complexity of the species' social system (Joffe, 1997). These patterns suggest that one critical purpose of childhood is to practice and refine folk-psychological competencies, such as using language to influence the behavior of other people.

Play, social interactions, and exploration of the environment and objects appear to be the mechanisms through which these emerging competencies are practiced and refined during development. Child-initiated social play, exploration, and so forth are intimately linked to cognitive and brain development, in that these activities provide experiences with the social, biological, and physical world. These experiences interact with the inherent but skeletal structure of folk modules and ensure their normal development and adaptation to local conditions. As an example, infants' inherent orientation to human faces will provide the experiences necessary to recognize their parents and to elicit caregiving and investment from these parents. In this view, children are biologically prepared to learn about other people and the biological and physical world, and are inherently motivated to seek out experiences that will facilitate this learning (Greenough et al., 1987; Scarr, 1992). The result is that these folk systems are fine-tuned to the nuances of the social group and ecology in which they live.

Implications for Education

Evolved folk abilities and associated behavioral strategies are sufficient for coping with the environments of our ancestors, but these abilities and strategies are not well suited to all contemporary conditions. This is a result of humans' extraordinary ability to create evolutionarily novel knowledge and competencies; these include reading, writing, and arithmetic (Geary, 2005, 2007). During the last several thousand years the cross-generational accumulation of evolutionarily novel knowledge (e.g., through books) has occurred at such a rapid pace (Richerson and Boyd, 2005) that the attentional and cognitive biases that facilitate the adaptation of folk abilities to local conditions during children's natural activities do not have evolved counterparts to facilitate the learning of secondary abilities to the same degree. The emergence of formal and mandatory schooling is a reflection of this mismatch and is a recent cultural innovation that is designed to better prepare children for living in the evolutionarily novel context, we call modern society.

Much of formal education is 'unnatural' in that what children are expected to learn in school involves tasks never encountered by our ancestors (Geary, 1995; Rozin, 1976). The basic goals of schools and schooling are thus to organize the activities of children so that they acquire competencies, such as the ability to read, and scientific and technical knowledge that lead to ways of understanding the world that differ from evolved heuristics and attributional biases. This is so that they can better function in a wider culture that contains many demands (e.g., need to understand compound interest) that have no evolutionary history (e.g., Geary, 2012b).

Evolution and Academic Development

Geary (1995) referred to language and other folk modules as *biologically primary abilities*, and skills that build upon these primary abilities but are principally cultural inventions, such are reading, as *biologically secondary abilities*. The mechanisms by which evolved systems are adapted to produce secondary competencies include the prefrontal cortex, the attentional and inhibitory control components of working memory, and the problem-solving competencies defined by fluid intelligence, but discussion is beyond the scope of this article (see Geary, 2005). One important point here is that because primary modular systems are inherently modifiable, they can – with sustained effort and practice – be modified or combined in unique ways to create biologically secondary abilities. The second point is that the creation of secondary abilities will require the explicit and effortful representation of the to-be-learned information in working memory, and oftentimes the inhibition of folk heuristics and attributional biases (Geary, 2012b).

The relation between language, a primary ability, and reading, a secondary ability, provides an example. Reading and writing are the principal means through which secondary knowledge has accumulated during the past several thousand years (i.e., through books) and are critical to the acquisition of this knowledge in school. From an evolutionary perspective, reading and writing systems initially emerged from motivation to communicate with and attempt to influence the behavior of other people, and therefore it is not surprising that they engage at least some of the same folk-psychological systems (e.g., Mann, 1984; Rozin, 1976). The following sections illustrate how this form of analysis can be used more generally with an evolutionary educational

psychology approach to academic learning, by placing learning to read and reading comprehension in the context of several folk-psychological domains.

Cognitive Mechanisms

The cognitive predictors of children's reading acquisition, the effectiveness of various types of reading instruction, and sources of individual differences in ease of learning to read support the prediction that many core reading competencies are dependent on language systems (e.g., Bradley and Bryant, 1983; Mann, 1984). The crucial early components include phonemic awareness – explicit awareness of distinct language sounds – and the ability to decode unfamiliar written words into these basic sounds. Decoding requires an *explicit* representation of the sound (e.g., *ba*, *da*, *ka*) in phonological working memory and the association of this sound, as well as blends of sounds, with corresponding visual patterns that are also explicitly represented in working memory (specifically, letters (e.g., *b*, *d*, *k*) and letter combinations, respectively (Bradley and Bryant, 1983)).

Children with a strong explicit awareness of basic language sounds are more skilled than are other children at associating these sounds with the symbol system of the written language and thus learn to read more quickly than other children (Wagner et al., 1994). Unlike natural language learning, the majority of children acquire these competencies most effectively with systematic, organized, and teacher-directed (or parent-directed) explicit instruction on phoneme identification, blending, and word decoding (e.g., Hindson et al., 2005). Skilled reading requires text comprehension, that is, an understanding of the meaning of the composition which in turn is dependent on several component skills; for instance, locating main themes and distinguishing highly relevant from less relevant passages. As with more basic reading skills, many children require explicit instruction in the use of these strategies to aid in text comprehension (Connor et al., 2004).

Another evolution-based prediction is that reading comprehension will be dependent in part on theory of mind and other folk-psychological domains (**Figure 1**), at least for literary stories, poems, dramas, and other genre that involve human relationships (Geary, 2007). Most of these stories involve the recreation of social relationships, more complex patterns of social dynamics, and even elaborate person schema knowledge for main characters.

Brain Mechanisms

The neuroanatomy of language is well known and overlaps considerably with the brain regions that support phonological decoding, reading fluency, and text comprehension (Price and Mechelli, 2005). Price (2000) concluded that passive processing of language sounds occurs in the traditional Wernicke's area (part of the temporal cortex); speech production involves Broca's area and areas that support word articulation (part of the frontal cortex); and, the representations of the meaning of spoken and heard utterances is distributed across the temporal and parietal cortices. Prefrontal working memory areas related to language are engaged when the speaker or listener has to make active decisions about the utterances (Poldrack et al., 1999).

Among other regions, tasks that tap phonological awareness or involve phonological decoding during the act of reading engage Wernicke's and Broca's areas, as do tasks that involve the reading of single words. Word reading also engages areas in the temporal and parietal cortices that support the comprehension of spoken utterances (Price and Mechelli, 2005). Early research on the comprehension of the syntax and the basic meaning of read sentences suggested engagement of the same brain regions that support the production and comprehension of spoken utterances, including Broca's and Wernicke's areas (Just et al., 1996).

These findings are consistent with the view that instruction in school, with reading as an example here, involves the modification of the brain and cognitive systems that support folk abilities in the creation of evolutionarily novel abilities that are culturally relevant. The example here touches only the surface of the complexity of the creation of biologically secondary abilities (for elaboration see Geary, 2005, 2007, 2012b).

Motivation to Learn

One very important implication of the evolutionary perspective is that the motivation to acquire school-taught secondary abilities is based on the requirements of the larger society and not always on the inherent interests of children. Given the recent emergence of near universal schooling in modern societies, there is no reason to believe that the skills that are taught in school are inherently interesting or enjoyable for all children to learn. This does not, however, preclude the self-motivated engagement in some secondary activities. Even though reading is a secondary ability that involves the modification of primary language systems, many children and adults are motivated to read. This motivation, however, is probably driven by the content of what is being read rather than by the process itself. Furthermore, the finding that intellectual curiosity is a basic dimension of human personality suggests that there will be a number of curious individuals who will pursue secondary activities (Berlyne, 1966). The more critical point is that the motivation to engage in the activities that will promote the acquisition of secondary abilities is not likely to be universal (e.g., Larson and Richards, 1998); in fact, curiosity is often fleeting and influenced by current knowledge, interests, and other factors (Loewenstein, 1994).

Traditional studies of achievement motivation include children's understanding of the relation between effort and ability on academic outcomes (Nicholls, 1984); valuation of academic learning in terms of mastery (i.e., motivated desire for a deep understanding of the material) or performance (e.g., standing relative to others) goals (Ames and Archer, 1988; Dweck and Leggett, 1988); academic self-efficacy (e.g., Bandura, 1993); and expectancy of success and attributions regarding the sources of success or failure (e.g., ability vs bad luck) in achieving academic goals (Weiner, 1990; Wigfield and Eccles, 2000). An integration of these models and the corresponding literature within an evolutionary framework is beyond the scope of this article, but I illustrate that such integration can be achieved by considering Bandura's (1997) model of social and cognitive self-efficacy.

"People make causal contributions to their own functioning through mechanisms of personal agency. Among the mechanisms of agency, none is more central or pervasive than people's beliefs about their capabilities to exercise control over their own level of functioning and over events that affect their lives" (Bandura, 1993: p. 118). Self-efficacy is an aspect of this personal agency and at its core is a self-referenced appraisal regarding the likelihood of success in various domains and through these influences, among other things, the pursuit of achievement in these domains and persistence in the face of failure. Bandura emphasizes one's explicit appraisal of efficacy and attributions regarding associated outcomes (e.g., cause of failure). From an evolutionary perspective, these map onto the folk-psychological domains of self-awareness, self-schema, and the ability to explicitly represent associated information in working memory and to problem solving on the basis of this information (see Geary, 2005, 2007).

In other words, Bandura's (1993, 1997) model of self-efficacy is consistent with an evolutionary perspective, but with different points of emphasis regarding children's academic motivations and corresponding self-evaluations. One area of difference is with respect to the prediction of domain-specific and inherent learning and motivational biases associated with folk knowledge and a much weaker motivational bias for evolutionarily novel academic learning such as mathematics. Further, the core of the self-schema is predicted, from an evolutionary perspective, to be referenced in terms of one's standing vis-à-vis peers and, importantly, more so for traits that have an evolutionary history than for culture-specific knowledge and abilities. The former includes physical abilities and attractiveness, social influence, and family status (Geary, 2010). These are predicted to be universal and to have substantial influences on the development of self-schemas and self-evaluations. This is contrasted with culturally specific activities, such as schooling, that are predicted to be important in these cultures but less central to most children's and adolescents' emerging self-schemas and evaluations. In keeping with this view is the finding that self-awareness and the emerging self-schema are embedded in a web of social relationships and that the best predictor of global self-esteem from childhood to adulthood is perceived physical attractiveness (Harter, 1998, 2006) and not, for instance, grades in high school mathematics classes.

From an evolutionary perspective, the valuation of academic achievement and the relation between achievement and self-esteem is predicted to be highly variable across and within cultures, and to be heavily dependent on explicit parental and cultural valuation of associated activities and outcomes (e.g., grades; Stevenson and Stigler, 1992), and heavily influenced by peers' valuation of academic achievement (Harris, 1995). From a social learning perspective (Bandura, 1986), many children will imitate parents and teachers who engage in academic activities (e.g., reading); many will come to focus on these activities because they provide access to culturally valuable resources such as a job and income; and, many will come to enjoy these activities in their own right, developing a mastery orientation (Winner, 2000). Children and adolescents will also develop a sense of academic self-efficacy in cultures with formal schooling. These outcomes also follow from an evolutionary perspective that includes evolved modes of cross-generational knowledge transmission.

One important difference comparing the social learning and evolutionary perspectives is with respect to the specificity of predictions: For instance, with successive grades, academic content will increasingly diverge from its evolved foundation, and thus academic learning is predicted to become more difficult and any motivation to engage in this learning is predicted to decrease; and this is the case (Eccles et al., 1993). Social living also becomes more complex and nuanced as people mature into adulthood, but motivational disengagement from social life is predicted to be far less common than disengagement from academic life. These differences follow logically from an evolutionary perspective, but less readily from a strictly social learning perspective.

If our goal is universal education that encompasses a variety of evolutionarily novel academic domains (e.g., mathematics) and abilities (e.g., phonetic decoding as related to reading), then we cannot assume that an inherent curiosity or motivation to learn will be sufficient for most children and adolescents. Children's and adolescent's explicit valuation of academic learning, the perceived utility of academic skills, and the centrality of self-efficacy in these areas to their overall self-esteem is predicted to be highly dependent on social–cultural valuation of academic competencies, such as explicit rewards for academic achievement (e.g., honor rolls) and valuation of cultural innovators (e.g., Edison). These types of cultural and social supports are critical because the lack of inherent motivation to learn in secondary domains and the difficulty of this learning are also predicted to result in negative attitudes and avoidance (e.g., of mathematics) for some students. In contrast, the child's and adolescent's valuation and perceived efficacy of their physical traits or social relationships are implicit features of their evolved folk psychology and will manifest with or without cultural supports, contrary to current assumptions regarding the source of the focus on these traits (e.g., Harter, 1998).

Implications for Instructional Activities

The disconnect between children's evolved motivational biases and the motivational focus needed to excel in modern academic domains and the gap between primary abilities and secondary abilities that is rapidly accumulating in modern societies is, at first blush, rather discouraging. Our lives and those of our children would be considerably easier, if all children found academic learning as engaging as they find interacting with their peers and as effortless as learning their native language. The value of an evolutionary approach to education is not only to better understand the roots of motivational and learning challenges that children face in modern schools, but also to provide a different perspective on what might be done to facilitate children's engagement of and success in academic learning.

There may be some conditions in which children's social play and discourse can be used to facilitate the acquisition of academic skills. The key may be to capitalize on the fuzzy boundary between the primary and secondary domains during the early years of schooling and children's motivation

to learn culturally important knowledge and to use these to build academic self-efficacy and other beliefs that will help to maintain effort and motivation in school learning in later years. Children's natural interest in novelty and their motivation to learn their culture will get them started but is not likely to maintain long-term academic learning.

Children's evolved biases, such as social play, and their observation of and learning from culturally successful adults result in the automatic and effortless modification of folk systems and the learning of the forms of culturally important activities, such as hunting, that have a long evolutionary history. For these domains and activities, there are corresponding motivational biases that ensure that children engage in the necessary activities. If we assume that children are inherently motivated to learn in secondary domains and learn them as effortlessly as they learn in primary domains, then we risk undervaluing the importance of focus and effort for secondary learning. Without an explicit assumption that learning will require effort, we risk having children assume that school learning will occur effortlessly. When effort is required for learning and children begin to experience failure, they are at risk for making attributions that may undermine their later engagement with school.

In closing, an evolutionary approach to children's learning and motivation in school can explain why learning to read is more difficult for children than learning their native language, among many other primary–secondary contrasts (Geary, 2007, 2012a, 2012b), and why many children are more motivated to socialize with their friends than to learn algebra. This perspective has profound implications for how to design instruction to address these differences between primary and secondary learning and to better understand and address the motivational and attributional (e.g., that learning will require effort) factors that will influence long-term engagement in academic learning.

See also: Cognitive Anthropology; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Cognitive Development in Infancy, Neural Mechanisms of; Cognitive Development: Child Education; Cognitive Development: Mathematics Learning and Instruction; Human Spatial Orientation, Neural Basis of; Language Development: Evolutionary Theories; Mind, Theories of; Neural Plasticity; Play, Anthropology of; Prefrontal Cortex Development and Development of Cognitive Function; Psychological Anthropology; School Achievement as Moderated by Sociodemographic Status: The US Case; Self-Efficacy: Education Aspects; Self-Regulated Learning: Theories, Measures, and Outcomes.

Bibliography

ABC Research Group, 1999. In: Gigerenzer, G., Todd, P.M. (Eds.), Simple Heuristics That Make Us Smart. Oxford University Press, New York.

Ames, C., Archer, J., 1988. Achievement goals in the classroom: students' learning strategies and motivation processes. Journal of Educational Psychology 80, 260–267.

Atran, S., 1998. Folk biology and the anthropology of science: cognitive universals and cultural particulars. Behavioral and Brain Sciences 21, 547–609.

Bandura, A., 1986. Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice-Hall, Englewood Cliffs, NJ.

Bandura, A., 1993. Perceived self-efficacy in cognitive development and functioning. Educational Psychologist 28, 117–148.

Bandura, A., 1997. Self-Efficacy: The Exercise of Control. Freeman, San Francisco.

Berlyne, D.E., 1966. Curiosity and exploration. Science 153, 25–33.

Bogin, B., 1999. Patterns of Human Growth, second ed. Cambridge University Press, Cambridge, UK.

Bradley, L., Bryant, P.E., February 3, 1983. Categorizing sounds and learning to read – a causal connection. Nature 301, 419–421.

Bugental, D.B., 2000. Acquisition of the algorithms of social life: a domain-based approach. Psychological Bulletin 126, 187–219.

Clement, J., 1982. Students' preconceptions in introductory mechanics. American Journal of Physics 50, 66–71.

Connor, C.M., Morrison, F.J., Petrella, J.N., 2004. Effective reading comprehension instruction: examining child X instruction interactions. Journal of Educational Psychology 96, 682–698.

Dweck, C.S., Leggett, E.L., 1988. A social–cognitive approach to motivation and personality. Psychological Review 95, 256–273.

Eccles, J., Wigfield, A., Harold, R.D., Blumenfeld, P., 1993. Age and gender differences in children's self- and task perceptions during elementary school. Child Development 64, 830–847.

Geary, D.C., 1995. Reflections of evolution and culture in children's cognition: implications for mathematical development and instruction. American Psychologist 50, 24–37.

Geary, D.C., 2005. The Origin of Mind: Evolution of Brain, Cognition, and General Intelligence. American Psychological Association, Washington, DC.

Geary, D.C., 2007. Educating the evolved mind: conceptual foundations for an evolutionary educational psychology. In: Carlson, J.S., Levin, J.R. (Eds.), Educating the Evolved Mind, Psychological Perspectives on Contemporary Educational Issues, vol. 2. Information Age, Greenwich, CT, pp. 1–99.

Geary, D.C., 2010. Male, Female: The Evolution of Human Sex Differences, second ed. American Psychological Association, Washington, DC.

Geary, D.C., 2012a. Evolutionary educational psychology. In: Harris, K.R., Graham, S., Urdan, T. (Editors-in-chief), Educational Psychology Handbook. In: Theories, Constructs, and Critical Issues, vol. 1. American Psychological Association, Washington, DC, pp. 595–620.

Geary, D.C., 2012b. The evolved mind and scientific discovery. In: Shrager, J., Carver, S.M. (Eds.), From Child to Scientist: Mechanisms of Learning and Development. American Psychological Association, Washington, DC, pp. 87–115.

Greenough, W.T., Black, J.E., Wallace, C.S., 1987. Experience and brain development. Child Development 58, 539–559.

Harris, J.R., 1995. Where is the child's environment? A group socialization theory of development. Psychological Review 102, 458–489.

Harter, S., 1998. The development of self representations. In: Eisenberg, N. (Vol. Ed.), Social, Emotional, and Personality Development. In: Damon, W. (Gen. Ed.), Handbook of Child Psychology, vol. 3, fifth ed. John Wiley & Sons, New York, pp. 1017–1095.

Harter, S., 2006. The self. In: Eisenberg, N. (Vol. Ed.), Social, Emotional, and Personality Development. In: Damon, W., Lerner, R. M. (Gen. Eds.), Handbook of Child Psychology, vol. 3, sixth ed. John Wiley & Sons, New York, pp. 505–570.

Hill, G.E., McGraw, K.J. (Eds.), 2006. Bird Coloration: Function and Evolution, vol. 2. Harvard University Press, Cambridge, MA.

Hindson, B., Byrne, B., Shankweiler, D., Fielding-Barnsley, R., Newman, C., Hine, D.W., 2005. Assessment and early instruction of preschool children at risk for reading disability. Journal of Educational Psychology 97, 687–704.

Horowitz, D.L., 2001. The Deadly Ethnic Riot. University of California Press, Berkeley, CA.

Joffe, T.H., 1997. Social pressures have selected for an extended juvenile period in primates. Journal of Human Evolution 32, 593–605.

Just, M.A., Carpenter, P.A., Keller, T.A., Eddy, W.F., Thulborn, K.R., October 4, 1996. Brain activation modulated by sentence comprehension. Science 274, 114–116.

Kahneman, D., Tversky, A., 1982. The simulation heuristic. In: Kahneman, D., Slovic, P., Tversky, A. (Eds.), Judgment Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, London, pp. 201–208.

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. Journal of Neuroscience 17, 4302–4311.

Larson, R., Richards, M., 1998. Waiting for the weekend: friday and saturday night as the emotional climax of the week. New Directions for Child and Adolescent Development 82, 37–51.

Loewenstein, G., 1994. The psychology of curiosity: a review and reinterpretation. Psychological Bulletin 116, 75–98.
Mann, V.A., 1984. Reading skill and language skill. Developmental Review 4, 1–15.
Nicholls, J.G., 1984. Achievement motivation: conceptions of ability, subjective experience, task choice, and performance. Psychological Review 91, 328–346.
Poldrack, R.A., Wagner, A.D., Prull, M.W., Desmond, J.E., Glover, G.H., Gabrieli, J.D.E., 1999. Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. NeuroImage 10, 15–35.
Povinelli, D.J., 2000. Folk Physics for Apes: The Chimpanzees Theory of How the World Works. Oxford University Press, New York.
Price, C.J., 2000. The anatomy of language: contributions from functional neuroimaging. Journal of Anatomy 197, 335–359.
Price, C.J., Mechelli, A., 2005. Reading and reading disturbance. Current Opinion in Neurobiology 15, 231–238.
Richerson, P.J., Boyd, R., 2005. Not by Genes Alone: How Culture Transformed Human Evolution. University of Chicago Press, Chicago, IL.
Rozin, P., 1976. The evolution of intelligence and access to the cognitive unconscious. In: Sprague, J.M., Epstein, A.N. (Eds.), Progress in Psychobiology and Physiological Psychology, vol. 6. Academic Press, New York, pp. 245–280.
Scarr, S., 1992. Developmental theories of the 1990s: developmental and individual differences. Child Development 63, 1–19.
Simon, H.A., 1956. Rational choice and the structure of the environment. Psychological Review 63, 129–138.
Stephan, W.G., 1985. Intergroup relations. In: Lindzey, G., Aronson, E. (Eds.), Handbook of Social Psychology, Special Fields and Applications, vol. II. Random House, New York, pp. 599–658.
Stevenson, H.W., Stigler, J.W., 1992. The Learning Gap: Why Our Schools are Failing and What We can Learn from Japanese and Chinese Education. Summit Books, New York.
Suddendorf, T., Corballis, M.C., 1997. Mental time travel and the evolution of the human mind. Genetic, Social, and General Psychology Monographs 123, 133–167.
Taylor, S.E., Klein, L.C., Lewis, B.P., Gruenewald, T.L., Gurung, R.A.R., Updegraff, J.A., 2000. Biobehavioral responses to stress in females: tend-and-befriend, not fight-or-flight. Psychological Review 107, 411–429.
Tomasello, M., Call, J., 1997. Primate Cognition. Oxford University Press, New York.
Wagner, R.K., Torgesen, J.K., Rashotte, C.A., 1994. Development of reading-related phonological processing abilities: new evidence of bidirectional causality from a latent variable longitudinal study. Developmental Psychology 30, 73–87.
Weiner, B., 1990. History of motivational research in education. Journal of Educational Psychology 82, 616–622.
Wigfield, A., Eccles, J.S., 2000. Expectancy-value theory of achievement motivation. Contemporary Educational Psychology 25, 68–81.
Winner, E., 2000. The origins and ends of giftedness. American Psychologist 55, 159–169.

Evolutionary Approaches to Human Behavior: Philosophical Aspects

Ben Fraser and Kim Sterelny, Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by K. Sterelny, volume 21, pp. 14543–14549, © 2001, Elsevier Ltd.

Abstract

Evolutionary approaches to human behavior date back to Charles Darwin, though they were somewhat eclipsed in the first half of the twentieth century. They began to become more prominent, as evolutionary studies of animal behavior matured as a scientific discipline in the middle of the twentieth century. The sociobiological approach prevalent in the 1970s and 1980s has been succeeded by a variety of approaches, including human behavioral ecology, evolutionary psychology, and gene–culture coevolution or dual inheritance theory. This article details the inception, prominent proponents, key features, representative case studies, and major criticisms of each of these three current approaches. It also notes the current trends in the field, toward greater use of computer models and experimental methods in studying human behavioral evolution.

Intellectual Background

The idea that it could be illuminating to take an evolutionary approach to understanding humans – our behaviors as well as our bodies – dates back at least to Charles Darwin. In his concluding remarks to *The Origin of Species*, Darwin predicted that, in the course of future developments of the theory he had just presented, "light will be thrown on the origin of man and his history" (1859: p. 647). Darwin followed up this brief allusion to human evolution in the *Origin* with a far more substantial treatment of the topic several years later in *The Descent of Man* and *The Expression of the Emotions in Man and Animals*. There, he defended what can be called a 'continuity' thesis: "the difference in mind between man and the higher animals, great as it is, certainly is one of degree and not of kind" (1871: p. 151). That is to say, the impressive mental abilities that underlie humans' complex ecological, social, and technological behavior – the abilities to imitate, reason, invent, remember, imagine, and more – do not constitute a deep divide between humans and the rest of the natural world. Rather, Darwin thought, such abilities have precursors in, and gradually evolved out of, more basic capacities found in other animals.

One idea of Darwin's has recently become the focus of renewed interest and debate: Darwin's appeal to group selection in his explanation of human mental and behavioral evolution. In particular, in speaking specifically about the human capacity for morality and its effectiveness in underwriting cooperative and altruistic behavior, Darwin claimed that "a high standard of morality gives but a slight or no advantage to each individual man and his children over the other men of the same tribe [but] an increase in the number of well-endowed men and an advancement in the standard of morality will certainly give an immense advantage to one tribe over another" (1871: p. 157). The key idea here is that the evolution of certain traits that are puzzling in terms of the individual bearer's success can be explained in terms of the advantages those traits provide to the bearer's group. Such group selectionist thinking was influential in evolutionary biology for a long time after Darwin, receiving a comprehensive modern statement in Vero Wynne-Edwards' (1962) work, *Animal Dispersion in Relation to Social Behaviour*. While Darwin's suggestion remains very controversial – though recently revived in contemporary guise by Samuel Bowles and Herbert Gintis (2012) – the puzzle of human cooperation remains very prominent in recent work.

The adoption in the 1960s of the so-called gene's-eye view saw the decline of group selectionist explanations (Williams, 1966; Maynard-Smith, 1964; Dawkins, 1976), though as noted above, the idea has made a return (Sober and Wilson, 1998). To take the gene's-eye view is to consider, in a given set of circumstances, what characteristics a gene would need to have in order to increase its frequency over generational time. The problem this perspective reveals in group-selectionist thinking is that groups of cooperative individuals, for instance, are vulnerable to undermining from within. In such a group, an individual with a gene that led it to take the benefits of others' cooperative behavior while refraining from itself paying the costs of cooperating would do better than its cooperative conspecifics. Over generational time, such individuals would outcompete and come to replace cooperators, and cooperation itself would collapse.

In the 1960s, William Hamilton pioneered an alternative approach to the evolution of social behavior (there were also important other theorists: especially John Maynard-Smith, George Williams, and Robert Trivers). Hamilton saw that genes for cooperative behavior could spread by promoting the success of kin, likely to carry the same genes, even when cooperative behavior imposed a cost on the cooperative agent. These ideas were originally developed to explain the extraordinary levels of cooperation among social insects, but were then extended to other forms of social life, including human social life, in a set of ideas that came to be known as human sociobiology, and that was prominently advocated by E.O Wilson in his *Sociobiology: The New Synthesis* (1975).

As noted above, an important element in this new tool kit was a reconceptualization of the notion of fitness: Hamilton showed that cooperating to help kin may evolve, despite decreasing the helper's 'personal' fitness provided it generates a compensating increases in the helper's 'inclusive fitness' (1964: p. 8). That is to say, if the cost of helping decreases the number of genes the helper contributes to the next generation

in the form of its own offspring by less than the help increases the number of genes (identical by descent to the helper's) passed on by the helper's relatives, then helping makes evolutionary sense. This inequality has come to be known as Hamilton's Rule, and this way of accounting for the evolution of certain kinds of cooperative social behavior, as kin selection.

Not all social behavior involves interactions between kin; humans, in particular, do not typically live in groups composed solely of close kin. Robert Trivers' (1971) theory of reciprocal altruism did much to help explain cooperative behavior outside the family (see also Axelrod, 1984). Trivers realized that such behavior could be fitness-advancing for the helper so long as several conditions were met. First, there must be an asymmetry between the cost and the benefit of helping, with the cost of helping being small compared to the benefit helping confers. Second, there must be the potential for individuals to interact repeatedly and to switch roles (i.e., helper vs. beneficiary). Third, individuals must have the ability to recognize those who have helped them in the past. Fourth and finally, individuals must have a tendency to help those who have helped them in the past and to withhold help from those who have failed to help them. Under these conditions, the immediate cost of helping another individual can be recouped (and more) over time as one's beneficiaries repay one's helpfulness in kind. Cooperation under these conditions can thus work out to a helper's long-term advantage.

Trivers' proposed explanation for apparently fitness sacrificing helpfulness has come to be called 'direct' reciprocity, since it involves pairs of individuals that trade favors with each other. Richard Alexander (1987) extended Trivers' work by suggesting that reciprocity might also be indirect, in which case "the return is expected from someone other than the recipient of the beneficence" (1987: p. 85). Alexander explained the fitness value of reputation, and of the value of investing in having a good reputation. Reciprocity, both direct and indirect, was another important explanatory resource for sociobiologists attempting to explain the evolution of human social behavior. Indirect reciprocity might be especially important in the human case, for it is not clear that Trivers' mechanism scales up to cooperative interactions that depend on more than two agents, for withholding help then becomes a very blunt instrument in response to cooperation failure: it imposes costs on the whole coalition rather than on the specific individuals who fail to cooperate (Sterelny, 2012).

The final chapter of Wilson's (1975) *Sociobiology* was devoted to humans, as was his Pulitzer Prize-winning *On Human Nature* (1978). In these publications, Wilson drew on the resources described above to develop sociobiological explanations for numerous behaviors he took to be human universals, including religious and ritual performance, the creation of art, engaging in trade, making war, and acting ethically.

The specifics of the sociobiologists' claims, and of the critical response to sociobiology, have been covered in great detail elsewhere (Segerstrale, 2000; Laland and Brown, 2002; Kitcher, 1985). More generally: (1) Sociobiology was behavior-oriented: the focus was on particular kinds of behavior, such as those mentioned above, rather than on the psychological mechanisms that generated those behaviors. The guiding assumption was that particular behaviors could be tuned by selection independently of one another, and hence human sociobiology was in increasing tension with the developing cognitive psychology, which saw particular behaviors as the result of interactions among cognitive mechanisms, each of which contributed to many behaviors, in many different contexts. (2) Sociobiology was adaptationist: the guiding aim was to explain its target behaviors as the results of natural selection, that is, as adaptations. (3) Sociobiology was, often, speculative: it had about it the air of just-so storytelling (Kitcher, 1985), with claims about evolution of certain human behaviors resting on appealing, perhaps plausible, but evidentially undersupported suggestions about adaptive advantage. This speculative nature of human sociobiology was particularly significant given that many of its claims were politically loaded, such as claims about homosexuality and territorial violence. (4) Finally, sociobiology was primarily concerned with identifying and explaining human behavioral similarities – the supposed human universals – not with diversity in human behavior.

Changes in Focus

Three current major evolutionary approaches to human behavior are detailed below. All of these approaches can be understood in contrast with – in some cases as reactions to or developments of – sociobiology.

The first approach, human behavioral ecology, represents a shift in focus compared to sociobiology, away from seeking universals and on to understanding diversity. The second, evolutionary psychology, was a self-conscious break with sociobiology for at least some of its practitioners, and contrasts with that earlier approach insofar as it focuses not on behavior itself, but on the psychological mechanisms that underlie and generate behavior. Finally, the third approach detailed below, gene–culture coevolution, represents a move away from sociobiology's gene-centric and adaptationist viewpoint by attempting to integrate cultural with genetic influences on human behavioral evolution, and also by recognizing the potential for cultural forces to lead to maladaptive outcomes over evolutionary time.

Current Knowledge

Human Behavioral Ecology

Behavioral ecologists attempt to understand the adaptive rationale of animal behavior. The discipline emerged in the 1960s and 1970s from the ethology that Tinbergen and Lorenz had established as a valid branch of the biological sciences, by integrating the formal machinery of evolutionary biology with the experimental and field methods of ethology. Human behavioral ecology is the application of those methods to humans. The discipline first appeared in the 1970s, and was pioneered and subsequently developed mostly by anthropologists, such as William Irons, Napoleon Chagnon, Irven DeVore, Kristen Hawkes, and Monique Borgerhoff Mulder.

The first key feature of human behavioral ecology is its emphasis on humans' capacity for adaptive behavioral plasticity in the face of environmental variation. Human behavioral ecologists make what has been dubbed the 'phenotypic gambit'

(Grafen, 1984), essentially betting that genetic constraints on human behavior are sufficiently loose that humans can flexibly adjust their behavior to maximize reproductive success across a wide range of different environments. Human behavioral ecologists' models assume that humans can recognize and assess, reasonably accurately, the options available to them, and that they prefer options (more food, more sex, safety, friendship) which as a matter of fact typically promote their fitness interests; subjective utility and objective interests mostly match. But their models typically leave these cognitive assumptions tacit: they explicitly focus on identifying the costs and benefits of different choices.

The second key feature of human behavioral ecology is its use of formal models. Optimality models in particular are central to this approach to human behavioral evolution. The process by which such models are employed is as follows. First, model construction, in which the behavior of interest is defined, and its costs and benefits, along with the relevant environmental parameters, are specified. The purpose of the model is to identify the behaviors that maximize fitness under the given environmental conditions. Next comes data collection. Here, the human behavioral ecologist operates much like a traditional anthropologist, directly observing members of a subject population, conducting interviews, and drawing on historical material, all with the aim of discovering how the members of that population actually behave in their environment. Finally, model testing occurs when the predictions of the optimality model are compared to the actual behaviors to determine whether the real-world people are behaving so as to maximize their fitness.

The particular model that a human behavioral ecologist uses will depend on the target behavior of interest. Models commonly used include optimal foraging theory, life history theory, sex allocation theory, and the polygyny threshold model.

Eric Alden Smith's (1985) research on the Inuit is an example of the use of optimal foraging theory to investigate human foraging practices. Optimal foraging models seek to quantify the costs and benefits of various foraging decisions – about prey choice, time allocation, patch choice, movement, and group formation, among others – in order to assign to the various possible foraging strategies a payoff that can serve as an adequate fitness proxy, such as net rate of energy capture. The guiding assumption is that foragers will act in ways that maximize this measure. Smith's model predicted that optimal group size for hunting geese was one, while optimal group size for hunting seals was three.

The Inuit did indeed hunt geese alone, but when hunting seals most often formed groups of four and sometimes as many as eight – far from optimal according to the model. Smith interestingly interpreted his results as partial confirmation for the view that the Inuit tailored their foraging behavior in ways that maximized their fitness, and also as an indication of the operation of previously unconsidered constraints on Inuit foraging behavior. In particular, Smith thought social factors, such as the costs of excluding individuals from group hunts, constrained optimization for some forms of foraging. This case study thus shows both how human behavioral ecology's guiding assumption of human adaptive behavioral flexibility, together with the formal modeling methods employed, can find confirmation in actual observed human behaviors, and also how even failures can be fruitful in suggesting refinements to models for future research.

Other examples of human behavioral ecology in action include the use of sex allocation theory and the polygyny threshold model to assess the adaptiveness of mating practices and patterns of parental investment in offspring. For example, Rebecca Bliege Bird (2007) has developed an elegant model to explain turtle hunting among the Meriam in North East Australia: male turtle hunting is (she argues) a signal of social and physical skill rather than mere resource acquisition.

Human behavioral ecology is not without its critics. Donald Symons (1990) criticized human behavioral ecologists' focus on the adaptiveness of behaviors – whether the behaviors were currently fitness enhancing – rather than on whether the behaviors were adaptations, that is, the products of a history of natural selection. Being adaptive and being an adaptation can come apart: not all currently adaptive traits are adaptations, and not all adaptations are currently adaptive. So, showing that certain behaviors are adaptive does not show that they are adaptations. There are two issues lurking here. First, how, exactly, does showing (say) that male turtle hunting among the Meriam is adaptive *explain* that form of social behavior? Given that behaviors themselves are not the result of natural selection, how does the adaptiveness of turtle hunting stabilize that social practice? Second, the fact that being adaptive and being an adaptation can come apart calls into question the phenotypic gambit. If human selective history and the adaptations it has shaped place significant constraints on human behavioral flexibility, then people now may sometimes – perhaps often – fail to behave in ways that maximize fitness. After all, there seem to be plenty of prima facie examples of biologically maladaptive behavior; for example, much of the western world prefers wealth to children. If true, this would undermine the guiding assumption of human behavioral ecology. Symons and Tooby are representatives of an approach to human behavioral evolution – evolutionary psychology – on which this possibility is thought actual. This approach is detailed below.

Evolutionary Psychology

Evolutionary psychology was conceived in the 1980s and championed by anthropologist John Tooby alongside Leda Cosmides, a psychologist. Cosmides and Tooby distinguished their approach to understanding human evolution from that of Wilson and the sociobiologists, and from the human behavioral ecologists, by focusing on psychological mechanisms for producing behavior rather than directly on behaviors themselves. Cosmides and Tooby considered the focus on behavior a "conceptual 'wrong turn'" (1987: p. 278), and one they aimed to correct. They were joined in this endeavor by the likes of Donald Symons, David Buss, Margo Wilson, Martin Daly, Steven Pinker and Dan Sperber.

The key features of evolutionary psychology are as follows. First, the approach is explicitly adaptationist, in that it supposes that the psychological mechanisms underlying human behavior are adaptations. Secondly, it stresses the fact that human psychological mechanisms for generating behavior were designed by selection to solve recurrent adaptive problems in

the evolutionary environment of our ancestors – the environment of evolutionary adaptation, or EEA – and hence, in slogan form, our modern skulls house Stone Age minds. This form of evolutionary psychology is a version of classical, nativist cognitive psychology: these psychological mechanisms are computational, information-processing mechanisms, and they are designed and prewired to solve specific, crucial, recurrent informational challenges faced by Pleistocene foragers. Methodologically, evolutionary psychologists first employ functional analysis to 'reverse engineer' the mind, meaning that they hypothesize adaptive problems faced in our evolutionary past, postulate solutions to those problems, then conduct studies much like nonevolutionary psychologists to determine whether the postulated mental mechanisms actually exist. The method of reverse engineering the mind leads evolutionary psychologists to a core commitment of their approach, namely, the 'massive modularity' thesis, according to which the human mind is made up of a large collection of individual mechanisms, 'modules', each designed to solve a distinct adaptive problem. Since they are dedicated to specific tasks, these mechanisms are said to be domain-specific. Finally, evolutionary psychologists claim that this suite of modules constitutes an evolved and universal human nature.

Leda Cosmides' (1989) work investigating how people perform across different versions of the Wason Selection Task provides a representative example of evolutionary psychology in action. Cosmides began by positing an adaptive problem important in human evolutionary history, namely, that of processing social information, and in particular detecting violations of social rules or cheating in social exchanges. She hypothesized that humans have evolved a mental module specialized for solving this adaptive problem.

To test this hypothesis, Cosmides had subjects attempt different versions of the Wason Selection task. In this task, subjects are told a conditional rule of the form 'if P then Q' and asked to determine when this rule may have been violated. Subjects receive information in the form of four cards. Each card is double-sided, one side stating whether the antecedent of the conditional is true or false (i.e., stating either P or not-P), the other side stating whether the consequent is true or false (i.e., stating either Q or not-Q). The subjects' information is incomplete, though; subjects can see only one side of each card. The task is to identify which cards must be turned over in order to determine whether the situation the card represents constitutes a violation of the rule. Logically, only a 'P and not-Q' combination can falsify the conditional rule, hence, the correct choice in the task is to turn over the P card (to see if it has a not-Q on the reverse) and the not-Q card (to check for a P).

When given nonsocial versions of the task, subjects performed poorly – only about 25% made the logically correct card choices – but, when the rule in question was a social rule, such as for instance "if a person is drinking alcohol, then s/he must be over 21," performance jumped markedly, with roughly 75% of subjects choosing to turn over the logically correct cards (in this case, the 'drinking beer' and 'under 21' cards). Cosmides interpreted this pattern of results as support for her hypothesis that humans have mental mechanisms specialized for detecting cheating and violation of social rules.

Other examples of evolutionary psychology in action include comparisons of male vs female mate preferences (Buss, 1989), studies of patterns of child abuse by step as compared to birth parents (Daly and Wilson 1999), and studies of the evolution of rape (Thornhill and Palmer 2000).

Many criticisms have been leveled at evolutionary psychology. One challenge comes in the form of attempts to explain the evolutionary psychologists' own data in ways that do not confirm their hypotheses. For instance, even some researchers who fit neatly enough into the evolutionary psychologists' own camp disputed the Wason selection task findings and interpretation (Sperber and Wilson, 1986). More recently, a prominent example of this kind of criticism is the work of David Buller, who in *Adapting Minds* (2005) takes on evolutionary psychologists case-by-case and offers alternative explanations for the data reported.

To take a specific example, evolutionary psychologists claim that females, more than males, should value high-status mates because such mates are more capable of providing resources during childbearing and -raising (important adaptive problems to be sure). Evolutionary psychologists such as David Buss (1989, 2005) offer data supposedly showing sex differences in mate preference, with females showing greater preference for high-status mates, and claim that these data reflect different evolved solutions to the different adaptive problems facing men and women in this arena. Buller considers the evolutionary psychologists' data and argues that an alternative explanation, namely, assortative mating by status, better explains that data than does the evolutionary psychologists' own hypothesis about evolved sex differences in preferences for status (Buller, 2005). Along similar lines, Garth Fletcher points out that while on average there are some gendered differences in human mate preferences, there is an enormous amount of overlap: men and women in general are looking for very similar virtues in their partners, and the overall message is one of similarity rather than differences (Fletcher et al., 2012). This kind of empirical engagement with evolutionary psychology remains ongoing.

There are more theoretical, global objections. One focuses on evolutionary psychology's picture of the EEA. A relatively easily dismissed version of this objection charges evolutionary psychologists with relying on a naive stereotype of an African savannah; some evolutionary psychologists simply deny that they, at least, were ever so naive (Tooby and Cosmides, 1990). A more trenchant version has it that, according to our best current knowledge, ancestral environments were too variable across time and space for the EEA concept to play the role it is supposed to in evolutionary psychological research. Crucially, part of the variability is due to humans' own role in creating their selective environments: we are not just passive subjects of selection, but active 'niche constructors' (Odling-Smee et al., 1996; Sterelny, 2012), meaning that at least some of the adaptive problems our ancestors faced were not stable but fluid, the results of an evolutionary 'arms race'. Under such circumstances, it may not be possible for selection to design enduring solutions in the form of special-purpose mental modules.

Evolutionary psychologists' 'massive modularity' thesis has often come under critical fire. To the claim that there is not enough time or information available during individual development for a person to learn on their own, how to solve the myriad adaptive problems life poses, it has been said that

socially scaffolded learning – epistemic niche construction – can explain humans' success at solving complex problems (Sterelny, 2012). This view is supported by the fact that humans now live, and act successfully, in social, economic, political, and physical environments vastly different from those of only 10 000 years ago, and our capacity to act successfully in these environments cannot be explained by appeal to innate, evolved modules that guide our responses to the challenges of mass society.

A common concern underlying the previous two objections is that evolutionary psychologists underestimate the extent to which social learning and cultural transmission have mattered in human evolution, and the extent to which human cognitive development depends on environmental shaping as well as genetic endowment (Heyes, 2012). The approach detailed below, gene–culture coevolution, addresses that worry.

Gene–Culture Coevolution

Sociologists see culture as a powerful shaper of human behavior. Sociobiologists by contrast suppose culture to be a relatively weak shaper of human behavior. Between these extremes, most researchers agree that both genetic and cultural influences matter in human evolution; the interesting controversy concerns the relative importance of each, and how to represent and model their interaction. Gene–culture coevolution, also known as dual inheritance theory, is the attempt to specify and quantify how genes and culture interact. The goal of this approach is to understand how cultural changes can affect genetic evolution, and how genetic changes can affect cultural evolution.

Two of the earliest proponents of this approach were E.O. Wilson and his collaborator Charles Lumsden (1981), however, their specific approach failed to take off. Another pair of contemporary researchers were also developing a gene–culture coevolutionary approach. The geneticists Marcus Feldman and Luca Cavalli-Sforza (1976) presented a model of trait evolution that integrated genetic and cultural inheritance systems, and followed this up with a book-length treatment, *Cultural Transmission and Evolution* (1981). Another very fruitful collaboration also began in this period, between anthropologists Robert Boyd and Peter Richerson (1985). They, their students and their collaborators have built an impressive research tradition, in partial overlap with human behavioral ecology. They use an impressive set of formal tools, drawing on models from population genetics and building those into transmission rules for cultural inheritance.

One key feature of the gene–culture coevolution approach is the recognition that humans reliably inherit information through two channels. One channel is genetic: we inherit information via transmission of DNA. The other is cultural. 'Culture' in this context can be ideological, consisting of information, or behavioral, consisting of skills, or even material, consisting of made objects such as tools. We inherit cultural elements via social transmission, through processes such as learning and imitation. There are a variety of directions in which social transmission can occur. One is, like genetic transmission, vertical, meaning from parents to offspring. Unlike genetic transmission (in humans), however, cultural transmission can also be oblique, from a member of a parental generation to a member of the offspring generation, or horizontal, between individuals within a single generation. In part because humans inherit cultural information from sources other than their parents, gene–culture coevolution can sometimes lead to maladaptive outcomes. Gene–culture coevolution can explain why things can go wrong, not just why they can go right.

Michael Tomasello, primatologist and developmental psychologist, has emphasized the fact that human informational inheritance is cumulative: humans inherit cultural information, transmit it, but sometimes with additions and improvements. The great apes are good enough at social learning to receive information from the previous generation, but they do not reliably add to it (Tomasello, 1999). In his view, the transition to accumulating culture was a watershed change in hominin evolution: humans have been genetically selected to learn culturally; there is mutual influence between these channels of inheritance; and as a consequence cultural conditions are hugely important in human evolutionary history (for more on this, see Shennan, 2003; Jablonka and Lamb, 2005; Whiten et al., 2011; Sterelny, 2012).

Research on dairy farming and lactose tolerance provides a good demonstration of the gene–culture coevolutionary approach. The relevant background information to this research is that there is variation worldwide across adult humans in the ability to digest dairy products such as milk. Some people can digest dairy products, in virtue of possessing a certain allele that governs the activity level of the enzyme lactase, which breaks down lactose. Other people – in fact, worldwide, most people – lack the necessary allele, do not produce sufficient lactase, and are made ill by consumption of dairy products. Moreover, possession of the allele for lactose tolerance is far more common in populations with a history of dairy farming than in populations without dairy traditions.

Feldman and Cavalli-Sforza (1989) developed a model of the evolution of lactose tolerance in which both genetic and cultural variation and transmission were included. In the model, lactose tolerance was under the control of a single gene, and the question was under which circumstances that lactose absorption allele would achieve high frequency in the model population. Interestingly, although, the ability to absorb lactose confers a fitness advantage when dairy products are available, the mere presence of dairy farming and the subsequent action of natural selection alone was not sufficient to drive the absorption allele to high frequency. Feldman and Cavalli-Sforza found that the evolution of lactose tolerance depended also, and critically, on cultural transmission. In particular, it must be the case that dairy consumers' offspring were very likely to also become dairy consumers. That is, there must be strong vertical transmission of the cultural trait of being a dairy consumer. In the model, natural and cultural selection acting in concert were required to drive the absorption allele to high frequency. Differences in the strength of cultural transmission of the practice of consuming dairy products may thus explain genetic variation in lactose tolerance across populations.

While the general approach taken by gene–culture coevolution researchers is promising, there are still very few empirically well worked out examples of this approach in

action. This dearth is partly due to a deficit of empirical knowledge, of how to distinguish, count, and track changing frequencies of cultural variants. It is also partly due to the formidable technical challenges of constructing adequate models of gene–culture coevolutionary target systems. One central challenge here is determining the extent to which cultural transmission and selection of cultural variants is a genuinely Darwinian process (Godfrey-Smith, 2009). Dan Sperber has long been skeptical, doubting that cultural learning typically has either the fidelity or chainlike character of true biological inheritance; he suggests that much social learning is diffuse, with subjects gaining a particular skill over many interactions from many sources (Sperber, 1996). Perhaps most significantly, there is currently a lack of integration between gene–culture coevolution and other relevant disciplines, such as paleoanthropology, as a consequence gene–culture coevolution ideas have not been much tested against the known historical record of hominin evolution. In this respect, though, there has been improvement, and the move toward greater integration is an emerging, and encouraging, trend.

Current and Emerging Trends

The increasing use of computer models and of experimental methods are two current and promising trends in the study of human behavioral evolution.

The use of computer models in this field is not new. It traces back at least to early sociobiology: John Maynard-Smith and George Price used game theory to model and understand animal conflict (1973), and Robert Axelrod (1984) held a simulated tournament between various behavioral strategies in the Prisoner's Dilemma. Recently, though, the use of models has been more extensive, and the models themselves more complex. A key figure in this respect is philosopher Brian Skyrms, whose books *The Evolution of the Social Contract* (1996), *The Stag Hunt and the Evolution of Social Structure* (2004), and *Signals* (2010) all apply sophisticated modeling techniques to understanding the evolution of human cooperation and communication, but there have been important contributions from evolutionary biology, human behavioral ecology, gene–culture coevolutionary theory, and evolutionary economics. Skyrms' approach is distinguished by the simplicity and generality of the basic models, the elegance through which these are elaborated, and his ability to situate his models in the context of broad debates both within philosophy and the empirical disciplines.

The key features of this approach are the use of evolutionary game theory and of mathematical models and computer simulation. One first specifies the various behavioral strategies to be assessed, for instance, a set of moves to be made in a cooperative game. Importantly, the 'agents' so modeled can be 'dumb', that is, they need to only behave and need not calculate or strategize. One must also specify the results of interactions between strategies (the 'pay offs'), the starting population composition (how many of each type of strategic agent are initially present), and likelihood of various types of agents encountering others (the degree of assortative interaction, if any). Then, simulations are run using replicator dynamics to see how the population changes over generations as a result of the differential success of the various strategies; doing this many times shows how the population is likely to evolve over time, in terms of both its trajectory and its final state.

This approach has been applied by Skyrms himself to understanding the evolution of fairness – in particular the human tendency to divide windfall resources roughly equally – and of honest and informative signaling. The extension of this approach to new phenomena is an ongoing and potentially very fruitful endeavor. Philosophically, there is a great deal of interest in the use of models as evidence and, relatedly, the relationship between models and the real-world systems they are meant to help us understand. The simplicity of the models makes the approach tractable and comprehensible, but it increases the gap between model and target systems (Weisberg, 2013).

Another recent and worthwhile development is the use of experimental methods. For example, experimental economics examines people's tendencies to divide resources fairly or unfairly, generously or selfishly, to reward fairness and generosity, and to punish unfairness and selfishness, across a range of carefully controlled conditions. One of the experimental economist's basic tools is the so-called money game and its many variants (Davis and Holt, 1992). In the Dictator game, one subject (the dictator) must decide how to allocate a sum of money between themselves and a partner (the receiver), who has no option but to accept the dictated split, even if the dictator chooses to keep the whole sum. In the Ultimatum game, one subject (the proposer) must propose a way of dividing a sum of money with their partner (the responder), who can either accept the offer (in which case proposer and responder share the money as proposed) or reject the offer (in which case neither proposer nor responder gets anything). In the Public Goods game, several subjects must individually decide whether to contribute their own resources to a common pool that will be multiplied by some factor then divided equally among all of the players.

A landmark study exemplifying this approach is that by Joseph Henrich et al. (2005), who collected cross-cultural data on money game play. Norms of fair division and contribution were found to vary significantly at the cultural group level, and this variation to be substantially explained by variation in different groups' socioeconomic histories and current structures. Other work in this relatively new tradition includes that by Ernst Fehr, Simon Gachter, Urs Fischbacher, and Samuel Bowles.

See also: Adaptation, Fitness, and Evolution; Comparative Evolutionary Psychology: Current Status and a Proposal for a More Integrated Future; Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Darwinism; Experimental Economics; Gene–Culture Coevolution; Human Behavioral Ecology; Human Cooperation, Evolution of; Human Mate Choice, Evolution of; Kin Selection; Sociobiology, History of; Sociobiology: Overview.

Bibliography

Alexander, R., 1987. The Biology of Moral Systems. Aldine de Gruyter, New York.
Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York.
Bliege Bird, R., 2007. Fishing and the sexual division of labor among the Meriam. American Anthropologist 109, 442–451.
Bowles, S., Gintis, H., 2012. A Cooperative Species: Human Reciprocity and Its Evolution. Princeton University Press, Princeton.
Boyd, R., Richerson, P., 1985. Culture and the Evolutionary Process. Chicago University Press, Chicago.
Buller, D., 2005. Adapting Minds. MIT Press, Massachusetts.
Buss, D., 1989. Sex differences in human mate preferences. Behavioural and Brain Sciences 12, 1–49.
Cosmides, L., Tooby, J., 1987. From evolution to behaviour: evolutionary psychology as the missing link. In: Dupre, J. (Ed.), The Latest on the Best: Essays on Evolution and Optimality. MIT Press, Massachusetts, pp. 277–306.
Cosmides, L., 1989. The logic of social exchange. Cognition 31, 187–276.
Daly, M., Wilson, M., 1999. The Truth about Cinderella: A Darwinian View of Parental Love. Yale University Press, New Haven.
Darwin, C., 1859[1998]. The Origin of Species. The Modern Library, New York.
Darwin, C., 1871[2004]. The Descent of Man. Penguin, London.
Davis, D., Holt, C., 1992. Experimental Economics. Princeton University Press, Princeton.
Dawkins, R., 1976. The Selfish Gene. Oxford University Press, Oxford.
Feldman, M.W., Cavalli-Sforza, L.L., 1976. Cultural and biological evolutionary processes, selection for a trait under complex transmission. Theoretical Population Biology 9, 238–259.
Feldman, M.W., Cavalli-Sforza, L.L., 1981. Cultural Transmission and Evolution. Princeton University Press, Princeton.
Feldman, M.W., Cavalli-Sforza, L.L., 1989. Evolution of lactose absorption. In: Feldman, M.W. (Ed.), Mathematical Evolutionary Theory. Princeton University Press, Princeton.
Fletcher, G., Simpson, J., Campbell, L., Overall, N., 2012. The Science of Intimate Relationships. John Wiley, West Sussex.
Godfrey-Smith, P., 2009. Darwinian Populations and Natural Selection. Oxford University Press, Oxford.
Grafen, A., 1984. Natural selection, kin selection and group selection. In: Krebs, J., Davies, N. (Eds.), Behavioural Ecology: An Evolutionary Approach. Blackwell, Oxford, pp. 62–85.
Hamilton, W., 1964. The genetical evolution of social behaviour. Journal of Theoretical Biology 7, 1–52.
Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N., Hill, K., Gil-White, F., Gurven, M., Marlowe, F., Patton, J., Tracer, D., 2005. "Economic man" in cross-cultural perspective. Behavioral and Brain Sciences 28, 795–815.
Heyes, C., 2012. Grist and mills: on the cultural origins of cultural learning. Philosophical Transactions of the Royal Society B 367, 2181–2191.
Jablonka, E., Lamb, M., 2005. Evolution in Four Dimensions. MIT Press, Massachusetts.
Kitcher, P., 1985. Vaulting Ambition. MIT Press, Massachusetts.
Laland, K., Brown, G., 2002. Sense and Nonsense: Evolutionary Perspectives on Human Behaviour. Oxford University Press, Oxford.
Lumsden, C., Wilson, E., 1981. Genes, Mind and Culture: The Co-evolutionary Process. Harvard University Press, Massachusetts.
Maynard-Smith, J., 1964. Group selection and kin selection. Nature 2010, 1145–1147.
Maynard-Smith, J., Price, G., 1973. The logic of animal conflict. Nature 246, 15–18.
Odling-Smee, F., Laland, K., Feldman, M., 1996. Niche construction. American Naturalist 147, 641–648.
Segerstrale, U., 2000. Defenders of the Truth. Oxford University Press, Oxford.
Shennan, S., 2003. Genes, Memes and Human History. Thames and Hudson, London.
Skyrms, B., 1996. The Evolution of the Social Contract. Cambridge University Press, Cambridge.
Skyrms, B., 2004. The Stag Hunt and the Evolution of Social Structure. Cambridge University Press, Cambridge.
Skyrms, B., 2010. Signals. Oxford University Press, Oxford.
Smith, E., 1985. Inuit foraging groups. Ethology and Sociobiology 6, 27–47.
Sober, E., Wilson, D., 1998. Unto Others: The Psychology and Evolution of Unselfish Behavior. Harvard University Press, Massachusetts.
Sperber, D., 1996. Explaining Culture. Blackwell, Oxford.
Sterelny, K., 2012. The Evolved Apprentice. MIT Press, Massachusetts.
Symons, D., 1990. Adaptiveness and adaptation. Ethology and Sociobiology 11, 427–444.
Sperber, D., Wilson, D., 1986. Relevance: Communication and Cognition. Blackwell, Oxford.
Thornhill, R., Palmer, C.T., 2000. Why men rape. The Sciences 40 (1), 30–36.
Tinbergen, N., 1963. On aims and methods of ethology. Zeitschrift fur Tierpsychologie 20, 410–433.
Tomasello, M., 1999. The Cultural Origins of Human Cognition. Harvard University Press, Massachusetts.
Tooby, J., Cosmides, L., 1990. The past explains the present: emotional adaptations and the structure of ancestral environments. Ethology and Sociobiology 11, 375–424.
Trivers, R., 1971. The evolution of reciprocal altruism. Quarterly Review of Biology 46, 35–57.
Whiten, A., Hinde, R., Laland, K., Stringer, C., 2011. Culture evolves. Philosophical Transactions of the Royal Society B 366, 938–948.
Williams, G., 1966. Adaptation and Natural Selection. Princeton University Press, Princeton.
Wilson, E., 1975. Sociobiology: The New Synthesis. Harvard University Press, Massachusetts.
Wilson, E., 1978. On Human Nature. Harvard University Press, Massachusetts.
Weisberg, M., 2013. Simulation and Similarity: Using Models to Understand the World. Oxford University Press, Oxford.
Wynne-Edwards, V., 1962. Animal Dispersion in Relation to Social Behaviour. Oliver and Boyd Ltd, Edinburgh.

Relevant Websites

http://philpapers.org/browse/evolution-of-culture – PhilPapers Collection for Evolution of Culture.
http://philpapers.org/browse/evolutionary-psychology – PhilPapers Collection for Evolutionary Psychology.
http://philpapers.org/browse/sociobiology – PhilPapers Collection for Sociobiology.
http://plato.stanford.edu/archives/spr2013/entries/evolution-cultural/ – Stanford Encyclopedia of Philosophy Entry on Cultural Evolution.
http://plato.stanford.edu/archives/fall2010/entries/evolutionary-psychology/ – Stanford Encyclopedia of Philosophy Entry on Evolutionary Psychology.
http://plato.stanford.edu/archives/win2013/entries/sociobiology/ – Stanford Encyclopedia of Philosophy Entry on Sociobiology.

Evolutionary Demography: A Darwinian Renaissance in Demography

Rebecca Sear, London School of Hygiene and Tropical Medicine, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Demography and evolutionary biology have a long history: Darwin was famously influenced by Malthus when developing his ideas on natural selection. The two disciplines remained closely associated throughout the early twentieth century. They disassociated after World War II, but in recent decades lines of communication between the disciplines have opened again. Initiatives from both evolutionary scientists and demographers have resulted in some genuinely interdisciplinary work, and evolutionary research is increasingly being published in demography journals and presented at demography conferences. This article will briefly survey the history of interaction between the disciplines and suggest a Darwinian renaissance in demography is underway.

Introduction

Demography is integral to evolutionary biology, given the latter's interest in reproductive success, a composite of survival and reproduction. But evolutionary biology has not always been integral to demography: demography currently considers itself a social science and few modern demographers have training in biology. This is despite the close association of the two disciplines at their inception. This article will outline the relationship between the two disciplines and attempt to demonstrate that, even though the two disciplines diverged during the middle decades of the twentieth century, a Darwinian renaissance is now underway in demography. That a resurgence of interest in evolutionary ideas is seen in demography, more so perhaps than most other social sciences, is no doubt due to a number of factors which ease communication between the two disciplines, including demography's empirical, quantitative basis. Most important, however, is likely to be acknowledgment from both disciplines that the other can be useful in their own endeavors. Many demographers have lamented that demography is a 'discipline without a theory,' leading some to turn to evolutionary theory to help fill this gap. Similarly, some evolutionary scientists have recognized the need to engage with the expertise of demographers; some have even argued that all evolutionary biologists should be demographers (Metcalf and Pavard, in their 2007 paper titled "Why evolutionary biologists should be demographers"). Demography has been described as "sex, death and passion wrapped in indicators" (van den Brekel, cited in Coleman, 2000, p. 27) but, when it comes to at least some of the human evolutionary sciences, there has been a tendency to focus too much on the sex and passion without enough emphasis on the indicators. Evolutionists could therefore learn much from one of demography's great strengths, which is its focus on high-quality data collection and methods, including accurate description before explanatory models are tested.

Darwin, Wallace, and Malthus

Initially, the two fields of evolutionary biology and demography were closely associated. Darwin was famously influenced by one of the earliest writers on population issues, Thomas Robert Malthus, as he developed his ideas on natural selection – the mechanism of evolution – while writing *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (first published in 1859). Darwin writes in his autobiography (1876, p. 120):

> In October 1838, that is, fifteen months after I had begun my systematic inquiry, I happened to read for amusement Malthus on Population, and being well prepared to appreciate the struggle for existence which everywhere goes on from long-continued observation of the habits of animals and plants, it at once struck me that under these circumstances favourable variations would tend to be preserved, and unfavourable ones to be destroyed. The results of this would be the formation of a new species. Here, then I had at last got a theory by which to work.

Darwin is referring to Malthus' 1798 work *An Essay on the Principle of Population*, a popular work which expounded Malthus' somewhat pessimistic view that human populations had an inevitable tendency to outgrow their resource base, written as a reaction to the Panglossian spirit which prevailed at the time that human society had never had it better. Malthus, in contrast to this optimism, believed that human society was doomed to perpetual misery because of the disparity between the rate of increase of human populations and that of their food resources. Given, he argued, that food production can only grow, at best, arithmetically $(2 + 2 + 2\ldots)$ but that human population can grow geometrically $(2 + 4 + 16 + 32\ldots)$ then human populations will inevitably grow faster than food production, leading to a never-ending cycle of rapid population growth being brought back into line with resources by the 'positive checks' of famine, war, and disease:

> The power of population is so superior to the power of the earth to produce subsistence for man, that premature death must in some shape or other visit the human race. The vices of mankind are active and able ministers of depopulation. They are the precursors in the great army of destruction, and often finish the dreadful work themselves. But should they fail in this war of extermination, sickly seasons, epidemics, pestilence, and plague advance in terrific array, and sweep off their thousands and tens of thousands. Should success be still incomplete, gigantic inevitable famine stalks in the rear, and with one mighty blow levels the population with the food of the world
>
> p. 126

 http://dx.doi.org/10.1016/B978-0-08-097086-8.81057-0

Malthus' purple prose undoubtedly had a part to play in the popularity of his essay, which was so well received that it was reprinted in six editions, one of which Darwin came across several decades after the first edition (the 'vices of mankind' that Malthus writes about, incidentally, are contraception and abortion, of which Malthus, as a clergyman, strongly disapproved). Alfred Russell Wallace, too, read Malthus, and was equally stimulated by his arguments about the carrying capacity of environments. From Wallace's acceptance speech on receiving the Darwin–Wallace Medal in 1908:

> Finally, both Darwin and myself, at the critical period when our minds were freshly stored with a considerable body of personal observation and reflection bearing upon the problem to be solved, had our attention directed to the system of positive checks as expounded by Malthus in his Principles of Population. The effect of this was analogous to that of friction upon the specially-prepared match, producing that flash of insight which led us immediately to the simple but universal law of the survival of the fittest, as the long-sought effective cause of the continuous modification and adaptation of living things.
>
> p. 9

Curiously, both Darwin and Malthus practiced what they preached in their own reproductive lives. Malthus, to foreshadow the discipline of demography to come, took a policy-oriented view and recommended that the dangers of unchecked population growth should be restrained through the mechanisms of late marriage and sexual abstinence (the title of the second edition of his essay was *An Essay on the Principle of Population: or a View of its Past and Present Effects on Human Happiness; with an enquiry into our Prospects respecting the Future Removal or Mitigation of the Evils which it occasions* demonstrating clearly both his interests in the practical applications of his work and a penchant for even longer book titles than Darwin). Malthus only had three children himself and has no living descendants today. Darwin, proponent of the individual-level perspective that natural selection preserves those 'favored' in the struggle for life, had 10 children and now has more than 100 currently living descendants. Wallace, more moderately, had three children and now has a handful of living descendants.

Evolutionary Biology and Demography in the Early Twentieth Century

Evolutionary biology and demography continued to develop in close association throughout the early twentieth century. Prominent figures during this period moved freely between these disciplines. Alfred Lotka, for example, a mathematician well known in evolutionary biology for his models of population biology (the Euler–Lotka and Lotka–Volterra equations are still widely used), also served as an officer of both the *Population Association of America (PAA)* and the *International Union for the Scientific Study of Population Problems* (*IUSSPP*, which later dropped the second P to become the *IUSSP*, now the largest international population organization). Raymond Pearl, a biologist who also worked on population problems, was involved in founding the biology journals *Quarterly Review of Biology* and *Human Biology* but also the demography associations *PAA* and *IUSSPP*. He is perhaps most well-known for some demographic work, particularly that on mortality and longevity, demonstrating in the 1920s, for example, an association between moderate alcohol consumption and long life.

Practitioners of both disciplines were also closely associated in the eugenics movement, aimed at encouraging the reproduction of those deemed to have desirable attributes and discouraging those deemed undesirable, which reached its peak during this period. Eugenics was promoted by some biologists who saw 'social Darwinism' as an obvious extension of Darwin's theories of evolution and a way to improve the human species (and was devised, incidentally, by Darwin's cousin, Francis Galton). Many prominent population activists, such as the pioneers of the birth control movement in the US and UK respectively, Margaret Sanger and Marie Stopes, were also proponents of eugenics. Stopes was such an ardent eugenicist, she supposedly cut her son out of her will when he married a short-sighted woman, arguing that his children may inherit the defect of myopia (Connelly, 2008). In the UK, the Eugenics Society was involved in the foundation of the Population Investigation Committee, a charity still based at the London School of Economics, which publishes one of the major, and eminently respectable, demographic journals, *Population Studies*.

Divergence after World War II

The two disciplines diverged after World War II, however. The eugenics movement may have been partly responsible for this cleavage. The postwar backlash against eugenics resulting from the explicitly eugenicist aims of Nazi Germany, as well as other human rights abuses linked to the eugenics movement, led to the social sciences striving hard to dissociate themselves from the linkage of biological theories to human affairs. In subsequent decades, the social sciences largely (and often vehemently) denied any role for biology in human behavior and biologists themselves steered clear of engaging with the social sciences. Demography became firmly situated within the social, and not the biological, sciences. In the United States today, for example, demography units are often situated within sociology departments. As a result of this divergence, many demographers have been raised in a social science tradition that views biological explanations of any human behavior with considerable wariness. As the demographer Richard Udry (1999, p. 911) puts it: "Our shared disciplinary [social science] immune systems recognize biological explanation of behavior as an infection, and reject it."

The reasons for this repositioning may not have been entirely political, however. The rapid demographic changes that took place during the twentieth century led many to the conclusion that demography must be more influenced by social factors than biology, since genetic change could not take place so rapidly. An additional factor may have been the very applied nature of demography in the post-World War II period. Much demographic data collection and analysis was focused on describing mortality and fertility levels and their determinants, so that policy could be directed at reducing mortality and fertility rates where they were still high. Arguably, such

applied science, at least in the early stages of mortality and fertility reduction, has relatively little need of theory, as substantial success can be achieved in, for example, reducing child mortality rates in low-income countries, in the absence of any overarching theoretical framework.

Whatever the reasons for the divergence of demography and biology, demography has now effectively shaken off the specter of eugenics, which biology has still not succeeded in doing. This is perhaps a little unfair given that demography was involved quite explicitly in promoting population control in the decades after World War II in the interests of curbing global population growth, in some cases at least with aims which skirted dangerously close to eugenics (see Matthew Connelly's 2008 history of the population control movement 'Fatal Misconception,' which was reviewed by journalist Fred Pearce as "…an investigative narrative of how individuals, NGOs, governments, and UN agencies colluded over decades to sideline the human rights of hundreds of millions of the world's poorest citizens"; New Scientist, 24 May 2008). Those applying an evolutionary perspective to human behavior in recent decades have in contrast taken care to emphasize the errors in eugenicist science and have been much less keen to engage in policy. Demography has since had its own debates about the controversial population control movement (the well-respected demography journal *Population and Development Review* published three highly critical book reviews of Fatal Misconception, including one by prominent demographer and recent president of the *IUSSP*, John Cleland), and now population policies typically have an explicit human rights, rather than population control, agenda. But it is still perhaps ironic that the social science community tends to regard demography as a respectable social science, whereas the application of evolutionary biology to human behavior is not yet wholly accepted as a reputable endeavor.

Renaissance

This postwar split led to several decades of nonfraternization between demography and evolutionary biology, but the rapprochement of the disciplines does now seem to be well underway. After a period of biologists' relative lack of interest in the social sciences, the 1970s saw the emergence of sociobiology – the application of evolutionary theory to social behavior – which stemmed from the work of (nonhuman) animal ethologists in the 1950s and 1960s (Segerstralle, 2001). E.O. Wilson controversially applied this approach to human behavior in the final chapter of his textbook on *Sociobiology: The New Synthesis* (published in 1975). Despite the protests of many within the social science community that this would inevitably lead to a second round of eugenics, the application of evolutionary theory to the human sciences has developed into a thriving and diverse field. While evolutionary psychologists have perhaps become numerically the most dominant of the evolutionary social scientists, there is also a successful discipline of evolutionary anthropology, as well as the integration of evolutionary theory into many other aspects of the human sciences, such as a movement to establish a field of evolutionary medicine.

The resurgence in enthusiasm for combining demography and evolutionary ideas began in the 1980s with a number of meetings aimed at bringing together biologists and demographers. Initially, the focus was on mortality, with workshops on 'Population & biology' in 1981, 'Upper limits to human life span' in 1987, 'Convergent issues in genetics and demography' in 1988, 'Biodemography of longevity' in 1991, and 'Life span: evolutionary, ecological and demographic perspectives' in 2001, though there was also a workshop on 'Biodemography of fertility and family behaviour' in 2002. These workshops have resulted in some exciting advances being made in evolutionary demography during the last few decades, as well as the increasingly frequent appearance of evolutionary ideas at population conferences. All the major population conferences, including the *IUSSP*, *PAA*, *European Association for Population Studies*, and the *British Society for Population Studies*, have now held sessions explicitly devoted to evolutionary demography. Short courses to train students in evolutionary demography have also now been held at the *International School for Demography* in Rostock, Germany, and at Stanford University. Labs in evolutionary demography, led by the Evolutionary Biodemography lab at The *Max Planck Institute for Demographic Research* in Rostock, Germany, are beginning to appear in the United States and Europe. What may be the first tenure-track position in human evolutionary demography was advertised in 2010 at Yale. Finally, and perhaps most notably, an Evolutionary Demography Society (http://www.evodemos.org/default.htm) was established in 2013, which held its first conference in October 2013.

Quantifying the Darwinian Renaissance in Demography?

As both demography and evolutionary biology are empirical disciplines, demographers and evolutionary biologists presumably require empirical evidence that a Darwinian renaissance may be underway. In order to provide some very crude evidence, I have attempted to determine whether biological terms may be increasing in frequency in demography journals over time. To do this, I analyzed the number of times the terms 'evolutionary' and 'biology' appeared in three top demography journals (*Demography*, *Population Studies*, and *Population and Development Review*) over the last few decades, using the jstor database. This database allows a count of the number of times a word appears in these journals within a certain time range (I restricted the analysis to search the full text of research articles only). The number of articles (and therefore the total number of words) published has also increased considerably in these journals over this time period, so to control for this I constructed a faux 'rate' for the appearance of each of these terms by dividing the number of times the term of interest was mentioned by the number of times the word 'population' appeared: making the assumption that the relative frequency of the term 'population' would have remained constant over time. For comparison, I performed the same calculations for the social science terms 'anthropology' and 'psychology.'

Figure 1 shows the absolute increase in the 'rate' of use for each term by decade in the jstor database (*Population* studies appear in jstor since 1947, *Demography* since 1964, and *PDR*

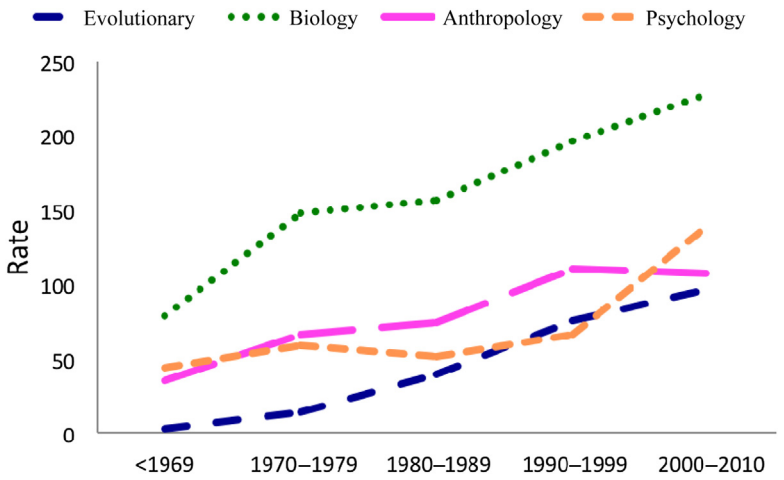

Figure 1 Frequency of terms in three demography journals over time.

since 1975; the analysis was only possible up to 2010 since the jstor database does not include the most recent issues of journals). This figure shows the frequency of all terms has increased relative to the term population over this time period, indicating that demography is becoming more diverse in the disciplines it interacts with. But there are differences in how frequently the different terms are mentioned, and how rapidly these rates have increased over time. 'Biology' is the most widely used term overall, perhaps suggesting that biology and demography have always been somewhat linked, and has increased substantially over time. 'Evolutionary' has seen a much larger increase over time, however. Initially, it was seen very infrequently in these journals, but showed the largest increase over time to become as common as the social science terms included. 'Anthropology' has steadily increased over time, perhaps reflecting the emergence of the field of anthropological demography. 'Psychology' showed little increase until the last decade, which may correspond to a relative lack of interest from demographers in psychology until very recently – perhaps, suggesting the emergence of a very new field of psychological demography.

Why Did the Renaissance Occur?

The reasons for this Darwinian Renaissance are undoubtedly many. The similarity in subject matter of evolutionary biology and demography – resulting from both fields being interested in fertility and mortality – is clearly very important. As the most biological of the social sciences, demography is likely to be rather more open to biological thinking than other social sciences. Demography is also a strongly empirical and largely quantitative discipline, and has not, therefore, been diverted into the cul-de-sac of postmodernism as have so many of the other social sciences. The demographer David Coleman (2000, p. 32) goes as far to say: "Demography without numbers is a bland form of social or historical waffle." Developments in genetics and physiology mean that it is increasingly straightforward to link the demographic processes of birth and death with biomarkers and genetic factors, which may have concentrated some demographers' minds on the importance of such biological factors: Kreager (2009, p. 469) suggests that "the tremendous development of genetics has recalled demographers' attention to evolutionary theory as an inescapable element of modern population thought."

But perhaps, one of the most important factors for demographic interest in incorporating biological theories is that, unlike some of the other social sciences, it is not wedded to any particular theory. In fact, it has been a recurrent complaint from within the field itself that demography is a discipline without a theory (which makes it somewhat ironic perhaps, that Darwin declared Malthus had "given him a theory by which to work": see quote from his autobiography above). As early as 1952, Vance asked in the title of his *PAA* presidential address "Is theory for demographers?" The prominent demographer Livi-Bacci has described the discipline as "a technique rather than a science" (Livi-Bacci, 1984, p. iii); and a plenary at *EAPS* in 1995 by Guillaume Wunsch was titled *God has chosen to give the easy problems to the physicists Or: Why demographers need theory*. (It should be noted that this is not a universally held view among demographers: for example, in Micklin and Poston's (2005) brief survey of the theoretical basis of demography they cite Gutman (1960), who argues that demography does offer *illuminating theoretical statements* (Micklin and Poston, 2005: p. 5). Your view of whether demography is theory free or not to some extent depends on your definition of theory: Burch (2003) argues that if a theory is redefined as a model, defined in turn as an approximation of reality, rather than a set of laws from which testable hypotheses can be generated, then demography is rich in theory, because it is rich in such models).

This lack of theory has left a void, which has become particularly noticeable as demographers have become increasingly interested in individual-level rather than population-level processes, and in causal explanations for demographic phenomena: Hobcraft (2006, p. 174) called for demographers "to tackle the difficult and interesting problems of understanding behaviour, rather than undertaking elaborate description," an endeavor which really requires theory. Demography has, in fact, not been shy to borrow from other

disciplines in the past: economic thinking has, for example, been particularly influential in demographic models of the fertility transition (the shift from high to low fertility which began in Europe in the eighteenth century and has now occurred almost worldwide). To quote David Coleman (2000, p. 27) again: demography has long been "a space where other specialists come for a while to do some of their work."

Finally, important in the reunifying of demography and biology is the commitment of individual researchers to work together to solve particular problems. Two of the most prominent scientists who have worked hard to reunite the two disciplines are Jim Vaupel and Jim Carey. A social and natural scientist respectively, they have been instrumental in organizing and developing many of the workshops, courses, and the new society mentioned above. They argue that a significant factor in the progress of evolutionary demography has been the identification of particular problems that both disciplines can contribute to:

> The good intentions of top scientists are not enough to integrate two fields with fundamentally different disciplinary histories, professional cultures, and epistemological frameworks. To make progress it is imperative to lay out a clear set of important (and ultimately fundable) questions that lie at the disciplinary interface. This is particularly important for integrating disciplines with disparate historical roots, such as demography with its roots in the social and analytical sciences and biology with its roots in the natural and experimental sciences.
>
> Carey and Vaupel, 2005, p. 633

The Incorporation of Evolutionary Biology into Demography

Carey and Vaupel (2005) argue that the first few seminars described above were not particularly successful, and did not produce much in the way of outputs, because a suitable problem had not yet been identified. In more recent years, evolutionary demography has really begun to take off, after the discipline began to make significant progress in the study of aging and mortality rates at old ages. A recognition by demographers that standard demographic models failed to accurately describe the pattern of mortality at very old ages, as average longevity increased in developed nations, led to the active forging of links with the biological sciences in the hope that evolutionary models may prove useful. Models that take account of how natural selection has shaped mortality trajectories are now being used to help explain the demographic puzzle of why mortality rates should begin to plateau or decline at old ages (when standard demographic models predict *increasing* mortality rates with increasing age). A particularly notable feature of this work is its comparative nature. Evolutionary demography is not just for humans: this field has produced considerable research on how mortality trajectories vary across nonhuman species (throughout the natural world) and even on inanimate objects. This work has demonstrated remarkable variability in the shape of age-specific mortality trajectories across species, suggesting that both demography and evolutionary theories on aging need to be rethought (Vaupel et al., 1998; Jones et al., 2014). This area has also benefitted from the attention that demographer Ron Lee (2003) has given to intergenerational transfers in the aging process. Lee has used evolutionary theory to develop models of aging which incorporate intergenerational transfers as a key component influencing mortality rates, which dovetails nicely with recent work in the evolutionary sciences emphasizing the importance of such transfers in our species (Hrdy, 2009).

Work on mortality may be the most significant success story for evolutionary demography so far, but evolutionary thinking is beginning to inform other areas of demography too. In the field of reproduction, David Coleman (2000) has highlighted areas of fertility research that require biological input, for example, to explain the puzzle of why people continue to have children even when they bring only material disadvantage and when effective contraception is readily available. Such behavior is not explicable using standard demographic, typically economic, models of fertility, and can really only be explained with recourse to evolutionary arguments. Children do not appear to make one healthy, wealthy, or happy, but we do descend from ancestors who had physiological and behavioral adaptations that resulted in reproduction: any individual without such physiological abilities or behavioral tendencies toward childbearing did not become an ancestor. Empirical work by demographers developing an interest in the evolutionary demography of reproduction has tended to focus on genetic influences on reproduction. For example, early publications stemming from the collaboration of demographer Hans-Peter Kohler with psychologist Joseph Rodgers used twin study designs to ask "is fertility behaviour in our genes?" (Kohler et al., 1999; this collaboration also resulted in the publication of an edited volume on the Biodemography of Reproduction: Rodgers and Kohler, 2003). The answer was: partially, yes. More recently, Melinda Mills has been making use of technological advances in genetics and is using GWAS (genome-wide association studies: Mills et al., 2013) to explore the genetic basis to fertility behavior; again, finding evidence that genes do matter.

Evolutionary demography is having an impact on demography not just in terms of publications but also the kinds of data that are collected through the large-scale surveys demographers typically use. Studies investigating biosocial influences on demographic events are only possible if data on the potential biological and social influences on demography exist. Demographic surveys have always collected detailed socioeconomic information, but are increasingly collecting detailed biological information too. John Hobcraft, for example, is a demographer who has been influential in promoting the inclusion of biomarkers in longitudinal surveys, so that it is easier to "elaborate pathways from biology through the brain/mind to behaviors and outcomes, particularly for social, demographic, economic, and health domains" (Hobcraft, 2009, p. 252).

I will end this section with another quote from Carey and Vaupel (2005), who have argued that incorporating more biology may be essential if demography is to remain an active and important discipline:

> …every discipline including demography is faced with the perennial struggle to define and renew itself and to ensure its relevance in an

> ever changing world. Like other social sciences, demography is slowly coming to terms with important truths that the biological sciences have provided beyond any doubt: that all aspects of humans – mind, behaviour, body – are products of biological evolution. It follows that this program, in particular, and biological demography in general, should help demography maintain a robust, energetic and creative presence in modern science.
>
> Carey and Vaupel, 2005, p. 631

What Evolutionists Can Learn from Demographers

> Nothing in biology, the eminent biologist Theodosius Dobzhansky (1900–1975) asserted, makes sense except in the light of evolution. It is equally valid to say that nothing in evolution can be understood except in the light of demography
>
> Vaupel, 2003, p. 84

The section above has described work done largely by demographers becoming interested in using evolutionary theory to inform their work, but enthusiasm for the reemergence of evolutionary biology comes from both disciplines: prominent evolutionary anthropologists, such as Hillard Kaplan and Monique Borgerhoff Mulder, have also engaged explicitly with the demographic community. They have formed links with demographers, not just publishing in demographic journals and organizing sessions on evolutionary demography at population conferences, but also by setting up working groups such as the IUSSP Panel for Evolutionary Perspectives on Demography (active 2004–10, and resulting in the publication of two special journal issues).

Evolutionists bringing a more demographic perspective into their work have tended to focus largely on reproduction (unlike demographers bringing evolution into their work). Kaplan and colleagues, for example, have written influential papers on theories of fertility, both theoretical and empirical (Kaplan, 1996; Kaplan and Lancaster, 2003). These papers draw on the demographic literature but incorporate elements of life history theory (the evolutionary framework relevant for studying demography), such as parental investment, and also give a historical perspective on reproductive behavior over time. Kaplan has also helped move forward debates in demography, by directly setting up tests of alternative hypotheses, where social and evolutionary explanations of fertility behavior are in conflict (though frequently evolutionary and social science explanations are not in conflict with one another, because they explain the same phenomena at different levels: Sear, 2015). For example, he empirically contrasted two models of intergenerational resource flows, which have been used in theories of the demographic transition. The influential demographer Jack Caldwell's "wealth flows" hypothesis for the demographic transition argues that in predemographic transition societies, net wealth flow is up generations, from children to parents. Evolutionary theory, on the other hand, predicts that net wealth flow will always be downward, from parents to children: wealth is always used in the service of reproduction, never the reverse. Kaplan (1994) tested these alternate models by measuring the economic costs and benefits of children in hunter-gatherer societies. His findings suggested that, contra Caldwell's wealth flows hypothesis, children were always a net economic cost to their parents, even in traditional subsistence societies. This does not mean that changes in the costs and benefits of children were not factors in the demographic transition, or that children never provide benefits to their parents, just that this one assumption of the wealth flows hypothesis was not supported.

The benefits of such cross-fertilization for the evolutionary community include borrowing from the methodological sophistication of demography, not just in formal but also statistical analysis of demographic data and the use of large-scale, longitudinal datasets. Kim Hill and Magdalena Hurtado produced one of the earliest monographs in evolutionary ecology, which made heavy use of statistical techniques and advances in demography, such as the use of event history models for demographic events. Their book *Ache Life History: The Ecology and Demography of a Foraging People* (Hill and Hurtado, 1996) was an impressive tour de force testing a wide range of hypotheses derived from life history theory using their longitudinal data set on the Ache, South American hunter-gatherers, and was very influential in shaping the developmental of the field of human evolutionary ecology (the study of physiology and behavior in ecological context, within the framework of evolutionary theory). Other evolutionary ecologists have followed in their footsteps, testing hypotheses on demographic data deriving from evolutionary theory but making use of the methods developed to analyze human demographic data. Increasingly, evolutionary ecologists are also making use of large-scale data sets collected by demographers, seeing an opportunity to test their hypotheses on existing large, rich data sets, rather than relying on small-scale data collection in 'traditional' populations (Nettle et al., 2013).

Remaining Differences of Opinion

The positive overview I have given so far should not be taken to mean that the discipline of evolutionary demography is entirely uncontroversial, nor that all demographers are comfortable with the idea that biology matters for human demography. In some quarters at least there still may be the view that biological and social explanations are alternates so that when trying to explain any human behavior "biology and the social are locked in an explanatory zero-sum game in which any ground ceded to the former diminishes the value of the latter" (Carey and Vaupel, 2005, p. 642–643). On the other side of the fence, some evolutionists have yet to fully engage with demographers to improve their understanding and analysis of the demographic phenomena they are interested in. There are also still conceptual differences between the two disciplines in that demography is a 'bottom-up' science, which starts with an empirical observation and subsequently attempts to explain that phenomenon (such as why does fertility decline?); whereas evolutionary biology is a 'top-down' science, which sees data as a means to an end to test hypotheses which derive from an overarching theoretical framework (Kaplan and Gurven, 2008). This primary interest in data versus theory means that the two disciplines will always have distinct and separate goals, but there is still sufficient overlap between the disciplines for the increasing interaction between the two to be scientifically fruitful.

Conclusion

Of all of the social sciences and humanities, demography seems perhaps to be one of those disciplines where a Darwinian renaissance is genuinely underway. Kreager, for example, wrote in 2009 (p. 471) that "evolutionary biology, *of course*, re-emerged in the 1990s as a potential source of evidence and theoretical insight into central demographic problems" (emphasis added). This is partly because renaissance is the right term to apply in the case of demography, given the close association of the two disciplines during their early development. But also because there is evidence of an emerging discipline of evolutionary demography, driven by both evolutionary scientists and demographers recognizing the need to draw on each other's expertise. Indeed, this discipline has not just emerged but is sufficiently well established to be "com [ing] of age" and "entering its adolescence" in the words of one author who contributed to a special issue on biodemography in the demography journal *Demographic Research* (Wachter, 2008, p. 1501). I will leave the last word to the demographer and biologist who have perhaps done the most to establish this discipline.

> Although still a modest subfield within demography, biodemography may be one of the fastest growing areas of demography and one of the most innovative and stimulating
>
> Carey and Vaupel, 2005, p. 625

See also: Darwinism; Demographic Models; Demography, Early History of; Demography: History Since 1900; Differential Aging; Eugenics as a Basis of Population Policy; Evolution, History of; Evolutionary Demography: A Synthesis of Two Population Sciences; Fertility Theory: Embodied-Capital Theory of Life History Evolution; Fertility Theory: Theory of Intergenerational Wealth Flows; Fertility Theory: Theory of Life History Evolution; Fertility Transition: Economic Explanations; Genetics, Disease, and Reproduction; Human Behavioral Ecology; Kinship, Evolution of; Mortality, Biodemography of; Sociobiology, History of.

Bibliography

Burch, T.K., 2003. Demography in a new key: a theory of population theory. Demographic Research 9 (11), 263–284.

Carey, J.R., Vaupel, J.W., 2005. Biodemography. In: Poston, D.L., Micklin, M. (Eds.), Handbook of Population. Kluwer Academic Press, New York, pp. 625–658.

Coleman, D., 2000. Demography in intellectual context: a subject in search of a home. In: Pavlik, Z. (Ed.), The Position of Demography among Other Disciplines. Department of Demography and Geodemography, Charles University, Prague, pp. 27–35 and other chapters in this volume on 'The Position of Demography among Other Disciplines'.

Connelly, M., 2008. Fatal Misconception: The Struggle to Control World Population. Belknap Press, Cambridge, MA.

Darwin, C., 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life, first ed. Available at Project Gutenberg: http://www.gutenberg.org/.

Darwin, C., 1876. The Autobiography of Charles Darwin. Available at Project Gutenberg: http://www.gutenberg.org/.

Hill, K., Hurtado, A.M., 1996. Ache Life History: The Ecology and Demography of a Foraging People. Aldine de Gruyter, New York.

Hobcraft, J., 2006. The ABC of demographic behaviour: how the interplays of alleles, brains, and contexts over the life course should shape research aimed at understanding population processes. Population Studies 60 (2), 153–187.

Hobcraft, J., 2009. Reflections on the incorporation of biomeasures into longitudinal social surveys: an international perspective. Biodemography and Social Biology 55 (2), 252–269.

Hrdy, S.B., 2009. Mothers and Others: The Evolutionary Origins of Mutual Understanding. Belknap Press, Cambridge, Massachusetts.

Jones, O.R., Scheuerlein, A., et al., 2014. Diversity of ageing across the tree of life. Nature 505 (7482), 169–173.

Kaplan, H., 1994. Evolutionary and wealth flows theories of fertility: empirical tests and new models. Population and Development Review 20 (4), 753–791.

Kaplan, H., 1996. A theory of fertility and parental investment in traditional and modern human societies. Yearbook of Physical Anthropology 39, 91–135.

Kaplan, H., Lancaster, J., 2003. An evolutionary and ecological analysis of human fertility, mating patterns and parental investment. In: Wachter, K.W., Bulatao, R.A. (Eds.), Offspring: Human Fertility in Biodemographic Perspective. National Academies Press, Washington, pp. 170–223.

Kaplan, H., Gurven, M., 2008. Top-down and bottom-up research in biodemography. Demographic Research 19 (44), 1587–1602.

Kohler, H.-P., Rodgers, J.L., Christensen, K., 1999. Is fertility behavior in our genes? Findings from a Danish twin study. Population and Development Review 25 (2), 253–288.

Kreager, P., 2009. Darwin and Lotka: two concepts of population. Demographic Research 21 (16), 469–502.

Lee, R.D., 2003. Rethinking the evolutionary theory of aging: transfers, not births, shape social species. Proceedings of the National Academy of Sciences of the United States of America 100 (16), 9637–9642.

Livi-Bacci, M., 1984. Foreward to Nathan Keyfitz. In: Population and Biology. Ordina Editions, Liège.

Malthus, T.R., 1798. An Essay on the Principle of Population. Available at Project Gutenberg: http://www.gutenberg.org/.

Metcalf, C.J.E., Pavard, S., 2007. Why evolutionary biologists should be demographers. Trends in Ecology & Evolution 22 (4), 205–212.

Micklin, M., Poston, D.L., 2005. Prologue: the demographer's kin: 50 years of growth and change. In: Poston, D.L., Micklin, M. (Eds.), Handbook of Population. Kluwer Academic/Plenum Publishers, New York, pp. 1–15.

Mills, M., Barban, N., Mandemakers, J., Snieder, H., Social Science Genetics Association Consortium, 2013. A meta GWAS (genetic wide association search) of fertility tempo and quantum. In: Paper Presented at the International Union for the Scientific Study of Population Conference, Busan, Korea. http://www.iussp.org/sites/default/files/event_call_for_papers/GWAS_IUSSP2013_Final_0.pdf.

Nettle, D., Gibson, M.A., Lawson, D.W., Sear, R., 2013. Human behavioural ecology: current research and future prospects. Behavioural Ecology 24 (5), 1031–1040.

Potts, Malcolm, Cleland, John, Hodgson, Dennis, 2008. Reviews of Connelly's fatal misconception. Book review symposium published in Population & Development Review 34 (3).

Rodgers, J.L., Kohler, H.P., 2003. The Biodemography of Human Reproduction and Fertility. Kluwer Academic Publishers, Boston.

Russell Wallace, A., 1908. Wallace's Remarks on Receiving the First Darwin-Wallace Medal on 1 July 1908. Printed in The Darwin-Wallace Celebration Held on Thursday, 1 July 1908, as published by the Linnean Society of London in 1909.

Sear, R., 2015. Evolutionary contributions to the study of human fertility. Population Studies.

Segerstralle, U., 2001. Defenders of the Truth: The Sociobiology Debate. Oxford University Press, Oxford.

Udry, R., 1999. How to alienate your natural allies and attract your enemies. Behavioural and Brain Sciences 22, 911.

Vance, R., 1952. Is theory for demographers? Social Forces 31, 9–13.

Vaupel, J.W., 2003. In: Demeny, P., McNicoll, G. (Eds.), Biodemography. Encyclopedia of Population. Thomson Gale, New York, pp. 84–87.

Vaupel, J.W., Carey, J.R., Christensen, K., Johnson, T.E., Yashin, A.I., Holm, N.V., Curtsinger, J.W., 1998. Biodemographic trajectories of longevity. Science 280 (5365), 855–860.

Wachter, K.W., 2008. Biodemography comes of age. Demographic Research 19 (40), 1501–1512.

Wilson, E.O., 1975. Sociobiology: The New Synthesis. Belknap Press, Cambridge, MA.

Wunsch, G., 1995. God has chosen to give the easy problems to the physicists, or why demographers need theory. In: Plenary Given at European Population Conference, Milan. http://www.un.org/popin/confcon/milan/plen6.html.

Evolutionary Demography: A Synthesis of Two Population Sciences

Daniel A Levitis, Max-Planck Odense Center on the Biodemography of Aging, University of Southern Denmark, Odense, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Demography is the quantitative study of population processes, while evolution is a population process that influences all aspects of biological organisms, including their demography. Demographic traits common to all human populations are the products of biological evolution or the interaction of biological and cultural evolution. Demographic variation within and among human populations is influenced by our biology, and therefore by natural selection and our evolutionary background. Demographic methods are necessary for studying populations of other species, and for quantifying evolutionary fitness. Comparisons among species allow for a greater understanding of how and why demographic patterns evolve.

Twin Population Processes

Shared Concepts and Focus

The field of demography arose as, and largely remains, a social science. It generally restricts itself to human populations, and the causative hypotheses it tests are based in sociology, economics, history, psychology, cultural anthropology, and other social sciences. Despite this cultural tendency, the underlying phenomena demography describes are strongly influenced by biology, and therefore, by biological evolution. Indeed, the central concepts of demography – including mortality, fertility, migration, populations, sociality, behavior, heterogeneity, heritability, stochasticity, resource flows, and environmental influences – are all also central to evolutionary biology. Evolutionary demography expands the scope of demographic research by considering nonhumans, and employing methods, concepts, and hypotheses from evolutionary biology. As such, it allows us to ask questions about the origins of demographic patterns that are otherwise unanswerable.

Separate literature regarding, for example, economic consequences of marriage markets in humans and fitness effects of assortative mating in all other species arise with distinct methods for quantifying extremely similar phenomena. Linguistic differences between fields (e.g., what demographers call 'fertility' and 'migration' biologists call 'fecundity' and 'dispersal') are symptomatic of a historical dearth of communication between social and natural scientists. (**Box 1** provides definitions of some relevant biological terms.) The study of evolutionary demography aims to cross these artificial barriers to mutual illumination by combining the concepts and tools of demography and evolution, thereby broadening the scope of both fields.

This article deals specifically with the intersection of evolution and demography, and (because the field of biology is much broader than the study of evolution) excludes much of the broader category of biodemography. For example, biomarkers of aging are often measured or the demography of nonhuman populations quantified without reference to concepts or tools derived from evolutionary biology. Other sections in this encyclopedia focus on topics in biodemography.

Evolution as Demography

While demography is defined as the quantitative study of population processes (Preston et al., 2000), evolution is a population process and, therefore, inherently demographic (Kohler and Rodgers, 2003). Evolution through natural selection can occur only when heritable traits influence the demography (survival or reproduction) of their bearers. The calculation of fitness requires consideration of not only the demographic effects of a trait, but also demographic characteristics of the population, including age structure, stochasticity, and population growth rate (Metcalf and Pavard, 2007). Evolutionary processes beyond selection, such as genetic drift and mutation, are strongly influenced by the size and characteristics of the population.

How Demography Evolves

In outline, the evolution of demographic traits is little different than that of morphological, physiological, or behavioral traits. Alleles that influence any of these phenotypic traits change in frequency over time through selection, recombination, mutation, and drift. The difference, in this context, between demography and these other aspects of the phenotype is that other traits influence fitness, and therefore selective outcomes, only via demography. More specifically, any trait that does not influence survival or reproduction has no fitness consequences, while one that increases survival or reproduction will tend to increase fitness. This implies both that demography is the aspect of phenotype most tightly tied to fitness, and that any response to selection influences the demography of the population.

The classic example in evolutionary demography is an allele that decreases susceptibility to a deadly disease at a particular age and therefore allows its bearers to continue reproducing, increasing the frequency of that allele and therefore the longevity of the population (see Kirkwood, 2005; Wachter et al., 1997 and references therein for review of evolutionary theories of aging). However, if selection strengthens the wood of a tree, or the linguistic ability of a primate, the demography of the population has also perforce been influenced. The resulting change may be quickly countered by density effects (such as food or space limitation) or the population may end up more longevous or fertile. Any instance of selective

Box 1 Definitions of key biological terms for evolutionary demography

Allele: One form of a gene. For example, two copies of the same chromosome may have slightly different forms of a gene that when functioning properly, helps to protect against breast cancer. Different alleles may offer different degrees of protection, with the null allele (a mutated form that fails to produce a functional product) offering no protection.

Dispersal: Movement of individuals from one site to another, generally noncyclical. This is what demographers would generally call migration.

Evolution: change in the genetic composition of lineages over time, and changes in the phenotype that result from this genetic change. Evolution results from selection, genetic drift, migration, mutation, and genetic recombination.

Fertility: The physiological capacity to reproduce, which demographers call fecundity.

Fecundity: How many offspring the individual actually produces, which demographers call fertility.

Fitness: The mean ability of individuals of a particular phenotype or genotype to get copies of their genetic material into future generations. Fitness is calculated in a variety of different ways (Metcalf and Pavard, 2007), but all depend on reproduction and survival. Fitness depends on both the characteristics of individuals and the particular environment in which it is measured.

Genetic drift: Change in allele frequencies that result not from selection or mutation, but rather stochastic changes in what proportion of the population carries each allele. Demographic stochasticity, whose effects decrease with population size, is a key driver of genetic drift.

Genotype: The set of alleles possessed by an individual, sometimes considered across all genes (the genome) but often in reference to only those genes that have bearing on the trait of interest.

Individual: The unit considered to be a separate entity, often by tacit analogy to an individual human. In some organisms, it is necessary to consider carefully how the individual is to be defined before demography can be examined.

Migration: Long-distance cyclical movements, often seasonal, back and forth between habitats.

Mutation: A change in the sequence of nucleotide (DNA or RNA bases) that constitutes the genome of an organism, or a new allele that results from such a change. While some mutations are beneficial and many are deleterious, most do not alter demography and, therefore, have no effect on fitness.

Phenotype: The organism's observable traits. Phenotype results from the interaction of the genotype with the environment.

Phylogeny: The evolutionary relationships among species or populations reflecting their common ancestry, often represented as a branching tree.

Selection: The differential survival or reproduction of individuals based on their phenotype. If these phenotypic differences are heritable, an evolutionary response to selection (change in allele frequencies) can occur.

Senescence: Where demographers usually define senescence in terms of declines in survival probabilities or reproductive rates with age, biologists more often refer to declines in function, such as slower healing or locomotor speed. Evolutionary demographers tend toward the demographic definitions unless otherwise specified.

evolution can, therefore, be thought of as favoring a morphological/physiological/and so on trait that is directly influenced by the genes, or as favoring the demographic effect of that trait. Selection for larger testes (which are useful in species that engage in sperm competition) can also be considered to be selection for higher investment in fertility.

So while evolution is a demographic process, the demography of any biological population is shaped by its evolutionary history. That humans begin reproducing after more than a decade, while small monkeys do so after one or a few years, for example, is a result of evolution. Indeed, any demographic trait that distinguishes humans from other species is at least in part the result of the evolved biological differences between humans and others. That human demography has changed so rapidly in the last centuries is surely an example of cultural evolution (Burger et al., 2012), but the fact that we have the linguistic, cognitive, and social abilities to change our culture so quickly (while chimpanzees, for example, do not) is the result of the biological evolution that has separated humans from related species. Biological evolution should not, therefore, be seen as an alternative to cultural evolution. Rather, our biological history constrains and allows for our cultural behaviors, and those behaviors determine our environment, and thereby, how our biological traits continue to evolve. Human culture and biology jointly and collaboratively drive the evolution of human demography.

In addition to demography and evolutionary biology, evolutionary demography intersects with a wide range of other fields, including anthropology, gerontology, epidemiology, evolutionary medicine, behavioral ecology, and reliability engineering.

Selective Explanations for Human-Demographic Traits

Optimality and Constraints

Evolutionary demography expands the scope of demography by considering evolutionary explanations for demographic traits. These are most often framed in terms of natural selection,

with the fundamental idea that selection pushes populations to achieve the optimal set of traits (Chu et al., 2008), with optimality defined as the maximization of fitness. Infinite reproduction starting immediately after birth coupled with zero mortality is the impossible limit toward which fitness maximization points. That this circumstance is quite impossible is because all organisms and populations operate within a variety of constraints imposed by their environment and internal biology. Increasing all fitness components simultaneously would certainly be favorable, but not feasible. Understanding the evolution of demography, therefore, frequently requires us to ask what has constrained that evolution.

The constraint that evolutionary demographers most often consider is that of resource limitation. An organism is modeled as partitioning its income at each age (in the currency of calories or time, for example) among a variety of investments (increasing current survival, current reproduction, growth, etc.). A unit of resource can buy so much growth, survival probability, and so on and income is limited by factors that may include size or investment in foraging (Chu et al., 2008). A variety of techniques are then available for finding the investment strategy that maximizes fitness. These techniques are often successful in predicting key features of life histories, such as the prereproductive growth phase, the onset of senescence at age of first reproduction, and sex ratios near parity.

This reasoning provides key hypotheses on the evolution of human demography. Onset of reproduction then occurs at that age where the tradeoff between fertility and continued growth shifts to favoring reproduction. Infant and child mortality, under this optimality-with-resource-constraints framework, result from strong selection for early investment in growth, at the cost of investment in current survival (Chu et al., 2008). Or they can be understood, by applying resource constraints at the level of the food-sharing group, rather than the individual, as resulting from the negative but increasing economic value (in terms of transferable calories) of dependent young (Lee, 2003). Women's postfertile survival can similarly be considered as resulting from their optimal allocation including the resources necessary to keep them alive to continue investing in the survival and reproduction of their children and grandchildren (reviewed in Pavard et al., 2008). Senescence, in this resource-constraint framework, is the result of investment in reproduction and current survival having a higher payoff than investment in traits that maintain future survival and reproduction.

However, resource limitation is far from the only constraint acting on the outcomes of selection as it shapes demography. Staying with the example of senescence, the genetic basis of demography must be considered. A single allele may affect multiple traits (pleiotropy), for example, increasing reproduction at one age while decreasing survival at all ages, or at a particular age. Selection can increase or decrease the frequency of this allele in the population but it cannot generally decouple the positive and negative effects. Where the total effect on fitness is positive, the allele will tend to become more common until it goes to fixation. For example, traits expressed early in life tend to have a greater effect on fitness than those occurring later in adult life, even if the magnitude of the effect is smaller. This is true both because a smaller proportion of individuals will be alive to express a trait later in life and because mortality later in life occurs after a larger proportion of the individual's reproduction has already occurred. Antagonistic pleiotropies (where some fitness effects are positive and other negative) tend to increase early reproduction and survival at the cost of later success, resulting in senescence. This is still fitness optimization, but with genetic rather than resource constraint.

A similar result can be obtained without assuming that optima are attainable, on the basis of mutation-selection balance. While selection tends to increase the frequency of alleles that maximize fitness, random mutations generally decrease fitness. The allele frequencies that result when selection and mutation act simultaneously will depend in part on how strongly selection acts to remove mutations. Among other things, this depends on the ages at which alleles have their fitness effects. Again, mutations that act later in life will tend to have less of an effect of fitness, and therefore be removed less efficiently, allowing accumulation of mutations that increase mortality and decrease fertility later in life. The constraint here derives from the fact that selection itself is only so powerful in determining the rates of available traits.

A further set of constraints arises from the fact that many potentially desirable phenotypes simply do not occur. The ability to reproduce asexually when no mate is available does not occur in any mammal and therefore probably would not occur in humans even if it was strongly selectively advantageous. While no demographer would assume facultative clonality in humans, our models do sometimes include traits that we have good reason to believe are impossible. For example, models in evolutionary demography frequently assume perfect age specificity of gene effects, such that selection can cause fertility to start increasing at an exact age, or mortality to approach infinity at the exact age when reproduction ends (see Levitis et al., 2013 and references therein). Perfect age specificity is a convenient simplifying assumption, but a false one. While average gene expression does change as a function of age, a remarkably small proportion of variance in expression is age specific. Furthermore, the demographic results of age-specific gene expression (whether genetic diseases or events in the normal healthy human life-history) often occur not when a gene is turned on or off, but after some product or effect of expression, influenced by environment and by feedback in regulator networks, has accumulated or attenuated, gradually increasing the probability of that event. Put more simply, even if gene expression changed rapidly as a function of age, phenotypic results may come more gradually. At the demographic level, this results in trajectories of gradually changing rates, rather than discontinuous changes (Cohen et al., 2012). An allele that would cause a particular abrupt change in an age profile of mortality or fertility might be optimal, but it also probably does not exist.

A related reason, the optimal trait may not exist can be found in Charles Darwin's (1859) favored description of evolution, 'descent with modification.' Evolution functions by gradually modifying traits of a lineage, and traits that are too dissimilar from parents have near zero probabilities of arising in descendants. If some small portion of the population already has the preferred trait, or something close to it, change can occur rapidly, but where heritable variation does not exist, or does not include traits preferable to the current norm, selection can do little. Imagine, for example, that in a given environment it would be highly selectively advantageous for women to cease reproducing at age 45 but then resume fertility at age 65, after

two decades of child rearing and physiological recovery. Natural variation in women's reproductive physiology does not include this or useful intermediate traits (there is no unmenopause), so the hypothetical trait cannot be selected for.

Traits Shared by All Human Populations

Causal hypotheses in demography most often focus on explaining variation within and among populations. Evolutionary demography, by contrast, has often focused on explaining specific portions of the patterns shared by all human populations. Looking at human age-specific mortality curves (**Figure 1**), for example, demographers most often ask why the curves differ from each other in some respect, or what characteristics of individuals determine who dies at each point along the curve. Why is infant mortality higher in some than others? Why does the bump in the teenage years get earlier in richer societies? Why do some adults die at the age of 40 years and others 90 years? Evolutionary demography often focuses on the similarity among these curves, explaining them with hypotheses based in selection and constraints on its actions. High but declining early mortality (i.e., ontogenescence) is explained in terms of optimality, as described above, but also in terms of mechanistic constraints and heterogeneous frailties (Levitis, 2011). The period of increased mortality observed early in adulthood and primarily in males (the adolescent mortality hump) is proposed to be the result of selection for risk-taking behaviors that increase reproductive opportunity (Kruger and Nesse, 2006). In contrast, no widely accepted hypothesis has been proposed to explain why senescent mortality is close to exponential in its increase over a wide range of adult ages.

Human fertility traits are also evolved, and aspects such as age at reproductive onset, interbirth intervals, twinning rates, offspring size, and age at reproductive cessation are addressed by evolutionary hypotheses (Wachter and Bulatao, 2003). One key aspect of human reproduction, the need for a large-headed baby to pass through a biped's narrow pelvis during birthing is both a legacy of evolution (all vertebrates that have pelvic bones pass their offspring through them) and may have had serious consequences for the evolution of human demography. This 'Obstetric Dilemma' may, by limiting infant head size, have increased the length and severity of the juvenile dependence, and women's evolved postfertile life stage likely evolved as a result of human socioecological responses to this dependence (Robson et al., 2006). That the ratio of offspring head size to maternal pelvic opening in humans is similar to that of related species has been taken as evidence against this hypothesis (Dunsworth et al., 2012), but can also be seen as evidence for a constraint that occurs more broadly than just humans. In other words, if human infant head size was already at the limit determined by primate pelvic morphology, but selection pushed for adults with bigger brains, this may have required an increase in the proportion of neurological development that occurs after birth.

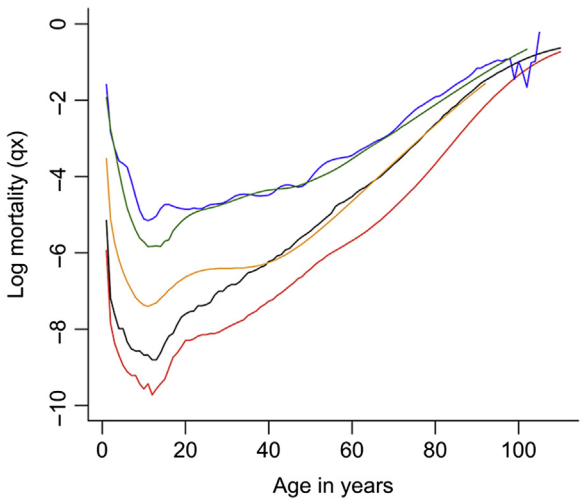

Figure 1 Human patterns of mortality over age are remarkably similar in form across populations. Data shown (from highest mortality to lowest) are Sweden in the 1750s, Taiwan in the 1920s, Sri Lanka in the 1980s, Belarus in the 1990s, and Japan in the 2000s. Cohort and male mortality curves have the same basic shape as these female period data (Human Life-Table Database and Human Mortality Database, 2013. Rostock, Germany and Berkeley, CA: UC Berkeley and Max Planck Institute for Demographic Research. www.mortality.org).

Few aspects of evolutionary demography receive as much attention from the public as the quest to understand women's postreproductive life span. While most other species seem, at least in their natural environments, to experience declines in fertility and survivorship roughly simultaneously, women across environments lose fecundity well before they die (or keep living after their fertility has declined to zero). A remarkable range of hypotheses (reviewed in Robson et al., 2006; Pavard et al., 2008; Levitis et al., 2013; Wachter et al., 1997) have been proposed to explain why. These focus on the roles of mothers and grandmothers in aiding younger kin, on the continued reproduction of older men and the genetic correlations between male and female and on avoidance of intergenerational competition, among many other selective arguments.

These are but a few examples: any demographic pattern that occurs across human populations should be interrogated as to its evolutionary basis. Doing so expands the scope of demography by allowing explanation of traits that can otherwise be noticed and measured but not explained.

Variation among Humans

Those demographic patterns that vary within and among human populations are less often subject to evolutionary hypothesizing for two good reasons. The first of these is that there exists a long troubling history of using evolutionary language and biased evolutionary arguments to rationalize discrimination, and many social scientists are, therefore, wary of biological explanations for behavioral or demographic differences among people. The second is that humans are tremendously culturally diverse but biologically fairly uniform, and the biggest demographic differences among us are, therefore, likely to be cultural in origin. Despite both of these good arguments, some demographically relevant human traits are appropriately explained in evolutionary terms. These differences can be placed into two rough categories: evolved differences and evolved plasticity.

When populations and individuals differ in their genetic predispositions to suffering from certain potentially deadly

diseases (e.g., certain cancers, sickle cell anemia), we can ask about potential fitness benefits of particular alleles, about the role of mutation, genetic drift, breeding patterns, and so on. In other words, we can understand these demographically significant differences in evolutionary terms. This same reasoning applies to many demography-influencing traits that are, or may be, heritable (Kohler and Rodgers, 2003).

Contrast this with evolved plasticity, which does not require heritable differences, but instead that the same set of genetic traits can be expressed in different (often selectively advantageous) ways in different environments. For example, the thrifty phenotype hypothesis (Hales and Barker, 2001) argues that individuals who undergo early development in a nutrient poor environment can respond to this by altering their development such that they are better able to deal with low-nutrient environments late in life.

Evolutionary reasoning can be helpful in understanding demographically important behaviors with more potential for controversy. Patterns of infanticide, for example, can be explained in terms of caretakers' degree and confidence of relatedness to those they are related to and the potential for redirection of parental resources (Dixson, 2013). Demographic transitions can be modeled as resulting from evolved economic behaviors and a changing environment (Borgerhoff Mulder, 1998). While there may be limits on the aspects of human demography that should be described in evolutionary terms, there are no obvious limits on the aspects that can be so described.

Demography beyond Humans

The Application of Demography to Nonhumans

Humans have unique patterns of mortality, fertility, and migration but are not unique in having mortality, fertility, and migration, and so it is not surprising that demographic methods (e.g., survival analysis) have long been adopted and adapted by biologists and ecologists to quantify the population processes of species beyond demography's traditional boundary. The motivation for these efforts varies from the applied (How many fish of what size can we catch sustainably? How quickly can this introduced pest population grow? Does this medicine extend life span in mice?) to the basic (Is sexual reproduction associated with higher offspring viability? Does copious food promote earlier natal dispersal?) and theoretical (If an organism could exponentially increase its reproductive output indefinitely, would it still be expected to evolve aging?). Indeed, demographic methods influenced by those originally developed to study humans are included in the basic quantitative toolboxes of wildlife biology, conservation biology, forestry management, population ecology, behavioral ecology, zoo management, and a wide range of other disciplines focused on nonhumans.

Challenges in Nonhuman Demography

It is not always straightforward to apply the methods of anthropocentric demography to other species. Simplifying assumptions that are safe for humans are often violated by other species. Particular nonhumans may reproduce by fission, switch sexes, experience extended periods of dormancy, have several very different life stages that reproduce in multiple ways, return to earlier stages of development, repeatedly lose and regenerate the majority of their tissues, return from apparent death, merge with or incorporate the genetic materials of others, and employ a dizzying array of tactics that confound any traditional demographic analysis. An attempt to study the longevity of sea sponges in the wild, for example, may be foiled by the difficulty of simply determining what an individual is, given ceaseless splitting, rejoining, fragmentation, disappearance, and reappearance (Tanaka, 2002). While age is a key variable in human demography, it is an ill-defined term of little use when individuals cannot be spatially and temporally demarcated. So while human-demographic methods may apply almost without modification to some species (many birds and mammals, for example) evolutionary demography employs and develops an array of modified or alternative methods for studying similar phenomena.

Adding to the need for alternative or modified methods are the particular data problems experienced when studying nonhumans. Nonhumans can generally not be expected to answer survey questions or report births and deaths. Individuals of some species are extremely difficult to recognize, mark, track, or even define. The underlying biology of a species may be poorly known. On the other hand, some species have distinct advantages for the researcher. Many trees provide a reliable means to determine an individuals' age (counting rings) without tracking over time, are unable to leave the study area and never decline to participate.

Methods for Nonhuman Populations

Where age is a key variable, human specific age-based methods can often be directly applied to nonhumans. Life tables, age-specific fertility, survival analyses with age-specific baseline functions, and many other common demographic tools are successfully applied to some animals, plants, fungi, bacteria, and so on. Where other key variables (stage, size-class, etc.) dominate or interact with age, some tools from the demographers' toolbox, including many regression methods, can still usefully be applied. But in these circumstances many tools less often employed by demographers, including a variety of purpose-built matrix models, are favored. These models often incorporate transition matrices mapping stochastic paths through several possible life stages. Where in human demography these stages could be never married/married/divorced, in crab demography they could be egg/planktonic larva/pre-breeder/breeding adult/nonbreeding adult. Some stage-based models are expanded to also include age. Integral Projection Models combine matrices with regressions to model the effects of continuously varying factors (e.g., rainfall) on the various matrix elements. Mark-recapture data, in which individually identifiable animals (zebras with identifying stripes or sea stars with tattoos) are repeatedly recorded and released are commonly used to study demography, and a variety of tools have been built for this purpose.

Studying nonhumans allows for the use of experimental methods that would otherwise be impractical, unethical, and illegal. These experiments can be used to document the demography of a particular species and how it depends upon

environment, to test a hypothesis about how demography evolves by applying a selective pressure over several generations, or to alter the genetic makeup of a population and examine how demography is affected.

Evolutionary Comparative Demography

Comparisons across Species

Evolutionary demographers often examine demographic variation among species. To understand how a demographic trait of humans (or any other species) has evolved, one needs to know what it evolved from (the evolutionary history of the trait), how widely it occurs, what traits it cooccurs with, what alternatives are biologically possible, and so on. Knowing these things requires comparisons across species.

While no other species' demography has been as thoroughly studied as that of humans, some aspects of demography have been documented in thousands of plants and animal species, allowing for broad comparison. The goal is generally to answer an evolutionary question. For example, Do higher metabolic rates lead to shorter life spans? Evolutionary demography overlaps extensively with life-history evolution, and many of the hypotheses tested derive from that literature.

Comparative Measures, Methods, and Data

Serious challenges exist not only in the types and quality of data that are available, but also in compiling the existing data from a wide variety of sources and building sensible comparisons between organisms that live in different ways and have been studied by people whose methods are often not designed for comparisons to other species. For example, maximum observed longevity, while far from an ideal measure of how long a species lives, has often been employed for comparative purposes because this is a relatively easy datum to record and compile and is a defined (if not known) quantity for every type of organism for which individuals can be identified. Because different species' demography occurs on different timescales, useful comparative measures are often dimensionless (Charnov, 1991) or explicitly corrected for the pace at which each species proceeds through life (Baudisch et al., 2013).

Once measures are chosen, a variety of methods from evolutionary biology are useful for comparative demographic studies. Allometric methods aid in the consideration of traits that scale with the size of the organism (e.g., longevity with body or brain mass) in a nonlinear fashion. Allometric relationships take the form $Y = a^{*}X^{Z}$. Y is the trait of interest, X is size, and a and z are fit coefficients. z, the scaling factor, describes how quickly Y increases with increase in X. Most life-history measures increase with size, but inherent physical constraints on the scaling of circulatory systems limit metabolic speeds such that larger organisms tend to use less energy per unit mass (West et al., 1997). As a result, z is often roughly 1/4, across a remarkable range of physiological and life-history measures and across many taxa. For example, the life span or age at first breeding of an organism of twice the size will tend to be about 1.2 times as long.

Phylogenetic methods, which explicitly take into account the evolutionary tree connecting the lineages being considered, are frequently required for comparative evolutionary demography (Stearns, 1983). Because closely related species are likely to share characteristics through common inheritance, it is not appropriate to treat them as independent samples in a cross-species analysis. The less close the relationship, the more independence we generally expect to see. A wide range of methods are available to estimate and correct for the degree to which this so-called 'phylogenetic inertia' is influencing a trait or the relationship between traits. Phylogenetic methods are used for a wide range of purposes, including estimating when a particular trait evolved, and thereby what other traits or events influenced its evolution.

Just as comparisons among humans are greatly aided by demographic databases, comparisons among species will become much easier as biodemographic databases come on line. Some, such as AnAge (offering primarily maximum observed longevity (De Magalhaes and Costa, 2009)) are already available. Others offering far greater demographic detail in the form of matrix models or life tables are under construction but not yet freely available.

Widely Occurring Traits

Comparative methods and comparative databases are particularly useful when a trait varies. If every species in a database has the same trait value, comparisons among them are not informative. A study of the demographic effects of sexual reproduction, for example, would not use a database on the demography of birds and mammals, all of which reproduce sexually. This is obvious, but has a potentially less obvious correlate: a widely occurring trait should be explained with a widely applicable hypothesis. A parsimonious explanation should explain sexual reproduction in all birds, and not just in moorhens. The same principle should be, but is often not, applied to humans. In order to ask why adult mortality increases exponentially with age, we should ask if this is true only in humans, or across all primates, mammals, vertebrates, or animals. To understand why women have a long post-reproductive life span, it is useful to ask if it is evolutionarily derived trait, or only a longer version of a trait that occurs across species (Levitis et al., 2013). Many, perhaps most, demographic traits that occur across all human populations have close parallels in related species in a way that implies homology (i.e., humans and our relatives coinherited, rather than independently deriving, these trait). One can always ask why humans have not evolved to modify or discard the inherited trait, but this is often fruitless, especially if the origin of the trait is not well understood. Where a trait is found beyond humans, a human-specific explanation is of limited value in explaining the origin of the trait (e.g., a pre-reproductive period), but can still be useful in explaining the details of why humans have a modification of the same trait (e.g., a longer prereproductive period).

So while evolutionary demography helps to expand the boundaries of demography (in terms of hypotheses, patterns, taxa considered, and methods used) applying evolutionary reasoning to demographic problems also usefully constrains the tendency to consider humans in a vacuum. It does so by placing us in the context of other species with which we share demographic traits, biological mechanisms, and evolutionary

history. That such interspecific comparisons are outside the tradition of demography only increases their capacity to expand our knowledge of population processes.

See also: Differential Aging; Early Life Influences on Health and Mortality in Adulthood; Fertility Theory: Embodied-Capital Theory of Life History Evolution; Fertility Theory: Theory of Intergenerational Wealth Flows; Infant and Child Mortality in Industrialized Countries; Life Table; Migration and Fertility; Mortality and the HIV/AIDS Epidemic; Mortality of the Oldest-Old; Mortality, Biodemography of; Mortality, Epidemiological, and Health Transitions; Period and Cohort Analysis in Demography.

Bibliography

Baudisch, A., Salguero-Gómez, R., Jones, O.R., Wrycza, T., Mbeau-Ache, C., Franco, M., Colchero, F., 2013. The pace and shape of senescence in angiosperms. Journal of Ecology 101, 596–606.

Borgerhoff Mulder, M., 1998. The demographic transition: are we any closer to an evolutionary explanation? Trends in Ecology & Evolution 13, 266–270.

Burger, O., Baudisch, A., Vaupel, J.W., 2012. Human mortality improvement in evolutionary context. Proceedings of the National Academy of Sciences 109, 18210–18214.

Charnov, E.L., 1991. Evolution of life history variation among female mammals. Proceedings of the National Academy of Sciences 88, 1134–1137.

Chu, C.Y.C., Chien, H.K., Lee, R.D., 2008. Explaining the optimality of U-shaped age-specific mortality. Theoretical Population Biology 73, 171–180.

Cohen, A.A., Martin, L.B., Wingfield, J.C., Mcwilliams, S.R., Dunne, J.A., 2012. Physiological regulatory networks: ecological roles and evolutionary constraints. Trends in Ecology & Evolution 27, 428–435.

Darwin, C., 1859. On the Origin of Species by Means of Natural Selection. John Murray, London.

De Magalhaes, J., Costa, J., 2009. A database of vertebrate longevity records and their relation to other life-history traits. Journal of Evolutionary Biology 22, 1770–1774.

Dixson, A.F., 2013. Male infanticide and primate monogamy. Proceedings of the National Academy of Sciences 110, E4937.

Dunsworth, H.M., Warrener, A.G., Deacon, T., Ellison, P.T., Pontzer, H., 2012. Metabolic hypothesis for human altriciality. Proceedings of the National Academy of Sciences 109, 15212–15216.

Hales, C.N., Barker, D.J., 2001. The thrifty phenotype hypothesis. British Medical Bulletin 60, 5–20.

Human Mortality Database, 2013. Max Planck Institute for Demographic Research, Rostock. Germany and University of California, Berkeley, CA. www.mortality.org.

Kirkwood, T.B., 2005. Understanding the odd science of aging. Cell 120, 437–447.

Kohler, H.-P., Rodgers, J.L., 2003. Education, fertility, and heritability: explaining a paradox. Offspring: Human Fertility Behavior in Biodemographic Perspective, 46–90.

Kruger, D.J., Nesse, R.M., 2006. An evolutionary life-history framework for understanding sex differences in human mortality rates. Human Nature 17, 74–97.

Lee, R.D., 2003. Rethinking the evolutionary theory of aging: transfers, not births, shape senescence in social species. Proceedings of the National Academy of Sciences 100, 9637–9642.

Levitis, D., Burger, O., Bingaman Lackey, L., 2013. The human postfertile life span in comparative evolutionary context. Evolutionary Anthropology 22, 66–79.

Levitis, D.A., 2011. Before senescence: the evolutionary demography of ontogenesis. Proceedings of the Royal Society B: Biological Sciences 278, 801–809.

Metcalf, C.J.E., Pavard, S., 2007. Why evolutionary biologists should be demographers. Trends in Ecology & Evolution 22, 205–212.

Pavard, S., Metcalf, C.J.E., Heyer, E., 2008. Senescence of reproduction may explain adaptive menopause in humans: a test of the "Mother" hypothesis. American Journal of Physical Anthropology 136, 194–203.

Preston, S., Heuveline, P., Guillot, M., 2000. Demography: Measuring and Modeling Population Processes. Wiley-Blackwell, Malden, MA.

Robson, S.L., Van Schaik, C.P., Hawkes, K., 2006. The derived features of human life history. In: Hawkes, K., Paine, R. (Eds.), The Evolution of Human Life History. School for Advanced Research Press, Sante Fe, pp. 17–44.

Stearns, S.C., 1983. The influence of size and phylogeny on patterns of covariation among life-history traits in the mammals. Oikos 41, 173–187.

Tanaka, K., 2002. Growth dynamics and mortality of the intertidal encrusting sponge Halichondria okadai (Demospongiae, Halichondrida). Marine Biology 140, 383–389.

Wachter, K.W., Bulatao, R.A., 2003. Offspring: Human Fertility Behavior in Biodemographic Perspective. National Academies Press.

Wachter, K.W., Finch, C.E., National Research COUNCIL Committee On, P, 1997. Between Zeus and the Salmon: The Biodemography of Longevity. National Academy Press, Washington, DC.

West, G.B., Brown, J.H., Enquist, B.J., 1997. A general model for the origin of allometric scaling laws in biology. Science 276, 122–126.

Evolutionary Developmental Psychology

Karin Machluf and David F Bjorklund, Florida Atlantic University, Boca Raton, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Evolutionary developmental psychology posits that natural selection operates across the lifespan and focuses on how psychological mechanisms develop. The basic tenets of evolutionary developmental psychology are outlined, along with detailed discussion of ontogenetic, deferred, and conditional adaptations. These are then applied to mental health, where the mental health model is pitted against the evolutionary developmental perspective of mental health. Lastly, adolescent risk-taking behaviors are examined from an evolutionary-informed perspective using life history theory to explore the influence of environmental harshness and unpredictability in early childhood.

Development and Evolution

The intimate connection between development and evolution has long been acknowledged. Darwin, in his controversial and revolutionary book *The Origin of Species* (1859), meticulously described the processes of embryology, as it was understood then, and used it as part of his indisputable evidence for evolution (*see* Darwinism). In a letter to a his colleague, Asa Gray, Darwin wrote, "Embryology is to me by far the strongest single class of facts in favor of change of forms, and not one, I think, of my reviewers has alluded to this" (1860). He concluded that if slight changes are introduced across different points of time, for different parts of the body, it would produce different forms that are adapted to those particular circumstances.

Darwin's assessment of the importance of development was correct but simultaneously premature. Darwin knew nothing of genetics, and the rediscovery of Mendel's work in the early 1900s and August Weismann's doctrine of the separation of the germ and somatic lines (what happens to the body cells of an organism during its lifetime cannot influence the content of its sex cells), resulted in the temporary parting of ways of development and evolution. The field became 'gene-centric,' integrating Mendelian genetics with Darwin's ideas of natural selection, leading to *neo-Darwinism*, or the *Modern Synthesis* (Huxley, 1942). This reflected the official acceptance by the scientific community that large-scale changes in form could be explained by small genetic changes acted upon by the process of natural selection, over long stretches of time, to produce variations within species. From this perspective, development did not matter, or, as more elegantly expressed by Richard Dawkins (1976/2006: p. 62), "The details of embryological developmental processes, interesting as they may be, are irrelevant to evolutionary considerations." Development, although of vital importance to the individual, was essentially inconsequential to natural selection and thus evolution.

In the past few decades, however, the focus on development has reemerged in evolutionary thinking, with, for example, Stephen Jay Gould's *Ontogeny and Phylogeny* (1977), reinvigorating the idea that modifications during development may influence evolution, Eva Jablonka and Marion Lamb's *Epigenetic Inheritance and Evolution* (1995), proposing the role of epigenetics in evolution (see also Jablonka and Lamb's 2005 book, *Evolution in Four Dimensions*), and Mary Jane West-Eberhard's *Developmental Plasticity and Evolution* (2003), extending epigenetics to developmental plasticity (*see* Epigenetic Inheritance). Over the last two decades, evolutionary thinking has entered the mainstream of psychology, and most evolutionary psychologists have modified the now-famous quote by Theodosius Dobzhansky (1973: p. 749) stating that "Nothing in biology makes sense except in the light of evolution" to "Nothing in psychology makes sense except in the light of evolution."

The central concept of Darwin's theory of evolution is *natural selection*, which assumes that more members of a species are born in a generation than will survive, resulting in competition for a set of limited resources (*see* Microevolution). Individuals vary, and those that have features associated with greater survival will pass those features to their offspring. In other words, heritable variations in characteristics of an individual interact with the local environment, and, over long periods of time, change in frequency, resulting, eventually, in specieswide traits in the population.

Evolutionary psychologists realized that natural selection could be used to explain species differences in psychological as well as physical features, and thus could be used to explain and predict many aspects of human behavior (e.g., Buss, 1995; Daly and Wilson, 1988; Tooby and Cosmides, 1992; *see* Evolution of Cognition: An Adaptationist Perspective). The principal focus of mainstream evolutionary psychology was understandably on adults, who do the reproducing, the ultimate measure of Darwinian success. Much like biologists of an earlier era, many evolutionary psychologists saw development as an epiphenomenon, interesting and important for the individual, but little more than 'noise' when it came to important psychological characteristics that typify *Homo sapiens*. Yet, individuals must survive the often-hazardous periods of infancy and childhood before reaching adulthood, where the reproducing (and subsequent rearing of children) takes place. In fact, although the likelihood of a newborn dying before reaching adolescence today is about 1% in Western cultures, it is closer to 50% in traditional societies and was likely as higher or higher for our ancestors (Volk and Atkinson, 2008). This makes childhood the crucible of evolution, such that any adaptations that increase the chance

 http://dx.doi.org/10.1016/B978-0-08-097086-8.81018-1

of surviving to adulthood during these earlier stages of life would certainly be favored by natural selection.

Evolutionary Developmental Psychology

Natural selection operates at all stages of the lifespan; however, it operated at some stages more than others, especially early in development. From this realization was born the field of *evolutionary developmental psychology* (e.g., Bjorklund et al., 2007; Bjorklund and Pellegrini, 2000, 2002; Bjorklund and Hernández Blasi, 2005; Burgess and MacDonald, 2005; Ellis and Bjorklund, 2005; Geary and Bjorklund, 2000). Evolutionary developmental psychology is defined as the application of the basic principles of evolution to explain contemporary human development: "It involves the study of the genetic and environmental mechanisms that underlie the universal development of social and cognitive competencies and the evolved epigenetic (gene–environment interactions) processes that adapt these competencies to local conditions; it assumes that not only are behaviors and cognitions that characterize adults the product of selection pressures operating over the course of evolution, but so also are characteristics of children's behaviors and minds" (Bjorklund and Pellegrini, 2002: p. 4). It posits that natural selection operates throughout ontogeny (i.e., the lifetime of the individual), not simply during an organism's reproductive years.

In this article, we first outline briefly some of the principles and assumptions of evolutionary developmental psychology (see Bjorklund and Pellegrini, 2000, 2002). We then look at how an evolutionary developmental psychological approach can be applied to the study of developmental psychopathology. We specifically contrast the canonical mental health model, which generally emphasizes *costs* and largely ignores any *benefits* of children's responses to risky environments, with an explicitly evolutionary developmental perspective. Finally, we discuss how this evolutionary developmental perspective can help better understand risk-taking and antisocial behaviors, especially in adolescence.

Some Basic Principles of Evolutionary Developmental Psychology

A central concern of evolutionary developmental psychology is how psychological mechanisms develop. In fact, this, some believed, is what mainstream evolutionary psychological theory was lacking. How do inherited, presumably genetically coded behaviors and cognitions get expressed in the phenotypes of adults? The obvious answer seems to be 'through development,' and in this section we examine some of the basic principles of evolutionary developmental psychology.

All Evolved Characteristics Develop via Continuous and Bidirectional Gene–Environment Interactions That Emerge Dynamically over Time

One problem that many people have with an evolutionary approach to human behavior is that there is the taint of *genetic determinism*, the belief that if a behavior or type of cognition evolved and is influenced by genes (which are, after all, the basic units of inheritance), then there is little one can do to change 'human nature'; that is, genes directly cause, or determine, behavior (e.g., Spencer et al., 2009; Lickliter and Honeycutt, 2003). Such an approach is generally anathema to developmental psychologists, accounting, in part, for the reluctance of some developmentalists to adopt an evolutionary perspective. However, evolutionary developmental psychologists have adopted a *developmental systems approach* (Gottlieb, 2002, 2007), which stipulates that genes and environment interact at all levels of organization, from the biochemistry of the cells through culture. From this perspective it is inappropriate to ask 'how much' of any feature (physical or psychological) is the result of 'nature' vs 'nurture,' but rather about ascertaining how do the biological and environmental factors interact to produce those features.

At the core of the developmental systems approach is the concept of *epigenesis*, which Gilbert Gottlieb (1991: p. 7) defined as "the emergence of new structures and functions during the course of development." New structures or behaviors are the result of the bidirectional relationship between all levels of biological and experiential factors, from the genetic through to the cultural. 'Experience,' from this perspective, includes events external to the individual, such as perceptual stimulation and social interaction, but also internal events, such as the presence of neurotransmitters, hormones, and even the firing of one neuron as it influences its neighbors and itself. Functioning at one level influences functioning at adjacent levels, with constant feedback between levels. This relationship can be expressed as follows:

$$\text{genetic activity (DNA} \leftrightarrow \text{RNA} \leftrightarrow \text{proteins)} \leftrightarrow \text{structural maturation} \leftrightarrow \text{function, activity.}$$

From this perspective, all development is the product of epigenesis, with complex interactions occurring among multiple levels.

Evolved psychological mechanisms, the central tenet of mainstream evolutionary psychology, can be viewed as inherited 'messages' that, following epigenetic rules, interact over time with the environment to produce behavior (see Bjorklund et al., 2007). Because the experiences of every individual are unique, there would be substantial individual differences (i.e., plasticity) in development. Despite this, most members of a species share many features. This is because individuals inherit not only a species-typical genome, but also a species-typical environment, starting with the prenatal environment. To the extent that individuals grow up in species-typical environments similar to those of their ancestors, development should follow a species-typical pattern.

Development Is Constrained by Both Genetic and Environmental Factors

Infants are born into a complex world of sights, smells, sounds, and other people. How do they learn to figure out what is important and what is not, and, in general, to make sense of the plethora of sensations that bombard them? It was once believed that infants were born (essentially) as blank slates –

they had the general human ability to learn and, via the operations of classical and operant conditioning, learned what is relevant and what is not. This seems not to be the case. Rather, newborns are prepared by natural selection to 'expect' certain types of environments and to process some information more easily than others. From this perspective, one can talk of infants and children inheriting biases that increase the likelihood of their developing adaptive responses to their environment, with such biases and constraints *enabling* learning.

For example, from shortly after birth, infants show biases in what they like to look at and listen to, and many of these are related to other people. For instance, at birth babies prefer the voices of women to those of men (DeCasper and Fifer, 1980) and prefer to listen to the language that their mothers speak rather than to other languages (Mehler et al., 1988), biases that are surely the result of prenatal auditory experiences. Such biases should not be thought of as 'innate' in the way the term is typically used, but rather as the nearly inevitable outcomes that are constrained by how genes and experience interact in a species-typical way during early development.

An Extended Childhood Is Needed in Which to Learn the Complexities of Human Social Communities

Humans spend more of their lives as 'prereproductives' than any other mammal, something that was surely risky, especially for our ancestors who could not expect a long life. In evolutionary biology, when there are great costs associated with a characteristic (here, an extended juvenile period) there should also be great benefits. For ancient and contemporary humans, many believe that those benefits were in terms of developing a large brain capable of acquiring the skills necessary to navigate the social world (Bjorklund et al., 2005; Dunbar, 1995, 2010). The complexities of human societies are highly varied across the globe, and it takes time for children to acquire the abilities to cooperate and compete with their fellow *H. sapiens*. Humans' extended period of immaturity, coupled with a big brain that permits a social learning ability unsurpassed in the animal world, allowed our species to become as socially sophisticated as we are.

Natural Selection Has Produced Adaptations of Infancy and Childhood: Ontogenetic, Deferred, and Conditional Adaptations

If natural selection operates at all life stages, then we must look specifically at adaptations associated with infancy and childhood, rather than focusing solely on adaptations in the reproductive adult. From a developmental perspective, however, not all adaptations are created equal. Some evolved characteristics of infants and children were specifically selected to serve an adaptive function at a specific time in development, termed *ontogenetic adaptations* (Bjorklund, 1997). In contrast, *deferred adaptations* are those aspects of childhood that were selected for their role in preparing the child for adulthood (Hernández Blasi and Bjorklund, 2003). And *conditional adaptations* influence the course of children's subsequent development, depending on conditions in their immediate environment (Boyce and Ellis, 2005).

Ontogenetic adaptations emerge at different times in development and adapt the child for challenges specific to that time in life, and may disappear when they are no longer functional. That is, some aspects of infancy and childhood are not preparations for later adulthood but were selected to serve an adaptive function at specific times in development. The most obvious ontogenetic adaptations are associated with prenatal functioning. The umbilical cord, for example, delivers nutrients to and removes wastes from the fetus but is discarded at birth, as infants' entire respiratory and consummatory systems become reorganized. An example of a postnatal ontogenetic adaptation is children's physical immaturity. Research has shown that children's immature physical features (i.e., 'babyness') can help them garner nurturing behaviors from their parents. Some features of physical immaturity include heads that are proportionally large relative to their bodies, foreheads that are large in relation to the rest of the face, a small flat nose, big eyes, and short limbs (Lorenz, 1943). Adults find these features 'cute' and rate children with these features as more sociable, easier to care for, and more competent (e.g., Alley, 1981; Glocker et al., 2009; Sprengelmeyer et al., 2009) than children without them. A recent study by Sherman et al. (2009) assessed how 'cuteness' influences behavioral carefulness, the increased control of and attention to fine motor movements, after viewing immature (puppies and kittens) and mature (dogs and cats) images. Sherman et al. (2009) concluded that viewing immature faces increased fine motor abilities in adults, which the authors suggest is important for caring for a small, delicate child (see also Nittono et al., 2012). In other research, *cognitive immaturity* has also been proposed as an ontogenetic adaptation. For example, children who express some forms of immature cognition ('The sun is not out because it is mad') are rated higher on measures of positive affect by adults (e.g., cute, likable) and lower on measures of negative affect (e.g., irritable, likely to lie) than children expressing more mature cognition (e.g., 'The sun is not out because wind blew clouds in front of it') (Bjorklund et al., 2010b; Periss et al., 2012).

In other studies, young children's tendencies to overestimate their physical, social, and cognitive abilities have been shown to have some positive consequences. Research has demonstrated that preschool children overestimate their memory abilities, their intelligence, and their social standing (see Bjorklund et al., 2010a for a review). This was historically thought of as a handicap, and that children's immature cognitions would eventually develop into functional, accurate, and mature cognitions (Bjorklund, 1997; Bjorklund and Green, 1992). From an evolutionary developmental perspective, however, this poor metacognition (thoughts about the one's cognitive abilities) may hold some possible benefits (Bjorklund and Green, 1992). For instance, Bjorklund et al. (1993) assessed preschool children's pretest predictions (how well they thought they would perform a task) and posttest evaluations (how well they thought they had performed the task) of performance on a series of imitation tasks (e.g., juggling one, two, or three balls). They then correlated the accuracy of children's estimations with their performance on the Vocabulary subset of the Wechsler Preschool and Primary Scale of Intelligence. Bjorklund et al. (1993) found that verbal

IQ in the 3- and 4-year-olds was higher in children who overestimated more, while the 5-year-olds with poor metacognition (higher overestimation) tended to have lower verbal IQ scores. Thus, for younger preschoolers, being out of touch with one's abilities was related to high levels of IQ. Another study by Shin et al. (2007) found that kindergarten, first-, and second-grade children who overestimated their memory performance more on earlier trials actually had higher levels of memory recall and strategy performance on later trials than more accurate children (see also Lopez et al., 1998). These findings suggest those children's immature cognition, in this case overestimation, may in fact serve an adaptive purpose, and is not simply a 'deficiency' that children need to grow out of.

In contrast to ontogenetic adaptations, deferred adaptations are those that prepare children for life as an adult and are perhaps most easily seen through the lens of gender differences in children's behavior. For example, boys and girls tend to differ in their forms of fantasy play (see Pellegrini, 2012). Girls engage in more play-parenting and nurturing roles, while boys play power and dominance games, usually centered on aggression or domination themes. There are also gender differences in the forms of aggression used by boys and girls. Boys are more physically aggressive while girls are more relationally aggressive, revealing a female bias for social–interactional conflict and a male bias for physical conflict. This is relevant from an evolutionary perspective considering that males have, over evolutionary history, competed with other males for mates and status, while females required greater social fluency in order to acquire resources to help rear their offspring. These gender differences can be seen as antecedents for the sexual roles children will assume as adults, or would have assumed in ancient environments (*see* Human Mate Choice, Evolution of).

Deferred adaptations are most likely to evolve when environmental conditions remain relatively stable over time. This provides children with reliable information about the environments that they are likely to encounter as adolescents and adults. However, although gross aspects of children's environments may be similar over time, and even similar to the environments experienced by ancestral children, many aspects of children's environments cannot be anticipated by natural selection. Some children may be born into resource-rich environments, whereas others may be born into resource-poor environments, along with the stresses associated with such environments. Children need behavioral and cognitive flexibility in order to direct their development in anticipation of adult environments. According to Belsky et al. (1991: p. 650), "a principal evolutionary function of early experience – the first 5 to 7 years – is to induce in the child an understanding of the availability and predictability of resources (broadly defined) in the environment, of the trustworthiness of others, and of the enduringness of close interpersonal relationships, all of which will affect how the developing person apportions reproductive effort." In other words, children need to be sensitive to environmental conditions early in life and adjust some aspects of their physical and psychological development in anticipation to what their future environment is apt to be in all likelihood.

Thus natural selection should favor plasticity and sensitivity to local environments, and this often takes the form of conditional adaptations. Boyce and Ellis (2005) defined conditional adaptations as "evolved mechanisms that detect and respond to specific features of childhood environments – features that have proven reliable over evolutionary time in predicting the nature of the social and physical world into which children will mature – and entrain developmental pathways that reliably matched those features during a species' natural selective history" (p. 290).

Perhaps the most studied topic from this perspective is the rate at which children (especially girls) attain puberty and their future adult mating strategies as a function of their early rearing environment. Belsky and his colleagues (1991) argued that children's early environments are good predictors of what their later environments will be like. An adaptive reproductive strategy ('adaptive' from an evolutionary perspective, not necessarily from the point of view of mainstream society or of mental health professionals) for children growing up in high-stress, low-resource, and unpredictable environments is to reproduce early and often, because it is substantially less likely that any one child will have the requisite resources needed to meet his or her reproductive (evolutionarily driven) goals. The opposite pattern would develop for children growing up in more predictable and supportive homes, with their most adaptive strategy being to postpone reproduction and to invest heavily in children. Nearly two decades of research has essentially confirmed this hypothesis, finding relationships between various types of socioemotional stress measured during childhood and the age at which girls attain menarche and the particulars of their subsequent sexual behavior (see Belsky, 2007; Del Giudice, 2009; Ellis, 2004, for reviews).

Applying Evolutionary Developmental Psychology to Mental Health

Although developmental psychologists were initially reluctant to incorporate evolutionary ideas into their theories and research, the role of evolutionary thinking in explaining important aspects of human development has increased over the past decade. Evolutionary theories can provide new insights into many aspects of development. One important contribution of such thinking is in the area of developmental psychopathology (see, e.g., articles in Ellis and Bjorklund, 2012; Ellis and Boyce, 2011).

The Mental Health Model

The current dominant scientific paradigm for explaining the relationship between high-risk environments and subsequent mental health could be referred to as the *mental health model*, or the *diathesis-stress model*, which assumes that high-risk environments adversely affect children's well-being, promoting *disturbances* in development, even if not clinical disorders per se, and contend that certain individuals are predisposed to develop psychopathologic conditions when exposed to stressful environments. According to this model, supportive environments tend to produce 'good' developmental

outcomes, such as good health, happiness, secure attachment, high self-esteem, emotional regulation, and educational success, whereas negative, stressful environments promote 'bad' developmental outcomes, such as illness, insecure attachment, behavioral problems, drug use, alcoholism, conduct problems, depression, early pregnancy, and limited education. This has led researchers to focus on the deleterious effects of familial and ecological stressors on mental health outcomes (e.g., adolescent onset of psychopathology). Dysfunctional behavior in childhood and adolescence is seen as the natural consequence of exposure to harsh, unpredictable, or uncontrollable socioecological contexts. As such, the mental health model emphasizes *costs* and largely ignores any *benefits* of children's responses to their life situations, making it difficult to explain children and adolescents' motives for their behavior. This leads to the view of children's reactions (e.g., conduct disorder in childhood, adolescent risk behavior) as inherently pathological, thereby requiring preventive interventions.

Evolutionary Developmental Psychological Models of Mental Health

In contrast, an *evolutionary developmental perspective* assumes that natural selection designed the developing brain to respond to different environmental contexts – good and bad – in an adaptive manner. When children encounter stressful environments, therefore, it does not disturb their development so much as it *directs* or *regulates* it toward strategies that are adaptive under stressful conditions – even if those strategies are currently harmful in terms of the long-term welfare of the individual. As stated by Ellis and his colleagues (2012: p. 600), "The evolutionary model posits that natural selection shaped our neurobiological mechanisms to detect and respond to the fitness-relevant costs and benefits afforded by different environments. Most important, these responses are not arbitrary but function adaptively to calibrate developmental and behavioral strategies to match those environments."

These adaptations may well be harmful in terms of the long-term welfare of the individual and society as a whole, but they are well designed to cope with the immediate circumstances. In some cases they are deeply ingrained in our psychology and physiology, based on many thousands of generations of genetic evolution. In other cases they are culturally ingrained, which does not make them any less compelling. Either way, when we understand the local sense in which children and adolescents' responses to high-risk environments are adaptive, a different agenda for research and intervention emerges, as compared with the mental health model.

Using an example summarized earlier, children growing up in high-stress environments with minimal resources and unstable homes would benefit (from an inclusive-fitness perspective) by reproducing early and often, rather than waiting to reproduce (as society would prefer), because it is unlikely that they will have the necessary resources needed to succeed in the long term. It is important to note that just because children and adolescents are adapting to harsh environments does not imply that these are ideal situations or that there are no costs. Harsh environments certainly have costs, sometimes even resulting in death, but this viewpoint recognizes the possible benefits of high-risk behaviors. Avoiding this naturalistic fallacy (i.e., thinking that 'natural' is 'good' or 'justifiable'), one can see how children adapting to high-stress environments allows them to make the best of a bad situation, although it may be seen as a negative outcome by conventional societal standards.

In this vein, several theorists have proposed specific evolutionary models based on the idea that conditional adaptations underpin development of *contingent* survival and reproductive strategies and thus enable individuals to function competently in a variety of different environments. For example, W. Thomas Boyce and Bruce Ellis' (Boyce and Ellis, 2005; Ellis and Boyce, 2008; Ellis et al., 2005, 2011) theory of *biological sensitivity to context* attempts to explain the adaptive relation between early life experience and stress reactivity. It proposes that early environment helps in priming the future stress reactivity, and that early experience with stress interacts with individual differences in aspects of children's personalities, sometimes producing adaptive long-term outcomes and sometimes even producing maladaptive long-term outcomes, depending on the quality of environment children live in. Both genetic dispositions and early experience influence how readily young organisms respond to events in the environment (that is, how sensitive they are to context), which in turn affects their long-term reactions to stress. Those individuals who are especially biologically sensitive to changes in the environment may gain special benefits or experience special deficits from extreme environments, relative to less-sensitive (and more stable) individuals; that is, sensitive individuals are impacted by their environment both *for better and for worse*. This contrasts with the mental health model's exclusive focus on possible negative implications of difficult environments.

This 'for better or for worse' phenomenon can be seen in research from Boyce and his colleagues (1998) with rhesus monkeys. Boyce et al. (1998) measured biological reactivity to novel experiences in a troop of rhesus monkeys living in a low-stress environment. Some monkeys were highly reactive to novel experiences, showing elevated cortisol levels, whereas others were less reactive. Furthermore, an examination of injuries over the course of a year indicated slightly more injures for animals classified as low in reactivity than those classified as high in reactivity. The monkeys then experienced a highly stressful period of confinement to a 1000-square-feet building. The rate in injuries between the high- and low-stress environments did not vary for the monkeys classified as low in biological reactivity. However, rates of injuries skyrocketed for the high-reactive monkeys. Whereas these monkeys had the lowest rate of injuries in the low-stress environment, they had the highest rate in the high-stress environment.

In a study with human children, Boyce and his colleagues (1995) measured the incidence of respiratory infections in 3- to 5-year-old children who were classified as either low or high on the basis of cardiovascular and immunologic reactivity. The incidence of respiratory infection for low-reactive children was about the same regardless if they came from low- or high-stress homes (a little less than three infections per 3-month period) (see also Obradović et al., 2010). In contrast, high-reactive children had the lowest

incidence of infections if they came from low-stress homes (slightly over two infections per 3-month period), but the highest rate if they came from high-stress homes (about 3.5 infections per 3-month period).

In describing high-reactive and low-reactive children, Boyce and Ellis (2005) used the metaphor of the *orchid* and the *dandelion*. Stemming from a Swedish idiomatic expression, *maskrosbarn*, the dandelion child refers to resilient children who are able to thrive in both negative and positive environments. In contrast, the *orkidebarn* refers to the orchid child, described as the "context-sensitive individual, whose survival and flourishing is intimately tied, like that of the orchid, to the nurturant or neglectful character of the ambient environment. In conditions of neglect, the orchid promptly declines, while in conditions of support and nurture, it is a flower of unusual delicacy and beauty" (Boyce and Ellis, 2005: p. 284).

In a related theory, Jay Belsky and his colleagues (Belsky, 2000, 2005, 2007; Belsky et al., 2007) proposed that some children are more sensitive to individual differences in parenting than others, or differentially susceptible to environmental influences. According to *differential susceptibility theory*, when environments are unpredictable, children who are easily able to adapt to a wide range of contexts (that is, who are highly responsive to differences in parenting behavior) will be able to adjust to novel, often less-than-optimal, environments (for example, father absence, insensitive parenting). They will also do particularly well in enriched environments, those providing better-than-average support, resources, and opportunities. Other children, however, are more stable and less influenced by extreme environments. Parents can hedge their bets, so to speak, by producing both types of children, some who will be receptive to change and others who will thrive in the 'expected' environment. Belsky believes that whether children are sensitive to parenting or not is primarily a function of genetics. Who are the children most sensitive to parents' endeavors? A number of researchers have shown that highly fearful, anxious, and 'difficult' children are more influenced by parental behavior (for example, maternal style of discipline) than less-anxious or 'average' children (Gilissen et al., 2008; Kochanska, 1993), as are infants with difficult temperaments (Stright et al., 2008; see Belsky, 2005; Belsky and Pluess, 2009).

In summary, the mental health model, the currently dominant scientific paradigm, is unable to fully explain the relationship between high-risk environments and subsequent mental health as it focuses exclusively on costs of risky actions and does not account for any of the potential evolutionary benefits that may result from engaging in such behaviors. The evolutionary developmental perspective, however, contends that natural selection has shaped children to be able to respond to both positive and negative environments in an adaptive manner. Children who are most susceptible, the orchid children, will thrive in supportive, stable environments, but will have greater adverse effects in less than ideal environments as well. Less susceptible children, the dandelions are more stable and are less likely to be influenced by external factors, irrespective of whether they are good or bad.

The Evolutionary Developmental Model and Risky Adolescent Behavior

Application of these differential sensitivity models can be seen in examinations of risky behavior in adolescents. Adolescents are more likely than any other population to engage in risky behaviors, such as sexual promiscuity, crime, and aggression (Steinberg, 2008). The mental health model would categorize these behaviors as maladaptive outcomes stemming from a negative environment and genetic disposition. This perspective, according to Ellis et al. (2012), only captures the downside of risk without considering why people take risks. Risky behaviors are not maladaptive if the expected benefits outweigh the expected costs.

Adolescent behavior, including risky behavior, should, according to the evolutionary developmental model, be examined for potential fitness benefits. It is important to note again the distinction between risky and maladaptive behaviors; risky behaviors may result in harm to the individual, but may still serve an evolutionary (i.e., fitness) purpose. Given this, however, risky adolescent behaviors still frequently produce negative consequences for the individual and the society as a whole and should therefore be examined in an attempt to minimize them. According to Ellis et al. (2012), viable solutions for interventions require understanding why adolescents engage in risky behaviors, rather than just categorizing them as negative and maladaptive.

Ellis et al. (2012) argue that adolescence serves as a critical period for individuals to develop and master their social milieu and reproductive strategies. Puberty is characterized by specific maturational changes such as the emergence of primary and secondary sexual characteristics, marked increase of appetite, height, and weight, increase in sensation-seeking, high-arousal experiences, and increases in peer aggression and parental conflict. Also apparent is the emergence of sexual motivation and desire to develop and maintain opposite-gender relationships. These changes correspond with many of the behaviors adolescents exhibit: the development of sexual characteristics makes an individual more attractive to the opposite gender, which corresponds to the increase in sexual drive and desire for high-arousal experiences. Higher levels of aggression and parental conflict help adolescents develop autonomy and attain social dominance among their peers, which helps explain the higher-stakes competition that occurs in adolescence over mates (or boyfriends and girlfriends), social status, and friendships.

Relative to adulthood, the adolescent years are associated with higher rates of depression, eating disorders, alcohol and drug use, violence, and sexual promiscuity. Ellis et al. (2012) argue that social feedback is particularly crucial to adolescents, with natural selection favoring increased emotional, behavioral, and neurobiological responses to activity within the social realm. For example, Chein et al. (2011) suggest that the presence of peers influences activation in the adolescent brain, but not the adult brain. Using functional magnetic resonance imaging, Chein et al. (2011) measured brain activity in adolescents and adults as they made decisions in a simulated driving task across two conditions. The first condition had the participant in the simulated driving task alone, whereas the second condition informed the participants

they were being watched by their peers. During the peer-observed condition, adolescent, but not adult, brains exhibited activity in the reward centers of the brain (ventral striatum and orbitofrontal cortex), which predicted subsequent risk-taking behaviors. Consistent with Ellis et al.'s (2012) position, these results suggest that peer influence plays a greater role in regulating social behaviors, particularly risk-taking behaviors, more during adolescence than in adulthood.

Of course, not all adolescents engage in risky behavior to the same extent. The mental health model has been successful in identifying some of the specific environmental conditions that generally produce stressful or supportive environments, which in turn are associated with adolescents' tendencies to engage in risky behaviors or poor mental health outcomes. Differences along the continuum of parental sensitivity, parental harshness, marital quality, parental mental health, and socio-economic status are just a few of the factors that the mental health model examines in order to understand behavioral outcomes. But as Ellis et al. (2012: p. 608) state, the mental health model has not been explicitly informed by evolutionary theory and, consequently, has not focused on or delineated *basic dimensions* of environmental stress and support that guide conditional adaptation. According to this position, guided by *life history theory* – which postulates that important developmental events in an animal's life are shaped by natural selection to promote inclusive fitness (*see* Fertility Theory: Embodied-Capital Theory of Life History Evolution) – individuals make trade-offs that prioritize resource and energy expenditures, such as reproduction (mating, parental investment) and bodily protection (predator defense, immune function), so that greater investment in one domain occurs at the expense of investment in other domains. Life history theory states that individuals will select a strategy that optimizes their trade-offs over the life course. These strategies lie on a slower–faster life continuum. Earlier pubertal timing, more sexual partners, earlier age of reproduction, more offspring with lower parental investment, and more risky behaviors typically characterize a faster life history strategy. A slower life history strategy, on the other hand, is characterized by slower rates of development, delayed age of first sexual encounter and reproduction, fewer offspring with higher parental investment, and exhibiting fewer risky behaviors (see Figure 1).

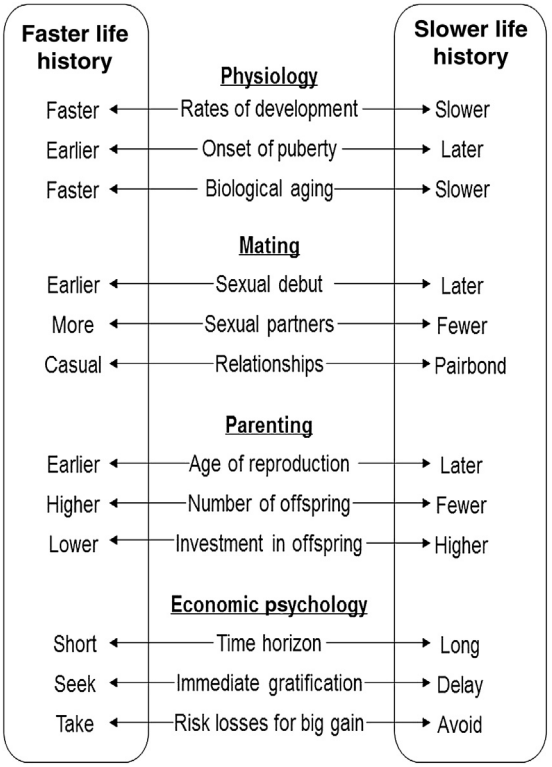

Figure 1 Faster versus slower life history strategies. Reproduced from Ellis, B.J., Dahl, R.E., Del Giudice, M., Dishion, T.J., Figuerdo, A.F., Gray, P., Hawley, P., Jackson, J.J., Jacobs, W.J., Volk, A.A., Wilson, D.S., 2012. The evolutionary basis of risky adolescent behavior: Implications for science, policy, and practice. Developmental Psychology 48, 598–623.

An Evolutionary Life History Framework: Harshness and Unpredictability

Recent research following these models has elaborated on children's sensitivity to context by going beyond the study of harsh environments and also examining the unique predictive value of variations in environmental predictability as well (Belsky et al., 2012; Ellis et al., 2009; Simpson et al., 2012). Ellis et al. (2009) proposed an evolutionary life history framework, stating how different environmental conditions early in development affect the fast versus slow life history trajectories, specifically the influence of harshness and early unpredictability. Ellis et al. (2009) defined harshness as age-specific rates of morbidity and mortality, predictably measured by socioeconomic status (SES); they defined unpredictability as the fluctuations of harshness across development. Unpredictability is measured by indicators such as frequent residential changes, parental divorce, and remarriage. Ellis et al. (2009) proposed that each of these environmental factors should uniquely contribute to differential outcomes in life history strategies. For example, in one recent study, family social disadvantage (harshness) predicted deviant, often criminal, behavior and the number of children one had in early adulthood (Dishion et al., 2012). Other research has shown that the predictability of a child's early environment can have an independent effect on a child's developmental trajectory. For example, Belsky et al. (2012) tested a model proposing that higher environmental harshness and higher levels of unpredictability in the first 5 years of the child's life will independently predict an accelerated life history trajectory. They found that unpredictability of a child's environment in the first 5 years is a more direct predictor of having a larger number of sexual partners, a signal of an accelerated life history strategy, at 15 years of age, while early harshness had a more indirect effect.

Simpson and his colleagues (2012) extended this line of research by looking at multiple life history variables, as opposed to just the number of sexual partners, and extended the outcome measure (number of sexual partners) to the age of 23. They also examined whether predictability and harshness are more influential on development when occurring in early (prenatal to the stage of 5-year-old children) vs later childhood (6- to 10-year-old children). Supporting the results in Belsky et al. (2012), they reported that exposure to harshness and

(especially) unpredictability in the first 5 years of childhood significantly and independently predicted an accelerated life history trajectory by having higher levels of aggression and incidences of delinquent behavior, more sexual partners, and higher likelihood of being associated with a criminal act. As expected, exposure to harshness and unpredictability in later childhood (6- to 10-year-old children) were not significant predictors of these outcomes.

If the environment is unstable and fluctuates often, it may be more beneficial for the individual to develop a faster life history strategy, reproducing faster and more often, in anticipation that future environments will be unpredictable. Alternatively, if one's early environment is stable, this is a good indication that future environments will also be stable, and a slower life history strategy would be more beneficial in the long run. From an evolutionary perspective, adolescents exhibiting risky behaviors, sexual promiscuity, and early reproduction are exhibiting adaptive behaviors that are best suited for their environmental context. Exactly these patterns have been observed in several recent studies (Ellis et al., 2009; Simpson et al., 2012), independent of measures of harshness, finding strong support for the evolutionary developmental models.

The Evolutionary Developmental Model and Antisocial Behavior

An evolutionary developmental perspective can also shed light on antisocial behaviors, such as bullying or stealing. Research by Hawley (1999, 2002, 2003) suggests that behaviors that would normally be characterized as antisocial are often quite successful in attainting resources for individuals exhibiting them, at least in some environments. Hawley (1999) posited three classes of children with respect to resource control: prosocial controllers, coercive controllers, and bistrategic controllers. Prosocial controllers gain and maintain access to resources by cooperating in a group, are socially skilled, and are typically friendly and well liked by peers. Coercive controllers are strictly aggressive in their attainment of resources. They tend to be impulsive, aggressive, and socially rejected. In contrast, bistrategic controllers use both prosocial/cooperative and antisocial/coercive strategies to attain and maintain resources, but tend to be socially savvy, attractive social partners, who also exhibit aggression and bullying similar to that of coercive controllers.

From this perspective, aggressive behaviors, although disruptive in a classroom and potentially leading to injuries, especially among adolescents, serve to obtain resources, access to mates or friends, and achieve social dominance (Sijtsema et al., 2009; Veenstra et al., 2010), with some children using aggressive behaviors strategically in combination with prosocial behavior. Furthermore, the adaptive benefits of bullying are evident by data that children identified as bullies tend to date at younger ages and engage in more dating activities than their nonbully counterparts (Connolly et al., 2000; Faris and Felmlee, 2011; Gallup et al., 2011).

In summary, aggression may have evolutionary benefits in attaining resources and social dominance, and ultimately, mating opportunities. The fact that aggressive behavior, often in the form of bullying, frequently has some benefits for children and adolescents in no way condones such behavior, but rather provides insights into why some intervention programs are destined to failure and others have a chance of succeeding.

Conclusion

Darwin's theory of natural selection is the driving force for modification and survival over time. Evolutionary explanations have been widely used to examine adult behavior and psychology; however, the field of evolutionary developmental psychology contends that natural selection acts throughout ontogeny and not simply only in adulthood. Individuals must survive the often-hazardous periods of infancy and childhood before reaching adulthood, suggesting that any adaptations that increase the chance of surviving to adulthood during the early stages of life would certainly be favored by natural selection.

An evolutionary approach to infancy, childhood, and adolescence provides researchers and practitioners with new insights with which to view both typical and atypical development. In particular, when 'problematic' behavior, such as risk-taking or aggression, is observed in children or adolescents, an evolutionary approach causes one to ask "What's in it for the child?". That is, although bullying or sexual promiscuity in children and adolescence, for example, are clearly behaviors that society has every right to want to minimize, an evolutionary developmental perspective recognizes that these behaviors may provide some adaptive value for some children in some specifiable contexts.

Applying evolutionary ideas to help understand both typical and atypical development does not mean that the other accounts (e.g., learning, social, cognitive) are unnecessary. Rather, an evolutionary perspective supplements other approaches to understanding development and, in the end, produces a more complete, and from an application perspective perhaps more useful, account of development. In fact, one might modify once again Dobzhansky's famous statement and say that "nothing in childhood and development make sense exception in the light of evolution" (cf Konner, 2010).

See also: Cognition, Evolution of; Darwinism; Epigenetic Inheritance; Fertility Theory: Embodied-Capital Theory of Life History Evolution; Human Mate Choice, Evolution of; Microevolution.

Bibliography

Alley, T.R., 1981. Head shape and the perception of cuteness. Developmental Psychology 17, 650–654.

Belsky, J., 2000. Conditional and alternative reproductive strategies: individual differences in susceptibility to rearing experience. In: Rodgers, J., Rowe, D., Miller, W. (Eds.), Genetic Influences on Human Fertility and Sexuality: Theoretical and Empirical Contributions from the Biological and Behavioral Sciences. Kluwer, Boston, MA, pp. 127–146.

Belsky, J., 2005. Differential susceptibility to rearing influences: an evolutionary hypothesis and some evidence. In: Ellis, B.J., Bjorklund, D.F. (Eds.), Origins of the Social Mind: Evolutionary Psychology and Child Development. Guildford, New York, NY, pp. 139–163.

Belsky, J., 2007. Experience in childhood and the development of reproductive strategies. Acta Psychologica Sinica 39, 454–468.
Belsky, J., Bakermans-Kranenburg, M.J., van Ijzendoorn, M.H., 2007. For better and for worse: differential susceptibility to environmental influences. Current Directions in Psychological Science 16, 300–304.
Belsky, J., Pluess, M., 2009. Beyond diathesis-stress: differential susceptibility to environmental influences. Psychological Bulletin 135, 885–908.
Belsky, J., Schlomer, G.L., Ellis, B.J., 2012. Beyond cumulative risk: distinguishing harshness and unpredictability as determinants of parenting and early life history strategy. Developmental Psychology 48, 662–673.
Belsky, J., Steinberg, L., Draper, P., 1991. Childhood experience, interpersonal development and reproductive strategy: an evolutionary theory of socialization. Child Development 62, 647–670.
Bjorklund, D.F., 1997. The role of immaturity in human development. Psychological Bulletin 122, 153–169.
Bjorklund, D.F., Causey, K., Periss, V., 2010a. The evolution and development of human social cognition. In: Kappeler, P., Silk, J. (Eds.), Mind the Gap: Tracing the Origins of Human Universals. Springer Verlag, Berlin, pp. 351–371.
Bjorklund, D.F., Cormier, C., Rosenberg, J.S., 2005. The evolution of theory of mind: big brains, social complexity, and inhibition. In: Schneider, W., Schumann-Hengsteler, R., Sodian, B. (Eds.), Young Children's Cognitive Development: Interrelationships among Executive Functioning, Working Memory, Verbal Ability and Theory of Mind. Erlbaum, Mahwah, NJ, pp. 147–174.
Bjorklund, D.F., Ellis, B.J., 2005. Evolutionary psychology and child development: an emerging synthesis. In: Ellis, B.J., Bjorklund, D.F. (Eds.), Origins of the Social Mind: Evolutionary Psychology and Child Development. Guilford, New York, pp. 3–18.
Bjorklund, D.F., Ellis, B.J., Rosenberg, J.S., 2007. Evolved probabilistic cognitive mechanisms: an evolutionary approach to gene-environment-development. In: Kail, R.V. (Ed.), Advances in Child Development and Behavior, vol. 35. Elsevier, Oxford, pp. 1–39.
Bjorklund, D.F., Gaultney, J.F., Green, B.L., 1993. I watch therefore I can do: the development of meta-imitation over the preschool years and the advantage of optimism in one's imitative skills. In: Pasnak, R., Howe, M.L. (Eds.), Emerging Themes in Cognitive Development, Competencies, vol. II. Springer-Verlag, New York, pp. 79–102.
Bjorklund, D.F., Green, B.L., 1992. The adaptive nature of cognitive immaturity. American Psychologist 47, 46–54.
Bjorklund, D.F., Hernández Blasi, C., 2005. Evolutionary developmental psychology. In: Buss, D. (Ed.), Evolutionary Psychology Handbook. Wiley, New York, pp. 828–850.
Bjorklund, D.F., Hernández Blasi, C., Periss, V., 2010b. Lorenz revisited: the adaptive nature of children's supernatural thinking. Human Nature 21, 371–392.
Bjorklund, D.F., Pellegrini, A.D., 2000. Child development and evolutionary psychology. Child Development 71, 1687–1708.
Bjorklund, D.F., Pellegrini, A.D., 2002. The Origins of Human Nature: Evolutionary Developmental Psychology. American Psychological Association, Washington, DC.
Bjorklund, D.F., Periss, V., Causey, K., 2009. The benefits of youth. European Journal of Developmental Psychology 6, 120–137.
Boyce, W.T., Chesney, M., Alkon–Leonard, A., Tschann, J., Adams, S., Chesterman, B., Cohen, F., Kaiser, P., Folkman, S., Wara, D., 1995. Psychobiologic reactivity to stress and childhood respiratory illnesses: results of two prospective studies. Psychosomatic Medicine 57, 411–422.
Boyce, W.T., Ellis, B.J., 2005. Biological sensitivity to context: I. An evolutionary-developmental theory of the origins and functions of stress reactivity. Development & Psychopathology 17, 271–301.
Boyce, W.T., O'Neill-Wagner, P., Price, C.S., Haines, M., Suomi, S.J., 1998. Crowding stress and violent injuries among behaviorally inhibited rhesus macaques. Health Psychology 17, 285–289.
Burgess, R., MacDonald, K. (Eds.), 2005. Evolutionary Perspectives on Human Development. Sage Publications, Thousand Oaks, CA.
Burkhardt, F., Pearn, A.M., Evans, S. (Eds.), 2008. Evolution: Selected Letters of Charles Darwin 1860–1870. Cambridge University Press, UK.
Buss, D.M., 1995. Psychological sex differences: origins through sexual selection. American Psychologist 50, 164–168.
Chein, J., Albert, D., O'Brien, L., Uckert, K., Steinberg, L., 2011. Peers increase adolescent risk taking by enhancing activity in the brain's reward circuitry. Developmental Science 14, F1–F10.
Connolly, J., Pepler, D., Craig, W., Taradash, A., 2000. Dating experiences of bullies in early adolescence. Child Maltreatment 5, 299–310.
Daly, M., Wilson, M., 1988. Homicide. Aldine, Hawthorne, NY.
Darwin, C., 1859. On the Origin of Species. J. Murray, London.
Dawkins, R., 1976/2006. The Selfish Gene. Oxford University Press, Oxford.
DeCasper, A.J., Fifer, W.P., 1980. Of human bonding: newborns prefer their mother's voices. Science 208, 1174–1176.
Del Giudice, M., 2009. Sex, attachment, and the development of reproductive strategies. Behavioral and Brain Sciences 32, 1–21.
Dishion, T.J., Ha, T., Véronneau, M.-H., 2012. An ecological analysis of the effects of deviant peer clustering on sexual promiscuity, problem behavior, and childbearing from early adolescence to adulthood: an enhancement of the life history framework. Developmental Psychology 48 (3), 703–717.
Dobzhansky, Y., 1937. Genetics and the Origin of Species. Colombia University Press, New York.
Dobzhansky, T., 1973. Nothing in biology makes sense except in the light of evolution. American Biology Teacher 35, 125–129.
Dunbar, R.I.M., 1995. Social networks, support cliques, and kinship. Human Nature 6, 273–290.
Dunbar, R.I.M., 2010. Brain and behaviour in primate evolution. In: Kappeler, P.M., Silk, J.B. (Eds.), Mind the Gap: Tracing the Origins of Human Universals. Springer, New York, pp. 315–330.
Ellis, B.J., 2004. Timing of pubertal maturation in girls: an integrated life history approach. Psychological Bulletin 130, 920–958.
Ellis, B.J., Bjorklund, D.F. (Eds.), 2005. Origins of the Social Mind: Evolutionary Psychology and Child Development. Guilford, New York.
Ellis, B.J., Bjorklund, D.F. (Eds.), 2012. Beyond mental health: an evolutionary analysis of development under risky and supportive environmental conditions: an introduction to the special section. Developmental Psychology 48, pp. 591–597.
Ellis, B.J., Boyce, W.T., 2008. Biological sensitivity to context. Current Directions in Psychological Science 17, 183–187.
Ellis, B.J., Boyce, W.T. (Eds.), 2011. Differential susceptibility to context: toward an understanding of sensitivity to developmental experiences and context. Development and Psychopathology 23, pp. 1–5.
Ellis, B.J., Boyce, W.T., Belsky, J., Bakermans-Kranenburg, M.J., van IJzendoorn, M.H., 2011. Differential susceptibility to the environment: an evolutionary-neurodevelopmental theory. Development and Psychopathology 23, 7–28.
Ellis, B.J., Dahl, R.E., Del Giudice, M., Dishion, T.J., Figuerdo, A.F., Gray, P., Hawley, P., Jackson, J.J., Jacobs, W.J., Volk, A.A., Wilson, D.S., 2012. The evolutionary basis of risky adolescent behavior: implications for science, policy, and practice. Developmental Psychology 48 (3), 598–623.
Ellis, B.J., Essex, M.J., Boyce, W.T., 2005. Biological sensitivity to context: II. Empirical explorations of an evolutionary–developmental theory. Development and Psychopathology 17, 303–328.
Ellis, B.J., Figueredo, A.J., Brumbach, B.H., Schlomer, G.L., 2009. Fundamental dimensions of environmental risk: the impact of harsh versus unpredictable environments on the evolution and development of life history strategies. Human Nature 20, 204–268.
Faris, R., Felmlee, D., 2011. Status struggles: network centrality and gender segregation in same- and cross-gender aggression. American Sociological Review 76, 48–73.
Gallup, A.C., O'Brien, D.T., Wilson, D.S., 2011. Intrasexual peer aggression and dating behavior during adolescence: an evolutionary perspective. Aggressive Behavior 37, 1–10.
Geary, D.C., Bjorklund, D.F., 2000. Evolutionary developmental psychology. Child Development 71, 57–65.
Gilissen, R., Bakersman-Kranenburg, M.J., van IJzendoorn, M.H., van der Veer, R., 2008. Parent-child relationship, temperament, and physiological reactions to fear-inducing film clips: further evidence for differential susceptibility. Journal of Experimental Child Psychology 99, 182–195.
Glocker, M.L., Langleben, D.D., Ruparel, K., Loughead, J.W., Gur, R.C., Sachser, N., 2009. Baby schema in infant faces induces cuteness perception and motivation for caretaking in adults. Ethology 115, 257–263.
Gottlieb, G., 1991. Experiential canalization of behavioral development: theory. Developmental Psychology 27, 4–13.
Gottlieb, G., 2002. Developmental-behavioral initiation of evolutionary change. Psychological Review 109, 211–218.
Gottlieb, G., 2007. Probabilistic epigenesis. Developmental Science 10, 1–11.
Gould, S.J., 1977. Ontogeny and Phylogeny. Harvard University Press, Cambridge.
Hawley, P.H., 1999. The ontogenesis of social dominance: a strategy-based evolutionary perspective. Developmental Review 19, 97–132.
Hawley, P.H., 2002. Social dominance and prosocial and coercive strategies of resource control in preschoolers. International Journal of Behavioral Development 26, 167–176.
Hawley, P.H., 2003. Prosocial and coercive configurations of resource control in early adolescence: a case for the well-adapted Machiavellian. Merrill-Palmer Quarterly 49, 279–309.

Hernández Blasi, C., Bjorklund, D.F., 2003. Evolutionary developmental psychology: a new tool for better understanding human ontogeny. Human Development 46, 259–281.

Huxley, J., 1942. Evolution, the Modern Synthesis. Harper Brothers, New York.

Jablonka, E., Lamb, M.J., 1995. Epigenetic Inheritance and Evolution. Oxford University Press, London.

Jablonka, E., Lamb, M.J., 2005. Evolution in Four Dimensions: Genetic, Epigenteic, Behavioral, and Symbolic Variation in the History of Life. MIT Press, Cambridge, MA.

Kochanska, G., 1993. Toward a synthesis of parental socialization and child temperament in early development of conscience. Child Development 64, 325–347.

Konner, M., 2010. The Evolution of Childhood: Relationships Emotions, Mind. Belknap Press, Cambridge, MA.

Lickliter, R., Honeycutt, H., 2003. Developmental dynamics: toward a biologically plausible evolutionary psychology. Psychological Bulletin 129, 819–835.

Lopez, D., Little, T.D., Oettingen, G., Baltes, P.B., 1998. Self-regulation and school performance: is there an optimal level of action-control? Journal of Experimental Child Psychology 70, 54–74.

Lorenz, K.Z., 1943. Die angeboren Formen moglicher Erfahrung (The innate forms of possible experience). Zeitschrift fur Tierpsychologie 5, 233–409.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., 1988. A precursor of language acquisition in young infants. Cognition 29, 143–178.

Nittono, H., Fukushima, M., Yano, A., Moriya, H., 2012. The power of Kawaii: viewing cute images promotes a careful behavior and narrows attentional focus. PLOSone 7, 1–7.

Obradović, J., Bush, N.R., Stamperdahl, J., Adler, N.E., Boyce, W.T., 2010. Biological sensitivity to context: the interactive effects of stress reactivity and family adversity on socio-emotional behavior and school readiness. Child Development 81, 270–289.

Pellegrini, A.D., 2012. Play. In: Zelazo, P. (Ed.), Oxford Handbook of Developmental Psychology. Oxford University Press, New York.

Periss, V., Hernández Blasi, C., Bjorklund, D.F., 2012. Cognitive "babyness": developmental differences in the power of young children's supernatural thinking to influence positive and negative affect. Developmental Psychology 48, 1203–1214.

Sherman, G.D., Haidt, J., Coan, J.A., 2009. Viewing cute images increases behavioral carefulness. Emotion 9, 282–286.

Shin, H.-E., Bjorklund, D.F., Beck, E.F., 2007. The adaptive nature of children's overestimation in a strategic memory task. Cognitive Development 22, 197–212.

Sijtsema, J.J., Veenstra, R., Lindenberg, S., Salmivalli, C., 2009. Empirical test of bullies' status goals: assessing direct goals, aggression, and prestige. Aggressive Behavior 35, 57–67.

Simpson, J., Griskevicius, V., Kuo, S., Sung, S., Collins, A., 2012. Evolution, stress, and social development: the influence of harsh vs. unpredictable environments on sex and risky behavior. Developmental Psychology 48, 674–686.

Spencer, J.P., Blumberg, M.S., McMurray, B., Robinson, S.R., Samuelson, L.K., Tomblin, J.B., 2009. Short arms and talking eggs: why we should no longer abide the nativist–empiricist debate. Child Development Perspectives 3, 79–87.

Sprengelmeyer, R., Perrett, D.I., Fagan, E.C., Cornwell, R.E., Lobmaier, J.S., Sprengelmeyer, A., Aasheim, H.B.M., Black, I.M., Cameron, L.M., Crow, S., Milne, N., Rhodes, E.C., Young, A.W., 2009. The cutest little baby face: a hormonal link to sensitivity to cuteness in infant faces. Psychological Science 20, 149–154.

Steinberg, L., 2008. A social neuroscience perspective on adolescent risk-taking. Developmental Review 28, 78–106.

Stright, A.D., Gallagher, K.C., Kelley, K., 2008. Infant temperament moderates relations between maternal parenting in early childhood and children's adjustment in first grade. Child Development 79, 186–200.

Tooby, J., Cosmides, L., 1992. The psychological foundations of culture. In: Barkow, J.H., Cosmides, L., Tooby, J. (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, New York, pp. 19–139.

Veenstra, R., Lindenberg, S., Munniksma, A., Dijkstra, J.K., 2010. The complex relation between bullying, victimization, acceptance, and rejection: giving special attention to status, affection, and sex differences. Child Development 81, 48–486.

Volk, T., Atkinson, J., 2008. Is child death the crucible of human evolution? Journal of Social, Evolutionary and Cultural Psychology 2, 247–260.

West-Eberhard, M.J., 2003. Developmental Plasticity and Evolution. Oxford University Press, New York.

Evolutionary Economics

Christian Cordes, University of Bremen, Bremen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

There are several ways to incorporate evolutionary concepts into economic thinking. This article reviews the most important transfers of this kind into evolutionary economics. I broadly differentiate between approaches that draw on an analogy construction to the biological sphere, those that make metaphorical use of Darwinian ideas, and avenues that are based on the fact that other forms of – cultural – evolution rest on foundations laid before by natural selection. It is shown that an evolutionary approach within economics informed by insights from cognitive science, evolutionary biology, and anthropology contributes to more realistic models of human behavior in economic contexts.

Introduction

Evolutionary economics is a paradigm for explaining economic change. In principle, there are several ways to incorporate evolutionary concepts into economic thinking to reach this explanatory goal. This problem has triggered an ongoing conceptual-methodological debate in the evolutionary economics community (see Knudsen, 2004; Hodgson and Knudsen, 2006; Cordes, 2006; Witt, 2003). In this article, I broadly differentiate between two potential channels of concept transfer. First, economics can be turned into an evolutionary science by introducing aspects of Darwin's explanatory model of biological evolution (*see* Darwinism) to economic theory via the construction of analogies or the metaphorical use of Darwinian ideas. Second, Darwinian concepts are relevant for explaining the basis on which economic evolution rests. On the one hand, biological evolution defines the long-term constraints on human-made economic evolution. On the other hand, innate dispositions of humans that have been shaped by natural selection earlier in humans' phylogeny define the cognitive foundations on which economic evolution rests. We call this second way of incorporating evolutionary concepts to economics a 'naturalistic approach.' Common to both approaches is the criticism of the widespread tradition in economics to derive human behavior from abstract principles without proper reference to insights from the evolutionary sciences.

Analogies to the Biological Sphere and the Metaphorical Use of Evolutionary Concepts in Evolutionary Economics

So-called neo-Schumpeterians within the evolutionary economics camp consider the concept of selection (*see* Adaptation, Fitness, and Evolution; Microevolution) as a core element of evolutionary economics. Selection processes, they argue, act on the diversity of behaviors and entail behavioral adaptations of agents who are only boundedly rational (i.e., their rationality is limited by the information available to them and by their particular cognitive and time constraints). This fact prevents optimal adaptation to external environments by the agents themselves. Nelson and Winter's (1982) as well as Metcalfe's (1994) approaches belong to this strand of thinking. Although the former relies on a largely metaphorical import of biological concepts to economics, the latter draws on a more direct application of a model of natural selection to economic competition. Finally, the idea of a generalized Darwinism is based on a close analogy to the biological realm and claims that a set of abstract Darwinian principles also can be applied to economic evolution.

Nelson and Winter's Concept of Organizational Routines

Nelson and Winter (1982) offer an evolutionary theoretic view of the firm that is inspired by previous works of Schumpeter (1934), Penrose (1959), Simon (1955), and Cyert and March (1992). Their theory is based on the notion of organizational routines and draws on a metaphorical use of the Darwinian theory of natural selection. Their behavioral assumptions deviate substantially from those of standard economics: Firms are not profit maximizers but follow rather rigid rules or routines, and agents, including managers, are only boundedly rational. For instance, there are two types of search rules for finding an improved technology: First, innovating by conducting a local search close to the currently applied technology, whereby the probability of success decreases with an increasing distance to the actual technology. Second, by imitating technologies of competitors in one's industry by observational learning, firms can adopt new technologies. In this context, an erroneous estimation of the new technology's input coefficients by managers is always possible.

Routines define the firm's core capabilities and are defined as regular patterns of deliberating and acting in organizational processes. They are partly unconscious or tacit and are questioned only in the case of insufficient performance ('satisficing behavior,' see Simon, 1955). Routines enable coordinated behavior (behavior of others can be anticipated) while avoiding great decision effort on the part of cognitively constrained individuals. There is a hierarchy of organizational routines: lower order organizational skills and higher order decision procedures. Routines serve as a firm's 'memory' and are based on metaindividual, partly implicit shared knowledge. On the one hand, routines allow for fast and effective behavior in well-defined situations. On the other

 http://dx.doi.org/10.1016/B978-0-08-097086-8.81019-3

hand, there are constraints in modifying, replicating, and imitating routines, which might entail dysfunctionality in changing business environments and organizational inertia.

Moreover, to innovate in a dynamic Schumpeterian environment, firms choose different strategies and have different core capabilities (e.g., in research and development), that is, there is variety within the population of firms, which is a prerequisite for market selection to take place. The focus is on competition between heterogeneous firms instead of a representative firm or macroproduction functions. In this context, firm strategies are as much a matter of, for example, faith of top management or company tradition, as they are of calculation, and there is no reason to argue a priori that firm commitments are optimal. Nelson and Winter's theory involves the general Darwinian principles (*see* Darwinism) of variation, retention, and selection: Variation is introduced by firms developing new routines, retention is facilitated by stable routines, and selection is represented by competition between firms endowed with varying sets of routines. The latter process determines the differential dissemination of routines, that is, modified 'gene frequencies' in the 'gene pool,' via differential firm growth through the reinvestment of profits. Routines are considered as a firm's genes and a population of firms represents an industry's 'gene pool.' Diversity in firm behaviors tends to be eroded by market selection.

Nelson and Winter's formal approach is based on a replicator dynamic – a concept stemming from population genetics – that describes how profitable firms' capital stocks grow and unprofitable firms' capital stocks shrink. They then carry through a series of computer simulations of their model of firm competition. The final modeling aim of Nelson and Winter is to reproduce the results of the neoclassical (Solowian) growth model in a formal approach that includes competition via innovation and imitation. Their approach amounts to a revival of Schumpeterian ideas in combination with a biological metaphor. Moreover, it provides an alternative growth model that is based on a more realistic description of firm behavior and competitive processes.

Metcalfe's Use of the Replicator Dynamic

Selectionist models (*see* Adaptation, Fitness, and Evolution; Microevolution) analogous to phylogenetic biological evolution are prominent in many works in the field of evolutionary economics. One prominent example is the work of Stanley Metcalfe (1988, 1994). He analyzes the diffusion of innovation in an economy with the help of a replicator dynamic (analogous to 'Fisher's principle' in population genetics). In general, diffusion analysis is concerned with how the economic significance of a new technology changes over time, which is an aspect of structural change: new technologies are integrated into the economy and impose changes upon its structure. Metcalfe shows that the relative profitability of competing technologies or innovations is important, not their absolute profitability, and that the former will change during the diffusion process.

His basic argument relies on technological variety across firms as the basis for competitive advantage and a determinant of how rival technologies diffuse relative to one another. The rate of diffusion of an innovation is linked directly to that innovation's distance from average-practice technology, that is, economic variety is correlated with the pace of selection: The change of a technology's market share is proportional to the distance between its own unit cost and population average unit cost. By this process, the lowest cost, best-practice technology is brought to dominance (other technologies dropping in relative terms to zero). Selection works on economic variety at a pace that is directly proportional to economic variety at any point in time, a phenomenon known as the 'distance from the mean principle' that underlies many selection processes. In the course of this process, selection destroys the variety on which it depends. There is, however, an endogenous relation between selection and variety generation: Selection will, via induced technological creativity, generate new variety – as is assumed in Schumpeterian economic development.

Metcalfe's model reproduces empirically observable stylized developments of variables that also have been the subject matter of neoclassical growth models (e.g., capital accumulation in Solow-type growth models), industrial economics (e.g., within-industry competition processes), or diffusion theory (e.g., the dominant stylized fact that the usage of new technologies over time typically follows an S-curve). Diffusion phenomena, Metcalfe argues, should be treated as part of the broader picture of competition and structural change in capitalist economies because they describe the unfolding of economic structures as a consequence of variety in behavior or innovative activity. Therefore, at the root of the replicator dynamic is the notion that competition is driven by technological or behavioral variety, implying that there is no representative firm or technology and no static equilibrium.

Generalized Darwinism: Applying a Core Set of Darwinian Principles

Generalized Darwinism shows how abstract principles derived from Darwin's explanatory model of biological evolution (*see* Darwinism) can be applied to cultural evolution. In his article "Darwinism in Economics: From Analogy to Ontology," Hodgson (2002) expounded his notion of this concept (labeled here as 'universal Darwinism'): a core set of Darwinian principles that, along with auxiliary explanations specific to each scientific domain, is considered applicable to a wide range of phenomena. Hodgson argued that evolutionary aspects of the biological and the cultural spheres both involve the general Darwinian principles of variation, inheritance, and selection. Generalized Darwinism seeks to apply Darwinian principles to cultural evolution (*see* Cultural Evolution: Overview; Evolutionary Epistemology), that is, it is argued that these principles also are valid for explaining other forms of evolution. Consequently, generalized Darwinism is proposed as a broad theoretical framework for the analysis of the evolution of all open, complex systems, including economic systems. Hodgson has argued that Darwinian evolution shares these common general features with economics due to the fact that both biology and the social sciences address evolving systems. Thus, he says, Darwin's explanatory model provides a theoretical framework and ontological precepts. Hence, generalized Darwinism suggests making substantial use of Darwinian concepts, that is, the latter are not applied in a heuristic or metaphorical form only.

The intention of a generalized Darwinism is to apply Darwinian principles as a general purpose tool to various forms of evolution sharing a common ontological basis (see Hodgson and Knudsen, 2004; Knudsen, 2004; Stoelhorst, 2008). These principles – variation, inheritance, and selection – are an abstract reduction derived from Darwinian theory in evolutionary biology and, by analogy, are transferred to economics. Knudsen argued, "that an appreciation of the causal structure of neo-Darwinism may help develop a general selection theory which can be used as a basis for a development of evolutionary economics" (Knudsen 2002, p. 446). Hodgson (2002, 2003) introduced socioeconomic replicators, such as habits, routines, and institutions, which are compared with DNA and that subsequently are subject to selection. Moreover, he stated that socioeconomic replicators make copies of themselves, although imperfect ones. Proponents of a generalized Darwinism argue that abstraction removes the particular content from the specific domain in which principles apply. Hence, they maintain, the abstract principles of Darwinian evolution can be applied in different evolutionary domains.

The Naturalistic Interpretation within an Evolutionary Economics

The approaches in this section all focus on the lasting influence of humans' biological inheritance on contemporary behavior and the constraints on economic evolution set by biological evolution.

American Institutionalism: Turning Economics into an Evolutionary Science

Since their emergence during the second half of the nineteenth century, Darwin-inspired, instinct-based theories of human agency have delivered the psychological foundations of American Institutionalism (e.g., Rutherford, 2000; Asso and Fiorito, 2004). In contrast to the modern ethological understanding of the term 'instincts,' institutionalists at that time did not consider them to be more or less fixed behaviors, but as a bundle of flexible dispositions or propensities that can be suppressed or diverted (Hodgson, 2004, p. 162). Present-day cognitive scientists would speak of 'psychological dispositions' when addressing the same phenomena. This understanding of instincts as evolved, innate human dispositions by institutionalists enabled cross-fertilization between economics, evolutionary biology, and psychology. Psychologist William McDougall (1910, 1924) laid the foundation of instinct theory. He argued that, directly or indirectly, instincts are the prime movers of all human activity by determining the latter's ends. McDougall defined instincts as innate psychophysical dispositions that determine what agents perceive, pay attention to, and experience as an emotional excitement, and the impulses to act on objects in a particular manner (1910: Chapter II). Humans' intellectual apparatus then is the instrument by which these impulses seek their satisfaction. Consequently, at the beginning of the twentieth century, early institutionalists like Thorstein Veblen, Wesley Mitchell, John R. Commons, and others employed instinct theory in their analysis of economic behavior and focused especially on the interaction between instinctive motivation and intentional economic action (Mitchell, 1910; Veblen, 1914; Clark, 1918; Tead, 1918; Commons, 1920; Parker, 1920).

A path-breaking strand of this interpretation of an evolutionary economics reaches back to Thorstein Veblen. In 1898, Veblen (1898) put the question "Why is economics not an evolutionary science?" to the audience and delivered an early theoretical concept for an evolutionary approach to economics, inspired by contemporary Darwinian thought. Veblen developed a theory of institutional evolution that takes into account the cumulative nature of the cause and effect sequence of the process of institutional development. The first part of Veblen's "Instinct of Workmanship" (1914) contains the most systematic formulation of his theory of institutional change and its psychological foundations. Instincts are a key component of Veblen's theory of institutional evolution (see also Ayres, 1958).

Veblen (1919, p. 231) criticized the hedonistic, selfish, and rationalistic psychology of orthodox economics. As an alternative, he suggested a psychology based on instincts and habits that was crucially inspired by McDougall's work and integrated it into his model of institutional change (see Mitchell, 1914; Rutherford, 2001). He considered the determinants of human behavior to be the main explanatory variables and driving forces of institutional change. Thereby, behavior responds to two kinds of determinants: the basic drive of instinctive factors and the drive to conform to habits. With respect to the capacity to motivate behavior, instincts and habits are functionally equivalent. Habits, as elements of the socially approved scheme of conduct and pursuit, become proximate ends of endeavor. They may even occupy the interests of agents to such an extent that the instinctively given ulterior purpose temporarily loses influence (Veblen, 1914, p. 7). Although the formation of habits relies on evolved sophisticated forms of social learning, instincts are based directly on innate cognitive dispositions. In general, Veblen identified instincts as specific innate tendencies of the mind that have evolved in the process of adaptation of species to their environment.

Veblen's theory of cumulative evolutionary change takes the following shape (see Rutherford, 1998): The starting point of institutional change always consists of the given instinctive endowment of humans – including, for example, the instinct of workmanship that underlies technological creativity, and social instincts, such as the parental bent that aims at the welfare of the family or group, including nonrelatives. Objectives that people pursue are, in line with McDougal, a matter of instinct (Veblen, 1914, p. 3). It is by the prompting of instinct that reflection, deliberation, and technological creativity come to be employed. The established way of life making up society's material and technical conditions gives rise to corresponding habits of action as well as thought and finally institutions. The latter are supported by social sanction and are passed on through socialization. Institutions are largely "a matter of tradition out of the past, a legacy of habits of thought accumulated through the experience of past generations" (Veblen, 1914, p. 7). If the existing institutional

circumstances are not too obstructive, the instinctive nature of humans will express itself in ways that result in new technological insights, new corresponding habits of action and thought, and ultimately new institutions.

The Evolution of Institutions: Hayek and North

In a manner similar to Veblen's ideas of human nature, F. A. Hayek – as a representative of the Austrian School – assumed a core set of human instincts that evolved in the primitive tribal societies of prehistory. Hayek's theory of their exact nature is, however, much less developed than Veblen's (see Hodgson, 2007). Following Hayek, innate behavioral dispositions can be assumed to have great influence on the evolution of institutions. He stated that frugal forms of social behavior, values, and attitudes became genetically fixed during human phylogeny. His conception of instincts starts from his notion of social rules: A stratification of rules of conduct rests on genetically inherited 'instinctive' drives that are determined by 'physiological structure' (1979, p. 159f). This system of layered rules includes traditions of the past and a thin layer of deliberately added 'made' rules. Hayek also explicitly stated that humans' instinctive proclivities evolved in the course of millions of years during which the biological constitution of humans was created by natural selection. Hayek did not, however, appreciate the continuous impact of instincts on intentionality, learning, and other forms of cognition (see Hodgson, 2007).

Moreover, Hayek did not specify the exact content of humans' innate instincts and how their working keeps influencing institutional evolution. Instead, he just claimed that instincts are 'tamed and checked' by the rules of conduct and institutions of civilized society (Hayek 1960, p. 60f). In Hayek's theory of human nature, there is therefore a constant conflict between innate instincts – adapted to life in primitive societies or tribes – and the necessity to behave according to a civilized society's abstract rules. The social rules of conduct are meant to restrain the 'natural' instincts to enable the 'open society' that is based on larger and ordered groups (Hayek, 1979, p. 160ff).

Hayek also shared a strong interest in theories of cultural evolution with Veblen. They both emphasized the unique human capability for cultural adaptations through cultural learning. Moreover, they both paid extraordinary attention to group-regarding proclivities. Hayek argued that an inherent part of human nature includes the welfare of others as a condition for individual happiness (Hayek 1960, p. 78). In particular, he mentioned the instincts of solidarity and altruism, which apply to the members of one's group (Hayek 1988, p. 12 and sections 3 and 4). Furthermore, Hayek's theory of societal evolution includes the idea of a process of cultural group selection (*see* Human Cooperation: Evolution of) – an idea that also was also prominent among American institutionalism (Cordes, 2007). According to Hayek, social learning processes form a layer 'between instinct and reason' consisting of rules of conduct as basic institutions. This part of his conception of cultural evolution is akin to Veblen's notions on the role of habits. The rules of conduct, Hayek argued, can emerge as collective, unintended outcomes and then affect a society's economic prosperity, which, in turn, drives the group selection process between societies (Hayek 1979, p. 160). Existing rules of conduct have developed because the groups practicing them were more successful and displaced others (Hayek 1973). Hayek therefore claimed that beneficial institutions and ideal rules of conduct are not a result of human reason but rather spring from a cultural evolutionary process – that is, group adaptation to past experiences (1979, p. 155ff).

A naturalistically inspired theory of economic change has been brought forward by Douglass North (2005). His approach puts human cognition center stage: The driving forces of economic evolution are, he argued, human intentionality, beliefs, cognitive learning, and knowledge creation, including its subsequent dissemination via social learning. Denzau and North (1994) have provided a conceptual link between human cognition and the evolution of institutions. They have argued that to understand the constraints on human choice, one has to explore the relationship between the categories and mental models that individuals construct to make sense out of their environment, on the one hand, and the institutions that develop in historical time to order social interactions, on the other. Categories and mental models are based on individual and collective learning processes, that is, evolved capabilities of social cognition contribute to the design and the passing on of these mental frameworks.

Thermodynamics: Constraints on Economic Evolution

Another author who took up a naturalistic stance on economics is Georgescu-Roegen (1966, 1976; see also later contributions in this spirit: Gowdy (1994) or Faber and Proops (1997)). His focus was on the long-term constraints biological evolution puts on economic evolution. Georgescu-Roegen labeled his approach 'bioeconomics.' He brought forward the idea that the economic process is a continuation of the biological one that involves exosomatic instruments to obtain low entropy in addition to nature's endosomatic instruments (1966, p. 98f). The exosomatic instruments, he maintained, are a result of humans' instincts, those of workmanship and of idle curiosity – a reference to Veblen's proposed set of instincts. Georgescu-Roegen suggested concrete contents of human preferences, including biological, social, and personal wants. He showed that human beings exhibit lexicographic preferences, for example, in ethical matters, so that maximum equality of liberty takes lexicographic priority over economic well-being. These preferences violate standard assumptions concerning substitutability in choice. When referring to social instincts, Georgescu-Roegen argued that utility to an individual not only is a function of her own consumption but also considers the welfare of her community (see Gowdy and Mesner, 1998).

Moreover, starting from thermodynamics, he argued that some basic laws of physics are widely ignored by economists: The closed systems of growth models, for example, do not account for the necessary continuous inflow of low entropy from the environment to maintain economic growth. The economic process is entropic: It transforms low into high entropy (1966, p. 96f). Georgescu-Roegen insisted that economic theory be grounded in biophysical reality. By accounting for these physical laws, he introduced some

fundamental naturalistic aspects to economics: natural limitations to economic evolution. By invoking the Second Law of Thermodynamics, Georgescu-Roegen presented a constraint on all economic activity: entropic degradation. He emphasized the necessity of recognizing human agents as members of a biological species bound by biophysical processes and as social beings shaped by institutional patterns.

The Indirect Evolutionary Approach and Game Theory

Several more recent approaches claim that certain basic features of economic behavior are best explained as a result of natural selection. They account for human behaviors, such as reciprocal altruism, notions of fairness, moral behavior, and other rules of conduct that presumably have a genetic background (e.g., Hofbauer and Sigmund, 1988) (*see* Human Cooperation: Evolution of). Güth and Yaari (1992), for example, suggested their so-called indirect evolutionary approach that can be used to explain what basic wants evolutionary processes have endowed us with. They argued that wants guide behavior, behavior determines reproductive success, and the latter, as a manifestation of genetic fitness, in turn drives the evolution of wants. Individuals whose wants made them relatively more successful in acquiring and efficiently tapping scarce resources under the prevailing circumstances will have a higher number of offspring. On the condition that the offspring inherit the genetically fixed wants, those wants ultimately will be disseminated within the population, which enable an individual to cope more successfully with an environment that contains scarce resources. In this Darwinian perspective, aspects of human behavior are viewed as a product of biological evolutionary history. Furthermore, in his theory of social contract, Binmore (1998, 2001) applied a game-theoretic analysis to specify some contents of humans' preferences regarding fairness, leadership, and justice that might have evolved during phylogeny.

Another Naturalistic Endeavor: The Continuity Hypothesis

Another naturalistic endeavor in evolutionary economics starts from a 'continuity hypothesis' (Witt, 2003, p. 15f, 2004; Ruprecht, 2005; Cordes, 2006; Buenstorf and Cordes, 2008; Woersdorfer and Kaus, 2011). This departure rejects the application of abstract principles derived from Darwinism to socioeconomic evolution. At some point in time, it is argued, Darwinian evolutionary theory (*see* Darwinism) lost its power to explain human behavior. After a period of coevolution with biological evolution (*see* Gene-Culture Coevolution), cultural evolution eventually allowed forms of human behavior to emerge that entailed a strong relative reproductive success, reducing selection pressure significantly and increasing behavioral variety. Other forms of evolutionary change take place within the freedom left by the constraints of Darwinian evolution, but with different means and according to their own regularities. The foundations of cultural evolution, however, have been laid before by natural selection in the form of humans' cognitive apparatus. This establishes an ontological continuity of evolution, yet without expanding the domain of Darwinian concepts to these new evolutionary phenomena. There are other cultural, domain-specific realizations of evolutionary processes. The continuity hypothesis shows the way in which Darwinian theory can be relevant for economics: The human species is a result of natural (Darwinian) evolution; biological evolution has shaped the ground and still defines the constraints for humanmade, or cultural, evolution.

Moreover, Darwinian theory explains the origins of economic evolution in human phylogeny and fosters the understanding of the lasting influence of innate elements, dispositions, and programs on behavior. Evolutionary selection has created a set of cognitive devices that participate in generating human behavior (*see* Cognition, Evolution of). The human behavioral repertoire is the basis on which other forms of evolution rest, not least of those that are socially contingent. From this perspective, the biologically evolved foundations of learning and reasoning directly enable and affect cultural evolution. Darwinian theories of evolution therefore are suited to explain the natural origins of, for example, human learning, intentionality, and deliberative behavior, but they are ill-suited to grasp the dynamics of cultural evolution – with its own modes of transmission – that are based on these evolved cognitive capabilities.

Among other things, a naturalistic approach that includes a set of evolved cognitive dispositions can account for the motivational underpinnings of human behavior within organizations (Nicholson and White, 2006; Langlois, 2003). Several naturalistic contributions are dedicated to an analysis of the evolution of business cultures that offer a behavioral enrichment of the theory of the firm (e.g., Witt, 1998; Cordes et al., 2008; Cordes et al., 2010). Some of these papers use formal models of cultural evolution that stem from anthropology (Cavalli-Sforza and Feldman, 1981; Boyd and Richerson, 1985) (*see* Cultural Evolution: Overview; Cultural Evolution: Theory and Models). It is shown that what successful firms do better than markets – besides economizing on transaction costs – is to establish a cooperative regime among its employees that 'crowds in' desirable behavior based on humans' unique evolved group-regarding social predispositions – behaviors that go beyond the self-regarding concerns emphasized in traditional economic and biological theory (e.g., Gintis, 2007). In this context, the theory of gene-culture coevolution serves as a central tool in this analysis (Boyd and Richerson, 1980; Richerson and Boyd, 2005) (*see* Gene-Culture Coevolution). It shows how nonselfish, cooperative human traits have evolved in humans' past and how they take effect in contemporary behavior (e.g., Henrich, 2004; Gintis et al., 2005). Moreover, several of these contributions tackle the task of understanding how entrepreneurs and other business leaders can actively shape behavior within a firm as prominent role models in cultural learning and how this affects firm performance and size – a cognitive dimension that is neglected in both transaction cost economics and the theory of the firm more generally (Witt, 2007).

The continuity hypothesis has been applied to several other areas in economics. A naturalistic concept of wants inspired by behavioral theory and cognitive psychology, for example, can explain the origins and development of – innate and learned – wants people pursue. Moreover, it is suited to grasp changes in wants, that is, to account for the fact that preferences are not constant (Witt, 2001; Ruprecht, 2005; Schubert, 2012).

Consumer behavior also has an essential role to play in any transition toward a more sustainable economy: 'green' consumption patterns, for example, can spread within a population via processes of social learning that are based on, as well as influenced by, evolved cognitive capabilities (see Buenstorf and Cordes, 2008; Woersdorfer and Kaus, 2011).

Conclusion

There are different possibilities for how Darwinian concepts can be harnessed to further develop economic theory. This article has presented approaches within evolutionary economics that are based on analogy to the biological sphere, concepts that draw on a merely metaphorical use of Darwinian ideas, and some avenues that argue that the Darwinian theory of biological evolution is relevant for the social sciences in a very basic sense: The human species is, after all, a result of natural (Darwinian) evolution and all other forms of cultural evolution rest on these foundations. In general, an evolutionary approach within economics informed by insights from modern cognitive science, evolutionary biology, and anthropology can be assumed to contribute to the development of more realistic models of human behavior in economic contexts (see Hodgson, 2004: Chapter 19).

Bibliography

Asso, Pier F., Fiorito, Luca, 2004. Human nature and economic institutions: instinct psychology, behaviorism, and the development of American institutionalism. Journal of the History of Economic Thought 26 (4), 445–477.

Ayres, C.E., 1958. Veblen's theory of instincts reconsidered. In: Dowd, D.F. (Ed.), Thorstein Veblen: A Critical Reappraisal. Cornell University Press, Ithaca, pp. 25–37.

Binmore, Ken, 1998. Just Playing: Game Theory and the Social Contract II. MIT Press, Cambridge, MA.

Binmore, Ken, 2001. Natural justice and political stability. Journal of Institutional and Theoretical Economics 157, 133–151.

Boyd, Robert, Richerson, Peter J., 1980. Sociobiology, culture and economic theory. Journal of Economic Behavior and Organization 1 (2), 97–121.

Boyd, Robert, Richerson, Peter J., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago, IL.

Buenstorf, Guido, Cordes, Christian, 2008. Can sustainable consumption be learned? A model of cultural evolution. Ecological Economics 67 (4), 646–657.

Cavalli-Sforza, Luigi L., Feldman, Marcus W, 1981. Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press, Princeton, NJ.

Clark, John M., 1918. Economics and modern psychology: I. Journal of Political Economy 26 (1), 1–30.

Commons, John R., 1920. Principles of Labor Legislation. Harper&Brothers, New York.

Cordes, Christian, 2006. Darwinism in economics: from analogy to continuity. Journal of Evolutionary Economics 16 (5), 529–541.

Cordes, Christian, 2007. The role of "Instincts" in the development of corporate cultures. Journal of Economic Issues 41 (3), 747–764.

Cordes, Christian, Richerson, Peter J., McElreath, Richard, Strimling, Pontus, 2008. A naturalistic approach to the theory of the firm: the role of cooperation and cultural evolution. Journal of Economic Behavior & Organization 68 (1), 125–139.

Cordes, Christian, Richerson, Peter J., Schwesinger, Georg, 2010. How corporate cultures coevolve with the business environment: the case of firm growth crises and industry evolution. Journal of Economic Behavior & Organization 76 (3), 465–480.

Cyert, Richard M., March, James G., 1992. A Behavioral Theory of the Firm. Blackwell, Cambridge, MA.

Denzau, Arthur T., North, Douglass C., 1994. Shared mental models: ideologies and institutions. Kyklos 47 (1), 3–31.

Faber, Malte, Proops, John L.R., 1997. Evolution, Time, Production and the Environment. Springer, Berlin.

Georgescu-Roegen, Nicholas, 1966. Analytical Economics: Issues and Problems. Harvard University Press, Cambridge, MA.

Georgescu-Roegen, Nicholas, 1976. Energy and Economic Myths. Pergamon Press, New York.

Gintis, Herbert, 2007. A framework for the unification of the behavioral sciences. Behavioral and Brain Sciences 30 (1), 1–16.

Gintis, Herbert, Bowles, Samuel, Boyd, Robert T., Fehr, Ernst, 2005. Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life. MIT Press, Cambridge, MA.

Gowdy, John M., 1994. Coevolutionary Economics: The Economy, Society and the Environment. Kluwer Academic Publishers, Boston, MA.

Gowdy, John M., Mesner, Susan, 1998. The evolution of Georgescu–Roegen's bioeconomics. Review of Social Economy 56 (2), 136–156.

Güth, Werner, Yaari, Menahem E., 1992. Explaining reciprocal behavior in simple strategic games: an evolutionary approach. In: Witt, U. (Ed.), Explaining Process and Change. The University of Michigan Press, Ann Arbor, pp. 23–34.

Hayek, Friedrich A., 1960. The Constitution of Liberty. University of Chicago Press, Chicago, IL.

Hayek, Friedrich A., 1973. Law, Legislation and Liberty: Rules and Order. University of Chicago Press, Chicago, IL.

Hayek, Friedrich A., 1979. Law, Legislation and Liberty: The Political Order of a Free People. University of Chicago Press, Chicago, IL.

Hayek, Friedrich A., 1988. The Fatal Conceit. Routledge, London.

Henrich, Joseph, 2004. Cultural group selection, coevolutionary processes and large-scale cooperation. Journal of Economic Behavior and Organization 53 (1), 3–35.

Hodgson, Geoffrey M., 2002. Darwinism in economics: from analogy to ontology. Journal of Evolutionary Economics 12, 259–281.

Hodgson, Geoffrey M., 2003. The mystery of the routine: the Darwinian destiny of an evolutionary theory of economic change. Revue Economique 54 (2), 355–384.

Hodgson, Geoffrey M., 2004. The Evolution of Institutional Economics. Routledge, London.

Hodgson, Geoffrey M., 2007. Instinct and habit before reason: comparing the views of John Dewey, Friedrich Hayek and Thorstein Veblen. Advances in Austrian Economics 9, 109–143.

Hodgson, Geoffrey M., Knudsen, Thorbjørn, 2004. The firm as an interactor: firms as vehicles for habits and routines. Journal of Evolutionary Economics 14 (3), 281–307.

Hodgson, Geoffrey M., Knudsen, Thorbjørn, 2006. Why we need a generalized Darwinism, and why generalized Darwinism is not enough. Journal of Economic Behavior and Organization 61 (1), 1–19.

Hofbauer, Josef, Sigmund, Karl, 1988. The Theory of Evolution and Dynamical Systems. Cambridge University Press, Cambridge.

Knudsen, Thorbjørn, 2002. Economic selection theory. Journal of Evolutionary Economics 12, 443–470.

Knudsen, Thorbjørn, 2004. General selection theory and economic evolution: the Price equation and the replicator/interactor distinction. Journal of Economic Methodology 11 (2), 147–173.

Langlois, Richard N., 2003. Cognitive comparative advantage and the organization of work: lessons from Herbert Simon's vision of the future. Journal of Economic Psychology 24 (2), 167–187.

McDougall, William, 1910. An Introduction to Social Psychology. John W. Luce & Co, Boston, MA.

McDougall, William, 1924. Can sociology and social psychology dispense with instincts? American Journal of Sociology 29 (6), 657–673.

Metcalfe, Stanley J., 1988. The diffusion of innovation: an interpretative survey. In: Dosi, G., Freeman, C., Nelson, R.R., Silverberg, G., Soete, L. (Eds.), Technical Change and Economic Theory. Pinter Publishers, London, pp. 560–589.

Metcalfe, Stanley J., 1994. Competition, Fisher's principle and increasing returns in the selection process. Journal of Evolutionary Economics 4, 327–346.

Mitchell, Wesley C., 1910. The rationality of economic activity: I. Journal of Political Economy 18 (2), 97–113.

Mitchell, Wesley C., 1914. Human behavior and economics: a survey of recent literature. Quarterly Journal of Economics 29 (1), 1–47.

Nelson, Richard R., Winter, Sidney G., 1982. An Evolutionary Theory of Economic Change. Belknap Press of Harvard University Press, Cambridge, MA.

Nicholson, Nigel, White, Rod, 2006. Darwinism - a new paradigm for organizational behavior? Journal of Organizational Behavior 27 (2), 111–119.

North, Douglass C., 2005. Understanding the Process of Economic Change. Princeton University Press, Princeton, NJ.

Parker, Carleton H., 1920. The Casual Laborer and Other Essays. Harper, Brace & World, New York.

Penrose, Edith T., 1959. The Theory of the Growth of the Firm. Basil Blackwell, London.

Richerson, Peter J., Boyd, Robert, 2005. Not by Genes Alone: How Culture Transformed Human Evolution. The University of Chicago Press, Chicago, IL.

Ruprecht, Wilhelm, 2005. The historical development of the consumption of sweeteners – a learning approach. Journal of Evolutionary Economics 15, 247–272.

Rutherford, Malcolm, 1998. Veblen's evolutionary programme: a promise unfulfilled. Cambridge Journal of Economics 22, 463–477.

Rutherford, Malcolm, 2000. Understanding institutional economics: 1918–1929. Journal of the History of Economic Thought 22 (3), 277–308.

Rutherford, Malcolm, 2001. Institutional economics: then and now. Journal of Economic Perspectives 15 (3), 173–194.

Schubert, Christian, 2012. Is novelty always a good thing? Towards an evolutionary welfare economics. Journal of Evolutionary Economics 23 (3), 585–619.

Schumpeter, Joseph A., 1934. The Theory of Economic Development. Harvard University Press, Cambridge, MA.

Simon, Herbert A., 1955. A behavioral model of rational choice. Quarterly Journal of Economics 69 (1), 99–118.

Stoelhorst, Jan-Willem, 2008. Darwinian foundations for evolutionary economics. Journal of Economic Issues 42 (2), 415–423.

Tead, Ordway, 1918. Instincts in Industry. Houghton Mifflin Co., Boston, MA.

Veblen, Thorstein B., 1898. Why is economics not an evolutionary science? Quarterly Journal of Economics 12, 373–397.

Veblen, Thorstein B., 1914. The Instinct of Workmanship, and the State of the Industrial Arts. MacMillian, New York.

Veblen, Thorstein B., 1919. The Place of Science in Modern Civilization and Other Essays. Huebsch, New York.

Witt, Ulrich, 1998. Imagination and leadership - the neglected dimension of an evolutionary theory of the firm. Journal of Economic Behavior and Organization 35 (2), 161–177.

Witt, Ulrich, 2001. Learning to consume – a theory of wants and the growth of demand. Journal of Evolutionary Economics 11, 23–36.

Witt, Ulrich, 2003. The Evolving Economy - Essays on the Evolutionary Approach to Economics. Edward Elgar, Cheltenham.

Witt, Ulrich, 2004. On the proper interpretation of 'evolution' in economics and its implications for production theory. Journal of Economic Methodology 11 (2), 125–146.

Witt, Ulrich, 2007. Firms as realizations of entrepreneurial visions. Journal of Management Studies 44 (7), 1125–1140.

Woersdorfer, Julia S., Kaus, Wolfhard, 2011. Will nonowners follow pioneer consumers in the adoption of solar thermal systems? Empirical evidence for northwestern Germany. Ecological Economics 70 (12), 2282–2291.

Evolutionary Epistemology

David L Hull,[†] Northwestern University, Evanston, IL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 7, pp. 5012–5017, © 2001, Elsevier Ltd.

Abstract

Traditionally, epistemology concerns justifying beliefs. What right do we have to think that our beliefs are true? Evolutionary epistemology is a form of naturalized epistemology. In naturalized epistemology, the justification provided turns on an appeal to something about the natural world, in this case the evolutionary process. Organisms that do not react appropriately to their environments are not likely to survive to reproduce. Evolutionary epistemology began as both epistemology and evolutionary. Evolutionary epistemology was 'epistemology' in the sense that it was designed to justify certain basic beliefs that people hold. Evolutionary epistemology was 'evolutionary' in the sense that it was based on evolutionary theory as a biological theory. The reason that our knowledge of the world of our experience is reasonably good is that our sense organs have been honed by natural selection, and the results of this honing are programmed into our genes. The early versions of evolutionary epistemology were gene based; that is, the relevant knowledge was programmed into our genes and thus passed from generation to generation. Things like behavior were viewed as phenotypic characteristics.

Traditionally, epistemology concerns justifying beliefs. What right do we have to think that our beliefs are true? Evolutionary epistemology is a form of naturalized epistemology. In naturalized epistemology, the justification provided turns on an appeal to something about the natural world, in this case the evolutionary process. Organisms that do not react appropriately to their environments are not likely to survive to reproduce. In some organisms this reaction can appropriately be termed 'knowledge.' Those organisms that are better at learning about their environments (which includes other organisms) are more likely to survive to reproduce than their less adept congeners. Thus, knowledge of those features of the environment that are relevant to survival and reproduction is to some extent justified. *Homo sapiens* as a species depends for its survival on playing the knowledge game better and more extensively than any other species. Evolutionary epistemologists propose to justify at least some of the beliefs that human beings hold by reference to survival value (for the most seminal paper in evolutionary epistemology, see Campbell, 1974).

Gene-Based Evolutionary Theory

In order to understand evolutionary epistemology, one must understand evolution, in particular the distinction between genotypes and phenotypes (*see* Genotype and Phenotype). The genotype of an organism consists in the sequence of bases in its DNA, or rarely, in its RNA. Everything else is part of its phenotype, from its blood type to its behavior. Variations in the genotype of an organism lead, through embryological development, to different phenotypic traits. These genetic variations are caused, totally caused. Among these causes are unequal crossover, increased temperature, and the presence of numerous mutagens. However, to count as Darwinian evolution, no correlation must exist between the variations an organism might need and those it might get; for example, it cannot be 'Lamarckian.' For Lamarckian evolution to occur, some feature of an organism's environment must change the phenotype of that organism in such a way that it is better adapted to the environmental feature that produced this change in the first place. This phenotypic change must then somehow be transmitted to the organism's genes to be passed on to subsequent generations via the genetic material. A mother dog giving fleas to her puppies does not count as Lamarckian inheritance.

Some of the traits exhibited by an organism make it better able to survive and reproduce. As a result, the genes that code for these traits are passed on differentially from generation to generation. However, as organisms become better-adapted to their environments, their environments frequently move out from under them. A few species succeed in surviving through the millennia largely unchanged, but most survive only by evolving. One feature of gene-based selection in biological evolution is that it can occur quite rapidly, as in viruses and bacteria, or quite slowly, as in most larger animals. However, in every case of Darwinian evolution, the relevant traits are inherited literally via the genetic material. Initially, the 'units' used in evolutionary biology were traits because evolutionary biologists had no access to the genetic material. As this access has increased, traits have been replaced by genes as units of inheritance. Unfortunately, the genetic material can be divided up in numerous ways. As a result, biologists concentrate on the 'information' contained in the genetic material.

Advocates of evolutionary epistemology in its most literal sense argue that the preceding story equally applies to some behavioral traits and predispositions. These traits and dispositions are as much phenotypic traits as are hair texture and skin color. Knowledge of the world in which organisms live gets programmed into an organism's genes. For example, fear of heights and the sucking reflex are in some sense programmed into the genes of human beings. The evolution of behaviors and behavioral predispositions has all the usual characteristics of gene-based biological evolution. As traits, behaviors can be adaptive. The inheritance involved is genetic, strictly Darwinian, not Lamarckian, and it can occur at a variety of rates from very slow to very fast.

[†] Deceased

The majority of epistemologists are not very impressed with evolutionary epistemology, in part because it is naturalistic, and only normative epistemologies are worth the title of epistemology. But a second weakness is that the only beliefs that evolutionary epistemology in this strict sense can even purport to justify are beliefs about entities and processes that the organism can perceive. Evolutionary epistemology might lend some justification for how human beings distinguish species of large organisms, but it cannot justify any beliefs that we have about the millions of tiny microorganisms that fill every nook and cranny of the world, let alone quantum indeterminacy.

Individual Learning

In 'eusocial organisms,' such as bees and termites, nearly all of the behaviors exhibited by these creatures evolved in a strictly Darwinian fashion. Relatively little learning takes place in the span of a single generation. But many organisms can learn from interacting with their environments. In a second form of evolutionary epistemology, knowledge does not get programmed into our 'genes,' but into our 'brains.' Gene-based selection in biological evolution can handle features of the environment that change relatively slowly, e.g., changes in temperature resulting from continental drift, but some changes are so rapid that organisms with longer generation times cannot adapt to them. One solution to this problem is the evolution of the 'central nervous system.' Organisms with highly developed central nervous systems can learn about their environments and react to changes in them with greater speed. This information does not get programmed into an organism's genes but into its brain. 'Operant learning' is the simplest form of individual learning. Each kind of organism has a range of behaviors that it tends to emit spontaneously. Feedback from its environment increases or decreases the frequency of these behaviors. Sequential environmental responses shape the behavioral repertoires of these organisms.

The preconditions for operant learning are certainly programmed into the genes of those organisms that are capable of individual learning, and arose through gene-based natural selection. However, the specific learned repertoires cannot be passed on to successive generations. Each organism must learn anew. For better or for worse, calculus is not programmed into the genes of any human being. Each generation must learn it from scratch. If over many generations the same learning regimen is repeated and produces some evolutionary advantage, what was once a learned behavior can to some extent become programmed into the genes. Within a single life-history of an organism, the mechanisms that produce increased knowledge can be viewed as a 'selection process.' However, in this selection process, 'brain states' take the place of genes. Behavioral units can vary from something as brief and discrete as pressing a lever to something as complex and protracted as mowing a lawn. We have yet to discover the size and shape of the brain states that are responsible for the retention of information.

A second solution to the problem posed by differences in reaction times is the 'immune system.' In order to keep up with the rapid rate of reproduction of such antigens as viruses, immune systems evolved in many larger organisms. Within the life of a single organism, selection in the immune system occurs at the cellular level. Antibodies are produced in huge numbers, each with its own reaction site. If any of these antibodies happens upon an antigen that contains a protein that matches its reaction site, this antibody proliferates and, in most cases, eliminates the antigen. Once again, the 'information' garnered through successive invasions by antigens cannot be passed on to later generations. After each invasion, a few memory cells lie in wait for any subsequent invasions. That is all. Although some workers in the early years of modern immunology thought that the process might contain some elements of Lamarckian inheritance, it does not.

Social Learning

Classic epistemology dealt with individual, isolated cognizers. How much can we learn about our environments in total isolation from all other people? How certain could Robinson Crusoe be, once marooned on his island, of any new knowledge that he obtained? After all, he had no one with whom to check his discoveries. But even the Robinson Crusoe example is not pure enough, because Crusoe brought with him language and a fund of socially acquired knowledge. What about someone who began life totally isolated from all other human beings? In the first place, such an isolated baby would rapidly die, but how about the science fiction case? How about a baby adopted by wolves or some other unlikely creature? How much could it know? How justified would its knowledge be? The answer to both questions is very little. Social organisms can learn more than solitary organisms and be more secure in this knowledge because of culture (*see* Cultural Evolution: Theory and Models).

Most discussions of evolutionary epistemology do not deal with strict gene-based biological evolution or individual learning such as operant behavior, but 'social' learning. In social learning, knowledge acquired by one organism in its lifetime can be passed on to other organisms. This knowledge is transmitted from organism to organism, not through genes, not through the passing on of any brain tissue, but through 'imitation' and 'communication.' As a result, social learning need not proceed vertically, as defined by the direction that genes take in passing from parent to offspring, but horizontally (see Cavalli-Sforza and Feldman, 1981). Education in the human species is an example of social learning. Through repeated efforts, teachers succeed in getting at least some of their pupils to be able to solve problems in calculus. Although the ability to learn in this way surely has a genetic component, social learning does not involve transmitting information genetically. As a result, it cannot count literally as Darwinian or Lamarckian. Of course, nothing keeps genes and culture from coevolving. In fact 'gene-culture coevolution' is likely to be of primary importance in human evolution (see Durham, 1991; for additional reading, see Boyd and Richerson, 1985; Cavalli-Sforza and Feldman, 1981; Eldredge and Grene, 1992; and Lumsden and Wilson, 1981).

Memetics

At this stage, advocates of evolutionary epistemology as social learning make a significant move. They cease treating bits of

knowledge, behavior, and the like as analogous to 'characters' but as analogous to 'genes.' Just as genes get passed on differentially in gene-based biological evolution, 'memes' get transmitted differentially in 'memetic' evolution (Blackmore, 1999; Dawkins, 1976; and Plotkin, 1994). Just replacing genes with memes may not sound like a very profound change, but the implications are significant. Once again, memetic evolution is not in the least Lamarckian. Literally, memetic evolution is not Lamarckian because nothing is being passed on 'genetically.' Metaphorically, it is not Lamarckian because what is being transmitted is not analogous to 'characters' but to 'genes.' It is the inheritance of acquired memes, not genes or even characters. The reason for emphasizing that memetic evolution is in no cogent sense Lamarckian, is that critics of memetic evolution think that they can dismiss this perspective simply by terming it 'Lamarckian.'

Critics of memetics provide a long list of objections to the memetic form of evolutionary epistemology. First, supposedly memetic evolution occurs so much more quickly than does gene-based biological evolution. But the same variations in speed of transmission can be found in both. Certainly viruses can spread as quickly as any meme. Second, critics complain that the analogy between genes and memes is imperfect. Genes are supposedly absolutely distinct, all of the same size and kind, and do not interact with each other. Supposedly memes exhibit none of these characteristics. Hence, the meme-gene analogy breaks down. But one need not know very much about genetics to know that genes possess none of the characteristics attributed to them by critics of memetics. Some genes are reasonably discrete; many are not, and much of the genetic material cannot be subdivided meaningfully into discrete genes. Genes vary tremendously in size, and numerous different sorts of genes exist (e.g., regulator genes). Also genes interact with each other in extremely complex ways in development. In addition, many elements of knowledge communication possess all of the preceding characteristics reasonably well developed; for example, e-mail addresses.

The chief problem with treating conceptual entities as memes is the difficulty of distinguishing *phenotypic characteristics* in this metaphorical usage. Behaviors count as characteristics in literal evolutionary epistemology. They are analogous to 'genes' in the metaphorical usage. Then what counts as phenotypic characters in the metaphorical usage? Memes get transferred from brains to air waves, to computer chips, to computer screens, to brains, etc. All of this counts as replication because information is being transmitted with relatively little loss. However, every once in a while, 'replication' is interrupted by 'translation.' Knowledge is connected causally to what it is knowledge of: you believe, for example, that barking dogs do not bite, so as you approach a barking dog, you put this belief to the test. The result will modify your belief system. The dog itself is not a meme.

As in other cases of selection, the notion of information plays a central role. One important feature of memetic replication is transmission with relatively little loss of information. One important feature of translation into the phenotype is significant loss of information. Instructions on how to make a Stradivarius violin can be transmitted from agent to agent. Antonio Stradivari might have written down these instructions, and even allowed an apprentice to watch him make a violin. If he had, this apprentice might have been able to make violins of the same caliber as did his master. He might have passed down this combination of knowledge and skill to the present. However, that is not what happened. All we have are Stradivari's violins. As repeated failures have shown, making violins based just on what we can learn from the finished product (reverse engineering) is extremely difficult, if not impossible. Future work in memetics is needed to present adequate analyses of 'information' and the 'genotype-phenotype' distinction.

A General Analysis of Selection

Early versions of evolutionary epistemology involved reasoning analogously from gene-based selection in biological evolution to knowledge acquisition. The analogical character of this activity has been the source of endless criticism. Evolutionary epistemology is 'merely' an analogy. In these arguments, the peculiarities of gene-based selection have played an inordinately important role in biological evolution. Any departure from gene-based selection in biological evolution, no matter how trivial, is viewed as a lethal disanalogy. However, an alternative view has been suggested periodically. Instead of reasoning analogously from one sort of selection process to another, present a 'general analysis of selection' that is equally applicable to all sorts of selection processes, whether gene-based selection in biological evolution, the reaction of the immune system to antigens, individual learning, social learning, memetics, etc. (Toulmin, 1972). Although the father of evolutionary epistemology, Donald Campbell (1974) was one of the strongest proponents of justifying knowledge by reference to gene-based selection in biological evolution. In his later writings he adopted this second program. Evolutionary epistemology is no longer an analogy, but it is a special case of selection. Just as the processes that result in functional organization have been isolated and analyzed as a special sort of process, selection processes warrant an equally careful treatment.

Starting with Lewontin's (1970) classic statement, successive attempts have been made to specify the essence of selection processes. According to Lewontin, the three necessary elements for selection to occur are phenotypic variation (different individuals in a population have different morphologies, physiologies, and behaviors), differential fitness (different phenotypes have different rates of survival and reproduction in different environments), and heritable fitness (there is a correlation between parents and offspring in the contributions of each to future generations). According to Hull's (1980) more general characterization, the selection process is made up of three subprocesses related in very specific ways: (1) replication of information in successive generations with only minimal modification, (2) environmental interaction that causes the preceding replication to be differential, and (3) some source of variation. The goal is to show that gene-based selection in biological evolution, the reaction of the immune system to antigens, individual learning, social learning, and other processes as well, all exhibit the basic structure of selection processes. However, such general analyses of selection do not provide any epistemological warrant. If evolutionary

epistemology is to remain 'epistemology,' then some other source of epistemological warrant must be provided (Hull, 2001).

Conclusion

Evolutionary epistemology has changed significantly from its slim beginnings. Initially, it dealt only with knowledge that is literally gene-based, e.g., a tendency to group organisms into discrete species. Evolutionary epistemology in this strict sense justified certain beliefs. However, this warrant was neither very strong nor very extensive. Evolutionary epistemology in this sense is limited only to those environmental interactions that are in the range of an organism's sense organs and that matter with respect to an organism's survival and reproduction. Evolutionary epistemology in this strict sense can provide no justification whatsoever for a belief in quantum indeterminacy.

However, the sort of evolutionary epistemology that has received the greatest attention is metaphorical in character. Knowledge is in no sense programmed into the genes to be transmitted to future generations. Instead the transmission is social. The entities being transmitted are in a broad sense memes. Numerous objections have been raised to this metaphorical evolutionary epistemology. First, so it is alleged, memes are not all that similar to genes. However, these objections stem from an unreal view of both genes and memes. The real problem is identifying something that can count as the phenotype in memetic evolution. In the absence of such a distinction, memetic evolution cannot be characterized as either Darwinian or Lamarckian. Once again, a mother dog giving her puppies fleas is not Lamarckian inheritance in either a literal or a metaphorical sense.

Whatever warrant metaphorical evolutionary epistemology provides arises from its analogical connections to literal evolutionary epistemology. Because many species of organism must be able to distinguish members of their own species to reproduce and survive, other beliefs more distantly related to reproduction and survival gain some analogical warrant. One problem with both sorts of evolutionary epistemology is that they do not distinguish between 'success' and 'truth.' Many false beliefs have been proven to be very successful with respect to survival and reproduction, e.g., the belief that God is on the side of one's own group in times of war. For this and many other reasons, a majority of professional epistemologists reject evolutionary epistemology as not being nearly good enough. Of course, professional epistemologists reach the same conclusion with respect to all other epistemological justifications as well, save their own.

A third sort of evolutionary epistemology has gained increased popularity through the years. Instead of reasoning analogically from gene-based to memebased evolution, a general analysis of selection is produced that equally applies to all sorts of selection processes from the reaction of the immune system to antigens to social learning. As promising as this research program might well be, it has one serious drawback. It supplies no epistemological warrant, at least not in any of the traditional senses. Just because a belief arose through a selection process, it does not follow that this belief is true. It all depends on the character of the selection process. For example, a belief that survives the sort of serious testing that goes on in science is likely to be true, while biblical exegesis is not likely to increase its warrant.

However, during this same period, what counts as 'epistemology' has expanded greatly. Epistemologists are no longer limited just to justifying knowledge. They are engaged in all sorts of other activities as well. For our purposes, the issue is whether or not providing a general analysis of selection can justify certain sorts of knowledge and, if not, whether it counts as 'epistemology' in any of its more recent senses.

See also: Adaptation, Fitness, and Evolution; Cognition, Evolution of; Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Darwin, Charles Robert (1809–82); Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Evolutionary Theory, Structure of.

Bibliography

Blackmore, S., 1999. The Meme Machine. Oxford University Press, Oxford, UK.
Boyd, R., Richerson, P.J., 1985. Culture and the Evolutionary Process. University of Chicago Press, chicago.
Campbell, D., 1974. Evolutionary epistemology. In: Schilpp, P.A. (Ed.), The Philosophy of Karl R. Popper. Open Court Publishers, La Salle, IL.
Cavalli-Sforz, L.L., Feldman, M.W., 1981. Cultural Transmission and Evolution: A Quantitative Approach. Princeton University, Princeton, NJ.
Dawkins, R., 1976. The Selfish Gene. Oxford University Press, Oxford, UK.
Durham, W.H., 1991. Coevolution: Genes, Culture, and Human Diversity. Stanford University Press, Stanford, CA.
Eldredge, N., Grene, M., 1992. Interactions: The Biological Context of Social Systems. Columbia University Press, New York.
Hull, D.L., 1980. Individuality and selection. Annual Review of Ecology and Systematics 11, 311–332.
Hull, D.L., 2001. In search of epistemological warrant. In: Campbell, D.T. (Ed.), Selection Theory and Social Construction. State University of New York Press, Stony Brook, NY, pp. 155–167.
Lewontin, R.C., 1970. The units of selection. Annual Review of Ecology and Systematics 1, 1–18.
Lumsden, C.J., Wilson, E.O., 1981. Genes, Mind, and Culture: The Coevolutionary Process. Harvard University Press, Cambridge, MA.
Plotkin, H., 1994. Darwin Machines and the Nature of Knowledge. Harvard University Press, Cambridge, MA.
Toulmin, S., 1972. Human Understanding. Princeton University Press, Princeton, NJ.

Evolutionary Group Dynamics

Tatsuya Kameda, Hokkaido University, Hokkaido, Japan
Mark Van Vugt, VU University Amsterdam, Amsterdam, The Netherlands
Robert Scott Tindale, Loyola University Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Evolutionary psychology adds many insights to the literature on group dynamics and small-group processes. First, groups are a fundamental aspect of human evolution, suggesting that humans have evolved a range of adaptations to deal with specific threats and opportunities afforded by living in groups. Second, an evolutionary perspective integrates knowledge from numerous behavioral science disciplines such as psychology, evolutionary biology, primatology, biological anthropology, social neuroscience, and behavioral economics that are all concerned with group dynamics. Third, an evolutionary analysis produces many novel hypotheses about different aspects of our group psychology. We show the generativity of an evolutionary psychological approach through discussing examples of research applying evolutionary thinking to understanding key adaptive group tasks such as coordination, social exchange, status, group cohesion, collective decision making, and intergroup relations.

Introduction

'Group dynamics' refers to interplays of people's social cognitions and behaviors operating within a group or between groups. The term was coined by a social psychologist, Kurt Lewin, who established the Group Dynamics Research Center in 1945 at the Massachusetts Institute of Technology. Yet, its scope has been interdisciplinary from the outset, encompassing psychology, sociology, and anthropology, among other disciplines. Recently, an increasing amount of research in this area has applied evolutionary principles to understand group dynamics (Kameda and Tindale, 2006; Van Vugt and Kameda, 2012). Applying an evolutionary perspective, group living can be seen as an adaptive strategy that increased the survival and reproductive success of ancestral humans.

One of the impetuses that promoted the shift was the social brain hypothesis (Dunbar 1993). Comparative studies with other primates and mammals have found a positive correlation between the size of the neocortex relative to the rest of the brain and the average group size of the species. Humans are ranked top, with a large neocortex and associated large average group size, which is extrapolated to be around 150 individuals from the brain data. Given the high metabolic costs of maintaining a large brain, the social brain hypothesis posits that humans have evolved large brains in order to make the most of living in large, complex social groups. In other words, many of our core cognitive and emotional faculties are likely to have been tuned to solve recurrent adaptive problems that ancestral humans have encountered in group life. The primary aim of an evolutionary approach to group dynamics is thus to identify and analyze specific adaptive group problems and the psychological architectures that have evolved to solve those problems.

Key Adaptive Problems in Group Life

Van Vugt and Kameda (2012) have proposed six key adaptive problems that ancestral humans were likely to have encountered recurrently in group life:

1. Coordinating members' activities (group coordination);
2. Exchanging resources (social exchange);
3. Negotiating group hierarchies (status);
4. Keeping groups together (group cohesion);
5. Making collective decisions (group decision making); and
6. Interacting with members of outgroups (intergroup relations).

This list is meant to be neither exhaustive nor mutually exclusive. However, it serves as a reasonable starting point for building an evolutionary framework for analyzing group processes. The list also corresponds closely to the core themes of group dynamics identified in textbooks such as Forsyth's *Group Dynamics* (2010).

Although none of these adaptive group challenges has been fully analyzed using an evolutionary framework, various research programs have contributed to developing such analyses by providing evidence for evolved psychological mechanisms that address a particular group challenge. The promise of an evolutionary group dynamics lies in the generativity and productivity of this approach in formulating novel hypotheses and providing supportive empirical evidence. This article provides an illustrative set of findings of evolutionary-inspired research programs for each of these core group challenges.

Group Coordination

How to coordinate members' cognitive and physical resources is a key to effective group performance. As a nomadic, group-living species, early humans would have had to solve problems associated with coordinating their activities with other individuals in their groups. For instance, when moving they would have had to decide where to move to and when and for how long to stay there. To solve this would have required mechanisms for identifying that situations require coordination, developing rules for how to achieve coordination (e.g., turn taking and leadership), and then carrying these rules out.

Leadership

There are multiple indications that leadership might be an adaptive solution to coordinate members' actions (Van Vugt, 2006). Whenever organisms must do things together to achieve their reproductive goals, they face a critical coordination problem, that is, how do they decide what to do when? A simple thought experiment illustrates that leadership – where one individual takes the initiative and others follow – is a powerful solution to such coordination problems. Suppose that two individuals who are thirsty must find a water hole to drink. They must stay together as a form of protection, but how do they decide on which water hole to go to? In such cases, it is adaptive for one individual to take the initiative to go to a particular hole, which leaves the other no option but to follow. Coordinating on the same water hole is the equilibrium solution to this game. An implication would be that leadership–followership interaction emerges spontaneously and does not require much cognitive computation. Unsurprisingly, then, the emergence of leadership has been documented across many different animal species that face functionally important coordination problems, including teaching in ants, choosing foraging sites in honeybees, movement in stickleback guppies, and peacekeeping in nonhuman primates (King et al., 2009).

Among humans, similar kinds of coordination problems also result, predictably, in the emergence of leader–follower relations. This occurs quickly and spontaneously across many different situations and cultures, suggesting evidence for adaptation (Brown, 1991). Of course, the exact system of leadership varies across different situations, and there is evidence for both highly democratic leadership structures and highly despotic leadership structures in humans. They likely represent different adaptive solutions to various local group conditions – for instance, dictatorial leadership might have emerged initially in response to an immediate crisis that required quick and decisive action (Van Vugt et al., 2008; Van Vugt, 2009).

Transactive Memory

Another coordination device that may have evolved to support group coordination is a transactive memory system. Wegner (1987) argued that groups can store and process more information than individuals because they can share the responsibility of knowledge storage. However, in order to retrieve the information efficiently, a shared knowledge system must exist identifying who in the group knows what – a transactive memory system. Once such a system exists, each individual member needs to store only information for which they are responsible, easing the memory load on each member but increasing the total amount of information available to the group. Division of cognitive labor is common in social insects, and one famous example is the waggle dance of the honeybee, a group decision-making device for selecting foraging sites (Seeley, 2010). Research on humans also suggests that members working in the same group often specialize in different areas, and group members are very quick at recognizing and using each other's expertise (Littlepage et al., 2008). People who are expert in a particular domain not only have more information on a given topic but also are the ones who are responsible for storing new information in their area of expertise. A set of experiments showed that teams performed better on a group task to the extent that the team members divided their cognitive tasks better. Furthermore, members of teams with better transactive memory systems also trusted each other's expertise more. There is also evidence that such systems begin to form quite quickly through normal group interaction (Moreland et al., 1996).

Structural and Social Focal Points

A game-theoretic concept discussed by Schelling (1960) – focal point – seems to play an important role in group coordination. Early work had shown that certain game solutions (e.g., equal outcomes – see Section Social Exchange) tend to receive greater support than would be expected by rational game-playing assumptions (Komorita and Chertkoff, 1973). Schelling argued that such outcome distributions 'stood out' and were salient because of their normative or focal nature. It appears that humans may have learned to use structural characteristics of their environments to help to coordinate action (e.g., meeting in the center of the village or at the water's edge). More recent research has shown that 'social focal points' are also useful in coordinated action (Abele and Stasser, 2008). Thus, normative preferences among the group members or exhibited by high-status members (leaders) can be used to guide action in ambiguous situations. Abele and Stasser (2008) showed that people were quite good at coordinating actions around social focal points, and successful coordination led to greater liking among group members. More recent research (Abele and Chartier, 2012) has shown that groups are substantially better than individuals at locating and using social focal points. Groups' superiority seems to stem from both majority processes as well as the group's ability to recognize appropriate focal points even when proposed by a single group member (see related discussions under Section Group Decision Making).

Following Group Norms

Group norms can also be helpful for coordinating behavior. Recent physiological evidence has supported the idea of evolved mechanisms to ensure greater conformity to ingroup norms (Stallen et al., 2012). Participants in their study were administered either oxytocin or a placebo and were asked to rate stimuli in terms of attractiveness. They were also shown normative ratings of each symbol by both ingroup and outgroup members. When normative ratings differed for the ingroup and outgroup, participants' ratings were more similar to the ingroup ratings but only when administered oxytocin. Thus, this hormone, which has also been related to ingroup trust and cooperation (Kemp and Guestella, 2011), also seems to help coordination by enhancing normative behavior.

Social Exchange

Social exchange – cooperation for mutual benefit – is a pervasive and cross-culturally universal feature of human group life. Exchanging vital resources with others is fundamental for any gregarious species, yet humans are unique in being able to establish large-scale cooperation with genetically unrelated individuals. Social norms related to cooperation are robust across various cultures ranging from hunter–gatherer, horticultural, tribal, and agricultural to highly industrialized societies (Henrich et al., 2004). Such norms specify how individuals should behave in a group situation where

incentives for free riding exist, including when to cooperate, how to divide group outcomes, and how to punish uncooperative group members.

Cooperation Norms

Human collective action is often governed by a norm of conditional cooperation (Fehr and Fischbacher, 2004) (*see* Human Cooperation, Evolution of). This norm dictates that an individual should cooperate if other group members cooperate, whereas he or she is allowed to not cooperate if the others defect. Fischbacher et al. (2001) examined participants' willingness to contribute in a one-shot public-goods experiment as a function of the average contribution of the other group members. Despite the economic incentives to free-ride (contribute nothing), 50% of the participants matched their contributions with the average contribution of other members. Furthermore, when participants simply observed the interaction of two players in a Prisoner's Dilemma game, they spent their own endowment to punish players who defected unilaterally but *not* players who defected bilaterally (Fehr and Fischbacher, 2004). This pattern indicates that the norm of conditional cooperation ('unilateral defection is not acceptable') is enforced by neutral observers ('third parties'). Such selective sanctioning includes not only physical punishment but also social exclusion, collectively denying the norm violator access to interpersonal relations in a group.

An evolutionary perspective suggests a novel hypothesis about the psychological mechanisms underlying enforcement of cooperation norms. Given that exchanges of valuable resources occur mainly within ingroups, violation of cooperation norms should be more serious to one's survival if committed by another ingroup member than if committed by an outgroup member. If so, then noncooperative behavior by ingroup members should be punished more severely than noncooperative behavior by outgroup members. Using the third-party punishment paradigm, Shinada et al. (2004) confirmed this prediction.

Distribution Norms

'Distribution norms' refer to a set of shared beliefs that prescribes how resources should be distributed among group members. Evidence suggests that motives for egalitarian sharing often operate strongly in resource distribution (Fehr and Schmidt, 1999; Kameda et al., 2010). For example, results from numerous one-shot Ultimatum Game experiments indicate that modal offers by a proposer for a responder's share are around 40–50% and that offers in this range are rarely rejected (Camerer, 2003). Although there are some cultural differences, extremely small offers (1–10%) are rarely seen in ultimatum-bargaining experiments conducted in a wide range of societies, including nonindustrialized as well as industrialized ones (Henrich et al., 2004). Violators of the egalitarian distribution norm are also punished. Henrich et al. (2006) conducted a third-party punishment experiment, where participants could spend their own endowments to punish an unfair proposer in the Ultimatum Game. The experimental results from 15 diverse populations showed that costly punishment by a third party was common, although the magnitude of punishment varied substantially across cultures.

Ethnographic studies show that egalitarian sharing of hunted meat constitutes a core feature of hunter–gatherer life. Compared to collected resources (e.g., cassava), hunted meat is often the target of communal sharing. Kaplan and Hill (1985) argued that the sharing system, once established, functions as a collective risk reduction device. While acquisition of collected resources is relatively stable and dependable, acquisition of meat is a highly variable, uncertain event. By including many individuals in the sharing group, the variance in meat supply decreases exponentially (Gurven, 2004; but see also Kameda et al., 2003a, 2005, for the potential free-rider problem in meat sharing). This may imply that our minds are built to be highly sensitive to cues of uncertainty in resource acquisition. Kameda et al. (2002) showed that such uncertainty cues promoted people's willingness to share with others beyond their personal distributive ideologies.

Status

Competition for status, especially among males, is a robust phenomenon across cultures. Even in modern ad hoc groups such as juries, status competition is often observed, where some jurors 'show off' their toughness for the purpose of establishing their prestige in the group while sacrificing factual discussions about the case (Hastie et al., 1983). Evolutionarily, with an increase in the size and complexity of social groups, competition for scarce resources including food, water, and sexual mates would have intensified, paving the way for the emergence of status hierarchies (Boehm, 1999; Dunbar, 2004). To climb the group hierarchy or maintain one's current status in the group, one would need to closely monitor one's relative standing in the hierarchy and to behave strategically in the group.

Signaling and Status Emotions

Along with other group-living animals, humans display various cues to signal their relative status to others. For example, nonverbal cues such as a firm handshake, a poised posture, gaze maintenance, or initiatives in conversations are among characteristics of high-status individuals when they interact with others. Some of these cues are culture specific, but other cues are easily recognizable by even outsiders who do not belong to the culture (Ridgeway et al., 1985), which may suggest their evolutionary origins.

Besides sensitivities to subtle status signals displayed during interaction, our own emotional experiences are also closely linked to changes in relative status in a group. For example, when people experience a status gain (e.g., winning an award), they tend to feel pride. Recent research has shown that, like the 'basic' emotions, pride is associated with a distinct, universally recognized, nonverbal expression, which is spontaneously displayed during pride experiences (Tracy and Robins, 2004). It has also been argued that self-esteem may be an evolved internal gauge ('sociometer') that monitors people's standing in a group and motivates actions when people feel their status is being threatened (Leary, 1999).

Competitive Altruism

When individuals can select social exchange partners freely in a group, status competition can arise for establishing a good reputation as a generous person. This phenomenon is called

'competitive altruism' (Roberts, 1998), as people compete to be chosen as exchange or coalition partners through their generosity. The anthropological literature documents various examples of excessive public displays of altruism and generosity. Experimental evidence also shows that generous individuals receive more status than nongenerous individuals and are preferred as group leaders (Hardy and Van Vugt, 2006).

Competitive altruism requires psychological mechanisms to monitor and enhance one's relative standing about generosity in a group. People are highly sensitive to social cues that suggest their reputation is under scrutiny. For example, experimental evidence shows that people behave more generously when they think they are being watched by others (Bateson et al., 2006; Haley and Fessler, 2005). Men also behave more generously to a stranger when they are watched by a potential female partner (Iredale et al., 2008).

Group Cohesion

Group cohesion has been defined as "the resultant of all forces acting on the members of a group to remain in the group" (Festinger, 1950, p. 274). In light of the importance of staying together as a unit in a hostile environment, our ancestors had to evolve mechanisms to preserve social cohesion. Furthermore, as human social networks increased in size over the course of human evolution, we would expect these bonding mechanisms to have become increasingly sophisticated. To maintain group cohesion would require specialized mechanisms to recognize oneself and others as belonging to the same group as well as mechanisms to feel emotionally connected with others in increasingly large groups.

Social Identity

Thinking of people who are not necessarily around all the time as belonging to the same group as you requires the capacity for symbolic thought whereby language or rituals become markers of shared membership. A symbolic social identity allowed our ancestors to connect with a large network of individuals who were spread around a particular area, and this may have been quite helpful in sharing resources as well as in competing with other groups. Human social identity is highly group based, and people spontaneously make us-versus-them categorizations (Tajfel and Turner, 1979). Preserving group cohesion also requires a sense of group loyalty whereby individuals are prepared to forgo attractive alternatives in favor of staying with their current group. A sense of loyalty is deeply ingrained in human psychology.

Religion, Music, and Dance

Humans have many specialized behavioral mechanisms for fostering group cohesion, which may have deep evolutionary roots. Religion, for instance, is an effective method to promote cohesion between strangers and mobilize them for joint action on behalf of a group (Atran, 2002). Similarly, dance and music may have ancient roots. According to some evolutionary anthropologists, dance and music may have evolved as adaptations for connecting large networks of genetically unrelated strangers (Dunbar, 2004).

Group Decision Making

Along with our highly sophisticated faculty for language, the ubiquity of group decision making across many societies may seem to suggest that group decisions are uniquely human. However, recent research on animal behavior suggests that this is not the case. Group decision making seems to be common in the animal kingdom as well, including in social insects (e.g., ants, termites, and honeybees), fish, and some mammals (Conradt and List, 2009; Seeley, 2010). Although 'animal group decision making' may look more like automated self-organization than deliberate coordination, the social aggregation processes (e.g., honeybees' waggle dances to recruit more fellow searchers) are in fact highly coordinated and often lead to efficient group-level outcomes. Evidently, some well-structured social coordination mechanisms that yield collective wisdom are an outcome of natural selection (Kameda et al., 2012).

Despotism versus Democracy

What is the key evolved decisional structure that enables collective wisdom in honeybees and some other animals? Conradt and Roper (2003) compared two contrasting structures, despotism and democracy, in the animal kingdom. Using a stochastic model, they showed that democratic decisions usually yield better fitness outcomes to group members than despotism – even when the despot is the most experienced group member, it pays other members to accept the despot's decision only when group size is small and the difference between their own and the despot's information is large. These findings may be extendable to human group decision making as well. Most naturally occurring environments for humans as well as other animals are characterized by large statistical uncertainties. Given that no single individual (despot) can handle these uncertainties alone even if he or she were highly experienced, the more viable and reliable decisional structure in the long run is to use groups as an aggregation device. By aggregating members' opinions, random errors in individual perceptions under uncertainty are canceled collectively, as implied by the law of large numbers in statistics (Surowiecki, 2004).

A recent study compared several decision rules, which differed in computational loads, in terms of their net efficiencies under uncertainty (Hastie and Kameda, 2005). These included the best member ('despotism') rule and the majority ('democracy') rule. Results from both computer simulations and laboratory experiments showed that the majority rule fared quite well, performing at levels comparable to much more computationally taxing rules. Furthermore, the majority rule outperformed the despotic best member rule, even when members were not forced to cooperate for group endeavor and free riding was possible (Kameda et al., 2011).

The Robust Beauty of the Majority Rule

These results indicate that, despite its computational simplicity, the majority rule can achieve surprisingly high levels of performance. Such observations may explain the popularity of the majority rule across the full spectrum of human groups from hunter–gatherer and tribal societies (Boehm, 1999) to modern industrial democracies (Kerr and Tindale, 2004), as well as the

animal cases in which democratic decisions are often more beneficial than despotism (Conradt and Roper, 2003). Of course, phylogenetically, humans are quite distantly related to honeybees and other social insects. Yet, the striking similarities in decision styles between two of the most social species on earth suggest that humans and honeybees have evolved structurally similar group aggregation mechanisms (e.g., utilizing the law of large numbers) to solve similar adaptive problems (e.g., foraging). Dealing with uncertainty is the key challenge underlying the evolution and use of these mechanisms.

Majority processes can be seen as a major instance of 'social sharedness' (Kameda and Tindale, 2006; Kameda et al., 2003b). Social sharedness is a description of the phenomena associated with shared cognitions and preferences playing an inordinate role in group decision making and problem solving. Majority rules are an instantiation of sharedness at the preference level. However, there is now a large amount of evidence that shared information, motives, and background knowledge can all play important roles in how groups reach consensus. It is often through shared cognitions and information that groups can perform even better than simple majority processes by a minority faction using such information to 'demonstrate' the correctness of their position to a majority that coalesced around a less-than-optimal choice alternative (see Laughlin, 2011; Tindale et al., 2012).

Intergroup Relations

A final problem that our ancestors faced was how to deal with members of other groups. As population densities increased in human evolution, so did the competition for scarce resources with members of rival groups. On the one hand, relations with outgroups provided opportunities for sharing resources such as food, mates, and information. On the other hand, intergroup relations could be a source of tension and conflict over scarce resources. As a consequence, humans likely possess highly specialized mechanisms that enable them to reap the benefits of intergroup relations while avoiding the costs.

Fear of Strangers

As part of this evolved intergroup psychology, humans are relatively suspicious and even fearful of strangers. Fear of strangers is an innate response that is seen among young children. Fear is strongest toward outgroup males, presumably because they constituted a considerable physical threat in ancestral times (McDonald et al., 2012). Outgroups posed not only a significant physical threat in ancestral environments but also a disease threat. A recent study showed that ethnocentrism is strongest among women who are in the early stage of pregnancy presumably because they (and their fetus) are most at risk of catching a disease (Navarrete et al., 2007).

Intergroup Aggression and Warfare

An adaptive solution to intergroup competition is engaging in organized violence against members of outgroups. Humans and chimpanzees use coalitional aggression to gain access to reproductively relevant resources such as territories and sexual mates (Brosnan et al., 2009). In both species, such coalitions usually consist of males, arguably because males have more to gain from participating in intergroup conflict – what has been dubbed the "male warrior hypothesis" (Van Vugt, 2009). Research on the male warrior hypothesis shows that men are more 'tribal' than women: they are more aggressive in intergroup encounters and have a stronger inclination to dehumanize outgroup members. Men are also more likely to make sacrifices on behalf of their group during intergroup conflict (Van Vugt et al., 2007). Consistent with the male warrior hypothesis, a recent study suggests that physically formidable men have a stronger preference for intergroup aggression and warfare than do less formidable men (Sell et al., 2009).

Recent game-theoretic simulations are consistent with the ideas of ingroup sacrifice and outgroup aggression (Choi and Bowles, 2007). These simulations showed that neither societies made up of mainly pure aggressors nor those of mainly pure altruists survived well over time. However, societies with mainly 'parochial altruists' (members who sacrifice for ingroup members but shun or aggress against outgroup members) proved to be evolutionarily stable. More recent work has begun to isolate the physiological and neurological correlates of these effects and has shown that oxytocin helps to regulate responses to both ingroup and outgroup members (De Dreu et al., 2010). Thus, behaving in ways that favor ingroup welfare appears to be adaptive for both the group and the individuals who depend on it.

The Promise of an Evolutionary Science of Group Dynamics

An evolutionary approach to group dynamics can be fruitful in at least four different ways. First, an evolutionary perspective can provide a more complete understanding of particular group processes by asking fundamental questions about the functions, origins, and evolution of these phenomena. A more complete account inevitably follows from rigorous attempts to establish conceptual links between evolutionary processes operating on ancestral populations and psychological processes operating within contemporary groups.

Second, an evolutionary perspective can help overcome biases and blind spots in the study of groups. It strikes us as odd that the social psychological literature on group decision making often focuses on what is wrong with groups, disregarding the fact that the group is the natural environment for humans. Examples include research on groupthink, brainstorming, group polarization, and information sharing. A cursory reading of this literature all too easily suggests that people are poor collective decision makers. Any such conclusion, however, is inaccurate (or, at the very least, overly simplistic), and we believe that an evolutionary perspective can produce more sophisticated and accurate conclusions about group decision making.

Third, an evolutionary approach is useful in yielding novel hypotheses about group behavior. As seen in this article, evolutionary reasoning has led to a number of new hypotheses about traditional group topics such as status, conformity, and social influence, which are unlikely to have been stimulated by other theoretical frameworks (Kenrick et al., 2003).

Finally, an evolutionary approach can expand the boundaries of scientific inquiry on group dynamics by suggesting important

group phenomena that have previously received little if any attention from group researchers. Laughter, language, gossip, dance, music, sports, culture, and religion are increasingly being understood as group-level adaptations, that is, as manifestations of psychological processes that connect individuals to each other in large and diverse groups, and these insights have benefited from evolutionarily informed inquiry.

In short, an evolutionary perspective reinforces our awareness that group dynamics are fundamental to the study of human nature. Furthermore, it provides a set of conceptual and empirical tools that can be used to understand and describe group processes more completely and accurately.

See also: Evolutionary Social Psychology; Human Cooperation, Evolution of.

Bibliography

Abele, S., Chartier, C., 2012. Groups Outperform Individuals in Tacit Coordination by using Consensual and Disjunctive Salience. Paper Presented at the 7th Annual Conference of the International Network of Groups Researchers. Chicago, IL.

Abele, S., Stasser, G.L., 2008. Coordination success and interpersonal perceptions: matching versus mismatching. Journal of Personality and Social Psychology 95, 576–592.

Atran, S., 2002. In Gods We Trust: The Evolutionary Landscape of Religion. Oxford University Press, Oxford, UK.

Bateson, M., Nettle, D., Roberts, G., 2006. Cues of being watched enhance cooperation in a real-world setting. Biology Letters 2, 412–414.

Boehm, C., 1999. Hierarchy in the Forest: The Evolution of Egalitarian Behavior. Harvard University Press, Cambridge, MA.

Brosnan, S.F., Newton-Fisher, N.E., Van Vugt, M., 2009. A melding of minds: when primatology meets personality and social psychology. Personality and Social Psychology Review 13, 129–147.

Brown, D., 1991. Human Universals. McGraw-Hill, Boston, Mass.

Camerer, C., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press, Princeton.

Choi, J., Bowles, S., 2007. The coevolution of parochial altruism and war. Science 318, 636–640.

Conradt, L., List, C. (Eds.), 2009. Theme issue: group decision making in humans and animals. Philosophical Transactions of the Royal Society B, vol. 364, pp. 719–852.

Conradt, L., Roper, T.J., 2003. Group decision-making in animals. Nature 421, 155–158.

De Dreu, C.K.W., Greer, L.L., Handgraaf, M.J.J., Shalvi, S., Van Kleef, G.A., Baas, M., Ten Velden, F.S., Van Dijk, E., Feith, S.W.W., 2010. The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. Science 328, 1408–1411.

Dunbar, R.I.M., 1993. Coevolution of neocortical size, group size, and language in humans. Behavioral and Brain Sciences 16, 681–735.

Dunbar, R.I.M., 2004. Grooming, Gossip, and the Evolution of Language. Faber & Faber, London, UK.

Fehr, E., Fischbacher, U., 2004. Third party sanctions and social norms. Evolution and Human Behavior 25, 63–87.

Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Quarterly Journal of Economics 114, 817–868.

Festinger, L., 1950. Informal social communication. Psychological Review 57, 271–282.

Fischbacher, U., Gächter, S., Fehr, E., 2001. Are people conditionally cooperative? Evidence from a public goods experiment. Economics Letters 71, 397–404.

Forsyth, D., 2010. Group Dynamics. Wadsworth, Belmont, CA.

Gurven, M., 2004. To give or not to give: an evolutionary ecology of human food transfers. Behavioral and Brain Sciences 27, 543–559.

Haley, K., Fessler, D., 2005. Nobody's watching. Evolution and Human Behavior 26, 245–256.

Hardy, C.L., Van Vugt, M., 2006. Nice guys finish first: the competitive altruism hypothesis. Personality and Social Psychology Bulletin 32, 1402–1413.

Hastie, R., Kameda, T., 2005. The robust beauty of majority rules in group decisions. Psychological Review 112, 494–508.

Hastie, R., Penrod, S.D., Pennington, N., 1983. Inside the Jury. Harvard University Press, Cambridge, MA.

Henrich, J., Boyd, R., Bowles, S., Gintis, H., Fehr, E., 2004. Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small Scale Societies. Oxford University Press, Oxford, UK.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J.C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, R., Tracer, D., Ziker, J., 2006. Costly punishment across human societies. Science 312, 1767–1770.

Iredale, W., Van Vugt, M., Dunbar, R., 2008. Showing off in humans: male generosity as mate signal. Evolutionary Psychology 6, 386–392.

Kameda, T., Takezawa, M., Hastie, R., 2003a. The logic of social sharing: an evolutionary game analysis of adaptive norm development. Personality and Social Psychology Review 7, 2–19.

Kameda, T., Takezawa, M., Hastie, R., 2005. Where do social norms come from? The example of communal sharing. Current Directions in Psychological Science 14, 331–334.

Kameda, T., Takezawa, M., Ohtsubo, Y., Hastie, R., 2010. Are our minds fundamentally egalitarian? Adaptive bases of different sociocultural models about distributive justice. In: Schaller, M., Norenzyan, A., Heine, S.J., Yamagishi, T., Kameda, T. (Eds.), Evolution, Culture, and the Human Mind. Psychology Press, New York, pp. 151–163.

Kameda, T., Takezawa, M., Tindale, R.S., Smith, C., 2002. Social sharing and risk reduction: exploring a computational algorithm for the psychology of windfall gains. Evolution and Human Behavior 23, 11–33.

Kameda, T., Tindale, R.S., 2006. Groups as adaptive devices: human docility and group aggregation mechanisms in evolutionary context. In: Schaller, M., Simpson, J.A., Kenrick, D.T. (Eds.), Evolution and Social Psychology. Psychology Press, New York, pp. 317–341.

Kameda, T., Tindale, R.S., Davis, J.H., 2003b. Cognitions, preferences, and social sharedness: past, present and future directions in group decision making. In: Schneider, S.L., Shanteau, J. (Eds.), Emerging Perspectives on Judgment and Decision Research. Cambridge University Press, New York, pp. 458–485.

Kameda, T., Tsukasaki, T., Hastie, R., Berg, N., 2011. Democracy under uncertainty: the wisdom of crowds and the free-rider problem in group decision making. Psychological Review 118, 76–96.

Kameda, T., Wisdom, T., Toyowaka, W., Inukai, K., 2012. Is consensus-seeking unique to humans? A selective review of animal group decision-making and its implications for (human) social psychology. Group Processes and Intergroup Relations 15, 673–689.

Kaplan, H., Hill, K., 1985. Food sharing among Ache foragers: tests of explanatory hypotheses. Current Anthropology 26, 223–246.

Kemp, A.H., Guastella, A.J., 2011. The role of oxytocin in human affect: a novel hypothesis. Current Directions in Psychological Science 20, 222–231.

Kenrick, D.T., Li, N.P., Butner, J., 2003. Dynamical evolutionary psychology: individual decision-rules and emergent social norms. Psychological Review 110, 3–28.

Kerr, N.L., Tindale, R.S., 2004. Group performance and decision making. Annual Review of Psychology 55, 623–655.

King, A., Johnson, D.D.P., Van Vugt, M., 2009. The origins and evolution of leadership. Current Biology 19, 1591–1682.

Komorita, S.S., Chertkoff, J.M., 1973. A bargaining theory of coalition formation. Psychological Review 80, 149–162.

Laughlin, P.R., 2011. Group Problem Solving. Princeton University Press, Princeton, NJ.

Leary, M.R., 1999. Making sense of self-esteem. Current Directions in Psychological Science 8, 32–35.

Littlepage, G., et al., 2008. Transactive memory and performance in work groups. Group Dynamics 12, 223–241.

McDonald, M.M., Navarrete, C.D., Van Vugt, M., 2012. Evolution and the psychology of intergroup conflict: the Male Warrior Hypothesis. Philosophical Transactions of the Royal Society - Biological Sciences 367, 670–679.

Moreland, R., Argote, L., Krishnan, R., 1996. Socially shared cognition at work. In: Nye, J., Brower, A. (Eds.), What's Social about Social Cognition. Sage, Thousand Oaks, CA.

Navarrete, C.D., Fessler, D.M.T., Eng, S.J., 2007. Elevated ethnocentrism in the first trimester of pregnancy. Evolution and Human Behavior 28, 60–65.

Ridgeway, C.L., Berger, J., Smith, L., 1985. Nonverbal cues and status: an expectation states approach. American Journal of Sociology 90, 955–978.

Roberts, G., 1998. Competitive altruism: from reciprocity to the handicap principle. Proceedings of the Royal Society of London: Series B 265, 427–431.

Schelling, T., 1960. The Strategy of Conflict. Harvard University Press, Cambridge, MA.

Seeley, T.D., 2010. Honeybee Democracy. Princeton University Press, Princeton.

Sell, A., Cosmides, L., Tooby, J., 2009. Human adaptations for the visual assessment of strength and fighting ability from body and face. Proceedings of the Royal Society of London: Series B 276, 575–584.

Shinada, M., Yamagishi, T., Ohmura, Y., 2004. False friends are worse than bitter enemies: "altruistic" punishment of in-group members. Evolution and Human Behavior 25, 379–393.

Stallen, M., De Dreu, C.K.W., Shalvi, S., Smidts, A., Sanfey, A.G., 2012. The herding hormone: oxytocin stimulates in-group conformity. Psychological Science. http://dx.doi.org/10.1177/0956797612446026.

Surowiecki, J., 2004. The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Doubleday, New York.

Tajfel, H., Turner, J., 1979. An integrative theory of intergroup conflict. In: Austin, W., Worchel, S. (Eds.), The Psychology of Intergroup Relations. Brooks/Cole, Monterey, CA, pp. 33–47.

Tindale, R.S., Talbot, M., Martinez, R., 2012. Group decision making. In: Levine, J.M. (Ed.), Group Processes. Psychology Press, New York, pp. 165–194.

Tracy, J.L., Robins, R.W., 2004. Show your pride: evidence for a discrete emotion expression. Psychological Science 15, 194–197.

Van Vugt, M., 2006. Evolutionary origins of leadership and followership. Personality and Social Psychology Review 10, 354–371.

Van Vugt, M., 2009. Despotism, democracy and the evolutionary dynamics of leadership and followership. American Psychologist 64, 54–56.

Van Vugt, M., De Cremer, D., Janssen, D., 2007. Gender differences in cooperation and competition: the male warrior hypothesis. Psychological Science 18, 19–23.

Van Vugt, M., Hogan, R., Kaiser, R., 2008. Leadership, followership, and evolution: some lessons from the past. American Psychologist 63, 82–196.

Van Vugt, M., Kameda, T., 2012. Evolution and groups. In: Levine, J.M. (Ed.), Group Processes. Psychology Press, New York.

Wegner, D.M., 1987. Transactive memory: a contemporary analysis of the group mind. In: Mullen, B., Goethals, G.R. (Eds.), Theories of Group Behavior. Springer-Verlag, New York, pp. 185–208.

Evolutionary Medicine

Alexandra Alvergne, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Evolutionary medicine is a set of concepts and tools with which to address medical problems. This young science asks novel questions, starting from why the human body is vulnerable to diseases to how to exploit evolutionary dynamics to improve cancer therapy. In an attempt to assess why, how, and to what extent evolutionary medicine sheds light on medical issues, this article reviews (1) the conceptual tools borrowed from evolutionary biology and their medical relevance, (2) how evolutionary medicine is evolving and its main criticisms, and (3) what practical medical applications are or can be derived.

Evolutionary medicine (EM), also known as Darwinian medicine, first emerged among the biological sciences following the motivation to understand, broadly, why the body is vulnerable to diseases (Nesse and Williams, 1996; Williams and Nesse, 1991). The value of an historical perspective for answering medical questions had first been suggested by Erasmus Darwin, a physician and the grandfather of Darwin, while the first attempt to apply evolutionary reasoning to assess medical practices dates back to the 1980s (Ewald, 1980). The field as it is currently known eventually emerged a decade later through the combined efforts of Randolph M. Nesse, a US practicing psychiatrist, and George C. Williams, a US professor emeritus of ecology and evolution, first with a major publication in the *Quarterly Review of Biology* in 1991 (Williams and Nesse, 1991) and second with a popular science book: "*Why we get sick: the New Darwinian Science of Medicine*" in 1995. Since then, the field has grown and evolutionary biologists mainly, but also some physicians, increasingly argue for the complementarity of an 'ultimate' perspective in answering medical questions (Nesse and Stearns, 2008; Nesse and Williams, 1996). It consists in understanding whether and the extent to which evolutionary processes such as natural selection and drift (e.g., hazard or chance) can account for why we are all vulnerable to diseases as well as patterns of variation in disease susceptibility among populations, individuals, and across the life span. Such an understanding enables us to question how current medical practices influence evolutionary processes (e.g., the evolution of resistance to antibiotics and chemotherapy), or might interfere with evolved predispositions (e.g., mother–infant separation at birth and the onset and maintenance of lactation). EM also has wider relevance for challenging the idea of norms in public health recommendations (e.g., nutrition), that of a perfectly designed body, or whether the pursuit of extending the life span is a sound objective when one considers how the body has been shaped by natural selection.

EM is a relatively young and fast growing area, which has, since its emergence, generated not only much promising research but also a few critiques. An undisputed benefit of applying evolutionary thinking to medical issues is in understanding the evolution of resistance to antibiotics and chemotherapy, which has obvious and critical implications for public health policies and medical practices. Given its potentially large domain of application, EM is increasingly prevalent within the biological sciences, as testified to by a number of edited books, scientific journals, workshops, and blogs. Yet, the 'e-word' has a tendency to be avoided in mainstream medicine (Antonovics et al., 2007) and the social sciences are generally suspicious of evolutionary approaches to human affairs, for several reasons. First, this avoidance is not independent of past societal experience and is likely a legacy of the infamous eugenics movement and the misuse of evolutionary theory for social purposes over the first half of the twentieth century (*see* Evolution, History of). This raises the need to define what EM is not, which has been clarified by Nesse and Williams (1996, p. 11): "We are by no means advocating that medicine should assist natural selection, nor do we suggest that biology can guide moral decisions […]. We are urging not an alternative to modern medicine but rather an additional perspective […]. We would be very much against the idea of EM as a kind of alternative cult." Second, others have questioned whether Darwin would recognize himself the premises that guide research and the extent to which adaptive hypotheses are favored over nonadaptive ones (Marks, in Elton and O'Higgins, 2008). Consequently, there is an increasing tendency to refer to evolutionary rather than Darwinian medicine, which acknowledges the fact that the discipline cannot claim any presumed recognition from Darwin himself. As for all sciences, evolutionary approaches to human affairs in general and to medicine in particular are themselves evolving. Since the launch of EM as we know it, some of the initial assumptions have been discussed, such as the validity of the EEA, i.e., the 'evolutionary environment of adaptiveness' (Foley, 1995). Ecology, a science focusing on the role of the social and biological environments in shaping disease variation, is increasingly being combined with the historical approach. Evolutionary thought now permeates research in various medical areas such as molecular biology, physiology, genetics, development, neuroscience, and psychiatry (**Table 1**). This demonstrates that EM is not a field (Stearns, 2012), but rather a series of concepts and tools to analyze medical questions from a different angle.

This article attempts to grasp why, how, and to what extent EM sheds new light on medical problems by presenting an overview of (1) the questions raised by an evolutionary perspective and their medical relevance; (2) the criticisms made of research in EM; and (3) some key fields of applications. By no means is the review exhaustive, and the reader will find more

Table 1 Evolutionary thought interacts with medical issues along a broad and diverse front

Evolutionary specialties	*Medical issues*	*Diseases*
Evolutionary genetics	Genetic basis of disease	Any with a genetic element
	Genomic conflicts	Growth disorders, cancer, mental diseases
Evolution of development	Fetal programming	Diseases of pregnancy
Evolutionary physiology	Hormonally mediated	Reproductive cancers
	Trade-offs affecting reproduction	Susceptibility to infection
	Trade-offs affecting aging	Degenerative diseases
Evolutionary neuroscience	Mental diseases	Mental disorders as disruptions of mental adaptations
	Addiction	Substance abuse
Molecular phylogenetics	Understanding origins	Cancer, emerging diseases
	Microbiomes	Forensic medicine
Evolutionary psychiatry	Mental diseases	Depression, anxiety
Evolutionary ecology	Microbiomes	Gut floras and disease
	Antibiotic resistance	Infectious diseases
	Chemotherapy resistance	Cancer

Taken from Stearns, S.C., 2012. Evolutionary medicine: its scope, interest and potential. Proceedings. Biological Sciences/the Royal Society 279(1746), 4305–4321. http://dx.doi.org/10.1098/rspb.2012.1326.

detailed information on the Evolutionary Medicine Review Web site (http://evmedreview.com) and in the burgeoning and extensive literature authored by R. Nesse, S. Stearns, L. Trevathan, G. Bentley, and P. Gluckman, to name a few.

Evolutionary Conceptual Tools and Their Medical Relevance

Mainstream Western medicine is usually regarded as taking a relatively mechanistic approach: it aims to understand how the body works and focuses on immediate (i.e., proximate) developmental, molecular, anatomical, or physiological causes of illness. It views disease as a malfunction of the body, the treatment of which generally aims to reset the body back to a 'normal' state (Nesse and Williams, 1996). An ultimate historical perspective aims to understand why we are all vulnerable to diseases in the first place. Put another way, why has natural selection left us so poorly equipped to fight pathogens and deal with new environmental conditions? As compared to the traditional proximate approach of modern medicine, an ultimate approach allows a shift in perspective to understand the flaws of bodies as the result of historical legacies or compromises between bodily functions that have been shaped by natural selection over time. An evolutionary perspective challenges the idea that the body is perfectly designed (Nesse and Stearns, 2008) as evolution is not 'an engineer' but 'a tinkerer,' with no preconceived plan, constrained to work with what is already there (Jacob, 1977). EM also emphasizes that natural selection is slow, such that the body is never entirely 'adapted' to its current environment (either pathogenic or nutritional), and maximizes individual reproduction over health (Nesse and Stearns, 2008). Given this, EM questions whether a particular bodily response to changing environmental conditions is an adaptation or a pathology (e.g., fever and child development), why the human body is equipped with apparently useless features (e.g., wisdom teeth and appendixes) and maladaptive ones, why new environmental conditions are associated with the emergence of new diseases, why we die, and whether we can win the fight against pathogens. More generally, proponents of EM consider the 'evolutionary significance' of each disease: why is it there? Why has it not been eliminated by natural selection? Why are some people susceptible to a particular type of disease while others are not? (Nesse and Williams, 1996)

The evolutionary sciences contribute both theoretical concepts and methods to address these questions. First, the evolutionary causes of diseases are generally grouped into the concepts of (1) arms races, to understand the evolution of the immune system, the evolution of pathogen and cancer resistance, as well as the expression of genomic conflicts; (2) life-history trade-offs and pleiotropic effects, reflecting the role of constraints between fitness functions across the life span and between sexes; (3) mismatch between current and past environments, which refers to the time lag human biology faces against both cultural and pathogen evolution, to understand the emergence of 'diseases of civilization'; (4) historical legacies or the idea that selection cannot start all over again and has to work with what is already there. One important method brought by evolutionary biology are phylogenetic trees (*see* Comparative Method in Evolutionary Studies), allowing us, for instance, not only to infer the genealogical relationships between clones in cancer clonal evolution (Yachida et al., 2010) but also to test hypotheses about the role of contingency (history) and constraint in the evolution of traits. Another method is population genetics (*see* Microevolution), enabling us to infer changes in allele frequency under the influence of evolutionary processes such as natural selection and drift. Most often, a combination of concepts is needed to provide a complete evolutionary account for a given disease. For the sake of simplicity, however, conceptual tools are presented here separately.

Evolutionary Arms Races: Host–Pathogen Coevolution and Genetic Conflict

Individuals compete not only with each other but also with members of other species, which leads to coevolutionary processes. To illustrate that the survival of individuals depends on their ability to constantly adapt to a changing environment, Leigh Van Valen (1973) proposed the Red Queen principle following what the Red Queen told Alice in Lewis Carroll's

piece *Through the Looking-Glass*: "now, here, you see, it takes all the running you can do, just to keep in the same place." This principle sheds light on the evolutionary processes underlying host–pathogen interactions, moving away from the idea that both parties are in symbiosis and rather to the more likely scenario that they are involved in an evolutionary arms race without end. In the battle to evolve weapons and counterweapons, humans are at a great evolutionary disadvantage. Pathogens reproduce much faster than we do; it takes thousands of years to produce 300 human generations, versus a few weeks in the case of bacteria and virus (Nesse and Williams, 1996). These much shorter generation times provide an advantage to bacteria and viruses, increasing the rate at which natural selection can favor mutations that are beneficial in the arms race. Yet, humans have evolved a myriad of defense mechanisms against pathogens, including general adaptations such as coughing, diarrhea, and vomiting as well as a sophisticated chemical weaponry expressed by the immune system deploying diverse and complex antibodies (Nesse and Williams, 1996: p. 52). In the cases of intense selective pressures, resistance to pathogens can evolve in just a few generations, and there is some evidence of selection for resistance to malaria, small pox, and tuberculosis, in the form of an increase in the frequency of an initially rarer genetic variant. However, if the outbreak is too severe and a mutation that confers counteradaptation does not already exist by chance, all infected individuals will die and the gene pool will not change. In this context, pathogens have another advantage, as the probability of a beneficial mutation already being present increases with the number of pathogens in the body, which can potentially be very high. As a result, we cannot evolve fast enough to escape from microorganisms (Nesse and Williams, 1996). Understanding processes of coevolution in host–pathogen interactions has implications across a wide range of topics, including the evolution of the immune system and defense mechanisms, the evolution of pathogen resistance and virulence, and the expression of genetic conflicts resulting from the battle between paternal and maternal genomes.

The Red Queen principle has direct implications for understanding how human interventions and practices impact on the evolution of pathogen resistance and virulence. For instance, long-term use of antibiotics, by exerting a strong selective pressure on resistant bacteria, will rapidly cause an initially rare antibiotic-resistant bacterium to increase in frequency (e.g., most strains of staphylococcal bacteria are now resistant to penicillin and ciprofloxacin). Similarly, the evolution of pathogen virulence is shaped by human practices that influence pathogen transmission: increased opportunities for transmission, decreasing the value of maintaining the host alive, are correlated with increased virulence. In the context of human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome (AIDS), for instance, virulence is expected to increase in response to increased opportunities of transmission, which are dependent on the number of intravenous drug users or sexual partners. This argument is supported by demographic, historical, and epidemiological data. First, the rarity of clinical AIDS in Central Africa prior to the HIV-1 pandemic that emerged in the 1980s, combined with the long-term presence of HIV in human populations, suggest that pandemic HIV-1 had increased virulence relative to the ancestral HIV (Ewald, 1994: p. 132). Second, demographic data show that the 1980s African HIV-1 pandemic has followed a decade of extensive migration from rural to urban areas, which brought about a higher demand for sexual commerce. Conversely, campaigns advocating the use of condoms have diminished the opportunity of transmission, favoring strains with decreased virulence (Ewald, 1994). Following this line of reasoning, whenever a disease is transmitted by a vector (e.g., mosquitoes for malaria and hands of doctors for puerperal fever) thereby increasing the opportunities for transmission as compared to personal contacts, virulent pathogens immobilizing the host will become more frequent (Ewald, 1994). The evolution of virulence as a response to practices can occur quickly because reproductive cycles are short in pathogens, and this can be exploited to decrease the incidence of virulent strains.

The arms race metaphor is also useful in questioning whether or not we should treat symptoms. In other words, is the bodily response to an infection such as coughing or diarrhea a manipulation of the pathogen to increase its probability of transmission (e.g., rabies) and/or an adaptive immune response to eliminate the pathogen (e.g., fever)? In many cases, the defense and the manipulation hypotheses are not mutually exclusive. For instance, diarrhea following infection by *Shigella* helps both the bacteria to spread and the host to resist the infection, with the administering of some treatment (diphenyloxalate) delaying the recovery (Ewald, 1994). Then, if a defense is adaptive, should we block it? It is important to note that the onset of defensive mechanisms is not without cost (for instance, high fever is related to increased probability of tissue damage) and if alternative treatment such as antibiotics and antipyretics could eliminate the infection without incurring the defense cost, then treating the symptom should be preferred (Ewald, 1980). However, if the pathogens become resistant to antibiotics and very high dosages are required, the decision should be reevaluated depending on the goal of the treatment (eliminating the disease or managing pain) (Ewald, 1980). In general, the symptomatic treatment of disease is more relevant with manipulative manifestations rather than defensive manifestations of the disease (Ewald, 1980). The EM approach can thus either recommend treatment or its abstention, which depends on testing the validity of the manipulation and defenses hypotheses.

Finally, the arms race concept is important for understanding diseases that result from a deregulation of the evolutionary conflict between the sexes, such as preeclampsia, Angelman syndrome (AS) (neurogenetic disorder associated with various symptoms including intellectual and developmental impairment), or Prader–Willi syndrome (PWS) (neurogenetic disease associated with a wide range of symptoms including excessive eating and morbid obesity). An arms race between the sexes may occur when maternal and paternal genes do not share the same interest over the amount of resources invested in each offspring (i.e., when maternal investment of resources in one offspring reduces the amount of remaining resources available for other offspring who do not share the same father), and when there is the opportunity to manipulate parental resources (i.e., when there is physical contact with the maternal parent such as in angiosperms and in mammals)

(Úbeda and Wilkins, 2008). The arms race results in the differential expression of the copies of a gene depending on their parental origin, with one parental copy being overexpressed and the other underexpressed, a phenomenon called genomic imprinting (Haig, 2002) (*see* Sociobiology: Overview). One of the best known examples of genomic imprinting modulating access to maternal resources in mammals is the gene coding for the insulin-like growth factor 2, mainly responsible for embryonic growth, where the paternal copy is overexpressed and the maternal copy underexpressed or silent. Most known imprinted genes are involved in growth, but there is also some evidence that they are involved in neural development, with paternally imprinted genes being associated with the hypothalamus (development of emotions) and maternally inherited genes being associated with the cortex (development of intelligence). A deleterious recessive mutation on an imprinted gene is likely to be associated with a strong phenotypic effect, as the copy from the other parent cannot maintain the normal phenotype. For instance, when there is a modification of an inherited allele on chromosome 15, individuals develop either AS if the modification is on the maternal copy or PWS if it is on the paternal copy. In AS, the loss of the maternal allele selected to reduce the demand for maternal resources is associated with increased activity and prolonged suckling of the new born. In addition, individuals with AS suffer mental retardation and difficulties with speech. Conversely, in PWS, the loss of the paternal allele selected to increase the demand of maternal resources leads to reduced activity and poor suckling, hyperphagia during childhood, obesity, and poor libido (Bentley & Aunger, in Elton and O'Higgins, 2008). Regarding mechanisms, an imprinted gene expression is associated with epigenetic modification on each allele, such as DNA methylation and histone modification (*see* Epigenetic Inheritance). These modifications are established during gametogenesis but can be reprogrammed later in life, and a failure of this system could lead to the development of diseases at any stage of life. Dietary conditions at critical development stages may result in a shortage or an excess of methyl donors, which might alter the pattern of expression of imprinted genes. Given the current state of knowledge, however, it is unclear whether and when such alteration is adaptive or pathological, and it has been suggested that simply supplying the missing nutrient at a later developmental stage may create a new set of disease conditions (Úbeda and Wilkins, 2008).

Evolutionary Constraints: Life-History Trade-Offs and Aging

An evolutionary approach views the body as resulting from a series of trade-offs between fitness functions: when resources are limited, resources are traded off between competing functions such as growth, survival, and reproduction. Allocating more resources to reproduction implies a smaller allocation to growth such that, for instance, across species and within species, those with an earlier age of first reproduction also tend to have the shortest growth (Stearns, 1992). It follows that there is no such thing as a Darwinian Demon (a hypothetical organism that can maximize all elements of its fitness simultaneously), as not all functions can be simultaneously optimized. Trade-offs vary within life cycles as defined by patterns of growth (infancy, juvenility, childhood, adolescence, and mature adult), and within any specific life cycle, natural selection is expected to have shaped the optimal allocation of resources between different functions depending on individual characteristics (e.g., sex) and environmental circumstances (e.g., availability of nutritional resources). At a proximate level, such trade-offs are mediated by hormones (steroids and metabolic), which vary in their expression between individuals and across the life span; for instance, testosterone is not only associated with enhanced competitive abilities and mating success but is also an immunosuppressant. The conceptual foundation that views organisms as having life cycles rather than having a life cycle (Bonner cited in Stearns, 1992) and that within each life cycle, phenotypic traits result from the action of natural selection on trade-offs between fitness functions has several theoretical and practical implications.

First, it is directly related to the antagonistic pleiotropy hypothesis for understanding the evolution of senescence (Williams, 1957), one of the most puzzling questions in evolutionary biology. Building on the work of Haldane (1941), who reasoned that diseases that are expressed late in life after reproduction might have already occurred (e.g., Huntington disease) cannot be eliminated by natural selection, and Medawar (1952), who pointed out that selection weakens with age as fewer individuals are still alive within older age groups, Williams proposed that genes causing death can be selected for if they provide a fitness advantage early in life (Williams, 1957). This hypothesis, which assumes that senescence is a life-history trait shaped by natural selection, is prevalent in the literature and is supported by studies showing (1) that there is heritable variation for senescence and (2) alleles that slow aging also often reduce reproduction (reviewed in Partridge and Gems, 2006). There is still controversy in the field as not all genes that cause senescence necessarily have a benefit, but EM has changed the way researchers view the aging process by stimulating the investigation of biological variation in aging rates among animal species and individuals (Partridge and Gems, 2006). Along these lines, studies of centenarians have revealed that heritable variation in inflammatory activity could be involved in shaping senescence rates. Senescence is now increasingly considered to be a by-product of selection for the inflammatory response, with the benefit of counteracting agents early in life being downsized by continuous and long-term proinflammatory condition later on, in a life stage largely unforeseen by evolution (Franceschi et al., 2000).

Second, a life-history perspective questions the idea of norms in child developmental medicine. Because of the diversity and unpredictability of environments, natural selection is expected to have favored the ability to adjust development to local socioecological conditions, a phenomenon known as developmental plasticity (*see* Evolutionary Developmental Psychology). For instance, across populations, the pace of development is correlated with subadult mortality in the local environment, controlling for energetic availability (i.e., the velocity of growth controlling for body size), the development being faster in high-mortality environments. It suggests that there has been selective pressure for accelerating life history when risk of mortality, and thus the risk of not reproducing, is high (Walker et al., 2006). In addition, growth

trajectories and the onset of puberty depend on the availability of resources, influencing the strength of trade-offs. Environmental effects such as low weight gain during childhood due to malnutrition or resulting from intense sport activity is correlated with delayed age at sexual maturation (see Gluckman and Hanson, 2006). Overall, those with the earliest age at menarche are those who were born in a high-mortality environment and encountered good conditions during childhood. This adjustment explains why girls born in developing countries and adopted at an early age in a developed country have a precocious development (Gluckman and Hanson, 2006). An evolutionary approach views early age of menarche as an evolved response to local environmental conditions rather than a pathology, which in turn questions the rationale for prescribing gonadotropin suppressors to retard sexual development. More generally, the idea of a norm for child development is challenged.

Environmental Mismatch and Genetic Quirks

The rationale underlying the mismatch hypothesis is that environmental change creates a new set of conditions that the genome has not experienced before. Some heritable variation will become helpful, harmful, or neutral depending on the environment. It follows that diseases are not caused by defective genes but by novel environments (Nesse and Williams, 1996). The mismatch may result from rapid and recent cultural change in diet and change in life history and it has been argued to account for the 'diseases of civilization' such as obesity, allergies, vitamin D deficiencies, or addiction (Nesse and Williams, 1996). For instance, a low availability of fat and sugar during human evolution has selected for an intense craving for those nutrients. Unfortunately, the exploitation of this innate taste by the food industry has contributed to increase the prevalence of obesity and diabetes in the recent past, where fat and sugar have been made available in large quantities. Similarly, a change in women's life history, i.e., earlier age at menarche, later age at first birth, and lower number of children, exposes women to different hormonal patterns, which is not independent of the occurrence of reproductive cancers (Eaton et al., 1994). Provided that the new environmental conditions are stable enough over generations, genetic evolution might eventually catch up with cultural change. For example, the reliance on dairy products for subsistence has selected for the ability to digest lactose during adulthood among populations who have practiced agriculture. However, at any point in time, there might be a mismatch between the current environmental conditions and those that have shaped the evolution of a particular trait.

One highly cited example of a genetic quirk (genes that have no deleterious effects except when exposed to novel environments (Nesse and Williams, 1996: p. 106)) is myopia, the development of which requires both genetic and environmental factors. Twin studies have revealed that myopia has a strong genetic basis, and the disease is expressed in about 25% of Americans (Nesse and Williams, 1996). This high frequency is puzzling, as surely this condition should have been selected against before the advent of modern vision correction devices. A longitudinal study shows the role of environmental novelty in the onset of the condition. Among the East Greenland Inuit, no cases of nearsightedness were reported in the 1950s, but by 1994, myopia had increased by 25 and 17% among the 5- to 9-year-olds and the 10- to 14-year-olds, respectively (Norn, 1997). Although the myopinizing factor is unknown, it has been suggested to be either schooling and/or change in diet as such cultural pressures were not present before 1942. The value of the mismatch hypothesis lies in both challenging the idea of an optimum body and highlighting the importance of cultural change for the onset of diseases. As Konner put it (cited in Nesse and Williams, 1996: p. 106) "the discovery of a genetic determination for a disorder may provide the best hope for an environmental treatment of it."

Historical Legacies and 'Scars' of Evolution

The features of human biology have been established by long-term historical processes, which combined with the impossibility for natural selection to start all over again, lead to the occurrence of some design flaws. For instance, we share with all vertebrates an air–food traffic problem in that food might go down 'the wrong pipe.' This results from the position of the esophagus, located behind the trachea, which must be closed by an epiglottis when food travels to the stomach. This anatomical flaw, leading to 1 in 100 000 people choking to death each year, is a legacy from a remote ancestor, a minute wormlike animal for whom the digestive system was co-opted to serve a respiratory function (Nesse and Williams, 1996). Historical processes have also been invoked to explain why 70–80% of US citizens are affected by back problems each year, why we have varicose veins and hemorrhoids, and why our cooling system is not good enough to protect us from sun stroke (for a detailed account of the 'scars of evolution,' see Morgan, 1994).

Of particular interest are the 'scars' associated with shifting from quadrupedal walking to bipedal locomotion during human evolution. Although some adaptive changes have occurred, some flaws still persist with regard to the skeleton and the regulation of blood pressure. First, the verticality associated with walking on two rather than four legs imposed new constraints on the originally single-arched spine. The human backbone now develops into an S shape during infancy, which partly deals with the new vertical weight-bearing stresses. At 4 months, an upward curve along the neck region allows babies to maintain their head up, and by 1 year, an additional backward curve down the lower trunk allows babies to walk. To be able to pivot on the front end, the backbone S shape has been accompanied by a change in the shape of vertebrae into that of a wedge (Morgan, 1994). This in turn weakens the backbone later in life in the region where the wedge shape is most pronounced, i.e., the lower back. Second, verticality imposes new gravitational constraints on circulation, as blood returning to the heart must overcome a higher gravitational pull than it did on four legs. This leads to varicose veins despite a system of valves in the legs to prevent blood reflux, and hemorrhoids, which does not even benefit from a valve system as the anus is not lower than the heart in a four-legged mammal (Morgan, 1994). Overall, historical legacies have been implicated in the etiology of acne, anorexia,

anxiety, back pain, cancer, obesity, sleep disorders, snoring, and sun stroke thrombosis (Morgan, 1994).

Criticisms

If the e-word has been avoided in medical publications when referring to evolutionary processes, the science of EM, as advocated by Nesse and Williams, has received little direct criticism so far (but see Marks in Elton and O'Higgins, 2008; Cournoyea, 2013). This situation likely results from the appeal of EM in providing a theoretical framework in which researchers, students, and patients can organize medical facts, combined with the more general acceptance and fashion of evolutionary theory in explaining the living world among scientists, journalists, and the general public alike. Yet EM is a science, and as such, requires a critical perspective. Concerns can be specific to the use of evolutionary theory in medicine (Marks, in Elton and O'Higgins, 2008) and can also be more general and echo those directed at evolutionary biology (e.g., 'the adaptationist program' Gould and Lewontin, 1979), biological anthropology and medical ecology (e.g., too little interest in the role of political economy in shaping human/environment coevolution: Singer, 1996). More generally, EM is questioned regarding (1) its assumptions about "biology, morality, and society" (Marks, in Elton and O'Higgins, 2008) and (2) its relevance to practical medicine (Cournoyea, 2013).

The first area of concern with EM is that as a science, it is not independent from the political and economic contexts in which it is practiced, which has implications for the emergence and shaping of scientific ideas as well as their applications (Marks, in Elton and O'Higgins, 2008). The use of evolutionary theory in medical matters has known several forms or meanings, from the British and American Eugenics societies of the 1920s aiming at improving the human race, to modern EM aimed at understanding the vulnerability to diseases. Yet Marks argues that it is not possible to distinguish between the use and misuse of evolution in medicine as it is always related to the political system in which it is expressed (Marks, in Elton and O'Higgins, 2008). For instance, following the idea that different populations have known different evolutionary histories, as exemplified by the phenotypic variation in the ability to digest milk, there has been an appeal for tailoring medicine to different ethnic populations. The economic potential of such a field has not escaped lobbyists, who soon developed the field of pharmacogenomics aimed at improving the efficacy of treatment by tailoring drugs to individual genetic makeup. This is exemplified by the first 'ethnic' drug, BIDIL, a therapy to treat heart failure among African Americans (a group known to suffer higher rates of heart diseases), which was approved by the US Food and Drug Administration in 2005. The rationale for the drug, however, ignores current knowledge on the sociocultural causes for health disparities among groups as well as genetic variation within groups. It turned out that BIDIL was medically inefficient, its mechanism was not really understood, and it became a financial flop. This example echoes some criticisms that have been put by 'critical medical anthropologists' to medical ecologists and biological anthropologists, pointing out that denying the political dimensions of science is itself a political act, which can be dangerous in an era where health is a public responsibility funded by tax payers (Morgan, 1993).

A second set of concerns touches on the premises on which research is built and how it might shape the interpretation of the causes of diseases. This criticism is more about the evolutionary sciences as a whole than about EM, and dates back to 1979. In the 'Spandrels of San Marco,' the biologists Gould and Lewontin famously cautioned against the 'adaptationist program,' which consists, among other things, in seeing all parts of all living things as having evolved by natural selection to fulfill a specific function (Gould and Lewontin, 1979). They argue in favor of a pluralistic approach in which adaptation is not viewed as the default explanation, calling for a consideration of the role of other evolutionary processes such as drift (when a change in allele frequency is due to stochastic events rather than selective processes) (*see* Microevolution). For instance, as history and chance can play a large part in the evolution of a trait (Jacob, 1977), data on the demographic history of a population (e.g., past population bottlenecks) and its genealogy (e.g., founder effects) are needed to tease apart whether a disease is present due to selective processes (e.g., sickle-cell anemia selected for its protective effects against malaria) or demographic history (e.g., Sachs disease among Ashkenazi Jews due to a founder effect). Focusing only on adaptive patterns is problematic when patterns of variation result from both selection and demographic effects. Thus, Marks wonders whether 'Darwin would recognize himself' in modern EM, as Darwin's own writings favored a pluralistic approach, and reminds us that since Darwin offers scientific legitimacy, this legitimacy can be exploited (Marks, in Elton and O'Higgins, 2008).

Third, resistance might have been fostered by the strong revelatory promises and deterministic feel associated with the human genome project, in which the role of current socio-ecological forces in shaping phenotypic variation are secondary. More generally, one area of contention is the greater importance given to phylogenetic (proximity of descent) as opposed to adaptive (shaped by the environment) divergence, or the role of ancestry versus ecology. While the genetic difference between humans and chimpanzees is very small, we are different in many respects. Are chimpanzees the most relevant models to understand humans (due to low genetic divergence)? If one wants to understand cooperative breeding and its implications for mother–infant bonding, or the relationship between testosterone, survival, and paternity, one might better turn toward the *Callitrichids*, more distant genetically but showing convergent evolution for those traits. The problem of disconnecting genetics from ecology extends to other areas of research, and in particular to what is now known as the Paleodiet. Since its first formulation, the Paleodiet has gained considerable credit in popular science, the media, and the public, and countless books advocating the Stone Age diet have flourished. Mainly based on studies of modern hunter-gatherers, the idea is that humans are adapted to Palaeolithic times, and therefore anything related to agriculture

and the modern environment is detrimental. In particular, it recommends a special ratio of plant and protein (although the ratio itself has varied depending on the way the data are analyzed). While an historical perspective is highly relevant to understand why there is little chance that many of us are adapted to a large quantity of glucose–fructose syrup, this perspective tends to ignore the tremendous climatic variation during the Pleistocene (200 000 years), the high ecological variation in the diet of modern hunter-gatherers (Elton and O'Higgins, 2008), and the fact that modern hunter-gatherers probably descend from a selected sample of historical populations who did not take up agriculture. The idea of a threshold is also 'antievolutionary' (coined by Marks in Elton and O'Higgins, 2008), as many instances of genetic adaptations have been documented since the advent of agriculture (e.g., lactase dehydrogenase). This exemplifies how nice stories can take over the public space even though they might not stand up to data.

Finally, the utility of the approach of EM in the everyday practice of the doctor has been questioned (Cournoyea, 2013), as medicine and evolutionary biology have different goals (e.g., treating patients versus understanding why diseases materialize in the first place). Some regret the lack of definitive answers and clear guidelines (Marks, in Elton and O'Higgins, 2008). This echoes the more general concern about the value of fundamental as opposed to applied research. Applied research cannot be designed before fundamental questions have been solved, however. For instance, understanding the phenomenon of infection transmission in the case of puerperal fever has led, although with some delay, to better hygiene in hospitals wards and ultimately, better health. As for the lack of definitive answers, one might contend that a major input of EM is to caution against clear normative guidelines without an understanding of the underlying processes. At this stage, EM is better viewed as a tool to assess current medical practices and public health initiatives. Illustrative examples of potential applications devised by an evolutionary perspective are detailed in the next section.

Applications

The science of EM has been about 20 years old at the time of writing this article. Attempts to draw concrete applications from EM, however, have taken off only recently. In 2008, an edited book by Elton and O'Higgins "*Medicine and Evolution: Current Applications, Future Prospects*" provided one the first attempts to face the challenge posed by translating theory into practice, discussing to what extent it is possible and what areas appear the most promising. In this line, peer-reviewed journals such as *Evolutionary Applications* and *Evolution, Medicine and Public Health* have recently been launched by the Evolution, Medicine, and Public Health Foundation. Overall, the young science of EM has tremendous potential to understand the etiology of diseases, assess current practices, and contribute to future diagnostic and therapeutic tools as well as public health programs. Such applications remain marginal to mainstream medicine, however, which is not independent from the fact that EM is not included in the curricula of medical schools, yet. Proponents of EM argue that rather than adding to the existing burden of students, including EM would provide medical students with a framework within which to organize facts (Nesse and Stearns, 2008). This aspect is not covered here; for more details on the place of EM in medical school, see Nesse and Schiffman (2003).

Antibiotic Management in Hospitals

Bacteria can evolve resistance to antibiotics very rapidly. After the initial therapeutic revolution brought about by antibacterial antibiotics, their intensive utilization and misuse led to a sharp increase in the frequency of resistant bacteria. We are now in the 'postantibiotic era,' in which the long time needed to develop new antibiotics may leave us short of therapeutic options. Bacteria resistance is particularly lethal in intensive care units, where the number of hospital-acquired (nosocomial) infections can be counted in the millions and the number of deaths in the thousands, not to mention the cost of prolonged stay in hospital. Here, an evolutionary approach can help understand and prevent the development of single- and multidrug resistance as well as investigate the reversibility of resistance. Using this approach to control the evolution of resistance generally involves the alteration of the environment to maintain the prevalence of sensitive strains. It relies on the effect of pleiotropy, when the expression of resistance to an antibiotic is also associated with a reproductive cost (for the bacteria) in the absence of the antibiotic, and fluctuating selection, where genetic variation is maintained (the presence of both sensitive and resistant strains) by rapidly changing the selective pressures, i.e., the antibiotic environments. As resistance to an antibiotic is most often associated with a cost in the absence of the antibiotic, removing the selective pressure allows sensitive strains to outcompete and displace resistant strains.

To reduce the rate of evolution of multidrug resistant bacteria, the first initiative has been to limit the total amount of antibiotic used in hospitals; however, medical, social, and economic forces have prevented the full use of this opportunity. Consequently, strategies making use of fluctuating selection, i.e., increasing the heterogeneity of antibiotics to limit the potential for natural selection to favor resistant strains, have been proposed. These are mainly antibiotic cycling, where the antibiotic is periodically replaced but all individuals within a ward are exposed to the same antibiotic at a given time, and antibiotic mixing, where the antibiotic is periodically replaced but where all individuals within a ward are exposed to different antibiotics at a given time (see **Figure 1** for an illustration). Empirical evidence for the efficiency of cycling strategies is controversial and most models have rejected it in favor of mixing strategies (Bergstrom et al., 2004) as the rate at which bacteria encounter new antibiotics is faster (i.e., increases heterogeneity), thereby reducing the amount of time during which resistance can rise to a high frequency (**Figure 1**). Mixing strategies, however, present the disadvantage of allowing horizontal gene transfer between bacteria that are resistant to different antibiotics and might therefore facilitate multiple drug resistance. Consequently, cycling strategies have seen a resurge of interest, with theoretical work showing that multiple resistance is less likely to occur if the antibiotic associated with the higher

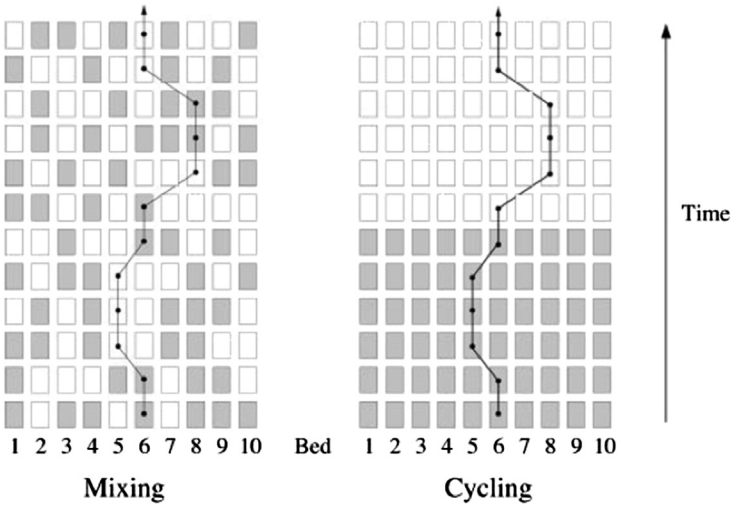

Figure 1 Effects of cycling and mixing on the selective conditions faced by a bacterial clone. Cycling offers greater heterogeneity at the level of the ward, but mixing offers greater heterogeneity at the level of the individual patient. Taken from Bergstrom, C.T., Lo, M., Lipsitch, M., 2004. Ecological theory suggests that antimicrobial cycling will not reduce antimicrobial resistance in hospitals. Proceedings of the National Academy of Sciences of the United States of America 101(36), 13285–13290. http://dx.doi.org/10.1073/pnas.0402298101.

resistance cost is used first, as the higher the cost the quicker the decrease in the frequency of resistant bacteria when the second antibiotic is used (Goulart et al., 2013).

The displacement of resistance remains challenging. It can take years due to phenomena such as compensatory evolution (the cost has been removed) and coselection (mutation conferring resistance is maintained through the selection of another portion of DNA to which it is linked), and it may not be practical given the time window available to the patient (Andersson and Hughes, 2010). In addition, the cost of resistance to an antibiotic varies among strains and from location to location, and therefore phylogeography of the pathogen would be required in designing strategies. Finally, the quantitative predictions of the models are dependent on the assumptions regarding the type of antibiotics used as well as their similarities. Overall, the very complex ecology of bacteria and their great ability to develop resistance, combined with the difficulty for models to make realistic predictions and the relatively small scope for experimentation in hospitals, together imply that we are not there yet, but evolutionary reasoning appears to be central if we are to make any step forward.

Cancer Therapy

Cancer was first described as an evolutionary process by Nowell in 1976. Most cancers are ecosystems of evolving clones (a clone is a group of cells related to each other by descent from unitary origin) that are escaping replication controls and DNA repair systems. The evolution of cancer cells is characterized by a sequence of clonal expansion, genetic diversification, and clonal selection. Most cancers are monoclonal in origin, but the expansion of population size and constant acquisition of new mutations lead to an increased genetic and phenotypic heterogeneity within a tumor (subclones). That not all mutations occur in the same cells explains partial tumor response to therapy and the emergence of drug-resistant cells. In this context, chemotherapy favors the development of metastatic cancers in two ways. First, it favors chromosomal instability thereby increasing the frequency of mutations, some of which potentially lead to resistance. Second, chemotherapy acts as a selective pressure against sensitive clones (decreasing their frequency) and for resistant ones (increasing their frequency). In other words, most single-target therapy will lead to the selective survival of a drug-resistant clone, which will expand and become dominant.

The opportunities for EM to contribute to cancer therapy reside in the exploitation of the evolutionary dynamics of cancers. While the oncologist can predict the future (which new selective pressure will be used), the cancer cannot: natural selection does not favor resistance for a chemotherapy that has never been experienced. Several potential avenues have been proposed (reviewed in Gerlinger and Swanton, 2010). First, assessing changes in clonal composition over time can tell us about which clone is resistant and which clone is sensitive to intervention. Most current-generation diagnostic genomic tests for cancer only inform about the presence/absence of cancer but do not inform about the proportion of each clone. This limits our ability to readjust therapy before progression occurs, or to track clonal bottlenecks (i.e., a decrease in clonal heterogeneity and thus in the potential to evolve) during the course of treatment by specifically targeting resistant clones. While assessing change in clonal composition is technically feasible through the advent of next-generation sequencing and new noninvasive techniques measuring heterogeneity in circulating tumor DNA in the plasma, the time and cost involved with this approach render it still not practical at the public health scale. Second, theoretical models and *in vivo* experiments propose a move away from a 'treatment for cure' strategy (generally a high-dose treatment aiming at killing as many cancer cells as possible) to a 'treatment for stability' strategy, a treatment that aims to prevent the elimination of sensitive clones. According to models and *in vivo* experiments, this latter approach would provide patients with a better life

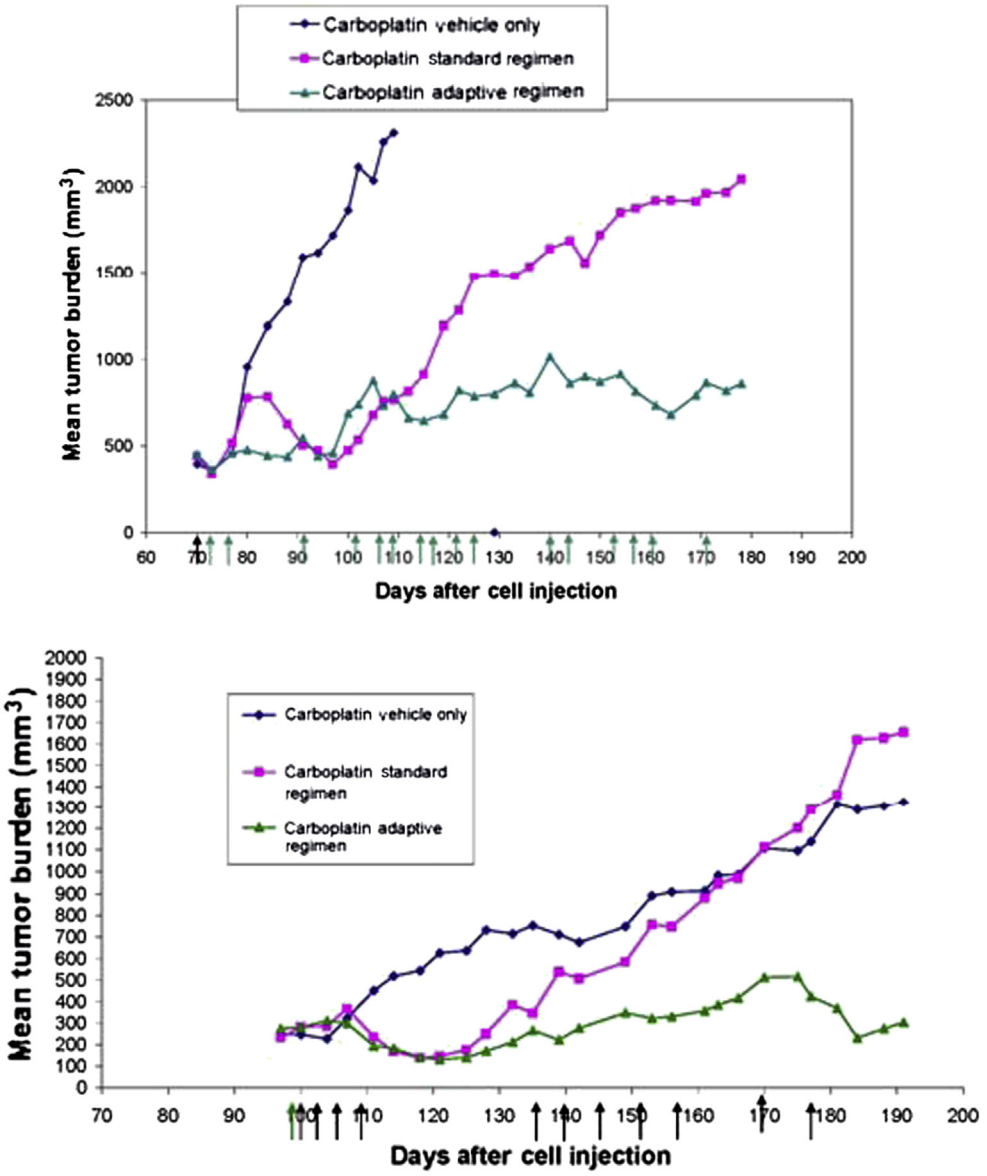

Figure 2 Adaptive versus standard therapies for mean tumor burden. To test the results of the models, experimental testing was conducted with mice inoculated with ovarian cancer cells. The two panels represent two different experiments. Each experiment included four animals in three experimental arms: (a) control (vehicle only); (b) 'standard' high-dose therapy consisting of 60 mg kg^{-1} each 4 days for three doses; (c) adaptive therapy that begins with a dose of 50 mg kg^{-1} and then adjusts the dose to maintain a stable tumor volume. The arrows on the x-axis represent days in which therapy was given in the adaptive group. The *y*-axis is the mean tumor volume for the four animals in each experimental group, and the *x*-axis is the time from inoculation of 10^7 tumor cells. Adaptive therapy can maintain small tumor and animals can survive indefinitely; Standard therapy is followed by substantial effect but it is followed by the recurrence of the tumor and the death of the animals. Adapted from Gatenby, R.A., Silva, A.S., Gillies, R.J., Frieden, B.R., 2009. Adaptive therapy. Cancer Research 69(11), 4894–4903.

expectancy as compared to high-dose chemotherapy (**Figure 2**; Gatenby et al., 2009). Third, the use of evolutionary tools such as phylogenies offer promising avenues for the early diagnosis of slowly evolving cancer at a stage where it is still, potentially, curable (Yachida et al., 2010). For instance, in the case of pancreatic cancer, a quantitative analysis of its genetic evolution shows that at least 10 years elapse between the initiating mutation and the birth of the nonmetastatic founder cell, and five more years are required to reach the metastatic state, after which patients are being diagnosed but the cancer is no longer curable (Yachida et al., 2010). These tools combined with advanced imaging and blood tests can offer an opportunity for early detection. Finally, it has also been suggested that drugs that promote genetic instability would be lethal for cancer (Gerlinger and Swanton, 2010). While phenotypic heterogeneity within clones is the necessary condition for the evolution of resistance, too much instability has costs for cancer in that it would produce too many deleterious mutations. More generally, all this research suggests that we must move away from the prevailing

monoclonal model and aggressive chemotherapy to develop strategies that take advantage of clonal heterogeneity (Gerlinger and Swanton, 2010).

Helminths (Parasitic Worms) and Crohn's Disease

Diseases resulting from an acute and mistaken response of the immune system such as allergies, Crohn's disease, and inflammatory bowel diseases are common in industrialized countries but rare in the developing regions of the world. According to the hygiene or 'old friends' hypothesis, these new diseases result from a mismatch between the environment in which the immune system has evolved and the modern urban environment where, following improvements in hygiene practices, helminthic infestation has dramatically reduced (Rook, 2012). However, the immune system has coevolved with parasitic worms so that the presence of helminths actually decreases the immune response. Because helminthic infestation offers protection from immunological diseases, poor helminthic infestation is believed to be the missing immunoregulatory factor that has led to a recent increase in autoimmune diseases (Rook, 2012).

There have been a number of studies looking at the role of helminthic therapy for a number of conditions including asthma, allergy, cancer, celiac disease, diabetes, and inflammation. Perhaps the most striking study to date is one that investigated the response of patients affected by Crohn's disease to helminthic therapy. The study, involving 29 patients, found that remission was increased by 72% after 24 weeks of treatment consisting of ingesting ova of *Trichuris suis*, an intestinal parasitic worm (**Figure 3**; Summers et al., 2005). Similar effects have been found for colitis, and no secondary effects of helminthic therapy have been observed to date. Since then, other worms have been identified, especially hookworm, which, if associated with anemia in poor regions, alleviate the symptoms of Crohn's disease in rich regions. Some have cautioned against the view that helminth (non)infection is the cause of all inflammatory diseases, particularly those in infants and those that present multiethnic prevalence (Reddy and Fried, 2009) as well as against the risk of multi-infections or coinfections. Yet, while more research on a larger scale is warranted to replicate those results, helminthic therapy offers exciting new avenues to treat inflammatory diseases.

Emergency Medicine and Intensive Care

It has been proposed that multiorgan failure observed in critically ill patients is, rather than a failure per se, a protective mechanism that would enhance the chances of survival of cells and organs in the face of an 'overwhelming insult' (Singer et al., 2004). Failure was previously envisaged as an abnormality resulting from an inflammatory response during an initial phase and followed by tissue hypoxia, cell damage, and organ dysfunction in a later phase. However, several observations challenge this view, among which are the facts that multiorgan failure is not associated with an increased necrosis of organs and that intervention on the critically ill such as blood transfusion and ventilation have been associated with adverse outcomes. It is increasingly being recognized that rather than hypoxia per se, the phenomenon corresponds to a reduced use of oxygen by cells which enter a dormant state in the face of reduced oxygen delivery to enhance the prospect of cell survival. A few studies suggest that interventions aiming at increasing/restoring the amount of oxygen in critically ill patients through ventilation or blood transfusion might actually be more damaging than restorative (see for review Bentley & Aunger, in Elton and O'Higgins, 2008). Viewing multiorgan failure as an adaptation rather than an abnormality might have important implications, as interventions might be harmful if they interfere with a response that is primarily protective. While a promising path, more studies directly assessing the adverse impact of routine interventions in palliative care are needed.

Birth and Parenting

Since EM has been launched (Nesse and Williams, 1996) as a complementary approach to mainstream medicine, one

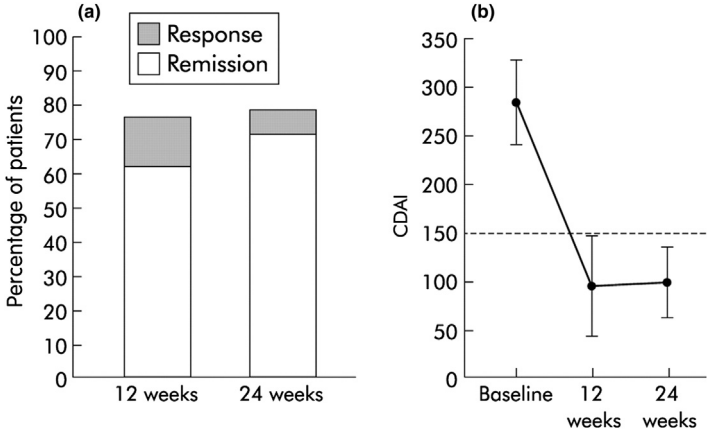

Figure 3 Crohn's disease and helminthic therapy. (a) Percentage of patients achieving remission or response at week 12 or 24 after initiating ova therapy. (b) Mean change in Crohn's disease activity index (CDAI, mean (SD)) for respondents to ova therapy. CDAI <150 is remission. $p < .0001$, week 12 or week 24 compared with baseline (time 0). Taken from Summers, R.W., Elliott, D.E., Urban, J.F., Thompson, R., Weinstock, J.V., 2005. Trichuris suis therapy in Crohn's disease. Gut 54(1), 87–90.

core interest has been to question practices in pregnancy, childbirth, and mother–infant relationships. First, EM questions the general practice of supplementing pregnant women with iron. Iron deficiency during pregnancy is thought to have evolved in response to pathogenic pressure (Nesse and Williams, 1996). Indeed, nearly all bacteria, fungi, and protozoa depend on a continuous supply of host iron for their growth and hosts have evolved a variety of defense mechanisms to preclude pathogens from acquiring essential material (Weinberg, 2009). It has thus been argued that low iron levels may protect against acute infection in young children (Nesse and Williams, 1996). Similar mechanisms exist in other species and, for instance, the iron-binding protein ovotransferrin in an egg white protects the yolk from most infections (Weinberg, 2009). If iron supplementation has potentially little effect in developed countries where parasitic load is low, such intervention might be more problematic in developing countries. Second, evolutionary studies have questioned the rationale underlying the treatment of jaundice in the preterm infant (Hammerman et al., 1998). It has been suggested that serum bilirubin, rather than being a toxic metabolic waste, is a protective antioxidant defense mechanism. This is especially relevant for preterm neonates for whom the concentration of antioxidant enzymes is reduced. Supporting this argument, a study involving 41 premature neonates revealed a relatively high correlation between serum bilirubin in premature and antioxidant potential during the first week of life (Hammerman et al., 1998; **Figure 4**). Third, EM has questioned mother–infant sleeping patterns at hospitals and the extent to which early sleeping arrangements influence long-term breast-feeding practices (Ball in Elton and O'Higgins, 2008). Evidence from comparative analyses suggests that the development of an early close physical relationship is critical for infant survival and normal development (Ball in Elton and O'Higgins, 2008). In particular, the onset, efficiency, and duration of breast-feeding are not independent from skin-to-skin contact in the immediate postnatal period. For instance, the amount of milk produced in the early days after birth is directly related to the intensity of nipple stimulation. Researchers from the United Kingdom used a randomized control trial enrolling >100 women to compare mother–infant interactions depending on whether the infant was sleeping in the same bed or beside it in a cot or a crib. Feeding frequency was found to be five times higher in the bed as compared to the cot condition and in addition, they observed a 20% difference in the proportion of mothers still breast-feeding their babies after 2 weeks (Ball in Elton and O'Higgins, 2008). Although more research has to be carried out, the study makes the case for rethinking current postnatal ward practices.

Public Health and Health Behavior Change

Standard public health campaigns aimed at preventing harmful behavior such as sun bathing or smoking usually focus on education and knowledge of the health consequences of behavior. The premises of those interventions are that unhealthy, irrational behavior results from a lack of information. However, the impact of such campaigns does not appear to be sufficient to lead to significant behavioral change. More recently, approaches from the field of evolutionary psychology have led to promising results. The rationale is that psychology is particularly sensitive to short-term social benefits as opposed to long-term health risks. Social motives and the desire to appear more attractive are indeed related to a number of health-related problems such as skin cancer, substance abuse, and eating disorders. The insight from an evolutionary approach here resides in the consideration that natural selection favors any heritable (psychological) behavior that increases reproduction over long-term health. Using this perspective, a study recruiting 147 college students followed change in the intentions for sun tanning after receiving a workbook describing the appearance-damaging effect of sun tanning. After 'treatment,' tendency to sun tan indoors was reduced by over 50%, a self-reported change that was still present 2 months after the intervention (**Figure 5**; Hillhouse and Turrisi, 2002). Although the study had some limitations (self-report, only 2 months follow-up), its impact on behavior is far more significant than that of any other interventions. Sun tanning is marketed as being glamorous and increasing attractiveness yet it greatly increases the risk of skin

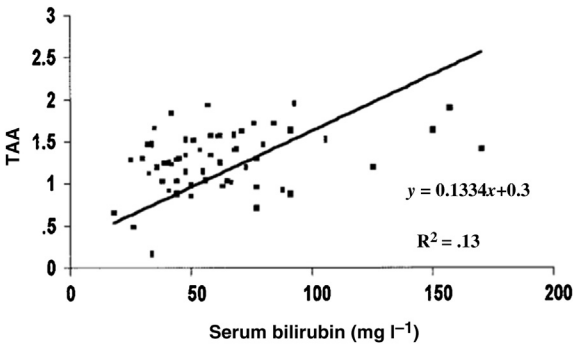

Figure 4 Correlation between serum bilirubin in premature neonates and antioxidant potential during the first week of life. Adapted from Hammerman, C., Goldstein, R., Kaplan, M., Eran, M., Goldschmidt, D., Eidelman, A.I., Gartner, L.M., 1998. Bilirubin in the premature: toxic waste or natural defense? Clinical Chemistry 44(12), 2551–2553.

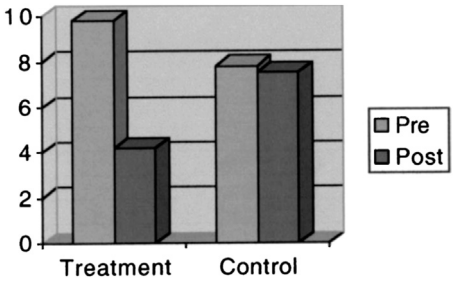

Figure 5 An appearance-focused intervention on sun tanning behavior. The graph depicts changes in indoor tanning behavior (e.g., number of indoor tanning visits reported for the previous 2 months) as a function of treatment group preintervention versus 2-month follow-up. Treatment consists in receiving a workbook describing the appearance-damaging effect of sun tanning. Adapted from Hillhouse, J.J., Turrisi, R., 2002. Examination of the efficacy of an appearance-focused intervention to reduce UV exposure. Journal of Behavioral Medicine 25(4), 395–409.

cancer. To tackle the problem, an 'R UV UGLY' campaign has been introduced in 2011/12 by Cancer Research UK. We are yet to assess the impact of this appearance-based intervention on sun tanning behavior, but it is stimulating for other initiatives targeting behavior that are both social and health damaging. For instance, how about changing 'Smoking Kills' to 'Smoking makes you Ugly' on cigarette packaging?

Summary

EM is a young, promising, and burgeoning science that offers a number of concepts and methods to both rethink current medical practices and devise new diagnostic and therapeutic tools. EM is increasingly interacting with other scientific fields such as neuroscience, genetics, physiology, psychiatry, and developmental studies, but its role in everyday practice remains marginal and it is not yet included in the medical curriculum of most universities. Indeed, EM does not offer ready-made answers to health professionals and patients. Its main insight is in challenging the way we think about diseases, pointing out that treating symptoms without an understanding of their ultimate causes might no longer be satisfactory if we are to make significant progress in improving health care. EM still has a long way to go, however, as research remains more theoretical than applied and more advanced in some medical subfields (e.g., pediatrics) than others (e.g., cardiology). Only time will tell the full impact of this complementarity approach on medical practice and global health.

See also: Adaptation, Fitness, and Evolution; Comparative Method in Evolutionary Studies; Darwinism; Epigenetic Inheritance; Evolutionary Developmental Psychology; Microevolution.

Bibliography

Andersson, D.I., Hughes, D., 2010. Antibiotic resistance and its cost: is it possible to reverse resistance? Nature Reviews Microbiology 8 (4), 260–271 http://dx.doi.org/10.1038/nrmicro2319.

Antonovics, J., Abbate, J.L., Baker, C.H., Daley, D., Hood, M.E., Jenkins, C.E., Vondrasek, J., 2007. Evolution by any other name: antibiotic resistance and avoidance of the e-word. PLoS Biology 5 (2), e30. http://dx.doi.org/10.1371/journal.pbio.0050030.

Bergstrom, C.T., Lo, M., Lipsitch, M., 2004. Ecological theory suggests that antimicrobial cycling will not reduce antimicrobial resistance in hospitals. Proceedings of the National Academy of Sciences of the United States of America 101 (36), 13285–13290. http://dx.doi.org/10.1073/pnas.0402298101.

Cournoyea, M., 2013. Ancestral assumptions and the clinical uncertainty of evolutionary medicine. Perspectives in Biology and Medicine 56, 36–52.

Eaton, S.B., Pike, M.C., Short, R.V., Lee, N.C., Trussell, J., Hatcher, R.A., Hurtado, A.M., 1994. Women's reproductive cancers in evolutionary context. The Quarterly Review of Biology 69 (3), 353–367.

Elton, S., O'Higgins, P., 2008. Medicine and Evolution: Current Applications, Future Prospects. CRC Press, Boca Raton.

Ewald, P.W., 1980. Evolutionary biology and the treatment of signs and symptoms of infectious disease. Journal of Theoretical Biology 86 (1), 169–176.

Ewald, P.W., 1994. The Evolution of Infectious Diseases. Oxford University Press.

Foley, R., 1995. The adaptive legacy of human evolution: a search for the environment of evolutionary adaptedness. Evolutionary Anthropology: Issues, News, and Reviews 4, 194–203.

Franceschi, C., Bonafè, M., Valensin, S., Olivieri, F., De Luca, M., Ottaviani, E., De Benedictis, G., 2000. Inflamm-aging. An evolutionary perspective on immunosenescence. Annals of the New York Academy of Sciences 908, 244–254.

Gatenby, R.A., Silva, A.S., Gillies, R.J., Frieden, B.R., 2009. Adaptive therapy. Cancer Research 69 (11), 4894–4903.

Gerlinger, M., Swanton, C., 2010. How Darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. British Journal of Cancer 103 (8), 1139–1143. http://dx.doi.org/10.1038/sj.bjc.6605912.

Gluckman, P.D., Hanson, M.A., 2006. Evolution, development and timing of puberty. Trends in Endocrinology and Metabolism 17 (1), 7–12.

Goulart, C.P., Mahmudi, M., Crona, K.A., Jacobs, S.D., Kallmann, M., Hall, B.G., Barlow, M., 2013. Designing antibiotic cycling strategies by determining and understanding local adaptive landscapes. PloS One 8 (2), e56040. http://dx.doi.org/10.1371/journal.pone.0056040.

Gould, S.J., Lewontin, R.C., 1979. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proceedings of the Royal Society of London. Series B. Biological Sciences 205 (1161), 581–598.

Haig, D., 2002. Genomic Imprinting and Kinship. Rutgers University Press, London.

Haldane, J.B.S., 1941. New Paths in Genetics. Allen and Unwin, London, UK.

Hammerman, C., Goldstein, R., Kaplan, M., Eran, M., Goldschmidt, D., Eidelman, A.I., Gartner, L.M., 1998. Bilirubin in the premature: toxic waste or natural defense? Clinical Chemistry 44 (12), 2551–2553.

Hillhouse, J.J., Turrisi, R., 2002. Examination of the efficacy of an appearance-focused intervention to reduce UV exposure. Journal of Behavioral Medicine 25 (4), 395–409.

Jacob, F., 1977. Evolution and tinkering. Science 196 (4295), 1161–1166.

Medawar, P.B., 1952. An Unsolved Problem of Biology. H.K. Lewis and Co., Oxford, UK.

Morgan, E., 1994. The Scars of Evolution. Oxford University Press on Demand.

Morgan, L., 1993. Comments on Wiley's "adaptation and the biocultural paradigm in medical anthropology: a critical review". Medical Anthropology Quaterly 7 (2), 199–201.

Nesse, R.M., Schiffman, J.D., 2003. Evolutionary biology in the medical school curriculum. Bioscience 53 (6), 585–587.

Nesse, R.M., Stearns, S.C., 2008. The great opportunity: evolutionary applications to medicine and public health. Evolutionary Applications 1 (1), 28–48.

Nesse, R.M., Williams, G.C., 1996. Why We Get Sick: The New Science of Darwinian Medicine. Vintage.

Norn, M., 1997. Myopia among the inuit population of East Greenland. Longitudinal study 1950–1994. Acta Ophthalmologica Scandinavica 75 (6), 723–725.

Partridge, L., Gems, D., 2006. Beyond the evolutionary theory of ageing, from functional genomics to evo-gero. Trends in Ecology & Evolution 21 (6), 334–340. http://dx.doi.org/10.1016/j.tree.2006.02.008.

Reddy, A., Fried, B., 2009. An update on the use of helminths to treat Crohn's and other autoimmunune diseases. Parasitology Research 104 (2), 217–221.

Rook, G.A., 2012. Hygiene hypothesis and autoimmune diseases. Clinical Reviews in Allergy & Immunology 42 (1), 5–15.

Singer, M., 1996. Farewell to adaptationism: unnatural selection and the politics of biology. Medical Anthropology Quarterly 10 (4), 496–515.

Singer, M., De Santis, V., Vitale, D., Jeffcoate, W., 2004. Multiorgan failure is an adaptive, endocrine-mediated, metabolic response to overwhelming systemic inflammation. Lancet 364 (9433), 545–548. http://dx.doi.org/10.1016/S0140-6736(04)16815-3.

Stearns, S.C., 1992. The Evolution of Life Histories. Oxford University Press Oxford.

Stearns, S.C., 2012. Evolutionary medicine: its scope, interest and potential. Proceedings. Biological Sciences/the Royal Society 279 (1746), 4305–4321. http://dx.doi.org/10.1098/rspb.2012.1326.

Summers, R.W., Elliott, D.E., Urban, J.F., Thompson, R., Weinstock, J.V., 2005. Trichuris suis therapy in Crohn's disease. Gut 54 (1), 87–90.

Úbeda, F., Wilkins, J.F., 2008. Imprinted genes and human disease: an evolutionary perspective. In: Genomic Imprinting. Springer, pp. 101–115.

Van Valen, L., 1973. A new evolutionary law. Evolutionary Theory 1, 1–30.

Walker, R., Gurven, M., Hill, K., Migliano, A., Chagnon, N., De Souza, R., Yamauchi, T., 2006. Growth rates and life histories in twenty-two small-scale societies. American Journal of Human Biology 18, 295–311.

Weinberg, E.D., 2009. Iron availability and infection. Biochimica et Biophysica Acta 1790, 600–605.

Williams, G.C., 1957. Pleiotropy, natural selection, and the evolution of senescence. Evolution. International Journal of Organic Evolution 11 (4), 398–411.

Williams, G.C., Nesse, R.M., 1991. The dawn of Darwinian medicine. Quarterly Review of Biology 66 (1), 1–22.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Nowak, M.A., 2010. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature 467 (7319), 1114–1117.

Evolutionary Social Psychology

Douglas T Kenrick, Arizona State University, Tempe, AZ, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 7, pp. 5022–5024, © 2001, Elsevier Ltd.

Abstract

Evolutionary social psychology is a framework that relies on principles of evolutionary theory to bridge various subdisciplines of psychology with other social sciences. From an evolutionary perspective, regularities in human social behavior reflect adaptive mechanisms designed to help our ancestors survive and reproduce. A number of different social psychological topics have been studied from an evolutionary perspective, including mating choice, aggression, altruism, and social cognition. Methodological approaches have included cross-cultural and cross-species comparisons, but the framework provides a basis for a range of experimental studies. Evolutionary psychologists assume that research on any topic in social behavior could benefit from an evolutionary perspective, and that other areas of psychology, including personality, psychopathology, and cognitive neuroscience, are all linked by this perspective. At the broadest level, the behaviors of humans and other animals are linked by general principles such as inclusive fitness, sexual selection, and differential parental investment. The field has been controversial, challenging traditional social science views, and it has been the subject of several misconceptions based on confusion about genetic determinism.

In 1908, McDougall's *Social Psychology* adopted an evolutionary perspective, following Darwin's (1872) suggestion that human social behaviors (including smiles, sneers, and other emotional expressions) were based in dispositions that helped our ancestors survive and reproduce. Evolutionary theory is founded on the idea of natural selection – that characteristics increasing chances of survival and reproduction proliferate because their bearers have more successful offspring. Characteristics well-suited to particular environmental demands are called adaptations. Evolutionary social psychology involves a search for evidence of adaptive mechanisms linked to human interaction.

Although people vary among themselves, both within and between cultures, certain regularities recur in human social interactions the world over. Members of all human groups communicate using complex spoken language, and share emotional expressions such as smiles and sneers, for example. Some social behavior patterns are unique to humans, some we share with many other species – the strong bonds between mammalian mothers and offspring, for instance. Similarly, other animals with slow-developing young often have biparental care of offspring, and other animals that live in small groups tend to have dominance hierarchies.

General Principles Applying across Species

Evolutionary theorists search for principles applying across animal species. *Differential parental investment* refers to differences between parents in the amount of resources invested in offspring (Trivers, 1972). As either sex increases investment in offspring, they become more selective about choosing mates. Members of the other sex consequently compete to be chosen. In mammals, females have an initially high investment (due to internal gestation and nursing). Because females in most mammalian species contribute substantially to offspring, they are in a position to choose among males demonstrating relatively superior health, attractiveness, or dominance. Hence, mammalian males, compared to females, tend to fight more among themselves, to be physically larger, and to possess more ornaments such as antlers. *Sexual selection* occurs when features are transmitted to offspring because they confer advantages in attracting the opposite sex (as in peacocks' feathers), or in competing with the same sex (as in rams' horns).

A central principle is *inclusive fitness maximization* – the idea that animals are designed not simply to survive and produce their own offspring, but to increase the probability their genes make it into future generations. Inclusive fitness explains many instances of apparently self-sacrificing behavior in animals. Ground squirrels who spot predators are more likely to make alarm calls if neighbors are closely related (Sherman, 1981). By slightly increasing their own risk, humans and other animals can gain a net genetic benefit – saving other individuals with copies of their genes.

Application to Human Social Behavior

Evolutionary theorists mine these broad principles for insights about human behavior, and to make predictions about circumstances under which people will be more or less likely to manifest aggression, altruism, or courtship behaviors. For example, men are slightly larger than women and reach reproductive maturity at a later age. Women commit substantially fewer homicides or assaults than men, and male violence is predominantly directed at other males. Across animal species, these are hallmarks of differential parental investment and polygyny. Of 849 societies in the cross-cultural area files, polygyny (one male and multiple females) was either usual or occasional in 708, polyandry (one female and multiple males) was found in only four, and the remaining 137 were exclusively monogamous (Daly and Wilson, 1983). Further, the circumstances associated with these variations follow principles found in other species (Daly and Wilson,

1983; Crook and Crook, 1988). For example, polyandry often involves brothers (who share genetic interests) living in resource-poor environments. Steep polygyny (as in harems) is found in circumstances of relative abundance, steep social hierarchy, and occasional famine (so a female's offspring have a better chance of surviving if she chooses an already married man from a wealthy powerful family over a single poor man). Although polygyny was an option across societies, most marriages are monogamous, because only highly successful men attracted multiple wives. Hence, ancestral males invested most of their reproductive effort in the offspring of a small number of women (often one). As a consequence, human males are selective about long-term mates, and females will compete with one another for desirable males (Buss, 1999; Kenrick and Trost, 1993).

Cross-Cultural Data

Because all human groups involve some common demands, evolutionary theorists assume that people the world over, despite cultural and ecological differences, share common human characteristics. Cross-cultural data reveal some general patterns consistent with principles of differential parental investment and sexual selection (Kenrick and Trost, 1993). Adult males have killed one another many times more frequently than adult females in societies ranging from hunter-gatherers in remote jungles to urbanites in modern Asia and Europe, as far back as historical data are available (Daly and Wilson, 1988). Likewise, women across societies place more emphasis than men on potential mates' resources and status (Buss, 1999).

Other data on mate selection reflect a unique feature of human biology – females reach sexual maturity earlier than males, show a fertility peak in the twenties and early thirties, and then completely cease reproduction despite many remaining years of life. Consistent with this difference in reproductive lifespan, women are attracted to mates several years older than themselves, whereas men change relative preferences as they age. Young men are attracted to women their own age or older; older men prefer women much younger than themselves (Kenrick et al., 1996). These patterns are found around the world, with somewhat stronger sex differences in societies in which women begin reproducing earlier (Kenrick and Keefe, 1992).

Experimental Data

Evolutionary principles can be used to generate experimental predictions. Experiments are not designed to test basic assumptions about natural selection, but to use the general principles as heuristics. For example, one team of researchers used inclusive fitness theory to generate predictions about helpfulness towards individuals varying in genetic relatedness (Burnstein et al., 1994). Experimental scenarios described helping situations ranging from small errands to rescuing someone in a burning building. For everyday helping, norms of politeness buffered discrepancies based on kinship and reproductive potential (grandmother was more likely to be helped than younger sister, for example). In life or death situations, however, help was focused more on closely related individuals with high reproductive potential (younger sister compared to grandmother or cousin, for example). Another series of studies examined how sex differences in mate selection criteria affect perceptual processes related to judgments of oneself and potential mates (Gutierres et al., 1999; Kenrick et al., 1994). Women rated commitment to male partners lower after exposure to socially dominant, but not physically attractive, men, whereas men's commitment showed the opposite effect. Conversely, women's self-ratings of their mating desirability were diminished more by physically attractive than by socially dominant women, with men's self-ratings showing the opposite effect (diminished by dominant, but not physically attractive men).

History and Future Directions

McDougall's (1908) *Social Psychology* assumed human social 'instincts.' The position was much misunderstood and abandoned without an empirical legacy. Decades later, European ethologists began successfully applying evolutionary models to social behaviors of nonhuman animals. In 1975, E.O. Wilson made a controversial suggestion that 'sociobiology' (the evolutionary analysis of animal social behavior) would eventually encompass the social sciences. This stimulated social psychologists to consider empirical applications to their field (e.g., Cunningham, 1981). Part of the controversy about the approach stemmed from misunderstandings about genetic determinism. Human social behaviors vary across contexts and cultures, and social scientists assumed sociobiology implied rigid instincts insensitive to environments. Some protested that evolutionary hypotheses were untestable, or required researchers to study genetic biology or archaeology. Evolutionary theorists clarified that environmentally insensitive instincts based on deterministic genes are misconceptions, hardly relevant to invertebrates, much less to mammals (e.g., Alcock, 1998; Kenrick, 1995). The brains of complex animals contain not rigid programs, but environmentally sensitive psychological mechanisms (Buss and Kenrick, 1998; Kenrick et al., 1998; Tooby and Cosmides, 1992). Questions about testability were addressed with empirical data and competing evolutionary hypotheses (e.g., Graziano et al., 1997; Gangestad and Thornhill, 1997). The position remains controversial, but has moved towards mainstream empirical science (e.g., Buss, 1999; Crawford and Krebs, 1998; Simpson and Kenrick, 1997). Evolutionary psychologists assume the perspective could enrich research on any topic in social behavior, and that other areas of psychology including personality, psychopathology, and cognitive neuroscience are all linked by this perspective. Although the implications for mating behaviors and sex differences have been most thoroughly explored, evolutionary models have implications for understanding kinship and family relationships, intergroup relations, and social cognition. Evolutionary psychologists believe the Darwinian perspective is as relevant to the study of human beings as it has been to the study of all other living species.

See also: Gender Differences in Personality and Social Behavior; Social Psychology: Research Methods; Writing, Evolution of.

Bibliography

Alcock, J., 1998. Animal Behavior, sixth ed. Sinauer Associates, Sunderland, MA.

Burnstein, E., Crandall, C., Kitayama, S., 1994. Some neo-Darwinian decision rules for altruism: weighing cues for inclusive fitness as a function of the biological importance of the decision. Journal of Personality and Social Psychology 67, 773–789.

Buss, D.M., 1999. Evolutionary Psychology: The New Science of Mind. Allyn and Bacon, Boston, MA.

Buss, D.M., Kenrick, D.T., 1998. Evolutionary social psychology. In: Gilbert, D.T., Fiske, S.T., Lindzey, G. (Eds.), Handbook of Social Psychology, fourth ed. McGraw-Hill, New York, pp. 982–1026.

Cosmides, L., Tooby, J., 1992. Cognitive adaptations for social exchange. In: Barkow, J.H., Cosmides, L., Tooby, J. (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, New York.

Crawford, C., Krebs, D.L., 1998. Handbook of Evolutionary Psychology: Ideas, Issues, and Applications. Erlbaum, Mahwah, NJ.

Crook, J.H., Crook, S.J., 1988. Tibetan polyandry: problems of adaptation and fitness. In: Betzig, L., Borgerhoff-Mulder, M., Turke, P. (Eds.), Human Reproductive Behaviour: A Darwinian Perspective. Cambridge University Press, Cambridge, UK.

Cunningham, M.R., 1981. Sociobiology as a supplementary paradigm for social psychological research. In: Wheeler, L. (Ed.), Review of Personality & Social Psychology. Sage, Beverly Hills, CA.

Daly, M., Wilson, M., 1983. Sex, Evolution, and Behavior, second ed. Willard Grant Press, Boston.

Daly, M., Wilson, M., 1988. Homicide. de Gruyter, New York.

Daly, M., Salmon, C., Wilson, M., 1997. Kinship: the conceptual hole in psychological studies of social cognition and close relationships. In: Simpson, J.A., Kenrick, D.T. (Eds.), Evolutionary Social Psychology. Erlbaum, Mahwah, NJ, pp. 265–296.

Darwin, C., 1872. The Expression of Emotion in Man and Animals. Murray, London.

Gangestad, S.W., Thornhill, R., 1997. Human sexual selection and developmental stability. In: Simpson, J.A., Kenrick, D.T. (Eds.), Evolutionary Social Psychology. Erlbaum, Mahwah, NJ, pp. 169–195.

Graziano, W.G., Jensen-Campbell, L.A., Todd, M., Finch, J.F., 1997. Interpersonal attraction from an evolutionary perspective: reactions to dominant and prosocial men. In: Simpson, J.A., Kenrick, D.T. (Eds.), Evolutionary Social Psychology. Erlbaum, Mahwah, NJ, pp. 141–167.

Gutierres, S.E., Kenrick, D.T., Partch, J.J., 1999. Beauty, dominance, and the mating game: contrast effects in self-assessment reflect gender differences in mate selection. Personality & Social Psychology Bulletin 25, 1126–1134.

Kenrick, D.T., 1995. Evolutionary theory versus the confederacy of dunces. Psychological Inquiry 6, 56–61.

Kenrick, D.T., Gabrielidis, C., Keefe, R.C., Cornelius, J.S., 1996. Adolescents' age preferences for dating partners: support for an evolutionary model of life-history strategies. Child Development 67, 1499–1511.

Kenrick, D.T., Keefe, R.C., 1992. Age preferences in mates reflect sex differences in human reproductive strategies. Behavioral and Brain Sciences 15, 75–133.

Kenrick, D.T., Neuberg, S.L., Zierk, K.L., Krones, J.M., 1994. Evolution and social cognition: contrast effects as a function of sex, dominance, and physical attractiveness. Personality & Social Psychology Bulletin 20, 210–217.

Kenrick, D.T., Sadalla, E.K., Keefe, R.C., 1998. Evolutionary cognitive psychology. In: Crawford, C., Krebs, D. (Eds.), Evolution and Human Behavior: Ideas, Issues, and Applications. Erlbaum, Mahwah, NJ, pp. 485–514.

Kenrick, D.T., Trost, M.R., 1993. The evolutionary perspective. In: Beall, A., Sternberg, R.J. (Eds.), Perspectives on the Psychology of Gender. Guilford Press, New York.

McDougall, W., 1908. An Introduction to Social Psychology. John W. Luce & Co. Boston, MA.

Sherman, P.W., 1981. Kinship demography and Belding's ground squirrel nepotism. Behavioral Ecology and Sociobiology 8, 251–259.

Simpson, J.A., Kenrick, D.T. (Eds.), 1997. Evolutionary Social Psychology. Erlbaum, Mahwah, NJ.

Tooby, J., Cosmides, L., 1992. The psychological foundations of culture. In: Barkow, J.H., Cosmides, L., Tooby, J. (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, New York.

Trivers, R.L., 1972. Parental investment and sexual selection. In: Campbell, B. (Ed.), Sexual Selection and the Descent of Man 1871–1971. Aldine, Chicago, IL.

Wilson, E.O., 1975. Sociobiology: The New Synthesis. Belknap Press of Harvard University Press, Cambridge, MA.

Evolutionary Sociology

Rosemary L Hopcroft, University of North Carolina-Charlotte, Charlotte, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E. Di Nuoscio, volume 21, pp. 14286–14290, © 2001, Elsevier Ltd.

Abstract

Evolutionary sociology (alternatively known as biosociology) is a new area in sociology that covers a wide range of topics, from microsociological to macrosociological, with the unifying feature being an acknowledgment of the role of evolved human biology in social life and hence an incorporation of biological factors into sociological scholarship. The field focuses on how evolved human biology interacts with particular social environments to simultaneously both produce, and respond to, social institutions and structures. In this article, I review the six major strands of research within the evolutionary sociology area.

Introduction

Evolutionary sociology (alternatively known as biosociology) is an area that has emerged relatively recently in sociology, and its emergence has not been without controversy. However, in the last 10 years the number of publications in the area has increased dramatically, and an Evolution, Biology and Society Section of the American Sociological Association was first created in 2004. The name evolutionary sociology covers a wide range of topics, from microsociological to macrosociological, with the unifying feature being an acknowledgment of the role of biology in human social life and hence an incorporation of biological factors into sociological scholarship. The field focuses on how evolved human biology interacts with particular social environments to simultaneously both produce and respond to social institutions and structures. Researchers in the area use a variety of sociological methodologies, as well as research results and methodologies, from an array of disciplines including anthropology, behavioral genetics, history, primatology, paleoanthropology, biology, psychology, and neuroscience. What follows discusses the background to the area and reviews six major strands of research within the evolutionary sociology field: (1) emotions and social behavior, (2) neurosociology, (3) evolution and social behavior, (4) genes and social behavior, (5) hormones and social behavior, and (6) societal evolution.

Background to the Field

The 1970s controversy over sociobiology (see Segerstråle, 2000 for details) percolated into sociology in the late 1970s, with sociologists such as Pierre van den Berghe and Joseph Lopreato becoming vocal proponents (van den Berghe, 1974; Lopreato, 1984). However, sociobiology was resoundingly rejected by the vast majority of sociologists, and by the 1990s had largely disappeared from the field (Machalek and Martin, 2004). It took root in anthropology, where anthropologists such as Napoleon Chagnon, Sarah Hrdy, David Barash, Monique Borgerhoff Mulder, Laura Betzig, Richard Alexander, William Irons, and others became leading proponents. In the late 1980s a revised form of sociobiology known as evolutionary psychology was born, with early leading proponents including the psychologist Leda Cosmides and the anthropologist John Tooby (Cosmides and Tooby, 1987). Evolutionary psychology took off as a subfield within psychology, and this change and the formation of the Human Behavior and Evolution Society invigorated the nascent evolutionary anthropology group. By the turn of the century, evolutionary psychology was well ensconced within the discipline of psychology, and an evolutionary anthropology section of the American Anthropological Association was formed. At about this time, the Evolution, Biology and Society Section of the American Sociological Association became a section in formation and a home for evolutionary sociologists. At first the section was called Evolution and Sociology. Later, as scholars studying the interaction between genes and social behavior joined the section, the name was changed to Evolution, Biology & Society.

Evolutionary sociology, like sociobiology before it, remains controversial in sociology. Many of the same arguments originally made against sociobiology are leveled against evolutionary sociology. An early critique of sociobiology is that in the past biology has been used to support the prejudices of the researcher and/or to support the existing social status quo (see Gould, 1981 for a review of this problematic work). For example, biology was used by social Darwinists in the early twentieth century to justify and explain existing social inequalities between groups. It should be noted that current evolutionary sociologists distance themselves from the prejudices, faulty reasoning, and methods that characterized much of this earlier work. The evolutionary view is that all individuals share the same universal biology and evolutionary legacy. This is buttressed by the finding that there is more genetic variation within groups than between groups (Witherspoon et al., 2007).

Other criticisms include the charge that the evolutionary or biosocial approach is reductionist (Rose, 2000) and unnecessary (Heuveline, 2004). Yet many sociologists acknowledge that there is a need for models of individual behavior (hence reductionism) in sociology (Crippen, 1994; Whitmeyer, 1994; Kiser, 1996; Massey, 2002, 2005; Lovett, 2006). Others claim that evolutionary/biosocial approaches embrace genetic determinism, or that scholars in the area fail to give evidence in support of their claims. Much of this is based on a misunderstanding of evolutionary approaches (Machalek and

Martin, 2004) and a lack of knowledge of research findings in the area, particularly in neighboring disciplines to sociology such as anthropology, psychology, and biology.

Yet the tide may be turning. Freese et al. (2003) was a watershed piece, a well-cited mainstream sociological publication advocating the incorporation of biology into the discipline. Early biosociologists such as Booth et al. (2000) had previously written a review of research on biosocial influences on the family, a follow-up was published by D'Onofrio and Lahey (2010). The introduction of biosociology into the mainstream was fully established by the publication of special issues on biosociology in two of the three major journals in the field: *Social Forces* in 2006 and the *American Journal of Sociology* in 2008. Guang Guo (2006) edited the special issue of *Social Forces*, and Peter Bearman (2008) edited the special issue of the *American Journal of Sociology*.

It is notable that many of these authors included in these special issues have particularly been proponents of examining genetic influences on behavior and have therefore largely sidestepped evolutionary questions. Yet genes, as are all biochemicals and in fact everything in biology, are products of evolution (Dobzhansky, 1973). Genes and social behavior are just one subarea within evolutionary sociology as it currently exists within sociology. The other subareas are emotions and social behavior, neurosociology, evolution and social behavior, hormones and social behavior, and societal evolution.

Emotions and Social Behavior

Jonathan Turner was an early proponent of evolutionary sociology, and much of his work has dealt with the sociology of emotion. He and coauthors have argued for the centrality of emotions in human actions, and they make the case for the grounding of human emotions in biology (Turner, 2007; Turner and Maryanski, 2005, 2008; Turner and Stets, 2006). Douglas Massey (2002) in his presidential address at the meetings of the American Sociological Association suggested that the evolution of human society and human cognition illustrates the creation and workings of the human emotional brain and the fact that it operates independently of, but strongly influences, the rational brain.

Turner and Maryanski (2008) in a work based on information from primatology, behavioral ecology, and paleoanthropology draw a picture of the evolved human that emerged some 150 000 years ago and has changed little since: an individualistic hominid, emotionally tied to group members, but resistant to domination by other hominids. They note that this is the human nature that cultures have built on but not eradicated.

Besides work on how individual emotions influence social relations, there are other scholars within this group who examine the reverse: how social relations and situations in turn influence emotions (Stets and Asencio, 2008; TenHouten, 2005). Robinson et al. (2004) discuss how experience in group interactions of being high and low status influences emotional and physiological states. Much of the research shows that occupying a low status position is more stressful than occupying a high status position. Similarly, Massey (2004) theorizes that exposure to stress leads to a variety of deleterious health and cognitive outcomes for African Americans (see also Davis and Were, 2008).

Neurosociology

Authors in the subarea of neurosociology argue for the linking of social psychology and social neuroscience. This is another emerging area, spearheaded by David Franks and Thomas Smith (Franks and Smith, 2009; Smith, 2004). A didactic session on the area was held at the Annual Meetings of the American Sociological Association in 2010.

David Franks and others (Franks, 2010; Hammond, 2004) draw on the neuroscience of emotion and cognition. Their research draws on findings from research using fMRI scanners and PET scans to examine which parts of the brain are used for cognition and emotions. They examine the role of emotions in social behavior, unconscious behavior, the role of mirror neurons in imitation and empathy, and the brain bases of the social self. Findings include the importance of normal social interactions for the development of the individual brain and the development of the self. Another finding is the fact that the large majority of what the brain does, perhaps as much as 80%, is unconscious to the individual. Last, contrary to many beliefs, emotion is the basis of rational decision making. For people to make sensible decisions, the available choices must have an emotional valence. When that emotional valence is missing, for whatever reason, the person cannot make rational decisions.

Franks (2010) notes that research on mirror neurons support the work of the Chicago pragmatists. Mirror neurons are neurons in the brain that mimic the neurons of others. So when a person falls and feels pain, neurons in his or her brain fire, and the same neurons in the brain of the person witnessing the event also fire. This means that to a certain extent the witness actually feels the fallen person's pain. As a result, mirror neurons are importantly involved in empathy. They are also involved in helping individuals see how others see them, and this is an important component of the development of self-identity and self-esteem. This is the mechanism for what Charles Horton Cooley referred to as the "looking glass self" (Franks, 2010). Other neurosociologists show how other social processes are based on the existing neural wiring of the human brain. For example, Hammond (2004) shows how other social experiences such as religious experiences piggyback on neural circuits evolved for different purposes.

Evolution and Social Behavior

This is probably the least well-accepted subtopic within evolutionary sociology in the mainstream of sociology. Pierre van den Berghe was one of the first sociologists working in the area, and he found little audience within sociology (van den Berghe, 1979, 1990). More recently there has been much greater acceptance of the evolutionary approach to human social behavior in sociology both in the United States and abroad (Whitmeyer, 1994, 1997; Lopreato and Crippen, 1999; Machalek and Martin, 2004; Sanderson, 2001, 2008; Runciman, 2008; Lacerda, 2009; Hopcroft, 2008, 2009 a,b). These authors all argue for a linkage of theory and research in sociology with theory from

evolutionary biology. Many of them deal with resistance among sociologists to this kind of thinking, as well as erroneous ideas about the nature of the area.

The first evolutionary approach to the family within sociology was the book *Human Family Systems* (van den Berghe, 1979), a complete review of human family systems in all known human societies and how they can be explained by an evolutionary perspective. van den Berghe notes the centrality of kin processes in all societies, as well as commonalities in mate choice and parenting across societies. For example, in all human societies women marry men who are older than they are, on average, although the age gap varies across cultures. In all human societies women assume the majority of child care. Since then, prominent work in the area has been done by family demographers (Biblarz and Raftery, 1999) who found the counterintuitive finding that single mothers are better at sponsoring the attainment of their children than reconstituted families.

Work by Hopcroft (2006), Fieder and Huber (2007), and Nettle and Pollet (2008) has refined our understanding of the relationship between status and fertility, showing that income earned by men is positively related to their fertility in advanced industrial societies, while income earned by women is negatively related to their fertility. There has also been research testing evolutionary hypotheses concerning investment in children. Hopcroft (2005) tested the Trivers–Willard hypothesis that high-status parents would invest more in sons, and low-status parents would invest more in daughters, and found support for the hypothesis – finding that sons of high-status parents received more years of education than daughters, and daughters of low-status parents received more years of education than sons. Freese and Powell (1999) tested the hypothesis by examining parental investments in adolescents and found no support for it. Hamilton et al. (2007) tested the evolutionary hypothesis that biological children would receive more investment than nonbiological children. They found this was not true in the case of adoptive children, although others have found it is true of stepchildren (Henretta et al., 2012).

Evolutionary biology posits sex differences, particularly in parenting behavior, with women being the more committed parents (Rossi, 1984). This has implications for many social phenomena, including inequality between the sexes (Huber, 2007; Hopcroft, 2009a). There are other sex differences that likely have an evolved, biological component, for example, the sex difference in depression (Hopcroft and Bradley, 2007). Ellis et al. (2008) have published a comprehensive list of sex differences, many of which likely have an evolutionary origin.

The potential contributions of evolutionary theory to sociological theory have been explored in different ways by a number of authors. Blute (2006, 2010) examine the possibility of using models from evolutionary biology to model cultural evolution (see also Mark, 2011). Sanderson has argued that materialist theories need help from evolutionary theories (see Sanderson, 2001, 2007; see also Horne, 2004). Hopcroft (2008, 2009b) has suggested that evolutionary theory can help create a unifying paradigm for sociology. In Europe a prominent proponent of the use of evolutionary theory in sociology is the Finnish sociologist J.P. Roos (see Roos and Rotkirch, 2005).

There is much interest in the evolutionary origins of religion in neighboring disciplines of anthropology and psychology. Sociologist Steven K. Sanderson (2008) has recently joined this enterprise. He suggests that the universal existence of religion across all human societies is because during the human evolutionary period religiosity was an adaptive trait. Although not specifically evolutionary, Miller and Stark (2002) suggest that gender differences in risk aversion (possibly an evolved trait) play a role in the sex difference in religiosity. Collett and Lizardo (2009) debate this point. Regarding crime, Ellis notes that universally young men are more involved in crime than any other group (Ellis, 2004, see also Savage and Vila, 2003). Ellis argues that there are evolutionary reasons to explain this. Last, one of the pioneers of the evolution and social behavior area, Pierre van den Berghe, has also explored the evolutionary bases of ethnic behavior (van den Berghe, 1981). A follow-up to this work was published by a student (Whitmeyer, 1997).

Genes and Social Behavior

Scholars in this area argue for an incorporation of genetic effects on behavior in sociological research. Shanahan et al. (2010) urges sociologists to join the research effort to examine the joint effects of genes and environmental contexts on social behavior. Conley (2009) and Guo et al. (2009) describe methods for including genetic information into social science research. Much of this research relies on the collection of DNA data as well as the more typical survey data in sociological data sets. An important data source is the National Longitudinal Study of Adolescent Health (Add Health), a study designed by Peter Bearman and Richard Udry, both at the time members of the Department of Sociology at the University of North Carolina at Chapel Hill.

This research has examined the relative influence of social contexts and genes on substance abuse (Button et al., 2009), drinking behavior (Pescosolido et al., 2008), and smoking (Boardman et al., 2010). Other research examines the role of biological potential and environment on income attainment (Conley and Bennett, 2001), educational attainment (Shanahan et al., 2008), and intellectual development (Guo and Stearns, 2002).

There are also a variety of studies that examine the interaction of genes and social contexts on physical and mental health (Adkins et al., 2008; Pescosolido et al., 2008; Kendler et al., 2011). Others have examined the relative effects of genes on happiness (Schnittker, 2008) and attitudes (Eaves et al., 2008). Using Add Health data, criminologist Kevin Beaver (Beaver, 2008; Beaver et al., 2008; Vaughn et al., 2009), as well as sociologist Guang Guo (Guo et al., 2008a), have examined the interaction between genes and environment on criminal and delinquent behavior.

Other researchers look at the gene–environment interactions for outcomes such as same-sex attraction (Bearman and Brückner, 2002), number of sexual partners (Guo et al., 2008b), age at first intercourse (Guo and Tong, 2006), and fertility behavior (Kohler et al., 1999, 2002). Powell et al. (2010) examine Americans' beliefs about the role of genes in same sex attraction and how this affects their opinions about same sex couples.

All these studies typically find both a substantial genetic effect and a gene–environment interaction on the social outcome being studied. One general finding that emerges

from a number of these studies is that genetic influence on behavior is stronger during affluent time periods when individuals (presumably) have more choice. This has been capitalized on by Nielsen (2006), who suggests that the degree to which an individual achieves their genetic potential can reflect the opportunities present in their society (see also Adkins and Vaisey, 2009). Although this generally rises with affluence, there is variation in such opportunities across developed societies.

Besides this work on gene–environment interactions, other scholars focus on epigenesis – how gene expression is regulated by the environment and how changes in gene expression can even be passed on to future generations (Lock, 2009; Niewohner, 2011; Singh, 2012; Spector, 2012). These studies stress the malleability of the organism and its development in response to environmental context.

Hormones and Social Behavior

Alan Booth has been a pioneer in sociology at including biomarkers such as testosterone in social science research in the family. His article with coauthors (Booth et al., 2006) both reviews research and advocates for continued research on testosterone's role in social behavior. These researchers use hormonal analysis of saliva samples and statistical methods. Findings show that many important social phenomena are a product of the interaction of social contexts with individual hormonal states (Mazur, 2004; Booth et al., 2008). Much of the research has been on the effects of testosterone. Of this research, most has been on the effects of testosterone in boys and men (Mazur and Booth, 1998; Dabbs and Dabbs, 2000; Updegraff et al., 2006). There have been comparatively fewer studies of the effects of testosterone on girls and women (Udry, 2000; Booth et al., 2005). Much of this research has been highly controversial within sociology, particularly when it pertains to sex differences in behavior (e.g., see Miller and Costello, 2001; Risman, 2001 for critiques of Udry, 2000).

Societal Evolution

The last subarea within the field of evolutionary sociology is a macrosociological area that examines large-scale societal change and the evolution of human societies. Gerhard Lenski, the pioneering scholar in this area, was the first to note the importance of subsistence technology in shaping social arrangements and the degree of social stratification (Lenski, 1984, 2005). Lenski's conclusions that a society's subsistence technology (hunting and gathering, horticulture, plow agriculture, or industrial technology) predicts their population (size), social organization, and ideology is well known and much used. The association of complexity of subsistence technology with complexity in occupational and political structures is also well known. Lenski (1984) finding that social inequality increases with increased complexity of subsistence technology until we reach industrial societies has also been widely recognized (Nielsen, 2004). Others note the implications of subsistence technology for gender stratification. While the amount of gender stratification in society generally tracks the degree of stratification in general, Blumberg (2004) suggests that the most important factor influencing gender stratification in a society is women's relative control of key economic resources. Women are more likely to have a higher status in societies where they contribute to the society's main productive activities, but women's work does not enhance gender equality unless it also results in control of significant economic resources.

Lenski and Nolan (2005) hypothesized that currently developing societies that were previously agrarian do better in terms of various measures of development than societies that were previously horticultural. The hypothesis is empirically strongly supported. Lenski's Ecological Evolutionary Theory (2005) also has implications for the failures of Marxist society. The failure of Marxist societies in Lenski's view is due to the failure of Marxism to fully understand human nature. Marxist societies ended with a concentration of power at the top, so Communist societies more closely resemble the monarchies and oligarchies in agrarian societies.

Unlike Lenski, who focuses on internal characteristics of societies for explaining societal evolution, other authors link the process of societal evolution primarily to the society's place in the modern world system and adopt a world system's theoretical perspective (e.g., Chase-Dunn, 2005). Thomas Hall and coauthors (Hall and Fenelon, 2009) have looked at the distinctiveness of indigenous movements and how they have responded to recent policy reforms and globalization. Findings from this research highlight the importance of the influence of global and historical forces on indigenous movements.

Conclusion

Evolutionary sociology is a fairly recent field within sociology but has become home to a group of scholars working in a variety of subareas. Of these, the most prominent are the scholars that study the interaction of individual genetic potential with social environments on social outcomes. As more information on the role of biology in social phenomena emerges, the discipline as a whole is likely to lose its biophobia and begin to more fully embrace work that examines the evolutionary origins of human biology.

See also: Altruism and Prosocial Behavior, Sociology of; Cooperation: Sociological Aspects; Darwinism: Social; Emotion, Evolution of; Emotions, Sociology of; Evolutionary Approaches to Human Behavior: Philosophical Aspects; Evolutionary Demography: A Synthesis of Two Population Sciences; Family as Institution; Gene–Culture Coevolution; Genetic Studies of Behavior: Methodology; Genetics and Society; Kin Selection; Kinship, Evolution of; Lifespan Development: Evolutionary Perspectives; Macrosociology-Microsociology; Marriage Systems, Evolution of; Methodological Individualism in Sociology; Morality: Evolution of; Neurosociology; Religious Belief, Evolution of; Sexual Orientation: Genetic Aspects; Social Epigenetics; Sociobiology and Sociology: A New Synthesis; Sociobiology, History of; Trust, Sociology of.

Bibliography

Adkins, Daniel E., Vaisey, Stephen, 2009. Toward a unified stratification theory: structure, genome, and status across human societies. Sociological Theory 27 (2), 99–121.

Adkins, Daniel E., Wang, Victor, Elder Jr., Glen H., 2008. Stress processes and trajectories of depressive symptoms in early life: gendered development. Advances in Life Course Research 13, 107–136.

Bearman, Peter (Ed.), 2008. Exploring Genetics and Social Structure. American Journal of Sociology, 114.3 (S1), pp. v–x (special issue).

Bearman, Peter S., Brückner, Hannah, 2002. Opposite-sex twins and adolescent same-sex attraction. American Journal of Sociology 107 (5), 1179–1205.

Beaver, Kevin M., 2008. Nonshared environmental influences on adolescent delinquent involvement and adult criminal behavior. Criminology 46 (2), 341–369.

Beaver, Kevin M., Wright, John Paul, Delisi, Matt, Vaughn, Michael G., 2008. Desistance from delinquency: the marriage effect revisited and extended. Social Science Research 37 (3), 736–752.

Biblarz, Timothy J., Raftery, Adrian E., 1999. Family structure, educational attainment, and socioeconomic success: rethinking the 'pathology of matriarchy'. American Journal of Sociology 105 (2), 321–365.

Blumberg, Rae Lesser, 2004. Extending Lenski's schema to hold up both halves of the sky – a theory-guided way of conceptualizing Agrarian societies that illuminates a puzzle about gender stratification. Sociological Theory 22 (2:), 278–291.

Blute, Marion, 2006. Gene–culture coevolutionary games. Social Forces 85 (1), 145–149.

Blute, Marion, 2010. Darwinian Sociocultural Evolution: Evolutionary Solutions to Dilemmas in Cultural and Social Theory. Cambridge University Press.

Boardman, Jason D., Blalock, Casely L., Pampel, Fred C., 2010. Trends in the genetic influences on smoking. Journal of Health and Social Behavior 51 (1), 108–123.

Booth, Alan, Carver, Karen, Granger, Douglas A., 2000. Biosocial perspectives on the family. Journal of Marriage and Family 62 (4), 1018–1034.

Booth, Alan, Granger, Douglas A., Mazur, Allan, Kivlighan, Katie T., 2006. Testosterone and social behavior. Social Forces 85 (1), 167–191.

Booth, Alan, Johnson, D., Douglas Granger, D., 2005. Testosterone, marital quality, and role overload. Journal of Marriage and the Family 67, 483–498.

Booth, Alan,, Granger, Douglas A., Shirtcliff, Elizabeth A., 2008. Gender- and age-related differences in the association between social relationship quality and trait levels of salivary cortisol. Journal of Research on Adolescence 18 (2), 239–260.

Button, Tanya M.M., Stallings, Michael C., Rhee, Soo Hyun, Corley, Robin P., Boardman, Jason D., Hewitt, John K., 2009. Perceived peer delinquency and the genetic predisposition for substance dependence vulnerability. Drug and Alcohol Dependence 100 (1–2), 1–8.

Chase-Dunn, Christopher, 2005. Social evolution and the future of world society. Journal of World-Systems Research 11 (2), 171–192.

Collett, Jessica L., Lizardo, Omar, 2009. A power-control theory of gender and religiosity. Journal for the Scientific Study of Religion 48 (2), 213–231.

Conley, Dalton, 2009. The promise and challenges of incorporating genetic data into longitudinal social science surveys and research. Biodemography and Social Biology 55 (2), 238–251.

Conley, Dalton, Bennett, Neil G., 2001. Birth weight and income: interactions across generations. Journal of Health and Social Behavior 42 (4), 450–465.

Cosmides, Leda, Tooby, John, 1987. From evolution to behavior: evolutionary psychology as the missing link. In: Dupré, John (Ed.), The Latest on the Best: Essays on Evolution and Optimality. The MIT Press, Cambridge, MA, pp. 276–306.

Crippen, Timothy, 1994. Neo-Darwinian approaches in the social sciences: unwarranted concerns and misconceptions. Sociological Perspectives 37, 391–401.

Dabbs Jr., James M., Dabbs, Mary G., 2000. Heroes, Rogues, and Lovers: Testosterone and Behavior. McGraw-Hill, New York.

Davis, Jeff, Were, Daniel, 2008. A longitudinal study of the effects of uncertainty on reproductive behaviors. Human Nature 19 (4), 426–452.

Dobzhansky, Theodosius, March 1973. Nothing in biology makes sense except in the light of evolution. American Biology Teacher 35, 125–129.

D'Onofrio, Brian M., Lahey, Benjamin B., 2010. Biosocial influences on the family: a decade review. Journal of Marriage and Family 72 (3), 762–782.

Eaves, Lindon J., Hatemi, Peter K., Prom-Womley, Elizabeth C., Murrelle, Lenn, 2008. Social and genetic influences on adolescent religious attitudes and practices. Social Forces 86 (4), 1621–1646.

Ellis, Lee, 2004. Sex, status, and criminality: a theoretical nexus. Social Biology 51 (3–4), 144–165.

Ellis, Lee, Hershberger, Scott, Field, Evelyn, Wersinger, Scott, Pellis, Sergio, Geary, David, Palmer, Craig, Hoyenga, Katharine, Hetsroni, Amir, Karadi, Kazmer, 2008. Sex Differences: Summarizing More than a Century of Scientific Research. Psychology Press.

Fieder, Martin, Huber, Susanne, 2007. The effects of sex and childlessness on the association between status and reproductive output in modern society. Evolution and Human Behavior 28, 392–398.

Franks, David, 2010. Neurosociology: The Nexus between Neuroscience and Social Psychology. Springer, New York.

Franks, David, Smith, Thomas, 2009. A neurosociological perspective on emotions. A review article by David Franks and Thomas Smith: mind, brain and society: toward a neurosociology of emotion. Sociologie 5 (2), 244–256.

Freese, Jeremy, Li, Jui-Chung Allen, Wade, Lisa D., 2003. The potential relevance of biology to social inquiry. Annual Review of Sociology 29, 233–256.

Freese, Jeremy, Powell, Brian, 1999. Sociobiology, status, and parental investment in sons and daughters: testing the Trivers–Willard hypothesis. American Journal of Sociology 104 (6), 1704–1743.

Gould, Stephen Jay, 1981. The Mismeasure of Man. Norton, New York.

Guo, Guang (Ed.), 2006. The Linking of Sociology and Biology. Social Forces, 85.1, pp. 145–149 (special issue).

Guo, Guang, Roettger, Michael, Cai, Tianji, 2008a. The integration of genetic propensities into social control models of delinquency and violence among male youths. American Sociological Review 73 (4), 543–568.

Guo, Guang, Tong, Yuying, Cai, Tianji, 2008b. Gene by social–context interactions for number of sexual partners among white male youths: genetics-informed sociology. American Journal of Sociology 114, 36–66.

Guo, Guang, Stearns, Elizabeth, 2002. The social influences on the realization of genetic potential for intellectual development. Social Forces 80 (3), 881–910.

Guo, Guang, Tong, Yuying, 2006. Age at first sexual intercourse, genes, and social & demographic context: evidence from twins and the dopamine D4 receptor gene. Demography 43, 747–769.

Guo, Guang, Elder, Glen H., Cai, Tianji, Hamilton, Nathan, 2009. Gene–environment interactions: peers' alcohol use moderates genetic contribution to adolescent drinking behavior. Social Science Research 38 (1), 213–224.

Hall, Thomas D., Fenelon, James V., 2009. Indigenous Peoples and Globalization: Resistance and Revitalization.

Hamilton, Laura, Cheng, Simon, Powell, Brian, 2007. Adoptive parents, adaptive parents: evaluating the importance of biological ties for parental investment. American Sociological Review 72 (1), 95–116.

Hammond, Michael, 2004. The enhancement imperative and group dynamics in the emergence of religion and ascriptive inequality. Advances in Group Processes 21, 167–188.

Henretta, John C., Wolf, Douglas A., Van Voorhis, Matthew F., Soldo, Beth J., July 2012. Family structure and the reproduction of inequality: parents' contribution to children's college costs. Social Science Research 41 (4), 876–887.

Heuveline, P., 2004. Review of animal traditions: behavioural inheritance in evolution by Eytan Avital and Eva Jablonka. American Journal of Sociology 109, 1500–1506.

Hopcroft, Rosemary L., 2005. Parental status and differential investment in sons and daughters: Trivers–Willard revisited. Social Forces 83 (3), 169–193.

Hopcroft, Rosemary L., 2006. Sex, status and reproductive success in the contemporary U.S. Evolution and Human Behavior 27, 104–120.

Hopcroft, Rosemary L., 2008. Darwinian conflict theory: alternative theory or unifying paradigm for sociology? In: Niedenzu, Heinz-Jürgen, Meleghy, Tamás, Meyer, Peter (Eds.), The New Evolutionary Science: Human Nature, Social Behavior and Social Change. Paradigm Publishers, Colorado.

Hopcroft, Rosemary L., 2009a. Gender inequality in interaction: an evolutionary account. Social Forces 87 (4), 1845–1872.

Hopcroft, R.L., 2009b. The evolved actor in sociology. Sociological Theory 27 (4), 390–406.

Hopcroft, Rosemary L., Bradley, Dana Burr, 2007. The sex difference in depression across 29 countries. Social Forces 85 (4), 1483–1507.

Horne, Christine, 2004. Values and evolutionary psychology. Sociological Theory 22 (3), 477–503.

Huber, Joan, 2007. On the Origins of Gender Inequality. Paradigm Publishers, Boulder.

Kendler, Kenneth S., Jaffee, Sara, Romer, Dan (Eds.), 2011. The Dynamic Genome and Mental Health: The Role of Genomes and Environments in Youth Development. Oxford University Press.

Kiser, Edgar, 1996. The revival of narrative in historical sociology: what rational choice theory can contribute. Politics and Society 24, 249–271.

Kohler, Hans-Peter, Rodgers, J.L., Christensen, K., 1999. Is fertility behavior in our genes? Findings from a Danish twin study. Population and Development Review 25, 253–288.

Kohler, Hans-Peter, Rodgers, J.L., Christensen, Kaare, 2002. Between nurture and nature: the shifting determinants of female fertility in Danish twin cohorts. Social Biology 49 (3–4), 218–248.
Lacerda, Andre Luis Ribeiro, 2009. Biosocial approaches in sociology: biosociology or evolutionary sociology? (Original Title: Abordagens biossociais na sociologia: biossociologia ou sociologia evolucionista). Revista Brasileira de Ciencias Sociais 24 (70), 155–165.
Lenski, Gerhard, 1984. Power and Privilege: A Theory of Social Stratification. University of North Carolina Press, Chapel Hill.
Lenski, Gerhard, 2005. Ecological–Evolutionary Theory: Principles and Applications. Paradigm Publishers, Boulder and London.
Lenski, Gerhard, Nolan, Patrick, 2005. Trajectories of development among third world societies. In: Lenski, Gerhard (Ed.), Evolutionary Theory: Principles and Applications. Paradigm Publishers, Boulder, CO, pp. 187–201.
Lock, Margaret, 2009. Part 3: the human in the body demoting the genetic body. Anthropologica 51 (1), 159–172.
Lopreato, Joseph, 1984. Human Nature and Biocultural Evolution. Allen and Unwin, Boston.
Lopreato, Joseph, Crippen, T.A., 1999. Crisis in Sociology: The Need for Darwin. Transaction, New Brunswick, NJ.
Lovett, Frank, 2006. Rational choice theory and explanation. Rationality and Society 18, 237–272.
Machalek, Richard, Martin, Michael W., 2004. Sociology and the second Darwinian revolution: a metatheoretical analysis. Sociological Theory 22 (3), 455–476.
Mark, Noah P., 2011. The cultural evolution of indiscriminate altruism in a large randomly matching population. The Journal of Mathematical Sociology 35 (4), 235–248.
Massey, Douglas S., 2002. A brief history of human society: the origin and role of emotion in social life. American Sociological Review 67, 1–29.
Massey, Douglas S., 2004. Segregation and stratification: a biosocial perspective. Du Bois Review: Social Science Research on Race 1 (1), 7–25.
Massey, D.S., 2005. Strangers in a Strange Land. Humans in an Urbanizing World. Norton, New York.
Mazur, Allan, 2004. Biosociology of Dominance and Deference. Rowman & Littlefield Lanham, MD.
Mazur, Allan, Booth, Allan, 1998. Testosterone and dominance in men. Behavioral and Brain Sciences 21, 353–363.
Miller, Eleanor M., Costello, Carrie Yang, 2001. The limits of biological determinism. American Sociological Review 66 (4), 592–598.
Miller, Alan S., Stark, Rodney, 2002. Gender and religiousness: can socialization explanations be saved? American Journal of Sociology 107 (6), 1399–1423.
Nettle, Daniel, Pollet, Thomas V., 2008. Natural selection on male wealth in humans. The American Naturalist 172 (5), 658–666.
Nielsen, Francois, 2004. The ecological–evolutionary typology of human societies and the evolution of social inequality. Sociological Theory 22 (2), 292–314.
Nielsen, Francois, 2006. Achievement and ascription in educational attainment: genetic and environmental influences on adolescent schooling. Social Forces 85 (1), 193–216.
Niewohner, Jörg, 2011. Epigenetics: embedded bodies and the molecularisation of biography and milieu. BioSocieties 6 (3), 279–298.
Pescosolido, Bernice A., Perry, Brea L., Scott Long, J., Martin, Jack K., Nurnberger Jr., John I., Hesselbrock, Victor, 2008. Under the influence of genetics: how transdisciplinarity leads us to rethink social pathways to illness. American Journal of Sociology 114 (S1), S171–S201.
Powell, Brian, Bolzendahl, Catherine, Geist, Claudia, Steelman, Lala Carr, 2010. Counted Out: Same-Sex Relations and Americans' Definitions of Family. American Sociological Association's Rose Series in Sociology. Russell Sage Foundation, New York.
Risman, Barbara J., 2001. Calling the bluff of value-free science. American Sociological Review 66 (4), 605–611.
Robinson, Dawn T., Rogalin, Christabel L., Smith-Lovin, Lynn, 2004. Physiological measures of theoretical concepts: some ideas for linking deflection and emotion to physical responses during interaction. Advances in Group Processes 21, 77–115.
Roos, J.P., Rotkirch, Anna, 2005. Habitus, naturaleza o educacion? Hacia un paradigma de la sociologia evolutiva (Habitus, nature or nurture? Towards a paradigm of evolutionary sociology). In: Monreal, Juan, Diaz, Capitolina, Garcia Escribano, Juan J. (Eds.), Viejas Sociedades, Nueva Sociologia. Centro de Investigaciones Sociologicas, Madrid, pp. 87–100.
Rose, H., 2000. Colonizing the social sciences? In: Rose, H., Rose, S. (Eds.), Alas, Poor Darwin Arguments against Evolutionary Psychology. Harmony Books, New York, pp. 127–154.
Rossi, Alice S., 1984. Gender and parenthood. American Sociological Review 49 (1), 1–19.
Runciman, W.G., 2008. Forgetting the founders. The Sociological Review 56 (3), 358–369.
Sanderson, Stephen K., 2001. The Evolution of Human Sociality. Rowman and Littlefield Publishers, Lanham, Maryland.
Sanderson, Stephen K., 2007. Marvin Harris, meet Charles Darwin: a critical evaluation and theoretical extension of cultural materialism. In: Kuznar, Lawrence A., Sanderson, Stephen K. (Eds.), Studying Societies and Cultures: Marvin Harris's Cultural Materialism and Its Legacy. Paradigm Publishers, Boulder, CO, pp. 194–228.
Sanderson, Stephen K., 2008. Adaptation, evolution, and religion. Religion 38, 141–156.
Savage, Joanne, Vila, Bryan J., 2003. Human ecology, crime, and crime control: linking individual behavior and aggregate crime. Social Biology 50 (1–2:), 77–101.
Schnittker, Jason, 2008. Happiness and success: genes, families, and the psychological effects of socioeconomic position and social support. American Journal of Sociology 114, S233–S259.
Segerstråle, Ullica Christina Olofsdotter, 2000. Defenders of the Truth: The Battle for Science in the Sociobiology Debate and Beyond. Oxford University Press, Oxford and New York.
Shanahan, Michael J., Bauldry, Shawn, Freeman, Jason, 2010. Beyond Mendel's ghost. Contexts 9 (4), 34–39.
Shanahan, Michael J., Vaisey, Stephen, Erickson, Lance D., Smolen, Andrew, 2008. Environmental contingencies and genetic propensities: social capital, educational continuation, and dopamine receptor gene DRD2. American Journal of Sociology 114, S260–S286.
Singh, I., 2012. Human development, nature and nurture: working beyond the divide. BioSocieties 7 (3), 308–321.
Smith, Thomas Spence, 2004. Where sociability comes from: neurosociological foundations of social interaction. In: Camic, Charles, Joas, Hans (Eds.), The Dialogical Turn New Roles for Sociology in the Postdisciplinary Age. Rowman & Littlefield, Lanham, MD, pp. 199–220.
Spector, Tim, 2012. Identically Different: Why You Can Change Your Genes. Weidenfeld & Nicolson, London.
Stets, Jan E., Asencio, Emily K., 2008. Consistency and enhancement processes in understanding emotions. Social Forces 86 (3), 1055–1078.
TenHouten, Warren D., 2005. Primary emotions and social relations: a first report. Free Inquiry in Creative Sociology 33 (2), 79–92.
Turner, Jonathan H., 2007. Human Emotions: A Sociological Theory. Routledge, New York.
Turner, Jonathan H., Maryanski, Alexandra, 2005. Incest: The Origin of the Taboo. Paradigm Press.
Turner, Jonathan H., Maryanski, Alexandra, 2008. On the Origins of Societies by Natural Selection. Paradigm Publishers, Boulder, Colorado.
Turner, Jonathan H., Stets, Jan E., 2006. Sociological theories of human emotions. Annual Review of Sociology 32, 25–52.
Udry, J. Richard, 2000. Biological limits of gender construction. American Sociological Review 65 (3), 443–457.
Updegraff, K., Booth, A., Thayer, Shawna, 2006. The role of family relationship quality and testosterone levels in adolescents' peer experience: a biosocial analysis. Family Psychology 20, 21–29.
van den Berghe, Pierre L., December 1974. Bringing beasts back in: toward a biosocial theory of aggression. American Sociological Review 39 (6), 777–788.
van den Berghe, Pierre L., 1979. Human Family Systems. Elsevier, New York.
Van den Berghe, Pierre L., 1981. The Ethnic Phenomenon. Elsevier, New York.
van den Berghe, Pierre L., 1990. Why most sociologists don't (and won't) think evolutionarily. Sociological Forum 5 (2), 173–185.
Vaughn, Michael G., Delisi, Matt, Beaver, Kevin M., Wright, John Paul, 2009. DAT1 and 5HTT are associated with pathological criminal behavior in a nationally representative sample of youth. Criminal Justice and Behavior 36 (11), 1103–1114.
Whitmeyer, Joseph M., 1994. Why actor models are integral to structural analysis. Sociological Theory 12, 153–165.
Whitmeyer, Joseph M., 1997. Endogamy as a basis for ethnic behavior. Sociological Theory 15 (2), 162–178.
Witherspoon, D.J., Wooding, S., Rogers, A.R., Marchani, E.E., Watkins, W.S., Batzer, M.A., Jorde, L.B., 2007. Genetic similarities within and between human populations. Genetics 176, 351–359.

Relevant Websites

http://www.cpc.unc.edu/projects/addhealth – Add-Health Dataset Website.
http://hbes.com/ – Human Behavior and Evolution Society Website.
http://www2.asanet.org/sectionevol/ – Section Website.

Evolutionary Theory, Structure of

Elisabeth Lloyd, Indiana University, Bloomington, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Charles Darwin proposed a general type of natural selection model that could explain a variety of particular cases of adaptation to local environments, once details of organismic traits and selection pressure were inserted. Much of evolutionary theory today, though not all, is represented through mathematical models, especially through the models of population genetics, of the evolution of states of a given system, both in isolation and interaction through time. This article discusses in detail various ways to describe the evolutionary models that make up evolutionary theory. The main items needed for this description are the model types of natural selection, drift, and so forth, most often described through the definition of a state space, state variables, parameters, and a set of laws of succession and coexistence for the system. Choosing a 'state space' (and, thereby, a set of state variables) for the representation of genetic states and changes in a population is a crucial part of population genetics theory. Claims about evolutionary models may be confirmed in three different ways: (1) through fit of the outcome of the model to a natural system; (2) through independent testing of assumptions of the model, including parameters and parameter ranges; and (3) through a range of instances of fit over a variety of the natural systems to be accounted for by the model, through a variety of assumptions tested, and including both instances of fit and some independent support for aspects of the model.

The Structure of Darwinian Evolutionary Models

The basic structure of Darwinian evolutionary models is deceptively simple and elegant; such simplicity can yield powerful change over the proposed time spans encountered by the biological systems involved. For evolution by natural selection, for example, we start with a basic very general model structure that includes a set of assumptions or components (Griesemer, 1984; Lloyd, 1983; Van Fraassen (1980: p. 44) has defined 'model type' as the description of a structure in which certain parameters are left unspecified in this way). The classic selection model type concerns a population of varied organisms, and these variations are assumed to be at least partially inheritable. In this population's environment (which is always defined relative to the organism's needs and sensitivities), there are demands and stresses affecting some variants more than others, resulting in differential reproduction. Thus, the basic natural selection model type leads to the change in the population variants' composition and structure over evolutionary time. (For discussion of multilevel selection model types, see Lloyd, 2012.) In 1859, Charles Darwin, in *On the Origin of Species*, proposed his basic abstract selection model type, and other evolutionary model types, and also filled in their details in various cases in various ways in the book and his correspondence; in this manner, he illustrated how to use and apply the models (*see* Darwinism).

More specifically, each of the model types serve as formats for explanation; the particular terms or factors in a model type vary in each application, depending on the outcome of the model and various assumed conditions (Griesemer, 1984; Lloyd, 1983). For example, in the *Origin*, Darwin needed evidence for his assumption that wild organisms spontaneously developed heritable variations some of which would be advantageous to their survival and reproduction. Such a general assumption of the existence of useful variations was necessary for a variable in the natural selection model type, which would then be instantiated in specific models by individual cases of variation. Darwin found empirical evidence for this assumption in the animal breeders' information, as he presents in the *Origin*. (Darwin kept up an avid and long-term correspondence with a number of pigeon breeders, from whom he learned detailed information about the vast variety of feather form, pattern, and color, as well as behaviors that spontaneously arose in their pigeons.) Elsewhere in his correspondence, Darwin also offers evidence to support an empirical assumption of a specific model constructed using the selection model type. In looking at wingless insects that appear on oceanic islands at a higher concentration than elsewhere, Darwin proposed to explain their frequency by a selection pressure of high winds blowing them off the islands, thus favoring insects without wings. He tested this model assumption by comparing an even smaller island to the other islands he was examining earlier: he found, as he predicted, an even higher percentage of wingless insects. His earlier assumption of a selection pressure was thus indirectly confirmed in his proposed model (Beatty, 1980: p. 401; Lloyd, 1983: pp. 121–122; Thompson, 1983: pp. 226–227). Section Confirming and Testing Models below provides a more detailed discussion about the confirmation of evolutionary models.

Note that the evolutionary selection models can range from the very abstract and general model structure with its high level assumptions, through to ever more specified models as we identify and fill in the necessary assumptions of selection pressures, types of variation, and evidence for heritability. Ultimately, our ever more detailed specifications of the assumptions result in a fully specified selection model, which anchors the most concrete end of the model continuum, from the most abstract model type to the most concrete model (Beatty, 1980: pp. 412–413; Griesemer, 1984: pp. 106–107; Lloyd, 1983: pp. 118–119).

The Structure of Population Genetic Evolutionary Models

While Darwin's evolutionary models and explanations were nonquantitative, much of evolutionary theory at the beginning of the twenty-first century is represented through mathematical models or equations of population genetics, which specify the evolution of states of a given system, both in isolation and interaction, through time. This is done by conceiving of the evolutionary model as capable of a certain set of states, with these states represented by elements of a certain mathematical space, the state space (Beatty, 1980; Lloyd, 1988/1994; Thompson, 1983). Generally speaking, 'models' and 'systems' always refer to ideal systems; when the actual biological systems are being discussed, they are called 'empirical' or 'natural' systems. The variables used in each mathematical model represent distinct measurable, or potentially quantifiable, physical/biological magnitudes. Classically, any particular configuration of values for these variables is a 'state' of the system, the 'state space' being the collection of all possible configurations of the variables.

The theory itself represents the behavior of the system in terms of its states: the rules or laws of the theory (i.e., laws of coexistence, succession, or interaction) can delineate various configurations and trajectories on the state space. A description of the structure of the theory itself therefore only involves the description of the set of models that make up the theory. In an application of the theory or models to the real world, we consider both 'data models' and the idealized models describing idealized systems and their resemblance. 'Data models' are simplified structures representing the natural world made up of measurements, observations, or the results of experiments; that is, numbers or values extracted from the real world, many of which are arranged in relations suggested by the theoretical models (Suppes, 1962). The ideal systems are usually gradually specified, as described in Section The Structure of Darwinian Evolutionary Models, to show a good match or similarity with the data models, to which they are compared during an application of the theory or model type (Giere, 1988; Van Fraassen, 1980).

Construction of a model within the theory involves assignment of a location in the state space of the theory to a system of the kind defined by the theory. Potentially, there are many kinds of systems that a given theory can be used to describe. Limitations come from the dynamical sufficiency (whether it can be used to describe the system accurately and completely) and the accuracy and effectiveness of the laws used to describe the system and its changes. Thus, there are two main aspects to defining a model. First, the state space must be defined, which involves choosing the variables and parameters with which the system will be described; second, coexistence laws must be defined, which describe the structure of the system, and laws of succession, which describe changes in its structure.

Defining the state space involves defining the set of all of the states the system could possibly exhibit. Certain mathematical entities – in the case of many evolutionary models, these are vectors – are chosen to represent these states. The collection of all the possible values for each variable assigned a place in the vector is the state space of the system. The system and its states can have a geometrical interpretation: the variables used in the state description (i.e., state variables) can be conceived as the axes of a Cartesian space.

The state of the system at any time may be represented as a point in that space, located by projection on the various axes. The family of measurable physical magnitudes, in terms of which a given system is defined, also includes a set of parameters. Parameters are values that are not themselves a function of the state of the system. Thus, a parameter can be understood as a *fixed value* of a variable in the state space – topologically, setting a parameter amounts to limiting the number of possible structures in the state space by reducing the dimensionality of the model.

Laws, used to describe the behavior of the system in question, must also be defined in a description of a model or set of models. Laws have various forms: in general, coexistence laws describe the possible states of the system, while laws of succession describe changes in the state of the system.

Let us discuss in more detail the description of the evolutionary models that make up current evolutionary theory. The main items needed for this description are the definition of a state space, state variables, parameters, and a set of laws of succession and coexistence for the system. Choosing a 'state space' (and, thereby, a set of state variables) for the representation of genetic states and changes in a population is a crucial part of population genetics theory. (Population genetics models are a highly technical, but not necessarily 'core,' part of evolutionary theory itself (Thompson, 1983)).

Paul Thompson suggests that the state space for population genetics would include the physically possible states of populations in terms of genotype frequencies. The state space would be "a Cartesian *n*-space where '*n*' is a function of the number of possible pairs of alleles in the population" (Thompson, 1983: p. 223). We can picture this geometrically as *n* axes, the values of which are frequencies of the genotype. The state variables are the frequencies for each genotype. Note that this is a one-locus system; that is, we take only a single gene locus and determine the dimensionality of the model as a function of the number of alleles at that single locus.

Another type of single locus system, used less commonly than the one described by Thompson, involves using single gene frequencies, rather than genotype frequencies, as state variables. Some of the debates about 'genic selectionism' center around the descriptive, dynamical, and parametric adequacy of this state space and its parameters for representing evolutionary phenomena (Dawkins, 1982; Lloyd et al., 2008).

With both genotype and gene frequency state spaces, treating the genetic system of an organism as being able to be isolated into single loci involves a number of assumptions about the system as a whole. For instance, if the relative fitnesses of the genotypes at a locus are dependent on other loci, then the frequencies of a single locus observed in isolation will not be sufficient to determine the actual genotype frequencies. Assumptions about the structure of the system as a whole can thus be incorporated into the state space in order to reduce its dimensionality. Lewontin (1974) offers a detailed analysis of the quantitative effects of dimensionality of various assumptions about the biological systems being modeled. Interactions between genotypes, and between one locus and another cannot be represented in a single locus model (with the exception of frequency-dependent selection), for the

simple reason that they involve more than one locus (Lewontin, 1974: pp. 273–281).

As far as the structure of the theory goes, although all single locus models should, in some sense, be grouped together, they are not all exactly the same model – each particular model has a different number of state variables, depending upon the number of alleles at that locus. Because a model type is simply an abstraction of a model, constructed by abstracting one or more of the models' parameters or variables, a single model can be an instance of more than one model type. This is an extremely important aspect of the flexibility of this approach to theory structure. Along similar lines, we may say that each model type can be associated with a distinctive state space type. In the preceding example, the single locus model is to be taken as an instance of a general state space type for all single locus models; that is, the different single locus model types are conceived as utilizing the same state space type. Alternatives, such as two-locus models, must be taken as instances of a different state space type.

'Parameters' are the values that appear in the succession and coexistence laws of a system that are the same for all possible states of the defined system. For instance, in the modification of the Hardy–Weinberg equation that predicts the frequencies of the genotypes after selection, the selection coefficient, s, or multiplier, appears as a parameter in the formula

$$p^2AA + 2pqAa + (1-s)q^2aa = 1 - sq^2.$$

There are a variety of methods of establishing the value at which a parameter should be fixed or set in the construction of models for a given real system; for example, simulation techniques can be used to obtain estimates of biologically important parameters. In some contexts, maximum likelihood estimations may be possible. Parameters can also be set arbitrarily, or ignored. This is equivalent to incorporating certain assumptions into the model for purposes of simplification (Levins, 1968: pp. 8, 89).

One expects the values of parameters to have an impact on the system being represented; but variations in parameter values can make a larger or smaller amount of difference to the system. For instance, take a deterministic model that incorporates a parameter for mutation, μ. The rates of change of this model are virtually insensitive to realistic variations in the value of the mutation parameter, μ. Yet the selection parameters play a crucial role in this same model. A very small amount of selection in favor of an allele will have a cumulative effect strong enough to replace other alleles (Lewontin, 1974: p. 267).

Population size is another case in which the value assigned to the parameter has a large impact on the model results. The parameter for effective population size, N, can play a crucial role in some models, because selection results can be quite different with a restricted gene pool size (Mayr, 1967: pp. 48–50). In many of the stochastic models involved in calculating rates of evolutionary change, the resulting distributions and their moments can depend completely on the ratio of the mean deterministic force to the variance arising from random processes (Lewontin, 1974: p. 268). This variance is usually proportional to $1/N$ and is related to the finiteness of population size. Thus, change in the value of the single parameter, N, can completely alter the structure represented by the theory.

The choice of parameters can also make a major difference to the model outcome. Theoreticians have choices about how to express certain aspects of the system or environment. The choice of parameters used to represent the various aspects can have a profound effect on the structure, even to the point of rendering the model useless for representing the real-world system in question. Group selection models provide a case in which choice of parameters not only alters the results of the models, but also can lead to the near disappearance of the phenomenon being modeled (Wade, 1978).

Some authors, when discussing genetic changes in populations, speak of the system in terms of a type of state space involving phenotypes. This makes sense, because the phenotype determines the breeding system and the action of natural selection, the results of which are reflected in the genetic changes in the population; quantitative genetics is the set of models that concentrate on phenotypic state spaces. In his analysis of the structure of population genetics theory, Lewontin (1974) traces a single calculation of a change in genetic state through both genotypic and phenotypic descriptions of the population.

But in addition, mating patterns, and probabilities of survival and reproduction, although influenced by genes, are a consequence of developmental events that are contingent on the environment of the developing organism, and involve more than any simple phenotype. In the most general case, environment includes influences of the phenotypes of previous generations by means of cytoplasmic inheritance through the egg. A complete general representation of genetic evolutionary processes then requires not two, but six spaces with sequential transitions within them and mappings from one to the other (Lloyd et al., 2008). Haploid spaces represent only the single chromosome or set of one sex's genes, while (sexually reproducing) diploid state spaces represent both sets. The complete set of six spaces, S, is thus:

S_1, a diploid phenotypic space;
S_2, a diploid genotypic space;
S_3, a diploid pair phenotypic space;
S_4, a diploid pair genotypic space;
S_5, a haploid phenotypic space; and
S_6, a haploid genic space.

(Assuming that organisms reproduce in discrete generations, with no overlap between generations.) While it is true that the system's transition from one generation to the next may be represented in any one of the six spaces, we need to know the entire loop – and all the parameter values in each of the model stages in that loop – in order to obtain a fully accurate representation of any chosen state space in the next generation, that is, to get the state transformation equation between generations *within any one of the spaces* (e.g., if we want to move from (S_6) to (S_6 in the new generation)). For instance, for the transition in allelic space, we must move out of that space, into genotypic space to define the fitness parameters, and back into allelic space in order to characterize the next generation (Lloyd et al., 2008: pp. 143–144).

Now consider a few particular aspects and forms of the 'laws' used in population genetic models. The most obvious differences, in laws as well as in state space types, are between deterministic and stochastic models. Coexistence laws describe the possible states of the system in terms of the state space. In the case of evolutionary theory, these laws would consist of

conditions delineating a subset of the state space that contains only the biologically possible states. Laws of succession describe changes in the state of the system. In the case of evolutionary theory, dynamic laws concern changes in the genetic composition of populations.

The laws of succession select the biologically possible trajectories in the state space, with states at particular times being represented by points in the state space. The law of succession is the equation of which the biologically possible trajectories are the solutions. The Hardy–Weinberg equation is the fundamental law of both coexistence and succession in population genetics theory. Even the dynamic laws of the theory are usually used to assess the properties of the equilibrium states and steady state distributions. The Hardy–Weinberg law is a very simple, deterministic succession law that is used in a very simple state space. As parameters are added to the equation, we get different laws, technically speaking. For example, compare the laws used to calculate the frequency p' of the A allele in the next generation. Plugging only the selection coefficient against a recessive into the basic Hardy–Weinberg law, we get the recursion for the dominant allele frequency as $p' = p/(1-sq^2)$. Addition of parameters for the mutation rates yields a completely different law, $p'' = 1-\mu p' + \nu q'$, where μ is the rate of mutation away from the dominant allele, and ν is the rate of mutation toward it. We could consider these laws to be of a single type – variations on the basic Hardy–Weinberg law – which are usually used in a certain state space type. The actual state space used in each instance depends on the genetic characteristics of the natural system, and not usually on the parameters. For instance, the succession of a system at Hardy–Weinberg equilibrium, and of a system that is not at equilibrium but is under selection pressure, could both be modeled in the same state space, using different laws.

A theory can have either deterministic or statistical laws for its state transitions. Furthermore, the states themselves can be either statistical or nonstatistical. In population genetics models, gene frequencies that often appear in the set of state variables are statistical entities, thus the states themselves are also statistical entities. In general, a law is deterministic if, when all of the parameters and variables are specified, the succeeding states are uniquely determined. In population genetics, this means that the initial population and parameters are all that is needed to get an exact prediction of the new population state.

In sum, the structure of evolutionary theory can be understood by examining the families of models it presents. In the case of population genetics theory, the set of model types – stochastic and deterministic, single locus or multilocus – can be understood as related families of models. The question then becomes how to define the exact nature of the relationships among them, and how they relate to the rest of the evolutionary models, such as phenotypic or evolutionary-developmental models.

Expansions and Alternative Views of the Structure of Evolutionary Theory

In philosophy of science, the above analysis instantiates an approach to theory structure known as the 'semantic view,' that focuses on models (Giere, 1988; Suppe, 1977, 2000; Suppes, 1962; Van Fraassen, 1980). Some authors have misrepresented this approach, while others reject the semantic view yet support what they call a 'model-based' approach, whose differences from the semantic view are unclear (Downes, 1992; Godfrey Smith, 2006). For instance, Downes (1992: p. 143), Godfrey Smith (2006: p. 731), and Love (2012: p. 7) all claim that the semantic view is committed to 'mathematized theories,' which is not useful for some biological contexts. But both Beatty (1980) and Lloyd (1983) analyzed Darwin's (nonmathematical) selection theory and models as their first presentations of the semantic view of evolutionary theory. So clearly the semantic view can be used to analyze 'nonmathematized' theories and biological contexts. Especially in the hands of Suppes (1962), the semantic view can be used to analyze any structures into set theoretic terms, and thus into 'mathematical' terms (mathematics concerns structures, and all structures can be seen as mathematical objects); but the whole purpose is to represent the system in the most accurate, most useful, or most beneficial, etc., way. This does *not* mean that the analyzed structures are themselves represented mathematically *before* the analysis, and also does not mean that they are represented in equations *after* the analysis. Some authors also neglect the various hierarchies of evolutionary models offered using the semantic view, as outlined in Section The Structure of Darwinian Evolutionary Models, with misleading results (Godfrey Smith, 2006: pp. 732, 739ff; Lloyd, 1983: pp. 118–119).

In contrast, biologist Samuel Scheiner (2010) has recently offered an expanded account of the structure of evolutionary theory that takes the semantic view as its touchstone, and goes on to include other areas of biology. And Alan Love (2010) has provided a very interesting expansion of the analysis of evolutionary theory through the inclusion of developmental biology models and model types. The analysis of the structure of the theory offered in Section The Structure of Population Genetic Evolutionary Models for population genetics can thus be advanced and expanded, (using a variety of detailed methods of describing models, if needed), to incorporate developmental biology, by extending the linkages and overlaps of the model families understood to make up evolutionary theory. Love's overall approach is extremely useful, as it focuses on which research problems in evolutionary biology subfields compose the theory as a whole (2010, p. 428). Just as on the earlier analysis of Darwinian models in Section The Structure of Population Genetic Evolutionary Models, in Love's view, the entire hierarchy of models is explored, from the very abstract or general, to the extremely concrete. He emphasizes that different researchers use midlevel models in different ways and toward different ends; the resulting fully filled-out and concrete models can even be incompatible with one another. Love offers lists of model constraints from cellular and developmental biology that add additional model structure that may be needed to complement evolutionary models in order to fulfill the requirements of an extended evolutionary synthesis (Love, 2012: p. 8). By combining different empirical generalizations, some developmental biologists create mechanisms for the 'evolution of form' that can then be tested. (The component principles of these generalizations are not, themselves, best considered mechanistic models, according to Love (2012: p. 8)). The relevant causal mechanisms operate

on the level of individual ontogeny of organisms, not on population-level mechanisms such as natural selection or genetic drift. Thus we get the concrete level models just mentioned. Having a flexible and robust view of theory structure that can account for this wide range of models is imperative, for Love. He contrasts the complex and flexible view he derives from James Griesemer's (1984) semantic view account, with that of biologists Maxissimo Pigliucci and Gerd Muller (2010).

Pigliucci and Muller (2010) discuss the possibility of an extended evolutionary synthesis that is intended to incorporate evolutionary–developmental biology (evo–devo) in significant and integrated ways. They see the need for such an expansion as based on new concepts and findings in evo–devo that cannot be forced into their view of the traditional evolutionary synthesis from the 1930s and 1940s. Pigliucci and Muller offer an onion-type diagram to show the structure of evolutionary theory: they place 'Darwinism' as a circle at the very center of the onion, which includes only natural selection, inheritance, and variation. In the next outer ring, they place the 'modern synthesis,' which includes things like 'Mendelian inheritance,' 'speciation and trends,' population genetics, and gene mutation, etc. In the last, outer-most ring, which they call the 'extended evolutionary synthesis' Pigliucci and Muller place evo–devo theory, niche construction, multilevel selection, genomic phenomena, etc.

One apparent problem with this proposed structure is: how does evo–devo theory interact with natural selection, when they are layers apart in the onion? It seems that they have no integration, on this picture. Alan Love complains also that this model assumes a 'core' of evolutionary population genetics, which is always in operation, even though many have no wish to make such an assumption, and may explicitly reject it (Lloyd, 1988/1994: p. 8; Love, 2012: p. 6). The general problem with the onion picture is thus the lack of integration and relations among its parts, which is available under analyses that approach the structure as hierarchical and intertwined families of models, which include the detailed new evo–devo models.

A valuable alternative approach to the structure of current evolutionary biology as well as its history is presented by Telmo Piavani (2011), who offers a reconstruction inspired by philosopher Imre Lakatos. Piavani's updated Lakatosian approach dispenses with some weaknesses in the original Lakatosian approaches, including any strong falsificationism, and analyzes contemporary evolutionary theory into a flexible structure that delineates the theory's 'explanatory core' and its 'protective belt.' The traditional 'explanatory core' includes such concepts as: evolution as a fact, the tree of life, common descent, and natural and sexual selection. The explanatory core of evolutionary theory serves as the basic theoretical and assumptive background when building evolutionary explanations, and can be combined with other explanatory and theoretical tools from the 'protective belt,' which has expanded and contracted over time, depending on our state of biological knowledge. Currently, Piavani argues that the evolutionary core includes two hierarchies, one a genealogical hierarchy, namely of nested levels of transmission of genetic materials, organisms, demes, species, and monophyletic taxa. The other hierarchy is ecological, made up of, for example, organisms, populations like avatars, local ecosystems, and regional ecosystems, the key issue being the transfer of energy. Using these two hierarchies in the explanatory core, Piavani then describes a 'pluralistic protective belt,' that is, a set of models and assumptions, in addition to the 'explanatory core,' that may be appealed to by an evolutionary biologist when constructing an explanation (Piavani, 2011: pp. 12–13). This set of models and assumptions includes evo–devo. Piavani emphasizes that the theoretical changes that evolutionary theory has undergone since the modern synthesis in the 1930s and 1940s has been a progressive one, and now the theory is very healthy. The current dual core of the theory, and its various models and features of the protective belt demonstrate the growth of evolutionary knowledge since the modern synthesis (although they do have the disadvantage, according to someone like Love, of privileging some models as the 'core' of evolutionary theory.) Piavani also emphasizes that much of the theory is capable of 'falsification,' but *has not been falsified*, which concerns the testability and confirmation of the theory, the topic of section Confirming and Testing Models.

Confirming and Testing Models

In general, one evaluates a model by comparing the outcome of the model with empirical observations, and by independently testing the assumed conditions, including observations not available at the time of model construction (Lloyd, 1983: pp. 116–117). In population genetics, the most obvious way to support a claim of the form "this natural system is described by this model," is to demonstrate the simple matching of some part of the model with some part of the natural system being described. (Throughout this discussion, let us assume that each 'match' of model to a natural system in the real world involves a specification of the model, by filling in variable and parameter values, plus a determination from the real world of a 'data model' into a similar format. We will skip these specifications in our discussion, henceforth, and for simplicity, simply say that a "model matches or describes the natural world.") In a population genetics model, the solution of an equation might yield a single genotype frequency value. The genotype frequency is a state variable in the model. Given a certain set of input variables (e.g., the initial genotype frequency value, in this case), the output values of variables can be calculated using the rules or laws of the model. The output set of variables (i.e., the solution of the model equation given the input values of the variables) is the 'outcome' of the model. Determining the fit between model and natural system involves testing how well the genotype frequency trajectory calculated from the model (the outcome of the model) matches that measured in the relevant natural (or experimental) populations. It can be evaluated by determining the fit of one curve (the model trajectory or coexistence conditions) to another (taken from the natural system) using ordinary statistical techniques of evaluating curve fitting.

Also, numerous assumptions are made in the construction of any model, and testing a model sometimes involves confirming some or all of these assumptions independently. These include assumptions about which factors influence the changes in the

system, what the ranges for the parameters are, and what the mathematical form of the laws is. On the basis of these assumptions, the models take on certain features. Many of these assumptions thus have potential empirical content. That is, although they are assumptions made about certain mathematical entities and processes during the construction of the theoretical model, when empirical claims are then made about this model, the assumptions may have empirical significance. For instance, the assumption might be made during the construction of a model that the population is panmictic, that is, that all genotypes interbreed at random with each other. The model outcome, in this case, is still a genotype frequency trajectory, for which ordinary curve-fitting tests can be performed on the natural population to which the model is applied. But the model can have additional empirical significance, given the empirical claim that a natural system is a system of the kind described in the model. The assumption of panmixia, as a description of the population structure of the system under question, must be considered part of the system description that is being evaluated empirically. Evidence to the effect that certain genotypes in the population breed exclusively with each other (i.e., evidence that the population is not, in fact, panmictic) would undermine empirical claims about the model as a whole, other things being equal, and even if genotype frequencies (i.e., the model outcome) provide good instances of fit. In other words, the assumption that genotypes are randomly redistributed in each generation is intrinsic to many population genetics model types. Hence, although the assumption that the population is panmictic often is taken for granted in the actual definition of the model type – that is, in the law formula – it is interpreted empirically and plays an important role in determining the empirical adequacy of the claim.

By the same token, evidence that the assumptions of the model hold for the natural system being considered will increase the credibility of the claim that the model accurately describes the natural system. Technically, we can describe this situation as follows. From the point of view of empirical claims made about it, the model has three parts: state variables, empirically interpreted background assumptions, and those aspects and assumptions of the model that are not directly empirically interpreted. Because the possible outcomes of the model (along with the inputs) are actually values of the state variables, we can understand the input and outcome of the model as a sort of minimal empirical description of the natural system. If a model is claimed to describe accurately a natural system, at the very least, this means that the variables in which the natural system is described change according to the laws presented in the theoretical model. Under many circumstances, such models, in which only the state variables are empirically interpreted, are understood to be 'mere' calculating devices, because the match with the ('data model' extracted from the) natural system is so limited.

Various aspects of the model – for instance, the form of its laws, or the values of its parameters – rely on assumptions made during model construction that can be interpreted empirically. The assumption of panmixia, discussed above, is an empirically interpreted background assumption of many population genetics models. Because it is empirically interpreted, the presence or absence of panmixia in the natural population in question makes a difference to the empirical adequacy of the model. Finally, there may be aspects of the model that are not interpreted empirically at all. For instance, some parameters appearing in the laws might be theoretically determined and might have no counterpart in the natural system against which the model is compared.

Given that there can be aspects of the model structure that are not directly tested by examining the fit of the state variable curve, direct testing of these other aspects would give additional reason to accept (or reject) the model as a whole. In other words, it is taken that direct testing provides a stronger test than indirect testing, hence there is a higher degree of confirmation if the test is supported by empirical evidence. Direct empirical evidence for certain empirically interpreted aspects of the model that are not included in the state variables (and thus are confirmed only indirectly by goodness-of-fit tests), therefore provides additional support for the application of the model.

The above sort of testing of assumptions involves making sure that the empirical conditions for application of various parts of the model description actually do hold. In other words, in order to accept an explanation constructed by applying the model, the conditions for application must be verified. The specific values inserted as the parameters or fixed values of the model are another important aspect of empirical claims that can be tested. In some models, mutation rates, etc., appear in the equation – part of the task of confirming the application of the model involves making sure that the values inserted for the parameters are appropriate for the natural system being described. Finally, there is a more abstract form of independent support available, in which some general aspect of the model, for instance, the interrelation between the two variables, or the significance of a particular parameter or variable, can be supported through evidence outside the application of the model itself.

Lastly, variety of evidence, of which there are at least three kinds, is an important factor in the evaluation of empirical claims. I discuss three kinds of variety of evidence here: (1) variety of instances of fit; (2) variety of independently supported model assumptions; and (3) variety of types of support, which include fit and independent support of aspects of the model.

First, let us consider variety of fit, that is, variety of instances in which the model outcome matches the value obtained from the natural system. Traditionally, variety of evidence has meant variety of fit. There are two issues that must be distinguished when considering the variety of fit. One involves the range of application of a model or model type, while the other involves the degree of empirical support provided for a single application. Both of these issues involve the notion of independence, which needs to be clarified before continuing. One point of any claim to variety of evidence is to show that there is evidence for the hypothesis in question from different or independent sources. The notion of independence here is not the sort of independence that is found in the frequency interpretation of probability theory, where the independence depends only on accepting a particular reference class. Rather, in the context of scientific theories, independence is relative to whatever theories have already been accepted.

Significantly, in evolutionary theory, independence is usually relative to some assumption about natural kinds. Empirical claims are often made to the effect that a model is

applicable over a certain range of natural systems. Inherent in these claims is the assumption that all of the natural systems in the range participate in some key feature or features that make it possible and/or desirable to describe them all with the same model type. Thus, there is some assumption of a natural kind (characterized by the key feature or features) whenever a range of applications arises. The scientist, in making a claim that the model type is applicable to such and such systems, is making an empirical assumption about the existence of the key feature or features in the range of natural systems under question. Hence, for evolutionary models, testing for variety of fit depends on accepting an assumption about what constitutes different or independent instances of fit; this, in turn, amounts to accepting a particular view of natural kinds. Part of what variety of evidence does, in a bootstrapping effect, is to confirm this original assumption about natural kinds. More technically, variety of evidence confirms the sufficiency of the parameters and state space.

Take the case in which a model type is claimed to be applicable over a more extended range than that actually covered by available evidence. This extrapolation of the range of a model can be performed by simply accepting or assuming the applicability of the model to the entire range in question. A more convincing way to extend applicability is to offer evidence of fit between the model and the data in the new part of the range. Provision of a variety of fit can thus provide additional reason for accepting the empirical claim regarding the range of applicability of the model.

For instance, a theory confirmed by 10 instances of fit involving populations of size 1000 (where population size is a relevant parameter) is in a different situation with regard to confirmation than a theory confirmed by one instance of each of 10 different population sizes ranging from 1 to 1 000 000. If the empirical claims made about these two models asserted the same broad range of applicability, the latter model is confirmed by a greater variety of instances of fit. That is, the empirical claim about the latter model is better confirmed, through successful application (fits) over a larger section of the relevant range than the first model. Variety of fit can therefore provide additional reason for accepting an empirical claim about the range of applicability of a model. Variety of fit can also provide additional reason for accepting a particular empirical claim, that is, one of the form, "this natural system is accurately described by that model." That is, variety of evidence can serve to increase confidence in the accuracy of any particular description of a natural system by a model.

Variety of fit is only one kind of variety of evidence. An increase in the number and kind of assumptions tested independently, that is, greater variety of assumptions tested, would also provide additional reason for accepting an empirical claim about a model. The final sort of variety of evidence involves the mixture of instances of fit and instances of independently tested aspects of the model. In this case, the variety of types of evidence offered for an empirical claim about a model is an aspect of confirmation.

According to the view of confirmation sketched above, claims about evolutionary models may be confirmed in three different ways: (1) through fit of the outcome of the model to a natural system; (2) through independent testing of assumptions of the model, including parameters and parameter ranges; and (3) through a range of instances of fit over a variety of the natural systems to be accounted for by the model, through a variety of assumptions tested, and including both instances of fit and some independent support for aspects of the evolutionary model.

See also: Adaptation, Fitness, and Evolution; Cultural Evolution: Theory and Models; Darwinism; Evolution, History of; Microevolution.

Bibliography

Beatty, J., 1980. What's wrong with the received view of evolutionary theory?. In: Asquith, P.D., Giere, R.N. (Eds.), PSA, vol. 2. Philosophy of Science Association, East Lansing, MI, pp. 397–426.

Dawkins, R., 1982. The Extended Phenotype. Oxford University Press, Oxford, UK.

Downes, S.M., 1992. The Importance of Models in Theorizing: A Deflationary Semantic View. In: PSA, vol. 1. University of Chicago Press, Chicago, IL, 142–153.

Giere, R., 1988. Explaining Science. University of Chicago Press, Chicago, IL.

Godfrey Smith, P., 2006. The strategy of model-based science. Biology and Philosophy 21, 725–740.

Griesemer, J.R., 1984. Presentations and the status of theories. In: Asquith, P.D. (Ed.), PSA, vol. 1. Philosophy of Science Association, East Lansing, MI, pp. 102–114.

Levins, R., 1968. Evolution in Changing Environments. Princeton University Press, Princeton, NJ.

Lewontin, R.C., 1974. The Genetic Basis of Evolutionary Change. Columbia University Press, New York.

Lloyd, E.A., 1983. The nature of Darwin's support for the theory of natural selection. Philosophy of Science 50, 112–129.

Lloyd, E.A., 1988/1994. The Structure and Confirmation of Evolutionary Theory. Princeton University Press, Princeton, NJ.

Lloyd, E.A., 2012. Units and Levels of Selection. Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/selection-units/ (August 2005. Revised 2012).

Lloyd, E.A., Lewontin, R.C., Feldman, M., 2008. The generational cycle of state spaces and adequate genetical representation. Philosophy of Science 75, 140–156.

Love, A., 2010. Rethinking the structure of evolutionary theory for an extended synthesis. In: Pigliucci, M., Muller, G.B. (Eds.), Evolution – The Extended Synthesis. MIT Press, Cambridge, MA, pp. 403–441.

Love, A., July 19, 2012. Theory is as theory does: scientific practice and theory structure in biology. Biological Theory. http://dx.doi.org/10.1007/s13752-012-0046-2.

Mayr, E., 1967. Evolutionary challenges to the mathematical interpretation of evolution. In: Moorehead, P.S., Kaplan, M.M. (Eds.), Mathematical Challenges to the Neo-Darwinian Interpretation of Evolution. Wistar Institute Press, Philadelphia, PA, pp. 47–54.

Piavani, T., 2011. An evolving research programme: the structure of evolutionary theory from a Lakatosian perspective. In: Fasolo, A. (Ed.), The Theory of Evolution and Its Impact. Springer-Verlag, Berlin, pp. 211–228. http://dx.doi.org/10-1007/978-88-470-19744_14.

Pigliucci, M., Muller, G.B., 2010. Elements of an extended evolutionary synthesis. In: Pigliucci, M., Muller, G.B. (Eds.), Evolution – The Extended Synthesis. MIT Press, Cambridge, MA, pp. 3–17.

Scheiner, S., 2010. Toward a conceptual framework for biology. The Quarterly Review of Biology 85, 293–318.

Suppe, F., 1977. The Structure of Scientific Theories, second ed. University of Illinois Press, Urbana, IL.

Suppe, F., 2000. Understanding scientific theories: an assessment of developments, 1969–1998. Philosophy of Science 67, S102–S115.

Suppes, P., 1962. Models of data. In: Nagel, E., Suppes, P., Tarski, A. (Eds.), Logic, Methodology, and Philosophy of Science. Stanford University Press, Stanford, CA, pp. 252–261.

Thompson, P., 1983. The structure of evolutionary theory: a semantic perspective. Studies in History and Philosophy of Science 14, 215–229.

Van Fraassen, B.C., 1980. The Scientific Image. Clarendon Press, Oxford University Press, Oxford, UK.

Wade, M., 1978. A critical review of the models of group selection. Quarterly Review of Biology 53, 101–114.

Exchange in Anthropology

Gustav Peebles, The New School, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C. Gregory, volume 8, pp 5039–5042, © 2001, Elsevier Ltd.

Abstract

Anthropological research into exchange stretches back to the origins of the discipline. In surveying the history and findings of this research, it is shown that anthropology's qualitative method of studying exchange can be seen as complementary to, rather than in competition with, the quantitative study of exchange within the discipline of economics. In the late 1980s, the study of exchange within anthropology turned upon several important texts that have helped to set the agenda leading up to the current day.

Introduction

In some ways, the study of exchange within anthropology still hews very closely to Malinowski's early attack against the "current economic text books" (Malinowski, 1961: p. 46). In those pages, he complained that the discipline of economics had proclaimed that "primitive man" was merely "actuated by pure economic motives of enlightened self-interest." Malinowski, on the other hand, was at pains to show that the Trobrianders of his own research were instead prompted to action by motives that were "highly complex, [and] not directed towards… the direct achievement of utilitarian purposes" (Malinowski, 1961: p. 46). Indeed, it could be argued that it is hard to grasp anthropological exchange theory without grasping its long battle with the discipline of economics. Despite occasional harmonic interludes, both disciplines produced and reproduced two influential, but separate, lines of inquiry that divided the world of exchange theory into qualitative and quantitative approaches.

In engaging this now age-old battle, the anthropological study of exchange built up a large, twentieth-century edifice upon the supposed distinction between 'nonmarket' and 'market' exchange. Once anthropology's bailiwick was circumscribed and fortified, the discipline became the suitable zone of expertise on all matters relating to the 'gift economy,' a space wherein one could find unquantifiable and nonutilitarian values circulating, such as love and communal commitment. Economics, meanwhile, focused on the market-based circulation of commodities, governed by the quantifiable value of money (see Hann and Hart, 2011: pp. 12–13).

But for at least 30 years, anthropological research has questioned why its theories of exchange must be cordoned off within the gift economy or the nonmarket sphere. Gifts and commodities are no longer so easily segregated from one another, nor is quantity from quality. Rather, much research has focused on how quantitative value assessments and exchanges are socially produced in the first place. Equally, one might say that Benjamin's (1999: p. 18, passim) belief in the quasi-mystical 'phantasmagoria' of commodities has become at least as vital to advances in the anthropological study of exchange as Mauss's enduring notion of the 'hau' of the gift. (To be fair, Mauss spoke of similar issues in his conclusion, when he wrote, "Things sold still have a soul" (1990: p. 66). But the more general neglect of these comments is symptomatic of the enduring divide described here.) This article, therefore, will first chart the developments that emerged out of the bifurcation between the qualitative and the quantitative, and then conclude by showing new trends within the anthropological study of exchange that have begun to productively breach these strongholds.

The Concept of Exchange

The old Rousseauian-Marxist concept of 'alienation' proves to be foundational when searching for an overarching rubric for generations of anthropological exchange theory. Both classical anthropological theory as well as more recent work relies on the idea that the only things that can be exchanged are things that can become detached from individuals or societies, even if there is disagreement about what, precisely, becomes detached (see below for an important challenge to this general paradigm). One exchange partner has an item of value that he or she can do without, while another seeks to obtain this precise item in exchange for something that this second exchange partner is equally willing to alienate. For this reason, classic anthropological studies diagram exchanges as shown by Gregory (2001: p. 5039), "where A gives B the object x in exchange for the object y":

$$A \underset{y}{\overset{x}{\rightleftarrows}} B$$

While this may state the obvious, alienation is intriguing as a starting point for at least two reasons. First, it illuminates culturally produced visions of self and other – along a wide spectrum from the individual to the family to the national level. A classic example emerges out of the Ricardian (1821) understanding of comparative advantage. A country in a northerly clime might decide that it could, indeed, grow grapes at all costs, by investing in heavy infrastructure such as greenhouses and energy. If it were to do so, it would be collectively deciding that it was unwilling to exchange anything for grapes; that is, it would be refusing to alienate the capacity to grow grapes to an Other outside of itself, for it

views the growing of grapes as constitutive of its own self-perception and internal needs. This vital item, therefore, cannot be cast into the more dangerous and tenuous field of exchange, but must remain self-produced and unalienated. A less frivolous example of this would be a country that insisted on retaining its capacity to produce military materiel, for fear that leaving this to the realm of exchange could leave it highly vulnerable. In this sense, exchange constantly affirms, reaffirms, and challenges the borders of self and other.

Second, the grounding concept of alienation is fruitful for it frames another crucial issue with regard to exchange: property. Exchange sets in motion an entire field of variant property relations – socially legitimated ways in which objects can be attached, detached, and reattached to different bodies, whether individual or corporate. Some of the most important contributions from anthropology have evolved out of its long-standing insistence that property itself can be a field of contestation or misunderstanding. This greatly complicates a more utilitarian approach to exchange, which typically can only envision fee simple and freehold ownership structures (so much so that the technical term 'usufruct' has largely disappeared from English vernacular). Anthropological studies of cooperatives, commonly held land, inalienable goods, and other such modes of property and nonproperty serve as reminders that the notion of property can come in many shapes and sizes. Since the 2008 financial crisis, when the general public learned that hedge funds have the capacity to sell assets that they do not even own, it has become increasingly clear that anthropological approaches to exchange can easily expand well beyond their old boundaries in 'native life.'

Both these foundational concepts – alienation and property relations – allow other vital aspects of social life to come to the fore when analyzing exchange, whether one is talking about the inversion of exchange (i.e., hoarding), the division of labor, the interplay between hierarchy and equality, the relationship between power and redistribution, and, finally, the role of value itself in the grounding of social life. Considering this, one can quickly see why a reduction of exchange to mere utilitarian price fails to explain the rich and complex panoply of issues that grow out of it.

Examples from the Corpus

A review of the primary advances in the theory of exchange within anthropology will quickly confirm the discipline's commitment to this nonreductionist, qualitative approach. Beginning with Boas (1966: pp. 77–104), who first studied the famous potlatch ceremony, one sees a drive toward studying exchange from a nonmaterialist perspective. How, he wondered, could people actively seek to extinguish wealth, rather than cultivate it? The Smithian (1976 [1776]) notion that exchange and the expansion of material prosperity went hand-in-hand seemed turned on its head by studies such as these.

Malinowski picked up this thread, by attempting to provide an extensive typology of exchange, ranging from "pure gifts" (characterized by the socially binding exchanges displayed within family life) all the way to "trade, pure and simple," which was governed by "mutual advantage" (Malinowski, 1961: pp. 177–191). He argued that, by way of placing any given exchange of values somewhere along this spectrum, the Trobrianders navigated and built the wide variety of social ties present in their day-to-day lives. Further, each type of trade was accompanied by a corresponding type of social behavior. For example, crude haggling prevailed during "trade, pure and simple," but highly specific religious taboos governed certain customary payments (Malinowski, 1961: pp. 332–333).

If Malinowski focused on the shifting role of humans in any given exchange, Mauss interceded in the debate by making the profound suggestion that objects themselves might also carry peculiar powers that help to motivate and circumscribe exchange: "Souls are mixed with things; things with souls…. Each emerges from their own sphere and mixes together. This is precisely what contract and exchange are" (Mauss, 1990: p. 20). Most famously, he picked up a Maori term – 'hau' – that has become rote knowledge for all anthropologists since. The 'hau' of an object represents a certain spirit of the original giver that has moved into the object itself. In particular, the 'hau' sets in motion the spirit of reciprocity and obligation, because this 'hau' is a spiritual element hearkening from the object's place of origin, to which it always seeks to return (Mauss, 1990: pp. 8–13).

Levi-Strauss swept into anthropological exchange theory with revelatory fervor, born out of a belief that structural linguistics could provide newfound objectivism to a discipline dominated by subjectivism. In *The Elementary Structures of Kinship* (1969), Levi-Strauss attributes a material value to women, thereby mobilizing them into the sphere of exchange. In his theory, they are but the supreme commodities (a "group's most important assets," p. 479), to be transferred among and across generations of men. The incest rule, in this telling, is much like a gold standard; because it proscribes access to a specific range of women in any given kinship system, Levi-Strauss sees it is a purely artificial method for ensuring that the availability of women is never allowed to inflate away in a sea of endless abundance. And yet, despite this reductionist approach, Levi-Strauss still follows Durkheim's dictum that the sphere of exchange had a meaning well beyond the mere satisfaction of material needs and wants. For him, exchange "provides the means of binding men together, and of superimposing upon the natural links of kinship the henceforth artificial links… of alliance" (Levi-Strauss, 1969: p. 480). Important critiques of his theory, influenced by feminist anthropology, can be found in Rubin (1975) and MacCormack and Strathern (1980).

Bourdieu's giant corpus (e.g., 1987 and 1991) can be seen as related to Levi-Strauss's, in that he also imports the quantitative spirit of economic theory into anthropological theory. While many critics of economics have adopted Bourdieu's insistence that there are many types of capital beyond the mere economic form, ultimately, his conceptual apparatus assumes that these multiple forms of capital will seek out self-maximizing returns. He qualifies capital by farming it out to a host of different types, but each is nevertheless actuated by the same 'enlightened self-interest' that Malinowski deemed to have insufficient explanatory power. Issues posed by ethnographic evidence such as the potlatch would pose no trouble for Bourdieu: the destruction of economic capital would merely be a mode of investing in symbolic capital.

On the heels of Levi-Strauss's open embrace of economic reductionism, anthropological exchange theory delved into its infamous 'formalist/substantivist debate,' wherein the discipline officially segregated itself by way of the quantitative and qualitative divide. The substantivists insisted upon a non-reductionist approach, and gradually came to dominate what is today known as cultural anthropology, while only a small group of formalist anthropologists continued to follow developments in the field of economics in order to see if these theories could be applied to their own field sites (but see Geertz, 1978). Bohannan's article on *"The impact of money on an African subsistence economy"* (1959) still stands today as a centrally influential piece in solidifying the substantivist approach. In it, he follows Polanyi's radical claim (1957) that, contrary to popular beliefs, money can be enshrouded in qualitative distinctions, thereby generating a diverse range of quantitative measuring sticks. Using the Tiv of West Africa as his example, Bohannan argues that the circulation of valuables can be circumscribed by a set of complex social rules, each with its own 'sphere of exchange' dominated by its own 'special purpose money.' However, as 'modern,' 'general purpose money' arrives, it reveals a stunning capacity to traverse these rules, thereby causing all the previously segregated qualities of the different valuables to only be assessed now on a single numeric scale. Hann and Hart (2011) have recently revisited the power of Polanyi's approach to markets and exchange, and argue for some of its insights to return to debate.

Sahlins also had an enduring impact on anthropological exchange theory with his book *Stone Age Economics* (1972). Several chapters have had broad influence, but in particular, the two final chapters stand as benchmarks. In the final chapter, Sahlins compellingly maps the role of social relations onto supply and demand, thereby insisting that "Lacking a trade contact, a man may not be able to get what he wants at any price" (1972: p. 298). From this vantage point, supply and demand take a distinct backseat to the power of social relations, and price is far from being the only variable that can determine whether a market clears. In the penultimate chapter, Sahlins builds upon Malinowski's claims, in order to reveal that his old typology of exchange correlates with kinship distance. In short, "the span of social distance between those who exchange conditions the mode of exchange" (p. 196). In this enhancement of the model, one comes to see that theft is practiced exclusively upon extreme outsiders, while near kinsmen would show "voluntary food-sharing," and "the expectation of a direct material return is unseemly" (p. 194).

Hart contributed to this vision by expanding into the distinctions between barter and ceremonial exchange, arguing that "one form [ceremonial exchange] is a temporary social framework erected in the relative absence of society; the other [barter] is an atomized interaction predicated on the presence of society" (1986: p. 648). Equally, Parry and Bloch's collected volume, *Money and the Morality of Exchange* (1989), added a temporal dimension to Sahlins' spatial one: in it, they announce that research findings from diverse locales suggest that many societies perceive two distinct modalities of exchange: those that are dictated by short-term and long-term time horizons. The former tend to be viewed as more individualistic and materialistic, while the latter tend to be seen as more selfless and community spirited. Despite the impact of this tradition that maps social distance onto types of exchange, Zelizer's important research, *The Social Meaning of Money* (1997), complicates much of this. In her study of the earmarking of money within the American household, she finds that even nuclear family members can sometimes treat each other as callous exchange partners, or as potential targets of thievery.

In his book *Gifts and Commodities* (1982), Gregory represents the high-water mark of the twentieth-century tendency to separate the world into market- and nonmarket-based exchanges; in it, he succinctly proposes a dichotomous definition for gifts and commodities: the former stand as "*inalienable* objects transacted by non-aliens," while the latter represented "*alienable* objects transacted by aliens" (1982: p. 43, emphasis in original). However, if his 1982 book was unfairly misunderstood by its critics, his later contribution, *Savage Money: The Anthropology and Politics of Commodity Exchange* (1997), has been broadly influential to students of anthropological exchange theory. In it, he deftly blends the tradition of political economy from Smith to Marx to Sraffa with the anthropological study of exchange from Malinowski to Levi-Strauss to Sahlins. One chapter productively re-imagines the ubiquitous 'quantity theory of money' as a 'quality theory of money,' while yet another forces the reader to question a centuries-old and well-nourished hatred of the village usurer. This latter work put to use the best of the anthropological tradition, by bringing ground-level ethnographic detail to an age-old problem. In so doing, Gregory explodes influential theories that have greatly impacted many people's lives, as they have been adopted by powerful institutions such as The World Bank.

New Directions

If the formalist/substantivist divide can be read as one that insisted upon a stark categorical divide between quantity and quality, then one might say that the last 30 years have evidenced the slow unraveling of these borders. Perhaps inspired by a Deleuzean postmodernist spirit, or perhaps by the influence of the transnational moment, anthropological exchange theory no longer dwells too much upon this divide. Instead, it has tried to move into new productive spaces – often the spaces that open up as a result of seeing the various transgressions across the quality/quantity border. These new spaces do not so much dismiss the vitality and significance of this divide, as much as they seek to find alternative pathways to different insights, since the insights from the quantity/quality divide are now well rehearsed, and in significant measure, well accepted.

In retrospect, one can see that a thicket of arguments published in the late 1980s and early 1990s set the stage for this new movement. Appadurai's much-cited chapter in his edited volume by the same name, *The Social Life of Things*, liberated a generation of scholars who had been weaned on the quantity/quality divide. Equally, it helped spur scholars to push anthropology well past the 'traditional' societies that had been its standard bailiwick, suggesting that

anthropological method and its theories of exchange could be applied fruitfully to the circulation of commodities in 'modern' societies as well. Moreover, Appadurai insisted that one not fetishize the gift economies of traditional society to the exclusion of noticing the proliferation of commodities in their midst (as Gregory had also probed). Commodities are not the special preserve of capitalism, he says. In other words, his long essay can be seen, not least, as an effort to fight against an anthropology that was "excessively dualistic" (Appadurai, 1986: p. 12). Instead, he turned to examples such as the kula to argue against the twentieth-century paradigm, explaining that there is a "difficulty in separating gift and commodity exchange even in preindustrial, nonmonetary systems, and it reminds us of the dangers in correlating zones of social intimacy too rigidly with distinct forms of exchange" (Appadurai, 1986: p. 22).

An array of scholars interested in the application of exchange theory well outside the bounds of the 'gift economy' has been highly influenced by Munn's (1986) *The Fame of Gawa* (not least, Appadurai, whose text hinges on her arguments). Munn brilliantly altered the focus of inquiry concerning exchange by demanding to know the meaning of 'value' itself, and how it is constituted and sustained, rather than assuming it as a given. Consequently, this text, especially when read in its entirety, provides a supple and nuanced incursion into decades of literature on exchange. Munn argues that 'value' is nothing more, nothing less, than the ability of subjects to move through space and time. The successful deployment of value thereby reinforces spatial and temporal pathways through which others can move as well. Contrariwise, the destruction of value causes the shrinkage of these same intersubjective pathways. Seen from this light, 'value' is a sort of metaphysical potency that cannot be denominated, but can nevertheless be studied as part of the vital effort on the part of subjects as they both cooperate and compete in order to build viable communities. A community without value devolves into atomism, just as a community that robustly allows for the increase in value will be gain in strength.

Strathern's (1988) *Gender of the Gift* likewise transformed questions concerning exchange by shining light upon the Western underpinnings of exchange theory itself: "I choose to show the contextualized nature of indigenous constructs by exposing the contextualized nature of analytical ones. This requires that the analytical constructs themselves be located in the society that produced them" (1988: p. 8). Most radically, Strathern questioned something at the very heart of a century of exchange theory – the supposed antinomy between society and individual. To do so, she reintroduced a term from Marriott (1976), 'dividual,' in order to point out that our own belief that a person is a bounded and pure 'individual' that can be circumscribed from other individuals is a Western social construct, not a universal truth about human life per se. Strathern then goes on to insist upon the concept of 'mediated' versus 'unmediated' exchanges, in order to discuss Melanesian notions of personhood that often rest upon circulating objects that are seen as 'partible' aspects of the giver. The object, therefore, does not carry a 'hau' as in the old Maussian argument, but rather, the valued object or substance is viewed *as an actual part* of the giver himself or herself. As Strathern makes clear, this proposal greatly undermines the common belief that alienation of persons and objects serves as a pillar of exchange itself, as described above. Objects are, instead, also subjects that help forge the actual personhood of givers and receivers as exchange continues apace.

Weiner (1992) intervened powerfully in exchange theory at this time as well, by building upon its long-standing neglect of 'inalienable possessions,' that is, the valued objects that are held in reserve and only rarely step out into the world of trade. She claimed that the mundane exchange of replicable and alienable goods only occurs as part of a competition (or, in Appadurai's terms, a "tournament of values") that is organized around the disputed ownership of one unique, inalienable good. Thus, exchange partners ranged in a hierarchy are all battling each other in order to wrest away control of an inalienable good hoarded by their more powerful trade partner. 'Inalienability' is only a relative concept here, since she readily admits that often these inalienable goods (e.g., family heirlooms) are only less *likely* to move into the sphere of exchange, rather than that they will never do so. In this sense, Weiner's text is intriguing for its unintentional, but harmonic, convergence with a long range of thought from economics with regard to fractional reserve banking, Gresham's law, and hoarding. Thus, with Weiner, one can see that anthropology is perhaps trending back toward the economic theory that it had banished back with the discipline's original emergence. Perhaps this is not so surprising, since Weiner herself insists that "It is imperative that we sever what is left of the Enlightenment's partition of humanity into two camps – the 'primitive' and the West. To achieve this goal we must recognize that creating inalienable possessions is a human endeavor, universally practiced, and that local manifestations, as diverse as they are, can be compared" (1992: p. 155).

At least some of Guyer's many contributions can be viewed in light of this trend as well. While still critical of much of economics literature, she still represents the beginning of a promising trend that envisions a more robust conversation with economics. Although her method is rigorously ethnographic and she resolutely refuses to import any reductive modeling, her analysis is influenced by ideas from within economics that anthropological theorists have often tried to dismiss in the past. For example, she embraces the idea that many social actors can be governed by the seeking of profit via 'instrumental calculation,' although she believes that it blossoms into something far more variable and complex than standard economic models would allow (Guyer, 2004: pp. 17–18, 25). She has also productively intervened in the age-old discussion concerning spheres of exchange (see above), by focusing on 'asymmetric exchange.' Even though much writing on exchange since Adam Smith has assumed that exchange only occurs when items of equal value are traded, Guyer probes exchanges where the goods traded are of manifestly unequal values. With this focus, the constitution of socially mediated space and time, and what economists refer to as 'path dependency,' comes more to the fore. To wit, "without questioning Bohannon's ethnography at all," Guyer can reinterpret the spheres of exchange argument by mapping it onto a network of spatially organized and hierarchical trade

networks (2004: p. 28). What initially appears as a localized moral coding involving the purity of various spheres can now be seen, instead, as the asymmetric flow of different objects of value across different political–economic spaces.

Similar to Guyer's reminder concerning asymmetrical exchanges, Roitman's study of exchange practices in the Chad Basin, entitled *Fiscal Disobedience* (2005), has provoked a new look at some age-old categories from within the literature. First, Roitman studies the shifting moral valences of debt, thereby complicating old folk – and analytic – beliefs about the moral and economic dangers of it. Likewise, she has challenged a central feature of much study of exchange in anthropology – the purported divide between formal and informal sectors of the economy. First coined by Hart (1973), the formal/informal divide relies on the idea that the state is involved only in the formal market, whereas the informal market is somehow 'outside' the state's gaze and management. Instead, much in keeping with the Deleuzean trends of recent years, Roitman sees this binary opposition as a zone of fruitful transgression rather than a clear-cut and obvious ideal type, and she reveals how the state plays a significant role in the informal sector, despite its supposed neglect. In so doing, she finds evidence that even the seemingly 'unregulated' layers of exchange are, indeed, regulated, but merely by a widely contested field of social actors.

Further breakdowns of straightforward binarisms such as quantity/quality, credit/debt, state/market, subject/object, real/artificial, gift/commodity, and market/nonmarket are characteristic of current thinking in anthropological exchange theory. Increasingly, anthropologists are studying the social constitution of such binarisms rather than presuming them as given; equally, they are confronting the many zones in which these supposedly stable categories bleed in and out of one another (intriguingly, these advances can also be seen in economics, in such developments as Behavioral Economics). Zaloom (2006), for example, finds that the most emphatically market-oriented zone – the Chicago Board of Trade – is riddled with social relationships. With fine-tuned attention to such things as the quantitative movements of the yield curve, Zaloom instead studies the way in which quantity and quality work together, rather than stand apart, in everyday exchange. Chu (2010), meanwhile, reminds us that people might even see exchange transactions as transcending death itself, as people seek out sacrificial exchanges with their ancestors. Shipton also pivots upon this issue, and his recent trilogy (2007, 2009, 2010) provides an excellent example of how exchange theory can provide new insights while still retaining a foundational interest in the discipline's original questions concerning alienation and property.

Maurer has boldly followed the tendency to transcend these binarisms until one witnesses their complete unraveling. By combining analyses of three institutions that might ordinarily be studied separately – offshore finance, Islamic finance, and the local currency movement – Maurer argues that the traditional focus that saw exchange theory grounded in the notion of 'adequation' needs to be redirected. Adequation relies upon Saussure's axiomatic belief in a harmonic overlap between sign and object; as such, it lays the groundwork for the idea that objects' value can be compared and made equivalent by way of a clear, quantitative sense of price. But just as the poststructural critique undermines this axiom within linguistics itself, so too does Maurer believe that exchange theory needs to move away from it. One can do so, he shows, by studying exchanges wherein adequation has failed – much like when Guyer insists upon studying asymmetric exchanges – and people continue to exchange despite their acknowledgment that there is no trade of equivalents. Indeed, he labels the concept of equivalence itself as "part of the fantasy work of fetishism" (Maurer, 2005: p. 112), just as he dismisses the search for a theory of value as an old modernist proclivity that should be abandoned (Maurer, 2005: pp. 21–22). If one can say, then, that much of anthropological exchange theory was built upon Saussure's axiom, then Maurer's work to integrate more recent interventions in linguistics by the likes of Deleuze, Derrida, and Zizek could mark a new turning point in how anthropology assesses exchange.

Conclusion

In spite of some of these recent movements toward more cross-fertilization, anthropology's ongoing debates with economics will hopefully never cease, since such contention can be extraordinarily productive for all parties. For example, Graeber (2010) has recently taken economics to task for its centuries-old failure to account for the vast empirical record regarding the origin of money in credit, rather than barter. Another exciting development born out of this tension is the welter of research emerging out of the Institute for Money, Technology and Financial Inclusion (IMTFI) (see Schwittay, 2011). By relying on the qualitative ethnographic method, researchers at this institute have reframed a highly consequential debate about the global poor's lack of access to financial services. Instead of assuming that change comes from researchers, think tanks, or corporations, IMTFI assumes that the end users of the financial services can themselves be the financial innovators, inventing and utilizing new forms of money and accounting that emerge from grassroots practice. But no matter the topic of study, the enormous changes in the world economy in the past 10 years ensure that anthropological practice and theory will continue to evolve well into the future, working off of the recent advances described here.

See also: Alienation, Sociology of; Economic Anthropology; Informal Sector; Informal and Underground Economics; Money, Sociology of; Money: Anthropological Aspects; Potlatch in Anthropology; Property Ownership; Property Rights, History of; Property: Anthropological Aspects; Trade and Exchange, Archaeology of.

Bibliography

Appadurai, Arjun (Ed.), 1986. The Social Life of Things: Commodities in Cultural Perspective. Cambridge University Press, Cambridge, UK.

Benjamin, Walter, 1999. The Arcades Project. Belknap Harvard University Press, Cambridge, MA.

Boas, Fraz, 1966. In: Codere, Helen F. (Ed.), Kwakiutl Ethnography. University of Chicago Press, IL.
Bohannan, Paul, 1959. The impact of money on an African subsistence. The Journal of Economic History 19 (4), 491–503.
Bourdieu, Pierre, 1987 [1984]. Distinction: A Social Critique of the Judgement of Taste. Harvard University Press, Cambridge, MA.
Bourdieu, Pierre, 1991. In: Thompson, John B. (Ed.), Language and Symbolic Power: The Economy of Linguistic Exchanges. Harvard University Press, Cambridge, MA.
Chu, Julie Y., 2010. Cosmologies of Credit: Transnational Mobility and the Politics of Destination in China. Duke University Press, Durham, NC.
Geertz, Clifford, May 1, 1978. The bazaar economy: information and search in peasant marketing. The American Economic Review 68 (2), 28–32.
Graeber, David, 2010. Debt: The First 5,000 Years. Melville House, Brooklyn, NY.
Gregory, C.A., 1982. Gifts and Commodities. Academic Press, UK, London.
Gregory, C.A., 1997. Savage Money: The Anthropology and Politics of Commodity Exchange. Psychology Press.
Gregory, C.A., 2001. Exchange in anthropology. In: Smelser, Neil J., Baltes, Paul B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Elsevier, UK, Oxford.
Guyer, Jane I., 2004. Marginal Gains: Monetary Transactions in Atlantic Africa. University of Chicago Press, IL.
Hann, Chris, Hart, Keith, 2011. Economic Anthropology. Polity.
Hart, Keith, March 1, 1973. Informal income opportunities and urban employment in Ghana. The Journal of Modern African Studies 11 (1), 61–89.
Hart, Keith, December 1, 1986. Heads or tails? Two sides of the coin. Man 21 (4), 637–656.
Levi-Strauss, Claude, 1969 [1949]. Elementary Structures of Kinship (J.H. Bell, J.R. von Sturmer, R. Needham, Trans.). Beacon, Boston. From French.
MacCormack, Carol, Strathern, Marilyn (Eds.), 1980. Nature, Culture and Gender. Cambridge University Press, MA.
Malinowski, Bronisław, 1961 [1922]. Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagoes of Melanesian New Guinea. Dutton, New York.
Marriott, McKim, 1976. Hindu transactions: diversity without dualism. In: Kapferer, B. (Ed.), Transaction and Meaning. ISHI Publications, Philadelphia, PA.
Maurer, Bill, 2005. Mutual Life, Limited: Islamic Banking, Alternative Currencies, Lateral Reason. Princeton University Press, NJ.
Mauss, Marcel, 1990 [1954]. The Gift: The Form and Reason for Exchange in Archaic Societies. Routledge, London.
Munn, Nancy D., 1986. The Fame of Gawa: A Symbolic Study of Value Transformation in a Massim (Papua New Guinea) Society. Cambridge University Press, Cambridge, UK.
Parry, Jonathan, Bloch, Maurice P. (Eds.), 1989. Money and the Morality of Exchange. Cambridge University Press, Cambridge, UK.
Polanyi, Karl, 1957. Trade and Market in the Early Empires: Economies in History and Theory. Free Press.
Rubin, Gayle, 1975. The traffic in women. In: Reiter, Rayna R. (Ed.), Toward an Anthropology of Women. Monthly Review Press, New York, pp. 157–210.
Ricardo, David, 1821. On the Principles of Political Economy, and Taxation. John Murray, London.
Roitman, Janet, 2005. Fiscal Disobedience: An Anthropology of Economic Regulation in Central Africa. Princeton University Press, Princeton, NJ.
Sahlins, Marshall David, 1972. Stone Age Economics. Transaction Publishers, New York.
Schwittay, Anke F., December 1, 2011. The financial inclusion assemblage: subjects, technics, rationalities. Critique of Anthropology 31 (4), 381–401.
Shipton, Parker MacDonald, 2007. The Nature of Entrustment: Intimacy, Exchange, and the Sacred in Africa. Yale University Press, New Haven, CT.
Shipton, Parker MacDonald, 2009. Mortgaging the Ancestors: Ideologies of Attachment in Africa. Yale University Press, New Haven, CT.
Shipton, Parker MacDonald, 2010. Credit between Cultures: Farmers, Financiers, and Misunderstanding in Africa. Yale University Press, New Haven, CT.
Smith, Adam, 1976 [1776]. An Inquiry into the Nature and Causes of the Wealth of Nations. University of Chicago Press, IL.
Strathern, M., 1988. The Gender of the Gift: Problems with Women and Problems with Society in Melanesia. University of California Press, Berkeley.
Weiner, Annette B., 1992. Inalienable Possessions: The Paradox of Keeping-While-Giving. University of California Press, Berkeley.
Zaloom, Caitlin, 2006. Out of the Pits: Traders and Technology from Chicago to London. University of Chicago Press, Chicago.
Zelizer, Viviana A. Rotman, 1997. The Social Meaning of Money: Pin Money, Paychecks, Poor Relief, and Other Currencies. Princeton University Press, NJ.

Exchange: Social

Karen S Cook, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social exchange theory is one of the major theories of social interaction in the social sciences. Homans, Blau, and Emerson were the key theorists who developed the original theories of social exchange. Theoretical and empirical developments include the extension of their work to the analysis of power and dependence, social networks, reciprocity, fairness, social cohesion, and solidarity. The work on social exchange in sociology has clear links to research on social capital, particularly work on networks, norms, and trust. Contributors to this tradition of research include Cook, Friedkin, Lawler, Molm, Willer, Yamagishi, and their collaborators.

Social exchange theory is one of the primary orientations to the analysis of social interaction and social structure. Social relations are viewed in terms of the primacy of the costs and benefits exchanged in interaction. Social structures are viewed as networks of connected social relations between individuals or groups, sometimes acting through agents. The behavioral basis of the theory varies depending on the specific theorist. Homans based his original formulation on the behaviorist tradition, as did Molm in her early work. Homans examined the reinforcement principles that play out in the behavior of actors in interaction with one another. Blau eschewed behaviorism and based his original formulation on cognitive concepts from psychology and microeconomics: preferences, interests, indifference curves, supply, and demand. Subsequent formulations, including the work of Emerson and Cook, tended to blend aspects of these two foundations. Initially a major focus of exchange theory was the study of power and its structural bases. The power-dependence perspective first developed by Emerson to apply to dyads was later extended to apply to the study of exchange networks. In Emerson's (1972) work coalitions, division of labor, specialization, and related processes were analyzed in terms of power dynamics (e.g., Cook and Gillmore, 1984). Some of the major developments in exchange theory include the analysis of power, networks of exchange relations, and processes of structural change and applications of these principles. More recent developments include the study of commitment, trust, and solidarity.

Social exchange theory has its origins in the field of sociology in the work of George C. Homans and Peter Blau. Homans' (1958) famous paper entitled 'Social Behavior as Exchange' laid the foundation of this orientation to the study of behavior. He subsequently expanded these ideas in *Social Behavior: Its Elementary Forms*, published in 1961 (1974). The primary focus of his work is what he referred to as 'subinstitutional' or elementary behavior, the actions of individuals in direct interaction with one another that form the bedrock of groups and organizations. He contrasted elementary behavior with institutionalized forms of behavior, such as that involved in conforming to norms or to role prescriptions for appropriate behavior. For Homans, institutionalized behavior and elementary forms of behavior were most often distinct. He argued that elementary behavior could sometimes 'crack the institutional crust,' forcing changes in the institutionalized ways of doing things (e.g., the changes in social behavior and institutions forged by rebellions and revolutions). This orientation set his work apart from some of the dominant thinking at the time, particularly the work of Parsons.

Major Theoretical Formulations: Blau and Homans

Blau (1964), in his famous book, *Exchange and Power in Social Life*, developed a much more extensive treatment of the links between microlevel social behavior and the groups, organizations, and institutions it constitutes. A major difference between the perspectives of Homans and Blau is the latter's recognition of 'emergent' processes at more complex levels of social organization. For Blau, the whole is more than the sum of its parts. Relationships between the parts or elements of the structure create emergent processes that evolve from the interaction of the parts, but are not reducible to properties of the individual elements. Homans, in contrast, adopted a reductionist orientation to explanation using exchange principles arguing that many of the 'laws' of social organization could be understood in terms of the behavior of the interacting parts, typically individuals. For this reason, according to Homans, psychological principles ultimately laid the foundation for the laws of elementary social behavior.

While Homans' efforts were aimed at understanding elementary behavior, Blau's aim was much more macro in focus. He intended to formulate a theory of social structure and institutions based on a sound microfoundation. He presented his book on social exchange as a 'prolog' to a theory of social structure. The origin of Blau's vision of social behavior was not behaviorism, as in Homans' case; rather, it was a more sociological version of microeconomics, which he labeled 'social exchange theory,' borrowing the term from Homans. (For a lengthy critique of Blau's use of microeconomic reasoning see Heath, 1976, among others.) In many ways, Blau's work was one of the first major theoretical efforts to extend the logic of microeconomics to the analysis of social behavior, a task that much later won Gary Becker the Nobel Prize in Economics. Homans instead based his analysis of the social behavior he called 'exchange' on principles of behavioral analysis derived

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32056-6

from the work of the behaviorists famous at the time he wrote. A major influence on his theoretical work was the empirical research of B.F. Skinner, his well-known colleague at Harvard.

Both Homans and Blau addressed indirect exchange processes, though such processes form a much more significant component of Blau's theoretical formulation. Homans acknowledged the importance of the indirect ties that actors have when they are in direct exchange with the same sets of individuals. For example, employees of the same employer have indirect exchange relations with each other through their direct contacts with their employer. Blau extended the analysis of indirect exchange, acknowledging the role it plays in larger and more complex collectivities as well as in collective action. In particular, Blau viewed indirect exchange as the basis for coordinated action in complex social structures (and the substructures of which they are composed). This coordination of action through indirect exchange in larger collectivities necessitates mediation by common values, which can be the focus of opposition and conflict, especially as the social structures involved become more complex. Blau discussed group formation, cohesion, integration, differentiation, opposition, conflict, and dissolution as significant forces of structural change and stability in complex social structures. These forces played a major role in his theory of social exchange as he endeavored to provide a comprehensive microfoundation for macrosociology. Homans and Blau were both engaged in distinctive efforts to produce a coherent theoretical alternative to Parsonian functionalism, the dominant paradigm in sociology until the late l960s.

The Roots of Social Exchange Theory

Somewhat independently, Homans and Blau developed their theories of social exchange even though each had been well versed in functionalism and Parsons' social systems approach to the analysis of society. The vestiges of both traditions are clearly evident in Blau's treatment of social exchange, though much less so in the work of Homans who labored skillfully to ground his theory of exchange explicitly on Skinnerian principles of behavioral analysis, repudiating the need for more macro-level concepts and principles. Blau's goal was to produce a theory of social structure by clarifying the nature of the social processes and associations that generate the more complex structures of organizations, communities, and societies. He explicitly attempted to link the sociology of everyday life popularized in the work of Simmel and Goffman with the broad theories of society characteristic of Weber and Parsons. Unlike Homans, however, his focus on the interpersonal was primarily as the basis for theorizing about more complex forms of association.

Psychological Roots

Besides functionalism and Parsons' version of systems theory other influences on the development of social exchange theory came from an eclectic set of sources including pragmatism, utilitarianism, behaviorism (for Homans), and conflict theory (in the case of Blau's work). Turner (1986) provides a fairly accurate and detailed account of the various roots of social exchange theory in his discussion of the development of this theoretical orientation and its impact in the field of sociology. In addition to these sources of influence, the major works of several psychologists and anthropologists provided further impetus to the development of the exchange paradigm in sociology. The dominant contribution from psychology is the work of John Thibaut and Harold Kelley (1959), *The Social Psychology of Groups*. This volume presents a lengthy treatment of the forms of interdependence among individuals engaged in various types of social interaction. In matrix format it lays out the rewards and costs attached to different behaviors under varying conditions of interdependence. The framework is used to make predictions about behavior given different types of control over the outcomes of the other actor(s) in the situation. For example, Thibaut and Kelley (1959) distinguish behavior control from fate control in social relationships using their basic orientation to interdependence. Their work formed the basis of much experimental research in psychology over the next several decades primarily on cooperation, competition, negotiation, and coalition formation. It also dovetailed nicely at that time with the emergence of game theory in the social sciences as a tool for understanding strategic interactions, particularly in psychology, economics, and political science.

Anthropological Roots

The early work of anthropologists like Levi-Strauss (1969 [1922]), Malinowski (1922), and Mauss (1925) influenced exchange theory in an entirely different way. Sociologists drew upon the fieldwork of these scholars to provide examples of the significance of social exchange, often contrasted directly with economic exchange. Malinowski (1922), for example, in his comprehensive analysis of the Kula Ring among the Trobriand Islanders, argued that the exchange of arm bands for necklaces across the islands over time was an important source of social solidarity, the significance of which far exceeded the economic value of the items exchanged. Blau (1964), and subsequently Richard Emerson (1972, 1981), developed conceptions of indirect exchange and generalized exchange based to some extent on Malinowski's extensive fieldwork. Mauss's (1925) complex analysis of the gift also provided input to the development of theory concerning generalized exchange and reciprocity (see Molm, 1997; Molm et al., 1999 on reciprocal or non-negotiated exchange; Yamagishi and Cook, 1993; Bearman, 1997; Molm et al., 2007 on generalized exchange).

Exchange Theory: Empirical Research

Homans and Blau popularized exchange theorizing in sociology in the late 1960s and early 1970s, but sustained empirical research on the topic did not begin until the mid-1970s largely as a result of the influence of Emerson's (1972) more formal theoretical work, based on his earlier treatment of power-dependence relations (1962, 1964). Ironically, Emerson's work on power-dependence relations also influenced Blau's (1964) discussions of power and power-gaining strategies. Emerson built a conception of social exchange around his fundamental insights concerning

power in social relations. He, like Blau, made power and inequality central to his treatment of exchange processes. Similarly, like Blau, he viewed his theory of social exchange as the beginning steps toward formulating a general theory of social structure. In Emerson's case, however, the structures of primary interest were the networks formed as a result of the connections among a set of actors engaged in social exchange with one another. Two of Emerson's distinct contributions to exchange theory in sociology are his fundamental insight into the relational nature of power and his extension of exchange theory to analyze the social networks created by exchange relations.

Before expanding on these two major contributions to the development of exchange theory it is worth noting that Emerson's first efforts to formulate a theory of social exchange had more in common with Homans's work than that of Blau, despite their common goal of building a theory of social structure. In the two papers written in 1967 (and subsequently published in 1972) in which Emerson laid out the foundations of his theory of exchange, he formulated a model of the actor built on principles of behavioral analysis adopted from Skinner via Homans. Key concepts include reward (positive reinforcement), cost (rewards foregone), alternative exchange relations, dependence, and balance. Propositions are formulated based on the reinforcement notions of behavior first introduced into exchange theory by Homans and subsequently developed not only by Emerson, but also in the early work of Molm (e.g., Molm and Wiggins, 1979; Molm, 1980).

Emerson's Theory of Exchange Relations

A central focus of Emerson's theoretical framework is the exchange relation as the unit of analysis. As Turner (1986) notes, this move made exchange theory less vulnerable to charges of tautology and reductionism. Emerson's goal was to present a theory of the formation, maintenance, and termination of exchange relations and the forms of social structure created by connected exchange relations (e.g., networks of exchange). Less critical in this theoretical formulation is the precise nature of the values of the actors involved or the cognitive structure of their interests. Assuming actors enter exchange relations for reasons of reward (or avoidance of negative reinforcement in alternative relations), the focus shifts to understanding aspects of the exchange relations and the structure of the networks or groups in which they are embedded. The emphasis on the social relation as the unit of analysis distinguishes Emerson's work from that of Homans and Blau and allowed Emerson to present a more coherent theory of social structure.

Power and Dependence in Exchange

The primary legacy of Emerson's work on social exchange is his conception of how power is determined by the dependence of one actor on another in an exchange relation and how dependence is structured by the nature of the connections among actors in networks of exchange opportunities that deny or provide access to exchange alternatives. The key definition of power in Emerson's formulation is: The power of actor A over actor B in the Ax:By exchange relation (where x and y represent resources of value) increases as a function of the value of y to A, and decreases proportional to the degree of availability of y to A from alternative sources (other than B). These two factors (resource value and availability) determine the level of B's dependence on A and thus A's power over B. That is, the power of A over B is a direct function of B's dependence on A in the A:B exchange relation. The more dependent B is on A, the more power A has over B. Embedding this relationship in a network of exchange opportunities creates the basis for a structural theory of power in exchange networks.

Exchange relations are connected to the extent that exchange in one relation affects or is affected by the nature of the exchange in another relation. The connection according to Emerson's work can either be positive or negative. A negative connection means that exchange in one relation reduces the amount or frequency of exchange in another exchange relation involving one of the same parties (e.g., the A–B and B–C exchange relations are negatively connected at B if exchange in the A–B relation reduces the frequency or amount of exchange in the B–C relation). A connection is positive if the amount or frequency of exchange in one relation increases the amount or frequency of exchange in an exchange relation involving at least one of the parties to both exchanges (e.g., the A–B relation is positively connected to the B–C relation if exchange in the A–B relation increases the frequency or amount of exchange in the B–C relation). These are modal cases; however, exchange in two relations may also be 'mixed,' involving both positive and negative exchange connections relating to different aspects (or dimensions) of exchange. Exchange in more complicated networks may involve both positive and negative connections (see Cook et al., 1983), often with different actors in the network of exchange relations. In addition, one form of exchange can be mixed with or embedded in networks in which exchanges of another type are dominant (see Molm et al., 2012).

Exchange Networks

Emerson (1972, 1976) (Cook and Emerson, 1978) expanded this important direction of development in exchange theory in later publications formulating what is now known as 'exchange network theory.' Markovsky et al. (1988) subsequently developed a similar perspective they called 'network exchange theory' (also referred to as 'exchange-resistance theory,' Heckathorn, 1983). This latter approach derives from Willer's (e.g., Willer and Anderson, 1981; Willer, 1999) 'elementary theory' of social relations, which has an affinity with exchange concepts, but is not derived from Emerson's power-dependence perspective. These traditions of work in recent research have much in common, despite the remaining subtle differences in assumptions about actor behavior, network connections, and bases of power. In all of these programs of research the main focus of the early empirical work was determining the structural determinants of the distribution of power in networks of varying configurations.

Early debates about the value of exchange theory in the social sciences were centered on the role of rationality, tautology, and reductionism (see Emerson, 1976). These debates gave way during the 1980s and 1990s to more sophisticated arguments in the field concerning theoretical strategy and the best ways to

develop microlevel models of exchange that inform macro-level processes. More complex theoretical formulations emerged during this period based on the results of careful empirical work, much of it experimental, testing the fundamental propositions derived from exchange theory. Emerson's formulation provided the impetus for much of this empirical work primarily because he formalized the theory to make it more precise and therefore testable. Subsequent formulations adopted a similar theoretical strategy (e.g., Markovsky et al., 1988; Bonacich, 1998; Bienenstock and Bonacich, 1992, 1993; Friedkin, 1992; Skvoretz and Willer, 1993; Whitmeyer, 1994; Yamaguchi, 1996; and, more recently, van Assen, 2003, and van Assen and van de Rijt, 2007, among others).

Key Assumptions of Exchange Theory

The key assumptions of exchange theory summarized by Molm (1997), (see also Molm and Cook, 1995: p. 210) include the following: (1) behavior is motivated by the desire to increase gain and to avoid loss (or to increase outcomes that are positively valued and to decrease outcomes that are negatively valued); (2) exchange relations develop in structures of mutual dependence (both parties have some reason to engage in exchange to obtain resources of value or there would be no need to form an exchange relation); (3) actors engage in recurrent, mutually contingent exchanges with specific partners over time (i.e., they are not engaged in simple one-shot transactions); and (4) valued outcomes obey the economic law of diminishing marginal utility (or the psychological principle of satiation). Based on these core assumptions, various predictions are made about the behavior of actors engaged in exchange and the effects of different factors on the outcomes of exchange. The power-dependence principle, in addition, allows for the formulation of predictions concerning the effects of increasing the value of the resources involved in the exchange and the availability of resources from alternate sources.

Power Inequalities in Networks

Empirical work since the 1980s established several primary facts about exchange in networks of social relations. Power inequalities in networks of exchange are determined by access to valued resources in the network. The networks structure this access by providing connections or links to certain actors and restricting access to others. The structure of the network determines the accessibility of resources and thus the nature of the distribution of power in the network (see e.g., Cook and Emerson, 1978; Cook et al., 1983; Markovsky et al., 1988; Whitmeyer, 1999; Yamagishi et al., 1988). Subsequent empirical work has extended this research, developing a number of new formalizations concerning the structure and size of the networks, the form of exchange, and the dynamic nature of the processes involved over time (e.g., Cook and Yamagishi, 1992; Markovsky, 1987, 1995; Markovsky et. al., 1993; Skvoretz and Lovaglia, 1995; Willer, 1992; Willer and Skvoretz, 1997). For example, van Assen (2003) examines a large number of networks up to size nine and demonstrates that power-dependence theory and exchange-resistance theory make similar predictions regarding the classification of the types of power exercised. Similarly, van Assen and van de Rijt (2007) analyze dynamic exchange networks, investigating the effects of adding and deleting links in the network over time on the payoffs to the actors involved and their 'neighbors.' Willer et al. (2012) have also begun exploring more general analytic procedures for extending exchange network theories to apply to large-scale networks.

In exchange relations characterized by power inequalities, the use of power clearly varies not only by network structure and size but it also varies depending on whether the actors control positive or negative outcomes. Extending the theory to deal with the exercise of coercive power within social exchange relations, Molm (1997) produced insights into the nature of power relations in situations in which the actors have not only reward power, but also punishment power. Comparing the behavioral effects of the use of different bases of power in social exchange relations, Molm demonstrated that the results that hold for the exercise of reward power do not hold for the use of punishment. Specifically, punishment power is used much less frequently, in part because it carries with it the risk of withdrawal from the relationship on the part of the target of the punishment to the extent that the relationship can be exited. (Molm focused primarily on voluntary exchange relations; that is, both parties have some degree of control over their involvement in the exchange and can exit if they so desire.) However, when punishment power is used it tends to be very effective in modifying behavior to make the relations more mutually rewarding in the long run. According to Molm (1997: p. 268): "When punishment is administered contingently and consistently, coercion is a powerful means of getting what one wants." However, coercive power is rarely used due to the risks involved.

Coercive Power

Integrating the analysis of coercive power with reward power in social exchange relations allowed for the investigation of a wider range of social situations in which more complex exchanges occur. Frequently, those involved in ongoing exchange relations have control over both rewards and punishments (even if it is only the withholding of rewarding behaviors). In addition, Molm's (1997) work made a major contribution to the analysis of the strategic use of structurally based power advantages and disadvantages. As she correctly noted, the missing link in Emerson's (1962, 1972) original work on power-dependence relations was the nature of the precise mechanisms that relate structural determinants of power with the actual use of power by those in positions of power. In this effort norms of fairness or justice and attitudes toward risk play a key role. Conceptions of fairness constrain the use of power under some conditions especially the use of coercive power, and aversion to risk makes some actors unwilling to use the structural power at their disposal for fear of loss. Fairness also affects the extent to which actors get committed to one another over time in exchange relations that differ in the degree of risk and uncertainty involved.

Risk and Uncertainty

Additional empirical research was subsequently conducted that extended the existing exchange theories to deal with the effects of

important factors such as uncertainty and risk on the nature and structure of social exchange. Facing uncertain environments, actors involved in exchange are more likely to seek to form committed exchange relations (Cook and Emerson, 1978; Kollock, 1994; Lawler and Yoon, 1996). Commitment reduces the extent to which actors seek exchange with alternatives and thus serves to reduce the power inequalities both within the exchange relation and within the network in which the relation is embedded. As Kollock's (1994) work demonstrated, uncertainty not only results in commitment formation as a means of reducing uncertainty, it also tends to be correlated with perceptions of trustworthiness of the actors involved in the exchange relations.

Yamagishi et al. (1998) reported that trust emerges in exchange relations under conditions of high uncertainty as actors form commitments to exclusive exchange with one another in an attempt to avoid the possibility of exploitation by unknown actors who enter the exchange opportunity structure. Under conditions of low uncertainty, actors are much more likely to continue to 'play the market' and to avoid forming commitments to specific partners in order to maximize their access to valued resources. (Uncertainty in these experimental studies referred to the likelihood of being exploited by a new partner in an exchange network of opportunities that changes over time. In Kollock's (1994) study, the uncertainty involved concerned the quality of the goods to be exchanged.) In research on trust, uncertainty and vulnerability are often defined as two of the key elements of situations in which trust considerations are paramount (e.g., Heimer, 2001; Cook, 2005; Cook et al., 2004). As Molm's recent work (e.g., Molm et al., 2009) demonstrates, different types of exchange vary in the nature and extent of the uncertainty and risk involved, which in turn affects the degree of trust that emerges. Reciprocal exchange, for example, typically results in greater trust than does negotiated exchange with binding agreements. It also results in more resilient, affective trust, and greater cohesiveness (Molm, 2010).

Relational Cohesion and Social Exchange

Lawler and his collaborators (e.g., Lawler and Yoon, 1993) developed a theory of relational cohesion based on principles of social exchange. This work examined the conditions under which social exchange relations develop out of opportunities for exchange and are linked to the emergence of positive emotions about the exchange relation. These positive emotions and subsequently feelings of cohesion or solidarity, Lawler argues, develop based on positive evaluations of the outcomes of the exchanges between actors and the frequency of their exchange. Low frequency and unfavorable (or less favorable) exchanges are much less likely to lead to commitment to the relation, to positive feelings about the exchange, and to feelings of cohesiveness or solidarity (i.e., what Lawler terms a 'we-feeling'). This line of research returned to some of the earlier anthropological work that examines the links between exchange and solidarity in social relations. It also integrated into existing theoretical conceptions of exchange theory important considerations of the emotional bases of exchange and commitment (see Lawler and Yoon, 1998; Lawler and Thye, 1999).

Building on his theory of relational cohesion and his affect theory of social exchange (Lawler, 2002), Lawler and his collaborators (e.g., Lawler et al., 2008) have explained how various types of exchange (reciprocal, negotiated, generalized, and productive) result in different levels of 'microsocial order,' reflecting the nature of the recurrent social interactions, emotional responses, perceptions of being a group, and affective sentiments directed at the group or network and subsequent attachment to the group. These are the elements of social order as it emerges at the microinteractional level. It is strongest when activities are joint (as is the case with productive exchange) and when the tasks involved generate a shared sense of responsibility (strongest in productive exchange and weakest in generalized exchange). The significance of this work is that it connects exchange theories explicitly to the broader study of social commitments (Lawler, 2009) and social order.

Applications of the Theory

Applications of exchange theory and power-dependence concepts extend from dyadic relations like dating and longer-term relationships including marriage (e.g., Sprecher and Metts, 1999) to often short-term relations characteristic of various practitioner–client interactions (such as physician–patient meetings under managed care, see Grembowski et al., 1998), in addition to more macro-level applications. The focus of analysis at the microlevel is often the nature of the relationship and the factors that predict satisfaction (or dissatisfaction) and longevity (or termination) of the relation. Two dimensions frequently investigated in this respect are the degree of power imbalance in the relationship and the extent to which the relationship is embedded in a network of other important exchange relationships. Power imbalance in partnerships often creates additional tension in the relation as actors try to manage the resource differences between the parties to the exchange and attempt to maintain commitments in the face of alternatives. Within an organizational structure such as that produced under managed care in the health care delivery system, choice of provider is often limited exacerbating patient concerns with their dependence on the available physicians. Mechanisms are put into place to provide patients in these settings access to redress if the power of the physician is abused. These include increased professional and utilization reviews by colleagues, administrators, and those who insure the patients. In all cases in which a power imbalance exists in an exchange relation there is the potential for exploitation, thus widely differing mechanisms have been created that serve to provide some monitoring to insure the equitable and fair treatment of those who are most dependent and thus vulnerable to the abuse of power. A prime example is the oversight demanded frequently in nursing homes that must deliver care to one of the most dependent populations in the USA, the infirm aged. Aspects of both power-dependence theory and exchange theory have proven useful in the study of power processes in many contexts, as well as in the analysis of trust and its breakdown, especially in organizations with the

potential for power abuse (e.g., those characterized by hierarchy, unequal power relations, and a lack of transparency or monitoring of those in power).

At the more macrolevel, exchange networks represent an important framework for analyzing the global transformation of corporate relations mediated by the Internet, as well as the reorganization of the workplace and the exponential increase over the past decade in interpersonal connections over the Internet (e.g., Facebook, Linked-In). The world of global business has been transformed as B2B (business to business) commerce has become a dominant mode of conducting transactions at the macrolevel. In addition, business to client (B2C) relations and P2P (peer to peer) connections have grown steadily creating a world-wide market not only for goods and services, but also for the recirculation and recycling of objects in second-hand markets. These relations represent complex networks of exchange and trade that include not only economic, but also political and social dimensions of transnational significance. For example, networks that link actors with common cause have been at the core of recent political transformations especially in the emerging democracies in various regions of the world. Future research will likely see the further expansion of exchange analyses of power and influence in global networks embedded in complex social and political realities, as well as new studies of exchange in what is now referred to as the 'sharing economy,' emerging forms of broad-based generalized exchange systems mediated by the Internet such as TaskRabbit, CouchSurfing, AirBnB, Uber, RelayRides, and FreeCycle.

See also: Conflict and War: Anthropological Aspects; Groups, Sociology of; Integration, Social; Interactionism, Symbolic; Labor, Division of; Malinowski, Bronislaw (1884–1942); Mauss, Marcel (1872–1950); Power in Society; Social Networks; Social Structure; Trust, Sociology of.

Bibliography

van Assen, M.A.L.M., 2003. Exchange networks: an analysis of all networks up to size 9. Advanced in Group Processes 20, 67–103.

van Assen, M.A.L.M., van de Rijt, A., 2007. Dynamic exchange networks. Social Networks 29 (2), 266–278.

Bearman, Peter, 1997. Generalized exchange. American Journal of Sociology 102, 1383–1415.

Bienenstock, E.J., Bonacich, P., 1992. The core as a solution to exclusionary networks. Social Networks 14, 231–243.

Bienenstock, E.J., Bonacich, P., 1993. Game-theory models for exchange networks: experimental results. Sociological Perspectives 36, 117–135.

Blau, P.M., 1964. Exchange and Power in Social Life. Wiley, New York.

Bonacich, P., 1998. A behavioral foundation for a structural theory of power in exchange networks. Social Psychology Quarterly 61, 185–198.

Cook, K.S., 2005. Networks, norms and trust: the social psychology of social capital. Social Psychology Quarterly 68 (1), 4–14.

Cook, K.S., Emerson, R.M., 1978. Power, equity and commitment in exchange networks. American Sociological Review 43, 721–739.

Cook, K.S., Emerson, R.M., Gillmore, M.R., Yamagishi, T., 1983. The distribution of power in exchange networks: theory and experimental results. American Journal of Sociology 89, 275–305.

Cook, K.S., Gillmore, M., 1984. Power, dependence, and coalitions. In: Lawler, E.J. (Ed.), Advances in Group Processes, vol. 1. JAI, Greenwich, CT, pp. 27–58.

Cook, K.S., Hardin, R., Levi, M., 2004. Cooperation without Trust? Russell Sage Foundation, New York.

Cook, K.S., Yamagishi, T., 1992. Power in exchange networks: a power-dependence formulation. Social Networks 14, 245–265.

Emerson, R.M., 1962. Power-dependence relations. American Sociological Review 27, 31–41.

Emerson, R.M., 1972. Exchange theory, part I: a psychological basis for social exchange; exchange theory, part II: exchange relations and network structures. In: Berger, J., Zelditch Jr, M., Anderson, B. (Eds.), Sociological Theories in Progress, vol. 2. Houghton Mifflin, pp. 38–87.

Emerson, R.M., 1976. Social exchange theory. Annual Review of Sociology 2, 335–362.

Emerson, R.M., 1981. Social exchange theory. In: Rosenberg, M., Turner, R. (Eds.), Social Psychology: Sociological Perspectives. Basic Books, New York, pp. 30–65.

Friedkin, N.E., 1992. An expected value model of social power: predictions for selected exchange networks. Social Networks 14, 213–229.

Grembowski, D., Cook, K.S., Patrick, D., Roussell, A., March 1998. Managed care and physician referral: a social exchange perspective. Medical Care Research and Review.

Heath, A.F., 1976. Rational Choice and Social Exchange: A Critique of Exchange Theory. Cambridge University Press, Cambridge, UK.

Heckathorn, D., 1983. Extensions of power-dependence theory: the concept of resistance. Social Forces 61, 1206–1231.

Heimer, C., 2001. Solving the problem of trust. In: Cook, K.S. (Ed.), Trust in Society. Russell Sage Foundation, New York, pp. 40–88.

Homans, G.C., 1958. Social behavior as exchange. American Journal of Sociology 62, 597–606.

Homans, G.C., 1961 [1974]. Social Behavior: Its Elementary Forms, second ed. Harcourt, Brace, & World, New York.

Kollock, P., 1994. The emergence of exchange structures: an experimental study of uncertainty, commitment and trust. American Journal of Sociology 100, 313–345.

Lawler, E.J., 2001. An affect theory of social exchange. American Journal of Sociology 107, 321–352.

Lawler, E.J., 2002. Micro social orders. Social Psychology Quarterly 65 (1), 4–17.

Lawler, E.J., 2009. Social Commitments in a Depersonalized World. Russell Sage Foundation, New York.

Lawler, E.J., Thye, Shane R., 1999. Bringing emotions into social exchange theory. Annual Review of Sociology 25, 217–244.

Lawler, E.J., Thye, Shane R., Yoon, Jeongkoo, 2008. Social exchange and micro social order. American Sociological Review 73, 519–542.

Lawler, E.J., Yoon, J., 1993. Power and the emergence of commitment behavior in negotiated exchange. American Sociological Review 58, 465–481.

Lawler, E.J., Yoon, J., 1996. Commitment in exchange relations: test of a theory of relational cohesion. American Sociological Review 61, 89–108.

Lawler, E.J., Yoon, J., 1998. Network structure and emotion in exchange relations. American Sociological Review 63, 871–894.

Levi-Strauss, C., 1969 [1922]. The Elementary Structures of Kinship, revised ed. Beacon, Boston.

Lovaglia, M.J., Skvoretz, J., Willer, D., Markovsky, B., 1995. Negotiated exchanges in social networks. Social Forces 74, 123–155.

Malinowski, B., 1922. Argonauts of the Western Pacific. Routledge, London.

Markovsky, B., 1987. Toward multilevel sociological theories: simulations of actor and network effects. Sociological Theory 5, 101–117.

Markovsky, B., 1995. Developing an exchange network simulator. Sociological Perspectives 38, 519–545.

Markovsky, B., Skvoretz, J., Willer, D., Lovaglia, M.J., Erger, J., 1993. The seeds of weak power: extending network exchange theory. American Sociological Review, 197–209.

Markovsky, B., Willer, D., Patton, T., 1988. Power relations in exchange networks. American Sociological Review 53, 220–236.

Mauss, M., 1925. Essai sur le Don en Sociologie et Anthropologie. Presses Universitaires de France, Paris.

Molm, L.D., 1980. The effects of structural variations in social reinforcement contingencies on exchange and cooperation. Social Psychology Quarterly 43, 269–282.

Molm, L.D., 1997. Coercive Power in Social Exchange. Cambridge University Press, Cambridge, UK.

Molm, L.D., 2010. The structure of reciprocity. Social Psychology Quarterly 73, 119–131.

Molm, L.D., Cook, K.S., 1995. Social exchange and exchange networks. In: Cook, K.S., Fine, G.A., House, J.S. (Eds.), Sociological Perspectives on Social Psychology. Allyn and Bacon, Needham Heights, MA, pp. 209–235.

Molm, L.D., Peterson, G., Takahashi, N., 1999. Power in negotiated and reciprocal exchange. American Sociological Review 64, 876–890.

Molm, L.D., Collett, J.L., Schaefer, D.R., 2007. Building solidarity through generalized exchange: a theory of reciprocity. American Journal of Sociology 113, 205–242.

Molm, L.D., Schaefer, D.R., Collett, J.L., 2009. Fragile and resilient trust: risk and uncertainty in negotiated and reciprocal exchange. Sociological Theory 27 (1), 1–32.
Molm, L.D., Whitham, M.M., Melamed, D., 2012. Forms of exchange and integrative bonds: effects of history and embeddedness. American Sociological Review 77, 141–165.
Molm, L.D., Wiggins, J.A., 1979. A behavioral analysis of the dynamics of social exchange in the dyad. Social Forces 57, 1157–1179.
Skvoretz, J., Lovaglia, M.J., 1995. Who exchanges with whom: structural determinants of exchange frequency in negotiated exchange networks. Social Psychology Quarterly 58, 163–177.
Skvoretz, J., Willer, D., 1993. Exclusion and power: a test of 4 theories of power in exchange networks. American Sociological Review 58, 801–818.
Sprecher, S., Metts, S., 1999. Romantic beliefs: their influence on relationships and patterns of change over time. Journal of Social and Personal Relationships 16, 834–851.
Thibaut, J.W., Kelley, H.H., 1959. The Social Psychology of Groups. Wiley, New York.
Turner, J., 1986. The Structure of Sociological Theory, fourth ed. Dorsey, Chicago.
Whitmeyer, J.M., 1994. Social structure and the actor: the case of power in exchange networks. Social Psychology Quarterly 57, 177–189.
Whitmeyer, J.M., 1999. Interest-network structures in exchange networks. Sociological Perspectives 42, 23–47.
Willer, D., 1992. Predicting power in exchange networks: a brief history and introduction to the issues. Social Networks 14, 187–211.
Willer, D., 1999. Network Exchange Theory. Praeger, Westport, CT.
Willer, D., Anderson, B., 1981. Networks, Exchange and Coercion: The Elementary Theory and Its Applications. Elsevier, New York.
Willer, D., Skvoretz, J., 1997. Games and structures. Rationality and Society 9, 5–35.
Willer, D., van Assen, M.A.L.M., Emanuelson, P., 2012. Analyzing large scale exchange networks. Social Networks 34 (2), 171–180.
Yamagishi, T., Cook, K.S., 1993. Generalized exchange and social dilemmas. Social Psychology Quarterly 56, 235–248.
Yamagishi, T., Cook, K.S., Watabe, K., 1998. Uncertainty, trust and commitment formation in the United States and Japan. American Journal of Sociology 104, 165–194.
Yamagishi, T., Gillmore, M.R., Cook, K.S., 1988. Network connections and the distribution of power in exchange networks. American Journal of Sociology 93, 833–851.
Yamaguchi, K., 1996. Power in networks of substitutable and complementary exchange relations: a rational-choice model and an analysis of power centralization. American Sociological Review 61, 308–332.

Executive Functions during Childhood, Development of

Sanja Šimleša and Maja Cepanec, University of Zagreb, Zagreb, Croatia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Executive functions (EFs) encompass higher order processes required for purposeful, goal-oriented activity. The processes most frequently included in this term are inhibition, working memory, cognitive flexibility, and planning. More recent studies show that EF processes have different developmental trajectories and start their intensive development at different periods. This article presents definitions of EF, changes in the concept and studies of EF during time, different development trajectories of EF and related changes in the brain through three developmental periods: (1) infants and toddlers, (2) early childhood, and (3) middle childhood, and finally methodological issues in the studies of EF in children.

Definitions of Executive Functions

Executive functions (EFs) play an important role in the individual's development, from language development, imaginative play and theory of mind in childhood, academic success in adolescence, to career and family planning and organization in young adulthood. Any alterations in the development of EFs, due to biological or environmental factors, reflect on the individual's social, academic, and professional development and social status. Although authors define the concept of EFs in different ways, they all agree that EFs encompass higher order processes required for purposeful, goal-oriented activity and adaptive responses to novel or complex situations (Anderson, 2002). In general, the processes most frequently mentioned in the literature are response inhibition, working memory, cognitive flexibility, and planning. According to some authors, this group also includes temporal organization of behavior, goal anticipation, activity initiation, activity monitoring, and response evaluation.

Changes in the Focus or Emphasis over Time

Changes in the Concept of EF (Unitary vs. Nonunitary Concept of EF)

In the early studies, EFs were viewed as a homogenous construct, i.e., a unitary, central EF rather than a number of different processes encompassed by the same term. More recent studies define EFs as an encompassing term including a whole range of interrelated processes required for a purposeful, goal-oriented behavior. This is evidenced by a number of findings from various disciplines: (1) from the clinical point of view, patients rarely suffer from a general executive dysfunction; (2) processes encompassed by the term EFs are localized in different areas of the prefrontal cortex; (3) measures of different aspects of EFs show low correlations; (4) factor analyses point to the existence of more factors; and (5) different processes of EFs have different developmental trajectories.

In their attempt to answer the question whether EFs form a unitary construct or encompass different, isolated processes, Miyake et al. (2000) found that correlations between different tasks that assess inhibition, cognitive set shifting, and working memory ranged from 0.4 to 0.6. Such relatively high correlations indicate that these are different constructs that partially share the common basis. The basis of these moderately high correlations according to some authors is inhibitory control, according to others it is working memory, and according to a more complex approach it is both inhibitory control and working memory.

Changes in the Studies of EF over Time

Until recently, EFs were exclusively studied in the frame of adult neuropsychology because it was believed that the frontal lobe remains 'functionally silent' until adolescence, which is why the tasks intended for EF assessment were appropriate for adults only. This led to neglecting of the research on developmental changes in the domain of EFs. According to the Scopus database, the first studies on EFs appeared in the 1970s, and from 1980s to 2000s there was an increase in the number of studies on EFs in children. While in the first half of the 1980s, there was only a single study on EFs that included the term 'children,' in the 1990s the share of such studies considerably increased, to as much as 22% from 1996 to 2010 (see Table 1).

With the increase in the number of studies, there also came advances in the methodological approach. Methodological advances brought about new findings on the maturation of the central nervous system and related cognitive development. Some of the EF aspects were at first intensively researched in animals, in order to experimentally examine the relationship

Table 1 Results from the Scoups database showing the number of studies with the keyword 'executive functions' (EFs), the number of studies with the keywords 'executive functions' and 'children,' and their ratio between 1981 and 2010

	Number of studies		*Percentage studies of EF involving keyword 'children'*
Years	*'EF'*	*'EF' & 'children'*	
1981–1985	14	1	7.1
1986–1990	25	3	12.0
1991–1995	154	22	14.3
1996–2000	847	188	22.2
2001–2005	2140	476	22.2
2006–2010	5412	1152	21.3

	ADHD	Autism	Language disorders/impairments	Dyslexia	Learning disabilities	Tourette	Down syndrome	Phenylketonuria	Fragile X
1981–1985	0	0	0	0	0	0	0	0	0
1986–1990	1	0	1	0	1	0	0	1	0
1991–1995	5	10	2	0	3	3	0	5	3
1996–2000	53	44	28	9	14	5	3	10	10
2001–2005	161	80	70	20	26	19	4	14	13
2006–2010	335	178	109	48	34	28	25	23	19

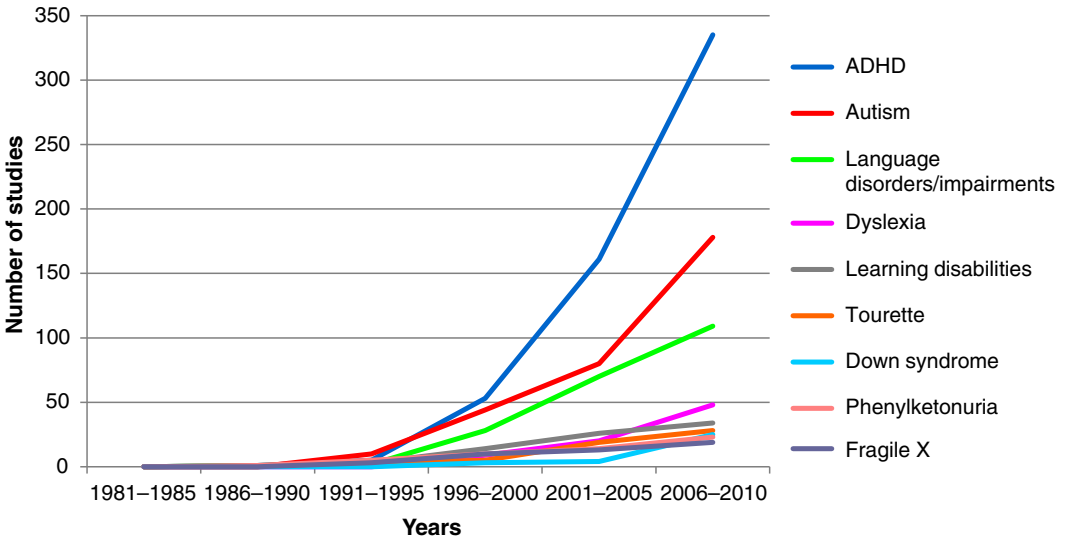

Figure 1 Increase in the number of studies (according to the Scopus database) on executive functions in clinical populations of persons with defined developmental disorders and/or syndromes according to the year of study.

between behavior (behavioral measures of EFs) and the structure and function of the central nervous system (especially in the prefrontal cortex). These were the studies that set the basis for modern developmental research on EFs, which closely relates behavioral and neurophysiological indicators.

Today, EFs start to be tested very early, mostly at the age of 7 months, and EFs' development can be followed until adult age. Simultaneous to the 'discovery' of the developmental aspect of EFs began the research on EFs in developmental disorders (Figure 1), especially in the population of children with attention-deficit/hyperactivity disorder, autism spectrum disorders, and language disorders/impairments. As expected, there were fewer studies on EFs in developmental disabilities that had a more clearly defined etiological mechanism (genetic and metabolic disorders).

Developmental approach to EF research showed that considerable changes in these abilities took place during childhood, which can be greatly attributed to the maturation of the central nervous system. Therefore, the development of EFs has also been extensively studied in children born preterm, showing the influence of immature or vulnerable nervous system on the development of higher developmental functions.

Development of Executive Functions

Different processes encompassed by the name 'executive functions' have different developmental trajectories and begin their intensive development and maturation at different time points. Welsh et al. (1991), by using different measures of EFs, tested children aged 3–12 years and found evidence of a staged development – some components of EFs start to mature earlier and others later. However, most of the studies show that the greatest changes in the performance on tasks that assess EFs are observed in early childhood. Significant changes in the abilities of EFs that take place in preschool age are associated with increasingly better behavioral self-control in children, and lead to the shift from external to internal mechanisms (resources) of self-regulatory behavior.

However, since the task complexity and the mode of EFs testing is different in the infant, preschool, and school period, a careful interpretation is required when making comparative analyses and studying the developmental trajectories of some EFs. In other words, it is impossible to test a certain function throughout the entire developmental period with the same test/task, and even if this is done, the mechanisms of task solving and the child's relying on different cognitive processes that are emerging or have already been acquired hinder interpretation and decrease the data reliability.

Development of EF in Infants and Toddlers (0–2 Years)

According to the research available so far, behaviorally measurable EF abilities begin to appear in the second half of the first year of life. Studies of EFs in infants and toddlers are mostly oriented to studying inhibitory control and working memory.

For investigation of cognitive development in children of that age, most authors use Piaget's research techniques, i.e.,

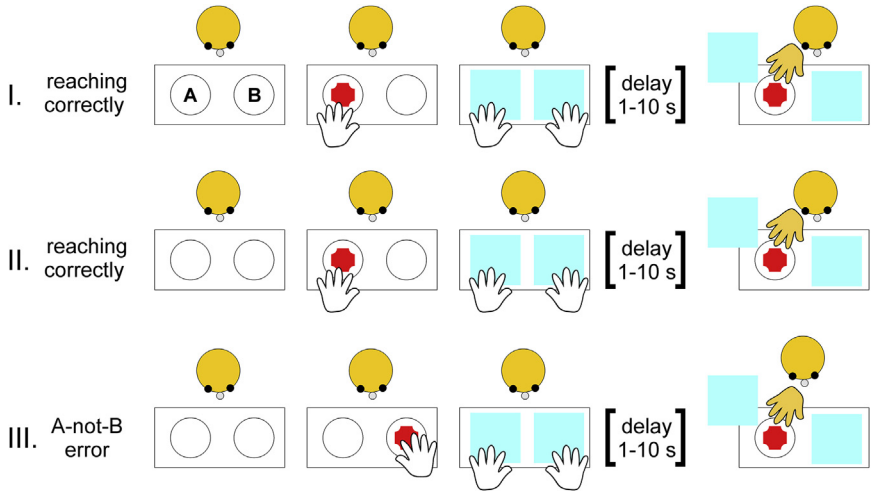

Figure 2 A-not-B task.

object permanence paradigm and object retrieval task. The most frequently used task for assessing early EFs is the A-not-B task (a variant of delayed response task), in which the child searches for an object that the experimenter has hidden in front of him or her in one of the two holes (Figure 2). After the child repeatedly finds the object in the hole A, the experimenter hides it in hole B. If the child continues to search for the object in the previous place (hole A), he or she makes the A-not-B error. The experimenter slowly increases the delay between hiding and searching. Piaget (1954) observed, which was later confirmed by most of the other researchers, that children aged 7–12 months make the A-not-B error, but in this period great differences are observed in the delay length required for the A-not-B error to occur (Diamond, 2002).

Piaget believed that the A-not-B error was not caused by the child's failure to remember, but by his or her inability to distinguish between the object and the location where it had first been found. Over the years, many different explanations have been given as to what mechanisms lead to making of the A-not-B error. There are controversial opinions regarding the validity of the task but it has to be pointed out that similar controversies have also been encountered, to a lesser or greater extent, in other tasks for EF assessment in all age groups (see Chapter 4 on methodological problems in EFs assessment).

In an attempt to determine the factors that influence the performance on the A-not-B task, researchers designed a number of task variants and varied the position and posture of the child, number of holes, the degree of the object's visibility, the mode of response making (looking vs. reaching), the cover color, etc. A contribution to answering the question which resources are needed for a successful task performance was also made by neuroscientific research. Diamond and Goldman-Rakic (1989) showed in a study on rhesus monkeys that the A-not-B error was made by adult monkeys with frontal lobe lesion, but not by those with either parietal lobe lesions or hippocampal lesions (Diamond et al., 1989). The fact that neither parietal nor hippocampal lesions led to making of the A-not-B error speaks in favor of the fact that a successful performance on this task is a result of a combination of abilities, i.e., it depends on the ability of maintenance of an object representation in memory (during the delay) and the inhibition of incorrect response(s), which are the functions of the frontal cortex.

As expected, further research showed a substantial and clear association between the performance on delayed response tasks and functioning of the dorsolateral prefrontal cortex. Therefore, Diamond (2002) says that "This is one of the strongest brain–behavior relations in all of cognitive neuroscience" (p. 468). Due to such a strong relation between the successful performance on tasks that assess EFs and functioning of prefrontal cerebral cortex (PFC) and PFC networks (*see* Cognitive Control (Executive Function): Role of Prefrontal Cortex), it is almost impossible to talk about EF development without taking into consideration the maturation changes in the central nervous system. As already mentioned, the frontal cortex was earlier considered to be functionally silent in childhood, which is why EFs were not studied in this period. However, more recent research in the field of neuroscience has shown that the first two years of life are a period when substantial and important processes of brain maturation and reorganization take place. At the end of the first year of life, when the first important changes in the performance on the A-not-B task can be behaviorally observed, other great structural and functional changes take place in the prefrontal cortex. It is the time when intensive processes of postnatal reorganization of cerebral cortex occur, which include the growth and maturation of pyramidal ('cognitive') neurons, formation of the basic cytoarchitectural characteristics of the prefrontal cortex, morphological differentiation of neurons (growth and formation of dendritic and axon branches), synaptogenesis (formation of neuronal connections), axon myelination, etc. Functionally, in the period of 6–12 months postnatally, an increase in glucose metabolism in the prefrontal cortex takes place, as well as an increase in the density of dopamine receptors, which is considered to be particularly important for the development of higher cognitive functions.

The A-not-B task has in many behavioral and neuroscientific studies served as a basis for EF research, both in children and in other primates. Some studies used it in a slightly changed form to assess other aspects of EFs, i.e., cognitive flexibility. The

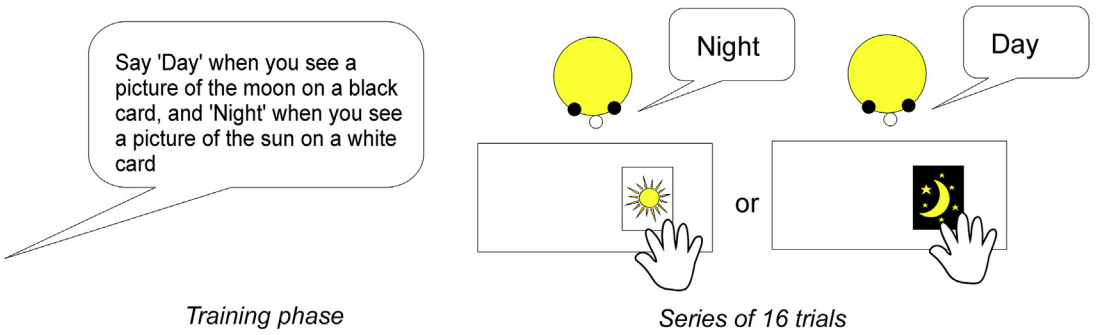

Figure 3 Day-night task.

starting point of such studies was the studying of perseveration, which according to some authors was a reflection of immature or partially developing skills.

Besides delayed response tasks, EF studies, especially those on planning in infants, use the object retrieval task (Diamond, 1981), in which the child can reach a goal only by creating an effective plan. In this task, the target object is placed inside a glass box and can be reached only if the child finds out which side of the box is open. If we take into consideration that the target object is visible the whole time, this task does not require short-term memory but good planning skills and self-control. It has been shown that in 6.5- to 7-month-old children the performance is fully guided by visual input, while in 11- to 12-month-old children it is planned. This implies that planned, purposeful, goal-oriented behaviors are already present in 12-month-old children, and that these abilities get substantially improved at the end of the first year.

EF development during the second year of life has been considerably less studied, especially because there are no widely accepted and used tasks for assessing EFs in that age group. McGuigan and Núñez (2006) experimentally examined executive functioning in 18- to 24-month-old children by using novel detour reaching tasks that included novel-causal and novel-arbitrary sequences. In this way, they controlled the saturation of the tasks by the demands of working memory and inhibitory control. The success of the task performance and the mechanism analysis of EF development were based on changing the task's complexity by varying the number of steps the child has to make to solve the task, the number of attention shifts, etc. Authors believe that early changes in top–down systems at the end of the second year of life may bring about key changes in complex executive functioning that will crystallize in the preschool year.

The progress in the working memory abilities, inhibitory control, and planning is also reflected in the area of self-regulation. The second year brings the development of the basic ability of self-control that includes postponement of the child's own demands as well as the beginning of behavioral regulation. Active regulation of behavior and emotions at the end of the second year is a good predictor of some aspects of social development.

Development of EF in Early Childhood (2–6 Years)

Preschool age is a period of intensive cognitive development. In children aged 3–6 years, progress is clearly visible in many aspects of EFs: inhibitory control, working memory, cognitive flexibility, and simple planning. Understanding of EFs in early childhood is also important because of their relation to the language development and theory of mind, which together play an important role in the social development of an individual.

In this age group, inhibition is most frequently assessed with Stroop-like tasks, such as day–night task (Gerstadt et al., 1994) and grass–snow task (Carlson and Moses, 2001), which test the verbal response ability. In the day–night task, the children are required to say 'day' when they see a picture of stars and moon, and to say 'night' when they see a picture of the sun, over 16 trials (Figure 3). In the grass–snow task the children are required to show a white color patch when the experimenter says 'grass' and a green color patch in response to the word 'snow' over 16 trials. Like the adult Stroop test, children have to (1) maintain the task instructions over a series of trials, (2) suppress a dominant response associated with a perceptual stimulus, while (3) selecting and executing a competing, conflicting subdominant response. Children aged 3–4 years often find it difficult to solve these tasks, while children older than 5 years solve these tasks easily (have more correct answers with less latency).

At the age of 5 years, the ability to retain and manipulate complex information, i.e., working memory, starts to develop increasingly and reaches a plateau in early adulthood. It is believed that in this age, working memory starts to divide into verbal and visuospatial working memory (Alloway et al., 2004), which clearly corresponds to the terms 'phonological loop' and 'visuospatial sketchpad' in Baddeley's model of working memory. The most frequently used task for working memory assessment in this age group is digit span task (Davis and Pratt, 1995). In this task, the experimenter shows the child a doll that repeats the experimenter's words backwards. If the experimenter says 'one, two' the doll repeats 'two, one,' and the experimenter asks the child to repeat the numbers backwards, like the doll. The experimenter begins with two numbers, adding more numbers until the child makes a mistake three times in a row. Children aged 3 and 4 years mostly have difficulties in completing this task, while children aged 5 years repeat two or three numbers backwards.

In order to test the combined action of working memory and inhibition, Luria (1966) developed the tapping task (Luria's tapping task) in which the child is required to tap the surface twice when the experimenter taps once, i.e., taps once when the experimenter taps twice. In this task, the child has

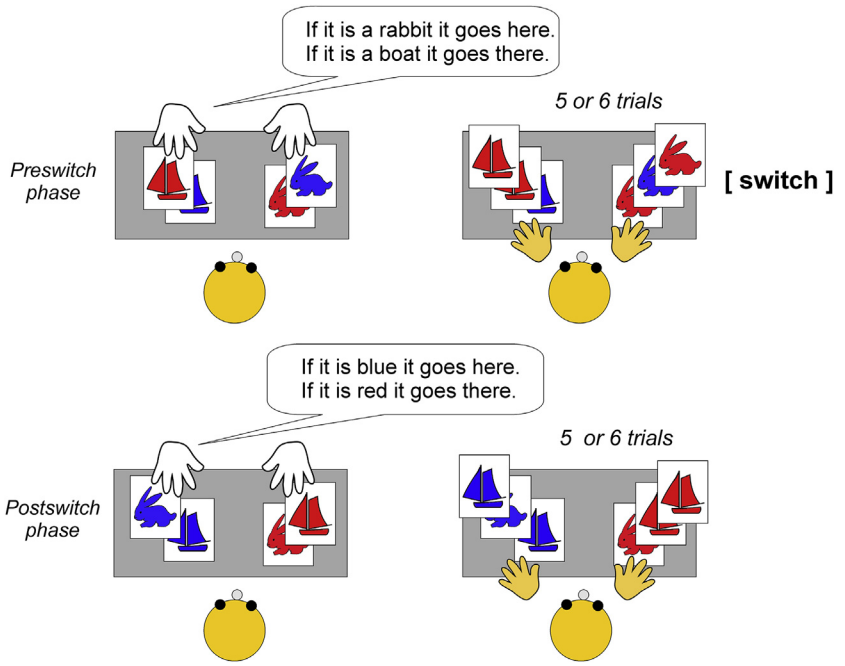

Figure 4 The dimensional change card sort.

to keep in mind two rules and inhibit the natural tendency to imitate the experimenter. The greatest progress in this task has been made in the age group 3–4 years, and the greatest progress in the response speed between 4.5 and 5 years.

Besides the development of the abilities of inhibition and working memory, the preschool age is the age of intensive development of cognitive flexibility. The most frequently used task for assessing cognitive flexibility in preschool age is the Dimensional Change Card Sort (DCCS) (Figure 4) (Frye et al., 1995). The child is shown a card with pictures of different shapes in different colors and is asked to sort the cards according to one dimension (e.g., shape), and after that according to another (e.g., color). This task tests the child's ability to switch the cognitive set. According to the authors of this task, most 3-year-olds can sort the cards according to one dimension, but experience difficulties when the instructions change. Most 4- to 5-year-old children do not experience difficulties with the cognitive set switch, i.e., they complete this task successfully.

At the end of early childhood, the planning ability improves and reaches a plateau in early adulthood. Many researchers believe that improvement in the planning ability is a consequence of the improvement in the ability of response inhibition and working memory. According to Senn et al. (2004), tasks for assessing inhibition and working memory explain 29% variance of planning tasks in children aged 2–6 years. The tasks frequently used for assessing the planning ability are Tower of London (TOL) (Shallice, 1982) and Tower of Hanoi (TOH) (Simon, 1975), in which the child is asked, on the basis of a model, to put three balls of different colors in a set configuration on three pegs of different heights. The two tasks are almost the same. They both require transferring of objects (balls or disk) on pegs from an initial start state to a goal end state in a minimum number of moves. Furthermore, both tasks impose some common rules restricting the manner in which the objects can be moved from peg to peg. Only one ball/disk can be moved at a time, and any ball/disk not being moved must remain on the peg. The difference between these two tasks is that in the TOH task, the minimum number of moves required for each trial is not specified, whereas in TOL, participants are instructed to achieve the configuration in a specified minimum number of moves for each trial, as quickly as possible. Also, the goal model in the TOH task is presented as a physical model, while in the TOL task it is presented pictorially. The TOL and TOH are useful as a measure of EF in children because the procedure offers a variety of difficulty levels. While solving this task, the participant goes through different processes: the initial state, or the state in which the problem solver sorts out the given objects; the goal state, or the solution state that the problem solver tries to achieve; and the steps that the problem solver takes to transform the initial state into the goal state, which initially may not be obvious. Research showed age-related improvements in TOL performance, as older children are more accurate and make fewer extra moves when solving TOL problems.

In conclusion, many behavioral studies showed considerable developmental changes in the performance on many tasks that assess the processes of EFs in early childhood. However, very little is known about the structural and chemical changes in the central nervous system that underlie these behavioral changes. Nevertheless, research points out that in this period the number of dendritic spines (the parts of pyramidal neurons' dendrites that contain the most of synapses) is still much higher than that in the adult brain. The fact is that a large number of synapses is created postnatally (much more than in the adult brain) and is later decreased over time (synaptic pruning) so that only those 'survive' that have been proved to be needed by experience or

active use. It seems that such 'excessive' number of synapses (which remains in the prefrontal cortex for an especially long period) creates an anatomical substrate for neural plasticity and some types of learning, and enables considerable changes in the way the brain processes information and in the development of higher cognitive functions.

Development of EF in Middle Childhood (6–12 Years)

Although in the last few years EF research has been oriented to early childhood, it is exceptionally important to examine EFs during middle childhood. Examining EFs through the entire childhood, even the entire lifetime, can help us more easily answer the question whether EF is a unitary construct, how the relationship between certain aspects of EFs changes developmentally, and what influence EFs have on school performance, social functioning, and emotional control.

Although early childhood brings intensive fundamental changes in the ability of prepotent response inhibition, middle childhood brings greater changes in the response speed and accuracy. In this period, inhibition is most frequently tested by computerized tasks, because they can precisely measure behavior variables (e.g., reaction time). In these tasks, the child is required to respond to certain stimuli presented on the computer screen while inhibiting the response to distracter stimuli. The results mostly show that 8-year-old children make fewer errors caused by inattention, impulsivity, and distractibility than children aged 10–12 years, i.e., during middle childhood this aspect of EF undergoes considerable maturation.

In children aged 6–12 years, the progress is also clearly visible in the abilities of working memory, cognitive flexibility (particularly flexible shifting from old to a new cognitive set and back), and planning, and especially in the increase in response speed. The abilities of working memory and cognitive flexibility appear to emerge in the preschool years, but linear progress is clearly visible only in middle childhood (e.g., on tasks for assessment of working memory, linear progress was observed between 4 and 15 years of age). Here, we have to have in mind that age-related changes in performance depend on the complexity of a particular task.

Cognitive flexibility in middle childhood is most frequently assessed by the Wisconsin Card Sorting Test (WCST). Compared to the DCCS task, this task requires children to find the correct sorting principle by themselves. In its conventional form (Heaton, 1981), the WCST consists of 4 key cards and 128 response cards with geometric figures that vary according to three perceptual dimensions (color, form, or number). The child is required to find the correct sorting principle by trial and error and with the help of examiner feedback. When the child 'discovers' the correct sorting principle, he or she has to maintain this sorting principle across changing stimulus conditions while ignoring the other, irrelevant, stimulus dimensions. After 10 correct responses in a row, the classification principle changes without warning. Using this task, some researchers showed that children aged 6–10 years demonstrated considerable performance improvements, but only when they reach the age of 13–15 years the performance equals that of adults (Welsh et al., 1991).

The EF that shows the largest gains in middle childhood is planning. Planning refers to the ability to formulate actions in advance and to approach a task in an organized, strategic, and efficient manner, which is precisely why this activity is a critical part of goal-oriented behavior. Tasks that are used for EFs assessing are the TOL (Shallice, 1982) and TOH (Simon, 1975) (for task description, see the article on early childhood). Executive planning and problem-solving abilities have been shown to be more advanced in older than in younger children, both in terms of the number of errors and the task-solving time. The time variables show age-related changes, with younger children (7–9 years) solving the task more slowly and having longer total problem-solving times than older children (10–12 years). However, Luciana and Nelson (2002) used the computer version of the TOL (Cambridge Neuropsychological Testing Automated Battery (CANTAB)) to compare a sample of healthy 4- to 12-year-olds with college students and reported that 12-year-olds did not reach young adult levels in terms of perfect solutions and the number of moves required to complete five-move problems.

In general, the understanding of typical EF development in middle childhood is important because the start of school brings a number of changes into the child's life – social environment is changed, expectations from the child are higher, and the complexity of everyday experiences is greater. The studies that aimed to determine the causality between EFs and school functioning have been inconclusive. However, they have found an association between EFs and social–emotional development (social and emotional self-regulation, effortful control, and the ability to resist temptation) as well as school success in math, reading, and writing.

In summary, we can say that middle childhood is the period of dynamic behavioral, cognitive, and emotional development. Anatomically, elimination of excessive synapses in the prefrontal cortex has still not been terminated. At the age of 7 years, neural density is a bit higher than the average value in adults, and even at the age of 11 years children have more synapses than adults (Huttenlocher, 1979). At the age of 9–10 years, the glucose mechanism level starts to decrease, and the adult level is reached at the age of 16–18 years (Chugani, 1994). Functionally, a large number of abilities that are under the responsibility of the frontal lobe have an extended development period.

In conclusion, in middle childhood and adolescence, progress is visible in the abilities of cognitive flexibility and planning, as well as response speed. Although the specifics of the developmental trajectory depend both on the complexity of the task and the scoring method, it seems that inhibition shows prominent improvement during preschool age and less changes later on, and that planning ability experiences the greatest maturation changes. Although the developmental pace is slower, changes in performance (especially in response speed) can be followed up to early adult age.

Methodological Issues in Studies on EF in Children

EF studies face various conceptual problems and challenges. The most important conceptual problem is the synonymous use of terms such as 'executive,' 'control,' 'frontal functions,' and 'control system,' which have not unambiguously operationalized.

A considerable methodological problem is the lack of solid evidence on the validity and reliability of the tasks used for EF assessment. In the assessment of EFs, researchers encounter two types of 'errors.' One is related to the multifactorial nature of executive tasks so it is not quite clear which aspects of EFs they measure (e.g., most of the tasks that measure inhibitory control also measure working memory), or at least to what extent the success on a certain task is influenced by a certain aspect of EFs. The second type of 'error' is related to the fact that the task performance can depend not only on EFs but also on other abilities and other levels of information processing and response making. Therefore, a successful performance on such tasks might include the processes of visual perception, motor skills, processing speed, etc. For all these reasons, poor task performance can be a consequence of EF impairment, or can have a secondary cause in the impairment of some other function.

A further challenge that researchers face is the issue of ecological validity. In an attempt to capture certain narrow segments of EFs, sometimes very simple single tasks are used that do not pick up EF deficits that a person may have in real life, where much more complex demands are present. For example, persons with frontal lobe damage have shown good performance on EF tests, but they are still impulsive and disorganized in real life. One of the ways to solve this problem, especially in children, is to broaden the nature of EF assessments by introducing parent questionnaires designed to index children's everyday EF skills.

It is important to point out that the assessment of EFs in early age mostly includes specifically designed but nonstandardized tasks. Even today there are only a few standardized instruments for assessing EFs in children, e.g., computerized test CANTAB and Attention and Executive Functions domain in a Neuropsychological Assessment-II, to name only some. The lack of standardized tests did not for a long time allow researchers to identify individual differences and detect deviations in the development of EFs, as well as to follow their developmental trajectories by longitudinal research.

Researchers face different methodological problems depending on whether their field of study is EFs in children or in adults (Hughes and Graham, 2002). Compared to adults, EF assessment in infants and children is to a different extent burdened by certain methodological problems, both conceptual and technical in nature.

One of the problems in the assessment of EFs in adults is the difficulty to distinguish between the use of automatic processes and the use of controlled processes during task solving. This distinction is not absolute, so it is better to describe it in terms of a continuum. The processes required to perform an unfamiliar task gradually change from controlled to automatic and even small changes in the task demands can cause the shift. This all leads to an inability to distinguish between executive and nonexecutive tasks. Also, when an 'executive' task is repeated, one cannot be sure whether its underlying processes stay the same, i.e., there is a weak test–retest reliability in adults. The difficulty in distinguishing between automatic and controlled processes is also present in the assessment of children, but it has a somewhat lesser influence. The fact is that children can perceive a novel and unfamiliar task as novel and unfamiliar for a longer period of time. Theoretically, only novel tasks can pick up deficits of EFs, so the corresponding validity and test–retest reliabilities are usually greater in EF assessment in children.

Although the complexity of standard 'executive' tasks is a problem in EF assessment in adults, in children this problem is less obvious, since the tasks, in order to be developmentally appropriate, are simpler, which reduces the influence of a number of different underlying processes.

However, researchers of EFs in children face other challenges, which are less pronounced in the assessment of adults, such as the problem of motivation and the influence of language abilities on the final performance. For example, inadequate understanding of verbal instructions for task solving, i.e., inadequate understanding of language, can influence the performance on EF tasks. Many tasks for assessing EFs in adults are inappropriate for children (e.g., Stroop word test) precisely because of inadequate language understanding.

Some methodological problems encountered in the assessment of children are technical in nature. For example, computerized tests that are conducted using touch screen are mostly considered very reliable and allow precise following of respondent's responses; however, in children, the screen sometimes is not sensitive enough to register the child's pressure or hand temperature (Luciana and Nelson, 2002). This is why all the results expressed as response latency in young children have to be considered with caution.

Due to all the listed factors that can influence the final score on the EF assessment tasks in children, in the analysis and interpretation of the results it is advisable to consider both quantitative (e.g., percent correct, time taken) and qualitative (e.g., frequency of rule violations, displays of frustration or of distractibility) measures of the performance. The fact is that in most EF assessment tasks, performance is evaluated as the sum of all accurately solved tasks. In this way, the total final quantitative result is affected by a range of both executive and nonexecutive abilities. The control of such factors, i.e., at least those that are reflected in the child's behavior and that are visible to the experimenter, could contribute to a better interpretation of the obtained results.

Finally, it is important to point out that in the last 20 years, possibly the greatest progress in the research of EFs, besides methodological progress in terms of computerized tests application, has been made in the integration of behavioral and neuroscience perspectives, which has led to a better recognition of the interactions between social and biological influences on the development of EFs in children.

Bibliography

Alloway, Tracy Packiam., Gathercole, Susan E., Willis, Catherine, Adams, Anne-Marie, 2004. A structural analysis of working memory and related cognitive skills in young children. Journal of Experimental Child Psychology 87, 85–106.

Anderson, Peter John, 2002. Assessment and development of executive function (EF) during childhood. Child Neuropsychology 8, 71–82.

Carlson, Stephanie M., Moses, Louis J, 2001. Individual differences in inhibitory control and children's theory of mind. Child Development 72, 1032–1053.

Chugani, Harry T., 1994. Development of regional brain glucose metabolism in relation to behavior and plasticity. In: Dawson, Geraldine, Fischer, Kurt W. (Eds.), Human Behavior and the Developing Brain. The Guilford Press, New York, pp. 153–175.

Davis, Helen L., Pratt, Chris, 1995. The development of children's theory of mind: the working memory explanation. Australian Journal of Psychology 47, 25–31.

Diamond, Adele, 1981. Retrieval of an object from an open box: the development of visual-tactile control of reaching in the first year of life. Society of Research in Child Development Abstracts 3, 78.

Diamond, Adele, 2002. Normal development of prefrontal cortex from birth to young adulthood: cognitive functions, anatomy and biochemistry. In: Stuss, Donald T., Knights, Robert T. (Eds.), Principles of Frontal Lobe Function. Oxford Universty Press, New York, pp. 466–503.

Diamond, Adele, Goldman-Rakic, Patricia S., 1989. Comparison of human infants and rhesus monkeys on Piaget's A-not-B task: evidence for dependence on dorsolateral prefrontal cortex. Experimental Brain Research 74, 24–40.

Diamond, Adele., Zola-Morgan, Stuart, Squire, Larry R., 1989. Successful performance by monkeys with lesions of the hippocampal formation on A-not-B and object retrieval, two tasks that mark developmental changes in human infants. Behavioral Neuroscience 103, 526–537.

Frye, Douglas, Zelazo, Philip David, Palfai, Tibor, 1995. Theory of mind and rule-based reasoning. Cognitive Development 10, 483–528.

Gerstadt, Cherie L., Hong, Y.J., Diamond, Adele, 1994. The relationship between cognition and action: performance of children $3^1/_2$ – 7 years old on Stroop-like day-night test. Cognition 53, 129–153.

Heaton, Robert K., 1981. The Wisconsin Card Sorting Test Manual. Psychological Assessment Resources Inc, Odessa.

Hughes, Claire, Graham, Andrew, 2002. Measuring executive functions in childhood: problems and solutions? Child and Adolescent Mental Health 7, 131–142.

Huttenlocher, Peter R, 1979. Synaptic density in the human frontal cortex: developmental changes and effects of aging. Brain Research 163, 195–205.

Luciana, Monica, Nelson, Charles A., 2002. Assessment of neuropsychological function through use of the Cambridge neuropsychological testing automated battery: performance in 4- to 12- year-old children. Developmental Neuropsychology 22, 595–624.

Luria, Aleksandr Romanovich, 1966. Higher Cortical Functions in Man. Basic Books, New York.

McGuigan, Nicola, Núñez, María, 2006. Executive functioning by 18–24-month old children: effects of inhibition, working memory demands and narrative in a novel detour-reaching task. Infant and Child Development 15, 519–542.

Miyake, Akira, Friedman, Naomi P., Emerson, Michael J., Witzki, Alexander H, Howerter, Amy, 2000. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. Cognitive Psychology 41, 49–100.

Piaget, Jean, 1954. The Construction of Reality in the Child. Basic Books, New York.

Senn, Theresa E., Espy, Kimberly Andrews, Kaufmann, Paul M, 2004. Using path analysis to understand executive function organisation in preschool children. Developmental Neuropsychology 26, 445–464.

Shallice, Tim, 1982. Specific impairments of planning. Philosophical Transactions of the Royal Society of London, B 298, 199–209.

Simon, Herbert A., 1975. The functional equivalence of problem solving skills. Cognitive Psychology 7, 268–288.

Welsh, Marilyn C., Pennington, Bruce F., Groisser, Dena B., 1991. A normative-developmental study of executive function: a window on prefrontal function in children. Developmental Neuropsychology 7, 131–149.

Executive Government

Wolfgang C Müller, University of Vienna, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In its modern understanding, executive government exists only after the successful separation of powers and the establishment of parliaments. The various types of executives are defined mainly by their method of creation and relations with the legislature. Constitutional and other rules' silence regarding them leaves much of the executive government's internal life for effective self-organization by incumbents. Institutional rules, functional requirements, and the party-political support base of governments lead to distinct models of executive government. The relative strength of government is largely rooted in institutional and party-political factors. Even weak governments tend to be the most important single actors in democratic states. The study of executive government's actual functioning, deeds, and omissions still lacks what other subfields of political science have achieved, which is partly due to the specifics of the research topic.

Executive Government

'Executive government' refers to the political part of the executive branch (which, in turn, is to be distinguished from the legislative and the judiciary branches). There is, however, no easy answer as to where precisely to draw the boundaries between the political part and the remainder of the executive.

The unique task of executive government is to govern the country. Governing means ruling, providing leadership, and transforming political will into public policy. It is not, as the term 'executive' might suggest, just implementing laws passed by the legislature. Rather, governing means executive governments have a strong imprint on the laws passed during their reign and, more generally, exercise overall control over a country, and determine its direction. Even weak governments are the most important single political actor in most regimes.

Origin

While government has existed ever since mankind has organized states (Finer, 1997), executive government in the modern sense emerged through the piecemeal split-off of state functions from a traditionally undivided central government (mostly a monarch) (King, 1975). In order to limit the government's power, judicial functions were transferred to courts and legislative functions to parliaments. This process began in twelfth- and thirteenth-century England. It had many national variations and in Europe was not completed before the twentieth century. The constitutional doctrine of the separation of powers – as developed first and foremost by Locke, Montesquieu, and Madison – provided a normative justification for the separation of legislative, judicial, and executive institutions, which was seen to guarantee liberty and justice (Vile, 1967).

In practice, the state functions were never as neatly separated as envisaged by normative theorists. The executive has retained important legislative functions, in particular, drafting legislation and issuing government decrees and ordinances (Carey and Shugart, 1998). With the full democratization of polities and the establishment of political parties as the main mechanism to structure elections and coordinate incumbents, executives have gained almost a monopoly in law-making in parliamentary systems. In presidential systems, this is not true to the same extent, but executives also exert large influence on legislation by proposing policies and exercising veto rights.

As nondemocratic regimes typically mimic the institutions of democratic ones, executive government here too prevails. Yet, separation of power exists only on paper, allowing holders of political power to intervene in all types of decisions. The real power holders often assume leading positions in executive government, but it may also be that the real center of power is elsewhere. Then executive government may be subject to the orders of the military leadership, the state party presidium, or a religious leader. Nondemocratic regimes thus show a great variety with regard to executive government and will not be further discussed here.

Types of Executive Governments

The executive's nature and executive–legislative relations have traditionally been crucial for the classification of regime types. In the contemporary world, the most important regime types are (1) presidential systems, (2) parliamentary or cabinet systems, and (3) semi-presidential systems. In presidential systems, the chief executive is directly elected for a fixed term and appoints and dismisses the members of his/her government. These cannot be dismissed by the legislature. In a parliamentary or cabinet system, the executive, whether appointed by a head of state or directly created by parliament, remains fully accountable to parliament. Although the prime minister typically has a gatekeeper function, at least through formal control of access to and dismissal from the cabinet, formal decision-making is collective for important matters. Less important matters fall under the jurisdiction of individual cabinet ministers. A semi-presidential system combines a directly elected and powerful head of state with a prime minister and government who are accountable to parliament.

There is, however, no full agreement on how to define these regime types precisely and whether the semi-presidential type indeed constitutes a distinct regime type (Sartori, 1997; Lijphart, 2012; Elgie, 1999; Strøm et al., 2003). In addition to

these widespread regime types, assembly government, collegial (or 'directorate') government, and government with a directly elected prime minister can be identified. In assembly government, decisions are indeed made in parliament, and the cabinet is not more than the assembly's executive secretariat. With the exception of the French Third Republic, it has existed in its pure form for only short periods, often after a revolutionary regime change. There is only one case of collegial (or 'directorate') government – Switzerland. There, the cabinet is elected by parliament for a fixed term, i.e., it is not accountable to the assembly and it is fully collegial; even the chairmanship rotates among its members. Only one case of government with a directly elected prime minister has existed (Israel 1996–2003); this is generally considered a failed constitutional experiment (Kretzmer, 2006).

While these five first-listed regime types have since long belonged to the arsenal of political science, a new demand for constitutional engineering (which came with the new waves of democratization of the 1970s and particularly the 1980s and 1990s) and the new theoretical research agenda of neo-institutionalism have spawned new interest in them (Weaver and Rockman, 1993; Linz and Valenzuela, 1994; Moe and Caldwell, 1994; Sartori, 1997; Lijphart, 2012; Strøm et al., 2003). The question that is common to constitutional engineers and neo-institutional theorists is: What types of output are these regime types likely to generate in terms of the maintenance of democracy, accountability of public officials, and public policy? Each regime type is likely to maximize some effects, but at the costs of others. Hence, the appropriateness of regime types and of specific institutions within the confines of a regime type depends on the most important challenges a polity faces (Sartori, 1997).

The Study of Executive Government

Research on the political executive is a large and increasingly diverse field, including studies on individual offices (such as presidents, prime ministers, cabinet ministers, and top bureaucrats), and government institutions and processes (including informal ones).

Classic institutionalism had not much to say about the executive per se. The executive was mainly characterized through its relations with other state institutions. These relations, in turn, were important for classifying regime types. The bulk of research on the executive, however, has always been historic-qualitative, building on memoires, media reports, the occasional interview, private papers of former members of the executive, and other archival sources. Notwithstanding some impressive individual scholarship, this research has traditionally suffered from three problems. First, with the exception of the US presidency, research on the executive was rare, particularly given its subject's relevance and compared with the attention other political phenomena have received. Secondly, compared with other research fields, fewer attempts at original and systematic data collection were made in executive studies. This partly reflects the research object's properties, namely small numbers (of chief executives, cabinets, cabinet ministers, and top bureaucrats) and limited access (because they are busy people with greater secrecy requirements). Finally, much research on the executive had also been quite idiosyncratic and more rooted in national peculiarities than, for instance, research on electoral behavior or parliaments.

Although these three points are still valid to some extent, executive research has made important achievements in recent decades. First, more countries and individual governments have been subjected to studies in the historic-qualitative tradition, and consequently, the pool of recorded qualitative information has been greatly extended (from the Pacific to Eastern Europe, see, e.g., Hayao, 1993; Blondel et al., 2007). While chief executives have always received much attention (e.g., Rose and Suleiman, 1980; Helms, 2005; Poguntke and Webb, 2005), more recently, the perspective on executive government has become broader and less leader-centered (e.g., Blondel and Müller-Rommel, 1993; Laver and Shepsle, 1994; Strøm et al., 2008). Secondly, the best country studies have become much more rigorous, building on, for instance, systematic interviews with former cabinet ministers (e.g., Burch and Holliday, 1996), ethnographic observation of political actors at work combined with interviews (Rhodes, 2011), and theory-guided case studies of executive decisions (e.g., Rhodes and Dunleavy, 1995). Third, quantitative data and statistical methods have gained further grounds in the study of executive government (with regard to parliamentary systems, this mostly applies to elite and coalition studies; for research on the US presidency, see Edwards et al., 2009; Edwards and Howell, 2009). Finally, executive studies have become more theory-guided and comparative. At the forefront of theory-inspired research, agency theory has been applied to executive–legislative relations (e.g., Epstein and O'Halloran, 1999; Strøm et al., 2003) and intra-executive relations (e.g., Andeweg, 2000). The fact that country expertise is probably more crucial here than in any other subdiscipline is reflected by much cross-national collaborative work. However, comparative work including executives from both sides of the Atlantic is still rare (Weaver and Rockman, 1993; Weller et al., 1997).

Given the centrality of executive government, it is often hard to disentangle the relevant research from the study of the country's governance *tout court*. Consequently, much research devoted to the study of other political phenomena intersects and is also relevant to understand what executive governments do and the causes and consequences of government action (e.g., research on parliaments, political parties, interest groups, business elites, the mass media, political agendas, and public policy).

The Organization and Functioning of Executive Government

Constitutional documents typically omit rules about the internal working and decision-making of government, and little is fixed by other formal rules. Such omission is not accidental but aims at providing current governmental actors to conceive their internal organization and decision-making modes according to their needs. Similar challenges to incumbents have often established conventions on how to organize and run executive governments typical to particular countries or regimes. But the executive can change such conventions by unilateral action, and, if gradual, such change may not be immediately recognized by the outside observer.

Ambitious politicians aim for executive government office not only because such positions are attractive as such, but also because they provide incumbents with a unique chance to make policy in line with their preferences. Even defending the status quo often requires government actions. While a portion of issues on the government agenda are thus due to the incumbents' desire to make policy, many more result from other actors' successful agenda-setting efforts, routine administrative processes, secular trends, and external events (including natural disaster, terrorist attacks, foreign government action, and market turbulence). Again, being able to exert strong influence on policy responses to such externally imposed issues is an executive government's privilege.

Effective policy-making requires collecting and evaluating information on the issues at stake, conceptualizing and vetting policy options, developing workable policy instruments (such as regulation, incentive systems, the provision of public services, or diplomatic moves), providing financial means for the resulting costs, mobilizing the required political and societal support, and making sure that the implementation does not suffer from bureaucratic drift or ignorance and resistance of target groups. Policy failure can result from any of these. Governing is thus a highly challenging task. To fulfill it, executive government adapts and builds new internal structures and processes. It mobilizes expertise and manpower to support the thin layer of career politicians in executive government by drawing on reorganizing and staffing the permanent bureaucracy (Lewis, 2008), recruiting political appointees (who leave office with their political masters), and employing external advisors on voluntary or professional basis. With the exception of the US presidency (e.g., Burke, 2000; Warshaw, 2004), research on executive government support structures is largely scattered in national language publications.

Depending on the regime type and the current balance of political forces, executive governments face different needs in terms of mobilizing support. One target is clearly the legislature, the support of which is required to pass legislation and, in case of parliamentary and semi-presidential regimes, to ensure the survival of the cabinet; it constitutes a particular challenge in multiparty systems and for minority and coalition governments. The literature on coalition governance has identified and analyzed a number of mechanisms employed by government parties to ensure the continued support of the cabinet parties and the legislature. These include written formal coalition agreements, the use of existing structures such as parliamentary committees and watchdog junior ministers to check the office behavior of the partners, and the establishment of extraconstitutional bodies such as coalition committees to force consensus and to manage and resolve conflict if it arises (Martin and Vanberg, 2011; Müller and Strøm, 2000; Strøm et al., 2008, 2010). While much remains to be done in describing and analyzing individual mechanisms of coalition governance, countries, and periods, the theoretical agenda is moving toward the analysis of coalition governance to issues of their overall design, to focus on the entire architecture of interparty institutions and mechanisms.

Another dimension of support is derived from mass media and the general electorate. Basically, maintaining the latter is crucial in any democracy, and mass media support can be instrumental. Both can also be helpful to overcome resistance within government parties and from the legislature. The study of presidential government has long included this dimension of executive leadership (Edwards, 2009), and it also figures prominently in the analyses of 'presidentialized' prime ministers such as Tony Blair and Gerhard Schröder.

To capture the actual functioning of executive government, political science has established a number of descriptive models. These are partly derived from the constitutional order but try to highlight how the institutional features and political forces interact and how executive government arrives at decisions. The following models thus capture to what extent actors are typically able to determine outcomes of government decision-making processes compared with other participants. They were initially developed with the background of the archetypal cases of presidential and parliamentary government – the US and the UK – and subsequently applied to other cases. As these cases are partly more complex, they have given reason for conceptual refinement.

Elgie (1997), for instance, distinguishes monocratic (presidential, prime ministerial), collective (cabinet), ministerial, and shared government (with an in-group of executive government members in the driving seat). Any of these can also manifest itself as bureaucratic government (where politicians only maintain the façade but *de facto* abdicate and leave real decision-making to bureaucrats). Finally, segmented government means that different modes may coexist, each one prevailing in one or more policy areas.

These government modes describe how executive government decides at given times. They are likely to change along with important structural and contextual parameters. Single-party cabinets, for instance, are likely to adopt the prime ministerial government mode, but coalition cabinets are not. Some changes are of more long-term nature. The authors identifying trends toward prime ministerial government or presidentialization have identified crucial changes in the tasks of the state, the mass media environment, international embedding, cleavage structures, and electoral politics as factors driving this process (Poguntke and Webb, 2005). Yet, such structural factors only provide opportunities and expectations: Prime ministerial government also requires suitable personalities, who are not always available. The ministerial government mode (Laver and Shepsle, 1994, 1996), where real power rests with individual ministers, and cabinet approval remains purely formal, is rooted in the institutional prerogatives of ministers in their policy domain, partisan interest and power sharing in coalition governments, and the impossibility to exercise full control for a single center of authority in single-party government.

Segmented government implies that each cabinet is likely to handle issues differently according to its policy domain (Andeweg, 1997; Elgie, 1997). Factors such as the policy substance (degree of technicality, financial resources involved), nature of the policy community, pressure on government from parties and interest groups, mass media interest, and the electoral benefits to reap or the damage inflicted if things go wrong can differ largely. This suggests different ways to handle these issues. Currently, 'segmented government' is an interesting idea based on some largely anecdotal evidence from a few countries that still might merit theoretical elaboration.

The frequently asserted transformation of cabinet government into prime ministerial government and/or ministerial government suggests some long-term change between models of government. This becomes more complex when further short-term situational factors are considered and routine decision-making is distinguished from decision-making in situations of crisis (e.g., Allison, 1971; Boin et al., 2005).

New research frontiers on executive government are defined by the increasing internationalization of governance. On the one hand, this means the building of regional centers of integration. The most important and deeply structured one is the European Union. Thus, executive studies have begun to explore how EU member states meet the challenge of multilevel governance, that is, the participation of all levels of the national executive in the decision-shaping and decision-making in EU institutions (e.g., Kassim et al., 2000). On the other hand, the increasing globalization of markets and environmental and security concerns have led to the establishment of global governance structures in which the chief executives of the most important 8 (G8) or 20 (G20) states participate (Hajnal, 2007). If anything, this has contributed to a further concentration of power at the top of national governments.

All of the above is about structures and processes. Beginning with the behavioral revolution, research has also focused on people in executive government. Many studies in the elite research tradition compiled and analyzed data on their demographics, social background, and career patterns (e.g., Blondel, 1980; Blondel and Thiébault, 1991; Krook and O'Brian, 2010). While certainly valuable to demonstrate cross-national differences and the impact of social and political change of great magnitude, it has remained doubtful that inferences can be made from these data on the behavior of incumbents. The gender issue might be a possible but still an empirically underresearched exception.

Executive Government Capacity, Output, and Performance

Basic constitutional choices, specific institutional constraints, the partisan makeup of government, and executive control over the legislature determine the policy-making capacity of government to a large degree.

Executive governments in presidential systems have traditionally been considered structurally weak under conditions of 'divided government' (i.e., when the legislature or individual chambers and the presidency are held by representatives of different parties). Such division of power is implicit in the design of presidentialism and intended to cause interinstitutional cooperation and partisan policy-moderation. In cases of ideological or severe policy conflict, however, cooperation is likely to fail and policy stalemate or deadlock to result (Binder, 2003). While presidential powers typically ensure that the country's foreign and military policy is not too greatly affected in such instances, other government business may suffer severely. Repeated shutdowns of most of the US government services in recent years resulting from the failure to pass budgets have angered citizens and sent unwelcome signals to the markets.

Divided government can also occur in semi-presidential systems when the presidential and the parliamentary elections result in different majorities. Typically, such incongruence is less consequential, as the prime minister/cabinet/parliamentary majority bloc clearly outweighs the president/parliamentary minority one (Elgie, 1999; Elgie and Moestrup, 2008). In parliamentary systems, single-party majority cabinets are the structurally strong governments, and coalition and minority cabinets the structurally weak ones. In all regime types, governing becomes more challenging when the government or its legislative support base suffers from great internal policy distance or internally divided parties (factionalized, ideologically incohesive).

Government capacity critically hinges on the quality of the incumbents and the visions and skills of the individuals. While suitable candidates are not always available, selection mechanisms – institutional and intraparty mechanisms – can privilege some individual qualities over others (Heclo, 1973).

It is also important to consider cabinet stability in this context. It is generally accepted that instable government is weak. Depending on the regime form and the research perspective, the theoretical concept of 'government stability' has been defined differently. Most definitions focus on government duration. Here parliamentary systems offer the greatest range of measurement options, as different criteria can define a government: formal acts (official in and out dates), the identity of the prime minister, elections, party composition, and even the cabinet's majority status (majority, 50%, minority). Each change is potentially important. The standard definition of cabinet includes elections, party composition, and identity of the prime minister, but more or less comprehensive definitions may suite a particular research purpose better. In multiparty presidential systems, the same set of indicators can be applied (Shin, 2013). As these indicators are unambiguous, smoothly translate into quantitative data, and are relatively easily collectable over time and space, government stability has become a well-researched topic (Warwick, 1994; Saalfeld, 2013; Schleiter and Morgan-Jones, 2009).

The other perspective on government stability is stability in executive leadership that focuses on either the one specific office (president, prime minister, or whatever title is used) (Bienen and van de Walle, 1991) or the group of top executives (the cabinet members) (Huber and Martinez-Gallardo, 2008; Fischer et al., 2012). This research shows that in parliamentary systems cabinet membership stability is only loosely related to the first type of cabinet stability. More traditional political scientists and game theorists alike have related ministerial instability to gains of the bureaucracy in terms of autonomy (Rose, 1976; Huber and Lupia, 2001). At the same time, each change potentially delays government business.

Many different mechanisms are behind change in cabinet membership and portfolio allocation, including changes in the party composition; shifts of power between the cabinet parties; death or illness of cabinet members; shifts in intraparty power; the firing, hiring, promotion, and demotion of individual ministers according to their performance or public standing; and symbolic action by the prime minister to regain momentum. Such a vast area of conceptually different and often hard-to-measure variables calls for country expertise. So far, this research tradition is often rooted in national case studies,

which do not, of course, preclude the use of quantitative data and statistical methods (Indridason and Kam, 2008; Dowding and Dumont, 2009; Berlinski et al., 2012).

Political observers often employ the term 'stable government' in yet another sense: its firm support, internal cohesion, and capability of making decisions and forcefully interacting with its environment. The presence of variables that explain cabinet instability in terms of turnover indicates that statistically, the government is more likely to fall than those without these characteristics. Crisis periods typically precede the fall of cabinets, internal conflict, and stalemate in decision-making. Given that considerably more such crises are to be expected than cabinet falls (as brinkmanship may succeed), structurally weak governments are likely to be unstable in this second sense of the word.

Yet, not too much should be inferred from the latter for actual performance of these governments. Government builders may have established governance mechanisms to at least partly compensate for structural weaknesses. A desideratum in research is therefore an attempt to conceptually and empirically come to grips with this commonly held idea of stable government. One way to address it would be the systematic analysis of nonterminal cabinet crises.

Turning from executive government's performance in terms of survival and stability to output and outcome performance, we need to keep in mind that executive governments are typically the most powerful political actors within states and exercise much and often critical influence on decisions that are important for the fate of nations. These include going to war or making peace and introducing major domestic reforms via legislative programs and devoting public resources to tasks such as education and welfare. Therefore, the performance of executive government is both politically and substantively important.

Actual government behavior manifests itself in individual policy measures, more complex reform packages, the overall legislative program, and budget allocations. International organizations such as the OECD or the EU provide many relevant indicators, most of which relate to country performance. Direct measures of government outputs are rare and mostly generated by individual researchers. While tailor-made for their specific purpose, they are often quite limited in their scope, geographic breadth, and the time covered and therefore only slightly useful for to studying other research questions and building time-series data sources for long-term accumulation of knowledge.

Given the unique position of executive government to exert influence on public policy and the frequent credit claims by incumbents for actual outcomes, researchers also consider the state of the world and how it has changed under a particular government when measuring its performance. Indicators include the classic economic ones (unemployment, economic growth, and inflation rates) and many others relating to all dimensions of human life (e.g., health, education, environmental quality, crime, access to means of communication, etc.). Citizens make similar comparisons, and survey researchers routinely ask about how their respondents' or the country has fared compared with the immediate past. Yet, outcomes such as country performance are influenced by a multitude of time-varying factors other than government policy and thus need to be considered in international comparison. One particular line of government performance evaluation is to measure it against the yardstick of the pledges that political actors made when campaigning for election to executive office. Most of this research after the pioneering study of Royed (1996) has been limited to single countries and a few cabinets, but attempts are under way to make this research comparative (Thomson et al., 2012). A final perspective on the performance of executive government is evaluative at a more abstract level based on survey data of government popularity, approval of the chief executive, trust in government, and, in a historical perspective, lasting reputation.

According to the outcome of legitimacy perspective, a system's legitimacy rests primarily on its achievements rather than the process it employs in producing those outcomes (Scharpf, 1999; Majone, 2006). In this perspective, executive government action remains crucial for system legitimacy.

The attractiveness of executive office on the one hand and the relevance of performance to the electorate on the other are the key mechanisms to tie public policy to the will of the people and hence to ensure democracy in substantive terms. Assuming myopic voters, governments therefore may try to manipulate the business cycle to make sure that the state of the world is good when voters are called or to time their expenditure and taxation policies according to electoral convenience rather than economic needs (Mueller, 2003). While it seems that globalization has largely removed the first of these strategies from the arsenals of most governments, the second one is a remaining concern.

The peril of entrusting one institution and a set of individuals with a great amount of factual power lies in its abuse. Such concerns have traditionally been at the heart of constitutional design. Yet, constitutional constraints and accountability mechanisms do not always work sufficiently. Sadly, many individuals in executive office fall prey to temptations that the office provides, and almost no government is fully immune against such fallibility. Executive government members thus may exploit their office for private or political group advantage by corruption and political patronage or use executive powers to cover up such misbehavior. They may also pervert political competition by *de facto* or formally changing the rules of the game. Such dismal behavior has been diagnosed as a major cause of economic underperformance and failure in societal development (as measured by the Human Development Index). This has given rise to increased attention to corruption, political patronage, and – contrarily – good governance in recent years (Rothstein, 2011).

In the worst case, countries characterized by executive power abuse find themselves on a slippery road to a nondemocratic regime. Such cases abound in the less wealthy world regions with countries such as Zimbabwe (under Robert Mugabe) or Venezuela (under Hugo Chavez) as relevant (though not equivalent) cases. Similar tendencies have been recorded even for European states such as Italy (under Silvio Berlusconi) and Hungary (under Viktor Orbán). The fear of government that has driven classic constitutional thinking thus has not lost any of its relevance today. How to enable the government to carry out its tasks and at the same time to keep it accountable – Madison's question – remains the key issue it has ever been in the history of democracy.

See also: Institutionalism; Parliamentary Government; Political Representation; Presidential Government.

Bibliography

Allison, G.T., 1971. Essence of Decision. Explaining the Cuban Missile Crisis. Little, Brown and Company, Boston.

Andeweg, R.B., 1997. Collegiality and collectivity: cabinets, cabinet committees and cabinet ministers. In: Weller, P., Bakvis, H., Rhodes, R.A.W. (Eds.), The Hollow Crown. Macmillan, London, pp. 58–85.

Andeweg, R.B., 2000. Ministers as double agents? The delegation process between cabinet and ministers. European Journal of Political Research 37, 377–395.

Berlinski, S., Dewan, T., Dowding, K., 2012. Accounting for Ministers. Scandals and Survival in British Government 1945–2007. Cambridge University Press, Cambridge.

Bienen, H., van de Walle, N., 1991. Of Time and Power. Leadership Duration in the Modern World. Stanford University Press, Stanford.

Binder, S., 2003. Stalemate. Brookings, Washington, DC.

Blondel, J., 1980. World Leaders. Sage, London.

Blondel, J., Thiébault, J.-L. (Eds.), 1991. The Profession of Government Ministers in Western Europe. Macmillan, London.

Blondel, J., Müller-Rommel, F. (Eds.), 1993. Governing Together. Macmillan, Basingstoke.

Blondel, J., Müller-Rommel, F., Malova, D., 2007. Governing New European Democracies. Palgrave Macmillan, Basingstoke.

Boin, A., 't Hart, P., Stern, E., Sundelius, B., 2005. The Politics of Crisis Management. Cambridge University Press, Cambridge.

Burch, M., Holliday, I., 1996. The British Cabinet System. Prentice Hall, London.

Burke, J.P., 2000. The Institutional Presidency: Organizing and Managing the White House from FDR to Clinton. Johns Hopkins University Press, Baltimore.

Carey, J.M., Shugart, M.S. (Eds.), 1998. Executive Decree Authority. Cambridge University Press, Cambridge.

Dowding, K., Dumont, P. (Eds.), 2009. The Selection of Ministers in Europe: Hiring and Firing. Routledge, London.

Edwards, G.C., 2009. Leading the public. In: Edwards, G.C., Howell, W.G. (Eds.), The Oxford Handbook of the American Presidency. Oxford University Press, Oxford, pp. 182–207.

Edwards, G.C., Kessel, J.H., Rockman, B.A. (Eds.), 2009. Researching the Presidency. University of Pittsburgh Press, Pittsburgh.

Edwards, G.C., Howell, W.G. (Eds.), 2009. The Oxford Handbook of the American Presidency. Oxford University Press, Oxford.

Elgie, R., 1997. Models of executive politics: a framework for the study of executive power relations in parliamentary and semi-presidential regimes. Political Studies 45, 217–231.

Elgie, R. (Ed.), 1999. Semi-presidentialism in Europe. Oxford University Press, Oxford.

Elgie, R., Moestrup, S. (Eds.), 2008. Semi-presidentialism in Central and Eastern Europe. Manchester University Press, Manchester.

Epstein, D., O'Halloran, S., 1999. Delegating Powers. Cambridge University Press, Cambridge.

Finer, S.E., 1997. The History of Government from the Earliest Times. Oxford University Press, Oxford.

Fischer, J., Dowding, K., Dumont, P., 2012. The duration and stability of cabinet members. International Political Science Review 35, 505–519.

Hajnal, P.I., 2007. The G8 System and the G20: Evolution, Role and Documentation. Ashgate, Aldershot.

Hayao, K., 1993. The Japanese Prime Minister and Public Policy. University of Pittsburgh Press, Pittsburgh.

Heclo, H., 1973. Presidential and prime ministerial selection. In: Matthews, D.R. (Ed.), Perspectives on Presidential Selection. Brookings, Washington, pp. 19–48.

Helms, L., 2005. Presidents, Prime Ministers and Chancellors. Palgrave Macmillan, Basingstoke.

Huber, J., Lupia, A., 2001. Cabinet instability and delegation in parliamentary democracies. American Journal of Political Science 45, 18–32.

Huber, J.D., Martinez-Gallardo, C., 2008. Replacing cabinet ministers: patterns of ministerial stability in parliamentary democracies. American Political Science Review 102, 169–180.

Indridason, I., Kam, C., 2008. Cabinet reshuffles and ministerial drift. British Journal of Political Science 38, 621–656.

Kassim, H., Peters, B.G., Wright, V. (Eds.), 2000. EU Policy Co-ordination: The National Dimension. Oxford University Press, Oxford.

King, A., 1975. Executives. In: Greenstein, F.I., Polsby, N.W. (Eds.), Handbook of Political Science, vol. 5. Addison-Wesley, New York, pp. 173–256.

Kretzmer, D., 2006. Experimenting with constitutional change: direct election of the prime minister in Israel. European Constitutional Law Review 2, 60–80.

Krook, M.L., O'Brien, D.Z., 2010. All the president's men? The appointment of female cabinet ministers worldwide. Journal of Politics 74, 840–855.

Laver, M., Shepsle, K.A. (Eds.), 1994. Cabinet Ministers and Parliamentary Government. Cambridge University Press, Cambridge.

Laver, M., Shepsle, K.A., 1996. Making and Breaking Governments. Cambridge University Press, Cambridge.

Lewis, D.E., 2008. The Politics of Presidential Appointments. Princeton University Press, Princeton.

Lijphart, A., 2012. Patterns of Democracy. Yale University Press, New Haven.

Linz, J.J., Valenzuela, A. (Eds.), 1994. The Failure of Presidential Democracy. Johns Hopkins University Press, Baltimore, MD.

Majone, G., 2006. The common sense of European integration. Journal of European Public Policy 13, 607–626.

Martin, L.W., Vanberg, G.S., 2011. Parliaments and Coalitions. Oxford University Press, Oxford.

Moe, T.M., Caldwell, M., 1994. The institutional foundations of democratic government: a comparison of presidential and parliamentary systems. Journal of Institutional and Theoretical Economics 150, 171–195.

Mueller, D.C., 2003. Public Choice III. Cambridge University Press, Cambridge.

Müller, W.C., Strøm, K. (Eds.), 2000. Coalition Governments in Western Europe. Oxford University Press, Oxford.

Poguntke, T., Webb, P. (Eds.), 2005. The Presidentialization of Politics. Oxford University Press, Oxford.

Rhodes, R.A.W., 2011. Everyday Life in British Government. Oxford University Press, Oxford.

Rhodes, R.A.W., Dunleavy, P. (Eds.), 1995. Prime Minister, Cabinet and Core Executive. Macmillan, London.

Rose, R., 1976. The Problem of Party Government. Pelican Books, London.

Rose, R., Suleiman, E.N. (Eds.), 1980. Presidents and Prime Ministers. American Enterprise Institute, Washington, DC.

Rothstein, B., 2011. The Quality of Government: Corruption, Social Trust, and Inequality in International Perspective. University of Chicago Press, Chicago.

Royed, T.J., 1996. Testing the mandate model in Britain and the United States: evidence from the Reagan and Thatcher Eras. British Journal of Political Science 26, 45–80.

Saalfeld, T., 2013. Economic performance, political institutions and cabinet durability in 28 European parliamentary democracies, 1945–2011. In: Müller, W.C., Narud, H. (Eds.), Party Governance and Party Democracy. Springer, New York, pp. 51–79.

Sartori, G., 1997. Comparative Constitutional Engineering, second ed. Macmillan, Houndmills.

Scharpf, F.G., 1999. Governing in Europe. Effective and Democratic? Oxford University Press, Oxford.

Schleiter, P., Morgan-Jones, E., 2009. Constitutional power and competing risks: Monarchs, presidents, prime ministers, and the termination of East and West European cabinets. American Political Science Review 103, 496–512.

Shin, J.H., 2013. Cabinet duration in presidential democracies. Political Science Quarterly 128, 317–339.

Strøm, K., 2000. Delegation and accountability in parliamentary democracies. European Journal of Political Research 37, 261–289.

Strøm, K., Müller, W.C., Bergman, T. (Eds.), 2003. Delegation and Accountability in Parliamentary Democracies. Oxford University Press, Oxford.

Strøm, K., Müller, W.C., Bergman, T. (Eds.), 2008. Cabinets and Coalition Bargaining: The Democratic Life Cycle in Western Europe. Oxford University Press, Oxford.

Strøm, K., Müller, W.C., Smith, D.M., 2010. Parliamentary control of coalition governments. Annual Review of Political Science 13, 517–535.

Thomson, R.R., Royed, T., Naurin, E., Artés, J., Ferguson, M., Kostadinova, P., Moury, C., 2012. The program-to-policy linkage: a comparative study of election pledges and government policies in ten countries. Paper Prepared for the 2012 Annual Meeting of the American Political Science Association, New Orleans, August 30–September 2.

Vile, M.J.C., 1967. Constitutionalism and the Separation of Powers. Oxford University Press, Oxford.

Warshaw, S.A., 2004. The Keys to Power. Managing the Presidency. Longman, New York.

Warwick, P.V., 1994. Government Survival in Parliamentary Democracies. Cambridge University Press, Cambridge.

Weaver, R.K., Rockman, B.A., 1993. Do Institutions Matter? Government Capabilities in the United States and Abroad. Brookings, Washington, DC.

Weller, P., Bakvis, H., Rhodes, R.A.W. (Eds.), 1997. The Hollow Crown. Macmillan, Basingstoke.

Expectancy-Value-Cost Model of Motivation

Kenneth E Barron, James Madison University, Harrisonburg, VA, USA
Chris S Hulleman, University of Virginia, Charlottesville, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Expectancy-Value models recognize the important role of two components in promoting overall motivation: having an *expectancy* of being successful in a task and having a *value* for engaging in the task. In this article, we highlight the additional role that the *cost* of engaging in an activity plays in influencing motivation and propose a revised Expectancy-Value-Cost model of motivation. Because cost has been less researched and written about, we pay particular attention to its past, present, and future role in an overall model of motivation.

Although numerous theories of motivation have been proposed over the past few decades, Expectancy-Value models of motivation stand out for their ability to synthesize multiple theoretical perspectives, capture the key components of what motivates an individual, and explain a wide range of achievement-related behaviors. In this article, we review the contemporary perspective of Expectancy-Value models used in education.

As the name suggests, Expectancy-Value models have centered on the importance of two components in promoting overall motivation: having an *expectancy* of being successful in a task and having a *value* for engaging in the task. In addition, as indicated by the title of this article, we propose promoting a third component into the overall name of the model to highlight the additional role that the *cost* of engaging in an activity plays in influencing motivation. Although cost is theorized to be a subcomponent in prior Expectancy-Value models, it has been largely ignored in the past empirical work (Wigfield and Cambria, 2010). Fortunately, cost as a factor has reemerged with recent research demonstrating its importance in capturing motivational dynamics that complement expectancy and value components.

For example, consider the following three students enrolled in a calculus course. Math is a challenging subject for Rory, but by putting in extra effort and adopting the appropriate study strategies he has been able to perform well in prior coursework. However, he finally met his match with calculus. Even with additional effort, he is unable to understand the material. As a result, he lacks confidence that he will do well, and his motivation for calculus has decreased. In contrast, math is an easy subject for Jeff. He has always scored at the top of his class and continues to do well in calculus. But this year, his motivation also has substantially decreased. Jeff struggles to see the utility of learning calculus and how he will use it in the future. Finally, there is Jessica, who excels in math and finds it to be one of her favorite courses. She is also interested in several science careers that use calculus. However, due to an ambitious academic and extracurricular schedule, she is struggling to find enough time to complete her schoolwork. In particular, her grades in math have suffered. She knows she could do it; she just cannot find the time. Now, she too admits her motivation for math is not what it used to be.

Each student faces a unique motivational challenge, reflecting one of the expectancy, value, and cost components. In Rory's case, he lacks the confidence (i.e., expectancy) that he can successfully learn calculus. In Jeff's case, he fails to see a reason or purpose (i.e., value) for learning calculus. However, Jessica's challenge is different. She possesses expectancy and value for math, but there are other barriers in her way (i.e., costs) that are thwarting her from being able to invest the time and energy to be successful. Thus for a comprehensive model of motivation, we propose including all three in the revised Expectancy-Value-Cost model.

Overview of Expectancy-Value Models

For more than 30 years, Eccles and her colleagues (Eccles et al., 1983; Wigfield and Eccles, 2000) have championed a contemporary version of Expectancy-Value models to better understand students' choices, persistence, and performance in education. This work grew out of earlier theoretical models of Expectancy-Value motivation (e.g., Atkinson, 1957; Lewin et al., 1944). Expectancy-Value frameworks also have been developed in other fields, such as Vroom's (1964) Valence-Instrumentality-Expectancy model for work settings, but are beyond the scope of this article focused on education.

There are three defining features of Eccles et al.'s contemporary version of the Expectancy-Value model: it is psychological, developmental, and integrative. First, in terms of being psychological, Expectancy-Value models are rooted in an individual's subjective beliefs. Eccles et al. (1983) stressed, "the model itself is built on the assumption that is not reality itself (i.e., past success or failures) that most directly determines children's expectancies, values, and behaviors but rather the interpretation of that reality" (p. 79). In particular, to be optimally motivated, Eccles et al. (1998) argued that a student has to answer "yes" to two fundamental questions: "Can I do the task?" and "Do I want to do the task?" The first question reflects a belief in having an expectancy to do a task, and the second reflects a belief in having a value or reason to do the activity. To truly understand a student's academic choices and behaviors, Eccles (2006) stressed the importance of needing to understand what a student is 'psychologically thinking.'

Second, in terms of being developmental, the model asserts that expectancy and value are shaped over time by individual and contextual factors (see Eccles et al., 1983, 1998 for reviews). These include personal and family demographics (e.g., gender, culture, and socioeconomic status), past experiences of success and failure, an individual's goals and self-concept, and the influence of different socializing agents (e.g., parents, teachers, peers, and schools). Expectancy and value beliefs then are hypothesized to influence the specific academic choices, persistence, and performance of a student in school. For example, Eccles originally developed her version of the Expectancy-Value model in the late 1970s and early 1980s to understand why girls were less likely to persist in taking higher-level math courses, even though they performed similarly to boys in early math courses. The bulk of research supporting this model has focused on large-scale, correlational studies where students self-report their expectancies and values and then are tracked over time to investigate achievement-related outcomes. Although expectancies and values are positively correlated with a wide range of adaptive achievement outcomes, a unique pattern emerges when both are tested simultaneously. Expectancy is more predictive of performance outcomes, and value is more predictive of continued interest and future course taking outcomes (Eccles et al., 1983; Wigfield and Eccles, 2000). Despite the clear associations between expectancy and value and important educational outcomes, the growing body of developmental work also reveals an overall negative picture of student motivation. Students' expectancies and values typically decline as they progress through school (Jacobs et al., 2002).

Third, in terms of being integrative, the Eccles et al.'s model synthesizes multiple perspectives. In her earliest writing (see Parsons, 1980), Eccles observed how the adoption of different theoretical perspectives led researchers to focus on distinct subsets of the possible factors affecting achievement motivation with no clear way to link or integrate discrepant findings. Thus, she set out to develop a comprehensive motivational framework that integrated different perspectives to better explain students' academic choices, persistence, and performance. She also emphasized the importance of identifying motivational factors that were modifiable, which in turn could be used to design interventions to enhance student motivation. This led her to an Expectancy-Value framework and the components of expectancy, value, and cost, which we elaborate on in separate sections below.

The Expectancy Component

As noted above, the question "Can I do the task?" parsimoniously captures the essence of the expectancy component (Eccles et al., 1998). When students believe that they can do something, they are more likely to engage in that behavior. Initially, Eccles and her colleagues argued that expectancy beliefs were multifaceted and that there was merit in distinguishing between two types of expectancies: *ability beliefs* that comprised someone's current/immediate beliefs about being able to complete a task and *expectancy beliefs* that reflect being able to do the task in the future. However, even though this clear, theoretical distinction is possible, empirical studies of these two subcomponents reveal that they are highly correlated and 'empirically indistinguishable' (Eccles and Wigfield, 1995). As a result, most investigations combine measures of ability and expectancy beliefs into a general expectancy scale. Although ability/expectancy beliefs fail to separate *within* a specific academic or nonacademic domain, students do hold different expectancy beliefs *across* different subjects in school (e.g., that they are better in math than English).

From the outset, Eccles et al. (1983) highlighted the close connection between expectancy and other constructs linked to students' beliefs about being able to complete a task, such as self-concept of ability, task difficulty, locus of control, and attributions. In subsequent writing, Eccles and Wigfield (1995, 2002) highlighted how expectancy was related to other theoretical perspectives such as Self-Efficacy Theory (Bandura, 1986), Self-Worth Theory (Covington, 1992), and Attribution Theory (Weiner, 1979). Similarly, we find it easy to link expectancy to more recently proposed motivational constructs, such as academic mindsets (Dweck, 2006). Thus, the expectancy component of Expectancy-Value models offers an overarching, umbrella construct that can capture and integrate a wide range of theoretical perspectives focused on the importance of believing that one can accomplish a task (see Eccles and Wigfield, 2002; Wigfield and Eccles, 2000 for reviews).

The Value Component

As noted above, the question "Do I want to do the task?" was offered as a parsimonious way to capture the essence of the value component (Eccles et al., 1998). When students hold the belief that they value something, they are more likely to engage in that behavior. Building on the work of earlier value researchers (e.g., Battle, 1966; Crandall et al., 1962; Feather, 1982; Rokeach, 1980), Eccles and her colleagues proposed many reasons why a task would hold value for an individual.

The first is *intrinsic* or *interest* value that reflects the inherent enjoyment an individual experiences from engaging in the task for its own sake. The second is *utility* value that reflects the usefulness of a task in helping achieve other short-term or long-term goals. In some of their work, Eccles and her colleagues labeled this extrinsic utility to emphasize engaging in a task as a means for achieving another end. The third is *attainment* value that reflects that the task affirms a valued aspect of an individual's identity and meets a need that is important to an individual. In particular, attainment value represented a way for Eccles and her colleagues to capture a wide-array of additional values suggested by other researchers (e.g., Rokeach, 1980), without assessing each of these additional values separately (e.g., relatedness value, competence value, and esteem value).

In contrast to the first three subcomponents of value that reflect positive reasons to want to engage in an activity, Eccles et al. (1983) proposed a fourth value labeled *cost*. Eccles et al. suggested that the overall value of a task can be negatively impacted by the perceived costs associated with performing the task. Initially, three types of cost were hypothesized: the amount of effort needed to be successful in the task, the time lost to engage in other valued activities, and negative

psychological states resulting from struggle or failure in the task. The first two types of cost were hypothesized as costs of success (e.g., having to give up time and energy for a task or having to give up doing other valued activities), whereas the third was linked to costs of failure (e.g., embarrassment or anxiety). It was predicted that the choice to want to do an activity would entail a cost/benefit analysis. As the level of cost increases, the overall value of the activity should decrease.

Thus, four different subcomponents of value were proposed to influence an individual's overall value of an activity: intrinsic value, utility value, attainment value, and cost. Eccles and Wigfield (1995) wrote "the first three [sub] components are best thought of as attracting characteristics that affect the positive valence of the task … cost, in contrast, is best thought of as those factors … that affect the negative valence of the activity" (p. 216). However, the work of Eccles and her colleagues remains largely silent on how to effectively measure cost or how researchers should weight the positive and negative subcomponents of value into an overall measure. Instead, their work concentrates on evaluating the positive subcomponents. Empirical work testing the factor structure supports the separation of the three positive types of value for students in fifth through twelfth grade (Eccles and Wigfield, 1995). However, in younger students, positive value beliefs only separate reliably into one or two factors (Eccles et al., 1993; Wigfield and Eccles, 1992).

Finally, like expectancy, the value component in the Eccles et al.'s model also offers an overarching, umbrella construct that captures and integrates a wide range of theoretical perspectives focused on the importance of wanting to engage in the task (see Eccles and Wigfield, 2002 for a review). These include Self-Determination Theory (Deci and Ryan, 1985), Interest Theories (Hidi and Renninger, 2006), Intrinsic–Extrinsic Motivation Theories (Sansone and Haraciewicz, 2000), Self-Worth Theory (Covington, 1992), and Achievement Goal Theories (Ames, 1992).

The Cost Component

In contrast to the extensive body of work on expectancy and the positively valenced value components, there has been far less empirical work on cost. To honor its theoretical placement in Eccles et al.'s model, an initial discussion of cost was included above in the section of the value component. However, as promised, we review here the past, present, and future role of cost in Expectancy-Value models. Importantly, we consider whether cost is worthy of being promoted from a subcomponent of value as in past theorizing of Expectancy-Value models to a major component that complements expectancy and value components. We will also consider whether an additional overarching question (such as "Can I do the task?" and "Do I want to do the task?") would be beneficial to propose to parsimoniously capture the role of cost.

Past Work

In terms of its past, we first reconsider the theoretical placement of cost in Eccles et al.'s Expectancy-Value model. In their earliest writing, cost was described as an important mediator of value (Eccles et al., 1983). However, in subsequent writing, cost was promoted to one of four subcomponents of value (along with intrinsic, utility, and attainment) that would be weighted in a cost/benefit analysis to determine one's overall value. This type of relationship implies that cost is better conceived of as a moderator variable of value instead. In a moderator relationship, the effect of one variable on an outcome depends on knowing the level of another variable. In other words, the overall effect of value on promoting motivation depends on knowing whether or not someone experiences high or low cost. Thus, incorporating cost into our Expectancy-Value models is critical to fully capturing the motivational dynamics of what attracts or detracts us from engaging in an activity.

Beyond influencing value, it is possible to deduce additional connections on how cost negatively affects expectancy from Eccles and her colleagues' writing. This suggests that cost has a more prominent role than previously indicated. For example, Eccles et al. (1983) linked expectancy beliefs to two variables that we believe serve as proxies for two of the three hypothesized dimensions of cost. First, with regard to cost resulting from the amount of effort required for an activity, Eccles et al. discussed task difficulty perceptions as one of the two main contributors to expectancy beliefs, with the other being an individual's ability beliefs in that domain. When developing initial measures, Eccles and colleagues measured different dimensions of task difficulty and ability beliefs suggesting that both would interact to predict an individual's overall expectancy. Ability beliefs were predicted to positively impact expectancy, while task difficulty was predicted to negatively impact expectancy. In particular, one of the task difficulty dimensions assessed the effort required by the activity, which seems clearly related to costs due to the amount of effort required by the activity.

If task difficulty is reevaluated in this light, past research by Eccles and her colleagues offers evidence how cost may fit into a modified Expectancy-Value-Cost framework. In their most rigorous test of the factor structure of the core components of their model, Eccles and Wigfield (1995) found that their analyses supported a structure of three task values (interest, importance, and utility), one combined ability/expectancy factor, and two task difficulty factors (perceptions of difficulty and perceptions of effort required to do well). Although no explicit measure of cost was included, if we use task difficulty as a proxy for cost (especially items assessing the perceptions of effort required to do well), we have support that cost separates into its own factor rather than loading negatively on either of the other value or expectancy factors. Furthermore, while value and expectancy scales were positively correlated with each other, they were all negatively related to the task difficulty scales. These findings are in line with predictions of how cost should be negatively related to both value and expectancy. Consequently, the work of Eccles and colleagues may not be so silent on offering empirical evidence for the distinction of a unique cost component that is related to both expectancy and value components.

Second, we also found Eccles and her colleagues linking expectancy to costs occurring from negative psychological

consequences. While reviewing expectancy beliefs, Eccles et al. (1983) highlighted a series of studies debating the relationship between expectancy and the negative psychological experience of math anxiety. Fennema and Sherman (1977) argued that math anxiety and expectancy were psychologically equivalent and simply the inverse of each other. In contrast, Meece (1980) argued instead that math anxiety resulted from a combination of low expectancy for success coupled with high psychological cost from a fear of failing. To test these competing ideas, Wigfield and Meece (1988) developed a measure designed to assess affective and cognitive dimensions of math anxiety while being careful to avoid item overlap that would artificially inflate the relationship between anxiety and expectancy. They found that the affective dimension of math anxiety shared moderate to strong negative correlations to expectancy and value, while the cognitive dimension had weak relationships. Moreover, while the affective dimension continued to be more strongly correlated to expectancy than value, it was not as strongly correlated as Fennema and Sherman (1977) suggested. These findings reveal that if math anxiety measures offer a proxy for assessing the negative psychological dimension of cost, we can infer that cost is clearly linked to expectancy.

Furthermore, Wigfield and Meece (1988) included measures of task difficulty, offering an additional test to evaluate how different hypothesized types of costs could be related. The affective anxiety scale was found to share strong, positive correlations with each of the different scales of task difficulty. Interestingly, they suggested that future research should explore the links between math anxiety, performance, and other components of the Expectancy-Value model more fully. They also argued that math anxiety should be considered conceptually distinct from expectancies. In particular, they highlighted the implications that this has for students:

> Our results show that the anxiety that students report represents more than a lack of confidence in math; rather it also centers on negative affective reactions to math. In regard to intervention efforts to alleviate math anxiety, we would suggest that techniques to build anxious students' confidence in their math ability may not be enough to alleviate the strong negative affective reactions to math that they experience. Math anxious students also need training to reduce their fear and dread of math. As has been found in the test anxiety area, intervention efforts focused on both the cognitive and affective components of math anxiety may prove to be the most effective way to reduce its debilitating effects (p. 214).

Thus, identifying cost as a subcomponent of value appears to lack both conceptual and empirical support. Instead, we propose it is time to promote cost to a major component that could be combined and interacted with both expectancy and value components to determine when someone is optimally motivated. Of course, this quickly becomes an empirical question. Although initial attempts were made to include cost from the outset of their work (Parsons, 1980), explicit measures of cost and analyses incorporating cost were not fully developed. Without well-developed and agreed upon measures of cost, we are unable to test models that evaluate the merits of keeping expectancy, value, and cost components separate, and how these three components relate to each other and to student outcomes.

Present Work

Fortunately, interest in studying cost is changing in the present. First, researchers are adopting qualitative approaches that support the inclusion of cost constructs in addition to expectancy and value constructs to more fully capture motivational dynamics. For example, Watkinson et al. (2005) interviewed elementary-aged students to determine the reasons why children chose to participate in or avoid different activities during recess. When discussing reasons that made them avoid participating, children described psychological costs (e.g., being teased) as well as physical costs (e.g., being uncomfortable) in addition to identifying reasons linked to expectancy and value components. Similarly, Xiang et al. (2006) asked children why they liked or disliked participating in particular programs in gym class (e.g., a running unit). Again, a major portion of comments for why students disliked particular programs focused on perceived costs. Finally, Chen and Liu (2009) used qualitative approaches as part of a mixed method study of student motivation to participate in university physical education classes. They measured expectancy and values quantitatively using modified items from Eccles and Wigfield (1995), but due to the lack of a validated scale for cost, they interviewed students to obtain qualitative data on perceived cost. Students offered numerous cost explanations for choosing to stop taking physical education, such as having a heavy workload and other pressing demands on their time.

In addition to qualitative investigations highlighting the importance of cost components, a number of researchers are beginning to measure cost quantitatively. Chiang et al. (2011) surveyed elementary students' about their expectancy, value, and cost beliefs and their willingness to participate in physical activity. They used existing items by Eccles and her colleagues to measure expectancy and value, but added three new items to assess cost (with one item for each of the three subtypes of cost proposed by Eccles et al., 1983). Factor analyses revealed cost loaded on a separate factor from a combined expectancy/value factor. When predicting levels of physical activity, students reporting higher costs were less likely to be active, whereas students reporting higher expectancy/values were more likely to be active.

Returning to a more traditional academic domain, Luttrell et al. (2010) reported the development of a new assessment inventory to measure interest value, utility value, attainment value, and personal cost for math. The cost items focused on two of the three types of cost proposed by Eccles et al. (1983): the amount of effort required by the activity and negative psychological consequences. Interestingly, many of the cost items were similar to the math anxiety items reported in Wigfield and Meece (1988). Factor analysis supported a four-factor solution. Cost was found to be negatively correlated to all three value scales, whereas the three value scales were positively correlated to each other. In addition, students who took three or more college courses rated math less costly than those who took zero, one, or two courses.

Conley (2012) included two new cost items in a large scale study to determine the most common motivation profiles for math students in middle school. Her new items focused on only one type of cost: the loss of engaging in valued alternatives. Again, cost formed a unique factor and was negatively correlated with expectancy, intrinsic value, utility value, and attainment value. Cluster analyses also revealed that cost played a critical role in discriminating more and less adaptive motivational profiles. Then on a measurement note, Conley offered a warning that totaling value and cost scales into an overall value index could mask important differences in motivational profiles of students. For example, the cluster of students who reported lower interest and lower cost would have the same overall value index as another cluster of students who had higher interest and higher cost, even though their profiles were quite different.

Finally, in a large-scale study of both mathematics and English, Trautwein et al. (2012) evaluated items designed to assess cost, expectancy, intrinsic value, utility value, and attainment value. Their cost measure included one item focused on the amount of effort necessary for the class and one item focused on the loss of engaging in valued alternatives. For both math and English, factor analyses revealed that cost and the three value factors supported a four-factor structure. Furthermore, cost was negatively correlated to all three values, and was the most negatively correlated to expectancy. Finally, analyses indicated that students' expectancy, value, and cost beliefs differed by academic domain (e.g., that students perceived different levels of cost for math than English).

In our own research program, we have also been active in reconsidering the role of cost in Expectancy-Value frameworks, and how to approach measuring it. We also adopted qualitative and quantitative approaches to inform our understanding. As part her master's thesis, Flake (2012) conducted a comprehensive literature review of the conceptualization and measurement of cost to determine the core dimensions of cost. Then she followed an in-depth scale development process that included additional qualitative data collection with students. Her goals were to reevaluate the proposed dimensions found in the literature, propose a new pool of items to assess each of the dimensions that emerge, and utilize outside experts to review items for their ability to assess each dimension. A number of major findings emerged that now shape our views about cost.

First, a review of the educational psychology literature consistently revealed three dimensions of cost that Eccles et al. (1983) proposed from the outset of their work: (1) the amount of effort required to be successful in the task, (2) the loss of being able to engage in other valued activities, and (3) negative psychological states resulting from struggle or failure in the task. However, review of the psychological literature outside of education (specifically in behavioral economics) suggested the merits of adding another dimension of cost. Having decreased motivation to engage in a task may result not only from the amount of effort required by the activity itself, but also from the amount of effort that an individual has to exert on other activities that he or she is engaged in. As a result, Flake proposed adding two types of effort costs: the amount of effort required by the task itself (i.e., effort-related cost) and the amount of effort required by other tasks (i.e., effort-unrelated cost). Most university faculty who try to pursue their research know these two types of effort costs all too well when they consider how to find enough time to dedicate to research (effort-related costs) when they have teaching and service responsibilities to juggle as well (effort-unrelated costs). Similarly, in our opening example of the three students taking a calculus class, we initially diagnosed Jessica's motivational problem as cost. But now if we extended the diagnosis to a particular type of cost, the original three dimensions of cost do not directly capture her motivational problem. Instead, it stems from taking on too many high-demanding classes and extracurricular activities that are thwarting her ability to put more time and energy into calculus (i.e., she is experiencing effort-unrelated cost).

In addition to proposing new dimensions of cost, a second major finding from Flake's (2012) work is how items need to be written to best capture cost. A defining feature of Expectancy-Value models rests in an individual's subjective appraisal of the core components. In her qualitative study, Flake asked students to describe features of their most and least motivating classes in college. Interestingly, students reported effort-related costs (e.g., the class was challenging) in both their most and least motivating classes. However, their appraisals were quite different. In their most motivating class, expending additional effort was perceived positively; but in their least motivating class, it was clearly perceived negatively. This informed how to best approach creating a pool of items to measure cost. For example, as we suggested above, task difficulty measures from prior research could serve as a proxy for effort-related cost. One possible way to assess task difficulty could be to measure how challenging a course is (e.g., 'This class is challenging.'). However, challenge can actually increase a student's value because it pushes a student to grow increasing his or her competence. Instead, we recommend writing cost items that capture a negative appraisal from the outset (e.g., 'This class is *too* challenging.'). To agree that a class is *too* challenging suggests it has surpassed a critical threshold and that it is now overwhelming and is perceived to have cost. Measuring cost objectively as the amount of effort or task difficulty alone is not enough. To be perceived as cost, it must be perceived negatively by the respondent.

In terms of quantitative approaches, we are currently involved in a grant project funded by the National Science Foundation to develop a rapid measure of expectancy, value, and cost that can be easily adapted and used across a wide age range and academic domains. We also are testing the measure to see the utility of measuring expectancy, value, and cost at different levels of specificity. For example, we have assessed expectancy, value, and cost at the level of a specific academic task (e.g., for a science unit on carbon), at the level of a specific class (e.g., biology), and at the level of academic semester (e.g., for your coursework this semester). We have three major preliminary findings to report (Flake et al., 2011; Getty et al., 2013; Lazowski et al., 2012). First, factor analyses strongly support the separation of expectancy, value, and cost components into three different scales. Second, expectancy and value are positively related, but in turn are both negatively related to cost. Third, expectancy, value, and cost provide unique

predictive validity on important educational outcomes when tested simultaneously in regression or path models. Expectancy is the strongest, positive predictor of performance outcomes (e.g., test score for a science unit or final grade in an academic class), but is unrelated to interest outcomes (e.g., continued interest in studying that topic or pursuing a career in that area). In contrast, value is the strongest predictor of interest outcomes, but is unrelated to performance outcomes. Cost, however, is a negative predictor of *both* interest and performance outcomes and improves our ability to predict both. We also initiated a longitudinal study to track middle school students' expectancy, value, and cost for their math coursework. Our initial wave of results found small downward trajectories for fifth through seventh grade on expectancy and value, and small upward trajectories for cost. However, large drops in expectancy and value and large increases in cost occur by the time students end eighth grade.

In sum, current qualitative research showcases the value of adding cost into our overall motivation framework, especially to better understand the barriers when individuals are unmotivated. In addition, more researchers are designing and incorporating new measures of cost in quantitative studies. This work is consistently revealing that cost separates into its own factor and is empirically distinguishable from expectancy and value factors.

The Future

With renewed energy and a growing body of researchers now investigating cost, we expect many advances to occur over the next few years. With this work, a number of key questions will need to be addressed as future research unfolds. First, we need to resolve how cost is conceptualized and measured. For example, in our work, we proposed adding at least one additional dimension of cost above and beyond the original theorizing of Eccles et al. (1983), but one key question is, are there other dimensions worthy of consideration? Furthermore, because multiple dimensions may contribute to cost, researchers need to pay particular attention to how a measure is designed and compared with other research. As evidenced by most of the cost work cited above, researchers assessed only one or two of the potential dimensions in a given study. Also, it is not clear how cost dimensions will function alone by themselves or in conjunction with each other. Perhaps the experience of one type of cost could be enough to deter a student from being motivated, thus creating measures of multiple dimensions and totaling them into a total score might be misleading and mask important effects.

Second, we need to consider how cost is tested in conjunction with expectancy and value. We propose that it is time to promote cost to a major component of Expectancy-Value models, so that researchers can simultaneously test the independent and joint effects of expectancy, value, and cost together. In other words, cost needs to enter our models in a way that allows a test of both its independent, additive effect as well as its potential moderator effect that can alter how motivated individuals are depending on their level of expectancy and/or value. This presents a major paradigm shift from early theorizing that explicitly suggested that cost was a mediator or moderator that led to an overall value belief. In addition, although much of the existing work on Expectancy-Value by Eccles and her colleagues had downplayed the interactive or moderating role of expectancy and value, recent methodological developments in modeling interaction terms have emerged to offer more powerful tests, and have found support that the interaction between expectancy and value adds additional variance in explaining outcomes (e.g., see Trautwein et al., 2012). Now it is time to extend these tests to look at interactions involving cost as well.

Final Thoughts

In this article, we have reviewed the current state of Expectancy-Value models in education, and the importance of developing a student's expectancy that he or she can do the task ("Can I do the task?") and a student's value that he or she wants to do the task ("Do I want to do the task?"). However, in addition to having students respond yes to both of these questions to be optimally motivated, a student needs to be free of costs in order to demonstrate motivated behavior.

In our revised Expectancy-Value-Cost model, cost is a distinct component, alongside expectancy and value, which determines motivated behavior. We argued that cost is multidimensional with at least four subcomponents: effort related to the task, effort unrelated to the task, loss of valued alternatives, and negative psychological experiences. In addition, the effort and resources required for an activity are costly only when they are perceived to be too much by a student.

Thus, we would like to close with a new third question that captures the importance of being free of costs that can prevent us from being motivated. Specifically, to be optimally motivated, we need to ensure an individual also says yes to the question: "Am I free of barriers that prevent me from investing my time, energy, and resources into the activity?" As our opening example of students suggests, a student could clearly hold a belief that "Yes, I can do a task" and "Yes, I value the task," but still remains unmotivated because the student's final answer is "No, I have barriers preventing me from engaging in the task." If we only ask two of these three questions, the scope of our models to predict and understand motivation will be limited.

Eccles and Wigfield (1995) highlighted in their article most extensively evaluating the measures of their model that "over time, there has been an evolution in the conceptualization of constructs linked to expectancy for success and task value, as well as a refinement in the components of each construct" (p. 217). We predict continued evolution once again, as renewed energy is focused on cost and on the question of how the components of expectancy, value, and cost work together to impact motivation.

See also: Academic Achievement Motivation, Development of; Academic Motivation and Performance: Task Value Interventions; Academic Performance, Effects of Socio-Economic Status on; Affect-Regulation Motivation; Avoidance and Approach Motivation: A Brief History; Flow in

Motivational Psychology; Gender and Academic Motivation; Grit; Health Self-Regulation, Motivational and Volitional Aspects of; Interest, Psychology of; Intrinsic Motivation: A Cultural Perspective; Leisure Activities Choices among Adolescents; Motivation and Actions, Psychology of; Motivation in Youth Sport and Physical Activity: Developmental Perspectives; Motivation, Familial Influences on; Motivation, Learning, and Instruction; Motivation: History of the Concept; Motivational Development, Systems Theory of; Race and Academic Motivation; School Achievement: Motivational Determinants and Processes; School Burnout and Engagement: Lessons from a Longitudinal Study in Finland; Schooling: Impact on Cognitive and Motivational Development; Self-Determination Theory; Test Anxiety and Academic Achievement; Values, Psychology of; Values, Social Psychology of; Vocational Interests, Values, and Preferences, Psychology of; Work Motivation.

Bibliography

Ames, C., 1992. Classrooms: goals, structures, and student motivation. Journal of Educational Psychology 84, 261–271.

Atkinson, J.W., 1957. Motivational determinants of risk-taking behavior. Psychological Review 64, 359–372.

Bandura, A., 1986. Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice-Hall, Englewood Cliffs, NJ.

Battle, E., 1966. Motivational determinants of academic task persistence. Journal of Personality and Social Psychology 4, 634–642.

Chen, A., Liu, X., 2009. Task values, cost, and choice decisions in college physical education. Journal of Teaching in Physical Education 28, 192–213.

Chiang, E.S., Byrd, S.P., Molin, A.J., 2011. Children's perceived cost for exercise: application of an expectancy-value paradigm. Health Education & Behavior 38, 143–149.

Conley, A.M., 2012. Patterns of motivational beliefs: combining achievement goals and expectancy-value perspectives. Journal of Educational Psychology 1, 32–47.

Covington, M.V., 1992. Making the Grade: A Self-worth Perspective on Motivation and School Reform. Cambridge University Press, New York.

Crandall, V.C., Katkovsky, W., Preston, A., 1962. Children's belief in their own control of reinforcement in intellectual-academic achievement situations. Child Development 36, 91–109.

Deci, E.L., Ryan, R.M., 1985. Intrinsic Motivation and Self-determination in Human Behavior. Plenum, New York, NY.

Dweck, C.S., 2006. Mindset. Random House, New York.

Eccles, J.S., Wigfield, A., Harold, R., Blumenfeld, P.B., 1993. Age and gender differences in children's self- and task perceptions during elementary school. Child Development 64, 830–847.

Eccles, J.S., 2006. A motivational perspective on school achievement: taking responsibility for learning, teaching, and supporting. In: Sternberg, R.J., Subotnik, R.F. (Eds.), Optimizing Student Success in School with the Other Three Rs: Reasoning, Resilience, and Responsibility. Information Age Publishing, Greenwich, CT, pp. 199–224.

Eccles (Parsons), J., Adler, T.F., Futterman, R., Goff, S.B., Kaczala, C.M., Meece, J.L., Midgley, C., 1983. Expectancies, values, and academic behaviors. In: Spence, J.T. (Ed.), Achievement and Achievement Motivation. W. H. Freeman, San Francisco, CA, pp. 75–146.

Eccles, J.S., Wigfield, A., 1995. In the mind of the actor: the structure of adolescents' academic achievement related-beliefs and self-perceptions. Personality and Social Psychology Bulletin 21, 215–225.

Eccles, J.S., Wigfield, A., Schiefele, U., 1998. Motivation. In: Eisenberg, N. (Ed.), Handbook of child psychology, fifth ed., vol. 3. Wiley, New York, pp. 1017–1095.

Eccles, J.S., Wigfield, A., 2002. Motivational beliefs, values, and goals. Annual Review of Psychology 53, 109–132.

Feather, N.T., 1982. Expectancy–value approaches: present status and future directions. In: Feather, N.T. (Ed.), Expectations and Actions: Expectancy–Value Models in Psychology. Erlbaum, Hillsdale, NJ, pp. 395–420.

Fennema, E., Sherman, J., 1977. Sex-related differences in mathematics achievement, spatial visualization, and affective factors. American Educational Research Journal 14, 51–71.

Flake, J.K., 2012. Measuring Cost: The Forgotten Component of Expectancy-Value Theory (Unpublished master's thesis). James Madison Univerisity, Harrisonburg, VA.

Flake, J., Barron, K.E., Hulleman, C.S., Grays, M., Lazoswki, R., Fessler, D., May 2011. Evaluating cost: the forgotten component of expectancy-value theory. In: Poster Presented at the Annual Meeting of Association for Psychological Sciences, Washington, DC.

Getty, S., Hulleman, C.S., Barron, K.E., Stuhlsatz, A.M., Marks, J.C., April 2013. Factors that affect learning in high school science: measuring motivation, achievement, and interest in science. In: Paper Presented at the Annual Meeting of National Association for Research in Science Teaching, San Juan, Puerto Rico.

Hidi, S., Renninger, K.A., 2006. The four-phase model of interest development. Educational Psychologist 41, 111–127.

Jacobs, J.E., Lanza, S., Osgood, D.W., Eccles, J.S., Wigfield, A., 2002. Changes in children's selfcompetence and values: gender and domain differences across grades one through twelve. Child Development 73, 509–527.

Lazowski, R.A., Hulleman, C.S., Barron, K.E., Getty, S., September 2012. Development of an expectancy-value scale for an online science curriculum. In: Paper Presented at the Motivation Retreat, University of Tubingen, Germany.

Lewin, K., Dembo, T., Festinger, L., Sears, P., 1944. Level of aspiration. In: Hunt, J.McV. (Ed.), Personality and the behavioral disorders, vol. 1. Ronald Press, New York, pp. 333–378.

Luttrell, V.R., Callen, B.W., Allen, C.S., Wood, M., Deeds, D.G., Richard, D.C.S., 2010. The mathematics value inventory for general education students: development and initial validation. Educational and Psychological Measurement 70, 142–160.

Meece, J., 1980. A theoretical framework for studying students' course selection in mathematics. In: Paper Presented at the Annual Meeting of American Educational Research Association, Boston, MA.

Parsons, J.E., 1980. Self-perceptions, task perceptions, and academic choice: origins and change. In: Unpublished Final Technical Report to the National Institute of Education, Washington, DC (ERIC Document Reproduction Service No. ED 186577).

Rokeach, M., 1980. Some unresolved issues in theories of beliefs, attitudes, and values. In: Page, M.M. (Ed.), Nebraska Symposium on Motivation. University of Nebraska Press, Lincoln, NE, pp. 261–304.

Sansone, C., Haraciewicz, J.M., 2000. Intrinsic and Extrinsic Motivation: The Search for Optimal Motivation and Performance. Academic Press, San Diego, CA.

Trautwein, U., Marsch, H.W., Nagengast, B., Ludtke, O., Nagy, G., Jonkmann, K., 2012. Probing for the multiplicative term in modern expectancy–value theory: a latent interaction modeling study. Journal of Educational Psychology 104, 763–777.

Vroom, V.H., 1964. Work and Motivation. Wiley, New York.

Watkinson, E.J., Dwyer, S.A., Nielsen, A.B., 2005. Children theorize about reasons for recess engagement: does expectancy-value theory apply? Adapted Physical Activity Quarterly 22, 179–197.

Weiner, B., 1979. A theory of motivation for some classroom experiences. Journal of Educational Psychology 71, 3–25.

Wigfield, A., Cambria, J., 2010. Students' achievement value, goals orientation, and interest development, and relation to achievement outcomes. Developmental Review 30, 1–35.

Wigfield, A., Eccles, J.S., 1992. The development of achievement task values: a theoretical analysis. Developmental Review 12, 265–310.

Wigfield, A., Eccles, J.S., 2000. Expectancy-value theory of achievement motivation. Contemporary Educational Psychology 25, 68–81.

Wigfield, A., Meece, J., 1988. Math anxiety in elementary and secondary school students. Journal of Educational Psychology 80, 210–216.

Xiang, P., McBride, R.E., Bruene, A., 2006. Fourth-grade students' motivational changes in an elementary physical education running program. Research Quarterly for Exercise and Sport 77, 195–207.

Expectations, Economics of

George W Evans, University of Oregon, Eugene, OR, USA; and University of St. Andrews, Fife, UK
Seppo Honkapohja, Bank of Finland, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Expectations are central to the determination of macroeconomic variables. Since the 1970s, 'adaptive expectations' have been replaced by 'rational expectations' (RE), with expectations modeled as optimal forecasts. Some economic models exhibit multiple RE, giving an independent role for expectations. The learning approach goes beyond RE to ask whether and how agents might coordinate on RE. Adaptive learning theories can also generate novel learning dynamics. Survey data, estimates of expectations based on market data, and experimental results are increasingly studied.

'Expectations' in economics refers to the forecasts or views that decision-makers hold about future prices, sales, incomes, taxes, or other key variables. The importance of expectations is due to their often substantial impact on the current choices of firms and households, and hence on current prices and the overall level of economic activity. For example, present value calculations in investment decisions are conditional on expected future prices and sales. Equity prices, interest rates, and exchange rates all clearly depend on expected future prices.

Modern economic theory recognizes that the central difference between economics and natural sciences lies in the forward-looking decisions made by economic agents. Therefore, expectations are a basic building block of economic theories. A central aspect of economic theories is that expectations influence the time path of the economy, and conversely one might reasonably hypothesize that the time path influences expectations. The rational expectations (RE) hypothesis, which portrays expectations as an equilibrium in this two-sided relationship, has been the standard methodology for modeling expectations for several decades. Recently, new models and theories have sought to extend and improve the RE approach.

Expectations in the History of Thought

Systematic economic analyses in which expectations play a major role began in the nineteenth century. Alfred Marshall is credited with the notion of 'static expectations' of prices. The 'cobweb' model of a market with a production lag was one of the first formal models with expectations in the 1930s. In the same decade, the temporary equilibrium approach explicitly introduced expectations of future prices influencing current demands and supplies. John Muth was the first to formulate the notion of RE in 1961 (reprinted in Lucas and Sargent, 1981).

In macroeconomic contexts, the importance of the state of long-term expectations of prospective yields for investment and asset prices was emphasized by John Maynard Keynes (1936) in his General Theory. Keynes stressed the central role of expectations for the determination of output and employment, but did not have an explicit model of how expectations are formed. He even sometimes suggested that attempting to forecast very distant future events can virtually overwhelm rational calculation.

In the 1950s and 1960s formal models of expectations were incorporated into almost every area of macroeconomics, including consumption, investment, money demand, and inflation, using adaptive expectations (AE) or related schemes. AE were formally introduced in the 1950s by Phillip Cagan, Milton Friedman, and Marc Nerlove.

RE made the decisive appearance in macroeconomics in the work of Robert E. Lucas Jr. and Thomas J. Sargent in the beginning of 1970s. Many of the key contributions are collected in Lucas and Sargent (1981) and Lucas (1981). The RE hypothesis became widely used in the 1970s and 1980s and it is currently the benchmark paradigm in both micro- and macroeconomics. Most recently, models of *learning behavior* and related research have gone beyond RE by incorporating explicit theories of data collection and forecasting into economic models.

In this article we review models of expectation formation, with an emphasis on RE and more recent developments. RE modeling is the subject of many books, see e.g., Sargent (1987) and Farmer (1999). Evans and Honkapohja (2001) is a treatise on the learning approach.

Traditional Models

We illustrate the modeling of expectations with some well-known simple models. The first example is the cobweb model of a single competitive market with a time lag in production. Demand d_t is assumed to depend negatively on the prevailing market price, while supply s_t depends positively on the expected price, i.e., at time t,

$$d_t = m_I - m_p p_t + v_{1t} \quad \text{and} \quad s_t = r_I + r_p p_t^e + v_{2t}$$

where m_p, $r_p > 0$ and m_I, r_I denote intercepts. Here v_{1t}, v_{2t} are exogenous *iid* (identically and independently distributed) random shocks with mean zero and constant variance. Random shocks of this type are called 'white noise.' The model is illustrated in **Figure 1** in which, for simplicity, we have assumed there are no random shocks, i.e., $v_{1t} = v_{2t} = 0$. Note

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71005-1

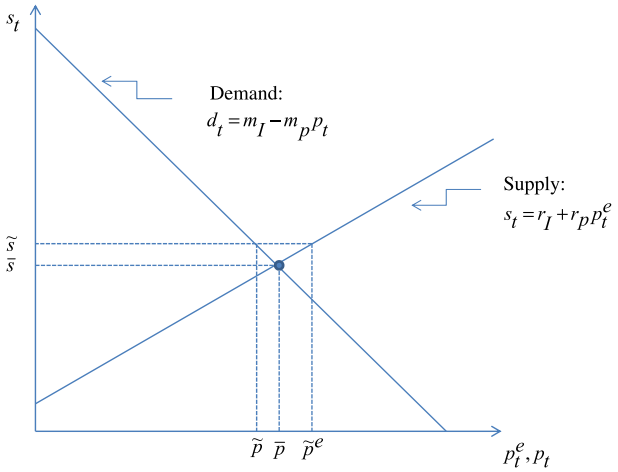

Figure 1 Cobweb model.

that expected price and market-clearing price are on the horizontal axis with corresponding quantities supplied and demanded on the vertical axis.

The supply function reflects a 1-period production lag. Supply decisions for period t are made in the previous period $t-1$ and depend on the price p_t^e expected at time $t-1$ given information available. For simplicity, we make the 'representative agent' assumption that all firms have the same expectation. Extensions to heterogeneous expectations have been analyzed in the literature.

We assume that markets clear. The price p_t at t is then obtained by equating s_t and d_t, which gives

$$p_t = \mu + \alpha p_t^e + \eta_t \qquad [1]$$

where $\mu = (m_I - r_I)/m_p$ and $\alpha = -r_p/m_p < 0$. Here $\eta_t = (v_{1t} - v_{2t})/m_p$ so that η_t is white noise.

In Figure 1, in which there are no random shocks, there is a single price $\bar{p}$ that has the property that expectations will be realized: if $p_t^e = \bar{p}$ then $p_t = \bar{p}$. Here $\bar{p}$ is called the 'perfect foresight' price because firms precisely forecast its value. The corresponding quantity produced is $\bar{s}$, and when this quantity is supplied to the market, the corresponding market-clearing price is indeed $p_t = \bar{p}$. If instead $p_t^e \neq \bar{p}$ then the actual price will be different from the one expected. This is illustrated in Figure 1: if $p_t^e = \tilde{p}^e > \bar{p}$ then firms will supply a large quantity $\tilde{s}$ and this will lead the market-clearing price in period t to be $p_t = \tilde{p} < \bar{p} < \tilde{p}^e$.

If random shocks to supply and demand are included then it is not possible to have perfect foresight. The corresponding concept is instead called RE. Under RE firms will make forecast errors, but these errors will be random and unsystematic, as discussed further below.

Note that the cobweb model leading to eqn [1] has negative feedback from expectations to outcomes, since $\alpha < 0$. This is why high expected price in Figure 1 leads to a low actual or realized price. Positive feedback versions of eqn [1] in which $\alpha > 0$, e.g., macroeconomic models based on the 'Lucas supply curve,' are equally important.

Equation [1] is an example of a temporary equilibrium relationship. This illustrates the central role of expectations by showing how the current market-clearing price depends on expected prices. Developments since the 1930s can be seen as different theories of expectations formation, i.e., how to close the model so that it constitutes a fully specified dynamic theory.

Static expectations and AE were the two models of expectations formation most widely used in the early literature. In the cobweb model static expectations take the form $p_t^e = p_{t-1}$. In effect static expectations assume that the most recently observed price is the best predictor of next period's price. Substituting into [1] one obtains $p_t = \mu + \alpha p_{t-1} + \eta_t$. If there are no random shocks this simply becomes $p_t = \mu + \alpha p_{t-1}$. In terms of Figure 1, and moving to the next time period, the expectation at time t of next period's price is given by $p_{t+1}^e = p_t = \tilde{p}$. Moving up to the supply curve, across to the demand curve and down gives the next period's price, which in the case shown will be somewhat higher than $\bar{p}$.

The sequence of prices over time is equivalently given by the dynamic equation $p_t - \bar{p} = \alpha(p_{t-1} - \bar{p})$, where the perfect foresight price is $\bar{p} = (1-\alpha)^{-1}\mu$. It can be seen that (1) if $\alpha < 0$ prices alternate between positive and negative deviations from $\bar{p}$, and (2) prices converge over time to perfect foresight if and only if $|\alpha| < 1$. If $\alpha < -1$ prices follow an explosive cycle that does not converge. Whether the stability condition $|\alpha| < 1$ is satisfied depends on the relative slopes of the demand and supply curves.

When there are random shocks and static expectations the sequence of prices given by $p_t = \mu + \alpha p_{t-1} + \eta_t$ is random and follows a stochastic process known as an autoregressive process of first order. This and related stochastic processes are studied in standard textbooks on time series analysis and econometrics. In the stochastic case the condition $|\alpha| < 1$ determines whether the price converges to a stationary stochastic process centered on $\bar{p}$.

Beginning in the early 1960s the static expectations hypothesis was largely replaced by AE. In the cobweb model this takes the form

$$p_t^e = p_{t-1}^e + \lambda\left(p_{t-1} - p_{t-1}^e\right)$$

where $0 < \lambda < 1$. Under AE expectations partially respond to the most recently observed price: expectation p_t^e is a weighted average of p_{t-1} and the previous expectation p_{t-1}^e, where λ is the weight on p_{t-1}.

AE supplanted static expectations for three reasons. First, it seems plausible that firms would use all past prices, not just the most recent one, in forming expectations. It can be shown that AE can equivalently be written as a distributed lag average of past prices with weights declining geometrically at rate $1-\lambda$ over past data. Second, it is more general than static expectations, which is equivalent to $\lambda = 1$.

Third, AE tends to be more stable than static expectations. In the nonstochastic case AE converges to perfect foresight if and only if $|1 - \lambda(1-\alpha)| < 1$. In the stochastic case with AE it can be shown that both expectations and prices converge to stationary stochastic processes centered on the perfect foresight price provided the last condition is met. Under AE the stability condition is satisfied for a wider range of parameters than under static expectations. For example, if $\lambda = 1/3$ stability holds if $-5 < \alpha < 1$. Below we show that under econometric learning stability holds for all $\alpha < 1$.

AE played a prominent role in macroeconomics in the 1960s and 1970s. For example, inflation expectations were often modeled this way in the analysis of the expectations augmented Phillips curve. Besides AE, other distributed lag formulations were used in the literature to allow for extrapolative or regressive elements.

Rational Expectations

The RE revolution begins with the observation that AE, or any other fixed weight distributed lag formula, may provide poor forecasts in certain contexts and that better forecast rules may be readily available. The optimal forecast method depends on the stochastic process of the variable being forecast and this implies interdependency between the forecasting method and the economic model. On this approach we write

$$p_t^e = E_{t-1}p_t$$

For the cobweb example $E_{t-1}p_t$ denotes the mathematical expectation of p_t conditional on variables observable at time $t-1$ (including past data).

RE is an equilibrium concept. The stochastic process followed by prices depends on the forecast rules used by agents, so that the optimal choice of the forecast rule by any agent is conditional on the choices of others. An RE equilibrium imposes the consistency condition that each agent's choice is a best response to the choices of others. In the 'representative agent' models these choices are identical.

For the model [1] we have $p_t = \mu + \alpha E_{t-1}p_t + \eta_t$. Taking conditional expectations E_{t-1} on both sides and rearranging yields $E_{t-1}p_t = \bar{p} = (1-\alpha)^{-1}\mu$ and we have $p_t = \bar{p} + \eta_t$. This is the unique way to form expectations that are 'rational' in the model [1].

Two related observations should be made. First, under RE the appropriate way to form expectations depends on the stochastic process followed by the exogenous variables, η_t in the example. If these are not *iid* processes then the RE will themselves be random variables, and they often form a complicated stochastic process. Second, neither static nor AE are in general rational, except in special cases.

Next, we consider some key aspects and implications of RE models.

Forward-Looking Character

In the cobweb model, expectations pertain only to current prices. Most economic models have forward-looking elements, so that expectations about future periods appear. A simple example is the Cagan model of inflation which postulates that the demand for money depends linearly on expected inflation

$$m_t - p_t = -\psi\left(p_{t+1}^e - p_t\right) + \zeta_t, \quad \psi > 0$$

Here m_t is the log of the money supply at time t, ζ_t is an *iid* money demand shock with zero mean, p_t is the log of the price level at time t, and p_{t+1}^e denotes the expectation of p_{t+1} formed in time t. Solving for p_t we get the equation

$$p_t = \alpha p_{t+1}^e + \beta m_t + v_t \qquad [2]$$

where $\alpha = \psi(1+\psi)^{-1}$, $\beta = (1+\psi)^{-1}$, and $v_t = -(1+\psi)^{-1}\zeta_t$. Note that $0 < \alpha < 1$ and $\beta = 1-\alpha$.

The same formal model arises in other contexts, e.g., the standard model of asset pricing with risk neutrality takes form [2] for suitable α and β. Under RE the model is

$$p_t = \alpha E_t p_{t+1} + \beta m_t + v_t \qquad [3]$$

If m_t follows an exogenous stochastic process, one can show that the expression

$$p_t = v_t + \beta \sum_{i=0}^{\infty} \alpha^i E_t m_{t+i} \qquad [4]$$

solves the model, provided the sum converges. Given a specific stochastic process for m_t, it is possible to compute the forecasts $E_t m_{t+i}$ and hence the RE solution [4].

An important feature of [4] is its forward-looking character. Current price is a sum of expected values of future money stocks. If, for example, there is a change in monetary policy that is anticipated to lower the money stock from some specified future date $T > t$, this will be reflected in the price level already at time t. Previously unexpected changes in the policy rule affect p_t from the moment they become known.

More general models than [3] allow for a dependence on lagged values of the endogenous variable p_t, expectations formed at different times and over various horizons, and more general exogenous processes. Generalizations to multivariate frameworks are particularly important in practice.

A number of techniques are available for obtaining solutions under RE. Undetermined coefficient methods are often particularly convenient. Frequently, as in the model [3] with $|\alpha| < 1$, there is a unique nonexplosive solution, given nonexplosive exogenous variables. In this case, the Blanchard–Kahn or other methods can be used to compute this solution.

Policy Evaluation: The Lucas Critique

During the 1950s and 1960s applied macroeconometric models explicitly or implicitly assumed that expectations were given by AE or by some related lag scheme. The estimated systems were used for policy analysis by evaluating the effects of alternative policies. This implicitly assumed that the estimated coefficients reflect a structure that is invariant to alternative policies. Lucas criticized this approach in a 1976 paper (see Lucas, 1981). He argued that if expectations are rational, then the coefficients relating target variables to policy instruments will change when the policy process changes.

As an example, consider the Cagan model [3] with $\beta = 1-\alpha$ and assume that money supply follows the rule

$$m_t = \mu + \rho m_{t-1} + w_t \qquad [5]$$

where $|\rho| < 1$ and w_t is white noise. In this case the solution [4] reduces to

$$p_t = \delta + \gamma m_t + v_t, \quad \text{where} \qquad [6]$$

$$\delta = (1-\alpha\rho)^{-1}\alpha\mu \quad \text{and} \quad \gamma = (1-\rho\alpha)^{-1}(1-\alpha) \qquad [7]$$

The 'traditional' approach to policy would estimate δ and γ from historical data and use the estimated version

of [6] to evaluate the effects of alternative policy sequences m_t on the price level. But it is clear from [7] that γ is not invariant to the policy rule since, for example, a change in the policy parameter ρ affects the value of γ. Thus, under RE the response of p_t to m_t depends on the form of the policy rule.

The effect of the 'Lucas critique' has been to shift attention to estimation of deep parameters (such as α in the above example), which do remain invariant to policy changes and to use these estimates to evaluate alternative policy rules under RE.

Estimation and Testing of RE Models

When data on expectations are not available, it is still often possible in the context of structural models both to estimate the models under the assumption of RE and to test the RE assumption. In this case the RE assumption is being jointly tested with the model. A simple example is the Cagan model [3] with money supply rule [5]. Equations [5] and [6] can both be estimated, but [7] implies nonlinear cross-equation parameter restrictions imposed by the RE hypothesis. The system can be estimated under these overidentifying restrictions. The restrictions can also be statistically tested, in effect testing RE jointly with the model.

Closely related is the estimation and testing of *Euler equations*. These are dynamic equations, relating e.g., current consumption to expected future consumption and interest rates, which arise as optimality conditions from dynamic utility maximization. Under RE testable overidentifying restrictions again arise.

Multiple Equilibria and RE

Many RE models can have multiple equilibria. The existence of rational speculative bubble solutions besides the fundamental equilibrium illustrates this phenomenon. The standard model of asset pricing is formally identical to the Cagan model [2] if we set $v_t \equiv 0, \beta = 1, \alpha = (1+r)^{-1}$, where r is the real net interest rate. Here m_t represents the dividend at t and p_t is the price of the asset. Then [4] is called the 'fundamental solution,' i.e., the present value of anticipated future dividends. A 'rational bubble solution' is the sum of the fundamental solution and a bubble, which is any process B_t satisfying $E_t B_{t+1} = \alpha^{-1} B_t$. The bubble solutions are 'explosive' since $0 < \alpha < 1$, and macroeconomists are often content to assume away 'explosive bubbles.' In models with carefully elaborated microfoundations, explosive solutions can often be ruled out theoretically. However, the empirical issue of whether rational (or near-rational) bubbles affect asset prices is much debated.

Other models have multiple RE solutions, none of which are explosive. For example, consider the model [3] with $|\alpha| > 1$. This case can arise from linearized versions of the overlapping generations model of money under some specifications of preferences. If m_t is *iid* with mean $\overline{m}$, from eqn [4] there is the RE solution

$$p_t = (1-\alpha)^{-1}\alpha\beta\overline{m} + \beta m_t + v_t \qquad [8]$$

However, there are also well-behaved nonexplosive solutions of the form

$$p_t = \alpha^{-1}p_{t-1} - \beta\alpha^{-1}m_{t-1} - \alpha^{-1}v_{t-1} + \varepsilon_t \qquad [9]$$

where ε_t is an arbitrary process, observable at time t and satisfying $E_{t-1}(\varepsilon_t) = 0$.

The variable ε_t in [9] is often called a 'sunspot' variable and the corresponding equilibrium a 'sunspot solution.' The term sunspot is used to indicate the arbitrary nature of the variable. The solution depends on ε_t only because everyone believes that it does. It does not appear in the basic model structure. Guesnerie (2001) provides a review of sunspot equilibria and endogenous fluctuations.

Sunspot solutions have been shown to exist in overlapping generations models with carefully constructed microfoundations. More recently it has been shown that, with increasing returns to scale and monopolistic competition, sunspot solutions can exist in variations of common business cycle models. These models raise the possibility that business cycle fluctuations may be due to random but rational fluctuations in household and firm expectations, which act as a 'self-fulfilling prophecy.' The empirical significance of self-fulfilling prophecies remains an open issue.

The preceding examples involve linear models. Other types of multiplicity of RE solutions arise in nonlinear models. Many nonlinear models can be put in the general form $y_t = F(y^e_{t+1})$, where random shocks have here been omitted for simplicity. If y_t is increasing in y^e_{t+1} there may be multiple steady states satisfying $\overline{y} = F(\overline{y})$ i.e., that occur at the intersections of the graph of $F(.)$ and the 45° line. This possibility can arise in models with increasing returns to scale in production, externalities, or monopolistic competition. y is often a measure of output or aggregate economic activity and the low steady states represent inefficient 'coordination failures.' Other specifications of model $y_t = F(y^e_{t+1})$ can have multiple perfect foresight equilibria taking the form of regular cycles in addition to a steady state.

Nonlinear models can also exhibit sunspot equilibria, taking the form of a finite-state Markov process. For example, in models with two steady states it can be shown that there are RE solutions, depending on a two-state Markov sunspot process, in which the solution alternates randomly between values near the perfect foresight steady states.

Learning and Expectations

The RE approach presupposes that firms and households have a great deal of knowledge about the economy. Even in simple examples, computing the RE coefficients requires full knowledge of the model structure, the values of its parameters and those of the stochastic process of the exogenous shocks. In empirical work economists do not themselves know the parameter values and must estimate them econometrically. A more plausible view of rationality might thus be that the agents also act like statisticians or econometricians when doing the required forecasting. This insight is the starting point of the adaptive learning approach to modeling expectations formation.

Taking this approach immediately raises the question of its relationship to RE. In many cases learning can provide at least an asymptotic justification for the RE hypothesis. For example, in the cobweb model [1], if agents estimate an unknown constant expected value by computing the sample mean from past prices one can show that expectations will converge over time to the RE value. In more general models convergence to RE can occur if agents use the appropriate econometric functional form and run regressions in the same way that an econometrician might.

Another advantage of the learning approach arises when there are multiple equilibria. Consider versions of the model [3] above in which there are two types of solutions [8] and [9]. For the pure RE approach this is a conundrum. Which solution should we and the agents choose? Similarly, in nonlinear models of the form $y_t = F(y^e_{t+1})$ with multiple steady states, cyclic solutions, or sunspots, one can ask which equilibria are stable under learning. The adaptive learning approach provides answers to all these questions, as discussed in the next section.

Finally, the process of learning adds dynamics that are different from RE dynamics. In the examples, just described these dynamics disappear asymptotically. However, there are various situations in which one can expect learning dynamics to remain important over time. For example, if the economy undergoes structural shifts from time to time then agents will need periodically to relearn the relevant stochastic processes. Recent research suggests that these new dynamics can be of empirical importance; see Evans and Honkapohja (2013: Section 2.3) for an overview.

Statistical Approach to Learning

In the adaptive learning approach economic agents behave like statisticians or econometricians when forecasting economic variables needed in their decision-making. At each time the economy is in a temporary equilibrium in which the current state of the economy depends on expectations. The learning approach to expectations formation makes the forecast functions and estimation of their parameters explicit. A novel feature of this approach is that expectations and forecast functions influence future data points.

As an illustration, consider again the cobweb model [1]. Assume agents believe the stochastic process for the market price takes the form $p_t = \text{constant} + \text{noise}$ i.e., the same functional form as the RE solution. The sample mean is the standard way for estimating an unknown constant. Combining the sample mean with eqn [1] leads to a fully specified stochastic dynamical system. It can be shown that the system under learning converges to the RE solution if $\alpha < 1$.

If the economic model incorporates exogenous or lagged endogenous variables, it is natural for agents to estimate the parameters of the perceived process by means of least-squares regressions. For example, suppose an observable exogenous variable w_{t-1} is introduced into the cobweb model, so that eqn [1] becomes

$$p_t = \mu + \alpha p^e_t + \delta w_{t-1} + \eta_t \qquad [10]$$

The unique RE solution is $p_t = \bar{a} + \bar{b} w_{t-1} + \eta_t$, where $\bar{a} = (1-\alpha)^{-1}\mu$ and $\bar{b} = (1-\alpha)^{-1}\delta$ and the corresponding rational forecast is $E_{t-1}p_t = \bar{a} + \bar{b} w_{t-1}$.

If RE is not postulated, it is natural to assume agents forecast price as a linear function of the observable w_{t-1}. Under least-squares learning the forecast of agents is

$$p^e_t = a_{t-1} + b_{t-1} w_{t-1}$$

where a_{t-1} and b_{t-1} are parameter estimates obtained by a least-squares regression of p_t on w_{t-1} and an intercept, using the data available through time $t-1$. It can be shown that (a_t, b_t) converges to the unique RE equilibrium (REE) given by $(\bar{a}, \bar{b})$ if $\alpha < 1$.

Least-squares learning can be generalized to allow for priors on the initial forecast model parameters, and implemented by recursive forms of Bayesian updating. Bayesian learning procedures can also allow for perceived parameter drift.

Stability under Learning

When expectations are modeled by least-squares learning there is convergence to the REE as $t \to \infty$ provided a stability condition is met. This condition can generally be obtained by the expectational stability (E-stability) approach, see Evans and Honkapohja (2001).

To illustrate this approach for the cobweb model [10], suppose expectations are based on the *Perceived Law of Motion* (PLM) $p_t = a + b w_{t-1} + \eta_t$ and hence given by $p^e_t = a + b w_{t-1}$, where (a, b) may not be the REE values. Substituting this into [10] we obtain the corresponding *Actual Law of Motion* (ALM).

$$p_t = (\mu + \alpha a) + (\delta + \alpha b) w_{t-1} + \eta_t$$

This yields a mapping from the PLM parameters (a, b) to the ALM parameters $T(a, b) = (\mu + \alpha a, \delta + \alpha b)$. At the REE $(\bar{a}, \bar{b})$ we have $T(\bar{a}, \bar{b}) = (\bar{a}, \bar{b})$. Expectational stability looks at whether the REE is the outcome of a process in which the parameters of the PLM are adjusted slowly toward the parameters of the induced ALM. Convergence requires fulfillment of the E-stability condition, which in the cobweb model is $\alpha < 1$. For details see Evans and Honkapohja (2001).

Even in fairly complicated models it is often straightforward to compute E-stability conditions and these appear quite generally to govern convergence of statistical learning rules. Local stability under adaptive learning provides a selection criterion in models with multiple RE equilibria. For example, in the nonlinear model $y_t = F(y^e_{t+1})$, only steady states with slope $F'(\bar{y}) < 1$ are E-stable. This model can also have sunspot solutions and an appropriate form of adaptive learning can converge to rational sunspot equilibria, provided the corresponding E-stability condition is satisfied, see Evans and Honkapohja (2001).

Other Approaches and Further Issues

Some other recent approaches to modeling expectation formation are closer to RE than adaptive learning. In 'eductive learning' agents engage in a process of reasoning using common knowledge assumptions and the learning takes place in logical or notional time, e.g., see Guesnerie (2005). In the cobweb model [1] it can be shown that if agents are fully rational and have full understanding of the structure, and if this is common knowledge, then such hyperrational agents will coordinate on the REE if $|\alpha| < 1$.

Another approach that retains aspects of RE but assumes costs and constraints to agents from acquisition of information is called rational inattention (or sticky information) in the recent literature, e.g., see Sims (2011) for a review.

In contrast, the adaptive learning approach assumes a form of bounded rationality. Bounded rationality in learning may result from pertinent issues in real-world forecasting. For example, applied econometricians do not know the true data-generating process. This usually implies that their forecasting models involve misspecification and heterogeneity. Another difficulty is structural change of unknown form that takes place in the economy. The recent literature has introduced these aspects into the adaptive learning approach.

Misspecification and heterogeneity raise the possibility that agents choose between competing forecasting models with the predictive performance of competing models influencing the choice of the model over time. This gives rise to the 'dynamic predictor selection' approach. This choice might also be between heuristic forecasting models, as in Hommes (2013). Another natural approach when multiple models are in play is that agents do some kind of forecasting model averaging.

If structural change of unknown form is a concern, a natural approach is to employ learning rules, such as constant gain algorithms, that can track structural changes more accurately than least-squares algorithms. Constant gain algorithms do not converge fully to an RE equilibrium even if the economic structure is in fact constant. The estimated parameter values contain some randomness asymptotically. In some cases these procedures can lead to escape dynamics in which parameter estimates stray far from the RE solution for an extended period.

The central message from the modified learning procedures just discussed is that variations of econometric learning, which reflect real-world concerns, can lead to new learning dynamics that are qualitatively different from RE dynamics. For example, regime switching and increased volatility can be explained by these procedures. An overview and references to the issues reviewed in this section is provided in Evans and Honkapohja (2013: Sections 2.2.2 and 2.4.2).

Economic Applications

Imperfect knowledge and learning have recently been shown to have major implications to a number of areas in macroeconomics and finance. For overviews of and references to the topics discussed in this section, see Evans and Honkapohja (2013: Sections 2.3, 2.4.2, and 2.5.

The learning viewpoint in the equilibrium models of business cycle fluctuations can mitigate implications of the RE hypothesis that do not match the empirical data. The influence of anticipations of government policies on the outcomes of policies, much emphasized by the RE hypothesis, also appears with adaptive learning by private agents who have knowledge of future policies but the learning dynamics can be quite different. Interestingly, Ricardian equivalence for fiscal policy financing may or may not hold under learning depending on the detailed assumptions.

The learning viewpoint has contributed to thinking about macroeconomic policy, given its role as a source of macroeconomic dynamics. Facilitating stability of private-sector learning is clearly an objective for monetary policy design and this has implications for the formulation of both Taylor-type empirical rules and optimal policy rules. Analysis of exceptional situations involving either rapid inflation or economies with 'liquidity traps' at near-zero interest rates and low or negative inflation has also benefited from the learning viewpoint. These situations are challenging for fiscal policy as well as monetary policy. For a review see Evans and Honkapohja (2009, 2013).

In finance, learning models provide a natural explanation for empirical aspects of stock prices such as mean reversion, excess volatility, persistence, and long-term predictability, which are problematic under RE. Learning has also been employed in recent research on speculative bubbles of asset prices, and can help in overcoming the forward premium puzzle in the empirics of exchange rates, as well as issues of volatility shifts and persistence.

Empirical Issues in Expectations Formation

In Section Economic Applications we discussed economic applications in which models of expectations play a central role. In addition there is extensive research that considers different aspects of empirical data on expectations and aims to test empirically different models of expectation formation.

Testing for Rationality

Given its prominence, testing for rationality is of special interest. If the RE hypothesis is correct, then expectations obey very strong assumptions. Specifically, the orthogonality principle states that if a forecast f_t of a variable p_t, is made based on information set I_{t-1}, then under RE the forecast errors $e_t = p_t - f_t$ must satisfy $E(e_t|I_{t-1}) = 0$. If survey data for f_t is available, the test for RE can be implemented by a regression $e_t = \phi' z_{t-1} + u_t$, where z_{t-1} is a vector of variables in I_{t-1}, and testing the null hypothesis $\phi = 0$.

In practice tests of rationality use survey data on forecasts (see the next section) and the tests typically lead to rejections of rationality, though the interpretation of these results is controversial. While various interpretations are possible, one possibility is that rejections of rationality arise from agents deviating from full RE due to learning, forecasting costs, and/or strategic uncertainty.

Fitting Survey Data on Expectations

There are numerous extensive data sets on expectations of macroeconomic time series, including the EU Commission Consumer Survey, the ECB Survey of Professional Forecasters, the FRB Philadelphia Survey of Professional Forecasters, and the University of Michigan Survey of Consumer Attitudes and Behavior. These survey data are regularly monitored and used in forecasting for economic decision and policy making in practice.

In economic research, these survey data have been used to estimate macroeconomic models in which expectation variables are explicitly included, for an example see Orphanides and Williams (2012). Another use of this data has been testing alternative hypotheses about the expectation formation process, see Pesaran and Weale (2006) and Sinclair (2010) for reviews and recent research.

Using Economic Frameworks to Extract Expectations

Survey data are very important as they provide direct measurements of expectations. The main alternative to measuring expectations is to employ economic models and other data to obtain measurements of expectations. There are two main approaches using this methodology.

Indirect measures of inflation expectations can be obtained from asset market data. The existence of a range of assets, such as indexed and nonindexed bonds, bonds of different maturities, interest rate swaps, and other inflation-indexed derivative assets, the processes of which respond differently to inflation expectations or to expectations of future interest rates, make it possible to extract the inflation or interest rate forecasts of market participants from financial market data. These forecasts are regularly used as information in economic and financial decision-making. A challenge in this data is the need to separate mean dynamics from a risk premium that can vary over time. See Christensen et al. (2010) for an example of measuring inflation expectations and risk premia from bond market data.

The other main methodology to elicit forecasts of economic variables is to employ laboratory experiments. A large literature has recently developed in this area. See, for example, Hommes (2013). One of the main purposes is to use this experimental data to test aspects of expectation formation in a population of agents. These experiments can also investigate the types of economic dynamics that can emerge in models with expectations.

See also: Behavioral Economics; Data Bases and Statistical Systems: Economics (Macroeconomics); Experimental Economics; Monetary Policy Since the 2007–2009 Financial Crisis.

Bibliography

Christensen, J.H.E., Lopez, J.A., Rudebusch, G.D., 2010. Inflation expectations and risk premiums in an arbitrage-free model of nominal and real bond yields. Journal of Money, Credit and Banking (Suppl. 42) 143–178.

Evans, G.W., Honkapohja, S., 2001. Learning and Expectations in Macroeconomics. Princeton University Press, Princeton, NJ.

Evans, G.W., Honkapohja, S., 2009. Expectations, learning and monetary policy: an overview of recent research. In: Schmidt-Hebbel, K., Walsh, C.E. (Eds.), Monetary Policy under Uncertainty and Learning. Bank of Chile, Santiago, pp. 27–76.

Evans, G.W., Honkapohja, S., 2013. Learning as a rational foundation for macroeconomics and finance. In: Frydman, R., Phelps, E.S. (Eds.), Rethinking Expectations: The Way Forward for Macroeconomics. Princeton University Press, Princeton, NJ, pp. 68–111.

Farmer, R.E., 1999. The Economics of Self-fulfilling Prophesies, second ed. MIT Press, Cambridge, MA.

Guesnerie, R., 2001. Assessing Rational Expectations, Sunspot Multiplicity and Economic Fluctuations. MIT Press, Cambridge, MA.

Guesnerie, R., 2005. Assessing Rational Expectations 2, Eductive Stability in Economics. MIT Press, Cambridge, MA.

Hommes, C., 2013. Behavioral Rationality and Heterogeneous Expectations in Complex Economic Systems. Cambridge University Press, Cambridge.

Keynes, J.M., 1936. The General Theory of Employment, Interest and Money. Macmillan, London.

Lucas Jr., R.E., 1981. Studies in Business Cycle Theory. MIT Press, Cambridge, MA.

Lucas Jr., R.E., Sargent, T.J. (Eds.), 1981. Rational Expectations and Econometric Practice. University of Minnesota Press, Minneapolis.

Orphanides, A., Williams, J.C., 2012. Monetary policy mistakes and the evolution of inflation expectations. In: Bordo, M., Orphanides, A. (Eds.), The Great Inflation. National Bureau of Economic Research and University of Chicago Press, Chicago, pp. 255–288.

Pesaran, M.H., Weale, M., 2006. Survey expectations. In: Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of Economic Forecasting, vol. I. Elsevier, Amsterdam, pp. 715–776.

Sargent, T.J., 1987. Macroeconomic Theory, second ed. Academic Press, New York.

Sims, C., 2011. Rational inattention and monetary economics. In: Friedman, B.M., Woodford, M. (Eds.), Handbook of Monetary Economics, vol. 3A. Elsevier, Amsterdam, pp. 155–181.

Sinclair, P. (Ed.), 2010. Inflation Expectations. Routledge, Oxford.

Experiential Psychotherapy

Leslie S Greenberg, York University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Experiential therapy focuses on promoting in-therapy experiencing of what is being talked about. Methods that stimulate emotional experience are used within the context of an empathic facilitative relationship. People are seen as active agents in the construction of their own realities and the self is seen as a process in a continual state of self-organization. The goal of treatment is to promote self-reorganization by integrating disparate parts of the self and by increasing awareness. Awareness of the processes that regulate what enters awareness and what does not is also emphasized. Experiential theorists have specified patterned sequences of change processes that occur in sessions, sequences such as attending to bodily felt sensation, arousing emotion followed by accessing needs. They attempt to facilitate these steps in therapy sessions. The therapist is seen as an expert in the use of methods that promote experiencing and facilitate new steps rather than as an expert on what people are experiencing. In addition, I–thou dialogues between client and therapist about experiences in the relationship are also seen as an important focus of therapeutic work.

Introduction

The term *experiential* is defined in the *Concise Oxford Dictionary* (1990) as "involving or based on experience" and experience is defined as the "actual observation of, or practical acquaintance with, facts or events" or "to feel or be affected by (an emotion, etc.)." A distinction between two ways of knowing, knowledge by acquaintance and knowledge by description, first made by St Augustine, and later emphasized in the epistemologies of William James and Bertrand Russell, helps describe the essence of experiential therapy. Experiential therapy is an approach to therapy that focuses on promoting knowledge by acquaintance. Here a person does not come to know something about him- or herself conceptually but rather has an emotional experience of it. In experiential therapy, the clients' experiencing process is kept as a continuous point of reference for all therapist responses and change is seen as occurring by the promotion of new in-session experience.

Although the term experiential therapy first arose to refer to Whitaker and Malone's approaches to psychodynamically oriented family and individual therapy (Whitaker and Malone, 1953), this approach has subsequently grown out of a variety of 'third force,' humanistic, approaches. These include client-centered (Rogers, 1959), Gestalt (Perls et al., 1951), and existential therapy (May and Yalom, 1995), as well as other approaches that emphasize working with the clients in-therapy experiencing process (Greenberg et al., 1998). Experiential therapy consists of those approaches that, within the context of a facilitative human relationship, focus on the creation of new meaning by the symbolization of experience in awareness. The two major foci of experiential therapy practice are the personal presence of the therapist in the provision of an emphatic and facilitatively genuine therapeutic relationship, and increasing clients' awareness of internal experience by focusing them on their moment-by-moment subjective experience.

View of Human Nature and Functioning

Experiential therapy adopts an existential view of human nature, one that sees existence as preceding essence. People are seen as active agents in the construction of their own realities and choice is seen as the final arbiter in human functioning. Individuals are seen as being born morally neutral with a penchant both for health and sickness, and good and bad. What is essential, however, is that people are seen as having innate worth, the ability to know the difference between good and evil, and the capacity to choose. Free will thus is the foundational principle and therapy involves facilitation of the actualization of the potential for healthy functioning. Health is seen as arising from the ability to symbolize experience in awareness and to integrate different parts of the self to create equilibrium and coherent meaning.

Experiential theory of personality proposes a process model of the self. The self is seen as a dynamic experiential system that is in a continual process of self-organization, i.e., organizing at any moment as the self one finds oneself to be as for example confident, unsure, humorous, or serious. Experiential theory is constructivist in nature and adopts a dynamic and dialectical view of human functioning. It emphasizes that meaning is created by human activity, in dialogue with others, that change is an inherent aspect of all systems, and that this occurs by a synthesis of parts. Meaning construction is viewed as being constrained by a bodily felt emotional experience but as influenced by language and culture and ultimately as created by a synthesis of experience and symbol (Gendlin, 1962; Greenberg et al., 1993; Greenberg and Pascual-Leone, 1995, 1997). Emotional experience is seen as both creating and being created by its conscious symbolization and expression. This view casts people as creators of the self they find themselves to be. In addition, experiential therapy, by adopting a relational view of functioning, sees the self as coming into existence at the boundary between inside and outside, between organism and environment, by a synthesis of bodily experience, symbol, and interpersonal validation.

Emotional experience, although seen as a basically healthy resource, is viewed as capable of either providing healthy adaptive information based on its biologically adaptive origins or, in certain instances, as having become maladaptive through learning and experience. The most basic process for the individual in therapy is thus one of developing awareness of emotion and discriminating which emotional responses are

healthy and can be used as a guide, and which are maladaptive, and need to be changed.

Goals

The general goals of treatment are to promote more adaptive, dynamic organizations and integrative self-organizations. Therapy focuses on the whole person, i.e., it is person-centered rather than problem- or symptom-focused, but within this holistic focus, the underlying determinants of different types of self-dysfunction are also an important point of focus. Thus both a change in manner of functioning of the whole self and changes in particular problems in self-organization are viewed as important. For example, a client may be seen as changing the manner of functioning both by becoming more empathic to the self and by being able to symbolize bodily felt experience, as well as changing by resolving a specific problem such resolving unfinished business with a significant other. Thus, treatment in addition to promoting self-acceptance and strengthening of the self also aims at solving particular problems of self-organization that emerge in treatment. The problems that are focused on emerge in a collaborative fashion over the course of treatment.

Basic Principles of Practice

Experiential therapy recognizes both the power of the understanding relationship and the importance of different in-therapy tasks in promoting different types of therapeutic change. The quality of the bond between participants as well as collaboration on the tasks and goals of therapy are seen as essential in creating a good therapeutic alliance. The bond is warm, respectful, empathic, and validating, while the goals and tasks focus on increasing awareness, deepening experience, and resolving specific in-session emotional problems.

The relational bond is seen as involving three main healing ingredients: first, a more transcendent aspect, the human *presence* of the therapist, witnessing and validating the other's humanness; second, a set of more explicit facilitative *attitudes* that create a safe working environment; and third, a set of specific interpersonal *behaviors* that facilitate growth and provide new interpersonal experience. Buber's (1958) 'I–thou' relationship, involving such elements as presence, commitment to dialogue, and nonexploitiveness, and the Rogerian triad (Rogers, 1957) of empathy, positive regard, and congruence together describe the general nature of the relationship. Empathy is seen as a complex cognitive/affective process of imaginative entry into the world of the other and involves understanding, in addition to being with, the other and the promotion of deeper experience. This process helps the client to feel connected, to regulate affect, and to construct new meaning.

A relational bond of this type is seen as both confirming the client as an authentic source of experience and as providing the optimal context for helping the client to attend to and become aware of implicit, not yet symbolized experience and to communicate and explore it without fear of evaluation. The facilitative relationship, in addition to being curative in and of itself, also provides a safe environment for working on particular problems. Finally, not only does the relationship serve as a confirming environment and as a context for specific forms of intrapsychic work, but it is itself also a medium for specific corrective interpersonal experiences. Thus, certain forms of work on the relationship between client and therapist are also seen as mutative.

In addition to the provision of a 'healing' relationship, facilitating work on particular therapeutic tasks is also seen as a core ingredient of experiential therapy (Greenberg et al., 1993). The most central task of experiential therapy is that of deepening the clients' experience. This involves helping clients focus on their internal experience, helping them to symbolize it in words, and creating a new meaning. The therapist promotes different internal processes at different times to aid experiential processing. The processes facilitated range from symbolizing a bodily felt sense, to evoking memories, to letting an intense feeling form, to expressing feelings, to reflecting on experience to create new meaning. The client in addition is encouraged to engage in such activities as psychodramatic enactments or exercises in imagination to help address particular emotional issues. Specific tasks such as resolving problematic reactions or resolving splits between parts of the self are worked on in therapy. Resolving problematic reactions, for example, involves re-evoking the situation in a vivid manner in order to reexperience what occurred in order to develop a new view of the self and situation. Resolving splits involves the promotion of separation and contact between two parts of the self in a dialogue to facilitate self-acceptance. In these tasks, the client is seen as an active agent engaging in an exploratory process, and the tasks are viewed as being done *by* clients rather than done *to* them.

A Process Theory

An important distinguishing characteristic of experiential therapy is that it offers a process theory of how to facilitate knowledge of acquaintance, rather than a content theory of personality or psychopathology. A process theory of this type specifies both the moment-by-moment steps in the client's process of change and the therapist interventions that will facilitate these steps. The emphasis in each step always is on how to promote the direct sensing of what is concretely felt in the moment to create new meaning.

There is an explicit assumption that within each individual there is a flow of experiencing to which the person can refer to in order to be informed about the personal meaning of particular experiences. The main principle of the experiential method is to have people check whatever is said or done against their own concretely felt experience. Change is seen as emerging from a growing awareness of previously unsymbolized experience and the bringing of this experience into dialectical interaction with words or symbols and other aspect of experience to create new meaning.

The key to experiential therapy is to have clients experience or become acquainted with any manner of content in a new way such that this new experience will produce a change in the way they view themselves, others, and the world. Experiential therapy thus adds to the other approaches the emphasis that

symbols, schemes, and even behavior must interact with the body based, experiential, level of existence in order to produce change. It thus offers a process theory of how body and symbol interact and a set of methods for promoting this process.

Experiential therapy theorists have specified patterned sequences of change processes and events, the explicit connections of which are spelled out by their experiential theories of change. Three such characteristic sequences are described. One sequence offered by Gendlin (1996) includes three fundamental client change events: the client focuses on a directly felt meaning; the client allows feelings, words, and pictures to arise from this inward focusing, and attends to the generated feeling; and the client receives a new felt meaning that emerges from the ensuing shift in body experience. Another experiential sequence (Mahrer, 1989) consists of four client basic change events: the client attains a level of strong feeling; the client welcomes and appreciates the accessed inner experiencing; the client becomes the inner experiencing in the context of earlier life scenes; and the client becomes and behaves according to the inner experiencing in the context of imminent future life scenes.

A third sequence for transforming distressed feelings (Greenberg and Paivio, 1997; Pascual-Leone and Greenberg, 2007) involves changing emotion with emotion (Greenberg, 2010). The process begins with a client in a state of emotional distress, feeling pain, hopeless, helpless, complaint, self-pity, confused, despair, or some other distressing secondary emotional experience. When clients in a state of global distress begin to elaborate and differentiate their thoughts and feelings they subsequently move into maladaptive fear and shame or the sadness of lonely abandonment. The path to resolution involves accessing the unmet in the painful emotion and this helps activate assertive anger or self-soothing and the expression of grief or hurt. All this facilitates a sense of acceptance and agency (Pascual-Leone and Greenberg, 2007).

Transformation occurs when maladaptive states are differentiated into *adaptive needs* which act to refute the core *negative evaluations* about the self embedded in the core maladaptive schemes. The inherent opposition of these two experiences "I am not worthy or lovable" and "I *deserve* to be loved or respected" in response to the same situation is overcome by the creation of new meaning in which a new more *positive evaluation* of the self emerges. Throughout the process of transformation, moderate to high emotional arousal is necessary but at a level that remains facilitative of the healing process. The therapist thus needs to facilitate emotional arousal but not so much that it is dysregulating or disorienting.

In these processes, the therapist is seen as an expert on how to facilitate new steps in the client's experience rather than an expert on the content of the client's experience. Therapists thus avoid interpretations of the content of a client's experiences that tell the patient why he or she does things and that are theory-driven rather than experience-near. Responses that are conceptual or explanatory, or are expressed as fact, or that give the message that truth comes from the therapists' professional knowledge are avoided in favor of phenomenologically refined exploration of the client's experience. Experiential therapy thus attempts to eliminate any interpretations that are based on the therapist's theory of how people are, or should be, or those that attempt to reveal hidden 'truths.'

The experiential approach rather places great emphasis on, and respect for, what the patient experiences, and pays special attention to what the patient experiences in the relationship with the therapist. Experiential therapy involves consistent listening from within the patient's frame of reference. Hence sustained empathic inquiry is a central part of the practice. However, although this empathic emphasis is necessary for good psychotherapy, it is not necessarily sufficient for the best psychotherapy. The most effective psychotherapy also requires therapist practices of a technical nature, such as phenomenological focusing, experimentation, and behavioral observation (Gendlin, 1996; Perls et al., 1951). Diagnostic understanding of the individual, and also understanding the social, cultural, and institutional forces affecting the individual, is required.

Notwithstanding the fact that what the patient experiences is the indispensable essence of psychotherapy, and that it is imperative that this be the subject of respectful ongoing inquiry by therapist and patient, what the patient does not experience is also an indispensable and critical component of what happens in therapy. Important factors that are kept from the patient's conscious experience that do not become figural in awareness also need to be the subject of phenomenological exploration. In experiential therapy, there is a special emphasis on bringing into awareness the processes that regulate this process of awareness/unawareness (Polster and Polster, 1973; Yontef, 1993). This work is codirected by therapist and patient and is not based on any alleged higher truth of the therapist. It is based on a joint empathic exploration that may include observation, experimentation, and dialogue.

Conclusion

In the experiential approach, an empathic focus on the patient's actual experience is seen as indispensable, but the therapist is also seen as making contributions in addition to sustained empathic inquiry. Therapists complement the empathic inquiry with a variety of therapeutic interventions that can help the patient learn how efficiently to focus awareness and distinguish actual phenomenological experience that can highlight the processes essential to the patient's characterological organization by sophisticated and competent understanding that includes understanding theories of healthy functioning and psychopathology.

Experiential therapy has been demonstrated to be effective in treating depression, anxiety, childhood maltreatment, and marital distress in addition to some personality disorders and a variety of other problems in living (Greenberg et al., 1994, 1998).

See also: Constructivist Theory and Psychotherapy; Culture and the Self: Implications for Psychological Theory; Interpersonal Psychotherapy; Motivation in Clinical Interventions; Multimodal Therapy in Clinical Psychology; Personality and Self;

Personality Assessment; Personality Development: Systems Theories; Psychological Treatment, Effectiveness of; Psychological Treatments Delivered Nontraditionally; Psychological Treatments, Empirically Supported; Psychology of the Real Self: Psychoanalytic Perspectives; Psychotherapy and Hermeneutics; Psychotherapy Integration; Self and Emotional Development in Adulthood and Later Life; Self and Identity Development During Adolescence Across Cultures; Self-Categorization Theory; Self-Concept: From Unidimensional to Multidimensional and Beyond; Selfconscious Emotions, Psychology of; Self-Development in Childhood and Adolescence; Self-Efficacy; Self-Esteem; Self-Regulated Learning; Self-regulated Learning: Theories, Measures, and Outcomes; Self-Regulation During Adolescence: Variations Associated with Individual – Context Relations; Selfregulation in Adulthood; Therapist–Patient Relationship.

Bibliography

Buber, M., 1958. I and Thou. Scribner, New York.

Concise Oxford Dictionary, eighth ed. 1990. Clarendon Press, Oxford, UK.

Gendlin, E.T., 1962. Experiencing and the Creation of Meaning. Free Press of Glencoe, New York.

Gendlin, E.T., 1996. A Focusing Approach to Psychotherapy. Guilford Press, New York.

Greenberg, L., 2010. Emotion-Focused Therapy: Theory and Practice. APA Press, Washington, DC.

Greenberg, L., Elliott, R., Lietaer, G., 1994. Research on experiential psychotherapy. In: Bergin, A.E., Garfield, S.L. (Eds.), Handbook of Psychotherapy, Behavior Change. Wiley, New York.

Greenberg, L., Paivio, S., 1997. Working with Emotion in Psychotherapy. Guilford Press, New York.

Greenberg, L., Pascual-Leone, J., 1995. A dialectical constructivist approach to experiential change. In: Neimeyer, R.A., Mahoney, M.J. (Eds.), Constructivism in Psychotherapy. American Psychological Association, Washington, DC, pp. 169–191.

Greenberg, L., Pascual-Leone, J., 1997. Emotion in the creation of personal meaning. In: Power, M., Brewin, C. (Eds.), Transformation of Meaning. John Wiley, Chichester, UK.

Greenberg, L.S., Rice, L.N., Elliot, R., 1993. Facilitating Emotional Change: The Moment by Moment Process. Guilford Press, New York.

Greenberg, L., Watson, J., Lietaer, G. (Eds.), 1998. Handbook of Experiential Therapy. Guilford Press, New York.

Mahrer, A.R., 1989. How to Do Experiential Psychotherapy: A Manual for Practitioners. University of Ottawa Press, Ottawa, Canada.

May, R., Yalom, I., 1995. Existential psychotherapy. In: Corsini, R., Wedding, D. (Eds.), Current Psychotherapies, fourth ed. Peacock, Itasca, IL, pp. 262–292.

Pascual-Leone, A., Greenberg, L., 2007. Emotional processing in experiential therapy: why "the only way out is through". Journal of Consulting and Clinical Psychology 75, 875–887.

Perls, F., Hefferline, R.F., Goodman, P., 1951. Gestalt Therapy. Dell, New York.

Polster, E., Polster, M., 1973. Gestalt Therapy Integrated. Brunner/Mazel, New York.

Rogers, C.R., 1957. The necessary and sufficient condition of therapeutic personality change. Journal of Consulting Psychology 21, 95–103.

Rogers, C.R., 1959. A theory of therapy, personality and interpersonal relationships, as developed in the client-centered framework. In: Koch, S. (Ed.), Psychology: A Study of a Science, vol. 3. McGraw-Hill, New York, pp. 184–256.

Whitaker, C.A., Malone, T.P., 1953. The Roots of Psychotherapy. Blakiston, New York.

Yontef, G.M., 1993. Awareness, Dialogue and Process: Essays on Gestalt Therapy. Gestalt Journal Press, Highland, NY.

Experimental Design: Bayesian Designs

Merlise A Clyde, Duke University, Durham, NC, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 8, pp. 5075–5081, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

This article provides an overview of experimental design using a Bayesian decision-theoretic framework. Scientific experimentation requires decisions about how an experiment will be conducted and analyzed. Such decisions depend on the goals and purpose of the experiment, but certain choices may be restricted by available resources and ethical considerations. Prior information may be available from earlier experiments or from conjectures which motivate the investigation. The Bayesian approach provides a coherent framework where prior information and uncertainties regarding unknown quantities can be combined to find an experimental design that optimizes the goals of the experiment.

This article provides an overview of experimental design using a Bayesian decision-theoretic framework. Scientific experimentation requires decisions about how an experiment will be conducted and analyzed. Such decisions depend on the goals and purpose of the experiment, but certain choices may be restricted by available resources and ethica1 considerations. Prior information may be available from earlier experiments or from conjectures which motivate the investigation. The Bayesian approach provides a coherent framework where prior information and uncertainties regarding unknown quantities can be combined to find an experimental design that optimizes the goals of the experiment.

Introduction

Experimentation plays an integral part in the scientific method. Conjectures and hypotheses are put forth based on the current state of knowledge. Experimental data may be collected to address unknown aspects of the problem. Finally, analysis of experimental results may lead to one hypothesis being favored over others or may lead to new questions and investigations, so that the process is repeated, with the accumulation of additional knowledge about the scientific process under investigation.

In some fields of scientific inquiry, physical models can be used to describe the outcome of an experiment given certain inputs with complete certainty. In the majority of applications, one cannot describe the scientific phenomena perfectly, leading to a distribution of possible outcomes, which can be described by probability models. For example, in comparing a new therapy to an existing treatment, individuals receive one of two treatments. The outcome or response is an indicator of 'success' or 'failure' of the given treatment, which can be modeled using Bernoulli distributions with unknown success probabilities. In psychological or educational testing, the outcome may be scores on a battery of tests. The outcomes or responses may depend on many other factors besides the assigned treatment, such as age, education, gender, or other explanatory variables. Hypotheses or quantities of interest are often phrased in terms of parameters θ of statistical models which relate the distribution of outcomes Y to levels of explanatory variables.

In conducting an experiment, there are many design issues to resolve, including deciding which treatments to study, which factors to control, and what aspects of an experiment to randomize. Other aspects of experimental design – such as how many experimental units are needed, how many observations should be allocated to each treatment, or what levels of other input or control variables should be used – have traditionally fallen under the umbrella of statistical design (see Experimental Design: Overview). Because of costs, ethics, or other limitations on resources or time, sample sizes are usually restricted, and efficient use of available resources is critical. The purpose of optimal experimental design is to improve statistical inference regarding the quantities of interest by the optimal selection of values for design factors under the control of the investigator, within, of course, the constraints of available resources. Decision theory provides a mathematical foundation for the selection of optimal designs. Prior information from earlier related experiments, observational studies, or subjective beliefs from personal observations, can be valuable in deciding how to allocate treatments efficiently, leading to more informative experiments. The Bayesian approach to experimental design provides a formal way to incorporate such prior information into the design process.

Bayesian Optimal Designs

The statistical aspects of an experiment e may be formally described by the sample space Ω (possible outcomes for the response Y), the parameter space Θ, and a probability model $p_e(y|\theta)$ that represents the distribution of observable random variables Y indexed by a parameter θ, an element of the parameter space Θ. Sample sizes, treatment levels, number of treatments, levels of explanatory variables, or other aspects of the design to be selected are implicitly contained in $p_e(Y|\theta)$. The primary goal(s) or terminal decision(s) of an experiment may include, but are not limited to, estimating θ or other quantities of interest that are functions of θ, predicting future

observations, selecting among competing models, or testing other hypotheses.

Lindley presented a two-part decision theoretic approach to experimental design, which provides a unifying theory for most work in Bayesian experimental design today. Lindley's approach involves specification of a suitable utility function reflecting the purpose and costs of the experiment; the best design is selected to maximize expected utility. In this framework, an experiment e is selected from the possible collection of experiments E (the first decision problem). After choosing an experiment e, outcomes Y are observed. Based on the observed data Y and experiment e, a terminal decision d is selected from possible decision rules D, which addresses the terminal goal(s) of the experiment. A utility function in the form $U(d, \theta, e, Y)$ encodes the costs and consequences of using experiment e and decision d with data Y and parameter θ. Assuming that the goals of an experiment and terminal decision can be formally expressed through a utility function, the Bayesian solution is to find the best design and best decision rule that maximize expected utility.

While the process of experimentation followed by inference/decision making proceeds in time order, it is easier to solve the optimal decision problem in reverse time order. The terminal-stage decision problem involves finding the best decision rule d given the observed data Y under experiment e that maximizes the *posterior* expected utility,

$$\max_d \int_\Theta U(d,\theta,e,Y)p(\theta|Y,e)\mathrm{d}\theta = U(e,Y) \qquad [1]$$

Here, the expectation or averaging over θ accounts for uncertainty regarding the unknown θ. The expectation is taken with respect to the posterior distribution of θ, which properly reflects uncertainty in θ at the terminal stage after Y has been observed under experiment e.

As the experiment e must be selected before data Y are observed, the second-stage optimization problem involves finding the best experiment e that maximizes the *preposterior* expected utility. The preposterior expected utility is obtained by integrating the result in Eqn. [1] over possible outcomes in the sample space Ω,

$$U(e) = \int_\Omega U(e,Y)p(Y|e)\mathrm{d}Y = \int_\Theta\int_\Omega U(e,Y)p_e(Y|\theta)p(\theta)\mathrm{d}\theta\mathrm{d}Y \qquad [2]$$

This integral is with respect to $p(Y|e)$, the marginal distribution of the data under experiment e, which is obtained by integrating $p_e(Y|\theta)$, over possible *prior* values for θ, described by prior distribution $p(\theta)$. The Bayesian solution to the experimental design problem is provided by the experiment e^* which maximizes $U(e)$:

$$U(e^*) = \max_e \int_\Omega \max_d \int_\Theta U(d,\theta,e,Y) = p(\theta|\gamma,e)p(\gamma|e)\mathrm{d}\theta\mathrm{d}\gamma \qquad [3]$$

This general formulation can be used to find optimal designs for a single experiment, and can be extended to optimal selection of a sequence of experiments and sequential decision making (Lindley 1972).

Choice of Utility Functions

It is important that utility functions be tailored to the goals of a given problem. Optimal designs for discriminating between two different models may be quite different from designs for prediction. In a one-way analysis of variance model, the best design for comparing k treatments to a control group, is not necessarily the optimal design for estimating the effects of $k+1$ treatments, as these experiments have different goals. While taking an equal number of observations in each of the $k+1$ treatment groups is a possibility, other arrangements may provide more information, particularly when data from previous experiments are taken into account. For example, several previous studies may be available using an existing therapy, but limited information may be available on a new treatment. Differential costs of treatment also need to be considered, which may lead to other allocations of sample sizes or choice of experiments. Ethical considerations may also be incorporated that may constrain the assignment of treatments (Kadane 1996).

Many of the papers in Bayesian design have based utility functions on Shannon information or quadratic loss. Motivation for these choices is discussed below. The reader is encouraged to refer to Chaloner and Verdinelli (1995) for additional details on these criteria, references, and relationships of other Bayesian utility functions to standard optimality criteria.

Shannon Information

Shannon information is appropriate for inference problems regarding θ or functions in θ, without specification of particular hypotheses. The expected utility function is based on the expected change in Shannon information or equivalently the Kullback-Leibler divergence between the posterior and prior distributions. As the prior distribution does not depend on the design, this simplifies to the expected Shannon information of the posterior distribution,

$$U(e) = \int_\Theta\int_\Omega \log\{p(\theta|U,e)\}p_e(\gamma|\theta)p(\theta)\mathrm{d}Y\mathrm{d}\theta \qquad [4]$$

thus the design goal is to find the design that maximizes the information provided by the experiment. In normal linear models with normal prior distributions, this leads to a criterion related to the well-known D-optimality from classical design,

$$U(e) \propto \log|(X_e^T X_e + R)/\sigma^2 \qquad [5]$$

where X_c is the design matrix for experiment e and R/σ^2 is the prior precision matrix (inverse of the prior covariance matrix).

If prediction of future observations is important, the expected gain in Shannon information for a future observation Y_{n+1},

$$U(e) = \int \log(p(Y_{n+1}|Y,e)p(Y_{n+1}|Y,e)p(Y|e)\mathrm{d}Y\mathrm{d}Y_{n+1} \qquad [6]$$

may be relevant. In normal linear models with prediction at a new point x_{n+1} this leads to

$$U(e) \propto \log\left[\left\{x_{n+1}^T(X_e^T X_e + R)^{-1}x_{n+1} + 1\right\}^{-1}\Big/\sigma^2\right] \qquad [7]$$

a Bayesian version of c-optimality.

Quadratic Loss

Point estimation based on quadratic loss leads to the expected utility function

$$U(e) = -\int_{\Theta}\int_{\Omega}(\theta+\widehat{\theta})^{-T}A(\theta-\widehat{\theta}) \qquad [8]$$

where A is a symmetric non-negative definite matrix. The matrix A can be used to weight several different estimation problems where interest may be in estimating individual components of θ or linear combinations of θ. Under the normal linear model and normal prior distribution this results in

$$U(e) = -\sigma^2 tr\left\{A\left(X_e^T X_e + R\right)^{-1}\right\} \qquad [9]$$

a Bayesian generalization of the A-optimality design criterion.

Other Utility and Loss Functions

Utility functions may be based on other loss functions besides quadratic loss. While a quadratic loss function may be appropriate in many cases, there are times when underestimation of a quantity incurs greater losses than overestimation. In such situations, asymmetric loss functions are more appropriate for design and inference (Clyde et al 1996). Discussion of other utility functions related to prediction, hypothesis testing and model discrimination and applications can be found in Chaloner and Verdinelli (1995).

Multiple Objectives

An experiment may often have several, possibly competing objectives which cannot be easily characterized by only one of the standard optimality criteria, and several design criteria may be appropriate. Weighted combinations of utility functions are still valid utility functions, so optimal designs for multiple objectives can be handled within the maximizing expected utility framework. For examples see Verdinelli (1992), Verdinelli and Kadane (1992), Clyde and Chaloner (1996). A difficulty with this approach is that utilities may be expressed in different scales which must be accounted for in the choice of weights. Equivalently, one can find the optimal experiment under one (primary) optimality criterion, subject to constraints on the minimal *efficiency* of the experiment under criteria for other objectives. These approaches are a potential way to address robustness of experimental designs under multiple objectives, models, and prior distributions.

Prior Distributions

Prior elicitation is an important step in designing Bayesian experiments, as well as analysis. Clyde et al (1996) use historical data from previous experiments to construct a hierarchical prior distribution to use for design of future experiments. Kadane (1996) considers many of the practical issues in subjective elicitation for clinical trials. Tsai and Chaloner (2001) describe a design problem where prior distributions are elicited from more than 50 clinical experts.

While some researchers may agree to using prior information to help design an experiment, they may want to have their final conclusions stand on their own, such as in a frequentist analysis. One of the goals of the experiment may be to convince a skeptic, who has different beliefs from the designer of the experiment, that a treatment is effective. In such cases the prior distribution used in constructing the posterior distribution $p(\theta|Y, e)$ in the expected utility Eqn. [1] for the terminal decision problem may be different from the prior distribution $p(\theta)$ used in finding the best design in Eqn. [2]. Etzioni and Kadane (1993) consider the problem of design when the designer and terminal decision maker have different prior beliefs corresponding to different decision rules.

Calculations

Calculation of expected utility, as in Eqns. [1] and [2], requires evaluation of potentially high-dimensional integrals, combined with difficult optimization problems, and in part, has limited the use of Bayesian optimal design in the practice. In normal linear models for many of the standard utility functions presented, the terminal decision rule can be solved in closed form and integrals can be computed analytically, leaving the optimization problem of finding the best experiment e^*. Except in special cases, numerical optimization must often be used to find the optimal design.

One can often relax the problem in the following way. Finding optimal 'exact' designs is often a difficult problem (similar to the traveling salesman problem). Designs can often be viewed in terms of a collection of support points (treatment levels) with weights that indicate the number of observations to be assigned at each support point. An exact design is one where the number of observations at each support point is an integer. A 'continuous' design is obtained by allowing the solution for the weights to be any real number. Rather than finding the optimal exact design, in the relaxed problem, the class of experiments is enlarged to include continuous designs. Mathematically, finding the optimal design in the continuous problem is easier to solve, and methods for checking optimality are often feasible. If the solution corresponds to an exact design, one can then show that it is the globally optimal design. While continuous designs cannot be used in practice, rounding of continuous designs provides exact solutions that are often close to optimal.

For nonlinear models, generalized linear models and other 'nonlinear' design problems (i.e., interest in nonlinear functions in an otherwise linear model), expected utility generally cannot be calculated in closed form and must be approximated. Many of the results for nonlinear design rely on asymptotic normal approximations in calculating expectations (Chaloner and Verdinelli 1995). Integrals may also be approximated by numerical quadrature, Laplace integration or Monte Carlo integration. With advances in computing resources, simulation-based optimal design is now an option, although methods often have to be designed for each specific application (see Müller 1999 for an overview of this area). As utility functions in many problems have been selected for their computational tractability,

simulation-based design may soon open the way for greater use of scientific-based utility functions that better match the goals of the experiment. Hierarchical models are becoming increasingly important in modelling latent and random effects and accounting for subject-to-subject variability, for example. Advances in Bayesian computation, such as Markov chain Monte Carlo, now mean that inference in such models can be carried out in real problems. It is becoming easier to accommodate hierarchical and other complex models in the design of experiments with simulation-based optimal design schemes. Clyde et al (1996) compare analytic approximations and Monte Carlo-based design schemes for a hierarchical binary regression model using an asymmetric loss function.

Applications

Pilz (1991) covers Bayesian design and estimation in linear models, although from a rather mathematical viewpoint. Atkinson (1996) reviews both classical and Bayesian optimal design for linear and nonlinear designs, and presents recent applications such as design for clinical trials. Chaloner and Verdinelli (1995) provide a thorough review of literature on Bayesian designs and applications, which includes experimental designs for linear models such as regression and analysis of variance models, factorial and factional factorial experiments, variance component models, mixtures of linear models, hierarchical models, nonlinear regression models, binary regression models, designs for clinical trials and sequential experimentation. The article includes several worked-out examples, including design for one-way analysis of variance. The article by Clyde et al (1996) explores simulation-based design for a hierarchical logistic regression model, and illustrates how to construct prior distributions based on previous experiments. Other examples of simulation-based design appear in Müller (1999). For an example related to social sciences, see Experimental Design: Large-Scale Social Experimentation.

The area of Bayesian design and analysis for clinical trails in both sequential and nonsequential designs is an exciting and active area, with many practical developments. For a survey of literature on sequential design, see Sequential Statistical Methods. The volume edited by Kadane (1996) describes a complete case study and discusses many of the critical issues in design and analysis of clinical trials, considering ethics, prior elicitation, randomization, treatment allocation, utilities, and decision making. Tsai and Chaloner (2001) describe the design of two large clinical trials, where prior distributions are based on eliciting prior opinions from over 50 clinicians. The design problem involves finding a sample size so that consensus of opinion is met with high probability, where consensus means that all clinicians would prescribe the same treatment based on their posterior opinions after the trial. Other recent examples include Rosner and Berry (1995), Simon and Freudman (1997).

While the increased use of Bayesian design in applications is especially encouraging, specialized software is often required. There is a growing need for reliable, user-friendly software for Bayesian design so that the methods can be more accessible and see greater applicability.

Example

The following design problem is concerned with choosing a sample size for an experiment to confirm results from an earlier study. All individuals in the first study had been diagnosed with breast cancer. Each individual had a biopsy of tumor tissue analyzed for expression levels or a protein implicated in progression of the disease; measured protein expression levels were scored between 0 and 8. All 135 individuals received a chemotherapeutic agent, with the goal of achieving pre-operative tumor regression. The measured outcome for each subject was a binary indicator of tumor regression. The researchers expected that the probability of tumor regression would increase as protein expression levels increased, and were surprised that the estimated probabilities reached a maximum around an expression level of 5, and then declined with higher levels. A logical regression model was used to relate expression 1evels to clinical outcome (indicator of tumor regression). The linear predictor in the logistic regression model included a term with a parameter θ, such that for $\theta \geq 1$ the probability of tumor regression increases with expression level, while for $\theta < 1$ the probability of tumor regression increases up to a level of 5, then decreases. From a statistical perspective the evidence in favor of $\theta < 1$ was weak with a Bayes Factor of 1.65 ('not worth a bare mention' on Jeffreys' scale of interpretation for Bayes Factors (Kass and Raftery 1995). The goal of the second study is to investigate whether this decline was a chance occurrence. If the decrease is real, then this may suggest that a higher dose is necessary at higher expression levels, leading to other experiments.

Spezzaferri (1988) presented a utility function for model discrimination which can be used in designs for testing $H_0{:}\theta \geq 1$ versus $H_1{:}\theta < 1$. Spezzaferri's criterion can also incorporate parameter estimation, but for simplicity only model discrimination is considered. The expected utility function for model discrimination can be represented as

$$U(e) = \int_{\Omega} pr(\theta < 1|Y, e)p(Y|e, H_1)\mathrm{d}Y$$

where $p(Y|e, H_1)$, is the marginal distribution of the new data under $H_1{:}\theta < 1$.

The experimental design involves selecting the total sample size N_e, as individuals with specific protein expression levels cannot be selected in advance due to costs of obtaining the results. The expected utility versus negative costs ($-\$2000 \times N$) is illustrated in **Figure 1**, and was constructed using smoothing of utilities generated using Monte Carlo experiments (Müller 1999). While analytic calculations are intractable, calculation of utilities for this model are straightforward via simulation. The marginal distribution for Y given an experiment with sample size N must incorporate averaging over different configurations of expression levels among the N subjects and the unknown parameters in the logistic regression model. Configurations of size N are generated using a multinomial distribution, with a probability vector π. The expression level distribution from the first study is used to construct a conjugate Dirichlet prior distribution for π. Data from the previous experiment is used to construct the prior distribution for θ and other parameters in the logistic model, and construct the

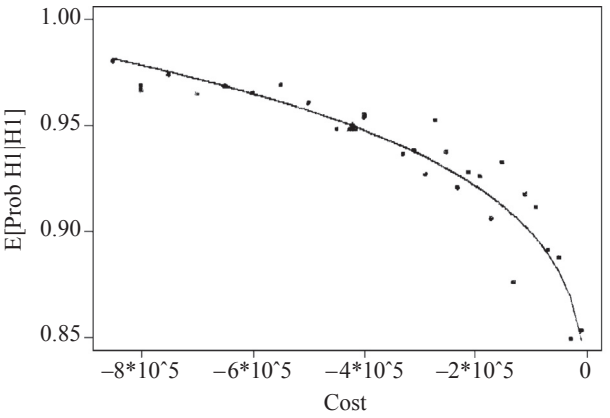

Figure 1 Smoothed expected utility for model discrimination vs. costs; the points indicate simulated expected utilities which exhibit Monte Carlo variation

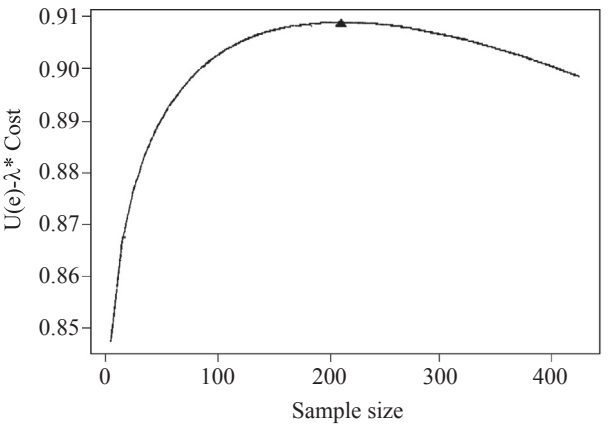

Figure 2 Combined utility for model discrimination with a penalty for costs; the triangle indicates the combined utility at the optimal design

predictive distribution of the new data given a vector of new expression levels. For each set of stimulated data under the new study $Pr(\theta < 1|Y, e)$ is calculated, which incorporates both the old data (in the prior distribution) and the new data. For each sample size N, protein expression levels and outcomes are repeatedly generated from their predictive distributions, and the average $Pr(\theta < 1|Y, e)$ over experiments e with the same sample size N_e provides an estimate of the expected utility for e. These averages were smoothed in order to eliminate Monte Carlo variation in their estimates and are plotted in **Figure 1**.

For the above criterion, expected utility will increase with the total sample size, and as formulated the design problem has an infinite solution. Each point on the curve (**Figure 1**) is a combination of $U(e)$ and cost. Expected utility has a maximum at 1, which is obtained as N goes to infinity (and with infinite costs). The experiment which minimizes costs is given by $N = 0$, where the posterior probability that $\theta < 1$ given the data from the previous study was 0.85. The researchers want to minimize total costs, but at the same time find a sample size such that the expected utility is above 0.95 (corresponding to a Bayes factor above 3, which would provide strong evidence in favor of H_1). Combining costs and model discrimination, the combined expected utility function can be written as

$$U(e) - \lambda(\$2000 * N_c) \qquad [10]$$

where λ reflects the trade-off between for discriminating between the two hypotheses and costs of subjects. In general, it is difficult to specify these trade-offs explicitly, as the different components are often measured in different units. Verdinelli and Kadane (1992) illustrate choosing λ when there are trade-offs between two components in a utility function. If the expected utility for model discrimination is selected as 0.95, then the corresponding cost is \$421 327 (the point represented by the triangle in **Figure 1**), and λ corresponds to the negative of the slope of the curve at this point (λ=0.0001961/\$2 000). **Figure 2** shows the combined utility given by Eqn. [10] using this value of λ. The maximum (marked by the triangle) corresponds to an optimal sample size of 211. For utilities which combine more than two objectives, graphical solutions may not be feasible, but a constrained optimization approach can be used to determine the trade-off parameters (Clyde and Chaloner 1996).

Summary

Bayesian experimental design is a rapidly growing area of research, with many exciting recent developments in simulation-based design and a growing number of real applications, particularly in clinical trials. By incorporating prior information, the Bayesian approach can lead to more efficient use of resources with less costly, and more informative, designs. Utility functions can be explicitly tailored to the given problem and can address multiple objectives. Models, prior distributions, and utility functions have often been chosen to permit tractable calculations. However, with increased computing power, simulation-based methods for finding optimal designs allow for more realistic model, prior and utility specifications. Optimal designs have often been criticized because the number of support points in the design equals the number of parameters (in nonlinear Bayesian design this is not always the case), which does not permit checking the assumed model form. Model uncertainly is almost always an issue in inference, and certainly at the design stage. Bayesian model averaging has been extremely successful in accounting for model uncertainly in inference problems, and has great potential for use in Bayesian experimental design for constructing more robust designs.

Bibliography

Amzal, B., Frédéric, Y.B., Parent, Eric, R., Christian P., 2006. Bayesian-Optimal Design Via Interacting Particle Systems. Journal of the American Statistical Association 101 (474), 773–785.

Atkinson, A.C., Donev, A.N., Tobias, R.D., 2007. Optimum Experimental Designs, with SAS. Oxford University Press.

Atkinson, A.-C., 1996. The usefulness of optimum experimental designs (discussion pp. 95–111). Journal of the Royal Statistical Society, Series B, Methodological 58, 59–76.

Chaloner, K., Verdinelli, I., 1995. Bayesian experimental design: A review. Statistical Science 10, 273–304.

Clyde, M., Chaloner, K., 1996. The equivalence of constrained and weighted designs in multiple objective design problems. Journal of the American Statistical Association 91, 1236–1244.

Clyde, M., Müller, P., Parmigiani, G., 1996. Inference and design strategies for a hierarchical logistic regression mode. In: Berry, D., Stangl, D. (Eds.), Bayesian Biostatistics. Marcel Dekker, New York, pp. 297–320.

Etzioni, R., Kadane, J.B., 1993. Optimal experimental design for an other's analysis. Journal of the American Statistical Association 88, 1404–1411.

Hinkelmann, Klaus, 2012. Design and Analysis of Experiments, Special Designs and Applications. John Wiley & Sons, Hoboken, NJ.

Kadane, J.B., 1996. Bayesian Methods and Ethics in a Clinical Trial Design. Wiley, New York.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.

Lindley, D.V., 1972. Bayesian Statistics—A Review. SIAM, Philadelphia, PA.

Müller, Peter, Sansó, Bruno, De Iorio, Maria, 2004. Optimal Bayesian Design by Inhomogeneous Markov Chain Simulation. Journal of the American Statistical Association 99 (467), 788–798.

Müller, P., 1999. Simulation based optimal design (with discussion). In: Berger, J., Bernardo, J., Dawid, A., Smith, A. (Eds.), Bayesian Statistics 6. Clarendon Press, Oxford, UK, pp. 459–474.

Pilz, J., 1991. Bayesian Estimation and Experimental Design in Linear Regression Models. Wiley, New York.

Rosner, G.L., Berry, D.A., 1995. A Bayesian group sequential design for a multiple arm randomized clinical trial. Statistics in Medicine 14, 381–394.

Simon, R., Freedman, L.S., 1997. Bayesian design and analysis of 2×2 factorial clinical trials. Biometrics 53, 456–464.

Spezzaferri, F., 1988. Nonsequential design for model discrimination and parameter estimation. In: Bernardo, J., DeGroot, M., Lindley, D., Smith, A. (Eds.), Bayesian Statistics 3. Clarendon Press, Oxford, UK, pp. 777–783.

Tsai, C.-P., Chaloner, K., 2001. Using prior opinions to examine sample size in two clinical trials. In: Gatsonis, C., Kass, R.E., Carlin, B., Carriquiry, A., Gelman, A., Verdinelli, I., West, M. (Eds.), Case Studies in Bayesian Statistics, Vol. V. Springer-Verlag, New York.

Verdinelli, I., Kadane, J.B., 1992. Bayesian designs for maximizing information and outcome. Journal of the American Statistical Association 87, 510–515.

Verdinelli, I., 1992. Advances in Bayesian experimental design (with discussion). In: Bernardo, J., Berger, J., Dawid, A., Smith, A. (Eds.), Bayesian Statistics 4. Clarendon Press, Oxford, UK, pp. 467–481.

Experimental Design: Compliance

Alix I Gitelman, Oregon State University, Corvallis, OR, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 8, pp. 5081–5086, © 2001, Elsevier Ltd.

Abstract

In randomized experiments designed to evaluate the relative efficacy of two or more treatments or interventions, subject noncompliance with (or nonadherence to) assigned treatment may seriously jeopardize the interpretation of results from such studies. Essentially, the noncompliant subject self-selects his or her treatment and/or treatment level, thereby disturbing the covariate balance (1995) and its associated safeguard against confounding factors that are afforded by randomized treatment assignment. In the case of educational or behavioral interventions, noncompliance may manifest itself as lowered attendance at intervention sessions or refusal of assigned (or offered) interventions. The causal interpretation of intervention effects may be seriously compromised by confounding factors introduced by subject noncompliance. Approaches to modeling causal treatment effects in light of the noncompliance problem are described using the counterfactual model for causal inference made popular by several studies.

Subject noncompliance with, or nonadherence to assigned treatment in randomized experiments designed to evaluate the relative efficacy of two or more treatments or interventions may seriously jeopardize the interpretation of results from such studies. Essentially, the noncompliant subject self-selects his or her treatment and/or treatment level, thereby disturbing the covariate balance (Rosenbaum, 1995: pp. 14–15) and accompanying safeguard against confounding factors afforded by randomization. Attributing subject outcomes to the treatments under study when subjects are noncompliant is stymied by confounding factors introduced by that noncompliance. For example, in a study to evaluate the relative efficacy of mental health interventions on the quality of life for breast cancer patients (Helgeson et al., 1999), women were asked to attend eight sessions of either an informational seminar or peer support group during the course of their adjuvant chemotherapy and/or radiation treatments. In this case, women's lowered attendance at their assigned interventions may have been related to how well they were tolerating their drug or radiation treatments, which at the same time may have affected their quality-of-life outcomes, so that the effects of the interventions are obscured.

Difficulties with Noncompliance

The complication of noncompliance necessitates the clear distinction between treatment *assignment* and treatment *exposure*. In randomized experiments with perfect compliance, these are the same; but when some subjects are noncompliant, assignment and exposure are different for those noncompliers. This distinction represents the crux of the problem induced by noncompliance: if a study is analyzed without regard to differences in assignment and exposure (e.g., using an intention-to-treat analysis) then the research question that is answered may not be the one that is posed.

Boruch (1997) takes a strong stand for analyzing studies as they are randomized (i.e., without accounting for noncompliance), since compliance behavior during a randomized trial is likely to approximate that same behavior in the general population were the interventions under study to be applied more broadly. Indeed, substantial noncompliance in itself may speak to the efficacy of assignment to the interventions under study. Nevertheless, understanding the relative efficacy of exposure to interventions and possible relationships between compliance behavior and subject outcome warrants a deeper examination of compliance.

Consideration of the statistical problems induced by subject noncompliance has taken place primarily in the field of clinical trials in cases where subjects take less of their assigned drug treatment and/or supplement their treatments with medications or activities other than the ones under study. A direct translation of the compliance problem from the medical setting to the behavioral or education setting comes from considering attendance at sessions of experimental interventions as a 'dose' or 'treatment level.' Alternatively, non-compliance with behavioral interventions might be as simple as subjects not taking advantage of an offered program, such as a family's decision to keep their child in public school even after they have been selected by lottery to attend a private school (see Barnard et al., 1999).

Regardless of the manifestation of noncompliance, questions regarding causal effects of interventions under study remain. Holland (1986) provides an extensive overview of causal inference in statistics, making use of the counterfactual model, the seeds of which date at least to Fisher (1918) and Neyman (1923), and which has been refined over several years by Rubin (1974, 1977, 1978, 1980). This model, relating exposure and outcome, is summarized here to provide a framework for illustrating the noncompliance problem and for describing methods for describing existing methods for addressing that problem (Dawid (2000) argues for causal modeling without using counterfactual outcomes, though addressing noncompliance is greatly facilitated by their use; see discussions following this article).

The Counterfactual Model

The foundation of the counterfactual model is the notion of potential outcomes, which are the collection of outcomes that would be observed for a subject if that subject were exposed to all possible treatments. The words counterfactual and potential suggest quantities that are not observed – outcomes 'counter to the fact,' or that 'might be observed.' Consider N subjects, and let T_i. denote the treatment assignment and X_i the treatment exposure for subject i. For simplicity, let T_i and X_i assume binary values, 0 and 1, representing a control and active treatment, respectively.

The specific quantities of interest for causal inference are subject outcomes under each treatment: $Y_i(0)$ is the outcome that would be observed for subject i if he or she were exposed to treatment 0, and $Y_i(1)$ is the outcome that would be observed for the same subject if he or she were instead exposed to treatment 1. Typically, only one of $Y_i(0)$ or $Y_i(1)$ is observed for each subject. The observed outcome for subject i, Y_i (or $Y_i(X_i)$), is linked to the counterfactuals via the consistency relation

$$Y_i = Y_i(1)X_i + Y_i(0)[1 - X_i] \qquad [1]$$

that is, the observed outcome *is* the potential outcome under the observed treatment exposure.

It should be noted that $Y_i(0)$ and $Y_i(1)$ are reduced from their general form as functions of the entire vector of treatment exposures via the *stable unit treatment value assumption* (*SUTVA*) of Rubin (1980). This assumption asserts that each subject's potential outcomes depend only upon his or her own treatment exposure and not those of any other subjects. In clinical trial settings, where subjects typically do not interact, SUTVA may be a reasonable assumption. In studies of behavioral or educational interventions, however, where subjects may interact with each other extensively, this assumption must be examined more carefully, and the causal conclusions that are based on it may be tenuous (Gitelman, 1999).

The potential outcomes just defined, and SUTVA, are sufficient for defining an individual causal effect (ICE). This is simply the effect, for subject i, of exposure to treatment 1 versus exposure to treatment 0

$$Y_i(1) - Y_i(0) \qquad [2]$$

Alternative individual causal effects such as the ratio of these individual potential outcomes or the difference of their logs could be considered, though the causal parameter obtained by taking the expectation of the difference in eqn [2] argues for using this simpler form. The average causal effect (ACE) is the individual causal effect averaged over the space of subjects

$$\theta = E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)] \qquad [3]$$

The probability distribution over which these expectations are taken is induced by the process by which subjects enter, or are selected for, the study.

In general, θ is nonidentifiable without further assumptions. Rubin (1978) showed, for instance, that θ is identifiable when treatment exposure is 'ignorable,' meaning that each subject's treatment exposure is stochastically independent of his or her collection of potential outcomes (this independence is expressed using the symbol $\coprod$), following Dawid (1979)

$$\{Y_i(0), Y_i(1)\} \coprod X_i \qquad [4]$$

Under randomized treatment assignment and complete compliance, treatment exposure ignorability holds, since $X_i \equiv T_i$ for all subjects i. In this case

$$\begin{aligned} \theta &= E[Y(1)] - E[Y(0)] \\ &= E[Y(1)|X = 1] - E[Y(0)|X = 0] \\ &= E[Y|X = 1] - E[Y|X = 0] \\ &= \alpha \end{aligned} \qquad [5]$$

where the second equality follows from ignorability and the third from the consistency relation in eqn [1]. α is called the apparent treatment effect, and it can be estimated by the difference in sample means of the group of subjects exposed to each treatment, perhaps appropriately scaled by a standard error estimate. Notice that α is typically identifiable regardless of treatment exposure ignorability – it is its equivalence to θ that is afforded by ignorability.

If treatments are randomly assigned, but some subjects are noncompliant, then treatment exposure is probably *not* ignorable. In that case, the apparent treatment effect is not a good estimate of the average causal effect. Specifically, there may be some characteristic of the subject (e.g., in a study comparing educational interventions, his or her academic ability) that predisposes him or her to be noncompliant with treatment assignment. This very characteristic may also be reflected in his or her potential outcomes under that treatment, and therefore his or her outcome cannot be attributed solely to the treatment in question.

One option when faced with noncompliance is to estimate the *intention-to-treat* effect, and use it as an estimate of the average causal effect. The intention-to-treat effect is given by

$$\alpha_{\text{itt}} = E[Y|T = 1] - E[Y|T = 0] \qquad [6]$$

which can be estimated by the difference in sample means of the two treatment groups. This effect explores the relationship between assignment and outcome. Alternatives to this approach that seek to illuminate the relationship between exposure and outcome are described next.

Simple Noncompliance

In the case of binary treatment assignment and binary treatment exposure, noncompliance with treatment assignment manifests as 'treatment switching.' That is, a subject assigned to the active treatment is noncompliant if he or she switches to the inactive treatment (for instance by not taking the treatment at all). Similarly, a subject assigned to the control treatment is noncompliant by switching to the active treatment. This is essentially the situation described in Barnard et al. (1999), wherein families participated in a lottery to receive funding and other support for private school attendance. Some families, even though they 'won' the lottery, decided to keep their children in public schools – hence, these families are the 'noncompliers.'

Modification of the Counterfactual Model

The counterfactual outcomes described previously depend upon treatment exposure only, whereas to account for noncompliance they must be altered to depend also on treatment assignment. Following Angrist et al. (1996), let $X_i(0)$ and $X_i(1)$ denote potential or counterfactual exposures for subject i; these are the treatments subject i would be exposed to (i.e., based on his or her compliance behavior) under assignment to treatments 0 and 1, respectively. Let $Y_i(T_i, X_i(T_i))$ denote the potential outcome of subject i, under treatment assignment T_i and exposure $X_i(T_i)$. Expression of both the potential exposures and potential outcomes again rely on Rubin's SUTVA. In the binary assignment and binary exposure case, each subject has four potential outcomes, one for each assignment/exposure pair

$$\{Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1)\} \quad [7]$$

For instance, $Y_i(1, 0)$ is the outcome subject i would experience had he or she been assigned to the active treatment and been noncompliant with that treatment, thereby receiving the control treatment.

As in the case of the simpler potential outcomes $Y_i(0)$ and $Y_i(1)$, the outcomes in eqn [7] are related to the observed outcome via a consistency relation

$$Y_i = \sum_{t=0}^{1} \sum_{x=0}^{1} Y_i(t,x) I_{\{T_i = t, X_i = x\}} \quad [8]$$

where $I_{\{\cdot\}}$ is the indicator function. As before, the observed outcome is the potential outcome that corresponds to the observed treatment assignment/exposure pair. Non-identifiability is more drastic in this setting, however, as there are now four counterfactual outcomes to consider for each subject instead of just two.

The causal parameters of interest are the expected values (with expectations taking over the space of subjects) of the potential outcomes in eqn [7]; namely

$$E[Y(0,0)], \quad E[Y(0,1)], \quad E[Y(1,0)], \quad \text{and} \quad E[Y(1,1)]$$

With noncompliance, exposure is not ignorable, and so it is not necessarily the case that $E[Y(0, 0)]$ equals $E[Y(1, 0)]$ or that $E[Y(0, 1)]$ equals $E[Y(1, 1)]$. That is, the average subject outcome of a noncompliant subject assigned to the treatment arm will not, in general, be the same as that of a *compliant* subject assigned to the control arm. Methods to address this complication are described next.

Bounds for ACE Using an Instrumental Variable

A variable T is an *instrumental variable* if the distribution of another variable, Y, depends on T only through a third variable, X. Instrumental variables have been popularized in the econometrics literature (*see* Instrumental Variables in Statistics and Econometrics); they have been used to address noncompliance in other settings (e.g., Angrist et al., 1996). Ideally, T and X should be correlated, and T and Y uncorrelated conditionally on X. Then T can be used as a surrogate or *instrument* for X. An instrumental variable, then, is a variable that is partly defined by a conditional independence relationship, known as the *exclusion restriction*, or *instrumental variable assumption*

$$T \coprod Y|X \quad [9]$$

and partly by the required correlation between T and X. Angrist et al. (1996) translate this assumption into the language of potential outcomes, using assignment as an instrument for exposure

$$Y_i(T_i, X_i(T_i)) = Y_i[X_i(T_i)] \quad [10]$$

that is, a subject's potential outcomes do not depend on the treatment assignment directly, but only through that assignment's effect on the treatment exposure. Angrist et al. (1996) show the equivalence of the counterfactual and structural (or simultaneous) equation models (*see* Simultaneous Equation Estimation: Overview).

The instrumental variable assumption in eqn [10] has the effect of reducing the number of potential outcomes from four, as in eqn [7] to two, simply $Y_i(0)$ and $Y_i(1)$. For each subject, however, each of $Y_i(0)$ and $Y_i(1)$ can be observed in one of two ways, either by the subject's compliance with assigned treatment or their noncompliance with the opposing treatment. By using assignment as an instrument for exposure the average causal effect, θ can be bounded by taking minimums or maximums over these two ways of observing each treatment exposure (of course, these bounds also depend upon the responses having bounded expectations). Balke and Pearl (1994) produce the tightest possible bounds for θ (under binary assignment, binary exposure, and binary outcome) using the instrumental variable assumption and linear programming results.

Angrist et al. (1996) use the instrumental variable assumption, combined with a monotonicity assumption, to estimate θ rather than bound it. This monotonicity assumption applies to the counterfactual exposures – it is assumed that a subject will only be noncompliant by switching from the active treatment to the inactive control, but not vice versa, whereby $Y_i(1, 0)$ is eliminated from the collection in eqn [7]. Because it relates exposure under an active treatment to that under a control treatment, this monotonicity assumption only is applicable in trials that compare an active treatment (or treatments) to a control.

Noncompliance as Underdosing

A more complicated noncompliance situation occurs when noncompliant subjects take a lower-than-intended dose of their assigned treatment (or attend fewer-than-intended sessions of their assigned intervention), rather than switching between treatments. For simplicity, treatment assignment remains binary in this discussion. The counterfactual exposures, $X_i(0)$ and $X_i(1)$, now take categorical or even continuous values, however, and there is no switching between treatments except perhaps in the case of a subject taking a zero dose of an active treatment. In a randomized dose-response experiment, subjects are randomly assigned to different doses of a treatment (i.e., dose is ignorable); but with noncompliance the subjects *choose* those doses (i.e., compliance levels), and

so the resulting treatment outcomes cannot be attributed to dose or compliance levels alone.

In a direct extension of the simple compliance case discussed previously, Goetghebeur and Molenberghs (1996) model counterfactual compliance-response curves (similar to dose-response curves) for categorical exposure levels and binary outcomes, Their model relies on a different monotonicity assumption (this one asserts that increased compliance cannot lead to decreased probability of a successful outcome); an instrumental variable type of assumption and a constant global odds ratio assumption. A subsequent contribution, Goetghebeur et al. (1998), relaxes this last assumption somewhat. An essential difficulty with these models, however, is that due to the need for modeling counterfactual (i.e., unobservable) outcomes, rather strong, untestable assumptions are needed. For example, the constant global odds ratio assumption is untestable, and it dictates an artificial relationship between the counterfactual compliance-response curves.

Efron and Feldman (1991) take the approach of modeling compliance behavior as a fixed attribute of the subject, rather than as a response to treatment assignment. As a fixed attribute, each subject's compliance behavior remains the same, regardless of the treatment to which he or she is assigned – this eliminates the problem of having to address counterfactual exposure levels. Compliance-response curves are estimated with this approach, and with additional modeling assumptions (e.g., linearity of the dose-response relationship), portions of the 'true' dose-response curves are recovered. Again, the true dose-response curves can be fully estimated only if dose level is ignorable.

Robins (1994) explores the slightly different setting of sequential randomized trials, in which subjects are assigned to treatment regimes in series, interspersed with assessment of their medical condition. In this article it is assumed that, conditional on all observable covariates, the treatment exposure regime is ignorable.

Addressing Noncompliance in Practice

When subjects are noncompliant with assignment to experimental interventions, a study loses the solid footing afforded by randomized treatment assignments. The essential question of interest regarding the effect of exposure to interventions is now compromised by the subject's selection of their intervention and/or intervention level. The counterfactual model provides a foundation for examining the relationship between exposure and outcome, though careful consideration of the assumptions of both that model and the additional modeling components needed for identifying causal parameters of interest are crucial and may depend upon the type of study.

When faced with compliance problems, or the possibility of compliance problems, the researcher may be best served by careful planning of intervention delivery with an eye toward averting attendance problems or other lapses in the integrity of the intended treatment (Meyer and Fienberg, 1992). Furthermore, by collecting as many relevant covariates as possible, differential, subject-selected treatment exposure can be alleviated, in part by adjusting for these covariates using, for example, propensity score matching (Rosenbaum and Rubin, 1983). If sufficient subject covariates are observed, it may be reasonable to assume that treatment exposure level is ignorable conditional on these covariates, whereby the causal parameters of interest can be identified. For compliance problems that cannot be addressed a priori, however, the bounding (e.g., Balke and Pearl, 1994) and modeling (e.g., Angrist et al., 1996; Barnard et al., 1999) methods described herein provide useful options.

Bibliography

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91, 444–472.

Barnard, J., Frangakis, C., Hill, J., Rubin, D., 1999. School choice in NY City: a Bayesian analysis of an imperfect randomized experiment. In: Kass, R.E. (Ed.), Proceedings of the Fifth Case Studies in Bayesian Statistics Workshop, vol. 5. Springer-Verlag, New York.

Boruch, R.F., 1997. Randomized Experiments for Planning and Evaluation, A Practical Guide. Sage, Thousand Oaks, CA.

Balke, A., Pearl, J., 1994. Counterfactual probabilities, computational methods, bounds and applications. In: Lopez de Mantaras, R., Poole, D. (Eds.), Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Mateo, CA, pp. 46–54.

Dawid, A.P., 1979. Conditional independence in statistical theory. Journal of the Royal Statistical Society, Series B 41, 1–31.

Dawid, A.P., 2000. Causal inference without counterfactuals. Journal of the American Statistical Association 95, 407–424.

Efron, B., Feldman, D., 1991. Compliance as an explanatory variable in clinical trials. Journal of the American Statistical Association 86, 9–26.

Fisher, R.A., 1918. The causes of human variability. Eugenics Review 10, 213–220.

Gitelman, A.I., 1999. Treatment integrity concerns in comparative education studies (Ph.D. thesis). Carnegie Mellon University, Pennysylvania.

Goetghebeur, E., Molenberghs, G., 1996. Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. Journal of the American Statistical Association 91, 928–934.

Goetghebeur, E., Molenberghs, G., Katz, J., 1998. Estimating the causal effect of compliance on binary outcome in randomized controlled trials. Statistics in Medicine 17 (3), 341–356.

Helgeson, V.S., Cohen, S., Schulz, R., Yasko, J., 1999. Education and peer discussion group interventions and adjustment to breast cancer. Archives of General Psychiatry 56, 340–347.

Holland, P.W., 1986. Statistics and causal inference. Journal of the American Statistical Association 8, 945–970.

Meyer, M.M., Fienberg, S.E., 1992. The Case of Bilingual Education Strategies. National Academy Press, Washington, DC.

Neyman, J., 1923. On the application of probability theory to agricultural experiments. Essay on principles. Koczniki Nauk Roiniczych X, 1–51. Reprinted in: Statistical Science 5, 463–480.

Robins, J.M., 1994. Correcting for noncompliance in randomized trials using structural nested mean models. Communications in Statistics A 23, 2379–2412.

Rosenbaum, P.R., 1995. Observational Studies. Springer, New York.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Rubin, D., 1974. Estimating causal effects of treatment in randomized and nonrandomized studies. Journal of Experimental Psychology 66, 688–701.

Rubin, D., 1977. Assignment of treatment group on the basis of a covariate. Journal of Educational Statistics 21, 1–26.

Rubin, D.B., 1978. Bayesian inference causal effects: the role of randomization. The Annals of Statistics 6, 34–58.

Rubin, D.B., 1980. Discussion of 'randomization analysis of experimental data: the Fisher randomization test,' by D. Basu. Journal of the American Statistical Association 75, 591–593.

Experimental Design: Large-Scale Social Experimentation

Carl N Morris, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C.N. Morris, C.W.F. Chiu, volume 8, pp. 5086–5090, © 2001, Elsevier Ltd.

Abstract

Large-scale social experiments became scientifically feasible in the 1960s, and the US government began funding them to obtain better information for policymakers. The rate of new experiments has risen steadily since then, although the newer experiments tend to be smaller and more focused. This article reviews some of the advantages, disadvantages, and design issues that accompany large-scale, long-term social experimentation.

An experiment is a deliberately planned process to collect data that enable causal inferences and legitimize treatment comparisons. Random assignments of subjects to treatments (including control groups) allow statistically significant differences among treatment outcomes to be attributed solely to the treatment differences. Without strong additional assumptions, observational studies and quasi-experiments cannot justify drawing causal inferences, because only experiments, by definition, assign the treatments randomly (*see* Interrupted Time Series Quasi-Experimental Designs; Experimental Design: Randomization and Social Experiments; Experimental Design: Compliance).

Public welfare relies on experiments for many purposes, including to approve medicines, medical procedures, new foods, and product reliability. Large-scale social experiments differ from most of these experiments, even from many that involve human subjects, in that the treatments in social experiments are social programs, the subjects are humans, and the responses measure outcomes for these people (or families) as they make decisions under different possible programs.

Large-scale public policy experiments in the United States generally are considered to have started with the New Jersey Negative Income Tax Experiment (NJ–NIT) 1968. Orr (1999) puts the number of social experiments (not all large scale) launched per decade in the US at six in the 1960s, at 49 and 84 in the next two decades, and then at 56 for 1991–95.

These experiments were fielded to gather scientific information so that policymakers could know as accurately as possible the costs and benefits of possible social programs. Most social experiments have been initiated and funded by agencies (national, state, or at other political levels) concerned with the welfare of large populations. While these large experiments have been costly and time-consuming, they have found sponsors partly because their costs were dwarfed by the costs of actual programs.

Some Examples

Much has been written about social experiments. Greenberg and Shroder (2004) summarize the duration, cost, treatments, measured outcomes, sample sizes, design, funding sources, developers, evaluators, and results of over 100 of them. Orr (1999), Fienberg et al. (1985), and others in the references here, review some of the main studies and list further sources. This article describes several of the better-known public policy social experiments, each of which monitored human subjects for several years as they made personal or family economic decisions while being eligible for benefits provided by an experimental policy to which they were randomly assigned.

The New Jersey Negative Income Tax Experiment

The NJ–NIT experiment (Fienberg et al., 1985; Greenberg and Shroder, 2004; Orr, 1999; Hausman and Wise, 1985) was inspired by the thesis of Ross (1966) while she was a PhD student in economics at MIT. Ross argued that only a social experiment could resolve disputes over how work patterns of the poor might change if the government were to adopt a Negative Income Tax (NIT) program to supplement their incomes. The Office of Economic Opportunity in the Department of Health, Education, and Welfare (HEW) funded the initial Income Maintenance Experiment (IME) to begin in 1968. Mathematica Policy Research, commissioned to design and field it, randomized a total of 725 experimental households into nine treatment groups, and another 632 households into a control group, with welfare and near-welfare households in four New Jersey cities being enrolled for 3 years each. The nine experimental treatments in this response surface design involved two main variables: one providing a range of income guarantees and the other providing a range of tax rate reductions to families as incentives to earn additional income. The NJ–NIT objectives were to estimate, mainly via regression analyses, how work hours, rates of high-school graduation, and other outcomes depended on these factors.

Difficulties encountered in the NJ–IME included significant nonstationary participant behavior at the outset caused by participants needing time to become familiar with their program incentives, and toward the end caused by some gaming behaviors. The NIT design deliberately correlated the random treatment assignments with family incomes in order to reduce experimental costs. That dangerous strategy complicated the analyses and risked compromising the experiment's randomization. Other design concerns included differences (lack of balance) between treatment and control participants in their initial income reporting, and use of total income as

an eligibility criterion at enrollment with a resulting underrepresentation of working wives.

The Health Insurance Experiment

The need for a health insurance experiment was recognized during the Nixon administration when for the first time the government was considering possible forms of a national health insurance program. Congressional leaders held widely divergent views about possible costs of a national program, and especially about the elasticity of demand for health care. If demand were completely inelastic, previously insured individuals would continue to purchase medical services at the same rate. That would make government sponsorship more affordable. While no experiment could provide all needed information, policymakers funded the health insurance experiment (HIE) because it would narrow the range of disagreements, and help avoid disastrous errors in a national program.

The HIE was fielded in 1974 as the income maintenance experiments wound down. The Office of Economic Opportunity, and later HEW, asked the Rand Corporation to design and conduct the HIE, ultimately detailed in *Free for All* (Newhouse, 1993) by its principal investigator.

The HIE was designed to assess the demand for health care and many other questions, including whether health benefits might derive from a national health insurance program. Nearly 3000 families, each for 3–5 years, were randomized to one of 14 insurance plans, including an HMO group. Insurance plans had two main dimensions, with varying family-level coinsurance rates, and with varying deductible limits (after reaching the limits families were exempted from any further expenses for the years remaining). Anticipating that the world might differ considerably after the results of an 8-year experiment became available, the HIE treatments included no particular proposal, but instead were chosen to provide a range of rates and benefits that could be used reliably to extrapolate HIE results to future legislative proposals.

Demand in the HIE was ultimately found to be elastic. Perhaps primary among HIE findings have been the elasticity estimates that legislators continued to use for years when comparing the costs of new health insurance proposals (Newhouse, 1993; Orr, 1999).

JOBSTART, a Demonstration

The rate of new social experiments continued to increase as the technology for doing them matures, and as new research groups and firms develop the skills to design, conduct, and analyze them. However, the newer experiments have tended to be smaller and more focused than the first ones, many having been designed to demonstrate the effectiveness of an existing program, and to identify areas for improvement. These 'demonstration' experiments are simpler to field than those, like the IME and the HIE, that were designed to provide information on a wide range of unspecified programs.

Perhaps politicians are partial to demonstrations because of their lower costs and because they focus on concrete legislation. Demonstration experiments cost less because some of the participant funding comes through an existing program's benefits, and because focus is exclusively on one treatment and a control group. A complication is that control group subjects, if matched to treatment subjects, will be eligible for the same program benefits and so must be allowed to access them. One way to counter this is to have the treatment be an encouragement for eligible subjects to take advantage of an underused program.

JOBSTART was a demonstration experiment to evaluate the Job Training Partnership Act (JTPA) of 1982 (Greenberg and Shroder, 2004; Cave et al., 1993). Sponsored jointly by the US Department of Labor and by several private foundations, JOBSTART was designed and evaluated by Manpower Demonstration Research Center (MDRC) between 1985 and 1992, and administered by the Local Service Delivery Areas in 13 geographically disparate US sites. Subjects were randomized into treatment and control groups, with about 1150 subjects in each. All subjects were JTPA eligible (a national program), being economically disadvantaged high-school dropouts aged between 17 and 21 years, and all were poor readers. Treatment interventions included two summers of vocational and educational training, support services, job placement assistance, and incentive payments. Subjects were monitored for 4 years.

JOBSTART found that the treatment group had a dramatically higher rate of high-school completion. Women who had no children before the demonstration, but who gave birth afterward, were less likely to receive Aid to Families with Dependent Children (AFDC) payments. The opposite result was found for women who started with children. No major long-term differences were found for employment, earnings, or welfare measures.

The Housing Assistance Supply Experiment (HASE)

Economic theory predicts an elastic supply response to increased demand, so that benefits received by individuals do not necessarily rise in proportion to the dollars provided. Even so, almost all large-scale social experiments have measured demand only, mainly because a supply response is extremely difficult to produce in an experiment. The Housing Assistance Supply Experiment (HASE) is a notable exception (Lowry, 1983). While Housing Allowance Demand Experiments had been fielded to evaluate new housing policies (Bradbury and Downs, 1981), evaluating the supply response to alternative housing programs requires a long-term saturation experiment in which all eligible members of an entire community participate in an experimental program. A supply experiment takes longer because the supply side (the construction industry in this case) needs adequate time to respond to any new demand, generally much longer than individuals need to respond to new opportunities.

The Department of Housing and Urban Development sponsored HASE to learn how the supply side would respond if individuals and families received increased government assistance for making home improvements. HASE ran from 1972 to 1982 in Green Bay, Wisconsin, and in South Bend, Indiana, selected as two small, stable communities in which a meaningful supply response might be stimulated. Long-term benefits were guaranteed to all eligible families for 15 years, well beyond the length of the experiment, to give the construction

community sufficient incentive to relocate workers and businesses into these regions.

Supply experiments are extremely ambitious, but the questions they address are crucial.

Alternatives to Social Experiments

Less expensive alternatives must be considered before undertaking a social experiment. These include expert (and nonexpert) opinions, theoretical models, surveys, observational studies, quasi-experiments (*see* Interrupted Time Series Quasi-Experimental Designs; Experimental Design: Compliance), and, if they exist, appropriate 'natural experiments.' Because these alternatives usually are cheaper, faster, and easier than an experiment, they will be preferred whenever they can provide valid predictions. Even if they cannot produce valid predictions, their consideration is vital as a precursor to designing an efficient social experiment.

These alternatives all have drawbacks. Expert opinions and theoretical models are no better than the experts or models. While surveys can sample randomly from a target population, they cannot assign treatments randomly, and surveyed subjects rarely would know how they would respond to a hypothetical future program. Nonexperimental data, quasi-experiments, and observational studies (both being nonrandomized studies) might be used to extrapolate from data on existing programs to predict outcomes of untried programs. Such predictions, however, are especially unreliable for major program changes, and self-selection in such data may bias results. Natural experiments might occur if a comparable country or region adopts a policy similar to one under consideration, but such often do not exist, and when one does occur, results still must be extrapolated to another country or population.

As none of these alternatives has randomly assigned treatments, selection bias will be present and causal inferences will be suspect. Drivers who use seat belts have lower traffic fatality rates, but those rates are confounded with the same drivers being generally safer drivers. Similarly, individuals who choose health insurance plans with less generous coinsurance rates generally consume fewer health services. Is that only because disproportionately many of them, knowing they are healthier, choose less generous insurance plans that they pay for? If so, their behavior in a future health insurance program that includes the entire population will not be reliably predictable from their past utilizations. A health insurance experiment (being randomized) offers an alternative option to using observational data for predicting the costs of proposed national health insurance programs.

Risks of Social Experiments

Being costly and time-consuming, social experiments have been used sparingly. They might not provide useful information for many years, risking their being irrelevant. Human subjects can refuse to enroll, and enrollees can drop out before completion. If that happens with regularity, the desired randomization fails, and so the experiment fails (*see* Nonsampling Errors).

If large experiments become too visible, their subjects may feel that the situation is artificial and act differently than they would when covered by a national program.

Drastic political, social, legislative, or economic changes can occur during an experiment and potentially invalidate it. However, a control group can be protective in such situations, to the extent that the shock affects all treatments and controls equally. Orr (1999), who as a government economist helped spawn and plan several early public policy experiments, reviews such concerns about social experiments, and he discusses how experiments can protect themselves. Besides the issues mentioned here, he discusses whether and how experiments have affected policy, their credibility, time considerations, communication issues, their generalizability, relevance of their results, the policy environment, and policy links.

Design Issues

Those who design social experiments make numerous crucial choices. They must choose treatments that span the policies that are likely to be considered without wasting too much of the budget by picking an overly large range that provides too little information on the policies that ultimately matter. But, how many sites are needed? Too many sites are difficult to manage, and for a fixed budget, increased management costs that must be paid for by decreased sample sizes. Too few sites restrict generalizability. Should sites be chosen randomly? Probably not if the number of sites must be kept small, because then randomization provides little basis for inference.

More subjects are better, but too many subjects make an experiment unfeasibly expensive. Without strong reasons to do otherwise, it helps to keep the percentage of subjects allocated to each treatment the same in every experimental site. Balanced samples (balance refers to matching the key characteristics of the sample across sites and treatments) have various optimal properties in experimental design, and they enjoy considerable face validity (Morris and Hill, 2000).

Enrollment participations must continue long enough for individuals to learn how to take advantage of experimental programs (and so to reach steady state), and perhaps to discount (or adjust for) the times at the beginning and/or the end of the experiment when subjects may act differently than they would in a continuing government program. However, overly long enrollments waste time and budgets because early results are better for policy, and also, when there are positive correlations within individuals, because for estimating mean responses, one individual for more years provides less information per year than does a larger number of individuals, each for fewer years, assuming an equivalent number of person-years. These transitory issues may make it necessary to field a longer experiment, or at least to guarantee treatment benefits to participants beyond the actual measurement period.

Decisions about what data to collect, about the design of questionnaires, and about how much and how frequently to interview subjects can be as crucial as other statistical design decisions. Interviewing subjects too often or too intensively (respondent burden) risks overstimulation, causing them to become overly aware of participating in an experiment and

inducing artificial behavior, termed 'Hawthorne effects.' Respondent burden can cause dropouts, careless responses, and selective nonresponse. Interviewing subjects too infrequently sacrifices potentially important data, so a balance must be struck. Experimenters can usefully gain a better understanding of respondent burden by subjecting themselves to the same interviewing processes that experimental subjects face.

While large-scale social experiments encounter difficulties beyond those of smaller experiments, they will have greater financial resources to deal with them. Some potential difficulties that concerned HIE designers included creating Hawthorne effects by frequent interviewing; activating participants by obtaining a medical screening examination at enrollment; stimulating health expenditures by having to pay participation incentives to families; and not being sure of the best time horizon (Newhouse et al., 1979). The HIE budget allowed for four (balanced) 'subexperiments' within the main experiment to measure these effects. In one subexperiment, half of the HIE subjects were chosen at random for weekly interviews, and half for biweekly. Similarly, 60% of the HIE subjects were randomly selected for initial screening exams, and the rest not; and 70% were randomly assigned for 3-year enrollments, and the other 30% for 5 years. This allowed testing for these effects and, if found, adjusting for them.

Further Reading

Orr (1999) provides an overview and first-hand examples of the design and implementation of social experiments. Greenberg and Shroder (2004) summarize 143 completed social experiments in the United States, and 75 others that were ongoing at the time of their publication. Each summary describes the target population, policies tested, experimental design, sites, findings, sources of further information, and public access to the data. Campbell's fundamental work in social experimentation is summarized in Campbell and Russo (1999). Boruch (1997) provides reference material on randomization in field experiments, and Neuberg (1989) discusses anomalies of social control experimentation.

See also: Experimental Design: Compliance; Experimental Design: Overview; Nonsampling Errors.

Bibliography

Aigner, D.J., Morris, C. (Eds.), 1979. Journal of Econometrics: 11. Experimental Design in Econometrics (1).

Boruch, R.F., 1975. On common contentions about randomized field experiments. In: Boruch, R.F., Riecken, H.W. (Eds.), Experimental Tests of Public Policy. Westview, Boulder, CO, pp. 108–145.

Boruch, R.F., 1997. Randomized Experiments for Planning and Evaluation: A Practical Guide. Sage, Thousand Oaks, CA.

Bradbury, K., Downs, A. (Eds.), 1981. Do Housing Allowances Work? Brookings Institution, Washington, DC.

Campbell, D.T., Russo, M.J., 1999. Social Experimentation. Sage, Thousand Oaks, CA.

Cave, G., Doolittle, F., Bos, H., Toussaint, C., 1993. JOBSTART: Final Report on a Program for School Dropouts. Manpower Demonstration Research Corporation, New York.

Ferber, R., Hirsch, W.Z., 1982. Social Experimentation and Economic Policy. Cambridge University Press, Cambridge, UK.

Fienberg, S.E., Singer, B., Tanur, J.M., 1985. Large-Scale Social Experimentation in the United States. In: A Celebration of Statistics: The ISI Centenary, vol. 287–326. Springer-Verlag, New York.

Greenberg, D., Shroder, M., 2004. The Digest of Social Experiments, third ed. Urban Institute Press, Washington, DC.

Hausman, J., Wise, D. (Eds.), 1985. Social Experimentation. University of Chicago Press, Chicago.

Lowry, I.S. (Ed.), 1983. Experimenting with Housing Allowances: The Final Report of the Housing Assistance Supply Experiment. Oelgeschlager, Gunn & Hain, Cambridge, MA.

Morris, C.N., 1979. A finite selection model for experimental design of the health insurance study. Journal of Econometrics 11, 43–61.

Morris, C.N., Hill, J.L., 2000. The Health Insurance Experiment: Design Using the Finite Selection Model. Public Policy and Statistics: Case Studies from RAND. Springer, New York.

Neuberg, L.G., 1989. Conceptual Anomalies in Economics and Statistics: Lessons from the Social Experiment. Cambridge University Press, Cambridge, UK.

Newhouse, J.P., 1993. Free for All? Lessons from the RAND Health Insurance Experiments. Harvard University Press, Cambridge, MA.

Newhouse, J.P., Marquis, K.H., Morris, C.N., Phelps, C.E., Rogers, W.H., 1979. Measurement issues in the second generation of social experiments: the health insurance study. Journal of Econometrics 11, 117–129.

Orr, L.L., 1999. Social Experiments: Evaluating Public Programs with Experimental Methods. Sage, Thousand Oaks, CA.

Ross, H., 1966. A Proposal for a Demonstration of New Techniques in Income Maintenance (Mimeo). University of Wisconsin Data Center Archives, Institute for Research on Poverty, Madison, WI.

Experimental Design: Overview

Angela M Dean, The Ohio State University, Columbus, OH, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 8, pp. 5090–5096, © 2001, Elsevier Ltd.

Abstract

Experimental design is the art of devising an experimental plan that maximizes the information that can be obtained on the effects of stimuli (treatments) presented to subjects. The article discusses basic features of good design, including comparison of treatments, randomization, avoidance of bias, and reduction of variability. Designs are classified as between-subjects designs, where subjects are assigned one treatment each, and within-subjects designs, where subjects are assigned sequences of treatments. Between-subjects designs include completely randomized designs, where subjects are assigned completely at random to treatments, and matched designs or block designs, where subjects are divided into groups (blocks) of like subjects and random assignment is done within each block separately. Within-subjects designs include crossover designs and split-plot designs. In these, there can be problems of carryover effects, where each measurement may depend not only upon the treatment given in that time period but also on the preceding treatment(s). Counterbalanced Latin squares, which have every treatment preceding every other treatment the same number of times, can be used to isolate treatment effects from carryover effects.

Many scientific discoveries are made by observing how a change in a stimulus that is presented to a subject or object affects the response (measurement) given by that subject or object. In an experiment, the investigator has direct control over which stimuli are presented to which subjects in which time periods. This control, when exercised correctly, enables the investigator to deduce a 'cause and effect' relationship, that is, the investigator can deduce that a given change in a stimulus causes a given change in the measured response. The plan of how an experiment is to proceed is called the 'design of the experiment.' The art of experimental design is the art of devising an experimental plan that maximizes the information that can be obtained on the effects of the stimuli.

Terminology

Experimentation is used in almost every branch of science, with the result that the terminology used in experimental design is not quite standardized. For example, in some fields, the subject or object which is to be presented with a stimulus and then to be measured is called a 'unit' or an 'experimental unit.' The stimulus, itself, may be called the 'treatment' or the 'level of a factor' or the 'level of an independent variable.' In factorial experiments, where a subject is presented with a combination of different types of stimuli (such as a particular light intensity together with a particular noise level), the combination may be called a 'treatment combination' but, for simplicity, the term 'treatment' may be taken to mean either a single stimulus or a combination of stimuli, depending on the context.

In some experiments, measurements are made on each subject over several time periods. These are known as 'repeated measurements.' The terminology concerning the associated designs again differs among disciplines. The term 'repeated measurements design' may refer solely to a design involving repeated measurements on a subject to whom a single treatment has been administered or it may include designs in which the treatment is changed before each measurement. The latter type of design is also known as a 'within-subjects design' or a 'block design' or a 'crossover design.' All of these designs are grouped under the heading of 'split-plot designs' by some authors, while others reserve this last term for a design with two types of treatment, one of which is held constant throughout the repeated measurements on a subject and the other of which changes before each measurement (see Section Split-Plot Designs).

The Purpose of Experimental Design

The pioneer in statistical experimental design was Sir R.A. Fisher (Fisher, 1951), who was concerned with maximizing the amount of information about agricultural crop production. In the social and behavioral sciences, the questions of scientific interest are very different, but the art of good experimental design is similar.

Every experiment has a budget (time and money) and a limit as to the number of subjects that can be recruited. Also, every experiment has inherent variability: subjects differ from one another in fundamental ways, technicians differ in how they read measuring instruments, subjects and instrument readings vary over time, and so on. Variability translates into uncertainty and uncertainty reduces the amount of useful information. Information gained from an experiment may be viewed either from a sampling perspective or from a Bayesian perspective. In the former, hypothesis tests and confidence intervals are generally used, and the most informative experimental designs yield the most powerful tests and the shortest intervals, while producing unbiased results. For a Bayesian analysis, the most informative designs are those that maximize the expected utility.

The purpose of designing an experiment is (1) to maximize the amount of information that can be gained within a given budget or (2) to minimize the budget required to obtain a given amount of information. Maximizing information is done by controlling and reducing the effect of extraneous

variables and by removing confounding variables and bias, as described in Sections Control of Extraneous and Confounding Variables and Removal of Bias.

Features of Good Design

Comparison

Experiments, by their nature, tend to be comparative. Questions of interest tend to be of the type 'does this treatment elicit a "better" response than that treatment?' and, if so, 'how much better?' Even if a single treatment appears to be the only one of interest, information about its effect on a subject is usually of no value without a comparison with other treatments. For instance, Moore and McCabe (1999: Chapter 3) cite an example of a medical study that showed a gastric freezing technique to be a good treatment for ulcer pain, but in later comparative experiments it was shown that the pain relief using freezing was no better than the relief achieved using the same technique but with no freezing solution. Apparently, the pain relief was due to nothing more than doctors showing concern for their patients (a placebo effect) and the freezing technique was abandoned. A good design for evaluating the effect of a single treatment will, therefore, always include a second treatment, called a control or control treatment, for comparison purposes. In some experiments, the control is the treatment in current use, and in other experiments it is the 'absence of a stimulus.' An experiment with more than one treatment needs no control since the experimental treatments can be compared among themselves. Nevertheless, a control can often add extra information. For example, in an experiment on how different types of background music (such as 'classical,' 'rock,' 'rap,' etc.) affect the time taken to learn a new task, a control treatment might be the absence of background music.

Control of Extraneous and Confounding Variables

A good experimental design allows a particular set of stimuli to be compared with each other with high accuracy. Therefore, any other variable (or factor) that causes the experimental measurements to vary at best reduces the efficiency of the experiment and at worst completely masks the true effect being investigated. For example, in an experiment to investigate the effect of employing different types of memory aid in memorization, the IQ of the subjects may play an important role. The variability of IQ of subjects in the experiment would then contribute to the variability in the measured memorization scores. One way of controlling the effect of such an extraneous variable is to hold the variable fixed during the experiment. For instance, IQ could be held more or less fixed by using as subjects only people with a tested IQ within a certain range. Although it reduces variability and avoids masking the effects of the memory aids, this strategy limits the applicability of the conclusions of the experiment since the results would apply only to the people in the population with IQ within this range. A preferable method of controlling extraneous variability is to use a matched design (see below).

An extraneous variable whose effect is completely muddled with that of the factor(s) of interest is called a confounding variable. In the above example, if all the subjects with high IQ were to be tested using the first memory aid and all those with low IQ were to be tested using the second, then, if high IQ is correlated with memorization ability, the first memory aid will inevitably appear to be the better one, regardless of its true merits.

The masking effect of a confounding variable can be reduced by randomization. The simplest form of randomization leads to a 'completely randomized design' (see Section Completely Randomized Designs); subjects are recruited from the general public as randomly as is possible (see Section Selection of Subjects) and then assigned to the stimuli at random in such a way that every subject has the same chance of being assigned to any one of the stimuli. In such a design, it is likely, although not guaranteed, that each stimulus will receive roughly the same distribution of values of the extraneous variable. If a completely randomized design is used in the above example, the memory aids should receive roughly similar ranges of subject IQs.

In all experiments there are confounding variables that have small effects, which are ignored by the experimenter, and confounding variables that are accidentally overlooked. The use of randomization helps to spread out the effects of these variables so that the response given to any one stimulus is less likely to be inflated upward or downward due to extraneous factors. For a discussion of random assignment using a random number table or a computer random number generator, see, for example, Dean and Voss (1999: Chapters 1 and 3).

A more foolproof method of ensuring similar distributions of IQ levels for each memory aid in the above example is to divide the subjects into groups so that subjects within the same group have similar IQs, and then to make the random assignment of subjects to stimuli within each group separately. The division into groups provides the control necessary for removing the variability and masking due to the extraneous IQ variable, while maintaining the applicability to the general population. This type of design is called a 'matched design' or 'block design' (see Section Matched Pairs and Block Designs). Matched designs and completely randomized design are examples of 'between-subjects designs' (see Section Between-Subjects Designs).

A further possibility is to measure each subject under a sequence of different treatments – a 'crossover design' or 'within-subjects design' (see Section Within-Subject Designs). The use of such a design in the above example would completely control the extraneous variable since it would ensure that the distribution of IQs assigned to each memory aid is identical. However, new extraneous variables have now been introduced such as fatigue on the part of the subject over the course of the experiment. If the effects of the new extraneous variables are likely to be small, they can be ignored, but if they are large, then the variables should be controlled by making sure that subjects are assigned the memory aids in different orders. The orders can be assigned at random for each subject separately – a 'block design,' or by deliberately making sure that each stimulus is viewed by the same number of subjects in each time period – a 'Latin square design.'

Removal of Bias

Since each experiment is run with a particular purpose in mind, experimenters tend to have built-in, although perhaps subconscious, biases toward or against certain treatments. A random

assignment of subjects to treatments and a random ordering of observations ensure that experimenter bias cannot consciously or unconsciously favor one treatment above another.

Subjects' own biases toward the treatments can affect their responses. It may or may not be possible or ethical to conceal the true nature of the experiment from the subjects. It may be possible, however, to mask from both the subjects and the person(s) running the experiment which of the stimuli is the experimental treatment(s) and which is the control (a 'double-blind experiment'). Leach (1991: Section 24, Appendix 1) discusses ethical issues of concealing information from subjects and lists the guidelines published by the British Psychological Society. Kirk (1982: Section 1.5) gives references to ethical guidelines put out by the American Psychological Association, the American Sociological Association, and other bodies.

Planning an Experiment

Guides to planning experiments can be found in many texts, for example, Myers (1979: Chapter 1), Cox (1958), Dean and Voss (1999: Chapter 2), and Leach (1991: Sections 5–9). A protocol, which gives in great detail, step-by-step, how the experiment is to proceed, is usually prepared well in advance. The protocol includes details about subject selection, measurement and data collection methods, preparation of materials, preparation of subjects, and a draft statistical analysis.

A pilot experiment in which a small number of observations are collected is often run early in the planning stage. Although these observations will usually be thrown away when the main experiment commences, the pilot experiment gives an opportunity to check that the experimental procedure will work as planned and that the required analyses are possible. It also gives indications of unexpected important confounding variables and the likely accuracy of the results. It allows problems to be detected and corrected before they arise in the main experiment.

Selection of Subjects

Ideally, the subjects taking part in the experiment are selected at random from the population to which the conclusions of the experiment will be applied (the 'target population'). Since this is not always possible, the experimenter may be forced to use volunteers who will inevitably come from a subset of the population. The results of the experiment may or may not then apply to the entire target population.

The number of subjects required to achieve the goals of powerful hypothesis tests and short confidence intervals, or of maximum expected utility, can be calculated using statistical formulae. The required number of subjects depends upon the design selected for the experiment, on the comparisons of interest, and on the variability of the responses of subjects when assigned the same treatment under identical experimental conditions. In general, 'between-subjects designs' require many more subjects than 'within-subjects designs.' Between-subjects designs are used in fields where subjects can be assigned only a single treatment (as in the evaluation of different teaching methods) and/or where subjects are sufficiently plentiful to offset the subject-to-subject variability. Within-subjects designs are preferred in fields where subjects are scarce or highly variable.

The length of the sequence of treatments that can be presented to any one subject depends upon the nature of the experiment. For example, in experimentation involving multiple visits to a laboratory, subject tolerance can be as low as two or three visits. In experiments in which treatments can be changed rapidly with no long-term effect on the subject, a much longer sequence of treatments can be used, requiring fewer subjects in total. There are exceptional circumstances in which an experiment is conducted on a single subject (see Wilson, 1995; Kratochwill and Levin, 1992), but such experiments cannot give conclusive information about the population as a whole and are not used in standard experimentation.

Models and Analysis

The model links the dependent (response) variable(s) to all of the factors (independent variables) that could influence the response, such as the various stimuli, the subjects, time periods, extraneous variables that were used in determining the design, and other important variables that can be taken into account only during the analysis (called 'concomitant variables' or 'covariates': see, e.g., Maxwell and Delaney, 1990: Chapter 9). All extraneous variables that were ignored in the design are grouped together in a single 'error variable.' A measurement on a subject during a time period before the experiment begins is called a 'baseline measurement' and can be used to increase the accuracy of the results (see, e.g., Jones and Kenward, 1989: Chapter 2, Section 4.4; Cotton, 1998: Chapters 9, 10).

If the distributions of the error variables are identical and approximately normal, then standard analysis of variance techniques can be used. Designs in which treatments are assigned the same numbers of subjects and the same numbers of time periods are the easiest to analyze and interpret. They are also the least sensitive to assumptions about normality of the error distributions and equal variances. When the errors do not follow a normal distribution, other types of analysis are needed such as analysis of generalized linear models and categorical data analysis (e.g., see Ratkowski et al., 1993: Chapter 7; Jones and Kenward, 1989: Chapter 3; Crowder and Hand, 1990: Chapter 8).

Repeated measurements on a subject under a particular treatment may require a time series or regression analysis (see Crowder and Hand, 1990). Bayesian techniques in analyzing crossover designs are mentioned by Jones and Kenward (1989: pp. 80, 235). In the case of correlated responses per subject, multivariate analysis is used for normally distributed errors (see, e.g., Winer et al., 1991: Chapter 4; Johnson and Wichern, 1992; Maxwell and Delaney, 1990: Chapters 13, 14; Crowder and Hand, 1990: Chapter 4; Jones and Kenward, 1989: Chapter 7; Myers, 1979: Chapter 18).

Between-Subjects Designs

Completely Randomized Designs

In a completely randomized design, each subject is assigned to just one treatment. The assignment is done completely at random so that each subject has exactly the same chance of

being assigned to each possible treatment (stimulus or combination of stimuli). Often a restriction is applied so that each stimulus receives the same number of subjects.

Completely randomized designs are simple to use and analyze. They are most suited to situations where subjects are plentiful, where subjects' responses would not be too variable if they were all given the same stimulus under the same experimental conditions, and where experimental conditions can be held constant throughout the experiment.

If there are extraneous variables that add variation to the responses and that can be measured during the experiment (such as age or IQ), their effects can be removed during the analysis (analysis of covariance) (e.g., Winer et al., 1991: Chapter 10). However, where it is possible for the extraneous variables to become confounding variables in an unfortunate randomization, a matched pairs design, block design, or within-subjects design would be preferred (see Sections Matched Pairs and Block Designs and Within-Subject Designs).

As in all designs, it is possible to take repeated measurements in completely randomized designs. Each subject is measured for some number of time intervals after administration of the single treatment. Cotton (1998) calls such a design a 'multigroup split-plot design.'

Matched Pairs and Block Designs

If there are just two treatments of interest, the variability in the responses (observations) due to differences in the subjects themselves can be reduced by pairing the subjects so that, within each pair, the subjects are as alike as possible. For each pair separately, the two subjects are assigned at random to the two treatments (a matched pairs design).

When there are t treatments with $t > 2$, the subjects are matched into groups (or blocks) of size t, and within each group the subjects are assigned at random to the t treatments. This type of design is usually known as a 'randomized block design.'

Block designs are also appropriate when the extraneous variation is due to variables unrelated to the subjects. Experimental conditions cannot always be held constant throughout the experiment, as for example, changes in the weather, use of different testing centers, use of different laboratory technicians, and so on. To combat this, the subjects would be put into groups of size t (not necessarily matched) and, apart from the assignment to different treatments, all subjects within a group would be tested under the same experimental conditions.

If the number of treatments, t, to be compared is large (as is often the case in a factorial experiment where the treatments are combinations of several stimuli), there may not be enough subjects to form groups of t similar subjects nor may it be possible to hold conditions constant for t measurements. In this case, an incomplete block design can be used, where each group of $s(<t)$ subjects is assigned at random to a preselected subset of s treatments (see, e.g., Winer, 1971: Chapter 9; Dean and Voss, 1999: Chapters 11, 13).

Within-Subject Designs

In a within-subjects design, each subject is essentially matched with himself or herself and assigned a sequence of some or all of the treatments. The order or presentation is decided using randomization for each subject separately. The comparison of any two treatments is made for each subject ('within' each subject) and then averaged over all the subjects. This has the advantage that subject-to-subject variability does not play a part in the comparison of the treatments.

In experiments where stimuli are presented to each subject in quick succession, 'carryover effects' can be a problem. For example, a subject asked to work in bright light followed by normal light may perceive the normal light to be darker than if it had been preceded by dim light. These carryover effects can be mitigated by separating the trials by a period of time, called a 'washout period,' in which the subject is asked to do something completely different in some control state. If a long enough washout period cannot be arranged, then the experiment usually is counterbalanced (see Section Latin Squares). Carryover effects are also known as 'residual effects' and the effect of the treatment administered in the current time period is called the 'direct treatment effect.'

Crossover Designs

In the simplest within-subjects design, any randomization of the order of treatments for any subject is accepted. However, if time period effects or carryover effects are thought to be important and are to be included in the model, then it is desirable to exercise control over which treatment sequences are used. In most cases, the carryover effect from a given treatment will be assumed to be the same no matter which treatment follows it. If the treatments interact, then this assumption may not be valid and larger designs and more complicated analyses are needed.

When the number of treatments is small, say two or three, and the subject can be measured over several time periods, then each subject can be assigned some or all of the treatments more than once. For three time periods and two treatments (one assigned to one period and one assigned to two periods), there are 6 possible treatment sequences; with four time periods and two treatments (each assigned to two periods), there are 14 possible treatment sequences; with four time periods and four treatments (each assigned to one period), there are 12 possible sequences, and so on. If there are sufficient subjects, all the possible sequences can be used an equal number of times. If the number of subjects is small, however, it may not even be possible to use each sequence once. Information must then be used about which set of sequences provides the best design. Among the possibilities are 'variance balanced designs' (see Section Balanced Crossover Designs) and 'Latin squares' (see Section Latin Squares).

In general, if it is possible to avoid them, two-period designs are not recommended. Not only can the carryover effects not be estimated independently of the treatment by period interaction but also three-period designs are considerably more efficient (Jones and Kenward, 1989: Section 4.16).

Balanced Crossover Designs

Crossover designs that allow differences between all pairs of treatments to be estimated with the same precision are

called 'variance balanced.' Variance balanced designs include crossover designs that use all possible treatment sequences, and counterbalanced Latin square designs (see Section Latin Squares). Variance balanced designs, efficient for comparing all pairs of treatments, are tabulated by Ratkowski et al. (1993: Chapter 5) and also by Jones and Kenward (1989: pp. 212–214, 223–224). The latter authors also list designs efficient for comparing test treatments with a control.

The treatment given in the last period of a crossover design can be repeated in an extra period. Such designs, called 'extra-period designs,' allow the carryover from the last treatment to be measured, thus increasing the precision of the treatment comparisons.

Balanced designs typically require a large number of subjects when the number of treatments is large and so cannot always be used. An alternative is to use an efficient 'partially balanced' design. These designs allow treatment differences to be estimated with two or three different precisions that are fairly close in value. When the treatments are factorial in nature, the effects of the individual factors (main effects) and the interactions between the factors are usually of primary interest. Variance balance is then desirable for the comparisons of the levels of each factor separately (see Jones and Kenward, 1989: pp. 222–228; Ratkowski et al., 1993: Chapter 6, for tabulated designs).

Latin Squares

A Latin square design is ideal for any experiment in which it is possible to measure each subject under every treatment and, in addition, it is necessary to control for changing conditions over the course of the experiment. A Latin square is a design in which each treatment is assigned to each time period and to each subject the same number of times (see Dean and Voss, 1999: Chapter 12). If there are t treatments, t time periods, and mt subjects, then m Latin squares (each with t treatment sequences) would be used.

Carryover effects are controlled by using Latin squares that are 'counterbalanced' (Cotton, 1993). This means that, looking at the sequences of treatments assigned to all the subjects taken together, every treatment is preceded by every other treatment for the same number of subjects. Counterbalanced Latin squares exist for any even number of treatments and for some odd numbers of treatments (e.g., $t = 9, 15, 21, 27$; see Jones and Kenward, 1989: Section 5.2.2, for references). For other odd numbers, a pair of Latin squares can be used, which between them give a set of $2t$ counterbalanced sequences.

If a carryover effect is expected to persist for more than one time period, then the counterbalancing must be extended to treatments occurring more than one time period prior to the current treatment.

Other Designs

Nested or Hierarchical Designs

It is not unusual for extraneous variables to be 'nested.' For example, if subjects are recruited and tested separately at different testing centers, the subjects are 'nested within testing center.' If subjects are animals such as mice or piglets, then the subjects are naturally nested within litters, which are nested within parent, which may be nested within laboratory. The nesting information can be used in matched designs, since the nesting forms natural groupings of like subjects. For within-subjects designs, the nesting information can be used during the analysis for examining the different sources of extraneous variation. Designs in which different levels of nesting are assigned different treatment factors are called 'split-plot designs' (see Section Split-Plot Designs).

A second type of nesting is a nesting structure within the treatment factors being examined. Examples given by Myers (1979) include memorization of words within grammatical class, time taken to complete problems within difficulty levels, and so on. Models and analyses used in such experiments must reflect the nested treatment structure.

Split-Plot Designs

An experiment with more than one type of stimulus (factor) can be run as a split-plot design with a level of one or more factors being applied to a subject throughout the course of the experiment (as for a between-subjects design), and the levels of the other factor(s) being changed for each time period (as for a within-subjects design). Such designs are sometimes called 'mixed designs.' The stimuli applied as the within-subjects design will generally be measured more accurately than those applied as the between-subjects design, since subject-to-subject variability enters into the comparison of the latter.

Split-plot designs are useful when it is difficult to change the levels of one of the factors. For example, Dean and Voss (1999: Chapter 19) cite an example of an optokinetic experiment on the drift of focus of a subject's eyes from the center of a rotating drum measured under two different lighting conditions. The change of lighting conditions was a time-consuming process, whereas it was simple to change the speed of rotation of the drum. Consequently, each subject was assigned a single lighting condition throughout an entire viewing session, and during the viewing session the subject was assigned a sequence of different speeds.

Optimality and Efficiency of Designs

As pointed out by Cotton (1998), designs that are best (i.e., optimal) for one purpose are not necessarily best for another purpose and compromises may need to be made. The optimal design for investigating the effects in one model may be totally unsuitable for a different model. For example, in a crossover design with two treatments and two time periods, a set of counterbalanced Latin squares provides an optimal design for estimating direct treatment effects and carryover effects. This design does not, however, allow estimation of both a carryover effect and an interaction between treatments and time periods. Thus, if both of these effects are required in the model, then more than two time periods must be used in the experiment.

As a general rule, the most balanced designs are optimal when interest lies equally in all treatment comparisons. The following features are typical characteristics of balance: every treatment is assigned the same number of subjects, every treatment is observed in every time period the same number of

times, and every treatment is preceded by every other treatment (including itself, if possible) for the same number of subjects and in the same number of time periods. When balance is not achievable, computer programs for generating optimal designs are commercially available. For other settings, where comparison of all treatments is not the main goal of the experiment, more sophisticated algorithms are needed; see, for example, Atkinson and Donev (1992) for a discussion of design optimality from a Bayesian perspective.

Bibliography

Atkinson, A.C., Donev, A.N., 1992. Optimum Experimental Design. Oxford Science Publications, Oxford, UK.

Breakwell, G.M., Hammond, S., Fife-Shaw, C. (Eds.), 1995. Research Methods in Psychology. Sage, London.

Cotton, J.W., 1993. Latin square designs. In: Edwards, L.K. (Ed.), Applied Analysis of Variance. Dekker, New York.

Cotton, J.W., 1998. Analyzing Within-Subjects Experiments. Erlbaum, Marwah, NJ.

Cox, D.R., 1958. Planning of Experiments. Wiley, New York.

Crowder, M.J., Hand, D.J., 1990. Analysis of Repeated Measures. Chapman and Hall, London.

Dean, A.M., Voss, D.T., 1999. Design and Analysis of Experiments. Springer Verlag, New York.

Edwards, L.K. (Ed.), 1993. Applied Analysis of Variance in Behavioral Sciences. Dekker, New York.

Fisher, R.A., 1951. The Design of Experiments, sixth ed. Oliver and Boyd, Edinburgh, UK.

Harris, P., 1986. Designing and Reporting Experiments. Open University Press, Milton Keynes, UK.

Johnson, R.A., Wichern, D.W. (Eds.), 1992. Applied Multivariate Statistics, third ed. Prentice Hall, Englewood Cliffs, NJ.

Jones, B., Kenward, M.G., 1989. Design and Analysis of Cross-Over Trials. Chapman and Hall, London.

Keppel, G., 1982. Design and Analysis: A Researcher's Handbook, second ed. Prentice Hall, Englewood Cliffs, NJ.

Kirk, R.E., 1982. Experimental Design: Procedures for the Behavioral Sciences, second ed. Brooks/Cole, Belmont, CA.

Kratochwill, T.R., Levin, J.R., 1992. Single-case Research Design and Analysis: New Directions for Psychology and Education. Erlbaum, Hillsdale, NJ.

Kuehl, R.O., 1994. Statistical Principles of Research Design and Analysis. Duxbury, Belmont, CA.

Leach, J., 1991. Running Applied Psychology Experiments. Open University Press, Milton Keynes, UK.

Maxwell, S.E., Delaney, H.D., 1990. Designing Experiments and Analyzing Data. A Model Comparison Perspective. Wadsworth, Belmont, CA.

Moore, D.S., McCabe, G.P., 1999. Introduction to the Practice of Statistics, third ed. Freeman, New York.

Myers, J.L., 1979. Fundamentals of Experimental Design, third ed. Allyn and Bacon, Boston.

Ratkowski, D.A., Evans, M.A., Alldredge, J.R., 1993. Cross-Over Experiments. Design, Analysis and Application. Dekker, New York.

Senn, S., 1993. Cross-Over Trials in Clinical Research. Wiley, New York.

Wilson, S.L., 1995. Single case experimental designs. In: Breakwell, G.M., Hamond, S., Fife-Shaw, C. (Eds.), Research Methods in Psychology. Sage, London.

Winer, B.J., 1971. Statistical Principles in Experimental Design, second ed. McGraw-Hill, New York.

Winer, B.J., Brown, D.R., Michels, K.M., 1991. Statistical Principles in Experimental Design, third ed. McGraw-Hill, New York.

Experimental Design: Randomization and Social Experiments

Robert F Boruch, Graduate School of Education, University of Pennsylvania, Philadelphia, PA, USA; and University of Pennsylvania, Philadelphia, PA, USA
Benjamin Brumley, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Randomized controlled experiments are designed so as to make fair comparisons about which of two or more interventions works better, and to produce the least equivocal evidence possible about the relative effects of the interventions. The studies of impact may also be called randomized trials or social experiments depending on the particular focus. This article covers these kinds of studies in diverse academic disciplines and service delivery areas, different countries, and various geopolitical jurisdictions.

Introduction

This article defines randomized controlled trials (RCTs) and explains their rationale. The emphasis is on field studies that are designed to produce less equivocal evidence than laboratory experiments about what works and what does not. This article is studded with examples from education, crime and delinquency, social welfare, health care, and other areas. The handling transcends geopolitical boundaries. Important aspects of such trials are considered, including their history and practical concerns in their implementation. The last section references access to further information.

Definition and Rationale

In the simplest RCT, individuals are randomly assigned to one of two or more interventions. The interventions may be different intensities of social service, different mixes of therapies, and so on. The interventions may include a 'control' condition in which no services, or services that are ordinarily available, are provided. For example, to understand whether a new employment and training program works well, one could identify a pool of eligible participants and then randomly allocate some of those eligible individuals to the new program. The remaining eligible individuals are assigned to the alternative intervention. This alternative may be a normally available training service, i.e., a 'control condition,' or a program with a different service.

This random assignment of individuals to a new employment and training program and to an alternative program permits one to make a fair comparison of the outcomes, like wage rates, achieved by the people who were assigned to each group. The value that is added by a new program can be estimated well because the people who were assigned to the conventional services and to the new program do not differ systematically. In statistical vernacular, the comparison is unbiased. Put less formally, the groups are equivalent. Furthermore, one can take into account the chance differences between the groups' outcomes using conventional statistical methods. Both the idea of a fair comparison and the idea of taking statistical chance results into account are matters of serious interest, often but not always, in judicial settings (Kaye and Freeman, 2011).

To estimate the relative effect of a particular intervention, one has to estimate the status of individuals had they not had the intervention. This status is a counterfactual. At times, but not often, a precise forecast can be made about how individuals would fare in the absence of the intervention. Also at times, one may be able to construct what appears to be a fair comparison group, a counterfactual that represents the state of people in the absence of the particular intervention. Under these circumstances, statistical methods that undergird 'quasi-experiments' or 'observational studies' might then be employed to generate evidence about the relative effectiveness of programs. These approaches try to make a fair comparison of the outcome of alternative interventions based on sophisticated statistical models of how people behave, and on conscientious attention to competing explanations about why a difference in the effects of interventions might appear. See Rosenbaum (2010) and Morgan and Winship (2007), for instance.

Often, the researcher *cannot* make good forecasts, or construct defensible ad hoc comparison groups, or account for competing explanations for a difference among interventions in a quasi-experiment or observational study. A randomized trial may then be warranted because it assures that the comparison is unbiased. Because people or entities are randomly allocated into two or more groups, their equivalence assures that competing explanations for people's behaviors are taken into account. Both groups are subject to the same influences, apart from the interventions.

At times entities, such as schools, organizations, or geopolitical jurisdictions, may be randomly assigned to different interventions, rather than individuals. Different phrases are typically used to identify such trials. These phrases include cluster randomized trials, group randomized experiments, and place-based trials, e.g., Hayes and Moulton (2009), Donner and Klar (2010), and Murray (1998). Cluster and place randomized trials have increased in their frequency in developed and developing countries.

A Brief History

Randomized trials in medicine antedate such trials in the social sector. For instance, Claude Bernard's (1957) *Introduction to the Study of Experimental Medicine*, first published in 1865 in France,

outlined a framework for controlled comparisons though he said nothing about the topic of random assignment. In the late 1890s researchers in Denmark appear to have first used randomized groups to understand whether certain diphtheria serum worked (Hrobjartsson et al., 1998). During the same period, psychophysical experiments in German laboratory studies employed random assignment (Stigler, 1986).

Modern RCTs have their origins in the 1940s and 1950s, in studies undertaken by Bradford Hill and others in the United Kingdom to generate high-quality evidence on the effectiveness of streptomycin in treating tuberculosis. Later, Jonas Salk, Paul Meier, and others in the United States successfully used randomized trials to discern the effects of vaccine on the incidence of poliomyelitis. For readers with an interest in the history of randomized experiments in health and other sectors, see Marks (1997) and Maynard and Chalmers (1997).

In the social sciences, precedents include the graduated work incentive trials of the 1970s, tests of employment and training programs in the 1980s, and trials on welfare programs in the 1990s. Fienberg et al. (1985) discuss these and other large-scale economic and policy experiments, and how substantive theory is important in the experiment's design. See the Web site for the Social Research Demonstration Corporation for information on related work in Canada (http://www.policyresearch.gc.ca) or the ERN Digest of Experiments (http://www.ssrn.com/update/ern/ern_random-social-experiments.html) for contemporary reports on economic experiments in many countries.

RCTs have also been employed in education research. The Perry Preschool/High Scope trial on preschool education in the 1960s, for instance, is a precedent which demonstrated that intensive early education services can produce lasting educational effects. The Tennessee STAR trial in the 1990s was another milestone: a large-scale randomized trial on the positive effects of reducing school class size (Mosteller, 1999; Finn and Achilles, 1990). Other uses of randomized trials in education have tested new approaches to tutoring in England and the United States (Fitz-Gibbon, 1990; Fuchs et al., 1997), radio-based mathematics education in Nicaragua (Jamison et al., 1980), cultural enrichment programs for children in the barrios in Colombia (McKay et al., 1978), and cooperative group learning in elementary grades in Australia (Gillies and Ashman, 1998).

Boruch (1997) and Oakley (2000) provide early histories and other illustrations of RCTs; the following covers activity since 2000.

More Recent Examples in Diverse Countries

Randomized trials have been mounted in a growing area of research on microfinance and microenterprise in developing countries so as to estimate sustained benefits of such programs (see Karlan and Appel, 2011). Additional examples of randomized evaluation of economic interventions in developing countries include entrepreneurship promotion efforts in Tunisia (Almeida et al., 2012) and evaluation of the effectiveness of advice via mobile phones on agricultural productivity in Gujarat, India (Cole and Fernando, 2012).

In education, randomized designs have been used to test the effectiveness of conditional cash transfers that pay for families to keep their children in school. For instance, trials have been mounted in Mexico under the PROGRESA program (Behrman et al., 2009). RCTs have been employed to address the effectiveness of early schooling and tutoring on academic outcomes (e.g., Tymms et al., 2011) and trials are increasingly utilized in the evaluation of early schooling programs (e.g., Fantuzzo et al., 2011). Other investigations reflect an increased interest in randomized trials in Liberia, India, and Pakistan to assess different ways of enhancing school accountability (Bruns et al., 2011).

In the criminal justice sector, randomized trials in the United States have been deployed to estimate the effects of police handling of misdemeanor domestic assault, new approaches to courts' handling of certain cases, intensive police patrol strategies, and other interventions (Weisburd and Eck, 2004; Braga, 2001). RCTs on the 'Scared Straight' delinquency prevention program in the United States have identified unintended negative consequences of well-intended programs with 'Scared Straight' participants exhibiting higher rates of delinquency than the control group (see Petrosino et al., 2000). Experimental work in crime prevention, reviewed by Farrington et al. (2010), is occurring in Germany, Switzerland, the United States, and Canada with such trials providing valuable evidence on the efficacy, or lack thereof, for prevention programs across these different political contexts. Killias (2006) reviewed the implications of US work on randomized trials for criminological research in Europe.

RCTs are being employed around the world to test the effectiveness of behavioral health and social welfare interventions. In Switzerland, for example, randomized trials have been carried out to understand the effects of heroin therapy versus conventional treatment for individuals with drug addiction (Killias et al., 2000). Many randomized field trials have been used in Burkina Faso and South Africa, among others, to test behavioral and social approaches to health care, in health testing, and to reduce the risk of various diseases (Auvert et al., 2005; Bisoffi et al., 2009). Additionally, trials are increasing in frequency in the social welfare arena. Solomon et al. (2009) give an account of randomized trials in studies that have been mounted to test programs on psychosocial intervention and on social welfare reform.

Newer areas in which such trials have been mounted include political science. In some of these, the aim has been to understand the effects of campaign financing efforts and voter registration in the United States (Green and Gerber, 2003). More generally, the *Cambridge Handbook of Experimental Political Science*, edited by Druckman et al. (2011), provides an account of cluster randomized trials on plebiscites versus jurisdictional meetings in Indonesia, antiviolence initiatives in Nigeria, corruption detection and prevention in India, reconstruction programs in Liberia, election campaign strategies in Benin, and women's representation innovations in parts of India and Brazil, among others. Kinder's (2011) chapter in that handbook provides a social scientific history, giving attention to the seminal contributions of Donald Campbell.

Implementation of Randomized Controlled Trials

When a simple randomized trial involving two interventions is well designed and deployed, the statistical analyses are usually straightforward; textbooks for the design and statistical analysis

of results of such trials are readily available for single level and clustered trial (multilevel) scenarios (Maxwell and Delaney, 2004; Raudenbush and Bryk, 2002). The operational issues in deploying a field trial, however, are less frequently covered in textbooks. Nonetheless, a number of good sources are available to researchers who aspire to design and conduct one well. These include a special issue of the *Annals of the Academy of Political and Social Sciences* (volume 599, issued in 2005) in which managerial, institutional, and ethical challenges are addressed in cases drawn internationally from health, criminology, education, and welfare, all in the context of cluster randomized trials. A special issue of the *Journal of Experimental Criminology*, edited by Heather Strang in 2012 (volume 8, number three), covers management of experiments in crime prevention and reduction, emphasizing the coalitions that are involved and the development of relationships that assure a well-run trial.

For community-based interventions in the psychosocial services area, Solomon et al. (2009) provide guidance on implementation of trials and the interventions they are designed to test. See Boruch (1997) for earlier work along these lines. Many of the cases discussed in these works are attentive to various considerations in trials including tactics for assuring that the randomization process is not corrupted, that the deployment of the intervention is properly done, that cooperation among essential stakeholders is assured, and that the incentives and disincentives for adherence to the experiment's design are identified.

Other types of reconnaissance before conducting trials are also recommended. Statistical methods exist to plan the number of participants needed for the trial, and meta-analysis and systematic review can inform intervention design. Those conducting trials can capitalize on contemporary statistical methods and software to design a trial that maximizes their capacity to detect statistically significant effects. See, for example, PowerUp! (Dong and Maynard, 2013) and Optimal Design (Raudenbush et al., 2011). While accumulating and synthesizing results of randomized trials are also important in building the knowledge base about the effectiveness of social innovations, advances in meta-analysis and in systematic reviews can help to do this (see Cooper et al. (2009) for the technical state of the art and the Campbell Collaboration for illustrative high-quality products at http://campbellcollaboration.org).

Ethics and Law

For many countries the ethical standards for research (including RCTs) on human subjects have been influenced heavily by the Nuremberg Code and the Helsinki Declaration on medical research. Elements of these have been adopted, at times, to assure that the rights of people are protected in sociological, criminological, educational research, and other social research. The codes are pertinent to RCTs as they are to other kinds of studies.

Such codes usually require, for instance, that individuals be informed that their engagement in a trial is voluntary, if indeed it is, and that they can decline to participate or, having agreed to participate, that they can withdraw. Codes regard it as essential to provide a statement of possible risks and benefits to the participant. In the developed countries, an institutional review board is usually required to examine the experiment's design in order to assure that ethics' standards are met.

Such professional codes have been embodied in government regulations. In the United States, for example, the US Code of Federal Regulations (45 CFR 46) requires that most research organizations in the United States abide by these kinds of standards. Canada, the United Kingdom, and the Nordic countries have analogous regulations. Ethical standards are often reified in law. That is, legislatures, diets, parliaments, and other law-making bodies can influence decisions about whether randomized trials are run. In the United States, for instance, welfare laws that were enacted during the 1990s permitted various states of the United States to depart from national rules. This permission to depart from national law was in the interest of testing unconventional approaches to enhancing the education and employment status of low-income families. Such 'waivers' are an important vehicle for deploying and testing innovations that allow for ordinary societal interest in social improvement.

Until recently, no code of ethics of any professional organization considered directly the topic of trials in which entire organizations or geopolitical jurisdictions or other entities were the targets of random allocation. The absence of such handling led to questions about whose consent ought to be elicited in a trial, whether informed consent could indeed be elicited, how one thinks about risks and benefits, and so on. A Canadian group has developed a series of papers and a consensus document to clarify issues and options (Weijer et al., 2012).

One of the paradoxes of codes of ethics for researchers is that codes of research ethics have not been applicable to professionals who experiment in an uncontrolled way by 'trying something out.' Despite the absence of unequivocal evidence on a treatment's efficacy, for example, a physician may use the treatment on a patient without any ethics board review and without the informed consent that is required in a randomized trial. Similarly, a teacher who tries out new ways of teaching may do so without oversight by any institutional review board.

The designer of a randomized field trial can use at least four simple strategies to tailor the trial's design so as to conform to good standards of ethics and law. In the first place, if a purportedly effective service is in short supply, relative to the demand for the service, random assignment of people to the service can often be justified. That is, many people regard a lottery as the fairest way to allocate a scarce resource.

Second, people recognize that the eligibility of individuals for a particular intervention is an important ingredient in their design of a trial. For example, some criminal offenders clearly need to be jailed. Other offenders clearly should not be put into jail. For still others, the effectiveness of jail would not be known. The randomized study would focus then only on those for whom the benefit of jail to the individual is uncertain. That is, one designs the trial so as to include only those for whom the intervention's effectiveness relative to prison is uncertain.

A third strategy to the ethical design of randomized field trials involves a wait-list approach. That is, equally eligible people are randomly allocated into a group that receives treatment early and others are randomized to receive treatment later. The experiment then compares the state of two groups after the first has completed treatment and before the second group has begun.

A fourth strategy recognizes that, in some cases, random assignment of individuals to treatments is unethical (or infeasible) but recognizes that the random assignment of entities to different interventions is ethical. The trial's designer, for instance, may find that local juvenile facilities are unwilling, on ethical grounds, to assign individuals randomly to different treatments in their facility. The trialists may also find, however, that the same facilities are willing to try out interventions whose effectiveness is unknown, so long as the entire facility is randomly assigned to the intervention. See examples elsewhere in this article on such cluster randomized trails.

Institutional Capacity for Randomized Controlled Trials

Doing a randomized field trial also requires the understanding of managerial issues in field research and how to handle the cultural, ethical, and political-institutional factors that must be taken into account. Over the past 20 years, organizations such as the American Institutes for Research, Manpower Demonstration Research Corporation, Abt Associates, and Mathematica Policy Research conduct large-scale experiments in the United States and at times in low- and middle-income countries. Reports on the issues involved in implementing such studies are given in newsletters issued by the organizations and on their Web sites. In Canada, the Social Research and Demonstration Corporation has done so.

Other organizational arrangements have been developed in universities to facilitate managing and conducting small- to medium-scale trials. In the United Kingdom, for instance, the University of York has created a specialized Institute for Effective Education to this end, and University of Durham has a cadre of people who conduct such trials in the UK and elsewhere in collaboration with researchers from other countries.

Access to Information about Randomized Controlled Trials

Until the early 1990s, there existed no central source of information on randomized trials undertaken by health care researchers. In 1993, the international Cochrane Collaboration (Maynard and Chalmers, 1997) was created to prepare, maintain, and disseminate systematic reviews of studies on the effects of health care programs. As part of its mission to produce systematic reviews, the collaboration evolved into an important major source of information on randomized trials. The Cochrane Library is accessible at www.cochrane.ac.uk.

Until the year 2000, no organization had been invented to provide ready access to randomized trials in education, crime and justice, social work and social welfare, and other social sectors. The international Campbell Collaboration was developed as a younger sibling to the Cochrane Collaboration. The Campbell Collaboration's mission, in part, is to identify trials, screen reports on experiments for their quality, and produce systematic reviews of them. See http://www.campbellcollaboration.org/.

Over the past decade, a number of specialized journals and organizations have been created to focus on design, execution, and reporting on such trials. The peer-reviewed journals that contain full reports on the design and results of trials and that are web accessible include the *Journal of Research on Educational Effectiveness* and the *Journal of Experimental Criminology*. The *Randomized Social Experiments eJournal* produces frequent newsletters in which abstracts of reports on new controlled trials, mainly in economics, are routinely provided, and hyperlinks to full reports are given where possible. See http://www.ssrn.com/update/ern/ern_random-social-experiments.html. *Clinical Trials* focuses mainly on the medical sector but does include coverage of social and behavioral aspects that influence the interventions being tested in the trials (http://ctj.sagepub.com/). Edited handbooks on social research nowadays typically include chapters on experiments, e.g., *Handbook of Quantitative Criminology*, *Handbook on Practical Program Evaluation*, and *Handbook on Applied Social Research*.

Standards are also necessary for reporting on randomized trials to determine their credibility of findings and the extent to which they were run properly. The CONSORT Statement in health is fundamental (http://www.consort-statement.org/consort-statement/overview0/). The guidelines have been adapted to social fields including education; see The What Works Clearinghouse, http://ies.ed.gov/ncee/wwc/. In the prevention research arena, guidelines have been expanded by Flay et al. (2005) on design and execution of studies of efficacy and effectiveness and worthiness of dissemination, and by Valentine et al. (2011) on replication standards.

Acknowledgments

This work was supported in part by the Institute of Education Sciences, US Department of Education, through Grant #R305B090015 to the University of Pennsylvania. The opinions expressed are those of the authors and do not represent the views of the Institute or the US Department of Education.

See also: Field Experiments; Fieldwork: Ethical Aspects; Psychological Treatments: Randomized Controlled Trials; Systematic Reviewing and Meta-Analysis.

Bibliography

Almeida, R., Brodmann, S., Grun, R., Barouni, M., 2012. Entrepreneurship Training and Self-Employment among University Graduates: Evidence from a Randomized Trial in Tunisia. World Bank Policy Research Working Paper, (6285).

Auvert, B., Taljaard, D., Lagarde, E., Sobngwi-Tambekou, J., Sitta, R., Puren, A., 2005. Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 Trial. PLoS Medicine *2* (11), e298.

Behrman, J.R., Parker, S.W., Todd, P.E., 2009. Schooling impacts of conditional cash transfers on young children: evidence from Mexico. Economic Development and Cultural Change 57 (3), 439.

Bernard, C., 1957/1865. An Introduction to the Study of Experimental Medicine (H.C. Greene, Trans.). Dover Publications, New York.

Bisoffi, Z., Sirima, B.S., Angheben, A., Lodesani, C., Gobbi, F., Tinto, H., Van den Ende, J., 2009. Rapid malaria diagnostic tests vs. clinical management of malaria in rural Burkina Faso: safety and effect on clinical decisions. A randomized trial. Tropical Medicine & International Health 14 (5), 491–498.

Boruch, R.F., 1997. Randomized Experiments for Planning and Evaluation: A Practical Guide. Sage Publications, Thousand Oaks, CA.

Braga, A.A., 2001. The effects of hot spots policing on crime. The Annals of the American Academy of Political and Social Science 578 (1), 104–125.

Bruns, B., Filmer, D., Patrinos, H.A., 2011. Making Schools Work: New Evidence on Accountability Reforms. The World Bank, Washington, DC.

Cole, S., Fernando, A., 2012. The Value of Advice: Evidence from Mobile Phone-Based Agricultural Extension. Harvard Business School Finance Working Paper, (13-047).
Cooper, H., Hedges, L.V., Valentine, J.C., 2009. Handbook of Research Synthesis and Meta-Analysis. Russell Sage Foundation.
Dong, N., Maynard, R.A., 2013. PowerUp!: a tool for calculating minimum detectable effect sizes and sample size requirements for experimental and quasi-experimental designs. Journal of Research on Educational Effectiveness 6 (1), 24–67. http://dx.doi.org/10.1080/19345747.2012.673143.
Donner, A., Klar, N., 2010. Design and Analysis of Cluster Randomized Trials in Health Research. Arnold, London.
Druckman, J.N., Green, D.P., Kuklinski, J.H., Lupia, A. (Eds.), 2011. Cambridge Handbook of Experimental Political Science. Cambridge University Press, Cambridge, UK.
Fantuzzo, J.W., Gadsden, V.L., McDermott, P.A., 2011. An integrated curriculum to improve mathematics, language, and literacy for Head Start children. American Educational Research Journal 48 (3), 763–793.
Farrington, D.P., Loeber, R., Welsh, B.C., 2010. Longitudinal-experimental studies. Chapter 24. In: Piquero, A., Weisburd, D. (Eds.), Handbook of Quantitative Criminology. Springer, New York, pp. 503–518.
Fienberg, S.E., Singer, B., Tanur, J.M., 1985. Large scale social experimentation in the United States. In: Atkinson, A.C., Fienberg, S.E. (Eds.), A Celebration of Statistics: The ISI Centenary Volume. Springer-Verlag, Berlin, pp. 287–326.
Finn, J.D., Achilles, C.M., 1990. Answers and questions about class size: a state wide experiment. American Educational Research Journal 27 (3), 557–577.
Fitz-Gibbon, C.T., 1990. Success and failure in peer tutoring experiments. In: Goodlad, S., Hirst, B. (Eds.), Explorations in Peer Tutoring. Basil Blackwell Ltd, Oxford, UK, pp. 26–57.
Fuchs, D., Fuchs, L.S., Mathes, P.G., Simmons, D.C., 1997. Peer-assisted learning strategies: making classrooms more responsive to diversity. American Educational Research Journal 34 (1), 174–206.
Flay, B., Biglan, A., Boruch, R., Castro, F., Gottfredson, D., Kellam, S., Moscicki, E., Schinke, S., Valentine, J., Ji, P., 2005. Standards of evidence: criteria for efficacy, effectiveness and dissemination. Prevention Science 6 (3), 151–175.
Gillies, R.M., Ashman, A.F., 1998. Behavior and interactions of children in cooperative groups in lower and middle elementary grades. Journal of Educational Psychology 90 (4), 746–757.
Green, D.P., Gerber, A.S., 2003. The underprovision of experiments in political science. The Annals of the American Academy of Political and Social Science 589 (1), 94–112.
Hayes, R.J., Moulton, L.H., 2009. Cluster Randomized Trials. CRC Press/Taylor and Francis Group, New York and London.
Hrobjartsson, A., Gotzsche, P.C., Gluud, C., 1998. The controlled clinical trial turns 100, Fibieger's trial of serum treatment of diphtheria. British Medical Journal 317, 1243–1245.
Jamison, D., Searle, B., Suppes, P., 1980. In: Friend, J. (Ed.), Radio Mathematics in Nicaragua. Research Report. Stanford University Department of Education, Stanford, CA.
Karlan, D.S., Appel, J., 2011. More than Good Intentions: Improving the Ways the World's Poor Borrow, Save, Farm, Learn, and Stay Healthy. Dutton, New York.
Kaye, D.H., Freeman, D.A., 2011. Reference Guide on Statistics. In Federal Judicial Center and National Research Council, Reference Manual on Scientific Evidence, third ed. National Academies Press, Washington, DC. 211–302.
Killias, M., 2006. Improving impact evaluations through randomized experiments: the challenges of the national research council report for European criminology. Journal of Experimental Criminology 2 (3), 375–379.
Killias, M., Aebi, M.F., Ribeaud, D., 2000. Learning through controlled experiments: community service and heroin prescription in Switzerland. Crime & Delinquency 46 (2), 233–251.
Kinder, D.R., 2011. Campbell's ghost. In: Druckman, J.N., et al. (Eds.), Cambridge Handbook of Experimental Political Science. Cambridge University Press, Cambridge, UK, pp. 525–529.
Marks, H.M., 1997. The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900–1990. Cambridge University Press, Cambridge, UK.
Maynard, A., Chalmers, I. (Eds.), 1997. Non-random Reflections on Health Services Research: On the 25th Anniversary of Archie Cochrane's Effectiveness and Efficiency. British Medical Journal (BMJ) Publishing Group, London.
Maxwell, S.E., Delaney, H.D., 2004. Designing Experiments and Analyzing Data: A Model Comparison Perspective. Routledge.
McKay, J.H., McKay, A., Sinnestera, L., Gomez, H., Lloreda, P., 1978. Improving cognitive ability in chronically deprived children. Science 200 (4), 270–278.
Mosteller, F., 1999. How does class size relate to achievement in schools? In: Mayer, S.E., Peterson, P.E. (Eds.), Earning and Learning: How Schools Matter. Brookings Institution Press, Washington, DC, pp. 117–130. Russell Sage Foundation, New York.
Morgan, S.L., Winship, C., 2007. Counterfactuals and Causal Evidence. Cambridge University Press, Cambridge.
Murray, D.M., 1998. Design and Analysis of Group-Randomized Trials. Oxford University Press, New York. Oxford, UK.
Oakley, A., 2000. Experiments in Knowing: Gender and Method in the Social Sciences. Polity Press, Cambridge, UK.
Petrosino, A., Turpin-Petrosino, C., Finckenauer, J.O., 2000. Well-meaning programs can have harmful effects! Lessons from experiments of programs such as Scared Straight. Crime & Delinquency 46 (3), 354–379.
Raudenbush, S.W., Bryk, A.S., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods, vol. 1. Sage Publications, Thousand Oaks, CA.
Raudenbush, S.W., Spybrook, J., Congdon, R., Liu, X., Martinez, A., Bloom, H., Hill, C., 2011. Optimal Design Plus Empirical Evidence (Version 3.0) (Software). Retrieved from: http://www.wtgrantfoundation.org/resources/consultation-service-and-optimal-design.
Rosenbaum, P., 2010. Design of Observational Studies. Springer, New York.
Schweinhart, L., 2003. Retrieved from: http://www.highscope.org/file/Research/PerryProject/Perry-SRCD_2003.pdf.
Solomon, P., Cavanaugh, M., Draine, J., 2009. Randomized Controlled Trials: Design and Implementation for Community-Based Psycho-social Interventions. Oxford, New York.
Stigler, S.M., 1986. The History of Statistics: The Measurement of Uncertainty before 1900. Belnap Press of Harvard University Press, Cambridge, MA.
Tymms, P., Merrell, C., Thurston, A., Andor, J., Topping, K., Miller, D., 2011. Improving attainment across a whole district: school reform through peer tutoring in a randomized controlled trial. School Effectiveness and School Improvement 22 (3), 265–289.
Valentine, J.C., Biglan, A., Boruch, R.F., Castro, F.G., Collins, L.M., Flay, B.R., Kellam, S., Moscicki, E., Schinke, S.P., 2011. Replication in prevention science. Prevention Science 12 (2), 103–117.
Weijer, C., Grimshaw, J.M., Eccles, M.P., McRae, A.D., White, A., Brehaut, J.C., Taljaard, M., Althabe, F., 2012. The Ottawa statement on the ethical design and conduct of cluster randomized trials. PLoS Medicine 9 (11), e1001346.
Weisburd, D., Eck, J.E., 2004. What can police do to reduce crime, disorder, and fear? The Annals of the American Academy of Political and Social Science 593 (1), 42–65.

Experimental Economics

Gigi Foster, University of New South Wales, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by V.L. Smith, volume 8, pp. 5100–5108, © 2001, Elsevier Ltd.

Abstract

This article provides an introduction to the method of laboratory experimentation in economics, an overview of a selection of experimental studies, and a discussion of the findings and frontiers of this area of research. Emphasis is placed on how lessons from the economics laboratory are related to and can complement ideas and theory in other scientific subfields.

Introduction

Experiments overseen by economists have multiplied exponentially over the past 30 years. Experiments were once the province of the 'hard sciences' and psychology, whereas they now enjoy a respected and stable place in the economist's methodological toolbox. This article summarizes the defining features and scientific contributions of the modern subfield of experimental economics, describes the relation of economic experimentation to other methodological approaches in economics and to related research streams in social and behavioral sciences, and discusses the potential of economic experiments for progressing frontiers in both scientific knowledge and policy.

The Potential of Experimentation in Economics

Why do scientists perform experiments? Ostensibly, they want to test theories in progress in a fashion that allows them some control over the setting in which that test occurs. A well-designed experiment is one whose results allow scientists to reject, overwrite, or refine their operating understanding of the key quantities and governing dynamics at work in generating observed phenomena. By its very nature hence, experimentation is inextricably bound to theory. In the social sciences, 'theory' means what Croson and Gächter (2010) describe as "a description of a social situation, which involves specifying the actors, the choices they face, their information, and how they evaluate each possible outcome" (p. 123). A good theory in social science is a stylized representation (model) of 'situations in general' that can be applied to many complex real-life situations to deliver reasonably accurate predictions of the behavior of humans in those situations. To test whether a working theory is consistent with observed behavior, testable implications can be derived from that theory and built into an experimental design. To isolate the impact of one particular aspect of an accepted theory, the experimental design aims to hold 'everything else' constant, allowing the scientist to record outcomes in response to variations in only the level or pitch of the focal aspect. To horse race the usefulness of two or more working theories, competing implications can be drawn from each and a design sought that enables a direct comparison of the relative strength of support for the implications of each model. These are the sorts of agendas pursued through the various experiment-based methods in economics, which include the design and implementation of field experiments, the ex-post exploitation of 'natural experiments' created by policy quirks or acts of nature, and the design and implementation of laboratory experiments, which are the focus of this article.

While the discipline of economics has arguably delivered the most powerful standard model of human behavior, which is used today in some form or another to guide governmental policies and corporate decisions around the world (on the macroeconomic level, the mainstream model is well-represented by the writings of some of its founders (e.g., Paul Samuelson), and is reflected in what is taught to students of the discipline; on the microeconomic level, the main thrust of mainstream economic argument is simple: individuals respond in predictable ways to incentives (including, but not limited to, those embodied in market signals), no modern social science offers an internally consistent theory that fully explains the extreme complexity of human behavior. Hence, experimentalists across the social sciences – including in economics – perenially find that their initial theories in progress are inadequate to embrace all their observations. This leads to a natural process of refining the theory, running more tests, doing more refining, running more tests, and so on, in the continual pursuit of a deeper and more perfect understanding of behavior. These processes pursued simultaneously by many researchers can ultimately lead to a fundamental restructuring of basic theory, which holds the potential for true scientific advance. Until this restructuring occurs, scientific discourse commonly revolves around particular previously unexpected yet replicated phenomena that then provide an engine for generating hypotheses that only indirectly connect with those initial theories that motivated the original experiments. Modern experimental economics, with its incredible variety of applications and its range of obliquely connected results, can be argued to be in such a stage today (see Smith, 2010; Roth, 2014 for extended and thoughtful discussions of this phenomenon). Future breakthroughs in core economic theory, the backbone of the economics discipline, may yet be driven in part by the contributions of experimental economists.

Disciplinary Context and Constituents of Economics Experiments

The term 'experimental economics' generally refers to the subset of explorations by economists that is informed by

laboratory experiments. Yet economists today are also very active in the design and implementation of field experiments and in the capitalization, sometimes in collaboration with government administrators or policy makers, upon events in the real world that simulate a controlled experiment – often referred to as a 'natural experiment.' Field experiments, often aimed at testing the effects of particular institutions or policy interventions on individual behavior or aggregate social outcomes, are used frequently by researchers in institutional economics (e.g., Olken, 2010), public economics (e.g., Lacetera et al., 2013), and development economics (see Duflo, 2006 for a review). Natural experiments, in which subjects do not even know that they are participating, have been sought and exploited by those working in fields such as labor economics (e.g., Card and Krueger, 1994; Siminski, 2013) and health economics (e.g., Chou et al., 2010). Some of the economics discipline's most well-known work, including much of what is presented in the blockbuster popular science book *Freakonomics* (Levitt and Dubner, 2005), is the product of research that has exploited field or natural experiments.

By contrast, laboratory experiments designed and run by economists typically involve creating stylized representations of reality that are then presented to real people (experimental participants) in a controlled laboratory setting. Experimenters typically want to determine whether, and under what conditions, models of individual decision making, strategic interaction among agents, and/or market dynamics 'work' in the sense that their implications are consistent with the observed behavior of experimental participants. To this end, experimenters design a stylized world that specifies with greater or lesser precision the institutions that constrain and support decision making and choices (e.g., rules about what types of decisions are feasible, and information technology, such as whether participants know what other participants own or are doing, or whether it is costly to acquire this information), the available resources, and the 'payoffs' that accrue to each participant in each possible state of the world that might arise in the experiment. These payoffs are typically denominated in experimental or real-world currency, made known to participants, and are paid out to participants in real-world currency at the conclusion of the experiment according to the realized state(s) of the world (naturally, experimental subjects may be motivated by factors other than their monetary payoffs in the experiment, and often this is implicitly offered as a reason why subjects' behavior is observed to deviate from what might be expected based on a rational calculus of monetary profit maximization). Some experiments include uncertainty about more than merely what other subjects will choose – for example, the process by which an outcome is generated may be affected by random chance, or participants may face a choice of investing in a risky lottery.

The Theoretical Backdrop of Experiments in Economics

The design of economic experiments is frequently informed by game theory, decision theory, theories of industrial organization, and/or nascent ideas in the relatively young subfield of behavioral economics. These overlapping subfields, together with fundamental tenets derived from standard economic theory (e.g., that people respond to incentives), either singly or in combination provide the lens through which most experimental results in economics are interpreted by practitioners. The relevance of these subfields to the design of economic experiments is briefly reviewed below.

Game theory is the study of models of multiagent social situations in which individuals face choices whose effects depend more or less directly upon the choices of particular others. Game theory expands the scope of mainstream economic theory, because the latter's presumption of large, well-functioning markets rejects a personal social feedback element: in the standard model, no one individual's actions should be able to affect the market signals (chiefly, prices) that others use as inputs to their decisions, meaning that no one individual should rationally be swayed toward or away from any particular course of action merely because of the action of another individual. In game theory, by contrast, focal players are finite in number (usually numbering in the single digits) and can typically affect, though not unilaterally determine, the outcomes for everyone. The stylized models built of players' interactions in such models, which include settings for institutions, information availability and precision, resource endowments, and individual objectives, are often mathematically complex and challenging to solve. Game theory has been seminal in motivating some of the original games played in the laboratory, including most famously the Prisoner's Dilemma game and other instances of 'social dilemmas' (described more fully below), and has also motivated experiments designed to observe behavior in auctions and bargaining settings. However, the implications of many modern game theoretic models are unhelpful for the design of economic experiments. This happens most glaringly when it is found that in most possible specifications of a model, the possible equilibrium outcomes (i.e., the set of constellations of choices from which no individual has an incentive to deviate) are infinite. Additionally, economists' theories of individual decision making that game theory models try to animate do not yet feature a widely accepted, simple, and satisfying (as opposed to ad hoc) incorporation of motivations other than pure monetary self-interest, despite the fact that the existence of such motivations is often deduced from experimental results. These are the two primary reasons why modern game theory models are not more widely used as scaffolds for the design of economic experiments today.

Industrial organization theory, a branch of applied game theory, has delivered predictions about behavior in particular concrete interactive settings, from simple dyadic games like principal agent problems up to and including entire markets. By focusing on the behavior of profit-maximizing firms and/or their representatives, partners, employees, or competitors in particular well-defined and constrained situations – an area where mainstream economic theory has proven its predictive validity time and again – industrial organization (I/O) theorists have largely been able to circumvent the morass of accurately capturing those elements of individual choice that may be driven by phenomena other than pure monetary self-interest. I/O theory has generated clear and robust predictions about behavior under contrasting market structures (e.g., monopoly vs oligopoly), information structures (e.g., asymmetric or perfect),

and transaction cost levels, which can be relatively easily incorporated into experimental designs and manipulated by the experimenter. The relative clarity, tractability, and specificity of the implications of these models have meant that they underpinned the design of some of the first economic experiments (see Plott, 1982 for a review) and are still frequently used today.

Decision theorists examine many dimensions of how individuals decide how to allocate their resources (e.g., money or effort) or how to set their beliefs (i.e., their mental models of the world). It is distinct from game theory inasmuch as the social element of choice is frequently included only obliquely: the main focus is often not on calculating or interpreting the final result of the many interdependent choices of an entire interacting band of decision makers, but rather on specifying the decision process of a single agent or a set of agents, operating independently. Models of individual decision making contributed by economists include most famously those that specify the internal mental models and/or preferences of individual agents (such as subjective expected utility theory) and also those that specify particular mental idiosyncrasies (such as a reliance on mental heuristics) that may taint choice. These aspects of the underpinnings of choice may guide experimental design and the interpretation of results, but are arguably impossible to manipulate in a laboratory setting, since experimental subjects come in the laboratory door with them already in place.

'Behavioral economics' is the catch-all term for studies of the behavior of individual decision makers that appears, on the face of it, to be at odds with the behavior that would be predicted by mainstream economic models in which individuals act rationally with the aim of maximizing their own material gains. The reasons behavioral economists have proposed for why observed behavior deviates from these mainstream predictions include elements such as emotion, bounded rationality, internal relations with the unseen, systematically biased information processing, and the satisfaction of psychological needs, none of which are included in mainstream economic models. The developing theories of institutional economists also frequently contain some of these elements and have also given rise to laboratory experiments (e.g., Dal Bó et al., 2010). However, as with decision theory, the direct experimental manipulation of these dimensions of human motivation and behavior is arguably impossible or at best highly challenging, meaning that experiments can deliver more suggestive evidence than conclusive proof that these dimensions exist and are important.

Games That Are Played

Within the modern experimental economics community, particular features of design and protocol have endogenously emerged as strong norms to which one is expected to adhere in order for one's research to be considered acceptable (similar guidelines for 'best practice' arguably also exist for field experiments conducted by economists (see List, 2011)). These norms include the following.

No Deception

How much information do experimental subjects have about what is happening in the experiment? At one end of the continuum, subjects might be lied to outright about what the experiment is about, the level or determinants of phenomena that theory predicts should influence their behavior, and so on. At the other end, subjects might be told everything, from the minutest detail of the mathematical process being used to deliver an observation on a particular phenomenon, up to and including the precise hypothesis that the experimenter aims to test. The norm in the field at present is somewhere between these two extremes: researchers are expected to describe to participants in reasonable detail what will actually be happening in the experiment, and not to lie to them outright; but it is neither considered necessary to tell participants every detail of every process nor is it always considered either necessary or desirable to reveal the purpose of the experiment, since in many cases one might argue that doing so would irrecoverably change subjects' behavior away from what they would otherwise be inclined to do. This norm is particularly striking in its opposition to norms in experimental psychology, where outright lying to subjects followed by ex-post debriefing, while not universally accepted, are far more common (see Korn, 1997 for a review).

Careful of 'Endowment Effects' and 'Framing Effects'

Experimenters have found repeatedly that subjects appear to attach more value to objects that they perceive as theirs already, than to objects that they do not (yet) own. This often-replicated phenomenon, underpinned by Kahneman and Tversky (1979) and more concisely and generically described as reflecting 'reference-dependent preferences' or 'loss aversion,' has given rise to a concern that no behavioral result found in the laboratory be interpreted as something other than reference dependence when it could be explained by that alone. Merely using language or signals that frame a resource as already in a subject's possession has been found to alter subjects' behavior. Many other types of 'framing effects' – changes in behavior as a result of using particular language or signaling, rather than as a result of changing the parameters of the material problem being faced by subjects – have been found by experimental economists and are hence another focal area of concern in interpreting results.

Careful of 'Learning Effects'

Many economic experiments involve facing a particular decision-making scenario repeatedly. The standard concern is to ensure that observed behavior is interpreted only after partialling out any changes that may have resulted from subjects' learning, over the course of repeating the game many times, how the game really works or how to perform better. This is a particular problem when the task is complex and/or when there is doubt about whether the subjects understood the instructions, meaning that extended delivery of clear instructions with plenty of time for subjects to ask questions and/or be tested on their understanding of the instructions are hallmarks of modern economic experiments.

Incentivize Focal Choices

As mentioned above, a central tenet of economics is that people respond to incentives, or, in other words, that the absence of incentives will often produce different behavior than what

would result when incentives are in place. For this reason, and because they wish to be able to use their results in thinking about real-life settings where incentives are in play, experimental economists typically associate monetary incentives for subjects with whatever decisions of subjects they aim to interpret. Again in contrast to many experiments in our sister social sciences, subjects in economics experiments do not merely consider what they might do in a hypothetical situation with no real monetary stake in the outcome; they act out stylized representations of real situations, with real money on the line.

Experimenters in economics also pay particular attention to the randomization and anonymity requirements of whatever experiment is being run. The concerns here involve ensuring that behavior generated in the laboratory is not tainted by preexisting relationships among subjects, tacit expectations between subjects and the experimenter ('experimenter effects'), or nonrepresentativeness of the laboratory sample. On this last point, regardless of how close the experimenter gets to mimicking reality in the laboratory, any experimental result is inevitably vulnerable to the critique of low 'external validity' – in other words, poor confidence that insights based on the results can be used to interpret situations outside the laboratory. This concern in greater in situations where subject descriptors like age, experience, or culture, whose distributions are often truncated in subject pools, are expected to play an important role in determining behavior or behavior-relevant beliefs. Partly for this reason, and also because any number of other particular details of an experiment may be unique, replication – as in any experimental science – is considered the gold standard for generating support for the robustness of results.

Variations on a suite of workhorse game types are frequently presented to participants in economic experiments. These games are often programmed in specialized software, of which two popular examples are zTree (Zurich Toolbox for Ready-made Economic Experiments) and Mouselab. A brief sketch of the most popular workhorse games and some of the results they typically generate are given below.

Individual Decision-Making Games

These games feature no interaction between subjects, and focus on exploration of various aspects of how people make decisions. A common example is investment games, which offer subjects a risky investment opportunity and study how they behave under varying settings of risk, return, market signals, uncertainty, and so on (for example, as discussed in Harrison and Rutström, 2009).

Trust Games

Ultimatum bargaining games (popularized by Güth et al., 1982) provide a resource endowment to a subject who is offered the opportunity to give a portion of it, of his choosing, to a partner. The partner who is offered a share of the endowment is then allowed to either accept the offer (in which case it is 'paid out,' meaning that the division of the endowment proposed in the offer is implemented for both participants), or reject the offer, in which case both participants receive nothing. Economic theory suggests that the recipient should accept any positive offer made, since something is better than nothing, and therefore that the offerer should offer a tiny amount and expect this offer to be accepted by the recipient. The standard finding in the laboratory is that while significant heterogeneity across subjects exists (see Oosterbeek et al., 2004), on average a substantial share but less than half the initial endowment is offered and is accepted by the recipient. Dictator games, an outgrowth of ultimatum bargaining games, begin like the ultimatum game but the partner has no option but to accept the dictator's choice. Economic theory as animated in noncooperative game theory suggests that persons with the opportunity to be dictators should simply keep the entire endowment for themselves. The standard finding in the laboratory is that, by contrast, most people in the role of dictator share some of their resources with their partner. See Engel (2011) for a review.

Coordination and Cooperation Games

These include variants of the Prisoner's Dilemma (depicted in Figure 1), and public goods games where participants are given an initial endowment and the choice to invest part of it in a social good from which everyone in the game will benefit. Where theory predicts that the socially best outcome is not an equilibrium (and hence will not be reached by laboratory participants), such as in the case of a one-shot Prisoner's Dilemma, the term 'social dilemma' is applied. The standard result in Prisoner's Dilemma games, whether one-shot or repeated, is that people do not always play 'confess.' This has sparked much discussion of why that is, and in what situations denial is, more or less likely (e.g., Andreoni and Miller, 1993).

Market Trading Games

These games aim to mimic an actual market, such as the market for pollution permits or the market in which workers are matched to employers, often in order to validate theories about how such markets should be designed so as to maximize social welfare. See Roth (2002) for a discussion.

Auction Games

These games implement auctions for objects of value, often in order to validate theories about how to design auctions so as to maximize social welfare. See Kagel and Levin (2008) for a review of work in this area.

Years of prison to serve: (Player 1, Player 2)		Player 2: Confess	Player 2: Deny
Player 1	Confess	(5, 5)	(0, 6)
	Deny	(6, 0)	(2, 2)

Figure 1 The canonical Prisoner's Dilemma.

Economists have also used many other types of customized games in their attempts to mimic particular real-life situations of interest. See Davis and Holt (1993) for a more extended discussion of the variety of game types that have been used, their theoretical underpinnings, and their analytical targets.

Crossing Disciplinary Boundaries in Economics Experiments

As noted above, mainstream economic models have been claimed to be inadequate to explain some of the behavior that economists and other social scientists have observed in the laboratory (see Erev and Greiner, 2014 for a fuller discussion). Any latent phenomena that do not appear in such models are potentially present in any experiment, as a result of the prior conditioning of subjects, and despite their best efforts, experimentalists cannot fully control these phenomena, if indeed they exist. Also, subject heterogeneity, the effects of which are controllable to an extent via randomization but that, nonetheless, has been found in many studies, is not the main focus of mainstream economic theory featuring agents that are either identical or differ only along particular proscribed dimensions.

For these reasons, some economists have reached outside-of-their-home discipline to look for ideas and explanatory frameworks with which to guide and interpret their work. In particular, economic theory is weak in capturing the full set of factors (including nonmaterial factors) that may influence a decision maker's behavior, and it is weak in capturing group-related phenomena, with industrial organization theory being a partial exception inasmuch as a firm is a group. Unsurprisingly then, the disciplines whose contributions are most heavily drawn upon by experimental economists to help interpret their results and even in some cases to help design their studies are psychology, sociology, and political science.

Emotion and Cognition

Modern experimental economists are keenly aware that their subjects may have mental models, habits, or limitations that influence their beliefs and behavior in the laboratory. A few aspects of these psychological and neurological phenomena that have drawn particular attention in the experimental community are mentioned briefly below.

Fairness, Altruism, or Inequality Aversion

People's observed behavior in trust games, reviewed above, and in other games in the laboratory has been suggested to be the result of inbuilt commitment to some notion of generosity or sense of fairness. Ideas and jargon formerly the province of psychology, such as 'empathy' and 'altruism,' have thus made their way into the frameworks and parlance of economic experimentalists (e.g., Andreoni and Rao, 2011).

Cognitive Load and Heuristics

It has been vociferously argued (e.g., Tversky and Kahneman, 1974; Gigerenzer et al., 1999; among many others) that due to limitations and tendencies of the human brain, information presented in laboratory experiments – and indeed in the real world – is processed in a biased or incomplete manner, a contention of psychology for generations. This comes with clear implications for framing information (discussed above) and the wisdom of running experiments in which the instructions are complex.

Functional magnetic resonance imaging studies have on occasion been drawn on to motivate or interpret experimental results run by economists (e.g., Delgado et al., 2008). The brain remains today a highly mysterious organ, and it would appear that significant scientific advances are required before neurological evidence can lead to fundamental advances in those strands of behavioral theory that claim particular mental or emotional traits to be important in explaining behavior.

Group Dynamics

As evidenced by the recent proliferation of work in economic subfields that touch upon humans' interactions with and within their social groups (e.g., institutional economics, cultural economics, the economics of identity), economists increasingly recognize the role of group context and dynamics for individual behavior. It is particularly difficult to apply standard economic theory to this area of investigation, because of its omission of group-level phenomena from anything but the rather cold and often anonymous realm of buyers and sellers. How might people's behavior respond to incentives arising from social norms within companies, families, or countries? How do these incentives arise in the first place, how can we measure them, and what keeps them binding? These are questions for which coherent answers require either an expansion of core economic theory (a recent example of an attempt to offer such an expansion is Frijters (2013)) or a borrowing from sister disciplines that have asked such questions for decades, like sociology and political science.

Studying the nature of the interplay between individuals and their real-world groups is difficult in the laboratory, since subjects come into the laboratory with their group-related identification, norms, and habits already set. However, some economic experiments have tried to create new groups in the laboratory that mimic the rules and context of real-world groups. A particularly popular approach is to set up a leadership scenario with the potential for redistribution or exploitation by the controlling elite of common-pool resources, complete with voting and/or resource control rules, intended to mimic modern democratic groups such as the nation-state (e.g., Cabrales et al., 2012). This type of setting has been used to try to isolate the contributions of institutions such as rotating leadership, anonymous choice, socially mediated punishment, and power inequality to individuals' behavior (including exploitative and prosocial behavior) and aggregate welfare outcomes.

The Usefulness of Economic Experiments in Progressing Policy and Science

Laboratory experimentation by economists has delivered important evidence both in favor of, and against, the predictive validity of standard economic theory. In the former arena lie

many results regarding the optimal design of auctions and markets, and these results have been used to design actual markets in the real world, with powerful real welfare gains. Prime examples of this are Al Roth's kidney exchange program and the design of spectrum auctions.

While it has generated a proliferation of results, where the experimental method is weaker at present is in providing unambiguous guidance about phenomena that could help to refine our scientific theories of human decision making and social behavior. Debates presently rage about whether laboratory results that have been interpreted by some as evidence of altruism, bounded rationality, group norms, or other phenomena that deviate from the mainstream economic view of the economic agent in fact require an expansion of core theory, or are instead the result of particular experimental settings and do not extend to real-world decision making. Arguably, even if such phenomena do influence human behavior, identifying and measuring their causes (such as loyalty to one's nation state) using experimental methods will surely prove challenging. Nonetheless, experimental economics today is a thriving subfield and should be expected to keep delivering useful, intriguing, and controversial findings for years to come that will gradually push the discipline toward progress.

See also: Auctions; Behavioral Economics; Cooperative Game Theory; Expectations, Economics of; Game Theory: Noncooperative Games; Information, Economics of.

Bibliography

Andreoni, James, Miller, John H., 1993. Rational cooperation in the finitely repeated Prisoner's Dilemma: experimental evidence. The Economic Journal 103, 570–585.

Andreoni, James, Rao, Justin M., 2011. The power of asking: how communication affects selfishness, empathy, and altruism. Journal of Public Economics 95 (7–8), 513–520.

Cabrales, Antonio, Nagel, Rosemarie, Rodríguez Mora, José V., 2012. It is Hobbes, not Rousseau: an experiment on voting and redistribution. Experimental Economics 15 (2), 278–308.

Card, David, Krueger, Alan, 1994. Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. American Economic Review 84 (4), 772–793.

Chou, Shin-Yi, Liu, Jin-Tan, Grossman, Michael, Joyce, Ted, 2010. Parental education and child health: evidence from a natural experiment in Taiwan. American Economic Journal: Applied Economics 2 (1), 33–61.

Croson, Rachel, Gächter, Simon, 2010. The science of experimental economics. Journal of Economic Behavior and Organization 73 (1), 122–131.

Davis, Douglas D., Holt, Charles A., 1993. Experimental Economics. Princeton University Press, New Jersey.

Dal Bó, Pedro, Foster, Andrew, Putterman, Louis, 2010. Institutions and behavior: experimental evidence on the effects of democracy. American Economic Review 100, 2205–2229.

Delgado, Mauricio, Schotter, Andrew, Ozbay, Erkut Y., Phelps, Elizabeth A., 2008. Understanding overbidding: using the neural circuitry of reward to design economic auctions. Science 321 (5897), 1849–1852.

Duflo, Esther, 2006. Field experiments in development economics. In: Blundell, Richard, Newey, Whitney, Persson, Torsten (Eds.), Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, vol. 2, Book 42. Cambridge University Press, Cambridge, pp. 322–348.

Engel, Christoph, 2011. Dictator games: a meta study. Experimental Economics 14 (4), 583–610.

Erev, Ido, Greiner, Ben, 2014. The 1-800 critique, counter-examples, and the future of behavioral economics. In: Frechette, G., Schotter, A. (Eds.), The Methods of Modern Experimental Economics. Oxford University Press.

Frijters, Paul, Foster, Gigi, 2013. An Economic Theory of Greed, Love, Groups, and Networks. Cambridge University Press, Cambridge.

Gigerenzer, Gerd, Todd, Peter M., The ABC Research Group, 1999. Simple Heuristics That Make Us Smart. Oxford University Press, Cambridge.

Güth, Werner, Schmittberger, Rolf, Schwarze, Bernd, 1982. An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization 3, 367–388.

Harrison, Glenn W., Elisabet Rutström, E., 2009. Expected utility theory and prospect theory: one wedding and a decent funeral. Experimental Economics 12, 133–158.

Kagel, John H., Levin, Dan, 2008. Auctions: experiments. In: Durlauf, Steven N., Blume, Lawrence E. (Eds.), The New Palgrave Dictionary of Economics, second ed.

Kagel, John H., Roth, Alvin E., 1995. Handbook of Experimental Economics. Princeton University Press, Princeton, NJ.

Kahneman, Daniel, Tversky, Amos, 1979. Prospect theory: an analysis of decision under risk. Econometrica 47 (2), 263–291.

Korn, James H., 1997. Illusions of Reality: A History of Deception in Social Psychology. SUNY Press, Albany.

Lacetera, Nicola, Macis, Mario, Slonim, Robert, 2013. Economic rewards to motivate blood donations. Science 340 (6135), 927–928.

Levitt, Steven D., Dubner, Stephen J., 2005. Freakonomics: A Rogue Economist Explores the Hidden Side of Everything. William Morrow/HarperCollins, New York.

List, John A., 2011. Why economists should conduct field experiments and 14 tips for pulling one off. Journal of Economic Perspectives 25 (3), 3–16.

Olken, Benjamin A., 2010. Direct democracy and local public goods: evidence from a field experiment in Indonesia. American Political Science Review 104 (2), 243–267.

Oosterbeek, Hessel, Sloof, Randolph, Kuilen, Gus van, 2004. Cultural differences in ultimatum game experiments: evidence from a meta-analysis. Experimental Economics 7, 171–188.

Plott, Charles R., 1982. Industrial organization theory and experimental economics. Journal of Economic Literature 20, 1485–1527.

Roth, Alvin E., 2002. The economist as engineer: game theory, experimentation, and computation as tools for design economics. Econometrica 70 (4), 1341–1378.

Roth, Alvin E., 2014. Is experimental economics living up to its promise? In: Frechette, Guillaume R., Schotter, Andrew (Eds.), The Methods of Modern Experimental Economics. Oxford University Press.

Siminski, Peter, 2013. Employment effects of army service and veterans' compensation: evidence from the Australian Vietnam-era conscription lotteries. Review of Economics and Statistics 95 (1), 87–97.

Smith, Vernon L., 2010. Theory and experiment: what are the questions? Journal of Economic Behavior and Organization 73 (1), 3–15.

Tversky, Amos, Kahneman, Daniel, 1974. Judgement under uncertainty: heuristics and biases. Science, New Series 185 (4157), 1124–1131.

Relevant Websites

http://www.mouselabweb.org/ – MouselabWEB.
http://www.orsee.org/ – Online Recruitment System for Economic Experiments.
http://veconlab.econ.virginia.edu – Virginia EconLab.
http://www.iew.uzh.ch/ztree – Zurich Toolbox for Readymade Economic Experiments.

Experimental Laboratories: Psychological

Daniel Greenberg, College of Charleston, Charleston, SC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Experimental laboratories in psychology are a bit different from the typical hard-science lab; in fact, to the casual visitor, they look a lot like ordinary office space. Nevertheless, these labs generally need certain features to be viable, including dedicated research space, reliable access to participants, and equipment for participant testing. The precise needs vary from investigator to investigator, and this article does not attempt to cover them all; rather, it seeks to provide a sense of the key issues that arise when investigators, department chairs, and administrators set out to build a new psychology lab.

Laboratory-Based Research on Human Participants in Psychology

Research in psychology can range from surgical experiments on nonhuman animals to observational research on social groups. This article will focus on noninvasive research on human beings – research that includes controlled experiments, pseudoexperiments, natural experiments, and observational research. (The lab needs of animal researchers and neuroimaging researchers are substantially different and will be covered in a separate article.) For instance, a psychology department's roster might include the following:

- A cognitive psychologist studying the heuristics people use to make decisions.
- A neuropsychologist investigating the effects of brain damage on learning and memory.
- A physiological psychologist measuring the galvanic skin response (GSR) elicited by emotionally evocative stimuli.
- A developmental psychologist conducting observational research on family dynamics.
- A social psychologist carrying out survey research examining health behaviors in college students.

This list is by no means comprehensive; it merely serves to provide a sense of the kinds of research that are done in psychology, and thus the kinds of facilities that might be needed.

The Fundamentals

Lab Location

Whatever the lab's focus might be, it must have a space that is consistently quiet. A loud sound at the wrong time can ruin a painstaking afternoon of data collection in an instant, and a lab situated in a noisy environment is useless. A modicum of soundproofing can be helpful, but it is particularly important to pay careful attention to what is located above, below, and behind the proposed laboratory space. Furthermore, potential participants must find it easy to access the space. For undergraduates, this is rarely a problem; however, other participants may require additional resources, such as convenient parking, an elevator, a waiting room, or a nearby restroom. These points may seem obvious, but they are critical: Participants will not come unless they find the experience agreeable, and a lab without participants cannot function.

Lab Design

Most psychology labs have a relatively straightforward design. The author's lab at the College of Charleston provides an example of the sort of space that is suitable for a midsize regional school that focuses on undergraduate education (a lab at a major research university would likely require more space). The lab is separated from the remainder of the building by a long hallway, and there is a small waiting room that is shared with another investigator. The lab itself consists of a main room with four small testing rooms attached, two on each side; the main room is roughly 18 m^2 (200 sq ft) whereas the testing rooms range between 5 and 9 m^2 (60–100 sq ft). The rooms have typical institutional furniture, including chairs reserved for participant use, and a number of desktop or laptop computers.

This lab is probably the simplest possible model; it focuses on paper- and computer-based cognitive testing of older and younger adults, so no elaborate equipment is required. Naturally, investigators with different research programs will have different needs. In developmental and observational research, for instance, a wall between the two testing rooms might contain a one-way mirror, which would allow investigators and research assistants to observe participants' interactions. In addition, cameras can be set up to record participants' activities, and microphones can be run through a drop ceiling. Researchers studying spatial navigation or driving might need a room dedicated to a virtual–reality setup. Other researchers rely on a battery of standardized psychological tests such as the Wechsler Adult Intelligence Scale (WAIS) or the Woodcock–Johnson battery. Still others might need specialized equipment to measure physiological variables such as blood pressure, pulse, or GSR (a change in skin conductivity indicative of physiological arousal). The requirements are limited only by the imagination of the investigator and (more relevantly) the resources of the institution.

Recruiting Participants

Of course, a well-built and well-stocked lab is useless without participants, so participant recruitment is one of the key concerns of any researcher in experimental psychology. Researchers need to be able to test enough participants to ensure adequate statistical power provide sufficient external

 http://dx.doi.org/10.1016/B978-0-08-097086-8.41028-7

validity, and allow for testing of both control and experimental groups. Consequently, researchers have developed a number of strategies for participant recruitment.

Undergraduate Participant Pools

Many studies in experimental psychology are conducted on undergraduates enrolled in introductory psychology classes (also known as the 'subject pool' or 'participant pool'). Investigators rely on participant pools because they are readily available and inexpensive to test because these participants are typically compensated by receiving credit toward their grade in their introductory psychology class; moreover, the participants are accustomed to taking tests and answering questions, and will typically perform with adequate diligence and effort. According to standard ethical principles, they must be given an alternative to participation in research, such as writing short research papers or summaries of journal articles.

At larger institutions, managing the participant pool becomes a logistical challenge, which is why most institutions rely on web-based systems. Experimenters who want to use undergraduate participants are given accounts on the system, allowing them to post experiment times, descriptions, durations, and locations. They can set a cutoff time by which participants must sign up to participate (for instance, an available time slot can be set to disappear if it is not filled 24 h in advance), making it easy for experimenters to plan out their day. The system can also be set up to send reminder e-mails, or even reminder text messages, at a specified time. Investigators can also set up prescreening surveys or indeed administer entire questionnaires online. The system also makes it easy to grant or deny credit, track how many students have participated, and so forth. Students in introductory classes are given accounts as well, and thus they can easily enroll in experiments that fit their schedule, confirm or cancel appointments, contact experimenters, and track how much credit they have earned. At the end of the semester, system administrators and instructors can run reports on the entire subject pool, allowing them to award the correct amount of credit. While no system is perfect (and none can entirely prevent a certain amount of squabbling over no-shows, cancellations, appropriate amounts of credit, etc.), these web-based systems are generally a worthwhile investment.

Concerns About Undergraduate Participant Pools

In most cases, the undergraduate participant pool is supposed to be educational – that is, the students are supposed to learn something about psychology from the experience. Although students and researchers alike are sometimes skeptical of this notion, existing studies of this issue generally paint a more optimistic picture. Landrum and Chastain (1995) used a short questionnaire to assess the views of a small percentage of participants, and discovered that most of them found the experience to be educational. Bowman and Waite (2003) found that students who participated in research were more satisfied than those who pursued the alternatives; they also found that students had a better understanding of research procedures and a more positive attitude toward psychology. Similarly, Elliott et al. (2010) found that students who participated in an experiment learned just as much as students who simply listened to a lecture or read an article about it, although the students generally found participation more enjoyable. Thus, on a pedagogical level, these experiences do seem to serve their stated purpose, at least when proper debriefing techniques are used.

On an experimental level, however, the reliance on undergraduates raises concerns about validity, in part because of continued skepticism about how representative these participants really are. As Henrich et al. (2010) have observed, undergraduates are generally 'WEIRD' – Western, Educated, Industrialized, Rich, and Democratic – and thus they are not a representative sample of humanity. In fact, as the university admissions office will be happy to point out, these psychology students are often not a representative sample of their age group either. They are not a representative sample of the students at their own university, where the majority of students may be enrolled in other subjects. They might not even be a representative sample of students in introductory psychology classes: a reasonable number of students choose the research alternative for a variety of reasons (see Elicker et al., 2010), and by definition nonparticipants are difficult to study. One investigation, however, has found that students who do participate are better students than those who do not (Padilla-Walker et al., 2005), suggesting that participants might differ from nonparticipants on tasks that draw on academic aptitude.

To make matters worse, the characteristics of participants are known to vary across the semester. Participants who show up earlier in the semester tend to be more conscientious and are more likely to be female (e.g., Witt et al., 2011). They may be more academically oriented, more compliant, and more introverted, and there is some evidence that they tend to have higher GPAs (see Aviv et al., 2002, for a review). Results for cognitive tasks have been mixed (see Zelenski et al., 2003, for a review). Nevertheless, the message is clear: A study can yield different results at different times in the semester.

Beyond the Participant Pool

These are longstanding concerns (Sieber and Saks, 1989; Sears, 1986, 2008), and they have led many researchers to seek out participants outside of the usual participant pool. In any event, undergraduates are simply not appropriate participants for certain lines of research, such as research on aging or early child development. The primary challenge in this approach involves recruitment. Naturally, participants will not magically show up at the lab door; they need to be recruited through advertisements (whether on paper or the Internet), flyers, and the like. In this way, researchers can build up a database of reliable participants. The word 'reliable' is important here, because investigators must avoid 'frequent fliers' – people who make at least part of their living by participating in research, and who are perfectly willing to be less than honest about whether they meet study criteria. A well-established and reliable database allows investigators to select participants who meet certain demographic criteria. It is important to note, however, that even a database like this is not static; it needs to be replenished on a regular basis as participants move out of the area, lose interest in participation, or pass away. The participants who do remain in the database will change with time and might one day fail to meet study criteria. Thus, recruitment is always ongoing – and so are the costs associated with it.

Participant Payments

These costs arise in part because participants outside the undergraduate pool typically need to be paid. For nonintrusive psychological research, the compensation is nominal – in 2014 US dollars, a participant might earn $10–15 an hour for participation in a simple behavioral study along with free parking. Nevertheless, participation rates drop substantially if compensation is not provided. Experience has shown that administrators and investigators should look into the method of paying participants *early* in the process of setting up the lab, because university financial offices are often unfamiliar with this practice and have been known to raise numerous objections. Here are some questions to ask: Can participants be paid in cash? Can they be given gift cards, gift certificates, or other items? Will they have to be paid by check, electronic transfer, or other similar mechanism? How long will it take to issue payment? Will participants have to disclose their address, social security number, tax ID number, or other personal information? Are there limits to how much participants can be paid in a single year? Who is responsible for tracking all of these? What sort of paperwork is required, what sort of records must be kept, and for how long? As a general rule, the more cumbersome and intrusive the paperwork, the less willing people will be to participate.

Hardware, Software, and Other Equipment

In the early days of psychology, researchers tested participants using all manner of clever mechanical equipment. These days, most testing can be done using fairly standard office computers equipped with specialized software. The brand names and the capabilities of the software packages change on a regular basis, but they all have one key feature in common: They provide experimenters with precise control (that is, on the order of tens of milliseconds) over the presentation of stimuli and the collection of responses. Ordinary office software cannot do this.

Most of these software packages have extensive capabilities that allow researchers to program almost any conceivable experiment. They can present sound clips, movies, images, or text. They can record responses using a keyboard, mouse, button box, joystick, and so on. They can even create 'branching' experiments in which the order of experimental trials depends on the person's responses – for instance, the experiment can be programmed to slow down automatically if the participants are making a large number of errors. Once the experiment has been run, the data can easily be exported to a variety of standard spreadsheet or database software, then analyzed with any of the standard statistical programs used in psychology (Matlab, Minitab, SPSS, SAS, R, etc.).

Online Testing

Other researchers test participants online. There are several ways to do this, including online surveys, and they do allow experimenters to collect massive amounts of data in a relatively short amount of time. Nevertheless, all of them involve several difficulties. For one, there is no straightforward way to know whether people are who they say they are, which only exacerbates the 'frequent flyer' problem. Second, online testing allows for less control over the participants themselves: one cannot readily know whether participants are completing the experiment while preparing dinner or using unauthorized resources to 'cheat.' Third, precise experimental control of timing parameters is more difficult online, particularly the millisecond timing that is important for many studies of cognition and perception. These limitations do not preclude online testing; they just mean that researchers must give some thought to the kinds of tasks that are appropriately administered online.

One possible avenue of administration involves Mechanical Turk, a service run by the web site amazon.com. Here, experimenters can post tasks for participants to complete – or, to use Mechanical Turk's terms, requesters can post human intelligence tasks (HITs) for workers. Experimenters can restrict their experiment to participants who meet specific criteria. Furthermore, participants are rated based on their performance (most relevantly, whether they completed the task promptly and adequately), which means that requesters can exclude participants who have proven unreliable in the past. Participants are paid through amazon.com, and the rates of payment are quite low – far lower than they are for in-person studies.

But are the results comparable? Several studies have examined this question. Crump et al. (2013), for instance, used Mechanical Turk to collect data on a series of standard cognitive tasks, such as the Stroop task, the flanker task, priming tasks, and so on. They found that the results were generally in accord with those obtained in the lab, but they also identified some important limitations. In particular, they recommend using some sort of measure to check whether participants understand the instructions, since participants have no easy way of asking questions during the task. They also recommend monitoring and reporting dropout rates, which tend to be higher in online experiments, and using caution when implementing experiments that require precise timing or detailed visual displays.

Conclusion

The traditional introductory psychology participant pool remains a common source of participants, but researchers are increasingly turning to online testing approaches to increase their sample size and recruit a broader and more representative group of participants. These changes might seem to foreshadow the end of in-person testing. Nevertheless, any number of experiments, such as those that involve millisecond timing, complex sensory stimuli, or certain forms of social interaction, simply cannot be run online with present technology. Even when running more straightforward experiments, there is simply no substitute for testing at least a few participants in person. Thus, the brick-and-mortar psychology lab will remain a fundamental part of the science for some time to come.

See also: Experimental Design: Overview; Experimental Laboratories: Social and Economic; External Validity;

Laboratory Studies: Historical Perspectives; Observational Studies: Overview.

Bibliography

Aviv, A.L., Zelenski, J.M., Rallo, L., Larsen, R.J., 2002. Who comes when: personality differences in early and later participation in a university subject pool. Personality and Individual Differences 33, 487–496.

Bowman, L.L., Waite, B.M., 2003. Volunteering in research: student satisfaction and educational benefits. Teaching of Psychology 30, 102–106.

Crump, M.J.C., McDonnell, J.V., Gureckis, T.M., 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PLOS One 8, e57410.

Elicker, J., McConnell, N.L., Hall, R.J., 2010. Research participation for course credit in Introduction to Psychology: why don't people participate? Teaching of Psychology 37, 183–185.

Elliott, L.J., Rice, S., Trafimow, D., Madson, L., Hipshur, M.F., 2010. Research participation versus classroom lecture: a comparison of student learning. Teaching of Psychology 37, 129–131.

Henrich, J., Heine, S.J., Norenzayan, A., 2010. The weirdest people in the world? Behavioral and Brain Sciences 33, 1–75.

Landrum, R.E., Chastain, G., 1995. Experiment spot-checks: a method for assessing the educational value of undergraduate participation in research. IRB 17, 4–6.

Padilla-Walker, L.M., Thompson, R.A., Zamboanga, B.L., Schmersal, L.A., 2005. Extra credit as incentive for voluntary research participation. Teaching of Psychology 32, 150–153.

Sears, D.O., 1986. College sophomores in the lab. Journal of Personality and Social Psychology 51, 515–530.

Sears, D.O., 2008. College student-it is redux. Psychological Inquiry 19, 72–77.

Sieber, J.E., Saks, M.J., 1989. A census of subject pool characteristics and policies. American Psychologist 44, 1053–1061.

Witt, E.A., Donnellan, M.B., Orlando, M.J., 2011. Timing and selection effects within a psychology subject pool: personality and sex matter. Personality and Individual Differences 50, 355–359.

Zelenski, J.M., Rusting, C.L., Larsen, R.J., 2003. Consistency in the time of experiment participation and personality correlates: a methodological note. Personality and Individual Differences 34, 547–558.

Experimental Laboratories: Social and Economic

Richard Gonzalez, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The experimental laboratory has become a key component in social and economic research. The article discusses how to prioritize the equipment and personnel needed to create a fully operational laboratory to answer research questions.

What is an experimental laboratory in social and economic sciences? A laboratory (lab) is defined as a place that contains equipment necessary for one's scientific work. Under this definition, the key phrase is 'equipment necessary.' In practice, the contents of a lab are dictated by the kind of research one does and the tools required to implement that research. The social psychologist studying communication patterns within a married couple requires a different lab than a social scientist studying the psychophysiological correlates of aggressive behavior or an experimental economist testing an implication of game theory. Sometimes a social science lab is merely a small room with a computer. As long as that space, and the equipment it houses, allows the research to be conducted, then the room can be called a lab. Much as Pygmalion carved his ideal mate out of ivory, a researcher creates a research environment from the equipment, setting, and people necessary to conduct research. Research labs will vary depending on the needs of the researcher.

The Computer

One piece of equipment that has become indispensable in just about all social and economic labs is the computer. At a minimum, the computer facilitates the editing of manuscripts. Writing experts claim that the ability to edit and rewrite one's prose is critical for clear, crisp writing. A sophisticated user of a word-processing program can format a manuscript in the manner required by a journal editor and easily construct a reference list from an electronic database. An Internet connection allows collaborators to exchange working drafts of articles either through electronic mail or shared disk space, and authors can submit manuscripts for publication electronically, as well as distribute reports to other interested parties. In this sense, the computer has replaced the typewriter as the machine used to write and the copier as the machine used to disseminate the written product of one's research. Dissemination of results is an important feature of all research projects – others need to benefit from the new knowledge that has been created.

There are, of course, more advanced uses of the computer in social and economic research. For instance, the computer can be useful in the selection of the research sample, data collection, and data analysis. Unlike the traditional paper-and-pencil questionnaire, administering a survey by computer permits automatic tailoring of follow-up questions depending on the answers the respondent provides to previous questions. Software can help to create and administer complicated sampling designs, and can facilitate elaborate laboratory procedures such as the presentation of research stimuli at a rapid rate, presenting questions in random order, or measuring the timing of a participant's response.

Some researchers have started placing their studies on the World Wide Web to complement their lab and field data collection (e.g., see http://www.socialpsychology.org/expts.htm and Amazon's Mechanical Turk Web site https://www.mturk.com/mturk/welcome). Researchers may allow unlimited access to their study or grant limited access to a select subsample. Care must be taken to create a representative sample. Sampling issues aside, conducting studies over the World Wide Web makes it convenient to collect data at multiple sites from any number of computer platforms. This permits data collection anywhere in the world at relatively little expense.

Computers are also used to analyze data. A simple spreadsheet program can be used for descriptive statistics and simple inferential tests, whereas specialized software is needed for more computationally intensive data analytic techniques such as bootstrapping and Bayesian algorithms. Even the analysis of some forms of qualitative data (such as narratives) can be facilitated through the use of software programs designed to code original text. For example, a word in the text could be coded as a noun, as a positive or negative word, or any number of other categories relevant to the particular research question.

The computer also aids lab management. For instance, scheduling and calendaring software can help organize interdependent parts of a research project on a time line. Such time lines are useful not only for facilitating the scheduling of lab facilities and lab meetings, but also for planning and writing grant proposals.

How many computers should a lab contain? What kind of computer should a lab have? These questions are difficult to answer in the abstract because so much depends on the research question. For example, one project may involve 200 handheld tablet computers because the study requires 200 married couples to answer questions each evening at home for a month. On the other hand, a researcher who studies social interaction through computer simulation may only need one computer with fast processors and extra memory. In sum, the needs of the research drive the equipment required in the lab.

What Is Your Research Question?

It is useful to consider three factors when designing a lab for social or economic research: the experimental context the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.41029-9

researcher needs to conduct the research, the methods the researcher will use, and the measurements the researcher will make. Each of these three issues is subsumed by the following more general question: what is the research question? If your research question addresses how Mayans conceptualize kinship terms, then you would need a special setting, special materials, and special measures – the lab will likely need to be portable. This research question requires a different lab setup than, say, a research question about trust in the context of an economic game that can be conducted in a controlled setting of lab located in a university building.

What Experimental Context Does the Research Require?

What context does the research require? For example, if one wants to study social behaviors in a bar, then it might be useful to design a lab setting that resembles a bar (complete with mood music, barstools, and liquor cabinets). If one wants to study jury deliberation, then it is useful to design a lab setting that resembles an actual jury room. These labs can be constructed to reproduce, or mimic, a real-world setting, and hence increase the external validity of the research, especially if these details influence behavior to be more realistic.

However, it is also possible to study social and economic behavior in an artificial setting, an analog setting that evokes similar processes as the real-world setting but is easier to create and easier to control. One example is the classic Milgram (1975) study on obedience. Milgram claimed that his obedience findings could help us understand how citizens in a totalitarian government may be pressured to follow superiors, *but* his research setting was low in external validity with respect to totalitarian governments. Milgram's research setting was successful at evoking the processes that make it difficult to refuse requests or refuse to follow orders. Milgram's lab setting involved an elaborate stage production that the experiment was about the effects of punishment on learning. *However*, the actual purpose of the study was to examine the social interaction between the participant (who believed he or she was playing the role of the 'teacher' in the study on learning) and the experimenter. Milgram's work is useful in demonstrating that much can be learned from artificial experimental settings. One must be careful, however, not to extrapolate too much from such artificial circumstances because of the loss of external validity.

Sometimes the researcher must assume the role of a stage manager in a theatrical company. The researcher attempts to create the setting that produces the reaction in the participant required by the research question. This may frequently include the use of stage props, as in the Milgram study that used a machine to administer (bogus) electric shock to the 'learner.' The role of a stage manager should not be underestimated. Social psychology experiments frequently are criticized for being low in external validity, but a study that engages the participant in the task is high in another important feature mundane realism (see Aronson et al., 1992, for a discussion of validity and mundane realism). As much as possible, studies should try to engage the participant because a bored, disengaged participant is likely to provide low-quality data.

Some research questions can be studied in a relatively simple setting. For instance, research investigating people's facial reactions to disturbing photographs would need a means for displaying the pictures (a slide projector, a computer, a photo album), so an appropriately lit room would be a sufficient setting. The participants could be videotaped while viewing the photographs and trained coders could rate the facial expressions. With the proper cameras and software it is also possible to use computer algorithms to automatically code the facial expressions in real time.

Tools for the Research Method

One of the defining features of an experiment is that it includes a manipulation, also known as an independent variable. The goal is to create at least two conditions that differ only on the particular feature that is manipulated. Random assignment of research participants to condition facilitates this goal.

Laboratory equipment may be needed to conduct the manipulation. For instance, work on eyewitness identification of a criminal suspect may require that the composition of the lineup be varied to include foils that are either high in physical similarity to the suspect or low in physical similarity. Thus, if the researcher uses a photolineup, then the lab should include a broad range of photographs so that many different lineups could be constructed.

A computer can be useful in carrying out experimental manipulations. For instance, if the manipulation requires varying the duration a stimulus is presented, then the computer can be used to control such durations. Experimental stimuli may involve presentation of video or animation and a computer can facilitate the creation, editing, and presentation of such material.

Sometimes the manipulation requires the use of drugs. For instance, research on arousal may require the participant in one condition to receive a dose of epinephrine whereas a participant in another condition to receive a dose of a neutral sugar pill. Some researchers in decision-making use oxytocin to manipulate trust while playing an economic game. Such research requires equipment necessary to store and administer the drugs and in some cases appropriate medical professionals on site. Careful attention should be given to proper tracking of the use of the substances such as storing them in a safe that requires two people with different keys to access. Participants should be cleared for safety prior to leaving the study. For example, a study examining behavior in an experimental bar that involves consuming alcohol should ensure that participants are not intoxicated when leaving the lab.

Measurement Devices

An important feature of empirical work is that something is measured. Even qualitative research has measurement, such as the recording and interpretation of a narrative. The lab must include the tools necessary to make and store such measurements.

Standard measurement devices include an audio recorder, a video recorder, a computer with interactive software, and pencils for participants to write responses. Some research may also require measurements of the electrophysiology of the brain or autonomic nervous system, which requires specialized equipment to store and analyze samples (e.g., measurement of cortisol level in saliva).

Some experiments may require the use of standardized scales such as personality tests, intelligence tests, or scales to measure depression. Because some of these scales are copyrighted, researchers may need to contact the developers of the scale for permission to use. Again, which scales are needed in the lab depend on the research question. For instance, if the research question examines the relation between intelligence and aggressive behavior, then scales to assess intelligence would be needed on hand.

Managing a Lab

Much like large offices, some experimental labs may require a lab manager to handle the day-to-day functioning of the lab. The manager's tasks may include scheduling the use of lab resources, monitoring research budgets, and staying abreast of manuscript and grant proposal deadlines. Sometimes a lab requires more specialized personnel, such as a technician when electrophysiological recording equipment is used or a programmer if specialized software is needed. Lab personnel may also handle maintenance of lab equipment and storage of data.

Summary

New researchers frequently ask, 'What kind of equipment should I include in my lab?' As argued in this article, one should reply with the following questions: What is the research question? What experimental setting is needed for the research? What equipment is needed to conduct the research? What equipment is needed to collect and analyze the measurements? How will the administration of the lab be handled?

Further Reading

For details on how to conduct experiments and an explanation of experimental methodology see Aronson et al. (1992). For an explanation of how to analyze data from an experiment see Gonzalez (2009). For suggestions on how to write an article describing empirical work in social psychology see Bem (1986) and the *Publication Manual of the American Psychological Association* (see http://www.apa.org). For information on the ethical treatment of research participants see professional society Web sites such as the American Psychological Association (e.g., http://www.apa.org/monitor/jan03/principles.aspx).

See also: Experimental Laboratories: Psychological; Experimental Laboratories: Social and Economic; Laboratory Studies: Historical Perspectives; Problem Selection in the Social Sciences: Methodology.

Bibliography

Aronson, E., Ellsworth, P., Carlsmith, J.M., Gonzales, M., 1992. Methods of Research in Social Psychology. Random House, New York.

Bem, D.J., 1986. Writing the empirical journal article. In: Zanna, M., Darley, J. (Eds.), The Compleat Academic: A Practical Guide for the Beginning Social Scientist. Random House, New York, pp. 171–201.

Gonzalez, R., 2009. Data Analysis for Experimental Design. Guilford Press, New York.

Milgram, S., 1975. Obedience to Authority: An Experimental View. Harper & Row, New York.

Experimental Methods in Survey Research in Demography

Duncan Thomas, Duke University, Durham, NC, USA
Elizabeth Frankenberg, Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Randomized controlled trials (RCTs) have been profitably used to identify causal effects in population research. However, the design and implementation of social experiments is not straightforward and it is not clear that it is either feasible or desirable to attempt to answer some questions in population using only the so-called 'gold standard' double-blind RCT. It seems likely that the integration of the creative use of theory with the advantages of both RCTs and nonexperimental study designs has the greatest hope of advancing scientific knowledge about population behaviors and processes.

Introduction

A goal of population research is to isolate causal effects of, for example, policies, programs, circumstances, and behavioral choices. For example, what is the causal effect of improved access to family planning services on the health and economic status of women? What is the impact of marriage on health and well-being? What is the impact of migration on those who move and those who are left behind? Answering these types of questions with observational studies is challenging because the counterfactual is typically not observed: what would have happened if the migrant had not moved is not known. It is difficult to separate the impact of the stimulus (the decision to migrate) from other, unobserved factors that distinguish those who experience the stimulus (those who choose to migrate) from those who do not (those who do not choose to migrate). The promise of social experiments is that they provide a method for putting aside these unobserved differences and, thereby, identify causal effects.

In the ideal randomized controlled trial (RCT), study subjects are randomly assigned to receive a stimulus (the treatment group) or a placebo (the control group). After the stimulus, the difference between the average outcome of the treatment group and the average outcome of the control group provides a measure of the causal effect of the stimulus on the outcome. Because the study subjects are randomly assigned to receive the stimulus, concerns about what drives selection into each group (or unobserved differences between the groups) are, by design, put aside.

Implementation of the ideal RCT is far from straightforward in the best of circumstances. Implementing social experiments to address important questions in the population sciences is very challenging. To examine the impact of migration, for example, one cannot simply move a randomly selected group of people to a new location and force them to remain there while forcing those in the control group to stay put. It may be possible, however, to subsidize the costs of migrating, broadly defined, or, in the case of international migration, adopt a policy of randomly assigning visas to a subset of those who apply. Aspects of each of these designs raise important issues regarding the question that is addressed and the interpretation of the estimates. The central point is that social experiments have the potential to be very valuable in population research but, as is the case with observational studies, they have strengths and weaknesses.

Indeed, beyond identifying causal effects, it is important to understand the mechanisms that underlie behavioral choices. To this end, it seems the greatest promise of RCTs and social experiments lies in their fuller integration with theory and observational designs to advance understanding of population processes, a position advocated by Heckman (2010).

Historical Perspective

RCTs have a long history that, according to Stigler (1992), dates back at least to Peirce's work in psychology (Peirce and Jastrow, 1885). The statistical foundations were formalized in the 1920s by Neyman (1923) and Fisher (1935) and the methods were used in the agriculture sciences to evaluate the impact of inputs and practices on output. RCTs were later adopted in medicine to evaluate the impact of treatments on patients and, it is often argued, that double-blind RCTs are the 'gold standard' for evidence in that field. Whereas randomization of agricultural plots is relatively easy, randomization with human subjects is more complicated, particularly when the stakes – life or death, for example – are high. Without question, randomized controlled social experiments are even more complicated to design and implement.

In the 1960s and 1970s, several large-scale, influential social experiments were fielded to examine, for example, the impact of providing an income guarantee through a negative income tax (Orcutt and Orcutt, 1968), of varying the costs of health insurance (in the RAND Health Insurance Experiment, Newhouse et al., 1996), and of providing job training to youth and unskilled adults (Bloom et al., 1997). In the population field, studies in demographic surveillance sites in Navrongo, Ghana and Matlab, Bangladesh pioneered the use of social experiments to investigate the impact of improved access to family planning and maternal and child health programs on health and well-being (Phillips et al., 1988, 2006).

In recent years, the adoption of RCTs has exploded, particularly in development economics, led in large part by Banerjee and Duflo along with their colleagues, many of whom are associated with the Abdul Latif Jameel Poverty Action Lab at MIT (Duflo, 2007; Duflo et al., 2008; Banerjee and Duflo, 2009, 2010, 2011). Deaton (2009, 2010) has been a leading critic of the

rise of RCTs in the field. Over the last 20 years, Heckman (1992, 2010) has played a pioneering role in exploiting the advantages and addressing the disadvantages of social experiments.

The Ideal RCT

We begin with the ideal or 'gold standard' RCT in which study subjects are randomly assigned, at the individual level, to a treatment or a control group, with neither the study subjects nor investigators knowing who has been assigned to which group (that is, the study is double-blind). To lay the foundation, consider the linear model

$$\theta_i = \beta_0 + \beta_1 T_i + \varepsilon_i \quad [1]$$

in which $T_i = 1$ if individual i is treated and 0 if not. The goal is to estimate β_1 to measure the causal impact of a stimulus or treatment, T_i, on the outcome of interest, θ_i. Put another way, we are interested in estimating

$$\beta_1 = E\left[\left(\theta_i^T - \theta_i^C\right)\middle|T_i = 1\right] \quad [2]$$

where θ_i^T is the outcome when an individual is treated and θ_i^C is the outcome when the individual is not treated. For example, we may be interested in measuring the causal impact of a treatment that improved physical health on hourly earnings. In this example θ_i^T is hourly earnings of a person who received the treatment (took iron supplements); θ_i^C is what the same person would have earned per hour in the absence of receiving the treatment. Obviously, the counterfactual, the outcome if the individual who was treated had not been treated (denoted θ_i^C in eqn [2]), is not observed. However, the difference between the average earnings of those who received the health-improving intervention and those who did not can be observed and so we can estimate

$$\widehat{\beta}_1 = E\left[\theta_i^T\middle|T_i = 1\right] - E\left[\theta_i^C\middle|T_i = 0\right] \quad [3]$$

Rewriting eqn [3]

$$\widehat{\beta}_1 = E\left[\left(\theta_i^T - \theta_i^C\right)\middle|T_i = 1\right] + E\left[\theta_i^C\middle|T_i = 1\right] - E\left[\theta_i^C\middle|T_i = 0\right] \quad [4]$$

we have

$$\widehat{\beta}_1 = \beta_1 + E\left[\theta_i^C\middle|T_i = 1\right] - E\left[\theta_i^C\middle|T_i = 0\right] \quad [5]$$

where the second and third terms in eqn [5] capture the effect of 'selection,' that is the impact on $\widehat{\beta}_1$ of unobserved heterogeneity, ε_i, correlated with the treatment, T_i, in eqn [1]. Returning to our example, the second term is the average earnings that the treatment group would have earned if they had not been treated, the counterfactual that cannot be observed. The third term is the average earnings of those who were not treated. If the difference between the second and third terms is zero then $E[T_i\varepsilon_i] = 0$ and $\widehat{\beta}_1$ is a consistent estimate of the causal impact of T_i on θ_i:

$$\text{plim}_{n\to\infty}\widehat{\beta}_1 = \beta_1 \quad [6]$$

On the one hand, if study participants who received the treatment would have earned more in the absence of the treatment than participants in the control group, then $E[\theta_i^C|T_i = 1] - E[\theta_i^C|T_i = 0] > 0$ and the treatment group would be positively selected, relative to the controls. In that case, $E[T_i\varepsilon_i] > 0$ and $\widehat{\beta}_1$ would be an upward biased estimate of the causal impact of T_i on θ_i. This might arise, for example, if subjects have to apply to participate in the treatment and those who apply are those who would have earned more even in the absence of the treatment. On the other hand, if, relative to controls, the study participants would have earned less in the absence of the treatment, then $E[\theta_i^C|T_i = 1] - E[\theta_i^C|T_i = 0] < 0$ and the treatment group would be negatively selected, $E[T_i\varepsilon_i] < 0$ and $\widehat{\beta}_1$ will be downward biased. This might arise if those with a lower value of time (and lower potential hourly earnings) are also more likely to apply for and participate in the treatment.

Experimental studies are designed so that the selection term $E[\theta_i^C|T_i = 1] - E[\theta_i^C|T_i = 0]$ is zero so that $E[T_i\varepsilon_i] = 0$ and eqn [6] is true. In a double-blind RCT, this can be achieved by randomly assigning subjects to either the treatment or control group and then exposing only the treatment group to the stimulus and, after the treatment has had an impact, comparing the outcomes (or changes in outcomes) of the treated and control subjects. In an RCT that is designed and implemented well, the second and third terms in eqn [5] will be zero by construction. The average treatment effect (ATE) is an unbiased estimate of the causal effect of the treatment on the outcome.

Key Features of the 'Ideal' RCT

There are at least three features of the 'ideal' or 'gold standard' RCT model that are critical for estimated treatment effects to be consistent. First, study subjects assigned to the treatment should be 'exchangeable with' those assigned to the control group; i.e., in terms of any characteristic that is related to the outcome of interest, the treatments and controls are effectively interchangeable.

In some studies, randomization is conducted after stratifying on observable characteristics (such as age, gender, race, education, or location) and study subjects are randomized within each subgroup. Analyses conducted on each subgroup satisfy the conditions necessary to interpret the estimated effects as causal. The reverse is not true: if study subjects are randomly assigned to treatment and control without stratifying by group, it is inappropriate to assume that the selection term in eqn [5] is zero within groups. For example, say you are interested in examining the effect of a treatment on males and females separately but randomization was not conducted separately by gender. We cannot assume that the selection term in eqn [5] is zero for all the study subjects and also for males as well as for females, taken separately. That may be true, and if so, then we are lucky. However, since it is not built into the study design, one cannot assume that it is true.

This also highlights that RCTs lean heavily on the assumption of linearity in eqn [1] since the mean of the difference between treatment and control outcomes is the same as the difference between the mean treatment and mean control outcome. Estimation of heterogeneity of effects calls for additional assumptions that may not be correct and will typically need to be justified by drawing on a theory or at least a mechanism that lies outside the RCT study design (Deaton, 2010; Heckman and Smith, 1995).

Second, attrition, which is a key concern in every longitudinal study, has the potential to be particularly pernicious in an RCT. If posttreatment follow-up is selected on benefits (or costs) to those in the treatment group relative to the benefits (and costs) to those in the control group, the attrition will potentially contaminate estimates of the treatment effect. Since outcomes of subjects who are lost to follow-up are not measured, the effects of attrition are necessarily absorbed in unobserved heterogeneity, ε, in eqn [1] and it is not obvious that it is reasonable to assume $E[T_i\varepsilon_i] = 0$. To address this concern, some RCT studies establish that attrition is unrelated to treatment status and, in some cases, to observed baseline characteristics. While such results are comforting, they do not establish that attrition is ignorable. For example, if attrition is related to the impact of the treatment – say those who do not benefit much from the treatment exit the study – then it may be difficult to detect this sort of selection in the absence of a posttreatment measure of the outcome of interest. In the 'gold standard' RCT there is no place for attrition; achieving that in a population-based study is a tall order.

The third critical feature of the double-blind individual RCT is that outcomes of the control group must not be affected by the treatment to assure that the second and third terms in eqn [5] are equal. This rules out, for example, any response by the controls to the fact that they were not selected for the treatment. In a double-blind RCT, neither the study subjects nor the study investigators know who is assigned to receive the treatment or who is assigned to the control group. In that case, neither the treatments nor the controls have the information necessary to change their behavior because they know they are (or are not) in the treatment group.

The 'gold standard' RCT study is double-blinded because a similar concern applies to study investigators who may treat those in each group differently, albeit implicitly. This sort of behavior has been described in some drug trials, even in double-blinded studies, as investigators have been able to infer who is in the treatment group (by examining biomarkers, for example) and have adjusted their own behavior or treatment in order to help the subject (see, for example, Lauritsen, 1990). Studies have also documented instances of nonrandom assignment of subjects to the treatment or control group. For example, in studies that have been implemented by people who have a stake in the outcome, or want the best for the subjects, and allocate those subjects whom they believe will benefit most from the treatment to that group. The double-blind design is intended to assure these types of contamination do not occur.

Implementation of Double-Blind RCTs in Population Studies

In an ideal study design, the RCT involves a comparison of the impact of a treatment with an identical looking placebo (which may be an alternative treatment) to assure the study is double-blind. This approach has been used in population research. For example, an RCT implemented in a district in Central Java, Indonesia, tested the hypothesis that health has a causal effect on economic productivity by comparing changes in hourly earnings of iron-deficient individuals who had been randomly assigned to receive iron supplements with earnings of controls who received an identical-looking sugar pill (Thomas et al., 2014).

However, many studies in the population sciences seek to measure the impact of treatments such as improved access to family planning or other health-related services; participation in an education or training program; access to credit, financing, or an income transfer program. In such cases, it is not possible to design a double-blind RCT because there is no placebo that can be given to the controls that appears to be identical to the treatment.

This raises concerns with contamination of the estimated treatment effect because of behavioral responses by study participants to random assignment or to the treatment itself. In a study of a training program, for example, if some of the subjects assigned to the treatment group choose to not participate in the program or drop out of the program before it is completed then it will not be possible to measure the program effect on those subjects. On the other side of the coin, if some of the subjects who are assigned to the control group participate in another, related training program, then the estimated treatment effect will be biased. One of the motivations for randomly assigning some people to participate in a program is because the program is oversubscribed. In this situation, treatments are more likely to participate in the program, but control subjects are likely to be more motivated to seek out alternatives to the program treatment. Move to Opportunity (MTO) is a major randomized housing mobility experiment that was implemented in the 1990s in which 4600 low-income families with children living in public housing in disadvantaged urban neighborhoods were offered the chance to move to private market housing in better neighborhoods. Families were randomized to receive a housing voucher that subsidized private market rents (with half of them carrying some constraints on where they could relocate) while the controls received no vouchers. Only about 50% of those who received a voucher managed to relocate to a better neighborhood and a substantial fraction of the control subjects also relocated (Ludwig et al., 2012).

Intent to Treat and Treatment on the Treated Estimates

One approach to addressing these concerns is to estimate the intent to treat effect which measures the impact of the program on all the subjects who were assigned to the treatment relative to all the subjects who were assigned to control, irrespective of whether they actually participated in the program. This contrasts with the estimate of the effect of the treatment on the treated (TOT) which includes only those subjects who complete the treatment and compares them with the controls. Clearly, the TOT estimate cannot be assigned a causal interpretation. Studies have used random assignment status as an instrument for participation in the treatment in an instrumental variables framework to identify the causal impact of the treatment on those who chose to participate. This does not provide evidence on the likely impact of the treatment on those who did not choose to participate (Heckman, 1992).

Level of Randomization

Another approach to addressing the concern that there may be behavioral responses by subjects to being assigned to the treatment or control group is to eschew randomization at the individual level and, instead, randomize at a higher level such as the community, school, health center, or microcredit agency. In this case, if the study is able to exclude people who are not in the community (or school etc.) at the time of randomization from receiving the treatment, then it is possible to mitigate the impact of behavioral responses to the treatment and control assignments. In practice, this can be difficult and may raise ethical issues. In addition, such a design may have unintended impacts on behavior.

PROGRESA is a very large and ambitious conditional cash transfer (CCT) that was implemented in rural Mexico after a rigorous randomized control trial. Parker et al. (2008) provide a description and review results from the evaluation. The evaluation study selected 506 low-income communities and randomly assigned about two-thirds of those communities to receive the CCT program in 1998 with the other third of communities being told they would receive the treatment 3 years later (although the treatment was provided to households in these communities 18 months after the evaluation began). The CCT was means tested at the household level based on a baseline survey. Over and above a basic transfer that depended on household size and composition, additional payments were made conditional on children's school attendance and participation by household members in activities at the health center. In some households, the payments were very large relative to their pretransfer income.

Studies have established that households that were eligible to receive the transfers had higher levels of spending, improved health and schooling outcomes, and invested more in productive assets. However, there is also evidence that noncoresident family members living in the same community as a beneficiary household also benefited from the program, indicating sharing of resources among extended families. This suggests that impacts of the CCT on outcomes in beneficiary households may be underestimated. On the other hand, the program was designed so that individuals and households could not move to a treatment community and receive the treatment; eligibility for the treatment was determined at the baseline survey. Moreover, if a member of a treatment household moved away from the community, the individual did not take the CCT with him or her and the baseline household payment was reduced accordingly. Migration out of poor rural communities is widespread in Mexico and it is a mechanism used to improve the well-being of both the migrant and those left behind. By design, the PROGRESA evaluation provides a disincentive for young adults in beneficiary households to move away in search of a better life. The same disincentive does not apply to young adults in households in control communities. This further complicates interpretation of the estimated impact of PROGRESA. Since the evaluation study did not follow movers, it is difficult to assess the likely magnitude or even direction of this effect. Moreover, attrition is a serious problem in the evaluation study: loss to follow-up is substantial (with about 20% of the study participants being lost by the second follow-up) and it is potentially selective. Nonetheless, PROGRESA has been judged a great success, was implemented across all rural and later to urban Mexico, and renamed Oportunidades.

Estimation of Variability

In model [1],

$$\theta_i = \beta_0 + \beta_1 T_i + \varepsilon_i$$

the variance of errors may not be the same among subjects in the treatment and control groups. Consistent estimation of the standard error of the treatment effect, $\widehat{\beta}_1$, is straightforward with an estimator that takes heteroskedasticity into account, such as the Huber estimator or the jackknife (Huber, 1967; Efron, 1982).

In studies, like PROGRESA, in which randomization is at the community (or group) level, rather than at the individual level, estimation of variability is more complicated. Rewrite model [1]

$$\theta_{ig} = \beta_0 + \beta_1 T_g + \varepsilon_{ig} \qquad [7]$$

where treatment, T, is defined at the group, g, level. Unobserved heterogeneity, ε_{ig}, varies across individuals and may be correlated within groups. Failure to take into account within-group correlations (or clustering) of residuals yields estimates of variability that are biased, typically (but not always) downwards. Adoption of an estimator like the Huber or jackknife estimator that allows for both heteroskedasticity and the block structure of the residuals provides consistent estimates of the standard error of the treatment effect.

Huber-type estimators do not perform well when the number of groups is small (say less than 15), the bias in estimators that ignore the problem can be quite large. A wild clustered bootstrap procedure has been shown to perform well in these situations (Cameron et al., 2008). While the concern arises in both experimental and nonexperimental studies, it is particularly salient in RCTs that randomize at the group level with few groups.

In some studies, the model is extended to include covariates, X, which may be measured at the individual or group level:

$$\theta_{ig} = \beta_0 + \beta_1 T_g + \beta_2 X_{ig} + \varepsilon_{ig} \qquad [8]$$

The idea is that, if randomization has not been violated, the covariates should have no impact on the estimated treatment effect, $\widehat{\beta}_1$, because they are, by design, orthogonal to the treatment. The role of the covariates is to absorb unobserved heterogeneity and thus increase efficiency of the estimates. The strategy is, however, not without controversy (Freedman, 2008).

Generalizability of Evidence

Population scientists are often concerned with the generalizability of results in both observational and experimental studies. However, it is discussed most often in the context of experiments that are conducted on selected subsamples. For example, some experimental studies are based on convenience samples of respondents who sign up for a study or are recruited for a study without paying attention to whether the study

sample is representative of a well-defined population of interest. The external validity or generalizability of results from such studies to a broader population is difficult to assess since there are potentially unobserved characteristics in the broader population that are not present in the study sample. In that case, one of the key advantages of the RCT design, that $E[T_i\varepsilon_i] = 0$, may not be true in the broader population and so little can be inferred about the generalizability of the evidence.

Experimental studies that use the oversubscription approach to recruit subjects are potentially prone to this criticism if the goal of the study is to provide evidence on the impact of the program beyond those who have signed up for it. For example, it is not clear that the results of the evaluation of a training program based on an oversubscription recruitment strategy will generalize to all the unemployed. This concern will arise if, relative to all the people who are eligible for a training program, those who sign up to participate are more motivated to do well, possibly because they have recently experienced a dip in earnings or a recent, short unemployment spell (Heckman, 1992). As with nonexperimental studies, it is possible to make additional behavioral assumptions in order to estimate bounds on the treatment effect and its distribution, potentially exploiting both the experimental variation itself and also the design of the experiment such as the potential outcomes for those who receive the treatment(s) and those who do not (Heckman et al., 1997).

Drawing out generalizable findings is more complicated when there are behavioral responses by the subjects to the study itself or to the fact that the subjects have been assigned to either the treatment or control group. These are often called Hawthorne effects which arise when study subjects respond to the stimulus of being studied. Again, this is a general concern in social research, particularly studies that involve intensive interaction between the subjects and investigators. The issue can be more complicated in experimental studies if subjects react to the fact that they were assigned to either the treatment or control group and draw inferences from that assignment. One of the difficulties with designing experimental studies is that it can be difficult to predict behavioral responses to the study or its design. For example, in an ambitious large-scale study of the impact of incentives on police performance in India, subjects in one arm of the study who were not selected for the treatment felt discouraged. This had a deleterious impact on their work effort and so the treatment (or lack of it) affected the behaviors and outcomes of the controls, substantially complicating interpretation of the results (Banerjee et al., 2014).

In many cases, observational and experimental research studies are small in scale which raises questions about both generalizability and whether the study results are likely to hold at a larger scale. For logistical and cost reasons, experiments are often conducted in a small number of study sites and, in some cases, the sites are selected because an implementing partner or institution is operating in that area. Generalizability will be compromised if the study results are driven by characteristics that are unique to the site or implementing partner. For example, if the implementing partner in a study of the impact of microcredit is a particularly effective microcredit provider in an area where the program has been established, it is not clear that study results will generalize to settings in which providers are not as effective or programs have not already been established. Similar issues arise with studies that involve schools or health providers that are committed to improve outcomes but are not representative of the full set of institutions that will be called upon to implement changes if the treatment is scaled up. These issues are familiar to population scientists and studies in the population field are often designed with subjects and study sites (and possibly partners and implementers) selected so that they are representative of a well-defined underlying population.

Equilibrium and Longer-Term Effects

Scaling up an experiment raises a related but different issue: there may be equilibrium effects that are not captured by measuring only the treatment effect. Consider an experiment in which the treatment that improves the health of subjects results in their being more productive (by providing iron supplements in an area where iron deficiency is widespread) and, as a result, say the treated subjects have more rice to sell on the market. There may well be equilibrium effects if the increased supply of rice results in a reduced price (if the local market is closed) or increased income in the community (if more rice is sold outside the local market) which may, in turn lead to increased opportunities for work. To explore these effects, the iron supplementation study in Central Java, Indonesia (Thomas et al., 2014), was designed so that the fraction of the population recruited for the study in each of the 146 desa (villages) varied and the fraction of the study subjects who were at risk of benefiting from the supplements also varied. In addition, some desa were randomly assigned to not receive either the treatment or the control so that they serve as additional controls, uncontaminated by the study itself.

In an RCT that provided rainfall index insurance to cultivators in India, demand for hired labor and market wages was more sensitive to rainfall with the insurance. This resulted in reduced uncertainty among cultivators but increased uncertainty among agricultural wage laborers. This effect would not have been observed if the study had focused only on the impact of the RCT on cultivator choices (Mobarak and Rosenzweig, 2014). The point is not that equilibrium effects are inherently bad (or good) but, as is the case in observational studies, drawing on theory to identify potential equilibrium effects and designing experimental studies to explore the empirical importance of such effects is likely to have a substantial pay-off.

Few experiments are designed to provide evidence on longer-term effects. From a theoretical perspective, the impact of a short-run increase in income is likely to be different from the impact of a permanent increase in income or an increase in income that is expected to last for the long term. It is difficult to infer the impact of a permanent increase in income with the PROGRESA design. This is an important limitation, inherent in many RCT studies that are designed to provide a treatment for a limited period of time. This is particularly important when effects of the intervention are likely to cumulate over time – as subjects save enough to start a business, for example. A similar issue arises in the RAND Health Insurance Experiment that randomly assigned health insurance with different co-payments and deductibles to subjects. The program was provided for 3 years. If the benefits of health insurance cumulate, the experiment is unlikely to

uncover the full benefits of the insurance. In fact, subjects who received the most generous benefits were more likely to have a birth during the 3-year window, indicating shifts in the timing of fertility in response to the time-limited nature of the experiment (Newhouse et al., 1996).

A related, but different, feature of many RCTs is that they are designed as stand-alone studies that do not follow subjects for the long term. In those cases, it is not possible to measure the longer-term impact of the specific (short-term) intervention. This is not a feature of RCTs, per se, but rather reflects the way many have been implemented, the costs of the studies, and the costs of longer-term follow-ups. In some cases, it is unlikely to be feasible to design and implement an RCT that will provide answers to question many years hence. For example, measurement of the impact of prenatal exposures or early childhood experiences on health and mortality at older ages is unlikely to garner much support. However, tracking subjects over the life course, and measuring the evolution of impacts over this period has the potential to be very profitable. Indeed the MTO evaluation as well as evaluations of the impact of early child interventions in adulthood (Hoddinott et al., 2008; Heckman et al., 2013; Campbell et al., 2014) are superb examples of the value added to science of following treatments and controls in RCTs over the long haul.

A complementary approach is to build randomized variation into broad-purpose longitudinal population-based surveys that are designed to follow individuals over the longer term, independent of the RCT. The Matlab Health and Socioeconomic Survey (MHSS) (Rahman et al., 1999) has been a pioneer in this regard. MHSS is a large-scale population-based longitudinal study that has been conducted in conjunction with a long-term demographic surveillance study conducted by "the International Centre for Diarrhoeal Disease Research, Bangladesh" in Matlab district, Bangladesh, and has been used to evaluate, for example, maternal and child health and family planning interventions (Schultz and Joshi, 2013).

Integrating Theory and Estimation with Experimental Designs

Recent studies have better integrated RCTs with theory and other research designs. An increasing number of experiments in the social sciences are designed to reach beyond measurement of an ATE of a program or policy and test theories about behavior or identify mechanisms that plausibly explain behaviors and outcomes. Many of these designs have roots in behavioral economics.

In an early example, Ashraf et al. (2006) designed a commitment savings product that was offered to a randomly selected sample of current and former clients of a bank in the Philippines in order to provide insights into intertemporal preferences and savings behaviors. Several studies have used experimental variation in combination with models to examine how price variation affects consumer demand; these studies have described, for example, the role that prices play in signaling information and how short-run subsidies affect consumption over the longer run (Dupas, 2014).

Another line of research exploits randomly assigned variation in experimental designs to validate estimation of complex structural models that have a strong link to the underlying theory and, by necessity, make potentially strong assumptions. Todd and Wolpin (2006) use data from the PROGRESA evaluation study to estimate a dynamic behavioral model of parental decisions about fertility and child schooling. Estimates are based on data from subjects collected prior to receipt of program income and model estimates are validated by comparing predictions of program effects from this model with experimental results. The model performs very well. This is important because, without this external validation, it is very difficult to confidently assess the quality of model estimates and, therefore, the assumptions underlying the model. Furthermore, the structural model is designed to conduct simulations of alternative policy designs – in this case conditional transfer schedules – and the authors conclude that the same program impacts could be achieved with fewer resources, a result that is consistent with conclusions, which Schultz (2004) drew based on estimation of PROGRESA program impacts.

Results from experimental studies have been profitably used to evaluate assumptions in observational studies (Lalonde, 1986). Currie and Thomas (1995) and Garces et al. (2002) use longitudinal nonexperimental survey data to examine the impact of Head Start on children, adolescents, and young adults by comparing siblings, one of whom attended Head Start and one did not. It is not obvious that this identification strategy is valid and so they assess the strategy by first replicating the main results from the Perry Preschool Project. That project is an RCT that measured the impact of early child education on poor children, and served as a model for Head Start. The use of nonexperimental data and methods provided evidence of the impact of Head Start on a broader and more diverse population than the Perry Preschool study population as well as the impact on outcomes and at ages that were not assessed in the Perry project. The credibility of the nonexperimental results depends critically on the validation from replication of the RCT results.

Conclusions

In principle, the ideal or 'gold standard' RCT is an extremely powerful tool for identification of causal effects in the population sciences. However, implementation of an RCT that meets this ideal is far from straightforward in medicine and is even more complicated in the context of social experiments. In practice, the design and implementation of an RCT in the population sciences is fraught with problems. But the same is true, arguably a fortiori, of observational studies. The key advantage of an RCT is that the investigator has the potential to randomly assign who is exposed to a stimulus or treatment that is of central interest to the study. One of the disadvantages is that the bigger the impact of the treatment, the more likely that study participants will seek to undo the advantages of randomization.

Garber et al. (2014) argue that "In the absence of theoretical or methodological breakthroughs, the only possibility for further learning comes from experiments, particularly experiments with strong external validity." It is unclear this is true: the argument that one approach to empirical inquiry dominates all others under all circumstances is extreme. Whereas some questions in the

population and social sciences are unlikely to be credibly answered without an experimental design, others are not amenable to experimental manipulation and cannot possibly be addressed with an RCT. RCTs are but one arrow in the population scientist's quiver and a potentially very powerful one at that. Rather than toss out all the other arrows, it seems likely that scientific knowledge about population behavior and processes will be advanced through the integration of creative use of theory with the clever design of both experimental and nonexperimental studies. Recent studies suggest that, at least in this regard, future prospects for the population sciences are very bright.

See also: Cluster Randomized Trials, Ethical Implications of; Experimental Design: Large-Scale Social Experimentation; Experimental Design: Randomization and Social Experiments; Experimental Economics; External Validity; Field Observational Research in Anthropology and Sociology; Internal Validity; Observational Studies: Overview; Optimal Experimental Design; Psychological Treatments: Randomized Controlled Trials; Reliability and Validity Assessment: New Approaches.

Bibliography

Ashraf, N., Karlan, D., Yin, W., 2006. Tying Odysseus to the mast: evidence from a commitment savings product in the Philippines. Quarterly Journal of Economics 121 (2), 635–672.

Banerjee, A., Duflo, E., 2009. The experimental approach to development economics. Annual Reviews of Economics 1, 151–178.

Banerjee, A., Duflo, E., 2010. Giving credit where it is due. Journal of Economic Perspectives 24 (3), 61–80.

Banerjee, A., Duflo, E., 2011. Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty. Perseus, Philadelphia.

Banerjee, A., Chattopadhyay, R., Duflo, E., Keniston, D., Singh, N., 2014. Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy and Training, Mimeo.

Bloom, H., Orr, L., Bell, S., Cave, G., Doolittle, F., Lin, W., Bos, J., 1997. The benefits and costs of JTPA Title II-A programs: key findings form the national job training partnership act study. Journal of Human Resources 32 (3), 549–576.

Cameron, A., Gelbach, J., Miller, D., 2008. Bootstrap-based improvements for inference with clustered errors. Review of Economics and Statistics 90 (3), 414–427.

Campbell, F., Conti, G., Heckman, J., Moon, S.H., Pinto, R., Pungello, E., 2014. Early childhood investments substantially boost adult health. Science 343 (6178), 1478–1485.

Currie, J., Thomas, D., 1995. Does head start make a difference? American Economic Review 85 (3), 341–364.

Deaton, A., 2009. Instruments of development: randomization in the tropics and the search for the elusive keys to economic development, Keynes lecture in economics. Proceedings of the British Academy 162.

Deaton, A., 2010. Instruments, randomization and learning about development. Journal of Economic Literature 48 (2), 424–455.

Duflo, E., 2007. Field experiments in development economics. In: Blundell, R., Newey, W., Persson, T. (Eds.), Advances in Economic Theory and Econometrics, Econometric Society Monograph, vol. 42. Cambridge University Press, Cambridge (Chapter 13).

Duflo, E., Kremer, M., Glennerster, R., 2008. Using randomization in development economics research: a toolkit. In: Schultz, T.P., Strauss, J. (Eds.), Handbook of Development Economics, vol. 4. Elsevier, Amsterdam (chapter 15).

Dupas, P., 2014. Short -run subsidies and long-run adoption of new health products: evidence from a field experiment. Econometrica 82 (1), 197–228.

Efron, B., 1982. The jackknife, the bootstrap and other resampling plans. In: SIAM CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia.

Fisher, R.A., 1935. The Design of Experiments. Macmillan, London.

Freedman, D.A., 2008. On regression adjustments to experimental data. Advances in Applied Mathematics 40, 180–193.

Garber, A.S., Green, D.P., Kaplan, E.H., 2014. The illusion of learning from observational research. In: Teele, D.L. (Ed.), Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences. Yale University Press, New Haven.

Garces, E., Thomas, D., Currie, J., 2002. Longer-term effects of head start. American Economic Review 92 (4), 999–1012.

Heckman, J., 1992. Randomization and social policy evaluation. In: Manski, C., Garfinkel, I. (Eds.), Evaluating Welfare and Training Programs. Harvard University Press, Cambridge, pp. 201–230.

Heckman, 2010. Building bridges between structural and program evaluation approaches to evaluating policy. Journal of Economic Literature 48 (2), 356–398.

Heckman, J.J., Smith, J.A., 1995. Assessing the case for social experiments. Journal of Economic Perspectives 9 (2), 85–110.

Heckman, J., Pinto, R., Savelyev, P., 2013. Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. American Economic Review 103 (6), 2052–2086.

Heckman, J., Smith, J., Clements, N., 1997. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. Review of Economic Studies 64, 605–654.

Hoddinott, J., Maluccio, J., Behrman, J., Flores, R., Martorell, R., 2008. Effect of a nutrition intervention during early childhood on economic productivity in Guatemalan adults. Lancet 371, 411–416.

Huber, P.J., 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 221–233. Berkeley.

Lalonde, R., 1986. Evaluating the econometric evaluations of training programs with experimental data. American Economic Review 76 (4), 604–620.

Lauritsen, J., 1990. Poison by Prescription: The AZT Story. Asklepios Press, New York.

Ludwig, J., Duncan, G., Gennetian, L., Katz, L., Kessler, R., Kling, J., Sanbonmatsu, L., 2012. Neighborhood effects on the long term well-being of low-income adults. Science 337 (6101), 1505–1510.

Mobarak, A.M., Rosenzweig, M., 2014. Risk, insurance and wages in general equilibrium. NBER Working Paper 19811.

Newhouse, J., the Insurance Experiment Group, 1996. Free for All? Lessons from the RAND Health Insurance Experiment. Harvard University Press, Cambridge.

Neyman, J., 1923. On the application of probability theory to agricultural experiments. Essays on principles. Section 9. Statistical Science 5 (4), 465–472.

Orcutt, G., Orcutt, A., 1968. Incentive and disincentive experimentation for income maintenance policy purposes. American Economic Review 58, 754–772.

Parker, S., Rubalcava, L., Teruel, G., 2008. Evaluating conditional schooling and health programs. In: Schultz, T.P., Strauss, J. (Eds.), Handbook of Development Economics, vol. 4, pp. 3964–4035 (Chapter 62).

Peirce, C.S., Jastrow, J., 1885. On small differences in sensation. Memoirs of the National Academy of Sciences 3, 73–83.

Phillips, J., Simmons, R., Koenig, M., Chakraborty, J., 1988. Determinants of reproductive change in a traditional society: evidence from Matlab, Bangladesh. Studies in Family Planning 19 (6), 313–334.

Phillips, J., Bawah, A., Binka, F., 2006. Accelerating reproductive and child health programme impact with community-based services: the Navrongo experiment in Ghana. Bulletin of the World Health Organization 84, 949–955.

Rahman, O., Menken, J., Foster, A., Peterson, C.E., Khan, M.N., Kuhn, R., Gertler, P., 1999. The 1996 Matlab Health and Socioeconomic Survey: Overview and User's Guide. RAND, Santa Monica.

Schultz, T.P., 2004. School subsidies for the poor: evaluating the Mexican progresa poverty program. Journal of Development Economics 74 (1), 199–250.

Schultz, T.P., Joshi, S., 2013. Family planning and women's and children's health: consequences of an outreach program in Matlab, Bangladesh. Demography 50 (1), 149–180.

Stigler, S., 1992. A historical view of statistical concepts in psychology and education research. American Journal of Education 101 (1), 60–70.

Thomas, D., Frankenberg, E., Friedman, J., Habicht, J.P., Ingwersen, N., McKelvey, C., Hakimi, M., Jaswadi, Pelto G., Sikoki, B., Seeman, T., Smith, J.P., Sumantri, C., Suriastini, W., Wilopo, S., 2014. Causal Effect of Health on Labor Market Outcomes: Experimental Evidence, Mimeo.

Todd, P., Wolpin, K., 2006. Assessing the impact of a school subsidy program in Mexico: using a social experiment to validate a dynamic behavioral model of child schooling and fertility. American Economic Review 96 (5), 1385–1417.

Relevant Websites

https://www.socialscienceregistry.org/ – AEA RCT Registry.
http://www.povertyactionlab.org/ – Abdul Latif Jameel Poverty Action Lab.
http://www.cochrane.org/ – The Cochrane Collaboration.

Experimentation in Psychology, History of

GCG Trudy Dehue, University of Groningen, Groningen, The Netherlands

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 8, pp. 5115–5120, © 2001, Elsevier Ltd.

Abstract

Standardized experimentation constitutes the core of psychology's scientific identity. Since the 1940s, the ideal experiment has been understood to be an experiment comparing randomly composed experimental and control groups. This article discusses the historical origination of this 'random group design.' Whereas psychology textbooks generally attribute its origination to natural science, historical research suggests that comparing randomly composed groups is firmly rooted in psychology's *own* experimental traditions. Moreover, these traditions have changed together with changing social relations. The article discusses the 'psychophysical' research practices of nineteenth-century scholars trying to discover the given laws of human perception, and elucidates how these practices were adopted and adapted by twentieth-century educational psychologists looking for technically useful data. It also discusses the random group design's further dissemination and its current status as mostly a guiding model rather than a viable strategy.

From the very beginning of their studies, psychology students in Western countries learn that the discipline specializes in quantitative research rather than 'armchair philosophy.' Methodological and statistical textbooks teach how to measure individual or group characteristics such as intelligence, neuroticism, or attitudes. Moreover, introductory and advanced textbooks spell out rules of experimentation for tracing the causes of the psychological characteristics that are measured or evaluating attempts at improving them. Standardized experimentation in particular constitutes the core of psychology's scientific identity.

The typical psychological experiment compares one or more treated experimental groups to an untreated control group. As natural groups may vary in many more respects than only the hypothesized cause or treatment to be tested, experimenters must compose their own groups equal in all respects but the factor concerned. The ideal way to create comparable groups is to assign subjects to groups on the basis of chance and thus cancel out unwanted between-group differences. From the 1940s in America and from the 1960s in Europe, generations of psychologists have been taught that the *exemplary* experiment is an experiment in which randomly composed experimental and control groups are compared. Statistical inference testing is employed for establishing the significance of any differences in the results.

The Origins of the Exemplary Design

Psychology textbooks typically discuss the experiment comparing randomly composed experimental and control groups (from now on 'the random group experiment') as being 'the' scientific experiment. They implicitly suggest or explicitly state that this form of experiment was borrowed directly from the natural sciences (more on the definition of an experiment in psychology textbooks can be found in Winston and Blais, 1996). Yet, as has been demonstrated by researchers of science, in natural sciences such as physics and chemistry there are various ways of experimenting, and these are mostly learned during apprenticeship and via exemplars rather than through isolated methodology books and courses (Kuhn 1962). Moreover, comparing random groups is not a key experimental practice in such sciences (Hacking, 1983; Galison, 1997).

Historical research suggests that comparing randomly composed groups is firmly rooted in psychology's *own* experimental traditions, traditions which have changed together with changing social relations. The discipline's present-day definition of what constitutes an ideal experiment was brought about bit-by-bit when methodological practices from nineteenth-century 'psychophysical' experimenters trying to discover the given laws of human perception were adopted and adapted by early twentieth-century educational psychologists looking for technically useful data. These educational psychologists then passed the random group design on to other branches of psychology and social sciences (Danziger, 1990; Winston and Blais, 1996; Dehue, 1997).

Experimentation in Nineteenth-Century Psychophysics

Psychological experimentation began as an endeavor of mid-nineteenth-century gentlemen-scholars to discover the given design of human nature. The word 'experiment' could refer to all kinds of research other than naturalistic observation. Experimentation involved manipulating human sensation and perception within artificial situations. An important founder of psychological experimentation was the German scholar Gustav Fechner, who during the years 1845 to 1849 lifted his arms with a weight in each hand over 67 000 times. After each lifting, Fechner meticulously recorded whether he had been able to distinguish the heavier weight from the lighter. In this way he established the 'just noticeable difference' or 'threshold of perception' as well as the mathematical relationship between the weights as measured and the weights as perceived.

In Fechner's time, the claim to measure anything as subjective as human sensation was novel and impressive. In the second half of the nineteenth century, innumerable 'threshold experiments' were conducted in biology, physiology, and

particularly the newly established psychological laboratories at European and American universities. Countless extra weights were lifted, and Fechner's methods of 'psychophysical' experimentation were employed for testing other senses as well. A popular variation was touching the skin with two pins at specific distances and trying to establish the minimum distance that can still be perceived (see **Figure 1**). The different sensitivities of foreheads, chins, shoulders, arms, fingers, hips, bellies, legs, knees, and ankles were discussed. Neophytes would complain about sensitive and scaly skins and apologetically argue that some 6000 pricks should be enough to draw a conclusion (Dehue, 1997).

Like Fechner, the first psychophysical researchers acted as their own experimental subjects. Soon, however, the idea arose that the stimuli should be applied in random order, and that random control trials were needed. Researchers in Germany and subsequently the USA argued that the perception of distances might be biased by expectations on the next distance to come. Such expectations might be induced by awareness of the actual distances in previous trials or inadvertent regularities in previous series of trials. Moreover, it was argued, the sheer knowledge that there are always two pins might arouse the idea of perceiving a distance even if the actual distance is below the threshold of perception. For these reasons, it was decided that it takes *two* to experiment. Assistants were employed to apply the stimuli in random order as well as randomly carry out control trials in which the skin was touched at one point only without knowledge of the 'observers' (later on the individuals providing the stimuli would be called the 'experimenters' and the observers became the 'subjects,' see Danziger, 1990 on the social meaning of such terminological changes). These were the first experiments employing randomization as a methodological means. In the 1880s, the American scholars Charles S. Peirce and Joseph Jastrow, moreover, introduced the first *artificial* randomizers. Distrusting the assistants' ability to establish truly random orders, they used playing cards or dice (for a more extensive discussion and references, see Dehue, 1997).

If the stimuli are indicated as 'S,' the responses as 'R,' and the randomly applied control trials (in which the actual double prick was withheld) as '*noS*,' then the design of these experiments was something like S-R-*noS*-R-S-R-S-R-*noS*-R-S-R-*noS*-R-S-R-S-R (etc.). In later educational experimentation, the randomly inserted control trials (*noS*) were no longer conducted with the same subjects as the experimental trials (S) but with randomly composed control *groups*. This change implied the constitution of the random group design. In order to explain why that happened, some extra information should first be given on the methods of early psychophysics as well as social changes taking place outside the laboratories.

The 'Transfer of Training' Phenomenon

Considerable concern about individual expectations and mistaken assessments became the hallmark of psychophysics. For the sake of replicability, utmost precision had to be exercised. The initiators of particular experiments extensively described the required laboratory setting, the instruments to be used, the handling of the instruments, as well as the required behaviors of assistants and observers both during the trials and in their spare time (Morawski, 1988; Benschop and Draaisma, 2000).

The researchers of just noticeable differences, however, were confronted with a rather obvious extra source of bias. After many trials, the *skill* of the experimental subjects improved, and this artificially lowered the threshold of perception. Switching to the other arm or leg did not solve the problem because the undesirable training seemed to transfer automatically to other parts of the body. The transfer phenomenon thwarted the experiments but also greatly intrigued the experimenters. Many of them became so fascinated by it that from a source of bias 'transfer of training' became a research topic of its own. At the end of the century, experiments were being conducted in American and European laboratories to establish the effects of training in all kinds of sensory, motor, and cognitive skills. Subjects were not only pinpricked. They were also asked to put needles through small holes with one hand and then checked for improvement with the other hand, or they were trained to tap their right big toe and tested for their ability with their left big toe (Bray, 1928).

One has a mental picture of distinguished-looking men in suits sitting on wooden laboratory chairs and fiddling with needles or concentrating on their bare feet. But most importantly for the historical origination of the random group design, in these transfer of training studies the design of the threshold experiments had to be adapted. Once the stimuli had

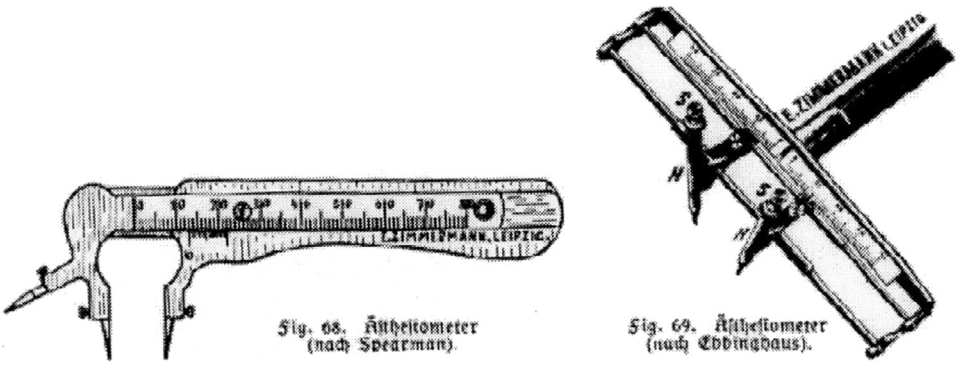

Figure 1 'Aesthesiometers' for pinpricks experiments. The one with three pins was especially designed for experiments with control trials (from Schulze, R., 1913. Aus der Werkstatt der experimentellen Psychologie und Pädagogik [From the workshop of experimental psychology and pedagogy]. Vogtländer Verlag, Leipzig, Germany.)

become trainings the original design of S-R-*noS*-R-S-R-S-R-*noS*-R-S-R-*noS*-R-S-R-S-R (etc.) was no longer feasible. Trainings cannot vary randomly as they cannot be haphazardly undone. Thus the standard design was limited to a particular 'cutout': (*noS*-) R-S-R. First a response (now an ability) was tested *without* preceding stimulus (now a training), next the training was given, and subsequently the ability was tested again. The transfer of training researchers worked with experimental groups whose abilities were compared before and after training.

Experimentation in Twentieth-Century Educational Psychology

While psychophysical experimenters were absorbed in their painstaking experiments, the world outside the laboratories was rapidly changing. In early-twentieth-century Western countries, the common ideology of *laissez-faire* politics was widely replaced by that of planned social progress. Social amelioration gradually turned from a local charity matter into a concern of centralized government. The concomitant emergence of bureaucratic administration was to generate significant changes in the kind of knowledge about human functioning and human relations that was produced.

Whereas the authority of nineteenth-century local officials depended on their personal trustworthiness, the newly appointed administrators had to justify their decisions in terms of impersonal data and overt standards. Moreover, progress had to come about as economically as possible. The expanding class of bureaucratic officials turned to professional social researchers for knowledge that could serve the needs of administrative standardization and efficiency. With the shift from aristocratic to bureaucratic domination the former gentleman-scholar trying to discover the given laws of nature was increasingly replaced by the administrative researcher looking for technically useful data (Danziger, 1990; Ross, 1991; Porter, 1995).

As far as psychology was concerned, the change was first brought about by psychophysical researchers taking an interest in educational administration. Around the turn of the twentieth century a debate blew up within American educational administrative circles on the utility of teaching subjects such as Latin and formal mathematics. School administrators wanted to abolish such subjects because of their apparent redundancy while proponents argued that 'formal discipline' strengthens general mental capacities. The issue reminded some American psychologists of the transfer of training phenomenon. Reasoning that formal discipline basically is a matter of transfer, they slightly adapted the former transfer experiments. They let subjects estimate weights or mark misspelled words, then trained them in estimating other weights or marking other misspelled words, and subsequently subjected them to the first tests again (*NoS*-R-S-R).

It may seem odd from a present-day perspective, but at the time a great deal of value was ascribed to the outcomes of such experiments. They seemed to provide an ideal basis for solid and impersonal decision-making. In the heat of the debate charges of experimental bias were soon made. In 1907, John E. Coover, a psychology student and former school principal, published an article on the issue (together with his teacher Frank Angell). The article argued that no valid causal conclusions can be drawn if the usual design of test-training-test (*noS*-R-S-R) was used (Coover and Angell, 1907). With this strategy, the authors emphasized, it is impossible to tell whether the outcomes are caused by the training or by any other factor. The first test, for instance, might serve as a training too.

In experiments done for Coover's 1905 master's thesis, and described in the 1907 article, the solution was used of comparing 'experimental reagents,' who had been given a training, with 'control reagents,' who had not. With Coover's design the usual sequence of (*noS*-) R-S-R was, so to speak, cut in two. The experimental group underwent the second part of S-R (training and test) and the control group only the first half of (*noS*-) R (a test without preceding training). In other words, the classical control *trial* in which the stimulus was withheld (*noS*), was now applied to special control *subjects* (more on the early introduction of the control group is given in Dehue, 2000)

Creating Equal Groups

In the 1910s, experimental educational research began to thrive. Educational experimenters acquired contracts from school administrators and left their laboratories for experimentation in real schools. 'Treatments' to be assessed now included methods of memorizing, ways of teaching, educational measures such as punishing and praising, and all kinds of classroom circumstances. The idea of using a control group also spread. In the 1916 volume of the *Journal of Educational Psychology*, 14% of the articles reported the use of control groups (Danziger, 1990: pp. 113–15).

In the schools, however, life was much messier than in the former locations of psychophysical experimentation. It was much harder to protect experimental results against charges of arbitrariness. Methodological scrutiny became ever more important. It soon turned out that a price had to be paid for comparing the results of groups rather than comparing subjects to themselves. Contaminating factors between the pretest and the post-test were replaced by invalidating differences between the groups.

At first it became customary to handle the new problem by hypothesizing which factors could bias the results, then subjecting subjects to preliminary tests for these factors, and next forming groups with equal test results. For instance, for testing the efficacy of a new way of teaching, equal groups were created by first measuring the pupils' intelligence and next composing equal groups as to mental capacity. With this 'matching' procedure, however, the choice of suspect factors depended very much on the researchers' personal imagination. Possibly the sex of the children also makes a difference, or their social background, or any other factor. Moreover, matching was very time and money consuming. For the combined reasons of saving money and cancelling out subjectivity – the twin principles of administrative rationality – experimental psychologists again invoked the methodological use of chance.

In 1923, the American psychologist William McCall published a manual entitled *How to Experiment in Education*. In his introduction, McCall estimated that increased efficiency in education could save a full year of teaching per person, and worked out that psychological experimentation would save '$134 680 000 000 000 for the next 100 generations of Americans.' He then he advanced a way of making his profession even more profitable to society, which was to compose the

groups by chance. McCall did not take the task of randomization lightly. He rejected the procedure of writing numbers on pieces of paper, because papers with larger numbers contain more ink and are therefore likely to sink further to the bottom of a container. He nevertheless maintained that 'any device which will make the selection truly random is satisfactory' (McCall, 1923: pp. 41–2).

Randomization, formerly used for creating random orders of stimuli, was now advanced as a cheap way of discarding unwanted differences between experimental groups. The random group design was first worked out in US educational psychology in the 1920s (on its later introduction in medicine, see Marks, 1997).

A Guiding Rather Than Practicable Principle

While most psychology textbooks ascribe the random group design to natural science, some credit the British biometrician Ronald Fisher with it. Fisher developed the design in the 1930s for agricultural research. His methodological handbook, *The Design of Experiments* (1935), was widely received indeed. However, psychologists did not learn to use randomly composed groups from Fisher. From the 1920s onwards, educational experimenters had struggled with the complexity of data resulting from experiments simultaneously testing various experimental groups and comparing them to one control group. Fisher offered statistical techniques for analyzing such complex between-groups variance and assessing the significance of the results (on Fisher's influence on psychology, see Gigerenzer and Murray, 1987). His warnings that, for *statistical* reasons, these techniques could only be validly used if the groups were randomly composed added extra authority to psychologists already promoting the employment of random groups. In the 1940s, a range of handbooks on research design and statistics, mostly written by educational psychologists, passed on Fisher's instructions to other branches of psychology. One of these methodological professionals was Quinn McNemar who would later observe that Fisher's ways quickly 'became epidemic just as though psychologists had never previously planned an experiment' (McNemar, 1960: p. 297).

At first, the combination of the random group design and significance testing seemed to provide an unquestionable experimental paradigm. Yet, in the context of experiments with humans, comparing random groups became an ideal mostly in the sense of a guiding principle rather than a viable strategy. Moral, legal, or practical reasons often prohibit people being given a treatment purely on the basis of chance. For example, the educational effects of prizes cannot be investigated by granting them at random, and differences in ethnic backgrounds among schools cannot be cancelled out by creating new classes on the basis of chance. Moreover, psychologists became involved with administrative research projects outside their traditional domain. From the 1930s, they were enlisted by other social scientists to measure the educational effects of radio talks and movies on issues such as voting, immigration, bootlegging, or war. In such contexts, it became all the more difficult to let the assignment of individuals to treatments depend purely on chance.

Renowned experiments by the US Army constitute a telling example of the unfeasibility of random assignment even with adult subjects living under command in a similar way to schoolchildren. During the Second World War, the Army's 'Morale Division' was charged with improving the soldiers' spirit. A film series entitled *Why We Fight* was made, and the Division's experimental section, chiefly staffed with psychologists, was asked to investigate the films' efficacy. The plan to evaluate the efficacy of these movies with the help of randomly composed groups did not work, as most soldiers became suspicious when some of them were summoned to see a movie while others were not. Random assignment thus appeared to elicit an unexpected new kind of bias. Although upon pretesting the existing army units turned out to differ in many consequential respects, the units had to be used as experimental and control groups (Hovland, 1949; for a history see Buck, 1985).

After the war, the army experimenters were ensured of jobs at the best US universities (Herman, 1995: p. 137). They widely propagated the standards of proper experimentation. Methodological treatises began to appear still advancing the random group design as the paragon form of experiment, but simultaneously indicating that unlike subjects in agricultural experimentation, people often cannot be assigned haphazardly to experimental conditions. Alternative designs and accompanying statistical techniques were developed.

Among the first methodologists to write in this vein was the young psychologist Donald Campbell, who also started his career in the army as an attitude and propaganda researcher. With the assistance of the statistician Julian Stanley, Campbell published a chapter entitled 'Experimental and Quasi-Experimental Designs for Research on Teaching.' In 1966 the chapter was reprinted as a separate 83-page booklet with the briefer but more-embracing title of *Experimental and Quasi-Experimental Designs for Research* (Campbell and Stanley, 1966).

The book was the starting point of Campbell's career as a renowned methodologist of social science. In the decades to come, innumerable publications and thick textbooks on 'true' and 'quasi' experimental designs by Campbell and many other authors in America and Europe were to follow. These works testify to the lasting high regard but limited applicability of random group experiments as well as their sociotechnical inspiration. As one of them summarized these three aspects in a striking comparison: 'Most people after all do not drive around in Rolls Royces or Cadillacs, even if they represent some peak of automobile engineering perfection' (Bulmer, 1986: p. 170).

See also: Experiment in Science and Technology Studies; Experimental Design: Overview; Experimental Laboratories: Psychological; Experimental Laboratories: Social and Economic; Experimenter and Subject Artifacts: Methodology.

Bibliography

Benschop, R., Draaisma, D., 2000. In pursuit of precision. The calibration of minds and machines in late nineteenth century psychology. Annals of Science 57 (1), 1–25.

Bray, C.W., 1928. Transfer of Learning. Journal of Experimental Psychology 11, 443–467.

Buck, P., 1985. Adjusting to military life: the social sciences go to war, 1941–1950. In: Smith, M.R. (Ed.), Military Enterprise and Technological Change: Perspectives on the American Experience. MIT Press, Cambridge, MA.
Bulmer, M., 1986. Evaluation research and social experimentation. In: Bulmer, M. (Ed.), Social Science and Social Policy. Allen and Unwin, London.
Campbell, D.T., Stanley, J.C., 1966. Experimental and Quasi-experimental Designs for Research. R McNally, Chicago.
Coover, J.E., Angell, F., 1907. General practice effect of special exercise. American Journal of Psychology 18, 328–340.
Danziger, K., 1990. Constructing the Subject: Historical Origins of Psychological Research. Cambridge University Press, New York.
Dehue, T., 1997. Deception, efficiency and random groups. Psychology and the gradual origination of the random group design. Isis 88, 653–673.
Dehue, T., 2000. From deception-trials to control-reagents. The introduction of the control group about a century ago. American Psychologist 55 (2), 264–268.
Fisher, R.A., 1935. The Design of Experiments. Oliver & Boyde, Edinburgh, UK.
Galison, P.L., 1997. Image and Logic: A Material Culture of Microphysics. University of Chicago Press, Chicago.
Gigerenzer, G., Murray, D.J., 1987. Cognition as Intuitive Statistics. Lawrence Erlbaum, Hillsdale, NJ.
Hacking, I., 1983. Representing and Intervening. Introductory Topics in the Philosophy of Natural Science. Cambridge University Press, New York.
Herman, E., 1995. The Romance of American Psychology: Political Culture in the Age of Experts. University of California Press, Berkeley, CA.
Hovland, C.I., Lumdsdaine, A.A., Sheffield, F.E., 1949. Experiments on Mass Communication. Princeton University Press, Princeton, NJ.
Kuhn, T.S., 1962. The Structure of Scientific Revolutions. Chicago University Press, Chicago.
Marks, H.M., 1997. The progress of experiment. Science and therapeutic reform in the United States, 1900–1990. Cambridge University Press, New York.
McCall, W.A., 1923. How to Experiment in Education. Mcmillan, New York.
McNemar, Q., 1960. At random: sense and nonsense. American Psychologist 15, 295–300.
Morawski, J.G. (Ed.), 1988. The Rise of Experimentation in American Psychology. Yale University Press, New Haven, CT.
Porter, T.M., 1995. Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton University Press, Princeton, NJ.
Ross, D., 1991. The Origins of American Social Science. Cambridge University Press, New York.
Schulze, R., 1913. Aus der Werkstatt der experimentellen Psychologie und Pädagogik [From the workshop of experimental psychology and pedagogy]. Vogtländer Verlag, Leipzig, Germany.
Winston, A.S., Blais, D.J., 1996. What counts as an experiment? A trans-disciplinary analysis of textbooks, 1930–1970. American Journal of Psychology 109, 599–616.

Experimenter and Subject Artifacts: Methodology

Ralph L Rosnow, Temple University, Philadelphia, PA, USA
David B Strohmetz, Monmouth University, West Long Branch, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R.L. Rosnow, volume 8, pp. 5120–5124, © 2001, Elsevier Ltd.

Abstract

'Experimenter artifacts' and 'subject artifacts' are uncontrolled systematic errors that can arise within the experimental situation. Experimenter artifacts can be classified as being noninteractional or interactional. Subject artifacts can result from demand characteristics within the experimental situation and from the motivations of the subjects themselves. While there are strategies for minimizing potential artifacts, experimenters must be mindful of ethical considerations.

'Experimenter artifacts' and 'subject artifacts' refer to systematic errors that can be attributed to uncontrolled experimenter and subject variables, respectively. The term 'experimenter' is used broadly in this context to refer not just to scientists who perform laboratory experiments, but also to researchers working in any area of empirical investigation involving humans. Discussed here are the definition and history of artifacts in the social and behavioral sciences, the nature and control of experimenter and subject artifacts, a meditational model for conceptualizing how artifacts can be introduced into an investigation, and the delicate balance between ethical accountability and the avoidance of artifacts in research with human subjects.

The Nature of Artifacts in Scientific Inquiry

The sociologist H. H. Hyman wisely cautioned researchers not to equate ignorance of error with lack of error. Because all scientific investigation is subject to error, it is prudent to be aware of this, to expose the sources in an attempt to control them, and to estimate the magnitude of such errors when they occur (Hyman, 1954). In the social and behavioral sciences, researchers study the conduct of participants (or 'subjects') to further our understanding of the complexities of human behavior. Some of the complexity of subjects' behavior, however, results from specifiable but uncontrolled extraneous variables, often social in nature. The effects of such variables are conceptualized as a set of systematic errors (or 'artifacts') to be isolated, quantified, taken into account when interpreting the results, and, when possible, eliminated. Thus, artifacts not only are inconsequential effects in the research design, but also may jeopardize the validity of the results. To be sure, today's artifact could become tomorrow's independent variable, as conditions discounted as 'nuisance' variables at one time may be redefined as substantive variables in their own right in some future study. For example, the placebo effect in medical research, where an inert substance or nonspecific treatment produces a 'healing' effect, can be defined as an artifact but can also be exploited as a potential 'treatment' for chronic pain disorder (Turner et al., 1994).

Experimenter and subject artifacts, because they are conceptualized as systematic errors, are further distinguished from random errors. To explain the difference, Rosnow and Rosenthal (1997) envisioned an imaginary grocer who weighed the same bunch of grapes a number of times in a row. In an ideal world, no matter how precise the measurements, they would always come out a bit differently. Some measurements would slightly overestimate the correct weight of the grapes, while others would slightly underestimate the correct weight. Because the various overestimates and underestimates occur haphazardly (rather than in a patterned fashion), we can assume that the random fluctuations (or random errors) will cancel one another over the long run. Thus, we can simply average the measurements to estimate the correct weight of the grapes. But suppose the grocer always weighed the grapes with his thumb on the scale, thereby introducing a systematic error by always tacking on a few extra ounces to the true weight. This is essentially the nature of artifacts, inasmuch as they tend to push measurements in the same direction, up or down, systematically causing the 'average value' be consistently too large or too small. The problem for researchers is to determine the direction in which a particular artifact might push the results, to try to improve the accuracy of the measurements, and, when possible, to eliminate the source of systematic error and therefore the artifact.

The discovery of artifacts in the social and behavioral sciences has a long history, but it was not until the 1950s and 1960s that investigation of experimenter and subject artifacts began in earnest (see Rosnow, 1981, for a detailed discussion of the historical context and wider implications of this work). Claims about the existence of experimenter and subject artifacts were sometimes met with skepticism, but by the late 1960s there was ample evidence to show the magnitude of the artifact problem in social and behavioral research (see Rosenthal and Rosnow, 1969). Seminal investigations conducted by R. Rosenthal and M. T. Orne, working independently of one another, provided the initial impetus behind a flood of relevant work in the 1960s and 1970s. On the assumption that typical subjects often try to penetrate the nature of the study in which they are participating, Orne proposed the expression "demand characteristics of the experiment," which he derived from K. Lewin's concept of *Aufforderungscharaktere* to describe the 'demand value' of uncontrolled task-orienting cues in an experiment that exert an influence on subjects' ideas about its purpose. About the same time, Rosenthal showed that the

experimenter's tacit hypothesis or expectancy can sometimes become a self-fulfilling prophecy of the subjects' experimental responses. Well before this work appeared, however, there had been intriguing clues about the possible existence of artifacts caused by uncontrolled variables when sentient subjects are the objects of investigation.

Earlier Clues about Artifacts

An early indication about artifacts may be lurking in the investigative procedures used in animal research involving a horse named 'Clever Hans.' At the turn of the twentieth century, an eminent German psychologist, O. Pfungst (1911), carried out a six-week investigation of this remarkable horse owned by a mathematics teacher named von Osten. Hans, who responded by tapping his hoof according to a code taught to him by von Osten, was purported to be able to reason and answer questions put to him in German, including solving arithmetic problems. It was unlikely that von Osten had fraudulent intent because he readily allowed others (even in his absence) to question Hans, and also did not profit financially from the horse's talents. Thus, it was possible to rule out intentional cues from von Osten as the reason for the horse's cleverness. In a series of elegant experiments, Pfungst demonstrated that Hans's accuracy diminished when (1) Hans was fitted with blinders, so he could not see his questioners; (2) the distance between Hans and his questioners was increased; or (3) the questioner did not know the answer. These results implied that Hans's performance was due to something other than his capacity to reason. Pfungst found that Hans was responding to subtle cues given by his questioners, not just intentional cues but also unwitting movements and mannerisms. This classic research provided an object lesson in the susceptibility of behavior (even animal behavior) to unconscious suggestion.

A half century later, another influential development fostered the suspicion that human subjects behave in special ways when they know they are the 'subjects' of investigation. This proposition, called the 'Hawthorne effect,' grew out of a series of human factors experiments begun in the 1920s by a group of industrial researchers at the Hawthorne works of the Western Electric Company in Cicero, Illinois (Roethlisberger and Dickson, 1939). The experiments were intended to study how workers' productivity and job satisfaction might be affected by specific workplace conditions (light, temperature, rest periods, etc.). The conclusion reached, however, was that increases in productivity were due not to the manipulated experimental conditions, but to the uncontrolled effects of being singled out for participation and the adventitious cues intrinsic in the experimental situation. This research has subsequently been taken to task because of ambiguities and irregularities that make it difficult to interpret. It is also charged that secondary accounts of the research were tarnished by sweeping generalizations embroidered by overly zealous authors (for an interesting behavioral interpretation of previously unreported circumstances of the Hawthorne research, see, e.g., Gillespie, 1988; Parsons, 1974). Nonetheless, the 'Hawthorne effect' entered into the vocabulary of the social and behavioral sciences as implying a kind of placebo effect (see Sommer, 1968). As a consequence, researchers were advised to employ 'Hawthorne control groups,' which received special attention for a meaningless task. Presumably, by examining obtained differences in performances between the Hawthorne control group and a no-treatment control group, any Hawthorne effects should be detectable.

Another important development, which went largely unnoticed for many years, involved a conceptual advance. Rosenzweig (1933) published an insightful critique of uncontrolled aspects of psychology experiments. In particular, he identified three distinct sources of artifacts, which he theorized were associated with (1) the experimenter's 'observational attitude,' (2) the subjects' motivational states, and (3) the experimenter's demeanor when interacting with the subjects. Rosenzweig sketched procedures that he thought might obviate these 'errors,' such as the use of 'simple deceptions' to distract subjects from the true purpose of the experiment. He cautioned, however, that it might be unclear whether the experimenter or the subject was the 'true deceiver,' because subjects typically carry on a train of thought in which they attempt to guess the purpose of the experiment, and then model their behavior accordingly. Interestingly, early psychologists referred to their subjects as 'reagents,' implying that they resembled the ingredients that chemists mixed in clean test tubes to produce predictable reactions. Rosenzweig argued that the 'test tubes' (i.e., research settings) and 'reagents' (the subjects) of psychological researchers were sullied by the needs, anxieties, self-deceptions, and intentions of human beings who knew very well that their behavior was being scrutinized as part of a scientific investigation.

The Role of Experimenter Artifacts

Rosenthal (1966) divided experimenter artifacts into two general classes, called 'noninteractional' and 'interactional.' Noninteractional artifacts operate, so to speak, in the mind, in the eye, or in the hand of the scientist, but do not directly affect the subjects' responses. Within this class are three subtypes, known as observer, interpreter, and intentional effects. 'Observer effects' refer to overestimates or underestimates of results during the observation and recording phase of an experiment, which are controlled by the use of mechanical recording devices or by independent replications to discover if similar results are observed. 'Interpreter bias' refers to an error in the interpretation of data, which is ostensibly controlled by allowing other scientists to analyze and interpret the original data for themselves (see also MacCoun, 1998). 'Intentional effects' imply that the data were fabricated or fraudulently manipulated. Such distortions are devastating to science because they undermine the integrity of the literature on which the advancement of science depends; safeguarding against dishonesty is presumed to be the ethical responsibility of every scientist (cf. Carey, 2011).

Interactional artifacts, the second general class, are those due to extraneous experimenter-related variables that can directly influence the reactions of the research subjects. These include 'biosocial effects' (biases due to biological or social attributes of the researchers, such as the sex, age, ethnicity, and race of the researcher), 'psychosocial effects' (e.g., the

researcher's personality and comportment), 'situational effects' (biases attributable to uncontrolled aspects of the setting), and modeling effects (i.e., the example set by the investigator in the role of a 'scientist'). Replications across settings, subjects, and experimenters are used to isolate interactional artifacts, although procedural and sampling differences may not be easy to interpret. Moreover, scientists, like all humans, are susceptible to the biases imposed by the limitations of perception and cognition.

One additional interactional effect, well known in the social and behavioral sciences, was named the 'experimenter expectancy effect' by Rosenthal, its discoverer. It has spawned a large number of studies (Rosenthal and Rubin, 1978), with various applications in education, the law, and clinical practice (see Blanck, 1993). An 'experimenter expectancy bias' means that the experimenter's hypothesis leads unintentionally to behavior toward the subjects that, in turn, increases the likelihood that the hypothesis will be confirmed (reminiscent of the Clever Hans phenomenon). One solution to this problem is to keep experimenters unaware of ('blind' to) which subjects are to receive the experimental treatment and which the control treatment. If the experimenters do not know what treatment the subject has received, then they cannot accurately communicate expectancies about the nature of the particular treatment. The necessity of keeping experimenters blind is well recognized in randomized drug trials, for example. In fact, no drug trial is taken completely seriously unless it has followed elaborate 'double-blind procedures,' in which neither the subjects nor the experimenters know who is in the experimental group and who is in the control group.

The Role of Subject Artifacts

In focusing on subject artifacts, researchers have generally been interested in the role of motivations that guide the research subject's behavior. Orne (1962) described the role of the 'good subject,' that is, the participant who is particularly attentive to, and compliant with, demand characteristics. In one fascinating demonstration study, Orne asked subjects to add hundreds of thousands of two-digit numbers – which they kept doing until finally the experimenter gave up! Even when subjects were told to tear each worksheet into a minimum of 32 pieces before going on to the next, they persisted. Orne explained this behavior as the 'role enactment' of individuals who reason that, no matter how trivial and inane the assigned task seemed to them, it must have some important scientific purpose or they would not have been asked to participate in the first place. Thus, Orne argued, they complied with the demand characteristics of the experiment in order to 'further the cause of science.'

Not all research participants are willing to play the 'good subject' role, but Rosenthal and Rosnow (1975) reported that many volunteers for research participation are likely to conform to this expectation. Volunteers for research are also usually better educated, higher in social class status, more intelligent, more approval motivated, and more sociable than nonvolunteers. These and other demographics and personality characteristics may result in a 'volunteer subject bias' in that the results of studies using subjects who specifically volunteered may not be simply generalizable to the population at large. For example, standardizing the norms of an intelligence test on typical volunteer subjects may produce biased estimates of true values in a more general population. Rosenthal and Rosnow described some strategies for improving recruitment procedures and the generalizability of results, and also logical procedures for estimating the direction of volunteer subject bias.

Other subject motivations may be operating, however, including what Rosenberg called 'evaluation apprehension,' referring to an anxiety-toned state in which subjects worry that the experimenter plans to evaluate an aspect of their psychological competence. As a result, they develop their own hypotheses about how to win the experimenter's approval or avoid disapproval. Experiments in which evaluation apprehension appears likely to occur are those that contain an element of surprise or have an aura of mystery. The more transparent and salient the task-orienting cues and the more control the experimenter has in administering a positive or negative evaluation, the greater may be the bias ostensibly resulting from the participant's evaluation apprehension (see Rosenberg's discussion in Rosenthal and Rosnow, 1969). One solution to this problem is to ensure the confidentiality of subjects' responses, on the assumption that individual subjects will be less apprehensive and more forthcoming in their responses. In research on sensitive topics, such as studies of sexual behavior, self-reported data nevertheless may be suspected of measurement error and participation bias (for a detailed discussion, see Cantania et al., 1990).

Other subject artifacts have been studied in the field of assessment, including a variation on evaluation apprehension known as 'social desirability bias.' In this case, the respondent to a personality inventory answers items in ways that seem to elicit favorable evaluations. One way to assess this problem is by means of a questionnaire that was specifically designed to tease out this response bias (Crowne and Marlow, 1964). Various other subject biases in assessment have been identified, and particular strategies have been developed to isolate or avoid those problems, such as the development of 'lie scales' to detect prevarication.

Minimizing Potential Artifacts

In his discussion of the social psychology of the experiment, Strohmetz (2008) described a meditational model earlier proposed by Rosnow for conceptualizing how artifacts may be introduced into an investigation. The first step is the degree to which subjects are receptive to the presence of potential demand characteristics in the experimental situation. These demand characteristics could stem from the methodology itself or from the experimenter present in the situation. The second step concerns the motivation of subjects to act upon these task-orienting cues. Volunteer subjects as well as those experiencing evaluation apprehension may be particularly motivated to alter their behaviors in response to perceived demand characteristics. The last step involves the capability of subjects to alter their behavior in response to the demand characteristics. Allen (2004) provided empirical support for the receptivity and motivation steps in this model.

The utility of this model is that it suggests opportunities for experimenters to identify and therefore minimize the potential for artifacts that can threaten the internal validity of a study and its conclusions. For example, experimenters could reduce receptivity by using deception to divert subjects' attention away from potential task-orienting cues. Interactions between the experimenter and subjects can be limited to control this source of demand characteristics. Conducting studies in naturalistic settings and utilizing unobtrusive measures can reduce the likelihood that the subjects are aware that they are being observed, making them less receptive to the presence of demand characteristics. Reducing subjects' motivation to act upon potential task-orienting cues can also reduce the potential for research artifacts. To minimize a volunteer bias, experimenters can implement inclusive strategies for recruiting research participants. Employing nonthreatening procedures can help reduce evaluation apprehension and social desirability concerns.

While this model is a useful heuristic when designing an empirical study, there is no assurance that artifacts have not been introduced into one's study. Orne (see his chapter in Rosenthal and Rosnow, 1969) suggested a strategy to ferret out potential demand characteristics by using 'quasi-control subjects' in addition to the usual control groups. Quasi-controls are individuals who are asked to step out of the traditional subject role; they reflect on the experiment in which they participated and suggest how their own experimental behavior was influenced by extraneous variables as opposed to only the controlled independent variable. It is now argued, however, that demand characteristics also exert influence on behavior beyond the experimental setting and should be recognized as a pervasive influence on all human interaction (for the role of demand characteristics in the context of psychotherapy, see Orne and Bauer-Manley, 1991).

Artifacts and Ethics in Collision

Until World War II, the popular conception of science as an 'endless frontier' unencumbered by moral constraints remained largely unshaken. Currently, however, investigators in even the most benign areas of the social and behavioral sciences are required to comply with stringent ethical guidelines laid done in professional and statutory codes overseen by review boards (see Rosnow et al., 1993). This situation can at times lead to knotty moral dilemmas as researchers try to strike a balance between ethical accountability and the technical demands of sound scientific practice (see Blanck et al., 1992; Kimmel, 1996). The effort to control for artifacts may, for example, come into collision with ethical imperatives such as informed consent and full disclosure. By way of illustration, in clinical trials using double-blind designs in which the control group is given a placebo, complying with full disclosure implies that the potential participants will be told they could be in a condition that will be deprived of possible benefits of the experimental treatment. Once they become aware of this fact, however, some subjects might surreptitiously share doses in the hope that this will increase the likelihood that each of them will receive at least some access to the experimental treatment. Such a situation could render any subsequent comparison of the experimental group and the 'untreated' control group meaningless. A strategy that can be used to avoid this problem, if there is an alternative treatment that is better than a placebo, is to provide the control group with this treatment rather than a placebo. This approach avoids the ethical bind of depriving control patients of any effective treatment, keeps the design double-blind, and focuses the research on other comparisons of practical and scientific importance (e.g., efficacy, relative cost, and side effects). As moral dilemmas proliferate in the sciences, the challenge faced by researchers is to expand our existing knowledge with valid data and yet abide by an evolving social contract that is responsive to current ethical sensitivities (for further discussion, see Rosnow, 1997).

See also: Data Collection: Interviewing; Internal Validity; Participant Observation.

Bibliography

Allen, C.T., 2004. A theory-based approach for improving demand artifact assessment in advertising experiments. Journal of Advertising 33, 63–73.

Blanck, P.D. (Ed.), 1993. Interpersonal Expectations: Theory, Research, and Applications. Cambridge University Press, New York.

Blanck, P.D., Bellack, A.S., Rosnow, R.L., Rotheram-Borus, M.J., Schooler, N.R., 1992. Scientific rewards and conflicts of ethical choices in human subjects research. American Psychologist 47, 959–965.

Cantania, J.A., Gibson, D.R., Chitwood, D.D., Coates, T.J., 1990. Methodological problems in AIDS behavioral research: influences on measurement error and participation bias in studies of sexual behavior. Psychological Bulletin 108, 339–362.

Carey, B., 2011, November 3. Fraud Case Seen as Red Flag for Psychology Research. New York Times. A3.

Crowne, D.P., Marlow, D., 1964. The Approval Motive. Wiley, New York.

Gillespie, R., 1988. The Hawthorne experiments and the politics of experimentation. In: Morawski, J.G. (Ed.), The Rise of Experimentation in American Psychology. Yale University Press, New Haven, CT, pp. 114–137.

Hyman, H.H., 1954. Interviewing in Social Research. University of Chicago Press, Chicago, IL.

Kimmel, A.J., 1996. Ethical Issues in Behavioral Research: A Survey. Blackwell, Oxford, UK.

MacCoun, R.J., 1998. Biases in the interpretation and use of research results. Annual Review of Psychology 49, 259–287.

Orne, M.T., 1962. On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implication. American Psychologist 17, 776–783.

Orne, M.T., Bauer-Manley, N.K., 1991. Disorders of self: myths, metaphors, and demand characteristics of treatment. In: Strauss, J., Goethals, G.R. (Eds.), The Self: Interdisciplinary Approaches. Springer Verlag, New York, pp. 93–106.

Parsons, H.M., 1974. What happened at Hawthorne? Science 183, 922–932.

Pfungst, O., 1911. Clever Hans (The Horse of Mr. von Osten). H. Holt, New York (reissued 1965 by Holt, Rinehart and Winston, New York).

Roethlisberger, F.J., Dickson, W.J., 1939. Management and the Worker. Harvard University Press, Cambridge, MA.

Rosenthal, R., 1966. Experimenter Effects in Behavioral Research. Appleton-Century-Crofts, New York.

Rosenthal, R., Rosnow, R.L. (Eds.), 1969. Artifact in Behavioral Research. Academic Press, New York.

Rosenthal, R., Rosnow, R.L., 1975. The Volunteer Subject. Wiley, New York.

Rosenthal, R., Rubin, D.B., 1978. Interpersonal expectancy effects: the first 345 studies. Behavioral and Brain Sciences 3, 377–386.

Rosenzweig, S., 1933. The experimental situation as a psychological problem. Psychological Bulletin 68, 337–354.
Rosnow, R.L., 1981. Paradigms in Transition: The Methodology of Social Inquiry. Oxford University Press, New York.
Rosnow, R.L., 1997. Hedgehogs, foxes, and the evolving social contract in psychological science: ethical challenges and methodological opportunities. Psychological Methods 2, 345–356.
Rosnow, R.L., Rosenthal, R., 1997. People Studying People: Artifacts and Ethics in Behavioral Research. W. H. Freeman, New York.
Rosnow, R.L., Rotheram-Borus, M.J., Ceci, S.J., Blanck, P.D., Koocher, G.P., 1993. The institutional review board as a mirror of scientific and ethical standards. American Psychologist 48, 821–886.
Sommer, R., 1968. Hawthorne dogma. Psychological Bulletin 70, 592–595.
Strohmetz, D.B., 2008. Research artifacts and the social psychology of psychological experiments. Social and Personality Psychology Compass 2, 861–877.
Turner, J.A., Deyo, R.A., Loeser, J.D., Von Korff, M., Fordyce, W.E., 1994. The importance of placebo effects in pain treatment and research. Journal of the American Medical Association 271, 1609–1614.

Experiment in Science and Technology Studies

Peter Dear, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Experiment' is a category of study that has long stood at the heart of much work in science and technology studies. Its meaning, however, has changed over time, depending upon the overall analytical setting within which it has been invoked and provided with its significance for the understanding of science. This article follows the development of ideas about 'experiment' from its role in the mainstream Anglo-American philosophy of science of the 1950s and 1960s through to its reconsideration in the sociologically oriented work of the 1970s and 1980s, until its new, nuanced place in the understandings of science as practice and process that were developed during the late 1980s and throughout the 1990s.

Experiment was a major concern of science and technology studies (STS) during the last quarter of the twentieth century. This is because certain views of experimentation as a privileged way of making knowledge about nature were important focuses of dissent during the specialty's early career. Various versions of the so-called 'received view' of science prevalent in Anglo-American philosophy of science in the period following World War II, with their roots in logical empiricism, held up experiments as the crucial empirical elements of scientific knowledge. Irrespective of whether the logic employed in their modeling of scientific reasoning was falsificationist or conformationist, versions of the 'received view' represented theories as conceptual structures that coordinate with the world through independent experimental tests of their predictions. This article will survey some of the major developments that have emerged to modify those older views.

Development of New Approaches to Experiment: Challenges to the 'Received View'

In the 1950s and early 1960s, one of the central assumptions of the 'received view' was seriously challenged. The presumed independence of experimental and observational data from contamination by the very theories that the data were intended to judge was challenged by W.V.O. Quine (Quine, 1953), Norwood Russell Hanson (Hanson, 1958), and Thomas S. Kuhn (Kuhn, 1970 [1962]). Quine showed, using an argument similar to one used by Pierre Duhem (1954), early in the twentieth century, that empirical data always underdetermine the theories that can rest upon them. In Duhem's hands, the argument rested on a position of 'conventionalism,' whereby the relationship between items of discrete empirical data and the theories employing those data was seen to be underdetermined; in particular, Duhem argued that there could be no such thing as a 'crucial experiment' that could decide definitively between two alternative theories. A result inconsistent with one of the theories need not be attributed to an inadequacy in the target theory itself, but might instead be attributed to error in a subsidiary theory (such as one relating to observational instruments) that had not initially been in question. Possibilities of that sort amounted to decisions, according to Duhem, and were consistent with a conventionalist view of scientific theories, whereby pure convention determined which in a network of theories would be preserved and which amended so as to fit experimental results.

Hanson and Kuhn went further, with their arguments in favor of the 'theory-ladenness of observation.' They held that a pure observation language was in principle impossible, and that, accordingly, the results of experiments and observations of nature always implicated theoretical concepts in their very formulation. This position blurred the distinction between empirical data on the one hand and the theories that are justified by those data on the other. Both Kuhn and Hanson likened empirical data to the psychologist's gestalt experiments, wherein, for example, the 'same' image can sometimes be seen as a duck and sometimes as a rabbit; the difference, they claimed, was that in science there is no possibility of resting one's description on an account of the underlying set of lines. Instead, the scientist sees a duck or a rabbit, with no further reduction to a common observational base being possible. All empirical data are already interpreted; that is, seen in a particular way (cf Fleck, 1979 [1935], an acknowledged source for Kuhn).

Sociological and Historical Exploitations of New Theoretical Arguments

These challenges to the 'received view' of science provoked much debate in the 1960s and into the 1970s. In its wake, however, there arose attempts to exploit ideas of such as those of Hanson and Kuhn in the service of sociological projects aimed at making sense of scientific knowledge as a social product. In the early days of the 'sociology of scientific knowledge' (SSK), the Duhem–Quine 'underdetermination' thesis became a standard in-principle justification for sociological studies of the acceptance of knowledge claims by scientific communities (e.g., Mulkay, 1979; Knorr-Cetina and Mulkay, 1983). This was so because it was argued that the inadequacy of empirical evidence to necessitate belief in any particular theory in science left room for social explanations for which belief among the manifold logical possibilities was actually chosen. At the same time, however, a more fundamental theoretical challenge to the self-sufficiency of 'received view' accounts of the role of experiment in science was being developed by H.M. Collins, drawing on a sociological reading of Wittgenstein (Winch, 1958; cf Bloor, 1983).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.85006-0

Wittgenstein had argued that rules do not contain rules for their own application; such a position, in the form of 'finitism,' also applies to practices of classification (see, for a summary of its relevance to STS, Barnes et al., 1996). The basic claim amounts to observing that every new situation, every new occurrence of some kind of event, is in its own way unique; it therefore requires indeterminate judgments of similarity and difference to decide whether a new event or instance is or is not a legitimate member of a given category. In articles from the mid-1970s onward, and culminating in a book in 1985 (Collins, 1985), Collins developed empirical case studies that exploited Wittgenstein's arguments to show how the participants in actual scientific controversies resolve problems of the ambiguity of experimental outcomes. In work on claims to the detection of high-flux gravitational radiation in the 1970s, as well as in the transfer of skills relating to the successful replication of transversely excited atmospheric (TEA)-laser technology, Collins developed the notion of the 'experimenters' regress,' which he saw as an application in the domain of the sociology of science of Wittgenstein's 'indeterminacy of rules.' In the case of an experimental result, a judgment must be made that involves deciding whether the experiment or experimental run was competently conducted. If yes, then the outcome can be used to assess an associated theoretical claim (which would then open up Duhem–Quine issues), whereas if not, then the outcome will be irrelevant to the claim. Collins showed that, routinely, such assessments of experimental competence were central to scientific controversies. The ultimate criterion for judging the competence of an experiment, according to Collins, was whether the experimental outcome was correct or not. In the case of experiments, the results of which were taken to depend on unknown or controversial aspects of the natural world, therefore, controversy over nature was coextensive with controversy over which experiments were fit to be taken into account in reference to that feature of nature. This, then, is the essential paradox designated by the term 'experimenters' regress': it is only possible to know whether an experiment was competently performed if it is known what the correct outcome should be, but (in an empiricist epistemology) the correct outcome can only be known as a consequence of the performance of competent experiments.

The sort of empirical studies of scientific practice conducted by Collins also yielded fruitfully to analysis in terms of the Duhem–Quine thesis. Trevor Pinch's 'Towards an Analysis of Scientific Observation' (Pinch, 1985) introduced the idea of 'externality' in understanding the role and fate within the scientific community of experimental work. Pinch noted that knowledge-claims made on the basis of experimental studies could themselves be couched in alternative ways that maximized or minimized the likelihood that the competence and reliability of the experiments themselves would be questioned. An experimentally based assertion with high externality would be one that used the experimental outcome to make far-reaching inferences about some domain of existing knowledge; an assertion with low externality would depart but little from a bald statement of the local circumstances of the experimental situation, and could provoke but little controversy. Andrew Pickering (1984) examined similar professional dynamics in the theoretical and experimental work in particle physics that established the existence and characteristics of quarks in the 1970s.

Together, the Duhem–Quine thesis, finitism, and the experimenters' regress had opened up, by the mid-1980s, a vast terrain within which sociologically informed study of the experimental dimensions of scientific knowledge could be conducted. A landmark study that fully exploited the new possibilities was Steven Shapin and Simon Schaffer's *Leviathan and the Air-Pump* (Shapin and Schaffer, 1985/2011). This historical study directly challenged the older 'received view' of the role of experiment in making scientific knowledge, and did so by investigating a classic, oft-cited historical example usually held to exemplify that view. The subject chosen in the book was the air-pump work of Robert Boyle in the 1660s. Boyle's work had been presented (Conant, 1970 [1957]) in one of the *Harvard Case Histories in Experimental Science*, dating from the late 1940s and 1950s. These were publications that, importantly, had shaped the teaching and scholarship in the history of science for many years, and which were consistent with the usual conception of experiments as unequivocal arbiters of theoretical ideas. Shapin and Schaffer thus accepted Boyle's air-pump experiments, for the sake of argument, as an exemplary model of scientific procedure. They then attempted to show how a close examination revealed that Boyle's experiments and their associated knowledge claims were far from unequivocal – were, in fact, strongly contested – and that the interpretation of them advanced by Boyle and his allies could also be understood as part of a political vision of the proper social constitution and conduct of a philosophical community. In order to achieve this purpose, Shapin and Schaffer used the criticisms of Boyle advanced by the philosopher and mathematician Thomas Hobbes, a man whose view of the proper way to make natural knowledge differed significantly from that of Boyle and whose associated conception of the place and legitimacy of knowledge in a polity also differed markedly from Boyle's. Hobbes's detailed criticisms of Boyle's air-pump work revealed an alternative set of values and commitments regarding the best way to achieve social stability and consensus, a way that invalidated experimental work as an effective means of securing knowledge claims. An alternative approach to similar issues (Dear, 1985, 1987, 1995) focuses on conventions of knowledge making that were less local than those examined by Shapin and Schaffer, and which support a view of much more mediated connections with social organization. Dear's main argument is that the modern 'experiment' as a historically specific, contrived event was a development of seventeenth-century Europe, and had close filiations with disciplinary approaches employed in mathematical sciences that stressed operationalism.

Laboratory Studies and Naturalistic Approaches

By the mid-1980s, in addition to the work already discussed, a number of so-called laboratory studies were being published. These studies also had a strong influence on views of experiment in science, but, in contrast to the approaches considered above, they tended to eschew, often explicitly, an overt engagement with the 'received view' philosophies of science. Instead, the methodological and/or epistemological

approaches that were central to laboratory studies drew from such traditions as those of anthropology, ethnomethodology, and phenomenology. The three crucial and exemplary lab studies from this period were Latour and Woolgar's *Laboratory Life* (Latour and Woolgar, 1979), Karin Knorr-Cetina's *The Manufacture of Knowledge* (Knorr-Cetina, 1981), and Michael Lynch's *Art and Artifact in Laboratory Science* (Lynch, 1985). These three have in common a naturalistic stance toward the work of the laboratory, rejecting presuppositions as to the character of that work. Thus, they have no a priori interest in studying laboratory work as a means of learning how 'reliable,' or 'true,' or 'epistemologically privileged' knowledge results from laboratory practices. Instead, they treat their particular laboratories as sites where activities are carried out, and attempt to describe and account for the nature of those activities. A notable feature of the rapidly changing intellectual landscape on which these laboratory studies were performed is the existence, in the first edition of *Laboratory Life*, of the subtitle *The Social Construction of Scientific Facts* (Latour and Woolgar, 1979), and its replacement in the second edition, by the less doctrinaire *The Construction of Scientific Facts* (Latour and Woogar, 1986). The change signals the way in which, in the years surrounding 1980, the rhetoric of SSK, with its insistence on 'social construction' as the crucial mechanism by which scientific knowledge was made, dominated science studies. By 1986, this insistence was already looking rather hackneyed and uninformative in some quarters.

The place of experiment in the tradition of laboratory studies is less clear-cut than it was in the mainstream SSK tradition, because the former were not reacting directly to the epistemological highlighting of experiment characteristic of the philosophical 'received view.' Thus, experiments were less likely to be seen as obviously individuated events; instead, they were seen as falling within an overall passage of scientific activity. The highlighting of particular experiments as especially significant cognitive elements of such passages now became contingent on the behavior and interactions of the scientists being studied, rather than being self-evidently central to an understanding of scientific work.

A collection of papers originating in a conference of 1985 on 'The Uses of Experiment' (Gooding et al., 1989; see also Le Grand, 1990 for a similar collection) attempted to shed light on experiment as a category used by scientists rather than as an epistemologically significant item singled out by philosophers. The contributors made use of the new work in science studies to lay out, chiefly by means of empirical studies of contemporary and past science, the multifarious ways in which experiment has been used as a rhetorical and practical tool in the establishment of knowledge-claims and in the consolidation of scientific communities of practice. One especially notable concept deployed in this volume, centrally in the contribution by Simon Schaffer (1989), was that of experimental 'transparency.' This concept designated situations in which an experimental apparatus becomes so unproblematic and theoretically unquestioned that the phenomena to which it affords access are seen directly, as it were, as if no theoretically justified experimental device intervened between observer and phenomena. In such cases, in effect, the apparatus has changed from being an experimental setup to being a scientific measuring instrument. For example, such was the fate of the mercury barometer in the seventeenth century, which began life as an experiment intended to examine the pressure of the air and the possibility of a vacuum but which, by the end of the century, had become an instrument for measuring atmospheric pressure as an aid to understanding the weather. These kinds of conceptualization also hark back to the much earlier work of the French philosopher Gaston Bachelard, who spoke of the artificial creation of phenomena by apparatus, using the term *phénoméno-technique* (e.g., Bachelard, 1938; see Gaukroger, 1976 for a lucid introduction).

The second half of the 1980s also saw the appearance of empirical studies of modern physical experimentation that adopted attitudes quite different from those of the advocates of SSK. Peter Galison, in his book *How Experiments End* (1987), examined several episodes in post-World War II high-energy physics in the United States. The book's intention was to describe the ways in which experimental closure was brought about in the episodes discussed, but in a way explicitly opposed to accounts given by Collins and Pickering, among others; that is, Galison's aim was to show that social mechanisms were insufficient to explain the outcome of the experimental work. Allan Franklin also published work on modern experimental physics with broadly the same aim (Franklin, 1986, 1990). In both cases, they aimed to stress the role of an independent physical world in creating sometimes-unexpected outcomes to experiments was emphasized, and (in Franklin's case, at least) to downplay the role of social actions and social structures in determining outcomes. Franklin argued that social considerations become irrelevant in light of the overwhelming influence of objective experimental outcomes and their correct rational assessment. Franklin's position, however, appears largely to ignore the epistemological problems that fueled SSK.

Discussions of experiments and the best ways of understanding their role in science also went well beyond issues of their direct epistemological function. Historical studies, preeminently Shapin and Schaffer's *Leviathan and the Air-Pump*, were by now having an impact on wider conceptions of the ideological structure and social meaning of science. The most important example of that impact was Yaron Ezrahi's (1990) *The Descent of Icarus*. This took Shapin and Schaffer's presentation of Robert Boyle's kind of experimental science as the model for science ever since the seventeenth century, and adapted their view of its political meaning to an account of the development of modern liberal ideology and, subsequently, modern liberal states from the eighteenth to the twentieth centuries. In Ezrahi's view, the legitimacy of experimental knowledge claims in science became the accepted norm for knowledge claims in the political arena as well. Shapin and Schaffer had stressed the role of 'witnessing' in Boyle's kind of experimental science, whereby many people, either directly or through the reading of detailed reports ('virtual witnessing'), came to attest the truth of an asserted experimental outcome and thereby its associated factual knowledge claim. Ezrahi suggested that a similar visibility, or transparency, of political actions and the reasons for them became the political ideal, along with the notion of the accountability

of political leaders. Ezrahi's ideal-type of such a polity, the United States in the twentieth century, thus placed great stress on the possibility of citizens seeing what their government was doing and in assessing the grounds for and wisdom of those actions (see also Latour, 1993 for different political extrapolations, with a focus on 'modernity' rather than specific nation states).

Latour on Experiments and Their Mobility

Other aspects of experimentation were also investigated in the 1980s and early 1990s, most notably in the work of Bruno Latour. Latour's naturalistic approach to scientific practice, which had been explicitly anthropological in inspiration in *Laboratory Life*, extended his representation there of laboratory science as an enterprise centrally concerned with the production of 'literary inscriptions' (any kind of trace inscribed on paper). This extension took the form, in his *Science in Action* (Latour, 1987), of an analysis of the structure of experimental papers. That structure, he argued, reveals the agonistic logic of experimental procedures through a kind of shadowboxing, whereby the paper's authors attempt to anticipate and head off skeptical criticisms of their claims, such as might be put forward by a determined opponent. Thus, the experimental paper/document (and, by implication, the experimental protocol itself) takes the form of one side of a dialog aimed at silencing the critic through anticipation of all plausible objections. Part of the power of this analytical approach supposedly resides in Latour's stress on semiotic analysis (Latour and Bastide, 1986), which connects with his emphasis on taking into account 'natural actants' in accounts of scientific activity (see also Latour, 1988). Thus semiotic actors in the experimental account, which are normally taken to refer to things in the natural world (such entities as subatomic particles, or microbes, etc.), are treated on a par with those semiotic actors normally taken to refer to things in the social world (preeminently, the scientists performing the experiment). Thus, they are understood in exactly the same way. Latour therefore claims that there is no justification for treating social actors as real while treating 'natural actors' as simply social constructs, which is what (in his view) the SSK approach advocates. Problems seem to arise in Latour's approach, however, when the question of shifting between the analytical level of the semiotic and the analytical level of the putatively 'real' is considered; Latour is obscure on the point.

One especially influential idea in Latour (1983, 1988) concerns a way of understanding how experimental outcomes can be mobile; that is, how a successful experiment conducted in a specific laboratory is made to be successful, or at least persuasive, at other locations. This issue has been referred to as the 'problem of constructivism' in Golinski (1998); although Latour would not count himself as a 'constructivist' in the SSK sense, his similar stress on the work that goes into creating an experimental 'fact' leaves him with the same problem. The difficulty that Latour purports to solve is to explain how, if experimental facts are made in situations whose specificity and locality ('natural' and 'social') are taken as crucial to explaining the particular outcome, that outcome is seen to be reproduced in significantly different settings. The traditional ('received view') answer is that the experimental outcome is simply a reflection of the way that nature works, and an implicit principle of the uniformity of nature therefore implies that the same experiment, in the sense of the same material setup and procedure, will yield the same result wherever it is conducted. The constructivist, however, because of a conviction that 'social' elements (however defined in detail) are essential to producing the outcome, appears to have difficulty in explaining this mobility of experiment (cf Collins, 1985 on the TEA-laser). Specifically, Latour explains the apparent paradox of experimental success outside the laboratory setting, using Louis Pasteur's field trials of his anthrax vaccine as an example, by denying that the laboratory has ever really been abandoned. Instead, when, for instance, Pasteur managed to create a vaccine found to work on animals in carefully controlled laboratory conditions, his demonstration of its effectiveness on actual farm animals involved, according to Latour, transferring his laboratory to the farm. It was not that Pasteur's vaccine worked in the laboratory and on the farm; the vaccine only ever worked in the context of the laboratory trial. Pasteur's trick lay in turning the farm into a laboratory while maintaining the illusion that it was still only a farm. Hence, experiments can be spread, on this view, only via networks of laboratories or laboratory-equivalents; they do not in fact work everywhere.

Experimental Systems

In the 1990s, there appeared a number of new conceptualizations of experiment, including an important collection of essays concerning the life sciences, *The Right Tools for the Job* (Clarke and Fujimura, 1992), and significant discussions of various experimental practices and regimes in Peter Galison's *Image and Logic* (Galison, 1997). These new conceptualizations had in common a more expansive and complex view of the ways in which experiment forms a part of elaborate systems of scientific practice. Andrew Pickering's *The Mangle of Practice* (Pickering, 1995) developed from and built upon his earlier SSK-influenced work by attempting to understand scientific practice as pragmatic interactions among experimental procedures, theoretical models, and physical reality. The resistance and accommodation arising through the interaction between these theoretical and material practices had become in Pickering's work an important and central feature that rejected any attempt at pure sociological reductionism (although whether this has ever truly existed in STS is a debated point). Pickering saw scientific work as an ongoing enterprise aimed at stabilization: theory and experiment cooperate, at least for a time, in producing mutually consistent, intelligible interactions between human beings and the world. By placing experiment within this more complex structure of scientific practice, Pickering thus endeavored to move beyond the sociological insights of science studies and to regain a connection with the natural (and obdurate) reality with which science has always claimed to deal.

Similar pragmatic moves had also appeared in an important study by the historian Robert Kohler. In his *Lords of the Fly* (Kohler, 1994), Kohler studied the early decades of *drosophila* (fruit-fly) research in the United States, centered on the laboratory of Thomas Hunt Morgan at Columbia University and subsequently at the California Institute of Technology. Kohler described the operations of Morgan's fly lab in terms of the exploitation of the capabilities of the materials, centrally including the flies, with their particular breeding and variation characteristics. The whole taken together, and crucially including material practices (given some precedence in this study over intellectual practices), was characterized by Kohler variously in terms of 'experimental practice,' 'experimental production,' or as a 'system of experimental mass production.' The term 'experimental system' was also used by Hans-Jörg Rheinberger in an ambitious study, *Toward a History of Epistemic Things* (Rheinberger, 1997). Rheinberger, following the same general tendencies as Kohler and others, and with an important nod to Fleck, represented his project as being directed against the old philosophical tendency to view experiments as isolated epistemological elements, instead focusing his attention on the integration of experiment within a multifaceted system of practices which together constitute the modern scientific enterprise (cf also Knorr-Cetina, 1999; especially pp. 39–43). Consequently, experiment has lost much of its former philosophical mystique.

The Disappearance of the Experiment from Science and Technology Studies

Although Collins (2004, 2011) has published further episodes in his long-term study of developments in gravity wave experiments, after the 1990s, there have been few, if any, major works in science and technology studies that explicitly treat 'the experiment' as a topic of investigation. A continued interest in scientific 'practice' (Schatzki et al., 2001) does not necessarily, as noted earlier, place 'the experiment' at its heart. In part, this decline of interest in the topic reflects a trend in the sciences themselves, away from 'pure' physics and toward biomedicine, information science, and hybrids of science and engineering (see Park Doing's, 2009 close ethnographic account of the ascendency of biologists at a facility once dominated by physicists). Studies of the relationship between laboratory practices and those of field sciences further illustrate the layered complexity that now typifies work involving experiment (Kohler, 2002). The relative lack of interest also reflects a move within science and technology studies toward the politics of science and questions about (as well as efforts to engender) public engagement with sciences. Particular agendas and multiple constituencies interact in, for example, agricultural field trials in which farmers tentatively allow scientists to conduct experiments in their fields (Henke, 2008). Other modes of 'impure science' (Epstein, 1996) are exemplified by clinical trials that involve agents and agendas from science, medicine, government, corporations, and social movements, which all have input into experimental designs and results. Such cases are of far greater current interest in science and technology studies than the classic mode of experimental practice. Nevertheless, the protocols and idioms of 'the experiment' are far from irrelevant. Even clinical trials can exhibit a highly creative mix of classical test and control conditions, together with ethical considerations, such as designing a cancer drug trial in a way that takes into account that it will be truncated as soon as evidence emerges of a likely beneficial effect of the tested combination of chemicals (Keating and Cambrosio, 2012).

Earlier research on experiments in history and social studies of science remains important, not perhaps as an explicit foundation for current research, but more as a resource that is invoked for arguing that the logics and practices of contemporary experiments are deeply intertwined with an entire array of social and cultural agents and agencies – logics and practices that do not adhere to the received view of experiment that prominent spokespersons for science, government officials, and even high court judges continue to use to legitimate and rationalize policies and decisions.

See also: Biomedical Sciences, History and Sociology of; Experimental Design: Overview; Kuhn, Thomas S. (1922–96); Laboratory Studies: Historical Perspectives; Logical Positivism and Logical Empiricism; Modernization and Modernity in History; Objectivity: Philosophical Aspects; Science and Technology Studies, Ethnomethodology of; Science and Technology, Anthropology of; Science, History of; Science: Constructivist Perspectives, History of; Scientific Knowledge, Sociology of; Social Constructivism; Social Properties (Facts and Entities): Philosophical Aspects; Social Science, The Idea of; Truth and Credibility in Science.

Bibliography

Bachelard, G., 1938. La formation de l'esprit scientifique. J. Vrin., Paris.
Barnes, B., Bloor, D., Henry, J., 1996. Scientific Knowledge: A Sociological Analysis. University of Chicago Press, Chicago.
Bloor, D., 1983. Wittgenstein: A Social Theory of Knowledge. Macmillan, London.
Clarke, A.E., Fujimura, J.H. (Eds.), 1992. The Right Tools for the Job: At Work in Twentieth-Century Life Sciences. Princeton University Press, Princeton, NJ.
Collins, H., 1985. Changing Order: Replication and Induction in Scientific Practice. Sage, London.
Collins, H., 2004. Gravity's Shadow: The Search for Gravitational Waves. University of Chicago Press, Chicago.
Collins, H., 2011. Gravity's Ghost: Scientific Discovery in the Twenty-First Century. University of Chicago Press, Chicago.
Conant, J.B., 1970 [1957]. Robert Boyle's experiments in pneumatics. In: Conant, J.B. (Ed.), Harvard Case Histories in Experimental Science, vol. 1. Harvard University Press, Cambridge, MA.
Dear, P., 1985. Totius in verba: rhetoric and authority in the early Royal Society. Isis 76, 145–161.
Dear, P., 1987. Jesuit mathematical science and the reconstitution of experience in the early seventeenth century. Studies in History and Philosophy of Science 18, 133–175.
Dear, P., 1995. Discipline and Experience: The Mathematical Way in the Scientific Revolution. University of Chicago Press, Chicago.
Doing, P., 2009. Velvet Revolution at the Synchrotron. MIT Press, Cambridge, MA.
Duhem, P., 1954. The Aim and Structure of Physical Theory. Princeton University Press, Princeton, NJ.
Epstein, S., 1996. Impure Science: AIDS, Activism, and the Politics of Knowledge. University of California Press, Berkeley, CA.

Ezrahi, Y., 1990. The Descent of Icarus: Science and the Transformation of Contemporary Democracy. Harvard University Press, Cambridge, MA.
Fleck, L., 1979 [1935]. Genesis and Development of a Scientific Fact. University of Chicago Press, Chicago.
Franklin, A., 1986. The Neglect of Experiment. Cambridge University Press, Cambridge, UK.
Franklin, A., 1990. Experiment, Right or Wrong. Cambridge University Press, Cambridge, UK.
Galison, P., 1987. How Experiments End. University of Chicago Press, Chicago.
Galison, P., 1997. Image and Logic: A Material Culture of Microphysics. University of Chicago Press, Chicago.
Gaukroger, S., 1976. Bachelard and the problem of epistemological analysis. Studies in History and Philosophy of Science 7, 189–244.
Golinski, J., 1998. Making Natural Knowledge: History of Science after Constructivism. Cambridge University Press, Cambridge, UK.
Gooding, D., Pinch, T., Schaffer, S. (Eds.), 1989. The Uses of Experiment: Studies in the Natural Sciences. Cambridge University Press, Cambridge, UK.
Hanson, N.R., 1958. Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science. Cambridge University Press, Cambridge.
Henke, C., 2008. Cultivating Science, Harvesting Power: Science and Industrial Agriculture in California. MIT Press, Cambridge, MA.
Keating, P., Cambrosio, A., 2012. Cancer on Trial: Oncology as a New Style of Practice. University of Chicago Press, Chicago.
Knorr-Cetina, K.D., 1981. The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science. Pergamon Press, Oxford.
Knorr-Cetina, K.D., 1999. Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press, Cambridge, MA.
Knorr-Cetina, K.D., Mulkay, M.J. (Eds.), 1983. Science Observed: Perspectives on the Social Study of Science. Sage, London.
Kohler, R.E., 1994. Lords of the Fly: Drosophila Genetics and the Experimental Life. University of Chicago Press, Chicago.
Kohler, R., 2002. Landscapes and Labscapes: Exploring the Lab-Field Border in Biology. University of Chicago Press, Chicago.
Kuhn, T.S., 1970 [1962]. The Structure of Scientific Revolutions, second ed. University of Chicago Press, Chicago.
Latour, B., 1983. Give me a laboratory and I will raise the world. In: Knorr-Cetina, K.D., Mulkay, M.J. (Eds.), Science Observed: Perspectives on the Social Study of Science. Sage, London, pp. 141–170.
Latour, B., 1987. Science in Action: How to Follow Scientists and Engineers through Society. Harvard University Press, Cambridge, MA.
Latour, B., 1988. The Pasteurization of France. Harvard University Press, Cambridge, MA.
Latour, B., 1993. We Have Never Been Modern. Harvard University Press, Cambridge, MA.
Latour, B., Bastide, F., 1986. Writing science–Fact and fiction. The analysis of the process of reality construction through the application of socio-semiotic methods to scientific texts. In: Callon, M., Law, J., Rip, A. (Eds.), Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World. Macmillan, London, pp. 51–66.
Latour, B., Woolgar, S., 1979. Laboratory Life: The Social Construction of Scientific Facts. Sage, London.
Latour, B., Woolgar, S., 1986. Laboratory Life: The Construction of Scientific Facts. Princeton University Press, Princeton, NJ.
Le Grand, H.E. (Ed.), 1990. Experimental Inquiries: Historical, Philosophical and Social Studies of Experimentation in Science. Kluwer, Dordrecht, The Netherlands.
Lynch, M., 1985. Art and Artifact in Laboratory Science: A Study of Shop Work and Shop Talk in a Research Laboratory. Routledge & Kegan Paul, London.
Mulkay, M.J., 1979. Science and the Sociology of Knowledge. George Allen & Unwin, London.
Pickering, A., 1984. Constructing Quarks: A Sociological History of Particle Physics. University of Chicago Press, Chicago.
Pickering, A., 1995. The Mangle of Practice: Time, Agency and Science. University of Chicago Press, Chicago.
Pinch, T., 1985. Towards an analysis of scientific observation: the externality and evidential significance of observational reports in physics. Social Studies of Science 15, 3–36.
Quine, W.V.O., 1953. Two dogmas of empiricism. In: Quine, W.V.O. (Ed.), From a Logical Point of View: Nine Logico-Philosophical Essays. Harvard University Press, Cambridge, MA.
Rheinberger, H.J., 1997. Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube. Stanford University Press, Stanford.
Schaffer, S., 1989. Glass works: Newton's prisms and the uses of experiment. In: Gooding, D., Pinch, T., Schaffer, S. (Eds.), The Uses of Experiment: Studies in the Natural Sciences. Cambridge University Press, Cambridge, UK.
Schatzki, T., Knorr Cetina, K., Savigny, E.v. (Eds.), 2001. The Practice Turn in Contemporary Theory. Routledge, New York.
Shapin, S., Schaffer, S., 1985/2011. Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life, second ed. Princeton University Press, Princeton, NJ.
Winch, P., 1958. The Idea of a Social Science and Its Relation to Philosophy. Routledge & Kegan Paul, London.

Expert Testimony

Margaret B Kovera and Jacqueline L Austin, John Jay College of Criminal Justice, CUNY, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by M. Kovera, volume 8, pp. 5139–5141, © 2001, Elsevier Ltd.

Abstract

Courts in civil and common law systems rely upon experts to help triers of fact render appropriate legal judgments. Expert testimony strongly influences evaluations of trial evidence and verdicts. Because expert testimony is so influential, the laws governing the admissibility of expert evidence are evolving. Despite the courts' best efforts to admit only unbiased and reliable expert evidence, judges may lack the skills and abilities necessary to accurately evaluate scientific methodology, and the adversarial nature of many legal systems may induce partisan bias in expert testimony. Therefore, court systems need procedural safeguards to help fact finders evaluate the quality of expert testimony.

To assist triers of fact in rendering appropriate legal judgments, courts may seek the assistance of experts. Courts admit expert testimony in legal proceedings when a judge or jury needs assistance evaluating a material fact in a court proceeding. Qualified by education, skill, or technical expertise, an expert provides fact finders with specialized information to educate them about case-relevant issues that are not within the common knowledge of the fact finder. Expert testimony differs from other types of testimony in legal proceedings because only experts may offer opinion testimony. As such, expert testimony is treated with care in the legal system.

Expert Testimony in Civil Law Systems

The primary distinction between the use of experts in civil and in common law systems is who retains the expert. Experts in civil law systems are generally called by the presiding judge as a court-appointed expert. Judges seek expert testimony when the issues to be resolved are beyond the understanding of the court. Judges may seek the opinion of one or multiple experts to answer a question. The expert may be required to conduct research to answer the question or may just be asked to provide known information that would help address the issue. Generally, the judge will provide a timeline for the expert to provide a written report that addresses the question posed to the expert. Upon the deadline, this report is delivered to the parties by the judge. Normally, there is no oral testimony of the expert unless the parties or the court call the expert to answer questions (Spencer, 1998). Although the general rule is that the judge will appoint an expert when necessary, parties may (but rarely do) call their own expert witnesses in civil law systems (Timmerbeil, 2003). However, when experts are appointed by a party rather than the court, they are often restricted from testifying and are only permitted to provide a written report (Timmerbeil, 2003).

Experts in the civil law system are often thought to be more neutral, more independent, and less biased than experts in common law systems because they are not accountable to any particular party or lawyer. However, judges in civil law systems also play a more active role in the adjudicative process. Thus, the judges in civil law jurisdictions are often viewed as less neutral than judges in common law jurisdictions, so there is some debate as to whether these experts are less biased than common law experts or just biased toward answering the judges' question in a manner that is consistent with the judges' opinion (see van Kampen, 2003).

Laws regarding the admissibility of expert testimony in civil law systems focus on the scope of the expert's duties and whether the expert's testimony or opinion is binding upon the presiding judge. For example, French courts maintain lists of experts who meet certain qualifications, although judges are free to appoint an expert who is not on the list (Taylor, 1996). In Korea and France, the findings of the expert are not binding upon the judge (Lee, 1997; Taylor, 1996).

Expert Testimony in Common Law Systems

In common law systems, it is typical for either party or both parties to proffer expert testimony. Much like civil law systems, judges also have the power to appoint court-appointed, neutral experts (Federal Rules of Evidence (FRE) 706); however, in practice, court-appointed experts are rare (Spencer, 1998). Because the parties' respective attorneys typically hire experts, experts in the common law system are often criticized for being hired guns. Indeed, attorneys may seek experts who will testify favorably with regard to their position, rather than providing testimony that is representative of the general scientific community. Further, experts are well paid in common law systems for their testimony.

In common law systems, judges must determine whether the testimony will be admitted into evidence. Each country in the common law system has its own unique laws regarding the admissibility of expert testimony, but there are several striking similarities across many of the countries, including the United States, Australia, England, and Canada (Gatowski et al., 1996). In each of these countries, judges may admit expert evidence only if it is relevant to a factual issue in contention and may admit expert evidence only if it will not confuse, mislead, or unfairly bias the jury. To be admitted at trial, the expert evidence must be able to assist the trier of fact in making a decision. If the expert evidence is within the common knowledge of the fact finder, then it should not be admitted. One concern with regard

to the admissibility of expert evidence is the reliability or scientific validity of the information that experts are providing in court. Court opinions in both the United States and Canada have addressed concerns about the proliferation of unreliable or junk science in court. These decisions have delineated specific criteria that judges must use to assess the reliability of expert evidence. Courts in England and Australia have not provided specific criteria for evaluating the reliability of expert evidence; however, they have ruled that expert evidence must be based on legitimate science or expertise (Gatowski et al., 1996).

Laws Governing the Admissibility of Expert Testimony in the United States

In the United States, expert testimony is required to meet certain criteria before being admitted into court. The purpose of these criteria is to ensure that unreliable expert testimony is prevented from entering the courtroom and influencing jurors' decisions. In 1923, federal courts evaluated scientific expert testimony on the general acceptance test, established in *Frye v. United States*, which specified that scientific evidence should be admitted if it is generally accepted within the relevant scientific community. This rule originated from a case in which James Frye was on trial for murder and his attorney wanted to present the results of a systolic blood pressure test, a predecessor to the polygraph, in support of his innocence. The trial court did not permit this testimony. Frye appealed and argued that the expert's testimony was improperly excluded. The District of Columbia (DC) Court of Appeals, however, agreed with the trial court and stated that because systolic blood pressure was not generally recognized as a method of lie detection within the scientific community, the trial court had met the appropriate grounds for exclusion. In essence, *Frye* placed the responsibility of evaluating the admissibility of expert testimony squarely on the shoulders of the relevant scientific community. Two main criticisms emerged from relying on *Frye*'s general acceptance test as the standard for admissibility. First, *Frye* excluded novel but valid research. Second, *Frye* was criticized for accepting invalid research that met the general acceptance test among a majority of researchers.

In the late 1960s, US Supreme Court Chief Justice Earl Warren and the US Congress worked together to create a series of new rules designed to help judges determine whether to admit evidence at trial, referred to as the FRE. The FRE included new rules governing the admissibility of expert testimony. These rules specify that expert evidence must be relevant to an issue in the trial (FRE 403), must not be within the common understanding of the juror, and must assist jurors with understanding a disputed issue in the trial (FRE 702). Further, in the Advisory Committee notes, the committee members specified that the rule is broadly written so as not to limit judges' evaluations 'merely to the scientific and technical but extend to all specialized knowledge. Similarly, the expert is viewed, not in a narrow sense, but as a person qualified by knowledge, skill, experience, training or education. Thus, within the scope of the rule are not only experts in the strictest sense of the word, e.g., physicians, physicists, and architects, but also the large group sometimes called skilled witnesses, such as bankers or landowners testifying to land values'. In essence, the FRE called on judges to evaluate all forms of expert testimony – whether it be specialized, technical, or scientific – when determining admissibility.

One shortcoming of the FRE was that the drafters did not address the *Frye* rule, which led to some confusion among judges in federal courts about whether the *Frye* standard or the FRE should govern the admissibility of expert testimony. This confusion was addressed by the US Supreme Court in 1993, when the Court clarified the standards for the admissibility of scientific evidence in *Daubert v. Merrell Dow Pharmaceuticals* (1993). Jason Daubert, Eric Schuler, and their parents sued Merrell Dow Pharmaceuticals for marketing a drug, Bendectin, given to mothers for morning sickness. The children suffered from serious birth defects and proffered testimony from eight experts to demonstrate the link between the cited birth defects and Bendectin; they presented research of *in vitro* and *in vivo* animal studies, pharmacological studies on the chemical structure of Bendectin showing that Bendectin had a chemical structure that mimicked other drugs known to cause birth defects, and a meta-analysis composed of epidemiological studies that compared the rate of birth defects among children whose mothers had or had not ingested Bendectin during pregnancy. The District Court granted Merrill Dow's motion for summary judgment because the research methodologies cited by the plaintiffs were not generally accepted. The Ninth Circuit Court of Appeals affirmed the District Court's ruling, again on grounds that the research was not generally accepted. Appealing to the US Supreme Court, the plaintiffs argued that the lower courts improperly applied the *Frye* rule when determining admissibility and stated that the FRE should govern the admissibility of expert testimony. The US Supreme Court granted *certiorari*, overturned *Frye's* general acceptance test as the singular rule for granting expert testimony, and stated that the more recently adopted FRE superseded *Frye*.

Daubert also specified that testimony must be both relevant and reliable to be admissible. To assist judges in the evaluation of reliability, the Court provided a list of four nonexhaustive factors that courts could use in their determination of reliability: whether the data can be falsified, whether the data have been subjected to peer review or publication, whether the known or potential error rates are available, and general acceptance of the findings in the relevant scientific community. Federal Rule of Evidence 702 was amended in 2000 to mirror the reliability factors in *Daubert*, stating that expert testimony should be admitted "provided that (1) the testimony is sufficiently based upon reliable facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case" (FRE 702, 2000). In addition to these changes, *Daubert* also placed squarely on the shoulders of the trial judge the responsibility for determining whether scientific evidence is reliable, rather than simply allowing them to evaluate the scientific field's conclusions. Essentially, this decision conveyed that judges are to be evidentiary 'gatekeepers' for scientific testimony. *General Electric Co v. Joiner* (1997) reiterated judges' gatekeeping role and stressed that judges must determine admissibility by evaluating reliability.

Daubert addressed only the admission of scientific testimony but was mute on the requirements for nonscientific

testimony. Therefore, judges could avoid the responsibility of evaluating evidentiary reliability by categorizing an expert's testimony as 'specialized' rather than 'scientific.' In *Kumho Tire Company v. Carmichael* (1999), the Court expanded judges' responsibility to also include evaluating the reliability of nonscientific expert testimony. In this case, the plaintiff was in a serious car crash caused by a tire blowout that resulted in serious injuries and a death. The plaintiff wanted to present testimony from an expert in tire failure analysis to demonstrate that the tire blowout was caused either by a defect in the tire's design or by a manufacturing defect. The trial court excluded the testimony on grounds that the testimony did not meet the standards established in the FRE or the reliability factors specified in *Daubert*. The Supreme Court disagreed with the trial court and instructed judges that their obligation to evaluate testimony goes beyond just the evaluation of scientific knowledge and extends to specialized and technical knowledge as well. Further, they specified that the factors listed in the *Daubert* decision do not apply to the evaluation of nonscientific testimony. In determinations of reliability for specialized and technical knowledge, the Court emphasized that the four factors listed in *Daubert* are not exhaustive, which gave trial judges a wide range of discretion in their determination of reliability. Following the federal courts, many state courts in the United States have adopted admissibility standards based on *Daubert* and its progeny.

Legal Decisions About the Reliability of Expert Evidence

Daubert requires judges to evaluate the reliability of expert testimony, but judges may lack the ability to apply the *Daubert* factors in the absence of scientific training. Studies of judges' abilities to evaluate expert evidence suggest that they have little to no knowledge of social science methods, making it unlikely that they are capable of evaluating empirical methods (Manuto and O'Rourke, 1991). A national survey of state trial court judges found that although 91% of the state judges believed that their role of gatekeeper was appropriate, only 52% felt their education had sufficiently prepared them to perform scientific evidence evaluations (Gatowski et al., 2001). In the survey, judges stated that the four *Daubert* factors would guide their decision to admit or deny expert testimony; however, they lacked the ability to operationalize and understand concepts central to *Daubert*'s criteria for evaluating reliability, such as error rate and falsifiability. When researchers rated the judges' understanding of the four factors, only 6% of judges understood the concept of falsifiability and only 4% understood what is meant by an error rate. Judges demonstrated a better understanding of peer review (71%) and general acceptance (81%).

In one study designed to answer the question of whether judges have the ability to make competent judgments about the reliability and validity of expert evidence, Florida state judges read a summary of an expert's testimony in a hostile work environment case and then ruled on whether they would admit that testimony in court (Kovera and McAuliff, 2000). The judges read proposed expert testimony that either was fundamentally flawed (i.e., was missing a control group, suffered from a confound, or suffered from experimenter bias) or was valid. Judges who had no scientific training admitted the expert evidence equally often regardless of study validity. However, judges with some prior training rated valid research as more legally admissible than invalid science. Judges without scientific training appear to lack the skills necessary to differentiate between valid and flawed scientific testimony and therefore may be ineffective gatekeepers, which could result in the admission of unreliable expert evidence.

Legal Safeguards Against Unreliable Expert Evidence

Although there are a number of procedural safeguards in US courts to protect against jurors basing their decisions on unreliable expert evidence (e.g., cross-examination, judicial instructions, and opposing expert testimony), empirical evidence questions their efficacy. Judicial instructions on the burden of proof likely make jurors more skeptical of all expert testimony and provide no means of better equipping jurors to evaluate the reliability and the validity of the evidence (Kovera et al., 2002). In at least one study, a cross-examination crafted to educate jurors about methodological issues did not improve jurors' sensitivity to variations in the methodological quality of a study on which an expert based her testimony, in comparison to a cross-examination that failed to address methodological issues (Kovera et al., 1999).

Opposing experts demonstrate varying success at educating jurors about scientific validity. Studies show that the presence of opposing experts increases juror skepticism regarding both scientific evidence and the experts who testify (Cutler and Penrod, 1995; Devenport and Cutler, 2004). In one study, opposing experts failed to sensitize jurors to the presence of internal validity flaws in the expert's research, even when their testimony specifically addressed threats to internal validity. Indeed, participants were more likely to convict when a prosecution expert testified, regardless of the scientific validity of the opposing defense expert's study (Levett and Kovera, 2008). However, another study demonstrated that opposing experts might increase both skepticism and sensitivity in a sexual harassment trial (Levett and Kovera, 2009). In a trial simulation, a plaintiff's expert based her testimony on a study that either lacked a control group or was valid. Different versions of the trial also had different types of opposing expert testimony (none, an opposing expert who did not address methodology, or an opposing expert who did address methodology). In the absence of opposing expert testimony, participants rated the defendant to be more liable and the initial expert as more competent and qualified than when an opposing expert also testified. However, when the opposing expert addressed the scientific methodology of the initial expert, jurors rated research that was valid to be more valid than invalid research and jurors rated the expert who presented valid science to be more competent than the expert who presented invalid science. Opposing expert testimony did not affect the main decisions in the case, such as the likelihood that the plaintiff faced a hostile work environment or verdict (Levett and Kovera, 2009). In sum, the available empirical data suggest that opposing experts consistently increase skepticism but either fail to educate jurors about unreliable scientific evidence or fail to influence trial decisions enough to influence key decisions.

Adversarial Allegiance in Expert Testimony

In theory, ethical standards maintain that expert witnesses will be unbiased, impartial, neutral, and objective when testifying in court. Indeed, experts should educate decision makers in a manner that informs and improves decision making. However, adversarial judicial systems may encourage the selection of biased experts. Attorneys are required to seek experts who will represent and support their client's case, which may push attorneys to seek expert witnesses who hold views favorable to their case rather than experts who hold views that are generally accepted within the scientific community. In a survey, judges and attorneys stated that the primary challenge for expert testimony post-*Daubert* is that experts may lose objectivity in a case by becoming advocates for the side that hired them (Krafka et al., 2002). Experts also opine that they feel pressure to provide evidence that supports the position of the party who retained them (Saks, 1990; Saks and Wissler, 1984). Further, it is not unusual for attorneys to push experts into advocacy, encouraging them to provide testimony that is favorable to their case and beyond what the experts' ethical standards should comfortably allow them to say.

This pressure to testify as an advocate may cause experts to overstate conclusions, present information in a biased manner, or present only information that is favorable to the client rather than evidence that is representative of the field's knowledge. Several studies suggest that adversarial allegiance (or allegiance to the side calling the witness) affects fact witnesses. In one study, participants playing the witnesses watched a simulated bar fight and then testified about what they saw one week later. During the intervening week, another participant playing the role of a neutral, court-appointed attorney or an adversarial attorney interviewed the witnesses. The witnesses' testimony was more biased when an adversarial attorney rather than a nonadversarial attorney conducted the intervening interview (Sheppard and Vidmar, 1980).

Despite the expert's ethical obligations to provide impartial testimony, adversarial allegiance appears to affect expert testimony in addition to the testimony of fact witnesses. In one study, graduate students learned that they had been retained by a defense or by a prosecution attorney to interpret an ambiguous criminal case. When the graduate students thought the defense had retained them, they were more likely to reach a conclusion that supported the defendant. When the graduate students thought the prosecution had retained them, they were more likely to reach a conclusion that supported the state. Thus, the retaining side affected their evaluations of case facts (Otto, 1989). In another study, witnesses played the role of either a nonadversarial witness (i.e., one who had been appointed by the court) or an adversarial witness who had been called by either the defense or the plaintiff to testify. Both the judge, who listened to the testimony and who was kept unaware of the study's purpose, and objective raters who listened to tape-recorded versions of the witnesses' testimony found that the testimony from adversarial witnesses was subtly biased toward the side who had called them in relation to the testimony from the neutral witness. Thus, testimony may simply be biased because a particular party called the witness (Vidmar and Laird, 1983). Even in real trials, experts' reports of clinical assessments using instruments with high rates of interrater reliability that are relatively objectively scored show signs of adversarial allegiance. When researchers examined all 23 cases in one state in which opposing experts differed in their assessments of a defendant's psychopathy with a highly reliable scale, experts' reports were more likely to favor the side that hired them (Murrie et al., 2008). Similar evidence of adversarial allegiance was found in a review of clinical psychologists' expert reports in civil sex offender commitment cases (Murrie et al., 2009).

Concurrent Expert Testimony as a Potential Remedy for Adversarial Allegiance

Partisan expert opinions and battling experts in legal settings have encouraged some courts to seek alternative methods of addressing adversarial allegiance among experts. For example, Australian and Canadian courts are modifying the way in which expert testimony is prepared and presented. This technique, referred to as concurrent expert testimony, is designed to reduce adversarial bias and improve the ability of fact finders to understand and evaluate trial evidence (Sanders, 2007). Concurrent expert testimony is a multistep process composed of both a pretrial conference and a two-part joint testimony session (Edmond, 2008, 2009).

The purpose of the pretrial conference is for experts to have an opportunity to indicate areas of agreement and resolve as many differences as possible prior to trial. After the experts have discussed their positions, they produce a joint report that outlines areas of agreement as well as areas of dispute, noting who holds which position (Edmond, 2008, 2009). All participating experts must endorse this report. The absence of the attorneys from the pretrial conference is thought to increase experts' independence from the parties who have retained them and enable the experts to render an impartial opinion (Edmond, 2008).

During the trial proceedings, the experts testify together in a two-part process. In the first phase of testimony, experts have a great deal of testimonial freedom. Each expert outlines his or her personal interpretation of the case along with supporting information. Following this initial presentation, experts are free to ask each other questions and comment on the opposing experts' testimony and evidence. The judge may play an active role in this fact-finding process by asking questions, suggesting additional topics, and directing commentary. Lawyers may also ask questions during this part of the process; however, their manner is less formal than conventional direct examination. The second segment of the concurrent testimony is more typical of adversarial proceedings in which lawyers are free to examine the experts. Lawyers may use this time to cross-examine experts who oppose their position. Cross-examination questions may be presented to one or many of the experts, and experts may also be asked to comment on another expert's answer.

Proponents of concurrent testimony argue that the process enhances neutrality and reduces partisanship on the part of experts because it moves them from close proximity to their lawyers to close proximity to other experts (Edmond, 2008, 2009; Garbis, 2003). Further, concurrent expert testimony is designed to enhance fact finders' comprehension and application of expert evidence to case facts (Downes, 2004; Edmond,

2008, 2009; Garbis, 2003; McClellan, 2007). However, research is needed to determine whether concurrent expert testimony does indeed enhance neutrality among experts, reduce partisanship, and improve fact finders' comprehension of expert evidence and application of it to case facts.

Empirical Research on the Effect of Expert Testimony on Jurors' Decisions

Psychological researchers rely primarily on trial simulation techniques to examine the effects of expert testimony on fact finders' judgments. These trial simulations involve presenting mock jurors (who may be college undergraduates, members of the community, or jurors who have reported to jury duty) with written, audiotaped, or videotaped trial stimuli and measuring behavioral opinions. The trial stimuli vary the presence or absence of expert testimony or some characteristic of the expert testimony.

Empirical Research on Adversarial Expert Testimony

Research on the effects of expert testimony on jury decisions became common in the 1980s. Most of this research has examined whether the testimony of an expert called by one of the parties (an adversarial expert) affected juror or jury decisions. For example, early research examined the question of whether expert testimony influenced juror judgments of guilt. Trial simulation studies demonstrated that when jurors heard expert testimony presented by the defense, they were more likely to return defense verdicts than were jurors who did not hear expert testimony (e.g., Schuller and Cripps, 1998). In contrast, jurors who heard expert testimony presented by the prosecution were more likely to convict the defendant than were jurors who did not hear prosecution expert testimony (e.g., Brekke and Borgida, 1988).

After establishing that expert testimony influenced juror decisions, researchers began to explore whether certain variables moderate the effects of expert testimony on juror judgments. For example, expert testimony has greater influence on jury decisions when it appears early rather than late in the trial (Brekke and Borgida, 1988; Schuller and Cripps, 1998). There is some speculation that early presentation is advantageous because it allows the expert evidence to serve as a framework for understanding other trial evidence. Several studies have also demonstrated that expert testimony is more influential if the expert explicitly links the scientific research to the facts of the case under consideration (Brekke and Borgida, 1988; Schuller, 1992).

Eventually, researchers began to study whether expert testimony improves the quality of juror decisions. Cutler and colleagues (Cutler et al., 1989) presented jurors with a trial simulation in which they manipulated factors affecting the reliability of eyewitness identifications (e.g., the presence of a weapon, disguised perpetrator, or retention interval). They also manipulated the presence of expert testimony describing how these factors influence the accuracy of identifications (e.g., people are less likely to recognize a perpetrator who wore a disguise). In the absence of expert testimony, jurors' decisions were unaffected by factors known to decrease the reliability of eyewitness identifications. After hearing expert testimony on the topic, jurors became sensitive to these factors and used this information in making decisions. Thus, experts may improve fact finders' abilities to make educated decisions. Further, this sensitization effect is more likely to occur when the expert explicitly links the scientific research and the facts of the case (Kovera et al., 1997).

Empirical Research on Nonadversarial Expert Testimony

There may be important differences between the effects of adversarial expert testimony and the testimony provided by a nonadversarial or court-appointed expert. Some have argued that because a nonadversarial expert is proffered by the court rather than one of the parties in a case, jurors may be less likely to critically evaluate the evidence provided by that expert. Indeed, testimony from a court-appointed expert does not have the same sensitizing effect that adversarial expert testimony has on jurors. In one study, nonadversarial expert testimony on the factors influencing the reliability of eyewitness identifications caused jurors to become skeptical of the reliability of all eyewitness identification rather than just those identifications made under unreliable conditions (Cutler et al., 1989). This pattern of effects suggests that jurors are not systematically processing the expert evidence presented by the court-appointed expert, and therefore do not understand the subtleties of that testimony (e.g., that only some eyewitness identifications are unreliable, but not all).

Another concern raised about the use of court-appointed or nonadversarial experts is that their testimony will be prejudicial because jurors may accord it undue weight. Trial simulation research comparing juror evaluations of nonadversarial and adversarial expert testimony demonstrated that jurors find both types of experts equally credible and that their testimony is given similar weight by jurors when making trial judgments (Brekke et al., 1991). Moreover, jurors seem to better remember evidence that is presented by an adversarial expert than evidence presented by a nonadversarial expert, providing additional evidence that jurors do not scrutinize nonadversarial expert testimony as closely as they scrutinize adversarial testimony.

Conclusion

Although expert testimony on many different topics is admitted in court, controversy still surrounds whether judges have the skills and abilities required to evaluate scientific methodology with the goal of preventing junk science from reaching a jury. Because judges receive little education that would enable them to make sophisticated judgments about the reliability and validity of expert evidence, it is important that the judicial system provides scientific training to judges – who are the gatekeepers tasked with keeping poor science out of evidence – and attorneys, who must make arguments before judges to exclude unreliable expert evidence. This sort of training is essential to ensure that these legal actors have the skills necessary to identify bad science. This training is particularly important in light of trial simulation evidence that expert evidence has a great influence on juror decisions, even when it lacks scientific validity, and that the

procedural safeguards in place to help jurors make informed decisions seem ineffective. Additional research is needed to identify other procedures that could help fact finders properly evaluate the quality of expert evidence.

See also: Courts and Adjudication; Forensic Psychology; Judges; Juries; Litigation; Science and Law; Scientific Evidence: Legal Aspects; Trials.

Bibliography

Brekke, N., Borgida, E., 1988. Expert psychological testimony in rape trials: a social-cognitive analysis. Journal of Personality and Social Psychology 55, 372–386.

Brekke, N., Enko, P., Clavet, G., Seelau, E., 1991. Of juries and court-appointed experts: the impact of nonadversarial versus adversarial expert testimony. Law and Human Behavior 15, 451–475.

Cutler, B., Dexter, H., Penrod, S., 1989. Expert testimony and jury decision making: how do jurors make decisions? Law and Human Behavior 20, 379–394.

Cutler, B., Penrod, S., 1995. Mistaken Identification: The Eyewitness, Psychology, and the Law. Cambridge University Press, New York, NY, US.

Daubert vs. Merrell Dow Pharmaceuticals, Inc. 1993 113 S. Ct. 2786.

Devenport, J., Cutler, B., 2004. Impact of defense-only and opposing eyewitness experts on juror judgments. Law and Human Behavior 28, 569–576.

Downes, G., February 2004. Concurrent Expert Evidence in the Administrative Appeals Tribunal: The New South Wales Experience. Paper presented at the Australasian Conference of Planning and Environmental Courts and Tribunals. Hobart, Australia.

Edmond, G., 2008. Secret of the "hot tub": expert witnesses, concurrent evidence and judge-led law reform in Australia. Civil Justice Quarterly 27, 51–82.

Edmond, G., 2009. Conventions in science and law: Merton and the hot tub: scientific conventions and expert evidence in Australian civil procedure. Law and Contemporary Problems 72, 159–189.

Federal Rules of Evidence. 1974, 1984, 2000. West, St. Paul, MN.

Frye vs. United States 1923. 293 F. Supp. 1013 (D.C. Cir).

Garbis, M., 2003. Aussie inspired musings on technological issues: of kangaroo courts, tutorials, and hot tub cross-examination. The Green Bag 6, 141–150.

Gatowski, S., Dobbin, S., Richardson, J., Ginsburg, G., Merlino, M., Dahir, V., 2001. Asking the gatekeepers: a national survey of judges on judging expert evidence in a post-*Daubert* world. Law and Human Behavior 25, 433–458.

Gatowski, S., Dobbin, S., Richardson, J., Nowlin, C., Ginsburg, G., 1996. The diffusion of scientific evidence: a comparative analysis of admissibility standards in Australia, Canada, England, and the United States, and their impact on social and behavioural sciences. Expert Evidence: The International Digest of Human Behaviour, Science and Law 4, 86–92.

General Electric Co. v. Joiner 1997, 522 U.S. 136, 118 S. Ct. 512.

Kovera, M., Gresham, A., Borgida, E., Gray, E., Regan, P., 1997. Does expert psychological testimony inform or influence juror decision making? A social cognitive analysis. Journal of Applied Psychology 82, 178–191.

Kovera, M., McAuliff, B., 2000. The effects of peer review and evidence quality on judge evaluations of psychological science: are judges effective gatekeepers? Journal of Applied Psychology 85, 574–586.

Kovera, M., McAuliff, B., Hebert, K., 1999. Reasoning about scientific evidence: effects of juror gender and evidence quality on juror decisions in a hostile work environment case. Journal of Applied Psychology 84, 362–375.

Kovera, M., Russano, M., McAuliff, B., 2002. Assessment of the commonsense psychology underlying *Daubert*: legal decision makers' abilities to evaluate expert evidence in hostile work environment cases. Psychology, Public Policy, and Law 8, 180–200.

Krafka, C., Dunn, M., Johnson, M., Cecil, J., Miletich, D., 2002. Judge and attorney experiences, practices, and concerns regarding expert testimony in federal civil trials. Psychology, Public Policy, and Law 8, 309–332.

Kumho Tire Co. v. Carmichael 1999 526 U.S. 137, 119 S. Ct. 1167.

Lee, E., 1997. Expert evidence in the republic of Korea and under the US federal rules of evidence: a comparative study. Loyola of Los Angeles International and Comparative Law Journal 19, 585–632.

Levett, L., Kovera, M., 2008. The effectiveness of opposing expert witnesses for educating jurors about unreliable expert evidence. Law and Human Behavior 32, 363–374.

Levett, L., Kovera, M., 2009. Psychological mediators of the effects of opposing expert testimony on juror decisions. Psychology, Public Policy, and Law 15, 124–148.

Manuto, R., O'Rourke, S., 1991. Federal judges' perceptions of social research in judicial decision-making. Communication Reports 4, 103–106.

McClellan, P., 2007. Contemporary challenges for the justice system – expert evidence. Paper presented at the Australian Lawyers' Alliance Medical Law Conference. Sydney, Australia.

Murrie, D., Boccaccini, M., Johnson, J., Janke, C., 2008. Does interrater disagreement on psychopathy checklist scores in sexually violent predator trials suggest partisan allegiance in forensic evaluations? Law and Human Behavior 32, 352–362.

Murrie, D., Boccaccini, M., Turner, D., Meeks, M., Woods, C., Tussey, C., 2009. Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: evidence of adversarial allegiance in forensic evaluation? Psychology, Public Policy, and Law 15, 19–53.

Otto, R., 1989. Bias and expert testimony of mental health professionals in adversarial proceedings: a preliminary investigation. Behavioral Sciences and the Law 7, 267–273.

Saks, M., 1990. Expert witnesses, nonexpert witnesses, and nonwitness experts. Law and Human Behavior 14, 291–313.

Saks, M., Wissler, R., 1984. Legal and psychological bases of expert testimony: surveys of the law and of jurors. Behavioral Sciences and the Law 2, 435–449.

Sanders, J., 2007. Ethical considerations in the use of expert testimony: expert witness ethics. Fordham Law Review 76, 1539–1584.

Schuller, R., 1992. The impact of battered woman syndrome evidence on jury decision processes. Law and Human Behavior 16, 597–620.

Schuller, R., Cripps, J., 1998. Expert evidence pertaining to battered women: the impact of gender of expert and timing of testimony. Law and Human Behavior 22, 17–31.

Sheppard, B., Vidmar, N., 1980. Adversary pretrial procedures and testimonial evidence: effects of lawyer's role and Machiavellianism. Journal of Personality and Social Psychology 39, 320–332.

Spencer, J., 1998. The Role of Experts in the Common Law and the Civil Law: A Comparison. Expert Witnesses in Child Abuse Cases: What Can and Should Be Said in Court. American Psychological Association, Washington, DC, US, pp. 29–81.

Taylor, R., 1996. A comparative study of expert testimony in France and the United States: philosophical underpinnings, history, practice, and procedure. International Law Journal 31, 181–213.

Timmerbeil, S., 2003. The role of expert witnesses in German and U.S. civil litigation. Annual Survey of International & Comparative Law 9, 162–187.

van Kampen, P., 2003. Expert Evidence: The State of Law in the Netherlands and the United States. Adversarial Versus Inquisitorial Justice: Psychological Perspectives on Criminal Justice Systems. Kluwer Academic/Plenum Publishers, New York, NY, US, pp. 209–234.

Vidmar, N., Laird, N., 1983. Adversary social roles: their effects on witnesses' communication of evidence and the assessments of adjudicators. Journal of Personality and Social Psychology 44, 888–898.

Expertise, Anthropological Approaches to

Dominic Boyer, Rice University, Houston, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Over the past two decades, anthropological research on experts and expertise has flourished. However, rather than treating experts and expertise as new foci of anthropological inquiry, this article explores the historical roots of this field in the anthropology of knowledge, in the postcolonial and reflexive turns of the 1970s and 1980s, and in the rise of science and technology studies as one of the greatest influences on anthropological research in the 1990s and 2000s. The article discusses anthropology's relative disinterest in proposing universal models of experts and expertise and its commitment instead to exploring the processual dimensions of how specialized knowledge and knowers are made, dimensions in which epistemic schemata are formulated, challenged, reproduced, circulated, and, sometimes, institutionalized as 'expertise' by social actors who reciprocally become identified as 'experts' in the process. Yet, the article does show that the anthropological study of experts and expertise extends and refines anthropological reflexivity by encouraging new kinds of interexpert exchanges and relations. In closing, the author predicts that this area of research will continue to be ethnographically and theoretically generative in the future.

Roots in the Anthropology of Knowledge

Experts and expertise are relatively new research subjects for anthropological inquiry; the field has only achieved a distinctive shape and substance since the 1990s. But like so many seemingly novel subjects of research, experts and expertise belong to a broad root system of associated themes with a deep disciplinary history. The anthropology of experts shares common research sites and populations with the anthropology of elites, with the anthropology of science and technology, and with the anthropology of modernity, to name a few of these associations. But what makes it distinctively old and new is its permutation of a long legacy of anthropological research on knowledge and ways of knowing. One is immediately reminded of Malcolm Crick's argument that the anthropology of knowledge "is not a subfield but rather a reminder of what anthropology is centrally concerned with" (1982: p. 287).

Indeed, engagement with certain aspects of knowledge (particularly rationality, logic, and cognition) represented one of the key watersheds in anthropology's disciplinary formation in the late nineteenth and early twentieth centuries. As anthropology began to explore its implications as a rigorous field science (and not solely a laboratory or historical–geographic science) of human culture and society at large, it needed to make sense of the diversity of different kinds of human knowledge it was encountering across the world. As is well known, universal-evolutionary schemas were popular among nineteenth- and early-twentieth century anthropologists as a way of ordering human epistemic diversity and also of elevating Northern/Western/Christian anthropological science above all other forms of human cultural and social knowledge. But, as is also well known, anthropology's integrity as an empirical field science came to militate against universalist paradigms in the twentieth century. The critical debates surrounding characterizations of 'primitive' knowledge in texts like E.B. Tylor's *Primitive Culture*, Sir James Frazer's *The Golden Bough*, and Lucien Lévy-Bruhl's *How Natives Think* are good cases in point. These debates were not only expressions of anthropological science at work, but actually also served to integrate and solidify an anthropological science in the making. They raised questions concerning the universality and relativity of human knowledge that created entire research programs within the fledgling discipline: did so-called primitives have the same logical and cognitive capabilities as 'moderns'? Were their knowledge works different, but equivalently valid, or instead erroneous? How can one comparatively evaluate the epistemic qualities of magic, science, and religion? Lévy-Bruhl argued, for example, that primitives had logical and cognitive capacities like us but that they were unrigorous and unscientific in their application of them. The primitive was thus imagined as antithetical to the modern (scientific) expert. However, Boasians like Paul Radin argued that there were primitive 'men of thought' who were just as rigorous in their application of logic and as creative in their knowledge works as moderns (1927). The advancing critique of universalizing evolutionary paradigms of human cognition and behavior further drew a distinction between anthropology as an empirical science of biological inheritances and effects and anthropology as an empirical science of inquiry into the plurality and unity of human social, cultural, and material forms. In other words, it completed a process that had begun already in the late eighteenth century to distinguish a 'pragmatic anthropology' (in Kant's terms), the study of what human beings make of themselves, from a 'physical anthropology,' the study of what human beings inherit from nature. As the antiuniversalist position became increasingly mainstreamed in anthropology in the twentieth century, the pragmatic study of diverse forms and ways of knowing remained a unifying focus for anthropological inquiry despite its proliferating research sites, topics, and conversations. The key point being that the study of knowledge helped make anthropology what it is today.

Given the long-standing centrality of research on knowledge to the anthropological enterprise, it is unsurprising that the recognition of specialized knowledge as social phenomenon also has a relatively long history in the discipline. Once again, comparative reflection on the status of European science with

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12237-8

other human systems of knowledge was the starting point. But 'experts' and 'expertise' began to appear in the peripheral vision of the field no later than Evans-Pritchard's designation of Nuer chiefs as 'ritual experts' (1940). By the 1960s, references to experts began to appear with more frequency within ethnography, particularly in the tradition of British social anthropology. However, where 'religious experts' and 'ritual experts' (see, e.g., Knutsson, 1963: p. 507) were mentioned, the motivating interest typically remained religion or ritual rather than expertise per se. 'The expert' was a relatively transparent social designation, not one that seemed to admit or reward further anthropological theorization.

In certain respects, this remains the case today (Boyer, 2008). Other fields in the human sciences, notably sociology (e.g., Abbott, 1988; Brint, 1994; Freidson, 2001), behavioral and cognitive psychology (Ericsson and Smith, 1991), and science and technology studies (Collins and Evans, 2002) have expended much more concerted effort on theorization of experts and expertise than anthropology has. There is thus, at least as of yet, no singular anthropological definition of the expert. However, drawing upon the phenomenological, praxiological, and semiological traditions of the discipline, one can at least schematize an expert as an actor who has developed skills in, semiotic-epistemic competence for, and attentional concern with, some sphere of practical activity or 'enactment' (Carr, 2010). By this definition, a car mechanic or a street performer are clearly experts in their respective crafts, although the qualitative dimensions of their expertise are very different (and valued differently) from more widely recognized and institutionally secure experts like doctors, lawyers, or scientists. Indeed, by linking expertise to the combination of skill, competence, attention, and practice, it becomes clear that there is no human being who is not 'expert' in some fashion, much as Gramsci wrote that all men are intellectuals even if not socially validated as such.

Although the very openness of this definition may appear to weaken its analytical capacity, the way it highlights the tension between the experiential-performative and social-institutional poles of skilled knowing and doing actually offers, from the point of view of anthropology, analytical traction in just the right place. What anthropology seeks is a comparative exploration of how specialized knowledge and knowers operate in complex social environments across the world. Among ethnographers working in these fields, there is a common concern with how experts and expertise are made and unmade through social and historical processes. For example, a recent study explores how news journalists are struggling to adapt their traditional expertise and authority as news makers to new scales, temporalities, and institutions of electronic information (Boyer, 2013). A general theory of expertise is not necessary to embark on such projects as anthropologists often find themselves uncompelled by the 'smooth illusion' (Jones, 2010) of universalist models of expertise and more keen to explore the diverse, complicated practices through which epistemic schemata are formulated, performed, authorized, challenged, reproduced, circulated, and, ultimately, institutionalized as 'expertise' by social actors who reciprocally become identified as 'experts' in the process. This deep comparativist attention to the processual dimension of the making of specialized knowledge (and knowledge makers) is what distinguishes the anthropological approach to experts and expertise within the human sciences.

The Mainstreaming of Anthropological Research on Experts and Expertise

One might fairly ask, why has the past two decades seen such rapid growth of anthropological research interest in experts and expertise? The reasons are several. The first factor is undoubtedly anthropology's move to reckon with its colonial legacies in the late 1960s and early 1970s (Hymes, 1969; Asad, 1973) as part of the rise of a broader postcolonial challenge to the legacies of European political, social, and of course also epistemic imposition. Decolonization in the 1950s and 1960s not only fundamentally impacted the rules of the game of anthropological field research and knowledge making across the world, but also challenged the discipline's historical association with the study of 'nonmodern' peoples. Calls increased for symmetrical and critical attention to 'modern' Northern and Western cultures and societies as well. The folkloristic-national and comparative-international wings of anthropological research were to be fused or at least brought into closer alignment. Northern and Western cultures and especially powerful groups in those cultures were no longer to be safely obscured from anthropological illumination. It was, in Laura Nader's famous phrase, time "to study up" (1969).

The validation of new anthropologies of modernity and elites certainly helped to open the gateway for experts and expertise to become legitimate subjects of inquiry. But an even more immediate propellant for an emerging 'anthropology of experts' was the enormous explosion of interest around anthropological science and technology studies in the 1990s and 2000s (e.g., Fischer, 2003; Fortun, 2001; Franklin and Roberts, 2006; Gusterson, 1996; Helmreich, 2000; Knorr Cetina, 1999; Rabinow, 1997; Sunder Rajan, 2006; Traweek, 1988). The circumstances surrounding the rise of science and technology studies to prominence within the human sciences are themselves complex. On the one hand, the inventive hybrid philosophical–empirical programs of figures like Michel Callon, Michel Foucault, and Bruno Latour could certainly be viewed as capable of inspiring cross-disciplinary research programs by themselves. On the other, it seems clear that the so-called postmodern shift to explore the contingencies and limits of scientific/expert knowledge making in the 1970s and 1980s resonated strongly with the crisis of Keynesian technocracy to guarantee an ever-improving modernity as the traditional colonial powers became increasing vulnerable and dependent (e.g., on the oil of an increasingly ungovernable Middle East).

But whatever the deeper cause of its popularity and intuitiveness, what is undeniable is that science and technology studies have massively impacted anthropological research horizons over the past two decades, the concepts and arguments spilling outside of obviously scientific and technical contexts into other domains, particularly into investigations of state making and technocratic contexts as well (e.g., Shore and Wright, 1997; Mitchell, 2002). In the 2010s, there are few domains of anthropological research today that science and technology studies have not touched in some way. At the

same time, it is important to recognize that an expanding 'anthropology of professions' exists that is not simply reducible to the extension of science studies' research into other areas of expertise. Recent ethnographic work on activists (Howe, 2013), advertisers (Mazzarella, 2003), journalists (Boyer, 2005; Hannerz, 2004), lawyers (Mertz, 2007), magicians (Jones, 2010), and politicians (Holmes, 2000), among many other forms of work and life, is exemplary for its plurality of analytical approaches. But it is also characterized by a common, close attention to practices and institutions of expert knowledge. Both the anthropology of professions and anthropological science and technology studies have contributed decisively to the critical mass of ethnographic case material needed for comparative exploration of the singularities and commonalities of expertise across the world. Even in cases where 'expert' and 'expertise' are not the dominant analytic categories, a strong case can be made that this research explores essentially the same dynamics and phenomena, however named. And, increasingly one finds 'expertise' emerging as a dominant analytic for anthropological studies of knowledge (see, e.g., Carr, 2010; Jones, 2010; Roosth, 2013).

The final factor that has helped shape and legitimate anthropological research on experts and expertise has been the methodological reevaluation of disciplinary norms of fieldwork and ethnography, a process that began overtly in the 1980s with the 'reflexive turn' of landmark works like James Clifford and George Marcus's *Writing Culture* but that has recently pursued more systematic methodological reflection on the changing expectations and practices of anthropological knowledge making (see, e.g., Faubion and Marcus, 2009). Although reflexivity is a venerable feature of knowledge in anthropology and human sciences, with better hindsight one can appreciate now that the significance of the reflexive turn was growing recognition that anthropological ethnography needed to be more methodologically self-aware to cope with different expectations for research in a very different world from the imperial era in which it originated. In a word, anthropology, as a matter of survival as a field science, needed to pay better attention to its own 'situatedness.' Although reflexivity is sometimes viewed, particularly by other social sciences, as a sign of empirical weakness or narcissism, anthropology's constant self-analysis of its own limits and patterns of understanding can also be interpreted as making it a more precise and rigorous empirical science. As discussed in the next section, the normalization of anthropological reflexivity over the past three decades has opened up the space for a new theorization of the interexpert relations of fieldwork and postfieldwork knowledge production.

'Paraethnography' and the Relations between Cultures of Expertise

In an important recent article on 'cultures of expertise,' Doug Holmes and George Marcus have noted that anthropological engagements of other experts inevitably bring anthropological knowledge into disquieting but generative juxtaposition with other modes of expert knowledge, some of them distinctively 'paraethnographic' or 'paraanthropological,' which exist outside the networks and institutions of academic anthropology. It is worth citing their evaluation of this situation at length:

> In our experience, ethnographers trained in the tradition of anthropology do not approach the study of formal institutions such as banks, bureaucracies, corporations, and state agencies with much confidence. These are realms in which the traditional informants of ethnography must be rethought as counterparts rather than 'others' – as both subjects and intellectual partners in inquiry. … Here we suggest a particular strategy for re-functioning ethnography around a research relation in which the ethnographer identifies a paraethnographic dimension in such domains of expertise – the de facto and self-conscious critical faculty that operates in any expert domain as a way of dealing with contradiction, exception, facts that are fugitive, and that suggest a social realm not in alignment with the representations generated by the application of the reigning statistical mode of analysis. Making ethnography from the found paraethnographic redefines the status of the subject or informant, asks what different accounts one wants from such key figures in the fieldwork process, and indeed questions what the ethnography of experts means within a broad, multi-sited design of research. (Holmes and Marcus, 2005: pp. 236–237)

It is worth noting that Holmes and Marcus's optimism regarding the prospect of collaboration with other experts in fieldwork is not shared by all anthropologists of experts. Some see instead the threat of an epistemological end point for anthropology in the potential doubling, collapse, and/or cancellation of analytical knowledge forms. For example, if similar social theories are shared by both the analyst and the informant, then what value or expertise is the academic observer really adding to the research process?

While such skepticism provokes interesting reflection, its more dire scenarios of anthropological superfluity remain unrealized. It is never the case, for example, that anthropology fully shares analytical practices let alone professional ideologies with other cultures of expertise. But the situation of contact between anthropologists and other experts can indeed generate friction and misunderstanding. The root of this friction is ultimately sociological and linked to the boundary management of jurisdictions, which Andrew Abbott terms the 'defining relation' in professional life (1988: p. 3). Abbott describes how professions jealously guard the jurisdictions upon which their social privilege depends. When cultures of expertise encounter each other and threaten to overlap, tensions and conflicts inevitably arise. Anthropologists are certainly no strangers to such tensions. Consider, for example, the expanding appropriation of ethnographic research techniques and culture theory in the business world, in government, even in the military. This is expert knowledge that is both ours and not ours, a phenomenon that simultaneously seems to confirm the importance of anthropological expertise and to threaten it. Sigmund Freud describes the anxiety-inducing effect of doubling well: "having been an assurance of immortality … [it] becomes the uncanny harbinger of death." As in all professional life, uncertain jurisdictions undermine identity and authority and often elicit fears of irrelevance or extinction. The concerns that arise from the practice of the anthropology of experts are thus real but could be reframed as follows: On what basis does the representative of one culture of expertise

(the anthropologist) claim legitimate analytical jurisdiction over the members of another culture of expertise and how is this claim enacted? How can an ethnographer document another expert culture without reframing their expert knowledge into the analytical categories of anthropology, thus quietly subverting one's interlocutors to the epistemic norms of the anthropological culture of expertise?

The anthropology of experts can be, in short, unsettling of traditional anthropological research relations. It is often hard to pay proper respect to 'emic' understandings when these represent alternative models of expertise sometimes competing for authority and legitimacy with one's own. Tara Schwegler, for example, found it is dangerously easy to critique the expertise of the neoliberal economists she researched in Mexico (2008), and much more challenging to properly appreciate the integrity and inventiveness of their knowledge making practices. It can also be the case that it is impossible to meaningfully absorb enough of another language and culture of expertise to understand its nuances properly. Truly coming to inhabit another culture of expertise amounts, after all, to a second professionalization process. Finally, even though the anthropology of experts tends to be a fairly cordial and collegial affair, the threats of exclusion and surveillance are never far away. Cultures of expertise are often socially privileged, quasisovereign, and able to restrict fieldwork access, to monitor the acquisition and subsequent circulation of their expert knowledge, and even, if they are so inclined, to police ethnographic and theoretical content. They offer obstacles like intellectual property rights, offices of corporate communications, and nondisclosure agreements, which would-be ethnographers must navigate. And, even in the best access scenarios, the highly specialized task sets, professional credentials, and organizational hierarchies of many cultures of expertise close doors to even the best prepared, most 'insider' ethnographers.

Such unsettling circumstances help explain why anthropological research on experts has spurred serious consideration of the limits and necessary renewal of anthropological knowledge (e.g., Rabinow, 1999), rising even into talk of epistemological dilemmas and crises. But the value of the anthropology of experts in this author's view is precisely that it sharpens and extends our reflexive awareness beyond the glosses of epistemological dilemma and crisis and toward a deeper appreciation of the intersubjective and contextual contingencies of making knowledge. Attention to the dynamics of interexpert exchange in the process of composing anthropological knowledge, for example, makes us more aware of the particularities of anthropology's jurisdiction in the human sciences. We realize how we are experts operating in an ecology of expertise, specialists in field research, in social-theoretical analysis, in ethnographic writing, who coexist and interact constantly with other kinds of skilled knowers and doers. We gain new appreciation of the 'situatedness' both of our knowledge and of others.' We come to understand how some epistemic practices and conditions inhabit many disciplines and jurisdictions and how others are more singular. We are almost by necessity forced to develop forms of reciprocity that balance the interests of analyst and informant in the situation of fieldwork. These are the ethics of what Hannerz has aptly termed studying sideways (2004). Anthropologists of experts have thus done inventive work at creating generative conversations across jurisdictions (e.g., Helmreich, 2009), engaging and translating other expertise to help extend our own into new research areas. In this respect, one feels that the study of experts and expertise is one of anthropology's most successful areas of conceptual and analytical innovation as we have the constant possibility of borrowing and repurposing tokens of expertise drawn from our interlocutors and collaborators.

The anthropology of experts, in sum, enriches and extends the horizon of anthropological reflexivity and anthropological knowledge, not only inviting new modes of interexpert collaboration but also deepening self-knowledge of the distinctive capacities of anthropology's epistemic practices and inspiring new experiments to expand those capacities.

Conclusion

As noted above, more and more anthropologists are becoming interested in research on experts and expertise. New additions to the ethnographic ranks (like Summerson Carr's addiction counselors, Graham Jones's magicians, and Sophia Roosth's molecular gastronomists, to name only a few of many examples) are being incorporated into the broader conversation. And, as anthropological science and technology studies, the anthropology of modernity, and the anthropology of professions explore their overlapping research interests, one finds a shared analytic terminology of 'experts' and 'expertise' becoming more widespread. Thus, it seems very likely that the next decade will see the anthropology of experts and expertise continue to flourish and that part of this flourishing will be more focused conceptual work on operationalizing experts and expertise from an anthropological perspective.

The future, in other words, looks bright. The author highlights one horizon toward which there are hopes the field will grow. Anthropologists still need to push experts outside of their professional comfort zones and to engage their lives more fully. It is still rare to find studies of cultures of expertise that follow experts outside their work spaces, that go home with them, to their children's playgrounds, to family gatherings, to bars and churches, and so on (for a counterexample, see Hugh Gusterson's (1996) resourceful ethnography of nuclear weapons scientists beyond their laboratories). Given the aforementioned difficulties of gaining access to cultures of expertise, it perhaps is not surprising that, once 'inside,' ethnographers have tended to encamp themselves in the relative security of the professional, institutional, and public dimensions of expert lives. Of course, this work has been very rewarding and significant ethnographically. The trouble is that it plays into professional–ideological conceits that what constitutes 'expert' life is fundamentally professional, very often mental, and that all other dimensions of expert lives are trivial, at least when it comes to the constitution of 'expertise.' This is the 'phenomenology of expertise' (Boyer, 2005) in action, how the continuous experience of intense professional and mental activity constitutes a truth horizon in which skilled knowers and skilled knowledge are posited as ontologically supreme. Professional ideologies thus tend to resist (strongly) the idea

that expertise also has roots in informal and situational understandings; that expert rationality possesses a halo of irrational desires, fantasies, and anxieties; that cultural and gender norms help constitute expert subjectivities; and so on. Yet, deeper ethnographic investigation has already shown that the content of expert knowledge is intimately tied to the social and phenomenological experience of the expert knower (Boyer, 2005, 2013). A fuller anthropological portrait of experts and expertise thus demands greater attention to these factors. It is not a question of diminishing anyone's expert authority but rather of subjecting experts to the same standards and practices of analysis anthropology routinely applies to everyone else.

See also: Industrial Societies; Intellectuals, Sociology of; Knowledge, Anthropology of; Knowledge, Sociology of; Modernity: Anthropological Aspects; Political Anthropology; Reflexivity in Anthropology; Science and Technology, Anthropology of; Scientific Knowledge, Sociology of; Sociological Theory.

Bibliography

Abbott, A., 1988. The System of Professions: An Essay on the Division of Expert Labor. University of Chicago Press, Chicago.
Asad, T. (Ed.), 1973. Anthropology and the Colonial Encounter. Humanities Press, Atlantic Highlands.
Boyer, D., 2005. Spirit and System: Media, Intellectuals, and the Dialectic in Modern German Culture. University of Chicago Press, Chicago.
Boyer, D., 2008. Thinking through the anthropology of experts. Anthropology in Action 15 (2), 38–46.
Boyer, D., 2010. Digital expertise in online journalism (and anthropology). Anthropological Quarterly 83 (1), 73–96.
Boyer, D., 2013. The Life Informatic: Newsmaking in the Digital Era. Cornell University Press, Ithaca.
Brint, S.G., 1994. In an Age of Experts: The Changing Role of Professionals in Politics and Public Life. Princeton University Press, Princeton.
Carr, E.S., 2010. Enactments of expertise. Annual Review of Anthropology 39, 17–32.
Collins, H.M., Evans, R., 2002. The third wave of science studies: studies of expertise and experience. Social Studies of Science 32 (2), 235–296.
Crick, M., 1982. Anthropology of knowledge. Annual Review of Anthropology 11, 287–313.
Ericsson, K.A., Smith, J. (Eds.), 1991. Toward a General Theory of Expertise: Prospects and Limits. Cambridge University Press, Cambridge.
Evans-Pritchard, E.E., 1940. The Nuer. Clarendon, Oxford.
Faubion, J.D., Marcus, G.E. (Eds.), 2009. Fieldwork Is Not What It Used to Be. Cornell University Press, Ithaca.
Fischer, M.M.J., 2003. Emergent Forms of Life and the Anthropological Voice. Duke University Press, Durham.
Fortun, K., 2001. Advocacy after Bhopal: Environmentalism, Disaster, New Global Orders. University of Chicago Press, Chicago.
Franklin, S., Roberts, C., 2006. Born and Made: An Ethnography of Preimplantation Genetic Diagnosis. Princeton University Press, Princeton.
Freidson, E., 2001. Professionalism: The Third Logic. University of Chicago Press, Chicago.
Gusterson, H., 1996. Nuclear Rites: A Weapons Laboratory at the End of the Cold War. University of California Press, Berkeley.
Hannerz, U., 2004. Foreign News: Exploring the World of Foreign Correspondents. University of Chicago Press, Chicago.
Helmreich, S., 2000. Silicon Second Nature: Culturing Artificial Life in a Digital World. University of California Press, Berkeley.
Helmreich, S., 2009. Alien Ocean: Anthropological Voyages in Microbial Seas. University of California Press, Berkeley.
Holmes, D.R., 2000. Integral Europe: Fast Capitalism, Multiculturalism, Neofascism. Princeton University Press, Princeton.
Holmes, D.R., Marcus, George E., 2005. Cultures of expertise and the management of globalization: toward the re-functioning of ethnography. In: Ong, A., Collier, S.J. (Eds.), Global Assemblages. Blackwell, Oxford.
Holston, J., 1989. The Modernist City: An Anthropological Critique of Brasilia. University of Chicago Press, Chicago.
Howe, C., 2013. Intimate Activism: The Struggle for Sexual Rights in Postrevolutionary Nicaragua. Duke University Press, Durham.
Hymes, D. (Ed.), 1969. Reinventing Anthropology. Vintage Books, New York.
Jones, C., 2010. Better women: the cultural politics of gendered expertise. American Anthropologist 112 (2), 270–282.
Knorr Cetina, K., 1999. Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press, Cambridge.
Knutsson, K.E., 1963. Social structure of the Mecca Galla. Ethnology 2 (4), 506–511.
Mathews, A., 2011. Instituting Nature: Authority, Expertise, and Power in Mexican Forests. The MIT Press, Cambridge.
Mazzarella, W., 2003. Shoveling Smoke: Advertising and Globalization in Contemporary India. Duke University Press, Durham.
Mertz, E., 2007. The Language of Law School: Learning to "Think Like a Lawyer". Oxford University Press, New York.
Mitchell, T.J., 2002. Rule of Experts: Egypt, Techno-Politics, Modernity. University of California Press, Berkeley.
Nader, L., 1969. Up the anthropologist – perspectives gained from studying up. In: Hymes, D. (Ed.), Reinventing Anthropology. Pantheon, New York.
Povinelli, E., 2002. The Cunning of Recognition: Indigenous Alterities and the Making of Australian Multiculturalism. Duke University Press, Durham.
Rabinow, P., 1997. Making PCR: A Story of Biotechnology. University of Chicago Press, Chicago.
Rabinow, P., 1999. French DNA: Trouble in Purgatory. University of Chicago Press, Chicago.
Radin, P., 1927. Primitive Man as Philosopher. Appleton, New York.
Roosth, S., 2013. Of foams and formalisms: scientific expertise and craft practice in molecular gastronomy. American Anthropologist 115 (1), 4–16.
Schwegler, T., 2008. Take it from the top (down)? Rethinking neoliberalism and political authority in Mexico. American Ethnologist 35 (4), 682–700.
Shore, C., Wright, S. (Eds.), 1997. Anthropology of Policy: Critical Perspectives on Governance and Power. Routledge, London.
Sunder Rajan, K., 2006. Biocapital: The Constitution of Postgenomic Life. Duke University Press, Durham.
Traweek, S., 1988. Beamtimes and Lifetimes: The World of High Energy Physicists. Harvard University Press, Cambridge.

Explanation: Conceptions in the Social Sciences

Hans Westmeyer, Free University of Berlin, Berlin, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 8, pp. 5154–5159, © 2001, Elsevier Ltd.

Abstract

Explanation is an important goal of any scientific inquiry. An explanation is a statement or account which makes what is to be explained clearer than it was before and promotes its understanding. There are different kinds of entities to be explained in the social sciences: events or phenomena; scientific laws or theories; the functioning of systems; the meaning of expressions, ideas, or behaviors; actions in the sense of their (moral) justification. The most important kind are events and phenomena. The question 'What is the proper structure and which are the characteristics of a good or adequate explanation of events or phenomena?' is at the heart of a lively debate between the proponents and opponents of different models of scientific explanation. The best known ones are the deductive-nomological model of scientific explanation and the statistical-relevance model of probabilistic explanation. These models presuppose a comprehensive and well articulated knowledge base which is not available in most of the social sciences. More realistic recent alternatives are the model of aleatory explanation and, as the most promising one, the embedding theory of scientific explanation which will presumably play a leading role in future discussions of the issue.

Explanation is an important goal of any scientific inquiry. Many social scientists would say that it is the most important one. Nevertheless, the expression 'explanation' as it is used within the social sciences is still a vague and ambigous term with at least five different connotations: (E1) the explanation of an event or a phenomenon, (E2) the explanation of a scientific law or theory, (E3) the explanation of the functioning of a system, (E4) the explanation of the meaning of an entity (e.g., an expression, an idea, a behavior), and (E5) the explanation of an action in the sense of its (moral) justification. In each case, an explanation is a statement or account which makes what is to be explained clearer than it was before and promotes its understanding. But how this is to be achieved for the different variants of the concept of explanation and what are the characteristics of a good or adequate explanation in each case is a controversial topic not only in the social sciences but in science in general (cf Stegmüller, 1983). The focus of discussion has been, from the beginning, on the first variant E1, i.e., the explanation of events or phenomena. This variant and the problems involved will be the main subject of this article.

Explanation-Seeking Questions

Explanations can be understood as answers to certain questions, the so-called *explanation seeking questions*. In the social sciences, a very important type of explanation seeking questions refers to actions of a certain kind performed by a certain person. An explanation of individual actions is not only of interest in itself, but it is also required for a proper understanding of many *macrosocial* phenomena that are construed as the aggregate result of individual actions of large numbers of persons. As Little (1991) puts it, all social processes, causal influences, systemic interactions, etc., are ultimately embodied in the actions of individual actors within a specific social and natural environment. This may be one reason why most publications about explanation in the social sciences primarily, or even exclusively, deal with the explanation of human actions.

Suppose a person p performed an action of kind a. There are at least the following eight questions relevant to the explanation of this singular action a of person p (cf Tuomela, 1977):

(Q1) Why did p do a?
(Q2) What was the reason of p for doing a?
(Q3) What was the motive (emotion, feeling, sensation, personality characteristic) which led p to do a?
(Q4) Which of the wants (desires, decisions, duties, obligations, norms) of p brought about his or her doing a?
(Q5) Which of the beliefs (thoughts, considerations, knowledge) of p brought about his or her doing a?
(Q6) What was the purpose (intention, aim, end, goal) of p for doing a?
(Q7) How is it possible that p did a?
(Q8) What does p's doing a mean (express)?

Answers to these questions are not independent from each other. Q1 could be answered by referring to causes and reasons for action a, whereas Q2 seems to accept only the latter. Q4 and Q5 which already mention certain kinds of reasons in the formulation of the question are included in Q2. Thus, an answer to Q4 and Q5 can be transformed into an answer to Q2. An answer to Q3 can be interpreted as an answer to Q1 or Q2. Tuomela assumes that answers to questions Q1–Q5 are statable as:

(A1) p did a because of c.

Here c gives the causes or reasons for p's doing a. All answers to Q6 can be paraphrased in the form

(A2) p did a in order to do g.

Here g is p's goal action for the purpose of which a was performed. An illustration of A2 would be: p saved money in order to enjoy life after retirement.

Tuomela claims that answers of the type A2 can be further reduced to answers of the type A1 and that the expression 'because of' in A1 signifies a causal relation between a and c. Answers to *how possible questions* of type Q7 are incomplete variants of A1. Explanations of individual actions by means of

sociological variables such as minority status, social class, or social position of an agent are good examples of answers to questions of type Q7.

Answers to questions of type Q8 are of a totally different nature. They belong to category E4, i.e., the explanation of the meaning of an entity, in that they reveal an additional understanding of action *a*. The so-called *speech acts*, that are linguistic actions, may illustrate this type: if action *a* of person *p* consists of the utterance 'I will do that!,' an appropriate answer to the question 'What does *p*'s saying "I will do that!" express?' might be 'This utterance of *p* is a promise, not only an announcement or a prediction.'

Answers of type A1 seem to be fundamental in the explanation of human actions. Differentiating between types of psychological explanations allows a more comprehensive treatment of component *c* of A1.

Types of Psychological Explanations

Bunge and Ardila (1987) suggest the following classification of psychological explanations: (T1) tautological explanations, (T2) teleological explanations, (T3) mentalist explanations, (T4) metaphorical explanations, (T5) genetic explanations, (T6) developmental explanations, (T7) environmental explanations, (T8) evolutionary explanations, (T9) physiological explanations, and (T10) mixed explanations. Each type shall be briefly explained (in the sense of E4) and illustrated by two simple examples.

T1: *Tautological explanations* refer to basic capabilities or mental faculties: person *p* is able to anticipate (or anticipates) events because of *p*'s forethought capability. Person *p* is able to think about (or thinks about) herself because of *p*'s self-reflective capability.

T2: *Teleological explanations* refer to goals or purposes (cf component *g* in A2): person *p* sold his shares in order to maximize his profit. Person *p* studied medicine in order to become a doctor.

T3: *Mentalist explanations* refer to mental events: person *p* developed a neurosis because *p* suffered from an intrapsychic conflict. Person *p* works so hard because *p* is highly motivated.

T4: *Metaphorical explanations* refer to analogies with physical or social processes or with animals or machines: person *p* processes information like a parallel processing computer does. Person *p*'s drivings follow the same principles as a steam boiler.

T5: *Genetic explanations* refer to the genetic endowment of a person: person *p* has a high intelligence because her or his parents also had a high intelligence. Person *p* suffers from Alzheimer's disease because *p* carries two apolipoproteine E type 4 alleles.

T6: *Developmental explanations* refer to stages or levels of biological or psychological development or to events in the persons's past: person *p* does not understand the lever laws because *p* is not yet in the stage of formal operations. Person *p* suffers from agoraphobia because *p* has experienced a severe trauma in early childhood.

T7: *Environmental explanations* refer to external conditions and factors: person *p* shows a high response rate because *p* is working under a variable ratio reinforcement schedule. The agoraphobic symptoms of *p* are weakened because *p* is massively exposed to the threatening stimuli.

T8: *Evolutionary explanations* refer to the survival value of a behavior or behavior tendency, its selective advantages or disadvantages: male persons show sexual jealousy in order to ensure paternity in putative offspring by preventing encroachment by competing males, and this behavior tendency has evolved because of the selective advantage of a greater reproduction rate of these persons (cf Buss, 1997). Persons have a high pain threshold under duress because of its survival value.

T9: *Physiological explanations* refer to physiological, especially neurophysiological and endocrinological, processes and mechanisms: introverts are more arousable than extraverts because introverts have a lower threshold for activation in the ascending reticular activating system than extraverts. The depressive person *p* experienced an elevation of her or his mood because *p* took a cyclic antidepressant which increased the chemical neurotransmitter serotonin.

T10: *Mixed explanations* are combinations of two or more of the above mentioned types of explanation. Mostly, explanations are not pure cases of one type, but combinations of at least two types, i.e., they are mixed explanations. Cases of one type usually provide only a partial answer to the respective explanation seeking question. In answers of the type A1 to explanation seeking questions and in most of the given examples of the different types of psychological explanations, the relation term 'because' occurs. It links that which is to be explained to that which explains. A characterization of this relation is provided by the so-called *models of explanation*.

Models of Explanation

Explanation is one of the main subjects of the philosophy of science. So far, many different models of explanation have been constructed, most of them incompatible with one another (cf Salmon, 1989). Each model attempts to answer at least two questions: (1) What is (the structure of) an explanation? and (2) What is a good (proper, appropriate, adequate) explanation? Some answers to these questions will be briefly outlined.

The Deductive-Nomological Model of Explanation

According to this classical view, an explanation is an *argument* which shows that the phenomenon to be explained can be inferred from certain other facts by means of specified general laws. This type of argument may be schematized as a deductive inference of the following form (cf Hempel, 1965):

(D-NE)	$L_1, L_2, \ldots, L_r$	General laws
	$C_1, C_2, \ldots, C_K$	Statement of antecedent conditions
	E	Description of the empirical phenomenon to be explained

E as the description of the phenomenon to be explained is called the *explanandum* (*sentence*). The statements of antecedent conditions, which make assertions about particular facts, and the general laws together form the *explanans*, i.e., that which explains. Explanations of this kind are called explanations by deductive subsumption under general laws, or *deductive-nomological explanations*. (The root of the term 'nomological' is the Greek word 'nomos' for law.)

D-NE gives an answer to the question 'What is (the structure of) an explanation?' A comparison between this structure and the one underlying the examples of the different types of psychological explanations reveals that most of the latter are stated in an elliptical form, viz., '*E* because of *c*' (cf A1). The component *c* refers only to a subset of the antecedent conditions $C_1, C_2, \ldots, C_k$ that are required for a proper explanation of the explanandum *E*, and the general laws, if there are any at all, are totally omitted. Thus, explanations of the form '*E* because of *c*' do not count as proper explanations.

To give an answer to the second question, i.e., 'What is a good explanation?,' requires the formulation of *conditions of adequacy*. In the case of the model of deductive-nomological explanation, there are four such conditions (cf Hempel, 1965):

(R1) The explanandum must be a logical consequence of the explanans, i.e., the explanandum must be logically deducible from the information contained in the explanans.
(R2) The explanans must contain general assumptions or hypotheses, and these must actually be required for the derivation of the explanandum.
(R3) The explanans must have empirical content, i.e., it must be capable, at least in principle, of test by experiment or observation.
(R4) The sentences constituting the explanans must be well confirmed.

It is not easy to find examples of deductive-nomological explanations in the social and behavioral sciences that display the proper structure and satisfy the conditions R1–R4. One example may be: the phenomenon to be explained is Peter's frequent aggressive behavior in the presence of his mother. The explanation is given with reference to the theory of operant behavior. The explanandum statement *E* can be formulated as 'Peter shows a high frequency of aggressive behavior if his mother is present.' The principle of positive reinforcement 'If a positive reinforcer is presented contingent on the occurence of a behavior, then the frequency of this behavior will increase' can be the law statement *L*. At least two statements of antecedent conditions are required, viz., A_1 'Contingent on Peter's aggressive behavior, his mother regularly shows attentive behavior,' and A_2 'His mother's attentive behavior is a positive reinforcer for Peter.'

In this case, the explanandum *E* seems to be deducible from the explanans, i.e., the conjunction of *L*, A_1, and A_2. Thus, R1 as the first condition of adequacy seems to be satisfied. But what about the conditions R2–R4? That depends on the status of *L*. If *L* would have empirical content and could be regarded as well confirmed, R2–R4 would be satisfied too. With regard to *L*, two problems arise: (1) According to its critics, *L* is no empirical law, but only a logical consequence of the definition of the concept of a positive reinforcer (the much discussed problem of circularity of the empirical 'law of effect'); (2) even if this were not the case, *L* is not a well confirmed universal law because its validity depends upon conditions which are not explicitly mentioned in *L*. One important condition is the following: (C1) there exists no other behavior in the respective situation contingent on the occurrence of which a more effective positive reinforcer is presented. If there exists only one such other behavior in the case of Peter and his mother – and this is not excluded by the antecedent conditions A_1 and A_2 – a quite different law statement would become relevant, viz., the 'law' of parallel reinforcement. If there exist even two or more of the behaviors excluded by C1 which seems to be the usual case in unrestricted natural environments, it is difficult to find an appropriate law statement at all.

This example is representative of the problems that conflict with any attempt to apply the model of deductive-nomological explanation to the social sciences. Universal deterministic laws as they are required by the model are hard to find in this domain. Nevertheless, Hempel (1965) who is the best known proponent of the model defended its significance for the social sciences and even for history. He considered any deviation from the standards set by the model as leading to insufficient and incomplete explanations. Nowadays, few social scientists are willing to share his view. But the idea of a *nomological* social science with strict laws and explanations of the deductive-nomological type is still alive (cf McIntyre, 1996).

The trust in the possibility of a change in the direction of a nomological social science is based on a certain feature of scientific explanations, viz., that any explanation of an event is an explanation of that event *under a certain description*. The explanandum statement *E* in a deductive-nomological explanation does not represent the event to be explained as such, but the event under a certain description. It might be that the current situation in the social sciences is the consequence of an unfavorable choice of modes of description – that, at least, is the argument. Other choices which would lead to a redescription of the subject matter of the social sciences could, perhaps, reopen the way to a truly nomological social science. Unfortunately, McIntyre (1996) who is a confirmed supporter of this view has not produced any laws, offered any new social theories, or even given any redescriptions so far.

The Statistical-Relevance Model of Probabilistic Explanation

It may be difficult to find genuine nomological laws in the social sciences, but it is easy to find statements which describe *statistical regularities* between events. '85 percent of the persons who have a panic disorder and who undergo an exposure therapy experience relief from their symptoms' is an example. Concrete percentages or probabilities do not occur in every statistical statement. Instead, fuzzy terms like 'nearly all,' 'most,' 'the majority,' 'more than half,' etc., are used. Hempel (1965) had recognized this situation very early and broadened the scope of his theory of explanation by proposing his much discussed *model of inductive-statistical explanation*.

Referring to the above example of a statistical law, an explanandum *E* could be 'Person *p* experiences relief from his or her symptoms of panic.' Appropriate antecent conditions would be A_1 'Person *p* has a panic disorder' and A_2 'Person *p* undergoes an exposure therapy.' The statistical law in conjunction with these two antecedent conditions constitutes

the explanans in this case. The inductive-statistical explanation of *E* could be expressed in two ways: (1) the inductive probability of the explanandum given the explanans amounts to 0.85. (2) The explanans inductively supports the explanandum in the degree 0.85 (cf Schurz, 1995/96).

The objections against Hempel's model of inductive-statistical explanation led Salmon (cf 1989) to the formulation of his *statistical-relevance model of probabilistic explanation*. Whereas in Hempel's model an explanation of an event is an inductive argument that confers upon the event to be explained a high inductive probability, an explanation of an event according to Salmon's model is an assemblage of factors that are statistically relevant to the occurrence of the event to be explained, accompanied by an associated probability distribution.

Salmon and Salmon (1979) gave the following example to clarify this notion: Albert, an American teenager, is convicted of stealing a car. A sociologist who attempts to explain that act of delinquency will presumably look for certain factors that are known to be relevant to delinquent behavior, e.g., sex, age, marital status of parents, religious background, socioeconomic status, residence (rural/urban/suburban), etc. Each of those factors, as well as many others, is assumed to be statistically relevant to delinquency in the sense that the probability of delinquency is different for boys than for girls, different for urban dwellers than for rural youngsters, different for children whose parents have remained married than for those from broken homes, etc. If any of these factors does not make some difference in the probability of delinquency, it will be discarded as irrelevant, and hence as lacking in explanatory value.

To explain Albert's delinquent behavior, the population of American teenagers to which Albert belongs is partitioned into a series of subclasses in terms of such relevant factors as those already mentioned. For each subclass in the resulting partition, the probability for delinquent behavior on the part of members of that subclass will be determined. Then, Albert will be assigned to one such subclass in terms of his actual attributes, e.g., 17-year-old male, suburban dweller, middle-class undivorced parents, etc., and the probability will be cited for that subclass. If the number of relevant factors which enter into the partition is large, the probability obtained in a given case may be quite low. For Salmon, that is not crucial. What matters is that the relevant factors be taken into account in partitioning, and that the probabilities associated with the various subclasses be accurate.

This model is confronted with two problems. (1) The requirement of a homogeneous relevant partition of the reference class (in the example: the population of American teenagers) is not easy to meet. In most cases in which explanations of single events are looked for in the social sciences, a comparable probability distribution will not be available. (2) The fact that an event counts as appropriately explained even if the probability value associated with the respective subclass of the partition may be very low comes into conflict with deeply rooted preconceptions about proper explanations. With regard to Tuomela's explanation seeking questions, the best that could be achieved by this kind of explanatory efforts is an answer to a question of type Q7, i.e., an answer to a how possible question.

These problems of the statistical-relevance model of explanation led others to challenge its claim to be a model for the explanation of single events. Schurz (1995/96) comes to the conclusion that the genuine explananda of statistical laws are not single events but sample frequencies. In his discussion of the explanatory value of causal modeling techniques in the social sciences, Woodward (1989) assumes that the statistical information in the delinquency example may contribute to the explanation of facts about population level parameters but not to the explanation of why any particular boy became a delinquent.

Even if this critique of Salmon's model were to be accepted, it would still be of value to the social sciences, since it provides a standard for a deeper statistical-causal understanding of social phenomena as, for example, juvenile delinquency, unemployment, or terrorism in certain reference classes. This standard could also be applied to the increasing use of meta-analytic methods for explanatory purposes in the social sciences when several explanatory variables are involved at the same time (cf Cook et al., 1994).

The Model of Aleatory Explanations

A more recent model of causal explanations of specific events is the model of aleatory explanation (see **Table 1**) introduced by Humphreys (cf 1989). He agrees with Salmon that causal explanations are possible within the realm of chancy, or aleatory, phenomena. But in his model, in contrast to Salmon's, no probability value is assigned to an explanation. The demands on the knowledge base are much lower than in Salmon's model, e.g., no complex probability distribution on the basis of a homogeneous relevant partition of a reference class has to be available.

What is required for a proper explanation of a specific event Y in S at t are two lists of causes of Y, i.e., a list of *contributing* and a list of *counteracting causes*. In Humphreys' model, for something to be a cause it must invariantly produce its effect. But causes in this model are probabilistic causes, and they produce changes in the value of the *chance* of the effect. A contributing cause of Y produces an increase, a counteracting cause of Y a decrease in the value of the chance of Y.

This model seems to be applicable in various contexts within the social sciences (cf Henderson, 1993). It takes account of the fact that social phenomena are usually the result of multiple causal influences. It does not presuppose the existence of complete lists of all influences which, positively or negatively, affect a given outcome. Aleatory explanations are conjunctive. They can be improved by including additional probabilistic causes which may come up in further research.

Table 1 The canonical form for casual explanations of specific events (cf Humphreys 1989, pp. 286 f.)

Request for the explanation:
What is the explanation of Y in S at t?
Appropriate explanation:
Y in S at t [occurs, was present] because of Φ despite Ψ
Notes
'Y' is a term referring to a property or change in property.
'S' is a term referring to a system.
't' is a term referring to a time.
'Φ' is a (non empty) list of terms referring to contributing causes of Y.
'Ψ' is a (possibly empty) list of terms referring to counteracting causes of Y.

The model has already been applied to explanations within a theory of social interaction in small groups (cf Westmeyer, 1996). But its importance goes well beyond that particular context. Take, for example, the interactionistic theory-frame of Magnusson (1980) under which many theories of social interaction can be subsumed. One fundamental assumption of this theory is that changes in the probabilities for behaviors of a person in a certain situation are a function of changes in the factors within this situation. Such determining within-situation factors are the physical situation, behaviors of the person concerned, and behaviors of the interaction partners.

In terms of Humphreys' model, 'Y' would refer to a change in the probability of a behavior of the person concerned; 'S' would refer to the setting in which the intrasituational dynamic interaction between the person and the interaction partners takes place; 't' would refer to a certain point or period in time at which the change in the probability of the behavior occurs; 'Φ' would refer to the set of all those determining within-situation factors, the changes of which facilitate (contribute to) the occurrence of Y; and 'Ψ' would refer to the set of all those determining within-situation factors, the changes of which inhibit (act counter to) the occurrence of Y. Contributing and counteracting causes could be identified by methods of event sequence or time series analysis.

The Embedding Theory of Explanation

Most of the models and notions of explanation dealt with so far are strongly connected to the so-called *statement view* of scientific theories that assumes theories to be sets of laws or theoretical assumptions. A very different picture is given by the *structuralist view* which is a variant of the model-theoretic approach to the understanding of scientific theories. Here, theories are reconstructed as complex set-theoretical structures. The structuralist program has already demonstrated its fruitfulness in the analysis of various theories from the social sciences (cf Stegmüller et al., 1982; Westmeyer, 1992).

From this point of view, explaining a phenomenon means to explain it *within the context of a certain theory*. This requires to show (1) that the phenomenon can be construed as an element of the set of intended applications of the theory, and (2) that the theory can be successfully applied to that element. In short, what has to be done is to embed the phenomenon into the theory. This leads Bartelborth (1996) who gives the most detailed account of this idea to speak of the *embedding theory of explanation*. For a concrete example of this type of explanation in the social sciences see Westmeyer (1996). He shows that it is possible to combine the embedding of a phenomenon into a certain theory with the application of Humphreys' model of aleatory explanation.

The embedding theory is the most promising variant of the *unification approach* (cf Kitcher, 1989) to the problem of explanation. 'Explaining as unifying' is a notion which increasingly attracts the attention of philosophers of science. It paves the way for an integrative treatment of explanation and understanding (cf Schurz and Lambert, 1994) which is especially of interest to those social scientists who are used to construe explaining and understanding as basically different scientific activities. It is in accordance with a coherence theory of truth and fully compatible with basic epistemological tenets of the more recent constructionist and constructivist approaches to the social sciences. These characteristic features of the embedding theory give rise to the expectation that this theory will play a leading role in future discussions of the issue of explanation in the social sciences.

See also: Classification, Understandings of in the Social Sciences; Cognitive Modeling: Connectionist Approaches; Constructivism and Constructionism: Methodology; Event Sequence Analysis: Methods; Functional Explanation: Philosophical Aspects; Generalization: Conceptions in the Social Sciences; Historical Explanation, Theories of: Philosophical Aspects; Interpretive Methods: Macromethods; Interpretive Methods: Micromethods; Rational Choice Explanation: Philosophical Aspects.

Bibliography

Bartelborth, T., 1996. Begründungsstrategien [Strategies of Justification]. Akademie Verlag, Berlin.
Bunge, M., Ardila, R., 1987. Philosophy of Psychology. Springer-Verlag, New York.
Buss, D.M., 1997. Evolutionary foundations of personality. In: Hogan, R., Johnson, J., Briggs, S. (Eds.), Handbook of Personality Psychology. Academic Press, San Diego, CA.
Cook, T.D., Cooper, H., Cordray, D.S., Hartmann, H., Hedges, L.V., Light, R.J., Louis, T.A., Mosteller, F., 1994. Meta-analysis for Explanation: A Casebook. Russell Sage Foundation, New York.
Hempel, C.G., 1965. Aspects of Scientific Explanation. Free Press, Albany, New York.
Henderson, D.K., 1993. Interpretation and Explanation in the Human Sciences. State University of New York Press, New York.
Humphreys, P.W., 1989. Scientific explanation: The causes, some of the causes, and nothing but the causes. In: Kitcher, P., Salmon, W.C. (Eds.), Scientific Explanation. University of Minnesota Press, Minneapolis, MN, pp. 283–306.
Kitcher, P., 1989. Explanatory unification and the causal structure of the world. In: Kitcher, P., Salmon, W.C. (Eds.), Scientific Explanation. University of Minnesota Press, Minneapolis, MN, pp. 410–505.
Kitcher, P., Salmon, W.C. (Eds.), 1989. Scientific Explanation. University of Minnesota Press, Minneapolis, MN.
Little, D., 1991. Varieties of Social Explanation. Westview Press, Boulder, CO.
Magnusson, D., 1980. Personality in an interactional paradigm of research. Zeitschrift für Differentielle und Diagnostische Psychologie 1, 17–34.
McIntyre, L.C., 1996. Laws and Explanation in the Social Sciences. Westview Press, Boulder, CO.
Salmon, W.C., 1989. Four decades of scientific explanation. In: Kitcher, P., Salmon, W.C. (Eds.), Scientific Explanation. University of Minnesota Press, Minneapolis, MN, pp. 3–219.
Salmon, W.C., 1998. Causality and Explanation. Oxford University Press, New York.
Salmon, M.H., Salmon, W.C., 1979. Alternative models of scientific explanation. American Anthropologist 81, 61–74.
Schurz, G., 1995/96. Scientific explanation: A critical survey. Foundations of Science 3, 429–465.
Schurz, G., Lambert, K., 1994. Outline of a theory of scientific understanding. Synthese 101, 65–120.
Stegmüller, W., 1983. Erklärung–Begründung–Kausalität [Explanation–Justification–Causality], second ed. Springer-Verlag, Berlin.
Stegmüller, W., Balzer, W., Spohn, W. (Eds.), 1982. Philosophy of Economics. Springer-Verlag, New York.
Tuomela, R., 1977. Human Action and its Explanation. D. Reidel, Boston, MA.
Westmeyer, H. (Ed.), 1992. The Structuralist Program in Psychology: Foundations and Applications. Hogrefe & Huber Publishers, Seattle, WA.
Westmeyer, H., 1996. A concept of explanation for social interaction models. In: Hegselmann, R., Mueller, U., Troitzsch, K.G. (Eds.), Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View. Kluwer, Dordrecht.
Woodward, J., 1989. The causal mechanical model of explanation. In: Kitcher, P., Salmon, W.C. (Eds.), Scientific Explanation. University of Minnesota Press, Minneapolis, MN, pp. 357–383.

Explanatory Style and Health

Athanassios Douzenis, Attikon University Hospital, Athens University Medical School, Athens, Greece
Dionysis Seretis, Attikon University Hospital, Athens University Medical School, Athens, Greece

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C. Peterson, volume 13, pp. 5162–5164, © 2001, Elsevier Ltd.

Abstract

Explanations or attributions of somatic feelings or events have mostly been investigated in reference to mental disorders such as panic disorder and hypochondriasis. These cognitions have salient effects on the diagnosis, management, and course of disorders according to various evidence-based psychological models. Each person experiences the events that happen to him/her and attempts to understand them. Explanatory style refers to the way individuals understand for themselves why an event, positive or negative, has happened. This understanding has many facets and the way people perceive an external event (or a bodily sensation) is influencing their reaction to it and consequently has an effect on them. Explanatory style is considered a psychological attribute and is a term used alternatively to attribution style. This sometimes causes confusion. Explanatory style, as defined, can influence diagnosis and treatment as well as illness behavior.

Explanatory Style and Treatment

Whereas attribution theories have their roots in social psychology – and go back to the seminal work of Fritz Heider (2013), early twentieth century – various models have been developed in the context of clinical psychology with a view to investigating the processes by which individuals attempt to explain the causes of behavior and events. Explanatory style has mostly been investigated in the context of certain mental disorders e.g., depression, phobias, panic disorder, hypochondriasis, and other somatic symptom disorders. In hypochondriasis, the belief that one suffers from a serious medical problem largely develops from the fact that patients misinterpret physical sensations, as symptoms of a serious disease. These perceptions lead to help-seeking behaviors, which typically aim to receive a diagnosis and potentially result in appropriate or inappropriate medical reassurance, information about the diagnosis, treatment, or further referral for specialist assessment and/or more medical tests. If the problem is not resolved, despite its short-term benefits in reduction of illness worry, diagnosis or reassurance acts as a maintaining factor for anxiety, allowing for the development of new seeking behaviors (Salkovskis, 1991, 1996) – which may in turn result in increased levels of medical consultations and investigations (Conroy et al., 1999). Kirmayer and Taillefer (1997) also suggest that attribution of physical sensations to physical causes results in help seeking. Interestingly, they argue misattribution alone is crucial toward the development of somatic symptom disorders as it deprives one of turning to oneself in order to seek alternative explanations and develop coping behaviors.

The explanatory style of an individual influences his/her adherence to treatment (Long et al., 2007). It has been shown in clinical practice that patients' compliance to the prescribed treatment (be it pharmaceutical or other health advice) is crucial in order to ensure recovery. Nonadherence incurs clinical failures and consequently financial burdens, not only for the individual but also for the health system and society as a whole. Clinicians attempting to reduce noncompliance have developed general and specific interventions. 'Compliance therapies' have been trialed by clinicians, based on cognitive behavioral techniques. Trained clinicians deliver specific therapeutic interventions that emphasize medication cessation and relapse, normalizing rationales for stigma and recognition of characteristic prodromal symptoms (Rief and Rojas, 2007). One of the most frequently used approaches involves meetings with the doctor, aiming to increase positive attitudes toward treatment, recognition of side effects and reporting them. It is widely acknowledged that among the most important parameters in achieving compliance is the doctor–patient relationship, i.e., the therapeutic alliance. In this respect, it has been shown that standard encounters with the physician (the psychiatrist in this particular research) involving nonspecific counseling, thereby stimulating positive attitudes toward treatment, can be as effective as more specific interventions for noncompliance in the long term (Jackson and Kroenke, 2008). Undisputedly, compliance is more difficult to achieve in the treatment of chronic (long term) disorders (Zolnierek and DiMatteo, 2009).

Explanatory Style: Cognitive Processes and Behavior

When someone suffers from an illness, the symptoms of the illness can influence his or her behavior. Additionally individuals suffering from a somatic illness are ascribed the role of 'the ill person,' which implies relief from certain everyday chores (absence from work support from their social milieu). Barsky and colleagues (Barsky and Wyshak, 1990; Barsky, 1992) ascribed a central role to cognitive processes including the process of perceiving somatic sensations as signs of a disease. They postulated that illness anxiety increases as a result of a vicious cycle of misattribution, selective attention to physical symptoms and 'amplification' of interoception. This amplification of interoceptive awareness, i.e., of perception of somatic sensations and/or of the physiological condition of the body (Craig, 2002) eventually leads to the intensification of physical symptoms. It is also worth noting that illness worry or health anxiety, which is now considered a dimensional

construct (Longley et al., 2010), has been linked to catastrophizing, avoidance, body vigilance, and attribution (explanation) (Rief and Broadbent, 2007; Henningsen et al., 2011).

Illness anxiety and illness behavior are studied extensively in the context of medically unexplained symptoms (MUS) and Somatic Symptom and Related Disorders. Several studies on illness perception have shown that patients' beliefs about the cause of their illness (i.e., somatic, psychological, normalizing/environmental explanations) play an important role in pathogenesis and maintenance of MUS and Somatic Symptom and Related Disorders such as Somatic Symptom Disorder, Illness Anxiety Disorder, and Conversion Disorder (Voigt et al., 2010). A number of these studies have reported that patients with MUS/somatic symptom disorders tend toward an organic causal illness perception style (MacLeod et al., 1998; Nimnuan et al., 2001; Rief et al., 2004; Henningsen et al., 2005; Douzenis and Seretis, 2013). However, many patients report multiple symptom attributions – including psychological and social – and employ complex explanatory resources to account for their symptoms (Rief and Broadbent, 2007; Douzenis and Seretis, 2013). Finally, research on the role of memory is bound to have important implications in the field of explanation of illnesses and illness behavior: memory influences top–down psychological processes in the symptom perception of patients with MUS (e.g., Houtveen and Oei, 2007; McFarlane et al., 2008; Bogaerts et al., 2010) and is also involved in doctor–patient encounters (Rief and Broadbent, 2007).

See also: Chronic Pain: Models and Treatment Approaches; Doctor–Patient Interaction in the West: Psychosocial Aspects; Health Risk Perception; Illness Behavior and Care-Seeking; Interoception; Personalized Medicine; Symptom Perception, Awareness and Interpretation.

Bibliography

Barsky, A.J., 1992. Amplification, somatization, and the somatoform disorders. Psychosomatics 33, 28–34.

Barsky, A.J., Wyshak, G., 1990. Hypochondriasis and somatosensory amplification. The British Journal of Psychiatry 157, 404–409. http://dx.doi.org/10.1192/bjp.157.3.404.

Bogaerts, K., Van Eylen, L., Li, W., Bresseleers, J., Van Diest, I., De Peuter, S., Stans, L., Decramer, M., Van den Bergh, O., 2010. Distorted symptom perception in patients with medically unexplained symptoms. Journal of Abnormal Psychology 119 (1), 226.

Conroy, R.M., Smyth, O., Siriwardena, R., Fernandes, P., 1999. Health anxiety and characteristics of self-initiated general practitioner consultations. Journal of Psychosomatic Research 46 (1), 45–50.

Craig, A.D., 2002. How do you feel? Interoception: the sense of the physiological condition of the body. Nature Reviews Neuroscience 3 (8), 655–666.

Douzenis, A., Seretis, D., 2013. Descriptive and predictive validity of somatic attributions in patients with somatoform disorders: a systematic review of quantitative research. Journal of Psychosomatic Research 75 (3), 199–210.

Heider, F., 2013. The Psychology of Interpersonal Relations. Psychology Press.

Henningsen, P., Fink, P., Hausteiner-Wiehle, C., Rief, W., 2011. Terminology, classification and concepts. In: Creed, F., Henningsen, P., Fink, P. (Eds.), Medically Unexplained Symptoms, Somatisation and Bodily Distress: Developing Better Clinical Services. Cambridge University Press, Cambridge, pp. 43–68.

Henningsen, P., Jakobsen, T., Schiltenwolf, M., Weiss, M.G., 2005. Somatization revisited: diagnosis and perceived causes of common mental disorders. The Journal of Nervous and Mental Disease 193, 85–92.

Houtveen, J.H., Oei, N.Y., 2007. Recall bias in reporting medically unexplained symptoms comes from semantic memory. Journal of Psychosomatic Research 62 (3), 277–282.

Jackson, J.L., Kroenke, K., 2008. Prevalence, impact, and prognosis of multisomatoform disorder in primary care: a 5-year follow-up study. Psychosomatic Medicine 70 (4), 430–434.

Kirmayer, L., Taillefer, S., 1997. Somatoform disorders. In: Hersen, S.M.T.M. (Ed.), Adult Psychopathology and Diagnosis. John Wiley, New York.

Long, J., Orner, M.B., Rothendler, J., Reddy, S., Kozak, W., Kroupa, L.P., Berlowitz, D.R., 2007. Hypertensive patients' race, health beliefs, process of care, and medication adherence. Journal of General Internal Medicine 22 (6), 768–774.

Longley, S.L., Broman-Fulks, J.J., Calamari, J.E., Noyes, R., Wade, M., Orlando, C.M., 2010. A taxometric study of hypochondriasis symptoms. Behavior Therapy 41, 505–514. http://dx.doi.org/10.1016/j.beth.2010.02.002.

MacLeod, A.K., Haynes, C., Sensky, T., 1998. Attributions about common bodily sensations: their associations with hypochondriasis and anxiety. Psychological Medicine 28, 225–228.

McFarlane, A.C., Ellis, N., Barton, C., Browne, D., Van Hooff, M., 2008. The conundrum of medically unexplained symptoms: questions to consider. Psychosomatics 49 (5), 369–377.

Nimnuan, C., Hotopf, M., Wessely, S., 2001. Medically unexplained symptoms: an epidemiological study in seven specialities. Journal of Psychosomatic Research 51 (1), 361–367.

Rief, W., Broadbent, E., 2007. Explaining medically unexplained symptoms-models and mechanisms. Clinical Psychology Review 27, 821–841. http://dx.doi.org/10.1016/j.cpr.2007.07.005.

Rief, W., Nanke, A., Emmerich, J., Bender, A., Zech, T., 2004. Causal illness attributions in somatoform disorders: associations with comorbidity and illness behavior. Journal of Psychosomatic Research 57, 367–371.

Rief, W., Rojas, G., 2007. Stability of somatoform symptoms – implications for classification. Psychosomatic Medicine 69 (9), 864–869.

Salkovskis, P.M., 1991. The importance of behaviour in the maintenance of anxiety and panic: a cognitive account. Behavioural and Cognitive Psychotherapy 19 (01), 6–19.

Salkovskis, P.M., 1996. The cognitive approach to anxiety: threat beliefs, safety-seeking behaviours and the special case of health anxiety and obsessions. In: Salkovskis, P.M. (Ed.), Frontiers of Cognitive Therapy. Guilford Press, New York, pp. 48–74.

Voigt, K., Nagel, A., Meyer, B., Langs, G., Braukhaus, C., Löwe, B., 2010. Towards positive diagnostic criteria: a systematic review of somatoform disorder diagnoses and suggestions for future classification. Journal of Psychosomatic Research 68 (5), 403–414.

Zolnierek, K.B.H., DiMatteo, M.R., 2009. Physician communication and patient adherence to treatment: a meta-analysis. Medical Care 47 (8), 826.

Exploratory Data Analysis: Univariate Methods

David C Hoaglin, Sudbury, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The exploratory approach, developed by John W. Tukey, focuses on steps that precede any formal inferences. At its core lie four main themes: display (usually graphical), reexpression (via a mathematical transformation), resistance (to unusual behavior in data), and residuals (from subtracting a summary or fitted model). The most familiar exploratory data analysis techniques include the stem-and-leaf display, the boxplot, and median polish. The article discusses themes and techniques and illustrates their application to data in four examples.

Since about 1970, the phrase 'exploratory data analysis' (EDA) has referred to the approach and techniques promoted by John W. Tukey and various collaborators. EDA emphasizes relatively simple numerical and graphical techniques for looking at a set of data, to see what 'appearances' are present, and removing 'partial descriptions' of those appearances, in order to look further. These steps are intended to precede any attempt at drawing inferences from the data, a process that Tukey labeled 'confirmatory data analysis.' Statisticians have used exploratory techniques since they began to work with data, but extensive developments in statistical inference (roughly during the first half of the twentieth century) had relegated exploration to a secondary role as 'descriptive statistics.' Tukey sought to restore the balance between exploratory and confirmatory data analysis. In the process, mainly in his book *Exploratory Data Analysis* (1970–71, 1977), he introduced a variety of new techniques. Within a few years basic EDA techniques, especially displays, were available in statistical software. By the end of the century, a number of those techniques had become part of statistical instruction at all levels, including elementary school.

Main Themes of EDA

At the core of the EDA approach lie four main themes: display, reexpression, resistance, and residuals.

1. Displays (most of them graphical) allow the analyst to see the behavior of the data and, as analysis proceeds, the behavior of residuals and various diagnostic measures. EDA emphasizes the frequent use of displays, to ensure that unexpected features are not overlooked. The most common displays include the stem-and-leaf display, the boxplot, scatterplots of residuals, and the spread-versus-level plot.
2. Reexpression involves applying a single mathematical transformation (such as the square root or the logarithm) to each data value. The aim is to simplify, and thus make more effective, the analysis of the data. Depending on the structure of the data, the simplification may come from promoting symmetry, constancy of variability, straightness of relationship, or additivity of effect. EDA asks, at nearly the first opportunity, whether the scale in which the data are presented is satisfactory. If not, it freely tries a likely transformation.

   The reexpressions most often used in EDA come from the family of functions known as power transformations, which take y into y^p (almost always with a simple value of p such as 1/2, −1, or 2), together with the logarithm (which, for data analysis purposes, fits into the power family at $p = 0$). For example, in a study of the levels of nicotine in the blood and urine of smokers and nonsmokers (Russell and Feyerabend, 1975), the analysis gained simplicity by reexpressing the nicotine concentrations in the logarithmic scale.
3. Resistance provides insensitivity to misbehavior or unusual behavior in data. An analysis or summary is resistant if an arbitrary change in any small part of the data produces only a small change in the analysis or summary. EDA emphasizes resistance because experience suggests that 'good' data seldom contain less than about 5% gross errors.

   The notion of robustness is akin to resistance, but not the same. Robustness generally implies insensitivity to departures from assumptions surrounding an underlying probabilistic model. (Resistance is one aspect of 'qualitative robustness.')

   In summarizing the location of a sample, the median is highly resistant. (In terms of efficiency, it is not so highly robust because other estimators achieve greater efficiency across a broader range of distributions.) By contrast, the mean is highly nonresistant. A number of exploratory techniques for more structured forms of data, such as the 3-group resistant line for (x,y) data and median polish for two-way layouts, provide resistance because they are based on the median.
4. Residuals are what remain of data after a summary or fitted model has been subtracted out, according to the schematic equation

$$\text{residual} = \text{data} - \text{fit}$$

   For example, if the data are the pairs (x_i, y_i) and the fit is the line $\widehat{y}_i = a + bx_i$, then the residuals are $r_i = y_i - \widehat{y}_i$.

   EDA regards a careful examination of the residuals as an essential step in analyzing a set of data. This emphasis gains from the tendency of resistant analyses to provide a clear separation between dominant behavior and unusual behavior in the data. When the bulk of the data follows a consistent pattern, that pattern determines a resistant fit. The residuals then contain any drastic departures from the pattern, as well as the customary chance fluctuations.

Unusual residuals suggest a need to check on the circumstances surrounding those observations. As in more traditional practice, the residuals can warn of systematic difficulties with the data – curvature, nonadditivity, and nonconstancy of variability.

The examples and discussion that follow illustrate displays (via stem-and-leaf displays and boxplots), resistance (median polish), residuals, and reexpression.

Example 1: Florida Voting (Stem-and-Leaf Displays)

The initial results of the US presidential election in 2000 focused intense attention on the state of Florida, where the difference in total votes cast for George W. Bush and Albert Gore was so small that state law required a recount and thus delayed the outcome. Among the unusual features of the Florida voting was the relatively large number of votes cast in Palm Beach County for Patrick Buchanan, the candidate of the Reform Party (3407 out of 430 762 votes). Critics argued that most of those votes had been intended for Gore, but were recorded for Buchanan because of a problem with the type of ballot.

A number of analyses compared the results in Palm Beach County against those in other counties. Without explaining the reasons for those choices, one analyst selected the data from the 30 counties in which Bush received more than 30 000 votes and calculated the ratio of the number of Buchanan votes to the number of Bush votes, concluding that the probability of the ratio in Palm Beach County by pure chance was extremely small.

For those 30 counties Figure 1 shows a stem-and-leaf display of the percentage of votes for Buchanan (out of the total number of votes for four candidates: Bush, Gore, Buchanan, and Ralph Nader). In overall outline this display resembles a histogram with an interval width of 0.1%, but it shows more detail of the data values by using the 0.01% digit of each percentage instead of merely enclosing a uniform amount of area. This feature makes it easy to go back from a part of the display to the individual data values (and their identities) in the listing of the data – a valuable step in EDA.

Several alternatives and refinements permit the stem-and-leaf display to accommodate a wide variety of data patterns. The interval width (that is, the range of data values represented by one line in the display) may be 2, 5, or 10 times a power of 10; and stray observations may be listed separately at each end so as not to distort the scale. The use of positive and negative stems readily handles sets of residuals.

```
0|9                Dade
1|33456679
2|13445677
3|0222778
4|023
5|5               Marion
6|2               Santa Rosa
7|9               Palm Beach
```

Figure 1 Stem-and-leaf display of the percentage of votes cast for Patrick Buchanan in the 30 Florida counties in which George W. Bush received more than 30 000 votes. The unit for the leaves is 0.01%.

In Figure 1 Palm Beach County's 0.79% for Buchanan is noticeably high (indeed, another EDA technique would flag it as 'outside' and hence deserving investigation). But the 0.55% for Marion County and the 0.62% for Santa Rosa County are also noticeable at the high end. The layout of the stem-and-leaf display makes it easy to include tags that identify a few data values at either end.

The analyst's choice of counties with more than 30 000 votes for Bush naturally raises a question about the extent of support for Buchanan in the other 37 counties of Florida. Figure 2 shows a stem-and-leaf display for all 67 counties. In the context of the whole state, Palm Beach County no longer stands out; seven counties had a percentage for Buchanan higher than 0.79%, and two of those were above 1.6%. Those seven counties, however, are small; most cast fewer than 8000 total votes for the four candidates. Also, geographically, they are all located in north Florida, far from Palm Beach County. Thus, among counties with larger numbers of voters, it still appears that Palm Beach County produced an unusually large percentage for Buchanan. A stem-and-leaf display of the vote totals (not shown) reveals a clear gap around 40 000. Modifying Figure 1 to include all counties with totals greater than 40 000 would add three counties, whose percentages for Buchanan were 0.21%, 0.26%, and 0.47%.

Even after this modification the position of Palm Beach County in Figure 1 does not form the basis for a conclusion. It does invite further examination. Other analysts responded, taking into account the voting in counties throughout the United States, precinct level voting in Palm Beach County, and voting in other Florida elections. For example, even if one uses only data from Florida counties, a scatterplot of the percentage for Buchanan against the percentage of Bush (or against the percentage of Gore) shows that Palm Beach County is very far from the other counties.

```
 0|9                  Dade
 1|334456679
 2|1134456666777
 3|00222677889
 4|0023367788
 5|1133456
 6|12225
 7|1569
 8|79
 9|
10|4                  Holmes
11|06                 Washington, Gulf
12|
13|
14|
15|
16|3                  Liberty
17|4                  Calhoun
```

Figure 2 Stem-and-leaf display of the percentage of votes cast for Buchanan in all 67 Florida counties. The unit for the leaves is 0.01%.

Example 2: Teen Drivers (Stem-and-Leaf Display, Boxplots)

In the United States drivers under 20 years of age are involved in fatal motor vehicle crashes at a much higher rate than older drivers. An analysis of data for 2003 calculated the rate of fatal crashes (per 100 000 licensed drivers of 16–19 years of age) for each state and the District of Columbia (O'Donnell, 2005). The stem-and-leaf display in Figure 3 shows large differences among states (New Hampshire was lowest at 29, and North Carolina was highest at 105), and labeling other low rates and high rates suggests a geographic pattern.

The broad grouping of the four census regions (Northeast, 9 states; Midwest, 12 states; South, 16 states plus DC; and West, 13 states) provides the basis for the parallel boxplots in Figure 4. In a boxplot (also called a 'schematic plot') the box extends from the lower hinge (an approximate quartile) to the upper hinge and has a line across it at the median. The 'whiskers' above and below the box show the extent of the data, except for observations that are likely strays (defined according to a rule based on the hinges). Such strays (when present) appear individually, in order to focus attention on them. The general intent is to indicate the median, outline the middle half of the data, and show the range, with more detail at the ends if needed.

One message of Figure 4 is that the typical (median) rates of fatal crashes were lowest in the Northeast and highest in the South with the Midwest and the West roughly halfway between. The vertical extent of the boxes shows greater variation in the Northeast and the South, but the Midwest and the West each have one stray. In the Northeast the short whiskers (relative to the extent of the box) and the low position of the median in the box suggest displaying the regions' data in detail.

The plot in Figure 5 uses the states' two-letter abbreviations as the plotting symbols. In the Northeast Rhode Island, Pennsylvania, and Maine had substantially higher rates than the other states. In the Midwest the stray is Missouri. In the South the highest rate (105 in North Carolina) was nearly twice the lowest (55 in Maryland), with no sizable gaps in between. And in the West, Utah was noticeably low, and Nevada, Arizona, and Montana were noticeably high. With such an initial exploration as background, an analysis might consider such factors as restrictive licenses for teen drivers (especially 16-year-olds), use of safety belts, enforcement of speed limits and alcohol laws, and characteristics of roads.

```
 2*|
 2.|9              NH
 3*|3              CT
 3.|7899           VT,MA,UT,NJ
 4*|33             NE,NY
 4.|
 5*|0022244
 5.|5666789
 6*|0011124
 6.|777
 7*|234            AL,TN,PA
 7.|557            AR,SC,TX
 8*|0134           MT,RI,AZ,WV
 8.|689            MO,GA,NV
 9*|0122           KY,MS,FL,DE
 9.|8              DC
10*|
10.|5              NC
```

Figure 3 Stem-and-leaf display of states' rates of fatal crashes in 2003 involving teen drivers (per 100 000 licensed drivers of 16–19 years of age). The unit for the leaves is 1.

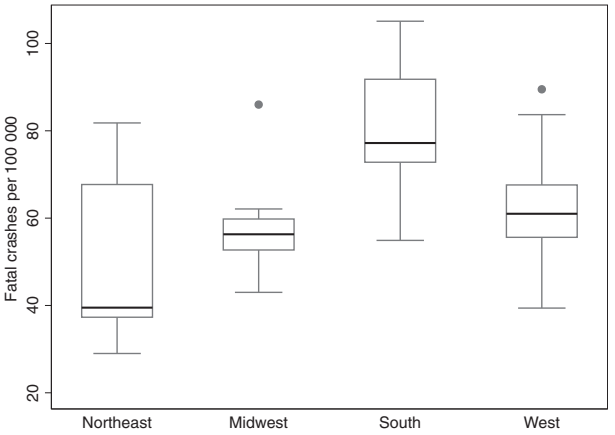

Figure 4 Parallel boxplots of the rates of fatal crashes by census region.

Example 3: Hotel Occupancy (Median Polish, Residuals)

In 1983, as several new hotels were under construction in Boston, operators of existing major hotels were concerned about the impact on their business. Table 1 shows the occupancy rates for 11 major Metropolitan Boston hotels, month by month, for January through June 1983. One approach to exploring such a two-way table uses an additive fit of the form

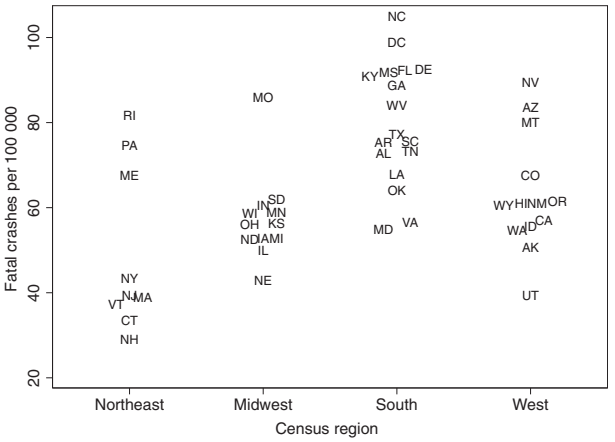

Figure 5 Plot of the rates of fatal crashes for all states by census region.

Table 1 Occupancy rates of major Metropolitan Boston hotels, 1983 (%)

Hotel	*January*	*February*	*March*	*April*	*May*	*June*
Bostonian	56	65	71	86	87	93
Colonnade	30	24	34	73	74	80
Copley Plaza	52	62	80	84	83	84
Hyatt	52	71	72	89	86	90
Lenox	38	45	55	71	78	78
Marriott Long Wharf	77	84	86	95	94	96
Meridien	54	63	61	73	81	82
Park Plaza	40	42	47	64	72	72
Parker House	57	60	60	78	82	85
Ritz-Carlton	45	53	52	75	83	82
Sheraton Boston	49	37	60	78	85	82

$$\widehat{y}_{ij} = m + a_i + b_j$$

with a row effect, a_i, for each hotel and a column effect, b_j, for each month. The EDA technique known as median polish arrives at such a fit by iteratively subtracting row medians and column medians, first from the table of data and then from the table left by the previous iteration. The common value, m, is determined by centering the effects so that the median of the a_i is 0 and the median of the b_j is 0. The use of the median gives the fit resistance: isolated behavior of the data in a few cells of the table will tend to have only a minor impact on the fit.

For the hotel occupancy rates **Table 2** shows the two-way table of residuals, $y_{ij} - \widehat{y}_{ij}$, bordered by the hotel effects, the month effects, and the common value. The month effects (ranging from −22.0 to +15.0) show occupancy rates in Boston hotels increasing substantially from January to June. Among the hotel effects the Marriott Long Wharf (at 21.5) stands out as having higher occupancy than the other hotels, and the Colonnade and the Park Plaza (at −12.0 and −12.5, respectively) were substantially lower. A stem-and-leaf display of the residuals (not shown) calls attention to three sizable negative residuals (−20.0, −19.0, and −15.0) and one sizable positive residual (+13.5). These indicate that the Colonnade (beyond its generally low occupancy) had unusually low occupancy in February and March, that the Sheraton Boston had unusually low occupancy in February, and that the Copley Plaza had unusually high occupancy in March.

Example 4: Everyday Reexpressions

In a number of applications from a variety of fields, a reexpression has become so customary that no one mentions it. By the time the data or results are presented, they are already in the transformed scale. Examples include established relationships in science and engineering, as well as contemporary data (see also Hoaglin, 1988).

For earthquakes the Richter scale expresses the strength in logarithmic units (base 10). The magnitude is defined as the logarithm of the amplitude recorded by a particular type of seismograph (at a specified distance from the epicenter).

In measuring the intensity of sounds, the decibel scale is logarithmically related to sound pressure.

Measurements of the fuel consumption of automobiles often involve determining the number of gallons of gasoline that a car used on a standard test course. Thus, the ratings in miles per gallon are the result of a reciprocal transformation.

Literature

The first published presentation of EDA was the preliminary edition (1970–71) of the book by John W. Tukey, whose 1977 edition represents the definitive account of the subject. Mosteller and Tukey (1977) also include substantial discussion of exploratory attitudes and techniques. Tukey's (1962) paper on the future of data analysis set the stage.

The two volumes of the *Statistician's Guide to Exploratory Data Analysis* (Hoaglin et al., 1983, 1985) discuss the rationale and development of a number of EDA techniques, explain and illustrate connections with classical statistical theory, and introduce new techniques.

Hoaglin et al. (1991) extend the basic attitudes and techniques of EDA to factorial analysis of variance.

Broader discussions of the roles of exploratory and confirmatory data analysis in scientific inquiry appear in Tukey (1972, 1980).

Table 2 Result of applying median polish to the hotel occupancy rates: table of residuals, bordered by effects and common value

Hotel	*January*	*February*	*March*	*April*	*May*	*June*	*Hotel effect*
Bostonian	−0.0	0.0	1.0	−0.0	−3.0	0.0	9.0
Colonnade	−5.0	−20.0	−15.0	8.0	5.0	8.0	−12.0
Copley Plaza	−0.5	0.5	13.5	1.5	−3.5	−5.5	5.5
Hyatt	−3.5	6.5	2.5	3.5	−3.5	−2.5	8.5
Lenox	−3.0	−5.0	0.0	0.0	3.0	0.0	−6.0
Marriott	8.5	6.5	3.5	−3.5	−8.5	−9.5	21.5
Meridien	7.0	7.0	0.0	−4.0	0.0	−2.0	−0.0
Park Plaza	5.5	−1.5	−1.5	−0.5	3.5	0.5	−12.5
Parker House	9.0	3.0	−2.0	−0.0	−0.0	0.0	1.0
Ritz-Carlton	0.0	−1.0	−7.0	0.0	4.0	0.0	−2.0
Sheraton Boston	2.0	−19.0	−1.0	1.0	4.0	−2.0	0.0
Month effect	−22.0	−13.0	−8.0	8.0	12.0	15.0	69.0

See also: Correspondence Analysis; Multidimensional Scaling I; Multidimensional Scaling II; Nonparametric Regression; Robustness.

Bibliography

Hoaglin, D.C., 1988. Transformations in everyday experience. CHANCE 1 (4), 40–45.
Hoaglin, D.C., Mosteller, F., Tukey, J.W., 1983. Understanding Robust and Exploratory Data Analysis, Wiley Classics Library ed. 2000. Wiley, New York.
Hoaglin, D.C., Mosteller, F., Tukey, J.W., 1985. Exploring Data Tables, Trends, and Shapes. Wiley, New York.
Hoaglin, D.C., Mosteller, F., Tukey, J.W., 1991. Fundamentals of Exploratory Analysis of Variance. Wiley, New York.
Mosteller, F., Tukey, J.W., 1977. Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley, Reading, MA.
O'Donnell, J., March 1, 2005. Deadly teen auto crashes show a pattern. USA Today 1A, 6A.
Russell, M.A.H., Feyerabend, C., 1975. Blood and urinary nicotine in non-smokers. Lancet 305, 179–181.
Tukey, J.W., 1962. The future of data analysis. Annals of Mathematical Statistics 33, 1–67. 812.
Tukey, J.W., 1970–71. Exploratory Data Analysis, limited preliminary ed. Addison-Wesley, Reading, MA [available from University Microfilms].
Tukey, J.W., 1972. Data analysis, computation, and mathematics. Quarterly of Applied Mathematics 30, 51–65.
Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, MA.
Tukey, J.W., 1980. We need both exploratory and confirmatory. The American Statistician 34, 23–25.

Exposure Therapies and Stress Inoculation: A Brief Overview

David Phillips and Dean Lauterbach, Eastern Michigan University, Ypsilanti, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Exposure therapies represent the most empirically supported treatments for anxiety and related disorders. Furthermore, stress inoculation training (SIT) uses exposure techniques to inoculate at-risk populations from anticipated stress. Unique iterations of exposure exist for each disorder, population, and patient, yet foundational elements span diagnostic categories. In the broadest sense, exposure therapies are a group of treatments that use the presentation of feared stimuli as a therapeutic tool. This article addresses the theoretical models, the common elements of practice, limitations, and future directions of exposure therapy.

Theoretical Models of Exposure Therapy

A variety of theoretical models attempt to explain the etiology of pathology and the efficacy of exposure therapies. Both cognitive and neurological explanations have been well articulated, but behavioral conditioning, operant and classical, is exposure's cornerstone. As such, the following section details the two most influential conditioning models: Mowrer's two-stage theory and the emotional processing model (EPM).

Mowrer's Two-Stage Theory

Mowrer's (1960) two-stage theory for the acquisition and maintenance of fear incorporates both classical and operant principles. The first stage of Mowrer's model suggests that neutral stimuli become feared stimuli when paired with already feared unconditional stimuli. The 'Little Albert' experiment (Watson and Rayner, 1920) is a classic example of this process. In a laboratory setting, 'Albert,' an infant, is presented with a white rat, which initially provokes no fear (the neutral stimulus). Fear of the rat (the conditioned stimulus) quickly develops when Watson and Rayner (1920) begin pairing the presentation of the rat with an alarming sound (i.e., the unconditioned stimulus). Mowrer argues that this process is adaptive when it allows organisms to generalize conditioning to situations that should be avoided, but this process becomes dysfunctional when the organism overgeneralizes to safe situations.

Across anxiety-related disorders, conditioned stimuli elicit distressing emotional responses that the organism actively avoids/seeks to escape (Richard et al., 2007). Therefore, avoidance behavior comes under operant control; the second stage of Mowrer's theory (1960). A common example is avoiding all forms of transportation following a traumatizing car accident. Automotive transportation (the conditioned stimulus) is avoided because it is associated with the fear of the collision (the unconditioned stimulus). Avoiding all forms of transportation mitigates anxiety, which is reinforcing in the moment but dysfunctional in the long term.

Although Mowrer's theory does not address how exposure works (Richard et al., 2007), the two-stage theory implies behavioral interventions that alter operant contingencies and the valence of conditioned stimuli (Forsyth et al., 2007); two key components to exposure therapies. This being said, the two-factor theory is not the primary basis of any modern exposure therapy. Critics of Mowrer's theory point to instances where individuals acquire fear responses but do not engage in avoidance behavior, or exhibit avoidance behavior despite reporting the absence of subjective distress (Richard et al., 2007). Also, many individuals diagnosed with anxiety or related disorders cannot recall a learning event that accounts for the first stage (i.e., the classical conditioning stage). Lastly, some argue that the two-stage theory does not adequately account for the complexities of nonphobic fears that often have intrapersonal and social consequences (Richard et al., 2007). Despite these shortcomings, Mowrer's theory serves as an elegant exemplar of how etiology and maintenance can have different mechanisms of causality.

Emotional Processing Model

Foa and Kozak's (1986) EPM posits that anxiety and related disorders develop from escape and avoidance behavior under stimulus control of a latent fear network (Rauch et al., 2012; Cahill et al., 2009; Foa and Kozak, 1986). This fear network contains all information, responses, and meaning related to the feared stimuli, and it is differentiated from other information structures by its reactivity and negative valence. For example, someone with agoraphobia might have a fear network that codes crowded places, elevators, and bridges as triggers for avoidance and escape. The fear network is adaptive in that it helps protect the individual from encountering stressful stimuli, but the pervasive overgeneralization to nondangerous situations hampers vocational, social, and health needs.

EPM suggests that activation of the fear network via exposure and subsequent habituation are essential therapy components. Activation and habituation are thought to decrease sensitivity to feared stimuli and mediate cognitive restructuring (Foa and Kozak, 1986). Some EPM therapies use explicit cognitive procedures to alter the fear network once activated (Resick and Schnicke, 1992), while others assert that only exposure is necessary to foster cognitive change (Foa and Kozak, 1986). EPM differs from Mowrer's theory in that it does not require single trial learning to account for the development and expansion of a fear network. Furthermore, EPM posits a latent cognitive construct to account for stimulus generalization and operant relationships. As such, EPM is a blend of cognitive and behavioral explanatory features.

Overall, EPM has considerable explanatory value, but current research on active mechanisms has challenged EMP's emphasis on habituation. Specifically, research suggests that new learning (i.e., inhibitory learning) is more predictive of treatment outcome than habituation. In fact, mandating in-session habituation may hinder the amount of new learning that can take place during a given session (Craske et al., 2008). Other critics challenge EPM's narrow etiological view, suggesting that interpersonal factors, such as low self-efficacy, maintain distress (Mineka and Thomas, 1999). Despite these criticisms, the EPM remains the most researched and widely applied theoretical model to date.

Treatment Characteristics of Exposure Therapy

Exposure therapies have been historically associated with treatment of syndromes previously classified as anxiety disorders in the Diagnostic and Statistical Manual of Mental Disorders, 4th edition, text revision (DSM-IV-TR; American Psychiatric Association, APA, 2000). Specifically, post-traumatic stress disorder (PTSD), obsessive-compulsive disorder (OCD), social phobia, specific phobia, panic disorder, and agoraphobia have responded well to exposure treatment. The DSM-5 (5th edition, APA, 2013) has moved PTSD into the new 'trauma- and stressor-related disorders' category and OCD into the new 'obsessive-compulsive and related disorders' category. According to this new codification, exposure's purview extends beyond anxiety disorders, but the specific disorders have not changed.

Unique iterations of exposure exist for each disorder (Richard and Lauterbach, 2007), but certain general procedures typify all applications. Specifically, each manualized treatment consists of exposure, generalization, and response prevention efforts. This section describes these general characteristics.

Exposure

Exposure procedures can be divided into *in vivo* exposure, imaginal exposure, and *in virtuo* (i.e., virtual reality) exposure. *In vivo* exposure occurs when an individual physically confronts a feared stimulus in its natural environment or artificially in session (Lauterbach and Reiland, 2007). For example, an individual with a phobia of flying might sit in a taxied plane as a preliminary form of *in vivo* exposure. To fully activate the fear network, *in vivo* exposures should mimic (or actually be) the feared stimuli (Craske and Barlow, 2008).

Imaginal exposure refers to the detailed recollection of feared stimuli (Cahill et al., 2009). During imaginal exposure individuals narrate their trauma in the present tense, even when narrating a past event, to increase intensity and emersion (Zayfert and Becker, 2007). For example, a person recounting a physical assault during imaginal exposure would be encouraged to say "I can see someone moving from my periphery" versus "I saw someone move from my periphery." Treatments often blend both *in vivo* and imaginal methods, presenting the chosen stimuli for prolonged periods and in multiple settings (Zayfert and Becker, 2007).

Advancements in technology have created a third exposure class known as *in virtuo* exposure, which utilizes virtual reality (VR) simulations to foster a therapeutic anxiety reaction (Bouchard et al., 2007). *In virtuo* exposure typically consists of head-mounted displays with motion trackers (Bouchard et al., 2007). Merits of *in virtuo* exposure include increased treatment control (Goncalves et al., 2012), higher acceptability (Garcia-Palacios et al., 2001), and a higher range of plausible exposure scenarios than *in vivo* (Bouchard et al., 2007). For example, in the treatment of combat PTSD, *in virtuo* exposure offers the ability to generate an interactive environment that mimics the patient's deployment setting. Thus far, *in virtuo* exposure has been successfully incorporated into treatments for phobias (Bouchard et al., 2007), PTSD (Rothbaum et al., 2001), panic disorder with agoraphobia (Pérez-Ara et al., 2010), and OCD (Kim et al., 2009). In short, VR holds the potential of revolutionizing exposure therapy.

In-vivo, imaginal, and *in virtuo* exposures share a common implementation protocol. At the outset of therapy, exposure events are chosen collaboratively by the therapist and client, and usually placed in a rank ordered anxiety hierarchy (Cahill et al., 2009). Treatment consists of progressing through this list, beginning with the least distressing and working toward the most distressing. Movement through the hierarchy is traditionally dictated by habituation, as measured via subjective units of distress (Richard et al., 2007).

Generalization

Generalization, in the context of therapy, refers to meaningful declines in anxiety, fear, and avoidance that persist across settings and locations (Sturmey et al., 2007). In exposure therapy, generalization of new learning into real world settings is the ultimate goal of therapy (Craske et al., 2008), and fostering generalization takes careful planning (Sturmey et al., 2007). In terms of the EPM, generalization requires activating the fear network's underlying structure (Foa and Kozak, 1986). As such, exposures should incorporate the most evocative components of the encoded fear (both physical and personal response aspects; Lang, 1977). For example, imaginal exposure of a motor vehicle accident should include detailed descriptions of the vehicle and logistics of the crash, as well as the smells, sounds, and physiological responses associated with the fear network.

The theme of generalization was embedded throughout the previous section on exposure. For example, the exposure hierarchy is an explicit method of ensuring that powerful and diverse exposure experiences occur. Failure to access essential aspects of the fear network can severely limit generalization (Foa and Kozak, 1986). As a simple example, consider a patient with a phobia of dogs. Exposure to a Yorkshire Terrier may work well for eliminating the fear response to small dogs, but might presumably do little for large breeds. To facilitate generalization across all breeds (i.e., settings), exposure should include a variety of dogs, with special attention to the most feared type. While varying exposure is a crucial element, generalization is also affected by the presence of avoidance behaviors during session. The following section describes this issue.

Response Prevention

Response prevention refers to the cessation of the avoidance behavior that has been maintaining the fear response

(Abramowitz et al., 2013). Exposure should always be paired with some form of response prevention to maximize treatment engagement and to put avoidance behavior on extinction (Abramowitz et al., 2013). For example, doing an exposure treatment for a germ phobia would require contact with a feared 'contaminated' surface (e.g., a restroom door handle), as well as the prevention of immediate hand washing, which is a form of escape. Failure to prevent the escape response impedes exposure and treatment outcome (Abramowitz, 1996). Response prevention also pertains to the thoughts and internal processes that maintain avoidance behavior such as recitation of songs or prayers, or conjuring distracting mental images (Purdon, 2004). For example, an individual suffering from repugnant sexual obsessions may recite a spiritual text to 'purify' his mind. Continuing this response during imaginal exposure blocks new learning/habituation, defusing exposure's therapeutic power. Response prevention also entails the careful management of fast acting anxiolytic medications, like Alprazolam, that function as avoidance behaviors and hamper exposure (Deacon, 2007).

A topic related to response prevention is safety signals, which are stimuli that diminish the negative valence of feared stimuli (Lohr et al., 2007). Learning that takes place in the presence of safety signals does not generalize to settings where safety signals are not present (Craske et al., 2008). Craske and Barlow (2008) note that people, food, money (e.g., to make emergency calls), medications, and familiar landmarks are potential safety signals that, like avoidance responses, disengage the individual from exposure, thus blocking the therapeutic effects (Craske et al., 2008). As such, detection and elimination of safety signals is integral to successful exposure treatment.

Stress Inoculation Training: Exposure as Prevention

Exposure procedures have also been incorporated into a preventative treatment for high-risk populations. Stress inoculation training (SIT) is a multifaceted cognitive behavioral intervention for stress and was developed in the 1980s. SIT is similar to the aforementioned exposure therapies in that it uses imaginal and *in vivo* exposure to build coping skills and increase functionality. Yet, SIT operates under a unique conceptual model and typically employs exposure as a preventative, rather than curative, procedure. SIT has been applied to multiple nonclinical (Saunders et al., 1996) and select clinical populations (Cahill et al., 2009) with varying degrees of success. Overall, SIT has proven to be an efficacious therapy that has amassed a considerable amount of research (Meichenbaum, 1985; Meichenbaum and Deffenbacher, 1988; Spencer, 2010). The following sections will detail SIT's conceptual model, procedure, and applications.

Conceptual Model of SIT

SIT's underlying framework is a transactional model of stress (Meichenbaum, 1985). According to SIT, stress emerges from a dynamic interplay of the individual, the environment, and emotional/physiological reflexes (Folkman, 1984). Just as outside forces introduce potentially stressful material, individuals engender stress via their interpersonal relationships and coping style (Meichenbaum, 1985). For example, an individual who responds to high job-place stress by drinking heavily during the week experiences stress as a result of both external (work demands) and personal (excessive drinking) factors. SIT places special attention on the role of cognitions, a mediator of the environment-behavior stress transaction (Meichenbaum, 1985).

Cognitions, in the context of SIT, can be understood as an appraisal of one's environment and self-efficacy (Folkman, 1984). Cognitions do not introduce new material, rather they make sense of the environment in a manner that produces or inhibits stress. Hopko and Hughes (2005) conceptualize this in terms of the well-known behavioral three-term contingency (A-B-C), suggesting that cognitions (B) evoke stress reactions (C), rather than the events themselves (A). As an example, imagine a single mother who works by day, goes to night school, and comes home to care for her son. According to SIT, the extent to which this situation (A) is internalized as stressful (C) depends on how the mother appraises her situation (B) and her ability to cope. It is easily seen that holding the belief "This is impossible; I am a failure" results in more stress than "Although this is hard, I am building a stable life that I can be proud of."

Implementation of Intervention

SIT is focused on changing distorted cognitive structures using a variety of methods. These cognitive structures, also known as schema, refer to underlying beliefs, assumptions, and perspectives about the world and one's self. Similar to typical exposure therapies, SIT relies on imaginal and *in vivo* exposure to alter cognitions. SIT also places a strong emphasis on cognitive restructuring and personalized treatment formulation (Meichenbaum, 1985). SIT organizes these techniques into three treatment phases: conceptualization, skills acquisition/rehearsal, and application (Meichenbaum, 1977). The following paragraphs detail the three phases of SIT, with special attention to exposure components.

The 'conceptualization phase' involves assessment of the client's problems, symptoms, and expectations, training in relaxation skills, and education about the nature of stress and the utility of previously used coping strategies (Meichenbaum and Deffenbacher, 1988). Imagery-based recall of recent stressful events, an exposure technique, is often used to identify the patient's stress patterns. Behavioral assessments, similar to *in vivo* exposure, are also used to test the environment and demonstrate how the client's typical behavior generates stress (Meichenbaum, 1985). This procedure encourages a proactive approach toward resolving stressors, which essentially inoculates against stress (Meichenbaum, 1996). For example, learning how to talk about work performance with one's boss prior to a crisis should alleviate stress in the event of a crisis. The information gathered from these procedures is aggregated and synthesized into a personalized clinical 'reconceptualization' of the client's stress. This revamped model emphasizes the transactional nature of stress and coping, the patients' control, and pinpoints symptoms for intervention.

'Skills acquisition/rehearsal' is the second phase, and dismantling studies suggest it is indispensible (Horan et al., 1977; Vallis, 1984). This phase of SIT focuses on building the client's skill-set to facilitate effective, resilient coping. Client input guides this phase, and the SIT therapist manual notes that: "the training approach adopted by SIT is to offer the variety of coping techniques in a cafeteria style, whereby clients can experiment in determining what works best for them" (Meichenbaum, 1985: p. 54). The skills acquired during this phase can be separated into problem- and emotion-focused coping.

Problem-focused coping attempts to address stress by directly altering environmental variables. These types of coping strategies are only appropriate in situations where environmental control is actually possible. In contrast, emotion-focused coping directly addresses the subjective meaning of the situation, and is often used in situations where environmental variables are beyond control (e.g., chronic pain or grieving loss). It is often the case that situations warrant both problem- and emotion-focused coping methods. The most typical SIT skills include cognitive skills, relaxation training, and self-instructional training.

SIT's third and final phase is 'application.' This phase is concerned with the generalization of the skills acquired during the acquisition phase via incremental induction (or inoculation) of stressful stimuli (Meichenbaum and Deffenbacher, 1988). Consistent with the flexible approach of the prior phases, the application and follow-through phase can entail a wide range of procedures to meet the needs of the specific individual. Techniques include imagery rehearsal (past and future events), behavioral rehearsal, and *in vivo* practice. The central theme of this phase is that generalization should not be left to chance. During therapy, the individual should come into contact with a wide variety of stressors to ensure mastery and application of SIT skills in daily life. SIT frames success as continued application of skills, not an elimination of experiences that could produce stress (Meichenbaum, 1985).

Nonclinical and Clinical Applications of SIT

The bulk of SIT research focuses on nonclinical populations in high stress environments. SIT has been effective for coping with stress-inducing environments involving anger, medical illness, surgery, performance evaluation, and vocational stress such as combat, police work, and teaching (see Meichenbaum, 1985 for a comprehensive review). SIT's research footprint has shrunk over the past decade, but dissertations continue to explore SIT's effectiveness. Most recently, SIT has been successfully applied to single mothers attending college (Spencer, 2010), members of a residential training center (Mayer, 2008), and college students (Ayers, 2010). These studies consistently find that SIT is effective for managing stress in nonclinical populations.

SIT has also been applied to traumatized populations (Cahill et al., 2009). Kilpatrick et al. (1982) created a manualized version of SIT to treat rape survivors. SIT has produced results equivalent to prolonged exposure (PE) for the treatment of rape victims, but follow-up analysis favored PE (Foa et al., 1991, 1999). SIT's utility in treating other PTSD populations has been mixed. For example, SIT has demonstrated limited efficacy in treating combat PTSD populations (Hensel-Dittmann et al., 2011). Yet, it is difficult to assess the efficacy of SIT for PTSD, because SIT is typically a control condition. In such studies, researchers restrict the use of exposure to make a clear distinction between SIT and the primary exposure intervention of interest (Foa et al., 1999, 1991). SIT is an intentionally dynamic intervention and these studies adulterate SIT's theoretical/treatment model and their conclusions should be analyzed with caution.

Obstacles to Implementing Exposure Therapies

The title of Richard and Gloster's (2007) chapter adeptly identifies the obstacles hindering exposure therapy's utilization: *Exposure Therapy Has a Public Relations Problem: A Dearth of Litigation Amid a Wealth of Concern*. Indeed, exposure therapy is suffering from a pervasive skepticism that has severely hampered its adoption in the clinical arena. A study of primary care patients with PTSD found that only 16% reported receiving exposure therapy (Rodriguez et al., 2003). VA rates are similar with less than 20% of PTSD specialists reporting consistent use of exposure techniques (Rosen et al., 2004). This problem is not limited to PTSD. Freihei et al. (2004) site that only 38 and 31% of surveyed therapists reported frequently using exposure for OCD and social anxiety disorder, respectively; furthermore, 79% reported rarely or never using exposure for panic disorder.

The poor utilization of exposure therapies is curious when considering the potential cost savings and efficacy of the various treatments. For example, a microsimulation of the financial impact of successfully treating Operation Iraqi Freedom veterans with PTSD found a projected savings of 165 million dollars over the course of 2 years, if all treatment-seeking veterans received empirically based exposure therapies (Kilmer et al., 2011). The American Psychological Association, the International Society of Traumatic Stress Studies, and the Department of Defense have fully endorsed exposure therapies as 'gold standard' treatments. Yet, the empirical literature reveals skepticism going back to early applications of exposure for PTSD (Hembree and Cahill, 2007). Select PTSD researchers have suggested that high levels of exposure are aversive and potentially retraumatizing (Pitman et al., 1991, 1996; Tarrier et al., 1999), whereas others have challenged these conclusions on the grounds of limited empirical support, lack of systematic research, and inadequate controls (see Hembree and Cahill, 2007, for a thorough review).

Nonetheless, those critical of exposure have made a lasting impact on the field. Becker et al. (2004) found that 59% of nonspecialized therapists and 48% of trauma-specific therapists report believing exposure increases a patient's desire to drop out. Furthermore, 87% of the main sample and 55% of specialists expressed the belief that exposure likely increases arousal. Consequently, exposure has accrued an almost medieval reputation, which one author likened to 'A Clockwork Orange' style behavioral modification (Prochaska and Norcross, 1999). Exposure's press in the public sector has also suffered. Slater (2003) quotes one journalist as stating that exposure is "torture, plain, and simple" (as cited in Richard and Gloster, 2007).

One reservation regarding exposure treatments is that patients will discontinue treatment prematurely. Hembree et al. (2003) analyzed 25 randomized controlled trails and found that attrition in exposure treatment (20.5%) did not statistically differ from cognitive therapy (22.1%) or EMDR (18.9%). Similarly, a meta-analysis of psychotherapy for PTSD (Bradley et al., 2005) found no significant differences in completion rate across different treatment types, and Imel et al. (2013) found that higher exposure does not result in higher attrition.

Furthermore, multiple studies have found a weak connection between exposure therapy and symptom exacerbation. In the most thorough analysis to date, Foa et al. (2002) reported that a minority of participants exhibited symptom worsening during treatment, and that such worsening was associated with the initiation of imaginal exposure. Importantly, those who showed symptom worsening during treatment were not more likely to drop out of treatment and did not have different treatment outcomes. As such, the evidence suggests that exposure may cause symptom exacerbation in a small percentage of patients, but this does not influence drop out or treatment outcomes (Hembree and Cahill, 2007).

The resistance to utilizing exposure, despite ample evidence dispelling skepticism, indicates that a major dissemination problem exists. Becker et al. (2004) report that only 13% of their sample attested to formal imaginal exposure training for PTSD and 47% reported limited or no familiarity. As such, serious attention should be given to generating and evaluating methods of dissemination (Gunter and Whittal, 2010).

Future Directions

Exposure and Pharmacotherapy

Pharmacotherapies such as selective serotonin reuptake inhibitors have demonstrated short-term effects comparable to exposure therapies (Deacon, 2007). Yet, research does not support the assertion that combined treatment is typically better than exposure alone (cf Marks et al., 1993). Furthermore, the paucity of research on combined treatment for PTSD, social phobia, GAD, and specific phobias preclude any definitive conclusions (Deacon, 2007). Multiple explanations have been offered to account for the weak findings. As previously stated, medications, especially benzodiazepines, are thought to inhibit exposure and operate as a form of medical escape. Relatedly, limited research suggests that learning during exposure with medications may not generalize to medication free settings (Deacon, 2007).

Fortunately, a new vein of pharmacotherapy has demonstrated promising preliminary findings. D-cycloserine (DCS), a partial *N*-methyl-D-aspartate (NMDA) agonist, in combination with exposure has produced impressive results (Ressler et al., 2007). DCS has been found to enhance cognitive learning in spatial memory tasks by acting on NMDA receptors, which are thought to mediate extinction and associative fear conditioning (Ressler et al., 2007). Studies of PTSD (de Kleine et al., 2014), specific phobia (Norberg et al., 2008; Ressler et al., 2007), OCD (Kushner et al., 2007), and generalized social anxiety disorder (Smits et al., 2013) have produced convincing results, suggesting that learning potentiators can dramatically enhance the effects of exposure therapy. Furthermore, DCS has been shown to accelerate remission (Ressler et al., 2007), decrease dropout (Kushner et al., 2007), and work with pediatric populations (Mataix-Cols et al., 2014). More research is needed to fully investigate this promising application of pharmacology and exposure.

Conclusion

Exposure therapies comprise the most efficacious treatments for anxiety and related disorders, as well as nonclinical stress. Exposure is a versatile technique that has been repackaged into numerous behavioral and cognitive treatments. Current theory suggests that exposure's utility stems from its ability to activate and foster amelioration of the latent fear network, thought to mediate distress. Traditional approaches have used a habituation model to describe exposure's affect on the fear network, but current research suggests that new learning (e.g., inhibitory learning) accounts for the therapeutic benefits of exposure. Regardless, exposure procedures that incorporate multiple modes of exposure, prevent avoidance responses, and program for generalizability engender positive outcomes. Unfortunately, dissemination of exposure into the clinical realm has been challenging, due to fears that exposure harms or unduly distresses patients. More research, training, and empirical approaches to dissemination are needed to rectify exposure treatment's perception in the public and clinical spheres. Nonetheless, exposure's future is promising with new pharmaceutical applications and innovative protocols for complex populations on the horizon.

See also: Behavior Therapy: Background, Basic Principles, and Early History; Evidence-Based Psychotherapy; Manual Based Treatments; Mass Trauma: Psychopathological Effects across the Life Span; Narrative Exposure Therapy; Obsessive-Compulsive Disorder across the Life Span; Phobias Across the Lifespan; Psychological Treatments, Empirically Supported; Virtual Reality in Psychotherapy.

Bibliography

Abramowitz, J.S., 1996. Variants of exposure and response prevention in the treatment of obsessive-compulsive disorder: a meta-analysis. Behavior Therapy 27 (4), 583–600.

Abramowitz, J.S., Fabricant, L.E., Jacoby, R.J., 2013. Obsessive-compulsive disorder. In: Craighead, W.E., Miklowitz, D.J., Craighead, L.W. (Eds.), Psychopathology: History, Diagnosis, and Empirical Foundations, second ed. John Wiley & Sons, Inc, Hoboken, NJ, pp. 80–107.

American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders, fourth ed., text rev. American Psychiatric Association, Washington, DC.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Association, Washington, DC.

Ayers, S.C., 2010. Achievement Motivation and Stress Inoculation Training: Coping through College (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (3403999).

Beck, A., 1976. Cognitive Therapy and the Emotional Disorders. International Universities Press, Oxford, England.

Becker, C.B., Zayfert, C., Anderson, E.R., 2004. A survey of psychologists' attitudes towards and utilization of exposure therapy for PTSD. Behaviour Research and Therapy 42 (3), 277–292.

Bouchard, S., Cote, S., Richard, D.C., 2007. Virtual reality applications for exposure. In: Richard, D.C., Lauterbach, D.L. (Eds.), Handbook of Exposure Therapies. Elsevier Inc., San Diego, CA, pp. 347–388.

Bradley, R., Greene, J., Russ, E., Dutra, L., Westen, D., 2005. A multidimensional meta-analysis of psychotherapy for PTSD. The American Journal of Psychiatry 162 (2), 214–227.

Cahill, S.P., Rothbaum, B.O., Resick, P.A., Follette, V.M., 2009. Cognitive-behavioral therapy for adults. In: Foa, E.B., Keane, T.M., Friedman, M.J., Cohen, J.A. (Eds.), Effective Treatments for PTSD: Practice Guidelines from the International Society for Traumatic Stress Studies, second ed. Guilford Press, New York, pp. 139–222.

Craske, M.G., Barlow, D.H., 2008. Panic disorder and agoraphobia. In: Barlow, D.H. (Ed.), Clinical Handbook of Psychological Disorders: A Step-by-step Treatment Manual, fourth ed. The Guilford Press, New York, pp. 1–64.

Craske, M.G., Kircanski, K., Zelikowsky, M., Mystkowski, J., Chowdhury, N., Baker, A., 2008. Optimizing inhibitory learning during exposure therapy. Behaviour Research and Therapy 46, 5–27.

Deacon, B., 2007. The effect of pharmacotherapy on the effectiveness of exposure therapy. In: Richard, D.C., Lauterbach, D.L. (Eds.), Handbook of Exposure Therapies. Elsevier Inc., San Diego, CA, pp. 311–333.

Foa, E.B., Dancu, C.V., Hembree, E.A., Jaycox, L.H., Meadows, E.A., Street, G.P., 1999. The efficacy of exposure therapy, stress inoculation training and their combination in ameliorating PTSD for female victims of assault. Journal of Consulting and Clinical Psychology 67, 194–200.

Foa, E.B., Kozak, M.J., 1986. Emotional processing of fear: exposure to corrective information. Psychological Bulletin 99, 20–35.

Foa, E.B., Rothbaum, B.O., Riggs, D., Murdock, T., 1991. Treatment of posttraumatic stress disorder in rape victims: a comparison between cognitive-behavioral procedures and counseling. Journal of Consulting and Clinical Psychology 59, 715–723.

Foa, E.B., Zoellner, L.A., Feeny, N.C., Hembree, E.A., Alvarez-Conrad, J., 2002. Does imaginal exposure exacerbate PTSD symptoms? Journal of Consulting and Clinical Psychology 68, 594–602.

Folkman, S., 1984. Personal control and stress and coping processes: a theoretical analysis. Journal of Personality and Social Psychology 46 (4), 839–852.

Forsyth, J.P., Barrios, V., Acheson, D.T., 2007. Exposure therapy and cognitive interventions for the anxiety disorders: overview and newer third-generation perspectives. In: Richard, D.C., Lauterbach, D.L. (Eds.), Handbook of Exposure Therapies. Elsevier Inc., San Diego, CA, pp. 61–108.

Freihei, S.R., Vye, C., Swan, R., Cady, M., 2004. Cognitive-behavioral therapy for anxiety: Is dissemination working? The Behavior Therapist 27, 25–32.

Garcia-Palacios, A., Hoffman, H.G., Kwong See, S., Tsai, A., Botella, C., 2001. Redefining therapeutic success of virtual reality exposure therapy. Cyberpsychology and Behavior 4 (3), 341–348.

Goncalves, R., Pedrozo, A.L., Coutinho, E.S.F., Figueira, I., Ventura, P., 2012. Efficacy of virtual reality exposure therapy in the treatment of PTSD: a systematic review, 7 (12), 1–7. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0048469.

Gunter, R.W., Whittal, M.L., 2010. Dissemination of cognitive-behavioral treatments for anxiety disorders: overcoming barriers and improving patient access. Clinical Psychology Review 30 (2), 194–202.

Hembree, E.A., Cahill, S.P., 2007. Obstacles to successful implementation of exposure therapy. In: Richard, D.C., Lauterbach, D.L. (Eds.), Handbook of Exposure Therapies. Elsevier Inc, San Diego, CA, pp. 389–408.

Hembree, E.A., Foa, E.B., Dorfan, N.M., Street, G.P., Kowalski, J., Tu, X., 2003. Do patients drop out prematurely from exposure therapy for PTSD? Journal of Traumatic Stress 16 (6), 555–562.

Hensel-Dittmann, D., Schauer, M., Ruf, M., Catani, C., Odenwald, M., Elbert, T., Neuner, F., 2011. Treatment of traumatized victims of war and torture: a randomized controlled comparison of narrative exposure therapy and stress inoculation training. Journal of Psychotherapy and Psychosomatics 80, 345–352.

Hopko, D.R., Hughes, F., 2005. Stress inoculation training. In: Hersen, M., Rosqvist, J., Gross, A.M., Drabman, R.S., Sugai, G., Horner, R. (Eds.), Encyclopedia of Behavior Modification and Cognitive Behavioral Therapy, vol. 1, pp. 570–574. Retrieved from: http://knowledge.sagepub.com.ezproxy.emich.edu/view/cbt/n154.xml.

Horan, J.J., Hackett, G., Buchanan, J.D., Stone, C.I., Demchik-Stone, D., 1977. Coping with pain: a component analysis of stress inoculation. Cognitive Therapy and Research 1, 211–222.

Imel, Z.E., Laska, K., Jakupcak, M., Simpson, T.L., 2013. Meta-analysis of dropout in treatments for posttraumatic stress disorder. Journal of Consulting and Clinical Psychology 81 (3), 393–404.

Kilmer, B., Eibner, C., Ringel, J.S., Pacula, R.C., 2011. Invisible wounds, visible savings? Using microsimulation to estimate the costs and savings associated with providing evidence-based treatment for PTSD and depression to veterans of operation enduring freedom and operation Iraqi freedom. Psychological Trauma: Theory, Research, Practice, and Policy 3, 201–211.

Kilpatrick, D.G., Veronen, L.J., Resick, P.A., 1982. Psychological sequelae to rape: assessment and treatment strategies. In: Doleys, D.M., Meredity, R.I., Ciminero, A.R. (Eds.), Behavioral Medicine: Assessment and Treatment Strategies. Plenum, New York, pp. 214–231.

Kim, K., Kim, C., Kim, S., Roh, D., Kim, S.I., 2009. Virtual reality for obsessive-compulsive disorder: past and the future. Psychiatry Investigations 6, 115–121.

de Kleine, R.A., Hendriks, G.J., Smits, J.A., Broekman, T.G., van Minnen, A., 2014. Prescriptive variables for d-cycloserine augmentation of exposure therapy for posttraumatic stress disorder. Journal of Psychiatric Research 48 (1), 40–46.

Kushner, M.G., Kim, S.W., Donahue, C., Thuras, P., Adson, D., Kotlyar, M., Foa, E.B., 2007. D-cycloserine augmented exposure therapy for obsessive-compulsive disorder. Biological Psychiatry 15 (62), 835–838.

Lang, P.J., 1977. Imagery in therapy: an information processing analysis. Behavior Therapy 8, 862–886.

Lauterbach, D., Reiland, S., 2007. Exposure therapy and post-traumatic stress disorder. In: Richard, D.C., Lauterbach, D.L. (Eds.), Handbook of Exposure Therapies. Elsevier Inc, San Diego, CA, pp. 153–168.

Lohr, J.M., Olatunji, B.O., Sawchuk, C.N., 2007. A functional analysis of danger and safety signals in anxiety disorders. Clinical Psychology Review 27 (1), 114–126.

Mataix-Cols, D., Turner, C., Monzani, B., Isomura, K., Murphy, C., Krebs, G., Heyman, I., 2014. Cognitive-behavioural therapy with post-session D-cycloserine augmentation for pediatric obsessive-compulsive disorder: pilot randomised controlled trial. British Journal of Psychiatry 204 (1), 77–78.

Marks, I.M., Swinson, R.P., Başoğlu, M., Kuch, K., Noshirani, H., O'Sullivan, G., Wickwire, K., 1993. Alprazolam and exposure alone and combined in panic disorder with agoraphobia: a controlled study in London and Toronto. British Journal of Psychiatry 162, 776–787.

Mayer, J.S., 2008. The Development, Implementation & Evaluation of a Stress Inoculation Training Program at a Residential Training Center (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (3370300).

Meichenbaum, D., 1977. Cognitive Behavioral Modification: An Integrative Approach. Plenum Press, New York.

Meichenbaum, D., 1985. Stress Inoculation Training. Pergamon Press, New York.

Meichenbaum, D., Deffenbacher, J.L., 1988. Stress inoculation training. The Counseling Psychologist 16 (1), 69–90.

Meichenbaum, D., 1996. Stress inoculation training for coping with stressors. The Clinical Psychologist 49, 4–7.

Mineka, S., Thomas, C., 1999. Mechanisms of change in exposure therapy for anxiety disorders. In: Dalgleish, T., Power, M.J. (Eds.), Handbook of Cognition and Emotion. John Wiley & Sons Ltd., New York, pp. 747–764.

Mowrer, O.H., 1960. Learning Theory and Behavior. John Wiley & Sons Inc., Hoboken, NJ.

Norberg, M.M., Krystal, J.H., Tolin, D.F., 2008. A meta-analysis of D-cycloserine and the facilitation of fear extinction and exposure therapy. Biological Psychiatry 63 (12), 1118–1126.

Pérez-Ara, M.A., Quero, S., Botella, C., Baños, R., Andreu-Mateu, S., García-Palacios, A., Bretón-López, J., 2010. Virtual reality interoceptive exposure for the treatment of panic disorder and agoraphobia. Annual Review of Cybertherapy and Telemedicine 154, 77–81.

Pitman, R.K., Altman, B., Greenwald, E., Longpre, R.E., Macklin, M.L., Poire, R.E., Steketee, G.S., 1991. Psychiatric complications during flooding therapy for posttraumatic stress disorder. The Journal of Clinical Psychiatry 52 (1), 17–20.

Pitman, R.K., Orr, S.P., Altman, B., Longpre, R.E., Poire, R.E., Macklin, M.L., 1996. Emotional processing during eye movement desensitization and reprocessing therapy of Vietnam veterans with chronic posttraumatic stress disorder. Comprehensive Psychiatry 37 (6), 419–429.

Prochaska, J.O., Norcross, J.C., 1999. Systems of Psychotherapy: A Transtheoretical Analysis, fourth ed. Books/Cole Publishing, Pacific Grove.

Purdon, C., 2004. Cognitive-behavioral treatment of repugnant obsessions. Journal of Clinical Psychology 60, 1169–1180.

Rauch, S.A.M., Eftekhari, A., Ruzek, J.I., 2012. Review of exposure therapy: a gold standard for PTSD treatment. Journal of Rehabilitation Research and Development 49 (5), 679–687.

Resick, P.A., Schnicke, M.K., 1992. Cognitive processing therapy for sexual assault victims. Journal of Consulting and Clinical Psychology 60 (5), 748–756.

Ressler, K.J., Davis, M., Rothbaum, B.O., 2007. Pharmacological enhancement of learning in exposure therapy. In: Richard, D.C., Lauterbach, D.L. (Eds.), Handbook of Exposure Therapies. Elsevier Inc., San Diego, CA, pp. 335–346.

Richard, D.C., Lauterbach, D., 2007. Handbook of Exposure Therapies. Elsevier Inc., San Diego, CA.

Richard, D.C., Gloster, A.T., 2007. Exposure therapy has a public relations problem: a dearth of litigation amid a wealth of concern. In: Richard, D.C., Lauterbach, D.L. (Eds.), Handbook of Exposure Therapies. Elsevier Inc., San Diego, CA, pp. 409–426.

Richard, D.C., Lauterbach, D., Gloster, A.T., 2007. Description, mechanisms of action, and assessment. In: Richard, D.C., Lauterbach, D.L. (Eds.), Handbook of Exposure Therapies. Elsevier Inc., San Diego, CA, pp. 1–28.

Rodriguez, B.F., Weisberg, R.B., Pagano, M.E., Machan, J.T., Culpepper, L., Keller, M.B., 2003. Mental health treatment received by primary care patients with posttraumatic stress disorder. Journal of Clinical Psychiatry 64 (10), 1230–1236.

Rosen, C.S., Chow, H.C., Finney, J.F., Greenbaum, M.A., Moos, R.H., Sheikh, J.I., Yesavage, J.A., 2004. VA practice patterns and practice guidelines for treating posttraumatic stress disorder. Journal of Traumatic Stress 17 (3), 213–222.

Rothbaum, B.A., Hodges, L.F., Ready, D., Graap, K., Alarcon, R., 2001. Virtual reality exposure therapy for Vietnam veterans with posttraumatic stress disorder. Journal of Clinical Psychiatry 62, 617–622.

Saunders, T., Driskell, J.E., Johnston, J.H., Salas, E., 1996. The effect of stress inoculation training on anxiety and performance. Journal of Occupational Health Psychology 1 (2), 170–186.

Smits, J.A.J., Rosenfield, D., Otto, M.W., Marques, L., Davis, M.L., Meuret, A.E., Hofmann, S.G., 2013. D-cycloserine enhancement of exposure therapy for social anxiety disorder depends on the success of exposure sessions. Journal of Psychiatric Research 47 (10), 1455–1461.

Spencer, M., 2010. Stress Inoculation Training for Single Mothers in College (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (3424833).

Sturmey, P., Ward-Horner, J., Marroquin, M., Doran, E., 2007. Structural and functional approaches to psychopathology and case formulation. In: Sturmey, P. (Ed.), Functional Analysis in Clinical Treatment. Academic Press, Boston, MA, pp. 1–21.

Tarrier, N., Pilgrim, H., Sommerfield, C., Faragher, B., Reynolds, M., Graham, E., Barrowclough, C., 1999. A randomized trial of cognitive therapy and imaginal exposure in the treatment of chronic posttraumatic stress disorder. Journal of Consulting and Clinical Psychology 67 (1), 13–18.

United States Department of Veterans Affairs, 2010. Management of Post-Traumatic Stress: VA/DoD Clinical Practice Guideline. Retrieved from: www.healthquality.va.gov/ptsd/CPGSummaryFINALMgmtofPTSDfinal021413.pdf.

Vallis, T.M., 1984. A complete component analysis of stress inoculation for pain tolerance. Cognitive Therapy and Research 8, 313–329.

Watson, J.B., Rayner, R., 1920. Conditioned emotional reactions. Journal of Experimental Psychology 3 (1), 1.

Zayfert, C., Becker, C.B., 2007. Cognitive-behavioral Therapy for PTSD: A Case Formulation Approach. The Guilford Press, New York.

Extended Schools

Alan Dyson, University of Manchester, Manchester, England

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Schools in many countries are developing extended roles. These take them beyond their core academic concerns and involve them more widely in the lives of their students and in families and communities. There are many approaches to developing these extended roles, but they can be seen as lying on a continuum from a focus on remedying deficits to a focus on developing assets. The evidence for the effectiveness of these approaches has limitations but is nonetheless promising. Ultimately, the development of extended roles raises fundamental questions about the role of schools and the purposes of education.

The core task of schools everywhere is to develop the cognitive capacities of their students and, particularly, to enable them to make progress through some more-or-less well-defined academic curriculum. However, in many places the role of schools is being extended beyond this core task. Schools are engaging with other, nonacademic aspects of their students' lives – with their social development, for instance, or their health and well-being. In order to do this, they are developing close partnerships with other, noneducational services, such as social care, youth work, or health. Often, they seek to work more closely with families, involving them in their children's education, helping them develop more effective parenting skills, and perhaps finding ways to support them through difficulties. They may also see themselves as having a role in the wider communities they serve, offering adult education provision, opening their facilities to community use, or contributing to community development and area regeneration.

In some cases, developments of this kind originate in the values and enthusiasms of particular school principals and their staff. In other cases, there are more formal local or national initiatives which support and encourage schools to move in this direction. In the USA, for instance, there have been a series of initiatives to develop what are variously known as 'full service' or 'community' schools (see, for instance, Dryfoos, 1994, 2005; Blank et al., 2003). In the Netherlands and Belgium, there are 'brede scholen' (Joos et al., 2006; Bakker, 2010). In Portugal, groups of schools have been encouraged to develop extended services and activities in the 'TEIP' initiative (Abrantes et al., 2013), which itself borrows from the Zones d'Éducation Prioritaires in France (Bénabou et al., 2005; Rochex, 2012). In Australia, schools have been supported in establishing links with community organizations (Gregoric, 2013). In England, meanwhile, there are 'extended' schools (Cummings et al., 2011b), and in Scotland 'new' or 'integrated' community schools (The Scottish Office, 1998).

Many of these initiatives are located in affluent countries. However, there is also a somewhat different tradition of community schooling in Africa, often supported by international organizations (Glassman et al., 2007; Miller-Grandvaux and Yoder, 2002), while some of the countries of central and eastern Europe have also been supported in developing community schools (International project 'International Standards for Community Schools: Strengthening the National Capacities,' no date). Moreover, although many of these initiatives are relatively recent, some places have a long history of schools with extended roles. In the USA that history can be traced back at least as far as the work of Jane Addams and John Dewey in the late nineteenth and early twentieth centuries (Benson et al., 2009), while in England the first recognizable community school initiative took the form of the Village Colleges developed in Cambridgeshire in the 1920s (Morris, 1925).

While this long history and extensive geographical reach makes it tempting to think in terms of a coherent movement, this would be misleading. Types of schools which at first glance appear to be similar, and which may even be labeled in the same way, can in fact be organized differently, can develop different practices, and, above all, can be based on very different rationales. The initiatives listed above, for instance (and there are many more), are concerned variously with making good shortfalls in state educational provision, involving communities in democratic processes, tackling the personal and social problems experienced by disadvantaged children in affluent countries, or ensuring the viability of rural communities. Some of them place an emphasis on community ownership of the school, others on colocating noneducational services on the school site, and others again on developing the school as a community resource. Moreover, even within the same initiative and the same broad purposes, individual schools are likely to develop distinctive characteristics as they respond to their local contexts (Cummings et al., 2011b).

This variability poses significant challenges in terms of building up a shared knowledge base on schools working with an extended role. If we cannot say what such schools are and how they operate – indeed, if there is no internationally agreed designation of such schools – it is difficult for one initiative to learn from another, and much of what we think we know may turn out to be illusory. However, while we may not be able to articulate a clear definition of what such a school 'really is,' we can at least map out some of the broad distinctions between different ways of thinking about how schools can extend their roles.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92130-5

Competing Rationales

An important set of distinctions is around the underpinning rationales for schools going beyond their core academic business. Although these rationales are extremely varied, it is useful to think in terms of a broad difference between deficit- and asset-based approaches (Foot and Hopkins, 2010). The former focus on the role schools can play in tackling the problems experienced by disadvantaged children, families, and communities. The latter focus on building upon the assets already available in the areas they serve so that those areas become better places to live and the people within them are able to live better lives.

The rationale for deficit-based approaches was articulated most clearly by Joy Dryfoos, one of the pioneers of 'full service' schools in the USA. In disadvantaged areas, she argued:

> … schools are failing because they cannot meet the complex needs of today's students … The cumulative effects of poverty have created social environments that challenge educators, community leaders, and practitioners of health, mental health, and social services to invent new kinds of institutional responses.
>
> Dryfoos, 1994, p. xvii

In such circumstances, a focus on the academic aspect of education alone is futile, since the other needs of children will undermine whatever the school seeks to achieve in this respect. Likewise, in the face of the complexity of these needs, there is little point in tackling problems in isolation from each other. Some form of coordinated approach is needed, which in turn implies that traditional models of service delivery through separate agencies are likely to be ineffective. There is therefore a role for the school in developing new models by extending the services it can offer from its own resources and by developing strong partnerships with other agencies. As the Coalition for Community Schools in the USA suggests:

> [These] schools recognize that many factors influence the education of our children and that various conditions need to be in place to produce positive results. This is why they work with partners to mobilize the assets of the school and the entire community to improve educational, health, social, family, economic, and related results.
>
> Shah et al., 2009, p. 9

It is probably true to say that some version of the deficit-based approach underpins the majority of attempts to extend the role of schools. This is particularly the case in affluent countries where education systems are struggling to tackle inequalities in outcomes between children experiencing disadvantage and their peers (OECD, 2012). However, such a rationale does not explain the emergence of community-managed schools in economically poorer countries, nor the focus on democratization through community involvement in countries undergoing political transformation. Such initiatives are more about enhancing the assets available to communities than providing services to tackle the problems besetting children. Moreover, even in affluent and politically stable countries asset-based approaches are occasionally evident, particularly where the pressures from relative disadvantage are not the focus of schools' work. The SchoolPLUS initiative in Saskatchewan, Canada, for instance, is based on a set of liberal values and is about enhancing quality of life and community relations (M. Tymchak (chair) Task Force and Public Dialogue on the Role of the School, 2001). As Regnier argues:

> The creation of new family, community, school, and agency relationships within the SchoolPLUS model can develop forms of working relationships with rootedness that sustain the belonging needed to explore possibilities and test limits in supportive, nourishing, and forgiving contexts that advance the best developmental interests of children and youth.
>
> Regnier, 2012, p. 76

The emphasis here is on nurturing environments for children and, more generally, on enabling citizens to live a good life rather than on tackling the social problems produced by inequality.

There is, of course, no reason why schools should not combine these two rationales in various ways. Harkavy et al. propose a definition of 'community schools' which sees them both as tackling the presenting problems of disadvantaged areas and energizing the democratic processes in such areas:

> 'Community schools' bring together multiple organizations and their resources not only to serve and educate young people but also to democratically engage all members of the community in which the school is located. Essentially, this idea extends and updates John Dewey's theory that the neighborhood school can and should function as the core neighborhood institution – one that provides comprehensive services and galvanizes community institutions and organizations to help solve the myriad problems individuals and communities confront in a rapidly changing world.
>
> Harkavy et al., 2013, p. 526

While such combined rationales are clearly possible at the level of rhetoric, whether such a combination undermines one or both of the rationales in practice is an open question.

Scope

A second dimension along which efforts to extend the role of schools can be differentiated is in terms of their scope, and specifically, the extent to which they focus on the perceived needs of their students or work also with family and community issues. Some are interested primarily in improving outcomes for school students. There are, for instance, many initiatives internationally which, like the 'all-day schools' in Germany (Stecher and Schüpbach, 2013), are concerned primarily with offering out-of-school-hours arts, sports, and leisure activities, or even more narrowly with offering additional opportunities to study. Other initiatives, such as City Connects, originating in Boston, USA (City Connects, 2011) add to this basic package a range of health, social, and educational services, although the focus remains on helping students achieve more educationally. Others again include

work with families, with adult learners, or with whole communities. Hence, the Harlem Children's Zone in the USA (The Harlem Children's Zone, 2009) and its variants elsewhere (Dyson et al., 2012) offer comprehensive packages of support to children, their families, and communities. The core aim may remain that of enabling school students to achieve, but at some point this becomes inseparable from enhancing their development more generally, supporting families, enriching the lives of local residents, and promoting community development and regeneration.

As the scope of initiatives expands, so typically does the range of partners involved. While schools may be able to manage some limited additions to their core provision themselves, more ambitious initiatives involve partnerships with other public, private, and voluntary agencies. These may begin with agencies providing services directly to children, such as psychologists, mentors, youth workers, and health workers. However, they expand to include family workers, community workers, and those responsible for the resources and well-being of local areas, such as housing providers, police, employers, and economic development agencies. As this happens, the individual school may well cease to be central to initiatives and becomes one among many partners. Very probably, area governance structures emerge to manage an area-level approach.

We can see this kind of expansion clearly in developments which took place in England during the 2000s. A series of initiatives based on individual 'extended schools' gave way to 'extended services' in which schools in an area operated as a cluster to offer a wider range of services (DfES, 2005). Typically, these clusters formed part of an extensive network of child and family services at area level managed by local authorities which, in some cases, were themselves linked to area regeneration strategies (Cummings et al., 2011a). All of these services were expected to pursue a common set of outcomes for children which went well beyond educational achievement and addressed health, safety, economic well-being, and citizenship (DfES, 2003).

It is possible to see the movement along this dimension as an inevitable progression through successive generations of initiatives as policy makers and practitioners learn more about the possibilities of schools' extended roles (Lawson, 2013). It is certainly the case that, as the distal factors which promote or inhibit educational outcomes become clearer, initiatives are likely to address issues ever-more distant from the school's core academic business. However, the variety of rationales and assumptions underpinning initiatives means that it is sometimes necessary to see different ones as pursuing different aims, rather than necessarily as learning how to pursue a common aim more effectively.

Evidence for Outcomes

Attempts to extend the role of schools involve considerable effort and expenditure of resource to enhance and reorganize services. Estimates of the cost of full service schools in England suggested that schools might be using resources up to the value of £2 m ($3.25 m) to support their work (Cummings et al., 2007, p. 134 ff.). It is therefore important to know what impacts these initiatives have on outcomes for children – and, where they are involved – for families, and communities.

Unfortunately, robust evidence on outcomes is difficult to come by. In part, this is because of the inherent difficulty of evaluating complex initiatives with multiple aims, implemented in the open and unstable environments of schools and areas. There are particular issues where schools are pursuing broadly asset-based approaches and are thus aiming at outcomes where few ready measures exist. Not surprisingly, therefore, most evaluations focus on the outcomes achieved by initiatives focusing on overcoming the effects of disadvantage. Even where this is the case, however, evaluation demands sophisticated and complex designs, and needs to track initiatives over a sufficient length of time for outcomes to become apparent (Dyson and Todd, 2010). Funders have often been reluctant to invest in such costly ventures and evaluators have tended to content themselves with snapshots of those outcomes – usually in relation to children's attainments – that are closest to hand (Henderson and Mapp, 2002; Wilkin et al., 2003).

Despite these limitations, there is something of a consensus that the outcomes of schools with extended roles are, at the very least, promising. Reviewing a range of 'community school' initiatives, Blank et al. (2003) conclude that they can produce positive outcomes in four areas:

> *Student learning*: Community school students show significant and widely evident gains in academic achievement and in essential areas of nonacademic development.
> *Family engagement*: Families of community school students show increased stability, communication with teachers, and school involvement. Parents demonstrate a greater sense of responsibility for their children's learning success.
> *School effectiveness*: Community schools enjoy stronger parent–teacher relationships, increased teacher satisfaction, a more positive school environment, and greater community support.
> *Community vitality*: Community schools promote better use of school buildings, and their neighborhoods enjoy increased security, heightened community pride, and better rapport among students and residents (Blank et al., 2003, pp. 1, 2, emphases in original).

However, this optimistic assessment needs a little unpicking. There is reasonably robust evidence that the 'components' of such extended approaches – such as family support, out-of-hours activities, and social and health interventions with children – can have positive impacts on specific outcomes for children, families, and communities (Dyson and Kerr, 2013; Heers et al., 2011). There is also evidence that school improvement efforts as part of an overall change package which includes the development of an extended role can have major impacts on school performance, and therefore on the educational attainments of children within schools (Dobbie and Fryer, 2011; Whitehurst and Croft, 2010). However, it is less clear that there is a multiplier effect from combining school improvement reforms, services, and activities in a holistic 'approach.' It is difficult to find convincing evidence that there is a community school or extended school 'effect' over and above the effects of the individual services and activities on particular outcomes.

Set against this is the caveat that interventions have tended to look for these multiplier effects in relation to educational achievement and other closely related educational outcomes rather than across the broad domains across which schools may be working. Certainly, there is evidence that interventions in one domain (such as health) can have impacts on outcomes in another (such as educational attainment) (Dyson and Kerr, 2013). Moreover, it seems likely that different aspects of the work of schools interact with each other so that, for instance, efforts to improve teaching and learning need work with families and communities in order to be maximally effective (Bryk et al., 2010). Looking for additive impacts on outcomes from each additional activity undertaken by a school may therefore be less illuminating than trying to understand how extended activities create the conditions under which actions more directly related to measurable outcomes can work. Without better evaluations it is impossible to be certain on these matters. In the meantime, however, it is reasonable to conclude that schools developing an extended role are likely to have positive impacts across a range of domains, but that they need to adopt a thoughtful approach in deciding how, precisely, that role will be configured, and need to monitor carefully the outcomes that result.

Two Wider Issues

Given the diversity of initiatives, it is entirely possible for the development of extended roles in schools to have rather limited aims and limited impacts on the nature of schooling. The addition of some out-of-hours tuition or the services of a health worker to a school is unlikely to trouble other aspects of the schools' work or to have significant impacts on the communities served by the school. However, more ambitious initiatives clearly have much wider implications. There are two in particular which deserve attention.

The first is the implication of schools' extended roles for the nature of child, family, and community services. Traditionally, such services have tended to operate separately from one another, coming together only for specific purposes (Brown and White, 2006). As schools take on extended roles, however, they increase the opportunities for interagency collaboration and effectively blur the boundaries between services. Ultimately this may lead to the kind of holistic approaches, suggested by the Harlem Children's Zone and the Every Child Matters agenda in England, that are predicated on collaborative action, shared outcomes, and area-level coordination. In turn, this has implications for whether schools are seen as purely educational institutions or as multidimensional organizations engaged in many aspects of the lives of children, families, and communities. Ultimately this raises questions about the nature and purpose of education (Fielding and Moss, 2011; Moss and Haydon, 2008).

The second is the implication for the relationship between schools and the children, families, and communities they serve. In their educational role, schools enjoy considerable power in, for instance, requiring attendance, imposing sanctions on behavior of which they disapprove, and determining what shall be taught. However, as they extend their roles, the balance inevitably shifts away from compulsion toward the offering of opportunities and services. If, moreover, schools adopt an asset-based approach, their dominant position becomes increasingly untenable and they need to develop 'relational power' (Warren et al., 2009) in which decision-making is shared collaboratively between professionals and those they seek to serve. Not only, therefore, does the extended role of schools give rise to questions about the nature and purpose of education, it also raises questions about the appropriate relationship between educational institutions and those on whose behalf they are supposed to work.

Some Conclusions

The history of schools' efforts to develop extended roles is a long but somewhat patchy one. Different initiatives in different places have sought to develop different versions of an extended role, based on different aims and assumptions. Not surprisingly, the evidence base is less robust than might be hoped, and there is a clear need for more sophisticated evaluations. Nonetheless, such evidence as exists is promising and there are good reasons for persevering with efforts of this kind. However, in order to avoid simply repeating the past, these efforts have to be undertaken in full knowledge of the multiple directions in which extended roles can develop and with an awareness of the more fundamental questions that are raised as schools change the way they work. At a time when many countries are troubled by their standing in international assessments of educational performance, and by the low attainments of their most disadvantaged students, it is perhaps inevitable that many initiatives will be narrowly conceived as remedial interventions. While this is a limited view of what extending the role of schools can achieve, it is at least a start. The challenge for policy makers and practitioners is to move beyond this to a more open exploration of what schools might be for.

See also: Equity and Education; Family and Schooling; School (Alternative Models): Ideas and Institutions; Social Inequality and Schooling; Socioeconomic Status and Academic Achievement.

Bibliography

Abrantes, P., Roldão, C., Amaral, P., Mauritti, R., 2013. Born to fail? Some lessons from a national programme to improve education in poor districts. International Studies in Sociology of Education 23, 17–38.

Bakker, P.P., 2010. Community Schools in the Netherlands. Jeugd Instituut, Utrecht, Netherlands.

Bénabou, R., Kramarz, F., Prost, C., 2005. The French Zones D'Education Prioritaire: Much Ado about Nothing?. CEPR Discussion Paper No. 5085, Centre for Economic Policy Research.

Benson, L., Harkavy, I., Johanek, M., Puckett, J., Summer 2009. The enduring appeal of community schools. American Educator 22–47.

Blank, M., Melaville, A., Shah, B., 2003. Making the Difference: Research and Practice in Community Schools. Coalition for Community Schools, Institute for Educational Leadership, Washington, DC.

Brown, K., White, K., 2006. Exploring the Evidence Base for Integrated Children's Services. Edinburgh.

Bryk, A.S., Sebring, P.B., Allensworth, E., Luppescu, S., Easton, J.Q., 2010. Organizing Schools for Improvement: Lessons from Chicago. University of Chicago Press, Chicago.

City Connects, 2011. City Connects: The Lasting Impact of Optimized Student Support. Boston College, Chestnut Hill, Massachusetts, Boston.

Cummings, C., Dyson, A., Muijs, D., Papps, I., Pearson, D., Raffo, C., Tiplady, L., Todd, L., Crowther, D., 2007. Evaluation of the Full Service Extended Schools Initiative: Final Report. Research report RR852. DfES, London.
Cummings, C., Dyson, A., Jones, L., Laing, K., Todd, L., 2011a. Extended Services Evaluation: The Role of Local Authorities: Thematic Review. DfE, London.
Cummings, C., Dyson, A., Todd, L., 2011b. Beyond the School Gates: Can Full Service and Extended Schools Overcome Disadvantage? Routledge, London.
DfES, 2003. Every Child Matters. Cm. 5860. The Stationery Office, London.
DfES, 2005. Extended Schools: Access to Opportunities and Services for All. A Prospectus. DfES, London.
Dobbie, W., Fryer, R.G., 2011. Are high-quality schools enough to increase achievement among the poor? Evidence from the Harlem Children's Zone. American Economic Journal: Applied Economics 3, 158–187.
Dryfoos, J., 1994. Full-Service Schools. Jossey-Bass, San Francisco.
Dryfoos, J., 2005. Full-service community schools: a strategy – not a program. New Directions for Youth Development 2005, 7–14.
Dyson, A., Kerr, K., 2013. Developing Children's Zones for England: What's the Evidence? Save the Children, London.
Dyson, A., Kerr, K., Raffo, C., Wigelsworth, M., with Wellings, C., 2012. Developing Children's Zones for England. London, Save the Children.
Dyson, A., Todd, L., 2010. Dealing with complexity: theory of change evaluation and the full service extended schools initiative. International Journal of Research & Method in Education 33, 119–134.
Fielding, M., Moss, P., 2011. Radical Education and the Common School. Routledge, London.
Foot, J., Hopkins, T., 2010. A Glass Half-Full: How an Asset Approach Can Improve Community Health and Well-Being. IDeA, London.
Glassman, D., Naidoo, J., Wood, F. (Eds.), 2007. Reaching the Unreached: Save the Children's Community Schools in Africa. Springer, New York.
Gregoric, C., 2013. School-Community Involvement, Mitcham. UNESCO-APNIEVE, South Australia.
Harkavy, I., Hartley, M., Axelroth Hodges, R., Weeks, J., 2013. The promise of university-assisted community schools to transform American schooling: a report from the field, 1985–2012. Peabody Journal of Education 88, 525–540.
Heers, M., Van Klaveren, C., Groot, W., Van Den Brink, H.M., 2011. Community Schools Unfolded: A Review of the Literature. TIER Working Paper Series. TIER WP 11/04. Top Institute for Evidence Based Education (TIER), Maastricht.
Henderson, A.T., Mapp, K.L., 2002. A New Wave of Evidence: The Impact of School, Family, and Community Connections on Student Achievement: Annual Synthesis 2002. Southwest Educational Development Laboratory, Austin, Texas.
The Harlem Children's Zone, 2009. The HCZ Project: 100 Blocks, One Bright Future. the Harlem Children's Zone, New York.
The Scottish Office, 1998. New Community Schools: The Prospectus. The Stationery Office, Edinburgh.
International Project "International Standards For Community Schools: Strengthening The National Capacities" (no date). Case study manual. http://communityschoolstandards.org/data/Case%20Study%20Quality%20Research%20Manual_ENG.pdf (accessed 20.12.13.).
Joos, A., Ernalsteen, V., Lanssens, A., Engels, M., 2006. Community Schools in Flanders and Brussels. A Framework for Development. Flemish Government, Brussels.
Lawson, H.A., 2013. Third-generation partnerships for P-16 pipelines and cradle-through-career education systems. Peabody Journal of Education 88, 637–656.
Miller-Grandvaux, Y., Yoder, K., 2002. A Literature Review of Community Schools in Africa. USAID, Washington, DC.
Morris, H., 1925. The Village College. Being a Memorandum on the Provision of Educational and Social Facilities for the Countryside, with Special Reference to Cambridgeshire. Cambridge University Press, Cambridge.
Moss, P., Haydon, G., 2008. Every Child Matters and the Concept of Education. Institute of Education, London.
OECD, 2012. Equity and Quality in Education: Supporting Disadvantaged Students and Schools. OECD Publishing, Paris.
Regnier, R., 2012. SchoolPLUS: creating community schools through integral development. in education 18, 6778.
Rochex, J.-Y., 2012. Twenty-five years of priority education policy in France: dubious specificity and disappointing results. In: Demeuse, M., Frandji, D., Greger, D., Rochex, J.-Y. (Eds.), Education Policies and Inequalities in Europe. Palgrave Macmillan, Basingstoke.
Shah, S.C., Brink, K., London, R., Masur, S., Quishuis, G., 2009. Community Schools Evaluation Toolkit. Coalition for Community Schools, Washington, DC.
Stecher, L., SchüPbach, M., 2013. Research on extended education in Germany – a general model with all-day schooling and private tutoring as two examples. Journal for Research on Extended Education 31–52.
Tymchak, M. (Chair) Task Force and Public Dialogue on the Role of the School, 2001. SchoolPlus: A Vision for Children and Youth. Toward a New School, Community and Human Service Partnership in Saskatchewan. Final report to the Minister of Education, Government of Saskatchewan. Government of Saskatchewan, Saskatchewan.
Warren, M.R., Hong, S., Rubin, C.L., Uy, P.S., 2009. Beyond the bake sale: a community-based relational approach to parent engagement in schools. Teachers College Record 111, 2209–2254.
Whitehurst, G.J., Croft, M., 2010. The Harlem Children's Zone, Promise Neighborhoods, and the Broader, Bolder Approach to Education. Brown Center on Education Policy at Brookings, Washington, DC.
Wilkin, A., White, R., Kinder, K., 2003. Towards Extended Schools: A Literature Review. DfES, London.

Relevant Websites

http://www.ond.vlaanderen.be/bredeschool – Brede School in Vlaanderen en Brussel (Community Schools in Flanders and Brussels, Belgium).
http://www.education.manchester.ac.uk/research/centres/cee/ – Centre for Equity in Education, University of Manchester, England.
http://www.bc.edu/schools/lsoe/cityconnects – City Connects, USA.
www.communityschools.org – Coalition for Community Schools, USA.
www.hcz.org – Harlem Children's Zone, USA.
http://www.bredeschool.nl – Landelijk Steunpunt Brede Scholen (National Centre for Community Schools, Netherlands).

External Validity

Laura C Leviton, The Robert Wood Johnson Foundation, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

External validity assesses the extent to which a causal relationship can be generalized to and across populations, settings, and treatment and measurement variables. External validity is almost wholly dependent on information outside of single studies, requiring judgments about the plausibility that results generalize beyond single studies. A confident assessment of external validity requires examination of a body of evidence from other research (causal and descriptive), from experience, and in applied settings, from reflective practitioners. Three frameworks permit assessment of external validity. The first is sampling across units, treatments, observations, and, where feasible, the larger social context. The second is decision making under uncertainty, in which prior information and judgments are used to assess the plausibility of generalization. The third is the probing of complex interactions that can obscure or illuminate the assessment of similarity across samples and studies.

Definition and Challenges

External validity (along with statistical conclusion validity, internal validity, and construct validity) provides a framework to conduct and interpret empirical studies of causal relationships. External validity assesses 'to what populations, settings, treatment variables, and measurement variables can (the effect) be generalized?' (Campbell and Stanley, 1966, p. 5). External validity was later redefined as assessing 'whether the causal relationship holds over variation in persons, settings, treatments, and measurement variables' (Shadish et al., 2002, p. 20). Rules for inference about external validity decrease uncertainty that a cause–effect relationship holds under particular conditions *outside* a single study. External validity is not defined by the *extent* to which a cause–effect relationship generalizes across conditions, any more than internal validity is defined by low p values or large effect sizes. After all, when we know that a causal relationship does not generalize, the information is useful to refine theory or to prevent the useless expenditure of resources, just as internally valid information on the absence of an effect is useful for these reasons. Sometimes researchers apply the term 'external validity' to correlational studies; however, this article will deal only with the external validity of causal relationships. Historically, researchers have faced three challenges in assessing external validity.

The Challenge of Inductive Inference

External validity requires inductive inference, relying on the accrual of studies and extrapolation beyond the conditions and findings of a single study. Inductive logic about causation is never fully justified (Campbell and Stanley, 1966) (*see* Generalization: Conceptions in the Social Sciences). By contrast, internal validity is established through a more deductive framework whereby causal relationships can be both corroborated and falsified within a single study (*see* Internal Validity). Both internal and external validity depend on information that originates outside the single study, especially regarding judgments about what is already known. However, external validity is much more dependent on information from outside the study.

Perceived Trade-offs between Internal and External Validity

Internal validity is compromised if research does not involve controls and the measurement of outcomes. But many researchers are concerned that the artificiality of such studies does not reflect prevalent real-world conditions (e.g., Barker, 1968; Brunswik, 1955). External validity can often be compromised by the very presence of design controls, obtrusive measures, and restricting the study to respondents who are willing to receive any treatment to which they might be assigned. Researchers concerned about external validity have divided into those who set less priority on the study of causes, in general, and those who would maximize external validity by studying causal relationships in field settings. Some researchers put a higher priority on internal validity, arguing that there is little point in generalizing a causal conclusion about which there is legitimate doubt (Campbell and Stanley, 1966; Cook and Campbell, 1979). Others maintain that external validity deserves priority because the research should be a reflection of prevalent real-world situations, especially those that involve testing the value of educational, public health, and social services (Cronbach, 1982). The challenge then arises: What is the best balance in a given study? Cannot both internal and external validity be optimized somehow?

The Challenge of Complex Interactions

Causal relationships in real-world settings are at the very least complicated and often dynamic and complex (Rogers, 2008). Statistical interactions of variables are assumed to be pervasive (e.g., Brunswik, 1955; Cronbach, 1982). This means that the strength of a causal relationship is assumed to vary with the population, setting, or time represented within any given study, and with the researcher's choices about treatments and measurement of outcomes. Without a sensitive assessment of such interactions, true effects can be obscured or causal claims can be overgeneralized to a wider range of people, settings, times, treatments, or outcome constructs than is warranted. Unfortunately, the number of possible interactions is often endless (Cook, 1993). The challenge is: How do we cope with complexity?

A practical first step is to assess some generic interactions that are both frequent and plausible to explain variation in outcomes and the interpretation of causal relationships. Interactions occur often between treatment, populations, measures, and context, posing plausible *threats to external validity* (Campbell and Stanley, 1966; Cook and Campbell, 1979). Because external validity specifies the conditions under which an internally valid relationship can be reproduced, threats to external validity necessarily invoke internal validity.

The interaction of testing and treatment suggests that an effect size varies by the conditions of measurement.

The interaction of selection and treatment hypothesizes that an effect size varies by population studied.

The interaction of setting and treatment describes effects that vary by setting.

The interaction of history and treatment deals with the extent to which a causal relationship replicates across different times and events.

Other interactions between features of the study and treatment sometimes occur, but their plausibility depends on the individual study context and on what is generally known about these study features (*see* Internal Validity).

External Validity Involves Inferences from Samples

Generalizing from Single Studies to Samples, Classes, Categories, and Constructs

The statistician would say that researchers generalize about a sample they have observed of a population, usually of persons or other units being studied. However, external validity treats inferences about treatments, measures, and settings just as seriously as inferences about human population groups. It therefore assumes that studies have to sample from these domains as well as from populations (Cook, 1993). A given study therefore represents not only a sample from a population of units but also a sample from a universe of potential settings, a universe of potential treatments, and a universe of potential measures. The sampling frame becomes explicit whenever an investigator makes choices about whom to study, what treatment to administer, or which measures upon which to base inferences. Beyond the individual study, bodies of research (literature reviews, meta-analyses, and programs of related studies) are now understood to be samples from the universe of potential studies. The available sample of findings both limits and guides inferences about external validity.

The sampling frame for external validity considers not only categories but also theoretical constructs (Cronbach, 1982). Construct validity is therefore intimately tied to external validity. This becomes clear when we note that a population of persons is akin to a category of persons; treatments often represent a set of underlying principles that are made operational in a given study; settings can be classified into those that facilitate or impede access to the population and moderate the 'dose' or fidelity of treatment; and outcome measures often reflect bigger, more general underlying constructs, such as 'access to medical care' or 'youth resiliency.' Specifying these larger classes, categories, and constructs is necessary to justify a generalization about causal relationships (Shadish et al., 2002).

Methodologists have adopted the following notation from Cronbach (1982). Uppercase notation represents the universe or domain, while lowercase represents an element in a specific study. We will use the prevention of HIV/AIDS to illustrate.

U represents the population of units (persons, sites, communities, organizations, states, nations, or other settings and aggregates) about which a conclusion is sought. In contrast, *u* refers to the sample of units in a study. Thus, 'people at high risk of HIV infection' include men who have unprotected sex with men, injection drug users (IDUs) and their sex partners, sex workers, runaway homeless youth, and, internationally, communities where HIV infection is prevalent.

T represents the plan and the set of realizations for a treatment, experimental manipulation, program, or other intervention, while *t* is the single realization of *T* for a given *u*. In 2012, there were over 100 evidence-based programs for HIV prevention, most of which include a core of common behavioral principles. Yet to be effective, a realization of *T* must be adjusted to be relevant to the community's situation. A naturally occurring event can also be denoted *t*, such as a celebrity announcing that he is HIV infected, and *T* is the set of similar events.

O represents the admissible procedures and conditions for obtaining data (and the data themselves). In causal research, it is often used to refer to the possible effect, thus using *O* to represent the outcome construct and *o* the actual measure of effect. HIV prevention commonly uses self-report and partner report measures, supplemented by unobtrusive measures such as condom purchases or redemption of vouchers for condoms.

S represents the context or setting, including the culture and times in which the study takes place. Some researchers use *s* to denote an individual study's setting and *S* to denote the accrual of such settings (Campbell and Stanley, 1966). Others maintain that *S* is always confounded with *U* and *T*, and cannot be deliberately sampled in the same way as *U*, *T*, and *O*. The debate becomes clear in HIV prevention, where the passage of time (*S*) has led many individuals at risk (*U*) to become complacent, requiring a different realization of *T*. Also, *T* is most effective when communities (*U*) trust the organizations (*S*) that carry out *T*. Yet these organizations vary in their capacity to carry out *T*, thus affecting the realization of *t* (dose or fidelity of the treatment).

UTOS is the domain about which a research question is asked. Once *U*, *T*, *O*, and *S* are specified concretely, that constitutes a description of the causal question being researched. In contrast, an individual *utoS* is the combination of *u*, *t*, and *o* that constitutes an instance on which data are collected. Thus, it is the operational or sampling level of a study; it is how the researcher proposes to answer the question about *T* as a cause, *O* as an effect, *U* as a population of persons, and *S* as the cultural and historical contexts in which the hypothesis is tested.

Within *UTOS*, replications are of three kinds: (1) exact replication using fresh samples of *U*, *T*, and *O*; (2) studies by other investigators trying to repeat the *UTO* procedures according to the best description available; and (3) conceptual replication by other investigators who choose different procedures to represent the same *U*, *T*, and *O* based on their own judgment about what constitutes instances of each. In all three

cases, the interest lies in ascertaining whether a class of causal relationships was replicated. For example, the policy maker may want to know whether a treatment for IDUs in New York City will produce a reduction in behaviors that lead to HIV transmission. Replication of the treatment across fresh samples of IDUs in New York and other northeastern cities increases confidence about treatment effectiveness.

At the level of *UTO*, the instances are joint distributions of units, treatments, and observations. However, some combinations are more common than others. For instance, most HIV prevention studies initially used testing and counseling approaches offered through highly trusted organizations to gay and bisexual men in the United States. Later on, a wider range of organizations addressed additional groups at risk, adding social marketing and community-based interventions to testing and counseling. By studying categories within these joint distributions, called sub-*UTOS*, we may ask whether a causal relationship holds with different categories within *UTOS*. These lower order categories often include characteristics of the populations, variations in the content or level of a treatment, or dimensions of some set of measured responses. For example, one may study behavior changes in IDUs versus women in African villages, or in cases that experience a treatment at high and moderate levels of intensity and duration. Sub-*UTOS* provide a key to external validity because patterns of variation within and across studies allow us to ascertain some of the conditions under which a causal relationship holds.

Inferences about *UTOS* are both retrospective and prospective. They rely on replication studies that already exist and on a definition of constructs that incorporates these studies. Generalization from *utoS* to *UTOS* is not entirely bound by the existing studies, since inferences sometimes need to be made about levels of variables that were not directly observed, for example, when fourth-grade and sixth-grade children are studied, but not those in the fifth grade, or when 4 h of therapy are given versus 8 h. Moreover, a critical understanding of variation in *UTOS* always depends on construct validity and on theory building, which usually go beyond the observed *utoS*. Nevertheless, at some point causal hypotheses undergo a test that then extends external validity to new variations within *UTOS* – or expands *UTOS*. By contrast, other questions about external validity concern new domains that were not studied, and these are entirely prospective, or even speculative.

Generalizing from Existing Studies to New Domains

Cronbach (1982) distinguished two basic types of external validity inferences: those concerning the reproducibility of findings in *UTOS*, and those that go beyond *UTOS* to new domains termed **UTOS*. The asterisk indicates that causal relationships in these domains have not yet been studied, or have not been studied jointly. Inferences about **UTOS* pose challenges above and beyond those in *UTOS*, whether the finding involves an important relationship in basic social science or the effect of a social program or policy. Uncertainty is much greater about whether the causal relationship holds in **UTOS*. Take the example of HIV prevention: results may impress policy makers who are looking for ways to combat the global epidemic. Will social marketing of condoms, when modified for use (**T*) with village women in another country (**U* and **S*), also lead to reductions in the (very different) behaviors that put them at risk of HIV transmission (**O*)? This requires asking whether the cause–effect relationship generalizes to **UTOS*. The answer is not definitive until cause-probing studies of **UTOS* are conducted.

Decision makers and practitioners in education, public health, and social programs face acute uncertainties about such issues all the time, and increasingly call for better ways to cope (Glasgow et al., 2006). The key is to find new ways to reduce uncertainty. There is always some uncertainty about causation in the social sciences, since plausible alternative explanations may be discovered for even the best replicated cause–effect relationship (*see* Internal Validity). However, uncertainty increases about the conditions under which causal relationships hold, as investigators move from an observed set of *utoS*, to unobserved variations within a well-defined *UTOS*, to **UTOS* (Cronbach, 1982).

Bayesian statisticians offer a useful way to cope with uncertainty about generalization: first incorporate information to assess the prior likelihood that a finding will generalize, assess the degree of uncertainty about the prior likelihood, and assess the likely consequences of a decision about generalization that is correct or incorrect (Cronbach, 1982). Where the consequences are important (expenditure of public resources, a prevalent challenge for implementing *T*, or a new direction for theory), new information is needed. After a decision is taken, additional information becomes available: either a test of causal relationships or descriptive and experiential information that may help to further reduce uncertainty.

Studies to Reduce Uncertainty about *UTOS* and **UTOS*

Inferences about external validity can proceed only so far in the context of a single study. The challenges can be addressed more adequately through literature reviews and programs of studies, planned or unplanned. Multiple studies allow us to improve the *quality* of inductive inference about external validity. They also permit us to recognize potential trade-offs between internal and external validity and to *optimize* both (where feasible). Finally, they improve our ability to *disentangle the web of interactions* among variables and so identify more of the conditions under which a causal relationship varies in magnitude and even sign.

Ideally, we would choose a sampling frame to represent the universe of all possible versions of *utoS*. The advent of 'Big Data,' in which so much data about social phenomena are being captured, may some day permit this (McAfee and Brynjolfson, 2012). For example, the Obama 2012 campaign conducted over 100 randomized experiments on donating through social media at fairly low cost. However, most causal studies in the social sciences operate under conditions that prohibit representative sampling. Concerning *S*, we cannot sample historical occasions at random, and expense usually prohibits choosing a random sample of all relevant social situations in order to implement a treatment therein. Certain *u* are inaccessible to us, or we are prohibited from studying them in certain ways. Moreover, a sampling frame – of any kind – implies that we know the dimensions of greatest relevance for *UTOS*; in many cases, however, we do not.

A pragmatic alternative approach is to use the less sophisticated methods and principles typically used for generalizing from operational representations of a cause or effect construct. Social scientists routinely generalize from *t* to *T* and from *o* to *O* without random sampling, and they seem to feel very confident about the inferences they draw because the *constructs* make logical sense. Shadish et al. (2002) identified five principles that scientists use in making generalizations. We first present the principles, then the implications for generalizing about a body of evidence.

Five Principles to Generalize about Causal Inferences

1. *Surface similarity*. Some *utoS* look like the *UTOS or *UTOS* to which one would like to generalize. This process is known as pattern matching (Campbell, 1988). Prior theory and other forms of consensus in social science help to identify prototypes and principles with which to classify groups (*U*), treatments (*T*), measures (*O*), and settings (*S*) (Berkowitz and Donnerstein, 1982). Just as the early naturalists classified birds, so social scientists can begin to classify and generalize about *UTOS*.
2. *Ruling out irrelevancies*. Across many studies, there is both relevant and irrelevant variation in *UTOS*. Variation is irrelevant when it does not change a causal generalization. Thus, a treatment might be implemented in different ways, only some of which matter; outcomes might be measured in different ways; and the demographics of respondents might differ even though they all belong in the category about which one wants to generalize. Variation in *UTOS* provides an opportunity to test whether a causal relationship holds across the irrelevancies actually sampled (Cook, 1993; Kruglanski and Kroy, 1976). It is therefore a test of whether the relationship is robust despite irrelevancy in how measurement occurs, a treatment is operationalized, a sample is drawn, and settings are selected. For example, we know that various groups at risk of HIV/AIDS (*U*) respond well to messages delivered by people similar to themselves (*T*), although the venues vary tremendously (e.g., street, classroom, bar, and homeless shelter), and the specific risk behaviors (*O*) vary as well. The risk behaviors form a larger class, as do the populations and treatments, and the findings replicate across cultures and decades (*S*). For generalizing to **UTOS*, this robust type of finding is also helpful. Given the consistent findings, what is the prior likelihood that a new group at risk of HIV/AIDS (**U*) would also change risk behaviors when exposed to an exemplar of *T*? It is certainly not zero.
3. *Making discriminations*. Certain features of *UTOS* help us to limit generalizations and thus sharpen and better define *UTOS*. This is especially the case when similar concepts can be differentiated with respect to a causal relationship. So for example, one has to be explicit about how liking differs from love, or how one school reform program differs from another. For generalization about *UTOS*, one has to demonstrate empirically that the causal proposition one wants to defend holds for one concept and not another similar concept. Both theory and measurement are needed to support the discrimination required. To differentiate what can be generalized in **UTOS*, boundary variables are helpful; these are defined as 'critical differences between the research setting and criterion setting which could alter the existence or magnitude of the relationship between two or more variables' (Bernardin and Villanova, 1986, p. 43). For example, HIV prevention is unlikely to work when provided through organizations that do not have the training and supervisory capacity to implement it well (Miller et al., 2003).
4. *Interpolation and extrapolation*. Few social or behavioral theories succeed in specifying the levels of a treatment that will usually bring about a desired effect. Parametric studies that vary many levels on the independent variable and then examine response curves are rare: instead, just two or three levels are manipulated, leaving us unclear about how the outcome variable would have behaved at points between or beyond the levels measured. In principle, such interpolation is easy. For generalizing about *UTOS*, it is desirable to sample a wide range of treatment levels and as many points in between as are feasible. In the same way, we might sample a wide range of persons (e.g., on age or educational attainment), time periods (e.g., years after discovery of the AIDS epidemic), and levels of outcome variables that might otherwise be truncated. Again, 'Big data' may eventually free external validity from this limitation (McAfee and Brynjolfsson, 2012)

Extrapolation of effects that are beyond the ranges available in studies is more problematic, because the shape of the curve or the nature of the effect itself may change at more extreme values on the independent variable. Thus, water turns to ice at 0 °C and to vapor at 100 °C. Between these points, it is water. But few social science relationships understood this well. We simply do not know where qualitative transformations come about in many social relations (e.g., how much life stress is needed to create clinically significant depression). Extrapolation to points closer at hand (*UTOS*) is generally less uncertain than extrapolation to more distant points (**UTOS*).

5. *Causal explanation*. One of the major goals of science is to explain phenomena, including causal relationships. Identifying the how or why of a relationship helps to replicate or extrapolate about the relationship, even in circumstances that are quite different. For instance, knowing why key messages cause people to protect themselves from HIV infection might help to adjust these messages to meet new circumstances over time. This is why causal explanation is held to be the major mechanism for moving from *utos* to **UTOS*, for generalizing to as-yet-unstudied circumstances (Cronbach, 1982). Thus, we judge social and behavioral theories by their capacity to specify constructs that moderate or mediate a causal relationship. However, it is rare to find studies that have manipulated all of these causal-explanatory contenders or even a goodly portion of them. Some are usually measured, though, and so it becomes important to use measurement and data analysis to establish which components are most likely to be causally responsible for an effect and so are essential to any practical generalizations that will be made from the data.

Using the Principles to Generalize from a Body of Evidence

Using these five principles, generalizations about *UTOS* can be made from existing research and through purposive sampling of new *utoS*. For extrapolating to **UTOS*, the research community

uses prior information and best judgment to determine the plausibility or implausibility of generalization (Cronbach, 1982).

In order to utilize the five principles for external validity, literature reviews and meta-analyses need to be thoroughly informed about the details of each *utoS*. It is on this basis that readers can be confident about the surface similarities of studies, the irrelevancies, and the features that discriminate key concepts. Yet many meta-analyses cannot go beyond the very surface of surface similarity, because the journal articles and other materials on which they rely do not provide enough detail, particularly about context and implementation (e.g., Ogrinc et al., 2008). Sometimes the details of *utoS* can have great importance – quality of implementation, for example (Lipsey, 2009). Also, causal explanations may emerge only as the body of research grows, and it is seldom possible to go back in time to study new mediators or moderators of the effect. In addition to these limitations, Shadish et al., (2002) offer a host of threats to the external validity of literature reviews, including sampling bias in *UTOS*, restricted variation in *UTOS*, failure to test for heterogeneity in effect sizes (which might lead to a premature conclusion that the effect is generalizable), lack of statistical power for studying sub-*UTOS*, and restricted variation in the irrelevancies that give confidence that the cause–effect relationship is a robust one.

A plan for new studies to probe external validity will sample purposively from the range of possible *utoS* based on surface similarity, by selecting modal or typical instances. Purposive sampling can offer a stronger test of external validity by introducing as much irrelevant variation as possible. And by sampling for specific cognate concepts, planned studies could help discriminate the limits of the causal relationship. With regard to interpolation and extrapolation, the Multiphase Optimization Strategy may be helpful (Collins et al., 2011). It uses preparatory steps to select varying doses of real-world social and health interventions, in a way that culminates in randomized experiments with the best potential to inform external validity. This strategy also assists discrimination by unpacking treatments and administering them with and without components that might be critical to the outcomes. Finally, causal explanation can be addressed by examining the deep or structural similarities and differences inherent in the samples of *utoS*. For example, cultural adaptation of an intervention may be required in order to include Latinos in a study; one wants to go beyond surface similarity to make sure that Latinos will respond to the adapted materials in the intended ways.

For generalizing to **UTOS*, it is necessary to rely on information beyond the corpus of cause-probing studies, by using sources such as surveys and qualitative studies. Applied fields do not generally do this very well, although good models exist. For example, social marketing in health promotion routinely employs a core set of behavioral principles, but through focus groups, rapid ethnography, and other information, the principles are systematically adapted to address the needs and concerns of the population and the setting (e.g., Leviton et al., 1999). For generalizing about the effects of social interventions, the researcher can also access the experience of expert, reflective practitioners (Schön, 1984). It is noteworthy that discussions of external validity in social programs almost completely exclude practitioners' opinions. Yet reflective practitioners have seen many examples of *utoS* or **utoS*, and as experts in service provision, they will perceive patterns in *UTOS* that the researcher does not recognize. Expert practitioners often have theories in use that can be elicited and utilized on behalf of external validity – an issue of particular importance for the quality of implementation, and for assessing complex and emerging contexts. In applied areas, all these sources of information can be incorporated into theories of change or logic models to assess the plausibility of achieving an effect in a particular **utoS* (Leviton et al., 2010). For example, condom promotion for HIV prevention may be unworkable in certain cultures – or must be greatly adapted to be feasible. Just as alternative explanations can be ruled out to assess internal validity, such information helps the decision maker rule out certain **utoS* with a fair degree of certainty ('You can't get there from here'). Thus, the limits of generalizing about **UTOS* are further described; the challenge comes in *ruling in* **utoS*!

Descriptive information and reflective practice can be equally useful after the fact. In the absence of a controlled study, experience provides useful information for external validity, just as the Bayesian statistician incorporates additional information into an *a posteriori* assessment of probabilities. It is on this basis that practitioner feedback about evidence-based practices should be assessed. Where there are prevalent barriers to implementation, these need to be assessed. Where there are prevalent contexts that moderate effects, we need to know about this. Yet we do not abandon an eventual test of **UTOS*; instead, we have a much better sampling frame to maximize useful information.

See also: Content Validity; Generalization: Conceptions in the Social Sciences; Internal Validity.

Bibliography

Barker, R.G., 1968. Ecological Psychology: Concepts and Methods for Studying the Environment of Human Behavior. Stanford University Press, Stanford CA.

Berkowitz, L., Donnerstein, E., 1982. External validity is more than skin deep. American Psychologist 37, 245–257.

Bernardin, H.J., Villanova, P., 1986. Performance appraisal. In: Locke, E.A. (Ed.), Generalizing from Laboratory to Field Settings. Lexington Books, Lexington, MA.

Brunswik, E., 1955. Representative design and probabilistic theory in a functional psychology. Psychological Review 62, 193–217.

Campbell, D.T., 1988. Methodology and Epistemology for Social Science: Selected Papers. University of Chicago Press, Chicago.

Campbell, D.T., Stanley, J.C., 1966. Experimental and Quasi-Experimental Designs for Research. Rand McNally, Chicago.

Collins, L.M., Baker, T.B., Mermelstein, R.J., Piper, M.E., Jorenby, D.E., Smith, S.S., Schlam, T.R., Cook, J.W., Fiore, M.C., 2011. The multiphase optimization strategy for engineering effective tobacco use interventions. Annals of Behavioral Medicine 41, 208–226.

Cook, T.D., 1993. A quasi-sampling theory of the generalization of causal relationships. In: Sechrest, L.B., Scott, A.G. (Eds.), Understanding Causes and Generalizing about Them. New Directions for Program Evaluation, vol. 57, pp. 39–82.

Cook, T.D., Campbell, D.T., 1979. Quasi-Experimentation. Rand McNally, Chicago.

Cronbach, L.J., 1982. Designing Evaluations of Educational and Social Programs. Jossey-Bass, San Francisco.

Glasgow, R.E., Green, L.W., Klesges, L.M., Abrams, D.B., Fisher, E.B., Goldstein, M.G., Hayman, L.L., Ockene, J.K., Orleans, C.T., 2006. External validity: we need to do more. Annals of Behavioral Medicine 31, 105–108.

Kruglanski, A.W., Kroy, M., 1976. Outcome validity in experimental research: a reconceptualization. Representative Research in Social Psychology 7, 166–176.

Leviton, L.C., Finnegan, J.R., Zapka, J.G., 1999. Formative research methods to understand patient and provider responses to heart attack symptoms. Evaluation and Program Planning 22, 85–398.

Leviton, L.C., Kettel Khan, L., Rog, D., Dawkins, N., Cotton, D., 2010. Evaluability assessment to improve public health. Annual Review of Public Health 31, 213–234.

Lipsey, M.W., 2009. The primary factors that characterize effective interventions with juvenile offenders: a meta-analytic overview. Victims and Offenders 4, 124–147.

McAfee, A., Brynjolfsson, E., 2012. Big data: the management revolution. Harvard Business Review 90, 60–66.

Miller, R.L., Bedney, B.J., Guenther-Grey, C., 2003. Assessing organizational capacity to deliver HIV prevention services collaboratively: tales from the field. Health Education and Behavior 30, 582–600.

Ogrinc, G., Mooney, S.E., Estrada, C., Foster, T., Goldmann, D., Hall, L.W., Huizinga, M.M., Liu, S.K., Mills, P., Neily, J., Nelson, W., Pronovost, P.J., Provost, L., Rubenstein, L.V., Speroff, T., Splaine, M., Thomson, R., Tomolo, A.M., Watts, B., 2008. The SQUIRE (Standards for QUality Improvement Reporting Excellence) guidelines for quality improvement reporting: explanation and elaboration. Quality and Safety in Health Care 17 (Supplement 1), i13–i32.

Rogers, P.J., 2008. Using programme theory to evaluate complicated and complex aspects of interventions. Evaluation 14, 29–48.

Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Sage, Thousand Oaks.

Schön, D., 1984. The Reflective Practitioner. Basic Books, New York.

Extraversion

Caroline Fielden, Lisa E Kim, and Carolyn MacCann, University of Sydney, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R.E. Lucas, volume 8, pp. 5202–5205, © 2001, Elsevier Ltd.

Abstract

Extraversion is a term often associated with a range of human behaviors and experiences. Biological, environmental, and even developmental features have variously been identified as important contributors to variance in these. Our understanding of how these features form an overall structure is, however, still developing. Attempts to better understand extraversion as a complex, dynamic system often one means by which the breadth of knowledge accumulated thus far might be integrated.

Extraversion is a personality trait that broadly describes differences between people in their social interactions, positive emotions, impulsivity, and energy levels. A highly extraverted person is generally thought to be outgoing, energetic, enthusiastic, and positive. Someone with low levels of extraversion (sometimes called introverted) is thought to be quiet, shy, calm, and slow-paced. Extraversion has long been recognized to be an important feature of human personality. From the typologies offered by the ancient Greeks and Romans, through to the sophisticated models of today, the behavioral and experiential variance described by extraversion has been considered an important descriptor of individual differences. The paragraphs below summarize several different approaches to understanding extraversion, including psychoanalytic, psychometric, and biological approaches. Different models of extraversion are compared, real-life outcomes associated with extraversion are described, and future directions in extraversion research considered.

Historical Overview of Extraversion

Ancient Typological Accounts

The earliest descriptions of personality were those of Hippocrates and Galen, who proposed that people could be classified as one of four personality types, each associated with a physical element: sanguine (air), choleric (fire), melancholic (earth), and phlegmatic (water). Each type was associated with particular physical features. Both sanguine and choleric types contained elements now considered aspects of extraversion. Sanguine types of people were social, expressive, and positive in outlook. Choleric types were energetic, wilful, and passionate.

Psychoanalytic Accounts

The earliest personality model to explicitly describe extraversion was Jung's analytical psychology. He proposed 16 personality types, these being combinations of four more basic, dichotomous sets of individual characteristics: extroversion/introversion, sensing/intuition, thinking/feeling, and judging/perceiving. According to Jung (1916), for each of these dichotomous pairs, individuals could be classified as one or the other: extrovert or introvert, judger or perceiver, and so on. This means that while Jung regarded extroversion to be a fundamental aspect of human personality, he described it as one that characterizes some, but not all people. Those who are not extroverts, Jung described as introverts. The fundamental distinction between these targeted differences in the value individuals attach to external versus internal stimuli. These differences were thought to account for important differences in how people behave. Thus, extroverts were thought to primarily value the external world, causing them to direct their energy outward. Introverts, on the other hand, were thought to primarily value the internal world, leading to an inner focus. Jung's theory formed the theoretical basis of the extraversion/introversion classification in the widely used Myers-Brigg's Personality Test.

Early Lexical Accounts

In contrast to the earlier approaches, the lexical approach to personality viewed extraversion as a quality that all people possess to different degrees. This is a dimensional approach rather than a typological approach to personality. The lexical approach uses everyday language as its starting point. The premise of this approach is that important psychological traits are likely to be encoded in everyday language. For example, the existence of trait adjectives such as 'energetic,' 'positive,' and 'excitable' suggests that differences in these traits provide useful information about what people are like. Allport and Odbert (1936) identified over 4500 such adjectives within the English language. Cattell (1943) later examined this word-list using factor analysis – a statistical technique for grouping a large number of variables into a smaller number of factors, based on the strength of association. By doing this, terms that describe the primary factors in which individuals differ were identified. For example, it was noted that people's ratings of others on the terms 'happy,' 'positive,' and 'enthusiastic' were generally similar, as were ratings on the terms 'reliable,' 'industrious,' and 'consistent.' This was taken as evidence that individual differences in two factors might explain differences in how people are rated in relation to these six adjectives.

Cattell's (1945, 1947) analyses suggested that somewhere between 12 and 20 factors could account for all personality differences between people. In these early analyses, extraversion did not emerge as a single factor but was represented by several related factors (e.g., dominance, liveliness). Later factor

analyses found that five factors best explain associations among adjective ratings, and extraversion emerged as one of these five key factors (Tupes and Christel, 1961; Norman, 1963). Newer lexical models suggest a six-factor model, but still include extraversion as a key personality factor (Lee and Ashton, 2004).

Early Biological Accounts: Eysenck

Eysenck (2008/1967) believed that differences in extraversion were due to physiological differences in brain systems that caused some people to be more easily aroused than others. Specifically, he proposed that the ascending reticular activating system (ARAS) regulated arousal. People differ in the sensitivity of this system, which causes them to respond differently to their environment. A moderately stimulating environment might cause people with a very sensitive ARAS to feel overstimulated and retreat (i.e., to behave like introverts), but people with a less sensitive ARAS to feel understimulated and seek out additional stimulation (i.e., to behave like extraverts). In this model, the key characteristic of extraversion is impulsivity – social characteristics like friendliness and gregariousness were not originally considered as part of the extraversion trait.

According to Eysenck (1965), differences in extraversion predict how easily people form automatic associations. One classic experiment presented a tone and a puff of air to the eyeball several times. The puff of air caused people to blink. When the tone had become associated with the puff of air, people would blink when they heard the tone, even if the puff of air was not present. People with low extraversion were quicker to form this automatic association, possibly because they were more aroused by the stimuli in the first place. That is, extraversion moderates the ease of classical conditioning.

Further Development of the Biological Approach

Reinforcement Sensitivity Theory

Gray (1970, 1987) extended Eysenck's research on extraversion and conditioning by distinguishing between positive and negative stimuli. He claimed that extraverts may be slower to form associations only for negative stimuli (like the puff of air), but are faster to form associations when stimuli are positive. That is, conditioning is controlled by two systems: (1) one that is reactive to positive stimuli and (2) one that is reactive to negative stimuli. The system representing reward sensitivity (i.e., reactivity to positive stimuli) was called the behavioral activation system (BAS), while the system representing punishment sensitivity (i.e., reactivity to negative stimuli) was called the behavioral inhibition system (BIS). This theory is known as reinforcement sensitivity theory (RST).

According to RST, these systems control the conditioning process by differentially reinforcing particular behaviors. Thus, those with a sensitive BAS system were thought to learn more readily from positive stimuli and therefore be more prone to impulsivity. Those, however, with a sensitive BIS system were thought to learn more readily from negative stimuli and therefore be more prone to anxiety and caution generally.

By this account then, extraversion was not regarded as a fundamental personality trait. Instead, impulsivity and anxiety were considered primary, and extraversion was described as a construct that combines both. RST then described extraversion in terms of a low tendency toward anxiety (or a low degree of punishment sensitivity) and a moderate tendency toward impulsivity (or a moderate degree of reward sensitivity).

Revised Reinforcement Sensitivity Theory

RST was later revised and is now known as revised reinforcement sensitivity theory (RST-r; Gray and McNaughton, 2000). In this revision, the two basic systems of reward sensitivity and punishment sensitivity were reinterpreted to represent approach and withdrawal tendencies. Approach tendencies were still thought to be expressive of the BAS, but withdrawal tendencies were thought to be expressive of a fight, flight, freeze system (FFFS). The BIS was still a feature of the theory, but was reconceptualized as a sophisticated system that resolves conflict between the BAS and FFFS. Such conflicts were thought to occur in two circumstances: (1) when more than one positive or negative stimulus is encountered, requiring a choice between the two and (2) when both systems are activated – for example when a person is attracted to someone, but is simultaneously afraid of rejection. When such a conflict occurs, the BIS was thought to inhibit both the approach and withdrawal systems in an iterative manner until a single behavior dominates. RST-r, therefore, suggests the personality system as a whole to be both complex (made up of many parts) and dynamic (changing over time).

This new theory led some researchers to reexamine whether extraversion is a primary personality trait. Yet many studies show that measures of the BAS are strongly related to extraversion. Such results suggest that extraversion may actually be the fundamental trait for reward sensitivity (Smillie et al., 2006).

The Underlying Structure of Extraversion

Extraversion is a broad personality trait that reflects an integrated network of narrower constructs (e.g., sociability, positive affect, impulsivity, and reward sensitivity). Perhaps the first empirical exploration of the underlying structure of extraversion was the factor analysis of Guilford and Guilford (1934). They found evidence for up to 18 underlying elements of extraversion. The four most important elements represented an approach tendency, an anxiety/withdrawal tendency, impulsiveness, and self-interest.

Many different personality models include extraversion as a primary trait. However, different models propose different underlying elements of extraversion, usually referred to as facets. Table 1 compares several different faceted models of extraversion, showing the conceptual similarity between different facets. All models include sociability, which represents a preference for social interaction. There is much less consensus regarding the other facets. For example, some models propose surgency (a type of reward sensitivity that produces positive affect) as the defining feature of extraversion, whereas other models do not include surgency at all (Goldberg, 1990). Similarly, while Tellegen (1985) proposed

Table 1 Conceptual representation of the facets of extraversion in various models of extraversion

	Eysenck personality questionnaire-revised (Eysenck et al., 1985)	*Sixteen personality factor questionnaire (Cattell et al., 1988)*	*Multidimensional personality questionnaire (Tellegen, 1985)*	*Revised NEO personality inventory (Costa and McCrae, 1992)*	*HEXACO personality inventory (Lee and Ashton, 2004)*
Sociability	Sociable	Warmth	Social closeness	Warmth Gregariousness	Sociability
Dominance	Dominant Assertive	Dominance	Social potency	Assertiveness	
Social boldness		Social boldness			Social boldness
Positivity	Surgent		Well-being	Positive emotions	
Energy	Active Lively	Liveliness		Activity	Liveliness Expressiveness
Excitement-seeking	Sensation-seeking Venturesome			Excitement-seeking	
Carefree	Carefree				
Self-reliance		Self-reliance			
Achievement			Achievement		

that positivity was the defining feature of extraversion and eliminated sensation-seeking elements from his model, the recent HEXACO model does not include positivity as part of extraversion at all.

DeYoung et al. (2007) recently suggested an intermediate level between narrow facets of personality (such as impulsivity) and broad domains of personality (such as extraversion). They propose that there are two aspects that underlie each personality domain. Their model is an extension of the NEO-PI-R model of personality shown in **Table 1**. In the NEO-PI-R model, there are six specific facets that underlie extraversion: warmth, gregariousness, assertiveness, positive emotions, activity, and excitement seeking. De Young et al. propose that these can be grouped into two aspects of enthusiasm and assertiveness. Enthusiasm is an aggregate of the gregariousness, positive emotions, warmth, and excitement-seeking facets. Assertiveness is an aggregate of the excitement seeking, activity, and assertiveness. Extraversion is the only one of the five domains where a facet (excitement seeking) contributes to both aspects.

Extraversion across the Lifespan

Although the domain of extraversion appears stable over time, the lower-order facets may show systematic age-related change (Soto et al., 2011). A longitudinal study of age-related personality change in adults found that (1) excitement seeking decreases after 30 years of age; (2) activity is largely stable until 40 years of age, when it begins to decrease; (3) positive emotions and warmth increase until about 50 years of age, with positive emotions decreasing from 60 years and warmth decreasing from 70 years; and (4) assertiveness increases until 60 years of age and then decreases slowly (Terracciano et al., 2005). Cross-sectional research found that while middle-aged people and older people have similar levels of gregariousness and assertiveness, older people have lower levels of warmth, activity, excitement seeking, and positive emotions than middle-aged people (Roepke et al., 2001). These results suggest that at the facet level, extraversion continues to change across the adult lifespan.

Extraversion and Valued Outcomes

Extraversion predicts valued outcomes at work, school, and in personal life. For work outcomes, the largest effects are for leadership and job satisfaction. Of the big five personality domains, extraversion shows the strongest relationship with transformational leadership and among the strongest with job satisfaction (Bono and Judge, 2004; Judge et al., 2002). Extraversion shows a positive association with academic performance in the early years of school but a negative association in high school and university (Poropat, 2009). This may be due to the different mechanisms underlying academic achievement at different stages. For example, socialization with teacher and peers may be a key mechanism for achievement in the early years of school, whereas a preference for socializing at university may have a detrimental effect on grades. Extraversion also relates to subjective well-being, such as life satisfaction and happiness (Cooper and DeNeve, 1998).

There is some empirical evidence that different facets of extraversion show different relationships with valued outcomes. For example, a recent meta-analysis predicted job performance from the six extraversion facets and two extraversion aspects of DeYoung et al.'s (2007) aspects model (Judge et al., 2013). Both aspects of extraversion showed a small association with performance, but this was mainly accounted for by the facets of positive emotions and activity. In fact, the excitement-seeking facet was associated with poorer job performance. These results highlight even when extraversion shows a zero or very small relationship with a valued outcome; some of the underlying elements (either facets or aspects) might still be important predictors.

Extraversion: A Complex, Dynamic System?

A complex, dynamic system is one with multiple underlying features that change over time. Extraversion can be regarded as a complex, dynamic system as (1) it has multiple facets underlying the construct as a whole (i.e., it is complex) and (2) these facets change over time (it is dynamic). For this reason, extraversion might be explained in terms of Cloninger's et al. (1993) Psychobiological Model of Temperament and Character. This

model describes personality as a complex dynamic system dependent on two sets of characteristics. First, four biologically based temperaments drive personality (novelty seeking, harm avoidance, reward dependence, and persistence). Second, the expression of these temperaments is modifiable by three character traits (self-directedness, cooperativeness, and self-transcendence). These character traits describe how an individual learns and adapts to their environment. Two of the temperaments (novelty seeking and reward dependence) relate to extraversion and one of the character traits (cooperativeness) relates to reward dependence (Stallings et al., 1996; DeFruyt et al., 2000). Therefore, changes in extraversion facets across the lifespan can thus be described in terms of a dynamic system involving such temperament and character traits.

Complex, dynamic systems have many properties that set them apart from simple systems. These vary according to the type of system being theorized, but include nonlinearity (producing outcomes that are disproportionate to their inputs), fractality (having self-similar patterns), and even chaos (being determined, but unpredictable) (Gregson and Guastello, 2010; van Geert and Steenbeck, 2008). The presence of any of these three features would have important ramifications for how extraversion is understood.

However, methodologies designed to identify these or related features of complex dynamic systems have yet to be applied to research on extraversion. It is possible that a complex systems approach may be able to integrate multiple findings into a single theory, explaining developmental changes in extraversion facets, different facets' relationships with valued outcomes (e.g., workplace performance and academic achievement), as well as behavior-environment contingencies. Complex system methods may thus represent a promising new direction for identifying the structures and operations that underlie extraversion.

See also: Emotional Regulation; Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis; Factor Analysis and Latent Variable Models in Personality Psychology; Five Factor Model of Personality, Assessment of; Five Factor Model of Personality, Facets of; Personality Assessment: Overview; Personality Development: Systems Theories; Personality and Educational Outcomes; Personality, Biological Models of; Personality, Trait Models of; Personality: Historical and Conceptual Perspectives; Self-Organizing Dynamical Systems; Temperament.

Bibliography

Allport, G., Odbert, H., 1936. Trait-names a Psycho-lexical Study. Psychological Review Company, Princeton, NJ.

Bono, J.E., Judge, T.A., 2004. Personality and transformational and transactional leadership: a meta-analysis. Journal of Applied Psychology 89, 901–910.

Cattell, R., 1943. The description of personality: basic traits resolved into clusters. The Journal of Abnormal and Social Psychology 38 (4), 476–506.

Cattell, R., 1945. The description of personality: principles and findings in a factor analysis. The American Journal of Psychology 58 (1), 69–90.

Cattell, R., 1947. Confirmation and clarification of primary personality factors. Psychometrika 12 (3), 197–220.

Cattell, R.B., Eber, H.W., Tatsuoka, M.M., 1988. Handbook for the sixteen personality factor questionnaire (16 PF). Institute for Personality and Ability Testing Champaign, Illinois.

Cloninger, C.R., Svrakic, D.M., Przybeck, T.R., 1993. A psychobiological model of temperament and character. Archives of General Psychiatry 50, 975–990.

Cooper, H., DeNeve, K.M., 1998. The happy personality: a meta-analysis of 137 personality traits and subjective well-being. Psychological Bulletin 124, 197–229.

Costa, P.T., McCrae, R.R., 1992. Four ways five factors are basic. Personality and Individual Differences 13 (6), 653–665.

DeFruyt, F., VanDeWiele, L., VanHeeringen, C., 2000. Cloninger's psychobiological model of temperament and character and the five-factor model of personality. Personality and Individual Differences 29, 441–452.

DeYoung, C.G., Quilty, L.C., Peterson, J.B., 2007. Between facets and domains: 10 aspects of the Big Five. Journal of Personality and Social Psychology 93, 880–896.

Eysenck, H., 1965. Extraversion and the acquisition of eyeblink and GSR conditioned responses. Psychological Bulletin 63, 258–270.

Eysenck, H., 1998. Dimensions of Personality. Transaction Publishers, New Brunswick, NJ.

Eysenck, H., 2008. The Biological Basis of Personality. Transaction Publishers, Somerset, NJ.

Eysenck, S.B.G., Eysenck, H.J., Barrett, P., 1985. A revised version of the psychoticism scale. Personality and Individual Differences 6 (1), 21–29. http://dx.doi.org/10.1016/0191-8869(85)90026-1.

Goldberg, L., 1990. An alternative "Description of Personality": The Big-Five factor structure. Personality Processes and Individual Differences 59 (6), 1216–1229.

Gray, J.A., 1970. The psychophysiological basis of introversion-extraversion. Behaviour Research & Therapy 8 (3), 249–266.

Gray, J.A., 1987. The Neuropsychology of Anxiety. Oxford University Press, Oxford.

Gray, J.A., McNaughton, N., 2000. The Neuropsychology of Anxiety, second ed. University Press, Oxford.

Gregson, A.M., Guastello, S.J., 2010. Introduction to nonlinear dynamical systems analysis. In: Guastello, S.J., Gregson, R.A. (Eds.), Nonlinear Dynamical Systems Analysis for the Behavioural Science Using Real Data. CRC Press, Florida, USA.

Guilford, J.P., Guilford, R.B., 1934. An analysis of the factors in a typical test of introversion/extroversion. The Journal of Abnormal and Social Psychology 28, 377–399.

Judge, T.A., Heller, D., Mount, M.K., 2002. Five-factor model of personality and job satisfaction: a meta-analysis. Journal of Applied Psychology 87, 530–541.

Judge, T.A., Rodell, J.B., Klinger, R.L., Simon, L.S., Crawford, E.R., 2013. Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. Journal of Applied Psychology 98 (6), 875–925.

Jung, C., 1916. The content of psychosis. In: Long, C.E. (Ed.), Analytical Psychology. Moffat Yard & Company, New York.

Lee, K., Ashton, M.C., 2004. Psychometric properties of the HEXACO personality inventory. Multivariate Behavioral Research 39, 329–358.

Norman, W.T., 1963. Toward an adequate taxonomy of personality attributes: replicated factor structure in peer nomination personality ratings. Journal of Abnormal and Social Psychology 66, 574–583.

Poropat, A., 2009. A meta-analysis of the five-factor model of personality and academic performance. Psychological Bulletin 135, 322–338.

Roepke, S.L., McAdams, L.A., Lindamer, L.A., Patterson, T.L., Jeste, D.V., 2001. Personality profiles among normal aged individuals as measured by the NEO-PI-R. Aging & Mental Health 5 (2), 159–164.

Smillie, L.D., Pickering, A.D., Jackson, C., 2006. The new reinforcement sensitivity theory: Implications for personality measurement. Personality and Social Psychology Review 10 (4), 320–355.

Soto, C., John, O., Gosling, S., Potter, J., 2011. Age differences in personality traits from 10 to 65: Big five domains and facets in a large cross-sectional sample. Journal of Personality and Social Psychology 100 (2), 330–348.

Stallings, M., Hewitt, J., Cloninger, C.R., Heath, A., Eaves, L., 1996. Genetic and environmental structure of the tridimensional personality questionnaire: three or four temperament dimensions? Journal of Personality and Social Psychology 70 (1), 127–140.

Tellegen, A., 1985. Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In: Hussain, T.A., Maser, J.D. (Eds.), Anxiety and the Anxiety Disorders. Lawrence Erlbaum Associates, Hillsdale, NJ, England, pp. 421–452.

Terracciano, A., McCrae, R., Brant, L., Costa, P., 2005. Hierarchical linear modeling analyses of the NEO-PI–R scales in the Baltimore longitudinal study of aging. Psychology and Aging 20 (3), 493–506.

Tupes, E.C., Christal, R.E., 1961. Recurrent Personality Factors Based on Trait Ratings (Technical Report No. ASD-TR-61–97). Air Force, Lack- land Air Force Base, TX.

Van Geert, P., Steenbeek, H., 2008. A complexity and dynamic systems approach to developmental assessment, modeling and research. In: Battro, A.M., Fischer, K.W., Léna, P. (Eds.), The Educated Brain: Essays in Neuro-education. Cambridge University Press, Cambridge UK, pp. 71–94.

Relevant Websites

http://www.apa.org/topics/personality – The American Psychological Association Page on Personality. This is a summary page of the American Psychological Association (APA) linking to research and practice on personality.

http://www.cymeon.com/rrst – CYMEON Research rRST Profile. At this link, you can take a personality test based on the revised Reinforcement Sensitivity Theory personality model and get feedback on your personality scores.

http://www.hexaco.org – The HEXACO website. This link describes the research, theory, and personality tests used in the recent six-factor HEXACO model of personality.

http://www.ipip.ori.org – The International Personality Item Pool. This website provides multiple public-domain personality test items representing the major measures of personality. These are free to use for research purposes.

http://www.personal.psu.edu/j5j/IPIP – The International Personality Item Pool NEO Personality Inventory. At this link, you can take a personality test based on the NEO-PI-R personality model and get feedback on your personality scores at the domain and facet level.

https://personality-project.org – The Personality Project. This is a collection of resources that provides a broad array of methods and theory for the study of personality.

http://www.youtube.com/watch?v=K-HSiZUxTIk – 1980 video footage of Hans Eysenck discussing the Biological Basis of Personality.

http://www.youtube.com/watch?v=Au-1LJP1bAU – 2013 video footage of William Revelle giving the Hans J Eysenck Lecture at the International Society for the Study of Individual Differences (ISSID) in Barcelona, 2013.

Eye Movement, Control of: Oculomotor Control

Bryan Kolb, University of Lethbridge, Lethbridge, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P.W. Glimcher, volume 8, pp. 5205–5208, © 2001, Elsevier Ltd.

Abstract

Six muscles control the movements of each eye, allowing two broad categories of movements: gaze-stabilization and gaze-shifting movements. Gaze-stabilization movements act to compensate for movements of the head and body, acting to stabilize the image of the visual world on the retina. Gaze-shifting movements shift the fovea into alignment with objects of interest in the visual world. Eye movements are controlled by three cranial nerve nuclei (the oculomotor (III), trochlear (IV), and abducens (VI)), which together control the six eye muscles. The oculomotor system provides an excellent model for understanding the principles of movement control.

Eye Movements

Five types of eye movements are generally recognized.

These movements fall into two broad categories: gaze-stabilization movements and gaze-shifting movements. Gaze-stabilization movements rotate the line-of-sight to compensate for movement of the head and body, stabilizing the visual world on the retina. In the absence of these movements, visual acuity is severely compromised because visual images move across the retina whenever the observer moves through the visual environment. In almost all vertebrates, two neural systems work together to achieve gaze stabilization: the vestibulo-ocular system and the optokinetic system. Gaze-shifting movements, unlike gaze-stabilization movements, shift the fovea into alignment with objects of interest in the visual world. The three classes of movements that achieve gaze shifts are saccades, smooth pursuit eye movements, and vergence eye movements. Saccades are rapid shifts of the line-of-sight from one point in the visual world to another. Smooth pursuit eye movements allow the line-of-sight to track moving objects, slowly rotating the eye to keep a visual stimulus, which is moving across the visual field aligned with the fovea. Together saccades and smooth pursuit movements are referred to as versional movements because they cause both eyes to move in a similar direction and at the same speed. Vergence eye movements that occur only in binocular animals like humans, shift the lines-of-sight of the two eyes with regard to each other (either converging or diverging the angle between the two eyes) so that a visual stimulus can be aligned with the same point on each retina, even if that stimulus moves toward or away from the observer. In the nineteenth century, Herring proposed that only vergence movements involve the independent control of both eyes.

For all other movements, he proposed that the two eyes always receive identical movement commands. This hypothesis, which still informs our basic view of oculomotor control, is known as Herring's law of equal innervation (but see recent challenges to the universality of the law, King, 2011).

The Musculature of the Eyes

In humans, all eye movements are rotations of the eye (never linear motions or translations) and are accomplished by six extraocular muscles that operate in three antagonistic pairs. One pair, the medial and lateral rectus muscles, is located on either side of each eyeball and controls the horizontal orientation of the eye. A second pair, the superior and inferior rectus muscles, controls the vertical orientation of the eye and a third pair, the superior and inferior oblique muscles, controls rotations of the eye around the line-of-sight. Although most people are unaware of them, rotations of the eye about the line-of-sight are quite common, though usually less than 10° in amplitude.

Three Cranial Nerve Nuclei Control Tension in the Extraocular Muscles

The tension generated in each muscle is regulated by the activity of pools of alpha motor neurons located in the oculomotor brain stem. The pools of neurons regulating the superior rectus, inferior rectus, medial rectus, and inferior oblique muscles each lie within separate subdivisions of the oculomotor, or third cranial nerve, nucleus. Neurons innervating the superior oblique muscle lie within the trochlear, or fourth cranial nerve, nucleus. Neurons innervating the lateral rectus muscle lie within the abducens, or sixth cranial nerve, nucleus. Because these six pools are the only source of input to the extraocular muscles, they serve as a final common pathway through which all eye movement control must be implemented. Physiological studies of these neurons made while monkeys produced each of the five types of eye movements described above indicate that all of these motor neurons participate in the generation of all of the five classes of eye movements; there is no specialization among the alpha motor neurons for any particular class of movement. Engineering models of the eye and its muscles developed in the 1970s (largely by David Robinson and colleagues at Johns Hopkins University, USA; Robinson, 1981) indicate that each of these motor neurons must generate two types of muscle force to

accomplish any of the five classes of movements: a pulsatile burst of force produced during the movement, which regulates the instantaneous velocity with which the eye rotates, and a long-lasting increment or decrement in maintained force, which persists after the movement is complete. This maintained step in force holds the eye stationary by resisting the elastic forces of the muscles and connective tissues surrounding the eye. Forces that act to passively pull the eye back toward a straight ahead position. Physiological experiments have demonstrated that each motor neuron produces both of these two types of force. In summary then, the control circuits associated with gaze-stabilizing and gaze-shifting movements all must activate the same set of alpha motor neurons and all of these circuits must produce signals, which regulate both the velocity and position of the eyes in a manner appropriate for that type of eye movement.

Gaze-Stabilization Movements

Gaze-stabilization movements achieve image stabilization when the head rotates in space by rotating the eyes at a velocity equal to that of the head but, in an exactly opposite direction. Gaze stabilization thus requires very precise information about the speed and direction of head rotation across a broad range of head speeds. At high speeds of head rotation the vestibulo-ocular system obtains information from the semicircular canals of the inner ear, which accurately encodes the velocity of head rotation. However, as the velocity of head rotation decreases, the vestibular system begins to underestimate velocity. At low speeds, the optokinetic system obtains information about the rate at which the whole visual field slides across the retina, information which accurately encodes the velocity of head rotation. In this system, as the velocity of head rotation increases head velocity is underestimated. Thus either system working alone could not fully compensate for all movements of the head and body, but working together these two systems can provide an accurate estimate of head velocity that can be used by the oculomotor control circuitry to counterrotate the eyes in the head across a broad range of head speeds. Of course, regardless of the speed of eye rotation, as the eyes move toward far eccentric positions in the orbit, it eventually becomes impossible for the eyes to move any farther. The gaze-stabilization systems respond to this challenge by very quickly moving the eyes in the opposite direction; this gives rise to a repetitive pattern of quick reset movements and slow stabilization phases usually called nystagmus.

Vestibulo-Ocular Response

In the vestibulo-ocular system (often called the vestibulo-ocular reflex (VOR)) the rotational velocity of the head is transduced directly by the semicircular canals of the inner ear (even in total darkness) and enters the brain via the vestibular nucleus. Robinson and colleagues recognized that because this signal already encodes the velocity of head rotation, in order to compute the desired direction and velocity for a gaze-stabilization movement it would only be necessary for the brain to take the inverse of this signal. Recalling that all oculomotor alpha motor neurons produce both a pulsatile burst of force that regulates the velocity of the eye movement and a long-lasting steplike increment or decrement in maintained force that holds the eye stationary, Robinson proposed that the vestibular signal could be channeled directly to the alpha motor neurons to produce the pulsatile burst of force. He then went on to suggest that changes to the long-lasting step in force required after each eye rotation could be computed from this pulse, or velocity, signal by the mathematical operation of integration. In the 1980s the lesion of a discrete brain area, the nucleus prepositus hypoglossi, was shown to eliminate only the long-lasting force change from the motor neurons without affecting eye velocity during movements, suggesting that the nucleus prepositus hypoglossi is a neural integrator employed by the oculomotor system (Leigh and Zee, 1991).

Optokinetic Response

In the optokinetic response (often called optokinetic nystagmus (OKN)) head velocity is computed based on the global motion of the retinal image. This computation, determining the speed and direction of image motion, is performed in the pretectum and yields a signal that also specifies the rotational velocity of the eyes necessary to achieve image stabilization. This visually derived signal passes to the vestibular nucleus where it has been shown to combine with the vestibular signal used by the VOR to yield a neural estimate of head velocity that is accurate across the entire range of speeds for which gaze stabilization can be accomplished. This combined signal is then processed by the same circuitry that Robinson and his colleagues described for the vestibulo-ocular system. Gaze stabilization thus appears to involve two sensory systems that are combined and passed to a single control circuit that governs eye rotation.

Gaze-Shifting Movements

In some sense, gaze-shifting movements oppose gaze stabilization. While the VOR and OKN stabilize images on the retina to maximize visual acuity, gaze-shifting movements redirect the line-of-sight, temporarily reducing the acuity with which some or the entire visual world is perceived.

Saccades

The saccadic system, in order to achieve a precise gaze shift, must supply the alpha motor neurons with a command that controls the amplitude and the direction of the desired change in the line-of-sight. Current evidence indicates that this command can originate in either of two brain structures: the superior colliculus of the midbrain or the frontal eye fields of the neocortex (Sparks, 1986). Both of these structures contain laminar sheets of neurons that code all possible saccadic amplitudes and directions in a topographic maplike organization and both project to brain stem regions that supply control signals to the oculomotor neural integrators and the motor neurons described above. Activation of neurons in these collicular and cortical maps has been shown to precede the onset of saccadic eye movements by about 30 ms. One group of theories proposes that these cortical and collicular signals

govern a brain stem feedback loop (incorporating the neural integrator), which accelerates the eye to a high velocity and keeps the eye in motion until the desired eye movement is complete while other theories make similar predictions but place this feedback loop outside the brain stem. In either case, it seems clear that the superior colliculus and frontal eye fields are important sources for these signals because if both of these structures are removed, no further saccades are possible. The superior colliculus and frontal eye fields, in turn, receive input from many areas within the visual and parietal cortices, the basal ganglia, and brain structures involved in audition and somatosensation. These areas are presumed to participate in the processes that must precede the decision to make a saccade, processes like attention and target selection.

Smooth Pursuit

In the smooth pursuit system, signals carrying information about target motion are extracted by motion processing areas in the visual cortex and then passed to the dorso-lateral pontine nucleus of the brain stem. There, neurons have been identified, which code either the direction and velocity of pursuit eye movements, the direction and velocity of visual target motion, or both. These signals are passed to the cerebellum where neurons have been shown to encode specifically the velocity of pursuit eye movements. These neurons, in turn, make connections with cells known to be upstream of the nucleus prepositus hypoglossi (the oculomotor integrator described above). As in the other oculomotor systems, the brain stem integrator appears to compute the long-term holding force from this signal and then to pass the sum of these signals to the motor neurons.

Vergence

A number of different cues are used by primates to determine the distance to a visual stimulus and from this information to compute the convergence angle necessary to fixate a target. The relative positions of a stimulus falling on the two retinas (the binocular disparity of a stimulus) can be used to compute how much additional convergence or divergence is required to cause a stimulus to fall on the same position in each retina. The state of the lens in the eye can also be used to determine the distance to objects both in and out of focus and thus to compute the required vergence angle. For that matter, even perspective cues can be used to compute a desired vergence angle. In any case, all of these cues are used to activate brain stem circuitry, which appears to be very similar to the brain stem circuitry employed by the saccadic system. This brain stem circuit appears to generate a vergence velocity command similar to the saccadic velocity command, which is in turn integrated and passed on to the alpha motor neurons. Of course, unlike the commands for versional movements, in this case each eye receives a very different command for both velocity and holding force.

Summary

While many important questions remain the focus of active research, the oculomotor system is a comparatively well-understood set of neural mechanisms that serves as an excellent model for understanding movement control in general.

See also: Cortical Areas Engaged in Movement: Neuroimaging Methods; Recovery of Function: Dependency on Age; Vestibulo-Ocular Reflex, Adaptation of the.

Bibliography

Berthoz, A., Melvill Jones, G. (Eds.), 1985. Mechanisms in Gaze Control: Facts and Theories. Reviews of Oculomotor Research, vol. 1. Elsevier, New York.

Buttner-Ennever, J.A. (Ed.), 1988. Neuroanatomy of the Oculomotor System. Reviews of Oculomotor Research, vol. 2. Elsevier, New York.

Carpenter, R.H.S., 1988. Movements of the Eyes, second ed. Pion, London.

Carpenter, R.H.S. (Ed.), 1991. Eye Movements. Vision and Visual Dysfunction, vol. 8. CRC Press, Boston, MA.

Fuchs, A.F., Kaneko, C.R.S., Scudder, C.A., 1985. Brainstem control of saccadic eye movements. Annual Review of Neuroscience 8, 307–337.

Glimcher, P.W., 1998. Eye movements. In: Zigmond, M.J., et al. (Eds.), Fundamental Neuroscience. Academic Press, New York, pp. 993–1010.

Horn, A.K., Leigh, R.J., 2011. The anatomy and physiology of the ocular motor system. Handbook of Clinical Neurology 102, 21–69.

King, W.M., 2011. Binocular coordination of eye movements: Hering's law of equal innervation or uniocular control? European Journal of Neuroscience 33, 2139–2146.

Kowler, E. (Ed.), 1990. Eye Movements and Their Role in Visual and Cognitive Processes. Reviews of Oculomotor Research, vol. 4. Elsevier, New York.

Leigh, R.J., Zee, D.S., 1991. The Neurology of Eye Movements, second ed. F A Davis, Philadelphia, PA.

Lisberger, S.G., Morris, E.J., Tychsen, L., 1987. Visual motion processing and sensory-motor integration for smooth pursuit eye movements. Annual Review of Neuroscience 10, 97–129.

Miles, F.A., Wallman, J. (Eds.), 1993. Visual Motion and its Role in the Stabilization of Gaze. Reviews of Oculomotor Research, vol. 5. Elsevier, New York.

Raphan, T., Cohen, B., 1978. Brainstem mechanisms for rapid and slow eye movements. Annual Review of Physiology 40, 527–552.

Robinson, D.A., 1981. Control of eye movements. In: Brooks, V.B. (Ed.), The Nervous System, Handbook of Physiology, vol. 2, Part 2. Williams and Wilkins, Baltimore, MD, pp. 1275–1320.

Sparks, D.L., 1986. Translation of sensory signals into commands for saccadic eye movements: role of primate superior colliculus. Physiological Reviews 66, 118–171.

Wurtz, R.H., Goldberg, M.E. (Eds.), 1989. The Neurobiology of Saccadic Eye Movements. Reviews of Oculomotor Research, vol. 3. Elsevier, New York.

Eye Movements in Reading

Keith Rayner, University of California, San Diego, La Jolla, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The primary characteristics of eye movements (fixations, saccades, and regressions) during reading are described. Much of the emphasis is on skilled reading, but the eye movement characteristics of beginning readers and dyslexic readers are also discussed. Research findings on the following topics are also discussed: (1) the size of the perceptual span (or area of effective vision) during a fixation in reading, (2) integration of information across eye fixations in reading, (3) eye movement control – where to fixate next, and (4) eye movement control – when to move the eyes.

The primary characteristics of eye movements (fixations, saccades, and regressions) during reading are described. Much of the emphasis is on skilled reading, but the eye movement characteristics of beginning readers and dyslexic readers are also discussed. Research findings on the following topics are also discussed: (1) the size of the perceptual span (or area of effective vision) during a fixation in reading, (2) integration of information across eye fixations in reading, (3) eye movement control – where to fixate next, and (4) eye movement control – when to move the eyes.

Reading is arguably the most important skill that human beings acquire during their school years. An integral part of the reading process, both for beginning readers and for skilled readers, is the eye movements that occur when we comprehend text. In this article, some basic facts about eye movements during reading are discussed. Developmental changes in eye movements are also discussed, as are the following topics: the perceptual span (or span of effective vision), the integration of information across fixations, and the control of eye movements.

Basic Facts about Eye Movements during Reading

Although it seems as though we move our eyes smoothly across the text when we read, this is an illusion since our eyes make a series of movements (called *saccades*) that last about 20–35 ms. Between the saccades, our eyes are relatively still for about 200–250 ms (these periods when the eyes are stationary are called *fixations*). No information is obtained during saccades because the eyes are moving so quickly across the text; all new information is thus encoded during the fixations. Reading is therefore a bit like a slide show where each new slide (with a different view of the text) appears every quarter of a second or so. The saccade is typically 7–9 letter spaces and is not affected by the size of the print as long as it is not too small or too large. The appropriate metric to use when discussing eye movements therefore is letter spaces, and not visual angle (generally, 3–4 letter spaces are equivalent to 1 degree of visual angle). The other primary characteristic of eye movements is that about 10–15% of the time readers move their eyes back in the text (*regressions*) to look at material that has already been read.

Eye movements during reading are necessary because of acuity limitations in the visual system. With respect to the fixation point, any line of text can be divided into three regions: foveal, parafoveal, and peripheral. In the foveal region (extending 1 degree of visual angle to the left and right of fixation), acuity is sharpest and the letters can be easily resolved. In the parafoveal region (extending to 5 degrees of visual angle on either side of fixation) and the peripheral region (everything on the line beyond the parafoveal region), acuity drops off markedly so that our ability to identify letters is not very good even in the near parafovea. The purpose of eye movements in reading therefore is to place the fovea on that part of the text to be processed next. Whereas a majority of the words in a text are fixated during skilled reading, many words are skipped so that foveal processing of each word is not necessary. For example, content words are fixated about 85% of the time while function words are fixated about 35% of the time. Function words are fixated less frequently than content words because they are generally quite short and there is a relationship between word skipping and word length: as length increases, the probability of fixating increases; whereas 2–3 letter words are fixated less than 25% of the time, words that are 8 letters or longer are almost always fixated (and often fixated more than once).

When reading passages of text that are hard or unfamiliar, the eye movements of skilled readers are somewhat different from when they read text that is easy or familiar. As text difficulty increases, there are systematic effects on eye movements: fixation durations increase, saccade lengths decrease, and regression frequency increases. An important fact about eye movements in reading is that the ease or difficulty associated with understanding the text strongly influences the durations of fixations, the length of saccades, and the frequency of regressions.

It is important to note that the values presented above for fixation duration, saccade length, and regression frequency are averages and there is considerable variability in all of the measures. Thus, although the mean fixation duration is 200–250 ms and the mean saccade length is about 8 letter spaces, for individual readers these values might be somewhat higher or lower. Thus, the mean fixation duration for individual readers generally ranges between 200 ms and 300 ms and the average saccade length generally ranges from 6 to 10 letters. This between-reader variability (which also exists for regression frequency) is perhaps not as important as the fact that there is considerable within-reader variability. In other

words, although a reader's mean fixation duration might be 250 ms, the range of individual fixation durations can be from under 100 ms to over 500 ms within a passage of text. Likewise, the variability in saccade length can range from 1 letter (and sometimes even less than a letter) to over 15 letter spaces (though such long saccades typically follow regressions).

While fixation durations are typically examined when global aspects of reading are considered, when more detailed analyses involving individual target words are necessary, measures other than the mean fixation duration are usually examined. If readers always made one and only one fixation on a word then the fixation duration on the word would be the unit of processing. However, as noted above, words are sometimes fixated more than once, and sometimes they are skipped. Thus, using the mean fixation duration to infer the processing time for a word is inadequate because it underestimates the time the eyes are actually on the word; that is, a 200 ms fixation and a 150 ms fixation on the same word would yield a mean fixation duration of 175 ms, but the actual amount of time on the word would be 350 ms. Likewise, using only words that are fixated once (*single fixation duration*) is problematic because many words are skipped. Thus, the two most frequently used measures of processing time for a word are *first fixation duration* (the duration of the first fixation on a word independent of the total number of fixations) and *gaze duration* (the duration of all fixations on a word prior to an eye movement to another word).

Developmental Changes in Eye Movements

When children learn to read, they must learn to systemically move their eyes through text. Children in first grade typically fixate for 300–400 ms and only move their eyes a few characters. Thus, they fixate on virtually every word in the text. However, as reading skill develops, eye movements become more like those of skilled readers. By the time children are in fourth or fifth grade, the basic characteristics of their eye movements are like those of skilled readers. While their fixation durations and saccade lengths are similar to those of skilled readers, they do make more regressions than skilled readers. Children who experience severe difficulties learning to read (*dyslexic* readers) do not follow this systematic progression in the development of eye movement characteristics. Indeed, adult dyslexic readers make longer fixations, shorter saccades, and more regressions than skilled readers.

The Perceptual Span

How much information does a skilled reader acquire on each fixation and what is the size of the perceptual span (or area of effective vision) on each fixation? In order to investigate this question, the *eye-contingent display change* paradigm (McConkie and Rayner, 1975; Rayner, 1975; Rayner and Bertera, 1979) was developed. In this paradigm, a reader's eye movements are monitored (generally every millisecond) by a highly accurate eye-tracking system. The eyetracker is interfaced with a computer, which controls the display monitor from which the reader reads and changes in the text are made contingent on the location of the reader's eyes. Generally, the display changes are made during saccades and the reader is not aware of the changes.

There are three primary types of eye-contingent paradigms: the *moving window*, *foveal mask*, and *boundary* techniques. With the moving window technique, on each fixation a portion of the text around the fixation point is available to the reader. However, outside of this window, the text is replaced by other letters, or by Xs (see Figure 1). When the reader moves his or her eyes, the window moves with the eyes. Thus, wherever the reader looks, there is readable text within the window and altered text outside the window. The rationale with the technique is that when the window is as large as the region from which a reader can normally obtain information, reading will not differ from when there is no window present. The foveal mask technique is very similar to the moving window paradigm except that the text and replaced letters are reversed. Thus, wherever the reader looks, the letters around the fixation are replaced by Xs while outside of the mask area the text remains normal (see Figure 1). Finally, in the boundary technique, an invisible boundary location is specified in the text and when

```
The most important skill that humans    Normal
           +
_____________________________________________

XXX Xost important sXXXX XXXX XXXXXX
           +                            MW
XXX XXXX XXXXXtant skill that XXXXXX
                     +
_____________________________________________

The most XXXXXXXnt skill that humans
           +                            FM
The most imporXXXXXXXill that humans
                +
_____________________________________________

The most important money that humans
    +       +                           B
The most important skill that humans
                     +
_____________________________________________
```

Figure 1 Examples of the moving window (MW), foveal mask (FM), and boundary (B) paradigms. The first line shows a normal line of text with the fixation indicated by the +. The next two lines show an example of two successive fixations with a window of 15 letter spaces (7 to the left and 7 to the right of the fixation). The next two lines show an example of two successive fixations with a foveal mask of 7 letter spaces. The bottom two lines show an example of the boundary paradigm. The first of these lines shows a line of text prior to the display change; the second line shows the same line after the display change. When the readers' eyes cross the invisible boundary location (just after the last letter of the word *important*), the initially displayed word (*money*) is replaced by the target word (*skill*). The change occurs during the saccade and readers do not notice it.

the reader's eyes cross the boundary, an originally displayed word or letter string is replaced by a target word (see Figure 1). The amount of time that the reader looks at the target word is computed as a function of (1) the relationship between the initially displayed stimulus and the target word, and (2) the distance that the reader was from the target word prior to launching a saccade that crossed the boundary location.

For readers of English (and other alphabetic writing systems printed from left to right), the span extends from the beginning of the currently fixated word, or roughly 3–4 letter spaces to the left of fixation, to about 14–15 letter spaces to the right of fixation. The span is thus asymmetric – it extends further to the right of fixation than to the left of fixation. For readers of languages printed from right to left (such as Hebrew), the span is asymmetric but in the opposite direction from English so that it is larger left of fixation than right. While the span is asymmetric, it is the case that no useful information is acquired below the line of text that is currently being read as readers apparently focus their attention only on the line being fixated.

Although the perceptual span extends about 14–15 letter spaces to the right of fixation, the area from which words can be identified on a given fixation (the word identification span) generally does not exceed 7–8 letter spaces to the right of fixation. However, neither the perceptual span nor the word identification span is fixed as both can be modulated by word length. For example, if three short words occur in succession, readers are able to identify all of them. If the upcoming word is highly constrained by the context, readers acquire more information from that word than from unpredictable words, and if the fixated word is difficult to process, readers obtain less information from the upcoming word. Finally, the span is attention based and not due strictly to acuity factors (though they do have an influence) since if the letters get progressively larger away from fixation (to offset acuity limitations), the span is still about 14–15 letters to the right of fixation.

Orthography also influences the size of the perceptual span. Hebrew readers have a smaller span than English readers, and Japanese and Chinese readers have even smaller spans. Hebrew is a more densely packed language than English, and Japanese and Chinese are both more densely packed than Hebrew; densely packed refers to the fact that it takes more letters per sentence in English than Hebrew, for example. For Chinese readers, the span extends about one character to the left of fixation to 2–3 to the right. Finally, reading skill influences the size of the perceptual span. Beginning readers (at the end of second grade) have a smaller span than skilled readers and adult dyslexic readers have smaller spans than skilled readers.

Integration of Information across Eye Fixations

It is clear from research using both the moving window and the boundary technique that there is a *preview* benefit from having the word to the right of fixation available on an eye fixation, and that information obtained about the parafoveal word on fixation n is combined with information on fixation $n + 1$ to speed identification of the word when it is subsequently fixated. What kind of information is integrated across consecutive eye fixations in reading?

Experiments using the boundary paradigm described above have varied the orthographic, phonological, morphological, and semantic similarity between an initially displayed stimulus and a target word in attempts to determine the basis of the preview effect. A number of major findings have emerged from these studies. First, there is facilitation due to orthographic similarity as having a preview of *chest* facilitates the processing of *chart*. However, the facilitation is not strictly due to visual similarity because changing the case of letters from fixation to fixation (so that *ChArT* becomes *cHaRt* on the next) has little effect on reading behavior. Second, the facilitation is in part due to abstract letter codes associated with the first few letters of an unidentified parafoveal word; although there may be some facilitation from other parts of the word to the right of fixation, the bulk of the preview effect is due to the beginning letters. The effect is not simply due to spatial proximity because there is facilitation from the beginning letters of words when readers are asked to read sentences from right to left, but with letters within words printed from left to right. Third, there is facilitation due to phonological similarity, since *beech* facilitates *beach* and *shoot* facilitates *chute*, with less facilitation in the latter than in the former case. Fourth, although morphological factors can influence fixation time on a word, they are not the source of the preview benefit in English. Finally, although somewhat controversial, it has generally been found that there is no facilitation for readers of English due to semantic similarity since *song* as the preview does not facilitate the processing of *tune*. It thus appears that orthographic (abstract letter) and phonological codes are the basis of the preview benefit and used in integrating information across fixations.

Eye Movement Control: Where to Fixate Next

There are two components to the issue of how eye movements are controlled during reading: (1) where to fixate next and (2) when to move the eyes. It appears that there are separate mechanisms involved in these decisions and they will accordingly be discussed separately.

Word length seems to be the primary determinant of where to fixate next in skilled reading. When length information about the upcoming word is not available, readers move their eyes a shorter distance than when such information is available. Also, the length of the word to the right of fixation strongly influences the size of the saccade. Furthermore, there is a *landing position* effect such that readers tend to fixate about halfway between the beginning and the middle of words. However, there is also a *launch site* effect such that where readers land in a word is strongly influenced by where the saccade came from. Thus, whereas the most frequent landing position is near the middle of the word, if the prior saccade was launched some distance (8–10 letters) from the target word then the landing position shifts to the left of center. Likewise, if the prior saccade was launched close (2–3 characters) to the beginning of the target word, the landing position shifts to the right of center. Given this, the optimal strategy would be to fixate near the middle of each successive word. However, since short words can often be identified when they are just to the right of fixation, they are often skipped. Factors such as these result in the landing position distribution being spread somewhat.

Eye Movement Control: When to Move the Eyes

What determines when we move our eyes? The amount of time a reader fixates on a word or segment of text reveals something about the cognitive processes associated with comprehending that word or segment. One reason for this is that information gets into the processing system very early in a fixation (thus leaving a lot of time for processes associated with word recognition and other necessary comprehension processes). Experiments using the foveal mask paradigm in which the onset of the mask is delayed following a saccade have demonstrated that reading proceeds quite normally if the reader has 50–60 ms to process the text prior to the onset of the mask. If the mask occurs earlier, reading is disrupted. Although readers typically acquire the visual information needed for reading during the first 50–60 ms of a fixation, they can extract information at other times during a fixation as needed.

There have been many demonstrations that various lexical, syntactic, and discourse factors influence fixation time on a word. In particular, all of the following variables influence fixation time on a word: (1) word frequency, (2) contextual constraint, (3) semantic relationships between words in a sentence, (4) anaphora and coreference, (5) lexical ambiguity, and (6) syntactic disambiguation. Thus, for example, readers look longer (the first fixation duration and gaze duration are both longer) at words that do not occur as frequently in the language than they look at more frequent words, and they look at words that are highly predictable for less time than they look at words that are not predictable from the preceding context. In essence, the amount of time that readers look at words is strongly influenced by the ease or difficulty associated with understanding that word. Other factors also influence how long readers look at words. When reading a sentence such as "While Mary was mending the sock fell off her lap" (referred to as garden-path sentences), readers fixate for a long time on *fell*. This is because they typically parse *the sock* as a direct object and then do not know what to do with *fell* when they get to it. Thus, they need to reanalyze the sentence. In this particular example, readers would not only fixate for a long time on the disambiguating word (*fell*), they would also make a regression back to the earlier part of the sentence where the misanalysis began.

In essence, the variability in fixation durations that was mentioned at the outset of this article is largely due to the ease or difficulty of processing the fixated words. Because the oculomotor system is apparently closely linked to the part of the brain responsible for processing language, eye movements are a very good way to study various questions related to language and comprehension. Some computational models of eye movements in reading (see Engbert et al., 2005; Reichle et al., 1998, 2009) have been developed that nicely capture many of the properties of skilled reading.

See also: Chinese Literacy Acquisition: A Multidimensional Puzzle; Eye Movement, Control of: Oculomotor Control; Neural Basis of Reading; Reading Nonalphabetic Scripts; Vision, High-Level Theory of; Vision, Low-Level Theory of.

Bibliography

Crowder, R.G., Wagner, R.K., 1992. The Psychology of Reading, second ed. Oxford University Press, New York.

Engbert, R., Nuthmann, A., Richter, E.M., Kliegl, R., 2005. SWIFT: a dynamical model of saccade generation during reading. Psychological Review 112, 777–813.

Frazier, L., Rayner, K., 1982. Making and correcting errors during sentence comprehension: eye movements in the analysis of structurally ambiguous sentences. Cognitive Psychology 14, 178–210.

Just, M.A., Carpenter, P.A., 1980. A theory of reading: from eye fixations to comprehension. Psychological Review 87, 329–354.

Just, M.A., Carpenter, P.A., 1987. The Psychology of Reading and Language Comprehension. Allyn and Bacon, Boston.

McConkie, G.W., Rayner, K., 1975. The span of the effective stimulus during a fixation in reading. Perception and Psychophysics 17, 578–586.

Pollatsek, A., Bolozky, S., Well, A.D., Rayner, K., 1981. Asymmetries in the perceptual span for Israeli readers. Brain and Language 14, 174–180.

Rayner, K., 1975. The perceptual span and peripheral cues in reading. Cognitive Psychology 7, 65–81.

Rayner, K., 1998. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin 124, 372–422.

Rayner, K., 2009. The thirty fifth Sir Frederick Bartlett lecture: eye movements and attention in reading, scene perception, and visual search. Quarterly Journal of Experimental Psychology 62 (8), 1457–1506.

Rayner, K., Bertera, J.H., 1979. Reading without a fovea. Science 206, 468–469.

Rayner, K., Pollatsek, A., 1989. The Psychology of Reading. Prentice-Hall, Englewood Cliffs, NJ.

Rayner, K., Pollatsek, A., Ashby, J., Clifton, C., 2012. The Psychology of Reading. Psychology Press, New York.

Reichle, E.D., Pollatsek, A., Fisher, D.L., Rayner, K., 1998. Toward a model of eye movement control in reading. Psychological Review 105, 125–157.

Reichle, E.D., Warren, T., McConnell, K., 2009. Using E-Z reader to model the effects of higher level language processing on eye movements during reading. Psychonomic Bulletin & Review 16, 1–21.

Eyeblink Classical Conditioning

Kevin L Brown and John H Freeman, The University of Iowa, Iowa City, IA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D. Lavond, volume 8, pp. 5214–5217, © 2001, Elsevier Ltd.

Abstract

Eyeblink conditioning is a classical conditioning procedure that has been used extensively to investigate the neural mechanisms underlying associative learning and memory. The cerebellum ipsilateral to the conditioned eye is essential for eyeblink conditioning. Conditional stimulus (CS) and unconditional stimulus (US) input arising from the pontine nuclei and inferior olive, respectively, converge upon the cerebellar cortex and deep nuclei. Combined input from the CS and US pathways during training trials induces synaptic plasticity in the cerebellar cortex and anterior interpositus nucleus. These two sites of synaptic plasticity constitute the memory necessary for the eyeblink conditioned response. A future direction for this research is to determine the cellular and molecular mechanisms underlying cerebellar learning.

Introduction

The ability to learn relationships between stimuli and events that often occur together is known as associative learning, a fundamental form of learning that is critical for adapting to our environment. Associative learning involving reflexive behaviors is particularly amenable to objective scientific research since reflexes are reliably elicited and can be measured quantitatively more easily than complex behaviors. Initial experimental demonstrations of associative learning of reflexive behaviors are credited to the Russian physiologist Ivan Pavlov, and his work with salivary conditioning in dogs provided the basis for classical conditioning research. In a typical classical (or 'Pavlovian') conditioning procedure, a stimulus that initially does not produce the response being measured (e.g., a ringing bell, termed the conditional stimulus, or CS) precedes and reliably signals the occurrence of a stimulus that elicits the response being measured (e.g., food, termed the unconditional stimulus, or US). With sufficient pairings of the CS and US, the CS comes to elicit conditional responses (CRs; e.g., salivation) that, in some cases, are similar in form to the reflexive unconditional response (UR; e.g., salivation) elicited by the US. Classical conditioning experiments allow for precise experimenter control of stimulus exposure, and the use of proper control procedures readily identifies nonassociative contributions to performance.

Behavioral Studies of Eyeblink Conditioning

Eyeblink conditioning was established as a conditioning procedure for humans in the 1920s (Cason, 1922). In eyeblink conditioning, repeated presentations of a tone or light CS paired with a US that elicits a blink reflex (mild shock or air puff to the eye) yields anticipatory CRs (conditioned eyelid closure) that typically peak (i.e., when the eyelid reaches maximum closure) near the onset time of the US. Though human eyeblink conditioning procedures were refined over subsequent decades (see Woodruff-Pak and Steinmetz, 2000), various shortcomings inherent with eyeblink conditioning in humans including the presence of voluntary blinks and short latency unconditioned startle responses to auditory CSs led to the development of a rabbit model of eyeblink conditioning.

Isidore Gormezano, Alan Wagner, and their colleagues modified the human eyeblink conditioning setup to make it suitable for rabbits. Animal studies are necessary for invasive neuroscience research, and rabbits in particular are well-suited for eyeblink conditioning studies since they show a high tolerance for restraint, display low spontaneous blink rates, and show virtually no alpha responses to CSs typically used in eyeblink conditioning (see Gormezano, 1966). In the initial studies, adult rabbits were restrained within a Plexiglas box and received 70–82 daily presentations of a 600 ms, 72 dB tone CS that preceded a 100 ms puff of air to the cornea (US). The CS and US ended simultaneously (an example of 'delay conditioning'), so the interval between the onset of the CS and the onset of the US within each trial was 500 ms. A potentiometer was used to identify either (1) movement of the external eyelid (Schneiderman et al., 1962), (2) movement of the nictitating membrane (tissue that moves laterally across the eye of rabbits; Gormezano et al., 1962), or eyeball retraction (Deaux and Gormezano, 1963), and responses that exceeded a fixed threshold (typically 1 mm of movement) were recorded. Classically conditioned rabbits in all three experiments showed a gradual increase in CRs over the course of the 8-day training period. Overall, the percentage of CRs in classically conditioned rabbits was minimal early in training (typically less than 10%), but by the end of training CR percentages were 70% or higher, with maximum conditioned eyelid or nictitating membrane closure (CR peaks) occurring at or near the onset of the US. In contrast, control groups receiving the CS only, the US only, or random presentations of the CS and US averaged 10% or fewer CRs over the course of training, thus demonstrating that the associative pairing of the CS and US in the classically conditioned groups was essential for producing robust CRs. Importantly, CS-alone presentations led to a gradual dissipation of CRs in subjects previously given paired CS–US training, a type of inhibitory learning termed 'extinction' by Pavlov. Through extensive subsequent research that characterized properties of the eyeblink response (CR and UR) and examined ranges of parameters (e.g., CS–US intervals, CS and US intensities) capable of supporting conditioning, a comprehensive behavioral analysis of rabbit eyeblink conditioning was provided (see

Gormezano et al., 1983). With appropriate methodological and parametric adjustments the basic properties of eyeblink conditioning in rabbits are observed in many other infrahuman species (e.g., mice and rats), and the underlying neural substrates initially identified in rabbits are highly conserved across species, including rodents and humans.

Identifying Brain Regions Essential for Eyeblink Conditioning

The pioneering behavioral work by Gormezano and others (see Wagner, 1969) using rabbit eyeblink conditioning provided a foundation for investigating the neurobiological mechanisms that underlie associative learning. Richard Thompson and colleagues recognized the advantages afforded by eyeblink conditioning and adopted it as a model system for identifying the most direct links possible between associative learning and memory storage.

Multiple neuroscience techniques and approaches were used by the Thompson laboratory to identify the 'engram' (the neural representation of learning and memory; Thompson, 1976). One way proposed to identify candidate structures was to record neuronal activity in a region of interest during acquisition of eyeblink conditioning; if a particular region is essential one would expect that learning-related changes in neuronal activity would be evident. Because the target response (CR) is a discrete, well-timed eyelid closure that emerges gradually over multiple trials, eyeblink conditioning allows for reliable identification of neuronal activity that corresponds to learning. Indeed, the pattern of neuronal activity in the hippocampus predicts and models the time course of the eyeblink CR in subjects trained in the standard delay eyeblink conditioning procedure but not in subjects receiving unpaired presentations of the CS and US (Berger et al., 1976; Berger and Thompson, 1978). However, removal of the hippocampus does not impair acquisition of delay eyeblink conditioning (Schmaltz and Theios, 1972; Solomon et al., 1983), suggesting that this structure is normally involved but not essential for eyeblink conditioning. In fact, CRs are observed in decerebrate preparations, suggesting that the critical site(s) is/are located in the hindbrain (Mauk and Thompson, 1987).

Converging evidence based on electrophysiological recording, electrical stimulation, lesion, and reversible inactivation indicates that the cerebellum is essential for eyeblink (eyelid or nictitating membrane) conditioning (Christian and Thompson, 2003; Freeman and Steinmetz, 2011). Briefly: (1) neurons in the cerebellum exhibit learning-related activity that precedes the CR, suggesting a causal mechanism (**Figure 1**), (2) electrical stimulation of specific cerebellar regions is sufficient to produce eyelid closure, and (3) cerebellar lesions abolish CRs while leaving the UR intact, demonstrating that the absence of the behavioral expression of learning does not simply reflect an inability to perform the target response (McCormick and Thompson, 1984). Recording, stimulation, and lesions of a relatively small region of the cerebellum – the anterior interpositus nucleus ipsilateral to the trained eye – are sufficient to produce these results. Perhaps the most compelling evidence suggesting that the memory (engram) is formed and stored in the cerebellum comes from reversible inactivation studies. Rabbits receiving infusions of muscimol – a gamma-aminobutyric acid (GABA) agonist that temporarily hyperpolarizes neurons – into the cerebellum prior to daily eyeblink conditioning sessions failed to show CRs throughout training (URs were unaffected; Krupa et al., 1993). Subsequent training without infusions revealed a gradual emergence of CRs that was identical to initial acquisition in subjects receiving infusions of saline (vehicle), indicating there was no carry-over

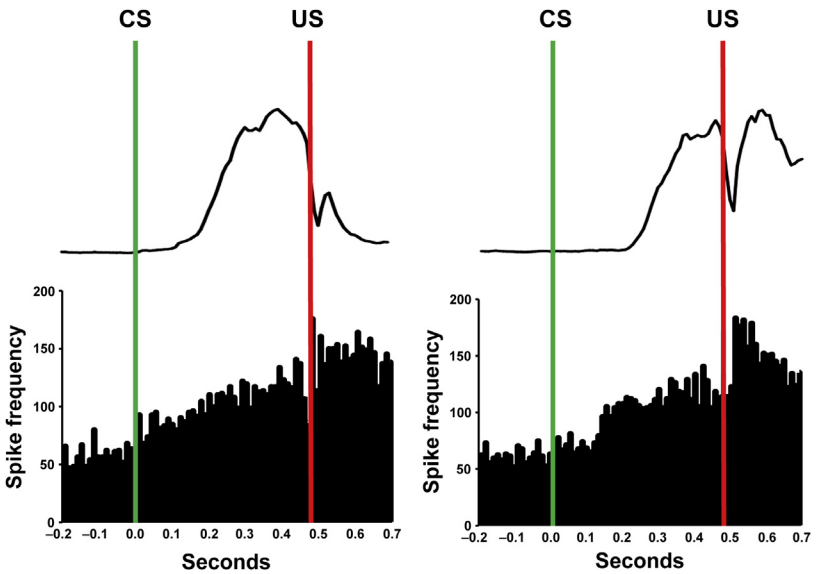

Figure 1 Examples of electrophysiological activity (spikes/s) recorded from the anterior interpositus nucleus and corresponding eyelid movement in well-trained rats during paired eyeblink conditioning training. The green line represents conditional stimulus (CS) onset and the red line represents unconditional stimulus (US) onset. Note the increase in activity after onset of the CS. Figure adapted from Freeman, J.H., Steinmetz, A.B., 2011. Neural circuitry and plasticity mechanisms underlying delay eyeblink conditioning. Learning and Memory 18, 666–677.

or 'savings' from prior CS–US training. In contrast, muscimol infusions into sites efferent to the cerebellum that are involved (directly or indirectly) in control of blink-related movement (contralateral red nucleus; cranial motor neurons) blocked CR expression but had no lingering effects on postinfusion CRs. That is, CRs in these subjects were at asymptotic levels at the start of subsequent training with no infusions, indicating that substantial learning had occurred during inactivation sessions (Krupa et al., 1993, 1996). Similar findings were reported using different reversible inactivation techniques (Chapman et al., 1990; Clark et al., 1992; Clark and Lavond, 1993). Additionally, reversible inactivation of the cerebellum shortly *after* daily eyeblink conditioning sessions blocks retention of CRs, presumably resulting from disruption of memory consolidation (Attwell et al., 2002; Cooke et al., 2004). These findings indicate that the memory underlying eyeblink conditioning is formed and stored in the cerebellum, and not in downstream targets of the cerebellum.

Mechanisms of Cerebellar Learning in Eyeblink Conditioning

Memory formation and storage results from experience-dependent changes ('plasticity') occurring within the cerebellar cortex and anterior interpositus nucleus. Genetic manipulations provide an opportunity to explore how noninvasive disruption of potentially important mechanisms affects eyeblink conditioning. Genetically modified mice lacking specific receptor subtypes (Aiba et al., 1994; Kishimoto et al., 2002) or glial cell proteins (Shibuki et al., 1996) highly expressed in the cerebellar cortex are impaired in delay eyeblink conditioning and show impaired long-term depression (LTD) in the cerebellar cortex. LTD is demonstrated using *in vitro* electrophysiology and refers to persistent decreases in synaptic strength at parallel fiber-Purkinje cell synapses (as evidenced by depressed Purkinje cell activity) brought about by repeated, simultaneous stimulation of two distinct cerebellar afferents (Ito et al., 1982) shown to transmit CS and US information necessary for eyeblink conditioning (see Section Conditioned and Unconditioned Stimulus Input Circuitry, below). Purkinje cell axons form the entire output of the cerebellar cortex, projecting primarily to cerebellar deep nuclei. Furthermore, Purkinje cells produce action potentials at high frequencies under baseline conditions and form inhibitory synapses on their targets. Therefore, a depression of Purkinje cell firing (as in LTD) results in a 'release' of deep nuclear (e.g., anterior interpositus) neurons from chronic Purkinje cell inhibition. A model of cerebellar associative learning derived from a variety of techniques, including reversible inactivation and computer simulations proposes that LTD at parallel fiber-Purkinje cell synapses during the latter part of the CS in well-trained subjects produces a transient pause in Purkinje cell firing that 'disinhibits' cerebellar deep nuclei (Medina et al., 2000). According to this model, the learning-related cerebellar cortical plasticity permits additional plasticity – long-term potentiation (LTP) in cerebellar deep nuclei (Racine et al., 1986) – resulting in robust activity in deep nuclear neurons, thus producing the target response (e.g., eyeblink CR).

Consistent with the aforementioned model, electrophysiological recordings of Purkinje cell activity during eyeblink conditioning in well-trained subjects show a pattern of decreased firing during the latter portion of the CS (Hesslow and Ivarsson, 1994; Green and Steinmetz, 2005; Jirenhed et al., 2007), a period that coincides with the CR. This pattern of Purkinje cell activity is not observed early in training (Jirenhed et al., 2007), suggesting that the decreased activity during the CS in well-trained subjects represents LTD-like learning-related plasticity. Plasticity in the cerebellar cortex is essential for appropriate timing of the CR, and blocking Purkinje cell output in well-trained subjects leads to early onset, poorly timed CRs (Garcia and Mauk, 1998). Ohyama et al. (2006) used pharmacological manipulations as well as electrical stimulation of cerebellar afferents to demonstrate that these poorly timed CRs reflect plasticity in cerebellar deep nuclei and not in downstream targets (i.e., red nucleus). CRs were associative in nature and did not generalize to auditory cues (e.g., white noise) that had no prior history with the US. The associative and input-specific nature of these poorly timed CRs is consistent with an LTP-like mechanism in cerebellar deep nuclei underlying eyeblink conditioning. Additional indirect evidence for LTD-like mechanisms in the cerebellar cortex and LTP-like mechanisms in the cerebellar deep nuclei underlying eyeblink conditioning comes from electron microscopy studies. Specifically, reductions in excitatory parallel fiber-Purkine cell synapses (Connor et al., 2009) and increases in excitatory synapses per neuron in anterior interpositus nuclei (Kleim et al., 2002) have been reported in subjects receiving paired CS–US training relative to controls.

Although there is compelling evidence for LTD-like and LTP-like mechanisms in the cerebellum underlying eyeblink conditioning, alternatives must be considered. For one, cerebellar LTD and LTP have been demonstrated *in vitro* but have not been verified *in vivo* during eyeblink conditioning. Furthermore, some mutant mice that show impaired cerebellar cortical LTD are unimpaired in delay eyeblink conditioning (Schonewille et al., 2011). Additionally, delay eyeblink conditioning can proceed – albeit slowly – following global loss of Purkinje cells (Chen et al., 1996; Nolan and Freeman, 2006), indicating that cerebellar cortical LTD-like mechanisms may be normally involved but are not essential for eyeblink conditioning. Multiple plasticity mechanisms – in addition to plasticity at the level of the synapse – at different points throughout the essential circuitry are likely involved in the formation and storage of memory (see Hansel et al., 2001). It is important to note that these proposed mechanisms do not apply solely to eyeblink conditioning. The neural substrates and mechanisms underlying eyeblink conditioning are believed to underlie many other forms of cerebellar-dependent learning as well.

Conditioned and Unconditioned Stimulus Input Circuitry

Neural signals transmitting CS and US information converge upon the cerebellum to establish plasticity essential for eyeblink conditioning. CS information is conveyed to the

cerebellum from the pontine nuclei contralateral to the conditioned eye via the middle cerebellar peduncle. Mossy fibers from the pontine nuclei synapse with (1) granule cells in the cerebellar cortex (granule cell axons – 'parallel fibers' – make synaptic contacts with many Purkinje cells) and (2) cerebellar deep nuclear neurons (Shinoda et al., 1992; Mihailoff, 1993). Lesions of the middle cerebellar peduncle abolish learning with auditory, visual, or tactile CSs (Lewis et al., 1987), and electrical stimulation of the pontine nuclei or the middle cerebellar peduncle serves as highly effective CSs when forward paired with a peripheral US (Steinmetz et al., 1986). Electrical stimulation of lateral pontine nuclei produces robust CRs that are timed similar to CRs using peripheral CSs (Steinmetz, 1990), CRs gradually diminish in subjects trained with electrical stimulation as a CS following removal of the peripheral US (extinction), and lesions of the interpositus nucleus abolish CRs using stimulation of pontine nuclei or middle cerebellar peduncle as a CS (Steinmetz et al., 1987). These findings provide evidence that electrical stimulation of mossy fibers mimics the effects of peripheral CSs typically used for eyeblink conditioning. One concern with electrical stimulation is that it can produce axon potentials traveling back toward, rather than away from, the cell body (antidromic activation). However, eyeblink conditioning proceeds using middle cerebellar peduncle stimulation as a CS while upstream portions of the middle cerebellar peduncle (closer to the pontine cell bodies) are simultaneously inactivated (Hesslow et al., 1999).

CS information from different sensory modalities is transmitted to pontine nuclei through different pathways. Since most eyeblink conditioning studies use auditory CSs, the auditory CS pathway to the pontine nuclei will be the focus here (for a review of visual and somatosensory CS pathways, see Freeman and Steinmetz, 2011). Steinmetz et al. (1987) identified a projection from the cochlear nuclei to the lateral pontine nuclei and found that lesions of the lateral pontine nuclei prevented CRs when a tone, but not a light, was used as a CS. Subsequent studies have shown that the inferior colliculus and the medial auditory thalamus also provide auditory CS input that is necessary for eyeblink conditioning. The inferior colliculus provides auditory input to the medial auditory thalamus (Campolattaro et al., 2007), and lesions of the inferior colliculus impair eyeblink conditioning (Freeman et al., 2007). Lesions of the medial auditory thalamus prevent CRs to a tone CS but leave conditioning using a light CS unaffected (Halverson and Freeman, 2006). Furthermore, electrical stimulation of the medial auditory thalamus as a CS produces robust conditioning when forward paired (but not when unpaired) with a peripheral US (Campolattaro et al., 2007). Medial auditory thalamic nuclei send an ipsilateral projection to the pontine nuclei (Campolattaro et al., 2007), and this projection appears to be a necessary part of the auditory CS pathway since CRs produced from medial auditory thalamic stimulation are severely impaired by reversible inactivation of the lateral pontine nuclei (Halverson and Freeman, 2010). In summary, the auditory CS pathway includes separate projections from auditory nuclei (e.g., cochlear nuclei and the inferior colliculus) to the medial auditory thalamus contralateral to the conditioned eye, a direct projection from the medial auditory thalamus to the lateral pontine nuclei, and mossy fiber projections emerging from pontine nuclei to the cerebellar cortex and cerebellar deep nuclei (**Figure 2**).

Approaches similar to those used to identify the CS pathway have been used successfully in identifying the US pathway. The inferior olive receives inputs from sensory areas transmitting information from the cornea and face (e.g., trigeminal nerve) and US information is sent from the inferior olive to Purkinje cells in the cerebellar cortex (as climbing fibers) and to cerebellar deep nuclear cells (Shinoda et al., 2000; Sugihara et al., 2001). Electrical stimulation of the dorsal accessory inferior olive serves as an effective US when paired with a peripheral CS under appropriate parametric conditions and CR acquisition rates generated from this protocol are similar to acquisition rates in subjects receiving a peripheral US (Mauk et al., 1986). Destruction of the cerebellar deep nuclei prevents CRs generated from CS-inferior olive stimulation, and unpaired presentations of the CS and inferior olive stimulation following acquisition produces extinction (Mauk et al., 1986). Additionally, lesions of the inferior olive in well-trained subjects lead to an elimination of CRs that is either (1) gradual, similar to the pattern of CR decrement observed with removal of a peripheral US (McCormick et al., 1985), or (2) immediate (Yeo et al., 1986), closer to effects seen with lesions or inactivation of the interpositus nucleus. Though the basis for the differences in the time course of postlesion CR loss is unclear, these findings provide evidence that the inferior olive transmits US information to the cerebellum essential for eyeblink conditioning.

Cerebellar Feedback to the CS and US Pathways

The CS and US pathways receive feedback projections from the cerebellum. An inhibitory projection from the cerebellar deep nuclei to the inferior olive has been identified (Bengsston and Hesslow, 2006), and neuronal activity in the dorsal accessory inferior olive decreases during the US period over the course of acquisition of eyeblink conditioning (Sears and Steinmetz, 1991). This decreased activity in well-trained subjects is

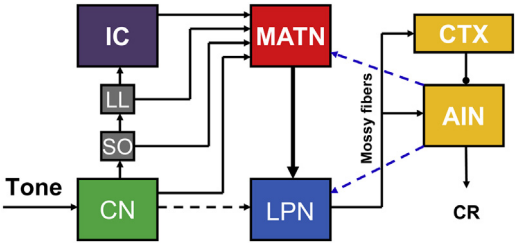

Figure 2 Hypothesized auditory conditional stimulus pathway for eyeblink conditioning. There are inputs into the medial auditory thalamic nuclei (MATN) from the cochlear nucleus (CN), inferior colliculus (IC), and other auditory brainstem nuclei (superior olive – SO – and nucleus of the lateral lemniscus – LL). Input is then projected to the lateral pontine nuclei (LPN), which projects to the cerebellar cortex (CTX) and anterior interpositus nucleus (AIN). Hypothesized feedback projections from the AIN to the LPN and MATN are also shown (blue dashed lines). Figure adapted from Freeman, J.H., Steinmetz, A.B., 2011. Neural circuitry and plasticity mechanisms underlying delay eyeblink conditioning. Learning and Memory 18, 666–677.

evident in CS–US, but not in US-alone trials, suggesting that output from the deep nuclei related to CR generation is the primary cause. Inhibitory feedback from the deep nuclei to the inferior olive prevents learning about redundant or otherwise irrelevant stimuli (see Kim et al. (1998) for a role in the 'blocking' effect). A model of cerebellar function proposes that inhibitory, CR-related feedback from cerebellar deep nuclei is critical in maintaining learning-related plasticity by helping to provide an 'equilibrium' level of activity in the cerebellar cortex (Mauk and Donegan, 1997). Consistent with predictions from this model, blocking the influence of inhibitory feedback by infusing a GABA antagonist (picrotoxin) into the inferior olive prevents extinction during CS-alone presentations in well-trained subjects (Medina et al., 2002).

Cerebellar feedback identified upon the CS pathway, in contrast, is excitatory and may serve to strengthen CS input to the cerebellum. Projections to lateral pontine nuclei from the cerebellar deep nuclei have been identified (Steinmetz and Sengelaub, 1992; see **Figure 2**), and learning-related neuronal activity evident in the pontine nuclei in well-trained subjects is eliminated following inactivation of cerebellar deep nuclei (Clark et al., 1997; Bao et al., 2000). This pontine neuronal activity is not abolished by inactivation of the motor nuclei that control the blink response (Clark et al., 1997), suggesting that the pattern of pontine neuronal activity in well-trained subjects does not simply reflect performance of the CR. Halverson et al. (2010) identified projections from cerebellar deep nuclei to the medial auditory thalamus and recorded neuronal activity in the anterior interpositus nucleus and medial auditory thalamus simultaneously during acquisition of eyeblink conditioning. Robust, learning-related increases in neuronal activity emerged in the interpositus nucleus earlier than in the medial auditory thalamus across training sessions, and within trials, learning-related increases in interpositus activity preceded the CR while learning-related medial auditory thalamic activity typically emerged after CR onset. These findings suggest that cerebellar feedback may be driving learning-related activity in the medial auditory thalamus (**Figure 3**).

Conclusions and Future Directions

Eyeblink conditioning is perhaps the most suitable model system for identifying neural mechanisms underlying associative learning and memory. The discrete, robust target response is easily established in a number of species and across the lifespan. Although the essential neural circuitry is well-defined (see **Figure 4**), there are many issues that require further investigation. For one, it is still unknown how cerebellar feedback on CS and US pathways influences acquisition and retention of eyeblink conditioning. The availability of increasingly sophisticated neuroscience techniques (e.g., optogenetics) offers unique opportunities to explore not only dynamic interactions between the cerebellum and its inputs, but also plasticity mechanisms within the cerebellar cortex and deep nuclei. Plasticity mechanisms have been hypothesized, but the cellular and molecular mechanisms underlying learning-related plasticity throughout the essential cerebellum–brainstem circuitry require further examination. Related to these potential mechanisms is the relatively unknown role of cerebellar cortical interneurons. Recent characterization of *in vivo* firing profiles of identified

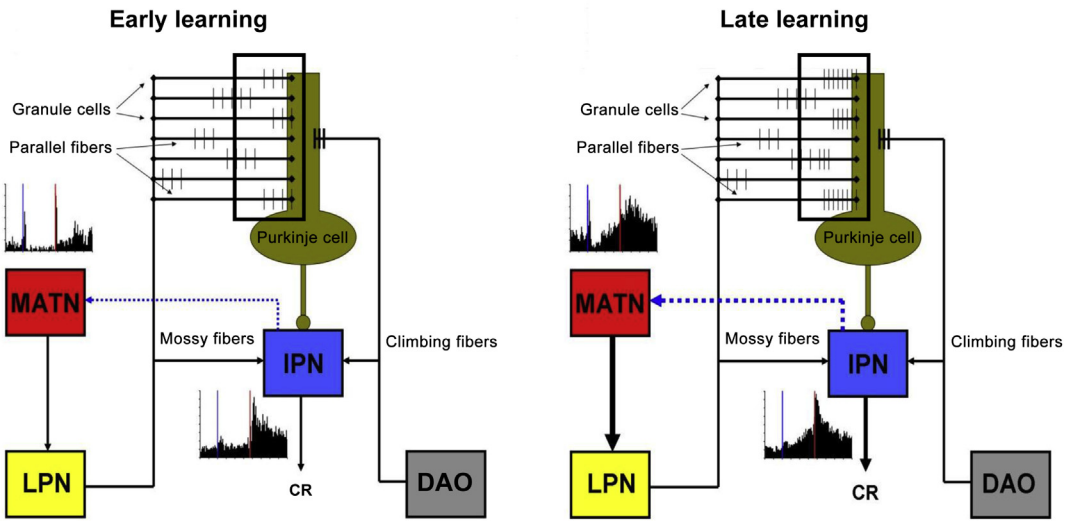

Figure 3 Hypothesized role of medial auditory thalamic plasticity in eyeblink conditioning. Learning is initiated by conditional stimulus-activated parallel fibers that arrive at Purkinje cells nearly simultaneously with climbing fiber input (window). No learning-related plasticity is evident in the interpositus nucleus (IPN) or medial auditory thalamic nuclei (MATN) during the initial training trials (*left*). Learning is potentiated by an increase (relative to initial training) in parallel fiber activity that occurs nearly simultaneously with climbing fiber input (window) and mossy fiber activity in the IPN. Learning-related plasticity is evident in IPN and MATN during later stages of learning (*right*). An increase in learning-related excitatory cerebellar feedback to the MATN (blue dotted line) may, in turn, increase the MATN output to the LPN and subsequently to the cerebellum to further facilitate learning. Figure adapted from Freeman, J.H., Steinmetz, A.B., 2011. Neural circuitry and plasticity mechanisms underlying delay eyeblink conditioning. Learning and Memory 18, 666–677.

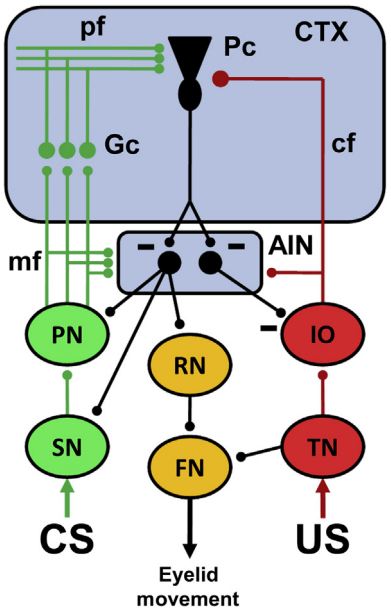

Figure 4 A simplified diagram of the neural circuitry underlying eyeblink conditioning. The cerebellar anterior interpositus nucleus (AIN) and Purkinje cells (Pc) in the cerebellar cortex receive convergent input from the conditioned stimulus (CS, green) and unconditioned stimulus (US, red) neural pathways. The CS pathway includes subcortical sensory nuclei (SN) that include the medial auditory thalamus for auditory conditioning, pontine nuclei (PN), mossy fiber (mf) projection to the AIN and cortical granule cells (Gc), and the parallel fiber (pf) projections to Purkinje cells. The US pathway includes the trigeminal nucleus (TN), dorsal accessory division of the inferior olive (IO), and the climbing fiber (cf) projection to the AIN and Pc. The output pathway for performance of the conditioned blink response (orange) includes the AIN projection to the red nucleus (RN) and its projection to facial motor nucleus (FN) which causes eyelid closure. Feedback projections from the AIN to the PN, SN, and IO regulate CS and US input to facilitate acquisition and maintain plasticity within the cerebellum. Inhibitory synapses are depicted by a minus sign. All other synapses are excitatory. Figure adapted from Freeman, J.H., Steinmetz, A.B., 2011. Neural circuitry and plasticity mechanisms underlying delay eyeblink conditioning. Learning and Memory 18, 666–677.

interneurons (Ruigrok et al., 2011) may lead to a greater understanding of how these cortical cells influence eyeblink conditioning. The study of eyeblink conditioning has made tremendous contributions to our understanding of the neural bases of learning and memory and promises to remain an exciting area of investigation for years to come.

See also: Cerebellum: Cognitive Functions; Learning and Memory, Neural Basis of; Long-Term Potentiation and Long-Term Depression; Neural Plasticity; Neurons and Dendrites, Integration of Information in.

Bibliography

Aiba, A., Kano, M., Chen, C., Stanton, M.E., Fox, G.D., Herrup, K., Zwingman, T.A., Tonegawa, S., 1994. Deficient cerebellar long-term depression and impaired motor learning in mGluR1 mutant mice. Cell 79, 377–388.

Attwell, P.J., Cooke, S.F., Yeo, C.H., 2002. Cerebellar function in consolidation of a motor memory. Neuron 34, 1011–1020.

Bao, S., Chen, L., Thompson, R.F., 2000. Learning- and cerebellum-dependent neuronal activity in the lateral pontine nucleus. Behavioral Neuroscience 114, 254–261.

Bengsston, F., Hesslow, G., 2006. Cerebellar control of the inferior olive. Cerebellum 5, 7–14.

Berger, T.W., Alger, B., Thompson, R.F., 1976. Neuronal substrates of classical conditioning in the hippocampus. Science 192, 483–485.

Berger, T.W., Thompson, R.F., 1978. Neuronal plasticity in the limbic system during classical conditioning of the rabbit nictitating membrane response. I. The hippocampus. Brain Research 145, 323–346.

Campolattaro, M.M., Halverson, H.E., Freeman, J.H., 2007. Medial auditory thalamic stimulation as a conditioned stimulus for eyeblink conditioning in rats. Learning and Memory 14, 152–159.

Cason, H., 1922. The conditioned eyelid reaction. Journal of Experimental Psychology 5, 153–196.

Chapman, P.F., Steinmetz, J.E., Sears, L.L., Thompson, R.F., 1990. Effects of lidocaine injection in the interpositus nucleus and red nucleus on conditioned behavioral and neuronal responses. Brain Research 537, 149–156.

Chen, L., Bao, S., Lockard, J.M., Kim, J.J., Thompson, R.F., 1996. Impaired classical eyeblink conditioning in cerebellar-lesioned and Purkinje cell degeneration (pcd) mutant mice. Journal of Neuroscience 16, 2829–2838.

Christian, K.M., Thompson, R.F., 2003. Neural substrates of eyeblink conditioning: acquisition and retention. Learning and Memory 11, 427–455.

Clark, R.E., Gohl, E.B., Lavond, D.G., 1997. The learning-related activity that develops in the pontine nuclei during classical eyeblink conditioning is dependent on the interpositus nucleus. Learning and Memory 3, 532–544.

Clark, R.E., Lavond, D.G., 1993. Reversible lesions of the red nucleus during acquisition and retention of a classically conditioned behavior in rabbits. Behavioral Neuroscience 107, 264–270.

Clark, R.E., Zhang, A.A., Lavond, D.G., 1992. Reversible lesions of the cerebellar interpositus nucleus during acquisition and retention of a classically conditioned behavior. Behavioral Neuroscience 106, 879–888.

Connor, S., Bloomfield, J., LeBoutillier, J.C., Thompson, R.F., Petit, T.L., Weeks, A.C., 2009. Eyeblink conditioning leads to fewer synapses in the rabbit cerebellar cortex. Behavioral Neuroscience 123, 856–862.

Cooke, S.F., Attwell, P.J., Yeo, C.H., 2004. Temporal properties of cerebellar-dependent memory consolidation. Journal of Neuroscience 24, 2934–2941.

Deaux, E., Gormezano, I., 1963. Eyeball retraction: classical conditioning and extinction in the albino rabbit. Science 141, 630–631.

Freeman, J.H., Halverson, H.E., Hubbard, E.M., 2007. Inferior colliculus lesions impair eyeblink conditioning in rats. Learning and Memory 14, 842–846.

Freeman, J.H., Steinmetz, A.B., 2011. Neural circuitry and plasticity mechanisms underlying delay eyeblink conditioning. Learning and Memory 18, 666–677.

Garcia, K.S., Mauk, M.D., 1998. Pharmacological analysis of cerebellar contributions to the timing and expression of conditioned eyelid responses. Neuropharmacology 37, 471–480.

Gormezano, I., 1966. Classical conditioning. In: Sidowski, J.B. (Ed.), Experimental Methods and Instrumentation in Psychology, vol. 1. McGraw-Hill, New York, pp. 385–420.

Gormezano, I., Kehoe, E.J., Marshall, B.S., 1983. Twenty years of classical conditioning with the rabbit. Progress in Psychobiology and Physiological Psychology 10, 197–275.

Gormezano, I., Schneiderman, N., Deaux, E., Fuentes, I., 1962. Nictitating membrane: classical conditioning and extinction in the albino rabbit. Science 138, 33–34.

Green, J.T., Steinmetz, J.E., 2005. Purkinje cell activity in the cerebellar anterior lobe after rabbit eyeblink conditioning. Learning and Memory 12, 260–269.

Halverson, H.E., Freeman, J.H., 2006. Medial auditory thalamic nuclei are necessary for eyeblink conditioning. Behavioral Neuroscience 120, 880–887.

Halverson, H.E., Freeman, J.H., 2010. Medial auditory thalamic input to the lateral pontine nuclei is necessary for auditory eyeblink conditioning. Neurobiology of Learning and Memory 93, 92–98.

Halverson, H.E., Lee, I., Freeman, J.H., 2010. Associative plasticity in the medial auditory thalamus and cerebellar interpositus nucleus during eyeblink conditioning. Journal of Neuroscience 30, 8787–8796.

Hansel, C.H., Linden, D.J., D'Angelo, E., 2001. Beyond parallel fiber LTD: the diversity of synaptic and non-synaptic plasticity in the cerebellum. Nature Neuroscience 4, 467–475.

Hesslow, G., Ivarsson, M., 1994. Suppression of cerebellar Purkinje cells during conditioned responses in ferrets. Neuroreport 5, 649–652.

Hesslow, G., Svensson, P., Ivarsson, M., 1999. Learned movements elicited by direct stimulation of mossy fiber afferents. Neuron 24, 179–185.

Ito, M., Sakurai, M., Tongroach, P., 1982. Climbing fibre induced depression of both mossy fibre responsiveness and glutamate sensitivity of cerebellar Purkinje cells. Journal of Physiology 324, 113–134.

Jirenhed, D.-A., Bengtsson, F., Hesslow, G., 2007. Acquisition, extinction, and reacquisition of a cerebellar cortical memory trace. Journal of Neuroscience 27, 2493–2502.

Kim, J.J., Krupa, D.J., Thompson, R.F., 1998. Inhibitory cerebello-olivary projections and blocking effect in classical conditioning. Science 279, 570–573.

Kishimoto, Y., Fujimichi, R., Araishi, K., Kawahara, S., Kano, M., Aiba, A., Kirino, Y., 2002. mGluR1 in cerebellar Purkinje cells is required for normal association of temporally contiguous stimuli in classical conditioning. European Journal of Neuroscience 16, 2416–2424.

Kleim, J.A., Freeman, J.H., Bruneau, R., Nolan, B.C., Cooper, N.R., Zook, A., Walters, D., 2002. Synapse formation is associated with memory storage in the cerebellum. Proceedings of the National Academy of Sciences USA 99, 13228–13231.

Krupa, D.J., Thompson, J.K., Thompson, R.F., 1993. Localization of a memory trace in the mammalian brain. Science 260, 989–991.

Krupa, D.J., Weng, J., Thompson, R.F., 1996. Inactivation of brainstem motor nuclei blocks expression but not acquisition of the rabbit's classically conditioned eyeblink response. Behavioral Neuroscience 110, 219–227.

Lewis, J.L., LoTurco, J.L., Solomon, P.R., 1987. Lesions of the middle cerebellar peduncle disrupt acquisition and retention of the rabbit's classically conditioned nictitating membrane response. Behavioral Neuroscience 101, 151–157.

Mauk, M.D., Donegan, N.H., 1997. A model of Pavlovian eyelid conditioning based on the synaptic organization of the cerebellum. Learning and Memory 3, 130–158.

Mauk, M.D., Steinmetz, J.E., Thompson, R.F., 1986. Classical conditioning using stimulation of the inferior olive as the unconditioned stimulus. Proceedings of the National Academy of Sciences USA 83, 5349–5353.

Mauk, M.D., Thompson, R.F., 1987. Retention of classically conditioned eyelid responses following acute decerebration. Brain Research 403, 89–95.

McCormick, D.A., Steinmetz, J.E., Thompson, R.F., 1985. Lesions of the inferior olivary complex cause extinction of the classically conditioned eyeblink response. Brain Research 359, 120–130.

McCormick, D.A., Thompson, R.F., 1984. Cerebellum: essential involvement in the classically conditioned eyelid response. Science 223, 296–299.

Medina, J.F., Nores, W.L., Mauk, M.D., 2002. Inhibition of climbing fibers is a signal for the extinction of conditioned eyelid responses. Nature 416, 330–333.

Medina, J.F., Nores, W.L., Ohyama, T., Mauk, M.D., 2000. Mechanisms of cerebellar learning suggested by eyelid conditioning. Current Opinion in Neurobiology 10, 717–724.

Mihailoff, G.A., 1993. Cerebellar nuclear projections from the basilar pontine nuclei and nucleus reticularis tegmenti pontis as demonstrated with PHA-L tracing in the rat. Journal of Comparative Neurology 330, 130–146.

Nolan, B.C., Freeman, J.H., 2006. Purkinje cell loss by OX7-saporin impairs acquisition and extinction of eyeblink conditioning. Learning & Memory 13, 359–365.

Ohyama, T., Nores, W.L., Medina, J.F., Riusech, F.A., Mauk, M.D., 2006. Learning-induced plasticity in deep cerebellar nucleus. Journal of Neuroscience 26, 12656–12663.

Racine, R.J., Wilson, D.A., Gingell, R., Sunderland, D., 1986. Long-term potentiation in the interpositus and vestibular nuclei in the rat. Experimental Brain Research 63, 158–162.

Ruigrok, T.J., Hensbroek, R.A., Simpson, J.I., 2011. Spontaneous activity signatures of morphologically identified interneurons in the vestibulocerebellum. Journal of Neuroscience 31, 712–724.

Schmaltz, L.W., Theios, J., 1972. Acquisition and extinction of a classically conditioned response in hippocampectomized rabbits (*Oryctolagus cuniculus*). Journal of Comparative and Physiological Psychology 79, 328–333.

Schneiderman, N., Fuentes, I., Gormezano, I., 1962. Acquisition and extinction of the classically conditioned eyelid response in the albino rabbit. Science 136, 650–652.

Schonewille, M., Gao, Z., Boele, H.-J., Vinueza Veloz, M.F., Amerika, W.E., Simek, A.A.M., De Jeu, M.T., Steinberg, J.P., Takamiya, K., Hoebeek, F.E., et al., 2011. Reevaluating the role of LTD in cerebellar motor learning. Neuron 70, 43–50.

Sears, L.L., Steinmetz, J.E., 1991. Dorsal accessory inferior olive activity diminishes during acquisition of the rabbit classically conditioned eyelid response. Brain Research 545, 114–122.

Shibuki, K., Gomi, H., Chen, L., Bao, S., Kim, J.J., Wakatsuki, H., Fujisaki, T., Fujimoto, K., Katoh, A., Ikeda, T., Chen, C., et al., 1996. Deficient cerebellar long-term depression, impaired eyeblink conditioning, and normal motor coordination in GFAP mutant mice. Neuron 16, 587–599.

Shinoda, Y., Sugiuchi, Y., Futami, T., Izawa, R., 1992. Axon collaterals of mossy fibers from the pontine nucleus in the cerebellar dentate nucleus. Journal of Neurophysiology 67, 547–560.

Shinoda, Y., Sugihara, I., Wu, H.S., Sugiuchi, Y., 2000. The entire trajectory of single climbing and mossy fibers in the cerebellar nuclei and cortex. Progress in Brain Research 124, 173–186.

Solomon, P.R., Solomon, S.D., Vander Schaff, E., Perry, H.E., 1983. Altered activity in the hippocampus is more detrimental to classical conditioning than removing the structure. Science 220, 329–331.

Steinmetz, J.E., 1990. Classical nictitating membrane conditioning in rabbits with varying interstimulus intervals and direct activation of cerebellar mossy fibers as the CS. Behavioural Brain Research 38, 97–108.

Steinmetz, J.E., Logan, C.G., Rosen, D.J., Thompson, J.K., Lavond, D.G., Thompson, R.F., 1987. Initial localization of the acoustic conditioned stimulus projection system to the cerebellum essential for classical eyelid conditioning. Proceedings of the National Academy of Sciences USA 84, 3531–3535.

Steinmetz, J.E., Rosen, D.J., Chapman, P.F., Lavond, D.G., Thompson, R.F., 1986. Classical conditioning of the rabbit eyelid response with a mossy-fiber stimulation CS: I pontine nuclei and middle cerebellar peduncle stimulation. Behavioral Neuroscience 100, 878–887.

Steinmetz, J.E., Sengelaub, D.R., 1992. Possible conditioned stimulus pathway for classical eyelid conditioning in rabbits. Behavioral and Neural Biology 57, 103–115.

Sugihara, I., Wu, H.S., Shinoda, Y., 2001. The entire trajectories of single olivocerebellar axons in the cerebellar cortex and their contribution to cerebellar compartmentalization. Journal of Neuroscience 21, 7715–7723.

Thompson, R.F., 1976. The search for the engram. American Psychologist 31, 209–227.

Wagner, A.R., 1969. Stimulus selection and a "modified continuity theory". In: Bower, G.H., Spence, J.T. (Eds.), The Psychology of Learning and Motivation, vol. 3. Academic Press, New York, pp. 1–41.

Woodruff-Pak, D.S., Steinmetz, J.E. (Eds.), 2000. Eyeblink Classical Conditioning. Applications in Humans, vol. 1. Kluwer Academic Publishers, Boston.

Yeo, C.H., Hardimann, M.J., Glickstein, M., 1986. Classical conditioning of the nictitating membrane response of the rabbit IV. Lesions of the inferior olive. Experimental Brain Research 63, 81–92.

Eyewitness Testimony

Robert A Nash, University of Surrey, Guildford, UK
Maciej Hanczakowski, Cardiff University, Cardiff, UK
Giuliana Mazzoni, University of Hull, Hull, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Eyewitnesses have always played a fundamental role in the legal process. However, eyewitness evidence is not always as reliable or as accurate as one might hope, given the weight that decision makers often ascribe to it. In this article, we review some of the many factors that can determine the reliability of witness testimony, including factors that are, and those that are not, within the control of investigative and judicial processes. In doing so, we consider how policies might assist in ensuring that witness testimony can play the most effective and informative role in the legal process as possible.

Eyewitness testimony refers to the descriptions that bystanders or victims give to investigators, or in the courtroom, to describe what they can recall about the incidents under investigation. Since antiquity, eyewitnesses' testimony has overwhelmingly constituted the main basis for courts' and juries' decisions to convict or to acquit. Even today, many court verdicts are based almost exclusively on evidence provided by eyewitnesses. This is true even when other available case evidence clearly contradicts the eyewitnesses' reports and, importantly by extension, even when the witnesses are factually incorrect. The severity of this situation is epitomized by data from the Innocence Project, a public policy organization based in the USA. Using DNA testing, more than 300 individuals have been found to be wrongfully convicted, most commonly of murder or rape, and have consequently been exonerated. It is important to consider that these cases undoubtedly represent only a fraction of the true number of wrongful convictions, not least because in many cases forensic trace evidence was never, or is no longer, available to be DNA tested. More importantly, of these 300-plus cases, more than 75% were convicted at least in part on the basis of mistaken eyewitness identification (www.innocenceproject.org; Wells and Quinlivan, 2009). This rather worrisome situation can be linked to the fact that eyewitnesses' reports are typically perceived as more convincing than many other types of legal evidence, especially when those reports are expressed with high levels of confidence. In this article, we outline some of the most important factors – including confidence – that have been studied in the context of understanding what leads witnesses to be more versus less reliable.

Any given testimony is the product of a number of cognitive (intrapersonal) and social (interpersonal) processes. Perceptual and attentional processes are involved during the moments when individuals actually witness or are victims of a crime. Memory processes are involved in encoding, storing, and later retrieving the details of what has been witnessed. Metacognitive processes are involved in witnesses' decisions whether to report specific details that they remember. And in each of these instances, the social context within which these processes occur also plays a vital role. In this article we only briefly cover, for the sake of completeness, the initial perceptual and attentional processes, with particular emphasis on a phenomenon termed the 'weapon focus effect.' We devote more space to the complex and highly debated issue of witnesses' susceptibility to giving inaccurate reports of events. In this regard, we discuss the repeated finding that witnesses' confidence in their memories is only poorly related to the accuracy of their reports, and we also consider the widely studied phenomenon of the misinformation effect. Research in this area shows us how readily witnesses' reports can be influenced by authority figures, and by other witnesses of an incident. We next turn to consider research that debates how best to gather useful information from witnesses, both in the context of investigative interviews, and in the context of identifying culprits in police lineup procedures. An examination of factors that affect judges' perceptions of the reliability and credibility of witnesses will represent the last part of this article.

Eyewitness Factors

Given the importance and the persuasiveness of eyewitness testimony, it is of utmost importance for witnesses' reports to be accurate and reliable. Unfortunately, as confirmed now by many hundreds of studies, the reliability of eyewitness reports depends on a countless number of variables, some of which can lead the testimonial evidence to be weak, and some of which can lead it to be factually inaccurate. These variables have traditionally been divided into two types: system variables and estimator variables.

System variables are those that can be controlled through the processes and procedures of eliciting witnesses' memory reports; for example, the amount of time that investigators invest in building rapport with the witnesses in advance of their formal interviews, or the instructions given to witnesses before they participate in a lineup identification procedure. We will discuss many such variables in further detail later in this article. Estimator variables, in contrast, are those that cannot be controlled, but that nevertheless play important roles in determining the reliability of the end product. For example, we are unable to control the age of the witnesses, or the amount of alcohol they had imbibed at the time the incident occurred. Other important estimator variables include the perceptual conditions under which the bystander witnessed the incident (e.g., the lighting conditions) and other attentional factors (e.g., the witness's level of concentration and distraction).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.51033-2

Among the most widely studied estimator variables is the presence of a weapon during the incident, which can lead to a perceptual effect known as 'weapon focus.' Weapon focus refers to the tendency of witnesses to pay particularly high levels of attention to weapons when one is present during a crime. This attention to the weapon, in turn, comes at the expense of attention paid to other elements of the scene. As a result, the presence of weapons has been shown to lead to a poorer ability of witnesses to describe a perpetrator's appearance and to identify the perpetrator at a later time (Fawcett et al., 2013). Among the multitude of studies on this phenomenon, many authors have argued that in fact the 'weapon focus' effect may have little to do with weapons per se. Instead, this effect may be related to how a weapon is perceived in its situational context. Pickel (1999), for instance, showed her participants different videos depicting a perpetrator with a gun. She found that participants who saw the armed perpetrator at a baseball field were subsequently far less able to describe him than those who saw him at a shooting range. Similarly, Mitchell et al. (1998) found a comparable 'weapon focus' effect among participants who saw a perpetrator holding a gun, as among different participants who saw a perpetrator holding a stick of celery. Together, these data suggest that it is the unusualness of a weapon, rather than the threat that it seems to pose, which attracts attention. What the 'weapon focus' effect illustrates well, though, is that the reliability of witnesses can be influenced quite easily by straightforward visual elements of the incident. With regard to estimator variables more broadly, even though these are factors that the justice system cannot control, empirical research can nonetheless help us to understand these factors as means to predicting and understanding the circumstances under which witnesses will be most reliable.

Confidence–Accuracy Relationship

We noted above that the confidence of a witness can greatly determine judges' assessments of the reliability of their testimony. Confidence, unlike the other factors we have outlined so far, can be considered both an estimator and a system variable. That is to say, whereas it is largely influenced by the circumstances under which an event was witnessed, it can also be shaped and molded by subsequent social and investigative processes. A major focus of research on witness confidence has been on the confidence–accuracy relationship; that is, the question of whether witnesses who demonstrate higher confidence for testimonies are more likely to be correct, compared to witnesses who demonstrate lower confidence.

A conviction nowadays widely held by eyewitness researchers is that the accuracy of people's memory reports and their confidence in those reports are only weakly related, with most empirically derived correlations ranging between 0 and around 0.3 – by convention a weak-to-moderate relationship at best (e.g., Bothwell et al., 1987). Recently, however, a conceptual shift in thinking has occurred, and a new consensus has started to emerge that confidence may constitute a somewhat better predictor of accuracy than had previously been assumed. Specifically, researchers have argued that the weak relationship often documented between accuracy and confidence stems in part from the narrow testing conditions that are employed to assess this relationship, and in part from inappropriate statistical methods used to assess this relationship. With regard to the former reason, experimental studies on eyewitness memory typically aim to create homogenous study and testing conditions for all participants; that is, all participants experience a to-be-remembered event under highly similar conditions. This homogeneity leads to minimal variability in participant–witnesses' memory performance, which is in contrast to the conditions that occur in real life where different witnesses might observe the same event from different perspectives and distances, with different amounts of exposure time, and so forth. When variability in viewing conditions is minimized, the correlation coefficients that are used to assess the confidence–accuracy relationship can be artificially reduced. Indeed, in studies in which highly variable encoding conditions were created, larger (and therefore more meaningful) confidence–accuracy correlations were observed (Lindsay et al., 1998).

With regard to the second reason, several recent investigations into the confidence–accuracy relationship have recruited alternative analytic approaches, in place of the correlation analysis that is conventionally used in this literature. One such approach is calibration (Juslin et al., 1996). Calibration statistics tell us about the probability, based upon obtained data, that a given memory report expressed with a given degree of confidence is accurate, as illustrated in **Figure 1**. By using the calibration approach, Neil Brewer and his colleagues have demonstrated that confidence judgments generally track the accuracy of memory reports rather more meaningfully than correlation analyses would suggest (Brewer, 2006). In the area of eyewitness identification, where this method is now increasingly applied, it has been shown that confidence judgments are a reasonably good marker of accuracy of

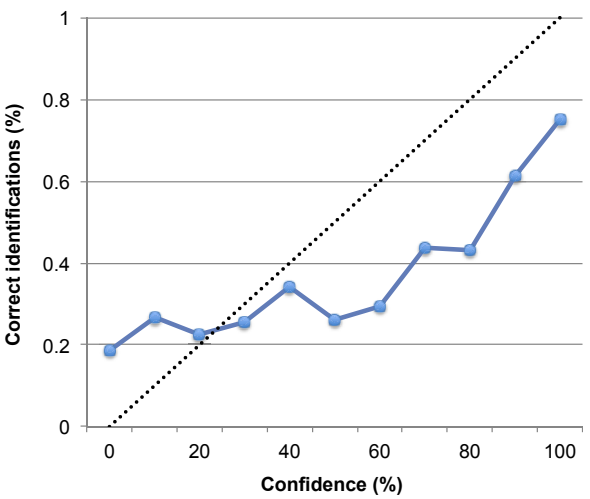

Figure 1 A calibration curve representing data from 1000 hypothetical eyewitnesses; the dotted line represents 'perfect' calibration of witnesses' confidence and accuracy. The hypothetical data illustrate the most common pattern of findings in this literature: the confidence–accuracy correlation for the data set is weak to moderate in magnitude ($r = 0.3$), yet the calibration curve shows that witnesses' confidence judgments generally track their accuracy quite well, with a degree of overconfidence observed.

identifications among people who actually pick a person from a lineup (so-called 'choosers'). However, this relationship is weaker and not meaningful for those who reject the lineup (nonchoosers) – a distinction that has also been obtained in studies that used the conventional correlation approach to analysis (Sporer et al., 1995).

In sum then, there is some evidence that the strength of the confidence–accuracy relationship has commonly been underestimated, and that at least in some conditions, confidence can be a fairly good index of accuracy. Nonetheless, we should note that, however, it is assessed, a perfect confidence–accuracy relationship could occur only if confidence judgments were based on exactly the same information that shapes the accuracy of a memory report. This, however, is certainly not the case: confidence is often determined by external factors that have no influence on memory performance. The most obvious example of such a situation is the case of postidentification feedback in eyewitness identification experiments. Confirmatory feedback received after the identification decision ('Good, you identified the suspect') generally elevates witnesses' confidence in the correctness of this decision, although it clearly cannot have any influence on the actual accuracy of the decision itself (Douglass and Steblay, 2006). Research has shown that confirmatory feedback can influence witnesses' confidence not only when it is provided verbally, but also when it is provided through nonverbal and paraverbal cues, for instance, through nods or shakes of the head, or voice intonation, which might not be intended or even consciously recognized by the investigator (Garrioch and Brimacombe, 2001). In cases such as these, confidence and memory accuracy are at least partly based on different types of information, which may cumulatively contribute to an imperfect confidence–accuracy relationship.

To summarize, there are a variety of situations in which confidence may become dissociated from accuracy. These situations arise whenever accuracy is determined by different factors than confidence. It has to be noted, though, that the abundance of examples of confidence–accuracy dissociations reported in the literature can be partly explained by methodological and statistical factors. In the recent studies that recruit more precise measures of the confidence–accuracy relationship, and those that avoid homogenizing witnesses' experiences, confidence tracks the accuracy of a memory report rather more meaningfully than has more usually been concluded. What is undisputed, though, is that the relationship between confidence and accuracy is both imperfect, and susceptible to influence.

The Effects of Misinformation

As well as it being relatively easy to influence witnesses' confidence in their memory reports, substantial evidence has now been amassed showing that the actual accuracy of these reports is also relatively easy to influence. That is to say, witnesses' reports and identifications can be manipulated and changed as a product of various suggestive procedures. In the 1970s, psychologist Elizabeth Loftus pioneered research into what has been called the *misinformation effect*, a term that refers to the distortion of people's beliefs about – and memories of – past experiences as a consequence of receiving erroneous information about those experiences ('postevent information').

Given the weight eyewitness reports have on jurors' decisions, tainted reports are undoubtedly consequential in legal proceedings. The misinformation effect is among the most robust effects in the legal psychology domain, and has been replicated and elucidated by generations of researchers the world over. Loftus' early work was influential in demonstrating the simplicity with which witnesses' memory reports could be altered, for example, simply through the use of minimally leading questions or through the incidental mentioning of incorrect details during questioning (e.g., Loftus, 1975). Hundreds of studies since have given us a deeper understanding of how postevent information works, under what conditions, and what the consequences might be. In particular, there has been considerable debate over the question of whether misinformation actually changes witnesses' memories, or whether it simply supplements the original memories, perhaps making them less easily accessed (e.g., McCloskey and Zaragoza, 1985). As well as having theoretical importance insofar as our understanding of human memory is concerned, the answer to this question is also of practical importance – if misinformation actually changes memories, then no interviewing practice would successfully retrieve the original memories. Even to date, there is no firm consensus on the answer to this question, and there has been considerable evidence for both sides. What is certainly clear, though, is that postevent information is sufficient to lead witnesses to confidently report details that are factually incorrect. Moreover, discriminating a correct from an incorrect witness report can be extremely difficult in practice. Recent studies have demonstrated that witnesses can be led to implicate an innocent person in a wrongdoing even when the witnesses know that their testimony will have negative consequences for the other person (Kaasa et al., 2013; Wade et al., 2010).

The question of exactly who is most prone to the effects of postevent information has been extensively studied and debated over the years. Validated scales have been created, and widely used, to understand both the individual and situational predictors of interrogative suggestibility (that is, individuals' susceptibility to suggestions; e.g., Gudjonsson, 1984). Among other applications, these scales have been used to demonstrate that interrogative suggestibility is increased when interviewers behave in a 'firm' rather than friendly manner (Baxter et al., 2006), and that a variety of demographic and personal characteristics of an interviewee – including young or old age, learning disabilities, and life adversity – can make individuals vulnerable in this sense (Clare and Gudjonsson, 1993; Drake, 2010). More recently, research with Chinese participants demonstrated, among other findings, that people who are more intelligent and have better working memory ability are somewhat more immune to the effects of misinformation than are their peers (Zhu et al., 2010). Frost et al. (2013) recently showed that specific combinations of individual difference factors can selectively enhance acceptance of misinformation at different temporal levels. They showed that people who scored highly on measures of compliance tended to be those who would accept misinformation immediately, whereas those who scored highly on measures of fantasy proneness were those for whom the misinformation had robust and lasting effects. In other words, susceptibility to misinformation might rely on the interactions of different traits that separately predict

the acceptance of, and the elaboration of, the suggested information.

Cowitness Misinformation

As illustrated by the above studies, the problem of misinformation in the legal process has typically been studied and discussed with a focus on postevent suggestions provided by investigators, or otherwise authoritative sources. But in recent years, many researchers have begun to explore the effects and consequences of misinformation that is provided by cowitnesses – that is, by other bystanders or victims who also observed the incident. Survey research shows us that in a large majority of cases in which a crime has been witnessed, there is in fact more than one witness. Furthermore, in a large majority of cases involving multiple witnesses, cowitnesses discuss with each other what they saw, prior to giving initial testimony (Paterson and Kemp, 2006). Researchers have expressed concern that these cowitnesses could perhaps influence each other's memory reports, and these researchers turned out to be quite justified in their concerns. There are now a considerable number of studies demonstrating that discussion between two or more cowitnesses can lead individuals to confidently report observing details that only their colleagues observed. This 'cowitness misinformation effect' seems to be robust, but is considerably more prevalent in cases where the cowitnesses have a preexisting friendship or romantic relationship (Hope et al., 2008), and in cases where one observer is perceived to be a 'better' witness than the other (French et al., 2011). The cowitness misinformation effect is of particular concern for legal and investigative processes. Whereas police officers, for example, can be rigorously trained to avoid using suggestive interview practices, little can be done to prevent witnesses from speaking to each other, especially prior to their first contact with investigators. There is some evidence, though, that explicitly encouraging witnesses not to report details that they only remember being told, not actually seeing, can mitigate the effects (Bodner et al., 2009).

Gathering Evidence from Witnesses

Given the overwhelming evidence that witnesses' testimonies are malleable and susceptible to external influence, an important question is what we might do to minimize the likelihood of obtaining unreliable testimonies. We consider this issue here within two broad contexts: First, we consider established interviewing techniques that have been proposed – and indeed applied – as 'best practice' for gathering information from witnesses. Second, we consider the procedures used for obtaining suspect-identification evidence, via the use of police 'lineups.'

Interviewing Techniques

Several specific interviewing techniques have been proposed by researchers to ensure the maximum informativeness of witnesses' reports (in terms of the amount of detail they report) while also maintaining their quality (in terms of the accuracy of that detail). Of those techniques, the Cognitive Interview is to date the most studied (Memon et al., 2010), and below we focus mainly on this approach, though it is worth noting that this approach shares many commonalities with other best-practice interviewing protocols (e.g., the National Institute of Child Health and Development's Investigative Interview Protocol; Orbach et al., 2000), and we outline these general commonalities first. Most recommended interviewing techniques share the same basic principles of good practice, which can be summarized in the following points (see, for example, the Memorandum of Good Practice published by the United Kingdom's Home Office (1992; see also subsequent extensions and variations, for example, Westcott and Jones, 1997; Ministry of Justice, 2011)). First, efforts should be taken to build rapport and trust with a witness before any direct interviewing about the incident in question takes place. This rapport building can put witnesses at ease, can reduce social pressure upon them, and has been shown in some studies to actually improve the quantity and quality of information they subsequently provide (Vallano and Schreiber Compo, 2011). Second, the interview should begin with an opportunity for the witness to provide a free and uninterrupted account of what happened. Third, following this account, if it is necessary to ask questions of the witness, then these questions should be as open-ended as possible so as not to lead the witness to give any implied or suggested response. One can proceed to ask closed questions only when the material provided using open questions is deemed insufficient, and when closed questions are asked, these should be carefully worded so as not to contain new and/or leading information. One element stressed in some of the interviewing techniques is to preface the interview with a clear statement of what is and what is not expected during the interview – often referred to as establishing 'ground rules' (Memon et al., 2010). For example, one important way to enhance the accuracy of the report is to clarify that the witness is allowed to say 'I do not know' in response to questions (Orbach et al., 2000). In general, research shows us that when police officers are trained to interview witnesses by following best-practice recommendations such as those we have described, they conduct better quality interviews as a result, and can elicit information from witnesses that is of greater investigative value (e.g., Pipe et al., 2008).

As we mentioned above, the Cognitive Interview is the investigative interviewing technique that has arguably received the most empirical scrutiny in the psychological literature. The technique, originally developed by psychologists Ed Geiselman and Ron Fisher in the 1980s (Geiselman et al., 1985) and since refined and augmented in further revisions, is specifically designed to maximize the amount of information that a witness can provide. The Cognitive Interview in its original format comprised four components that could be used to assist witnesses' recollections, each of which was conceived on the basis of psychological research on remembering. The first is known as mental reinstatement of context, and involves encouraging the witness to carefully visualize the incident in their mind's eye, and to bring to mind the sounds, smells, thoughts, and feelings that they had experienced at the time. During this exercise, witnesses might be encouraged to close their eyes, and recent research has shown that this too can enhance recollection (Vredeveldt and Penrod, 2013). The second component is termed 'report everything,' and involves

instructing the witness not to selectively edit their accounts on the basis of what they personally think is important, or what they think the investigator is looking for. Reporting everything is important not only because ostensibly trivial details can often turn out to be critical, but also because recalling these trivial details can sometimes trigger the recollection of other, less trivial details. The third and fourth Cognitive Interview components require recalling the incident in different ways – from different perspectives (trying to visualize the events from a different viewpoint than one's own) and in different orders (trying to recall in reverse order, or starting from the middle rather than from the beginning of the event). These latter stages have rather fallen out of favor in recent times, however, as studies show not only that police rarely use the techniques (Dando et al., 2008), but also that removing them from the protocol can speed up interviewing times without any meaningful loss of information reported, as illustrated in **Figure 2** (Dando et al., 2009).

Substantial research has validated the Cognitive Interview method. Dozens of studies conducted both in laboratory settings and in the field have shown that using the Cognitive Interview components substantially increases the amount of correct detail that witnesses recall, although a smaller concomitant increase in incorrect detail is also often obtained (Memon et al., 2010). In short, witnesses interviewed using the Cognitive Interview technique tend to give more complete and overall more accurate accounts. No wonder, then, that this approach is now formally used in police training in several countries including the United Kingdom, New Zealand, and Israel.

One serious practical issue is that even among these proven effective interviewing techniques, none can necessarily be administered immediately after an incident occurs; that is to say, due to limitations in police resources, it may often be several days or even weeks after a crime occurs before investigators are able to properly interview the witnesses. The problem then remains that during this interval, considerable amounts of useful information might be forgotten, and external sources might interfere with or add to witnesses' recollections. To obviate this problem, a new method has recently been proposed, labeled the 'Self-Administered Interview' (Gabbert et al., 2009), which has now been extensively tested in the field. The method is based upon the principles of the Cognitive Interview, but translated into a paper-based interview protocol that a witness completes independently at their earliest opportunity, in advance of a formal police interview that might take place at a later time. Gabbert et al. (2009) showed – as have several researchers since – that completing a Self-Administered Interview can reduce the extent of forgetting of incident-related information during a delay of around 1 week. Using this tool, therefore, represents one straightforward and effective way by which investigators might maintain the quality of witness evidence even in the face of difficult-resourcing problems. Nash et al. (2014) recently proposed that conducting remote interviews – without the physical copresence of the interviewer and witness – might be another possible solution to this delay problem in some instances. They argued that in some cases, such remote interviewing might circumvent some of the causes of delays. Nash et al.'s data supported the notion that conducting a remote interview sooner would lead to better recall than would conducting a face-to-face interview later; however, further empirical exploration of this modern technological solution is clearly needed. In general, it is clear that research can play a considerable role in finding and testing solutions to difficult and ingrained problems in investigative processes, which stand to improve the integrity and completeness of eyewitness evidence.

Lineups

Lineup procedures, referred to in some countries as 'identity parades,' are an example of a recognition task in which

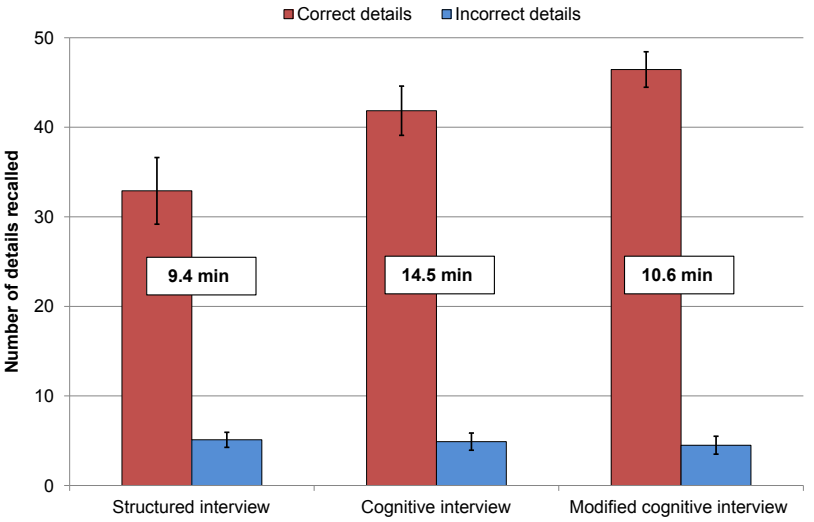

Figure 2 Results from Dando, C.J., Wilcock, R., Milne, R., Henry, L. (2009. A modified cognitive interview procedure for frontline police investigators. Applied Cognitive Psychology 23, 698–716), in which 60 eyewitnesses were interviewed about a recent event using either a standard interview protocol (i.e., not involving the cognitive interview components), a cognitive interview, or a modified cognitive interview that omitted the change-order and change-perspective components. Error bars are 95% confidence intervals and mean interview durations are shown for each condition.

a witness has to decide whether a previously seen person is present among a selection of possible candidates. Lineups can be distinguished from 'showups,' in which a witness is presented with just one person and asked to decide whether this person is the perpetrator or not. Even though it is significantly more laborious to administer a lineup than a showup, the popularity of lineups is based on a general consensus that lineups are fairer procedures than are showups (Wells and Quinlivan, 2009).

Emerging from decades of studies on lineups, a number of guidelines have been proposed pertaining to how lineups should best be administered so as to obtain reliable identification decisions (e.g., Wells et al., 1998). Among other types of advice, these guidelines cover the instructions that should be given to witnesses prior to the procedure, the person who should administer the lineup, and the way by which the foils (that is, people who are known to be innocent) in the lineups should be chosen. There is a general consensus that the instructions given to a witness should be unbiased, that is, they should stress that a perpetrator may or may not be present in the lineup (Steblay, 1997). Studies show that providing biased instructions leads to a higher rate of both correct and incorrect identifications, which, therefore, means that such instructions increase the risk that an innocent suspect will be prosecuted. Researchers also agree that a person administering the lineup should not be aware of which lineup member is the suspect, so that the administrator could not deliberately or unwittingly influence a witness through nonverbal or verbal cues (Wright et al., 2010). Finally, it has been proposed that foils in a lineup should be chosen not based on similarity to the suspect but based on a witness's description of a perpetrator (Luus and Wells, 1991). One good reason for this rule is that if the foils are all chosen to look as similar as possible to the suspect, this might mean the witness's task is made unnecessarily difficult. So long as the foils and suspect all match the witness's description, however, there should be no unfair bias against the suspect, but the recognition cues by which the witness might make a confident identification will be preserved and made more salient. In other words, using a culprit-description strategy rather than a suspect-similarity strategy should improve the likelihood of identifying the perpetrator, without unfairly leading the witness to pick the suspect.

A very large number of research studies on lineups have been devoted to the issue of the format of presentation. Usually, there are two possible formats by which a lineup could be administered. In a simultaneous lineup, a suspect and all foils are presented at the same time, and a witness is asked to make a single decision: to either identify one member of the lineup as being the perpetrator, or to reject the lineup by saying that the perpetrator is not present. In a sequential lineup, a suspect and foils are instead presented one-by-one, and the witness is asked to decide for each presented person whether he or she is the perpetrator. The interest in these two formats of presentation was sparked by reports showing that sequential lineups lead witnesses to use more reliable decision-making processes than do simultaneous lineups (e.g., Lindsay and Wells, 1985). Specifically, it has been argued that sequential lineups prevent witnesses from using a relative judgment strategy, whereby the members of the lineup are compared to each other with the aim of determining, which looks most similar to the perpetrator. This relative judgment strategy is thought to underlie many false identifications, whereby the suspect looks like but is not in fact the perpetrator. Reports on the superiority of sequential lineups have been convincing enough to persuade some legal authorities to promote the sequential approach in their actual investigative proceedings. Nevertheless, similar to the discussion above regarding the confidence–accuracy correlation, some researchers have more recently suggested that the sequential superiority effect is actually an artifact of the statistical techniques used to examine the data. Using what they argue are more appropriate statistical techniques, Mickes et al. (2012) have shown evidence that the simultaneous procedure is, in fact, the better method for discriminating between the presence and absence of a perpetrator in a lineup. Despite the abundance of evidence thus far for preferring sequential lineups, arguments such as these will no doubt lead to considerable empirical and policy scrutiny in the years to come.

When considering lineup identification data, it is important to distinguish between the general propensity of a witness to select a person in a lineup, and the actual reliability of these decisions. The relevance of this distinction extends beyond the discussion on the sequential vs. simultaneous debate. Indeed, in a recent overview of the guidelines proposed to enhance reliability of lineup identifications, Clark (2012) argued that the best-practice guidelines outlined above, including unbiased instructions and the 'double blind' administrator rule, affect witnesses' *tendency* to identify somebody rather than the *reliability* of their identification decisions. In other words, various techniques can make witnesses less ready to identify someone in the lineup, and these techniques will, therefore, result in a lower rate of wrongful identifications but at the important cost that fewer correct identifications would also be made. Countering this argument, other authors pointed out that many of the correct identifications made under suggestive or otherwise inadequate lineup procedures might not be legitimate 'hits.' Rather, they might well include guesses and selections based upon external 'hints' (Newman and Loftus, 2012). These authors argue that we do not gain from increasing correct identifications for the wrong reasons. In short, there is clearly a degree of ambiguity regarding what kinds of identifications are desirable; that is, to what extent the 'ends justify the means.' This ambiguity is compounded by the difficulty of knowing in practice, when the perpetrator is not known, whether a witness's choice is actually based upon true recognition, rather than on unwanted influence.

Jurors' Appraisals of Witness Reliability

Whereas witnesses undoubtedly vary in their accuracy, it is the task of legal decision makers such as judges and juries to determine how much weight to ascribe to a given testimony. Whether a testimony or identification that is made by a witness will influence judges and jurors depends on the extent to which that witness is perceived to be reliable, that is, the extent to which he or she is judged to have an accurate recollection of the incident.

The accuracy of jurors' appraisals of witness reliability will depend partly on the accuracy of their beliefs concerning how

human memory works. For example, if jurors believe that postevent information would not affect the accuracy of a testimony – a belief starkly contrasting with experts' knowledge about eyewitness testimony – then the jurors' perception of the reliability of a given memory report may be severely distorted. A number of studies have, therefore, looked at how potential jurors' lay theories of memory match the expert consensus informed by empirical studies. A recent meta-analysis of these studies, conducted by Desmarais and Read (2011), suggests that in some respects potential jurors' beliefs match quite well what experts know about the reliability of testimonies, but that they clearly contrast with experts' knowledge in other respects. However, one issue that must be taken into account is that in the majority of studies on this topic, juror knowledge has been measured through respondents' ability to correctly answer multiple-choice questions on various eyewitness topics. The extent to which the ability to correctly answer such questions is an adequate index of 'knowledge,' and the extent to which this knowledge would be spontaneously applied by jurors in a real legal case, are disputable. As a result, one might conclude that the estimates of juror knowledge obtained in these studies are 'best-case scenarios,' and further research with different measures of knowledge would clearly be useful.

Besides the knowledge of factors that may affect the reliability of testimony, jurors' appraisals of testimony are also determined by the features of the testimony itself. One factor determining perceptions of reliability is the sheer amount of details provided in a testimony. Research shows that a testimony is often perceived to be more accurate when it contains more details, even if these details are trivial and of little relevance to the central issue about which the witness testifies (Bell and Loftus, 1988). Also, a large number of studies clearly indicate that perceived reliability is strongly affected by witnesses' declared confidence in their own testimony (e.g., Whitley and Greenberg, 1986). The more confidently the statement seems to be made, the more accurate and reliable it is perceived to be. As discussed in the section above concerning confidence–accuracy relationship, this strategy – although reasonable under many circumstances – could lead to misplaced reliance on witnesses whenever their confidence and accuracy have become dissociated.

With these findings in mind, it is worth noting that jurors do seem to appropriately balance the persuasiveness of the details provided in a testimony, and that of witnesses' confidence, against the persuasiveness of evidence from other sources. One major clue to the reliability and credibility of an eyewitness often lies in this congruency of evidence. When an eyewitness is contradicted even on a trivial detail, the assessed reliability of that eyewitness is starkly reduced (Wells and Leippe, 1981). Also, when a witness is proven wrong on a detail that was reported with high confidence, this witness's perceived reliability is undermined by a much greater extent than when a witness is proven wrong on a detail reported with low confidence (Tenney et al., 2007). Therefore, although confidence is often considered by judges and juries as an indicator of reliability, the apparent relationship between witnesses' confidence and accuracy on individual elements of their reports can also shape the appraisal of their testimony.

Conclusion

As we noted at the beginning of this article, law courts in many countries have always readily based their decisions on witness evidence, and this continues to be true today. This is most likely for three reasons. First, in many circumstances no other evidence is available. Second, variables that are capable of modifying witnesses' reports are so ubiquitous that discarding evidence that is potentially unreliable might lead to discarding most witnesses' reports. Finally, witness evidence is typically interpreted as being highly convincing, both by judges and by members of the jury. For these reasons, it remains highly important that investigators, judges, jurors, and policymakers all understand the circumstances under which witnesses are most likely to be reliable, and know the investigative procedures that can help us to maximize this accuracy. Such understanding will help us to ensure that witnesses play the most effective and informative role in the legal system as possible.

See also: Crime and Criminal Justice: Applied Research from Routine Monitoring to Evidence-Based Practices; Episodic Memory; False Memories; Injustice: Legal Aspects; Juries; Justice and Law; Justice: Social Psychological Perspectives; Policing; Policy Analysis: Evidence Based Policy-Making; Recognition; Reconstructive Memory, Psychology of; Science and Law; Source Monitoring; Wrongful Convictions.

Bibliography

Baxter, J.S., Boon, J.C., Marley, C., 2006. Interrogative pressure and responses to minimally leading questions. Personality and Individual Differences 40, 87–98.

Bell, B.E., Loftus, E.F., 1988. Degree of detail of eyewitness testimony and mock juror judgments. Journal of Applied Social Psychology 18, 1171–1192.

Bodner, G.E., Musch, E., Azad, T., 2009. Reevaluating the potency of the memory conformity effect. Memory & Cognition 37, 1069–1076.

Bothwell, R.K., Deffenbacher, K.A., Brigham, J.C., 1987. Correlation of eyewitness accuracy and confidence: optimality hypothesis revisited. Journal of Applied Psychology 72, 691–695.

Brewer, N., 2006. Uses and abuses of eyewitness identification confidence. Legal and Criminological Psychology 11, 3–23.

Clare, I.C.H., Gudjonsson, G.H., 1993. Interrogative suggestibility, confabulation, and acquiescence in people with mild learning disabilities (mental handicap): implications for reliability during police interrogations. British Journal of Clinical Psychology 32, 295–301.

Clark, S.E., 2012. Costs and benefits of eyewitness identification reform: psychological science and public policy. Perspectives on Psychological Science 7, 238–259.

Dando, C., Wilcock, R., Milne, R., 2008. The cognitive interview: Inexperienced police officers' perceptions of their witness/victim interviewing practices. Legal and Criminological Psychology 13, 59–70.

Dando, C.J., Wilcock, R., Milne, R., Henry, L., 2009. A modified cognitive interview procedure for frontline police investigators. Applied Cognitive Psychology 23, 698–716.

Desmarais, S.L., Read, J.D., 2011. After 30 years, what do we know about what jurors know? A meta-analytic review of lay knowledge regarding eyewitness factors. Law and Human Behavior 35, 200–210.

Douglass, A.B., Steblay, N., 2006. Memory distortion in eyewitnesses: a meta-analysis of the post-identification feedback effect. Applied Cognitive Psychology 20, 859–869.

Drake, K.E., 2010. Interrogative suggestibility: life adversity, neuroticism, and compliance. Personality and Individual Differences 48, 493–498.

Fawcett, J.M., Russell, E.J., Peace, K.A., Christie, J., 2013. Of guns and geese: a meta-analytic review of the 'weapon focus' literature. Psychology, Crime & Law 19, 35–66.

French, L., Garry, M., Mori, K., 2011. Relative–not absolute–judgments of credibility affect susceptibility to misinformation conveyed during discussion. Acta Psychologica 136, 119–128.

Frost, P., Nussbaum, G., Loconto, T., Syke, R., Warren, C., Muise, C., 2013. An individual differences approach to the suggestibility of memory over time. Memory 21, 408–416.
Gabbert, F., Hope, L., Fisher, R.P., 2009. Protecting eyewitness evidence: examining the efficacy of a self-administered interview tool. Law and Human Behavior 33, 298–307.
Garrioch, L., Brimacombe, C.E., 2001. Lineup administrators' expectations: their impact on eyewitness confidence. Law and Human Behavior 25, 299–315.
Geiselman, R.E., Fisher, R.P., MacKinnon, D.P., Holland, H.L., 1985. Eyewitness memory enhancement in the police interview: cognitive retrieval mnemonics versus hypnosis. Journal of Applied Psychology 70, 401–412.
Gudjonsson, G.H., 1984. A new scale of interrogative suggestibility. Personality and Individual Differences 5, 303–314.
Home Office, 1992. Memorandum of Good Practice on Video Recorded Interviews with Child Witnesses for Criminal Proceedings. Author, London.
Hope, L., Ost, J., Gabbert, F., Healey, S., Lenton, E., 2008. "With a little help from my friends…": the role of co-witness relationship in susceptibility to misinformation. Acta Psychologica 127, 476–484.
Juslin, P., Olsson, N., Winman, A., 1996. Calibration and diagnosticity of confidence in eyewitness identification: comments on what can be inferred from the low confidence–accuracy correlation. Journal of Experimental Psychology: Learning, Memory, and Cognition 22, 1304–1316.
Kaasa, S.O., Cauffman, E., Clarke-Stewart, K.A., Loftus, E.F., 2013. False accusations in an investigative context: differences between suggestible and non-suggestible witnesses. Behavioral Sciences & the Law 31, 574–592.
Lindsay, D.S., Read, J.D., Sharma, K., 1998. Accuracy and confidence in person identification: the relationship is strong when witnessing conditions vary widely. Psychological Science 9, 215–218.
Lindsay, R.C., Wells, G.L., 1985. Improving eyewitness identifications from lineups: simultaneous versus sequential lineup presentation. Journal of Applied Psychology 70, 556–564.
Loftus, E.F., 1975. Leading questions and the eyewitness report. Cognitive Psychology 7, 560–572.
Luus, C.E., Wells, G.L., 1991. Eyewitness identification and the selection of distracters for lineups. Law and Human Behavior 15, 43–57.
McCloskey, M., Zaragoza, M., 1985. Misleading postevent information and memory for events: arguments and evidence against memory impairment hypotheses. Journal of Experimental Psychology: General 114, 1–16.
Memon, A., Meissner, C.A., Fraser, J., 2010. The Cognitive Interview: a meta-analytic review and study space analysis of the past 25 years. Psychology, Public Policy, and Law 16, 340–372.
Mickes, L., Flowe, H.D., Wixted, J.T., 2012. Receiver operating characteristic analysis of eyewitness memory: comparing the diagnostic accuracy of simultaneous versus sequential lineups. Journal of Experimental Psychology: Applied 18, 361–376.
Ministry of Justice, 2011. Achieving Best Evidence in Criminal Proceedings: Guidance on Interviewing Victims and Witnesses, and Guidance on Using Special Measures. Retrieved from: www.cps.gov.uk/publications/docs/best_evidence_in_criminal_proceedings.pdf.
Mitchell, K.J., Livosky, M., Mather, M., 1998. The weapon focus effect revisited: the role of novelty. Legal and Criminological Psychology 3, 287–303.
Nash, R.A., Houston, K.A., Ryan, K., Woodger, N., 2014. Remembering remotely: would video-mediation impair witnesses' memory reports? Psychology, Crime & Law 20, 756–768.
Newman, E.J., Loftus, E.F., 2012. Clarkian logic on trial. Perspectives on Psychological Science 7, 260–263.
Orbach, Y., Hershkowitz, I., Lamb, M.E., Sternberg, K.J., Esplin, P.W., Horowitz, D., 2000. Assessing the value of structured protocols for forensic interviews of alleged abuse victims. Child Abuse & Neglect 24, 733–752.
Paterson, H.M., Kemp, R.I., 2006. Co-witnesses talk: a survey of eyewitness discussion. Psychology, Crime & Law 12, 181–191.
Pickel, K.L., 1999. The influence of context on the "weapon focus" effect. Law and Human Behavior 23, 299–311.
Pipe, M.E., Orbach, Y., Lamb, M., Abbott, C.B., Stewart, H., 2008. Do Best Practice Interviews with Child Abuse Victims Influence Case Processing. Unpublished report. US Department of Justice.
Sporer, S.L., Penrod, S., Read, D., Cutler, B., 1995. Choosing, confidence, and accuracy: a meta-analysis of the confidence-accuracy relation in eyewitness identification studies. Psychological Bulletin 118, 315–327.
Steblay, N.M., 1997. Social influence in eyewitness recall: a meta-analytic review of lineup instruction effects. Law and Human Behavior 21, 283–297.
Tenney, E.R., MacCoun, R.J., Spellman, B.A., Hastie, R., 2007. Calibration trumps confidence as a basis for witness credibility. Psychological Science 18, 46–50.
Vallano, J.P., Schreiber, C.N., 2011. A comfortable witness is a good witness: rapport-building and susceptibility to misinformation in an investigative mock-crime interview. Applied Cognitive Psychology 25, 960–970.
Vredeveldt, A., Penrod, S.D., 2013. Eye-closure improves memory for a witnessed event under naturalistic conditions. Psychology, Crime & Law 19, 893–905.
Wade, K.A., Green, S.L., Nash, R.A., 2010. Can fabricated evidence induce false eyewitness testimony? Applied Cognitive Psychology 24, 899–908.
Wells, G.L., Leippe, M.R., 1981. How do triers of fact infer the accuracy of eyewitness identifications? Using memory for peripheral detail can be misleading. Journal of Applied Psychology 66, 682–687.
Wells, G.L., Quinlivan, D.S., 2009. Suggestive eyewitness identification procedures and the Supreme Court's reliability test in light of eyewitness science: 30 years later. Law and Human Behavior 33, 1–24.
Wells, G.L., Small, M., Penrod, S., Malpass, R.S., Fulero, S.M., Brimacombe, C.E., 1998. Eyewitness identification procedures: recommendations for lineups and photospreads. Law and Human Behavior 22, 603–647.
Westcott, H.L., Jones, J., 1997. Perspectives on the Memorandum: Policy, Practice, and Research on Investigative Interviewing. Arena, Aldershot.
Whitley, B.E., Greenberg, M.S., 1986. The role of eyewitness confidence in juror perceptions of credibility. Journal of Applied Social Psychology 16, 387–409.
Wright, D.B., Carlucci, M.E., Evans, J.R., Schreiber Compo, N., 2010. Turning a blind eye to double blind line-ups. Applied Cognitive Psychology 24, 849–867.
Zhu, B., Chen, C., Loftus, E.F., Lin, C., He, Q., Chen, C., Dong, Q., 2010. Individual differences in false memory from misinformation: cognitive factors. Memory 18, 543–555.

Relevant Websites

www.slate.com/articles/news_and_politics/jurisprudence/features/2011/getting_it_wrong_convicting_the_innocent/how_eyewitnesses_can_send_innocents_to_jail.html – Brandon Garrett on 'How Eyewitnesses Can Send Innocents to Jail.'
www.innocenceproject.org – Innocence Project.
www.ted.com/talks/elizabeth_loftus_the_fiction_of_memory.html – 'Professor Elizabeth Loftus'; 2013 TED Talk, 'The Fiction of Memory.'
www.selfadministeredinterview.com – The Self-Administered Interview.
https://www.ncjrs.gov/pdffiles1/nij/178240.pdf – US Department of Justice's 'Eyewitness Evidence: A Guide for Law Enforcement (1999).

F

Face Perception Development During Early Childhood

Gudrun Schwarzer and Claudia Freitag, University of Giessen, Giessen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article addresses the development of face processing during the first 4 years of life. Beginning with newborns' preference for faces we describe how and when infants begin to categorize and recognize different aspects of faces like gender, attractiveness, and familiarity. We also examine the extent to which young children rely on featural and configural information when they process faces. Further on we focus on the processing of social information in faces like emotional expressions and gaze direction. We finally provide evidence that experience plays a crucial role in the development of face expertise in infants and young children.

Introduction

Faces are highly complex visual stimuli. They simultaneously convey information regarding the identity of a person and a variety of social information, such as emotions or social intentions. Face identity refers to the relatively stable visual information in a face that comprises individual facial features, such as the eyes, nose, and mouth, as well as the configuration of these features. The emotions or intentions of a person, in contrast, correspond to rather dynamic or transient visual cues, including emotional expressions or eye gaze.

Despite the complexity, adults are experts at processing faces. They are remarkably accurate and fast at recognizing previously encountered people by their faces (Sergent, 1989). They can store and recognize hundreds of faces (Bahrick et al., 1975) and can quickly decode emotional expressions and direction of gaze. It is not surprising that extensive research has been carried out in an attempt to understand the developmental origin of such an impressive ability. In everyday life when young children interact with other people they have to process the variety of face information, which challenges their perceptual and cognitive abilities. Previous research suggests that both nature and nurture play a crucial role in the development of face processing (Park et al., 2009). For example, there is evidence that infants display an innate preference for faces. When offered different visual stimuli, newborns are especially attracted to faces, preferring faces to other patterns (e.g., Mondloch et al., 1999; Valenza et al., 1996). However, other studies have demonstrated that experience also plays a significant role in the development of face processing. Early experience with specific types of faces (faces of a specific species or race) leads to lasting advantages in the processing of those faces (e.g., Pascalis et al., 2002).

The aim of this article is to review the body of evidence on the nature of the face processing system during infancy and up to the first 3–4 years of life. We first comment on infants' early preference for faces. Then, we describe the development of two different processes, the categorization of faces as belonging to a specific category, e.g., attractiveness or gender, and the recognition of a certain face on an individual level. We also examine the extent to which young children rely on featural and configural information when they process faces. Then, we review the role of social face information, and finally, we conclude with the role of experience in children's face processing.

Preference for Faces

The human face is most likely the first social stimulus with which newborns interact. They not only attend to faces, but prefer to look at faces shortly after birth. Several studies showed that newborns looked longer at schematic stimuli with face-like arrangements than at stimuli with similar complexity but without face-like configurations (Frantz, 1961; Goren et al., 1975; Johnson et al., 1991; Maurer and Young, 1983; Mondloch et al., 1999; Simion et al., 1998a,b; Turati et al., 2002). A configuration of two small quadrates symbolizing the eyes arranged above a centered quadrate symbolizing the mouth, surrounded by a circle, seemed to be essential for evoking such a preference (Johnson and Morton, 1991). Newborns also prefer to look at photos of real faces when presented upright, compared to photos of faces presented inverted (180° rotation, see Figure 1), again indicating their preference for the face-like configuration of eyes above the mouth (Macchi Cassia et al., 2004). Di Giorgio et al. (2012) reported that newborns' preferences for faces were not restricted to human faces but also applied to nonhuman primate faces. During the first weeks of life, face preference seems to narrow toward real human faces compared to schematic faces. Face preference for schematic faces was less

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23234-0

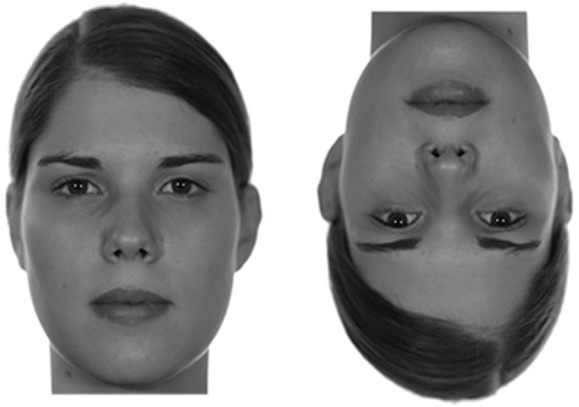

Figure 1 Example of an upright and inverted face.

pronounced at the end of the first month and was no longer existent in 3- to 5-month-old infants (Johnson et al., 1991).

The mechanisms underlying early face preference are controversial. The dual mechanism theory of Morton and Johnson (1991) presumes that newborns have an innate mechanism that includes knowledge of face structures. Such a mechanism, termed CONSPEC, triggers the infant's attention to actively search for faces during the first months of life. The information gathered by CONSPEC is processed in the primitive subcortical brain structures. Beginning at approximately 2 months of age, a second independent mechanism of face processing, termed CONLEARN, develops and processes face information in cortical brain structures. Based on increasing experience with faces, CONLEARN enables the perception of individual differences between faces and the recognition of familiar faces.

In contrast to the dual mechanism theory, infant face preference has been explained by nonspecific perceptual constraints arising from the general properties of visual processing shortly after birth (Turati, 2004). Here, it is assumed that infants are sensitive to a specific stimulus configuration, termed up-down asymmetry and prefer to look at stimuli comprising more information in the upper half than in the lower half (see Figure 2, Simion et al., 2001). This visual configuration is considered to be typical and is most often provided in faces but is not unique to faces alone (Banks and Salapatek, 1981; Simion et al., 2003). In summary, both theories presume that newborns are prepared to orient to faces shortly after birth but they differ in how this early preference for faces is explained, i.e., due to an innate representation of faces or due to less specific attentional biases.

Categorization and Recognition of Faces

Not only do infants prefer to look at faces, they also begin to categorize faces according to attractiveness, gender, or emotional expression. Categorizing in general implies the existence of a prototype that reflects the average of different exemplars within one category (Quinn and Slater, 2003). A face prototype is characterized as the internal representation of an average face based on all of the individual faces that a person has processed

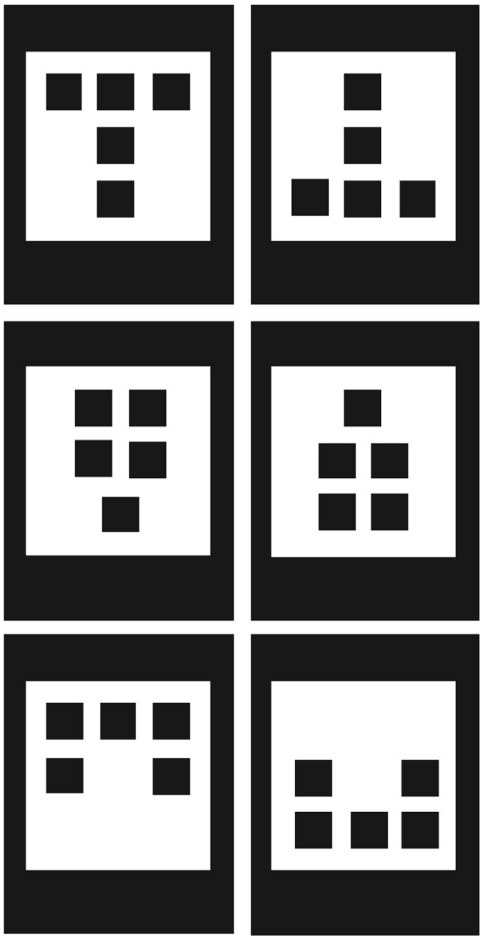

Figure 2 Non-face-like stimuli used to study infants' preference for up-down asymmetry. Simion, F., Macchi Cassia, V., Turati, C., Valenza, E., 2001. The origins of face perception: specific versus non-specific mechanisms. Infant and Child Development 10 (1–2), 59–65. Reproduced with permission.

so far (Valentine, 1991, see later). At the age of 3 months, infants are able to build face prototypes while before that age each face is stored separately with no reference to the internal representation of other faces (de Haan et al., 2001). The authors habituated 1- and 3-month-old infants to four faces and then tested them using a morphed face representing the average (the prototype) of all four faces. While those aged 3 months recognized the prototype, the 1-month-old infants did not. Thus, infants are able to form face categories at approximately 3 months of age due to their ability to build prototypes and their growing experience with faces.

The attractiveness of faces as classified by adults is a categorization that infants are sensitive to. By 6 months of age, infants prefer to look at attractive faces compared to faces classified as less attractive, independent of the ethnicity, gender, and age of the faces (Langlois et al., 1991). Such an early preference can be due to the fact that infant face prototypes resemble adult face prototypes, as infants mostly experience adult faces, and that attractive faces are almost identical to the face prototypes of the relevant gender in adults.

Another facial category that infants learn to discriminate is gender. The majority of 3-month-old infants tested in a preferential-looking paradigm preferred to look at female faces, compared to male faces (Quinn et al., 2002). The authors explained this result with the observation that many of the infants were raised by female caregivers, leading to greater levels of experience with female faces. This hypothesis was supported by the finding that infants raised by male caregivers preferred to look at male faces.

The ability to categorize emotional facial expressions arises at approximately 7 months of age (Nelson et al., 2006). At that age, infants begin to discriminate emotional expressions in faces of different ages and genders (Bornstein and Arterberry, 2003; Nelson, 2001), indicating that they are able to generalize emotional expressions across multiple exemplars of faces. Nelson et al. (2006) demonstrated that happiness was the first emotional expression that was recognized in different faces, followed by the expression of fear at the end of the first year. At the same time as infants begin to categorize emotional expressions they also become sensitive to the intensity of an emotional facial expression and are able to categorize mild and strong expressions of happiness and anger (Striano et al., 2002).

The skill needed to discriminate distinct emotional expressions within an identical face, which is more basic than that needed to categorize facial expressions, arises earlier, at approximately 3 months of age. Based on studies using a habituation paradigm, 3-month-old infants were able to discriminate between happiness and surprise within an identical face but not between happiness and sadness (Young-Browne et al., 1977). It seems that familiarity facilitates the discrimination of emotional expressions in young infants, as 3.5-month-old infants were able to discriminate between expressions of happiness and anger in their mothers' faces but not in unfamiliar female faces (Kahana-Kalman and Walker-Andrews, 2001). Some weeks later, at 4–5 months of age, infants seem to be able to discriminate between happiness, anger, and neutral expressions (La Barbera et al., 1976), as well as between sadness and fear within identical unfamiliar faces (Schwartz et al., 1985). Beyond the first year of life, infants and toddlers become more and more efficient at recognizing emotional expressions in different poses (Bruce et al., 2000) based on inner facial features only (Mondloch et al., 2003). Three year olds are almost error-free in recognizing basic emotions, with the highest accuracy in recognizing happiness (Strand et al., 2008), while the categorization of more complex emotional facial expressions, such as shame and abashment, were not achieved until the middle childhood years (Markham and Adams, 1992).

Familiarity of faces is another crucial category that infants have to learn, as it is essential for emotional binding and social communication to be able to discriminate between familiar and unfamiliar faces and to recognize personally familiar faces. It is not surprising that the caregiver's face is the first to be recognized by infants. At 2–36 h after birth, newborns' sucking responses were higher when they saw a photo of their mother's face compared to a photo of an unfamiliar female face (Walton et al., 1992). When mothers were present but did not speak, 4-day-old newborns looked longer to their mother's face than to a stranger's face (Field et al., 1984). Similar results were found in 49 h olds, even when the individual smell of the mother was masked by a flavor (Bushnell et al., 1989). Pascalis et al. (1995) replicated those findings and, in addition, found that the preference for the mother's face disappeared when the outer facial features, such as the hair and chin contour, were covered by a scarf. Such a dominance of the outer facial features in recognizing mothers' faces was replicated in 19- to 25-day-old infants (Bartrip et al., 2001), while 35- to 40-day-old infants recognized their mothers' face when presented with inner features only.

One important requirement for recognizing the mother's face shortly after birth seems to be the simultaneous perception of the mother's voice and face. Sai (2005) reported that only newborns who experienced their mother's face and voice together recognized the mother's face 3–7 h after birth, while infants who saw their mother's face but did not hear her voice could not discriminate between their mother's face and an unfamiliar female face. These results suggest that experience with the mother's voice during pregnancy primes the infant's recognition of the mother's face and points to intermodal learning in very young infants.

What about the recognition of other faces? Early competency in face recognition has been observed in 4-day-old newborns. Infants were habituated to one face, and recognition was tested immediately after habituation and after a delay of 2 min by presenting the habituation face together with an unfamiliar face. At both testing times, the infants looked longer at the unfamiliar face, indicating recognition of the habituated face, immediately and after a 2-min delay (Pascalis and de Schonen, 1994). A more demanding challenge within face recognition is the ability to recognize faces in varied poses. To investigate this skill, 3 and 6 month olds were habituated to a single face in different poses that also varied in emotional expression. After 2 min and 24 h, the 6 month olds and the male 3 month olds recognized the habituation face despite the variations in poses and emotional expressions (Pascalis et al., 1998). Based on the results from behavioral studies and neurophysiological findings, Nelson (2003) concluded that face recognition is relatively robust at the age of 6 months. Despite the impressive face recognition abilities of infants, it is recognized that the face processing system undergoes development during the first decade of life (Schwarzer et al., 2003). Face recognition during the transition from infancy to childhood has been less investigated. The few existing studies show that when young children were tested with a small number of faces they show high recognition accuracy; e.g., 2 year olds made only few mistakes in recognizing a single face across variations in poses and emotional expressions (Brace et al., 2001). Similar findings were made in 3 year olds using two target faces (Freitag and Schwarzer, 2011).

Featural and Configural Processing

Whereas most investigators agree on the definition of featural processing of faces in terms of a concentration of a single facial feature, configural processing can be divided into three types (see Maurer et al., 2002): (1) sensitivity to first-order relations, i.e., identifying that a stimulus is a face because the features are

arranged with two eyes above a nose, which in turn is located above a mouth; (2) holistic processing, i.e., gluing together the features into a gestalt; and (3) sensitivity to second-order relations, i.e., perceiving specific distances among the features. In the following paragraphs, we describe the extent to which infants and young children show these three types of configural processing.

In principle, all studies demonstrating young infants' preferences for face-like stimuli indicate their sensitivity to first-order relational information in faces (see earlier text). For example, Simion et al. (1998a,b) reported that newborns could discriminate between schematic faces that consisted of small elements arranged in either a face-like or a non-face-like configuration and that they tended to prefer the face-like pattern. Thus, previous research indicates that sensitivity to first-order relations in faces exists from the beginning of life.

The extent to which infants process faces holistically as a gestalt was first studied by Cohen and Cashon (2001) and Cashon and Cohen (2004). Using a so-called switch design, the authors contrasted holistic processing with featural processing. First, they habituated the infants to two adult female faces. Then, the infants were tested with a familiar habituation face, a 'switch-face,' and a novel face. The switch-face was a composite of the two habituation faces, consisting of the internal features or a subset of internal features of one face and the external features of the other face. The faces in the habituation and test phases were presented either upright or inverted. To assess whether the infants used featural or holistic processing, it was crucial to determine whether they looked longer at the switch-face or the familiar habituation face. The authors reasoned that the infants would look longer at the switch-face only if they were processing more than the independent features and were integrating at least some of the internal and external features. However, if they were processing the facial sections independently, the infants would not respond to the switch-face. The results showed that holistic processing seemed to follow an N-shaped developmental pattern for upright faces and an inverted U-shaped pattern for inverted faces across an age range of 3–7 months. This means that the youngest age group, the 3 month olds, processed upright and inverted internal and external sections of faces independently of each other, i.e., by feature. Recently, however, Turati et al. (2010) investigated holistic processing of the top and bottom halves of faces in 3 month olds using a modified version of the composite face task. They showed that differentiation between the two top halves of faces was impaired when the halves were combined with new bottom halves in an aligned condition, compared to a condition in which the face halves were misaligned. Those results can be interpreted as an indication of holistic face processing of the top and bottom halves of faces. Thus, at 3 months of age, holistic processing of faces seems to be dependent on which facial sections (internal/external, top/bottom) are taken under consideration. In the study by Cashon and Cohen (2004), the 4 month olds processed both upright and inverted internal/external sections holistically, while the 6 month olds processed both upright and inverted external and internal sections independently of each other; i.e., featurally. Finally, the 7 month olds processed the upright internal and external facial sections holistically and the inverted sections featurally. Schwarzer et al. (2007) and Schwarzer and Zauner (2003) followed up the Cashon and Cohen study and investigated which parts of a face create a switch effect. They compared changes in the eyes or the mouth with changes in both the eyes and the mouth. The authors found that 8- and 10-month-old infants processed all of the changes in the face holistically, whereas 6 month olds processed eyes as features but integrated the mouth into the whole face. Four month olds processed both the eyes and the mouth as features. Thus, while it is assumed that 7- to 8-month-old infants process faces holistically, the results from younger infants are mixed. The majority of the results suggest that young infants tend to process faces by feature, rather than as a whole.

Holistic processing has also been demonstrated in older children (4 year olds) using a part-whole task (Pellicano and Rhodes, 2003). In this task, children learned to identify a set of faces and were then asked to identify single features separately and in the context of the whole face. They identified the features more easily and more correctly in the whole-face condition than in the separate condition, indicating holistic processing. De Heering et al. (2007) and Macchi Cassia et al. (2009) showed holistic processing in 3- and 4-year-old children using the composite task. Children more quickly and correctly recognized the top half of a face when presented with a misaligned bottom half, but were slower to recognize the top half of a face when the top half was aligned with the bottom half of a different face. However, when 2- to 5-year-old children were asked to categorize faces they did not categorize them holistically but, instead, categorized them by single features (Schwarzer, 2002). Thus, young children are sensitive to holistic changes in a face but are still unable to explicitly use holistic face information to categorize faces.

There are also studies demonstrating infants' sensitivity to second-order relations with the regard to perceiving specific distances among the features. Bhatt et al. (2005) habituated 5-month-old infants to a face in which the eyes were moved slightly up and the mouth was moved slightly down. After habituation, the infants looked longer at the unaltered face and failed to discriminate the faces when they were inverted. Similarly, Thompson et al. (2001) and Bertin and Bhatt (2004) provided evidence of infant sensitivity to second-order relations at 6 and 7 months of age. Leo and Simion (2009) showed that even newborns were sensitive to specific changes in a face. They demonstrated that newborns could discriminate upright Thatcher faces from non-Thatcher faces, although those faces differed only in their second-order configural information.

Pellicano et al. (2006) provided evidence of sensitivity in young children to the second-order relations of a face at approximately 4 years of age using the part-whole task. They found that children were better able to recognize facial features with the same configuration than corresponding facial features with new spacing, indicating sensitivity to changes in internal facial features. Macchi Cassia et al. (2011) revealed signs of sensitivity to second-order relations in 3-year-old children. They tested sensitivity to variations in the spacing of features in faces and non-face objects using

a two-alternative forced choice matching-to-sample task. Three year olds performed above chance on the face and non-face objects discrimination task but only when tested with larger spacing changes in a task that required no memory demands.

Taken together, previous studies on infant face processing suggest that young infants process all three types of configural information of faces. Sensitivity to first-order relations remains functional throughout development, whereas holistic processing and sensitivity to second-order relations only develops during the second half of the first year. The extent to which holistic processing and sensitivity to second-order relations are used in young childhood strongly depends on the demands of the tasks.

Role of Social Face Information in Face Processing

Faces comprise a variety of social information, including gender, age, emotional expression, and eye contact. Emotional facial expressions and gaze direction are highly relevant for nonverbal social communication in each age group and in infants in particular. Bowlby (1969), for example, reported that a smile on a caregiver's face induces the expression of joy in infants and, in turn, enhances the caring behavior of the adult. Particularly in situations that are unpredictable to infants, infants make use of their caregivers' emotional expressions to learn about a potential threat and to regulate their behavior in relation to the caregivers' emotional facial expression (e.g., Sorce et al., 1985).

The processing of emotional expression does not only impact infants' social communication, but also seems to influence infant processing of facial identity. For infants and young children, results indicate that emotional facial expressions can enhance face recognition (Pascalis et al., 1998; Turati et al., 2004). Seven-month-old infants who were habituated to a frontal view of a face recognized the habituation face in a 3/4-pose only when it was presented with a smile or anger expression but not when it was presented with a neutral expression (Gross and Schwarzer, 2010). The influence of emotional expression on face recognition was also shown in 3 year olds who were more proficient at recognizing faces in varied emotional expressions when the faces were encoded with smile or anger expressions compared to neutral expressions (Freitag and Schwarzer, 2011). These results suggest that emotional facial expressions enrich a face, facilitating the processing of face identity and making face recognition more robust against variations in poses or emotions. This interpretation is supported by neurophysiological data indicating that faces expressing emotions activate more complex and interactive brain areas compared to faces with neutral expressions (Vuilleumier and Pourtois, 2007).

Gaze direction and eye contact are also crucial for social communication (see Frischen et al., 2007). For infants, it is an important developmental goal to attend to the gaze of their interaction partners and to learn to coordinate their own visual attention with that of others. The ability to focus with an interaction partner on the same object at the same time enables shared attention (Tomasello, 1999), which can be seen as a general principle for social learning. It was reported that newborns tested 37 h after birth prefer to look at faces with eyes open compared to eyes closed (Batki et al., 2000). Newborns are also sensitive to the gaze direction of other people and prefer to look at faces with direct gaze (Farroni et al., 2006) compared to averted gaze. The finding that infants discriminate gaze direction in others is supported by the result that the brain activity of 4 month olds was substantially different for faces with eye contact compared to faces with averted gaze (Johnson and Farroni, 2003). At 3 months of age, infants begin to follow the eye movements of a social agent with their own eyes. While young infants gaze following is limited to situations in which the social agent's gaze is directed to objects very close to the infant, gaze following occurs with more distant objects at approximately 6 months of age (Butterworth and Itakura, 2000). The ability to coordinate gaze with that of others becomes robust at the end of the first year (Farroni et al., 2002).

The extent to which infants and toddlers follow the eye movements of others is also influenced by situational aspects. For example, it was shown that infants respond more often to the eye movements of others when they had previously been interacting directly (Collicott et al., 2008; Johnson et al., 1998). Another example showed that young children focused on the eye movements of others when they had to find an object that was hidden from their own perspective (Moll and Tomasello, 2004).

Role of Experience in Face Processing

Previous research has demonstrated that experience with faces plays a crucial role in young children's processing of faces. In the following, we elucidate how early visual experience shapes the functionality of the face processing system and how the specific faces to which infants are exposed refine the recognition of faces.

Studies with patients deprived of early visual experience by bilateral cataracts showed that visual experience during the first weeks and months of life is critical for the normal development of some, but not all, aspects of face processing (see Mondloch et al., 2010, for an overview). For example, infants treated for congenital cataracts after 6 weeks of age showed a pattern of face preferences on the day that they could first see similar to that of newborns. This result indicates that the postnatal sensitivity to first-order relations (see earlier text) is driven by visual input but does not require early visual input (Mondloch et al., unpublished data, cited by Mondloch et al., 2009). However, such an early visual deprivation was also associated with permanent deficits in the later development of face processing, such as in the holistic processing of faces (Le Grand et al., 2004) and in the processing of second-order relations in faces. For example, adult patients deprived of early visual experience had difficulty distinguishing faces that differed only in the spacing of the features (Le Grand et al., 2001) and in the recognition of faces from a novel point of view (Geldart et al., 2002). In contrast, later development of the processing of facial features was not impaired. They were able to distinguish faces that differed only in the shape of

individual features or in the contour of the external face (Le Grand et al., 2001).

The role of experience in the development of face processing is also evident when exploring how experience with individual faces or certain face categories affects the recognition of faces in children with normal early visual experience (see Pascalis et al., 2011, for an overview). Most of the studies have focused on the role of the infants' specific experience with faces on the recognition of faces belonging to their own or another species or race. With regard to species, Pascalis et al. (2002) demonstrated that 6 month olds could recognize different human and monkey faces, whereas memory for individual monkey faces disappeared at 9 months of age. However, this loss was attenuated when infants were trained to recognize monkey faces (Pascalis et al., 2005). Scott and Monesson (2009) showed that not only experience with monkey faces but also the development of exemplar-based individual memory could prevent such a loss. They trained 6–9 month olds in three different conditions. One group was trained at the individual level (each monkey face was individually labeled), a second group was trained at the categorical level (each monkey face was labeled 'monkey'), and a third group was shown monkey faces without labels. Only the infants trained at the individual level maintained the ability to discriminate among the monkey faces at 9 months of age. Infants who were otherwise trained lost the ability.

With regard to race, similar results on the effects of experience have been revealed. Whereas 3-month-old Caucasians discriminated between faces from four ethnic groups (African, Middle Eastern, Chinese, and Caucasian), 6-month-old Caucasians discriminated only between Chinese and Caucasian faces and 9-month-old Caucasians did so only within Caucasian faces (Kelly et al., 2007). Kelly et al. (2009) found a similar developmental course for 3-, 6-, and 9-month-old Chinese infants when presented with Chinese, Caucasian, and African faces. Recently, infants' specific experiences with other races have been shown to reverse this other-race effect. Heron-Delaney et al. (2011) demonstrated that exposing Caucasian infants to Chinese faces through perceptual training with picture books for 1 h between 6 and 9 months of age enabled the Caucasian infants to maintain the ability to discriminate Chinese faces at 9 months of age. A recent study by Spangler et al. (2012) produced similar results. Here, the other-race effect of Caucasian infants at 3, 6, and 9 months of age was investigated longitudinally using Caucasian and African faces. The authors found a similar ability to differentiate African faces at all ages, even at 9 months of age, whereas another sample of 9-month-old Caucasians were not able to differentiate African faces. The results were interpreted to mean that repeated exposure to African faces in the longitudinal sample at 3 and 6 months seemed to provide sufficient experience with African faces to prevent the onset of the other-race effect at 9 months of age.

Thus, it can be concluded that, according to Nelson's (2001) concept of perceptual narrowing, training or experience with faces of other races or other species slows or modifies the fine-tuning toward faces of one's own race or species. Such fine-tuning, occurring with increasing experience, can be related to the face model of Valentine (1991). He proposed that faces are encoded and represented in a multidimensional cognitive map, a perceptual face space. In this space, individual faces are coded relative to the average of previously encountered faces. The distance among faces represents their perceived similarity. Typical faces are represented close to each other, while distinctive faces are more spread out in the face space. Pascalis et al. (2011) assumed that face spaces are tuned toward faces present in the environment and are less effective at processing faces from different ethnic groups or different species. The authors suppose, according to Furl et al. (2002), that frequent exposure to a certain face category, such as a certain race or species, could cause a warp in the face space that enhances the differences in faces within a familiar and often experienced face category, leading to an advantage in differentiating such faces. At the same time, this space limits the quality of the face representations for other-race or other-species faces, leading to low differentiation between them.

Whereas a huge amount of research has focused on the role of experience in infant face processing, far less data are available for the following years of childhood. One of the few studies, for example, Sangrigoli and de Schonen (2004), demonstrated the other-race effect in 3 year olds. However, this study also showed that the effect was highly fragile at that age. When faces were presented for 1000 ms, the other-race effect appeared, but when the presentation time was reduced to 500 ms the effect disappeared. This might be because developing face processing abilities are highly intertwined with concurrently developing general cognitive abilities at that age. This can complicate the assessment of face processing abilities and can produce inconsistent results. Future research needs to examine the role of experience in face processing while considering the general cognitive abilities of children.

See also: Early Emotional Development and Cultural Variability; Emotion in Cognition; Face Recognition Models: Computational Approaches; Face Recognition, Psychological and Neural Aspects; Facial Emotion Expression, Individual Differences In; Infant Gaze Behaviors; Visual Memory, Psychology of; Visual Perception, Neural Basis of.

Bibliography

Bahrick, H.P., Bahrick, P.O., Wittlinger, R.P., 1975. Fifty years of memory for names and faces: a crosssectional approach. Journal of Experimental Psychology: General 104, 54–75.

Banks, M., Salapatek, P., 1981. Infant pattern vision: a new approach based on the contrast sensitivity function. Journal of Experimental Child Psychology 31, 1–45.

Bartrip, J., Morton, J., de Schonen, S., 2001. Responses to mother's face in 3-week to 5-month-old infants. British Journal of Developmental Psychology 19, 219–232.

Batki, A., Baron-Cohen, S., Wheelwright, S., Connellan, J., Ahluwalia, J., 2000. Is there an innate gaze module? Evidence from human neonates. Infant Behavior and Development 23, 223–229.

Bertin, E., Bhatt, R.S., 2004. The Thatcher illusion and face processing in infancy. Developmental Science 7 (4), 431–436.

Bhatt, R.S., Bertin, E., Hayden, A., Reed, A., 2005. Face processing in infancy: developmental changes. Child Development 76 (1), 169–181.

Bornstein, M.H., Arterberry, M.E., 2003. Recognition, discrimination and categorization of smiling by 5-month-old infants. Developmental Science 6 (5), 585–599.

Bowlby, J., 1969. Attachment and Loss. In: Attachment, vol. 1. Basic Books, New York.

Brace, N.A., Hole, G.J., Kemp, R.I., Pike, G.E., Van Duuren, M., Norgate, L., 2001. Developmental changes in the effect of inversion: using a picture book to investigate face recognition. Perception 30, 85–94.

Bruce, V., Campbell, R.N., Doherty-Sneddon, G., Import, A., Langton, S., McAuley, S., Wright, R., 2000. Testing face processing skills in children. British Journal of Developmental Psychology 18, 319–333.

Bushnell, I.W.R., Sai, F., Mullin, J.T., 1989. Neonatal recognition of the mother's face. British Journal of Developmental Psychology 7, 3–15.

Butterworth, G., Itakura, S., 2000. How the eyes, head, and hand serve definite reference. British Journal of Developmental Psychology 18, 25–50.

Cashon, H.C., Cohen, L.B., 2004. Beyond u-shaped development in infants' processing of faces: an information-processing account. Journal of Cognition and Development 5 (1), 59–80.

Cohen, L.B., Cashon, C.H., 2001. Do 7-month-old infants process independent features or facial configurations? Infant and Child Development 10, 83–92.

Collicott, C., Collins, S., Moore, C., 2008. The development of gaze following in a third party context. Infancy 14 (3), 363–376.

de Haan, M., Johnson, M.H., Maurer, D., Perrett, D.I., 2001. Recognition of individual faces and average face prototypes by 1- and 3-month-old infants. Cognitive Development 16, 659–678.

De Heering, A., Houthuys, S., Rossion, B., 2007. Holistic face processing is mature at 4 years of age: evidence from the composite effect. Journal of Experimental Child Psychology 96, 57–70.

Di Giorgio, E., Leo, I., Pascalis, O., Simion, F., 2012. Is the face-perception system human-specific at birth? Developmental Psychology 48 (4), 1083–1090.

Farroni, T., Johnson, M.H., Brockbank, M., Simion, F., 2002. Infants' use of gaze direction to cue attention: the importance of perceived motion. Visual Cognition 7, 705–718.

Farroni, T., Menon, E., Johnson, M.H., 2006. Factors influencing newborns' preference for faces with eye contact. Journal of Experimental Child Psychology 95, 298–308.

Field, T.M., Cohen, D., Garcia, R., Greenberg, R., 1984. Mother-stranger discrimination by the newborn. Infant Behavior and Development 7, 19–25.

Frantz, R.L., 1961. The origin of form perception. Scientific American 204, 66–72.

Freitag, C., Schwarzer, G., 2011. Influence of emotional facial expressions on 3–5-year-olds' face recognition. Cognitive Development 26, 230–247.

Frischen, A., Bayliss, A.P., Tipper, S.P., 2007. Gaze cueing of attention: visual attention, social cognition, and individual differences. Psychological Bulletin 133, 694–724.

Furl, N., Phillips, P.J., O'Toole, A.J., 2002. Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. Cognitive Science 26, 797–815.

Geldart, S., Mondloch, C.J., Maurer, D., de Schonen, S., Brent, H., 2002. The effect of early visual deprivation on the development of face processing. Developmental Science 5, 490–501.

Goren, C., Sarty, M., Wu, P.Y.K., 1975. Visual following and pattern discrimination of face-like stimuli by newborn infants. Pediatrics 56, 544–549.

Gross, C., Schwarzer, G., 2010. Face recognition across varying poses in seven- and nine-month-old infants: the role of facial expression. International Journal of Behavioral Development 34 (5), 417–426.

Heron-Delaney, M., Anzures, G., Herbert, J.S., Quinn, P.C., Slater, A., Tanaka, J.W., Lee, K., Pascalis, O., 2011. Perceptual training prevents the emergence of the other race effect during infancy. Plos One 6 (5), e19858. http://dx.doi.org/10.1371/journal.pone.0019858.

Johnson, M.H., Dziurawiec, S., Ellis, H., Morton, J., 1991. Newborns' preferential tracking of face-like stimuli and its subsequent decline. Cognition 40, 1–19.

Johnson, M.H., Farroni, T., 2003. Perceiving and acting on the eyes: the development and neural basis of eye gaze perception. In: Pascalis, O., Slater, A. (Eds.), The Development of Face Processing in Infancy and Early Childhood. Nova Sciences Publishers, Inc., New York, pp. 155–167.

Johnson, M.H., Morton, J., 1991. Biology and Cognitive Development: The Case of Face Recognition. Blackwell, Oxford, UK.

Johnson, S.C., Slaughter, V., Carey, S., 1998. Whose gaze will infants follow? Features that elicit gaze following in 12-month-olds. Developmental Science 1, 233–238.

Kahana-Kalman, R., Walker-Andrews, A.S., 2001. The role of person familiarity in young infants' perception of emotional expressions. Child Development 72, 352–369.

Kelly, D.J., Liu, S., Lee, K., Quinn, P.C., Pascalis, O., Slater, A.M., et al., 2009. Development of the other-race effect during infancy: evidence toward universality? Journal of Experimental Child Psychology 104, 105–114.

Kelly, D.J., Quinn, P.C., Slater, A.M., Lee, K., Ge, L., Pascalis, O., 2007. The other-race effect develops during infancy: evidence of perceptual narrowing. Psychological Science 18, 1084–1089.

La Barbera, J.D., Izard, C.E., Vietze, P., Parisi, S.A., 1976. Four- and six- month-old infants' responses to joy, anger, and neutral expressions. Child Development 47, 535–538.

Langlois, J.H., Ritter, J.M., Roggman, L.A., Vaughn, L.S., 1991. Facial diversity and infant preferences for attractive faces. Developmental Psychology 27, 79–84.

Le Grand, R., Mondloch, C.J., Maurer, D., Brent, H.P., 2001. Early visual experience and face processing. Nature 410, 890.

Le Grand, R., Mondloch, C.J., Maurer, D., Brent, H.P., 2004. Impairment in holistic face processing following early visual deprivation. Psychological Science 15, 762–768.

Leo, I., Simion, F., 2009. Face processing at birth: a Thatcher illusion study. Developmental Science 12, 492–498.

Macchi Cassia, V., Picozzi, M., Kuefner, D., Bricolo, E., Turati, C., 2009. Holistic processing for faces and cars in preschool-aged children and adults: evidence from the composite effect. Developmental Science 12, 236–248. http://dx.doi.org/10.1111/j.1467–7687.2008.00765.x.

Macchi Cassia, V., Turati, C., Schwarzer, G., 2011. Sensitivity to spacing changes in faces and nonface objects in preschool-aged children and adults. Journal of Experimental Child Psychology 109, 454–467.

Macchi Cassia, V., Turati, C., Simion, F., 2004. Can a non-specific bias toward top-heavy patterns explain newborns' face preference? Psychological Science 15 (6), 379–383.

Markham, R., Adams, K., 1992. The effect of type of task on children's identification of facial expressions. Journal of Nonverbal Behavior 16, 21–39.

Maurer, D., Le Grand, R., Mondloch, C.J., 2002. The many faces of configural processing. Trends in Cognitive Science 6, 255–260.

Maurer, D., Young, R., 1983. Newborns' following of natural and distorted arrangements of facial features. Infant Behaviour and Development 6, 127–131.

Moll, H., Tomasello, M., 2004. 12- and 18-month-old infant follow gaze to spaces behind barriers. Developmental Science 7, 1–9.

Mondloch, C.J., Geldart, S., Maurer, D., Le Grand, R., 2003. Developmental changes in face processing skills. Journal of Experimental Child Psychology 86, 67–84.

Mondloch, C.J., Le Grand, R., Maurer, D., 2010. Development of expertise in face recognition. In: Gauthier, I., Tarr, M.J., Bub, D. (Eds.), Perceptual Expertise: Bridging Brain and Behavior. Oxford University Press, Oxford, pp. 67–106.

Mondloch, C.J., Lewis, T.L., Budreau, D.R., Maurer, D., Dannemiller, J.L., Stephens, B.R., et al., 1999. Face perception during early infancy. Psychological Science 10 (5), 419–422.

Morton, J., Johnson, M.H., 1991. CONSPEC and CONLERN: a two-process theory of infant face recognition. Psychological Review 98 (2), 164–181.

Nelson, C.A., 2001. The development and neural bases of face recognition. Infant and Child Development 10 (1–2), 3–18.

Nelson, C.A., 2003. The development of face recognition reflects an experience-expectant and activity-dependent process. In: Pascalis, O., Slater, A. (Eds.), The Development of Face Processing in Infancy and Early Childhood: Current Perspectives. Nova Science Publishers, New York, pp. 79–97.

Nelson, C.A., Parker, S.W., Guthrie, D., 2006. The discrimination of facial expressions by typically developing infants and toddlers and those experiencing early institutional care. Infant Behavior and Development 29, 210–219.

Park, J., Newman, L.I., Polk, T.A., 2009. Face processing: the interplay of nature and nurture. Neuroscience Update 15 (5), 445–449.

Pascalis, O., de Haan, M., Nelson, C.A., 2002. Is face processing species-specific during the first year of life? Science 296, 1321–1323.

Pascalis, O., de Haan, M., Nelson, C., de Schonen, S., 1998. Long-term recognition memory for faces assessed by visual paired comparison in 3- and 6-month-old infants. Journal of Experimental Psychology: Learning, Memory and Cognition 24, 249–260.

Pascalis, O., De Martin de Viviès, X., Anzures, G., Quinn, P.C., Slater, A.M., Tanaka, J.W., Lee, K., 2011. Development of face processing. Advanced Review 2, 666–675.

Pascalis, O., de Schonen, S., 1994. Recognition memory in 3–4-day-old human infants. NeuroReport 5, 1721–1724.

Pascalis, O., de Schonen, S., Morton, J., Deruelle, C., Fabre-Grenet, M., 1995. Mother's face recognition in neonates: a replication and an extension. Infant Behavior and Development 18, 79–86.

Pascalis, O., Scott, L.S., Kelly, D.J., Shannon, R.W., Nicholson, E., Coleman, M., Nelson, C.A., 2005. Plasticity of face processing in infancy. Proceedings of the National Academy of Sciences 102 (14), 5297–5300.

Pellicano, E., Rhodes, G., 2003. Holistic processing of faces in preschool children and adults. Psychological Science 14 (6), 618–622.

Pellicano, E., Rhodes, G., Peters, M., 2006. Are preschoolers sensitive to configural information in faces? Developmental Science 9 (3), 270–277.

Quinn, P.C., Slater, A., 2003. Face perception at birth and beyond. In: Pascalis, O., Slater, A. (Eds.), The Development of Face Processing in Infancy and Early Childhood: Current Perspectives. Nova Science, New York, pp. 3–12.

Quinn, P.C., Yahr, J., Kuhn, A., Slater, A.M., Pascalis, O., 2002. Representation of the gender of human faces by infants: a preference for female. Perception 31, 1109–1121.

Sai, F.Z., 2005. The role of the mother's voice in developing mother's face preference: evidence for intermodal perception at birth. Infant and Child Development 14 (1), 29–50.

Sangrigoli, S., de Schonen, S., 2004. Effects of visual experience on face processing: a developmental study of inversion and non-native effects. Developmental Science 7, 74–87.

Schwartz, G., Izard, C., Ansul, S., 1985. The 5-month-old's ability to discriminate facial expressions of emotions. Infant Behavior and Development 8, 65–77.

Schwarzer, G., 2002. Processing of facial and non-facial visual stimuli in 2–5-year-old children. Infant and Child Development 11, 253–269.

Schwarzer, G., Zauner, N., 2003. Face processing in 8-month-old infants: evidence for configural and analytical processing. Vision Research 43, 2783–2793.

Schwarzer, G., Zauner, N., Jovanovic, B., 2007. Evidence of a shift from featural to configural face processing in infancy. Developmental Science 10 (4), 452–463.

Schwarzer, G., Zauner, N., Korell, M., 2003. Face processing during the first decade of life. In: Schwarzer, G., Leder, H. (Eds.), The Development of Face Processing. Hogrefe & Huber, Cambridge, MA.

Scott, L.S., Monesson, A., 2009. The origin of biases in face perception. Psychological Science 20, 676–680.

Sergent, J., 1989. Structural processing of faces. In: Young, A.W., Ellis, H. (Eds.), Handbook of Research on Face Processing. Elsevier Science, Amsterdam, pp. 57–91.

Simion, F., Macchi Cassia, V., Turati, C., Valenza, E., 2001. The origins of face perception: specific versus non-specific mechanisms. Infant and Child Development 10 (1–2), 59–65.

Simion, F., Macchi Cassia, V., Turati, C., Valenza, E., 2003. Non-specific perceptual biases at the origins of face processing. In: Pascalis, O., Slater, A. (Eds.), The Development of Face Processing in Infancy and Childhood: Current Perspectives. Nova Publishers, New York, pp. 13–25.

Simion, F., Valenza, E., Umiltà, C., 1998a. Mechanisms underlying face preference at birth. In: Simion, F., Butterworth, G. (Eds.), The Development of Sensory, Motor and Cognitive Capacities in Early Infancy: From Perception to Cognition. Psychology Press, Hove, pp. 87–101.

Simion, F., Valenza, E., Umiltà, C., Dalla Barba, B., 1998b. Preferential orienting to faces in newborns: a temporal-nasal asymmetry. Journal of Experimental Psychology: Human Perception and Performance 24 (5), 1399–1405.

Sorce, J.F., Emde, R.N., Campos, J., Klinnert, M.D., 1985. Maternal emotional signaling: its effect on the visual cliff behavior of 1-year-olds. Developmental Psychiatry 21, 195–200.

Spangler, S.M., Schwarzer, G., Freitag, C., Vierhaus, M., Teubert, M., Fassbender, I., Lohaus, A., Kolling, T., Graf, F., Goertz, C., Knopf, M., Lamm, B., Keller, H., 2012. The other-race effect in a longitudinal sample of 3-, 6-, and 9-month-old infants: evidence of a training effect. Infancy. http://dx.doi.org/10.1111/j.1532-7078.2012.00137.x.

Strand, P.S., Cerna, S., Downs, A., 2008. Shyness and emotion-processing skills in preschoolers: a 6-month longitudinal study. Infant and Child Development 17 (2), 109–120.

Striano, T., Brennan, P., Vanman, E.J., 2002. Maternal depressive symptoms and 6-month-old infants' sensitivity to facial expressions. Infancy 3 (1), 115–126.

Thompson, L.A., Madrid, V., Westbrook, S., Johnston, V., 2001. Infants attend to second-order relational properties of faces. Psychonomic Bulletin & Review 8, 769–777.

Tomasello, M., 1999. The Cultural Origins of Human Cognition. Harvard University Press, Cambridge, MA.

Turati, C., 2004. Why faces are not special to newborns: an alternative account of the face preference. Current Directions in Psychological Science 13, 5–8.

Turati, C., Di Giorgio, E., Bardi, L., Simion, F., 2010. Holistic face processing in newborns, 3-month-old infants, and adults. Child Development 88 (6), 1894–1905.

Turati, C., Sangrioli, S., Ruel, J., de Schonen, S., 2004. Evidence of the face inversion effect in 4-month-old infants. Infancy 6, 275–297.

Turati, C., Simion, F., Milani, I., Umiltà, C., 2002. Newborns preference for faces: what is crucial? Developmental Psychology 38, 875–882.

Valentine, T., 1991. A unified account of the effects of distinctiveness, inversion, and race in face recognition. Quarterly Journal of Experimental Psychology 43A (2), 161–204.

Valenza, E., Simion, F., Macchi Cassia, V., Umiltà, C., 1996. Face preference at birth. Journal of Experimental Psychology: Human Perception and Performance 27, 892–903.

Vuilleumier, P., Pourtois, G., 2007. Distributed and interactive brain mechanisms during emotional face perception: evidence from functional neuroimaging. Neuropsychologia 45, 174–194.

Walton, G.E., Bower, N.J.A., Bower, T.G.R., 1992. Recognition of familiar faces by newborns. Infant Behavior and Development 15, 265–269.

Young-Browne, G., Rosenfeld, H.M., Horowitz, F.D., 1977. Infant discrimination of facial expressions. Child Development 48, 555–562.

Face Recognition Models: Computational Approaches

Zhen Lei and Stan Z Li, Chinese Academy of Sciences, Beijing, China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Face recognition by machine is a technique that automatically recognizes human beings from images. It is a cross topic involving techniques from pattern recognition, computer vision, image processing, and so on. This article presents an introduction to face recognition from a computational model view. The concept of face recognition is briefly introduced. This is followed by a description of four main computational modules in the process, namely face detection, face normalization, face feature extraction, and feature matching. Finally, existing challenges and future research directions in real world face recognition are discussed.

Face Recognition

Face recognition is a task that humans perform routinely and effortlessly in our daily lives. Face recognition by machine is a computational technique that automatically recognizes human beings from images. Aimed at different applications, face recognition can be roughly categorized into face identification and face verification. Face identification is to find the identity from a set of registered candidates whose stored face images are the most similar to the input face image. Face verification is used to examine whether two face images come from the same person or not. In real world applications, there are also cases that involve simultaneous face identification and verification. For example, in a video-based watch list surveillance system, the captured face image is both identified (to whom it belongs) and checked to see if it is on the watch list.

In the past several decades, face recognition has attracted much attention and has been developed rapidly. The face recognition test report in Phillips et al. (2010) showed that the state-of-the-art face recognition algorithms have boosted face recognition performance by 10–100 times better than the best performances in years 1997 and 2002. This is solid proof that face recognition techniques have made great progress. The report also indicated that face recognition with still frontal face images captured in a controlled environment has achieved excellent performance, while its performance in uncontrolled conditions is still far from satisfactory. Phillips et al. (2010) also found that some of the algorithms can already outperform human participants in recognizing faces, which is a very promising achievement for the face recognition research field.

Generally speaking, current mainstream face recognition methods consist of the following four modules (see Figure 1): face detection, face normalization, face feature extraction, and feature matching. Given an image, the face is first detected and segmented; operations of alignment and preprocessing are then applied to normalize the face image; salient face features that are useful and robust for face recognition are then extracted from the normalized face; finally, the extracted features are compared with the templates stored in face database to compute the similarity between different faces.

Face Detection and Landmark Localization

Face detection is the first step for whole face biometrics, and its accuracy greatly affects the performance of sequential operations. Regarding this issue, the algorithm proposed by Viola and Jones (2004) is probably the most successful and pioneering contribution. In their method, a cascade of Adaboost classifier with Haar-like feature is designed for face detection. The Haar-like feature can be fast computed by integral image technique. The Adaboost learning is able to select most effective features from a large feature pool to form a strong classifier. With the cascade structure (Figure 2), most nonface samples are rapidly rejected at the former stages, so that face detection process is greatly accelerated. Nowadays, not only the frontal face detection but also many more multiview (or multipose) face detection algorithms have been developed using a similar framework.

The methodology of face detection can also be applied to landmark (e.g., eyes, nose tip, and mouth) localization, which

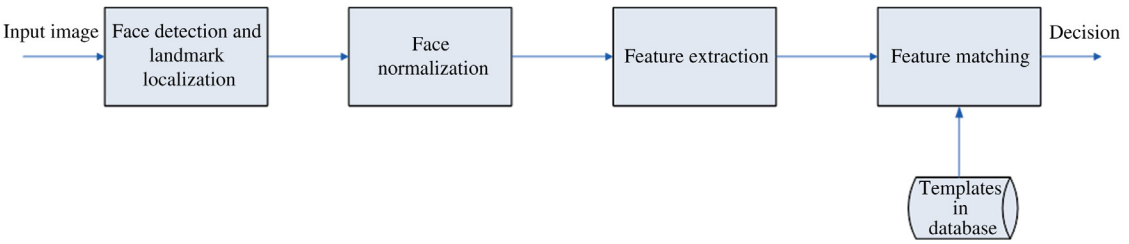

Figure 1 Face recognition flowchart.

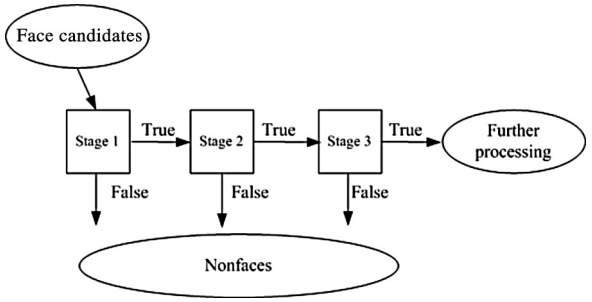

Figure 2 Cascade structure of face detection. The nonface samples are rapidly rejected by the former classifiers.

can then be utilized to face geometrical normalization. Active Shape Model (ASM) and Active Appearance Model (AAM) (Cootes et al., 1995) are also popular approaches to locating facial landmarks. In ASM, facial landmarks are first detected by searching local regions, and then these landmarks are jointly optimized by global shape constraints. In AAM, not only the shape constraint but also the texture constraint are utilized to improve the accuracy of localization.

Face Normalization

Face normalization includes face geometric and photometric normalization. The geometric normalization transforms the face image into a predefined size so the pose and scale variations can be reduced. The photometric normalization process extracts the intrinsic face information so the external illumination variation is removed. Some prevailing processing methods include histogram equalization, self-quotient image, Retinex, logarithmic total variation, and so on. Many of these methods focus on extracting the face reflectance component, which is considered as the intrinsic factor and is invariant to illumination variation.

Feature Extraction

After the preprocessing steps such as face detection and face normalization, we move to the next module to extract discriminant features for face recognition. In real application, the face images are easily affected by expression, lighting, aging, and so forth. Therefore, how to extract features that are robust to these variations is a critical issue for face recognition.

Up to now, many face representation approaches have been introduced, which can be roughly categorized as subspace-based holistic features and local appearance features (Li and Jain, 2011). Holistic methods model the face image from a global view and extract the features based on the whole image. In contrast, local descriptors capture the local structures of the face image, respectively, and then these features are combined to represent the whole face.

Face images reside in a subspace of the input image space. For example, a 60×60 grey image with 3600 pixels can express a large number of pattern classes, such as trees, cars, and faces. However, among this huge number ($256^{3600} > 10^{8100}$) of possible configurations, only a tiny fraction corresponds to faces. Therefore, there is high redundancy in the pixel-based image representation; it could be greatly reduced when only the face patterns are considered. Subspace learning is a tool that finds the embedded subspace of a specific object (e.g., face) from the original image space.

Typical subspace learning methods include Principal Component Analysis (PCA) (Turk and Pentland, 1991) and Linear Discriminate Analysis (LDA) (Belhumeur et al., 1997). PCA provides an optimal linear transformation from the original image space to an orthogonal eigenspace with reduced dimensionality in sense of the least mean square reconstruction error. PCA was first applied in face recognition in the 1990s and greatly promoted face recognition development at the time. PCA successfully extracts the most distinct variations of different faces and reduces the noise components of face images so the face recognition performance is significantly improved. Given a data set $X = [X_1, X_2, \ldots, X_N]$, where N is the number of samples, PCA aims to learn projections (PCs) W that maximizes the total covariant matrix $S_t = \sum_{i=1}^{N}(X_i - m)(X_i - m)^T$, where m is the mean vector over the sample set. The PCs can be derived by solving the eigenvalue problem $S_t W = \lambda W$ with its leading eigenvalues.

Although PCA is optimal for modeling the face distribution, it may not be the best for classification because it fails to utilize any class label information. On the contrary, LDA is proposed to find a subspace that gathers the samples from the same person and meanwhile enlarges the margin among different persons. In particular, it seeks a linear transformation that maximizes the ratio of between class variance to within class variance. The between-class scatter matrix S_b and the within-class scatter matrix S_w are defined as follows,

$$\begin{aligned} S_b &= \sum_{i=1}^{c} N_i (U_i - U)(U_i - U)^T \\ S_w &= \sum_{i=1}^{c} \sum_{X_k \in C_i} (X_k - U_i)(X_k - U_i)^T \end{aligned} \quad [1]$$

where U_i is the data mean of the ith class; U is the data mean over the total data set. N_i is the number of samples in ith class. By solving the generalized eigenvalue problem $S_b W = \lambda S_w W$, the solution of LDA can be obtained with its leading eigenvalues. In practice, because of the limited sample size and high dimension data, the within class scatter S_w is usually singular. One possible way is first to reduce the dimension with PCA and then apply LDA to find the discriminant subspace for classification.

Local appearance features, as opposed to holistic features like PCA and LDA, provide another alternative to face representation. They are extracted from the local regions of face images, respectively, and hence are usually more robust to local changes (e.g., expression and lighting variations) than holistic features. Among all possibilities, Gabor and local binary patterns (LBPs) are two representative features. Daugman (1985) first proposed a 2D Gabor filter and successfully applied it to accurate iris recognition. The profiles of Gabor filters mimic the human visual perception system, thus Gabor is expected to be effective to represent face images in

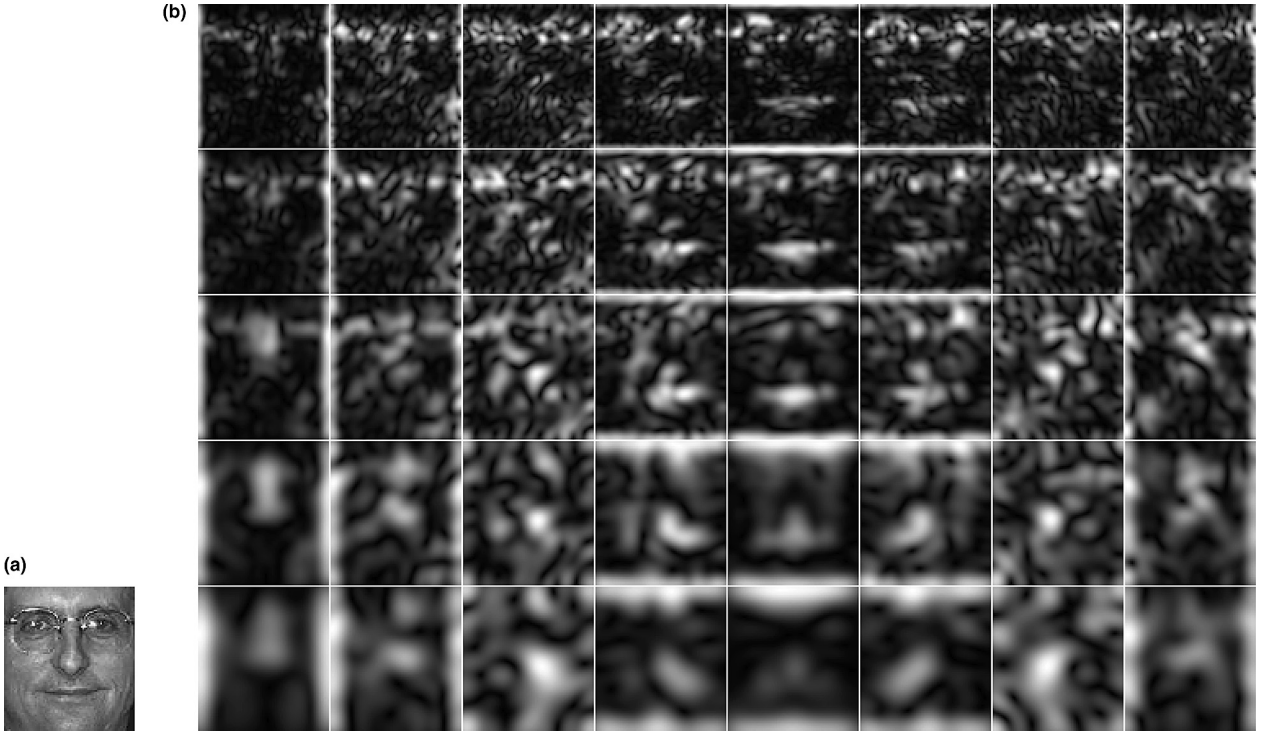

Figure 3 Face image (a) and its Gabor magnitude response image (b).

a biological manner. Moreover, Gabor wavelets have also proved to be optimal for measuring local spatial frequencies mathematically. By applying a bank of various Gabor filters at different scales and orientations, it is capable of capturing local structures corresponding to specific spatial frequency (scale), spatial locality, and selective orientation. Many well-known face recognition methods like Elastic Bunch Graph Matching (Wiskott et al., 1997) and Gabor Fisher Classifier (Liu and Wechsler, 2002) adopt the Gabor feature as face representation, validating the effectiveness of Gabor representation. Particularly, Gabor magnitude response is more popular in face recognition since it is more stable to local image variations. Figure 3 shows the magnitude images filtered by a bank of Gabor filters.

LBP (Ahonen et al., 2006) is another simple, yet effective face representation method. It models the pixel relationship in the neighborhood. Figure 4 illustrates the process of basic LBP operator. For each image pixel, the neighboring pixels are compared with the central one, and the difference is thresholded and concatenated into a binary string. This string can be converted into a decimal value, which is named an LBP label. For the sake of accuracy and generality, usually the distribution of LBP labels within certain regions is calculated to produce a local histogram, and all local histograms are concatenated to form the global representation. Due to its binary operator, LBP is invariant to monotone transformation of images and is robust to illumination changes to some extent. Nowadays, many LBP variations like Multi-Block LBP (Liao et al., 2007) and Gabor volume-based LBP (Lei et al., 2011) have been proposed to further improve the discriminative power of LBP descriptor.

The holistic and local features can also be used together to boost the face recognition performance. After extracting local features, holistic-based methods like LDA are applied on these extracted local features to find a discriminant subspace to classify faces. This is one of the state-of-the-art algorithms for face recognition.

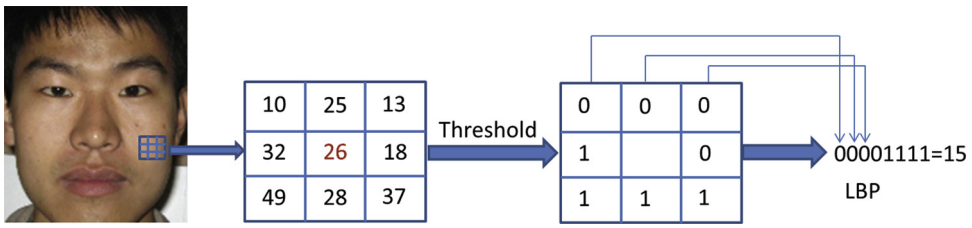

Figure 4 Local binary pattern (LBP) extraction.

Recently, Kumar et al. (2011) have proposed attribute and simile representations for face recognition. The attribute means the describable aspects of visual appearance (like gender, race, and age) and uses binary classifiers to recognize them. For simile, the test image (regions of the image) is compared to the images (regions) in a reference set, and the similarity between them is used as the face representation. The attribute and simile are finally combined to form a compact face description. This representation is shown to be effective and robust to real word face recognition.

Feature Matching

In the feature matching phase, efficient classifiers should be designed to measure the similarities of different samples and judge whether two images come from the same person or not. In many face recognition applications, the number of identities (classes) is changeable, so many traditional classification methods in pattern recognition cannot be directly applied to face recognition. The most popular and flexible classification method is the nearest neighboring classifier, which computes the similarity (or distance) of sample pairs using a specific metric. Classification result is made by finding the class label with the smallest distance. By transforming face recognition to a two-class classification problem, other popular classification algorithms can also be applied to face recognition. For example, Moghaddam et al. (2000) propose the concepts of intra- and interface samples space. This converts the multiclass face recognition problem to a two-class one by classifying whether two images come from the same person or not. In this way, some efficient classifiers like Adaboost and support vector machine (SVM) can be used to distinguish this sample pairs.

Recently, sparse representation-based classifier (SRC) (Wright et al., 2009) has been introduced in face recognition. It treats face recognition as the problem of searching the sparse combination of templates in a database to represent the input image (Figure 5). SRC assumes that images in the database, which share the same identity with the input image, would make more contributions to the representation. Because of the sparse characteristic, the SRC is robust to facial appearance variations, especially when the face is occluded. To guarantee a high accuracy, SRC requires the gallery set to contain as many diverse face images (including different illuminations, expressions, occlusions, etc.) as possible. Yet in real world applications, it is very difficult to collect many face images, which limits the application of the SRC method.

Challenges in Face Recognition

Variations such as lighting, pose, and occlusion are still main issues that affect the face recognition performance. For cooperative scenarios, frontal face detection and recognition in normal lighting environment is a reasonably well-solved problem. For noncooperative scenarios, the non-frontal face detection has been well solved, but the unconstrained face recognition is still an open problem. The face images in these scenarios (Figure 6) always suffer from one or several factors like lighting, pose, occlusion, low resolution, blur, and so forth, which lead face recognition to be a challenging task.

Face recognition across aging is another unsolved problem. Current face recognition performs well when the age range is within 5 years. However, the performance decreases dramatically when the range exceeds 10 years. Aging is challenging because it is a combination of intrinsic facial variation and external factors including expression, pose, and lighting variations. Addressing the aging problem means it must address all other variants as well. One possible way is to use a 3D face model to simulate the face appearance changing across aging and synthesize the virtual face images with similar age for face recognition. Another approach is to match the two age-

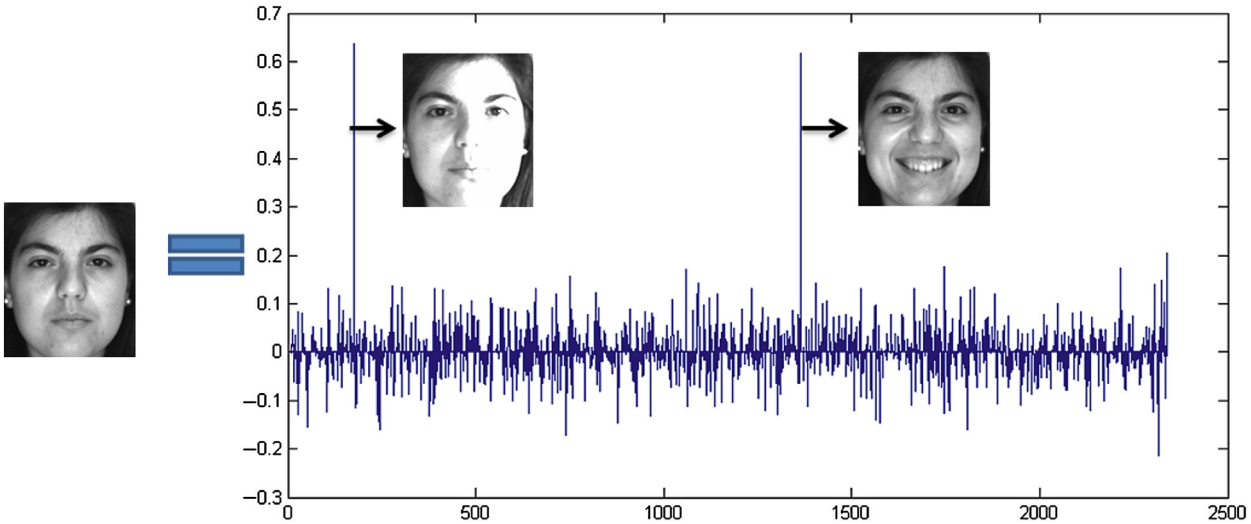

Figure 5 Process of sparse representation classifier. The test image is represented as a sparse linear combination of training images. The responses of the images that are of the same identity with the test image usually have stronger responses than others.

Figure 6 Face images captured from a face video surveillance system. There are variations on lighting, pose, occlusion, blurring, and so forth.

separated images directly. Both age-invariant feature extraction and effective classifier learning are open problems.

See also: Bayesian Theory, History, Applications, and Contemporary Directions; Learning: Mathematical Learning Theory, History of; Vision, High-Level Theory of; Vision, Low-Level Theory of.

Bibliography

Ahonen, T., Hadid, A., Pietikainen, M., 2006. Face description with local binary patterns: application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 2037–2041.

Belhumeur, P., Hespanha, J., Kriegman, D., 1997. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7), 711–720.

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active shape models: their training and application. CVGIP: Image Understanding 61, 38–59.

Daugman, J.G., July 1985. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. Journal of the Optical Society of America A 2, 1160–1169.

Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K., 2011. Describable visual attributes for face verification and image search. IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (10), 1962–1977.

Lei, Z., Liao, S., Pietikáinen, M., Li, S.Z., January 2011. Face recognition by exploring information jointly in space, scale and orientation. IEEE Transactions on Image Processing 20 (1), 247–256.

Li, S.Z., Jain, A.K. (Eds.), 2011. Handbook of Face Recognition, second ed. Springer.

Liao, S., Zhu, X., Lei, Z., Zhang, L., Li, S.Z., 2007. Learning multi-scale block local binary patterns for face recognition. In: Proceedings of IAPR/IEEE International Conference on Biometrics, pp. 828–837.

Liu, C., Wechsler, H., 2002. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Transactions on Image Processing 11 (4), 467–476.

Moghaddam, B., Jebara, T., Pentland, A., 2000. Bayesian face recognition. Pattern Recognition 33 (11), 1771–1782.

Phillips, P.J., Scruggs, W.T., O'Toole, A.J., Flynn, P.J., Bowyer, K.W., Schott, C.L., Sharpe, M., 2010. FRVT 2006 and ICE 2006 large-scale experimental results. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (5), 831–846.

Turk, M.A., Pentland, A.P., June 1991. Face recognition using eigenfaces. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Hawaii, pp. 586–591.

Viola, P., Jones, M., 2004. Robust real-time face detection. International Journal of Computer Vision 57, 137–154.

Wiskott, L., Fellous, J., Kruger, N., von der Malsburg, C., 1997. Face recognition by elastic bunch graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7), 775–779.

Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y., January 2009. Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2), 210–227.

Face Recognition, Psychological and Neural Aspects

Aina Puce, Indiana University, Bloomington, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Face recognition can be defined as the ability to identify an individual from gazing only at their face, or an image of their face. This ability remains largely unaffected even with extreme changes in lighting level, viewing angle, and distance as experienced in daily life. In cases of brain injury, individuals may develop an inability to recognize previously familiar faces – a condition known as prosopagnosia. The observation that prosopagnosic individuals never recover this function argues for the existence of a specialized brain circuitry that deals with this important stimulus category. Physiological studies in primates originally identified cells responsive specifically to faces, prompting the search for similar tissue in humans. Noninvasive brain blood flow and neurophysiological studies indicate that this is also the case in humans. These data also suggest that the ability to recognize a face proceeds independently and in parallel with the ability to read facial expressions and gestures, and that these functions are handled by a so-called 'extended face network' in the brain. There is debate as whether 'face-specific' brain regions respond specifically to faces, or whether they merely reflect the activity of neurons trained to individuate exemplars from object categories with highly similar member items. It is important to note that this ability develops in early life in a maturing brain.

Behavioral Studies, Early Theoretical Models, and Investigations in Brain-Injured Individuals

Groucho Marx once said, "I never forget a face, but in your case I'll make an exception." This statement is remarkable in that we cannot actively choose not to recognize or remember a face. This process proceeds to completion without a conscious effort on our part, and the complexity of this operation only becomes apparent when it breaks down, e.g., when a face appears familiar, but cannot be associated with a name or context of the original interaction. What is truly remarkable is that people can still recognize individuals who have not been seen for long periods of time and who have also changed their appearance as a function of aging.

After birth, probably one of the first few objects seen most often is a face. Infants actually pay more attention to faces than other stimulus categories (Morton and Johnson, 1991). Being able to recognize the face of a caregiver is important – the infant is totally dependent on them for nourishment and shelter. Facial recognition in children as young as 4 years of age has been postulated to be qualitatively similar to that of adults (McKone et al., 2009), however, different brain responses to various face-processing tasks show differential changes from childhood to adolescence, as indicated by their impaired recognition performance when the faces are presented upside down (Taylor et al., 2004). The inability of adults to successfully recognize inverted faces (Yin, 1969) had been demonstrated previously as a hallmark of specialized processing of the upright face. While it has always been thought that face-processing skills in adults are stable across the lifespan, recent evidence suggests that face naming skills peak during the early twenties, whereas the ability to learn and recognize faces without having to name them peaks in the early thirties of age (Germine et al., 2011).

Studies in adults suggest that successful face recognition proceeds in a series of stages, based on behavioral studies of normal individuals and those with brain injury. This well-known and now classic model of familiar face recognition was first proposed by Bruce and Young in the year 1986. Face perception, or detection, i.e., the ability to see a presented object as a face, and not a chair, formed the first stage of this process, the so-called structural encoding stage. The face and its features are processed holistically, with the output of the structural encoder feeding the so-called face recognition units (FRUs). At this stage, the familiarity judgment is made. Next, FRU output activates so-called person identity nodes (PINs), allowing the information previously stored about that individual (e.g., gender, age, profession, relationship to observer, usual interaction context – i.e., work or home, specific details of pleasant or unpleasant previous contact, etc.) to be accessed. Finally, the output from the PINs activates the name representation for that individual. This multipart familiar facial recognition model has in the past explained many errors in facial recognition seen in everyday life (e.g., the face is familiar, but the person's name cannot be accessed), and in cases of brain injury. A second pathway in Bruce and Young's model dealt with the processing of facial expressions, however, at the time of exposition, this was not a well-developed part of this model.

Prosopagnosia is the inability to recognize previously familiar faces, and can occur following a stroke or as a result of a brain tumor. The individual can no longer recognize even the faces of their spouse or children. Prosopagnosia can co-occur with other visual deficits, such as a loss of color vision (achromatopsia) and an inability to recognize everyday objects (object agnosia) (Meadows, 1974). These visual deficits can co-occur in brain injury, as brain regions selectively processing these stimulus attributes are located near one another. Apperceptive prosopagnosia is so named because the source of face recognition difficulty occurs largely to disruptions in basic visual perception. This form of prosopagnosia can co-occur with object agnosia, and patients often describe a degraded, fragmented visual scene. Alternatively, basic visual perception may be fairly intact, and a face is seen as a 'face,' but the individual's name or their personal details cannot be accessed, i.e., associative prosopagnosia. The ability

to recognize facial expressions is dissociable from facial recognition (Humphreys et al., 1993), and this dissociation was originally proposed in early models of face processing (Bruce and Young, 1986).

Individuals with prosopagnosia do not appear to recover the ability to recognize faces once the critical areas of the brain have been damaged. Face recognition may be 'hardwired' in the brain even in the absence of postnatal experience with faces, as illustrated by the intriguing case of prosopagnosia in a 16-year-old boy who sustained his brain injury at one day of age (Farah et al., 2000). Even more intriguing are the documented cases of 'developmental prosopagnosia,' where in the presence of apparently normal brain structure individuals are born with the inability to recognize faces, an ability that also appears not to develop in life (Le Grand et al., 2006). The different forms of prosopagnosia have stimulated researchers to hunt for specialized brain circuitry that processes faces. Additionally, recordings from single cells in visually sensitive regions of monkey brains show specific responses to faces, and not other object classes (for a review, see Desimone, 1991; Milders and Perrett, 1993). Given that humans and monkeys are both social animals, and that faces are an important stimulus in this context, it was thought likely that the human brain possesses brain cells with similar response properties.

Neuroimaging and Neurophysiological Studies and a Brain Network for Face Processing

In the latter part of the twentieth century many human physiological studies were dedicated to investigating the neural mechanisms underlying facial perception (see review by Haxby et al., 2000). This was prompted, in part, by the development of neuroimaging techniques such as positron emission tomography (PET) and, more recently, functional magnetic resonance imaging. Both methods effectively measure focal changes in brain blood flow during perception and cognition. One of the first investigations of face perception and recognition was performed by Justine Sergent working at the Montreal Neurological Institute in 1992 (Sergent et al., 1992). Sergent and her colleagues performed a PET study examining differences in cerebral blood flow when healthy adults viewed pictures of faces and discriminated between various facial attributes. For example, subjects made gender judgments (deciding whether a face was male or female), remembering if a particular face had been shown to the subject previously, and so on. The blood-flow patterns in these conditions were contrasted to conditions where subjects viewed visual material such as gratings (grids of black and white lines). These studies identified regions of the occipital and temporal lobe on the underside of the brain as being selectively active when subjects viewed and discriminated between faces. Since then many investigators have followed suit and studied other aspects of facial processing (reviewed by Haxby et al., 2000), and the studies show concordance with this initial pioneering investigation. Taken together the literature to date suggests that the brain possesses parallel pathways that deal with facial identity and facial gesture, respectively (Allison et al., 2000), and that there is a distributed network of brain structures that work together to handle these different functions (Haxby et al., 2000), as depicted in **Figure 1**. While 'face-selective' regions in both hemispheres possess the capability to process faces, it is the right hemisphere that appears to be more important for this process. In prosopagnosia, for example, if the lesion occurs on one side of the brain it is usually on the right side (De Renzi et al., 1994).

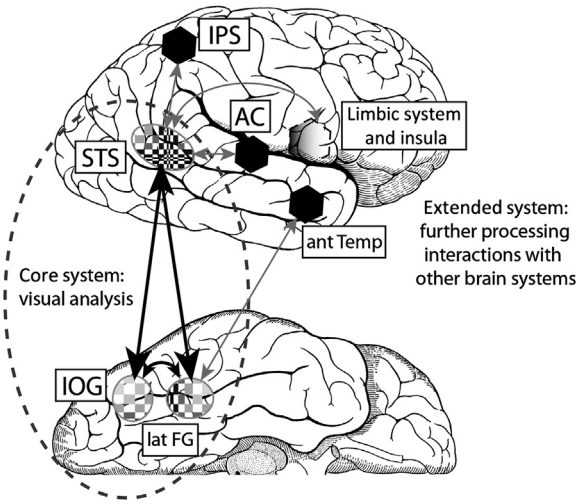

Figure 1 The distributed brain network for processing faces as proposed by Haxby, J.V., Hoffman, E.A., Gobbini, M.I., 2000. The distributed human neural system for face perception. Trends in Cognitive Sciences 4 (6), 223–233 from an extensive review of the literature. The distributed brain network consists of two major parts. A core system performs visual analysis of faces and consists of the inferior occipital gyri (IOG), the lateral fusiform gyrus (lat FG) and the superior temporal sulcus (STS). The IOG encodes basic facial features, the lat FG is thought to deal with invariant aspects of faces, which aid in the gleaning of unique identity. The STS deals with the perception of dynamic aspects of faces such as eye gaze and facial expression changes, as well a lip movement. An extended system derives input from the core system and performs further face processing and interacts with other brain systems. The extended system includes the anterior temporal lobe (ant Temp), which takes input from the fusiform gyrus so that personal identity, name and biographical information can be linked to the seen face. Other parts of the extended system include the intraparietal sulcus (IPS) (thought to deal with the aspects of direction attention in one's space), the auditory cortex (AC) (assisting in speech perception), and the limbic system and the insula (dealing with emotional aspects of the face and affective context).

Our understanding of what is active in the human brain when a face is being evaluated clearly has evolved as a result of the now substantial neuroimaging and neuropsychological lesion literature on face processing. For example, the extended face-processing network (**Figure 1**) now includes explicit brain structures that are linked to particular face recognition processes. Additionally, the alternate pathway for processing facial expression has now been more completely specified, relative to the original model proposed by Bruce and Young in 1986.

Blood flow and neuropsychological lesions studies clearly show which brain regions are active, however these methods cannot characterize these changes well over the course of time. Recordings of electrical or magnetic activity of the brain electroencephalography or magnetoencephalography have resolution to thousandths of a second (millisecond) and can

therefore chart the dynamics of face perception processes accurately. A brain response to seeing a face occurs at around 1/5th of a second, which has been named as N170 (Bentin et al., 1996). The robustness of the N170/N200 under a large number of perceptual manipulations and the seemingly automatic way in which the response is generated suggests that this may be a neural correlate of the structural encoder of Bruce and Young's (1986) original model, and is a vital first part in the visual analysis of the face in Haxby et al.'s model (2000). The N170 data are consistent with behavioral studies of face perception, where healthy subjects can readily detect faces relative to other object categories, despite stimulus degradation, fragmentation, rotation, inversion, manipulations of light and shade, and so on (Bentin et al., 1996). A similar neural response can also be recorded directly from the surface of the fusiform gyrus in human brain in epilepsy surgery patients (so called N200), and like its scalp counterpart occurs irrespective of whether the observer attempts to recognize the face or not, and is largely independent on lighting conditions, size, orientation of the face, gender, or familiarity of the face (Puce et al., 1999). Brain activity that follows the N170/N200 is more likely to be attributable to higher order processes such as recalling the face or characterizing it according to some category (Puce et al., 1999).

Despite our excellent abilities to recognize faces under varied environment, in the laboratory setting facial recognition abilities can be challenged and be shown to be impaired under different extreme condition. Individuating one person's face from another requires that the features that are unique to that particular (familiar) individual are extracted and matched to a preexisting 'template.' Manipulations that impair our ability to extract subtle spatial differences will affect successful facial recognition. For example, Groucho Marx had a very idiosyncratic and easily recognizable face (see **Figure 2**). Yet, when he is upside down he is more difficult to recognize (compare **Figure 2(a)** with **Figure 2(b)**). Similarly, a negative image may make a previously familiar face unrecognizable (**Figure 2(c)**). Similarly, manipulations of spatial frequency content or amount of detail of the face can also impair facial recognition (**Figure 2(d) and 2(e)**). Under extreme conditions such as these we are forced to rely on idiosyncratic, incidental details that might accompany the face, such as the cigar and moustache so that we have learned to associate with Groucho Marx's face. Experimental manipulations such as these also help to emphasize how expert we are with identifying faces in the natural setting – the process is effortless and automatic.

On a daily basis, in natural conditions, our expert ability to discriminate between individual faces is based on detecting changes in subtle spatial configurations in a homogeneous object category, unlike any other object category dealt with on a daily basis. Our specialized facial recognition skills are so honed that behavioral studies have repeatedly demonstrated an own-race advantage for facial recognition across different ethnic groups, i.e., Caucasian, Asian (Brigham, 1986). These data suggest that there really might be a basis for the often-heard comment from travelers that the faces of people of other races look alike. Different ethnic groups have idiosyncrasies in their facial features that an individual member of that particular group learns to differentiate. The expertise that develops with the individual's own ethnic group may hence not necessarily be generalizable to another ethnic group. It has been said that the same type of bias might apply to viewing individuals in one's own age range also.

Does 'Face-Specific' Cortex Participate Only in Face Processing?

Are faces a special stimulus category? There is no doubt that we are experts with faces. However, there is debate about the nature of this expertise, and there are currently many unanswered questions regarding these issues. For example, do 'face-specific' regions of the brain deal only with faces, or are they also active in individuals who are experts with other object categories? i.e., are these neurons functioning specifically for detecting and recognizing faces, or are they a more general expert 'individuator' of categories of objects with highly similar member items (Gauthier and Logothetis, 2000)? Is the expertise with faces, and associated brain circuitry, that develops with the developing brain throughout childhood, different to expertise acquired with other stimulus categories in adult life?

How can these various interrelated processes be disentangled, given that we cannot test people who do not have a lifetime's exposure to faces? There are a number of approaches that are currently being undertaken in order to try and unravel these issues. First, some insight may come from studying patients with developmental prosopagnosia (Le Grand et al., 2006). This is an extremely rare disorder, where

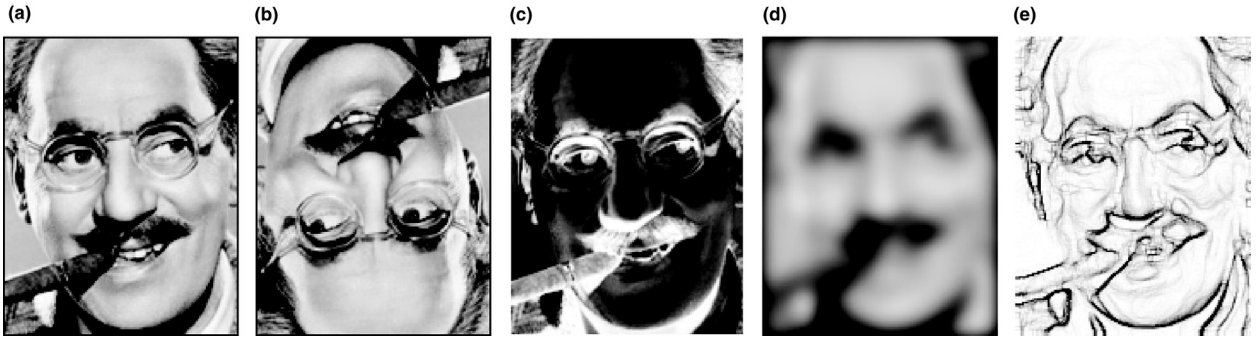

Figure 2 The many faces of Groucho Marx. (a) The unaltered face; (b) an inverted orientation; (c) inverted gray-scale palette; (d) removal of the high-spatial frequency content of the image; and (e) removal of the low-spatial frequency content.

the individual has never developed the ability to recognize faces. Physiological and behavioral studies of face and object recognition in these individuals relative to both healthy and brain-injured subjects may shed some light on these questions. Second, face perception and recognition studies using cutting-edge neuroimaging techniques may be helpful. Direct recordings of electrical activity from the brain indicate that face-sensitive regions exist in a patchy mosaic with regions responsive to objects, and to words, for example. The relatively coarse spatial resolution in most neuroimaging studies to date could produce blood-flow measures containing contributions from different kinds of category-specific regions, making it difficult to evaluate exactly how these brain regions deal with facial information. Studies of facial recognition performed with (high field strength) functional MRI combined with recordings of the electrical activity of the brain in the same subject may finally shed some light on why it is impossible to forget a face, despite our best attempts to do so. Finally, many investigators are turning to the developing brain for answers to these questions. There are many knowledge gaps related to how the young human brain develops and hones its facial recognition skills from infancy to adolescence.

See also: Face Recognition Models: Computational Approaches; Facial Emotion Expression, Individual Differences In; Neural Representations of Objects; Prosopagnosia; Visual Perception, Neural Basis of.

Bibliography

Allison, T., Puce, A., McCarthy, G., 2000. Social perception from visual cues: role of the STS region. Trends in Cognitive Science 4 (7), 267–278.

Bentin, S., Allison, T., Puce, A., Perez, A., McCarthy, G., 1996. Electrophysiological studies of face perception in humans. Journal of Cognitive Neuroscience 8, 551–565.

Brigham, J.C., 1986. The influence of race on face recognition. In: Ellis, H.D., Jeeves, M.A., Newcombe, F., Young, A. (Eds.), Aspects of Face Processing. Martinus Nijhoff, Dordrecht, The Netherlands, pp. 170–177.

Bruce, V., Young, A., 1986. Understanding face recognition. British Journal of Psychology 77 (Pt 3), 305–327.

De Renzi, E., Perani, D., Carlesimo, G.A., Silveri, M.C., Fazio, F., 1994. Prosopagnosia can be associated with damage confined to the right hemisphere – an MRI and PET study and a review of the literature. Neuropsychologia 32 (8), 893–902.

Desimone, R., 1991. Face-selective cells in the temporal cortex of monkeys. Journal of Cognitive Neuroscience 3 (1), 1–8.

Farah, M.J., Rabinowitz, C., Quinn, G.E., Liu, G.T., 2000. Early commitment of neural substrates for face recognition. Cognitive Neuropsychology 17 (1), 117–123.

Gauthier, I., Logothetis, N.K., 2000. Is face recognition not so unique after all? Cognitive Neuropsychology 17, 125–142.

Germine, L.T., Duchaine, B., Nakayama, K., 2011. Where cognitive development and aging meet: face learning ability peaks after age 30. Cognition 118 (2), 201–210.

Haxby, J.V., Hoffman, E.A., Gobbini, M.I., 2000. The distributed human neural system for face perception. Trends in Cognitive Sciences 4 (6), 223–233.

Humphreys, G.W., Donnelly, N., Riddoch, M.J., 1993. Expression is computed separately from facial identity, and it is computed separately for moving and static faces: neuropsychological evidence. Neuropsychologia 31 (2), 173–181.

Le Grand, R., Cooper, P.A., Mondloch, C.J., Lewis, T.L., Sagiv, N., de Gelder, B., et al., 2006. What aspects of face processing are impaired in developmental prosopagnosia? Brain and Cognition 61 (2), 139–158.

McKone, E., Crookes, K., Kanwisher, N., 2009. The cognitive and neural development of face recognition in humans. In: Gazzaniga, M.S. (Ed.), The Cognitive Neurosciences, fourth ed. MIT Press, Cambridge.

Meadows, J.C., 1974. The anatomical basis of prosopagnosia. Journal of Neurology, Neurosurgery, and Psychiatry 37 (5), 489–501.

Milders, M.V., Perrett, D.I., 1993. Recent developments in the neuropsychology and physiology of face processing. Bailliere's Clinical Neurology 2 (2), 361–388.

Morton, J., Johnson, M.H., 1991. CONSPEC and CONLERN: a two-process theory of infant face recognition. Psychological Review 98 (2), 164–181.

Puce, A., Allison, T., McCarthy, G., 1999. Electrophysiological studies of human face perception. III: effects of top-down processing on face-specific potentials. Cerebral Cortex 9 (5), 445–458.

Sergent, J., Ohta, S., MacDonald, B., 1992. Functional neuroanatomy of face and object processing. A positron emission tomography study. Brain 115 (Pt 1), 15–36.

Taylor, M.J., Batty, M., Itier, R.J., 2004. The faces of development: a review of early face processing over childhood. Journal of Cognitive Neuroscience 16 (8), 1426–1442.

Yin, R.K., 1969. Looking at upside-down faces. Journal of Experimental Psychology 81 (1), 141–145.

Facial Emotion Expression, Individual Differences in

Andrea Hildebrandt, Ernst-Moritz-Arndt University Greifswald, Greifswald, Germany
Sally Olderbak and Oliver Wilhelm, Ulm University, Ulm, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M.G. Frank, volume 8, pp. 5230–5234, © 2001, Elsevier Ltd.

Abstract

In this contribution, we present theories discussing the origin and function of facial emotion expressions, research regarding their neural architecture, evidence for individual differences, and applied measurement methodologies. We then describe a series of assumptions regarding the factorial structure of facial emotion expression ability, showing that all of these postulations still need comprehensive empirical research in order to better understand individual differences in facial expressivity. Finally, we discuss the nomological net surrounding the production of facial emotion expressions including the relation of this ability with related skills, such as the ability to perceive emotion cues from faces.

For over a century scholarly literature has recognized that facial expressions play a key role in humans' communication of affect. Since Darwin's classic book, *The Expression of Emotions in Man and Animals*, the scientific reflection on spontaneous and voluntary facial activity has inspired a series of currently unresolved debates over the origin and function of facial emotion expressions in humans. Many researchers (e.g., Barrett, 2011) have strongly questioned that the *prototypic* emotional expressions – *fear*, *sadness*, *disgust*, *anger*, *happiness*, and *surprise* – are biologically hardwired, have well-defined functions for human survival, and thus exist in everyday life. Instead researchers have asked whether they "have created a science of emotional symbols by routinely using *prototypical expressions* in experiments" (p. 402, italics added).

In this article, we review debated theoretical positions over the origin of facial expressions and emphasize that a contemporary view on the nature of facial emotion displays should also recognize that to a strong degree facial expressions are voluntarily used to shape social interactions and that facial displays are socially learned: "Emotional expressions might be learned and, like other symbol-based communication that is socially learned, this would be evolutionary significant, immediately functional in individual instances, and adaptive for a species" (Barrett, 2011: p. 406). Because expressions are assumed to be, to a strong degree, defined by culture and interpersonal interactions, a contemporary view on emotion expressions recognizes within-emotion variability across persons, situations, and cultures as well as between-emotion similarities in facial displays. Such variations were also recognized by Darwin (1872/1965) and there is a growing account for such variability in the recent scientific literature (see, e.g., Barrett, 2009; Reisenzein et al., 2013).

In addition to theoretical considerations regarding the origin, nature, function, and neural bases of facial emotion expressions, we will stress the role of facial displays in everyday communication and define the voluntary production of facial expressions as a basic socioemotional competence that can and should be measured as a prominent individual differences construct. We will review the literature on the nomological net of facial expressivity and indicate that a multivariate and psychometric approach is needed to assess emotion communication ability via faces and its correlates in a conclusive way.

Theories on Facial Expressions of Emotion

Particularly Allport (1924), Tomkins (1962), Ekman (1972), and Izard (1971) theorized that facial emotion expressions are shaped throughout humans' evolutionary history and play a functional role in humans' adaptation to the physical and social environment. This functionalist interpretation of facial expressions has been widely accepted in later research and proponents have emphasized the need for studying spontaneous expressions. From the theoretical point of view that emotions are discrete, research suggests that prototypical emotion-specific response configurations will occur in the face as a consequence of the emotional experiences of that organism (e.g., Ekman, 1992a). In parallel, sociobiologically oriented theories emerged, suggesting facial expressions are purely social signals used for strategic or deceptive purposes (Dawkins and Krebs, 1978). These theories accentuated the important role of emotion displays in social communication. However, the cultural stereotype of affect displays, but also recent evidence based upon a comparison of sensory acquisition during fear versus disgust expressions (Susskind et al., 2008), to some extent support the functional utility of facial emotion expressions.

Hence, contemporary theorists argue that expressive signals can only be successfully used for strategic communicative purposes in a perspicuous manner because they are also often states of the organism. Scherer (2009a) defined *pull factors* of expression as voluntarily produced signals that are culturally coined and have specific structures that can be well understood by others. Furthermore, it is assumed that *pull factors* produce highly similar outputs like *push factors*, which are factors determined by the internal state of the organism. Thus, Scherer (2009b; also Scherer et al., 2013) emphasized that emotion expression is multifunctional; this is in line with Bühler's postulation regarding the trifold facets of signals: signals that operate like a *symptom* can be understood as a *symbol* and are often used as an *appeal* (Bühler and Goodwin, 1990). Scherer (2007) further suggested that emotion expressions have a strong strategic aspect which is more or less pronounced depending on the situation and that the *voluntary* and *purposeful* use of emotional expressions is a central element of communicative socioemotional skills, where individual differences may arise.

Past research on facial expressions produced an abundance of scientific studies that utilized a static viewpoint on faces. Furthermore, these studies mainly used expression recognition paradigms to study the nature and function of facial expressions. Marking the onset of a new research epoch, a series of contributions to a recent special section published in *Emotion Review* 2013 emphasized (1) the need for a theoretical model to explain the expression *production* mechanism and formulated (2) a desideratum for considering the dynamics of facial expression production (e.g., Fernández-Dols, 2013; Scherer et al., 2013).

Basic emotion theories (e.g., Ekman, 1972) that dominated facial expression research in the past postulated prototypical emotion-specific response configurations in the face as a consequence of a particular emotional experience of the person. Within this theoretical framework it has been assumed that expressions originate from so-called affect programs (AP). Importantly, the AP view considered affect display configurations to arise at a discrete time point, and it assumes possibly no differences in the activation timing of action units (AUs, functionally independent facial muscles coded in the Facial Action Coding System (FACS), see below for details).

In contrast, the componential theoretical view on expression production, recently described by Scherer et al. (2013), proposed that facial AUs or AU configurations emerge as a consequence of appraisal results. This perspective considers the production of facial expressions to be a dynamic process so that predictions regarding the sequential appearance of AUs in response to appraisal checks can be made. For instance, if *novelty* is attributed to a stimulus in the context of the appraisal domain *relevance* and the stimulus is further characterized by its sudden appearance, presumably AUs 1 + 2 and 5 will be activated, a configuration that is typically considered to represent fear. In other instances – depending on person-specific characteristics and situative variables – a variety of anger, joy, or even disgust expressions may appear to be a consequence of the same appraisal checks. Furthermore, if the appraisal check leads to some kind of intrinsic pleasantness, AUs 12 + 25 and 38 may be activated and a joy expression will emerge (examples from Scherer et al., 2013).

Reisenzein et al. (2013) listed additional alternative theories that try to explain the emergence of specific emotion-related facial expression configurations. These include Fridlund's (1994) theory that humans have combined AP, which is supported by research from the behavioral ecology field, Fernández-Dols' (1999) situationist view of emotion expression, Frijda and Tscherkassof's (1997) theory that facial expressions are components of evolutionary action tendencies, Parkinson's (2013) theory that emotional expressions are inherited and learned relational orientations, Russell's (2003) dimensional view, which explains that facial expressions are the bases for pleasure–arousal experiences, and Camras et al. (1997) proposal that facial expressions are coordinative motor structures. The authors concluded that alternative explanations to AP (including the componential view by Scherer et al., 2013) are better suited to account for some of the findings on a generally limited coherence between emotion and facial expressions; however, none of the theories are yet advanced enough to account for a series of available empirical findings (Reisenzein et al., 2013).

The Neural Architecture of Facial Expressions of Emotion

Research regarding the neural architecture of facial expression production is still in its infancy (Korb and Sander, 2009; but see Rinn, 1984). Emotional facial expressions are produced by about 20 pairs of facial muscles innervated by the seventh cranial nerve: the facialis originates in the motor nucleus of the facial muscle and is situated in the ventrolateral region of the inferior pons. The existence of two dissociable facial pareses – the emotional versus the voluntary facial paresis (e.g., Hopf et al., 1992) – indicate at least partly different neural networks and mechanisms for spontaneous and voluntary expression production. In addition to patient data, recent neuroimaging studies investigated the functional involvement of brain networks during spontaneous compared with voluntary facial expressions (e.g., Wild et al., 2006). However, artifact problems during PET and fMRI measurement due to facial movements is still an issue and a strong limitation of noninvasive methods.

In a nutshell, these studies suggest that voluntary expressions are controlled mainly by the primary motor cortex (M1) and the lateral premotor cortex (LPMC). These brain areas control the contralateral muscles of the lower face and both contra- and ipsilateral muscles of the upper face, which leads to greater problems in performing asymmetric facial movements with the upper parts of the face, for example the eyebrows. In addition to M1 and LPMC, there are further cortical motor areas involved in facial expressions: the supplementary motor area (M2), and the rostral and caudal cingulate motor area (M3 and M4, respectively; Morecraft et al., 2001). Spontaneous facial expressions, on the other hand, are traditionally believed to be produced by subcortical areas including the basal ganglia and the amygdala; the activity of these areas may be modulated by rostal and medial cortical structures (Rinn, 1984). However, it also seems that emotions may lead to spontaneous facial expressions via the cingulate motor areas (Morecraft et al., 2004).

The facial feedback hypothesis (McIntosh, 1996) assumes that the intentional uptake of an affect display can generate a subjective feeling of the target's affect. Based on this observation, Ekman (1992b) speculated about connections between the motor cortex and additional brain areas involved in emotion processing. The finding that brain regions responsible for affective behavior (e.g., insula, amygdala) are functionally involved during the intentional imitation of emotion expressions and the absence of such activation during chewing and licking in an fMRI setting led Lee and colleagues to conclude that voluntary imitation is interrelated with the affective representation of the emotion (Lee et al., 2006). Due to the sparse knowledge about the neural mechanisms of voluntary versus emotional facial expressions, and the remaining methodological problems inherent in imaging methodologies, we may conclude that the most consistent finding in this area is the involvement of the M1 in the production of voluntary emotion displays whereas its involvement seem to be unnecessary for spontaneous expressions of emotion (Korb and Sander, 2009).

Movements and actions can be studied with a number of neuroscientific methods (beside imaging techniques susceptible

to movements) of which very few have been applied to facial expressions. Individual differences in such measures are not investigated. Scalp-recorded movement-related cortical potentials start about 2 s prior to self-paced voluntary movements (Kornhuber and Deecke, 1965) and include several components: (1) a gradual and bilaterally symmetrical negativity (Bereitschaftspotential: BP); (2) a steeper asymmetrical negativity (negative slope: NS′); (3) a small premotion positivity (PMP); (4) a transient small negative deflection (motor potential: MP); and (5) a large positive deflection (reafferent potential: RAP). The negative slow wave, including the BP and the NS′, is commonly referred to as the readiness potential. As a post-movement component similar to MPs (premovement component), the N+50 (Shibasaki et al., 1980) and frontal peak of motor potential (Tarkka and Hallett, 1991) was reported to have maximal activation over frontal regions, contralateral to the responding limb. These differentiated analyses of movement-related potentials have been measured for finger movements. There seems to be only one study that applied analyses of movement-related event-related brain potentials (ERPs) to voluntary facial expressions: Korb et al. (2008) compared the BP before self-paced unilateral and bilateral smiles with the BP before unilateral finger movements. The BP before bilateral smiles was comparable in size to the BP for right-hand movements (in right-handers) and it showed bilateral symmetry.

The BP is presumably part of a further ERP component, which is elicited when a warning stimulus or cue announces a second stimulus, which requires a response. Usually, negativity with a maximum at the central recording site occurs between the warning stimulus and the response signal. This negative potential has been termed the contingent negative variation (CNV; Walter et al., 1964). When the interstimulus interval is long enough, a first negativity, following the warning signal, can be distinguished from a negativity preceding the response signal. The latter has been related to the BP even if such negativities can be observed also in situations without response requirements. The CNV has been shown to be useful in exploring motor programming. Motor programming refers to the fact that elements of a movement sequence can be prepared in advance of the movement and can be executed without external feedback (Keele, 1968). Very recently, the CNV has been investigated in the context of voluntary facial expressions (Recio et al., 2014). The authors tested the effects of motor planning and reprogramming of facial expressions of happiness and anger and found no evidence for emotion specificity (anger versus joy) in the CNV (initial preparation). However, there was a greater allocation of processing resources for the inhibition of the preactivated motor plan and the updating of a new one in the case of anger compared with happiness expressions.

Individual Differences in the Facial Expression of Emotion

With respect to emotion-related abilities, Scherer (2007) proposed four major areas of competencies where individual differences may occur. The *production* competence refers to the ability of an individual to generate appropriate reactions in emotion-eliciting situations. Among others, the spontaneous expression of emotions in the face falls into this branch of emotion-related competencies. *Response preparation* and *emotion regulation* are further classes of competencies proposed by Scherer. The fourth class of competence, the *communication of emotion*, embraces aspects of productive (encoding) and receptive (decoding) abilities. The voluntary expression of emotion in the face in response to a command can be considered a relevant facet of basic socioemotional communicative abilities. However, the study of the psychometric structure of these abilities, the investigation of their location within a taxonomic model of socioemotional skills, and their relationship to further constructs are still questions that need to be addressed empirically in multivariate studies. Available univariate data, however, can help to postulate psychometric models of facial expressivity.

The structure of individual differences in emotional expressivity has been mainly investigated via self-report measures using structural equation modeling (e.g., Friedman et al., 1980; Gross and John, 1995). Self-report scale scores have then been related to univariate behavioral indicators of facial expressivity. The estimated first-order correlations were, however, zero or of a low magnitude for different self-report scales, ranging between −0.04 and 0.37. Thus, the shared variance between persons' beliefs about their expressivity and their objectively measured expressive skills is maximally 14% (e.g., Gross and John, 1995). Investigating individual differences based on objective measures is therefore a crucial demand for a better understanding of emotion-related production skills in the facial modality.

Individual differences studies applying behavioral observations of spontaneous and voluntary (posed) facial expressions are rather rare and multivariate studies on large samples that allow extracting latent variables that are adjusted for measurement error and method specificity do not exist at all. As far as we know from zero-order correlations of manifest variables, spontaneous and voluntary expressivity in the facial modality are correlated, 0.46–0.62 (Cunningham, 1977; Zuckerman et al., 1976). Thus, there may be a factorial differentiation between spontaneous production of facial expressions and voluntary nonverbal communicative skills. Furthermore, because humans have more or less ample control over their facial expressions, studying the structure of emotional communication skills in the facial modality may even lead to additional difficulties. Measurement paradigms need to include tasks that tap voluntary neutralization, masking, and simulation abilities. The question to be addressed is whether the ability to voluntarily inhibit emotional expressions in emotion-laden situations (neutralization) is distinguishable from the ability to show a different, socially desirable emotion when another emotion is felt (masking), and from the ability to simulate a socially desirable emotion if no emotion is present in a given situation (simulation). Furthermore, emotion category specificity within all these theoretically conceivable ability domains also needs systematic investigations from a latent variable modeling perspective in order to understand individual differences in facial expressivity in a conclusive way.

The Measurement of Facial Expression Production

By classifying observational methods of emotion measurement, Reisenzein et al. (2014) distinguished between (1)

objective behavior coding systems, (2) theory-based observation methods, and (3) intuitive observer judgments of emotions. Whereas the first just codes emotion-related behaviors, the second also makes inferences about emotions based on the coded behavior following some kind of theoretical assumptions. The intuitive observer judgment of emotions relies on an observer's implicit belief about the relationship between different emotions and their behavioral manifestations. All of these types of observational methods have been applied in facial emotion expression research.

The FACS (Ekman and Friesen, 1978) is an objective system for the measurement of facial behavior that is based on anatomical knowledge about the facial muscles and their movement configurations. The FACS codes expressions in terms of AUs that are defined as the smallest possible, functionally independent motions of the facial muscles (for details on coding AUs we refer to Matsumoto, 2009). FACS is an atheoretical coding system of AUs; thus, it does not make any inferences about observable emotion displays based on a particular observed configuration of AUs. Examples of theory-based coding systems are the well-known emotion FACS, which is based on the basic emotion theory and the Facial Expression Coding System (Kring and Sloan, 2007). The FACS system is also used as a component in the Specific Affect Coding System (Coan and Gottman, 2007), which incorporates facial emotion expressions with verbal and other nonverbal behaviors, and tends to code positive and negative emotion constructs (see Reisenzein et al., 2014 for details).

When coding recorded data from behavioral studies on facial expressions of emotion, researchers also adopt observer judgments above objective- and theory-based systems. Such judgments are mainly used for studying individual differences in spontaneous expressivity. For instance, in a slide-viewing paradigm, Buck (1979) asked participants to watch emotionally laden images and to rate their emotional reactions on several emotion scales. During this task, the senders' spontaneous facial expressions were filmed by a hidden camera and presented to receiver participants who were asked, based on the senders' facial expressions, to judge what kind of slide the sender viewed. Generally, the measurement of spontaneous expressivity is based upon induction paradigms using emotional films, pictures, situation descriptions, etc. (e.g., Rottenberg et al., 2007; Harmon-Jones et al., 2007), so that the observed persons are most often unaware that their expressive behavior is recorded (Buck, 1984). In contrast, studies on voluntary expressivity ask participants to do their best at displaying certain expressions or imitating facial emotion displays that are presented.

Recently, we have made attempts to measure individual differences in the ability to express emotion signals in the face using multiple task paradigms. The multiple task approach is essential for assessing abilities. Five experimental tasks were designed asking participants to either produce a specific emotional face, or with varying conditions, imitate a posed facial expression (e.g., static picture or a video). The expressions have been recorded for a period of 5 s, resulting in time series data for a single trial (see **Figure 1**). Before these tasks participants were asked to show a neutral face, in order to assess baseline emotion expression, and before and after the tasks to move specific parts of their face (e.g., raise eyebrows) in order to later assess general facial plasticity. A review of emotion expression coding methodologies suggests that automated software programs (e.g., FACET) are useful for coding numerous images. We compared four scoring methods (arithmetic mean across the 5 s video data, geometric mean, average area under the curve, and max score) as well as four data treatments (untreated, loss smoothing, controlling for baseline facial emotion expression, and controlling for facial plasticity) and found that controlling for baseline facial emotion expression, facial plasticity, and applying a max scoring technique was the best treatment for the data. Controlling for baseline emotion expression accounts for individual differences in base rate emotion expression, which potentially makes some emotions easier for an individual to express compared with others. Controlling for facial plasticity accounts for individual differences in the extent to which one can express particular AUs, thus achieving particular emotion expressions. Finally, during a trial participants may vary in the intensity of their emotion expression; if a score for best performance is desired, the max scoring identifies a participant's highest score and, unlike the other scoring methods, is not affected by that variability (see Figure 1 and Olderbak et al., 2013 for more details).

Toward a Psychometric Model Representing the Structure of the Voluntary Facial Expression Production Ability

It is challenging to postulate a factorial structure of voluntary facial emotion expression (VfEE) as an aspect of socioemotional communicative skills, because research lacks sufficient knowledge regarding the specificity of individual differences in this domain. This is because no available study on VfEE goes beyond the use of univariate measures, reports of mean performance differences across experimental situations, zero-order correlations between spontaneous versus voluntary expressions, or reports of their manifest relationship with decoding abilities and personality traits. Thus, postulations in this area are mostly based on theoretical assumptions.

Conceptually there may be a difference between the *production* and *imitation* of expressions. Production could be defined as displaying facial expressions in response to a particular emotion label, whereas imitation could be defined as voluntarily mimicking the facial expression of another person. It remains an empirical question whether such a differentiation would influence only the mean performance or also the rank order of persons, thus the correlational structure between production and imitation measures. If the rank order is affected, a single latent expression factor might not sufficiently account for the variability in production and imitation performance (see Panel A of **Figure 2** as compared with Panel B). Regarding performance levels, it is plausible to expect higher expression scores, when imitating prototypical expressions, compared with expression scores when producing a facial expression in response to a label. This hypothesis relies on the assumptions that production reflects a higher amount of idiosyncratic (person-specific) expression characteristics which are also a weaker display of prototypical emotion expressions. Socializations theories (e.g., Halberstadt et al., 2001) would postulate differentiable VfEE factors, because the expressivity

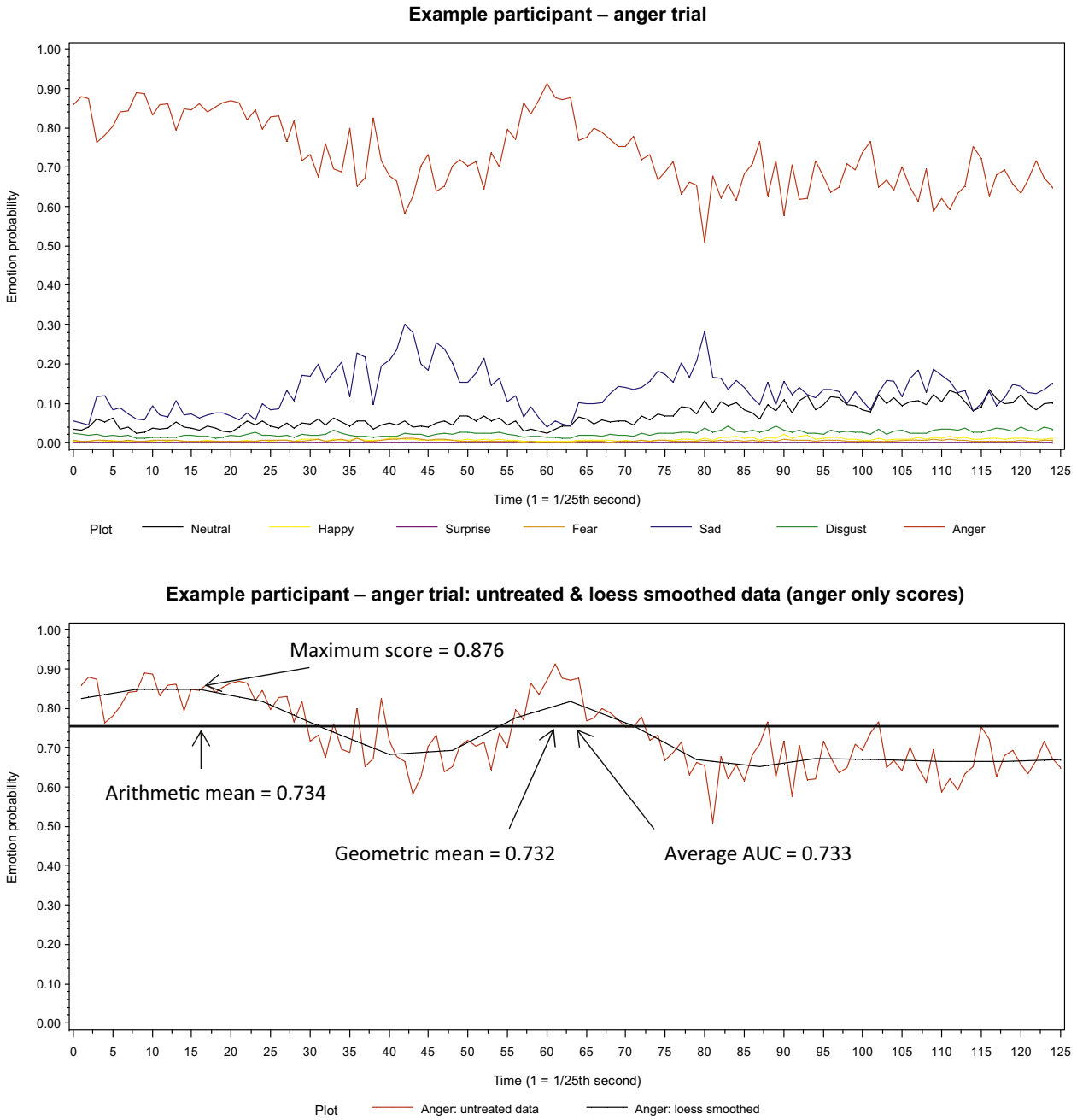

Figure 1 Trial level data from tasks measuring emotion expression ability over a time frame of 5 s.

level of the environment in which a person grew up may influence the persons' production ability, but probably the imitation performance less. Imitation performance may instead be related to facial plasticity and extraversion-related personality traits.

In addition to basic production and imitation factors, it can be theorized that the abilities to neutralize (i.e., suppress), mask, and simulate facial expressions are distinct abilities that may be more dependent on situation-specific knowledge about emotion-related display rules, emotion regulatory competence, as well as agreeableness and emotionality-related personality traits. Not named indicators in Figure 2 may represent performance measures of neutralizing, simulating, and masking facial expressions.

Furthermore, it is important to test whether emotion category specific factors besides the more general factors are needed to explain individual differences in production, imitation, neutralization, masking, and simulation measures, or whether global expressivity factors are sufficient to account for the correlations across such measures (see Panel C of Figure 2 as compared with Panels A and B). Again, theories that explain the development of facial expressivity mainly

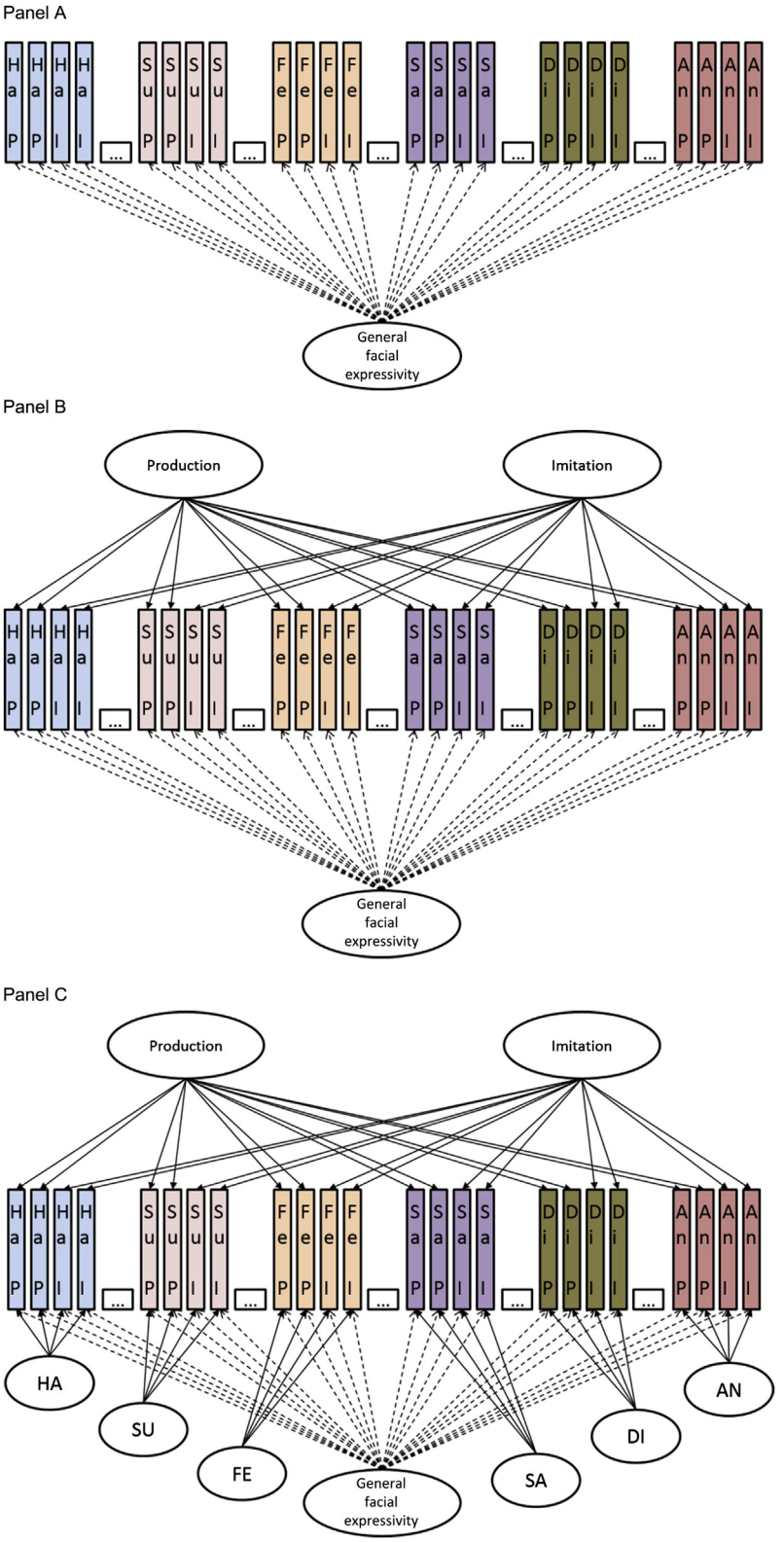

Figure 2 Schematic representation of a series of conceivable measurement models on the structure of individual differences in facial expressivity measured with multiple behavioral tasks. Boxes represent measured variables, e.g., average performance across a series of trials of an expression task; directed arrows represent directed effects; thus the models assume latent variables (represented as ellipses) to account for the common variance across manifest variables. Panel A represents a model that assumes a general factor to account for the systematic variance across persons in all measures of facial expressivity; Ha, happiness expression trials; Su, surprise expression trials; Fe, fear expression trials; Sa, sadness expression trials; Di, disgust expression trials; An, anger expression trials; empty indicator boxed represent further conceivable measures of facial expressivity beside production (P) and imitation (I). Panel B displays a model that assumes general expressivity not to be sufficient to explain the common variance across all expressivity indicators and would test whether specific variance due to production vs. imitation competence may enhance the model. Panel C displays a more complex model that assumes specific performance differences between persons depending on which emotion needs to be expressed.

through environmental influences, and theories that stress the role of automatic and voluntary mimicking, would predict discrete emotion-specific factors. Because specific environments have a higher probability of eliciting specific emotional displays for certain emotion categories, individuals may become better at expressing some specific emotions compared to others. This would partly explain differences in the rank order of persons for different expression categories.

On the Relation of Facial Expression Production and Their Reception – Exploring the Nomological Net

So far we have reviewed theoretical assumptions and available univariate data that postulate expectations concerning the internal structure of individual differences in behavioral measures of emotion expression abilities related to the facial modality. This leads to an additional important question: how are receptive and productive abilities related within a conceivable taxonomic model of emotional abilities? This question is important for deriving a nomological net around facial expressivity measured as a performance-based construct.

Because encoding and decoding facial information are both crucial aspects of social interactions, they can be considered "opposite sides of the same coin" (Elfenbein and Eisenkraft, 2010, p. 301). Theories and empirical data on their relationship are, however, equivocal, with positive, zero, and negative correlations postulated and reported. Elfenbein and Eisenkraft (2010) summarized the literature that addressed this question and conducted a meta-analysis in the hope to solve the old mystery of the threefold empirical finding. Here, we will briefly summarize these theoretical perspectives and outline the results of the meta-analysis thereafter.

Performance-based emotional intelligence theories and Scherer's theory on emotional competence (Scherer, 2007) postulate a positive relationship between receptive and productive abilities as part of a broader socioemotional communication factor. Furthermore, potential reinforcement might occur between the two sets of abilities. Thus, some individuals may become more expressive because they show stronger mimicking of their counterparts with whom they empathize. Due to their more pronounced empathy, compared to other individuals, they are also better at recognizing emotional displays of others, which is a prerequisite of explicit mimicry (e.g., Osgood, 1966). The theory of emotional contagion also postulates a similar mechanism of reciprocal feedback. Hatfield et al. (2011) reviewed the literature on emotional contagion suggesting that persons tend to automatically mimic the expressions of their counterparts and thereby empathize with the emotions of others, which has the consequence of a better emotion expression recognition triggered by the feedback provided by self-experience due to mimicry. This theory suggests that stronger mimicry behavior leads to more pronounced expressivity that facilitates emotion recognition.

Neurological and psychiatric patient studies of right-brain damaged, Parkinson, schizophrenic, or unipolar depressive individuals also suggest a double dissociation between receptive and productive emotion communication abilities in the facial and vocal modality, as compared with normal controls (e.g., Borod et al., 1990). Halberstadt et al. (2001) advanced a developmental theory of affective social competence that, similarly to Scherer's theory (Scherer, 2007), postulates three dynamically interacting aspects of sending, receiving, and experiencing affect. The authors discuss three developmental disorders characterized by impaired emotional competencies (autism, behavior disorder, and children who suffered from maltreatment) that both receptive and productive aspects of affect communication are impaired in all those cases. Such findings on child populations characterized by emotional malfunctioning also seem to support the view that emotion perception and expression are positively related.

One line of explanations postulating a negative relation between receptive and productive aspects of affective communication emerged from socialization theories advanced earlier by Halberstadt (1986), who argued that children growing up in environments that are low in expressiveness develop a high sensitivity for recognizing affect, even in very subtle displays of emotion, and vice versa. This explanation is embedded into more general theories describing psychological compensation processes across skills (e.g., Salthouse, 1995) and in person–environment interactions. Another line of explanation has been advanced by Zuckerman et al. (1976) who found negative correlations between encoding and decoding, but only for affect displays of different emotion categories, and positive correlations within the same affect class. The authors attempted to explain the negative correlations through mechanisms of interpersonal accommodation, postulating that persons having troubles with perceiving a given affect would assume that other persons face the same difficulty and would seek to express that affect more clearly to facilitate its decoding by their counterparts.

Likewise, Elfenbein and Eisenkraft (2010) advanced possible explanations for a zero correlation between encoding and decoding abilities. They proposed that individual differences in the ability to compensate for family environments might moderate the relationship between encoding and decoding; thus, good compensators may show a negative relationship and those who are less susceptible to compensation would show a positive encoding–decoding correlation. Treating those individuals as one group results in an apparent zero relationship. Furthermore, they listed neurobiological findings suggesting separate brain locations for perceiving and expressing affect. However, there are at least two shortcomings to the brain-behavior explanations. First, the neural networks involved in emotional facial expression seem to partly overlap with brain regions involved in the perception of affect, and second, the neural locations of expressive behavior are still less understood.

In their meta-analysis based on 40 primary studies, Elfenbein and Eisenkraft (2010) explicitly distinguished results for posed versus elicited expressions. They defined elicited spontaneous expressions as being caused by feelings and physiology and posed expressions as being produced through the social intention to communicate states. Overall, the analyses yielded a positive correlation for posed emotion expressions and affect recognition of 0.19 (on the basis of 24 available studies), but no relationship for spontaneous expressions.

However, this positive correlation, which suggests a possible relation between VfEE and decoding abilities, may be

an underestimate due to methodological shortcomings of the available primary studies at the time. These are as follows: the single-task approach used to measure both receptive and productive abilities, the inclusion of a series of neurological and psychiatric patient samples, possibly associated with ceiling effects in recognition tasks for normal controls, and no control for face plasticity and facial expression bias at parameterizing expressive abilities. We argue that a multivariate approach and a more sophisticated modeling of both abilities in a latent variable framework that accounts for task specificity and measurement error will be beneficial to determine the relationship of individual differences between VfEE and decoding abilities by disattenuating it.

Conclusions

Empirical research to date suggests that studying facial emotion expressions is a fruitful line of research for assessing emotion-related abilities. Existing studies suggest the measurement of facial expressions is feasible; however, given the paucity of adequate study designs and measurement tools, much research is needed on this topic. Research produced thus far, however, suggest there are many promising perspectives for this field, including the applied ones. Future research should explore the relation between explicit facial emotion expression and incidential facial mimicry, in addition to their relations with similar constructs. Likewise, facial emotion expression ability and related socioemotional competencies may be deficient in individuals with particular mental disorders (e.g., autism, psychopathy), suggesting these populations provide a unique opportunity to study individual differences in facial emotion expression. Finally, given the importance of expressing emotions facially in everyday communication, and the importance of first impressions in numerous situations (e.g., interviews), research should explore options for training individuals to do better in facial emotion expression, such as producing more intense expressions, or more situationally appropriate expressions.

See also: Emotion and Expression; Emotional Intelligence and Competencies; Emotional Regulation; Facial Perception; Factor Analysis and Latent Variable Models in Personality Psychology; Imitation During Infancy and Early Childhood; Neonatal Imitation; Personality Assessment; Social Intelligence and Competencies; Social and Emotional Development in the Context of the Family.

Bibliography

Allport, F., 1924. Social Psychology. Houghton Mifflin, New York.
Barrett, L.F., 2009. Variety is the spice of life: a psychologist constructionist approach to understanding variability in emotion. Cognition & Emotion 23, 1284–1306.
Barrett, L.F., 2011. Was Darwin wrong about emotional expressions? Current Directions in Psychological Science 20, 400–406.
Borod, J.C., Welkowitz, J., Alpert, M., Brozgold, A.Z., Martin, C., Peselow, E., Diller, L., 1990. Parameters of emotional processing in neuropsychiatric disorders – conceptual issues and a battery of tests. Journal of Communicative Disorders 23, 247–271.
Buck, R., 1979. Measuring individual differences in the nonverbal communication of affect: the slide-viewing paradigm. Human Communication Research 6, 47–57.
Buck, R., 1984. The Communication of Emotion. Guilford Press, New York.
Bühler, K., Goodwin, D.F., 1990. Theory of Language: The Representational Function of Language. John Benjamins, Amsterdam.
Camras, L.A., Lambrecht, L., Michel, G.F., 1997. Infant "surprise" expressions as coordinative motor structures. Journal of Nonverbal Behavior 20, 183–195.
Coan, J.A., Gottman, J.M., 2007. The specific affect coding system (SPAFF). In: Coan, J.A., Allen, J.B. (Eds.), Handbook of Emotion Elicitation and Assessment. Oxford University Press, New York, pp. 106–123.
Cunningham, M.R., 1977. Personality and the structure of the nonverbal communication of emotion. Journal of Personality 45, 564–584.
Darwin, C., 1965. The Expression of the Emotions in Man and Animals. Chicago University Press, Chicago, IL (Original work published 1872).
Dawkins, R., Krebs, J.R., 1978. Animal signals: information or manipulation. In: Krebs, J.R., Davies, N.B. (Eds.), Behavioural Ecology: An Evolutionary Approach. Blackwell Scientific Publications, Oxford, pp. 282–309.
Ekman, P., 1972. Universal and cultural differences in facial expression of emotion. In: Cole, J.R. (Ed.), Nebraska Symposium on Motivation. Nebraska University Press, Lincoln, NE, pp. 207–283.
Ekman, P., 1992a. Facial expressions of emotion: new findings, new questions. Psychological Science 3, 34–38.
Ekman, P., 1992b. Facial expressions of emotion: an old controversy and new findings. Philosophical Transactions of the Woxal Society of London. Series B: Biological Sciences 335, 63–69.
Ekman, P., Friesen, W.V., 1978. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto, CA.
Elfenbein, H.A., Eisenkraft, N., 2010. The relationship between displaying and perceiving nonverbal cues of affect: a meta-analysis to solve an old mystery. Journal of Personality and Social Psychology 98, 301–318.
FACET® 2.1 [Computer Software]. Emotient. Retreived from: http://www.emotient.com/products#FACETVision.
Fernández-Dols, J.M., 1999. Facial expression and emotion: a situationist view. In: Philippot, P., Feldman, R.S., Coats, E.J. (Eds.), The Social Context of Nonverbal Behavior. Cambridge University Press, Cambridge, pp. 242–261.
Fernández-Dols, J.-M., 2013. Advances in the study of facial expression: an introduction to the special section. Emotion Review 5, 3–7.
Frijda, N.H., Tscherkassof, A., 1997. Facial expressions as modes of action readiness. In: Russell, J.A., Fernández-Dols, J.M. (Eds.), The Psychology of Facial Expression. Cambridge University Press, Cambridge, pp. 78–102.
Fridlund, A.J., 1994. Human Facial Expression: An Evolutionary View. Academic Press, San Diego, CA.
Friedman, H.S., Prince, L.M., Riggio, R.E., DiMatteo, M.R., 1980. Understanding and assessing nonverbal expressiveness: the affective communication test. Journal of Personality and Social Psychology 39, 333–351.
Gross, J.J., John, O.P., 1995. Facets of emotional expressivity: three self-report factors and their correlates. Personality and Individual Differences 19, 555–568.
Halberstadt, A.G., 1986. Family socialization of emotional expression and nonverbal communication styles and skills. Journal of Personality and Social Psychology 51, 827–836.
Halberstadt, A., Denham, S.A., Dunsmore, J., 2001. Affective social competence. Social Development 10, 79–119.
Harmon-Jones, E., Amodio, D.A., Zinner, L.R., 2007. Social psychological methods of emotion elicitation. In: Coan, J.A., Allen, J.J.B. (Eds.), Handbook of Emotion Elicitation and Assessment. University Press, Oxford, pp. 91–105.
Hatfield, E., Rapson, R.L., Le, Y.L., 2011. Primitive emotional contagion: recent research. In: Decety, J., Ickes, W. (Eds.), The Social Neuroscience of Empathy. MIT Press, Boston, MA, pp. 19–30.
Hopf, H.C., Muller-Forell, W., Hopf, N.J., 1992. Localization of emotional and volitional facial paresis. Neurology 42, 1918–1923.
Izard, C.E., 1971. The Face of Emotion. Appleton-Century-Crofts, New York.
Keele, S.W., 1968. Movement control in skilled motor performance. Psychological Bulletin 70, 387–403.
Korb, S., Grandjean, D., Scherer, K., 2008. Motor commands of facial expressions: the Bereitschaftspotential of posed smiles. Brain Topography 20, 232–238.
Korb, S., Sander, D., 2009. The neural architecture of facial expression. In: Sander, D., Scherer, K. (Eds.), The Oxford Companion to Emotion and Affective Sciences. Oxford University Press, Oxford, pp. 173–175.
Kornhuber, H.H., Deecke, L., 1965. Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. Pflügers Archiv 284, 1–17.

Kring, A.M., Sloan, D.M., 2007. The facial expression coding system (FACES): development, validation and utility. Psychological Assessment 19, 210–224.
Lee, T.-W., Josephs, O., Dolan, R.J., Critchley, H.D., 2006. Imitating expressions: emotion-specific neural substrates in facial mimicry. Social Cognitive and Affective Neuroscience 1, 122–135.
Matsumoto, D., 2009. Action unit. In: Sander, D., Scherer, K. (Eds.), The Oxford Companion to Emotion and Affective Sciences. Oxford University Press, Oxford, pp. 2–3.
McIntosh, D.N., 1996. Facial feedback hypotheses: evidence, implications, and directions. Motivation and Emotion 20, 121–147.
Morecraft, R., Louie, J.L., Herrick, J.L., Stilwell-Morecraft, K.S., 2001. Cortical innervation of the facial nucleus in the non-human primate: a new interpretation of the effects of stroke and related subtotal brain trauma on the muscles of facial expression. Brain 124, 176–208.
Morecraft, R., Stilwell-Morecraft, K.S., Rossing, W.R., 2004. The motor cortex and facial expression: new insights from neuroscience. The Neurologist 10, 235–249.
Olderbak, S.G., Hildebrandt, A., Pinkpank, T., Wilhelm, O., Sommer, W., 2013. Psychometric challenges and proposed solutions when scoring facial emotion expression codes. Behavior Research Methods. Published online 6 December 2013 at: http://link.springer.com/article/10.3758/s13428-013-0421-3/fulltext.html.
Osgood, C.E., 1966. Dimensionality of the semantic space for communication via facial expressions. Scandinavian Journal of Psychology 7, 1–30.
Parkinson, B., 2013. Contextualizing facial activity. Emotion Review 5, 97–103.
Recio, G., Shmuilovich, O., Sommer, W., 2014. Should I smile or should I frown? An ERP study on the voluntary control of emotion-related facial expressions. Psychophysiology 51, 789–799.
Reisenzein, R., Junge, M., Studtmann, M., Huber, O., 2014. Observational approaches to the measurement of emotions. In: Pekrun, R., Linnenbrink-Garcia, L. (Eds.), International Handbook of Emotions in Education. Taylor & Francis/Routledge, pp. 580–606.
Reisenzein, R., Studtmann, M., Horstmann, G., 2013. Coherence between emotion and facial expression: evidence from laboratory experiments. Emotion Review 5, 16–23.
Rinn, W.E., 1984. The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions. Psychological Bulletin 95, 52–77.
Rottenberg, J., Ray, R.D., Gross, J.J., 2007. Emotion elicitation using films. In: Coan, J.A., Allen, J.J.B. (Eds.), Handbook of Emotion Elicitation and Assessment. University Press, Oxford, pp. 9–28.
Russell, J.A., 2003. Core affect and the psychological construction of emotion. Psychological Review 110, 145–172.
Salthouse, T.A., 1995. Refining the concept of psychological compensation. In: Dixon, R.A., Bäckman, L. (Eds.), Compensating for Psychological Deficits and Declines: Managing Losses and Promoting Gains. Erlbaum, Mahwah, NJ, pp. 21–34.
Scherer, K., 2007. Componential emotion theory can inform models of emotional competence. In: Matthews, G., Zeidner, M., Roberts, R. (Eds.), Science of Emotional Intelligence: Knowns and Unknowns. Oxford University Press, Oxford, pp. 101–126.
Scherer, K., 2009a. Push/pull effects. In: Sander, D., Scherer, K. (Eds.), The Oxford Companion to Emotion and Affective Sciences. Oxford University Press, Oxford, p. 326.
Scherer, K., 2009b. Principles of emotional expression. In: Sander, D., Scherer, K. (Eds.), The Oxford Companion to Emotion and Affective Sciences. Oxford University Press, Oxford, pp. 167–169.
Scherer, K.R., Mortillaro, M., Mehu, M., 2013. Understanding the mechanisms underlying the production of facial expression of emotion: a componential perspective. Emotion Review 5, 47–53.
Shibasaki, H., Barrett, G., Halliday, E., Halliday, A.M., 1980. Components of the movement-related cortical potential and their scalp topography. Electroencephalography and Clinical Neurophysiology 49, 213–226.
Susskind, J.M., Lee, D.H., Cusi, A., Feiman, R., Grabski, W., Anderson, A.K., 2008. Expressing fear enhances sensory acquisition. Nature Neuroscience 11, 843–850.
Tarkka, I.M., Hallett, M., 1991. Topography of scalp-recorded motor potentials in human finger movements. Journal of Clinical Neurophysiology 8, 331–341.
Tomkins, S.S., 1962. Affect, Imagery, Consciousness. Vol. 1. The Positive Affects. Springer, New York.
Walter, W.G., Cooper, R., Aldridge, V.J., McCallum, W.C., Winter, A.L., 1964. Contingent negative variation: an electric sign of sensorimotor association and expectancy in the human brain. Nature 203, 380–384.
Wild, B., Rodden, F.A., Rapp, A., Erb, M., Grodd, W., Ruch, W., 2006. Humor and smiling: cortical regions selective for cognitive, affective, and volitional components. Neurology 66, 887–893.
Zuckerman, M., Hall, J.A., DeFrank, R.S., Rosenthal, R., 1976. Encoding and decoding of spontaneous and posed facial expressions. Journal of Personality and Social Psychology 34, 966–977.
Zuckerman, M., Larrance, D.T., Hall, J.A., DeFrank, R.S., Rosenthal, R., 1979. Posed and spontaneous communication of emotion via facial and vocal cues. Journal of Personality 47, 712–733.

Relevant Websites

http://emr.sagepub.com/content/5/1.toc – Emotion Review journal; January 2013; 5 (1) table of contents.

http://imotionsglobal.com/software/add-on-modules/attention-tool-facet-module-facial-action-coding-system-facs/ – Attention Tool – Emotient Module: Facial Expression Analysis.

http://www.noldus.com/human-behavior-research/products/facereader – Facial expression recognition software: FaceReader.

Facial Perception

Oliver Wilhelm, University Ulm, Ulm, Germany
Andrea Hildebrandt, Ernst-Moritz-Arndt University Greifswald, Greifswald, Germany
Werner Sommer, Humboldt-Universität zu Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, we summarize popular theories of face identity and facial expression processing. Next, we present some general findings on individual differences in this field including considerations regarding more common and specific face identity and facial expression processing assessment principles. We then discuss available measures and present a task-taxonomy, followed by a short summary of key findings in the field. We then present a summary of research results concerning neural underpinnings in identity and emotion processing. Finally, we conclude with some remarks about the emerging field of identity and expression investigations and some research heuristics.

Faces provide indispensable information for human communication. They inform the perceiver about the familiarity and identity of others and about those persons' emotional states and intentions. The mechanisms of facial information processing and the respective neural correlates received wide attention in the last years. Most of this research is based on experimental findings in healthy participants and evidence from persons with innate or acquired disorders of face processing. This article will focus on variability in these abilities across persons and the structure of individual differences in facial information processing, after providing a summary of theories built upon experimental and clinical research.

Theories of Face Identity and Facial Expression Processing

Functional models of facial information processing (e.g., Haxby et al., 2000; Young and Bruce, 2011) postulate a hierarchical structure, including view-centered pictorial and view-independent structural encoding of the face, access to stored representations of known faces in memory, and to associated (biographical) knowledge about the person and his or her name. In addition to identity processing, these models allow predictions about how the face processing system deals with variable information provided by faces – in particular, emotional expressions, facial movements accompanying speech, and gaze direction. Further, socially relevant information that can be derived from the invariant face structure independent of the person's identity – i.e., age, gender, and judgments about attractiveness – is mapped in these models as visually derived semantic codes. Early theorizing assumed independent streams of identity versus expression-related information. This initial idea was recently modified in favor of partial dependence views (e.g., Calder, 2011; Young and Bruce, 2011). However, it is not yet fully clarified how these systems interact.

Research on the general physiological implementation of face-based processes distinguishes between a core system and an extended system (Haxby et al., 2000). The core system consists of three bilateral, hierarchically organized brain regions. During initial stages of face perception, facial features are processed in the inferior occipital gyri. The inferior occipital gyri provide input to the lateral fusiform gyrus, where invariant aspects of faces – arguably cues of unique identities – are processed. This region has been found to respond more strongly to the perception of faces than to the perception of nonface objects (Kanwisher, 2000). Based on this and similar findings, the so-called fusiform face area (FFA), situated in the fusiform gyrus, has been postulated to be a specialized module for face perception. If attention is directed to changeable aspects of faces, like emotion expression or lip movement, the superior temporal sulci are activated. Visually derived semantic codes, biographical, and name information are processed in the extended system (Haxby et al., 2000). Beyond the core system, an extended neural system – specialized on different cognitive functions – is considered to interact with face-responsive regions (Haxby et al., 2000). We highlight four such systems. First, the intraparietal sulcus is considered to be in charge of processing spatially directed attention and thought to aid the perception of eye gaze direction. Second, the audiovisual integration in speech perception (see the McGurk effect), which is supported by the interplay of the auditory and superior temporal cortex. Third, during emotion processing the superior temporal sulcus maintains connections with the amygdala, insula, and limbic system, which are the neural systems involved in the processing of emotion. Fourth, access to semantic and biographical information during face processing appears to be mediated by anterior temporal regions. Taken together, face processing is supported by multiple brain regions that are spatially disparate, yet act together in a coordinated way. The mechanisms of facial information processing are supposed to be hierarchically organized and actual models assume an interplay between the identity and expression processing routes. Furthermore, Haxby et al. (2000) emphasize that brain systems for facial information processing do not work in isolation but contribute to other cognitive functions such as object recognition.

Individual Differences in Face Identity and Expression Processing

Neurocognitive theories of face identity and facial expression processing describe general mechanisms and the functional organization of the facial information processing system. The

 http://dx.doi.org/10.1016/B978-0-08-097086-8.25060-5

taxonomy, stability, and predictive validity of individual differences in these cognitive mechanisms have been rather recently – at least partly – addressed as a set of independent research questions.

Given the neuroanatomic evidence on at least partly distinct brain localizations of core aspects of face identity and facial expression processing one might be tempted to infer that the abilities to process neutral and emotional faces are highly distinct and independent. However, all of the proposed tasks for measuring facial information processing have attributes of prototypical ability tests: Participants are willing to do their best, there is an objective correct response, and performance is assessed as degree to which responses are correct. It is well known that any two such measures will covary positively in unselected samples – a phenomenon labeled positive manifold in ability research. Therefore, it is unrealistic to expect that (1) measures of face perception that capture individual differences will be completely independent of general cognitive abilities measures and (2) measures of facial emotion processing will be orthogonal to measures of face identity processing. Additionally, when taking a closer look at emotion recognition tasks Adolphs (2002) concluded that emotion identification may be carried out as a combination of processes reflecting perception, semantic knowledge, and reasoning. The independence and redundance of performance in a variety of measures is best investigated by studying their factorial structure.

How do we understand the emotional states of others? Lipps (1907) suggested that persons imitate (mimic) observed affective behavior and that this mimicry elicits a corresponding state in the observer, facilitating or enabling affect understanding. This idea is echoed in modern embodied simulation theories. Indeed, facial mimicry of emotional expressions is frequently observed, although in some cases only with the aid of electromyographic recordings. Mimicry reactions are often involuntary, automatic, and occur within a fraction of a second (Dimberg et al., 2000). The idea that simulation is fundamental for the perception of emotions in others is in line with findings that hindering incidental mimicry leads to performance impairment in emotion perception (e.g., Neal and Chartrand, 2011). However, whether persons who show more incidental mimicry are also better emotion perceivers and whether they do so in all situations is less clear. Hess and colleagues did not observe such a relationship between mimicry and emotion perception in young adults (e.g., Hess and Blairy, 2001). Therefore, facial mimicry may not be necessary for the classification of prototypical emotional displays but may be helpful for social binding or when emotion recognition is difficult. In line with the latter view, Künecke et al. (2014), using a multivariate approach, showed a positive relationship between the degree of facial corrugator response and the ability of emotion perception.

In the popular four-branch theory of emotional abilities, face-based emotion processing is included as one of the key components of emotional intelligence (Mayer et al., 2008). In the component process models (Scherer, 2007) face-based emotion processing is included as receptive decoding. Wilhelm and colleagues (2010) claimed that face identity perception and recognition are core aspects of such emotional abilities. If emotional intelligence exists (Borsboom et al., 2004) individual differences in perceiving neutral and emotional faces ought to be considered lower order factors in a higher order model of socioemotional abilities. Based upon the process theories delineated in the introduction facial information-related decoding abilities may be further subdivided into perceptual and mnestic aspects of performance. Additionally, it is not yet settled whether or not the perception of neutral and emotional faces needs to be distinguished on an individual differences level. Importantly, broader theories of socioemotional abilities that postulate a series of emotion receptive performance components ought to include more or less specific abilities of processing identity and emotion in the voice, gestures, and postures. Obviously, systematically deriving ability measures of such components that are grounded in basic research and incorporating them into factor models of socioemotional abilities is a major challenge for future research.

Apart from distinguishing perceptual and mnestic demands of tasks challenging either identity or expression processing of faces, there is one more crucial task attribute requiring consideration: whether performance is expressed as speed or accuracy. In experimental research many of the tasks used to provoke face identity and facial expression processing are easy to solve. Comparisons between experimental conditions are then usually based on the latency of responses. Within ability research such measures would be considered processing speed tasks because participants from the application population solve a very high proportion of trials correctly and individual differences are primarily manifest in the response latency. It is important to note that easy tasks and hard tasks usually load on different ability factors. Latency-related ability factors may or may not arise as a function of the information content. Indeed, measures challenging the speed of object processing are essentially perfectly collinear with measures challenging the speed of face identity and facial expression processing (Hildebrandt et al., 2012, 2013). This is not true for measures of face identity and facial expression processing scored by the accuracy of responses (Wilhelm et al., 2010). Given the distinction between speed and accuracy tasks that is well established in ability research and the high collinearity of speed-related abilities that lack facets of identity or emotion processing, we recommend avoiding measures of face identity and facial expression processing exclusively challenging the speed of performance.

After narrowing down the face identity and expression processing ability domains based on well-funded experimental research, it is instrumental to consider prior work on socioemotional abilities from other fields of individual differences research. The so-called jingle fallacy refers to instances in which constructs are deemed the same, although they only share a similar verbal label. Conversely, the jangle fallacy refers to instances in which constructs are deemed different although the only obvious distinction is the verbal label established within a scientific community. The recently popularized *theory of mind* and measures arguably capturing individual differences in this ability are probably an instance of a jangle fallacy. The *Reading the Mind in the Eyes* Test displays photographs of the eye region of 36 actors and actresses. Participants are to select which of several response options best matches what those persons think or feel. Obviously, the stimuli and the task presented here are akin

to provoking facial expression processing. It is therefore important to solidly clarify the status of individual differences in such *theory of mind* tasks. Presupposing adequate psychometric characteristics of the measures, we would predict very high correlations between theory of mind and emotion identification (i.e., a jangle fallacy). In that case, the theory of mind ability may not need distinct theorizing.

Measures of Face Identity and Facial Expression Processing

One way to investigate the status quo of individual differences research in this area is to go through available ability measures and search for tasks that use faces as stimuli. We will describe a few such measures – beginning with tasks challenging identity processing and proceeding to measures of facial expression processing – and highlight their positive and negative attributes. It is important to note that we restrict the presentation here to protypical ability measures; meaning the behavior of participants is assessed by comparing responses to an objective standard of performance.

In the famous faces test, participants are asked to pick the correct name out of a few alternatives for each famous face shown. As a measure of identity recognition, this type of test is conceptually problematic because it neglects differences between persons in prior exposure to the face stimuli and knowledge about the displayed famous persons. In the Warrington Recognition Memory for Faces Test no control concerning stimulus attributes irrelevant for face cognition, like clothing, hair, or paraphernalia such as glasses, is applied. Therefore, these irrelevant stimulus attributes might be used to solve the task. Presumably, this is specifically true for participants with severe problems in processing faces, such as prosopagnosic patients. In the Benton Facial Recognition Test, participants are presented a target photo and are asked to select the target identity from six simultaneously presented response alternatives. In some items participants need to identify an identical picture, in others they have to pick the correct identity when pictures are taken from different angles or in different lighting conditions. Although stimulus control is much better here, participants can (and do) still rely on eyebrows and hair lines in processing faces. The objections raised above also apply to most of the available measures of socioemotional abilities and learning and recognition of face-tasks included in many intelligence and memory test batteries. Additionally, in many of these tasks components unrelated to identity and expression processing contribute to performance.

In order to solve issues with face external stimulus components, Duchaine and Nakayama (2006) developed the Cambridge Face Perception Task (CFPT) and the Cambridge Face Memory Test (CFMT). These tests include only face internal features and the perception task successfully minimizes mnestic demands. The CFPT applies morphing: the creation of new faces through blending two different faces into each other. Target faces are morphed with several other faces varying the similarity of the morphed faces with the target face. The target face is presented along with the morphs and participants are asked to order the morphs according to their similarity with the target face. In the CFMT, participants are asked to learn and – after a very brief retention interval – recognize pictures of six faces in different conditions (identical, different angles, different lighting, and degraded through visual noise). Both of the Cambridge tests are certainly superior to many earlier measures.

Despite this somewhat mixed assessment concerning ability tasks relying on neutral faces as stimuli, recent developments in the field are encouraging. On the one side, we recently presented a task battery that allows measuring individual differences in perceiving and recognizing neutral faces and were able to establish and replicate the distinction of these abilities from object recognition and established ability factors (Wilhelm et al., 2010). Within the task battery, an additional aim was to distinguish between tasks challenging either speed or accuracy of processing – a prominent distinction in intelligence research outlined above (Carroll, 1993). The multitask approach for measuring face identity processing-related abilities presented by Wilhelm et al. (2010) includes task paradigms widely used in experimental research on face processing (Part–Whole, Composite Faces, and the Inversion paradigms), different memory paradigms (delayed nonmatching, immediate recall, delayed recall, implicit memory), and over 500 face identities across all tasks. The factorial structure of the tests has been established across a series of independent samples.

With respect to facial expression processing, many measures have an intrinsic component of perceptual speed because stimulus exposition is limited in time (i.e., Japanese and Caucasian Brief Affective Recognition Test (Matsumoto et al., 2000)). The popular Diagnostic Analysis of Nonverbal Accuracy (DANVA; Nowicki and Carton, 1993), which challenges recognition of happiness, sadness, anger, and fear in emotional faces, also has a strictly time-limited stimulus exposition. The manipulations of expression intensity in the DANVA would probably allow the derivation of accuracy scores to express performance. However, the tasks have not yet been probed as measures of accuracy.

The *Profile of Nonverbal Sensitivity* (Rosenthal et al., 1979) is another popular measure, which uses recordings of faces, voices, and bodies as stimuli to assess emotion recognition from multimodal and dynamic stimuli. The use of dynamic facial stimuli is an interesting aspect of the measure but an important drawback here is the limitation to only two emotion categories (positive versus negative).

One out of several measures designed for the assessment of potential disorders of affect processing is the *Florida Affect Battery* (Bowers et al., 1999), which uses static facial, vocal, and cross-modal stimuli and multiple methods (discrimination, naming, selection, and matching). This test is arguably too easy for unimpaired subjects.

The *Multimodal Emotion Recognition Test* relies on 10 categories of emotion expressions but the measure is limited to affect naming as a single observational method. The *Geneva Emotion Recognition Test* (Schlegel et al., 2014) features dynamic and multimodal actor portrayals of 14 emotions – amusement, anger, disgust, despair, pride, anxiety, interest, irritation, joy, fear, pleasure, relief, surprise, and sadness – but exclusively relies on affect naming.

Two tasks challenging perceptual matching and explicit labeling of expression with satisfying psychometric background were recently presented (Palermo et al., 2013). Both tasks build upon the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise). The first task applies an odd-man-out paradigm (i.e., participants have to indicate which out of three simultaneously presented pictures displays a different emotion). The second task requires participants to explicitly identify the emotion in a target face by selecting one of six written labels.

The validity and utility of highly specific information processing paradigms like the *Emotional Stroop Test*, for example, (Matthews et al., 2003) is still unclear.

Most of the above approaches do not allow abstracting from task-specific variance – a crucial step toward studying latent variables (ability estimates) rather than manifest raw scores. Unfortunately, multivariate approaches are more often an exception than the rule in the field. Additionally, stimulus sampling seems to be particularly relevant when creating items and compiling tests (Judd et al., 2012). Very recently Wilhelm et al. (2014) proposed a task battery for measuring facial expression perception and recognition. The same distinction of speed versus accuracy-based measures for face identity processing has also been implemented for the expression processing tasks. The battery includes five tasks for measuring facial expression perception and identification that are based on the composite paradigm, the face inversion paradigm, a measure of emotion discrimination in morphed continua between facial expressions that share action units, and a visual search paradigm with emotional expressions of different intensity. The battery also includes recognition measures that successfully avoid confusing identity and emotion processing and applies different pictures in the learning versus recognition stages. In order to avoid biases arising through applying the same methods for performance assessment repeatedly, multivariate achievement tests should vary memory cues, viewing angles, expression intensity, and the like.

Summary of Key Findings

Theories of face-based receptive communicative abilities – as reviewed above – are mainly originating from the cognitive-experimental literature and were investigated primarily with experimental, neurophysiological, and neurobiological data. However, the structure and predictive validity of mental functions such as face perception, face memory, emotion perception, and emotion memory can only be determined by studying individual differences in a variety of adequate tasks. Overarching theories based on such data – for example, the four-branch theory of emotional intelligence (e.g., Mayer et al., 2008) – are conceptually and theoretically controversial, mainly because their grounding in more basic science is insufficient. Based on the theoretical distinctions and models reviewed above, we postulate a taxonomic model of individual differences in facial information processing that is – arguably – supported by available data.

Following the well-established distinction in the field, a first differentiation of emotional abilities is between receptive (decoding) and productive (encoding) performance. A second critical distinction is between speed performance (i.e., latencies of correct responses) and the accuracy of performance (i.e., proportion of correct responses in tasks that are so hard that some participants cannot solve them regardless of how much time is available). A third important categorization refers to predominant perceptual or mnestic demands of tasks.

Several larger studies critically test these distinctions. For example, based on two larger studies Wilhelm et al. (2010) provided evidence for three factors of face cognition: (1) the accuracy of face perception, (2) the accuracy of face memory, and (3) the speed of perceiving and recognizing unfamiliar faces. These factors were replicated in an independent sample of younger adults with new identities as face stimuli (Hildebrandt et al., 2013). These studies show a clear distinction between speed and accuracy factors, which were nearly completely independent of each other. The shared variance between the performance accuracy and speed of face cognition was less than 5%. This finding has a strong implication for theories and research on face processing, because it shows that research outcomes can tell completely different stories, depending on whether speed or accuracy is analyzed. Additionally, these studies show unequivocally that perception of and memory for faces are distinguishable abilities on the basis of accuracy indicators, but not on the basis of speed indicators. The expectation of a distinction between perception and memory was based on the assumption stressed in face processing models, that at a perceptual stage structural encoding of faces is central, whereas at a memory stage, the access to face representations stored in memory is critical. Obviously, memory performance will partly depend on successful perception. Interestingly, such a perception versus memory differentiation is not feasible on the basis of speed data. Individual differences in the swiftness of recognizing well-learned faces are perfectly collinear with face perception speed. In order to further clarify the status of the speed factor of face cognition, the convergent validity of this factor was studied. On the level of latent variables the speed of face cognition was not distinct from the speed of processing facial emotional expressions and both facial speed factors could not be distinguished from the speed of object cognition. Therefore, the swiftness of performance in these three domains has zero specificity (Hildebrandt et al., 2013).

With respect to the two factors that describe and explain performance in the perception and memory accuracy tasks there is strong and replicated evidence that they need to be dissociated from general cognition, object cognition, immediate and delayed memory for words, numbers and figures, and mental speed. In several samples of young adults, a large amount of variance that is specific for the face cognition factors was replicated. In the first of these studies there was 46% face perception specific variability and 52% face memory specific variability. In an additional study (Hildebrandt et al., 2011), covering an age range from 18 to 88 years, there was no moderator effect of age on the relationship between face perception and face memory versus general cognition, memory, and mental speed. Therefore, this study shows the absence of aging-related dedifferentiation between face cognition on one side and several factors of general cognitive functioning on the other side. Thus, the specificity of face cognition did

not diminish across adult age, despite the expected substantial age-related performance declines. Taken together these results support the assumption that individual differences in face cognition constitute abilities with substantial uniqueness across the adult range.

Conceptually, face perception and face memory can be considered necessary but not sufficient for emotion perception (identification) and memory for emotional expressions. Their empirical distinctness is an essential question to be addressed. Contemporary neurocognitive models of face processing describe partial overlap between the identity and expression processing routes. Recently, Palermo et al. (2013) established correlational evidence in favor of a common first step of identity and expression processing. A factorial model that approximates models of facial information processing may have a nested structure. A theoretical model is depicted in **Figure 1**. To simulate the hierarchical structure of processing face-related information – differentiated through successive processing stages in neurocognitive models – the model first postulates a general factor of face perception. This general factor captures interindividual variability of face perception tasks and accounts for face perception-related variability in all other indicator tasks measuring face memory and facial emotion processing. Second, to represent a partial overlap of the routes processing face identity versus expression that also holds at later stages, the factorial model includes a broad face memory factor that accounts for variability in all memory tasks. Importantly, this factor is purified from variance due to perceptual processing. Two further more specific factors account for variability due to processing emotion-related information only.

The model (1) implies a partial dependency of emotion processing from identity processing and (2) allows for specificity of emotion processing over and above identity cognition. However, the degree of specificity of emotion processing-related factors partialled for face identity processing-related variance is still an unanswered research question. The postulated model needs to be tested on the basis of accuracy data because speed measures of facial emotion processing and the speed of face cognition are perfectly collinear with speed measures lacking interpersonally relevant content.

Once a nontrivial degree of specificity of emotion processing-related factors is established, the emotion specificity is an important research question. It will be interesting to see how strongly indicators of performance in different emotion categories (happiness, surprise, fear, sadness, disgust, and anger) will be related with each other. Although it has been argued that persons may differ in their accuracy to identify specific emotional expressions, because persons may be predominantly faced with different affects during early socialization (e.g., Tomkins and McCarter, 1964), first evidence suggests a lack of emotion category specificity of individual differences (Schlegel et al., 2014). Multivariate research designs will be able to provide more conclusive evidence concerning this question.

Neural Underpinnings of Individual Difference in Face Cognition

The neural mechanisms underlying individual differences in face cognition may provide a deeper understanding of why people differ in these abilities. Performance measures used to derive ability factors are end products of the information processing occurring within the neurocognitive system. In contrast, neuroscientific methods can provide information about the success, organization, and timing of intermediate processing steps that lead to the manifest, end products of performance.

Neuroimaging studies on individual differences in face processing-related abilities are surprisingly rare and confined to small samples of usually not more than 30 participants. The responsivity in bilateral anterior temporal poles and right anterior hippocampus correlated positively with self-rated preexperimental familiarity of face stimuli. Rotshtein et al. (2007) reported that the behavioral impact of second-order (configural) information of faces correlated with the activation of bilateral inferior occipital gyrus and right fusiform gyrus. Both studies rely only on small samples. For a mixed group of normal and prosopagnosic participants Furl et al. (2011) reported that better performance in facial identity processing was positively associated (1) with face

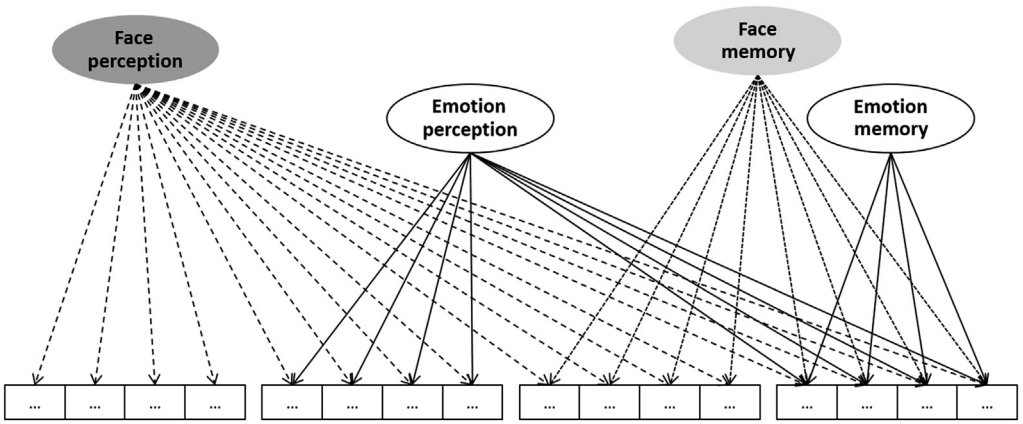

Figure 1 Schematic theoretical model of facial identity and emotion processing; ovals represent latent variables and squares depict indicators used to measure the latent variables.

selectivity in fusiform gyri bilaterally and left anterior temporal lobe, (2) with peak face selectivity in the right and left individually defined FFA, and (3) with the size of the right FFA.

Efforts to tap the neurocognitive mechanisms underlying face cognition in large samples using multivariate modeling are currently restricted to the analysis of event-related brain potentials (ERPs). ERPs are complex responses of the brain to discrete events, such as the presentation of a stimulus, and consist of distinct components that can be related to specific cognitive subprocesses. Amplitudes and latencies of ERPs reflect the intensity and the timing of the processes reflected in these components. Therefore, it is of interest whether individual differences in specific ability factors are related to amplitudes and latencies of specific performance-relevant ERPs. Of particular interest here are ERPs that have been related to the putative stages of face processes as postulated in the functional models outlined above.

The occipital P100 is related to the processing of domain-general, low-level stimulus features and is characterized by a positive peak at occipital scalp sites around 75–135 ms; it is thought to be generated in extrastriate visual brain areas. The N170 is characterized by a negative peak at occipitotemporal sites around 150–190 ms, typically larger for faces than for other objects, but still unaffected by the familiarity of faces. Therefore, it is considered to reflect the configural encoding of facial features and their integration into a holistic percept. Various brain areas have been proposed as sources of the electrical N170, including fusiform, lingual, or occipitotemporal gyri. The Early Repetition Effect (ERE) or N250r is usually operationalized as the difference wave between ERPs to primed and unprimed faces in priming tasks; it is most pronounced at temporal-parietal sites around 260–330 m and presumably originates in the fusiform gyrus. The ERE and Late Repetition Effect (LRE) (see below) are more pronounced for familiar as compared to unfamiliar faces and for personally familiar than for famous faces. Hence, the ERE is associated with the activation of structural representations of faces in long-term memory and with the identification of familiar faces (e.g., Schweinberger and Burton, 2011). Finally, the LRE or N400 – also typically observed in priming experiments – has been related to the activation of person-related knowledge in long-term memory (e.g., Schweinberger and Burton, 2011). It is characterized as the difference wave between ERPs to primed and unprimed faces, usually consisting of a parietal positivity between 330 and 500 m. According to the model for familiar face recognition (see above) the neural generators of the LRE are considered to be part of the extended neural system involved in the retrieving of person-related knowledge (i.e., anterior temporal cortex, precuneus, posterior cingulate cortex).

Kaltwasser et al. (2014) investigated how these ERP components are related with face cognition abilities and with general cognitive functioning. All ERP components could be modeled as unidimensional reliable latent variables. Several such experimentally independent ERPs were only moderately correlated, suggesting that they ought to be considered indicators of specific, essentially independent neurocognitive processes. Shorter latencies of the N170 component along with shorter latencies and larger amplitudes of the ERE and LRE for learned faces have been associated with better face cognition performance. The amplitude of the ERE for learned faces was the strongest predictor of face cognition. The behavior versus ERP latency associations were not face specific, because they diminished after partialing general cognitive functioning.

A few studies investigated the neural underpinnings of individual differences in emotional facial expressions. Kim et al. (2003) reported that more intense subjective experience of surprise is correlated with stronger amygdala activation as measured in fMRI. Using an extreme group design Corden et al. (2006) found less activation of amygdala, fusiform, and anterior temporal cortex in response to neutral faces for participants scoring low on a fear recognition task as compared to a group with normal scores in this task. An ERP study by Tamamiya and Hiraki (2013) found that N170 amplitude predicted the recognition accuracy of happy and angry faces.

Conclusions

Research on face identity and facial expression processing is a rapidly moving field with exciting new developments in many areas. These developments include elaborations on neurocognitive models into the more applied field of individual differences. One consequence of this elaboration is that more and better measurement instruments will be available for research purposes and applied assessment. Research results emerging from this field also have the potential to dramatically improve our current concepts of socioemotional abilities. In opposition to ex-cathedra ideas about the status and organization of socioemotional abilities, an approach grounded in neurocognitive work on identity and emotion processing is obviously a more solid foundation of psychological research. Based upon such a foundation, a variety of intriguing research questions can be proposed. Some of these questions concerning face identity and facial expression processing not addressed in the present contribution include behavior genetics, molecular genetics, sex differences, relevance for a variety of mental disorders, developmental trajectories in childhood, youth, and aging, training and interventions based upon identities and expressions, and many more. Research in all of these areas hinges upon adequate measurement instruments. Due to the relevance of these instruments, we would like to conclude with statements that might seem evident, obvious, and somewhat boring but nevertheless have not been adhered to sufficiently in the past.

The unfolding individual differences research agenda in the field of face identity and facial expression processing is best advised to adhere to the following simple heuristics: Human abilities should be measured with ability indicators. Self-report measures are not ability indicators. Therefore, self-report measures should not be used to study socioemotional abilities. Theoretical claims about abilities should be based on evidence that shows adequate coherence and dependability within measures. Several such measures designed to capture the same ability should show coherence and dependability across measures. Once combinations of established abilities are partialed from such a set of measures a nontrivial amount of residual variance should be present.

This residual variance should predict criteria or outcomes of relevance. The presently available evidence summarized in this contribution endorses the idea that some of the proposed socioemotional ability factors are distinct from established ability factors. This is no small achievement given the fate of many initially intriguing but soon abandoned ideas brought forward in ability research.

See also: Emotion and Expression; Emotion, Perception and Expression of; Emotional Intelligence and Competencies; Emotional Regulation; Emotions, Sociology of; Emotions: Methods of Assessment; Facial Emotion Expression, Individual Differences in; Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis; Factor Analysis and Latent Variable Models in Personality Psychology; Personality Assessment; Personality, Trait Models of; Social Intelligence and Competencies; Social and Emotional Development in the Context of the Family.

Bibliography

Adolphs, R., 2002. Neural systems for recognizing emotion. Current Opinion in Neurobiology 12, 169–177.

Borsboom, D., Mellenbergh, G.J., van Heerden, J., 2004. The concept of validity. Psychological Review 111, 1061–1071.

Bowers, D., Blonder, L.X., Heilman, K.M., 1999. Florida Affect Battery, A Manual. Centre for Neuropsychological Studies, Cognitive Science Laboratory, University of Florida.

Calder, A.J., 2011. Does facial identity and facial expression recognition involve separate visual routes? In: Calder, A.J., Rhodes, G., Johnson, M.H., Haxby, J.V. (Eds.), The Oxford Handbook of Face Perception. Oxford University Press, Oxford, pp. 427–448.

Carroll, J.B., 1993. Human Cognitive Abilities: A Survey of Factor-Analytic Studies. Cambridge University Press, Cambridge, United Kingdom.

Corden, B., Critchley, H.D., Skuse, D., Dolan, R.J., 2006. Fear recognition ability predicts differences in social cognitive and neural functioning in men. Journal of Cognitive Neuroscience 18, 889–897.

Dimberg, U., Thunberg, M., Elmehed, K., 2000. Unconscious facial reactions to emotional facial expressions of emotions. Psychological Science 11, 86–89.

Duchaine, B., Nakayama, K., 2006. The Cambridge Face Memory Test: results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. Neuropsychologia 44, 576–585.

Furl, N., Garrido, L., Dolan, R.J., Driver, J., Duchaine, B., 2011. Fusiform gyrus face selectivity relates to individual differences in facial recognition ability. Journal of Cognitive Neuroscience 23, 1723–1740.

Haxby, J.V., Hoffman, E.A., Gobbini, M.I., 2000. The distributed human neural system for face perception. Trends in Cognitive Sciences 4, 223–233.

Hess, U., Blairy, S., 2001. Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy. International Journal of Psychophysiology 40, 129–141.

Hildebrandt, A., Wilhelm, O., Schmiedek, F., Herzmann, G., Sommer, W., 2011. On the specificity of face cognition compared to general cognitive functioning across adult age. Psychology and Aging 26, 701–715.

Hildebrandt, A., Schacht, A., Sommer, W., Wilhelm, O., 2012. Measuring the speed of recognising facially expressed emotions. Cognition and Emotion 26, 650–666.

Hildebrandt, A., Wilhelm, O., Herzmann, G., Sommer, W., 2013. Face and object cognition across adult age. Psychology & Aging 28, 243–248.

Judd, C.M., Westfall, J., Kenny, D.A., 2012. Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. Journal of Personality and Social Psychology 103, 54–69.

Kaltwasser, L., Hildebrandt, A., Recio, G., Wilhelm, O., Sommer, W., 2014. Neurocognitive mechanisms of individual differences in face cognition: a replication and extension. Cognitive, Affective, and Behavioral Neuroscience 14, 861–878.

Kanwisher, N., 2000. Domain specificity in face perception. Nature Neuroscience 3, 759–763.

Kim, H., Somerville, L.H., Johnstone, T., Alexander, A.L., Whalen, P.J., 2003. Inverse amygdala and medial prefrontal cortex responses to surprised faces. Neuroreport 14, 2317–2322.

Künecke, J., Hildebrandt, A., Recio, G., Sommer, W., Wilhelm, O., 2014. Facial EMG responses to emotional expressions are related to emotion perception ability. PLOS ONE 9 (1): e84053.

Lipps, T., 1907. Das Wissen von fremden Ichen. In: Lipps, T. (Ed.), Psychologische Untersuchungen. Engelmann, Leipzig, pp. 694–722.

Matthews, G., Roberts, R.D., Zeidner, M., 2003. Development of emotional intelligence: a skeptical- but not dismissive perspective. Human Development 46, 109–114.

Matsumoto, D., LeRoux, J., Wilson-Cohn, C., Raroque, J., Kooken, K., Ekman, P., Yrizarry, N., Loewinger, S., Uchida, H., Yee, A., Amo, L., Goh, A., 2000. A new test to measure emotion recognition ability: Matsumoto and Ekman's Japanese and Caucasian Brief Affect Recognition Test (JACBART). Journal of Nonverbal Behavior 24, 179–209.

Mayer, J.D., Roberts, R.D., Barsade, S.G., 2008. Human abilities: emotional intelligence. Annual Review of Psychology 59, 507–536.

Neal, D.T., Chartrand, T.L., 2011. Embodied emotion perception: amplifying and dampening facial feedback modulates emotion perception accuracy. Social Psychological and Personality Science 2, 673–678.

Nowicki, S., Carton, J., 1993. The measurement of emotional intensity from facial expressions; the DANVA FACES 2. Journal of Social Psychology 133, 749–751.

Palermo, R., O'Connor, K.B., Davis, J.M., Irons, J., McKone, E., 2013. New tests to measure individual differences in matching and labelling facial expressions of emotion, and their association with ability to recognise vocal emotions and facial identity. PLoS ONE 8, e68126.

Rotshtein, P., Geng, J.J., Driver, J., Dolan, R.J., 2007. Role of features and second-order spatial relations in face discrimination, face recognition, and individual face skills: behavioral and functional magnetic resonance imaging data. Journal of Cognitive Neuroscience 19, 1435–1452.

Rosenthal, R., Hall, J.A., DiMatteo, M.R., Rogers, P.L., Archer, D., 1979. Sensitivity to Nonverbal Communication: The PONS Test. John Hopkins University Press, Baltimore.

Schlegel, K., Grandjean, D., Scherer, K.R., 2014. Introducing the Geneva emotion recognition test: an example of Rasch-based test development. Psychological Assessment 26, 666–672.

Scherer, K., 2007. Componential emotion theory can inform models of emotional competence. In: Matthews, G., Zeidner, M., Roberts, R.D. (Eds.), Science of Emotional Intelligence: Knowns and Unknowns. Oxford University Press, Oxford, pp. 101–126.

Schweinberger, S.R., Burton, A.M., 2011. Person perception 25 years after Bruce and Young (1986): an introduction. British Journal of Psychology 102, 695–703.

Tamamiya, Y., Hiraki, K., 2013. Individual differences in the recognition of facial expressions: an event-related potentials study. PLoS ONE 8.

Tomkins, S., McCarter, R., 1964. What and where are the primary affects? Some evidence for a theory. Perceptual and Motor Skills 18, 119–158.

Wilhelm, O., Hildebrandt, A., Manske, K., Schacht, A., Sommer, W., 2014. Test battery for measuring the perception and recognition of facial expressions of emotion. Frontiers in Psychology, Emotion Science 5, 404.

Wilhelm, O., Herzmann, G., Kunina, O., Danthiir, V., Schacht, A., Sommer, W., 2010. Individual differences in face cognition. Journal of Personality and Social Psychology 99, 530–548.

Young, A.W., Bruce, V., 2011. Understanding person perception. British Journal of Psychology 102, 959–974.

Relevant Websites

http://www.faceblind.org/ – Prosopagnosia Research Centers at Harvard University and University College London.

http://www.face-rec.org/ – Face Recognition Homepage.

http://scienceaid.co.uk/psychology/cognition/face.html – Science aid: face recognition.

http://ukcatalogue.oup.com/product/9780199559053.do – Oxford Handbook of Face Perception.

Factionalism

Jeremy Boissevain, University of Amsterdam, Amsterdam, The Netherlands

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 8, pp. 5236–5239, © 2001, Elsevier Ltd.

Abstract

Factionalism refers to competition between coalitions or groups – factions - within a larger organization or community. Factions compete for resources which provide power chances, such as economic assets, office, honor, ideology and behavioral norms. Since unity and consensus are widely held organizational and community ideals, factions are generally perceived as self-interested and disruptive and members condemned as disloyal. The central focus of a faction is the leader, who has recruited its members. Ties between leader and followers are usually personal, although followers may recruit others on behalf of their leader. Anthropologists first examined factionalism as a structural phenomenon in the late 1950s. It was then perceived as resulting from rapid social change. Factions were regarded as symmetrically organized conflict groups in dynamic equilibrium, a fundamental tenet of functionalism. During the 1970s, reflecting theoretical developments, belief in dynamic equilibrium was gradually abandoned. Competing factions came to be seen as asymmetrical, sometimes manifesting class or ethnic bias and differing in access to resources and strategy, tactics, internal organization and ideology. Factions that persist over time may evolve into corporate groups. Factionalism takes many forms and has been observed in all parts of the world. It occurs in most organizations and communities. It is an intrinsic part of political life.

Factionalism refers to dissension between rival subgroups – factions – within a larger social unit. Factionalism can take many forms and has been observed in all parts of the world. It is a basic political process dynamically related to social change.

Factions and Organizations

Factionalism, regardless of where it is classified on the conflict continuum, is conflict between factions. Factions are coalitions of persons or subgroups that compete over specific issues within a larger organization or community. The central focus of a faction is the leader who coordinates its activities and recruits its members. Ties between leader and followers are usually personal, although some followers may recruit others on behalf of the leader. The issues about which factions compete are diverse. They generally concern scarce resources the control over which provide power chances, such as economic assets, office and new laws. But they may also involve honor, ideology and behavioral norms. Loosely structured, factions are non-corporate groups that generally dissolve when the particular issues that gave rise to them are resolved. But if the issues remain unresolved, factions may acquire a range of cultural trappings such as property, symbols, ideology and bureaucratic organization. They can then evolve into permanent corporate groups, such as ritual moieties, political parties or other formal associations, for which the term faction is no longer appropriate.

Factionalism takes place within the framework of an established social entity, whether village, school, political party, office, club, kin group, etc., that have clear norms of behaviour. These norms generally include notions of unity, consensus, and cooperation.

The covert maneuvering of factions bent on achieving their own, often self-serving aims, contradict these norms. Hence factionalism has a pejorative connotation. It is seen as subverting organizational rules and goals. Faction members are consequently viewed as disloyal, obstructive persons whose pursuit of narrow, short-term advantage endangers the wider, long-term goals of the organization. The divisiveness inherent in factionalism also hampers the day to day affairs of the organization or community that depend on cooperation. Furthermore, factionalism jeopardizes the good name and image of unity and harmony it seeks to project to the outside world.

The Study of Factionalism

The study of factions and factionalism developed slowly, shadowing theoretical shifts in the social sciences. Although Linton (1936) long ago suggested that factions presented an interesting but unexplored field, little was done until the 1950s. Political anthropology until then was dominated by the functionalist paradigm elaborated in *African Political Systems* (Fortes and Evans-Pritchard, 1940). This viewed politics as maintaining order through consensus, harmony, and balanced opposition. The political groups on which functionalists focused were enduring units, corporate groups. Conflict, if examined, was viewed as reinforcing the social structure. Loosely structured, temporary coalitions such as factions patently did fit into this conception of politics. The theoretical hegemony of Africanist political anthropologists began to be challenged in the 1950s. Not surprisingly, Raymond Firth, given his interest in individual choice, had for some time been uncomfortable with the functionalist paradigm. He and colleagues, who had observed factionalism in Indian communities, were the first to examine factionalism theoretically (Firth, 1957). They treated factions as informal counterparts of more formal political formations whose members were recruited according to structurally diverse principles. They also noted that factions tended to become activated on specific occasions and not as regularly recurring features. Other studies of factions and factionalism swiftly

followed (Siegel and Beals, 1960; Boissevain, 1964, 1974; Nicholas, 1965; Bailey, 1969; Thoden van Velzen, 1973; Bujra, 1973; Alavi, 1973). Most employed a transactionalist perspective, viewing political activity as an arena in which entrepreneurs transact personal relations for political and economic gain. The study of factionalism culminated in the late 1970s with the volume edited by Silverman and Salisbury (1977). This is still considered the definitive work on the subject. Following its publication and the demise of functionalist social science – in which trans- actionalists played a significant role – academic debate moved on to puzzles related to symbolic, cognitive, and discursive approaches to politics. Factions and factionalism have become accepted concepts whose characteristics are no longer debated. They are also proving useful to related disciplines (see Brumfiel and Fox, 1994).

Issues

Much of the work on factionalism in the 1960s, although using a transactionalist perspective, continued to be strongly influenced by functionalism. Factionalism was seen as occurring as a result of rapid sociocultural change. Change was seen as coming from the surrounding environment, not as the result of tensions inherent in the community or society in which factionalism arose. Factionalism was generally viewed as occurring because the system's equilibrium was disturbed, and it operated to restore its dynamic equilibrium. Factions were regarded as structurally similar and in balanced opposition. These viewpoints were debated and successfully challenged during the 1970s.

Many of the societies in which factionalism was observed were in fact not subject to rapid social change. But if factionalism was not always a 'result' of change, it seemed always to be 'about' change: changes in power resources, ideology, rules and/or ways of doing things. Factions competed about who was to be boss, about which normative concepts were to be followed, about whose will was to prevail, and thus about which rules were to be followed. Moreover, a closer examination of rival factions revealed that far from being similar or evenly matched, they differed in their access to resources, strategy, tactics, internal organization, ideology, and social composition.

Factions: Structure, Symmetry, and Balance

Factions often form in opposition to or in defense of some issue or some pre-existing source of power and authority within a community or organization. The distribution of such resources is binary. Some persons have more and some have less. Those with more normally constitute the local establishment that coalesces around a leader or a dominant personage such as a headman, mayor, parish priest, or club president. They and their supporters form the 'establishment' faction. Those who are dissatisfied with the establishment's exercise of its power constitute a category from which a rival or 'opposition' faction can be recruited.

Reports also mention conservative and progressive factions. Because the local establishment normally defends the 'status quo,' from which it derives its superordinate position, it is often labeled as conservative. Since the opposition faction challenges the established defenders of the 'status quo' it is labelled progressive. These labels often do not reflect reality, as when a progressive opposition faction defeats its rival and becomes the dominant, establishment faction. There is a further reason why opposition factions come to be regarded as progressive. Since they are weaker, they are perhaps more receptive to new resources if and when these become available in the wider society. With these they can challenge their rivals. In rapidly changing societies these may include new government and commercial offices, new laws and new ideologies. Such new resources rapidly tend to change the balance of power. Factionalism thus does not necessarily result from the availability of new resources, as some authors have suggested. Rather, new resources are used in ongoing competitions for power and prestige and tend to escalate the conflict. The use of new resources is not random. A faction will use new resources when it seems likely that they will strengthen its position. It is then labelled progressive, in the sense of favoring change of the 'status quo.'

There is evidence that conflict groups – whether faction, ritual moiety, or political party – differ organizationally. When the opposition consists merely of a category of persons disgruntled with the dominant power elite, they are obviously less organized than the establishment. The internal structure of the local establishment faction, whose members are used to networking to maintain their position, will normally have a more developed exchange circuit than the opposition faction. But if a conflict between the two persists over time, the opposition faction may well become better organized than its rival may. Good organization is a valuable resource and one of the ways a weaker faction can successfully challenge its rival. It is thus more open than the establishment faction to organizational innovation. Because of its superior resources, the dominant faction also tends to be more wasteful. It does not need to husband its resources to the extent that its weaker rival does.

Like most coalitions, factions have core and peripheral members. Core members cluster around the leader and have multiple links to each other. The peripheral members are often linked only to the leader or to a single member of the core group. Where there is a strong core, the faction often acquires some of the characteristics of corporate groups noted previously.

Factions are not necessarily ideologically neutral, as some authors have suggested. The differences between establishment and opposition, like those between conservative and progressive, are not random. They have ideological implications. There is evidence that opposition factions recruit more support than their rivals from weaker or even marginal social categories do. Since opposition leaders usually lack the network ties and material resources that establishment leaders use to recruit followers, they cannot afford to be too particular about the nature of their support. The strength of a faction is usually a function of its size. Just as they often turn to new ideologies and tactics – which because they are new are often viewed as socially unacceptable and subversive – opposition leaders also recruit supporters from among those who are less influential or are regarded as social or morally inferior. Followers are followers.

Opposition factions consequently often develop or adopt an overarching ideology or symbol to bind their heterogeneous members into a unity. They also often align themselves with political parties that defend the interests of their socially weaker supporters. Most often, opposition factions seek links to parties that embody an emancipatory ideology. Their establishment rivals develop relations with political parties representing vested interests in the wider society. Political parties, on the other hand, also consciously make use of local factions to recruit support at the grass roots level.

Conclusion

If factionalism is not necessarily a product of social change, it appears to be always about change. Factions are coalitions that compete for power to determine, and thus to change, what is to be accepted as normal. Rival factions, because they have different access to power chances, are not evenly matched, structurally similar groups. Their asymmetry is fundamental to understanding the nature of factionalism and its dynamic for long term change.

Structural asymmetry and competition for power are also characteristic of class-based conflict groups. This suggests that the line of cleavage between faction, class or party cuts across moral categories and socioeconomic classes, not at right angles, as most functionalist and class-oriented analysts postulate, but diagonally. Where the line approaches the vertical, forming conflicting coalitions with a generally even spread across socio-economic classes, it is reasonable to speak of factionalism, in the case of face-to-face groups, and party conflict in the case of conflict on a broader scale. Where the line of cleavage approaches the horizontal, forming conflict groups that are more clearly differentiated according to socio-economic criteria, the term class conflict seems appropriate. But in every case the axis of cleavage must be determined by empirical investigation. It should not be taken for granted.

See also: Charisma, Social Aspects of; Conflict and Conflict Resolution, Social Psychology of; Conflict and Consensus; Conflict: Anthropological Aspects; Conflict: Organizational; Groups, Sociology of; Issue Networks: Iron Triangles, Subgovernments, Policy Communities, Policy Networks; Solidarity, Sociology of.

Bibliography

Alavi, H., 1973. Peasant classes and primordial loyalties. Journal of Peasant Studies. 1, 232–262.

Bailey, F.G., 1957. Caste and the Economic Frontier. Manchester University Press, Manchester, UK.

Bailey, F.G., 1969. Stratagems and Spoils. Basil Blackwell, Oxford, UK.

Boissevain, J., 1964. Factions, parties and politics in a Maltese village. American Anthropologist 66, 1275–1287.

Boissevain, J., 1974. Friends of Friends: Networks, Manipulators and Coalitions. Basil Blackwell, Oxford, UK.

Brumfiel, E.M., Fox, J.W. (Eds.), 1994. Factional Competition and Political Development in the New World. Cambridge University Press, Cambridge, UK.

Bujra, J., 1973. The dynamics of political action: A new look at factionalism. American Anthropologist 75, 132–152.

Firth, R., 1957. Introduction to factions in Indian and overseas Indian societies. British Journal of Sociology. 8, 291–295.

Fortes, M., Evans-Pritchard, E.E. (Eds.), 1940. African Political Systems. Oxford University Press, London.

Linton, R., 1936. The Study of Man. D. Appleton-Century, New York.

Nicholas, R.W., 1965. Factions: A comparative analysis. In: Banton, M. (Ed.), Political Systems and the Distribution of Power. Tavistock, London.

Siegel, B., Beals, A.R., 1960. Pervasive factionalism. American Anthropologist. 62, 395–417.

Silverman, M., Salisbury, R.F. (Eds.), 1977. A House Divided? Anthropological Studies ofFactionalism. Memorial University, Newfoundland, Canada.

Thoden van Velzen, H.U.E., 1973. Coalitions and network analysis. In: Boissevain, J., Mitchell, J.C. (Eds.), Network Analysis: Studies in Human Interaction. Mouton, The Hague, The Netherlands, pp. 219–250.

Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis

Ralph O Mueller, University of Hartford, West Hartford, CT, USA
Gregory R Hancock, University of Maryland, College Park, MD, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R.O. Mueller, volume 8, pp. 5239–5244, © 2001, Elsevier Ltd.

Abstract

Confirmatory factor analysis (CFA) is a quantitative data analysis method that belongs to the family of structural equation modeling (SEM) techniques. CFA allows for the assessment of fit between observed data and an a priori conceptualized, theoretically grounded model that specifies the hypothesized causal relations between latent factors and their observed indicator (manifest) variables. Because population-level equivalence between data and model cannot be shown with sample data, CFA should be viewed as a mainly disconfirmatory technique. That is, CFA facilitates the statistical rejection – or, at best, a tentative retention – of a specific theory regarding the factor(s) responsible for the observed relations in the data. If, on the other hand, the investigator's intentions are a mostly ungrounded exploration of relations suggested by the data, classical exploratory factor analysis is the more appropriate approach. In this article, typical steps in a CFA are introduced theoretically and via example: from model specification and identification, to parameter estimation, data-model fit assessment, and potential model modification. Didactic references are provided for a more in-depth study of CFA and SEM techniques in the social and behavioral sciences.

Overview

The term *factor analysis* describes a host of methods, all of which have the purpose of facilitating a better understanding of the unobserved variables (factors) that underlie a set of directly measurable and observed variables (for a nontechnical overview, see Bandalos and Finney, 2010 and references therein). These factors are often believed to represent constructs, psychological or otherwise, that have a direct bearing on the measured variables; as such, they are assumed to motivate (and in turn be inferable from) the pattern of covariances (unstandardized correlations) among those observed variables. In the late 1960s, works by Karl Jöreskog (e.g., 1966, 1967) articulated a method for *confirmatory factor analysis* (CFA), an application of normal theory maximum likelihood estimation to factor models with specific theoretical latent structures. Such structures could include the a priori specification of the number of factors, their orthogonality or obliquity, and which variables had zero and nonzero relations with those factors. This distinguishes CFA from well-known *exploratory factor analysis* (for classic introductions, see Gorsuch, 1983; Mulaik, 1972) wherein the number and nature of the factors emerge from the observed variables' data through a mathematical algorithm, largely blind to any substantive theory. Most crucial in Jöreskog's CFA work was the provision for a formal statistical χ^2-test of the fit between the pattern of relations among the measured variables and the theorized factor model, thereby facilitating the disconfirmation or tentative confirmation of a hypothesized factor model. Soon after, Jöreskog and others put forth a more general framework for the integration of measured and latent variables into causal networks, serving as the foundation for what is often known as *structural equation modeling* (SEM). CFA, which may be considered a special case within the more general SEM framework, is the focus of the current article.

CFA Model Specification, Identification, and Parameter Estimation

Suppose an educational researcher wishes to investigate the possibility of a low positive relation between reading and mathematics ability for fifth grade students, which is measured by standardized tests such as the Stanford Achievement Test or the Iowa Test of Basic Skills. The model shown in Figure 1 might be hypothesized. Measured variables X_1 through X_6, shown in rectangles, are believed to be caused by the latent factors ξ_1 and ξ_2, shown in ellipses. Here, ξ_1 and ξ_2 represent

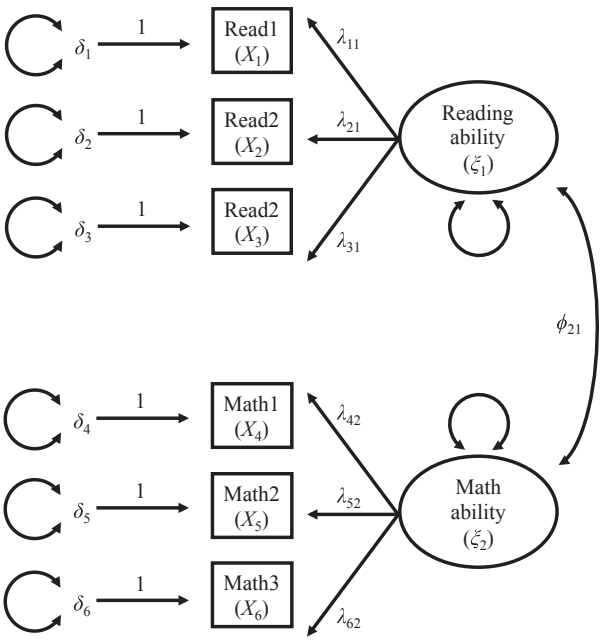

Figure 1 Hypothetical CFA model of reading and mathematics ability.

true latent (unobserved) reading and mathematics ability, respectively, with X_1 through X_3 being standardized reading test measures (Read1 through Read3) and X_4 through X_6 being standardized mathematics test measures (Math1 through Math3). Table 1 includes a simulated variance/covariance matrix for the observed variables X_1 through X_6 based on test data from $n = 1200$ fifth graders. Our example's focus is the noncausal covariance between reading and mathematics ability, ϕ_{21}. In SEM path diagrams, a covariance is indicated by a two-headed arrow connecting the two constructs, and – because a variance is a covariance of a variable (observed or latent) with itself – it, too, is depicted by a two-headed arrow, in this case from the variable to itself.

A factor's hypothesized causal impact on its measured indicator variables, its *loading*, is symbolized by an arrow from the factor to the variable with magnitude λ_{ij}, where i denotes the observed variable and j denotes the latent factor. Note that such a model explicitly posits the factors as causing the variables, rather than the variables causing the factors; the latter type of model, in which the factor is characterized as *emergent* rather than *latent*, is much less common and beyond the scope of this article (but see Kline, 2013). In many cases, there is no arrow from a factor to a variable, such as from ξ_1 to X_4; this implies that reading ability has no theoretical causal bearing on the Math1 variable. Finally, to the extent that the factors do not perfectly explain each variable, a residual term, δ_i, is included as an influential contributor (with its variance shown by a two-headed arrow from δ_i to itself). This residual might consist of variable-specific measurement error as well as other influences. Thus, each observed variable is the sum of two parts, that attributable to the latent factor(s) and a residual part unique to the variable.

The causal relations of the hypothesized model shown in Figure 1 may be expressed as a system of six regression-like *structural equations*:

$$X_1 = \lambda_{11}\xi_1 + 0\xi_2 + \delta_1 \quad [1]$$

$$X_2 = \lambda_{21}\xi_1 + 0\xi_2 + \delta_2 \quad [2]$$

$$X_3 = \lambda_{31}\xi_1 + 0\xi_2 + \delta_3 \quad [3]$$

$$X_4 = 0\xi_1 + \lambda_{42}\xi_2 + \delta_4 \quad [4]$$

$$X_5 = 0\xi_1 + \lambda_{52}\xi_2 + \delta_5 \quad [5]$$

$$X_6 = 0\xi_1 + \lambda_{62}\xi_2 + \delta_6 \quad [6]$$

Equivalently, these equations can be represented in matrix form, as in

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{bmatrix} \quad [7]$$

That is,

$$\mathbf{X} = \mathbf{\Lambda\xi} + \boldsymbol{\delta} \quad [8]$$

where $\mathbf{X}$ is a column vector of observed variables, Λ is a matrix of factor loadings, $\boldsymbol{\xi}$ is a vector of latent constructs, and $\boldsymbol{\delta}$ is a column vector of residuals.

The implication of Figure 1 and the accompanying structural equations is that the population variance/covariance matrix for the X variables, $\mathbf{\Sigma}$, is a function of (1) the λ_{ij} loadings in matrix Λ, (2) the variances and covariance of the latent factors in a matrix $\mathbf{\Phi}$, and (3) the variances and covariances among the residuals in a matrix $\mathbf{\Theta}_\delta$ (note that in Figure 1 all residual covariances are zero as implied by the absence of two-headed arrows between the δ terms). More specifically, if all model parameters (loadings, variances, and covariances) are contained in a single column vector $\boldsymbol{\theta}$, the population variance/covariance matrix of the observed variables that is implied by the model and its parameters, $\mathbf{\Sigma}(\boldsymbol{\theta})$, is given by

$$\Sigma(\boldsymbol{\theta}) = \mathbf{\Lambda\Phi\Lambda}' + \mathbf{\Theta}_\delta \quad [9]$$

A vector of parameter estimates, $\widehat{\boldsymbol{\theta}}$, can be derived so that the model-implied variance/covariance matrix $\mathbf{\Sigma}(\widehat{\boldsymbol{\theta}})$ is as similar as possible to the observed variance/covariance matrix, $\mathbf{S}$, provided that model identification has first been ensured. To that end, each parameter in a model must be expressible as a function of the variances and covariances of the observed variables. When a system of such relations can be uniquely

Table 1 Simulated data and selected parameter estimates for the reading and mathematics ability model in Figure 1

	X_1	X_2	X_3	X_4	X_5	X_6
Variance/covariance matrix ($n = 1200$)						
	129.96					
	79.75	192.65				
	694.20	871.11	12038.48			
	307.03	391.75	3402.04	9876.38		
	230.99	415.74	2476.89	4815.24	12126.41	
	37.85	53.36	416.71	740.98	656.84	135.722
Standardized factor loadings and indicator reliability estimates						
ξ_1	0.70*	0.73*	0.79*	0	0	0
ξ_2	0	0	0	0.75*	0.60*	0.85*
R^2	0.49	0.53	0.62	0.56	0.35	0.73
Data-model fit indices						
$\chi^2 = 16.98$, $df = 8$ $p = 0.030$		CFI = 0.996	SRMR = 0.016		RMSEA = 0.031	

Note. $^*p < 0.05$.

solved for the unknown parameters, the model is *just-identified*. When multiple such expressions exist for one or more parameters, the model is *overidentified* and, in that case, a best-fit (although not unique) estimate for each parameter is derived. If, however, at least one parameter cannot be expressed as a function of the observed variables' variances and covariances, the model is *underidentified*, and some or all parameters cannot be estimated on the basis of the data alone. This underidentification might be the result of the researcher attempting to impose a model that is too complex relative to the number of variances and covariances of the observed variables. Additionally, empirical underidentification might arise when unfortunate estimates for select parameters (e.g., values of zero for factor covariances) render subsets of model parameters inestimable. Fortunately, in most CFA applications, it suffices to ensure that (1) the total number of parameters to be estimated, t, does not exceed the number of unique variances and covariances of the observed variables, u, and (2) each latent factor has an assigned unit of measurement. To accomplish the latter condition for the model in **Figure 1**, we set the factor variances to unity (alternatively, for each of the two factors we could have specified one of the factor loadings to equal unity, thereby setting each factor's units equal to those of the observed variable). The model in **Figure 1** is overidentified with $u = 6(7)/2 = 21$ nonredundant observed variances and covariances and $t = 13$ parameters to be estimated: six factor loadings, one covariance between the two latent factors, and six variances of the residual terms associated with the observed variables.

Given that a model is just- or (preferably) overidentified, sample estimates can be obtained through a variety of estimation methods. These include maximum likelihood and generalized least squares, both of which assume multivariate normality and are asymptotically equivalent, as well as asymptotically distribution-free estimation methods that generally require a substantially larger sample size. These methods iteratively minimize a function of the discrepancy between $\mathbf{S}$ and $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$, where $\mathbf{S}$ is the unrestricted variance/covariance matrix of the observed X variables and $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ is the model-implied variance/covariance matrix reproduced from the iteratively changing parameter estimates. The standardized maximum likelihood estimates of key parameters in the reading and mathematics ability model are presented in **Table 1**. Before focusing on the estimate of the example's main parameter of interest (ϕ_{21}), we should consider whether or not there is any evidence suggesting data-model misfit, any statistical – and theoretically justifiable – rationale for modifying the hypothesized model, or any indication of factor unreliability.

Data-Model Fit Assessment and Model Modification

One of the advantages of CFA is the ability to assess the quality of the fit of the data to the model. A multitude of measures exists that assist the researcher in deciding whether to reject or tentatively retain an a priori specified overidentified model (see Marsh et al., 1988; Tanaka, 1993). In general, measures to assess the fit between the variances and covariances observed in the data and those implied by the model can be classified into three categories: absolute, parsimonious, and incremental. *Absolute* fit indices are those that improve as the overall discrepancy between $\mathbf{S}$ and $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ decreases. Examples of such measures include the model χ^2 statistic that tests the stringent null hypothesis H_0: $\boldsymbol{\Sigma} = \Sigma(\boldsymbol{\theta})$, the standardized root mean-square residual (SRMR) that roughly assesses the average standardized discrepancy between observed and model-implied variances and covariances, and the goodness-of-fit index (Jöreskog and Sörbom, 1996) designed to evaluate the amount of observed variance/covariance information that can be accounted for by the model.

Parsimonious fit indices take into account not just the overall absolute fit, but also the degree of model complexity required to achieve that fit. Indices such as the adjusted goodness-of-fit index (Jöreskog and Sörbom, 1996), the Akaike Information Criterion (Akaike, 1974), and the root mean-square error of approximation (RMSEA, Steiger and Lind, 1980) indicate greatest data-model fit when data have reasonable absolute fit and models are relatively simple. Finally, *incremental* fit indices such as the normed fit index (Bentler and Bonett, 1980) and the comparative fit index (CFI, Bentler, 1990) gauge the data-model fit of a hypothesized model relative to that of a more restrictive baseline model with fewer parameters.

The three types of fit indices together help the researcher to converge upon a decision regarding the CFA model's acceptability. In our example, the statistically significant maximum likelihood $\chi^2 = 16.98$ ($df = u - t = 8$, $p < 0.05$) indicates that the observed variance/covariance matrix would occur rarely if our model correctly depicted the true population relations. This absolute fit statistic, however, is notoriously sensitive to very small and theoretically trivial model mis-specifications (e.g., slight amounts of error covariance) under large sample conditions. As such, other fit indices are generally preferred for model evaluation. According to Hu and Bentler (1999), for example, CFI values of 0.96 or greater together with SRMR values less than 0.09 (or with RMSEA values less than 0.06) point to acceptable data-model fit. The indices in **Table 1** suggest no appreciable data-model inconsistency given the Hu and Bentler recommendations.

After the data-model fit has been assessed, a decision about that model's worth must be reached. Acceptable fit indices usually lead to the conclusion that no present evidence exists warranting a rejection of the model or the theory underlying it. This is not to say that the model and theory have been confirmed, much less proven as correct; rather, the current factor model remains as one of possibly many that satisfactorily explain the relations among the observed variables (see Hershberger and Marcoulides, 2013). On the other hand, when fit indices indicate a potential data-model misfit, one might be reluctant to dismiss the model entirely. Instead, attempts are often made to modify the model post hoc so that acceptable fit indices can be obtained. Such modifications could include the addition of cross-loadings, allowing a given variable to load on multiple factors, or of error covariances in which variables' residuals are allowed to covary in order to reflect some potential variable relation above and beyond that motivated by the factor structure itself.

Most CFA software packages will facilitate such model improvement by providing modification indices (Lagrange multiplier tests) indicating what changes in the model could

reap the greatest increase in absolute fit, that is, decrease in the model χ^2 statistic. While such indices constitute a potentially useful tool for remedying incorrectly specified models, it seems imperative to warn against an atheoretical hunt for the model with the best fit. Many alternative models exist that can explain the observed data equally well; hence, attempted modifications must be based on a sound understanding of the specific theory underlying the model. Furthermore, when modifications and reanalyzes of the data are based solely on data-model misfit information, subsequent fit results might be due largely to chance rather than true improvements to the model. Modified structures therefore should be cross-validated with an independent sample whenever possible. If a new sample is not available but the initial sample is large enough, one can randomly split the sample into calibration and validation subsamples and compute Cudeck and Browne's (1983) cross-validation index. When the initial sample is too small, Browne and Cudeck (1989) also offered a single sample alternative, an estimate of the expected value of the cross-validation index. From the analysis of the model in **Figure 1**, none of the modification indices suggested changes to the model that would result in a significant improvement in data-model fit (i.e., a significant decrease in the model χ^2 statistic); thus, we did not report them in **Table 1**.

Before drawing conclusions regarding the relation between reading and mathematics ability, the question of the quality (i.e., reliability) of the factors should be addressed. Traditionally, this has been accomplished (1) by focusing on the reliability of scores from individual indicator variables, R^2 (i.e., the proportion of variability in an observed variable that can be accounted for by the underlying factor), or (2) by assessing the reliability of scores from linear composites of the indicator variables (e.g., Miller, 1995; Raykov, 1997). Alternatively, Hancock and Mueller (2001) suggested a measure (coefficient H) computable most easily from the standardized factor loadings that can be used to assess the reliability (i.e., replicability) of a latent construct itself as reflected by scores from its multiple observed indicator variables. One of coefficient H's advantages over traditional construct reliability measures is that it is never less than the best indicator variable's reliability (R^2), thereby drawing information from all indicators in a manner commensurate with their own ability to reflect the construct. For the current example, $\widehat{H} = 0.79$ and $\widehat{H} = 0.82$ for the reading and mathematics ability constructs, respectively, while the R^2 values for the respective factors' indicator variables in **Table 1** range between 0.49–0.62 and 0.35–0.73. Thus, the two factors exhibit reasonable and satisfactory levels of construct reliability, given that about 80% of their variance is explainable by their respective indicators. **Table 1** also lists the standardized factor loadings, that is, the standardized versions of the λ_{ij} parameters in the $\mathbf{\Lambda}$ matrix that indicate the strength and direction of the a priori specified causal influences of the latent factors on the observed variables. All factor loadings are positive and statistically significantly different from zero ($p < 0.05$).

Finally, our primary research question can be addressed. The CFA estimate of the correlation between the two latent constructs reading and mathematics ability is 0.51 and statistically significantly different from zero ($p < 0.05$). This estimate indicates a low to moderate positive association between the two constructs of interest, as hypothesized.

Conclusion

CFA has become established as an important analytical tool for many areas of the social and behavioral sciences. It belongs to the family of SEM techniques that allow for the investigation of causal relations among latent and observed variables in a priori specified, theory-derived models. The main advantage of CFA lies in its ability to aid researchers in bridging the common gap between theory and observation. For example, an instrument might be developed by creating multiple items for each of several specific theoretical constructs (**Figure 1**). Instead of analyzing data with an exploratory factor analysis (where each item is free to load on each factor) and potentially facing a solution inconsistent with initial theory, a CFA can give the investigator valuable information regarding the fit of the data to the specific, theory-derived measurement model (where items load only on the factors they were designed to measure), and point to the potential weakness of specific items. CFA is best understood as a process, from model conceptualization, identification, and parameter estimation, to data-model fit assessment and potential model modification. As opposed to exploratory methods, CFA's strength lies in its disconfirmatory nature: models or theories can be rejected, but results might also point toward potential modifications to be investigated in subsequent analyses.

A recent treatment of CFA for applied researchers is Brown (2009). Popular textbooks on more general SEM – those that include many CFA examples from disciplines covered in this encyclopedia and that utilize commonly available software (i.e., AMOS, EQS, LISREL, and MPlus) – are Bollen (1989), Byrne (1998, 2006, 2010, 2012), Kline (2011), Loehlin (2004), Raykov and Marcoulides (2006), and Schumacker and Lomax (2010). Social science journals that publish many CFA and SEM applications and methodological developments include *Educational and Psychological Measurement*, *Multivariate Behavioral Research*, *Journal of Experimental Education*, *Psychological Methods*, *Sociological Methodology*, *Sociological Methods and Research*, and *Structural Equation Modeling: A Multidisciplinary Journal*.

See also: Big Five Factor Model, Theory and Structure; Emotions and Aging; Factor Analysis and Latent Variable Models in Personality Psychology; Five Factor Model of Personality, Assessment of; Five Factor Model of Personality, Facets of; Five Factor Model of Personality, Universality of.

Bibliography

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723.

Bandalos, D.L., Finney, S.J., 2010. Factor analysis: exploratory and confirmatory. In: Hancock, G.R., Mueller, R.O. (Eds.), The Reviewer's Guide to Quantitative Methods in the Social Sciences. Routledge, New York, NY, pp. 93–114.

Bentler, P.M., 1990. Comparative fit indexes in structural models. Psychological Bulletin 107, 238–246.

Bentler, P.M., Bonett, D.G., 1980. Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin 88, 588–606.

Bollen, K.A., 1989. Structural Equations with Latent Variables. Wiley, New York, NY.

Brown, T.A., 2009. Confirmatory Factor Analysis for Applied Research. Guilford Press, New York, NY.

Browne, M.W., Cudeck, R., 1989. Single sample cross-validation indexes for covariance structures. Multivariate Behavioral Research 24, 445–455.
Byrne, B.M., 1998. Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS. Lawrence Erlbaum, Mahwah, NJ.
Byrne, B.M., 2006. Structural Equation Modeling with EQS: Basic Concepts, Applications, and Programming, second ed. Lawrence Erlbaum, Mahwah, NJ.
Byrne, B.M., 2010. Structural Equation Modeling with AMOS, second ed. Routledge, New York, NY.
Byrne, B.M., 2012. Structural Equation Modeling with Mplus. Taylor & Francis, New York, NY.
Cudeck, R., Browne, M.W., 1983. Cross-validation of covariance structures. Multivariate Behavioral Research 18, 147–167.
Gorsuch, R.L., 1983. Factor Analysis, second ed. Lawrence Erlbaum Associates, Hillsdale, NJ.
Hancock, G.R., Mueller, R.O., 2001. Rethinking construct reliability within latent variable systems. In: Cudeck, R., du Toit, S., Sörbom, D. (Eds.), Structural Equation Modeling: Present and Future—A Festschrift in Honor of Karl Jöreskog. Scientific Software International, Lincolnwood, IL, pp. 195–216.
Hershberger, S.L., Marcoulides, G.A., 2013. The problem of equivalent models. In: Hancock, G.R., Mueller, R.O. (Eds.), Structural Equation Modeling: A Second Course, second ed. Information Age Publishing, Charlotte, NC, pp. 3–39.
Hu, L., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal 6, 1–55.
Jöreskog, K.G., 1966. Testing a simple structure hypothesis in factor analysis. Psychometrika 31, 165–178.
Jöreskog, K.G., 1967. Some contributions to maximum likelihood factor analysis. Psychometrika 32, 443–482.
Jöreskog, K.G., Sörbom, D., 1996. LISREL 8. Scientific Software International, Chicago.
Kline, R.B., 2011. Principles and Practice of Structural Equation Modeling, third ed. The Guilford Press, New York, NY.
Kline, R.B., 2013. Reverse arrow dynamics: Feedback loops and formative measurement. In: Hancock, G.R., Mueller, R.O. (Eds.), Structural Equation Modeling: A Second Course, second ed. Information Age Publishing, Charlotte, NC, pp. 41–79.
Loehlin, J.C., 2004. Latent Variable Models, fourth ed. Lawrence Erlbaum, Hillsdale, NJ.
Marsh, H.W., Balla, J.R., McDonald, R.P., 1988. Goodness-of-fit indexes in confirmatory factor analysis: the effect of sample size. Psychological Bulletin 103, 391–410.
Miller, M.B., 1995. Coefficient alpha: a basic introduction from the perspectives of classical test theory and structural equation modeling. Structural Equation Modeling: A Multidisciplinary Journal 2, 255–273.
Mulaik, S.A., 1972. The Foundations of Factor Analysis. McGraw-Hill, New York, NY.
Raykov, T., 1997. Estimation of composite reliability for congeneric measures. Applied Psychological Measurement 21, 173–184.
Raykov, T., Marcoulides, G.A., 2006. A First Course in Structural Equation Modeling, second ed. Lawrence Erlbaum, Mahwah, NJ.
Schumacker, R.E., Lomax, R.G., 2010. A Beginner's Guide to Structural Equation Modeling, third ed. Routledge, New York, NY.
Steiger, J.H., Lind, J.M., 1980. Statistically based tests for the number of common factors. In: Paper Presented at the Annual Meeting of the Psychometric Society. Iowa City, IA.
Tanaka, J.S., 1993. Multifaceted conceptions of fit in structural equation models. In: Bollen, K.A., Long, J.S. (Eds.), Testing Structural Equation Models. Sage, Newbury Park, CA, pp. 10–39.

Relevant Websites

www.ibm.com – AMOS Software.
www.mvsoft.com – EQS Software.
www.ssicentral.com – LISREL Software.
www.statmodel.com – MPlus Software.

Factor Analysis and Latent Structure Analysis: Overview

David J Bartholomew, London School of Economics, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

For over 100 years, factor analysis has probably been the most widely used multivariate technique among social scientists. It is still not well known that it is one of a family of methods which share the same theoretical framework and pose similar problems of interpretation. These are known, collectively, as latent variable methods. The family also includes latent class analysis and latent profile analysis. The family connection is described here within a framework which is designed to show what they have in common. Latent variables play a fundamental role in social science and some other occurrences are briefly touched upon.

Introduction

The terms factor analysis and latent structure analysis refer to two aspects of essentially the same problem. In practice, a 'factor,' as the term is used in factor analysis, is simply a latent variable and the failure to recognize this equivalence has often caused much misunderstanding. In ordinary statistical usage the term 'factor' is used, rather loosely, to refer to almost any variable. The reason for this imprecision is historical in that factor analysis is, possibly, the oldest multivariate statistical method. It was invented by Charles Spearman around 1904, and was developed and initially used almost exclusively by psychologists. It is now possible to see the two strands in their proper relationship.

A latent variable is any variable occurring in a statistical problem, which cannot be observed. This may be because it is inconvenient, or impracticable, to observe it or it cannot be observed, even in principle. The latter case includes what we might call hypothetical variables, which are introduced into a statistical problem because they simplify the interpretation of the results. For example, one of the earliest and most enduring examples of a latent variable is 'general ability.' This is closely related to the subject of Spearman's original work on general intelligence. It is commonly found that scores on tests of arithmetic, for example, made on the same individual, say, are highly correlated with one another. This is usually put down to the fact that all such scores depend on an underlying ability to perform such tests. By introducing a latent variable representing arithmetical ability, the interpretation of the data can be considerably simplified. Sometimes the object of the analysis is less specific and we merely inquire whether there might be any latent variable, which plays such an explanatory role without knowing, in advance, what it is. The problem then is to identify what that latent variable might be in terms which make substantive sense.

Viewed in this way, latent variable analysis is a special case of a much wider range of statistical problems concerning unobserved variables, which has been treated in Bartholomew (2013). These include missing values, simple mixtures, and, when viewed appropriately, time series analysis. Indeed, the whole of statistics is included if we approach the subject from a Bayesian perspective where parameters are treated, in effect, as latent variables.

The Basic Idea

Although factor analysis and other latent variable methods are often regarded as relatively advanced multivariate methods, they all depend on a very simple idea. This idea is very familiar to students, in elementary statistics courses, under the guise of what is called 'spurious' correlation. There it is explained that a correlation between two variables may not be due to a real causal link between them, but be caused by the presence of a third variable which is related to both. This common influence results in a correlation which is properly called 'spurious.' To take an old example where the relationships are now relatively well understood, consider the variables involved in cancer of the lung. It has long been known that the amount of tobacco smoke inhaled by an individual is positively correlated with the risk of cancer of the lung. But it could be argued that this correlation could be spurious because it might be induced by some other variable. For example, living in an urban environment, with its polluted atmosphere, might be the cause of both smoking and lung cancer. Had this been the case, the correlation between smoking and lung cancer would have vanished if the analysis had been done for subgroups whose members all had the same exposure to polluted air. In essence, the general aim is to find other variables which, if held constant, would have eliminated the correlation between smoking and cancer.

In any latent variable analysis the aim is to 'explain' the correlations between observed variables by showing that they could have arisen because of their dependence on other, unobserved, variables.

This idea may be further illustrated by a numerical example using only binary variables called A and B and expressing their relationship in the familiar form of a 2×2 table.

	$A = 1$	$A = 0$	*Total*
$B = 1$	700	400	1100
$B = 0$	300	600	900
Total	1000	1000	2000

This table shows a positive association between A and B; if $A = 1$, B is more likely to be 1 than if $B = 0$. Suppose now there is a third binary variable called C. We might then look at the association between A and B for each of the two values taken by C. Suppose the breakdown gives the following two tables.

$C = 1$	$A = 1$	$A = 0$	*Total*
$B = 1$	640	160	800
$B = 0$	160	40	200
Total	800	200	1000

$C = 0$	$A = 1$	$A = 0$	*Total*
B = 1	60	240	300
B = 0	140	560	700
Total	200	800	1000

It is clear that there is no association between A and B in either of the last two tables; the column frequencies in the first two columns are in exactly the same proportion as in the final column. This implies that the association in the first table is entirely spurious because, when we hold C constant, the association vanishes.

Although this highly simplified example conveys the basic idea, real problems are much more complicated. The variables need not be binary but could be continuous, categorical – with and without ordering of the categories – or any mixture of the two. In our example, there was only one latent variable, C. In practice there may be several and, as with the observed variables, they may be continuous or categorical. In all cases the essential idea is to find whether variables exist, like C, such that, when we condition on them, the observed variables become independent.

We now go over this ground again but this time in greater generality.

Formal Statement of the Conceptual Framework

The last section sets out the key idea underlying all latent variable models. Now, the general approach is expressed more formally in mathematical terms. This is essential if the full generality is to be grasped but, to ease the path of the nonmathematical reader, we treat only the case of continuous variables and simplify the presentation by omitting details such as the limits of integration. Full details will be found in Bartholomew et al. (2011), but the reader who is interested only in specific models may pass immediately to later sections.

Denote the set of observed variables by the p-vector, $\mathbf{x}$. Its elements, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ may be continuous or categorical or a mixture of the two. These are the generalization of the binary variables A and B of the last section and they are often referred to as 'manifest' variables. The latent variables are similarly denoted by the q-vector, $\mathbf{z}$.

The essential thing about this kind of model is that $\mathbf{z}$ is not observed. For this reason everything must be expressed in terms of the distribution of the observable variables $\mathbf{x}$. The relationship between the two can be obtained by using standard probability methods as

$$f(\mathbf{x}) = \int f(\mathbf{x}, \mathbf{z})\mathrm{d}\mathbf{z} \qquad [1]$$

Here, as already noted, the equation is written as if the variables were all continuous – hence the integral in eqn [1]. This is purely for convenience and the translation needed to cover other cases does not affect the logic of the argument. The next step is to write

$$f(\mathbf{x}, \mathbf{z}) = f(\mathbf{z})f(\mathbf{x}|\mathbf{z}) \qquad [2]$$

This decomposition provides us with the two elements to which the description of the last section referred – namely a statement of how the manifest variables would vary if $\mathbf{z}$ were known, given here by $f(\mathbf{x}|\mathbf{z})$, and the prior distribution of $\mathbf{z}$. If both parts of eqn [2] can be specified, $f(\mathbf{x})$ can be determined and thus the way is open to estimate any parameters of the model from a likelihood function constructed from $f(\mathbf{x})$. To scale $\mathbf{z}$, that is to find a value for $\mathbf{z}$ for an individual with a given $\boldsymbol{x}$, we may use the posterior distribution given by

$$f(\mathbf{z}|\mathbf{x}) = f(\mathbf{z})f(\mathbf{x}|\mathbf{z})/f(\mathbf{x}) \qquad [3]$$

A scale value, or factor score as it is known, for a particular $\mathbf{z}$ can then be found by using some measure of location of its posterior distribution such as $E(\mathbf{z}_i|\mathbf{x})$. This specification is deceptively simple. It merely spells out some of the consequences of regarding $\mathbf{x}$ and $\mathbf{z}$ as random variables. To turn it into a model we have to specify $f(\mathbf{z})$ and $f(\mathbf{x}|\mathbf{z})$. Here we face a fundamental problem because the decomposition of eqn [3] is not unique – there are infinitely many possible pairs, each leading to the same $f(\mathbf{x})$ and hence indistinguishable empirically from one another.

The choice of $f(\mathbf{z})$ is resolved by appealing to the fact established above that, since $\mathbf{z}$ is a latent variable, it can be constructed to have any convenient distribution. For instance, as a matter of convention one may choose the $\mathbf{z}$'s to be standard, normal, and mutually independent.

The core of the model-building exercise centers on the choice of the conditional distribution $f(\mathbf{x}|\mathbf{z})$. This has to be considered in the context of individual problems but there is one further simplification which applies to all models. This simplification involves making explicit what is already implicit in our formulation. The fact that the $\mathbf{x}$'s are supposed to be indicators of $\mathbf{z}$ implies that the $\mathbf{x}$'s will be mutually correlated – because of their common dependence on some, at least, of the $\mathbf{z}$'s. Hence, if $\mathbf{z}$ were to be fixed, the source of those correlations would vanish. Conditional on $\mathbf{z}$, the $\mathbf{x}$'s would then be independent. If this were not so, it could be inferred that there was at least one other $\mathbf{z}$, not already included, giving rise to the correlations remaining. Sufficient $\mathbf{z}$'s therefore have to be included in the model to eliminate the correlations after conditioning of $\mathbf{z}$. It then follows that

$$f(x|z) = \prod_{i=1}^{n} f_i(x_i|z) \qquad [4]$$

This is often referred to as the assumption of conditional (or local) independence but it is not an assumption in the usual

sense. It is, rather, a statement of what is meant by saying that $\mathbf{z}$ explains the interdependence of the $\mathbf{x}$'s. It is, essentially, a definition of q as the smallest integer for which an equation of the form [4] can be constructed.

How this may be done in particular cases will be shown in the following sections, but there is an important general principle which offers a fruitful way forward. Distributions of the exponential family are widely used in statistical modeling because they lead to estimates with highly desirable properties such as sufficiency. They provide a rationale for many of the most important models used in statistics, such as generalized linear models.

Essentially the same properties can be utilized here. They are set out in the article entitled *The General Linear Latent Variable Model* in Bartholomew et al. (2011). The essential idea is to make what is sometimes called the 'natural' or 'canonical' parameter of the one-parameter exponential family distribution, a linear function of the latent variables. If that is done, it turns out that the posterior distribution of $\mathbf{z}$ given $\mathbf{x}$ of eqn [3] depends on the $\mathbf{x}$'s only through q linear combinations of the $\mathbf{x}$'s, which we may write

$$X_j = \sum_{i=1}^{p} \alpha_{ij} x_i \quad (j = 1, 2, \ldots, q) \qquad [5]$$

This is a remarkable result, which justifies the common practice of 'estimating' latent variables from sums, or linear combinations of observed variables.

Structure of the Family

As we have already noted, latent variables can be classified into several types. Many, like intelligence, are conceived of as being continuous, in which case we are looking for a scale on which individuals can be located. In other contexts, it is more appropriate to think of the latent variable as categorical. In that case individuals are supposed to belong to one of several classes, which may or may not be ordered. What is true of the latent variables is, of course, true for the manifest variables and the only essential difference between the various methods is in the types of variables for which they are appropriate. A convenient way of displaying the relationship between the methods is to introduce the fourfold classification of **Table 1**. We classify the manifest and the latent variables as 'metrical' or 'categorical.' In the former case we mean that they take values on some continuous or discrete scale of measurement; in the latter they fall into categories.

This classification is not exhaustive. It does not, for example, include cases where the two kinds of variables occur among the manifest (or latent) variables. Neither does it distinguish between continuous and discrete metrical variables nor between ordered and unordered categorical variables. Nevertheless, it does serve to provide a broad convenient framework in which to place the most widely used methods covered in this article. The terminology of the subject reflects the diverse origins of the subject accumulated over almost a century.

Table 1 Classification of latent variable methods

	Manifest variables	
	Metrical	*Categorical*
Latent variables		
Metrical	Factor analysis	Latent trait analysis
Categorical	Latent profile analysis	Latent class analysis

Members of the Family

Factor Analysis

This is the oldest and most widely used latent variable method. The model is usually written as

$$\mathbf{x} = \mu + \Lambda \mathbf{z} + \mathbf{e} \qquad [6]$$

where $E(\mathbf{z}) = E(\mathbf{e}) = 0$; $\mathrm{cov}(\mathbf{e}) = \Psi$, a diagonal matrix; $\mathrm{cov}(\mathbf{z},\mathbf{e}) = 0$; and $\mathbf{z} \sim N(0, 1)$. An equivalent way of writing the model, which conforms more closely to the general treatment above, is to specify the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ as

$$\mathbf{x}|\mathbf{z} \sim N(\mu + \Lambda \mathbf{z}, \psi) \qquad [7]$$

which makes it clear that $\mathbf{z}$ influences $\mathbf{x}$ only through its mean, which is linear in $\mathbf{z}$. The marginal distribution of $\mathbf{x}$ is $N(\mu, \Lambda\Lambda' + \Psi)$ and it is from this that the parameters Λ and Ψ must be estimated. The matrix $\Lambda\Lambda' + \Psi$ is sometimes referred to as the covariance structure.

Fitting the model consists in determining Λ and Ψ, which make the fitted and observed covariances of the $\mathbf{x}$'s as close as possible. This can be done without invoking the distributional assumptions at all since the covariance structure derived from eqn [6] is the same, regardless of the assumptions about the form of the distributions of $\mathbf{z}$ and $\mathbf{e}$. However, without such assumptions nothing can be said about goodness of fit or the sampling variation of the estimators.

Posterior analysis allows the distribution of $\mathbf{z}$ given $\mathbf{x}$ to be determined, which for the factor model leads to

$$\mathbf{z}|\mathbf{x} \sim N\left(\Lambda \sum^{-1} \mathbf{x} - \mu\right), \left(I + \Lambda' \psi^{-1} \Lambda\right) \qquad [8]$$

where $\sum = \Lambda\Lambda' + \Psi$. As predictors of $\mathbf{z}$ given $\mathbf{x}$, we can use the mean values of this distribution, replacing the parameters by their sample estimates.

Interpretation and Rotation in the Case of the Factor Model

The interpretation of factors involves procedures that are common to all linear latent variable models. They are most familiar in the context of factor analysis so, although treated here, their generality must be emphasized. The problem may arise in two forms. If there is some idea of what factors to expect (as in the arithmetical ability example above), the question is how to confirm their existence. If, on the other hand, the analysis is purely exploratory, further guidance is needed on how to assign meaning to any factor that is found. The essential idea for doing this is to look at the relationships between the latent and manifest variables. This is facilitated by noting that the element λ_{ij} of Λ may be interpreted as the correlation between $\mathbf{x}_i$ and $\mathbf{z}_j$.

Manifest variables, which are highly correlated with $\mathbf{z}_i$ say, have a lot in common with $\mathbf{z}_i$ which, in turn, is highly

influential in determining those **x**'s. It must then be asked what it is that those **x**'s have in common with one another which they also share with $\mathbf{z}_i$. Interpretation is often facilitated by a process known as 'rotation.' Thus far the treatment has supposed that there was only one solution to the problem of fitting the model. In fact there are infinitely many models, which all predict the same covariance structure. These solutions can be generated from one another by a procedure known as 'rotation' (because of its geometrical interpretation). Each such solution represents an equivalent way of describing the latent space, but some ways may be easier to interpret than others. For example, a solution for Λ which has, at most, one nonzero value in each row implies that the **x**'s have been divided into disjoint groups, each group depending on a single latent variable. That latent variable is then interpreted in the light of what the members of that subset have in common.

Latent Class Models

If the latent space consists of a finite (usually small) set of classes and if the manifest variables are also categorical, the bottom right cell of Table 1 applies. The commonest case is where the manifest variables are binary. If there are c latent classes, the prior distribution $f(\mathbf{z})$ becomes a discrete probability distribution over the classes, η_j being the probability of belonging to class j. In this case these probabilities can be treated as unknown parameters to be estimated. For convenience, the two values which each $\mathbf{x}_i$ takes are coded as 1 and 0, the conditional distributions $f_i(\mathbf{x}_i|\mathbf{z})$ consists of two probabilities: $\Pr\{\mathbf{x}_i = 1|\mathbf{z}\}$ and $\Pr\{\mathbf{x}_i = 0|\mathbf{z}\}$. The obvious choice for the distribution of $\mathbf{x}_i$ is the Bernoulli distribution written as

$$f_i(x_i|j) = \{\pi_i(j)\}^{x_i}\{1-\pi_i(j)\}^{1-x_i} \quad x_i = 0,1 \quad j = 1,2,\ldots,c \qquad [9]$$

The joint probability distribution of **x** is then

$$f(x) = \sum_{j=1}^{c}\eta_j\prod_{i=1}^{p}\{\pi_i(j)\}^{x_i}\{1-\pi_i(j)\}^{1-x_i} \qquad [10]$$

This can be used to construct the likelihood function from which the parameters may be estimated.

Posterior analysis for this case is concerned with allocating individuals to latent classes after **x** has been observed and this is done easily by substituting into eqn [3]. This gives us the probability that an individual with observed vector **x** belongs to each of the classes. The number of parameters in this model increases rapidly with c and serious problems of identifiability arise if there are more than, say, four or five classes. Also the standard errors of the parameters estimates also increase rapidly as c increases.

Latent Trait Models

These models were devised, primarily, mainly for use in educational testing where the latent trait refers to some ability. Thus there is usually only one latent variable, representing that ability, and many indicators. The latter are often binary, corresponding to responses to the items being 'right' or 'wrong.' Latent trait modeling has become a specialized field, often referred to as item response theory with a literature and notation of its own. The model may also be used in other fields, such as sociology, where it may be more appropriate to introduce several latent variables.

A latent trait model with binary **x**'s is similar to a latent class model. The prior distribution is now continuous and will usually be taken as standard normal. The response variables will be Bernoulli random variables but now they depend on the continuous latent variables. Since the Bernoulli distribution is a member of the exponential family the appropriate form for $\pi_i(\mathbf{z})$ turns out to be

$$\text{logit}\ \pi_i(\mathbf{z}) = \alpha_i + \alpha_{i1}\mathbf{z}_1 + \alpha_{i2}\mathbf{z}_2 + \cdots + \alpha_{iq}\mathbf{z}_q \qquad [11]$$

Other versions of the model that exist are widely used in which the logit function on the left-hand side of eqn [11] is replaced by $\phi^{-1}(.)$, (the inverse standard normal distribution function). These give very similar results to eqn [11] but they lack the sufficiency properties. If $j = 1$ and if the α_{ij}s are the same for all i, the model is a random effects version of the Rasch model (see, for example, Bartholomew, 1996). In the latter model the trait values are taken as parameters rather than random variables so, in the strict sense of classical inference, the Rasch model is not a latent variable model.

Other Models

The latent profile models of Table 1 have not been included here because they have found relatively little practical application. However, it has also been shown by Molenaar and von Eye (1994) that they are virtually indistinguishable from factor models in the sense that they have an equivalent covariance structure. Nor have we considered hybrid models in which manifest and latent variables may be mixed type – discrete or continuous. A particular advantage of our general approach is that it lends itself to the treatment of all such models as exemplified, for example, in Bartholomew et al. (2011).

Relationships between Latent Variables

In much sociological research, the interest is not so much in the latent variables themselves as individual entities, but in the relationships between them. The interrelationship of manifest variables is a major subject of investigation in social science covering techniques such as path analysis, regression, log-linear analysis, and graphical models. It is natural to want to extend these models to include latent variables. This has given rise to linear structural relations modeling, which can be implemented using widely available software packages such as LISREL, Mplus, AMOS, and EQS.

In essence, such models have two parts: (1) a measurement model which supposes that each latent variable is linked to its own set of indicators through a factor model and (2) a structural model which specifies the (linear) relationships among the latent variables. Such a model is fitted by equating the observed and theoretical covariance matrices. A much more general framework, which allows a wider range of models, is provided by Bartholomew et al. (2011). Although these methods are very widely used, serious questions have been raised about the identifiability of the models (see Anderson

and Gerbing, 1988; Croon and Bolck, 1997; Bartholomew et al., 2011). These authors have suggested that it may be better to separate the measurement and structural parts of the analysis. This can be done by constructing 'estimates' of the latent variables and then exploring their interrelationships by the traditional methods used for manifest variables.

Estimation of Parameters

The feature of latent variable models which has posed considerable practical problems in the past has been the large number of unknown parameters that have to be estimated. For example, a standard factor analysis model with p variables and q factors involves $pq + q$ parameters and since p may easily exceed 10 or, even 100 this leads, inevitably, to large numbers of parameters. In the early days of factor analysis, great ingenuity had to be exercised to devise methods of estimating the parameters which were computationally feasible. The calculation of standard errors was almost impossible and this led, perhaps, to overoptimistic claims about what had been established. Modern computing facilities have radically altered the situation but words of caution are still in order.

General methods such as maximum likelihood are, of course, available and they have the great advantage of yielding asymptotic standard errors and covariances of the estimators. Bayesian methods are, likewise, not restricted, in principle at least, by the number of parameters. The latter provide estimates of the posterior distributions of the parameters, which provide a more complete picture of the uncertainty surrounding the estimates. Indeed, in a certain sense, the two approaches may be said to have converged when there is a large number of parameters to be estimated; the justification for this remark will become clearer as we proceed.

The standard asymptotic theory of maximum likelihood estimation depends on the behavior of the likelihood in the immediate neighborhood of the maximum. Empirical work with the method has shown that the sampling distribution only approaches the normal form implied by the asymptotic theory, if the sample size is very large indeed. This means that the asymptotic standard errors may not be an adequate approximation in many practical situations. Furthermore, the approach to the maximum of the standard iterative methods may be extremely slow. With problems involving relatively small number of parameters and with adequate computing power, this is not usually a serious problem and the software, which most practical applications use has been based, has often been equal to the task. Nevertheless, there are considerable attractions in moving to Bayesian methods, especially as the number of parameters increases.

Bayesian methods involve approximating the posterior distribution of the parameters. This enables one to find the maximum of this distribution and, if required, other properties of the posterior distribution. Provided that the prior distribution chosen for the parameters is fairly 'flat' there will be very little difference between the posterior distribution and the likelihood function in the neighborhood of the maximum. Hence the method is thus, for all practical purposes, an approximation to the method of maximum likelihood.

The method, known as Markov Chain Monte Carlo, generates the empirical posterior distribution from which summary values, like the maximum, can be estimated. The Markov Chain part of the name comes from the fact that the posterior distribution emerges as the steady state of a certain Markov chain. Repeated empirical realizations of the chain then generate a random sample from the posterior distribution. The method, illustrated numerically, is described in Bartholomew et al. (2011).

The particular attractiveness of the method is that it involves a very large number of very simple steps – a task for which computers are ideally suited. It therefore offers a way forward in what have, hitherto, been extremely intractable numerical problems.

Historical Background

Factor analysis was invented by Spearman (1904) for the specific purpose of operationalizing the notion of general intelligence. He remained wedded to the view that there was a single dominant latent variable of human ability which he named 'g' which accounted for most of the variation in observed performance. This idea was challenged by Godfrey Thomson who introduced what he first called the 'sampling of bonds' model which had an identical correlation structure (see, for example, Thomson, 1938). Over the years, Thomson was at pains to emphasize that his model had been produced primarily to show that there were more than one model which would account equally well for the data rather than that it corresponded to reality. More recently, what has become known as Thomson's 'bonds model,' has received more attention (see, for example, Bartholomew et al., 2009) as a serious competitor for Spearman's model in that there is no good biological ground for distinguishing between them. It can also be generalized in various ways which have considerable practical plausibility.

Following Spearman, Thurstone (1947) emphasized the multifactorial approach and, for that purpose, generalized Spearman's treatment to allow for several factors. At that time, factor analysis was ahead of its time in two senses. First, it was a sophisticated multivariate technique introduced long before the statistical technology was available. In consequence, its development was somewhat idiosyncratic and its separation from the statistical mainstream gave it an alien appearance, which persists to this day and still causes some suspicion and misunderstanding. Second, it is a computer-intensive technique whose full potential could only be realized with the arrival of the powerful electronic computers in the 1980s. This lack distorted the early development of the subject by the need to concentrate on the search for computational simplifications.

Latent structure analysis, which covers the remaining cells of **Table 1**, was introduced by Lazarsfeld after the World War II, with a view to its applications in sociology. A book-length treatment, by Lazarsfeld and Henry (1968), has held the field for many years as the definitive source of information but this has now been superseded by a new generation of books of which Heinen (1996) is both comprehensive and relatively recent.

Latent structure analysis came on the scene half a century after factor analysis, and in a different disciplinary context. For that reason there was, until very recently, little cross-fertilization between the two fields. When viewed from a level of sufficient generality, the methods are essentially the same.

In the postwar era, more powerful methods became available but the whole field has largely remained a set of distinct subspecialisms each with its own literature and technology. It is over 40 years since Anderson (1959) pointed out that all latent methods shared a common conceptual framework but his insight has been slow to yield its full potential. The classification of methods as given in **Table 1** was made on the basis of a common approach in Bartholomew (1984), the culmination of which is to be seen in Bartholomew et al. (2011).

Future Developments

The advent of massive computer power has changed the practice of multivariate analysis radically, and of latent variable analysis in particular. It is not possible to foresee all of the consequences of this development. The limiting factor is no longer computing power but of getting data of sufficient quality and quantity to justify the very complicated models which the theory provides and computers can handle. The precision with which models with many parameters can be estimated may often be very low unless the sample size runs into many thousands. This makes it all the more important to estimate the sampling variability adequately of the estimates of the parameters on which the interpretation depends. Traditionally this has been done by finding asymptotic standard errors but these can be very imprecise, as we have already remarked. However, it is now possible to supplement these results by resampling methods such as the bootstrap or other straightforward simulation methods.

There are other fields in which latent variable models are, or may be, used which currently exist in isolation. Some of these are covered in Bartholomew (2013). One generalization, which may not be obvious, is the latent time series. Some work has been done for the case where the latent process is a Markov chain. In this area the term 'hidden' is often used instead of 'latent' and this helps to conceal the family connections (for an introduction see Zucchini and MacDonald, 2009). An application in a more traditional time series context will be found in Harvey and Chung (2000). Also, there has been a great deal of important work done by economists on what they call unobserved heterogeneity. This essentially involves the introduction of latent variables into econometric models and there is substantial scope for new work in this area.

Notes on Bibliography

The literature in this field is enormous. On the one hand, there are classics dating back to the early years of the last century or, in the case of latent structure models, almost as long. These are distinguished from current work by three things. First they were written before the computer era and much attention is typically given to the complications of computation. Second, they were written, largely, from within particular disciplinary perspectives and hence suffer from lack of recognition of the wider context. Third, being mainly the work of those whose interests lay in substantive fields, they are strong on substantive relevance but sometimes weak in technical generality. Some examples are provided in the following Bibliography but they should only be read in the light of these limitations.

A useful source on factor analysis is Cudeck and McCullum (2007). This publication marked the 100th anniversary of Spearman's introduction of the method and it looks backward and forward. A very wide-ranging treatment, covering latent structure methods more generally, is provided by Skrondal and Rabe-Hesketh (2004). A useful text on the theoretical basis of latent variable methods is provided by Bartholomew et al. (2011). At the textbook level, for students in the social and behavioral sciences, there is a good deal relating to latent variable analysis in Bartholomew et al. (2008). The references in these books will lead on to much reliable information for those new to the field.

See also: Bayesian Graphical Models and Networks; Bayesian Models in Neuroscience; Bayesian Statistics; Bayesian Theory, History, Applications, and Contemporary Directions; Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis; Factor Analysis and Latent Structure: Irt and Rasch Models; Latent Structure and Causal Variables; Markov Chain Monte Carlo Methods; Monte Carlo Methods and Bayesian Computation: Importance Sampling; Monte Carlo Methods and Bayesian Computation: Overview; Psychometrics: Preference Models with Latent Variables.

Bibliography

Anderson, T.W., 1959. Some scaling models and estimation procedures in the latent class model. In: Grenander, U. (Ed.), Probability and Statistics. Wiley, New York, pp. 9–38.

Anderson, J.C., Gerbing, D.W., 1988. Structural equation modeling in practice: a review and recommended two-step approach. Psychological Bulletin 103, 411–423.

Bartholomew, D.J., 1984. The foundations of factor analysis. Biometrika 71, 221–232.

Bartholomew, D.J., 1996. The Statistical Approach to Social Measurement. Academic Press, San Diego, CA.

Bartholomew, D.J., 2013. Unobserved Variables: Models and Misunderstandings. Springer, Heidelberg.

Bartholomew, D.J., Deary, I.J., Lawn, M., 2009. A new lease of life for Thomson's bonds model. Psychological Review 116, 567–579.

Bartholomew, D.J., Knott, M., Moustaki, I., 2011. Latent Variable Models and Factor Analysis: A Unified Approach, third ed. John Wiley and Sons Ltd., Chichester.

Bartholomew, D.J., Steele, F., Moustaki, I., Galbraith, J.I., 2008. Analysis of Multivariate Social Science Data, second ed. CRC Press, Boca Raton, FL.

Croon, M., Bolck, A., 1997. On the Use of Factor Scores in Structural Equations Models. Technical Report 97.10.102/7. Work and Organization Research Centre, Tilburg University.

Cudeck, R., McCullum, R.C. (Eds.), 2007. Factor Analysis at 100: Historical Developments and Future Directions. Lawrence Erlbaum Associates Inc., Mahwah, NJ.

Harvey, A., Chung, C.-H., 2000. Estimating the underlying change in unemployment in the UK. Journal of the Royal Statistical Society A 163, 303–339.
Heinen, T., 1996. Latent Class and Discrete Latent Trait Models: Similarities and Differences. Sage, Thousand Oaks, CA.
Lazarsfeld, P.F., Henry, N.W., 1968. Latent Structure Analysis. Houghton Mifflin, New York.
Molenaar, P.C.W., von Eye, A., 1994. On the arbitrary nature of latent variables. In: von Eye, A., Clogg, C.C. (Eds.), Latent Variables Analysis: Applications for Developmental Research. Sage, Thousand Oaks, CA.
Skrondal, A., Rabe-Hesketh, S., 2004. Generalised Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Chapman and Hall/CRC, Boca Raton, FL.
Spearman, C., 1904. General intelligence, objectively determined and measured. American Journal of Psychology 15, 201–293.
Thomson, G.H., 1938. The Factorial Analysis of Human Ability. University of London Press, London.
Thurstone, L.L., 1947. Multiple Factor Analysis. University of Chicago Press, Chicago.
Zucchini, W., MacDonald, I.L., 2009. Hidden Markov Models for Time Series: An Introduction Using R. Chapman and Hall/CRC Press, Boca Raton, FL.

Factor Analysis and Latent Structure: IRT and Rasch Models

Brian W Junker, Carnegie Mellon University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 8, pp. 5244–5249, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the editor.

Abstract

Educational assessments, psychological scales, and social surveys often generate data in the form of a two-way array of discrete response data for N subjects (examinees, respondents, etc.) and J probes (test, scale, or survey items). Subjects' responses to the probes may be viewed as fallible measures of one or more underlying constructs, represented generically by a scalar or vector-valued index {latent variable} for each subject. When the response variables are discrete-valued, item response theory (IRT) and its variants provide a set of modeling and estimation tools for making inferences about the relationships between the X_{ij} and the latent variable(s) θ_i and about subjects' location in the latent structure represented by θ. This article reviews standard parametric IRT models and estimation methods, extensions to compelx data analysis setting by way of a hierarchical Bayes formulation, nonparametric approaches to IRT, and selected applications in education, psychology, the social sciences, and biostatistics. Extensions of IRT to handle cognitively diagnostic assessment are briefly sketched, and similarities between IRT and Bayesian inference networks are pointed out.

In education and the social sciences, we often ask subjects to respond to a set of items (questions, statements, or tasks) on survey forms, self-report inventories, and mental tests, that are coded as discrete – often dichotomous – variables. In many settings it is natural to think, in analogy with factor analysis (*see* Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis; Factor Analysis and Latent Variable Models in Personality Psychology), that there is one or more continuous latent variables for each subject – such as political agency, extrovertedness, or ability in some area of mathematics – that can be measured or estimated using the positive and negative responses to these items. Item Response Theory (IRT), and specializations such as the Rasch Model, refer to a set of statistical modeling and estimation tools for making inferences about one or more continuous latent variables from this kind of multivariate discrete data.

Introduction: the Rasch Model

As an example, consider a test of transitive reasoning in school children (e.g., Sijtsma and Verweij, 1999). An item may consist of three rods, A, B, and C, of differing lengths. The child is given evidence that Length(A) < Length(B), and Length(B) < Length(C), and is asked to deduce a relationship between Length(A) and Length(C); his or her answer is scored 1 if it is correct and 0 if it is incorrect. Let us say that J such items, involving different types of objects and attributes (balls of varying weight, disks of varying area, etc.) and different numbers of objects, make up a test given to N school children. If the scored responses were continuous we might build a two-way additive ANOVA model in which child i's response to item j, X_{ij}, is the sum (or difference) of the child's achieved transitive reasoning level, and the item's innate difficulty, as determined by the attribute on which objects are compared and the number of objects being compared.

However, since the responses are dichotomous $X_{ij} = 0$ or 1, we need a model specifying the *probability* that $X_{ij} = 1$, in terms of the child's transitive reasoning ability and the item's difficulty. One of the earliest such models, developed by the Danish mathematician Georg Rasch (Rasch 1960), specifies

$$P_j(\theta_i) \equiv P\left[X_{ij} = 1 \middle| \theta_i, \beta_j\right] = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}, \quad i = 1, \ldots, N, j = 1, \ldots, J \qquad [1]$$

where θ_i represents child i's transitive reasoning level, and β_j represents the *difficulty* of the item: as the latent variable θ_i increases, the probability of getting the item right increases, and as the difficulty parameter β_j increases, the probability of getting the item right decreases. The Rasch model, together with similar models using different forms for $P_j(\theta_i)$ developed by Jane Loevinger (Loevinger, 1948) and Frederic M Lord (Lord, 1952), became the core of what is now item response theory.

On the logit scale, the Rasch model has exactly the two-way additive ANOVA structure we sought:

$$\text{logit } P_j(\theta_i) = \ln \frac{P_j(\theta)}{1 - P_j(\theta_i)} = \theta_i - \beta_j.$$

We further assume local independence among item responses, i.e., all X_{ij} are conditionally independent, given θs and βs. Letting x_{ij} denote the observed values of X_{ij}, we may write the likelihood for the Rasch model as

$$\begin{aligned} P\left[X_{ij} = x_{ij}, i = 1, \ldots, N, j = 1, \ldots, J \middle| \theta_1, \ldots, \theta_N; \beta_1, \ldots, \beta_J\right] \\ = \prod_{i=1}^{N} \prod_{j=1}^{J} P_j(\theta_i)^{x_{ij}} \left[1 - P_j(\theta_i)\right]^{1 - x_{ij}}. \end{aligned} \qquad [2]$$

An up-to date account of the Rasch model and its various extensions is given by Fischer and Molenaar (1995), to whom we also refer for historical notes and references.

The Rasch model in Eqn. [1] has the interesting invariance property that the odds ratio comparing two items j and k is

$$\frac{P_j(\theta_i)}{1-P_j(\theta_i)}\cdot\frac{1-P_k(\theta_i)}{P_k(\theta_i)} = \exp\beta_k - \beta_j$$

independent of θ_i. Hence in our example, the 'difficulty' of two items involving different transitive reasoning settings can be compared, in principle, using any convenient sample of subjects regardless of their transitive reasoning level (in practice, subjects still contribute at least 'small sample bias' to the estimated βs). Similarly, the odds ratio comparing two subjects is independent of the item difficulty parameters. These two invariance properties are instances of 'specific objectivity,' a property Rasch considered important in defining a measurement model and from which Eqn. [1] can be derived (see Fischer and Molenaar, 1995). The Rasch model is also closely related to the Bradley–Terry model of paired comparisons. Consider two items j and k for which $\beta_j < \beta_k$. For any subject i who got only one item right, it follows from Eqn. [1] that the probability that the easier item was correct is

$$P\left[X_{ij} = 1 \middle| X_{ij} + X_{ik} = 1, \theta_i, \beta_j, \beta_k\right] = \frac{\exp(\beta_k - \beta_j)}{1+\exp(\beta_k - \beta_j)} \quad [3]$$

independent of θ_i; this is exactly the Bradley–Terry model with parameters $\phi_j = -\beta_j$. The lack of dependence on θ_i on the right can also be used to construct empirical tests of the Rasch model, since the probability in Eqn. [3] can be estimated without assuming the Rasch model and should be the same in any subpopulation of subjects in which the Rasch model applies.

Estimation in the Rasch Model

Two estimation methods are traditionally associated with the Rasch model, joint maximum likelihood (JML) and conditional maximum likelihood (CML). One can maximize Eqn. [2] jointly in the θs and βs, by numerically solving the 'normal equations' obtained by setting all partial derivatives of the logarithm of Eqn. [2] equal to zero; the maximizing values $\widehat{\theta}_i$ and $\widehat{\beta}_j$ are called joint maximum likelihood estimators (JMLEs). However, the JMLEs turn out to be *inconsistent*, or asymptotically biased: suppose we are primarily interested in estimating β_js; it need not be the case that each $\widehat{\beta}_j$ converges to β_j as we increase the subject sample size N, keeping the number of items J fixed. Indeed, as N increases, more θ_is are also added to the model: intuitively, some information must be expended on estimating them, and there is not enough information left over to improve the estimates of the β_is. One solution to this problem is to exploit the fact that in the Rasch model, a sufficient statistic for each θ_i is subject i's total number correct score, $X_{i+} = \sum_{j=1}^{J} X_{ij}$. If the likelihood in Eqn. [2] is divided by the joint likelihood for the observed x_{i+}s, the resulting conditional likelihood contains only the β_j parameters, and the conditional maximum likelihood estimators (CMLEs) $\widehat{\beta}_j$ obtained by maximizing this new likelihood are consistent (asymptotically unbiased) as N grows and J stays fixed. For this reason, CML estimates are usually preferred over JML estimates. Additional details are provided by Andersen (1980), whose earlier work demonstrated the inconsistency problem with JML, and established consistency of CML for the Rasch model.

Another approach to solving the inconsistency problem with JML is to assume the θ_i are independent random effects following a common (discrete or continuous) distribution $f(\theta|\lambda)$. Integrating over θ_i for each subject yields the marginal likelihood

$$\prod_{i=1}^{N}\int\prod_{j=1}^{J} P_j(\theta_i)^{x_{ij}}\left[1-P_j(\theta_i)\right]^{1-x_{ij}} f(\theta_i|\lambda)\mathrm{d}\theta_i \quad [4]$$

and maximizing this with respect to β_is and λ (using an E-M algorithm, for example) yields what are called maximum marginal likelihood (MML) estimates $\widehat{\beta}_j$ and $\widehat{\lambda}$. These MML estimates are also consistent (asymptotically unbiased). The method based on Eqn. [4] can be interpreted as an empirical Bayes method; it also links the Rasch model directly with other latent variable approaches such as factor analysis, where the latent variable θ_i is treated as an unobserved random variable, or equivalently as completely missing data (*see* Factor Analysis and Latent Variable Models in Personality Psychology).

Finally, it can be shown that under Eqn. [4], the marginal distribution of the data follows a log-linear model for the 2^J table of the form

$$\ln p(x_{i1},\ldots,x_{iJ}) = \alpha - \sum_{j=1}^{J}\beta_j x_{ij} + \sum_{k=0}^{J}\gamma_k \chi(x_{i+} = k)$$

where $x(S) = 1$ if statement S is true, and 0 otherwise. This is the log-linear model of quasisymmetry; the connection between the Rasch model and the quasi-symmetry model has been independently discovered innumerable times in the literature. Lindsay et al. (1991) give a reasonably complete account of the connection with quasisymmetry and consequences for the identifiability of the distribution $f(\theta \mid \lambda)$ in a semiparametric formulation of Eqn. [4].

Parametric IRT

Other parametric IRT models have been developed for situations in which the Rasch model does not fit, while retaining the ideas of dichotomous ($X_{ij} \in \{0, 1\}$) or at least polytomous ($X_{ij} \in \{0, 1, \ldots, K-1\}$) item scores, a unidimensional (or low-dimensional) latent variable θ_i characterizing subjects, a low-dimensional parameter β_j characterizing items, and local (conditional) independence of all item responses X_{ij} given all parameters. An accessible review of developments in parametric IRT in the 1960s, 1970s and 1980s is given by Hambleton et al. (1991).

A simple generalization of the Rasch model that is often useful for dichotomously-scored items,

$$P_j(\theta_i) = \frac{\exp\left[\alpha_j(\theta_i - \beta_j)\right]}{1+\exp\left[\alpha_j(\theta_i - \beta_j)\right]} \quad [5]$$

is called the two-parameter logistic (2PL) model. In the 2PL model, the slope parameters $\langle_{jk}$ are called 'discrimination parameters,' and they control the Fisher information for estimating the θs. Again, the location parameters β_j play the role of

'difficulty parameters'; they also determine the value of θ maximizing the Fisher information. The three-parameter logistic (3PL) model adds a nonzero lower asymptote to Eqn. [3], to model random or exchangeable response behavior by low-θ subjects. Further generalizations, discussed in the next several paragraphs, are treated at length by Fischer and Molenaar (1995) and van der Linden and Hambleton (1997).

Additional covariates and other structures can be incorporated into these models as regression terms. For example, in the linear logistic test model (LLTM), the difficulty parameters β_j in the Rasch model are rewritten as linear combinations of K basic parameters $/_k$ with weights q_{jk}, and

$$\text{logit } P_j(\theta_i) = \theta_i - \sum_{k=1}^{k} q_{jk}\eta_k.$$

The matrix $Q = [q_{jk}]$ is usually obtained as a regression model matrix from *a priori* analysis of the items; for example, in the transitive reasoning example described in Sect. 1, the q_{jk} might indicate what kind and how many objects were used for each item. By contrast, multidimensional compensatory IRT models decompose the unidimensional θ_i parameter into an item-dependent linear combination of underlying traits, e.g.,

$$\text{logit } P_j(\theta_i) = \sum_{d=1}^{D} \alpha_{jd}\theta_{id} - \beta_j.$$

For example, mathematics word problems may involve both a math component, θ_{i1}, and a verbal component θ_{i2} in different proportions in different problems, as determined by the discrimination parameters $\langle_{jd}$. In some settings, the $\langle_{jd}$ may be fixed *a priori*, like the q_{jk} in the LLTM model, and in other settings they may be estimated analogously to factor loadings in Factor Analysis. Multiplicative or conjunctive IRT models combine unidimensional models for components of response in a product $P_j(\theta_i) = \Pi_d^D = 1\, P_{jd}(\theta_{id})$ where $P_{jd}(\theta_{id})$ are parametric unidimensional dichotomous response functions. The usual interpretation is that the $P_{jd}(\theta_{id})$ are probabilities of correctly performing component skills or subtasks, which must be done in conjunction in order to generate a correct response to the item itself; this model is sometimes referred to as the multicomponent latent trait model (MLTM).

Probit versions of all the above models, in which the logit function is replaced with $\Phi^{-1}(p)$, the inverse standard normal distribution function, are also in use. Aside from a trivial rescaling of the parameters, there is little difference in practice between the probit and logit models; the choice depends mostly on computational expedience.

IRT model can also be extended to handle polytomous or multicategory items, $X_{ij} \in \{0, 1, 2, \ldots, K-1\}$, in much the same way that logistic regression models are extended polytomous data: in the 'graded response model' (GRM) the cumulative logit, logit $P[X_{ij} > c|\theta_i, \beta_j]$, is linear in θ_i; in the 'sequential model' (SM) the continuation-ratio logit, logit $P[X_{ij} > c|X_{ij} > c-1, \theta_i, \beta_j]$, is linear in θ_i; and in the 'partial credit model' (PCM) the adjacent-category logit, logit $P[X_{ij} = c+1|X_{ij} \in \{c, c+1\}, \theta_i, \beta_j]$, is linear in θ_i. Again, probit reformulations of these models are straightforward. When the discrimination parameters are all non-negative, the cumulative response curves in any of these models is nondecreasing (Hemker et al. 1997), a condition referred to as 'monotonicity.' By contrast, IRT formulations of direct-response probabilistic unfolding models, also known as proximity models or parallelogram models, have unimodal response curves that peak near the location β_j of the item.

As was the case for the Rasch model discussed in Sect. 2, JML estimates for IRT models are relatively easy to set up but generally inconsistent (asymptotically biased); CML methods can be used when simple sufficient statistics are available. MML methods based on applying an E-M algorithm to Eqn. [4], with $P_j(\theta_i)$ redefined according to one of the above parametric forms, are most common, however. The recent development of Markov Chain Monte Carlo (MCMC) integration methods for Bayesian statistics (*see* Markov Chain Monte Carlo Methods) provides an alternative to these methods. Patz and Junker (1999) sketch a general MCMC methodology for parametric IRT models, and Johnson and Albert (1999) survey a range of related applications from a working Bayesian statistician's perspective.

The IRT toolbox has been pushed forward in recent years by the needs of both large-scale educational assessment surveys (e.g., Johnson et al. 1994) and cognitively based assessment models (e.g., Nichols et al. 1995). To accommodate finer structure in the latent space, as well as to incorporate subject and item covariates in the model, it is useful to reformulate the IRT framework as a two-way hierarchical Bayes structure for N individuals and J response variables, as in **Figure 1**. In addition to the item response functions $P_j(\theta_i) = P(\theta_i; \beta_j)$ in **Figure 1**, f_i and g_j are prior distributions on the latent variables θ and item parameters β respectively (both of which may be multidimensional), and $\lfloor_f$ and $\lfloor_g$ represent sets of hyperparameters needed to specify these prior distributions.

$$\text{First level:} \quad X_{ij} \sim P(\theta_i; \beta_j)^{X_{ij}}\,[1 - P(\theta_i; \beta_j)]^{(1-X_{ij})},$$

$$i = 1,\ldots, N;\ j = 1,\ldots, J$$

$$\text{Second level:} \quad \theta_i \sim f_i(\theta|\lambda_f), \quad \text{each } i$$

$$\beta_j \sim g_j(\beta|\lambda_g), \quad \text{each } j$$

$$\text{Third level:} \quad \lambda_f \sim \phi_f(\lambda_f)$$

$$\lambda_g \sim \phi_g(\lambda_g)$$

Figure 1 Two-way hierarchical structure for N individuals and J dichotomous response variables. Factors in the first level are independent and are multiplied together to form a likelihood for the response data matrix $[X_{ij}]$. Factors in the second and third levels are also independent, and covariates and special structure for the latent variables and item parameters may be introduced in these levels.

The model in **Figure 1** is expressed for dichotomous items but can be generalized easily to polytomous items or combinations of item types, incomplete data, and so forth (e.g., Patz and Junker 1999). The MML formulation discussed in the previous paragraph is obtained by taking $f_i(\theta)$ to be a single population distribution not dependent on j for the latent trait, and taking g_j to be flat priors not depending on j for the item parameters.

Nonparametric IRT

Nonparametric IRT refers to several related methodologies for working with item response data without fully committing to any particular well-known family of parametric item response theory models. Some topics in nonparametric IRT are considered in van der Linden and Hambleton (1997) and Boomsma et al. (2001), and several current issues are surveyed in a special issue of *Applied Psychological Measurement* (Junker and Sijtsma, 2001).

Modern nonparametric scaling methods include the 'essential independence' (EI) approach of Stout and his students and colleagues (e.g., Stout et al., 1996), and the Mokken scaling approach (e.g., Hemker et al., 1995); both methods are related to Cronbach's alpha bound on reliability. Mokken scaling is very good at finding groups of items that are highly discriminating among subjects; Stout's methods tend to identify groups of items that narrowly satisfy LI/EI, M, and U better. Ramsay (e.g., Ramsay, 1996) has developed computational tools for estimating item response functions nonparametrically, and for visualizing the dimensionality of θ in terms of the surface in the J-dimensional unit cube generated by the joint likelihood for $X_{i1}, \ldots, X_{iJ}$ as θ_i varies.

Some Applications

Parametric and nonparametric IRT methods are routinely used to assess the quality of individual items and sets of items in educational measurement work (Linn, 1989) and to elucidate experimental design and statistical estimation issues when performances of examinees who sit for different versions of the same exam must be compared; such work goes under the name test equating. A similar problem arises in the scoring of sequentially-designed, computerized adaptive tests (e.g., Sands et al., 1997). IRT modeling has also elucidated research into the assessment of sociological bias (differential item functioning) on standardized test items (Holland and Wainer, 1993) and the detection and diagnosis of subject outliers (Meijer, 1996). Related work includes the study of rater effects (e.g., Patz et al. 2000) and matrix-sampling and the incorporation of group effects and other item and subject covariates in educational assessment survey work (Johnson et al., 1994).

In the broader social sciences and related areas, IRT has also played a substantial role. The Rasch model has long been applied in social survey work; see, for example, Duncan (1985). The Rasch model and other IRT models can also be applied to model subject heterogeneity in estimating a closed population from multiple-recapture census data (*see* Censuses: History and Methods); Fienberg et al. (1999) recently surveyed this methodology and compared Bayesian IRT models with standard log-linear models in this context. Applications to multiple outcomes of designed experiments and to panel data are indicated in the volumes by Fischer and Molenaar (1995) and van der Linden and Hambleton (1997). Sijtsma and Verweij (1999) apply nonparametric IRT to scale construction in developmental psychology; and Nichols et al. (1995) collect many applications of IRT and related Bayesian inference network models (*see* Bayesian Graphical Models and Networks; Mislevy, 1996) in cognitive diagnosis. Applications of parametric and nonparametric IRT to psychiatric scales include Gibbons et al. (1985), Santor et al. (1995) and Kim and Pilkonis (1999). In biostatistics Legler and Ryan (1997) use IRT to model multiple physical outcomes in the study of birth defects.

Summary

Item response theory (IRT) has grown from its roots in postwar mental-testing problems, through intensive use in educational measurements in the 1970s, 1980s, and 1990s, to become a mature statistical toolkit for modeling of multivariate discrete response data using subject-level latent variables. Applications of IRT can be found throughout the social sciences and related areas, from education, psychology, economics, and demography to medical research.

Most parametric IRT models would be recognizable by modern statisticians as mixed-effects multivariate generalized linear models, but IRT has benefited from interaction with all parts of the statistical community. Almost any assessment phenomenon – from between-subject dependence due to institutional or sociological factors, to behavioral aspects of raters, to the analysis of item responses in terms of requisite subject or item features – can be expressed in the hierarchical mixture/Bayes modeling framework of **Figure 1**, because of its conceptual simplicity. Recent advances in computation, and MCMC methods in particular, have made it possible to estimate a vastly wider variety of these models than would have been imaginable even in the early 1990s however the more complicated models typically also make higher sample-size demands. Nonparametric IRT approaches also rely on computationally intensive methods, including spline and kernel smoothing and bootstrap techniques, to estimate and test probability inequalities, stochastic ordering properties, and similar features of the models. Questions motivating both parametric and nonparametric IRT modeling inevitably involve identifying the phenomena that are worth detailed modeling, and seeing if the computational and data collection machinery can be pushed to be informative about reasonable models of these phenomena.

Hambleton et al. (1991) provide a fairly straightforward introduction to IRT in educational testing; Andersen (1980) presents the Rasch model in the context of other statistical models for the social sciences; and Fischer and Molenaar (1995) and van der Linden and Hambleton (1997) provide modern technical accounts. A broad range of current research issues in IRT is collected in Boomsma et al. (2001).

Bibliography

Andersen, E.B., 1980. Discrete Statistical Models with Social Science Applications. North-Holland, New York.

Andrich, David, 2004. Controversy and the Rasch Model: A Characteristic of Incompatible Paradigms? Medical care 42 (1), 1–7.

Béguin, A.A., Glas, C.A.W., 2001. Mcmc Estimation and Some Model-Fit Analysis of Multidimensional Irt Models. Psychometrika 66 (4), 541–561.

Boomsma, A., Snijders, T.A.B., van Duijn, M.A.J. (Eds.), 2001. Essays on Item Response Theory. Springer-Verlag, New York.

De Boeck, P., 2008. Random Item Irt Models. Psychometrika 73 (4), 533–559.

Duncan, O.D., 1985. Some models of response uncertainty for panel analysis. Social Science Research 14, 126–141.

Fienberg, S.E., Johnson, M.S., Junker, B.W., 1999. Classical multilevel and Bayesian approaches to population size estimation using multiple lists. Journal of Royal Statistical Society: Series A 162, 383–405.

Fischer, G.H., Molenaar, I.W. (Eds.), 1995. Rasch Models: Foundations, Recent Developments, and Applications. Springer-Verlag, New York.

Gibbons, R.D., Clark, D.C., von Ammon Cavanaugh, S., Davis, J.M., 1985. Application of modern psychometric theory in psychiatric research. Journal of Psychiatric Research 19, 43–55.

Hambleton, R.K., Swaminathan, H., Rogers, H.J. (Eds.), 1991. Fundamentals of Item Response Theory. Sage, Newbury Park, CA.

Hemker, B.T., Sijtsma, K., Molenaar, I.W., 1995. Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. Applied Psychological Measurement 19, 337–352.

Hemker, B.T., Sijtsma, K., Molenaar, I.W., Junker, B.W., 1997. Stochastic ordering using the latent trait and the sum score in polytomous IRT models. Psymka 62, 331–347.

Holland, P.W., Wainer, H., 1993. Differential Item Functioning. Erlbaum, Hillsdale, NJ.

Johnson, V.E., Albert, J., 1999. Ordinal Data Modeling. Springer-Verlag, New York.

Johnson, E.G., Mislevy, R.J., Thomas, N., 1994. Theoretical background and philosophy of NAEP scaling procedures. In: Johnson, E.G., et al. (Eds.), Technical Report of the NAEP 1992 Trial State Assessment Program in Reading. OERI, US Dept. of Ed, Washington, DC, pp. 133–146. Chap. 8.

Junker, B.W., Sijtsma, K., 2001. Nonparametric IRT in action: An overview of the special issue. Applied Psychological Measurement 25.

Kim, Y., Pilkonis, P.A., 1999. Selecting the most informative items in the IIP scales for personality disorders: An application of item response theory. Journal of Personality Disorders 13, 157–174.

Legler, J.M., Ryan, L.M., 1997. Latent variable models for teratogenesis using multiple binary outcomes. Journal of the American Statistical Association 92, 13–20.

Li, Feiming, Cohen, Allan S., Kim, Seock-Ho, Cho, Sun-Joo, 2009. Model Selection Methods for Mixture Dichotomous Irt Models. Applied psychological measurement 33 (5), 353–373.

Lindsay, B., Clogg, C.C., Grego, J., 1991. Semiparametric estimation in the Rasch model and related exponential response model, including a simple latent class model for item analysis. Journal of the American Statistical Association 86, 96–107.

van der Linden, W.J., Hambleton, R.K. (Eds.), 1997. Handbook of Modern Item Response Theory. Springer-Verlag, New York.

Linn, R.L. (Ed.), 1989. Educational Measurement, third ed. Macmillan, New York.

Loevinger, J., 1948. The technique of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. Psychological Bulletin 45, 507–530.

Lord, F.M., 1952. A theory of test scores. Psychometric Society, New York.

Meijer, R.R., 1996. Person-fit research: An introduction. Applied Measurement in Education 9, 3–8.

Mislevy, R.J., 1996. Test theory reconceived. Journal of Educational Measurement 33, 379–416.

Nichols, P.D., Chipman, S.F., Brennan, R.L. (Eds.), 1995. Cognitively Diagnostic Assessment. Erlbaum, Hillsdale, NJ.

Patz, R.J., Junker, B.W., 1999. Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. Journal of Educational and Behavioral Statistics 24, 342–366.

Patz, R.J., Junker, B.W., Johnson, M.S., 2002. The hierarchical rater model for rated test items and its application to large–scale educational assessment data. Journal of Educational and Behavioral Statistics 27, 341–384.

Ramsay, J.O., 1996. A geometrical approach to item response theory. Behmka 23, 3–17.

Rasch, G., 1960. Probabilistic models for some intelligence and attainment tests. University of Chicago Press, Chicago.

Rizopoulos, Dimitris, 2006. Ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses. Journal of Statistical Software 17 (5), 1–25.

Sands, W.A., Waters, B.K., McBride, J.R., 1997. Computerized Adaptive Testing: From Inquiry to Operation. American Psychological Association, Washington, DC.

Santor, D.A., Zuroff, D.C., Ramsay, J.O., Cervantes, P., Palacios, J., 1995. Examining scale discriminability in the BDI and CES-D as a function of depressive severity. Psychiatry Assessment 7, 131–139.

Sijtsma, K., Verweij, A.C., 1999. Knowledge of solution strategies and IRT modeling of items for transitive reasoning. Applied Psychological Measurement 23, 55–68.

Stout, W., Habing, B., Douglas, J., Kim, H.R., Roussos, L., Zhang, J., 1996. Conditional covariance-based nonparametric multidimensionality assessment. Applied Psychological Measurement 20, 331–354.

Factor Analysis and Latent Variable Models in Personality Psychology

Ralf Schulze, Anna-Lena Hilger, and Philipp M Engelberg, Bergische Universität Wuppertal, Wuppertal, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A generalized model for latent structures is introduced as a general framework for the factor model to give an overview about modeling options in personality research. It is briefly described and it is shown that the factor model is one of many instances of the generalized linear latent variable model. Factor analysis as one of the most important multivariate methods in personality research is then presented with an emphasis on central issues in exploratory factor analysis. The presentation also addresses the role of confirmatory factor analysis as a method in personality research.

Introduction

Among the most central research questions of personality research are the following: What are the fundamental personality constructs and what are their relationships or what is their structure? Since constructs are commonly considered as unobserved or even unobservable variables (Bollen, 2002), personality psychology, like many other subdisciplines of psychology, can mainly be characterized as the investigation of stable psychological attributes that are not directly observed in empirical research. Nevertheless, it is always assumed that the constructs of personality psychology somehow relate to observable behavior, and that these relationships are the basis to make inferences about the existence and relations between the constructs.

The major models that are used in personality psychology to formalize the relationship between unobserved variables and observed indicators are described in this article. Unobserved variables are interchangeably called latent variables or factors in what follows. The focus will be put on basic statistical models with latent variables instead of (1) more philosophical approaches to the question of what the nature of the latent variables actually may be (cf Bollen, 2002; Borsboom et al., 2003) and (2) the many qualitatively different latent structures, like hierarchical and higher order models, for example, that are prominent in personality psychology and ability research (see Schulze, 2005).

Examples of latent variables are personality traits (e.g., extraversion) or abilities (e.g., fluid intelligence) which relate to observable or so-called manifest variables. Typical examples for manifest variables are the responses to the self-report items of a personality test. A basic premise of factor analytic models and the latent variable models considered in this article is that the manifest variables are influenced by one (or more) latent variable(s). That is, the factors are considered to be the cause for the responses. In this sense, the latent variables determine the responses and thereby the value of the indicators. If many indicators are determined by the same factor, which is a common design feature of measurement instruments in personality psychology, then the latent variable is called a common cause or common factor. Given that such common factors exist, the indicators share common variance, the extent of which depends on the strength of the relationships of the indicators with the common factors. A crucial function of factor analytic models therefore is the explanation of the shared variance (i.e., covariances) between manifest variables by the introduction of latent variables.

There are many different factor models that can be distinguished with respect to several theoretical and statistical features. To understand the differences and to be able to make an appropriate choice for a certain measurement purpose, it is helpful to examine the most important features. Among these features are the number and the properties of both latent and manifest variables.

In the statistical latent variable models, both manifest and latent variables can be defined as random variables, and they can be either discrete (categorical) or continuous. A common well-known classification scheme of variables is the concept of level of measurement (e.g., nominal, ordinal, and metric). However, since statistical methods do technically not depend on the level of measurement but mainly on distributional assumptions (Gaito, 1980), the distinction of discrete and continuous random variables is adopted in this article for the classification of both manifest and latent variables.

The distributional assumptions of the variables depend on this classification. A typical example for a discrete manifest variable is an item from an ability measure that can only be answered correctly or incorrectly. If this is the case, then the manifest variable is discrete with two possible values 0 and 1 only (i.e., dichotomous) and it follows a Bernoulli distribution. Also typical for personality research are Likert-type items that are ordinarily treated as categorized continuous variables and assumed to be sampled from a normal distribution.

The classification of latent variables as categorical or continuous depends entirely on theoretical assumptions. Intelligence, for example, is commonly considered to be a continuous trait following a normal distribution. On the other hand, latent variables can be conceptualized as categorical, like in theories of personality types or as is common in the area of personality disorders. In such cases, the latent variables follow a distribution for discrete random variables.

In the following passages, a general framework for latent variable models that includes both categorical and continuous manifest and latent variables will first be described. Next, a particular instance, the factor analytic model, will be presented in more detail as it is the most relevant in terms of frequency of use in personality research, and also because of its historical importance for the field.

A General Framework for Latent Variables

Latent variable models have a rather long and partly fragmented tradition as statistical models for psychological assessment. As a consequence, many psychometric models have been proposed that include different numbers and types of manifest and latent variables. Despite the partial fragmentation, the equivalence has been proven for some models from very different traditions (Takane and de Leeuw, 1987; Kamata and Bauer, 2008). Modern approaches integrate characteristics of diverse models and provide a general framework for latent variable modeling (McDonald, 1999; Skrondal and Rabe-Hesketh, 2004; Bartholomew et al., 2011). Among the benefits of using a general framework is that researchers are more flexible and can make sure that their model fits the data characteristics, instead of adjusting the data to fit the model requirements. It also facilitates insights into the commonalities of seemingly or traditionally different models of measurement. Even if special instances of the general model have been more common in personality research so far, the current development in personality research indicates that a wide range of models will be applied in future (Morizot et al., 2007).

The general framework for latent variables can best be described as a generalized linear model for latent variables (Skrondal and Rabe-Hesketh, 2004). Broadly speaking, the generalized linear framework allows us to model linear relations of both categorical and continuous dependent (manifest) and independent (latent) variables, or more generally, of variables sampled from distributions of the exponential family (see, e.g., Gill, 2000). The exponential family is a group of distributions whose functions can be rewritten in an exponential form. Exemplary members of the exponential family that are familiar to most researchers are the normal and the binomial distribution. Generalized linear latent models provide a framework for measurement models where either categorical or continuous observed responses are predicted by either categorical or continuous latent variables.

The Generalized Model for Latent Variables

The generalized model for latent variables consists of three components which will be very briefly introduced in the following passages. The first component of the model is a matrix of dependent observed variables Y (e.g., responses to items) sampled from a probability function of the exponential family. These random variables can be characterized by their respective expected value. For a normally distributed item (e.g., a self-report item) the expected value would simply be the mean for all respondents in the population. For a Bernoulli-distributed item (e.g., a dichotomous item scored as either correct or incorrect), the expected value would be the probability of choosing the correct option.

The second element of the model is the linear component v that represents an interim dependent variable in the model that is predicted by the latent explanatory variables η using the corresponding weights Λ. The latent explanatory variable can be continuous or categorical.

The third component of the generalized model for latent variables is the link function. With the link function it is possible to link the expected value of the distribution of Y to the linear component v. If Y is assumed to be normally distributed, the link function is the identity link. With the identity link, the expected value of Y can directly be linked with the linear component, resulting in a standard regression model with a latent predictor. If the dependent variable Y is not normally distributed because it is categorical, for example, different link functions have to be used. The expected value of the Bernoulli-distributed dichotomous variable, the probability of success can be linked with the linear component either with a logit link or a probit link. In other words, the link function is the part of the model where the generalization takes places. The use of different link functions provides a means to model not only normally distributed dependent variables but any dependent variables sampled from a distribution of the exponential family.

For the purpose of illustration assume that l respondents answered q items of a personality questionnaire which are determined by k latent variables. The response scale is treated as continuous and the responses are assumed to follow a multivariate normal distribution. The same assumption applies to the latent variables. In this case, the linear component v can be written in matrix notation as

$$v = \Lambda\eta \qquad [1]$$

where η is a $k \times l$ vector of values of the latent variables for every respondent and Λ is a $q \times k$ matrix of weights for every item. Given the assumptions stated above, the linear component can directly be linked to the expected values of the observed variables as follows:

$$E(Y) = \mu = \Lambda\eta \qquad [2]$$

$$Y = \Lambda\eta + \Psi E \qquad [3]$$

Equation [2] is consistent with the generalized linear latent model presentation. In eqn [3] the values of the observed variables are modeled and, hence, the error term E weighted by Ψ is part of the equation.

In the dichotomous case, the observed variables follow a Bernoulli distribution. Nevertheless, an underlying latent variable may still be normally distributed but the above given link function would no longer be appropriate in this case. The logit link is one of two options to link the expected values of Y, the probability of success, to the linear component:

$$E(Y) = P(Y = 1) = \frac{\exp(\Lambda\eta)}{1 + \exp(\Lambda\eta)} \qquad [4]$$

Alternatively, the probit link can be used:

$$E(Y) = P(Y = 1) = \phi(\Lambda\eta) \qquad [5]$$

where ϕ is the standard normal cumulative distribution function. The two models in eqns [2] and [4] differ in type and distribution of the observed variables but the same distributional assumptions apply to the explanatory latent variables in both models.

In sum, the generalized linear model for latent variables provides a very flexible framework for measurement models in psychology. It encompasses many popular models for latent variables as special cases that have often been seen as distinct. For example, the model for dichotomous responses in eqns [4] or [5] is a two-parameter Item Response Theory (IRT) model or

a factor model for categorical (dichotomous) variables (see Christoffersson, 1975; Takane and de Leeuw, 1987; Kamata and Bauer, 2008). Yet another example for a popular special case is the first model in eqns [2] and [3] that is known as a factor model, which will be presented in more detail in the next section. It should also be noted that, in addition to IRT models and factor models, the framework includes latent class models, longitudinal models, multilevel models, and structural equation models (Skrondal and Rabe-Hesketh, 2004).

Factor Analysis

Factor analysis is more than 100 years old and has gradually evolved over time into a multifarious statistical technique that is part of the generalized linear latent variable model class (see Bartholomew, 2007). It is probably not an overestimation of its impact on personality research in particular that without its invention personality theory would not be the same as we know it. Hence, at least a basic understanding of factor analysis is vitally important to understand current personality research. In the following, a brief overview of the most important steps is given. A far more detailed treatment of the technique and associated issues can be found in Mulaik (2010), for example.

Exploratory and Confirmatory Approaches

Factor analysis may be used with different research goals in mind. On the one end of a continuum from exploratory to confirmatory analyses, the goal may be to infer the number and relationships between latent variables underlying an empirical data set, free from any preconceived notions about the true latent structure. Exploratory factor analysis (EFA) is a form of factor analysis that is well suited for this research goal.

On the other end of the continuum, the goal of confirmatory factor analysis (CFA) is to empirically test or assess the tenability of a hypothesized latent structure for a set of observed variables. The major difference between CFA and EFA is that the latent structure has to be specified before any analyses can be done with CFA. Theoretical assumptions and previous empirical findings may be sources of knowledge to specify a factor model in advance before even the data collection begins.

Although the (common) factor model of both EFA and CFA is basically the same, there are many procedural differences that serve to accomplish the different goals of analysis. Hence, it appears to be useful to classify a factor analytic approach as EFA or CFA even when the basic model is the same and the specific goal may be located somewhere between the endpoints of the continuum. The focus of the following presentation will lean toward the EFA end because it has been the approach that is descriptively more characteristic for personality research, although this might well change in the proximate future. Terminology, issues, and problems of using the factor model in personality research will be highlighted in the subsequent sections.

Factor Model

Consistent with the presentation of the generalized model for latent variables above, the factor model can be stated in matrix notation as

$$\mathbf{Y} = \mathbf{\Lambda\eta} + \mathbf{\Psi E} \qquad [6]$$

This function is called the fundamental equation of factor analysis. The vector $\boldsymbol{\eta}$ contains the common factors that are assumed to be causal for the manifest variables and, hence, for their covariances and correlations. $\mathbf{E}$ is a vector of unique factors with expected values of zero and variances of unity. The unique factors may actually be composed of more than one latent variable and/or random errors. It is essential, however, that any latent variable that is an element of $\mathbf{E}$ is causal for only one measured variable and cannot be considered to be common for two or more manifest variables.

Common factors and unique factors are weighted by their respective factor pattern matrices $\mathbf{\Lambda}$ and $\mathbf{\Psi}$, where the former is the matrix of so-called loadings and the latter is a diagonal matrix because unique factors are assumed to be uncorrelated. Since common factors and unique factors are uncorrelated, the correlations between manifest variables ($\mathbf{R}_{YY}$) are only due to (1) correlations among the common factors, noted as $\mathbf{R}_{\eta\eta}$, (2) regression coefficients of common factors and manifest variables, noted as $\mathbf{\Lambda}$, and (3) the squared factor pattern vector of unique factors, $\mathbf{\Psi}^2$. Taken together, this equation is known as the *fundamental theorem of factor analysis*:

$$\mathbf{R}_{YY} = \mathbf{\Lambda R}_{\eta\eta}\mathbf{\Lambda}' + \mathbf{\Psi}^2 \qquad [7]$$

In factor analysis, the relations between the manifest variables are explained in terms of the latent variable structure. The correlation matrix of manifest variables that represents the relations entirely explained by common factors is the so-called reduced correlation matrix given in eqn [8]:

$$\mathbf{R}_C = \mathbf{R}_{YY} - \mathbf{\Psi}^2 \qquad [8]$$

Since $\mathbf{\Psi}^2$ is diagonal, the correlations between the manifest variables are unaffected by the subtraction in eqn [8]. The diagonal of $\mathbf{R}_C$ contains values that are called communalities. The communalities are the proportions of the variances of the q manifest variables that are determined by the k common factors. The communality of a single variable j is defined as the sum of its squared factor loadings in the case of uncorrelated (i.e., orthogonal) common factors, which is a starting assumption in EFA:

$$h_j = \lambda_{j1}^2 + \lambda_{j2}^2 + \ldots + \lambda_{jk}^2 \quad (j = 1, 2, \ldots, q) \qquad [9]$$

The reduced correlation matrix $\mathbf{R}_C$ is the basis for the extraction of common factors.

Extraction of Factors

The model stated in eqns [6] and [7] includes a number of unknown parameters like factor loadings, correlations between factors, and so forth. The first step in EFA is the estimation of the model parameters, also known as the extraction of common factors. The reduced correlation matrix $\mathbf{R}_C$ is the starting point for this estimation process. It should be noted that the communalities are already necessary at this point although they can be determined only when knowing at least the loadings of the model (see eqn [9]). As a result, the reduced correlation matrix $\mathbf{R}_C$ that is the starting point for the estimation process includes preliminary estimates for the communalities. The final communalities are determined after the model parameters are estimated.

The extraction of common factors can be done with a number of different statistical techniques. Only two of the

most prominent approaches can be described briefly here: the principal axes and the maximum likelihood method.

The essence of the principal axes method in the common factor model is the calculation of factors that explain a maximum of the variance of the manifest variables, while minimizing the residuals. Summing up the squared factor loadings of the factor p gives the eigenvalue γ of this factor:

$$\gamma_p = \lambda_{1p}^2 + \lambda_{2p}^2 + \ldots + \lambda_{qp}^2 \quad (p = 1, 2, \ldots, k) \qquad [10]$$

When using the principal axes method, the reduced correlation matrix $\mathbf{R}_C$ is decomposed into its eigenvalues and eigenvectors, which serve as a basis to determine the factor loadings. A detailed description of the mathematical procedure can be found in the pertinent literature (e.g., Mulaik, 2010). Principal component analysis is another popular procedure at this stage of the process. There are a number of differences between principle component analysis and factor analysis as described here. Two of the important differences are that (1) principal component analysis is used with the goal to find a minimum number (rank) of dimensions that allow to reproduce a maximum of the variance in the raw scores of the variables, and, most importantly in the present context, and (2) the model of principal component is fundamentally different from the one presented in eqn [6] (for an elaborate comparison see, e.g., Widaman, 2007).

Another important method for the extraction of common factors is the maximum likelihood method. One of the fundamental differences to other methods of extraction in EFA, including the principal axes method, is that the maximum likelihood method provides statistical estimates of model parameters for a population in which the factor model is assumed to hold. When the parameters are estimated with this method, the likelihood of the observed correlations between the manifest variables, given estimates of the parameters, can be calculated. The procedure is repeated until the factorial structure for which the observed correlations have the maximum likelihood is found.

As Fabrigar et al. (1999) pointed out, both methods have several advantages and disadvantages. In contrast to the principal axes method, the maximum likelihood approach provides a statistical model test and the possibility to calculate several fit indices to assess the appropriateness of the factor model. The model test and indices may also be used for the comparison of models with different numbers of latent variables, for example. Statistical tests are also provided for the model parameters by using the maximum likelihood. One of the disadvantages of this method lies in the requirement to make the assumption of a multivariate normal distribution of the manifest variables. The principal axes method, in contrast, does not require distributional assumptions. Although the statistical tests may be rendered useless by the violation of the distributional assumption with maximum likelihood, the parameter estimates are not necessarily negatively affected in such a case.

Determination of the Number of Factors

One of the fundamental issues with regard to the latent structure is the question of how many factors there are. In the process of determining this number, two possible errors of misspecifications can occur. Firstly, overfactoring is given when the process results in a model with more factors than necessary. Secondly, underfactoring describes a result with a model that includes too few latent variables as common causes for the observed variables. In the latter case, the model is too parsimonious, that is, the preferred number of common factors is too small. With respect to the calculation or estimation of the model parameters, it appears to be advisable to avoid the error of underfactoring more strongly than overfactoring.

Arguably, the three most popular methods for the determination of the number of factors in EFA are:

- The eigenvalues-greater-than-one rule, also called the Kaiser–Guttman rule. This rule demands the extraction of as many factors as there are eigenvalues greater than one after the initial extraction of factors based on the correlation matrix $\mathbf{R}_{YY}$ instead of the reduced correlation matrix $\mathbf{R}_C$. In literature this method was often misapplied by referring to the wrong matrix (Fabrigar et al., 1999).
- A graphical procedure for determining the number of factors is the so-called scree test. Here, the size of the eigenvalues (based on $\mathbf{R}_{YY}$ or $\mathbf{R}_C$) is plotted against the number of eigenvalues in descending order, and adjacent points are connected by a straight line. The graph is inspected to identify the point where, from left to right, the slope of the connecting lines markedly levels off. The number of factors is determined as the number of eigenvalues to the left of the point identified.
- A simulation-based approach to determine the number of factors is parallel analysis. The idea is to repeatedly determine the eigenvalues in a simulation study that arise in a specific situation simply by chance. The situation in the simulation study is specified as identical in number of observed variables and respondents with the situation of interest. The average (or 0.95-quantile) of the simulated eigenvalues is computed for each ordered eigenvalue. The mean (or quantile) is then compared with the empirical eigenvalues. The number of factors to extract is given by the number of empirical eigenvalues that are greater than the simulated counterparts, because they are 'above chance.'

Because none of the methods is fail-safe, it appears to be reasonable to apply multiple criteria. Fabrigar et al. (1999) suggest using the scree test and parallel analysis in conjunction to find the appropriate number of factors. If the common factors are extracted with the maximum likelihood method, it is more natural to use the result of the model test and additional fit indexes to determine the number of latent variables. Of course, depending on where on the continuum from exploratory to confirmatory the specific approach is, theoretical plausibility may also be taken into consideration.

Identification and Rotational Indeterminacy

If a unique set of values can be determined for the model parameters, a model is said to be identified. When the number of common factors is larger than 1 in EFA, this is not the case. This so-called rotational indeterminacy alludes to fact that different sets of loadings in the model may be chosen that cannot be distinguished with respect to their quality of

reproducing the correlation matrix $\mathbf{R}_{YY}$. It is also possible under rotational indeterminacy to choose a transformed set of loadings without changing the factor model equation fundamentally. Suppose that eqn [6] with extracted factors may be written as $\mathbf{Y} = \mathbf{\Lambda}^*\boldsymbol{\eta}^* + \mathbf{\Psi}\mathbf{E}$. This equation may be modified as

$$\begin{aligned} \mathbf{Y} &= \mathbf{\Lambda}^*\mathbf{M}\mathbf{M}^{-1}\boldsymbol{\eta}^* + \mathbf{\Psi}\mathbf{E} = \mathbf{\Lambda}\boldsymbol{\eta} + \mathbf{\Psi}\mathbf{E} \\ \text{where } \mathbf{\Lambda} &= \mathbf{\Lambda}^*\mathbf{M} \\ \text{and } \boldsymbol{\eta} &= \mathbf{M}^{-1}\boldsymbol{\eta}^* \end{aligned} \qquad [11]$$

Hence, there is an infinite number of $k \times k$ transformation matrices $\mathbf{M}$ that may possibly be used for adjustment of the common factors (see also Mulaik, 2010: 275 ff.). Prima facie, rotational indeterminacy may appear to be a fundamental problem of EFA but, in fact, it is actually productively exploited by most researchers when the matrix of loadings is transformed to arrive at a solution that is more interpretable.

In most cases this means to target a solution where (1) every common factor shows high factor loadings on a subset of manifest variables and low factor loadings on the remaining manifest variables, and (2) every manifest variable shows high loadings on a very small number of common factors. Such a structure would roughly be called a simple structure.

A fundamental question that has to be answered when rotating factor loadings is whether the factors are forced to remain uncorrelated or not, that is, whether $\mathbf{R}_{\eta\eta}$ is forced to remain diagonal or not. Orthogonal rotation techniques provide factor models whose factors are forced to be uncorrelated whereas oblique rotation techniques allow correlations between common factors.

There are numerous available rotation techniques, both orthogonal and oblique (for a systematic overview, see Browne, 2001). Fabrigar et al. (1999) pointed out that oblique rotation techniques do not necessarily provide high correlations between common factors. If latent variables are uncorrelated in the population, oblique rotations will give solutions with correlations close to zero. For that reason, the use of oblique rotations seems advisable in EFA.

The last step in EFA after rotation is the interpretation of the common factors. The interpretation is based on the mathematical properties of the factor model and on the meaning of the set of manifest variables that show high loadings on the same factor. For the purpose of interpretation the researcher may refer to the factor pattern matrix and to the factor structure matrix. Both matrices contain information about the relationships between the factors and the manifest variables of the model. The factor loadings $\mathbf{\Lambda}$ are regression weights that apply when predicting the observed variables from the common factors. They can be found in the factor pattern matrix. The factor structure matrix contains the covariances between the manifest variables and the factors. If all variables are standardized, as is customary in EFA, this matrix shows the correlations between indicators and latent variables. If the final factor model evinces a simple structure, this will be reflected especially in the factor pattern matrix. Only if an orthogonal rotation was conducted or an oblique rotation results in orthogonal factors, the factor pattern and the factor structure matrix would be identical.

Blurring the Distinction between EFA and CFA: Recent Developments

A vast number of useful extensions and enhancements of the basic factor model as presented has been published, including nonlinear, longitudinal, and even meta-analytic factor analysis, for example. Two of the more recent developments are highlighted here to illustrate how progress with the method has been fruitful for personality research. Both developments are indicative of the gradually becoming more fuzzy distinction between confirmatory and exploratory factor analytic approaches.

Firstly, stepwise exploratory factor analysis (SEFA; Kano and Harada, 2000) is a method that enhances the use of EFA for scale construction in personality research by providing a means for variable selection that helps to improve the fit of the factor model. SEFA is focused on the model fit, an aspect that is of utmost importance in confirmatory approaches, but it helps to improve fit by the elimination and addition of observed variables in a stepwise fashion. An example of the use of this method in personality research is given by Schulze and Roberts (2006).

Secondly, exploratory structural equation models (ESEM; Asparouhov and Muthén, 2009) is a method that helps to overcome the need for strict loading restrictions in CFA that often lead to factor models with an inadequate fit to personality item data. In contrast to SEFA, the data basis is not altered with ESEM. Instead, the loadings as model parameters in particular are more flexibly determined than is common in CFA by reintroducing rotational techniques into the confirmatory approach. An example of the application of ESEM using the big five model of personality is given by Marsh et al. (2010).

Overall, both approaches exemplify different methods between the endpoints of purely EFA and CFA. It appears that for personality research as it has been done for more than a century, such approaches may lead the way to enhanced instruments and models.

Synopsis

The (exploratory) factor model serves as an example for a special latent variable model which can be integrated in the framework of the generalized linear model for latent variables. For all models that include latent variables, the following steps have to be considered: (1) specifying the mathematical model, (2) estimating the model parameters, (3) (when possible) evaluating the fit of the model, and (4) interpreting the results. For EFA, these steps were briefly described above while taking into account that the EFA framework has special characteristics. That is, based on the choice of estimation method, the fit of the model is not always evaluated (e.g., with principal axes extraction). In addition, due to the issue of rotational indeterminacy, the factor loadings can be transformed to satisfy the need for a better interpretation of the loading pattern. Used with great care and understanding of the statistical theory, EFA can identify the structure of latent variables that explain the correlation of manifest variables. For more information about factor analysis see, for example, Mulaik (2010). For more information about the generalized linear model for latent variables, see Skrondal and Rabe-Hesketh (2004).

See also: Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis; Factor Analysis and Latent Structure Analysis: Overview; Factor Analysis and Latent Structure: IRT and Rasch Models; Five Factor Model of Personality, Assessment of; Intelligence: Assessments of; Intelligence: Central Conceptions and Psychometric Models; Personality Assessment; Psychometrics.

Bibliography

Asparouhov, T., Muthén, B., 2009. Exploratory structural equation modeling. Structural Equation Modeling 16, 397–438.

Bartholomew, D.J., 2007. Three faces of factor analysis. In: Cudeck, R., MacCallum, R.C. (Eds.), Factor Analysis at 100: Historical Developments and Future Directions. Erlbaum, Mahwah, NJ, pp. 9–21.

Bartholomew, D.J., Knott, M., Moustaki, I., 2011. Latent Variable Models and Factor Analysis: A Unified Approach, third ed. Wiley, New York.

Bollen, K.A., 2002. Latent variables in psychology and the social science. Annual Review of Psychology 53, 605–634.

Borsboom, D., Mellenbergh, G.J., van Heerden, J., 2003. The theoretical status of latent variables. Psychological Review 110, 203–219.

Browne, M.W., 2001. An overview of analytic rotation in exploratory factor analysis. Multivariate Behavioral Research 36, 111–150.

Christoffersson, A., 1975. Factor analysis of dichotomized variables. Psychometrika 40, 5–32.

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., Strahan, E.J., 1999. Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods 4, 272–299.

Gaito, J., 1980. Measurement scales and statistics: resurgence of an old misconception. Psychological Bulletin 87, 564–567.

Gill, J., 2000. Generalized Linear Models: A Unified Approach. Sage, Thousand Oaks, CA.

Kamata, A., Bauer, D.J., 2008. A note on the relation between factor analytic and item response theory models. Structural Equation Modeling 15, 136–153.

Kano, Y., Harada, A., 2000. Stepwise variable selection in factor analysis. Psychometrika 65, 7–22.

Marsh, H.W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A.J.S., Trautwein, U., Nagengast, B., 2010. A new look at the big five factor model through exploratory structural equation modeling. Psychological Assessment 22, 471–491.

McDonald, R.P., 1999. Test Theory: A Unified Treatment. Erlbaum, Mahwah, NJ.

Morizot, J., Ainsworth, A.T., Reise, S.P., 2007. Toward modern psychometrics: application of item response theory models in personality research. In: Robins, R.W., Fraley, R.C., Krueger, R.F. (Eds.), Handbook of Research Methods in Personality Psychology. Guilford Press, New York, pp. 407–423.

Mulaik, S.A., 2010. Foundations of Factor Analysis, second ed. Chapman & Hall, London.

Schulze, R., 2005. Modeling structures of intelligence. In: Wilhelm, O., Engle, R.W. (Eds.), Handbook of Understanding and Measuring Intelligence. Sage, Thousand Oaks, CA, pp. 241–263.

Schulze, R., Roberts, R.D., 2006. Assessing the big five: development and validation of the openness conscientiousness extraversion agreeableness neuroticism index condensed (OCEANIC). Zeitschrift für Psychologie 214, 133–149.

Skrondal, A., Rabe-Hesketh, S., 2004. Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Chapman & Hall/CRC, Boca Raton, FL.

Takane, Y., de Leeuw, J., 1987. On the relationship between item response theory and factor analysis of discretized variables. Psychometrika 52, 393–408.

Widaman, K.F., 2007. Common factors versus components: principals and principles, errors and misconceptions. In: Cudeck, R., MacCallum, R.C. (Eds.), Factor Analysis at 100: Historical Developments and Future Directions. Erlbaum, Mahwah, NJ, pp. 177–203.

False Memories

Rebecca M Nichols, Daniel Bogart, and Elizabeth F Loftus, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Our memories are not always the perfect representation of past events. We are certainly prone to forgetting, but somewhat less obviously, our memories are also susceptible to distortion. We are even capable of developing whole, new memories that are entirely false. False memories have consequences for people in their daily lives. Scientists have studied false memories in a laboratory setting in order to understand how they develop. Here, we describe various research methodologies and theoretical explanations for false memory phenomena.

For many of us, the experience of forgetting where we put our keys or where we parked our car is an all too familiar one. What routine occurrences such as these illustrate is that our memory system is not perfect: forgetting occurs quite frequently. Forgetting is a process that is uncontroversial. In contrast, other types of memory errors are less intuitive. Unlike errors of omission, which involve forgetting, errors of *commission* involve the incorporation of incorrect information into our memories. On one end of the continuum, false memories can involve the addition of incorrect details, such as misremembering the friend that was with you when you learned about the attacks of 11 September 2001. On the other end, false memories can be for entire events that never occurred, such as being lost in a shopping mall or attacked by an animal as a child.

What do false memories look like? For one thing, false memories are typically accompanied by the belief that the remembered information is true. False memories sometimes go beyond merely having mistaken beliefs about the past. One of the most common distinctions that memory researchers draw when assessing memory is between 'remembering' and 'knowing.' For example, in some studies, people memorize word lists and later are tested on them. During the memory test, people are probed about whether they merely have a sense that the word had been presented before (i.e., they *know* they have seen it before), or whether they actually have a distinct memory of having seen it before (i.e., they *remember* seeing it). Full-fledged false memories are accompanied with a real sense of recollection of the event in question, not merely the general sense that an event occurred.

How do we know false memories exist? Decades of research and real-world examples serve as proofs. If, at this point, the idea that false memories exist as a common occurrence seems surprising, let us turn to a few quick tests we can take to support the notion. Firstly, many, if not all, of us have had the experience of recalling some shared memory with a friend or family member, only to find out that there were major inconsistencies between the two accounts. In instances such as these, where both accounts cannot be simultaneously correct, at least one party must be experiencing a false memory (assuming both are being honest about what they remember). As a separate, but related, type of situation that demonstrates the existence of false memories, perhaps some of you have had the embarrassing experience of recounting an exciting memory of some personal experience you had only to have the person listening to you exclaim that he or she was the one whom that happened to, not you!

In addition to instances such as these, in which there are two or more conflicting memories for the same event, there are some details of individual memories that we can introspectively observe as not possible to be true. To see this, think back to a memory you have from your early years. Take a moment to visualize the scene, and try to visualize as many details as you can. Here's the test: when picturing the scene, where is your vantage point? In other words, are you seeing the event from the perspective of where your eyes were looking at the time, or do you have a more zoomed out view, looking upon your younger self from the perspective of a third party? Autobiographical memories with this latter quality (called the *observer perspective*, as opposed to *field perspective*) are clear examples that memories do not remain stored in the exact way they are initially experienced.

Our memory system does not operate in the same way as a DVD, loyally and comprehensively storing an objective recording of past events. Rather, details for the past can be distorted in a number of ways. These distortions can happen as a result of many causes, such as: viewing and incorporating photographs of past events into memory; discussing past events with others; vividly imagining events or event details that later get confused with the original memory; suggestive questioning used in interviews and therapeutic techniques; and cognitive, emotional, and motivational biases, to name a few (for more on laboratory experiments that examine these factors, see the Studying Memory Distortion section).

Here we review some instances of false memories that have occurred in the real world. Next we dissect some of the techniques and methodologies used to study false memories in laboratory settings, and what those experiments have demonstrated. Subsequently, we review some theoretical accounts regarding why false memories occur, and what factors contribute to a person being more or less susceptible to their development.

False Memories in the Real World

False memories occur to varying degrees in real world and in everyday life. In the broadest sense of the term, false

memories refer to instances when we get certain aspects of remembered events wrong. This might happen when we misremember how a movie ended, or who attended last year's Thanksgiving dinner. These types of false memories for details can (and do) happen while overall memory for the general event – i.e., seeing a movie, and getting together for a Thanksgiving dinner – remains intact. But even general memories for autobiographical events are, at times, fallible. The richest, most robust forms of false memories involve just such instances of memory error; not only do we sometimes forget that entire events occurred, but we also sometimes remember whole events that never did (for more on this phenomenon, *see* Confabulation).

In the real world, memory errors such as forgetting or mixing up details of past events happen quite frequently. These errors are typically innocuous, and often go unnoticed or uncorrected in the social situations in which they are recounted. From a functional standpoint, some researchers have argued that this makes complete sense, as a primary purpose of using memory to tell stories, for instance, might not be the accuracy of the account, but rather its entertainment value or a transmission of cultural or group values (Newman and Lindsay, 2009; Levine et al., 2009). Regardless of whether or not such a flexible memory system might serve some overall adaptive purpose, real-world examples of memory errors have, at times, devastating repercussions.

False memories for details of criminal events, for example, occur in cases of mistaken eyewitness identification. One contributing factor to false eyewitness memory involves a process called *unconscious transference*. In cases of unconscious transference, an eyewitness inadvertently confuses the memory of the assailant's appearance with that of another person, perhaps a bystander or other innocent person who was seen around the time of the crime. In one particularly ironic demonstration of this effect, a psychological expert in the field of eyewitness testimony was accused of rape by a victim who had seen a TV interview in which the psychologist was featured (Schacter, 1996)! Luckily, the expert had an airtight alibi in this instance and was able to prove his innocence, but such is not always the case. In fact, according to the Innocence Project, eyewitness misidentification is the number one contributing cause in cases of wrongful conviction, implicated in nearly 75% of such cases (Eyewitness Misidentification, 2012). The point of these examples is that false memories in the real world can have devastating effects.

While false memories can and do develop naturally, sometimes their formation is facilitated by other factors such as social manipulation (purposeful or otherwise), or even genuine efforts to treat those who are psychologically distressed. As an example of the former, when police officers provide positive feedback after an eyewitness identifies a suspect from a lineup, the confidence in the witness's selection can become inflated to a level well beyond what it would have been otherwise, and one that is wholly independent of the memory's actual veracity. This is what happened to a woman named Jennifer Thompson, who received just such feedback from law enforcement officers after she misidentified Ronald Cotton as the man who had raped her. The consequences were dire for all parties; Cotton spent 11 years of his life in prison before being exonerated by DNA evidence and Thompson has had to grapple with the fact that she identified an innocent man. For a personal account of the story from Thompson-Cannino et al. (2009), see the book they coauthored entitled *Picking Cotton* and *see* Eyewitness testimony for more information about the psychological implications of witnesses' accounts.

False memories have also been created by the use of questionable therapeutic techniques. Fanning the flames of this controversy, these memories were not for mundane events such as brushing one's teeth or getting a haircut, but rather for highly emotionally charged episodes of child sexual abuse. During the late 1980s and early 1990s the number of people coming forward with such 'recovered' memories of child sexual abuse was staggering. Self-help books such as *The Courage to Heal* asserted bold claims like: "if you think you were abused and your life shows the symptoms, then you were" (p. 22). Operating under similar assumptions to those found in books like this, some therapists employed techniques like hypnosis in order to relax the patients and relieve them of their psychological defenses, thus freeing them to remember long forgotten episodes of abuse.

Or so the story goes. In what has been dubbed the most contentious debate in the history of psychology, the 'memory wars' were a vitriolic argument over whether techniques common to clinical settings were leading to high rates of false memories in therapy patients (Diamond, 1980; Spanos et al., 1999). Some scientists examined these suspicious techniques in laboratory settings. One technique that has been considered problematic is a type of 'guided imagination.' Several studies have shown that imagining that an event took place makes it more likely that people come to believe that the event actually did take place, a phenomenon called 'imagination inflation' (Garry et al., 1996; Thomas and Loftus, 2002). Because hypnosis is often paired with instructions to imagine certain events occurring, and because it leaves the person in the hypnotic state particularly vulnerable to suggestion, it brings with it the real risk of creating false memories.

What resulted from these practices were hundreds of patients who entered therapy and left with newly 'recovered' memories of past abuse. Believing that they had been abused as children, men and women began litigating against their parents, or former neighbors, or teachers or others whom they now believed had abused them. In one case, a man named Paul Ingram was accused first by one of his daughters, Ericka, and later by his other daughter, Julie, and eventually one of his sons, Chad, of having inflicted just such abuse over a decade prior. Ingram and his family were devoutly religious and regularly attended church. Ingram was also a cop. As time progressed, the content of accusations against Ingram got more and more grisly; they included claims from Julie that her father and some of his fellow cops regularly raped the daughters after their weekly poker games. When first asked whether satanic rituals had occurred, Julie replied that they had not. Months later, however, Julie remembered more details: she now recalled being burned with scalding irons and being cut with knives so blood could be obtained for sacrificial rituals (Watters, 1991).

Initially Paul Ingram had no memory of any of these occurrences, but he asserted that he had raised his daughters to

tell the truth, and reasoned that what they were saying must be true. Over time, and with the suggestion of interrogators and his pastor, Ingram also began 'remembering' taking part in these satanic sex rituals. Despite the mounting reports and the confession from Ingram, many of the details of the case still were not adding up. In order to resolve some of the inconsistencies, the prosecution brought in psychologist Richard Ofshe, an expert on cult practices. While interviewing Paul Ingram, Ofshe became skeptical and decided to test Ingram's reliability. In order to do so, Ofshe invented a piece of evidence and asked Ingram if he remembered it: *your son and daughter said that you forced them to have sex with each other while you watched*. Consistent with the pattern observed to this point, Ingram at first had no recollection of this, but after praying on it for a few days, the memory came to him and he confessed that he had, in fact, orchestrated such abuse. Neither son nor daughter corroborated this claim.

As this case progressed, physical evidence – which would be expected given the scope and regularity of the supposed cult rituals taking place – was noticeably absent. It became clear to many that the details were too outlandish, and the discrepancies too large to be believed. Paul Ingram eventually realized that his confessions had not been based on true memories, but were rather the result of suggestion and imagination. This realization came, however, after he pled, and was found to be, guilty of the crimes. Unfortunately, as is often true in cases of false confession (where a person confesses to a crime he did not commit), the confession is so powerful that it contaminates how the other pieces of evidence are evaluated (Kassin et al., 2012). Cases like these, centered around false confessions, ruin lives and waste thousands of dollars in resources.

The cases involving Ronald Cotton/Jennifer Thompson or Paul Ingram illustrate an important point: false memories can have potentially dire consequences in real-world settings. This is true whether they involve errors in memory details, such as the face of the perpetrator in the Cotton/Thompson case, or the remembering of entire events that never occurred, such as with the Paul Ingram case. Ingram shows us that false memories can even lead a person to internalize beliefs of guilt and confess to a crime he never committed. In light of the important consequences that false memories can have, it is crucial to view memory skeptically and explore the roles that suggestion, imagination, and other techniques such as hypnosis play in increasing the likelihood that they develop.

Studying Memory Distortion

Over the course of the last few decades, researchers have been able to demonstrate that that our memories do not always operate as reliably as we would like them to. Memory does not work like a DVD or video recorder; we instead construct and reconstruct our memories, and it is during this process that memories may become susceptible to damage and distortion. For more than 40 years, psychological scientists have created false memories in human subjects using an assortment of clever techniques. These tactics vary tremendously in their methods and styles, and because of this, the memories that develop from each of them take on very distinctive qualities.

Early memory distortion work examined the effects of question wording on subsequent memory reports. For example, Loftus and Palmer (1974) showed subjects a video of an automobile accident. Afterward, some subjects were asked "How fast were the cars going when they hit each other?" while other subjects received the same question but with 'bumped,' 'contacted,' 'collided,' or 'smashed' written in place of 'hit.' Subjects' estimates of speed varied depending on the verb used in the question: subjects who read 'bumped,' for example, estimated a slower speed on average than subjects who read 'hit,' who reported a slower speed on average than the subjects who read 'smashed.' Subjects who read 'smashed' were also most likely to report having seen broken glass in the video, though no broken glass was ever present. The wording of the question seemed to imply something about the content of the video itself, and this, in turn, affected subjects' memory for the witnessed event.

The smashed-hit study shows that a leading question can affect the answer a witness gives and also affect their memory. But other forms of suggestion can also affect witness memory, particularly when the suggestion contains misinformation. Misinformation can impair a witness's memory for the details of an event, a phenomenon, which is now known as the 'misinformation effect' (Ayers and Reder, 1998; Loftus, 2005). In a typical misinformation effect study, subjects witness some type of event, are later given some misleading information about that event, and then are tested on their memory for original event details. In one study, subjects viewed a series of slides that depicted an automobile accident. Afterward, the subjects were given some misinformation about the event. For example, some subjects were asked "Did another car pass the red Datsun while it was stopped at the yield sign?" though the car was actually stopped at a *stop sign* in the slideshow. When subjects' memories were tested for elements from the original slideshow, those who received the misinformation were more likely to report having seen a yield sign than those who did not (Loftus et al., 1978). Research on the misinformation effect has persisted for more than 30 years, and researchers continue to produce the effect in slideshow events (Okado and Stark, 2005) as well as video events (Takarangi et al., 2006) and live events (Morgan et al., 2013). The misinformation effect is particularly useful for understanding factors that can influence eyewitness memory; eyewitnesses may be exposed to incorrect details about a crime or other witnessed event from cowitnesses, law enforcement officers, or other legal professionals (Wright et al., 2009). To the extent that witnesses encounter misinformation from these and other sources, their memories may be in jeopardy, and this can have profound consequences for legal outcomes and for justice.

Another widely studied false memory research paradigm is known as the Deese–Roediger–McDermott (DRM) list-learning task (Roediger and McDermott, 1995; Stadler et al., 1999). In the DRM, subjects attempt to memorize lists of semantically related words (e.g., *note*, *sound*, *piano*, *sing*, *radio*, *band*, etc.). Afterward, subjects' memory is tested for studied words (e.g., *note*), nonstudied, nonrelated control words (e.g., *pocket*), and, importantly, nonpresented but semantically related critical items (e.g., *music*). At test, subjects recognize having studied these critical items at substantially large rates. Critical

items can also be freely recalled (Gallo, 2010). The type of errors produced by the DRM truly shows the associative nature of our memories: we may be at risk for holding a false memory if it seems to fit in *schematically*, or within the overall mental representation, of a true memory.

One criticism of the misinformation and especially of the DRM studies, however, is that they operate within the confines of a highly restrictive laboratory environment. It has been argued that because they do not closely approximate real-world situations, they therefore do little to explain how memories can be distorted for perceptually rich, personally relevant, meaningful autobiographical events. Some have even decried the use of the term *false memory* to describe the type of errors that the DRM tends to produce (DePrince et al., 2004; Freyd and Gleaves, 1996). Researchers have also been able to demonstrate, however, that memories from our own past can be created in experimental settings. In the first of these studies, Loftus and Pickrell (1995) led approximately one-quarter of subjects to believe, over the course of three interviews, that, as children, they had been lost in a shopping mall and rescued despite confirmation from their family members that this had never actually happened. Since then, other types of these rich false memories have been planted, such as spilling punch at a wedding (Hyman et al., 1995), suffering from a vicious animal attack (Porter et al., 2000), and getting sick after eating boiled eggs or dill pickles (Bernstein et al., 2005). A plethora of techniques for creating rich false memories has been developed over the years, which include dream interpretation (Loftus and Mazzoni, 1998), guided visualization (Paddock et al., 2000), hypnosis (Spanos et al., 1999), false feedback (Bernstein et al., 2005; Laney and Loftus, 2008), doctored photographs (Wade et al., 2002), doctored video (Nash and Wade, 2009), and guided imagination (Garry et al., 1996).

Despite the large assortment of rich false memory techniques developed, false autobiographical memory research has not come without criticism. Some have claimed that the memories that result from typical laboratory-based experiments are plausible and therefore relatively easy to accept as true events from one's past (Pezdek and Blandón-Gitlin, 2011). It may be difficult, therefore, to apply the results of these studies to explain high-profile repressed memory cases such as the Paul Ingram case. However, events with somewhat low plausibility (such as receiving an enema in childhood) and even extremely low plausibility (such as a seeing a ghost) have been successfully planted in experimental subjects (Sharman and Scoboria, 2009, 2011), and this suggests that memory distortion processes could also be at work in cases of repressed memories of highly unusual or improbable events.

The scientific psychological literature is replete with techniques developed to induce the formation of false memories. The memories that result vary wildly in terms of their complexity and meaning: a false memory from the DRM does not possess the richness and personal significance of a false memory from one's childhood. When taken together, however, the corpus of scientific false memory manipulations shows just how malleable our memories can be, and they help address questions about how and under what conditions our memories can be distorted.

False Memory Theories

Given the prevalence of false memories in controlled research settings and in the real world, it is useful to understand the theoretical explanations for false memory phenomena as a means for understanding why false memories develop and how we can potentially predict them in a number of circumstances.

Perhaps one of the first theoretical contributions to explain false memory formation arose from the work of Bartlett (1932), who began to distinguish *reproductive* from *reconstructive* memory processes. Reproductive processes operate similarly to playing back recorded video; they produce memories that are accurate representations of the material that was encoded. Reconstructive processes, on the other hand, involve a mixture of accurately recalled material from the original event as well as an active rebuilding of the missing pieces. The resultant memory can be, therefore, a combination of both accurate and inaccurate details. Bartlett believed that reconstructive processes were more likely to take place when recalling events that were particularly lengthy or detailed. Bartlett's musings, which bucked the trend at the time about how memory works, continue to weave their way through contemporary theories about memory generally and false memory in particular.

As scientists began to move away from thinking about memory recall as pushing 'play' on a recorded video, false memory theories became more refined. For example, Underwood (1965) first proposed an associative theory of false memory. His implicit associative response theory states that we all hold semantic networks, or webs of information with nodes (concepts) that are connected to other semantically related nodes. The strength of these connections depends on the relatedness of the items and the frequency with which those items are activated. For example, in a word-list task, subjects may be asked to memorize the word 'sleep.' Implicitly, or without conscious awareness, subjects may simultaneously activate the node that represents 'bed.' Underwood found that when tested, subjects were more likely to falsely remember nonpresented words if they were semantically associated to the target word. More generally, associative strength has important implications for memory reconstruction: to the extent we are missing pieces of a memory but have activated related items, we are at serious risk for reconstructing our memory with schematically consistent errors of commission.

Another predominant explanation for false memory phenomena is that of Fuzzy Trace Theory (Brainerd and Reyna, 2002). This theory states that when we encode events, we encode two separate but parallel components of the event: a broad, general, overall 'gist' meaning of the event and specific, precise, 'verbatim' representations of event components. While both gist and verbatim traces both support true memory recall, false memories can arise when we properly encode the gist trace but encode the verbatim traces less well. Information from the gist is then used to fill in gaps in our memory for specifics about the event.

Activation theories and Fuzzy Trace Theory support false memories that develop spontaneously from mental processes within the rememberer. But what causes false memories to form in the presence of some kind of stimulus, such as a suggestion

about an event? One of the most widely accepted theories to explain these kinds of false memories is source-monitoring theory (Johnson et al., 1993; *see* Source Monitoring). According to this theory, subjects encode both the original event, such as the stop sign in the slideshow event from the early Loftus et al. (1978) misinformation study, and the suggestion, such as the yield sign present in the suggestive questioning about that event. False memories occur when subjects misattribute the source of the misinformation to be the original event and not the actual source of the suggestion. As described, source-monitoring theory is directly applicable to studies demonstrating the misinformation effect. However, this theory can also help explain false memory development when the suggestion is generated internally. In an imagination inflation exercise, for example, subjects are asked to rate the confidence they have that a number of events happened to them during childhood. They are then asked to imagine a selected event, in full perceptual detail, after which they are asked to rate their confidence again for all events. An increase in confidence from preimagination to postimagination may reflect the development of a false memory for that event (Garry et al., 1996). Source-monitoring theory would predict this phenomenon; for subjects who experience an increase in confidence, they might be misattributing the source of the imagined event to be their own childhood and not a recent imagination experience.

Components of multiple theories have also been combined to account for false memories in particular paradigms. For example, Roediger et al. (2001) proposed activation-monitoring theory to explain false memories that are brought about in the DRM task. This theory assumes that unstudied but associated critical items (like *music*) are activated during encoding, and monitoring processes during retrieval affect how critical items are misremembered at test. In other words, subjects may misremember *music* because they remember the slideshow as the source of the word rather than as a product of their own internal activation.

These false memory theories serve an important purpose for false memory research: they help address questions about the mechanisms that can ultimately distort our memories. Advancing knowledge about why false memories occur not only furthers our understanding about real world and lab-based false memories alike, but they can also help scientists predict when and under what conditions memory distortion is likely to develop in the future. This sort of prediction is no easy feat, but as false memory theories become more refined, so too will our understanding of false memory phenomena.

False Memory and the Brain

False memory research conducted within the last few years has spanned a number of useful and interesting topics. For example, technological advances in brain imaging have allowed researchers to examine how different behaviors or behavioral patterns might also be associated with different patterns of brain activation. In one study, researchers were curious as to whether or not there would be unique brain activation patterns for true memories when compared to false memories; if so, neuroimaging tools could be a worthwhile means for distinguishing true and false memories in different contexts. In the study, researchers subjected participants to a typical misinformation paradigm while information about brain activity was captured by a functional magnetic resonance imaging (fMRI) machine. These fMRI scanners are able to acquire patterns of blood flow in specific areas of the brain at different points in time, and where there is increased blood flow, there is increased activity of neurons. In the study, subjects were presented a visual slideshow of events and were later provided with a narrative of the event, presented aurally, that contained misinformation. When the subjects' memories were tested, true memories and false memories produced relatively similar patterns of activation, though true memories of the event (which was presented visually) were accompanied by more activation in the visual cortex while false memories (which were derived from there misinformation presented aurally) were accompanied by more activation in the auditory cortex (Stark et al., 2010). Despite this promising early work, we are a long way away from being able to take an individual memory and reliably determine whether it is true or false based on neural activity.

Final Remarks

As we have seen, memory is quite malleable. While it serves us well, our memories are probably filled with not only truths but also some fictions: small ones and large ones. Sometimes these do not matter very much, as when we tell a friend we had chicken last night instead of fish. But at other times, false memories can be devastating for the lives of innocent people. Scientists are passionate about understanding false memories, not only for the sake of scientific discovery, but also with the hope that this work will lead us to new ways to reduce the false memory problem.

Bibliography

Ayers, M.S., Reder, L.M., 1998. A theoretical review of the misinformation effect: predictions from an activation-based memory model. Psychonomic Bulletin & Review 5, 1–21. http://dx.doi.org/10.3758/BF03209454.

Bartlett, F.C., 1932. Remembering: A Study in Experimental and Social Psychology. Cambridge University Press, Cambridge, England.

Bernstein, D.M., Laney, C., Morris, E.K., Loftus, E.F., 2005. False beliefs about fattening foods can have healthy consequences. Proceedings of the National Academy of Sciences of the United States of America 102, 13724–13731.

Brainerd, C.J., Reyna, V.F., 2002. Fuzzy-trace theory and false memory. Current Directions in Psychological Science 11 (5), 164–169. http://dx.doi.org/10.2307/20182800.

DePrince, A.P., Allard, C.B., Oh, H., Freyd, J.J., 2004. What's in a name for memory errors? Implications and ethical issues arising from the use of the term "false memory" for errors in memory for details. Ethics & Behavior 14 (3), 201–233. http://dx.doi.org/10.1207/s15327019eb1403_1.

Diamond, B., 1980. Inherent problems in the use of pretrial hypnosis on a prospective witness. California Review 68, 313–349.

Eyewitness Misidentification, 2012. The Innocence Project. Available from: http://www.innocenceproject.org/understand/Eyewitness-Misidentification.php (retrieved 01.12.2005).

Freyd, J.J., Gleaves, D.H., 1996. "Remembering" words not presented in lists: relevance to the current recovered/false memory controversy. Journal of Experimental Psychology: Learning, Memory, and Cognition 22 (3), 811–813. http://dx.doi.org/10.1037/0278-7393.22.3.811.

Gallo, D.A., 2010. False memories and fantastic beliefs: 15 years of the DRM illusion. Memory & Cognition 38 (7), 833–848. http://dx.doi.org/10.3758/MC.38.7.833.

Garry, M., Manning, C.G., Loftus, E.F., Sherman, S.J., 1996. Imagination inflation: imagining a childhood event inflates confidence that it occurred. Psychonomic Bulletin and Review 2, 208–214.

Hyman, I.E., Husband, T.H., Billings, F.J., 1995. False memories of childhood experiences. Applied Cognitive Psychology 9, 181–197.

Johnson, M.K., Hashtroudi, S., Lindsay, D.S., 1993. Source monitoring. Psychological Bulletin 114 (1), 3–28. http://dx.doi.org/10.1037/0033-2909.114.1.3.

Kassin, S.M., Bogart, D., Kerner, J., 2012. Confessions that corrupt: evidence from the DNA exoneration case files. Psychological Science 23, 41–45. http://dx.doi.org/10.1177/0956797611422918.

Laney, C., Loftus, E.F., 2008. Emotional content of true and false memories. Memory 16, 500–516. http://dx.doi.org/10.1080/09658210802065939.

Levine, L.J., Lench, H.C., Safer, M.A., 2009. Functions of remembering and misremembering emotion. Applied Cognitive Psychology 23, 1059–1075. http://dx.doi.org/10.1002/acp.1610.

Loftus, E.F., 2005. Planting misinformation in the human mind: a 30-year investigation of the malleability of memory. Learning and Memory 12, 361–366.

Loftus, E.F., Mazzoni, G.A.L., 1998. Using imagination and personalized suggestion to change people. Behavior Therapy 29, 691–706.

Loftus, E.F., Miller, D.G., Burns, H.J., 1978. Semantic integration of verbal information into visual memory. Journal of Experimental Psychology: Human, Learning, and Memory 4, 19–31.

Loftus, E.F., Palmer, J.C., 1974. Reconstruction of automobile destruction: an example of the interaction between language and memory. Journal of Verbal Learning and Verbal Behavior 13 (5), 585–589.

Loftus, E.F., Pickrell, J.E., 1995. The formation on false memories. Psychiatric Annals 25, 720–725.

Morgan III, C.A., Southwick, S., Steffian, G., Hazlett, G.A., Loftus, E.F., 2013. Misinformation can influence memory for recently experienced, highly stressful events. International Journal of Law and Psychiatry 36 (1), 11–17. http://dx.doi.org/10.1016/j.ijlp.2012.11.002.

Nash, R.A., Wade, K.A., 2009. Innocent but proven guilty: eliciting internalized false confessions using doctored-video evidence. Applied Cognitive Psychology 23, 624–637. http://dx.doi.org/10.1002/acp.1500.

Newman, E.J., Lindsay, D.S., 2009. False memories: what the hell are they for? Applied Cognitive Psychology 23, 1105–1121 http://dx.doi.org/10.1002/acp.1613.

Okado, Y., Stark, C.E.L., 2005. Neural activity during encoding predicts false memories created by misinformation. Learning and Memory 12 (1), 3–11. http://dx.doi.org/10.1101/lm.87605.

Paddock, J.R., Terranova, S., Kwok, R., Halpern, D.V., 2000. When knowing becomes remembering: Individual differences in susceptibility to suggestion. The Journal of Genetic Psychology 161 (4), 453–468.

Pezdek, K., Blandón-Gitlin, I., 2011. Imagining implausible events does not lead to false autobiographical memories: commentary on Sharman and Scoboria (2009). Applied Cognitive Psychology 25 (2), 341–343. http://dx.doi.org/10.1002/acp.1704.

Porter, S., Birt, A.R., Yuille, J.C., Lehman, D.R., 2000. Negotiating false memories: interviewer and rememberer characteristics relate to memory distortion. Psychological Science 11 (6), 507–510. http://dx.doi.org/10.1111/1467-9280.00297.

Roediger, H.L., McDermott, K.B., 1995. Creating false memories: remembering words not presented in lists. Journal of Experimental Psychology: Learning, Memory, and Cognition 21 (4), 803–814.

Roediger III, H.L., Balota, D.A., Watson, J.M., 2001. Spreading activation and the arousal of false memories. In: Roediger III, H.L., Nairne, J.S., Neath, I., Surprenant, A.M. (Eds.), The Nature of Remembering: Essays in Honor of Robert G. Crowder. American Psychological Association, Washington, DC, pp. 95–115.

Schacter, D.L., 1996. Searching for Memory: The Brain, the Mind, and the Past. Basic Books, New York, NY.

Sharman, S.J., Scoboria, A., 2009. Imagination equally influences false memories of high and low plausibility events. Applied Cognitive Psychology 23, 813–827.

Sharman, S.J., Scoboria, A., 2011. Event plausibility and imagination inflation: a reply to Pezdek and Blandon-Gitlin. Applied Cognitive Psychology 25 (2), 344–346. http://dx.doi.org/10.1002/acp.1705.

Spanos, N.P., Burgess, C.A., Burgess, M.F., Samuels, C., Blois, W.O., 1999. Creating false memories of infancy with hypnotic and non-hypnotic procedures. Applied Cognitive Psychology 13, 201–218.

Stadler, M.A., Roediger, H.L., McDermott, K.B., 1999. Norms for word lists that create false memories. Memory and Cognition 27, 494–500.

Stark, C.E.L., Okado, Y., Loftus, E.F., 2010. Imaging the reconstruction of true and false memories. Learning and Memory 17, 485–488.

Takarangi, M.K.T., Parker, S., Garry, M., 2006. Modernising the misinformation effect: the development of a new stimulus set. Applied Cognitive Psychology 20 (5), 583–590. http://dx.doi.org/10.1002/acp.1209.

Thomas, A.K., Loftus, E.F., 2002. Creating bizarre false memories through imagination. Memory & Cognition 30, 423–431.

Thompson-Cannino, J., Cotton, R., Torneo, E., 2009. Picking Cotton. St. Martin's Press, New York, NY.

Underwood, B.J., 1965. False recognition produced by implicit verbal responses. Journal of Experimental Psychology 70, 122–129.

Wade, K.A., Garry, M., Read, J., Lindsay, D.S., 2002. A picture is worth a thousand lies: using false photographs to create false childhood memories. Psychonomic Bulletin & Review 9 (3), 597–603. http://dx.doi.org/10.3758/BF03196318.

Watters, E., July–August 1991. The devil in Mr. Ingram. Mother Jones 30–33, 65–68.

Wright, D.B., Memon, A., Skagerberg, E.M., Gabbert, F., 2009. When eyewitnesses talk. Current Directions in Psychological Science 18 (3), 174–178. http://dx.doi.org/10.1111/j.1467-8721.2009.01631.x.

Familial Studies: Genetic Inferences

Vincent P Diego, Jack W Kent, Jr., and John Blangero, Texas Biomedical Research Institute, San Antonio, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C. Capron, volume 8, pp. 5259–5265, © 2001, Elsevier Ltd.

Abstract

Most human characteristics (including behavioral and neurocognitive phenotypes) are influenced to some extent by heritable genetic variation. The concept of heritability, which is a measure of the relative amount of observed phenotypic variance that is attributable to genes, has had a long and controversial history in human genetics. Inference about heritability obtained from familial relationships predated and informed the discovery of the cellular and biochemical mechanisms of inheritance. Our recent ability to obtain complete DNA sequence of individual genomes reveals the full extent of human genetic diversity, and allows for identification of empirical patterns of kinship when family structure is unknown. In this context, the tools of statistical genetics remain essential for quantifying, identifying, and interpreting causal relationships of genetic variation to traits of interest. This article reviews key concepts and controversies in statistical genetics and introduces recent advances in efficient inference from complex pedigrees.

The broad strategy of human genetics is to characterize heritable genomic variation that contributes to the correlation between relatives for traits of interest, as a tool for revealing the biological mechanisms underlying these traits. This strategy is both simple and complex. It is simple because an individual's genome is a sample of each of its parental genomes via cellular processes that are well understood: the set of possible causal variation is large but finite. It is complex because we expect most traits to be the net result of variation in multiple genes whose expression is modulated by environmental exposures, and also because, for ethical and practical reasons, our initial investigations must be observational rather than experimental. Statistical genetics provides the framework for these initial investigations to define focused hypotheses that *are* amenable to experimental pursuit, and it also accounts for the effects due to genetic relatedness (which often may not be the primary focus) when interpreting experimental results. Because the strategy is both simple and complex, the tools of statistical genetics must be correspondingly both parsimonious and extensible.

Introduction to Statistical Genetics and the Heritability Concept

Human statistical genetics begins with the linear variance components (VCs) model developed by Fisher (1918) (Hopper, 1993; Hopper and Visscher, 2005). Since Fisher (1918), a dominant theme in statistical genetics has been the partitioning of the total phenotypic variance initially into uncorrelated genetic and residual environmental VCs, where the genetic variance may be further partitioned into additive, dominance, and epistatic VCs (i.e., reflecting the incremental contribution of alleles, interaction between alleles, and interaction between genes). To the extent that putative environmental causes are measurable, VCs for similarity of environmental exposures and genotype–environment interaction (GEI) may be added to the model as appropriate to clarify the genetic signal. From the genetic VCs, we get the broad-sense and narrow-sense heritability, which are respectively defined as the ratio of the total genetic variance to the total phenotypic variance, denoted by H^2, and the ratio of the additive genetic variance to the total phenotypic variance, denoted by h^2. If the dominance and epistatic VCs are negligible, then the broad-sense and narrow-sense heritabilities converge.

The simplest form of the VCs model makes the following four assumptions common to all linear models: (1) observations are normally, and (2) independently distributed, (3) effects are additive, and (4) the variance is constant or homogeneous across observations. We contend that all the major criticisms of heritability estimation (Capron and Vetta, 2001; Vetta and Capron, 2001; Vetta and Courgeau, 2003/2004) can be addressed by noting that the respective targets of the criticisms are departures from the assumptions of the simple model that can easily be accounted for by existing modeling approaches. Fisher (1918) himself initiated this approach when he extended the basic model to account for assortative mating, which is a source of nonindependence due to correlated mating behavior, and for dominance and epistasis, which are respectively nonadditivity of alleles within loci and nonadditivity among loci. Reviews of VCs models that include such extensions and more can be found in Hopper (1993) and Hopper and Visscher (2005). In the case of dominance and epistasis, however, recent authoritative reviews have concluded that such departures from nonadditivity in most cases can be considered negligible (Hill et al., 2008; Crow, 2010). Similarly, Hill (2010) observed that in genomewide association studies (GWAS) of 10 000 or more individuals, none of the identified variants showed evidence of dominance or epistasis. Because of space limitations, we cannot adequately address how violations in the assumptions of the VCs model may be accounted for. Suffice it to say that we take the position that while the VCs modeling approach can easily account for departures from its basic assumptions, it often is not empirically warranted to do so.

The mathematical formulation of the VCs model for a phenotype vector is given as follows (Blangero et al., 2001, 2013; Almasy and Blangero, 2010):

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times j}\boldsymbol{\beta}_{j\times 1} + \mathbf{g} + \mathbf{e}, \quad [1]$$

where $\mathbf{y}$, the phenotype vector of interest, $\mathbf{X}$, a design matrix of covariate effects, and $\boldsymbol{\beta}$, a vector of regression coefficients, are of dimensions $n \times 1$, $n \times j$, and $j \times 1$, respectively, and n and j give the numbers of individuals in the pedigree and of fixed-effects parameters, respectively; and $\mathbf{g}$, and $\mathbf{e}$ are unobserved vectors of random genetic and environmental effects, respectively. On assuming that the genetic and environmental effects are uncorrelated and that dominance and epistasis are negligible, the polygenic model for the phenotypic covariance matrix is as follows:

$$V[\mathbf{y}] = \boldsymbol{\Sigma} = \sigma_{\mathrm{p}}^2\left[\mathbf{K}h^2 + \mathbf{I}\left(1 - h^2\right)\right], \quad [2]$$

where $\mathbf{K}$ is a genetic relationship kernel or matrix (GRK), the elements of which are relatedness coefficients for individuals i and j, $\mathbf{I}$ is the identity matrix, and $h^2 = \frac{\sigma_{\mathrm{g}}^2}{(\sigma_{\mathrm{g}}^2+\sigma_{\mathrm{e}}^2)} = \frac{\sigma_{\mathrm{g}}^2}{\sigma_{\mathrm{p}}^2}$ is the standard additive genetic heritability, where σ_{g}^2, σ_{e}^2, and σ_{p}^2 are the additive genetic, residual environmental, and total phenotypic variances, respectively. For this basic model, $\mathbf{K} = 2\boldsymbol{\Phi}$, where $\boldsymbol{\Phi}$ is the expected kinship matrix generally derived directly from pedigree information. Assuming that the trait follows a multivariate normal (MVN) distribution, the model ln-likelihood function is given as follows:

$$\ln L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) = -\frac{1}{2}\left[N \ln 2\pi + \ln\left|\sum\right| + \boldsymbol{\delta}'\sum^{-1}\boldsymbol{\delta}\right], \quad [3]$$

where $\boldsymbol{\theta} = [\boldsymbol{\beta}, \sigma_{\mathrm{p}}^2, h^2]'$, which denotes the parameter vector, is expressed as conditional on the data made up of $\mathbf{y}$ and $\mathbf{X}$, and $\boldsymbol{\delta} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ is the vector of residuals. The likelihood function is required for inferential procedures, including estimation of model parameters and model likelihoods via maximum likelihood (ML), and hypothesis testing by way of the likelihood ratio test (LRT) statistic.

Pregenomics Phase of the Heritability Controversy

Relative to other statistical genetic concepts, the notion of heritability has received the lion's share of controversy, which conveniently can be divided into a pregenomics phase and a genomics phase. The pregenomics phase of the heritability controversy took place during a time when, for lack of detailed genomic information, complex-trait genetics was by necessity inferential rather than mechanistic. In human genetics, controversy centered on political and social inferences from genetic data, notably Jensen's (1969) argument that consistently high estimates of heritability for measures of human intelligence implied that academic performance was determined genetically and therefore not amenable to improvement by changes in education practices or social policy. This assertion was recognized widely as employing misinterpretations of heritability that agricultural geneticists already had learned to avoid (Visscher et al., 2008). For a comprehensive treatment of this phase of the heritability controversy, we refer to Sesardic (2005) and other sections of this encyclopedia (see '*See also*' section).

Although the historical context of the pregenomic controversy has passed, one aspect of the reaction to Jensen (1969) continues to inform the current debate. Lewontin (1974, 2001), a key figure for this line of thought, took the tack of rejecting the applicability of heritability estimation to human genetics on theoretical grounds. Lewontin argued (1) that nonadditive effects have a substantial effect on traits but are not reflected in heritability estimates, (2) that heritability estimates are constrained by stochastic allele frequencies and environmental exposures of the sampled individuals and therefore are study specific; in Lewontin's view, this specificity eliminated the explanatory value of the estimates, and (3) the existence of GEI renders heritability estimates under a simple model erroneous. Similar objections to heritability and VCs estimation in general have been raised by others (Capron and Vetta, 2001; Vetta and Capron, 2001; Vetta and Courgeau, 2003/2004; Vitzthum, 2003; Rose, 2006; Rende, 2004). Recent philosophical analysis of this controversy can be found in Rende (2004), Oftedal (2005), Sesardic (2005), Tabery (2009), and Downes (2010).

The first argument of Lewontin (1974) is addressed at length in the following section, as accumulating genomic data has provided evidence to clarify this long-standing theoretical concern. The second argument deserves special attention. Agricultural geneticists have long recognized that heritability estimates are sample-specific, although the range of estimates for specific traits is often remarkably similar across studies (Falconer and Mackay, 1996, pp. 161–163). By the mid-twentieth century, heritability had long been a useful tool for agricultural breeding programs, but its predictive and explanatory value for human studies was limited, as Lewontin observed. What Lewontin (1974) failed to note, however, was that heritability is useful *because* it is sample-specific: It provides an estimate of the genetic signal available in a sample – which, upon replication, identifies traits that are amenable to subsequent genetic analysis. More important, the VCs model explicitly accounts for the *within-sample* genetic and environmental correlations among study subjects (including empirical kinship among so-called 'unrelated' individuals; Yang et al., 2010, 2011a,b). In the present genomics era, when we are able to observe the full extent of human genetic diversity directly, modeling this nonindependence is essential to separate true fixed effects of genetic variants of interest from other sources of phenotypic similarity. As for Lewontin's third criticism, we have consistently found that the incorporation of GEI in the VCs model, rather than rendering the heritability estimate erroneous, in fact and to the contrary clarifies the genetic signal measured by heritability, thereby enhancing gene detection and localization (Diego et al., 2003, 2009; Vinson et al., 2008; Voruganti et al., 2011; Kent et al., 2012).

Genomics Phase of the Heritability Controversy

The genomics phase of the heritability debate followed the availability of reference sequence data from the Human

Genome Project (http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml) and the use of high-throughput genotypic data in human genetic studies, including GWAS (Visscher et al., 2008). Briefly, GWAS analysis tests the fixed effects on a trait of interest of a large set of common genetic markers (typically, hundreds of thousands to a few million markers per genome). These markers are chosen as proxies for expected common *functional* genetic variation in nearby chromosomal regions, based on a controversial theory that common diseases should have common causes (Chakravarty, 1999). Much of the appeal of this approach is an expectation that, if genetic causes are common across population strata, their fixed effects should account for the majority of the relevant genetics of the traits of interest. Genetic analysis therefore could be carried out in 'unrelated' individuals drawn from the general population (Risch and Merkingas, 1996), the random effects of kinship, the logistically challenging family studies needed to provide these data, and the technically complex VCs models required to estimate these effects could all be avoided. Opposition to the concept of heritability shifted from the assertion that it is wrong to the assertion that it is obsolete.

Previous heritability estimates for many traits of interest remain in the literature. Following a decade of GWAS on a plethora of complex traits, it widely is found that the fixed effects of marker variants detected in these analyses usually account for a small fraction of the heritability – often *collectively* accounting for less than 5–10% of the total trait variance (e.g., height; Lango et al., 2010). Given that complex-trait heritabilities routinely are reported to be much higher than the GWA estimates, there is the problem of having to explain the 'missing heritability' (Manolio et al., 2009).

The Impact of Sequence Data

The remarkable technological breakthroughs that allowed the Human Genome Project to generate the reference sequence for the human genome have opened the possibility of affordable sequencing on a substantial number of individuals, either via exome sequencing (covering only the protein-coding regions of known genes; Abecasis et al., 2012) or that ultimate objective, whole-genome sequencing (WGS). The wealth of data now available from these studies allows, for the first time in the history of human genetics, empirical tests of statistical genetics theory, including hypotheses about the origin of 'missing heritability'.

A leading hypothesis for the explanation of the missing heritability is that much of the genetic variance for these traits is due to interactions – especially GEI or gene–gene interaction (epistasis) (Manolio et al., 2009; Zuk et al., 2012). Arguably, Zuk et al. (2012) have offered the most forceful analysis and modeling to advance the hypothesis that the incorporation of interaction terms, especially epistasis, in the model would explain the missing heritability. Zuk et al. (2012) introduced novel limiting-pathway (LP) models of epistasis, and they showed that under these models a significant portion of the missing heritability was explained. Thus, they called the gap between naïve GWA and expected heritability estimates 'phantom heritability' because the difference was never missing after all (under their argument) but rather was simply undetected because of the lack of the epistasis VC. This phantom component is really the result of contaminated estimation of the most focal heritabilities. This primarily is the result (although not explicitly acknowledged by Zuk et al., 2012) of the overreliance on twin studies and nuclear family designs that are most susceptible to the bleeding of any epistatic component into the purely additive heritability. To our knowledge, there have been at least two developments that show that it is premature to abandon the classical most parsimonious VCs model in favor of more complicated versions of the model, such as those proposed by Zuk et al. (2012).

The first development stems from the work by Visscher et al. (2006, 2007). These investigators developed a method for assumption-free heritability estimation. What they really meant is that their method did not assume fixed amounts of genetic similarity as a function of relationship class but empirically estimated the actual observed sharing of genetic material between relatives. As shown in Visscher et al. (2007) and noted in Visscher (2009) and Hill (2010), there was a linear relationship between chromosome length and per-chromosome estimates of heritability, and no evidence of nonadditivity involving loci across chromosomes.

The second development also was initiated by Visscher et al. (Yang et al., 2010, 2011a,b). Visscher, Yang, and colleagues appropriated methods from animal breeding genetics (VanRaden, 2008) to estimate an empirical GRK based on a set of genomewide SNPs in a sample of unrelated individuals. Because the GRK is estimated on unrelated individuals, its elements (i.e., the relatedness coefficients) usually are not confounded by shared familial environmental effects (Yang et al., 2010, 2011a,b; Zaitlen and Kraft, 2012). Zaitlen and Kraft (2012) simulated data using a variety of LP epistasis models of Zuk et al. as the generating models, and then they estimated heritability using the Yang–Visscher GRK. They found that the heritability estimates were not biased by the lack of an epistasis VC in the estimating model, thus refuting the claims made by Zuk et al. (2012) regarding phantom heritability. This is a direct result of relying on more distant relationships in which epistasis has little influence on expected correlations between individuals. It can be shown that this same property is effectively observed in studies of large extended pedigrees when heritability inference is limited to relatives beyond first degree (e.g., when monozygotic twins, parent–offspring, and sibling relationships are not used for heritability estimation).

Several points should be mentioned in regard to heritability estimates in samples of unrelated individuals, especially using the GRK computed under the Yang–Visscher approach. This approach is inefficient relative to pedigree-based methods requiring 5000–10 000 individuals or more to accurately estimate heritability with modest effects (Zaitlen and Kraft, 2012). In other words, this approach has little statistical power. Furthermore, although it is true that such heritability estimates will not be influenced by shared environment, they can be upwardly biased due to hidden between-group variation (i.e., population substructure) (Zaitlen and Kraft, 2012). Conversely, when calculated using standard GWA SNP panels, this approach underestimates the true heritability of the sample population because it fails to capture the genetic

variance resulting from rare sequence variation. Improved estimates can be expected, however, in WGS studies that cover the whole spectrum of common and rare alleles. Further still, we note that the GRK under the Yang–Visscher approach is potentially biased (or at least less interpretable as a measure of heritability) because their estimates allow for negative relatedness coefficients. In this regard, we suggest the GRK estimation algorithm developed by Speed et al. (2012) as it accounts for linkage disequilibrium and estimates proper relatedness coefficients.

To return to the missing heritability problem, a second hypothesis suggests that missing heritability in GWAS is due to that method's focus on common genetic variants. Opinions differ about the exact threshold separating 'rare' and 'common' variation; for the single nucleotide polymorphisms (SNPs) typically employed in association studies, variants with a minor allele frequency less than about 1–5% generally are classified as rare. In direct contrast to the 'common disease, common variant' hypothesis that informed the GWAS strategy, population genetic theory has long predicted that the majority of genomic variation should be rare (**Figure 1**; Pritchard, 2001; Pritchard and Cox, 2002). The argument is as follows: Chance mutation introduces a single copy of a variant in a single individual; that individual has a 50% chance of passing the new variant to each of its offspring; but in the absence of strong natural selection favoring the variant, its spread through the population will be slow. On the other hand, unless it has a large effect on the trait, a deleterious rare allele will not be subject to strong purifying selection that could force its allele frequency to zero. The expected result is the accumulation of mutations at essentially every nucleotide in the genome, although the number of individuals carrying a specific mutation (and thus its minor allele frequency) will remain low.

The availability of WGS provides direct evidence that, consistent with theory, most human sequence variation is rare

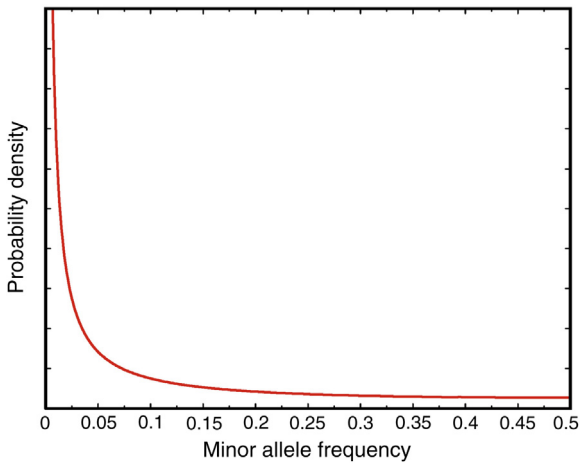

Figure 1 Predicted distribution of minor allele frequencies in a population at equilibrium, assuming a biologically plausible balance between mutation and selection. For a polygenic trait, the effect sizes of most variants will be small and subject to no or weak purifying selection; consequently, most variants are rare (minor allele frequency <0.05).

(Abecasis et al., 2012) – so rare that much sequence variation is effectively private (or lineage-specific). That fraction of the genome that we are most interested in, the phenotypically functional component, is even more likely to be dominated by such rare variation. In humans, rare functional variation is best studied in large pedigrees, an approach that represents an implicit enrichment strategy for identifying and studying rare functional variants. Mendelian transmissions from parents to offspring maximize the chance that multiple copies of rare variants exist in the pedigree (**Figure 2**). Alternatively, studies of unrelated individuals like those typically performed in the now-receding GWAS era that has focused only on common sequence variation never can capture more than one copy of a 'private' variant. Although there are accumulating methods to examine the joint effects of sequence variation in a gene-centric manner that may be of value in the study of unrelated individuals, a large part of human genetics will stay focused on the rapid identification of specific rare variants of moderate to large effect on disease risk because such variants more rapidly lead to functional experimental validation and causal gene discovery with all of its concomitant benefits. Thus, it is apparent that the coming WGS era of human genetics will require a return to our fundamental roots with a refocus on pedigree-based studies of human variation (Blangero, 2004; Frazer et al., 2009; Schork et al., 2009; Ott et al., 2011).

An Eigensimplification Approach for Study Designs Involving Large Families

Utilizing optimal large pedigrees for the examination of complex human phenotypic data represents a formidable statistical problem. We have developed a method that can rapidly analyze SNPs from WGS data sampled from large extended families, which is based on the spectral or eigenvalue decomposition (EVD) of the covariance matrix (Dyer et al., 2009; Blangero et al., 2013). We call this general process the eigensimplification of the likelihood function.

Assuming for convenience and with complete generality that $\sigma_p^2 = 1$, the EVD of the covariance matrix can be written as follows:

$$\mathbf{\Sigma} = \mathbf{S}\mathbf{D}_p\mathbf{S}' = \mathbf{S}\left[h^2\mathbf{D}_g + \left(1-h^2\right)\mathbf{I}\right]\mathbf{S}' = \mathbf{S}\left[\mathbf{I} + h^2\left(\mathbf{D}_g - \mathbf{I}\right)\right]\mathbf{S}', \quad [4]$$

where $\mathbf{S}$ is an orthogonal matrix of eigenvectors, and $\mathbf{D}_p = \mathrm{diag}\{\lambda_{pi}\}$ and $\mathbf{D}_g = \mathrm{diag}\{\lambda_{gi}\}$ are respectively diagonal matrices of phenotypic and additive genetic eigenvalues. Likelihood computations now can be greatly simplified by employing a linear transformation of the vector of residuals to the eigen-basis of the covariance matrix, which we denote by $\boldsymbol{\tau}$:

$$\boldsymbol{\tau} = \mathbf{S}'\delta. \quad [5]$$

Under the transformation, the likelihood of the basic VCs model is given as follows:

$$\ln L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) = -\frac{1}{2}\sum \ln\left[1 + h^2\left(\lambda_{gi} - 1\right)\right] - \frac{1}{2}\sum \frac{\tau_i^2}{1 + h^2\left(\lambda_{gi} - 1\right)}, \quad [6]$$

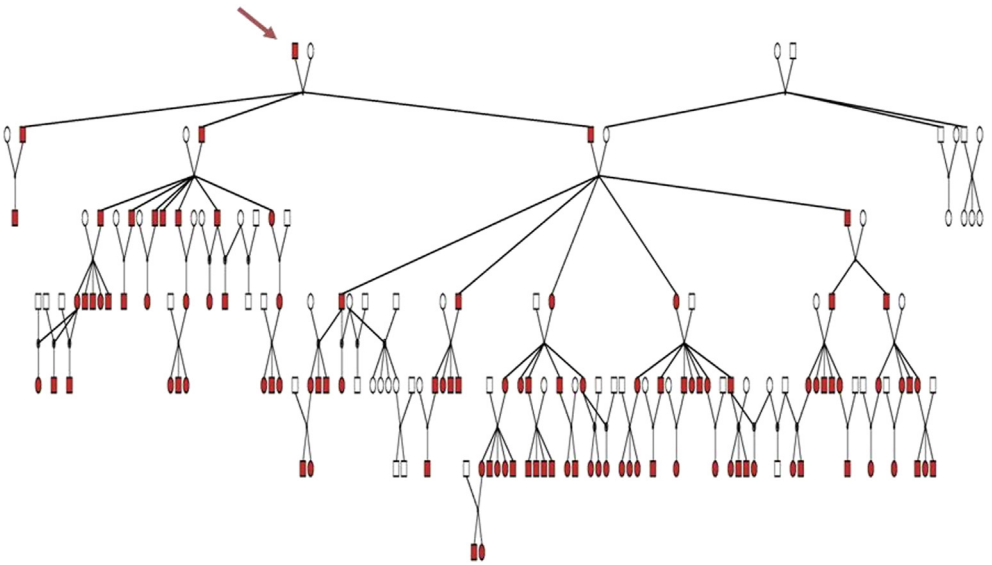

Figure 2 A San Antonio Family Heart Study (SAFHS) extended pedigree ($N = 171$ individuals). Suppose the founder indicated by the red arrow harbors a private variant. It is possible that this one family can have up to 105 copies due to Mendelian transmission.

where λ_{gi} and τ_i are respectively the additive genetic eigenvalue and transformed trait for the ith individual, and the summations are taken over n. One of the virtues of this eigensimplification approach can be understood on comparing eqns [3] and [6]. Under the likelihood function of eqn [3], the requirement of the inverse covariance matrix makes the associated computations nearly intractable for large family studies and for a large number of SNPs. Large family studies translate into a large covariance matrix, which must be inverted some number of times more than the number of SNPs. This is because to find the maximum of the likelihood function, ML estimation algorithms must 'explore' via numerical iteration the complex likelihood surface, and each iteration requires a new inversion of the covariance matrix. Moreover, this computationally intensive iterative process must be performed for each SNP. Under eqn [6], however, large matrix inversion is avoided. Thus, the original complex multivariate likelihood is simplified to a sum of univariate likelihoods, and consequently all associated likelihood computations are greatly simplified. In fact, likelihood analysis of families under the eigensimplification approach is commensurate in terms of speed with that of a sample of unrelated individuals of equal size.

Another virtue of the eigensimplification approach is that it leads to elegant simplifications in the test statistics and in power analysis. This has become an extremely useful tool for understanding the efficiency of different study designs (such as twin designs versus extended pedigree designs).

The expected likelihood ratio test statistic (ELRT) for testing heritability, derived from application of eqn [6], is given as follows:

$$\begin{aligned}\mathrm{ELRT}\left[\widehat{h}^2 > 0\right] &= E\left[-2\left(\ln L(h_0^2 = 0) - \ln L\left(\widehat{h}^2\right)\right)\right] \\ &= -\sum \ln\left[1 + \widehat{h}^2(\lambda_{gi} - 1)\right],\end{aligned} \quad [7]$$

where h_0^2 and $\widehat{h}^2$ denote the heritabilities under the null and alternative hypotheses, respectively. To write the ELRT for association testing, we note the following widely used expression for the variance at a quantitative trait nucleotide (QTN), denoted by σ_q^2:

$$\sigma_q^2 = 2p(1-p)a^2, \quad [8]$$

where p and $(1-p)$ are the allele frequencies at a single genetic locus with two alleles, and a is the displacement between genotypic means. Using the eigensimplification of the MVN likelihood, we find for testing association with a DNA sequence variant at a QTN:

$$\begin{aligned}\mathrm{ELRT}\left[h_q^2 > 0 : \text{pedigrees}\right] &= E\Big[-2\Big(\ln L\left(h_q^2 = 0\right) \\ &\quad - \ln L\left(h_q^2 > 0\right)\Big)\Big] \\ &= -n\ln\left(1 - h_q^2\right) \\ &\quad + \sum \ln\left[1 + h_t^2(\lambda_{gi} - 1)\right] \\ &\quad - \sum \ln\left[1 + h_r^2(\lambda_{gi} - 1)\right] \\ &= \mathrm{ELRT}\left[h_q^2 > 0 : \text{unrelateds}\right] \\ &\quad - \left(\mathrm{ELRT}\left[h_t^2 > 0 : \text{pedigrees}\right]\right. \\ &\quad \left. - \mathrm{ELRT}\left[h_r^2 > 0 : \text{pedigrees}\right]\right),\end{aligned} \quad [9]$$

where $h_t^2 = \frac{\sigma_g^2 + \sigma_q^2}{\sigma_p^2}$, $h_q^2 = \frac{\sigma_q^2}{\sigma_p^2}$, and $h_r^2 = \frac{h_t^2 - h_q^2}{1 - h_q^2}$ are the total, QTN, and residual heritabilities, respectively, and where we express the last form of the statistic in terms of the expected statistics for the QTN-specific heritability in a sample of unrelateds, the total heritability in pedigrees, and the residual heritability in pedigrees, respectively.

The ELRTs in eqns [7] and [9] are ostensibly simpler and easier to compute than other comparable asymptotic test statistics.

Following the methods detailed in Blangero et al. (2013), we can use these ELRTs for heritability and association testing to compute their power functions. We present in **Figure 3** a schematic diagram of several relationship–pedigree structures, including monozygotic (MZ) twins, sibships, and nuclear families. An example of a large extended family is already provided in **Figure 2**. This extended family is made up of 171 individuals across 6 generations (this is an actual family from our San Antonio Family Heart Study (SAFHS) sample that has undergone WGS). In **Figure 4**, we plot at the top panel the ELRTs for three different relationship structures, namely MZ twins, nuclear families with three siblings (NF), and the SAFHS example pedigree (EP), all scaled to 250 individuals for easy comparison. At the bottom panel of **Figure 4**, we show the scaled power functions for heritability estimation for the same relationship structures. Notably, the extended-family design is most powerful in the region of the null. For heritabilities above approximately 0.47, however, the MZ twins design becomes most powerful. Our results indicate that the conventionally and widely held belief (especially among behavioral geneticists) that monozygotic twins constitute the most powerful design for estimating heritability is fundamentally incorrect in the most important part of the parameter space (i.e., in the local area of the null hypothesis). Although we are comparing the two designs, it is noteworthy that the MZ twins design also potentially suffers from confounding of genetic and shared environmental or common household effects, whereas such confounding is much less likely in extended-family designs in which the subordinate nuclear family units typically reside in different households. As mentioned previously, twin designs also suffer from being much more susceptible to biased narrow-sense heritability estimates due to contamination from epistatic components.

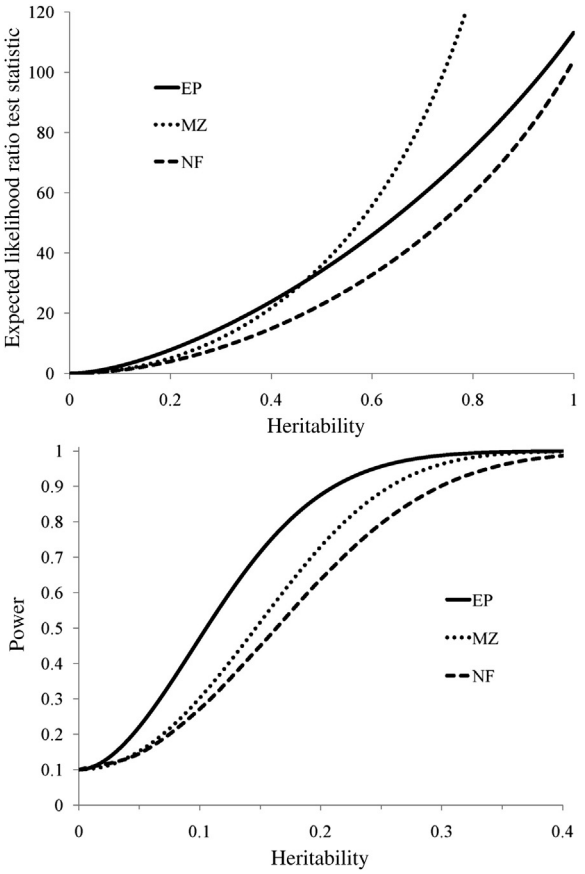

Figure 4 Expected likelihood ratio test statistic and power for heritability estimation. For both panels: SAFHS extended pedigree (EP) (solid line), monozygotic twins (MZ) (dotted line), and nuclear family (NF) (dashed line) all scaled to 250 individuals.

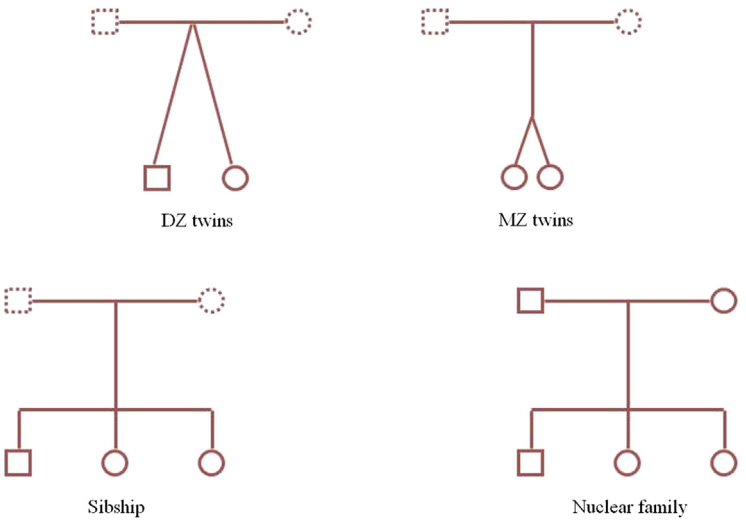

Figure 3 Some simple relationship and pedigree structures.

We examined similar study design comparisons in relation to association testing using eqn [9]. In association testing, we are estimating the heritability attributable to a given genetic marker or sequence variant. For a total heritability of 0.1, we plotted the association ELRTs for the four relationship structures, and power to detect association for the same fixed sample sizes in Figure 5. Figure 5 shows that for this fixed low total heritability and a reasonable range of QTN-specific heritabilities, power to detect associations is greatest in unrelated samples as expected. Loss of power is greatest in the extended pedigree due to the substantial correlation between subjects, however, even for this design the loss of power is low. Figure 6 shows the effect of total heritability on power to detect association for the extended pedigree. Power to detect association is influenced by total heritability with the power loss being maximized at a residual heritability of 0.50. Interestingly, power loss as seen in the ELRT is minimal both near the null region for heritability (as expected) and is somewhat counterintuitively near the maximum of heritability (at 1). Regardless, our theoretical results show little loss of power in the association analysis of even large and complex pedigrees. Furthermore, when considering the increased focus on the analysis of rare variants, power actually is substantially increased in large pedigrees due to the accumulation of multiple copies (and hence, increased genotypic variance) of rare variants incurred through Mendelian transmission in variant-harboring lineages.

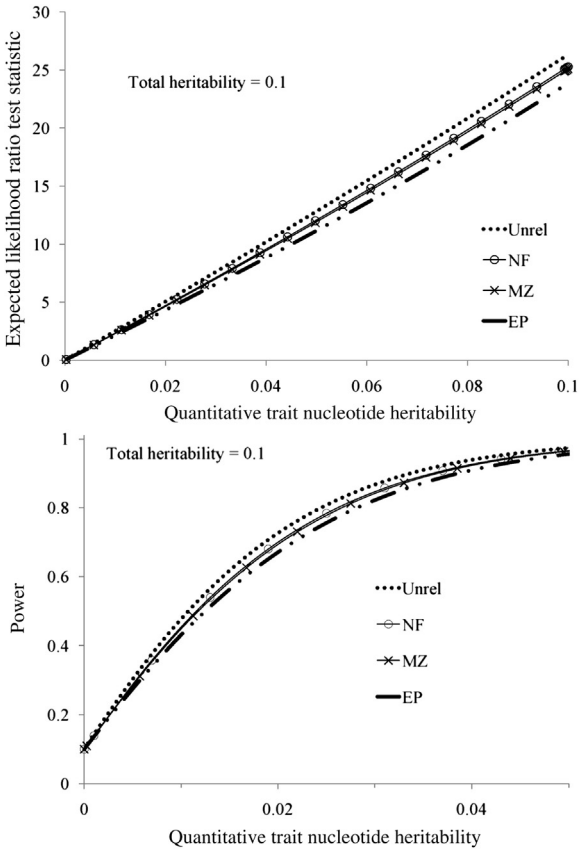

Figure 5 Expected likelihood ratio test statistic and power for association testing. For both panels: Unrelateds (dotted line), nuclear family (NF) (thin solid line with open circles), monozygotic twins (MZ), (thin solid line with cross marks), and SAFHS extended pedigree (EP) (thick solid line) all scaled to 250 individuals.

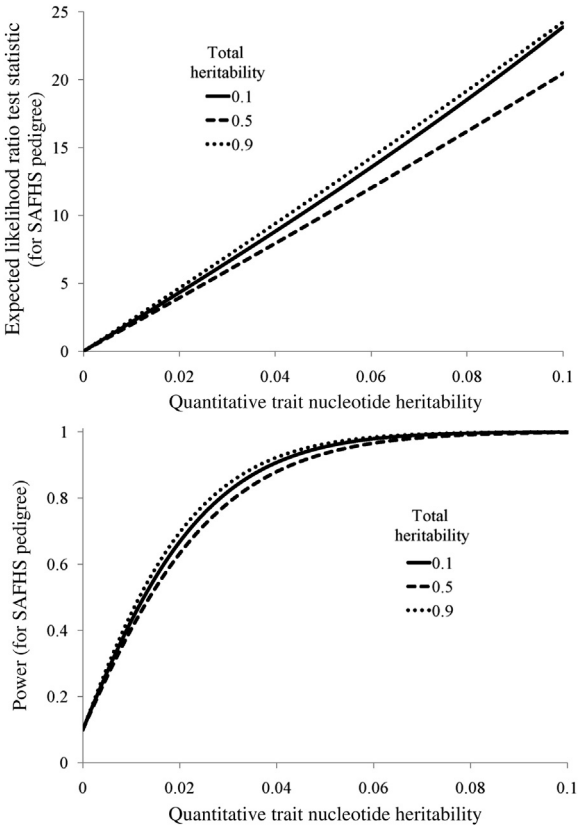

Figure 6 Effect of varying the total heritability on the expected likelihood ratio test statistic and power for association testing all scaled to 250 individuals. Total heritabilities are equal to 0.1 (solid line), 0.5 (dashed line), and 0.9 (dotted line).

Some Examples of the New Genetics and a Practical Extension

In Table 1, we highlighted results from several studies that used the Yang–Visscher approach to estimate heritability under eqn [2] using an empirical GRK computed on GWA SNP data from a sample of unrelated individuals. For crude comparisons with existing heritability estimates, we also included a column for heritability estimates from twin studies. The traits shown span a wide range of interests relevant to social and behavioral scientists. Importantly, together they show that heritable genetic effects can have an unequivocal impact on overall trait variation.

The issue of common and rare variation emerges in each of these studies. Since these were, from a genetics perspective, GWA studies, they used common SNP variants. As can be gleaned from a closer examination of these results, only some

Table 1 Heritability estimates from twin studies and GWA SNP markers

	Heritability estimate; references	
Trait	*Twin design*	*GWA SNPs*[a]
Crystallized-type intelligence	0.74; Cattell et al. (1981)	0.40; Davies et al. (2011)
Fluid-type intelligence	0.56; Cattell et al. (1981)	0.51; Davies et al. (2011)
Neuroticism	0.41–0.58[b]; Bouchard and McGue (2003)	0.06; Vinkhuyzen et al. (2012)
Extraversion	0.49–0.53[b]; Bouchard and McGue (2003)	0.12; Vinkhuyzen et al. (2012)
Education attainment	0.42; Branigan et al. (2013)	0.16; Benjamin et al. (2012a)
Trust	0.33 in males; 0.39 in females; Oskarsson et al. (2012)	0.24; Benjamin et al. (2012a)
Economic policy	0.20–0.60[c]; Hatemi et al. (2011)	0.34; Benjamin et al. (2012a)
Foreign policy	0.34–0.56[c]; Cranmer and Dawes (2012)	0.35; Benjamin et al. (2012a)
Schizophrenia	0.83; Sullivan et al. (2003)	0.23; Lee et al. (2012)
Major depressive disorder	0.33–0.54[b]; Shih et al., 2004	0.32; Lubke et al. (2012)
Bipolar disorder	0.71–0.77[c]; Edvardsen et al. (2008)	0.40; Lee et al. (2011)

[a]Computed using the Yang–Visscher approach (see text).
[b]Range across different studies.
[c]Range across related traits.

of the missing heritability was explained. For reasons given earlier in this entry, we are confident that WGS studies on families will be able to more fully recover the missing heritability.

One widely held position regarding the potential contribution that the new genetics can make to the social and behavioral sciences is that genetic data, especially of the GWA or WGS variety, may be used fruitfully as control variables (Beauchamp et al., 2011; Benjamin et al., 2012a,b; Fowler and Dawes, 2013). Just following, we outline how genetic data may be used as control variables in the context of the VCs model. Perhaps somewhat counterintuitively, heritability can be used to clarify inferences on the effect of nongenetic variables of interest to social and behavioral scientists. Let γ denote the ratio of nongenetic to genetic effects; that is, $\gamma = (1 - h^2)/h^2$. Then, the vector of estimated genetic values (EGVs), denoted by $\widehat{\mathbf{g}}$, is given as follows (Quillen et al., in press):

$$\widehat{\mathbf{g}} = \left(\mathbf{I} + \mathbf{K}^{-1}\gamma\right)^{-1}\boldsymbol{\delta}. \qquad [10]$$

Since $\widehat{\mathbf{g}}$ is the estimated version of $\mathbf{g}$ in eqn [1], it can be thought of as the genetic contribution to the phenotype. Furthermore, with this estimate we can sweep out genetic effects to obtain a derived trait that highlights environmental or, more precisely, nongenetic effects. Specifically, we can do this simply by rearranging eqn [1] as follows:

$$\widehat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \widehat{\mathbf{g}} = \boldsymbol{\delta} - \widehat{\mathbf{g}}. \qquad [11]$$

This equation should prove useful in investigations of social and behavioral traits known to be partially under genetic influence as in those in **Table 1**.

Conclusion

As we have shown, the concept of heritability has gone, in the genomics phase of its controversial history, from being perceived as 'irrelevant' to now being recognized as a matter of central concern. This is true for at least two reasons related in different ways to the missing heritability problem. For one, we showed that the missing heritability problem stirred up interest in the consideration of more complex VCs models that include interactions (e.g., dominance or epistasis). As we noted, however, the accumulating evidence on this matter seems to vindicate use of the simple polygenic model limited to additive genetic effects. Related to the missing heritability problem, we showed how there is an empirically motivated shift from study designs optimized for the detection of common variants (i.e., GWAS in samples of unrelated individuals) to those for the detection of rare variants (i.e., WGS in extended families). With the return to family studies for the optimal detection of rare variants, the eigensimplification approach (Blangero et al., 2013) reviewed herein will prove useful in the study design stage as well as in subsequent likelihood analysis. Moreover, as we demonstrated in our description of this approach, heritability is a fundamental parameter in the likelihood function and associated ELRTs, and is a canonical determinant of power to detect genetic effects.

Under the Yang–Visscher approach discussed earlier that focuses on the use of unrelated individuals, all the SNPs from a GWAS are used to estimate an empirical GRK, which in turn allows heritability estimation. As we noted, however, Zaitlen and Kraft (2012) showed that this approach has little power to detect modest effect sizes, requiring 5000–10 000 individuals or more. Historically, this method was developed because GWAS advocates and proponents of the importance of common genetic variants wanted to explain the missing heritability problem (cf. the concerns in the introduction sections in Yang et al., 2010, 2011a,b). This method and associated study design, however, does not represent the optimal strategy to replace the twin paradigm that has dominated human behavioral genetics. Given the statistical inefficiency for both heritability estimation and the detection of rare sequence variant effects that is so important for causal gene identification, we contend that continuing the expansion of heritability studies of human phenotypes under this approach of focusing on unrelated individuals would not be a wise use of resources.

See also: Genetic Studies of Behavior: Methodology; Genetics: The New Genetics; Intelligence, Genetics of: Cognitive Abilities; Linear Mixed Models; Personality Differences and Development: Genetic and Environmental Contributions; Scientific Controversies.

Bibliography

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A., 1000 Genomes Project Consortium, 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65.

Almasy, L., Blangero, J., 2010. Variance component methods for analysis of complex phenotypes. Cold Spring Harbor Protocols 2010 (5). http://dx.doi.org/10.1101/pdb.top77.

Beauchamp, J.P., Cesarini, D., Johannesson, M., van der Loos, M.J.H.M., Koellinger, P.D., Groenen, P.J.F., Fowler, J.H., Rosenquist, J.N., Thurik, A.R., Christakis, N.A., 2011. Molecular genetics and economics. Journal of Economic Perspectives 25, 57–82.

Benjamin, D.J., Cesarini, D., van der Loos, M.J., Dawes, C.T., Koellinger, P.D., Magnusson, P.K., Chabris, C.F., Conley, D., Laibson, D., Johannesson, M., Visscher, P.M., 2012a. The genetic architecture of economic and political preferences. Proceedings of the National Academy of Sciences of the United States of America 109, 8026–8031.

Benjamin, D.J., Cesarini, D., Chabris, C.F., Glaeser, E.L., Laibson, D.I., Guðnason, V., Harris, T.B., Launer, L.J., Purcell, S., Smith, A.V., Johannesson, M., Magnusson, P.K., Beauchamp, J.P., Christakis, N.A., Atwood, C.S., Hebert, B., Freese, J., Hauser, R.M., Hauser, T.S., Grankvist, A., Hultman, C.M., Lichtenstein, P., 2012b. The promises and pitfalls of genoeconomics. Annual Review of Economics 4, 627–662.

Blangero, J., 2004. Localization and identification of human quantitative trait loci: king harvest has surely come. Current Opinion in Genetics & Development 14, 233–240.

Blangero, J., Williams, J.T., Almasy, L., 2001. Variance component methods for detecting complex trait loci. In: Rao, D.C., Province, M.A. (Eds.), Advances in Genetics, vol. 42. Academic Press, pp. 151–181.

Blangero, J., Diego, V.P., Dyer, T.D., Almeida, M., Peralta, J., Kent Jr., J.W., Williams, J.T., Almasy, L., Göring, H.H.H., 2013. A kernel of truth: statistical advances in polygenic variance component models for complex human pedigrees. In: Friedmann, T., Dunlap, J.C., Goodwin, S.F. (Eds.), Advances in Genetics, vol. 81. Elsevier, Inc., pp. 1–31.

Bouchard Jr., T.J., McGue, M., 2003. Genetic and environmental influences on human psychological differences. Journal of Neurobiology 54, 4–45.

Branigan, A.R., McCallum, K.J., Freese, J., 2013. Variation in the Heritability of Educational Attainment: An International Meta-analysis. Institute for Policy Research, Northwestern University. Working Paper Series, WP-13-09.

Capron, C., Vetta, A., 2001. Comments on "Why have children in the 21st century?". European Journal of Population 17, 23–30.

Cattell, R.B., Schuberger, J.M., Ahern, F.M., Kameoka, V., 1981. The heritability of fluid and crystallized intelligences: by the MAVA design and OSES analysis. Australian Journal of Psychology 33, 355–374.

Chakravarti, A., 1999. Population genetics – making sense out of sequence. Nature Genetics 21, 56–60.

Cranmer, S.J., Dawes, C.T., 2012. The heritability of foreign policy preferences. Twin Research and Human Genetics 15, 52–59.

Crow, J.F., 2010. On epistasis: why it is unimportant in polygenic directional selection. Philosophical Transactions of the Royal Society, B: Biological Sciences 365, 1241–1244.

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S.E., Liewald, D., Ke, X., Le Hellard, S., Christoforou, A., Luciano, M., McGhee, K., Lopez, L., Gow, A.J., Corley, J., Redmond, P., Fox, H.C., Haggarty, P., Whalley, L.J., McNeill, G., Goddard, M.E., Espeseth, T., Lundervold, A.J., Reinvang, I., Pickles, A., Steen, V.M., Ollier, W., Porteous, D.J., Horan, M., Starr, J.M., Pendleton, N., Visscher, P.M., Deary, I.J., 2011. Genome-wide association studies establish that human intelligence is highly heritable and polygenic. Molecular Psychiatry 16, 996–1005.

Diego, V.P., Almasy, L., Dyer, T.D., Soler, J.M., Blangero, J., 2003. Strategy and model building in the fourth dimension: a null model for genotype × age interaction as a Gaussian stationary stochastic process. BMC Genetics 4 (Suppl. 1), S34.

Diego, V.P., Rainwater, D.L., Wang, X.L., Cole, S.A., Curran, J.E., Johnson, M.P., Jowett, J.B., Dyer, T.D., Williams, J.T., Moses, E.K., Comuzzie, A.G., Maccluer, J.W., Mahaney, M.C., Blangero, J., 2009. Genotype x adiposity interaction linkage analyses reveal a locus on chromosome 1 for lipoprotein-associated phospholipase A2, a marker of inflammation and oxidative stress. American Journal of Human Genetics 80, 168–177.

Downes, S.M., 2010. Heritability. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy, Fall 2010 ed. http://plato.stanford.edu/archives/fall2010/entries/heredity

Dyer, T.D., Diego, V.P., Kent Jr., J.W., Göring, H.H., Blangero, J., 2009. Rapid exact likelihood-based quantitative trait association analysis in large pedigrees. In: Poster Presented at the American Society of Human Genetics Annual Meeting, Honolulu, HI, 20–24 October 2009.

Edvardsen, J., Torgersen, S., Røysamb, E., Lygren, S., Skre, I., Onstad, S., Oien, P.A., 2008. Heritability of bipolar spectrum disorders. Unity or heterogeneity? Journal of Affective Disorders 106, 229–240.

Falconer, D.S., Mackay, T.F.C., 1996. Introduction to Quantitative Genetics, fourth ed. Pearson Education, Ltd., Essex, England.

Fisher, R.A., 1918. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh 52, 399–433.

Fowler, J.H., Dawes, C.T., 2013. In defense of genopolitics. American Political Science Review 107, 1–13.

Frazer, K.A., Murray, S.S., Schork, N.J., Topol, E.J., 2009. Human genetic variation and its contribution to complex traits. Nature Reviews Genetics 10, 241–251.

Hatemi, P.K., Dawes, C.T., Frost-Keller, A., Settle, J.E., Verhulst, B., 2011. Integrating social science and genetics: news from the political front. Biodemography and Social Biology 57, 67–87.

Hill, W.G., 2010. Understanding and using quantitative genetic variation. Philosophical Transactions of the Royal Society, B: Biological Sciences 365, 73–85.

Hill, W.G., Goddard, M.E., Visscher, P.M., 2008. Data and theory point to mainly additive genetic variance for complex traits. PLoS Genetics 4 (2), e1000008.

Hopper, J.L., 1993. Variance components for statistical genetics: applications in medical research to characteristics related to human diseases and health. Statistical Methods in Medical Research 2, 199–223.

Hopper, J.L., Visscher, P.M., 2005. Variance component analysis. In: Armitage, P., Colton, T. (Eds.), Encyclopedia of Biostatistics, vol. 8. Wiley, Inc., pp. 5676–5685.

Jensen, A., 1969. How much can we boost IQ and scholastic achievement? Harvard Educational Review 39, 1–123.

Kent Jr., J.W., Göring, H.H., Charlesworth, J.C., Drigalenko, E., Diego, V.P., Curran, J.E., Johnson, M.P., Dyer, T.D., Cole, S.A., Jowett, J.B., Mahaney, M.C., Comuzzie, A.G., Almasy, L., Moses, E.K., Blangero, J., Williams-Blangero, S., 2012. Genotype x age interaction in human transcriptional ageing. Mechanisms of Ageing and Development 133, 581–590.

Lango, A.H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., Ferreira, T., Wood, A.R., Weyant, R.J., Segrè, A.V., Speliotes, E.K., Wheeler, E., Soranzo, N., Park, J.H., Yang, J., Gudbjartsson, D., et al., 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467, 832–838.

Lee, S.H., Wray, N.R., Goddard, M.E., Visscher, P.M., 2011. Estimating missing heritability for disease from genome-wide association studies. American Journal of Human Genetics 88, 294–305.

Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Schizophrenia Psychiatric Genome-Wide Association Study Consortium (PGC-SCZ), International Schizophrenia Consortium (ISC), Molecular Genetics of Schizophrenia Collaboration (MGS), Sullivan, P.F., Goddard, M.E., Keller, M.C., Visscher, P.M., Wray, N.R., 2012. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nature Genetics 44, 247–250.

Lewontin, R.C., 1974. The analysis of variance and the analysis of causes. American Journal of Human Genetics 26, 400–411.

Lewontin, R.C., 2001. The Triple Helix: Gene, Organisms and Environment. Harvard University Press, Cambridge, MA.

Lubke, G.H., Hottenga, J.J., Walters, R., Laurin, C., de Geus, E.J., Willemsen, G., Smit, J.H., Middeldorp, C.M., Penninx, B.W., Vink, J.M., Boomsma, D.I., 2012. Estimating the genetic variance of major depressive disorder due to all single nucleotide polymorphisms. Biological Psychiatry 72, 707–709.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F.C., McCarroll, S.A., Visscher, P.M., 2009. Finding the missing heritability of complex diseases. Nature 46, 747–753.

Oftedal, G., 2005. Heritability and genetic causation. Philosophy of Science 72, 699–709.

Oskarsson, S., Dawes, C., Johannesson, M., Magnusson, P.K., 2012. The genetic origins of the relationship between psychological traits and social trust. Twin Research and Human Genetics 15, 21–33.

Ott, J., Kamatani, Y., Lathrop, M., 2011. Family-based designs for genome-wide association studies. Nature Reviews Genetics 12, 465–474.

Pritchard, J.K., 2001. Are rare variants responsible for susceptibility to common diseases? American Journal of Human Genetics 69, 124–137.

Pritchard, J.K., Cox, N.J., 2002. The allelic architecture of human disease genes: common disease – common variant… or not? Human Molecular Genetics 11, 2417–2423.

Quillen, E.E., Voruganti, V.S., Chittoor, G., Rubicz, R., Peralta1, J.M., Almeida, M.A.A., Kent Jr., J.W., Diego, V.P., Dyer, T.D., Comuzzie, A.G., Göring, H.H.H., Duggirala, R., Almasy, L., Blangero, J. Evaluation of estimated genetic values and their application to genome-wide investigation of systolic blood pressure. BMC Proceedings, in press.

Rende, R., 2004. Beyond heritability: biological process in social context. In: Coll, C.G., Bearer, E.L., Lerner, R.M. (Eds.), Nature and Nurture: The Complex Interplay of Genetic and Environmental Influences on Human Behavior and Development. Lawrence Elbaum Associates, Inc., pp. 107–126.

Risch, N., Merikangas, K., 1996. The future of genetic studies of human complex disease. Science 273, 1516–1517.

Rose, S.P.R., 2006. Commentary: heritability estimates—long past their sell-by date. International Journal of Epidemiology 35, 525–527.

Schork, N.J., Murray, S.S., Frazer, K.A., Topol, E.J., 2009. Common vs. rare allele hypotheses for complex diseases. Current Opinion in Genetics & Development 19, 212–219.

Sesardic, N., 2005. Making Sense of Heritability. Cambridge University Press, Cambridge, UK.

Shih, R.A., Belmonte, P.L., Zandi, P.P., 2004. A review of the evidence from family, twin and adoption studies for a genetic contribution to adult psychiatric disorders. International Review of Psychiatry 16, 260–283.

Speed, D., Hemani, G., Johnson, M.R., Balding, D.J., 2012. Improved heritability estimation from genome-wide SNPs. American Journal of Human Genetics 91, 1011–1021.

Sullivan, P.F., Kendler, K.S., Neale, M.C., 2003. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Archives of General Psychiatry 60, 1187–1192.

Tabery, J., 2009. Making sense of the nature–nurture debate. Biology & Philosophy 24, 711–723.

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91, 4414–4423.

Vetta, A., Capron, C., 2001. Familial studies: genetic inferences. In: Smeiser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Elsevier, Inc.

Vetta, A., Corgeau, D., 2003/2004. Demographic behaviour and behaviour genetics. Population (English Edition) 58, 401–428.

Vinkhuyzen, A.A., Pedersen, N.L., Yang, J., Lee, S.H., Magnusson, P.K., Iacono, W.G., McGue, M., Madden, P.A., Heath, A.C., Luciano, M., Payton, A., Horan, M., Ollier, W., Pendleton, N., Deary, I.J., Montgomery, G.W., Martin, N.G., Visscher, P.M., Wray, N.R., 2012. Common SNPs explain some of the variation in the personality dimensions of neuroticism and extraversion. Translational Psychiatry 2, e102. Erratum: 2012;2:e125.

Vinson, A., Mahaney, M.C., Diego, V.P., Cox, L.A., Rogers, J., VandeBerg, J.L., Rainwater, D.L., 2008. Genotype-by-diet effects on co-variation in Lp-PLA2 activity and LDL-cholesterol concentration in baboons fed an atherogenic diet. Journal of Lipid Research 49, 1295–1302.

Visscher, P.M., 2009. Whole genome approaches to quantitative genetics. Genetica 136, 351–358.

Visscher, P.M., Medland, S.E., Ferreira, M.A., Morley, K.I., Zhu, G., Cornes, B.K., Montgomery, G.W., Martin, N.G., 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genetics 2 (3), e41.

Visscher, P.M., Macgregor, S., Benyamin, B., Zhu, G., Gordon, S., Medland, S., Hill, W.G., Hottenga, J.J., Willemsen, G., Boomsma, D.I., Liu, Y.Z., Deng, H.W., Montgomery, G.W., Martin, N.G., 2007. Genome partitioning of genetic variation for height from 11,214 sibling pairs. American Journal of Human Genetics 81, 1104–1110.

Visscher, P.M., Hill, W.G., Wray, N.R., 2008. Heritability in the genomics era–concepts and misconceptions. Nature Reviews Genetics 9, 255–266.

Vitzthum, V.J., 2003. A number no greater than the sum of its parts: the use and abuse of heritability. Human Biology 75 (4), 539–558.

Voruganti, V.S., Diego, V.P., Haack, K., Cole, S.A., Blangero, J., Göring, H.H., Laston, S., Wenger, C.R., Ebbesson, S.O., Fabsitz, R.R., Devereux, R.B., Howard, B.V., Umans, J.G., MacCluer, J.W., Comuzzie, A.G., 2011. A QTL for genotype by sex interaction for anthropometric measurements in Alaskan Eskimos (GOCADAN Study) on chromosome 19q12–13. Obesity 19, 1840–1846.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Goddard, M.E., Visscher, P.M., 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42, 565–569.

Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., Hill, W.G., Landi, M.T., Alonso, A., Lettre, G., Lin, P., Ling, H., Lowe, W., Mathias, R.A., Melbye, M., Pugh, E., Cornelis, M.C., Weir, B.S., Goddard, M.E., Visscher, P.M., 2011a. Genome partitioning of genetic variation for complex traits using common SNPs. Nature Genetics 43, 519–525.

Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M., 2011b. GCTA: a tool for genome-wide complex trait analysis. American Journal of Human Genetics 88, 76–82.

Zaitlen, N., Kraft, P., 2012. Heritability in the genome-wide association era. Human Genetics 131, 1655–1664.

Zuk, O., Hechter, E., Sunyaev, S.R., Lander, E.S., 2012. 'The mystery of missing heritability: genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences of the United States of America 109, 1193–1198.

Families and Households, Formal Demography of

Frans Willekens, Max Planck Institute for Demographic Research, Rostock, Germany; and Netherlands Interdisciplinary Demographic Institute (NIDI), The Hague, The Netherlands
Evert van Imhoff[†]

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E. Van Imhoff, volume 8, pp. 5271–5275, © 2001, Elsevier Ltd.

Abstract

'Family and household demography' differs from traditional demography in that it explicitly recognizes and studies *relationships between individuals*. Formal demography focuses on the definition and measurement of families and households, and modeling of types, number, and composition of families and households in the past, current, and future. The entry briefly covers issues in definition, measurement, and modeling. The evolution of modeling from headship rate models to agent-based models is discussed.

'Family and household demography' differs from traditional demography in that it explicitly recognizes and studies *relationships between individuals*. It studies how the core processes of birth, death, marriage, divorce, and migration occur to multiple persons interacting with each other. In doing so, additional relevant processes like entering or leaving coresidence are also studied. In short, relationships between individuals are the central focus of family and household demography. This article outlines the main subfields within the broad area of family and household demography: demography of kinship, households, families, and marriage, respectively. It briefly discusses data issues and definitional problems with a direct bearing on family and household demography. The main part of the article is concerned with formal modeling of households and families, in particular with strategies to overcome the inherent complexity of these models. Three classes of strategies are discussed in more detail: severing the link between related individuals; macro- versus micromodels; static versus dynamic models.

'Family and household demography' differs from traditional demography in that it explicitly recognizes and studies *relationships between individuals*. In traditional demography, the core processes of birth, death, marriage, divorce, and migration are typically studied as occurring to individuals in isolation. In contrast, family and household demography studies how these processes occur to multiple persons interacting with each other. In doing so, additional relevant processes like entering or leaving coresidence are also studied. In short, relationships between individuals are the central focus of family and household demography. This article outlines the main subfields within the broad area of family and household demography, as well as the most important formal concepts.

Classification by Subject Matter

Three types of relationships between individuals can be distinguished (Ryder, 1987):

1. *Consanguineal* relationships run via the fundamental parent–child relation. Any two persons sharing a common ancestor, i.e., linked by a sequence of birth events, are said to be related by blood.
2. *Coresidence* relationships are between persons living together. A foolproof definition of coresidence does not exist (think, for instance, of student housing, lodgers, or part-time households), but 'sharing roof and manger' is not bad for a start.
3. *Conjugal* relationships are between individuals who are married to each other. In fact, living together in marriage is a special form of coresidence. To many people, the bond of marriage is so special that it is (almost or fully) put on the same footing as the consanguineal bonds mentioned under (1). In modern societies, the traditional marriage is gradually replaced by cohabitation (consensual unions). In these societies, 'being married' is more and more replaced by the more neutral 'living together as a couple,' which includes both formal marriages and consensual unions, and frequently also includes homosexual cohabitation. In the first edition of the recommendations for population and housing censuses (Revision 1), the United Nations Statistical Division recommended that "Couples living in consensual unions should be regarded as married couples" (United Nations, 1998, p. 67). In the second revision published in 2008, the recommendation changed. A distinction is made between marital status and *de facto* union status. Marital status is the personal status of each individual in relation to the marriage laws or customs of the country. *De facto* union status refers to extralegal unions (including some consensual unions) of varying degrees of stability common in some countries. The recommendation is: "Couples living in consensual unions may, where appropriate, be regarded as constituting a family nucleus" (United Nations, 2008, p. 131).

The broad field of family and household demography studies all these links between individuals, in their own right but also in relation to other demographic processes (e.g., migration of household, mortality patterns in kin groups). Within the broad field, several subfields can be identified, according to the type of relationship on which they concentrate:

1. *Demography of kinship* studies links of types (1) and (3).
2. *Demography of households* studies links of type (2), but also of type (3) since 'living together as a couple' is a special form of 'living together'.

[†] Original author deceased.

3. *Demography of the family* studies the simultaneous links of types (1) and (3), on the one hand, and of type (2), on the other: a family is a group of coresident kin.
4. *Demography of marriage* studies links of type (3).

Data Issues

The data sources for household and family demography are the same as for demography in general: censuses, population registers, surveys (in kinship studies, genealogies are also sometimes used). However, *definitional problems* are particularly strong in household and family data.

There are two sets of definitional problems. The first relates to criteria for establishing *whether or not a certain relationship exists* between two persons. Examples of such problems include:

1. Two persons (e.g., a mother and a student child) live together during weekends only. Do they belong to the same household?
2. Two persons form a household. How to judge whether they 'live together as a couple?'
3. Two students live in an apartment, occupying separate rooms but sharing kitchen and bathroom. Do they count as one or as two households?

The second group of definitional problems relates to criteria for *classifying households and/or family into categories*, like those used in published household statistics. Examples include:

1. A woman lives with her 30-year-old son. Should this household be termed a lone-parent household?
2. A lone mother cohabits with a new partner. Is this a 'couple with children' household, or rather a 'one-parent family plus other(s)?'

Such definitional degrees of freedom result in actual definitional differences, the consequences of which are potentially very large (Keilman, 1995). Therefore, demographers must be well aware of the precise definitions underlying their household and family data. In descriptive studies, definitional differences can seriously distort international or intertemporal comparisons. In modeling exercises, it is crucial that model parameters are derived from data that are definitionally consistent with the variables in the model. In order to stimulate international comparability of household and family data, the United Nations has specified detailed, extremely useful recommendations for data collection, definition, and tabulation in population and household censuses (United Nations, 2008).

Unit of Analysis

A major problem in the formal demography of families and households is whether the group or the individual should be taken as the primary unit of analysis. Almost by definition, the behaviors of the individuals within the group are highly interdependent. Therefore, analyzing the behavior of the group cannot be done without proper account of the characteristics of its members; conversely, analyzing the behavior of the individual cannot be done without proper account of the characteristics of the group (including those of the other individuals in the group).

Today, most family and household demographers agree that it is most convenient to work at the level of *individuals*. The main reason for this is that, unlike individuals, groups do not have a clearly defined origin and destination in time since members can enter and leave all the time. Thus, modern practice is to define a group in terms of its members and the relationships between them, rather than defining individuals in terms of the group to which they belong.

Because of strong interdependence of within-group behavior, proper modeling of individual family or household members' behavior requires sufficient information on the characteristics of the other individuals involved. In microlevel data, this implies the inclusion of *pointers* to other individuals in the data set, with sufficient detail on the type of the relationship to that other individual. Usually, a reference person is selected, e.g., the oldest person in the household or the head of the household. A main advantage of such data is that it can provide very rich descriptive information on different family and household types, as opposed to standard census breakdowns by just a few main categories. Because of the problems associated with the selection of a reference person in the household to whom all other household members report or designate their relationship, a new method has been proposed for mapping household structures. It is the household relationship matrix method (Maté and Miller, 2003). The household relationship matrix allows for the collection of all relationships between all household members. The method was developed by the census offices in the United Kingdom in preparation of the 2000 census. The Conference of European Statisticians recommended the method for the 2010 Censuses of Population and Housing (United Nations Economic Commission for Europe, 2006).

Naturally, when using sample survey data, proper account should be given to the representativeness of the data at different units of analysis. For example, a random sample of individuals is *not* representative at the household level, since large households will be overrepresented. Weighting schemes must be used to correct this and other potential distortions.

Models in Family and Household Demography

A model is a simplified, quantitative description of reality. In family and household demography, models provide a formal description of the quantitative relationships between family or household characteristics and their demographic determinants and consequences.

At the heart of any modeling exercise lies the specification of the *state space*: the representation of the components of the system of interest. At the level of the *individual*, the state space consists of a number of characteristics or attributes (e.g., age, sex, position in the household, age of spouse, number of children), each of which can take a certain value. At the population level, the state space consists of all possible combinations of attribute values: it is a breakdown of all individuals comprising the population by relevant characteristics.

The ultimate purpose of model building is to obtain an understanding about this multidimensional breakdown of

the total population. A classical demographer is satisfied with a breakdown by age and sex. A marriage demographer would want to add at least 'marital status' as an additional variable. A kinship demographer would want to add a complete breakdown by number of living children, grandchildren, siblings, parents, and so on. A household demographer would want to have, next to age and sex, variables like household size and household type, in a classification that is sufficiently detailed for the research or policy problem at hand. All these demographers have in common that they want to understand why an observed population breakdown is how it is; how particular changes in behavior (for example, total fertility level) would change this breakdown; or how the currently observed breakdown will develop over time (forecasting).

The particular problem in household and family models is that the relevant state space is potentially very large and complicated. This has forced family and household demographers to sacrifice all kinds of descriptive, behavioral, and/or conceptual detail. Evidently, the result is that existing household and family models reflect a wide range of different modeling strategies. Four important classes of these strategies are discussed in the sections below.

Severing the Link

As said before, family and household demography is all about relationships between individuals. A particularly important type of 'sacrifice' frequently made by family and household demographers is to discard most of the information on the characteristics of the individuals *with whom* a particular relationship exists. For example, in the case of a marriage model, a woman aged 40 and married to a man aged 42 is classified as being a married woman aged 40: the information about the age of the husband is discarded.

An example of this 'severing the link' strategy is provided by marital-status models of the multistate type (*see* Multistate Transition Models in Demography). Here, individuals are classified by sex s (m for men, f for women), age x, and marital status i, where i can take the following values: N for never married, M for married, D for divorced, W for widowed. The development over time of the population is described in terms of *events*, which are changes in the attributes of the individual members of the population. For example, next year's 31-year-old married female population is given by (ignoring migration)

$$\begin{aligned} P(f,31,M,t+1) = {} & P(f,30,M,t) + E(f,30,N,M,t) \\ & + E(f,30,D,M,t) + E(f,30,W,M,t) \\ & - E(f,30,M,D,t) - E(f,30,M,W,t) \\ & - E(f,30,M,\text{dead},t) \end{aligned}$$

where $P(s, x, i, t)$ denotes the population at the start of year t of sex s, age x, in marital status i, and $E(s, x, i, j, t)$ denotes the number of individuals of sex s and age x who during year t experience an event from position i to position j. The various numbers of events depend on the size of the relevant population at risk of experiencing the event, and on the rate at which individuals in the population at risk experience the event. For example, the number of women who marry for the first time at age 30 in year t, i.e., $E(f, 30, N, M, t)$, equals the product of two factors: the number of never-married 30-year-old women $P(f, 30, N, t)$, and the first-marriage rate for 30-year-old women $M(f, 30, N, M, t)$.

Given a full set of age-, sex-, origin-, and destination-specific rates, the population by age, sex, and marital status can be projected forward in age and time. Also, a marital-status life table can be calculated. However, a usual characteristic of this model is that marital-status events for one sex are independent of the demographic situation for the other sex. As a result, in this model it is not guaranteed that the resulting population flows for men and women will be mutually consistent. For example, the number of married men having a divorce should equal the number of married women having a divorce, but there is no mechanism in the model to ensure this equality. This problem is known as the *two-sex problem* in nuptiality models (Keilman, 1985a). It is a direct consequence of the fact that the explicit link between marriage partners is missing. There are several solutions to the two-sex problem, each of which introduces fresh problems:

1. To restore the link between marriage partners. This greatly expands the state space, since married persons must be simultaneously classified by own age *and* by the age of their spouse. It also greatly complicates the modeling of numbers of events, since rates must be defined for all combinations of male *and* female ages and marital statuses (*see* Population Dynamics: Two-Sex Demographic Models).
2. A slightly different way of restoring the link between marriage partners is to recast the model in terms of *microsimulation* (see Section Macro- versus Micromodels).
3. To concentrate on one sex (usually: women) only. This is convenient, but obviously unrealistic.
4. To introduce an adjustment mechanism, by which aggregate numbers of events (i.e., aggregated over all ages) are forced to satisfy certain requirements. For example, if the model initially produces 1000 male divorces and 800 female divorces, male divorces are adjusted downward and female divorces upward to result in, say, 900 divorces for each sex. Such procedures are called *consistency algorithms* (Keilman, 1985b; Van Imhoff, 1992). Although such algorithms can be given a satisfactory behavioral interpretation, they nevertheless have an artificial ring and are certainly rough approximations only.
5. A combination of the previous two 'solutions' leads to consistency algorithms that are one-sex dominant. For example, in a female-dominant marital-status model, the total number of male divorces is determined by the female married population and female divorce rates.

Marital-status models are by no means the only models in which 'severing the link' leads to consistency problems. All household and family models which are of the multistate type suffer from essentially the same problem.

Macro- versus Micromodels

In principle, an algebraic representation of a model can be manipulated to study the model's properties. If the model is not too complicated, it could be analyzed in this way. Because family and household models easily get quite complex,

analytic models are rare in family and household demography. A good example of an analytic model (although valid under a number of quite restrictive assumptions only) is the kinship model of Pullum (1987).

For more complex models, analytic manipulation becomes infeasible, so that *numerical simulation* methods have to be used to study the implications of the model. Here, two fundamentally different strategies can be distinguished: *macro*simulation versus *micro*simulation (Van Imhoff and Post, 1998; *see* Microsimulation in Demographic Research).

In macromodels, numbers of events (e.g., 'first marriages to women aged 25') are obtained by applying rates (e.g., 'first-marriage rate for 25-year old women') to a *group* of individuals of a certain size (e.g., 'number of never-married women aged 25'). In micromodels, events are obtained by applying probabilistic decision rules (Monte Carlo experiments) to *individual persons*, separately for a number of individual persons. To follow the example of first marriages: for each individual never-married woman aged 25 in the database, draw a random number between 0 and 1; if this random number is less than the first-marriage probability, the woman is deemed to marry, otherwise she remains unmarried.

In macrosimulation, the population is cross-classified by attributes and calculations are carried out in terms of the cells in the cross-classification: for each cell, the model should evaluate how the number of individuals will change over time. Microsimulation, on the other hand, concentrates on individuals in a population: for each individual, the attribute vector is updated according to the specifications of the model and the results of the Monte Carlo experiments. The storage and retention of information in microsimulation occurs via a *list* of individuals and their attributes; in macrosimulation this is done via the aggregate cross-classification *table*. If there are K attributes and M_i categories for attribute $i = 1 \ldots K$, the table at the macrolevel consists of $M_1 \times M_2 \times \ldots \times M_K$ cells; in contrast, at the microlevel a total population of N individuals can be described by a matrix with $N \times K$ cells. For most applications where a relatively large number of attributes is considered, the size of the aggregate table is much larger than the size of the list. This is particularly the case in household and family models, where characteristics of the individuals to whom relationships exist need to be included. In macromodels this would lead to incredibly large tables.

In micromodels, contrary to macromodels, it is quite easy to keep links between individuals intact, simply by including in the individual records of the database some reference numbers (pointers) to other records (persons) in the database. As a result, the consequences of an event that is simulated for one person can be easily determined and updated for the other persons involved. Thus, the two-sex problem or more generally the consistency problem discussed in the previous section is more or less 'automatically' solved in micromodels. The other side of the coin is that, for events that involve creating a *new* link between individuals, a matching problem has to be solved (e.g., if a woman is simulated to marry, a concrete husband must be identified).

Another main advantage of microsimulation is that it can provide very *rich output*. The output of a microsimulation model consists of a database with individual data that can be aggregated in an almost infinite number of ways. Compare this with the situation in macromodels, where the aggregation scheme is fixed once the model (i.e., the state space) has been specified. Apart from detailed cross-sectional counts, the microsimulation database can also be used to construct longitudinal information, e.g., in the form of individual biographies.

In short, microsimulation offers three main advantages that are of particular relevance in family and household demography: it can handle a large state space; relationships between individuals can easily and explicitly be retained; and it provides richer output.

The best-known microsimulation model in the field of family and kinship is SOCSIM (SOCiodemographic micro-SIMulation), developed by Hammel and Wachter at Berkeley with Laslett at Cambridge University (Wachter et al., 1978; Hammel et al., 1990). The SOCSIM model uses microsimulation to build up kinship links for an initial population of unrelated individuals that is subject to rates of fertility, mortality, and nuptiality (including divorce). Partners are found by a scoring algorithm that has a preferred spousal age difference of about 2 years and restrictions on the age difference of partners. Murphy (2004) used SOCSIM to model the population of England and Wales during two periods of 600 years: 1250 to 1850, and 1750 to 2350. Murphy uses historical rates and contemporary rates to produce the demographic events. Hammel (2005) applied SOCSIM to study the effects of demographic shocks in anthropological populations. A model that is closely related to SOCSIM is CAMSIM (CAMbridge SIMulation), developed for the simulation of kin counts (Smith, 1987). A kin count is a set of numbers expressing how many kin of various types exist for an individual at some time. Zhao (2006) compares the SOCSIM and CAMSIM models.

Dynamic versus Static Models

Populations change over time, and so do household and family structures. Thus, if a model is to describe this change, it must always contain the *time* element in one way or another. In this sense, all demographic models are dynamic by definition. However, there are many ways in which time can be included into a model. Merely adding an index t to all model variables and parameters hardly warrants the term 'dynamic.' A truly dynamic model should not only specify what the system looks like 10 years from now, but also how the system is supposed to get from here to there. In other words, the *processes* that underlie the changes in the system variables should be explicitly included in the model. In a truly dynamic model, the focus is on the *events* of formation and dissolution of relationships between individuals. The LIPRO (LIfestyle PROjections) household model (Van Imhoff and Keilman, 1991; Van Imhoff, 1995; Van Imhoff, 2005) and the ProFamy model (Zeng et al., 1997, 2014) belong to that category. For a comparison of these two models, see Willekens (2009).

In contrast, *static* models do little more than compare household and family structures at different points in time, thus essentially treating the underlying processes of household and family formation and dissolution as a black box. The classical headship ratio models belong to this category. The headship ratio is the proportion of household heads in a population category, e.g., '80 percent of all 40-year old males are

head of a household.' If from a population forecast (*see* Population Forecasts) we know that the number of 40-year-old males will increase from 1000 to 1500, and we extrapolate that the headship ratio for 40-year-old males will increase from 80% to 90%, then it follows that the number of households headed by a 40-year-old male will increase from 800 to 1350. With headship ratios for all ages and both sexes, a full population forecast can in this way be easily converted into a forecast of the number of (private) households. Several generalizations of the classical headship ratio model exist (Kono, 1987; Wilson, 2013). However, they all suffer from the disadvantage that the headship ratio is a static concept: the model cannot tell us why the current ratio of 80% will increase to 90%. The major limitation is that these models do not consider the processes of household formation and dissolution. In a way, the dynamic vs. static issue is gradual rather than fundamental. For instance, if a demographic forecaster hypothesizes that marriage rates will fall by 10% between now and 10 years into the future, one could argue that the marriage component of the model is static since the demographic forecaster does not specify *how* the marriage rates are going to fall. Similarly, the headship ratio method is generally termed static since the changes in age-specific headship are not explicitly specified; however, the changes in the age-specific population size to which these headship ratios are applied *are* explicitly modeled, so the headship ratio model is dynamic at least to some extent.

In fact, leaving some model variables static rather than 'truly' dynamic is an effective strategy to reduce the complexity of the model. The decision of where to stop making variables dynamic is governed by the fundamental tradeoff between information intensity on the one hand, and the capacity to make meaningful predictions on the other. This tradeoff must be faced in every effort of modeling human behavior.

Agent-Based Models

The solutions to the two-sex problem, listed in Section Severing the Link, are all second-best solutions. The two-sex problem typically arises when sex-specific marriage rates are used while interactions between the sexes in the marriage market are disregarded. The two-sex problem can be resolved by explicitly modeling interactions and their changes over time. The theoretical basis is the theory of partner search. During specific stages of life, individuals search for a partner and when a partner meets an individual's preferences, the (one-sided) search is successful. To result in a match, the two partners should have a successful search, i.e., the success should be mutual. A match may result in a stable relationship, cohabitation, and marriage. Becker (1981) shows that in a perfect marriage market the total gain over all marriages is maximized when individuals marry other individuals with similar traits, "regardless of whether the trait is financial, biological, or psychological" (Becker, 1981, p. 71). The process by which likes match with likes is known as *assortative mating*. Partner search is a matching process and assortative mating is one possible outcome of that process. Explicit modeling of the partner search mechanism involves the assignment of behavioral rules and decision rules to individuals and to model the interactions and relationships. This type of model is known as individual-based or agent-based models (ABM). Todd et al. (2005) present such an agent-based model of partner search. Billari et al. (2007) model the behavior of marriageable persons in a social network (the 'Wedding Ring'). The social pressure they experience to marry depends on the proportion married and their duration of marriage.

The formal demography of families and households is a relatively young subdiscipline of demography. It has a history of about 30 years. The subdiscipline faces many challenges (Willekens, 2009). A first challenge is to improve data availability and data quality. Longitudinal data are ideal because they track individuals over time and allow estimating rates of transition that depend on personal attributes. A second challenge is to introduce behavioral theory in family and household models and to discover the individual heuristics that produce the patterns of change in families and households. ABM is the way to go. The nature and structure of families and households vary between cultures and change over time. What is universal, however, is that relationships and ties between individuals result in relatively stable social units the demography of which is highly relevant to society. The demography of these groups is the subject of family and household demography.

See also: Family Theory: Economics of Marriage and Divorce; Household Archaeology; Households in Anthropology; Kinship in Anthropology.

Bibliography

Becker, G., 1981. A Treatise on the Family. Harvard University Press, Cambridge, MA.

Billari, F.C., Prskawetz, A., Diaz, B.A., Fent, T., 2007. The "Wedding-Ring": an agent-based marriage model based on social interaction. Demographic Research 17 (3), 59–82.

Grebenik, E., Höhn, C., Mackensen, R. (Eds.), 1989. Later Phases of the Family Cycle. Clarendon Press, Oxford, UK.

Hammel, E.A., 2005. Demographic dynamics of kinship in anthropological populations. Proceedings of the National Academy of Sciences 102 (6), 2248–2253.

Hammel, E.A., Mason, C., Wachter, K.W., 1990. SOCSIM II, a Sociodemographic Microsimulation Program, Rev. 1.0, Operating Manual: Graduate Group in Demography. Working Paper No. 29. University of California, Institute of International Studies, Program in Population Research, Berkeley, California.

Keilman, N.W., 1985a. Nuptiality models and the two-sex problem in national population forecasts. European Journal of Population 1, 207–235.

Keilman, N.W., 1985b. Internal and external consistency in multidimensional population projection models. Environment and Planning A 17, 1473–1498.

Keilman, N.W., 1995. Household concepts and household definitions in Western Europe: different levels but similar trends in household developments. In: Van Imhoff, E., Kuijsten, A., Hooimeijer, P., Van Wissen, L. (Eds.), Household Demography and Household Modeling. Plenum, New York/London.

Keilman, N.W., Kuijsten, A., Vossen, A. (Eds.), 1987. Modelling Household Formation and Dissolution. Clarendon Press, Oxford, UK.

Kono, S., 1987. The headship rate method for projecting households. In: Bongaarts, J., Burch, T.K., Wachter, K.W. (Eds.), Family Demography. Clarendon Press, Oxford, UK.

Maté, I., Miller, G., 2003. The UK 2001 census question on within household relationships. Statistical Journal of the United Nations Commission for Europe 20 (1), 27–37.

Murphy, M., 2004. Tracing very long-term kinship networks using SOCSIM. Demographic Research 10 (7), 171–196.

Pullum, T.W., 1987. Some mathematical models of kinship and the family. In: Bongaarts, J., Burch, T., Wachter, K. (Eds.), Family Demography:Methods and Their Application. Oxford University Press, Oxford.

Ryder, N., 1987. Discussion. In: Bongaarts, J., Burch, T.K., Wachter, K.W. (Eds.), Family Demography. Clarendon Press, Oxford, UK.

Smith, J., 1987. The computer simulation of kin sets and kin counts. In: Bongaarts, J., Burch, T.K., Wachter, K.W. (Eds.), Family Demography. Methods and Their Application. Clarendon Press, Oxford, pp. 249–266.

Todd, P.M., Billari, F., Simao, J., 2005. Aggregate age-at-marriage patterns from individual mate- search heuristics. Demography 42 (3), 559–574.

United Nations, 1998. Principles and Recommendations for Population and Housing Censuses, Revision 1. United Nations Department of Economic and Social Affairs, New York. Statistics Division, Statistical Papers Series M No. 67/Rev.1.

United Nations, 2008. Principles and Recommendations for Population and Housing Censuses, Revision 2. United Nations Department of Economic and Social Affairs, New York. Statistics Division, Statistical Papers Series M No. 67/Rev.2.

United Nations Economic Commission for Europe, 2006. Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing. United Nations Publication ECE/CES/STAT/NONE/2006/4, United Nations, New York, ISSN:0069–8458.

Van Imhoff, E., 1992. A general characterization of consistency algorithms in multidimensional demographic projection models. Population Studies 52, 159–169.

Van Imhoff, E., 1995. LIPRO: a multistate household projection model. In: Van Imhoff, E., Kuijsten, A., Hooimeijer, P., Van Wissen, L. (Eds.), Household Demography and Household Modeling. Plenum, New York/London.

Van Imhoff, E., 2005. LIPRO 4.0 Tutorial. Netherlands Interdisciplinary Demographic Institute, The Hague.

Van Imhoff, E., Keilman, N., 1991. LIPRO 2.0: An Application of a Dynamic Demographic Projection Model to Household Structure in the Netherlands. Swets & Zeitlinger, Amsterdam/Lisse.

Van Imhoff, E., Post, W., 1998. Microsimulation methods for population projection. Population 52, 889–932.

Wachter, K.W., Hammel, E.A., Laslett, P., 1978. Statistical Studies of Historical Social Structure. Academic Press, New York.

Willekens, F., 2009. Family and household demography. In: Zeng, Yi. (Ed.), Demography. In: Encyclopedia of Life Support Systems (EOLSS), Developed under the Auspices of the UNESCO. Eolss Publishers, Oxford, UK. http://www.eolss.net/.

Wilson, T., 2013. The sequential propensity household projection model. Demographic Research 28 (14), 681–712.

Zeng, Y., Vaupel, J.W., Wang, Z., 1997. A multidimensional model for projecting family households – with an illustrative numerical application. Mathematical Population Studies 6 (3), 187–216.

Zeng, Y., Land, K., Wang, Z., Gu, D., 2014. Household and Living Arrangement Projections. The Extended Cohort-Component Method and Applications to the U.S. and China. Springer, New York.

Zhao, Z., 2006. Computer microsimulation and historical study of social structure: a comparative review of SOCSIM and CAMSIM. Revista de Demografia Historica XXIV (2), 59–88.

Families as Educational Settings

Beth Kurtz-Costes, University of North Carolina, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sociocultural theory, social cognitive theory, and bioecological systems theory depict how children's learning occurs within the family. Physical characteristics of the home and parent language use provide a backdrop for children's education. Parents' beliefs that shape their educationally relevant behaviors include perceptions of the child's competencies, expectations, stereotypes, causal attributions, and values. Children learn from direct instruction and by observing behaviors modeled by their parents and siblings. Learning is also shaped by parents' school involvement and selection of children's out-of-home activities. Current scholarship explores cultural, national, and social class diversity in these processes using complex models and sophisticated data analytic strategies.

Introduction

As the primary context for child development, families – including the home environment; parents' beliefs, values, goals, and expectations for their children; parent-child and child-sibling interactions – are an important setting for children's education. Education is viewed here in broad terms, reflecting not only academic standing in the formal setting of school, but also children's cognitive development, including aspects of cognition such as acquisition of language, problem-solving skills, memory development, and general knowledge. In that vein, family influences on children's educational progress are myriad, as family not only provides the primary context for children's development but also reflects the broader societal and cultural contexts within which the family is embedded.

Classic Theories

The educational functions of the family in shaping children's development can be understood by drawing from three classic theories: Vygotsky's sociocultural theory (Vygotsky, 1978), social learning theory (Bandura, 1997), and Bronfenbrenner's (1979) bioecological systems theory. Each of these is summarized below.

Vygotsky's Sociocultural Theory

Vygotsky described children's knowledge acquisition as 'internalization,' reflecting that knowledge is culturally based, and that children's knowledge and skills are acquired through their experiences with social partners. Children's internalization of shared cultural knowledge occurs through the process of guided participation, by which parents and other social partners impart knowledge to children through direct instruction and by showing children how to perform actions. Given the socially constructed nature of children's cognition, at any given time, a child is capable of performing tasks within a given range, which Vygotsky named the 'zone of proximal development' (ZPD). At the lower end of the zone, the child can perform tasks independently. With assistance from a more knowledgeable social partner, the child can successfully complete a level that is higher within the zone. As the child learns more through guided participation, the zone shifts to reflect the child's greater skill and understanding. In order to tailor instruction to the developmental level of the child, parents and other social partners use scaffolding. Scaffolding refers to adjusting the level of support or instruction according to the cognitive level of the child in order that the child might learn new information. Thus, appropriate scaffolding enables the child to perform at a higher level within the ZPD, and might also result in shifting the child's ZPD upward.

For Vygotsky, development occurs as a result of children mastering the tools (e.g., language, counting systems) that are used and valued within their society. Thus, according to sociocultural theory, learning occurs primarily within social interactions and what is learned depends upon the child's culture. Vygotsky argued that both the content and the process of children's educational progress are inherently social in nature. This perspective is quite different from that of Piaget, who viewed the child as an active, independent problem-solver who acquires knowledge about the world through manipulation and active experimentation.

Bandura and Social Cognitive Theory

The social basis of learning is also stressed within social cognitive theory. In contrast to classical learning theory, which posited that individuals must be reinforced to learn a behavior, Bandura argued that many behaviors are learned through observation. For example, in Bandura's famous Bobo doll experiments, children who viewed an adult model using a blown-up doll as a punching bag were more likely to hit the doll themselves than children who did not see such aggression. Thus, Bandura showed that many behaviors are learned by watching a model perform the behavior, and then by imitating the behavior afterward. However, not all behaviors that are observed are learned by the child. Whether or not an observed behavior is subsequently imitated depends on factors linked to the learner (e.g., whether or not the learner was paying attention), characteristics of the model (e.g., whether or not the model is admired), and the situation (e.g., whether the behavior is rewarded). Because behaviors are more likely to be learned when they are modeled by powerful, highly regarded

models, children are particularly likely to learn behaviors that are modeled by their parents and older siblings.

The likelihood that a behavior is learned is shaped by cognitive processes and moral beliefs. In the Bobo doll experiments, children who viewed an adult model punching the Bobo doll and being subsequently punished were less likely to hit the doll themselves than children who did not see the adult punished. Bandura also showed that morality and cognitive abilities cause children to perform learned behaviors selectively. For example, a child might learn that it is acceptable to use certain colorful expressions with siblings and parents but not in front of grandparents.

Bronfenbrenner and Bioecological Systems Theory

Bronfenbrenner's bioecological systems theory posits that child development occurs as a result of interactions between the genetic/biological characteristics of the child and a complex set of environmental systems that have the child at the center (Bronfenbrenner and Morris, 2006). The microsystem includes the contexts in which the child is found, such as home and school. Given that microsystem events are experienced directly by the child, this system usually has the strongest immediate influence. The mesosystem represents interactions between aspects of the microsystem but in the child's absence. For example, a parent's contact with the child's teacher occurs within the mesosystem. The exosystem comprises contexts that do not include the child, but which influence others with whom the child has contact. A parent's experiences in his or her workplace occur within the exosystem. The macrosystem encompasses the broader culture in which the child resides. Therefore, the macrosystem shapes aspects of the other systems and influences the child directly and indirectly. For example, a girl who lives in a culture that values domestic talents in women might be more likely than her brother to learn from their mother how to cook and clean the house. Finally, the chronosystem represents the dimension of time. Timing shapes children's development and learning at an individual level (e.g., with age, the child might spend more time with siblings and less time with parents) and at a societal level (e.g., children born in the 1990s were more likely to learn to use computers at a young age than children born in the 1970s).

Bioecological systems theory depicts how the family context, in which the child resides, is at the center of these overlapping systems which together form the context for the child's learning and development. Examples of these processes are given in the following summary, which describes three broad aspects of the home: the home environment, parents' beliefs, and parent and sibling behaviors.

Fostering Children's Learning: The Home Environment

As noted within bioecological systems theory, the microsystem or contexts within which children reside are particularly influential in shaping their development (Bronfenbrenner and Morris, 2006). Thus, the day-to-day environment in which the child lives, breathes, and plays shapes learning profoundly. Two aspects of the home environment are discussed here: the physical characteristics of the home environment and parent language use.

Physical Characteristics of the Home Environment

In attempts to understand individual differences in children's academic skills at the point of school entry and as they proceed through primary and secondary school, researchers have noted differences in characteristics of the home environment. Home environments that provide the best opportunities for children to learn are those in which children are safe from environmental toxins and trauma, have access to nutritious food and safe drinking water, are cared for by responsive caregivers, and enjoy residential stability (i.e., neither the individuals who constitute the family unit nor the location where they reside changes frequently) (Duncan et al., 2010). Children who live in neighborhoods with high levels of noise due to an adjacent airport or highway score lower on measures of attention, memory, and reading comprehension than children from quiet neighborhoods. Crowding – having a large number of individuals within a small space – is also associated with lower cognitive skills in children (Evans, 2006).

Children's exposure to television and other technology is recognized as a home characteristic that shapes learning and cognitive development. Although watching educational television programming has some benefits, particularly for children who live in impoverished environments, in general television viewing is associated with poorer child outcomes (Schmidt and Anderson, 2007). In contrast, research shows that use of computers enhances learning and cognitive development, even at early ages (Hofferth, 2010). It is not yet clear to what extent and in what ways availability of technology such as smart phone applications influences children's learning and development.

Children's educational progress is also shaped by the availability of formal and informal learning materials such as books, games, and toys. These materials foster children's development when they are appropriate for the child's age (Bradley et al., 1988). In infancy, children benefit from viewing complex patterns and manipulating objects. After infants learn to crawl, and later to walk, they benefit from adequate space in the home where they may freely explore without danger of injury. Age-appropriate games, books, and activities foster the child's development of verbal, spatial, and problem-solving skills. Researchers have noted that low-income families often do not have as many learning materials available to children in the home as higher-income families (Duncan et al., 2010). These differences in home environments lead to differences in children's skills at the point of school entry – differences that usually remain stable as children proceed through school.

Parent Language

Large differences have been found across families in parents' use of language in the home. The total amount of spoken language, parents' use of different words (i.e., the size of their vocabularies), the length of parents' statements to children, and the frequency with which parents respond to child-initiated conversation are all related to children's intelligence at a later

age (Hart and Risley, 1995). Parents' language use with infants guides the development of the infant's ability to distinguish among phonemes. Children speak at an earlier age if their parents speak directly to them before the infant produces recognizable speech.

Many caregivers spontaneously use 'motherese,' which is speech that is characterized by short utterances with common words, simple syntax, and repetition, usually using a high-pitched voice. Such infant-directed speech provides the scaffolding to help the young child to learn words and concepts. Sensitive caregivers often provide stimulation that fosters verbal learning – for example, by introducing new words and concepts at an appropriate rate so that the child can learn them.

For immigrant children, educational progress is shaped by whether or not English is their first language. On average, Latino children in the United States who speak a language other than English at home begin kindergarten with significantly lower math and verbal skills than their non-Hispanic White peers (Reardon and Galindo, 2009). However, these achievement gaps narrow between kindergarten and Grade 5, reflecting that as children gain proficiency in English, their school achievement catches up to that of peers. Language use in the home also influences the formal educational progress of immigrant children indirectly through parents' school involvement. Research with Latino parents has shown that communication difficulties for parents who do not speak English, combined with cultural misunderstandings (e.g., hesitancy to ask a teacher questions for fear of showing disrespect) lead to less school involvement on the part of Latino parents (Hill and Torres, 2010). However, Latino parents may compensate for less school involvement by providing more monitoring in the home of children's homework completion (Aud et al., 2010).

Parents' Beliefs and Values

Parents' beliefs affect children's educational outcomes both through their influences on parental behaviors and on children's developing beliefs (Okagaki, 2001). For instance, a parent will speak to her 6-month-old daughter more often if she believes the infant can benefit from 'conversations' with an adult partner. Parents' beliefs also shape the developing beliefs of their children, many of which serve a motivational role for the child. Thus, a child whose father conveys that a failure was due to lack of effort rather than lack of ability might be more likely to persevere on a future task (Weiner, 1979). The following types of parental beliefs are described here: parents' expectations for their children and perceptions of children's competence, cultural stereotypes about group differences in skills, causal attributions about the reasons for children's behaviors and their academic outcomes, and parents' values and interests.

Parents' Expectations of the Child and Perceptions of Children's Competencies

Whether or not they have been verbalized, parents have expectations for the behaviors they expect of their children at given ages. Most parents also have expectations for their children's performance in school and eventual educational attainment.

Parents' Expectations for Mature Behavior

Parents vary tremendously in their beliefs about children at different ages (McGillicuddy-De Lisi and Sigel, 1995). Some parents expect that 18-month-olds can be potty-trained, whereas other parents are nonplussed when the 3-year-old continues to have 'accidents.' Some parents may expect their 3-year-olds to master rudimentary concepts of number and quantity that other parents would not expect before the age of seven. Although parents may not be conscious of many of these beliefs, such beliefs nevertheless shape everyday activities in the household, family rules, and the messages children are given – both blatant and subtle – regarding their abilities. For example, parents who believe that an 8-year-old is capable of planning might expect their child to pack his/her school lunch; select weather-appropriate clothes; and budget time to include homework, soccer practice, and leisure in a given evening.

Parental expectations that infants and toddlers are able to self-regulate their behavior (e.g., refrain from crying when they experience distress) have been linked to inappropriately punitive parental behavior and more negative child outcomes (Burchinal et al., 2010). Among older children, expectations for mature behavior that are clearly communicated to the child and that are accompanied by warmth and nonpunitive discipline are associated with better child outcomes (Baumrind, 1973).

Parents' Perceptions of Children's Abilities and Expectations for Educational Attainment

Parents' perceptions of their children's abilities influence the activities they choose for their children and the extent to which they encourage their children in certain pursuits. For example, a parent who views a child as talented athletically might encourage the child's participation in team sports. In this case, parent perceptions of the child's competence influence whether or not the child is encouraged to try out for a soccer team. Parent perceptions also shape the level of challenge that the parent provides for the child in learning situations (Vygotsky, 1978). Parents balance their level of support and challenge based on their view of the child's current competence level.

In terms of parents' expectations for their children's academic success and their eventual academic attainment, it is clear that children's competencies inspire parents' expectations (i.e., parents have higher expectations of more competent children). In general, the research literature also indicates an independent effect of parent expectations on children's achievement: Parents with more ambitious expectations have, on average, children who excel academically (Zhang et al., 2011). Parents who have high educational attainment expectations for their children and who perceive that their children are capable are likely to foster achievement striving in their children and to provide supportive learning environments. Both parents' perceptions of their children and their expectations for their children's future accomplishments are often shaped by cultural stereotypes, which are discussed next.

Cultural Stereotypes

Within any society, cultural stereotypes shape expectations of and behaviors with the child. Cultural stereotypes associated with gender differences are common in the United States and many other countries. For example, girls are assumed to be better attuned to emotions and social relationships than boys, and therefore parents are more likely to talk about feelings with daughters than with sons. Although stereotypes are often based to some extent in actual group characteristics (e.g., many Scandinavians are tall and blonde), stereotypes sometimes lead to biased perceptions of group members, memory distortions, and differential treatment.

Parents' gender stereotypes are particularly powerful in shaping family processes. Because daughters tend to be viewed as more vulnerable than sons, boys are more likely to be encouraged to pursue activities outside the home that promote agency and autonomy. Girls, in contrast, might be encouraged to spend more time with parents, and usually have greater responsibility for domestic tasks such as cooking, cleaning, and care of younger siblings (Crouter et al., 2001). The gendered nature of children's experiences within the family is more pronounced in families where parents hold traditional gender beliefs. Children learn the behaviors modeled by their parents and often adopt these behaviors themselves as they reach adulthood.

Gender stereotypes – as well as other stereotypes based on social groups – vary across cultural groups. Within Latino families, for example, men are traditionally expected to be the economic providers for their family and to be responsible for most important decisions. Women, in contrast, are traditionally expected to be submissive to their husbands, passive, and self-sacrificing for the good of the family. The extent to which these traditional stereotypes are upheld within the family as well as attitudes in the child's mesosystem, exosystem, and macrosystem will influence the degree to which the child internalizes gendered norms and will practice such behaviors as an adult.

Parents' Causal Attributions regarding Their Children's Behaviors

Classic psychological theories (e.g., Rotter, 1966; Weiner, 1979) and empirical research demonstrate that individuals vary in their beliefs about the reasons for behaviors and for performance outcomes, and that such attributions shape both responses to others and expectations about future behaviors and outcomes. Two aspects of parents' attributions are discussed here: Parents' assumptions about the reasons underlying children's behaviors and misbehaviors, and parents' beliefs about the reasons for children's academic successes and failures.

Parents' Attributions about Child Compliance and Misbehaviors

Parents' beliefs about children's behavior shape their discipline practices, how and to what extent autonomy is fostered, and age-related expectations of their children. For example, a parent who believes that an infant is crying because she is willful and angry with her caregivers might be punitive (Burchinal et al., 2010). In contrast, a parent who attributes the crying to fatigue, hunger, or another source of physical distress will be more likely to comfort the infant and eliminate the source of distress. Parents of teenagers might attribute their children's misbehavior to peer pressure, poor judgment exercised on the road to learning autonomy, disrespect, or stupidity. Parents usually choose responses to child misbehaviors that reflect their attributions.

A substantial body of literature indicates more positive cognitive and social outcomes for children whose parents show warmth while exhibiting firm yet flexible control, a style called democratic parenting by Baldwin (1955), and later described by Baumrind (1973) as authoritative parenting. Authoritative parents rely primarily on reasoning as a discipline tool rather than punishment and view their children's good behavior as the primary goal of their parenting rather than obedience. Thus, parents who value their children's autonomy, and whose knowledge and beliefs are consistent with an authoritative parenting style are more likely to foster educational progress in their children, as marked by school grades, low incidence of discipline problems, and healthy self-esteem.

Some recent research indicates cultural variation in the relationship between parenting style and children's outcomes, both in the importance of autonomy encouragement and the possible detrimental effects of harsh discipline. Harsh discipline may be less detrimental to children in cultures in which it is sanctioned and in environments where misbehavior might put the child in danger (Dodge et al., 2005). The extent to which autonomy is valued and encouraged also differs across cultural contexts.

Parents' Attributions about Children's Academic Successes and Failures

For children in a formal educational setting, children and parents form beliefs about the reasons underlying the child's academic outcomes. When the child's academic success is attributed to ability, expectations for future successes are high because ability is viewed as an internal and stable characteristic of the child. In contrast, if success is attributed to an external, uncontrollable cause such as luck or an easy task, then parents and children have less confidence in future successes (Weiner, 1979). Causal attributions influence children's outcomes through their motivational effect on children's persistence in the face of failure and/or challenging tasks as well as through parents' and teachers' behaviors with children. Parents' attributions are related to their behaviors with their children such as homework support and encouragement to take advanced coursework.

Because academic stereotypes by their very nature include attributional content (i.e., one social group is assumed to be more talented than another), researchers have investigated linkages among cultural stereotypes, causal attributions, and outcomes such as parent behaviors with their children. This research has shown that parents have different expectations of the math successes of sons as compared to daughters, and that these expectations are played out in the types of activities that are provided to children in middle childhood (Jacobs et al., 2005).

Parents' Values and Interests

According to expectancy-value theory, goal-directed actions are guided by expectations that certain behaviors are likely to lead

to a desired outcome and by the extent to which the outcome is valued (Fishbein and Ajzen, 2010; Wigfield and Eccles, 2000). Therefore, a child will pursue a task (and parents will encourage children in that pursuit) when they believe the child is capable of mastery and when the task is valued. According to Wigfield and Eccles (2000), a goal is valued if the individual has high interest in the domain, if the goal is viewed as useful and important, and the cost of achieving the goal is not prohibitive.

Parents' values and interests comprise an important mechanism that leads to differences in children's learning both within and across cultures. Whereas knowledge of which basketball teams have made it to the Final Four is an important information to be learned in some families, other families place little value on sports knowledge and instead might have political discussions at the dinner table each evening. Parents' values and interests have the largest impact on children's learning in early childhood. By adolescence, youths' goal pursuits are increasingly shaped by peers and their own interests. However, families continue to be an important source of influence.

Parents' and Siblings' Behaviors

The beliefs and values that were discussed in the prior section influence children's educational outcomes through several mechanisms. Parents' beliefs and values are reflected in the structuring of the home environment – for instance, in parents' selection of toys for the child. Parents' beliefs also shape their behaviors – both direct interaction with the child, and behaviors that the child observes. In this section three aspects of parent behaviors are discussed: parents' and siblings' direct instruction and behaviors as role models for the child, parents' school involvement, and selection of out-of-home activities.

Parents and Siblings as Role Models and Teachers

As depicted in social learning theory, one way in which families have profound influences on children's learning is through the modeling of behaviors. Children who observe their parents reading are more likely to read themselves. Children are more likely to learn about organized sports if parents and/or siblings are involved in such sports. Children become 'socialized,' learning the acceptable behaviors of their culture, by observing parents and siblings. Thus, for example, a young Korean child will learn that a loud belch is an appropriate way to show appreciation for an excellent meal, whereas a child in Kansas will learn that such noises should be suppressed at the table.

Through observing parents and other family members, children learn cognitive skills such as language, social skills such as how to greet a stranger, and cultural attitudes such as appropriate dress for a party. The toddler observes an older sibling's interactions with friends and imitates that behavior later in other social situations. A young adolescent internalizes a social stereotype after observing his parent consistently becomes tense when encountering members of another racial group.

In addition to these examples of observational learning, children learn from the direct instruction of their parents and siblings. From birth onward, parents and siblings are an important source of children's knowledge. Through everyday conversations and activities with family members, children acquire concepts of schemas (e.g., what a visit to the doctor's office entails), grammar, and mathematical concepts (e.g., the early distinction between 'one' and 'more'; how to fairly divide the last piece of cake). Children learn memory strategies (e.g., where can I put my show-and-tell item so I will not forget it tomorrow morning), metacognitive knowledge (I cannot concentrate on my homework while big brother is listening to that music), and problem solving (How should I clean up this spilled sugar?).

Instruction that takes place in the home is markedly different from that found in the school setting, in that it is largely unplanned, informal, and context-grounded, rather than the planned, formal, and context-independent instruction that occurs in school. Most instruction in the home occurs spontaneously based on the specific events, feelings, and needs of a given moment in the family life. Because it is so highly context-dependent, this instruction from siblings and parents provides the basis for most of the child's knowledge and skills of how to survive in the day-to-day world.

Parents' School Involvement

One way in which parents model educationally oriented behavior to their children and shape the child's microsystem and exosystem is through their actions directed toward the child's formal education. Such school involvement takes many forms, occurring both inside and outside the home. In the United States, parents who volunteer in the child's classroom, attend school events, and find other ways to interact with teachers and other school personnel are more likely to be viewed positively by school personnel, have greater knowledge of school rules and events, and are more aware of what their children experience at school. Such school-directed involvement is associated with more positive achievement outcomes for children, perhaps because of parents' improved ability to provide scaffolding for their children, or perhaps because teachers tend to interact more with children whose parents are involved at school.

Other parental school involvement occurs within the home setting, where parents monitor children's homework completion, encourage and support them when children are challenged by difficult academic tasks, and provide tangible supports for children's academic progress. Researchers have noted that parents who are not at ease in the school setting – perhaps because of having little formal education themselves, or perhaps (for immigrant parents) because they are not fluent in the language used at school – are often likely to be involved in children's schooling in these ways within the home.

Low income, racial/ethnic minority, and immigrant parents face many barriers to school-focused involvement in their children's schooling (Nzinga-Johnson et al., 2009). In addition to the language barriers faced by many immigrant parents, educational practices within the host country may be foreign to immigrants (Hill and Torres, 2010). For example, extracurricular activities such as sports, academic clubs, and arts events (theater, music, dance) that are common in most schools in the United States are uncommon in many other countries. Single parents and parents who work more than

one job usually find it more challenging than dual-parent families to attend school events or volunteer in the classroom. Ethnic/racial minority parents who have had prior discriminatory experiences with school personnel may avoid interactions with their child's teacher, or might have interactions that are distrustful or hostile. Whereas such parental school involvement is encouraged and often expected in the United States, in many other countries parental presence in the school setting is not expected and in some cases is viewed negatively.

Out-of-Home Activities

As noted by Bronfenbrenner in bioecological systems theory, the immediate microsystem of the home is only one context in which families provide the backdrop for children's learning. Parent employment, which is usually part of the child's exosystem, shapes the child's development in several ways. Early investigations in parent employment and children's outcomes were mostly group comparisons aimed at determining whether or not maternal employment had a detrimental effect on children. These investigations indicated no differences in quantity or quality of parenting systematically related to maternal employment status. More recent research has identified process variables related to parent employment that have both positive and negative influences on children. According to scarcity or stress theory, parents have a limited amount of time and energy, and parental employment is deleterious to children's developmental outcomes to the extent to which it decreases parents' time, energy, and patience. Stress theory reflects a strong trend in the United States and many other industrialized nations: Namely, in spite of the impressive technological advances of the past century, most parents work harder than ever (Schor, 1991). Particularly as maternal employment rates have increased, the everyday lives of children have become harried. Thus one consequence of parental employment for children is the resulting stress associated with parental employment-related absence, whether that absence is actual or psychological, because stressed parents are preoccupied, tired, and impatient. A second consequence of demographic changes in employment patterns is that children spend a greater amount of time with other caretakers. Therefore, parents' selection of day care for infants and preschoolers, and their choices of afterschool activities for school age children have real impact on children's developmental outcomes.

Parental employment can also benefit children's educational progress. Parents gain in income, self-esteem, competency, and social support through their employment; these benefits for parents may translate into benefits for children. Mothers who are employed in professional positions serve as positive role models for daughters. In addition, parents' work characteristics are related to the quality of their parenting. Parents who enjoy their work, who are socially rewarded for it, and who are intellectually challenged in the work setting are more likely to show warmth and greater flexible but firm control with their children than parents who have less satisfying employment.

Parents of infants and preschoolers decide at what age the child will begin to spend significant amounts of time outside the home, and where that time will be spent. In middle childhood and adolescence, the child's afterschool activities and vacation time might be selected jointly by the child and parents, as well as other family members. Because of the interests, occupations, and habits of parents, siblings, and extended family members, one child may spend free afternoons playing basketball, while another goes to the shopping mall. Such influences demonstrate ways in which the mesosystem (e.g., at a school open house, a teacher suggests to the parent that a child enroll in a summer science camp) and exosystem (e.g., because of the parent's work expertise, the family uses vacation time to attend a national decathlon featuring solar-powered cars) might shape the child's microsystem.

Current Research Directions

In recent decades, researchers have made a concerted effort to expand our understanding of children's educational progress to cross national, racial, cultural, and social class lines. Whereas the majority of research studies in the twentieth century were conducted with White, middle class Americans, scholars are attempting to understand ways in which children's development varies in other countries and cultures, and the extent to which cultural context shapes such processes. In the United States and elsewhere, immigration patterns are changing the characteristics of the population, and therefore scholars are interested not only in how children's development differs across cultural groups, but also how the acculturation process shapes children's experiences. Thus, for instance, a child whose parents are not fluent in the language of the host country might have responsibilities for communication between his parents and school officials.

Another change in scholarship is a move toward more complex models in understanding child learning and development. The development of more sophisticated data analytic procedures has permitted researchers to address questions such as how change over time differs within and across groups, and how such change is differentially influenced by multiple levels of children's contexts. For example, current data analytic techniques permit researchers to determine how race composition of the child's classroom – which changes from year to year – is related to change in the child's academic performance, how this change differs according to the race of the child's teacher, and how these effects differ across schools and across racial groups within schools. Whereas 50 years ago researchers were more likely to focus on a single context or relationship (e.g., parent–child interaction during problem solving), researchers are increasingly examining how development emerges from the interplay of numerous agents (e.g., parent, sibling, and peer influences operating together) and at multiple levels (biological, behavioral, and societal).

The greater complexity in research designs is a reflection of changes in the theoretical approaches to understanding development, in which behavioral processes are viewed as emerging from the interplay of various systems ranging from the genetic to the cultural. Developmental researchers recognize that developmental processes such as children's educational progress can be understood best when taking into account these

multiple levels, and that interactions across the levels occur across time and lead to new patterns of organization in context (Spencer et al., 2009).

See also: Academic Self-Concept and Achievement; Family and Schooling; Intelligence in Cultural, Social and Educational Context; Intrinsic Motivation: A Cultural Perspective; Learning and Instruction: Social-Cognitive Perspectives; Parents and Teachers as Partners in Education; Sex Role Development and Education; Social Inequality and Schooling; Vygotsky's Theory of Human Development and New Approaches to Education.

Bibliography

Aud, S., Fox, M.A., KewalRamani, A., 2010. Status and Trends in the Education of Racial and Ethnic Groups (NCES 2010-015). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Baldwin, A.L., 1955. Behavior and Development in Childhood. Dryden Press, Ft. Worth, TX.

Bandura, A., 1997. Self-efficacy: The Exercise of Control. Freeman, New York.

Baumrind, D., 1973. The development of instrumental competence through socialization. In: Pick, A.D. (Ed.), Minnesota Symposium on Child Psychology, vol. 7. University of Minnesota Press, Minneapolis, MN, pp. 3–46.

Bradley, R.H., Caldwell, B.M., Rock, S.L., 1988. Home environment and school performance: a ten-year follow-up and examination of three models of environmental action. Child Development 59, 852–867.

Bronfenbrenner, U., 1979. The Ecology of Human Development: Experiments by Nature and Design. Harvard University Press, Cambridge, MA.

Bronfenbrenner, U., Morris, P.A., 2006. The bioecological model of human development. In: Lerner, R.M., Damon, W. (Eds.), Handbook of Child Psychology, sixth ed., vol. 1. Wiley, Hoboken, NJ, pp. 793–828.

Burchinal, M., Skinner, D., Reznick, J.S., 2010. European American and African American mothers' beliefs about parenting and disciplining infants: a mixed method analysis. Parenting Science and Practice 10, 79–96.

Crouter, A.C., Head, M.R., Bumpus, M.F., McHale, S.M., 2001. Household chores: under what conditions do mothers lean on daughters? In: Fuligni, A.J. (Ed.), Family Obligation and Assistance during Adolescence: Contextual Variations and Developmental Implications. Jossey-Bass, San Francisco, CA, pp. 23–41.

Dodge, K.A., McLoyd, V.C., Lansford, J.E., 2005. The cultural context of physically disciplining children. In: McLoyd, V.C., Hill, N., Dodge, K.A. (Eds.), African American Family Life: Ecological and Cultural Diversity. Guildford Press, New York, pp. 245–263.

Duncan, G.J., Ziol-Guest, K.M., Kalil, A., 2010. Early childhood poverty and adult attainment, behavior, and health. Child Development 81, 306–325.

Evans, G.W., 2006. Child development and the physical environment. Annual Review of Psychology 57, 423–451.

Fishbein, M., Ajzen, I., 2010. Predicting and Changing Behavior: The Reasoned Action Approach. Taylor and Francis, New York, NY.

Hart, B., Risley, T.R., 1995. Meaningful Differences in the Everyday Experience of Young American Children. Brooks, Baltimore, MD.

Hill, N.E., Torres, K., 2010. Negotiating the American dream: the paradox of aspirations and achievement among Latino students and engagement between their families and school. Journal of Social Issues 66, 95–112.

Hofferth, S.L., 2010. Home media and children's achievement and behavior. Child Development 81, 1598–1619.

Jacobs, J.E., Davis-Kean, P., Bleeker, M., Eccles, J.S., Malanchuk, O., 2005. 'I can, but I don't want to': the impact of parents, interests, and activities on gender differences in math. In: Gallagher, A.M., Kaufman, J.C. (Eds.), Gender Differences in Mathematics: An Integrative Psychological Approach. Cambridge University Press, New York, pp. 246–263.

McGillicuddy-De Lisi, A.V., Sigel, I.E., 1995. Parental beliefs. In: Bornstein, M.H. (Ed.), Handbook of Parenting, Status and Social Conditions of Parenting, vol. 3. Erlbaum, Hillsdale, NJ, pp. 333–358.

Nzinga-Johnson, S., Baker, J.A., Aupperlee, J., 2009. Teacher-parent relationships and school involvement among racially and educationally diverse parents of kindergartners. Elementary School Journal 110, 81–91.

Okagaki, L., 2001. Parental beliefs, parenting style, and children's intellectual development. In: Grigorenko, E.L., Sternberg, R.J. (Eds.), Family Environment and Intellectual Functioning: A Life-Span Perspective. Erlbaum, Mahwah, NJ, pp. 141–172.

Reardon, S.F., Galindo, C., 2009. The Hispanic-White achievement gap in math and reading in the elementary grades. American Education Research Journal 46, 853–891.

Rotter, J.B., 1966. Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs 80 (609).

Schmidt, M.E., Anderson, D.R., 2007. The impact of television on cognitive development and educational attainment. In: Wartella, E. (Ed.), Children and Television: Fifty Years of Research. Lawrence Erlbaum, Mahwah, NJ, pp. 65–84.

Schor, J.B., 1991. The Overworked American. Basic Books, New York.

Spencer, J.P., Samuelson, L.K., Blumberg, M.S., McMurray, B., Robinson, S.R., Tomblin, J.B., 2009. Seeing the world through a third eye: developmental systems theory looks beyond the nativist–empiricist debate. Child Development Perspectives 3, 103–105.

Vygotsky, L.S., 1978. Mind in Society. Harvard University Press, Cambridge, MA.

Weiner, B., 1979. A theory of motivation for some classroom experiences. Journal of Educational Psychology 71, 3–25.

Wigfield, A., Eccles, J.S., 2000. Expectancy-value theory of achievement motivation. Contemporary Educational Psychology 25, 68–81.

Zhang, Y., Haddad, E., Torres, B., Chen, C., 2011. The reciprocal relationships among parents' expectations, adolescents' expectations, and adolescents' achievement: a two-wave longitudinal analysis of the NELS data. Journal of Youth and Adolescence 40, 479–489.

Family and Kinship, History of

Tamara K Hareven, University of Delaware, Newark, DE, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 8, pp. 5290–5298, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article discusses the emergence and development of the historical study of the family and interprets them in the larger context of social history. It addresses the following issues: the emergence of the field of family history; household structure; developmental perspectives; kinship; the family's interaction with the process of industrialization; family strategies; and future directions in the field of family history.

This article examines the emergence of the history of the family as an interdisciplinary field in Europe and in the United States. It traces the field to its origins in France and England and its subsequent development in other European countries, Japan, and the United States. It discusses the characteristics of the field due to its origins from both historical demography and social and cultural history, particularly from the 'new' social history. The article examines the various dimensions of the historical study of the family, starting with the discovery of the continuity of a nuclear household structure since preindustrial society, and the presence of 'modern' characteristics of demographic behavior (such as family limitations) in Western Europe and North America, prior to industrialization. It proceeds to explore subsequent discoveries of the importance of kinship networks outside the household, and the salient role of kin in the process of industrialization and in adaptation to industrial life. Examining the development of scholarship in the History of the Family in the 1980s and 1990s, the article traces the emergence of the life-course approach, in response to the need for a developmental perspective and its utility in examining the interaction of 'individual time,' 'family time,' and 'historical time.' The article further examines the family's interaction with the process of industrialization. Contrary to prevailing myths about social breakdown, it emphasizes the family's role as an active agent in that process. In doing so, it discusses both the earlier stages of 'protoindustrialization' and the emergence of the factory system. In assessing the family's role as an active agent, it explores family strategies as an important process of family decision-making. Finally, the article examines critically, earlier studies of 'grand' changes in the family over time and the meaning of 'modern' family patterns. The article concludes with suggestions for continued and future research.

Emergence of the Field

Recent historical research on the family has revised some widely held myths about family life in the past, as well as generalizations about the impact of the grand processes of social change on the family and society. Family history has complex roots in both the historical demography of the early 1960s and the 'new social history' of the same period. Particularly in the United States, it has shared with the latter a commitment to reconstructing the family patterns of ordinary people and viewing them as actors in the process of change. Contemporary historians of the family have sought to reintroduce human experience into historical research and to emphasize the complexity of historical change (see Hareven, 1971, 2000a; Tilly, 1987). Hence, research in the history of the family explores previously neglected dimensions of human experience such as growing up, courting, getting married, bearing and rearing children, living in families, becoming old and dying. The formidable goal has been to understand the family in various contexts of change and at different points in historical time. In short, the historical study of the family represents an effort to understand the interrelationship between 'individual time,' 'family time,' and 'historical time' (Elder, 1978).

Before systematic historical research on the family began, various social science disciplines had generated their own myths and grand theories about continuities and changes in family behavior in the past. Sociologists in particular argued that, in preindustrial societies in the west, the dominant household form had contained an extended family, often involving three coresident generations, and that the 'modern' family – characterized by a nuclear household structure, family limitation, the spacing of children, and population mobility – was brought about by industrialization. Associated with these generalizations was also the popular myth that industrialization broke up traditional family patterns and community life. By providing a perspective on change over time, as well as on family behavior within specific social and cultural contexts in discrete time periods, historical research on the family has led to the rejection of these assumptions and to the resulting questioning of the role of industrialization as a major watershed in this field for American and European history (Wrigley, 1972; Goode, 1963).

Over the three-and-a-half decades of its existence, family history has moved from a limited view of the family as a static unit at one point in time to an examination of the family as a process over the life course of its members; from a study of discrete domestic structures to the investigation of the nuclear family's relations with the wider kinship group; and from a study of the family as a separate domestic unit to an examination of its interaction with the worlds of religion, work, education, correctional and welfare institutions, and with processes such as migration, industrialization, and urbanization (Vinovskis, 1977; Hareven, 1982).

As research in this field developed, new findings and approaches have led to the revision and augmentation of the pioneers' approaches. Efforts to explore decision-making processes within the family have led to an investigation of strategies and choices that individuals and family groups make. The life-course approach has added an important developmental dimension by focusing on age and cohort comparisons in ways that link individual and family development to historical events. Research also expanded chronologically to cover antiquity, and geographically to encompass southern Europe, eastern Europe, the Mediterranean, Japan, China, and Latin America, in addition to Western Europe, North America, and Japan. The cumulative impact of studies in the history of the family has been to revise simplistic views of both social change and family behavior (Reher, 1997; Andorka, 1994; Plakans and Wetherell, 1999; Todorova, 1997; Halpern et al., 1996; Hayami and Uchida, 1972; Kurosu, 1996; Lee and Campbell, 1997; Lee and Feng, 1999; Borges, 1992; Kuznesof, 1997).

The emergence of the history of the family as a special area of inquiry received its major impetus from the publication of Philippe Aries's *Centuries of Childhood* (1960 in French and 1962 in an English translation). Aries argued that childhood as we know it originated only in the early modern period and that its discovery was closely linked to the emergence of the 'modern' or conjugal family, in which parents' private relationships with their children were more important "than the honor of a line, the integrity of an inheritance, or the age and permanence of a name" (Aries, 1962: p. 393). By linking the 'discovery' of childhood to transformations in family and social structures as well as economic and demographic changes, Aries inspired a whole generation of scholars.

Along another line of research, historical demographers in France in the early 1960s provided family historians with a powerful 'new weapon' for the analysis of demographic processes. Louis Henry and Pierre Goubert had developed a family reconstitution technique in the 1950s. Using genealogies, and marriage, baptismal, and death records from parish registers, this method enabled demographers to reconstruct aggregate patterns of fertility, nuptiality, and mortality for vast numbers of people and, in some instances, over several generations (Henry, 1953).

From the 1960s on, historical demography and family history in France developed into two parallel but interrelated streams. One continued to concentrate on demographic analysis, along the lines of Henry and Goubert; the other, influenced by Aries, anthropology, and the French social history tradition, integrated demographic analyses with patterns of family and sexuality, linking community and social and cultural variables with *mentalité*, as exemplified in the work of Emmanuel LeRoy Ladurie (1978), André Burguière (1987), Jean-Louis Flandrin (1976), and Martine Segalen (1991), among others. Family reconstitution subsequently became a powerful tool in the hands of the Cambridge Group for the History of Population and Social Structure.

Established in 1964, the Cambridge Group adapted the family reconstitution method to English parish registers, while also pursuing analysis of seventeenth-century nominal household registers. Wrigley found in the decline in fertility in the seventeenth century and in its recovery in the eighteenth century, evidence that rural births and marriages responded to changing economic conditions. He concluded that the demographic transition involved a change "from a system of control through social institution and custom to one in which the private choice of individual couples played a major part in governing the fertility rate" (Laslett, 1977; Wrigley and Schofield, 1981).

Such demographic analyses for France and England, and subsequently Sweden, Austria, Germany, and The Netherlands, revealed that, in the preindustrial period, people married later than had been generally assumed, couples practiced some form of family limitation and child spacing as early as the seventeenth century, households were predominantly nuclear rather than extended, and preindustrial populations experienced extensive geographic mobility (Laslett, 1965; Goubert, 1977). Late marriage served as a method of family limitation. It was also closely related to the expectation that a newly wed couple would establish a separate household and be self-supporting, while continuing to contribute to their families of origin. Linking late age at marriage to the nuclearity of the household, John Hajnal (1965) developed his model of the 'West European Marriage Pattern,' which served as the basic model for the analysis of West European families until it came under criticism. Hajnal claimed that marriage patterns in historic Europe could be differentiated with a line drawn approximately from St. Petersburg on the Baltic Sea to the Adriatic Sea. West of this line marriages tended to be relatively late (late 20s) for both genders, and substantial portions of the population did not marry at all. East of this line, marriage for both genders was relatively early (late teens) and marriage was nearly universal. He identified the 'western' pattern as the 'European' pattern and suggested that the 'eastern' pattern was shared by many developing countries (in the 1960s).

Household Structure

Using nominal census records, Laslett found evidence of continuity in nuclear household structure in England, at least since the sixteenth century. A landmark volume published by Laslett and Wall (1972), containing articles on household structure for various countries, including Japan, concluded that preindustrial households in Western Europe and North America affirmed these findings and dispelled the previously held assumption that industrialization brought about a nuclear family form. Although Laslett's early work emphasized the persistence of a nuclear 'family,' the major unit discussed was the household, not the family. A household is not identical with a family, as the domestic group may include nonrelatives as well. Nor is a 'family' restricted to the household, as extended family ties transcended the household unit (Goody, 1972; Hareven, 1982).

Subsequent studies showed that coresidence with extended kin tended to increase, not decrease, after the 'industrial revolution,' because of the need of newly arrived migrants to industrial cities to share housing space (Laslett and Wall, 1972; Hayami and Uchida, 1972; Anderson, 1971). David Herlihy (1987) and Richard Smith (1979), respectively, traced a nuclear household structure in medieval Tuscany back to the twelfth century, as well as in England; and, in the medieval

village of Montaillou, LeRoy Ladurie (1978) found a variety of forms of family structures in the *domus* (household), ranging from ordinary nuclear families to various complex configurations. Household structure varied over the life course of its members, in relation to the economy and migration.

Such early discoveries also led to the formation of new generations and stereotypes: by combining the findings about the nuclearity of the household with the Hajnal thesis on the Western marriage pattern, Laslett developed a typology of the 'western family,' which he characterized as having a simple nuclear household; a relatively late child-bearing age; a relatively narrow age gap between husband and wife, with a relatively high proportion of wives older than their husbands; and the presence in the household of 'life cycle' servants unrelated to the family with whom they were residing. By contrast, he characterized the South European and Mediterranean family type as having complex household structures and early marriage, in which nuclear households are formed by the breakup of extended ones, rather than through marriage (Laslett, 1977). These typologies, along with the Hajnal thesis, later came under criticism. David Kertzer refuted the thesis that links early female marriage to joint household organization as a characteristic of the 'South European' family pattern. He showed variability in household structures in Italy, depending on types of landholding and sharecropping. Similarly, Haim Gerber, on the basis of his analysis of urban Turkish households in the seventeenth century, questioned stereotypes of the traditional 'Middle Eastern family' (Kertzer and Hogan, 1989; Gerber, 1989).

In New England, when interpreted in the context of economic and social institutions (such as landholding, inheritance, community structure, and religion), demographic patterns served as the backbone of complex family and community analyses. In his analysis of household and family structure in colonial Andover, Massachusetts, Philip Greven revealed the power relations between fathers and sons and the limits of the sons' autonomy based on inheritance. He found that households were primarily restricted to members of the nuclear family and nonrelatives, but a 'complex web' of family connections reinforced by proximity of residence, often on the same land, permeated the community (Greven, 1970). Demos, in his study of colonial Plymouth, Massachusetts, used family reconstitution, the analysis of nominal census records and a great variety of other sources to reconstruct family and community patterns (1970). He also found a predominance of nuclear households, but these differed considerably from contemporary ones. Families were larger then and included nonrelatives, such as servants, apprentices, and boarders and lodgers in the households. Age configurations were also different; in some families, the oldest child in the household would be an adult while the youngest was still at the breast.

A major critique of early studies of household structure was their reliance on 'snapshots' in the household schedules listed in the census at one point in time. On the basis of his reconstruction of peasant households in the villages of Heidenreichstein in northern Austria in 1763, Berkner (1972) provided evidence that household structure changed several times over the course of the family's life, as it moved through its cycle – from a stem family to a nuclear one, and later to a stem family again, in relation to inheritance. Households were not static. At the same time, Demos' reconstruction of the stages of the life cycle in Plymouth demonstrated dynamic movements of individuals within and across families and households, in relation to the religious and cultural prescriptions of Puritan society (Demos, 1970).

The presence of 'life cycle' servants – young men and women in their teens who lived and served in other people's households in the transitional period between leaving home and marrying – which Laslett found in English villages, as well as the presence of other unrelated individuals in the household, proves the remarkable flexibility of households in the past. Nuclear households included nonrelatives such as servants, apprentices, boarders and lodgers. Similarly, in nineteenth-century American urban households, Modell and Hareven (1973) found a high proportion of boarders and lodgers, who provided important substitutes for children who had left home. Households expanded and contracted over the life course and in accordance with the family's needs.

Developmental Perspectives: The Life Course

The discovery of boarders and lodgers in nineteenth-century households and of the overall fluidity of household structure led historians of the family to search for a framework that would allow them to capture the movement of individuals through various household forms over their lives and the changes in the composition of the family and the household under various historical conditions. The family cycle model, which was used originally to identify changes in family structure, became unsatisfactory, because the stages of the family cycle were derived from contemporary American middle-class families, and were a priori. As Glen Elder explained, the family cycle identifies stages of parenthood rather than the more dynamic aspects of individual transitions into and out of various family roles (Elder, 1978).

Dissatisfaction with the family cycle model led historians and sociologists to develop the life-course approach, which has introduced a dynamic dimension into the historical study of the family, and has moved analysis from a simplistic examination of stages of the family cycle to individuals' and families' timing of life transitions in relation to historical time (Hareven, 1978; Elder, 1978). According to the life-course paradigm, the pace and definition of the timing of life transitions, such as starting work, leaving home, and getting married, are determined by their social and cultural context. On the familial level, timing involves the synchronization of individual life transitions with collective family ones (Elder, 1978). For example, using both demographic evidence and prescriptive literature, Modell, Funstenberg, and Herschberg found a major historical change in American society over the past century, from the 'erratic' timing of life transitions governed by collective family needs to more individualized timing in conformity with age norms (Modell et al., 1976).

Life-course research also illuminates the links between behavior and perception. While the actual timing of life transitions can be reconstructed from demographic records, its meaning to the individual and family members undergoing these transitions hinges on the examination of qualitative, subjective sources (Neugarten and Hagestadt, 1976;

Bourdelais, 1993; Hareven and Masaoka, 1988; Hareven, 2000b). A life-course perspective has also made an important contribution to the study of kinship by directing attention to the changing configurations of kin with whom individuals 'travel' together over their lives. Such configurations are formed and reformed; they change in their composition, in their relationship to the individual and the nuclear family, and in relation to each other over the life course.

Kinship

During the second decade of the development in the field of family history, kinship ties outside the household received increasing attention. Rather than being isolated, nuclear households were embedded in extended kinship ties outside their confines. Members of the nuclear family were engaged in various forms of mutual assistance, collaboration, and rituals with extended kin. Even though aging parents did not generally reside in the same household with their married adult children, they lived in the vicinity, often on the same land (Greven, 1970; Berkner, 1972; Smith, 1979; Sabean, 1998).

For nineteenth-and twentieth-century urban populations, Anderson (1971) and Hareven (1982), respectively, have documented the central role of family members and more distant kin in organizing migration from rural areas to industrial cities, in facilitating settlement in urban communities, and in helping migrants adapt to new working and living conditions. In her study of late-nineteenth-and-twentieth-century textile workers in Manchester, New Hampshire, Hareven found that kinship networks in communities of origin were reinforced by the back-and-forth migration of individual members, and the transfer of resources. Following 'chain migration' routes, villagers who went to work in industrial communities spearheaded migration for other relatives by locating housing and jobs. Those who remained in the communities of origin often took care of aging parents and other relatives who stayed behind (Anderson, 1971; Hareven, 1982). Hareven found that despite the scale of the modern factory system, workers' kinship networks cushioned the adaptation of immigrant workers to new industrial working conditions. Kin acted as brokers between the workers and their industrial employers. Rather than disrupting kinship ties, migration often strengthened them and led to the development of new functions for kin in response to changing economic and employment conditions in response to the requirements of the industrial system (Hareven, 1982).

The Family's Interaction with the Process of Industrialization

An examination of the family's interaction with the process of social and economic change enables us to understand more precisely what occurred internally within the family and how such changes were accomplished on a societal level. It provides new insights into the process of industrialization and urbanization, into how labor markets functioned, and how industrial work processes and labor relations were organized. Historical research on the family has led to the rejection of simplistic models of social change, especially of modernization theory. How the family both initiates change and adapts to it and how it translates the impact of larger structural changes to its own sphere are crucial questions for the understanding of the process of social change.

The important principle underlying these questions is a view of the family as an active agent. The family planned, initiated, or resisted change; it did not just respond passively. Historical research over the past two decades, highlighting the family's role as an active agent, has led to revising questions such as "What was the impact of industrialization on the family?" to "What was the family's impact on the industrial system?" (Hareven, 1982, cf Goode, 1963). The question has been further cast into a more complex form: Under what circumstances was the family more able to control its destiny and to affect the larger social processes, and under what circumstances did the family succumb to declining markets, changing modes of production, business cycles, and other external forces?

The family had a central role in charting the transitions to various modes of industrial production by following diverse routes: from a rural economy to a protoindustrial or industrial one, or from a rural economy directly to an industrial one. In all these cases, while following its own priorities, when possible, the family facilitated the advance of industrialization by releasing the labor force needed for the newly developing factories and by organizing migration to industrial centers (Smelser, 1959; Tilly and Scott, 1978; Quataert, 1985). The family's interaction with the industrial system has received attention from studies of family production systems in artisanal and protoindustrial settings and the study of family and work in the context of large industrial enterprises.

Protoindustrialization was a household-based form of production in the countryside and in some urban areas in England, France, Belgium, Switzerland, and Austria, from the seventeenth to the nineteenth centuries that preceded the 'industrial revolution.' It was characterized by the production of goods in local cottages for a capitalist employer who sold the products in external markets. Rudolph Braun (who first identified these patterns in Switzerland) and other historians developed a model of demographic behavior they claimed was typical of 'protoindustrial' families. This consisted of earlier age at marriage, higher fertility, and delays in children's departure from the home, all of which were conducive to maximizing a family labor force. Marriage formed the key to the spread of the cottage industry, because it led to the formation of a new nuclear family unit, which was also a 'unit of labor' (Braun, 1974; Mendels, 1972; Medick, 1976).

In the 1980s and 1990s, the demographic model of protoindustrialization has come under criticism. On the one hand, scholars have questioned whether demographic behavior associated with protoindustrialization had been sufficiently widespread to warrant such generalizations. They found that demographic behavior in response to new economic conditions "differed according to the environment (i.e., land ownership, social structure, and inheritance) and the exact nature of the economic change which took place" (Gutmann and Leboutte, 1984: p. 589). Regional variation, interaction

with the local agricultural structure, as well as the type of cottage industry (whether weaving or spinning), made a difference (Mitterauer, 1992).

Like research on the family in protoindustrial regimes, the study of family in the factory system has provided considerable evidence documenting the family's role as an active agent in its interaction with the process of industrialization. Such research effectively undermined the theory of 'social breakdown' that had haunted sociology and social history. According to this notion, migration from rural to urban centers had led to the uprooting of people from their traditional kinship networks. The pressures of industrial work and urban life caused a disintegration of the family unit, and adaptation to industrial life stripped migrants of their traditional culture (Parsons, 1955; Wirth, 1938).

Historians and sociologists alike have challenged the theory of social breakdown, by providing contrary evidence. Neil Smelser, in his pioneer study of the family in the early stages of industrialization in Britain, documented the recruitment of entire family groups as work units in the early textile factories. Fathers contracted for their children, collected their wages, and were responsible for their performance in the factory. Entire families depended on the factory as their employer; the factories, in turn, depended on the recruitment of family groups to maintain a continuous labor supply (Smelser, 1959). Anderson documented a continuity of recruitment in family units and in the vital role of kin among workers in the textile industry in Lancashire, England, at least through the middle of the nineteenth century. The practice of family members working together in the factory was carried over into the United States, as the factory system developed there, and was still present among immigrant workers in the twentieth century. The very success of the early industrial system depended on the recruitment of labor along kinship lines (Anderson, 1971; Hareven, 1982). As mentioned above, Hareven's study of family and work in the Amoskeag Mills in Manchester, New Hampshire, has documented the active role of the family and the kinship group in relation to the industrial corporation. It showed that the family type most 'fit' to interact with the factory system was not an 'isolated' nuclear family, but rather one embedded in extended kinship ties (Hareven, 1982).

The functions that kin fulfilled in their interaction with the factory system were not merely archaic carryovers from rural society but a selective use of premigration patterns of kin assistance patterns in response to the needs dictated by industrial conditions. The encounter of immigrant workers with the modern factory system led neither to the abandonment of premigration family traditions, nor to a rigid adherence to them. Rather, the workers adapted their customs and social organization to the new conditions they confronted. In doing so, they addressed the factory system on its own terms. The family selected those aspects of its premigration culture that were most useful in coping with industrial conditions and adapted them to new needs. These patterns of selectivity called into question the linear view of social change advanced by modernization theory. Family behavior did not modernize at the same pace as workers' conduct in the factory. The family was both a custodian of tradition and an agent of change (Hareven, 1982).

Family Strategies

To better understand the family's role in these processes of social change, it is necessary to examine how families charted their strategies in relation to external opportunities and constraints. Family strategies involved both the decisions individuals or families made and the actual timing of such decisions in relation to opportunities and constraints. Strategies included, at times, calculated trade-offs in order to find employment, achieve solvency, buy a house, facilitate children's education or their occupational advancement, control or facilitate a child's marriage, save for the future and provide for times of illness, old age, and death. Strategies were part of a larger life plan. As people encountered new circumstances, they modified and reshaped their plans and strategies in the context of their own culture and traditions (Hareven, 1990).

As research on family strategies developed further, the concept also underwent important revisions. Under the influence of feminist historians, the question was raised: 'Whose strategies are family strategies?' This question is still part of an important research agenda about the process of decision making within the family, particularly concerning women's participation. Recent research suggests, though, that there may have been a considerable variation along gender lines and among family members in the forging of strategies.

Future Directions

The multidimensional research effort in family history has produced impressive results and identified new lines of investigation. Research still needs to pursue established topics that have not received sufficient attention, including the family's relationship to social space and a more systematic study of the family's relationship to religion, the state, and the legal system.

There is a need for more systematic studies of the 'edges' of family life, such as solitary individuals, orphans and foundlings, and the process of family breakdown through divorce or death and of family reconstitution through remarriage. A second direction is the forging of more systematic linkages between demographic behavior and household structure with internal family dynamics; a closer integration of the study of the household with nonresident kin; a more careful linkage of kinship and household patterns with various processes such as work and migration; and the relationship between demographic and family patterns with cultural dimensions.

In addition to this specific agenda, the larger questions of historical changes in the family over time still need to be explored systematically in several directions: First, if industrialization did not cause the emergence of 'modern' family behavior in Western Europe and the United States, what did cause it? One plausible explanation that has been advanced is the commercial revolution of the sixteenth and seventeenth centuries. This does not explain, however, the existence of nuclear household patterns in the Middle Ages. While Stone (1977), like Aries, dated the emergence of the 'modern family' in the late seventeenth and early eighteenth centuries, Shorter placed it in the late eighteenth, and early nineteenth centuries. Carl Degler (1980) dated the emergence of the American modern family in the late eighteenth or early

nineteenth centuries as well. Following in Aries's footsteps, Stone, Shorter (1976) and Degler have focused on the rise of 'affective individualism' as the major criterion of 'modern' family life. They generally agreed that the 'modern family' is privatized, nuclear, domestic, and based on emotional bonding between husband and wife, and between parents and children. The correlates of its emergence were the weakening influence of extended kin, friends, and neighbors on family ties and an isolation of the family from interaction with the community. There is some disagreement and, at times, lack of clarity about which class initiated these changes. Aries, Stone, and, more implicitly, Degler viewed the bourgeoisie and the gentry as the vanguard, while Shorter assigned a crucial role to peasants and workers.

Related to the emergence of the family as a private, intimate entity is the widely accepted interpretation that this transformation was based on the family's surrender of earlier-held functions of production, welfare, education, and social control to other institutions and its withdrawal into domestic privacy has become one of the standard clichés of family history and sociology. We need to understand the process by which these functions were transferred to other institutions. A systematic exploration of the family's relationship to public agencies and institutions – specifically institutions of education, welfare, and social control, after the transfer of its functions – will help historians escape the trap of viewing the family as an isolated domestic entity.

Historical studies in Europe in the 1990s have emphasized an integrated view of the family both as a private entity and an object of the state. *The History of Private Life*, under the general editorship of Philippe Aries and Georges Duby, which focuses primarily on French society, deserves particular attention. The fourth volume, *From the Fires of Revolution to the Great War*, edited by Michelle Perrot, presents the experiential history of the family in relation to public life. The authors view the family's private life and sexuality as inseparable from the state (Perrot, 1990).

As further research continues to emerge, it becomes necessary to develop a comprehensive model of change in family behavior that does justice to its complexity. The most important dimension still absent from studies of long-term changes in the family involves more systematic distinctions by social classes, rather than generalizing for the entire society on the basis of the middle class. We need a more detailed understanding of the historical process by which other classes adopted middle-class family behavior, if indeed that was the case, and of what class differences have survived such a process. How did patterns of family behavior that first emerged in the middle class transfer to other classes in society and by what processes? Equally important, at the end of that process, had rural and working-class families really adopted middle-class family forms, or did differences persist underneath apparent similarities?

The question about when and how change takes place needs to be raised again. Various strands of change in family history still need to be traced individually to the point of their culmination as visible transformations. The process is rendered more difficult by the fact that not all the strands undergo change at the same pace. As the family is not a monolith, different members within it may initiate change or accept change at different points. For example, women were agents of change in certain areas of family life, such as in the practice of family limitation, while men were innovators in others. Even children were innovators in certain areas. They may have brought literacy into the family, new perceptions of behavior they learned in school or new work habits and new technologies. Thus, when we examine the family's role in the process of social change, we need to differentiate among the respective roles of members within the family, in interacting with these processes. This is an important task for future decades.

See also: Family as Institution; Gender History; Industrialization, Typologies and History of; Kinship in Anthropology; Labor History; Life Course in History; Sociology, Epistemology of; Work, Anthropology of; Work, History of.

Bibliography

Anderson, M.S., 1971. Family Structure in Nineteenth-century Lancashire. Cambridge University Press, Cambridge, UK.

Andorka, R., 1994. The historical demography of a proper Hungarian village: At any in the eighteenth and nineteenth centuries. Journal of Family History 19, 311–331.

Aries, P., 1962. Centuries of Childhood: A Social History of Family Life (R. Baldick, Trans.). Knopf, New York.

Berkner, L.K., 1972. The stem family and the developmental cycle of the peasant household: an eighteenth century Austrian example. American Historical Review 77, 398–418.

Borges, D.E., 1992. The Family in Bahia, Brazil, 1870–1945. Stanford University Press, Stanford, CA.

Botelho, L.A., Ottaway, S.R., 2008. The History of Old Age in England, 1600–1800. Pickering & Chatto, London.

Botelho, L.A., Thane, P., 2001. Women and Ageing in British Society since 1500. Pearson Education, Harlow, England.

Bourdelais, P., 1993. La Nouvelle Age de la vieillesse: Histoire du vieillessment de la Population. O. Jacob, Paris.

Braun, R., 1974. Early industrialization and demographic change in the Canton of Zurich. In: Tilly, C. (Ed.), Historical Studies of Changing Fertility: Quantitative Studies in History. Princeton University Press, Princeton, IN.

Burguiére, A., 1987. The formation of the couple. Journal of Family History 12, 39–56.

Coster, W., 2001. Family and Kinship in England, 1450–1800. Longman, New York.

Coster, W., 2002. Baptism and Spiritual Kinship in Early Modern England. Ashgate, Aldershot, Hants, England.

Degler, C.N., 1980. At Odds: Women and the Family in America from the Revolution to the Present. Oxford University Press, New York.

Demos, J., 1970. A Little Commonwealth: Family Life in Plymouth Colony. Oxford University Press, New York.

Elder Jr., G.H., 1978. Family history and the life course. In: Hareven, T.K. (Ed.), Transitions: The Family and the Life Course in Historical Perspective. Academic Press, New York, pp. 17–64.

Flandrin, J.-L., 1976. Families in Former Times: Kinship, Household, and Sexuality. Cambridge University Press, Cambridge, UK.

Foyster, E.A., 2005. Marital violence: An English Family History, 1660–1875. Cambridge University Press, Cambridge.

Gerber, H., 1989. Anthropology and family history: the Ottoman and Turkish families. Journal of Family History 14, 409–421.

Goode, W., 1963. World Revolution and Family Patterns. Free Press, New York.

Goody, J., 1972. Evolution of the family. In: Laslett, P., Wall, R. (Eds.), Household and Family in Past Time. Cambridge University Press, Cambridge, UK.

Goubert, P., 1977. Family and province: a contribution to the knowledge of family structures in early modern France. Journal of Family History 2, 223–236.

Greven, P., 1970. Four Generations: Population, Land, and Family in Colonial Andover, Massachusetts. Cornell University Press, Ithaca, NY.

Groundwater, A., 2010. The Scottish Middle March, 1573–1625: Power, Kinship, Allegiance. Royal Historical Society, London, England.
Gutmann, M., Leboutte, R., 1984. Rethinking protoindustrialization and family. Journal of Interdisciplinary History 14, 587–607.
Hajnal, J., 1965. European marriage patterns in perspective. In: Glass, D.V., Eversley, D.E.C. (Eds.), Population in History. E Arnold, London.
Halpern, J.M., Kaser, K., Wagner, R.A., 1996. Patriarchy in the Balkans: Temporal and cross-cultural approaches. History of the Family:. An International Quarterly 1, 425–443.
Hareven, T.K., Masaoka, K., 1988. Turning points and transitions: perceptions of the life course. Journal of Family History 13, 271–289.
Hareven, T.K., 1971. The history of the family as an interdisciplinary field. Journal of Interdisciplinary History 2, 273–299.
Hareven, T.K. (Ed.), 1978. Transitions: The Family and the Life Course in Historical Perspective. Academic Press, New York.
Hareven, T.K., 1982. Family Time and Industrial Time: The Relationship Between the Family and Work in a New England Industrial Community. Cambridge University Press, New York.
Hareven, T.K., 1990. A complex relationship: family strategies and the process of economic and social change. In: Friedland, R., Robertson, A.F. (Eds.), Beyond the Marketplace: Rethinking Economy and Society. Aldine de Gruyter, New York.
Hareven, T.K., 2000a. Families, History, and Social Change: Life-Course & Cross-Cultural Perspectives. Westview, Press, Boulder, CO.
Hareven, T.K., 2000b. Formal timing and subjective timing: Historical changes and the cultural construction of the life course. In: Bideau, A., Bourdelais, P., Brunet, G. (Eds.), Structures Par Âge et Âges de la Vie: Deux Visions du Monde? Cahier de Annales de Demographie Historique, Paris.
Hayami, A., Uchida, N., 1972. Size of household in a Japanese country throughout the Tokugawa era. In: Laslett, P., Wall, R. (Eds.), Household and Family in Past Time. Cambridge University Press, Cambridge, UK.
Henry, L., 1953. Une richesse démographique en fiche: Les registres paroissiaux. Population 8, 281–290.
Herlihy, D., 1987. The family and religious ideologies in medieval Europe. Journal of Family History 12, 3–18.
Herman, E., 2008. Kinship by Design: A History of Adoption in the Modern United States. University of Chicago Press, Chicago.
Johnson, P., Thane, P., 1998. Old Age from Antiquity to Postmodernity. Routledge, London.
Kertzer, D.I., Hogan, D., 1989. Family, Political Economy, and Demographic Change: The Transformation of Life in Casalecchi, Italy, 1961–1921. Life Course Studies. University of Wisconsin Press, Madison, WI.
Kingston, J., 2011. Japan in Transformation, 1945–2010. Longman, Harlow, England.
Klein, H.S., 2004. A Population History of the United States. Cambridge University Press, Cambridge, UK.
Kurosu, S., 1996. Leaving home in a stem family system: Departures of heir and nonheirs in preindustrial Japan. History of the Family 1, 329–352.
Kuznesof, E.A., 1997. Who were the families of 'natural' children in nineteenth-century Rio de Janeiro, Brazil? A comparison of baptismal and census records. The History of the Family: An International Quarterly 2, 171–182.
Laslett, P., Wall, R. (Eds.), 1972. Household and Family in Past Time. Cambridge University Press, Cambridge, UK.
Laslett, P., 1965. The World We Have Lost. Methuen, London.
Laslett, P. (Ed.), 1977. Family Life and Illicit Love in Earlier Generations: Essays in Historical Sociology. Cambridge University Press, Cambridge, UK.
Lee, J., Campbell, C., 1997. Fate and Fortune in Rural China: Social Organization and Population Behavior in Lioning 1774–1873. Cambridge University Press, Cambridge, UK.
Lee, J., Feng, W., 1999. One Quarter of Humanity: Malthusian Mythology and Chinese Realities, 1700–2000. Harvard University Press, Cambridge, MA.
LeRoy Ladurie, E., 1978. Montaillou: The Promised Land of Error. Braziller, New York (B. Bray, Trans.).
Majumdar, J.A.P., Burton, A.M., 2003. Family History. Oxford University Press, New Delhi.
Maranda, P., 2006. French Kinship: Structure and History. J.-M. Tremblay, Chicoutimi.
Maynes, M.J., Waltner, A.B., 2012. The family: A World History. Oxford University Press, Oxford.
Medick, H., 1976. The protoindustrial family economy. Social History 3, 291–315.
Mendels, F., 1972. Protoindustrialization: the first phase of the industrialization process. Journal of Economic History 32, 241–261.
Mitterauer, M., 1992. Peasant and nonpeasant family forms in relation to the physical environment and the local economy. Journal of Family History 17, 139–159.
Modell, J., Hareven, T.K., 1973. Urbanization and the malleable household: boarding and lodging in American families. Journal of Marriage and the Family 35, 467–479.
Modell, J., Furstenberg, F., Hershberg, T., 1976. Social change and transitions to adulthood in historical perspective. Journal of Family History 1 (1), 7–32.
Neugarten, B., Hagestadt, G., 1976. Age and the life course. In: Binstock, R., Shanas, E. (Eds.), Handbook of Aging and the Social Sciences. Van Nostrand Reinhold, New York.
Parsons, T., 1955. The American family: its relation to personality and to the social structure. In: Parsons, R., Bales, R. (Eds.), Family, Socialization, and Interaction Process. Free Press, Glencoe, IL.
Perrot, M. (Ed.), 1990. A History of Private Lives. From the Fires of Revolution to the Great War (A. Goldhammer, Trans.). vol. 4. Beklnap Press of Harvard University Press, Cambridge, London.
Plakans, A., Wetherell, C., 1999. Fertility and cultural diversity: an exploratory essay concerning the borderlands problem in the nineteenth century Baltic area. Continuity and Change 1, 33–56.
Quataert, J., 1985. Combining agrarian and industrial livelihood: rural households in the Saxon Oberlausitz in the nineteenth century. Journal of Family History 10, 145–162.
Reher, D.S., 1997. Perspectives on the Family in Spain, Past and Present. Oxford University Press, Oxford, UK.
Rogers, J., 1993. Nordic family history: themes and issues, old and new. Journal of Family History 18, 291–314.
Sabean, D.W., 1998. Kinship in Neckarhausen, 1700–1870. Cambridge University Press, Cambridge, UK.
Segalen, M., 1991. Fifteen Generations of Bretons: Kinship and Society in Lower Brittany, 1720–1980. Cambridge University Press, New York.
Shorter, E., 1976. The Making of the Modern Family. Basic Books, New York.
Smelser, N., 1959. Social Change in the Industrial Revolution: An Application of Theory to the British Cotton Industry. University of Chicago Press, Chicago.
Smith, R., 1979. Kin and neighbors in a thirteenth-century Suffolk community. Journal of Family History 4, 219–256.
Stone, L., 1977. The Family, Sex, and Marriage in England, 500–1800. Harper and Row, New York.
Szonyi, M., 2002. Practicing Kinship: Lineage and Descent in Late Imperial China. Stanford University Press, Stanford, CA.
Tilly, L., Scott, J., 1978. Women, Work, and Family. Holt, Rinehart and Winston, New York.
Tilly, C., 1987. Family history, social history, and social change. Journal of Family History 12, 319–330.
Todorova, M., 1997. Imagining the Balkans. Oxford University Press, New York.
Uhlmann, A.J., 2007. Family, Gender and Kinship in Australia: The Social and Cultural Logic of Practice and Subjectivity. Ashgate Pub, Farnham.
Vinovskis, M., 1977. From household size to the life course: some observations on recent trends in family history. American Behavioral Scientist 21, 263–287.
Wall, R., Hareven, T.K., Ehmer, J., 2001. Family history revisited: Comparative perspectives. University of Delaware Press, Newark, DE.
Winslow, M., 2001. War, Resettlement, Rooting and Ageing: An Oral History Study of Polish Émigrés in Britain.
Wirth, L., 1938. Urbanism as a way of life. American Journal of Sociology 44, 1–24.
Worby, S., 2010. Law and Kinship in Thirteenth-century England. Royal Historical Society/Boydell Press, Woodbridge, Suffolk, UK.
Wrigley, E., Schofield, R.S., 1981. The Population History of England, 1541–1871. Harvard University Press, Cambridge, MA.
Wrigley, E., 1972. The process of modernization and the industrial revolution in England. Journal of Interdisciplinary History 3, 225–260.

Family and Psychiatric Illness: Family Therapy

Judith Lask, Kings College London, London, UK; and United Kingdom Council for Psychotherapy, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Leff, volume 8, pp. 5298–5303, © 2001, Elsevier Ltd.

Abstract

The practice of Family Therapy and its theoretical underpinnings have been evolving for more than 60 years. Family Therapy, Family Interventions, and Systemic Family Practice are in many parts of the world accepted as key components of mental health services. There is increasing evidence to support the value of systemic approaches and an increasing commitment from practitioners to develop and work from a sound evidence base. This article gives a brief overview of Family Therapy and other family interventions, and the theories informing their practice. Evidence for efficacy is examined and current applications are discussed.

The practice of Family Therapy and its theoretical underpinnings have been evolving for more than 60 years and Family Therapy, Family Interventions, and Systemic Family Practice are in many parts of the world accepted as key components of mental health services. There is increasing evidence to support the value of systemic approaches and an increasing commitment from practitioners to develop and work from a sound evidence base (Eisler and Lask, 2009; Carr, 2014a,b). This article provides a brief overview of Family Therapy and other family interventions, and the theories informing their practice. The evidence for their efficacy is examined and current applications are discussed.

The beginnings of Family Therapy can be found in the fields of both adult and child mental health. In the 1940s and 1950s there was an interest in the way in which communication within families might give rise to schizophrenia in a family member (Singer and Wynne, 1966). Although this idea has been shown to be erroneous, the focus on patterns of communication, including 'the double bind theory,' has been influential and added to our understanding of the way in which families can have profound effects on the mental health of individuals. Unfortunately the idea that there is a direct (often one way) relationship between family dysfunction and individual problems persisted for some time. However over the last 25 years there has been a growing awareness that this explanation is a serious over-simplification of a very complex process. In current practice, families are seen as resources and working partners rather than the main cause of problems. Genes, personality, life experience, social and economic environments, and health together with families and other relationships are joint contributors to both psychiatric illness and well being.

Work in the areas of child development, attachment, and adult mental health has highlighted the importance of family and other relationships in the development of good mental health (Rutter, 2006). Conversely relationship dysfunction can give rise, and contribute, to mental health problems such as depression, anxiety, attachment disorders, and phobias. In Child and Adolescent Mental Health there is unsurprisingly, a greater awareness of the importance of families, carers, and other systems around the child and the need to include them in any treatment plan. In the 1960s, Bowlby, among others, began to include parents in treatment. By the 1980s the profession of Family Therapist was developing in a number of countries and appointments of Family Therapists were being made in Child and Adolescent Mental Health Services. Family Therapy has now become established as one of the main evidence based treatments available to children, young people, and their families. In adult mental health the developments have been more uneven. This is understandable when the role of the family is less clear; however, the value of Family Therapy and Family Interventions has proved important in supporting families and engaging them in helping a family member with mental health problems. Systemic approaches are also helpful in working with complex professional systems and other care groups such as hostels and day centers.

What Is Family Therapy?

This is a more complicated question than it seems. Although the idea of the system being important is foundational, family therapy is broad based and contains many approaches to thinking about and working with families and other systems. Some approaches focus on working with patterns of behavior, others with beliefs or emotional attachments, and yet others with trans-generational influences. Some psychiatric problems are more likely to respond to one approach rather than another but there is also an emphasis on adapting the approach to best fit a particular family. Value is placed on engagement and collaborative working, both of which are increasingly accepted as key components of successful and ethical psychological treatments. Family Therapists also have a broad definition of the family and will pragmatically work with the most important people around a referred child or adult. This may involve seeing the whole family together, parts of the family (sub-systems), or the family without the referred patient. Decisions are made on the basis of who is available and the agreed focus of the work. Sometimes interventions may be made with other systems such as professional systems or friendship networks or, in Multi-Family Therapy, a number of families may be seen together (Asen and Scholz, 2010). Fundamental to Systemic Family Therapy is the idea that we are all affected by relationships with others and if the systems

around us work well and support recovery this will make a positive difference to outcome, The next section will contain a broad overview of the foundational ideas of current practice followed by a discussion of the evidence base of some of these approaches and an outline about how they may be used in practice.

Theoretical Overview

Systemic Family Therapy is concerned with the functioning of families and other systems and different approaches focus on particular aspects of systems. Early approaches were influenced by cybernetics and the idea of self-regulating systems. The focus was on mechanisms that moderated the dual aspects of change and stability. In this way problems were often seen as attempts to keep a system stable. For example, a young person with eating disorder may keep a family together by requiring family members to focus on an ill child because of fear that the family may break apart once child-rearing is accomplished. In this sense the problem is seen to perform a function for the system. There was also a focus on patterns of communication – for example, patterns which lead to escalation but do not resolve the problem (Bateson, 1972). Elements of both of these approaches are to be found in current systemic therapies but no longer is each seen as the 'answer' but one part of a complex jigsaw of cause, influence, and effect.

Provision of a context that allows for both change and stability is important as families help members to adapt to developmental and contextual changes. Family Life Cycle Theory (Carter and McGoldrick, 2004) influences all major approaches in Family Therapy. It is a way of encapsulating the interaction between family, wider contexts, and the developmental needs of individuals. It outlines the way in which all families have to adapt and change over time. Unexpected events such as untimely death, forced migration, or unemployment will impose extra demands. When the demands for change are too great or there are conflicting demands (such as need for adolescent independence when there are serious external threats) individuals may develop problems such as depression, anxiety, or substance abuse. Flexibility and adaptability are important aspects of individual resilience and they are equally important in families. Thus, those who are able to adapt their roles and responsibilities, beliefs, and behavior to fit new circumstances will be more resilient to adversity (Rutter, 1999).

The idea of circularity is another key idea in systemic work. This refers to the way in which patterns of behavior can become repetitive circular processes. The following example is influenced by attachment ideas which have been an important recent influence on Family Therapy. An anxious husband seeks reassurance from his wife who is preoccupied with work, which leads his wife to become irritated and to resist his demands which in turn makes him more anxious and more 'needy' which, in turn, is more intolerable for his wife who has a need to feel independent. This may be an oft-repeated pattern that can be described as 'stuck' and does not lead to any resolution or change. Meanwhile the husband's anxiety symptoms may lead to a psychiatric referral. Family patterns passed down from one generation to another may also be a factor in maintaining these unhelpful interactions.

Another fundamental idea is that what happens to one part of the system inevitably affects other parts of the system. In this way we are all impacted by what happens to those with whom we are in relationship and what we do affects them. Family therapy in child psychiatry began to develop because of observations that children in individual therapy were unable to move on because of the way their family behaved toward them. Approaches at this time were more influenced by psychodynamic ideas but over time these ideas have been used to a lesser degree in Family Therapy practice.

The 1960s saw the development of structural Family Therapy which focuses on the importance of clear organizational structures, roles, and responsibilities in families. This approach suggested that problems were most likely to develop when the structure of the family was unclear, relationships were either too distant or too enmeshed, or children were parentified. Sometimes individuals become 'triangulated' or caught up in the problems between others (perhaps a child caught up in parental conflict). As with family life cycle theory, criticisms were made that it presented a very narrow, white middle class view of family life. However the ideas are still very useful as long as account is taken of family differences including class and culture. Interventions are often carried out within family sessions with the aim of helping families to relate in more positive ways. For example where an adult daughter is caught up in the conflictual relationship between her parents, the therapist may encourage the parents to have a discussion while the therapist ensures that the daughter does not intervene. This allows them all to have the experience of communicating in a different way.

The same period saw the development of Strategic Family Therapy that focuses on stuck behavior patterns, natural resistance to change, and the need to devise creative solutions to encourage people to make those changes. It assumes that insight is not always possible and does not inevitably lead to change. In this approach between sessions family tasks are often set. The focus is change in repetitive behavior patterns. For example a therapist may ask a couple to 'put off' their daily arguments until after the evening meal and then to sit down and argue for 15 min. The idea is that if an argument is taken out of its usual context it may have a different flavor, some of the high emotion will be decreased and the instruction to do so may lead to some reflection and even rebellion! As time progressed concern grew about the possibility of this approach being manipulative but the ideas can be used in an open and transparent way with due care and attention to ethical practice.

Another approach that focuses on behavior change is Family Interventions in Psychosis (Lam, 1991). This has been shown to be an effective approach in reducing relapse rate for those who have developed a psychotic illness and focuses on psycho-education and reduction of high emotional intensity, especially negative and critical comments. This is a good example of the therapist taking a collaborative position and working with the family as a team in order to help a family member with serious mental health difficulties. In the early days of Family Therapy there tended to be a focus on the family as the cause for schizophrenia and other psychoses. Patients were seen as being caught in a 'double bind.' The double bind

was defined carefully as a contradiction within the verbal content of a message, or between the words used and the feelings expressed. In addition, the recipient of the message is not allowed to comment on the contradiction or leave the relationship (the doubling of the bind) (Singer and Wynne, 1966). At this point little account was taken of the effect on families of living with a member with serious psychiatric illness. Studies into the consequences of the discharge of chronic psychiatric patients into the community found that relapse rates were higher when the discharge was to close relatives (Leff and Vaughn, 1985). This led to the study of high emotional intensity as an important factor in relapse (Butzlaff and Hooley, 1998). National organizations were set up by relatives of patients – such as the National Schizophrenia Fellowship in the UK and the National Alliance for the Mentally Ill in the US. This further raised the profile of families who were often judged negatively or marginalized by mental health services. Research in the 1970s and 1980s showed that a program of psycho-education help, with reducing negative comments and increasing problem solving, reduced relapse from 50% to 20%.

During the 1970s and 1980s some important developments took place in Milan in Italy where the approach called Milan Systemic Therapy was developed (Boscolo, 1987). This approach highlighted the importance of beliefs as determinates of behavior, relationships, and emotions. The context for the development of these ideas was closure of large psychiatric hospitals and the need to treat people as out patients and involve their families. This approach developed questioning techniques (Fleurides et al., 1986) and encouraged the therapist to retain curiosity as they explored the connections between problems and relationships. Techniques included Socratic questioning similar to that utilized in cognitive behavior therapy but also included 'circular questions' which explored the connections between relationships, beliefs, and behavior. These questions were used to gather information about how the system worked, explore hypotheses, and also act as an intervention to help family members become observers of themselves and explore new ways of seeing things. This approach also gave rise to developments in team working and the use of a one-way screen through which team members observed the therapy session and had mid-way discussions with the therapist in order to share ideas and present alternative viewpoints. The next developments often called 'Post Milan' took a postmodern position and rejected simplistic cause and effect. More account was taken of the value of open dialogue in therapy, the exploration of meanings, and the evolution of new ideas. Therefore, for example, a depressed family member may come to be seen as bravely struggling against depression rather than as a failure. This will alter the way other family members relate and behave. This 'reframing' is a key intervention in a number of approaches.

Narrative approaches to therapy developed from the 1980s and focused on the way that all of us develop 'narratives' to make sense of the world and emphasizes the way in which those narratives shape the way we live our lives (Morgan, 2000). There is a particular focus on the way in which wider systems such as community, culture, professional systems, education systems, media, and so on contribute to, and shape those narratives. The role of the therapist is to help individuals, couples, and families look at alternative narratives which better fit their lives and help them to move forward. The process of understanding how narratives are formed challenges their truth and opens them up to examination. The therapist becomes a witness to new developments in the patient's life and celebrates and highlights positive changes, emphasizing the patient's personal agency in making the changes. An intervention from this approach, widely used with people with psychiatric illness, is 'externalization.' The illness is separated from the person by speaking of it as an external entity. Questions will be asked about how the 'anorexia' affects everyone's lives and there follows an exploration of the ways in which it has been possible to control 'anorexia.' The aim is to identify 'unique outcomes' which serve to question the assumption that the person is totally inhabited by the illness and is helpless. These unique outcomes are developed in order to generate useful tools to help reduce the power of 'anorexia.' A similar approach, used on its own or as part of an integrative approach, is Brief Solution Focused Therapy (De Shazer, 1985) which identifies and helps reinforce a person's own preferred ways of fighting symptoms. For example, someone might state that if their partner smiles at them in the morning it 'gets the day off to a better start.' The partner may be asked to smile even though they do not feel like it. Embedded within many of these approaches is the idea that a small change can have a beneficial ripple effect throughout the family.

These 'postmodern' approaches have led to therapy teams being used in different ways. They can be used in narrative therapy as groups of witnesses commenting on new developments in someone's life or perhaps as reflecting team who come into the therapy room and share different perspectives on the issues being discussed.

Open dialogue approaches in the treatment of serious mental illness (Seikkula, 2006) is a recent development. This approach involves family members and members of the professional network who meet together to talk. The assumption is that solutions will emerge out of the unstructured conversations. This has proved to be helpful at times of crisis and enables everyone including the patient to be heard and for new ways forward to be decided.

More recently Family Therapy has developed a greater interest in emotions and the way they are connected to beliefs, relationships, and behavior. Emotionally Focused Couple Therapy is an example of such an approach. It combines attachment ideas, communication coaching, and trans-generational ideas with systems theory (Johnson, 2004). Narrative Attachment Therapy developed by Dallos and Vetere has a similar focus.

Values and Ethics in Current Practice

Systemic approaches emphasize the importance of context in determining the way we live our lives and the practice of Family Therapy has of course also been influenced by the societal and professional contexts. The focus currently is on creating a strong therapeutic alliance (Escudero et al., 2008), working collaboratively with family members and other clients, and respecting individual aspirations. There is a strong focus on respect for, and taking account of difference in relation to gender, culture,

religion, ability, and other dimensions of difference (Flaskas et al., 2005). There is also a strong emphasis on working from a strength and resilience perspective and appreciating the capacity of family members to overcome most difficulties as they go through life. There is more humility about the fact that no approach works effectively with everyone. Systemic Psychotherapists work hard at understanding their own issues and experience and the way in which these affect their work. Experience can be a resource but at times may constrain the work and, for example, lead to an avoidance of certain questions or a fear of upsetting a particular family member. Many systemic psychotherapists do have their own therapy but work on the interface between personal and professional is usually done in groups as part of their course and not termed 'therapy.'

When a number of individuals have to be taken into account, the ethics of practice become more complex especially when there are power differentials and conflicting needs. There is a clear remit to prioritize those who are most vulnerable, and the safeguarding of children is a priority. Most people have a commitment to integrative working although some will focus on one particular approach. There is a clear requirement for Family Therapists to work within the boundary of their own training and skill and to consider other approaches and treatments which would be helpful as an alternative or adjunct to their work.

What Is the Evidence for Systemically Based Psychological Therapies?

Especially where psychological therapies are delivered in public sector services it is important to gain as much information as possible about their efficacy and more specifically which aspects of therapy are most useful and in what circumstances. Health economics assessments of approaches are now a common requirement. In a series of US studies it was shown that the medical cost offset associated with Family and Couple Therapy at least covered the cost of the therapy and sometimes exceeded it (Crane and Christenson, 2012). The links between childhood psychiatric disorder and adult disorder are well documented so that although this section is divided into adult and child, it is clear that this is more a reflection of the organization of services than the reality of life. Research into work with families is particularly complex because the focus is rarely exclusively on the referred patient. Measures into family functioning are not entirely satisfactory and can be cumbersome. The development of the SCORE measure of family functioning provides a new, easy to use, clinically based measure which will be of assistance in much needed research into Systemic Family Therapy especially in Child and Adolescent Mental Health (Stratton et al., 2014).

Depression and Anxiety

The incidence of depression and anxiety is high with almost a fifth of adults in the UK reporting signs of depression and anxiety. The condition is highly distressing for the person suffering from depression and also for other family members, including any children, In the UK the incidence of anxiety and depression in children is on the rise. Although current research indicates that Cognitive Behavior Therapy (CBT) is a preferred first line approach this does not work, or does not work sufficiently for everyone. Narrative reviews of controlled trials of systemic interventions for depression support the effectiveness of outpatient and inpatient systemic couples therapy, Family Therapy based on the McMaster Model, Emotionally Focused Couples Therapy, Behavioral Couple Therapy, Cognitive Marital Therapy, and Conjoint Interpersonal Therapy (Barbato and D'Avanzo, 2008). All of these approaches to Couples Therapy require fewer than 20 sessions and focus on relationship enhancement, mood management, stress reduction, and damage limitation for other family members.

Studies by Leff et al. (2000) and Lemmens et al. (2009) show that Systemic Couple Therapy is more effective than medication at 2-year follow up. Lemmens et al. utilized a mix of multi-family groups, couple sessions, and family sessions and found that the outcomes were much better than medication alone.

Early on in the history of Family Therapy there was an interest in the origins of psychosis and the possible ways family interactions could exacerbate problems. While it is clear that there is no straightforward causal link between psychosis and family interaction, work on family interventions in psychosis has shown the value of an approach which combines psycho education with lowering the levels of expressed emotion. Pfammatter et al. (2006) carried out a review of three meta-analyses of psycho-educational Family Therapy together with a new meta-analysis of the 30 best trials. They found that multimodal programs including both psycho-educational Family Therapy and medication led to lower relapse rates and less hospitalization.

Family based therapies have been shown to be effective with agoraphobia and panic disorder, and with OCD. Systemic interventions create a context in which families can support recovery and a forum within which family interaction patterns maintaining symptoms can be addressed (Renshaw et al., 2005).

Behavioral Couples Therapy is an established approach in working with people with alcohol problems. Typically, treatment includes a sobriety contract, help with problem solving, communication training, and relationship enhancement work (O'Farrell and Clements, 2012).

As the focus of this is psychiatric illness it is worth mentioning the role of Couple Therapy, Family Therapy, and mediation in the process of separation and divorce. The relationship stress associated with the process of divorce and separation and the risks for the mental health of children and future adults makes it important to consider the value of systemic approaches which not only work with the couple but also with the children. They will keep the needs of the children paramount and help parents to collaborate in creating the best outcomes despite their own strong emotions.

It is not surprising that family approaches have a significant part to play in the treatment of child and adolescent psychiatric problems. Although the differences in many studies are small the evidence for family interventions is growing, with a number of recent well designed and large scale studies. A number of recent reviews provide a general overview of the evidence for the field as a whole (Carr, 2014a,b) as well as for problem specific areas such as alcohol and drug abuse

(Hogue and Liddle, 2009), conduct disorder and delinquency, eating disorders, anxiety disorders, and mood disorders (Cosgrove and Chang, 2013).

With most problems the differences between Family Therapy and other active treatments are relatively small but with some there is consistent evidence of Family Therapy leading to better outcomes. This is particularly the case for behavior problems, youth offending (Sydow et al., 2013), and substance misuse (Tanner-Smith et al., 2013) where the evidence from around 30 RCTs provides clear support for the use of Family Therapy both in terms of outcome and engagement in other parts of the therapeutic program (Stanton, 2004).

Similarly, there is now compelling evidence for Family Therapy in the treatment of eating disorders (Couturier et al., 2013) with recent larger studies confirming its superiority in the treatment of anorexia nervosa in comparison with an adolescent focused individual therapy as well as its role in the treatment of bulimia nervosa.

For other problems such as anxiety disorders (including OCD), depression, or bipolar disorder there is evidence that family interventions are effective but the comparison with other treatments is less straightforward. Some studies have shown that where the parents themselves are anxious the family intervention may be more effective and there are data suggesting that even when there is no difference between changes in anxiety symptoms, the family intervention leads to greater changes in general functioning both at the individual and family levels (Khanna and Kendall, 2009). Two groups have evaluated Systemic Family Therapy for depression in children and adolescents and shown them to be effective in comparison with control (Diamond et al., 2002) and comparable to psychodynamic psychotherapy (Trowell et al., 2007).

Clinical Practice

Although there is good evidence that Family Therapy can be effective with a range of problems there has to be caution in translating research findings into practice. They can indicate useful ways to proceed but individual family factors have to be taken into account and culture, gender, and age can profoundly affect work with families. As previously explained, systemic work does not always have to include the whole family and there are times when this is contraindicated. Where there are safeguarding issues it is important to see a child or a young person on their own. High degrees of conflict may indicate that it is better to begin with some individual sessions and then move to family sessions. Sometimes particular family members do not want to be involved. A number of approaches to working with client groups have been manualized but these manuals are usually less specific than for some therapies and allow room for adaptation. For example, in work with conduct disorder in children the engagement process is particularly important and may take a number of sessions. Family work in this group tends to focus on family roles, clear communication, lines of authority, and problem solving. There is also an emphasis on embedding new learning in everyday life. Work with depression and self-harm in children and young people will focus on encouraging safeguarding and enabling communication of emotions as well as exploring attachment patterns. Work with eating disorders will begin by prioritizing refeeding and getting to a safe weight. It will involve 'externalizing' and helping parents to feel more skilled with their young person. Work with psychosis will usually involve some psycho-education and emotionally focused couple work for depression will focus on managing strong emotions and meeting each other's attachment needs.

The number of sessions is sometimes prescribed (often between 6 and 12) but some pieces of work go on for longer and there may be variable lengths of time between sessions. Engagement of all family members is always a crucial aspect of the work and very often the completion of a genogram will help to track information about the family and identify patterns over time. Family Therapy may be a treatment on its own or an adjunct to other treatments and can be useful in engaging clients and encouraging compliance with other approaches. It is a method that over many years has given to client feedback and collaborative working to identify what each family and family member finds helpful. There are a range of nonverbal interventions such as genograms, drawing, and timelines. When small children are in sessions play and art materials are used to ensure the child is involved in a session and able to communicate in a way that is age appropriate. Experiential techniques such as family sculpting or role-play may be used. Between session tasks, diaries and research tasks are also commonly utilized.

All psychiatric illness affects others and can be adversely and positively affected by those around the ill person. Family Therapy has a key role in helping recovery and lessening the risk of relapse, as also in limiting potential damage to other members of the family.

See also: Anxiety and Anxiety Disorders; Behavioral Problems, Effects of Parenting and Family Structure on; Bowlby, John (1907–90); Child Protection; Child and Adolescent Psychiatry, Principles of; Children and Families in Social Work; Cultural Views of Life Phases; Depression; Eating Disorders; Emotional Development, Effects of Parenting and Family Structure on; Evidence-Based Psychiatry; Family as Institution; Obesity: Cultural and Biological Factors; Repartnering and Stepchildren; Schizophrenia; Social and Emotional Development in the Context of the Family; Stress in Infancy and Early Childhood: Effects on Development.

Bibliography

Asen, E., Scholz, M., 2010. Multi Family Therapy: Concepts and Techniques. Wiley Blackwell, Oxford.

Barbato, A., D'Avanzo, B., 2008. Efficacy of couple therapy as a treatment for depression: a meta-analysis. Psychiatric Quarterly 79, 121–132.

Bateson, G., 1972. Steps to an Ecology of Mind. University of Chicago Press, Illinois.

Boscolo, L., 1987. Milan System Family Therapy: Conversations in Theory and Practice. Basic Books, New York.

Butzlaff, R.L., Hooley, J.M., 1998. Expressed emotion and psychiatric relapse: a meta-analysis. Archives of General Psychiatry 55, 547–552.

Carr, A., 2014a. The evidence base for couple therapy, family therapy and systemic interventions for adult-focused problems. Journal of Family Therapy 36, 158–194.

Carr, A., 2014b. The evidence base for family therapy and systemic interventions for child-focused problems. Journal of Family Therapy 36, 107–157.

Carter, B., McGoldrick, M. (Eds.), 2004. The Expanded Family Life Cycle: Individual, Family and Social Perspectives, third ed. Allyn and Bacon, New York.

Crane, D.R., Christenson, J.D., 2012. A summary report of the cost-effectiveness of the profession and practice of marriage and family therapy. Contemporary Family Therapy 34, 204–216.

Cosgrove, V.E., Chang, R.D., 2013. Bipolar depression in pediatric populations: epidemiology and management. Paediatric Drugs 2115, 83–91.

Couturier, J., Kimber, M., Szatmari, P., 2013. Efficacy of family-based treatment for adolescents with eating disorders: a systematic review and meta-analysis. International Journal of Eating Disorders 46, 3–11.

De Shazer, S., 1985. Keys to Solutions in Brief Therapy. Norton, New York.

Diamond, G.S., Diamond, G.M., Levy, S.A., 2002. Attachment-Based Family Therapy for Depressed Adolescents. APA, Washington.

Eisler, I., Lask, J., 2009. Family interviewing and family therapy. In: Rutter, M., Bishop, D., Pine, D., Scott, S., Stevenson, J., Taylor, E.A., Thapar, A. (Eds.), Rutter's Child and Adolescent Psychiatry, fifth ed. Blackwell Publishing Ltd, Oxford, UK.

Escudero, V., Friedlander, M.L., Varela, N., Abascal, A., 2008. Observing the therapeutic alliance in family therapy: associations with participants' perceptions and therapeutic outcomes. Journal of Family Therapy 30, 194–214.

Flaskas, C., Mason, B., Perlesz, A. (Eds.), 2005. The Space between Experience, Context and Process in the Therapeutic Relationship. Karnac, London.

Fleurides, C., Nelson, T., Rosenthal, D., 1986. The Evolution of Circular Questions: Training Family Therapists, vol. 12, pp. 113–127.

Hogue, A., Liddle, H.A., 2009. Family-based treatment for adolescent substance abuse: controlled trials and new horizons in services research. Journal of family therapy 31, 126–154.

Johnson, S., 2004. The Practice of Emotionally Focused Couple Therapy. Brunner-Routledge, New York.

Lam, D.H., 1991. Psycho-social family intervention in schizophrenia: a review of empirical studies. Psychological Medicine 21, 423–441.

Leff, J., Vaughn, C., 1985. Expressed Emotion in Families: Its Significance for Mental Illness. Guilford, New York.

Leff, J., Vearnals, S., Brewin, C.R., Wolff, G., Alexander, B., Asen, E., Dayson, D., Jones, E., Chisholm, D., Everitt, B., 2000. The London intervention trial: an RCT of antidepressants versus couple therapy in the treatment and maintenance of depressed people with a partner: clinical outcome and costs. British Journal of Psychiatry 177, 95–100.

Lemmens, G.M., Oenema, A., Klepp, K.L., 2009. Therapeutic factors in systemic multi-family groups for major depression: patients' and partners' perspectives. Journal of Family Therapy 31, 250–269.

Morgan, A., 2000. What Is Narrative Therapy: An Easy to Read Introduction. Dulwich Centre Press, Adelaide Aus.

O'Farrell, T., Clements, K., 2012. Review of outcome research on marital and family therapy in the treatment of alcoholism. Journal of Marital and Family Therapy 38, 122–144.

Pfammatter, M., Junghan, U., Brenner, H., 2006. Efficacy of psychological therapy in schizophrenia: conclusions from meta-analyses. Schizophrenia Bulletin 32 (Suppl. 11), S64–S80.

Renshaw, K., Steketee, G., Chambless, D., 2005. Involving family members in the treatment of OCD. Cognitive Behaviour Therapy 34, 164–175.

Rutter, M., 1999. Resilience concepts and findings; implications for family therapy. Journal of Family Therapy 21, 179–194.

Rutter, M., 2006. The promotion of resilience in the face of adversity. In: Clarke-Stewart, A., Dunn, J. (Eds.), Families Count: Effects on Child and Adolescent Development. Cambridge University Press, Cambridge.

Seikkula, J., 2006. Dialogical Meetings in Social Networks. Karnac, london.

Singer, M.T., Wynne, L.C., 1966. Communication styles in parents of normals, neurotics and schizophrenics. Psychiatry Research Reports 20, 25–38.

Stratton, P., Lask, J., Bland, J., Nowotny, E., Evans, C., Singh, R., Janes, E., Peppiatt, A., 2014. Detecting therapeutic improvement early in therapy: validation of the SCORE-15 index of family functioning and change. Journal of Family Therapy 36, 3–19.

Sydow, K., et al., 2013. The efficacy of systemic therapy for childhood and adolescent externalizing disorders: a systematic review of 48 RCTs. Family Process 52, 576–661.

Family and Schooling

Thomas Kellaghan, St Patrick's College, Dublin, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Discontinuities between home and school experienced by children, following which research designs to study relationships between family and schooling are described. The main conclusions of empirical studies of relationships between family socioeconomic status (the most frequently used measure of family circumstances) and children's school learning are presented. Family characteristics that affect student achievement are conceptualized in terms of economic, cultural, and social capital. Home support initiatives relating to health and nutrition and home–school partnerships for families characterized by high rates of educational failure, and the response of schools to problems associated with socioeconomic disadvantage, are described.

There is an obvious difficulty in considering relationships between families and other institutions at a time when families across the world are changing rapidly and radically in their structure, functions, and internal dynamics. While some investigations have addressed aspects of those changes, much research on family–school relationships does not take account of them. Perhaps this research is more properly regarded as providing information about homes, which can encompass a variety of family types, than about families.

Up to relatively recently, the family took major responsibility for the care, education, and training of children. In it, children acquired skills to deal with day-to-day tasks, including those of work. During the Industrial Revolution, as work moved into factories and became more specialized and skilled, and knowledge accumulated, families could no longer provide children with required knowledge and skills, and the need for an institution dedicated to formal learning arose. Increases in migration and urbanization served to underline the need, and so in the nineteenth century, nations came to accept education and training in school as their responsibility.

Discontinuities between Home and School

While one might expect families and schools to play complementary and mutually reinforcing roles in children's education, they may not always do so because of differences in their goals, structure, demands, activities, values, and the competencies they foster. These differences will be experienced to a greater or lesser extent by all children as they move from home to school: from life in a small group to life in large groups; from an environment marked by informality and freedom in the use of time and space to one in which children are allocated to classrooms and seats and work to a daily schedule; from an environment in which relationships are intimate, continuous, and particularistic to one in which they are less familiar, more transitory, and universalistic; from an environment in which learning occurs in a natural and realistic context as the need arises to one in which learning is formal, decontextualized, deliberate, and conscious, taking place day in, day out, whether or not it seems to have any particular purpose (Kellaghan et al., 1993).

Most children will in time adjust to the demands of the classroom. The ease with which they do will probably be, at least to some extent, a function of the compatibility of practices in the developmental contexts of homes and schools. When differences between homes and schools are large, children may experience a serious dysfunction in socialization when they go to school, which will affect their ability to adapt and to learn. This may occur in the case of children living in lower socioeconomic status (SES) homes; children living in developing countries who attend Western-type schools; ethnic or language minority children; and children of immigrant families. Most home support initiatives are designed to help children in these situations bridge the gap between home and school.

Research Designs to Study Family and Schooling

Research on the family factors that influence children's scholastic achievement (knowledge, skills, competencies, usually assessed by standardized tests) and attainment (level reached in the education system, operationalized by qualifications obtained, highest grade completed, or number of years in the education system), much of it in the context of policies of providing equality of educational opportunity, has been carried out, mostly in industrialized countries, since 1900 in several disciplines: education, psychology, sociology, economics, and behavior genetics. A variety of designs have been used. A great many studies used cross-sectional surveys in which data on student achievement were correlated with data on families obtained at the same time. Some studies used a longitudinal design, collecting data at varying points in children's life cycles. Other investigations capitalized on the conditions provided in 'natural experiments' to examine the development of sibling pairs with varying degrees of genetic relatedness (monozygotic twins, dizygotic twins, full siblings) in various types of family.

The limitations of correlational studies in disentangling the effects of variables that are intercorrelated to make inferences about causality are widely recognized. Even when efforts are made in analysis to control for the influence of variables considered to be relevant, there is always the possibility that investigators will have failed to obtain information on key variables.

Alternative methodologies to the 'objective' quantitative approach of empirical studies adopted the view that since acts imply purposeful constructions by actors, they can only be understood from the actor's point of view. In this context, ethnographic and case study procedures have been used to explore individuals' interpretations of family–school relationships. The approach is supported by the findings of psychological research that perceptions of family supports are better predictors of personal adjustment than actual supportive behaviors provided by network members.

Family SES and Children's School Learning

In a great majority of studies of the relationship between family variables and children's school performance, a measure of a family's SES (or social class) was used to represent family influence. SES is a multidimensional construct, and studies have varied in the weight they ascribed to one or more of its three main indicators (parental income, parental level of education, and parental occupation) or to other aspects of the students' environment, such as family and community resources, which have received increasing attention in recent studies.

In modeling the relationships between SES and its indicators, SES is commonly considered to be a latent variable in the formative sense since it is determined by its indicators (which have no necessary relationships with each other) (see Edwards and Bagozzi, 2000). This contrasts with a reflective relationship in which variation among indicators is assumed to be determined by the latent variable (as is the case in classical test theory). Very different causal structures are hypothesized in the two models with the consequence that choice of model has implications for the content validity of a construct and the representation of structural relations between it and its components.

Not all studies endorse the view that the relationship between SES and its indicators is formative. For example, SES is treated as a latent variable with reflective indicators in the Programme for International Student Achievement (PISA) (OECD, 2010). It may be that a mixed model is required to describe the relationship. Thus, causality/determination would be perceived to flow from first-order constructs (income, educational level, occupation) to indicators (items in questionnaires or other indices) in a reflective fashion and from the constructs to a second-order construct (SES) in a formative fashion (as indicated by the arrows in **Figure 1**).

The following eight conclusions can be drawn from studies of the relationship between SES and student learning (Bornstein and Bradley, 2010; Sirin, 2005; White, 1982). First, SES is positively, but not very strongly, related to children's performance on measures of scholastic ability and achievement. Children from high SES homes tend to perform better than children from low SES homes. Second, the magnitude of the relationship between SES and student achievement is contingent on several methodological factors in studies: (1) the way in which SES was measured; (2) the unit of analysis for socioeconomic data (aggregated data produce greater effect sizes); (3) the source of data (relationships are stronger when parents provide information than when students provide it); (4) the achievement measure (single subject achievement measures, such as measures of mathematics or science achievement, exhibit stronger relationships than general achievement measures, such as grade point average or a composite achievement test). Third, differences in cognitive skills between children from varying socioeconomic backgrounds are in evidence before children enter school.

Fourth, the length of time a student spends in the education system (attainment) is a function of his or her family SES. This is so even for students of the same level of general scholastic ability. Fifth, when secondary school programs are differentiated, major selection in terms of socioeconomic background occurs at the point of differentiation. Students from high socioeconomic backgrounds tend to enter an academic type of program that will lead to third-level education, while children from low socioeconomic backgrounds tend to enter technical, vocational, or short-cycle programs. The effects of SES are stronger at earlier than at later transitions.

Sixth, SES is a stronger predictor of achievement of white children than of minority students in the United States. It has been suggested that this may be because neighborhood and school SES exert a stronger influence than family SES on student achievement in minority (especially African American) communities. Seventh, there appears to have been a decline in the strength of the relationship between SES and student achievement over the last three decades of the twentieth century. This may be due to improved social and economic conditions, an increased number of minority and urban students in more recent studies, or the use of more sensitive measures of social and economic circumstances.

Finally, international studies reveal large differences between education systems in the strength of the association between SES and student attainment and achievement. For example, Sieben and De Graaf (2003) found significant differences between 11 countries in family effects on educational attainment. In PISA 2009, the percentage of variance in the reading literacy scores of 15-year-olds explained by a measure of student socioeconomic background (based on parental levels of education and the presence of cultural and educational resources, including books, in the home) varied from a low of 1.8 in Macao-China to a high of 27.4 in Peru (OECD, 2010). The fact that the variation was greater in 'partner' (non-OECD) countries than in more industrialized OECD (Organisation for Economic Co-operation and Development) countries raises questions about the appropriateness of using the measures of achievement and SES in the former. If measures of SES have been found to be limited in their capacity to capture minority students' social and economic background in industrialized countries (Sirin, 2005), one would expect even greater problems when they are used in countries that are culturally very different and less well developed economically.

Familial Factors Affecting the Development of Children's Competencies and Dispositions

The consistent finding that children's development, including their scholastic performance, is related to their socioeconomic background has led to a large number of studies designed to identify the precise factors contributing to this situation. A significant finding of such research is that variables that

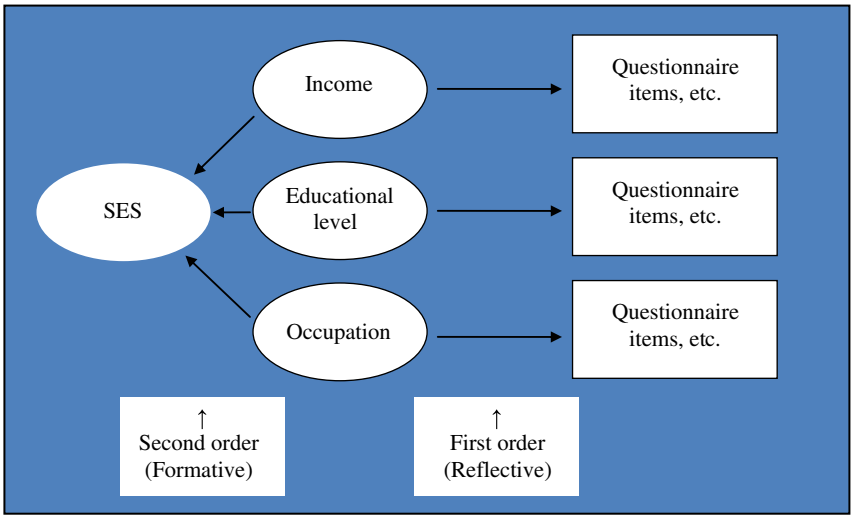

Figure 1 First- and second-order relationships between SES and its components.

describe the ethos and teaching styles of families, although related to SES, are better predictors of student learning than measures of SES alone. While studies of the relationship between SES and scholastic ability and achievement, when individuals are the unit of analysis, typically yield correlations of about 0.2, the correlations between the ethos and teaching styles of families and measures of student ability and achievement are 0.55 and higher (White, 1982). Thus, it is concluded that what parents do is more important than their status.

Familial factors that affect the development of children's competencies and dispositions that facilitate scholastic progress have been conceptualized in terms of three types of 'capital': economic or financial capital, cultural capital which relates primarily to the conditions that foster cognitive and scholastic development, and social capital which is primarily related to conduct, identity, social behavior, attitudes, and motivation. Though the three types may be considered separately, they are interrelated. This is evident when, for example, parents' expectations and encouragement are used in definitions of both cultural and social capital. Other aspects of the relationship are in evidence when social capital is considered to enhance cultural capital, or when cultural capital is considered convertible under certain circumstances to economic capital.

Economic Capital

Economic capital relates to the material, particularly financial, resources available to families and communities. Its absence constitutes poverty, which is reflected in an inability to provide the basic resources required to support children's development. Among the factors that may affect children's ability to attend school and to learn are hunger, malnutrition, and a variety of diseases. Differences in the health of children have been found to explain a significant portion of the achievement gap between social classes (CDC, 2002). This is so even in industrialized countries, but is more pronounced in developing countries.

The role of the home is perhaps most obvious in the early years of children's development when homes take care of children's physical needs for food, nutrition, and health. However, all homes may not do this successfully, due to either lack of resources or ignorance. Families and communities are often unaware of the effects of nutrition, in particular that poor nutrition affects student achievement (CDC, 2002).

Nearly half of children in the developing world are underweight or stunted; a third are infected with parasitic worms; and a similar proportion suffer from micronutrient deficiencies. Malnutrition is increasing in Sub-Saharan Africa. While it is decreasing in South Asia, that region still has the highest rates and the largest numbers of malnourished children. This all adds up to a picture in which hundreds of millions of children are affected by ill health and hunger in ways that have consequences for their education (Bundy, 2010).

Cultural Capital

In addition to taking care of the physical needs of children, homes also lay the foundation for children's school learning by fostering the development of language, self-concepts, interpersonal skills, values, and motivation to engage in the formal learning tasks of the school. These activities are facilitated by cultural capital which exists in three forms: (1) in embodiments of cognitive and noncognitive competencies that actively organize future experience and may be inherited through socialization or may be consciously acquired; (2) in an 'objectified' state in cultural artifacts (pictures, books, instruments); and (3) in an institutionalized state (academic credentials or qualifications). Language forms an important part of cultural capital since, in addition to being a means of communication, it provides, together with a richer or poorer vocabulary, a system of categories, which enables one to decipher and manipulate complex logical and aesthetic structures (Bourdieu, 1986).

Separate strands of research in education (Kellaghan et al., 1993) and psychology (Moos and Moos, 1994) have identified a number of behaviors and conditions in the family that may be considered expressions of cultural capital: (1) modeling (in use of complex language; in planning and organization to ensure that time and space are well structured and used; intellectual-cultural orientation in activities; moral-religious emphasis); (2) providing motivation and reinforcement (encouraging and rewarding school-related activities and independence in decision-making); (3) holding high academic aspirations and expectations; (4) providing direct instruction (guiding and supporting academic work; helping with homework); and (5) ensuring that the activities engaged in are developmentally appropriate.

Social Capital

In separate, but mutually supporting research in psychology and sociology, interpersonal relationships have been identified as another area in which the family's influence on children's development can be discerned. In psychology, the focus is on cognitive aspects of support (that incorporate expectations about how others will respond when assistance is needed) and on actual supportive behavior (Pierce et al., 1996). Sociologists, on the other hand, speak of social capital, which is perceived to inhere in formal and informal networks of acquaintance and recognition between and among individuals and organizations (Coleman, 1987). Three forms have been identified: (1) shared obligations and expectations; (2) useful information that is passed between members of a social network; and (3) a value or norm consensus between adults, coupled with a willingness to monitor behavior and enforce norms that control and shape children's activities. These conditions are seen as contributing to the development in children of basic attachments, attitudes, willingness to make certain efforts, and concepts of self, all of which work to children's advantage when they go to school. It may be, however, that empirical research has overestimated the effects of social capital variables because of failure to take account of the tendency of similar people to become friends with each other. A variety of individuals and organizations can act as sources of social capital: siblings, grandparents, and other members of the extended family, neighbors, churches, and various social and cultural groups.

Variation in Family Structure

Research on the effects of variation in family structure and function on children's school learning does not provide clear-cut findings, which is not surprising given the complexity and variety of the conditions being investigated. It is, for example, difficult to design studies that will isolate the effects of family circumstances (e.g., divorce, maternal employment, blended families, lone parenting) from the effects of many conditions frequently associated with these circumstances. One consistent finding, however, is that lone parenting is often associated with school problems for children when the parent (usually a mother) is young, financially poor, and poorly educated. Single motherhood reduces children's future life chances and, since the condition is more common among low-income mothers, it most likely contributes to the intergenerational persistence of inequality (McLanahan and Percheski, 2008).

Many studies report a low negative correlation between family size and scholastic abilities, achievements, and attainment, due probably to variation in the availability of familial resources (including parental attention) for each individual child. At any rate, increased spacing between children tends to reduce the decrement associated with increase in size. The impact of family size has been found to vary with cultural context and type of outcome.

Changes over time in family structure and activities, reflected, for example, in a reduction in the number of adults (in the extreme case, in a lone-parent family) and in a decline in the range of exchanges between adults and children, are perceived by some commentators as having had a negative impact on children's educational development (Coleman, 1987). The changes are represented as reflecting an erosion of social capital.

Home Support Initiatives

Policy interest in helping homes support their children's development increased in the 1960s in the United States with implementation of Head Start, Follow Through, and Title 1 programs in preschool and early elementary grades. Policy was in part a response to dramatic changes in the structure and functions of families, as well as to the cumulative findings of research that home environments and out-of-school activities contribute significantly to students' school learning. It also was a reaction to the limited success, despite considerable investment, of school-based efforts to raise significantly the educational achievements of children living in areas described as economically and socially disadvantaged. Home support initiatives have focused on health and nutrition and on building partnerships between homes and schools.

Health and Nutrition

It was in recognition of the important role of health and nutrition that school health (medical inspection) and school feeding programs were among the first social protection programs to emerge at the beginning of the twentieth century in industrialized countries. Such programs, with a focus on improving the quality of school meals, are almost universal now in high- and middle-income countries. In low-income countries, there has been considerable growth in national school health programs over the last decade, though their coverage is often limited and in many cases regressive, providing most benefits to urban elites.

Programs include provision of food to children; micronutrient fortification (vitamins and minerals, especially vitamin A, iodine, iron, zinc, folic acid); deworming; malaria control; HIV prevention and care; and promotion of hygienic practices. Key elements of programs that have contributed to success include a focus on education outcomes; collaboration between education and health agencies; consultation with stakeholders (community-based organizations, nongovernment organizations, faith-based organizations, parent–teacher

associations); and use of existing infrastructures (especially school-based management systems) (Bundy, 2010).

Home–School Partnerships

In the United States, there is a long tradition of involving parents and communities in children's formal education in, for example, parent–teacher associations. In Europe, the situation is more complex because of differences between countries in legal arrangements and ideology. However, in many countries, recent efforts, in some cases supported by legislation, go well beyond traditional parent–teacher associations in the role they propose for parents, involving them in curriculum decision-making and in teaching and learning activities in the classroom (e.g., as classroom assistants) or in the home (e.g., in developing children's skills in problem solving and language and in supervision of homework). Many of the efforts are directed toward families in areas in which children face economic, educational, and social problems, and are considered to be at risk of educational failure, school dropout, and intergenerational poverty.

As is the case with family-support initiatives in general, parent involvement programs are designed to improve the developmental capacity and educational options available to families. They are based on a variety of models and approaches, have few theoretical frameworks, and vary in the extent to which empirical evidence is available on their effectiveness. Most programs can be described as instrumental-technical, in that they are designed to provide parents with a range of skills that are considered relevant to their children's education, addressing one or more of the aspects of family environments that are considered to impact on students' scholastic performance. Alternative approaches are based on an interpretative paradigm which, in recognition of the uniqueness of parents' experience, attempt to facilitate them in interpreting and understanding that experience. Other approaches go a step further, again helping parents interpret their situation, but to do so critically in the context of presumed determining social structures and ideologies.

There is an enormous and varied literature on programs designed to help parents provide a more stimulating and enriched environment for children's cognitive growth (especially in the preschool years) and to teach and reinforce the concepts, skills, and attitudes required in school (see, e.g., Bornstein and Bradley, 2010; Epstein, 1991; Henderson and Mapp, 2002; Kellaghan et al., 1993). Programs can be categorized by the extent to which they focus on children's formal learning and teaching. In some, the focus is distal, for example, involving parents in governance and advocacy on behalf of the school, engaging in voluntary work at school, and fund raising. In other programs, reflecting research findings on expressions of cultural capital, the focus may be considered proximal in that they attempt to increase parents' involvement in their children's learning. Such programs encompass a variety of objectives: to develop parents' understanding that the home environment has a profound effect on the school learning of children; to develop parents' sense of efficacy and belief that they can exert positive influence on their children's educational outcomes; to develop parents' self-confidence in establishing a home environment that will provide rich learning experiences for children; to encourage parents to engage in a range of activities that will contribute to the development of children's learning (e.g., engage in conversation at an appropriate level; provide access to reading material/books); to assist parents to develop strategies designed to supervise children and monitor how they spend their time; to encourage parents to talk to their children about school and what they are learning and to discuss future plans; to demonstrate specific strategies that parents can use (e.g., how to interact with children in dealing with homework); and to invite and provide opportunities for parent-teacher contacts, for example, to discuss students' scholastic progress.

Research findings on the effects of a variety of home interventions and parent involvement programs indicate that efforts to help parents become stronger partners in their children's learning can have a significant impact on children's cognitive development and school performance. The greater the involvement of parents in proximal activities, the greater the impact on student learning. To cite just one example, assisting parents to become involved in their children's homework has been found to result in higher rates of homework completion, fewer homework problems, and, possibly, improved academic performance among elementary school children (Patall et al., 2008).

Efforts to increase the involvement of parents in their children's education are not limited to industrialized countries. Such efforts find encouragement in research findings from across the world that document a positive relationship between parental involvement and a variety of school variables. For example, in 26-country international reading literacy study, degree of parental cooperation was found to be the strongest of 56 predictors of school effectiveness (Postlethwaite and Ross, 1992). Parental engagement has also been associated with reduced grade failure and repetition (Mexico); increased teacher efforts (El Salvador, Nicaragua); and higher student achievements on national primary school examinations (Kenya), tests of language and mathematics (Colombia), and life skills and writing (Bangladesh) (Gertler et al., 2006).

While evaluations of individual initiatives to improve family involvement have reported a variety of effects on parents and children, it has proved difficult to identify the precise aspects of partnerships that are effective, or the conditions in which they are most effective, because of variation in the goals of programs, in the situations in which programs are delivered, in the actual delivery of services, and in the recipients of services, together with design, methodological, and analytic flaws in studies designed to evaluate the effectiveness of programs (see, e.g., Mattingly et al., 2002; van Steensel et al., 2011).

A number of problems associated with efforts to promote partnerships have been identified. (1) Despite rhetoric regarding choice, partnership, and parents' role as primary educators, many efforts mirror the traditional home–school power relationships that they were designed to alter. (2) In many cases, schools limit the involvement of parents, denying them any real input into the construction of 'shared understandings.' (3) Many approaches are based on an implicit view of parents as being inadequate in parenting skills, or as being 'hard to reach.' (4) Many parents are untouched by the

partnership, and remain distanced and alienated. (5) The extent of parental influence depends on parents' social, cultural, and material resources. (6) Strategies generally fail to reach fathers (Vincent and Tomlinson, 1997).

The Response of Schools

The considerable effort that has been invested in programs designed to assist families, in particular parents, in providing children with experiences that will support their school learning may seem to place the onus for change firmly on the family. However, few would deny that schools also have a role to play in helping children deal with any discontinuities they may encounter in bridging the gap between their home and community experiences and the demands of schooling.

Many commentators have pointed to the need for teachers to guard against stereotyping, low expectations, and generally negative attitudes when faced with children from socioeconomically disadvantaged backgrounds who come to school with poor scholastic readiness skills. But a need has also been identified for a more dynamic response from schools. This might involve a range of instructional strategies designed to improve student learning or what has been called 'culturally responsive teaching' or 'culturally relevant pedagogies' that include instructional interventions that serve as a bridge between home literacy and school literacy practices (see, e.g., Li, 2009). It has also been suggested that if school personnel are familiar with children's home backgrounds, they will become aware of the categories of meaning that children bring with them to school, and the competencies they have developed, which can then be taken into account in teaching. Language is a major resource in these activities since it can be used to encode relevant relationships and to point out important connections between contexts and tasks that children may not identify without assistance. When connections are verbalized and made explicit, children are helped to see relationships between tasks in different contexts and to appreciate how skills developed in one context (the home) can be applied in another (the school).

Conclusion

That family factors (including SES) are related to children's school learning is one of the most consistent findings of social science. This does not mean, however, that issues do not remain in understanding the nature of the relationship. Two issues, which merit further investigation, can be identified. One relates to the uniqueness of each child's experience in the home. The second is the complexity of relationships between families and schools.

Sufficient research evidence exists to challenge the assumption underlying studies of the relationship between family factors and student learning that key influences on children's development (e.g., general intellectual climate of the home, parenting style, parental encouragement for achievement, provision of books) are equally shared by children growing up in the same family. If this were the case, we would expect the children in a family to be more alike than they actually are. Behavioral genetics, in addressing this issue, while recognizing that certain aspects of children's genetic endowment and environment (inside and outside the home) are shared by children in a family, points out that other aspects are unique to each child, and that when genetic relatedness is taken into account, it is the nonshared aspects that contribute to variation in children's characteristics (Plomin and Daniels, 2011).

Identification of significant aspects of nonshared environments, however, has proved elusive. Some systematic sources of differences within families, such as birth order, gender, sibling interaction, and differential treatment by parents, are obvious enough. However, they have not been found to account for a great deal of variance in children's school achievements. Neither have unsystematic stochastic development processes; the impact of single environmental events tends to be very small.

Most studies of relationships between family factors and students' school learning fail to take adequate account of how schools and families interact. Other research findings, however, should serve to underline the complexity of the interactions. First, family background (including SES) helps determine the kind of school and classroom environment to which a student will have access. Schools attended by students from lower socioeconomic backgrounds are likely to have more limited resources and less experienced teachers. Second, family background influences the quality of the relationship between school personnel and parents, particularly access. Third, teachers may have lower expectations for, and negative attitudes toward, students from lower socioeconomic backgrounds. Fourth, in school, student achievement will be negatively affected by the presence of increasing densities of students from lower socioeconomic backgrounds (social context effect). This situation will be exacerbated if students are tracked/streamed at an early age on the basis of measures of achievement or scholastic ability. Greater understanding of these issues could have important implications for addressing problems that can arise in the family–school interface and in the design of strategies to address them.

See also: Cross-National Comparisons in Education: Findings from Pisa; Educational Sociology; Ethnicity and Educational Achievement; Families as Educational Settings; Parents and Teachers as Partners in Education; Social Inequality and Schooling; Socioeconomic Status and Academic Achievement.

Bibliography

Bornstein, M.H., Bradley, R.H. (Eds.), 2010. Socioeconomic Status, Parenting, and Child Development. Psychology Press, New York.

Bourdieu, P., 1986. The forms of capital. In: Richardson, J.G. (Ed.), Handbook of Theory and Research for the Sociology of Education. Greenwood Press, Westport, CT, pp. 241–258.

Bundy, D. (Ed.), 2010. Health, Equity, and Education for All. How School Health and School Feeding Programs Are Levelling the Playing Field. World Bank, Washington, DC.

CDC (Center for Disease Control and Prevention), 2002. Pediatric Nutrition Surveillance 2001 Report. US Department of Health and Human Sciences, Washington, DC.

Coleman, J.S., 1987. Families and schools. Educational Researcher 16 (6), 32–38.

Edwards, J.R., Bagozzi, R.P., 2000. On the nature and direction of relationships between constructs and measures. Psychological Methods 5, 155–174.

Epstein, J.L., 1991. Effects on student achievement of teachers' practices of parent involvement. In: Silvern, S.B. (Ed.), Advances of Reading/Language Research, Literacy through Family, Community, and School Interaction, vol. 5. JAI Press, Greenwich, CT, pp. 261–276.

Gertler, P., Patrinos, H.A., Rubio-Codina, M., 2006. Empowering Parents to Improve Education: Evidence from Rural Mexico. World Bank, Washington, DC.

Henderson, A.T., Mapp, K.L., 2002. A New Wave of Evidence. The Impact of School, Family and Community Connections on Student Achievement. Annual Synthesis 2002. National Center for Family & Community Connections with Schools, Austen, TX.

Kellaghan, T., Sloane, K., Alvarez, B., Bloom, B.S., 1993. The Home Environment and School Learning. Promoting Parental Involvement in the Education of Children. Jossey-Bass, San Francisco.

Li, G. (Ed.), 2009. Multicultural Families, Home Literacies, and Mainstream Schooling. Information Age Publishing, Charlotte, NC.

Mattingly, D.J., Prislin, R., McKenzie, T.L., Rodriguez, J.L., Kayza, B., 2002. Evaluating evaluations: the case of parent involvement programs. Review of Educational Research 72, 549–576.

McLanahan, S., Percheski, C., 2008. Family structure and the reproduction of inequalities. Annual Review of Sociology 34, 257–276.

Moos, R.H., Moos, B.S., 1994. Family Environment Scale Manual. Consulting Psychologists Press, Palo Alto, CA.

OECD (Organisation for Economic Co-operation and Development), 2010. PISA 2009 Results: Overcoming Social Background. Equity in Learning Opportunities and Outcomes, vol. II. OECD, Paris.

Patall, E.A., Cooper, H., Robinson, J.C., 2008. Parent involvement in homework: a research synthesis. Review of Educational Research 78, 1039–1101.

Pierce, G.R., Sarason, B.R., Sarason, I.G., 1996. Handbook of Social Support and the Family. Plenum Press, New York.

Plomin, R., Daniels, D., 2011. Why are children in the same family so different from one another? International Journal of Epidemiology 40, 563–582.

Postlethwaite, T.N., Ross, K.N., 1992. Effective Schools in Reading: Implications for Educational Planners. International Association for the Evaluation of Educational Achievement, The Hague.

Sieben, I., De Graff, P.M., 2003. The total impact of the family on educational attainment. A comparative sibling analysis. European Society 5, 33–68.

Sirin, S.R., 2005. Socioeconomic status and academic achievement: a meta-analytic review of research. Review of Educational Research 75, 417–453.

van Steensel, R., McElvany, N., Kurvers, J., Herppich, S., 2011. How effective are family literacy programs: results of a meta-analysis. Review of Educational Research 81, 69–96.

Vincent, C., Tomlinson, S., 1997. Home-school relationships: 'the swarming of disciplinary mechanisms'? British Educational Research Journal 23, 361–377.

White, K.R., 1982. The relation between socioeconomic status and academic achievement. Psychological Bulletin 91, 461–481.

Family and Work

Marisa Young, McMaster University, Hamilton, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, I discuss focal interconnections between family and work. First, I outline the ways in which labor market participation is stratified by family statuses, depending on marital status, presence, and age of children, from an international lens. Second, I discuss how family and work roles may interact negatively, resulting in what scholars refer to as 'work – family conflict.' I discuss recent literature on work – family conflict, including key definitions and prevalence rates cross-culturally, as well as the health, family, and work-related antecedents and consequences of this modern-day stressor.

Introduction

This article considers a fundamental subject of family and work. That is, the interaction and conflict between these two role sets. Generally referred to as 'work – family conflict,' the negative interaction of competing demands associated with work and family roles can result in deleterious health, family, and work-related consequences. In this article, I first outline how labor market patterns vary by marital and parental statuses, before discussing in detail the topic of work – family conflict. I dedicate most of this article to outlining key definitions related to the topic of work – family conflict, along with its prevalence rates, antecedents, and consequences. **Figure 1** provides a visual overview of my discussion. In closing, I identify three previously neglected areas of research for the advancement of work – family scholarship.

Labor Market Patterns and Family Statuses

Labor market patterns vary widely depending on family statuses. For example, in some countries married individuals are more likely to be employed, compared to the unmarried (Bureau of Labor Statistics, 2011; Statistics Canada, 2011). In most countries, women without children or with older children are more likely to be employed full time, compared to women with younger children, who tend to work fewer hours, if at all (Solera, 2009). These trends depend on geographical region, however. I highlight key patterns in the following section with a concentration on gender differences in single and married parents' labor market participation.

Canada and the United States

As of May 2014, the labor market participation rate in the United States was around 63% (Bureau of Labor Statistics, 2011 http://www.marketwatch.com/story/us-labor-force-participation-rate-falls-sharply-in-april-2014-05-02). This value was slightly higher in Canada during that same time (Canadian Labor Force Survey, 2014). Many of these workers come from dual-earner families. A recent U.S. study reports that in close to half of married-couple heterosexual families (with and without children), both the husband and wife work. In approximately 20% of these same families, the husband is the only spouse working, whereas in 8% of these families the wife is the only spouse working (Bureau of Labor Statistics, 2014). These patterns are similar to reports from Canada (Ferraro, 2010).

Among families with children, over 85% have at least one parent working for pay. However, the rates of participation tend to be less among single parent households, and vary by gender. For example, among single mothers in the United States, 68.2% are employed, whereas 81% of single fathers are employed. Similar patterns are observed in Canada: single fathers are more likely to be employed; however, single mothers are slowly increasing their labor market participation – from 67% in 1995 to 80% in 2005 (Luffman, 2006).

Additional differences in labor market participation can be found among mothers in the United States and Canada, depending on children's ages. For example, the labor market participation rate overall for mothers with children under the age of 18 is approximately 70%; however, participation rates are lower among those mothers with a spouse present.

Mothers with younger children in the United States are less likely to be employed compared to those with older children. In 2013, the labor market participation rate for mothers with children under 6 years was 64%, compared to a participation rate of 75% for those mothers whose youngest child was 6–17 years old. Among mothers with infants, 57% participated in paid work during this time. These figures have risen, compared to previous years (Lavery, 2011).

Similar trends are documented in Canada, where the number of women in the labor force with children under 6 years grew from 31.4% in 1976 to 67.8% in 2012 (Statistics Canada, 2013). Canadian statistics, however, suggest that single mothers are less likely to be employed compared to mothers in married-couple families, albeit this gap has decreased over the past two decades. This difference in participation rates is most apparent among those with younger children: in 2009 only 45% of single mothers with an infant at home were employed, compared to 67% of married mothers (Ferraro, 2010).

Europe, Asia, and Australia

The patterns of employment in Europe, Britain, Asia, and Australia vary somewhat compared to those of the United States and Canada, especially with regards to mothers' employment. Recent statistics suggest that the number of single mothers in paid employment in Eastern European countries is greater than married mothers in most countries, specifically in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.94023-6

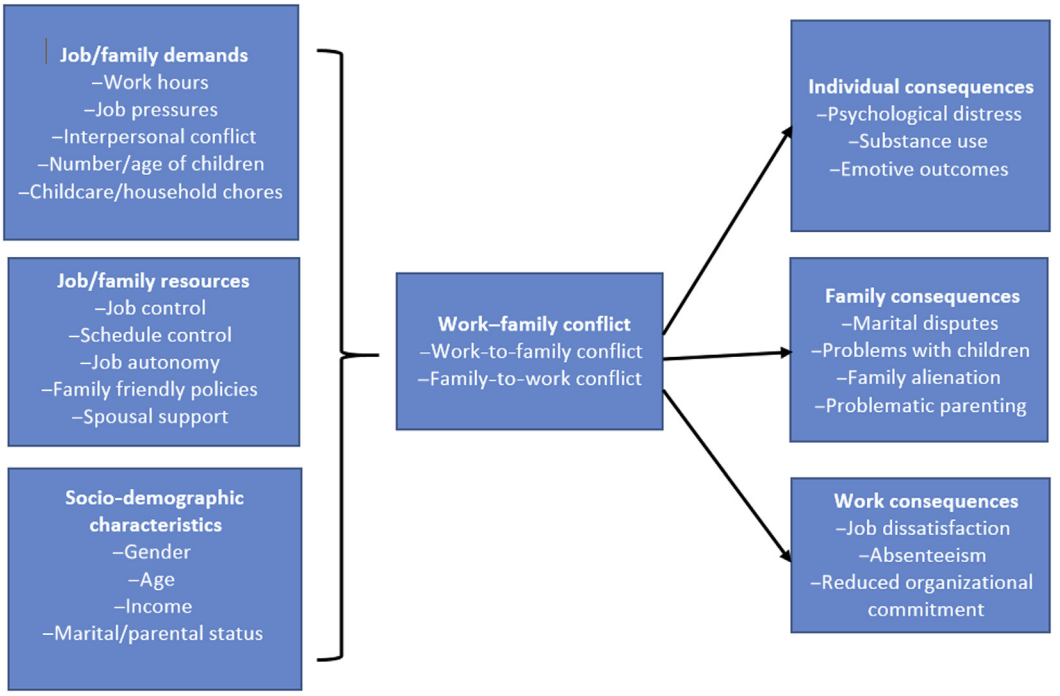

Figure 1 Theoretical model of work–family conflict antecedents and consequences.

the Slovak Republic, Spain, Luxembourg, Italy, and Greece. The difference in participation rates hover around 20%, and are likely due to levels of financial support provided by the state to single mothers. Countries that provide greater support tend to report fewer single parents in paid work, such as in Australia, Ireland, or the United Kingdom. A recent report on all single parents suggests that employment rates among this group are highest in Denmark, Iceland, Japan, Luxembourg, and Sweden and lowest in New Zealand, Turkey, and the United Kingdom (OECD, 2013).

Labor market participation rates across these countries also vary by parental status and the age of children, similar to Canada and the United States. Parents with older children are more likely to be working full time, compared to parents with younger children. The majority of parents with children older than 6 years work full time (30 h or more per week), whereas the majority of parents with at least one child under 6 years work part-time (less than 30 h per week). There are exceptions among single parents with younger children, who are more likely to work full time, again, depending on the level of state support afforded to sole-parent families (Eurostat, 2012).

Combining Family and Work: Conflict and Consequences

Greater numbers of parents – both married and single – in paid work across various geographical regions have led to growing experiences of work – family conflict. Work – family conflict refers to "a form of inter-role conflict in which the role pressures from the work and family domains are mutually incompatible so that participation in one role [family] is made more difficult by participation in another role [work]" (Greenhaus and Beutell, 1985: p. 77). Work – family conflict is considered bidirectional, including when work interferes with family (work-to-family conflict), or when family interferes with work (family-to-work conflict). Work – family conflict may be synonymous with work – family interference and work – nonwork conflict (Greenhaus et al., 2006). The combination of work and family can also lead to positive experiences, or what has been called 'work – family facilitation.' Defined as "the extent to which participation at work (or home) is made easier by virtue of the; experiences, skills and opportunities gained or developed at home (or work)" (Frone, 2003: p. 145). Similar to work – family conflict, work – family facilitation is bidirectional, and can be initiated by either domain. Despite recent attention to this topic, the details of work – family facilitation are beyond the scope of this article.

Work – family conflict may occur as some combination of time-, strain-, or behavioral-based conflict across competing role sets. Time-based conflict reflects the incompatibilities between work and family based on objective 'time' restrictions to one domain. Simply put, time spent in one domain makes it difficult to fulfill time obligations in the other domain. Strain-based conflict refers to strain experienced in one domain that carries over and interferes with responsibilities in the opposite domain. Behavioral-based conflict reflects processes by which behaviors in one domain restrict or contradict behaviors in the other domain, resulting in conflict. For example, leadership roles in the workplace are often characterized by assertiveness, aggressiveness, objectivity, and achievement. These behaviors may contradict family role expectations of the nurturing, caregiving, comforting spouse and/or parent, leading to greater conflict (Greenhaus and Beutell, 1985).

Prevalence and Consequences of Work – Family Conflict

Using similar definitions to those outlined previously, researchers have found moderate to high reports of work – family conflict, internationally. A recent Canadian study on the topic highlights that over 30% of Canadians report high work-to-family conflict, while approximately 40% report at least moderate levels (Duxbury and Higgins, 2005; Duxbury et al., 2008). When asked about the extent to which family life interfered with work life, 43% of this same sample reported high conflict across these domains.

American-based studies report similar findings, including the National Study of the Changing Workforce (Galinsky et al., 2008). Experiences of work – family conflict have been increasing among Americans – especially males. In 1977 only 34% of men reported work – family conflict, whereas 49% did in 2008. These values are higher among fathers in dual-earner households, 60% of which report some to high exposure to conflict today (Aumann et al., 2011).

Reports of work – family conflict in other parts of the world vary depending on the cultural context, with emphasis on the prevalence of dual versus single-earner families, and national-level work – family policies. For example, work – family conflict may be less prevalent in countries with more affordable child care or more gender egalitarian states, such as in Scandinavian or Nordic countries. Alternatively, parts of Asian and Middle Eastern countries that uphold more traditional family models of the male breadwinner and female domestic provider may report less work – family conflict, given the separation of domestic and economic spheres in these contexts (Gornick and Meyers, 2008). Europeans, on average, report similar work – family conflict levels as Canadians and Americans. A recent study by the EU found that over 25% experienced some form of work – family conflict. These figures are particularly high among women, where over 90% reported at least some pressure in either direction (European Agency for Safety and Health at Work, 2012: p. 2; see Kotowska et al., 2010; for the study details).

The consequences of work – family conflict are also well documented internationally, and generally fall under three typologies: individual health, family, and work-related outcomes (Bellavia and Frone, 2005).

Health-related consequences of work – family conflict include psychological distress, depression symptoms, and overall mental instability (Frone, 2000; Mirowsky and Ross, 2003). In fact, some find that work – family conflict is among the leading causes of psychological distress among North Americans (Young and Schieman, 2012; Young et al., 2014). Others have found positive associations between work – family conflict and anxiety, poor physical health, substance abuse, and negative emotions, such as guilt, exhaustion, shame and anger (Duxbury and Higgins, 2005; Hochschild, 1997; Korabik et al., 2008; Schieman, 2006; Simon, 1995).

Family-related consequences of work – family conflict include spousal disputes, reduced marital satisfaction, family distress, family alienation, and problems with children – including social, mental, and academic outcomes (Frone et al., 1992; Milkie et al., 2009; Senecal et al., 2001). Both family-to-work conflict and work-to-family conflict have been shown to negatively affect family outcomes of these types. Family-related tardiness and absenteeism, reduced emotional support, and problematic parenting are also shown to result from work – family conflict (see Bellavia and Frone, 2005).

Work-related consequences of work – family conflict include factors such as reduced job satisfaction (Bruck et al., 2002), increased job distress (Behnson, 2002), and positively predicts turnover intentions and absenteeism (Alsam et al., 2013; Bellavia and Frone, 2005). Some scholars also find that work – family conflict can reduce organizational commitment, overall (Greenhaus et al., 2006).

Theories/Perspectives of Work – Family Conflict

There are several theories used to explain the processes of work – family conflict. Among the most well-known is *role conflict theory*, which is traditionally rooted in role theory. Role theory suggests that individuals occupy a variety of roles on a daily basis. These roles may be inherently incompatible, given that they comprise various time and behavioral expectations. The clash between obligations results in what others have called 'interrole conflict.' Others adopted this theory in application to work – family conflict. Simply put, individuals experience work – family conflict when attempting to simultaneously occupy or engage in both types of roles (Greenhaus and Beutell, 1985; Kahn et al., 1964).

Demand-resource models have also been used to explain the antecedents and processes of work – family conflict. Demand-resource models suggest that the focal determinants of work – family conflict include the demands and resources at home and work (Bakker and Geurts, 2004; Voydanoff, 2007). For example, greater work and family demands, including job pressures, work hours, or number of children may increase work – family conflict compared to otherwise. Similarly, greater work and family resources, such as work flextime, schedule control, a supportive spouse, or paid domestic help may decrease work – family conflict (see Bellavia and Frone, 2005; Voydanoff, 2007; Young, 2011).

Other scholars further delineate work/family demands and resources as either 'within' or 'between' domain. Within-domain demands/resources are specific to work or family, respectively. Examples include generalized job pressures and in-house child care obligations. Between-domain demands/resources refer to those which can transfer and disrupt or aid demands in neighboring spheres. An example of a between-domain demand might be contact from supervisors at home. Alternatively, a between-domain resource could be job flexibility, because it provides individuals the temporal and spatial ability to simultaneously engage in work and family activities (i.e., taking work home; multitasking). Thus, while within-domain demands might reduce the time available for opposite domain-expectations, between-domain demands/resources might increase the extent to which work and family obligations interact (Voydanoff, 2007).

The distinction of within and between-domain demands/resources relates to a third, more recent perspective of work – family conflict processes: border/boundary theories. *Border/boundary theories* try to isolate the conditions that lead to the blurring of boundaries between work and family domains (Clark, 2000; Nippert-Eng, 1996). From this perspective, work and family obligations are not always incompatible and at

times may overlap or blur, complicating work – family interactions. This theory focuses on the continuum of overlap between work and family – ranging from total segmentation to full integration of spheres. **Figure 2** presents a visual representation of this idea, as outlined by Clark's (2000) version of Border Theory. The greater the integration, the greater the work – family conflict experienced.

To help explain these processes, border/boundary theories draw upon terms of *flexibility* and *permeability*. 'Flexibility' refers to the degree that an individual's work responsibilities can be completed outside the expected spatial and temporal parameters of the workplace. In other words, flexibility refers to "the extent to which a border may contract or expand, depending on the demands of one domain or the other" (Clark, 2000: p. 757). Likewise, 'permeability' entails "the degree to which a role allows one to be physically located in the role's domain but psychologically and/or behaviorally involved in another role" (Ashforth et al., 2000: p. 474).

It may be that permeability and flexibility help facilitate transitions between roles. However, they may also weaken the work – family boundary and influence role-blurring activities like multitasking on work and family activities, or working at home – each of which may contribute to work – family conflict (Clark, 2000; Olson-Buchanan and Boswell, 2006). Scholars refer to this as the *role-blurring hypothesis*, which predicts additional role-blurring activities, and therefore greater work – family conflict among those who experience greater integration, and whose work – family border is more flexible and permeable (Schieman and Young, 2010).

A final perspective of work – family conflict borrows from mental health scholarship in order to conceptualize work – family conflict as a *chronic stressor*, defined within the Stress Process Model (Pearlin et al., 1981). This perspective establishes a link between the experience of work – family conflict and the idea that it is embedded in broader social origins, including social statuses, biography, and social contexts (such as the neighborhood, workplace, etc.). These contexts affect the meaning of the stressor to the individual. Chronic stressors are also unique in that, by definition, they have dual objectives and subjective elements (Wheaton, 1999); whether or not there is an objective basis for the stressor, the "perception of it operates as a stressor" (Wheaton, 1997: p. 60). In the case of work – family conflict, the antecedents and consequences of perceived versus actual conflict between work and family differ. Objective influences, like work hours or child obligations, may lead to specific instances of time-related conflict. Yet, the social roles individuals assume and the immediate context in which they find themselves may

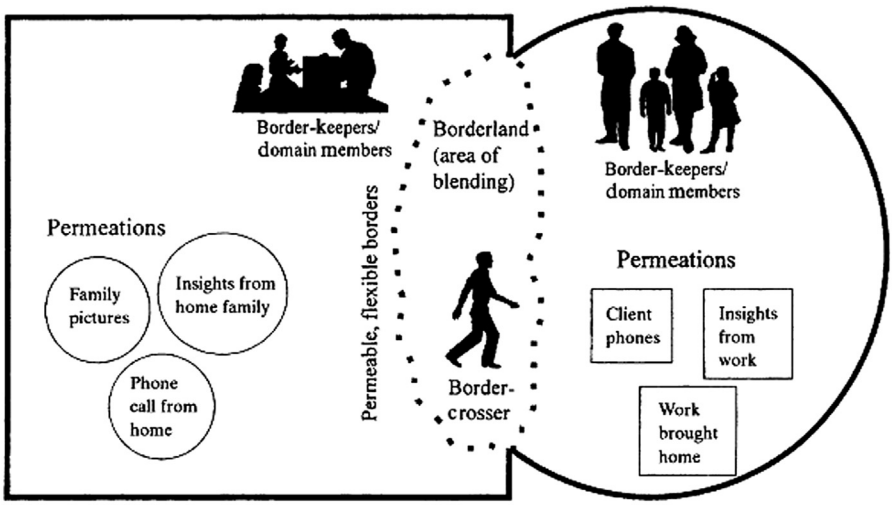

Domains

Extent of segmentation and integration
Overlap of valued means and ends
Overlap of cultures

Borders

Border strength
Permeability
Flexibility
Blending

Border-crossers

Peripheral vs central domain membership
Identification
Influence

Border-keepers & other domain members

Other-domain awareness
Commitment to border-crossers

Figure 2 Conceptualization of border theory. Reproduced from Clark, S.C., 2000. Work/family border theory: a new theory of work/family balance. Human Relations 53, 747–770.

influence whether they perceive these situations as conflictual and problematic. From this view, the subjective component of work – family conflict and its consequences may be sensitive to alternate antecedents, such as the social and psychological context in which one finds themselves (Young and Wheaton, 2013: p. 482).

Commonly Documented Influences of Work – Family Conflict

Within these various theories there are standard sets of work – family conflict antecedents reported. *Work-related antecedents* – including work-related demands and resources – are more strongly associated with work-to-family conflict compared to family-to-work conflict (Byron, 2005). Work-related demands that tend to increase work-to-family conflict include work hours, job pressures or demands, role ambiguity, work dissatisfaction, or stress, and interpersonal conflict with coworkers or supervisors. Time spent at work and job pressures tend to be the most common predictor of work-to-family conflict (Schieman et al., 2009; Michel et al., 2010). Essentially, greater objective expectations placed upon an individual increases the likelihood that work will spill over and disrupt activities in the family sphere.

Work-related resources, alternatively, tend to decrease work-to-family conflict. Resources relate to job control, support, and rewards (Karasek, 1979: p. 289). Supportive coworkers and supervisors may provide instrumental and emotional support when dealing with work-related demands. Despite this assumption, evidence is mixed. Some researchers suggest that the availability of supportive supervisors and coworkers reduces the deleterious effects of work demands on work – family conflict, while others find no significant effect, or find that support actually increases the effect of demands (Ayman and Antani, 2008; Cooper et al., 2001; Wallace, 2006).

Job control reflects schedule flexibility, decision-making latitude, supervisory authority, and/or autonomy. Job control essentially allows individuals to dictate when, where, how, and what work is performed (Karasek, 1979). The ability to manipulate one's work demands to meet necessary family demands reduces experiences of conflict between the two spheres. As posited in the job-demand/resource model, job control fosters flexibility and the efficacious completion of responsibilities (Bakker and Geurts, 2004). Evidence of the advantages of job control for work – family conflict, however, is mixed. What constitutes work-related resources has been problematized among scholars, more recently. Current research highlights that so-called resources may in fact act like 'demands' that might increase rather than decrease work – family conflict (Schieman and Young, 2010). Control over work may allow individuals to more easily engage in role-blurring activities, which can increase permeability between work and family domains, and increase work – family conflict levels.

These findings closely mirror the associations found between job rewards, such as earnings and authority, and work – family conflict. Higher earnings and greater authority might help offset work demands, thereby reducing interference with family life. However, more rewarding jobs often come with greater control and more demands – both of which have been shown to increase work – family conflict (Schieman et al., 2009).

The workplace itself – including its policies and culture – may also influence experiences of work – family conflict. Family-friendly policies provide greater opportunities for employees to integrate work and family roles by offering flexible work schedules, job sharing options, on-site child care facilities, and paid family leave. However, the use of these policies by employees, and therefore their effectiveness, depend upon the workplace culture. More traditional workplace cultures might penalize workers for using the said policies, compared to more family-friendly workplaces. Consequences endured by workers may include reduced pay, or promotion opportunities. Thus, polices may be ineffective in reducing work – family conflict unless coupled with a family-friendly culture that encourages employees to take advantage of such opportunities (Thompson et al., 1999).

Family-related antecedents are more strongly associated with family-to-work conflict, compared to work-to-family conflict. These predictors can also be divided into demands and resources. Family-related demands might include number of children, presence of young children, time spent on child care and household chores, marital dissatisfaction, or spousal disputes, to name but a few (Michel et al., 2010; Young et al., 2014). Researchers have also documented the deleterious consequences of family role ambiguity for work – family conflict (Carlson et al., 2000).

Family-related resources can be categorized into physical and emotional resources. Physical resources comprise access to safe, affordable child care facilities, quality schools, convenient work locations in proximal distance to home (Mirowsky and Ross, 2003; Voydanoff, 2007). Emotional resources refer to the support provided by one's spouse and other immediate or extended family members (Frone et al., 1997). As might be expected, the absence of these types of supports may also constitute demands.

Experiences of work – family conflict may also vary depending on *sociodemographic characteristics*, including individual's gender, age, and socioeconomic status. Men and women, for example, have been shown to report comparable levels of conflict. This is likely because of men's growing reports of family obligations over the years, at least in North America (Aumann et al., 2011).

Despite what might seem like a convergence of quantitative reports of work – family conflict among men and women, the experience of such conflict might still vary by gender. Robin Simon – a gender and mental health scholar – explains that this might be the case since men and women attribute different interpretation of how work and family roles intersect (1995: p. 183):

> [h]istorically, the meaning of work and family roles has differed for males and females, and there has been greater interdependence between work and family obligations for men than for women. Whereas men's family roles largely involved economic support, women's family roles were based on emotional support and nurturance.

Because women's work and family roles are considered more independent, conflict between them may be more

problematic, compared to men's roles, which seem overlapping (i.e., the contribution to the family is in fact financial provision itself). Different interpretations of how work and family fit with one another can influence experiences of interaction between them.

Work – family conflict experiences also vary across the life course. Parents with young children experience the greatest work – family conflict (Mirowsky and Ross, 2003). This situation is mostly applicable to middle aged individuals (i.e., approximately, 20–40 years old, depending on national context). For example, the average childbearing age for women in North America is between 25 and 30 years old (Mathews and Hamilton, 2009; Milan, 2008). Younger and older people are less likely to have children living at home, and are also less likely to be working (Schieman et al., 2001).

Socioeconomic status also predicts variable work – family conflict experiences. Research has concentrated on either side of the spectrum – the poor and the professional. Families who are less off financially may hold variable work schedules, in precarious positions, with few benefits or job rewards. These individuals must work many hours to earn a living wage. This type of situation makes it difficult to negotiate family schedules, child care, and other nonwork-related obligations. The worst affected tend to be single mothers, with few financial resources, because they often lack emotional and instrumental support from others (Williams and Boushey, 2010).

On the other end of the spectrum, are higher socioeconomic status professionals. In most developed countries, there has been a growing emphasis on working longer hours among higher status occupations. The average American family, for example, worked 11 h more per week in 2006 compared to 1979 (Gornick and Meyers, 2008). Higher status positions are also associated with greater flexibility, autonomy, and control. All of these features can lead to what Schieman and colleagues (2006) refer to as the 'stress of higher status': higher status positions require longer hours and more demands, coupled with more job control and flexibility to meet these demands. As discussed previously, the combination of high demands and control can result in more work – family conflict, since it leads to greater permeability between work and family borders (Clarke, 2000).

Thus, the poor and professional might experience similar levels of work – family conflict. The difference is the extent to which each is free to choose the amount of interference across domains; the poor have less choice than the professional. Moreover, professionals maintain the financial means to offset these conflicts, and their deleterious consequences.

Related to the other antecedents mentioned thus far, marital and parental statuses influence experiences of work – family conflict, as well. Single parents, for example, endure greater work and family demands and have fewer family resources to draw upon compared to married couples or individuals in extended family households (Duxbury and Higgins, 2005). There may also be differences between married and cohabiting couples and unattached individuals. For example, researchers argue that the family domain and its associated demands can be less salient to cohabiting couples (Nock, 1995), which may result in lower levels of conflict between role domains. Moreover, compared to coupled parents, unattached individuals may experience less work – family conflict because they have fewer family-related demands that could potentially conflict with their paid work responsibilities (Byron, 2005).

Survey Measures of Work – Family Conflict

Research in the area often measures work – family conflict through quantitative surveys using a series of Likert items. Separate sets of questions are asked to measure work-to-family conflict, and family-to-work conflict. For example, to assess the extent of family-to-work conflict, respondents are asked how often in the past 3 months "Did your family or personal life drain you of the energy you needed to do your job?," "How often did your family or personal life keep you from concentrating on your job?," "How often did you not have enough time for your job because of your family or personal life?" (Bond et al., 2003). Similar questions are asked to assess work-to-family conflict, but revised accordingly. Separate scales are then generated from each set of questions, respectively.

Despite the rigorous validation of these items, they still have flaws. For example, some scholars point out that the frequency-based responses offered may be too vague, and do not identify whether an answer of 'very often,' for example, refers to daily or weekly experiences. This problem may be remedied by asking respondents to report the number of times they experience said conflict over some designated period; providing additional response categories to capture more detailed responses (i.e., 3–5 days per week, 6–7 days per week; or, the use of time-diary methods to track individuals' behavior hourly, daily, or weekly) (Bellavia and Frone, 2005).

Directions for Future Research on Work – Family Conflict

Despite progress on the topic of work – family conflict to-date, there is much research to be done. Work – family conflict is not an isolated experience, and instead, may impact surrounding others, including family members and coworkers/supervisors. Research, however, has focused mostly on consequences of work – family conflict from an individual-level, rather than from a dyadic, family, or organizational-level. Research on the crossover stress of work – family conflict to others is growing (Bakker et al., 2009; Young et al., 2014). However, the mechanisms through which these processes unfold, and the long term consequences crossover stress has for others, remain unclear. Researchers must ask broader questions about how parents' work – family conflict affects children throughout the life course – from childhood to adulthood. Scholars must also address how the work – family conflict of one spouse impacts another's day-to-day routines and mental health. Finally, studies of workplaces must further question how work – family conflict experiences affect dynamics among coworkers at an individual and group-level. These sorts of studies would help clarify the broader impact of work – family conflict on society, more generally.

A second neglected area of research includes studies on the contextual effects on work – family conflict – specifically the impact of residential neighborhood and community contexts. An emphasis on these areas complies with current calls for research. For example, Voydanoff highlights that "… few studies have examined the relationships of most interest … that is, the effects of community demands and resources on work-family conflict …," and concludes that "… proposed relationships regarding the role of community demands, resources, and strategies require examination through studies that are methodologically sound and that incorporate a range of community characteristics" (2005: p. 591). While some scholars have begun to address this gap (Voydanoff, 2007; Young and Wheaton, 2013), much more is needed to effectively advance literature on work – family conflict.

Finally, the definition of 'family' in the phrase work – family conflict must expand to include alternative family forms and activities outside the traditionally defined nuclear family. Work – family conflict is largely associated with the modern image of the dual-career couple with young children. However, unmarried nonparents across the globe also experience work – family conflict, based on the obligations from others in one's social network, like extended family members, friends, and neighbors. Moreover, the definition of what constitutes family should be inclusive of alternative family forms, such as same-sex couples with and without children. Research on work – family conflict – including definitions and research methods – must evolve to encompass these less prevalent experiences.

Summary

This article considered a central topic related to family and work; namely, the conflict that arises when individuals experience competing demands between these two role sets. 'Work – family conflict,' as it is commonly referred to in the literature, may be experienced as time-, strain-, or behavioral competition between spheres, and is considered bidirectional in nature. Work-related antecedents are more strongly predictive of work – to family conflict, while family-related antecedents are more strongly predictive of family-to-work conflict. Multiple theories have been used to explain the complex processes of work – family conflict, including more traditional views of role conflict theory, and recent views of border-boundary theories and the Stress Process Model. Work – family conflict may also vary depending on a number of sociodemographic features, including gender, age, and income. The topic has become of focal interest to work – family scholars in recent decades, as well as health researchers, as a result of the documented deleterious consequences arising from this modern-day stressor. Yet, despite the plethora of scholarship on work – family conflict to-date, researchers have neglected to consider the impact of work – family conflict on other members of the family and workplace, as well as how work – family conflict is influenced by broader contexts, such as the workplace structure and neighborhood of residence. Advancements in work – family scholarship may be made by exploring these areas of research further.

See also: Aging and the Labor Market; Family and Schooling; Female Power at Work; Gender and Education; Old Age Work Participation; Socioeconomic Status and Academic Achievement; Time Use and Resource Consumption; Youth Joblessness.

Bibliography

Alsam, N., Imran, R., Anwar, M., Hameed, Z., Kafayat, A., 2013. The impact of work family conflict on turnover intentions: empirical evidence from Pakistan. World Applied Sciences Journal 24 (5), 628–633.

Ashforth, B.E., Kreiner, G.E., Fugate, M., 2000. All in a day's work: boundaries and microrole transitions. Academy of Management Review 25, 472–491.

Aumann, K., Galinsky, E., Matos, K., 2011. The New Male Mystique. Families and Work Institute [Online] Available from: http://Familiesandwork.Org/Site/Research/Reports/Newmalemystique.Pdf (accessed 09.06.14.).

Ayman, R., Antani, A., 2008. Social Support and Work Family Conflict. In: Korabik, K., Lero, D.S., Whitehead, D.L. (Eds.), Handbook of Work-Family Integration. Academic Press, London, pp. 287–304.

Bakker, A.B., Geurts, S.A., 2004. Toward a dual-process model of work-home interference. Work and Occupations 31, 345–366.

Bakker, A.B., Westman, M., Emmerik, H.V., 2009. Advances in crossover theory. Journal of Managerial Psychology 24, 206–219.

Behnson, S.J., 2002. Which dominates? The relative importance of work-family organizational support and general organizational context on employee outcomes. Journal of Vocational Behavior 61, 53–72.

Bellavia, G., Frone, M., 2005. Work-family conflict. In: Barling, J., Kelloway, K., Frone, M.J. (Eds.), Handbook of Work Stress. Sage Publications, Thousand Oaks, pp. 113–149.

Bond, J.T., Thompson, C., Galinsky, E., Prottas, D., 2003. Highlights of the National Study of the Changing Workforce. Families and Work Institute, New York.

Bureau of Labor Statistics, 2011. Labor Market Statistics from the Current Population Survey [Online] Available from: http://www.bls.gov/cps/lfcharacteristics.htm (accessed 09.07.14.).

Bruck, C., Allen, T., Spector, P., 2002. The relationship between work-family conflict and job satisfaction: a finer grained analysis. Journal of Vocational Behavior 60, 336–353.

Byron, K., 2005. A meta-analytic review of work-family conflict and its antecedents. Journal of Vocational Behavior 62, 169–198.

Carlson, D.S., Kacmar, M.K., Williams, L.J., 2000. Construction and validation of a multidimensional measure of work-family conflict. Journal of Vocational Behavior 56 (2), 249–276.

Clark, S.C., 2000. Work/family border theory: a new theory of work/family balance. Human Relations 53, 747–770.

Cooper, C.L., Dewe, P.J., O'Driscoll, M.P., 2001. Organizational Stress: A Review and Critique of Theory, Research and Applications. Sage Publications, UK.

Duxbury, L., Higgins, C., 2005. Who Is at Risk? Predictors of Work-life Conflict. The, 2001 National Work-life Conflict Study. Report 5. Health Canada, Ottawa.

Duxbury, L., Lyons, S., Higgins, C., 2008. Too much to do, and not enough time: an examination of role overload. In: Korabik, K., Lero, D.S., Whitehead, D.L. (Eds.), Handbook of Work-Family Integration: Research, Theory, and Best Practices. Academic Press/Elsevier, London, UK, pp. 125–140.

Duxbury, L., Lyons, S., Higgins, C., 2009. Work-life Conflict in Canada in the New Millennium: Key Findings and Recommendations from the, 2001 National Work-life Conflict Study. Report 6. Health Canada, Ottawa.

European Agency for Safety and Health at Work, 2012. Family Issues and Work-family Values [Online] Available from: https://Osha.Europa.Eu/En/Publications/E-Facts/E-Fact-57-Family-Issues-Work-Life-Balance (accessed 01.06.14.).

Eurostat, 2012. Eurostat Labour Force Data Base [Online] Available from: http://Epp.Eurostat.Ec.Europa.Eu/Portal/Page/Portal/Employment_Unemployment_Lfs/Documents/Eulfs_Database_Userguide_2012_0.pdf (accessed 28.06.14.).

Ferrao, V., 2010. Paid Work. Women in Canada: A Gender Based Statistical Report. Statistics, Canada.

Foley, S., Linnehan, F., Greenhaus, J.H., Weer, C.H., 2006. The impact of gender similarity, racial similarity, and work culture on family-supportive supervision. Group & Organization Management 31 (4), 420–441.

Frone, M.R., Russell, M., Barnes, G.M., 1996. Work-family conflict, gender and health related outcomes: a study of employed parents in two community samples. Journal of Occupational Health Psychology 1, 57–69.

Frone, M.R., Yardley, J.K., Markel, K.S., 1997. Developing and testing an integrative model of the work-family interface. Journal of Vocational Behavior 50, 145–167.

Frone, M.R., 2003. Work–family balance. In: Quick, J.C., Tetrick, L.E. (Eds.), Handbook of Occupational Health Psychology. American Psychological Association, Washington, DC.
Frone, M.R., Russell, M., Cooper, M.L., 1992. Antecedents and outcomes of work-family conflict: testing a model of the work-family interface. Journal of Applied Psychology 77, 65–78.
Frone, M.R., 2000. Work–family conflict and employee psychiatric disorders: the national comorbidity survey. Journal of Applied Psychology 85 (6), 888–895.
Galinsky, E., Bond, J.T., Sakai, K., Kim, S., Giuntoli, N., 2008. National Study of Employers. Families and Work Institute, New York.
Glass, J., Riley, L., 1998. Family responsive policies and employee retention following childbirth. Social Forces 76, 1401–1435.
Gornick, J.C., Meyers, M.K., 2008. Creating gender Egalitarian societies: an agenda for reform. Politics & Society 36 (3), 313–349.
Greenhaus, J.H., Allen, T.D., Spector, P.E., 2006. Health consequences of work-family conflict: the dark side of the work-family interface. In: Perrewé, P.L., Ganster, D.C. (Eds.), Employee Health, Coping and Methodologies: Research in Occupational Stress and Well-Being, vol. 5. Emerald Group Publishing Limited, Bingley, UK, pp. 61–98.
Greenhaus, J.H., Beutell, N., 1985. Sources of conflict between work and family roles. The Academy of Management Review 101, 76–88.
Grzywacz, J.G., Marks, N., 2000. Reconceptualizing the work-family interface: an ecological perspective on the correlates of positive and negative spillover between work and family. Journal of Occupational Health Psychology 5, 111–126.
Grzywacz, J.G., Almeida, D., McDonald, D., 2002. Work-family spillover and daily reports of work and family stress in the adult labor force. Family Relations 51, 28–36.
Hochschild, A., 1997. The Time Bind. Henry Holt, New York.
Jacobs, J.A., Gerson, K., 2004. The Time Divide: Work, Family, and Gender Inequality. Harvard University Press, Cambridge, MA.
Kahn, R., Wolfe, D.M., Quinn, R.P., Snoek, D., Rosenthal, R.A., 1964. Organizational Stress: Studies in Role Conflict and Ambiguity. Wiley, New York.
Karasek, R., 1979. Job demands, job decision latitude, and mental strain: implications for job redesign. Administrative Science Quarterly 24, 285–306.
Korabik, K., Lero, D., Whitehead, D.L., 2008. Handbook of Work-family Integration: Research, Theory, and Best Practices. Academic Press/Elsevier, London.
Kotowska, E.I., Matysiak, A., Styrc, M., Paillhe, A., Solaz, A., Vignoli, D., Vermeylen, G., Anderson, R., 2010. Second European Quality of Life Survey: Family Life and Work. Office for Official Publications of the European Communities, Luxembourg.
Lavery, D., 2011. More mothers of children in U.S. Workforce. Population Reference Bureau [Online] Available from: http://www.Prb.Org/Publications/Articles/2012/Us-Working-Mothers-With-Children.Aspx (accessed 30.06.14.).
Luffman, J., 2006. The core-age labour force. Perspectives on Labour and Income 7 (9) [Online] Available from: http://www.Statcan.Gc.Ca/Pub/75-001-X/10906/9297-Eng.htm (accessed 30.06.14.).
Matthews, T.J., Hamilton, B.E., 2009. Delayed childbearing: more women are having their first child later in life. National Center for Health Statistics Policy Brief 21, 1–8.
Michel, J.S., Kotrba, L.M., Mitchelson, J.K., Clark, M.A., Baltes, B.B., 2010. Antecedents of work–family conflict: a meta-analytic review. Journal of Organizational Behaviour 67 (2), 169–198.
Milan, A., 2008. Fertility overview. In: Report on the Demographic Situation in Canada. Component of Statistics Canada Catalogue No. 91-209-X.
Milkie, M.A., Raley, S.B., Bianchi, S.M., 2009. Taking on the second shift: time allocations and time pressures of U.S. parents with preschoolers. Social Forces 88 (2), 487–517.
Mirowsky, J., Ross, C., 2003. Social Causes of Psychological Distress, second ed. Aldine De Gruyter, New York.
Nippert-Eng, C., 1996. Home and Work. University of Chicago Press, Chicago.
Nock, S.L., 1995. A comparison of marriages and cohabiting relationships. Journal of Family Issues 16 (1), 53–76.
OECD, 2013. Maternal Employment Y Partner Status [Online] Available from: http://www.Oecd.Org/Els/Family/Lmf1_3_Maternal_Employment_By_Family_Status_July2013.pdf (accessed 29.06.14.).
Olson-Buchanan, J.B., Boswell, W.R., 2006. Blurring boundaries: correlates of integration and segmentation between work and nonwork. Journal of Vocational Behavior 68, 432–445.
Pearlin, L.I., Menaghan, E.G., Lieberman, M.A., Mullan, G.T., 1981. The stress process. Journal of Health and Social Behavior 22, 337–356.
Schieman, S., Gundy, K.V., Taylor, J., 2001. Status, role, and resource explanations for age differences in psychological distress. Journal of Health and Social Behavior 42, 80–96.
Schieman, S., Whitestone, Y., Van Gundy, K., 2006. The nature of work and the stress of higher status. Journal of Health and Social Behavior 47, 242–257.
Schieman, S., Young, M., 2010. Is there a downside to schedule control for the work-family interface? Journal of Family Issues 31 (10), 1391–1414.
Schieman, S., Milkie, M., Glavin, P., December 2009. When work interferes with life: work- nonwork interference and the influence of work-related demands and resources. American Sociological Review 74, 966–988.
Schieman, S., 2006. Anger. In: Stets, J.E., Turner, J.H. (Eds.), The Handbook of the Sociology of Emotions. Springer, New York.
Senecal, C., Vallerand, R.J., Guay, F., 2001. Antecedents and outcomes of work-family conflict: toward a motivational model. Personality and Social Psychology Bulletin 27, 176–186.
Simon, R., 1995. Gender, multiple roles, role meaning, and mental health. Journal of Health and Social Behavior 36, 182–194.
Solera, C., 2009. Women in and Out of Paid Work: Changes across Generations in Italy and Britain. The Policy Press, Portland.
Statistics Canada, 2011. Census of Canada: National Household Survey. http://www12.statcan.gc.ca/census-recensement/index-eng.cfm (retrieved 01.14.).
Statistics Canada, 2013. Labour Force Survey [Online] Available from: http://www23.Statcan.Gc.Ca/Imdb/P2sv.Pl?Function=Getsurvey&Sdds=3701 (accessed 09.07.14.).
Thompson, C., Beauvais, L., Lyness, K., 1999. When work-family benefits are not enough: the influence of work-family culture on benefit utilization, organizational attachment, and work-family conflict. Journal of Vocational Behavior 54, 392–415.
Voydanoff, P., 2007. Work, Family and Community: Exploring Interconnections. Lawrence Erlbaum Associates, NJ.
Wallace, J.E., 2006. Can women in law have it all: a study of motherhood, career satisfaction and life balance. In: Greenwood, R., Suddaby, R. (Eds.), Research in the Sociology of Organizations, Professional Service Firms, Volume 24. Elsevier Ltd., pp. 283–306.
Wheaton, B., 1997. The Nature of Chronic Stress. In: Gottlieb, B.H. (Ed.), Coping with Chronic Stress. Plenum Press, New York, pp. 43–73.
Wheaton, B., 1999. Social stress. In: Aneshensel, C., Phelan, J. (Eds.), The Handbook of the Sociology of Mental Health. Kluwer Academic/Plenum Publishers, New York.
Williams, J.C., Boushey, H., 2010. The Three Faces of Work-family Conflict: The Poor, the Professionals, and the Missing Middle. Center for American Progress. Work Life Law, San Francisco.
Young, M., Schieman, S., 2012. When hard times take a toll: the distressing consequences of economic hardship and life events within the family-work interface. Journal of Health and Social Behavior 53 (1), 84–98.
Young, M., Wheaton, B., 2013. The impact of neighborhood social composition on work-family conflict and distress. Journal of Health and Social Behavior 54 (4), 481–497.
Young, M., Schieman, S., Milkie, M., 2014. Spouse's work-to-family conflict, family stressors, and mental health among dual-earner mothers and fathers. Society & Mental Health 4 (1), 1–20.
Young, M., 2011. Neighborhood Effects on Family-to-work Conflict and Distress. Comparative Program on Health and Society Working Paper Series. The Munk Centre, University of Toronto, Toronto.

Relevant Websites

http://www.cprn.org/doc.cfm?doc=52& – Canadian Policy Research Network.
http://www.oecd.org/employment/emp/onlineoecdemploymentdatabase.htm – OECD Labor Force Survey.
http://www.workfamilyconflict.ca/ – Project 3535.
http://workfamily.sas.upenn.edu/content/welcome-1/ – Work–Family Research Network.

Family as Institution

Linda J Waite, University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The family is a mode of social organization often characterized by marriage and children. Permanence, joint production, coresidence, and the social recognition of a sexual and child rearing union are, perhaps, the most important characteristics of the institution of marriage. Although most children are born to married couples, increasingly childbearing occurs outside of the structure of marriage. The definition of family has changed to include nontraditional structures such as gay and lesbian partnerships, cohabiting couples, and nonmarried childbearing. In the United States, the share of adults who are unmarried has risen as more people delay first marriage and fewer remarry following the divorce from or the death of their spouse. The structure of the family differs around the globe, but remains one of the key social institutions in every society.

Family Mode of Social Organization

The family is "a social network, not necessarily localized, that is based on culturally recognized biological and marital relationships" (Thornton and Fricke, 1989). For the most part, this is still true. However, the rise of alternative family structures, including single parents, gay and lesbian partnerships, and cohabiting couples, sometimes including children, raises the question of whether a family must have a culturally recognized biological or marital relationship. In most cultures, a cohabiting couple living with their own child would constitute a family. But what of a woman, her child, and her cohabiting boyfriend? Does the family consist of the woman and the child, or does it include the boyfriend who has no marital, legal, or biological relationship to either one but lives with them and shares resources?

In most times and places, families were responsible for the production, distribution, and consumption of commodities; for reproduction and socialization of the next generation; and for coresidence and transmission of property. In many ways, the economic role of the family has remained unchanged. However, production of commodities and household specialization plays a decreasingly important role, due to increased longevity and declining fertility (which correspond to the fact that most of one's adult life is spent living in a household without one's own children) and increased rates of female labor outside of the home (Stevenson and Wolfers, 2007).

In most industrialized societies, the structure of the family looks quite different than it did a half-century ago. Fewer people live in families as traditionally defined. Fewer people are married, and alternative family forms such as single parenthood, cohabitation, and same-sex unions are on the rise. Changes in family structures affect family members, especially dependent children and elderly relatives. The departure from traditional marital unions seems to also coincide with the decline in fertility rates. And lower fertility rates, in turn, alter family structures, reducing family size and the number of relatives – from siblings to cousins to aunts and uncles and grandchildren – that people have.

Marriage

Permanence, joint production, coresidence, and the social recognition of a sexual and child rearing union are, perhaps, the most important characteristics of the institution of marriage. When two adults make a legally binding promise to live and work together for their joint well-being, and to do so, ideally, for the rest of their lives, they tend to specialize, dividing between them the labor required to maintain the family. This specialization allows married men and women to produce more than they would if they did not specialize. The coresidence and resource sharing of married couples create substantial economies of scale; at any standard of living, it costs much less for people to live together than it would if they lived separately. The economies of scale and the specialization of spouses both tend to increase the economic well-being of family members living together.

To put it another way, the sharing of economic and social resources through the institution of marriage acts as a sort of risk-spreader. Spouses act as a small insurance pool against life's uncertainties, reducing their need to protect themselves against unexpected events. Marriage also connects spouses and family members to a larger network of help, support, and obligation through their extended families, friends, and others. The insurance function of marriage increases the economic well-being of family members, and the support function of marriage improves their emotional well-being.

The institution of marriage also builds on and fosters *trust.* Since spouses share social and economic resources and expect to do so over the long term, both gain when the family unit gains. This reduces the need for family members to devote scarce resources to making sure that others in the family are acting in the family's best interests, thus increasing efficiency.

Marriage changes the behavior of spouses and thereby their well-being. The specialization, economies of scale, and insurance functions of marriage all increase the economic well-being of family members, and the increase is typically quite substantial. Generally, married people produce more of the goods and services consumed in the home and accumulate more assets than unmarried people (Lupton and Smith, 2003). Married people also tend to have better physical and emotional health than single people, at least in part because they are married (Mirowsky and Ross, 1989; Waite and Gallagher, 2000). The social support provided by a spouse, combined with the economic resources produced by the marriage, facilitate both the production and the maintenance of health.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32057-8

In most societies, marriage circumscribes a large majority of sexual relationships. Data from the United States show that almost all married men and women are sexually active and almost all have only one sex partner – their spouse. Unmarried men and women have much lower levels of sexual activity than the married, in part because a substantial minority have no sex partner at all (just under a quarter of the unmarried men and a third of the unmarried women who were not cohabiting had no sex partner in the previous year). Men and women who are cohabiting are at least as sexually active as those who are married, but are less likely to be sexually exclusive.

Married people and those in formally recognized cohabitations such as those in the Nordic countries (Kiernan, 2000) have higher fertility rates than unmarried and unpartnered. Indeed, a key function of marriage is the bearing and raising of children. In addition, children in married couples are better off, on average, on many dimensions than other children. A plausible explanation is that the institution of marriage directs the resources of the spouses and their extended families toward the couple's children, increasing child well-being.

Finally, marriage differs from other, less formal relationships primarily in its legal status. As previously mentioned, marriage is a legally binding contract. Spouses acquire rights and responsibilities with marriage, enforceable through both the legal systems and through social expectations and social pressure. As such, the treatment of marriage in the law shapes the institution and the extent to which it ensures family stability. Legally recognized cohabitations such as those in the Nordic countries share many of the same benefits (Kiernan, 2004).

Social and cultural traditions affect what constitutes a marriage and who is permitted to marry, and the degree to which marriage is considered a financial contract or a romantic one differs by region, religion, and culture. For example, consanguinity rates remain quite high in Sudan, Libya, and Saudi Arabia, where 40–50% of ever-married women ages 15–49 are wed to their first cousins (Rashad et al., 2005). On the other hand, some trends, like increasing age at marriage, pertain to nearly every region.

The Increase in Age at Marriage

Both men and women in the United States and many European countries are now marrying later than at any other time in at least the last 50 years. In the United States between 1970 and 2010, the median age of first marriage for American women increased by almost 6 years, from 20.8 to 26.7, and for men the median age increased by more than 5 years, from 23.2 to 28.7 (U.S. Census Bureau, 2010). From 1970 to 2000, the proportion of women who had never been married increased from 36 to 73% among those 20–24 years old and from 6 to 22% among those 30–34 years old. Similar increases occurred for men.

For African Americans, the delay in first marriage has been especially striking. The median age at first marriage increased to 30.8 for African American men and 30.3 for African American women (U.S. Census Bureau, 2011, Families and Living Arrangements). This represents an 8-year delay for African American men and a 10-year delay for African American women since the 1960s. And, in 2010, 47% of African American women and 50% of African American men ages 15 and older had never married (U.S. Census Bureau, 2010). Among families with black householders, only 44.4% were married couples compared with 80.8% of white family households (U.S. Census Bureau, 2011, Families and Living Arrangements).

Similar changes in marriage patterns have taken place in most European countries; recent cohorts are marrying at older ages and over a wider range of ages than in the past. Notably, European countries differ substantially in marriage ages. The Nordic countries of Sweden, Denmark, and Iceland show the highest average ages at marriage for women (around age 32). The Eastern European countries of Moldova, Ukraine, and Belarus show the lowest (around age 24) (UNECE, 2012, Statistical Database). The age at first marriage has risen substantially in Mediterranean countries over the last several decades. The mean age for first marriage for women was 22.3 in Greece, 24.1 in Italy, and 23.9 in Spain in 1980. In 2008, the ages increased to 28.9 in Greece, 29.9 in Italy, and 30.2 in Spain.

Delays in first marriage appear to have taken place in nearly every region on the globe, including the Middle East. For example, in the early 1970s, around 40% of women ages 15–19 were married in Kuwait and Libya, but these figures had dropped by the mid-1990s to 5 and 1%, respectively (Rashad et al., 2005). In Japan, woman's age at first marriage increased from 25.3 in 1987 to 27.4 in 2005 (Kaneko et al., 2008). The countries with the lowest age (between 17 and 18) at first marriage for women are Mali and Niger (United Nations, 2009b, World Marriage Data 2008).

The Decline of Marriage

Age at marriage has risen substantially, divorce rates are high and stable, and rates of remarriage have fallen, so that a larger proportion of adults are unmarried now than in the past. In 1970, unmarried people in the United States made up 28% of the adult population. In 2010, 44% of all adults were unmarried. In fact, the shift away from marriage has been so dramatic for blacks that in 2000, only 39% of black men and 31% of black women were married, compared to 59% of white men and 56% of white women (Fields and Casper, 2003). The decline of marriage has been largest among women with little education and almost nonexistent among college graduates, creating a divide in the family building experience along lines of class as well as race.

Though marriage rates have declined in the United States and other developed countries, the vast majority of adults do marry at some time in their lives. In the United States, more than 95% of men and women marry by age 50. And 75% of women who divorce remarry within 10 years. Relatively high proportions of men and women have not married by their late 40s in the Nordic countries where cohabitation is common (Kiernan, 2004). Caribbean countries such as Jamaica and Barbados, countries characterized by a long history of visiting relationships, in which couples are sexually active but live separately, show even higher rates of never married adults. In Sweden, for example, 24% of men and 16% of women in their late 40s had never married, and in Jamaica, 48% of men and

46% of women had never married by these ages. Delayed and foregone marriage appear to contribute to declining fertility trends, even as children (and childbearing) feature prominently in alternative family forms.

Cohabiting Families

Many contemporary heterosexual couples begin their lives together not in a marriage, but when they begin sharing a residence. They form intimate unions by cohabiting, with marriage usually following at some later point unless the relationship dissolves. In the United States, unmarried heterosexual cohabitation has become so common that most young adult Americans cohabit at some point in their lives. Rates of women cohabiting before age 25 have climbed from 7% of those born in the late 1940s to 55% among those born in the late 1960s. According to a recent estimate, 46.6% of American women between ages 15 and 45 will cohabit at some time in their lives (Heuveline and Timberlake, 2004).

Yet cohabitations tend to be relatively short-lived in the United States, with most either transformed into marriages or dissolved within a few years. The most recent estimates for the United States suggest that about 55% of cohabiting couples marry and 40% end their relationship within 5 years of moving in together. Only about one-sixth of cohabiting couples remain together without marrying for at least 3 years, and only 1 in 10 last for 5 years or more without either splitting up or marrying (Bumpass and Lu, 2000). As a result, the proportion currently cohabiting is only about 10% among women ages 20–29, compared to 27–53% currently married at these ages, respectively. At older and younger ages, fewer than 1 woman in 10 is currently cohabiting.

In the United States, cohabitation has become an increasingly common step in the courtship process, preceding the majority of marriages and remarriages. The percentage of marriages preceded by cohabitation increased from about 10% for those marrying between 1965 and 1974 to more than 50% of those marrying between 1990 and 1994. And cohabitation is especially common for those whose first marriage dissolved.

Although a number of European countries have experienced similar increases in cohabitation, some have experienced much more and some much less (Kiernan, 2004). Cohabitation is strikingly common in the Nordic countries of Denmark, Sweden, and Finland; France also shows fairly high levels, with over 30% of the women and men ages 25–34 in cohabiting unions. By contrast, a group of countries including the Netherlands, Great Britain, Germany, and Austria shows moderate levels of cohabitation, with 15–22% of women from ages 20 to 34 are in this type of union. And in the Southern European countries of Italy, Greece, and Portugal, cohabitation is rare, with less than 10% cohabiting among women in this same age group (Kiernan, 2004).

In some countries, cohabitation is a socially recognized form of partnership that is quite stable and seems to function more as a permanent alternative to marriage rather than a substitute for it. Sweden, France, and Canada fall into this category. In other countries, such as Belgium, Austria, and Germany, cohabitation appears to be more like a prelude to marriage or a stage in the courtship process. By contrast, cohabitation appears to be relatively rare in Italy, Poland, and Spain. In Latin America and the Caribbean, a significant number of girls aged 15–19 elect to live in consensual unions, with rates of almost 24% in Brazil, 20% in Nicaragua, and 18% in Dominican Republic (United Nations, 2009a, Demographic Yearbook data collections).

Cohabiting couples share a residence and a sex life, but they need to share little else. At least in the United States, cohabiting couples are much less likely than married couples to combine financial resources, less likely to be sexually exclusive, less likely to share leisure time and a social life, less likely to have children, and less likely to remain together. The requirements for establishing or ending a cohabiting union are minimal, with no legal or religious or community formalities involved. Thus, some scholars argue that cohabitation and marriage are distinct social institutions. According to some scholars, cohabitation is much less institutionalized than marriage, at least in the United States and other countries in which it has become common relatively recently, because it is not covered by clear expectations or norms, and the legal rights and responsibilities of cohabiting partners have not been established (Waite and Gallagher, 2000).

So while many people choose cohabitation over marriage, a number of conditions will need to be met before cohabitation becomes indistinguishable from marriage. Cohabitation will provide a comprehensive substitute for marriage only when couples tend to select either cohabitation or marriage at the beginning of their relationship and rarely make the transition from cohabitation to marriage; when cohabitations become about as stable as marriages; when cohabiting couples become socially recognized as permanent unions and treated as married; and when childbearing takes place in cohabiting relationships at about the same rate as in marriage.

Although not all of the above conditions apply, the gap between marriage and cohabitation is narrowing in some countries, such as in the Nordic countries, where cohabitation has much the same legal and social status as marriage. But even in the Nordic countries, cohabiting relationships are less stable than marriages, with at least one in five of the unions dissolving before the child reaches the age of 5 if the cohabitation does not lead to marriage (Kiernan, 2004).

Unmarried Childbearing

Changes in marriage have played a central role in increases in unmarried childbearing, which has reached historically unprecedented levels (Bachrach, 1998); in 2010, 41% of all births and 44% of all first births in the United States occurred to women who were not married (Martin et al., 2011). The proportion of births to unmarried women has increased without interruption from 1997 through 2009. This level compares with 33.2% in 2000 and 18.4% in 1980. Contrary to public perception, teenagers under age 20 accounted for just 21% of all nonmarital births in 2009, down from 29% a decade earlier (Martin et al., 2011). In fact, births to unmarried women increasingly occur in cohabiting households. According to the National Center for Health Statistics based on data collected from 2006 through 2010, the percentage of first births to women living with a male partner jumped from 12% in 2002

to 22% in 2006–10 – an 83% increase. Likewise, the percentage of cohabiting new fathers rose from 18 to 25%. In some regions of the United States, these numbers are even more striking. In an analysis that surveyed 250 families with unmarried parents in Oakland, California, and tracked the parents' relationships for 1 year from the time of their child's birth, half of the parents were living together when their child was born (Waller, 2001).

The proportion of births to unmarried women depends on the share of all women who are married, the fertility of married women, and the fertility of unmarried women. Marital fertility is relatively low in the United States and many European countries. Where nonmarital fertility is also low, as it is in Spain, Italy, and Japan, total fertility is substantially below replacement, but those low-fertility countries with fairly high levels of nonmarital childbearing, like the United States and France, tend to have higher total fertility (Rindfuss et al., 2000).

Unmarried childbearing varies substantially among racial and ethnic groups in the United States. The percentage of births to unmarried women is highest for black women (73%), and lowest for Asian and Pacific Islander Americans (17%), with Hispanic women (53%) and non-Hispanic whites intermediate between these two extremes at 29% (Martin et al., 2011). Unmarried childbearing also varies considerably across states and regions, but these differences can primarily be attributed to the racial compositions of these areas.

As we might expect from differences in the proportion of women who are single and in the proportion in cohabitational unions, unmarried childbearing also varies dramatically across countries in Europe. The Nordic countries, which have a high rates of cohabitation, also have quite high percentages of births to unmarried women; in Norway, for example, 46% of births in recent years occurred to unmarried women, most of whom were cohabiting. More than half of births in Sweden and almost half in Norway occur to unmarried women. In 1980, Mediterranean countries saw rates of birth to unmarried women under 5%; in the last several decades, these percentages have more than quadrupled, with 28% of births in Spain and 21% in Italy to unmarried mothers (Ventura, 2009). The social implications of unmarried childbearing depend on both the extent to which these births take place within stable, socially recognized cohabiting couples, like in the Nordic countries, and the extent to which social welfare programs cushion the financial impact on the family of having a single parent. Where few supports for single parents exist, like in the United States, children in these families do less well than children raised in two-parent families, but even under these circumstances, most children raised by single parents are happy and successful as adults (Amato and Booth, 1997).

Alternative Family Forms

The married, two-parent family has historically been the most common family form in the United States and other industrialized countries for some centuries. However, as just discussed, the proportion of women who are married has declined substantially over the last half-century in developed countries. At the same time, the proportion of U.S. families consisting of a married couple with children fell from 43% in 1950 to 20% in 2010 (Frey, 2010). Even at the height of the married-couple family era, many people lived in other types of families, most often due to the death of one member of the couple before all the children were grown. When death ended many marriages relatively early in life, remarriage and stepfamilies were common, as were single-parent families caused by widowhood (Watkins et al., 1987).

High rates of divorce combined with relatively low rates of remarriage, especially for women with children, have led to sizable proportions of families headed by a divorced single mother. In addition, the rise of cohabitation and nonmarital childbearing, discussed in detail previously, has meant that unmarried-couple families and never-married mother families are now common alternative family forms. 'Nonfamily' households are also on the rise; these consist of a person living alone or with nonrelatives only, such as roommates or an unmarried partner. In addition to delayed marriage, nonmarriage, continued high rates of marital disruption, and lower rates of remarriage, the rise in nonfamily living can also be traced to earlier nest-leaving by young adults and to increases in independent living at older ages. In 2010, 34% of all households did not contain a family compared to 19% in 1970. And people living alone formed 27% of households in 2010, up from 17% in 1970 (U.S. Bureau of Census).

Gay and Lesbian Families

One alternative family form consists of two adults of the same sex, sometimes raising children. Historically and traditionally, a family consisted of people related by blood or marriage in a culturally recognized social network of biological and marital relationships. In many places, marriage continues to be defined as a legal relationship between an adult man and an adult woman who form a new family. Gay and lesbian families, sometimes based on a socially recognized and sometimes on a legally recognized relationship, challenge these definitions. Attitudes toward sex between two adults of the same sex have become substantially more accepting in the United States over the past several decades, with about half of Americans agreeing that homosexuality should be accepted by society (Wike and Menasce Horowitz, 2007, Pew Research Center). At the same time, extension of the definition of family and marriage to same-sex couples has been hotly contested and fiercely debated.

In the United States, many attempts to extend access to marriage and family rights to same-sex couples have been turned back by legislators or voters, with some notable exceptions. Same-sex marriage has become legal in six states: Connecticut, Iowa, Massachusetts, New Hampshire, New York, and Vermont. Washington, D.C. and two Native American Tribal jurisdictions (Coquille and Suquamish) perform same-sex marriages as well. Maryland and Rhode Island recognize same-sex marriages from other states. Still, 9 states prohibit same-sex marriage in statute and 30 prohibit it in their constitution. Several European countries have moved furthest on these issues. For example, France allows same-sex couples to register their partnerships as civil unions; Denmark has extended child custody rights to same-sex couples; and perhaps most definitively, the Netherlands became the first country to grant same-sex couples full and equal rights to marriage in 2000. Since then, 10 countries (Argentina, Belgium, Canada, Denmark, Iceland,

Norway, Portugal, Spain, South Africa, and Sweden) and parts of Mexico have begun to allow same-sex couples to marry.

The National Survey of Family Growth reported that 3.4% of women and 3.6% of men identify as gay, lesbian, or bisexual, based on data from 2006–08 (Chandra et al., 2011). However, it can be incredibly difficult to ascertain rates of homosexuality due to the fact that people respond based on identity, activity, and attraction. For example, in the National Survey of Family Growth, 11% of adults reported attraction to members of the same sex and 8.8% reported same-sex behavior, but only 3.7% of adults identify as lesbian, gay, or bisexual (Chandra et al., 2011). Additionally, in the United States, more people typically identify as bisexual, especially women, than in other international studies. According to revised estimates from the 2010 Census, there were 131 729 same-sex married couple households and 514 735 same-sex unmarried partner households in the United States, roughly 0.56% of U.S. households (U.S. Census Bureau, 2010). This census is the first to collect data on same-sex married couple households.

Single Adults and Single Parents

Given that many couples in the United States and similar countries begin their intimate lives together by cohabiting rather than by marrying, it seems that the form of the union has changed more than its existence. But even when we consider both marriage and cohabitation, young adults are less likely to have formed a union now than in the past. Among young women born in the early 1950s, about one-quarter had not formed a union by age 25, compared to one-third of those born in the late 1960s. More adults are remaining single longer.

In many European countries, women typically are in either cohabiting or marital unions by their mid- to late 20s. However, 55% of Italian women and 41% of Spanish women have never been in a partnership – neither cohabited nor married at these ages, compared to around one in three Portuguese and Greek women. In the Nordic countries and France, about one-third of women ages 25–34 are single, one-third are cohabiting, and one-third are married. Union dissolution contributes to the numbers of single men and women, as does postponement of union formation, cohabitation, and marriage (Kiernan, 2004).

A number of large-scale social changes are implicated in the postponement and decline in marriage and the increase in cohabitation, divorce, and unmarried childbearing. These include the development of effective contraceptives, especially the birth control pill, and legalization of abortion, both of which contributed to the sexual revolution. These changes lowered the costs of sex between unmarried partners at the same time that attitudes toward sex between unmarried people liberalized dramatically. The sexual revolution also lowered the costs of delaying marriage, since one could have an active sex life while single, and gave women and couples much more control over the occurrence and timing of births. This latter change dovetailed neatly with women's increasing labor force participation, which was, in turn, both facilitated by and a consequence of rises in women's investment in their own education (Goldin and Katz, 2002).

Sex

In spite of the sexual revolution, marriage circumscribes the vast majority of sexual relationships. Almost all married men and women are sexually active and almost all have only one sex partner – their spouse. Unmarried men and women have much lower levels of sexual activity than the married, and frequently have no sex partner at all. Cohabiting couples are at least as sexually active as married couples, but are much less likely to be sexually exclusive (Laumann et al., 1994). Thus, the married couple remains the locus of the vast majority of sexual activity.

Social and cultural attitudes structure sex and partnership throughout the life course. For example, Latina adolescents tend to engage in sexual activity with partners who are 'substantially older' (four or more years) at much greater levels than other adolescent girls (Martinez et al., 2011). African American adolescents are more likely than other teen populations to have stronger attachment to their families than their friends, perhaps explaining why African American youth tend to have longer lasting but less intimate romantic relationship styles than other youth (Smetana and Gettman, 2006). Differences in culture, religion, and socialization remain important influences of sexual partnerships throughout the life course.

Physical health and sexual function structure sex and partnership patterns as well. Sexual attitudes and experiences change over the life course for men and women. Patterns in the availability of a partner over the life course are gender differentiated, primarily due to women's greater longevity and likelihood that they are married to men older than themselves. Men's likelihood of being in a marital or cohabiting relationship remains over 80% and stable into their 60s. Among women, however, partnership prevalence drops sharply after the early 50s (Waite et al., 2009).

Conclusions

Many public commentators and some scholars of the family argue that the family is under siege from legal, economic, and social change and in decline as a result (Popenoe, 1993). As evidence, they point to low marriage rates, high divorce rates, and low marital and high nonmarital fertility. And this evidence is compelling, as far as it goes. But most adults are married (although it is a remarriage for many), most have children, most rate their marriage as very happy and place a high value on family life, and most have only one sex partner – their spouse. Most single parents raise happy and successful children. This evidence suggests that the family remains a key social institution, even in its altered state.

See also: Children, Value of; Cohabitation: United States; Families and Households, Formal Demography of; Family Size Preferences; Family Theory: Competing Perspectives in Social Demography; Family Theory: Economics of Childbearing; Family Theory: Economics of Intergenerational Relations; Family Theory: Economics of Marriage and Divorce; Fertility Theory; Fertility Transition: Economic Explanations; Fertility of

Single and Cohabiting Women; Marriage and the Dual-Career Family.

Bibliography

Amato, P.R., Booth, A., 1997. A Generation at Risk: Growing up in an Era of Family Upheaval. Harvard University Press, Cambridge, Mass.

Bachrach, C., 1998. The changing circumstances of marriage and fertility in the United States. In: Moffitt, R.A. (Ed.), Welfare, the Family and Reproductive Behavior: Research Perspectives. National Academy Press, Washington, D.C., pp. 9–32.

Bumpass, L., Lu, H., 2000. Trends in cohabitation and implications for children's family contexts in the United States. Population Studies 54 (1), 29–41.

Chandra, A., Mosher, W.D., Copen, C., 2011. Sexual behavior, sexual attraction, and sexual identity in the United States: data from the 2006–2008 National Survey of Family Grown, 36. National Health Statistics Reports.

Cherlin, A., 2010. The Marriage-Go-Round: The State of Marriage and the Family in America Today. Vintage Books, New York.

Coontz, S., 2000. The Way We Never Were: American Families and the Nostalgia Trap, second ed. Basic Books, New York, NY.

Coontz, S., 2006. Marriage, a History: How Love Conquered Marriage. Viking, New York, NY.

Edin, K., Kefalas, M., 2005. Promises I Can Keep: Why Poor Women Put Motherhood before Marriage. University of California Press, Berkeley, London.

Fields, J., Casper, L.M., 2003. America's Families and Living Arrangements: March 2000. In: Current Population Reports. U.S. Census Bureau, Washington, DC. P20–537 (esp. Table A1).

Frey, W.H., 2010. Households and Families. State of Metropolitan America. Brookings Institution, Washington D.C., pp. 90–103.

Gerson, K., 2009. The Unfinished Revolution: Coming of Age in the New Era of Gender, Work, and Family. Oxford University Press, USA.

Goldin, C., Katz, L.F., August 2002. The power of the pill: oral contraceptives and women's career and marriage decisions. Journal of Political Economy 110, 730–770.

Heuveline, P., Timberlake, J., 2004. The role of cohabitation in family formation: the United States in comparative perspective. Journal of Marriage and Family 66, 1214–1230 (esp. Table 4, p. 1226).

Kaneko, R., Sasai, T., Kamano, T., Iwasawa, M., Mita, F., Moriizumi, R., 2008. Marriage process and fertility of Japanese married couples. The Japanese Journal of Population 6 (1), 24–50.

Kiernan, K., 2000. European perspectives on union formation. In: Waite, L., Bachrach, C., Hindin, M., Thomson, E., Thornton, A. (Eds.), Ties that Bind: Perspectives on Marriage and Cohabitation. Aldine de Gruyter, New York.

Kiernan, K., 2004. Unmarried cohabitation and parenthood in Britain and Europe. Law and Policy 26 (1).

Lareau, A., 2011. Unequal Childhoods: Class, Race, and Family Life, second ed. University of California Press, Berkeley.

Laumann, E.O., Gagnon, J.H., Michael, R.T., Michaels, S., 1994. The Social Organization of Sexuality. University of Chicago Press, Chicago, IL.

Lupton, J., Smith, J.P., 2003. Marriage assets, and savings. In: Grossbard-Shechtman, S. (Ed.), Marriage and the Economy: Theory and Evidence from Advanced Industrial Societies. Cambridge University Press, Cambridge, Mass.

Martin, J.A., Hamilton, B.E., Ventura, S.J., Osterman, M.J.K., Kirmeyer, S., Mathews, T.J., Wilson, E.C., 2011. Births: final data for 2009. National Vital Statistics Report 60 (1).

Martinez, G., Copen, C.E., Abma, J.C., 2011. Teenagers in the United States: sexual activity, contraceptive use, and childbearing, 2006–2010. National Survey of Family Growth. Vital and Health Statistics 23 (31). Hyattsville, Maryland: U.S. Department of Health and Human Services, Centers For Disease Control and Prevention, National Center for Health Statistics.

Merrill, D.M., 2011. When Your Children Marry: How Marriage Changes Relationships with Sons & Daughters. Rowman & Littlefield Publishers, Inc, Lanham.

Mirowsky, J., Ross, C.E., 1989. Social Causes of Psychological Distress. Aldine de Gruyter, New York.

Moore, M.R., 2011. Invisible Families: Gay Identities, Relationships, and Motherhood among Black Women. University of California Press, Berkeley.

Popenoe, D., 1993. American family decline, 1960–1990: a review and reappraisal. Journal of Marriage and the Family 55, 527–555.

Popenoe, D., 2009. Families without Fathers: Fathers, Marriage and Children in American Society. Transaction Publishers, US.

Rashad, H., Magued, O., Roudi-Fahimi, F., 2005. Marriage trends in the Arab world. Population Reference Bureau.

Regnerus, M., Ueker, J., 2011. Premarital Sex in America: How Young Americans Meet, Mate, and Think about Marrying. Oxford University Press.

Rindfuss, R.R., Benjamin, K., Morgan, S.P., 2000. The Changing Institutional Context of Low Fertility. Carolina Population Center, University of North Carolina, Chapel Hill, NC.

Risman, B.J. (Ed.), 2009. Families as They Really Are. W. W. Norton & Company, New York.

Smetana, J.G., Gettman, D.C., 2006. Autonomy and relatedness with parents and romantic development in African American adolescents. Developmental Psychology 42, 1347–1351.

Stevenson, B., Wolfers, J., 2007. changes and their driving forces (No. w12944). National Bureau of Economic Research Marriage and divorce.

Thornton, A., Fricke, T.E., 1989. Social change and the family: comparative perspectives from the West, China and South Asia. In: Stycos, J.M. (Ed.), Demography as an Interdiscipline. Transaction, New Brunswick, NJ, pp. 128–161, esp. 130.

UNECE (United Nations Economic Commission for Europe), 2012. UNECE Statistical Database, Gender Country Profiles, http://www.unece.org/stats/stats_h.html.

U.S. Census Bureau, 2010. American Community Survey. U.S. Government Printing Office, Washington, D.C.

U.S. Census Bureau, 2011. Families and Living Arrangements. Table F1. U.S. Government Printing Office, Washington, D.C.

United Nations, 2009a. Demographic Yearbook Data Collections (accessed in September 2012).

United Nations, 2009b. World Marriage Data 2008. United Nations Population Division, New York.

Ventura, S.J., 2009. Changing Patterns of Nonmarital Childbearing in the United States. NCHS Data Brief, No. 18. National Center for Health Statistics, Hyattsville, MD.

Waite, L.J., Gallagher, M., 2000. The Case for Marriage. Doubleday, New York.

Waite, L.J., Laumann, E.O., Das, A., Schumm, L.P., 2009. Sexuality: measures of partnerships, practices, attitudes, and problems in the National Social Life, Health and Aging Study. Journals of Gerontology: Social Sciences 64B, i56–i66.

Waller, M.R., 2001. Unmarried Parents, Fragile Families: New Evidence from Oakland. Public Policy Institute of California, San Francisco, CA.

Wallerstein, J., Lewis, J., Blakeslee, S., 2000. The Unexpected Legacy of Divorce: A 25 Year Landmark Study. Hyperion, New York.

Watkins, S.C., Meken, J.A., Bongaarts, J., 1987. Demographic foundations of family change. American Sociological Review 52, 346–358.

Wike, R., Menasce Horowitz, J., 2007. The Pew Global Attitudes Project. Pew Research Center Project, p. 35.

Family Business

Guido Corbetta, Bocconi University, Milan, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Family business is an interdisciplinary field devoted to study the structural and transitional problems of family firms. Family firms can be defined as companies of various sizes controlled by one or more owners tied by family relationship or solid alliances. The recent important theoretical and empirical developments have crystallized a double-edged vision of the family firm, which includes both the positives and the negatives of family control descending. Future studies should propose research designs that take into account the need to go beyond simple – although powerful – variables such as family ownership and leadership. As in most of the managerial fields, the simultaneous search for rigor and relevance should dominate the scholarly agenda over the next few years.

Family business is an interdisciplinary field, born as an autonomous one only in the early 1980s, involving all professionals who advise, study, and work in family firms addressing their structural and transitional problems. Family firms are dynamic systems including two subsets, the family and the firm. Traditionally, family firms have been considered as a residual of the past that inhibits the development of more efficient 'managerial' firms with dispersed shareholding. However, different factors have persuaded individuals and institutions to work on them in research and practice: high and persisting relevance in modern economies, evidence that family firms are able to grow up to very large sizes, and the need to support them to overcome the difficulties to go beyond the first generation. Today, there are 57 endowed Chairs in Family Business mostly in the United States and Europe.

Family firms can be defined as companies of various sizes controlled by one or more owners tied by family relationship or solid alliances (see Corbetta and Salvato, 2012). Despite this general tenet, there are different possible definitions of the family firm. Westhead and Cowling (1998), for instance, were among the first to notice how previous research relied on alternative views about what the family firm is. Such definitions were based on (1) the entrepreneur perception of being a family firm, (2) the family ownership, (3) the family management, and (4) the intergenerational ownership transition (Westhead and Cowling, 1998). Chua et al. (1999) noticed how different researchers use varying definitions of family involvement (in the ownership, in the business, or both) to qualify a family firm, and hence proposed an alternative theoretical definition based on the differential characteristics of family firms vis-à-vis the others. More precisely, they define the family firm as "a business governed and/or managed with the intention to shape and pursue the vision of the business held by a dominant coalition controlled by members of the same family or a small number of families in a manner that is potentially sustainable across generations of the family or families" (Chua et al., 1999: 25). The same scholars in a later study further enucleate the four distinguishing points of the family business definition, which are (1) the intention to maintain family control by the dominant coalition, (2) a set of resources and capabilities arising from family involvement and interactions among family actors, (3) a vision set by the family controlling the firm, and (4) the pursuance of such vision. Along the same line, Astrachan et al. (2002) proposed an index of family involvement called Family-Power, Experience, and Culture (F-PEC), which enables the assessment of family influence on a continuous scale rather than restrict its use to categorical variables (yes/no), and allows comparisons across businesses concerning levels of family involvement, and its effects on performance as well as other business behaviors.

While most of the empirical research refers to family ownership, ownership-based definitions adopted in different studies are largely dependent upon size and other firm characteristics, as well as upon the institutional context. For instance, while Anderson and Reeb (2004) in their study of large US firms used family control thresholds ranging from 5 to 10%, other studies relied on a considerably larger threshold of 50% of shares for private, reduced to 30 or 25% for listed firms (e.g., Minichilli et al., 2010).

The relevance of family business has recently been recognized also by important institutions such as the European Commission, which set a Family Business Group to discuss and research on family firms in individual countries, and the critical issues they face. One of the most important outcomes of the activity of this group has been to come up with a definition, included in the 2009 Report of the EU Commission. According to this definition, "A firm, of any size, is a family enterprise if: 1. The majority of votes is in possession of the natural person(s) who established the firm, or in possession of the natural person(s) who has/have acquired the share capital of the firm, or in possession of their spouses, parents, child, or children's direct heirs. 2. The majority of votes may be indirect or direct. 3. At least one representative of the family or kin is involved in the management or administration of the firm." This definition also applies to listed companies that meet the definition of family enterprise if the person(s) who established or acquired the firm (share capital) or their families or descendants possess 25% of the right to vote mandated by their share capital (EU Commission, Report, 2009). This was the first time that an official, high-level body adopted a definition of the family firm in order to establish a framework for successive

policies to promote the development of these companies. We can take this as a clear sign of the prominent position of family firms, at least in the European context.

Origins and Development of Family Business Field

Even if some family business articles appeared in the 1950s, 1960s, and 1970s, only at the beginning of the 1980s was family business born as a field acknowledged by a number of North American scholars and practitioners. A fundamental contribution was made by a management researcher (Richard Beckhard, Management and Organization Behavior Professor at MIT Sloan School of Management for more than 20 years), by a psychology researcher (Harry Levinson, professor at the Psychiatry Department at Harvard Medical School), and by a consultant (Léon Danco, CEO of the Center for Family Business in Cleveland). In the mid-1990s, they have been recognized publicly as family business pioneers and interviewed in the 1994 *Family Business Review* (FBR) issue (vol. 2) dedicated to the field's history; in these interviews, they all admit to how difficult it was to make the scientific community accept the existence of family business as a special discipline.

The first studies in family business were case descriptions without any interpretative framework, which made it difficult to open a scientific debate. Most family business professionals were practitioners at that time; as such, they were not well respected as researchers by the scientific community. Further, they were not very committed to publishing in order to share their knowledge and experiences with others.

Generally speaking, family firms traditionally have been considered – and not only by academics – as a residual of the past to overcome, beginning from the study by Berle and Means (1932) who considered separation between ownership and control as a fundamental condition of development for modern economies. Not surprisingly, the media were used to dealing with family businesses only to highlight how disputes among family owners or nepotistic behaviors negatively impacted the business.

In 1984, something of decisive importance in the development of family business occurred: a group of North American scholars and practitioners, coordinated by Barbara Hollander, gave rise to the Family Firm Institute (FFI). By means of its various services, FFI has become a special forum for knowledge sharing and development. Particularly, annual conferences and the FBR – founded by Ivan Lansberg in 1988 – have been fundamental tools that helped develop theory and practice. In the leading article of the first issue, Lansberg himself provided significant data about how early the stage of the field's development was: according to a study conducted on the Social Science Citation Index from 1977 to 1987, only 53 articles had been published in the United States on family business issues, as compared with the 813 on entrepreneurship.

In 1990, the Family Business Network (FBN) was founded in Europe. Despite its different membership policies (there are more family business members and less professionals), FBN fulfills a similar mission by spreading family business knowledge and fostering a scientific and cultural debate on this topic.

The increasing number of monographic issues on family business published in such reviews as Organizational Dynamics (summer 1983), Business History (October 1993), and Entrepreneurship Theory and Practice (1 November 1994) shows that rising attention has been devoted to this new discipline by scholars in other fields as well.

In the United States and in other parts of the world, some universities followed the increasing interest on the part of scholars and other stakeholders toward family business.

Different factors have persuaded individuals and institutions to growingly invest resources in exploring family business. Among these is the diffusion of data about the importance of family businesses in modern economies, evidence that family firms can grow big and cannot just be identified with small and medium enterprises (SMEs), and consciousness of the difficulties for family firms to last beyond the first generation. Last but not least, empirical evidence suggests that family firms represent the most common type of organization around the world. Although the different national figures vary according to the operational definition being used, family firms account for (1) 80–90% of all business enterprises in the United States, with a remarkable presence of those firms also in very large businesses (about one-third of the Fortune 500 firms are family controlled, with this number rising up to 50% for the Fortune 1000 firms); (2) more than two-thirds of firms in Australia, and the majority of businesses in the United Kingdom; (3) the vast majority of companies in most of the European Countries, for example, 83% in France, 82% in Italy, and 79% in Germany; and (4) around 95% of all firms in Asia and the Middle East, as well as the majority in other emerging countries such as Brazil (FFI global data points; Gomez-Mejia et al., 2011). This widespread presence secures an estimated percentage of the global annual gross domestic product ranging from 70 to even 90%, employing the majority of the active workforce in most of the industrialized countries in the world (FFI global data points).

Research Areas and Achieved Results

The first family business studies mostly concerned founder succession issues, which is either transition from the controlling owner archetype to another or recycling of the same archetype with different actors. Such studies (1) considered small- and medium-sized firms with simple managerial structures, which typically dealt with simpler family situations; (2) focused on leadership transitions from the founder to the heir, without an explicit acknowledgment of other family and business members, and often overlooking the importance of ownership changes; (3) considered the family as an obstacle to the effective and efficient management of the firm, suggesting a sharp distinction between family and enterprise as a fundamental condition for the business continuity.

Even if most of these studies, heavily rooted in a rationalistic approach, have been revisited or have even become outdated, they achieved the important result of opening up a new research stream by explicitly recognizing the differences – in terms of roles, goals, and needs – and the mutual influence between the family and the business.

However, the growth of the family business field occurred mostly in the past 20 years, with several important conceptual

and empirical developments that can be summarized as follows:

1. The definition of the main dynamic relationships between family and business using interpretative frameworks based on different stages of the firms, people, and families' life cycles. Along this line, Miller and Le Breton-Miller (2005) identified four elements that should secure the family businesses' success in the long run. They are (1) continuity, that is, the will of family firms to constantly and passionately realize their mission; (2) community, achieved through a cohesive team of employees united by strong values; (3) connections, since relationships with partners vastly exceed the time span and scope of ordinary market transactions; and (4) command, since family business decision makers are able to act independently, rapidly, and without rigid restrictions from shareholders;
2. A more in-depth (and insightful) approach to family business successions, admitting complex situations characterized by multiple family and nonfamily leadership candidates, and subsequently stating the importance to base the successor's selection on specific criteria such as motivation, managerial abilities, career development, outside work experience, formal education, training programs, etc.;
3. A closer attention to the differences between different types of family firms, going beyond the traditional overlap with small- and medium-sized enterprises. Along this line, Corbetta proposed four archetypes of family firms corresponding to different sizes and ownership concentration levels, and also showing typical evolutional patterns from one stage to another. Consistent to this, recent studies are increasingly arguing for the importance of sampling issues, both in terms of firm size and family ownership concentration, to influence actual research results (e.g., Miller, 2013);
4. A theoretical and empirical quest to understand the direct family ownership effects on firm behavior and performance, leading to conceptualize both the negative sides due to increasing agency costs related to problems of altruism and self-control (Schulze et al., 2001), as well as the positive sides of 'distinctive familiness,' which can lead to family-based advantages and transgenerational wealth creation (Habbershon et al., 2003). Despite the long-lasting search, however, a recent large-scale meta-analysis showed that on average family ownership has either no effect or at best offers a small performance advantage relative to other types of ownership.
5. A stronger focus on formal aspects of family governance and leadership, including the functioning of boards of directors (Anderson and Reeb, 2004), but mostly the effect of family leadership on firm performance (Anderson and Reeb, 2004; Minichilli et al., 2010). While most of these studies report a positive effect of family leadership, others raise the importance to contextualize these results within specific family firm 'types' or 'configurations' (Miller et al., 2013).
6. The importance of recognizing different levels of analysis in family business research, going beyond the importance of individuals and individual leaders, but also considering (1) interpersonal or group-level relations (e.g., between different groups of family and nonfamily employees and managers); (2) organizational level, pointing out five different sources of capital such as human, social, survivability, patient capital, and family governance structures; and (3) the role of family firms in society and at a societal level (Sharma, 2004);
7. The distinctive behavior of family firms, with respect to emotions, exploring various constructs such as emotional capital, emotional ownership, emotional returns and costs, and emotional value. These ideas have been recently conceptualized in the socioemotional wealth (SEW) approach in studying family business (Gomez-Mejia et al., 2011);
8. The influence of emotions and distinctive family (economic and noneconomic) goals on strategic objectives and decisions in family firms. According to Chrisman et al. (2005), family firms show relevant specificities in terms of (1) strategic objectives, which go beyond profit maximization and include set nonfinancial goals such as family satisfaction and sustainability of the business across the family's lifetime; (2) strategic choices, which include also quality and long-term relationships with suppliers, customers, and local communities; and (3) strategic processes, which tend to be based on trust mechanisms, cohesion, and a value-driven business orientation made possible by higher levels of goal alignment between owners and managers;
9. The subsequent strategic alternatives available to family firms, providing insights on the business or corporate strategies used by family firms to exploit their unique resources and capabilities. As such, various studies considered the relationship between family involvement in the business and choices such as diversification, acquisitions, corporate social responsibility, and financing choices, among others.

These important theoretical and empirical developments have crystallized a double-edged vision of the family firm, which includes both the positives and the negatives of family control descending from the previous considerations, also characterized as the 'bright' and the 'dark' side of family involvement in business (Minichilli et al., 2010). According to a recent review, the positive aspects include (Gomez-Mejia et al., 2011) the following:

- Patient capital that values long-term returns over short-term gains;
- Affective commitment of owners, managers, and employees;
- A culture that fosters employee identification with the organization and lower employee turnover, thereby ensuring continuity and the capacity to pursue long-term projects;
- Accumulation of social capital, which becomes a 'moral resource';
- Continuity of leadership, offering greater resilience in hard times and faster decision making.

These aspects turn into negative aspects of the family when one (or more) of the following phenomena occur (Gomez-Mejia et al., 2011):

- Appropriation of firm resources to satisfy parochial family desires;

- Nepotism, cronyism, and incompetent family members on the payroll;
- Excessive effort expended on family emotions and conflict, which distract from paying attention to market forces and business matters;
- The 'spoiled kid syndrome,' which can create feelings of inequity among employees;
- Parental altruism, overuse of business perks, and biased hiring decisions.

The intellectual curiosity to further explore the pros and cons of family involvement in business in different types of firms, as well as in different contexts, resulted in a growing number and quality of scientific articles both in field and in top-tier generalist journals. This is witnessed by the impact factor (IF) trend of the field journal (FBR), which obtained the first IF in 2007 (equal to 0.67), growing up to 2.42 in 2010, at comparable levels of more established management journals. At the same time, the interest in the topic has been marked by a wealth of articles published in journal special issues. Among the most recent experiences, Schulze and Gedajlovic edited in 2010 a special issue for the *Journal of Management Studies* entitled 'The family and enterprise: Unpacking the connections,' and exploring a number of emerging issues in family business research such as the impact of top management teams' composition and diversity on performance, the diversification decisions in family-controlled firms, or the effect of family control and ownership in different institutional settings. Interestingly, then, Steir, Lumpkin, and Wright edited in 2011 a further special issue in the *Strategic Entrepreneurship Journal* entitled 'Strategic Entrepreneurship in Family Business,' in which they discuss the ways in which the influence of family matters to strategic entrepreneurship, and how strategic entrepreneurship can contribute to understanding and strengthening of family firms, discussing concepts such as entrepreneurial orientation and portfolio entrepreneurship in family firms, mostly in a resource-based framework.

Finally, the interest in family business research has overcome in the past decade the boundaries of management research, with a growing number of contributions from influential scholars from the fields of economics, finance, and even sociology. Economists for the most part have tended to focus on questions relating to the governance of the family firm and its influence within the global economy (La Porta et al., 1999), focusing on the variety of mechanisms that family-controlled firms use to enhance their power, while finance-oriented scholars mostly concentrated attention on the role of family managers' style on firm performance, as well as the impact of family and nonfamily CEOs' succession (Bennedsen et al., 2007).

From the Present to the Future of Family Business Studies

The Theoretical Arena

A first important element to identify future directions of development for the field relates to the theoretical arena, which today characterizes family business studies. Despite the impressive number of published conceptual articles, indeed, it seems that family business scholars mostly used established theories, applying them to family firms for the purpose of understanding and explaining their uniqueness.

According to agency theory (Jensen and Meckling, 1976; Schulze et al., 2001), one of the common assumptions of economic-based investigations of family firms is that they are believed to lower agency costs due to the alignment of interests between owners and agents. Nevertheless, recent interpretations and substantive works by Bill Schulze and colleagues emphasized the negative aspects of close alignment between owners and agents, and highlighted the risk of the so-called parental altruism,' that is, the tendency of parents to attribute benefits to family members (e.g., higher compensation, consumption of perks, and generous or 'familial' appointments), in general, serving all the specific interests of the family.

The 'failure' of the agency theory to satisfactorily explain the family executive's behaviors led scholars to rely more on psychological aspects of human beings to support evidences from family business studies, such as the stewardship theory (Corbetta and Salvato, 2004). According to the stewardship theory, in serving their corporations, family executives are driven by both economic and noneconomic rationales, and experience satisfaction from the ability to share, help, and support other cherished family members in the process. They are also motivated by a need to gain intrinsic satisfaction through effectively performing exigent work, and thereby to gain recognition from peers and superiors. This contributes to these managers' sense of identity and self-esteem deriving by identification with their company and their fulfillment of family obligations.

The growth in rigor, relevance, and width of family business studies over years progressively attracted attention from strategy scholars, with the natural consequence of cross-fertilization among studies and theories. One of the most notable results has been the rise of the resource-based view (RBV) (Barney, 1991) as a theoretical lens to identify resources and capabilities that are unique, scarce, valuable, hardly imitable, and with significant rent potentials within family firms. Under this view, family involvement (so-called familiness) contributes to the building up of competitive advantage over 'faceless' competitors for the firm through higher stocks of social capital and 'patient' financial resources that enhance the firm's economic sustainability and longevity (Miller and Le Breton-Miller, 2005). These resources can also contribute indirectly to firm performance through the achievement of outcomes such as the preservation of family ties and transgenerational value creation (Sharma, 2004).

Nevertheless, the inability of the agency, stewardship, and RBV to univocally predict the positive or negative side of family businesses led a number of scholars to propose novel interpretative theoretical lenses to investigate family business phenomena.

On the one hand, some studies tried to combine agency and stewardship predictions in order to reconcile predictions stemming from the economic and psychological rationales that characterize, respectively, those two theories. As a result, many other studies suggest that family involvement is an unambiguously positive or negative influence, whereas the interpretive framework proposed by these scholars significantly conditions both those expectations. The underlying thesis here is that both

these well-argued views of family business may hold, but under different circumstances, largely characterized by the degree to which a business and its top executives are embedded within a family.

On the other hand, another group of scholars tried to introduce a new paradigm in the family business field, which again combines psychological and economic rationales in family firms' decision making, but using the preservation of the 'socioemotional endowment' of a family as the economic reference point for strategic decisions. Introduced by Gomez-Mejia et al. (2011) as a variant of the behavioral agency model, the SEW perspective refers to nonfinancial aspects of the firm that meet the family's affective needs, such as identity, the ability to exercise family influence, and the perpetuation of the family dynasty. Based on the assumption that family firms may give priority to socioeconomic rather than purely economic rationales, the preservation of SEW is a crucial feature of family firms.

The SEW perspective has recently been used in the context of family firms to explain their uniqueness with regard to a range of topics, such as managerial entrenchment or executive compensation. The general conclusion from this research is that controlling family owners tend to be guided by a set of motives different from those for other types of owners.

Theory Building or Theory Testing: Rigor and Relevance

The theoretical prolificacy represents one of the distinctive features of the family business field, as well as one of its potentially problematic issues. Searching for a unique (and unifying) conceptual framework to understand family firms' characteristics and behavior, scholars have traditionally attempted to propose new concepts, ideas, and framework, with the result that empirical testing of newly elaborated constructs has been often downplayed. Among recent studies published in the FBR, for instance, constructs and related propositions on family emotions, family social capital, and knowledge integration have been proposed, but not tested yet. This evidence parallels calls from the SEW growing literature: despite SEW being so far used as an interpretative framework to explain outcome variables, few or no studies actually measured the socioemotional dimension of families, and the challenge to measure SEW is definitely next.

This impression is confirmed in a recent work (Tognazzo, 2012), which considers 279 articles published in the FBR from 1998 to 2010, for the purpose of distinguishing between theory building (introducing new constructs, new relationships, or variables) and theory testing (test hypothesis on the bases of previous theories using large samples). Interestingly, trends of theory building and testing in FBR are stable over the past decade, with theory building dominating over theory testing, exactly the opposite of what the author found in more traditional journals such as the *Academy of Management Journal*, *Administrative Science Quarterly*, and *Strategic Management Journal*, in which theory testing instead dominates. At the same time, however, the relevance for practice, which traditionally characterized FBR articles, significantly dropped in the past decade, with the article practical relevance score halved during the period considered for the analysis. It confirms a tendency toward a proliferation of new constructs and relationships, at the expense of more 'rigorous' tests of established paradigms on large samples, also in order to gain legitimacy in the field, as well as of 'relevant' research with practical implications. As in most of the managerial fields, the simultaneous search for rigor and relevance should dominate the scholarly agenda over the next few years.

Some Insights on the Next Developments in the Family Business Field

Drivers of Development

A recent survey of 80 family business scholars by Litz et al. (2012) about the state of the art and future developments of the field supports the above insights and considerations. Specifically, the authors contend how, although the field has achieved a substantial consensus, this consensus is at best of a penultimate nature (Litz et al., 2012). In other terms, the impressive growth experienced by the family business field in the past two decades has potential downsides that inhibit proper consensus. On the one hand, the fast rise in numbers of doctoral students and scholars interested in the topic determined a scattered growth characterized by spotty referencing and fragmented views of the field, casting doubts about whether this quantitative growth has always been accompanied by a qualitative development; consequently, while there is definite consensus that questions relating to family business are important, there is much less consensus as to the answers (Litz et al., 2012). This lack of consensus is mirrored by additional insights stemming from the survey. As the authors effectively summarize, "the field will need to develop and innovate by expanding its conceptual boundaries to mode deliberately include the family, enlarge its temporal boundaries to more reflectively appreciate the past, open its international boundaries to welcome findings from a growing array of international and ethnic contexts, and enrich its inherent complexity to appreciate works that present novel variables and engage in increasingly diverse set of topics" (Litz et al., 2012, p. 30).

In addition to consensus, a second important driver of development might be an increasing reliance on an interdisciplinary approach to the field. In terms of necessary competencies, many family business scholars believe that most research topics need interdisciplinary team effort; in particular, researchers in different branches of management, psychology, and sociology should cooperate to capture the distinctive facets of family business. For instance, organizational behavior and social psychology can help one to better understand the role of emotions in these firms, finance can contribute in helping one to understand the constraining role of capital and financial markets on future growth, and sociology can help one to unravel the influence of institutional pressures and family networks on firm survival (Gomez-Mejia et al., 2011). The reason why interdisciplinary collaboration is not well developed is most likely because of the high costs of joint research due to the very different starting points in terms of knowledge, languages, and methodologies. Nevertheless, since in many fields radical advances come from collaboration between authors in different disciplines and professions, such collaboration should be strongly encouraged.

Relevant Topics and Methodological Challenges

In terms of content, the future development of the field will be characterized by a number of relevant issues, as well as by important methodological challenges.

On the one hand, companies' strategic context is significantly changing because of the impact of new technologies and globalization, as well as the impact of the economic crisis that started in 2008. As such, the family businesses' capability to compete in such a context will represent a relevant research topic. Another key issue will be the impact of increasing managerial and financial sophistication on the culture and behavior of family businesses, as well as the responses of family firms to increasing institutionalization of their governance forms. This is evident in the recent development of codes of good governance for private firms (Buysse Code in Belgium; EcoDa Recommendations; Institute of Directors' Code in the United Kingdom), and specifically for family firms (IFC Family Business Governance Handbook – World Bank; Governance für Familienunternehmen in Switzerland; Guía Colombiana de Gobierno Corporativo para Sociedades Cerradas y de Familia in Colombia).

This is also related to what Chrisman et al. (2010) propose in their review and research agenda. According to these scholars, three important emerging issues in the family business research are (1) family governance systems, devoting more attention to monitoring mechanisms, informal family system versus formal ones, boards of directors in family enterprises, with the need to investigate further their composition, functioning, and performance; (2) owner–owner agency costs, investigating situations in which controlling owners may extract benefits from the firm at the expense of minority owners; and (3) noneconomic goals, which include how the pursuit of noneconomic goals can provide competitive advantage or, on the contrary, more conservative behaviors (Chrisman et al., 2010).

Finally, empirical evidence from the first four editions of the annual Italian Observatory on medium and large family-controlled firms (in collaboration with AIdAF, The Italian Association of Family Firms, Unicredit Bank, and the Chamber of Commerce of Milan) suggest additional research topics for the near future. Consistent also with the theoretical review above, these topics include the following: (1) Family firms' heterogeneity, which should stress the importance of critical dimensions such as firm size, ownership concentration, generation in command, and family pervasiveness at different leadership levels. The adoption of a configurational approach would be of help to reassess all the traditional issues of family business investigation (including the effects of governance, leadership, succession, strategy, and performance) into specific configurations or 'types' of family firms. (2) Leadership (or coleadership) models: the limited attention to coleadership is surprising in light of the importance of coleadership in family firms from a conceptual standpoint, as well as in terms of magnitude of the phenomenon. Conceptually, family business literature indicated coleadership as a leadership model particularly suited for family firms at different stages of the firm development and/or generations. Coleadership may occur among cofounders, with parents, brothers, or cousins, who share the responsibility to lead a new venture. (3) Multilevel analyses: most family business studies followed the tendency in the broader strategic management literature to investigate the impact of managerial characteristics on firm outcomes, and this tendency has been replicated in the family business field. Nevertheless, the intrinsic interplay among actors (family and nonfamily) within family firms, and the sophistication of analytical techniques and new statistical tools allow now for the possibility of further delving into this, and more attention to multilevel investigations should be devoted. (4) Family firms' behaviors: future research may benefit from considering more in-depth behaviors of family firms and family actors. Again, behaviors can be effectively investigated in different types of family firms, and include entrepreneurship (in smaller firms), corporate social responsibility, or stakeholder attention, as well as more strategic issues such as adaptability and responses to changing environments.

On the other hand, proper investigation of these issues open important methodological challenges. First, the vast majority of existing studies relied on a rather limited set of variables, including family control and family leadership mostly. Despite the challenge posed by measurement issues, future studies should propose research designs that take into account the need to go beyond simple – although powerful – variables such as family ownership and leadership. The more accurate test of constructs related to behaviors of various family and nonfamily actors inside family firms would also be of extreme help to solve some of the theoretical tensions between economic- or psychological-based theories discussed above, and to allow a better investigation of family firms' and family actors' behaviors. At the same time, however, the will to more deeply investigate family businesses raise concerns on information availability. Entrepreneurs and their families are known to be very reserved and protective of their privacy. Moreover, most family businesses do not have to disclose much information about their structure or about their performances. Researchers are trying to overcome this problem by collaborating with institutions such as banks that can provide family business databases or by designing research projects based on primary sources and making use of interviews, questionnaires, and longitudinal case studies. Sponsorship will also be essential for this purpose: the strong link with family businesses and institutions will be more and more helpful in making the relationship between theory and practice mutually beneficial, and to fund large, cross-country databases. Obviously scholars should avoid focusing on short-term needs and commit themselves to identifying long-term research goals and strategies.

See also: Bourgeoisie and Middle Classes, History of; Entrepreneurship and Organizations; Ethnicity and Ethnic Groups: Historical Aspects; Family as Institution.

Bibliography

Anderson, R.C., Reeb, D.M., 2004. Board composition: balancing family influence in S&P 500 firms. Administrative Science Quarterly 49, 209–237.

Astrachan, J.H., Klein, S.B., Smyrnios, K.X., 2002. The F-PEC scale of family influence: a proposal for solving the family business definition problem. Family Business Review 15 (1), 45–58.
Barney, J., 1991. Firm resources and sustained competitive advantage. Journal of Management 17 (1), 99–120.
Bennedsen, M., Nielsen, K., Perez-Gonzalez, F., Wolfenzon, D., 2007. Inside the family firm: the role of families in succession decisions and performance. The Quarterly Journal of Economics 122, 647–691.
Berle, A.A., Means, G.C., 1932. The Modern Corporation and Private Property. Macmillan, New York.
Chrisman, J.J., Chua, J.H., Sharma, P., 2005. Trends and directions in the development of a strategic management theory of the family firm. Entrepreneurship Theory and Practice 29 (5), 555–575.
Chrisman, J.J., Kellermanns, F.W., Chan, K.C., Liano, K., 2010. Intellectual foundations of current research in family business: an identification and review of 25 influential articles. Family Business Review 23 (1), 9–26.
Chua, J.H., Chrisman, J.J., Sharma, P., 1999. Defining the family business by behaviour. Entrepreneurship Theory and Practice 23 (4), 19–39.
Corbetta, G., Salvato, C., 2012. Strategies for Longevity in Family Firms: A European Perspective. Palgrave MacMillan, Basingstoke, UK.
Corbetta, G., Salvato, C., 2004. Self-serving or self-actualizing? Models of man and agency costs in different types of family firms: a commentary on "comparing the agency costs of family and non-family firms: conceptual issues and exploratory evidence". Entrepreneurship Theory and Practice 28 (4), 355–362.
Gomez-Mejia, L.R., Cruz, C., Berrone, P., De Castro, J., 2011. The bind that ties: socioemotional wealth preservation in family firms. Academy of Management Annals 5, 653–707.
Habbershon, T.G., Williams, M.L., MacMillan, I., 2003. A unified systems perspective of family firm performance. Journal of Business Venturing 18 (4), 451–465.
Jensen, M.C., Meckling, W.H., 1976. Theory of the firm: managerial behavior, agency costs and ownership structure. Journal of Financial Economics 3, 305–360.
La Porta, R., Lopez-de-Silanes, F., Shleifer, A., 1999. Corporate ownership around the world. Journal of Finance 54 (2), 471–517.
Litz, R.A., Pearson, A., Litchfield, S., 2012. Charting the future of family business research: perspectives from the field. Family Business Review 25 (1), 5–15.
Miller, D., Le Breton-Miller, I., 2005. Managing for the Long Run. Lessons in Competitive Advantage from Great Family Businesses. Harvard Business School Press, Boston, MA.
Miller, D., Minichilli, A., Corbetta, G., 2013. Is family leadership always beneficial? Strategic Management Journal 34 (5), 553–571.
Minichilli, A., Corbetta, G., MacMillan, I., 2010. Top management teams in family controlled companies: 'familiness', 'faultlines' and the impact on financial performance. Journal of Management Studies 47 (2), 205–222.
Schulze, W.S., Gedajlovic, E., 2010. Whiter family business. Journal of Management Studies 27, 1919–204.
Schulze, W.S., Lubatkin, M.H., Dino, R.N., Buchholtz, A.K., 2001. Agency relationships in family firms: theory and evidence. Organization Science 12, 99–116.
Sharma, P., 2004. An overview of the field of family business studies: current status and directions for the future. Family Business Review 17, 1–36.
Tognazzo, A., 2012. Entrepreneurship and Performance in Family SMEs: From Theoretical Models to a Competency Approach. PhD Thesis. Università di Padova.
Westhead, P., Cowling, M., 1998. Family firm research: the need for a methodological rethink. Entrepreneurship Theory and Practice 23 (1), 31–56.

Relevant Websites

www.fbn-i.org – Family Business Network.
fbr.sagepub.com – Family Business Review.
www.ffi.org – Family Firm Institute.
www.ifera.org – International Family Enterprise Research Academy.

Family Law in the United States

Martha A Fineman, Emory University – School of Law, Atlanta, GA, USA
June Carbone, University of Minnesota Law School, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M. Maclean, volume 8, pp. 5328–5332, © 2001, Elsevier Ltd.

Abstract

Sociolegal scholars maintain that the role of family law is to promote the 'institutionalization' of the family. Institutionalization involves the creation of normative expectations, the coordination of behavior, and the regularization of roles associated with family formation, conduct, and dissolution. Two long-term developments currently are reshaping families and transforming family law: greater autonomy for women and growing economic inequality in Western societies. These changes have eroded the formerly near-universal view of marriage as the only appropriate site for childbearing. Growing inequality has created a menu of options in family formation that are importantly shaped by class. Cherlin (2004) sees the growth in nonmarital cohabitation and same-sex unions, as representing the 'deinstitutionalization' of American marriage. Such normative 'innovations' have also been viewed as undermining the very institution of family.

Introduction

Sociolegal scholars maintain that the role of family law is to promote the 'institutionalization' of the family. Institutionalization involves the creation of normative expectations, the coordination of behavior, and the regularization of roles associated with family formation, conduct, and dissolution. Carl Schneider (1992) described this institutionalization as accomplished through the 'channeling function' of family law. Law's channeling in the family context is less coercive than it is in, for example, criminal law, but more directive than the reliance on voluntary assumption of obligations found in areas such as contract law.

Two long-term developments currently are reshaping families and transforming family law: greater autonomy for women and growing economic inequality in Western societies. These changes have eroded the formerly near-universal acceptance of marriage as the only appropriate site for childbearing. Growing inequality has created a menu of options in family formation that are importantly shaped by class. Today, marriages tend to occur much later in life, if at all. Some commentators suggest marriage has become a 'status symbol,' attainable only by those who have achieved both maturity and financial stability.

Sociologist Andrew Cherlin (2004) has also noted a move away from the social norms that once guided young people into marriage and kept them there. These norms imposed gendered roles that marked entry into adulthood across society (and which also fostered the dependence of wives on wage-earning husbands). Modern relationships are seen as part of a quest for individual expression and fulfillment rather than societally mandated institutional obligations. Cherlin describes related societal changes, such as the growth in nonmarital cohabitation and same-sex unions, as representing the 'deinstitutionalization' of American marriage. Such normative 'innovations' in coupling have also been viewed as undermining the very institution of family.

The Traditional Family of the Industrial Era

The 'traditional' family is the middle class family of the Enlightenment and the manufacturing era (Carbone, 2000). The industrial age, with greater urbanization and an expanded middle class, remade the foundations of family life. Marriage was no longer an arranged affair that cemented the property ties between two families, but a more voluntary enterprise in which the couple made choices for themselves (Stone, 1977; Coontz, 2006). In the process, women's roles were transformed. No longer subordinate to their husband's oversight of the family's commercial and personal production, wives were mistresses of a remade, middle-class domestic realm. The family ideal changed from that of an independent 'little commonwealth' (Grossberg, 1985) to a 'haven' charged with the moral instruction of the young and maintenance of the family's spiritual keel (Lasch, 1995).

At the height of these changes in Europe, philosophers like Immanuel Kant and Friedrich Hegel struggled to rationalize family bonds resting on the agreement of the couple, rather than the wisdom of their elders. Kant argued that marriage must be understood as a contract in which a man and a woman give themselves to each other unconditionally for life to satisfy the principle that people should be treated as ends in themselves rather than means to an end. Hegel resisted Kant's use of 'contract,' and asserted that if marriage were to be understood as compact, it was an agreement that transcended the limited perspective of contract. He maintained the complementary natures of male and female merged in marriage, creating a new union greater than the sum of the two individuals. The couple's commitment to each other was not merely a private compact they could dissolve at will, but a compact that united the couple with the community in a shared enterprise.

Notwithstanding their disagreements, Kant and Hegel agreed that the family necessarily rested on marriage, which was defined as the permanent union between one man and one woman. The relationship needed to be indissoluble in nature, a commitment to the community by the couple that accepted responsibility for the consequences of their sexual relationship, particularly pregnancy and childbirth, with their accompanying economic and social burdens. The marital family was the societal institution whereby dependency was privatized (Fineman, 2004). This new family ideal set the stage for greater policy emphasis on the nuclear family, as well as fostering greater middle class investment in children (Carbone, 2000).

The legal institutionalization of the marital family was bolstered by laws punishing and stigmatizing nonmarital cohabitation, fornication, and child-bearing. Other laws burnished marital commitments by making divorce extremely difficult, tying paternity to the marital family, and confining women's social and economic roles largely in terms of their marital roles as wife and mother. Within this system, the universality of marriage rested upon gender inequality. Conservative commentator William Kristol (1991) observed that for marriage to work women must be taught "to grasp the following three points: the necessity of marriage, the importance of good morals, and the necessity of inequality within marriage." Feminist Adrienne Rich (1980), writing a decade earlier, similarly explained that, "[w]omen have married because it was necessary, in order to survive economically, in order to have children who would not suffer economic deprivation or social ostracism, in order to remain respectable, in order to do what was expected of women."

Family law reinforced gendered expectations, and the subordination that accompanied them. Notions of family privacy cloaked violence that occurred behind closed doors (Fineman, 1995). Forcing unwed mothers to put their children up for adoption held the line on single parent families (Solinger, 1992). Family law for the wealthy and sophisticated allowed them to manage relationships on their own terms, (Mnookin and Kornhauser, 1979) while the poor who failed to realize the marital ideal of 'independence' and 'self-sufficiency' became visible in highly stigmatized systems of social welfare (Tenbroek, 1965; Fineman, 1995).

The 'Deinstitutionalizing' of the Traditional Family

The deinstitutionalization of the family occurred in response to three 'revolutions' that took place in the mid-twentieth century: the gender equality, sexual liberation, and no-fault divorce movements. These forces redefined the relationship between men and women, remade expectations for work and family, and decoupled marriage and reproduction (Jacob, 1988).

During the late 1960s and early 1970s, states in the United States began to lower the age of political majority from 21 to 18. This culminated in a constitutional amendment that lowered the age of majority on a national level, and eventually a Supreme Court ruling that extended the privileges of adulthood to teens. This change in the legal age had the inadvertent effect of making the newly available birth control pill legal on college campuses without parental approval. Economists Goldin and Katz indicate that the greater availability of the pill correlated with a significant delay in the age of marriage, women staying in school longer, and a dramatic increase in female enrollment in graduate and professional schools (Goldin and Katz, 2002). The legalization of abortion in 1973 saw adoptions, which peaked in 1970, cut in half by 1975 as teen births steadily declined.

These changes had a profound effect on family formation. First, education increased women's independence – even the mothers of small children could conceivably support themselves. Second, as women directed more energy to market labor, men were expected to contribute more to childrearing. Third, women gained greater control over their own sexuality, as much of the stigma associated with nonmarital sexuality was removed and pregnancy prevention became accessible. The transformation of women's roles that came with the waning of the industrial economy prompted a revolution in family law (Jacob, 1988).

The divorce law of the industrial era permitted marital dissolution only if one party (and one party alone) was at 'fault.' Flouting well-defined marital obligations justified freeing the other spouse from the bonds of what was seen as an already defunct union. If both parties were at fault, however, neither could obtain a divorce. Perhaps in light of the growing independence of women, demand for divorce rose. Spouses colluded and 'divorce factories' rose up to help them (Jacob, 1988). It was said that adultery was proved with certainty only where it had not occurred, as where one of the spouses might agree to be photographed with a paid model outside a hotel to 'prove' that divorce grounds existed. These practices discredited the process and increased the demand for divorce reform over the course of the twentieth century. When the dam obstructing reform efforts gave way, divorce reform swept the United States, with every state liberalizing the grounds for divorce between 1969 and 1985. The incidence of divorce grew dramatically, and single parents became a common occurrence.

As the stigma associated with nonmarital sexuality declined, so did the legal distinctions between marital and nonmarital children. In England, 'illegitimate' children were said to be *filius nullius*, literally, the child of no one, unable to inherit from their mother or father. In 1972, the US Supreme Court struck down state laws distinguishing between marital and nonmarital children with respect to inheritance (*Levy v. Louisiana*). In the same year, it also held unconstitutional the refusal to recognize as a legal parent an unmarried, biological father who had lived with the mother and children (*Stanley v. Illinois*). In the years that followed, family law eliminated most of the remaining distinctions between marital and nonmarital children.

Finally, custody law shifted from a model that presumed that children of tender years would be better off with their mothers, to one that assumes that children benefit from the continuing involvement of both parents. This legal shift dismantled the gendered roles of 'mother' and 'father' and replaced them with a more neutral parenting model.

The Emergence of Marriage as a Marker of Class and Legal Parentage as the Foundation of Family Law

The evolution in women's roles produced a fundamental change in the ways in which men and women match up in the choice of long term partners. Between 1960 and 2000, both men and women increased the emphasis they placed on a prospective partner's earning potential (Stone et al., 2007), with a much more dramatic change seen among men. During the same period, the degree of 'assortative mating,' that is, the tendency to marry partners with similar levels of education and income, has increased substantially (Schwartz, 2013). In addition, the average age of marriage has risen (Hymowitz et al., 2013) and, with a later age of marriage, it is easier for potential mates to determine who will realize their youthful

promise and who will fail to meet societal expectations for success. Taken collectively, these changes mean that investment in education and income potential pays off not only in terms of career potential, but in terms of both men and women's attractiveness to potential partners.

Paralleling these changes, however, has been a dramatic increase in income inequality, with a greater differential among men than among women. Significantly, at the bottom of the income distribution, men are losing ground, as employment opportunities for the less skilled have declined and employment stability has been harder to come by (Carbone and Cahn, 2014). In addition, the gendered wage gap between men and women, which in 1990 was smaller for female college graduates, today has increased substantially for the most highly educated women, even as it has been shrinking for other groups (Carbone, 2000). Changes in both income and inequality have remade the marriage market. At the top of the income distribution pyramid is a large group of successful men who would like to marry the small group of elite women. At the bottom, chronic unemployment and mass incarceration render a large number of men effectively unmarriageable. In the middle, women's income and employment stability have improved, while the fortunes of comparable men have worsened.

While attitudes about gender roles have become more flexible (Amato, et al., 2009), 'gender performance' remains important in mate selection and both men and women agree that a man who cannot hold a steady job is not a good marriage prospect (Cherlin, 2009). Further, while both men and women would like to marry someone who earns about what they do, many feel more comfortable if the man is the higher earner. The couples who report the highest marital quality manage either dual earner or male primary breadwinner relationships; working class couples where the wife works outside the home more than she would like because the husband cannot earn enough are among the most divorce prone (Amato, et al., 2009).

Taken together, greater inequality and persistent gender performance expectations have made it more likely that the women at the top will find an attractive spouse than the women at the bottom, while the women in the middle seek to pair with a shrinking group of 'comparable' men. The result remakes the terms of adult relationships, and family law ratifies the new terms, but often indirectly. In the meantime, the top, the bottom, and the middle play out their differences in terms of the legal connection to children.

The Redefinition of Adult Ties – From Dependent to Interdependent

Marriage in the industrial era was premised on female dependence; women had relatively little power within marriage (Okin, 1991) and, at divorce, the common law system typically allowed the men to retain the majority of the family's assets while the women might receive support. The common law system accorded property at divorce in accordance with title. Thus, women reclaimed any property they held before the marriage or through gift or inheritance and men kept the property produced during marriage and titled in their names. Couples might choose to put some assets, such as the family home, in their name jointly and if they did such assets would be split at divorce. Community property states, in contrast, treated all property acquired as a result of the marital couples' labor as jointly owned, irrespective of title (Harris et al., 2009).

Marriage in the information age, in contrast, is premised on interdependence (Cahn and Carbone, 2010). Husbands and wives are expected to commingle their assets and efforts; at divorce, these contributions are presumed to be equal and subject to equal division. Support, if it awarded at all, tends to be short term and transitional (Fineman, 1991; Baker, 2010). The new system treats men and women as equally capable of self-support and assumes that even long-term homemakers will reenter the labor force following a breakup.

At the same time, men and women are both expected to contribute to child care. While the 'best interest of the child' remains the lynchpin for custody awards, almost all US jurisdictions presume that continuing contact with both parents is in the child's interests. Some jurisdictions create a presumption in favor of 'joint custody,' that is, the award of legal decision-making power (joint legal custody) and shares of the child's time (joint physical custody) to both parents (Fineman, 1991). Some states have even considered a presumption in favor of shared custody principles that attempt to award roughly equal periods of time with each parent. Other states, while not necessarily favoring joint or shared custody, have enacted 'friendly parent' provisions that presume it is in the child's interests for the parent most likely to facilitate the involvement of the other to receive primary custody. In accordance with these principles, almost all parents who pursue the issue receive at least visitation; parents involved in violence, abuse, or dangerous behavior may still receive supervised visitation that allows them to spend time with the child under the supervision of a third party such as the grandparents, or in more serious cases, a trained supervisor (Meier, 2003).

Child support, in turn, has become set in accordance with relatively mechanical formulas based on both parents' income. Most states automatically adjust support in accordance with the amount of time the child spends with each parent. Increasingly, emphasis on shared custody reduces the amount of support paid to the other parent as the courts assume that each custodial parent devotes resources to the child's care while the child is in that parent's custody.

The more egalitarian nature of the law of marriage and divorce no longer assigns marital roles, support obligations, or child care responsibilities in accordance with gender stereotypes. This change in norms has also facilitated greater recognition of same-sex marriage as prohibitions on the recognition of same-sex loving relationships become pointless and punitive. Some argue that the dismantling of explicit gender roles undermines marriage and contributes to its deinstitutionalization (Cherlin, 2004). Others maintain that the recognition of same-sex unions helps remake marriage in terms of a modernist emphasis on flexible roles rather than an insistence on traditionalist or religious dictates (Cahn and Carbone, 2010; Rauch, 2010).

The transformation of marriage has also affected the willingness to marry. In 2013, a study of 18- to 29-year-old cohabitants found that better educated men and less educated women expressed the greatest reluctance. Among those without a high-school diploma, 67% of men versus only 47% of women said that they expect to marry their current partner.

Among those with at least some college, in contrast, 68% of women versus only 46% of men expected to marry their current partner (Hymowitz et al., 2013). Better educated men often hope to 'do better,' especially if their current partner does not share a similar level of income or ambition, and they are wary of early commitments that might distract them from their career objectives (Regnerus and Uecker, 2011). Yet, once they do commit, an interdependent marriage model gives them greater security than one that presumes wives' dependence. Premarital agreements that protect the assets of the higher earning party, more typically the male among elite couples, have become more common, and more likely to be enforced (Younger, 2007). Marriages of relatively short durations sometimes trigger spousal support limitations. Some states expressly restrict long-term support to marriages longer than 10 years and other states do so through judicial decisions (Baker, 2010). Moreover, the preference for shared custody arrangements protects male rights in children. Women initiate roughly two-thirds of divorces and they are less likely to do so in states that routinely award joint custody (Brinig and Allen, 2000). For all couples, the share of male custodial time with children has increased, and the more income the father has, the less likely the mother is to receive sole custody after a break-up (Brown and Cook, 2012).

The fact that well-educated women can find acceptable men and that men enjoy greater protections within an interdependent marriage model has increased the emphasis on marriage among American college graduates. This group marries relatively late, but continues to raise children primarily within marriage. The top 10% of women measured by income used to be among the least likely to marry; at the end of the twentieth century, they were the only group not to see their marriage rates decline in comparison with 1970 levels (Looney and Greenstone, 2012). For college graduates as a group, divorce rates in 2004 had fallen back to the levels of the mid-1960s, before the enactment of no-fault divorce (McLanahan, 2004). In 2006–08, white female college graduates had the same nonmarital birth rate they did in 1982 (2%). And for both whites and African-Americans, the likelihood that a 14-year-old girl was living with both parents rose between the early 1980s and the first decade of the twenty-first century (Wilcox and Marquardt, 2010). For everyone else, however, divorce, nonmarital births and relationship instability continued to rise.

Interdependent marriage, while attractive to the top group in society, poses a threat to others, and particularly to the increasing number of women in relationships with less reliable men. Male layoffs, job instability, low wages, contacts with the criminal justice system, violence, and substance abuse all contribute to the likelihood that a relationship will end, and all have increased for working class men since the mid-1980s (Murray, 2013). These factors, which were once concentrated among isolated groups such as inner city African-Americans, today affect a larger percentage of society and have decreased the supply of marriageable men. Moreover, women's share of housework follows a U-shaped curve. The more income a woman makes the less housework she does – up to the point where she earns more than her husband and then she does more (Schneider, 2011). The likelihood that a husband earns more than the wife in turn corresponds to class: in families with dual earners, the wife earns more than the husband in 70% of marriages in the bottom quintile of families, in comparison with 34% of wives in families with incomes in the top 20% (Glynn, 2010). Working class women have thus become more likely to be both the more reliable breadwinner *and* the primary caretaker. They are also more likely than men to decide to end a relationship. If they were to marry their partners and divorce, they would find that (1) it is more difficult to sever ties premised on interdependent finances; (2) the assets to be divided in the settlement, however meager, may be ones they think of as 'theirs' and/or as the assets on which they are relying to provide for their children; and (3) husbands are more likely than other fathers to receive shared custody (Brown and Cook, 2012). In these circumstances, less educated women have reason to be wary about marriage. While the divorce rates of college graduates declined after 1990, those without college degrees continued to rise at least until the financial crisis. Marriage rates for the same group declined and declined even more after 2007, and nonmarital births have steadily increased over the last 30 to 40 years.

Those who have children together without marrying describe their relationships in contingent terms rather than in terms of the unconditional commitment that Kant and Hegel described as the hallmark of marriage. Men and women often view each other with a measure of distrust, the men wary of women who expect a meal ticket and the women distrustful of the men's fidelity and reliability (Burton et al., 2009; Edin and Nelson, 2013, Wilson, 1987). These couples distinguish marriage – and its requirement of unconditional commitment – from cohabitation, and the differences in behavior correspond to the differences in attitudes. Married spouses are more likely than cohabitants to visit each other in prison, to stay together through difficult times, and to share child-rearing responsibilities (Braman, 2007; Bowman, 2010). Cohabitants engage in shorter term, more transactional exchanges. Mothers tend to stay with fathers so long as they contribute financially, remain faithful, avoid violence and criminality, and help with the child (Baker, 2010; Edin and Nelson, 2013). If the mothers do not support the fathers' involvement in the child's life, the father moves on to a new partner and often new children (McLanahan, 2004).

In this system, marriage is simply one of a number of choices about how to structure adult relationships. Pregnancy and childbirth do not necessarily imply an obligation to marry, stay together or even share childrearing responsibilities. Instead, couples individually negotiate the terms of their relationships. The more critical (and less individual) legal determination becomes the establishment of parentage.

The Primacy of the Parent–Child Tie

In the industrial era, unmarried motherhood was suspect and commanded little support. Legal paternity, in turn, did not automatically follow from biology. Instead, it required marriage. Today, legal motherhood ordinarily follows from childbirth, with the principal exception involving use of a gestational carrier (Harris et al., 2009), while legal fatherhood can be established through marriage, voluntary acknowledgment, or a paternity action. The combination of DNA testing and relationship instability, however, has made explicit that

which was hidden in other eras: biology and functioning as a parent do not always coincide. When they do not, there is a choice to be made as to which tie – biology or function – should be the basis for an allocation of decision-making power.

Marriage, both historically and today, facilitates paternity establishment. When a married woman gives birth, the husband is treated as the legal father even if neither husband nor wife takes any steps to secure official recognition. (The birth certificate, whether it names or fails to name a father, constitutes evidence of legal parentage, but it does not in itself establish either biological or legal parenthood nor necessarily serves as a barrier to a later challenge.) If the two divorce, the court will ask the couple to list any 'children of the marriage' and will not grant the divorce without an order allocating responsibility for the children in terms of both custody and support. That order, in turn, may collaterally estop each of them from later denying the husband's parental status (Carbone and Cahn, 2011). Historically, the opportunities to challenge the fact of the husband's biological paternity were limited. Both historically and today, if no one brings such a challenge, the husband remains the child's legal father irrespective of the circumstances of the child's conception.

An unmarried father, in contrast, did not historically have rights as a legal parent. He could not seek custody or visitation or prevent the mother from placing the child for adoption. While he could be held responsible for support, the ability to seek support rested with the state, rather than with the mother, and was used in an effort to offset the resources the state provided to the child (Harris et al., 2009). A man who cared about his child was expected to marry the mother or to forego any say in the child's life. Indeed, as recently as in Georgia, the only way to establish paternity outside of marriage was for the mother to swear out a criminal arrest warrant in 'fornication and bastardy' against the father.

The increase in nonmarital births and growing public expenditures for the Aid for Families with Dependent Children program led the US government to provide incentives for states to streamline paternity establishment. In the middle of the twentieth century, a pregnancy, particularly among whites, typically triggered a decision to marry or break-up. By the 1990s, unmarried mothers of all races had become more likely to be in a continuing relationship with the father, and the father had become more likely to be present with the mother at the hospital at the time of the child's birth (McLanahan, 2004). The states, recognizing these patterns, provided for 'voluntary acknowledgments of paternity' to be signed by both mother and father preferably before the child left the hospital. The acknowledgments had the force of a paternity judgment without either a court action or a paternity test (Harris et al., 2009). Like the marital presumption, they established legal parentage based on a presumption of biological paternity. Whatever the facts, the presumption remained in force absent an action to rescind the acknowledgment, in accordance with the limited circumstances and time periods provided by the statutes.

In addition to these provisions, the alleged father, the mother, or the state can initiate a paternity action to establish parentage. These actions typically require a DNA test to establish biological parentage. The state often brings such actions to compel child support payments, particularly for children receiving state benefits. Fathers also bring such actions to establish their legal status where the mothers will not cooperate with a voluntary acknowledgment.

These procedures exist to secure recognition for biological parents. Historically, adoption served as the only way to recognize an adult as a legal parent outside of a presumption of biological parenthood. Today, however, several states recognize parental status on the basis of 'functioning like a parent.' California, for example, extends a presumption of parenthood to those who "welcome the child into their households and hold out the child as [their] … own." While the statutory provision had been initially treated as a presumption of biological paternity, the California Supreme Court ruled in 2002 that proof that the presumed father was not the biological father did not necessarily rebut the presumption (*In re Nicholos H.*). In subsequent cases, the court has extended the presumption to women as well as men and indicated a clear preference for recognizing those who have assumed parental functions over those whose parenthood was based on either biology or marriage alone (Carbone and Cahn, 2014). Other states have similarly recognized parentage on the basis of function, under the labels of *de facto* or psychological parenthood or parenthood by estoppel. The courts nonetheless tend to give more weightage to biology in establishing responsibility for support while giving more weightage to function in granting rights to custody or visitation (Harris et al., 2009).

Class and Parental Ties

Taken together, these provisions for establishing parenthood parallel the class-based changes in marriage. Where couples marry before they have children, the mother effectively consents to the father's continuing involvement in the child's life. Paternity establishment follows automatically, the father has substantial rights over the child that the courts allocate automatically in the event of divorce, and fathers in fact invest more time and money in their children.

Where the parents sign a voluntary acknowledgment at the time of the child's birth, the parents typically expect mutual involvement in the child's life and, in fact, these fathers are more likely than the fathers who do not sign voluntary acknowledgments to continue to see the child after a break-up with the mother and to pay support voluntarily (McLanahan, 2004). While voluntary acknowledgments establish paternal rights identical to those established through marriages or judgments, they are less likely to lead to custodial orders granting the child time with the father. In the absence of marriage, the couple may go their separate ways without any involvement with the courts and without any allocation of custody rights. Instead, they tend to end up in court only if the custodial parent seeks support or if the noncustodial parent seeks a share of the child's time. Particularly if the child receives state benefits, the state may help with child support enforcement, but there is rarely much if any assistance in securing custodial rights. Going to court may be costly and enforcing a subsequent judgment difficult.

The same factors hold with even greater force to those who do not sign voluntary acknowledgments. If the mother refuses to acknowledge the father, the father must bring a court action

that will turn on a paternity test. The tests can cost several hundred dollars in themselves.

Custody statutes no longer distinguish between married and unmarried fathers, and legal parents who go to court can expect to receive custodial rights. The likelihood of doing so nonetheless corresponds with marriage – married fathers are dramatically more likely to have a custodial order giving them a share of the child's time than unmarried fathers, and those with voluntary acknowledgments are more likely to have such orders than those with paternity judgments (Brown and Cook, 2012). In addition, while the state facilitates child support enforcement, state-initiated actions have proven to be counterproductive in encouraging greater paternal involvement. Instead, fathers are more likely to remain involved with their children through cooperation with the mother; involuntary enforcement actions often increase the conflicts underlying such relationships and undercut the relationship with the child (Harris, 2011). Racial differences in relationship stability narrow or disappear once the ratio of men to women is taken into account; cultural differences, however, affect continuing contact with the child. The involvement of white fathers continues to depend more on a continuing relationship with the mother while African-American fathers' relationship with the child is more likely to be independent of adult ties, though the differences between the two groups are narrowing.

A contentious issue underlying parentage is the continuing status of the marital presumption. Courts in Texas and Kentucky acknowledge that an uncritical application of the marital presumption allows the mother to decide who the father will be and, indeed, in some Kentucky cases the mother appears to have married or remained married primarily to cut the biological father out of the child's life (Carbone and Cahn, 2014). These states have made it easier for biological fathers to challenge the presumption. In contrast, Utah and Louisiana tend to apply the presumption primarily to reward faithful husbands over 'adulterous' fathers, but have made it easier for husbands to challenge paternity at or after divorce. California, on the other hand, acknowledges that the marital presumption gives the mother effective decision-making power, but continues to apply the presumption in the context of a preference for upholding the parental rights of those who have assumed parental functions. A few states have begun to recognize more than two parents, including Louisiana and California, but it remains to be seen how these developments will affect recognition of step-parents.

All of these decisions involve a regulation of the mother's decision-making power. With marriage, the mother can be said to have consented to her husband's parental role. Giving voluntary acknowledgments, she consents to paternity, but without any necessary understandings about the terms of the relationship. Where paternity has not been established, the mother's decision to end a relationship typically ends the father's involvement, unless he satisfies her terms for continuing involvement with the child (Edin and Nelson, 2013). The California cases accept the mother's ability to choose the second parent, but limit her ability to change her mind later. The Texas and Kentucky decisions subject the mother to the biological father's ability to demand a role in the child's life – if he can succeed in getting to court and enforcing a judgment.

The irony is that these developments simultaneously make formal family law more supportive of fathers' rights and make the ability to enforce them more dependent on fathers' income and status. The better educated male cohabitants who express skepticism about their partners do so because elite men have become pickier about their choice of spouses, but the law enforces their rights once they do marry. Less educated women have become warier about marriage because they enjoy a smaller choice of reliable men in the context of more interdependent relationships. Less elite men report, in turn, that "[i]n a startling reversal of the way gender typically operates in American society, unwed childbearing seems to offer mom, and not dad, all the power: 'it's her way or the highway,' in the words of one father" (Edin and Nelson, 2013).

Conclusion

Were this simply a matter of personal choice, the consequences might be termed "postmodern family" law (Stark, 2001) and the creation of greater family variety celebrated as personal expression. The class-based nature of these changes, however, should be cause for concern as they skew the allocation of resources to the next generation. Class-based differences in family resources have grown dramatically. In 1970, for example, college graduate and high school graduate parents spent about the same amount of interactive time with their children. By 2000, however, college graduate parents were spending dramatically more time with their children than the less educated. In 1970, parental education made a difference in children's performance, but income-based differences were relatively small. In contrast, today income differences have grown dramatically even among parents with the same education, producing even greater gaps in cognitive performance and children's well-being (Reardon, 2011).

The state has responded to these developments by making them worse. State policies increase the advantages of those with the most income and the neoliberal state has reduced rather than increased the assistance to low income families. The result recreates reduced social mobility and threatens to impoverish the lives of the next generation.

See also: Children and Law; Empirical Legal Studies; Family Theory: Economics of Marriage and Divorce; Family as Institution; Family: The Anthropology of the Concept and Its History; Gender and the Law; Property: Legal Aspects of Intergenerational Transmission.

Bibliography

Amato, Paul, Booth, Alan, Johnson, David R., Rogers, Stacy J., 2009. Alone Together: How Marriage in America is Changing. Harvard University Press, Cambridge, MA.

Baker, Katharine K., 2010. The stories of marriage. Journal of Law and Family Studies 12 (1).

Bowman, Cynthia G., 2010. Unmarried Couples, Law, and Public Policy. Oxford University Press, New York.

Braman, Donald, 2007. Doing Time on the Outside: Incarceration and Family Life in Urban America. University of Michigan Press, Ann Arbor, MI.

Brinig, Margaret F., Allen, Douglas W., 2000. "These boots are made for walking": why most divorce filers are women. American Law Economics Review 2, 126, 128, Tbl. 1, 136–137.

Brown, Patricia, Cook, Steven T., 2012. Children's Placement Arrangements in Divorce and Paternity Cases in Wisconsin. Institute for Research on Poverty Report. September 2011 (rev. November 2012). www.irp.wisc.edu/research/childsup/cspolicy/pdfs/2009-11/Task4A_CS_09-11_Final_revi2012.pdf.

Burton, Linda M., Cherlin, Andrew, Winn, Donna-Marie, Estacion, Angela, Holder-Taylor, Clara, 2009. The role of trust in low-income mothers' intimate unions. Journal of Marriage and Family 71 (5), 1107–1124. www.ncbi.nlm.nih.gov/pmc/articles/PMC2788951/pdf/nihms151618.pdf.

Cahn, Naomi, Carbone, June, 2010. Red Families v. Blue Families: Legal Polarization and the Creation of Culture. Oxford University Press, New York.

Carbone, June, 2000. From Partners to Parents: The Second Revolution in Family Law. Columbia University Press, New York.

Carbone, June, Cahn, Naomi, 2011. Marriage, parents and child support. Family Law Quarterly 45, 219.

Carbone, June, Cahn, Naomi, 2014. Marriage Markets: How Inequality is Remaking the American Family. Oxford University Press, New York.

Cherlin, Andrew J., 2004. The deinstitutionalization of American marriage. Journal of Marriage and Family 66 (4), 848–861.

Cherlin, Andrew, 2009. The Marriage Go-round: The State of Marriage and Family in America Today. Vintage Books, New York.

Coontz, Stephanie, 2006. Marriage, a History: From Obedience to Intimacy, or How Love Conquered Marriage. Viking, New York.

Edin, Kathryn, Nelson, Timothy J., 2013. Doing the Best I Can: Fatherhood in the Inner City. University of California Press, Berkeley, CA.

Fineman, Martha A., 1991. The Illusion of Equality: The Rhetoric and Reality of Divorce Reform. University of Chicago Press, Chicago.

Fineman, Martha A., 1995. The Neutered Mother the Sexual Family, and Other Twentieth Century Tragedies. Routledge, New York.

Fineman, Martha A., 2004. The Autonomy Myth: A Theory of Dependency. The New Press.

Glynn, Sarah J., 2010. The New Breadwinners, 2010 Update. Center for American Progress. http://www.americanprogress.org/issues/2012/04/epd_breadwinners.html.

Goldin, Claudia, Katz, Lawerence, 2002. The power of the pill: oral contraceptives and women's career and marriage. Journal of Political Economy 110 (4), 730–770.

Grossberg, Michael, 1985. Governing the Hearth: Law and Family in Nineteenth-Century America. University of North Carolina Press, Chapel Hill and London.

Harris, Leslie J., 2011. Questioning child support enforcement policy for poor families. Family Law Quarterly 45, 157.

Harris, Leslie Joan, Carbone, June, Teitelbaum, Lee E., 2009. Family Law, fourth ed. Aspen, New York.

Hegel, Georg W.F., 1952. Philosophy of Right (1821) (T.M. Knox, Trans.). Oxford University Press, London.

Hymowitz, Kay, Carroll, Jason S., Wilcox, W. Bradford, Kaye, Kelleen, 2013. Knot Yet: The Benefits and Costs of Delayed Marriage in America. National Marriage Project Report, 8. http://nationalmarriageproject.org/wp-content/uploads/2013/03/KnotYet-FinalForWeb.pdf.

Jacob, Herbert, 1988. Silent Revolution: The Transformation of Divorce Law in the United States. University of Chicago Press, Chicago.

Kant, Immanuel, 1887. The Philosophy of Law: An Exposition of the Fundamental Principles of Jurisprudence as the Science of Right (W. Hastie, Trans.), T. & T. Clark, Edinburgh, xxxvi, 265 pp.

Kristol, William, 1991. Women's liberation: the relevance of Tocqueville. In: Masugi, Ken (Ed.), Interpreting Tocqueville's Democracy in America, 480, p. 491.

Lasch, Christopher, 1995. Haven in a Heartless World: The Family Besieged. W.W. Norton & Company, New York.

Looney, Adam, Greenstone, Michael, February 2012. The Marriage Gap: The Impact of Economic and Technical Change in Marriage Rates. Hamilton Project Paper, Brookings Institution. www.hamiltonproject.org/papers/the_marriage_gap_the_impact_of_economic_and_technological_change_on_ma.

McLanahan, Sara, 2004. Diverging destinies: how children are faring under the second demographic transition. Demography 41 (4), 607–622.

Meier, Joan S., 2003. A domestic violence, child custody, and child protection: understanding judicial resistance and imagining the solutions. American University Journal of Gender, Social Policy & the Law 11, 657.

Mnookin, Robert, Kornhauser, Lewis, 1979. Bargaining in the shadow of the law: the case of divorce. Yale Law Journal 88, 950.

Murray, Charles, 2013. Coming Apart: The State of White America, 1960–2010. Random House, New York.

Okin, Susan M., 1991. Gender, Justice and the Family. Basic Books, New York.

Rauch, Jonathan, 2010. Red families, blue families, gay families, and the search for a new normal. Law and Inequality 28, 333–344.

Reardon, Sean F., 2011. The widening academic achievement gap between the rich and the poor: new evidence and possible explanations. In: Duncan, Greg J., Murnane, Richard J. (Eds.), Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances. Sage, New York, pp. 91–116.

Regnerus, Mark, Uecker, Jeremy, 2011. Premarital Sex in America: How Young Americans Meet, Mate, and Think about Marrying. Oxford University Press, New York.

Rich, Adrienne, 1980. Compulsory heterosexuality and lesbian existence. Signs 5, 631–654.

Schneider, Carl, 1992. The channelling function in family law. Hofstra Law Review 20, 495.

Schneider, Daniel, 2011. Market earnings and household work: new tests of gender performance theory. Journal of Marriage and Family 73, 845.

Schwartz, Christine R., 2013. Trends and variation in assortative mating: causes and consequences. Annual Review of Sociology 39, 451–460.

Solinger, Rickie, 1992. Wake Up Little Susie: Single Pregnancy and Race before Roe v. Wade. Routledge, New York.

Stark, Barbara, 2001. Marriage proposals: from one-size-fits-all to postmodern marriage law. California Law Review 89, 1479–1548.

Stone, Emily A., Shackelford, Todd K., Buss, David M., 2007. Sex ratio and mate preferences: a cross-cultural investigation. European Journal of Social Psychology 37, 288–294.

Stone, Lawrence, 1977. The Family, Sex and Marriage in England, 1500–1800. Harper & Row, London.

Tenbroek, Jacobus, 1965. California's dual system of family law: its origin, development, and present status, 16 Stanford Law Review 257 (1964), 16 Stanford Law Review 900 (1964). Stanford Law Review 17 (pts. I, II, & II), 614.

When marriage disappears: the new middle America. In: Wilcox, W.B., Marquardt, E. (Eds.), National Marriage Project and Institute for American Values. http://stateofourunions.org/2010/SOOU2010.pdf.

Wilson, William J., 1987. The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy. University of Chicago Press, Chicago.

Younger, Judith T., 2007. Lovers' contracts in the courts: forsaking the minimum decencies. William and Mary Journal of Women and the Law 13, 349.

Family Planning: Program Effects

Grant Miller, Stanford University, Stanford, CA, USA
Kimberly S Babiarz, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews empirical evidence on the microlevel consequences of family planning programs in middle- and low-income countries. In doing so, it focuses on fertility outcomes (the number and timing of births), women's health and socioeconomic outcomes (mortality, human capital, and labor force participation), and children's health and socioeconomic outcomes throughout the life cycle. Though effect sizes are heterogeneous, long-term studies imply that, in practice, family planning programs may only explain a small share of fertility decline in real-world settings. Family planning programs may also have quantitatively modest – but practically meaningful – effects on the socioeconomic welfare of individuals and families.

Introduction

The twentieth century witnessed the birth of modern 'family planning' and its evolution from an early target of antiobscenity laws (Comstock, 1873) to a focus of global efforts to improve human welfare (Cleland et al., 2006). Many observers date the modern family planning movement to the famous 1916 opening of New York City's first birth control clinic. Mid-century, the International Planned Parenthood Federation and the Population Council were instrumental in both the development of modern contraceptive technologies (including the first oral contraceptive, Enovid, and plastic intrauterine devices) and in promoting their widespread distribution, ushering in the 'modern' era of fertility control.

Heightened concern about world's unprecedented rate of population growth (and its potential macroeconomic and environmental consequences) also arose mid-century (Coale and Hoover, 1959) – and Neo-Malthusian alarm grew in tenor with Paul Ehrlich's publication of *The Population Bomb* in 1968. In response, aid agencies including USAID and the World Bank joined with the Population Council, the International Planned Parenthood Foundation (IPPF), and others in supporting the establishment of large-scale family planning programs worldwide. Global funding for family planning tripled during the 1970s and early 1980s – and by the mid-1990s, large-scale family planning programs were active in 115 countries (Cleland et al., 2006). Remarkably, the total fertility rate in developing countries also fell by more than half over this period (Sinding, 2007).

Academics have long debated the primary forces responsible for global fertility decline – and in particular, the contribution made by family planning (Kaiser, 2011). One view largely credits family planning programs directly (Bongaarts et al., 1990; Robey et al., 1993). Others give them less credit, arguing that fertility decline is a response to economic development and industrialization, even if aided by family planning programs (Becker, 1960; Davis, 1967; Voigtlander and Voth, 2013). A third view, promoted in part by the European Fertility Project (although its methodology has been debated (Brown and Guinnane, 2003)), emphasizes the importance of social networks that facilitated the diffusion of 'cultural innovations' including the social acceptability of small families (Coale and Watkins, 1986). Finally, some propose that falling mortality rates play a central role (Palloni and Rafalimanana, 1999).

As dire predictions about Malthusian demographic catastrophe failed to materialize (Lam, 2011), population policy became more nuanced during the 1980s and 1990s (Kelley, 2003). (Recognition of the importance of population age structure, not just population size and growth, grew during this period, and includes work on the 'demographic dividend' (Bloom and Williamson, 1998; Bloom et al., 2003; Bloom and Canning, 2009). Stances on the relationship between population growth and economic development became more neutral (Finkle and Crane, 1985), changing the motivation for family planning programs. At the 1994 International Conference on Population and Development in Cairo, macroeconomic rationales for population policy were replaced with feminist, rights-based frameworks. Family planning programs were recast to emphasize individual needs and desires, stressing socioeconomic and health benefits for women their families (Glasier et al., 2006). Also, as the definition of reproductive health expanded, so did the scope of family planning programs - growing to encompass safe delivery, prevention and treatment of sexually-transmitted infections, safe abortion, and violence against women (Greene and Merrick, 2005).

In the spirit of contemporary emphasis on broad, individual- level benefits, this article reviews empirical evidence on the microlevel consequences of family planning programs in middle- and low-income countries. In doing so, it focuses on three types of outcomes: (1) fertility outcomes (the number and timing of births), (2) women's health and socioeconomic outcomes (mortality, human capital, and labor force participation), and (3) children's health and socioeconomic outcomes throughout the life cycle. Importantly, we omit studies that simply estimate the effect of family planning programs on contraceptive use and knowledge, and we also omit studies that simply relate the number, timing, and spacing of births to downstream socioeconomic outcomes (given debate about the influence of family planning programs on fertility).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31036-4

The remainder of this article is organized as follows. Section Background Controversy: Supply vs Demand and the Pritchett (1994) Critique summarizes background debate about the relative importance of supply- vs demand-side factors in explaining fertility decline. The third section reviews empirical studies of the relationship between family planning programs and fertility outcomes, while the fourth covers program effects on health and socioeconomic outcomes among women and their children. Pooling across studies, the fifth section concludes by summarizing ranges of estimates that emerge from the empirical literature. Overall, family planning programs may explain only a small share of the fertility decline in real-world settings, but it may also have meaningful effects on the socioeconomic welfare of individuals and families.

Background Controversy: Supply vs Demand and the Pritchett's (1994) Critique

Among the competing explanations for fertility decline described above, an important distinction emerges between demand- and supply-side factors. While there is little debate about the association between use of modern contraceptives and fertility, the underlying determinants of contraceptive use are controversial. The crux of the debate pits the relative importance of changes in the demand for children (and for contraceptives) against changes in the availability and price of contraceptives (generally accomplished through family planning programs). Before reviewing microlevel empirical studies of family planning in Sections Family Planning Programs and the Number, Timing, and Spacing of Births and Family Planning Programs and Health and Socio-Economic Outcomes among Women and Children, we first summarize this controversy.

Lant Pritchett's 1994 article entitled "Desired Fertility and the Impact of Population Policies" published in *Population and Development Review* is a landmark paper in this debate. The paper challenges conventional wisdom about the centrality of contraceptive supply in determining total fertility rates in developing countries (Bongaarts et al., 1990). Pritchett begins by deconstructing the conceptual fallacy equating contraceptive prevalence (or use) to contraceptive supply (or availability). Because use is jointly determined by both supply and demand (and not by supply alone), the close relationship observed between a country's contraceptive prevalence rate and its total fertility rate fails to isolate the importance of either demand- or supply-side factors.

Across three different measures of desired fertility (or the demand for children), the article establishes a nearly one-to-one correspondence between a country's total fertility rate and the number of children that women report wanting (regression coefficients range from 0.78 to 0.95 – with a positive constant of about 1–1.5). Pritchett argues that these relationships demonstrate the primacy of demand and that contraceptive supply has little influence on a country's total fertility rate. He then addresses concerns about *ex post* rationalization and about the dependence of desired fertility on contraceptive availability using two related instrumental variables (IVs) methodologies. First, he uses the proportion of women with four children who desire no more (a forward-looking statistic free of ex post rationalization) to instrument for a retrospective measure of the demand for children fertility (DTFR). Second, he uses a retrospective measure of demand (DTFR) to instrument for a forward-looking measure (WTFR). The resulting IVs estimates are again close to one, which Pritchett interprets to strengthen the conclusion that women are essentially able to have the number of children that they want – and contraceptive supply can at best explain 10% of the cross-country variation in fertility rates.

Pritchett's IVs approach requires two important assumptions that we highlight to meet the necessary 'exclusion restriction' (Angrist and Kreuger, 1991). The first is that the share of women in a country who have a given number of children and report not wanting more is unrelated to the country's total fertility rate except through desired fertility (demand). When answering questions about wanting more children, if women have in mind the costs of fertility control that they actually face in practice (despite survey instructions to the contrary - Pritchett argues that DHS surveys aim to elicit information about desired fertility at zero contraceptive cost), this assumption would be violated. However, these costs may be difficult for respondents to disregard – the interpersonal costs that women face in bargaining over fertility with their partners, for example. The second related assumption is that past and current/future costs of fertility control are uncorrelated. Because such a correlation is plausible, current/future costs of fertility control could in fact be related to past fertility decisions. Ultimately, these assumptions are not directly testable.

On a more general level, Pritchett argues that the financial cost of modern contraceptives is too small relative to the importance of fertility decisions to exert meaningful influence. However, these costs may nonetheless be prohibitive for poor households facing tight credit or liquidity constraints, and importantly, the total cost of fertility control extends far beyond the financial cost of contraceptives – bargaining with spouses over fertility and direct disutility from using of contraceptive devices, for example – and some empirical evidence suggests that these costs may be substantial. On bargaining, Ashraf et al. (2014) show that among the subset of couples with different fertility preferences, offering vouchers for contraceptives to women privately is significantly more effective than offering vouchers to them in the presence of their husbands (reducing unwanted births by 57%). On the direct disutility of modern contraceptives, research on commercial sex markets suggests that men are willing to pay 23% more for sex without a condom (and double that for unprotected sex with attractive commercial sex workers) (Gertler et al., 2005).

Despite some unresolved concerns with Pritchett's (1994) analysis, its overarching point – that the demand for children may be more important than supply-side factors (including contraceptive supply) in determining fertility – may nonetheless stand. Important determinants of the demand for children studied in the empirical literature on fertility (which are not independent of each other) include infant and child mortality rates (Angeles, 2010; Kalemli-Ozcan, 2002; Schultz, 1985), the opportunity cost of time (both generally and for women) (Breierova and Duflo, 2004; Lavy and Zablotsky, 2011; Schultz, 1985), anticipated increases in future demand for human

capital (Galor and Weil, 2000), and women's bargaining power within households (Rasul, 2008).

Family Planning Programs and the Number, Timing, and Spacing of Births

The empirical literature on family planning is vast. To aid in this article, we used a strategy akin to the 'systematic review' methodology in the biomedical sciences, although we emphasize that we did not undertake a formal systematic review. Using explicit criteria to search four major databases of indexed journals (together with a 'snowball' approach), our initial search yielded 9501 studies for consideration. We then reviewed abstracts, applying inclusion criteria for methodological rigor, direct micro-level analysis of family planning program effects in developing countries, and an explicit focus on at least one of our three types of primary outcomes. For nonexperimental (observational) studies, we restricted this article to those using commonly accepted methodologies for addressing the role of confounding factors and endogenous program placement/intensity. These include (but are not limited to) conventional panel data techniques, difference-in-difference estimation, and IVs strategies using plausibly exogenous instruments. Our final list includes 28 studies meeting these criteria (see the accompanying electronic appendix (available at www.stanford.edu/~ngmiller) for details).

Randomized Controlled Trials

Arguably, the most prominent study of family planning conducted to date is the famous Matlab family planning experiment. Beginning in 1977, the experiment randomly assigned 141 villages in Matlab, Bangladesh to either the study's 'treatment' arm (70) or a control arm (71). In treatment villages, female reproductive health workers visited the homes of married women of childbearing age every 2 weeks to educate women about reproductive health, provide nutrition counseling, and offer modern contraceptives free of charge. Maternal and child health services were integrated into the experimental treatment in 1982 (Muhuri, 1995) and safe motherhood services followed (Chowdhury et al., 2009).

The Matlab experiment has produced a large volume of academic papers. Early papers report a 25% reduction in the general fertility rate over the first 2 years of the experiment (Phillips et al., 1982) and these reductions lasted throughout the 1980s (Foster and Roy, 1997; Koenig et al., 1992). Longer term follow-up suggests that these fertility effects persisted for at least 20 years, reducing the number of children ever born by 1–1.5 children and extending birth intervals by 8–13 months (Joshi and Schultz, 2007). Estimates of implied lifetime fertility reductions range from 13% (Sinha, 2005) to 23% (Joshi and Schultz, 2013).

The Matlab experiment has also generated debate. First, critics note that the intensity and expense of the program would be unrealistic on a large scale. In its first 10 years, program costs were approximately $180 per birth averted (120% of Bangladesh's GDP per capita in 1987) (Simmons et al., 1991) – nearly 10 times more than mean family planning spending in developing countries at the time (Pritchett, 1994). Second, bundling child health services into the family planning treatment effectively reduced the 'price' of child survival – and could have therefore exerted independent influence on fertility (Becker, 1991; Caldwell, 1976; Sah, 1991). Third, Schultz (2009) suggests that possible deviations from true randomization in the experiment's implementation.

There are few other randomized controlled trials of family planning services. The Navrongo experiment in Ghana studied 37 communities randomized across four treatment arms (Binka et al., 1995; Phillips et al., 2006). Although balance across arms was not achieved, treatment arms combining family planning service training and community outreach were associated with a 15% reduction in the total fertility rate among married women (Debpuur et al., 2002; Phillips et al., 2006). Effects on parity progression persisted for 15 years (and have largely been attributed to the promotion of contraceptive use through community organizations) but have declined over time (Phillips et al., 2006, 2012). Two other experiments integrated family planning services into existing programs: a microcredit program in Ethiopia (Desai and Tarozzi, 2011) and an HIV treatment program in Kenya (Kosgei et al., 2011). Both experiments find null results in the first few years following implementation.

Observational Studies

Two influential observational studies analyze contemporaneous effects of Indonesian National Family Planning Coordinating Board's programs during the late 1970s and 1980s. Gertler and Molyneaux (1994) combine birth history data with community-level information about family planning program activity to study program effects on community-level birth hazards between 1982 and 1987. Accounting for changes in the demand for children and community fixed effects, the authors show that family planning programs account for just 4–8% of Indonesia's decline in fertility (a 1–2% reduction in the monthly birth hazard) during the study period (Gertler and Molyneaux, 1994). Relatedly, using both survey and population census data and explicitly modeling village-level family planning program placement, (Pitt et al., 1993) fail to find that the program had any significant effect on fertility between 1976 and 1986.

Also estimating short-run program effects, two additional studies use strategies that exploit program variation due to idiosyncratic public finance rules governing local government resource allocation. Specifically, Molyneaux and Gertler (2000) find that family planning subsidies were associated with 3–6% lower quarterly birth hazards in Indonesia between 1985 and 1994. Studying family planning in Ethiopia, Portner et al. (2011) find that among uneducated women, the presence of a family planning program in a woman's 1990 district of residence was associated with roughly one fewer child ever born in the 1994 Ethiopian population census (a reduction of about 20% in this subgroup).

Other studies examine the cumulative effects of family planning programs over longer periods of time. To study the long-run effects of PROFAMILIA, Colombia's predominant family planning provider from 1965 until the 1980s, Miller (2010) combines 1985 and 1993 population census data with information about its staggered geographic expansion

across the country to estimate changes in fertility outcomes associated with age-specific program exposure. Controlling for cohort and birth area fixed effects as well as area-specific time trends, he finds that the presence of family planning services explains about 6–7% of Colombia's fertility decline over this period (postponing first births and reducing completed fertility by approximately 0.25–0.33 children).

Three studies by Angeles, Guilkey, and Mroz also use spatial and temporal variation in program exposure to estimate long-run effects on cumulative fertility. In Tanzania and Peru, the authors use random effects models to estimate the relationship between community-level program placement and individual birth hazards over two decades. They find 25–35% reductions in children ever born in Peru (Angeles et al., 2005a) and 10–20% reductions in Tanzania, reporting that half of these effects are linked to program exposure during adolescence (Angeles et al., 1998). In Indonesia, the authors model both indirect and direct pathways through which family planning might influence a woman's lifetime fertility. Jointly estimating equations for educational attainment, marriage, and fertility, they find that the long-term presence of family planning programs is associated with a 20% reduction in completed fertility (or about one child) (Angeles et al., 2005b).

Another set of studies examines the role of family planning in Iran's striking fertility decline during the 1980s and 1990s (as Iran's total fertility rate declined from over 7 births per woman to about 2 (Hashemi and Salehi-Isfahani, 2013)). Family planning was banned at the time of the Islamic Revolution, but in 1989, Iran abruptly legalized modern contraceptives and launched family planning programs through its large, preexisting cadre of community health workers (Behvarz). Salehi-Isfahani et al. (2010) combine data from the Iranian Ministry of Health with population census records from 1986 to 1996; using a difference-in-difference approach to study variation in the timing and location of community health worker activity, they find a 4% decline in the child to woman ratio associated with program activity, explaining 8–20% of the decline during the study period. Analyzing age-specific program exposure among cohorts of women in the 2006 Iranian census, Modrek and Ghobadi (2011) find an 18% reduction in children ever born among women first exposed to family planning between 20 and 34 years of age and a 28% reduction among women first exposed between 15 and 19 years of age. Using a similar approach, Hashemi and Salehi-Isfahani (2013) report that the presence of a program was associated with a 5–7% increase in birth spacing.

Finally, a recent group of studies examines the short-run consequences of unanticipated 'shocks' to family planning programs. We note that the interpretation of these papers' estimates is complex because longer-term fertility behavior (which may offset the consequences of short-lived shocks) is not observed. In Eastern Europe, where abortion was historically the predominant form of fertility regulation, many countries limited or banned abortion during the 1980s and 1990s. Examining the abrupt end of Romania's ban on abortion and family planning in 1989, Pop-Eleches (2010) uses survey data on births among reproductive age women 2 years before and after the end of the ban to employ a single difference approach (akin to a regression discontinuity design). The author finds that lifting the ban was associated with a 30% decline in fertility – and conducting simulations using his estimates, he suggests that the cumulative effect of the abortion ban was an increase of approximately 0.5 children (a 25% increase) among women who spent all of their reproductive years under the ban (Pop-Eleches, 2010).

Studying disruptions in the supply of free condoms to Filipino provinces due to supply chain irregularities, Salas (2013) finds that a 6% reduction in contraceptive supply is associated with a 15% increase in short-run birth hazard. Similarly, analyzing fluctuations in contraceptive supply in Ghana due to the instatement and repeal of the United States' "Mexico City Policy," Jones (2013) finds that the policy led to a 10% increase in pregnancies (or a 0.2 percentage point increase in the monthly pregnancy hazard) among rural women and women with little education.

Family Planning Programs and Health and Socioeconomic Outcomes among Women and Children

Beyond their impact on fertility, family planning programs may also have important health and socioeconomic benefits for mothers and their children (Canning and Schultz, 2012). Intuitively, these benefits could stem from changes in the number, timing, and spacing of births or from changes in sibship size and composition – but isolating the specific mechanisms or pathways through which they are produced is often not possible. Studying these benefits is also difficult because some do not emerge for many years, requiring long study periods.

Benefits among Mothers

Given high maternal mortality rates in many low- and middle-income countries, reductions in the number of births may mechanically reduce maternal mortality rates simply by reducing the number of times women are at risk of maternity-related death (Rahman and Menken, 2012). However, if family planning programs also selectively reduce the relative incidence of riskier pregnancies (such as higher parity births), they may also reduce the maternal mortality ratio (Cleland et al., 2012; Jain, 2011; Winikoff and Sullivan, 1987).

Few studies analyze the direct relationship between family planning programs and maternal mortality. One exception is an early study of the Matlab experiment reporting that the maternal mortality rate in treatment areas declined to about half of the rate in control areas between 1976 and 1985 (Koenig et al., 1988). However, this treatment effect was due to relative declines in fertility – there was no change in the maternal mortality ratio (and hence no change in average mortality risk conditional on pregnancy).

Among surviving women, long-term studies of the Matlab experiment find anthropometric gains associated with the treatment. Canning and Schultz (2012) report that in 1996, reproductive age women in treatment areas had a 1 kg m^{-2} higher Body Mass Index (BMI) than women in control areas. Adjusting for characteristics of women and their households,

Joshi and Schultz (2013) find that this BMI advantage among women in treatment villages is only present at older ages. However, they also find that women in treatment areas are roughly 2 kg heavier on average and are less likely to be underweight (defined as BMI<18 kg m^{-2}). In interpreting these findings, we emphasize that improvements in women's health could be due to bundling of women's health services into the family planning 'treatment' beginning in 1982.

By allowing women greater control over the number, timing, and spacing of births, family planning programs may also influence their educational attainment, labor force participation, and lifetime earnings (Greene and Merrick, 2005). Studying Colombia, Miller (2010) reports that the presence of a local family planning program early in women's reproductive years is associated with a 1% increase in educational attainment (0.05 years of schooling) and a 4–7% increase in the likelihood formal sector employment (1–2 percentage points). Using simulations, Angeles et al. (2005b) report larger effects on educational attainment in Indonesia, suggesting that lifetime exposure to family planning programs is associated with gains of 25–27% (or an additional 1.3 years of schooling).

Benefits among Children

Family planning programs could also influence children's health and socioeconomic outcomes through a variety of mechanisms (through changes in birth spacing and sibship size or through gains in women's educational attainment, for example). Although a large literature links various dimensions of fertility to child welfare (DaVanzo et al., 2005; Molyneaux and Gertler, 2000; Rutstein, 2005), few studies focus directly on family planning programs.

Studying child survival, two experiments find that family planning programs are associated with substantial reductions in child mortality under 5 years of age (decreases between 30% and 50%) (Joshi and Schultz, 2007; 2013; Phillips et al., 2006). However, we emphasize that these programs bundled family planning services together with other health services – and, in particular, ones targeting infant and child health, making it difficult to isolate the contribution of family planning services per se. Focusing more directly on family planning services alone, Miller (2010) and Portner et al. (2011) report no significant program effects on infant and child mortality.

Several other studies examine anthropometric measures of child health. Joshi and Schultz (2007) report that girls in Matlab's treatment villages have BMI z-scores that are 0.4 standard deviations larger on average than those of their control village counterparts. Studying children in the Philippines, Rosenzweig and Wolpin (1986) use longitudinal data to estimate within-child changes in anthropometrics associated with the share of a child's life for which family planning services were available (accounting for both regional and family characteristics potentially correlated with birth spacing). They find that lifetime exposure to family planning is associated with a 7% gain in child height and a 12% gain in child weight.

Beyond child health, family planning may benefit children in other ways throughout their lifetime, too – by increasing parents' investments in their education, for example (Becker, 1991). Evidence on the direct relationship between family planning programs and children's educational attainment is again thin, however, and limited to long-term studies of the Matlab experiment. Foster and Roy (1997) report that the Matlab treatment increased both girls' and boys' years of completed schooling (by 6–15% and 5–12%, respectively), an effect they link to reductions in sibship size. Joshi and Schultz (2007) find that relative to their peers in control villages, gains in educational attainment of 0.5 standard deviations persisted among boys (but not girls) in treatment villages.

Finally, a study of Romania's 1966 ban of abortion and family planning services compares cohorts born just after the ban (which include additional 'unwanted' children) with cohorts born just before it (Pop-Eleches, 2006). Controlling for family characteristics, children born after the ban were less likely to complete either high school or college (by 1.7% and 1.4%, respectively) and were instead 2.1% more likely to complete a lower-return vocational degree. This article attributes these changes in educational outcomes not only to the 'wantedness' of children, but also to crowding in schools among post-ban birth cohorts (the role of sibship size is not explicitly studied) (Pop-Eleches, 2006).

Conclusion

This article reviews empirical studies of family planning programs in developing countries, focusing on fertility and on the health and socioeconomic welfare of women and children. Because the family planning literature is vast, we used an informal approach akin to the 'systematic review' methodology in the biomedical sciences in an effort both to be comprehensive and to explicitly define our focus.

Pooling across studies, we propose several generalizations about family planning program effects. First, we find that program effects on fertility vary between 5% and 35% fewer children ever born and 5–7% longer birth spacing (see **Table 1** for a summary). Estimates from the famous Matlab experiment generally fall in the upper-end of this range – but we note that the experimental intervention is likely too intensive to be replicated on large scale, limiting the generalizability of its results – and they also presumably reflect reductions attributable to other health services bundled into the experimental treatment. Long-term studies of real-world programs yield more modest effect sizes, implying that, in practice, family planning programs may only explain about 8% of fertility decline in real-world settings. Overall, it is difficult to argue that family planning programs alone explain a large share of the fertility decline in developing countries over the past half century. Demand-side factors may play a more important role (Gertler and Molyneaux, 1994; Pritchett, 1994).

Second, we find evidence that family planning programs may have quantitatively modest – but practically meaningful – effects on the socioeconomic welfare of individuals and families. Excluding studies of family planning programs that bundle maternal and child health services into their service packages, there is little evidence about program effects on maternal and child health outcomes. However, long-term studies of socioeconomic outcomes suggest that family planning programs may raise educational attainment among

Table 1 Estimates of the effect of family planning programs of fertility outcomes.

Country	*Study period*	*Fertility measure*	*Marginal or incremental effect size*	*Relative effect size*	*Background fertility change during study period*	*Percent change explained*	*Citation*	*Notes*
Bangladesh	1974–80	General fertility rate	1.8	25%			Phillips et al., 1982	
Bangladesh	1967–90	Children 0–8	0.41	25%	25%		Foster and Roy, 1997	
Bangladesh	1974–96	Children ever born	1–1.5	15%			Joshi and Schultz, 2007	
Bangladesh	1974–96	Lifetime fertility	0.5	13%	40%	33%	Sinha, 2005	Calculated as the difference in lifetime fertility among women in treatment areas before and after treatment (7.21 vs 4.29)
Bangladesh	1974–96	Child to woman ratio	0.12	17%	46%	37%	Joshi and Schultz, 2007	Roughly calculated as the average change in fertility/1974 fertility level
Bangladesh	1974–96	Children ever born	1	23%			Joshi and Schultz, 2013	Calculated as the marginal change in children ever born/ overall sample mean children ever born
Colombia	1965–93	Children ever born	0.25–0.33	5%	51%	6-7%	Miller, 2010	
Ethiopia	2003–06	Total births during study period	–	–	–	–	Desai and Tarozzi, 2011	Results on total births during the 3-year study period are mixed, and in some cases point estimates are positive.
Ethiopia	1990–2005	Children ever born	0–0.9	0–20%			Portner et al., 2011	Effect size varies by educational attainment with effects concentrated among uneducated mothers
Ghana	1996–2003	Total fertility rate	1	15%			Debpuur et al., 2002, Phillips et al., 2006	
Ghana	1995–2010	Total fertility rate	0.42	42%			Phillips et al., 2012	
Indonesia	1970–93	Children ever born	0.9	18%			Angeles et al., 2005a	
Indonesia	1982–87	Quarterly birth hazard	0.07	1.8%	23%	4-8%	Gertler and Molyneaux, 1994	Calculated roughly, using overall 1980–85 fertility declines (declines between 1982 and 1987 not given)
Indonesia	1985–94	Quarterly birth hazard	0.03–0.06				Molyneaux and Gertler, 2000	
Iran	1986–96	Child to woman ratio	0.04–0.08	4–8%	46%	8–20%	Salehi-Isfahani et al., 2010	Calculated using the overall mean child to woman ratio in 1986 (0.95) and 1996 (0.51)
Iran	1970–2004	Children ever born	0.63–0.99	18–28%			Modrek and Ghobadi, 2011	Calculated roughly using overall mean children ever born
Iran	1970–2000	Yearly birth hazard	0.02–0.05				Hashemi and Salehi-Isfahani, 2013	Marginal effect size observed for higher parity births
Kenya	2005–09	Incident pregnancy	–	–	–	–	Kosgei et al., 2011	
Peru	1972–91	Children ever born	0.93–1.3	25–35%			Angeles et al., 2005b	Marginal effect size varies depending on educational attainment
Romania	1988–92	Yearly birth hazard	0.07				Pop-Eleches, 2010	
Romania	1988–92	Total fertility rate	0.39–1.1	28–34%			Pop-Eleches, 2010	Marginal effect size varies depending on educational attainment
Tanzania	1970–91	Children ever born		10–20%			Angeles et al., 1998	

Note: Column 4 reports the change in fertility associated with a one unit change in the relevant measure of family planning program, Column 5 gives the effect size relative to fertility level at baseline, Column 6 reports the overall change in fertility observed during the period of study, and Column 7 reports the implied percent of the overall fertility decline explained by family planning.

women (by 1–30%) and among children (by 5–18%). Although socioeconomic effects at the bottom end of these ranges may seem small, we note that they are not dissimilar in magnitude to gains associated with programs explicitly aiming to boost educational attainment.

Finally, we conclude by highlighting the considerable heterogeneity in the magnitude of program effects in the family planning literature and emphasize the importance of further work to explain them.

Acknowledgments

We are grateful to Colin Boyle, Sepideh Modrek, and T. Paul Schultz for helpful comments and suggestions.

Bibliography

Al, J.F., 2012. Grandmultiparity: a potential risk factor for adverse pregnancy outcomes. Journal of Reproductive Medicine 57, 53–57.

Angeles, G., Dietrich, J., Guilkey, D., Mancini, D., Mroz, T., 2001. A Meta-analysis of the Impact of Family Planning Programs on Fertility Preferences, Contraceptive Method Choice and Fertility. Carolina Population Center, Chapel Hill, NC.

Angeles, G., Guilkey, D.K., Mroz, T.A., 1998. Purposive program placement and the estimation of family planning program effects in Tanzania. Journal of the American Statistical Association 93, 884–899.

Angeles, G., Guilkey, D.K., Mroz, T.A., 2005a. The determinants of fertility in rural Peru: program effects in the early years of the national family planning program. Journal of Population Economics 18, 367–389.

Angeles, G., Guilkey, D.K., Mroz, T.A., 2005b. The effects of education and family planning programs on fertility in Indonesia. Economic Development and Cultural Change 54, 165–201.

Angeles, L., 2010. Demographic transitions: analyzing the effects of mortality on fertility. Journal of Population Economics 23, 99–120.

Angrist, J.D., Krueger, A.B., 1991. Does compulsory school attendance affect schooling and earnings. Quarterly Journal of Economics 106, 979–1014.

Ashraf, N., Field, E., Lee, J., 2014. Household bargaining and excess fertility: an experimental study in Zambia. American Economic Review 104, 2210–2237.

Bailey, M.J., 2006. More power to the pill: the impact of contraceptive freedom on women's life cycle labor supply. Quarterly Journal of Economics 121, 289–320.

Bailey, M.J., Hershbein, B., Miller, A.R., 2012. The opt-in revolution? Contraception and the gender gap in wages. American Economic Journal-applied Economics 4, 225–254.

Becker, G.S., 1960. An economic analysis of fertility. In: NBER (Ed.), Demographic and Economic Change in Developed Countries. Columbia University Press, New York.

Becker, G.S., 1991. A Treatise on the Family. Harvard University Press, Cambridge.

Binka, F.N., Nazzar, A., Phillips, J.F., 1995. The navrongo community-health and family-planning project. Studies in Family Planning 26, 121–139.

Bloom, D.E., Canning, D., 2009. Population, poverty reduction, and the Cairo agenda. In: Reichenbach, L., Roseman, M.J. (Eds.), Reproductive Health and Human Rights: The Way Forward. University of Pennsylvania Press, Philadelphia.

Bloom, D.E., Canning, D., Sevilla, J., 2003. The Demographic Dividend: A New Perspective on the Economic Consequences of Population Change. RAND, Santa Monica, CA.

Bloom, D.E., Williamson, J.G., 1998. Demographic transition and economic miracles in emerging Asia. World Bank Economic Review 12, 419–455.

Bongaarts, J., Mauldin, W.P., Phillips, J.F., 1990. The demographic-impact of family-planning programs. Studies in Family Planning 21, 299–310.

Boserup, E., 1965. The Conditions of Agricultural Growth: The Economics of Agrarian Change under Population Pressure. Allen & Unwin, London.

Breierova, L., Duflo, E., 2004. The Impact of Education on Fertility and Child Mortality: Do Fathers Really Matter Less than Mothers? NBER Working Paper Applied Demography National bureau of Economic Research, Cambridge.

Brown, J.C., Guinnane, T., 2003. Two Statistical Problems in the Princeton Project on the European Fertility Transition. Center discussion paper No. 869. Yale University Economic Growth Center, New Haven, CT.

Caldwell, J.C., 1976. Toward a restatement of demographic transition theory. Population and Development Review 2, 321–366.

Canning, D., Schultz, T.P., 2012. The economic consequences of reproductive health and family planning. Lancet 380, 165–171.

Chowdhury, M.E., Ahmed, A., Kalim, N., Koblinsky, M., 2009. Causes of maternal mortality decline in matlab, Bangladesh. Journal of Health Population and Nutrition 27, 108–123.

Cleland, J., Bernstein, S., Ezeh, A., Faundes, A., Glasier, A., Innis, J., 2006. Family planning: the unfinished agenda. The Lancet 368 (9549), 1810–1827.

Cleland, J., Conde-Agudelo, A., Peterson, H., Ross, J., Tsui, A., 2012. Contraception and health. Lancet 380, 149–156.

Coale, A.J., Hoover, E.M., 1959. Population growth and economic development in low-income countries. American Economic Review 49, 436–438.

Coale, A.J., Watkins, S.C. (Eds.), 1986. The Decline of Fertility in Europe: The Revised Proceedings of a Conference on the Princeton European Fertility Project. Princeton University Press, Princeton, NJ.

Comstock, A., 1873. An act for the suppression of trade in, and circulation of, obscene literature and articles of immoral use. ch. 258, § 2 Statistics 17, 599, 42nd United States Congress, 3rd Session.

Conde-Agudelo, A., Belizan, J.M., 2000. Maternal morbidity and mortality associated with interpregnancy interval: cross sectional study. British Medical Journal 321, 1255–1259.

DaVanzo, J., Razzaque, A., Rahman, M., 2005. The Effects of Birth Spacing on Infant and Child mortality., Pregnancy Outcomes and Maternal Morbidity and Mortality in Matlab, Bangladesh. RAND Working Paper. The RAND Corporation, Santa Monica, CA.

Davis, K., 1967. Population policy: will current programs succeed? Science 158, 730–739.

Debpuur, C., Phillips, J.E., Jackson, E.F., Nazzar, A., Ngom, P., Binka, F.N., 2002. The impact of the navrongo project on contraceptive knowledge and use, reproductive preferences, and fertility. Studies in Family Planning 33, 141–164.

Desai, J., Tarozzi, A., 2011. Microcredit, family planning programs, and contraceptive behavior: evidence from a field Experiment in Ethiopia. Demography 48, 749–782.

Ezegwui, H., Onoh, R., Ikeako, L., Onyebuchi, A., Umeora, O., Ezeonu, P., Ibekwe, P., 2013. Investigating maternal mortality in a public teaching hospital, abakaliki, ebonyi state, Nigeria. Annals of Medical and Health Sciences Research 3, 75–80.

Finkle, J.L., Crane, B.B., 1985. Ideology and politics at Mexico-City – the United-States at the 1984 international-conference on population. Population and Development Review 11, 1–28.

Foster, A., Roy, N., 1997. The dynamics of education and fertility: evidence from a family planning experiment. Unpublished manuscript.

Galor, O., Weil, D.N., 2000. Population, technology, and growth: from Malthusian stagnation to the demographic transition and beyond. American Economic Review 90, 806–828.

Gertler, P., Shah, M., Bertozzi, S.M., 2005. Risky business: the market for unprotected commercial sex. Journal of Political Economy 113, 518–550.

Gertler, P.J., Molyneaux, J.W., 1994. How economic-development and family-planning programs combined to reduce indonesian fertility. Demography 31, 33–63.

Glasier, A., Gulmezoglu, A.M., Schmid, G.P., Moreno, C.G., Van Look, P.F., 2006. Sexual and reproductive health 1-sexual and reproductive health: a matter of life and death. Lancet 368, 1595–1607.

Goldin, C., Katz, L.F., 2002. The power of the pill: oral contraceptives and women's career and marriage decisions. Journal of Political Economy 110, 730–770.

Greene, M.E., Merrick, T., 2005. Poverty Reduction: Does Reproductive Health Matter?. Health, nutrition and population discussion paper The World Bank, Washington, DC.

Hashemi, A., Salehi-Isfahani, D., 2013. From health service delivery to family planning: the changing impact of health clinics on fertility in rural Iran. Economic Development and Cultural Change 61, 281–309.

Hesketh, T., Zhu, W.X., 1997. Health in China – the one child family policy: the good, the bad, and the ugly. British Medical Journal 314, 1685–1687.

Jain, A.K., 2011. Measuring the effect of fertility decline on the maternal mortality ratio. Studies in Family Planning 42, 247–260.

Jain, A.K., Ross, J.A., 2012. Fertility differences among developing countries: are they still related to family planning program efforts and social settings? International Perspectives on Sexual and Reproductive Health 38, 15–22.

Jayachandran, S., Lleras-Muney, A., 2009. Life expectancy and human capital investments: evidence from maternal mortality declines. Quarterly Journal of Economics 124, 349–397.

Jones, K., 2013. Contraceptive Supply and Fertility Outcomes: Evidence from Ghana. International Food Policy Research Institute, Washington, DC.

Joshi, S., Schultz, T., 2007. Family Planning as an Investment in Development: Evaluation of a Program's Consequences in Matlab, Bangladesh (Yale University Economic Growth Center Discussion Paper).

Joshi, S., Schultz, T.P., 2013. Family planning and women's and children's health: long-term consequences of an outreach program in matlab, Bangladesh. Demography 50, 149–180.

Kaiser, J., 2011. Does family planning bring down fertility? Science 333, 548–549.

Kalemli-Ozcan, S., 2002. Does the mortality decline promote economic growth? Journal of Economic Growth 7, 411–439.

Kelley, A.C., 2003. The population debate in historical perspective: revisionism revised. In: Birdsall, N., Kelley, A.C., Sinding, S.W. (Eds.), Population Matters: Demographic Change, Economic Growth, and Poverty in the Developing World. Oxford University Press, Oxford.

Koenig, M.A., Fauveau, V., Chowdhury, A.I., Chakraborty, J., Khan, M.A., 1988. Maternal mortality in Matlab, Bangladesh: 1976-85. Studies in Family Planning 19, 69–80.

Koenig, M.A., Khan, M.A., Wojtyniak, B., Clemens, J.D., Chakraborty, J., Fauveau, V., Phillips, J.F., Akbar, J., Barua, U.S., 1990. Impact of measles vaccination on childhood mortality in rural Bangladesh. Bulletin of the World Health Organization 68, 441–447.

Koenig, M.A., Rob, U., Khan, M.A., Chakraborty, J., Fauveau, V., 1992. Contraceptive use in matlab, Bangladesh in 1990-Levels, trends, and explanations. Studies in Family Planning 23, 352–364.

Kohler, H.-P., 2012. Copenhagen Consensus 2012: Challenge Paper on "Population Growth". Population Studies Center Working Paper Series, PSC 12-03. University of Pennsylvania Scholarly Commons, Philadelphia.

Kosgei, R.J., Lubano, K.M., Shen, C.Y., Wools-Kaloustian, K.K., Musick, B.S., Siika, A.M., Mabeya, H., Carter, E.J., Mwangi, A., Kiarie, J., 2011. Impact of integrated family planning and HIV care services on contraceptive use and pregnancy outcomes: a retrospective cohort study. Jaids-journal of Acquired Immune Deficiency Syndromes 58, E121–E126.

Lam, D., 2011. How the world survived the population bomb: lessons from 50 years of extraordinary demographic history. Demography 48, 1231–1262.

Lavy, V., Zablotsky, A., 2011. Mother's Schooling and Fertility under Low Female Labor Force Participation: Evidence from a Natural Experiment. NBER Working Paper NO. 16856. National Bureau of Economic Research, Cambridge.

Levine, P.B., Staiger, D., 2004. Abortion policy and fertility outcomes: the Eastern European experience. Journal of Law & Economics 47, 223–243.

Lundberg, S., Pollak, R.A., 1996. Bargaining and distribution in marriage. Journal of Economic Perspectives 10, 139–158.

Menken, J., Duffy, L., Kuhn, R., 2003. Childbearing and women's survival: new evidence from rural Bangladesh. Population and Development Review 29, 405.

Miller, G., 2010. Y contraception as development? New evidence from family planning in Colombia. Economic Journal 120, 709–736.

Modrek, S., Ghobadi, N., 2011. The expansion of health houses and fertility outcomes in rural Iran. Studies in Family Planning 42, 137–146.

Molyneaux, J.W., Gertler, P.J., 2000. The impact of targeted family planning programs in Indonesia. Population and Development Review 26, 61–85.

Muhuri, P.K., 1995. Health programs, maternal education, and differential child mortality in matlab, Bangladesh. Population and Development Review 813–834.

Palloni, A., Rafalimanana, H., 1999. The effects of infant mortality on fertility revisited: new evidence from Latin America. Demography 36, 41–58.

Phillips, J.F., Bawah, A.A., Binka, F.N., 2006. Accelerating reproductive and child health programme impact with community-based services: the Navrongo experiment in Ghana. Bulletin of the World Health Organization 84, 949–955.

Phillips, J.F., Hossain, M.B., 2003. The impact of household delivery of family planning services on women's status in Bangladesh. International Family Planning Perspectives 29, 138–145.

Phillips, J.F., Jackson, E.F., Bawah, A.A., Macleod, B., Adongo, P., Baynes, C., Williams, J., 2012. The long-term fertility impact of the navrongo project in Northern Ghana. Studies in Family Planning 43, 175–190.

Phillips, J.F., Stinson, W.S., Bhatia, S., Rahman, M., Chakraborty, J., 1982. The demographic impact of the family planning–health services project in matlab, Bangladesh. Studies in Family Planning 131–140.

Pitt, M.M., Rosenzweig, M.R., Gibbons, D.M., 1993. The determinants and consequences of the placement of government programs in Indonesia. World Bank Economic Review 7, 319–348.

Pop-Eleches, C., 2006. The impact of an abortion ban on socioeconomic outcomes of children: evidence from Romania. Journal of Political Economy 114, 744–773.

Pop-Eleches, C., 2010. The supply of birth control methods, education, and fertility evidence from Romania. Journal of Human Resources 45, 971–997.

Portner, C.C., Beegle, K., Christiaensen, L., 2011. Family Planning and Fertility: Estimating Program Effects Using Cross-sectional Data. Policy Research Working Paper. World Bank.

Pritchett, L.H., 1994. Desired fertility and the impact of population policies. Population and Development Review 20, 1–55.

Rahman, M.O., Menken, J., 2012. Reproductive health. In: Merson, M.H., Black, R.E., Milles, A.J. (Eds.), Global Health: Diseases, Programs, Systems and Policies. Jones & Bartlett Learning, Burlington, MA.

Rasul, I., 2008. Household bargaining over fertility: theory and evidence from Malaysia. Journal of Development Economics 86, 215–241.

Robey, B., Rutstein, S.O., Morris, L., 1993. The fertility decline in developing-countries. Scientific American 269, 60–67.

Rogers, E.M., Vaughan, P.W., Swalehe, R.M.A., Rao, N., Svenkerud, P., Sood, S., 1999. Effects of an entertainment-education radio soap opera on family planning behavior in Tanzania. Studies in Family Planning 30, 193–211.

Rosenzweig, M.R., Wolpin, K.I., 1986. Evaluating the effects of optimally distributed public programs – child health and family-planning interventions. American Economic Review 76, 470–482.

Rutstein, S.O., 2005. Effects of preceding birth intervals on neonatal, infant and under-five years mortality and nutritional status in developing countries: evidence from the demographic and health surveys. International Journal of Gynecology & Obstetrics 89, S7–S24.

Sah, R.K., 1991. The effects of child-mortality changes on fertility choice and parental welfare. Journal of Political Economy 99, 582–606.

Salas, J.M.I., 2013. Consequences of Withdrawal: Free Condoms and Birth Rates in the Philippines. University of California, Irvine.

Salehi-Isfahani, D., Abbasi-Shavazi, M.J., Hosseini-Chavoshi, M., 2010. Family planning and fertility decline in rural Iran: the impact of rural health clinics. Health Economics 19 (S1), 159–180.

Schultz, T., 1992. Assessing family planning cost-effectiveness. In: Phillips, J.E., Ross, J.A. (Eds.), Family Planning Programmes and Fertility. Oxford University Press, New York.

Schultz, T.P., 1985. Changing world prices, womens wages, and the fertility transition – Sweden, 1860–1910. Journal of Political Economy 93, 1126–1154.

Schultz, T.P., 1994. Human-capital, family-planning, and their effects on population-growth. American Economic Review 84, 255–260.

Schultz, T.P., 2009. How Does Family Planning Promote Development? Evidence from a Social Experiment in Matlab, Bangladesh—1977–1996. Yale University, Economic Growth Center, New Haven, CT.

Simmons, G.B., Balk, D., Faiz, K.K., 1991. Cost-effectiveness analysis of family planning programs in rural Bangladesh: evidence from matlab. Studies in Family Planning 22, 83–101.

Simon, J.L., 1981. The Ultimate Resource. Princeton University Press, Princeton.

Sinding, S., 2007. The global family planning revolution: overview and perspective. In: Robinson, W.C., Ross, J.A. (Eds.), The Global Family Planning Revolution: Three Decades of Population Policies and Programs. The World Bank, Washington, DC.

Sinha, N., 2005. Fertility, child work, and schooling consequences of family planning programs: evidence from an experiment in rural Bangladesh. Economic Development and Cultural Change 54, 97–128.

Tsui, A.O., 2001. Population policies, family planning programs, and fertility: the record. Population and Development Review 27, 184–204.

Tsui, A.O., Mcdonald-Mosley, R., Burke, A.E., 2010. Family planning and the burden of unintended pregnancies. Epidemiologic Reviews 32, 152–174.

Voigtlander, N., Voth, H.-J., 2013. How the west "Invented" fertility restriction. American Economic Review 103, 2227–2264.

Westoff, C.F., 2013. The recent fertility transition in Rwanda. Population and Development Review 38, 169–178.

Winikoff, B., Sullivan, M., 1987. Assessing the role of family planning in reducing maternal mortality. Studies in Family Planning 18, 128–143.

Family Planning Programs: Feminist Perspectives

Carmen Barroso, International Planned Parenthood Federation/Western Hemisphere Region, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since the 1960s, governments and private organizations implemented family planning programs, largely driven by fears of rapid population growth. By the 1990s, the rationale for these programs shifted to the human right to autonomy in sexuality and reproduction. With the 20th anniversary of the United Nations International Conference on Population and Development Programme of Action approaching, this article examines the feminist critique of population programs that led to this historic shift. It examines gaps that remain in the implementation of the Programme of Action, and emerging issues – including youth sexual rights – that feminists are prioritizing as key areas for population and development policies and programs.

Population Paradigms

Since the 1960s and even earlier in some cases, governments and private organizations in many poorer countries have developed and implemented family planning programs. By 1985, 37 governments had established national family planning programs, and an additional 33 countries had significant nongovernmental programs with the direct or indirect support of governments (Bongaarts et al., 1990). These efforts were largely driven by fears of rapid population growth, and the reduction of fertility was their highest priority. By the 1990s, however, critics of these programs succeeded in causing, what many have called, a revolution in the rationale for family planning programs. The human rights to autonomy in matters of sexuality and reproduction was proposed by feminists from the global north and south as the cornerstone for these efforts. As a result of this critique, 179 countries agreed to a new approach to population policies and family planning programs at the United Nations International Conference on Population and Development (ICPD), held in Cairo in 1994.

Family Planning Programs and the Erstwhile Population Approach

To fully understand the significance of the Cairo accord and the influence feminist thinking had on it, it is necessary to look at the philosophies and practices it replaced. The 1960s, with the advent of new contraceptive technologies and an increase in fears of a population explosion, saw a significant increase in support for family programs in both industrialized and developing nations. In 1967, the United Nations Fund for Population Activities (later renamed the United Nations Population Fund) was established to support population and family planning programs. At the UN population conference in Bucharest in 1974, many policy makers – mainly from Northern nations – viewed unchecked population growth as the root cause of poverty, environmental degradation, and social unrest. Many developing countries resisted this line of thinking, suggesting instead that economic development was a more straightforward solution to societal ills than family planning. While women's rights and empowerment gained some attention, they were not seen to be central arguments for family planning. Meanwhile, on the ground, family planning programs continued, largely unaffected by the debate at the international forum (Demeny, 1985).

At the 1984 International Conference on Population in Mexico City, demographic concerns continued to be the driving force in discussions. However, in a reversal of its previous policy, the powerful US delegation now asserted that population growth had a neutral, not harmful, effect on economic development. Furthermore, antiabortion politics gained center stage as the US launched an attack on abortion rights (Demeny, 1985). Still, family planning programs on the ground continued to be guided by demographic goals and to use mechanistic measures such as reaching target numbers of 'contraceptive acceptors.' The principal goal driving these programs was the distribution of contraception, chiefly to married women. Perhaps more accurately called population programs than family planning programs, these efforts usually did not address other health needs, even reproductive health needs.

The Feminist Critique of Population and Family Planning

As these programs proliferated throughout the 1970s and 1980s, women around the world began focusing on the negative effects of family planning programs that were based on demographics. While not denying the benefits of family planning, feminists rejected the depiction of women as targets of programs designed to limit births (Pitanguy, 1995). The feminist critique concentrated on several aspects of program practices, including their low quality, the vulnerability of women to coercion, and their sole concentration on fertility regulation to the exclusion of other women's health needs.

The vulnerability to coercion was a major element of the feminist critique. In some countries, health care providers were under pressure to achieve targets or be subject to penalties, and workers pressed people into accepting contraception or sterilization against their will. With the success of family planning programs measured in terms of numbers of acceptors, some agencies and governments devised incentive and disincentive schemes to induce couples to participate in family planning

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31030-3

programs. For some time, there was a debate about the fine line separating incentives from coercion. Feminists argued that, while milder than sheer force, the incentive practice was coercive, and that incentives were aimed primarily at the poor, for whom an economic inducement could be a vital necessity rather than a free choice (Hartmann, 1995). The Cairo Programme of Action later renounced incentives.

Furthermore, women – and women were the objects of these campaigns in much greater numbers than men – were given little contraceptive choice. The focus was on sterilization or other modern, provider-controlled methods of contraception such as the injectable Depo-Provera. The risks of these methods, feminists argued, were largely ignored and even exacerbated as a result of poor screening procedures and the lack of follow-up examinations. Safer methods, such as the condom and diaphragm, were neglected, even though barrier methods have the ability to help prevent the spread of sexually transmitted infections. Indeed, apart from fertility regulation, family planning programs provided little else in the way of services. Furthermore, male contraceptive methods were ignored in the research sphere and male participation was neglected in family planning programs, except for sterilization campaigns in some countries (Dixon-Mueller, 1993).

Feminists insisted that reproductive health programs be comprehensive in nature with family planning being only one item in a wide range of women's reproductive health needs. Other reproductive health services included skilled care during pregnancy; delivery and postpartum care; prevention of infertility; counseling for sexual dysfunction; safe abortion; prevention and management of sexually transmitted infections; reproductive tract infections and reproductive system cancers; and treatment of postmenopausal health problems (HERA, n.d.).

Furthermore, feminists demanded that women's health advocates should be represented at all levels of policy and decision-making. They also maintained that men's responsibilities in the reproductive realm should not be ignored (Women's Declaration on Population Policies, 1992).

Perhaps most central to the feminist critique is the assertion that women have the ability to make appropriate reproductive choices. In the feminist view, reproductive choices belong at the individual level, not with the health practitioner and especially not with the state. Noncoercive family planning programs aim to prevent unwanted pregnancies – as defined by couples, not by the government or other outside parties – through the provision of competent services that respect dignity and privacy. The essence of the feminists' perspective is the view that reproductive rights are human rights.

Another fundamental premise of the feminist perspective on family planning programs sees women's empowerment and equality as a desirable goal in its own right, rather than a means to a demographic end. Feminists also emphasize men's responsibilities in sexual and reproductive matters (García-Moreno and Claro, 1994).

Articulating these views, women in many countries around the globe began advocating for a wholesale transformation of the approach to population policies and family planning programs. By the time of the 1994 United Nations International Conference on Population and Development (ICPD), an international women's health movement was organized and had developed the expertise and the professional skills needed to influence policy makers and incite the transformation of global population and family planning programs (Klugman, 2000). "The Cairo Programme of Action sums up the agenda of priorities which women from all over the world, through their leadership networks, built gradually during the years of preparation for the conference" (Berquó, 1997).

The Cairo Programme of Action and the 'Feminist Imprint'

The Cairo conference ushered in a new era of population policy and family planning programming. Its resulting Programme of Action reflects a clear 'feminist imprint' (Petchesky, 1995). The conference document presents a new perspective, in which high rates of population growth are understood as an interdependent and aggravating factor – rather than the cause – of problems such as poverty and environmental degradation. Even more critically, it places women's well-being at the center of population policy, and points to the human rights of individuals to determine and plan family size. Unlike documents from previous population conferences, the ICPD Programme of Action includes chapters on 'Reproductive Rights and Reproductive Health,' and 'Gender Equality, Equity and the Empowerment of Women.'

Echoing the feminist critique, the comprehensive nature of reproductive health is elucidated in Section 7.2 of the Programme of Action, as follows:

> Reproductive health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity, in all matters relating to the reproductive system and to its functions and processes. Reproductive health therefore implies that people are able to have a satisfying and safe sex life and that they have the capability to reproduce and the freedom to decide if, when and how often to do so. Implicit in this last condition are the right of men and women to be informed and to have access to safe, effective, affordable and acceptable methods of family planning of their choice, as well as other methods of their choice for regulation of fertility which are not against the law, and the right of access to appropriate health-care services that will enable women to go safely through pregnancy and childbirth and provide couples with the best chance of having a healthy infant.

While previous population conferences had viewed women's development as a tool for achieving lowered fertility rates, the Cairo Programme of Action framed women's empowerment as a goal worthy in its own right. 'The empowerment and autonomy of women,' the agreement reads, "and the improvement of their political, social, economic and health status is a highly important end in itself" (*Programme of Action*: 4.1). In another clear echo of the feminist critique, the document goes far in its call for gender equality (*Programme of Action*: 4.1):

> The power relations that impede women's attainment of healthy and fulfilling lives operate at many levels of society, from the most personal to the highly public. Achieving change requires policy and programme actions that will improve women's access to secure

livelihoods and economic resources, alleviate their extreme responsibilities with regard to housework, remove legal impediments to their participation in public life, and raise social awareness through effective programmes of education and mass communication.

Critiquing the Critique

The feminist critique has its detractors and the outcome of the Cairo Conference was not universally applauded. Vocal critics were mainly religious fundamentalists, especially the leadership of the Roman Catholic Church (Shannon, 1994). A much quieter reaction was registered among some demographers and others whose concerns about high rates of population growth led to their dismay at what they saw as the dismantling of population policies (Dixon-Mueller and Germain, 2000). For example, Malcom Potts maintained that the "flaw in the Cairo process was to move from a well-tested middle ground of supplying family planning to adopting a much broader agenda" (Potts, 1995). Jason Finkle and Alison McIntosh claimed that women's groups 'co-opted' the ICPD. The conference, they maintained, was "not a step forward for the population movement, which for many years has given primacy to efforts to limit population growth" (Finkle and McIntosh, 1996).

Even some supporters of the reproductive health paradigm adopted in Cairo found some faults in its processes and outcomes. In a review of responses "that tend to temper the initial feeling of euphoria surrounding the triumph of the feminist agenda," Huda Zurayk writes that while a large number of women from Third World countries were present in the ICPD process, "the strings were definitely being pulled by the well-organized Western feminist groups" (Zurayk, 1997). Others, however, point to the significant contribution of women from the Global South – particularly women from Brazil, India, Egypt, Nigeria, Mexico, and South Africa – to the ICPD agreement.

Beyond Cairo: The Feminist Critique Moves Forward

After the Cairo Conference, the feminist critique expanded to include a vision of policies, governing not only family planning programs but also the broader panoply of issues related to reproduction and sexuality. One year after the ICPD, in 1995, the United Nations held another major international conference in Beijing focusing on women's rights. Again the ensuing conference document – the Beijing Platform for Action – reflected feminist input. While the Platform did not explicitly use the phrase 'sexual rights,' the agreement nonetheless endorsed the basic tenets of such rights, and placed them within a human rights framework. "The human rights of women include their right to have control over and decide freely and responsibly on matters related to their sexuality, including sexual and reproductive health, free of coercion, discrimination and violence," reads paragraph 96 of the *Platform for Action*.

There is little question that feminist analysis has had significant influence on policy worldwide, principally through the actions of women's nongovernmental organizations. Feminist scholars have pressed for a similar impact on the discipline of demography, where gender issues have remained marginal. While ideological resistance presents a considerable obstacle, "intellectually, the incorporation of gender issues in demography promises to enhance the science by offering a deeper understanding of social process and the application of multilevel methodologies" (Presser, 2000).

As the 20th anniversary of the ICPD approaches, feminists are calling attention to the glaring gaps and inequalities that remain from the uneven implementation of the Cairo agenda, and from issues that were not fully addressed in Cairo. Four themes have been the focus of the feminist critique of late: sexual rights, unsafe abortion, comprehensive sexuality education, and government accountability. The one that is getting more traction in international fora is the sexual rights of youth, which also encompasses the right to sexuality education.

Today's generation of youth make up the largest cohort of young people ever, and addressing their needs is an increasingly urgent concern of family planning agencies. The unmet need for contraception among adolescents aged 10–19 is more than twice that of adult women (Guttmacher, 2010), while nearly a fifth of young women in developing countries become pregnant before they turn 18 (UNFPA, 2013). Another egregious violation of young women's sexual rights is the practice of early and forced marriage. In developing countries, some 16% of girls marry before they are 18 years old, and they are often pressured to conceive immediately, with harmful implications for both the mother and baby (UNFPA, 2013).

The Cairo Programme of Action (Paragraph 7.3) called upon governments to commit "to the promotion of mutually respectful and equitable gender relations and particularly to meeting the educational and service needs of adolescents to enable them to deal in a positive and responsible way with their sexuality." Yet young people still face formidable social and psychological barriers to the exercise of a healthy and pleasurable sexuality. Some of these barriers are common to those affecting all people especially those living in poverty, such as lack of sufficient resources for reproductive health programs and persistent opposition from conservative sources deeply opposed to reproductive autonomy. In fact restrictions to young people's access to services have been a major rallying cry of conservative forces, which seem to have lost ground on the power they exerted in controlling people's behaviors in their daily lives in many parts of the world, but are very active exerting political power both at national and international levels.

The family planning programs are now aware of the need to cater to young people, but all too often, reproductive health programs for youth are focused primarily on preventing teen pregnancy and hence are subject to a similar feminist critique to that which was applied against early family planning programs. These programs for youth see young people narrowly as mere contraceptive users, rather than recognizing them as sexual beings with sexual rights. As a result, young people end up alienated from information and services, and hence do not have the resources they need to fully exercise their rights.

A further impediment to young people's rights is the proliferation of legislation establishing high age of consent, below which any sexual relation is considered statutory rape,

and laws requiring parental consent for minors to access a range of sexual and reproductive health services, including information, counseling, contraception, and abortion. While these laws are intended to protect children, they are actually not very effective in preventing the abuse of minors or fostering dialogue between parents and adolescents. In fact, these laws are counterproductive because they impede adolescents' access to the contraceptive technologies they need to safely and responsibly exercise their sexuality. They also infringe upon young people's autonomy. Article 5 of the United Nations Convention on the Rights of the Child establishes that, as children grow older, their evolving capacity for decision-making must be taken into account, increasingly allowing minors to exercise their rights on their own behalf according to their capacity (United Nations, 1989).

As the development community prepares to negotiate a new post-2015 global development agenda, young people themselves are taking the position of leadership exercised by feminists in 1994 in their call for the primacy of a rights perspective. At the 2012 Global Youth Forum in Bali, some 3000 youth participated in drafting a declaration that echoes the feminist agenda in Cairo and takes it one step further. For example, the declaration's text on sexual rights states:

> Governments must fund and develop, in equal partnership with young people and health care providers, policies, laws, and programs that recognize, promote, and protect young peoples' sexual rights as human rights. This must be developed in accordance with the principles of human rights, non-discrimination, respect, equality and inclusivity, with a gendered, multicultural and secular approach … Cultural and religious barriers such as parental and spousal consent, and early and forced marriages, should never prevent access to family planning, safe and legal abortion, and other reproductive health services – recognizing that young people have autonomy over their own bodies, pleasures, and desires.
>
> UNFPA, 2012.

One obstacle to integrating a rights perspective into the development agenda is that international health donors have become increasingly focused on measurable results. In principle, this shift is a positive one. The problem is the narrow definition of results prevailing in this new paradigm. The standard indicator used to measure the success of family planning programs is the number of years a couple will be protected against unwanted pregnancies. This indicator encourages the promotion of long-acting contraceptives, such as the intrauterine device and implants. Many women do want to use long-acting contraceptives, and therefore increasing access to these methods is a positive move. However, promoting them to the exclusion of other methods is problematic in that they do not offer protection against sexually transmitted infections, including HIV, and they may be less appropriate for women who have sporadic sexual relations or multiple partners, as well as women who may be unknowingly exposed to sexually transmitted infections from their spouse or partner. Furthermore, these long-acting methods have significance for many women's rights advocates who recall a bygone era when population growth was seen as the major driver of economic and social ills, and when family planning programs driven by numerical targets related to contraceptive use led to the neglect of proper health education and informed consent, particularly for women and men in developing countries.

This concern was at the root of the feminist reaction to the London Family Planning Summit, which was cohosted by the Bill and Melinda Gates Foundation and the UK Department for International Development in 2012. The purpose of the Summit was to draw attention to the family planning agenda, which had become a neglected area of health and development both in terms of funding and research. The Summit was successful in gaining both political and economic support for reducing the unmet need for family planning. Feminists and sexual and reproductive health and rights advocates were very vocal in raising concerns in the run up to the event, when the narrow focus on family planning seemed to diminish the hard-won gains of Cairo to place family planning within the context of women's broader, sexual, and reproductive health needs and rights. Furthermore, the Summit was criticized for its exclusion of the issue of unsafe abortion, and exclusive focus on numerical targets for access to contraception (Guttmacher, 2012). If the only measure of success is new contraceptive users, feminists and rights advocates argued, providers would feel pressured to meet those targets without regard for rights. In particular, the government of India, one of the countries that made commitments at the Summit, drew heavy criticism given the persistent use of incentives for sterilization quotas, health workers are still expected to fill in many states, in spite of the abolition of centralized targets soon after the Cairo conference (Human Rights Watch, 2012). The organizers of the Summit were quick to respond to these criticisms and included repeated references to the importance of women's rights, and rousing speeches were made pledging to give women and girls the power to decide. FP 2020 was created to follow-up the commitments of the Summit. Its first report, launched in November of 2013, mentions the development of a framework to measure and report on the autonomy, equity and human rights-based dimensions of family planning programs but it is still unclear how this will be implemented or measured (Family Planning, 2020).

See also: Gender and Health Care; Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective.

Bibliography

Berquó, E., 1997. The ICPD programme of action and reproductive health policy development in Brazil. In: Proceedings of the International Population Conference. International Union for the Scientific Study of Population, Liége, Belgium.

Bongaarts, J., Parker, M.W., Phillips, J.F., 1990. The demographic impact of family planning programs. Studies in Family Planning 21 (6), 299–310.

Correa, S., Petchesky, R., 1994. Reproductive and sexual rights: a feminist perspective. In: Sen, G., Germain, A., Chen, L. (Eds.), Population Policies Reconsidered: Health, Empowerment and Rights. Harvard University Press, Boston, pp. 107–123.

Demeny, P., 1985. Bucharest, Mexico city, and beyond. Population and Development Review 11 (1), 99–106.

Dixon-Mueller, R., 1993. Population Policy and Women's Rights: Transforming Reproductive Choice. Praeger, Westport, CT.

Dixon-Mueller, R., Germain, A., 2000. Reproductive health and the demographic imagination. In: Presser, H., Sen, G. (Eds.), Women's Empowerment and Demographic Processes: Moving Beyond Cairo. Oxford University Press, Oxford, UK.

Family Planning 2020, 2013. FP2020 Partnership in Action. From: http://www.familyplanning2020.org/progress (retrieved 20.12.13.).

Finkle, J., McIntosh, A., 1996. Cairo revisited: some thoughts on the implications of the ICPD. Health Transition Review 6, 110–114.

Garcia-Moreno, C., Claro, A., 1994. Challenges from the women's health movement: women's rights versus population control. In: Sen, G., Germain, A., Chen, L. (Eds.), Population Policies Reconsidered: Health, Empowerment, and Rights. Harvard University Press, Boston, pp. 27–46.

Guttmacher Institute and International Planned Parenthood Federation, 2010. Facts on the Sexual and Reproductive Health of Adolescent Women in the Developing World. From: http://www.guttmacher.org/pubs/FB-Adolescents-SRH.pdf (retrieved 28.11.13.).

Guttmacher Institute, 2012. London summit puts family planning back on the agenda, offers new lease on life for women and girls. Guttmacher Policy Review 15 (3). From: http://www.guttmacher.org/pubs/gpr/15/3/gpr150320.html (retrieved 28.11.13.).

Hartmann, B., 1995. Reproductive Rights and Wrongs: The Global Politics of Population Control. South End, Boston.

HERA: Health, Empowerment, Rights and Accountability, n.d. Women's Sexual and Reproductive Rights and Health Action Sheets. HERA, New York.

Human Rights Watch, 2012. India: Target-Driven Sterilization Harming Women. From: http://www.hrw.org/news/2012/07/12/india-target-driven-sterilization-harming-women (retrieved 28.12.13.).

Jain, A., Bruce, J., 1994. Quality: the key to success. In: Mazur, L. (Ed.), Beyond the Numbers. Island, Washington, DC.

Klugman, B., 2000. Empowering women through the policy process: the making of health policy in South Africa. In: Presser, H., Sen, G. (Eds.), Women's Empowerment and Demographic Processes: Moving Beyond Cairo. Oxford University Press, Oxford, UK, pp. 95–118.

Petchesky, R.P., 1995. From population control to reproductive rights: feminist fault lines. Reproductive Health Matters 6, 152–161.

Pitanguy, J., 1994. Feminist politics and reproductive rights: the case of Brazil. In: Sen, G., Snow, R. (Eds.), Power and Decision: The Social Control of Reproduction. Harvard University Press, Boston.

Pitanguy, J., 1995. From Mexico to Beijing: a new paradigm. Health and Human Rights 1 (4), 454–460.

Potts, M., 1995. Cairo's skewed consensus. Population Today 23 (11).

Presser, H., 2000. Demography, feminism, and the science-policy nexus. In: Presser, H., Sen, G. (Eds.), Women's Empowerment and Demographic Processes: Moving Beyond Cairo. Oxford University Press, Oxford, UK.

Shannon, D., 1994. All roads led to Cairo. Conscience 15 (4), 3–10.

UNFPA, 2012. Bali Global Youth Forum Declaration, 11. From: http://icpdbeyond2014.org/uploads/browser/files/bali_global_youth_forum_declaration.pdf (retrieved 24.12.13.).

UNFPA, 2013. Motherhood in Childhood: Facing the Challenge of Adolescent Pregnancy. State of the World Population 2013. United Nations Population Fund, New York.

United Nations, 1989. Convention on the Rights of the Child, A/RES/44/25. United Nations, New York. From: http://www.un.org/documents/ga/res/44/a44r025.htm (retrieved 28.12.13.).

Women's declaration on population policies, 1992. In: Mazur, L. (Ed.), Beyond the Numbers. Island, Washington, DC.

Zurayk, H., 1997. Reproductive health in population policy: a review and a look ahead. In: International Population Conference. International Union for the Scientific Study of Population (IUSSP), Liége, Belgium.

Family Policy

Mary Daly, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Family policy, at its most developed in Europe, refers to state policies oriented to the welfare of children and the support of family as an institution or way of life. Most widely, family polices concern themselves with demographic matters, the financial and other resources available to families, gender equality, the links between family life and employment, and the role and significance of family as a social organization. The most common constituent elements of family policy consist of cash benefits for families with children, child care and early education services, and leaves from employment for family-related purposes. Family policy is characterized by great variation and, in the last decades, considerable reform.

Defining the Field

Family policy is not at all clear-cut as a field of social policy and there is no agreement on a definition. It is in fact a broad field – Gauthier (1999: 32) terms it a 'wide umbrella of policies.' As such, critical decisions are required about what to include and exclude. One way of approaching this is to think in terms of what society considers as family. From this perspective, families with children would be at the center of any definition of family policy. A societal understanding of family also suggests that family policies should encompass policies on the care of older people, given that in many societies this is seen to be the responsibility of families. However, if we take the stance of the state – which is the customary starting point for policy analysis – policies for older people are not widely regarded or made as family policies. They most often fall within the domains of health or pensions/retirement policy. One of the earliest definitions of family policy is that used by Kamerman and Kahn (1978). They defined it as deliberate governmental actions taken toward families with children. Both scholarship and policy have moved on from this, especially in recognizing that family policy is not limited to that which is focused specifically on family. While the focus more or less remains on families with children, it is now customary to consider as family policy a range of instruments that governments use to support, influence, and regulate behaviors and welfare in different types of families with children.

This contribution takes the broader route in several respects, especially in recognizing the context within which family policies are designed and operate. In this regard, I would emphasize two things. First, seeing the broader context means recognizing that family policy is implicated in organizing the interrelations among the family, state, and market. In other words, while it may appear to be micro and rather small scale, family policy is actually part of the governance of foundational matters relating to how the economy and society are organized. Second, it is important not to underestimate the complexity of family. Rather than a mere structure, families are constituted by sets of interrelationships and divisions, especially along gender and generational lines. So family policies are intimately involved in governing these and other considerations. In addition, understanding family as complex means that we cannot view it as a passive phenomenon on which state policies stamp their mark without any resistance. Rather, family has to be seen as a domain and sphere of activity in its own right, with its own dynamic, relationships, and interests.

In the analysis to follow with the main focus on Europe, attention is centered especially on the typical approaches and how these vary across countries, the drivers and political background, the types of policy instrument in place (cash benefits, tax allowances, leave, and services), the motives and objectives, financing and other arrangements, and general trends.

The History of Family Policy in Europe

Taking Europe as the canvas, one can, generalizing somewhat from Bahle (2009) and Daly (2010), identify two general contexts in the history of engagement between states, markets, and families. The first was of industrialization and nation-building wherein the state and political actors, from the late nineteenth century on, endeavored to respond to the needs generated by the shift from agricultural to industrial forms of production and economic organization. From a family policy perspective, the central intent here was the support of the working man and his family. With the male breadwinner family as the orienting image, early social and family policies linked closely to the idea of a family wage, and provided subsidies for both marriage and the 'dependants' of the breadwinner (Crouch, 1999). The dynamic in the second period, commencing in the 1970s, lay in the expansion of the service economy and, associated with this and other factors, the large-scale movement of women into the labor market. This turned the focus on a broader meaning of family support, one tied into the question of how family-related roles and functions could be reconfigured in order to enable economic and social reform. Gender politics sat alongside class politics as shaping influences in this period when services substituting for maternal caregiving and financial support for families grew rapidly, as did state interest in the early education and care of children.

There are many possible goals of family policy just as there are many interpretations of how family policy affects family life and vice versa. As Ferrarini (2006: 1) points out, when legislating for services or transfers policy makers modify the bundle of choices available to an individual and so delineate the

range of citizens' personal agency. In addition to individuals' behaviors around family life and employment, other primary 'targets' for intervention on the part of family policy include the relational and caring aspects of family life and the functioning of family as a social and economic institution. Kaufmann (2002: 426–428) identified the following classic motivations for family policy within and across countries:

- the natalist motive – the importance of keeping the birth-rate up and managing demographic change;
- the economic motive – the improvement of the stock of human capital and bolstering the family's contribution to economic functioning;
- the socioeconomic motive – balancing of employment (self-sufficiency) and support for the family;
- the societal motive – maintaining the significance of the family as a form of social capital and a source of social stability;
- the sociopolitical motive – compensating for the costs of having families;
- the institutional motive – to protect the family as a valued institution;
- the gender equality motive – to affect the distribution of resources and roles between women and men; and
- the children's welfare motive – to ensure that children are reared with sufficient resources and to improve their welfare through a range of services and provisions.

Family policies were slow to develop. It is generally agreed that state responsibility for families developed later than other areas of social policy (Gauthier, 1996). Among the first family policies were those offering financial assistance to families with children, either through subsidies to wages or through cash allowances for children (known as 'child benefits'). These dated from the 1940s in many countries (and earlier in some) but only in the late 1960s did countries first start to expand family and child care services (e.g., Denmark, Sweden, and later Norway, Finland, Belgium, and France) and other countries universalize child benefits (e.g., Germany, Austria, and Luxembourg). Any slowness was partly due to the difficulty of making policy for such a value-laden sphere as family – which is generally seen as private – and also because there were few actors pushing for family policy and, indeed, a number actively pushing against it (such as in some countries the Catholic Church, which wanted to keep the family as a sphere where Catholic moral influence would prevail).

A range of different actors and providers are involved in family policy. These include, in addition to the public authorities (national, regional, and local), the nonprofit and voluntary sectors (nongovernmental organizations). In some countries, the latter (including church-related bodies) are major providers of family-related services and in others such as France they have played a big role in politicizing family policy. Joining these in service provision are firms and enterprises that have become more important over time.

Different Types of Family Policy Approaches

A comparative approach is almost essential for family policy because it varies so widely. One of the big contributions of comparative welfare state research has been to demonstrate that family policy is part of a much broader set of arrangements and 'settlements' (on the part of powerful actors in society especially) around the functioning of economy and society (Ferrarini, 2006; Hantrais, 2004). This means that, while family policy usually arises to meet needs or solve 'problems,' it is influenced by ideological and normative positions and is almost always tied to a particular arrangement of state, market, polity, and society.

The strength of particular underlying motivations and the diversity of national philosophical orientations toward the family has led to a number of distinct models of family–state–market relations in Europe historically (Daly, 2010; Hantrais, 2004).

A first approach, typified by (West) Germany and Austria, developed in a context where family as an institution had high social value and constituted a legitimate ground for political action. Indeed, the claims of the family were elevated and placed above those of other institutions or social categories by the principle of subsidiarity. The latter decrees that, while the purpose of state support is to enhance family well-being and functioning, such state action should not undermine the autonomy of the family. Social policy in this mold idealized and bolstered the family as an essential source of stability. There was a strong preference for the male breadwinner/female homemaker family model. The mother at home symbolized the appropriate moral order. This, together with the widespread belief that young children should be cared at home, made for a strong gender division of labor. However, it was a widely held belief, and one slow to recede, that families should be financially supported with child rearing and the cost burden of children shared by the society. Hence, this model of the welfare state tended to organize its support of families in the form of income supplements rather than services that tied in with the ideology of mother as carer.

The second classical model was egalitarian, wherein the objective of state policies was to support employment on the part of both female and male parents; and in order to do so, families had to be given access to high-quality child care and other services. This model, found in the Scandinavian countries, did not operate with a strong concept of family as a collective or even separate institution but was much more individualistic (Ellingsaeter and Leira, 2006). Women's role and identity as workers was written into the institutions of state and market as was the principle of individual independence and autonomy. On the one hand women were enabled to be workers and on the other the service-intensive state more or less took over the care of children beyond the toddler stage. Family membership might be a source of emotional stability and identity but family as a privileged social unit was much less supported as compared with other parts of Europe. As a result, this type of welfare state model tended to make public what in other countries is placed beyond the reach of the state. Service provision for families was widespread and children (as well as other members of society) had extensive social rights.

Finally, there are countries such as the United Kingdom (and other liberal countries such as Australia and Canada) where, instead of generic family support, policy draws mainly from an antipoverty and economic functioning orientation.

State intervention in family lacked legitimacy in these countries, apart from preventing or managing social ills such as poverty and family malfunction. The individualism of these cultures and the value placed on financial independence and self-sufficiency, on the part of both individuals and families, tended to veto state intervention for the purposes of generic family support. This rendered family policy relatively underdeveloped; where it existed it was oriented mainly to the prevention of poverty and 'crises' occasioned by familial 'dysfunction.' Policy stayed well away from the social recognition of the family as practiced by French and German policies (O'Connor et al., 1999).

Other models also exist. For example, in the Mediterranean countries, the state tended to desist from putting in place anything other than rudimentary support for families. Without much in the way of cash benefits or tax allowances for their children, families tend to be left to their own devices, which among other things makes for a vibrant flow of support and considerable dependence across the generations.

While these different models might appear like historical artifacts nowadays, in fact the originating philosophy and historical pattern of family policy tend to endure. Hence, new developments are shaped by a country's historical framing of the relationship between family and the state on the one hand and family and economy/society on the other. That said, there is considerable reform and powerful influences are pushing for convergence across countries.

Expenditures on Family Policy

In 2009, the Organisation for Economic Co-operation and Development (OECD) countries on average spent 2.6% of their gross domestic product (GDP) on family policy (OECD Family Database). Across the OECD area, the total public expenditure on family policy ranged from less than 1% to about 4% of GDP. Expenditure on cash benefits was about 1.5% in the OECD and EU countries. However, countries organize their spending quite differently. For example, Denmark and Sweden put an emphasis on parental leave compensation and therefore exhibit only average spending on family allowances despite an overall high expenditure on total cash benefits. The majority of countries spend a higher proportion on cash benefits than on services or tax benefits. Exceptions include Chile, Denmark, France, Iceland, Israel, Italy, Korea, Mexico, Netherlands, Norway, Spain, Sweden, and the United States, where spending on services is the same or higher. Also, the proportion spent on tax breaks toward family is of considerable size in Belgium, the Czech Republic, France, Germany, Japan, the Netherlands, Slovenia, and the United States (more than 0.5% of GDP).

Main Constituents of Family Policy

As mentioned, family policy in most countries has mainly concerned itself with supporting families with children and child-rearing. Raising children incurs both direct and indirect costs, the former arising from the additional costs involved with children and the latter mainly from income foregone from employment because of childbearing and child-minding. Welfare states have long compensated for the former but they were much slower to take on the latter, although moves toward gender equality and concerns about work–family balance have opened policy makers' eyes to indirect costs and family members' preferences. Given the variety of innovations being introduced in family policy, the discussion that follows considers only the main elements of family policy, namely, child benefits, early education and child care services, leaves from employment, and provision by employers or enterprises.

Child Benefits

Historically across Europe child benefits or family allowances were the main pillar of family policy. Designed to assist families with some of the costs of raising children, they date mainly from the period around World War II. At the time of their introduction and subsequently, they represented a truly innovative form of social right since they tended to have no conditions attached to receipt (Montanari, 2000: 309). Usually paid until the child reaches school-leaving age, these allowances exist in every country of the EU today, although their levels vary considerably and they are not universally available to all families. While most countries grant family allowances without income-testing, nine EU member states means-test them (Blum and Rille-Pfeiffer, 2010: 11). Means-testing of these allowances tends to be characteristic of the Mediterranean and postsocialist countries, although the current Conservative/Liberal government in the United Kingdom has also introduced reforms that target the benefit by family income. When countries target, they tend to do so usually on the basis of family income, although some countries also target by the age of the child and number of children. Child benefits are usually funded from general taxation revenues and in this and other ways are an expression of societal support for people who are raising children.

At average earnings, the family benefit package of a family in the OECD countries is worth on average about 10% of the average wage (OECD Family Database). In many countries, the size of the package depends on earnings. In some countries, the family benefit package differs sharply by family size. On average, the benefit for the third child is about 2 percentage points (of an average wage) higher than that for the first child (Lohmann et al., 2009: 54). The situation for single parents in terms of the family benefits package is not advantageous in all countries. Comparing the family benefit packages by family type shows that in Belgium, Sweden, and the Slovak Republic, taxes and benefits place single parents in a less advantageous position as compared with couples without children (Lohmann et al., 2009: 54). In comparison to couples with children, the benefits for single parents are the same or lower in a large number of countries, with some notable exceptions like Austria and Norway. On average, the family benefit packages of single parents are about the same as those of couples with children at the same wage level in the OECD countries (−0.2) and are smaller on average in the EU countries (Lohmann et al., 2009: 54).

Early Childhood Education and Care

This area of provision has become much more important in recent years, associated with both the growth of children's

rights and countries' concerns about the educational performance and general behavior and achievements of their populations. The provision of day care for children is the classic policy response here. However, looking across countries, this is a differentiated phenomenon as some countries view it as care for children, whereas others regard it in terms of education and early learning (Scheiwe and Willekens, 2009). Typically, ECEC focuses on children under compulsory school age and refers to services provided outside the home (in nursery or crèche settings, for example). As Lohmann et al. (2009: 42) point out, the term ECEC originates from the term 'early childhood development,' which emphasizes a holistic approach to the child's physical, emotional, social, and cognitive development. Another commonly used term, 'early childhood education' is favored by educational authorities, who wish to emphasize the learning aspect, and is used especially in the United States (Scheiwe and Willekens, 2009). ECEC services also differ in terms of which ministry they are under; in some countries it is the ministry of education, whereas in others it is the welfare ministry.

The origins and background of ECEC differ fundamentally (in line with the models outlined earlier). In some countries, such as the Nordic countries, ECEC arrangements were initially set up to provide care for the entire age group from birth to compulsory school age during the times when parents were at work. In other countries, such as Germany, ECEC provisions originated in the objective of offering educational or pedagogical services to children aged 3 years and over before they entered primary school. In still other countries such as the United Kingdom, the goal was to offer care and early education services for children from disadvantaged backgrounds. While Lohmann et al. (2009) see the influence of encouraging maternal employment in the services existing today, they also point out that the institutional design of services for young children still in many ways reflects fundamental differences in originating objectives.

When compared across countries, there are other key lines of variation in ECEC policy as well. The main ones include whether there is a legal entitlement, the age range of the children covered, the level of enrollment or service availability, the cost or affordability, and the opening hours (see Plantenga and Remery, 2009). These will be briefly discussed in turn.

Some countries have introduced legal entitlements to day care. Statutory entitlements to child care have been enacted in a number of countries (e.g., Finland, Sweden, Denmark, Norway, Germany, and the United Kingdom) while in Austria, Cyprus, Hungary, Luxembourg, and Poland, the last preschool year has been made compulsory. Most of these guarantee day care services to children over the age of 3 years, but some of the Nordic countries also provide these services to children under age 3 years. As Lohmann et al. (2009: 43) point out, though, a guarantee of service provision does not mean that adequate services are actually available; there may, for example, be long waiting lists to get into day care programs. Most European (and many other) countries saw enormous growth in day care provision in the last 20 years. Among the EU countries, this was prioritized as an area for growth, not least because the European Union set targets for child care provision in 2002; known as the Barcelona targets, these set a target that 33% of the 0- to 2-year-old cohort should be in child care and 90% of those aged between 3 and 6 years by 2010. Child care for older children is more accepted than that for younger children. It is common for countries to differentiate between services for children aged 0–2 years and those between the age of 3 years and school age. Cultural norms around motherhood and parenthood affect the demand for day care and also to some extent the supply. Other factors that have an impact are the levels of employment and unemployment and how these are distributed between women and men.

On the basis of figures for 2008, overall enrollment at a relevant facility is 30% on average for children aged 0–2 years in the OECD and 28.2% in the EU-27 countries (OECD Family Database Chart PF 2.3.A and Table PF 3.2.A). For children aged between 3 and 5 years, on average 77.3% in the OECD area and 81.8% in the EU-27 are enrolled in some kind of care or educational service (Lohmann et al., 2009). Enrollment rates vary widely although the Scandinavian countries tend generally to have the highest levels of enrollment.

In addition to enrollment rates, another source of variation in regard to funding in turn affects the costs to parents. Public funding for ECEC for 0- to 2-year-olds covers on average 82% of total costs in the European Union and in the OECD countries alike (Lohmann et al., 2009). The public share is highest in Ireland (100%) and lowest in the Netherlands (64.5%). For older children, aged 3–6 years, the European Union and the OECD average public funding covers some 94% of the cost.

The average opening hours for facilities for children aged between 0 and 2 years vary from 7 h per day (Spain) to 11.5 h (in Sweden), with an EU and OECD average of 9.3 h a day (Lohmann et al., 2009). For older children, much provision is offered part-time, and opening hours for 3- to 6-year-olds vary from 4 h per day (Greece and Ireland) to 11.5 h (Sweden).

The widespread trend of child care expansion is in some places accompanied by what might be thought of as a counter move: the introduction of a choice-oriented home care allowances. Finland was a pioneer in this regard when in 1985 it instituted a policy that entitled any parent who stayed at home with their 1- to 3-year-old child to receive compensation from the government. A cash benefit for child care is available after the period of paid parental leave has run out and until the child reaches the age of three provided the child does not attend state subsidized child care. Norway introduced a similar provision in 1998. Other countries also are encouraging diversity and choice. France, for example, gives financial support to families to provide child care at home, either by the parent or through the employment of a home-based child care worker. Spain, too, has a benefit encouraging home care – families with three or more children can reduce social security contributions by 45%, if they hire a child minder or domestic worker (Daly, 2011).

Parental and Maternal/Paternal Leaves

Maternity leave is one of the oldest social security benefits, given that maternity has long been recognized as an income and health risk, especially when women are employed. The classic maternity leave – which dates in some parts of Europe from the end of the nineteenth century – was introduced so as to allow women a respite for some weeks prior to and after the child's birth. The leave typically guarantees a period away from employment during which the leave holder's employment

will be protected and also gives some income replacement for foregone wages during leave.

However, Europe is seeing innovation not so much in maternity leave as in other types of family-related leave. From the 1990s on parental leave has become more important. This is a job-protected term of leave provided to enable parents to care for their child, normally during the first year of life. The right to take leave is typically available to the mother and the father, who are usually left to decide between themselves how they will utilize and divide the leave. Sometimes, a certain period is set aside for the second parent.

Increasingly now, countries also offer paternal leave, which is job-protected leave provided to the employed father, normally to be taken near the time of childbirth in order for him to spend time with his newborn child and offer support to the mother and other children in the home. Typically it is of short duration – a number of weeks – and paid at a flat rate. The motivations for its introduction vary. It has been taken forward especially in the Nordic countries by instituting an individual, nontransferable right on the part of each parent to the leave. The so-called father's quota was pioneered in Norway in 1992 to encourage fathers to take child-related leave. Although it is not universal across the Nordic countries, it is regarded as a very important measure toward greater gender equality and improving father–child relationships.

In regard to leave length, the analysis by Lohmann et al. (2009) of provisions in the 38 most developed countries in the mid-2000s indicates that the average length of maternity leave in the EU countries is 17.7 weeks and in the OECD countries 16.1 weeks. Of the countries that provide paid paternity leave, the longest leave is available in Iceland (13 weeks) and the shortest in Austria, Greece, Ireland, Luxembourg, the Netherlands, and Spain, which each award less than 1 week of paid paternity leave. On average, fathers in the EU countries have 1.9 weeks of paid paternity leave available and in the OECD countries 2.8 weeks. Parental leave periods vary from 12 weeks (Belgium and Italy) to 156 weeks (Poland, France, the Czech Republic, and Slovakia), with an EU average of 82.4 weeks and an OECD average of 69.4 weeks.

Employer/Enterprise Provision

As Lohmann et al. (2009: 59) point out, new approaches to national family policy focus much more on employers than former policy approaches did. However some countries, such as the Netherlands, have a tradition of employer support for families and many of the Continental European countries have a history of providing wage subsidies for fathers.

These authors suggest that one can distinguish between sets of workplace practices that either are enforced by law on the national or subnational level or are agreed upon between unions and employers or within the firms themselves. At the firm level, four types of family-friendly working arrangements can be distinguished: flexible working arrangements, special leave arrangements, special child care arrangements, and other supportive arrangements (the framework was originally developed by Den Dulk, 2001).

With regard to flexible working arrangements, which are increasingly common (although this may be changed by the recession), the measures favored include part-time work, flexitime, telework, and a system that allows for banking or saving hours. The European Union played a role by stipulating the principle of nondiscrimination of part-time work in terms of working conditions in the EU part-time work directive 97/81/EC 21. Often part-time arrangements are settled at the level of the firm, but some countries have national legislation.

With regard to leave and child care arrangements, the measures provided at the firm level are often complementary to parental leave and child care provided on the national level. Thus, leave arrangements offered by employers are often provided for family reasons (e.g., sick child). Child care arrangements at the firm level may include workplace nurseries, contractually guaranteed slots in outside child care facilities, or financial assistance (for an overview, see Den Dulk, 2001). Other supportive measures offer possibilities such as work–family management training or employee counseling. However, the general consensus is that these services are not widespread among firms, and there are hardly any representative internationally comparable data on this (Lohmann et al., 2009: 60).

An Overview of General Trends

As we have seen, family policy has a number of different goals and is a dynamic area of policy within and across countries. Until relatively recently, governments were able to take for granted the existence of the traditional family, in terms of both its form and the gendered arrangements of earnings and care on the part of the husband/father and the wife/mother. However, during the last 20 or so years economic and family changes have become major drivers of policy change. Expenditure reflects this, with a spike in spending on family policy in most countries up to the recent recession (OECD, 2011). Spending on young children, and especially in relation to early child care services, was the main focus of growth.

Over time, one sees new emphases in policy and new rationales. While there is considerable variation in reforms being introduced, it is possible to identify a number of common trends.

1. There is a move toward a more diversified set of family supports wherein cash benefits sit alongside tax allowances, services, leave, and employment-based measures to comprise the typical family policy portfolio. So family policy is not just becoming more active, it is also becoming more diverse. This is true both within countries and across them. The field of family policy is broadening and deepening. Diversification is also seen in terms of the providers of services to families with an increased role for the private sector (especially for the profit sector and employers) and a reduced role for the public authorities as providers of services (although they continue to be major funders).
2. There is also a discernible move toward children's rights and a greater concern with children's early education, development, and well-being. The general concerns around children and childhood have led to both improved access for children to education and welfare services and the introduction of social and political rights for children (with a special emphasis on their right to participate in decisions that directly affect them). A notable development has been the growth and in some cases intensification of services oriented

to children's early development (e.g., ECEC, parenting support) and an increased focus on measures to combat poverty and disadvantage among children. As the OECD (2011: 66) puts it, policy is increasingly treating early childhood as a distinct period of life.

3. A further identifiable trend is toward a closer focusing of state support on parenthood and parenting. Law and policy today tend to emphasize parental or biological responsibilities over relations of marriage. The financial and care-related obligations of fathers are especially in focus. There is also a move away from a tax privileging of marriage and a tendency to grant nonmarried couples similar rights and responsibilities as married couples in regard to children.

Particular rationales underlie these and other developments. It seems to me that at the present time, two approaches are predominant. One is the social investment approach. While it has other defining emphases, scholars like Jane Jenson (2004) and Ruth Lister (2006) have read the focus on children and child-rearing as centripetal to a new welfare paradigm in Europe. This 'new' welfare architecture is oriented toward children and the reproduction of the labor force, viewing social expenditures as most justifiable when they lead to the accumulation of 'capital.' Childhood has become politicized and children a potential asset to be 'harvested' in the future.

A second rationale, more a policy aim than a worked out paradigm, is work–family reconciliation. The general intent here is to balance or 'reconcile' working with family life. This has had two main expressions in policy. On the one hand, employed parents are given increased incentives and support to take time off work to care for their children. Both parents can be targeted in this regard in an attempt to affect the share of unpaid done by each but there is increased interest in encouraging fathers to take (short periods of) time off work to care for their very young children (as per the fathers' leave mentioned above). On the other hand, there is a general move toward activating people to be employed so as to raise the level of gainful activity (and associated sense of personal responsibility). This could be said to have a 'work–family reconciliation' character in that both female and male parents are increasingly encouraged to be economically active. It will be obvious that there is a certain tension – if not contradiction – between the two types of 'reconciliation.'

International policy actors (especially the European Union and OECD) have played a major role in driving forward and legitimizing these (and other) philosophies. While their main concern lies with broader economic and/or political goals, both the European Union and the OECD have been encouraging countries to develop a family policy that removes the barriers to labor market participation on the part of different family member (mothers especially) and enables child development. Of the two, the EU has gone furthest – setting ambitious targets (on both child care and female employment) and introducing laws compelling employment leaves for the purpose of child care.

All of this leads to one major set of conclusions. Family policy is a complex field and a domain of policy that is contested and fought over. It is also a sphere of social policy that reaches deep into society; in other words, it has many causes and consequences and normally involves economic, political, and social concerns.

See also: Family Theory: Competing Perspectives in Social Demography; Family as Institution; Female Power at Work; Gender, Economics of; Income Distribution and Inequality: Measurement Issues; Welfare Economics; Welfare Policy Comparisons; Welfare Retrenchment.

Bibliography

Bahle, T., 2009. Public child care in Europe: historical trajectories and new directions. In: Scheiwe, K., Willekens, H. (Eds.), Child Care and Preschool Development in Europe – Institutional Perspectives. Palgrave Macmillan, London, pp. 23–42.

Blum, S., Rille-Pfeiffer, C., 2010. Major Trends of State-Family Policies in Europe. Family Platform, Working Report April 2010.

Crouch, C., 1999. Social Change in Western Europe. Oxford University Press, Oxford.

Daly, M., 2010. Families versus state and market. In: Castles, F.G., Leibfried, S., Lewis, J., Obinger, H., Pierson, S. (Eds.), The Oxford Handbook of the Welfare State. Oxford University Press, Oxford, pp. 139–151.

Daly, M., 2011. What adult worker model? A critical look at recent social policy reform in Europe from a gender and family perspective. Social Politics 18 (1), 1–23.

Den Dulk, L., 2001. Work-Family Arrangements in Organisations: A Cross-National Study in the Netherlands, Italy, the United Kingdom and Sweden. Rozenberg Publishers, Amsterdam.

Eisenstadt, N., 2011. Providing a Sure Start How Government Discovered Early Childhood. Policy Press, Bristol.

Ellingsaeter, A.-L., Leira, A., 2006. Politicising Parenthood in Scandinavia: Gender Relations in Welfare States. The Policy Press, Bristol.

Ferrarini, T., 2006. Families, States and Labour Markets. Institutions, Causes and Consequences of Family Policy in Post-war Welfare States. Edward Elgar, Cheltenham.

Gauthier, A.H., 1996. The State and the Family. A Comparative Analysis of Family Policies in Industrialized Countries. Clarendon Press, Oxford.

Gauthier, A.H., 1999. The sources and methods of comparative family policy research. Comparative Social Research 10, 31–56.

Hantrais, L., 2004. Family Policy Matters. Responding to Family Change in Europe. Policy Press, Bristol.

Jenson, J., 2004. Changing the paradigm: family responsibility or investing in children. Canadian Journal of Sociology 29 (2), 169–192.

Kamerman, S., Kahn, A.J., 1978. Family Policy: Government and Families in Fourteen Countries. Columbia University Press, New York.

Kaufmann, F.-X., 2002. Politics and policies towards the family in Europe: a framework and an enquiry into their differences and convergences. In: Kaufmann, F.-X., Kuijsten, A., Schulze, H.-J., Strohmeier, H.-J. (Eds.), Family Life and Family Policies in Europe, Problems and Issues in Comparative Perspective, vol. 2. Oxford University Press, Oxford, pp. 419–490.

Lister, R., 2006. Children (but not women) first: new labour, child welfare and gender. Critical Social Policy 26 (2), 315–335.

Lohmann, P., Peter, F.H., Rostgaard, T., Spiess, C.K., 2009. Towards a Framework for Assessing Family Policies in the EU, OECD Social, Employment and Migration Working Papers, No. 88. OECD, Paris. http://dx.doi.org/10.1787/223883627348.

Montanari, I., 2000. From family wage to marriage subsidy and child benefits: controversy and consensus in the development of family policy. Journal of European Social Policy 10 (4), 307–333.

O'Connor, J.S., Orloff, A.S., Shaver, S., 1999. States, Markets, Families Gender, Liberalism and Social Policy. Cambridge University Press, Cambridge. OECD Family Database at: www.oecd.org/social/family/database.

OECD, 2011. Doing Better for Families. OECD, Paris.

Plantenga, J., Remery, C., 2009. The Provision of Childcare Services. A Comparative Review of 30 European Countries. European Communities, Brussels.

Scheiwe, K., Willekens, H. (Eds.), 2009. Child Care and Preschool Development in Europe. Palgrave Macmillan, Basingstoke.

Relevant Websites

http://www.coe.int/t/dg3/familypolicy/database/default_en.asp – Council of Europe Family Policy Database.

http://ec.europa.eu/social/main.jsp?catId=815&langId=en – MISSOC the EU's Mutual Information System on Social Protection.

http://www.oecd.org/social/family/database – OECD Family Database.

Family Size Preferences

Elizabeth Thomson, Stockholm University, Stockholm, Sweden; and University of Wisconsin–Madison, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Family size preferences may be represented by family size desires, intentions, expectations, and ideals. Desired family size is the number of children wanted in one's lifetime and is viewed as a measure of the demand for children which, in combination with the supply of children and contraception determines the number of children born. Family size desires are conceptually and empirically distinct from family size intentions, expectations, and ideals. Desires, intentions, and expectations are often uncertain and change over the life course in response to new experiences and conditions. Ideals are shaped by the larger society and therefore change more slowly. Because childbearing is a two-person event, desires and intentions of sexual partners must be considered. As predictors of aggregate or individual fertility, family size desires must be adjusted to take into account partners' desires and changes over the life course, as well as fecundity, fertility control, and other constraints on individual behavior or outcomes.

Conceptualization and Measurement

Desired, Intended, Expected, and Ideal Family Size

Desired family size is conceptualized as the demand for children, that is, "the number of children parents would have if there were no subjective or economic problems involved in regulating fertility" (McClelland, 1983: p. 288). Desired family size is the basis for estimating the level of wanted versus unwanted fertility (Bongaarts, 1990, 2001). Family size intentions, on the other hand, reflect a decision to implement specific behaviors required to achieve a particular number of births (Miller, 2011). As such behaviors occur in sequence over a relatively long period of time, the concept of intention is better suited to a particular birth than to cumulative family size. Family size expectations incorporate both family size desires and perceptions of the likelihood that preferences can be translated into outcomes. Intentions and expectations are conceptually very close, but expectations incorporate a greater degree of uncertainty, especially with respect to changes in desires and circumstances and degree of fertility control. In settings with high levels of fertility control and for persons with high levels of efficacy, intentions and expectations may be empirically indistinguishable. The gap between desires and intentions or expectations arises from partner's desires, perceived fecundity and fertility control, or other known constraints on fertility behavior or outcomes (Bongaarts, 1990; Miller, 2011).

Ideal family size is of a quite different character. Personal ideas would be equivalent to desires, if there were no constraints on having children and perfect fertility control. In low-fertility contexts where costs of children and fertility control are high, personal ideal family size is larger than expected family size (Testa, 2006). Family size ideals are most often, however, specified for the population at large and in that respect constitute a social norm (Trent, 1980; Mason, 1983; Hagewen and Morgan, 2005; Testa and Grilli, 2006). Livi Bacci (2001) suggests that when fertility is very low, desired family size may not be a good indicator of fertility demand because it is strongly influenced by stereotypes of the normative two-child family. The same argument could be made for social contexts in which fertility is very high (Nauck, 2007).

Family size desires, intentions, expectations, and ideals are generally specified in terms of a single number of children. Terhune and Kaufman (1973) argued that family size desires should be conceptualized as a utility function with varying values across parities. For some, the function might have a clear peak, perhaps even only one family size with positive value. For others, the function might be relatively flat, with only extreme family sizes producing negative value. A similar function could be specified for family size expectations where each possible family size is assigned a probability, ranging from zero to one. The fact that one cannot intend two alternative outcomes at the same time makes it impossible to conceive family size intentions in a parallel fashion, again suggesting that intentions are better suited to sequential rather than cumulative childbearing. Ideal family size by definition requires a single choice among family sizes, but the aggregate distribution in a population of ideal family size can be viewed as a functional representation of the strength as well as content of a childbearing norm.

Measurement

Two types of questions are most commonly used to measure family size desires. The first asks respondents how many children they would want to have in their lifetime, if they could choose the exact number. For childless respondents, this question measures desires prospectively; for parents, the question includes a retrospective assessment of previous births with a prospective assessment of potential future births. Retrospective assessments of birth wantedness are subject to social desirability bias or rationalization and therefore produce upwardly biased estimates of desired family size (Bongaarts, 1990; Williams and Abma, 2000). Another version of this question attempts to give respondents permission to report previous births as unwanted by asking something like: "If you could start life over again, how many children would you want to have?" McClelland (1983) suggests that such questions produce less reliable and valid measures of fertility demand than prospective questions.

The second type of question asks respondents how many more children they would like to have; for childless respondents, the answer would presumably be the same as for the first type of

question, but for parents, retrospective assessments of previous births are excluded. To obtain an estimate of desired family size, some analysts simply add the respondent's current number of children to the additional number desired. This strategy assumes that all previous births were wanted, producing an even stronger upward bias than does the first type of measure.

Measures of family size desires that allow for a range of acceptable, possibly equally preferred family sizes require more questions (Terhune and Kaufman, 1973; Coombs, 1974; Ballweg and Klemmack, 1984; Hin et al., 2011). Scales based on unfolding theory (I-scales) distinguish preferences for number (IN) and sex (IS) of children (Coombs et al., 1975). Despite their superiority to single-choice measures in terms of reliability and predictive validity (see below), such measures have seldom been used in large-scale surveys.

Family size intentions, expectations, and ideals are also almost uniformly measured with questions that ask for a single choice. As intentions or expectations are oriented to the future, family size intentions or expectations are measured as the sum of the number of children already born and the number of any additional intended or expected children.

When family size desires, intentions or expectations are specified in terms of a specific birth, measurement is generally more precise. Desires may be measured along dimensions of intensity, centrality, and certainty (Schaeffer and Thomson, 1992; Thomson and Brandreth, 1995) and may incorporate both positive and negative components (Miller and Pasta, 1995). Intentions may be specified in terms of certainty (Thomson, 1997) and expectations in terms of likelihood or probability (Thomson and Brandreth, 1995).

Nonnumerical Responses and Uncertainty

In high-fertility societies, small but significant proportions of respondents do not provide a numerical response to questions about desired family size, reporting instead that the number of children they have is 'up to God' or that they will accept whatever number of children they have. These responses represent a completely flat family size utility function both theoretically (Terhune and Kaufman, 1973) and empirically (Jensen, 1985). Because such respondents are willing to express numerical family size ideals or preferences for one family size over another, and do report unwanted or mistimed births (McClelland, 1983; Riley et al., 1993), the 'up to God' response might reflect moral objections to, or lack of possibilities for fertility control. In a context of high child mortality, 'up to God' may also refer to uncertainties about the number of children who will survive (Sandberg, 2005).

Even where individuals have high control over childbearing, they may be uncertain about how many children they want, whether they want any children at all, and whether they want a second, third, or higher order birth (Schaeffer and Thomson, 1992). They might perceive a close balance between the positive and negative outcomes of childbearing or simultaneously experience strong positive or negative motivations for children (Miller, 2011), and therefore be uncertain how they really feel or what they really want.

When respondents are asked to state birth intentions – i.e., a plan to carry out behaviors that will result in a birth or not – they often express uncertainty. Uncertainty is inherent in the theoretical construct of fertility expectations as unknown changes in external conditions and events must be taken into account. Uncertainty in intentions or expectations may arise from uncertain or weak desires or from uncertainty about future conditions for rearing children or success with fertility control (Ní Bhrolcháin and Beaujouan, 2012). As Ní Bhrolcháin and Beaujouan (2011: p. 120) point out, "… uncertainty is a reasonable response to the many contingencies and unknowns that people encounter through the life cycle …."

Uncertainty about ideal family size should theoretically arise only if one is ignorant of the conditions and behaviors in the society at large. Those who do not know or refuse to provide an ideal family size are most likely to be expressing acceptance of a range of family sizes. In many countries, a small but significant proportion of respondents assert that there is no ideal family size (Hagewen and Morgan, 2005; Testa, 2006).

Changes in Family Size Desires, Intentions, Expectations, and Ideals

A major criticism of the concept of desired or intended or expected family size is that births usually come one at a time. Each birth provides new information that may influence future childbearing desires, intentions, and expectations (Namboodiri, 1983). Other experiences also intervene to change the perceived benefits and costs and the desire, intention, or expectation of having a particular number of children.

Miller (2011) argues that desires for children are more stable than intentions and expectations because they are linked to personality traits. Ní Bhrolcháin and Beaujouan (2012) argue to the contrary that the pervasive ambiguity in fertility desires is evidence that they are socially constructed over the life course. They go so far as to suggest that desires for children may be constructed in the process of answering survey questions (see also Schaeffer and Thomson, 1992). In both high- and low-fertility contexts, desires for children shift over relatively short periods of time (Bankole and Westoff, 1998; Miller and Pasta, 1995; Heiland et al., 2008).

The evidence is clear that family size intentions or expectations change over the life course (e.g., Freedman et al., 1965; Westoff and Ryder, 1977; Quesnel-Vallée and Morgan, 2003; Hayford, 2009; Liefbroer, 2009; Morgan and Rackin, 2010; Iacovou and Tavares, 2011). Changes are associated with experiences of the first birth, the influence of partners or educational and career trajectories. Changes in intentions or expectations generally occur within the range of normative family size (Freedman et al., 1965; Bankole and Westoff, 1998; Morgan and Rackin, 2010).

Changes in ideal family size may also occur at the individual level, but as indicators of change in social norms are of greater interest at the societal level. In low-fertility countries, ideal family size has been exceptionally stable, with the vast majority specifying two children. In recent years, significant downward shifts have been identified in a few European contexts (Sobotka, 2009). Goldstein et al. (2003) argued that persistent fertility under two children per women could shift ideal family size to well under two.

Her, His, and Their Family Size Desires, Intentions, and Expectations

Most data on family size desires, intentions, or expectations come from interviews with women, while ideal family size is more often asked of women and men in general-purpose social surveys. The assumptions underlying women-only fertility surveys are that (1) women and men share the same conditions of life and will therefore want, intend, and expect the same number of children; (2) because women bear children and are primary child rearers, their desires and intentions are those that determine fertility; and/or (3) women's desires, intentions, and expectations will incorporate those of their partners, reflecting resolution of any initial disagreements. In the aggregate, women's and men's respective family size desires and intentions are more similar than different (Mason and Taj, 1987; Bankole and Singh, 1998; Testa, 2006). Data collected from couples, on the other hand, consistently show that substantial proportions disagree about desires, intentions, or expectations for further childbearing or completed family size (e.g., Coombs and Fernandez, 1978; Thomson et al., 1990; Becker, 1996; Thomson, 1997; Bankole and Singh, 1998; Mason and Smith, 2000). Furthermore, when partners express diverging desires for children, they influence one another's intentions to have a child (Morgan, 1985; Miller and Pasta, 1996; Thomson, 1997). Thus, couple disagreement is a primary source of gaps between desires and intentions or expectations.

Family Size Desires, Intentions, and Births

The relationship between family size desires and births depends on relationships between desires and intentions, intentions and behavior, and behavior and outcomes. In addition to couple disagreement, a key source of gaps between desires and intentions or expectations is the perception of control over fertility. Those who believe themselves to be infecund may desire more children than they intend or expect; those who do not perceive themselves to be able to control fertility may intend or expect more children than they desire. Perceived social norms are also potential constraints on the translation of desires into intentions and subsequent behavior (Billari et al., 2009).

Measures of family size desires, particularly the desire for additional children, are strongly associated with eventual births at the aggregate levels (Bongaarts, 1990; Pritchett, 1994). Lee (1980) has demonstrated that a 'moving target' model of desired and achieved fertility produces even stronger correspondence between family size desires and period fertility. Bongaarts (1990) developed a model for adjusting average desired family size for upward and downward biases; the adjusted value is even more strongly associated with the total fertility rate than the initial value.

At the individual level, the relationship between childbearing desires or intentions and subsequent births is also quite strong (Westoff and Ryder, 1977; Schoen et al., 1999). The strength of the relationships depends on the length of time between observation of desires or intentions and births, the strength of desires or certainty of intentions, the capacity for fertility control, the strength of opposing social norms and other conditions that might inhibit the translation of desires into behaviors (Davidson and Jaccard, 1979; Philipov, 2009; Hayford and Agadjanian, 2012). A key element is an agreement between the sexual partners. In virtually every study that includes the desires or intentions of both partners, their joint childbearing is influenced by the desires or intentions of each (e.g., Becker, 1996; Miller and Pasta, 1996; Thomson, 1997; Thomson and Hoem, 1998; Bankole and Singh, 1998; Becker, 1999; Mason and Smith, 2000).

Estimated relationships between fertility desires, intentions, expectations, and births also depend on measurement quality. The IN-scale discussed above provides more accurate measures of preferences and therefore has larger estimated effects on fertility behavior than does the single most preferred number of children (Coombs, 1979). Measures of the desire for children that incorporate dimensions of attitude strength have stronger effects on childbearing intentions and behavior than do yes-no-do not know responses to direct questions about desire for children (Thomson and Brandreth, 1995). Similarly, intentions or expectations measured with some degree of certainty or uncertainty are better predictors of births than a simple dichotomy (Thomson, 1997).

See also: Attitudes and Behavior; Children, Value of; Demographic Measurement: General Issues and Measures of Fertility; Norms; Second Demographic Transition.

Bibliography

Ballweg, J., Klemmack, D., 1984. Dimensions and crystallization of family-size norms. Sociological Spectrum 4, 1–15.
Bankole, A., Singh, S., 1998. Couples' fertility and contraceptive decision-making in developing countries: hearing the man's voice. International Family Planning Perspectives 24, 15–24.
Bankole, A., Westoff, C.F., 1998. The consistency and validity of reproductive attitudes: evidence from Morocco. Journal of Biosocial Science 30, 439–455.
Becker, S., 1996. Couples and reproductive health: a review of couple studies. Studies in Family Planning 27, 291–306.
Becker, S., 1999. Measuring unmet need: wives, husbands or couples? International Family Planning Perspectives 25, 172–180.
Billari, F.C., Philipov, D., Testa, M.R., 2009. Attitudes, norms and perceived behavioural control: explaining fertility intentions in Bulgaria. European Journal of Population 25, 439–465.
Bongaarts, J., 1990. The measurement of wanted fertility. Population and Development Review 16, 487–506.
Bongaarts, J., 2001. Fertility and reproductive preferences in post-transitional societies. Population and Development Review 27 (Suppl.), 260–281.
Coombs, L.C., 1979. Reproductive goals and achieved fertility: a fifteen-year perspective. Demography 16, 523–534.
Coombs, L.C., 1974. The measurement of family size preferences and subsequent fertility. Demography 11, 587–611.
Coombs, C.H., Coombs, L.C., McClelland, G.H., 1975. Preference scales for number and sex of children. Population Studies 29, 273–298.
Coombs, L.C., Fernandez, D., 1978. Husband-wife agreement about reproductive goals. Demography 15, 57–73.
Davidson, A.R., Jaccard, J.J., 1979. Variables that moderate the attitude-behavior relation: results of a longitudinal survey. Journal of Personality and Social Psychology 37, 1364–1376.
Freedman, R., Coombs, L.C., Bumpass, L., 1965. Stability and change in expectations about family size: a longitudinal study. Demography 2, 250–275.
Goldstein, J., Lutz, W., Testa, M.R., 2003. The emergence of sub-replacement family size ideals in Europe. Population Research and Policy Review 22, 479–496.
Hagewen, K.J., Morgan, P.S., 2005. Intended and ideal family size in the United States, 1970–2002. Population and Development Review 31, 507–527.

Hayford, S.R., 2009. The evolution of fertility expectations over the life course. Demography 46, 765–783.
Hayford, S.R., Agadjanian, V., 2012. From desires to behavior: moderating factors in a fertility transition. Demographic Research 26, 511–542.
Heiland, F., Prskawetz, A., Sanderson, W.C., 2008. Are individuals' desired family sizes stable? Evidence from West German panel data. European Journal of Population 24, 129–156.
Hin, S., Gauthier, A., Goldstein, J., Buehler, C., 2011. Fertility preferences: what measuring second choices teaches us. Vienna Yearbook of Population Research 9, 131–156.
Iacovou, M., Tavares, L.P., 2011. Yearning, learning, and conceding: reasons men and women change their childbearing intentions. Population and Development Review 37, 89–123.
Jensen, E., 1985. Desired fertility, the 'up to God' response, and sample selection bias. Demography 22, 445–454.
Liefbroer, A.C., 2009. Changes in family size intentions across young adulthood: a life-course perspective. European Journal of Population 25, 363–386.
Lee, R.D., 1980. Aiming at a moving target: period fertility and changing reproductive goals. Population Studies 34, 205–226.
Livi Bacci, M., 2001. Comment: desired family size and the future course of fertility. Population and Development Review 27 (Suppl.), 282–289.
Mason, K.O., 1983. Fertility norms. In: Bulatao, R.A., Lee, R.D. (Eds.), Determinants of Fertility in Developing Countries, Supply and Demand for Children, vol. 1. Academic Press, New York.
Mason, K.O., Taj, A.M., 1987. Differences between women's and men's reproductive goals in developing countries. Population and Development Review 13, 611–638.
Mason, K.O., Smith, H.L., 2000. Husbands' versus wives' fertility goals and use of contraception: the influence of gender context in five Asian countries. Demography 37, 299–311.
McClelland, G.H., 1983. Family-size desires as measures of demand. In: Bulatao, R.A., Lee, R.D. (Eds.), Determinants of Fertility in Developing Countries, Supply and Demand for Children, vol. 1. Academic Press, New York.
Miller, W.B., 2011. Differences between fertility desires and intentions: implications for theory, research and policy. Vienna Yearbook of Population Research 9, 75–98.
Miller, W.B., Pasta, D.J., 1995. How does childbearing affect fertility motivations and desires? Social Biology 42, 185–198.
Miller, W.B., Pasta, D.J., 1996. Couple disagreement: effects on the formation and implementation of fertility decisions. Personal Relationships 3, 307–336.
Morgan, S.P., 1985. Individual and couple intentions for more children: a research note. Demography 22, 125–132.
Morgan, S.P., Rackin, H., 2010. The correspondence between fertility intentions and behavior in the United States. Population and Development Review 36, 91–118.
Namboodiri, N.K., 1983. Sequential fertility decision making and the life course. In: Bulatao, R.A., Lee, R.D. (Eds.), Determinants of Fertility in Developing Countries, Supply and Demand for Children, vol. 1. Academic Press, New York, pp. 444–472.
Nauck, B., 2007. Value of children and the framing of fertility: results from a cross-cultural comparative survey in 10 societies. European Sociological Review 23, 615–629.
Ní Bhrolcháin, M., Beaujouan, É., 2011. Uncertainty in fertility intentions in Britain, 1979–2007. Vienna Yearbook of Population Research 9, 99–129.
Ní Bhrolcháin, M., Beaujouan, É., 2012. How Real Are Reproductive Goals? Uncertainty and the Construction of Fertility Preferences. Paper presented at the annual meetings of the Population Association of America, San Franciso, CA.
Philipov, D., 2009. Fertility intentions and outcomes: the role of policies to close the gap. European Journal of Population 25, 355–361.
Pritchett, L., 1994. Desired fertility and the impact of population policies. Population and Development Review 20, 1–55.
Quesnel-Vallee, A., Morgan, S.P., 2003. Missing the target? Correspondence of fertility intentions and behavior in the U.S. Population Research and Policy Review 22, 497–525.
Riley, A.P., Hermalin, A.I., Rosero-Bixby, L., 1993. A new look at the determinants of nonnumeric response to desired family size: the case of Costa Rica. Demography 30, 159–174.
Ryder, N.B., Westoff, C.F., 1967. The trend of expected parity in the United States: 1955, 1960, 1965. Population Index 33, 153–168.
Sandberg, J., 2005. The influence of network mortality experience on nonnumeric response concerning expected family size: evidence from a Nepalese mountain village. Demography 42, 737–756.
Schaeffer, N., Thomson, E., 1992. The discovery of grounded uncertainty: developing standardized questions about strength of fertility motivation. Sociological Methodology 22, 37–82.
Schoen, R., Astone, N.M., Kim, Y.J., Nathanson, C.A., Fields, J.M., 1999. Do fertility intentions affect fertility behavior? Journal of Marriage and the Family 61, 790–799.
Sobotka, T., 2009. Sub-replacement fertility intentions in Austria. European Journal of Population 25, 387–412.
Terhune, K.W., Kaufman, S., 1973. The family size utility function. Demography 10, 599–618.
Testa, M.R., 2006. Childbearing Preferences and Family Issues in Europe. Special Eurobarometer 253/Wave 65.1, TNS Opinion & Social. European Commission.
Testa, M.R., Grilli, L., 2006. The influence of childbearing regional contexts on ideal family size in Europe: a multilevel analysis. Population 61, 109–137.
Thomson, E., 1997. Couple childbearing intentions and births. Demography 34, 343–354.
Thomson, E., Brandreth, Y., 1995. Measuring fertility demand. Demography 32, 81–96.
Thomson, E., Hoem, J.M., 1998. Couple childbearing plans and births in Sweden. Demography 35, 315–322.
Thomson, E., McDonald, E., Bumpass, L.L., 1990. Fertility desires and fertility: hers, his, and theirs. Demography 27, 579–588.
Trent, R.B., 1980. Evidence bearing on the construct validity of 'ideal family size'. Population and Environment 3, 309–327.
Westoff, C., Ryder, N., 1977. The predictive validity of reproductive intentions. Demography 14, 431–453.
Williams, L., Abma, J., 2000. Birth wantedness reports: a look forward and a look back. Social Biology 47, 147–163.

Family: The Anthropology of the Concept and Its History

Jean La Fontaine, University of London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The family is both a significant institution in Western culture and a technical term used in analysis. This article deals with both aspects. A consideration of its significance in Western social life first establishes its importance. Then, a history of the anthropology of the family shows the difficulties entailed in trying to find a neutral approach to facilitate a more cross-cultural and theoretical understanding.

Introduction

The family is both a significant institution in Western culture and a technical term in cross-cultural analysis. From the start, its great importance in Western culture has given it a central place in anthropology. However, a series of problems arise when 'the family' is used as a basis of comparison with other societies. Simplifying the concept by identifying one or other of its component elements as fundamental has not solved the problems in cross-cultural analysis, but has preserved the illusion that 'the family' is a useful theoretical tool. A history of the anthropology of the family shows the difficulties entailed in trying to find a neutral approach to it and a general theoretical understanding of the ideas and practices it entails.

The Family in Western Society

The significance of the family in Western societies is obvious in daily life. Children in primary school are asked to describe their families as a writing exercise and politicians vie with one another to be 'the party of the family.' Christianity, the majority religion of Western society, celebrates the Holy Family and its priests proclaim the duty of Christians to emulate it in their domestic life. The private life of individuals is identified with 'the family'; it is 'the building block of society,' given equal weight with the public institutions of the state, since the division into public and private spheres of life is one of the fundamental social distinctions in Western culture.

The social unit to which the term family is usually applied consists of parents and children; the importance of this unit is signaled by the fact that the terms used by its members for one another are not applied to any other relatives, which is not the case in other societies; the relationships these terms designate are seen as unique. Sometimes labeled the 'nuclear family' this form may be described as 'natural,' the consequence of the long period of maturation required by the human child, rather than a socially defined unit and those adults who have neither spouses nor children might once have been regarded as not fully adult.

The idea that the family is a natural unit and perceived as such has been a difficulty in cross-cultural comparison. This idea as a misperception can be readily shown by listing the multiple uses of the term in common speech. Despite its importance, the word 'family' has no single, clearly defined meaning. In twenty-first century England, in ordinary speech, 'family' may evoke any one of a cluster of meanings, depending on context. Harris (1983: p. 31) lists six different meanings ranging from parents and children to status and historical continuity. It is difficult to claim that any of these meanings is the 'true' meaning and others are merely derivative usages, although some people might argue that the unit of parents and children is the primary sense of the term. However, there is evidence to show that this meaning is relatively recent in English history. Davidoff and Hall (1987) trace its importance to the rise of the industrially based middle-class in the nineteenth century. Before that time, 'family' referred to the residents of a single household, including the kin and servants of its recognized head. The association of residence with 'family' persisted and the residential dimension was transferred to the new 'nuclear family' despite the fact that household and family were not always identical in composition. As historical research has subsequently shown, households have long varied in composition from small nuclear units to large communities containing servants and other employees, relatives, and foster-children.

The family is not an isolate but part of a system of relations traced more widely, but based on the same concepts. This may be referred to as the extended family but anthropologists use the term kinship. Networks of kin ramify and are spread widely, genealogically and geographically (*see* Kinship in Anthropology). In most societies in the world, kinship is more significant than 'family' and kinship relations may underlie social frameworks of much wider social significance than the household. The comparison has also shown that the distinction made between relationships within and outside the family has obscured the significance of the kinship system in Europe, and it has drawn very little attention even from anthropologists.

The Anthropology of the Family

The issues with which the anthropology of the family has been associated reflect ideas implicit in the Western concept. The polyvalence of the meanings associated with 'the family' results in a variety of conceptual problems for the would-be neutral anthropologist.

At the outset the focus was developmental; it gave priority to the setting up of 'a family' that is marriage, which was then considered to be the 'foundation' of the family. In the nineteenth century, the scholars who developed anthropology regarded monogamous lifelong marriage as a characteristic of 'civilization,' the culmination of human evolution. The Western family was seen as a late development not to be found among peoples who represented earlier stages of the evolutionary process. These latter were characterized by plural marriage and their classification of related individuals was thought to show failure to understand the nature and significance of biological links. Nineteenth century evolutionism has long been recognized as lacking a sound empirical basis for its conjectures, but the idea that the present form of 'the family' is a product of social changes over time has not been entirely abandoned. It survived in a modified form in the view that the present form of 'the family' is a product of modernization, of capitalism, or of industrial society and, by contrast, in the myth that Europe's peasant forebears all lived in extended family groups.

By the middle of the twentieth century, some anthropologists argued an opposed view, claiming that the family could be said to be universal, even though contained within many different forms of residential unit. The basis for this conclusion was the assumption that 'the family,' so far from being associated with marriage in a particular cultural form, was based on a universal natural function, reproduction. It was this 'biological fact' that formed the basis of the family, a view characteristic of the societies from which anthropologists came. 'The family,' a couple with their children, might be incorporated in various different domestic arrangements, but was itself a biological 'given.' Whether any individual society recognized this or not was said to depend on their culture, a view not totally unlike to that of the evolutionists with whom they disagreed. Murdock (1949) proclaimed the nuclear family to be universal, a coresidential unit with 'sexual, economic, reproductive, and educational' functions. His conclusions depended on ignoring a good deal of information on the wide variety of social units associated with these functions. The procedure showed all too clearly the looseness of the term 'family.' Malinowski, an earlier proponent of this functional view, had nevertheless cautioned that "…the essential features of the individual family depend upon the general structure of a given society…" (1913: p. 6).

A consideration of social structure built up on the roles associated with membership of a family introduced new factors. Starting from marriage as the founding of a new family, the tie between spouses can also be seen as linking distinct 'families,' since husband and wife do not lose their roles in the family in which they were children. Individuals are thus members of at least two 'families': one in which they are children and (usually) siblings and one in which they are spouses and (usually) parents. Sociologists have labeled these the family of origin and the family of procreation, which emphasizes the different methods by which the focal individuals are recruited, i.e., birth and marriage.

Anthropological research in other societies revealed different associations of kinship and household structure which made it necessary to develop a terminology for cross-cultural comparison. The establishment of different types of 'families' such as the 'stem family' of Japan (consisting of one married pair from each of two, or even three, successive generations and their unmarried children it is also to be found in Eastern Europe and extending from the rural areas of Eastern and Southern Europe) or the 'extended family,' meaning large households containing three or more generations, which can still be found in India and Africa. The term 'matrifocal family' was coined by R.T. Smith to describe the Caribbean household of mother daughter(s) and grandchildren with men only visiting or temporarily resident (Smith, 1973). This form is now also found among poor black families in both the USA and Great Britain. A large number of related households living associated as one community may also be referred to as an extended family, identifying their common locality as similar to residence in a single household. All these uses of the term became current in academic works (Janssens, 1993 (it is also to be found in Eastern Europe and extending from the rural areas of Eastern and Southern Europe)).

A solution to the theoretical problems raised by the data on domestic organization was to distinguish between 'the family' as a reproductive unit and the household as a residential group. One of the first to do this was Fortes (1958) who used the term domestic group, referring to the 'family' as the reproductive nucleus within it. His distinction emphasized a number of contrasts between the concepts: 'the family' consists of parents and children, whereas the domestic group might include more than two generations and other kin; the family is recruited through marriage and parenthood, whereas many different kinds of tie may draw individuals into living together. The family is concerned with social and physical reproduction; the domestic group 'is essentially a householding and housekeeping unit' (Fortes, 1958: p. 8). According to Fortes, the differences in the size and structure of domestic groups that can be observed in a particular society do not reflect breaches of any 'rule of residence' as had previously been thought, but reflect the developmental cycle of growth and maturation in the reproductive nucleus, or, in some cases, nuclei (Fortes 1949). They may also display the effect on household structure of the political and economic institutions of the society in question. These processes and the cultural reactions that they entail are the driving forces behind household changes.

The distinction between 'family' and 'household' also drew attention to economic production as a household activity. Property rights and labor needs were shown to influence the form of the domestic group as much as the developmental cycle of the reproductive core. Households were shaped by the demands for different kinds of labor associated with their subsistence activities that might vary according to the season, and the association of related people within different households concerned economic viability as well as kinship obligations (Stenning, 1958). Goody went further and showed that interpersonal relations in the 'reproductive core' were also affected by these factors (Goody, 1956, 1976). Marxist anthropology subsequently attempted to construct different modes of domestic production and reveal the power structures entailed in them, but these largely ignored reproductive processes and were not widely adopted.

In recent years, attention to relationships in households and localities has emphasized the association based on criteria other than kinship. This underlines the difference between family and household by focusing on the generation of relationships among individuals and the meanings attached to the process rather than a formal structure of roles (Carsten, 1995, 1997).

The view that the family was a 'natural' phenomenon had been reflected in a similar view of the universal existence of rules restricting sexual relations within the 'nuclear family' to the conjugal pair. These rules, insofar as they affected the family, were known as the incest taboo; the use of the term 'taboo' denoted their strict nature and implied the pollution that in many societies was believed to follow any breach. It was argued by Westermarck (1891) that these relationships were such that sexual attraction 'naturally' did not arise because the nature of daily interactions destroyed such feelings. Others, like Malinowski, argued *per contra* that the prohibition of sex between siblings and between parents and children was essential to the maintenance of authority and the prevention of conflict within a basic social unit and so was not natural but fundamentally social (Malinowski, 1926, 1940).

The later development of sociobiology returned to the support of the 'natural' basis of the incest taboo, claiming that the incest taboo could also be observed to exist among the higher primates and was therefore genetic, rather than a feature of human family relations. This, like other forms of biological determinism, was vigorously contested by social anthropologists (Schneider, 1965; Sahlins, 1976). In particular, it was pointed out that paternity is not identified by the higher primates and among human beings it is always socially determined. The accumulation of data on other societies made clear that sexual prohibitions of an identical nature extend outside the family in all societies and include unrelated people such as affines or step-relations. That this was also the case in Western society had been largely ignored (La Fontaine, 1990). Incest taboos, considered as total systems of prohibitions, varied widely cross-culturally. The prohibition on incest could not therefore be considered biologically determined or 'natural' (Goody, 1956).

Few social or cultural anthropologists now accept that the incest taboo is genetic or universally the same, but the conviction that among close kin it is only rarely broken and that incest is 'unnatural' lingered on. By the end of the twentieth century, however evidence that breaches of the incest taboo were infrequent, but not rare in Western societies and became public knowledge with the reporting of cases of the sexual abuse of children by their fathers or (less often) brothers. This offered proof that the 'taboo' was neither genetic nor unbroken (La Fontaine, 1990). The reality of incest was not easily accepted, whether by the public at large or by social scientists, showing that the assumptions underlying it and also the concept of the family it upheld were deeply rooted in Western culture.

A significant finding of anthropological research in many parts of the world was that the ways in which relatives might be classified was both varied and amenable to analysis as systems, which might then be compared. Kinship categories cut across both reproductive and domestic units so that, for example, it was (and is) not uncommon for women to be addressed as 'mother' who have no biological connection with the speaker. This is not merely a courtesy title as might be the case in some parts of England for example. While the means of identifying the 'birth mother' exists in most societies, what Westerners perceive as a biological tie between two individuals is not as significant in other societies as their relative positions within the kinship system as a whole. In effect, the classification of kinship in other societies and the terminology in which it was expressed did not single out the nuclear family, nor distinguish 'biological' from other kinship relations.

In addition, it became clear that kinship systems incorporated a local view of what constituted distance and closeness of relationships that might be linked to a perception of biology different from that underlying the western family. For example, distinctions are made in many systems between relatives linked through men and those related through women. Father's brother and mother's brother are defined in many societies as different kind of kinsmen and the justification for this may be the conviction that the tie between a child and each of its parents is not equally 'natural.' The extreme view, found in some Australian and Melanesian societies, denies the male contribution to conception altogether so that a father is not believed to have any biological link to his children, their relationship is established by his marriage to their mother. The opposite view, the denial of maternal contribution to the substance of her child occurs in a few patrilineal societies whose members consider women merely incubate the child placed in their womb by its begetter. If the family were to be considered universal, the recognition of a 'natural' unit, then the existence of societies that explicitly deny this would have to be reconciled with the theory.

At intervals during the last 50 years, anthropologists have pointed out the deficiencies of the concept of the family as an analytical tool (Fortes, 1958; Yanagisako, 1979; Schneider, 1984; La Fontaine, 1985). Robertson's solution to the problem is an original one. He starts with an account of the importance of the idea of the family in Western culture; later, he draws attention to social scientists' 'fixation on the family' (1991: p. 154). As he points out these two phenomena are undoubtedly connected. He argues that reproductive functions were wrongly attributed to the family, which is merely part of one version of the social construction maintaining the continuity of any society by ensuring the regular replacement of its aging members by the next generation. In order to compare the various cultural forms that enable reproduction, the concept of the family is discarded. Reproduction, social and physical, is universal but the cultural form that ensures it is society as constituted in particular forms. Robertson's approach implies that all that can be compared as reproductive systems are social wholes, which had been briefly tried and rejected by anthropologists as unwieldy and unproductive (Benedict, 1934). A logical consequence of this position is that the function of physical reproduction should be ignored both as the origin of the family and as an issue in anthropology. Robertson argues that the idea of the family, like other

cultural values with which it is associated, imposes a permanence and order on a reality that is always in flux.

Another approach came to the same conclusion as Robertson but from a different angle. Under the rather more materialist influence of feminism, anthropologists in the 1980s showed how the 'natural' processes of childbirth seemed to be very widely used to symbolize women and legitimize their subordination to men. Within the domestic sphere, also associated with women, the division of labor was seen as reflecting the natural capabilities of men and women; the products of their labor were shared with their children. However, the distribution of goods and the allocation of labor within the household actually reflected its internal power structure rather than a pooling of family resources or a 'natural' division of family labor. The realities of power were presented as unquestionable by being associated with the natures of men and women. Concepts of nature were thus shown to be more than mental constructs; they were also intimately bound up with the social distribution of power and authority, which they helped to legitimize and reproduce. The 'family' and its reproductive function was revealed as a symbolic representation legitimizing household organization (Harris, 1981). Nature stands as a symbol for what is a fundamentally social unit.

These views were stimulated by the work of the French anthropologist Lévi-Strauss' work on symbolism and meaning. His revolutionary approach to kinship used it to dispose of some of the questions that have been raised. He argued (Levi-Strauss, 1958, tr.1963) that exchange of goods and of women, not biological reproduction, was the basis of social life. Hence, marriage rules are integral to the wider social structure, not merely the basis of the family or a means of recruitment to households. Systems where marriage with relatives was the ideal showed how the patterns of distribution of women in marriage constituted social structures. His 'atom of kinship' (Levi-Strauss, 1958, tr. 1963) contained the same elements as 'the nuclear family,' but rearranged so as to emphasize the relation between a man and his wife's brother created by marriage. He argued that the pairs of opposites that were the elements of the 'atom of kinship' reflected basic oppositions in systems of thought, the deep structure of culture.

While Lévi-Strauss' structuralist approach undoubtedly revealed the significance of cultural understandings of the world, he was less interested in distinguishing between the meanings drawn from the analysts' own Western culture and those of the societies they were studying or in comparing them. His work, immensely influential though it was, took the comparative study very little further. However Lévi-Strauss' emphasis on the symbolism of cultural concepts, including that of nature, was fruitfully applied to the study of kinship systems. Schneider analyzed American kinship in terms of the conjunction and contrast between concepts of nature and concepts of law (Schneider, 1968, 1972). 'The family' is the source of basic elements that categorize kinship relations beyond it, extending to relations with affines and more distant kin. Marriage, conceived of as taking place between unrelated persons established the legal–cultural relationship between spouses whereas kinship between parents and children was deemed natural – though only in the sense that the natural world was perceived through the medium of culture. There is thus in Schneider's approach, a distinction to be made between the material nature of the world and cultural constructions of that world (the cultural construct of nature). Ethnoscience, the cultural understanding of the natural world, is to be distinguished from formal science where evidence and proof support conclusions. Schneider makes a further distinction when he states that cultural concepts of nature can be used as symbols to enhance the significance of important elements of culture and this is what he perceives to be the role of 'nature' underlying the concept of the family. Human reproduction conceptualized culturally as parenthood represents nature while marriage, being purely cultural, denotes 'law.' In fact, Schneider remarks "kinship is biology" (Schneider, 1968: p. 116).

Barnes in his critique of Schneider (Barnes, 1973) starts from a Lévi-Straussian opposition between nature and culture, the latter being virtually the same as Schneider's 'law.' He points out that parenthood is not entirely 'natural' as Schneider would have it; 'the relations of nature to fatherhood and motherhood are different' (Barnes, 1973: p. 72). Whereas physical motherhood (the status of genetrix) is demonstrated at birth, no matter who rears the child, no such event defines physical paternity (the genitor). Conception cannot be perceived and was not understood in Europe until relatively late in history, but it was believed to involve both partners long before science had demonstrated the process. If we then consider the cultural elements in the family it becomes clear that fatherhood must be associated with marriage as cultural, not natural. It is marriage that defines the paternity of children whatever the physical reality of their conception. Considered comparatively, paternity is a 'freer symbol' (Barnes, 1973: p. 71), owing less to biology than maternity and thus can show greater cultural variation, as, indeed it does. Thus the mother–child tie is the only kinship relationship that can be said to be in any sense, 'natural' (see Fortes, 1958 for an earlier statement of this point of view) and concepts of 'the family' and of 'kinship' partake of both elements.

Marilyn Strathern subsequently argued that in providing new methods of 'assisted parenthood' science will undermine the idea of the family as being fundamentally 'natural' (Strathern, 1992) and destroy the concept of relatedness seen as extending outward from the nuclear family. In particular, the possibility of 'surrogate motherhood' whereby a woman gives birth to a child which is the genetic offspring of another woman to whom it will be given will undermine the notion of motherhood as based in biology (Edwards and Strathern, 2000). An earlier discussion of the same topic (Wolfram, 1987) came to rather different conclusions. Looking at historical and literary evidence as well as more recent legal debates, she concluded, "change can often be accommodated in existing categories" (Wolfram, 1987: p. 214), and her discussion takes into account the role of law as equally involved. One might also add that children had been brought up by people who were not their genetic parents for centuries before the law on adoption in 1928 made it legal and often enough it was not known that the women they called mother were not biological relations.

The idea of 'the family' has always been a poor fit with social reality. Mothers have adopted the illegitimate children of their daughters as their own or kinswomen have taken in children whose mothers died or were unable to manage large numbers of children. Adultery and extramarital conception resulted in legitimate paternity not being matched by the genetic constitution of children. Early deaths made step-relations common. Yet the concept of 'the family' survived these diversions by reality. Perceptions may have changed among patients of the clinics offering reproductive technologies where research has largely been done, but it seems just as likely that in the wider society, the understandings of human reproduction which they accept as science-based views of nature, but which anthropologists argue are moderated by culture, still remain relatively unchanged.

At the same time, a rediscovery of the significance of relationships based on locality and residence emphasized other means of creating relationships than those traditionally considered. The collection of essays edited by Carsten and Hugh-Jones (1995) was published at a time when interest in the family and kinship had declined and helped revive interest in domestic life. Demonstration of the many different ways in which membership of a domestic group can be established and the household related to its neighbors seemed to make the study of kinship alone too narrow and partial.

As a result of these separate approaches to the study of the family a focus on the folk concepts of relatedness was suggested (Carsten, 1995: p. 225; Carsten, 2000). Close ties may be based on shared residence, shared economic ties, nurturing or sharing food, gifts, and indebtedness or other forms of interaction. Interaction between meaning and action is thus reestablished as significant. This is not to say that kinship disappears however. Lambert argues that in North India, "consanguineal kinship constitutes the ideal model of relatedness and optative forms of relation invoke this model rather than being entirely identified with it" (Lambert, 2000: p. 85). Folk theories of relatedness constitute important sources of data for the anthropologists but they are not explanations that can contribute to anthropological theory, which must have general relevance if it is to be valid as a discipline.

While it may be generally accepted that the concept of the family is unhelpful, the term continues to be used. Thus, Segalen and Zonabend, having noted that under the influence of anthropology the 'overfluid' concept of the family has been replaced by analytical concepts such as household and kinship (1987: p. 112), then use 'family' unreflectively and do not discuss why or in what sense. Goody, the doyen of anthropological studies of domestic kinship, continues to write of the family and assert the European view that it is based on biological necessity (Goody, 1983, 2000: p. 1). This seems accurate, descriptively, of the European ethnography but it does not facilitate comparative study.

A final strand in late twentieth century anthropological thinking on 'the family' has been contributed by research in Europe and on issues of public concern (La Fontaine, 1990; Simpson, 1998). Simpson argues for the study of divorce, not as a social problem or a symptom of family or social breakdown, but as the reorganization of interpersonal relationships in new forms of kindred that are just as much a part of present day culture as 'the family.' Simpson's rejection of 'the family' as a theoretical model is explicit, but his continued use of the term, even qualified as 'the unclear family,' implies, as does the usage of Segalen and Zonabend, a tacit attachment to certain connotations of it that have yet to be made explicit. The concept of the family may have been discarded from the repertoire of many anthropologists but it is too soon to regard its relegation to the scrap heap as final.

See also: Divorce, Sociology of; Environmental Stress and Health; Family Law in the United States; Family and Kinship, History of; Family and Psychiatric Illness: Family Therapy; Family as Institution; Fatherhood; Feminist Legal Theory; Households in Anthropology; Kinship Terminology; Kinship in Anthropology; Love and Intimacy, Psychology of; Marital Quality and Health; Marriage; Matrifocality; Parenting Attitudes and Beliefs across Cultures; Partnership Formation and Dissolution in Western Societies.

Bibliography

Barnes, J., 1973. Genitor: genitrix: nature: culture. In: Goody, J. (Ed.), The Character of Kinship. Cambridge University Press, Cambridge, UK.
Benedict, R., 1934. Patterns of Culture. Houghton Mifflin, New York.
Carsten, J., 1995. The substance of kinship and the heat of the hearth: feeding, personhood, and relatedness among Malays in Pulau Langkawi. American Ethnologist 22, 223–241.
Carsten, J., 1997. The Heat of the Hearth. Clarendon Press, Oxford, UK.
Carsten, J. (Ed.), 2000. Cultures of Relatedness. Cambridge University Press, Cambridge.
Carsten, J., Hugh-Jones, S., 1995. About the House: Levi-Strauss and Beyond. Cambridge University Press, Cambridge.
Davidoff, L., Hall, C., 1987. Family Fortunes. University of Chicago Press, Chicago.
Edwards, J., Strathern, M., 2000. Including our own. In: Carsten, J. (Ed.), Cultures of Relatedness. Cambridge University Press, Cambridge.
Fortes, M., 1949. Time and social structure: an Ashanti case-study. In: Fortes, M. (Ed.), Essays Presented to A.R. Radcliffe-Brown. Clarendon Press, Oxford, UK.
Fortes, M., 1958. Introduction. In: Goody, J. (Ed.), The Developmental Cycle in Domestic Groups. Cambridge University Press, Cambridge, UK.
Goody, J., 1956. A comparative approach to incest and adultery. British Journal of Sociology 7, 286–305.
Goody, J., 1976. Production and Reproduction: A Comparative Study of the Domestic Domain. Cambridge University Press, Cambridge, UK.
Goody, J., 1983. The Development of the Family and Marriage in Europe. Cambridge University Press, Cambridge, UK.
Goody, J., 2000. The European Family. Blackwel, Oxford.
Harris, O., 1981. Households as natural units. In: Young, K., Wolkowitz, C., McCullagh, R. (Eds.), Of Marriage and the Market. CSE Books, London.
Harris, C., 1983. The Family in Industrial Society. George Allen and Unwin, London.
Janssens, A., 1993. Family and Social Change. Cambridge University Press, Cambridge, UK.
La Fontaine, J.S., 1985. Anthropological perspectives on the family and social change. The Quarterly Journal of Social Affairs 1 (1), 29–56.
La Fontaine, J.S., 1990. Child Sexual Abuse. Polity Press, Cambridge, Oxford, UK.
Lambert, H., 2000. Sentiment and substance in north India. In: Carsten, J. (Ed.), Cultures of Relatedness. Cambridge University Press, Cambridge.
Levi-Strauss, C., 1958. Les Structures Elémentaires de la Parenté (The Elementary Structures of Kinship, tr. 1963).
Malinowski, B., 1913. The Family among the Australian Aborigines. University of London Press, London.
Malinowski, B., 1926. Crime and Custom in Savage Society. Routledge and Kegan Paul, London.

Malinowski, B., 1940. A sociological study of the rationale of the prohibited degrees of marriage. appendix 3. In: Kindred and Affinity as Impediments to Marriage. Church of England Commission, London.
Murdock, G.P., 1949. Social Structure. Macmillan, New York.
Reining, P. (Ed.), Kinship Studies in the Morgan Centennial Year. Anthropological Society of Washington, Washington, DC.
Robertson, A., 1991. Beyond the Family. Polity Press, Cambridge, UK.
Sahlins, M., 1976. The Use and Abuse of Biology. University of Michigan Press, Ann Arbor, MI.
Schneider, D.M., 1965. Kinship and biology. In: Coale, A.J., Levy, M.J., Fallers, L.A., Schneider, M.A., Tomkins, S.S. (Eds.), Aspects of the Analysis of Family Structure. Princeton University Press, Princeton, NJ.
Schneider, D., 1968. American Kinship: A Cultural Account. Englewood Cliffs, New Jersey.
Schneider, D., 1972. What is kinship all about? In: Reining, P. (Ed.), Kinship Studies in the Morgan Centennial Year. Anthropological Society of Washington, Washington, DC.
Schneider, D.M., 1984. A Critique of the Study of Kinship. University of Michigan Press, Ann Arbor, MI.
Segalen, M., Zonabend, F., 1987. Social anthropology and the ethnology of France: the field of kinship and the family (J. Benthall, trans.) ASA Monographs 25. In: Jackson, A. (Ed.), Anthropology at Home. Tavistock, London.
Simpson, B., 1998. Changing Families. Berg, Oxford, UK.
Smith, R.T., 1973. The matrifocal family. In: Goody, J. (Ed.), The Character of Kinship. Cambridge University Press, Cambridge, UK.
Stenning, D., 1958. Household viability among the pastoral Fulani. In: Goody, J. (Ed.), The Developmental Cycle in Domestic Groups. Cambridge University Press, Cambridge, UK.
Strathern, M., 1992. After Nature: English Kinship in the Late Twentieth Century. Cambridge University Press, Cambridge.
Westermarck, E., 1891. The History of Human Marriage. MacMillan, London.
Wolfram, S., 1987. In-laws and Outlaws: Kinship and Marriage in England. Croom Helm, Beckenham, Kent.
Yanagisako, S.J., 1979. Family and household: the analysis of domestic groups. Annual Review of Anthropology 8, 161–205.

Family Theory: Competing Perspectives in Social Demography

Susan L Brown, Bowling Green State University, Bowling Green, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by V.K. Oppenheimer, volume 8, pp. 5367–5373, © 2001, Elsevier Ltd.

Abstract

Demographic research on families has relied on two key theoretical approaches: sex-role specialization in marriage and the economic significance of an independent nuclear family for shaping family-related behaviors. In this article, I address how these two approaches are linked to economic behavior and consider their potential ramifications for shifts in marriages and families.

This article examines two competing theoretical perspectives on recent changes in marriage and family behavior in industrial societies, their determinants, and consequences. Each focuses on the role of one of two fundamental pillars underlying most demographic research on the family: (1) the implications for family behavior of an independent nuclear family system and (2) the importance of sex-role specialization in marriage. The specialization perspective argues that the desirability of marriage is based on sex-role specialization with women focusing on domestic production (and reproduction) and men on market work. Hence it concludes that the rapid postwar rise in women's employment made them increasingly independent of marriage, leading to a retreat from marriage as indicated by sharply declining marriage and fertility rates, and rising marital instability as well as nonmarital cohabitation and unwed childbearing.

The specialization position has been subject to several major criticisms. Some of the most important are as follows: contrary to the theory's argument, empirical research shows that women in a better labor-market position are usually *more* rather than *less* likely to marry; the theory is one of *non*marriage while the phenomenon observed is primarily *delayed* marriage; and sex-role specialization in an independent nuclear family system tends to be a high-risk strategy and historically often required the use of the family's adolescent and young adult children to achieve economic stability, a role that now is taken over by wives. The independent nuclear family perspective has focused on how the changing economic position of young people has affected the ability of couples to be economically self-supporting. It has argued that the recent well-documented rise in economic inequality has increased the difficulty of assortative mating and of establishing an independent household. Studies indicate that men's employment stability and earnings have a strong positive effect on their marriage chances; moreover, the length of time it takes men to achieve career 'maturity' varies substantially among and within socioeconomic groups, creating significant differences in the age at marriage. A growing body of research demonstrates that these processes largely account for marriage trends in recent decades. Rising economic inequality in society coincides with the bifurcation of marriage and family patterns by education strata.

Two competing perspectives on the effect of postwar changes in the economic behavior of men and women on the family have characterized recent demographic research. Each primarily focuses on one of two fundamental theoretical pillars which underlay most demographic research on families in Western societies: (1) the economic implications for family behavior of an independent nuclear family system and (2) the importance of sex-role specialization in marriage. This article will undertake an examination of each of these perspectives in terms of these theoretical underpinnings, how they are affected by changing economic behavior, and the implications of this for changing patterns of marriage and family.

Overview

A number of major changes in marriage and family behavior have occurred in the West since the early 1970s. Age at marriage and marital instability has risen considerably while fertility has declined substantially (Blossfeld, 1995; Mason and Jensen, 1995; Oppenheimer, 1994). The magnitude of these changes has fostered considerable debate about their determinants and consequences and what these imply for the future of the family. However, in most cases, the debate is not so much about major theoretical differences on the nature of the family but differences in which aspects of two long and commonly accepted underpinnings should be emphasized the most. One is that marriage as an institution, as well as marriages as functioning social forms, is based on sex-role specialization. Although there are variations in how this specialization argument is conceptualized, the general idea is that the husband's role is largely economic while the wife's lies in the care of the home and in the reproduction and socialization of the children.

The second theoretical underpinning to most discussions is that an independent nuclear family system is a long-standing feature of populations of Western European origin. Such a family system is characterized by husband–wife families which are economically independent and live separately in their own household, although a single-aged parent might also have sometimes been included in the past. This is opposed to an extended family system which is usually characterized by joint family households consisting of several nuclear families. In general, historical demographic research has indicated that marriage and fertility behavior have often been quite different

in populations with independent nuclear as opposed to joint family systems. In the former case, marriage is contingent on the couple's ability to be self-supporting and able to set up a separate household; in the latter, the economic ability of the household as a whole to absorb new members has tended to be more important. Since young people's economic position is often precarious, the requirement of economic independence has historically led to somewhat later marriages and also to periodic fluctuations in marriage timing, depending on variations in economic circumstances. This has often resulted in substantial short- and long-term effects on fertility, especially in periods before the deliberate control of marital fertility became common.

This article focuses on two opposing positions regarding the effect of shifts in the economic behavior of men and women. One emphasizes sex-role specialization, arguing that the rising employment of women has been the major factor leading to the observed changes in marriage and family behavior because it has reduced specialization and hence permitted women to become economically independent of marriage, thereby reducing the gain to marriage. Assuming women's employment is unlikely to decrease, the result will be a major and irreversible transformation of the family in modern industrialized societies.

The opposing position is that the family has not been transformed as much as the women's independence argument maintains nor has women's rising employment had a negative effect on marriage formation, although it undoubtedly mitigates against high fertility. On the other hand, there is some evidence that deterioration in young *men's* economic position has had an important effect on marriage and family behavior because it has seriously affected the ability of young couples to set up an independent household. If so, then the assumed irreversibility of present trends is open to question, although it is extremely unlikely that a return to postwar patterns of the 1950–70s will occur.

These two alternatives do not exhaust the many possible reasons that have been proposed for changing family behavior, although they cover the major debates related to the effect of changes in economic behavior. Nevertheless, it is often the case that different explanations may be implicitly, if not explicitly, dependent on one or both of these positions. For example, it has been argued that modernization and the rise of individualism is the major force behind the recent demographic changes (*see* Second Demographic Transition). However, neither men nor women can be autonomous individualists if they are still economically dependent on traditional marital and kinship ties.

Independent Nuclear Family System

Given the assumption of sex-role specialization, demographic research on marriage and family behavior in independent nuclear family systems in preindustrial and early industrial populations has tended to focus primarily on the economic position of men, although there has long been a general recognition that women could help facilitate the formation of a marriage via earnings before marriage. In farming communities, marriage was often considerably delayed as the heirs had to wait for the father to die or retire before they inherited the property and with it the means of marrying. Even where leasing rather than owning land was the rule, as in England, there were often extensive delays until a man could save enough to lease a farm and set up a household. The effect of this pattern on marriage timing – and hence on fertility in historically non-contracepting populations – was often substantial with high proportions not marrying until their thirties and relatively substantial proportions never marrying (*see* Arensberg and Kimball, 1968; Goldstone, 1986).

With the growth of wage labor in the cities and on the farms as industrialization progressed, young men could much more easily become economically independent at a relatively young age. This promoted a reduction in the age at marriage, although offsetting factors were the need to save up to establish an independent household and prior obligations to contribute to the parental household, especially if one or both parents had died or were incapacitated in some way. In general, economic development also increased the sensitivity of young men's economic position to business cycles and structural changes in the economy. Recessions or depressions reduced marriage propensities, especially among the young, and recoveries led to sharp rises in rates. Moreover, the structural changes in the economy associated with economic development changed the skill levels and types of skills demanded of workers. This has placed an increasing premium on education and technical training in modern industrial societies which has, in turn, postponed the full-time entry of young people into adult-level employment and, with it, entry into marriage and the achievement of the degree of economic stability usually considered necessary for starting to raise a family.

Role of Relative Cohort Size

Postwar analyses of the influence of men's economic position on marriage formation and fertility was initially dominated by the work of Richard Easterlin who hypothesized that fluctuations in relative cohort size had a major impact on young men's labor market position, thereby affecting marriage timing and marital fertility (*see* Easterlin, 1987). Thus, Easterlin argued that the postwar baby boom, such as the one that the US experienced in the early postwar period, was largely caused by the favorable economic position of the small depression cohorts entering the labor market at that time. This encouraged them to marry earlier and have more children than previous cohorts. However, the relatively large cohorts they produced crowded the labor market, negatively affecting their lifetime earnings. This led to postponed marriage and childbearing, lower completed fertility, and greater marital instability. In addition, more wives would be working to compensate for their husband's poorer economic position. The result was smaller birth cohorts in the next generation – the baby bust cohorts – who presumably would be in a much more favorable relative labor-market position. According to the theory, this should lead once again to earlier marriage and childbearing, somewhat higher fertility, increasing marital stability, and the slowing of the growth in married women's employment. In short, the theory posited a self-perpetuating cycle.

The relative cohort size argument ran into empirical difficulties as research in the 1980s showed that postwar marriage and fertility fluctuations were largely *period* rather than *cohort* driven, i.e., fluctuations in demographic rates occurred simultaneously for numerous birth cohorts, regardless of cohort size. Moreover, the theory essentially accepted the premise that marriage is based on sex-role specialization. Women's labor-market behavior was considered to vary primarily in response to men's earnings position rather than in response to women's own employment opportunities, independent career goals, or family cycle variations in economic needs. However, women's employment continued to rise rapidly throughout the postwar period, regardless of variations in cohort size. As a result, the focus of research interest began to shift to the familial consequences of the transformation of women's economic roles.

Sex-Role Specialization

At the same time that relative cohort size argument (and hence male-oriented explanations of changes in marriage and family behavior) began to lose favor, the continuing rapid rise of married women's employment has seemed to provide a ready explanation for the ongoing changes in marriage and family behavior. Hence, the theoretical emphasis shifted from a concern with the ability to set up and maintain an economically independent household to the critical importance of sex-role specialization for family formation and cohesion.

The centrality of differentiated sex roles as an explanation of changes in marriage and family behavior reflects a long-standing and influential tradition in the social science literature on the importance of sex-role specialization for a stable marriage system. In the early postwar period, this idea was emphasized by Talcott Parsons (1949) who argued that sex-role segregation was a functional necessity for marital stability and even for the viability of society itself. More recently, this same theme has been elaborated somewhat differently in the economic theory of marriage by Gary Becker (*see* Family Theory: Economics of Marriage and Divorce). In an argument very similar to one Durkheim made over a century ago, Becker maintains that the major gain to marriage lies in the mutual dependence of spouses, each of whom specializes in certain functions – the woman in domestic production (and reproduction), the man in market work. Marriage then involves trading the fruits of these different skills. However, in response to economic growth and the rising wages it produces, women's market work also rises. The result is that women become less specialized and, as their earnings approach those of potential mates, more economically independent, leading, in turn, to a decline in the desirability of marrying or of staying married as well as the desirability of high fertility.

Many scholars do not necessarily agree with Becker's argument in its entirety. Nevertheless, an economic independence argument of one sort or another remains appealing among sociologists as well as economists and is currently one of the major contenders in trying to explain recent marriage and family trends in the United States (Cherlin, 1992; Oppenheimer, 1997). One reason is that the model seems to have tremendous face validity. Not only does the notion of specialized sex roles fit in with long-held views of how the traditional family historically functioned, but the rapid changes in marital behavior appeared to have followed very closely upon the rapid rise in married women's employment, especially that of young women. Nevertheless, the independence hypothesis has also run into a number of serious theoretical and empirical problems.

Mismatches

One problem the theory has encountered is that there are number of mismatches between the phenomena it claims to explain and the behavioral changes that are actually occurring. First, the independence hypothesis is essentially a theory of *non*marriage for it argues that if, by their own endeavors, women can achieve approximately the same income as a prospective spouse there is not much to gain by marrying and specializing in home production. However, so far, while there has been a considerable increase in *delayed* marriage, as yet there is little evidence of a substantial rise in *non*marriage (Oppenheimer, 1994; Blossfeld, 1995). In fact, Americans are unique for their proclivity to marry and remarry. According to Cherlin (2009), Americans are on a marriage-go-round. Although the age at first marriage in the US is at a historic high point, the vast majority eventually marry. Hence one reason for the apparent support for the hypothesis is due to a serious confounding of *delayed* marriage with *nonmarriage* in the statistics. Moreover, to date, there has also been little effort to articulate how the independence argument could lead to *delayed* instead of *non*marriage.

A major difficulty in assessing the outcome of recent marriage trends lies in the recency of the trends and the necessity of waiting 20 years or more before one can accurately determine whether there actually will be a significant rise in nonmarriage. Efforts to estimate the future proportions never marrying are often misleading because they typically use the current behavior of older people as a guide. However, during a period of rapid timing shifts, the *current* behavior of somewhat older people, who are the relatively sparse and atypical remains of much earlier marrying cohorts, is often a particularly poor guide to the future behavior of the much larger and less select group of younger people who have been delaying marriage in more recent years.

A second kind of mismatch lies in using the family behavior of the 1950s and 1960s in the US (and somewhat later in Europe) as typifying the traditional family patterns when sex-role specialization prevailed; compared to this, the large subsequent changes in marriage and family behavior certainly appear to represent a major retreat from marriage and the family. However, this was the baby boom era when fertility was unusually high compared to the previous 50 years; age at marriage was unusually low – far lower than at the turn of the century – and, in the US at least, even marital disruption was somewhat less likely than in previous years. In short, this benchmark was highly atypical historically and unrepresentative of the supposedly 'traditional' family (Coontz, 2004). Hence using the 1950s, termed the golden age of the family, as a benchmark, has led to a considerable exaggeration of the extensiveness of the subsequent break with historically

traditional patterns and of the historical origins of many recent trends.

There is still another type of mismatch characteristic of a theory which posits that the stability of the family is founded on women specializing in home production. The proposed model seems strangely out of tune with the needs and characteristics of modern societies. Much of the specialized home production of women in the past was devoted to bearing and rearing children who never survived to adulthood. For women to be equally occupied in contemporary low-mortality societies would result in large families. However, even moderate family sizes in a low-mortality society lead to rapid population growth. Moreover, couples do not just want to produce children per se, they want to produce children like themselves – that is, they are interested in social, not just biological reproduction. But the cost of social reproduction is high in a society where increasingly substantial and lengthy investments in human capital for each child are required. Hence, high fertility is usually not a realistic option for most people in modern industrial societies. With low fertility, however, the female specialist would not be doing anything highly productive for much of her life. In short, specialization in home production appears to be a rather anachronistic institutional role for women in industrial societies. Moreover, recent empirical work demonstrates that female labor force participation is *positively* associated with fertility in industrialized nations (Brewster and Rindfuss, 2000). Women appear to prefer to combine work and family.

Empirical Status of the *Independence* Hypothesis

The women's independence hypothesis has not been well supported by empirical evidence. At any given time, empirical analyses show that indicators of women's economic independence have not generally had a negative effect on marriage formation. For example, under an independence argument, better-educated women should be more economically independent of marriage; however, microlevel regression analyses show that, once school enrollment is taken into account, they have a *higher* rather than a *lower* propensity to marry. Although more educated women tend to delay marriage entry, they are ultimately most likely to ever marry. Indeed, women's education, earnings, and employment are all positively related to getting and staying married (Lichter et al., 1992).

The role of women's economic independence in divorce appears to be conditional on marital quality. Although women's employment and earnings are often positively associated with divorce, this relationship is actually more complex because it depends on the quality of the marriage. Among happily married couples, women's employment is not associated with divorce. Among unhappily married couples, women's employment is positively related to divorce, suggesting that economic independence allows women a pathway to exit an unsatisfying marriage (Schoen et al., 2002). This pattern has been established in other studies, too, including one showing that the effect of economic independence is a function of gender role ideology (Sayer and Bianchi, 2000). Marital quality was more closely linked to divorce than was women's economic independence, leading to the same substantive conclusion: women's economic independence allows women to leave bad marriages. Independence per se is not a driving factor in the likelihood of divorce. Still other work suggests that wives increase their market work in anticipation of divorce (Johnson and Skinner, 1986).

Coping with Risk and Change

Although rarely considered in explications of the theory, historical research on independent nuclear families indicates that extreme sex-role specialization in marriage is essentially a high-risk and inflexible family strategy unless accompanied by supplementary support mechanisms (Oppenheimer, 1994). Even with such supports, specialization often entailed considerable individual and social costs. An inherent problem is that the temporary or permanent loss of one specialist in a family can mean that functions vital to the well-being of the complementary specialist and children are not being performed. Husbands/fathers can die or become ill or disabled; they can lose their jobs and have difficulty finding another one; they could desert the family for a variety of reasons or become an alcoholic; and so on. The result is that the family is left without its major source of income. Except for employment-related shifts, there are similar problems involving the wife–mother specialist. In that case, there could be no one to take care of the children or the home. Specialization may be a feasible strategy in a large extended family household where no particular individual is indispensable because of the redundancy in personnel that can characterize such a system. However, for independent nuclear families and their individual members, specialization entails considerable risks.

Extreme sex-role specialization is also not a very flexible way of dealing with the varying needs of nuclear families over their developmental cycle. Since individuals' consumption needs and productive capabilities vary markedly by age, a basic feature of nuclear families is that the ratio of consumers to producers, and hence the family's level of living, can vary substantially over the family's developmental cycle. Hence, specialization involves a potentially serious inflexibility in dealing with changes in both a family's internal composition as well as the stresses posed by its environment.

The large literature on family history in Western societies indicates that a variety of strategies were developed to maintain economic stability over the family's developmental cycle and in the event of the temporary or permanent loss of specializing parents (Goldin, 1981; Haines, 1979; Rowntree, 1922; see Oppenheimer, 1994 for a review). In the nineteenth and early twentieth centuries these strategies usually involved utilizing the productive labor of children, daughters, as well as sons. The evidence also indicates that utilizing the labor of one's children could exact substantial costs. A sufficient number of children old enough to make an economic difference was generally not available until the middle or later stages of the family cycle. Families who temporarily or permanently lost the contribution of the father early in the family cycle were not greatly helped by such a strategy. If the mother was lost when the children were young, the family might break up with children being parceled out among relatives or even going to orphanages. In general, this economic reliance on one's offspring often led to a pattern of 'life cycle poverty' where periods of poverty and comparative plenty alternated over the life cycle of workers. Another

well-known disadvantage to the extensive employment of children to supplement their family's income was that it tended to discourage schooling and hence had a negative effect on children's adult socioeconomic status.

Even aside from the drawbacks of using children's work, the economic advantages of this strategy were eventually bound to decline during industrialization as the adult labor force became more skilled and hence the relative economic contribution unskilled children could make declined. This suggests that other equilibrating mechanisms were likely to develop. One mechanism may be governmental social support systems. At the family level, however, from a historical perspective, the rise in married women's employment might be viewed as a functional substitute for the work of their children, facilitating the more extensive schooling of the next generation and thereby fostering upward intergenerational social mobility. In addition, wives' employment is not limited to the later stages of the family's developmental cycle and wives, being adults, and usually having educational attainments roughly similar to their husbands, can command a much higher wage than unskilled children could. Moreover, recent research indicates that wives' earnings are currently playing an important role in offsetting the less favorable earning position of their husbands (Levy, 1998). It seems likely, then, that wives' employment could be a source of cohesion in marriages; in difficult times, a helpmate may be of much greater value than a complete economic dependant.

Even aside from wives' employment during episodic periods of an increased need for additional income, wives who work regularly may also increase the interdependence of spouses. Two-earner families where the couple pool their resources are families whose level of living is based on their *combined* income. Even when the income contribution of each spouse is equal, given economies of scale, neither partner could live as well on his/her own or save and invest as much. Hence, instead of the asymmetric dependence of the wife which is characteristic of complete specialization, there is a symmetric dependence where each spouse depends on the other to maintain living standards. What this may also indicate is that the combined income of the two-earner family has come to form the social standard rather than the husband's income alone. To the extent this is the case, it becomes increasingly difficult for single earners and married couples with a more traditional division of labor to achieve the same level of living as the two-earner family. Hence, the mutual dependence of the two-earner family may not only contribute to their own gain to marriage but may also reduce the relative gain to being single and to marriages characterized by a specialized division of labor.

In sum, it is unclear that specialization is the only, or even the optimum, method of creating cohesion in marital relationships in modern industrial societies. The specialization model itself seems highly problematic and women's rising employment may actually be promoting a more collaborative model of marriage.

Rising Economic Inequality

Given the empirical and theoretical limitations of the specialization model of marriage, research attention has been returning once again to the role of economic difficulties in setting up and maintaining an independent nuclear family. While the effect of relative cohort size earlier dominated research in this area, the more recent focus is on such issues as changing economic structure, chronically high unemployment rates (particularly among young adults), and rising earnings inequality (again especially among the young but also among the moderately-to-less educated as well). These analyses have also returned to the question of men's changing economic position which had been neglected for quite some time.

Virtually all recent research has shown that men who are more likely to be working full-time/year-round, had higher current earnings, and/or greater future career prospects were more likely to marry in any given short time interval (Clarkberg, 1999; Huinink, 1995; Oppenheimer, 1997; Oppenheimer et al., 1997). Some research in this area has also started to look at the role of economic factors in the formation of cohabiting unions and this indicates that here too, some level of economic stability is important, especially for the transition to marriage among cohabitors (Brown, 2000; Clarkberg, 1999).

Oppenheimer (1988) has theoretically and empirically elaborated on these issues in a variety of ways. First, she has conceptualized the transition to work as a career entry *process* rather than a *step*, which was more commonly the approach in previous life-course analyses. This change in approach emphasizes that embarking on a relatively stable occupational career takes time, during which a young man is in a relatively 'immature' career state. One aspect of career immaturity directly impeding the formation of an independent household is relatively low income; however, another is uncertainties regarding what kind of stable employment may eventually be achieved, when this may occur, what lifestyle the individual may ultimately expect or desire, and so forth. Such uncertainties can impede mate selection and affect the individual's own desirability as a prospective marriage partner. Hence, career immaturity is not only likely to foster delayed marriage for economic reasons but for social-psychological ones as well. Second, viewing the transition to work as a *process* expands the emphasis beyond the effect of career immaturity at any single point in time to the length of time it takes to achieve some career stability and the cumulative impact this will have on marriage timing. The delaying effect of only a modest effect of career instability in any given year, for example, may nevertheless have a large cumulative impact if the instability persists over many years. Moreover, this approach is well suited for examining how the length and difficulty of career development could vary substantially among different groups and over time as well, thereby facilitating analyses of socioeconomic differentials in marriage timing and trends in marriage behavior.

Using 12 years of an American longitudinal survey of young people, Oppenheimer et al. (1997) found that not only did a variety of indicators of career immaturity have a substantial effect in any given year on marriage formation, but that the cumulative impact was substantial when the career development process was followed over time. Moreover, there were large race–schooling differences in the length and difficulty of the career entry process which, in turn, produced major differences in marriage timing. Because the analysis was limited to the young adult life course of young men born between 1957 and 1964, how much of the rise in delayed marriage over

the past 25 years can be accounted for by career development problems has yet to be directly examined. However, there has been a substantial and well-documented rise in inequality in the US since the 1970s, and this has had a particularly severe effect on moderately-to-less educated younger workers, suggesting that career entry difficulties may be playing an important role in these trends. Indeed, there appear to be diverging destinies for young men and women as marriage increasingly becomes the province of the most highly educated (Cherlin, 2009; McLanahan, 2004). Marriage is now viewed by many as a luxury good or capstone experience and the bar for marriage has risen among all groups, effectively preventing many from achieving this milestone. To the extent that couples view financial security and good relationship quality as prerequisites for marriage (Gibson-Davis, 2009), these barriers can be insurmountable for a growing share of society as widening inequality hollows out the middle class. Many of the benefits of marriage, including co-residence and fertility, can be achieved outside of marriage through other pathways such as cohabitation. These developments give new urgency to a question that has vexed scholars in recent decades: to what extent is delayed marriage actually foregone marriage? Future research must address how and why the timing of marriage is changing.

See also: Family Theory: Economics of Marriage and Divorce; Marriage and the Dual-Career Family; Reproductive Rights in Affluent Nations.

Bibliography

Arensberg, C.M., Kimball, S.T., 1968. Family and Community in Ireland. Harvard University Press, Cambridge, MA.

Blossfeld, H.-P. (Ed.), 1995. The New Role of Women: Family Formation in Modern Societies. Westview Press, Boulder, CO.

Brewster, K.L., Rindfuss, R., 2000. Fertility and women's employment in industrialized nations. Annual Review of Sociology 26, 271–296.

Brown, S.L., 2000. Union transitions among cohabitors: the significance of relationship assessments and expectations. Journal of Marriage and Family 62 (3), 833–846.

Cherlin, A.J., 2009. The Marriage-go-round: The State of Marriage and the Family in America Today. Knopf, New York.

Cherlin, A.J., 1992. Marriage, Divorce, and Remarriage. Harvard University Press, Cambridge, MA.

Clarkberg, M., 1999. The price of partnering: the role of economic well-being in young adults' first union experiences. Social Forces 77, 945–968.

Coontz, S., 2004. The world historical transformation of marriage. Journal of Marriage and Family 66 (4), 974–979.

Easterlin, R.A., 1987. Birth and Fortune. University of Chicago Press, Chicago.

Gibson-Davis, C.M., 2009. Money, marriage, and children: testing the financial expectations and family formation theory. Journal of Marriage and Family 71, 146–160.

Goldin, C., 1981. Family strategies and the family economy in the late nineteenth century: the role of secondary workers. In: Hershberg, T. (Ed.), Work, Space, Family and Group Experience in the Nineteenth Century. Oxford University Press, New York, pp. 277–310.

Goldstone, J.A., 1986. The demographic revolution in England: a re-examination. Population Studies–Journal of Demography 49, 5–33.

Haines, M., 1979. Industrial work and the family life cycle, 1889–1890. Research in Economic History 4, 449–495.

Huinink, J., 1995. Education, work, and family patterns of men: the case of West German. In: Blossfeld, H.-P. (Ed.), The New Role of Women: Family Formation in Modern Societies. Westview Press, Boulder, CO, pp. 247–262.

Johnson, W.R., Skinner, J., 1986. Labor supply and marital separation. The American Economic Review, 455–469.

Levy, F., 1998. The New Dollars and Dreams: American Incomes and Economic Change. Russell Sage, New York.

Lichter, D.T., McLaughlin, D.K., Kephart, G., Landry, D.J., 1992. Race and the retreat from marriage: a shortage of marriageable men? American Sociological Review 57, 781–799.

Mason, K.O., Jensen, A.-M., 1995. Gender and Family Change in Industrialized Countries. Clarendon Press, Oxford, UK.

McLanahan, S., 2004. Diverging destinies: how children are faring under the second demographic transition. Demography 41, 607–627.

Oppenheimer, V.K., 1988. A theory of marriage timing. American Journal of Sociology 94, 563–591.

Oppenheimer, V.K., 1994. Women's rising employment and the future of the family in industrial societies. Population and Development Review 20, 293–342.

Oppenheimer, V.K., 1997. Women's employment and the gain to marriage: the specialization and trading model. Annual Review of Sociology 23, 431–453.

Oppenheimer, V.K., Kalmijn, M., Lim, N., 1997. Men's career development and marriage timing during a period of rising inequality. Demography 34, 311–330.

Parsons, T., 1949. The social structure of the family. In: Anshen, R. (Ed.), The Family: Its Function and Destiny. Harper, New York, pp. 173–201.

Rowntree, B.S., 1922. Poverty: A Study of Town Life. Longmans, Green, London.

Sayer, L.C., Bianchi, S.M., 2000. Women's economic independence and the probability of divorce: a review and reexamination. Journal of Family Issues 21, 906–943.

Schoen, R., Astone, N.M., Kim, Y.J., Rothert, K., Standish, N.J., 2002. Women's employment, marital happiness, and divorce. Social Forces 81 (2), 643–662.

Family Theory: Economics of Childbearing

John F Ermisch, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Economic theories of childbearing relate decisions about family size and the timing of births over people's lives to elements of the constraints on choice that people face, particularly prices and incomes. This article presents the basic concepts used in these theories, discusses their main lines of argument, and summarizes some of their empirical implications.

Economic theories of childbearing relate decisions about family size and the timing of births over people's lives to elements of the constraints on choice that people face, particularly prices and incomes. This article presents the basic concepts and approaches used in these theories, discusses their main lines of argument, and summarizes some of their empirical implications. It first examines 'static' theories of completed family size and then discusses dynamic theories of the timing of childbearing. While mainly concerned with theories of fertility in stable unions, it concludes with a discussion of nonmarital fertility.

'Child Quality' and Fertility

Economic theories of fertility have been around for over 200 years. The ones advanced by Thomas Malthus are the best known. In a few words, Malthus' view was that fertility was determined by the age at marriage and frequency of coition during marriage. He argued that an increase in people's income would encourage them to marry earlier and have sexual intercourse more often. Modern economic theories of fertility date from a paper by Becker (1960), which generalized and developed the Malthusian theory. Until Section Nonmarital Fertility, it is assumed that fertility decisions are made by a couple in a stable union.

Becker's point of departure is that children are usually a source of psychic satisfaction for parents. An important aspect of his seminal contribution was to point out that this satisfaction is likely to depend on the amount of resources that parents spend on children as well as the number of children that they have. The basic idea is that if parents voluntarily spend more on a child, it is because they obtain additional satisfaction from doing so. For instance, parents are likely to receive higher satisfaction if their children obtain a university degree than if they leave school as teenagers. For short, he calls children who have more resources devoted to them 'higher quality' children, hastening to add "that 'higher quality' does not mean morally better" (Becker, 1960: p. 211).

The characterization of 'child quality' in subsequent analyses has moved in the direction of identifying it with the lifetime well-being of the child ('quality of life'), which can be increased by investing more in the child's 'human capital' (i.e., his or her ability to earn income as an adult) or by the direct transfer of wealth to the child. That is, parents receive more satisfaction from having children who are better off throughout their lives, and they make monetary transfers and human capital investments to influence their child's lifetime standard of living.

Having incorporated the quality dimension of reproduction decisions, the well-developed model of consumer choice is applied to them. Parents are assumed to choose the best combination of the number of children, their quality, and the parents' own standard of living subject to parents' lifetime income and the prices that they face in markets. An increase in parents' income is expected to increase both the quantity and quality of children. Analogous to the tendency for people to increase the quality of most consumer durable goods proportionately much more than the quantity of such goods when their income increases, Becker argued that this is also likely to be the case for children. In economics, the 'income elasticity' of a good is defined as the percentage increase in the demand for that good in response to a 1% increase in income. Thus, using this terminology, Becker's contention is that the income elasticity of the number of children ('quantity') is small compared to the income elasticity of child quality. Put somewhat differently, an increase in parents' income may increase the amount spent on children substantially, but this would mainly take the form of higher quality rather than more children. The introduction of the child quality concept generalizes the Malthusian theory, which only allows responses to income along the quantity dimension.

Quantity–Quality Interaction

Becker and Lewis (1973) and Willis (1973) added a new twist to the quantity–quality model. They assumed that parents view child quantity and quality as substitutes and they treat all their children equally, in the sense that child quality is the same for each child. Expenditures on their children's human capital acquisition and monetary transfers to their children represent expenditure on child quality, which is the *product* of the number of children, quality per child and the full cost per unit of child quality (which is discussed later). This model can be generalized to incorporate costs of the number of children that do not depend on child quality and costs of quality that do not depend on the number of children. As an example of the former, a decline in the cost of averting births, say because of the introduction of the oral contraceptive pill, would increase the marginal cost of a birth without affecting the marginal cost of child quality.

The product in the budget constraint implies that the cost (or 'shadow price') of an additional child is proportional to the level of child quality, and the cost ('shadow price') of raising child quality is proportional to the number of children parents have. As a consequence, there is an important interaction between family size and child quality. Suppose, for example, that there is a *small* exogenous decrease in fertility, say because of lower contraception costs. This lowers the shadow price of child quality, which in turn raises child quality, which raises the shadow price of children, which lowers family size further, and so on. Similarly, a higher return to human capital increases investment per child. It raises the shadow price of children, which lowers family size and the price of child quality, thereby raising child quality further. Thus, increases in the returns to human capital investment associated with technical change may lead to simultaneous *large* reductions in fertility and increases in human capital investment in children. This is consistent with important stylized facts of economic development and links family economics with the study of economic growth and development (Rosenzweig, 1990). The interaction between child quality and fertility does not derive from assuming that child quality and quantity are *close* substitutes for one another. Indeed, just the opposite must be the case, because if both fertility and child quality are to be nonzero, then they cannot be close substitutes. Thus, family size can be highly responsive to changes in prices and incomes, even though children have no close substitutes.

The quantity–quality interaction may also produce a 'high fertility–low child investment trap,' in which low quality produces a low price of children, high fertility and a high shadow price of child quality. Higher parents' income increases fertility and the price of child quality, keeping child investment low. It may take some policy or technological development that alters the prices of quantity or quality independently, such as a family planning intervention or a large increase in the return to human capital investment, to 'spring the trap.' Once sprung, if the quality income elasticity exceeds the one for quantity, then the ratio of quality to the number of children rises with higher income, thereby increasing the shadow price of an additional child relative to the shadow price of child quality. The substitution effect induced by this increase may be sufficiently large to produce a decline in fertility when income increases, even though children are 'normal goods,' in the sense that parents want more of them when parental income increases.

Child Mortality and Fertility

The ultimate manifestation of low child quality is a child not surviving to adulthood. Scientific advance or a policy intervention, such as better water supply or public health, increases the probability of child survival, which has conflicting impacts on fertility. On the one hand, because a cost may be incurred by a birth regardless of whether or not the child survives, lower child mortality reduces the price of a surviving birth, thereby encouraging *higher* fertility. But if parents can influence the chances that their own children survive to become adults by spending more on each child, then it is *possible* that better chances of child survival reduce fertility, provided that exogenous factors affecting child survival substitute for parents' expenditure (Cigno, 1998). Such a relationship may help account for the 'demographic transition' – the change from a high fertility–high child mortality environment to a low fertility–low mortality one.

Another possible mechanism that may cause improvements in child survival to reduce fertility was proposed by Sah (1991). He argued that parents may wish to have more births when survival chances are low, because some may die before adulthood and they will not have enough time to replace them in their fecund period. This 'hoarding effect' on fertility dominates the fertility-discouraging effect from a higher price of surviving birth when the child mortality risk is high if parents are sufficiently 'risk averse,' or there is a 'target fertility level.' If either of these conditions holds, better survival chances for children tend to reduce fertility because their incentive to practice such 'hoarding' is less.

Fertility and Support in Old Age

If people do not have access to a capital market, an extended family transfer network including three generations at different stages of life could substitute for one (Cigno, 2000). In effect, it arranges 'loans' to its young members from its middle-aged ones and enforces repayment later when the young borrowers have become middle aged and the middle-aged lenders have become old. Because everyone needs the family transfer rules to survive in old age, the prospect of receiving no support from the family in old age deters any member from disobeying the rules, causing the rules to persist over generations until there is a change of circumstances outside the family. These rules also determine fertility. Larger transfers to the elderly relative to those to children increase the fertility required to sustain the family system.

The opening of a capital market with a sufficiently high interest rate offers the middle aged an alternative to this family transfer system. A threat of no support from the family in old age is no longer a deterrent, because they can make their own provision for old age through the market. In broad terms, this prediction is consistent with the observation that the growth of the financial sector, or the introduction of a state pension system, tends to coincide with a sharp decline in private transfers from the middle aged to their elderly parents and a fall in fertility. The fact that childbearing does not cease, suggests that the demand for children is not entirely derived from the need for transfers from them to finance consumption in old age.

The Cost of Children

Consumer theory tells us that, like any other price, a higher full cost of child quality per unit ought to reduce both fertility and child quality. Elucidation of the factors affecting this cost was another important contribution to the economics of fertility. In particular, it acknowledged the key role of parental time, especially the mother's, in the rearing of and investment in children. The foundations for this analysis were laid in a paper by Mincer (1963) and in Becker's (1965) theory of time allocation, and Willis (1973) built on them in his analysis.

The concept of 'household production' is important in understanding what influences the cost of child quality. Following Becker (1965), the parents' standard of living, the number of children and quality per child yield satisfaction to the parents, but they cannot be purchased directly in the market. Rather, the parents' standard of living and child quality are both produced within the household by combining parents' own time and purchased goods and services. It is the costs of these inputs that determine the cost of child quality relative to the cost of the parents' living standard. Parental time in the production of child quality is disproportionately the mother's time, and the rearing of children is assumed to be *time intensive* relative to other home production activities. That is, the proportion of the total cost of producing children represented by the value of the mother's time input to children is larger than the proportion of the total cost of producing the parents' standard of living represented by the value of her time input to that production. Thus, the unit cost of child quality relative to the cost of the parents' living standard is directly related to the mother's cost of time. If she has ever been in paid employment, her cost of time is the wage she could earn in employment (i.e., her forgone earnings). The higher her wage, the higher the cost of quality per child and of an additional child.

This analysis suggests, therefore, that men's and women's wages should have different effects on fertility. Higher wages mean higher income for the parents, encouraging them to have more children and to invest more in the human capital of each child or to make larger monetary transfers to them (i.e., higher quality). But higher wages also mean more income lost by those caring for children and investing in their human capital. Because child-rearing is mainly done by the mother, higher women's wages raise the opportunity cost of a child as well as increasing family income, while higher men's wages mainly affect childbearing through their effect on the couple's income. If the opportunity cost effect of women's wages is larger than their effect on fertility through higher family income, then higher women's wages reduce family size and child quality. It then follows that an increase in the ratio of women's to men's wages would reduce fertility.

In this model, the fraction of the woman's time supplied to the labor market is also a choice variable, which depends on the wage that the woman can earn and her partner's wage. There would tend to be a negative correlation between family size and the fraction of their life they spend in paid employment. This is a correlation between two choice variables. Variations in each partner's wages and in tastes drive both variables.

Purchased childcare can weaken the link between a woman's wage and the cost of an additional child. In Ermisch (1989), it is treated as an increasingly imperfect substitute for the mother's time in child-rearing as the amount of it increases. Mothers with high wages tend to purchase a much larger proportion of childcare time. For them, higher wages have less effect on the cost of children, making it more likely that a higher wage increases fertility by raising family income. Similarly, in countries with heavily subsidized childcare, mothers contribute much less of childcare themselves, making it more likely that women earning higher pay have larger families. At low to moderate levels of wages, a higher mother's wage reduces fertility, but its negative impact attenuates as her wage rises, or the price of childcare falls, because mothers purchase a larger proportion of childcare time under these conditions. The impact of the price of childcare on fertility displays a similar interaction, becoming more negative as the mother's wage rises.

Nevertheless, when examining changes in fertility over time, the cost of purchased childcare and women's wages tend to move together, because women's labor is such an important input to the provision of childcare services. Thus, over time women's pay relative to men's and fertility should be negatively related, because higher women's pay raises the cost of children.

Dynamic Fertility Models

The theories discussed so far are 'static' in nature, the relevant time unit being the parents' lifetime. In order to consider decisions about the timing of births, imperfect fertility control, divorce risk, or the consequences of unexpected outcomes like birth-control failure or child mortality, a dynamic model is needed. First, some features of dynamic life-cycle theories are discussed, and then a model of fertility dynamics across generations is introduced.

Life-Cycle Dynamics

Dynamic models incorporating a sequence of decisions and imperfect, costly fertility control are more realistic than the static theories discussed above. Their predictions often depend on what is assumed about parents' ability to borrow and save. One of two polar assumptions has been made in the literature: either parents can borrow and save as they choose ('perfect capital markets') or they cannot do so at all (i.e., they must consume all of their current income).

In one of the earliest dynamic fertility models (Heckman and Willis, 1976), a couple's fertility is controllable by the contraceptive strategies they adopt, but control is not perfect. Because it is assumed that couples start obtaining satisfaction from children as soon as they are born, they wish to have their children early. Couples may, however, find it best to adopt a 'precautionary' contraceptive strategy early in their partnership in order to reduce the risk of having more children than they would have chosen to have, if fertility control were perfect and free. Such a strategy would result in postponing the age at first birth. If people cannot borrow or save, and their incomes rise with age, they would also wish to practice contraception in order to try to postpone childbearing to a time when the marginal benefit of income for parental consumption is lower (when they can 'better afford' children). Subsequent spacing of births (rather than having children as quickly as possible) would also be a response to this tension between their desire to have children early and the economic incentive to have them later, when parents' income is higher. The more rapid the increase in income with age, the greater the incentive to postpone the start of childbearing and to space births. In other words, births are timed in order to 'smooth' parental consumption over time.

Another common (but not universal) feature of dynamic models is allowance for the possibility that a woman's current participation in paid employment improves her future

earning capacity, and that absence from employment may reduce her pay when she returns to employment. Thus, a mother's absence from paid employment because of a birth has two costs: a loss of current earnings and a loss of future earning potential. An early example of such a model is that of Moffitt (1984).

Even when there are perfect capital markets, couples may choose to postpone childbearing because these opportunity costs of childbearing are high when the woman is younger and on the steep part of her career-earnings profile. Women in jobs in which work experience has less effect on earnings have less incentive to postpone. In addition, if the care of young children is particularly intensive in the mother's time, then a birth produces a temporary increase in the value of her time in the home to a level above the value of her time in the labor market, causing her to leave paid employment. It also increases the cost of an additional child, encouraging her to postpone the next birth until the value of her time declines sufficiently through the aging of the child. An unexpected birth, because of contraceptive failure, would generate similar incentives for the mother to leave paid employment temporarily and postpone having another child. Thus, there is an economic incentive to space births.

The risk of divorce can also affect fertility, and its effect is also associated with women's opportunity costs of childbearing in a dynamic framework. If, for example, more participation in paid employment raises future wages, the risk of divorce can encourage more paid employment by the mother during marriage and, by raising the cost of child quality, lower expenditure on children and lower fertility. These 'defensive investments' in future earning potential are undertaken to increase welfare later, when outcomes are uncertain because of the possibility of divorce (Weiss, 1997).

The death of a child encourages its replacement, because the marginal benefit of an additional child is likely to increase substantially when a child dies. Replacement makes a woman's number of births increase with the experience of child mortality. Unless child mortality is expected when the mother is older, replacement is a better response to child mortality risk than 'hoarding' (discussed earlier) because the latter involves a larger deviation from what the parents would have done in the absence of child mortality. Wolpin (1984) was the first to incorporate uncertainty about child survival into a sequential, dynamic fertility model, which permits analysis of replacement in response to child deaths and of the effect of infant mortality risk on fertility.

Detailed surveys by Hotz et al. (1997) and Schultz (1997) discuss the empirical strategies for applying the theories discussed above and the results of many empirical studies.

Easterlin's Relative Income Hypothesis

In Easterlin's (1980) view, an important factor affecting a young couple's willingness to marry and have children is their outlook for supporting their material aspirations. If they expect their earning potential to be high relative to their aspirations, they will be more optimistic and will feel freer to marry and have children. If they expect that they will find it difficult to achieve the standard of living to which they aspire, they will hesitate to marry and have children. The ratio of the couple's potential earning power to their material aspirations is called their 'relative income.'

A person's material aspirations are formed through experience; for a young person, the primary relevant experience is the standard of living experienced in his/her parents' home. Young people's expectations concerning their earning potential are largely shaped by their experience of working.

A person's relative income may be related to the size of his/her generation relative to the size of previous generations. There are important differences between the characteristics of new labor force entrants and experienced workers. The former have not yet acquired a high level of skills and are engaged in a considerable amount of job search with consequent high-job turnover. Older workers are more skilled and occupy career jobs with relatively low-job turnover. As a consequence of these differences, the ability to substitute between young and older workers is limited. Thus, when young workers are scarce relative to older workers, their wages are higher relative to older workers and their chances of upward mobility improve. When they are in relative abundance, their relative wage declines and their promotion chances deteriorate. If there is a steady growth in labor demand, the earnings of young people (say, aged 15–29 years) relative to older people (say, aged 30–64 years) tends to be inversely related to the number of young people relative to older people ('relative generation size'), which suggests that their relative income also moves inversely with relative generation size.

The relative income theory of fertility would then lead us to expect that people from smaller birth cohorts would have higher fertility than people from larger cohorts. If couples feel constrained to choose family size within certain limits, say two to four children, then 'scarce' generations, who start childbearing earlier because of their favorable economic prospects relative to their material aspirations, would cut back later as they reach their upper limits, and people from 'surplus' generations would catch up later to attain their lower limits. Such behavior tends to make the fertility rates of women of different ages (different generations) move together. The period total fertility rate would then also move inversely with relative generation size, and it would fluctuate much more than completed family size.

Nonmarital Fertility

The rising proportion of births outside marriage in many countries has produced a shift away from the paradigm of a couple making fertility decisions as a single entity. It suggests that men and women should be treated as individual agents in fertility decisions. Higher divorce rates would reinforce such a treatment.

The process of finding a spouse is one in which information must be gathered by costly search. These market frictions affect who marries whom, the gains from each marriage and the distribution of gains between spouses. Marriage market frictions open the possibility of a marriage market equilibrium in which there is childbearing outside marriage (Ermisch, 2003: chapter 7). When a man and women meet, the man can choose to marry the woman, or not, if she will accept him. While a woman faces the same choice when she meets a man, she can

also choose to have a child by the man and then raise it without the father. Depending on the amount of support from the father or from the social welfare system, a woman's welfare when raising a child by herself may be greater than what she obtains when single and childless. But there are also costs in terms of marriage market prospects associated with raising a child alone. A single woman with a child may find it more difficult to contact potential husbands while looking after a child. Taking this perspective, a woman who contacts a man whom she does not wish to marry, or who will not marry her, would choose to have a child by the man if the short-run gain exceeds the long-term costs in terms of her marriage prospects. Those women who expect to obtain a significant increase in welfare when they marry suffer a greater long-term cost by having a child while single than women whose marriage prospects are such that they expect to gain little from marriage. Thus, women with poorer marriage prospects are more likely to have children outside marriage. If marriage prospects are correlated with education or potential income, then low-educated or low-income women would be the ones that have children outside marriage.

Similar predictions are produced in a model without market frictions in which people either marry or remain single forever. Willis (1999) shows there can be what he calls an 'out-of-wedlock equilibrium' in which men with low incomes seek to father children outside marriage and women oblige them in preference to remaining childless. If the number of women exceeds the number of men and if women at the lower end of the income distribution have incomes, which are sufficient for them to want to raise children with their own resources (including transfers from the welfare system), then fatherhood outside marriage would be free. If low-income men can father children by more than one woman, nonmarital fatherhood may be more advantageous for them than forming a union with a low-income woman, even though the mother and child would be better off in the union. The single low-income mothers gain by becoming mothers rather than remaining childless. Thus, both theories explain nonmarital fertility as a rational choice by low-income men and women, while at the same time higher income couples choose to have children within marriage.

There has been little theoretical analysis of childbearing within cohabiting unions, which has grown in importance in many countries. In unions that are expected to have a short life, it may be akin to the analysis of nonmarital childbearing discussed above. More generally, it shares some features of the analysis of marital fertility, but there are also unique attributes of cohabiting unions that are probably important for fertility decisions. One is their high risk of dissolution (see discussion of divorce risk in Section Life-Cycle Dynamics). In light of it, it is important to treat the cohabiting couple's fertility decision as problem of choice under uncertainty in which each partner is an individual agent whose preferences, expectations, and resources influence the decision. These are new challenges in the economic theory of fertility.

See also: Family Size Preferences; Family Theory: Competing Perspectives in Social Demography; Fertility Theory: Embodied-Capital Theory of Life History Evolution; Fertility Theory: Theory of Intergenerational Wealth Flows; Fertility of Single and Cohabiting Women; Mortality, Biodemography of; Motherhood.

Bibliography

Becker, G., 1960. An economic analysis of fertility. In: Demographic and Economic Change in Developed Countries. National Bureau of Economic Research, Princeton University Press, Princeton, NJ, pp. 209–231.

Becker, G., 1965. A theory of the allocation of time. Economic Journal 75, 493–517.

Becker, G., Lewis, H.G., 1973. On the interaction between the quantity and quality of children. Journal of Political Economy 81, S279–S288.

Cigno, A., 1998. Fertility decisions when infant survival is endogenous. Journal of Population Economics 11, 21–28.

Cigno, A., 2000. Self-enforcing family constitutions. In: Mason, A., Tapinos, G. (Eds.), Sharing the Wealth: Intergenerational Economic Relations and Demographic Change. Oxford University Press, Oxford.

Easterlin, R., 1980. Birth and Fortune. Grant-McIntyre, London.

Ermisch, J., 1989. Purchased child care, optimal family size and mother's employment. Journal of Population Economics 2, 79–102.

Ermisch, J., 2003. An Economic Analysis of the Family. Princeton University Press.

Heckman, J., Willis, R., 1976. Estimation of a stochastic model of reproduction: an econometric approach. In: Terleckyj, N. (Ed.), Household Production and Consumption. Columbia University Press, New York.

Hotz, V.J., Klerman, J., Willis, R., 1997. The economics of fertility in developed countries. In: Rosenzweig, M., Stark, O. (Eds.), Handbook of Population and Family Economics. Elsevier, Amsterdam, pp. 241–347.

Mincer, J., 1963. Market prices, opportunity costs and income effects. In: Measurement in Economics: Studies in Mathematical Economics in Honor of Yehuda Grunfeld. Stanford University Press, Stanford, CA, pp. 67–82.

Moffitt, R., 1984. Optimal life-cycle profiles of fertility and labor supply. Research in Population Economics 5, 29–50.

Rosenzweig, M.R., 1990. Population growth and human capital investments: theory and evidence. Journal of Political Economy 98, S38–S70.

Sah, R., 1991. The effect of child mortality changes on fertility choice and parental welfare. Journal of Political Economy 99, 582–606.

Schultz, T.P., 1997. Demand for children in low income countries. In: Rosenzweig, M., Stark, O. (Eds.), Handbook of Population and Family Economics. Elsevier, Amsterdam, pp. 349–430.

Willis, R., 1973. A new approach to the economic theory of fertility behavior. Journal of Political Economy 81, S14–S64.

Weiss, Y., 1997. The formation and dissolution of families: why marry? Who marries whom? And what happens upon divorce? In: Rosenzweig, M., Stark, O. (Eds.), Handbook of Population and Family Economics. Elsevier, Amsterdam, pp. 82–123.

Willis, R., 1999. A theory of out-of-wedlock childbearing. Journal of Political Economy 107, S33–S64.

Wolpin, K., 1984. An estimable dynamic stochastic model of fertility and child mortality. Journal of Political Economy 92, 852–874.

Family Theory: Economics of Intergenerational Relations

Alessandro Cigno, University of Florence, Firenze, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Voluntary transfers between different generations of the same family may be attributed to either altruistic or nonaltruistic motivations. There is evidence that a substantial part of these transfers, including some apparently altruistic ones, are actually part of an unspoken deal involving three generations at a time.

In economics, the term 'transfer' denotes the delivery of money or other assets without counterpart. That usually is or appears to be the case when giver and receiver are members of the same family. Taxes and government subsidies are referred to as (public) transfers if there is no direct link between what any particular firm or household pays in taxes, and what it gets back in subsidies. A gray area is represented by certain types of old-age pensions, where the size of the individual benefits received are related in some way to the individual contributions made, but the link is either weak or subject to change by political decision.

Here, we are primarily concerned with transfers between different generations of the same family, but we are interested also in how such payments are affected by public transfers. **Figure 1**, drawn from Cigno et al. (2006) shows the distribution, by age of the household head, of monetary help given to and received from noncoresident 'friends or relatives' by Italian households. It transpires from the graph that the probability of making a transfer is highest, and that of receiving one lowest, when the household head is in the 40–59 age range. Similar pictures can be constructed for other countries. What can be the reason for such a pattern? The economic explanations fall into two broad categories. One follows what is called the altruistic approach, whereby individuals are assumed to derive direct utility (satisfaction) from the well-being of other family members. The other assumes that individuals are ultimately self-interested. Transfers are straight, no-strings-attached gifts according to the first explanation, part of a strategy that is expected to bring direct or indirect benefits to the donor according to the second.

Altruistic Explanations

Simple altruistic explanations assume that sentiment is unidirectional: either from parents to children (*descending altruism*) or from children to parents (*ascending altruism*). More complicated versions allow for altruism in both directions.

Descending altruism underlies Richard Barro's and Gary Becker's theories of private transfers (Barro, 1974; Becker, 1974). This assumption allows them to model transfers within the family as the outcome of a single optimization, whereby parents decide how much each family member will consume subject to two constraints. The first of these constraints is that the present value of the consumption expenditures of all family members cannot exceed that of their incomes. The second is that the transfers received (the difference between consumption expenditure and income in any given period) by any family member other than the decision-maker(s) must be positive or zero, otherwise they would be rejected. Under standard conditions, the model predicts that transfers will not decrease (increase) as the income of the donor (beneficiary) increases. This is consistent with the observation (see **Figure 1**) that the probability of making a transfer is highest, and that of receiving one lowest, in middle life, when a person's earning capacity is at its peak. It is difficult to reconcile, however, with the observation that people continue to receive *inter vivos* transfers after the age of 60 (or even 75), when few of them will have living parents.

Descending altruism has an important policy implication. If it is assumed that parents make transfers to their children in the form of bequests, rather than *inter vivos*, the descending-altruism

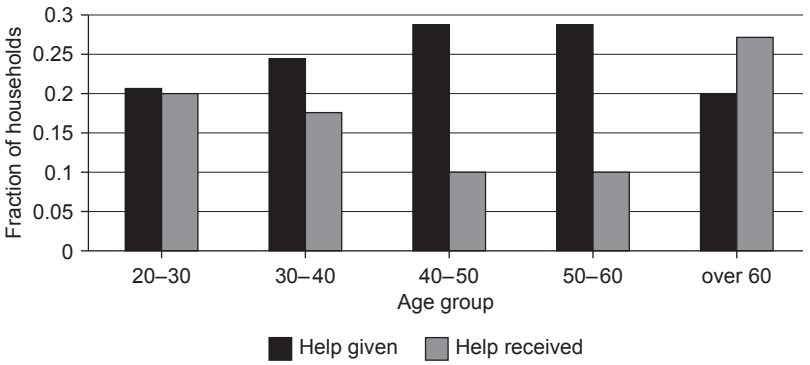

Figure 1 Help given/received by age group.

model predicts that household saving is increased, dollar for dollar, by any current government deficit (Barro, 1974). That is because, conscious of the government intertemporal budget constraint, altruistic taxpayers will perceive any current deficit as a tax on their descendants (this is the so-called *Ricardian equivalence*, from the name of David Ricardo, who first detected it). As taxpayers will then respond to a current government deficit by saving more, in order to be able to leave their descendants bequests large enough to pay for the extra taxes, any net public transfer will be matched by private transfers of opposite sign. Ricardian equivalence does not hold exactly, however, if fertility is endogenous (i.e., if parents can choose how many children to have or, more realistically, condition the probability distribution of births through birth control) and parents derive satisfaction from the number of children, as well as from how much each of these children will consume (Becker and Barro, 1988).

One of the mechanisms that could generate a current government deficit is a pay-as-you-go public pension system, where current pensions are paid out of current contributions. Suppose that the government were to take a dollar off the social security contributions due by each of the present taxpayers, without any corresponding reduction in their pension entitlements. Each member of the next generation of taxpayers would then have to pay more in taxes. Knowing that, present taxpayers would then want to compensate their own children by making larger bequests to each of them. But this would make it more expensive to have an extra child than to raise the well-being of existing children. Under standard conditions, the present generation would then respond to the policy by having fewer children, and *more* than compensating each child for the heavier tax burden he or she will have to shoulder. Of course, the change in fertility behavior would result in a lower number of future taxpayers, and thus raise the burden on each of them. Given that the aggregate effects of a couple's actions are infinitesimal, however, the present generation will disregard them. Hence, the fertility rate will fall. Depending on whether the rise in transfers per child is proportionally larger or smaller than the reduction in the number of children, the saving rate could go either up or down.

By contrast, if the government were to promise higher benefits to current taxpayers already covered by old-age security, or extend coverage to new categories, and simultaneously raise taxes or contributions so as not to force a transfer in favor of present workers, there would be no effect on either saving or fertility. There is evidence (Cigno and Rosati, 1996 among others) that old-age security per se (i.e., in the absence of intergenerational spillovers) encourages household saving and discourages fertility, while deficits discourage saving. This empirical finding is consistent with the prediction of the Becker–Barro model that a deficit *could* result in lower savings, but contradicts that model's other prediction that, in the absence of intergenerational transfers, savings and fertility would remain the same.

The implications of ascending altruism are examined in Nishimura and Zhang (1992). Under this assumption, the direction of transfers is not from working-age parents to young children, but from working-age children to retirement-age parents. That is consistent with the observation that old people receive *inter vivos* transfers, but difficult to reconcile with the one that they also make such transfers. Given ascending altruism, however, parents view their children as an asset. If fertility is endogenous, working-age individuals (or couples) will then choose the composition of their portfolios so as to equate the marginal return to having children to that of holding conventional assets (capital). This carries the implication that any form of government generosity toward current taxpayers at the expense of future ones would induce the former to save more, not in order to leave larger bequests to their children as in the descending-altruism model, but in order to put some of the windfall aside for their own old age. By contrast, a fully funded expansion in pension coverage (a form of forced saving) would reduce voluntary saving. Both these predictions clash with the empirical finding, already mentioned, that deficits discourage voluntary household saving, while an increase in social security coverage per se encourage it.

Whichever its direction, unilateral altruism is thus an inadequate explanation of what we observe. Can bilateral altruism (parents love children, children love parents) do better? If family members were unanimous in their preferences, intrafamily transfers could again be described as the outcome of the maximization of a single preference function (this time common to all family members), subject only to the constraint that the sum of the discounted expenditures cannot be greater than the sum of the discounted incomes of all family members. There is now no restriction on the direction of money flows: depending on life-cycle patterns, transfers from middle-aged children to elderly parents may be positive in some families and negative in others. This is consistent with evidence of two-way flows between generations. But the model has also another behavioral implication: since the amount consumed by each family member depends on *total* family income, and not on his or her own individual income, changing the intrafamily distribution of income should not change the intrafamily distribution of consumption. This carries the strong policy implication that redistributive policies will be ineffective: taking a dollar off the rich (e.g., the parents) and giving it to the poor (e.g., the children) would reduce the amount voluntarily transferred from the former to the latter by exactly one dollar. Statistical tests based on this property (Cox, 1987; Altonji et al., 1992) reject the hypothesis.

Nonaltruistic Explanations

Nonaltruistic explanations are based on the assumption that individuals are interested only in their own lifetime consumption stream. If a person voluntarily surrenders money to a relative, it will be in return for something: either profit, or an improvement in the allocation of consumption (over the life cycle, or across states of the world). There does not seem to be any difference of substance between doing this and giving money to the grocer in exchange for a dozen eggs. Nonetheless, it is customary to classify as transfers any money or other assets changing hands within the family whenever there is no explicit *quid pro quo* (as there would be, for example, if a father paid his child to wash his car). If we want to find a nonaltruistic explanation for intrafamily transfers, we must then look for tacit agreements likely to bring benefits to all concerned.

In the presence of capital market imperfections, an opportunity for mutually advantageous arrangements among members of the same family stems from life-cycle income variations. Standard microeconomic theory predicts that individuals smooth consumption over the life cycle by borrowing from the market when they are young, and lending (buying assets) when they are middle-aged. In reality, however, minors cannot sign contracts, and young adults have difficulty in obtaining credit against future earnings. The very fact that people survive to adulthood in such circumstances is thus evidence that the young receive support from someone. But whom? Self-interested relatives, parents included, do not make presents. If the risk-adjusted return is at least as high as that offered by the formal capital market, they may make loans. There is an enforcement problem, however, in that elderly parents have no means of compelling a now middle-aged child to honor a promise made 20–30 years earlier (all the more so if there was only an implicit commitment, and no promise was actually uttered).

A possible solution is offered by Cigno (1993). Imagine a set of unwritten family rules, call it a 'family constitution,' prescribing the amount that a parent must pay a young child, and the amount that a grown-up child must pay an elderly parent, subject to the clause that nothing is due to a parent who disobeyed the rules. The clause makes paying one's own parents a condition for getting the same treatment from one's own children. A family constitution is self-enforcing if its rules are so designed that complying with them is at least as advantageous, for a person in middle life, as the alternative of having no children, giving nothing to parents, and providing for old age by buying capital. Evidence of the existence of self-enforcing family constitutions is reported in Cigno et al. (2006).

Mutual insurance arrangements ('I will help you if things go well for me and badly for you, you will help me if it is the other way round') are another possible source of intrafamily transfers. Here too, however, there is an enforcement problem. Why should anyone want to enter into such an informal arrangement if it is possible to buy insurance, perhaps even at a lower premium, from the market with all the safeguards afforded by the law? The insurance market has an advantage over the family in that individual risks are less likely to be correlated in a large heterogeneous population than in a small homogeneous group. On the other hand, however, informational (adverse selection and moral hazard) problems are less severe within the family because, compared with total strangers, family members have privileged information about one another's characteristics, and lower costs of monitoring one another's behavior. If the market does not offer insurance to certain individuals or for certain types of risk, or offers it only on very unfavorable terms, entering into some kind of intrafamily arrangement might then be the only or the cheapest way of getting insurance (Di Tella and MacCullogh, 2002).

If an intrafamily insurance agreement involves members of different generations, it will give rise to intergenerational transfers. An example is provided by the case where a family member, say a parent, transfers wealth or precommits to leave a bequest to a younger family member, say to a child, in exchange for a promise of indefinite old-age support (Kotlikoff and Spivak, 1981). Given that the duration of life is uncertain, such an arrangement will insure the parent against the risk of running out of money if he lives to a very old age. The problem is again enforcement, and the solution could again be a family constitution as in the case of intergenerational credit.

Let us now suppose that a family constitution is in place, and imagine that the government decides to raise pension entitlements. The present middle-aged will then want to shift less of their current spending capacity to old age than they would otherwise. Since the amount they are supposed to pay to their elderly parents under present family rules has not changed, however, some of them will find it advantageous to disobey the rules, and consequently give up the idea of receiving support from their own children in the future. They will then pay nothing to their own parents, have no children and possibly top up their expected pension benefits by buying conventional assets. Conversely, if the government were to cut current taxes or pension contributions without reducing pension entitlements, some middle-aged individuals, who would otherwise have found it advantageous to disobey family rules, will now have children instead of buying conventional assets. This is consistent with the already mentioned empirical evidence that aggregate household saving responds positively to government provided old-age security, and negatively to government deficits. It also explains why fertility responds negatively to publicly provided old-age security.

See also: Age: Anthropological Aspects; Family and Kinship, History of; Family as Institution; Family: The Anthropology of the Concept and Its History; Family Theory: Competing Perspectives in Social Demography; Family Theory: Economics of Childbearing; Kinship in Anthropology; Parenting Attitudes and Beliefs across Cultures.

Bibliography

Altonji, J.G., Hayashi, F., Kotlikoff, L.J., 1992. Is the extended family altruistically linked? Direct evidence using micro data. American Economic Review 82, 1177–1198.

Barro, R.J., 1974. Are government bonds net wealth? Journal of Political Economy 82, 1095–1118.

Becker, G.S., 1974. A theory of social interactions. Journal of Political Economy 82, 1063–1093.

Becker, G.S., Barro, R.J., 1988. A reformulation of the economic theory of fertility. Quarterly Journal of Economics 103, 1–26.

Cigno, A., 1993. Intergenerational transfers without altruism: family, market and state. European Journal of Political Economy 7, 505–518.

Cigno, A., Giannelli, G.C., Rosati, F.C., Vuri, D., 2006. Is there such a thing as a family constitution? A test based on credit rationing. Review of Economics of the Household 4, 183–204.

Cigno, A., Rosati, F.C., 1996. Jointly determined saving and fertility behaviour: theory, and evidence for Germany, Italy, UK and USA. European Economic Review 40, 1561–1589.

Cox, D., 1987. Motives for private income transfers. Journal of Political Economy 95, 508–546.

Di Tella, R., MacCullogh, R., 2002. Informal family insurance and the design of the welfare state. Economic Journal 112, 481–503.

Kotlikoff, L.J., Spivak, A., 1981. The family as an incomplete annuities market. Journal of Political Economy 89, 372–391.

Nishimura, K., Zhang, J., 1992. Pay-as-you-go public pensions with endogenous fertility. Journal of Public Economics 48, 239–258.

Family Theory: Economics of Marriage and Divorce

Yoram Weiss, University of Tel-Aviv, Tel-Aviv, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article summarizes the economic approach to marriage and divorce. The economic gains from marriage arise from sharing and coordination. Durable and mutual care help to support efficient outcomes. The gains from marriage and their division influence the decisions to marry and to stay married. Competition in the marriage market determines who will marry whom and the division of the gains. Complementary marital traits lead to a positive assortative matching. Frictions and search explain why, in any given moment in time, part of the population is single and why individuals enter into imperfect unions that dissolve when a better match is found or an unanticipated shock occurs.

From an economic point of view, marriage is a voluntary partnership for the purpose of joint production and joint consumption. As such, it is comparable to other economic organizations that aim to maximize some private gains, but are subject to market discipline. This article describes the working of the marriage market, as a feedback system in which the aggregate divorce and marriage rates influence and are influenced by individual choices, and explains the features of this market that have led to the sharp increase in marital turnover.

The Gains from Marriage

The economic analysis of the family naturally begins with the gains from marriage. These gains are a major determinant of the decision to form a new union or to break an existing relationship. The main economic advantages of being married, rather than being single, are as follows:

1. Children. Although children can be produced and raised outside the family, the family has a substantial advantage in carrying out these activities. There are two interrelated factors that cause this advantage: by nature, parents care about their own children and, because of this mutual interest, it is more efficient that the parents themselves determine the expenditure on their children. Because children are viewed as a public good by their parents, an efficient allocation of family resources between public and private uses requires some coordination between the parents. If the parents live separately, as either single or remarried, the noncustodian parent loses control of child expenditures and is less likely to contribute. In addition, lack of contact may reduce the parent's benefits from the children's 'quality' (i.e., their well-being and success).
2. Division of labor. Family members coordinate their work activities to exploit comparative advantage and increasing returns. For instance, one partner works in the market while the other works at home. This pattern is economically efficient if the partners have different market wages or different productivity at home. Such differences may be inherent or, more likely, acquired through investment in skills. Investment in human capital is more profitable the higher its rate of utilization. This type of increasing returns leads to specialization that can, in principle, create mutual gains through redistribution.
3. Sharing. There are collective (nonrival) goods that can be jointly consumed by both partners: for example, child quality, common leisure activities, housing expenditures, and shared information. The share of such public goods in family expenditures can be substantial. Lazear and Michael (1980) estimate that two single individuals can almost double their purchasing power by forming a union.
4. Credit and investment. Married partners often engage in implicit loans that facilitate investment activities. For example, one partner works when the other is in school, expecting to share in the future gains from the investment. Similarly, parents often finance the schooling of their children, in part in the anticipation of old-age support. Such internal transfers can support investments that are mutually beneficial and would not be carried out because of market imperfections.
5. Risk pooling. A family can spread the risks of idiosyncratic shocks to individual members. For example, one partner can work when the other is sick or unemployed. The family jointly accumulates precautionary savings that can serve either member in case of need.

Realization of these potential gains from marriage requires coordination between the two partners and durable relationship. Durable relationships are motivated by the long-term investment in children and the accumulation of marital capital, which is lost or diminished in value if separation occurs. However, to induce the actions that generate the gains from marriage, it is necessary for the partners to exchange commitments. For instance, a young wife is unlikely to support her husband by working while he is in medical school, if she expects him to leave her and marry a young nurse when he gets his degree. To some extent, emotional factors such as love and guilt replace formal commitments and facilitate efficient investments and effort levels. As an example, consider the role of an altruistic family head. Suppose that the head is given control over family resources and can make transfers as the head sees fit. The only requirement for becoming a head is that they should care about all family members, willing to raise their income, whenever the total resources at the head's disposal rise. Under these conditions, selfish family members

voluntarily act in the interest of the group. The reason is that any productive (destructive) action which increases (decreases) total family resources is rewarded by an increased (decreased) transfer from the family head (Becker, 1991).

There is plenty of evidence for division of labor within families. Typically, husbands work more in the market than do their wives. In addition, married men work longer hours in the market and have substantially higher wages than unmarried men. Similarly, married women have lower wages and work more at home than unmarried women. This division of labor is influenced by the higher market wages of males, which are reinforced by their higher participation in the labor market. Over the years, husbands have increased their share in household work and wives increased their share in market work. But the different household role for men and women are still maintained.

The family is not the only organization that fulfills these functions. Markets and government institutions compete with and complement the family in this regard. If all goods and work activities are marketable, there is no need to form marriages to enjoy increasing returns or to pool risks. Similarly, with good medical or unemployment insurance one does not need to rely on one's spouse. Generally, the role of the family varies depending on market conditions and government policies. In addition, human partnerships need not be confined to couples of opposite sexes. One also observes 'extended families' of varying structures that coordinate the activities of their members and provide self-insurance. The prevalence of male–female partnerships has to do with sexual attraction, which triggers some initial amount of blind trust and the production of children. These emotional and biological considerations are sufficient to bring into the family domain some activities that could be purchased in the market. Then, the accumulation of specific 'marital capital,' such as children and shared experience, increases the costs of separation and creates incentives for a lasting relationship. In this sense, there is an accumulative effect in which economic considerations and investments reinforce the natural attachment.

The potential gains from marriage generally depend on the personal traits of the two partners. Some individual traits are complements, and the gains from marriage rise if the partners have similar endowments. For instance, similarity in tastes or education facilitates agreement on collective goods, such as the education of the children. In other cases, the traits are substitutes, and gains from marriage rise if the partners have different traits. For instance, the gains from division of labor and from credit are higher when the wages or incomes of the partners differ.

Family Decision-Making

The existence of potential gains from marriage is not sufficient to motivate marriage and to sustain it. Prospective mates need to form some notion as to whether families realize the potential gains and how they are divided. Because families consist of several members with potentially conflicting interests, there is a question as to how families reach decisions. The old notion that consensus leads to joint maximization of a common objective is too narrow. However, because of their proximity and repeated interaction, it is plausible to assume that married partners reach some sort of agreement that specifies an efficient allocation of resources and a division of the gains. These two principles replace the principle of rationality in individual choices.

In a special case, referred to as transferable utility, it is possible to separate the issues of efficiency and distribution. This situation arises if there exists a commodity (say money) that, upon changing hands, shifts utilities between the partners at a fixed rate of exchange. In this case, the family decision process can be broken into two steps: first choose an action that maximizes a weighted sum of the individual utilities, then use money to divide the jointly maximal outcome. The basic idea is that the partners can agree on most aspects of their choice and restrict their bargaining to transfers of money, which do not detract from total resources.

In general, the problems of efficiency and distribution are intertwined. The family can still be described as maximizing a weighted sum of the individual utilities, but the weights depend on the partners ability to affect the family choice. There is no need, however, to specify a particular bargaining procedure. It is sufficient to assume that partners have different 'decision powers' within the family that may depend on individual attributes, such as earning capacity, subjective factors such as impatience and risk aversion, and on market conditions, such as the sex ratio and availability of alternative mates. The 'collective approach' assumes that partners always reach an efficient outcome and examines family choices under varying circumstances based only on the assumption that the family has a stable efficient rule of behavior (Browning et al., 2014: Chapter 3).

It is empirically testable whether families maximize a joint utility function, whether they act efficiently, and it is also possible to recover the sharing rule from observed family choices. Recent evidence shows that, holding total family income constant, an increase in the income of one member shifts the allocation of consumption goods in their favor. These findings reject the hypotheses that the family maximizes a joint objective, which would imply that within family, allocations are invariant under such redistribution of total resources, but are consistent with both cooperation and noncooperation. However, only cooperation implies efficiency, which yields testable cross-equation restrictions on family demand. Based on data on consumption and work patterns within families, efficiency is not rejected, suggesting that cooperation is the relevant model (Browning et al., 2014: Chapter 5).

In the context of uncertainty, however, efficiency has some further implications. Efficient risk sharing among linked individuals implies that, holding aggregate consumption constant, the consumption of each family member is independent of idiosyncratic shocks such as fall into unemployment or bad health. Stated differently, all individuals in the household are affected by a random shock to any individual income and their consumption levels move together. Here again commitment plays an important role. If commitment is incomplete the consumption of each spouse will depend also the varying 'decision powers' of the partners. Using data from Consumer Expenditure Survey, Mazzoco (2007) finds that full commitment is strongly rejected.

Divorce and Its Economic Consequences

Modern marriage markets are characterized by high turnover; men and women divorce and remarry at higher rates than in the past. However, different individuals have different marital histories; they marry, divorce, and remarry at different rates (Bruze et al., 2011). To explain this variation it is crucial to understand the two-sided aspect of marriage markets. Marriages form and dissolve based on preferences and expectations of two *different* individuals who operate in a 'marriage market' with many competing agents. As time passes, new information on match quality and outside options is accumulated, and each partner decides whether to dissolve the partnership or to continue the marriage. In making this choice, partners must consider the expected value of each alternative, where the value of remaining married includes the option of later divorce, and the value of divorcing includes the option of later remarriage. Divorce occurs endogenously whenever one partner has an alternative option that the current spouse cannot, or is unwilling to, match by a redistribution of the gains from marriage.

The model outlined above yields several testable implications:

1. It is the unanticipated changes in the characteristics of the partners or the quality of match which trigger divorce. It is clear that falling out of love can destroy the gains from the current marriage and cause divorce. It is less obvious how unanticipated changes in personal attributes, such as earning capacity, influence divorce, because they affect both the gains from the current marriage and the outside options. However, the partners were matched, based on their traits as observed at the time of marriage, any unanticipated change reduces the gains from interaction. Thus, if the husband is unexpectedly wealthy, he may seek a better wife, and if he becomes unexpectedly poor, his wife may seek a better husband. Note that both positive and negative 'surprises' can disrupt the marriage.
2. If the gains from marriage are substantial, small shocks will not lead to divorce. Therefore, the probability of divorce will be lower among couples who are well matched. Anticipating that, couples sort into marriage according to characteristics that are likely to enhance the stability of the marriage. In this respect, the decisions to marry and divorce are linked.
3. By the same logic, costs of divorce mitigate the impact of unanticipated shocks on marital dissolution. The costs arise from loss of specific capital, such as information about the current spouse, and loss of coordination, especially with regard to child expenditures. These costs of divorce are influenced by the postdivorce transfer of resources in the form of alimony and child support. There are also emotional and legal costs associated with the break up.
4. Somewhat more controversial is the role of divorce laws, in particular whether the legal possibility unilaterally to walk away from a marriage influences the divorce rate. With transferable utility, such a change should only affect the shares in the gains from marriage but not the decision to separate, because the partner who wishes to continue the marriage can compensate the one who wishes to leave by giving up part of the gains from the marriage. Conversely, the person who wants to leave can pay damages to the one left behind. In general, the outcome depends on the ability of partners to transfer resources within marriage and following marriage.

These implications have been tested empirically. There is evidence that surprises matter. Specifically, an unexpected increase in the husband's earnings capacity reduces the divorce hazard, while an unexpected increase in the wife's earning capacity raises the divorce hazard. The quality of match also matters. For instance, couples in which both partners are highly educated tend to be more stable (Weiss and Willis, 1997; Bruze et al., 2011). There is only weak evidence that divorce rates are higher in states where 'fault' is not a prerequisite for divorce (Stevenson and Wolfers, 2007).

The Marriage Market

Individuals in society have many potential partners. This situation creates competition over the potential gains from marriage. In modern societies, explicit price mechanisms are not observed. Nevertheless, the assignment of partners and the sharing of the gains from marriage can be analyzed within a market framework, because an undesired marriage can be avoided or replaced by a better one.

Any market solution is a particular assignment of males to females. The major questions of interest are the following: what types of matches are likely to form in equilibrium; for instance, would a class structure emerge where the rich marry the rich? What determines the rates of marital turnover and the proportions of married and single individuals in society? Does the marriage market operate efficiently and, if not, what type of intervention is required?

The marriage market determines not only the assignment of partners but also the division of resources and activities within the family. As market conditions change, a shortage of suitable partners of a particular kind leads to an increase in their gains from marriage. For instance, there is evidence that an increase in the demand for men, created by faster population growth combined with the tendency of men to marry younger women (a marriage squeeze), has led to an increase of dowries in rural India. In modern societies, up-front payments are rare and market forces are mostly revealed by the division of labor and consumption patterns within families. There is a recent literature in which marital shares are either inferred from marriage patterns, as in (Choo and Siow, 2006; Chiappori et al., 2012) or estimated directly from data on leisure and consumption within couples as in (Chiappori et al., 2002; Mazzocco et al., 2007). Additional information on the (expected) gains from marriage is contained in the decisions to enter marriage and to stay married (Bruze et al., 2011).

Matching models provide a starting point for such analysis. These models investigate the mapping from preferences over prospective matches into a stable assignment (Roth and Sotomayor, 1990; Browning et al., 2014: Chapter 7). Reflecting the assumption that marriage is voluntary, an assignment is said to be stable if no married person would rather be single and no two (married or unmarried) persons

prefer to form a new union. To illustrate, assume that each male is endowed with a single trait, m, and each female is endowed with a single trait, f, which positively affects the gains from marriage, denoted by $z = g(m, f)$. Suppose, first, that z is a public good that both partners can consume jointly. Then, the only stable assignment is that in which males with high m marry females with high f, and, if there are more (less) eligible men than women, the men (women) with the lowest endowments remain unmarried, a positive assortative mating. However, if one assumes, instead, that z can be divided between the two partners, a positive (negative) assortative mating occurs only if the two traits are complements (substitutes). Matching models refer to these two situations as nontransferable and transferable utility, respectively. With transferable utility, a man with low m may obtain women with high f by giving up part of his private share in the gains from marriage. Complementarity (substitution) means that the two traits interact in such a way that the benefits from a woman with high f are higher (lower) for a male with high m than for a male with low m. The type of interaction in the gains from marriage determines the willingness to pay for the different attributes. An important lesson is that in a marriage market, with sufficient scope for compensation within marriage, the best man is not necessarily the one married to the best woman because, with negative interaction, either one of them can be bid away by the second best of the opposite sex (Becker, 1991).

The process of matching in real life is characterized by scarcity of information about potential matches. The participants in the process must spend time and money to locate their best options. The realized distribution of matches and the division of the gains from each marriage are therefore determined in an equilibrium which is influenced by the costs of search and the search policies of other participants. The main ingredients of the search model are as follows. There is a random process that creates meetings between members of society of the opposite sex. When a meeting occurs, the partners adopt a reservation policy where the match is accepted only if it provides expected gains above some prespecified value. Otherwise, they depart and wait for the next meeting to occur (Mortensen, 1988).

Because meetings are random and sparse in time, those who actually meet and choose to marry enjoy a positive rent. An important issue is the division of these rents between the partners. The literature mentions two considerations that determine the division of the gains from marriage: outside options, reflected in the value of continued search, and the self-enforcing allocation that would emerge if the marriage continued without agreement. Combining these two considerations, the sharing rule will be influenced by both the value of search as singles (outside the current match) and the value of continued search during the bargaining process, including the option of leaving when an outside offer (whose value exceeds the value of potential agreement) arrives. In this way, a link is created between the division of gains and the market conditions. For instance, if there is excess supply of women in a particular marriage market, the value of continued search by females is reduced and consequently their share in the gains from marriage declines.

With frictions, there is still a tendency to positive (negative) assortative mating, based on the type of the interaction in traits. If the traits are complements, individuals of either sex with higher endowment will adopt a more selective reservation policy and will be matched, on the average, with a highly endowed person of the opposite sex. However, with sufficient friction, it is possible also to have negative assortative mating under complementarity. The reason for this result is that, because of the low frequency of meetings, males with low m expect women with high f to accept them, and if the gains from such a match are large enough, they will reject all women with low f and wait until a high f arrives (Burdett and Coles, 1999; Shimer and Smith, 2000).

Without frictions, one would expect negative sorting on wages and positive sorting on nonwage income, because the gains from division of labor are larger when wages differ, while the gains from sharing goods are larger when incomes are similar (Becker, 1991). However, empirical findings suggest positive sorting on both wage and nonwage income. In particular, there is a substantial correlation in the schooling achievements of partners to marriage.

Frictions and search explain why, in any given moment in time, part of the population is single, even though there are positive gains from marriage. The aggregate stock of unattached individuals is determined by the rates of entry and exit into marriage implied by the arrival rates of offers and the optimal reservation policies chosen by individuals. These decisions, in turn, depend on the aggregate proportion of singles, because if one meets an unattached person it is more likely that a new union will be formed than upon meeting an attached person, who already has some positive rents. Therefore, the higher the proportion of singles (including divorcees), the higher is the private expected gains from divorce. This type of reinforcement can lead to multiple equilibria, whereby both high and low aggregate divorce rates can be sustained through the induced change in the individual incentives to marry and divorce.

The time pattern of aggregate divorces is consistent with multiple equilibria. For instance, the divorce rate in the USA doubled during the decade between 1965 and 1975, standing at roughly 10% from 1940 to 1965 and at roughly 20% from 1975 to 1995, suggesting a switch from a low to a high equilibrium. Explanations for the timing of the change include the appearance of the contraceptive pill, break up of norms, and legal changes (Michael, 1988). The main conceptual point, however, is that any change in exogenous factors, even a small one, can have marked impact on the aggregate divorce rate, because of the inherent reinforcements in marriage markets.

Broadly viewed, divorce is a corrective mechanism that enables the replacement of bad matches by better ones. The problem, however, is that private decisions may lead to suboptimal social outcomes because of the various externalities that infest search markets. These externalities exist at the level of a single couple and the market at large. In the models outlined above, those who initiate the divorce fail to internalize the interest of their spouse in continued marriage and the custodial mother that remarries fails to internalize the impact on the father. Moreover, children may suffer even if their welfare is fully internalized, because the child's welfare is only one consideration in the parents' marital choices. At the market level, a person who chooses to divorce fails to take into account the impact on the remarriage prospects of others, and if commitments are made, on the quality of prospective mates.

Some Policy Issues

Despite its firm roots in nature and its antiquity in human society, the future of the family institution has been recently put into question. This was motivated by recent trends, common to all Western societies, of declining marriage rates, declining fertility, higher divorce rates, and a rise in alternative arrangements such as cohabitation, single-parent households, and single-mother families. Social observers view these trends with considerable alarm especially because of the potential harm to children (more than one million children in the USA are involved in a divorce every year).

Although there is ample evidence that divorce reduces the welfare of single wives and of children with single or stepparents, this is only one part of the picture. Continuation of marriage under adverse conditions can have equally harmful results, although these are harder to identify.

It has been observed that divorced husbands, even if relatively well to do, fail to support their exwives and their children at the standard to which they were accustomed during marriage (Weiss and Willis, 1997; Chiappori and Weiss, 2007). Consequently, divorced women and children in their custody suffer a large decline in income. The possible explanations for this phenomenon are: lack of binding marriage contracts, the inability of noncustodial parents to monitor expenditures by the custodian, and loss of interest in the child by the noncustodian parent. In most cases, the wife obtains custody and controls the expenditure on children. This situation creates an agency problem, because out of every dollar transferred to the custodial wife with the intention of raising the welfare of the child, part is used for her own consumption. Because of this 'tax,' noncustodial fathers reduce their transfer, and consequently child quality is reduced. The assignment of custody to the wife is usually motivated by her comparative advantage in child care, but the agency problem implies that such a practice has shortcomings that are not easy to overcome, because the courts cannot verify within-household allocations.

The adequacy of transfers must be judged not only on the basis of the efficiency of allocation of family resources in the aftermath of divorce, but also on the basis of partners' options and expectations at the time of marriage. From the latter point of view, efficiency requires that the consumption levels during marriage and divorce are tied together. Courts do take into account these two broad considerations in awarding divorce settlements. These are usually based on postdivorce incomes and the accustomed consumption of the wife and children during marriage. However, the evidence suggests that actual transfers fall short of the efficient level from both the ex post and ex ante points of view.

With deficient transfer mechanisms, the partners must better prepare for the event of divorce. One important instrument is the allocation of time within marriage. By investing in human capital, each partner can be less dependent on transfers in the event of divorce. However, such investments may detract from marital output. For instance, a wife who works is better defended against divorce but has less time to spend with children. Indeed, there is evidence that women tend to increase their investment in market work in anticipation of divorce. Thus, if the risk of divorce rises, children may suffer even before the marriage breaks (Browning et al., 2014: Chapter 6).

Because meetings are random and uncoordinated, the search behavior and the decision whether to marry or divorce of each individual influences the marriage (and remarriage) prospects of all other members of society who are potential matches. Therefore, legal intervention must take into consideration the impact of the law not only on particular families but also on the marriage market as a whole. These two considerations may be in conflict. As an illustration, consider an increase in the size, or the enforcement, of child-support payments. Holding the aggregate divorce and remarriage rates fixed, each family in isolation may be better off as a result of the additional insurance. However, the reduced propensity to divorce can have a negative effect on the remarriage prospects of those who have suffered a large negative shock to their quality of match and who wish to separate. Another externality that must be considered, and may operate in the opposite direction, is that the availability of alternative mates can reduce investments that are specific to the current match and detract from its quality. This is indeed the logic behind the laws and religious norms aimed at restricting divorce.

The links between poverty and family take a particularly strong form in the black communities in the US. It has been argued by Wilson (1987) that black women delay or avoid marriage and have children out of wedlock because of a shortage of marriageable black men with a steady job and good income. As pointed out by Willis (1999), low-income men may also forego marriage in order to father children out of wedlock with low-income women who voluntarily bear and rear these children. This type of analysis illustrates the general point that it is not simply the number of men and women that determines marital and fertility outcomes but also their respective distributions of traits. Moreover, shifts in the distribution of traits of men can affect the marriage behavior and work behavior of both men and women. For instance, Seitz (2009) shows that if black men had the same wage distribution as whites, they would work more than black women (as among whites) and, surprisingly, the fraction of marriages in the black population would decline. This occurs in her simulations because the decision power of black women within marriage declines and, consequently, the share that the wife receives from the marital gains declines too when black men receive higher wages.

The observed changes in family arrangements are closely related to the dramatic changes in labor-market participation and education of women, and the associated technological advances in the workplace and in the home. Whatever the causal relationships, it is clear that these processes reinforce each other and generate complex dynamics. While the reduction in marriage has been fairly smooth, the changes in divorce appear discontinuous, suggesting search externalities and perhaps the breaking of norms in the processes governing divorce and remarriage. These abrupt changes cause a special adjustment problem that requires social and legal restructuring. However, the expected outcome for the early twenty-first century is that the marriage market will settle at a new equilibrium with higher marital turnover, reflecting the lower gains from marriage as the disparity in the market productivity of men and women gradually declines. Although reduced, the gains from marriage are unlikely to disappear and the family, in one form or another, will continue to have a major role.

See also: Family as Institution; Marriage and the Dual-Career Family.

Bibliography

Becker, G., 1991. A Treatise on the Family, enl. ed. Harvard University Press, Cambridge, MA.

Browning, M., Chiappori, P.A., Weiss, Y., 2014. Economics of the Family. Cambridge University Press, Cambridge.

Bruze, G., Svarer, M., Weiss, Y., 2011. The Dynamics of Marriage and Divorce. Aarhus University.

Burdett, K., Coles, M.G., 1999. Long-term partnership formation: marriage and employment. Economic Journal, Series F 109, 307–334.

Chiappori, P.A., Weiss, Y., 2006. Divorce, remarriage, and welfare: a general equilibrium approach. Journal of the European Economic Association 4, 415–426.

Chiappori, P.A., Weiss, Y., 2007. Divorce, remarriage, and child support. Journal of Labor Economics 25, 37–74.

Chiappori, P.A., Fortin, B., Lacroix, G., 2002. Marriage market, divorce legislation, and household labor supply. Journal of Political Economy 110, 37–72.

Chiappori, P.A., Salanie, B., Weiss, Y., 2012. Partner Choice and the Marital College Premium. Columbia university.

Choo, E., Siow, A., 2006. Who marries whom and why. Journal of Political Economy 114, 175–201.

Lazear, E.P., Michael, R.T., 1980. Family size and the distribution of real per capita income. American Economic Review 70, 91–107.

Mazzoco, M., 2007. Household intertemporal behavior: a collective characterization and a test of commitment. Review of Economic Studies 74, 857–895.

Mazzocco, M., Ruiz, C., Yamaguchi, S., 2007. Labor Supply, Wealth Dynamics and Marriage Decisions. UCLA.

Michael, R., 1988. Why did the US divorce rate double within a decade? Research in Popular Economics 6, 367–399.

Mortensen, D.T., 1988. Matching: finding a partner for life or otherwise. American Journal of Sociology 94 (Suppl.), S215–S240.

Roth, A., Sotomayor, M., 1990. Two Sided Matching: A Study in Game-Theoretic Modeling and Analysis. Cambridge University Press, Cambridge, UK.

Seitz, S., 2009. Accounting for racial differences in marriage and employment. Journal of Labor Economics 27, 385–437.

Stevenson, B., Wolfers, J., 2007. Marriage and divorce: changes and their driving forces. Journal of Economic Perspectives 21 (2), 27–52.

Shimer, R., Smith, L., 2000. Assortative matching and search. Econometrica 68, 343–370.

Weiss, Y., Willis, R.J., 1997. Match quality, new information, and marital dissolution. Journal of Labor Economics 15, S293–S329.

Willis, R.J., 1999. A theory of out-of-wedlock childbearing. Journal of Political Economy 107 (Part 2 Suppl.), S33–S64.

Wilson, W.J., 1987. The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy. University of Chicago Press.

Fashion, Sociology of

Sellerberg Ann-Mari, Lund University, Lund, Sweden
Aspers Patrik, Uppsala University, Uppsala, Sweden

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M. Sellerberg, Volume 8, pp. 5411–5415, © 2001, Elsevier Ltd.

Abstract

Fashion is a social phenomenon *par excellence*. A simple sociological definition of fashion is 'being first with the latest.' Its underlying principle is revealed in the link between the *modus* derivatives and the term 'modern,' with its original meaning of 'now' or 'for today.' Fashion revolves around imitation and diffusion, and it applies to all domains of social life, though most of the research is on garments. The field of fashion has been neglected, but since the mid-1990s, it has attracted sociologists studying production of fashion, identity, and the body in relation to fashion. What is lacking is theory development.

Introduction

Fashion is a social phenomenon *par excellence*. As a phenomenon it may be as old as mankind, but became common with the emergence of consumer society in the nineteenth century. It was then that social scientists started to show interest in fashion. A simple sociological definition of fashion is 'being first with the latest.' The definition implies that fashion is about change, and that an object, style, or activity stands out against a backdrop of stability. A fashion is understood as being right in a specific time and context. What is right is 'defined' by a clique of people, fashion leaders, who others see as having the status to determine what is in fashion. Fashion needs followers, but it is also necessary that some people are out of fashion, since not everyone, or everything, can be 'in fashion.' In fact, once something is too common it is likely to be replaced; if not, fashion cannot be used to make distinctions. Thus, fashion is a recurrent phenomenon that creates its own history, from which ideas may be reused or reinterpreted.

While the word 'fashion' in English comes originally from the Latin *facere*, to make, in many European languages it derives from *modus*, to measure. From the sixteenth century, fashion refers to conformation to existing taste, which is also when the notion got a more contemporary meaning. Mode refers to manner, and there is a strong connection to 'modern' and 'modernity,' concepts that are connected to capitalism and development. Its underlying principle is revealed in the link between the *modus* derivatives and the term 'modern,' with its original meaning of 'now' or 'for todays.' Fashion is a general social phenomenon of diffusion, but which in both everyday life and in the social sciences mainly means garments, and in particular haute couture. However, already Blumer outlined a much larger scope of fashion: "painting, music, drama, architecture, household decoration, entertainment, literature … medical practice, business management, political doctrines, philosophy, psychological and social science, and even such redoubtable areas as the physical science and mathematics" (Blumer, 1968:342). It is, in fact, hard to *a priori* exclude an area from the analysis of fashion.

Fashion is different from innovation, style, and trend. Fads refer to processes of diffusion of things or activities that normally include many actors, such as when the skateboard was introduced. Fads do not have a history of replacing each other. Fashion, furthermore is different from innovation. Both fashion and innovation replace something existing, such as an older fashion or an existing practice or technique, but innovations permanently alter the social practice or things, for example, the material people use to insulate houses. An innovation, hence, has lasting effects, and may be seen as a progress according to a scale of evaluation. Fashion is different from style, which refers to a multidimensional self-referential aesthetic system produced and extended over. The styles of punk or Goth are examples. A trend, finally, is a process with an identifiable direction over time, such as when people are dining out more frequently. A trend, for example, of how one should present food, thus, may encompass several fashions, of where one 'should' eat out in various cities.

Fashion was once a limited phenomenon, with very few in a society able to participate. In modern societies, however, economic well-being has made it possible for the vast majority in many countries around the globe to participate in the 'fashion race.' Wealth is not the only factor; however, people also enjoy far greater psychological mobility, and the new form of communication it generates makes fashion more important today. In *Psychologie économique* (1902), Tarde coined the term 'inter-psychologie.' Society is the sum of people's mental condition in constant interaction. Free time and consumption become increasingly significant vehicles for such communication. New needs are born out of this mental interaction.

Social science research has for almost 100 years treated fashion as a field of lesser importance. Researchers, not the least sociologists, have over the last 15 years paid more attention to fashion, especially to the production of fashion, and there is today also an emerging discipline called fashion studies. It is at the same time clear that the field needs more theoretical development.

Fashion in Different Societies

Fashion's universal nature does not preclude strong variations in its social significance from society to society. Thus, Tarde (1903) differentiated between customary and fashionable

societies. In the former, people see *custom* as governing their lives; they are more passionate about their country than about their period because custom prizes the past above all. In periods where *fashion* predominates, people are instead prouder of their era than their country.

Variations between societies are thrown up by different conditions. In economically stagnant periods, conventional stratification variables such as class and income are more decisive in people's lifestyles. In economically expansive periods, increasing importance is laid on fashion instead. Modern society is often contrasted with earlier epochs, and it is argued that fashion's role has markedly increased, although many reasons are given for the change. An influential argument is that the differences in lifestyle of different groups have increased in number and scope, making it possible for people to exhibit their group's distinctive nature.

Simmel sees the 'need for distinction' and the 'need for union' as preconditions for fashion. Certain societies therefore lack the necessary motivation. Simmel (1904) mentions that in fourteenth-century Florence, male dress was devoid of fashion. The 'need for union,' the need to express how certain groups were distinct from others with the help of fashion, was absent, and everyone could dress according to personal preference. Another of Simmel's examples was Venice at the same time. Again, we find that fashion is absent from upper-class dress. Venetian aristocrats were ordered to dress in black so that the lower classes would not know how few they were. They did *not* want to differentiate themselves, and thus the 'need for distinction' that fashion requires was missing. Their 'uniform' can be interpreted as one dimension of what constitutes their status group, as Weber would call it.

Variations in fashion's importance have also been tied to women's standing in society. According to Veblen (1899), it was woman's role to be a vicarious consumer; her consumption symbolized the man's wealth, and thus she became the vehicle of fashion. In Simmel (1904), however, we find an early formulation of compensation theory. Fashion acts as a safety valve. Deny women self-expression in other areas, and the only thing left for them is fashion. In fourteenth- and fifteenth-century Germany, the chance of personal development had increased, which, however, was denied to women. Rarely have more 'hypertrophic' modes of female dress been seen than in this period.

Simmel (1904) further explains fashion differences between societies. Fashion has two fundamental, symbolic functions: an ability to mark difference and an ability to express community. In many nonindustrialized societies, the need for union is acute, while the need for distinction is weak. Simmel mentions certain African tribes who lacked both class differentiation and the concomitant need to signal differences. Where people in these societies want to note differences, it is often a matter of open hostility, which in turn is sterile ground for fashion. Given that fashion functions best where the distance between groups is otherwise only vaguely established, in instances where open hostility already exists, fashion as a sociological form becomes meaningless.

A similar argument can be found in König (1974), who identifies three different stages in society, each with a different type of fashion. The first significant historical change was the emergence of a fixed class hierarchy; at this stage, fashion is the preserve of exclusive, feudal nobility. In the second stage, rigid differences between upper and lower classes are weakened; lower classes adopt the fashion of the elite, ultimately forcing it to alter fashion constantly to retain its individuality. The third stage sees the development of egalitarian democracies and industrial technology. This modern form of society brings with it the spread of fashion through mass consumption.

Fashion as a Process of Change

Fashion is about change, but this change presumes that other things are ordered; if not, changes would not be perceived. The question of what drives fashion's continuous changes has kept researchers busy, and a series of different theories has been given in answer. In Simmel's analysis of fashion, we find a foretaste of many of the theories that were to follow.

Social Differentiation

Many have stressed fashion's origins in class. Though class has its origin in production, its cultural and symbolic meaning is expressed by means of consumption. The analysis by Veblen (1899) of patterns of fashion dispersal has become a classic. A fashion begins with the 'leisure class,' the group with most money. It is attractive to this class because fashion items cost a great deal more than regular items. By degrees, the fashion is adopted by the upwardly aspiring middle classes. Simmel (1904) also describes how fashion must constantly be supplied with new content if it is to give social expression both to similarities within groups and differences between them. To describe fashion's downward spread through society, the *trickle-down* model was adopted by Barber and Lobel (1953), who dismissed the argument that fashion is irrational as obscuring its connections with the economic system. Meanwhile, named for the imitation of the elite that brings in turn an alteration in elite habits, we have the chase and flight model.

Bell (1976) follows Veblen in seeing economic competition as a force behind changes in fashion. Dress for the middle and upper classes is an expression of wealth. This wealth can be expressed in the form of *conspicuous consumption*, *conspicuous leisure*, and *conspicuous waste*. To these, Bell adds his own category of *conspicuous outrage* to convey the fashion-setting class-conscious choice of clothing that does *not* conform with prevailing notions of good taste.

The theory that fashion is the means of social distinction for the upper classes in a hierarchical society has had great influence. Mimicry of social superiors and an urge to be different from 'inferiors' are often noted as the reasons for fashion's changes, while an individual's choice of style is seen as a significant means of securing his or her position in society. Such choices often establish their distance from those below them in the social pecking order and express subtle distinctions of degree within one and the same class. Researchers who present other models often start from the theory of social differentiation, even if they distance themselves from its application to today's society. Some argue that the idea that fashion was the distinguishing mark of the elite was probably valid in the eighteenth and nineteenth centuries, but this is less

so today. It has been argued that the traditional upper class has lost its fashion primacy to be replaced, for example, by today's powerless elite. This new elite, termed *Divi* by Alberoni (1967), uses the mass media to present new lifestyles and new consumption. Theories stressing that ideas through the means of trend scouts also move from the street to the designers and thus affecting or even determining fashion have been presented. Others have discussed how subgroups, for example, Goths (Hodkinson, 2002) have developed their own fashions. It is nonetheless fair to say that one cannot speak of one fashion only, but of fashions.

Fashion's Inner Dynamic

Simmel's analysis included a further suggestion of what drives fashion. Fashion is made up of internal, conflicting forces that necessarily lead to change; it is distinguished by differentiation *and* union, superficiality *and* profundity, freedom *and* dependence, individuality *and* imitation, destruction *and* construction, leadership *and* submission. Antagonistic forces are united in fashion in a way that ensures continual change. Nedelmann (1990) develops the idea of the inherently contradictory nature of fashion, describing its *Eigendynamik*. Fashion, by its nature, contains a stimulus for change derived from polar opposites. People like to imitate, thus spreading a certain fashion, but are equally drawn to differentiation. Imitation leads to the spread of fashion, while new ideas and differentiation lead to its contraction. Nedelmann invokes Simmel, writing that the greater the degree of concord within the upper classes, the more frantic the search for imitation from beneath, and the more constant the search for new fashions by the upper classes. On the other hand, the more the members of the lower classes use style to differentiate between themselves, the less the upper classes are inclined to come up with new fashions. Fashion in this society thus has less significance.

Davis (1992) has also noted the ambivalence of fashion, and identifies its source in sexuality. Davis writes of a continuous dialectic in clothing between the erotic and the modest. Sellerberg (1994) similarly stresses the significance of antagonism as a driving force of change. Fashion reduces *and* creates complexity. It is a continuous break with authority, but its break is authoritative: 'this is what matters, for now.' In this contradiction one finds the motor of change.

System Theories

Davis (1992) differentiates between two types of sociological model used to explain the dissemination of clothing styles. The first, the *populist* model, centers on the consumer's role and the laymen's innovations in dress. Researchers who use the populist model study general consumption, for example, how groups such as teenagers, surfers, gays, skateboarders, or feminists set their own stamp on fashion. The other model is a *system* model. Some system models are predicated on the existence of established social centers for innovation, for example, Paris or Milan, with their fashion shows, designers, editors, wholesalers, and stylists. The fashion system is here seen as a closed circle of producers, distributors, and consumers (Kawamura, 2004).

The trickle-down theory has been criticized by system theorists because it does not focus on the complex structures of organizations and marketing that channel and mediate the fashion process. Those who use a system model aim to analyze the influence of the different elements of the fashion system. One subject of controversy has been whether designers are creative interpreters or merely passive intermediaries in the commercially vital process of heeding consumers' wishes and reconciling them with the producers' requirements.

Fashion as a Break with Convention

The destruction of what exists and the construction of something new in the process are the elements that unite in fashion, according to Simmel. Researchers who today see fashion as breaking current rules usually identify groups other than the socioeconomic upper class as leading fashion. Campbell (1992) has criticized the class differentiation theorist's explanations and identifies different types of 'new': the new, the innovative, and the novel. The novel – new as different – is central to fashion as it breaks with what already exists. The necessary conditions for creating novelty exist in peripheral groups such as bohemians and artists. It is these groups who set fashion in motion, *not* an upper class. To argue that fashion is unconventional, however, is also to argue that fashion continuously reestablishes social conventions and social order. Blumer (1969) sees both similarities and differences between fashion and custom. In both cases, it is a question of social ordering. Custom, however, is static or changes only slowly, while fashion is a constant creator of social order. In Blumer's argument, every fashion emerges from an intensive process of 'collective selection.' The alternatives that are ultimately selected become fashion in different areas. Fashion conjures order out of the vast jumble of possible styles where no objective norm can determine 'right' or 'wrong.'

Fashion as a Psychological and Social Expression

Psychologically oriented theories of fashion frequently center on general human needs such as the wish to be appreciated. It is often pointed out that people's need to display and strengthen their ego is particularly strong. Fashion has great shock value in the hunt for self-expression; it provides a tangible 'look at me.'

In a discussion of fashion as an expression of personality, we can return to Simmel. Fashion makes it possible to express opposites. While conveying personality (paradoxically in the extreme imitation of the modern), it also offers the chance to *avoid* conveying individuality, and thus functions as the ego's 'iron mask.'

Fashion offers a means of psychological expression, but also a social expression of protest. People express in fashion the impulses and wishes that are *not* sanctioned by normal social conventions. There is thus a constant tension between the more established 'domain of culture' and the marginal 'domain of fashion.' In fashion, young people express their stance through wearing what König (1973) calls 'anti-fashion.'

The idea of fashion as a vehicle of social protest has been criticized, however. Protest styles are usually incorporated into fashion and exploited economically (Emberley, 1987).

Others argue that fashion's break with convention is actually a pseudodeviation. Although certainly defying convention, each break loses its original edge as soon as it is expressed in fashion (Wilson, 1985).

Fashion has been interpreted as an expression of the search for a personal identity in an insecure world, and 'fads and fashions' as an expression of the worry that typifies our time. This has sometimes been interpreted as a negative phenomenon. Blumer (1969) thinks that the operation of fashion becomes natural in a changing world where people must continuously cut their ties with the past. Individuals are liberated from earlier models, and orient themselves toward the future. The importance of fashion is as clearly seen among those who follow or try to determine fashion, some of whom turn into 'fashion victims,' as among those who actively argue for an antifashion. Regardless of which attitude a person has, it is hard to stay 'outside' of fashion.

The Production of Fashion

Fashion is not only an effect of the increasing wealth, it is a social phenomenon that drives the economy. Since the mid-1990s sociologists have studied the production of fashion, and looked into the conditions and effects of how fashion is produced. This research has focused on the garment industry, but in contrast to much of the earlier sociological work, consumption and social distinction have not been in the foreground. But the focus has not only been on garment production, but on the institutions, actors, and logic of fashion production – how it comes to be that some garments, models, photographers, or styles come into fashion, and others not (Entwistle, 2009). The general point being made is that fashion is a collective, though not a concerted, product, and that various producers, including, designers, models, fashion photographers, magazines, buyers, and many other coproducers of fashion, are involved in this production. Several authors have contributed to this field. One study that represents this field of research is a sociological analysis of fashion with a focus on global garment markets (Aspers, 2010). The study shows how the final consumer market for fashion garments in Western Europe is intimately tied to the markets for production of these garments, located in countries where the cost of production is low. Others have studied fashion magazines (Moeran, 2006) and what they do to affect the current fashion and fashion models. The fashion industry is one of the largest in the world, and the economic side of fashion, however, is not only restricted to garments.

Fashion's Cycles

Some researchers have been less interested in fashion's social dynamic than in an in-depth descriptive study of its cycles. Richardson and Kroeber (1940) use empirical data to identify six different formal aspects in women's dress taking, for example, the changes in length and width of skirts over 300 years. A number of researchers have followed Kroeber's empirical tradition, in which specific changes in fashion are seen to occur within a continuous cycle of longer periods. For example, Robinson (1976) studied beard fashions between 1842 and 1972 as time series analyses. To understand the elapse between the reappearance of the same fashion, Robinson argues that certain ground rules apply for fashion's cyclical changes, one being that as long as a number of people still follow a certain fashion, it prevents younger people from adopting it.

Fashion in this perspective is not born, but is rediscovered. Cross-cyclical changes receive particular attention as demonstrating sharp reactions to extreme styles. For example, the streamlined form's soft curves are contrasted with the ensuing period's particularly angular design. Cycles are not only restricted to garments. Stanley Lieberson (2000) has shown that personal names have a strong cyclical pattern.

Ambivalence Toward Fashion: The Vitally Unimportant

Fashion is dubious, and especially economists, for example, Pesendorfer (1995) have often criticized fashion for being the primary example of irrationality – and thus something that should be abolished. Fashion's symbolic significance has been contrasted with its lack of practicality. This has led many today to express ambivalence toward fashion, and there is a tendency to legitimize it using functional arguments; the rational and practical is acceptable where the impractical is not. Function and symbolism can sometimes be reconciled in fashion, however. Aerodynamics has been described as a discovery of great significance for industrial design, and for a time all household appliances were streamlined. 'Streamlining' symbolized America's enlightened practice of throwing over old conventions to create a new and dynamic civilization. Those who chose, for example, a streamlined kettle, demonstrated that they also chose the rational, the future, and the practical.

Fashion as Unpredictable

Increasingly, fashion's analysts, as stated by Aspers and Godart (2013) have come to emphasize the inner dynamic in fashion. Individual theories are run together to reflect fashion's inherent contradictions: it is coercion *and* individual choice; it is social order *and* a break with social order; it is union *and* differentiation. The nature of fashion is seen as ever more complex.

There is a more simplistic perspective, however, common to those who base their interpretation of fashion on industrial products. Here one finds phrases such as 'the most thorough logic of planned aging,' and 'industry is based on extolling a perpetual neology.' Fashion, however, is to be found everywhere, not only in areas that are subject to commercial planning, while the aging of fashion is impossible to predict precisely, however much industry demands precise guidelines. The problems stem from fashion's most central characteristic: because it symbolizes 'now,' it can never be foreseen with any certainty. Fashion pinpoints what is right in the present, but the moment it can be predicted accurately, it is no longer a matter of fashion. The transient symbols of what is right for now can never be nailed down in advance.

See also: Art and Culture, Economics of; Collective Action; Collective Behavior, Sociology of; Consumer Psychology; Consumption, Sociology of; Culture, Sociology of; Differentiation: Social; Fashion, Sociology of; Hegemony and Cultural Resistance; Simmel, Georg (1858–1918); Social Identity in Sociology; Values, Sociology of.

Bibliography

Alberoni, F., 1967. Consumi e società. Società editrice il Mulino, Bologna, Italy.

Aspers, P., 2010. Orderly Fashion, A Sociology of Markets. Princeton University Press, Princeton, NJ.

Aspers, P., Godart, F., 2013. Sociology of fashion: order and change. Annual Review of Sociology 39, 171–192.

Barber, B., Lobel, L.S., 1953. Fashion in women's clothes and the American social system. In: Bendix, R., Lipset, S.M. (Eds.), Class, Status and Power. Free Press, Glencoe, IL, pp. 323–332.

Bell, Q., 1976. On Human Finery. Hogarth Press, London.

Blumer, H., 1968. Fashion. In: Sills, D. (Ed.), International Encyclopedia of the Social Sciences, vol. 5. The Macmillan Company, London, pp. 341–345.

Blumer, H., 1969. Fashion: from class differentiation to collective selection. Sociological Quarterly 10, 275.

Campbell, C., 1992. The desire for the new. Its nature and social location as presented in theories of fashion and modern consumerism. In: Silverstone, R., Hirsch, E. (Eds.), Consuming Technologies: Media and Information in Domestic Spaces. Routledge, London, pp. 48–66.

Davis, F., 1992. Fashion, Culture and Identity. University of Chicago Press, Chicago, IL.

Emberley, J., 1987. The fashion apparatus and the deconstruction of post-modern subjectivity. Canadian Journal of Political Social and Theory 11, 38.

Entwistle, J., 2009. The Aesthetic Economy of Fashion: Markets and Value in Clothing and Modelling. Berg, Oxford.

Hodkinson, P., 2002. Goth, Identity, Style and Subculture. Berg, Oxford.

Kawamura, Y., 2004. The Japanese Revolution in Paris Fashion. Berg, Oxford.

König, R., 1973. The Restless Image: A Sociology of Fashion. Allen & Unwin, London.

König, R., 1974. A la Mode: On the Social Psychology of Fashion. Seabury Press, New York.

Lieberson, S., 2000. A Matter of Taste: How Names, Fashions, and Culture Change. Yale University Press, Yale.

Moeran, B., 2006. More than just a fashion magazine. Current Sociology 54 (5), 725–744.

Nedelmann, B., 1990. Georg Simmel as an analyst of autonomous processes: the merry-go-round of fashion. In: Kaern, M., Phillips, B.S., Cohen, R.S. (Eds.), Georg Simmel and Contemporary Sociology. Kluwer, Dordrecht, The Netherlands.

Pesendorfer, W., 1995. Design innovation and fashion cycles. The American Economic Review 85 (4), 771–792.

Richardson, J., Kroeber, A.L., 1940. Three Centuries of Women's Dress Fashions: A Quantitative Analysis. University of California, Berkeley, CA.

Robinson, D.E., 1976. Fashions in shaving and trimming of the beard: the men of the illustrated London news, 1842–1972. American Journal of Sociology 81, 1133–1141.

Sellerberg, A.-M., 1994. A Blend of Contradictions: Georg Simmel in Theory and Practice. Transaction Publishers, New Brunswick, NJ.

Simmel, G., 1904. Fashion. Fashion International Quarterly 10, 130–155.

Tarde, G., 1902. Psychologie Economique. F. Alcan, Paris.

Tarde, G., 1903. Laws of Imitation (Trans. from the 2nd French Edition, Parsons E.C.). In: Custom and Fashion. Holt, New York (Chapter vii).

Veblen, T.B., 1899. The Theory of the Leisure Class: An Economic Study in the Evolution of Institutions. MacMillan, New York.

Wilson, E., 1985. Adorned in Dreams. Virago, London.

Fatherhood

Sandra Hofferth and Frances Goldscheider, University of Maryland, College Park, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Interest in understanding fatherhood has grown in industrialized countries as men have become more involved in rearing children. This appears to be a continuation of the gender revolution that was seen as stalled because, as women reduced their household work, men did not appear to take up the slack. In contrast to other forms of work at home, involvement with children has high rewards. Men's increased involvement will contribute to improving family life in the future because of the benefits it can bring to themselves and their families. This brief article summarizes some of the current knowledge and issues in fatherhood research.

Interest in understanding fatherhood has grown enormously in industrialized countries, starting in the final decades of the twentieth century, with a vast increase in the numbers of publications and a new journal on the topic (*Fathering*). In part, this reflects the growing complexity of men's fatherhood experiences as union dissolution became more common, increasing the proportions of absent fathers and stepfathers, as well as the small numbers of male custodial parents (Goldscheider and Sassler, 2006). Also, of course, the interest in understanding fatherhood reflects men's increased involvement with their coresidential, biological children (Furstenberg, 1988; Hofferth et al., 2013) as well as their noncoresidential children (Amato et al., 2009), together with the recognition that this is important for both children and fathers (Pleck, 2007; Eggebeen, 2001).

The Gender Revolution and Changes in Fatherhood and Motherhood

At the broadest historical level, the increase in involved fatherhood and the growth in scholarly interest reflect the 'second half' of the gender revolution in the more industrialized countries (Goldscheider, 2012). In the first half, once thought of as the only gender revolution, the extremes of the separate spheres of the 1950s and women increasingly joined men in the public sphere, the labor market. Given their increased education and potential earnings and the stagnating earnings of men, they could better provide for their families' needs in the workforce than they could by remaining in the home. The second half, however, which many find to be 'incomplete' (Esping-Andersen, 2009), 'unfinished' (Gerson, 2010), or even 'stalled' (Hochschild and Machung, 1989; England, 2010) is now underway across the industrialized world, and consists in men's increasingly participating in the private sphere of the home (Kan et al., 2011).

Most studies of men's increased involvement in the tasks of the home that became difficult for women to manage as their employment levels increased, with longer hours and younger children, have focused on housework, with child care often ignored (Sullivan, 2013). However, as many have recognized, time with children is experienced as far more personally rewarding than other household tasks (Krueger et al., 2009). Most people also find housework easier to outsource or forego than child care, hence the growth in meals away from home and carry-out restaurant food, as well as the decline in food preparation and overall housework time (Raley et al., 2012). In contrast, people attempt to minimize outsourcing child care to the hours both parents are at work, pursuing active parenting to the extent possible; fathers now spend as many or more hours with children as mothers on weekends (Yeung et al., 2001).

This trend toward men's greater involvement in the tasks of the home, now advancing in all industrialized countries, has been strengthening families, reducing union dissolution (Oláh and Gähler, 2014) and 'lowest-low' fertility (Kaufman, 2000; Torr and Short, 2004; Duvander and Andersson, 2006). However, the many decades that separated women's increased entry into the public sphere and men's increased involvement in the private sphere were difficult for families, reducing fertility (Waite and Stolzenberg, 1976) and increasing union dissolution (Ruggles, 1997). As a result, men and women have become less committed; they exit relationships more easily and increasingly choose to cohabit rather than marry, even as they pursue parenthood (Cherlin, 2010). Our view, that the second half of the gender revolution will lead to increasingly stable families with more children, differs from the more prevalent theoretical approach, that of the second demographic transition (SDT), which predicts continued family weakness, with high levels of nonmarriage, nonparenthood, and cohabitation (e.g., Lesthaeghe, 2010). Another change over the past decade that may be strengthening committed family relationships, particularly for children, is the increased acceptance and now legitimation in many states of same sex relationships and the increased proportion of these new couples who are rearing children (Williams Institute, 2013).

Nevertheless, both the SDT and gender revolution approaches recognize that an important consequence of the increase in female labor force participation was a sea change in the relationships between men and women that upset the low divorce, high fertility regime that characterized so many industrialized countries in the decades immediately after World War II. As a result, too many children have had to deal with their parents' unstable relationships, with absent fathers (and increasingly absent mothers), and with their custodial parents' new coresidential relationships.

Changes in the Living Arrangements of Children and Parents, 1996–2009

The most recent changes in the coresidential arrangements of children (less than 18) and parents are shown in Table 1, revealing the continued growth between 1996 and 2009 in the proportion of children living with unmarried (and often nonbiological) parents and the decline in the proportion living with two, married biological parents. Overall, the decline in living with two parents was a relatively small 2% points (from 73.4 to 71.4%), although the decline in the proportions living with two married biological parents was larger (4.1% points). Hence, the overall decline was offset by the growth in living with parents in less traditional relationships: unmarried biological parents, which increased from 1.8 to 3.2%; a married biological parent and a stepparent, which increased from 6.7 to 7.1%; and the least traditional parental structure, two unmarried parents, only one biological, which increased from 2.5 to 2.8%.

There was much less change in the proportion of children living with a lone parent, which increased overall from 22.7 to 24%. Most of this increase was due to the increase in children living in father-only families (from 2.1 to 3%); the increase in the share of children living in lone-mother families was much smaller, from 20.6 to 21%. Finally, likely reflecting the early impact of the Great Recession of 2008, there was also an increase from 3.9 to 4.6% in the share of children who were not living with either of their parents and were likely living with grandparents or other relatives, or were in the foster care system.

It is likely, however, that these relatively small changes, mostly continuing the trend over the past several decades, would look different with a control for social class, given the increased inequality of life chances and family relationships that have characterized the United States in recent decades. Children of parents who have at least some college education are far more likely to be living with both biological parents, particularly married parents (McLanahan, 2004). Given the increased recognition of the importance of fathering for children, these changes have put the children of less educated parents at considerable risk.

Table 1 Percentage of children <18 living with parents, 1996 and 2009

Family type	*1996*	*2009*
Two-parent family	73.4	71.4
Two married, biological parents	62.4	58.3
Two unmarried, biological parents	1.8	3.2
Two married parents, one nonbiological (step)	6.7	7.1
Two unmarried parents, one nonbiological	2.5	2.8
One-parent family	22.7	24.0
Mother only	20.6	21.0
Father only	2.1	3.0
No-parent family	3.9	4.6
Total	100.0	100.0

Sources: 1996: Federal Interagency Forum on Child and Family Statistics, 2000. America's Children: Key National Indicators of Well-being. U.S. Government Printing Office, Washington, DC, 2007: Kreider, R., Ellis, K., 2011., Living Arrangements of children 2009 126 (P70). U.S. Census Bureau, Washington, DC.

The Transition to Fatherhood

Most men eventually become fathers. However, early fatherhood and less conventional fatherhood are more common among men of lower socioeconomic status or unstable family backgrounds. Men who experienced an unstable childhood family have been found to transition to biological parenthood early, particularly to nonresidential fatherhood and, among residential fathers, to cohabiting fatherhood rather than married fatherhood (Hofferth and Goldscheider, 2010).

Further, men who experienced a disrupted family structure in childhood were both more willing to become stepfathers (Goldscheider and Kaufman, 2006), and more likely to do so (Goldscheider et al., 2009). Entry into single fatherhood is also influenced by childhood family structure. Having experienced four or more parental transitions as a child was associated with a male being likely to become a single father (Hofferth and Goldscheider, 2010).

Motivation for and Barriers to Father Involvement

Father involvement encompasses three components: (1) engagement, (2) accessibility, and (3) responsibility (Lamb et al., 1985). Engagement is fathers' direct interaction with children, such as bathing, playing games, and reading a book. Accessibility refers to the ease with which the father can be reached, such as being at home in another room or contactable by phone. Responsibility is the extent to which the father is involved in managing care for the child, such as making doctor appointments, buying clothes, selecting and coordinating child care, and arranging after school activities. In addition, researchers often incorporate measures of the quality of father–child interactions or relationships (e.g., warmth responsiveness and control) in definitions of father engagement (Pleck, 2010).

What are fathers' motivations for involvement? The literature has conceptualized motivations as biological/evolutionary and psychosocial. Men are motivated to father and rear children in order to maintain the family line; this implies also that men will be more motivated to spend time with biological children, as they have the best long-term payoff in evolutionary biological terms (Waynforth, 2013). In humans, the motivation may be to improve the chances of securing a new partner, that is, a mating investment. Males of many species have been shown to parent nonbiological offspring (Hofferth and Anderson, 2003), so involved stepfathers are not unique.

Situation/structural and legal barriers to father involvement are numerous. These include inflexible employment schedules, strong gender-role stereotypes, and legal divorce settlements that limit fathering rights to biological fathers (Pleck and Masciadrelli, 2004). Research has shown that father involvement is strongly influenced by the mother, both in terms of gatekeeping and modeling, but mother involvement is less dependent on the father (Hofferth and Pleck, 2008).

Paternal Influences on Child Development

Because of the difference in access to children who are nonresidential, the literature has generally separately examined

the influence of fathers on children according to whether or not the father and child are living together. Among those who coreside, the literature classifies, first, whether or not the child is a biological child of the father and, second, whether or not the parents are legally married.

Level of achievement and behavior problems are the most commonly explored developmental outcomes associated with father involvement. Among residential fathers, more research supports the association of parenting with child behavior than with child achievement. Children living in stable two biological married parent families uniformly have better behavioral outcomes than children living in other family types. The benefits include greater economic resources, more effective parenting, and more stability. Most research suggests that fathers contribute independently, although the specific contribution often depends upon characteristics of the child and of the couple relationship. A recent study indicates that positive paternal interactions and warmth for the child reinforces maternal supportiveness of the child. Mothers' parenting is not as effective in reducing toddlers' behavior problems without paternal support (Verhoeven et al., 2010). Another recent study suggests that paternal presence is much more important in reducing child behavior problems when the child has a difficult than an easy temperament (Cabrera et al., 2014). It truly is the interaction between two well-functioning parents taking the child temperament into account that matter to child behavior. Although biological parents have advantages, the research suggests that stepfathers can be excellent partners and parents. The benefits of living with the biological father also appear to extend to unmarried biological fathers. Such fathers tend to be younger and less advantaged; controlling for these differences reduces the disadvantage to the child. Only in recent years have there been sufficient numbers of families in which to examine this association (Hofferth, 2006; Hofferth and Anderson, 2003).

Based on a meta-analysis of literature on nonresidential fathers and child outcomes up to the late 1990s, Amato and Gilbreth (1999) found that the provision of child support, greater emotional closeness, and authoritative parenting were consistently associated with greater academic success and fewer externalizing behavior problems, whereas contact was related to academic success but not to externalizing problems. Internalizing behavior was not consistently implicated for any of the involvement measures. Of course, the age of the child is important. Hawkins et al. (2007) suggest that by the time the child reaches adolescence, the involvement of nonresidential fathers may make much less of a difference; at that point, fathers respond more to the child than child to father. Because there are clear differences across study time periods and across child ages, it is time for another such careful meta-analysis to be conducted.

Discussion and Conclusion

Men's involvement in the private sphere of the family and, in particular, the expansion of their paternal roles is clearly proceeding unevenly. Children are increasingly living with parents in unstable relationships, which are less likely to provide the security children need as they are growing up. This is the case for those living with cohabiting biological fathers and those living with stepfathers. There is substantial concern that some men lack motivation or the knowledge and resources to adequately parent their children. Research indicates that parenting has intergenerational consequences; how men were parented matters to how they parent their own children (Hofferth et al., 2012). Yet there are signs of men's increasing commitment to their children, both in terms of fathers' increased involvement with both their coresidential and noncoresidential children and in the growth in children living with their fathers but not their mothers.

Nevertheless, the message from the research and reviews conducted to date on father involvement is that more than biological, marital, or actual residential structure, it is the quality of the emotional relationship between the child and the father that matters the most for the child's behavior, both in childhood and adolescence (Cabrera et al., 2012; Hofferth et al., 2013). Another important point is that parents and children have joint influences. Parents respond to children's behavior and children to the actions of parents, though it is likely that parental influence is stronger on younger than older children. Finally, children's temperament and the mother–father and mother–child relationship also matter and need to be considered. Fathers matter more under some conditions than others, in particular, when a child is less able to control his/her emotions.

Research on same sex couples, whose strengths for child-rearing are mostly unknown, is needed (but see Goldberg and Smith, 2013). Because same sex couples are eagerly marrying and adopting children, this will be an important context for children in the future. Another important area is that of immigrant families. We have not addressed racial, ethnic, generational, and cultural differences in this article. The research does not suggest that the main conclusion about the quality of the relationship mattering most would differ across groups, although attitudes and values regarding the role of the father may vary.

Further, there is little research on whether more intensive fatherhood is leading men to become even more involved in the private sphere of the family. How do varying amounts or types of father involvement affect the likelihood of providing care for sick children, and of providing a healthy environment within which children can develop to their full potential? Finally, fatherhood has been shown to benefit men as well as their families. It is clear that involved fatherhood is an important step in breaking down the separate spheres, further advancing the second half of the gender revolution.

See also: Child Care and Development across Cultures; Fathers and Fathering: Social Work Perspectives; Gender History; Gender, Economics of; Gender, Time-Use, and Fertility Recovery in Industrialized Countries; Labor, Division of; Repartnering and Stepchildren; Residential Child Care; Time Use Research: Recent Developments; Time Use and Gender.

Bibliography

Amato, P., Gilbreth, J., 1999. Nonresident fathers and children's well-being: a meta-analysis. Journal of Marriage and the Family 61, 557–573.

Amato, P., Meyers, C., Emery, R., 2009. Changes in nonresident father-child contact from 1976 to 2002. Family Relations 58, 41–53.

Cabrera, N.J., Cook, G.A., McFadden, K., Bradley, R., 2012. Father residence and father-child relationship quality: peer relationships and externalizing behavior. Family Science 2, 109–119.

Cabrera, N., Hofferth, S., Hancock, G., 2014. Family structure, maternal employment, and Children's externalizing behavior change: differences by age and self-regulation. European Journal of Developmental Psychology 11, 136–158. http://dx.doi.org/10.1080/17405629.2013.873716.

Cherlin, A., 2010. Demographic trends in the United States: a review of research in the 2000s. Journal of Marriage and Family 72, 403–419.

Duvander, A., Andersson, G., 2006. Gender equality and fertility in Sweden: a study of the impact of father's uptake of parental leave on continued childbearing. Marriage and Family Review 39 (1/2), 121–142.

Eggebeen, D., 2001. Does fatherhood matter for men? Journal of Marriage and Family 63, 381–393.

England, P., 2010. The gender revolution: uneven and stalled. Gender and Society 24, 149–166.

Esping-Andersen, G., 2009. The Incomplete Revolution: Adapting to Women's New Roles. Polity Press, Oxford, Cambridge.

Federal Interagency Forum on Child and Family Statistics, 2000. America's Children: Key National Indicators of Well-being. U.S. Government Printing Office, Washington, DC.

Furstenberg, F., 1988. Good dads–bad dads: two faces of fatherhood. In: Cherlin, A. (Ed.), The Changing American Family and Public Policy. The Urban Institute, Washington, pp. 193–218.

Gauthier, A. Social class and parental investment in children. In: Scott, R.A., Kosslyn, S. (Eds.), Emerging Trends in the Social and Behavior Sciences. Sage, in press.

Gerson, K., 2010. The Unfinished Revolution: Coming of Age in a New Era of Gender, Work, and Family. Oxford University Press, New York.

Goldberg, A., Smith, J., 2013. Predictors of psychological adjustment in early-placed adopted children with lesbian, gay, and heterosexual parents. Journal of Family Psychology 27, 431–442.

Goldscheider, F., 13 June 2012. The gender revolution and the second demographic transition: understanding recent family trends in industrialized societies. In: Keynote Address to the Bi-annual Meetings of the European Association of Population Studies, Stockholm, Sweden.

Goldscheider, F., Kaufman, G., 2006. Willingness to stepparent: attitudes towards partners who already have children. Journal of Family Issues 27, 1415–1436.

Goldscheider, F., Kaufman, G., Sassler, S., 2009. Navigating the "New" marriage market: how attitudes towards partner characteristics shape union formation. Journal of Family Issues 30, 719–737.

Goldscheider, F., Sassler, S., 2006. Creating stepfamilies: integrating children into the study of union formation. Journal of Marriage and Family 68, 275–291 erratum, number 3.

Hawkins, D.N., Amato, P.R., King, V., 2007. Nonresident father involvement and adolescent well-being. American Sociological Review 72, 990–1010.

Hochschild, A., Machung, A., 1989. The Second Shift: Working Parents and the Revolution at Home.

Hofferth, S.L., 2006. Residential father family type and child well-being: investment versus selection. Demography 43, 53–77.

Hofferth, S.L., Anderson, K., 2003. Are all dads equal? Biology vs. marriage as basis for paternal investment. Journal of Marriage and Family 65, 213–232.

Hofferth, S.L., Goldscheider, F., 2010. Family structure and the transition to parenthood. Demography 47, 415–437.

Hofferth, S.L., Pleck, J., Goldscheider, F., Curtin, S., Hrapczynski, K., 2013. Family structure and men's motivation for parenting in the United States. In: Cabrera, N., Tamis-LeMonda, K. (Eds.), Handbook of Father Involvement: Multidisciplinary Perspectives, Second ed. Routledge, New York, pp. 57–80.

Hofferth, S.L., Pleck, J., 2008. Mother involvement as an influence on father involvement with early adolescents. Fathering 6, 267–286.

Hofferth, S.L., Pleck, J., Vesely, C., 2012. The transmission of parenting from fathers to sons. Parenting: Science and Practice 12, 282–305.

Kan, M., Sullivan, O., Gershuny, J., 2011. Gender convergence in domestic work: discerning the effects of interactional and institutional barriers from large-scale data. Sociology 45, 234–251.

Kaufman, G., 2000. Do gender roles matter? Family formation and dissolution among traditional and egalitarian men and women. Journal of Family Issues 21, 128–145.

Kreider, R., Ellis, R., 2011. Living Arrangements of Children: 2009 126 (P70). U.S. Census Bureau, Washington, DC.

Krueger, Alan, Kahneman, D., Schkade, D., Schwartz, N., Stone, A., 2009. National time accounting: the currency of life. In: Krueger, A. (Ed.), Measuring the Subjective Well-being of Nations. University of Chicago Press, Chicago, pp. 9–86.

Lamb, M., Pleck, J., Charnov, E., Levine, J., 1985. Paternal behaviour in humans. American Psychologist 25, 883–894.

Lesthaeghe, R., 2010. The unfolding story of the second demographic transition. Population and Development Review 3, 211–251.

McLanahan, S., 2004. Diverging destinies: how children are faring under the second demographic transition. Demography 41, 607–627.

Oláh, L., Gähler, M., 2014. Gender equality perceptions, division of paid and unpaid work and partnership dissolution in Sweden. Social Forces. http://dx.doi.org/10.1093/sf/sou066.

Pleck, J., 2007. Why could father involvement benefit children? Theoretical perspectives. Applied Developmental Science 11, 1–7.

Pleck, J., 2010. Paternal involvement: revised conceptualization and theoretical linkages with child outcomes. In: Lamb, M. (Ed.), The Role of the Father in Child Development, fifth ed. Wiley, New York, pp. 67–107.

Pleck, J., Masciadrelli, B.P., 2004. Paternal involvement by U.S. residential fathers: levels, sources, and consequences. In: Lamb, M. (Ed.), The Role of the Father in Child Development, fourth ed. John Wiley & Sons, New York, pp. 222–271.

Raley, S., Bianchi, S., Wang, W., 2012. When do fathers care? Mothers' economic contribution and fathers' involvement in child care. American Journal of Sociology 117, 1422–1459.

Ruggles, S., 1997. The rise of divorce and separation in the United States, 1880–1990. Demography 34, 455–466.

Sullivan, O., 2013. What do we learn about gender by analyzing housework separately from childcare? Some considerations from time use evidence. Journal of Family Theory and Review 5, 72–84.

Torr, B., Short, S., 2004. Second births and the second shift: a research note on gender equity and fertility. Population and Development Review 30, 109–130.

Verhoeven, M., Junger, M., van Aken, C., Dekovic, M., van Aken, M.A., 2010. Mothering, fathering, and externalizing behavior in toddler boys. Journal of Marriage and Family 72, 307–317.

Waite, L., Stolzenberg, R., 1976. Intended fertility and female labor force participation of young women: insights from nonrecursive models. American Sociological Review 41, 235–252.

Waynforth, D., 2013. Evolutional perspectives on father involvement. In: Cabrera, N., Tamis-LeMonda, K. (Eds.), Handbook of Father Involvement: Multidisciplinary Perspectives, second ed. Routledge, New York, pp. 23–26.

Williams Institute, 2013. Percent of Same-Sex Couples Raising Children in Top Metropolitan Areas. http://williamsinstitute.law.ucla.edu/research/census-lgbt-demographics-studies/infographic-msas-may-2013/ (accessed 08.08.13.).

Yeung, W., Sandberg, J., Davis-Kean, P., Hofferth, S., 2001. Children's time with fathers in intact families. Journal of Marriage and Family 63, 136–154.

Fathers and Fathering: Social Work Perspectives

Jonathan Scourfield, Cardiff University, Cardiff, Wales, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Fathers' is a very diverse category and there is considerable variation in fathering within and between countries. Fathers have not traditionally been the focus of much attention in social work practice, but some committed practitioners are now developing innovative interventions. There is relatively little evidence about which interventions result in good outcomes, but this article attempts to provide an overview of what is known about effective practice with fathers.

Introduction

Men comprise around half the population, but only a minority of the users of social work services. For a variety of reasons, men are much less likely than women to have contact with social workers and this gender difference is perhaps especially acute in relation to child and family services. This article will provide an overview of social work with fathers. It begins by setting the social and cultural context of contemporary fathering, before moving on to critiques of social work practice, initiatives to improve the inclusion of fathers, and the evidence for the effectiveness of interventions.

In this article, the term 'fathers' is used inclusively, recognizing that there are a range of relationships to children to which the term could apply – biological fathers, adoptive fathers, stepfathers, and other less formalized social father roles. It is important to begin this article by acknowledging diversity, since as well as a range of different kinds of relationship to children, there is also a range of fathering practices and the meaning of fatherhood status is not fixed but will vary between and within cultures and locations.

Social workers can encounter men as fathers in many different roles. Child and family settings range from universal services through to casework where children are at risk of significant harm. Obviously, fathers can also have other social needs and therefore fathering can be relevant to many other social work issues such as, for example, substance misuse, mental health, offending, and disability.

There are several reasons why it is important to work with fathers. These include sharing the responsibility for children's welfare, rather than putting undue pressure on mothers, and respecting children's rights to a relationship with men who are significant to them. There is evidence from cohort studies of the importance of fathers to children's development. For example, Flouri and Buchanan's (2002) analysis of the UK National Child Development Study found that father involvement at age 16 predicted diminished psychological distress at age 33 for women. Lamb (2010), in summarizing the greatly expanded field of research on father influence, notes that the effects mothers and fathers have on children are more similar than they are different. Fathers' individual characteristics are less important in determining the influence on child development than their quality of relationship with their children. The amount of time fathers spend with children is probably less important than the quality of that contact and how it is perceived by all parties. Finally, fathers' relationships with their partners and ex-partners are also important for positive paternal influences.

There is very limited evidence that including fathers as well as mothers leads to more positive outcomes for family services. The meta-analysis of parent training evaluations by Lundahl et al. (2008) found that programs involving both mothers and fathers achieved more desirable parenting practices and more positive changes in children's behavior than programs involving only mothers. Also, Bellamy's (2009) study of child welfare cases from the US National Survey of Child and Adolescent Well-Being found that the involvement of a noncustodial parent, which was most often a biological father, was associated with reduced likelihood of children being placed into out-of-home care.

The Social and Cultural Context of Fathering

A thorough overview of the contemporary social and cultural context of fathering, with salient connections to social work, is provided by Featherstone (2009). Here, I simply summarize a few important issues. There is, of course, substantial international and intranational variation in the context of fatherhood. Differences in local economy and social class, ethnic culture, and religion are just some of the variables which can influence fathers' involvement in family life. A dominant new discourse at least in Western countries is that of intimate fatherhood (Dermott, 2008) wherein fathers are expected to forge an emotional bond with their children. This is clearly a change from previous generations where financial provision was seen as paramount. However, as Miller's (2011) British research shows, despite fathers' increasing involvement in hands-on care, it is usually mothers who continue to carry the load of '24/7 caring and thinking responsibility' (p. 171). Many European countries – and especially Nordic ones – have moved social policy in the direction of supporting work–family balance and the greater involvement of fathers in child care, although there is some way to go and tensions remain (O'Brien and Moss, 2010).

Social workers tend to intervene where there are social problems. It is inevitable, therefore, that although the fathers encountered by social workers can be resources for the care and nurturing of children, the behavior of some fathers puts women and children at risk. This problematic behavior also

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28032-X

indicates these men's own psychosocial vulnerability (Featherstone, 2009). In many countries, social workers have an important role in protecting children from maltreatment, so it is inevitable that some of their contact with fathers will be in this context. There is a large body of research on gender in relation to child maltreatment and a review of this evidence is beyond the scope of this article. Here, I simply note that the latest UK prevalence study (Radford et al., 2011) found an overall gender symmetry in the maltreatment of children under 18, but men were more likely than women to perpetrate *severe* maltreatment of children in this age range. Men were also more likely than women to maltreat young people aged 18–24. When it came to different kinds of abuse, men were more implicated in severe physical abuse and in all kinds of sexual abuse. The research by Berger et al. (2009) using the US Fragile Families and Child Wellbeing Study found that children were more likely to come into contact with child protection services in families with men not biologically related to the children and families with men who are not romantically involved with the children's mothers. Bellamy's (2009) study of US child welfare cases found that caseworkers perceived nonparental males (i.e., men who were neither biological, step-, or adoptive fathers) in households to pose a greater risk to children than parental males. Children living with nonparental males were also significantly more likely to come into out-of-home care. However, the presence of a nonparental adult male in the household was not associated with future maltreatment reports. This could be, as Bellamy acknowledges, because men putting children most at risk have left the family by the time of the later survey wave. However, as Dubowitz (2009) notes, it might suggest unhelpful practitioner stereotyping of nonparental males. This leads us into the next subsection of critiques of current social work practice.

Social Work Approaches to Fathering

The rest of the article provides an overview of social work approaches to fathering, starting with a summary of critiques of current practice and then moving on to more positive terrain with an overview of attempts to improve the engagement of fathers and the evidence base on the effectiveness of interventions.

Critiques of Current Practice

There is a broad consensus that current social work practitioners tend not to do a very good job of engaging with fathers. Initial engagement is one aspect of the social work process, but fathers also need to enroll in social work interventions and then stay involved (Gordon et al., 2012).

There are problems with men's involvement in all kinds of family services, but perhaps the difficulties are most acute in child protection, where the stakes are especially high. The barriers are many, and concern the attitudes of mothers and the reluctance of fathers themselves in many cases, as well as the roles of practitioners (Maxwell et al., 2012). It is important to understand the different levels on which barriers to father involvement operate. Gordon et al. (2012) provide a useful review and ecological analysis which separates out individual, family, service provider, program, community, and policy levels. Although Gordon et al. see the evidence base as slender, in fact a body of qualitative research with practitioners has emerged in recent years, albeit some of this research has not been published in journals but in 'gray' literature or in student dissertations. One important theme from qualitative research has been practitioner culture, which acts against the involvement of fathers in fairly complex ways. Scourfield's (2003) ethnographic research from the United Kingdom described a mixture of gendered discourses in social work office culture. Some of these discourses were in tension with each other, but the combined effect was to make the engagement of men much less likely than the more familiar practice of working with mothers. Strega et al. (2008) provide a relatively rare quantitative picture from their mixed method research in Canada. They found that social workers deemed fathers to be irrelevant to mothers and children in 50% of cases and only 50% of those fathers who were seen as an asset to children were contacted.

Where the evidence base on barriers undoubtedly is very thin is in relation to fathers' own accounts of social work contact (or the lack of it). One example is the Family Rights Group, Fathers Matter research (Ashley et al., 2006), in which fathers spoke of seeing workers as firmly pro-women and anti-men. In another study of men whose children were in the care of the state, Dominelli et al. (2011) describe fathers' own accounts as transcending the good father–bad father binary which tends to be dominant in practice (Brandon et al., 2009).

Initiatives to Improve the Engagement of Fathers

As was noted at the beginning of this article, there is a range of services for fathers from those which are universal for parents in a locality, through to highly targeted interventions for fathers whose behavior is putting children at risk. Most common would be for fathers to be involved in services for parents – i.e., mothers and/or fathers – such as informal support via a family center or attendance at a parent training course. Less common would be specific services for fathers, although there are many examples of father-only services which have been designed to combat their reluctance to come to mainstream settings (Ghate et al., 2000).

As an example of within-country variation, a recent survey I undertook on interventions with fathers across the United Kingdom (see Scourfield, 2013) found there to be large numbers of men attending very informal 'stay and play' sessions – some with mothers and some for fathers only, as well as healthy numbers of men attending universal parenting courses alongside mothers. There were also examples of targeted interventions with healthy numbers of men attending, courses in prison and programs for men whose behavior was putting children at risk. Most practitioners were offering one-off local groups, set up perhaps by individuals with a particular interest in fathers. Many of these were unstructured social support groups.

Magill-Evans et al. (2006) conducted a systematic review of interventions for fathers of young children. Although their interest was in health care and health promotion more than social work, the diversity of approaches they describe indicates the creative range of interventions which is also

found in social care. They found studies of discussion groups, infant massage, kangaroo care (infant held skin-to-skin), observation and modeling of interactions with infants, participation in a preschool program, and parent training programs. McAllister et al. (2012) took a consciously international approach to their 'critical review of best practice' and identified 35 programs which have promoted or facilitated the involvement of men with children up to 8 years of age. Some of these (e.g., Triple P parenting courses) are used extensively in many different countries, whereas others are country-specific. They found little published evidence of projects in the Global South engaging with men as fathers, although there were several examples of interventions with men (e.g., on violence prevention or sexual health), from which lessons could be learned for work with fathers.

In addition to these dedicated interventions for fathers, some men-only and some alongside mothers, there have also been attempts to improve mainstream social work with fathers, in response to the criticisms noted above. For example, the Fathers Matter projects instigated by the Family Rights Group in the United Kingdom (see, e.g., Ashley et al., 2006) have tackled statutory social work services for children and families and have been followed by a Fatherhood Institute initiative to improve the engagement of fathers in child safeguarding in England.

The politics of social work with men can be fraught (Featherstone et al., 2007), with some opposing views which seem difficult to reconcile. For example, some feminist practitioners and academics emphasize the risk to women and children that can result from men's continued involvement with children after domestic abuse (e.g., Harne, 2011). Men's rights advocates, in contrast, such as the UK organization Families Need Fathers (see http://www.fnf.org.uk/), tend to argue for keeping biological fathers involved with children's lives in almost every case. Whereas some have tried to engage violent men in behavior change through their relationships with their children, other practitioners are unhappy that such approaches do not tackle violence against women in a systematic way (Featherstone and Fraser, 2012).

Evidence of Effectiveness

The general conclusion of McAllister et al. (2012) after reviewing the international evidence on social interventions with fathers is that there is little robust evidence on outcomes. Bronte-Tinkew et al. (2007) similarly note from their review of interventions in Western countries that most studies have small sample sizes, there are few experimental studies, and few programs found to be effective have been replicated. There is a large body of evidence on the effectiveness of parenting programs, but McAllister et al. (2012) note that there is a tendency not to disaggregate findings by sex of parent.

There are some findings about parent training programs which paint an optimistic picture for father engagement. As noted earlier, Lundahl et al. (2008) have observed that programs involving fathers as well as mothers are more effective than those for mothers only. However, there is also some evidence that fathers gain less from these interventions than mothers. For example, Nowak and Heinrichs's (2008) meta-analysis of studies of Triple P, the program with the strongest international evidence base, found fathers reporting lower levels of improvement than mothers or teachers for parenting, parental well-being, and child problems. Generally, attendance at parent training is much lower for men than for women. The study by Lindsay et al. (2011) of parents attending one of five evidence-based parenting programs in England – data on gender were available for 6095 parents – found that 15% of parents attending programs were men.

Magill-Evans et al. (2006) considered interventions for parents where there were separate outcomes reported for fathers and also interventions for fathers only. They were reasonably optimistic about outcomes, noting improvement in fathers' interactions with children and perceptions of children. However, they found no evidence of impact on child development and it should be acknowledged that some of the studies they reviewed had very small samples indeed and some had no comparison group, randomized or otherwise. Philip and O'Brien (2012) recently reviewed the evidence on outcomes of interventions for fathers after separation and divorce. Like Magill-Evans et al. (2006), they included interventions for both parents as well as those for fathers only and the review encompassed a range of study designs. Most studies did not disaggregate results by sex of parent. Some positive effects were reported, however. Parental conflict reduced in 11 of the 14 interventions and in half of the interventions, there were improvements in the co-parental relationship and in children's internalization of problems (i.e., sadness and low affect). As with the review by Magill-Evans et al. (2006), it should be noted that Philip and O'Brien (2012) did not limit their review to the most robust evaluations.

A rare example of a robust evaluation is the study by Cowan et al. (2009). Couples ($n = 289$) were randomly allocated to one of three interventions: a 16-week group program for couples, the same 16-week group program for fathers only, and a one-off information session. The group interventions were seen as preventative, so high-risk families were not included. Cowan et al. describe (p. 669) the group program, which was the same for couples as for fathers only, as occupying the middle ground of the continuum between open-ended group theory and psycho-educational teaching of specific skills. Couples in the two-group programs had better outcomes than those who only had a single meeting, in terms of father engagement with children, couple's relationship quality, and children's behavior. More consistent and longer-term effects were found in those who attended the couple's group than in those where only the fathers attended. This might suggest the importance of involving fathers alongside mothers in parents' or couples' interventions where possible. Practitioners work hard to do this, but others respond to the difficulty of enrolling men in parenting interventions by setting up modified versions for fathers only or by developing a new program from scratch which they think will appeal to fathers (Scourfield, 2013). Although some criticism has been leveled at the most widely used parenting programs for not culturally adapting to the need of different ethnic communities and therefore being less successful in recruiting minorities, it should be noted that two-thirds of families taking part in the Cowan et al. (2009) study were Mexican-American (the rest were European-American).

There is very little evidence of the effectiveness of interventions with fathers to prevent child maltreatment. The systematic review by Smith et al. (2012) unearthed 15 studies, but only two of these reported father-specific outcomes. Most of the interventions they identified addressed parenting function and capacity. All except two were mostly attended by women or there was little follow-up data on men. Most of the studies only used pre–post testing of those who received the intervention, with no comparison group. Some of these limitations are beginning to sound fairly familiar.

Addressing other problem behaviors as well as parenting may hold some promise for work with fathers. For example, in the study by Lam et al. (2009), 30 fathers with alcohol problems were randomly allocated to one of three different intervention approaches: behavioral couples therapy; parent skills with behavioral couples therapy; and traditional individual-based treatment for substance misuse. Parent training with behavioral couples therapy was the approach which resulted in the greatest change in contact with child protection services, albeit with a small to medium effect size. Lam et al. acknowledge a larger study is needed to provide stronger evidence on these approaches.

Although there is a lack of robust evidence on the effectiveness of interventions with fathers, process evaluations which describe the theoretical basis of interventions and mechanisms to effect change are also useful. There are some examples of such work on interesting interventions – for example, Scott and Crooks (2007) on 'Caring Dads,' a program for men who are putting their children at risk, and McGuire (2009) on 'Family Man' a prison-based program to enhance father involvement. The next stage with such programs is to measure outcomes in a comparative study design.

Finally, it is worth noting that there is also very little evidence about initiatives to improve the engagement of men in mainstream services. The little work that has been published is summarized here. English et al. (2009) put in place a half-day training course for all practitioners in child welfare services in one US state, as well as optional additional training modules which were taken up by some staff. Pre–post testing suggested some gains in father engagement as evidenced by agency self-assessment and case file review. Scourfield et al. (2012) developed a two-day training course on the initial engagement of fathers for child protection social workers in South Wales which included a day of skills training, via an introduction to motivational interviewing. They found a significant increase in trainees' self-efficacy for work with fathers and also in their self-reported engagement of some fathers on caseloads. There was a large self-reported increase in the engagement of nonresidential fathers and a smaller but significant increase in the engagement of nonrisky men living with children. There was, however, no change in the engagement of men who are putting children at risk. Both studies, English et al. (2009) and Scourfield et al. (2012), only used pre–post testing and no comparison group.

Conclusion

There has been a modest flourishing of practice developments in this new social work field in recent years. These developments do not always come from those who call themselves social workers, but the scope of this article has included all social interventions and a broad definition of 'social work' has been used. The UK survey I have undertaken (see Scourfield, 2013) shows great practitioner creativity, especially in universal family support programs. Attempts to get men involved with social or parenting support have included the use of activities traditionally associated with masculinity, such as sport and Lego (a building toy), and also less traditional activities such as art and cooking. It seems, however, that many of these creative interventions may rely on the dedication of individual practitioners. Although in some areas of practice there has been support at a policy level for more engagement of fathers – for example, the Sure Start early childhood initiative in the United Kingdom – there is much work to be done in getting the importance of work with fathers explicitly recognized in policies. Unless major public and third-sector organizations make father work a policy priority, it is likely to proceed in a piecemeal fashion and remain on the margins of social work.

There will always be debate about appropriate methods in this field. It has already been noted that the politics of fatherhood are controversial. Practitioners and commentators also disagree about whether or not it is right to use traditional aspects of masculinity as a way to engaging men in services (see Featherstone et al., 2007). Most would see the sense of some pragmatism in this regard and there is room for the kind of experimentation mentioned earlier. Given the diversity of fathers and fathering, a range of different intervention approaches is needed. However, it is also important to establish the effectiveness of interventions, even if those that are found to be effective should not be regarded as universally applicable. It is important now that we add to the existing process evaluations and pre–post testing of intervention participants by setting up experimental and quasiexperimental studies which compare the outcomes of the intervention with outcomes for normal services. There is particular need perhaps for robust evaluation of those interventions where the stakes are highest, namely those for fathers whose behavior is putting children at risk.

I end this article by reproducing the 10 'characteristics of effective fatherhood programs' identified by Bronte-Tinkew et al. (2007: pp. 10–12) on the basis of their review of evaluated programs in Western countries. In identifying these characteristics, these authors are drawing on the wider evidence base on effective practice in social interventions. Effective programs are those which incorporated teaching methods and materials appropriate for fathers and the cultures of the populations served; selected teachers or leaders who believed in the program they were implementing and then provided them with training; had a higher staff–participant ratio; used targeted curricula; used theoretical approaches that have been effective in influencing parenting behaviors in other contexts; employed a variety of teaching methods designed to focus on the fathers as individuals, and thereby personalized the information; lasted a sufficient amount of time to complete important core activities adequately; had staff who engaged in one-on-one relationships with fathers; encouraged the use of an incentive to fathers and/or their families; and replicated their curricula with fidelity.

These criteria do not fit all intervention approaches. They do not, for example, sit comfortably with father-led community development approaches. However, they are a useful starting point for discussion about what makes an effective approach to social work with fathers.

See also: Afrocentric Approaches to Social Work; Child Protection: International Issues; Children and Families in Social Work; Gendered Care and Social Work Practice; Masculinity and Social Work Practice; Motherhood; Parenting Attitudes and Beliefs across Cultures.

Bibliography

Ashley, C., Featherstone, B., Roskill, C., Ryan, M., White, S., 2006. Fathers Matter: Research Findings on Fathers and Their Involvement with Social Care Services. Family Rights Group, London.

Bellamy, J., 2009. A national study of male involvement among families in contact with the child welfare system. Child Maltreatment 14 (3), 255–262.

Berger, L.M., Paxson, C., Waldfogel, J., 2009. Mothers, men, and child protective services involvement. Child Maltreatment 14 (3), 263–276.

Brandon, M., Bailey, S., Belderson, P., Gardner, R., Sidebotham, P., Dodsworth, J., Warren, J., Black, J., 2009. Understanding Serious Case Reviews and Their Impact: A Biennial Analysis of Serious Case Reviews 2005–7, Research Report DCSF-rr129. Department for Children, Schools and Families, London.

Bronte-Tinkew, J., Carrano, J., Allen, T., Bowie, L., Mbawa, K., Matthews, G., 2007. Elements of Promising Practice for Fatherhood Programs: Evidence-based Research Findings on Programs for Fathers. National Responsible Fatherhood Clearinghouse, Gaithersburg, MD.

Cowan, P.A., Cowan, C.P., Pruett, M.K., Pruett, K., Wong, J.J., 2009. Promoting fathers' engagement with children: preventive interventions for low-income families. Journal of Marriage and Family 71 (3), 663–677.

Dermott, E., 2008. Intimate Fatherhood. A Sociological Analysis. Routledge, Abingdon.

Dominelli, L., Strega, S., Walmsley, C., Callahan, M., Brown, L., 2011. 'Here's my story': fathers of 'looked after' children recount their experiences in the Canadian child welfare system. British Journal of Social Work 41 (2), 351–367.

Dubowitz, H., 2009. Commentary on fathers and children and maltreatment: relationships matter most. Child Maltreatment 14 (3), 291–293.

English, D.J., Brummel, S., Martens, P., 2009. Fatherhood in the child welfare system: evaluation of a pilot project to improve father involvement. Journal of Public Child Welfare 3 (3), 213–234.

Featherstone, B., 2009. Contemporary Fathering: Theory, Policy and Practice. Policy Press, Bristol.

Featherstone, B., Rivett, M., Scourfield, J., 2007. Working with Men in Health and Social Care. Sage, London.

Featherstone, B., Fraser, C., 2012. Working with fathers around domestic violence: contemporary debates. Child Abuse Review 21 (4), 255–263.

Flouri, E., Buchanan, A., 2002. The role of father involvement in children's later mental health. Journal of Adolescence 26, 63–78.

Ghate, D., Shaw, C., Hazel, N., 2000. Engaging Fathers in Preventive Services: Fathers and Family Centres. Joseph Rowntree Foundation.

Gordon, D.M., Oliveros, A., Hawes, S.W., Iwamoto, D.K., Rayford, B.S., 2012. Engaging fathers in child protection services: a review of factors and strategies across ecological systems. Children and Youth Services Review 34 (8), 1399–1417.

Harne, L., 2011. Violent Fathering and the Risks to Children. Policy Press, Bristol.

Lam, W.K.K., Fals-Stewart, W., Kelley, M.L., 2009. Effects on substance use, parental relationship, parenting, and CPS involvement. Child Maltreatment 14 (3), 243–254.

Lamb, M.E., 2010. How do fathers influence children's development? Let me count the ways. In: Lamb, M.E. (Ed.), The Role of the Father in Child Development, fifth ed. Wiley, Hoboken, NJ, pp. 1–26.

Lindsay, G., Strand, S., Cullen, M.A., Cullen, S., Band, S., Davis, H., Conlon, G., Barlow, J., Evans, R., 2011. Parenting Early Intervention Program Evaluation. Department for Education, London. Research Report DFE-RR121(a).

Lundahl, B.W., Tollefson, D., Risser, H., Lovejoy, M.C., 2008. A meta-analysis of father involvement in parent training. Research on Social Work Practice 18, 97–106.

Magill-Evans, J., Harrison, M.J., Rempel, G., Slater, L., 2006. Interventions with fathers of young children: systematic literature review. Journal of Advanced Nursing 55 (2), 248–264.

Maxwell, N., Scourfield, J., Featherstone, B., Holland, S., Tolman, R., 2012. Engaging fathers in child welfare services: a narrative review of recent research evidence. Child & Family Social Work 17 (2), 160–169.

McAllister, F., Burgess, A., Kato, J., Barker, G., 2012. Fatherhood: Parenting Programs and Policy - a Critical Review of Best Practice. Fatherhood Institute/Promundo/MenCare, London/Washington, DC.

McGuire, J., 2009. Family Man. An Outline of the Theoretical Basis of the Program. Division of Clinical Psychology, University of Liverpool, Liverpool.

Miller, T., 2011. Making Sense of Fatherhood: Gender, Caring and Work. Cambridge University Press, Cambridge.

Nowak, C., Heinrichs, N., 2008. A comprehensive meta-analysis of Triple P positive parenting program using hierarchical linear modeling: effectiveness and moderating variables. Clinical Child and Family Psychology Review 11, 114–144.

O'Brien, M., Moss, P., 2010. Fathers, work and family policies in Europe. In: Lamb, M.E. (Ed.), The Role of the Father in Child Development, fifth ed. Wiley, Hoboken, NJ, pp. 551–577.

Philip, G., O'Brien, 2012. Supporting Fathers after Separation or Divorce: Evidence and Insights. University of East Anglia Centre for Research on the Child and Family, Norwich.

Radford, L., Corral, S., Bradley, C., Fisher, H., Bassett, C., Howat, N., Collishaw, S., 2011. Child Abuse and Neglect in the UK Today. National Society for the Prevention of Cruelty to Children, London.

Scott, K.S., Crooks, C.V., 2007. Preliminary evaluation of an intervention program for maltreating fathers. Brief Treatment and Crisis Intervention 7 (3), 224–238.

Scourfield, J., 2003. Gender and Child Protection. Palgrave Macmillan, London.

Scourfield, J. 2013. Working with fathers - what's going on out there? Survey report at http://workingwithfathers.weebly.com/practitioners-survey.html.

Scourfield, J., Tolman, R., Maxwell, N., Holland, S., Bullock, A., Sloan, L., 2012. Results of a training course for social workers on engaging fathers in child protection. Children and Youth Services Review 34 (8), 1425–1432.

Smith, T.K., Duggan, A., Bair-Merritt, M.H., Cox, G., 2012. Systematic review of fathers' involvement in programs for the primary prevention of child maltreatment. Child Abuse Review 21 (4), 237–254.

Strega, S., Fleet, C., Brown, L., Dominelli, L., Callahan, M., Walmsley, C., 2008. Connecting father absence and mother blame in child welfare policies and practice. Children and Youth Services Review 30 (7), 705–716.

Relevant Websites

http://www.fatherhoodinstitute.org/ – Fatherhood Institute (UK) (accessed 14.06.13.).

http://www.fatherhood.gov/about-us/nrfc-resources/nrfc-promising-practices – National Responsible Fatherhood Clearing House (USA) promising practice reports (accessed 14.06.13.).

http://workingwithfathers.weebly.com/ – Working with fathers to improve children's well-being (the author's own site) (accessed 14.06.13.).

Fear Conditioning: Overview

Herta Flor, Heidelberg University, Mannheim, Germany
Niels Birbaumer, University of Tübingen, Tübingen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Fear conditioning refers to the pairing of an initially neutral stimulus with an aversive fear-eliciting stimulus. The conditioned fear response is described in terms of subjective, behavioral, and physiological responses. The psychobiological basis of fear conditioning in animals has been well established, and the amygdala has been identified as a central structure. Genetic variables as well as cognitive factors also contribute to fear conditioning. The significance of fear conditioning for the understanding of many mental disorders has been investigated with a focus on extinction learning and the consolidation and reconsolidation of extinguished memories.

Basic Terms

The pairing of an initially neutral stimulus (the conditioned stimulus, CS) with a biologically relevant stimulus (the unconditioned stimulus, US) comes to elicit a response (conditioned response, CR) that is usually, but not always, similar to the response previously associated with the unconditioned stimulus (the unconditioned response, UR). In fear conditioning, the US is an aversive fear-eliciting stimulus such as painful electric shock or loud noise, the CS is a neutral tone or light stimulus. The UR and the CR consist of changes on the subjective, behavioral and physiological level and include (in humans) enhanced subjective fear and (in animals) freezing, but also changes in heart rate and skin conductance, the release of stress hormones, altered pain sensitivity, and startle reflex potentiation.

The development of the CR is based on the formation of an association between a neutral stimulus and a stimulus with innate biological significance that then predicts the US (Rescorla, 1988). Most studies involving fear conditioning have used cue conditioning, i.e., discrete CSs such as tones or graphic displays were presented. Usually delay conditioning, where the CS overlaps with and often terminates with the US, has been used rather than trace conditioning, where the CS and the US are separated in time. In addition to cue conditioning, context conditioning has been studied, where the CS is not a discrete cue but an internal or external context (Maren et al., 2013) such as a room or a certain mood state (**Figure 1**).

Fear can be viewed as a specific reaction to threatening stimuli. It can turn into an anxiety disorder when the fear becomes disproportionate to the stimulus that elicits it or when fear is experienced in inappropriate situations. Cue conditioning is thought to elicit fear, whereas context conditioning has been associated with anxiety. Due to the unpredictable nature of the US, anxiety rather than fear seems to be the main behavioral correlate.

Mechanisms of Fear Conditioning in Animals

The psychobiological basis of fear conditioning in animals, especially rodents, has been well established (LeDoux, 2000, 2014). When a tone is used as CS and electric shock as US, the auditory CS reaches the lateral nucleus of the amygdala from the thalamus or the cortex, probably, depending on the complexity of the CS. When contextual stimuli are employed as CS, the hippocampus is also involved and transmits the signal to the basolateral and the accessory basolateral nucleus of the amygdala. The nociceptive US also reaches the lateral nucleus of the amygdala via thalamic and cortical pathways. Conditioning seems to require an intact lateral nucleus of the amygdala when cue conditioning is involved and an intact basolateral and accessory basolateral nucleus when context conditioning is involved. The central nucleus of the amygdala receives input from the lateral and basolateral nuclei, and it projects to brainstem nuclei that initiate the expression of peripheral fear responses. While the lateral or basolateral nuclei are thus important for the acquisition of the CR, the central nucleus is essential for the expression of the CR (see **Figure 2**). Further processes that occur are changes in synaptic strength as well as altered gene expression and protein synthesis, which results in long-term memories (Johansen et al., 2011).

The initial site of plasticity seems to be the lateral nucleus of the amygdala involving both excitatory and inhibitory circuits (Ehrlich et al., 2009); however, thalamic and cortical reorganization has also been shown during fear conditioning (cf Marek et al., 2013). The amygdala plays an important role in the formation of the memory trace, but memory traces are also established elsewhere, especially for long-term memory storage. If the amygdala is inactivated during fear conditioning, no CR is formed. If the inactivation occurs after conditioning, then the memory trace is not abolished (Pape and Pare, 2010).

Fear Conditioning in Humans

In humans, the role of the amygdala is less clear than in animals. In contrast to rodents, humans possess a much larger prefrontal cortex and it is likely that it takes over some of the functions that are related to the amygdala in rodents. Damage to the amygdala in humans, has, however, also been related to deficits in fear conditioning, whereas lesions in the hippocampus have been shown to interfere with declarative learning

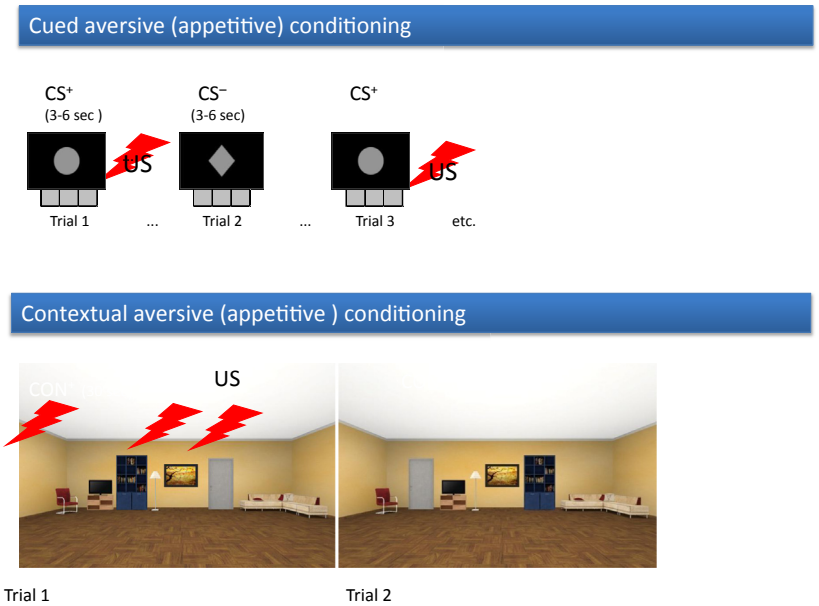

Figure 1 Cued fear conditioning as depicted in the first row involves discrete neutral cues such as circles or diamonds that serve as conditioned stimuli (CS) and unconditioned stimuli (US) such as painful electrical stimulation. The CS precedes the US in a lawful manner and usually coterminates with the US. In differential fear conditioning, one CS (the CS+) is paired with the US and serves as danger signal, whereas another CS (the CS−) is never paired with the US and serves as safety signal. In contextual cue conditioning, a context such as a room serves as CS and the US is unpredictable because there is no fixed time between the onset of the CS and the US.

and contextual fear conditioning (cf Bechara et al., 1995). Most studies using functional magnetic resonance imaging and differential conditioning paradigms in humans found amygdala activation, which was positively correlated with activations in the primary sensory cortices, the prefrontal cortex, and the thalamus (cf Sehlmeyer et al., 2009). The notion that a direct pathway from the thalamus to the amygdala also exists in humans was supported by showing that activation of the right amygdala in a masked conditioning paradigm was associated with activity in the superior colliculus and pulvinar, but not with activity at cortical sites (Morris et al., 1999). In addition to the amygdala, the acquisition of a fear response creates activation in the cingulate gyrus and the premotor cortex. Activation of the insulae has also been reported when painful stimuli were used as US. When trace or context conditioning was employed, activation in the hippocampus in addition to activation in the amygdala and cortical areas was noted (Maren et al., 2013).

Based on the observation that not all stimuli are equally easy to condition, Seligman (1971) suggested that phylogenetically relevant stimuli such as snakes or spiders might serve as 'prepared stimuli' and be more conditionable than others. Subsequently, Öhman (for a review see Öhman et al., 2000) has shown that conditioned skin conductance responses to a prepared CS such as an angry face or a picture of a snake persist for a longer time (even when the presentation is masked) than when a neutral stimulus such as a picture of a mushroom is used as CS and combined with electric shock as US. Moreover, masked presentations of these stimuli also lead to the acquisition of conditioned responses suggesting that unconscious learning of these stimulus associations is possible.

More recent accounts of fear conditioning have emphasized the role of extinction and extinction memory (see below), as well as cognitive variables of fear including expectancy and prediction error. Prediction error refers to the circumstance that an obtained outcome differs from a predicted outcome, in fear learning, for example, the probability of the occurrence or nonoccurrence of the aversive US neurons in several brain structures, including the amygdala, code prediction, and violations of prediction. Computational models of fear conditioning have shown that prediction error coding is a fundamental principle of fear learning (McNally et al., 2011).

Extinction and Extinction Memory

Extinction and extinction memory have increasingly become a focus of research, since these processes are thought to be seminal for the understanding of learning but also of anxiety disorders. Extinction occurs when the CS is repeatedly presented without the US, leading to a gradual reduction of the conditioned fear response. While pavlovian fear conditioning may be evolutionarily adaptive and should thus be strong in all persons, extinction of fear is greatly context-dependent and may show larger interindividual variability. Extinction is not only the inhibition of fear, but also a new learning process, where a new association between the CS and the absence of the US is formed (Schiller and Delgado, 2010). Extinction has been separated in early and late stages, with the late stage (after 6 h) being related to protein synthesis, and thus more stable memories. The infralimbic cortex in rodents and its medial prefrontal counterpart in humans play a major role in extinction; hypoactivation or reduced cortical thickness in these

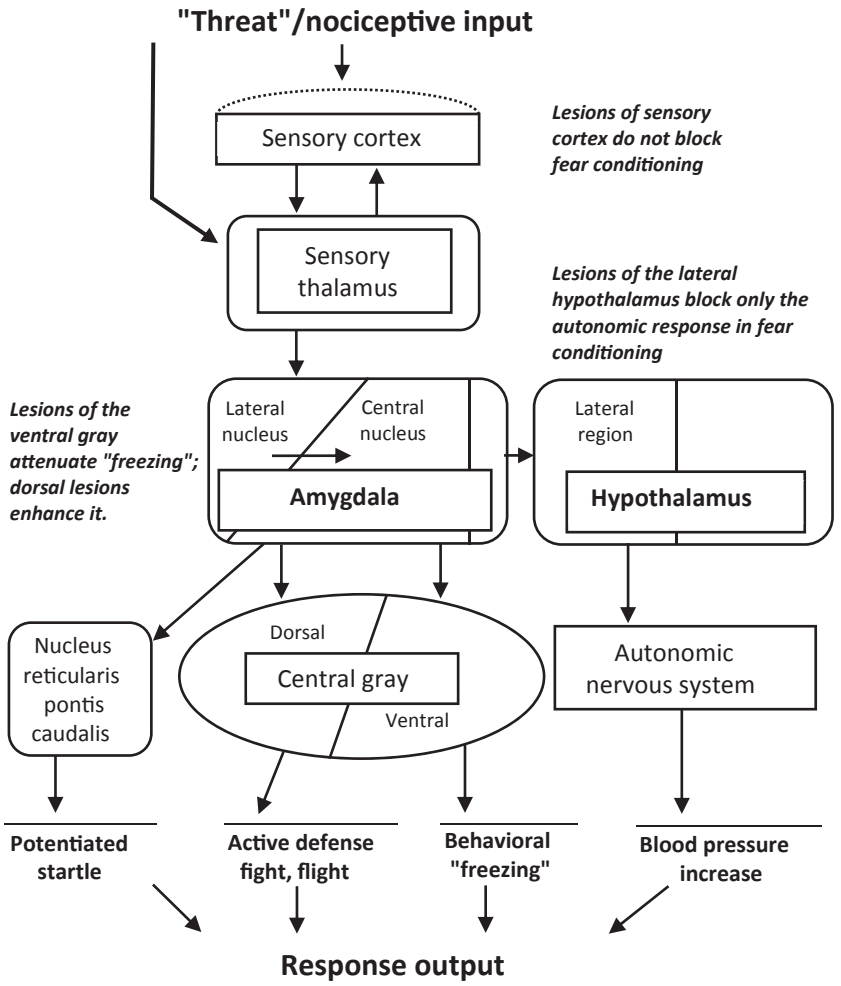

Figure 2 Fear stimuli proceed from sense receptor systems to the sensory cortex and/or sensory thalamus and then to the amygdala. There are three important connections efferent to the amygdala: (a) a projection from the central amygdala to the lateral hypothalamic area that mediates the autonomic components of the emotional response; (b) projections to the midbrain central gray region mediates defensive freezing and escape behavior and (c) a direct projection to the nucleus reticularis pontis caudalis modulates the startle reflex. Based on Bradley, M.M., Lang, P., 2000. Measuring emotion: behavior, feeling and physiology. In: Lane, R.D., Nadel, L. (Eds.), Cognitive Neuroscience of Emotion. Oxford University Press, New York.

regions interferes with extinction, and stress-induced impairments of extinction are related to atrophy and a loss of GABAergic inhibition in the infralimbic cortex (cf Akirav and Maroun, 2007). There is also evidence that the prelimbic cortex (or dorsal anterior cingulate in humans) is hyperactive when extinction is deficient. The prefrontal cortex inhibits fear via inhibition of the amygdala, which has specialized extinction encoding neurons, and its disruption may be important for the lack of fear extinction such as in post-traumatic stress disorder (PTSD). Extinction is dependent on the *N*-methyl-D-aspartate (NMDA) receptor. Several phenomena that disrupt extinction have been described. Recovery of fear refers to the process that fear may be spontaneously reactivated and extinction may be forgotten with the passage of time. Renewal refers to the fact that an extinguished memory of fear becomes active when a change of context occurs (Vervliet et al., 2013). Finally, reinstatement refers to the return of fear when the unconditioned stimulus is presented again. The hippocampus plays an important role in extinction and extinction memory related to the fact that extinction but not acquisition is context-dependent (Bouton et al., 2006). Extinction can be facilitated by the use of pharmacological enhancers. For example, the partial NMDA receptor antagonist D-cycloserine has been employed to speed up the extinction process in both animals and humans; however, the result has been equivocal. Recent accounts of extinction have also emphasized the similarity of extinction learning and reward processing, and this may led to new insights about the mechanisms of the extinction process.

Learning results in the consolidation of fear memories, but these memories can become labile when they are recalled and can then be enhanced or diminished in this process of 'reconsolidation.' Reactivation can occur by the presentation of a memory associated environmental cue (such as a fear conditioned CS). Monfils et al. (2009) and Schiller et al. (2010) have shown that a previously conditioned CS presented early

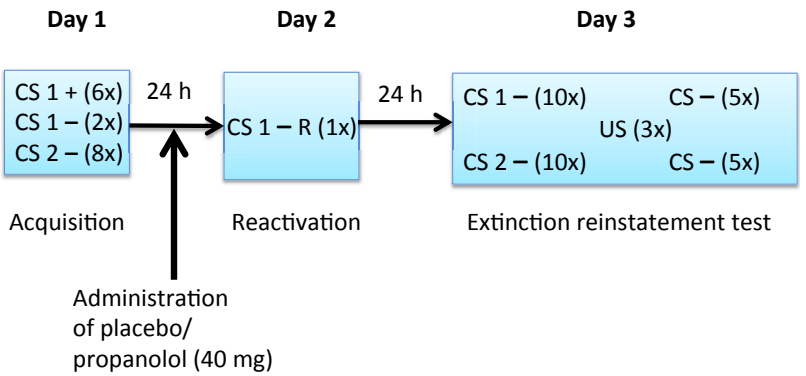

Figure 3 Design of the study that used the lability of the fear trace in the reconsolidation window by administering propanolol before the memory was reactivated by the presentation of a conditioned stimulus that reactivated the fear. When later extinction and reinstatement were tested, the disruption of fear reconsolidation by propanolol led to better extinction and less reinstatement. Based on Kindt, M., Soeter, M., Vervliet, B., 2009. Beyond extinction: erasing human fear responses and preventing the return of fear. Nature Neuroscience 12, 256–258.

but not late (>6 h) after acquisition and prior to extinction can eliminate the spontaneous recovery of fear, reinstatement, renewal, and fast reacquisition. The idea is that the CS presentation reactivates the memory and makes the fear-memory trace labile in a reconsolidation window that lasts up to 6 h, so that it can easily be overwritten by extinction. This could have important implications for the treatment of fear and anxiety, because one could make use of this reconsolidation window to develop better treatments of anxiety disorders. For example, Kindt et al. (2009) showed that the use of propranolol shortly before the reactivation of a fear memory erased the later expression of fear and also prevented its reactivation (see Figure 3).

Genetic and Epigenetic Factors

Genetic factors seem to play a role in the development in anxiety disorders in humans. Familial studies, twin studies, and adoption studies show genetic variances of 20–40%. The same was found for fear conditioning. Fear conditioning ability and predisposition to learn fear responses were extensively studied in animals. In rodents, fear processes or anxiety measured in the open field test or avoidance conditioning have revealed a strong genetic basis explaining most of the interindividual variance. However, the genes responsible for that trait have only partially been identified. Important genes that modulate fear conditioning involve the serotonin transporter (5HTT) and the catechol-*O*-methyltransferase (COMT) gene. Moreover, genes involved in the regulation of the hypothalamus–pituitary–adrenal axis such as the glucocorticoid receptor gene or the corticotropin-releasing hormone genes, but also the dopamine transporter (DAT1) and the pituitary adenylate cyclase 1 receptor (ADCYAP1R1) genes have been implicated (Lonsdorf and Kalisch, 2011). Controversial data exist for the brain-derived neurotrophic factor (BDNF) gene. There is also mounting evidence that epigenetic mechanisms such as DNA methylation and histone modification contribute to fear conditioning (Zovkic and Sweatt, 2013). For example, histone acetylation and DNA methylation were found to be crucial for fear memory consolidation and synaptic plasticity in the lateral amygdala. Dias and Ressler (2014) showed that prenatal fear conditioning odors led to enhanced startle responding, neuroanatomical changes, and hypomethylation of an odor receptor encoding gene in the offspring of the trained animals, thus explaining transgenerational effects in fear learning.

Significance of Fear Conditioning for the Understanding of Anxiety Disorders

Although classical fear conditioning has been implicated as an important mechanism in the development of most anxiety disorders, there are few studies that support the etiological significance of fear conditioning. Retrospective data from either adult or child patients or the parents of children with specific phobias revealed that about 40–75% of the subjects or the parents could recall a traumatic event (e.g., Menzies and Clarke, 1993); a substantial number of the remaining subjects experienced vicarious conditioning or had been verbally informed about potential dangers related to the feared object. In evaluating these data, it must be kept in mind that all these reports are retrospective and that fear conditioning may be nondeclarative and therefore not be accessible to conscious recollection. Of greater relevance are, therefore, studies that examine fear conditioning in clinical groups or analogue samples. In social phobics, several studies used differential conditioning with faces as CS and aversive odor or shock as US and found either preexisting abnormalities or abnormal fear learning. In persons with psychopathy and antisocial personality disorder, disturbed fear conditioning seems to be present with a concomitant complete absence of orbitofrontal, insular, and cingulate activation during fear conditioning (e.g., Birbaumer et al., 2000). This absence of associative learning with respect to the aversive consequences of a stimulus might be the basis of the antisocial and often criminal behavior of these persons. In PTSD, higher conditionability as evident in enhanced and more difficult to extinguish fear responses has been observed. There is some evidence that delayed extinction of fear before

the disorder is predictive of the onset of PTSD later on. In addition, deficient context conditioning seems to co-occur with enhanced cue conditioning, lack of fear extinction, and excessive generalization (Flor and Nees, 2014; Milad and Quirk, 2012), and this dissociation of cue and context conditioning may be a crucial factor in the maintenance of aversive and disruptive trauma memories. More recently, pharmacological enhancers of extinction have gained much attention based on the increasing evidence of the molecular mechanisms involved in fear acquisition and extinction circuits such as the NMDA receptor, opioids, or dopamine (Fitzgerald et al., 2014).

Future Perspectives

Future research must focus on the elucidation of the neuronal correlates of fear conditioning in healthy humans and patients. It is not clear to what extent the amygdala plays the same role in fear conditioning in humans as it does in rodents. Furthermore, the role of awareness of the CS and the CS–US relationship has become more important with the view that the prediction of aversive events and the computation of prediction error is a crucial characteristic of fear learning. The advent of new methods such as optogenetics to specifically manipulate neural cell populations with high temporal and spatial precision, permits the elucidation of fear circuit activity in great detail (Johansen et al., 2012). The importance of fear conditioning for the understanding of anxiety disorders needs to be explored using paradigms that involve longitudinal measurements. Finally, the role of cognitive and genetic variables and their relationship to fear conditioning needs to be further explored.

See also: Anxiety and Anxiety Disorders; Anxiety and Fear Conditioning, Neural Basis of; Eyeblink Classical Conditioning; Fear-Potentiated Startle; Fear: Psychological and Neural Aspects.

Bibliography

Akirav, I., Maroun, M., 2007. The role of the medial prefrontal cortex-amygdala circuit in stress effects on the extinction of fear. Neural Plasticity 30873. http://dx.doi.org/10.1155/2007/30873.

Bechara, A., Tranel, D., Damasio, H., Adolphs, R., Rockland, C., Damasio, A.R., 1995. Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. Science 269, 1115–1118.

Birbaumer, N., Veit, R., Flor, H., Patrick, C.J., 2000. Excess and lack of fear: an fMRI study with social phobias and psychopaths. Society for Neuroscience 26, S1072. Abstract.

Bouton, M.E., Westbrook, R.F., Corcoran, K.A., Maren, S., 2006. Contextual and temporal modulation of extinction: behavioral and biological mechanisms. Biological Psychiatry 60, 352–360.

Bradley, M.M., Lang, P., 2000. Measuring emotion: behavior, feeling and physiology. In: Lane, R.D., Nadel, L. (Eds.), Cognitive Neuroscience of Emotion. Oxford University Press, New York.

Dias, B.G., Ressler, K.J., 2014. Parental olfactory experience influences behavior and neural structure in subsequent generations. Nature Neuroscience 17, 89–96.

Ehrlich, I., Humeau, Y., Grenier, F., Ciocchi, S., Herry, C., Lüthi, A., 2009. Amygdala inhibitory circuits and the control of fear memory. Neuron 62, 757–771.

Fitzgerald, P.J., Seemann, J.R., Maren, S., 2014. Can fear extinction be enhanced? A review of pharmacological and behavioral findings. Brain Research Bulletin 105, 46–60.

Flor, H., Nees, F., 2014. Learning, memory and brain plasticity in posttraumatic stress disorder: context matters. Restorative Neurology and Neuroscience 32, 95–102.

Johansen, J.P., Cain, C.K., Ostroff, L.E., LeDoux, J.E., 2011. Molecular mechanisms of fear learning and memory. Cell 147, 509–524.

Johansen, J.P., Wolff, S.B., Lüthi, A., LeDoux, J.E., 2012. Controlling the elements: an optogenetic approach to understanding the neural circuits of fear. Biological Psychiatry 71, 1053–1060.

Kindt, M., Soeter, M., Vervliet, B., 2009. Beyond extinction: erasing human fear responses and preventing the return of fear. Nature Neuroscience 12, 256–258.

LeDoux, J.E., 2000. Emotion circuits in the brain. Annual Review of Neuroscience 23, 155–184.

Ledoux, J.E., 2014. Coming to terms with fear. Proceedings of the National Academy of Sciences of the United States of America 111, 2871–2878.

Lonsdorf, T.B., Kalisch, R., 2011. A review on experimental and clinical genetic associations studies on fear conditioning, extinction and cognitive-behavioral treatment. Translational Psychiatry 1, e41.

Marek, R., Strobel, C., Bredy, T.W., Sah, P., 2013. The amygdala and medial prefrontal cortex: partners in the fear circuit. Journal of Physiology 591, 2381–2391.

Maren, S., Phan, K.L., Liberzon, I., 2013. The contextual brain: implications for fear conditioning, extinction and psychopathology. Nature Reviews Neuroscience 14, 417–428.

McNally, G.P., Johansen, J.P., Blair, H.T., 2011. Placing prediction into the fear circuit. Trends in Neurosciences 34, 283–292.

Menzies, R.G., Clarke, J.C., 1993. The etiology of childhood water phobia. Behaviour Research and Therapy 31, 499–501.

Milad, M.R., Quirk, G.J., 2012. Fear extinction as a model for translational neuroscience: ten years of progress. Annual Review of Psychology 63, 129–151.

Monfils, M.H., Cowansage, K.K., Klann, E., LeDoux, J.E., 2009. Extinction-reconsolidation boundaries: key to persistent attenuation of fear memories. Science 324, 951–955.

Morris, J.S., Öhman, A., Dolan, R.J., 1999. A subcortical pathway to the right amygdala mediating "unseen" fear. Proceedings of the National Academy of Sciences of the United States of America 96, 1680–1685.

Öhman, A., Flykt, A., Lundqvist, D., 2000. Unconscious emotion: evolutionary perspectives, psychophysiological data and neuropsychological mechanisms. In: Lane, R.D., Nadel, L. (Eds.), Cognitive Neuroscience of Emotion. Oxford University Press, Oxford, pp. 296–327.

Pape, H.C., Pare, D., 2010. Plastic synaptic networks of the amygdala for the acquisition, expression, and extinction of conditioned fear. Physiological Reviews 90, 419–463.

Rescorla, R.A., 1988. Pavlovian conditioning. It's not what you think it is. American Psychologist 43, 151–160.

Schiller, D., Delgado, M.R., 2010. Overlapping neural systems mediating extinction, reversal and regulation of fear. Trends in Cognitive Sciences 14, 268–276.

Schiller, D., Monfils, M.H., Raio, C.M., Johnson, D.C., Ledoux, J.E., Phelps, E.A., 2010. Preventing the return of fear in humans using reconsolidation update mechanisms. Nature 463, 49–53.

Sehlmeyer, C., Schoning, S., Zwitserlood, P., Pfleiderer, B., Kircher, T., Arolt, V., Konrad, C., 2009. Human fear conditioning and extinction in neuroimaging: a systematic review. PLoS One 4, e5865.

Seligman, M.E.P., 1971. Phobias and preparedness. Behavior Therapy 2, 307–320.

Vervliet, B., Baeyens, F., Van den Bergh, O., Hermans, D., 2013. Extinction, generalization, and return of fear: a critical review of renewal research in humans. Biological Psychology 92, 51–58.

Zovkic, I.B., Sweatt, J.D., 2013. Epigenetic mechanisms in learned fear: implications for PTSD. Neuropsychopharmacology 38, 77–93.

Relevant Websites

http://www.yasni.info/ext.php?url=http%3A%2F%2Fwww.bcf.uni-freiburg.de%2Fpeople%2Fassociated%2Fresearch%2Fd10&name=Andreas+L%C3%BCthi&showads=1&lc=de-de&lg=de&rg=de&rip=de – Andreas Lüthi Lab.

http://ppw.kuleuven.be/home/english/research/clep/people/00035504 – Bram Vervliet (CLEP Center).

http://neuroscience.nih.gov/Lab.asp?Org_ID=374 – Christian Grillon Lab.

http://www.dnalc.org/view/1238-Fear-Learning.html – DNA Learning Center.

http://fanselowlab.psych.ucla.edu/Main/Home.html – Fanselow Lab.

http://www.cns.nyu.edu/corefaculty/LeDoux.php – LeDoux Lab.

http://www.md.rcm.upr.edu/quirk/Home.html – The Quirk Lab.

Fear of Crime

Jodi Lane, University of Florida, Gainesville, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Fear of crime is more common than victimization and is important because it can lead to stress and behavioral precautions that impinge on quality of life. This article discusses definitions of fear as an emotional reaction (not a cognitive one) and enumerates how scholars have worked to better their fear measurements over time. The article also discusses the personal characteristics that predict fear and describes the theoretical ideas developed to explain it, including vulnerability (especially for women and the elderly), neighborhood factors (subcultural diversity, disorder, and community concern), and indirect victimization. Finally, we briefly discuss current and emerging trends in research.

Fear of Crime: Importance and Definitions

When people think about crime, they often consider its causes – or why people do it – and the consequences for others involved in the immediate criminal event, including victims and witnesses. Yet, crime has broader consequences beyond the people directly affected by or involved in specific incidents. One consequence is fear of becoming a victim, which is much more common among the population than victimization itself (Warr, 2000). While very few people report being victimized by violent crime (fewer than 3%) or property crime (about 14%) (Truman and Planty, 2012), a recent Gallup poll found that 38% of Americans were afraid to walk in the area around their home at night (Gallup, 2011). Research in other countries also finds that fear of crime is high compared to one's risk of personal victimization (European Crime Prevention Network, 2004).

Fear of crime is an important social issue generally and to policy makers because it affects quality of life and can lead to regular anxiety and behavioral changes, including avoiding specific areas, restricting daily activities, fortifying one's home, and carrying weapons to protect oneself (Ferraro, 1995; Warr, 1985). At the extremes, it might lead someone to do something passive but life altering, like refuse to leave his or her home, or active, like join a gang or commit peremptory attacks on others to prevent personal victimization (Lane and Fox, 2012).

As discussed in this article, over time, scholars have been inconsistent in how they measure, and therefore operationally define, fear of crime, but in the last 20 or so years, researchers have agreed that fear refers to an emotional reaction (not a thought process). For example, Ferraro (1995: p. 23) defined fear of crime as "an emotional response of dread or anxiety to crime or symbols that a person associates with crime" (see also Ferraro and LaGrange, 1987). Warr (2000: p. 453) further clarified the definition as "an emotion, a feeling of alarm or dread caused by an awareness or expectation of danger." While Farrall et al. (2009) agree that fear can be an emotion with an associated physical response, they argue that it is rare for people to have such an intense feeling. Rather, they argue that most people respond to crime emotionally in terms of worry about something happening to them now or in the future, and they note that it is important for researchers to measure the regularity and intensity of emotional reactions to crime.

Changes in Measurement of Fear over Time

Fear of crime has been measured in a variety of ways, depending on the study and its purpose. Before the 1980s, it was very common for researchers to measure fear of crime with one of two questions. The first, "How safe do you feel or would you feel being out alone in your neighborhood at night? (or during the day)," was designed for use in the National Crime Survey, now called the National Crime Victimization Survey. The second, "Is there any area right around here–that is, within a mile–where you would be afraid to walk alone at night," was developed for the General Social Survey (GSS) (Ferraro, 1995). In the 1980s, key fear-of-crime experts, including Ken Ferraro, Randy LaGrange, and Mark Warr, began criticizing these measures because they (1) confounded perceived risk of victimization, a cognitive assessment, with fear of crime, an emotional reaction; (2) did not mention the word crime; (3) did not distinguish among fears of different crimes (homicide, assault, burglary, etc.); and (4) asked about fear in hypothetical situations ('Would you?') rather than in everyday ones (Ferraro and LaGrange, 1987; LaGrange and Ferraro, 1987; Warr, 1984; Warr and Stafford, 1983; see also Farrall et al., 1997). While the GSS still uses its question and Gallup (2011) still uses a variation of the GSS question, "Is there any area near where you live–that is, within a mile–where you would be afraid to walk alone at night?," most scholarly studies in the last 20 years have focused on (1) measuring both perceived risk and fear of crime, (2) measuring perceived risk and fear of multiple offenses, and (3) ensuring that words that specify fear are included in the questions (e.g., 'afraid,' 'worry,' or 'safety'). In addition, some scholars have recently added more specificity to measures by arguing that fear may vary not only based on types of offenses but also based on types of offenders. Specifically, Lane and Meeker have argued that perpetrator characteristics can add another layer to our understanding of fear, and researchers in the last decade have conducted studies examining fear of gang members, terrorists, and strangers versus known offenders, for example (see Lane and Meeker, 2000, 2003b, 2003c, 2004; Nellis, 2009a, 2009b; Wilcox et al., 2006).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45058-0

Changes in Research Focus over Time

Personal Characteristics

Much of the early research on fear of crime focused on measuring the prevalence of fear and understanding the characteristics of people who were afraid. Early studies, using the general but problematic questions, showed that victims, minorities, and people living in urban (and especially lower income) areas were more afraid. In addition, results also indicated that women and the elderly were more afraid than men and younger people, even though they faced less actual risk of victimization (see, e.g., Baumer, 1978; Garofalo and Laub, 1978; Skogan and Maxfield, 1981; Warr, 1994, for a review). It made intuitive sense to researchers that people who lived in urban and low-income areas, often minorities, would be more afraid of crime, because crime was more common in those areas. Yet, early researchers were puzzled by the fact that women and the elderly were expressing more fear than males and young people, which led them to focus their work on understanding and explaining causes of crime-related fear other than objective risk of victimization. One approach, as mentioned here, was to improve the measurement of fear of crime, first by separating perceived risk from fear and by examining fear of specific offenses (e.g., LaGrange and Ferraro, 1987; Warr, 1984, 1985; Warr and Stafford, 1983). Interestingly, many of the studies that have examined fear of specific offenses in the last 30 years or so have continued to find that women were more afraid than men; yet, in contrast, they also found that younger people, not older ones, were more afraid, lending more support to expressed concern over the value of earlier measures (see, e.g., Ferraro, 1995, 1996; Lane and Meeker, 2003b; Warr, 1994, for a review). Other approaches were theoretical, and studies focused on finding factors other than demographic characteristics that could explain fear of crime. The next few sections will discuss the theoretical approaches that emerged over time and the research conducted to examine those theoretical ideas.

Vulnerability

One of the ideas to explain differential fear among people was argued by scholars such as Skogan and Maxfield (1981) and focused on physical and social vulnerability. Skogan and Maxfield (1981) noted that women and the elderly faced more physical vulnerability than males and young people. 'Physical vulnerability' refers to the possibility that they may be less likely to be able to protect themselves if attacked and to be physically injured if they were attacked, due to their generally smaller statures and comparative lack of physical strength. Women as a group tend to be smaller and have less physical power. The elderly have an additional burden of losing their physical strength and ability to move quickly as they age, and many of them live alone, especially women, since men tend to die younger. Skogan and Maxfield (1981) noted that many elderly would have no one to take care of them if they were injured, meaning that they might face being put into institutions if they were severely injured and could not care for themselves. According to Skogan and Maxfield (1981), people who are socially vulnerable, in contrast, are those who live in places where victimization risk is higher, that is, where crime is higher. In the United States, these people are more likely to be minority and poor. Even though many people in these groups may not have yet been victimized and may not be, their daily experiences in their neighborhoods mean that exposure to crime is higher in terms of what they witness and know about, and therefore their crime-related stress is higher.

Theoretical Perspectives on Gender Differences in Fear

Because women are consistently more afraid than men, no matter how fear is measured, some scholars have focused their theoretical arguments on specifically explaining greater fear among women. There are three related but distinct perspectives beyond the simple physical vulnerability argument, and all of them at some level point to feelings or experiences of vulnerability for women (Lane, 2013). The first is "differential socialization," which argues that the reason that women do not feel strong enough to protect themselves from victimization is that "women are socialized to be weak and submissive while men are socialized to be strong and powerful" by their parents and society at large (Lane, 2013: p. 58; see also De Groof, 2008). Hollander (2001) elaborated this argument when she found that her respondents just accepted that women were weaker, even though in reality some women are stronger than some men. She argued that being weak was part of being feminine and being strong and dangerous was considered an important element of masculinity. Interestingly, a recent study in Belgium found that the gender identity mattered more than biological sex did. Specifically, people who were attitudinally and behaviorally more masculine were less afraid than people who were more feminine, regardless of their biological sex (Cops and Pleysier, 2011; see Lane, 2013, for a review).

The second theoretical perspective on women's fear, originally specified by Kenneth Ferraro (1995, 1996), is the "shadow of sexual assault" thesis. Ferraro argued that women are more afraid of crime more generally because they continually worry that nonsexual crimes (e.g., harassment and burglary) can turn into sexual assault. Rape and other offenses served as what Mark Warr (1985: p. 245) called "perceptually contemporaneous offenses," because they were tied together in women's minds. They might be tied, in part, because women have been socialized particularly to fear rape, for example, by their parents who worried about how they dressed or with whom they went out (see Warr, 1984). Some other scholars (e.g., Gordon and Riger, 1989) have argued that fear of rape is a particular predicament that applies only to women. Studies aimed at testing this theoretical idea have consistently found that fear of rape is tied to fear of other crimes for women, and some studies have also found that it predicts fear of other crimes for men, too (Cook and Fox, 2012; Fisher and Sloan, 2003; Hilinski, 2009; Lane and Fox, 2013; Lane et al., 2009; Lane and Meeker, 2003c; May 2001b; Wilcox et al., 2006; see Lane, 2013, for a review).

The third theoretical idea focused on explaining greater fear of crime among women focuses on patriarchy, arguing that men use the threat of rape as a tool purposely to dominate and control women (Brownmiller, 1975). Betsy Stanko, a scholar who now heads the research unit for the London Metropolitan Police, has conducted qualitative research

studies with men and women about their fear of crime. She believes that women's greater fear of crime makes sense, because their 'ordinary experiences' in everyday life show them that they should be afraid (Stanko, 1985: p. 9). She believes that in a male-dominated society, men regularly terrorize women through the threat of physical and sexual harm. She also believes that typical crime statistics do not capture the subtle ways that men victimize women or women's crime experiences with acquaintances or family members, therefore underestimating the criminal victimization that women face (Stanko, 1985; see also Young, 1992). Stanko (1985, 1990) and Esther Madriz (1997) both find problematic the societal expectation that women can protect themselves from men if they just behave appropriately (e.g., dressing conservatively or avoiding certain places or men), because this puts the responsibility for violence against women on the women rather than the male perpetrators who actually commit the violence and hurt them (see Stanko, 1998).

In sum, the vulnerability argument points to either personal or environmental vulnerabilities that lead people to worry that they will become victims of crime, and some scholars believe these explanations easily can explain differential fear levels between women and men. As this section has illustrated, some scholars believe that social vulnerability arises for women because they are treated differently in society. Yet, other research efforts have focused on understanding other environmental factors, such as neighborhood characteristics and media consumption, which can create feelings of social vulnerability and increase fear of crime. Next, we turn to these ideas.

Neighborhood Factors

There are three distinct but connected theoretical ideas about neighborhood or community characteristics that can prompt people to fear crime. These ideas are set in the context of social disorganization theory, a macrolevel theory that posits that crime is more common in areas where residents cannot work together and where poverty, racial and ethnic diversity, and residential mobility are evident (Bursik, 1988; Bursik and Grasmick, 1993; Sampson, 1993; Shaw and McKay, 1972/1942). Fear-of-crime researchers borrowed from this macrolevel theory to create three microlevel theories about neighborhood factors that lead people to fear crime. These perspectives involve subcultural diversity, disorder and incivilities, and community concern and decline.

The first perspective, subcultural diversity, focuses on racial heterogeneity and asserts that people who live near others who look or act differently than they do are afraid because they do not understand the behaviors and mannerisms of those who are different (see Lewis and Salem, 1986; Merry, 1981). That is, they are unsure about how those others will act. Although most scholarly work examining this perspective has considered it in terms of its connections with the other ideas (e.g., disorder and community concern), there is support for this idea. For example, Madriz (1997) found that fear of crime was related to stereotypes about people who were racially and ethnically different, including people from other countries. Lane's (2002) qualitative work on the fear of gangs also showed that people's concern about primarily undocumented immigrants encroaching on their neighborhoods (rather than staying in the barrios) led them to worry about disorder and consequently decline in their neighborhoods, which in turn led them to fear gangs. Quantitative work also supports these findings that racial, ethnic, and cultural heterogeneity matter. For example, studies have shown that concerns about diversity, racial stereotypes, and actual neighborhood racial composition are predictive of fear (e.g., Chiricos et al., 1997b; Lane and Meeker, 2000; Liska et al., 1982; St. John and Heald-Moore, 1996).

The disorder or incivilities perspective asserts that when people see evidence of social (e.g., homelessness, gangs, and kids hanging out) and physical disorder (e.g., trash, unkempt property, graffiti, and bars on windows), they assume that crime is also there and that they are at risk of victimization, leading them to fear crime. Studies have generally shown that perceptions of disorder are important predictors of fear (e.g., Brunton-Smith and Sturgis, 2011; Ferguson and Mindel, 2007; Franklin et al., 2008; Kanan and Pruitt, 2002; LaGrange et al., 1992; Lane and Meeker, 2003a; Lewis and Maxfield, 1980; Robinson et al., 2003; Rountree and Land, 1996; Scarborough et al., 2010; Skogan and Maxfield, 1981; Taylor and Hale, 1986; Taylor, 2001). Fewer have found that objective indicators of disorder matter (Perkins et al., 1990, 1992; Perkins and Taylor, 1996).

The third environmental perspective, community concern, argues that people who are afraid are primarily affected by the fact that their neighborhood is in decline or no longer the way it used to be (Conklin, 1975; Garofalo and Laub, 1978). As Lane (2002) found, this was particularly relevant for long-term residents who had seen their neighborhoods change over time and no longer felt comfortable leaving their doors unlocked or sending children to the park to play alone. Much of the work on community concern links it to disorder, arguing that one of the reasons why disorder is predictive of fear is that it leads people to believe that their community is declining (see Garofalo and Laub, 1978; Skogan, 1990; Skogan and Maxfield, 1981). Lane's (2002) qualitative work showed this connection. Moreover, when quantitative studies examine this theoretical argument, they show support (see Lane and Meeker, 2003a, 2005, 2011; Taylor and Hale, 1986).

Indirect Victimization

In addition to the three microlevel theoretical ideas focused on neighborhood factors as key issues leading people to fear crime, there is a fourth one called 'indirect victimization.' There are multiple ways that one could be indirectly victimized, but research tends to focus on people learning about crime from friends, neighbors, community policing officers, television, and/or newspapers and therefore becoming afraid. Skogan (1986) argued that one of the key sources of information about crime is social networks – that is, talking to people we know (see also Tyler, 1980, 1984). Research on indirect victimization from other people is not as common as research on the impact of media sources, but those that have been conducted confirm that people can become more afraid or worried after talking to others about crime (see Lane, 2002; Covington and Taylor, 1991; Taylor and Hale, 1986). Research on the effects of community policing in particular sometimes shows that it

reduces fear of crime (e.g., Kelling et al., 1981; Roh and Oliver, 2005; Zhao et al., 2002), yet other research shows that it increases fear (e.g., Hinkle and Weisburd, 2008) or does not affect it (e.g., Scheider et al., 2003).

Researchers more commonly examine the impact of newspapers and television on fear of crime. Some research on newspaper effects finds that stories that are (1) more salient (e.g., near the front of the paper) or (2) more sensational or random and occurring locally increase fear (Heath, 1984; Liska and Baccaglini, 1990; Williams and Dickinson (1993)). However, other studies find that reading newspapers does not increase fear (Chiricos et al., 1997a; Perkins and Taylor, 1996) or that it decreases it for some people (e.g., Whites but not Latinos) (Lane and Meeker, 2003a). Research on the impact of television on fear typically finds that results depend on the audience characteristics and type of programming. In the 1970s, George Gerbner and his colleagues argued that people who watched a lot of television would be more afraid, regardless of their personal characteristics (Gerbner and Gross, 1976; Gerbner et al., 1977, 1978). This became known as the 'cultivation hypothesis,' and subsequent studies testing the idea found that once personal characteristics were taken into account, television viewing often did not impact fear (e.g., Hirsch, 1981; Hughes, 1980; Potter, 1986; Reith, 1999; Sacco, 1982; Tyler, 1984; Wober, 1978). Other studies, however, found that television mattered only for certain people or in certain areas (e.g., urban or high-crime areas) (e.g., Chiricos et al., 1997a, 2000a; Doob and Macdonald, 1979; Gerbner et al., 1980; O'Keefe and Reid-Nash, 1987; Schlesinger et al., 1991).

Current and Emerging Trends in Research and Theory

In summary, over time research has changed focus from personal characteristics alone to developing theoretical ideas about why personal characteristics might matter and to continually improving the validity, reliability, and specificity of fear-of-crime measures. There have been some recent trends that deserve additional mention. First, some research has focused on explaining how the individual theoretical models, while distinct, actually connect in some people's minds (e.g., Lane, 2002; Taylor and Hale, 1986; Lane and Meeker, 2003b, 2011). Second, in an effort to increase our understanding of fear and its impact on different people, research has focused on further refining samples, measures, and analysis approaches. One trend has been to examine fear in groups other than the general population of adults, including juvenile and adult offenders, who have more personal experience with crime (Lane, 2006, 2009; Lane and Fox, 2012, 2013; May, 2001a; May et al., 2002), and young people, who have less life experience (e.g., Brown and Benedict, 2004; De Groof, 2008; Irwin, 2004; May and Dunaway, 2000). Another trend has been to disaggregate samples to understand the differential impact of the theoretically important factors on different subsets of the population (e.g., women versus men, old versus young, minorities versus whites, and gang members versus nongang members) (e.g., Chiricos et al., 1997a, 1997b, 2000b; Cook and Fox, 2012; Ferraro, 1996; Lane and Fox, 2012, 2013; Lane, 2009; Lane and Meeker, 2003c, 2004, 2005, 2011; May 2001b). Moreover, recent research has pointed to the idea that although people often are more afraid than their victimization risk would warrant, the frequency and intensity of these fears are critical pieces of the puzzle that have not received enough attention. For example, Farrall and Gadd (2004) argued that people are not frequently afraid, having only a few episodes of fear per year. They also found that feelings of fear are not necessarily that intense when they occur.

In fact, Farrall et al. (2009) put forth the most recent theoretical ideas about fear of crime, beyond the incremental revisions of earlier ideas that have resulted from the wealth of studies on fear that have occurred over time. These scholars argued that fear of crime can be experienced in many ways, and they focus especially on the important differences between 'experiential' and 'expressive' fear. In their opinion, experiential fear occurs for those people who live in disordered and high-crime areas and face real objective risk of victimization, or real social vulnerability, as Skogan and Maxfield (1981) long ago discussed. That is, these people have experienced crime and know what it can do. They face real and powerful threats to their own safety and have episodes of intense emotional fear. According to Farrall et al. (2009), expressive fear of crime, in contrast, is much more common and is more about worries that people have about societal issues more generally and about their general feelings of insecurity. These issues of concern might include lack of social control, a lack of positive interactions with unknown people, and, as Merry (1981) first noted, a feeling that they cannot trust people who are different than they are to behave appropriately. These scholars argue that people have more general uneasiness rather than a lot of specific moments of intense emotion (Farrall et al., 2009), an idea that is reminiscent of Wilson's (1968) argument that people feel an "urban unease," which at its base is a concern about the failure of their communities and a lack of trust in others to follow accepted societal codes of conduct.

Summary and Conclusions

This article has focused on the study of fear of crime, which is more common than victimization. Fear is important because it has the potential to lead to behavioral precautions, which at one end of the spectrum can be passive but life altering (such as refusing to leave the home) or, at the other end, can be active and criminal (such as carrying weapons or joining a criminal street gang). This document first discussed the definition of fear (an emotional reaction) and then discussed how scholars have worked to better their measurements of the construct over time. Next came a discussion of the personal characteristics that are related to crime fear and the theoretical ideas that have been developed to explain the connections between demographic factors and fear. These ideas include vulnerability, neighborhood factors (subcultural diversity, disorder, and community concern), and indirect victimization. Finally, we discussed current and emerging trends in research and theory, including studying fear in different types of samples, disaggregating samples to examine the differential impacts of the theoretical factors on different groups, and pushing for research that

examines the frequency and intensity of fear emotions and realizes that fear can mean different types of reactions depending on the person.

See also: Age and Crime; Communities and Crime; Crime and Gender; Crime, Immigration, and Ethnicity; Crime, Race and Ethnicity; Feminist Criminology and Gender Studies; Media, Crime and Crime Waves; Spatial Dynamics and Crime.

Bibliography

Baumer, T.L., 1978. Research on fear of crime in the United States. Victimology 3, 254–264.

Brown, B., Benedict, W.R., 2004. Bullets, blades, and being afraid in Hispanic high schools: an exploratory study of the presence of weapons and fear of weapon-associated victimization among high school students in a border town. Crime & Delinquency 50, 372–394.

Brownmiller, S., 1975. Against Our Will: Men, Women and Rape. Simon & Schuster, New York.

Brunton-Smith, I., Sturgis, P., 2011. Do neighborhoods generate fear of crime? An empirical test using the British Crime Survey. Criminology 49, 331–369.

Bursik Jr, R.J., 1988. Social disorganization and theories of crime and delinquency. Criminology 26, 519–551.

Bursik Jr., R.J., Grasmick, H.G., 1993. Neighborhoods and Crime: The Dimensions of Effective Community Control. Lexington Books, New York.

Chiricos, T., Eschholz, S., Gertz, M., 1997a. Crime, news, and fear of crime: toward an identification of audience effects. Social Problems 44, 342–357.

Chiricos, T., Hogan, M., Gertz, M., 1997b. Racial composition of neighborhood and fear of crime. Criminology 35, 107–131.

Chiricos, T., Padgett, K., Gertz, M., 2000a. Crime, news, and fear of crime: toward an identification of audience effects. Social Problems 44, 342–357.

Chiricos, T., Padgett, K., Gertz, M., 2000b. Fear, TV news, and the reality of crime. Criminology 38, 755–786.

Conklin, J., 1975. The Impact of Crime. Macmillan, New York.

Cook, C., Fox, K.A., 2012. Testing the relative importance of contemporaneous offenses: the impacts of fear of sexual assault versus fear of physical harm among men and women. Journal of Criminal Justice 40, 142–151.

Cops, D., Pleysier, S., 2011. "Doing gender" in fear of crime: the impact of gender identity on reported levels of fear of crime in adolescents and young adults. British Journal of Criminology 51, 58–74.

Covington, J., Taylor, R.B., 1991. Fear of crime in urban residential neighborhoods: implications of between- and within-neighborhood sources for current models. The Sociological Quarterly 32, 231–249.

De Groof, S., 2008. And my mama said: the (relative) parental influence on fear of crime among adolescent girls and boys. Youth & Society 39, 267–293.

Doob, A.N., Macdonald, G.E., 1979. Television viewing and fear of victimization: is the relationship causal? Journal of Personality and Social Psychology 37, 170–179.

European Crime Prevention Network, 2004. A Review of Scientifically Evaluated Good Practices for Reducing Feelings of Insecurity or Fear of Crime in the EU Members States. http://www.eucpn.org/pubdocs/submitted for publicationreview_reducing_feelings_insecurities_fear_crime_en.pdf (accessed 13.02.11.).

Farrall, S., Bannister, J., Ditton, J., Gilchrist, E., 1997. Questioning the measurement of the 'fear of crime'. British Journal of Criminology 37, 658–679.

Farrall, S., Gadd, D., 2004. The frequency of fear of crime. British Journal of Criminology 44, 127–132.

Farrall, S., Jackson, J., Gray, E., 2009. Social Order and the Fear of Crime in Contemporary Times. Oxford University Press, Oxford.

Ferguson, K.M., Mindel, C.H., 2007. Modeling fear of crime in Dallas neighborhoods: a test of social capital theory. Crime & Delinquency 53, 322–349.

Ferraro, K.F., 1995. Fear of Crime: Interpreting Victimization Risk. State University of New York Press, New York.

Ferraro, K.F., 1996. Women's fear of victimization: shadow of sexual assault? Social Forces 75, 667–690.

Ferraro, K.F., LaGrange, R., 1987. The measurement of fear of crime. Sociological Inquiry 57, 70–101.

Fisher, B.S., Sloan III, J.J., 2003. Unraveling the fear of victimization among college women: is the "shadow of sexual assault hypothesis" supported? Justice Quarterly 20, 633–659.

Franklin, T.W., Franklin, C.A., Fearn, N.E., 2008. A multilevel analysis of the vulnerability, disorder, and social integration models of fear of crime. Social Justice Research 21, 204–227.

Gallup, 2011. Most Americans Believe Crime in U.S. is Worsening. http://www.gallup.com/poll/150464/Americans-Believe-Crime-Worsening.aspx (accessed 13.02.11.).

Garofalo, J., Laub, J., 1978. The fear of crime: broadening our perspective. Victimology 2, 242–253.

Gerbner, G., Gross, L., 1976. Living with television: the violence profile. Journal of Communication 26, 173–199.

Gerbner, G., Gross, L., Jackson-Beeck, M., Jeffries-Fox, S., Signorelli, N., 1977. TV violence profile no. 8: the highlights. Journal of Communication 27, 171–180.

Gerbner, G., Gross, L., Jackson-Beeck, M., Jeffries-Fox, S., Signarelli, N., 1978. Cultural indicators: violence profile no. 9. Journal of Communication 28, 176–207.

Gerbner, G., Gross, L., Morgan, M., Signorelli, N., 1980. The "mainstreaming" of America: violence profile no. 11. Journal of Communication 30, 10–29.

Gordon, M.T., Riger, S., 1989. The Female Fear: The Social Cost of Rape. University of Illinois Press, Urbana.

Heath, L., 1984. Impact of newspaper crime reports on fear of crime: multimethodological investigation. Journal of Personality and Social Psychology 47, 263–276.

Hilinski, C.M., 2009. Fear of crime among college students: a test of the shadow of sexual assault hypothesis. American Journal of Criminal Justice 34, 84–102.

Hinkle, J.C., Weisburd, D., 2008. The irony of broken windows policing: a micro-place study of the relationship between disorder, focused police crackdowns, and fear of crime. Journal of Criminal Justice 36, 503–512.

Hirsch, P.M., 1981. On not learning from one's own mistakes: a reanalysis of Gerbner et al.'s findings on cultivation analysis, part II. Communication Research 8, 3–37.

Hollander, J.A., 2001. Vulnerability and dangerousness: the construction of gender through conversation about violence. Gender and Society 15, 83–109.

Hughes, M., 1980. The fruits of the cultivation analysis: a reexamination of some effects of television viewing. Public Opinion Quarterly 44, 287–302.

Irwin, K., 2004. The violence of adolescent life: experiencing and managing everyday threats. Youth & Society 35, 452–479.

Kanan, J.W., Pruitt, M.V., 2002. Modeling fear of crime and perceived victimization risk: the (in)significance of neighborhood integration. Sociological Inquiry 72, 527–548.

Kelling, G.L., Pate, A., Ferrara, A., Utne, M., Brown, C.E., 1981. Newark Foot Patrol Experiment. The Police Foundation, Washington, DC.

LaGrange, R.L., Ferraro, K.F., 1987. The elderly's fear of crime: a critical examination of the research. Research on Aging 9, 372–391.

LaGrange, R.L., Ferraro, K.F., Supancic, M., 1992. Perceived risk and fear of crime: role of social and physical incivilities. Journal of Research in Crime and Delinquency 29, 311–334.

Lane, J., 2002. Fear of gang crime: a qualitative examination of the four perspectives. Journal of Research in Crime and Delinquency 39, 437–471.

Lane, J., 2006. Exploring fear of general and gang crimes among juveniles on probation: the impacts of delinquent behaviors. Youth Violence and Juvenile Justice 4, 34–54.

Lane, J., 2009. Perceptions of neighborhood problems, fear of crime and resulting behavioral precautions: comparing institutionalized girls and boys in Florida. Journal of Contemporary Criminal Justice 25, 264–281.

Lane, J., 2013. Theoretical explanations for gender differences in fear of crime: research and prospects. In: Renzetti, C.M., Miller, S.L., Gover, A.R. (Eds.), Routledge International Handbook of Crime and Gender Studies. Routledge, New York, pp. 57–67.

Lane, J., Gover, A.R., Dahod, S., 2009. Fear of violent crime among men and women on campus: the impact of perceived risk and fear of sexual assault. Violence and Victims 24, 172–192.

Lane, J., Fox, K.A., 2012. Fear of crime among gang and non-gang offenders: comparing the impacts of perpetration, victimization, and neighborhood factors. Justice Quarterly 29, 491–523.

Lane, J., Fox, K.A., 2013. Fear of property, violent, and gang crime: examining the shadow of sexual assault thesis among male and female offenders. Criminal Justice and Behavior 40, 472–496. (First published online January 24, 2013) http://dx.doi.org/10.1177/0093854812463564.

Lane, J., Meeker, J.W., 2000. Subcultural diversity and the fear of crime and gangs. Crime & Delinquency 46, 497–521.

Lane, J., Meeker, J.W., 2003a. Ethnicity, information sources, and fear of crime. Deviant Behavior 24, 1–26.

Lane, J., Meeker, J.W., 2003b. Fear of gang crime: a look at three theoretical models. Law and Society Review 37, 425–456.

Lane, J., Meeker, J.W., 2003c. Women's and men's fear of gang crimes: sexual and nonsexual assault as perceptually contemporaneous offenses. Justice Quarterly 20, 337–371.

Lane, J., Meeker, J.W., 2004. Social disorganization perceptions, fear of gang crime, and behavioural precautions among Whites, Latinos, and Vietnamese. Journal of Criminal Justice 32, 49–62.
Lane, J., Meeker, J.W., 2005. Theories and fear of gang crime among Whites and Latinos: a replication and extension of prior research. Journal of Criminal Justice 33, 627–641.
Lane, J., Meeker, J.W., 2011. The combined impact of theoretical models on perceived risk and fear of gang crime among Whites and Latinos. Victims & Offenders 6, 64–92.
Lewis, D.A., Maxfield, M.G., 1980. Fear in the neighbourhoods: an investigation of the impact of crime. Journal of Research in Crime and Delinquency 17, 160–189.
Lewis, D.A., Salem, G., 1986. Fear of Crime: Incivility and the Production of a Social Problem. Transaction Books, New Brunswick, NJ.
Liska, A.E., Baccaglini, W., 1990. Feeling safe by comparison: crime in the newspapers. Social Problems 37, 360–374.
Liska, A.E., Lawrence, J.J., Sanchirico, A., 1982. Fear of crime as a social fact. Social Forces 60, 760–770.
Madriz, E., 1997. Nothing Bad Happens to Good Girls: Fear of Crime in Women's Lives. University of California Press, Berkeley.
May, D.C., 2001a. Adolescent Fear of Crime, Perceptions of Risk, and Defensive Behaviors: An Alternative Explanation of Violent Delinquency. Edwin Mellen Press, Lewiston, NY.
May, D.C., 2001b. The effect of fear of sexual victimization on adolescent fear of crime. Sociological Spectrum 21, 141–174.
May, D.C., Dunaway, R.G., 2000. Predictors of fear of criminal victimization at school among adolescents. Sociological Spectrum 20, 149–168.
May, D.C., Vartanian, L.R., Virgo, K., 2002. The impact of parental attachment and supervision on fear of crime among adolescent males. Adolescence 27, 267–287.
Merry, S.E., 1981. Urban Danger: Life in a Neighborhood of Strangers. Temple University Press, Philadelphia.
Nellis, A.M., 2009a. Fear of terrorism. In: Borgeson, K., Valeri, R. (Eds.), Terrorism in America. Jones and Bartlett Press, Boston, pp. 117–144.
Nellis, A.M., 2009b. Gender differences in fear of terrorism. Journal of Contemporary Criminal Justice 25, 322–340.
O'Keefe, G.J., Reid-Nash, K., 1987. Crime news and real-world blues. Communication Research 14, 147–163.
Perkins, D.D., Florin, P., Rich, R.C., Wandersman, A., Chavis, D.M., 1990. Participation and the social and physical environment of residential blocks: crime and community context. American Journal of Community Psychology 18, 83–115.
Perkins, D.D., Meeks, J.W., Taylor, R.B., 1992. The physical environment of street blocks and resident perceptions of crime and disorder: implications for theory and measurement. Journal of Environmental Psychology 12, 21–34.
Perkins, D.D., Taylor, R.B., 1996. Ecological assessments of community disorder: their relationship to fear of crime and theoretical implications. American Journal of Community Psychology 24, 63–107.
Potter, W.J., 1986. Perceived reality and the cultivation hypothesis. Journal of Broadcasting and Electronic Media 30, 159–174.
Reith, M., 1999. Viewing crime drama and authoritarian aggressions: an investigation of the relationship between crime viewing, fear, and aggression. Journal of Broadcasting and Electronic Media 43, 211–221.
Robinson, J.B., Lawton, B.A., Taylor, R.B., Perkins, D.D., 2003. Multilevel longitudinal impacts of incivilities: fear of crime, expected safety, and block satisfaction. Journal of Quantitative Criminology 19, 237–274.
Roh, S., Oliver, W.M., 2005. Effects of community policing upon fear of crime: understanding the causal linkage. Policing: An International Journal of Police Strategies & Management 28, 670–683.
Rountree, P.W., Land, K.C., 1996. Perceived risk versus fear of crime: empirical evidence of conceptually distinct reactions in survey data. Social Forces 74, 1353–1376.
Sacco, V.F., 1982. The effects of mass media on perceptions of crime: a reanalysis of the issues. Pacific Sociological Review 25, 475–493.
Sampson, R.J., 1993. The community context of violent crime. In: Wilson Jr, W.J. (Ed.), Sociology and the Public Agenda. Sage, Newbury Park, pp. 259–286.
Scarborough, B.K., Like-Haislip, T.Z., Novak, K.J., Lucas, W.L., Alarid, L.F., 2010. Assessing the relationship between individual characteristics, neighborhood context, and fear of crime. Journal of Criminal Justice 38, 819–826.
Scheider, M.C., Rowell, T., Bezdikian, V., 2003. The impact of citizen perceptions of community policing on fear of crime: findings from twelve cities. Police Quarterly 6, 363–386.
Schlesinger, P., Tumber, H., Murdock, G., 1991. The media politics of crime and criminal justice. British Journal of Sociology 42, 397–420.
Shaw, C.R., McKay, H.D., 1972/1942. Juvenile Delinquency and Urban Areas. University of Chicago Press, Chicago.
Skogan, W., 1986. Fear of crime and neighborhood change. In: Reiss, A.J., Tonry, M. (Eds.), Communities and Crime. University of Chicago Press, Chicago, pp. 203–229.
Skogan, W.G., 1990. Disorder and Decline: Crime and the Spiral of Decay in American Neighborhoods. Free Press, New York.
Skogan, W.G., Maxfield, M.G., 1981. Coping with Crime: Individual and Neighborhood Reactions. Sage, Beverly Hills.
St. John, C., Heald-Moore, T., 1996. Racial prejudice and fear of criminal victimization by strangers in public settings. Social Inquiry 66, 267–284.
Stanko, E.A., 1985. Intimate Intrusions: Women's Experience of Male Violence. Routledge & Kegan Paul, Boston.
Stanko, E.A., 1990. Everyday Violence: How Women and Men Experience Sexual and Physical Danger. Pandora Press, London.
Stanko, E.A., 1998. Warnings to women: police advice and women's safety in Britain. In: Miller, S.L. (Ed.), Crime Control and Women: Feminist Implications of Criminal Justice Policy. Sage, Thousand Oaks, pp. 52–71.
Taylor, R.B., 2001. Breaking away from Broken Windows: Baltimore Neighborhoods and the Nationwide Fight against Crime, Grime, Fear and Decline. Westview Press, Boulder.
Taylor, R.B., Hale, M., 1986. Testing alternative models of fear of crime. The Journal of Criminal Law and Criminology 77, 151–189.
Truman, J.L., Planty, M., 2012. Criminal Victimization, 2011. U.S. Department of Justice, Washington, DC.
Tyler, T.R., 1980. Impact of directly and indirectly experienced events: the origin of crime-related judgments and behaviors. Journal of Personality and Social Psychology 39, 13–28.
Tyler, T.R., 1984. Assessing the risk of crime victimization: the integration of personal victimization experience and socially transmitted information. Journal of Social Issues 40, 27–38.
Warr, M., 1984. Fear of victimization: why are women and the elderly more afraid? Social Science Quarterly 65, 681–702.
Warr, M., 1985. Fear of rape among urban women. Social Problems 32, 238–250.
Warr, M., 1994. Public perceptions and reactions to violent offending and victimization. In: Reiss Jr., A.J., Roth, J.A. (Eds.), Understanding and Preventing Violence: Consequences and Control, vol. 4. National Academy Press, Washington, DC, pp. 1–66.
Warr, M., 2000. Fear of crime in the United States: avenues for research and policy. In: Duffee, D. (Ed.), Measurement and Analysis of Crime and Justice: Criminal Justice 2000, vol. 4. National Institute of Justice, Washington, DC, pp. 452–489.
Warr, M., Stafford, M., 1983. Fear of victimization: a look at the proximate causes. Social Forces 61, 1033–1043.
Wilcox, P., Jordan, C.E., Pritchard, A.J., 2006. Fear of acquaintance versus stranger rape as a "master status": towards refinement of the "shadow of sexual assault." Violence & Victims 21, 355–370.
Williams, P., Dickinson, J., 1993. Fear of crime: read all about it?: the relationship between newspaper crime reporting and fear of crime. British Journal of Criminology 33, 33–56.
Wilson, J.Q., 1968. The 'urban unease': community vs. city. The Public Interest 12, 25–39.
Wober, J.M., 1978. Televised violence and paranoid perception: the view from Great Britain. Public Opinion Quarterly 42, 315–321.
Young, V., 1992. Fear of victimization and victimization rates among women: a paradox? Justice Quarterly 9, 419–441.
Zhao, J.S., Scheider, M., Thurman, Q.C., 2002. The effect of police presence on public fear reduction and satisfaction: a review of the literature. The Justice Professional 15, 273–299.

Fear-Potentiated Startle

Alfons O Hamm, University of Greifswald, Greifswald, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Davis, volume 8, pp. 5425–5428, © 2001, Elsevier Ltd.

Abstract

The mammalian startle response is a cranial-to-caudal spreading wave of flexor movements elicited by abruptly occurring sensory events. This protective reflex is reliably potentiated when elicited during fear states in both animals and humans a phenomenon that has been named the *fear-potentiated startle*. Systematic research in rodents has uncovered the neural circuitry that is involved in the fear-induced potentiation of the startle reflex. Numerous studies have pinpointed the amygdala as the key structure involving the basolateral amygdala (input region) and the central amygdala (output region) that projects via the periaqueductal gray to the brain stem modulating the neural transmission in the primary acoustic startle pathway. Measuring the eyeblink component of the acoustic startle reflex is has been demonstrated that the human startle reflex is reliably potentiated when elicited during cues that have been previously paired with aversive events. This fear-potentiated startle is impaired in patients with unilateral lesions of the amygdala and completely blocked with bilateral lesions of the amygdala. In an effort to translate these encouraging empirical findings obtained in parallel both from animals and humans to the clinic this neuroscience and emotion science-based approach and methodology has been used to investigate emotional reactivity, fear learning, and memory in patients across a broad spectrum of emotional disorders. Accumulating evidence across laboratories suggests that the defense system is overactive to threat cues in the early development of an anxiety disorder. Defensive response mobilization generalizes more easily and is particularly enhanced in contexts in which unpredictable and uncontrollable aversive events occur. With increasing comorbidity, chronicity, and severity of the disorder blunted defensive responses occur.

Potentiation of the Mammalian Startle Response during Fear

In its full expression the mammalian startle response is a cranial-to-caudal spreading wave of flexor movements along the neural axis that is elicited by an abruptly occurring sensory event of a certain intensity (e.g., a sudden noise, a sharp movement, or a sudden tactile stimulus; Landis and Hunt, 1939). The startle response has been observed in many species and appears to be a protective reflex to prevent injury. In humans, the startle response consists of the following set of muscle movements: closure of the eyelid, forward head movement, raising and drawing forward of the shoulders, abduction of the upper arms, bending the elbows, pronation of the lower arms, forward movement of the trunk, contraction of the abdomen, and bending of the knees (see Davis, 1984). One of the most striking features of the startle response is that it is reliably potentiated if the organism is in a state of fear.

In their pioneering fear-conditioning experiment with rodents, Brown et al. (1951) used a light-tone compound stimulus as the conditioned stimulus (CS) and a footshock as the unconditioned stimulus (US), reasoning that anticipation of impending painful stimulation results in a conditioned fear response (fear was considered to be a learned drive at that time, Miller, 1948). The startle eliciting stimulus in this experiment was the sound resulting from firing a toy pistol two feet from the confinement box with drilled holes on top of the box (Brown et al., 1951). The whole-body startle response to the explosive sound was measured by scoring the deflections of the platform at the bottom of the box. Startle probes presented during the CS that has previously been associated with the painful US-evoked larger startle responses of the rats than probes presented during the same stimulus in the control group where this stimulus was never paired with the US. A series of follow-up studies provided evidence that this phenomenon – which has been termed the "*fear-potentiated startle effect*" – is indeed caused by a central fear state and not by other factors. Davis and Astrachan (1978) found a robust potentiation of the startle reflex after fear conditioning, independent of whether the shock US was administered to back or to the foot of the animal. Thus, the alternative explanation that a special body posture (e.g., crouching; Kurtz and Siegel, 1966) might have caused the startle potentiation was ruled out. Moreover, drugs that have anxiolytic effects in humans (e.g., diazepam) reduce the fear-potentiated startle effect in animals (Davis, 1979). By contrast, substances that induce fear in nonanxious individuals and increase symptoms in anxiety disorder patients (e.g., yohimbine) lead to increase in startle potentiation (Davis et al., 1979; see Davis, 1986, for review).

Neural Networks of the Fear-Potentiated Startle

Our understanding of the brain's neural circuitry underlying the fear-induced potentiation of the startle response comes primarily from neuroscience research with rodents and has been extensively investigated by Davis and colleagues (for reviews see Davis, 1998, 2000; Davis and Whalen, 2001; Davis et al., 2010). This neural circuitry is sketched panel (a) of **Figure 1**.

The *primary acoustic startle pathway* in rodents consists of three synapses (Davis, 1998). When the acoustic probe stimulus reaches the ear the afferent neural volleys travel from the spiral ganglion cells in the cochlea to very large cells embedded in the cochlear nerve in rodents (and also in humans) named 'cochlear root neurons,' and then project via thick axon

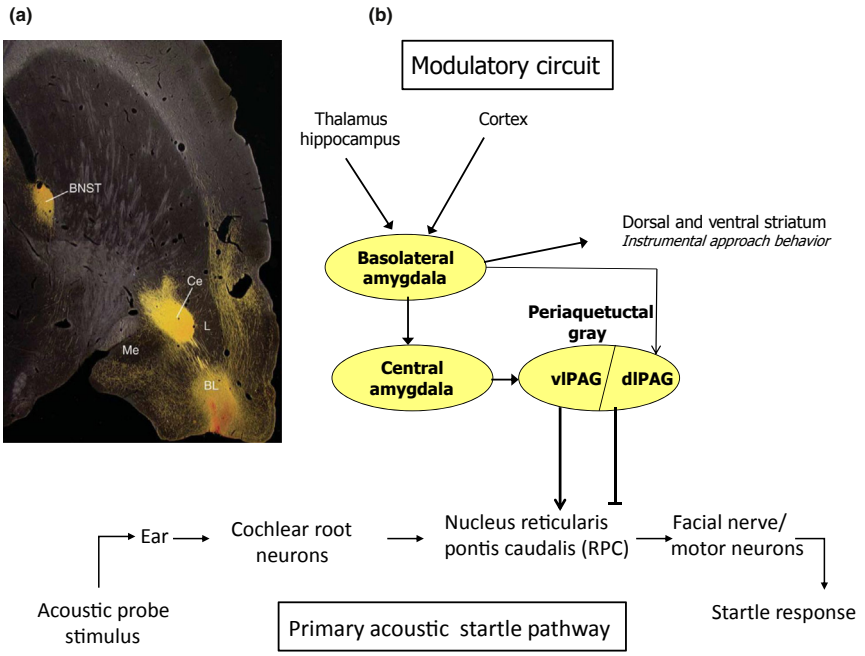

Figure 1 (a) Photomicrographs prepared and provided by Dr Chungjun Shi of 30 μm horizontal sections through the rat brain, cut at a slight angle to include the amygdala and the bed nucleus of the stria terminalis (BNST) in the same sectional plane. The anterograde tracer biotinylated dextran-amine (BDA) was infused in the basolateral amygdala (BL). Strong connections exist to the central nucleus of the amygdala (Ce) and the BNST (reprinted from Davis et al., 2010. Neuropsychopharmacology Reviews, Nature Publishing Group). (b) Schematic diagram of the neural structures involved in the fear-potentiated startle reflex. Conditioned stimuli activate the basolateral amygdala. In the case of fear-potentiated startle outputs of the basolateral amygdala impinge on the primary startle circuit via projections from the central nucleus of the amygdala along the periaqueductal gray. The ventral part activates attentive freezing and startle potentiation while the dorsal part activates flight behavior.

collaterals directly to the nucleus reticularis pontis caudalis (RPC), the sensorimotor interface in the pons. Efferent projections pass from the RPC through the facial motor nucleus (blink reflex) or the spinal cord (whole body startle response) to the reflex effectors (Davis, 1998). The anatomical structures forming the primary acoustic startle pathway in animals are found in humans as well (Martin et al., 1990).

The phenomenon of startle potentiation during fear conditioning suggests the existence of an additional *modulatory circuit*. Converging evidence now suggests that the amygdala and its many efferent projections may represent one of the key structures within this modulatory circuit. The basolateral amygdala receives sensory information from the thalamus, the hippocampus and the cortex. Depending on its previous association with either appetitive or aversive events the baso-lateral amygdala activates or modulates synaptic transmission in target areas appropriate to the reinforcement signal (or motivational valence) with which the sensory information has been associated. In case of appetitive cues, the basolateral amygdala activates the striatum organizing approach behavior toward the pleasant cue (Davis and Whalen, 2001; Lang and Bradley, 2010; see **Figure 1**).

If the sensory event has been associated with aversive stimuli, such as pain or dyspnea (shortness of breath), this sensory event changes neural transmission in the basolateral amygdala which projects to the central amygdala. Pathways from the central amygdala project to a variety of hypothalamic, midbrain and brain stem target areas producing the somatic, autonomic and endocrine signs of fear, including the *fear-potentiated startle*. Indeed electrical stimulation of the central amygdala elicits fear like behaviors including startle potentiation. In contrast, lesions of the central amygdala block fear-potentiated startle to cues that have been associated with shock (Hitchcock and Davis, 1986, 1991). The central amygdala projects via the ventral amygdalofugal pathway to the nucleus RPC in the brainstem, modulating the neural transmission in the primary acoustic startle pathway. There is evidence from Koch (1999) suggesting that a synaptic connection between the amygdalofugal pathway and the periaqueductal gray (PAG) is required for the fear-potentiated startle. Recent evidence suggests that there are parallel downstream pathways from the central amygdala that activate distinct patterns of a fear response (Gross and Canteras, 2012). Efferents from the central amygdala to the ventrolateral part of the PAG are crucial for interrupting ongoing behavior, promoting attentive freezing, and potentiation of the startle reflex. In contrast, activation of the lateral and dorsolateral part of the PAG initiates a flight behavior associated with sympathoexcitatory responses like tachycardia, hypertension, and redistribution of the blood flow (Benaroch, 2012). Moreover, infusion of low doses of potassium cyanide (KCN) facilitates escape behavior elicited by electrical stimulation of the dorsolateral PAG, while lesions of the dorsolateral PAG abolish KCN-evoked escape behavior (Schimitel et al., 2012). These animal findings are complemented by recent human data from our laboratory showing that freezing is associated

with potentiation of the startle reflex while escape behavior is indeed associated with a relative inhibition of the startle response (Hamm et al., 2014).

The Role of the Bed Nucleus of the Stria Terminalis (BNST)

In 1997 Walker and Davis made the observation that in rats startle responses were not only potentiated when elicited during cues that have previously been associated with aversive US conditions but also when elicited during exposure to light compared to startle responses to the same acoustic probes administered in the dark. This phenomenon has been named "*light-enhanced-startle*" effect. Interestingly this light-enhanced startle was reduced by both benzodiazepine and non-benzodiazepine anxiolytics (Walker and Davis, 1997, 2002). Interestingly, while fear-potentiated startle to a short (3.7 s) light CS that has been paired with a painful footshock is blocked by lesions or the central amygdala or by pharmacological blockade of the alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptors in this area, startle facilitation during long-lasting (20 min) exposure to bright light was unaffected by these manipulations. Lesions or local infusion of the AMPA antagonist NBQX into the BNST-blocked light-enhanced startle but left fear-potentiated startle to cues associated with aversive events intact. Since rodents are nocturnal and exhibit anxiety in open and illuminated spaces, one could argue that this environment creates a state of uncertainty about potential danger, typically described as anxious apprehension or worry. More recent animal data suggest that light-enhanced startle can be inhibited in a dose-dependent way by an oral administration of the corticotropin-releasing hormone (CRF) receptor 1 antagonist. Infusions of CRF into the lateral cerebral ventricle also markedly increased acoustic startle amplitudes in rats. Lesions of the lateral BNST completely blocked CRF-enhanced startle suggesting that this area is activated during states of anxious apprehension in uncertain contexts while the amygdala is more important for startle potentiation during short stimuli that were associated with aversive events (see for a review Davis et al., 2010). Interestingly, humans who are diurnal show the same facilitation of their startle response to probes presented during darkness (Grillon et al., 1997; Melzig et al., 2007), or during context conditions where unpleasant events occur in temporally unpredictable way (Grillon and Davis, 1997).

Fear-Potentiated Startle in Humans

To study the phenomenon of fear-potentiated startle in humans the movement of the eyelid, the most reliable and fastest component of the startle reflex sequence is recorded. Although many different procedures have been used to quantify the startle-eyeblink response (such as potentiometric, photoelectric, vertical electrooculographic, and magnetic search coil methods), most researchers currently record the electromyographic activity of the orbital portion of the orbicularis ocular muscle using miniature surface electrodes (see Lang et al., 1990; Blumenthal et al., 2005). Since the startle reflex can be elicited by any abruptly occurring sensory event of a certain intensity, electrical, mechanical (e.g., discrete taps or airpuffs to skin areas innervated by the trigeminal nerve), and visual (sudden increase in illumination) stimuli have been used in human research to elicit the startle reflex. However, the most commonly used startle-eliciting stimulus in human research is the acoustic probe stimulus, a 50-ms burst of white noise at an intensity of 95–100 dB presented with an instantaneous rise time.

Using this methodology the startle reflex has been used extensively to probe fear states of human beings. First in an effort to translate findings in rodents to humans it has been demonstrated that the startle eyeblink response is reliably potentiated in humans – like in rodents – when elicited during a cue that has previously been paired with an aversive electric shock. That is, during extinction startle responses are greater in magnitude (potentiated) to acoustic probes presented during previously fear-conditioned stimuli than during nonreinforced control stimuli or during the intertrial interval (ITI) (Hamm et al., 1993). Animal research suggests that startle potentiation might provide a direct metric for the assessment of defense system activation modulated by neural circuits including the amygdala. Thus, startle potentiation should be specific for fear learning and should not occur if behaviorally significant but nonaversive US conditions are associated with the conditioned stimuli. Following this reasoning Hamm and Vaitl (1996) completed a conditioning study with 100 participants in which for one group one (CS+) of two conditioned stimuli (pictures of neutral faces) was followed by an aversive electric shock during acquisition while the other neutral face (CS−) was never paired with the aversive US. To discriminate the acquisition of a fear response from pure cognitive learning of the contingencies another group saw the same two faces as conditioned stimuli, but in this case the CS+ was associated with a motivationally significant but nonaversive US (participants had to press a button when sensing a vibratory stimulus at their forearm and received a financial reward depending upon their reaction time), while the CS – was again never paired with this stimulus. The main findings of this experiment are depicted in the upper panel of **Figure 2**.

As in rodents, there was a clear potentiation of the startle reflex to probes that were presented during the reinforced stimulus (CS+) both compared to the nonreinforced CS – and to the probes that were presented in the absence of any visual foreground stimulus (ITI). Most importantly, the potentiation of the human startle reflex only occurred if the conditioned stimulus was associated with an aversive event and was absent if the CS was an imperative stimulus for a reaction time task. In contrast skin conductance responses – an index of orienting to significant (e.g., novel or arousing) stimuli (for review see Bradley, 2009) – were larger for the CS+ relative to CS – in both groups, i.e., independent whether the US was aversive or not. These data that have been replicated many times (Nitschke et al., 2002; Weike et al., 2007, 2008) clearly support the view that the startle potentiation is indeed a direct metric for the acquisition of a fear response.

Animal data show that bilateral lesions of the amygdala abolish fear-potentiated startle suggesting that the amygdala is

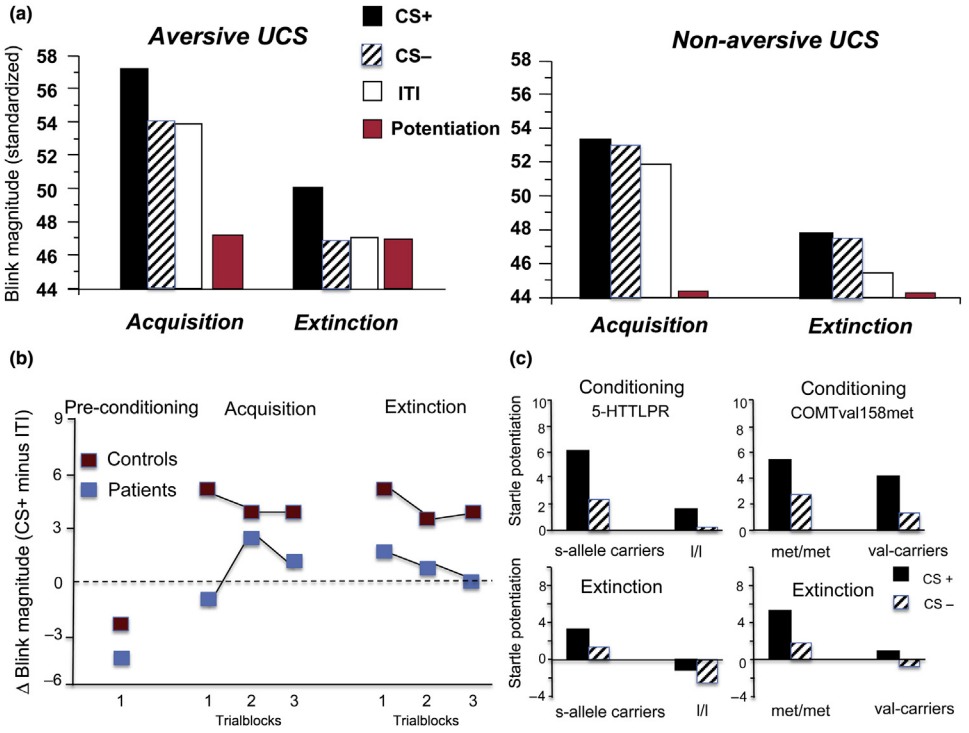

Figure 2 (a) Mean startle blink response magnitudes elicited during the presentation of the conditioned stimuli (CS+ and CS−) and the intertrial interval (ITI) and potentiation scores (CS+ minus CS−) during acquisition and extinction using an aversive shock (left panel) or a nonaversive reaction time task (right panel) as unconditioned stimulus in a differential delay conditioning experiment with healthy individuals. (b) Mean conditioned startle blink response potentiation for patients with unilateral temporal lobectomy and age matched controls, before, during, and after classical fear conditioning. (c) Potentiation (black bars: CS+ minus ITI; striped bars: CS− minus ITI) of startle response magnitudes as a function of genotype and stimulus. Results are shown for 5-HTTLLPR genotype groups (left) and for COMTval158met genotype groups (right) during conditioning (upper panel) and extinction (lower panel).

one key component of the neural circuitry modulating fear-potentiated startle. Following up on these animal findings Weike et al. (2005) completed a fear-conditioning experiment (identical to the one reported above) with 30 patients suffering from medically intractable unilateral temporal lobe epilepsy who had to undergo neurosurgery. For 18 patients the resections of the temporal lobe (amygdalahippocampectomy) involved the left anterior part of the temporal lobe (4.5 cm from the left temporal pole) in 12 patients the right anterior temporal lobe (5.5 cm from the right temporal pole). The bottom panel of Figure 2 (panel (b)) shows the main findings of this study.

Fear-conditioned startle potentiation was not completely abolished but significantly impaired in patients with unilateral lesions of the amygdala compared to healthy controls. These impairments were independent of lesion sites. Importantly, those patients who showed clear learned discrimination in skin conductance also exhibited clear impairment in fear-potentiated startle, suggesting that impairment of fear-potentiated startle was not due to a general learning deficit in these patients. Intact skin conductance conditioning for these patients was also reported by Coppens particularly when attention was explicitly allocated to the conditioning task by instructing the patients to provide shock expectancy ratings during each learning trial (Coppens et al., 2009). In contrast, and supporting the findings described above fear-potentiated startle was completely blocked in a group of four female subjects with Urbach–Wiethe disease from the Northern Cape of South Africa although half of these subjects learned the contingencies as assessed by verbal report data (Klumpers, 2012).

Recent behavioral genetic studies suggest that acquisition and also the extinction of fear-potentiated startle is modulated by genetic polymorphisms. Hariri et al. (2002) showed that a polymorphism in the promoter region of the human serotonin transporter (5-HTT) gene (*SLC6A4*) is related to amygdala activity evoked by threat cues, i.e., carriers of the short allele (a 43bp deletion) of the 5-HTT promoter (5-HTTLPR) show increased amygdala activation during viewing of fearful faces compared to long allele carriers. Based on these findings Lonsdorf et al. (2009) tested fear-potentiated startle in 5-HTTLPR s-allele carriers ($N = 30$) and l/l homozygotes ($N = 18$). In this study, on day 1 (Acquisition) one of two angry faces was associated with a highly annoying aversive unconditioned electrical stimulus while the other angry face was never paired with the US. After consolidation of the associative fear memory over night, extinction (both faces were presented without any US) started on the second day. Again startle probes were presented both during CS+, CS− and the ITI both during acquisition and extinction. The main findings of this study are depicted in panel (c) of Figure 2. As can be seen on left side of panel (c), only carriers of the 5-HTTLPR s allele acquired a significant

potentiation of the startle during cues that were associated with an aversive event through Pavlovian conditioning. In contrast long/long carriers did not show significant fear-potentiated startle. In this study, individuals were also screened for a second single nucleotide polymorphism on a gene coding for catechol-*O*-methyltransferase (COMT). Carriers of the met-allele have a fourfold reduction in enzyme activity compared with carriers of the val-allele, and as a result have higher extracellular dopamine levels particularly in the prefrontal cortex. Interestingly, acquisition of fear-potentiated startle was not affected by this genetic polymorphism. In contrast extinction of fear-potentiated startle was stronger for the val-allele carriers but was not affected by the 5-HTTLPR clearly supporting animal data showing that networks involved in fear learning differ from those involved in fear extinction (Quirk and Mueller, 2008). These genetic data are in line with findings from functional imaging studies showing that the acquisition of fear is associated with increased activation of the amygdala while extinction of the fear response requires activation of medial prefrontal cortical areas (Büchel and Dolan, 2000; Phelps et al., 2004; Milad et al., 2006). Klumpers et al. (2011) also found stronger fear-potentiated startle in individuals carrying at least one short allele for the 5-HTTLPR polymorphism compared to the long allele homozygotes, thus replicating and extending the findings of Lonsdorf et al. (2009). Moreover, using a context conditioning design in which aversive events occurred in one context at unpredictable times while no aversive stimuli were presented in another safe context, Glotzbach-Schoon et al. (2013) found stronger startle potentiation during the anxiety context in carriers of the short allele for the 5-HTTLPR who also carried the risk allele of the neuropeptide S receptor. Using also a context conditioning paradigm Heitland et al. (2013) again found increased potentiation of the startle reflex during aversive context conditions in short allele 5-HTTLPR carriers, who also carried the risk allele for the CRF receptor 1, a transmitter system which is particularly involved during startle enhancement in uncertain context conditions.

Using Fear-Potentiated Startle to Study Clinical Problems

In an effort to translate these basic findings from animal and human research to clinical problems, Hamm et al. (1997) first demonstrated that individuals with animal phobia showed a clear potentiation of the startle reflex when confronted with pictures of their feared animals (either snakes or spiders). Potentiation of the startle reflex evoked during viewing of these phobia-specific cues was significantly stronger than the facilitation of the startle reflex that was observed during viewing of other unpleasant but nonphobia-relevant pictures. These findings were replicated and extended by McTeague et al. (2012) who found a marked potentiation of the startle reflex in patients with specific phobia when probes were presented during imagery of personal narratives representing each patient's primary clinical fear. Again startle potentiation exceeded increase in startle magnitudes evoked during imagery of standard negative text (e.g., survival threat). Interestingly, this fear specific startle potentiation was more pronounced for patients who were diagnosed with a single specific phobia as principal diagnosis. Patients who were diagnosed with a different principal diagnosis (e.g., generalized anxiety disorder, adjustment disorder, etc.) but had a specific phobia as a comorbid diagnosis did show less startle potentiation during imagery of their personal fear scenes compared to patients with specific phobia as their principal diagnosis. Moreover, these patients with specific phobia as a comorbid disorder also showed more general negative affectivity and stronger general impairment. These data suggest that the assessment of startle potentiation during activation of fear-networks might serve as a meaningful phenotype to characterize different subgroups of patients within one diagnostic category possibly providing some neuroscience-based evidence for distinct psychopathological mechanisms.

Interestingly, a parallel pattern of results was observed for patients with social phobia (McTeague et al., 2009). Using the same script-driven imagery paradigm McTeague and coworkers found again increased startle potentiation during imagery of personal fear scenes in those patients who reported circumscribed focal social fears (like public speaking, performance fears, etc.), paralleling the findings for patients with circumscribed specific phobia. In contrast, those patients who had a more generalized social phobia, potentiation of the startle reflex was not specific for the personal fear but also generalized to other more standardized social situations. These data are in line with studies investigating fear-potentiated startle in patients with post-traumatic stress disorder (PTSD). Grillon and coworkers (Grillon and Morgan, 1999; Grillon, 2002) used a fear-conditioning paradigm to compare fear-potentiated startle in Gulf war veterans with and without PTSD. While PTSD patients did not differ from controls in the fear-potentiated startle during the cue that predicted the shock, startle responses elicited during the context in which the conditioning experiment took place was larger in the veterans with PTSD. Patients with PTSD had also larger startle magnitudes compared to patients with generalized anxiety disorder (GAD) when probes were presented during a context in which aversive events are unpredictable (Grillon et al., 2009). These data indicate that unpredictable and uncontrollable aversive events particularly activate defense response mobilization in these patients. Moreover, patients with PTSD also show impaired safety learning (Jovanovic et al., 2012). Inhibition of fear-potentiated startle by a safety cue presented in compound with the previously conditioned stimulus was reduced in PTSD patients. McTeague et al. (2010) found increased fear-potentiated startle in a large group of PTSD patients from a community sample with different trauma types during imagery of their personal traumatic narratives. In this PTSD-group a subgroup of patients was only exposed to a single trauma while another subgroup comprised patients exposed to two or more of such traumatic events (multiple-trauma subgroup). Enhanced startle potentiation was only observed in those PTSD patients with a single trauma. Multiple trauma patients showed no startle potentiation and a generally blunted physiological emotional response during imagery of their personal trauma. The multiple trauma patients had an earlier onset of their PTSD symptoms, reported more severe symptoms and showed greater comorbidity.

Following up on these findings, Limberg et al. (2011) investigated 40 nonmedicated patients with borderline personality disorder again using the script-driven imagery paradigm described above. Eighty percent of these patients had been exposed to traumatic events in the past and 26 patients in this group were also diagnosed with a comorbid current PTSD. Overall, startle modulation did not differ between borderline personality disorder patients and healthy controls, supporting previous findings (Herpertz et al., 1999), however, patients exhibited an increased potentiation of their startle responses during imagery of scenes describing rejection and abandonment, emotional schemas that are specifically relevant for these patients (Russel et al., 2007). Interestingly, blunted startle responses during imagery of personal trauma scenes was again observed in this study for those borderline patients with severe comorbid PTSD and the exposure to multiple traumas, probably reflecting a compromised defensive system after chronic exposure to severe stressors. Such explanation would be in line with animal data showing compromised defensive response mobilization in rats exposed to highly intense, prolonged, and inescapable shock.

Fear-potentiated startle has also been used to study psychopathology in panic disorder patients. Panic disorder patients report recurrent unexpected panic attacks (brief periods of intense fear accompanied by at least four out of 13 key symptoms including somatic symptoms) accompanied by at least 1 month of anxious apprehension about the implications or the consequences of the attacks or about having additional attacks in the near future. Importantly, for this group of anxiety patients the potential threat comes from inside the body. Therefore it might not come as a complete surprise that patients with panic disorder overall did not differ from controls in their potentiation of the startle reflex during anticipation of external nociceptive events (Grillon et al., 1994; Melzig et al., 2007). In contrast, if startle probes were presented during the perception of unexpected mild body symptoms, a clear potentiation of the startle response was observed for patients with panic disorder (Hamm et al., 2014) and in individuals who are at risk to develop a panic disorder (Melzig et al., 2011). In these studies, a voluntary hyperventilation procedure was employed where participants were guided to increase their respiratory rate and/or tidal volume exceeding the physiological demand, leading to a rapid drop in blood partial pressure of carbon dioxide, thus evoking typical body symptoms such as heart racing, dizziness, etc. Melzig et al. (2011) presented startle probes during early recovery from this hyperventilation procedure, because the pCO_2 level is still below 30 mmHg for approximately 2 min and body symptoms are still present without the interference of the task itself. As predicted, startle response magnitudes were substantially potentiated in individuals scoring high on the anxiety sensitivity index a risk factor for the development of a panic disorder. Accordingly patients with panic disorder and agoraphobia also showed increased potentiation of the startle reflex relative to age-matched controls when probes were presented during a 10-min recovery period after hyperventilation. Importantly, the amount of fear-potentiated startle during the perception of unexpected mild body symptoms was larger for those patients with a more severe symptom score on the panic and agoraphobia scale (Hamm et al., 2014). Panic disorder patients not only show increased defense response mobilization during perception of mild body symptoms they also show increased fear-potentiated startle during context conditions in which unpredictable and uncontrollable aversive events occur (Grillon et al., 2008). Moreover these patients also exhibit stronger generalization of their fear-potentiated startle to cues that resemble the original cue that was associated with an aversive event (Lissek et al., 2010). These data suggest that these patients show enhanced anxious apprehension toward unpredictable and uncontrollable events (like panic attacks) and that this defensive mobilization generalizes very easily to different situations thus providing a possible mechanism for the generalized avoidance behavior that is typically seen in patients with panic disorder and agoraphobia.

This translation of the fear-potentiated startle paradigm into the clinical research domain has provided very promising results leading to a better understanding of psychopathology from a neuroscience perspective. Moreover, fear-potentiated startle has also been used recently to elucidate the mechanisms involved in eliminating fear reactions to aversive events. Kindt et al. (2009) demonstrated that the oral administration of the beta-adrenergic antagonist just prior to the reactivation of the fear-conditioned memory abolished expression of fear-potentiated startle 1 day later, thus preventing reconsolidation of fear memories. Following up on these findings Sevenster et al. (2013) demonstrated that this disruption of reconsolidation and blockade of fear-potentiated startle requires active new learning, i.e., inducing a prediction error, the experience that an expected aversive event is no longer associated with the cue. The consolidation and reconsolidation of fear memories as indexed by fear-potentiated startle is modulated not only by the monoaminergic system but also by the glutamatergic system. In a series of experiments it was demonstrated that extinction of fear-potentiated startle was facilitated by administering the *N*-methyl-D-asparate (NMDA) receptor (glutamate acts on the NMDA receptor) partial agonist D-cycloserine either systemically or directly into the basolateral amygdala (Walker et al., 2002). Clinical studies that systematically test the effects of such pharmacological support of exposure therapies are under way (Davis et al., 2005; Hofmann et al., 2006; Otto et al., 2010).

Acknowledgments

The research reported in this article is supported by grants from the German Research Society (Ha 1593/15-1; Ha 1593/18-1) and by the German Federal Ministry of Education and Research (project number 01GV0615).

See also: Anxiety and Fear Conditioning, Neural Basis of; Fear Conditioning: Overview; Fear: Psychological and Neural Aspects; Genetic Approaches to Memory; Learning and Memory, Neural Basis of.

Bibliography

Benaroch, E.E., 2012. Periaqueductal grey: an interface for behavioral control. Neurology 78, 210–217.

Blumenthal, T.D., Cuthbert, B.N., Filion, D.L., Hackley, S., Lipp, O.V., van Boxtel, A., 2005. Committee report: guidelines for human startle eyeblink electromyographic studies. Psychophysiology 42, 1–15.

Bradley, M.M., 2009. Natural selective attention: orienting and emotion. Psychophysiology 46, 1–11.

Brown, J.S., Kalish, H.I., Farber, I.E., 1951. Conditioned fear as revealed by magnitude of startle response to an auditory stimulus. Journal of Experimental Psychology 41, 317–328.

Büchel, C., Dolan, R.J., 2000. Classical fear conditioning in functional neuroimaging. Current Opinion in Neurobiology 10, 219–223.

Coppens, E., Spruyt, A., Vandenbulcke, M., Van Paesschen, W., Vansteenwegen, D., 2009. Classical conditioned fear responses are preserved following unilateral temporal lobectomy in humans when concurrent US-expectancy ratings are used. Neuropsychologia 47, 2496–2503.

Davis, M., 1979. Diazepam and flurazepam: effects on conditioned fear as measured with the potentiated startle paradigm. Psychopharmacology 62, 1–7.

Davis, M., 1984. The mammalian startle response. In: Eaton, R.C. (Ed.), Neural Mechanisms of Startle Behavior. Plenum Press, New York, pp. 287–351.

Davis, M., 1986. Pharmacological and anatomical analysis of fear conditioning using the fear-potentiated startle paradigm. Behavioral Neuroscience 100, 814–824.

Davis, M., 1998. Are different parts of the extended amygdala involved in fear versus anxiety. Biological Psychiatry 44, 1239–1247.

Davis, M., 2000. The role of the amygdala in conditioned and unconditioned fear and anxiety. In: Aggleton, J.P. (Ed.), The Amygdala, vol. 2. University Press, Oxford, pp. 213–287.

Davis, M., Astrachan, D.I., 1978. Conditioned fear and startle magnitude: effects of different footshock or backshock intensities used in training. Journal of Experimental Psychology: Animal Behavior Processes 4, 95–103.

Davis, M., Whalen, P.J., 2001. The amygdala: vigilance and emotion. Molecular Psychiatry 6, 13–34.

Davis, M., Redmond, D.E., Baraban, J.M., 1979. Noradrenergic agonists and antagonists: effects on conditioned fear as measured by the potentiated startle paradigm. Psychopharmacology 65, 111–118.

Davis, M., Myers, K.M., Ressler, K.J., Rothbaum, B.O., 2005. Facilitation of extinction of conditioned fear by D-cycloserine: Implications for psychotherapy. Current Directions in Psychological Science 14, 214–219.

Davis, M., Walker, D.L., Miles, L., Grillon, C., 2010. Phasic vs. sustained fear in rats and humans: role of the extended amygdala in fear vs. anxiety. Neuropsychopharmacology Reviews 35, 105–135.

Glotzbach-Schoon, E., Andreatta, M., Reif, A., Ewald, H., Tröger, C., Baumann, C., Deckert, J., Mühlberger, A., Pauli, P., 2013. Contextual fear conditioning in virtual reality is affected by 5-HTTLPR and NPSR1 polymorphisms: effects on fear-potentiated startle. Frontiers in Behavioral Neuroscience 7. Article 31, 1–14.

Grillon, C., 2002. Startle reactivity and anxiety disorders: aversive conditioning, context, and neurobiology. Biological Psychiatry 52, 958–975.

Grillon, C., Davis, M., 1997. Fear-potentiated startle conditioning in humans: explicit and contextual cue conditioning following paired vs. unpaired training. Psychophysiology 34, 451–458.

Grillon, C., Morgan, C.A., 1999. Fear-potentiated startle conditioning to explicit and contextual cues in Gulf war veterans with posttraumatic stress disorder. Journal of Abnormal Psychology 108, 134–142.

Grillon, C., Pellowski, M., Merikangas, K.R., Davis, M., 1997. Darkness facilitates the acoustic startle reflex in humans. Biological Psychiatry 42, 453–460.

Grillon, C., Ameli, R., Goddard, A., Woods, S., Davis, M., 1994. Baseline and fear-potentiated startle in panic disorder patients. Biological Psychiatry 35, 431–439.

Grillon, C., Pine, D.S., Lissek, S., Rabin, S., Vythilingarn, M., 2009. Increased anxiety during anticipation of unpredictable aversive stimuli in posttraumatic stress disorder but not in generalized anxiety disorder. Biological Psychiatry 66, 47–53.

Grillon, C., Lissek, S., Rabin, S., McDowell, D., Dvir, S., Pine, D.S., 2008. Increased anxiety during anticipation of unpredictable but not predictable aversive stimuli as a psychophysiologic marker of panic disorder. American Journal of Psychiatry 165, 898–904.

Gross, C.T., Canteras, N.S., 2012. The many paths to fear. Nature Reviews, Neuroscience 13, 651–658.

Hairiri, A.R., Mattay, V.S., Tessitore, A., Kolachana, B., Fera, F., Goldman, D., Egan, M.F., Weinberger, D.R., 2002. Serotonin transporter genetic variation and the response of the human amygdala. Science 297, 400–403.

Hamm, A.O., Greenwald, M.K., Bradley, M.M., Lang, P.J., 1993. Emotional learning hedonic change and the startle probe. Journal of Abnormal Psychology 102, 453–465.

Hamm, A.O., Vaitl, D., 1996. Affective learning: awareness and aversion. Psychophysiology 33, 698–710.

Hamm, A.O., Cuthbert, B.N., Globisch, J., Vaitl, D., 1997. Fear and the startle reflex: blink modulation and autonomic response patterns in animal and mutilation fearful subjects. Psychophysiology 34, 97–107.

Hamm, A.O., Richter, J., Pané-Farré, C.A., 2014. When the threat comes from inside the body: a neuroscience based learning perspective of the etiology of panic disorder. Restorative Neurology and Neuroscience 32, 79–93.

Heitland, I., Groenink, L., Bijlsma, E.Y., Oosting, R.S., Baas, J.M.P., 2013. Human fear acquisition deficits in relation to genetic variants of the corticotropin releasing hormone receptor 1 and the serotonin transporter. PLoS ONE 8, e63772.

Herpertz, S.C., Kunert, H.J., Schwenger, U.B., Sass, H., 1999. Affective responsiveness in borderline personality disorder: a psychophysiological approach. American Journal of Psychiatry 156, 1550–1556.

Hitchcock, J.M., Davis, M., 1986. Lesions of the amygdala but not of the cerebellum block conditioned fear as measured with the potentiated startle paradigm. Behavioral Neuroscience 100, 11–22.

Hitchcock, J.M., Davis, M., 1991. The efferent pathway of the amygdala involved in conditioned fear as measured with the fear-potentiated startle paradigm. Behavioral Neuroscience 105, 826–842.

Hofmann, S.G., Meuret, A.E., Smits, J.A.J., Simon, N.M., Pollack, M.H., Eisenmenger, K., Shiekh, M., Otto, M.W., 2006. Augmentation of exposure therapy with D-cycloserine for social anxiety disorder. Archives of General Psychiatry 63, 298–304.

Jovanovic, T., Kazana, A., Bachevalier, J., Davis, M., 2012. Impaired safety learning may be a biomarker of PTSD. Neuropharmacology 62, 695–704.

Kindt, M., Soeter, M., Vervliet, B., 2009. Beyond extinction: erasing human fear responses and preventing the return of fear. Nature Neuroscience 12, 256–258.

Klumpers, F., 2012. Fear Not: Neurobiological Mechanisms of Fear and Anxiety. Ipskamp Drukkers B.V., The Netherlands.

Klumpers, F., Heitland, I., Oosting, R.S., Kenemans, J.L., Baas, J.M.P., 2011. Genetic variation in serotonin transporter function affects human fear expression indexed by fear-potentiated startle. Biological Psychology 89, 277–282.

Koch, M., 1999. The neurobiology of startle. Progress in Neurobiology 59, 107–128.

Kurtz, K.H., Siegel, A., 1966. Conditioned fear and the magnitude of the startle response: a replication and extension. Journal of Comparative and Physiological Psychology 62, 8–14.

Landis, C., Hunt, W.A., 1939. The Startle Pattern. Farrar, New York.

Lang, P.J., Bradley, M.M., Cuthbert, B.N., 1990. Emotion, attention, and the startle reflex. Psychological Review 97, 377–395.

Lang, P.J., Bradley, M.M., 2010. Emotion and the motivational brain. Biological Psychology 84, 437–450.

Limberg, A., Barnow, S., Freyberger, H.J., Hamm, A.O., 2011. Emotional vulnerability in borderline personality disorder is cue specific and modulated by traumatization. Biological Psychiatry 69, 574–582.

Lissek, S., Rabin, S., Heller, R.E., Lukenbaugh, D., Geraci, M., Pine, D.S., Grillon, C., 2010. Overgeneralization of conditioned fear as a pathogenic marker of anxiety disorder. American Journal of Psychiatry 167, 47–55.

Lonsdorf, T.B., Weike, A.I., Nikamo, P., Schalling, M., Hamm, A.O., Öhman, A., 2009. Genetic gating of human fear learning and extinction. Psychological Science 20, 198–206.

Martin, G.F., Holstege, G., Mehler, W.R., 1990. Reticular formation of the pons of the pons and medulla. In: Paxinos, G. (Ed.), The Human Nervous System. Academic Press, San Diego, CA, pp. 203–220.

Melzig, C.A., Holtz, K., Michalowski, J., Hamm, A.O., 2011. Interoceptive threat leads to defensive mobilization in highly anxiety sensitive persons. Psychophysiology 48, 745–754.

McTeague, L.M., Lang, P.J., Laplante, M.-C., Cuthbert, B.N., Strauss, C.C., Bradley, M.M., 2009. Fearful imagery in social phobia: generalization, comorbidity, and physiological reactivity. Biological Psychiatry 65, 374–382.

McTeague, L.M., Lang, P.J., Laplante, M.-C., Cuthbert, B.N., Shumen, J.R., Bradley, M.M., 2010. Aversive imagery in posttraumatic stress disorder: trauma recurrence, comorbidity, and physiological reactivity. Biological Psychiatry 67, 346–356.

McTeague, L.M., Lang, P.J., Wangelin, B.C., Laplante, M.-C., Bradley, M.M., 2012. Defensive mobilization in specific phobia: fear specificity, negative affectivity, and diagnostic prominence. Biological Psychiatry 72, 8–18.

Melzig, C.A., Weike, A.I., Zimmermann, J., Hamm, A.O., 2007. Startle reflex modulation and autonomic responding during anxious apprehension in panic disorder patients. Psychophysiology 44, 846–854.

Milad, M.R., Rauch, S.L., Pitman, R.K., Quirk, G.J., 2006. Fear extinction in rats: implications for human brain imaging and anxiety disorders. Biological Psychology 73, 61–71.

Miller, N.E., 1948. Studies of fear as an acquirable drive: I. Fear as motivation and fear-reduction as reinforcement in the learning of new responses. Journal of Experimental Psychology 28, 89–101.

Nitschke, J.B., Larson, C.L., Smoller, M.J., Navin, S.D., Pederson, A.J., Ruffalo, D., Mackiewicz, K.L., Gray, S.M., Victor, E., Davidson, R.J., 2002. Startle potentiation in aversive anticipation: evidence for state but not trait effects. Psychophysiology 22, 254–258.

Otto, M.W., Tolin, D.F., Simon, N.M., Pearlson, G.D., Basden, S., Meunier, S.A., Hofmann, S.G., Eisenmenger, K., Krystal, J.H., Pollack, M.H., 2010. Efficacy of D-cycloserine for enhancing response to cognitive-behavior therapy in panic disorder. Biological Psychiatry 67, 365–377.

Phelps, E.A., Delgado, M.R., Nearing, K.I., LeDoux, J.E., 2004. Extinction learning in humans: role of the amygdala and vmPFC. Neuron 43, 897–905.

Quirk, G.J., Mueller, D., 2008. Neural mechanisms of extinction learning and retrieval. Neuropsychopharmacology Reviews 33, 56–72.

Russel, J.J., Moskowitz, D.S., Zuroff, D.C., Sookman, D., Paris, J., 2007. Stability and variability of affective experience and interpersonal behavior in borderline personality disorder. Journal of Abnormal Psychology 116, 578–588.

Schimitel, F.G., Almeida, G.M., Pirol, D.N., Armini, R.S., Tufik, S., Schenberg, L.C., 2012. Evidence of a suffocation alarm system within the periaqueductal gray matter of the rat. Neuroscience 200, 59–73.

Sevenster, D., Beckers, T., Kindt, M., 2013. Prediction error governs pharmacologically induced amnesia for learned fear. Science 339, 830–833.

Walker, D.L., Davis, M., 1997. Anxiogenic effects of high illumination levels assessed with the acoustic startle paradigm. Biological Psychiatry 42, 461–471.

Walker, D.L., Davis, M., 2002. Light enhanced startle: further pharmacological and behavioral evaluation. Psychopharmacology 159, 304–310.

Walker, D.L., Ressler, K.J., Lu, K.-T., Davis, M., 2002. Facilitation of conditioned fear extinction by systemic administration or intra-amygdala infusions of D-cycloserine as assessed with fear-potentiated startle in rats. Journal of Neuroscience 22, 2343–2351.

Weike, A.I., Hamm, A.O., Schupp, H.T., Runge, U., Schroeder, H.W.S., Kessler, C., 2005. Fear conditioning following unilateral temporal lobectomy in humans: dissociation of conditioned startle potentiation and autonomic learning. Journal of Neuroscience 25, 11117–11124.

Weike, A.I., Schupp, H.T., Hamm, A.O., 2007. Fear acquisition requires awareness in trace but not delay conditioning. Psychophysiology 44, 170–180.

Weike, A.I., Schupp, H., Hamm, A.O., 2008. In dubio pro defensio: initial activation of conditioned fear is not cue specific. Behavioral Neuroscience 122, 685–696.

Fear: Psychological and Neural Aspects

Blake A Pellman and Jeansok J Kim, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.J. Kim, K.M. Myers, volume 8, pp. 5428–5433, © 2001, Elsevier Ltd.

Abstract

Fear is an innate, behavioral–motivational system that allows humans and animals to rapidly enact specific defensive behaviors in response to stimuli that the system, either through natural selection or learning, has come to recognize as potentially dangerous. The fear system likely evolved due to the reproductive success it conferred animals that could most aptly avoid or escape encounters with predatory or environmental threats. While fear to certain kinds of stimuli is unlearned and seemingly biologically prepared, fear can also be acquired rapidly and lastingly to stimuli associated with harmful outcomes, allowing animals to respond adaptively to new environmental threats. This article will present some of the behavioral and physiological aspects of fear that make it a useful model system for understanding the psychological and neurobiological mechanisms of emotion, learning, and memory, as well as the development of a variety of human fear-related disorders.

Fear – Definition and Background

All organisms are subject to threats and dangers from their environments, such as predation, dominance, and territoriality. Through the process of natural selection, evolution has endowed all organisms with defensive mechanisms (e.g., chemical toxins in plants; fight–flight responses in animals) for self-preservation. Fear serves this same purpose. In humans and other higher-order animals, 'fear' as a scientific term describes a behavioral–cognitive–emotional state in which a set of adaptive responses is activated in the presence of perceived danger. As the term is more commonly used, 'fear' refers to a sensation (or feeling) of strong unpleasant emotion aroused by stimuli or events that are perceived as threatening or dangerous.

This more anthropomorphic (psychological) sense of the word 'fear' may be useful in describing a person who is experiencing an environmental threat (e.g., snakes or extreme heights) and exhibiting certain stereotyped behaviors such as a fearful facial expression, autonomic arousal, behavioral readiness, and other defensive responses. However, it is probably inaccurate to use the term to describe the response of a goldfish (or any other lower-order animal) that is exhibiting similarly stereotyped behaviors (e.g., avoidance or escape) in response to painful physical stimuli such as electric shocks. In this article, fear will be defined as the coordinated pattern of brain activities, bodily changes, and adaptive behaviors that occurs within an animal in response to perceived threats and dangers.

A repertoire of innate fear responses that is characteristic of a particular animal species is often referred to as species-specific defensive reactions (SSDRs; Bolles, 1970) (**Figure 1(a)**). For instance, the universal expressions of fear in humans include particular facial expressions (seen in members of all cultures), changes in autonomic nervous system activities (e.g., increased heart rate, increased blood pressure, rapid respiration, perspiration, elevated pain threshold, pupil enlargement, and decreased digestive activity) and behavioral readiness (e.g., freezing, increased propensity to startle, and sharpening of cognition) (see LeDoux, 1996). These fear responses in humans are 'hardwired' (innate) as indicated by the observation that newborn infants, when exposed to aversive stimuli (e.g., a loud noise) for the very first time, exhibit responses such as sudden eye closure, rapid breathing, clutching of the hands, and crying (e.g., Watson and Rayner, 1920). The innateness of the fear response is also demonstrated in other animals; for example, laboratory-reared rats confronted with a live predator (e.g., a cat) or its odor for the very first time exhibit changes in the autonomic nervous system and behavioral readiness (responses that are remarkably similar to those seen in humans), and emit ultrasonic (22 kHz) distress vocalizations (a response that is an example of an SSDR) (Bolles, 1970). The advantage to survival, and therefore reproduction, conferred by such automatic responses is a logical consequence, but only recently have researchers begun to study the ecology of fear.

The ability to learn about potential threats or dangers is another important adaptive aspect of the fear system. For instance, although newborn infants innately exhibit fear to certain stimuli, they do not show innate fear to flame or height (two stimuli that most children learn and adults remember to avoid). A stimulus that initially does not evoke a fear or defensive response can develop that ability after the organism experiences the pain and/or physical stress associated with encountering that stimulus. This allows an organism to develop a rapid, automatic response to new threats in the environment. Fear learning, as opposed to reward learning (*see* Reinforcement: Neurochemical Substrates), can develop within as few as a single experience with the stimulus and its consequences. It can also be quite lasting, potentially for an entire lifetime, as is often seen with untreated cases of fear-related disorders (discussed later). Indeed, this calls into question the extent to which the fear system can generate adaptive behaviors, as fear-related disorders are often extremely debilitating to the individual.

Fear as an emotion has been a topic of interest to many subdisciplines of psychology, including psychoanalysis, behaviorism, cognitive psychology, and biological psychology. Historically, two opposing theories of emotion, the James–Lange theory and the Cannon–Bard theory, have been especially influential (**Figure 2**).

According to the James–Lange theory, threatening or dangerous stimuli trigger physiological arousal and behavioral responses (via the autonomic nervous system), which are

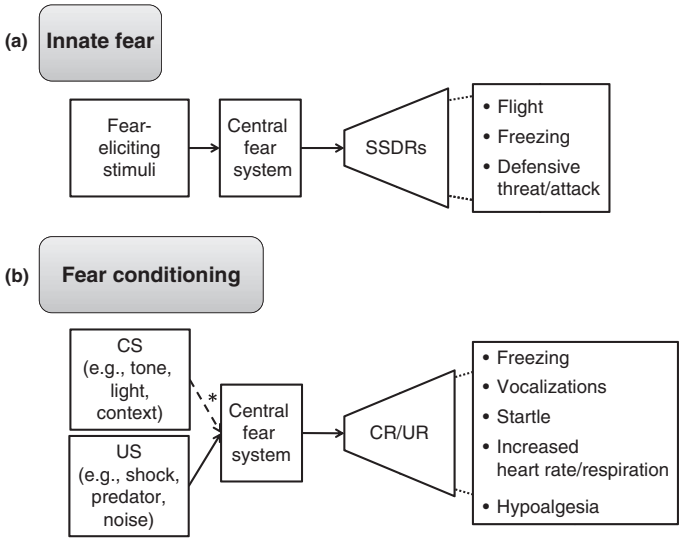

Figure 1 Simple models of innate fear and fear conditioning. (a) Innate fear: perceived danger triggers a coordinated set of programmed behavioral and physiological responses, which prepares the animal to contend with the potential threat. (b) Fear conditioning: fear of novel stimuli can be acquired through the process of classical conditioning. By virtue of association with the fear-eliciting US and its UR, the CS comes to activate a fear CR. *The putative locus of neural plasticity involved in fear conditioning.

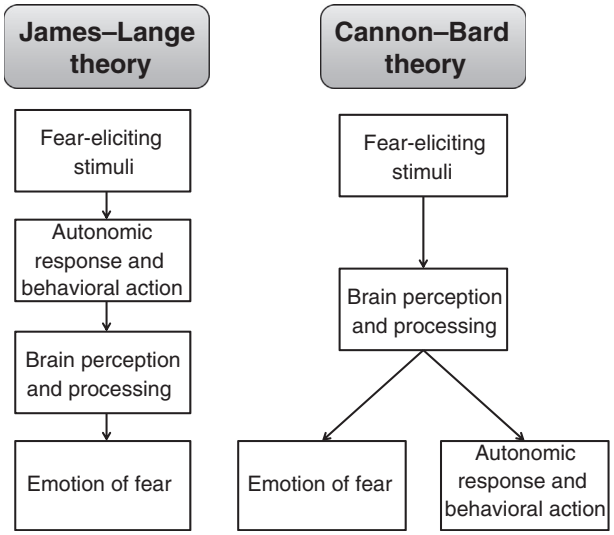

Figure 2 The James–Lange and Cannon–Bard theories propose different accounts by which threats and danger trigger fear emotions. The James–Lange Theory posits that fearful emotion is triggered when the brain detects the body responding reflexively to threatening or dangerous stimuli. In contrast, the Cannon–Bard Theory emphasizes the independence of fearful emotion from physiological arousal.

perceived and then processed as the experience of fear by the brain (emphasizing the importance of physiological feedback to the brain). In contrast, the Cannon–Bard theory posits that threats and dangers simultaneously trigger brain activities representing fear as well as initiate physiological arousal and behavioral readiness (emphasizing the importance of brain activity). Thus, the critical distinction between the two theories is whether the appearance of immediate danger (e.g., a bear) is postulated to activate the central fear system in parallel with fleeing and autonomic responses (Cannon–Bard theory), or to first trigger fleeing and autonomic responses, which are then detected and interpreted as fear by the brain (James–Lange theory). There is evidence in support of both theories, suggesting that central processes and bodily feedback both contribute to the emergence of fear as an emotion.

Because humans share many biological and behavioral features with other animals, experimental studies with nonhuman animals provide valuable information toward understanding the physiological as well as the psychological aspects of fear. Charles Darwin postulated in *The Expression of Emotions in Man and Animals* (1872) that emotions such as fear are innate, highly specific to situations, and the product of coordinated brain activities, and that different animal species with diverse environmental demands (or histories) have evolved dissimilar specialized sets of fear (or defensive) responses to maximize survival. It follows from this line of reasoning that more related species on the phylogenetic tree will exhibit more similar fear responses than will less related species, whose nervous systems share fewer characteristics. Indeed, studies have found similar patterns between humans and rodents in defensive responses to threatening stimuli and situations. Thus, it is highly likely that the neuronal substrates underlying fear (or other defensive mechanisms) in humans are more similar to those of other mammals (e.g., chimpanzees and rats) than those of nonmammalian animals (e.g., fish and birds).

Neuronal Substrates of Fear

The modern neurobiological study of emotion began in 1937 when James Papez published his description of a neuronal theory of emotion, connecting specific brain circuitry to emotional phenomenon, which came to be known as the Papez circuit. Intricate dissections of a human brain have found that this circuit begins in the hippocampus and runs through the fornix to

the mammillary bodies, to the anterior nucleus of the thalamus through the mammillothalamic tract, through the cingulum and to the entorhinal cortex, which connects back to the hippocampus to complete the circuit. In 1949, Paul D. MacLean expanded the Papez circuit in his conception of the limbic system (*see* Limbic System), a primitive but conserved autonomic emotional system, to include structures such as the amygdala, cingulate cortex, and areas of the prefrontal cortex. However, attempts to classify a specific, discrete circuit subserving emotional processes have proven to be difficult, though the efforts have generated a vast amount of information regarding the specific properties of emotion, particularly fear responses. Several lines of evidence point to temporal lobe structures (specifically the amygdala, an almond-shaped group of nuclei buried deep within the temporal lobes) as the key neural system underlying the central state of fear (see LeDoux, 1996). The classic studies of Klüver and Bucy (1937), who examined the effect of temporal lobe ablations in rhesus monkeys, were pioneering in this respect. Klüver and Bucy noted dramatic postlesion behavioral changes, including hyperactivity, hypersexuality, and a so-called 'psychic blindness' that was manifested as an apparent indifference to previously frightening and/or arousing stimuli. While the monkeys were wild and aggressive prior to surgery, they were remarkably tame and agreeable to human contact afterward, suggesting that a crucial emotive brain center had been removed or significantly affected by the lesions. Subsequent work, such as that of MacLean and Delgado (1953) using focal stimulation of temporal lobe structures in awake, behaving monkeys, pointed to the amygdala as the critical site. The importance of the amygdala in both innate fear responses and fear learning has been reconfirmed in numerous studies employing diverse species of animals (e.g., rats, mice, cats, rabbits, monkeys, humans) and various experimental techniques (e.g., lesions, stimulation, drug infusions, recordings), and most current theories of amygdalar function reflect the central role of the structure in the processing of fearful stimuli and enacting defensive behaviors. Consistent with this view, human neuropsychological and brain-imaging studies have found that patients with amygdalar damage are selectively impaired in the recognition of facial expressions of fear, and that there is a significantly increased amount of blood flow to the amygdala (as measured by functional magnetic resonance imaging) when normal subjects are presented with pictures of fearful faces (Adolphs, 2010).

The amygdala, as one of the principal structures of the limbic system, has access to sensory input from diverse areas of the brain (e.g., olfactory cortex, hippocampus, thalamus, hypothalamus, neocortex), and projects to various autonomic and somatomotor structures that mediate specific fear responses (e.g., bed nucleus of stria terminalis for activating stress hormones, periaqueductal gray matter for defensive behavior, lateral hypothalamus for sympathetic activation). Several subnuclei within the amygdala have been identified as having distinct functions, but the functions have yet to be fully characterized. The lateral amygdala (LA) receives input from the thalamus about sensory stimuli, and projects to the central nucleus (CeA) of the amygdala. Other subregions of the amygdala include the medial (MeA), basal (BA), and intercalated (ITC) nuclei. Different nuclei appear to be involved depending on the sensory modality of the fear-eliciting stimulus, and they each project to different structures throughout the brain, such as the hypothalamus and periaqueductal gray.

Given this pattern of connectivity, it is not surprising that damage to the amygdala eliminates stereotypical fear responses, such as defensive freezing (characterized as a motionless, crouching posture that functions to reduce the probability of detection by predators) exhibited by rodents in the presence of predators or perceived threats (Lavond et al., 1993). Nevertheless, while the amygdala is apparently an important component of the central fear-processing system, it does not necessarily follow that the amygdala is the seat of fear as an emotion (or a feeling). Given its intricate reciprocal connections with the neocortex, a structure implicated in higher cognitive functioning, it is very likely that the interactions between amygdala and neocortex (particularly sensory and limbic association cortices) are responsible for the emotion of fear. Thus, although the amygdala is essential in activating various fear responses (or SSDRs) both in humans and in rats, the fact that there are great differences in the neocortices of humans and rats suggest that perhaps there might be important differences in fear as experienced by the two species. Nonetheless, novel methods of physiological measurement and ways to analyze behavior are constantly being generated to solve the unanswered questions about fear and the brain.

Contemporary Fear Research

Classical or Pavlovian fear conditioning has been widely employed for studying the mechanism(s) by which fear is acquired and processed in the brain. Fear conditioning occurs when initially neutral 'conditional' stimuli (CSs), such as tones, lights, or distinctive environmental settings, are contingently paired with aversive, 'unconditional' stimuli (USs) such as electric shocks or loud noises, which reflexively activate unconditional fear responses (URs). Through CS–US association formation, the CS comes to elicit various conditional fear responses (CRs) that are similar to innate fear responses (**Figure 1(b)**). Perhaps the best known example of fear conditioning is the Little Albert experiment by Watson and Rayner (1920). Little Albert was an 11-month-old infant who initially exhibited curiosity (no fear) to a white rat by touching and playing with it. Watson and Rayner attempted to condition Albert to fear the rat by striking a steel bar with a hammer behind his head (US) as his hand touched the rat, causing him to startle, fall forward, and cry (UR). A week later, when the rat (CS) was placed near Little Albert's hand, he withdrew his hand and began to cry (CR). This fear of the rat was also alleged to be observed in the presence of other white furry animals and objects (e.g., rabbits, dogs, fur muffs), indicating that acquired fearful associations can generalize to other similar stimuli.

Modern investigations of fear conditioning typically employ rats as subjects and use a tone as a CS and a mild footshock as a US. Under these circumstances, a single CS–US pairing results in robust (rapid and enduring) fear learning, as evidenced by a variety of conditional fear responses exhibited upon subsequent presentations of the tone. Thus, the biological significance of fear conditioning is reflected by the quick and nearly permanent nature of changes in the behavior toward

the initially neutral stimulus that has been paired with an aversive experience.

Fear can also be rapidly acquired through instrumental or operant conditioning, in which the presentation of an aversive stimulus is contingent upon the behavior of the animal. A widely employed procedure with rodents is the passive (or inhibitory) avoidance task. In this paradigm, the animal's response (e.g., entering a dark compartment of a box when placed in an adjacent lighted compartment, or stepping down from a platform onto a grid floor) is paired with an aversive experience (e.g., a footshock). As a function of this response–stimulus pairing, the animal learns to avoid making the response that was followed by the aversive experience.

While the benefits to survival conferred by fear learning are obvious, the mechanisms underlying the associations between CS and UR have been evolutionarily constrained in the kinds of associations that can be formed. For example, Garcia and Koelling (1966) demonstrated that while rats can associate a light/sound with a footshock as well as tastes with illness, they cannot learn associations between tastes and footshocks or lights/sounds with illness. The mechanisms underlying different kinds of fear learning have evolved as a consequence of common, natural associations encountered by ancestral organisms and do not have the facility to form associations between just any CS–US pairing. This principle of 'selective association' is important to consider when trying to understand the neuronal substrates of fear learning, as it suggests there are different neuronal circuits underlying the learning of different kinds of defensive responses.

Similar to innate fear, evidence from lesion, pharmacological, recording, and stimulation studies indicates that the amygdala is essential not only to innate fear, but also to the acquisition of fear, as with fear conditioning (see Lavond et al., 1993; LeDoux, 1996). Thus, amygdalar lesions (particularly lesions of the CeA) impair the acquisition and retention of various fear-related behaviors in rats, rabbits, and cats. Neurons in the amygdala are responsive to both CSs and USs and undergo plastic changes during fear conditioning. Furthermore, restricted stimulation of the amygdala elicits specific fear responses (e.g., freezing). A novel and increasingly popular method is 'optogenetics,' which has allowed a high degree of neuronal specificity for studying the substrate of fear. This technique employs viral gene insertions to manipulate genetic transcription within neurons so that they generate receptors that are uniquely sensitive to light stimuli, thus permitting researchers to activate very specific neuronal circuits or subpopulations with implanted optical fibers (see Deisseroth, 2011). In a recent study, Johansen et al. (2010) used optogenetic stimulation in rats to demonstrate that activation of neurons within the LA drives the association between an innocuous stimulus and a fear response during fear learning.

Recent findings from human studies also indicate that the amygdala is critically involved in fear learning and fear responses. Ralph Adolphs and colleagues identified a patient (SM) as having rare, bilateral amygdalar lesions who exhibited a remarkable lack of overt and subjective fear responses to common fear-eliciting stimuli such as live snakes and spiders. Furthermore, many studies show that blood flow to the amygdala is increased during fear conditioning in normal subjects, further adding to the evidence that the amygdala is crucially involved in fear learning (see Adolphs, 2010).

One hypothesis of amygdalar function in fear conditioning is presented schematically in **Figure 3**. In this model, an emotional stimulus (such as a tone that has been paired with a footshock) is presented, and its sensory characteristics are processed by way of two parallel pathways that ultimately converge in the amygdala. One of these pathways is 'direct,' and provides relatively 'raw,' unprocessed information about the stimulus very quickly. This pathway projects through the thalamus to the amygdala. The second pathway is 'indirect,' and provides highly processed information about the stimulus but does so at the cost of an increased processing time (LeDoux, 1996). If the neocortex is the seat of conscious awareness, the thalamic pathway to the amygdala can be considered the subconscious (nondeclarative or implicit) fear route while the cortical pathway can be considered the conscious (declarative or explicit) fear route. This cortical pathway consists of projections through the primary sensory cortices to higher-order association areas as well as the hippocampal formation, which then projects to the amygdala. Upon reaching the amygdala (specifically the lateral/basolateral nuclei), the two converging representations of the stimulus are subject to additional processing within the amygdala itself. Most current theories assume that the essential plasticity underlying this form of learning occurs in the amygdala and corresponds to a neural instantiation of the association between the CS (the tone) and the US (the footshock) long assumed by psychologists to lie at the heart of classically conditioned behavior. This plasticity is widely believed to be mediated by the synthesis of new proteins within amygdalar neurons (*see* Synaptic Mechanisms and Protein Synthesis in Learning, Memory). The regions involved in forming and maintaining CS–US

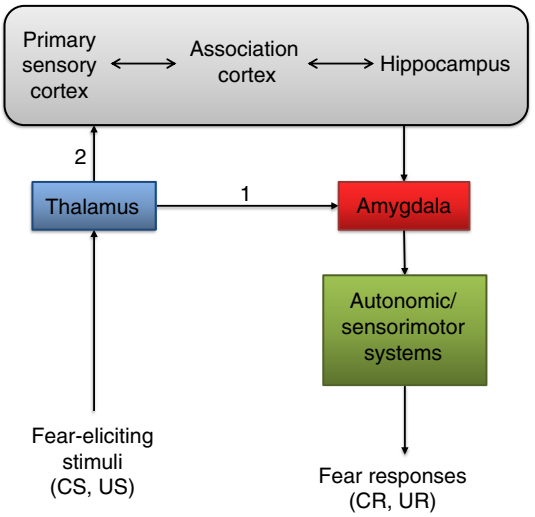

Figure 3 A simplified putative fear-conditioning circuit. Conditional stimuli (e.g., tone or light) are processed via two separate pathways to the amygdala, where associations between the US and CS are thought to form. One pathway is via the direct (1) thalamoamygdalar projections (subcortical) and the second is via the indirect (2) corticoamygdalar projections (cortical).

associations are still unclear, though new clues are being discovered by researchers investigating how humans and animals update memories of fearful associations in a process called 'reconsolidation.'

Reconsolidation, though a controversial topic among researchers, is theorized to occur when a previously acquired long-term memory trace is reactivated (or retrieved), becomes labile, and is therefore subjected to new modifications, likely through changes in neuron-specific protein transcription. The nature of the modification may be to strengthen the memory, to change the properties of the stimulus associated with a given aversive outcome, or to weaken the association (as in 'unlearning'). This hypothesized process of unlearning, which is essentially an erasure of the original memory, might be possible through blocking reconsolidation of a memory after it has become labile (see Section Fear Pathologies and Clinical Implications). Unlearning is distinct from inhibition of the original memory (i.e., the conditional response to a CS), which requires new learning and the original memory remains intact. This process of learned inhibition is called 'extinction' and occurs when the CS no longer predicts the US (Pavlov, 1927). This is readily modeled in fear-conditioning experiments when a US stops occurring with a CS, and, over many trials, the animal reduces the magnitude and frequency of its fear responses to the CS. The animal learns that it no longer needs to be afraid of the CS, and the conditional response is 'extinguished.' Extinction learning permits an animal to readapt to environments where particular threats or dangers are no longer present, providing new opportunities for foraging and mating. There are limitations of extinction learning, as it is highly context specific and readily overridden by renewal of the original fear response.

Extinction learning is largely attributed to the workings of the prefrontal cortex, a relatively new structure in the evolution of the mammalian brain, which has many intricate, reciprocal connections with the amygdala. Two areas in particular have been implicated in regulating the amygdala: the infralimbic cortex and the prelimbic cortex. Stimulation and inactivation studies have begun to characterize these two areas as having opposing influences on the amygdala (Quirk et al., 2010). The infralimbic cortex carries inhibitory projections to the amygdala, apparently driving extinction learning, but the prelimbic cortex has an excitatory effect on the amygdala, increasing fear responses. The precise functions of the inputs to and outputs from the amygdala, as well as the intraamygdaloid processing of CS- and US-related information, remain a topic of intensive research.

Fear Pathologies and Clinical Implications

Fear conditioning and extinction serve as a model system for understanding not only the neural basis of acquired fear in nonhuman animals, but also the etiology of a variety of fear-related disorders in humans (e.g., anxiety, phobias, post-traumatic stress disorder (PTSD), and panic disorder). Since the characteristic behavioral and physiological symptoms of human anxiety disorders closely resemble many of the fear responses observed in fear-conditioning paradigms, it is possible that anxiety and fear are closely related (**Table 1**) (*see* Anxiety and Fear Conditioning, Neural Basis of). It is important to note, however, that the fact that the symptomatic manifestation of anxiety disorders is similar to the fear responses observed in fear conditioning does not necessarily mean that the underlying neuronal substrates of fear and anxiety are the same (post hoc, ergo propter hoc). While anxiety disorders can arise unpredictably, it is known that most patients experience their first symptoms in early adult life, indicating that maladaptive anxiety may be at least partly acquired. Indeed, clinicians from the time of Freud have speculated that anxiety patients have acquired their disorder as a result of traumatic learning experiences. Since the paradigm *par excellence* of traumatic learning is fear conditioning, fear conditioning appears to be useful in understanding the genesis of anxiety disorders (**Figure 3**).

The question of whether anxiety is analogous to fear conditioning might be one of semantics; the difference is between normal, adaptive fear and abnormal, maladaptive fear (anxiety, phobias). However, there is an important distinction between anxiety disorders and fear conditioning (e.g., Seligman, 1971) that might be taken to suggest that the two have very different bases. That is, while conditioned fear responses may be extinguished by, for example, actively preventing an animal from making escape or avoidance responses to a danger-signaling stimulus, fear-related disorders in humans (e.g., phobias) are quite resistant to extinction. In fact, anxiety disorders are characterized by fears that the patients recognize to be irrational but are unable to control.

Based on **Figure 3**, it is interesting to speculate that fear-related disorders (irrational fears) might be due to unbalanced activities in the cortical and thalamic pathways to the amygdala. For example, in most people a snake seen behind a glass window will elicit a different response than will a snake encountered in the woods. Both cortical and subcortical routes will process information about the snake, but the cortical (conscious) route will integrate the contextual information of the glass window and assess the safety of the situation (possibly

Table 1 Behavioral and physiological responses associated with anxiety disorders and fear conditioning

Anxiety disorders	*Fear conditioning*
Palpitations, accelerated heart rate	Increased blood pressure, tachycardia
Sensations of shortness of breath or suffocating	Increase in respiration, shallow breathing
Sweating	Increase in electrical conductance of skin
Nausea or abdominal distress	Suppressed appetite
Paresthesia (numbing or tingling sensations)	Hypoalgesia or analgesia
Hyperreactivity	Potentiated startle responses

acting as a gate). However, if the cortical pathway is somehow disrupted, the subcortical activity will be left unchecked to activate fear responses, even in an inappropriate context. It is possible that this type of unregulated subconscious activation of fear contributes to the etiology of fear-related disorders.

Figure 4 illustrates a hypothetical model of normal, adaptive fear and abnormal maladaptive fear. It is conceivable that genetic predisposition and environmental factors (e.g., stress) interact in such a way as to disrupt the normal balance between cortically activated fear and subcortically activated fear, leading to fear-related disorders. If it is then supposed, as some have suggested (LeDoux, 1996) that extinction is achieved through cortical inputs to the amygdala, and direct thalamic inputs are relatively resistant to extinction, it follows that fear disorders based on overactivity in the subcortical pathway should be 'irrational' in the sense that they do not subside even when the patient recognizes them as such. A more comprehensive understanding of the neuronal substrates and, subsequently, the molecular mechanisms of the fear system will be useful in understanding and developing treatments for fear-related disorders.

Several clinical populations have provided valuable information for understanding the contribution of different brain regions to fear-related disorders. Patients with Urbach–Wiethe disease, a rare, recessive genetic condition characterized in part by bilateral calcifications of the interior medial temporal lobes (particularly of the amygdalae), are cognitively comparable to normal patients, though they typically exhibit deficits in emotional processing (Adolphs, 2010). Patients with phobias exhibit a persistent fear to a specific object or situation (e.g., spiders, snakes, water, heights) and actively avoid encountering such object or situation, though often even the patients recognize their fears to be irrational, exaggerated responses. Phobias can develop as a result of traumatic experiences (e.g., being bitten by a snake, falling from a high place), or, as seen with innate fear responses, be biologically hardwired, though it is uncertain for what fears humans may be predisposed (see Seligman, 1971). Patients diagnosed with PTSD exemplify the interactions between stress, anxiety, and fear quite distinctly. PTSD develops as a consequence of direct or indirect (witnessing) traumatic experiences such as natural disasters, physical or sexual assault, or war; aspects of which (e.g., smells, sounds, and/or locations) become fear- and anxiety-inducing stimuli that are particularly resistant to extinction. Though resembling phobias in their persistence, such stimuli need only be reminiscent of the original trauma to elicit strong fears and anxiety in those suffering from PTSD. Treatment for patients suffering from fear-related disorders may be achieved through cognitive–behavioral and pharmacological interventions. Drawing heavily from the extinction literature, exposure therapy involves exposing the patient to their object or situation of fear in controlled, nonthreatening environments in order to dissociate the stimulus from the fear of danger. Extensive exposure therapy over the patient's lifetime is often needed to maintain an extinguished fear since the fears can be spontaneously reinstated. Other therapies often utilize structured social support groups and routine-based behavioral changes to help patients control their fear and anxiety. Pharmacological therapies have also drawn on extinction literature, using drugs that enhance inhibitory learning in order to extinguish fear responses. The majority of pharmacological treatments involves anxiolytic (anxiety-reducing) drugs, and may work best for managing the symptoms of a disorder in combination with cognitive–behavioral therapies. More recent pharmacological approaches have attempted to either block fear memories from forming (during initial consolidation) or erase fear memories after they form by blocking reconsolidation. Immediate application of beta blockers (e.g., propranolol) after a traumatic experience may prevent abnormal fear memories of the trauma from developing, though their success seems extremely time sensitive. Intensive research is currently being conducted to identify pharmacological and behavioral therapies that may block or disrupt reconsolidation and erase previously acquired fear memories.

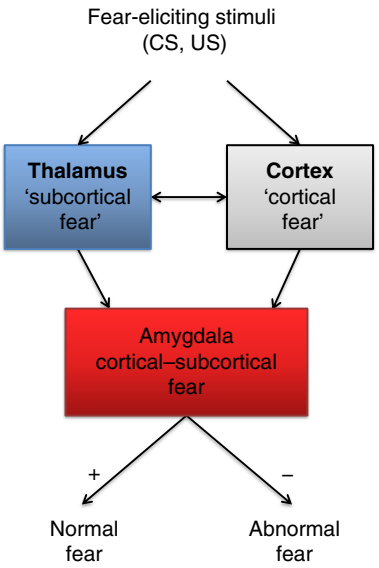

Figure 4 A simple model of normal and abnormal fear based on the putative circuitry of cortical and subcortical fear-conditioning pathways. The subcortical fear route provides a quick but imprecise representation of sensory information for activating fear (a nonconscious pathway), whereas the cortical fear route provides relatively slower but more detailed information for activating fear (a conscious pathway). In normal fear expression, there is a balance between the cortical fear and subcortical fear pathways, such that the cortical component is stronger than or equal to the subcortical component. Abnormal fear might arise when this balance is disrupted such that the activity of the subcortical fear system overwhelms the activity of the cortical fear system.

Conclusion

The study of fear is but one means of understanding emotion, but it is the most fruitful one to have been employed to date. This is likely due to the primacy of fear: unlike other emotions, fear is apparently shared by a wide variety of species. This ubiquity speaks to its importance as a mechanism for self-preservation. Thus, fear has been the intense focus of researchers over the past century, and much has been learned about the biological substrate underlying innate and learned fear. However, precise understanding of how the various brain

regions known to be involved interact with each other to collectively function as a system – identifying potentially threatening stimuli, determining the appropriate response, and enacting that response – remains a topic of lively debate. Furthermore, little is known about the ways in which fear influences the daily behavior of many animals, including humans, though clinical populations clearly demonstrate such effects. Indeed, it seems that therapies for fear-related disorders have room for improvement yet, promising a future of continued research from diverse areas of study.

See also: Anxiety and Fear Conditioning, Neural Basis of; Emotion, Neural Basis of; Fear-Potentiated Startle; Hippocampus and Related Structures; Learning and Memory, Neural Basis of; Limbic System; Reinforcement: Neurochemical Substrates; Synaptic Mechanisms and Protein Synthesis In memory.

Bibliography

Adolphs, R., 2010. What does the amygdala contribute to social cognition? Annals of the New York Academy of Sciences 1191, 42–61.

Bolles, R.C., 1970. Species-specific defensive reactions and avoidance learning. Psychological Review 71, 32–48.

Darwin, C., 1872. The Expression of the Emotions in Man and Animals. Murray, London.

Deisseroth, K., 2011. Optogenetics. Nature Methods 8, 26–29.

Garcia, J., Koelling, R.A., 1966. Relation of cue to consequence in avoidance learning. Psychonomic Science 4, 123–124.

Johansen, J.P., Hamanaka, H., Monfils, M.H., Behnia, R., Deisseroth, K., Blair, H.T., LeDoux, J.E., 2010. Optical activation of lateral amygdala pyramidal cells instructs associate fear learning. Proceedings of the National Academy of Sciences of the United States of America 108, 12692–12697.

Klüver, H., Bucy, P.C., 1937. 'Psychic blindness' and other symptoms following bilateral temporal lobectomy in rhesus monkeys. American Journal of Physiology—Regulatory Integrative and Comparative Physiology 119, 352–353.

Lavond, D.G., Kim, J.J., Thompson, R.F., 1993. Mammalian brain substrates of aversive classical conditioning. Annual Review of Psychology 44, 317–342.

LeDoux, J., 1996. The Emotional Brain: The Mysterious Underpinning of Emotional Life. Simon and Schuster, New York.

MacLean, P.D., Delgado, J.M.R., 1953. Electrical and chemical stimulation of frontotemporal portion of limbic system in the waking animal. EEG Clinical Neurophysiology 5, 91–100.

Papez, J.W., 1937. A proposed mechanism of emotion. Archives of Neurology and Psychiatry 38, 725–743.

Pavlov, I.P., 1927. Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex. Oxford University Press, London.

Quirk, G.J., Paré, D., Richardson, R., Herry, C., Monfils, M.H., Schiller, D., Vicentic, A., 2010. Erasing fear memories with extinction training. The Journal of Neuroscience 30, 14993–14997.

Seligman, M.E.P., 1971. Phobias and preparedness. Behavioral Therapy 2, 307–320.

Watson, J.B., Rayner, R., 1920. Conditioned emotional reactions. Journal of Experimental Psychology—Human Perception and Performance 3, 1–14.

Federalism

Arthur Burris, California State University, Hayward, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Burris, volume 8, pp. 5440–5442, © 2001, Elsevier Ltd., with revisons made by the Editor.

Abstract

Federalism refers to a set of institutional arrangements dividing power between a central government and regional governments. The term also refers to a philosophy of government that advocates federal arrangements as a desirable system of government. The relationship between federalism and related terms such as unitary government, federal system, and intergovernmental relations is clarified. The normative and empirical branches of the study of federalism are distinguished. Major arguments for and against federal systems of government are reviewed. Empirical studies of federalism are divided into those focusing on description and those focusing on causality. Within the descriptive realm, the core features of federal systems are delineated, and major federal systems of government are listed. Scholarly efforts to characterize variation within federal systems are described. The causes of the formation and dissolution of federal systems are discussed, along with efforts to understand the causes and consequences of change within federal systems. Areas likely to be of particular emphasis in future research are indicated.

'Federalism' refers to both an arrangement of political institutions and a philosophy of government. Federalism as a set of institutional arrangements denotes a political system in which power is shared between a central government and regional governments. Federalism as a philosophy of government holds that such federal structures are a desirable governmental form.

Federalism and Related Terms

The distinction between 'federalism' and related terms should be clarified. A government displaying the institutional characteristics of federalism is referred to as a 'federal system,' and the institutional structures themselves are described with the adjective 'federal.' A federal system is contrasted with a 'unitary system' of government in which all major decisions are made by the central government; some degree of administrative decentralization, however, may be present in a unitary system. The terms 'federalism' and 'federal system' have been used as overarching terms which encompass a variety of more specific types of political arrangements, including federations, confederacies, and leagues (Riker, 1975). The origins of the word 'federal' can be traced to the Latin *foedus*, denoting a league, treaty, or covenant. The term 'intergovernmental relations' refers to interactions between different governmental units within a political system. This term includes but is not limited to the study of interactions between central and regional governments within federal systems.

The Normative Study of Federalism

The study of federalism can be usefully divided into normative and empirical branches. The normative branch, which dates back centuries and remains very important today, includes the arguments in favor of the US constitution put forth in *The Federalist*. This normative literature consists of systematic consideration of federalism as a philosophy of government, including the enumeration of possible advantages and disadvantages of federal systems.

Proponents of federalism argue that federal systems are desirable because they may produce a number of advantageous consequences. In particular, advocates of federalism suggest that federal systems protect citizens against tyranny by dividing power among central and regional governments; expand possibilities for political participation by increasing the number of governments within a political system; allow for regional variations in public policy that are reflective of geographic differences in citizen preferences; and provide multiple governmental 'laboratories' in which innovative policy responses to public problems can be tested (Shapiro, 1995).

Other scholars have argued that federal systems may produce a variety of negative consequences. For example, federal systems may encourage regional inequalities in the distribution of the costs and benefits of public policies; allow for the oppression of local minorities by local majorities within the subnational jurisdictions; and produce an undesirable 'race to the bottom,' in which regional governments attempt to attract economic enterprises by lowering regulatory standards and reducing governmental taxation and expenditure (Shapiro, 1995).

The Empirical Study of Federalism

The empirical branch of the study of federalism has focused on the description and measurement of federal systems, and also on the establishment of relevant causal relationships.

Descriptive Approaches

In the descriptive realm, scholars have been concerned above all with identifying the core characteristics of federalism, characterizing variation in federal systems, and describing the mechanisms through which federal systems operate. At the

heart of the federal idea is the division of powers and responsibilities between an overarching central government exercising authority over a broad territory, on the one hand, and a number of constituent governments each exercising authority over a particular subset of that territory, on the other. This division of power is generally enshrined in a constitution that guarantees the existence of all of the governments and provides at least a rough outline of the division of power among them.

Within these broad parameters, scholars have offered a range of more specific definitions of federalism, and no universal agreement about the term's precise definition has been reached. Different definitions produce somewhat differing classifications of political systems into federal and nonfederal categories. The result is that the term 'federal' can be applied to a variety of political systems in which power is shared between general and constituent governments. The political institutions of ancient Israel and the Greek city-states displayed federal characteristics. The US federal system, established in 1789, has served as the prototype of modern federalism (Elazar, 1968). At the turn of the twenty-first century, significant federal elements existed in political systems across the globe, including Argentina, Australia, Austria, Belgium, Brazil, Canada, the European Union, Germany, India, Malaysia, Mexico, Nigeria, Pakistan, the Russian Federation, South Africa, Spain, Switzerland, the USA, and Venezuela (Elazar, 1994).

Among federal systems, there is variation in the degree and character of the distribution of power between national and regional governments. This variation can be observed both in formal constitutional arrangements and in the working relationships among different levels of government within a political system. One component of the variation between federal systems, and within individual systems over time, is the extent to which powers and responsibilities rest with the central government or with the regional governments. Federal systems also vary in the extent to which these powers are either strictly divided between levels of government or shared between those levels of government.

With respect to the first of these dimensions, scholars have not yet reached a consensus about a precise way to measure the relative centralization and decentralization of federal systems (Watts, 1998). The formal distribution of powers set forth in constitutions may provide some insight into the comparative degree of centralization between systems. Such a constitutional focus has limitations, however, and is unlikely to be as useful in measuring changes over time within a federal system. Here, close attention must be paid to legislative and judicial decisions altering the balance of power between the central government and the constituent governments. One summary measure that is commonly examined is the proportion of governmental expenditures undertaken by the central government, either in total or in particular policy areas (Peterson, 1995). Efforts to characterize variation in the distribution of powers between general and constituent governments in federal systems have also led to the development of a wide range of metaphors and conceptual models (e.g., 'dual federalism,' 'cooperative federalism'). These metaphors and models attempt to characterize either the degree of centralization in a federal system or the extent to which powers are strictly divided or cooperatively shared between levels of government, or both. In all, Stewart (1982) found more than 300 such models and metaphors that had been proposed in the literature.

Descriptive studies have also examined, both statically and dynamically, the mechanisms through which federal systems operate. Core areas of concern here include the constitutional, legislative, and judicial processes through which powers are allocated between levels of government, and the fiscal (e.g., grants-in-aid), regulatory, and political instruments through which governments within a federal system influence each other's actions (Wright, 1988).

Causal Approaches

The causal branch of the empirical study of federalism addresses two major areas of concern: the causes and consequences of the establishment and dissolution of federal systems, and the causes and consequences of changes in how powers are divided between governments within federal systems.

Federal systems can be established either by joining together previously independent governments, or by decentralizing power in a previously unitary state. The formation of federal systems involves somewhat different processes and conditions in each of these two cases. In all cases, however, federal systems emerge under political conditions that contain strong pressures for both political unity (in the form of a national government) and political autonomy (in the form of regional governments with meaningful powers).

The pressures for unity most commonly result from a desire for military or economic advantage that constituent governments feel cannot be achieved independently. When federal systems are formed from previously independent governments, the reluctance of these governments to cede too much power to a newly formed central government creates pressure for regional autonomy. More generally, pressures for autonomy most commonly involve protecting constituent governments' diverse interests, whether based on nationality, language, ethnicity, religion, economy, or other variables. Against the backdrop of these general dynamics, scholars have debated what conditions are necessary and sufficient to produce a federal system, with particular discussion of cultural, economic, political, social, and historical variables (Riker, 1975).

Scholars have sought to understand not only why federal systems are formed, but also what the consequences of federal systems are. For example, federal systems can have important consequences for the management of ethnic conflict, either mitigating or exacerbating ethnic tensions, depending on the particular institutional arrangements in place.

The late twentieth century witnessed the disintegration of some political systems that were federal in principle or practice, including the Soviet Union, Yugoslavia, and Czechoslovakia. The survival of federal systems is imperiled when institutional arrangements reinforce rather than crosscut existing lines of division within the system, or when one group or constituent government representing a supermajority dominates the overall system (Watts, 1998; Elazar, 1968).

At a less macroscopic level, the causal study of federalism focuses to a significant extent on the evolution of federal

systems: why the distribution of powers changes in federal systems, and with what consequences. There are numerous studies of the proximate impact of legislative, judicial, and executive decisions on the distribution of powers in federal systems. Other analysts have focused on the role of underlying forces such as changes in technology, culture, ideology, society, and economy (Anton, 1984).

In all, there is significant dissatisfaction among scholars with the current state of our ability to construct robust theories of why federal systems evolve in the ways that they do (Beam et al., 1983). Explanations focusing on proximate causes have been criticized for being atheoretical and for too often relying on single case studies. Analyses focusing on underlying causes have not produced a clear consensus (Anton, 1984). Understanding the causes of centralization and decentralization in federal systems will remain an important area of investigation in the decades ahead. Scholars may increasingly attempt to link explanations focusing on proximate and underlying causes, constructing theories that specify how underlying forces such as culture or economy might in turn produce centralizing or decentralizing actions by proximate institutions such as legislatures or courts.

Scholars have also discussed the consequences of centralization and decentralization within federal systems. For example, there is evidence that decentralized systems make decentralized political party structures more likely (Riker, 1975). In another realm, some scholars have argued that decentralization of redistributive and regulatory functions may lead to a competitive spiral in which subnational governments reduce their interventions in these areas (Peterson, 1995).

Future Directions

Several areas of the study of federalism will likely be of particular importance in the first decades of the twenty-first century, as the literature continues to respond to important political developments. Scholars will be particularly interested in: the interaction between economic globalization and federal structures the emergence and evolution of supranational federal arrangements, including the European Union and the relationship between federal structures and the management of conflict in multiethnic political systems.

Editor's Note

Political scientists continue to debate the virtues of the federalist form of government and of federalism as a normative or ideological position. See Alison L. LaCroix, 2010; see also Daniel L. Rubinfeld and Robert P. Inman, 2014.

See also: Electoral Systems; Political Representation; Political Thought, History of.

Bibliography

Anton, T.J., 1984. Intergovernmental change in the United States: An assessment of the literature. In: Miller, T. (Ed.), Public Sector Performance: A Conceptual Turning Point. Johns Hopkins University, Baltimore, MD, pp. 15–64.

Beam, D.R., Conlan, T.J., Walker, D.B., 1983. Federalism: The challenge of conflicting theories and contemporary practice. In: Finifter, A. (Ed.), Political Science: The State of the Discipline. American Political Science Association, Washington, DC, pp. 247–279.

Elazar, D.J., 1968. Federalism. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences, vol. 5. Macmillan, New York, pp. 353–367.

Elazar, D.J. (Ed.), 1994. Federal Systems of the World, second ed. Longman, Harlow, UK.

LaCroix, Alison L., 2010. The Ideological Origins of American Federalism. Harvard University Press, Cambridge, MA.

Peterson, P.E., 1995. The Price of Federalism. The Twentieth Century Fund, New York.

Riker, W.H., 1975. Federalism. In: Polsby, N.W., Greenstein, F.I. (Eds.), Handbook of Political Science, vol. 5. Addison-Wesley, Reading, MA, pp. 93–172.

Rubinfeld, Daniel L., Inman, Robert P., 2014. Economics of Federalism. In: Oxford Handbook of Law and Economics. Oxford University Press.

Shapiro, D.L., 1995. Federalism: A Dialogue. Northwestern University, Evanston, IL.

Stewart, W.H., 1982. Metaphors, models, and the development of federal theory. Publius: The Journal of Federalism 12, 5–24.

Watts, R.L., 1998. Federalism, federal political systems, and federations. Annual Review of Political Science 1, 117–137.

Wright, D.S., 1988. Understanding Intergovernmental Relations, third ed. Brooks/Cole, Pacific Grove, CA.

Female Genital Mutilation

Aud Talle, University of Oslo, Oslo, Norway

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 8, pp. 5447–5451, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Female genital mutilation (FGM) is a collective term for various surgeries performed on young girls while growing up, often performed as part of initiation rituals into adulthood. Other frequently used terms are female genital cutting (FGC) and female circumcision (FC). The article gives an overview of the practice: its extent and distribution, operation practices, historical roots, and cultural explanations. FGM is particularly widespread in northeastern Africa and in parts of eastern and western Africa. It is estimated that some 130 million women in the world today are genitally mutilated. The most common form of mutilation practice is clitoridecomy. During the last few decades large-scale immigration to Europe and the USA from African countries has extended the geographical location of the practice.

Female genital mutilation is a collective term for various surgeries performed on young girls, often done as part of initiation rituals into adulthood. Other terms commonly used are female genital cutting, female circumcision, clitoridectomy, infibulation or Pharaonic circumcision, excision, and *sunna*. Except for female genital cutting and female circumcision which also are general terms and used interchangeably with female genital mutilation, the other terms are more specific, indicating the particular type of operations being performed.

Extent and Distribution

Sources estimate that perhaps as many as 130 million women in the world today have been mutilated genitally (Toubia, 1995). Since relatively little extensive statistical research has been conducted on the topic, prevailing figures quoted tend to vary. Figures usually are derived from approximate population censuses from areas where mutilation is known to be practiced.

As a cultural practice, female genital surgeries are particularly widespread on the African continent and to some extent on the Saudi Arabian peninsula, but are only sporadically documented from the rest of the world. Instances of genital mutilation are reported from selected communities in the Middle East and from Muslim Malaysia. In Europe and the USA, clitoridectomy was until the 1930s medically administered in diagnosis of excessive masturbation, nymphomania, and hysteria (cf, e.g., Roheim, 1932; Bonaparte, 1953). Cosmetic surgery of enlarged clitorises and labia minora has been reported recently both from the USA and Europe (cf Walley, 1997). In Europe and the USA, female genital mutilation was primarily a single instance phenomenon and not a social convention. During the last few decades of the twentieth century, however, large-scale immigration to Europe, the USA, and elsewhere from African countries where genital operations are mandatory, has extended the geographical location of female genital mutilation (cf Lionnet, 1992)

In Africa, the practice of genital surgeries is contained to certain regions only, most notably to the northeastern and eastern parts of the continent, from Egypt in the north to Tanzania in the south and across the continent along the Sahelian countries from Ethiopia to Senegal. It is practiced in some 28 countries and by a wide variety and number of ethnic groups. The extent and type of surgical intervention as well as its cultural meaning and social context vary considerably making it difficult to generalize across the societies in which female genital surgeries are the norm. Female genital mutilation occurs among Muslims, Christians (particularly among Copts), as well as adherents of traditional systems of belief, among agricultural and nomadic peoples alike and in rural as well as urban communities. The highest regional concentration, that is, where a majority of women in a particular area are mutilated, is in Northeast Africa; in the Horn (Somalia, Djibouti, and Eritrea), across Ethiopia, and into northern Sudan and Egypt. In Somalia and northern Sudan female genital mutilation is almost universal (Gruenbaum, 1996). Elsewhere in Africa, genital surgeries as a cultural practice are scattered and alternate with ethnic affiliation, religious faith, economic class, and level of education. Within the same community, practices may even vary between subgroups and individual families, causing a situation where mutilated women live side by side with nonmutilated females.

As a general rule, where female genital mutilation is performed, male circumcision is also the norm. People themselves often draw a symbolic and social parallel between boys' and girls' circumcision. This is, among other things, reflected in a common term for the interventions in the local vernacular. While a number of societies in Africa and beyond practice male circumcision without cutting their women, the opposite, however, is never the case.

Operation Practices

Mutilation practices vary greatly. In some areas, such as Amharic Ethiopia, girls are operated upon within 40 days after birth, while in neighboring Somalia the operation usually is performed when the girls are between 6 and 8 years old (Hicks, 1993; Talle, 1993). In Kenya and Tanzania the girls are cut routinely just before puberty and often in association with

elaborate initiation ceremonies (cf Murray, 1974). In a few, relatively exceptional cases reported among others from West Africa, the surgery is performed after puberty. It also has been reported that married women who have defied traditions and remained unmutilated, in cases of barrenness have undergone the surgery after being married in order to promote fertility. Not only the girl's age, but also the surgical intervention, the performance context, and ritual elaboration exhibit substantial variations. In places in Egypt and among devoted Muslims in towns in the Sudan, for instance, only the prepuce or the tip of the clitoris is excised. This operation is referred to as *sunna* ('duty' in Arabic, i.e., to follow tradition). Surgically, *sunna* is the mildest form of mutilation and physiologically comparable to the removal of the foreskin at male circumcision. Although encouraged by religious leaders, health personnel, and opinion makers as a commendable mutilation form, it is still uncommon. What people label *sunna*, however, in practice is often really a clitoridectomy.

The most widespread form of female genital mutilation in fact is clitoridectomy, an intervention that implies removal of the whole or part of the clitoris, sometimes together with the *labia minora*. In the latter case the operation is also referred to as excision. Various forms of clitoridectomy-cum-excision are practiced from Senegal and Mauritania in West Africa, across Sahel to Egypt in the northeast, and southwards to Tanzania in East Africa. It is carried out across the whole area where female genital surgeries occur.

The surgically most severe form of operation is called infibulation (from Latin *fibulae*, clip) or Pharaonic circumcision. This is a form of surgery where the *labia minora* and parts of *labia majora* are excised wholly or partly, and often the clitoris; the sides of the vulva are then stitched together. A tiny orifice is left open serving as an outlet for urine and menstrual blood. The surgery transforms the genitals of a woman into a flattened, smoothed, and closed vulva with an oblong scar and a miniscule opening at the lower part. The practice of infibulation is particularly prevalent in the Horn of Africa, in the Sudan (Northern), and in southern (Nubian) parts of Egypt. It also occurs widely in Mali, in northern Nigeria, and among ethnic Somalis in other East African countries. The infibulated women are opened (defibulated) at marriage, either surgically – often by the practitioner who did the closing – or by the husband at penetration.

In medical and anthropological literature it is common to differentiate between four types of genital surgeries: *sunna*, clitoridectomy, excision, and infibulation. The types are classified by a gradient of surgical severity. Variations and modifications of the different types may occur over time as well in individual cases within the same geographical location. In the Sudan, for instance, an intermediate type, a compromise between excision and infibulation, sprung up during the colonial area as a local response to legislation against the practice imposed from outside (El Dareer, 1982). The skills and techniques of the practitioners as well as the preferences of individual families may contribute to variation and differences in mutilation styles.

Genital operations usually are performed in the home of the girl and, as a rule, undertaken by an elderly woman skilled in the task. Often these women are local birth attendants as well. In Somalia, the practitioners belonged traditionally to a ritually ambiguous and despised subclan within the community. In other places, and these days even in Somalia, any adult woman with the necessary courage and skills may engage in the practice.

The operation normally is performed without any extensive form of anaesthetic or antiseptic precautions. Different herbs, animal fat, dung, or other substances believed to have homeostatic and cicatrizing properties are applied to the wound after the operation to prevent excessive bleeding and infection. Depending upon the routine, skills, and preference of the practitioner, the instruments used for the surgery are either a small knife prepared for that kind of work, or a razorblade. In the case of infibulation, the severed parts are stitched with thorns of the acacia species, or recently, with catgut. In urban areas, medically trained personnel or paramedics commonly perform 'circumcision' on both boys and girls. In addition to giving prophylactic antibiotics and tetanus injections, the urban practitioners normally apply local anesthetics when operating. Many parents prefer these modern operations as they tend to reduce the pain and minimize adverse effects, such as hemorrhaging or infection from the operation. With the medicalization of the genital surgeries, men have also become involved in operating on girls. Traditionally, men operated only on boys, while women operated on girls. The same-sex pair of circumcizer and circumcized reflected the thinking of male circumcision as an analogy of female circumcision.

Among peoples in Kenya and Tanzania and in large parts of West Africa, the surgical intervention is associated with elaborate rituals of initiation into adulthood (e.g., Murray, 1974). In northeastern Africa, where infibulation is widespread, the operation, however, is not normally followed by any extensive ceremonial celebration. The operated girl is usually given a few presents and the family mark the day with some extra food and drinks, but the event is not a ceremony of great societal worth. In cases of physical cutting being part of some more significant rites of passage into adulthood, the waning of such traditions has not necessarily led to a decrease in the surgery itself. Instead of being part of a larger event, genital mutilation is now often a single act performed on girls at almost any age.

Historical Roots

The existence of female genital mutilation dates back to pre-Islamic and pre-Christian time. Although the practice as such does not stem from religious affiliation, many people make an association between circumcision and religious faith. This is particularly the case in West Africa where genital surgeries both in men and women are more common in Muslim than Christian communities (Koso-Thomas, 1987).

The term Pharaonic circumcision originates from the folk legend that the pharaohs of ancient Egypt performed similar operations on their women. Whether the pharaohs actually infibulated their women or not will probably never be substantiated, but the term nonetheless indicates that we are dealing with a cultural phenomenon of great antiquity.

Greek merchants and travelers who crossed the Mediterranean in classical time mention that female genital mutilation occurred among Egyptians as well as among the peoples along

the Red Sea. The sparse descriptions emanating from these travelers, however, do not allow us to ascertain whether it was infibulation or clitoridectomy, or both, people practiced in earlier times.

European travelers visiting Egypt much later reported on the occurrence of clitoridectomy (Seligmann, 1913). However, female slaves from the south (perhaps Nubians from the Sudan or southern Egypt), whom the wealthy Egyptians favored as mistresses, were infibulated (Widstrand, 1964). Some authors held that these girls were mutilated upon arriving in the north by the slave merchants or by their masters to avoid loose living or unwanted pregnancies. This explanation is disputed as early reports suggest that infibulation occurred for some time among the peoples of the upper Nile in northern Sudan. Burckhardt who visited these areas early in the nineteenth century claimed that infibulated girls ('virgins') received a higher price on the Egyptian slave market than noninfibulated ones (Burckhardt, 1819).

The beginning of female genital mutilation remains obscure. It has not been possible to determine whether mutilation originated in one area and from there spread to other areas or whether it originated independently in various places. With reference to the African continent, however, it is not unlikely that the practice arose in a culturally, commercially, central area and from there reached and influenced other parts of the continent. This assumption is based upon the fact that the practice of female genital mutilation appears to have had varying historical origins depending when people came into contact with each other (cf, e.g., Murray, 1974). Historical and linguistic sources point to the Chusitic and Semitic-speaking peoples living along the Red Sea Coast as a core area of female genital mutilation (Seligmann, 1913; Hicks, 1993).

Recently, the practice of female genital mutilation is reported to have spread from one ethnic group to another in countries such as Chad, the Sudan, and Tanzania (cf Leonard, 2000; Gruenbaum, 1996). In the Sudan, lower hierarchy groups often assimilate the practice of mutilation in order to be able to marry those of higher rank. Young girls in Chad are reported to perform clitoridectomy as a mark of modernity; since around 1980 the operation has become a kind of fashion in at least one reported area (Leonard, 2000). Studies from places in West Africa suggest that the expansion of Islam is conducive to the spread of female genital mutilation. Among the ethnic Meru in Northern Tanzania, clitoridectomy has been revitalized as a sign of cultural pride and resistance against rapid modernization. These sketchy, contemporary examples show that the meaning of female genital surgeries may assume various values following shifts in social reality. To know its origin does not explain its present existence (Gruenbaum, 1996). One distinguishable trait in the spread of female genital mutilation, however, appears to be the copying of cultural practices of higher status groups by people of lower rank (cf Mackie, 1996).

The Cultural Explanation

The practice of female genital mutilation is embedded deeply in cultural and moral preferences, gender identity and person forming, perceptions of body aesthetics, and ethnic marking. The unsuccessful attempts made by the British colonial administration during the 1930s and 1940s to abolish or modify the most severe types of female genital mutilation in the Sudan and in Kenya testify to the social importance of the practice (cf Murray, 1974; El Dareer, 1982). Anthropologists and others often associate the tradition of female genital mutilation with a patriarchal social structure, practically and symbolically privileging female reproductive powers over female sexuality. (cf Boddy, 1982; Talle, 1994).

People themselves give many explanations to defend the practice. For instance, the pastoral Maasai of Kenya and Tanzania claim that a girl can neither be married nor give birth to healthy children if she is not clitoridectomized. Should she deliver before she is operated upon, the child will die or be born disabled (Talle, 1994). In the Sudan and Egypt, uncut female genitals are considered ugly and dirty. Men are reluctant to marry such girls because by not being 'purified' by genital excision, they are held to be immoral in the widest sense. For the Bantu-speaking Kikuyu in Kenya a girl must have her clitoris removed to be considered an adult and responsible person. And so on. People present a plethora of meanings and motives for upholding the practice. Although seldom elaborated on explicitly, the control of female sexuality is at stake. References to moral looseness, unmarriagability, and fertility failures speak strongly to the societal value of female chastity and of male control of female fertility and sexuality (cf, e.g., El Sadaawi, 1980; Gordon, 1991; Talle, 1993).

In Somalia, where the practice of infibulation is particularly widespread, the operation is performed in two stages: the excision and the stitching. First, the clitoris is excised. This is considered an act of purification (*xalaalays*; from the Arabic *xalaal*, allowed by religion). The clitoris (*kintir*) is a part of the body which is associated with dirt and childhood. A girl who has not been operated on is referred to as *kintirleey* (childish, i.e., having clitoris) and teased by peers. To adult Somali women the word *kinterleey* has a very negative and somewhat embarrassing ring. Another Somali synonym for *kintir* is *awle* which means approximately 'with (or having) father.' This term suggest that the Somalis, as a number of other peoples performing genital surgeries, see the clitoris as a male part in the female body and thus held to be unclean. When excised, women become more feminine (Talle, 1993).

Where infibulation is practiced, virginity is not held to be a natural condition in women but has to be forcibly implanted in girls while they are still young. The hymen of a woman, which attracts so much interest and attention in many other Muslim societies and evokes strong sanctions if broken before marriage, has little social or cultural significance among infibulation practicing people. To them a chaste girl means a 'sewn' girl. A tight infibulation is a sign of distinction and moral excellence. Symbolically, the closing is more important than the excision as the cutting of the flesh is irrevocable while the closing of the vulva can, in principle, be executed over and over again. Although never as tight as the first time, later surgeries (after childbirth and at divorce) return the woman's body to its previous state by remaking her virginity (El Dareer, 1982). Virginity under such circumstances is not an anatomical facticity but a social category (Hayes, 1975).

The closing of her vulva is looked upon as a protection for the girl against indecency and immorality. A reason often given,

for instance by Somalis, as to why originally they began to infibulate their girls was to prevent sexual assault of young girls while out herding alone. The infibulation, or the creation of the artificial virginity, is seen as a safeguard against unwanted pregnancies and unidentified paternity. Both in northern Sudan and in Somalia a nubile, 'virgin' daughter upholds and marks the moral standing of the partrilineal group (Hayes, 1975; Kennedy, 1970; Talle, 1993). To leave a daughter or sister 'uncut' is held to be unnatural; the vision of such a woman is beyond comprehension, something which cannot even be expressed in words.

A Global Concern

During the feminist wave of the 1970s in Europe and the USA, female genital surgeries became a politicized issue. At the Copenhagen conference in 1980 during the United Nations Decade of Women (1975–85) the issue was brought to widespread international attention. Opponents of the practice asserted that female genital mutilation was a severe abuse of women's and children's rights and demanded legislation against it. The term 'female circumcision' was substituted by 'female genital mutilation' (Hosken, 1982). This change of name from circumcision which is used for genital operations on men, was meant to draw attention to the severity of intervention, primarily in terms of health risks, on girls and women as compared with that of men (cf, e.g., Toubia, 1995).

Legislation against female genital mutilation is on the agenda of international organizations such as Save the Children, Amnesty International, WHO, UNICEF, and other UN agencies. To prevent the spread of such surgeries into their countries, European governments are passing laws criminalizing all forms of genital mutilation (cf Lionnet, 1992). African countries, Egypt, the Sudan, and Kenya among others, have also passed legislation against female genital surgeries.

The global discourse on female genital mutilation has often been moralizing, racially prejudiced, and without sufficient understanding of the social and cultural context of the practice. Many African women have reacted to the paternalistic attitude that often permeates the global debate. The term female genital mutilation, which has received widespread acclaim, but which holds the implicit assumption that parents and kin deliberately intend to harm children, is contested (cf Walley, 1997). A more neutral term proposed recently is 'female genital cutting' (Shell-Duncan and Hernlund, 2000). African women – themselves frequently victimized by genital surgery – have, however, also been among the strongest opponents to the practice (e.g., Abdalla, 1982). The latest voice in a string of personal testimonies of an excruciatingly painful intervention with lifelong suffering comes from an internationally celebrated Somali model, now a UN ambassador against female genital mutilation (Dirie and Miller, 1998).

Although opposition is growing everywhere, women continue to favor the practice. They do it not necessarily because they like it, but because they see no other alternative. Africans in exile, exposed to a majority society where mutilation is neither a norm nor an ideal and, additionally, relieved of the social pressure to perform it, may spearhead the eventual eradication of female genital mutilation.

See also: Children, Rights of: Cultural Concerns; Cultural Relativism, Anthropology of; Decision Making, Psychology of; Feminisms and Acculturation around the Globe; Gender History; Human Rights and Social Work; Human Rights, Anthropology of; Human Rights: Legal Aspects; Near Middle East/North Africa Studies: Gender; Reproductive Rights in Affluent Nations; Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective; Ritual.

Bibliography

Abdalla, R.M.D., 1982. Sisters in Affliction: Circumcision and Infibulation of Women in Africa. Zed Press, London.
Boddy, J., 1982. Womb as oasis, the symbolic context of Pharaonic circumcision in rural northern Sudan. American Ethnologist 9 (4), 682–698.
Bonepart, M., 1953. Female Sexuality. International Universities Press, New York.
Burckhardt, J.L., 1819. Travels in Nubia. Darf Publishers, London.
Dareer, A., 1982. Women, Why Do You Weep? Circumcision and its Consequences. Zed Press, London.
Diop, N.J., Moreau, A., Benga, H., 2008. Evaluation of the Long-term Impact of the TOSTAN Programme on the Abandonment of FGM/C and Early Marriage: Results from a Qualitative Study in Senegal. The Population Council, New York.
Dirie, W., Miller, C., 1998. Desert Flower: The Extraordinary Journey of a Desert Nomad. William Morrow, New York.
Genital cutting remains an important issue in the world human rights and children's rights movements a decade and a half after Talle's entry was first published. Important recent writings on the subject include:
Gordon, D., 1991. Female circumcision and genital operations in Egypt and the Sudan: A dilemma for medical anthropology. Medical Anthropology Quarterly 5 (1), 3–14.
Gruenbaum, E., 1996. The cultural debate over female circumcision: The Sudanese are arguing this one out for themselves. Medical Anthropology Quarterly 10 (4), 455–475.
Gruenbaum, E., 2001. The Female Circumcision Controversy: An Anthropological Perspective. University of Pennsylvania Press, Philadelphia.
Gruenbaum, E., 2005. Socio-Cultural Dynamics of Female Genital Cutting: Research Findings, Gaps, and Directions. Culture. Health & Sexuality 7 (5), 429–441.
Hansen, H.H., 1972/1973. Clitoridectomy: Female circumcision in Egypt. FOLK 14 (15), 15–26.
Hayes, R.O., 1975. Female genital mutilation, fertility control, women's roles, and the patrilineage in modern Sudan: A functional analysis. American Ethnologist 2 (4), 627–637.
Hayford, S., Trinitapoli, J., 2011. Religious Differences in Female Genital Cutting: A Case Study from Burkina Faso. Journal for the Scientific Study of Religion 50 (2), 252–271.
Hernlund, Y., Shell-Duncan, B. (Eds.), 2007. Transcultural Bodies: Female Genital Cutting in Global Context. Rutgers University Press, New Brunswick, NJ and London.
Hicks, E.K., 1993. Infibulation: Female Mutilation in Islamic Northeastern Africa. Transaction Publishers, New Brunswick and London.
Hosken, J.P., 1982. The Hosken Report: Genital and Sexual Mutilation of Females. Women's International News Network, Lexington, MA.
Johnsdottar, S., 2014. FGM in Sweden: Swedish Legislation Regarding "Female Genital Mutilation" and Implementation of the Law. Lund University. Research Report EC Daphne Project.
Kennedy, J.G., 1970. Circumcision and excision in Egyptian Nubia. Man 5, 175–191.
Koso-Thomas, O., 1987. The Circumcision of Women: A Strategy for Eradication. Zed Press, London.
Leonard, L., 2000. We did it for pleasure only': Hearing alternative tales of female circumcision. Qualitative Inquiry 6 (2), 212–228.
Lionnet, F., 1992. Identity, sexuality, and criminality: 'Universal rights' and the debate around the practice of female excision in France. Contemporary French Civilization 16 (2), 294–307.
Mackie, G., 1996. Ending footbinding and infibulation: A convention account. American Sociological Review 61 (6), 999–1017.
Momoh, C., 2005. Female Genital Mutilation. Radcliffe Publishing, Oxford.
Mulongo, Peggy, McAndrew, Sue, Martin, Caroline Hollins, 2014. Crossing Borders: Discussing the Evidence Relating to the Mental Health Needs of Women Exposed to Female Genital Mutilation. International Journal of Mental Health Nursing 23 (4), 296–305.

Murray, J.M., 1974. The Kikuyu Female Circumcision Controversy with Special Reference to the Church Missionary Society's 'Sphere of Influence. University of California, Los Angeles. Ph.D. Available from Silverprint-Order Entry.

Obermeyer, C.M., 2005. The Consequences of Female Circumcision for Health and Sexuality: An Update on the Evidence. Culture. Health & Sexuality 7 (5), 443–461.

Roheim, G., 1932. Psycho-analysis of primitive cultural types. International Journal of Psycho-analysis XIII, 1–221.

Sadaawi, N., 1980. The Hidden Face of Eve: Women in the Arab World. Zed Press, London.

Seligmann, C.G., 1913. Some aspects of the Hamitic problem in the Anglo-Egyptian Sudan. Journal of the Anthropological Institute XLIII, 593–705.

Shell-Duncan, B., Hernlund, Y., 2000. Female 'circumcision' in Africa: Dimensions of the practice and the debates. In: Shell-Duncan, B., Hernlund, Y. (Eds.), Female 'Circumcision' in Africa: Culture, Controversy, and Change. Lynne Rienner Publishers, Boulder, CO.

Shell-Duncan, Bettina, Wander, Katherine, Hernlund, Ylva, Moreau, Amadou, 2013. Legislating Change? Responses to Criminalizing Female Genital Cutting in Senegal. Law & Society Review 47 (4), 803–835.

Smith, Courtney, 2011. Who Defines "Mutilation"? Challenging Imperialism in the Discourse of Female Genital Cutting. Feminist Formations 25 (1), 25–46.

Talle, A., 1993. Transforming women into 'pure' agnates: Aspects of female infibulation in Somalia. In: Broch-Due, V., Rudie, I., Bleie, T. (Eds.), Carved Flesh/Cast Selves: Gendered Symbols and Social Practices. Oxford/Providence: Berg.

Talle, A., 1994. The making of female fertility: Anthropological perspective on a bodily issue. Acta Obstetricia et Gynecologica Scandinavica 73, 280–283.

Toubia, N., 1995. Female Genital Mutilation. A Call for Global Action. Women Inc., New York.

United Nations Children's Fund (UNICEF), 2013. Female Genital Mutilation/Cutting: A Statistical Overview and Exploration of the Dynamics of Change. UNICEF, New York.

Walley, C.J., 1997. Searching for 'voices': Feminism, anthropology, and the global debate over female genital operations. Cultural Anthropology 12 (3), 405–438.

Widstrand, C.G., 1964. Female infibulation. Studia Etnografica Upsaliensia XX, 95–124.

Female Power at Work

David J Maume, University of Cincinnati, Cincinnati, OH, USA
Leah Ruppanner, University of Melbourne, Parkville, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews theory and research on the increasing presence of women in management and examines how this affects gender inequality among subordinates. Whereas it is true that women in upper management are more likely to reduce inequality among subordinates, most female managers are located at the lower rungs of corporate hierarchies where they have little impact on prevailing patterns of gender inequality. We call for further research on the location of women in management hierarchies and how this could affect their careers of their subordinates.

Introduction

Theories of economic inequality have long implicated men's dominance of authority positions as a mechanism for the enhancement of gender inequality in the labor market (Kanter, 1977; Reskin, 1988). It is thought that men in power will hire, evaluate, pay, and promote subordinates in ways that benefit men and disadvantage women. This belief is so widespread that the term 'glass ceiling' has entered both popular discourse and scholarly writings, describing the dearth of women in leadership positions from which they could act to reduce gender inequality.

Whereas it is true that relatively few women are CEOs or sit on corporate boards, women have dramatically increased their presence in management. For example, in the United States, figures from the Current Population Survey (a monthly survey of the labor force) show that women's representation in management rose from just over one-fifth in 1975 to nearly half by 2005 (cited in Cohen et al., 2009: p. 320). Similarly, women account for over half of the managers and senior officials (including legislators) in many European countries, including Germany, Finland, and Denmark (see **Figure 1**). Yet, significant variation exists among the most developed countries, with Japan reporting lower female representation in high-status positions than many of the Middle Eastern nations including Qatar and the United Emirates. This indicates that high economic development does not equate to gender equality and identifies important country-to-country variation in women's status reflected, in part, through women's status on the labor market. Of course, given that government statistics often combine lower-, middle-, and upper-managers into the same occupational category of manager, it is difficult to determine whether women's increased presence in management represents real progress in ascending to authority positions within firms. A related question is whether once they are in power, female managers are able and willing to combat gender inequality in job assignments, pay, and promotions among their subordinates.

This article will review theory and research on women in management. We discuss the primary reasons why women have increased their representation in management and review findings on the impact that female managers have on gender inequality (specifically job segregation), pay inequality, and employees' well-being. We largely confine our review to the

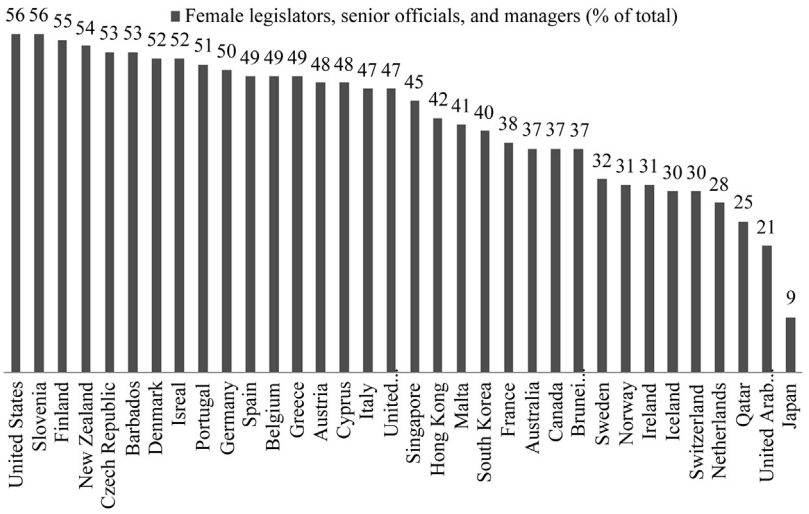

Figure 1 Percentage of women in high-status positions by country (United Nations Development Report, 2009).

United States and Europe because most of the research on female managers has been conducted in these countries. We conclude with a call for additional research that will more clearly elucidate the effect that female managers have on gender inequality.

The Rise of Female Managers

In the United States, the feminist movement in the 1960s inspired trends toward managerial diversity. Universities increasingly responded to women's demands for inclusion by admitting them into previously male-dominated fields of study such as engineering, law, the sciences, and business. With the decline in the sex typing of college majors, the gender gap in college attainment similarly declined and by the early 2000s, rates of college graduation among young adults ages 25–29 years were higher for women than men (Cotter et al., 2004). As women upgraded their skills in previously male-dominated fields of study, they were well positioned to take managerial jobs, as they pursued their careers. In addition, more managerial opportunities were open to women as corporations diversified their managerial ranks in response to legislation mandating equal treatment of workers and policies that awarded government contracts to businesses that practiced diversity were in place.

Only in a stagnant economy would more female managers necessarily mean fewer male managers. However, over the last 50 years the economy did grow and newer industries emerged, especially in care work and service provision. The transition from a goods-providing economy to a service-providing economy has increased the number of low-wage and part-time jobs and has drawn more women into the labor force. It is plausible that men retained their managerial positions in industries where they had traditionally worked and that female managers were incorporated into care work and service industries, largely supervising other women who were low-paid and marginally attached to the labor force. Indeed, Stainback and Tomaskovic-Devey's (2009) analysis of company reports to the U.S. Equal Opportunity Commission showed that between 1966 and 2000, white men were over-represented (relative to their proportion of the labor force) in management in the old higher-paying industries, such as construction, manufacturing, and transportation/utilities, whereas women's gains in management were most pronounced in the newer lower-paying retail trade and personal and social service industries. Moreover, this study also found that the gender composition of the firm's payroll significantly determined access to management, in that firms with a higher percentage of female employees had proportionately more female managers.

The rise of female managers in Europe largely parallels that of the United States, albeit with important country-to-country differences. Indeed, the number of female managers globally has increased sizably but with rates plateauing since 2007 (Thornton, 2014). Further, women's managerial representation has not kept pace with women's participation in the labor force (Lyness and Terrazas, 2006). Yet, the European policy context varies greatly from the United States reflecting a long legacy of equal rights amendments. For example, the European Union adopted equal pay legislation in 1975 (Council Directive 75/117/EEC of 10 February) (Europa, 2014), followed by an equal access to employment clause in 1976 (Council Directive 76/207/EEC) (Europa, 2014). Clearly, Europe, headed by the European Union, has a long history of commitment to gender equality as reflected by legislation. Yet, the effectiveness of these policies, notably women's representation in managerial positions, varies by country. This partially reflects country-specific variations in the definition of managers (Vinnicombe, 2000). The segregation of female managers in Europe parallels that in the United States. Even in the most gender-equal countries where equal opportunity laws are strictly enforced, female managers cluster in the public and banking sectors (Adler, 1994). Finally, structural impediments also limit women's position in high-status jobs. Indeed, the female manager rates are often paradoxical to gender empowerment. In other words, women hold fewer than expected managerial positions in the gender-empowered countries. Two explanations are offered: first, gender role expectations frame women as being 'not serious' about work, which despite being merely a perception carries weight in promotion decisions (Catalyst, 2002). Secondly, the cost of generous welfare state provisions, notably long maternity leaves, prohibits the promotion of women into high-status positions (Vinnicombe, 2000; Thornton, 2014). This suggests that even strong anti-discriminatory policy provisions are not sufficient to overcome organizational gender inequality.

Turning our gaze globally, important variation in women's employment exists by region. Notably, South East Asia has some of the highest percentages of female managers (32%) ahead of Europe's mean of (24%) (Thornton, 2014). This, in part, reflects family processes – notably the reliance on the extended family for childcare that enables women's continuous employment. By contrast, patriarchal societies, including Japan, report some of the lowest rates of female managerial representation. In sum, women's global managerial representation reflects complex family, gender, and work dynamics.

Some are skeptical that the rise of female managers represents real progress for women. First, because of government oversight of their personnel practices, many US firms may have re-titled a 'glorified secretary' into an office manager in order to increase diversity in the ranks of management. Secondly, as service-sector organizations evolved over time, a process of title-inflation created many more lower-level 'managerial' jobs out of previously 'clerical' and 'administrative' jobs. Thirdly, women who did become managers (e.g., in government, education, and health care) were entering jobs vacated by men, who were increasingly entering more prestigious male-only managerial jobs (e.g., in financial and professional services) a process known as 'managerial re-segregation.'

Jacobs (1992) examined these arguments using census and survey data on managers between 1970 and 1988. Using the US General Social Survey, he found that three-fourths of female managers had subordinates, which contrasts with the 'glorified secretary' image of a female manager with no supervisory duties. Survey findings also showed that the jobs of female and male managers were similarly rewarding – only one-fourth of male and female managers "… would quit work if they had enough money." From the census data, Jacobs found that the integration of managerial jobs continued over time, in contrast to the argument that managerial jobs were re-segregating;

and he found no evidence that title inflation was reducing the standing of managers, as the manager/non-manager ratios in median pay held steady over time. On the other hand, Jacobs found that compared with men, female managers occupied the lower rungs on corporate job ladders; more than half of the men but only one-third of the women had subordinates who, in turn, supervised lower-level subordinates. And, perhaps most important, Jacobs reported that the gender pay gap among managers, although declining over time, was still larger than in the labor force as a whole. From these results, Jacobs concluded "… the predominant trend has been toward real, if slow progress into management on the part of women." This is an important conclusion because if the rise of female managers represents real progress toward gender equality, then these organizational leaders would be positioned to ameliorate gender inequality among subordinates. It is to that question that we now turn.

Do Female Managers Reduce Gender Inequality?

On the one hand, organizational demographic theory and research suggests that female managers act to reduce gender inequality in their work environments. Assuming preferences for in-group members, organizational demographers find that female supervisors evaluate and reward female subordinates more favorably (Tsui and O'Reilly, 1989). Others argue that boss–worker homophily is important in another way, as shared experiences with gender discrimination and work-family conflict increases the likelihood that female superiors will bond with and mentor female subordinates (Fagenson, 1993). Finally, homophily structures social and communication networks such that female subordinates gain access to more information, resources, and career-enhancing job assignments when they report to a female manager (Hultin and Szulkin, 2003).

On the other hand, research into segregation and status characteristics suggests that female managers have no impact on gender inequality. That is, female workers in general are stereotypically seen as less competent and more family-focused than men and will be crowded into occupations that pay less and have higher turnover (Ridgeway, 2011; Reskin, 1988). However, employers must secure the commitment of devalued female workers, especially in service industries where they may be in direct contact with clients and customers. Thus, firms assign female supervisors to those working in female-dominated jobs. This process of 'bottom-up ascription' sends a symbolic message to devalued female subordinates that mobility into management is possible for them via hard work and commitment to their employers (Elliott and Smith, 2001). If so, the gender pay gap among subordinates may have more to do with processes that segregate women into lower-paying female-dominated jobs and may be largely unaffected by the sex of supervisors.

Even at the upper reaches of corporate hierarchies, some question women's ability to reduce gender inequality. First, there is the tendency to 'think manager–think male,' so that even if women do ascend to management they are presumed to be less competent than male managers (Ridgeway, 2011; Schein and Müller, 1996; Wajcman, 1998). Not only must female managers outperform their male counterparts to gain recognition but corporate work cultures dictate that strivers work long hours and subjugate their personal lives to the business imperatives of their employers (Kanter, 1977). This further disadvantages female managers who, unlike men, are assumed to have domestic responsibilities that compete with work obligations. Thus, in a male-dominated corporate culture, female managers may have ascended to their positions because they distanced themselves from women's issues and refused to address gender inequalities among their subordinates. And, by 'managing like a man,' female superiors often earn the 'queen bee' label from female subordinates who are resentful that their female bosses do not do more to address gender inequality in the workplace (Wajcman, 1998).

Several studies report that female managers ameliorate gender inequality. For example, studies in the United States (Cohen and Huffman, 2007), Sweden (Hultin and Szulkin, 2003), and Portugal (Cardoso and Winter-Ember, 2010) all find that as the proportion of female managers rises in firms or in labor markets, the wage gap among subordinates declines. Another study shows that gender segregation within firms declines with female representation in management (Huffman et al., 2010); cross-national data (Lyness and Kropf, 2005) find that managers in gender-empowered countries (i.e., with a higher percentage of female managers in the labor force) are more likely to report the importance of work–family balance and be willing to offer employees flexible work arrangements. It is important to note, however, that most of these studies use hierarchal modeling techniques, showing that the size of the gender wage gap among individuals is conditional upon women's representation in management in the corporate or labor market. Because these studies do not directly link individuals to managers, their results may be spurious if some other factor (e.g., government oversight, firm size) explains both women's representation among managers and higher pay for employees.

When studies directly link subordinates to the sex of supervisor, the results tend to support the proposition that female managers are powerless to reduce gender inequality. In samples of small businesses (Penner and Toro-Tulla, 2010) and retail grocers (Penner et al., 2013), the gender gap in pay was unrelated to the sex of the owner or store manager. Further, in a survey of the US labor force, Maume (2011) found support for a 'queen bee' female manager effect, in that male (but not female) subordinates reported more job-related support and were more optimistic about their promotion chances when they reported to female supervisor. One potential criticism of these studies is that by asking subordinates to report on the sex of their supervisors, they are largely reporting on first-line supervisors, not upper-level managers who have more power to ameliorate gender inequality within the firm. Indeed, in a survey of Korean workplaces, Stainback and Kwon (2012) found that sex segregation in job assignments declined with women's representation in upper management but female representation among first-line supervisors was associated with higher levels of sex-segregated job assignments among subordinates. In effect, highly placed female managers acted to reduce gender inequality but sex segregation among subordinates generated sex-segregated supervisory roles.

Summary

If female managers are to reduce gender inequality in the labor market, they must first gain access to powerful positions within firms and then act upon their desire to hire, pay, and promote female subordinates, similar to the existing position with regard to men. In this article, we have found that research on female managers is growing but the evidence on these two propositions is equivocal. First, in regard to access to management positions, the literature quite clearly implicates government efforts to promote equality as an impetus to women's increased presence in management. Yet, men have retained their management positions in the older male-dominated higher-paying manufacturing and financial service sectors of the economy, whereas women have increasingly gained access to management in newer female-dominated and lower-paying personal, trade, and social service sectors of the economy. Second, in regard to their effect on inequality when women manage, the evidence does not resolve the question of whether female managers act as 'change agents' or are 'cogs in the corporate machine.' Ecological studies report that women's managerial representation at the firm level reduces wage inequality at the level of subordinates; but when studies directly link sex-of-supervisor with subordinates, the wage gap is unaffected and there is even some indication that female superiors foster men's careers at the expense of their female subordinates. But ecological studies are plagued by concerns about spuriousness and the 'null sex-of-supervisor effect' on the gender wage gap of subordinates is likely because these studies are only observing first-line supervisors; if samples had more workers reporting to upper managers, the findings may be more supportive of the proposition that female managers are change agents.

In the end, the salient question may not be whether female managers matter in ameliorating inequality, but rather how their effectiveness varies across industries and the organizational context. Where women are represented in upper management, their presence and policies appear to reduce gender segregation and the gender pay gap. Yet, women's representation in the ranks of lower- and middle-managers appears to be an organizational strategy to match supervisors to workgroups that are largely female and poorly paid. From these lower-level supervisory positions, women have little power to affect change and reduce inequality. This returns us to the question of whether women's presence in management represents real or vacuous progress toward gender equality. Jacobs' (1992) conclusion that the rise of female managers represents slow but steady progress toward equality is more than two decades old and there is clearly a need to further investigate women's location in contemporary corporate hierarchies. This is especially true in light of research by Cohen et al. (2009), which drew on US census data to show that among managers, sex segregation increased between 1990 and 2000. This suggests that in contrast to past trends toward integration, managerial jobs were in the process of a re-segregation that allocated the poorest managerial jobs to women and the best managerial jobs to men.

We suspect that macroeconomic trends may be responsible for this process of managerial re-segregation. Beginning in the 1970s but accelerating in the last two decades, the 'service-oriented' employers have altered the employment relationship by hiring more contingent and part-time workers, freezing wages, and downsizing and reorganizing their work operations (Kalleberg, 2009). If women managers are over-represented in industries that employ low-wage and contingent labor whereas male managers are found in high-wage industries with stable workforces, then women's status as managers vis-à-vis their male counterparts may have eroded in the last few decades. This is a clear clarion call for further research on the relative position (and rewards) of male and female managers.

Finally, nearly all of the research cited above is cross-sectional in design, leaving open the question of how career trajectories are affected by male versus female managers. More clearly identifying the organizational actors affecting the decisions (hiring, job assignment, pay, promotion, etc.) that impact gender inequality and by observing how these vary over the course of a career is needed in future research. This will require difficult and ambitious data collection efforts but such projects are necessary if we are to better understand how and in what organizational context, female managers ameliorate gender inequality among subordinates.

See also: Family and Work; Gender and Technology: From exclusion to inclusion?; Management: General; Wage Differential and Structure.

Bibliography

Adler, N.J., 1994. Competitive frontiers: women managing across borders. Journal of Management Development 13, 24–41.

Cardoso, A.R., Winter-Ember, R., 2010. Female-led firms and gender wage policies. Industrial and Labor Relations Review 64, 143–163.

Catalyst, 2002. Women in Leadership: A European Business Imperative. Catalyst Research Center, New York.

Cohen, P.N., Huffman, M.L., 2007. Working for the woman? Female managers and the gender wage gap. American Sociological Review 72, 681–704.

Cohen, P.N., Huffman, M.L., Knauer, S., 2009. Stalled progress? Gender segregation and wage inequality among managers, 1980–2000. Work and Occupations 36, 318–342.

Cotter, D.A., Hermsen, J.M., Vanneman, R., 2004. Gender Inequality at Work. Russell Sage Foundation, New York.

Elliott, J.R., Smith, R.A., 2001. Ethnic matching of supervisors to subordinate work groups: findings of bottom-up ascription and social closure. Social Problems 48, 258–276.

Europa, 2014. Summaries of EU Legislation, from: http://europa.eu/legislation_summaries/employment_and_social_policy/equality_between_men_and_women/c10905_en.htm (retrieved 23.01.14.).

Fagenson, E.A., 1993. Women in Management: Trends, Issues, and Challenges in Managerial Diversity. Sage, Newbury Park, CA.

Huffman, M.L., Cohen, P.N., Pearlman, J., 2010. Engendering change: organizational dynamics and workplace gender segregation, 1975–2005. Administrative Science Quarterly 55, 255–277.

Hultin, M., Szulkin, R., 2003. Mechanisms of inequality: unequal access to organizational power and the gender wage gap. European Sociological Review 19, 143–159.

Jacobs, J.A., 1992. Women's entry into management: trends in earnings, authority, and values among salaried managers. Administrative Science Quarterly 37, 282–302.

Kalleberg, A.L., 2009. Precarious work, insecure workers. American Sociological Review 74, 1–22.

Kanter, R.M., 1977. Men and Women of the Corporation. Basic Books, New York.

Lyness, K.S., Kropf, M.B., 2005. The relationships of national gender equality and organizational support with work-family balance: a study of European managers. Human Relations 58, 33–60.

Lyness, K.S., Terrazas, J.M., 2006. Women in management: an update on their progress and persistent challenges. International Review of Industrial and Organizational Psychology 21, 267.

Maume, D.J., 2011. Meet the new boss…same as the old boss? Female supervisors and subordinate career prospects. Social Science Research 40, 287–298.

Penner, A.M., Toro-Tulla, H.J., 2010. Women in power and gender wage inequality: the case of small businesses. In: Williams, C., Dellinger, K. (Eds.), Research in the Sociology of Work: Gender and Sexuality in the Workplace. Emerald, Bingley, UK, pp. 83–105.

Penner, A.M., Toro-Tulla, H.J., Huffman, M., 2013. Do women managers ameliorate gender differences in wages? Evidence from a large grocery retailer. Sociological Perspectives 55, 365–381.

Reskin, B.F., 1988. Bringing the men back in: sex differentiation and the devaluation of women's work. Gender and Society 2, 58–81.

Ridgeway, C.L., 2011. Framed by Gender: How Gender Inequality Persists in the Modern World. Oxford University Press, New York.

Schein, V.E., Müller, R., 1996. Think manager—think male: a global phenomenon? Journal of Organizational Behavior 17, 33–41.

Stainback, K., Kwon, S., 2012. Female leaders, organizational power, and sex segregation. The Annals of the American Academy of Political and Social Sciences 639, 217–235.

Stainback, K., Tomaskovic-Devey, D., 2009. Intersections of power and privilege: long-term trends in managerial representation. American Sociological Review 74, 800–820.

Thornton, G., 2014. Women in Senior Management: Still Not Enough, from: http://www.grantthornton.com.au/Publications/Research/ibr_12.asp (retrieved 23.01.14.).

Tsui, A.S., O'Reilly, C.A., 1989. Beyond simple demographic effects: the importance of relational demography in superior-subordinate dyads. Academy of Management Journal 32, 402–423.

Vinnicombe, S., 2000. The position of women in management in Europe. Women in Management: Current Research Issues 2, 9–25.

Wajcman, J., 1998. Managing Like a Man: Women and Men in Corporate Management. University Park, Pennsylvania, Penn State Press.

Female Quotas in Parliament: Do They Make a Difference?

Mona L Krook, Rutgers University, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Electoral gender quotas have emerged as widespread political reform in recent years, having been adopted in more than 130 countries worldwide as of 2014. This article presents an overview of gender quota policies and surveys their effects in five respects: numbers of women elected, legislative diversity, policy-making, public attitudes, and political engagement. Overall, the evidence suggests that the introduction of quotas has been important for rethinking the role of gender in political life, enabling women to play a greater role in politics around the globe. However, the impact of quotas on women's empowerment more broadly continues to be debated.

Countries around the world have witnessed dramatic increases in the proportion of women elected to national parliaments since the 1990s. These changes are due in large part to the introduction of electoral gender quotas, establishing minimum thresholds for the nomination of female candidates. Although the first quotas were adopted in the 1930s, these policies became increasingly popular in the 1990s and 2000s in all major regions of the world (Dahlerup, 2006; Krook, 2009). The recent and global nature of these developments has sparked both scholarly and popular interest, leading work on quotas to become the fastest growing area of research on women and politics (Krook, 2006b), as well as a key focus of the gender-related programing of international organizations (Krook and True, 2012; Towns, 2010). Quotas have become the subject of a rich and expanding literature, with the potential to affect how social and behavioral scientists study a host of questions related to elections and political representation.

Gender Quotas

Gender quotas constitute the widest-reaching electoral reforms in recent years. Between 1930 and 1980, only 10 countries established quota provisions, followed by 12 states in the 1980s. Over the course of the 1990s, however, quotas appeared in more than 50 countries and have been joined by more than 50 more since the year 2000 (for an updated list, see http://www.quotaproject.org). As a result, quotas now exist in more than 130 countries around the world, but the vast majority of these measures have been passed within the last 15 years. The main reasons given for quota adoption include the mobilization of women's groups, the strategic incentives of political elites, and the support of international organizations and transnational networks (for a review, see Krook, 2009). One of the few countries where these debates have seen no echoes is the United States, where there is strong normative resistance to mechanisms of affirmative action and little pressure to conform to emerging international norms. In addition, the system of primary elections employed in the United States makes quotas for women especially difficult to apply (Krook, 2006a).

The three main types of quotas are reserved seats, party quotas, and legislative quotas. *Reserved seats* involve reforms to constitutions, and occasionally electoral laws, which set aside seats for women. They are found in Africa, Asia, and the Middle East. Although they first emerged in the 1930s and were the main type of quota adopted through the 1970s, since the year 2000 they have been passed in various countries that otherwise have very few women in politics. In most instances, they provide for low levels of female representation, usually between 1 and 10% of all representatives, although since 2000, several countries have instituted much larger provisions of 30%. In some cases, like India, reserved seats apply to single-member districts in which only women may run for election. In others, such as Afghanistan, they are allocated in multi-member districts to the designated number of women that win the most votes. In yet others, like Uganda, women are selected to these seats several weeks after the general elections by members of the national parliament.

Party quotas, in contrast, entail voluntary changes to party statutes committing a particular party to aim for a specific proportion of women among its candidates. They were first adopted in the early 1970s by a small number of socialist and social democratic parties in Western Europe. Over the course of the 1980s and 1990s, however, these measures began to appear in a diverse array of parties in all regions of the world, such that today they are the most common type of gender quota. They typically set a goal of between 25 and 50% female candidates. All the same, the phrasing of this requirement varies: some policies identify women as the group to be promoted by the quota, for example in Argentina, South Africa, and Spain, while others set out a more gender-neutral formulation, as in Italy and several Nordic countries. Party quotas govern the composition of party lists in countries with proportional representation electoral systems, which is the case in much of the world, and are directed at collections of single-member districts in countries with majoritarian arrangements, such as the United Kingdom.

Legislative quotas, finally, are measures enacted through reforms to electoral laws and sometimes constitutions requiring all parties to nominate a certain percentage of female candidates. They tend to be found in developing countries, especially Latin America, and postconflict societies, primarily in Africa, the Middle East, and Southeastern Europe. They are the newest kind of gender quota, appearing first only in the 1990s. They generally call for women to form between 25 and 50% of all candidates, and in most instances, the language of these

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64064-3

measures is gender-neutral, speaking of women and men together or making reference to the 'underrepresented sex.' Yet, legislative quotas vary in terms of how strictly their goals are articulated: some speak vaguely about 'facilitating access,' as is the case in France, while others offer concrete guidelines regarding the selection and placement of female candidates, like in Argentina, Belgium, and Costa Rica. Similar to party quotas, these policies are implemented in different ways depending on the electoral system, applying alternatively to party lists, the case in much of the world, or to a broader group of single-member districts, as in France. However, given their status as law, a distinctive feature of these quotas is that they may contain sanctions for noncompliance and be subject to oversight from external bodies, as in Mexico and Portugal. Quotas are therefore subject to extensive within- and across-category variation, in addition to being adopted for a range of different reasons.

Numbers Elected

One of the puzzling features of quota policies is that they have not resulted in a uniform rise in the proportion of women elected: some have produced dramatic jumps, while others have led to more modest changes and even setbacks in women's representation (Krook, 2009). In addition to differences in the features of specific policies, which affect their likelihood of being implemented, cross-national comparisons are complicated by the fact that quotas are introduced when variations already exist in the percentage of women in national parliaments. Outcomes are thus the combined result of quotas, where these are present, and other factors that were at work before quotas were established. As a result, quotas do not simply lead to gains proportional to the quota policy, but also interact in various ways with features of the broader political context.

Three reasons have been offered to untangle these effects. The first concerns the details of quota measures themselves. Although reserved seats guarantee changes in the numbers of women elected, some claim that party quotas are more effective than other types of quotas because they are voluntary measures, adopted from concerns about electoral advantage. Others insist that legislative quotas have more force because they bind all parties, not simply those who choose to adopt quotas, and are enforced by state bureaucracies and the courts, rather than party leaders (Jones, 1998; Tripp and Kang, 2008). However, the impact of particular policies also stem from their wording, requirements, sanctions, and perceived legitimacy, all of which may have intended and unintended effects (Krook, 2009). In France, for example, the financial penalties associated with the 50% quota law create distinct incentives for parties of different sizes: larger parties tend to ignore the requirements, while smaller parties are more likely to comply, for the simple reason that the latter are under greater pressure than the former to maximize the amount of state funding they receive (Murray, 2007).

A second explanation relates to the 'fit' between quota measures and other political institutions. For example, quotas often have the greatest impact in countries with proportional representation electoral systems with closed party lists and high district magnitudes (Jones, 1998; Schwindt-Bayer, 2009). In Sweden, for example, multiple seats are available in each constituency and candidates are elected from lists put forward by political parties. In contrast, it is more difficult to apply quotas where only one seat is available per district, unless the quota entails reserved seats, as in Tanzania. Quotas also tend to improve women's representation in countries where several parties coexist and larger parties respond to policy innovations initiated by smaller parties, as well as in parties with left-wing ideologies where the party leadership is better able to enforce party or national regulations (Kittilson, 2006). Further, quotas are often more successful in countries where the political culture emphasizes sexual difference and group representation, and less successful where it stresses sexual equality and individual representation (Krook et al., 2009). In part, this is because quotas have been challenged in some countries as a violation of constitutional principles of equality, as was the case in Italy and the United Kingdom in the 1990s and Mexico and Spain in the 2000s.

A third and final account points to the role of various actors vis-à-vis quota implementation. Party elites, who are often overwhelmingly male, are the group most directly responsible for variations in quota impact, since the effective application of quotas largely hinges around elites' willingness to recruit female candidates. In a large number of cases, elites take various steps to mitigate quota impact, ranging from passive refusal to enforce quotas to more active measures – including large-scale electoral fraud – to subvert their intended effects. Elites in Bolivia, for example, went so far as to change male names in female ones as a means to circumvent the 30% quota law. However, other actors may play a direct or indirect role in enforcing quota provisions, including women's organizations inside and outside parties, national and international courts, and ordinary citizens, all of whom may monitor party compliance with quota measures in ways that lead elites to ignore or honor, and possibly even exceed, quota requirements (Krook, 2009). Policy design, fit with other electoral institutions, and the presence or absence of political will to implement quota provisions thus all play key roles in shaping the effects of quotas on women's opportunities to be elected.

Legislative Diversity

Although increasing the numbers of women in political office is a central focus of quota reform, a closer look at the debates surrounding quota adoption suggests that their spread is not simply linked to concerns to increase women's political presence. Supporters argue, for example, that such measures will increase diversity among the types of women elected, raise attention to women's issues in policy-making, change the gendered nature of the public sphere, and inspire female voters to become more politically involved. At the same time, opponents express concerns that quotas will facilitate access for 'unqualified' women, bring individuals to office with little interest in promoting women's concerns, reinforce stereotypes about women's inferiority as political actors, and deter ordinary women's political participation (Franceschet et al., 2012). In other words, quota campaigns revolve around competing claims about what quotas might 'do' if enacted,

suggesting that quotas may have a host of positive and negative effects, above and beyond their impact on the numbers of women elected.

One common debate regards the types of women who might be chosen via quotas: advocates propose that women from more diverse backgrounds will be elected, while opponents express concerns about elite and 'unqualified' women benefiting from quota provisions. The findings thus far reveal a mixed picture, tied at least in part to difficulties in defining 'qualifications' for political office. Moreover, there can be tensions in the preferences of different constituencies: whereas women's movements might hope that the women elected via quotas more closely resemble 'ordinary women' and/or have feminist credentials, political elites might opt for female candidates who are more similar to the men already in office on the grounds that these characteristics conform with their already existing criteria of 'merit' (Franceschet et al., 2012).

Some studies indicate that quotas lead to the recruitment of elite women, especially those with ties to powerful men, and women with close loyalties to their parties (Cowley and Childs, 2003; Goetz and Hassim, 2003). Definitions of 'elite,' however, may also include women with high educational attainment, which, by some measures, would suggest high levels of qualification (Sater, 2007). Other research finds that quotas promote greater diversity in candidate selection, with those benefiting from quotas being relatively young and often coming from marginalized groups (Hughes, 2011). Almost all studies observe, in one way or another, that the women who accede to office via quotas have less overall political experience than their nonquota counterparts, both male and female. While this might be taken as evidence that quota women are less 'qualified,' it might also be viewed in terms of women's exclusion from electoral politics not affording them the opportunity to accumulate similar levels of political experience (Murray, 2010). Further, many of these women bring with them backgrounds in community organizing, suggesting that women may simply have different kinds of political experience (Franceschet et al., 2012).

Policy-Making

A second set of claims put forward during quota campaigns relates to how quotas might shape the legislative behavior of women in parliaments, above and beyond the effects of sex and gender. Most supporters make their case for quotas on the grounds that politics would change as a result of women's increased inclusion. They argue that policy-making would shift as a result of women's distinct policy priorities and, by allowing new voices into policy debates, would foster enhanced democratic legitimacy and good governance. Because women elected via quotas are elected because they are women, these debates may trigger a 'mandate effect,' leading citizens to anticipate that quota women are more likely than their nonquota counterparts to promote women's concerns. At the same time, however, negative connotations surrounding quotas may also generate a 'label effect,' leading women elected via quotas to ignore women's issues – or face stigmatization by other legislators due to their mode of selection, undercutting their legislative initiatives (Franceschet and Piscopo, 2008).

Initial research lends support to both of these intuitions. On the one hand, women elected through quotas have reported feeling obligated to act for women as a group, which has inspired them to bring new issues to the political agenda (Schwartz, 2004). On the other hand, others have sought to disassociate themselves from the quota and women's issues to demonstrate that they are 'serious' politicians (Childs, 2004). At the same time, many have been accused of acting only as proxies for male relatives and of being excessively loyal to party leaders (Goetz and Hassim, 2003). Both of these trends – deliberate lack of attention to women's issues and blind adherence to the policy priorities of political sponsors – lead to inattention to women's concerns in the policy process.

These dynamics, however, are again rooted in women's lack of presence in the political sphere. When quotas are introduced in a context where women have largely been absent, and especially where political life is governed by dynamics of patronage, quota women often do not have the skills or resources that would make them less vulnerable to manipulation (Cornwall and Goetz, 2005). In some cases, further, quota women may need to tread carefully in response to harassment, intimidation, or security concerns (Wordsworth, 2007). Despite these negative findings, there is some evidence to suggest that quota women are making a difference 'behind the scenes,' proposing more and more bills related to women's rights (Franceschet and Piscopo, 2008), participating disproportionately in debates on issues concerning women (Wang, 2013), and lobbying ministers and party leaders to pursue women's rights initiatives (Cowley and Childs, 2003).

Public Attitudes

A third and fourth group of assertions made during quota debates concerns the broader meaning of quotas for gendered ideas about the 'public' and the 'private.' As feminist theorists have long noted, men have tended to be associated with the public sphere of the economy and politics, while women have often been relegated to the private sphere of the home and family (Elshtain, 1981). The power of this divide has meant that women's movements have needed to mobilize in order to open up access for women to public sphere activities, including work and politics, which have traditionally been viewed as 'male' spaces. In such a context, quotas may serve an important symbolic purpose in legitimizing women as political actors, overcoming enduring gender gaps in women becoming candidates and elected officials. Opponents contend, however, that quotas may instead prove to be counterproductive: through 'artificial' solutions, quotas may reinforce negative stereotypes about female politicians as not belonging in politics or being able to make it there 'on their own.'

Some evidence bolsters the view that quota introduction may alter traditional gendered views about women in political life. A series of studies on India indicate that exposure to female leaders as a result of quotas can weaken gender stereotypes, eliminating negative bias in how the performance of female leaders is perceived, especially among male constituents (Beaman et al., 2009); enhancing the electoral prospects of women once seats are no longer reserved for women (Bhavnani, 2009); and even altering the aspirations

and levels of educational attainment that parents envision for their daughters (Beaman et al., 2012). Focus groups in Rwanda reveal, along similar lines, that gender quotas have led to a broader transformation in citizens' relationships in the private sphere, empowering women within their families as well as communities (Burnet, 2011).

Other work finds, however, that outward acceptance of the legitimacy of quotas often masks continued resistance. This is perhaps most obvious in the failure to fully implement quotas, whether these are state- or party-mandated (see above and Hinojosa, 2012). Resistance to quotas, however, may also take on less overt forms. Research finds, for example, that even years after quota introduction, many male (but not female) elites still do not accept quotas, denying the role of structural patterns of discrimination in favor of attributing female underrepresentation to choices made by individual women (Meier, 2008). Others simply 'forget' to implement quota policies, even when they do not openly speak out against them (Holli et al., 2006).

Political Engagement

Another way in which quotas may contribute to the erosion of the public/private divide is by inspiring female citizens to get more politically involved. As a signal that politics is open to women, these policies may, for example, lead women to turn out to vote, engage in protests, or discuss politics with friends at greater rates than prior to quota adoption, as well as provide new role models that lead them to put themselves forward as political candidates. On the flipside, skeptics suggest that the ongoing controversies surrounding quotas may serve instead to reinforce negative stereotypes about female politicians and thus further discourage women's political participation (Franceschet et al., 2012).

On balance, the evidence points to more positive than negative effects. Case studies indicate, for example, that quotas increase the rate at which female voters contact their political representatives (Childs, 2004; Kudva, 2003). Others find that the adoption of quotas has the effect of encouraging women to begin a political career, acquire political skills, and develop sustained political ambitions (Geissel and Hust, 2005). Still other work observes that quotas help build support for women's movement organizing more generally (Sacchet, 2008). At the same time, however, other scholars conclude that quotas have little or no effect on women's political activities, such as their willingness to sign petitions or participate in protests (Zetterberg, 2009). Even more troubling, some suggest that the presence of quotas may in fact be associated with the decreased strength (Britton, 2005) and increased repression of women's groups (Longman, 2006).

Conclusion

Gender quotas are a recent global phenomenon, taking diverse forms and reaching the political agenda for both feminist and nonfeminist reasons. Although women still do not occupy half of all seats in national parliaments (for up-to-date statistics, see http://www.ipu.org/wmn-e/classif.htm), a survey of global trends is encouraging: women appear to have shattered the political glass ceiling in countries with a diverse array of social, economic, cultural, and political characteristics. Quotas have played a key role in propelling these dramatic gains, interrupting existing dynamics of candidate selection that have tended to favor men over women. Quotas have not, however, always had their intended effects: the numbers of women elected in the wake of quota provisions has in some cases dramatically improved, while in others the level of female representation has stagnated and even declined. The broader effects of quotas also tend to be the subject of ongoing discussion, raising questions as to whether or not they empower women or make a difference in national political life.

More systematic research is needed in order to determine the effects of quotas on the election and empowerment of women. At least three new directions of research will be valuable in helping to parse out the conditions under which quotas might best serve women as a group – as well as contribute to positive social, economic, and political transformation in the countries in which they are introduced. One is to more closely consider questions of context: not only may quotas 'mean' different things across countries, but a variety of factors – age of democracy, level of development, and nature of party system, among others – may provide distinct opportunities and constraints for cultivating a diverse pool of representatives, promoting greater attention to women's policy concerns, and reorienting how citizens view women in politics and the political process more generally.

A second direction is to more explicitly incorporate the issue of time into quota analyses: a growing number of studies suggest that the impact of quotas may evolve from election to election, fostering more positive or negative effects over time. The conclusions drawn from one study may thus not generalize across all time points, with a broader time horizon being required in order to gauge how and why quotas have one impact as opposed to another at any given time. A third direction for future work is to address in greater depth the interconnections between different indicators of impact. The profile of legislators, for example, may shape the prospects to represent women's interests in policy-making – in turn shaping the symbolic effects of quotas on the public at large (see Franceschet et al., 2012). Gender quotas are thus likely to remain an important topic in social and behavioral research for a long time to come.

Finally, it is worth noting that quotas alone may not suffice to achieve gender equality in elected office. On the one hand, the application of quotas may be enhanced by programs expanding the pool of potential female candidates, supporting women's legislative capacities, and promoting broader transformation in public views toward women in politics. On the other hand, formal quotas may not be an option in all states and political parties, most often due to institutional barriers and ideological objections to the concept of positive action. Further, poorly designed policies can prove ineffective in achieving their stated goals. In these cases, alternative strategies are required to foster greater gender equality.

In contrast to the extensive literature on quotas, however, nonquota measures have not been subject to systematic documentation or research. Initial research suggests that actors on the ground in various countries have devised a number of

creative interventions for overcoming women's exclusion from political office, including public financing, political party initiatives, civil society programs, and parliamentary reforms (Krook and Norris, 2014). The sheer variety of these measures reveals a wide menu of options for promoting women in politics, while also underscoring the need for a multifaceted approach to tackle the multidimensional and diverse nature of obstacles to women's political inclusion.

See also: Electoral Systems; Feminist Political Theory and Political Science; Gender History; Gender and the Law; Organizations and Culture; Parliamentary Government; Proportional Representation.

Bibliography

Beaman, Lori, Duflo, Esther, Pande, Rohini, Topalova, Petia, 2012. Female leadership raises aspirations and educational attainment for girls: a policy experiment in India. Science 335 (6068), 582–586.

Beaman, Lori, Chattopadhyay, Raghabendra, Duflo, Esther, Pande, Rohini, Topalova, Petia, 2009. Powerful women: does exposure reduce bias? Quarterly Journal of Economics 124 (4), 1497–1540.

Bhavnani, Rikhil R., 2009. Do electoral quotas work after they are withdrawn? Evidence from a natural experiment in India. American Political Science Review 103 (1), 23–35.

Britton, Hannah, 2005. Women in the South African Parliament: From Resistance to Governance. Lynne Rienner, Boulder.

Burnet, Jennie E., 2011. Women have found respect: gender quotas, symbolic representation, and female empowerment in Rwanda. Politics & Gender 7 (3), 303–334.

Childs, Sarah, 2004. New Labour Women MPs: Women Representing Women. Routledge, London.

Cornwall, Andrea, Goetz, Anne Marie, 2005. Democratizing democracy: feminist perspectives. Democratization 12 (5), 783–800.

Cowley, Philip, Childs, Sarah, 2003. Too spineless to rebel? New labour's women MPs. British Journal of Political Science 33 (3), 345–365.

Dahlerup, Drude (Ed.), 2006. Women, Quotas, and Politics. Routledge, New York.

Elshtain, Jean Bethke, 1981. Public Man, Private Woman: Women in Social and Political Thought. Princeton University Press, Princeton.

Franceschet, Susan, Krook, Mona Lena, Piscopo, Jennifer M. (Eds.), 2012. The Impact of Gender Quotas. Oxford University Press, New York.

Franceschet, Susan, Piscopo, Jennifer M., 2008. Gender quotas and women's substantive representation: lessons from Argentina. Politics & Gender 4 (3), 393–425.

Geissel, Brigitte, Hust, Evelin, 2005. Democratic mobilisation through quotas: experiences in India and Germany. Commonwealth & Comparative Politics 43 (2), 222–244.

Goetz, Anne Marie, Hassim, Shireen (Eds.), 2003. No Shortcuts to Power: African Women in Politics and Policy Making. Zed Books, New York.

Hinojosa, Magda, 2012. Selecting Women, Electing Women: Political Representation and Candidate Selection in Latin America. Temple University Press, Philadelphia.

Holli, Anne Maria, Luhtakallio, Eeva, Raevaara, Eeva, 2006. Quota trouble: talking about gender quotas in Finnish local politics. International Feminist Journal of Politics 8 (2), 169–193.

Hughes, Melanie M., 2011. Intersectionality, quotas, and minority women's political representation worldwide. American Political Science Review 105 (3), 604–620.

Jones, Mark P., 1998. Gender quotas, electoral laws, and the election of women: lessons from the Argentine provinces. Comparative Political Studies 31 (1), 3–21.

Kittilson, Miki C., 2006. Challenging Parties, Changing Parliaments: Women and Elected Office in Contemporary Western Europe. Ohio State University Press, Columbus.

Krook, Mona L., 2006a. Gender quotas, norms, and politics. Politics & Gender 2 (1), 110–118.

Krook, Mona L., 2006b. Reforming representation: the diffusion of candidate gender quotas worldwide. Politics & Gender 2 (3), 303–327.

Krook, Mona L., 2009. Quotas for Women in Politics: Gender and Candidate Selection Reform Worldwide. Oxford University Press, New York.

Krook, Mona L., Lovenduski, Joni, Squires, Judith, 2009. Gender quotas and models of political citizenship. British Journal of Political Science 39 (4), 781–803.

Krook, Mona L., Norris, Pippa, 2014. Beyond quotas: strategies to promote gender equality in elected office. Political Studies 62 (1), 2–20.

Krook, Mona L., True, Jacqui, 2012. Rethinking the life cycles of international norms: the United Nations and the global promotion of gender equality. European Journal of International Relations 18 (1), 103–127.

Kudva, Neema, 2003. Engineering elections: the experiences of women in Panchayati Raj in Karnataka, India. International Journal of Politics, Culture and Society 16 (3), 445–463.

Longman, Timothy, 2006. Rwanda: achieving equality or serving an authoritarian state? In: Britton, Hannah, Bauer, Gretchen (Eds.), Women in African Parliaments. Lynne Rienner, Boulder, pp. 133–150.

Meier, Petra, 2008. A gender gap not closed by quotas: the renegotiation of the public sphere. International Feminist Journal of Politics 10 (3), 329–347.

Murray, Rainbow, 2007. How parties evaluate compulsory quotas: a study of the implementation of the 'parity' law in France. Parliamentary Affairs 60 (4), 568–584.

Murray, Rainbow, 2010. Second among unequals? A study of whether France's 'quota women' are up to the job. Politics & Gender 6 (1), 93–118.

Sacchet, Teresa, 2008. Beyond numbers: the impact of gender quotas in Latin America. International Feminist Journal of Politics 10 (3), 369–386.

Sater, James N., 2007. Changing politics from below? Women parliamentarians in Morocco. Democratization 14 (4), 723–742.

Schwartz, Helle, 2004. Women's representation in the Rwandan parliament (M.A. thesis). University of Gothenburg, Sweden.

Schwindt-Bayer, Leslie A., 2009. Making quotas work: the effect of gender quota laws on the election of women. Legislative Studies Quarterly 34 (1), 5–28.

Towns, Ann E., 2010. Women and States: Norms and Hierarchies in International Society. Cambridge University Press, New York.

Tripp, Aili M., Kang, Alice, 2008. The global impact of quotas: on the fast track to increased female legislative representation. Comparative Political Studies 41 (3), 338–361.

Wang, Vibeke, 2013. (Ph.D. dissertation). University of Bergen, Norway.

Wordsworth, Anna, 2007. A Matter of Interests: Gender and the Politics of Presence in Afghanistan's Wolesi Jirga. Issues Paper Series. Afghanistan Research and Evaluation Unit.

Zetterberg, Pär, 2009. Do gender quotas foster women's political engagement? Lessons from Latin America. Political Research Quarterly 62 (4), 715–730.

Femininities

Elroi J Windsor, Salem College, Winston-Salem, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Within the broad framework of heterosexuality, scripts of femininity operate as a way to organize gendered and sexual relationships between women and men. Girls and women learn to be feminine, and this socialization relates to interpersonal and institutional gender and sexual inequalities. The concept of hegemonic, or normative, femininity relates to masculinity constructs, as well as alternative forms of femininity marked by race, class, and sexual identity.

Scripts of Femininity

'Femininity' is a familiar term. Conversations about being feminine are common in everyday life and many people use the word 'feminine' to describe themselves and others. They may equate femininity with being a woman who embodies characteristics like being nurturing, sensitive, demure, or sweet. But femininity cannot be understood as a fixed set of essential traits that characterize all women. As a scholarly concept, femininity can carry diverse meanings with numerous interpretations. Within the context of heterosexual relationships, performances of femininity can employ different scripts. These scripts act as guidelines for individual behavior and social interaction. They are learned at an early age and reinforced throughout the life course.

When examined as a whole, individual expressions of femininity reveal distinct patterns. These themes become reinforced throughout different social institutions such as media, education, religion, sports, and the workforce. Studying these institutional or macrolevel forms makes it possible to see how ideas about femininity represent a much larger concept than simply wearing makeup and high heels while smiling coyly and sitting with one's legs crossed. Pulling back the conceptual lens reveals that femininity is a socializing ideology that defines and organizes material ways of life, particularly practices related to gender and sexual relationships. This ideological concept also casts some forms of femininity as hegemonic, or dominant, while marginalizing and subordinating other versions of femininity. These manifestations of femininity illustrate the varied ways femininities have been defined in Western cultures.

Learning to Be Feminine

Every child learns how to be a girl or a boy. The society in which a child is born contains plentiful information about how to express one's gender. These gender expectations are conveyed through numerous social institutions, such as media, law, and school. Gender socialization also occurs interpersonally, where family members, friends, and peers instruct children about appropriate gendered behavior. This process is known as the social construction of gender, and it gives meanings to words like girl/boy, woman/man, and feminine/masculine (Lorber, 1994).

When a young girl is taught how to act like a young lady, she is being taught the scripts of femininity. These messages reinforce what it means to be feminine, and, consequently, what it means to be a girl. In turn, boys receive a different set of messages about masculinity, which reference everything that a girl or woman is not. The social construction of gender relies on a binary gender system that positions men and masculinity as fundamentally different from women and femininity.

Each person then continues to 'do' gender throughout their lives, ensuring that they are signaling the right cues to be read as the gender they wish to present to others. Girls and women, then, do gender by demonstrating conventionally feminine characteristics (West and Zimmerman, 1987). Learning these femininity scripts is crucial for doing gender successfully. These scripts teach women and girls how to attract men's sexual attention. Consequently, the relationship between gender socialization and heterosexual relationships has been an important site for scholarly inquiry.

Theorizing Femininities

Scholars have illustrated how gender socialization and learning about femininity manifest in heterosexual relationships. They identify gender inequality as foundational in establishing relationships between women and men. Within most contemporary Western societies, hierarchical gender relationships shape much of social life. Male domination can be seen throughout social institutions like religion, science, and government. In this patriarchal context, women are subordinated to men. This gender stratification system also relies on conventional meanings of femininity, which include characteristics such as "physical vulnerability, an inability to use violence effectively, and compliance" (Schippers, 2007: p. 91). These defining features relegate femininity to a subordinate form compared to the more powerful construct of masculinity.

In some ways, gender scholarship has contributed to the continued subordination of femininity. Unlike the theoretical and empirical studies of masculinity, the concept of femininities has been undertheorized. Sociologist and feminist scholar R.W. Connell introduced the theoretical study of 'hegemonic masculinity' in 1987 (Connell, 1987), which sparked numerous theoretical and empirical studies that further illustrated, enriched, and solidified the concept as a paramount

feature of gender studies. Yet it took 20 more years before gender scholars dedicated similar analytical attention to the parallel concept of hegemonic femininity. Mimi Schippers first theorized about hegemonic femininity in her 2007 publication, 'Recovering the Feminine Other: Masculinity, Femininity, and Gender Hegemony'. In this work, Schippers describes hegemonic femininity as "the characteristics defined as womanly that establish and legitimate a hierarchical and complementary relationship to hegemonic masculinity and that, by doing so, guarantee the dominant position of men and the subordination of women" (Schippers, 2007: p. 94). This theoretical definition implicates femininity and women in hierarchical gender systems where hegemonic masculinity is continually able to sustain men's dominance over women.

In these conceptualizations of hegemonic gender expressions, men are expected to be powerful, strong, independent, and emotionally detached. A core component of hegemonic masculinity is men's desire for, and possession of, their binary opposite: the feminine object as embodied through women (Connell, 1987). Femininity is organized as antithetical to masculinity, and so is unable to embody the same characteristics without becoming stigmatized. In order to maintain men's dominance, the characteristics of femininity must not threaten the patriarchal gender order. Men who exhibit any indicators of femininity are ridiculed and negatively labeled as weak and womanly. Their exclusion from the hegemonic ideal results in marginalization and ostracism.

As ideological constructs, hegemonic standards of masculinity and femininity inform individual men's dominance over individual women, as well as the stratified patriarchal gender order as a whole (Schippers, 2007). By establishing characteristics of masculinity and femininity as binary opposites, where women and femininity are subordinate to men and masculinity, hegemonic constructs of these gender expressions form the basis for heterosexual interactions. This binary gender system of difference serves as the foundation for heterosexual erotic power within Western societies, where desire is thought of as resulting from attractions to difference. This heterosexual matrix functions as a natural consequence of the complementary and hierarchical relationship between masculinity and femininity, the meanings of which are constructed through these concepts of difference (Butler, 1990).

In the heterosexual matrix, conventional femininity is heralded as necessary for all women. A core assumption of hegemonic femininity is that women need to embody specific physical characteristics in order to successfully attract men. Within hetero-patriarchy, compliance with gender norms becomes vital and heterosexuality is compulsory (Butler, 1990). Women who seek men's attention must work to ensure they present an aesthetically pleasing amalgamation of face, body, dress, and mannerism. They may give in to a 'beauty myth' where, in order to possess adequate cultural or body capital, women need to conform to conventional standards of attractiveness (Wolf, 2002). The beauty myth is fueled by a profit-motivated industry that markets endless supplies of products designed to 'improve' one's natural appearance (Wolf, 2002). Women's pursuit and consumption of beautification technologies, such as plastic surgery and dieting, reflect cultural ideals about attractiveness and the feminine body. By attempting to embody these standards of attractiveness, women compete with each other for men's attention. These practices, with institutional support, reproduce a narrow version of heterosexual femininity as normative (Bordo, 1993).

Consequences of Upholding Femininity

Through institutionalized gender inequality, femininity becomes transformed as a material good expected from all women – but not without consequences. Women who do not conform to hegemonic femininity standards risk having their womanhood and sexuality questioned. Men, as well as other women, may view less feminine women as unattractive, lazy, apathetic, and lesbian (Wolf, 2002). Interpersonally, women who are less conventionally attractive are more lonely, less popular, and less sexually experienced than normatively attractive women. Compared to attractive women, unfeminine women are less likely to marry and less likely to marry men who belong to higher socioeconomic statuses (Sullivan, 2001). These stereotypes affect women's ability to connect with other women and men.

In addition to less rewarding personal relationships, less feminine women can also expect to experience significant losses in other central areas of life. In the workforce, women who are conventionally attractive are more often hired, promoted, and paid higher salaries than women who are perceived as unattractive (Sullivan, 2001). Within sports, women athletes often must display masculine traits such as competiveness, aggression, and dominance to excel in their games. Yet they face pressures to submit to a contrary sexualized discourse within sports that demands players embody femininity and heterosexuality (Mean and Kassing, 2008). The unforgiving arena of politics holds little compassion for women candidates, as exemplified in criticisms directed at Hillary Clinton for being too aggressive and unattractive and dismissals of Sarah Palin for being too sexy and dim-witted. Based on negative media representations, both of these high-profile politicians failed at straddling the precarious line of being assertive but not too masculine, and attractive, but not too feminine, when running for political office (Farmer, 2009). These institutional barriers inform the hetero-patriarchy that structures the gender stratification system.

Women often internalize these macrolevel messages. They may focus on maintaining an attractive outward appearance to garner approval from men and other women and to succeed in their careers. Yet a commitment to upholding normative standards of femininity can compromise women's physical and psychological well-being. Failure to comply with normative femininity can result in lower self-confidence, disordered eating, anxiety, and depression (Wolf, 2002). Thus, the dialectical relationships between cultural ideals of femininity and the practices women pursue to achieve them represent a systematic form of social control (Bordo, 1993). This social control of gender carries over into women's sexual relationships.

Femininity and Sexuality

Ideas about femininity, womanhood, and heterosexuality inform a person's sexual scripts. The concept of 'sexual scripts'

was first introduced in 1973 by John H. Gagnon and William Simon. These sociologists developed a theory of sexual interaction in response to biological explanations for sexual behavior that they deemed inaccurate (Simon and Gagnon, 2003). Gagnon and Simon believed that people negotiated sex and sexuality based on cultural messages, interactional cues, and internalized understandings, not biological urges or innate desire. They argued that social scripts shaped a person's sexual attitudes, behaviors, and identities (Gagnon, 1973; Gagnon and Simon, 1973; Simon, 1973; Simon and Gagnon, 2003). As gender and sexuality are intertwined, sexual scripts closely relate to constructions of femininity. Heterosexual women interpret norms about femininity and gender to negotiate their intimate relationships with men.

Gender inequality pervades much of social life and society's institutions, so it is inevitable that these dynamics will affect romantic and sexual relationships between women and men. Hegemonic femininity contributes to hetero-patriarchal imaginings of sexuality that render women submissive to men's desires. Just as women receive lessons about gender at a very young age, so too do they absorb how these femininity scripts relate to heterosexuality. Baby girls are often described as 'flirting' with boys and men who pay them attention, and girls learn early on that their (heterosexual) wedding day will be the best days of their lives. Research shows that adolescent girls consume media that emphasize looks and finding a boyfriend. Magazines marketed to teen girls showcase editorials and advertisements that advise readers how to be feminine and beautiful. Even though contemporary magazines pepper these messages with empowering rhetoric more so now than in past decades, the ends are the same: girls must be pretty and conventionally feminine to capture and retain men's desire (Gengler, 2011).

Research on young women's gender and sexual identities suggests that girls heed these media messages. Girls' body projects cater to the whims of male pleasure. Typically, girls need to give in to compulsory heterosexuality in order to forge successful gender and sexual identities. They police their own and other girls' bodies, desires, and behaviors to conform to hegemonic standards of hetero-femininity (Renold, 2000). Yet young women must be careful to avoid being seen as too sexually available, lest they be negatively labeled as sluts or whores. While teenage boys are often encouraged to actively desire sex, young women are expected to walk a fine line between being seen as sexually appealing but not acting in sexually promiscuous ways. This sexual double standard makes it difficult for young women to assert sexual agency and denies them gender equality in their intimate relationships (Tanenbaum, 2000).

Attempts to challenge the sexual double standard are not always effective. Girls who resist the idea that they should focus on appeasing male desire are only successful when they are able to demonstrate their heterosexuality (Renold, 2000). Similarly, teens who try to assert a postfeminist version of 'girl power' often can only do so in ways that bolster class-privileged, white, and heteronormative femininity (Allan, 2009). Thus, girls may mobilize sexual subjectivities through promoting neoliberal femininities (Charles, 2010). However, other research suggests that young women may challenge normative heterosexual scripts that require submission to men's desires. With a critical consciousness, young women may be able to assert agency through initiating sex, setting the terms of virginity loss, and openly communicating and acting upon their sexual desires (Stewart, 1999).

In adulthood, limited scripts of femininity often remain the same for heterosexual women. These women may sacrifice their own sexual desires and pleasures if they internalize messages that reinforce the sexual double standard (Crawford and Popp, 2003). Yet some adult women do try to project a sexual autonomy for themselves and their partners. Some women proudly claim a conventionally feminine appearance while also advocating feminist principles. They argue that women do not need to discard the trappings of femininity in order to challenge gender oppression (Stoller, 1999). Although these individual philosophies may be personally empowering, research suggests that they are insufficient to challenge the hetero-patriarchal gender order. For example, one ethnographic study on women's attendance at a strip club that featured male dancers found that the club's workers fostered heterosexist behaviors that diminished women's sexual autonomy. Although some women found the erotic space of the club to be a rare site for exhibiting their own sexual aggression, many did not (Pilcher, 2011). These disparate experiences illustrate that gender inequality in heterosexual contexts persists.

Feminine Sexual Bodies

Patriarchal constructs of binary gender also affect how people think about sexual bodies. Heterosexual sex is often framed as the penile–vaginal coital act that culminates with male ejaculation. This action positions women's sexual bodies as soft, passive, and compliant. Men use their harder, stronger bodies to penetrate women. They are thought of as acting on their more aggressive desires that must be fulfilled through orgasmic climax (Johnson, 1997). Women's vaginas are portrayed as receptacles to men's penises, instead of framing them as actively taking in hardness and ejecting flaccidity. One example of this paradigm can be seen in female genital cosmetic surgery. This procedure promotes reconstructing the vulva to accommodate men sexually. Yet it is often paradoxically marketed as sexually liberating for women (Braun, 2005). Such procedures increasingly medicalize women's sexual bodies, with healthcare professionals prescribing a range of treatments based on assumptions about women's desire. Treatments include off-label uses for drugs like Viagra, which are usually prescribed to men who seek relief from erectile dysfunction (Hartley, 2006). These surgical and medical interventions treat women's sexual bodies as passive objects of men's desires. They do little to understand the depth of women's sexuality or encourage whole-body approaches to sex that transcend penile–vaginal intercourse.

Rethinking Normative Hetero-Femininity: Race, Class, and Sexual Identity

In scholarship on femininities and heterosexuality, the influences of race, class, and sexual identity minorities are often mentioned, but not sufficiently developed. In response, some

scholars have chosen to articulate categorically specific theories and studies of non-hegemonic femininities that enrich current scholarship. These insights provide more nuanced understandings of marginalized communities.

Hierarchies of Race and Class

Hegemonic femininity reflects not only gender stratification, but hierarchies based on race, class, and sexual identity as well. Like hegemonic masculinity, the most privileged feminine gender expression is embodied by women who are also white, middle- to upper-middle class, and heterosexual. Alternative forms of feminine expression become constructed as inferior when they are believed to be more representative among women of color and women of lower socioeconomic statuses. Women in these communities may encounter negative sanctions for disrupting white expressions of femininity (Schippers, 2007). The conventional meanings of femininity as passive, dependent, and materialistic have been critiqued as representing white, middle-class women's values (Collins, 2000).

One alternative theoretical construction of femininity has been proposed by feminist theorist Patricia Hill Collins. In her groundbreaking work, *Black Sexual Politics: African Americans, Gender, and the New Racism*, Collins theorizes Black femininity. She illustrates how controlling images of Black women in media reflect racist stereotypes about Blackness. These images are also class-differentiated. Working-class Black women are portrayed in media as aggressive, domineering, and hypersexual, while middle-class portrayals evoke images of Black women as appropriately ambitious yet simultaneously subservient to white male authority. Middle-class Black women are also represented as more respectable sexually. These controlling images help justify racism and sexism against Black women. They promote a sexual politics of respectability that helps perpetuate gender inequality and partner violence in relationships between Black women and men (Collins, 2005).

Collins's ideas are supported in research that shows the effects of media messages on African-American preadolescents, especially within hip hop music and video media. For example, popular representations of Black women in these media influence youth to equate women's sexual desire with men's expectations (Stephens and Few, 2007). These images compromise a more autonomous sexuality among Black women. Yet some argue that it is possible for Black women to invoke these representations within Black culture to resist or challenge oppressive ideologies, as the forum of hip hop often mobilizes ideas of self-promotion (Collins, 2005).

Another alternative construction of femininities focuses on Asian American femininities. Research with Asian Americans reveals how women in this heterogeneous race group can internalize messages about the authority of normative white femininity, constructing Asian femininities as subordinated to the dominant version. In one study with Korean and Vietnamese Americans, these women used race as the primary basis for explaining what femininity meant to them. In white normative contexts, they felt pressure to submit to stereotypes about Asian American women as submissive and docile even though doing so disgusted them. In other situations, they felt compelled to distance themselves from stereotypes about Asian femininity in order to be taken seriously by dominant groups. Both behaviors subjugated Asian femininities to normative white standards of femininity, reifying the latter as hegemonic and personally ideal (Pyke and Johnson, 2003).

Overall, theoretical and empirical alternatives to hegemonic femininity have become more culturally available. They complicate normative representations of hetero-femininity. However, the increased visibility of different femininities does not always represent a positive shift in portrayals. Instead, representations within media often perpetuate race and class hierarchies. For example, one study on advertisements within the pages of teen girls' fashion magazines showed how the magazine equated beauty with whiteness, hypersexuality with Blackness, and technological skills with East Asians (Sengupta, 2006). These images reflected stereotypes about race and gender. In addition, media representations are not just abstract images; they can produce genuine effects in everyday life. One study on Latina women and media showed how Latina girls negotiated sexual agency within the complex realities of racism, patriarchy, poverty, and heterosexual privilege. In this study, some Latino families used media as a forum to initiate dialog about sexuality with daughters (Garcia, 2012). These race-based constructions of femininity offer a more complex understanding of hegemonic femininity.

Queering Femininities

Women who are not heterosexual have also proposed alternative femininity scripts that are useful to examine. In addition to expanding understandings of marginalized sexual communities, these scripts of femininity offer alternative visions of gender that include possibilities for heterosexual women who seek to challenge hetero-patriarchy.

Like femininity among heterosexual women, meanings of femininities vary among sexual minorities. For women who describe their sexual identities as lesbian, gay, bisexual, or queer (LGBQ), femininity can take on different meanings. Some feminine women who are sexual minorities embrace normative standards of femininity and internalize misogynist messages. Although these women may appear to meet hegemonic femininity standards, their sexual identities complicate the expectations of women who look conventionally feminine. Other LGBQ women, however, use their femininity in ways that try to reject both heteronormativity and heterosexism (Palder, 2008). Women who identify as 'femme' often seek to use feminine characteristics, or embody a hyperfemininity, as a source of power (e.g., Newman, 1995; Rose and Camilleri, 2002; Volcano and Dahl, 2008). Femme women invoke purposeful feminine expressions and assertive queer sexualities to challenge hetero-patriarchal male gazes. They resist viewing the trappings of femininity as inherently oppressive and argue that people of all genders can use femininity as a source of gender and sexual empowerment.

With a similar logic, popular writer, activist, and biologist Julia Serano asserts that femininity can and should be a source of empowerment for people, regardless of gender or sexuality. Writing as a transsexual woman, Serano proposes "putting the feminine back into feminism" (2007: p. 319). Serano denounces the hatred of femininity espoused within popular culture and by some feminist groups. She argues for a heterogeneous understanding of femininity that is decoupled from

femaleness. Her ideas 'queer' femininity by envisioning this form of gender expression and embodiment as potentially liberating for everyone.

Conclusions

Scholarship on femininity has contributed a wealth of knowledge to broader studies of gender and sexuality. As an ideological concept, hegemonic femininity theorizes the gender expressions and experiences of girls and women as learned. Institutions and individuals teach the scripts of femininity, and this education occurs throughout life as girls and women are socialized into conventional expressions of gender which treat them as distinct from, and subordinate to, the gender expressions of boys and men. The social construction of gender then reinforces binary classifications of gender which replicate gender inequalities. As a dominant form, hegemonic femininity also subordinates alternative versions of femininity, especially those marked by differences in race, class, and sexual identity.

These gender constructions are intimately connected to women's sexual relationships with men. The binary gender system serves as foundational for the hetero-patriarchal erotic matrix, where men continue to hold power over women. In this context, women compromise their own sexual agency to attract, please, and retain men. This dynamic creates a sexual double standard where women get punished for exhibiting a more active and desiring sexuality. Yet some scholarship suggests that women and girls do exact agency in negotiating sexuality, albeit with limited success. In addition, critiques from the margins of race, class, and sexual identity complicate understanding femininity as a monolithic, homogeneous form applicable to all women at all times. Future directions in the scripts of femininities will continue to complicate the interconnected relationship between gender and sexuality.

See also: Adolescent Sexual Risk; Beauty and the Labor Market; Erotic Capital; Feminisms and Acculturation around the Globe; Feminist Political Theory and Political Science; Gender and Women's Studies, Applied Research On; Gender: Gendering of Categories; Heterosexuality; Lesbian, Gay, Bisexual and Trans-sexual Butch-Femme Subcultures; Masculinity, Scripts of; Parallel Relationships; Queer Theory and Intersectionality; Sexuality Over the Life Course; Sexuality, Theories of; Social Constructivism; Social Stratification; Teens, Gender, and Self-Presentation in Social Media; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Allan, A.J., 2009. The importance of being a 'lady': hyper-femininity and heterosexuality in the private, single-sex primary school. Gender and Education 21 (2), 145–158.
Bordo, S., 1993. Unbearable Weight: Feminism, Western Culture and the Body. University of California Press, Berkeley and Los Angeles.
Braun, V., 2005. In search of (better) sexual pleasure: female genital 'cosmetic' surgery. Sexualities 8 (4), 407–424.
Butler, J., 1990. Gender Trouble: Feminism and the Subversion of Identity. Routledge, New York.
Charles, C.E., 2010. Complicating hetero-femininities: young women, sexualities and 'girl power' at school. International Journal of Qualitative Studies in Education 23 (1), 33–47.
Collins, P.H., 2000. Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment. Routledge, New York.
Collins, P.H., 2005. Black Sexual Politics: African Americans, Gender, and the New Racism. Routledge, New York.
Connell, R.W., 1987. Gender and Power: Society, the Person and Sexual Politics. Polity Press, Cambridge.
Crawford, M., Popp, D., 2003. Sexual double standards: a review and methodological critique of two decades of research. Journal of Sex Research 40 (1), 13–26.
Farmer, A., 2009. The political pillorying of pantsuits: the media's gender bias in the 2008 presidential campaign. Perspectives: A Magazine for and About Women Lawyers 17 (3), 4–7. http://www.americanbar.org/content/dam/aba/publishing/perspectives_magazine/women_perspectives_Winter09_media_bias.authcheckdam.pdf (accessed 13.08.14.).
Gagnon, J.H., 1973. Scripts and the coordination of sexual conduct. In: Cole, J.K., Deinstbier, R. (Eds.), 1972 Nebraska Symposium on Motivation, vol. 23. University of Nebraska Press, Lincoln, pp. 27–59.
Gagnon, J.H., Simon, W., 1973. Sexual Conduct: The Social Sources of Human Sexuality. Aldine Books, Chicago.
Garcia, L., 2012. Respect Yourself, Protect Yourself: Latina Girls and Sexual Identity. New York University Press, New York.
Gengler, A.M., 2011. Selling feminism, consuming femininity. Contexts 10, 68–69.
Hartley, H., 2006. The 'pinking' of viagra culture: drug industry efforts to create and repackage sex drugs for women. Sexualities 9 (3), 363–378.
Johnson, A.G., 1997. The Gender Knot: Unraveling Our Patriarchal Legacy. Temple University Press, Philadelphia.
Lorber, J., 1994. Paradoxes of Gender. Yale University Press, London.
Mean, L.J., Kassing, J.W., 2008. 'I would just like to be known as an athlete': managing hegemony, femininity, and heterosexuality in female sport. Western Journal of Communication 72 (2), 126–144.
Newman, L. (Ed.), 1995. The Femme Mystique. Alyson Publications, Los Angeles.
Palder, A., 2008. So, who feels pretty? Negotiating the meaning of femininity in a nonheterosexual community. Georgia State University, Unpublished doctoral dissertation. http://digitalarchive.gsu.edu/sociology_diss/36 (accessed 13.05.15.).
Pilcher, K.E.M., 2011. A 'sexy space' for women? Heterosexual women's experiences of a male strip show venue. Leisure Studies 30 (2), 217–235.
Pyke, K.D., Johnson, D.L., 2003. Asian American women and racialized femininities: 'doing' gender across cultural worlds. Gender & Society. 17 (1), 33–53.
Renold, E., 2000. 'Coming out': gender, (hetero)sexuality and the primary school. Gender and Education 12 (3), 309–326.
Rose, C.B., Camilleri, A. (Eds.), 2002. Brazen Femme: Queering Femininity. Arsenal Pulp Press, Vancouver.
Schippers, M., 2007. Recovering the feminine other: masculinity, femininity, and gender hegemony. Theory and Society 36 (1), 85–102.
Sengupta, R., 2006. Reading representations of black, East Asian, and white women in magazines for adolescent girls. Sex Roles 54 (11–12), 799–808.
Serano, J., 2007. Whipping Girl: A Transsexual Woman on Sexism and the Scapegoating of Femininity. Seal Press, Emeryville.
Simon, W., 1973. The social, the erotic and the sensual: the complexities of sexual scripts. In: Cole, J.K., Deinstbier, R. (Eds.), 1972 Nebraska Symposium on Motivation, vol. 23. University of Nebraska Press, Lincoln, pp. 61–82.
Simon, W., Gagnon, J.H., 2003. Sexual scripts: origins, influences and changes. Qualitative Sociology 26 (4), 491–497.
Stephens, D.P., Few, A.L., 2007. Hip hop honey or video ho: African American preadolescents' understanding of female sexual scripts in hip hop culture. Sexuality & Culture 11 (4), 48–69.
Stewart, F.J., 1999. Femininities in flux? Young women, heterosexuality and (safe) sex. Sexualities 2 (3), 275–290.
Stoller, D., 1999. Feminists fatale: BUST-ing the beauty myth. In: Karp, M., Stoller, D. (Eds.), The BUST Guide to the New Girl Order. Penguin Books, New York, pp. 41–47.
Sullivan, D.A., 2001. Cosmetic Surgery: The Cutting Edge of Commercial Medicine in America. Rutgers University Press, New Brunswick.
Tanenbaum, L., 2000. Slut! Growing up Female with a Bad Reputation. HarperCollins Publishers, New York.
Volcano, D.L., Dahl, U. (Eds.), 2008. Femmes of Power: Exploding Queer Femininities. Serpent's Tail, London.
West, C., Zimmerman, D.H., 1987. Doing gender. Gender & Society 1 (2), 125–151.
Wolf, N., 2002[1991]. The Beauty Myth: How Images of Beauty are Used against Women. Harper Perennial, New York.

Feminism in Anthropology

Louise Lamphere, University of New Mexico, Albuquerque, NM, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Feminism in anthropology in the United States emerged in the mid-1970s with the publication of two collections that examined women's subordination, power, and influence. By the 1980s, US anthropologists were rediscovering their feminist foremothers and turning their attention from only women to an analysis of gender. Feminists paid more attention to history, class, race, and sexuality as gender relations. Since 2000, feminists have taken a global perspective in studying immigration, sexuality, masculinity, new reproductive technologies, neoliberalism, and gender violence. Despite some claims that the twenty-first century is a postfeminist era, US feminist anthropologists are participating in a variety of activist movements and collaborating with feminists on other continents.

The 1970s saw the flowering of feminism within US anthropology. (Due to space limits, the author has confined this article to feminist anthropologists from the United States. The author has also not covered feminists whose specialty is either biological anthropology or language.) Influenced by the American feminist movement, women scholars, including professors, graduate students, and faculty wives, began teaching courses and delivering papers on women's roles in a wide variety of cultures. Two collections, published in 1974 and 1975, marked this new interest in women. *Woman, Culture and Society* (Rosaldo and Lamphere, 1974) coalesced around Michelle Z. Rosaldo, Jane Collier, and others at Stanford University and Louise Lamphere at Brown University. Another collection edited by Rayna Reiter (later Rapp) emerged from a number of feminist anthropologists at the University of Michigan (Reiter, 1975).

Many women in both of these networks were involved in the Vietnam antiwar movement as well as in the broader women's movement. They took part in consciousness-raising groups, testified at hearings on abortion rights, participated in rallies and marches, and attended conferences (Lamphere et al., 1997). Both collections focused on issues of equality and subordination and how to think about women's position cross-culturally. Feminists in other disciplines looked to anthropology to answer questions like, "Are women subordinate in every culture (as well as in our own)? Were there matriarchies where women ruled? Is there a biological basis for sex differences? And what is the role of the state in women's subordination? Analytical frameworks (structural, cultural, evolutionary, Marxist) reflected the graduate school training of many authors and many had not focused their field research on women's roles."

In 40 years since these two collections were published, feminist anthropology in the United States has changed. The number of feminist anthropologists has increased and diversified as the topics they study have multiplied. At least two new cohorts of scholars have conducted feminist research. More of these are minority women, men (both minority and white), women raised in non-Western traditions, and self-identified gay and lesbian anthropologists. Feminist anthropology has become international. Feminist analysis has blossomed especially in archaeology and in biological anthropology and linguistic anthropology. The emphasis has shifted. Gone is the focus on binary categories and sweeping universals. Writing about gender has replaced writing about women. More attention is paid to problematizing categories, attending to historic contexts, and experimenting with innovative writing strategies. Since 2000, feminist anthropologists are more engaged with issues of identity and difference, and the intersection of race, class, gender, and sexuality. Feminists speak about 'engendering' the discipline, and scholars have brought feminist perspectives to a wide range of topics such as globalization and transnationalism, new reproductive technologies, sexuality, gender violence, immigration, and the study of neoliberalism, the financial sector, and economic crisis.

Early Frameworks

The essays in *Woman, Culture and Society* emerged from a course at Stanford University in the Spring 1971, as well as papers delivered at the 1971 American Anthropological Association meetings, and the networks of the editors. The introductory essays by Rosaldo, Ortner, and Chodorow offered an integrated set of explanations for women's universal subordination of women, that is, in terms of social structure, culture, and socialization. All three argued that in every society women bear and raise children, and women's role as mother provides the basis for subordination. Women are confined to a domestic sphere (Rosaldo) or associated with nature rather than culture and seen as ambiguous, dangerous, and devalued (Ortner).

The Reiter collection also emerged out of a class, taught by women graduate students in 1971. Several articles reflected the four-field evolutionary and even Marxist orientation of the Michigan Department, in contrast to what many say is the structuralist and symbolic approach of the Rosaldo and Lamphere volume. Two articles had a lasting impact. Gayle Rubin's article "The Traffic in Women" pioneered the use of theory (Marx, Levi-Strauss, Lacan) to understand women's subordination and the creation of a normative heterosexuality. Reiter argued that the domestic/public divide coincided with the rise of the state. It was the first of several important treatments of women's subordination in relation to class divided, patriarchal states (Sacks, 1979; Silverblatt, 1987; Gailey, 1987).

The articles in both attested to the variety of positions women held in cultures throughout the world. Feminists offered examples where women held power or influence (McCormack, Brown) and case studies where women strategized to achieve their own goals even though men wielded power in patrilineal descent groups (Wolf, Collier, Silverman, Lamphere). Still others found cultures where men and women could be considered equal (Sacks, Gough, Draper). As Gayle Rubin put it "So they [the books] were locally inflected [structural and cultural versus materialist and evolutionary], but the project was a common one, and it was one of those historical tectonic shifts where you don't understand the forces that are impinging on all these different people in different places but they clearly were" (Lamphere et al., 2007: 417). Feminists felt, as Rapp said, "we and our whole generation were reconfiguring the field."

The three introductory essays in *Woman, Culture and Society* gave rise to a debate over women's subordination and the usefulness of dichotomies like domestic/public and nature/culture. Some scholars argued that in small-scale foraging and horticultural societies, women and men held roles that were "complementary but equal" (Schlegel, 1972). Working primarily from a Marxist perspective, Leacock (1980) (Eleanor Leacock was an important analyst of women's roles and mentor to feminists of the 1970s generation. In addition to her collected essays on women's status, she coedited collections on women and colonization and on women's work. She also published a critique of the culture of poverty and edited a new edition of Engels' *The Origin of the Family, Private Property and the State.*), Sacks (1979), and others took the position that colonialism had transformed and undercut many native economies, creating inequality where autonomy had been the norm.

The Turn to Gender, History, and Intersectionality

By 1980, feminists began to turn from 'finding women' to examining relationships the concept of "gender." Rosaldo's article "The Use and Abuse of Anthropology" (1980) was a first step in this direction. However, she continued to argue that women have been universally subordinate. She also began to break apart the categories o 'female' and 'male,' analyzing differences between unmarried young men and women and their married elders. Working with Jane Collier, she said "We are concerned to stress not the activities of women – or of men- alone; instead we are attempting to convey the ways in which a sexual division of labor in all human social groups is bound up with extremely complex forms of interdependence, politics, and hierarchy" (Rosaldo, 1980: 412). This same emphasis on gender and hierarchy is the unifying theme of *Sexual Meanings* edited by Sherry Ortner and Harriet Whitehead (1981). They call for an analysis of 'prestige structures' and suggest that gender systems are themselves prestige structures. Statuses such as warrior, statesman, Brahmin, or elder are seen as part of a male ranking system, while women get defined as wives and mothers, unmarked positions with lower prestige. These four authors pose models that are comparative and not based on universal dichotomies.

Micaela di Leonardo's collection *Gender at the Crossroads of Knowledge* (1991) represents this trend to examine gender rather than simply 'women.' In her Introduction, di Leonardo argues for a culture and political economy approach, one that respects many histories (rather than one evolution), takes a social constructionist position, and sees gender as embedded in larger institutions and other forms of hierarchy (race, class, ethnicity, and sexuality). Ann Stoler's historical analysis of the policing of sexuality in Indonesia and colonial Asia (Stoler, 2002), Patricia Zavella's comparative research on Hispana/Chicana women factory workers in California and New Mexico (1987; Lamphere et al., 1997) and Irene Silverblatt's work on women and the state in the Inca period (1987) are excellent examples of this approach to gender.

Discovering a Legacy

Simultaneous with the turn to gender and the importance of history, feminists began researching the lives of early female anthropologists and discovered a long lineage of women scholars and feminists. They reexamined the lives of well-known anthropologists like Ruth Benedict, Margaret Mead, and Elsie Clews Parsons and discovered the role of Franz Boas and the Columbia Anthropology Department in producing 20 women PhDs between 1920 and 1940. They recovered Nancy Lurie's history of the nineteenth-century Women's Anthropology Association outlining the beginnings of a female presence in the emerging discipline (1966). Several graduate students spearheaded the publication of *Women Anthropologists: A Biographical Dictionary* (Gacs et al., 1988), and *Hidden Scholars* (Parezo, 1993) examined the impact of women's research on the study of Native Americans in the American Southwest.

Adding to these sources that emphasize white women has been the new scholarship on minority women, many of whom did not have PhDs or spent their careers in minority institutions or outside the academy. One now views Zora Neale Hurston, Katherine Dunham, Ella Deloria, Christine Quintasket (Morning Star), Catherine Bond Day, Irene Diggs, Manet Fowler, and Vera Green as pioneering women in anthropology (Finn, 1995; Harrison and Harrison, 1999).

Full biographies have been written on a number of women anthropologists: Alice Fletcher, Elsie Clews Parsons, Ruth Benedict, Margaret Mead, Ruth Landes, Cora Dubois, and Zora Neal Hurston. These materials have made it possible for feminists to bring these women in from the margins and reevaluate their contributions (Lamphere, 2004). For example, Elsie Clews Parsons, Gladys Reichard, and Zora Neale Hurston can now be appreciated for their more dialogical writing that places the anthropologist within the text and creates more open, polyphonic descriptions of concrete interactions and conversations. Each used these textual strategies long before the appearance of the Clifford and Marcus collection (1986), often seen as the original call for a more dialogical approach to ethnography.

Diversifying Feminist Anthropology

The 1980s and 1990s brought many more minority and lesbian women into graduate programs anthropology and those who

obtained PhDs in the late 1970s began holding faculty positions. This laid the groundwork for a more diverse feminist anthropology. Johnnetta Cole's collection *All American Women* (1986) featured feminist writing and the biographic sketches of minority and working class women, as well as white women. African American feminists analyzed the impact of globalization and neoliberalism on women's lives in the Caribbean (Bolles, 1996; Harrison, 1997), collected the narrative histories of women (McClaurin, 1996), documented the impact of poverty (Mullings and Wali, 2000), and explored gender relationships in African cultures (Shaw, 1995). The forging of Black feminist anthropology is well documented in *Black Feminist Anthropology* (McClaurin, 2001). Among the Chicano/Latina woman who have studied women's lives are Patricia Zavella and Ruth Behar. Gina Perez, Monica Russel y Rodriguez, and Anna Ochoa O'Leary are among those who have made contributions during the first decade of this century. Bea Medicine's innovative scholarship on Lakota women warriors and Nancy Mithlo's research on Native American women artists are examples of contributions by Native Americans.

Another aspect of diversity has been the flowering of gay and lesbian anthropology. In the 1970s, this was initiated by Esther Newton's research on female impersonators, Ellen Lewin's (1993) study of *Lesbian Mothers*, and Elizabeth Kennedy and Madeline Davis's oral histories of lesbian women in Buffalo, New York. Ellen Lewin and William Leap have published three collections on the impact of gay/lesbian anthropology. *Out in the Field* narrates how gay and lesbian anthropologists negotiated their sexual identities during field research; *Out in Culture* emphasizes gay/lesbian theorizing; and *Out in Public* takes a reflexive public anthropology stance by exploring the ways both anthropologists and the gay/lesbian populations they study have engaged the public sphere and pushed for change.

Forging a New Theoretical Language

The important contributions of minority and lesbian feminists in the 1980s and 1990s cleared the way for a continued emphasis on what came to be called 'intersectionality,' the multiple interpenetration of gender, race, class, nationality, and other forms of difference. The attention to gender in relation to class and especially race owes much to the critique of white feminism by minority and Third World feminists often labeled as part of 'third-wave feminism.' The authors mentioned in the previous section are very much part of this same moment in theorizing gender.

The focus on intersectionality is also evident in studies of working class women, poverty and welfare, many influenced by Marxism and using a political economy approach. The collection *Women, and the Politics of Empowerment* (Bookman and Morgen, 1988) brings together feminist analysis of women's resistance, organizing and political consciousness in workplace, educational and community contexts across working class and minority communities. More extensive studies include Karen Sack's analysis of women hospital workers in *Caring by the Hour*, Ida Susser's focus on women in her New York ethnography *Norman Street*, and Mary Anglin's study of women's resistance in the Mica industry in *Women, Power and Dissent in the Hills of Carolina*. Two very excellent collections on women and welfare reform are edited by Catherine Kingfisher (*Western Welfare in Decline*) and by Francis Fox Piven, Joan Acker, Margaret Hollack, and Sandra Morgen (*Work, Welfare and Politics*).

In contrast to the political economy approach, postmodern feminists have followed the theories of Lyotard and Derrida who critique 'modernism' and its commitment of science, objectivity, and truth. However, many feminist anthropologists have been wary of a postmodern stance that dissolves the speaking subject and focuses primarily on discourse rather than on material reality (Mascia-Lees et al., 1989). Anthropologists remain committed to understanding the 'empirical reality' of women's lives and the analysis of social, political, and cultural structures. However, postmodernism has led feminist anthropologists to be conscious of the colonial heritage of their discipline, of their positionality as middle class observers and (often) white Western women. Many feminists have become advocates of critique, deconstructing our concepts as historically situated, culturally bound, and open to revision.

Several other trends emerged in the late 1980s, influenced by Marx, Foucault, Bourdieu, and Butler. First, Ortner (1989) drawing on the Marxist definition of 'praxis' and Bourdieu's concept of 'habitus' argued for an emphasis on practice or concrete actions emerging from structure that could also transform it (Ortner, 1989: 12). Second, anthropological feminists have used Foucaultian notions power as productive and capillary. Power, through structures like the panopticon, shapes the self and inscribes habits of self-policing and states of subjugation forming 'docile bodies' (Foucault, 1977). Aihwa Ong's study of Malaysian women factory workers (*Spirits of Resistance and Capitalist Discipline*), Ann Stoler's research on sexuality in Southeast Asia (*Carnal Knowledge and Imperial Power*), and Saba Mahmood's book on the Egyptian women's Mosque movement (*The Politics of Piety*) are some of the most important examples of this approach.

A third trend has been research on the body and embodiment. Studies by historians and feminist biologists have pointed to the Western bias in viewing the body and sexuality as biologically based, historically static, and as only situated in a two-sex, two-gendered fixed system. Rather than the body being 'presocial' and a blank slate, feminists using Foucault and Butler view sex and gender as malleable, culturally structured, and performative. Mascia-Lees and Patricia Sharpe's collection of essays exemplify this foray into analyzing the body in literature, painting, and contemporary US culture (2000).

Writing New Feminist Ethnographies

A central impact of postmodernist and Marxist theory has been the emphasis on rethinking the power dynamics embedded in feminists' relationship to their subjects. This began by interrogating the boundary any observer must confront – that between the Self and the Other, that is, between the investigator and those with whom she interacts. Feminist ethnographers questioned whether anthropology could be constructed differently when feminists study women. Stacey is deeply skeptical, stressing the class divisions and manipulative aspects of the feminist/subject relationship (1988). Marilyn Strathern

argues for an ambivalent relationship between feminism and anthropology since each mocks the way the contrasting discipline conceives of the Self/Other dichotomy (1987). However, Abu-Lughod is much more hopeful, suggesting that feminist ethnography can disrupt boundaries and replace the presumption of a universal female experience with a grounded sense of our commonalities and differences (1990: 27).

Other feminist anthropologists like Kirin Narayan and Patricia Zavella have explored the complexity of being partial 'insiders' (observers with kin or ethnic connections to those under study) yet being viewed as 'outsiders' because of national, class, and cultural differences. These essays recognize what Donna Haraway calls "situated knowledges" a feminist epistemological stance that acknowledges the social location of the researcher where knowledge is always partial and seen through the perspectives of active subjects (Haraway, 1991: 183–201).

This reevaluation has also led to different strategies for ethnographic writing. These include presenting women's voices in detail, including the author in narratives, and historically contextualizing ethnographic material. These writing strategies, also pioneered by Parsons, Reichard, and Hurston, move anthropology toward creating more dialogic rather than objectifying accounts. Lila Abu-Lughod's ethnography of Bedouin women's lives, *Writing Women's Worlds*; Ruth Behar's life history, *Translated Woman*; and Margery Wolf's *A Thrice-Told Tale* are examples of this growing trend.

The collection *Women Writing Culture* (Behar and Gordon, 1995) is in direct contrast to *Writing Culture* (Clifford and Marcus, 1986) that ignored feminism and women's writing. Several articles explore the writing of earlier women anthropologists, while other contributions attest to the women's past marginalization. There are also innovative examples of fiction, play writing, diaries, and narratives that suggest that feminist ethnography has taken new form and will reach broader audiences.

Engendering the Past through Feminist Archaeology

Feminist approaches to prehistory emerged more slowly than in social/cultural anthropology, but as they have grown, they have begun to transform archaeological theory and practice. *Engendering Archaeology: Women and Prehistory* (Gero and Conkey, 1991) is one of the first comprehensive responses foregrounding the analysis of gender. Contributors examine gender in a variety of archaeological contexts (e.g., pottery production, weaving, cooking, lithics, and art). The intention is to examine gender in relationship to power structures, household organization, the deployment of labor, and the organization of space.

Interest in gender continued to expand, but many archaeologists have been content with just "finding women in the past." In contrast, feminists examine gender relations and aim to change archaeological theory and practice. This means increasing the visibility of human agency, admitting the partial and situated nature of knowledge, and allowing for multiple interpretations rather than top-down, hierarchically positioned ones (Conkey and Gero, 1997: 429).

By the mid-1990s, there were a number of good examples of this new feminist research. Spector (1993) supplements her archaeological analysis of women's activities in a Wahpeton Village by using Dakota categories for found objects and engaging indigenous consultants. She also constructs an imaginative narrative of a how a woman's incised bone awl might have been made, used, and then lost. Elizabeth Brumfiel makes gender, class, and faction visible in her analysis of Azec figurines and tribute cloth (made by women weavers) (Brumfiel, 1996). Rosemary Joyce draws on Judith Butler's analysis of embodiment and subjectification to examine Mayan construction of gender as well as political hierarchy (Joyce, 2002), while Meg Conkey draws on feminist and indigenous archaeologies to explore domination in the archaeological record (Conkey, 2005).

The collection edited by Pamela Geller and Stockett (2007) contains articles that also take an intersectional approach by focusing on identity and difference, two themes that Geller emphasizes in her review (2009). A special issue of the *Journal of Archaeological Method and Theory* (Conkey and Wylie, 2007) brings together feminist thinking on archaeological fieldwork, practice, and theory. Articles also probe the 'backlash' by scholars of gender who wish to distance themselves from feminist approaches (Sorensen, 2000) and the continued reluctance by many (often male) archaeologists to acknowledge the contributions of feminists.

Emerging Issues in the Twenty-First Century

Since the 1990s, feminist anthropologists have become increasingly engaged with the study of the social issues that impact women in a globalizing world. Financial capital has become more mobile, production processes have moved to the Global South, and labor migration has increased. Worldwide computer networks and satellite connections as well as breakthroughs in biotechnology have transformed everything from television, film, and music to medicine, reproduction, and sexual norms. For anthropological feminists, the emerging issues of labor migration and women's work roles, reproduction and sexuality, gender violence, and human rights are all viewed through a global lens.

Gunawardena and Kingsolver's collection (2007) provides some important examples of how women's work is changing. From young factory workers in Sri Lanka to female traders in Ghana, domestic workers in Italy, and hotel workers in San Francisco, women are finding it difficult to make a living, but nevertheless forging survival strategies and contesting their situations. Migration is also gendered. Deborah Boehm shows how young Mexican males migrate to become adult providers while girl friends and wives remain to raise children or come to cook and keep house for a host of male relatives (2012) while Patricia Zavella's transnational study (2012) focuses on how Mexican migrant women and men struggle with displacement, instability, gendered racialization, and discrimination at work. Anthropologists are also studying the children of immigrants and their adaptation to the United States as well as women's role in helping their families hold on to language, customs, and identities (Coe et al., 2011).

International adoptions and the accompanying issues of culture and identity have been studied (Yngvesson, 2010; Howell, 2006). And as there are more immigrants from Middle Eastern cultures and Islamic states to European

Countries and to the United States, the maintenance of language differences, the veiling of women, and the practice of Islam have all become points of controversy (Abu-Lughod, 2008; Scott, 2007; Mahmood, 2005).

Feminist analysis of the body has also spawned an important literature within medical anthropology beginning with Emily Martin's *The Woman in the Body* (1987). *Conceiving the New World Order* (Ginsburg and Rapp, 1995) contains much of the new work on stratified reproduction across a wide range of countries from issues of abortion, assisted reproduction, and state controls on fertility to lesbian motherhood, single headed households, teenage mothers, and child care workers. Other book-length studies include Rapp's study of amniocentesis in New York, *Testing Women, Testing the Fetus*; Ragone's book *Surrogate Motherhood: Conception in the Heart*; Franklin's study on assisted reproduction, *Embodied Progress*; and Inhorn's ethnography of infertility in Egypt, *Local Babies, Global Science*. In the next decade, feminist anthropologists expanded their research on the new reproductive technologies and surrogacy to developing countries where their use and meaning are often very different (Rapp, 2011).

The body is also central to feminist and queer analysis of sexuality and sexual orientation an area that has grown and blossomed in the last two decades. Following on the ethnographies of gay and lesbian communities in the 1990s, more recent books, like those by Ellen Lewin, have focused issues of gay marriage and gay fatherhood. Others have continued the early work on gay and lesbian sexuality pioneered by Roger Lancaster in *Life Is Hard: Machismo, Danger and the Intimacy of Power in Nicaragua* and Evelyn Blackwood and Saskia Wieringa's collection, *Female Desire: Same-Sex Relations and Transgender Practices across Cultures*. Cymene Howe examines the gay and feminist movements in post-Sandanista Nicaragua (*Mediating Sexuality*) and Martin Manalansan portrays the lives of Filipino gay immigrants in New York (*Global Divas*). David Valentine examines the elastic term 'transgender,' (along with 'gay,' 'fem-queen,' or 'girl') in his book *Imagining Transgender* about the lives of those male-to-females in New York who see it as part of their identity, while Gayle Rubin's essays on sexuality (*Deviations*) and Margot Weiss' book on bondage (*Techniques of Pleasure*) explore bondage, discipline, sadism and masochism (BDSM) and alternative sexualities, theorizing about gender, hierarchy, dominance, and pleasure.

Anthropological studies of masculinity flourished in the 1990s with the work of Matthew Gutmann (*The Meanings of Macho*), David Gillmore (*Manhood in the Making*), and others. More recent research ranges from ethnographic analysis of male struggles with poverty, drugs, and race (Bourgois and Schonberg, *Righteous Dope Fiend*) to masculine images of colonial terror such as the pishtaco in Peru (Weismantel, *Cholos and Pishtacos*) to shifting definitions of masculinity in the Middle East that have emerged through the impact of new reproductive technologies (Inhorn, *The New Arab Man*). Mascia-Lees (2011) brings together all the major authors who have written on the diverse aspects of the body in a wide-ranging collection that sums up various approaches to embodiment from sexuality studies, to medical anthropology, to those interested in colonialism, racialization, and border politics.

The focus on globalization has encouraged the analysis of economic policy and institutions both at the high and low ends of the income spectrum. Feminists have written on neoliberalism and its impact on gender in Nicaragua (Florence Babb, *After Revolution: Mapping Gender and Cultural Politics in Neoliberal Nicaragua*), on the practices of governing and citizenship (Aihwa Ong, *Neoliberalism as Exception*), and on the creation of 'desiring subjects' with the appropriate needs, aspirations, and desires in China (Lisa Rofel, *Desiring China*). In the United States, financial crisis and income disparity have encouraged women anthropologists to write about the banking industry, for example, Karen Ho's book, *Liquidated*, and Gillian Tett's riveting account of the 2008 financial crisis, *Fool's Gold*. Melissa Fisher has written *Wall Street Women* and while Sandra Morgen and Mimi Abromowitz have argued that *Taxes Are a Women's Issue* in their 2005 study.

Finally, there has been continued research on gender violence. Sally Merry's *Gender Violence: A Cultural Perspective* (2009) uses personal stories and case studies to provide an overview of the forms of gender violence globally and efforts to diminish it. Hillary Haldane and Jennifer Weiss have headed a group of young women scholars conducting research in a variety of venues across a variety of countries (Wies and Haldane, 2011).

Feminist anthropology is now global. Faye Harrison (2009) outlines some of the differences between African feminisms and those of the West, and describes the impulse to decolonize Western feminisms though forging cooperation and solidarity across cultural, national, and hemispheric lines of difference, a position also supported by Janet Mancini Billson and Carolyn Fleuhr-Lobban (2005). Harrison shows how organizations like the International Union for Anthropological and Ethnographic Sciences can offer spaces for exchanges and collaboration and the building of a broader and more diverse set of feminisms.

Collaboration and Activism

In the second decade of the twenty-first century, many feminist anthropologists feel the feminist movement has faded. Some theorists argue that this is a postfeminist moment. However, there are still protests against rape (the 'slut walks' of 2012), against the narrowing of the right to abortion (protests against defunding of Planned Parenthood in 2012), and struggles concerning access to birth control. Internationally, protests against the treatment of women and girls have been vigorous in India, Egypt, and Afghanistan.

Younger scholars may be reluctant to call themselves 'feminists' but there are many who would identify as queer theorists, those who study women and development, or those who examine gender. Anthropologists have been among the leaders in nongovernmental organizations dedicated to policy-oriented research. Linda Basch, who was director of the National Council for Research on Women between 1996 and 2012, headed an organization that spearheaded research projects on women and girls in science, engineering, and technology; taxes as a woman's issue; retaining women of color in leadership positions; and the importance of publically funded child care. The Institute for Women's Policy Research, where

Jane Henrici, Study Director and President of the Association for Feminist Anthropology (2011–13) is employed, is a public policy 'think tank' that provides rigorous research and public dialogue on the needs of women. A collection edited by Craven and Davis *Feminist Activist Ethnography* (2013) outlines the issues that engage women activists (reproductive rights, food activism, gay marriage, and breast cancer). Feminist anthropologists also describe the organizing strategies their subjects embrace in order to improve women's lives. This form of engaged or public anthropology is increasingly adopted by feminists and is part of a much larger trend within the entire discipline.

Feminist anthropology in the United States has been transformed since its initial expansion in the 1970s. There is no longer a preoccupation with universals or assessing women's status, perhaps mirroring the discipline's move away from grand theories. Instead, there are careful historically contextualized studies that grapple with changes brought by increased globalization and transnational movements of capital and labor. Gender and its intersections with race, class, sexuality, and citizenship provide the lens through which feminists examine and engage with the past and with the issues confronting women both in the United States and across the world. Feminists are more frequently tackling issues that have public policy implications and are more committed to collaborating with and giving voice to individual women, women's movements, and women's issues areas of economy, reproduction, and sexuality. These studies are making an impact in many parts of anthropology including medical anthropology, studies of immigration, urban anthropology, the anthropology of work, and the practice of archaeology. Feminist anthropology is moving in the direction of engendering the discipline.

See also: Embodiment and Culture; Feminisms and Acculturation around the Globe; Feminist Criminology and Gender Studies; Feminist Political Theory and Political Science; Gender History; Gender and Feminist Studies in Geography; Gender and Place; Gender, Feminism, and Sexuality in Archaeological Studies; Postmodernism in Sociology; Queer Theory and Intersectionality.

Bibliography

Abu-Lughod, L., 1990. Can there be a feminist ethnography? Women and performance. A Journal of Feminist Theory 5 (1), 7–27.

Abu-Lughod, L., 2008. Do muslim women really need saving: reflections on cultural relativism and its others. American Anthropologist 104 (3), 893–990.

Behar, R., Gordon, D. (Eds.), 1995. Women Writing Culture. University of California Press, Berkeley and Los Angeles.

Boehm, D., 2012. Intimate Migrations: Gender, Family, and Illegality Among Transnational Mexicans. New York University Press, New York.

Coe, C., Reynolds, R.R., Boehm, D.A., Hess, J.M., Rae-Espinosa, H., 2011. Everyday Ruptures: Children, Youth, and Migration in a Global Perspective. Vanderbilt University Press, Nashville.

Bolles, L.A., 1996. Sister Jamaica: A Study of Women, Work and Household in Kingston. University Press of America, Lanham, MD.

Bookman, A., Morgen, S., 1988. Women and the Politics of Empowerment. Temple University Press, Philadelphia.

Brumfiel, E., 1996. Figurines and the Aztec state: testing the effectiveness of ideological domination. In: Wright, R. (Ed.), Gender and Archaeology. University of Pennsylvania Press, Philadelphia, pp. 143–166.

Clifford, J., Marcus, G. (Eds.), 1986. Writing Culture. University of California Press, Berkeley and Los Angeles.

Cole, J., 1986. All American Women: Ties that Divide, Ties that Bind. Free Press, New York and London.

Conkey, M.W., 2005. Dwelling at the margins, action at the intersection? Feminist and indigenous archaeologies. Journal of the World Archaeological Congress 1 (1), 9–59.

Conkey, M.W., Gero, J.M., 1997. Programme to practice: gender and feminism in archaeology. Annual Review of Anthropology 26, 411–437.

Conkey, M.W., Wylie, A. (Eds.), 2007. Doing Archaeology as a Feminist. Journal of Archaeological Method and Theory (special issue).

Craven, C., Davis, D.-A. (Eds.), 2013. Feminist Activist Ethnography: Counterpoints to Neoliberalism in North America. Lexington Press, Lanham.

Finn, J.L., 1995. Ella Cara Deloria and Mourning Dove: writing for cultures, writing against the grain. In: Behar, R., Gordon, D. (Eds.), Women Writing Culture. University of California Press, Berkeley and Los Angeles, pp. 131–147.

Foucault, M., 1977. Discipline and Punish: The Birth of the Prison.

Gacs, U., Kahn, A., McIntyre, J., Wienberg, R. (Eds.), 1988. Women Anthropologists: A Biographical Dictionary. Greenwood Press, Westport.

Gailey, C., 1987. Kinship to Kingship: Gender Hierarchy and State Formation in the Tongan Islands. University of Texas Press, Austin.

Geller, P.L., 2009. Identity and difference: complicating gender in archaeology. Annual Review of Archaeology 38, 65–81.

Geller, P.L., Stockett, M.K. (Eds.), 2007. Feminist Anthropology: Past, Present, and Future. University of Pennsylvania Press, Philadelphia.

Gero, J.M., Conkey, M.W. (Eds.), 1991. Engendering Archaeology: Women and Prehistory. Basil Blackwell, Cambridge, MA and Oxford, England.

Ginsburg, F., Rapp, R. (Eds.), 1995. Conceiving the New World Order: The Global Stratification of Reproduction. University of California Press, Berkeley and Los Angeles.

Gunewardena, N., Kingsolver, A., 2007. The Gender of Globalization: Women Navigating Cultural and Economic Marginalization. School of Advanced Research Press, Santa Fe.

Haraway, D., 1991. Simians, Cyborgs, and Women: The Reinvention of Nature. Routledge, New York and London.

Harrison, F., 1997. The gendered politics and violence of structural violence: a view from Jamaica. In: Lamphere, L., Ragone, H., Zavella, P. (Eds.), Situated Lives: Gender and Culture in Everyday Life. Routledge, New York, pp. 451–568.

Harrison, F., 2009. Reconciling perspectives on the world's diverse women and cultures of gender: toward new syntheses for the 21st century. In: Nas, P.J.M., Jijiao, Z. (Eds.), Anthropology Now: Essays by the Scientific Commissions of the IUAES & History of the IUAES. Intellectual Property Publishing House, Beijing, pp. 103–116.

Harrison, F.V., Harrison, I.E., 1999. African-American Pioneers in Anthropology. University of Chicago Press, Chicago.

Howell, S., 2006. The Kinning of Foreigners: Transnational Adoption in a Global Perspective. Berghahn, New York.

Joyce, R., 2002. Gender and Power in Prehispanic Mesoamerica. University of Texas Press, Austin.

Lamphere, L., 2004. Unofficial histories: a vision of anthropology from the margins. American Anthropologist 106 (1), 126–139.

Lamphere, L., Zavella, P., Gonzales, F., Evans, P., 1997. Sunbelt Working Mothers: Reconciling Factory and Family, Cornell University Press, Ithaca.

Lamphere, L., Rapp, R., Rubin, G., 3 September 2007. "Anthropologists are talking" about feminist anthropology. Ethnos 72, 408–426.

Leacock, E.B., 1980. Myths of Male Dominance: Collected Articles on Women Cross-Culturally. Monthly Review Press, New York.

Lewin, E., 1993. Lesbian Mothers: Accounts of Gender in American Culture. Cornell University Press, New York.

Lurie, N.O., 1966. Article on women's anthropology association. In: Helm, J. (Ed.), Pioneers of American Anthropology. University of Washington Press, Seattle.

Mahmood, S., 2005. The Politics of Piety: The Islamic Revival and the Feminist Subject. Princeton University Press, Princeton.

Martin, E., 1987. The Woman in the Body: A Cultural Analysis of Reproduction. Beacon Press, Boston.

Mascia-Lees, F.E., 2011. A Companion to the Anthropology of the Body and Embodiment, Wiley Blackwell, New York and London.

Mascia-Lees, F.E., Sharpe, P., 2000. Taking a Stand in a Postfeminist World: Toward an Engaged Cultural Criticism. SUNY Press, New York.

Mascia-Lees, F.E., Sharpe, P., Cohen, C.B., 1989. The postmodernist turn in anthropology: cautions from a feminist perspective. Signs: Journal of Women, Culture, and Society 15 (1), 7–33.

McClaurin, I., 1996. Women of Belize: Gender and Change in Central America. Rutgers University Press, New Brunswick.

McClaurin, I. (Ed.), 2001. Black Feminist Anthropology. Rutgers University Press, New Brunswick.
Merry, S., 2009. Gender Violence: A Cultural Perspective. Wiley-Blackwell, West Sussex, UK.
di Leonardo, M., 1991. Gender at the Crossroads of Knowledge: Feminist Anthropology in the Postmodern Era. University of California Press, Berkeley and Los Angeles.
Mullings, L., Wali, A., 2000. Stress and Resilience: The Social Context of Reproduction in Central Harlem. Plenum Publishers, New York.
Ortner, S.B., 1989. High Religion: A Cultural and Political History of Sherpa Buddhism. Princeton University Press, Princeton.
Ortner, S.B., Whitehead, H. (Eds.), 1981. Sexual Meanings: The Cultural Construction of Gender and Sexuality. Cambridge University Press, Cambridge.
Parezo, N., 1993. Hidden Scholars: Women Anthropologists in the Native American Southwest. University of New Mexico Press, Albuquerque.
Rapp, R., 2011. Reproductive entanglements: body, state, and culture in the dys/regulation of childbearing. Social Research 78 (3), 693–718.
Reiter, R.R. (Ed.), 1975. Toward an Anthropology of Women. Monthly Review Press, New York.
Rosaldo, M.Z., 1980. The use and abuse of anthropology: reflections on feminism and cross-cultural understanding. Signs: A Journal of Woman, Culture and Society 5 (3), 389–417.
Rosaldo, M.Z., Lamphere, L., 1974. Woman, Culture and Society. Stanford University Press, Stanford.
Sacks, K., 1979. Sisters and Wives: The Past and the Future of Sexual Equality. Greenwood Press, Westport.
Schlegel, A., 1972. Male Dominance and Female Authority: Domestic Authority in Matrilineal Societies. Human Relations Area Files Press, New Haven.
Scott, J., 2007. Politics of the Veil. Princeton University Press, Princeton.
Shaw, C.M., 1995. Race, Sex, and Class in Kenya. University of Minnesota Press, Minneapolis.
Silverblatt, I., 1987. Sun, Moon, and Witches: Gender Ideologies and Class in Inca and Colonial Peru. Princeton University Press, Princeton.
Sorensen, M.L.S., 2000. Gender Archaeology. Polity Press, Cambridge.
Spector, J., 1993. What This Awl Means: Feminist Archaeology at a Wahpeton Dakota Village. Minnesota Historical Society Press, Saint Paul.
Stacey, J., 1988. Can there be a feminist ethnography? Women's Studies International Forum 11 (1), 21–27.
Stoler, A.L., 2002. Carnal Knowledge and Imperial Power: Race and the Intimate in Colonial Rule. University of California Press, Berkeley and Los Angeles.
Strathern, M., 1987. An awkward relationship: the case of feminism and anthropology. Signs: A Journal of Women, Culture and Society 12 (2), 276–294.
Wies, J.R., Haldane, H.J. (Eds.), 2011. Anthropology at the Front Lines of Gender-Based Violence. Vanderbilt University Press, Nashville.
Yngvesson, B., 2010. Belonging in an Adopted World: Race, Identity and Transnational Adoption. University of Chicago Press, Chicago.
Zavella, P., 1987. Women's Work and Chicano Families: Cannery Workers of the Santa Clara Valley. Cornell University Press, Ithaca.
Zavella, P., 2012. I'm Neither Here Nor There: Mexicans' Quotidian Struggles with Migration and Poverty. Duke University Press, Durham.

Feminisms and Acculturation around the Globe

Akosua Adomako Ampofo, Edwin Asa Adjei, and Maame Kyerewaa Brobbey, University of Ghana, Legon, Ghana

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.C. Lam, volume 15, pp. 10163–10169, © 2001, Elsevier Ltd.

Abstract

Feminist theorizing and praxis has had a bumpy, yet ultimately successful voyage over several 'waves.' These waves have sought to respond to women's experiences and needs in changing social and political circumstances, eventually recognizing that context matters. Feminism has not only struggled to dismantle patriarchy and transform societies, but it has also responded to the unique cultural contexts in which people live. Today, some argue that we live in a postfeminist age in which feminism is no longer relevant. Others argue that this is a short-sighted view born out of a limited appreciation of the unique cultural contexts on which women's experiences are based, and feminist consciousness is born. This article discusses the multicultural and global trajectories and dimensions of feminisms, paying particular attention to non-Western forms that have received less attention.

Introduction

Women's status, social, economic, and political rights have been defined, defended, established, and then contested. Recent debates about Beyonce Knowles' feminist identity, as well as the assertions by powerful pop icons that they are not feminists have led some to argue that in reality we live in a postfeminist age. During a panel discussion at New York's New School (7 May 2014), feminist author and social activist Bell Hooks expressed qualms about the feminist perspective of Beyonce's projects. Particularly concerned about Beyonce's visual representation and her impact on young women, Hooks noted, "I see a part of Beyonce that is, in fact, anti-feminist, that is assaulting, that is a terrorist ..." (The Hufington Post; http://www.huffingtonpost.com/2014/05/09/beyonce-anti-feminist_n_5295891.html). And in a recent interview with Time magazine actress Shailene Woodley was asked if she considered herself a feminist. Ms. Woodley, 22, answered in the negative, explaining, "Because I love men, and I think the idea of 'raise women to power, take the men away from the power' is never going to work out because you need balance." (The New York Times, 21 May 2014; http://www.nytimes.com/2014/05/22/fashion/who-is-a-feminist-now.html). Others have pointed out that the very fact of a debate underscores the salience of feminist discourse and praxis. Today, it is generally accepted that there is not one but a variety of feminisms, each with unique cultural emphases. However, this contemporary taken-for-grantedness of feminism's multicultural nature did not come without contestations. This article traces the paths to contemporary multicultural feminisms across socio-geo-political contexts.

In 1967, the (Chicago) Women's Liberation Union was formed, birthing the women's liberation movement. In the United States, women's liberation gave way to feminism, a term that has been ascribed to the work of French physician Madeline Pelltier best known for her activism related to women's sexual and reproductive rights and the voting right. However, women had fought for equality before the concept was thus named.

Feminist scholarship enjoys rich and diverse theoretical and methodological traditions (Adomako Ampofo et al., 2004; Essof, 2001; LeGates, 2001; Imam et al., 1997; Mbilinyi, 1992; Mohanty, 1986). This brief review emphasizes the impact of its diverse global, especially Third World adherents. In a piece titled "Sarah Baartman, Invisible!" Adesanmi produces a fictitious letter written by Baartman to Sandra Gilbert and Susan Gubar who produced the Norton Anthology of Literature. Adesanmi (via Baartman) laments the 'excision of African theories and theorists' (2011, p. 109) from this 2452-page 2-volume collection. Sarah Baartman was the most famous of at least two Khoikhoi women who were exhibited as curiosities in nineteenth-century Europe under the now discredited name Hottentot Venus. African feminists have used her story to draw attention to the double burdens of sexism and racism endured by black women. The term feminism is used to refer to the notion of an agenda for women's rights, generally, while feminisms refer to the diverse forms that subsequently emerged in response to cross-cultural differences. The sources for this article include scholarly as well as popular sources published in English, and a Facebook© discussion with black feminist scholars in Germany, the United Kingdom, and the United States. (This is credited to Kara Keeling, Maisha Eggers Peggy Piesche, and Wangui Wa Goro.) The rest of the article maps feminism's global and, hence, multicultural trajectory, tracing feminism's waves and signposting contestations and convergences until the current 'postfeminism' era.

Enculturing Feminism

Elements of feminist theorizing and activism around social justice for women can be found in every modern culture. This section discusses the trajectory of feminism and its 'enculturing.' Feminist theorizing and politics emerged as an organized movement in Western societies in the latter half of the nineteenth century. Histories of Western feminism are often discussed in the context of three waves; each wave focusing on issues influenced by progress in women's rights, and the

inadequacies of an earlier wave. Of particular significance are the contestations between Western and Black/Third World feminists, leading to the formation of diverse feminisms privileging difference.

First-Wave Feminism

Western feminism of the nineteenth and early twentieth centuries has been labeled as first-wave feminism, in relation to the subsequent issues and approaches of the 1960s to 1980s. It emerged in the context of industrialization, the growing organization of the working class in Europe and America, and the so-called liberal politics of the time. This period was principally concerned with equal opportunities for women, and was anchored in a universalist paradigm in which women were undifferentiated. The period was characterized by activism and campaigns – including hunger strikes, picketing, conventions, parades, and marches – for women's education, employment, and legal rights; and for reforms in family life and sexual values. Specific issues included the push for equal pay for equal work, women's rights to vote, and rights to family planning services. This period also confronted female stereotypes and became known as the women's liberation movement. In the United States, women's liberation struggles begun as early as in the 1840s, and were linked with other movements, such as the abolitionist movement. However, the movement was generally by, and for white, educated, middle-class women about their social and political discomforts. For example, socialist feminists constructed marriage and the home as a site of oppression for women, a major point of later disagreement with Third World feminists. Betty Friedan's (1963) 'The Feminine Mystique' for instance, examines the conditions of middle-class American women isolated and excluded from social and productive life. Even so, the nineteenth century struggles were supported by Black women abolitionists, such as Maria Stewart, Frances E.W. Harper, Ida B. Wells, and Sojourner Truth who, with her famous "Ain't I a woman" call agitated for the rights of women of color (Krolokke and Sorensen, 2005). Significant during this wave of feminism was the suffragist movement, which was combined with the struggle to legalize contraception (Antrobus, 2004). Questions of the connections between the institutions of private property, control over assets, and female–male relations within families and society were founded within the social-democratic and communist debates of the late nineteenth and early twentieth centuries, particularly in Eastern Europe. A socialist/Marxist feminism also developed in workers' unions in the United States and social-democratic parties in Europe. Socialist feminism insists on locating feminism within wider movements for economic transformation and social justice. It is this view that paved the way for second-wave feminism.

UN Feminism

A brief detour between the first and second waves to reflect on the role of the United Nations (UN) in the feminist movement is in order here. Technically, UN feminism spans the first and second waves, even stretching beyond the boundaries of second-wave feminism in some cases. The United Nation's role in globalizing feminism, and thus catalyzing multicultural feminisms is undeniable. Meetings held in Mexico City in 1975 marking International Women's Year, and the beginning of the UN Decade on Women, eventually internationalized women's issues. While government representatives of member states attended the official UN meetings, women's groups attended parallel nongovernmental forums. With these NGO meetings began the tradition of measuring states' commitments to, and implementation of, gender equality policies. Follow-up meetings were held in Copenhagen in (1980), Nairobi in (1985), and Beijing in (1995) and women's issues and feminism gained global momentum. This helped women coalesce around common concerns such as discrimination generally, and gender-based violence more specifically. It also highlighted the specific experiences and concerns of women around the world, frequently locating these in discussions of development and underdevelopment, and the complicity of women from the industrialized (Western) world in the exploitation of women in the Third World. Third World feminists objected to underdevelopment being conceptualized as a Third World problem, and this led to a conference and the subsequent well-cited 1991 volume edited by Indian feminist Chandra Mohanty et al. titled, "Third World Women and the Politics of Feminism". Feminist development theorists note that the interaction between feminism and development has taken five main forms: Women in Development (WID); Women and Development (WAD); Gender and Development (GAD); Women, Environment, and Development (WED); and Postmodernism and Development. The WID approach, drawing on modernization theory, advocated for the integration of women into existing development projects and also setting up women-only projects that addressed everyday needs (Kabeer, 1994; Tinker, 1990; Rathgeber, 1990). In many African countries, this agenda was taken over by first ladies and has been described as femocracy (Mensah-Kutin et al., 2000).

The WAD perspective argued that it was precisely women's link with modernization that had impoverished them. Thus, WAD drew much more from dependency and neo-Marxist approaches to questions of underdevelopment. Concerns with the origins of patriarchy, the intensification of patriarchy with the spread of capitalism, and Engel's analysis of the rise of private property and the agricultural revolution formed the background to this approach, which was particularly favored by Third World feminists who supported the social role of the postcolonial state (Kabeer, 1994; Rathgeber, 1990).

In the GAD approach gender (power) relations became the main analytical category. This perspective looked at the structures and processes giving rise to women's disadvantaged position, including the ideology of male superiority. The assumptions in WID and WAD of women as a homogeneous group were critiqued and GAD argued in favor of intersectionality, recognizing that women are differentiated by class, race, religion, and other social characteristics.

The WED perspective was first introduced in the 1975 Mexico-City UN Conference by Vandana Shiva, and was founded on the notion of sustainable development, linking notions of equity between generations, the balance between economic and environmental needs to conserve nonrenewable resources, and the idea of reducing industrial waste and pollution. Sustainable development challenged the notion that development automatically translates into economic growth.

Offshoots of WED are Ecofeminism and, later, feminist political ecology. The late Kenyan Nobel Peace laureate Wanguri Mathai was a proponent of Ecofeminism, and she built the Green Belt movement in Kenya to support women and the environment.

The United Nations and the international discussions it facilitated have shaped the progression of feminist agendas and actions, as well as the recognition that women are not a monolithic category. The United Nations, perhaps through the contestations that erupted during the parallel NGO meetings of the international conferences, and the women's bureaux that were set up in different countries, may also be credited with catalyzing feminism's enculturation. A transparent and collective approach that accepts and understands differences, the power of discourse, and the importance of consultation so that women can articulate their own needs and agendas is favored.

Second-Wave Feminism

Second-wave feminism which began in the early 1960s, and is generally recognized as lasting into the late 1980s, was a response to the perceived lack of politics of the first wave. This wave was influenced by student protests; gay and lesbian movements; the anti-Vietnam war and civil rights and Black power movements in the United States, linking activism with theory-building (Antrobus, 2004; Krolokke and Sorensen, 2005). Theorizations of the subjugation of women were linked to broader critiques of patriarchy, capitalism, imperialism, normative heterosexuality, and women's gender roles; scholars distinguished biological sex from gender.

The second wave is responsible for the concept of 'women's empowerment' that remains influential, and for slogans such as 'sisterhood is powerful' and 'the personal is political,' as well as consciousness-raising efforts (Krolokke and Sorensen, 2005; Ali et al., 2000). The period was linked with, and provided the foundation for, the establishment of formal sites for feminist scholarship around the world – women's gender, and queer units in universities (Krolokke and Sorensen, 2005). The University of Maryland, Baltimore, lists more than 900 women's/gender/feminist study programs, departments, and research centers around the world with web sites. Many more will not have active web sites so this is a very conservative count. While the majority are to be found in the United States, in many places in the global south women's or feminist studies may be subsumed under other programs or within other disciplines and, thus, not be as easy to track via a cursory web search.

Confronted with 'cultural difference' – such as women wearing the (Muslim) veil – many Western feminists perceived themselves as rescuing Third World women from their cultures. This approach exposed tensions between Western and non-Western feminisms. Third World feminists, such as Gayatri Spivak (1988) criticized Western feminists for speaking on behalf of Third World women. Some, like Mohanty (1986), even opined that the glossing over of the diverse experiences, cultures, and histories among Third World women was a political project to subjugate Third World women. They further criticized Western feminists for speaking naively on behalf of Third World women. Thus, a fundamental shift in second wave feminism was the insistence on the importance of differences among women. Questions of the body, sexuality, female genital mutilation, the family, and motherhood emerged as key points of contention between Western and non-Western feminists. Radical Western feminists argued that women (should) have rights over their bodies, advocating a sexuality disconnected from the obligations of marriage and motherhood. Third World feminists disagreed and nowhere did this present itself more sharply than during the International Conference on Population and Development, ICPD held in Cairo in 1990 where many African representatives insisted, without apology, that abortion was not a key concern, and that motherhood remained central to their identities especially as it linked them to, or provided respite from, lineages as the case might be. Motherhood also provided status, and emotional and economic support. Third World feminists also insisted that the (over) emphasis on sexuality was misplaced and that they had the right to privilege their own concerns. African women in particular were very unhappy with a focus on female genital mutilation to the near-exclusion of other issues and without any attempt to understand the cultures in which the practice was located. The delicate home of global sisterhood based on a notion that feminism could create an alliance based on the commonality of womanhood and the patriarchal societies they lived in crumbled. At one meeting organized by the Nigerian feminist Nnaemeka Obioma in Nsukka in as late as 1992, African Americans and some African participants even objected to the participation of white women. The factions that emerged necessitated a questioning of the comfortable notion of global sisterhood (Nnaemeka, 1998). It saw the beginning of 'standpoint feminism' (Chodorow, 1978) and identity politics. Women of color, working-class, Third World, and lesbian feminists questioned the predominantly White, middle-class, and heterosexual feminist agenda of liberal, socialist, and radical feminists, accusing them of "ignoring, trivializing, or distorting the lives of women who were 'different' through other forms of othering" (Friedman, 1995, p. 9). The concept woman could no longer "bear the weight of all contents and meanings ascribed to it" (Gillis et al., 2004, p. 1). These accusations caused essential reconceptualizations in the feminist theory.

Third-Wave Feminism

Third-wave feminism emerged in the mid-1990s, the period of many of the UN World Conferences on Women and the signing of international agreements such as the Convention on the Elimination of all Forms of Discrimination against Women (CEDAW). These meetings meant that feminists from all over the world had begun to interact. It was also the beginning of an ICT society, meaning women could now meet virtually as well. Third-wave feminism finally addressed difference, speaks of feminisms, and makes room for contradictions. Terms such as the body, gender, sexuality, and 'heteronormativity' were deconstructed. This wave concerned itself with interrelated issues including environmentalism; human rights; domestic violence; women in politics; eating disorders, body image, and reconstruction; hegemonic knowledge production; globalization; and the effects of racism and classism. Third-wave feminists advocate self-location and positionality and an

antihegemonic global feminism. An important contribution to Third-wave feminist thinking is the notion of 'transversal politics' launched by Yuval-Davis (1997), a British Jew. Inspired by Spivak's theory of 'strategic essentialism' and Collins's theory of the 'partiality of standpoints' and 'situated knowledge' Yuval-Davis presents possibilities of dialogue between women across national, ethnic, and religious boundaries. Yuval-Davis was also inspired by the politics of feminist activist groups such as the London-based Women Against Fundamentalism (WAF) – which includes Christians, Jews, Muslims, Sikhs, Hindus – and others such as the Bologna feminists, who work with women from groups in conflict, such as Serbs and Croats or Palestinian and Israeli Jewish women who work across typical identity schisms. Transversal politics recognizes differences in nationality, ethnicity, or religion – and hence in agenda – but also requires a commitment to listen and participate in dialogue.

Black and Third World Feminisms

The distinguishing lines are very blurred, however, Black feminism typically focuses on the Diaspora experience. North American Black Feminism has its roots in nineteenth century abolition movements articulating Black resistance and advocacy for social, political, and economic equality (Krolokke and Sorensen, 2005). It is laden with accusations of racism in white feminism that ignores the particular histories of slavery, and neglects the racial oppression and invisibility of women of color and Black feminists (Collins, 1990). White feminists' concerns for the right to abortion, for instance, ignored the subject of the forced sterilization of Black women. Influential texts include Barbara Smith's "Toward a Black Feminist Criticism" (1977), Alice Walker's "In Search of Our Mothers' Gardens" (1983), Bell Hooks' "Ain't I A Woman? Black Woman and Feminism" (1981), Hazel Carby's "White Woman Listen!: Black Feminism and the Boundaries of Sisterhood" (1982), and Audre Lorde's "Sister Outsider" (1984). Black feminists argued that their political reality was one of a simultaneity of oppression, underscoring the notion of interlocking identities or intersectionality. Indeed, in her landmark text "Black Feminist Thought," Collins (1990) insists that in order to appropriately examine and respond to black women's lives, (black) feminist theory must be built simultaneously on the experiences and effects of race, gender, and class. Crenshaw (1989) coined the term 'intersectionality' to describe how 'multiple oppressions' are experienced, and Collins (1990) introduced 'the matrix of domination' to analyze how these intersecting oppressions are actually organized. Black feminists formed organizations to fight these multiple oppressions including the National Black Feminist Organization (NBFO) of 1973 and the Boston-based Black lesbian feminist Combahee River Collective of 1974. The Collective also expanded the by-now popular feminist mantra, "the personal is political," to include the notion that the personal is also cultural. They note that "Even our Black women's style of talking, testifying in Black language about what we have experienced, has a resonance that is both cultural and political" (Combahee River Collective, 1982, p. 17). Further, unlike white lesbian feminist separatism, the Collective advocated an inclusive political strategy that would enable black women and men to fight together against racism while simultaneously challenging sexism. Indeed, Black feminists have sought to differentiate their feminisms via naming, to include the traditions and values within black communities. Some reject the term feminism altogether preferring to discuss issues of gender. The unease with the term feminist goes beyond culture as Purkayastha et al. note: "unease arises from how the conceptualization of individuals is taken for granted. A sharp distinction is assumed to separate the individual from any collectivity" (2003, p. 506). Along comes 'womanism' a key feature of which is the need for community – including men – to fight injustice, and not merely an emphasis on individual advancement or emancipation. Alice Walker is credited with introducing the term 'womanist' in 1980. The term was adopted by many African feminists independently of Walker (see Ogunyemi, 1985). 'Womanism' is equated to Black feminism, and is portrayed as superior to White feminism – promoting relations between women and men; privileging motherhood, community, and spirituality. A fourth definition, according to Walker, and where she deviates from African womanism, speaks of romantic love between women, thereby making a claim for Black lesbian feminism. African women outside of Southern Africa have generally not embraced the LGBT movement, even though same-sex love between women, including romantic love, is increasingly acknowledged and discussed (see Tamale, 2011). Audre Lorde, lesbian feminist activist, was instrumental in linking heterosexuality with discussions of women's oppression. She used her writings and speeches to argue that heterosexuality is institutionalized to perpetuate the social power of men across class and race. In Germany, Eggers referenced a series, 'Adefra,' with Peggy Piesche and Katja Kinder where they talk about The Black Women's Movement in Germany. On the level of practice, Black Germans, through 'Adefra,' have also always found it important to "share activist spaces with brothers, sometimes children and grown children of Black women Activists, sometimes siblings, sometimes fellow activist comrades from other organizations" (Maisha Eggers, personal communication on Facebook, 2 March 2014). This leads to conflicts with the White Women's Movement in Germany who had a strict no-men policy. '"Adefra' is also a space that embraces Black Lesbians. In the United Kingdom also there are fractures born of race and class differences – white women do not generally accept class issues based on race, while some black women do not find issues of sexuality a priority. In the Facebook© discussion, Wa Goro points to the longest surviving organization, the Southhall Black Sisters set up in 1979 to meet the needs of black (Asian and African-Caribbean) and minority ethnic women (http://www.southallblacksisters.org.uk/). Although the term womanist has declined in usage, the term did enjoy a brief renaissance with the publication of the Womanist Reader (Phillips, 2006). Although black women insisted on their special circumstances, African and African-American feminists also built effective relationships with White feminists on specific issues such as access to services including day care, reproductive health care, maternity leave, and prevention of and support for domestic/gender-based violence. This was admitted during the first conference of the National Black Feminist Organization in 1973.

Third World Feminisms

The Third World has been defined severally (Spivak, 1988; Sangari, 1990). Third World women have diverse experiences with different local, national, and regional contexts. However, the commonality of the shared experiences of slavery, colonization, imperialism, and capitalism unifies them in unique ways (Spivak, 2006). Third World feminisms have been heavily influenced by globalization, international funding for national NGOs, transnational advocacy groups, and the UN organizations. These further expose, influence, and increase the tensions that exist between feminisms. Early encounters between feminists from different countries during the nongovernmental forums of the UN women conferences resulted in claims that Third World women were not feminists – or feminist enough – as they did not prioritize the issues that Western feminists considered critical. In *Under Western Eyes*, Mohanty et al. (1991) contest the Eurocentric gaze that privileges Western notions of progress and arbitrarily homogenizes Black women as victims of ignorance and restrictive cultures (Mohanty, Spivak (1988)). Other Third World scholars contest Western feminism's naivety in speaking on behalf of Third World women, and about non-Western social structures they do not understand (Oyewumi, 1997).

Today, the categories that would be considered Third World are numerous and fluid – Asian, Arab, African, Caribbean, Latina, Indigenous peoples, etc. The problematics around the terms, for example, 'Asian women' has been discussed at length by scholars such as Arisaka (2000). At best, it is an ambiguous category, at worst, laden with racial (racist undertones such as non-Christian (i.e., heathen or barbaric)). There are many influential texts reflecting the diverse Third World experiences, approaches, and theories. However, one could summarize the following as critical areas that have received much attention: women's visibility and voice in politics; violence against women; sexualities; and relations with the state. Useful reviews/accounts include those by Bayoumi and Sidawi (2014) and Saadawi (2007) on the Arab world; Loomba and Lukose (2012) on South Asia; Purkayastha et al. (2003) and Maitrayee (2005) on India; and Mihesuah (2003) and Pettman (1992) on Native Americans and Australian women, respectively.

African Feminism

The origins of African feminism(s) cannot be specifically dated, and we distinguish this from Black feminism to underscore a more conscious link to the continent. It can be argued that it predates colonialism certainly in purpose if not in name as one is dealing with several languages that predate the introduction of European languages in the continent. This is evidenced by women's specific struggles for their gender concerns. African feminism has been defined as an intellectual and activist movement, rooted in the experiences of slavery, colonialism, and patriarchy and striving to end interlocking systems of multiple oppressions resulting from these experiences and their conditions (Adomako Ampofo and Arnfred, 2009; Adomako Ampofo et al., 2004; Mama, 2011; Steady, 1981). Significant attention is paid to describing precolonial and traditional systems of gender relations to help understand connections to contemporary experiences, thus it is very much grounded in relating lived experiences to theorizing, and addressing issues such as health, gender-based violence, sexuality(ies), education, the family, work, politics, the state, and women's organizing (Adomako Ampofo et al., 2004). This important criterion for scientific enquiry has often led to the unfortunate accusation that African (scientists) feminists are mere empiricists. It is recognized as heterogeneous, bearing the marks of its multiple and diverse colonial contexts, and influenced by a multiplicity of civilizations (Islamic, Christian, and indigenous) before being further shaped by an array of anticolonial, anti-imperialist and nationalist movements, and interactions with (white) Western feminism (Norwood, 2013; Mama, 2011). African feminisms have been "fundamentally deconstructive, inherently rebellious and unapologetically anti-imperialist, antiracist, antisexist, and (recently, since the third wave) antihomophobic" (Norwood, 2013, p. 226). They also have problematized White Western feminism(s) and the production of knowledge on Africa, introducing alternative concepts, methodologies, and theorizations that valorize African-centered knowledge. Particularly significant have been the scholar-activist collaborations that have addressed the social and political concerns of women and that have been instrumental in women mobilizing. Adomako Ampofo (2009) maps some of the significant moments in these collaborations, significant among them the setting up of the electronic network, GWS, at the then African Gender Institute at UCT, and the subsequent establishment of the flagship e-journal, *Feminist Africa*. The journal's subject matter reflects the diversity of issues of concern as well as its activist edge. Along with other units, the African Gender Institute has since been merged into a School for African & Gender Studies, Anthropology & Linguistics (see more at: http://www.humanities.uct.ac.za/hum/departments/axl#sthash.A6ZNgE2b.dpuf). A recent issue of *Feminist Africa* (2006, p. 7) has been devoted to Diaspora voices in which Reddock, a well-known Caribbean feminist scholar wrote the Editorial, and Barriteau, Rajack-Talley, Safa and Boyce Davies, fellow Caribbean feminists are contributors. In addition to defining, problematizing, and celebrating *diaspora* (Reddock), the contributors provide an eloquent summary of the status of black feminism for women in the Caribbean and in Brazil. As with continental Africa, much of the literature engages with the state's neglect of women's concerns, or their direct complicity in women's oppression. Additionally, Caribbean feminists have been concerned with issues of identity (e.g., Mohammed, 1998) and masculinities. Nnaemeka (1998) and Reddock (2006) provide excellent summaries of some of the better-known conferences and anthologies that have taken place about African women's issues, particularly in Africa, and that have provided opportunities for important collaborations.

Post–Third-Wave Feminisms

This period is so named to reflect emerging issues and the technological tools of the new millennium, as well as the post-9/11 culture-of-war. There are increasing engagements with states and international organizations in discussions around citizenship, identities, globalization, peace and security, terrorism, climate change, and food security; and feminists have recognized the need to coalesce around these. This wave

includes intergenerational feminisms, and the use of social media in feminist engagements. As part of Facebook© discussions with Black feminists in Europe, Maisha Eggers, in Germany, paid tribute to the power of digital campaigns for black women's spaces in Europe. She offered the example of a campaign, "My feminism will be intersectional, or it will be bullshit."

The narratives and methods of a younger generation's experiences have introduced new issues and approaches. We find a feminism that manifests itself in 'grrl' rhetoric, which challenges the notion of 'universal womanhood' and embraces ambiguity. Younger feminists of this era adopt body expressions of femininity that earlier feminists identified with a male oppressive culture (Baumgardner and Richards, 2000; Gillis et al., 2004). The attitude is that, "it's possible to have a push-up bra and a brain at the same time" (Pinkfloor cited in Krolokke and Sorensen, 2005, p. 20). Popular culture is engaged with, while sexist language and inappropriate terms such as 'slut' and 'bitch' are criticized. Discussions on feminism and popular culture are linked with cultural studies and gender roles, exploring the ways in which feminism and popular culture influence each other. The interactions have revealed the sexist representations of women that societies consume and take for granted. Of particular importance has been the objectification of women in audiovisual media and advertising. Munford and Waters point out that popular culture celebrates models of privileged 'housewifery' that encourage 'fairy-tale idealizations of feminine idleness' which is a concern in feminist discourses (2013, p. 93). Significant in this 'post-era' are discourses around religion and fundamentalism, Euro-American attitudes toward Islam particularly, and the resistances to these attitudes that have fed feminist theorizing and activism about the place of Muslim 'others' in Europe. Rottmann and Ferree (2008) compare feminist discourses around Europe's antidiscrimination legislation and headscarf cases in Germany, showing that the relationships among race, religion, gender, and the role of the state influences German feminists' discussions, particularly feminists of color. Traditionally, Islamic women who wear the veils or headscarf have been viewed as not 'feminist enough' or worse, as culturally backward (Mernissi, 1987). Ignored is the role of the veil to protect, or even empower women in contexts where their movement is otherwise circumscribed or even limited, or worse, where they are exposed to sexual harassment (Odeh, 1993). The discourse addresses the confrontations among globalization, terrorism, feminism, and war. In their edited volume, *Feminism and War: Confronting US Imperialism*, Riley et al. note the annihilating effects of war on the lives of women and how women's liberation is invoked "to justify, legitimize and continue wars through the use of racialized discourses of male supremacy and female helplessness" (2008, p. 11).

This 'post-era' also enjoys convergences. The issue of motherhood, for instance, has now gained such prominence in North America that an entire series is dedicated to it by Demeter Press. Convergences do not lead to a (re)universalization of womanness, however. Rather, global feminisms are approached in terms of similarities and differences, and exchanges of ideologies, human resources, political movements, and transnational feminist collaborations and activisms (Lal et al., 2010). One such successful collaborative project is the Pathways of Women's Empowerment project coordinated through the Institute for Development Studies, IDS, Sussex, and with partners in four regional hubs – South Asia, Latin America, North Africa, and West Africa.

Challenges remain, however. Finding a nuanced position that can specify women's particular social locations, their universal human rights, and a feminist identity is not easy. As Mohanty (1986) noted, a focus on universal gender oppression can mean that particulars such as race and class must become invisible in order for gender to become visible. Valorizing differences – sexualities, religion, global location, class, generation– while keeping the larger project intact requires deep sensitivity and skillful balancing.

There are also backlashes. Wa Goro notes, "As with 'Black' advancement, these [feminist] issues seem to take on cyclical and 'one step forward two steps back' kind of patterns. These are also differentiated in various contexts at any time, with some progressive pockets in some arenas with simultaneous backlashes or backsliding elsewhere as in all struggles... and of course the mass media have a role to play in deciding whether to advance these ideas or not, based on the wider political/economic reality..." (personal communication on Facebook discussion, 3 February 2014). As long as gender-based inequalities remain we will need feminism. But for feminism to survive it must be inclusive. In the words of Bryson, "what we dare insist on for the survival of feminism is that we share this movement towards an open-minded, generous, and inclusive feminism that moves beyond modern/postmodern, local/global, theory/practice and academic/real world dichotomies in what some might describe as a dialectical resolution that builds upon the best of earlier thought and experience to suggest ways forward for the future. Rather than suggesting that feminism is moribund, they show the breadth and diversity of feminist politics and ideas, the ability to develop even in a hostile climate, and the continuing importance of feminist concerns." (Bryson, 2002, p. 239).

See also: Afrocentric Approaches to Social Work; Ecofeminism; Feminism in Anthropology; Feminist Criminology and Gender Studies; Feminist Legal Theory; Feminist Political Theory and Political Science; Feminist Psychotherapy; Feminist Theory: Radical Lesbian; Gender and Feminist Studies in Geography; Gender, Feminism, and Sexuality in Archaeological Studies; Situated Knowledge, Feminist and Science and Technology Studies Perspectives.

Bibliography

Antrobus, 2004. The Global Women's Movement: Issues and Strategies for the New Century (Global Issues) by Peggy Zed books.

Adesanmi, Pious, 2011. You're Not a Country, Africa. Penguin Books, Johannesburg.

Adomako Ampofo, Akosua, 2009. One who has truth, she has strength: the feminist activist inside and outside the academy in Ghana. In: Adomako Ampofo, Akosua, Arnfred, Signe (Eds.), African Feminist Research and Activism-Tensions, Challenges and Possibilities. Nordic Africa Institute, Uppsala, pp. 28–51.

Adomako Ampofo, Akosua, Beoku-Betts, Josephine, Osirim, Mary, Njambi, Wairimu, 2004. Women's and gender studies in English Sub-Saharan Africa: a review of research in the social sciences. Gender and Society 685–714.

Ali, S., Coate, K., Wa Goro, Wangui (Eds.), 2000. Global Feminist Politics: Identities in a Changing World. Routledge, London.

Arisaka, Yoko, 2000. Asian women: invisibility, locations, and claims to philosophy. In: Zack, N. (Ed.), Women of Colour in Philosophy. Blackwell, New York, pp. 1–18.
Baumgardner, Jennifer, Richards, Amy, 2000. Manifesta: Young Women, Feminism and the Future. Farrar, Straus and Giroux, New York.
Bayoumi, Noya, Sidawi, Rafif Rida, 2014. Arab Feminisms: Gender and Equality in the Middle East (Contemporary Arab Scholarship in the Social Sciences. I. B. Tauris.
Bryson, Mary, 2002. Me/no lesbian: the trouble with "troubling lesbian identities". International Journal of Qualitative Studies 15 (3), 373–380.
Collins, Patricia Hill, 1990. Black Feminist Thought: Knowledge Consciousness, and the Politics of Empowerment. Unwin Hyman, Boston.
Collins, Patricia Hill, 2001. What's in a name: womanism, black feminism, and beyond. Black Scholar 26 (1), 9–17.
Chodorow, Nancy, 1978. The reproduction of Mothering. University of California Press, Oakland, California.
Crenshaw, Kimberle, 1989. Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination, doctrine, feminist theory and antiracist politics. The University of Chicago Legal Forum 140, 139–167.
Ebunoluwa, Sotunsa Mobolanle, 2009. Feminisms. The Journal of Pan African Studies 3 (1), 227–234.
Essof, Shereen, 2001. African feminisms: histories, applications and prospects. African Feminisms 50 (1), 124–127.
Friedman, S., 1995. Beyond white and other: relationality and narratives of race in feminist discourse. Signs 21, 1–49.
Gillis, Stacy, Munford, Rebecca, 2004. Geneologies and generations: the politics and praxis of third wave feminism. Women's History Review 13 (2), 177–178.
Hooks, Bell, 1981. Ain't I a Woman: Black Women and Feminism. South End Press, Moston, MA.
Imam, Ayesha, Mama, Amina, Sow, Fatou (Eds.), 1997. Engendering African Social Sciences. CODESRIA, Dakar.
Kabeer, Naila, 1994. Reversed Realities: Gender Hierarchies in Development Thought. Verso, London.
Krolokke, C., Sorensen, A.S., 2005. Three waves of feminism: from suffragettes to grrls. In: Krolokke, C., Sorensen, A.S. (Eds.), Gender Communication Theories and Analyses: From Silence to Performance. Retrieved from http://www.sagepub.com/upm-data/6236_Chapter_1_Krolokke_2nd_Rev_Final_Pdf.pdf.
Krolokke, Charlotte, Sorenson, Anne Scott, 2006. Gender Communication Theories and Analyses: From Silence to Performance. Sage Publications Inc, California.
Lal, Jayati, et al., 2010. Recasting global feminisms: toward a comparative historical approach to women's activism and feminist scholarship. Feminist Studies 36 (1), 13–39.
Legates, Marlene, 2001. In Their Time: A History of Feminism in Western Society. Routledge, London.
Loomba, Ania, Lukose, Ritty A., 2012. South Asian Feminisms. Duke University Press, Durham, NC.
Lorde, Audre, 1984. Sister Outsider: Essays and Speeches. Crossing Press, California.
Maitrayee, Chaudhuri (Ed.), 2005. Feminisms in India. Zed Books, London.
Mama, Amina, 2011. What does it mean to do feminist research in African contexts? In: Feminist Review's Conference Feminist Review, e1–e20.
Mbilinyi, Majorie, 1992. Research methodologies. In: Meena, Ruth (Ed.), Gender in Southern Africa. SAPES Books, Harare.
Mensah-Kutin, Rose, et al., 2000. The National Machinery for Women in Ghana: An NGO Evaluation. Third World Network-Africa, Accra.
Mernissi, Fatima, 1987. Beyond the Veil: Male-Female Dynamics in Modern Muslim Society. Indiana University Press, Bloomington and Indianapolis.
Mihesuah, Devon Abbott, 2003. Indigenous American Women: Decolonization, Empowerment, Activism. University of Nebraska Press (Contemporary Indigenous Issues).
Mohammed, Patricia, 1998. Towards indigenous feminist theorizing in the Caribbean. Feminist Review 59.
Mohanty, Chandra Talpade, Russo, Anne, Torres, Lourdes, 1991. Third World Women and the Politics of Feminism. Indiana University Press, Bloomington.
Mohanty, Chandra Talpade, 1986. Under Western eyes: feminist scholarship and colonial discourses. Boundary 12 (3), 33–58.
Munford, Rebecca, Waters, Melanie, 2013. Feminism and Popular Culture: Investigating the Postfeminist Mystique. I.B.Tauris, London.
Nnaemeka, Obioma, 1998. The Politics of (M)othering: Womanhood, Identity and Resistance in African Literature. Routledge, New York (Opening out: Feminism for Today).
Norwood, C., 2013. Perspective in African feminism: exploring expressions of black feminism/womanism in the African diaspora. Sociology Compass 7 (3), 225–236.
Odeh, Lama Abu, 1993. Post-colonial feminism and the veil: thinking the difference. Feminist Review 26–37.
Ogunyemi, Chikwenye Okonjo, 1985. Womanism: the dynamics of the contemporary black female novel in English. Signs 11 (1), 55–68.
Oyewumi, Oyeronke, 1997. The Invention of Women: Making an African Sense of Western Gender Discourse. University of Minnesota Press, Minneapolis.
Pettman, Jan, 1992. Living in the Margins: Racism, Sexism and Feminism in Australia. Allen & Unwin, London.
Purkayastha, Bandana, Subramaniam, Mangala, Desai, Manisha, Bose, Sunita, 2003. The study of gender in India: a partial review. Gender & Society 17 (4), 503–524.
Phillips, Logan, 2006. The Womanist Reader. Routledge, New York.
Reddock, Rhoda, 1998. Women's organizations and movements in the Commonwealth Caribbean: the response to the global economic crisis in the 1980s. Feminist Reviews 59.
Rathgeber, Eva, 1990. WID, WAD, GAD: trends in research and practice. Journal of Development Areas 24, 489–502.
Rottman, Susan B., Ferree, Myra Marx, 2008. Citizenship and intersectionality. Social Politics 15 (4), 481–513.
Saadawi, Nawal El, 2007. The Hidden Face of Eve: Women in the Arab World. Zed Books.
Sangari, Kumkum, Vaid, Sudesh, 1990. Recasting Women: Essays in Indian Colonial History. Rutgers University Press, New Brunswick.
Spivak, Gayatri, 1988. Selected Subaltern Studies. Oxford University Press, Oxford.
Spivak, Gayatri, 2006. In Other Worlds: Essays in Cultural Politics. Routledge Classics.
Steady, Philomena, 1981. The Black Woman Cross-Culturally. Schenkman, Rochester, VT.
Tamale, Sylvia, 2011. African Sexualities: A Reader. Pambazuka Press, Cape Town.
Tinker, Irene (Ed.), 1990. Persistent Inequalities: Women and World Development. Oxford University Press, Oxford.
Walker, Alice, 1983. In Search of Our Mother's Garden: Womanist Prose. Harcourt, California.
Yuval-Davis, Nira, 1997. Gender and Nation (Politics and Culture Series). Sage Publications Ltd, London.

Relevant Websites

http://www.adefra.com/ – Adefra.
www.awdf.org/ – African Women's Development Fund; AWDF.
http://hoodfeminism.com/ – Hood Feminism.
http://bitch.org/ – Bitch: Feminist Response to Pop Culture.
cegensa.ug.edu.gh/ – Centre for Gender Studies and Advocacy; CEGENSA.
agi.ac.za/feminist-Africa/ – Feminist Africa.
www.womenaid.org/press/ – Green Belt Movement.
www.blackpast.org/ – National Black Feminist Organization.
http://userpages.umbc.edu/~Korenman/wmst/programs.html – Southhall Black Sisters: University of Maryland Baltimore County. (2014, March). Women's Studies Programs, Departments, & Research Centers.
www.ids.ac.uk/project/pathways-of-women's-empowerment-research-project-consortium/ – Women's Empowerment Research Project Consortium (WE-RPC).
http://www.wluml.org/ – Women Living Under Muslim Laws.

Feminist Criminology and Gender Studies

Susan F Sharp, University of Oklahoma, Norman, OK, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Feminist criminology refers to approaches to the study of crime, victimization, and the criminal justice system using a critical, feminist lens. There are many different theoretical approaches, including standpoint feminism, pathways theory, and engendering mainstream theories. These approaches have a critical analysis of how gender inequities shape crime and responses to crime, emphasizing the role of male hegemony in society. Feminist criminology began in the 1970s, with liberal feminist approaches, followed by examination of the roles of victimization and abuse in female offending. These early feminist works were informed by second-wave feminism. More recently, third-wave feminism has informed multicultural and postmodern approaches in feminist criminology. Feminist criminology may use quantitative, qualitative, or mixed methods approaches. The subject matter includes women as victims, women as offenders, women working in the criminal justice system, and masculinities and crime.

In the field of gender studies, criminology has remained more androcentric (male centered) than most social sciences (Britton, 2000), in part because men commit both more crimes and more serious crimes than women commit (Flavin, 2001; Steffensmeier and Allan, 1996). Criminology focused for decades on crime by males and theories derived from the study of males, largely ignoring women. Most theories of crime that were developed prior to the 1970s were based solely on research with males, and their applicability to females was ignored (Heidensohn, 1969). Criminology was criticized mid-twentieth century by feminists for ignoring gender, where men were seen as the norm. Thus, feminist criminology is a relatively new approach to the study of victimization, criminal offending, and the criminal justice system. Feminist criminology is not simply the study of crime by women, nor is it just the study of women and crime. There are males who are feminist criminologists, and there are many female criminologists who are not. Additionally, there are many examinations of women and crime that are not feminist. Feminist criminology explicitly takes into account the unequal power of males and females in its approach to the study of crime and justice.

Overview of Feminism in Criminology

Feminist criminology might be more aptly named feminist 'criminologies,' feminist perspectives in criminology, or feminist epistemologies of crime. The field comprises diverse approaches and theories. What these have in common is the focus on gender as central to criminological inquiry. Feminist criminology is not a specific theory but rather application of feminist perspectives to the field of criminology, with multidisciplinary roots and perspectives. One of the hallmarks of feminist criminology is that it is grounded in the examination of gender relations in society. As in any feminist theorizing, the central focus of feminist criminology is the understanding of how gender informs the understanding of social actions, in this case criminal behavior and victimization. The integration of feminist thought into a primarily male-oriented discipline has resulted in a range of loosely related theories and approaches.

Feminist criminologists have also criticized mainstream criminology for the gendered reproduction of knowledge that keeps criminology focused on the experiences of men (Renzetti, 1993). The problem is that feminist perspectives are usually separated from other theories in textbooks and given only cursory coverage. They also are rarely published in mainstream journals. Thus, generations after generations of students are unaware of feminist criminology and its contributions. They in turn teach their own students mainstream theories, with little, if any, reflection on more explicitly feminist approaches. Although it is becoming more mainstream today, feminist work is still largely relegated to specialty journals such as *Violence against Women*, *Women and Criminal Justice*, and *Feminist Criminology*, in special issues of other journals, or in trade books and texts on feminist criminology. This leads to a reproduction of knowledge that largely overlooks feminist thought. However, some mainstream theorists such as Agnew (Broidy and Agnew, 1998) have pointed out the importance of incorporating feminist approaches, not just women, into criminology as a field of study.

Feminist Criminological Approaches and Methodologies

Drawing from feminist theories in general, feminist criminologists tend to fall into three broad camps. Liberal feminist and radical feminist approaches undergirded much of the first two decades of theorizing and research in feminist criminology, while postmodern approaches have gained prominence in the last two decades.

In liberal feminism, the focus is on obtaining equality in the public sphere. This has led to efforts to achieve gender parity in prison educational and therapeutic programming. While this is beneficial, women prisoners' needs do not mirror those of men. Simply adding the same or similar programming and services

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45059-2

may fail to address women's health and mental health needs, which differ significantly from those of men. Thus, an explicitly feminist approach is needed to determine the needs of women prisoners. There is a dark side to liberal feminist efforts, with parity being translated as applying equally punitive approaches to women's corrections.

In contrast, the work of radical feminists has largely focused on women as victims. Radical feminists have largely focused on how societal inequality and patriarchy lead to gendered violence. Another important aspect of radical feminist criminology is the focus on activism to bring about social change. Radical feminists have worked to address reproductive justice and the criminalization of women's bodies, as well as on violence against women.

As a part of third-wave feminism, a third approach has gained prominence in feminist criminology over the last two decades. Postmodern feminism focuses on the intersections of gender, class, and race, arguing that there is not one overarching experience for women. Instead, an individual is placed in a variety of social positions. The impact of each of these (race, class, sex, nationality, age, and sexual orientation) is not simply additive. Instead, oppression is often multifaceted. To accurately describe and report the experiences of women as victims or offenders, it is important to understand these interlocking systems of oppression.

Within the broader approaches of second- and third-wave feminism, there are feminist empiricists and standpoint feminists (Ngaire, 1996). Feminist empiricists adhere to mainstream scientific methods, incorporating women and their viewpoints in the analyses. Feminist empiricism has helped broaden the knowledge about crime and the criminal justice system overall by incorporating the perspectives of women into research (Flavin, 2001). Empirical feminist work strives to measure some agreed upon reality in the social world (Daly, 1988), and may or may not be feminist. Standpoint feminists go further, challenging the validity of knowledge from empiricism. The standpoint perspective argues that knowledge is shaped by the power relations in society and thus never completely objective. They begin with the standpoint or position of the individuals under study (Daly, 2008; Flavin, 2001). Both are considered feminist approaches. Finally, some feminists take a deconstruction approach to challenge either–or binary thinking patterns about gender and other social placements, focusing instead upon the complex nature of social location (Daly, 2008).

There is also no single methodological approach that is considered to be feminist, either. Feminists may do quantitative analyses or qualitative work. However, simply adding gender as a control variable does not make the work feminist. Instead, careful research designs are utilized based on theoretical perspectives that take into account the hierarchical gender structure of society. In-depth interviews and life histories are types of qualitative approaches, with an emphasis on depth and meaning rather than description. Qualitative work focuses far more on meaning and process. Daly (2008) suggests that using both qualitative and quantitative approaches is desirable in the development of theory.

Reflexivity is a final important aspect of feminist research. It is intricately linked to both theory and method. Reflexive work critically analyzes the underlying assumptions involved in the work as well as the process. Feminist criminologists consider this an important part of doing research, because the goal is to identify bias introduced into the process. For some feminist researchers, reflexivity may include examining their own reactions to the research (Flavin, 2001). Additionally, feminist researchers are aware of the imbalance between the researcher and subject. In particular, feminism cautions against the potential for exploitation (Flavin, 2001).

It is readily apparent that theories and methods vary widely among those who consider themselves to be feminist criminologists. Additionally, the degree of reflexivity can range from identifying potential biases to reflection on the researcher's role in the process. Feminist criminology, however, has added to the knowledge about why both women and men commit crime, as well as the roles of personal history and social placement. It has drawn attention to inequities in the system, and usually called for change. Activism and praxis are both parts of feminist approaches to criminology, which sometimes put feminist criminology in conflict with a more mainstream approach that assumes research should be objective and detached.

Development of Feminist Perspectives in Criminology

While there were a few studies of female crime in the past, the role of hierarchical gender relations was largely ignored. The focus was on women offenders as women who deviated from socially accepted gender roles. For example, Lombroso and Ferrero (1893, translation by Rafter, 2004) tried to explain women's crime by suggesting that women who committed crime were more like men than like other women. Their argument linked women's sexual behavior with criminality, suggesting that women who were more sexually aggressive were more masculine. They believed this masculinity led to female offending. Likewise, in *The Unadjusted Girl*, Thomas (1923) focused on female sexuality and equated promiscuity with delinquency. Little if any attention was paid to the relative powerlessness of women or how patriarchal society controlled their sexuality and work. In a more negative analysis, Pollak (1950) argued that women were more devious than men, and their crimes were more hidden. Thus, their underrepresentation in the criminal justice system was due to their cunning as well as chivalrous and preferential treatment, not lower levels of offending. Recent research, however, has underscored the reality that females offend less often and less seriously than men. Feminist criminologists point out that understanding women's offending and their lower offending rates can aid in understanding male offending.

During the second wave of feminism, two books on women and crime were published that illustrate the difference between the study of women and crime and what is more generally considered to be feminist criminology. Adler (1975) published *Sisters in Crime*, arguing that as women gained more opportunities in mainstream society, this would translate to masculinization of women and more crime. She envisioned higher levels of property crimes by women as a result of their 'liberation.' Similarly, Simon (1975) argued that women's 'emancipation' would create more opportunities for women. The response of feminists to these two approaches was that neither considered the role of a patriarchal society in which

males exercised control over women's sexuality as well as their workforce participation. Furthermore, this approach was based on the faulty assumption that women had achieved equality. Thus, the contributions of Adler and Simon are generally not considered to be part of the body of feminist criminology, although some consider their work to exemplify liberal feminism due to the focus on legal and economic equality (Daly and Chesney-Lind, 1988). At the same time, second-wave feminism gave birth to a growing awareness of the importance of placing women at the center of the study of crime. This led to further work that was more overtly feminist.

Although there were important contributions prior to the publication of Carol Smart's (1976) critique of mainstream criminology, the emergence of feminist criminology is often attributed to her book. Her work was the first explicitly feminist approach to the study of women and crime and critique of mainstream criminology. Following its publication, a body of research and theory began emerging that is considered to be feminist criminology.

In an important treatise on feminist criminology, Daly and Chesney-Lind (1988) point out that there are two important issues that should be addressed by criminological theory. First, theories should be able to address the gender gap or gender ratio issue. In other words, since males commit more crime and more serious crime, theories should help us understand why this occurs. This requires careful consideration of gender roles and gendered imbalances in power that help predict the higher rates of male offending and lower rates of female offending. Second, criminology needs to address the generalizability issue. It should be able to explain not only men's, but also women's offending. Daly and Chesney-Lind (1988) further delineate five elements of feminist thought that must be taken into account for a theoretical approach to be considered truly feminist. First, it must take into account that gender is a complex set of cultural and social aspects that are related to biological sex differences. Second, feminist thought is based on the understanding of the ways that gender organizes social life and social relations. Third, this organization of social life and relationships is hierarchical. Males have more power than females. Fourth, this male hegemony results in the systems of knowledge reflecting the male view of the world. Thus, scholarship and knowledge maintains a strong gendered aspect. Finally, feminism argues that women should be at the center of intellectual inquiry (Daly and Chesney-Lind, 1988).

Much of early feminist work focused on the overt control of women's sexuality, the misrepresentation of female offenders in mainstream criminology, and the dearth of scholarship that focused on the role of patriarchal society in female offending. During the 1970s and early 1980s, considerable work began emerging that argued that it was the explicit control of girls' and women's lives that resulted in their criminal justice involvement. In particular, Chesney-Lind and others began delineating the gender inequities in the juvenile justice system. Chesney-Lind focused on the overdetention of minor girls for status offenses and how that was related to efforts to control young girls' sexuality. Furthermore, girls often engaged in status offenses because of abuse, especially sexual abuse in their homes. She found that girls were far more likely than boys to be detained for offenses like running away and curfew violations. In contrast, boys were more likely to be detained for criminal offenses and rarely entered into the system for status offenses. Taking an overtly feminist stance, Chesney-Lind (1973 and other work) argued that these status offenses were often due to girls' responses to abuse in the home. Girls who were abused protected themselves by running away from the abusive environment and consequently found themselves caught up in the juvenile justice system and labeled delinquent.

During the 1970s and 1980s, feminist criminological work tended to focus on three main areas: women as victims, women and girls as offenders, and institutional reactions to women's and girls' offending. At the forefront of the emerging body of feminist criminology, important work on intimate partner and sexual violence began appearing. This work was largely an outgrowth of the second wave of feminism. The decidedly masculine nature of both physical and sexual violence and the fact that victims of intimate violence and sexual violence were overwhelmingly female underscored much of the work during the late 1970s and 1980s. Susan Brownmiller's (1975) classic study of rape brought attention to the power dimension of rape, redefining the crime as one of male dominance. Feminist criminologists brought a variety of perspectives to bear on male victimizers, female victims, and the ways in which the gendered social structure facilitated violence against women. Indeed, so much research was published in this area and feminists took such a strong activist stance that the Violence Against Women Act was first passed in 1994, with reauthorization in 2000, 2005, and 2013. In 1995, the academic journal *Violence against Women* was launched, focusing on primarily feminist research in this area. One of the contentions of feminist criminologists who focused on violence against women was that women's victimization was qualitatively different from men's victimization because of their lesser power and the high probability that the perpetrators were related to the women through familial or romantic liaisons. Activism by radical feminists was successful in achieving redesign of the National Crime Victimization survey to specifically ask about sexual victimization and victimization by family members and intimates. This has added considerably to the knowledge about violent and sexual victimization.

Feminist work has also focused on domestic or intimate partner violence. In reaction to research suggesting that women were as abusive toward romantic partners as males, feminist work has focused on the power differentials between men and women and on women as victims. Early work (Dobash and Dobash, 1979) focused more narrowly on the abuse of wives by husbands. More recent feminist thought argues this is too narrow, adopting the term 'intimate partner abuse,' which includes physical, sexual, and psychological abuse, as well as destruction of pets and property (Tong, 1984). Feminists have also sought to dispute research that suggests that women are as violent toward their partners as men. Much of that body of work uses the Conflict Tactics Scale developed by Straus and Gelles. Feminists take issue with two aspects of that instrument. First, it fails to take actions in context. Because it is a simple count of actions taken by the subject and the subject's partner, it is impossible to determine the context. The question of whether the counted action was in reaction to force from the partner is unanswered. More importantly, mere counts of aggressive behaviors do not

capture the level of seriousness and the harm caused by the action. Feminists argue that the data show that abuse of females by males leads to more serious harms.

By the middle of the 1980s, feminist criminologists were developing theories to explain female offending as well as examining the experiences of women caught up in the criminal justice system. Early works by Carlen (1983) and others sought to provide a more accurate portrayal of who women who offended were and how their life histories explained their offending. By the 1990s, there were numerous books available on women and offending, many from a feminist perspective. The role of childhood maltreatment began emerging as a major explanatory factor in the offending of girls and women (Acoca, 1997; Chesney-Lind, 1973). Research began focusing on the early lives of female offenders, noting the high prevalence of sexual, physical, and psychological abuse as well as neglect. Feminist pathways approaches have focused on the lives of women prior to their offending, attempting to address the reasons why women and girls offend. Early pathways research utilized interviews with prostitutes, while more recent work has used incarcerated women and girls to further develop the understanding of female offending. An important aspect of pathways approaches is the use of the voices of the girls and women themselves. Pathways approaches have documented the link between victimization and offending (cf Acoca, 1998; Belknap, 2007). Abuse in both childhood and adulthood has emerged as a common theme. Women who were abused as children enter abusive relationships in adulthood. Mental health problems often occurred, and substance abuse was frequent (Owen, 1998).

During the same time frame, work began emerging that focused on women's experiences in the criminal justice system. In one of the most thorough treatments, Owen (1998) provides an in-depth examination of the experiences of women in prison. Others have focused on questions about unequal treatment in crime processing. Still others have utilized the knowledge gained about the backgrounds of female offenders to challenge to programming available to women and girls in the criminal and juvenile justice systems. Because of the intricate relationship between women's victimization and their offending, gender-sensitive and trauma-informed programming has been introduced into various systems around the world.

The next important development in feminist criminology was recognition of the differences among women rather than focusing on gender as a unilateral trait. This postmodern approach has emphasized that race, ethnicity, class, national origin, religion, immigrant status, and sexual orientation play both independent and interlocking roles in the lives of women (cf Richie, 1996). Informed by emerging critical race feminism in other fields, feminist criminology began incorporating a multifaceted lens that took into account the multiple social placements of women. Potter's (2006) Black Feminist Criminology focuses on the unique experience of African-American women, while Burgess-Proctor (2006) focused on the need for an approach that examines multiple systems of power and oppression.

Feminist criminology focuses in part on the gender gap in offending, examining the role of gender in the occurrence of crime. As described above, criminology should not only describe why women offend less but also why men offend more. Thus, it has also contributed to studies of masculinity and crime. Messerschmidt (1993) has produced a body of work on masculinities and crime that is considered to be feminist. His work focuses on three elements of the social structure that increase the likelihood of male offending: the gendered division of labor, male hegemony, and sexuality. His approach takes into account not only gender but also social placement, especially sexuality, race, and social class. Because the patriarchal social structure expects men to be more successful, this becomes the goal of young men. Those males born into a more privileged class (middle), race (White), and who exhibit the preferred sexual orientation (heterosexual) have access legitimately to success, through education and occupation. For those who are not as fortunate, nonlegitimate avenues of success, such as crime and delinquency, are used to establish success. From Messerschmidt's perspective, masculinity is what best explains the commission of crimes, because males feel pressure to demonstrate power. In general, women do not have this issue, so they offend at lower rates and for different reasons. While Messerschmidt has taken a feminist approach to explaining male criminality, his work does not provide an explanation for why females commit crime; thus, it addresses the gender ratio issue but not the generalizability issue raised by Daly and Chesney-Lind (1988).

It is also worth noting that some feminist criminologists focus on gender and work in the criminal justice field. Because of the masculine nature of crime and male hegemony, very few women worked in most areas of the criminal justice system until the late 1970s and early 1980s (Belknap, 2007). The blockage of women from opportunities has been blamed at times on the lesser physical strength of females. This presupposes that women are incapable of performing as police officers or working in prisons. Despite Title VII of the Civil Rights Act, women can still be legally barred from occupations through application of bona fide occupational qualifications (Belknap, 2007). These can be used to argue that a man should be hired for the position because it is more rational to do so. Thus, even in the twenty-first century, women remain grossly underrepresented in both policing and corrections, although their representation is higher.

Working in prisons can often be difficult for women, particularly if they are oriented toward helping. Belknap (2007) takes the stance that one should not use the term correctional officers because of the paucity of rehabilitation in the prisons and jails as well as the lack of rehabilitative training by prison and jail workers. Additionally, women may face difficulties with promotion opportunities. One line of thinking suggests that women should be employed in women's prisons because the prisoners have been traumatized by males and will respond better to females. The problem is that there are far fewer women's prisons and therefore fewer chances for advancement. When women are relegated primarily to women's facilities, they often find themselves unable to advance in their careers, while similarly placed males in men's facilities are promoted (Baunach and Rafter, 1982). Another argument is that it is

a violation of privacy for women to work in men's facilities (Belknap, 2007).

Women have been in the field of policing longer, but they have also faced numerous restrictions there. Initially, they worked as matrons, and during the first decades of the twentieth century they often were confined to specialized roles in line with tradition gender roles. Their positions were often with youth or women and were primarily custodial (Belknap, 2007). As late as 2005, only approximately 10% of police officers were female (Belknap, 2007). Reasons include perceptions that women are less capable, incompatibility of shift work with raising families, physical limitations, and resistance by male officers' (Natarajan, 2001) requirements.

Feminist Criminology and Professional Organizations

The two major academic organizations in criminology, the American Society of Criminology (ASC) and the Academy of Criminal Justice Sciences (ACJS), both have subdivisions that are profeminist. The Division on Women and Crime (DWC) of the ASC grew out of an informal group that emerged into a caucus in the larger professional organization. The DWC was granted divisional status in late 1981, with its bylaws accepted in 1983 and its first board of officers elected in 1984. Currently, the division has over 300 members. In 2006, the DWC launched the first issue of its own journal, *Feminist Criminology*, dedicated to feminist research and theory in criminology. While not all the members of the division are overtly feminist, many are. Areas of scholarship cover a wide range of topics, including the roles of both gender and sexual orientation in crime, victimization, and criminal justice responses to crime.

The ACJS is the other national academic organization in criminology. The Minorities and Women section was born during the early 1990s due to a commitment to affirmative action. A constitution was drafted and approved by the ACJS in 1991. The initial planning for long-range goals occurred in the summer of 1992. The section was developed as a more liberal feminist approach to criminology, focusing mainly on issues of diversity and inclusiveness. However, the organization has a large number of feminist members.

Summary

Feminist criminology is a descriptive term for a wide range of topics, theories, and methodologies. There is no unified theoretical approach but rather an emphasis on placing theory and analysis in the framework of male hegemony. Feminist criminologists may utilize mainstream theories in a feminist manner or ascribe to more overtly feminist approaches. Research may be quantitative, qualitative, or both, but it is characterized by reflexivity and awareness of power imbalances between researcher and researched. Feminist criminology examines women as offenders, women as victims, the relationship between victimization and offending, the gendered nature of criminal justice, and women working in the system. Despite the wide variety in approaches, however, the field is unified by a few important points. First, women are normally placed in the center of inquiry. Second, patriarchal society and its role in creating offending, victimization, and work opportunities is considered. Third, activism and emphasis on creating social change through research is emphasized. While some work may embrace only one or two of these, for criminology to be feminist, there must be at least one, and usually more, of these present.

See also: Crime and Gender; Critical Criminology; Intersectionality and Crime; Rape and Sexual Coercion.

Bibliography

Acoca, L., 1998. Outside/inside: the violation of American girls at home, in the street and in the juvenile justice system. Crime and Delinquency 44, 569–581.

Adler, F., 1975. Sisters in Crime. McGraw-Hill, New York.

Baunach, P.J., Rafter, N.H., 1982. Sex-role operations: strategies for women working in the criminal justice system. In: Rafter, N.H., Stanko, E.A. (Eds.), Judge, Lawyer, Victim, Thief. Northeastern University Press, Stoughton, MA, pp. 341–358.

Belknap, J., 2007. The Invisible Woman: Gender, Crime and Justice, third ed. Wadsworth, Belmont, CA.

Britton, D.M., 2000. Feminism in criminology: engendering the outlaw. The Annals of the American Academy of Political and Social Science 57, 57–76.

Brownmiller, S., 1975. Against Our Will: Men, Women and Rape. Simon and Schuster, New York.

Burgess-Proctor, A., 2006. Intersections of race, class, gender and crime: future directions for feminist criminology. Feminist Criminology 1, 27–47.

Carlen, P., 1983. Women's Imprisonment: A Study in Social Control. Routledge and Kegan Paul, London, UK.

Chesney-Lind, M., 1973. Judicial enforcement of the female sex role: family court and the female delinquent. Issues in Criminology 8, 51–69.

Chesney-Lind, M., 1986. The female offender. Signs 12, 78–96.

Daly, K., 2008. Feminist perspectives in criminology: a review with gen y in mind. In: McLaughlin, E., Newburn, T. (Eds.), The Handbook of Criminal Theory. Sage, London.

Daly, K., Chesney-Lind, M., 1988. Feminism and criminology. Justice Quarterly 5 (4), 9–48.

Dobash, R.E., Dobash, R., 1979. Violence against Wives. Free Press, New York.

Flavin, J.M., 2001. Feminism for the mainstream criminologist: an invitation. Journal of Criminal Justice 29, 271–285.

Heidensohn, F., 1969. The deviance of women: a critique and an enquiry. British Journal of Sociology 19, 160–176.

Lombroso, C., Ferrero, G., 2004. Criminal Woman, the Prostitute, and the Normal Woman (A new translation with an introduction and annotations by Nicole Hahn Rafter and Mary Gibson). Duke University Press, Durham, NC.

Messerschmidt, J.W., 1993. Masculinities and Crime. MD Rowman and Littlefield, Lanham.

Miller, J., Mullins, C., 2006. The status of feminist theories in criminology. In: Cullen, F., Wright, J.P., Blevins, K. (Eds.), Taking stock: the status of criminological theory/advances in criminological theory, vol. 15. Transaction, New Brunswick, NJ, pp. 217–250.

Naffine, N., 1996. Feminism and Criminology. Temple University Press, Philadelphia, PA.

Natarajan, M., 2001. Women police in a traditional society. International Journal of Comparative Sociology 42, 211–233.

Owen, B., 1998. "In the Mix": Struggle and Survival in a Women's Prison. SUNY Press, Albany, NY.

Pollak, O., 1950. The Criminality of Women. University of Pennsylvania Press, Philadelphia.

Potter, H., 2006. An argument for black feminist criminology: understanding African American women's experiences with intimate partner abuse using an integrated approach. Feminist Criminology 1, 106–124.

Renzetti, C.M., 1993. On the margins of the malestream (or, they still don't get it, do they?): feminist analyses in criminal justice education. Journal of Criminal Justice Education 4, 219–234.

Richie, B.E., 1996. Compelled to Crime: The Gender Entrapment of Battered Black Women. Routledge, New York.
Simon, R.J., 1975. Women and Crime. Lexington Books, Lexington, MA.
Smart, C., 1976. Women, Crime and Criminology: A Feminist Critique. Routledge and Kegan Paul, London.
Steffensmeier, D., Allan, E., 1996. Gender and crime: toward a gendered theory of female offending. Annual Review of Sociology 22, 459–487.
Tong, R., 1984. Women, Sex and the Law. Rowman and Allanheld, Totawa, NJ.

Relevant Websites

http://ascdwc.com/ – Division on Women and Crime of the American Society of Criminology.
http://www.acjs.org/minorities_and_women_section.cfm – Minorities and Women Section of the Academy of Criminal Justice Sciences.

Feminist Legal Theory

Genevieve R Painter, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by L. Obiora, R. Perry, volume 8, pp. 5464–5469, © 2001, Elsevier Ltd.

Abstract

Feminist legal theory manifests through writing and speaking about 'law' and 'women,' in an effort to promote and improve understanding about justice. Feminist legal theory is a set of ideas, an activity engaged in by thinkers in and outside academia, and an intellectual and political movement. Developments in feminist legal theory emerged through engagement with problems rooted in inequalities, experienced by individuals and communities, at the hands of people, corporations, or the state. This article draws out key areas of tension within the field of feminist legal theory, focusing on English-language feminist legal theory and spanning the field of national jurisdictions and international human rights.

Introduction

Feminist legal theory manifests through writing and speaking about 'law' and 'women,' in an effort to promote and improve understanding about justice. Much of that writing and speaking debates the meaning of the terms 'law' and 'women.' Feminist legal theory is a set of ideas, appearing as truth claims, critiques, and/or advocacy messages, an activity engaged in by thinkers in and outside academia, and an intellectual and political movement. In its life span, it has gone through several evolutions, to produce a plural, complex, and fractured body of thought and practice. Evidence for the maturity of feminist legal theory, as a field, can be found in a number of edited volumes and dedicated academic journals. Defining feminist legal theory as a set of ideas and practices situates it across academia, nongovernmental organizations, law firms, think tanks, community groups, and public office. Developments in feminist legal theory have emerged through engagement with problems connected to inequalities, as experienced by individuals and communities at the hands of people, corporations, or the state. The problem-driven impetus of feminist legal theory contributes to its present-day heterogeneity. It also explains how a 'field' of feminist legal theory survives, despite the enduring debates about pillar concepts like 'woman' or 'equality' and resistance to the normalizing and disciplining effects of definition.

The separation line between feminism and feminist legal theory is porous. While feminist legal theory takes law as a primary focus, the whole field of feminist studies regularly engages legal issues because the state is central to feminism's reformist agenda. Contemporary Western feminism's roots lie in nineteenth-century campaigns to abolish slavery and secure the right to vote for women and enslaved people. Although some US campaigners fought to end both slavery and women's oppression, the post-Civil War movement for women's suffrage splintered over decisions to prioritize the voting rights of white middle-class women.

As an intellectual and political enterprise, feminist legal theory has been a movement embedded in the political currents of its time. The twentieth-century's struggle between liberalism and Marxism set the stage for feminist debates about prioritizing civil equality rights over material inequalities. The contemporary rollback of the welfare state, the foregrounding of the nuclear family as the caretaker of human vulnerability, and the disproportionate impact of austerity policies on marginalized women provide the context for the critical attention among feminist legal theorists to issues ranging from the feminization of the global care market to the queer critique of same-sex marriage.

Though feminist theory and activism gained strength into the twentieth century, it was not until the 1970s that 'feminist legal theory' came to be labeled as a field, with different schools of thought appearing and receding over time. Six strands are signposted here and developed further throughout the article. The equality stage of feminist legal theory, sometimes labeled 'liberal feminism' (1) and dated to the 1970s, focused on formal inequalities between women and men and called for ending legal discrimination. By the 1980s, radical or dominance feminism (2) and difference feminism (3) shifted attention to the differences between women and men and the social production of inequality. Both schools of thought were criticized for single-minded attention to the concerns of class- and race-privileged women. An antiessentialist movement (4) introduced thinking and action on multiple and intersecting forms of discrimination. Postmodernism (5) turned attention from dichotomous and categorical identities toward questions of social construction and subjectivity. Queer legal theory (6) challenged foundationalist understanding of sex and sexuality.

This article addresses English-language feminist legal theory in North America, the UK, and Australia, with some additional international representation by authors writing in English. It draws out key areas of tension within the field of feminist legal theory oriented toward domestic and national jurisdictions, and then surveys the state of critique in international human rights.

Method and Epistemology

Feminist legal critique typically starts from the position that understanding material, structural, and ideological conditions requires an attention to sex and gender and their interaction with other axes of differentiation. A common analytic technique focuses on exposing features of an issue that other

methods overlook, by, for example, identifying the gender implications of rules and practices that seem neutral (Bartlett, 1990). However, some critics reject the idea that feminist analysis constitutes the addition of a 'gender lens' on some 'social real' defined on other primary material, ontological, or legal grounds (Morgensen, 2012).

Feminist legal theory has responded to concomitant shifts in legal scholarship and feminism. First through legal realism, then through law and society research and critical legal studies, the gaze of common law scholars turned from the theoretical ideal of law on the books and toward the law in action. While the new empiricism brought concern for making law more effective, it also challenged the hermetic seal between 'law' and 'society' and introduced the notion of the coconstitution of the law, society, and the subject.

Feminist theory traversed similar epistemological terrain. Liberal feminist legal theory revealed the male bias in formal law and legal processes and argued for the reform of law from a neutral perspective. Radical feminists exposed law's neutrality as a patriarchal sleight-of-hand. According to MacKinnon, male dominance's "point of view is the standard for point-of-viewlessness; its particularity the meaning of universality" (MacKinnon, 1989: pp. 116–117). A critique based on standpoint theory argued that this quest for neutrality presumed a view from nowhere – and that there can be no such thing. Rather, a truth could best be reached by treating women as authorities about their own lives and experiences (hooks, 1989). Feminist legal theorists seized the invitation to listen to women's voices in the legal arena, by, for example, analyzing how trial processes heard and shaped women's narratives of violence (Sheppard and Westphal, 2000) or thinking about whether women bring a unique perspective to judging (Minow, 1992). The epistemology of postmodern feminism questioned the whole project of an authoritative perspective on truth, calling instead for a nonfoundational, nonessentialist analysis. Mariana Valverde challenged the law's claims about the unity and coherence of categories like the 'legal subject,' and instead studied the law as a site for discourses and processes that shape what is taken as knowledge (Valverde, 2003). Feminist legal theory also drew on queer theory, a scholarship that postures itself against, or at variance with, the normative. According to queer legal theorist Adam Romero, by jettisoning foundational epistemologies in favor of a positionality consistently opposed to the norm, it becomes difficult to "know definitively what feminist legal theory and queer legal theory, as such, substantively are or do" (Romero, 2009: p. 197). Thus, an attention to epistemology exposes the challenges of the bases of feminism as a political and critical project.

The Human

The human is the starting point for most feminist legal theory. The field of posthuman feminist legal thought reaches beyond the human to nonhuman life, the environment, and the material. Ecofeminism challenges traditional environmental law (Rochette, 2002). Maneesha Deckha has argued that understanding animal oppression is essential to both understanding human forms of oppression and developing a coherent concept of justice (Deckha, 2006). Materialism was largely thrown out with the Marxist bathwater of the late twentieth century (Conaghan, 2013), but feminist scholars drawing from Science and Technology Studies are bringing the material back into view, exploring the enmeshment of mind and body, nature and culture, and human and nonhuman forms of agency (Conaghan, 2013). For example, ultrasound allows us to 'see' a fetus on a screen, an intermingling of machine, body, and senses that informs ideas of the fetus as an autonomous human life (Barad, 2007). The turn to materiality provides a new lens in the debates among feminist legal scholars on abortion and reproductive technologies.

Equality

Because US law historically established and upheld gender inequalities, American liberal feminists took to the courts to dismantle them. Explicit distinctions between men and women were the target of the Equal Rights Amendment, a proposed amendment to the US Constitution that aimed at invalidating gender-biased statutes and common law rules. Though ultimately not passed into law, the goals of the Equal Rights Amendment took effect through litigation led by the Women's Rights Project at the American Civil Liberties Union. Its lead litigator, Ruth Bader Ginsburg – who later became a US Supreme Court justice – fought a line of cases that attacked gender differences in the law, whether they served to harm women or men (Chamallas, 2003). On her understanding of equality, equality under the law required treating like persons alike and unlike persons alike.

The campaign for neutrality in legal standards ran aground when it reached situations where women and men seemed not to be similarly situated, like in scenarios involving pregnancy or physical strength. Some judges thought that pregnancy might be a 'disability'; others argued it was sex-specific and deserved 'special' treatment (Littleton, 1987). Courts, legislatures, and activists attempted to define a 'normal' standard against which 'difference' could be identified (Scales, 2006). The trap lay in framing equality as either sameness or difference and taking men as the standard referent for such difference.

A critique of this binary approach led to a redefinition of equality beyond purely formal terms. Debates about substantive equality reintroduced Marxist questions of the distribution of material resources. From within a liberal rights perspective, courts worried about the limits of a state's responsibilities to provide minimal guarantees or standards of living, with courts frequently deferring to legislatures on questions about redistribution of wealth (Cossman and Fudge, 2002).

Theorists reframed the equality debates as a debate about the meaning of justice. Building on the work of liberal philosopher John Rawls, some feminists argued that if justice is about distribution, then injustice is about the imperfect distribution of material resources, rights and liberties, and opportunities. Communitarians argued that just distribution should include attention to culture and community. Martha Nussbaum advanced a capabilities approach, focused on identifying the minimum conditions that states must provide to ensure human flourishing (Nussbaum, 2000). Carol Gilligan, a leading thinker of difference feminism, introduced the distinction between an ethic of care, in which moral judgment is based on context, relationships, care, and connection, and an ethic of justice, in which judgment

proceeds from abstract principles, rights, and individual autonomy (Gilligan, 1982). Though Gilligan's model has been criticized because it associated men with an ethic of justice and women with an ethic of care, it opened up a field of inquiry about feminist understandings of justice.

Nancy Fraser observed that these opposing views all depart from the 'justice as distribution' paradigm, as neither theory of justice emphasizes transformation or makes an effort to "restructure the underlying generative framework" (Fraser, 1997: p. 23). She argued for a transformative understanding of justice, so that the processes that produce injustice, rather than their effects, can be analyzed and tackled.

Difference

'Difference' signals the categories and lived experiences that are raised, interpreted, and used to produce divisions among people. Difference has been a central challenge throughout feminist legal theory.

In the history of feminism and feminist legal theory, race has arguably been the oldest, thorniest difference. Since the late nineteenth century, women of color have questioned feminism's prioritization of the concerns of white, upper-class women. Early feminism's focus on formal equality in public life and the workplace was criticized for its inadequate attention to the conditions of material inequality experienced by poor and racialized women. Feminists of the 1980s were criticized for essentializing the category 'woman,' whether they preferred ending patriarchal domination or celebrating women's unique gifts. Both law and feminist legal theory had failed to theorize women of color's experiences of discrimination along multiple axes of identity. Kimberlé Crenshaw introduced the concept of intersectionality to challenge the default whiteness of feminist analysis (Crenshaw, 1991). Legal tests often require categorizing harms based on 'race' or 'gender,' with the result that some plaintiffs fall between categories, while others receive no legal recognition of the compounding interactions among different forms of discrimination (Williams, 1987).

Could there be a feminist movement if the category of woman was infinitely fractured along plural lines of identity? This crisis within feminism took place during the apogee of identity politics and the emphasis on multicultural tolerance. Political theorist Susan Moller Okin suggested that multiculturalism was "bad for women" (Okin et al., 1999). Wendy Brown (1995) and Leti Volpp (2001) cautioned against a feminism that depicted women as passive victims of intrinsically oppressive 'culture.'

The synergies between feminism and the global war on terror positioned culture, gender, and sexuality as productive sites of state power. The US nonprofit Feminist Majority Foundation campaigned against the Taliban's treatment of Afghan women, launching what became a simple, media-friendly rallying cry for a deeply complex US war in Afghanistan (Hirschkind and Mahmood, 2002). In 2010, a public controversy erupted when Gita Sahgal, the head of Amnesty International's Gender, Sexuality, and Identity unit, resigned in response to the organization's partnership with Moazzam Begg, a former Guantanamo detainee and UK citizen with ties to the Taliban. Some read Amnesty's collaboration as a welcome attempt to repair relations with a demonized Muslim population, while others, including Sahgal, interpreted it as a threat to Amnesty's core values of equality and universal human rights (Gopal, 2013).

Legislators and judges in the West wrestled with the state's interface with a nonsecular 'other,' on issues ranging from Muslim veiling to the incorporation of religious law into family law disputes (Fishbayn, 2008). Some feminists are deeply troubled by 'accommodation' and integration of the so-called cultural norms in state law, as this is interpreted as ceding ground to rising religious fundamentalisms – both Muslim and Christian. Other feminist legal theorists consider this a welcome step in exposing the white, secular, patriarchal bias of Western law (Fournier and See, 2012). While many feminist critiques focused on race, gender, and religion, other voices called attention to the heteronormative underpinnings of these debates. For example, queer theorist Jasbir Puar studied the connections between the West's war on terrorism, nationalism, and gay rights discourse. Puar argued that the West contrasts its tolerance of homosexuality and its concern for the vulnerable Muslim queer to a demonized Middle Eastern masculinity, thus justifying Western Islamophobia and fueling a belligerent homonationalism (Puar, 2007).

Law

Feminist legal theory often aligns with the canonical categories it challenges, resulting in feminist critiques of contract, criminal law, tort law, tax law, and so forth (Bottomley, 1996; Brooks, 2011). An alternative approach is problem-based, producing bodies of scholarship and activism around categories like reproductive justice, sexual and gender-based violence, workers' rights, and the criminalization of poverty and migration (Bridgeman and Millns, 1998; Schneider, 2000; Conaghan and Rittich, 2005). These efforts map the ways that law intersects with women's lives, fails to respond to women's realities, shapes their opportunities, and constructs their sense of self. 'Law in action' studies inquire about the law's simultaneous role in constructing disadvantages, remedies, identities, and understandings of justice, as in Catherine Albiston's research on the effects of the US Family and Medical Leave Act in changing the behavior and values of employers and workers (Albiston, 2010). A canonical example of the feminist analysis of legal doctrine is Catharine MacKinnon's study of rape law, which shows how rape law starts from a male point of view (MacKinnon, 1983). A further approach connects feminist theory to philosophical debates about rights. Frances Olsen, for instance, warns against a turn to rights for the purpose of women's emancipation. She argues that the resolution of conflict in rights terms can oversimplify complex power relations, deprive conflicts of their political significance and force, and fail to take into account structural power inequalities (Olsen, 1984). Through this engagement with law, feminist legal theory has unseated the neutrality of law's foundations.

Feminist legal theorists take a range of views about what law 'is' and what prospects it offers for achieving the goals of feminist reformers. Liberal feminists, most concerned with formal inequalities, define the law as residing in cases and statutes and implemented by lawyers, judges, and legislators. Departing from the view that the law is fundamentally

grounded in fairness, proponents like Ruth Bader Ginsburg believe that the law could be reformed to better respect 'women's' rights.

Radical feminists, attentive to power relations, are less sanguine about the nature of law and its potential for positive social change. They find the law in the inert words of statutes and cases, but also in norms, habits, and entrenched attitudes. Typified by Catharine MacKinnon's work (MacKinnon, 1983), these scholars see the law as embedded in patriarchal, stratified social relations. The law is susceptible to reform through feminist litigation and campaigning, but it is highly resistant to change. For example, in her work on domestic violence, Reva Siegel has shown how law is more liable to continuity through transformation than to genuine reform (Siegel, 1997).

Postmodern theorists adopted a multidimensional view. For example, for Carol Smart, although the law is defined as a socially powerful place for producing knowledge claims about justice and distribution, it is not universally or intrinsically helpful or harmful to the subjects who seek its remedies (Smart, 1989). Queer critics like Janet Halley raise concerns about the law as an institution premised on defining and separating the normal from the deviant, thus casting doubt on the role of law in emancipatory political projects (Halley, 2006).

State

Feminist legal theorists disagree about the nature and role of the state in emancipatory political projects. For liberal legal theorists, the state appears variously as a site for protection, a guarantor of individual rights and liberties, an agent of limited redistribution, a legitimate source of punishment, and an arena for fair debate. On this view, the state's laws are open to change through legislative or judicial processes (Davies, 1994).

Difference feminist theorists are skeptical about reforming the neutral state. Since women and men are fundamentally different, with different starting points in life and different needs, reform efforts are best focused on remedies like affirmative action, quotas, maternity leave, and public child care (Nott, 2000; Gilligan, 1982; Davies, 1994). These differences aside, both liberal and difference theory envisioned a strong state at the helm of justice projects.

These understandings of the state were challenged through the influence of Marxism and postmodernism. It became less fashionable to see the state as either a positive or negative space for feminist engagement.

Radical and postmodern feminist theory draws on these insights. According to radical feminist theorists and their readings of Marx, the very idea of a line between the state, the market, and the family manufactures the zones of state noninterference – in contractual relations and family relations – that expose people to patriarchal and capitalist violence (Olsen, 1985; Rose, 1987). Postmodern feminism, influenced by philosopher Michel Foucault, describes the state as a site for the production of discourse, with normalizing and disciplinary effects caused by self-governance rather than overt coercion or domination (Boyd, 1997). From this perspective, every choice by the state is a form of regulation; even a decision to 'do nothing' is a regulatory decision, because the practical consequence of 'nonregulation' is a normalization of existing distributions of power and resources.

More recently, the analytical lens has broadened, the state has been decentered, and a plurality of sometimes contradictory, context-driven views have emerged. Many contemporary theorists treat the nation-state as one of many spaces in diffuse local and global assemblages of power. Janet Halley's concept of governance feminism describes the installation of feminists and feminist ideas within institutional and legal power (Halley, 2006). Those fighting to end gender-based violence in intimate relationships push for a stronger police force, swifter prosecution, and mandatory sentencing (Kelly, 2005). Other theorists advocate for considerably 'less' state – for example, those concerned with the criminalization of poverty, swelling prison populations, violent working conditions of sex workers, and life chances of trans people (Spade, 2011; Bumiller, 2008).

Freedom

Complex debates over the notion of freedom course through feminist legal theory, bringing in their wake questions about agency, power, choice, and the nature of the subject. As Marx discussed in his classic essay "On the Jewish Question," the state's abolition of formal distinctions based on birth and the decree of individual equality leaves untouched all the material, cultural, and social structures of domination. Liberalism's vision of freedom is synonymous with formal equality rights. Equating freedom with equality rested on the introduction of the abstract, liberal, self-sufficient universal subject (Fineman, 2004).

This liberal subject is the object of concerted study across feminist legal theory. Liberalism's canonical accounts start with a natural, unified, and autonomous (implicitly or explicitly male) individual who yearns for a social world in which he can pursue his self-interest. Feminists disagree with this model, arguing that the self is neither a preconstituted individual nor wholly determined by membership in a social group. Rather, the subject is socially constructed and constituted relationally, through involvement with – and incorporation of – significant others and integration into communities (Young, 1990).

Discarding the idea of any essential or natural identity, Judith Butler argues that "woman itself is a term in process … an ongoing discursive process" (Butler, 1990: p. 33). Gender identity is performed, as opposed to intrinsic or natural, and "genders can neither be true nor false, but are only produced as the truth effects of a discourse of primary and stable identity" (Butler, 1990: p. 136). On this account, the body does not preexist as a biological 'real,' but is itself materialized through discourse.

Considering the subject as the effect of ongoing social processes and discourses muddies the waters of freedom and agency. How do people know what they want if there is no core, autonomous self who directs decisions, their available choices are constrained by material realities, and their wants and desires are shaped by social conditions?

Commercial surrogacy contracts offer an example of such dilemmas. It was the absence of choice that led radical feminist Andrea Dworkin to compare surrogacy to prostitution (Dworkin, 1983), while Christine Overall questioned whether the choice to enter a surrogacy contract could be a free one

(Overall, 1987). On the other hand, some feminists argued that commercial surrogacy arrangements were in fact the manifestation of a woman's right to make choices about her body and enter into contracts for financial gain (Shalev, 1989). An empirical survey of existing research found that women of color are greatly underrepresented as surrogate mothers (Busby and Vun, 2010). Such empirical findings bolster the argument that respecting women's agency requires attending to their own accounts of their lives and capacities. On issues ranging from surrogates to sex workers to women in military academies, courts, legislators, activists, and theorists have struggled to define women as victims or agents and have debated whether policies should reduce or expand women's field of choices.

Sex and Sexuality

The women's movement spent the 1960s campaigning for reproductive rights like access to abortion and contraception, and the 1970s lesbian feminist movement brought greater attention to women's rights to sexual pleasure with other women. Adrienne Rich denounced feminism for assuming a standard norm of female heterosexuality (Rich, 1980). By the late 1970s, feminists began to emphasize the violence and danger associated with heterosexual sexuality and the late 1970s and early 1980s saw feminist legal theory gripped by the 'sex wars.'

Debates divided many feminists into two opposing positions, often labeled as radical and sex positive. According to the radical feminist point of view, under conditions of patriarchy, sex and sexuality intrinsically perpetuated violence against women. According to a sex-positive point of view, sex and sexuality were sites of personal pleasure and political possibilities for subverting patriarchal definitions of femininity.

The 'sex wars' of the 1980s took shape around an emerging antipornography movement, underpinned by radical feminism and its theory of gender oppression. Radical feminists, such as MacKinnon and Dworkin, campaigned for the use of censorship and other legal prohibitions in order to outlaw pornography and reduce sexual violence against women. Pornography was seen to shape the cultural devaluation of women, define women's sexuality, and eroticize violent relations of domination. Sex-positive feminists reacted to this legal campaigning by decrying sexual repression, calling for greater sexual freedom, and defining sexuality as a site of political resistance.

Sexual practices and the meaning of consent were focal points of disagreement. For radical feminists, pornography, prostitution, sadomasochism, and sexual role-play like butch–femme relationships amounted to the repetitive entrenchment of male dominance. For example, MacKinnon argued that heterosexuality "institutionalizes male sexual dominance and female sexual submission" (1989: p. 113). In contrast, sex-positive feminists supported consensual sexual practices and they criticized radical feminists for repressing sexual desire and stigmatizing sexual minorities. Carlin Meyer argued for a brand of sex-positive feminism that could theorize sexuality as a space for autonomous and liberating choices (Meyer, 1994), while Robin West called for theories of sex as valuable to women's subjective well-being (West, 1987).

Questions about consent also underpinned these debates. For radical feminists, there could be no authentic consent or agency in a world of choices defined by male dominance and violence. In contrast, sex-positive feminists viewed women as capable of consenting and exerting agency and they criticized radical feminists for defining women's identities as effects of male domination (Abrams, 1995). However, this work was slow to percolate into legal theory. Katherine Franke commented that feminist legal theorists, "have done a more than adequate job of theorizing the right to say *no*, but we have left to others the task of understanding what it might mean to say *yes*" (Franke, 2001: p. 181).

The heated debates about sex and sexuality in the 1980s generated both divisions and new political formations. Gayle Rubin argued that feminism had little to say about sexuality, notwithstanding its expertise on gender-based subordination (Rubin, 1992). Eve Kosofsky Sedgwick said that gender and sexuality were distinct analytic axes – and distinct intellectual arenas (Sedgwick, 1990). Their argument for a new intellectual terrain, beyond the sex/gender distinction, gave birth to the field of queer studies.

Through engagement with queer theory, feminist legal theory has moved toward an increasingly antiessentialist understanding of sex and gender and a more diverse account of sexuality. Theorists, including Judith Butler, opined that people were born with a biological sex, but acquired a self-conception of gender through processes of socialization (Butler, 1990). This insight severed the 'natural' or 'inherent' from the socially contingent, so that the socially produced could be legally and politically challenged. Queer theorists argued that sexuality was also socially produced and culturally constructed. Queer legal theorists challenged the law's reliance on a male/female binary. They argued that a multiplicity of sexes, genders, and sexualities exist along a spectrum. In recent years, some trans legal advocates have celebrated the legal permission to move across the male/female divide, while others argue for a world that accepts the 'in-between' reality of the trans-gendered, intersexed, bigendered, and androgynous (Cowan, 2013).

While sexual minorities and dissidents achieved some equality gains in the late twentieth and early twenty-first centuries, the story of progress nevertheless rests on an economy of normal and deviant, included and excluded. Campaigners applauded the legal recognition of same-sex marriage as a victory for formal legal equality, but others saw the same-sex marriage project as yet another instance of the yearning for recognition by a state, and with it, the attribution to the state of power to police everyone's sexualities (Spade, 2011).

Marriage and Social Reproduction

Marriage and social reproduction have been sites for the formation and consolidation of state authority. The law's construction of the private sphere as an autonomous realm served to uphold and generate gender subordination. Key themes have included the rights to bodily autonomy, choice, privacy, and the public/private divide.

Under the common law doctrine of coverture, a married woman was located under the authority and protection of her husband, preventing her from making contracts or wills, controlling her earnings, or owning real property. In the nineteenth century, in the United States, the state regulation of

marriage permitted the definition of a private sphere, shielded from state intervention. As informal, *de facto* marriages were brought into the formal institution of marriage, people were legally bound in defined obligations, thus widening the scope of state authority. Incorporating more relationships – including the nonnuclear families of freed slaves – into the institution of marriage ensured that dependents would be the responsibility of husbands, rather than the state. A gender ideology of 'separate spheres' emerged, in which each sex had different natures with unique and complementary moral attributes. Post-Civil War married women's earning acts gave a married woman property rights in her own labor, while simultaneously defining household labor as an obligation of marriage, rather than a form of work (Cott, 2000).

Feminist legal theorists challenged the public/private divide enshrined in the regimes of coverture and separate spheres ideology. The US feminists leveraged the constitutional right to privacy as a route to contraceptive sales and abortion, but they also sought to pierce the shield of privacy around the marital home to compel state intervention in domestic violence.

Feminist legal theory challenged the role of law in creating and upholding the heterosexual family as the setting for the distribution and fulfillment of natural and inevitable dependencies. Activists in the 1970s demanded wages for women's domestic work and pointed out that 'free' care work in the home creates, sustains, and subsidizes the labor force. According to Martha Albertson Fineman, "the sexual family is considered the 'natural' form for the social and cultural organization of intimacy, its form ordained by divine prescription and perpetuated by opinion polls" (Fineman, 1995: p. 143). Approximately, 70% of families in 1940 fit the traditional male breadwinner–female housewife model; by 1988, however, the so-called traditional families accounted for only about 20% of all families and today there is no longer a typical or dominant family form (Albiston, 2007). Nevertheless, the ideology of the nuclear family defines both normal and deviant configurations of social reproduction. A vibrant literature has emerged in recent years, studying the globally distributed care economy and the resulting alliances and conflicts of interest between and among women in the global North and global South (Williams, 2000; Johnson, 2002; Conaghan and Rittich, 2005).

Feminist Legal Theory and International Human Rights

Feminist critiques of international human rights can be divided into three categories: (1) normative and institutional calls for the inclusion of women into human rights, (2) study of the role of gender in the architecture of the human rights system, and (3) critique about the perils and pitfalls of human rights.

The statement that "Women's rights are human rights," from the 1993 Vienna World Conference on Human Rights, confirmed the migration of women's interests from the isolated backwater of 'special issues' into the mainstream of human rights (Friedman, 1995). Fields of feminist activism and critique include development, transitional justice, international humanitarian law, peace and conflict, violence against women, reproductive and sexuality rights, refugees and migration, and international economic policy (Askin and Koenig, 1999). Christine Chinkin and Hilary Charlesworth observe that while there has been much feminist critique of the use of rights at a domestic level, there has been relatively little feminist critique of the normative and tactical value of human rights at an international level (Charlesworth and Chinkin, 2000). Their explanation is that feminists have been hesitant to critique what is so new and so hard-won, echoing Patricia Williams' comment that, from the African-American perspective, "rights feel so new in our mouths" that they should not be impugned (Williams, 1987).

The first generation of feminist critique of human rights sought to explain why women were excluded from the protection of human rights and what had to be done to remedy this. One explanation was that there were no 'general' human rights; by this reckoning human rights are, in themselves, men's rights (Charlesworth, 1995). State sovereignty has been theorized as an extension of masculinity in the international sphere (Ruskola, 2010), while the primacy given to civil and political rights over economic, social, and cultural rights has been explained as a source of women's exclusion from human rights protections (Sullivan, 1995).

Another body of empirical scholarship illuminates the ways that gender appears in the architecture of contemporary human rights and their antecedents. During the founding of the United Nations, founders were unwilling to grant rights to women specifically, but they saw gender as a metric of development. In the newly formed United Nations, the status of women in territories administered by member states was a key indicator of a country's capacity for self-governance (Knop, 2002).

A third body of critique invited caution about the feminist project in international law. In line with the "Third World Approaches to International Law" movement, this new literature exposes the complexities, distorting conditionalities, and unintended consequences of feminist gains in international law (Chinkin et al., 2005; Otto, 2006). First, scholars note that the feminist human rights movement imagines the international as an imaginary and utopian site of justice. International elite human rights advocates (from both the North and South) construct the particular as the problem in order for the universal to be the solution, while simultaneously avoiding the injustice born within the international economic and political system (Nesiah, 1993).

Second, theorists study the effects of injecting feminist principles into international institutions. Examples of governance feminism in international law include the recognition of sexual violence crimes in international criminal law and a focus on the criminalization of sex trafficking (Halley et al., 2006). Although feminists have achieved goals like gender mainstreaming in international law, the institutionalization of feminism in international law has adverse effects.

Third, the ascent of women's rights, primarily driven by campaigning on violence against women, was conjoined with the demonization of culture as an intrinsically oppressive force. Ratna Kapur has criticized the ways that international law rests on an essentialized understanding of gender and culture, which combine to produce a 'helpless' native victim of human rights law (Kapur, 2002). Sally Engle Merry has shown how UN rights

processes adopt a static definition of culture, which is then counterposed to make universal claims for gender equality (Merry, 2006). For Leslye Obiora, the international and First World obsession with culture obscures the primary cause of women's rights violations: global inequality and structural economic exploitation (Obiora, 1997). From such a perspective, human rights and feminism might well be Trojan horses. They may bring some gains for women's equality and well-being, but other, new harms, like the demonization of culture and disregard for global material inequalities, might creep in as well.

Conclusion

Recent surveys about the state of the field in feminist legal theory have produced numerous edited volumes – signaling the field's depth and diversity – while at the same time sparking reflection about the dispersal of the field, due to productive recentering around key words like sexuality or masculinity or the mainstreaming of its key analytic concerns across mainstream legal scholarship (Abrams, 2011). Discussions about the health of the field reveal the character of feminist legal theory as a movement that provokes deep personal investments, rather than as an arms-length academic exercise. Scholars who are seen to have harmed the feminist legal project, as though it were a mortal being, risk extreme vitriol. Janet Halley's (2006) suggestion to "take a break from feminism" provoked such a reaction (Halley, 2006). Halley defined legal feminism as reliant on a gender polarity between women and men, a subordination of women to men, and a commitment to acting for women. This account, according to Halley, both diminishes feminist victories and glosses over the casualties on the way to power. For some, Halley's invitation was tantamount to apostasy. Although feminist legal theory has matured, diversified, and become more plural, there remains the belief that there is a battle to be fought, and some speech and action count as crossing enemy lines.

Critique today seems to require moving across the confines of jurisdictions – both national and intellectual – and abandoning monolithic theories of the state and single-method studies. As a field that has, from its inception, convened a rebellious rabble of activist and theoretical voices, feminist legal theory is experienced in crossing boundaries. One such boundary separates the humanist and the empirical. Political theorists, rather than feminist legal theorists, led the way in debates about the meaning of justice. This is surprising given legal theory's penchant for normative questions, but less surprising given that feminist legal scholars have devoted their energies to empirical investigation into the 'impact' of law on women and analysis aimed at winning legal victories by exposing the 'true' gendered nature of law. According to Marianne Constable, questions about the meaning of justice may be occluded by accepting the model of a world that 'is' and a law that 'does' things and 'tells' people what to do (Constable, 1994). The next chapter of feminist legal theory could attend to life – in all its pluralities – as it ought to be or could become.

See also: Critical Legal Studies; Critical Race Theory; Domestic Violence: Legal Perspectives; Gender and the Law; Human Rights: Legal Aspects; Law: The Socio-Legal Perspective; Legal Realism; Sexual Harassment: United States and Beyond.

Bibliography

Abrams, Kathryn, 2011. Introduction: the distinctive energies of 'normal science'. In: Abrams, Kathryn (Ed.), Legal Feminism Now, Issues in Legal Scholarship, vol. 9. Issue 2, Article 1.
Abrams, Kathryn, 1995. Sex wars redux: agency and coercion in feminist legal theory. Columbia Law Review 95 (2), 304–376. Columbia Law Review Association, Inc.
Albiston, Catherine T., 2007. Institutional perspectives on law, work, & family. Annual Review of Law & Social Science 3, 397.
Albiston, Catherine R., 2010. Institutional Inequality and the Mobilization of the Family and Medical Leave Act. Cambridge University Press, Cambridge.
Askin, Kelly D., Koenig, Dorean M., 1999. Women and International Human Rights Law. Transnational, Ardsley, NY.
Barad, Karen M., 2007. Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning.
Bartlett, Katharine T., 1990. Feminist legal methods. Harvard Law Review 103, 829.
Bottomley, Anne (Ed.), 1996. Feminist Perspectives on the Foundational Subjects of Law. Cavendish Publishing Ltd, London.
Boyd, Susan B., 1997. Challenging the public/private divide: an overview. In: Boyd, Susan B. (Ed.), Challenging the Public/Private Divide: Feminism, Law, and Public Policy. University of Toronto Press, Toronto, p. 3.
Bridgeman, Jo, Millns, Susan, 1998. Feminist Perspectives on Law: Law's Engagement with the Female Body. Sweet & Maxwell, London.
Brooks, Kim (Ed.), 2011. Challenging Gender Inequality in Tax Policy Making: Comparative Perspectives. Hart Pub., Oxford.
Brown, Wendy, 1995. States of Injury: Power and Freedom in Late Modernity. Princeton University Press, Princeton, NJ, 52–76.
Bumiller, Kristin, 2008. In an Abusive State: How Neoliberalism Appropriated the Feminist Movement against Sexual Violence. Duke University Press, Durham.
Busby, Karen, Vun, Delaney, 2010. Revisiting the Handmaid's tale: feminist theory meets empirical research on surrogate motherhood. Canadian Journal of Family Law 26, 13.
Butler, Judith, 1990. Gender Trouble: Feminism and the Subversion of Identity. Routledge, New York.
Chamallas, Martha, 2003. Introduction to Feminist Legal Theory. Aspen Publishers, New York.
Charlesworth, Hilary, Chinkin, Christine, 2000. The Boundaries of International Law: A Feminist Analysis. Manchester University Press, Manchester.
Charlesworth, Hilary, 1995. Human rights as men's rights. In: Peters, Wolper, Peters, Wolper, Andrea (Eds.), Women's Rights Human Rights: International Feminist Perspectives. Routledge, New York, p. 103.
Chinkin, Christine, Charlesworth, Hilary, Wright, Shelley, 2005. Feminist approaches to international law: reflections from another century. In: Buss, Doris, Manji, Ambreena (Eds.), International Law: Modern Feminist Approaches. Hart Publishing, Oxford, p. 17.
Conaghan, Joanne, 2013. Feminism, law and materialism: reclaiming the 'tainted' realm. In: Davies, M., Munro, V. (Eds.), The Ashgate Research Companion to Feminist Legal Theory. Ashgate, pp. 31–50.
Conaghan, Joanne, Rittich, Kerry (Eds.), 2005. Labour Law, Work, and Family: Critical and Comparative Perspectives. Oxford University Press, Oxford.
Constable, Marianne, 1994. Genealogy and jurisprudence: nietzsche, nihilism, and the social scientification of law. Law & Social Inquiry 19, 551–590.
Cossman, Brenda, Fudge, Judy (Eds.), 2002. Privatization, Law, and the Challenge to Feminism. University of Toronto Press, Toronto.
Cowan, Sharon, 2013. 'What a long, strange trip it's been': feminist and queer travel with sex, gender and sexuality. In: Davies, M., Munro, V. (Eds.), The Ashgate Research Companion to Feminist Legal Theory.
Cott, Nancy, 2000. Public Vows: A History of Marriage and the Nation. Harvard University Press, Cambridge, MA.
Crenshaw, Kimberlee, 1991. Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics. In: Bartlett, Katharine T., Kennedy, Rosanne (Eds.), Feminist Legal Theory: Readings in Law and Gender. Westview Press, Boulder.
Davies, Margaret, 1994. 'Feminisms': Asking the Law Question. Law Books, Sydney.

Deckha, Maneesha, 2006. The salience of species difference for feminist theory. Hastings Women's Law Journal 17 (1), 14–19.
Dworkin, Andrea, 1983. Right-Wing Women. Perigee Books, New York.
Fineman, Martha, 2004. The Autonomy Myth: A Theory of Dependency. New Press, New York.
Fineman, Martha, 1995. The Neutered Mother, the Sexual Family, and Other Twentieth Century Tragedies. Routledge, New York.
Fishbayn, Lisa, January 2008. Gender, multiculturalism and dialogue: the case of Jewish divorce. Canadian Journal of Law and Jurisprudence XXI (1), 71–96.
Fournier, Pascale, See, Erica, 2012. The 'naked face' of secular exclusion: Bill-94 and the privatization of belief. Windsor Yearbook of Access to Justice 30 (1), 63–76.
Franke, Katherine M., 2001. *Theorizing yes*: an essay on feminism, law and desire. Columbia Law Review 101, 181.
Fraser, Nancy, 1997. Justice Interruptus: Critical Reflections on the "Post-socialist" Condition. Routledge, New York.
Friedman, Elisabeth, 1995. Women's human rights: the emergence of a movement. In: Peters, Julie, Wolper, Andrea (Eds.), Women's Rights Human Rights: International Feminist Perspectives. Routledge, New York, p. 17.
Gilligan, Carol, 1982. In a Different Voice. Harvard University Press, Cambridge, MA.
Gopal, Priyamvada, Spring 2013. Speaking with difficulty: feminism and antiracism in Britain after 9/11. Feminist Studies 39 (1), 98.
Halley, Janet E., 2006. Split Decisions How and Why to Take a Break from Feminism. Princeton University Press, Princeton, NJ.
Halley, Janet, Kotiswaran, Prabha, Shamir, Hila, Thomas, Chantal, 2006. From the international to the local in feminist legal responses to rape, prostitution/sex work, and sex trafficking: four studies in contemporary governance feminism. Harvard Journal of Law and Gender 29 (2), 335–423.
Hirschkind, Charles, Mahmood, Saba, 2002. Feminism, the Taliban, and politics of counter-insurgency. Anthropological Quarterly 75 (2), 339–354.
hooks, bell, 1989. Talking Back: Thinking Feminist, Thinking Black. South End Press, Boston, MA.
Johnson, Rebecca, 2002. Taxing Choices: The Intersection of Class, Gender, Parenthood, and the Law. UBC Press, Vancouver.
Kapur, Ratna, 2002. The tragedy of victimization rhetoric: resurrecting the 'native' subject in international/post-colonial feminist legal politics. Harvard Human Rights Journal 15, 1–37.
Kelly, Liz, 2005. Inside outsiders: mainstreaming violence against women into human rights discourse and practice. International Feminist Journal of Politics 7 (4).
Knop, Karen, 2002. Diversity and Self-determination in International Law. Cambridge University Press, Cambridge, UK.
Littleton, Christine A., 1987. Reconstructing sexual equality. California Law Review 75, 1279–1337.
MacKinnon, Catharine A., 1989. Toward a Feminist Theory of the State. Harvard University Press, Cambridge, MA.
MacKinnon, Catharine, 1983. Feminism, Marxism, method and the state – toward feminist jurisprudence. Signs: Journal of Women in Culture and Society 8 (4), 635–658.
Merry, Sally E., 2006. Human Rights and Gender Violence: Translating International Law into Local Justice. University of Chicago Press, Chicago.
Meyer, Carlin, 1994. Sex, sin, and women's liberation: against porn suppression. Texas Law Review 72, 1097.
Minow, Martha, 1992. Stripped down like a runner or enriched by experience: bias and impartiality of judges and jurors. William & Mary Law Review 33, 1201.
Morgensen, Scott L., 2012. Introduction: theorizing gender, sexuality, and settler colonialism. In: Erai, Michelle, Morgensen, Scott L. (Eds.), Karangatai: Calling Out Gender and Sexuality in Settler Societies. Special Issue: Settler Colonial Studies 2 (2).
Nesiah, Vasuki, 1993. Towards a feminist internationality: a critique of us feminist legal scholarship. Harvard Women's Law Journal 16, 189.
Nott, Sue, 2000. Accentuating the positive: alternative strategies for promoting gender equality. In: Beveridge, Fiona, Nott, Sue, Stephen, Kylie (Eds.), Making Women Count: Integrating Gender into Law and Policy-Making. Ashgate, Dartmouth, UK, p. 277.
Nussbaum, Martha, 2000. Women and Human Development: The Capabilities Approach. Cambridge University Press, Cambridge.
Obiora, L. Amede, 1997. Feminism, globalization, and culture: after Beijing. Global Legal Studies Journal 4, 355–406.
Okin, Susan Moller, Cohen, Joshua, Howard, Matthew, Nussbaum, Martha Craven, 1999. Is Multiculturalism Bad for Women? Princeton University Press, Princeton, NJ.
Olsen, Frances, 1985. The myth of state intervention in the family. University of Michigan Journal of Law Reform 18, 835.
Olsen, Frances, 1984. Statutory rape: a feminist critique of rights analysis. Texas Law Review 63, 387.
Otto, Dianne, 2006. Lost in translation: re-scripting the sexed subjects of international human rights law. In: Orford, A. (Ed.), International Law and its Others. Cambridge University Press, Cambridge, pp. 318–356.
Overall, Christine, 1987. Ethics and Human Reproduction: A Feminist Analysis. Allen & Unwin, Boston.
Puar, Jasbir K., 2007. Terrorist Assemblages: Homonationalism in Queer Times. Duke University Press, Durham.
Rich, Adrienne, 1980. Compulsory heterosexuality and lesbian existence. Signs: Journal of Women in Culture and Society 5 (4), 631.
Rochette, Annie, 2002. Stop the rape of the world: an ecofeminist critique of sustainable development. University of New Brunswick Law Journal 51, 145.
Romero, Adam P., 2009. Methodological descriptions: "feminist" and "queer" legal theories. In: Fineman, M., Jackson, J.E., Romero, A.P. (Eds.), Feminist and Queer Legal Theory: Intimate Encounters, Uncomfortable Conversations. Ashgate, Farnham, UK.
Rose, Nikolas, 1987. Beyond the public/private division: law, power and the family. Journal of Law and Society 14 (1), 61.
Rubin, Gayle, 1992. Thinking sex: notes for a radical theory of the politics of sexuality. In: Vance, Carole S. (Ed.), Pleasure and Danger: Exploring Female Sexuality. Pandora, London, pp. 267–293.
Ruskola, Teemu, 2010. Raping like a state. UCLA Law Review 57, 1477.
Scales, A., 2006. Legal Feminism: Activism, Lawyering, and Legal Theory. New York University Press, New York.
Schneider, Elizabeth M., 2000. Battered Women & Feminist Lawmaking. Yale University Press, New Haven.
Sedgwick, E.K., 1990. Epistemology of the Closet. University of California Press, Berkeley.
Sheppard, Colleen, Westphal, Sarah, 2000. Narratives, law and the relational context: exploring stories of violence in young women's lives. Wisconsin Women's Law Journal 15, 335.
Shalev, Carmel, 1989. Birth Power: The Case for Surrogacy. Yale University Press, New Haven.
Siegel, Reva, 1997. Why equal protection no longer protects: the evolving form of status-enforcing state action. Stanford Law Review 49, 1111–1113.
Smart, Carol, 1989. Feminism and the Power of Law. Routledge, London.
Spade, D., 2011. Normal life: Administrative Violence, Critical Trans Politics, and the Limits of Law. South End Press, Brooklyn, NY.
Sullivan, Donna, 1995. The public/private distinction in international human rights law. In: Peters, Julie, Wolper, Andrea (Eds.), Women's Rights Human Rights: International Feminist Perspectives. Routledge, New York.
Valverde, Mariana, 2003. Law's Dream of a Common Knowledge. Princeton University Press, Princeton, NJ.
Volpp, Leti, 2001. Feminism versus multiculturalism. Columbia Law Review 101, 1181.
West, Robin L., 1987. The difference in women's hedonic lives: a phenomenological critique of feminist legal theory. Wisconsin Women's Law Journal 3 (81), 81.
Williams, Patricia J., 1987. Alchemical notes: reconstructing ideals from deconstructed rights. Harvard Civil Rights-Civil Liberties Law Review 22, 401.
Williams, Joan, 2000. Unbending Gender: Why Family and Work Conflict and what to Do about it. Oxford University Press, Oxford.
Young, Iris M., 1990. Justice and the Politics of Difference. Princeton University Press, Princeton.

Feminist Political Theory and Political Science

Joni Lovenduski, Birkbeck College, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Lovenduski, volume 8, pp. 5475–5478, © 2001, Elsevier Ltd., with an updated Bibliography supplied by the Editor.

Abstract

Feminist scholars have shown that few aspects of politics are unaffected by gender. The feminist study of politics is interdisciplinary and radical. There are three main dimensions to this critique of politics: feminist political theory which is predicated on a critique of contemporary liberalism, challenges the conventions of political theory; feminist political science, a critique of political institutions and the ways they are typically understood; and feminist accounts of gender and politics that combine explorations of gender as a category, as a hierarchy, and as performance in studies of how gender relations are implicated in power. In the study of politics, feminist use of the concept of gender involves redefinition of the split between public and private life and the re-examination of conventional political science for gender dimensions in order to construct a gender informed understanding of politics.

In popular discourse the term feminism is normally used to mean support for women's rights, while a feminist is someone whose central concern is with the position of women and their struggle for emancipation (Delmar, 1986). The definitions are contested by feminist scholars and subject to change. In the 1970s feminism was frequently defined as opposition to the dominance or subjugation of women. By the end of the 1990s it was impossible to define feminism without reference to the concept of gender and it was unusual to refer to feminism in the singular. Feminisms are therefore standpoints. They are also political practices (most often social movements), and sets of ideas and theories. To complete the definition it is necessary to add the concept of gender. Gender is a category that was developed to explore what counts as 'woman' and what counts as 'man' (Squires, 1999, p. 54). Thus, one useful definition of feminisms declares them to be oppositions to gender hierarchy, capturing the different situations and interests of different women in the pluralization of 'feminism,' while the social construction of masculinity and femininity is signaled by the term 'gender.'

Gender Studies and Feminist Scholarship

Challenging the approaches and preoccupations of mainstream scholarship has proved to be a continuing focus of feminist politics. Originally mainly concerned with 'women's studies,' the attention of feminist scholarship gradually shifted to the study of gender and especially to gender relations and the nature of masculinity and femininity. Feminist writings include critiques of most areas of contemporary scholarship calling into question their sensitivity to issues of gender. The shift had important implications for the social sciences including the study of politics. In the social sciences an important feminist concern is to establish the category of gender as a fundamental social division. The feminist critiques of political science are part of this tendency, addressing the standard themes of research on politics, questioning narrow definitions of the 'political,' and also drawing upon research and thinking about gender in other disciplines.

Feminist political science is therefore interdisciplinary and radical, offering a broad definition of the field. Feminist political studies take three main forms: (1) feminist political theory which is predicated on a critique of liberalism and challenges the conventions of political theory, (2) feminist political science which is a critique of political institutions and the ways they have been understood, and (3) feminist accounts of gender and politics which combine explorations of gender as a category, as a hierarchy, and as performance in studies of how gender relations are implicated in power. Each area is work in progress, complicated and contested. Similar patterns of development have characterized each of the three kinds of discussion.

Feminist Political Theory

The women's liberation movement of the 1960s and 1970s gave a new impetus to feminist politics. This was at first new social movement politics, suspicious of formal hierarchical organizations which were regarded not only as undemocratic but also as sexist. However, feminism was, and is, profoundly political. The early slogan 'the personal is political,' questioned the nature of power and challenged the conventional restriction of politics to the 'public sphere.' Initially, feminist political theory was a matter of unmasking sexist assumptions and the neglect of women by Western Political Theory. Carol Pateman argued that the subjugation of women was a central underpinning of modern political thought and that gender difference made a public sphere possible at the expense of women. She contended that the distinction between state and civil society is only visible if the family is ignored. If civil society is labeled private then family is forgotten in political discussion (Pateman, 1988). Hence the addition of women to the concerns of political theory brought the central assumptions of liberalism into question, notably the separation between public and private life.

Recognizing that writing gender differences into political theory required a major reconsideration of political ideas, feminist theorists addressed issues of equality, justice and

representation, initiating debates that grew increasingly sophisticated. Originally theoretical arguments were built around binary oppositions or dualisms such as equality/difference, nature/nurture, justice/care, inclusion/exclusion, all thought to parallel male/female dichotomies. Arguments in the equality/difference debates considered whether men and women should be treated the same or whether differences should be taken into account in order to treat both fairly. The justice/care debates addressed a perceived difference between an ethic of justice (in the liberal distributive sense) and an ethic of care (in a maternalist sense), argued to be male and female, respectively.

Representation arguments considered whether women need women to represent them, reasoning that the identity of the speaker matters to the interpretation of what is said (Phillips, 1995; Young, 1990). Discussions of justice, equality, and representation tended to converge into arguments about citizenship in which feminists seek to redefine politics in terms that integrate public and private concerns, breaking through the public/private divide (Dietz, 1985; Mouffe, 1995; Lister, 1997). Since the 1980s a tendency to contest the oppositions has become more prevalent as theorists seek to transcend or displace the hierarchies, often by redefining issues and concepts (Squires, 1999, pp. 226–33).

Feminist Political Science

Although there were very few links with feminist political theory, feminist political science has followed a similar trajectory, developing from the inclusion of information about women to the subject matter of politics to an intellectual concern to establish a gendered discipline. Feminist political science first sought to eliminate the 'bad science' of exaggerating and misconstruing differences between women and men, assuming a male political universe and using data without their original accompanying qualifications. They were troubled by the distortions present in the mainstream of behavioral political science arising, they believed, from an epistemology that separated facts and value and privileged so-called value-free factual accounts. The effect of concern only with putative facts was to allow behavioral political science to factor women out of consideration as political subjects. An overemphasis on measurable political behavior prevented recognition of '… not only the … exclusion of women from what is traditionally political, but also the inclusion of politics in what women have traditionally done' (Nelson, 1989, quoted in Carroll and Zerilli, 1993). Families, communities, and voluntary groups all went largely unresearched as constitutive of political life. Inevitably, feminist political scientists sought to expand definitions of politics, lending strength to other challenges to overconcern with electoral behavior and elite participation, at the same time as examining differences in men's and women's political behavior. Discussion of the significance of the public/private split and the confinement of women to the private realm were an important challenge to narrow constructions of political power. So too were studies of the crucial part played by the state in regulating sexuality (see Carroll and Zerilli, 1993; Lovenduski, 1998).

Once it was necessary (and possible) to study the effects of women on previously male-dominated political institutions and organizations, the need for an adequately theorized notion of gender was accepted. As increasing numbers of women appeared in legislatures, councils, and other representative bodies as well as in executive organizations and public administrations of various kinds, a considerable debate developed about the causes, stability, and results of gender differences in political activity. An especially powerful line of argument for both feminist activists and their opponents was the contention that women should not seek equal presence to men in powerful positions because they would be 'incorporated' and changed by it, they would become unable to represent women because they had been turned by male created institutions into political men. The counterclaim was that once a significant number and diversity of women occupy decision making positions in previously 'masculine' institutions they would be transformed, their gendering altered.

This argument, so important in feminist debate, has echoes in the conventions of the study of political processes (Randall, 1991). Once arrangements, practices, and rules are institutionalized, actors learn them, invest in them, and make decisions accordingly. Particular institutions become associated with particular behavior and outputs that may be classified as masculine or feminine (Duerst-Lahti and Kelly, 1995). The practices that insulate the positions of typical members and are notoriously difficult to change are gendered.

Feminism and Gender Studies

Feminist intellectual concern about women and power generated research in the arts, humanities, and social sciences. In the fields of history, sociology, social psychology, literature, and cultural studies, a substantial body of work about different aspects of gender and power was produced. In addition a significant field of 'new' men's studies examined masculinity as a category both of privilege and oppression (Carver, 1996). Analysis of masculinities has also received considerable attention from feminist theorists who also claimed that differences among women were at least as important as differences between women and men, that race, class, age, ethnicity, sexuality, and physical ability were, like sex, constitutive of political identity and therefore of political behavior and power.

Feminism, Gender Studies, and Political Science

When feminist political scientists contested narrow definitions of politics, it was because they wanted to expand their research to aspects of 'private' life and to examine the role of government and the state in constructing and enforcing certain kinds of relationships between women and men. As it became more widely understood that both women and men have gender, debates across disciplinary boundaries became a common feature of feminist research on politics. Once perceived, the effects of gender cannot be left to one side or held constant for the sake of other experiments. As it has come to be defined, gender is embedded in individuals, relationships, institutions, and organizations. The gender debates are part of a continuing

process of definition refinement and theory production that enhances understanding of political institutions and processes. A 'gendered' political science looks 'not just at particular theories and assumptions as they have been applied to women, but as they are constructed and apply in any case' (Sapiro, 1991). Institutional research traditions in political science require the close description of contexts, norms, processes, and cultures to which a gendered vocabulary may be applied. It follows that the gendered interests present in political arrangements should be acknowledged and analyzed.

In summary, feminist scholars have shown that it is difficult to think of an aspect of politics that is unaffected by gender. However, gender is not a neat and tidy concept to be slotted into unchanging research strategies. Using the concept of gender involves rejection of the definitions of politics implied by the conventions of the public/private split, re-examination of the productions of conventional political science for gender dimensions and, on that foundation, the construction of a political science that is gender informed.

Editor's Note

For the author's further thoughts and updated references, please see Joni Levenduski, 2005.

See also: Ecofeminism; Femininities; Feminism in Anthropology; Feminisms and Acculturation around the Globe; Feminist Criminology and Gender Studies; Feminist Legal Theory; Feminist Theory: Radical Lesbian; Gender and Feminist Studies in Geography; Gender, Feminism, and Sexuality in Archaeological Studies.

Bibliography

Carroll, S.J., Zerilli, M.G., 1993. Feminist challenges to political science. In: Finifter, A.W. (Ed.), Political Science the State of the Discipline, II. American Political Science Association, Washington, DC.

Carver, T., 1996. Gender is not a Synonym for women. L. Reiner, Boulder, CO.

Dahlerup, D., 1988. From a small to a large minority: women in Scandinavian politics. Scandinavian Political Studies 11, 275–298.

Delmar, R., 1986. What is feminism? In: Mitchell, J., Oakley, A. (Eds.), What is Feminism? Blackwell, Oxford, UK, pp. 8–33.

Dietz, M., 1985. Citizenship with a feminist face: the problem of maternal thinking. Political Theory 13 (1).

Duerst-Lahti, G., Kelly, R.M. (Eds.), 1995. Gender Power, Leadership and Governance. University of Michigan Press, Ann Arbor, MI.

Levenduski, Joni, 2005. Feminizing Politics. Polity Press, London.

Lister, R., 1997. Citizenship: Feminist Perspectives. Macmillan, London.

Lovenduski, J., 1998. Gendering research in political science. Annual Review of Political Science 1, 333–356.

Mouffe, C., 1995. Democracy, power and the political. In: Benhabib, S. (Ed.), Democracy and Difference: Contesting the Boundaries of the Political. Princeton University Press, Princeton, NJ, pp. 245–257.

Pateman, C., 1988. The Sexual Contract. Polity Press, Cambridge, UK.

Phillips, A., 1995. The Politics of Presence. Clarendon Press, Oxford, UK.

Randall, V., 1991. Feminism and political analysis. Political Studies 39, 513–532.

Sapiro, V., 1991. Gender politics, gendered politics: the state of the field. In: Crotty, W. (Ed.), Political Science: Looking to the Future. Northwestern University Press, Evanston, IL, pp. 165–187.

Squires, J., 1999. Gender in Political Theory. Polity Press, Cambridge, UK.

Young, I.M., 1990. Justice and the Politics of Difference. Princeton University Press, Princeton, NJ.

Feminist Psychotherapy

Carolyn Zerbe Enns, Cornell College, Mount Vernon, IA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.A. Katzman, volume 8, pp. 5478–5480, © 2001, Elsevier Ltd.

Abstract

This article reviews the basic tenets of feminist psychotherapy, its history and core values, techniques, and approach to integrative psychotherapy. Originally an approach that focused primarily on the concerns of women, feminist psychotherapy has increasingly emphasized the needs of all people and their multiple identities related to race/ethnicity, gender, sexual orientation, nationality, culture, age, disability, and other social identities. Feminist psychotherapy is characterized as a living, evolving approach that emphasizes the importance of understanding individuals within their biopsychosocial, ecological, transnational contexts.

Introduction

Feminist psychotherapy has its roots in the civil rights and women's movements of the 1960s. Over the past 50 years, it has evolved substantially and has increasingly transformed psychological practice. For example, the American Psychological Association's (APA, 2007) *Guidelines for Psychological Practice With Girls and Women* reflects feminist values and reveals ways in which egalitarian and social change values have been incorporated within basic assumptions about optimal psychological practice. Rather than being a specific set of theories or techniques for working with clients, feminist therapy can be described as a superordinate structure or framework that orients a practitioner's worldview and assumptions about what matters in competent and empowering practice with individuals, especially those persons whose experiences have often been marginalized within the culture. Brown (1994) noted that "what makes practice feminist is not who the clients are or the specific techniques used, but how the therapist thinks about what she does" (p. 2).

A feminist psychotherapist draws from a range of nonsexist psychotherapy approaches that can be integrated with feminist values. Thus, feminist therapists are typically technically eclectic, drawing from a diverse range of theoretical orientations such as cognitive behavioral, humanistic, psychodynamic, and family systems approaches. For example, a feminist psychotherapist may describe her or his approach as a feminist multicultural narrative approach, a feminist existential approach, or a multicultural feminist cognitive therapist.

Feminist psychotherapists are informed about the social justice issues within their environments, and continue to refine the meaning of feminist psychotherapy and practice in a changing world. The work of feminist therapists is often integrated with diverse interdisciplinary perspectives (e.g., political science, sociology, philosophy). Feminist therapists define their roles as therapists, advocates, and social change agents who support the holistic empowerment of persons with whom they work.

What is Feminism and Feminist Therapy?

Feminism is defined in *Webster's New World Dictionary* (1978) as (1) the principle that women should have political, economic, and social rights equal to those of men; and (2) the movement to win such rights for women. While this is a relatively benign definition, the range of reactions and beliefs attached to the word feminism is vast and often emotionally laden, with some holding misconceptions that feminists typically reject femininity, are antimale, and define their goals in extreme or inflexible ways. These stereotypes are not consistent with the goals or definition of feminism. Feminist theorist Gloria Jean Watkins, whose chosen pen name is bell hooks (2000), clarified the breadth and depth of feminism when she defined it inclusively as a commitment to ending all forms of oppression that interact with sexism, including heterosexism, colonialism, ethnocentrism, white supremacy, ageism, racism, classism, ableism, and other forms of marginalization. The goal of feminism is to empower all individuals, including men, to create a world in which equality is experienced in individual, interpersonal, institutional, national, and transnational contexts.

The relevance of bell hooks' (Gloria Jean Watkins) definition can be seen in Laura Brown's description of feminist therapy, which appears on her Web site (http://www.drlaurabrown.com/feminist-therapy, see also Brown, 1994): "I define feminist therapy as the practice of therapy informed by feminist political philosophies and analysis, grounded in multicultural feminist scholarship on the psychology of women, men and gender, which leads both therapist and client toward strategies and solutions advancing feminist resistance, transformation and social change in daily personal life, and in relationships with the social, emotional and political environments."

Contemporary feminist therapists adopt an intersectional perspective, which means that in order to understand constructs such as sex, gender, and gender-related biases, it is necessary to understand how gender-related concepts interact with and are informed by other social categories such as race/ethnicity, sexual orientation, sexuality, nationality, culture, class, and religion. Intersectionality also implies that a given individual may experience some social identities that confer oppression (e.g., sexism or classism), and they may also experience some social identities that confer privilege (e.g., white and/or heterosexual privilege). The degree to which one experiences privilege or disadvantage is modified by social contexts, such that an individual may experience privilege in some

settings (e.g., as a middle-class white female leader within a group with other women) and oppression in others (e.g., as a token woman in a corporate environment). Thus, a woman may experience discrimination in the corporate work environment but may perpetrate racist or classist behaviors toward persons with less privileged social status. The experience of marginalization, oppression, privilege, and equality are complex and changeable, adding to the complexity of practicing psychotherapy that focuses on supporting the equality of all people. As a foundation for considering the complexity of power, privilege, and oppression, the integration of multicultural and feminist perspectives is especially valuable.

It is noteworthy that multiple approaches to feminism exist and are defined by terms such as liberal feminism, multicultural feminism, radical feminism, socialist feminism, postmodern feminism, critical race feminism, third wave feminism, and queer theory. Although gender equality is a shared foundation of each of these theories, there are differences among these theories with regard to their assumptions about how and why oppression exists, how inequality can be best eradicated, and how equality should be defined and achieved. Exploration of these diverse approaches helps increase the feminist therapist's self-awareness of values, and provides a foundation for developing a consistent and well-informed philosophical approach that is integrated effectively with compatible psychotherapy techniques.

The Origins and Evolution of Feminist Therapy

The civil rights and women's movements of the 1960s prompted a reconsideration of roles for women and men and other disenfranchised groups in western society. Consciousness raising groups provided a forum for women to make the link between their personal experience and the political and social context in which these experiences were constructed. One of the outcomes of consciousness raising groups was the emergence of feminist psychotherapists who sought to implement the empowering aspects of consciousness raising groups in their work with individuals who experienced problems in living. Over time, feminist therapists integrated the values and practices of consciousness raising groups with nonsexist and antiracist psychotherapy systems.

Early feminist therapists noted that most of the then available theories, research, and diagnostic approaches established the lives of privileged men as normative, resulting in women's experiences or perspectives being seen as deficient in comparison to those of men (Gilligan, 1982). Chesler (1972) characterized the typical client–therapist relationship as establishing the therapist (typically male) as an expert or patriarch, who defined the reality of the patient and often reinforced the unequal status between men and women in the larger culture. Other criticisms and concerns (Worell and Remer, 2003) raised by feminist therapists focused on:

- the general omission of women from the knowledge base of psychology,
- the failure to recognize that psychopathology may reflect power imbalances in addition to intrapsychic and interpersonal difficulties,
- the disregard of many health professionals for women's self-reported experience,
- the practice of attributing blame and responsibility to women for their experience of sexual and physical violence,
- the tendency to pathologize women's interdependence and interpersonal involvement, such as by blaming mothers or characterizing women as interpersonally enmeshed, and
- the practice of medicalizing women's psychological problems, including diagnosing their concerns in narrow terms or overprescribing medications.

As an outgrowth of their concerns, feminist therapists sought to redefine the psychotherapy in egalitarian ways and to emphasize their clients' strengths, resilience, and agency rather seeing them as individuals who merely needed to be cured (Gilbert, 1980). One of the outcomes of this redefinition was the gradual implementation of informed consent principles as standard aspects of all psychotherapy practice. Over time, psychological and social service practice has benefited from an explosion of research and theory about the lives and experiences of diverse women (Hyde, 2013), and this research has provided a strong research foundation for the work of feminist psychotherapists. Another transformative outcome of the work of feminist researchers and practitioners can be seen in the *Guidelines for Psychological Practice With Girls and Women* (APA, 2007; discussed later in this article), which are based on at least 40 years of research and are intended as guides for optimal psychological practice for all psychological practitioners, not only persons who identify themselves as feminist. It is important to note that whether contemporary psychotherapists define themselves as feminist or not, feminist therapy values and assumptions have permeated the practice of those who seek to provide exemplary services to women and men.

Core Assumptions and Principles of Feminist Therapy

Although they may vary in their application, feminist approaches are characterized by the following assumptions: (1) the personal is political; (2) equality and egalitarianism are central to the practice and goals of psychotherapy; and (3) knowledge of diversity, complexity, and intersectionality are necessary for working toward empowerment (Brown, 1994, 2010; Enns, 2004, 2012; Worell and Remer, 2003).

The personal is political (as well as the political is personal) reflects the notion that human concerns cannot be solved in isolation from gender and other identity politics, which often result in women's and other marginalized persons' lower social status and oppression in most societies. Individual concerns need to be understood holistically and within a biopsychosocial and ecological context that considers microsystems, such as the systems and groups with which one interacts on daily basis; exosystems, such as social, governmental, and political institutions and policies that affect one's options and degrees of freedom; and macrosystems, or cultural influences, values, and worldviews (Ballou et al., 2002).

Feminist therapists bring a questioning stance toward status quo understandings of emotional distress, seeking to understand how a client's problems and strengths are related to or

embedded in her or his social milieu. Problems are understood as emerging from efforts to cope with challenges of living. For example, a person who experiences disempowerment may have learned to rely on indirect methods of coping, such as dependency, acquiescence, or manipulation. Although coping efforts such as dependency or acquiescence may support survival in difficult contexts, the costs associated with such strategies are often high and may result in symptoms of anxiety and depression. When these troubling symptoms are reframed as the client's coping skills that have outlived their usefulness, rather than as personal deficiencies, clients are more likely to move beyond self-blame and work toward developing skills that support growth and enhancement.

The feminist therapist recognizes that when problems exist in complex contexts, a client's individual efforts alone will not be sufficient for supporting long-term growth. A client's involvement in group therapy or other self-help options can facilitate the recognition of links between the personal and political by allowing the client to see commonalities as well as differences in the patterns of concerns of group members. In addition, helping clients gain access to community resources that can address additional aspects of their concerns and social contexts (e.g., educational or legal institutions) is important. A willingness on the part of feminist therapists to engage in advocacy on behalf of clients may be crucial to helping individuals access services that can support their empowerment. Thus, investing efforts in the service of social change and social justice is an important extension of the notion that the personal is political. Collective action is often necessary to achieve the social and political change that underpins many stated 'mental' problems.

A second core assumption emphasizes the importance of egalitarianism as a foundation for the psychotherapy relationship and as a goal or outcome of counseling and psychotherapy. The feminist therapist seeks to counteract some of the effects of inequities, oppression, and power abuses present in contemporary society by modeling a relationship based on equality. As such, the feminist therapist works to be transparent, open, and aware of her or his values. Self-knowledge on the part of the therapist helps to ensure that she or he does not inadvertently project biases and stereotypes on the client, which might limit the client's possibilities for growth. The client is seen as the 'best expert' on her or his life, and the ability of the therapist to show respect for the client's expertise contributes to her or his ability to move toward wholeness and health.

To support the value of equality, feminist therapists seek to use straightforward, genuine, and jargon-free language, which then support the client's ability to collaborate in identifying and building on strengths and resilience. Through brief self-disclosure and clear communication, the therapist helps to create a climate that facilitates the client's ability to raise questions and develop skills of assertiveness, which serve as an antidote against disempowerment that she or he may have experienced in the past. The feminist therapist also recognizes that his or her training and expertise make it impossible for the psychotherapy relationship to be fully equal. Discussions between the therapist and client about roles as well as power and knowledge differences within the psychotherapy relationship contribute to the client's ability to deal effectively with power differences beyond the psychotherapy relationship. The goal of equality is also a foundation for informed or empowered consent, which helps to ensure that clients are knowledgeable about the counselor's values and approach to working with individuals, and thus, are able to make clear decisions about whether work with the feminist therapist is likely to be in her or his best interest.

In addition to its importance as a psychotherapy relationship value, and major goal of feminist therapy is to help clients understand power dynamics and power abuses in the real world and to assist them in working toward equality in various spheres of life. Power analysis and gender/social identity analysis, which will be discussed briefly in the section on techniques, become tools for helping clients understand their own positions vis-á-vis sources of power and privilege as well as their experiences of disadvantage or oppression. Feminist psychotherapists assist clients in considering how factors such as race, class, sexual orientation, gender, religion, and other social identities affect their access to power and equality, and work toward helping clients plan how to negotiate equitable and satisfying roles in a wide range of life spheres, such as work, interpersonal, family, educational, and leisure roles. Empowerment is a major component of the goal of equality, and may include self-esteem and self-valuing, positive daily functioning, gender and culture awareness, perceived self-efficacy and control, self-nurturing capacities, problem-solving skills, behavioral flexibility, assertiveness, the ability to access community resources, and participation in social activism (Worell, 2001). Consistent with these core values, research has found that one's self-identification as a feminist psychotherapist predicted greater emphasis on psychotherapy goals related to empowerment and the personal is political (Moradi et al., 2000).

A third major assumption related to feminist therapy is the theme of diversity and intersectionality. Understanding and valuing the diversity of women's and men's lives is central to competent practice. As members of complex multicultural societies, feminist therapists celebrate the myriad interactions among multiple social statuses and categories such as nationality, culture, gender, sexual orientation, race/ethnicity, class, age, and disability. Many social identities (e.g., race, class, gender) and related oppressions (e.g., racism, classism, sexism) operate simultaneously. Although women typically share some common goals and life experiences, gender is only one of many social identities that influence the life course and experiences of empowerment or oppression. Furthermore, gender may seem less salient to some women or men than other markers of identity, privilege, and power, such as race/ethnicity or sexual orientation (Shields, 2008). In general, the complex effects of gender in interaction with other social identities or categories cannot be adequately captured through statistical analysis of interactions. One facet of identity may modify another, making it impossible to tease out the unique effects of gender alone.

The concept of intersectionality reflects the fact that aspects of identity are interrelated and may modify each other or be fused with each other. Thus, for example, gender and race may be inseparable and may be best characterized as gendered racism (Essed, 2001). Alternatively, some aspects of identity may be salient, highly influential, and part of the foreground in some situations (e.g., race or gender), but may be of limited

relevance and only part of the background in other situations. Different social contexts may also elicit an individual's awareness of specific aspects of identity. Thus, an eye to complexity and variability, even when considering the experiences of a single person, is central to competent feminist multicultural practice (APA, 2007).

The Feminist Therapy Institute Code of Ethics (2000) underlines the relevance of diversity to ethical practice, noting that feminist therapists seek to be knowledgeable of their own cultural backgrounds, to broaden their knowledge of cultural diversities, and to monitor their interactions with diverse clientele, being careful to identify and correct personal attitudes and behaviors that may limit their effectiveness.

Feminist Psychotherapy Assessment and Techniques

This section identifies some of the unique techniques and approaches to assessment that characterize feminist therapy. Although feminist therapists rely on many of the tools and techniques associated with other psychotherapies (e.g., cognitive, humanistic, existential, narrative), feminist therapists have developed distinctive methods for helping individuals gain awareness of their multiple identities and place difficulties in context. By placing problems in context, clients are better prepared to direct efforts toward aspects of their lives that they can control or influence and are less likely to internalize negative self-beliefs about areas that they may have limited ability to change, especially in the absence of advocacy and change of institutional structures and cultural attitudes. Tools such as gender and social identity analysis, power analysis, reframing, and strengths analysis are especially important as part of assessment and intervention.

Feminist psychotherapists have been cautious about relying primarily on traditional diagnostic assessments such as the *Diagnostic and Statistical Manual* (American Psychiatric Association, 2013) for a number of reasons, including the fact that traditional diagnostic labels tend to emphasize intrapsychic factors, focus primarily on deficit behaviors, may contribute to the medicalization of problems, and may reinforce subtle gender and other cultural biases. Furthermore, an emphasis on intrapsychic symptoms may contribute to 'pigeonholing' misery rather than contributing to creative methods of addressing pain and supporting growth. Although feminist psychotherapists recognize the value of mainstream labels and assessments (e.g., depression, anxiety disorders) for supporting efficient communication and a common language, they often note the deficiencies of traditional diagnosis for gaining a social-contextual, ecological perspective on problems in living. As a result, gender and social identity analyses have become a hallmark of feminist therapy and facilitate an understanding of how social categories relevant to gender, race, and other social-contextual factors are related to an individual's problems. The assessment of social identities is designed to raise the consciousness of the client about how social roles and categories may be related to his or her current challenges as well as to facilitate collaborative ways of working toward solutions.

Two sets of acronyms are especially useful for organizing social identity and gender role analysis: the ADDRESSING framework (Hays, 2008) and RESPECTFUL assessment (D'Andrea and Daniels, 2001). The acronym ADDRESSING guides exploration of experiences related to age (A), developmental experiences (D), disability/ability (D), religion (R), ethnicity (E), social class (S), sexual orientation (S), indigenous heritage (I), national origin (N), and gender and sex (G). Similarly, the acronym RESPECTFUL can be used to explore religious beliefs (R), ethnic–cultural–racial background (E), sexual orientation and identity (S), psychological mindedness (P), economic and social class factors (E), chronological and developmental patterns (C), threats to health and well-being (T), family history and dynamics (F), unique physical characteristics (U), and location or geographical residence (L). In general, open-ended questions are useful for gathering information about these social categories. Clients are encouraged to reflect on what rules they learned from their cultures about specific identities (e.g., gender-related expectations) and to consider ways in which they experience empowerment, privilege, or disadvantage related to these facets of experience and identity. By clarifying some of the barriers they may encounter, they are able to develop more realistic plans for how to direct their energies.

With regard to the social category of gender, relevant questions may include: What rules did you learn about what it means to be a woman or man? Where and how were these rules conveyed? What are the rewards for abiding by these rules and consequences of breaking rules? How is gender relevant in your work world? What other important aspects of your identity are relevant to the skills you hold and the limitations you experience?

In addition to asking open-ended questions, the therapist may use follow-up prompts that help the client understand or explore biases or privileges and their consequences. To emphasize clients' resilience and strength, it is also useful to inquire about ways in which clients have coped effectively with barriers. Although the acronyms presented above identify potential areas for identity analysis, each therapist develops methods for inquiring about these areas in ways that are natural and integrated with the therapist's style.

Other important feminist therapy techniques include power analysis and strength analysis, including reframing. As a part of power analysis, clients are encouraged to examine the role that power plays in their lives, including power to (personal agency and initiative), power with (collaborative power), and power over (dominance). Key questions include: when and how do clients experience abuses of power, how can abuses of power be challenged, and how can clients learn to experience productive forms of power or empowerment that support positive personal action or collaborative efforts? Strength analysis encompasses any methods for identifying strengths as well as methods for using these strengths in support of health. Reframing, or considering how problem behaviors or situations can be redirected or reconstrued in the service of positive health, is closely related to strength analysis. Strength and coping interventions, which have been components of feminist therapy since its inception, have now been widely adopted by persons who label their work as positive psychology.

Feminist Therapy as an Integrative Approach

Since its earliest days, feminist therapy practice has consisted of integrating feminist philosophical values with a wide variety of techniques that are compatible with a feminist worldview. Any psychotherapy techniques and theories may be integrated with a feminist approach, provided that these techniques do not lend support to biases or outcomes that are inconsistent with egalitarian values. Prior to blending specific psychotherapy approaches with a feminist framework, Worell and Remer (2003) recommended that feminist psychotherapists explore and evaluate specific theories (e.g., humanistic, psychodynamic) by considering the following questions: What forms of bias are embedded in this approach as they pertain to its historical evolution, theoretical concepts, assumptions about the nature of problems, assessment methods, techniques, and the roles of counselor and client? If biases exist, is it possible to restructure aspects of the approach so that it supports egalitarian and social justice goals? How can the approach be enhanced to support individual empowerment as well as social change that will prevent future problems?

The integration of feminist values and other psychotherapy approaches is often seen in approaches directed toward addressing specific issues such as sexual violence, domestic violence, and body image and eating disorders. In these cases, psychoeducation about cultural values such as rape and domestic violence myths or beliefs that support eating disorders is an important component of treatment that enlightens clients about how the personal is political. Through the use of open-ended questions and the discussion of cultural biases, clients are likely to experience greater awareness of their concerns and experience the effects of consciousness raising. A wide range of approaches (e.g., cognitive therapy, interpersonal therapy) can be paired with new awareness of cultural myths and beliefs and used to challenge nonproductive belief systems or develop new coping or behavioral skills. Although outcome research on feminist therapy is limited, it is important to note that the pairing of evidence-based treatments with a feminist worldview provides a solid foundation for competent practice.

Over time, convergence between the goals of feminist therapy and other approaches has occurred. For example, feminist therapies and multicultural therapies increasingly share an emphasis on eradicating injustice. Feminist therapy values related to informed consent (Hare-Mustin et al., 1979) are now widely adopted as important ethical aspects of psychological practice and have evolved as a result of growing understanding of clients as collaborators and shared partners. The impact of feminist values on psychotherapy practice in general can also be seen in APA's (2007) *Guidelines for Psychological Practice With Girls and Women*, which have been adopted as characteristics of optimal practice that transcend specific theoretical approaches. The 11 guidelines are divided into three clusters of (1) diversity, social context, and power; (2) professional responsibility; and (3) practice applications.

The first cluster of guidelines promote an integrated, biopsychosocial, ecological approach to understanding women's lives, and emphasize the importance to addressing socialization, multiple and intersecting identities, and experiences of privilege and disadvantage or oppression. The guidelines also encourage psychologists to be aware of ways in which biases can be institutionalized in health systems, education, the workplace, religious groups, legal systems, and family systems. By being mindful of these biases, psychologists are able to help clients separate internal from external causes of distress and to resist confining cultural socialization. In addition, psychologists are encouraged to integrate community resources with their psychotherapy work and to emphasize prevention and resilience (APA, 2007; Enns et al., 2015).

The second cluster of guidelines, which focus on professional responsibility, advises practitioners to be aware of the strengths and limitations of traditional diagnoses, assessment processes, and interventions. These guidelines encourage psychologists to implement methods that build on client strengths and that challenge barriers to wholeness. The benefit of inclusive, flexible definitions of mental health and goals is highlighted, as are recommendations that contextual and cultural aspects of people's lives receive emphasis. The importance of therapist self-awareness and self-monitoring is also identified as a major feature of professional responsibility (APA, 2007; Enns et al., 2015).

The third and final cluster speaks to specific interventions. Practitioners are encouraged to use approaches that support collaborative, power sharing relationships, and empowerment and client initiative. Practitioners are advised to recognize that optimal growth is likely to be enhanced through the use of community resources, educational services and centers, shelters and safe houses, work resources, and public assistance, and to draw on these resources to expand services available to clients. Finally, the importance of devoting attention to changing systemic and institutional biases is discussed. Practitioners are encouraged to extend the purview of their work to include social justice, advocacy, and public policy initiatives (APA, 2007; Enns et al., 2015).

Conclusion and Future Directions

The initial focus of feminist therapy was to name biases and to create a validated discipline to address them. Consciousness raising groups increased personal insights, which led to the use of these methods in feminist psychotherapy practice. Over time, feminists have examined ways in which early 'grand theories' of feminism are limited for understanding the complexity and diversity of individual lives, especially when considered in transnational context. Feminist psychotherapies based on 'locational' feminisms are emerging in multiple regions of the world (Enns, 2004, 2012; Enns et al., 2015), and are incorporating culture-specific practices and knowledge. Feminist multicultural and transnational psychotherapies are likely two major areas of growth during the twenty-first century.

See also: Diagnostic Processes in Clinical Psychology and Psychotherapy; Feminisms and Acculturation around the Globe; Interpersonal Psychotherapy; Psychotherapy Integration; Psychotherapy Integration; Therapist–Patient Relationship.

Bibliography

American Psychiatric Association, 2013. Diagnostic and Statistical Manual, fifth ed. Author, Washington, DC.

American Psychological Association, 2007. Guidelines for psychological practice with girls and women. American Psychologist 62, 949–979.

Ballou, M., Matsumoto, A., Wagner, M., 2002. Toward a feminist ecological theory of human nature: theory building in response to real-world dynamics. In: Ballou, M., Brown, L.S. (Eds.), Rethinking Mental Health and Disorder: Feminist Perspectives. Guildford Press, New York, pp. 99–141.

Brown, L.S., 1994. Subversive Dialogues: Theory in Feminist Therapy. Basic Books, New York.

Brown, L.S., 2010. Feminist Therapy. American Psychological Association Press, Washington, DC.

Chesler, P., 1972. Women and Madness. Doubleday, New York.

D'Andrea, M., Daniels, J., 2001. RESPECTFUL counseling: an integrative model for counselors. In: Pope-Davis, D., Coleman, H. (Eds.), The Interface of Class, Culture and Gender in Counseling. Sage, Thousand Oaks, CA, pp. 417–466.

Enns, C.Z., 2004. Feminist Theories and Feminist Psychotherapies: Origins, Themes, and Diversity. The Haworth Press, New York.

Enns, C.Z., 2012. Feminist approaches to counseling. In: Altmaier, E.M., Hansen, J.C. (Eds.). The Oxford Handbook of Counseling Psychology, New York, Oxford, pp. 434–459.

Enns, C.Z., Rice, J.K., Nutt, R.L. (Eds.), 2015. Psychological Practice with Women: Guidelines, Diversity, Empowerment. American Psychological Association, Washington, DC.

Essed, P., 2001. Understanding Everyday Racism: An Interdisciplinary Theory. Sage, Thousand Oaks, CA.

Feminist Therapy Institute, 2000. Feminist Therapy Institute Code of Ethics [revised 1999]. Feminist Therapy Institute, San Francisco.

Gilbert, L.A., 1980. Feminist therapy. In: Brodsky, A., Hare-Mustin, R.T. (Eds.), Women and Psychotherapy. Guilford Press, New York, pp. 245–265.

Gilligan, C., 1982. In a Different Voice. Harvard University Press, Cambridge, MA.

Hare-Mustin, R.T., Marecek, J., Kaplan, A.G., Liss-Levinson, N., 1979. Rights of clients, responsibilities of therapists. American Psychologist 34, 3–16.

Hays, P.A., 2008. Addressing Cultural Complexities in Practice, second ed. American Psychological Association, Washington DC.

hooks, b., 2000. Feminism Is for Everybody. South End Press, Cambridge, MA.

Hyde, J.S., 2013. Half the Human Experience, eighth ed. Wadsworth, Belmont, CA.

Moradi, B., Fischer, A.R., Hill, M.S., Jome, L.M., Blum, S.A., 2000. Does "feminist" plus "therapist" equal "feminist therapist"? Psychology of Women Quarterly 24, 285–296.

Shields, S.A., 2008. Gender: an intersectionality perspective. Sex Roles 59, 301–311.

Worell, J., 2001. Feminist interventions: accountability beyond symptom reduction. Psychology of Women Quarterly 25, 335–343.

Worell, J., Remer, P., 2003. Feminist Perspectives in Therapy—Empowering Diverse Women, second ed. Wiley, New York.

Feminist Research and Social Work

Ruth Phillips, The University of Sydney, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Growing from challenges to dominant, masculine thinking and action for social change for women, feminism offered frameworks for practice and research. Social work's unique relationship to women as the majority of its workforce and as clients fits well with evolutions in feminist thought. Key debates within feminism are echoed in debates within social work, including challenges to traditional research modalities, views of the domestic sphere, conceptualisations of ethics and ideas of the self-aware researcher. This article provides an overview of what feminist social work research is, how it is informed by feminist theories and research methodologies.

Introduction

The nature of social work as both a profession of mostly women and as a profession that works with mostly women means that it is a practice and scholarly discipline that has been strongly influenced by the women's movement and feminism. As prominent feminist social work researcher Lena Dominelli observed, arguments about the purpose of social work are bound up with debates about women's position in the social order, thus feminist debates have conceptual frameworks that respond to and fit with 'the changing nature of women's lives' (Dominelli, 2002). Sands and Nuccio (1992) also observed that many social work researchers have found a high level of compatibility between feminist theories and social work. Supporting this, is a detailed meta-analysis of empirical studies on the effectiveness of feminist social work practice methods by Gorey et al. (2008) who found that feminist interventions were more effective than those based on other practice methods.

However, social work research has been historically dominated by nonfeminist research and debates have been sustained around the application of experimental/empirical and nonexperimental approaches with no reference to gender relations (Gibbs, 2001; Trinder, 2000: p. 10). More recently, the focus of social work research has shifted from national, large-scale surveys on the "structural determinants of needs and problems to a narrower focus on the effectiveness of different interventions; an analysis of the dynamics and meanings in relationships; and a developing research agenda of the politics of identity and individual agency" (Gibbs, 2001: p. 691). Critical or standpoint feminist research has been historically dominant in feminist social work research, drawing its theoretical principles primarily from Marxism and radical feminism (Trinder, 2000: p. 44; Swigonski, 1994: p. 391). However, since the early 1980s, there has been a strong trend toward postmodern/poststucturalist feminist research (Trinder, 2000; Fawcett et al., 2000; Sands and Nuccio, 1992). This shift in methodology has been highly influenced by key feminist thinkers such as Judith Butler (1990), who posed many theoretical challenges related to feminist advocacy, the construction of gender and meaning, and language and ideas within feminist research. The postmodern turn has been influential across all social work research as the desire to see research participants as having fluid identities and being active agents, not only to understand, but also to empower participants in the research process, has become part and parcel of more self-reflective social work theory and practice. Williams (1996) argued, however, that feminism as a broad modernist epistemology had already done much of the work of postmodernism in that it provided a critique of universalist thought that had been wholly focused on male thinkers taking a male-centered view. This challenged the binary view of male/female by calling for equality; the private/public divide by making both spheres political; becoming suspicious of 'meta-narrative' by listening to women's voices and their experiences of power; and showing the power of the ideological representation of women as whores or mothers (Williams, 1996: pp. 64–65).

Broadly, feminist theory both motivates research and provides incentives for doing research, as well as integrating the relationship between activism, advocacy, research, and social change. This can be seen most clearly in relation to key social work fields such as child sexual abuse and domestic violence. Through consciousness raising and solidarity, women's liberation movements from the late 1960s to the late 1980s opened up spaces for women to speak of their experiences of violence and sexual abuse previously hidden in the private sphere of the home and to demand that previously unspoken and hidden abuses of gendered power were placed on governments' social policy and legislative agendas (Dominelli, 2002: p. 37). Feminist research has played a key role in bringing about reforms for appropriate responses to and training for authorities such as the police, the courts, and medical professionals.

Since the early 1990s, feminism in general has battled to maintain a public discourse that recognizes gender inequalities in both private and public spheres in Western countries where neoliberal agendas focusing on markets over social objectives have dominated politics (Hannam, 2007: p. 161). However, this has not occurred internationally where human rights and development policies have increasingly taken account of and included feminist objectives for social change (Lazzari et al., 2009: p. 348). Nevertheless, feminism has had a profound influence on social work practice, theory, and research (Fawcett et al., 2000). Like feminism, social work is far from homogenous and approaches to social work research reflect a wide range of disciplines and epistemologies. Feminist social

work research, while working within specific disciplines such as sociology, political science, ethnography, or psychology, shares the wider task of feminism, which seeks to transform traditions within disciplines to create an intellectual revolution that contribute to broad social change objectives (De Vault, 1990).

Feminist social work research is conducted across all fields of social work practice as well as in teaching and learning and theory building. Prominent fields of feminist social work research are where feminists have had a groundbreaking influence in social and political responses to social problems. For example, research on violence against women has resulted in significant changes within legal, policy, and practice responses (Laing et al., 2013). The field of care has also been heavily researched and influenced by feminist social work researchers. This arose from feminist demands, on the state and society, to recognize the value of care work and women's dominant role in caring for children, people with disabilities, and older persons and to fundamentally challenge the ethics of care. Other roles in society dominated by women are also of great interest to feminist social work researchers, such as poverty, child abuse, volunteering, community work/development, international development, sexual assault services, aging, youth services, and mental and wider health services (Kurz et al., 2005; Kisor and Kendal-Wilson, 2002; Featherstone, 2006; Hanisch and Moulding, 2011).

Feminist Research: Objectives and Methods

Following the development of the women's movement across the world during the 1960s and 1970s, the scholarly framing of specifically identifiable feminist research methodology and research objectives emerged during the early 1980s to the 1990s (Eichler, 1987; De Vault, 1990; Wolf, 1996; Reinharz, 1992). Feminist research has since adhered to specific ways of researching that have supported the objectives of the women's movement and scholarly feminism in its ever-transforming manifestations. Feminist research has been characteristically concerned with the oppression and exclusion of women; identity and difference; the value of subjectivity and personal experience; setting itself apart from reductionist scientific research; women's ways of knowing; and listening to and hearing women's voices.

Reinharz (1992) viewed feminist research as methods used in research by self-identified feminists; as research published in feminist journals; and as research methods externally recognized as feminist research. This may exclude feminist researchers who approach their research with a feminist agenda, but do not necessarily identify as feminist (Dankoski, 2000: p. 7). Reflecting a postmodern framework, Lather (1988) suggested that feminist research puts the social construction of gender at the center of research inquiry. A further differentiating feature of feminist research is that it 'moves the researcher into the role of change agent,' using research findings 'to influence or change the lives of those studied and of other members of the group' (Cummerton, 1986: p. 95). Harding argues that the focus of feminist research is not on methods, but on the way feminists do research. She observed that they "listen carefully to how women informants think about their lives and men's lives, and critically to how traditional social scientists conceptualize women's and men's lives' and tend to focus on activities previously thought of as insignificant" (1987: p. 2).

Due to the diverse epistemological differences in feminist research, it is impossible to propose a singular definition. However, feminist researchers are primarily concerned with understanding why inequality between men and women exists and why there is the overall domination of men over women. The point on which the different feminist approaches disagree, however, is about where to find the answers to these questions and how to proceed in transforming women's lives through liberation from oppressions and inequalities (Letherby, 2003).

The pluralisms of feminisms are politically important to feminist researchers as they challenge a singular, universal feminism or feminist approach as an essential aspect of contemporary feminist thought. The various strands of feminism include liberal, radical, lesbian, liberationist, Marxist, socialist, ecofeminist, standpoint, third world, postcolonial, poststructuralist, and postmodern (Hannam, 2007; Reinharz, 1992: p. 6). These theories or political identities are often associated with a struggle between the structuralist and poststructuralist or modernist and postmodernist-based feminisms, reflecting the persistently, debated nature of feminist thought and practice. Broadly, the key difference between modernist feminist theory and postmodernist feminist theory is around how to explore the ways women exist or are placed in relation to men.

For structuralist feminists, the solutions are in targeting the structures that oppress women and maintain their inequality, from the interpersonal to the broad structure of the patriarchal state and the law. Therefore, structuralist feminists seek to establish universal truths about women's oppression. For poststructuralist feminists, the struggle is over meaning and the power of meaning that is primarily conveyed through language or discourse and is based on the fluid concept that there is no single truth or generalizable experience of being a woman. Therefore, the strategic approach to change is in the theorization and critique of discourse, knowledge, truth, reality, rationality, and the idea of the socially constructed subject (Sands and Nuccio, 1992). This approach has led to the primacy of individual narratives in social work practice and research, which in turn, has raised concerns for structuralist feminists because it ignores collective communities that are so integral to many women's lives (Dominelli, 2002: p. 35).

Despite these debates, and the generalized trend to frame feminism within the more general terms of gender, gender equality, and gendered society, rather than a focus on women per se, key feminist concerns persist. Such concerns are at the core of feminist social work research and are based on ways of understanding broader social injustices and inequalities between men and women by taking a women-centered or gender-aware approach. This includes a strong adherence to taking into account key intersections of difference such as class, disability, ethnicity, race, sexual preference, or religion.

Feminist Social Work Research Methodologies

Feminist social work researchers involved in knowledge production and aiming to make a difference to women's lives,

utilize a wide range of research methods, including qualitative and quantitative approaches. Within these broad approaches, there are defined feminist methodologies of research and knowledge production. Feminist researchers approach research from a political perspective that acknowledges inequalities between men and women and that traditionally the research process has been exploitative and hierarchical (Letherby, 2003; Reinharz, 1992). Feminist methodologies are informed by the drive to address women's position in society and ensure that the research process is transformed so as not to repeat wider oppressions that women experience. This includes methodologies such as feminist informed interview practice, which draws on a body of feminist ethics that deliberately separates itself from the scientific ethic of a 'detached' and powerful differentiation between the researcher and the subject. This approach is to counter the historical experience of women's voices and experiences being excluded from research as well as traditional scientific exploitation of subjects in the research process, and forms the epistemological basis of feminist social work research (Letherby, 2003).

Standpoint feminist social work research commonly applies qualitative methods, but also utilizes a quantitative approach that intrinsically seeks a collaborative position between the researcher and the research participants. Originally drawing from Marxism, standpoint feminism sought to uncover and challenge women's oppression, but transformed to take account of difference, reflecting an epistemological break from modernism (Heckman, 1997: pp. 356, 362). In standpoint feminist theory, the original tension between social construction and universal truth was challenged by deconstructing it and conceiving of knowledge as situational and part of a discursive process (Heckman, 1997: pp. 356–357). It represents the early dominant structuralist or modernist feminist as well as more recent postmodernist approaches in social work research, particularly in the field of domestic violence.

Also reflecting the influence of the postmodern turn is an increasing application of 'intersectionality' as an analytical tool that is concerned with the intersection between race, gender, and class, and allows for a focus on the relevance of inequalities and differences between and within social or economic groups (Jordan-Zachery, 2007). Although many authors credit Kimberlé Crenshaw (1989) with originating 'intersectionality' for feminist research, Butler's critique of the universalized term women/woman as a fixed identity also gave rise to the idea of intersectional approaches, when she observed that

> …the term fails to be exhaustive, not because a pregendered "person" transcends the specific paraphernalia of its gender, but because gender is not always constituted coherently or consistently in different historical contexts and because gender intersects with racial, class, ethnic, sexual, and regional modalities of discursively constituted identities. (1990: p. 4)

Butler is critical of intersectional feminist methodology because it presents a list of fixed identities, emphasized by the often added 'etc.' (Butler, 1990: p. 143). Intersectionality is utilized by both structuralist and poststucturalist feminists because it addresses 'the most central theoretical and normative concern within feminist scholarship,' that of the recognition and acknowledgment of differences among women (Davis, 2008: p. 70) thus addressing the major critique of dominant early structural feminism's exclusions of certain groups of women through the universalization of the category of women.

Mehrotra made the observation that "as an eclectic, evolving body of scholarship that is committed to social justice and the empowerment of all women in diverse communities, feminist social work is well-positioned to engage with questions regarding intersectionality theorizing and its relevance to social work scholarship and practice" (2010: p. 418). Mehrotra (2010) highlighted four key characteristics of intersectional analysis: placing the lived experience of people of color and other marginalized groups as central to development theory; revealing how diversity within groups is often ignored and essentialized while exploring complexities of identity within groups; showing how social inequality and oppression are manifested in domains of power; and promoting social justice and social change through the research process.

Further to the broad methodological approach of intersectionality, there are specific feminist research theorizations about the positioning of the researcher as self-aware and self-reflective. For example, Mehrotra describes herself a 'South Asian queer woman' who focuses her research on lesbian, gay, bisexual, transgender, and queer communities of color and South Asian migrant communities, thus demonstrating the relationship between the researcher and the research subjects and the lived experience of intersectionality' (2010: p. 418). This is connected to the core radical feminist theory of the personal as political and is manifested in research processes. However, it is also an approach shared by postmodern feminist researchers who adopt a view that their experiences will inevitably affect the research and cannot be seen as separate from the data derived from the research (Wendt and Boylan, 2008: p. 602). This approach draws largely on the highly influential theorization of reflexivity in practice by Jan Fook, who defines reflexivity as the process of being able to locate oneself in the context of work or research and appreciate how one affects or influences the research or practice through one's own beliefs and perspectives and social, economic, and political background (2002: pp. 157–158). Importantly, these components of the self-reflective researcher are related to understanding the exercise of power through knowledge and discourse.

For international research that takes place in the Global South and is conducted by feminist researchers from rich nations, similar concerns arise in fieldwork. Feminist theoretical approaches provide guidance on how to conduct research in this context that is sensitive to power imbalances and does not repeat the errors of previous exploitative scientific approaches that resulted in a form of colonization or recolonization (Wolf, 1996). Postcolonial feminist research methodology specifically addresses the issues of power and inequality between researcher and research participant by prioritizing, regardless of their economic or social position, that women's voices should be heard in research and that it is a key task of the researcher to facilitate listening to their voices.

Contemporary feminist social work research is disseminated widely through key international and national refereed journals. For example, *Affilia: A Journal of Women and Social Work* is a prominent, dedicated feminist social work journal publishing peer-reviewed research papers since 1986. Other leading social work journals are important vehicles for the dissemination of feminist social work research, such as the *British Journal of Social Work*, *International Social Work*, and *Australian Social Work*.

Despite its presence in prominent journals, there is also research about the lack of representation of feminist research, and feminist consciousness in wider published social work research, compared to the presence of feminist social work practice. For example, Barretti (2011) conducted a content analysis of abstracts of 17 United States–based social work journals over two 10-year periods up to 1996 and then up to 2007, and found a decrease in the latter period in both 'women's' and feminist content, suggesting that the strength of the relationship between feminist research and social work research had diminished in that time. A search in 2012 of the US National Association of Social Work's journal *Social Work Research*, published since 1995, showed a low presence of feminist social work research. The number of papers that included the words 'women' and/or 'feminist' in the main text was 268 over a 17-year period, a relatively low representation compared with 2168 in the comparable *British Journal of Social Work* over the same period, perhaps indicating a stronger presence of feminism in British social work research than in the United States. Barretti noted that the presence of the designated feminist social work research journal *Affilia* in the United States, which has also seen a decrease (by 24%) in its articles informed by feminism, adds to the sense that women's content across social work research in general can be supplemental, and seems to be suffering retrenchment (2011: p. 275). Feminist researchers and others, who have taken up issues raised by feminists, have also turned the research lens on social work journals and wider research to demonstrate not only the absence of feminist research but inherent sexism within the published research and tendencies to ignore the implications of gendered social issues in practice (Pease, 2011: pp. 408–410; Parton, 2003: pp. 10–11; Dietz, 2000).

The Focus and Future of Feminist Social Work Research

Feminist social work research is pervasive across all human service and social policy areas. However, there are some areas of social work research in which feminist researchers are more evident or visible than others (such as domestic violence, family support, child protection, childcare, and 'care'), often due to the high presence of women as clients in key areas of social work practice. It is also the nature of social work that it overlaps with other professional domains and scholarly disciplines. For example, in health and mental health, feminist social work practitioners and researchers play significant roles, but in such a huge field, they are often overshadowed by the dominance of biomedical practice and research, or other 'care' professions. Just three core areas are discussed below.

Aging

Social work with older people is an area of rapid growth due to worldwide population aging and consequent demands for greater research. This is particularly important for feminist research, as women constitute a dominant group among elders within most communities and the majority of older women have lived unequal lives due to their gender and caring responsibilities and their experiences have been largely ignored in mainstream research (Bozalek and Hoolyman, 2013). Feminist social work research has focused on older women and homelessness; abuse and exploitation of older women; caring roles of older women; intersections of race, sexual orientation, human immunodeficiency virus infection/acquired immunodeficiency syndrome, gender, age and class; effects of work and retirement; older women and autonomy and agency; and older women and welfare (Browne, 1995; Bayliss, 2000; Langley, 2001).

For feminist social work researchers, there are not only previously neglected areas of research on aging and the needs of older women, but also an impetus to apply the range of feminist approaches to lift knowledge and social work responses to the needs of older women to achieve emancipatory change. An example is Patterson's (2003) feminist study exploring the life courses of 25 Asian and Caribbean born elderly women who migrated as adults to the United States, which directly links its findings to improve social work practice with diverse older women from diverse backgrounds. The study emphasized the 'participant as expert' in their own lives approach and the potential for such documented experience to teach effective and appropriate social policy and social work practice.

Future feminist social work research on aging will continue to resist the dominance of biomedical knowledge production about aging; to reveal the complexity of women's life narratives to combat ageist and stereotypical ideas of women's histories; to explore the impact of menopause, equal to fertility and reproduction; to reveal distinct experiences of older women's poverty and secondary economic status; to establish knowledge of strengths in women's aging and counter marginalization of older women; and to do research that integrates developments into wider social justice issues linked to the diverse places and roles of older women.

Domestic Violence

Findings in early feminist research on domestic violence had a profound impact on how social workers responded to women as victims of violence. A study by Dobash and Dobash (1979) found that in the United Kingdom, police, court personnel, and other helping professions functioned as obstacles to the safety of women suffering domestic violence, establishing that social worker responses lacked resources and reflected beliefs that the violence was the woman's problem and that the social worker's role was to maintain the nuclear family structure. Equally significant was the research by Stark and Flitcraft (1996) that revealed the extent of domestic violence through the rate of injuries of women presenting to hospitals and established that feminist advocacy and safe refuge are a key elements of women's safety from domestic violence. This kind of research was crucial for naming

the issue of domestic violence and developing a research practice that began to listen to women's voices about their experiences of violence.

The focus of feminist social work research has been on establishing the complexity of violence perpetrated by intimate partners, to include emotional, financial, and physical dimensions; assessing the extent, cost, and impact of domestic violence across diverse communities; establishing links between domestic violence and risks for children; mental health issues for women arising from experiences of domestic violence; the relationships between housing, homelessness, and domestic violence; the impact of sexual orientation, social, and cultural roots of violence against women; lesbian experiences of intimate partner violence; and legal issues and deaths related to women's protection and security after leaving a violent relationship. An example of effective feminist research in this field is a study by Laing et al. (2012) that draws on a large research enquiry of four separate studies that explored collaboration between mental health and women's domestic violence services and resulted in enhanced collaboration across the domestic violence and mental health sectors, resulting in important benefits for women dealing with domestic violence and mental health issues.

International studies on domestic violence represent the scope of intersectional feminist social work research (Sokoloff and Pratt, 2005). It also shows a lack of progress by governments in responding to domestic violence in many countries and that civil society organizations are most active in achieving gains in social policy responses (Htun and Weldon, 2012). The future will see a growth in international feminist social work research on domestic violence as social work research catches up with activist feminist advocates working in the civil society sector. Analyses of key global social policy responses to violence against women, indigenous peoples' demands for survival and gender equality, and critiques of policies such as the United Nations Millennium Development Goals (MDGs), which have done little to improve women's equality status globally, will become an increasing focus. Dominelli (2010) also points to wider concerns for social work in the global context in noting that the MDGs are unlikely to meet their goals, including challenges brought about through financial pressures such as the global financial crisis and environmental or ecological crises due to climate change and industrial globalization. These issues, plus a need for 'green social work,' are part of the future social work research agenda (Dominelli, 2010, 2012).

Care

Although early feminist engagements with the notion of care were primarily concerned with the critique of care as gendered labor, more recent conceptualizations have valorized women's capacity and resourcefulness in caring activities. Feminist social work research into 'care' has been conducted within the range of practices and policies that inform its broad conceptualization. This includes a theoretically postmodern, emancipatory shift for the 'cared for' in the idea that care should no longer be viewed as a one-way process of dependency, as many studies have revealed that there is a higher level of mutual benefit in care relationships and that often the person constructed as the cared for also cares for their carer (Tronto, 1993; Berridge, 2012). Within social work, care encompasses the care of children, within families, fostering and communities and by the state as well as private and institutional care for children or adults with disabilities and care for frail-aged people. A major early focus for feminism was the taken for granted status of gendered nature of care, particularly a mother's care for children. More recently, focus has been on the fact that it is commonly a woman's role to care for disabled family members of all ages, elders, and grandchildren. Feminist research in care also engages with the status of the cared for, particularly in the area of aging as women's longevity has led to a higher demand for the care of frail aged women.

As a core challenge to the traditional philosophy of ethics and to engage with the gendered nature of care, and reform of how care is conceived and emphasize its societal contribution as a service, feminists have developed a feminist ethics of care (Parton, 2003; Sevenhuijsen, 1998). Like all areas of feminist thought, there is not a single set of ethics of care; however, there are key principles. The first articulations of feminist ethics were in countering classical liberal theories of the isolated individual and women taking responsibility for others as a form of moral strength and social interconnectedness (Gilligan, 1982: p. 17). Thus, "the moral subject in the discourse of care always already lives in a network of relationships, in which s/he has to find balances between different forms of responsibility" (Sevenhuijsen, 2000: p. 10). Carol Gilligan described a structuralist positioning for a feminist ethics of care as an "ethic of resistance to the injustices inherent in patriarchy … where it transcends gender binaries and hierarchies that structure patriarchal institutions and cultures" (Ethics of Care, 2011: p. 1).

Further developments in the feminist ethics of care saw an emphasis on relationships over rules and rights, as care was recognized as intrinsically bound to social practices, feelings, and opinions, and seen as a moral activity rather than a gendered one, therefore determined as a woman's role (Tronto, 1993). Feminist social work research took the framework of the ethics of care into all spheres of care research, including social policy analyses; the positioning of mothers in families where children have been sexually abused, child welfare and child protection; care and carers of people with disabilities; women's labor in the care industry; international perspectives on women as carers; and the experience of being a carer (Featherstone, 2006; Hanisch et al., 2011; Parton, 2003). An example of the latter research is Anne Petrovich's studied account of becoming a 'sandwich carer,' where she explores complex demands of working in the paid workforce, care as a mother; care as daughter of frail aged parents; and care and death, raising key issues related to self-determination, responsibility, and the complexity of care relationships (Petrovich, 2008).

Conclusion

For feminist social work research to maintain its momentum for social change that includes progress toward consistent, shared gender equality, it appears that a reaffirmation of the

role of feminist social work theory and research is required. Feminist methodologies that are currently applied have overlapping characteristics that reflect the evolution of feminisms more broadly. Regardless of epistemological identity, feminist research in social work has moved beyond failures of early Western feminism to recognize diversity and difference in the experiences of being a woman or women and has transformed the way that research is done by actively resisting the normative scientific approach that historically ignored the voices of women. It also refused to subject research participants to strict hierarchies encompassed by the professional researcher being the subject while the participant was the object of the research. Diversity across a range of social, cultural, and economic divisions is also a key theme within the practice of feminist social work research as its web of engagement extends to all areas of human services and social issues and contributes to change in social work and social policy practice and theorization.

Despite a general disquiet that feminism had stalled in the West, debates about feminist approaches to social work research continue to be present in social work education and, there appears to be an ongoing commitment by educators to ensure that feminist approaches are not lost and remain diverse. Further, the increasing internationalization of feminist research due to rapid changes in the populations with whom social workers work with across the world, related to growth in the number of refugees and rapid changes in immigration, poses fresh demands for increased expansion of feminist social work research (Kreitzer, 2002; Erez et al., 2009).

See also: Domestic Violence: Sociological Perspectives; Ecofeminism; Family Planning Programs: Feminist Perspectives; Feminism in Anthropology; Feminisms and Acculturation around the Globe; Feminist Criminology and Gender Studies; Feminist Psychotherapy; Gender and Feminist Studies in Geography; Gender and Social Research; Gender, Feminism, and Sexuality in Archaeological Studies; Intersectionality and Crime; Masculinity and Social Work Practice; Queer Theory and Intersectionality.

Bibliography

Barretti, M.A., 2011. Women, feminism, and social work journals 10 years later: 1998–2007. Affilia 26 (3), 264–277.
Bayliss, K., 2000. Social work values, anti discriminatory practice and working with older lesbians. Social Work Education: The International Journal 19 (1), 45–53.
Berridge, C., 2012. Envisioning a gerontology-enriched theory of care. Affilia 27 (1), 8–21.
Bozalek, V., Hoolyman, N., 2013. Ageing and intergenerational care: critical/political ethics of care and feminist gerontology perspectives. Agenda: Empowering Women for Gender Equality 26 (4), 37–47.
Browne, C., 1995. Empowerment in social work with older women. Social Work 40 (3), 358–365.
Butler, J., 1990. Gender Trouble, Feminism and the Subversion of Identity. Routledge, New York, Oxon.
Crenshaw, K., 1989. Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracisit politics. The University of Chicago Legal Forum 14, 538–554.
Cummerton, J., 1986. A feminist perspective on research: what does it help us see? In: Van Den Bergh, N., Cooper, L. (Eds.), Feminist Visions for Social Work. National Association of Social Workers Inc., Maryland, pp. 80–100.
Dankoski, M., 2000. What makes feminist research? Journal of Feminist Family Therapy 12 (1), 3–19.
Davis, K., 2008. Intersectionality as buzzword: a sociology of science perspective on what makes feminist theory successful. Feminist Theory 9 (1), 67–85.
De Vault, M.L., 1990. Talking and listening from women's standpoint: feminist strategies for interviewing and analysis. Social Problems 37 (1), 96–116.
Dietz, C.A., 2000. Responding to oppression and abuse: a feminist challenge to clinical Social Work. Affilia 15 (3), 369–389.
Dobash, R.E., Dobash, R.P., 1979. Violence against Wives: A Case against the Patriarchy. Free Press, New York.
Dominelli, L., 2002. Feminist Social Work Theory and Practice. Palgrave, Basingstoke and New York.
Dominelli, L., 2010. Globalization, contemporary challenges and social work practice. International Social Work 53 (5), 599–612.
Dominelli, L., 2012. Green Social Work. Polity Press, Cambridge.
Eichler, M., 1987. Nonsexist Research Methods. Routledge, New York, London.
Erez, E., Adelman, M., Gregory, C., 2009. Intersections of immigration and domestic violence: voices of battered immigrant women. Feminist Criminology 4 (1), 32–56.
Ethics of Care, 2011. Carol Gilligan [Online]. University of Tilburg, University of Humanistic Studies at Utrecht, Tilburg, Utrecht. Available: http://ethicsofcare.org/interviews/carol-gilligan/ (accessed 13.01.13.).
Fawcett, B., Featherstone, B., Fook, J., Rossiter, A., 2000. Practice and Research in Social Work, Postmodern Feminist Perspectives. Routledge, London, New York.
Featherstone, B., 2006. Why gender matters in child welfare and protection. Critical Social Policy 26 (2), 294–314.
Fook, J., 2002. Social Work Critical Theory and Practice. Sage, London.
Gibbs, A., 2001. The changing nature and context of social work research. British Journal of Social Work 31 (5), 687–704.
Gilligan, C., 1982. In a Different Voice: Psychological Theory and Women's Development. Harvard University Press, Cambridge.
Gorey, K., Daly, C., Richter, N., Gleason, D., McCallum, M.J., 2008. The effectiveness of feminist social work methods. Journal of Social Service Research 29 (1), 37–55.
Hanisch, D., Moulding, N., 2011. Power, gender, and social work responses to child sexual abuse. Affilia 26 (3), 278–290.
Hannam, J., 2007. Feminism. Pearson Education, Harlow.
Harding, S. (Ed.), 1987. Feminism and Methodology. Indiana University Press, Bloomington.
Heckman, S., 1997. Truth and method: standpoint theory revisited. Signs 22 (2), 341–365.
Htun, M., Weldon, S., 2012. The civic origins of progressive policy change: combating violence against women in global perspective, 1975–2005. Perspectives on Politics 106 (3), 548–569.
Jordan-Zachery, J., 2007. Am I a black woman or a woman who is black? A few thoughts on the meaning of intersectionality. Politics and Gender 3 (2), 254–263.
Kisor, A.J., Kendal-Wilson, L., 2002. Older homeless women: reframing the stereotype of the bag lady. Affilia 17 (3), 354–370.
Kreitzer, L., 2002. Liberian refugee women: a qualitative study of their participation in planning camp programmes. International Social Work 45 (1), 45–58.
Kurz, B., Malcolm, B., Cournoyer, D., 2005. In the shadow of race: immigrant status and mental health. Affilia 20 (4), 434–447.
Laing, L., Humphreys, C., Cavanagh, K., 2013. Social Work and Domestic Violence. Sage, London.
Laing, L., Irwin, J., Toivonen, C., 2012. Across the divide: using research to enhance collaboration between mental health and domestic violence services. Australian Social Work 57 (4), 419–421.
Langley, J., 2001. Developing anti-oppressive empowerment social work practice with older lesbian women and gay men. British Journal of Social Work 31 (6), 917–932.
Lather, P., 1988. Feminist perspectives on empowering research methodologies. Women's Studies International Forum 11 (6), 569–581.
Lazzari, M., Colarossi, L., Collins, K., 2009. Feminists in social work, where have all the leaders gone? Affilia 24 (4), 348–359.
Letherby, G., 2003. Feminist Research in Theory and Practice. Open University Press, Buckingham.
Mehrotra, G., 2010. Toward a continuum of intersectionality theorizing for feminist social work scholarship. Affilia 25 (4), 417–430.
Parton, N., 2003. Rethinking professional practice: the contributions of social constructionism and the feminist 'ethics of care'. British Journal of Social Work 33 (1), 1–16.
Patterson, F.M., 2003. Heeding new voices: gender-related herstories of Asian and Caribbean-born elderly women. Affilia 18 (1), 68–79.
Pease, B., 2011. Men in social work: challenging or reproducing an unequal gender regime? Affilia 26 (4), 406–418.
Petrovich, A., 2008. Lessons learned in the sandwich. Affilia 23 (3), 223–230.
Reinharz, S., 1992. Feminist Methods in Social Research. Oxford University Press, New York.

Sands, R.G., Nuccio, K., 1992. Postmodern feminist theory and social work. Social Work 37 (6), 489–494.
Sevenhuijsen, S., 1998. Citizenship and the Ethics of Care: Feminist Considerations on Justice, Morality and Politics. Routledge, London, New York.
Sevenhuijsen, S., 2000. Caring in the third way; the relation between obligation, responsibility and care in Third Way discourse. Critical Social Policy 20 (1), 5–37.
Sokoloff, N., Pratt, C., 2005. Domestic Violence at the Margins: Readings on Race, Class, Gender, and Culture. Rutgers University Press, Piscataway, NJ.
Stark, E., Flitcraft, A., 1996. Women at Risk, Domestic Violence and Women's Health. Sage Publications, Thousand Oaks, London, New Delhi.
Swigonski, M.E., 1994. The logic of feminist standpoint theory for social work research. Social Work 39 (4), 387–393.
Trinder, L., 2000. Postmodern feminism and the 'doing' of research. In: Fawcett, B., Featherstone, B., Fook, J., Rossiter, A. (Eds.), Practice and Research in Social Work, Postmodern Feminist Perspectives. Routledge, London, New York.
Tronto, J., 1993. Moral Boundaries: A Political Argument for an Ethic of Care. Routledge, London.
Wendt, S., Boylan, J., 2008. Feminist social work research engaging with post-structural ideas. International Social Work 51 (5), 599–609.
Williams, F., 1996. Postmodern feminism and the question of difference. In: Parton, N. (Ed.), Social Theory, Social Change and Social Work. Routledge, London, New York, pp. 61–76.
Wolf, D.E., 1996. Feminist Dilemmas in Fieldwork. Westview Press, Boulder and Oxford.

Feminist Theory: Radical Lesbian

Wendy L Lee, Bloomsburg University of Pennsylvania, Bloomsburg, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Reworked via socialist feminist and ecologically motivated objectives, radical lesbian feminism offers a powerful theoretical, practical, and morally sound framework for the contemporary ecofeminist movement. First, by appealing to the strengths of earlier analyses, radical lesbian feminism offers a critique of compulsory heterosexuality. Second, what distinguishes radical lesbian feminism from liberal feminism is the former's commitment to reclaim women's experiences, desires, bodies, and lives as meaningful in themselves. An ecologically oriented radical lesbian feminism offers an avenue for comprehending that experience informed, for example, by the ever-growing threat posed by climate change. Lastly, the emancipation of lesbian identities is essentially revolutionary in that it aims to undermine institutions responsible for oppression. Such a revolution has a specifically ecological signature in that the most globally threatening configuration of oppression to date – oppression premised on access to drinkable water and breathable air – is itself the product of heteropatriarchally dominated multinational energy corporations.

Introduction: Three Vital Foci of Radical Lesbian Feminism for a Twenty-First Century World

While some might insist that the heyday of radical lesbian feminism concluded with the end of the twentieth century, in what follows I'll argue that, suitably reconfigured via socialist feminist and ecologically motivated objectives, radical lesbian feminism offers among the most powerful theoretical, practical, and morally sound frameworks for the contemporary ecofeminist movement. Three objectives or foci will govern my argument: first, appealing to the strengths of earlier analyses, radical lesbian feminists offer what remains the 'gold standard' critique of compulsory heterosexuality that defines male sexual access to women's bodies as 'natural,' thus institutionalizing male sexual and economic entitlement. Such analysis, I'll argue, remains vital to contemporary theorizing and activism because it demonstrates in what heteropatriarchal prerogative consists and how it dominates institutions and practices affecting women's lives, particularly with respect to essential resources like drinkable water and breathable air.

Second, what distinguishes radical lesbian feminism from liberal feminism is the former's commitment to reclaim women's experiences, desires, bodies, and lives as meaningful in themselves. As opposed to the more or less unquestioning adoption of male-defined values by liberal feminism, radical lesbian feminism interrogates whether, for example, aspiring to become the first female chief executive officer of Dutch Royal Shell is a goal worth pursuing. Radical lesbian feminism also takes women's particular experiences – as mothers, daughters, wives, and as lesbians – not merely as defined in connection to male counterparts, but as culturally, geographically, religiously, and sexually specific. An ecologically oriented radical lesbian feminism offers a new avenue for comprehending women's experience as informed by the ever-growing threat from extreme forms of, among other things, fossil fuel extraction. Both inside women's bodies in the form of carcinogens like benzene associated with breast cancer (one of the multitude of additives to drilling fluids and hence to waste water), and with respect to responsibilities traditionally shouldered by women especially in the 'developing' world, such as the increasing scarcity of fresh water, women's experience of the heteropatriarchal domination of energy production and usage informs not only theorizing about women's experience of deforestation, desertification, pollution, and ultimately climate change, but also for strategizing an activism aimed at emancipating women, children, and indigenous peoples from this domination. Because radical lesbian feminism makes women's experience central to its philosophical framework, it is in a position to provide a powerful voice to those impacted by the multinational corporatism of industrial-level fossil fuel extraction.

Third, I'll argue that the aims of radical lesbian feminists include at least the subversion of heteropatriarchal institutions or – better – their end in that such institutions either require the denial of lesbian identity or insist on defining it in terms of male desire and privilege (such as male-focused female-on-female pornography). In other words, insofar as 'lesbian' does not exist for a sexuality defined in terms of male access to women's bodies, radical lesbian feminists have sought to redefine and radicalize 'sexual identity' to not merely include 'lesbian,' but to deepen and broaden sexual identity well beyond notions of male pleasure or male access. As a forceful expression of the feminist slogan 'the personal is political,' radical lesbian feminists seek the emancipation of lesbian identities as part and parcel of the emancipation of sexual and/or gender identities themselves. Such emancipation is, in fact, a revolutionary reach beyond sexual and/or gender expression: the emancipation of lesbian identities just is a revolution that aims to undermine institutions responsible for oppression. Such a revolution, moreover, has a specifically ecological signature in that the most globally threatening configuration of oppression to date, namely, oppression premised on access to the necessary conditions of life – drinkable water and breathable air – is itself the product of heteropatriarchally dominated multinational energy corporations.

In light, then, of its critique of compulsory heterosexuality, its focus and revaluation of women's experience, and its revolutionary understanding of sexuality and gender expression, radical lesbian feminism offers, I'll argue, a theoretical framework for an ecologically oriented feminism aimed at substantive social, economic, and environmentally effective change.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10201-6

The Radical Feminist Lesbian Future

Emerging from mainstream liberal feminism of the late 1960s and early 1970s, radical feminists challenge the prevailing view that liberating women consists in reforming social institutions such as marriage, family, or the organization of work. They argued that insofar as a deeper analysis of these institutions demonstrates the extent to which some men (white, Western, middle-class, and mostly Christian) continue to be privileged over other men and virtually all women, reform alone cannot achieve the equality liberal feminists envision. Economic equality in a world whose prevailing values and institutionalized practices systematically empower some at the expense and exploitation of those defined as other, inferior, or marginalized can never be more than cosmetic. Radical feminists thus promote not reform but revolutionary change in the ways the deeper and more endemic inequality conceive gender, sexual identity, and sexuality. The aim is to end the oppression of women by creating first awareness and then resistance not only to male-dominated or patriarchal institutions, but also to the conceptual frameworks that sustain them.

Since that time – some 30–40 years later – much has changed for both lesbian feminists, and for radical lesbian feminists. Indeed, the feminist movement has become both more revolutionary in some important ways and less so in others. While, for example, the critique of heteropatriarchal institutions has been embraced by many feminist theorists and activists, those same institutions, particularly at the level of military, corporate, and government hegemony, continue to expand in ways that reaffirm and empower some men. To take one striking instance, the global fossil fuel industry has come to epitomize the domination of white, affluent, mostly men over existential necessities such as water. In the face of climate change, combined with the fact that women – particularly in the developing world – remain primarily responsible for tasks that require clean water, such domination could effectively undermine many of the gains made over the twentieth century not only by women, but also by developing world peoples.

Climate change, in other words, functions as an effective barometer of the success of twentieth century social movements – including the radical feminist movement – in that the same institutions responsible for the production and emission of potentially catastrophic quantities of greenhouse gases – military, government, and corporatist – are also responsible for the ongoing oppression of women, children, indigenous peoples, and the economically marginalized. Hence, analyses of the causal, political, and economic conditions responsible for climate change are crucial to any twenty-first century analyses of marriage, work, and family. As many socialist feminists point out, these are precisely the institutions at once reinscribed as units of economic reproduction by global corporatism and profoundly threatened in their capacity to reproduce labor and laborers as access to clean water becomes increasingly scarce. Indeed, it adds another layer of meaning to show that institutions that have functioned as sites of oppression for women, marriage, for example, also support – quite literally – the manufacture of global and potentially ecologically catastrophic conditions. Because radical lesbian feminists offer some of the most trenchant analyses of the concepts sustaining this heteropatriarchal 'Old Boys Club,' they are arguably in the best position to expand analyses to incorporate more explicitly environmentally oriented strategies of resistance.

If any single motif governs feminism as a whole and its specific domains of theory and practice, it is the necessity to expand beyond the issues governing the movement's third wave – including matters of sexual identity, expression, and politics. This is not to say that such sexual issues have become less important. Quite to the contrary, it is to say that twenty-first century feminist analyses must take into consideration subject matter not explicitly thematized in earlier incarnations of the movement, but that affects how its central themes are understood in the face of new challenges – especially environmental challenges that affect the quality of human life, and the mobility of human populations. To take one example, the endocrine system disruptions to which we are exposed in so-called feminized fish, solicit a set of contentious questions concerning not merely why we continue to use this terminology, but why it continues to resonate with us epistemically. That we all know to what such terms refer implies precisely that fixed set of characteristics are identified as 'feminine' or 'masculine,' and thus offers a radical lesbian critique a potentially fruitful new context toward offering more acute analyses of a language use that reinforces oppressive conceptions of gender.

To take a different kind of example, consider the ways in which transexuality has come to challenge the ways in which we conceive 'nature.' When the 'natural' can become the end result of highly specific, technologically mediated, intervention, the word 'nature' arguably forfeits the meaning 'original' or 'real.' But this contentious use of 'nature' has a potent analogue in contemporary environmental discourse over whether, for instance, ecological 'restoration' can restore an ecosystem to its 'natural' – original – state, or whether such interventions on damaged or destroyed ecologies amount to no more than cosmetic change. The transsexual may believe that he or she is realizing – potentially even restoring – the sexual and/or gender identity he or she was 'meant' to have. The restoration ecologist hopes their work restores an ecology to an original 'as it was meant to be' state. But given the technologically mediated nature of both appeals to 'nature,' it is a murky business what 'nature' could usefully mean. Such, however, are precisely the questions to which radical lesbian feminists have offered acute attention since the 'natural' and the 'unnatural' function as conceptual as well as political, social, and economic sites of oppression in virtue of sexual and/or gender identity.

The Critique of Liberal Feminism

Perhaps what most obviously differentiates mainstream liberal feminism from its radicalized sister is that while the former continues to hope that the master's tools (e.g., capitalism, extractive technologies) can be deployed to dismantle and then rebuild a better version of the master's house (e.g., economy, environment), the latter has largely quit this project of rebuilding and jettisoned its tools in favor of building not merely a house but an entire conceptual 'civilization' whose collective mission it is to embody the experiences, values, and desires of women. That the master's tools cannot be deployed

to dismantle his house could not be made clearer than by climate change; for whatever relative equality some women have gained, there is little reason to believe women's emancipation has contributed to slowing the demand for fossil fuel extraction, preventing deforestation, or dealing effectively with water and air pollution. In fact, the opposite is likely to be true; in the absence of any sustained critique of the competitive and avaricious tools of heteropatriarchal institutions, there are no good reasons to think women will behave in any more environmentally conscious a way than men. Climate change is simply evidence to that end.

If what we mean by 'radical' directs us to analyses of the tools – the concepts – out of which the master's house is built, radical lesbian feminists are especially well positioned; as Audra Lorde put it, the master's tools cannot be used to dismantle the master's house. By refusing such tools in favor of those that probe first and foremost what the master's tools are for, radical feminists may occupy the best position to see that the master's house is merely the 'McMansion' whose supersized energy consumption epitomizes the use of a 'tool set' that, while it can accommodate a few tokenized women, can be put to building only one kind of house – the house that reaffirms the master. In other words, while the primary focus of liberal feminism has been the evaluation of patriarchal institutions in the interest of retooling them to affect greater inclusion for women within them, radical lesbian feminism takes its point of departure from the observation that tools used for the purposes of exploitation and oppression may not be suitable for emancipation or, as we now see, the construction of a sustainable future.

Perhaps, however, the fundamental insight for contemporary radical lesbian feminism was the recognition that it could never be enough to agitate for inclusion as long as women's value is based on men's interests and desires. This too comes into sharp focus via the facts of climate change. The merely cosmetic improvements for women produced through liberal feminism's focus on reconfiguring the master's tools fails to address very real-life and death issues – the having and rearing of children, for example – under conditions that expose mother and child to environmental toxins, scarce clean water, and potential mass migrations from shrinking coastlines. In this light, liberal feminism must be understood not simply as cosmetic but as complicitous in the oppression of women and compromised others while more privileged women enjoy the fruits of that exploitation of labor and resources. What radical lesbian feminists insist is that the change necessary to produce lasting improvement in all women's lives – including the effort to mitigate climate change – requires understanding how heteropatriarchal institutions are not merely oppressive, but threaten the conditions of human and nonhuman life itself.

Some radical lesbian feminists argue that the notion 'equality' is tethered to a view of human nature that so privileges attributes traditionally associated with men (aggression, hierarchical forms of organization, rugged individualism, competition, and an ethics focused on impersonal rule-following over relationality), that women are only able to be included within such conceptual frameworks as honorary men. It is little wonder, they argue, that examples offered to young women by liberal feminists as role models and heroines are modeled after male heroes; the accomplishments of women like Harriet Tubman, or Marie Curie (or movie heroines like those found in *Terminator II*, *Aliens*, *Femme Nikita*, *The Hunger Games*, or *The Girl with the Dragon Tattoo*) are valued according to a standard that rewards domination over domesticity, conquest over caring, and militarism over mothering – masculinity over femininity traditionally conceived. This is not to suggest that these are not examples of very real valor, but that what is rewarded in them goes further to reinforce a male-centered vision of the good than it does to foster a vision centered on 'feminine' identified values such as cooperation, compromise, collective decision making, or compassion. It is thus simply not enough, argue radical lesbian feminists, for women to settle for reformist programs that aim at an equality based on access to masculine values. Given the existing conceptual constraints, such programs will either fail or, even if they succeed, will do so only at the cost of continuing to devalue qualities associated with women and femininity.

Liberal feminism does not necessarily represent the desires or interests of all women, whether those desires are to be considered culturally constructed or endowed by nature. This thus brings us to what is commonly referred to as the essentialism versus antiessentialism debate; if what radical lesbian feminist conceptual analysis shows is that the ways we conceive of gender, sexual identity, and sexuality privilege men and we consider these to be a reflection of natural propensity, should not we then aim to elevate the value of feminine nurturing, emotional responsiveness, and the tendency to focus on relationships, over abstract autonomy? Or, alternatively, should we eschew traditional femininity characteristics as the construction of patriarchal culture, arguing that there are no essential qualities that importantly distinguish men from women?

The meanings of androgyny are instructive here. One meaning of androgyny favors the promotion of the masculine and the feminine in all of us, acknowledging that while gender may be constructed, such qualities may still be salvageable when reinterpreted in a nonsexist context. Another meaning of androgyny, however, suggests that we ought to abandon notions of femininity and masculinity altogether as unworkable in favor of some other notion of personhood. Here, the question that arises concerns what 'personhood' would look like, and whether it is possible to envision a 'human being' that can comprehend biological differences without reproducing an atomistic individual that it becomes subject to the same criticism radical feminists level at liberal feminism. Can we, in other words, envision a concept of 'embodied person,' which, irreducible to the determinations of either biology or culture, is rich enough to account for differences relevant to and affected by both?

A contemporary radical lesbian feminism is well equipped to address these questions since the atomistic personhood informing the essentialism debates is itself part and parcel of a 'nature' against which 'human being' is conceived as radically separate. A liberal feminist perspective might raise questions about the qualities described as 'masculine' or 'feminine,' whereas a radical lesbian feminist approach queries whether the atomistic conception of 'person' is itself defensible. Well illustrated by the facts of climate change, the radical feminist critique shows that the atomistic conception of personhood divorced from the necessary conditions of life – clean water and breathable air – is likely vulnerable to conceiving itself as, well,

invulnerable. When masculinized 'not dependent,' 'autonomous,' and 'self sufficient' describe the atomistic 'person,' nature can only be conceived as a resource whose value is not existential, but commodifiable.

Moreover, a personhood conceived as invulnerable is also likely to be conceived as infinitely adaptable – a human, all too human, trait that is both beneficial and damning. We may be able, for example, to adapt to some of the environmental and geopolitical demands of a warming climate, but like the proverbial frog in the pot of warming water, we will reach a point where adaptation will not prevent disaster. By demonstrating that the essentialist debates over 'masculinity' and 'femininity' are really symptomatic of deeper issues concerning how we conceive 'human being,' radical feminists show how one such conception – atomism – has become institutionalized not only as male privilege over women, but human privilege over nature – with potentially catastrophic consequences.

Radical Lesbian Feminism and a Radicalized Socialist Lesbian Ecofeminism

However we settle the essentialist debates, the history of heteropatriarchy shows that the oppression of women is sanctioned through the denigration of qualities identified as feminine. Whether we seek liberation through the elevation of, or the abolition of, femininity, what remains clear is that reformist programs for inclusion can never be adequate. Another possible avenue of critique of heteropatriarchy, one that skirts the essentialist debates, is that offered by Marxist/socialist feminists who show, given the marriage of patriarchy to capitalism, that it is illuminating to conceive of women not as a group but as a class whose oppression is maintained through economic means. This notion fits neatly with analyses that identify atomistic personhood with the commodification of nature, since to conceive of women as a class is to conceive women in terms of the commodity value of their labor. As Frederick Engels argued, the very first commodity is women's capacity to produce – paid work and unpaid domestic labor – and reproduce – produce labor for the next generation. So too nature; commodified first as resource and then conceived as interminable or endless, it is really no wonder that for marketing purposes, nature is feminized and women are naturalized. Both effectively marry patriarchy to capitalism in that both are subject not merely to domination, but exploitation conceived as having no natural point of exhaustion. Ironic it is then, that it is female-sexed reproductive systems that are especially vulnerable to cancers linked to environmental pollutants – for example, diesel, associated with climate change.

Combined with a radical lesbian feminist analysis of patriarchal concepts of motherhood and domestic labor, a Marxist/socialist account shows how divisions of labor are sexed in at least two ways: first, because women's *labor qua* procreation is assigned to the 'natural' it is not counted as labor at all, and second, because the domestic family-related labor performed predominantly by women has become 'naturalized' in a capitalist economy that identifies all labor that is not paid or wage as not labor. Cast as the vehicle for the reproduction and sustenance of life, 'woman' becomes a paradox. 'She' is unexploitable since her labor is not counted as labor under capitalism, yet 'she' also epitomizes the very prototype of exploitability in that 'her' labors are systematically and maximally undervalued, which, in a capitalist system means un/undercompensated.

A necessary point can thus be made without entering the essentialism/antiessentialism debate, the fact that women's oppression is intimately linked to the relationship between her economic status and her status as provider of hearth, home, and progeny for men, well supports the radical lesbian feminist claim that there are no social institutions in which men do not have a controlling interest, including economy and family. What recommends this approach over liberal feminism, then, is the emphasis on the role of economic class. The best that liberal feminists can hope for is a reform leading to an equality of competition between men and women within capitalism, whereas the radical Marxist/socialist feminists are free to imagine a revolution leading to a utopian world within which neither sex nor class oppression are operative – much less determinative – of social status.

While a radicalized Marxist/socialist feminism goes further than liberal or radical lesbian feminism alone to evaluate capitalist heteropatriarchy and offer an alternative, even it ultimately fails the project of genuine emancipation. For while such a critique of capitalism shows something important about the sexed parameters of economic oppression, there remains considerable work to show the relationship between economic and other forms of oppression such as sex, sexual orientation, race, age, ability, and so on. Here too, however, an expanded radical lesbian feminism has something important to offer; because each form of oppression is linked in to conceptions of 'nature,' each may be evaluated in terms of the ways in which they support not only oppression, but exploitation of the natural and nonhuman animal world. That persons of African descent are, for example, commonly animalized in pre–Civil War defenses of slavery while nonhuman animals are humanized in stereotypically heterosexualized ways, indicates that what governs the ways in which we conceive of race is essentially heteropatriarchal.

Given the need for this level of analysis, many radical lesbian feminists distinguish themselves from both mainstream liberal feminism and Marxist/socialist feminism by taking a more skeptical, self-reflective approach to theorizing, arguing that it is at least an open question whether any patriarchal institution is salvageable in any form. In a world dominated by climate change, this question becomes especially pointed. There is little more revolutionary a declaration than that multinational capitalism, including Big Energy, Big Box, and Big Pharmaceutical, must be overthrown, but if this is not accompanied by the overthrow of other patriarchal institutions such as the legal definition of family as male-headed, the status of women could remain unchanged or worsened as men's conditions improve. What revolutionary changes would institutions like the family need to undergo to achieve equality not only with respect to gender, but also with respect to any number of other axes of oppression, including race, ability, economic class, indigenous culture status, age, sexual identity, and sexuality?

Some radical lesbian feminists such as the late Mary Daly argued that patriarchy must be overthrown or abandoned in

favor of separatist or women-only cultures and societies. In her 'radical elemental feminist manifesto,' *Quintessence*, Daly describes the fundamental ideology of patriarchal institutions as rapist, "characterized by invasion, violation, degradation, objectification, and destruction of women and nature; the fundamental paradigm of sexism, racism, classism, speciesism, and all other oppressive-isms" (1998: p. 8). Not unlike the radical Marxist/socialist feminist's view that the sexed division of labor lay at the root of economic exploitation, Daly argued that at the root of all forms of oppression lay a very particular form of sexual domination – the male prerogative to rape – and that this form of sexual enslavement then becomes the naturalized condition of virtually all nonmale, nonwhite, nonheterosexual, non-Western, nonhuman, and nonaffluent persons.

If Daly is correct, (1) all forms of oppression and exploitation take sex as their essential archetype. Hence, (2) feminism's emancipatory objectives cannot be achieved without serious attention paid to the ways in which sexism informs the other 'isms' that sustain a patriarchy that (3) is oppressive as well as violent in all of its past and present forms, and in which its primarily male beneficiaries continue to have a stake. Finally, (4) to whatever extent sex oppression is archetypal of all oppression so too must heterosexuality become compulsory, that is, a characteristic of the heteropatriarchal way we define sex and hence sexuality such that other sexualities are either completely occluded or, when acknowledged at all, identified with the perverse and unnatural.

Within compulsory heterosexuality, an action only counts as sex if (1) it involves a penis penetrating a vagina, (2) it is compulsory in the sense that no other option counts as an acceptable choice, and (3) it negatively associates abstinence from such sex with frigidity, mental illness, or homosexuality. On this view, then, the domination of sex cannot be accomplished without the domination of sexual identity and sexuality. Given the combined compulsion of (1)–(3), moreover, it is little wonder that Daly characterizes heteropatriarchy as rapist and thus resistance to patriarchy requires resistance to heterosexuality. Radical feminism thus becomes radical lesbian feminism as both the rejection of male sexual access and as the radical embrace of the value, lives, and bodies of women. Such is the meaning of dismantling the master's house with new tools. That is, radical lesbian feminists (or radical feminist lesbians) argue for two distinct but intimately related claims: that compulsory heterosexuality must be repudiated, and that lesbianism must be theorized as a positive, life-affirming, morally responsible choice.

One common objection to this view is that it is possible to embrace the political goals of radical lesbian feminism without jettisoning sexual relationships with sympathetic men. However, because even the most sympathetic of men are still disproportionately empowered in heterosexual relationships, such intercourse cannot fail to remain compulsory, argues Daly (among others), and hence self-defeating. This is not to say that heterosexual sex is inherently harmful to women, but rather that because male sexual access is one of the primary sources of women's oppression, cauterizing this source of harm is instrumental to ending heteropatriarchal power. Indeed, whether heterosexual sex is inherently harmful or not is beside the point given that, from a radical lesbian feminist point of view, it is hard to imagine conditions under which the potential for harm could be sufficiently ameliorated. Nothing, moreover, requires us to theorize men's 'essential' natures as aggressive or women's as passive; it is enough to recognize that history and culture are so saturated with the glorification of male sexual conquest and violence, that the prospects for change and the likelihood of that change being initiated by men is almost nil. Like Marx's conception of the permanent revolution of the working class, radical lesbian sexual separatism constitutes more than a political strategy; it constitutes the bedrock of a growing lesbian culture, a revolutionary positive choice to live one's life among women, not among women who are waiting for men to learn to behave better. Indeed, to opt merely for the latter is tantamount to abandoning in advance the worth of having chosen the former.

Feminists such as Sarah Hoagland argue that a constitutive feature of a lesbian ethic includes the repudiation of all forms of domination and subordination. To the extent, then, that heterosexuality is entrenched in and defined by these values, it can find few comfortable bedfellows – literally or figuratively – among radical lesbian feminists. Is a nonoppressive, non-dominating heterosexuality possible? In theory, no doubt it is. But, as feminist philosopher Marilyn Frye points out, the practical likelihood of this seems illusive, given the stakes men are likely to continue to have in heteropatriarchy. Lesbians, she argues, are simply impossible within the heteropatriarchal framework. Frye goes on to argue that given the biological, social, and political expectations attached to reproduction and child rearing, it is no wonder that exercising control over pregnancy, birth, and lactation becomes imperative for men who, without control over these, would experience the unjust withdrawal of privilege. As Frederick Engels had shown more than a century earlier, men's investment in sexual domination is rooted first and foremost in men's investment in the products of their own (hetero)sexuality, namely, the offspring who represent a form of property, a reflection of self-interest, and the potential for future inheritance. To deny men sexual access, then, is effectively to deny them the opportunity to control the conditions of women's reproductive lives, and hence of women's lives in general. Note that here too, an element of Marxist/socialist critique helps to forge the tools of radical feminist critique, especially, concepts like 'property,' 'product,' and 'inheritance.'

The affirmation of women's lives, bodies, values, and choices forms the second critical axis of any radical lesbian feminist theory. Yet, while this affirmation is often theorized as necessitating some separation from men – separatism – it is unclear what exactly this means, for whom it is realistically practicable, and to what extent it requires physical and material independence from both men and patriarchal institutions. This much, however, seems clear: the choice to no longer engage with men sexually constitutes a minimal criterion of separatism, whereas the choice to embrace a life composed primarily or even exclusively of 'women loving women' goes much further and is more difficult to realize in practice. There are several reasons why this separation is often impracticable: first, because we live in a multinational capitalist economy, wherewithal is economic wherewithal. Hence, the creation of any women-only space, permanently (e.g., a communal organic farm) or temporarily (e.g., women's music festivals) is an economic undertaking that is always at risk of excluding

some in favor of those who can afford separatism. As some lesbian feminist women of color point out, separatism continues to be a luxury of whiteness to whatever extent economic affluence does. Second, because every economic, social, cultural, medical, and political institution, to which we have access, is imbued with heteropatriarchal values, it remains somewhat obscure what separatism could mean short of wholesale abandonment. Third, short of the invention of parthenogenesis, a contribution of male-produced sperm is still requisite to human reproduction.

Such a characterization of separatism may, however, be guilty of straw fallacy; maybe it appears impracticable only because its parameters have been defined too narrowly or too simplistically. Whether this is so, however, depends precisely on what kind of conceptual work such a notion is supposed to be able to do, for whom, and for what duration. The range of other potentially separatist choices is very wide and includes at least the following: the formation of one's primary affective relationships, sexual or otherwise, with women and girls only; choosing a woman-centered spirituality; adopting and rearing girls within a lesbian culture; creating woman-centered works of art, dance, music, poetry, comedy, film, philosophy, and literature; rejecting male-centered practices such as nonhuman animal flesh consumption, hunting, or other hierarchical forms of organization; choosing political affiliations and activities that are consistent with a woman-centered life. The difficulty here, of course, is twofold: first, how ought what counts as woman-centered or lesbian (and do these mean the same thing?) be defined? How ought the reproduction of the same disabling forms of heteropatriarchal oppression be avoided within such a lesbian culture? (Should preoperative transsexual men-to-women be allowed into women-only spaces? Can a woman-centered life include rearing sons? Should radical lesbian feminists form political alliances with gay men?)

Second, if 'separatist' is itself subject to evolving cultural, ethnic, and historical interpretation, surely it is as dangerous to treat it ahistorically as it is to treat 'woman,' 'man,' and so on. What, then, does separatism require beyond the refusal of sexual access by men? Can lifelong lesbians who do not identify as feminists be considered separatists? Can women who have left long-standing heterosexual relationships and now identify as lesbian qualify as separatists in midlife? How separate is separate? The point, of course, is that many questions attend the endeavor to create new tools, and much more attention needs to be paid to the ways in which oppressions related not only to sex, but to race, ability, age, gender expression, and class intersect with compulsory heterosexuality. However, the questions that continue to haunt lesbian separatism do not detract from the power of radical lesbian feminist analyses of heteropatriarchal institutions. Indeed, questions about any utopic vision of the future – lesbian feminist or otherwise – are central not only to feminist theory but also to environmental philosophy. Whether future environmental conditions can sustain communities in any fashion desirable to humans and for a diversity of nonhuman life is an open question. If anything, in other words, makes the vision of a lesbian utopia seem naïve or unrealistic, it may not be whether separatism should or should be a goal; it may be whether future conditions for life are sustainable at all.

All human communities are threatened by climate change – some more than others. An ecologically oriented radical lesbian feminism need not set aside issues concerning heteropatriarchal privilege such as the attractions of separatism, transsexual identities, and so on, but must theorize them in light of the reemergence of a heteropatriarchal power that is unprecedented in human history, namely, that of multinational corporations that effectively control the conditions of life itself. I have made my focus industrialized fossil fuel extraction, but we could also have included the power wielded by corporations like Monsanto, factory farming, and the pharmaceutical giants. In any of these cases, radical lesbian feminism occupies, I suggest, an excellent position not only to evaluate the heteropatriarchal character of these global corporate enterprises, but to offer something essential in light of the threat posed by climate change, namely, the conceptual, moral, and practicable framework for the revolution necessary to end what amounts to corporate fascism, and with it a trajectory toward ecocide.

See also: Ecofeminism; Feminist Political Theory and Political Science; Gay and Lesbian Movements;
Gender History; Gender, Economics of; Heterosexism and Homophobia; Human Ecology; Lesbians: Social and Economic Situation; Sexism; Sexual Minority Families of Choice and Family Networks: We Are Family; Sexual Orientation: Historical and Social Construction; Social Movements: A Social Psychological Perspective; Social Protest.

Bibliography

Allen, J. (Ed.), 1986. Lesbian Philosophy: Explorations. Institute for Lesbian Studies, Palo Alto, CA.

Balsamo, A., 1999. Forms of technological embodiment: reading the body in contemporary culture. In: Price, Janet, Shildrick, Margrit (Eds.), Feminist Theory and the Body: A Reader. Edinburgh UP, Edinburgh, pp. 278–290.

Butler, J., 1990. Gender Trouble. Routledge, New York.

Butler, J., 1993. Bodies that Matter: On the Discursive Limits of "Sex". Routledge, New York.

Butler, J., 2003. Performative acts and gender constitution: an essay in phenomenology and feminist theory. In: McCann, Carole R., Kim, Seung-Kyung (Eds.), Feminist Theory Reader: Local and Global Perspectives. Routledge, New York, pp. 415–427.

Card, C. (Ed.), 1991. Feminist Ethics. University of Kansas Press, Lawrence, KS.

Clark, D., 1993. Commodity lesbianism. In: Ableove, Henry, Barale, Michele Aina, Halperin, David (Eds.), The Gay and Lesbian Studies; Reader. Routledge, New York, pp. 186–201.

Daly, M., 1998. Quintessence: A Radical Feminist Manifesto. Beacon Press, Boston.

Frye, M., 1983. The Politics of Reality: Essays in Feminist Theory. Crossing Press, Freedom, CA.

Halberstam, J., 1999. F2M: the making of female masculinity. In: Price, Janet, Shildrick, Margrit (Eds.), Feminist Theory and the Body: A Reader. Edinburgh UP, Edinburgh, pp. 125–133.

Hartmann, H., 1990. Capitalism, patriarchy, and job segregation by sex. In: Hansen, K.V., Phillipson, I.J. (Eds.), Women, Class, and the Feminist Imagination. Temple University Press, Philadelphia, PA.

Hoagland, S., 1988. Lesbian Ethics: Toward New Value. Institute of Lesbian Studies, Palo Alto, CA.

Hoagland, S.L., Frye, M. (Eds.), 2000. Feminist Interpretations of Mary Daly. Penn State Press, College Park, MD.

Hooks, B., 2000. Feminist Theory from Margin to Center. South End Press, Cambridge, MA.

Lee, W., 2001. On Marx: A Critical Introduction. Wadsworth Press. History of Philosophy Series.

Lee, W., 2003. But one man opens his seeing eye: the epistemic ubiquity of anthropomorphizing and its implications for conceptions of gender, race, and sexual identity. In: Heyes, C. (Ed.), The Grammar of Politics. Cornell University Press, Ithaca, pp. 167–185 (2005. Translated into Spanish and republished in Mujeres and Ecology: History, Thought, and Society. Ed. Alicia H. Puleo, Cristina Segura, and Maria Luisa Canana).

Lee, W., 2010. Contemporary Feminist Theory and Activism: Six Global Issues. In: Sager, A. (Ed.). Broadview Publishing.

Lee, W., 2011. Commentary on Eric Cave's "Marital pluralism: making marriage safer for love," sex, love, and friendship. In: Soble, A. (Ed.), The Society for Sex and Love, vol. 2. Paragon House.

Lee, W., 2001. Queering ecological feminism: Erotophobia, commodification, art, and lesbian identity. Co-authored with Laura Dow Ethics and the Environment 6 (2), 1–21.

Lee-Lampshire, W., 1999. Spilling all over the wide fields of our passions. Hypatia 14 (3), 1–16.

Lorde, A., 1984. Uses of the erotic: the erotic as power. In: Lorde, A. (Ed.), Sister Outsider. Crossing Press, Trumansburg, New York.

Lugones, M., 1987. Playfulness, world-traveling, and loving perception. Hypatia 2 (2), 3–19.

Martin, E., 1999. The egg and the sperm: how science has constructed a Romance based on stereotypical male-female roles. In: Price, J., Shildrick, M. (Eds.), Feminist Theory and the Body: A Reader. Edinburgh University Press, Edinburgh, pp. 179–189.

Rich, A., 1994. Compulsory heterosexuality and lesbian existence. In: Jaggar, A. (Ed.), Living with Contradictions. Westview Press, Boulder, CO.

Wilchins, R.A., 1997. Read My Lips: Sexual Subversion and the End of Gender. Firebrand Books, Ithaca, New York.

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF
JAMES D. WRIGHT
University of Central Florida, Orlando, FL, USA

VOLUME 9

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser–Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 9

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 9 *xlv*

VOLUME 9

Fertility and Culture: Anthropological Insights 1
Caroline H Bledsoe

Fertility and Social Networks 6
Hans-Peter Kohler

Fertility Change: Quantum and Tempo 10
John Bongaarts

Fertility of Single and Cohabiting Women 13
Lawrence L Wu

Fertility Theory 19
Nicoletta Balbo

Fertility Theory: Embodied-Capital Theory of Life History Evolution 28
Hillard S Kaplan, John A Bock, and Paul L Hooper

Fertility Theory: Theory of Intergenerational Wealth Flows 35
Kristin Snopkowski and Hillard Kaplan

Fertility Theory: Theory of Life History Evolution 40
Gert Stulp and Louise Barrett

Fertility Transition: China 46
Wang Feng

Fertility Transition: Cultural Explanations 49
Jennifer Johnson-Hanks

Fertility Transition: East Asia 55
Noriko O Tsuya

Fertility Transition: Economic Explanations 60
Paul T Schultz

Fertility Transition: Latin America and the Caribbean 68
Fatima Juárez and Cecilia Gayet

Fertility Transition: Middle East and North Africa 73
John B Casterline

Fertility Transition: South Asia 80
John Cleland

Fertility Transition: Southeast Asia 84
Patrick Heuveline and Charles Hirschman

Fertility Transition: Sub-Saharan Africa 92
David Shapiro

Fertility Trends in the Formerly Socialist Countries of Europe 99
Elizabeth Brainerd

Fetishism 105
Roger Sansi Roca

Feudalism 111
Levi Roach

Feyerabend, Paul (1924–94) 117
Daniel Kuby

Fiducial and Structural Statistical Inference 124
Donald AS Fraser

Field Experiments 128
Ruth Ditlmann and Elizabeth Levy Paluck

Field Observational Research in Anthropology and Sociology 135
Roger Sanjek

Field Theory 140
Gisèle Sapiro

Fieldwork in Social and Cultural Anthropology 149
Moshe Shokeid

Fieldwork: Ethical Aspects 153
Patricia Marshall

Film and Video Industry 156
Douglas Gomery

Film: Genres and Genre Theory 160
Ib Bondebjerg

Film History 165
Charles O'Brien

Finance, Anthropology of 171
Karen Ho

Finance, Geography of 177
Gordon L Clark

Financial Crises, History of 183
Giovanni Zanalda

Financial Regulation 191
Lucia Quaglia

Fire, Human Use, and Consequences 197
Johan Goudsblom

First Language Acquisition, Developmental Psychology of 201
RA Berman

First Language Acquisition, Linguistic Theory of 206
Rosalind Thornton

First World War, The 212
Jay Winter and Charles J Stille

Fisher, Irving (1867–1947) 218
William J Barber

Fisher, Ronald A (1890–1962) 222
AWF Edwards

Fishing Societies 226
Gisli Palsson

Five Factor Model of Personality, Assessment of 230
Michael J Boudreaux and Daniel J Ozer

Five Factor Model of Personality, Facets of 236
Ryan Y Hong and Sampo V Paunonen

Five Factor Model of Personality, Personality Disorder 242
Cristina Crego and Thomas A Widiger

Five Factor Model of Personality, Universality of 249
Velichko H Fetvadjiev and Fons JR van de Vijver

Flashbulb Memories 254
Daniel L Greenberg

Flow in Motivational Psychology 260
Jeanne Nakamura and Matthew Dubin

Fluency and Voice, Disorders of 266
David B Rosenfield

Focus Groups 274
Martha Ann Carey

Folklore 280
Barbro Klein

Food Deserts in US Cities 285
Amy M Donley and Marie C Gualtieri

Food in Anthropology 290
Steffan I Ayora-Diaz

Food Preferences, Psychology and Physiology of 296
Paul Rozin

Food Production, Origins of 300
Gary W Crawford

Food Security and 'Green Revolution' 307
William G Moseley

Food, Sociology of 311
Alain Drouard

Forecasting 316
Kenneth C Land

Foreign Investment: Direct 324
Klaus E Meyer

Foreign Language Teaching and Learning 327
Aleidine J Moeller and Theresa Catalano

Forensic Archaeology 333
John J Schultz

Forensic Genetic Databases: Ethical and Social Dimensions 339
Barbara Prainsack and Jay D Aronson

Forensic Psychiatry 346
John Bradford and Graham Glancy

Forensic Psychology 351
Dap Louw

Formal Methods of Cultural Analysis 357
John W Mohr and Craig Rawlings

Foucault, Michel (1926–84) 368
Arpad Szakolczai

Fourier Analysis 374
Rainer Picard

Frankfurt School: Institute for Social Research 380
Dustin B Garlitz and Hans-Herbert Kögler

Freedom, History of 387
Frank Lovett

Freedom of the Press 392
Kyu Ho Youm

Freedom: Political 398
Timothy O'Hagan

Free Will and Action 404
Michael Smith

French Revolution, The 409
L Hunt

Freud, Sigmund (1856–1939) 413
Clifford Yorke

Freyre, Gilberto (1900–87) 420
Peter Burke

Friendship, Anthropology of 425
Bettina Beer and Don Gardner

Friendship During Adolescence and Cultural Variations 432
Hildegunn Fandrem

Friendship During Childhood and Cultural Variations 442
Kelsey B Borner, Laurie A Gayes, and Jeffrey A Hall

Friendship During Infancy and Early Childhood and Cultural Variations 448
Kenneth H Rubin, Maureen A Wimsatt, Sara Heverly-Fitt, and Matthew G Barstead

Friendship Experiences of Lesbian, Gay, Bisexual, and Transgender Individuals 454
MP Galupo

Fromm, Erich (1900–80) 458
Iga Mergler, Neil McLaughlin, and Ismaël Traoré

Frontiers in History 464
Alfred J Rieber

Functional Approaches to Grammar 470
William Croft

Functional Brain Imaging of Language processes 476
Brigitte Stemmer and Frank A Rodden

Functional Data Analysis 514
James O Ramsay and Bernard W Silverman

Functional Explanation: Philosophical Aspects 519
Graham Frank Macdonald

Functional Magnetic Resonance Imagery, Analysis of 525
Martin A Lindquist

Functionalism in Anthropology 532
Joan Vincent

Functionalism, History of 536
Richard Münch

Functioning, Disability and Health, International Classification of 543
Alarcos Cieza and Jerome E Bickenbach

Funding of Social Science, History of 550
Thomas König

Future Thinking 555
Stefania de Vito

Future Time Perspective and Motivation 561
Willy Lens and Rachel Seginer

Gaia 567
Stephan Harding

Galton, Sir Francis (1822–1911) 569
John CW Waller

Game Theory 573
Guillermo Owen

Game Theory: Noncooperative Games 582
Eric van Damme

Gangs, Sociology of 592
Lorine A Hughes and James F Short, Jr

Garfinkel, Harold (1917–2011) 598
Stacy L Burns

Gastrointestinal Diseases: Psycho-social Aspects 605
Antonina Mikocka-Walus and Lesley Graff

Gated Communities 612
Renaud Le Goix

Gatekeepers in Social Science 618
Barbara Hoenig

Gauss, Carl Friedrich (1777–1855) 623
Richard W Farebrother

Gay and Bisexual Latinos in US Gay Culture 628
Xavier L Guadalupe-Diaz

Gay and Lesbian Movements 632
Amy L Stone

Gay and Lesbian Tourism 637
Kevin Markwell and Gordon Waitt

Gay Men's Emancipation Movement in Germany: 1897–1933 644
Jason Crouthamel

Gay, Lesbian, and Bisexual Youth 649
Ritch C Savin-Williams and Kenneth M Cohen

Gays, Lesbians, and Sports: No Longer on the Sideline 656
Eric Anderson and Mark McCormack

Geertz, Clifford (1926–2006) 661
Fred Inglis

Geist (Spirit): History of the Concept 666
Eric R Lybeck

Gellner, Ernest (1925–95) 671
Chris Hann

Gender and Academic Motivation 677
Charlotte A Agger and Judith L Meece

Gender and Education 682
Erin Pahlke and Priscilla Goble

Gender and Feminist Studies in Geography 689
Janice Monk

Gender and Health Care 694
Ute Habel and Birgit Derntl

Gender and Language: Cultural Concerns 698
Ruth Wodak

Gender and Physical Health 704
Gerdi Weidner

Gender and Place 709
Daphne Spain

Gender and Social Research 715
Fatima A Castillo

Gender and Technology: From Exclusion to Inclusion? 723
Vivian A Lagesen

Gender and the Law 729
Martha Chamallas

Gender and Women's Studies, Applied Research on 736
Mitchell Brown and Regina Moorer

Gender Development: Cultural Differences 742
Deborah L Best and Kelsey L Luvender

Gender Differences in Personality and Social Behavior 750
Marco Del Giudice

Gender, Economics of 757
Francine D Blau

Gender, Feminism, and Sexuality in Archaeological Studies 764
Ericka Engelstad

Gender History 768
Kathleen Canning

Gender, Time-Use, and Fertility Recovery in Industrialized Countries 775
Javier García-Manglano, Natalia Nollenberger, and Almudena Sevilla

Gender: Anthropological Aspects 781
Beverly Stoeltje

Gender: Gendering of Categories 788
Zorica Siročić

Gender: Grammatical 795
Greville G Corbett

Gendered Care and Social Work Practice 800
Joan Orme

Genealogy in Anthropology 806
Mary Bouquet

Gene–Culture Coevolution 813
Jeremy R Kendal

Gene–Environment Interactions in Health and Well-Being 819
Jason Schnittker

Gene–Environment Interplay 825
Jonathan Daw and Jason D Boardman

Generalizability Theory 834
Georg E Matt and Marisa Sklar

Generalization: Conceptions in the Social Sciences 839
Thomas D Cook

Generalized Linear Mixed Models 845
Charles E McCulloch and John M Neuhaus

Generation in Anthropology 853
Sarah Lamb

Generations in History 857
Anthony Esler

Generations: Political 861
Paul R Abramson

Generations, Sociology of 866
Shmuel N Eisenstadt and Bryan S Turner

Generative Grammar 871
Manfred Bierwisch

Genes and Behavior: Lessons from the Fruit Fly 879
Ralph J Greenspan

Genetic Admixture 887
Oscar Lao and Mannis van Oven

Genetic Ancestry Testing 898
Jessica Bardill and Nanibaa' A Garrison

Genetic Approaches to Memory 905
Alcino J Silva

Genetic Counseling: Historical, Ethical, and Practical Aspects 908
Jörg Schmidtke, Heather Skirton, Irmgard Nippert, and Gerhard Wolff

Genetic Engineering 915
David J Harris

Genetic Factors in Organizational Psychology 919
Nikolaos Dimotakis, Jeff Schatten, and Richard Arvey

Genetic Screening for Disease 926
Lidewij Henneman and Danielle Rosa Martina Timmermans

Genetic Studies of Behavior: Methodology 932
Carol A Prescott, Thalida Em Arpawong, and John J McArdle

Genetics and Anthropology 939
John H Moore and Jonathan Marks

Genetics and Disaster Victim Identification 945
Caroline Bennett

Genetics and Economic Behavior 951
Björn Wallace

Genetics and Forensics 955
Simon A Cole and Barbara Prainsack

Genetics and Indigenous Communities: Ethical Issues 962
Emma E Kowal

Genetics and Social Justice 969
Lindsay A Smith

Genetics and Society 975
Peter Glasner

Genetics and Sociology 984
Dalton Conley, Jason Fletcher, and Christopher Dawes

Genetics and the Media 990
Kate O'Riordan

Genetics, Disease, and Reproduction 994
Carrie Friese

Genetics: Legal Aspects 1001
Shawn HE Harmon

Genetics of Personality 1009
Sharon Niv and Laura A Baker

Genetics: The New Genetics 1015
Karen-Sue Taussig

This page intentionally left blank

CONTRIBUTORS TO VOLUME 9

Paul R. Abramson
Michigan State University, East Lansing, MI, USA

Charlotte A. Agger
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Eric Anderson
University of Winchester, Winchester, UK

Jay D. Aronson
Carnegie Mellon University, Pittsburgh, PA, USA

Thalida Em Arpawong
University of Southern California, Los Angeles, CA, USA

Richard Arvey
School of Business, National University of Singapore, Singapore

Steffan I. Ayora-Diaz
Universidad Autónoma de Yucatán, Mérida, Yucatán, Mexico

Laura A. Baker
University of Southern California, Los Angeles, CA, USA

Nicoletta Balbo
Dondena Centre, Bocconi University, Milan, Italy

William J. Barber
Wesleyan University, Middletown, CT, USA

Jessica Bardill
Stanford Center for Biomedical Ethics, Stanford, CA, USA

Louise Barrett
University of Lethbridge, AB, Canada; and UNISA, Applied Behavioural Ecology and Ecosystems Research Unit, Johannesburg, South Africa

Matthew G. Barstead
University of Maryland, College Park, MD, USA

Bettina Beer
Universität Luzern, Switzerland

Caroline Bennett
University of Kent, Canterbury, Kent, UK

R.A. Berman
Tel Aviv University, Israel

Deborah L. Best
Wake Forest University, Winston-Salem, NC, USA

Jerome E. Bickenbach
Leiter Disability Policy Group, Schweizer Paraplegiker-Forschung, Nottwil, Switzerland

Manfred Bierwisch
Humboldt Universität Berlin, Berlin, Germany

Francine D. Blau
Cornell University, Ithaca, NY, USA

Caroline H. Bledsoe
Northwestern University, Evanston, IL, USA

Jason D. Boardman
University of Colorado Boulder, Boulder, CO, USA

John A. Bock
California State University, Fullerton, CA, USA

Ib Bondebjerg
University of Copenhagen, Copenhagen S, Denmark

John Bongaarts
The Population Council, New York, NY, USA

Kelsey B. Borner
University of Kansas, Lawrence, KS, USA

Michael J. Boudreaux
University of California Riverside, Riverside, CA, USA

Mary Bouquet
Utrecht University, Utrecht, The Netherlands

John Bradford
University of Ottawa, Ottawa, ON, Canada

Elizabeth Brainerd
Brandeis University, Waltham, MA, USA

Mitchell Brown
Auburn University, Auburn, AL, USA

Peter Burke
Emmanuel College, Cambridge, UK

Kathleen Canning
University of Michigan, Ann Arbor, MI, USA

Martha Ann Carey
Kells Consulting, Media, PA, USA

John B. Casterline
Institute for Population Research, The Ohio State University, Columbus, OH, USA

Fatima A. Castillo
University of the Philippines Manila, Quezon City, Philippines

Theresa Catalano
Teaching, Learning and Teacher Education, University of Nebraska-Lincoln, Lincoln, NE, USA

Martha Chamallas
The Ohio State University, Moritz College of Law, Columbus, OH, USA

Alarcos Cieza
University of Southampton, Southampton, UK

Gordon L. Clark
Smith School of Enterprise and the Environment, University of Oxford, Oxford, UK; and Monash University, Caulfield, VIC, Australia

John Cleland
University of London, London, UK; and London School of Hygiene & Tropical Medicine, London, UK

Kenneth M. Cohen
Cornell University, Ithaca, NY, USA

Simon A. Cole
University of California, Irvine, CA, USA

Dalton Conley
New York University, New York, NY, USA

Thomas D. Cook
Northwestern University, Evanston, IL, USA

Greville G. Corbett
University of Surrey, Guildford, UK

Gary W. Crawford
University of Toronto Mississauga, Mississauga, ON, Canada

Cristina Crego
University of Kentucky, Lexington, KY, USA

William Croft
University of New Mexico, Albuquerque, NM, USA

Jason Crouthamel
Grand Valley State University, Allendale, MI, USA

Jonathan Daw
University of Colorado Boulder, Boulder, CO, USA

Christopher Dawes
New York University, New York, NY, USA

Marco Del Giudice
University of New Mexico, Albuquerque, NM, USA

Birgit Derntl
RWTH Aachen University, Aachen, Germany

Stefania de Vito
University of Edinburgh, Edinburgh, UK; and Suor Orsola Benincasa University, Naples, Italy

Nikolaos Dimotakis
J. Mack Robinson College of Business, Georgia State University, Atlanta, GA, USA

Ruth Ditlmann
WZB Berlin Social Science Center, Berlin, Germany

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Alain Drouard
National Center for Scientific Research, Paris, France

Matthew Dubin
Claremont Graduate University, Claremont, CA, USA

A.W.F. Edwards
Gonville and Caius College, Cambridge, UK

Shmuel N. Eisenstadt
Van Leer Jerusalem Institute, Jerusalem, Israel

Ericka Engelstad
Institute of Archaeology and Social Anthropology, University of Tromsø, Tromsø, Norway

Anthony Esler
College of William and Mary, Williamsburg, VA, USA

Hildegunn Fandrem
University of Stavanger, Stavanger, Norway

Richard W. Farebrother
Victoria University of Manchester, Manchester, UK

Wang Feng
University of California, Irvine, CA, USA

Velichko H. Fetvadjiev
University of Pretoria, Pretoria, South Africa

Jason Fletcher
La Follette School of Public Affairs, University of Wisconsin—Madison, Madison, WI, USA

Donald A.S. Fraser
University of Toronto, Toronto, ON, Canada

Carrie Friese
London School of Economics and Political Science, London, UK

M.P. Galupo
Towson University, Towson, MD, USA

Javier García-Manglano
Nuffield College, University of Oxford, Oxford, UK

Don Gardner
Universität Luzern, Switzerland; and CASS, Australian National University, ACT, Australia

Dustin B. Garlitz
University of South Florida, Tampa, FL, USA

Nanibaa′ A. Garrison
Stanford Center for Biomedical Ethics, Stanford, CA, USA

Laurie A. Gayes
University of Kansas, Lawrence, KS, USA

Cecilia Gayet
FLACSO México, México D.F., Mexico

Graham Glancy
University of Ottawa, Ottawa, ON, Canada

Peter Glasner
Cardiff University, Cardiff, UK

Priscilla Goble
School of Social and Family Dynamics, Arizona State University, Tempe, AZ, USA

Renaud Le Goix
University Paris 1 Panthéon-Sorbonne, UMR Géographie-cités, Paris, France

Douglas Gomery
University of Maryland, College Park, MD, USA

Johan Goudsblom
University of Amsterdam, Amsterdam, The Netherlands

Lesley Graff
Department of Clinical Health Psychology, Faculty of Medicine, University of Manitoba, Winnipeg, MB, Canada

Daniel L. Greenberg
College of Charleston, Charleston, SC, USA

Ralph J. Greenspan
University of California, San Diego, La Jolla, CA, USA

Xavier L. Guadalupe-Diaz
Framingham State University, MA, USA

Marie C. Gualtieri
University of Central Florida, Orlando, FL, USA

Ute Habel
RWTH Aachen University, Aachen, Germany

Jeffrey A. Hall
University of Kansas, Lawrence, KS, USA

Chris Hann
Max Planck Institute for Social Anthropology, Halle/Saale, Germany

Stephan Harding
Schumacher College, Dartington, Devon, UK

Shawn H.E. Harmon
University of Edinburgh, Edinburgh, United Kingdom

David J. Harris
Boston Children's Hospital, Boston, MA, USA

Lidewij Henneman
VU University Medical Center, Amsterdam, The Netherlands

Patrick Heuveline
University of California, Los Angeles, CA, USA

Sara Heverly-Fitt
University of Maryland, College Park, MD, USA

Charles Hirschman
University of Washington, Seattle, WA, USA

Karen Ho
University of Minnesota, MN, USA

Barbara Hoenig
Innsbruck University, Innsbruck, Austria

Ryan Y. Hong
National University of Singapore, Singapore

Paul L. Hooper
Santa Fe Institute, Santa Fe, NM, USA; and Emory University, Atlanta, GA, USA

Lorine A. Hughes
University of Nebraska at Omaha, Lincoln, NE, USA

L. Hunt
University of California, Los Angeles, CA, USA

Fred Inglis
University of Sheffield, Sheffield, UK; University of Warwick England, UK; and School of Social Science, Institute of Advanced Study, Princeton, NJ, USA

Jennifer Johnson-Hanks
University of California, Berkeley, Berkeley, CA, USA

Fatima Juárez
El Colegio de México, México D.F., Mexico

Hillard S. Kaplan
University of New Mexico, Albuquerque, NM, USA

Hillard Kaplan
University of New Mexico, Albuquerque, NM, USA

Jeremy R. Kendal
Durham University, Durham, UK

Barbro Klein
Swedish Collegium for Advanced Study (SCAS), Uppsala, Sweden

Hans-Herbert Kögler
University of North Florida, Jacksonville, FL, USA

Hans-Peter Kohler
Population Studies Center, University of Pennsylvania, Philadelphia, PA, USA

Thomas König
Vienna Science and Technology Fund, Vienna, Austria

Emma E. Kowal
University of Melbourne, VIC, Australia

Daniel Kuby
University of Vienna, Vienna, Austria

Vivian A. Lagesen
Norwegian University of Science and Technology, Trondheim, Norway

Sarah Lamb
Brandeis University, Waltham, MA, USA

Kenneth C. Land
Department of Sociology, Duke University, Durham, NC, USA

Oscar Lao
Erasmus MC — University Medical Center Rotterdam, Rotterdam, The Netherlands

Stacy L. Burns
Loyola Marymount University, Los Angeles, CA, USA

Willy Lens[†]
University of Leuven, Leuven, Belgium; and University of the Free State, Bloemfontein, South Africa

Martin A. Lindquist
Johns Hopkins University, Baltimore, MD, USA

Dap Louw
University of the Orange Free State, Bloemfontein, South Africa

Frank Lovett
Washington University in St. Louis, St. Louis, MO, USA

Kelsey L. Luvender
Wake Forest University, Winston-Salem, NC, USA

Eric R. Lybeck
Girton College, University of Cambridge, Cambridge, UK

Graham Frank Macdonald
University of Canterbury, Christchurch, New Zealand; and University of Manchester, Manchester, UK

Jonathan Marks
University of North Carolina—Charlotte, Charlotte, NC, USA

Kevin Markwell
Southern Cross University, NSW, Australia

Patricia Marshall
Loyola University of Chicago, Maywood, IL, USA

Georg E. Matt
San Diego State University, San Diego, CA, USA

John J. McArdle
University of Southern California, Los Angeles, CA, USA

Mark McCormack
Durham University, Durham, UK

Charles E. McCulloch
University of California, San Francisco, San Francisco, CA, USA

Neil McLaughlin
McMaster University, Hamilton, ON, Canada

Judith L. Meece
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Iga Mergler
McMaster University, Hamilton, ON, Canada

[†] Deceased.

Klaus E. Meyer
China Europe International Business School, Pudong, Shanghai, The People's Republic of China

Antonina Mikocka-Walus
Mental Health and Addiction Research Group, Department of Health Sciences, University of York, York, UK

Aleidine J. Moeller
Teaching, Learning and Teacher Education, University of Nebraska-Lincoln, Lincoln, NE, USA

John W. Mohr
University of California, Santa Barbara, Santa Barbara, CA, USA

Janice Monk
University of Arizona, Tucson, AZ, USA

John H. Moore
University of Florida, Gainesville, FL, USA

Regina Moorer
Auburn University, Auburn, AL, USA

William G. Moseley
Macalester College, Saint Paul, MN, USA

Richard Münch
Otto-Friedrich-Universität Bamberg, Bamberg, Germany

Jeanne Nakamura
Claremont Graduate University, Claremont, CA, USA

John M. Neuhaus
University of California, San Francisco, San Francisco, CA, USA

Irmgard Nippert
Westfaelische Wilhelms-Universitaet Muenster, Muenster, Germany

Sharon Niv
University of Southern California, Los Angeles, CA, USA

Natalia Nollenberger
School of Business & Management, Queen Mary University of London, London, UK

Charles O'Brien
Film Studies/School for Studies in Art and Culture, Carleton University, Ottawa, ON, Canada

Kate O'Riordan
University of Sussex, Falmer, Brighton, UK

Joan Orme
Glasgow School of Social Work, Glasgow, Scotland, UK

Timothy O'Hagan
University of East Anglia, Norwich, UK

Guillermo Owen
Naval Postgraduate School, Monterey, CA, USA

Daniel J. Ozer
University of California Riverside, Riverside, CA, USA

Erin Pahlke
Whitman College, Walla Walla, WA, USA

Gisli Palsson
University of Iceland, Reykjavik, Iceland

Elizabeth Levy Paluck
Princeton University, Princeton, NJ, USA

Sampo V. Paunonen
University of Western Ontario, London, ON, Canada

Rainer Picard
Technische Universität Dresden, Dresden, Germany

Barbara Prainsack
King's College London, Strand, London, UK

Carol A. Prescott
University of Southern California, Los Angeles, CA, USA

Lucia Quaglia
Derwent College, University of York, York, UK

James O. Ramsay
McGill University, Montreal, QC, Canada

Craig Rawlings
University of California, Santa Barbara, Santa Barbara, CA, USA

Alfred J. Rieber
The Central European University, Budapest, Hungary

Levi Roach
University of Exeter, Exeter, UK

Frank A. Rodden
Abteilung für Persönlichkeitspsychologie und Diagnostik, University of Zurich, Psychologisches Institut, Zürich, Switzerland

David B. Rosenfield
Neurological Institute, Houston Methodist Hospital, TX, USA; and Weill Medical College, Cornell University, New York, NY, USA

Paul Rozin
University of Pennsylvania, Philadelphia, PA, USA

Kenneth H. Rubin
University of Maryland, College Park, MD, USA

Roger Sanjek
City University of New York, Flushing, USA

Roger Sansi Roca
Universitat de Barcelona, Barcelona, Spain; and Goldsmiths, University of London, Greater London, UK

Gisèle Sapiro
École des hautes études en sciences sociales, Paris, France; and CNRS (Centre Européen de Sociologie et de Science Politique), Paris, France

Ritch C. Savin-Williams
Cornell University, Ithaca, NY, USA

Jeff Schatten
J. Mack Robinson College of Business, Georgia State University, Atlanta, GA, USA

Jörg Schmidtke
Hannover Medical School, Hannover, Germany

Jason Schnittker
University of Pennsylvania, Philadelphia, PA, USA

John J. Schultz
University of Central Florida, Orlando, FL, USA

Paul T. Schultz
Yale University, New Haven, CT, USA

Rachel Seginer
University of Haifa, Haifa, Israel

Almudena Sevilla
School of Business & Management, Queen Mary University of London, London, UK

David Shapiro
The Pennsylvania State University, PA, USA

Moshe Shokeid
Tel Aviv University, Tel Aviv, Israel

James F. Short
Washington State University, Pullman, WA, USA

Alcino J. Silva
Integrative Center for Learning and Memory Brain Research Institute, UCLA Los Angeles, CA, USA

Bernard W. Silverman
University of Bristol, Bristol, UK

Zorica Siročić
University of Graz, Graz, Austria

Heather Skirton
Plymouth University, Taunton, United Kingdom

Marisa Sklar
San Diego State University/University of California San Diego Joint Doctoral Program in Clinical Psychology, VA San Diego Health Care System, La Jolla, CA, USA

Lindsay A. Smith
University of New Mexico, Albuquerque, NM, USA

Michael Smith
Princeton University, Princeton, NJ, USA

Kristin Snopkowski
Boise State University, Boise, ID, USA

Daphne Spain
University of Virginia, Charlottesville, VA, USA

Brigitte Stemmer
Université de Montréal, Montreal, QC, Canada; and Brock University, St. Catherines, ON, Canada

Charles J. Stille
Yale University, New Haven, CT, USA

Beverly Stoeltje
Indiana University, Bloomington, IN, USA

Amy L. Stone
Trinity University, San Antonio, TX, USA

Gert Stulp
University of Groningen, Groningen, The Netherlands; and London School of Hygiene and Tropical Medicine, London, UK

Arpad Szakolczai
University College, Cork, Ireland

Karen-Sue Taussig
University of Minnesota, Minneapolis, MN, USA

Rosalind Thornton
ARC Centre of Excellence in Cognition & Its Disorders (CCD), Australian Hearing Hub, Macquarie University, NSW, Australia

Danielle Rosa Martina Timmermans
VU University Medical Center, Amsterdam, The Netherlands

Ismaël Traoré
McMaster University, Hamilton, ON, Canada

Noriko O. Tsuya
Department of Economics, Keio University, Tokyo, Japan

Bryan S. Turner
The City University of New York, New York, NY, USA

Eric van Damme
Tilburg University, Tilburg, The Netherlands

Fons J.R. van de Vijver
Tilburg University, Tilburg, The Netherlands; North-West University, Potchefstroom, South Africa; and University of Queensland, St Lucia, QLD, Australia

Mannis van Oven
Erasmus MC — University Medical Center Rotterdam, Rotterdam, The Netherlands

Joan Vincent
Columbia University, New York, NY, USA

Gordon Waitt
University of Wollongong, Wollongong, NSW, Australia

Björn Wallace
Faculty of Economics, University of Cambridge, Cambridge, UK

John C.W. Waller
University College, London, UK

Gerdi Weidner
San Francisco State University, Tiburon, CA, USA

Thomas A. Widiger
University of Kentucky, Lexington, KY, USA

Maureen A. Wimsatt
University of Maryland, College Park, MD, USA

Jay Winter
Yale University, New Haven, CT, USA

Ruth Wodak
Lancaster University, Lancaster, UK

Gerhard Wolff
Goethestraße, Freiburg, Germany

Lawrence L. Wu
New York University, New York, NY, USA

Clifford Yorke
South Moreton, UK

Kyu Ho Youm
University of Oregon, Eugene, OR, USA

Giovanni Zanalda
Duke University, Durham, NC, USA

This page intentionally left blank

Fertility and Culture: Anthropological Insights

Caroline H Bledsoe, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Between the fertility event that increases a population and the cultural ideas that shape how it is perceived and analyzed lies an engaging set of issues. To grapple with them, demographically oriented approaches to fertility, and to reproduction more broadly, have increasingly sought insight from sociocultural anthropology and its longtime concern with the ideas by which people understand their worlds. The results, despite the vast differences in method and philosophical orientation that typically separate demography and anthropology, are forging new ways to look at fertility. This article addresses the recent history of the anthropology and demography relationship in studies of fertility, focusing especially on developments that take culture as a fluid and contested set of mental schemes.

In its narrowest demographic meaning, fertility refers to the production of a living child that adds one person to a population. Sociocultural anthropology's mainstream has never been interested in the question of numbers that this definition implies, at least to the extent that demography has, in its focus on population's stocks and flows. Conversely, the anthropological penchant for rich, contextualized cultural meanings has drawn limited interest from demographers seeking parsimonious ways to describe and explain fertility's quantitative patterns. Despite their profound differences in core methods and philosophical orientation, though, both demography and its 'softer' neighbors appear to have benefited mutually, in recent decades, from their collective engagement with reproduction. This article focuses on demographically oriented studies of fertility specifically, and of reproduction more broadly, that have drawn cultural insight from anthropology, which this article will take to mean sociocultural anthropology. Contributions by authors in related fields such as sociology and history that employ anthropology's ethnographic methods (participant observation, open-ended interviews, text analysis, and so on) will be noted as well. This article focuses on anthropology's classic strengths in culture and its insistence on close attention to local perceptions of demographic phenomena. It will show that between the fertility event that increases a population and the cultural ideas that shape how it is perceived and analyzed lies an engaging set of issues. (Biological anthropology's emphases on the genetic and evolutionary causal bases of fertility are amply covered elsewhere in this Encyclopedia.)

Demography seeks to identify and measure key variables related to population and to demonstrate relations among them. To frame its questions and explain its findings, it has tended to draw on disciplines such as sociology, economics, history, and statistics. In recent years, some demographers have turned as well to anthropology and its methods. One source of inspiration was the enormously important Princeton European Fertility Project (Coale and Watkins, 1986), in which the decline in marital fertility in Europe was found to be related less to conventionally expected economic factors, especially levels of income and development, than to religion, dialect, ethnicity, and region – factors the project termed 'cultural.' According to Lesthaeghe (1977), for example, Belgian villages experienced fertility decline according to their linguistic groupings. Walloon villages reached low fertility earlier than did Flemish villages, irrespective of their economic characteristics or locations within the country. The implication was that information about fertility reduction methods, the values placed on children, and so on, diffused more readily within culturally or linguistically similar regions than among those that differed significantly. The second reason why demography turned to anthropology has been an increasing interest in patterns of fertility and mortality in developing countries. As early as the 1960s, questions were emerging about the social supports for high fertility and about the tremendous variations in observed reproductive practices that researchers trained in Western quantitative disciplines felt poorly equipped to handle. Because anthropology is the social science that has so routinely dealt with non-Western societies, anthropologists' empirical work on these societies began to draw the interest of demographers. Increasingly, moreover, many demographic studies have gone on to directly incorporate the methods and concepts of the discipline they initially saw as such an exotic oddity. Many demographic studies are planned from the outset to incorporate, along with their conventional analyses of surveys and censuses, ethnographic field components and interviews. Even the largest demographic survey in history, the International Demographic and Health Surveys, began some years ago to offer a 'focus group' component in which statements and explanations, rather than simply numeric and short-answer responses, could be elicited.

How Has Sociocultural Anthropology Been Applied to Demographic Studies of Fertility?

Sociocultural anthropology, like other social sciences, seeks to understand why people think and behave the way they do; more so than most of its sister disciplines, however, anthropology has long seen thought and behavior as shaped by local socioeconomic contexts and by the power structures that determine advantage within them. In addition, anthropological work begins with the premise that the cultural frames from which our understandings of human motivation and action spring are likely partial, if not altogether wrong. They therefore

have the potential to be dangerously misleading. This is true, no matter how familiar the culture appears to be to the researcher. In fact, the more familiar the culture and the instruments that purport to describe it, the greater the risks of being misled by one's preconceptions, and the more wary the analyst should be. To a field whose practitioners start with the expectation of being wrong, the most valued methods are those that offer the most promise for capturing local people's own subjective experiences and the logic underlies their choices in the worlds in which they operate. Doing so, it is felt, should lead not simply to better understandings of the realities people experience, whether they live in close or distant places, but to a more encompassing view of one's own reality and hence to better theory. Though many anthropologists might disagree, there is no reason why Western-based quantitative methods and data should not be as useful as intensive, open-ended interviews for trying to understand experience and to unravel local logics and the insights they can lay bare. When they do study local beliefs, however, Western 'quantifying' traditions has tended to treat these beliefs as the end of the story: as objects to discover and report. Local beliefs are also frequently depicted in demographic works as 'cultural barriers' to more rational understandings, as Watkins (2001) and Peters (2010) point out, rather than as potential windows of ever-deeper insight for unraveling puzzles that Western scholarship has poorly understood.

Operationalizing Cultural Traits in Statistical Taxonomies

As promising as anthropology's approach and methods can be in demographic research, what anthropology has usually offered the demographic enterprise has largely reflected the state of its own field. In the middle of the twentieth century, for example, a particular methodological approach, sometimes called the cross-cultural method, dominated American anthropology. This was a kind of statistical taxonomy in which societies were described by traits that could be 'operationalized' – cast in the form of discrete variables that could be coded and counted in order to test quantitative hypotheses and look for cross-cultural patterns. The primary data used in this endeavor, the Human Relations Area Files (HRAF) (Murdock, 1975), contained detailed codes of the organizational features of hundreds of ethnic groups. Prevalent as well in American anthropology at this time were questions about the relationship of culture to personality from Freudian perspectives. A number of anthropologists sought to combine this theoretical agenda with cross-cultural methodologies. Harvard anthropologists led by John and Beatrice Whiting (Whiting et al., 1958), for example, argued from HRAF data that life under harsh environmental conditions required parents to protect a child by long periods of breastfeeding and a ban on sexual relations. For boys, such practices, because they were thought to produce strong psychological attachments between mother and child, were said to necessitate dramatic manhood initiation rituals that separated boys from their mothers and fostered the development of their masculinity. Though personality development itself drew limited demographic interest, there was considerable interest in the potentials of such studies to shed light on observed durations of breastfeeding and postpartum sexual abstinence, which had more obvious connections to fertility through its biological bases. Applying this kind of analysis to data on kinship and economic subsistence, demographer Frank Lorimer (1954) argued that populations like those on small islands, with highly constrained and intensively utilized resources, developed forms of social organization and customs that effectively checked family size. In Polynesia, he maintained, such customs included coitus interruptus, abortion, infanticide, the encouragement of sea voyaging by young men, and banning further reproduction to the mother of a married son.

Culture as Ideas and Mental Schemes

The taxonomic approach drew considerable interest in demography. Yet within the mainstream of anthropology itself, it drew increasing antipathy. Critics charged that this approach, in attempting to extract practices from the contexts in which they drew meaning and make them into checklists of discrete, modular traits effectively froze what were in reality multifaceted, fluid phenomena. Particularly with the impact of Clifford Geertz (1973), anthropology's increasing emphasis on the importance of culture and ideas shifted decisively away from such approaches and toward the in-depth search for shared mental schemes and symbols that gave clues to local interpretations of the world. By contrast to efforts to operationalize cultural data and to break them down into codified pieces that could be incorporated directly into a set of equivalent universal categories, most anthropologists now stress that culture cannot be broken down into separable pieces and that meaning must be understood within a context. Anthropologists also recognize that apparent cultural oddities may be important clues to some facet of a reality that the analyst has not seen and whose careful coding systems may therefore bury this reality.

Anthropologists' insistence on rescuing culture and the ideas it generates from the domain of operationalizable concepts has not always been easy to explain to other disciplines. The work of demographer John Caldwell, however, took a notable stand in its efforts to make culture figure more explicitly as ideas or attitudes. The most notable example is his description of the spread of Westernization throughout the developing world: the rise of new attitudes about work, family relations, and fertility practices, particularly through schooling and its perceived influence in changing values of authoritarian family structures, a distant and severe father, and marriage as an arrangement between elders. Caldwell (1980) describes primary school textbooks in three African countries as powerful statements of new values of family organization: parental closeness to children; kindly, understanding fathers; and love-match marriages. So powerful were these ideas said to be, that the shift to mass Western-style education may have generated a new cultural climate among all members of these populations, even those who had never attended school. These ideational changes, he argued, in turn facilitated a shift in values to the nuclear family and to goals of fewer children.

The evolutionary implications associated with traditional/modern frameworks have fallen out of favor with most sociocultural anthropologists, who reject any necessary assignation

of traditionalism or modernity to societal forms, and they recognize all societies as inherently fluid and changing. However, Caldwell's determination to take on culture as a set of shared mental schemes that might bear on fertility may have helped to pave the way for new efforts to theorize fertility and to understand the historical and social contexts that gave rise to different ways of seeing it (e.g., Johnson-Hanks et al., 2001; Charbit, 2010; Szreter and Fischer, 2010).

An engaging example that sought explicit links between belief and behavior was Cambridge historian Musallam's (1983) account of the allowable uses of contraception for men in classical Islam. Whereas Western analysts had seen Islam as opposing the use of contraceptives its practice, Musallam's emphasis on culture and the logical strands connecting it to empirical reality described a more complex picture. Contracepting, he explained, was indeed seen as an act that fell short of Islamic religious perfection. In practice, however, it allowed a surprisingly wide range of exceptions. Because contraception was allowed if it was intended to preserve health and well-being, a man could contracept if his actions were intended to protect a young nursing child from the nutritional competition posed by a new pregnancy. He could also take contraceptive measures to protect his wife, especially if she was young or sickly, from the dangers of childbirth, or simply to protect her health and beauty from the wearing effects of childbearing. In another major domain of concern, that of economic welfare, a man who found himself tempted to engage in theft to meet his financial responsibilities to his children could practice contraception to limit the number of his offspring to a manageable size. Yet exceptions emerged in debates over the ends to which one might legitimately go to safeguard his property. Because a man could not sell a slave after she became the mother to his child, using contraception was seen as a lawful measure to retain liquidity of his slave assets. Conversely, when having sex with another man's slave, it was argued that a man could use contraception to ensure that his own children would not be born into slavery. And, in parallel to this was the alleged need to safeguard the welfare of the Islamic community under conditions of warfare. If a man had sex with a woman behind enemy lines, he was expected to take contraceptive precautions, lest he risk not only losing his own children to slavery by procreating with her, but swelling the numbers of the infidels.

Culture in Flux; Culture in Contestation

Despite the intrinsic appeal of pursuing cultural pathways of logic like this, contemporary anthropologists and their like-minded demographic colleagues seldom pose such unproblematic straight lines between cultural belief and behavior. Furthermore, as Johnson-Hanks (2005) shows, the future is a far murkier place than older models of economic rationality have implied. There are surprisingly thin connections between reproductive intents at the beginning of reproductive life and later on (Johnson-Hanks and Miguel, 2011). Rather than pursuing singular cultural templates that constrain action; the future is best handled by managing flexible sets of interpretive resources as new challenges come into view (Johnson-Hanks et al., 2001). Even the most apparently unbending bodies of religious canon can be argued to have numerous potential interpretations (Musallam, 1983).

Indeed, one of the most lively examples of research that has demanded a fluid, resourceful vision of culture is the topic of male fertility. Fertility research is by far the largest subdiscipline in contemporary demography. Yet until the 1990s, women comprised are almost exclusive objects of study. Men, if they appeared at all, were shadows: partners-by-implication of those engaged in childbearing. At the same time, there are tremendous challenges posed in efforts to analyze male fertility. While some men claim they do not know how many children they have or deny connection to offspring born outside of wedlock, others claim to have fathered children they have never seen, while infertile Egyptian men in even the most apparently conservative of circumstances may seek to resort to whatever new global technologies promise to yield children, or even to fashion new masculine identities from local cultural repertoires (Inhorn, 2012). The fact that men so often try to manage aspects of parenthood can thus work to the disadvantage of children in instances where men seem to be edging out of commitments to children. But men's diverse parenting energies can also have clear advantages for children. Townsend's study of rural Botswana (2000) showed that in this place, where the demands of migrant labor were so taxing, a young man who lived full time with his children was an oddity, and hence many children grew up hardly knowing their father at all. As they grew older, however, and unable to hold such strenuous jobs in the minds, men provided considerable support for many children, including those of other men, whether these were uncles, grandfathers, or fictional cousins. As this example suggests, persistent perceptions in the international population policy world of massive male disengagement from parenthood responsibilities, along with Western cultural preoccupations with the biological side of paternity, have masked some critical variations in the cultural handling of fatherhood.

One of the most important points of debate in current sociocultural theory that is now being brought to demography is the extent to which demographic patterns stem from politics and economics versus individual agency. That is, to what extent are people's choices limited by wider powers beyond their control? Can they, on the other hand, make their own decisions about reproductive actions by resisting or evading state powers? Carefully conceived studies that analyze demographic questions through anthropological lenses increasingly suggest that both forces are invariably at work, and that cultural ideologies are often at the center of the contestation. Greenhalgh and Li's account (1995) of the heavily enforced Chinese state policy of one child per couple that was implemented in 1979 reveals a dynamic of mutual influence. As their analyses show, women began to resist the formal one-child policy, strategizing to have two children, especially if the first was a girl. To disguise their efforts to ensure a boy, women attempted to adopt out daughters, move to another city to give birth to a second child, remove state-imposed intrauterine devices from their bodies, conceal unauthorized pregnancies until it was too late for a state-required abortion, and (as ultrasound technology spread) engage in sex-selective abortions. By 1991, the state, faced with such resistance, was forced to begin to rewrite the policy. It 'encouraged' one child for all and prohibited third child, yet allowed two children if the first

was a girl. As Greenhalgh and Li caution, however, individual women may have been liberating themselves individually from unwanted daughters, but their strategies of resistance effectively reinforced the state's gender hierarchy.

The fact that culture is increasingly seen in anthropology as a resource that people may put to use is best examined in a context in which consciousness about the effects of action on fertility and reproductive volition in general have been seen as low. For this reason, some of the best examples are found in topics related to high fertility: a pattern that has been seen in demographic convention as an outcome of biology and custom and not as the calculus of conscious choice (Coale, 1973). In populations said to have had what has been called natural fertility (Henry, 1961), the demand for children is high, and the number of children women produce is assumed to be limited only by how quickly they can be supplied, given the biological rhythms resulting from specific customary regimes. In sub-Saharan Africa, for example, substantial numbers of people continue to report on surveys that they want 'all the children God gives them,' and typical reactions to suggestions to reduce fertility range from polite interest to outrage. By contrast, in societies where desires for large numbers of children have abated, couples are said to engage in parity-specific behaviors, especially the use of modern contraceptives, in order to control completed family size.

However, ethnographic accounts of practices such as infanticide, child fosterage, child abandonment, adoption, and abortion make it clear that reproduction can be subject to many means of control, and that people may readily resort to them. Whereas Western models assume, for example, that reducing fertility is the best way to address economic troubles, families in many African societies see the diverse skills and social ties that they manage to cultivate through a diverse range of children as the key to their future capacity to cope with economic and political adversity. The widespread use of child fostering is a case in point. Young children are commonly fostered to 'grannies'; later, when they are older, they may be sent to urban households that offer training in specialized apprenticeships or a chance to attend a respected school. Adults, in short, treat the original biological event of parenthood not as a secure hold on a child's future support, but as the beginning of a long, continuously negotiated relationship in which they try to influence children's obligations. While Western family planning programs in developing countries continue to emphasize child numbers and the economic woes they intensify, many African families would find this as a drastic solution. They prefer to have a number of children, cultivating differentiated opportunities for them (Bledsoe, 1994).

Another example of people's efforts to exert control over childbearing appears in the cultural reinterpretation of contraceptive technologies in sub-Saharan Africa (Bledsoe, 2002). In Western cultural perceptions, contraceptives are devices that limit fertility, and people who do not wish to limit births would be unlikely users of devices that would do so. In West Africa, however, people see contraceptives more as devices to ensure the survival of many children by spacing their births carefully than to limit them. Some of the most common contraceptive users are thus breastfeeding women who want to avoid overlapping children, one in the womb and the other nursing, but who have resumed their menses before the breastfeeding child can be weaned. Families exert prodigious efforts to control the timing and circumstances of children's births, though, again, not necessarily to limit numbers of children.

Reconciling Local Cultural Knowledge and Western Science

Pursuing threads of counterintuitive cultural logic that are opened up by scrutinizing demographic patterns can yield not simply observations about culture but empirical truths that have been lost to contemporary scientific view. The example of contraceptive use in West Africa (Bledsoe, 2002) plays out in striking ways, shedding new light on the cultural ideas that inevitably influence the development and interpretation of even the rigorous Western quantitative methods.

Contemporary Western writings, both scientific and humanistic, depict the ability to bear children as one limited by time. That is, Western culture posits a time-dependent ability to conceive as the key to achieving fertility. Accordingly, the number of children a woman may bear is calculated by dividing the total amount of time she can conceive by the interval required for each pregnancy. The use of contraception plays directly into this presumed calculus. To the extent that the use of contraception or abstinence blocks off time in this potential reproductive span, a woman's fertility will remain low. In the African cultural view, however, the ability to reproduce is limited not by time but by bodily resources. Among the forces that drain bodily resources and terminate the ability to reproduce, childbirth is by far the most powerful. Each birth takes a mounting toll on a woman's physical reserves, cumulatively 'aging' her body (or wearing it out, in many local translations), and eventually making further childbearing dangerous, irrespective of her chronological age. In this vision, aging is not the same thing as age. Although aging tends to be equated with the passage of time in Western analyses, its pace, in the African view, is contingent on the physical ordeals of the body suffers. Difficult and/or closely spaced births, combined with a poor diet, illness, and strenuous work responsibilities, can exact a disproportionate toll. In such situations, a woman who is trying to have all the children God gives her will almost certainly run up against the anatomical and energy limits of the body before she runs into any temporal boundaries on the ability to conceive. Because contraceptives allow women more opportunity to recuperate between births, and because reproductive capacity is seen as governed by taxing events rather than by time, it is clear why contraceptives can be used to produce more, rather than fewer, children and even to stem the pace of aging.

The African vision of contingent wear and cumulative physical toll over a reproductive lifetime might appear to be 'just common sense.' It is more accurate, however, to say that our increasing personal distance from the high-and-risky end of the fertility horizon, combined with our unprecedented levels of preventive care and copious backup in case of emergencies, has sharply truncated our vision of the tolls that reproduction can take under conditions that we no longer experience or fully understand.

Future Possibilities for Anthropology and Demography in Studies of Reproduction

As these examples suggest, anthropology's style of engagement with demography and the insights to gain from such mergers will continue to change as the fields themselves change. In the evolving relationship between anthropology and demography, moreover, the spectrum of topics continues to broaden. New common interests continually arise in the politics of fertility control, new fertility technologies, gender and sexuality studies, the negotiation of fertility decisions among spouses, and the continual return to scrutinize the taken-for-granted bases of fertility (Petit, 2013; Verheijen, 2013).

Finally, while anthropology may have made itself as a discipline on the study of exotic people living in distant places, increasing numbers of practitioners are coming 'back home,' studying social and cultural phenomena in the urban United States and Europe. They are also experimenting with new forms and blends of data. These trends are facilitating bold new demographic and cultural projects, involving studies of almost unprecedented scope and creativity (Kertzer and Arel, 2002; Watkins and Swidler, 2009). Some of the most productive areas inspired by the determination on the parts of many to break down the quantitative/qualitative barriers lie in studies of migration. New anthropological interests in globalization are coming at the time when demographers are taking up the challenges of studying international migration policies that may separate parents and children, or create entirely new ones (Coe, 2013; Leinnaweaver, 2013). Indeed, the opening of the European Union and, in the United States, the removal of national-origin quotas for immigrants that have spawned migration across national borders on such an unprecedented scale are creating virtually uncharted forms of reproduction and reproductive careers. The continual emergence of patterns like these underscores the utility of past insights into population and culture; at the same time, it decisively challenges any confidence we might have entertained that these notions will suffice to understand population and its reproductive novelties across time and space that unfold almost daily in the national and international press.

See also: Abortion, Demography of: United States; East Asian Studies: Culture; Family Theory: Economics of Childbearing; Fertility Transition: China; Fertility Transition: Cultural Explanations; Heterosexuality in China; Migration and Fertility; Sex Selective Abortion; Sexual Politics in the Arab World.

Bibliography

Bledsoe, C., 1994. Children are like young bamboo trees: potentiality and reproduction in sub-Saharan Africa. In: Kiessling, K.L., Landbcrg, H. (Eds.), Population, Economic Development, and the Environment. Oxford University Press, New York, pp. 105–138.

Bledsoe, C.H., 2002. Contingent Lives: Fertility, Time, and Aging in West Africa. University of Chicago Press, Chicago, IL.

Caldwell, J.C., 1980. Mass education as a determinant of the timing of fertility decline. Population and Development Review 6 (2), 225–255.

Charbit, Y., 2010. The Classical Foundations of Population Thought: From Plato to Quesnay. Springer, Dordrecht.

Coale, A., 1973. The demographic transition reconsidered. In: International Union for the Scientific Study of Population. International Population Conference, vol. 1. Liege IUSSP, Liege, pp. 53–72.

Coale, A.J., Watkins, S.C. (Eds.), 1986. The Decline of Fertility in Europe. Princeton University Press, Princeton, NJ.

Coe, C., 2013. The Scattered Family: Parenting, African Migrants, and Global Inequality. University of Chicago Press, Chicago.

Geertz, C., 1973. The Interpretation of Cultures: Selected Essays by Clifford Geertz. Basic Books, New York.

Greenhalgh, S., Li, J.L., 1995. Engendering reproductive policy and practice in peasant China: for a feminist demography of reproduction. Signs: Journal of Women in Culture and Society 20 (3), 601–641.

Henry, L., 1961. Some data on natural fertility. Eugenics Quarterly 8, 81–91.

Inhorn, M.C., 2012. The New Arab Man: Emergent Masculinities, Technologies, and Islam in the Middle East. Princeton University Press, Princeton, NJ.

Isiugo-Abanihe, U.C., 1985. Child fosterage in west africa. Population and Development Review 11 (1), 53–73.

Johnson-Hanks, J., 2005. When the future decides – uncertainty and intentional action in contemporary Cameroon. Current Anthropology 46 (3), 363–385.

Johnson-Hanks, J., Bachrach, C., Morgan, S.P., Kohler, H.P., 2001. Understanding Family Change and Variation: Toward a Theory of Conjunctural Action. Springer Verlag.

Johnson-Hanks, J., Miguel, E., 2011. The stability and predictive ability of fertility intentions in western Kenya (with Edward Miguel). Paper presented at the Population Association of American Annual Meeting, Washington DC. March 31-April 2, 2011.

Kertzer, D.I., Arel, D. (Eds.), 2002. Census and Identity: The Politics of Race, Ethnicity, and Language in National Censuses. Cambridge University Press, Cambridge.

Leinnaweaver, J., 2013. Adoptive Migration: Raising Latinos in Spain. Duke University Press, Durham, NC.

Lesthaeghe, R.J., 1977. The Decline of Belgian Fertility, 1800–1970. Princeton University Press, Princeton, NJ.

Lorimer, F., 1954. Culture and Human Fertility. UNESCO, Paris.

Malthus, T.R., 1798. An essay on the principle of population. J. Johnson, London.

Murdock, G.P., 1975. Outline of World Cultures, fifth ed. Human Relations Area Files, New Haven, CT.

Musallam, B.F., 1983. Sex and Society in Islam: Birth Control before the Nineteenth Century. Cambridge University Press, Cambridge, UK.

Peters, P.E., 2010. Contestations over "tradition" and "culture" in a time of AIDS. Medical Anthropology 29 (3), 278–302.

Petit, V., 2013. Counting Populations, Understanding Societies: Towards a Interpretative Demography. Dordrecht, New York.

Szreter, S., Fischer, K., 2010. Sex before the sexual revolution. Intimate Life in England 1918–1963.

Townsend, N.W., 2000. Male fertility as a life time of relationships: contextualizing men's biological reproduction in Botswana. In: Bledsoe, C., Lerner, S., Guyer, J.I. (Eds.), Fertility and the Male Life Cycle in the Era of Fertility Decline. Oxford University Press, Oxford, UK.

Verheijen, J., 2013. Balancing Men, Morals and Money: Women's Agency between HIV and Security in a Malawi Village. African Studies Centre, Leiden.

Watkins, S.C., 2001. Fertility transition: cultural explanations. In: Smelser, Neil J., Baltes, Paul B. (Eds.), International Encyclopedia of the Social Sciences, second ed., vol. 8. Elsevier, Amsterdam, pp. 5572–5575.

Watkins, S.C., Swidler, A., 2009. Hearsay ethnography: conversational journals as a method for studying culture in action. Poetics 37 (2), 162–184.

Whiting, J.W.M., Kluckhohn, R., Anthony, A.S., 1958. The function of male initiation ceremonies at puberty. In: Maccoby, E.E., Newcomb, T., Hartley, E. (Eds.), Readings in Social Psychology. Holt, New York, pp. 359–370.

Fertility and Social Networks

Hans-Peter Kohler, Population Studies Center, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H.P. Kohler, C. Buhler, volume 21, pp. 14380–14384, © 2001, Elsevier Ltd.

Abstract

Social networks receive an increasing emphasis in theories of fertility change. In contexts where modern family planning constitutes an innovation, as well as in contexts where fertility has declined to very low levels, interaction in social networks can affect the spread of new ideas and methods of fertility limitation. Social interactions can also shape norms and schema related to childbearing. This article first explores the theoretical arguments that explain why social networks are relevant for understanding fertility behavior. It then presents evidence that social networks exhibit important influences on contraceptive behavior and fertility, and it discusses the implications of social networks for understanding fertility dynamics.

Explanations offered for fertility decline in both developed and developing countries underwent fundamental changes in recent years. In addition to more refined economic reasoning and the introduction of cultural aspects, theories of the fertility transition now routinely reserve a place for diffusion and/or social interactions effects. Such effects arise because individuals are themselves members of larger groups. The information that is known to group members, the choices they make, the schemas they hold, and the norms and values they adhere to, can all exert a powerful influence on individual incentives to innovate and change behaviors. In settings in which fertility has been high, for example, innovation might take the form of adopting modern contraception and fertility limitation. In low-fertility settings, social networks can also shape norms and schema related to childbearing, or affect individuals' perceptions about the trade-offs entailed with having and raising children. Social interaction with group members represents one important avenue along which such innovative demographic behavior can diffuse: peers, community members or relatives may provide relevant information to potential adopters of family planning, they can affect the schema that guide individuals' behaviors with respect to life-course planning and parenthood, and they also represent the institutional and social structures that may facilitate or constrain individual fertility or family planning decisions.

Theories of diffusion and social interactions usually encompass two central elements. First, they emphasize a particular set of social interactions that are associated with the adoption of innovations, such as social networks, opinion leaders, public media, etc. Second, diffusion models have an explicit dynamic dimension in which the adoption of an innovation is often associated with 'contagion' or 'tipping point' phenomena, which can accelerate social change; there may also be 'conjunctures' that represent specific short-term social contexts within which decision-making often occurs.

Because fertility transitions are often rapid, irreversible, and only weakly associated with the specific socioeconomic conditions in a population, such diffusionist explanations have appealed to demographers ever since the European Fertility Project was carried out (Coale and Watkins, 1986). At the same time, diffusionist approaches have been criticized for several key reasons, such as the unclear definition of what diffuses the unclear connection with economic models of behavior or other models of individual decision-making, the mostly indirect empirical evidence, and the insufficient knowledge about the preconditions for the onset of the diffusion process. Several of these potential problems can be overcome by investigating the diffusion process with the tools of social network analysis.

Why Social Networks Matter for Understanding Fertility

The theoretical perspective of social networks rests on one central insight: human action is not carried out by individuals who make their decisions in isolation and independence, but by socialized actors who are deeply rooted in their social environments. These environments consist of other individuals with whom a focal actor stays in various relationships such as kinship, love, friendship, competition, or business transactions. Within these relationships, norms and schema about fertility and parenthood are negotiated and enforced, and information and goods are exchanged (Mitchell, 1973; Johnson-Hanks et al., 2011). On the one hand, this implies that a network member has access to resources owned by other network partners. In addition, a network enables its members to use the resources of community members to whom they are tied only indirectly. Social networks, therefore, constitute a pool of resources, which form opportunity structures for action (Lin, 1999). The kind and quality of these resources are decisive for the goals that are attainable for the network members. On the other hand, social relationships create, maintain, and modify attitudes (Erickson, 1988). People feel uncomfortable, if the attitudes of their network partners are too different from their own. Social networks, therefore, bring the attitudes of their members in accordance with each other. Alternatively, if the network fails to bring attitudes into conformity, the relationships become less intensive or are abandoned.

The resources that are available in a network, and the tendencies toward similar attitudes, depend on the characteristics of individuals' direct relationships with one another as well as on the structure of the entire network. One basic characteristic of network relations is their strength (Granovetter, 1973). Strong relationships, i.e., relationships that are characterized by a high frequency of contact or by a high level of emotional intensity, can be influential because they tend to transfer more

valuable goods, more binding norms and more trustworthy – but also more redundant – information. On the other hand, a fundamental dimension of the network structure is its density. A dense network, i.e., a network in which a high proportion of the individuals are in direct contact with one another, can lead to a high degree of normative pressure on its members and it can provide intensive support for members in situations when help is needed. Less-dense networks are characterized by a greater degree of heterogeneity. Such networks provide more opportunities for individual development, and they are the source of more diverse and heterogeneous information.

The mechanisms by which social networks affect the diffusion process can be summarized under the headings 'social learning,' 'joint evaluation,' and 'social influence' (Bongaarts and Watkins, 1996; Montgomery and Casterline, 1996). At early stages, potential adopters need to be informed about the existence of new methods of fertility control, their availability, costs and benefits, and possible side effects. 'Social learning' is defined as the process by which individuals (1) learn about the existence and technical details of new contraceptive techniques, and (2) reduce the uncertainties associated with the adoption of these methods by drawing on the experience of network partners.

The spread of information about modern contraceptives is always accompanied by the spread of new ideas about the social role of women, the status of the family, and the desired number of children. 'Joint evaluation' refers to the fact that social networks (1) facilitate the transformation of these new ideas into terms that are meaningful in the local context, and (2) assist in the evaluation of the respective social changes. A woman's decision to adopt a modern form of contraception, however, rests not only on information, conviction, and discursive evaluation, but also on the normative pressures of dominant groups or actors. 'Social influence' describes the process by which an individual's preferences, attitudes, and schema related to parenthood are influenced by the prevailing norms and values of the social environment. The direction of this social influence can differ in different stages of fertility decline. In the early stages, social influence may constrain the adoption of contraception, which demographers may interpret as being in a woman's best interest, because deliberate control of fertility is still considered deviant. Once a noticeable fertility decline has set in, however, the direction of social influence can shift. As low fertility becomes commonplace, the interaction among women or couples is likely to encourage a decline in the demand for children and behaviors that lead to a limiting of fertility, while at the same time, social interactions start to have important reinforcing influences on the postponement of fertility to older ages (Kohler et al., 2002).

Empirical Evidence

The earliest systematic evidence for the influence of diffusion processes and social networks on fertility decisions is found in two seminal analyses of demographic change. First, the European Fertility Project (Coale and Watkins, 1986) showed that areas which are culturally similar, in language or ethnicity, for example, often exhibit similar fertility patterns despite substantially different socioeconomic conditions and development levels. Thus, new ideas about fertility as well as new contraceptive methods seem to have diffused within cultural and linguistic boundaries. More sophisticated approaches employing similar methodological ideas have used the correlation pattern in pooled time series to show support for the role of social interactions in the adoption of family planning over time (Montgomery and Casterline, 1993). Second, the Taichung family planning experiment in Taiwan (Freedman and Takeshita, 1969) used a randomized experiment and longitudinal follow-up to show how direct interventions of a family planning program targeted to some individuals spread to others who were not directly contacted by the program. The study thus suggests that family planning interventions extend beyond the direct effect on the contacted individuals and that interaction in social networks constitutes an important aspect in the diffusion of new contraceptive methods and ideas.

The above evidence, although intriguing, is far from conclusive. In all of these data, the evidence is indirect and based on correlated fertility developments in geographically and/or culturally close populations. Related early evidence on microdata has often not been able to conclusively document influences of social networks on fertility decisions because the analyses identified associations between social networks and fertility outcomes, which may not represent causal effects of networks on fertility outcomes when networks are formed endogenously and are characterized by homophily and/or network members are subject to shared unobserved environmental influences (Manski, 1993). To overcome these limitations, the recent literature aimed at identifying social interaction effects on fertility and fertility-related behaviors has therefore used longitudinal data on social networks and family planning use (Behrman et al., 2002), focused on interactions among siblings (Lyngstad and Prskawetz, 2010), used instrumental variable approaches with the identification assumptions that some aspects – like physical attractiveness – directly affect friends' fertility behaviors but only indirectly affects respondents' fertility behaviors through social interactions (Card and Giuliano, 2013), exploited nonlinear variation in child allowances that provide differential incentives for having more additional children (Manski and Mayshar, 2003) or used 'natural experiments' in which structural reforms (like education reforms) differentially affected the childbearing decisions of siblings (Monstad et al., 2011). There has also been renewed attention from qualitative researchers in trying to unpack the relationship between interactions in social networks and fertility behaviors. In the context of relatively high fertility, the work of Watkins and collaborators in South Nyanza, Kenya, for example, demonstrated that casual conversations about family planning occur frequently as women go about their daily activities (Kohler et al., forthcoming). The network partners in these conversations are usually geographically and socially close to one another and are of a similar age or education. Frequently, they also know each other and share multiple relations, e.g., stemming from money lending, help in the household, or family ties. Related recent studies have highlighted that similar influences of social networks affect fertility behavior also in low-fertility contexts (Bernardi, 2003), including particularly the intentions to have children (Balbo and Mills, 2011) and the differential effect of strong and weak ties on fertility decision-making (Keim et al., 2013).

Besides having been challenged in terms of providing an adequate identification of causal effects of networks on fertility

outcomes, a further criticism of existing empirical work has been that the link between the theoretical arguments of why social interaction matters and the respective empirical implementation is weak in many analyses. One possible way to overcome this limitation is to use additional information about the structure of the social network, e.g., the density. Dense networks impose more behavioral constraints on the adoption or nonadoption of family planning than do sparse networks, which are more open to innovative information and behavior. Thus, density can be used as an indicator for the mechanisms of 'social influence' and 'social learning.' If the prevalence of contraceptive use in dense networks is more relevant for the decision to adopt contraception than it is in sparse networks, then social influence is the most relevant aspect of social networks. If it is the other way around, however, then it is social learning that is the dominant component. In the South Nyanza district of Kenya, for instance, evidence for both pathways of influence can be found. Moreover, the relative importance of these two mechanisms depends on the socioeconomic context of the regions. Social learning seems to be relatively more important in areas with higher levels of market activity (Kohler et al., 2001).

Although the empirical evidence supporting the argument that social networks play an important role in fertility decisions is accumulating from various sources, it remains inconclusive in important aspects, e.g., as regards the empirical identification of causal network effects or the mechanisms through which social networks affect fertility and related behaviors. There is also a need for new theories that conceptualize fertility behavior as an unfolding process over the life course that is influenced by schemas, material conditions, and social interactions (Johnson-Hanks et al., 2011). Nevertheless, the extensive efforts currently in progress to collect data, combined with both an improved theoretical understanding of social interaction and a more sophisticated empirical methodology, are likely to reveal further insights.

Implications of Social Networks for Fertility Change

Current research on social networks and fertility is in part stimulated by the fact that the dynamic implications of fertility models that include social interactions seem to be more consistent with the empirical evidence on fertility transitions than individualistic models. However, specific theoretical analyses of social interactions and fertility decisions are required as further support for this conclusion. Early formal models treat the diffusion of low fertility in a similar manner to the way the spread of contagious diseases is treated in epidemiological models (Rosero-Bixby and Casterline, 1993). More recent developments have used stochastic processes or agent-based simulation (Kohler, 2001; Diaz et al., 2011), with models explicitly considering the interplay between social networks, fertility decisions, schemas, and economic incentives within an evolving social context (Johnson-Hanks et al., 2011).

The main consequence of interactions in social networks is that fertility decisions of women or couples in a population become interdependent and exert mutual influences on each other. The implications of this interdependence can be illustrated in a simple nonlinear model of social interaction (**Figure 1**). The discussion focuses on social network influences on contraceptive adoption, and the framework has been analogously applied to explain the emergence of lowest-low fertility (Kohler et al., 2002; Lois and Arránz Becker, 2014) and teenage childbearing (Card and Giuliano, 2013; Monstad et al., 2011). For illustration, we focus on contraceptive choice in high-fertility context and assume that the probability that a woman ever adopts modern contraception is given by,

$$\begin{aligned} &\mathrm{P}(\text{woman adopts modern contraception}|x,\bar{y}) \\ &\quad = F(\alpha(\bar{y}-0.5)+\beta x+\gamma) \end{aligned} \qquad [1]$$

where α and β are positive parameters, γ represents individual characteristics, x is the level of family planning effort, F is the cumulative logistic function, and $\bar{y}$ is the proportion of women in the population who use modern contraception in the reference period. Hence, the social utility term $\alpha\ (\bar{y}-0.5)$ captures the dependence of a woman's contraceptive choice on the fertility behavior in her network, where we assume for the sake of simplicity that this network includes all members of the population. The model implies that social effects are absent in a population in which half of the members use modern contraception and the other half do not. If the proportion of contraceptive users exceeds (is below) 0.5 then the social network increases (decreases) the probability of using contraception. This simple model already reveals that the aggregate properties of fertility dynamics in the presence of

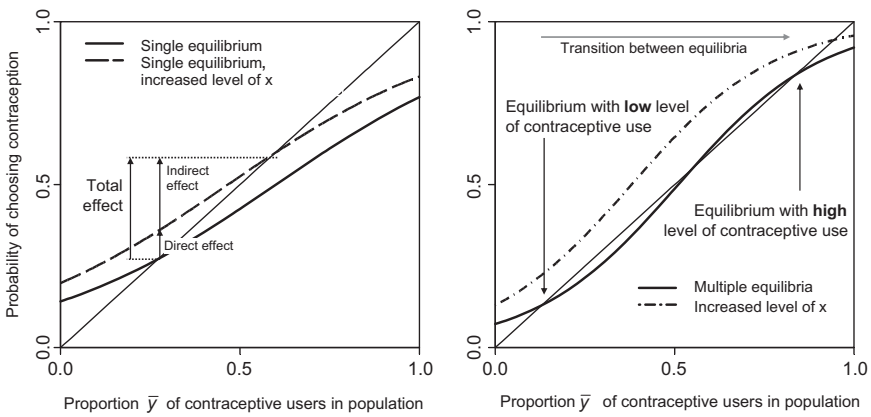

Figure 1 Implications of social networks for population dynamics. In the left-hand panel, social networks exert a relatively weak influence on individual contraceptive decisions ($\alpha = 3$), while in the right-hand panel the social network effect is relatively strong ($\alpha = 6$).

social interaction are distinct from the fertility dynamics in a purely individualistic model. In particular, the relevance of social interaction arises from three factors: social multiplier effects, multiple equilibria, and dense versus sparse networks.

Social Multiplier Effects

Social interaction implies the presence of multiplier effects, which change how the fertility level adjusts to changes in family planning programs. An increase in family planning effort x, for instance, increases the probability of choosing modern contraception in two distinct ways (Figure 1). The direct effect reflects the increase in the propensity to choose modern contraception while holding the population prevalence $\bar{y}$ constant. However, because the increased level of x affects all community members, the prevalence of modern contraception in the population tends to increase as well. This change in the community has a feedback or social multiplier effect on individual behavior in future periods via the social utility term in eqn [1]. This means that the total change in the prevalence of family planning exceeds the direct effect. The left-hand panel of Figure 1 shows how this total effect is composed of a direct effect and an indirect effect through social interaction. The interaction in social networks thus enhances the changes in contraceptive behavior that follow changes in family planning efforts or other socioeconomic changes.

Multiple Equilibria

Many formal fertility models assume that there is only one aggregate fertility level, denoted as 'equilibrium,' associated with any given level of socioeconomic conditions. When the interdependence of fertility decisions in social networks is sufficiently strong, however, multiple equilibria can emerge (right-hand panel of Figure 1). This possibility of there being multiple equilibria is relevant because in this case a population can be 'stuck' at a low level of contraceptive use, even though a second equilibrium exists with a higher level of contraceptive use and higher individual utility levels. In this case, increases in the policy level x (or other socioeconomic changes) can induce a transition from the low-contraceptive use to the high-contraceptive use equilibrium. These transitions between equilibria are often thought to occur at a rapid pace, resulting in significant changes in fertility behavior within a relatively short space of time. Moreover, both the high- and low-contraceptive use equilibria are stable, so that a transitory increase in the family planning effort can yield sustainable long-term changes in fertility levels.

Populations With Dense versus Sparse Networks

An important question in diffusion models is whether 'more' interaction in networks always leads to larger multiplier effects and more rapid diffusion. It can be shown that already in simple models such as eqn [1], social networks can be status-quo enforcing: the long-term effect associated with changes in the program effort can decrease with an increase in parameter α, which reflects the importance of the network influences in individual fertility decisions. The changes in contraceptive use resulting from changes in the program effort can therefore be smaller for very densely knit societies than for more sparse societies.

See also: Family Planning: Program Effects; Fertility Transition: Cultural Explanations; Fertility Transition: Sub-Saharan Africa; Models of Influence in Networks; Network Analysis, History of; Network Analysis; Second Demographic Transition; Sexual Behavior and Social Networks; Social Networks.

Bibliography

Balbo, N., Mills, M., 2011. The influence of the family network on the realisation of fertility intentions. Vienna Yearbook of Population Research 9, 179–205.

Behrman, J.R., Kohler, H.-P., Watkins, S.C., 2002. Social networks and changes in contraceptive use over time: evidence from a longitudinal study in rural Kenya. Demography 39 (4), 713–736.

Bernardi, L., 2003. Channels of social influence on reproduction. Population Research and Policy Review 22 (5–6), 527–555.

Bongaarts, J., Watkins, S.C., 1996. Social interactions and contemporary fertility transitions. Population and Development Review 22, 639–682.

Card, D., Giuliano, L., 2013. Peer effects and multiple equilibria in the risky behavior of friends. Review of Economics and Statistics 95 (4), 1130–1149.

Coale, A.J., Watkins, S.C., 1986. The Decline of Fertility in Europe. Princeton University Press, Princeton, NJ.

Diaz, B.A., Fent, T., Prskawetz, A., Bernardi, L., 2011. Transition to parenthood: the role of social interaction and endogenous networks. Demography 48 (2), 559–579.

Erickson, B.H., 1988. The relational basis of attitudes. In: Wellman, B., Berkowitz, S.D. (Eds.), Social Structures: A Network Approach. Cambridge University Press, Cambridge, pp. 99–121.

Freedman, R., Takeshita, J.Y., 1969. Family Planning in Taiwan: An Experiment in Social Change. Princeton University Press, Princeton, NJ.

Granovetter, M.S., 1973. The strength of weak ties. American Journal of Sociology 73, 1361–1380.

Johnson-Hanks, J.A., Bachrach, C.A., Morgan, S.P., Kohler, H.-P., 2011. Understanding Family Change and Variation: Toward a Theory of Conjunctural Action. Springer Verlag, New York/Berlin.

Keim, S., Klärner, A., Bernardi, L., 2013. Tie strength and family formation: which personal relationships are influential? Personal Relationships 20 (3), 462–478.

Kohler, H.-P., 2001. Fertility and Social Interactions: An Economic Perspective. Oxford University Press, Oxford.

Kohler, H.-P., Behrman, J.R., Watkins, S.C., 2001. The density of social networks and fertility decisions: evidence from South Nyanza District, Kenya. Demography 38, 43–58.

Kohler, H.-P., Billari, F.C., Ortega, J.A., 2002. The emergence of lowest-low fertility in Europe during the 1990s. Population and Development Review 28 (4), 641–681.

Kohler, H.-P., Helleringer, S., Behrman, J.R., Watkins, S.C., forthcoming. The social and the sexual: networks in contemporary demographic research. In: Kreager, P., Capelli, C., Ulijaszek, S., Winney, B. (Eds.), Population in the Human Sciences: Concepts, Models, Evidence. Oxford University Press, Oxford. Also available as PSC Working Paper 13-01 at http://repository.upenn.edu/psc_working_papers/41.

Lin, N., 1999. Building a network theory of social capital. Connections 22, 28–51.

Lois, D., Arránz Becker, O., 2014. Is fertility contagious? Using panel data to disentangle mechanisms of social network influences on fertility decisions. Advances in Life Course Research 21, 123–134.

Lyngstad, T.H., Prskawetz, A., 2010. Do siblings' fertility decisions influence each other. Demography 47 (4), 923–934.

Manski, C.F., 1993. Identification of endogenous effects: the reflection problem. Review of Economic Studies 60 (3), 531–542.

Manski, C.F., Mayshar, J., 2003. Private incentives and social interactions: fertility puzzles in Israel. Journal of the European Economic Association 1 (1), 181–211.

Mitchell, J.C., 1973. Networks, norms, and institutions. In: Boissevain, J., Mitchell, J.C. (Eds.), Network Analysis: Studies in Human Interaction. The Hague, Holland: Mouton, pp. 2–35.

Monstad, K., Propper, C., Salvanes, K.G., 2011. Is Teenage Motherhood Contagious? Evidence from a Natural Experiment. CEPR Discussion Paper No. DP8505. URL: http://ssrn.com/abstract=1908553.

Montgomery, M.R., Casterline, J.B., 1993. The diffusion of fertility control in Taiwan: evidence from pooled cross-section time-series models. Population Studies 47, 457–479.

Montgomery, M.R., Casterline, J.B., 1996. Social learning, social influence learning and new models of fertility. Population and Development Review 22, S151–S175.

Rosero-Bixby, L., Casterline, J.B., 1993. Modelling diffusion effects in fertility transition. Population Studies 47, 147–167.

Fertility Change: Quantum and Tempo

John Bongaarts, The Population Council, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The 'quantum' and 'tempo' dimensions of fertility usually refer, respectively, to the average number of births per woman during the reproductive years and the average age at childbearing. These measures can be estimated for cohorts of women who have completed their childbearing years, as well as for specific time periods. Unfortunately, quantum measures of period fertility (e.g., the total fertility rate) can be distorted, because its quantum and tempo dimensions interact. Assessment of period fertility levels and trends therefore should include adjustments to remove tempo distortions caused by changes in the timing of childbearing.

The 'quantum' and 'tempo' dimensions of fertility usually refer, respectively, to the average number of births per woman during the reproductive years and the average age at childbearing (Ryder, 1980; Pressat, 1985). Using standard demographic methods these measures can be calculated for birth cohorts of women who have completed their childbearing years, as well as for specific time periods. More generally, the term 'quantum' can refer to the average numbers of births that have occurred in a group of women who have experienced a common event, such as marriage, or its measurement can be limited to births of a certain order, such as first births. Similarly, the term 'tempo' can refer to the average time elapsed between a birth and a specific preceding event (e.g., marriage) or its measurement can be limited to births of a certain order.

Estimation of the quantum and tempo of childbearing for a cohort of women is straightforward provided the required data are available and accurate. These variables quantify key dimensions of a cohort's reproductive history but they have the considerable disadvantage of representing past experience: women currently aged 50 did most of their childbearing two to three decades ago when they were in their 20s and early 30s. Trends in cohort measures therefore do not reflect current conditions, and they are not often used to monitor recent developments because fertility can change substantially in a short period of time.

An alternative quantum measure, the period total fertility rate (PTFR), is used far more widely as an indicator of current fertility. It is defined as the average number of births a woman would have if she were to live through her reproductive years (ages 15–49) and bear children at each age at the rates observed in a particular year or period. It is a hypothetical measure because no real group of women has experienced or will necessarily experience these particular rates. The advantage of the PTFR is that it gives up-to-date information on levels and trends in fertility. The PTFR is a quantum indicator of period fertility, and the mean of the age-specific fertility schedule is a tempo measure.

The simplicity and wide availability of the PTFR have contributed to a neglect of some deficiencies in this fertility indicator. The demographic literature on the measurement of fertility includes a number of criticisms of the PTFR (Ní Bhrolcháin, 1992). The most important of these is the distortion of the PTFR that occurs when the timing of childbearing changes. During years in which women delay childbearing, period fertility rates are depressed relative to cohort fertility; in years when childbearing is accelerated, period fertility is raised (Hajnal, 1947; Ryder, 1956, 1964, 1980, 1983). The quantum and tempo dimensions of period fertility interact.

To demonstrate how these tempo distortions operate, it is helpful to examine a set of simple illustrations as presented in Figure 1. In these hypothetical illustrations, every woman is assumed to have exactly one birth so that the cohort total fertility rate equals one for all cohorts, and these births are assumed to take place at exactly the same age within each cohort. Figure 1(a) depicts the reference situation in which this age is set at exactly 20 years. Since there are no tempo effects, period fertility as measured by the PTFR is also constant and equal to one birth per woman. Assume next that the age at childbirth is raised by 1 year starting at time T, as shown in Figure 1(b). In this case, women who would have given birth at age 20 immediately following T have to delay their childbearing for a year until they reach their 21st birthday. As a consequence, no births occur during the year between T and $T+1$ and the PTFR then equals 0. In other words, 1 year's worth of births is lost because the mean age at childbearing increased by 1 year. In reality, the mean age at childbearing rarely changes suddenly and rapidly. A more realistic situation is summarized in Figure 1(c), which illustrates the effect of a slow rise in the age at childbearing. Starting at time T, successive cohorts delay childbearing so that the mean age at first birth rises linearly from 20 to 21 years over a 10-year period from T to $T+10$. In this case, the PTFR drops from 1 to 0.9 at time T, remains at 0.9 from T to $T+10$, and finally returns to 1 after time T. The PTFR is distorted by 10% or 0.1 birth per woman during the interval T to $T+10$. Over the 10-year period, 1 year's worth of fertility is lost from the PTFR as a result of a 1-year increase in the mean age at childbearing. Conversely, as shown in Figure 1(d), if the mean age at first birth declines linearly from 20 to 19 years over the same period of 10 years, the PTFR rises from 1 to 1.1 at time T, remains at 1.1 from T to $T+10$, and returns to 1 birth per woman after T. Bongaarts and Feeney (1998, 2006) present a fuller discussion of these effects.

The existence of timing distortions is readily documented when the age at childbearing is declining significantly. In that case, implausible results are often obtained for the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31072-8

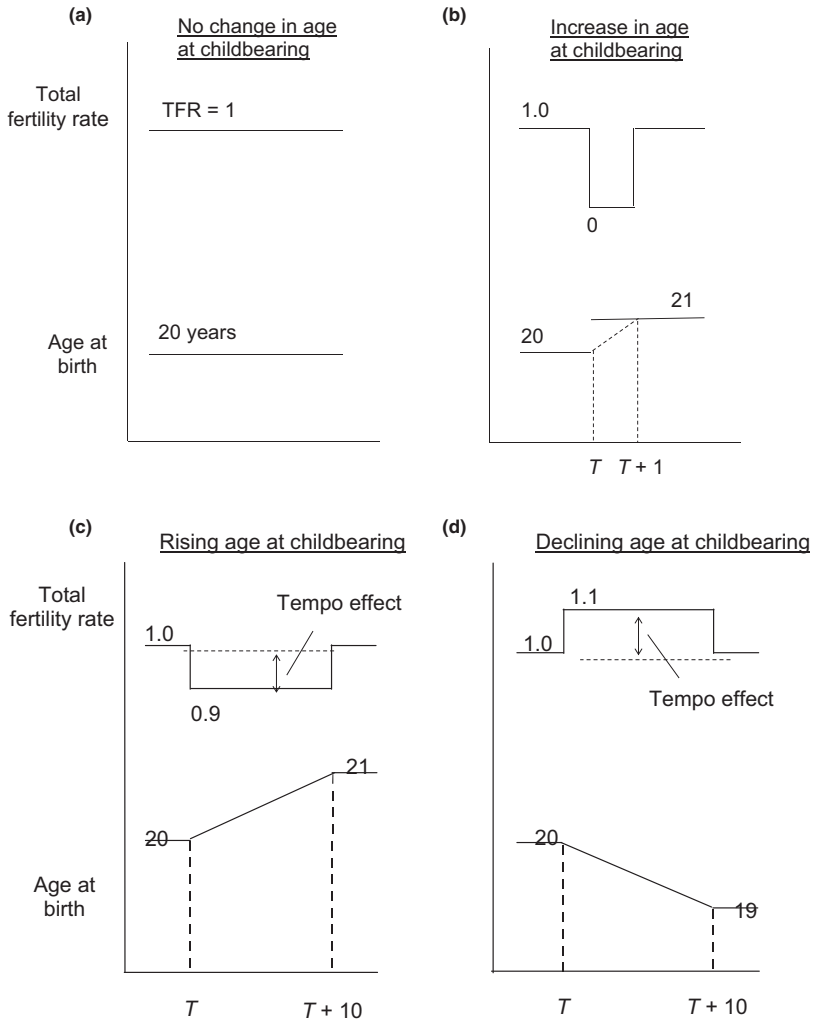

Figure 1 (a), (b) Hypothetical patterns of childbearing, with and without tempo effects. (c), (d) Fertility impact of changes in age at childbearing.

birth-order components of the PTFR, computed as the sum over age of the period-specific fertility rates of the second kind by age and birth order. For example, the component for births of order one ($PTFR_1$) equals the average number of first births women would have by age 50 if they were to bear first births at the age-specific rates of the second kind observed in a given year or period. In most years during the late 1940s and early 1950s, the $PTFR_1$ in the US exceeded 1 (Whelpton, 1954), a fact which seems to imply that women had more than one first birth on average. This is impossible and these $PTFR_1$ estimates must therefore be reinterpreted. The main explanation of why $PTFR_1$ exceeded 1 during the baby-boom years is that the age at childbearing declined, resulting in a temporary inflation of the PTFR and its additive birth-order components:

$$\mathrm{PTFR} = \sum_n \mathrm{PTFR}_n \qquad [1]$$

A detailed analysis of the effects of variations in the childbearing tempo on period fertility indicators was undertaken by Ryder in a series of influential papers (Ryder, 1956, 1964, 1980, 1983). Ryder proposed 'translation' equations to calculate the period fertility measures from corresponding cohort measures when the timing of cohort fertility is changing. These procedures have not found wide acceptance for two main reasons. First, in his work on the translation issue, Ryder assumes that the tempo and quantum of cohort fertility are the determinants of the PTFR and other period fertility measures. However, extensive empirical analysis of this issue has demonstrated that this is not the case (Ní Bhrolcháin, 1992). Period effects are clearly the prime source of variation in fertility rates, and the completed fertility of cohorts is largely determined by the period fertility rates during the cohort's reproductive years, rather than the other way around. Second, the two dimensions of aggregate cohort fertility – quantum and tempo – are not independent in practice. When cohorts reduce their fertility, they do this primarily by reducing childbearing at higher birth orders. As a result, the mean age at childbearing for all births to the cohort declines even if the timing of individual births does not change. In other words, a decline in the cohort quantum can lead to changes in the cohort's mean age at childbearing

that do not cause tempo distortions. Such distortions only occur if the mean ages at births of specific orders rise or fall. Ignoring this effect – as Ryder does in much of his writing – therefore gives biased results except when cohort fertility is constant. Fortunately, this second problem can be solved by analyzing fertility trends separately for each birth order rather than for overall fertility.

A procedure for removing tempo effects from the PTFR was proposed by Bongaarts and Feeney (1998). The approach is an outgrowth of Ryder's original translation equation. The above objections to Ryder's approach are addressed by assuming that fertility is only affected by period effects, and by applying adjustments separately to the birth-order components (PTFR_n) of the total fertility rate rather than to the total fertility rate itself. Under these conditions, it is possible to estimate a tempo-free total fertility rate in any given year (or period) from the conventional PTFR_n at each birth order n with the formula

$$\mathrm{PTFR}'_n = \mathrm{PTFR}_n/(1 - m_n) \qquad [2]$$

where PTFR_n is the observed total fertility rate component for birth order n, PTFR'_n is the tempo-free total fertility rate component for birth order n, and m_n is the annual *change* in the mean age of the age-specific fertility schedule for birth order n. Dividing the observed total fertility rate by $(1–m_n)$ at any given birth order n, produces an estimate of the total fertility rate that would have been observed if there had been no change in the timing of childbearing. Adding up results for the various birth orders gives the overall tempo-free total fertility rate.

$$\mathrm{PTFR}' = \sum_n \mathrm{PTFR}'_n \qquad [3]$$

The PTFR′ equals the distortion-free quantum component of the PTFR, and the difference PTFR′–PTFR equals the tempo distortion.

According to this argument, an annual increase of one-tenth of a year in the mean age at childbearing ($m_n = 0.1$) reduces the PTFR_n by 10% below its tempo-free level, because $\mathrm{PTFR}_n = 0.9\ \mathrm{PTFR}'_n$. Similarly, an annual decline in the mean age at a rate of just 0.1 years per year ($m_n = -0.1$) inflates the PTFR_n by 10%. Clearly, rather modest changes in the timing of childbearing at any birth order can produce substantial changes in observed fertility.

The above simple formula for removing tempo distortions assumes that the shape of the age-specific period fertility rates at each birth order does not change over time. Kohler and Philipov (2001) have suggested a more general equation that incorporates variance effects. Their formula reduces to the simpler version presented here when the shape of the fertility schedule is invariant. Further discussion of different tempo adjustment procedures can be found in Bongaarts and Sobotka (2011).

The existence of tempo distortions of period fertility measures has recently become an important issue in contemporary developed countries. In most of these societies, fertility has dropped well below the replacement level. If future fertility remains at these low levels, populations will decline in size and age rapidly. The fact that ages at childbearing have been rising quite rapidly since the 1970s implies that the PTFRs of these countries contain a tempo distortion. Sobotka (2004) estimates the tempo effects for many posttransitional societies in the late 1990s and finds downward distortions ranging up to 0.55 births per woman in the Czech Republic. In several countries the tempo effect is the primary reason why the PTFR is below the replacement level.

Assessment of fertility levels and trends should include an analysis of the tempo effects because the conventional PTFR gives a misleading estimate of the actual rate of childbearing when the age at childbearing is changing. This is the case in many posttransitional populations. Childbearing is being delayed and women are actually bearing more children than is indicated by the observed PTFRs. This distortion in the PTFR will continue for as long as the mean age at childbearing rises, but once the deferment ends the distortion is removed, and the PTFR will then rise unless there are offsetting declines in the quantum component of period fertility. This is in fact the main reason why the PTFR has risen in the decade up to 2008 in many European countries (Bongaarts and Sobotka, 2011).

See also: Family Size Preferences; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective.

Bibliography

Bongaarts, J., Feeney, G., 1998. On the tempo and quantum of fertility. Population and Development Review 24, 271–291.

Bongaarts, John, Feeney, Griffith, 2006. The quantum and tempo of life cycle events. ViennaYearbook of Population Research 115–151.

Bongaarts, John, Sobotka, Tomáš, 2011. Demographic explanations for the recent rise in European fertility: analysis based on the tempo- and parity-adjusted total fertility. Population and Development Review 38, 83–120.

Hajnal, J., 1947. The analysis of birth statistics in the light of the recent international recovery of the birth-rate. Population Studies 1, 137–164.

Kohler, H.-P., Philipov, D., 2001. Variance effects in the Bongaarts-Feeney formula. Demography 38, 1–16.

Ní Bhrolcháin, M., 1992. Period paramount? A critique of the cohort approach to fertility. Population and Development Review 18, 599–629.

Pressat, R., 1985. The Dictionary of Demography. Basil Blackwell, Oxford, UK.

Ryder, N.B., 1956. Problems of trend determination during a transition in fertility. Milbank Memorial Fund Quarterly 34, 5–21.

Ryder, N.B., 1964. The process of demographic translation. Demography 1, 74–82.

Ryder, N.B., 1980. Components of temporal variations in American fertility. In: Hiorns, R.W. (Ed.), Demographic Patterns in Developed Societies. Taylor and Francis, London, pp. 15–54.

Ryder, N.B., 1983. Cohort and period measures of changing fertility. In: Bulatao, R.A., Lee, R.D. (Eds.), Determinants of Fertility in Developing Countries. Academic Press, New York, pp. 737–756.

Sobotka, Tomáš, 2004. Is lowest-low fertility explained by the postponement of childbearing? Population and Development Review 30 (2), 195–220.

Whelpton, P.K., 1954. Cohort Fertility: Native White Women in the United States. Princeton University Press, Princeton, NJ (Reissued 1973 Kennikat Press, Port Washington, NY).

Fertility of Single and Cohabiting Women

Lawrence L Wu, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides an overview of current theory and empirical research on the fertility of single or cohabiting parents, with particular emphasis on research in the fields of demography, economics, history, and sociology. Topics covered include: historical shifts in usage concerning the terms employed to refer to these behaviors; measurement issues that arise in the study of nonmarital fertility; demographic trends; functionalist, historical, and policy literatures; empirical research to date on the socioeconomic determinants, including within-country and cross-national variation in the association between nonmarital fertility and socioeconomically disadvantage; consequences for women and children; and the continuing conceptual and theoretical challenges posed by the study of nonmarital fertility.

Historically, fertility outside of formal marriage has been referred to variously as bastardy, illegitimacy, or out of wedlock childbearing, with such terms connoting the inferior status of children born in such circumstances. More recently, the use of these terms has declined, in part because they presume an answer to the question of whether 'inferior status' is a necessary consequence of being born to an unmarried woman or whether some or all of the disadvantage typically accompanying such a birth might stem from other causes. Thus, terms such as nonmarital fertility or childbearing by single or cohabiting parents, which carry more neutral connotations, have seen increasing usage.

Measurement

The measurement of nonmarital fertility requires knowledge of at least two demographic statuses – fertility and marriage – and ideally others – whether a birth occurs within or outside of a cohabiting union. The first-order issue of identifying births that occur within or outside of formal marriage would appear to be a straightforward matter given sufficiently detailed data. Still, it is useful to identify potential inconsistencies across alternative definitions of nonmarital fertility, for example, 'the fertility of single or cohabiting women' versus 'childbearing to unwed parents.' Consider a birth conceived within a marital union in which the father dies before the child is born. Viewed one way, the child was conceived to married parents; viewed another way, the mother was single at birth by virtue of the death of her spouse. Similar issues arise for legitimated births or births to separated parents. In the latter case, because the birth occurs prior to the dissolution of a couples' formal marital union, one might view it as a marital birth. However, the child may have been conceived to the woman and a man other than her husband, either before or after marital separation; thus, another definition might classify such a birth as nonmarital. Note, furthermore, that the period between separation and divorce can be lengthy, with the timing of a formal divorce decree often a response to various factors, including the outcome (e.g., childbearing) under study. Consequently, some researchers have adopted an operational definition that shifts emphasis from formal to *de facto* marital status, thus including as nonmarital those births that occur between separation – marking the informal but behavioral dissolution of a marriage – and divorce – marking the formal dissolution of a marriage.

As a general accounting principle, one might suppose that the number of births to males and females should, in the aggregate, be identical, as when data are available from official birth registers. However, this principle need not hold under some definitions of nonmarital fertility (e.g., when both parents are married, but not to one another) given the types of data typically available to the analyst. Moreover, comparisons of survey and registration data often reveal that marriage and fertility data supplied by male respondents are markedly inferior in quality to those from female respondents, particularly with respect to male nonmarital fertility (Joyner et al., 2012). One consequence has been that there has been far less empirical research on male nonmarital fertility (for exceptions, see, e.g., Barber, 2001; Huang et al., 2005; Hofferth and Goldscheider, 2010; Carlson et al., 2013; Edin and Nelson, 2013). Although such an emphasis is often sensible, in that childrearing responsibilities continue to fall most heavily on women, it has nevertheless produced an asymmetry in research findings.

Additional measurement issues arise when attempting to distinguish between births to cohabiting couples versus those that occur in neither a marital nor cohabiting union. It is sometimes difficult to determine whether or when a cohabiting union has begun or ended, and both contemporaneous and retrospective reports of cohabitation may contain systematic biases (Teitler et al., 2006). These issues have led some to argue that cohabitation might be better conceptualized not as a discrete status but rather in terms of a continuum capturing the extent to which a couple might be said to coreside. Nevertheless, estimates from alternative data sources are often in substantial agreement, providing evidence, albeit of an indirect nature, suggesting that female survey respondents can provide reasonably accurate reports on their union status at birth, even retrospectively.

Demographic Trends

Nonmarital fertility has increased substantially in recent decades in both Europe and the United States. In the United

States, official statistics suggest that as of 2012, more than two in five births (40.7%) were to unmarried women, a level roughly four times higher than that in 1970, at which time one in nine births (10.7%) were to unmarried women. Estimates of cohort trends in the United States likewise reveal monotonic increases in various components of nonmarital fertility (Wu, 2008). Period levels and trends in much of Europe mirror those in the United States. As of the late 1990s, more than every third birth occurred outside of marriage in France, the United Kingdom, Finland, and East Germany, and roughly every second birth was nonmarital in the Scandinavian countries of Iceland, Sweden, Norway, and Denmark (Kiernan, 2001).

Nonmarital, as compared to marital, fertility, continues to occur at relatively early ages in the United States. Increases in women's age at first marriage have necessarily implied corresponding increases in women's ages at a marital first birth; by contrast, period trends in women's ages at a nonmarital first birth between 1970 and the mid-1990s were far more modest (Wu et al., 2001). Thus, differences in the age structure of nonmarital and marital fertility remain substantial in the United States. For example, in 2010, 88.2% of all births to teen women were nonmarital; for women 20–24, nearly two of three births (64.1%) were nonmarital. U.S. nonmarital fertility has also increased substantially among women over 19 years of age. In 1970, one half of all nonmarital births were to teenage women, but by the early 1990s, nonmarital births to women of 20 years of age and older accounted for more than two-thirds of all nonmarital births. These period trends have also been accompanied by marked declines in the likelihood of marriage in certain social and demographic groups, leading some to argue that observed trends result from a decline in male 'marriageability' stemming from a deterioration in men's economic circumstances (Wilson, 1987; Oppenheimer, 1994, 2003) or in declines in the relative wages of males and females (Schultz, 1994; Willis and Haaga, 1996; Willis, 1999; Moffitt, 2001). By contrast, two studies present evidence suggesting that period increases in nonmarital fertility from the late 1960s to the early 1980s (Akerlof et al., 1996) or cohort trends in premarital first births for women born during baby boom (England et al., 2013) were due primarily to decreases in the tendency to respond to a premarital conception by marrying before the birth, leading the former to interpret these trends as a behavioral response to a changing contraceptive regime (e.g., the birth control pill and abortion legalization) and the latter to question hypotheses concerning male marriageability.

Nonmarital fertility in the United States continues to differ substantially by race and ethnicity. In 2010, nonmarital births comprised nearly 3 in 10 of all births to non-Hispanic white women (29.0%), over half of all births to Hispanic women (53.4%), and nearly three of four of all births to non-Hispanic black women (72.5%). These differences also extend to births in cohabiting unions. Births to cohabiting white women have risen sharply in the last two decades (Bumpass and Lu, 2000; Raley, 2001; Kennedy and Bumpass, 2008), accounting for virtually all of the observed increase in nonmarital fertility. A continuing puzzle, however, is that births to cohabiting black women account for a far smaller proportion of nonmarital black births. Fewer than one out of five of recent nonmarital births occurred to cohabiting black women and, over time, births within cohabiting unions have constituted a declining fraction of nonmarital births to black women (Wu et al., 2001).

Nonmarital fertility in the United States and Europe differs in several important respects. The overlap between teen and nonmarital fertility, characteristic historically of the United States, is largely absent in the European context. In addition, births within stable cohabiting unions constitute a far larger fraction of European than of U.S. nonmarital births, with the overwhelming proportion of nonmarital births in Europe occurring within cohabitation (Kiernan, 2001). As a result, recent cohorts of U.S. children are likely to spend more years, on average, growing up without a father present in the child's household than are their European counterparts. Nonmarital fertility in the United States has also been linked to multipartner fertility, family diversity, and family complexity (Meyer et al., 2005; Wu and Li, 2005; Carlson and Furstenberg, 2006; Cherlin, 2010; Furstenberg, 2014). It is not clear if these patterns are distinct to the United States or if they also hold in Europe and elsewhere.

One might expect that the decoupling of both sexual activity and childbearing from marriage and the accompanying erosion of norms proscribing nonmarital fertility (Bumpass, 1990; Pagnini and Rindfuss, 1993; Cherlin, 2010; Smock and Greenland, 2010; Regnerus and Uecker, 2011; Murray, 2012) would imply that nonmarital fertility will have increasingly diffused beyond poor and disadvantaged segments of society to the nonpoor and more highly educated (Whitehead, 1993; Cherlin, 2001). The available empirical evidence is mixed on this point. The prevalence of cohabiting births in Europe and, in particular, Scandinavia has been interpreted by some as consistent with this view (Popenoe, 1996); however, Kiernan (2004) finds that cohabiting births are far more prevalent for the less educated compared to the more highly educated, a pattern that holds across all European countries Kiernan examined and which also holds in the United States and Great Britain, where nonmarital fertility continues to occur predominantly among those with less education and where there is accumulating evidence of diverging trends in the sociodemographic characteristics of married and unmarried mothers (Hoffman and Foster, 1997; Wu et al., 2001; Haveman et al., 2001; McLanahan, 2004; Murray, 2012).

Functionalist, Historical, and Policy Literature

The distinction between social and biological parenthood implicit in the discussion of measurement of nonmarital fertility was an explicit emphasis of early functionalist authors in anthropology and sociology (see, e.g., Malinowski, 1930; Davis, 1939). A classic statement due to Malinowski is:

> The most important moral and legal rule concerning the physiological state of kinship is that no child should be brought into the world without a man – and one man at that – assuming the role of sociological father, that is guardian and protector, the male link between the child and the rest of the community. I think that this generalization amounts to a universal sociological law, and as such I have called it … *The Principle of Legitimacy*. (pp. 137–38).

Accumulating empirical evidence of violations of this principle led Goode (1960) to argue that it was more likely to be followed or enforced when fathers were high in status, as when a father possessed substantial income or assets, or when his stake in the child was high.

By contrast, historians have often used nonmarital fertility as a rough empirical proxy for premarital or illicit sexual activity, while also emphasizing the ebbs and flows in both the incidence of 'bastardy' in different historical periods and the tenuous connection between levels of nonmarital fertility and the intensity of social or moral disapproval (Laslett et al., 1980; Stone, 1979). Moreover, while public and moral authorities typically viewed nonmarital childbearing as concrete evidence of fornication or adultery, real concern is argued by some to have centered around the financial burdens posed to the community by children born out of wedlock. Wells (1979), in particular, concludes that local authorities in colonial America were by and large "unwilling to punish children for the sins of their parents," noting, for example, that "both William Franklin, the last royal governor of New Jersey, and Alexander Hamilton were able to achieve political eminence … despite the fact that their parents were not legally married."

The question of whether the sins of parents should be visited on children remains highly relevant to contemporary policy debates. A continuing policy dilemma is that social and economic resources directed to children born outside of marriage also in effect provide assistance to parents – it is often difficult to target social policies and provisions in ways that benefit children but not parents. This concern has been less evident among policy-makers in Europe, with many European nations having had a longer and more generous tradition of social welfare provisions. It has, however, played a central role in U.S. policy debates, where concern that welfare programs such as Aid to Families with Dependent Children (AFDC) may have had the unintended consequence of increasing nonmarital births was a major factor in the dismantling of these programs in favor of those intended to discourage nonmarital childbearing.

The specific issue of whether more generous AFDC benefits increased nonmarital births in the United States spawned an extensive research literature. The available evidence, constructed largely from cross-sectional U.S. data (see Moffitt, 1998 for a review), finds support for such an association, with most researchers finding that this association is small in magnitude. However, a persistent puzzle generated by these cross-sectional findings is that they appear inconsistent with period trends in nonmarital births and welfare generosity, with nonmarital births increasing and the inflation-adjusted value of AFDC and other benefits falling in recent decades (Ellwood and Jencks, 2004). One possibility is that there is in fact no real inconsistency between the time series and cross-sectional evidence, in that falling AFDC benefits may have acted to reduce nonmarital births, but that other factors may have operated in the opposite direction to increase nonmarital births. Surprisingly little work has been done with respect to this issue, save for Moffitt (2001), who presents evidence in support of this possibility.

These issues were much in evidence when AFDC provisions in the U.S. for poor single mothers and their children were replaced in 1996 by a program entitled 'Temporary Assistance for Needy Families' (TANF), which imposed strict time limits on cash aid to poor single mothers and their children. Potential evidence that time limits did indeed discourage nonmarital fertility is at best mixed. On the one hand, trends in the period nonmarital fertility ratio, defined as the proportion of births to unmarried mothers relative to all births in a given calendar year, ceased increasing in the years immediately following the introduction of TANF; however, such evidence would also appear to suggest that such an effect was confined to a brief 3-year period, with increases in the nonmarital fertility ratio resuming in 2000. It is worth noting that the nonmarital fertility ratio (as well as more general period measures of overall fertility that combine marital and nonmarital births) appears to have been sensitive to economic conditions, with the nonmarital fertility ratio remaining essentially unchanged between 2008 and 2012 (the most recent year for which data were available), thus spanning a 4-year period that coincides with the beginning and aftermath of the serious economic recession of 2008.

Studies that have employed designs intended to better identify a causal effect of welfare reform have yielded mixed results to date. Two studies (Kaestner et al., 2003; Lopoo and DeLeire, 2006) report evidence suggesting a causal link between welfare reform and reductions in teen fertility, while two other studies (Horvath-Rose and Peters, 2001; Hao and Cherlin, 2004) report evidence inconsistent with any such causal link. Yet another study reports evidence suggesting no causal link between welfare reform and observed declines in welfare caseloads (Klerman and Haider, 2004).

Socioeconomic Determinants

A consistent finding in the U.S. literature (see, e.g., McLanahan and Sandefur, 1994) is that the risk of a nonmarital first birth is associated with family structure, with the standard finding being that these risks are higher for women who resided in a nonintact or single-mother family at some point during adolescence. This empirical association appears remarkably robust, for example, to adjustments for unmeasured factors shared by siblings (Powers and Hsueh, 1997) or the tendency for women with premarital first births to have initiated sexual activity slightly earlier than other women (Wu and Martin, 2009).

A difficulty with many of these studies is their reliance on a 'snapshot' of family structure, typically measured at 14 or 16 years of age, which makes problematic the interpretation of what exactly it is about family structure that may account for the higher observed risk of a nonmarital first birth. Somewhat surprisingly, evidence to date does not support widespread expectations concerning the intergenerational transmission of nonmarital childbearing – that being born outside of marriage raises the risk that a young woman will herself bear a child outside of marriage (Wu and Martinson, 1993; Wu, 1996; Haveman et al., 2001). Instead, what appears to be transmitted across generations is the propensity to bear a child at an early age (Barber, 2001; Haveman et al., 2001; Wu and Martin, 2009). Another consistent finding is that low-income and low educational attainments of parents appear substantially associated with premarital first birth

risks (Haveman and Wolfe, 1994; McLanahan and Sandefur, 1994; Wu, 1996; Ermisch, 2001). Four studies find that turbulent family environments, in the form of repeated changes in family structure, are associated with higher risks of a premarital first birth for females (Wu and Martinson, 1993; Haveman and Wolfe, 1994; Wu, 1996) or for males but not females (Hofferth and Goldscheider, 2010).

Consequences for Women and Children

Nonmarital childbearing is commonly thought to carry adverse consequences, both for the woman bearing the child and for the children born into such families. An early view capturing this sentiment (Campbell, 1968) held that:

> The girl who has an illegitimate child at the age of 16 suddenly has 90% of her life's script written for her. She will drop out of school; even if someone else in her family helps to take care of her baby, she will probably not be able to find a steady job that pays enough to provide for herself and her child; she may feel impelled to marry someone she might not otherwise have chosen. Her life choices are few, and most of them are bad.

Although this view remains highly influential in many respects, the research community has increasingly regarded it with substantial skepticism. Furstenberg et al. (1987) find substantial variability in the life course outcomes for women who began childbearing while in their teen years, many of whom had births while unmarried, leading many to question if a woman's life script is in fact '90% written' as a consequence of teen (and nonmarital) childbearing. Wu et al. (2001), using nationally representative U.S. data, document similar variability for women with nonmarital births, suggesting that the results of Furstenberg and colleagues may also hold for more recent cohorts of women. Another basis for skepticism is the realization by researchers that previous studies may have overstated the magnitude of the negative effects of teen childbearing. Indeed, a considerable controversy has arisen over whether teen childbearing might be held to have any causal relationship to the later social and economic disadvantages observed for teen mothers (see, e.g., Geronimus and Korenman, 1993; Grogger and Bronars, 1993; Hoffman et al. 1993a,b; Rosenzweig and Wolpin, 1995; Hotz et al., 2005; Lee, 2010; Stange, 2011).

Similar issues arise when evaluating findings from the extremely large empirical literature that has examined outcomes for children who resided at some point with a single parent in that this literature rarely distinguishes between whether such families were formed through divorce or nonmarital fertility (Seltzer, 2000). Thus, much of this literature is in fact silent on outcomes for children born to unmarried mothers, aside from documenting that these children are substantially more likely to grow up in poverty. As noted above, several studies find no empirical evidence in support of the transmission of nonmarital childbearing across generations. Korenman et al. (2001) find that the higher prevalence of unintended births among unmarried mothers relative to married mothers appears to account for little of the observed differential in parenting behaviors, infant health, or child development between marital and nonmarital births. Haveman et al. (2001) find that children born to unmarried mothers are significantly less likely to complete high school than children born to married parents.

The research literature has also been slow in providing evidence on the consequences of being born to cohabiting parents. One well-documented finding is that mean income is higher for married relative to cohabiting families and lowest in single-mother families, with these differences affecting the economic resources available to children residing in these families (see, e.g., Manning and Lichter, 1996; Carlson and Danziger, 1999; McLanahan, 2004; Manning and Brown, 2006). Empirical studies of Great Britain and the United States (Ermisch, 2001; Wu et al., 2001; Manning et al., 2004; Raley and Wildsmith, 2004; Wu and Musick, 2008) find that children born within a cohabiting union are far more likely to experience the loss of a parent – most typically the father – through the dissolution of their parent's union than are children born within marital unions. One study, comparing cohabitors who married before a first birth with cohabitors who married after a first birth, found no significant difference between these two groups in union stability following first birth, suggesting perhaps that decisions concerning marriage and childbearing may be made jointly by at least some cohabiting couples (Wu and Musick, 2008). Findings concerning the greater instability for children born to cohabiting versus married parents must also be weighed in light of evidence suggesting that this is far less true in continental Europe than in the United States or Great Britain (Kiernan, 2001, 2004). It is also worth emphasizing that what is observed – the presence or absence of a marital or cohabiting union – is but a crude proxy for the economic, social, and emotional commitments of the parents to the child, commitments which may, in turn, vary substantially over time.

Conceptual and Theoretical Challenges

Unlike much empirical research, theoretical work has taken seriously the 'two-sex' nature of nonmarital fertility – for example, how nonmarital fertility might respond to both men's and women's characteristics. Notable examples include Wilson's (1987) argument, reviewed above, concerning a dearth of 'marriageable' males in the U.S. black community and Willis' economic models of nonmarital fertility (Willis and Haaga, 1996; Willis, 1999), in which different distributions in the relative wages of men and women may produce behavioral equilibria in which men move from a pattern of high to low paternal investment in the children they father. Neal (2001) proposes a model that implies that the prevalence of nonmarital fertility can vary substantially if the choice set for women is expanded to consist not only of giving birth within or outside of marriage, as assumed by both Wilson and Willis, but also of the option of foregoing a birth entirely.

The study of nonmarital fertility poses other unusual challenges, both theoretically and empirically. One is that nonmarital fertility lies at the intersection of several joint (and potentially

endogenous) processes by which families are formed – relationships, cohabitation, marriage, and childbearing. A second challenge concerns the likely life course dynamics underlying nonmarital fertility. Empirical research is complicated by the fact that many key aspects relevant to this process are either unobserved or at best partially observed by the researcher, including the timing of initial (and ensuing frequency of) sexual activity, contraception among those who are sexually active, abortion, and how the resolution of a nonmarital pregnancy is regarded by the woman and by other relevant actors. A further challenge is that recent trends in nonmarital fertility may well reflect an evolving response by individuals to fundamental changes in the long-term contract represented by marriage, a contract that has itself changed in response to other social forces, including the near-universal levels of sexual activity and cohabitation prior to marriage, changes in the nature and levels of labor force participation of men and women, and continuing high levels of divorce and union instability (Seltzer, 2000; Lundberg, 2001; Cherlin, 2010; Furstenberg, 2014).

Acknowledgment

Research funding from NICHD (R01 HD 29550) is gratefully acknowledged.

See also: Abortion, Demography of: United States; Cohabitation: United States; Concentrated Disadvantage; Demographic Measurement: General Issues and Measures of Fertility; Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Family Planning: Program Effects; Family Theory: Competing Perspectives in Social Demography; Family Theory: Economics of Childbearing; Family Theory: Economics of Marriage and Divorce; Family as Institution; Fatherhood; Fertility Theory; Population Dynamics: Two-Sex Demographic Models; Poverty, Sociology of; Sexual Debut.

Bibliography

Akerlof, G.A., Yellen, J.L., Katz, M.L., 1996. An analysis of out-of-wedlock childbearing in the United States. Quarterly Journal of Economics 111, 277–317.

Barber, J.S., 2001. The intergenerational transmission of age at first birth among married and unmarried men and women. Social Science Research 30, 219–247.

Bumpass, L.L., 1990. What's happening to the family: Interaction between demographic and institutional change. Demography 27, 483–498.

Bumpass, L.L., Lu, H.-H., 2000. Trends in cohabitation and implications for children's family contexts. Population Studies 54, 29–41.

Campbell, A.A., 1968. The role of family planning in the reduction of poverty. Journal of Marriage and the Family 30, 236–245.

Carlson, M.J., Danziger, S., 1999. Cohabitation and the measurement of child poverty. Review of Income and Wealth 45, 179–191.

Carlson, M.J., Furstenberg Jr., F.F., 2006. The prevalence and correlates of multipartnered fertility among urban U.S. parents. Journal of Marriage and the Family 68, 718–732.

Carlson, M.J., VanOrman, A.G., Pilkauskas, N.V., 2013. Examining the antecedents of U.S. nonmarital fatherhood. Demography 50, 1421–1447.

Cherlin, A.J., 2001. New developments in the study of nonmarital childbearing. In: Wu, L.L., Wolfe, B. (Eds.), Out of Wedlock: Causes and Consequences of Nonmarital Fertility. Russell Sage Foundation, New York, pp. 390–402.

Cherlin, A.J., 2010. Demographic trends in the United States: a review of research in the 2000s. Journal of Marriage and the Family 72, 403–419.

Davis, K., 1939. Illegitimacy and social structure. American Journal of Sociology 45, 215–233.

Edin, K., Nelson, T.J., 2013. Doing the Best I Can: Fatherhood in the Inner City. University of California Press, Berkeley.

Ellwood, D.T., Jencks, C., 2004. The uneven spread of single-parent families: what do we know? Where do we look for answers? In: Neckerman, K.M. (Ed.), Social Inequality. Russell Sage Foundation, New York, pp. 3–78.

England, P., Wu, L.L., Shafer, E.F., 2013. Cohort trends in premarital first births: what role for premarital conception vs. the retreat from marriage? Demography 50, 2075–2104. Online version http://dx.doi.org/10.1007/s13524-013-0241-1.

Ermisch, J., 2001. Cohabitation and childbearing outside of marriage in Britain. In: Wu, L.L., Wolfe, B. (Eds.), Out of Wedlock: Causes and Consequences of Nonmarital Fertility. Russell Sage Foundation, New York, pp. 109–139.

Furstenberg, F.F., 2014. Fifty years of family change: from consensus to complexity. The Annals of the American Academy of Political and Social Science 654, 12–30.

Furstenberg Jr., F.F., Brooks-Gunn, J., Morgan, S.P., 1987. Adolescent Mothers in Later Life. Cambridge University Press, Cambridge.

Geronimus, A., Korenman, S., 1993. The socioeconomic costs of teenage childbearing: evidence and interpretation. Demography 30, 281–296.

Goode, W.J., 1960. Illegitimacy in the caribbean social structure. American Sociological Review 25, 21–30.

Grogger, J., Bronars, S.G., 1993. The socioeconomic consequences of childbearing: findings from a natural experiment. Family Planning Perspectives 25, 156–161, 174.

Hao, L., Cherlin, A.J., 2004. Welfare reform and teenage pregnancy, childbirth, and school dropout. Journal of Marriage and the Family 66, 179–194.

Huang, C.-C., Mincy, R.B., Garfinkel, I., 2005. Child support obligations and low-income fathers. Journal of Marriage and the Family 67, 1213–1225.

Haveman, R., Wolfe, B., 1994. Succeeding Generations: On the Effects of Investments in Children. Russell Sage Foundation, New York.

Haveman, R., Wolfe, B., Pence, K., 2001. Intergenerational effects of nonmarital and early childbearing. In: Wu, L.L., Wolfe, B. (Eds.), Out of Wedlock: Causes and Consequences of Nonmarital Fertility. Russell Sage Foundation, New York, pp. 287–316.

Hofferth, S.L., Goldscheider, F., 2010. Family structure and the transition to early parenthood. Demography 47, 415–437.

Hoffman, S.D., Foster, E.M., 1997. Nonmarital births and single mothers: cohort trends in the dynamics of nonmarital childbearing. History of the Family 2, 255–275.

Hoffman, S.D., Foster, E.M., Furstenberg Jr., F.F., 1993a. Reevaluating the costs of teenage childbearing. Demography 30, 1–13.

Hoffman, S.D., Foster, E.M., Furstenberg Jr., F.F., 1993b. Reevaluating the costs of teenage childbearing: response to Geronimus and Korenman. Demography 30, 291–296.

Horvath-Rose, A., Peters, H.E., 2001. Welfare waivers and nonmarital childbearing. In: Duncan, G.J., Chase-Lansdale, P.L. (Eds.), For Better and for Worse: Welfare Reform and the Well-being of Children and Families. Russell Sage Foundation, New York, pp. 222–245.

Hotz, V.J., McElroy, S.W., Sanders, S.G., 2005. Teenage childbearing and its life cycle consequences: exploiting a very natural experiment. Journal of Human Resources 40, 683–715.

Joyner, K., Peters, H.E., Hynes, K., Sikora, A., Rubenstein, J., Rendall, M., 2012. Quality of male fertility data in major U.S. surveys. Demography 49, 101–124.

Kaestner, R., Korenman, S., O'Neill, J., 2003. Has welfare reform changed teenage behaviors? Journal of Policy Analysis and Management 22, 225–248.

Kennedy, S., Bumpass, L.L., 2008. Cohabitation and children's living arrangements: new estimates from the United States. Demographic Research 19, 1663–1692.

Kiernan, K., 2001. European perspectives on nonmarital childbearing. In: Wu, L.L., Wolfe, B. (Eds.), Out of Wedlock: Causes and Consequences of Nonmarital Fertility. Russell Sage Foundation, New York, pp. 77–108.

Kiernan, K., 2004. Unmarried cohabitation and parenthood: here to stay? European perspectives. In: Moynihan, D.P., Smeeding, T.M., Rainwater, L. (Eds.), The Future of the Family. Russell Sage Foundation, New York, pp. 66–95.

Klerman, J.A., Haider, S.J., 2004. A stock-flow analysis of the welfare caseload. Journal of Human Resources 39, 865–886.

Korenman, S., Kaestner, R., Joyce, T.J., 2001. Unintended pregnancy and the consequences of nonmarital childbearing. In: Wu, L.L., Wolfe, B. (Eds.), Out of Wedlock: Causes and Consequences of Nonmarital Fertility. Russell Sage Foundation, New York, pp. 259–286.

Laslett, P., Oosterveen, K., Smith, R.M. (Eds.), 1980. Bastardy and its Comparative History. Harvard University Press, Cambridge, MA.

Lee, D., 2010. The early socioeconomic effects of teenage childbearing: a propensity score matching approach. Demographic Research 23, 697–736.

Lopoo, L.M., DeLeire, T., 2006. Did welfare reform influence the fertility of young teens? Journal of Policy Analysis and Management 25, 275–298.
Lundberg, S., 2001. Nonmarital fertility: lessons for family economics. In: Wu, L.L., Wolfe, B. (Eds.), Out of Wedlock: Causes and Consequences of Nonmarital Fertility. Russell Sage Foundation, New York, pp. 383–389.
Malinowski, B., 1930. Parenthood, the basis of social structure. In: Calverton, V.F., Schmalhausen, S.D. (Eds.), The New Generation: The Intimate Problems of Modern Parents and Children. Macaulay, New York, pp. 113–160.
Manning, W.D., Brown, S., 2006. Children's economic well-being in married and cohabiting parent families. Journal of Marriage and the Family 68, 345–362.
Manning, W.D., Lichter, D.T., 1996. Parental cohabitation and children's economic well-being. Journal of Marriage and the Family 58, 998–1010.
Manning, W., Smock, P., Majumdar, D., 2004. The relative stability of cohabiting and marital unions for children. Population Research and Policy Review 23, 135–159.
McLanahan, S.S., 2004. Diverging destinies: how children are faring under the second demographic transition. Demography 41, 607–627.
McLanahan, S.S., Sandefur, G., 1994. Growing Up with a Single Parent. Harvard University Press, Cambridge, MA.
Meyer, D.R., Cancian, M., Cook, S.T., 2005. Multiple-partner fertility: Incidence and implications for child support policy. Social Service Review 79, 577–601.
Moffitt, R.A., 1998. The effect of welfare on marriage and fertility. In: Moffitt, R.A. (Ed.), Welfare, the Family, and Reproductive Behavior: Research Perspectives. National Academy Press, Washington, DC, pp. 50–97.
Moffitt, R.A., 2001. Welfare benefits and female headship in U.S. time series. In: Wu, L.L., Wolfe, B. (Eds.), Out of Wedlock: Causes and Consequences of Nonmarital Fertility. Russell Sage Foundation, New York, pp. 143–172.
Murray, C., 2012. Coming Apart: The State of White America, 1960–2010. Crown Forum, New York.
Neal, D., 2001. The Economics of Family Structure. NBER Working Paper No. 8519.
Oppenheimer, V.K., 1994. Women's rising employment and the future of the family in industrial societies. Population and Development Review 20, 293–342.
Oppenheimer, V.K., 2003. Cohabiting and marriage during young men's career development process. Demography 40, 127–149.
Pagnini, D., Rindfuss, R.R., 1993. The divorce of marriage and childbearing: changing attitudes and behavior in the United States. Population and Development Review 19, 331–347.
Popenoe, D., 1996. Life Without Father. Free Press, New York.
Powers, D.A., Hsueh, J.C.-T., 1997. Sibling models of socioeconomic effects on the timing of first premarital birth. Demography 34, 493–511.
Raley, R.K., 2001. Increasing fertility in cohabiting unions: evidence for the second demographic transition in the United States? Demography 38, 59–66.
Raley, R.K., Wildsmith, E., 2004. Cohabitation and children's family instability. Journal of Marriage and the Family 66, 210–219.
Regnerus, M., Uecker, J., 2011. Premarital Sex in America: How Young Americans Meet, Mate, and Think about Marrying. Oxford University Press, New York.
Rosenzweig, M.R., Wolpin, K.I., 1995. Sisters, siblings, and mothers: the effect of teen-age childbearing on birth outcomes in a dynamic family context. Econometrica 63, 303–326.
Schultz, T.P., 1994. Marital status and fertility in the United States: welfare and labor market effects. Journal of Human Resources 29, 636–669.
Seltzer, J.A., 2000. Families formed outside of marriage. Journal of Marriage and the Family 62, 1247–1268.
Smock, P.J., Greenland, F.R., 2010. Diversity in pathways to parenthood: patterns, implications, and emerging research directions. Journal of Marriage and the Family 72, 576–593.
Stange, K., 2011. A longitudinal analysis of the relationship between fertility timing and schooling. Demography 48, 931–956.
Stone, L., 1979. The Family, Sex, and Marriage, abridged ed. Harper, New York.
Teitler, J.O., Reichman, N.E., Koball, H., 2006. Contemporaneous versus retrospective reports of cohabitation in the Fragile Families survey. Journal of Marriage and the Family 68, 469–477.
Wells, R.V., 1979. Illegitimacy and bridal pregnancy in colonial America. In: Laslett, P., Oosterveen, K., Smith, R.M. (Eds.), Bastardy and its Comparative History. Harvard University Press, Cambridge, MA, pp. 349–361.
Whitehead, B.D., April 1993. Dan Quayle was right. Atlantic Monthly 47–84.
Willis, R.J., 1999. A theory of out-of-wedlock childbearing. Journal of Political Economy 106, S33–S64.
Willis, R.J., Haaga, J.G., 1996. Economic approaches to understanding nonmarital fertility. In: Casterline, J.B., Lee, R.D., Foote, K.A. (Eds.), Fertility in the United States: New Patterns, New Theories, pp. 67–86.
Wilson, W.J., 1987. The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy. University of Chicago Press, Chicago.
Wu, L.L., 1996. Effects of family instability, income, and income instability on the risk of a premarital birth. American Sociological Review 61, 386–406.
Wu, L.L., 2008. Cohort estimates of nonmarital fertility. Demography 45, 193–207.
Wu, L.L., Bumpass, L.L., Musick, K., 2001. Historical and life course trajectories of nonmarital childbearing. In: Wu, L.L., Wolfe, B. (Eds.), Out of Wedlock: Causes and Consequences of Nonmarital Fertility. Russell Sage Foundation, New York, pp. 1–48.
Wu, L.L., Li, J.-C.A., 2005. Historical roots of family diversity: marital and childbearing trajectories of American women. In: Settersten, R.A., Furstenberg Jr., F.F., Rumbaut, R.G. (Eds.), On the Frontier of Adulthood: Theory, Research, and Public Policy. University of Chicago Press, Chicago, pp. 110–149.
Wu, L.L., Martin, S.P., 2009. Effects of exposure on prevalence and cumulative risk: direct and indirect effects in a recursive hazard model. Sociological Methodology 39, 185–232.
Wu, L.L., Musick, K., 2008. Stability of marital and cohabiting unions following a first birth. Population Research and Policy Review 27, 713–727.
Wu, L.L., Martinson, B.C., 1993. Family structure and the risk of a premarital birth. American Sociological Review 58, 210–232.

Fertility Theory

Nicoletta Balbo, Dondena Centre, Bocconi University, Milan, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides a review of fertility research in advanced societies, where birth control is the default option. The central aim is to provide a comprehensive review that summarizes how contemporary research has explained ongoing and expected fertility changes across time and space. The core determinants of fertility are grouped according to the analytical level in which they operate. Determinants are positioned at the individual/couple level (micro-level), social networks (meso-level), and by cultural and institutional settings (macro-level). The focus is on both quantum and tempo of fertility.

Introduction

During the last two decades, research about fertility in advanced societies – societies in which birth control is the default option – has flourished.

Since the 1980s, an increasing number of European and Asian countries reached very low fertility levels (Caldwell and Schindlmayr, 2003) and virtually all advanced societies witnessed a 'postponement' transition (Kohler et al., 2002a). For this reason, the topic of low-fertility-dominated fertility research in advanced societies. During the late 2000s, a reversal of the fertility decline in most advanced countries – albeit with great heterogeneity – has drawn considerable attention (Goldstein et al., 2009; Myrskylä et al., 2009; OECD, 2011). Another emerging topic is the impact of economic uncertainty and the recent economic recession on fertility (Mills and Blossfeld, 2005; Sobotka et al., 2011).

In order to illustrate the state of fertility research and the main fertility theories, this article provides a contemporary review of research and delineation of the core determinants of fertility. Determinants are categorized according to the analytical level at which they operate. Specifically, three levels can be identified: the micro-level, that is, the individual and/or couple level; the meso-level, which encompasses social relationships and social networks (i.e., characteristics pertaining to the network of friends, coworkers, relatives, and/or the relationship that links them to the individual); and the macro-level, which refers to the cultural and institutional settings where individuals and couples are embedded.

While distinguishing the different analytical levels, the two main components of fertility will as well be taken into account: tempo (i.e., the timing of childbirth) and quantum (i.e., the total number of children). When examining tempo, it is known that OECD countries have witnessed a rising mean age at first birth since the 1970s, coupled with an increasing proportion of births among mothers at advanced ages, albeit with considerable country-level variation (Billari et al., 2007; Sobotka et al., 2010; Mills et al., 2011). This process is generally referred to as the postponement of childbearing, which is the central focus of tempo studies in fertility research. The emergence of a 'postponement transition' of higher ages at first birth is a core phenomenon that has emerged in the last decades (Kohler et al., 2002a; Billingsley, 2010).

Quantum is related to the number of children (including childlessness). While demographic transition theory had implied that the quantum of fertility would stabilize around the replacement level of around 2.1 children per woman (e.g., Bongaarts, 2002), during the last decades of the twentieth century, advanced societies witnessed low, or even the 'lowest-low' or 'ultra-low' fertility levels close to 1 (e.g., Kohler et al., 2002a; Frejka et al., 2010). The novelty of this phenomenon captivated demographers and raised the concerns of policy-makers about both the long-term demographic sustainability of their populations and concerns about the growing gap between desired and achieved fertility (Chesnais, 1996; Bongaarts, 2001; Goldstein et al., 2003; OECD, 2011). Just as there appeared to be a consensus that advanced societies would either remain at low fertility levels or reach even lower levels (e.g., Lutz et al., 2003), new research produced evidence demonstrating onset of a fertility reversal during the 2000s (Goldstein et al., 2009; Myrskylä et al., 2009; OECD, 2011).

The remainder of this article is structured as follows. The following section focuses on the determinants of fertility at the micro-level, followed by a section illustrating factors acting at the meso-level. The final section reviews fertility research at the macro-level.

Micro-Level Determinants of Fertility

Research at the micro-level focuses on the individual and/or couple decision-making process underlying the choice to have a child and investigates which circumstances affect decisions regarding the timing and number of children. Choices are often characterized as a rational response to uncertainty and/or as conformity to the prescribed 'sequencing' of life course events. A rich body of research has studied the link between life course circumstances and trajectories (mainly, partnership, education, employment, and economic conditions) and fertility behavior. Other research has instead focused on the role of social class and family, or origin, which in turn shapes an individual's values and preferences. A major obstacle in this line of research is the challenge to establish causality, or in other words, the ability to empirically determine whether these life course factors are actual determinants of fertility or whether certain life course factors and fertility behavior are simultaneously affected

by the presence of other common determinants (endogeneity or reverse causality).

Fertility Intentions

Many scholars have used fertility intentions as a proximate determinant for actual fertility behavior, examining which factors influence the formation, realization, and/or change in fertility intentions (e.g., Westoff and Ryder, 1977; Schoen et al., 1999; Quesnel-Vallée and Morgan, 2003; Berrington, 2004; Liefbroer, 2009; Philipov, 2009; Spéder and Kapitány, 2009). Others have identified the systematic gap between intended and actual fertility as one of the causes of low fertility (Morgan and Taylor, 2006), since it reveals an 'unmet need' for children (Liefbroer, 2009).

Next to critical approaches that downplay the explanatory power of intentions and less individual-specific measures such as ideal family size (e.g., Westoff and Ryder, 1977; Quesnel-Vallée and Morgan, 2003), other studies argue that intentions are effective predictors of actual fertility (Schoen et al., 1999; Berrington, 2004). The majority of these latter studies draw upon the social psychological literature, often employing the Theory of Planned Behavior (TPB) (Ajzen, 1991). According to the TPB, intentions are the culmination of a combination of three antecedents: (1) attitudes (i.e., perceived costs and benefits); (2) subjective norms (e.g., influence of close friends and relatives); and (3) perceived control over behavior (i.e., extent to which behavior is perceived as subject to control by the individual).

Partner and Partnership

Changes in partnership dynamics experienced in the past decades across advanced societies has been linked to the postponement of parenthood. A growing number of studies have shown a parallel tendency to delay union formation and parenthood (Corijn and Klijzing, 2001), an increased frequency to have several partners before the first child (Wu and Schimmele, 2005), and a rise in unmarried cohabitation, which has been associated with a later age at entry marriage (Bumpass et al., 1991; Mills, 2004), if not a 'retreat from marriage' (Gibson-Davis et al., 2005).

Partnership status (i.e., cohabitation and marriage) (Hobcraft and Kiernan, 1995; Philipov et al., 2006; Testa, 2006) as well as a variety of other characteristics of both partners or of the couple have likewise been found to influence childbearing (e.g., Thomson et al., 1990; Corijn et al., 1996; Thomson and Hoem, 1998; Jansen and Liefbroer, 2006).

With the growth of more unstable relationships and higher levels of separation and divorce, another relevant topic has been the influence of union (in)stability and/or low relationship quality on childbearing. Although causal links are complex and there are important feedback mechanisms (Waite and Lillard, 1991), the existing literature provides evidence for two opposing mechanisms. Some studies find a negative relationship between low quality/instability of partnership and having children (Thornton, 1978; Myers, 1997). On the other hand, Friedman et al. (1994) argue that union instability leads to earlier childbearing since children are seen as a source of uncertainty reduction and thereby operates to enhance marital solidarity.

Gendered Division of Labor

Another important factor influencing fertility is the gendered division of domestic labor of couples within the household. Contemporary work builds on McDonald's (2000a,b) gendered fertility theory, which argues that very low fertility is the result of a hiatus of sustained gender inequity in family-oriented social institutions. Esping-Andersen (2009), also drawing on the work by the economist Goldin (2006), sees low fertility as a consequence of the 'incompleteness' of the revolution that transformed women's roles.

Stepfamily Fertility

The increase in unstable and multiple unions has also brought a growth in the study of stepfamily fertility. This body of research demonstrates that partners who already have children from previous unions are more likely to have a child together, often considered as a union commitment effect (e.g., Vikat et al., 1999; Buber and Prskawetz, 2000; Stewart, 2002; Prskawetz et al., 2003). Given that one child is enough to indicate commitment in a partnership, multiple relationships and subsequent partnerships might significantly contribute to total fertility.

Income, Education, and Human Capital

Socioeconomic circumstances of individuals have also been studied as determinants of fertility quantum and timing. Income (and wages in particular) has attracted considerable research interest in economics. Depending on the economic model that is adopted, the effect might be different. The family economics approach, pioneered by Becker (1960), maintains that individuals obtain direct pleasure from having and raising children, and from their well-being. Children, and possibly their quality level, thus resemble a consumption good in the utility function of their parents. While the initial formulation of this theory implies a positive link between income and number of children, the large body of literature that followed focused on a negative relationship between income and fertility, emphasizing two aspects. First approach focuses on the quality–quantity tradeoff, proposed by Becker and Lewis (1973) and Willis (1973). Here the argument is that an increase in income may lead to fewer children. This is attributed to the fact that parents with a higher income value children's quality, but a focus on higher-quality raises the cost of having (and raising) children, thereby potentially reducing fertility levels. A second approach focuses on the opportunity cost of having children, especially for women. Since raising children requires parental (and especially maternal) time, fertility is more costly for higher-income mothers, who are therefore expected to have fewer children (e.g., Kravdal, 1992). This is related to the literature that demonstrates a 'motherhood wage penalty,' with postponement of motherhood providing considerable earnings returns for higher educated women or those in professional occupations (Van Bavel, 2010; Begall and Mills, 2012). Others have extended Becker's static model by setting up dynamic economic models of the optimal timing of first birth (mostly focusing on women), based on the minimization of opportunity costs of childbearing, 'wage penalty,' and income loss (Happel et al., 1984; Cigno and Ermisch,

1989). They theorized and demonstrated that the higher the accumulation of human capital during education or the higher the returns to education, the later the transition to motherhood.

A similar approach has been adopted by the rich body literature that focuses on the relationship between education and labor market trajectories and the timing of first birth. The argument is that due to the accumulation of human capital, women with higher levels of education are more likely to pursue careers and increase their earning power. This likewise releases them from the pressure to get married and have a child for economic reasons. As the opportunity costs of childbearing and childrearing increase with human capital, highly educated women are more likely to postpone marriage and births. An important critique of this approach has been put forth by Oppenheimer (1994), who argues that highly educated women are more likely to find partners who are highly educated as well. This in turn operates as an incentive for women (because they can further pool economic resources) to enter into a union and subsequently have children once they complete their education. Although higher educated women have their first child later than their lower educated counterparts, some studies (Sobotka, 2004; Kravdal and Rindfuss, 2008) have highlighted that the higher educated are also more likely to recuperate at a later age.

The results linking education to fertility are, however, mixed, with recent empirical results instead showing a nonrelevant association between education and fertility (e.g., Skirbekk, 2008).

Other researchers have instead focused on the importance of educational enrollment, as opposed to the highest achieved level of education. Here findings show that individuals who are still enrolled in education are at a lower risk of having a child, likely attributed to the presence of a 'sequencing norm' of first finishing education, followed by parenthood (Hoem, 1986; Blossfeld and Huinink, 1991). Others have extended this research to also examine the importance of the educational field of study in relation to either socialization or self-selection effects into later occupations, which in turn impact fertility (e.g., Van Bavel, 2010; Begall and Mills, 2012).

Economic and Employment Uncertainty

Further, mostly sociological studies focus on the importance of employment status and, particularly, economic uncertainty on fertility outcomes. Theories of (largely economic) uncertainty are reminiscent of Easterlin's (1976) theory of economic deprivation, which posits that in historical periods of general economic uncertainty and rising unemployment, the tendency to marry and have children appears to diminish. This also relates to Oppenheimer's (1988, 2003) work on the impact of uncertainty in social and economic roles on the timing of family transitions. An increasing number of studies link economic uncertainty – often in the form of unemployment and precarious labor market situations – to the postponement of parenthood (Mills and Blossfeld, 2005; Kreyenfeld, 2010).

Fertility Preferences

An individual's fertility decisions are shaped by his or her own preferences, which several authors emphasize, are shaped early in an individual's life. Catherine Hakim's Preference Theory (2003) positions the heterogeneity of women's lifestyle preferences at the heart of fertility (and labor market) choices in advanced societies. Hakim assumes that lifestyle preferences are rather constant across the life course, with three main types: career oriented, family oriented, and those oriented toward combining work and family. These lifestyle preferences are seen as the main driver – with policies required that take into account this heterogeneity.

Research on preferences for the sex composition of children shows an effect of sex preferences on the probability to have more children (Andersson et al., 2006; Mills and Begall, 2010), albeit with considerable variation across countries (Hank and Kohler, 2000).

Intergenerational Transmission of Values and Behavior

The similarities of fertility histories across successive generations has also been a core area of research, focusing mainly on the stable result of a positive correlation between the number of siblings and number of own children (Thornton, 1980; Anderton et al., 1987; Axinn et al., 1994; Murphy and Wang, 2001) or between the age at first birth of parents and that of their children (Rijken and Liefbroer, 2009). The majority of studies concerning tempo focus on teenage motherhood, demonstrating that having had a young mother increases the risk of having a child at a young age (Furstenberg et al., 1990; Manlove, 1997). Within this body of literature, the intergenerational transmission of behavior is considered to be driven by intra-familial socialization processes that occur during childhood and adolescence (Thornton, 1980; Axinn et al., 1994; Murphy and Wang, 2001). The assumption is that parents transmit family values, preferences, and attitudes, as well as contraceptive knowledge.

Biodemography of Fertility

Next to socialization mechanisms, biological and genetic factors have also been used to explain intergenerational similarities in fertility preferences and behavior (Wachter and Bulatao, 2003). A series of studies have linked biological and genetic components to fertility behavior (Kohler et al., 1999, 2002a; Kohler and Rodgers, 2003). This body of research focuses on studying the interplay between fertility, environment, and genetic makeup of individuals and demonstrates that differences in the genetic composition of individuals affect their fertility outcomes and fertility-related behavior. There is also growing evidence that genetic variance changes over time and across educational levels, meaning that the importance of social norms and individual decision-making change across time (Kohler et al., 1999, 2002a; Kohler and Rodgers, 2003). The biology of fertility has also been revolutionized by the rapid diffusion of various types of assisted reproductive technologies (ART).

Socioeconomic Status and Cultural Context of Family of Origin

A related stream of research investigates the effect of the socioeconomic and cultural context of the family of origin on an individual's age at childbearing and fertility quantum decisions. Some studies have shown that there is a negative relationship between parents' educational level (especially father's education) and age at first birth (Blossfeld and

Huinink, 1991; Billari, 2001) as well as the number of children (Murphy and Wang, 2001; Rijken and Liefbroer, 2009). A negative relationship has also been found between parents' financial situation (i.e., job status) and (expected) number of children (Thornton, 1980; Murphy and Wang, 2001) and age at first birth (Rijken and Liefbroer, 2009). These findings suggest that in higher educated and high-status families, other goals beyond family formation are more easily transmitted, together with aspirations for material goods (Easterlin, 1969). Therefore, if consumption aspirations are high, parenthood will be reduced or at least postponed. For the opposite reason, parents' religiosity is positively associated with their children's fertility (Rijken and Liefbroer, 2009).

Meso-Level Determinants of Fertility

In recent years, increasing attention has been devoted to developing and applying theories that take into account the fact that individuals are positioned as social actors who make decisions and act while embedded in a web of social relationships with kin and peers. While some studies focus on the role of interpersonal interactions in shaping an individual's fertility decisionmaking, others investigate how the place of residence is associated with reproductive choices. Finally, another body of research looks at the social network as a source of social capital in the form of emotional and material aid. However, the acknowledgment of the importance of the social network in explaining observed fertility patterns has not as of yet been coupled with a convincing body of empirical research. The central reason rests with the lack of suitable data and the difficulty to model and properly identify social interaction effects. The social context or other individual-level factors common among individuals can also explain similarities in behavior (e.g., same timing of childbearing), and the social network may be chosen according to desired behaviors and changes over time.

Social Interaction

The impact of social interaction on fertility choices has received considerable attention (e.g., Bongaarts and Watkins, 1996; Montgomery and Casterline, 1996). This literature has predominately focused on the diffusion of contraceptive methods in developing countries (Kohler et al., 2001) and identified two channels via which social interaction takes place. The first is social learning, or how individuals gain knowledge from others, and the second is social influence, which is how consensus in peer groups constrains attitudes and behavior.

In advanced societies, research illustrates that personal network contacts such as coworkers, siblings, or friends are pivotal in shaping both fertility timing as well as quantum decisions (Bernardi et al., 2007; Axinn et al., 1994; Manski and Mayshar, 2003; Lyngstad and Prskawetz, 2010; Balbo and Mills, 2011; Balbo, 2012).

Place of Residence

Several studies document fertility differences by place of residence along several lines. First, fertility continues to vary across regions (Caltabiano, 2008; Kertzer et al., 2009). According to Kulu (2011), differences in desired family size explain fertility differentials between small towns and rural areas on the one hand, and urban areas on the other. Second, within urban areas, suburbs are consistently found to have higher fertility (Kulu et al., 2009), with single-family households associated with higher fertility (Kulu and Vikat, 2007). These differentials persist when controlling for the socioeconomic composition of areas (Kulu and Boyle, 2009), suggesting that contextual effects shape fertility decisions. However, the association between the place of residence and fertility choices raises causality concerns, since the relation might be spurious and the effect can work in both directions.

Social Capital

The body of research focusing on the social network as a source of social capital (Bühler and Philipov, 2005; Philipov et al., 2006) stems largely from sociological theory (Lin et al., 2001; Flap and Völker, 2004). Social capital is defined as the resources that individuals have access to via personal relationships and can include goods, as well as information, money, capacity to work, influence, power, or active help (Granovetter, 1973; Bourdieu, 1986; Coleman, 1988). Building on previous sociological applications (Granovetter, 1973; Bourdieu, 1986; Lin et al., 2001; Flap and Völker, 2004), some demographic studies have demonstrated how social capital (or the lack thereof), next to economic and cultural resources, shapes fertility decision-making (e.g., Schoen et al., 1997; Astone et al., 1999; Bühler and Philipov, 2005). This research looks at long-term, stable and trusting relationships (e.g., family members, grandparents, close friends, or relatives) that can provide the individual or the couple with fertility-relevant supportive resources.

The majority of existing studies on social capital as a determinant of (low) fertility has focused on ex-communist Eastern European countries. This region has undergone a period of extreme socioeconomic uncertainty after the end of the Soviet Union. In such a context, supportive network relationships and resources play a crucial role in an individual's fertility decision making (Bühler and Philipov, 2005). Consistent with this argument, several empirical studies demonstrate that the greater the social capital, the higher the probability to have a child (Philipov et al., 2006; Bühler and Fratczak, 2007) and to have the child sooner (Bühler and Philipov, 2005; Philipov et al., 2006).

The US literature focuses on the support that kin provide to mothers, including childcare and help with raising children, especially for single mothers – here kin are seen as providing 'safety nets.' This is particularly significant given the high rates of partnership instability and single parenthood and is a strong focus of the 'Fragile Families' study (Harknett et al., 2001). Harknett and Knab (2007) find that multipartnered fertility, through the difficulties of maintaining kin networks, lowers the financial, housing, and childcare support to mothers.

Macro-Level Determinants of Fertility

A vast literature focuses on how the cultural and institutional setting in which individuals and couples are embedded affects their fertility decision making. Here an 'economy versus culture' dichotomy can be identified (Billari, 2004). While several

studies investigate how economic trends, social policies, institutional constraints, and welfare regimes influence fertility tempo and quantum, other contributions focus on the influence of values, attitudes, and culture in reproductive behavior. Moreover, next to these two broad approaches, other macro-level studies look at the role of contraceptive technologies in fertility dynamics. Once again, the main challenge of this literature rests on how to deal with endogeneity and causality issues. Difficulties in disentangling the impact of policies from other observable or unobservable factors have often frustrated the effort to uncover policy effects on fertility. It is difficult to separate the impact of any specific policy from the broad range of policy instruments that potentially influence fertility, and it is problematic to empirically establish whether a specific policy was successful due to the temporal lag between policy initiation and take-up. Moreover policies might be endogenous, in that they may not only impact fertility and induce change, but are often a reaction to changes in fertility and an integral feature of these changes.

Economic Trends

Period effects of economic trends on fertility are usually investigated by linking the total fertility rate (TFR, which represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with current age-specific fertility rates) to macroeconomic indicators (e.g., GDP and unemployment rate). The relationship between TFR and GDP is ambiguous. Several studies (e.g., Martin, 2004) find a pro-cyclical relationship between economic growth and fertility in the developed world. Fertility decline during economic recessions are seen as a result of childbearing postponement, especially of first births, which can later be largely compensated during times of economic prosperity. Similar arguments can be found in Kohler et al. (2002a), Mills and Blossfeld (2005) and Sobotka et al. (2010), who maintain that an economic downturn brings about uncertainties that in turn, lead to postponement. Other studies, however, find contrasting results. Butz and Ward (1979) suggest that economic upswings bring about the increased employment of women, making children more expensive during times of economic prosperity. Therefore, fertility trends are likely to be countercyclical (Macunovich, 1996). The negative relationship between economic growth and postponement also seems to be contradicted by some recent studies (Billingsley, 2010; Aassve et al., 2006).

A related approach that has recently received great attention is the one that explains fertility patterns and cross-country fertility differences in terms of socioeconomic development. Myrskylä et al. (2009) provide evidence for a fundamental change in the well-established negative relationship between fertility and development. They find that while low and medium levels of the Human Development Index (HDI) are associated with persistent low fertility, higher HDI levels seem to promote fertility, reversing the declining pattern.

(Un)employment Trends

Economic uncertainty has also been studied by examining the effects of unemployment trends on the TFR. Findings consistently showed a negative association: the higher the unemployment, the lower the quantum of fertility (Macunovich, 1996) or the higher the postponement, which was found for first and second births (Adserà, 2010, 2011).

Adopting a complementary approach, other studies focus on the relationship between female labor force participation (LFP) and TFR, showing that in OECD countries, this association has changed from negative (where countries with higher LFP had lower TFR) to positive during the 1980s (Pampel, 2001; Ahn and Mira, 2002). It is, however, challenging to assess whether this implies a change in the causal relationship between the two variables. Mishra et al. (2010), engaging in a macro-econometric analysis aimed at ruling out endogeneity in order to unravel causation, find that causality runs from changes in fertility (TFR) to changes in LFP.

Another approach on the effects of economic trends has been developed by Easterlin (1961, 1968). According to Easterlin, cyclical changes in fertility are mainly due to fluctuations in birth rates and cohort size. Members of larger cohorts face more competition and thereby reduced economic opportunities, leading to lower fertility.

Policy Measures

A second stream of research has studied the impact of policy measures (e.g., labour-market, fiscal, family, or housing policies) on the timing of parenthood, as well as on fertility quantum. There is mixed evidence regarding the effectiveness of social policies on fertility (Gauthier, 2007; Hoem, 2008). Gauthier (2007) argues that their effects, although small, seem to affect the timing of fertility rather than the number of children.

A large number of studies investigate the effects of childcare provision on fertility. Most of the empirical research shows mixed findings. While some studies find that regions with poor childcare coverage have higher fertility (e.g., Rosen, 2004), others find that public availability of childcare has a positive effect on fertility (Rindfuss et al., 2010). Social security systems, and the reform in social security, have been discussed as determinants of fertility (Cigno and Rosati, 1992). The relevance of social security for fertility choices is also linked to the idea that children may provide security in old age, also in advanced societies (Rendall and Bahchieva, 1998).

Similar to economic factors, with which they are closely interrelated (Hoem and Hoem, 1997), the effect of family policies varies according to the institutional context and individual-level determinants.

Welfare Regimes

Building on the work of Esping-Andersen (1990, 1999), numerous scholars have explained cross-country differences in fertility and life course patterns by linking them to different institutional constellations. As described in Mills and Blossfeld (2005), who link different welfare regimes to fertility postponement, differences between welfare regimes manifest themselves in the priority of: (1) active employment-sustaining labor market policies (i.e., the commitment to full employment); (2) welfare-sustaining employment exit policies (i.e., support for those who are outside of the labor market such as

youth, unemployed, ill, poor, family care workers, and pensioners); (3) the scope and generosity of family allowances and services (i.e., maternity/paternity leave and childcare); and (4) the share of the public sector in the labor force. This constellation of policies in turn impacts the safety net that individuals can draw upon. Employment regulations and family-related services (e.g., childcare and leave) enable or constrain decision making about entry into parenthood or having additional children.

Defamilialized regimes, where the households' welfare and caring responsibilities are largely supported by the welfare state (and not the family), such as Nordic socio-democratic countries, or market provision (Anglo-Saxon liberal market regimes), enable higher fertility. Conversely, familistic states (Conservative and especially Southern European regimes), where the majority of the economic and caring responsibilities rest on the family, where institutions also support a traditional division of the domestic labor (i.e., the so-called male-breadwinner model), constrain fertility, resulting in lower fertility levels.

Value and Attitude Changes

This stream of research largely stems from the 'Second Demographic Transition' (SDT), developed by Lesthaeghe and van de Kaa (1986) to interpret demographic changes in industrialized societies. According to this theory, ideational changes, that mainly consist of the rejection of institutional control, accentuation of individual autonomy, and the rise of self-realization needs are the driving forces of new family arrangements and behaviors, among which are fertility postponement, reduced number of children, or childlessness, that have developed since the 1960s. As highlighted by Van de Kaa (1997), such ideational changes may occur in different periods and at a different intensity across diverse areas. The SDT framework has been used as an alternative (Lesthaeghe, 1983) or complementary explanation (Billingsley, 2010), next to economic cycle effects, to fertility quantum variations.

Other studies focus on the impact of changing social norms on fertility. Several researchers have documented the relevance of age deadlines for childbearing (i.e., ages after which it is not socially acceptable to become a parent) (Liefbroer and Billari, 2010). Billari et al. (2011) illustrate that age deadlines are positively associated with the prevalence of ART in a given country.

Historical and Cultural Continuities

Some studies identify historical and cultural continuities – or path dependency – as the roots of present fertility behaviors, reaching similar conclusions to that of Esping-Andersen (Reher, 1998; Dalla Zuanna, 2001). However, this literature, by strongly emphasizing the importance of cultural background, assumes that culture has shaped institutional settings (Pfau-Effinger, 1999).

This body of research can be distinguished along East–West and North–South divides. The East–West divide in Europe, running along an imaginary line connecting Trieste and St. Petersburg, was first noticed by Hajnal (1965). On the West side of the 'Hajnal line,' areas were characterized by late and not universal marriage, whereas to the East of the line, marriage was early and widespread. Historical continuities are then assumed to explain why birth happen earlier to the East of the Hajnal line. The North–South divide, first elaborated by Reher (1998), considers the strength of intergenerational family ties: while Southern countries are characterized by strong family ties, Northern areas generally have weak family ties. The main argument is that systems characterized by strong kinship and intergenerational relationship (e.g., Southern European familistic countries) are those where couples have lower fertility (Dalla Zuanna, 2001) and young people delay the transition to adulthood, in turn implying a postponement of childbearing (Billari, 2004). It is essential to note that familistic regimes, both from an institutional as well as cultural point of view, are not 'per se' detrimental to fertility. It is rather the interaction of these systems with the recent increased female status in the educational and labor market system and lack of institutional support to combine work and family that is the root cause of low fertility (Mills et al., 2008).

Studies on immigrants, linking fertility in the place of origin to the behavior of individuals in a 'destination' country have also shown the relevance of cultural continuities (Fernández and Fogli, 2006, 2009), with continuity in behavior. Nevertheless, migrants often show behavior that converges to their place of destination, demonstrating that adaptation prevails on selectivity (Kulu, 2005).

Contraceptive and Reproductive Technologies

Fertility differentials at the macro-level are not only explained by 'economy' and 'culture.' Researchers have widely studied the crucial role of the 'contraceptive revolution' on fertility quantum (e.g., Frejka, 2008) and fertility postponement (Sobotka, 2004). The spread of modern contraception, and especially the pill, has radically changed the nature of the fertility decision making and contributed to the reduction in the number of children and the postponement of childbearing (Goldin, 2006). On the other hand, the development of assisted reproductive technology (ART) seems to challenge the biological limits to postponement (Billari et al., 2007; Leridon, 2008).

Acknowledgment

With kind permission from Springer Science+Business Media B.V., this article is based on: Balbo, N., Billari, F.C., Mills, M., 2013. Fertility in advanced societies: a review. European Journal of Population 29 (1), 1–38.

See also: Abortion, Demography of: United States; Children, Value of; Cohabitation: United States; Family Size Preferences; Family Theory: Competing Perspectives in Social Demography; Family Theory: Economics of Childbearing; Family Theory: Economics of Intergenerational Relations; Family Theory: Economics of Marriage and Divorce; Fertility Change: Quantum and Tempo; Fertility Theory: Embodied-Capital Theory of Life History Evolution; Fertility Theory: Theory of Intergenerational Wealth Flows; Fertility Transition: China; Fertility Transition: Cultural Explanations; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Fertility Transition: Latin America and the Caribbean; Fertility Transition: Middle East and North Africa; Fertility Transition: South Asia; Fertility Transition:

Southeast Asia; Fertility Trends in the Formerly Socialist Countries of Europe; Fertility and Culture: Anthropological Insights; Fertility and Social Networks; Fertility of Single and Cohabiting Women; Infertility: Demographic Aspects; Repartnering and Stepchildren; Second Demographic Transition.

Bibliography

Aassve, A., Billari, F.C., Speder, Z., 2006. Societal transition, policy changes and family formation: evidence from Hungary. European Journal of Population 22 (2), 127–152.

Adserà, A., 2010. Where are the babies? Labor market conditions and fertility in Europe. European Journal of Population 27 (1), 1–32.

Adserà, A., 2011. The interplay of economic uncertainty and education in explaining second births in Europe. Demographic Research 25 (16), 513–544.

Ahn, N., Mira, P., 2002. A note on the changing relationship between fertility and female employment rates in developed countries. Journal of Population Economics 15, 667–682.

Ajzen, I., 1991. The theory of planned behavior. Organizational Behavior and Human Decision Processes 50 (2), 179–211.

Anderton, D., Tsuya, N., Bean, L., Mineau, G., 1987. Intergenerational transmission of relative fertility and life course patterns. Demography 24 (4), 467–480. http://dx.doi.org/10.2307/2061386.

Andersson, G., Hank, K., Rønsen, M., Vikat, A., 2006. Gendering family composition: sex preferences for children and childbearing behavior in the Nordic countries. Demography 43 (2), 255–267.

Astone, N.M., Nathanson, C.A., Schoen, R., Kim, Y.J., 1999. Family demography, social theory, and investment in social capital. Population and Development Review 25 (1), 1–31.

Axinn, W.G., Clarkberg, M.E., Thornton, A., 1994. Family influences on family size preferences. Demography 31, 65–79.

Balbo, N., 2012. Family, Friends and Fertility Thesis. University of Groningen.

Balbo, N., Billari, F.C., Mills, M., 2013. Fertility in advanced societies: a review. European Journal of Population 29 (1), 1–38.

Balbo, N., Mills, M., 2011. Social capital and pressure in fertility decision-making: second and third births in France, Germany and Bulgaria. Population Studies 65 (3), 335–351.

Becker, G.S., 1960. An economic analysis of fertility. In: Becker, G.S. (Ed.), Demographic and Economic Change in Developed Countries. Princeton University Press, Princeton, NJ, pp. 209–231.

Becker, G.S., Lewis, H.G., 1973. On the interaction between the quantity and quality of children. Journal of Political Economy 81 (2), S279–S288.

Begall, K., Mills, M.C., 2012. The influence of educational field, occupation, and occupational sex segregation on fertility in the Netherlands. European Sociological Review. http://dx.doi.org/10.1093/esr/jcs051.

Bernardi, L., Keim, S., von der Lippe, H., 2007. Social influences on fertility: a comparative mixed methods study in eastern and western Germany. Journal of Mixed Methods Research 1, 23–47.

Berrington, A., 2004. Perpetual postponers? Women's, men's and couple's fertility intentions and subsequent fertility behaviour. Population Trends 117, 9–19.

Billari, F.C., 2001. A log-logistic regression model for a transition rate with a starting threshold. Population Studies 55 (1), 15–24.

Billari, F.C., 2004. Becoming an adult in europe: a macro(/micro)-demographic perspective. Demographic Research 3 (2), 13–44.

Billari, F.C., Kohler, H.P., 2004. Patterns of low and lowest-low fertility in Europe. Population Studies 58 (2), 161–176.

Billari, F.C., Goisis, A., Liefbroer, A.C., Settersten, R.A., Aassve, A., Hagestad, G., et al., 2011. Social age deadlines for the childbearing of women and men. Human Reproduction 26 (3), 616–622.

Billari, F.C., Kohler, H.-P., Andersson, G., Lundström, H., 2007. Approaching the limit: long-term trends in late and very late fertility. Population and Development Review 33, 149–170.

Billingsley, S., 2010. The post-communist fertility puzzle. Population Research and Policy Review 29 (2), 193–231.

Blossfeld, H.-P., Huinink, J., 1991. Human capital investments or norms of role transition? How women's schooling and career affect the process of family formation. American Journal of Sociology 97, 143–168.

Bongaarts, J., 2001. Fertility and reproductive preferences in post-transitional societies. Population and Development Review 27, 260–281.

Bongaarts, J., 2002. The end of the fertility transition in the developed world. Population and Development Review 28 (3), 419–443.

Bongaarts, J., Watkins, S., 1996. Social interactions and contemporary fertility transitions. Population and Development Review 22 (4), 639–682.

Bourdieu, P., 1986. The forms of capital. In: Richardson, J.G. (Ed.), Handbook of Theory and Research for the Sociology of Education. Greenwood, New York, pp. 241–258.

Buber, I., Prskawetz, A., 2000. Fertility in second unions in Austria: findings from the Austrian FFS. Demographic Research 3 (2).

Bühler, C., Fratczak, E., 2007. Learning from others and receiving support: the impact of personal networks on fertility intentions in Poland. European Societies 9 (3), 359–382.

Bühler, C., Philipov, D., 2005. Social capital related to fertility: theoretical foundations and empirical evidence from Bulgaria. Vienna Yearbook of Population Research 2005, 53–81.

Bumpass, L., Sweet, J., Cherlin, A., 1991. The role of cohabitation in declining rates of marriage. Demography 53, 913–927.

Butz, W.P., Ward, M.P., 1979. The emergence of countercyclical U.S. Fertility. The American Economic Review 69 (3), 318–328.

Caldwell, J.C., Schindlmayr, T., 2003. Explanations of the fertility crisis in modern societies: a search for commonalities. Population Studies 57 (3), 241–263.

Caltabiano, M., 2008. Has the fertility decline come to an end in the different regions of Italy? New insights from a cohort approach. Population 63 (1), 157–172.

Chesnais, J.-C., 1996. Fertility, family, and social policy. Population and Development Review 22 (4), 729–739.

Cigno, A., Ermisch, J., 1989. A microeconomic analysis of the timing of first births. European Economic Review 33, 737–760.

Cigno, A., Rosati, F.C., 1992. The effects of financial markets and social security on saving and fertility behaviour in Italy. Journal of Population Economics 5 (4), 319–341.

Coleman, J.S., 1988. Social capital in the creation of human capital. American Journal of Sociology 94 (Suppl.), S95–S120.

Corijn, M., Klijzing, E., 2001. Transitions to Adulthood in Europe. Kluwer, Dordrecht.

Corijn, M., Liefbroer, A.C., De Jong Gierveld, J., 1996. It takes two to tango, doesn't it? The influence of couple characteristics on the timing of the birth of the first child. Journal of Marriage and the Family 58, 117–126.

Dalla Zuanna, G., 2001. The banquet of Aeolus: a familistic interpretation of Italy's lowest low fertility. Demographic Research 4 (5), 133–162.

Easterlin, R.A., 1961. The American baby boom in historical perspective. American Economic Review 51, 869–911.

Easterlin, R.A., 1968. Population, Labor Force, and Long Swings in Economic Growth: The American Experience. Columbia University Press, New York.

Easterlin, R.A., 1969. Towards a socioeconomic theory of fertility: a survey of recent research on economic factors in American fertility. In: Behrman, S.J., Corsa Jr., L., Freedman, R. (Eds.), Fertility and Family Planning: A World View. University of Michigan Press, Ann Arbor, pp. 127–156.

Easterlin, R.A., 1976. The conflict between aspirations and resources. Population and Development Review 2 (3), 417–425.

Esping-Andersen, G., 1990. The Three Worlds of Welfare Capitalism. Polity Press, Oxford.

Esping-Andersen, G., 1999. Social Foundations of Postindustrial Economies. Oxford University Press, Oxford.

Esping-Andersen, G., 2009. The Incomplete Revolution: Adapting Welfare States to Women's New Roles. Polity Press, Cambridge, UK.

Fernández, R., Fogli, A., 2006. Fertility: the role of culture and family experience. Journal of the European Economic Association 4 (2–3), 552–561.

Fernández, R., Fogli, A., 2009. Culture: an empirical investigation of beliefs, work, and fertility. American Economic Journal: Macroeconomics 1 (1), 146–177.

Flap, H., Völker, B., 2004. Creation and Returns of Social Capital: A New Research Program. Routledge, London.

Frejka, T., 2008. Birth regulation in Europe. Completing the contraceptive revolution. Demographic Research 19, 73–84.

Frejka, T., Jones, G.W., Sardon, J.-P., 2010. East asian childbearing patterns and policy developments. Population and Development Review 36, 579–606.

Friedman, D., Hechter, M., Kanazawa, S., 1994. A theory of the value of children. Demography 31, 375–401.

Furstenberg Jr., F.F., Levine, J.A., Brooks-Gunn, J., 1990. The children of teenage mothers: patterns of early childbearing in two generations. Family Planning Perspectives 22, 54–61.

Gauthier, A., 2007. The impact of family policies on fertility in industrialized countries: a review of the literature. Population Research and Policy Review 26, 323–346.

Gauthier, A., Hatzius, J., 1997. Family benefits and fertility: an econometric analysis. Population Studies 51, 295–306.

Gibson-Davis, C.M., Edin, K., McLanahan, S., 2005. High hopes but even higher expectations: the retreat from marriage among low-income couples. Journal of Marriage and the Family 67, 1301–1312.
Goldin, C., 2006. The quiet revolution that transformed women's employment, education, and family. American Economic Review 96 (2), 1–21.
Goldstein, J.R., Sobotka, T., Jasilioniene, A., 2009. The end of lowest-low fertility? Population and Development Review 35 (4), 663–700.
Goldstein, J., Lutz, W., Testa, M.R., 2003. The emergence of sub-replacement family size ideals in europe. Population Research and Policy Review 22 (5), 479–496.
Granovetter, M., 1973. The strength of weak ties. American Journal of Sociology 78 (6), 1360–1380.
Hajnal, J., 1965. European marriage patterns in perspective. In: Glass, D.S., Eversley, D.E.C. (Eds.), Population in History. Edward Arnold, London, pp. 101–143.
Hakim, C., 2003. A new approach to explaining fertility patterns: preference theory. Population and Development Review 29 (3), 349–374.
Hank, K., Kohler, H.-P., 2000. Gender preferences for children in Europe: empirical results from 17 FFS countries. Demographic Research 2 (1), 133–144.
Happel, S.K., Hill, J.K., Low, S.A., 1984. An economic analysis of the timing of childbirth. Population Studies 38, 299–311.
Harknett, K., Knab, J., 2007. More kin, less support: multipartnered fertility and perceived support among mothers. Journal of Marriage and Family 69 (1), 237–253.
Harknett, K., Hardman, L., Garfinkel, I., McLanahan, S.S., 2001. The fragile families study: social policies and labor markets in seven cities. Children and Youth Services Review 23 (6–7), 537–555.
Hobcraft, J., Kiernan, K., 1995. Becoming a parent in Europe. In: EAPS-IUSSP (Ed.), European Population Conference, vol. 1. Franco Angeli, Milan, pp. 27–65.
Hoem, B., Hoem, J.M., 1997. Sweden's family policies and roller-coaster fertility. Jinko Mondai Kenkyu (Journal of Population Problems) 52, 1–22.
Hoem, J.M., 1986. The impact of education on modern family-union initiation. European Journal of Population 2, 113–133.
Hoem, J.,M., 2008. Overview chapter 8: the impact of public policies on European fertility. Demographic Research Special Collection 7 (19), 249–260.
Jansen, M.D., Liefbroer, A.C., 2006. Couples' attitudes, childbirth and the division of labor. Journal of Family Issues 27, 1487–1511.
Kertzer, D.I., White, M.J., Bernardi, L., Gabrielli, G., 2009. Italy's path to very low fertility: the adequacy of economic and second demographic transition theories. European Journal of Population 25 (1), 89–115.
Kohler, H.P., Billari, F.C., Ortega, J.A., 2002a. The emergence of lowest-low fertility in Europe during the 1990s. Population and Development Review 28, 641–681.
Kohler, H.P., Rodgers, J.L., 2003. Education, fertility, and heritability: explaining a paradox. In: Wachter, K.W., Bulatao, R.A. (Eds.), Offspring: Fertility Behavior in Biodemographic Perspective. National Academies Press, Washington, DC, pp. 46–90.
Kohler, H.-P., Behrman, J.R., Watkins, S.C., 2001. The density of social networks and fertility decisions: evidence from South Nyanza district, Kenya. Demography 38 (1), 43–58.
Kohler, H.-P., Rodgers, J.L., Christensen, K., 2002b. Between nurture and nature: the shifting determinants of female fertility in Danish twin cohorts 1870–1968. Social Biology 49, 218–248.
Kohler, H.P., Rodgers, J.L., Christensen, K., 1999. Is fertility behavior in our genes? Findings from a Danish twin study. Population and Development Review 25, 253–288.
Kravdal, Ø., 1992. The emergence of a positive relation between education and third birth rates in Norway with supportive evidence from the United States. Population Studies 46 (3), 459–475.
Kravdal, Ø., Rindfuss, R.R., 2008. Changing relationships between education and fertility: a study of women and men born 1940 to 1964. American Sociological Review 73, 854–873.
Kreyenfeld, M., 2010. Uncertainties in female employment careers and the postponement of parenthood in Germany. European Sociological Review 26 (3), 351–366.
Kulu, H., 2005. Migration and fertility: competing hypotheses re-examined. European Journal of Population 21 (1), 51–87.
Kulu, H., 2011. Why do fertility levels vary between urban and rural areas? Regional Studies.
Kulu, H., Boyle, P.J., 2009. High fertility in city suburbs: compositional or contextual effects? European Journal of Population 25 (2), 157–174.
Kulu, H., Vikat, A., 2007. Fertility differences by housing type: the effect of housing conditions or of selective moves? Demographic Research 17 (26), 775–802.
Kulu, H., Boyle, P., Andersson, G., 2009. High suburban fertility: evidence from four northern european countries. Demographic Research 21 (31), 915–944.
Leridon, H., 2008. A new estimate of permanent sterility by age: sterility defined as the inability to conceive. Population Studies 62 (1), 15–24.
Lesthaeghe, R., van de Kaa, D., 1986. Twee demografische transities? (Two demographic transitions?). In: Lesthaeghe, van de Kaa (Eds.), Bevolking – Groei en Krimp, Mens en Maatschappij. Van Loghum Slaterus, Deventer, pp. 9–24.
Lesthaeghe, R., 1983. A century of demographic and cultural change in Western Europe: an exploration of underlying dimensions. Population and Development Review 9 (3), 411–435.
Liefbroer, A.C., 2009. Changes in family size intentions across young adulthood: a life-course perspective. European Journal of Population 25 (4), 363–386.
Liefbroer, A.C., Billari, F.C., 2010. Bringing norms back in: a theoretical and empirical discussion of their importance for understanding demographic behaviour. Population, Space and Place 16 (4), 287–305.
Lin, N., Cook, K.S., Burt, R.S., 2001. Social Capital: Theory and Research. Aldine Transaction, New Brunswick.
Lutz, W., O'Neill, B.C., Scherbov, S., 2003. Europe's population at a turning point. Science 299, 1991–1992.
Lyngstad, T.H., Prskawetz, A., 2010. Do siblings' fertility decisions influence each other? Demography 47 (4), 923–934.
Macunovich, D.J., 1996. Relative income and price of time: exploring their effects on US fertility and female labor force participation. Population and Development Review 22 (Suppl.), 223–257.
Manlove, J., 1997. Early motherhood in an intergenerational perspective: the experiences of a British cohort. Journal of Marriage and the Family 59, 263–297.
Manski, C.F., Mayshar, J., 2003. Private incentives and social interactions: fertility puzzles in Israel. Journal of the European Economic Association 1 (1), 181–211.
Martin, J., 2004. The ultimate vote of confidence': fertility rates and economic conditions in Australia, 1976–2000. Australian Social Policy 31–54, 2002–2003. Canberra: Commonwealth of Australia.
McDonald, P., 2000a. Gender equity in theories of fertility transition. Population and Development Review 26 (3), 427–439.
McDonald, P., 2000b. Gender equality, social institutions and the future of fertility. Journal of Population Research 17, 1–16.
Mills, M., 2004. Stability and change: the structuration of partnership histories in Canada, the Netherlands and the Russian federation. European Journal of Population 20, 141–175.
Mills, M., Begall, K., 2010. The impact of gender preferences on third births in europe: a multilevel examination of men and women. Population Studies 64 (1), 77–95.
Mills, M., Blossfeld, H.-P., 2005. Globalization, uncertainty and the early life course: a theoretical framework. In: Blossfeld, H.-P., Klijzing, E., Mills, M., Kurz, K. (Eds.), Globalization, Uncertainty and Youth in Society. Routledge, London, New York, pp. 1–24. Advances in Sociology Series.
Mills, M., Mencarini, L., Tanturri, M.L., Begall, K., 2008. Gender equity and fertility intentions in Italy and the Netherlands. Demographic Research 18 (1), 1–26.
Mills, M., Rindfuss, R.R., McDonald, P., te Velde, E., 2011. Why do people postpone parenthood? Reasons and social policy incentives. Human Reproduction Update 17 (6), 848–860.
Mishra, V., Nielsen, I., Smyth, R., 2010. On the relationship between female labour force participation and fertility in G7 countries: evidence from panel cointegration and Granger causality. Empirical Economics 38 (2), 361–372.
Montgomery, M.R., Casterline, J.B., 1996. Social influence, social learning, and new models of fertility. In: Casterline, J., Lee, Ronald, Foote, Karen (Eds.), Fertility in the United States: New Patterns, New Theories, pp. 87–99.
Morgan, S.P., Taylor, M.G., 2006. Low fertility at the turn of the twenty-first century. Annual Review of Sociology 32, 375–399.
Murphy, M., Wang, D., 2001. Family-level continuities in childbearing in low-fertility societies. European Journal of Population 17, 75–96.
Myers, S.M., 1997. Marital uncertainty and childbearing. Social Forces 75, 1271–1289.
Myrskylä, M., Kohler, H.-P., Billari, F.C., 2009. Advances in development reverse fertility declines. Nature 460 (7256), 741–743.
Nauck, B., 2007. Value of children and the framing of fertility: results from a cross-cultural comparative survey in 10 societies. European Sociological Review 23 (5), 615–629.
OECD, 2011. Doing Better for Families. OECD Publishing, Paris.
Oppenheimer, V.K., 1988. A theory of marriage timing. American Journal of Sociology 94, 563–591.
Oppenheimer, V.K., 1994. Women's rising employment and the future of the family in industrial societies. Population and Development Review 20, 293–342.
Oppenheimer, V.K., 2003. Cohabiting and marriage during young Men's career-development process. Demography 40, 127–149.
Pampel, F.C., 2001. The Institutional Context of Population Change. The University of Chicago Press, Chicago.
Pfau-Effinger, B., 1999. Change of family policies in the socio-cultural context of European Societies. In: Leira, A. (Ed.), Family Policies. Yearbook Comparative Social Research. JAI press, Stamford, pp. 135–159.

Philipov, D., 2009. Fertility intentions and outcomes: the role of policies to close the gap. European Journal of Population 25, 355–361.
Philipov, D., Speder, Z., Billari, F.C., 2006. Soon, later, or ever? The impact of anomie and social capital on fertility intentions in Bulgaria (2002) and Hungary (2001). Population Studies 60 (3), 289–308.
Prskawetz, A., Vikat, A., Philipov, D., Engelhardt, H., 2003. Pathways to stepfamily formation in Europe: results from the FFS. Demographic Research 8 (5), 107–149.
Quesnel-Vallée, A., Morgan, S.P., 2003. Missing the target? Correspondence of fertility intentions and behavior in the U.S. Population Research and Policy Review 22, 497–525.
Reher, D.S., 1998. Family ties in western Europe: persistent contrasts. Population and Development Review 24 (2), 203–234.
Rendall, M.S., Bahchieva, R.A., 1998. An old-age security motive for fertility in the United States? Population and Development Review 24 (2), 293–307.
Rijken, A.J., Liefbroer, A.C., 2009. The effects of relationship quality on fertility. European Journal of Population 25, 27–44.
Rindfuss, R.R., Guilkey, D.K., Morgan, S.P., Kravdal, O., 2010. Child-care availability and fertility in Norway. Population and Development Review 36 (4), 725–748.
Rosen, M., 2004. Fertility and public policies – evidence from Norway and Finland. Demographic Research 10 (6), 143–170.
Schoen, R., Astone, N.M., Kim, Y.J., Nathanson, C.A., Fields, J.M., 1999. Do fertility intentions affect fertility behavior? Journal of Marriage and the Family 61 (3), 790–799.
Schoen, R., Kim, Y., Nathanson, C., Fields, J., Astone, N.M., 1997. Why do Americans want children? Population and Development Review 23 (2), 333–358.
Skirbekk, V., 2008. Fertility trends by social status. Demographic Research 18 (5), 145–180.
Sobotka, T., Billari, F.C., Kohler, H.-P., 2010. The Return of Late Childbearing in Developed Countries: Causes, Trends and Implications. Vienna Institute of Demography, Vienna.
Sobotka, T., 2004. Postponement of Childbearing and Low Fertility in Europe. Dutch University Press, Amsterdam.
Sobotka, T., Skirbekk, V., Philipov, D., 2011. Economic recession and fertility in the developed world. Population and Development Review 37 (2), 267–306.
Spéder, Z., Kapitány, B., 2009. How are time-dependent childbearing intentions realized? Realization, postponement, abandonment, bringing forward. European Journal of Population 25, 503–523.
Stewart, S.D., 2002. The effect of stepchildren on childbearing intentions and births. Demography 39 (1), 181–197.
Testa, M.R., 2006. Childbearing preferences and family issues in Europe. Special Eurobarometer 253/Wave 65.1-TNS Opinion & Social.
Thomson, E., Hoem, J.M., 1998. Couple childbearing plans and births in Sweden. Demography 35, 315–322.
Thomson, E., McDonald, E., Bumpass, L.L., 1990. Fertility desires and fertility: hers, his, and theirs. Demography 27, 579–588.
Thornton, A., 1978. Marital dissolution, remarriage, and childbearing. Demography 15, 361–380.
Thornton, A., 1980. The influence of first generation fertility and economic status on second generation fertility. Population and Environment 3, 51–72.
Van Bavel, J., 2010. Choice of study discipline and the postponement of motherhood in Europe: the impact of expected earnings, gender composition and family attitudes. Demography 47, 439–458.
Van de Kaa, D.J., 1997. Options and sequences: Europe's demographic patterns. Journal of the Australian Population Association 14, 1–29.
Vikat, A., Hoem, J.M., Thomson, E., 1999. Stepfamily fertility in contemporary Sweden: the impact of childbearing before the current union. Population Studies 53, 211–225.
Wachter, K.W., Bulatao, R.A., 2003. Offspring: Human Fertility Behavior in Biodemographic Perspective. National Research Council, Committee on Population, Division on Behavioral and Social Sciences and Education. The National Academies Press, Washington, DC.
Waite, L.J., Lillard, L.A., 1991. Children and marital disruption. American Journal of Sociology 96 (4), 930–953.
Westoff, C., Ryder, N., 1977. The predictive validity of reproductive intentions. Demography 4, 431–453.
Willis, R.J.A., 1973. New approach to the economic theory of fertility behavior. Journal of Political Economy. 81 (2), S14–S64.
Wu, Z., Schimmele, C.M., 2005. Repartnering after first union disruption. Journal of Marriage and the Family 67, 27–36.

Fertility Theory: Embodied-Capital Theory of Life History Evolution

Hillard S Kaplan, University of New Mexico, Albuquerque, NM, USA
John A Bock, California State University, Fullerton, CA, USA
Paul L Hooper, Santa Fe Institute, Santa Fe, NM, USA; and Emory University, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The embodied-capital theory generalizes existing life history theory by treating the processes of growth, development, and maintenance as investments in stocks of somatic or embodied capital. The theory provides a means of understanding relationships between economic behavior, social behavior, and demographic traits such as fertility and survival. The approach sheds light on the long-term evolution of human life histories (including slow development, long life span, and menopause) and social structure (including high levels of parental investment and intergenerational transfers). It also provides an explanation of the demographic transition to low fertility that has accompanied modernization. This article reviews data from traditional small-scale societies and modern nations bearing on the theory, and addresses the changing roles of physiological, psychological, and cultural mechanisms that underlie fertility behavior in traditional and modern contexts.

The goal of this article is to provide an introduction to the embodied-capital theory of life history evolution and the empirical research upon which it is based. The embodied-capital theory generalizes existing life history theory in evolutionary biology by treating the processes of growth, development, and maintenance as investments in stocks of somatic or embodied capital (see Kaplan, 1996, 1997 and Kaplan et al., 2000 for more complete treatments). In a physical sense, embodied capital is organized somatic tissue – muscles, digestive organs, brains, etc. In a functional sense, embodied capital includes strength, immune function, coordination, skill, knowledge, and other abilities. Since such stocks tend to depreciate with time in response to metabolic activity (e.g., the production of free radicals and mutations associated with mitosis) and aging, allocations to maintenance can also be seen as investments in embodied capital. The logic of embodied capital theory provides a means for understanding the coevolution of the age schedules of fertility, mortality, and kin investment, and thus a range of phenomenon from the evolution of life span and menopause to modern demographic transition. In this article, the general theory is introduced and then applied to human hunting and gathering societies, which characterized most of our species' evolutionary history. This framework is then used to develop models of modern fertility behavior as the result of an interaction between historically novel conditions and the evolved psychological, cultural, and physiological mechanisms governing human parental investment and fertility.

The Embodied-Capital Theory of Life History Evolution

Hominids lived as hunter-gatherers for the vast majority of their evolutionary history (more than 2 million years). Agriculture originated only 10 000 years ago and has been practiced by the majority of the world's population for just two or three millennia, a relatively brief period of time for selection to act. The shift to a fully mercantile, highly urbanized economy with competitive wage–labor markets and a vast array of consumer goods since the mid-nineteenth century is completely novel. Since natural selection is a historical process, humans should be well adapted to the foraging lifestyle, but there is no reason to expect that they will respond adaptively (in the sense of maximizing reproductive fitness) to today's novel circumstances (see Kaplan et al., 1995a; Goodman et al., 2012 for evidence that modern fertility behavior does not maximize fitness). Therefore, an evolutionary explanation of human fertility and parental investment has three requirements: it must be consistent with our general understanding of evolution by natural selection; it must explain the behavior of humans living under traditional hunting and gathering conditions characteristic of our evolutionary history with adaptive models; and it must be able to predict the pattern of modern responses to novel conditions.

Our approach to satisfying these requirements has been to follow a four-step procedure (Kaplan, 1996; Kaplan and Lancaster, 2000). The first step is to develop a model of natural selection on fertility and parental investment for organisms in general. The second step is to apply that model to the specific selection pressures acting on human hunter-gatherers. The third step is to develop a model of the proximate physiological, psychological, and cultural mechanisms produced by selection and how they would interact to determine fertility and parental investment behavior. The fourth step is to generate predictive models of how those mechanisms would respond to conditions currently prevailing in the world. The theory is both ecological and psychological, integrating socioeconomic influences with a theory of mind to identify the processes by which salient features of the social and physical environment relate to people's reproductive and parenting behavior across diverse social, cultural, economic, and ecological contexts.

The Action of Natural Selection on Fertility and Parental Investment

Natural selection acts on variability in the traits of individual organisms within populations. Traits (and the genes that code

for those traits) whose *average* effects on those who possess them act to maximize their 'inclusive fitness' will tend to increase in frequency relative to other traits. *Inclusive fitness* includes effects both on the reproductive success of the individual bearing the gene and on other individuals, related by common descent, who also bear the gene. For example, selection on genes affecting alarm calls in response to predators depends both on their effects on the reproductive fitness of the caller – who may risk a greater threat of predation – and on relatives bearing those genes – whose lives may be saved by the call.

Fertility is the most direct contributor to an organism's fitness (i.e., the number of descendants it produces). In fact, all other fitness components, such as mortality, only affect fitness through their effects on fertility (for example, mortality rates affect fitness by affecting the probability of living to the next reproductive event). All else constant, any increase in fertility increases an organism's fitness. However, there are two trade-offs affecting natural selection on fertility.

The first is the trade off between present and future reproduction. By growing, an organism can increase its energy-capture rates in the future and thus increase its future fertility. For this reason, organisms exhibiting determinant growth typically have a juvenile phase in which fertility is zero until they reach a size at which some allocation to reproduction increases fitness more than growth. Similarly, among organisms that engage in repeated bouts of reproduction (humans included), some energy during the reproductive phase is diverted away from reproduction and allocated to maintenance so that it can live to reproduce again. Natural selection on age of first reproduction and on the adult reproductive rate is expected to maximize total allocations of energy to reproduction over the life course, and it depends on ecological factors, such as mortality rates and the effects of body size on energy capture and vulnerability to predation (Werner, 1986; Charnov, 1993; Kaplan and Robson, 2009).

The second trade-off is between quantity and quality of offspring, where quality is a function of parental investment in offspring and reflects its ability to survive and reproduce. In general, natural selection on offspring number and investment per offspring is expected to maximize the number of offspring that survive to reproduce themselves (Smith and Fretwell, 1974). Ecological factors, such as food supply, disease, and predation rates, also affect the optimal fertility rates and optimal expenditures on offspring.

Virtually, all complex organisms exhibit flexibility in both the age at first reproduction and fertility rates. Natural selection has resulted in physiological and psychological mechanisms by which individuals adjust fertility onset and fertility rates in relation to changing environmental conditions. For example, plants often exhibit many 1000-fold variations in fertility rates in relation to yearly and seasonal variations in rainfall and other environmental factors. Extensive research on many bird species has shown that this phenotypic plasticity tracks fitness fairly well (see Godfray et al., 1991; Bennett and Owens, 2002 for reviews). Birds under variable conditions adjust clutch sizes in ways that tend to maximize the number of surviving young produced during the life course. The pervasiveness of adaptive phenotypic plasticity in fertility among nonhumans suggests that humans are also likely to possess adaptations to adjust fertility onset and fertility rates in relation to changing environmental conditions.

The Embodied-Capital Theory and Its Application to Fertility and Parental Investment among Traditional Humans

The embodied-capital theory treats the present–future reproduction trade-off in terms of optimal investments in own embodied capital versus reproduction, and the quantity–quality trade-off in terms of investments in the embodied capital of offspring versus number produced. Models based on the theory allow us to address problems that have not been addressed with standard life history models. First, an exclusive focus on physical growth per se is an impoverished way of understanding human development. The large human brain, a stock of embodied capital, supports a great deal of learning and knowledge acquisition during both the juvenile and adult periods. Second, parental investment among humans affects not only survival to adulthood, but also the adult socioeconomic status of offspring. This is true not only in modern societies, but traditional hunter-gatherers also exhibit high levels of parental investment in offspring (Kaplan et al., 2000).

These models of investment in embodied capital have produced some fundamental results. At their most expansive, the models establish a general theory for the evolution of senescence and aging, through consideration of the costs and benefits of the developing and maintaining the quantity and quality of embodied capital over the life span (Kaplan and Robson, 2009). Of central interest here, they have also showed that investments in embodied capital affecting adult income or energy capture coevolve with investments affecting mortality and longevity. The longer the time spent growing and learning prior to reproducing, the more natural selection favors investments in staying alive to reap the benefits of those investments. Similarly, any investments that produce increased energy-capture rates later in life select for additional investments to reach those older ages. The converse is also true. Ecological features or investments that increase the probability of survival to older ages also produce selection for greater investments in income-related embodied capital (Kaplan and Robson, 2002). These coevolutionary effects have been particularly important in human life history evolution.

The human adaptation is broad and flexible, in one sense, and very narrow and specialized, in another. It is broad in the sense that as hunter-gatherers, humans have existed successfully in virtually all of the world's major habitats. This has entailed eating a very wide variety of foods, both plant and animal, both within and among environments. It also has entailed a great deal of flexibility in the contributions of different age and sex classes of individuals. The relative contributions of men and women to food production vary from group to group, as do production profiles by age.

The human adaptation is narrow and specialized in that it is based on extremely high investments in brain tissue and learning. In every environment, human foragers consume the largest, highest quality, and most difficult-to-acquire foods, using techniques that take years to learn (Kaplan et al., 2000).

This adaptation has profound effects on fertility, mortality, and wealth flows. Compared with other primates and mammals, there are at least four distinctive characteristics of human life histories: (1) an exceptionally long life span; (2) an extended period of juvenile dependence, resulting in families with multiple dependent children of different ages; (3) support of reproduction by older postreproductive individuals; and (4) male support of reproduction through the provisioning of females and their offspring.

The coevolution of brains, learning and life histories is evident in **Figure 1**. This figure plots net expected cumulative productivity by age (cumulating the probability of being alive at each age times the net productivity at that age) for human hunter-gatherers and wild-living chimpanzees, averaging over available data from different groups (see Kaplan et al., 2000 for details). The long human training period is evident when the troughs in the human and chimpanzee curves are compared. Humans are not net producers until about age of 18–20 years, but are very productive as adults, whereas chimpanzees begin to produce as much as they consume by about age of 5 years and are much less productive as adults. In fact, peak production rates for both human males and females do not occur until the mid-thirties for the most difficult-to-acquire resources. Humans also have an expected adult life span of about 40 years when they become net producers, whereas chimpanzees have an expected adult life span (beginning at about age of 13 years) of 14 years. The dashed line is a hypothetical cross of human production profiles with chimpanzee survival rates. It shows that the human production profile would not be viable with chimpanzee survival rates, because expected lifetime net production would be negative. The human adaptation depends on about 20 years of parental support and a long life span to recoup the investments in learning.

This body of theory can also account for the cessation of fertility in the form of menopause in terms of four principal considerations. First, oocytes and the other reproductive physiology constitute stocks of somatic capital, which are developed and maintained at a cost before and during the reproductive career. Second, childbearing itself entails direct costs in terms of maternal mortality and maternal depletion. Third, continued reproduction often trades off with economic productivity, which we know to be high at advanced ages in both sexes in traditional human societies. Fourth, humans' emphasis on quality over quantity of oocytes and offspring compared to other primates interact with these costs to determine a typical age schedule of reproductive activity and cessation, which precedes general somatic aging. In effect, selection has favored the programmed shutdown of reproduction as oocyte quality decreases with age in order to realize the gains to late-age production and economic transfers to young. These downward transfers support both the reproduction and embodied capital development of family members. Empirically, there is evidence that people adjust transfers on the basis of generation, need, and genetic kinship (Hooper, 2011), as would be predicted theoretically on the basis of inclusive fitness theory. Where monogamous unions predominate in old age, many men in traditional societies experience behavioral menopause at the same time as their wife, and fulfill considerable economic roles for close kin and family.

'The major demographic transition separating our species from apes was thus a shift to low mortality, an increased life span, a long dependent development period, a large commitment to learning and intelligence, menopause, and an age profile of production shifted toward older ages with downward wealth flows across generations.'

Figure 1 Cumulative expected net caloric production by age: humans and chimpanzees.

Ecology and the Proximate Mechanisms Governing Fertility and Parental Investment

The mortality and production profiles discussed above have direct implications for fertility regulation. Since human parents and grandparents provision children, it is likely that natural selection would have produced mechanisms by which fertility could respond to the number of children parents could afford to raise at any given time. Figure 2 presents a conceptual model of the evolved mechanisms of fertility regulation in humans and its responsiveness to physical, biological, and social conditions (*socioecology*). The figure depicts two classes of socioecological factors, one affecting survival and the other embodied capital, and three domains of evolved responses, psychology and culture, behavior, and reproductive physiology. The pathways in solid black refer to the traditional hunting and gathering lifeway, which we consider first. The pathways in striped gray include additional factors and causal pathways existing under modern conditions.

Traditional Foraging Societies

There are two main causal pathways regulating fertility that are both related to the energy constraints in provisioning offspring, one involving breastfeeding and the other involving the age profiles of work and food distribution. Natural selection on the timing and rate of weaning should reflect a balance between quantity effects (favoring earlier weaning) and quality effects on survival and growth of offspring (favoring later weaning). Socioecological conditions, such as the availability of easy-to-digest weaning foods and disease threats, are likely to impact the relationship between age at weaning and survival. Similarly, infant condition is likely to be individually variable and differentially affected by weaning. We expect that natural

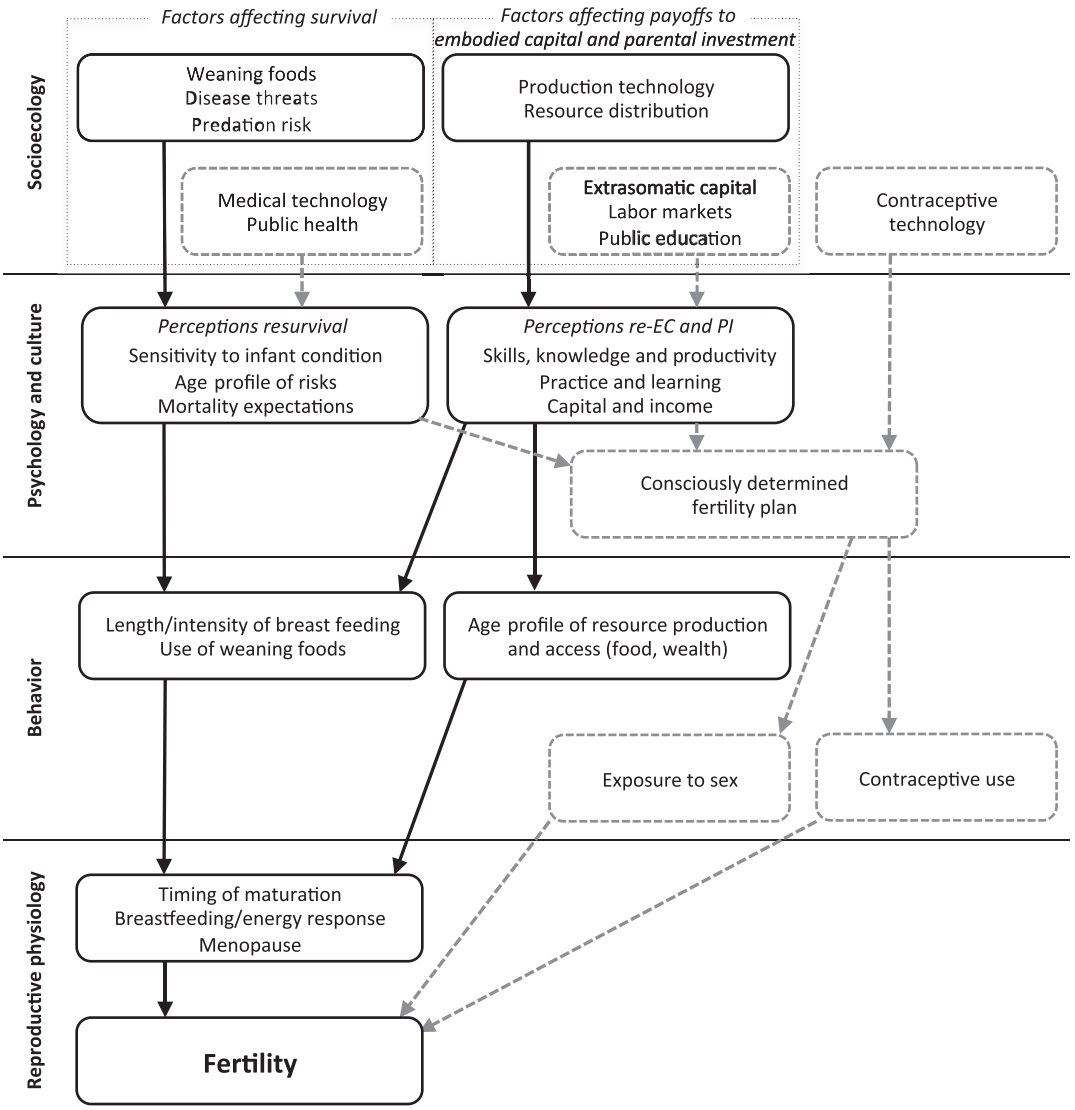

Figure 2 A mode of fertility determinants under traditional and modern conditions. Pathways in solid black are predominant under conditions of traditional natural fertility, while those in striped gray are predominant under modern economic and technological circumstances.

selection has acted on both infant and maternal psychology to make them responsive to those variable conditions. For example, sickly infants may be more demanding of the breast, and larger and healthier babies more interested in weaning foods. In the human case, maternal perceptions regarding breastfeeding are not only guided by her individual experience and sensitivity to her infant's condition, but also by the cumulative experience of her group, as reflected in cultural conventions and ideas about proper breastfeeding practices. Thus, there is likely an interaction between maternal and infant psychological processes in determining the length and timing of breastfeeding. The evolved reproductive physiology in the hypothalamic–pituitary axis translates this pattern of breastfeeding into differing degrees of fecundity.

The second causal pathway is due to the additional energy constraints involved in provisioning children. The age/sex profile of work and productivity determines the total energy available for consumption. Given some age profile of food distribution and of mortality, the fertility rate must be constrained so that energy consumed does not exceed energy produced. The longer children are provisioned and the greater proportion of their food needs subsidized by parents and grandparents, the lower the reproductive rate that can be supported with a given adult income.

Natural selection is therefore expected to act both on psychological abilities to assess the short- and long-term costs and benefits of different activity profiles and on fertility regulation so that energy flows are balanced. One fundamental trade-off with respect to activity profiles is between present and future production (Kaplan, 1996; Bock, 2002). If children engage in easy-to-perform foraging activities, such as fruit collection, this will provide immediate caloric benefits and lower the parental subsidy required. However, time allocation to those activities detracts from time spent learning more difficult activities, such as hunting and extractive foraging. During the learning phase, those activities produce very less food, but generate high rates of production in the future. Another fundamental trade-off is between production and mortality risks, since foraging exposes people to predation, accidents, and getting lost. This suggests that parental psychology should be responsive to age-specific mortality rates, how these rates are affected by alternative activity profiles, and their short- and long-term consequences for production/productivity.

There is a growing body of evidence supporting the view that people in foraging societies are sensitive to these trade-offs and adjust behavior accordingly. In a seminal series of papers, Blurton Jones and coworkers (Blurton Jones et al., 1994a,b; Hawkes et al., 1995) show that !Kung hunter-gatherer parents consciously assess foraging risks to children and that differences between Kung and Hadza children's productivity are due to features of the local ecology related to ease of resource extraction and dangers associated with productive tasks. Bock (1995, 2002), in a study of forager/agropastoralists in Botswana, shows that the age profiles of children's activity budgets reflect immediate skill and strength requirements as well as opportunities to learn. In all foraging societies for which data are available, people are aware of the many years of learning required to become a competent hunter and young men are provisioned while they learn those skills (Kaplan et al., 2000; Walker et al., 2002; Gurven et al., 2006).

These time-allocation decisions affect fertility through maternal energetics. In addition to breastfeeding effects, fecundity responds to women's net energy flows. Seasonal changes on food intake (Bailey et al., 1992), net energy expenditure (Hurtado and Hill, 1987), and weight (Bailey et al., 1992; Tracer, 1996; Kaplan et al., in press) all have been shown to predict fecundity, as do individual differences among women (Hill and Hurtado, 1996). These results are reviewed in Ellison's (2001) classic synthesis.

To summarize, in this model fertility is determined by the interaction of individual psychology and cumulative cultural knowledge, beliefs and practices, behavior, and reproductive physiology. These domains necessarily coevolved by natural selection, producing an ability to respond to ecological variability in mortality regimes and payoffs to parental subsidization of skill acquisition. The connection between psychology and fertility is indirect in that psychology adjusts parental investment (through productive behavior, wealth flows, and breastfeeding) and reproductive physiology translates parental investment decisions into fertility. The key to this system is that maximizing lifetime expected resource production through the optimal allocation of activities and wealth flows will tend also to maximize fitness when all wealth is in the form of food and extra food translates into higher fertility.

Fertility under Modernizing Conditions

In **Figure 2**, the additional socioecological factors and causal pathways proposed to result in fertility transition are shown in striped gray. Our proposal is that this fertility regime results from the interaction of new socioecological conditions with the psychological processes governing parental investment and wealth flows, evolved in response to the hunting and gathering lifeway. The two critical new factors are the existence of extrasomatic capital and changes in the payoffs to embodied capital, especially in the form of education.

Extrasomatic physical capital, such as land and money, is stored and invested in children in the form of inheritance and parental support for education. Unlike the traditional case where all wealth is embodied, there is no guarantee that a wealth-maximizing psychology will maximize fitness, because much wealth may never be converted into offspring and remain stored in other physical forms. In addition, perceptions about the optimal amount of physical capital to transfer to offspring may require lower fertility than the energy-based physiology of women will produce, stimulating a desire to control unwanted fertility through other means.

Changes in the payoffs to embodied capital occur for two reasons. Changes in the technology of production stimulating the development of education-based labor markets have led to very high returns on parental investments in children's education. There are multiple inputs into the production of education-based embodied capital. Parental time inputs prepare children for school and affect their progress in school. Variation in the quality of those inputs, perhaps based on parental education, may predict both levels of parental investment and fertility. Teachers, educational

materials, and the quality of student preparation all affect rates of learning (Card and Krueger, 1992).

Changing medical technology and public health efforts greatly reduced mortality rates for all age groups. Increased survival rates during the period of parental investment increase the expected costs per child born, favoring further increases in offspring quality (Kaplan, 1996). Increased survival during the adult period increases the expected years of return of educational investments, further increasing the incentive to invest in children's education.

It will be especially productive to investigate the public and private dynamics involved in this process. Increased private interest in education may have inspired more public investment in education. The quality of public education, in turn, increased the rate of return on private investments in education. Similarly, increased private interest in health (perhaps inspired, in part, by increased investment in education) also stimulated increased public investment in health. This, in turn, increased payoffs to educational investments.

The increased importance of physical capital and education appear to result in desired fertility levels far below those which would be produced by evolved physiology alone. Throughout most of the world, fertility is now regulated by the use of birth-control technology and/or controlled exposure to sex, as part of what is often a consciously determined childbearing plan. In response to the increased payoffs to investments in education and the expected costs of those investments, parents determine the number of children they can afford to raise, given their wealth. The low mortality rates also allow parents to plan reproduction at the outset, because the number of children born accurately predicts the number of children that will be raised to adulthood. Reproductive physiology affects fertility in the present in that it sets the broad limits within which a childbearing plan can be realized, and may actually lower fertility when plans include significant delays to first reproduction (Kaplan et al., 2001).

A great deal of further research is necessary, however, before we can understand why these changes in socioecology have resulted in such low levels of fertility and high levels of parental investment and wealth consumption. It does not appear that people are simply maximizing family wealth because net wealth of families would be maximized by higher fertility than is currently observed. It is also clear that people are not maximizing personal consumption; in that case, they would have no children. One possible hypothesis is that the social dynamics of small groups in hunting and gathering economies resulted in greater fitness for those of higher social standing and selected for a psychology in which relative social position of self and offspring is valued highly. Such a psychology would also emerge if people judged the well-being of themselves and their children based upon the wealth and consumption of others. This psychology may have been fitness maximizing under traditional conditions. If relative, as opposed to absolute, wealth and social standing guide human decisions regarding wealth flows, parental investment, and fertility, it is possible that 'run-away' consumption and investment in children's education result from the interaction of this psychology and modern education-based labor markets and consumption possibilities.

Implications of Embodied-Capital Theory for Understanding Current Fertility Transitions

Embodied-capital theory links several existing theories of demographic transition and, in so doing, provides a new perspective on each one. It rationalizes the shift from natural fertility to parity-specific fertility in terms of a changing quality/quantity trade-off, as a transition from high fertility/low parental investment to low fertility/high parental investment. It also shows why the shift to lower mortality rates can lead to lower fertility, but for reasons that are very different from what traditional demographic-transition theory proposes. According to the latter, lower mortality precedes and drives lower fertility, because fewer children are required to maintain the population equilibrium and cultural ideologies evolve to maintain population equilibrium. According to the present view, mortality rates are but one important factor in the overall rate of return on investments in embodied capital.

This model can further serve to organize research on fertility transitions, with the fundamental expectation that exogenous factors affecting the rate of return on parental investment in the embodied capital of offspring will predict the timing of fertility decline and its distribution within populations. Understanding these dynamics may provide insights into the course of fertility reduction and its variation across countries and among individuals within countries. For example, most studies of predemographic-transition societies show that there is a positive association between wealth and fertility (Borgerhoff Mulder, 1992). There is some evidence for a negative relationship during the demographic transition (Kaplan et al., 1995a; Vining, 1986) and no relationship after the transition is complete (Kaplan et al., 1995a). It may be that only the wealthy can take advantage of high levels of investment in children's embodied capital when there are few schools and limited public expenditures on health and education. The process of demographic transition may progress through different sectors of the society as more and more people can take advantage of increased rates of return due to changes in the economy and changes in public investments. These same dynamics, in combination with international transfers of educational and health technology, may also explain the national differences in the timing of fertility reduction.

Understanding these dynamics also requires an understanding of how people process the mix of information derived from direct assessment, observation of others, and cultural guidelines. The rapid change in economics of production, public education, and public health may lead people to reject existing cultural models, look for new models, and assess costs and benefits of alternative courses of action more consciously on the basis of direct experience. International family planning programs present one set of new models, promoting lower fertility and greater investment in children's education. When those models are consistent with individual assessments of costs and benefits, the direction of change is clear. It is less clear, however, how people will respond to family planning messages in the short and long run if the payoffs to investment in embodied capital are low, either because schools are poor or because employment opportunities for educated people are limited (as appears to be the case in many sub-Saharan African countries).

Conclusion

According to embodied-capital theory, fertility and parental investment behavior result from the interaction of exogenous conditions with an evolved response system shaped by natural selection. In addition, a distinctive feature of the learning-intensive nature of human psychology is that information about the environment and guidelines for behavior are obtained through cultural sources as well as through direct personal assessment (Bock, 1999, 2001). We still know very little about the mix of physiological, psychological, and cultural processes underlying decisions regarding wealth flows, educational investments and consumption patterns, or how those processes relate to fertility decisions. Understanding the complex interplay of exogenous conditions, cultural forces, and evolved response systems in the determination of family structure and fertility presents a formidable challenge, but one that promises great rewards.

See also: Life Course: Sociological Aspects.

Bibliography

Bailey, R.C., Jenike, M.R., Ellison, P.T., Bentley, G.R., Harrigan, A.M., Peacock, N.R., 1992. The ecology of birth seasonality among agriculturalists in Central Africa. Journal of Biosocial Science 24, 393–412.

Bennett, P., Owens, I., 2002. Evolutionary Ecology of Birds: Life Histories, Mating Systems, and Extinction. Oxford University Press, Oxford, UK.

Blurton Jones, N.G., Hawkes, K., Draper, P., 1994a. Differences between Hadza and !Kung children's work: original affluence or practical reason. In: Burch, E.S. (Ed.), Key Issues in Hunter–Gatherer Research. Berg, Oxford, UK, pp. 189–215.

Blurton Jones, N.G., Hawkes, K., Draper, P., 1994b. Foraging returns of !Kung adults and children: why didn't !Kung children forage? Journal of Anthropological Research 50, 217–248.

Bock, J.A., 1995. The Determinants of Variation in Children's Activities in a Southern African Community. Ph.D. thesis. University of New Mexico.

Bock, J., 1999. Evolutionary approaches to population: implications for research and policy. Population and Environment 21, 193–222.

Bock, J., 2001. Evolutionary theory and the search for a unified theory of fertility. American Journal of Human Biology 14 (2), 145–148.

Bock, J., 2002. Learning, life history and productivity: children's lives in the Okavango Delta, Botswana. Human Nature 13 (2), 161–197.

Borgerhoff Mulder, M., 1992. Reproductive decisions. In: Smith, E.A., Winterhalder, B. (Eds.), Evolutionary Ecology and Human Behavior. Aldine, Hawthorne, NY, pp. 339–374.

Card, D., Krueger, A.B., 1992. Does school quality matter? Returns to education and the characteristics of public schools in the United States. Journal of Political Economy 100, 1–40.

Charnov, E.L., 1993. Life History Invariants. Oxford University Press, Oxford, UK.

Ellison, P., 2001. On Fertile Ground. Harvard University Press, Cambridge, MA.

Godfray, H.C.J., Partridge, L., Harvey, P.H., 1991. Clutch size. Annual Review of Ecology and Systematics 22, 409–429.

Goodman, A., Koupil, I., Lawson, D.W., 2012. Low fertility increases descendant socioeconomic position but reduces long-term fitness in a modern post-industrial society. Proceedings of the Royal Society B: Biological Sciences 279 (1746), 4342–4351.

Gurven, M., Kaplan, H., Gutierrez, M., 2006. How long does it take to become a proficient hunter? Implications for the evolution of delayed growth. Journal of Human Evolution 51, 454–470.

Hawkes, K., O'Connell, J.F., Jones, N.G.B., 1995. Hadza children's foraging: juvenile dependency, social arrangements, and mobility among hunter–gatherers. Current Anthropology 36, 688–700.

Hill, K., Hurtado, A.M., 1996. Ache Life History: The Ecology and Demography of a Foraging People. Aldine, Hawthorne, NY.

Hooper, P.L., 2011. The Structure of Energy Production and Redistribution among Tsimane' Forager-horticulturalists. Ph.D. thesis. University of New Mexico.

Hurtado, A.M., Hill, K.R., 1987. Early dry season subsistence ecology of the Cuiva foragers of Venezuela. Human Ecology 15, 163–187.

Kaplan, H.S., 1996. A theory of fertility and parental investment in traditional and modern human societies. Yearbook of Physical Anthropology 39, 91–135.

Kaplan, H.S., 1997. The evolution of the human life course. In: Wachter, K., Finch, C. (Eds.), Between Zeus and the Salmon: The Biodemography of Longevity. National Academy of Sciences, Washington, DC, pp. 175–211.

Kaplan, H.S., Lancaster, J.B., Anderson, K.G., Tucker, W.T., 2001. An evolutionary approach to below replacement fertility. American Journal of Human Biology 14 (2), 233–256.

Kaplan, H.S., Lancaster, J.B., 2000. Skills-based competitive labor markets, the demographic transition, and the interaction of fertility and parental human capital in the determination of child outcomes. In: Cronk, L., Irons, W., Chagnon, N. (Eds.), Human Behavior and Adaptation: An Anthropological Perspective. Aldine, New York, pp. 283–322.

Kaplan, H.S., Lancaster, J.B., Bock, J.A., Johnson, S.E., 1995a. Does observed fertility maximize fitness among New Mexican men? A test of an optimality model and a new theory of parental investment in the embodied capital of offspring. Human Nature 6, 325–360.

Kaplan, H.S., Lancaster, J.B., Bock, J.A., Johnson, S.E., 1995b. Fertility and fitness among Albuquerque men: a competitive labor market theory. In: Dunbar, R.I.M. (Ed.), Human Reproductive Decisions: Biological and Social Perspectives. St. Martin's Press, New York, pp. 96–136.

Kaplan, H.S., Lancaster, J.B., Hill, K.R., Hurtado, A.M., 2000. A theory of human life history evolution: diet, intelligence, and longevity. Evolutionary Anthropology 9, 156–185.

Kaplan, H.S., Robson, A.J., 2002. The emergence of humans: the coevolution of intelligence and longevity with intergenerational transfers. Proceedings of the National Academy of Sciences of the United States of America 99 (15), 10221–10226.

Kaplan, H.S., Robson, A.J., 2009. We age because we grow. Proceedings of the Royal Society B: Biological Sciences 276, 1837–1844.

Kaplan, H.S., Gurven, M., Winking, J., Hooper, P.L., Stieglitz, J., 2010. Learning, menopause, and the human adaptive complex. Annals of the New York Academy of Sciences 1204, 30–42.

Kaplan, H.S., Hooper, P.L., Stieglitz, J., Gurven, M. The Causal Relationship between Fertility and Infant Mortality: Prospective Analyses of a Population in Transition. Chapter in press at Oxford University Press.

Smith, C.C., Fretwell, S.D., 1974. The optimal balance between size and number of offspring. American Naturalist 108, 499–506.

Tracer, D.P., 1996. Lactation, nutrition, and postpartum amenorrhea in Lowland Papua, New Guinea. Human Biology 68, 277–292.

Vining, D.R., 1986. Social versus reproductive success: the central theoretical problem of human sociobiology. Behavioral and Brain Sciences 9, 167–216.

Walker, R., Hill, K., Kaplan, H., McMillan, G., 2002. Age-dependency in hunting ability among the Ache of eastern Paraguay. Journal of Human Evolution 42 (6), 639–657.

Werner, E.E., 1986. Amphibian metamorphosis: growth-rate; predation risk; and the optimal size at transformation. American Naturalist 128, 319–341.

Fertility Theory: Theory of Intergenerational Wealth Flows

Kristin Snopkowski, Boise State University, Boise, ID, USA
Hillard Kaplan, University of New Mexico, Albuquerque, NM, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Caldwell's theory of wealth flows explains fertility decline as a rational decision by parents based on the direction of intergenerational transfers. In high-fertility contexts, this theory proposes that children produce more than they consume and therefore provide net wealth to parents. In contrast, in low-fertility contexts, parents invest more in children, resulting in children being a net economic cost. Empirical tests of this hypothesis have not found evidence to show that children are net providers to parents in high-fertility contexts. A weaker prediction of the model, that when children are more expensive, parents desire fewer of them, has been supported.

Introduction

Caldwell's wealth flows theory (1976a, 1980, 2005) was developed to explain the demographic transition from high to low-fertility regimes. The theory, which focuses on the flow of resources between generations, has stimulated a large body of empirical research, and has been modified in response to those empirical findings and in response to worldwide decreases in fertility that have occurred since it was first proposed. In addition, during the last two decades, intergenerational wealth flows has become an increasing focus of research by evolutionary anthropologists who apply ecological models to explain both parental investment and fertility jointly.

This article reviews Caldwell's original proposal, its extensions, and the empirical results it has stimulated. It then provides an overview of new approaches being applied by evolutionary anthropologists, concluding with a synthesis of the current state of our knowledge and new directions for research.

Wealth Flows Theory

Caldwell's theory of wealth flows proposes that fertility behavior is rational decision-making by parents. The theory predicts that in high-fertility contexts, parental wealth is increased by having more children. In those contexts, children work at young ages (Caldwell, 1976b) and care for elderly or sick parents in adulthood (Caldwell, 1966). Caldwell (1976a) argues that children's work and support in old age together result in contributions of children to parents being greater than those from parents to children over their joint life spans.

In addition to the quantifiable help children provide, large families can be beneficial through their increased political power (especially adult children married into other families). Large families also provide insurance against risk; if, for example, an individual family has crop failure, they can call on extended family members for support (Caldwell, 1981; Caldwell et al., 1985). In such circumstances, children are a source of wealth for parents and therefore parents choose to have as many children as they can.

A corollary prediction is that parents will choose to have lower fertility when children no longer provide wealth to parents and the transfers of wealth instead flow from parents to children. The demographic transition from high-fertility to low-fertility regimes begins as values and behavioral norms prioritize individual accomplishments, freedom, and well-being over contribution to the extended family. As individuals shift their emotional connection to their spouse and children over extended kin, they begin increasing investments in children, making them more expensive, leading parents to desire fewer or no children. This transition occurred first in Western Europe, the United States, and Australia as schooling became universal, transmitting those values.

Caldwell (1980) argues that the primary determinant of the onset of the fertility transition is universal schooling's effect on the family social structure. While education increases the cost of childbearing because of lost production while children are at school, the cost of school fees and uniforms, and the demand by children for consumer goods, it also changes the way society views childhood. As societies incorporate western ideology of nuclear families from mass education and mass media, couples begin to prioritize their spouse and children over the extended family. It is this change that leads to fertility decline.

Schooling makes it clear that society views children as the future and children need to be valued as more than simply child laborers. Usually the first generation of mass schooling is enough to initiate fertility decline, as parents recognize that they will not receive more from children than they will have to invest in them. In societies where the extended family structure can be maintained even with a large number of children attending school, fertility decline may not occur until those children who attended school become parents.

Different researchers have defined levels of education in different ways, such as literacy rates or the percentage of individuals educated, but Caldwell argues that it is mass education (or universal schooling) that really matters. It is not until all (or the vast majority) of children are educated that the values surrounding childhood and nuclear families change, and it is the education of women that is particularly important in forcing a shift toward nucleation (as women expect close emotional bonds with their spouse) and higher investment in

children. Mass education has paved the way for the expansion of mass media, where books, newspapers, movies, radio, and television all reinforce western cultural values of nuclear families and smaller family size.

Following the directional change of the intergenerational wealth flows, where parents provide more for children than is reciprocated, economic rationality dictates no fertility, but social and psychological forces keep fertility from falling to zero. Fertility decline, as Caldwell describes it, is essentially about changing attitudes and values about family structure, which influences the economics of the family and results in a changing directional flow of goods. We can observe societies where there are large urban populations, some proportion of children are attending school, many people are engaged in a market economy, but if the flows of wealth remain from children to adults, then high fertility remains.

In other regions of the world, traditional value systems dominated modernization for some time. As most regions of the world have begun to evidence reduced fertility, the theory was modified to focus on the cultural transmission of western values via mass schooling and cultural diffusion. Thus, in its most recent form, the theory examines fertility transition as involving three elements. 1.Cultural diffusion of individualist western values affects the economic costs of children, proceeding across regions of the world. 2.When such values permeate the cultural context, children no longer provide net flows of wealth to parents and cease to increase parental wealth. 3.The economic effects of value change, in turn, affect fertility desires, leading people to adopt practices that lower the number of children they have.

Empirical Applications of the Theory

Caldwell's theory of wealth flows has led to extensive empirical research. Wealth flows theory has been addressed empirically in two ways. 'Weak' correlational tests of the theory have been conducted to determine whether fertility levels respond to the economic contributions of children (De Tray, 1983; Dow and Werner, 1983). Those tests tend to support the theory in that fertility tends to decrease as children spend less time in productive activities.

'Stronger' tests have also been conducted in which wealth flows across generations are actually measured to determine whether, in fact, more wealth flows from children to parents when fertility is high. Those tests quantify production and consumption rates across age groups to determine net flows between generations over the course of the entire life span. Almost universally, the stronger tests do not support the theory.

The evidence shows that in high-fertility contexts children do not produce enough before they marry to overcome the huge investment they received from parents at young ages, and most elderly individuals still produce more than they consume (Cain, 1982; Kaplan, 1994; Lee and Kramer, 2002; Lee, 2000; Stecklov, 1999).

Kaplan (1994) measured the caloric yield by age and sex in three hunter-gatherer societies and found that children do not produce more than they consume until approximately 20 years of age, an age at which many individuals already had children of their own, and that elderly individuals are usually net producers. Many other studies have come to similar conclusions. Mueller (1976) studied 27 traditional agricultural communities and found that children did not produce as much as they consumed until late teens and never produce enough to cover their consumption. Lee and Kramer (2002) came to similar conclusions studying the Maya agriculturalists. Stecklov (1999) and Cain (1982), both analyzing economic returns to childbearing in developing countries, again found that child labor did not produce enough to compensate for consumption.

Finally, Lee (2000) compared intergenerational transfers across different economic and cultural systems, including hunter-gatherers, horticulturalists, agriculturalists, intensive agriculturalists, and one post-industrialized context. He found that the only context where there were upward wealth flows was the post-industrialized country, specifically, the contemporary United States, exactly opposite to the wealth flows prediction. This is due to the social security system, which allows older Americans the opportunity to retire and become net consumers. The data on consumption and production curves suggest that the only time wealth flows upwards is in low-fertility contexts with extensive social security programs, contrary to Caldwell's predictions.

Another critique of wealth flows theory is that Caldwell's description of family relations in pretransition societies may describe the African context well, but it describes other contexts, such as Asia, quite poorly, as Asian countries have reduced their fertility even though a strong extended family system remained intact (Mason, 1997). Several authors have argued that old-age care of parents is not an important determinant of fertility as some countries initiated pension schemes without a drop in fertility and others experienced large declines in fertility, even though children were still responsible for parents in old age (Cleland, 1993; Galor, 2012). Thadani (1978) argues that it is hard to test empirically the direction of causality between emotional nucleation and economic change and that it is possible that economic conditions lead to changing family structures instead of the reverse.

In spite of this growing body of evidence, a definitive conclusion will require more research. Existing studies, unfortunately, have not measured the benefits of increased political power and insurance that may accrue to large families, because they are difficult to measure. Caldwell has argued that those added effects could shift the results of empirical studies from the conclusion that children consume more than they produce to children actually becoming a net benefit to parents (Caldwell, 2005).

In addition, even though children may not provide net economic benefits when fertility is high, as was first predicted, many studies show that fertility is reduced when parents experience higher actual or perceived costs of children, which supports a weak version of Caldwell's theory (Clay and Vander Haar, 1993; De Tray, 1983; Dow and Werner, 1983; Nag et al., 1978).

Caldwell's theory was revolutionary in combining the role of cultural influences and economic optimization in fertility decision-making, which has led to much of the current fertility research among biodemographers. Today, researchers seek to understand how individuals make fertility decisions in

changing conditions (Kaplan et al., 2002; Kaplan and Lancaster, 2000; Shenk et al., 2013; Shenk, 2009) and understand how changing from extended kin networks of family organization to more isolated nuclear families influences fertility outcomes (Mace and Sear, 2005; Mathews and Sear, 2013a; Sear and Mace, 2008; Turke, 1989).

Evolutionary Approaches to Wealth Flows and Family Relations

Evolutionary anthropologists also focus on the links between fertility and intergenerational transfers. They share Caldwell's view that fertility will be negatively associated with wealth flows to children. However, their theoretical approach differs from traditional wealth flows theory on several fronts. The basic underlying assumption governing evolutionary approaches is that natural selection has acted on the physiological and psychological processes that govern both fertility and parental investment. The expectation is that natural selection favors genetic variants which tend to maximize the production of descendants in the long run. Overall, higher fertility results in greater number of descendants, but increased fertility trades off against offspring quality, which, in turn, affects their ability to survive and reproduce (see Fertility Theory: Embodied-Capital Theory of Life History Evolution).

This leads to a major difference in conceptualization of the factors determining fertility and parental investment. Evolutionary anthropologists expect that environmental and socioeconomic factors will determine the relationship between fertility and child outcomes, and therefore predict both parental investment and fertility (Kaplan, 1996). For example, the nature of child mortality risks alters how parents should allocate resources to offspring. Some risks may not be preventable by increased parental protection or provisioning, and others may be buffered by parental assistance. If increases in parental investment will be ineffective, as in the former case, maximization of fitness will be associated with higher fertility. In contrast, when increased parental investment greatly increases offspring survival, lower fertility is the expected outcome (Kaplan et al., 2002). Similarly, and perhaps more importantly, the relationship between parental investment and offspring income in adulthood is also likely to affect fertility decisions. Again, evolutionary anthropologists expect that the impact of parental investment on wealth of offspring will impact fertility decisions; the greater the impact the lower the fertility, holding all else constant.

Other socioeconomic factors may also determine the relationship between child outcomes and fertility. The impact of parental investment on offspring income, and hence the offspring's ability to produce and invest in children, is expected to affect parental investment, and hence fertility, indirectly. As parental help becomes increasingly effective in affecting child outcomes, the expectation is that fertility will decrease accordingly, holding all else constant. The role of culture in evolutionary models is informational. The relationship between the environment and offspring outcomes as a function of fertility, especially in the context of socioeconomic change, is not necessarily obvious to people. The behavior of other people can serve as a cue about those relationships. Thus, the causal relationships linking culture to fertility in evolutionary models is different from the view proposed by Caldwell. In the Caldwell model, culture change is a prime driver of behavior change, whereas in evolutionary models, culture change is driven by changes in the socioeconomic conditions people face, and hence is rather the effect of changing conditions rather than a causal determinant (Snopkowski and Kaplan, 2014).

There has been a long-standing appreciation that fertility behavior may be the result of economic factors and/or social network effects, it is only recently that evolutionary anthropologists have tried to explore how microenvironments influence fertility behavior. Within a local environment, individuals are exposed to different network influences, which lead to heterogeneity across communities. This allows for tests of the relative contribution of social network effects over and above those of individual characteristics. Recent work by Kaplan et al. (in press) compares individual-level and community-level effects on fertility outcomes among the Tsimane of Bolivia and find that individual-level effects explain most of the variance in fertility outcomes, while community-level predictors have little effect. In contrast, Colleran and Mace (2014), using a Polish sample, find that community-level predictors have significant effects on fertility outcomes, even after individual-level predictors are controlled for. These studies demonstrate that the role of cultural norms in predicting fertility outcomes is still not completely understood, but researchers are devising new and unique ways to test their effects.

There are still many areas of study for future research. With increasing globalization, local conditions lead to niches in the global economy, which, in turn, influence women's roles and have a direct effect on fertility outcomes. For example, in contexts with large garment industries that primarily employ women, such as Bangladesh, women face a steep tradeoff between work and child care. Women with children typically have to send children to their natal villages to be raised by kin. This is in sharp contrast to contexts where women's work has been integrated with child care and there is less conflict between women's work and child care. There are extensive possibilities to explore how socioecological factors affect fertility at a community level.

Role of the Family in Fertility Decision-Making

While Caldwell conceptualized the extended family as a family structure that required transfers from young to old members, other researchers have argued that extended kin operate to provide additional resources for childbearing (Hrdy, 2005). The loss of the extended family structure may mean that the costs of children become larger for parents because they cannot be dispersed to extended kin members (Turke, 1989) or that pronatal messages, which may come disproportionally from kin, are reduced as individuals are located further from extended kin members (Newson et al., 2005).

Evidence has been mounting for the positive effects extended kin (usually parents or in-laws) have on the survivorship of children and fertility rates. Children are more likely to survive in many contexts if grandparents are alive, with effects generally being strongest for maternal grandmothers (Beise

and Voland, 2002; Beise, 2005; Hadley, 2004; Kemkes-Grottenthalef, 2005; Lahdenperä et al., 2004; Sear et al., 2000; Sear, 2008; Tymicki, 2004). There is also evidence that grandmothers have positive effects on children's nutritional status (Gibson and Mace, 2005; Sear et al., 2000). In several contexts, grandmothers provide needed help to children and grandchildren; grandmothers reduce mother's work energy expenditure and reduce maternal direct child care among the Aka foragers of central Africa (Meehan et al., 2013), they reduce risk of grandchild mortality and low birth weight when they are the primary source of support for mothers in Puerto Rico (Scelza, 2011), and they relieve daughters of heavy domestic tasks in rural Ethiopia (Gibson and Mace, 2005). Finally, there is evidence that individuals who have close bonds with parents are more likely to engage in reproduction (Mathews and Sear, 2013a,b; Waynforth, 2012) and that having kin available who provide child care increase the likelihood of additional births (Bereczkei, 1998; Kaptijn et al., 2010). This thriving research area has demonstrated the positive effects grandparents have on grandchild outcomes, again providing evidence that resources flow from parents to children and grandchildren instead of the reverse.

Given that the variation in kin effects across contexts is not well understood and we expect kin to have differing effects depending on the local fertility norms and socioecologies, this provides a thriving area for future research. Further, we may expect variation depending on the type of kin member, as some kin are more closely related than others and some kin have their own reproductive opportunities, which may lead to kin reproductive conflict instead of cooperation. Empirical evidence shows mothers-in-law tend to have a positive effect on fertility outcomes for daughters-in-law (more so than mothers on daughter's fertility) (Sear and Coall, 2011), but we do not truly understand why this occurs. Both social and economic hypotheses have been brought forward as potential explanations, but future work will likely explore this evolutionary puzzle.

Conclusion

Research conducted in the past 30 years has provided support for the claim that parents adjust their fertility based on the actual and perceived costs of raising children, where higher costs and lower levels of productive work by children result in reduced fertility. Strong tests of the wealth flows theory provide evidence that the flow of wealth has not traditionally been from children to adults. In spite of this, Caldwell's theory of intergenerational wealth flows has spurred much research into the interrelationship of cultural values and economic decision-making, the amount of production and consumption across different age categories, and tests of differing hypotheses of the demographic transition. The interaction of economics and culture is still at the core of our understanding of fertility decline and understanding their interaction continues to generate new research. Finally, the evidence shows that grandparents provide tangible benefits to children and grandchildren, which promotes reproduction for children.

See also: Family Size Preferences; Fertility Theory: Embodied-Capital Theory of Life History Evolution; Fertility Theory; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Fertility and Social Networks.

Bibliography

Beise, J., 2005. The helping and the helpful grandmother: the role of maternal and paternal grandmothers in child mortality in the seventeenth- and eighteenth-century population of French settlers in Quebec, Canada. In: Voland, E., Chasiotis, A., Schiefenhovel, W. (Eds.), Grandmotherhood: The Evolutionary Significance of the Second Half of Female Life. Rutgers University Press, New Brunswick, New Jersey, and London, pp. 215–238.

Beise, J., Voland, E., 2002. A multilevel event history analysis of the effects of grandmothers on child mortality in a historical German population. Demographic Research 7, 469–498. http://dx.doi.org/10.4054/DemRes.2002.7.13.

Bereczkei, T., 1998. Kinship network, direct childcare, and fertility among Hungarians and Gypsies. Evolution and Human Behavior 19 (5), 283–298. http://dx.doi.org/10.1016/S1090-5138(98)00027-0.

Cain, M., 1982. Perspectives on family and fertility in developing countries. Population Studies 36 (2), 159–175. http://dx.doi.org/10.1080/00324728.1982.10409026.

Caldwell, J.C., 1966. The erosion of the family: a study of the fate of the family in Ghana. Population Studies 20 (1), 5–26. http://dx.doi.org/10.1080/00324728.1966.10406081.

Caldwell, J.C., 1976a. Toward a restatement of demographic transition theory. Population and Development Review 2 (3/4), 321. http://dx.doi.org/10.2307/1971615.

Caldwell, J.C., 1976b. Fertility and the household economy in Nigeria. Journal of Comparative Family Studies 7 (2), 193–253.

Caldwell, J.C., 1980. Mass education as a determinant of the timing of fertility decline. Population and Development Review 6 (2), 225–255.

Caldwell, J.C., 1981. The mechanisms of demographic change in historical perspective. Population Studies 35 (1), 5–27. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/11630767.

Caldwell, J.C., 2005. On net intergenerational wealth flows: an update. Population and Development Review 31 (4), 721–740.

Caldwell, J.C., Reddy, P.H., Caldwell, P., 1985. Educational transition in rural South India. Population and Development Review 11 (1), 29. http://dx.doi.org/10.2307/1973377.

Clay, D.C., Vander Haar, J.E., 1993. Patterns of intergenerational support and childbearing in the Third World. Population Studies 47 (1), 67–83. http://dx.doi.org/10.1080/0032472031000146736.

Cleland, J., 1993. Equity, security and fertility: a reaction to Thomas. Population Studies 47 (2), 345–352. http://dx.doi.org/10.1080/0032472031000147056.

Colleran, H., Jasienska, G., Nenko, I., Galbarczyk, A., Mace, R., 2014. Community-level education accelerates the cultural evolution of fertility decline. Proceedings of the Royal Society B: Biological Sciences 281, 2013–2732. http://dx.doi.org/10.1080/0032472031000147056.

De Tray, D., 1983. Children's work activities in Malaysia. Population and Development Review 9 (3), 437. http://dx.doi.org/10.2307/1973317.

Dow Jr., T.E., Werner, L.H., 1983. Prospects for fertility decline in rural Kenya. Population and Development Review 9 (1), 77–97.

Galor, O., 2012. The demographic transition: causes and consequences. Discussion Paper Series, Forschungsinstitut zur Zukunftder Arbeit, No. 6634, http://hdl.handle.net/10419/58611

Gibson, M.A., Mace, R., 2005. Helpful grandmothers in rural Ethiopia: a study of the effect of kin on child survival and growth. Evolution and Human Behavior 26 (6), 469–482. http://dx.doi.org/10.1016/j.evolhumbehav.2005.03.004.

Hadley, C., 2004. The costs and benefits of kin: kin networks and children's health among the Pimbwe of Tanzania. Human Nature 15 (4), 377–395.

Hrdy, S.B., 2005. Evolutionary context of human development: the cooperative breeding model. In: Carter, C.S., Ahnert, L., Grossmann, K.E., Hrdy, S.B., Lamb, M.E., Porges, S.W., Sachser, N. (Eds.), Attachment and Bonding: A New Synthesis. MIT Press, Cambridge, MA, pp. 9–32.

Kaplan, H., 1994. Evolutionary and wealth flows theories of fertility: empirical tests and new models. Population and Development Review 20 (4), 753–791.

Kaplan, H., 1996. A theory of fertility and parental investment in traditional and modern human societies. Yearbook of Physical Anthropology 39, 91–135.

Kaplan, H., Hooper, P.L., Gurven, M. The causal relationship between fertility and infant mortality: prospective analyses of a population in transition. In: Kreager, P., Winney, B., Ulijaszek, S., Capelli, C. (Eds.), Population in the Human Sciences: Concepts, Models, Evidence. Oxford University Press, Oxford.

Kaplan, H., Lancaster, J.B., Tucker, W.T., Anderson, K.G., 2002. Evolutionary approach to below replacement fertility. American Journal of Human Biology: the Official Journal of the Human Biology Council 14 (2), 233–256. http://dx.doi.org/10.1002/ajhb.10041.

Kaplan, H.S., Lancaster, J.B., 2000. The evolutionary economics and psychology of the demographic transition to low fertility. In: Cronk, L., Chagnon, N., Irons, W. (Eds.), Adaptation and Human Behavior: An Anthropological Perspective. Aldine de Gruyter, New York, pp. 283–322.

Kaptijn, R., Thomese, F., van Tilburg, T.G., Liefbroer, A.C., 2010. How grandparents matter: support for the cooperative breeding hypothesis in a contemporary Dutch population. Human Nature 21 (4), 393–405. Retrieved from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2995872&tool=pmcentrez&rendertype=abstract.

Kemkes-Grottenthalef, A., 2005. Of grandmothers, grandfathers and wicked step-grandparents. Differential impact of paternal grandparents on grandoffspring survival. Historical Social Research 30 (3), 219–239.

Lahdenperä, M., Lummaa, V., Helle, S., Tremblay, M., Russell, A.F., 2004. Fitness benefits of prolonged post-reproductive lifespan in women. Nature 428 (6979), 178–181. http://dx.doi.org/10.1038/nature02367.

Lee, R., 2000. A cross-cultural perspective on intergenerational transfers and the economic life cycle. In: Mason, A., Tapinos, G. (Eds.), Sharing the Wealth: Demographic Change and Economic Transfers between Generations, vol. 54. Oxford University Press, Oxford, pp. 17–56. http://dx.doi.org/10.2307/2058948.

Lee, R.D., Kramer, K.L., 2002. Children's economic roles in the Maya family life cycle: Cain, Caldwell, and Chayanov revisited. Population and Development Review 28 (3), 475–499.

Mace, R., Sear, R., 2005. Are humans cooperative breeders? In: Voland, E., Chasiotis, A., Schiefenhoevel, W. (Eds.), Grandmotherhood: The Evolutionary Significance of the Second Half of Female Life. Rutgers University Press, New Brunswick, NJ, pp. 143–159.

Mason, K.O., 1997. Explaining fertility transitions. Demography 34 (4), 443–454. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/9545624.

Mathews, P., Sear, R., February 2013a. Does the kin orientation of a British woman's social network influence her entry into motherhood? Demographic Research 28, 313–340. http://dx.doi.org/10.4054/DemRes.2013.28.11.

Mathews, P., Sear, R., 2013b. Family and fertility: kin influence on the progression to a second birth in the British Household Panel Study. PloS One 8 (3), e56941. http://dx.doi.org/10.1371/journal.pone.0056941.

Mueller, E., 1976. The economic value of children in peasant agriculture. In: Ridker, R.G. (Ed.), The Search for Selective Interventions. RFF Press, New York, pp. 182–266. Population and Development.

Meehan, C.L., Quinlan, R., Malcom, C.D., 2013. Cooperative breeding and maternal energy expenditure among Aka foragers. American Journal of Human Biology 25, 42–57. http://dx.doi.org/10.1002/ajhb.22336.

Nag, M., White, B.N.F., Peet, R.C., Bardhan, A., Hull, T.H., Johnson, A., Tax, S., 1978. An anthropological approach to the study of the economic value of children in Java and Nepal. Current Anthropology 19 (2), 293–306.

Newson, L., Postmes, T., Lea, S.E.G., Webley, P., 2005. Why are modern families small? Toward an evolutionary and cultural explanation for the demographic transition. Personality and Social Psychology Review 9 (4), 360–375. http://dx.doi.org/10.1207/s15327957pspr0904_5.

Scelza, B., 2011. The place of proximity: social support in mother-adult daughter relationships. Human Nature 22 (1–2), 108–127. http://dx.doi.org/10.1007/s12110-011-9112-x.

Sear, R., Mace, R., 2008. Who keeps children alive? A review of the effects of kin on child survival. Evolution and Human Behavior 29 (1), 1–18. http://dx.doi.org/10.1016/j.evolhumbehav.2007.10.001.

Sear, R., Mace, R., McGregor, I.A., 2000. Maternal grandmothers improve nutritional status and survival of children in rural Gambia. Proceedings of the Royal Society B 267 (1453), 1641–1647. http://dx.doi.org/10.1098/rspb.2000.1190.

Sear, R., 2008. Kin and child survival in rural Malawi. Human Nature 19 (3), 277–293. http://dx.doi.org/10.1007/s12110-008-9042-4.

Sear, Rebecca, Coall, D., 2011. How much does family matter? Cooperative breeding and the demographic transition. Population and Development Review 37 (Suppl. 1), 81–112. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/21280366.

Shenk, M.K., 2009. Testing three evolutionary models of the demographic transition: patterns of fertility and age at marriage in urban South India. American Journal of Human Biology: the Official Journal of the Human Biology Council 21 (4), 501–511. http://dx.doi.org/10.1002/ajhb.20943.

Shenk, M.K., Towner, M.C., Kress, H.C., Alam, N., 2013. A model comparison approach shows stronger support for economic models of fertility decline. Proceedings of the National Academy of Sciences of the United States of America 1–6. http://dx.doi.org/10.1073/pnas.1217029110.

Stecklov, G., 1999. Evaluating the economic returns to childbearing in Cote d'Ivoire. Population Studies 53, 1–17.

Snopkowski, K., Kaplan, H., 2014. A synthetic biosocial model of fertility transition: testing the relative contribution of embodied capital theory, changing cultural norms, and women's labor force participation. American Journal of Physical Anthropology 154, 322–333.

Thadani, V.N., 1978. The logic of sentiment: the family and social change. Population and Development Review 36 (2), 457–499.

Turke, P.W., 1989. Evolution and the demand for children. Population and Development Review 15 (1), 61–90.

Tymicki, K., 2004. Kin influence on female reproductive behavior: the evidence from reconstitution of the Bejsce parish registers, 18th to 20th centuries. Poland. American Journal of Human Biology 16 (5), 508–522. http://dx.doi.org/10.1002/ajhb.20059.

Waynforth, D., 2012. Grandparental investment and reproductive decisions in the longitudinal 1970 British cohort study. Proceedings of the Royal Society B 279 (1731), 1155–1160. http://dx.doi.org/10.1098/rspb.2011.1424.

Fertility Theory: Theory of Life History Evolution

Gert Stulp, University of Groningen, Groningen, The Netherlands; and London School of Hygiene and Tropical Medicine, London, UK
Louise Barrett, University of Lethbridge, AB, Canada; and UNISA, Applied Behavioural Ecology and Ecosystems Research Unit, Johannesburg, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Variation in fertility patterns in relation to environmental factors and mortality schedules can be understood using evolutionary life history theory (LHT). In addition to explaining differences within and between human populations, LHT allows us to make direct comparisons with other species. Here we discuss LHT, the concept of trade-offs, and the importance of this theory as applied to humans. We end by discussing the ways in which LHT potentially explains patterns of low fertility in contemporary industrialized society.

> Life histories lie at the heart of biology; no other field brings you closer to the underlying simplicities that unite and explain the diversity of living things and the complexities of their life cycles.
> Stearns, 1992: 9

In 2014, the world population topped 7 billion, with almost all observed population growth occurring in the less-developed world. At the same time, evidence suggests that global fertility rates are actually beginning to decline, with almost all developing nations showing some sign of entering a demographic transition (e.g., Sanderson, 2013). In the industrialized world, where fertility has been in decline for over a century, total fertility rates hit an all-time low in the 1990s, dipping below replacement levels. Yet, recent fertility projections suggest that this pattern may now be reversing, with fertility set to rise to replacement levels, and perhaps beyond, raising the intriguing possibility that reproductive strategies may be shifting in response to the conditions of the post-industrial world (Myrskylä et al., 2009, 2013). What determines the number of children a woman will have in her lifetime? How can we make sense of the fertility patterns we see, and understand why they vary in both time and space?

These questions have long engaged demographers, economists, and sociologists, but completely satisfactory answers continue to elude us. These are also questions that, in broad perspective, evolutionary biologists seek to answer about all species, as they relate directly to measurements of fitness (approximated by the number of descendants that contribute to future generations), and hence to the action of natural selection, the process by which adaptive evolutionary change occurs. Our goal here is to show how aspects of evolutionary thinking can be brought to bear on questions relating to human fertility patterns, and that, although we are not yet able to solve all the puzzles of human fertility variation, we can begin to address these issues in ways that productively augment more traditional social science approaches.

Life History Theory and the Cost of Reproduction

Life history theory (LHT) provides a robust evolutionary framework in which to situate studies of human reproductive and mortality schedules, in ways that also allow us to make direct comparisons with other species. An organism's life history is the combination of the probabilities of survival and fecundity that a species displays in its natural environment (Partridge and Harvey, 1988); it is therefore closely tied to the study of demography (Stearns, 1992). The theory of life history evolution posits that fitness is maximized by natural selection over the life span of an animal. This is not as easy as simply maximizing fertility at every age, as fertility in early life may be traded off for decreases in either fecundity, fertility, or survival in later life. That is, there is a cost to reproduction. (By fertility we mean the total number of offspring produced, whereas fecundity refers to the potential to reproduce.)

Why is reproduction costly? One reason is that there are ecological costs associated with the act of reproducing. For most (if not all) animals, the environment is a dangerous place: acquiring resources, mating, giving birth, and raising young are all activities that leave animals prone to physical injuries, exposed to disease, or at risk of death by predation (Roff, 1992). More importantly, reproduction is costly because it requires energy that must be diverted away from other bodily processes (the 'Principle of Allocation'). An animal's life history is a consequence of three biological processes that compete for resources in the body: repair (or maintenance), growth, and reproduction. Investment in reproduction inevitably decreases the amount of energy that can be invested in repair and growth and, as a consequence, decreases future fecundity and survival. These trade-offs – the costs paid in terms of fitness when a beneficial change in one trait is linked to a detrimental change in another (Stearns, 1992) – are a fundamental concept in LHT.

The cost of reproduction, and its associated trade-offs, not only affects an individual's own fecundity and future survival, but also that of its offspring. In an elegant study, Dijkstra et al. (1990) manipulated brood size (i.e., the number of chicks produced in a single breeding attempt) in kestrels, a small bird of prey, either adding or removing two nestlings to some nests (as well as leaving some nests untouched). They found that adding nestlings increased the total number of fledglings produced, but that the offspring in such nests had the slowest growth and highest nest mortality rates compared to nests that were reduced in size. Moreover, when raising an enlarged brood, male parents spent much more

 http://dx.doi.org/10.1016/B978-0-08-097086-8.34012-0

time in flight and flight-hunting, whereas female parents were lighter, and lost weight more rapidly. Most strikingly perhaps, the parents of enlarged broods also displayed lower survival. The cost of reproduction, therefore, is multifaceted, with both intragenerational effects (decreased survival of the parents) and intergenerational effects (reduced offspring survival, decreased mass of offspring).

Life histories are characterized by these kinds of trade-offs that, in one way or another, serve to balance the energetic resources allocated to current versus future survival and reproduction. The form such trade-offs take, and hence the life history trajectory followed, depends critically on the nature and quality of the environment an organism encounters because this will differentially influence the chances of survival and reproduction.

How Large and How Soon?

Many organisms must first reach a minimum size before they are able to reproduce, simply because a certain level of bodily development is needed to support any kind of reproductive function. Growth is often extended beyond this minimum body size, however, because a larger body size is beneficial for fecundity and juvenile survival. All else being equal, then, animals try to maximize body size. The amount of time an organism spends growing to achieve larger body size hinges directly on environmental conditions. Better quality environments lead to increased growth and hence larger, more fecund adult animals. In poor environments, growth is slowed due to reduced energy intake, which may lead to reduced adult size, but large body size may also be less favorable under such conditions, as this incurs greater costs of maintenance and repair. For a number of species, including humans, growth ceases entirely once reproduction begins (the so-called 'determinate' growth), and selection pressures and trade-offs change dramatically once this point is reached; hence the switch from the growth phase to the reproductive phase (i.e., the age of sexual maturity) is a particularly crucial transition in an animal's life history.

Sexual maturity is usually defined as age at first birth in LHT, rather than by any physiological measure of maturation (Stearns, 1992). The benefit of maturing at an early age is largely self-evident: it reduces the chances of dying before an individual has had the chance to reproduce. Furthermore, offspring of early-maturing animals will themselves mature earlier than those of later-maturing animals and hence start reproducing sooner. Delaying maturity when mortality risk is low can also be advantageous, however, because extended growth and larger size may increase the chances of survival, as well as initial and future fecundity, the number of reproductive events, and the quality of the offspring or parental care provided.

Environment-specific mortality risks thus strongly influence age at sexual maturity. Higher mortality reduces life expectancy, such that animals cannot 'afford' to postpone reproducing, and hence will mature at an early age, trading this off against a longer growth period and larger adult body size. Conversely, low mortality in later life results in later maturation, allowing individuals to reach a larger adult size. The relationship between low mortality and delayed age at first birth has consistently been documented, across species, across populations within species, and within populations. For instance, among humans, cross-national average age at first birth shows a strong negative association with mortality rates in those countries. Even at quite small scales, such as within a city, mortality predicts the timing of first birth (Nettle, 2011). Although Nettle (2011) did not consider the relationship between mortality and growth as the source of these differences, other studies have shown an interaction between patterns of growth, mortality, and age at maturity across populations (Walker et al., 2006).

How Varying Environments Lead to Reaction Norms

As the above makes clear, the trade-off between age and size at maturity, which in essence represents the trade-off between current and future reproduction, is environment specific. If a species occupies an environment that is either very stable over time or the species' range is very narrow, natural selection will likely have solved the optimality problem such that all animals will mature at a specific body size X and age Y (although, of course, there may be some genetic variation between individuals that gives rise to phenotypic or trait variation). Now imagine that a species occupies a broader range of environments and is likely to face more variable environmental conditions as a result. This variation produces variation in growth rates and age- and size-specific mortality rates (Stearns and Koella, 1986). Under poor environmental circumstances, for example, growth will be slowed substantially. If the animal still matured at body size X under such conditions, it would take considerable time to reach this point; this would be risky given that mortality may strike before the animal reaches this size. If, however, the animal matured at an earlier age, Y, it would do so at a very small adult size, which would also impose a cost in terms of future survival and fecundity. A different compromise between X and Y is therefore needed to reach an optimal outcome for this specific environment. When the environments that animals encounter are variable in this way, 'reaction norms' are likely to evolve. A reaction norm is a set of phenotypes produced by a single genotype across a range of environmental conditions, and allows animals to track the optimal trade-off. Thus, phenotypic variation across environments may represent the expression of an evolved reaction norm, which ensures an appropriate trade-off between life history components. It is also true that some of this variation can arise from genetic variation across individuals in the reaction norms themselves.

The finding that rates of mortality across neighborhoods within a city predict age at first birth potentially can be interpreted as an expressed reaction norm (Nettle, 2011), because these phenotypic differences are unlikely to be (solely) due to genetic differences. Cross-population differences in age at first birth and patterns of mortality also fit well with LHT predictions: as the chance of reaching adulthood decreases, earlier onset of sexual maturity is generally observed (Walker et al., 2006). These patterns cannot, however, be unambiguously interpreted in terms of a reaction norm. In some Pygmy groups, for instance, sexual maturity occurs early and at very short

adult stature in environments of high mortality, and these characteristics are linked to distinct hormonal profiles with a genetic basis, suggesting the action of natural selection on early reproduction, rather than a plastic response to the environment. Indeed, testing for reaction norms remains an ongoing challenge as we still know surprisingly little about population-level patterns and individual-level biological processes in relation to age, growth rates, and size at maturity (Walker et al., 2006; Stearns, 1992).

How Many and How Much to Invest?

The next decision an animal faces once it has reached the reproductive phase is how many offspring it should produce, and how much it should invest in them. Any energy invested in the current offspring entails costs in terms of future survival and offspring production (Dijkstra et al., 1990). Although in many species, clutch (or litter) size can be adjusted, the same does not hold for all animals, including humans. Humans are generally restricted to a litter size of one, as are most other primates. Current theorizing suggests that producing a low number of offspring on a regular basis is adaptive in primates, because of their long life span and their reliance on a highly variable resource base, both of which leave them vulnerable to unpredictable shortfalls in available resources – and hence to low juvenile survival – during any particular breeding attempt (Jones, 2011). By limiting their reproductive effort to a single offspring at a time, mothers are able to invest more in their own survival, thereby increasing their life span and ensuring a larger number of reproductive events. In this way, reproductive failure in any one year is not catastrophic in terms of fitness (Jones, 2011).

There is, however, a rather unusual case in humans, which sheds light on the cost of reproduction, namely twinning – a relatively rare event (0.6–4.5%; Lummaa, 2007). Compared to women who have singleton births, mothers of twins have a higher risk of dying during childbirth, longer subsequent birth intervals, are more likely to terminate reproduction entirely, are more likely to fail to raise their next offspring, and are more likely to die earlier in the postreproductive phase (Lummaa, 2007). Moreover, twins are usually smaller in size than singleton births and experience lower survivorship (Sear et al., 2001). Mothers who give birth to twins are also larger in size than those who give birth to singletons (Sear et al., 2001), suggesting that twinning is indeed costly and that larger maternal body size is needed to bear these increased costs. The cost of twinning extends to both the parents (intragenerational) and the offspring themselves (intergenerational). Twinning, however, may also provide advantages in terms of fitness, if the costs of reproduction are offset by the benefits of gaining two offspring from a single birth event (Sear et al., 2001). For example, Lummaa et al. (1998) observed that lifetime reproductive success of mothers of twins was higher in areas of high resource availability, but lower in areas of low resource availability compared to mothers of singletons. Moreover, they showed that rates of twinning were higher in those areas in which twins provided fitness advantages, providing some support to the notion that natural selection may maintain, or even favor, twinning under some circumstances (Lummaa et al., 1998).

Although humans generally produce only one offspring per reproductive event, they are rather unusual in that, due to a long period of offspring dependence, parents usually raise several offspring of varying ages simultaneously, with resources distributed over all children raised. Indeed, humans are different from other primates in a number of respects: in addition to a lengthy developmental period and dependence on caregivers, we also possess large expensive brains, have a relatively long adult life, and display intergenerational systems of resource flow. All these traits have been argued to stem from the fact that humans are adapted to a particularly skill-intensive foraging niche, in which many years of learning are needed to reach high levels of competence (Kaplan et al., 2002). Attempting to raise a series of children whose demands on parental care overlap in time inevitably leads to conflicts, given that resources are finite. As children are both heavily dependent and energetically costly in their first years of life, the biggest conflicts often arise when births are very closely spaced in time. Indeed, there is substantial evidence that shorter interbirth intervals increase child mortality (Hobcraft et al., 1983; Koenig et al., 1990). The negative effects of twinning on survival as discussed above – a case in which the lowest possible interbirth interval is achieved – also fit well with these patterns.

The adverse effects of short interbirth intervals are generally considered to be a consequence of either 'maternal depletion syndrome' (i.e., mothers do not have sufficient time to recover from the physiological and nutritional demands associated with pregnancy, parturition, and breast-feeding) or the competition between siblings over parental resources (Koenig et al., 1990). These effects are arguably strong enough to have led to the evolution of lactational amenorrhea: a postnatal period of infecundity that arises as a consequence of hormonal changes induced by breast-feeding (Howie and McNeilly, 1982). Indeed, breast-feeding is the most widely used method of controlling birth spacing in populations without modern contraception; the energetic demands of milk production are high, and nursing offspring represents a significant constraint on the availability of metabolic energy to support a new pregnancy (Ellison, 2003). As LHT predicts, then, investment in current offspring strongly limits females' ability to invest in future offspring.

Fertility, Longevity, and the Difficulties of Measuring Life History Trade-Offs

In principle, identifying trade-offs among different life history trade-offs should be straightforward, given an understanding of underlying theory and a species' ecology. In practice, however, it is not so easy. A consideration of the trade-off between fertility and survival illustrates this point perfectly. For species that reproduce repeatedly, like humans and other primates, this trade-off is dependent on both the spacing of births and on the total number of dependent young that need to be raised to adulthood. Increasing resources in reproduction detracts resources from both parental investment into children (which should reduce their quality

and hence their chances of survival) and also from bodily maintenance (which should lead to reduced parental survival). Studies of the trade-offs between fertility, child survival, and parental survival, however, reveal anything but consistent results (Lawson et al., 2012; Gagnon et al., 2009), with both positive as well as negative relationships observed.

Although there may, of course, be plausible theoretical reasons why fertility can potentially increase child or parental survival (for example, under some conditions, children may be successful at procuring resources for themselves and the family, changing the dynamic of the trade-offs in such ecologies: Lawson et al., 2012), the variability seen may also reflect the difficulty of accurately measuring trade-offs. Obviously, experimental methods (such as manipulating clutch size or breeding experiments; Roff, 1992) are not possible in humans, and we are left with assessing either phenotypic correlations between life history traits or genetic correlations between traits based on family data. These correlational methods necessarily set boundaries on the confidence we can place on the results obtained. For example, unidentified confounding variables may give rise to relationships that are counter to life history predictions (it should be noted, however, that this problem is most definitely not restricted to humans; Partridge and Harvey, 1988; Stearns, 1992). One such confound is the quality of the individual, or more specifically, the energy budget the individual has at its disposal. To put this in everyday terms, one might predict that most people would be able to afford either an iPad or an iPhone, given the expense of such items, and so would have either one or the other, but not both (i.e., there would be a trade-off). For individuals with a larger bank balance, however, such compromises are unnecessary; they can afford to buy both an iPad and an iPhone. The presence of such differences across individuals in the resources available can easily obscure the true nature of the trade-offs operating (Noordwijk and Jong, 1986). Another complicating factor is that our ability to identify trade-offs may be clouded by the presence of genotype by environment interactions, such that different genotypes respond differently to different environments (see Stearns (1992) for further discussion). These underlying difficulties may account for variation in, or lack of, observed trade-offs in a population.

Even when a predicted trade-off is observed, we often cannot infer causality because of the correlational methods used. For instance, increased reproductive effort (i.e., higher fertility) should come at a cost of parental effort (i.e., levels of parental investment), and indeed recent studies show that high fertility is often negatively related to child survival (Lawson et al., 2012). It is also possible, however, that the causality here is reversed: when mortality is high, women may increase their fertility in an effort to 'replace' deceased children, and/or they may preemptively increase fertility because they expect to experience a certain level of child mortality (Lawson et al., 2012; Winterhalder and Leslie, 2002). Although research on humans is unlikely to completely overcome the measurement difficulties identified above, the greater availability of databases containing LHT-relevant information, and their higher quality in terms of both sample size and level of detail, along with more advanced statistical methods, may well overcome these limitations to some degree (Lawson et al., 2012; Gagnon et al., 2009).

Quantity, Quality, and the Demographic Transition

Among industrialized societies, advances in medical science and public health have reduced child mortality to very low levels. Consequently, any reduction in parental investment may not manifest itself in terms of reduced offspring survival, but may be more nuanced. For example, high fertility is related to many adverse consequences for offspring due to competition between siblings. A larger number of siblings has been found to reduce grades at school, adult education and income, height, and health (Lawson and Mace, 2011), providing strong evidence for the trade-off between the quantity and quality of children.

It is exactly this trade-off that has been suggested, in the evolutionary literature, as the main explanation for the demographic transition (the switch from high to low fertility and mortality regimes that accompanies economic development and industrialization). Kaplan's (1996; Kaplan et al., 2002) embodied capital theory is worth highlighting in this regard as it explicitly incorporates aspects of prominent economic theories of quantity–quality trade-offs (Becker, 1960; Shenk et al., 2013), along with demographic theories of fertility (Bongaarts, 1978), into an evolutionary life history model. The theory treats growth, development, and maintenance as investments in stocks of embodied capital in both self and offspring. Importantly, this embodied capital includes knowledge and skills, as well as somatic tissue, thus bringing cognitive abilities into the realm of LHT. For humans, growth of knowledge may be as important as bodily growth with respect to adult competence and ultimately fitness maximization. There are a number of other theories of the demographic transition in the literature (see e.g., Shenk et al. (2013) for review), particularly those that focus more heavily on cultural explanations, but we will not discuss these here, as our aim is to illustrate how LHT attempts to explain this phenomenon in terms of the relevant trade-offs.

Kaplan's (1996; Kaplan et al., 2002) starting point is to argue that fertility regulation mechanisms are adapted to the selection pressures of the learning-dependent, skill-intensive hunter-gatherer foraging niche. Natural selection is hypothesized to have acted on human psychology such that people are sensitive to the environmental variation in learning opportunities and their likely payoffs, and levels of parental investment are calibrated accordingly. Individuals adaptively adjust levels of parental investment via breast-feeding (i.e., via one of the primary proximate determinants of fertility), lactational infecundability, as modeled by Bongaarts (1978), and nutritional flows to women (which also relates to lactational infecundability, both directly by affecting milk production, and indirectly via women's workloads and energy balance). Under conditions where all 'wealth' comes in the form of energy from food, the accumulation of embodied capital is automatically translated into offspring, via female reproductive physiology. Kaplan et al. (2002) demonstrate that, when applied to hunter-gatherer lifeways, a model of embodied capital maximizes fitness.

Kaplan (1996; Kaplan et al., 2002) then considers contemporary societies, asking what factors would prompt a decline in fertility. One key factor is argued to be the emergence of extrasomatic forms of wealth (livestock, land, money), which breaks the link between parental investment and reproductive physiology. Whereas somatic wealth translates directly into fertility under hunter-gatherer conditions, because our evolved psychological preferences for resources under such conditions amount solely to those energetic resources that can be stored in the body, this same psychology leads us astray when extrasomatic wealth can also be accumulated; this is argued to be because there is no mechanism in place to ensure that this evolutionary novel form of wealth is translated into fertility (Kaplan et al., 2002). Indeed, the model shows that a preference for wealth maximization under these conditions can lead to maladaptive behavior of the kind identified by Becker (1960), where people favor income and consumer goods over children. It should be noted, however, that this argument rests on the assumption that we possess an evolved psychological mechanism of this nature, and that this has failed to keep pace with the cultural and economic changes that characterize industrial societies. More stringent tests are needed to demonstrate the existence and operation of such a mechanism, and show that it is unable to respond adequately to wealth-accumulating environments.

A second factor linked to declining fertility that Kaplan (1996; Kaplan et al., 2002) identifies relates to the changing payoffs of embodied capital investments in low-mortality industrial contexts compared to hunter-gatherer lifeways. In essence, low mortality, combined with increased demands for skilled workers, is argued to result in high expected returns from investing heavily in the embodied capital of each child produced (in terms of educating them to the level required to be competitive in such societies). As family wealth is finite, this requires a trade-off between producing a few highly educated, skilled offspring versus a larger number of poorly educated, less-skilled, less-competitive offspring. Under such conditions, people calculate the number of children they believe they can afford to raise, given their wealth, and regulate their fertility accordingly.

Thus, even though in physiological terms, individuals have the capacity to produce large numbers of children, Kaplan (1996; Kaplan et al., 2002) argues that the high costs of providing them with the kinds of embodied capital needed to compete successfully, combined with the distorting effects of extrasomatic wealth on people's preferences, result in very few children actually being produced. Kaplan et al. (2002) therefore suggest that, although individuals trade off offspring quality against quantity exactly as predicted by LHT, in modern environments this number falls below that required to maximize fitness.

Yet, when extrasomatic energy use is considered (i.e., the use of fossil fuels, and the infrastructure required to support these) modern human fertility has decreased relative to other primate species exactly as expected given our increased energy consumption (Burnside et al., 2012). More specifically, human birthrates display an allometric relationship with energy expenditure in line with that of other species, and this pattern is quantitatively consistent with the life history patterns seen in other mammals, as well as nonindustrial human populations. As an example, the energy use of a woman in the United States nowadays is equivalent to the metabolic rate of a hypothetical 30 000 kg primate, and fertility rate is similar to what one would expect for a primate of this size (Burnside et al., 2012). Although industrial societies have larger more complex networks that can distribute larger amounts of energy resources than nonindustrial societies (pipelines, power grids, railway and road networks), such networks also require longer transport times for such energy to be distributed, and experience greater infrastructure costs. Burnside et al. (2012) therefore argue that, when all the relevant costs of raising children are factored into the equation, it becomes clear that offspring in industrial societies require exceptionally heavy investment, resulting in an extreme trade-off whereby very few children of very high quality are produced. When viewed from this perspective, the issue of whether people are behaving maladaptively and failing to maximize fitness remains open.

Indeed, there are a number of mathematical models demonstrating that small families may maximize fitness. Mace (1998) shows that when inherited resources contribute strongly to fitness and the cost of raising children is high, very low levels of fertility can be favored. Moreover, reducing fertility may enhance fitness in the far future if 'snowballing' effects of wealth are possible such that income generates income (Hill and Reeve, 2004), when long-term reproductive outcomes are more predictable (Winterhalder and Leslie, 2002) or when population dynamics are taken into account, such as (environmental) calamities resulting in demographic crashes (Boone and Kessler, 1999).

The most convincing empirical study to date, however, suggests that lowering fertility may not, in fact, maximize long-term fitness. Using data on over 14 000 Swedes born in the early 1900s, Goodman et al. (2012) showed that low fertility increased descendants' socioeconomic position (in line with the quantity–quality trade-off), but decreased their biological fitness. As the period covered by this study includes many demographic and societal changes that individuals had never encountered before, it remains possible that people's decision-making was in a state of flux, leading to maladaptive behavior in response to these changeable circumstances. Indeed, people's ability to track and respond to risk is not perfect (Winterhalder and Leslie, 2002). Although highly speculative, it seems possible that, as shifts in mortality schedules and costs of raising children may now have stabilized, people may be able to make more accurate assessments of environmental circumstances, and adjust their fertility in ways that ultimately contribute to higher fitness. Recent forecasts indicating that fertility may again be on the rise in some European nations (Myrskylä et al., 2013) can perhaps be interpreted in just this fashion.

Conclusion

The social sciences have been very successful in identifying the relevant determinants of fertility, both within and between populations (Balbo et al., 2013), but these determinants are often studied in a disjunct fashion. A life history approach posits that reproduction is inextricably linked to other processes

in the life cycle of an animal, with ecology determining how energy is allocated to these processes, and it thus argues that all aspects of the life cycle should be studied in tandem. The currency that is optimized in LHT, and more generally evolutionary theory, is the number of surviving offspring (or even the number of surviving offspring that contribute to the next generation). Incorporating such measures into studies of fertility, in addition to the number of children born per woman, would bring evolutionary and social science approaches closer together. A union of these approaches will augment and enhance our understanding of fertility, while placing humans in their appropriate evolutionary context, allowing us to make direct comparisons with other species.

See also: Adaptation, Fitness, and Evolution; Demographic Measurement: General Issues and Measures of Fertility; Evolutionary Demography: A Darwinian Renaissance in Demography; Evolutionary Demography: A Synthesis of Two Population Sciences; Fertility Theory: Embodied-Capital Theory of Life History Evolution; Fertility Theory: Theory of Intergenerational Wealth Flows; Fertility Theory; Historical Demography; Human Behavioral Ecology; Life Course in History; Life Table; Lifespan Development: Evolutionary Perspectives; Reproduction and Cultural Anthropology.

Bibliography

Balbo, N., Billari, F.C., Mills, M., 2013. Fertility in advanced societies: a review of research. European Journal of Population 29 (1), 1–38.

Becker, G.S., 1960. An economic analysis of fertility. In: Demographic and Economic Change in Developed Countries. National Bureau of Economic Research, Inc, pp. 209–240.

Bongaarts, J., 1978. A framework for analyzing the proximate determinants of fertility. Population and Development Review 4 (1), 105–132.

Boone, J.L., Kessler, K.L., 1999. More status or more children? Social status, fertility reduction, and long-term fitness. Evolution and Human Behavior 20 (4), 257–277.

Burnside, W.R., et al., 2012. Human macroecology: linking pattern and process in big-picture human ecology. Biological Reviews 87 (1), 194–208.

Dijkstra, C., et al., 1990. Brood size manipulations in the kestrel (*Falco tinnunculus*): effects on offspring and parent survival. Journal of Animal Ecology 59 (1), 269–285.

Ellison, P.T., 2003. Energetics and reproductive effort. American Journal of Human Biology 15 (3), 342–351.

Gagnon, A., et al., 2009. Is there a trade-off between fertility and longevity? A comparative study of women from three large historical databases accounting for mortality selection. American Journal of Human Biology 21 (4), 533–540.

Goodman, A., Koupil, I., Lawson, D.W., 2012. Low fertility increases descendant socioeconomic position but reduces long-term fitness in a modern post-industrial society. Proceedings of the Royal Society B: Biological Sciences 279 (1746), 4342–4351.

Hill, S.E., Reeve, H.K., 2004. Low fertility in humans as the evolutionary outcome of snowballing resource games. Behavioral Ecology 16 (2), 398–402.

Hobcraft, J., McDonald, J.W., Rutstein, S., 1983. Child-spacing effects on infant and early child mortality. Population Index 49 (4), 585–618.

Howie, P.W., McNeilly, A.S., 1982. Effect of breast-feeding patterns on human birth intervals. Journal of Reproduction and Fertility 65 (2), 545–557.

Jones, J.H., 2011. Primates and the evolution of long, slow life histories. Current Biologyy 21 (18), R708–R717.

Kaplan, H., 1996. A theory of fertility and parental investment in traditional and modern human societies. American Journal of Physical Anthropology 101 (23), 91–135.

Kaplan, H., et al., 2002. Evolutionary approach to below replacement fertility. American Journal of Human Biology 14 (2), 233–256.

Koenig, M.A., et al., 1990. Birth intervals and childhood mortality in rural Bangladesh. Demography 27 (2), 251–265.

Lawson, D.W., Alvergne, A., Gibson, M.A., 2012. The life-history trade-off between fertility and child survival. Proceedings of the Royal Society B: Biological Sciences 279 (1748), 4755–4764.

Lawson, D.W., Mace, R., 2011. Parental investment and the optimization of human family size. Philosophical Transactions of the Royal Society B: Biological Sciences 366 (1563), 333–343.

Lummaa, V., 2007. Life-history theory, longevity and reproduction in humans. In: Dunbar, R.I.M., Barrett, L. (Eds.), Oxford Handbook of Evolutionary Psychology. Oxford University Press, Oxford, pp. 397–413.

Lummaa, V., et al., 1998. Natural selection on human twinning. Nature 394 (6693), 533–534.

Mace, R., 1998. The coevolution of human fertility and wealth inheritance strategies. Philosophical transactions of the Royal Society B: Biological sciences 353 (1367), 389–397.

Myrskylä, M., Kohler, H.-P., Billari, F.C., 2009. Advances in development reverse fertility declines. Nature 460 (7256), 741–743.

Myrskylä, M., Goldstein, J.R., Cheng, Y.A., 2013. New cohort fertility forecasts for the developed world: rises, falls, and reversals. Population and Development Review 39 (1), 31–56.

Nettle, D., 2011. Flexibility in reproductive timing in human females: integrating ultimate and proximate explanations. Philosophical Transactions of the Royal Society B: Biological Sciences 366 (1563), 357–365.

Noordwijk, A.J. van, Jong, G. de, 1986. Acquisition and allocation of resources: their influence on variation in life history tactics. The American Naturalist 128 (1), 137–142.

Partridge, L., Harvey, P.H., 1988. The ecological context of life history evolution. Science 241 (4872), 1449–1455.

Roff, D.A., 1992. The Evolution of Life Histories: Theory and Analysis. Chapman & Hall, New York.

Sanderson, W.C., 2013. The End of World Population Growth in the 21st Century: New Challenges for Human Capital Formation and Sustainable Development. Routledge, Abingdon.

Sear, R., et al., 2001. The fitness of twin mothers: evidence from rural Gambia. Journal of Evolutionary Biology 14 (3), 433–443.

Shenk, M.K., et al., 2013. A model comparison approach shows stronger support for economic models of fertility decline. Proceedings of the National Academy of Sciences of the United States of America 110 (20), 8045–8050.

Stearns, S.C., 1992. The Evolution of Life Histories. Oxford University Press, Oxford.

Stearns, S.C., Koella, J.C., 1986. The evolution of phenotypic plasticity in life-history traits: predictions of reaction norms for age and size at maturity. Evolution 40 (5), 893–913.

Walker, R., et al., 2006. Growth rates and life histories in twenty-two small-scale societies. American Journal of Human Biology 18 (3), 295–311.

Winterhalder, B., Leslie, P., 2002. Risk-sensitive fertility – the variance compensation hypothesis. Evolution and Human Behavior 23 (1), 59–82.

Fertility Transition: China

Wang Feng, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

China's fertility transition followed a long historical tradition as Chinese families resorted to a variety of mechanisms to regulate the size and sex composition of their offspring long before the modern transition. The fertility transition was highly compressed in time, completed mostly in only a decade time. While China's transition was facilitated by a strong government birth control program, most of it was completed when China launched the extreme birth control policy known as the one-child. Such a compressed fertility transition, followed by more than three decades enforcement of the One-Child policy, has led to a series of long-term social and economic consequences.

Between 1970 and 1980, in one decade's time, China's national fertility level was cut in half. Following a decade of fluctuating around the level of two children per couple in the 1980s, fertility level in China dropped to below the replacement level and has stayed at that low level for over decades since the early 1990s. China's fertility transition, most notably in the 1970s, is the most drastic fertility decline observed among all human populations of significant size, aside from wars, epidemics, and famines. As the largest population in the world, China's fertility transition has also had the greatest impact on the slowing down of the world's population growth. This article reviews the process, historical background, the roles of the government programs and other factors, and the consequences of this transition.

Process of Fertility Transition

At the national level, China's fertility transition began in the early 1970s. In 1970, the total fertility rate was 5.8, close to the highest national level recorded for China. In the following decade, a most drastic reduction in fertility occurred. China's total fertility rate declined to 3.6 by 1975 and to 2.3 by 1980. The fertility level fluctuated between 2 and 3 throughout the 1980s. For the first time, it dropped to below the replacement level of 2.1 in the early 1990s and soon afterward, to a level of 1.5 and has stayed around such a level ever since. China, in other words, completed its fertility transition in two decades, from 1970 to 1990.

In China's urban areas and among the more educated subgroups of the population, the fertility transition began much earlier, dating back to the early 1950s (Lavely and Freedman, 1990). Since 1957, urban Chinese fertility declined almost consistently, with only one exception when fertility rebounded in 1963 after a plunge caused by the Great Leap Forward famine. By 1966, urban total fertility dropped to 3.1, and since 1974, it has never been higher than 2. In Shanghai, China's most cosmopolitan city, the total fertility rate dropped to below the replacement level (of 2.1 births per woman) for the first time in 1964, and to below 1 in 1975. By 1970, on the eve of China's first national family planning program, use of contraception and abortion among young women had already risen to 35 and 20%, respectively, in cities. Even in the countryside, over 15% of women at age 35 had already used a modern contraceptive method, and 7% had had at least one induced abortion.

In China's rural areas, where close to 80% of the population resided in the 1970s, fertility transition did not begin until much later, around 1970. In fact, during most years between 1950 and 1970, rural fertility showed signs of increase under the collective commune system. Fertility generally stayed at a level of above 6, with the exception of the Great Leap Famine years of 1959–61, when rural fertility was nearly halved. Consistent fertility decline in rural China began after 1970. The total fertility rate dropped to 5 by 1973, to 4 by 1975, and to 3 by 1979. During the 1980s, the total fertility rate fluctuated mostly between 2 and 3. In the early 1990s, it dropped to the below replacement level as was the case nationally.

There were also regional variations in the fertility transition in rural China. In provinces with a more advanced economy and a stronger tradition of population control, such as the lower Yangtze provinces of Jiangsu and Zhejiang, rural fertility declined earlier. In 1973, when the national level was 5, the total fertility rate was only 2.8 and 3.5 in these two provinces. By comparison, rural fertility was much above the national average in economically less developed provinces: it was 7.4 in Guizhou and 6.5 in Gansu (Coale and Chen, 1987).

Historical Background

China's large population size, along with a strong cultural preference for sons and a practice of early and universal marriage among Chinese women, have often led many to believe that the Chinese did not attempt to control fertility before China's contemporary fertility transition. Malthus, for instance, stated that "To account for this population, it will not be necessary to recur to the supposition of Montesquieu, that the climate of China is in any peculiar manner favourable to the production of children, and that the women are more prolific than in any other part of the world." Among the causes he identified is "The extraordinary encouragements that have been given to marriage, which have caused the immense produce of the country to be divided into very small shares, and have, consequently, rendered China more populous, in proportion to its means of subsistence, than perhaps any other country in the world" (1872/1971: p. 101, 103).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31077-7

Contrary to such earlier observations, many Chinese controlled their demographic outcomes in the past by relying on multiple mechanisms. The best-documented and most widely cited examples pertain to infanticide, especially female infanticide (Lee, 1981; Lee et al., 1994; Lee and Campbell, 1997). Chinese families killed some of their newborns in order to achieve their desired goals of the number and sex composition of the offspring. Poor harvests and economic difficulties also exacerbated the practice of infanticide.

Moreover, studies of Chinese historical population data also reveal that the Chinese fertility within marriages was significantly lower than their pretransition European counterparts (Barclay et al., 1976; Lee and Wang, 1999). Such a lower marital fertility level was achieved by a much longer interval between marriage and first birth, longer interbirth intervals, and earlier stopping. The same rationales that motivated infanticide also operated for fertility control behavior. Chinese parents sped up or slowed down reproduction according to the number and sex composition of existing children, and to their familial economic circumstances (Wang et al., 1995; Lee and Campbell, 1997; Zhao, 1997; Lavely, 2007). Fertility control in the past relied mainly on sexual abstinence, prolonged breast-feeding, and to a lesser extent, abortion and traditional contraceptives (Lee and Wang, 1999).

These behaviors suggest that controlling births is not a novel idea to the Chinese population. The contemporary fertility transition did not require new ideas of regulating population behavior, but rather a new environment in which fertility control became necessary and feasible. In contrast to traditional birth control, which relied on both prenatal and postnatal methods, the modern fertility transition is mostly prenatal, facilitated by new contraceptive and abortive technologies. Historical traditions of demographic control not only lay the foundation for contemporary fertility transition in mainland China, they also help explain the similarities in the pace and patterns of fertility transitions observed in several populations. In spite of having different political and economic systems, mainland China, Taiwan, Hong Kong, and Singapore all exhibited highly similar patterns of fertility transition, partly due to a shared cultural tradition (Coale and Freedman, 1993). The Chinese fertility transition process is therefore fundamentally a continuous process, extending a long tradition of demographic control.

Government Programs and Other Factors

The most prominent feature of China's fertility transition is the role played by government family planning programs. The most drastic fertility decline coincided with the launching of forceful government policies and programs of population control. Development of government policy can be traced to the early 1950s. Early concerns of rapid population growth and the need for a government family planning program emerged in the early 1950s, when China's 1953 census confirmed its large population size. Abortion was legalized and contraceptives began to be provided on a limited scale. Unlike later programs, which were motivated by government birth control goals, these early provisions were largely to meet an increasing demand among the population for birth control. These early efforts however were soon entangled in the ideological debates of whether China needed to control its population growth and they were interrupted by a series of political movements. As a result, it was not until the 1960s that the government seriously promoted birth control in urban and densely populated rural areas. Overall, government family planning programs at this early stage may have facilitated the early decline, but was not the main driving force.

Government intervention in reproduction reached a new phase in the early 1970s when the policy of *wan* (later marriage), *xi* (longer birth intervals), and *shao* (fewer births) was formed. Although largely accepted by most people on a voluntary basis, the implementation of the policy became increasingly intensified and regularized. The government utilized its control over the media to promote birth control as a national policy and mobilized a whole system of vertically controlled work and political organizations to implement the policy. Special government organizations in charge of birth control were also organized and expanded during this period. Official marriage age was stipulated to be 25 and 23 for urban males and females, and 22 and 20 for rural youths. Two children were considered the proper size for an urban couple and three for most rural couples. Under this program, Chinese ages at marriage rose sharply and fertility declined swiftly. Nationally, the female mean age at first marriage rose from 21 in 1970 to 23 in 1979, and the fertility level was halved (Banister, 1987).

In spite of this rapid fertility decline of the 1970s, the Chinese government was not content. Instead, motivated by a keen desire to increase quickly the per capita living standard within a short period of time, it launched an even more ambitious and a much more controversial birth control program in 1980. Known as the One-Child policy, it required each couple to have only one child, and was an emergency measure to further reduce China's population growth rate. Under the pressure of such an unprecedented and extreme policy, officials and especially birth control cadres resorted to all methods, including physical abuse in some locales, to meet the goals established by various levels of government. Women with two children, sometimes even only one, were required to have an intrauterine contraceptive device inserted, and women with three or more births were required to undergo sterilization. Pregnancies not meeting the requirement of birth control had to be aborted. Birth control campaigns were also launched in the 1980s, resulting in sharp increases in the numbers of abortions and sterilization (Hardee-Cleaveland and Banister, 1988).

This extreme policy met strong resistance from the Chinese population from the very beginning, especially from the vast population in the countryside. Whereas the earlier policies of birth control were by and large acceptable to Chinese farmers, this new extreme measure required them not only to forego their traditional perceptions of family and kinship, but also to reduce or even to remove their reliance on children for labor and old-age support. Popular resistance led to relaxation in the implementation in the Chinese countryside and to changes in the policy. As a result, Chinese couples in cities (accounting for about 20% of the population in the late 1970s) obeyed the One-Child policy, whereas most rural couples had two and some of them more than two children (Feeney and Yuan, 1994; Gu et al., 2007).

The draconian government policy may have masked the roles of many other factors underlying China's fertility transition process. In addition to the historical tradition of demographic regulation, many contemporary social and economic changes contributed to China's fertility transition. These changes include a drastic decline in mortality, especially infant mortality, the rise in nonagricultural and nonfamilial employment, rapid reduction in female illiteracy, and increased provision of social welfare, especially to the urban population (Peng, 1989). A changing social and economic context thus produced rapid adaptation to changes in childbearing in the Chinese mainland just as it did elsewhere in East Asia (Greenhalgh, 1988). At both the regional and the individual level, fertility differentials associated with employment type, residence, education, and income existed since as early as the mid-1950s (Lavely and Freedman, 1990). Factors other than government policy similarly have been shown to be crucial for China's below replacement fertility after the 1990s (Cai, 2010).

Consequences

With its speed and intensity, the Chinese fertility transition has also produced several distinctive consequences. The first is a unique social consequence, with a whole generation of single children and of parents with only one child. Among couples who would have completed their reproductive lives by the year 2010, 40% will have only one child. Among urban couples, for whom the One-Child policy was more effectively implemented, nearly 90% end up with only one child (Wang et al., 2012). Children grown up in these families have no siblings and in the future their children will not have any uncles, aunts, and cousins, all of whom are important parts of the traditional Chinese kinship network. Moreover, there will be a severe shortage in familial support at the old age. At least half of all parents of these single children will not be able to coreside with their married children in the future, when emotional and functional support from children is needed in old age.

Second, China's rapid fertility transition, with its resulting changes in population age structure, has had a profound impact on China's economy. Demographic dividend, the economic consequence of a favorable age structure, contributed at least 15% of China's historical economic boom in the closing days of the twentieth century (Wang and Mason, 2008). As that dividend diminishes, China faces the prospect of a rapid decline in the size of young labor force and an accelerated process of population aging. In less than 20 years and by 2030, China's ratio of working age population aged 20–59 to retirement age population aged 60 and over changes from 5:1 to 2:1. Even with China's record economic growth, its population has grown old before getting rich.

Lastly, China's rapid fertility transition has also led to an acutely unbalanced sex ratio between male and female births. Sex ratio at birth rose steadily, from 107 boys to 100 girls in 1981 to 115 boys to 100 girls in 1991, and to close to 120 boys to 100 girls in the first decade of the twenty-first century. The sex ratio among births of higher orders rose even more. Among second births, the recorded sex ratio rose from 107 in 1981 to 116 in 1985, and to 130 in 1993 (Gu and Roy, 1995). Much of this increased sex ratio is due to the widespread use of sex-selective abortion. Whereas the increase in the sex ratio may be attributable in part to its government program, the fact that similar increases in sex ratio are also observed in several other East Asian populations suggests this may at least be a temporary by-product of the fertility transition process in these societies. In China, such a lopsided sex ratio will have long lasting effect on marriage patterns and on population growth in the long run.

See also: Fertility Theory; Fertility Transition: Cultural Explanations; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Fertility Transition: Latin America and the Caribbean; Fertility Transition: Middle East and North Africa; Fertility Transition: South Asia; Fertility Transition: Southeast Asia; Fertility Transition: Sub-Saharan Africa; Fertility and Culture: Anthropological Insights.

Bibliography

Banister, Judith, 1987. China's Changing Population. Stanford University Press, Stanford.
Barclay, George W., Coale, Ansley J., Stoto, Michael A., Trussell, James, 1976. A reassessment of the demography of traditional rural China. Population Index 42, 606–635.
Cai, Yong, 2010. China's below-replacement fertility: government policy or socio-economic development? Population and Development Review 36 (3), 419–440.
Coale, Ansley J., Chen, Shengli, 1987. Basic Data on Fertility in the Provinces of China, 1942–1982. East-West Population Institute Paper Series, Honolulu.
Coale, Ansley J., Freedman, Ronald, 1993. 'Similarities in the fertility transition in China and three other East Asian populations'. In: Leete, Richard, Alam, Iqbal (Eds.), The Revolution in Asian Fertility: Dimensions, Causes, and Implications. Clarendon Press, Oxford.
Feeney, Griffith, Yuan, Jianhua, 1994. Below replacement fertility in China? A close look at recent evidence. Population Studies 48, 381–394.
Greenhalgh, Susan, 1988. Fertility as mobility: sinic transitions. Population and Development Review 14, 629–674.
Gu, Baochang, Roy, Krishtina, 1995. Sex ratio at birth in China, with reference to other areas in East Asia: what we know. Asia-Pacific Population Journal 10, 17–42.
Gu, Baochang, Feng, Wang, Zhigang, Guo, Erli, Zhang, 2007. China's local and national fertility policies at the end of the twentieth century. Population and Development Review 33 (1), 129–147.
Hardee-Cleaveland, Karen, Banister, Judith, 1988. Fertility policy and implementation in China, 1986–88. Population and Development Review 14, 245–286.
Lavely, William, 2007. Sex, breastfeeding, and marital fertility in pretransition China. Population and Development Review 33 (2), 289–320.
Lavely, William, Freedman, Ronald, 1990. The origins of Chinese fertility decline. Demography 27 (3), 89–116.
Lee, Bernice, 1981. Infanticide in China. In: Guisso, Richard, Johannson, Stanley (Eds.), Women in China. Philo Press, Youngstown, NJ, pp. 163–177.
Lee, James, Campbell, Cameron, 1997. Fate and Fortune in Rural China: Social Organization and Population Behavior in Liaoning, 1774–1873. Cambridge University Press, Cambridge.
Lee, James, Wang, Feng, 1999. One Quarter of Humanity, Malthusian Mythology and Chinese Realities 1700–2000. Harvard University Press, Cambridge, MA.
Malthus, Thomas R., 1872/1971. An Essay on the Principle of Population, seventh ed. Augustus M. Kelley, New York.
Peng, Xizhe, March 1989. Major determinants of China's fertility transition. The China Quarterly 117, 1–37.
Wang, Feng, Lee, James, Campbell, Cameron, 1995. Marital fertility control among the late imperial Chinese nobility: implications for two types of preventive check. Population Studies 49, 383–400.
Wang, Feng, Mason, Andrew, 2008. The demographic factor in China's transitions. In: Brant, Loren, Rawski, Thomas (Eds.), China's Great Economic Transformations. Cambridge University Press, pp. 136–166.
Wang, Feng, Cai, Yong, Gu, Baochang, 2012. Population, policy, and politics: how will history remember China's one-child policy? Population and Development Review 38 (Suppl.), 115–129.
Zhao, Zhongwei, 1997. Deliberate birth control under a high-fertility regime: reproductive behavior in China before 1970. Population and Development Review 23, 729–768.

Fertility Transition: Cultural Explanations

Jennifer Johnson-Hanks, University of California, Berkeley, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S. Cotts Watkins, volume 8, pp. 5572–5575, © 2001, Elsevier Ltd.

Abstract

'Culture' refers to shared systems of meaning or value, and their instantiation in material forms such as documents, artifacts, and institutions. Because 'culture' can refer to many things, explanations of fertility transition we call 'cultural' are a diverse set: some stress new contraceptive technologies, others social structures and institutions, others the movement of ideas or information between people, and still others the values and attitudes of individual social actors. This article provides a terse history of cultural approaches to fertility transition over the past seven decades.

'Fertility transition' refers to one element of demographic transition, the process by which mortality and fertility have fallen from approximate equilibrium at moderate or high levels to approximate equilibrium at low levels in most human populations over the past 200 years. As such, fertility transition is a special case of fertility change, characterized in its ideal-typical form as sustained decline in marital fertility following mortality decline with some lag, in association with increasing modernization, and with population equilibrium at the beginning and end of the process. Other kinds of fertility change, such as increasing proportion of births to unmarried women or changes in birth timing, have in recent years also been called 'fertility transitions,' although this usage remains contested. For the purposes of this article, the contestation is not important, as fertility transition narrowly understood and fertility change in a broader sense are amenable to the same kinds of cultural explanations.

'Culture' refers to shared systems of meaning or value, and their instantiation in material forms such as documents, artifacts, and institutions. Thus religion, gender, and even some kinds of politics are forms of culture, and the considerable majority of contemporary explanations of fertility are at least partially cultural ones. However, there is a second sense of the phrase 'cultural explanations of fertility transitions,' grounded in the intellectual history of population studies. Since approximately the 1970s, explanations of fertility decline have often been categorized into two opposing types – 'economic' and 'cultural' – where 'cultural' is defined almost entirely in opposition to its more coherent converse. Economic explanations posit that fertility outcomes are the product of individuals' utility maximization, usually about child numbers, so that fertility declines in response to the changing relative costs and benefits of children. What we have called cultural explanations, by contrast, are a diverse set. Under many other circumstances, these various explanations would not be classified as belonging to a single type: some stress technology, others social structures, others the movement of ideas or information between people, and still others cognitive states of individual social actors (see commentary in Mason, 1992: 1). However, they share a common opposition to the premise that fertility transition can be sufficiently explained by economics.

Cultural explanations of fertility transition include critiques of each aspect of the economic model. Some begin with the premise that the very idea of optimizing is a cultural construct that exists in certain times and places but not others. Others propose that individuals may indeed all be optimizing, but under such radically different circumstances as to make that claim irrelevant: if people's reproductive behavior is aimed at optimizing their relationships with ancestor-spirits, for example, then most economic models of fertility are going to be misspecified. Still others propose that reproductive behavior is so profoundly shaped by the flow of information, evaluation, and influence through social interaction that these flows must themselves be central to the analysis of fertility. Thus, some explanations we call cultural are in direct conflict with economic explanations – arguing against the premise of reproductive optimizing itself – while others are not necessarily inconsistent with the idea of means-ends rational decision-making about reproduction, but differ in emphasis. If people are seen as optimizing utility with culturally defined tastes under a complex set of cultural constraints, there can be an intellectual division of labor between those trying to understand the tastes and constraints on the one hand, and those trying to model the decision on the other.

Fertility transition sustains these competing explanatory models because the empirical corpus is complex and multiply construable. In favor of an economic perspective, nearly all data sets show a negative association between economic development and fertility. In favor of cultural explanations, that relation is highly variable across time and space in slope, strength, and the presence or absence of thresholds. Cultural explanations often draw attention to this diversity, to the ways that the relationship itself has changed over time, and to cases of asynchronous change: where fertility fell much sooner or much later than would have been predicted by economic drivers alone. Bongaarts and Watkins articulate the now common view: "variation in widely available standard measures of development … provides only a partial explanation of fertility differences between countries" (1996: 642), and the remainder of the explanation lies in social interaction – one of the processes commonly called 'cultural.'

Economics and Culture United in Institutions

Many early descriptions of fertility transition integrate economic and cultural perspectives. Although commonly cited

as attributing fertility decline to economic development alone, Notestein (1945) has a richly nuanced interpretation of fertility, which gives an important role to cultural change. He argues that fertility decline followed from the "growing individualism and rising levels of popular aspiration developed in urban industrial living" through which "the individual came to depend less and less on the status of his family for his place among his fellows" (1945: 40). Notestein saw the changes on fertility caused by economic development to occur largely *through* cultural transformations, both ideational (values and tastes) and institutional (constitution of the family and household, assignment to social roles).

The 1950s saw a substantial increase in studies into the relationships between culture, institutions, and demographic outcomes as anthropologists, sociologists, and demographers worked together to make sense of population processes in poor countries. Many of these authors made explicit links between the historical fertility transition in Europe and then-incipient fertility transitions in poor countries. Kingsley Davis and Judith Blake (1956) proposed a framework for thinking about the relationship between social structures – such as norms against intercourse during breastfeeding, or practices linking first marriage to inheritance – and fertility, through the now-famous intermediate fertility variables. While Bongaarts' reformulation of this idea in 1978 significantly downplayed the importance of social organization, the original Davis and Blake paper is a sociological argument about how roles, norms, and institutions influence fertility regardless of individual fertility intentions.

The functionalist approach to culture and fertility found in Davis and Blake was widespread in the 1950s. Among the prototypical works of this period is Frank Lorimer's 1954 book *Culture and Human Fertility*, published by the IUSSP in conjunction with UNESCO, including a tepid introduction by Frank Notestein, and contributions by anthropologists Meyer Fortes and Audrey Richards among others. The book is important less for its conclusions than for its approach, which is to treat fertility as an integral part of social systems, intimately linked to institutions of kinship and family, land tenure and inheritance, religion, education, and cultural attitudes. Most of the book focused on contemporary poor countries. However, in Chapter VI Lorimer focuses on the demographic transition in England and France, asking why the decline of fertility was so early in France and so late in England. He suggests that the cultural ideas introduced by the French revolution were critical to fertility change in that country, and concludes that "the record is clear on two negative points of theoretical importance: first, a trend toward increased control of fertility is not necessarily dependent on large-scale industrialization; secondly, the rapid introduction of mechanical industries into a previously non-industrial society does not automatically bring a trend toward increased control of fertility" (1954: 217). Whereas Notestein argued that economic changes would drive down fertility through changes in cultural practices, institutions, and ideas, a decade later Lorimer was arguing that economic changes were neither necessary nor sufficient for fertility declines, and suggesting that changing cultural ideas and cultural institutions could cause fertility to fall.

For Lorimer as for Davis and Blake, 'cultural' explanations of fertility depend very heavily on institutions – particularly those of kinship and family, but also broader social institutional arrangements, such as those concerning wage work, political participation, and the church. In this way, their arguments are largely consonant with claims of universal economic rationality, but where culture sets both the preferences and the constraints within which that rationality applies. This same view is seen several decades later in the influential work of John Caldwell.

Australian Caldwell had his first academic position at the University of Ghana (1962–64), and Ghanaian family life dramatically shaped his thinking about fertility. In the paper 'The Cultural Context of High Fertility in sub-Saharan Africa,' Pat and John Caldwell argue that African cultural systems are "seamless wholes … molded by stress on ancestry and descent" (1987: 410). For Caldwell, culture is the unity of institutions and ideas that together determine what individuals will try to achieve, and what constraints they will face. He argued therefore that "in societies of every type and stage of development, fertility behavior is rational, and fertility is high or low as a result of economic benefit" but that "[w]hether high or low fertility is economically rational is determined by social conditions" (1976: 355). These 'social conditions' include both institutions and values and attitudes, particularly concerning how children should be raised, and what obligations they have to their parents, which result in net wealth flows going either from parents to children (in low-fertility societies) or from children to parents (in high-fertility societies). Thus Caldwell's work continues the tradition set out by Notestein in which cultural and economic explanations of fertility transition are complementary and compatible, as people make choices that are rational within their own cultural context, and culture is largely defined by institutional forms.

Culture as (Potential) Interaction

A quite different view of culture emerged out of the European Fertility Project (EFP), started in 1963 by Ansley Coale at Princeton University. Coale sought to create a quantitative record of the European fertility transition, and to determine the temporal associations between social and economic circumstances and the timing of fertility decline across some 600 provinces in Europe. Collaborating with a large number of graduate students, Coale determined that the timing of fertility decline in relation to economic development varied considerably across the continent, but was geographically and linguistically clustered. Years later, Coale summarized the findings: "Adjacent provinces sharing the same language or ethnic origin were often alike in fertility even though different in educational levels, infant mortality, and the proportions living in cities. At the same time, adjacent provinces that had different cultural background … were not alike in fertility" (Coale, 1987: 21). That is, Coale and his students concluded that demographic change occurred through social interactions, facilitated by cultural similarity. The contribution of this remarkable work was not, as is often stated, to introduce demographers to the idea of culture. That had been done by Davis and Blake, Lorimer, and others in the 1950s. Rather, the key contribution of the EFP was to expand our understanding of culture in two ways. First, after the EFP, demographic understandings of fertility had to include the channels of information and influence along which changes in

ideas and values could flow. This lead in a relatively direct way to the contemporary foci on social networks and diffusion (*see* Fertility and Social Networks). Second, and more tentatively, several of Coale's students opened up ways of thinking about microlevel variation in ideas and values, whether through a focus on fertility decline forerunners (Livi-Bacci) or by pushing the implications of the empirical model to their theoretical conclusions (Watkins). This view of culture as plural and interactive is strikingly different from the bounded, stable, and institutional models of the functionalists of a decade before.

As Coale sought to make sense of the early findings of the EFP, he proposed a rethinking of demographic transition theory along cultural lines. In 1973, he presented three preconditions for a sustained decline in marital fertility. First, child numbers must enter into the 'calculus of conscious choice' – that is, parents must be able to entertain the idea that they could choose, even approximately, the number of children that they bear. If this precondition is not universally met, then it means that some parents or potential parents are not optimizing utility through their reproductive behavior, in striking contrast to the view of Davis or Caldwell. The second precondition states that lower child numbers must be perceived as advantageous by parents. Note here the key word 'perceived': while the objectively measured changing costs and benefits are surely related to parents' perceptions of the relative advantages of different family sizes, these two are not the same. Finally, Coale's third precondition states that parents must have access to an effective and acceptable means of fertility limitation. That is, some pregnancies arise because people either lack the information about how to prevent them or are unwilling to use available methods to prevent pregnancy. This is where technological innovation becomes part of a cultural explanation of fertility decline. In theory, any couple at any time in history could have used periodic or permanent abstinence or withdrawal to reduce fertility; Coale suggests that they either did not know how or were unwilling to accept such dramatic limitations on how and when they had sex.

Economists have argued that the burdens of a child so outweigh any potential burdens associated with fertility limitation that Coale's third precondition must always be met. But several demographers have shown that innovations in contraceptive technologies or knowledge do matter for fertility behavior, at least in some circumstances. For example, Shorter, Knodel, and van de Walle (1971) use data from the EFP to show that in a number of European countries, marital and nonmarital fertility declined in parallel. Since there were essentially no incentives favoring nonmarital fertility in Europe at this time, this suggests that: "the behavior of unmarried and married women was modified … as a result of obscure changes in the attitudes toward reproduction, and in the knowledge and acceptability of contraception and abortion" (Shorter et al., 1971: 393).

The EFP continued from 1963 until 1986. It produced to at least 10 books, dozens of articles, and nearly an entire generation of eminent demographers. Of the students of the EFP, none has been more influential on the topic of culture and fertility than Susan Cotts Watkins. Her 1991 monograph *Provinces into Nations* shows that from the middle of the nineteenth century to the middle of the twentieth, European fertility not only fell, but also became far less variable, particularly within national boundaries. She posits that fertility behavior is always influenced by social norms about how many children to have and when to have them, norms shared through social communication. However, Watkins argues, the mechanisms through which this social communication occurs have changed: in 1870, local gossip was the primary mechanism, whereas by 1960, newspapers and other national media dominated; as a result, the homogenizing effects of social communication on fertility are larger, and often correspond with the borders of nations. For Watkins, culture is thus centrally about interaction, both as a source of new ideas or information, and also as an "evaluative cloud of conversation" (Hammel, 1990: 467) which influences individual action. As a result, we cannot think about culture only through institutions, nor can we think about it as an individual trait. Culture is therefore something that *occurs* rather than something that endures, and it occurs particularly in communication.

In the mid-1990s, Watkins began work on contraceptive and fertility change in Africa, first in Kenya, then later in Malawi. This work continued many of the same themes developed in her European research, particularly concerning the importance for fertility of local systems of meaning and the role of social communication in their continuation and change, but with the advantage that she could collect data 'in real time' as new ideas diffused and changed (see Watkins, 2000). In order to show how social communication works, Watkins set out to collect rich data on networks over time: who talks to whom, about what, and to what effect? With Ann Swidler, she has also explored collecting data about natural conversations overheard in public (Watkins and Swidler, 2009). The projects in Kenya and Malawi have been important not only for the quality and quantity of data and publications, but also – like the EFP – as a training ground for a whole generation of talented and culturally oriented demographers, including Alex Weinreb, Nancy Luke, Jenny Trinitapoli, and many others.

Culture as Schemas, Attitudes, Values, and Beliefs

Whereas Watkins and others have focused on culture as interaction, seeing as central the conversations and social relationships through which ideas and influence flow, others have focused more on the *content* of culture, particularly as it is held in the mind: schemas, attitudes, values, and beliefs. An explicit study of the relation between fertility and these kinds of mental states first came to prominence in an attempt to understand reproductive differences and short-term change among married couples in the United States in the work of Princeton professors Norman Ryder and Charles Westoff on the National Fertility Studies (NFS) beginning in 1965, about the same time as Coale's EFP. The NFS framed reproduction as a decision made by a woman and her husband on the basis of a range of attitudes, values, and beliefs. Thousands of married American women were interviewed about their ideas and beliefs, some related closely to contraceptive practice (e.g., Should the federal government support birth control programs? Do you approve or disapprove of sterilization operations for men?), and others more broadly cultural, and particularly cultural-religious (e.g., What is your religion? How religious-minded are you?).

The NFS led to a considerable body of work on the values, attitudes, and beliefs related to fertility, both directly

(e.g., Ryder and Westoff, 1971; Westoff and Potvin, 1967), and indirectly (e.g., Frejka and Westoff, 2008; Simons, 1995). The early work in this genre is well-represented by Clifford, who argues that socioeconomic forces "exert their influence [on fertility] through the determination and reinforcement of certain value orientations which, in turn, have varying degrees of compatibility with particular fertility patterns" (1971: 38). In some later work, the causal link that Clifford asserts between socioeconomic conditions and values is attenuated, or even absent. Lesthaeghe, for example, argues that changes in values and ideas were largely responsible for the European fertility decline, and that these changes were largely autonomous of socioeconomic change (1983: 412), arising instead from the spread of secular individualism from the sphere of politics to the family. Fertility fell because people came to pursue "personal goals devoid of reference to a cohesive and overarching religious or philosophical construct" (1983: 415) after the social regulation of fertility gave "way to the principle of individual freedom of choice" (1983: 411). Lesthaeghe's more recent work on what he calls the "Second Demographic Transition" also proposes a central role for values in explaining fertility change (Lesthaeghe and Neidert, 2006). Similarly, Thorton's 'Developmental Idealism' stresses how fertility has changed in response to the spread and acceptance of Western values (2001). While Developmental Idealism has diffusion as its presumed mechanism, the model stresses the content of ideas – who values what – rather than the process through which those values move.

A highlight in the development of the study of fertility values is the edited volume *Dynamics of Values in Fertility Change* (Leete, 1999). The papers vary considerably in the degree to which they portray values as derived from socioeconomic circumstance, and also in the degree of causal power that they attribute to values. However, there is substantial agreement around three points noted in Casterline's terse conclusion. First "specific values are normally linked to other specific values and symbolic constructs, forming value systems and larger cultural systems" (Casterline, 1999: 358). It is therefore difficult to trace direct lines from individual values to particular reproductive behaviors. Second "all fertility behavior is value-driven, but fertility *change* is not necessarily driven by value *change*" (Casterline, 1999: 362, italics in the original), because, for example, fertility change may emerge as people continue to try to practice old values under new circumstances. Finally, "the values that have often been most decisive in determining fertility transition … have not explicitly referred to fertility, and, indeed, have sometimes been rather remotely related to fertility" (Casterline, 1999: 365). This is not only because values are often linked together, but also because family size is rarely a central value that organizes people's lives, whereas managing uncertainty, family solidarity, or belief in the church do so to a much greater degree. Taken together, these points suggest that understanding fertility behavior requires thinking about values as parts of cultural structures, rather than in isolation or only at the individual level.

Anthropology and Demography

Given that the discipline of anthropology created the contemporary concept of culture, it is noteworthy that anthropologists have featured so little in this article thus far. The reason is empirical. In the 1950s, Lorimer was in close conversation with structural-functional anthropologists in mapping out the social institutions, particularly of family and kinship, in which fertility is situated. But from the mid-1960s until the mid-1980s, even as many demographers were becoming increasingly interested in culture, anthropologists were largely absent from conversations about fertility and fertility transition. The conversation between demographers of fertility and sociocultural anthropologists was revitalized in the 1980s and blossomed in the 1990s with five major edited collections, at least a half-dozen major books, and several dozen consequential articles (the edited volumes are Basu and Aaby, 1998; Bledsoe and Pison, 1994; Greenhalgh, 1995; Kertzer and Fricke, 1997; Lesthaeghe, 1989; an especially important article is Hammel, 1990). The vitality of this period was intellectual, but also promoted by institutional resources: the Mellon Foundation provided substantial funding for training in anthropological demography starting in 1994, and the International Union for the Scientific Study of Population established a Committee on Anthropological Demography in 1998.

What distinguishes anthropological research on fertility in the 1990s is its characterization of reproduction as an activity grounded simultaneously in systems of meaning and systems of power. The word 'activity' here stresses that sex, contraception, abortion, and birth are things that people *do*, that they can do in new ways, that take time, and that are related to other, ongoing activities through which people constitute their lives differently in different times and places. To write about reproduction as a meaningful activity requires understanding how people think about births, but also how they think about a great many other things, such as gender, marriage, aging, work, religion, class, and progress. To write about reproduction as activity embedded in systems of power – whether primarily class (Kertzer, 1993; Schneider and Schneider, 1996), the state (Greenhalgh, 1994), or gender, marriage, and kinship (Bledsoe 1990, 1993) – entails understanding what about reproduction matters, to whom, and how access to reproduction is stratified.

This perspective is perhaps best accomplished in Bledsoe's 2002 monograph *Contingent Lives: Fertility, Time, and Aging in West Africa*. Here Bledsoe argues that high fertility in rural Gambia is an achievement for which women struggle and strategize, timing pregnancies in relation to their own bodily resources and social – particularly marital – circumstances in order to bear the full complement of children with which God endowed them. For Bledsie, 'pretransition' fertility is therefore not outside the 'calculus of conscious choice,' but its calculus is different than that assumed in most demographic texts, and therefore is hard to recognize. In the work of demographic anthropologists, culture became construed as a contested set of symbolic resources, constantly being reconfigured to suit new situations as people struggle to bear children when and how they want.

Recent Developments

Over the past 10–15 years, the concept of culture has become more central in the demographic literature on fertility than

perhaps ever before. Building on nearly seven decades of research, many in the present generation of fertility scholars agree that cultural forces and factors shape the kinds of reproductive contexts people face and the kinds of fertility behaviors that feel natural to them (see, e.g., Bernardi, 2003; Bernardi and Hutter, 2007; Fisher, 2006; Hayford and Trinitapoli, 2011; Hayford and Morgan, 2008; Johnson-Hanks, 2002, 2006; Obermeyer, 2000. Smith, 2001, 2004; Yeatman and Trinitapoli, 2008). The 'culture verses economics' framework that long organized debates about how fertility falls is often absent from this work. Also absent, often remarkably so, is the term 'culture' itself, particularly from titles, but often also as an explicit theoretical category. Trinitapoli and Weinreb's (2012) *Religion and AIDS in Africa*, for example, is a profoundly cultural book, which compellingly argues that to understand demographic outcomes we need to take religious practices and communities seriously, rather than assuming behavior on the basis of stated ideology, or studying only either individual-level behaviors or national differences. Examining how systems of meaning, practices, and resources are organized between individual, across communities and nations, Trinitapoli and Weinreb have culture in the center of their analysis, but tacitly: the term hardly appears. Similarly, Victor Agadjanian (e.g., 2001a,b) regularly writes about gender, religion, and ethnicity – all shared systems of meaning instantiated in material forms – but rarely uses the word 'culture.' Paradoxically, there is now so *much* culture in studies of fertility that the word has become almost obsolete.

Instead of arguing of the importance of culture as opposed to economic conditions, many contemporary scholars of fertility seek to understand the relationships among different parts of culture, and between these components of culture and reproduction. Whereas in 1970, demographers often said 'culture' to mean, 'something not explained by economic development,' today, the term 'culture' has a positive meaning, but its meaning is vast. It refers to shared systems of meaning or value, and the material forms that those meanings take, including institutions, performances, and artifacts. That is, culture structures social life, including reproductive practice. It is now clear that social networks, schemas and values, politics and social positions, institutions, religiosity, and the flow of ideas through media all matter for fertility. All of these things are part of culture. But they are also a disparate set, that work at different levels, in different ways, sometimes jointly and sometimes in opposition.

Moving forward, cultural explanations of fertility transition have to address the interplay of these multiple aspects. One attempt to do this is Johnson-Hanks et al. (2011), which proposes that we think about reproductive behavior as structured jointly by materials (objects, artifacts, institutions, and other kinds of resources) and by schemas (ideas that organize perception and action). Individual reproductive behavior in turn reinforces or alters the material and schematic structures in which it is embedded. This offers one way to integrate the material and schematic aspects of cultural systems without treating culture as a monolithic and inflexible whole. Simultaneous with new theoretical approaches, recent years have seen substantial innovations in the methods and data needed to study the relationships between structures of culture and fertility practices. For example, Margaret Frye (2013) has used card sort and optimal matching; Watkins and Swidler (2009) have used indigenous ethnographers; Bachrach (2013) has argued that the networked character of schemas means implies that new innovations in computational social science, dynamic systems modeling, 'big data,' and machine learning offer new opportunities for studying culture in relation to fertility. Since fertility itself is so deeply embedded in culture, explanations of fertility transition must also be cultural, and more than ever, we have a robust tool kit to take on that challenge.

See also: Fertility Transition: China; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Fertility Transition: Latin America and the Caribbean; Fertility Transition: Middle East and North Africa; Fertility Transition: South Asia; Fertility Transition: Southeast Asia; Fertility Transition: Sub-Saharan Africa; Fertility Trends in the Formerly Socialist Countries of Europe; Fertility and Culture: Anthropological Insights; Fertility and Social Networks.

Bibliography

Agadjanian, Victor, 2001a. Religion, social Milieu and the contraceptive revolution. Population Studies 55 (2), 135–148.

Agadjanian, Victor, 2001b. Negotiating through reproductive change: gendered social interaction and fertility regulation in Mozambique. Journal of Southern African Studies 27 (2), 291–309.

Bachrach, Christine, 10–14 April 2013. Presidential address. In: Annual Meeting of the Population Association of America. New Orleans.

Basu, Alaka M., Aaby, Peter (Eds.), 1998. The Methods and Uses of Anthropological Demography. Clarendon Press, Oxford.

Bernardi, Laura, Hutter, Inge, 2007. The anthropological demography of Europe. Demographic Research 17, 541–566.

Bernardi, Laura, 2003. Channels of social influence on reproduction. Research and Policy Review 22 (5–6), 527–555.

Bledsoe, Caroline, Pison, Gilles (Eds.), 1994. Nuptuality in Sub-Saharan Africa: Contemporary Anthropological and Demographic Perspectives. Clarendon Press, Oxford.

Bledsoe, Caroline, 1990. The politics of children: fosterage and the social management of fertility among the Mende of Sierra Leone. In: Handwerker, W.P. (Ed.), Births and Power: Social Change and the Politics of Reproduction. Westview Press, Boulder, CO, pp. 81–100.

Bledsoe, Caroline, 1993. The politics of Polygyny in Mende education and child fosterage transactions. In: Miller, B.D. (Ed.), Sex and Gender Hierarchies. Cambridge University Press, Cambridge, pp. 170–192.

Bledsoe, Caroline, 2002. Contingent Lives: Fertility, Time and Aging in West Africa. University of Chicago Press, Chicago.

Bongaarts, John, Watkins, Susan Cotts, 1996. Social interactions and contemporary fertility transitions. Population and Development Review 22 (4), 639–683.

Caldwell, John, Caldwell, Pat, 1987. The cultural context of high fertility in sub-Saharan africa. Population and Development Review 13 (3), 409–437.

Caldwell, John C., 1976. Toward a restatement of demographic transition theory. Population and Development Review 2 (3/4), 321–366.

Caldwell, John, 1982. Theory of Fertility Decline. Academic Press, London.

Casterline, John, 1999. Conclusion. In: Leete, R. (Ed.), Dynamics of Values in Fertility Change. Oxford University Press, Oxford, pp. 357–369.

Coale, Ansley, Watkins, Susan Cotts (Eds.), 1986. The Decline of Fertility in Europe. Princeton University Press, Princeton, NJ.

Coale, Ansley, 1987. Demographic transition. In: Eatwell, John, Milgate, Murray, Newman, Peter (Eds.), Social Economics. Palgrave Press.

Davis, Kingsley, Blake, Judith, 1956. Social structure and fertility: an analytic framework. Economic Development and Culture Change 4 (4), 190–211.

Fisher, Kate, 2006. Birth Control, Sex, and Marriage in Britain 1918–1960. Oxford University Press, Oxford.

Frejka, Thomas, Westoff, Charles F., 2008. Religion, religiousness and fertility in the US and in Europe. European Journal of Population 24, 5–31.

Frye, Margaret, 2013. Sexual Behavior and Scholastic Success: Moral Codes and Behavioral Outcomes (Doctoral dissertation in sociology and demography). UC Berkeley.

Greenhalgh, Susan (Ed.), 1995. Situating Fertility: Anthropology and Demographic Inquiry. Cambridge University Press, Cambridge.

Greenhalgh, Susan, 1994. Controlling births and bodies in Village China. American Ethnologist 21 (1), 3–30.
Hammel, Eugene A., 1990. A theory of culture for demography. Population and Development Review 16 (3), 455–485.
Hayford, Sarah, Trinitapoli, Jenny, 2011. Religious differences in female genital cutting: a case study from Burkina Faso. Journal for the Scientific Study of Religion 50 (2), 252–271.
Hayford, Sarah, Morgan, S. Philip, 2008. Religiosity and fertility in the United States: the role of fertility intentions. Social Forces 86 (3), 1163–1188.
Johnson-Hanks, Jennifer, Bachrach, Christine, Morgan, S. Philip, Kohler, Hans-Peter, 2011. Understanding Family Change and Variation: Structure, Conjuncture, and Action. Springer Verlag.
Johnson-Hanks, Jennifer, 2002. The modernity of traditional contraception. Population and Development Review 28 (2), 229–249.
Johnson-Hanks, Jennifer, 2006. Uncertain Honor: Modern Motherhood in an African Crisis. University of Chicago Press, Chicago.
Kertzer, David, Fricke, Thomas (Eds.), 1997. Anthropological Demography. University of Chicago Press, Chicago.
Kertzer, David, 1993. Sacrificed for Honor: Italian Infant Abandonment and the Politics of Reproductive Control. Beacon Press, Boston.
Leete, Richard (Ed.), 1999. Dynamics of Values in Fertility Change. Oxford University Press, Oxford.
Lesthaeghe, Ron, Neidert, Lisa, 2006. The second demographic transition in the United States: exception or textbook example? Population and Development Review 32 (4), 669–698.
Lesthaeghe, Ron (Ed.), 1989. Reproduction and Social Organization in Sub-Saharan Africa. University of California Press, Berkeley.
Lesthaeghe, Ron, 1983. A century of demographic and cultural change in Western Europe: an exploration of underlying dimensions. Population and Development Review 9 (3), 411–435.
Mason, Karen Oppenheim, 1992. Culture and the fertility transition: thoughts on theories of fertility decline. Genus 48 (3/4), 1–14.
Obermeyer, Carla Makhlouf, 2000. Risk, uncertainty, and agency: culture and safe motherhood in Morocco. Medical Anthropology 19 (2), 173–201.
Ryder, Norman, Westoff, Charles, 1971. Reproduction in the United States, 1965. Princeton University Press, Princeton, NJ.
Schneider, Jane, Schneider, Peter, 1996. Festival of the Poor: Fertility Decline and the Ideology of Class in Sicily, 1860–1980. The University of Arizona Press, Tucson.
Simons, John, 1995. Fertility and values in 15 western countries during the 1980s. In: Moor, Ruud de (Ed.), Values in Western Societies. Tilburg University Press.
Smith, Daniel J., 2001. Romance, parenthood, and gender in a modern African society. Ethnology 40 (2), 129–151.
Smith, Daniel J., 2004. Premarital sex, procreation, and HIV risk in Nigeria. Studies in Family Planning 35 (4), 223–235.
Shorter, E., Knodel, J., van de Walle, E., 1971. The decline of non-marital fertility in Europe, 1880–1940. Population Studies 25 (3), 375–393.
Thornton, Arland, 2001. The developmental paradigm, reading history sideways, and family change. Demography 38 (4).
Trinitapoli, Jenny, Weinreb, Alex, 2012. Religion and AIDS in Africa. Oxford University Press.
Watkins, Susan Cotts, Swidler, Ann, 2009. Hearsay ethnography: conversational journals as a method for studying culture in action. Poetics 37 (2), 162–184.
Watkins, Susan, Cotts, 1991. Provinces into Nations: The Demographic Integration of Western Europe, 1870–1960. NJ. Princeton University Press, Princeton, NJ.
Watkins, Susan Cotts, 2000. Local and foreign models of reproduction in Nyanza Province, Kenya, 1930–1998. Population and Development Review 26 (4), 725–759.
Westoff, Charles, Potvin, R., 1967. College Women and Fertility Values. Princeton University Press, Princeton, NJ.
Yeatman, Sara, Trinitapoli, Jenny, 2008. Beyond denomination: the relationship between religion and family planning in rural Malawi. Demographic Research 19, 1851–1881.

Fertility Transition: East Asia

Noriko O Tsuya, Department of Economics, Keio University, Tokyo, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Japan, Korea, and Taiwan experienced rapid transitions from high to the replacement level of fertility in the earlier postwar decades, followed by further declines to well below-replacement levels of childbearing in the recent decades. While the earlier fertility transitions were due mainly to reduction of marital fertility through contraception and induced abortion, the more recent transitions to very low levels were due largely to decreasing marriage among young women. The declining marriage and fertility were associated with increases in educational attainment and paid employment among women at peak reproductive ages in the context of rapid economic transformations.

East Asia experienced extremely rapid declines in fertility after World War II. In the last decades of the twentieth century, fertility declined further to below-replacement levels in some countries of the region, notably in Japan, the Republic of Korea (called 'Korea' hereafter for linguistic simplicity), and the Republic of China (called 'Taiwan' hereafter for the same reason). These three East Asian countries share the Confucian cultural heritage, which places heavy emphasis on obligations to family and society, leaving little room for self-interest (Bumpass et al., 2009). Because the ultimate goal of their traditional family systems was the biological or social continuation of the family line, the most important obligation, especially for women, was to marry and have children (Greenhalgh, 1988; Tsuya and Bumpass, 2004).

Focusing on these three countries with similar familial cultural backgrounds, this article looks at the patterns and factors associated with their postwar fertility transitions. After an overview of trends in fertility levels, the article discusses changes in the two major demographic factors that directly influence fertility: the age pattern of women's marriage and fertility within marriage. The article then discusses major socioeconomic factors that affect fertility, including women's education and employment in the context of rapid economic growth, urbanization, and industrial transformation. The article ends with a summary of findings and discussion of their demographic implications, focusing on population aging and population decline.

Fertility Trends

Although there were differences in timing and magnitude, Japan, Korea, and Taiwan all experienced dramatic downturns in fertility after World War II (see **Figure 1**). In Japan, the total fertility rate (TFR) declined from 4.5 children per woman in 1947 to 2.0 in 1957. After this rapid decline, fertility stabilized at the replacement level (a TFR of 2.0–2.1 children per woman) for nearly two decades until it started to decline again in the mid-1970s. Since then Japan's fertility has been steadily declining to below-replacement levels, reaching 1.5 children per woman in the 1990s, and then 1.3–1.4 children per woman in the 2000s.

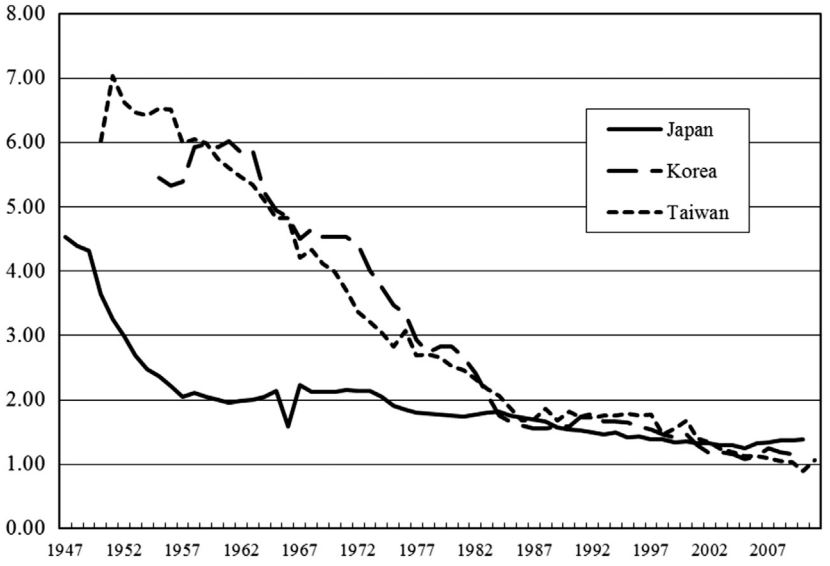

Figure 1 Total fertility rates in Japan, Korea, and Taiwan after World War II.

High fertility had persisted in Korea until rapid economic development began in the early 1960s. Then the TFR dropped precipitously from 5.9 children per woman in 1963 to 1.6 in 1986. This remarkable decline was followed by a modest upturn in the early 1990s until it started to slip again in the mid-1990s to very low levels (a TFR of 1.1–1.2 children per woman) in the 2000s. Taiwan's fertility transition was even more dramatic. Starting in the early 1950s, the TFR declined steadily and dramatically from 7.0 children per woman in 1951 to 1.7 in 1986. After a short pause, fertility began to decline again at the end of the 1990s, reaching a low of 1.0–1.1 children per woman in the late 2000s.

In summary, Japan experienced two fertility transitions, the first a rapid decline from historically high to replacement levels of childbearing shortly after World War II, and the second an ongoing decline since the mid-1970s from the replacement to very low levels of fertility. Korea and Taiwan both experienced phenomenal and almost continuous fertility transitions from very high to below-replacement levels from the early postwar decades until the mid-1980s, followed by further declines to extremely low levels of childbearing since the late 1990s. While many Western countries and some countries in other parts of Asia have also experienced or have been experiencing fertility declines to below-replacement levels, the fertility decline in East Asia is notable in its rapid pace and sheer magnitude. Today the levels of fertility in these three East Asian countries are all at or below what Kohler et al. (2002) considered as the 'lowest-low' fertility (a TFR of 1.3 children per woman).

Demographic Factors in Fertility Decline

In all three of these East Asian countries, the fertility transition is related directly to two demographic factors, namely, decreasing marriage among women and declining fertility within marriage. Because out-of-wedlock childbearing is minimal in East Asia, these two factors determine the overall levels of fertility. Demographic analysis indicates that roughly 90% of the decline in Japan's TFR between 1950 and 1975 was related to a reduction in marital fertility, and the remaining 10% was related to postponement of marriage among women. Similarly, around 70% of the TFR decline in Korea from 1960 to 1985 was due to declining fertility within marriage, and the remaining 30% was due to delayed marriage among women. In Taiwan around 75% of the decline in fertility between 1960 and 1985 was related to a decline in marital fertility, and the remaining 25% to later marriage. By contrast, the fertility decline in Japan after 1975 was due almost entirely to decreases in women's marriages. In a corresponding manner, the fertility declines from 1985 to 2010 in both Korea and Taiwan can also be attributed largely to later and less marriage among women.

In summary, the rapid transitions from high to the replacement level of fertility in Japan, Korea, and Taiwan after World War II were mainly the product of declining fertility among married women. On the other hand, unlike many countries in the West in which unmarried childbearing is common (van de Kaa, 1987), the more recent transitions to below-replacement fertility in these East Asian countries were, in the context of little childbearing outside marriage, due largely to declines in marriage among young women.

Reduction in Marital Fertility

The dramatic fertility declines from high to replacement or subreplacement levels in Japan, Korea, and Taiwan were caused primarily by reductions in fertility among married women, who gained control of their family size through contraception and induced abortion. The proportion practicing contraception among married women under the age of 50 years increased steadily in Japan from around 20% in 1950 to 61% in 1975 (Population Problems Research Council, 2000). After 1975, the prevalence of contraception stabilized at roughly 60%. With the expansion of government family planning programs, contraceptive practice spread rapidly in Korea and Taiwan from the 1960s to the early 1980s (Choe and Park, 2006; Chang et al., 1987). Contraceptive use among Korean wives aged between 15 and 44 years increased dramatically from only 9% in 1964 to 70% in 1985 (United Nations, 2000). In Taiwan contraceptive use among married women aged between 15 and 44 years also spread very rapidly from mere 4% in 1965 to 54% in 1985 (Sun, 2001).

Induced abortion is a traditional means of fertility control long practiced in East Asia. Along with the rapid increase in contraceptive use, rates of induced abortion also increased in the three East Asian countries. In Japan, the proportion of married women under age of 50 years reporting having experienced an induced abortion increased from 15% in 1952 to 41% in 1961 (Muramatsu, 1990). From the early 1960s to the 1980s, this proportion remained in the range of 30–40%. In Korea, the proportion of married women of reproductive age reporting having experienced an induced abortion increased from around 10% in the mid-1960s to nearly 40% in the late 1970s (Cho et al., 1982). Although the rate declined somewhat in the 1980s, induced abortion remains in the country as a viable option for terminating unwanted pregnancies. Although induced abortion had been illegal in Taiwan until 1985 (Wang, 1995), the reported use of abortion among married women of reproductive age rose steadily from the 1960s to the 1980s. The proportion of married women aged 35–39 years who ever had an induced abortion increased from 13% in 1965 to 42% in 1985 (Freedman et al., 1994).

Decreasing Marriage

As noted above, declining marriage has played a major role in the transitions of fertility to below-replacement levels in the three East Asian countries under consideration. Declining marriage for women in these countries is remarkable in light of the East Asian tradition of universal marriage and the cultural expectation that women should marry by their mid-twenties (Brinton, 1992). In Japan, the proportion never married was relatively stable among women in their twenties and thirties until the mid-1970s, when it started to increase precipitously. Between 1975 and 2010, the proportion never married rose dramatically from 21 to 60% among women aged 25–29 years, and from 8 to 35% among women aged 30–34 years (National Institute of Population and Social Security Research, 2013). Since the 1990s, there is also a sign of increasing nonmarriage: the proportion never married among Japanese women at age of 50 years – that is a commonly used indicator of the prevalence of permanent nonmarriage in

a population – edged up slowly from around 3% in 1975 to 5% in 1995, followed by a notable jump to 11% in 2010.

Korean women, too, are marrying increasingly later and less. The proportion never married among women aged 25–29 years rose slowly but steadily from mere 3% in 1955 to 18% in 1985. From 1985 to 2010, the proportion single jumped from 18% to 59% among women aged 25–29 years, and from 4% to 19% among women aged 30–34 years (Statistics Korea, 2011; Tsuya, 1991). Unlike Japan, however, there is not yet a clear sign of departure from the traditional pattern of universal marriage in Korea. The proportion never married among women at age of 50 years is still under 3% in 2010. Hence, while marriage has been increasingly delayed, universal marriage still remains intact in the country, although we cannot deny a possible 'spill over' of the increasing delay of marriage among younger age groups to older age groups, leading to increases in permanent nonmarriage among women in the coming years.

Taiwan has experienced, if not more, an equally dramatic decline in women's first marriage since the mid-1980s. The proportion never married slowly edged upward during the 1950s to the early 1980s: from 5% in 1956 to 23% in 1985 among women aged 25–29 years, and from only 2% to 9% among women aged 30–34 years. Between 1985 and 2010, increases in the proportion never married among women at the peak reproductive ages then accelerated: from 23 to 70% among women aged 25–29 years, and from 9 to 37% among women aged 30–34 years (Ministry of the Interior, 2011; Tsuya, 1991). Furthermore, similar to Japan, we can see a sign of increasing nonmarriage in Taiwan in recent years. While the proportion never married among women aged 45–49 years had been less than 3% until the mid-1980s, it increased notably from about 4% in 1985 to 10% in 2010.

Socioeconomic Factors in Fertility Decline

Although the time frame is different, Japan, Korea, and Taiwan all went through dramatic economic transformations after World War II, as characterized by the several decades of extraordinary economic growth and very rapid paces of urbanization and industrialization. This 'East Asian economic miracle' (McNicoll, 2006) preceded and progressed hand in hand with their fertility transitions, making them forerunners of the 'fertility revolution in Asia' (Leete and Alam, 1993). Concomitant with and as a result of the rapid economic development, educational levels and paid employment have increased dramatically in East Asia, and such gains have been especially remarkable for young women. These improvements in women's economic opportunities are among the major socioeconomic factors in fertility declines in the three East Asian countries under study (Chang and Li, 2011; Tsuya and Mason, 1995; Choe and Retherford, 2009).

Macro Economic Changes

Although the timing is different, Japan, Korea, and Taiwan all experienced extremely rapid economic growth after World War II. Starting in the mid-1950s, Japan's GDP per capita in constant prices grew at a phenomenal pace, quadrupling every 10–15 years until the late 1980s (Nakamura, 1981; World Bank, 1991). Between the mid-1960s and the mid-1990s, Korea's per capita GDP in constant prices also quadrupled every 10–15 years (World Bank, 1997). After doubling every 10 years from the mid-1960s to the mid-1980s, Taiwan's real per capita DDP quadrupled from the mid-1980s to the mid-1990s (Council for Economic Planning and Development, 2011).

In the course of rapid economic growth, these three countries experienced rapid urbanization. In Japan, the proportion of the population living in urban areas jumped from 37% in 1950 to 63% in 1960, and further to 91% in 2010 (National Institute of Population and Social Security Research, 2013). Korea has experienced similarly rapid urbanization, with the urban population rising from 21% of the population in 1950 to 83% in 2010 (United Nations, 2012). In Taiwan the proportion of the population living in cities of 150 000 or more inhabitants increased from 28% in 1965 to 72% in 2011 (Ministry of the Interior, 1976; Directorate-General of Budget, Accounting and Statistics, 2012a).

Along with rapid economic growth and urbanization, the three East Asian countries experienced dramatic changes in the structure of their industries. In Japan, the proportion of the workers employed in primary industries fell from 49% in 1950 to mere 4% in 2010 (National Institute of Population and Social Security Research, 2013). In Korea, the proportion of workers engaged in primary industries plummeted from 66% in 1960 to 8% in 2010. In Taiwan, the proportion of workers employed in primary industries dropped from 56% in 1952 to 5% in 2010 (Council for Economic Planning and Development, 2011).

Women's Educational Attainment

The dramatic economic transformations after World War II brought about important changes in women's status in the three East Asian countries under study. Educational attainment and paid employment increased, and these changes have affected fertility. Women's education is a key factor affecting fertility directly through the timing of marriage and marital fertility, and indirectly through employment and earnings (Cochrane, 1979). In Japan, a large majority of women were already receiving primary education before World War II; the most notable postwar change has been the steady, rapid increase in the proportion of women with higher education. The proportion of female high-school graduates advancing to higher education increased dramatically from 5% in 1955 to 56% in 2010, surpassing the rate of men during the 1990s (National Institute of Population and Social Security Research, 2013). The proportion of women aged 25–29 years who had higher education also increased from 4% in 1960 to 53% in 2010.

Unlike Japan, the educational attainment of Korean women was very low in the early postwar years and a sizable proportion of older women had no formal education. This situation has since changed dramatically, and the pace of increasing educational attainment was even more rapid than their Japanese counterparts. Only 3% of women born in 1937–41 had higher education, compared with 65% of women born in 1976–80 (Korea National Statistical Office, 2009). Taiwan has made

similarly impressive improvements in women's education. Like their Korean counterparts, educational attainment among Taiwanese women was very low in the early postwar years, with a sizable proportion of older women illiterate. But the female educational status has improved remarkably. While only 1% of women born in 1950–59 had higher education, the proportion with higher education increased to 28% among women born in 1970–79 (Chang and Li, 2011).

Women's Employment

With economies shifting rapidly from primary to secondary and tertiary industries, female employment underwent notable changes in the three East Asian countries under study. In Japan, although the overall female labor force participation rate has remained roughly 50% over the past 40 years, the proportion of working women in paid employment increased from 41% in 1960 to 84% in 2010 (National Institute of Population and Social Security Research, 2013). Furthermore, the age pattern of employment among women of reproductive age has been changing since the 1970s. The labor force participation rate of women aged 15–19 years has been declining steadily owing to increasing advancement to higher educational institutions and the rate of women aged 20–24 years has remained at roughly 70% throughout the postwar years, the labor force participation rates among women at the peak reproductive ages of 25–34 years increased sharply between 1975 and 2010: from 44 to 70% among women aged 25–29 years; and from 43 to 62% among women aged 30 to 34 years. Although the labor force participation rate among women at the peak childbearing ages is still lower than those of women in the younger and older ages thus preserving the well-known M-shaped age pattern of women's employment in the country, it has become much less notable since the mid-1970s. Although not as dramatically as among women aged 25–34 years, the labor force participation rate of women in their late thirties and forties also rose substantially during the same period. Hence, although Japanese women still tend to stop working during their peak childbearing and child-rearing years, such a tendency is weakening considerably in the recent decades.

In Korea, employment of women in their peak childbearing years began to increase precipitously in the 1980s. While the labor force participation among women aged 25–29 years had been stable at roughly 30% during the 1960s and the 1970s, it rose from 32% in 1980 to 68% in 2010 (Statistics Korea, 2011; Tsuya, 1991). The corresponding rate among women aged 30–34 years also increased from 41 to 56%, reaching levels similar to those of older women. The labor force participation rate in 2010 was 55% among women aged 35 to 39, and 62% among women aged 40–44 years. Consequently, the once notable M-shaped age pattern of women's employment has largely disappeared in Korea by 2010.

Unlike in Japan and Korea, there has been no M-shaped age pattern of female employment in postwar Taiwan, and the labor force participation rates of women of reproductive age have been increasing steadily since the mid-1950s, except for women aged 15–24 years whose employment rate decreased during the last few decades due to rising educational attainment. On the other hand, the labor force participation rose dramatically and steadily among women at peak childbearing and child-rearing ages. From 1966 to 2010, the labor force participation rate rose fourfold – from 21 to 84% – among women aged 25–29 years, almost four times from 19 to 77% among women aged 30–34 years, and three and half times from 21 to 74% among women aged 35–39 years (Directorate-General of Budget, Accounting and Statistics, 2012b; Tsuya, 1991).

Concluding Remarks

Japan, Korea, and Taiwan all went through rapid fertility transitions from high to the replacement level in the earlier postwar decades, followed by further and ongoing declines to well-below replacement levels of childbearing in the recent decades. While the earlier fertility transitions in these East Asian countries were due mainly to reduction of fertility among married women through deliberate control of their family size using contraception and induced abortion, the more recent transitions to very low levels of fertility were due largely to dramatic decreases in marriage among women at peak reproductive ages. These decreases in marriage among Japanese, Korean, and Taiwanese women at peak childbearing ages were associated with rising educational attainment among young women as well as increasing and prolonging paid employment among women at peak childbearing and child-rearing ages in the larger context of rapid economic development.

Although the timing, tempo, and scale of the fertility transitions are somewhat different among these three East Asian countries, dramatic reductions in their fertility to well-below replacement levels has resulted in or will soon bring about extremely rapid population aging and foreshadow dramatic population decline. With the proportion of population aged 65 and above years at 23% in 2010, Japan is currently the most aged country in the world, and its population aging is estimated to proceed further with the proportion of the elderly reaching 40% in 2060, according to the country's latest official population projection (based on the medium variants of fertility and mortality) (National Institute of Population and Social Security Research, 2013). Japan's population also started to decline in 2005, and the pace of the decline is projected to accelerate during the next half century. Japan's population size is estimated to shrink by almost one-third from approximately 128 million in 2010 to 87 million in 2060.

Although the populations of Korea and Taiwan are not yet as aged as Japan's, they are expected to age much more rapidly. In Korea, the proportion of population aged 65 years and above is projected to increase nearly four times from 11% in 2010 to 40% in 2060 (Statistics Korea, 2013) and, similarly, the proportion of the elderly in Taiwan is estimated to rise almost fourfold from 11% in 2010 to 39% in 2060 (Directorate-General of Budget, Accounting and Statistics, 2012a). While their population sizes are projected to grow slightly until the late 2020s to early 2030s, the populations of Korea and Taiwan are expected to shrink thereafter. Korea's population is estimated to edge up from around 49 million in 2010 to 52 million in 2030, followed by a sharp decline to around 44 million in 2060. Taiwan's population is projected to be stable at around 23 to 24 million between 2010 and 2030, and then to take a notable downturn to around 19 million in 2060.

See also: Abortion, Demography of: United States; Age Structure; East Asian Studies: Culture; East Asian Studies: Economics; East Asian Studies: Gender; Education and Economic Growth; Education and Employment; Family and Work; Fertility Change: Quantum and Tempo; Fertility Theory; Fertility Transition: China; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Fertility Transition: South Asia; Fertility Transition: Southeast Asia; Fertility and Culture: Anthropological Insights; Population Aging: Economic and Social Consequences; Population Forecasts; Second Demographic Transition.

Bibliography

Brinton, M.C., 1992. Christmas cakes and wedding cakes: the social organization of Japanese women's life course. In: Lebra, T.S. (Ed.), Japanese Social Organization. University of Hawaii Press, Honolulu, pp. 79–107.

Bumpass, L.L., Rindfuss, R.R., Choe, M.K., Tsuya, N.O., 2009. The institutional context of low fertility: the case of Japan. Asian Population Studies 5 (3), 215–235.

Chang, M.C., Freedman, R., Sun, T.H., 1987. Trend in fertility, family size preferences, and family planning practice: Taiwan, 1961–85. Studies in Family Planning 9, 54–70.

Chang, Y.C., Li, J.A., 2011. Trends and Educational Differentials in Marriage Formation among Taiwanese Women. RAND Labor and Population Working Paper Series WR-891.

Cho, L., Arnold, F., Kwon, T.H., 1982. The Determinants of Fertility in the Republic of Korea. National Academy Press, Washington, DC.

Choe, M.K., Park, K.A., 2006. Fertility decline in South Korea: forty years of policy-behavior dialogue. Korea Journal of Population Studies 29, 1–26.

Choe, M.K., Retherford, R.D., 2009. The contribution of education to South Korea's fertility decline to below 'lowest-low' level. Asian Population Studies 5 (3), 267–288.

Cochrane, S.H., 1979. Fertility and Education: What Do We Really Know? Johns Hopkins University Press, Baltimore.

Council for Economic Planning and Development, 2011. Taiwan Statistical Data Book 2001. Executive Yuan, tThe Republic of China, Taipei.

Directorate-General of Budget, Accounting and Statistics, 2012a. Statistical Yearbook of the Republic of China 2011. Executive Yuan, Republic of China, Taipei.

Directorate-General of Budget, Accounting and Statistics, 2012b. Annual Report of Manpower Survey Statistics 2011. Executive Yuan, Republic of China, Taipei.

Freedman, R., Chang, M.C., Sun, T.H., Weinstein, M., 1994. The fertility transition in Taiwan. In: Thornton, A., Lin, H.-S. (Eds.), Social Change and the Family in Taiwan. The University of Chicago Press, Chicago, pp. 264–304.

Greenhalgh, S., 1988. Fertility as mobility: Sinic transitions. Population and Development Review 14, 629–674.

Kohler, H.-P., Billari, F.C., Ortega, J.A., 2002. Tempo-adjusted period parity progression measures, fertility postponement and completed cohort fertility. Demographic Research 6 (6), 91–144.

Korea National Statistical Office, 2009. 2009 Social Indicators in Korea. National Statistical Office, Republic of Korea, Daejeon.

van de Kaa, D.J., 1987. Europe's second demographic transition. Population Bulletin 41 (1), 1–64.

Leete, R., Alam, I. (Eds.), 1993. The Revolution in Asian Family: Dimensions, Causes, and Implications. Oxford University Press, Oxford.

McNicoll, G., 2006. Policy lessons of the East Asian demographic transition. Population and Development Review 32 (1), 1–25.

Ministry of the Interior, 1976. 1975 Taiwan-Fukien Demographic Fact Book. Ministry of the Interior, Republic of China, Taipei.

Ministry of the Interior, 2011. 2010 Demographic Fact Book, Republic of China. Ministry of the Interior, Republic of China, Taipei.

Muramatsu, M., 1990. Induced abortion. In: Population Problems Research Council (Ed.), Summary of Twentieth National Survey on Family Planning. Mainichi Shimbun, Tokyo, pp. 83–97.

Nakamura, T., 1981. The Postwar Japanese Economy: Its Development and Structure. University of Tokyo Press, Tokyo.

National Institute of Population and Social Security Research, 2013. Latest Demographic Statistics 2013. National Institute of Population and Social Security Research, Tokyo.

Population Problems Research Council, 2000. The Population of Japan: An Overview of the 50 Postwar Years. Mainichi Shimbun, Tokyo.

Statistics Korea, 2011. Results of the 2010 Population and Housing Census. Statistics Korea, Daejeon.

Statistics Korea, 2013. Population Projections for Korea: 2010–2060. http://www.kostat.kr/portal/english/news/1/9/index (accessed 13.02.25.).

Sun, T.H., 2001. The impacts of a family planning program on contraceptive/fertility behavior in Taiwan. Ren Kou Zue Kan (Journal of Population Studies) 23, 49–92.

Tsuya, N., 1991. Trends and correlates of fertility decline in the NIES (in Japanese). Jinkōgaku Kenkyū (The Journal of Population Studies) 14, 49–66.

Tsuya, N.O., Bumpass, L.L., 2004. Introduction. In: Tsuya, N.O., Bumpass, L.L. (Eds.), Marriage, Work, and Family Life in Comparative Perspective. University of Hawaii Press, Honolulu, pp. 1–18.

Tsuya, N.O., Mason, K.O., 1995. Changing gender roles and below-replacement fertility in Japan. In: Mason, K.O., Jensen, A. (Eds.), Gender and Family Change in Industrialized Countries. Clarendon Press, Oxford, pp. 137–167.

United Nations, 2000. Levels and Trends of Contraceptive Use as Assessed in 1998. United Nations, New York.

United Nations, 2012. World Urbanization Prospects: The 2011 Revision. United Nations, New York.

Wang, P.D., 1995. Induced abortion in Taiwan. Perspectives in Public Health 115 (2), 100–108.

World Bank, 1991. World Tables 1991 Update. Johns Hopkins University Press, Baltimore.

World Bank, 1997. World Development Indicators 1997. World Bank, Washington, DC.

Fertility Transition: Economic Explanations

Paul T Schultz, Yale University, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Economic explanations for the fertility transition focus on the increase in wage premiums for better educated workers, especially for women, that encourage them to obtain more schooling and participate more of their time outside of the household. The private opportunity cost of mother's time employed in child-rearing has increased more rapidly than the cost of many other less time-intensive consumption goods and productive activities. These developments motivate parents to find less costly substitutes for additional births, and it is hypothesized that investments in the health, nutrition, and education of their children are such substitutes for more births. Empirical studies of individuals and countries offer a bridge connecting growth in labor productivity associated with the Industrial Revolution and the Fertility Transition. Yet the identification of well-measured forces outside of the family that set into motion these changes in family behavior remain controversial. What precisely increased the productivity of women's time outside of the home, made women's schooling more valuable and added to their life cycle earnings, and increased their children's survival? Variation in fertility has not been explained by a single widely accepted conceptual framework, in part because historical data over time on the forces outside the control of the family, which might drive these changes in reproductive behavior, are difficult to measure. Productive returns to schooling, for example, are rarely quantified before the start of the twentieth century, and even then wage returns are inferred for samples of individuals who report monetized labor earnings, the productive characteristics of those engaged in home production, and self-employment are expected to differ from those of wage earners. The underlying mechanism causing better educated workers to be more productive than their peers can be plausibly explained, but not confidently tested, such as the growing capital–labor intensity of production, or changes in the composition of output from industrial products to services, or increases in the relative price of goods and services that more educated females produce, or technical change that is biased toward employing better educated workers (Kuznets, 1966; Galor and Weil, 2000). States have also become more involved in the regulation of trade and industry, and the repercussions of social welfare policies on the family reinforce the incentives to reduce fertility. This article describes different lines of historical research mostly on the European demographic transition of the nineteenth century, and research findings from the study of low-income countries in the last 50 years.

Understanding the Economic Determinants of the Fertility

The collection of representative household surveys in the twentieth century and the reconstitution of community records of marriages, births, and deaths in earlier periods have allowed one to match at the individual level how marriage, fertility, schooling, economic activities, family wealth and income evolve together, and may help to account for how these forms of life cycle behavior at the family level are coordinated. Unexplained variations in these family behaviors are nonetheless likely to be interdependent since they are constrained by a family's shared endowments, market opportunities, and uncertainty introduced by weather, droughts, epidemics, etc. One approach to quantifying the external determinants of fertility is to measure and analyze the effects of shocks from outside of the family as a force that perturbs exogenously the family's environment and opportunities, and thereby simulates how family behavior responds to some key outcomes, and suggests how historical forces might affect fertility, such as changes in productive technology and trade that are expected to impact family outcomes during the fertility transition. One such independent biological event is for a woman in the first pregnancy to have twins, presumably an increase in her numbers of potential births that can be interpreted as a 'natural experiment' from which to estimate from individual data how this unplanned variation in the number of children affects the quality of subsequent children, such as their school attainment (Rosenzweig and Wolpin, 1980). Twins also provide a means for estimating how the resulting variation in family size affects women's commitments to the labor force, and affects the transfers a mother can expect to receive from her children in her old age, as recently documented in large panel surveys and censuses from China and Indonesia (Oliveira, 2013).

At the end of the nineteenth century, social legislation was first enacted for pensions for older workers, and a safety net established for the disabled, unemployed, and widowed that are hypothesized to have reduced the need for parents to have their own children to support them in old age. At about the same time, communities and governments began to coordinate tax support for public compulsory schools, child health programs, and restrictions on child labor. These public subsidies for child schooling and health are likely to affect private returns to these forms of child human capital investments and encouraged parents to substitute them for having more births. However, the inverse correlation between these important forms of human capital investments per child and the number of children parents have is not necessarily a causal effect in one direction or the other, unless some outside force, such as twins, is specified that would identify a plausible causal behavioral effect in only one direction.

Economic Approaches to the Fertility Transition

With the establishment of modern economic growth, fertility has tended to decline; first across high-income countries starting largely in the last quarter of the nineteenth century, and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31076-5

then even more rapidly across most low-income countries in the last quarter of the twentieth century (Kuznets, 1966; Schultz, 1981, 1997). Economic models of the allocation of time and wealth of families among production and consumption activities suggest several hypotheses linking changes in the economic environment in which people live to their reproductive demands and fertility (Becker, 1960; Mincer, 1963; Schultz, 1974). Empirical counterparts for the price, income, and environmental constraints on people's fertility choices have been used to explain aspects of fertility and other related outcomes of the family, strengthening the view that this household demand model sheds light on the determinants of the historic fertility transition (Schultz, 1974; Becker, 1981).

This economic approach to fertility determinants has been criticized by those who believe that fertility does not systematically adapt to changing economic conditions (Cleland and Wilson, 1987). The proposed alternative to the economic approach, that is called 'ideational change,' emphasizes social interactions and cultural barriers to the transmission of knowledge of birth control technology, but has not evolved to a form that it can be empirically distinguished from the predictions of the economic framework. However, research that seeks to quantify the interactive role of couples and household coresidents, or even peers, neighbors, or members of social networks can postulate in cooperative and non-cooperative game strategies within these groups of individuals. But because these groups are generally self-selected, it is complex to specify the factors that affect the formation of the group and do not also affect the behavior of individuals who decide to join such groups (Heckman, 1979; Manski, 1995). There is an active debate on how to apply the economic framework and to exploit statistically cross-sectional and time series variation in fertility to improve our understanding of these more diffuse determinants of fertility (e.g., Schultz, 1997).

There is a fundamental methodological divide between researchers who want to describe a 'structural model' that involves interactions between choices made at the household level, including those choices spanning an adult's lifetime, such as between a woman's fertility, marriage, and labor force history, and those who investigate a 'reduced form model' where only outside constraints on household choice are considered as potential explanatory variables, or commonly, market prices, predetermined income sources, technologies, and the physical–biological environment. It is widely recognized that the associations between lifetime choices represent more (or less) than a simple causal effect of one choice on another. Covariations between family choices embody possible effects transmitted in both directions, effects of unobserved factors that impact both outcomes, and heterogeneity due to latent differences among families, in individual constraints or preferences (Schultz, 1981). Identification restrictions are generally required to be estimated from nonexperimental data, the critical parameters to 'structural models,' and therefore the empirical conclusions drawn from them are often controversial, because they depend on untested and debatable modeling specifications (Manski, 1995). The intuitive appeal of studies of the consequences of 'twinning' is that they rely on a demographic treatment that is not readily controlled by the family, at least not until recent times.

The alternative approach is to estimate fertility as a function of explanatory variables that are not family choice variables or are not likely to be impacted by unobserved factors that could also affect fertility. By delimiting the potential list of explanatory variables, these 'reduced form' relationships represent a summation of the direct effect of the environmentally fixed constraints on the couple's fertility, plus the possible indirect effect of the same constraints on fertility that operate through the modification of other choice variables, such as the woman's age at marriage, extent of labor force activity outside of her home, sexual behavior, and contraceptive choice, etc. The limitation of reduced form estimates is that they do not necessarily clarify the mechanism by which the constraints impact fertility, for example, through the timing of marriage, or the adoption within marriage of various so-called 'natural' or 'unnatural' forms of birth control. The strength of reduced form estimates is that they try to provide an unbiased approximation for the total effect of the constraint on fertility, which is often a central concern of the social scientist, demographic forecaster, or policy maker.

Some historical demographers assign special importance to the moment when married women begin to restrict their birth rates in response to their prior number of births, or surviving children (Coale and Watkins, 1986). Although this may be a well-defined and interesting threshold that predicts the onset of the long-term decline in fertility in certain European populations, other populations may adopt other mechanisms to reach their changing reproductive goals, such as a change in median age at marriage (i.e., Malthus' preventive check), an increase in celibacy or the proportion who never marry in a birth cohort, an increase in the spacing of births within marriage, as may be evident in Africa today (Caldwell et al., 1992; Shapiro and Tabashe, 2003), or a change in nonmarital (illegitimate) fertility. Even in northern Europe, standard measures of the emergence of parity-specific fertility control within marriages are not precisely defined by existing data (Craft, 1989; Bengtsson and Dribe, 2006; Guinnane et al., 1994). Thus, economist generally focus on cohort fertility such as the number of children ever born, and do not single out one component, such as duration-specific marital fertility, that implicitly presumes the timing of entry into marriage is unrelated to reproductive goals or subsequent lifetime behavior of the couple. Because many studies of the fertility transition mechanically hold constant the age at marriage or the age composition of married women, to estimate what is then interpreted as the behavioral determinants of fertility, this methodology explicitly assumes that age at marriage is not in part a choice that could be responding to economic determinants of desired lifetime fertility.

Stages of the Demographic Transition

The 'demographic transition' refers to a sequence of three periods. In the first period, fertility and mortality rates are high and mortality tends to be highly variable, with population growth fluctuating widely about a moderate long-run trend of

growth or decline. Short-run responses in fertility (and mortality) can be attributed to cycles in weather, harvests, epidemics, or possibly to other exogenous variables, and the analysis of these shocks with time series statistical methods, such as vector-autoregression. In the second period, age-specific mortality rates decrease gradually, raising life expectancy at birth from 30–35 years in preindustrial Europe, to 70–80 years today in the high-income countries, and from 25–30 years in the low-income countries in the 1920s, to levels which today range from about 45–75, excluding countries ravaged by war or the onset of the AIDS epidemic. This second period is one of accelerating population growth, which some observers such as Malthus attributed to improved food supplies, and more specifically to the effects of nutrition on mortality, often initiated by technical progress in agriculture, followed by improved industrial productivity, and advances in transportation and communication, or the opening of the national economy to international trade (Fogel, 2004). A free press capable of publicizing food shortfalls and imminent famines has also been proposed as an institutional arrangement in the twentieth century that further mitigated excess mortality in the wake of periodic droughts and potential famines. Improvements in private and public health technologies are assigned an important role in reducing mortality, but some scholars argue that medical care was not very important in determining life expectancy before the start of the twentieth century with the introduction of antibacterials (sulfanilamide and penicillin) after the 1930s (Fogel, 2004).

In the third period of the transition, the secular decline in fertility begins, after which the number of births per woman falls by more than half, from five or six to about two, in the high-income countries, and from six to eight to about two or three in most low-income countries. Demographic experts conclude that when high-fertility levels have declined by more than 10%, they are not expected to rise again on a sustained basis (National Research Council, 2000). Medium-term swings in fertility are nonetheless distinguished from secular time series trends, occurring in response to major business cycles (e.g., the Great Depression followed by the post-World War II baby boom), wars, and economic shocks due to fundamental institutional changes, as occurred during the transition from centrally planned to market-oriented economies in the 1990s in eastern Europe and Russia, or in Iran after its Islamic Revolution, or during economic reversals in sub-Saharan Africa.

Economic Determinants of the Fertility Transition

Notestein (1945) described the early dimensions of the demographic transition, and speculated on the economic pressures, which he expected to reduce fertility in the low-income world: rising cost of children in urban life, individualistic promotion of child health and education. Economic models of a lifetime fertility choice emphasize different facets of this decision process, but the trade-off between quantity of children and their quality (i.e., education and health) has become a central theme (Becker, 1960, 1981; Schultz, 1981). Because women's time devoted to childcare is a large fraction of the modern private (opportunity) cost of fertility, her allocation of time is treated as jointly determined with fertility. Many cross-sectional and a few time series studies have related statistically this constraint to fertility outcomes.

A couple's lifetime wealth (endowments of human capital and nonhuman capital) is expected to increase with modern economic growth, and to encourage parents to demand more children, if the relative costs and functions of children did not change (Becker, 1960; Schultz, 1981). However, for reasons discussed below, the opportunity cost of children increased along with the cost of other goods and activities that complement the rearing of children, whereas the private price of some substitutes for children, such as old-age pensions and health care insurance, have been socialized and subsidized by governments. Marked improvements in the technology of birth control were developed in the 1950s, translating into the rapid diffusion of the pill (oral steroid contraceptive) and the interuterine device (IUD) to many parts of the world. The monetary and psychic cost and inconvenience (independent of coitus) of avoiding too many births was reduced, and should have reduced thereby the number of births. Family planning programs are not, clearly, a necessary condition for the occurrence of a fertility transition, because birth rates declined in some high-income countries without the benefit of new control technologies, or of public family planning extension activities designed to facilitate the adoption of birth control.

Finally, improvements in child health technology increased the proportion of births that survived to adulthood. Under two plausible assumptions these improvements in child survival would have fostered a homeostatic compensatory decline in fertility: (1) parent reproductive goals are defined primarily in terms of the number of surviving children parents have, rather than their number of births, and (2) these demands for survivors are highly price inelastic, such that the decline in child mortality, which reduces the opportunity costs for parents to produce a survivor, does not outweigh the substitution of parent resources into investments in higher child quality (Schultz, 1981). The actual expenditures parents commit to each of their children is not an indicator of the fixed private 'market price' of a child, but reflect the combination of the price of child inputs and also the parent's demand for child 'quality.' The market-determined prices of the inputs that households combine with their time to rear children are exogenous market prices. There is little consensus on which inputs to children are relevant, or how constant-quality input prices for children have changed, or whether they play a major role in the secular decline in fertility.

If Becker's (1960, 1981) hypothesis that quality and quantity of children are viewed by parents as substitutes is valid, and much of the variation in child quality is embodied in the education parents give their child, then the net decline in private price of schooling would affect positively the demand for quality and depress fertility. There is growing evidence in the twentieth century that wage returns to schooling increased in a widening circle of countries, and this can be viewed as a decline in the private price of schooling, which encourages well-informed parents and societies to invest more resources per child in the education of children. The increase in wage returns to schooling has been attributed to a technological bias

favoring skilled labor, or to the complementarity in production between the schooling of workers and the accumulating stock of technical knowledge available in the global economy.

Becker (1981) also assumes that the 'income effect' on the demand for child quality is a larger positive magnitude than the 'income effect' on the demand for child quantity. This might encourage parents to substitute further child quality for child quantity as their incomes increased from one generation to the next. This is an additional explanation for the anomalous fall in fertility in a period when per capita income was increasing and the private costs of rearing a surviving healthy child was falling.

The Labor Intensity of Children and Their Opportunity Cost

In the twentieth century, the fraction of national income paid out to labor has increased, despite the increase in capital employed per worker and the decrease in hours worked per year per worker (Kuznets, 1966; Fogel, 2004). Wage rates have grown relative to per capita income until the 1970s, implying that consumption activities that use more time of the consumer have become relatively more expensive over time. Children are thought to be a more 'time-intensive commodity' than the average of all consumption commodities. Therefore, the opportunity costs for household's to produce and benefit from a child have risen as parent wages have risen, offsetting some part of the increased demand for children that might have been otherwise attributed to rising per capita income (Mincer, 1963; Becker, 1965).

The household production–consumption model of family labor supply and resource allocation drew attention to gender specialization in household production, and to the 'female time-intensity' of children, thereby suggesting the woman's value of time could be a key component in the rising price of children (Mincer, 1963). In environments where the opportunity cost of female (i.e., mother's or childcare giver's) time is higher, fertility tends to be lower, other things equal (Schultz, 1981). The market wage available to women might thus be a useful 'opportunity price' variable for a child, and it has notably increased in the twentieth century at a time when fertility has fallen. But just as 'parental expenditures per child' do not represent a fixed market-determined price of a child, the woman's market wage does not represent a fixed price of a woman's lifetime input to child-rearing. First, the wage a woman receives in the labor force is affected by her prior job experience, which accumulates at different rates depending on her preferences among types of work and the allocation of her time. Consequently, a woman's wage is an endogenous choice variable closely linked to her lifetime allocation of time, which in turn is intimately related to her fertility decisions (Mincer, 1963).

A second problem with the woman's wage variable is that it is observed only for a self-selected fraction of women who work for a wage in the survey period, and this fraction has tended to rise over time in many populations, and to differ by socioeconomic strata within a population. It is possible that women who work for wages are those who receive an unexplained high-market wage offer, suggesting that the observed average wage of working women overstates the increasing value of time of an average woman. Conversely, women who are unusually productive in nonwage work, and thus are less likely to participate in the wage labor force, may still have an above average market-wage offer that they declined, in which case the observed average wage of working women would understate the increasing value of time of an average woman. Sample selection bias could distort the observed wage as a measure of the average value of women's time, and thus bias estimates of the 'woman's wage effect' on fertility (Heckman, 1979; Schultz, 1981).

The Value of Women's Time and the Cost of Children

If the observed wage of women is endogenously affected by women's life cycle behavior and cannot be treated as an exogenous market determinant of her life cycle fertility, women's wage cannot be directly included in a reduced form equation to predict its causal effect on fertility. A structural model may then be needed to estimate how productive opportunities for women to work outside of the home change, that are expected to lead to women's greater participation in the formal labor force, on-the-job advancements, raising productivity and wages that could reduce their fertility (e.g., Wanamaker, 2012). One approach to this problem is to employ the woman's schooling as an 'instrument' for her wage; schooling tends to be a good predictor of women's wages in the cross-section, just as schooling predicts men's wages. For most women, their human capital investment in schooling is completed before childbearing starts, and thus the schooling decision can be tentatively viewed as predetermined with respect to the choice of fertility; this is certainly plausible in nineteenth-century Europe, if not today in parts of sub-Saharan Africa where pregnancy may terminate the schooling of many girls. If the schooling decision is made by the child's parents, it may also be argued that the woman's schooling is relatively unaffected by her adult preferences between a labor market career and early fertility, in which case schooling may be a satisfactory 'instrument' for predicting the adult lifetime wage opportunities for adult women caused by exogenous forces outside the immediate control of adult women.

A second approach is to treat the structure of wages in the local labor market as exogenous from the individual's perspective, and the average local wage can be assumed independent of the individual woman's heterogeneity in her preferences between working outside the home and fertility. By averaging local individual wages at the community level and thus mitigating the correlation between idiosyncratic individual variation in preferences and wages, one can assess the fertility of women as a function of the regional average wage. But another source of bias could arise if general equilibrium pressures on the local wage due to unobserved regional conditions affected the local labor supply of women. Such an omitted variable that increases the labor supplied by women in a region should, other things equal, depress the market-clearing wage women receive in that locality, and could thereby boost fertility through the effects on the opportunity price of children. Consequently, such regional-omitted labor supply variables could bias the estimated effect of woman's regional

wage on fertility, masking the theoretically expected negative impact of aggregate derived demand-induced variation in women's wages on fertility.

To deal with this regional-omitted variable problem a regional instrument for wages may be useful that shifts only the aggregate regional economy's demand for female labor (relative to the demand for male labor), and thereby affects the local wage available to women (relative to men), but is not otherwise expected to influence fertility. Variables of this form that shift the aggregate regional derived demand schedule for women's labor are additional candidates for use as an 'instrument' for women's wages. Several types of variables might serve this purpose. One such demand variable is the regional location of natural resources; a second would be the composition of industries, which could also be affected by location of resources and the transportation network; a third could rely on the local composition of agricultural crops that are most profitable in the local agro-climatic conditions and soils and lead to the engagement of different proportions of female and male workers; and a fourth would rely on changes over time in internationally determined export–import prices for crops or industrial commodities that have a distinctive gender labor intensity, such as cash crops produced by males in some areas, such as coffee in Kenya, or Cacao in Ghana and Cote d'Ivoire, or tea in India and China. Finally, longer term evolutions of productive technology, such as higher yielding agricultural varieties and livestock or mechanization that shift the demand for labor in agriculture between men and women. A few illustrative studies can show how derived demand determinants can affect the relative productivity of women's labor.

Coal mines employ mostly men, and communities dependent on coal mines for their livelihood tend to have relatively low wages for women. Historically, these coal communities in the UK, United States, and elsewhere report relatively high-fertility levels, other things being equal (Haines, 1979). In the northern provinces of Sweden where the price and quantity of timber exports increased at the end of the nineteenth century, men's wages increased relative to women's wages, and fertility remained high well into the twentieth century. By contrast, fertility in the southern provinces of Sweden declined after the 1860s, as women's wages increased relative to those of men, in response to higher prices for exports of livestock, dairy, and food manufactured products, which relied relatively heavily on female labor (Schultz, 1985).

Natural resource exports also strengthen government finances by providing an inelastic source of tax revenues, notably in the case of oil and mineral exports. If traditional elites capture the rents on such exports, other forms of direct personal taxation can be reduced, though government revenues may then be less often allocated to support social services such as public health and schooling. Oil and mineral exporting countries tend to have relatively low levels of female paid labor force participation, and high fertility, holding constant for other factors including per capita income, and religion (Schultz, 1994).

Labor economists have noted that in regions where the share of employment in certain industries, for example, the services, are relatively large, female labor force participation rates tend to be higher, other things being equal. The importance of part-time, less-skilled, and physically less-demanding jobs are often associated with a larger share of women in the labor force, and also with a rise in the women's wages relative to men's, and lower fertility, other things being equal. Changes in fertility over time can then be potentially explained by changes in factors external to the household sector that govern the expansion of paid employment opportunities in industries that hire a disproportionate share of women (or men). The international price of a country's output (terms of trade) is another externally determined factor driving domestic demand that affects the composition of output and possibly gender differences in wages (e.g., Schultz, 1985). However, these changes in relative prices and employment prospects by gender must be perceived by couples as a long-term trend that they need to incorporate into setting their lifetime reproductive goals. On the other hand, year to year fluctuations in the price of coffee, which may employ much female labor in Colombia in the 1960s, might not have a notable impact on long run expectations and thereby influence cohort completed fertility.

The European Fertility Transition and Low Income Countries

With the opening of the US Erie Canal in 1825 and the reduction of ocean transportation costs, the increased global supply of grain from the middle west and from southern Russia depressed grain prices in European markets. The United Kingdom repealed their 'corn laws' in 1846, and removed their prohibitive tariffs on grain, and this change in trade policy improved opportunities for industrialization in the UK by reducing the cost of food and tending to increase the real wage of workers in export sectors such as textiles, which were relatively female labor intensive (e.g., Craft, 1989; Galloway et al., 1994). Some European countries continued to protect domestic grain producers, such as France, whereas others adapted to low-cost grain such as Sweden, Denmark, and the Netherlands. Even within agriculture, the composition of crops can affect the productive opportunities of women relative to men and their relative wages. As noted earlier in Sweden, dairy products and the export of butter encouraged employment of women and raised locally women's wages, whereas the decline in prices of grains depressed the agricultural wages of men. But different climate and soils of the different provinces of Sweden intermediated where these trade-induced movements in commodity prices affected most directly women's relative wages and fertility. In southern Sweden where the dairy industry was concentrated, fertility first declined, whereas in northern Sweden where forestry dominated the economy, male wages increased relative to female. Higher real wages for men were related to younger childbearing, but not to a significant change in total fertility rates. About one-third of the 25% decline in total fertility rates for Sweden as a whole in the period from 1860 to 1910 can be attributed to trade-driven increases of women's wage opportunities, whereas the rest of the decline in Swedish fertility in this half century is associated with urbanization and the increase in child survival, under the assumption that these were exogenous

developments (Schultz, 1985). Swedish parish registry data also confirm region-specific responses of fertility of individual women to these movements in international trade prices (Madestam and Simeonova, 2013). The Princeton European Fertility Project that sought to account for the fertility transition in Europe focused on Census indicators of industrialization, and not on the realignment in the prices of food grains and other agricultural and industrial commodities that may have impacted employment opportunities for European women by agricultural region (Coale and Watkins, 1986).

Agricultural productivity of men and women also differ markedly in low-income countries, and these differences have been attributed to natural resources, climate, and technology and the domestication of crops and livestock (Boserup, 1970). For example, Lavy (1985) attributes regional differences in relative wages for men and women in Egyptian agriculture to the regional comparative advantages in wheat or cotton. Female labor is relatively less valuable in wheat cultivated with plow technology than in cotton where women are valuable as cultivators and harvesters, and fertility is accordingly higher in the wheat growing regions.

When Mao relaxed the requirement that all regions in China be self-sufficient in food, tea production revived in those areas where geography and climate favored tea cultivation. Women's wages tend to be higher relative to men's in areas of tea cultivation in many parts of the world, due presumably to women's comparative advantage in plucking tea. The gender ratio of surviving females to males in the 2000 Census of China from cohorts born after Mao's policy increased in traditional tea regions, as did the schooling of girls relative to boys, even when these regions are compared to regions that specialized in the cultivation of orchards rather than tea. Survival and schooling of girls improves relative to boys as a response to the shift in composition of crops toward the female labor intensive tea (Qian, 2008).

In Ghana, cultivation of tree crops engage disproportionately more male than female labor, whereas cereals, fruits, and vegetables employ a greater share of female labor, given the local reliance on 'hoe technology.' If the differences in climate, soils, and agricultural technology account for the choice of crops by region, and explain gender differences in employment and wages by region, these agroclimatic factors may be a neglected key to explain women's participation in the family labor force and their fertility (Schultz, 2013). Crop composition also appears to be a local determinant of the prevalence of polygyny in neighboring Cote d'Ivoire as it also affects the family productivity of female labor (Jacoby, 1995). These studies presume that sex-specific interregional migration does not arbitrage away these gender differences in local wages or labor productivity affected by agroclimatic advantages for growing specific crops. Sex-specific migration may be restrained, depending on the country, by ethnicity, language barriers, and the slow diffusion of crop-specific agricultural productive knowledge between regions, or in the previous case of tea in China, by the government's system of residential registration that restrains the movement of the rural population.

The fertility transition in Prussian Germany has also been analyzed by region from 1816 to 1910. When regional persistent or regional fixed effects are included in the analysis of a pooled times series of cross sections, strong partial effects of industrialization and urbanization are associated with declining fertility from 1850 to 1910 (Richards, 1977). A regional fixed-effect specification for fertility has been estimated for smaller administrative units of Prussia from 1875 to 1910, but female labor force participation rates are added to explain the change in fertility (Galloway et al., 1994). This measure of labor force participation of women is plausibly a decision made jointly with the timing of marriage and fertility, and therefore when women's participation in the formal labor force is included as though it were an exogenous control, it accounts for the largest fraction of the fertility variation across Prussia during its fertility transition. These estimates are likely to be subject to simultaneous equation bias and are difficult to interpret as causal evidence of the impact of industrialization.

County-level data from the Prussian Censuses of 1816, 1847, and 1867 have subsequently been studied and they suggest that before the onset of the general fertility transition, fertility is already inversely related to school enrollment rates (Becker et al., 2013). However, this early indication of a trade off between quantity and quality of children does not establish causation without an exogenous explanation for why there is variation across communities in their investment in public schools or willingness to effectively enforce compulsory schooling laws. One hypothesis has been advanced that concentration of land ownership diminishes the effective demand for schooling in rural areas, because large landlords do not support the public provision of schooling (Galor et al., 2009). Assuming the landed aristocracy preferred lower wage farm labor and not better educated workers, the degree of concentration of land ownership by county would diminish school building. School enrollment rates predicted by the county are 'instrumented' by land concentration, which are inversely related to early fertility declines. This logic is extended further by ordering the census variables for counties over time. Land inequality in 1816 is used as an instrument to explain lower county enrollment of girls in 1816, which is associated with higher fertility of older women in 1847 and 1867 (who are presumably less well educated). As a further consistency check, changes between censuses in these county-level variables are used to account for changes in female enrollment and fertility (Becker et al., 2013). Nonetheless, many of these investigations of Prussian fertility, as with the European Fertility Project, control for the timing or prevalence of marriage, and even for the labor force participation of women, which are undoubtedly jointly determined choices along with fertility.

The Global Advancement in Women's Schooling

Throughout the world the advance of women's schooling tends to be associated with lower fertility, especially when controls are included for men's schooling or per capita income or consumption. This empirical regularity is often interpreted as a consequence of the increased opportunity cost of children when educated women enjoy improved opportunities to work outside of the home. By the start of the twentieth century, the

gender gap in schooling in the United States had diminished, except for college education. But this gap remained substantial in most countries in southern and eastern Europe until after the World War I, and women's years of schooling attainment began to catch up to that of men in low-income countries mainly in the aftermath of World War II. This trend in female schooling has been particularly slow in rural parts of South Asia where the fertility transition has also been delayed compared with East and Southeast Asia or Latin America. The wages of women with an additional year of schooling tend to increase by at least the same percentage as they do for men, which implies as the gender absolute gap in years of schooling diminishes, women's wages tend to increase relative to men's wages, as is seen in many parts of the developing world (Schultz, 1995).

The factors that have triggered these differential trends in the schooling of women and men have received relatively little study, and are often attributed to prevailing customs, and not explained by local exogenous changes in endowments, economic growth, composition of output, or globalization. But whatever the role of preexisting conditions and policy reforms, the expansion of women's schooling closely parallels the decline in fertility across high- and low-income countries since 1960, at both national and individual levels. Indeed, changes in women's schooling accounts for the largest share of the changes in total fertility rates observed in low-income countries since 1960, and in Africa since 1980 (Schultz, 1994). Even in an environment where economic progress has been minimal since the 1960s – Kinshasa City in the Congo – the increasing rate of female education is the most obvious factor explaining different levels of fertility across the population, and accounting for the decline in fertility in the 1980s and 1990s (Shapiro and Tombashe, 2003).

However, educated women could have lower fertility for many reasons, in addition to the greater opportunity cost of their time in child-rearing. The social and intellectual advantages that educated women may enjoy could help them in deciphering, adopting, and using effectively, new and old forms of birth control, and thereby avoiding births they do not want.

Female schooling may also affect fertility through its impact on the risks of child mortality. Censuses and surveys after World War II have documented for dozens of countries that child survival rates tend to be about 5% higher for mothers who have completed an additional year of schooling (Schultz, 1981, 1997). The proportionate decline in fertility associated with women's schooling is nonetheless larger than the proportionate increase in their child survival rates. Better educated women thus have smaller numbers of surviving children, which implies better educated women contribute less to the overall intergenerational rate of population growth than do their less educated peers (Schultz, 1994).

Schooling of women is also closely related to their labor force participation outside of their family, holding constant for their husband's income or education (Schultz, 1981). Household models of production and consumption generally assume that family labor supply choices are coordinated or jointly determined by family members, and women's education is expected to affect not only lifetime fertility, but also the timing of marriage and the positive matching of the productive characteristics between husband and wife in the marriage 'market.' Within such a framework of family-coordinated decision-making, it is a challenge to assess empirically how one of these jointly determined life-cycle choices made by women affects another of their choices.

To decompose the variation in fertility into that which is associated with the proximate timing of marriage, marital fertility rates, and non-marital fertility rates is a descriptive accounting exercise, but should not be used to infer causal connections. These three demographic components are likely to be behaviorally interdependent, because unobserved characteristics of women and their environments are likely to influence all of these components. The goal of explaining the fertility transition requires clarification of what external environmental developments or shifts in policy trigger or restrain the timing of the fertility transition.

Conclusions

The equalization of schooling of women and men and the reduction in the relative gender gap in wages are linked to the quality–quantity trade-off that relates the investment of parents in their children's schooling and health and the decline in fertility. This empirical regularity appears to be an important bridge between the Industrial Revolution and the Fertility Transition (Notestein, 1945; Galor and Weil, 2000). However, to understand the origins and causal significance of this inverse association requires in addition the identification of factors external to the household and community that can be accurately measured and are plausible determinants of private demand for, or public supply of, schooling for women and men. In addition, this factor must not otherwise affect the private demand for births or the biological determinants of reproductive potential or the supply of births, if the data are to single out the mechanisms by which variation in society contributes to time trends in fertility.

Economic explanations for the fertility transition emphasize the increasing returns to schooling since the Industrial Revolution, especially for women, which has encouraged women to obtain more schooling relative to men, and drive the advancement in women's wages absolutely and relative to the wages of men. These general attributes of economic development are not equally evident in all countries. Declines in fertility have therefore proceeded quite unevenly, first across the high-income countries, and more recently across the low-income countries, but following a pattern of economic developments that has been widely quantified in household surveys and suggestively followed in cross-country comparisons. Two other economic factors linked to the fertility transition are the decline in mortality, particularly among children, and the increased value of employment opportunities available to women outside of their homes.

At an aggregate level, many states have simultaneously legislated socialized support systems to provide for the consumption requirements and medical care of the elderly, and for the school and health investment requirements of the young, which may have on balance reduced the private demand of parents for children. Conversely, the provision of these health and pension services by governments may ultimately reflect the revealed preferences of the population to rely

on the public sector rather than on their own children for such old-age supports. With modern economic growth associated with the emergence of new specialized jobs that rewards workers for their geographic mobility, the intergenerational family support network is placed at a growing disadvantage as a social insurance scheme that depends to a great extent on residential proximity of parent and child generations.

See also: Children, Value of; Fertility Theory: Embodied-Capital Theory of Life History Evolution; Fertility Theory: Theory of Intergenerational Wealth Flows; Fertility Transition: Cultural Explanations; Gender and Feminist Studies in Geography.

Bibliography

Becker, G.S., 1960. An economic analysis of fertility. In: Demographic and Economic Change in Developed Countries. Princeton University Press, Princeton, NJ.

Becker, G.S., 1965. A theory of the allocation of time. Economic Journal 70, 493–517.

Becker, G.S., 1981. A Treatise on the Family. Harvard University Press, Cambridge, MA.

Becker, S., Cinnirella, F., Woessmann, L., 2013. Does women's education affect fertility: evidence from pre-demographic transition Prussia. European Review of Economic History 17 (1), 24–44.

Bengtsson, T., Dribe, M., 2006. Deliberate control in a natural fertility population: southern Sweden 1766–1864. Demography 43, 727–746.

Boserup, E., 1970. Women's Role in Economic Development. St. Martin's Press, New York.

Caldwell, J.C., Orubuloye, I.O., Caldwell, P., 1992. Fertility decline in Africa: a new type of transition. Population and Development Review 18 (2), 211–242.

Cleland, J., Wilson, C., 1987. Demand theories of the fertility transition: an iconoclastic view. Population Studies 41 (1), 5–30.

Coale, A.J., Watkins, S.C. (Eds.), 1986. The Decline of Fertility in Europe. Princeton University Press, Princeton, NJ.

Craft, N.F.R., 1989. Duration of marriage, fertility and women's employment opportunities in England and Wales in 1911. Population Studies 43, 325–335.

Fogel, R.W., 2004. The Escape from Hunger and Premature Death, 1700–2100. Cambridge University Press, Cambridge, UK.

Galloway, P.R., Hammel, E.A., Lee, R.D., 1994. Fertility decline in Prussia 1875 to 1910: pooled cross section time series. Population Studies 48 (1), 135–158.

Galor, O., Moav, O., Vollrath, D., 2009. Inequality in Landownership, the emergence of human capital promoting institutions and the great divergence. Review of Economic Studies 76 (1), 143–179.

Galor, O., Weil, D.N., 2000. Population, technology, and growth: from Malthusian stagnation to the demographic transition and beyond. American Economic Review 90 (4), 806–828.

Guinnane, T.W., Okun, B.S., Trussell, J., 1994. What do we know about the timing of the European fertility transition. Demography 31 (1), 1–20.

Haines, M.R., 1979. Fertility and Occupation: Population Patterns in Industrialization. Academic Press, New York.

Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47 (1), 153–161.

Jacoby, H.G., 1995. The economics of Polygyny in sub-Saharan Africa: female productivity and the demand for wives I Cote d'Ivoire. Journal of Political Economy 103 (5), 938–971.

Kuznets, S., 1966. Modern Economic Growth: Rate Structure and Spread. Yale University Press, New Haven, CT.

Lavy, V., 1985. Cropping pattern, mechanization, child labor and fertility behavior in rural Egypt. Economic Devevelopment and Cultural Change 33 (4), 777–791.

Madestam, A., Simeonova, E., 2013. Gender Empowerment in 19th Century Sweden: Using Historical Trade Shocks. paper American Economic Association meetings, San Diego CA.

Manski, C.F., 1995. Identification Problems in the Social Sciences. Harvard University Press, Cambridge, MA.

Mincer, J., 1963. Opportunity costs and income effects. In: Christ, C., Friedman, M., Goodman, L., Griliches, Z., Harberger, A., Liviatan, N., Mincer, J., Mudlak, Y., Nerlove, M., Patinkin, D., Telser, L., Theil, H. (Eds.), Measurement in Economics. Stanford University Press, Stanford, CA.

National Research Council, 2000. Beyond Eight Billion. National Academy Press, Washington, DC.

Notestein, F.W., 1945. Population in the long view. In: Schultz, T.W. (Ed.), Food for the World. University of Chicago Press, Chicago.

Oliveira, J.M. de, 2013. The Value of Children: Intergenerational Transfers, Fertility, and Human Capital, Ph.D. Dissertation. Yale University, New Haven, CT.

Qian, N., 2008. Women and the price of tea in China: the effect of sex specific earnings on sex imbalance. Quarterly Journal of Economics 123, 1257–1285.

Richards, T., 1977. Fertility decline in Germany: an econometric appraisal. Population Studies 31 (3), 537–553.

Rosenzweig, M.R., Wolpin, K.I., 1980. Testing the quantity-quality fertility model. Econometrica 48, 227–240.

Schultz, T.P., 1981. Economics of Population. Addison-Wesley, Reading, MA.

Schultz, T.P., 1985. Changing world prices, women's wages and the fertility transition: Sweden, 1860–1910. Journal of Political Economy 93 (6), 1126–1154.

Schultz, T.P., 1994. Human capital, family planning and their effects on population control. American Economic Review 83, 255–260.

Schultz, T.P., 1995. Investment in Women's Human Capital. University of Chicago Press, Chicago.

Schultz, T.P., 1997. The demand for children in low-income countries. In: Rosenzweig, M.R., Stark, O. (Eds.), Handbook of Population and Family Economics, vol. 1A. North Holland, Amsterdam, Netherlands. Chap. 8.

Schultz, T.P., April 12, 2013. Women's Work Opportunities, Choice of Job Type, Earnings and Fertility in Ghana: Evidence from 2005/06. Population Association of America Meetings, New Orleans.

Schultz, T.W. (Ed.), 1974. Economics of the Family. University of Chicago Press, Chicago.

Shapiro, D., Tombashe, B.O., 2003. Kinshasa in Transition. University of Chicago Press, Chicago.

Wanamaker, E.A., 2012. Industrialization and fertility in the nineteenth century: evidence from South Carolinas. Journal of Economic History 72, 168–196.

Fertility Transition: Latin America and the Caribbean

Fatima Juárez, El Colegio de México, México D.F., Mexico
Cecilia Gayet, FLACSO México, México D.F., Mexico

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G. Rodríguez, volume 8, pp. 5584–5588, © 2001, Elsevier Ltd.

Abstract

In Latin America and the Caribbean, when mortality declined, the population growth rate rapidly increased, reaching levels of 3.4% per year, and then the onset of fertility initiated. Between the 1960s and 1970s, there was a shift in the trend of total fertility rates in all countries. The fertility decline has been attributed to diverse factors, such as improvement in women's education and their role in the labor force. A characteristic of the region is that contraceptive use increased, but remains low among young people.

Fertility Trends in Latin America and the Caribbean

The continued decline in mortality in Latin America and the Caribbean in the early twentieth century was not accompanied by an immediate decline in fertility in most countries. Latin American governments, concerned with the large population growth, fueled policies to reduce fertility. One of the ideas at the time was that in order to achieve the development of the countries, it was necessary to slow population growth. In contrast, others felt that economic and social development of the countries would lead to a reduction in fertility. In general, most countries initiated family planning policies to achieve a fertility decline, without waiting for the prerequisite of social development to take place. At that time, it was believed that in order to achieve development of the countries, it was necessary to slow population growth. In contrast, others felt that economic and social development of the countries would lead to a reduction in fertility. In general, most countries initiated family planning policies to achieve a fertility decline, without waiting for the prerequisite of social development to take place. Thus, for example, in the 1970s, Mexico presented an annual population growth rate of 3.4%, with a doubling time of 20 years (Ordorica, 2010, 2005), which led to the implementation of a vigorous contraception program. The 1960s and 1970s showed a shift in the trend of the global fertility rates of countries in the region, whose decline far exceeded expectations (CEPAL, 2012; see **Figure 1**). The demographic transition in Latin America occurred in an accelerated manner that led to drastic changes in the age structure of populations and gave way to the current aging process (Livi Bacci, 2012; Tapinos, 1990). The global fertility rate for the period 2005–10 was less than two children per woman in 11 of the 38 countries of Latin America and the Caribbean (United Nations, 2014; see **Figure 2**). Projections indicate that by 2025, half of the countries in the region will have a fertility rate

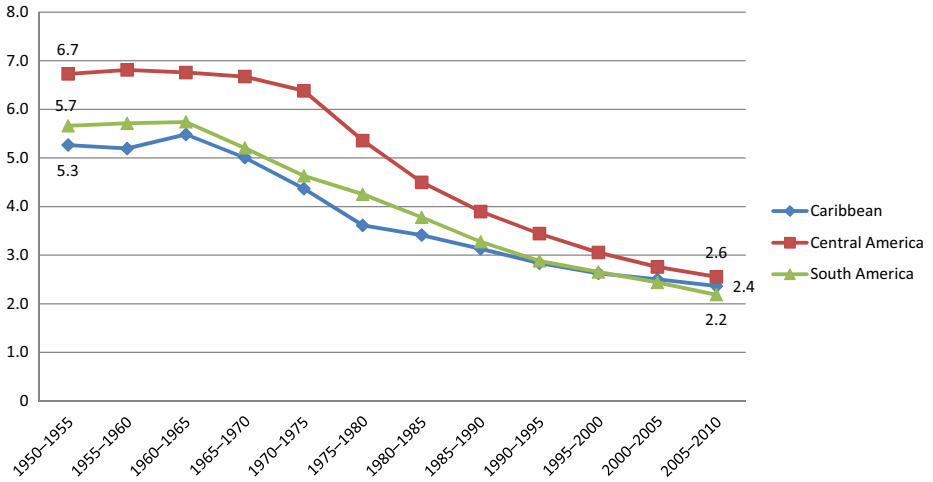

	1950–1955	1955–1960	1960–1965	1965–1970	1970–1975	1975–1980	1980–1985	1985–1990	1990–1995	1995–2000	2000–2005	2005–2010
Caribbean	5.3	5.2	5.5	5.0	4.4	3.6	3.4	3.1	2.8	2.6	2.5	2.4
Central America	6.7	6.8	6.8	6.7	6.4	5.4	4.5	3.9	3.4	3.1	2.8	2.6
South America	5.7	5.7	5.7	5.2	4.6	4.3	3.8	3.3	2.9	2.7	2.4	2.2

Figure 1 Total fertility (children per woman) in Latin America and the Caribbean. Source: United Nations, Department of Economic and Social Affairs, Population Division, 2014. World Population Prospects: The 2012 Revision. http://esa.un.org/wpp/ (accessed 12.03.14.).

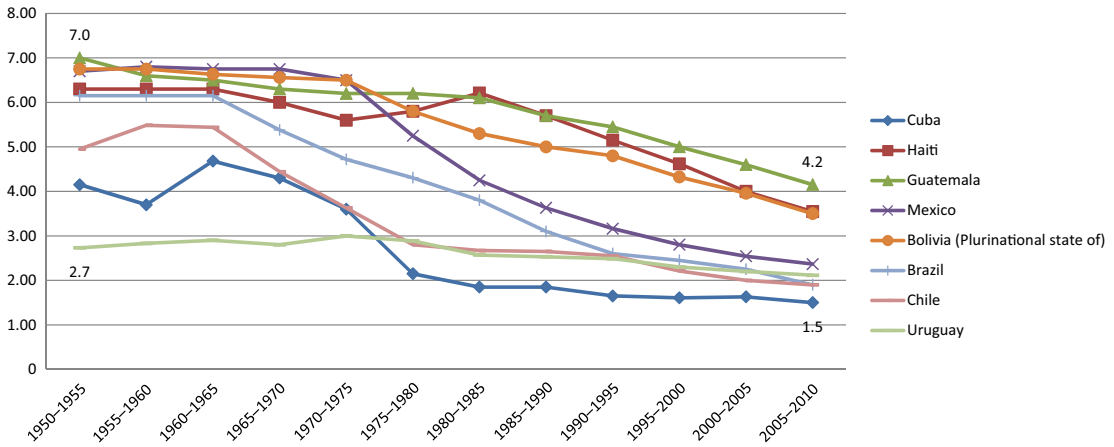

	1950–1955	1955–1960	1960–1965	1965–1970	1970–1975	1975–1980	1980–1985	1985–1990	1990–1995	1995–2000	2000–2005	2005–2010
Cuba	4.15	3.70	4.68	4.30	3.60	2.15	1.85	1.85	1.65	1.61	1.63	1.50
Haiti	6.30	6.30	6.30	6.00	5.60	5.80	6.21	5.70	5.15	4.62	4.00	3.55
Guatemala	7.00	6.60	6.50	6.30	6.20	6.20	6.10	5.70	5.45	5.00	4.60	4.15
Mexico	6.70	6.80	6.75	6.75	6.50	5.25	4.25	3.63	3.16	2.80	2.54	2.37
Bolivia (Plurinational state of)	6.75	6.75	6.63	6.56	6.50	5.80	5.30	5.00	4.80	4.32	3.96	3.50
Brazil	6.15	6.15	6.15	5.38	4.72	4.31	3.80	3.10	2.60	2.45	2.25	1.90
Chile	4.95	5.49	5.44	4.44	3.63	2.80	2.67	2.65	2.55	2.21	2.00	1.90
Uruguay	2.73	2.83	2.90	2.80	3.00	2.89	2.57	2.53	2.49	2.30	2.20	2.12

Figure 2 Total fertility (children per woman) in LA & C. Selected countries. Source: United Nations, Department of Economic and Social Affairs, Population Division, 2014. World Population Prospects: The 2012 Revision. http://esa.un.org/wpp/ (accessed 12.03.14.).

below replacement level (United Nations, 2014). Considering the information of recent surveys, it is also possible to argue that this projection will be exceeded in some countries, where the actual decline rate is even greater than expected. For example, the National Demographic and Health Survey of 2010 from Colombia reports a global fertility rate for that year of 2.1, while the UN projections have estimated a rate of 2.3 for the period 2010-15.

The factors related to fertility decline are diverse. On the one hand, social development in these countries, including steadily rising levels of women's education and their increased participation in the labor market, has played an important role. Also, the urbanization process experienced by Latin American countries has been important as it has led to improvements in access to health services and the expansion of mass media. With regards to schooling, between 2000 and 2007, the gross enrollment rate in secondary education increased from 83% to 89% and tertiary education accounted for the largest increase in comparison with other regions of the developing world, going from 23% to 34% between those years (Altbach et al., 2009). Barro and Lee (2010) estimate that in Latin America and the Caribbean, the average years of schooling of women over age 15 increased from 2.38 in 1950 to 8.21 in 2010. The participation of women in the labor market has also been growing over time. Around 1960, it was estimated that in some countries in the region, the labor participation rate of women aged 15–44 years was, on average, between 25% and 35% (Elizaga, 1974), by 2000 it was estimated to be 48%, and by 2010, 53% (World Bank, 2012). With regards to urbanization, it is noted that the percentage of urban population increased from 50% in the early 1960s to 79% by 2010 (World Bank, 2012). Births attended by medical personnel increased from 75% to 86% between 1990 and 2000, and as an example of the expansion of mass media, it has been estimated that the proportion of the Internet users in the region grew from zero in 1990 to 43% in 2012 (World Bank, 2012).

On the other hand, other aspects that have also played an important role in modifying fertility levels are the age at marriage (both formal and consensual union) and first birth, the increasing use of contraceptives, and changes in reproductive ideals. Although, as previously mentioned, all Latin American countries show a trend of decline in fertility rates, in some of them the rates have not declined at the same pace and level, as is the case in Guatemala, Haiti, and Bolivia (Plurinational State of) with 4.2, 3.6, and 3.5 children per woman, respectively. Additionally, there are differences within countries, with the lowest social strata having higher rates of fertility (Juarez et al., 2013a). Moreover, fertility rates have not had the same reduction in all age groups. One characteristic of Latin America and the Caribbean has been the slow pace of decline in adolescent fertility, which has become a specific research problem and policy focus (CEPAL, 2012). This contrasts with the evidence of a growth in the proportion of women who have not had any children up to age 30 or even at the end of the reproductive period (nulliparity), which shows a polarization in reproductive timing (CEPAL, 2012).

Proximate Determinates of Fertility

According to the approach of Davis and Blake (1956), there exists a set of factors, which have a direct impact on fertility.

Thus, changes in the age of sexual initiation may shorten or lengthen the time of exposure to risk of pregnancy, and its impact on fertility will depend on the use or the lack of usage of contraceptives. The situation in this respect is different in various countries of the region. In some, an early age at first intercourse has been observed, such as in Colombia, while in others it seems that, in general, there have been no major changes, as in the case of Mexico (Gayet et al., 2013). Compared with the experience of European countries, sexual onset in Latin America is earlier in most countries. The proportion of women aged 15–19 who have had sex in a group of countries with recent DHS surveys ranges from 10% in Guatemala to 50% in Colombia (Anderson et al., 2013). If the timing of sexual initiation, the onset of marriage, and age at first birth in countries with recent surveys are analyzed, it can be seen that the proportion of women aged 20–24 who started their sexual life before age 16 and before age 18 is considerable in Colombia and Guatemala (**Table 1**). In Colombia, sexual initiation is clearly separated from marriage, which is not so in other countries where fertility is still high, such as Guatemala. Before turning 18, 60.4% of women in Colombia had started sexual activity, while only 23% had married. In contrast, in Guatemala, 46.6% have had sex before age 18, compared to 37.9% who had married. The distances between the proportion of those sexually active and those married are markedly different. In turn, the proportion of those who had a child before that age is in some cases close to the proportion who had married, and quite large in countries like Guatemala and Honduras, where women who had a child before age 18 represent 27.1% and 22.2% of the women aged 20–24, respectively. This shows that there are different profiles between countries, but a common feature is early marriages and fertility, unlike what has been observed in European countries.

Table 1 Percentage of women aged 20–24 who have had first sexual intercourse, first marriage, and first birth before age 16 and age 18 in selected countries

% Of women 20–24	*Countries*					
	Colombia	*Mexico*	*Peru*	*Honduras*	*Bolivia*	*Guatemala*
Before age 16						
First sexual encounter	29.2	10.7	14.7	20.2	15.6	23.5
First marriage	10.0	6.5	7.1	15.1	7.6	19.1
First birth	5.7	3.2	3.5	6.1	5.2	7.6
Before age 18						
First sexual encounter	60.4	29.6	34.3	44.2	40.0	46.6
First marriage	23.0	18.3	17.5	33.6	21.7	37.9
First birth	19.7	13.1	13.8	22.2	20.0	27.1

Source: Juarez, F., Gayet, C., Mejía, G., 2013a. New and Old Paradigms on Fertility and Reproductive Health in Latin America. Paper presented in XXVII IUSSP International Population Conference, Busan, South Corea, 29.08.2013., estimated based on recent DHS and ENADID 2009 for Mexico.

The levels of modern methods of contraception use in the region are not low. For those countries with recent surveys, of women aged 15 to 24 who were married at the time of the interview, about 50% were using a modern method (**Figure 3**). Bolivia and Guatemala standout with lower levels and a significant proportion of women who were not using any method at all. In addition to the overall analysis of modern contraceptive methods, a key element is the type of method being used by women. In the literature, various studies have indicated that the use of methods that have daily intake or application, either before or during intercourse (e.g., pills, or barrier methods such as diaphragm and condoms), are less effective than long-term reversible methods. With the exception of Mexico, in the other countries surveyed the most used methods among married women aged 15–24 are injectables followed by tablets. Colombia stands out because young married women make greater use of long-term reversible methods than others (21% injections, implants 8.4%, and 6.8% IUD). Also in Mexico, the prevalence of IUD used among married women is notable (18.1%). Despite being young women, an important proportion has been sterilized in Colombia and Mexico, 7.0 and 4.2%, respectively (Juarez et al., 2013a).

An appropriate indicator to analyze the effectiveness of public policies is the unmet need for contraception. This number reflects the proportion of women who do not want

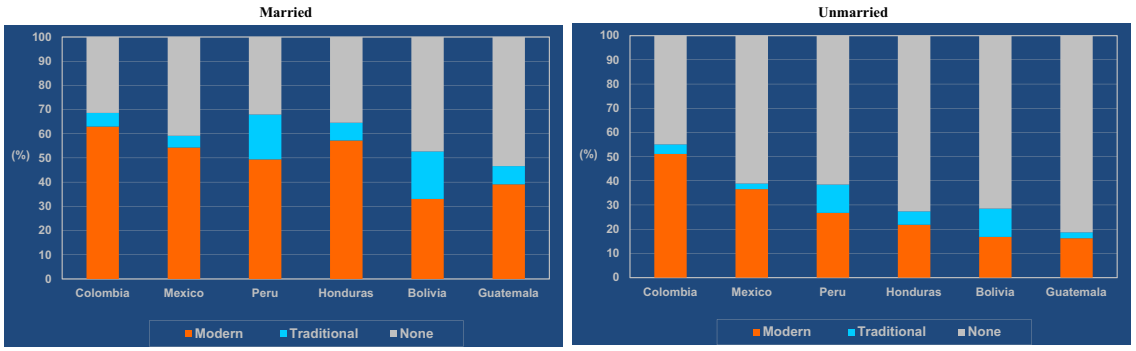

Married	Colombia	Mexico	Peru	Honduras	Bolivia	Guatemala
Modern	63.0	54.3	49.4	57.2	33.0	39.1
Traditional	5.7	4.8	18.5	7.3	19.7	7.5
None	31.4	40.9	32.1	35.5	47.3	53.4

Unmarried	Colombia	Mexico	Peru	Honduras	Bolivia
Modern	51.1	36.6	26.8	21.9	16.8
Traditional	4.0	2.3	11.7	5.5	11.7
None	44.9	61.1	61.5	72.6	71.5

Figure 3 Current contraceptive use among married and never-married women aged 15–24 years, by types of methods, in selected countries.

Table 2 Proportion of married women aged 15–24 with unmet needs for contraception in selected countries

Unmet need	*Colombia*	*Mexico*	*Peru*	*Honduras*	*Bolivia*
To space	10.3	14.8	10.1	10.6	17.7
To limit	3.7	10.4	3.1	2.3	12.3
Total	14.0	25.2	13.2	13.0	30.0

Source: Juarez, F., Gayet, C., Mejía, G., 2013a. New and Old Paradigms on Fertility and Reproductive Health in Latin America. Paper presented in XXVII IUSSP International Population Conference, Busan, South Corea, 29.08.2013., estimated based on recent DHS and ENADID 2009 for Mexico.

children and who are not using contraceptive methods to prevent pregnancy. In Mexico and Bolivia, 25% of young married women do not use the contraceptives they need (Table 2).

Sexually active, never-married young women are the ones who most require the use of contraception; however, their levels of contraceptive use are lower than among married women. For example, in the case of Colombia, 51.2% of never-married women aged 15–24 use modern contraception as compared to 63.0% of married women in the same age group. In Guatemala, 16.3% of never-married young women use modern methods, against 39.1% of married women. The case of Mexico stands out, a country of average development, where only 36.6% of young never-married women use modern methods compared to 54.3% of married women (Figure 3). One explanation is that health services are better prepared for married women, a situation that leads to young unmarried women having unintended pregnancies. The patterns of use of different modern contraceptives by young never-married women are quite different than that of those who are married. Condoms are the most used method in every country, followed by a much lower level of injections and pills (Juarez et al., 2013a). This widespread condom use occurs in these countries especially due to prevention programs of HIV/AIDS, and it should be noted that it is not the most effective method to prevent pregnancy, given that its use is not always consistent, which means that the responses to condom use would overestimate the use of modern methods.

Both, the lack of contraceptive use or the use of less effective methods, result in unplanned pregnancies and abortions. The level of abortions in Latin America and the Caribbean is the highest among the regions of the world, and most of the abortions are unsafe abortions (Sedgh et al., 2012). Estimates of induced abortion rates in some Latin American countries show that the level is even higher than 30 per 1000 women aged 15–44 years, much higher than that of some European countries (Sedgh et al., 2012; Juarez et al., 2013b). There exist a limited number of countries that have relatively reliable information on estimates of induced abortion by age of women. In a recent publication, it can be observed that in the case of Mexico, women of younger ages are those with the highest rates of abortion, reaching 55 abortions per 1000 women in the age group 20–24 years (Juarez et al., 2013b). In countries where fertility has declined and the motivation to have fewer children is strong, and contraceptive methods are not available or they are inconsistently used, the induced abortion rates are higher.

Focusing on Reproductive Rights

Finally, it is important to note that since a few decades ago the studies of fertility in Latin America and the Caribbean, as well as the provision of health services have been framed within a human rights perspective. Some countries since the 1980s and most of them later following the International Conference on Population and Development convened by the United Nations in Cairo in 1994, have since focused their contraception programs in the context of the right to reproductive health, leaving behind the exclusive emphasis on fertility in decline. The Plan of Action agreed at the conference included the right to universal access to reproductive health, in particular family planning and sexual health, with emphasis on equity and equality between the sexes. Also, a focus on reducing maternal and infant mortality was included. Furthermore, the proposed approach considers the economic, social, and cultural development for achievement of reproductive health. This perspective changed the study of fertility that in the past was only considered in terms of population growth and gave way to a broader conceptualization that included the measurement of unwanted fertility and unmet need for contraception in the context of the right to health. Countries have made progress in implementing the Action Plan, but there is still the challenge of putting into practice these rights, especially for the population not in union.

See also: Fertility Theory; Fertility Transition: China; Fertility Transition: Cultural Explanations; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Fertility Transition: Middle East and North Africa; Fertility Transition: South Asia; Fertility Transition: Southeast Asia; Fertility Transition: Sub-Saharan Africa; Fertility Trends in the Formerly Socialist Countries of Europe; Fertility and Culture: Anthropological Insights; Second Demographic Transition.

Bibliography

Altbach, P.G., Reisberg, L., Rumbley, L.E., 2009. Trends in Global Higher Education: Tracking an Academic Revolution. A Report Prepared for the UNESCO 2009 World Conference on Higher Education. United Nations Educational, Scientific and Cultural Organization, Paris. http://unesdoc.unesco.org/images/0018/001831/183168e.pdf (accessed 12.03.14.).

Anderson, R., Panchaud, C., Singh, S., Watson, K., 2013. Demystifying Data: A Guide to Using Evidence to Improve Young People's Sexual Health and Rights. Guttmacher Institute, New York.

Barro, R.J., Lee, J.W., 2010. A New Data Set of Educational Attainment in the World, 1950–2010. In: NBER Working Paper Series. Working Paper 15902. http://www.nber.org/papers/w15902.

CEPAL, 2012. Panorama social de América Latina 2011. Capítulo II. Panorama actual y perspectivas futuras de la fecundidad en América Latina, pp. 81–105. http://www.cepal.org/publicaciones/xml/1/45171/PSE2011-Panorama-Social-de-America-Latina.pdf (accessed 12.03.14.).

Davis, K., Blake, J., 1956. Social structure and fertility: an analytical framework. Economic Development and Cultural Change 4, 211–235.

Elizaga, J., 1974. The Participation of Women in the Labour Force of Latin America: Fertility and Other Factors, 109 Int'l Lab. Rev. 519 1974.

Gayet, C., Juarez, F., Bozon, M., 2013. Sexual practices of latin america and the caribbean. In: Baumle, A. (Ed.), International Handbook on the Demography of Sexuality, International Handbooks of Population, 5. Springer, pp. 67–90.

Juarez, F., Gayet, C., Mejía, G., 2013a. New and Old Paradigms on Fertility and Reproductive Health in Latin America. Paper presented in XXVII IUSSP International Population Conference, Busan, South Corea, 29.08.2013.

Juarez, F., Singh, S., Maddow-Zimet, I., Wulf, D., 2013b. Unintended Pregnancy and Induced Abortion in Mexico:Causes and Consequences. Guttmacher Institute, New York. http://www.guttmacher.org/pubs/Unintended-Pregnancy-Mexico.pdf (accessed 12.03.14.).

Livi Bacci, M., 2012. Historia Mínima de la Población Mundial. Ariel, Barcelona.

Ordorica, M., 2010. Las proyecciones de la población hasta la mitad del siglo XXI. In: García, B., Ordorica, M. (Eds.), Los grandes problemas de México, vol. I. Población, pp. 29–51.

Ordorica, M., May–August 2005. La población, sus ondas y su momentum demográfico. Boletín de los Sistemas Nacionales Estadístico y de Información Geográfica 1 (1), 18–28.

Sedgh, G., Singh, S., Shah, I.H., Åhman, E., Henshaw, S.K., et al., 2012. Induced abortion: incidence and trends worldwide from 1995 to 2008. Lancet 379, 625–632.

Tapinos, G., 1990. Elementos de demografía. Espasa Calpe, Madrid.

United Nations, Department of Economic and Social Affairs, Population Division, 2014. World Population Prospects: The 2012 Revision. http://esa.un.org/wpp/ (accessed 12.03.14.).

World Bank, 2012. World Development Indicators 2012. http://data.worldbank.org/sites/default/files/wdi-2012-ebook.pdf (accessed 12.03.14.).

Fertility Transition: Middle East and North Africa

John B Casterline, Institute for Population Research, The Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H.M. Yousif, volume 8, 5588–5593, © 2001, Elsevier Ltd.

Abstract

Fertility in the Middle East and North Africa declined by more than one-half in the period from 1960 to 2010, most rapidly in the two decades from the 1970s to the 1990s. At present seven of the 22 countries have fertility below replacement level (two births per woman), while the fertility rate exceeds three births per woman in six countries. Changes in marriage patterns have made an unusually large contribution to fertility decline in this region, roughly matching the contribution of improved birth control within marriage. Further decline will depend on future trends in marriage and the availability of the most effective methods of birth control (sterilization and induced abortion).

Introduction

The 22 countries considered here stretch from Western Sahara on the Atlantic coast of Africa to Iran in western Asia. The countries, and their estimated population sizes in 2010, are shown in Table 1. Nineteen of the 22 countries are commonly regarded as Arab states, as indicated by their membership in the Arab League. Some of these Arab countries contain substantial subgroups that are non-Arab according to criteria such as past and present culture (including language), historical patterns of migration and marriage, and/or self-identification. These include Berber population in Morocco and Algeria, Kurdish population in Iraq (and in the non-Arab countries of Iran and Turkey as well), and Asian labor migrants in the Persian Gulf states. Three of the countries are non-Arab: Iran, Israel, and Turkey. Note that Israel contains a large Arab minority population, however. With the exception of Israel, the majority of the population in all countries of the Middle East and North Africa is Muslim, but there are large Christian minorities in Egypt and in most of the countries in West Asia.

Table 1 Listing of countries and estimated population size (in millions) in 2010

Country	*Population*	*Country*	*Population*
North Africa		*West Asia*	
Algeria	37.1	Bahrain	1.2
Egypt	78.1	Iran	74.5
Libya	6.0	Iraq	31.0
Morocco	31.6	Israel	7.4
Sudan	35.6	Jordan	6.5
Tunisia	10.6	Kuwait	3.0
Western Sahara	0.5	Lebanon	4.3
		Oman	2.8
		Palestine	4.0
		Qatar	1.8
		Saudi Arabia	27.2
		Syria	21.5
		Turkey	72.1
		United Arab Emirates	8.4
		Yemen	22.8

Source: UN Population Division: World Population Prospects, 2012 Revision.

In 2010, the total population of these 22 countries was estimated as 488 million persons, with almost one-half (46%) in the three demographic giants: Egypt, Iran, and Turkey. A large fraction of the population in this set of countries resides along the Mediterranean Sea in northern Africa and western Asia. There are also substantial population concentrations in Morocco on the Atlantic Ocean, Iraq in the Mesopotamian River plain, Turkey on the Anatolian plateau and eastern mountains, Iran in the plateaus and mountains of southwest Asia, and Saudi Arabia with population along both the Persian Gulf and the Red Sea. Throughout this geographic expanse annual rainfall is relatively light, and this explains the concentration of population along seas and rivers. This climatic factor also helps account for the high level of urbanization as rural settlements engaged in intensive agriculture is feasible on a large scale in only a few locations (Atlantic coast of North Africa, Nile River Valley, Mesopotamia, and Anatolian plateau).

Population dynamics in the Middle East and North Africa during the past century have been intertwined with political upheaval and uneven economic development (the discussion in this paragraph and the next draws on Courbage (1999), Fargues (1989), and Tabutin and Schoumaker (2005)). Most of the countries were subject to European imperial rule, either directly or through externally imposed local political regimes, beginning in the nineteenth century (Egypt) and continuing until a decade or so after World War II. The breakup of the Ottoman Empire at the end of World War I was especially consequential for the political history of Western Asia, with reverberations to the present (Iraq, Palestine, and Syria). The nationalism of the 1950s and 1960s resulted in generous social policies and programs in some countries (e.g., Egypt), with consequences for demographic behavior (fertility and migration). In the period since 2000, the region has experienced excessive civil strife – Iraq, Palestine, and Syria, with the strife in Palestine a continuation of conflict which began in the late 1940s. Note that the extreme civil strife in these countries spills over into neighboring countries, thereby affecting a large fraction of the population of the region (all three of the demographic giants have been directly affected by conflict in nearby countries). The period since 2010 has also witnessed historically significant internal political upheaval – the 'Arab Spring' – which began in Tunisia and has rippled through the region. The ultimate outcome of this upheaval is not yet

known, in particular the longer-term political influence of Islamist parties whose programs often include explicit pro-natalist prescriptions about family life.

As for economic development, the oil wealth of the Gulf States has had a profound impact on the entire region from the 1970s to the present. Non-oil-producing states have benefited from oil wealth via labor migration and tourism (i.e., holiday visits by Gulf State nationals). More generally, international migration is a dominant feature of the landscape throughout the region. This includes intra-regional labor migration to the Persian Gulf states and northward migration to western Europe, especially from the Maghreb countries in northern Africa (Morocco, Algeria, and Tunisia). The Middle East and North Africa have also participated in the global surge in formal schooling in recent decades. The high rates of women's secondary and university schooling in post-1980 birth cohorts represent a sharp departure from the past, with implications for marriage and childbearing.

Fertility Decline to Date

Overview

The level of fertility in these 22 countries in the period from the early 1950s to the present is summarized in Table 2. In this table the measure of fertility is the total fertility rate (TFR), which is the sum of age-specific fertility rates for women of reproductive age during the specified historical period. The TFR can be viewed as the average number of births at the termination of a woman's reproductive career, were this career characterized by childbearing at the rates observed during the historical period in question. That is, this is a lifetime total for a hypothetical (or 'synthetic') cohort, not an actual birth cohort of women (whose reproduction would extend over three historical decades).

In the early 1950s, when the region had largely recovered from the severe disruption caused by World War II, almost one-half of the countries in the region had TFRs which exceeded seven births per woman, and only two countries (Israel at 4.3 and Lebanon at 5.7) had TFRs below six births per woman. The average for the region was roughly seven births per woman. With the exception of Israel and Lebanon (TFR = 5.7), these can be regarded as pre-decline levels of fertility which had obtained, with slight variation, for many generations in the past. Available evidence indicates that pre-decline fertility in the Middle East and North Africa was higher than in other major regions, including other subregions of Asia, Sub-Saharan Africa, and Latin America and the Caribbean (Casterline, 2011; Rashad, 2000). This coincides with other quantitative and qualitative evidence of the fundamental pro-natalism of these societies in the past: relatively early and nearly universal marriage, short postpartum abstinence, and a high valuation placed on children and childbearing.

Six decades later in 2010–14, 16 of the 22 countries have TFRs below three births per woman, and the average fertility for the region is less than three births per woman, a decline in the level of fertility from the early 1950s of more than one-half. It is also evident from Table 1 that in the majority of countries the decline was concentrated in the two decade period from the early 1970s to the early 1990s, although to be sure further substantial decline has occurred from the early 1990s to the present. From a comparative perspective, fertility decline in this region was late to start; in the early 1970s, only five countries had TFRs below six births per woman (Bahrain, Egypt, Israel, Lebanon, and Turkey), and only two of these (Israel and Lebanon) were below five births per woman. Among major regions, only Sub-Saharan Africa experienced later onset of fertility decline. Through the 1960s and 1970s (and even into the 1980s), the Arab region appeared to be very reluctant to join in the gathering international movement toward lower levels of fertility, a fact frequently commented upon in the demographic literature and in policy forums; see Obermeyer (1992).

Once commenced, the declines proceeded rapidly – as summarized in Table 3. During the 1980s, when the pace of decline was most rapid, 11 of the 22 countries experienced a decline in their TFRs that exceeded 1.5 births per woman. Of the more populous countries, only Sudan and Yemen were not caught up in the rapid fertility decline in this period (indeed, Yemen's TFR is thought to have increased from the early 1970s to the early 1990s).

As of 2010–14, seven countries in the Middle East and North Africa have TFRs which are equal to or below replacement level of 2.1 births per woman; these countries are Bahrain, Iran, Lebanon, Qatar, Tunisia, Turkey, and United Arab Emirates. The low fertility in the Persian Gulf States (Bahrain, Qatar, and United Arab Emirates) is striking. At the other end of the spectrum, four countries retain TFRs above four births per woman; these countries are Iraq, Palestine, Sudan, and Yemen. Each of these countries has its own distinctive political and demographic history, but it is difficult to ignore the fact that three of the four (excepting Yemen) have experienced severe civil strife in recent decades. Political demography is more tangible in this region than most (Fargues, 2000; Khawaja, 2003).

Table 2 Distribution of country-level TFR by historical period

TFR	*1950–55*	*1970–75*	*1990–95*	*2010–14*
>=7.0	10	9	1	0
6.0–6.9	10	8	3	0
5.0–5.9	1	3	3	0
4.0–4.9	1	1	3	4
3.0–3.9	0	1	8	2
2.2–2.9	0	0	3	9
<=2.1	0	0	1	7
Total	22	22	22	22

Source: UN Population Division: World Population Prospects, 2012 Revision.

Table 3 Rate of decline in TFR: births decline per decade

TFR	*1960s*	*1970s*	*1980s*	*1990s*	*2000s*
TFR increase	8	3	0	1	3
0.00–0.49	6	2	2	1	8
0.50–0.99	5	8	4	7	10
1.00–1.49	3	8	5	7	0
>=1.50	0	1	11	6	1
Total	22	22	22	22	22

Source: UN Population Division: World Population Prospects, 2012 Revision.

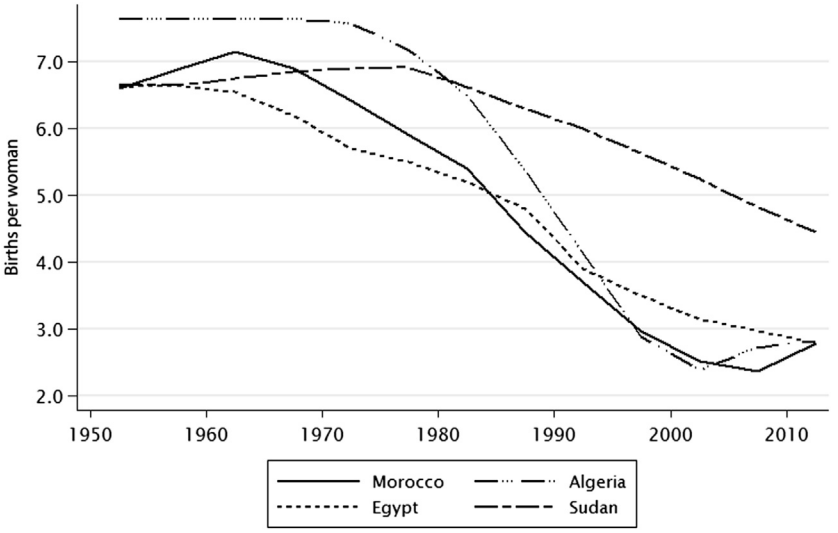

Figure 1 Fertility decline in populous North African countries. Source: UN Population Division.

Fertility Decline by Country

There is considerable inter-country diversity in fertility decline – both the historical timing of onset and the pace of decline. This becomes clear from examining trajectories country-by-country. Fertility trends for 12 of the 22 countries are shown in Figures 1–3.

The fertility declines in North Africa (Figure 1) have several interesting features. The declines in Algeria and Morocco were rapid, especially in Algeria which from the mid-1970s to mid-1990s experienced one of the fastest declines on record (four-birth decline in merely 20 years). By comparison, the declines in Egypt and Sudan have proceeded more slowly, especially in Sudan which after 1990 has among the highest levels of fertility in the region. A final intriguing feature of Figure 1 is the increase in fertility observed in both Algeria and Morocco in the period after 2000; in both countries, the estimate for the most recent period (2010–14) has risen to nearly three births per woman after declining to around 2.5 births per woman in the early 2000s. This recent increase has surprised demographers, who fully expected fertility in both countries to continue downward to replacement level or below, following the pathway of most countries once rapid and sustained decline is underway. The Algerian case has been examined in some detail by Ouadah-Bedidi and Vallin (2013), who conclude that the rise is principally due to a fall in the average age at first marriage. Note that fertility increase in both countries preceded the political upheaval of the Arab Spring.

The same diversity is evident in the populous countries of West Asia. Iran experienced a spectacularly rapid fertility decline from the mid-1980s to the early 2000s, following a slight rise after the 1979 Revolution (the new government

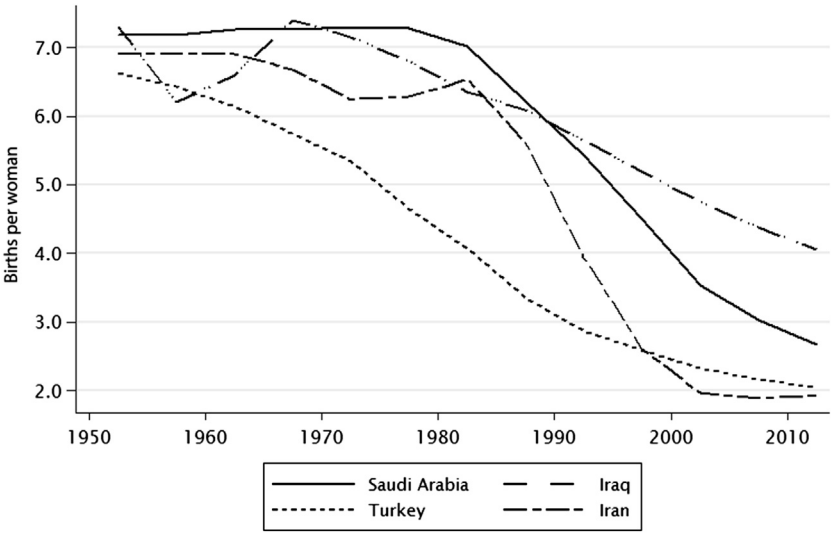

Figure 2 Fertility decline in populous West Asian countries. Source: UN Population Division.

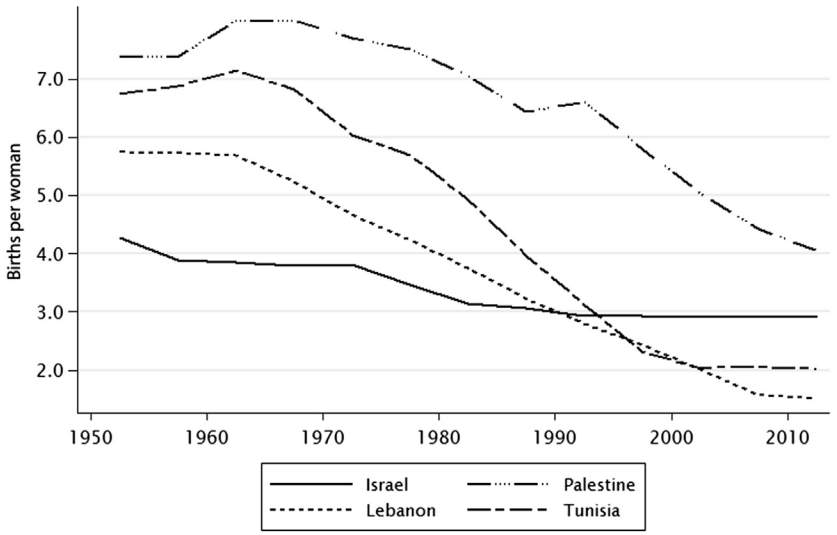

Figure 3 Fertility decline in select countries. Source: UN Population Division.

was initially explicitly pro-natalist). Abbasi-Shavazi et al. (2009) provide a comprehensive demographic portrait of this exceptional decline. Less remarked on is the substantial and relatively rapid fertility decline in Saudi Arabia in the period since 1980. In the most recent period fertility is still well above replacement level (the TFR estimate for 2010–14 is 2.7), so it remains uncertain whether fertility in Saudi Arabia will continue downward to replacement level (as has occurred in Iran and Turkey; the latter via slow and steady decline from the early 1950s to the present) or settle at a level above replacement or even increase (as has occurred in Algeria and Morocco). Iraq has experienced slow decline since the late 1960s; the UN estimates notwithstanding, its fertility patterns during the past decade of extreme civil unrest are murky due to disruption of demographic data collection.

Figure 3 shows four countries that are, in one respect or another, idiosyncratic. Israel has a majority Jewish population, and the majority of them are first- or second-generation emigrants from Europe or North America. Hence it is not surprising that Israel's level of fertility was markedly lower than other countries in the region through the 1950s and 1960s. What is perhaps unexpected is the stability of Israel's TFR at around three births per woman in the period from the 1980s to the present. This national-level TFR is an average of widely varying levels within the Israeli population, with sub-replacement fertility among the more secular and European segments offset by high fertility among Orthodox Jews (Fargues, 2000). Fertility in Palestine, by contrast, was quite high throughout the period from the early 1950s to the 1990s, and although it declined substantially after 1990 it remains moderately high up to the present (TFR = 4.0 in 2010–14) (Khawaja, 2003). Apart from Israel, Lebanon is usually regarded as the forerunner for fertility decline in this region, and its trajectory in **Figure 3** supports this. In the most recent period Lebanon's fertility is far below replacement (TFR = 1.5 in 2010–14), uniquely low among countries in this region. Finally, Tunisia is also often labeled as a fertility decline forerunner (Rashad, 2000), and this is indeed the case by comparison with its North African neighbors Algeria, Egypt, and Morocco (see **Figure 1**). But Tunisia's decline seems neither early nor rapid by comparison with several countries in West Asia (Iran, Lebanon, and Turkey).

Explanation

Demographic

A search for explanations for fertility decline in the Middle East and North Africa begins with demographic analysis of the direct determinants of fertility, namely sexual exposure and active birth control.

With regard to the former, because sexual activity in the region is confined to formal marital unions – probably less so than in the past but to a larger extent than in most other regions (Tabutin and Schoumaker, 2005) – the timing of first marriage and the ultimate proportion marrying have marked effects on fertility levels. Several analyses demonstrate convincingly that nuptiality change has made a substantial contribution to the fertility decline to date in this region (Rashad, 2000; Tabutin and Schoumaker, 2005). Considering the region as a whole, the nuptiality contribution has been on the order of one-half of the fertility decline, with the other one-half due to more effective birth control within marital unions. This breakdown varies across countries; in particular, the nuptiality contribution has been especially large in the Maghreb countries such as Algeria, Morocco, and Tunisia – where the nuptiality change has been most dramatic – and Lebanon as well, and the nuptiality contribution has been of lesser magnitude in Egypt, Iran, and Turkey.

The most direct evidence of the nuptiality change underway are declines in the proportion of women of reproductive age – especially under age 30 – who have ever married. Many of these declines are relatively recent, and hence it is too soon to determine whether the predominant phenomenon is delayed entrance to first marriage or an increase in celibacy (Rashad and Osman, 2003). This is an important distinction, not

necessarily from a mechanistic standpoint (lost reproductive exposure) but rather because of the societal implications, which in turn may indirectly bear on fertility. One might posit that an increase in celibacy is the more revolutionary of the two. If meaningful fractions of women (e.g., 10% or greater) remain unmarried through the reproductive years, almost necessarily this must force an acceptance of primary roles for adult women other than motherhood. To the extent such acceptance occurs, this could affect the formulation of childbearing goals of women in marital unions.

Recent demographic survey data show considerable cross-country variation in marriage patterns (Casterline, 2011). If one considers, for example, the percentage of women married under age 30, the percentage of those ever married at ages 20–24 ranges from 15% in Tunisia to nearly 60% in Palestine and Yemen, while at ages 25–29 the percentages range from below 50% in Algeria, Tunisia, and Lebanon to above 80% in Egypt and Yemen. The difference in reproductive exposure between, say, Algeria (16% married at ages 20–24) and Egypt (54% married at ages 20–24) is enormous. These are the ages at which reproductive unions tend to be most fertile.

Turning to the second major direct determinant of fertility, that is, active birth control, one of the noteworthy developments in the demography of this region in the period since 1960 has been the growing prevalence of contraception as a means of birth control (Tabutin and Schoumaker, 2005). The majority of women live in countries where at least one-half of the currently married women are using contraception (Casterline, 2011). Pursuing this in more detail, one can distinguish contraceptive methods according to their efficacy in preventing pregnancy. This investigation reveals several distinctive features of this region. First, ineffective 'traditional' methods are relatively prevalent in some countries. More specifically, withdrawal is popular in western Asia (Jordan, Lebanon, Syria, and Turkey), as has been evident in all fertility surveys in this subregion since the 1960s. Secondly, small percentages of women are sterilized, with the exception of several Gulf states (Oman and United Arab Emirates) and Iran. This is not surprising, as a common view among Arabs (most of whom are Sunni) is that sterilization is contrary to Islam. This is popular opinion, and not a view actively propounded by Islamic scholars (although it is the case that sterilization is rare in most Islamic countries). Recourse to sterilization to terminate childbearing once a desired number has been achieved is, therefore, regarded as an unappealing solution in most subgroups in this region. That leaves the inter-uterine device (IUD) and the oral contraceptive as the effective methods of choice; both are highly effective when used diligently but subject to error and deliberate discontinuation.

A further method of birth control is induced abortion. Induced abortion is legal under most circumstances in Israel, Tunisia, and Turkey, in limited circumstances (woman's physical and mental health) in most countries in the region, and only if the woman's life is endangered in a few countries (including Egypt and Iran) (United Nations Population Division, 2013b). Whatever the legal standing, there is a common assumption throughout these largely Islamic societies that induced abortion is contrary to Islam. There are no sound estimates of incidence (Tabutin and Schoumaker, 2005), but it can be assumed to be relatively low except in Iran, Lebanon, Tunisia, and Turkey. The incidence of induced abortion is probably highest for pregnancies of unmarried women, but these are also the most strenuously kept secret. In all likelihood, induced abortion has contributed less to fertility decline in the Middle East and North Africa than in other regions. Possibly increased demand for birth control will lead to far more reliance on induced abortion in the future; this possibility cannot be simply dismissed out of hand, especially given the increased availability of medical abortion (which can be procured from a pharmacy and administered privately at home). But to this point there is little indication that social norms against induced abortion are weakening, and, in the absence of concerted political efforts to legalize abortion, access to safe surgical abortion will remain difficult and costly.

Other Factors

Scholars have attributed fertility decline in this region to the constellation of factors posited to cause fertility decline in other major regions in the period since 1950: dramatic improvements in child survival, higher levels of formal schooling (especially for women), income growth and growth of secondary and tertiary sectors of the economy, and the emergence of norms of family life incompatible with high fertility. Existing scholarship contains no rigorous assessments of the relative weight of these and other hypothesized explanatory factors, but a number of scholars have rendered judgments on the basis of quantitative investigation (e.g., Rashad, 2000; Tabutin and Schoumaker, 2005).

From a comparative perspective, the causes of fertility decline in the Middle East and North Africa appear to be distinctive in several respects. First of all, adult women's participation in the formal labor market was relatively low in the past and up to the present is low from a comparative standpoint. This would seem to diminish the likely force of opportunity costs of childbearing, that is, the financial cost to the household of women foregoing income-generating economic activity in order to bear and raise children. Second, the establishment of expansive social welfare policies and programs in many countries in the region in the 1950s and 1960s created an unusual context for fertility decline. These programs included guaranteed employment in the public sector for persons with formal schooling credentials and subsidization of fuel and food. Directly and indirectly, these diffused the household-level cost of bearing and raising children. Arguably it was the weakening of the social welfare state beginning in the 1980s, as it became unaffordable and/or as the neoliberal economic policy came to dominate, which stimulated fertility decline in this region (see Courbage (1994) for the case of Syria).

There is dispute about the causal contribution to fertility decline of the provision of family planning services through public sector programs. Tunisia is often cited as an historically early demonstration of fertility reduction due in part to enhanced availability of good quality family planning services (Brown, 2007). Egypt has supported the largest and longest standing public-sector program in the region; Robinson and El-Zanaty (2006) give it much credit for the fertility decline to date, and Rashad and Eltigani (2005) somewhat less credit. From a demographic standpoint and as compared to other

countries in the region, a larger fraction of the Egyptian decline has been due to change in marital fertility than to change in nuptiality, a fact supportive of a relatively larger contribution of its family planning program. The Iranian decline too has been driven disproportionately by decline in marital fertility, and hence it is natural to credit effective provision of family planning services at the community level throughout Iran beginning in the late 1980s. But Abbasi-Shavazi et al. (2009) place more explanatory weight on shifts in the demand for children due to economic aspirations running ahead of economic opportunities. And the fertility decline in a few countries occurred in the absence of an active public-sector family planning program; Algeria is a notable instance (Ouadah-Bedidi and Vallin, 2013).

Three salient features of the demography of the Middle East and North Africa since the 1950s suggest a distinctive causal structure underlying its fertility decline. The first is the relatively large contribution to the fertility declines in most countries of postponement of entry to first marriage, that is, nuptiality change. This is evidence in support of the argument that increased formal schooling of women has been a decisive factor driving fertility decline. The second demographic feature is the relatively high rates of international migration. It is plausible that this has disrupted existing high-fertility regimes through multiple pathways, including transformation of household economics and economic aspirations, changes in intra-family power structure, and diffusion of new norms about family life. A third salient feature is the preexisting patriarchal character of the family system which directly and indirectly was strongly pro-natalist. Some scholars have argued that undermining this system was a prerequisite for fertility decline in these societies (Courbage and Todd, 2011; Fargues, 2005). For this reason, formal schooling of women may have assumed more importance as a cause of fertility in this region than elsewhere. Other factors noted above, including economic transformation and international migration, also may have served to weaken the patriarchal family system.

Prospects for Future Trends in Fertility

As of 2010–14, seven of 22 countries in the Middle East and North Africa have sub-replacement fertility, while six of the 22 countries have a TFR in excess of three births per woman. What does the evidence in hand suggest about prospects for future fertility change? This might consist of fertility increase in countries with the lowest fertility, and fertility decline in countries where fertility is well above replacement level.

Considering first the subset of countries with sub-replacement fertility, there is only one country with fertility far below replacement (i.e., TFR less than 1.8): Lebanon with a TFR of 1.5. Outside the Middle East and North Africa, the period since 2000 has witnessed substantial rebound in fertility in some countries with sub-replacement fertility, especially in northern Europe (most notably Sweden and the United Kingdom). But the more common pattern in western Europe, southern Europe, and East Asia has been rather a modest increase from TFRs far below replacement level (United Nations Population Division, 2013a). One might posit that Lebanon will adhere most closely to the Mediterranean pattern (Greece, Italy, and Spain) of sustained low fertility (although, to be sure, both Italy and Spain have recovered from TFRs well below 1.5). To this point – historical experience with fertility far below replacement is limited – there is no evidence that fertility below replacement level constitutes an unstable equilibrium.

More consequential is the future of fertility in those countries with recent TFRs in excess of three births per woman. Five of these six countries are in western Asia (Jordan, Iraq, Palestine, Syria, and Yemen), and the sixth is Sudan. Data from demographic surveys provide some basis for assessing the potential for further decline. These surveys routinely gather information on fertility desires. Over the decades, controversy has surrounded the measurement and interpretation of such data. But given that fertility is highly purposive, fertility desires are likely to be informative about fertility trends in the short-term, even recognizing the constraints which couples face in implementing their desires. Elsewhere (Casterline, 2011) the author has examined survey estimates of the level of unwanted fertility. These suggest that unwanted fertility is relatively common throughout the region – the fraction of unwanted births ranges between one-sixth (Lebanon) to one-third (Yemen) – and therefore fertility should decline further if women become more successful in avoiding unwanted births. At the same time, attitudinal data on fertility ideals makes clear that in most Middle East and North African countries younger cohorts of women (under age 30) continue to find families with three or more children appealing. Mean ideal number of children of 3.4 in Algeria, 3.1 in Lebanon, and 4.3 in Syria are especially striking, as each is substantially higher than the recent TFR. Almost no women choose zero or one child as their ideal. These and other attitudinal data are emphatic in documenting weak attachment to small-family ideals (i.e., two children or less) (Casterline, 2011). Other data reveal that most women recognize many advantages of stopping at two children and also perceive various costs of having three or more children. Even so, a large fraction of women – in some countries a plurality, even among the younger cohorts – desire to have three (or more) children. In short, while the attitudinal data reveal the prevalence of motivation in most countries to reduce fertility below existing levels, these data also provide no grounds for projecting that fertility will fall to two births per woman (or lower) in all countries in the region in the near future.

The behavioral factors that bear directly on fertility (see Section Demographic above) can also be examined for indications of the potential for further decline in fertility. There are, first of all, patterns of union formation. It was noted above that nuptiality change has made a substantial contribution to fertility decline in this region; does potential remain for further nuptiality change which would further depress fertility? Casterline (2011) examines the evidence and concludes that nuptiality change in the future has the potential to cause substantially more fertility decline in the region, especially in Egypt (most populous country in the region and characterized by relatively early marriage) and also in Morocco, Syria, and Yemen. And further nuptiality change does indeed seem likely; nuptiality change in the region appears to be driven by the (perceived) unaffordability of marriage, itself a function of heightened expectations about the material resources one

must bring to marriage (including the wedding itself) simultaneous with declines in economic opportunities for young adults (Rashad and Osman, 2003). Both could ease – the increase in expectations and the decline in economic opportunity – but this seems unlikely, at least in the short run.

The second behavioral factor which bears directly on fertility is active birth control. The heavy reliance on ineffective methods in some countries (see Section Demographic) leaves room for fertility reduction as a consequence of shifting the contraceptive method mix toward more effective methods. But the evident resistance to sterilization and induced abortion as a means of birth control, especially in the cluster of countries where fertility remains high, presents an obstacle to decline of fertility to two births per woman (or lower). An important unknown factor is whether medical induced abortion will become a common resolution of unintended pregnancies.

Finally, fertility prospects in Israel and Palestine resist conventional demographic analysis because of the unique confluence of political and demographic dynamics. It is reasonable to posit fertility decline in both countries, but it is not difficult to envision inter- and intra-country political developments which generate pressure toward maintenance of fertility levels substantially above replacement level.

Summing up this assessment of prospects for further decline in fertility, there are factors which facilitate and factors which hinder such further decline. Chief among the former is the potential for further nuptiality decline, and also the possibility of more recourse to medical abortion. Factors working against fertility decline include weak attachment to the two-child norm, and the fact that to this point neither sterilization nor surgical abortion are readily available as methods of birth control in those countries in the region which at present have fertility well above replacement level. Fertility rates at replacement level (or below) are almost never attained without heavy reliance on one or the other of these two methods of birth control.

See also: Children, Value of; Fertility Theory: Embodied-Capital Theory of Life History Evolution; Fertility Theory: Theory of Intergenerational Wealth Flows; Fertility Transition: Cultural Explanations; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations.

Bibliography

Abbasi-Shavazi, M., McDonald, P., Hosseini-Chavoshi, M., 2009. The Fertility Transition in Iran: Revolution and Reproduction. Springer, New York.

Brown, G., 2007. Tunisia: the debut of family planning. In: Robinson, W., Ross, J. (Eds.), The Global Family Planning Revolution: Three Decades of Population Policies and Programs. the World Bank, Washington, pp. 59–70.

Casterline, J., 2011. Fertility Prospects in the Arab World. Population Division Expert Paper No. 2011/6. United Nations, New York.

Courbage, Y., 1994. Fertility transition in Syria: from implicit population policy to explicit economic crisis. International Family Planning Perspectives 20 (4), 142–146.

Courbage, Y., 1999. Economic and political issues of fertility transition in the Arab world—answers and open questions. Population and Environment 20, 353–380.

Courbage, Y., Todd, E., 2011. A Convergence of Civilizations: The Transformation of Muslim Societies around the World. Columbia University Press, New York.

Fargues, P., 1989. The decline of Arab fertility. Population E 44 (1), 147–175.

Fargues, P., 2000. Protracted national conflict and fertility change among Palestinians and Israelis. Population and Development Review 26 (3), 441–482.

Fargues, P., 2005. Women in Arab societies: challenging the patriarchal system? Reproductive Health Matters 13 (25), 43–48.

Khawaja, M., 2003. The fertility of Palestinian women in Gaza, the West Bank, Jordan and Lebanon. Population-E 58 (3), 273–302.

Obermeyer, C., 1992. Islam, women and politics: the demography of Arab countries. Population and Development Review 18 (1), 33–66.

Ouadah-Bedidi, Z., Vallin, J., 2013. Fertility and population policy in Algeria: discrepancies between planning and outcomes. Population and Development Review 38 (Suppl.), 179–196.

Rashad, H., 2000. Demographic transition in Arab countries: a new perspective. Journal of Population Research 17 (1), 83–101.

Rashad, H., Eltigani, E., 2005. Explaining fertility decline in Egypt. In: Jones, G., Karim, M. (Eds.), Islam, the State and Population. C. Hurst & Co., London, pp. 174–198.

Rashad, H., Osman, M., 2003. Nuptiality in the Arab countries: changes and implications. In: Hopkins, N. (Ed.), The New Arab Family, Cairo Papers in Social Science. American University Press, Cairo, pp. 20–50.

Robinson, W., El-Zanaty, F., 2006. The Demographic Revolution in Modern Egypt. Lexington Books, Lanham, MD.

Tabutin, D., Schoumaker, B., 2005. The demography of the Arab world and the Middle East from the 1950s to the 2000s: a survey of changes and a statistical assessment. Population-E 60 (5–6), 505–616.

United Nations Population Division, 2013a. World Population Prospects 2012. United Nations, New York.

United Nations Population Division, 2013b. World Abortion Policies 2013. United Nations, New York.

Fertility Transition: South Asia

John Cleland, University of London, London, UK; and London School of Hygiene & Tropical Medicine, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Between 1950 and 2010, fertility in South Asia fell from about 6 to 2.5 births per woman. In Sri Lanka and South India, this transition started in the late 1950s and childbearing has reached replacement level. By contrast, fertility in Pakistan changed little until the late 1980s and was still high at about 3.5 births in 2010. The most remarkable and unexpected development has been the steep declines in Bangladesh and Nepal, two of the poorest countries in the world. The diversity in the timing and speed of transition cannot be explained in terms of material prosperity. Progressive social policies – resulting in universal schooling, high life expectancy, and state welfare provision – offer a compelling explanation for the early declines in Sri Lanka and Kerala, and gender inequality has been advanced as the major barrier to fertility decline in north India and Pakistan. Vigorous state promotion of family planning has probably accelerated the pace of decline in Sri Lanka, India, and Nepal. In Bangladesh, the family planning program may have triggered the onset of reproductive change, while in Pakistan the program has been ineffective because of weak political support.

Between 1950 and 2010, fertility in South Asia fell from about six to less than three births per woman. In Sri Lanka and South India, this transition started in the late 1950s and childbearing has reached replacement level. By contrast, fertility in Pakistan changed little until the late 1980s and was still well above three births in 2010. The most remarkable and unexpected development has been the steep decline in Bangladesh, one of the poorest countries in the world. The diversity in the timing and speed of transition cannot be explained in terms of material prosperity. Progressive social policies – resulting in universal schooling, high life expectancy, and state welfare provision – offer a compelling explanation for the early declines in Sri Lanka and Kerala, and gender inequality has been advanced as the major barrier to fertility decline in north India, Nepal, and Pakistan. Vigorous state promotion of family planning has probably accelerated the pace of decline in Sri Lanka, India, and Nepal. In Bangladesh, the family planning program may have triggered the onset of reproductive change, while, in Pakistan, the program has been ineffective because of weak political support.

South Asia is defined here as the countries of former British India (India, Pakistan, and Bangladesh) plus Nepal and Sri Lanka. Their combined population was estimated to be 1600 million in the year 2010, equivalent to 23% of the global total. Fertility trends in this region are thus a crucially important determinant of future global population size. In this article, the course of fertility over the past 60 years will be described, attempts to explain changes in reproductive behavior will be reviewed, and prospects for future fertility decline assessed.

Data Sources

Though only Sri Lanka possesses a vital registration system with high coverage of births, there is nevertheless a relative abundance of demographic data sources in South Asia, including decennial censuses and large-scale single-round demographic surveys. In terms of regular demographic monitoring, the outstanding achievement of the region is India's Sample Registration System (SRS). The SRS has yielded annual demographic estimates for most states in India since the early 1970s.

The collection of accurate demographic data in South Asia is difficult and all national data sets are flawed to some extent. This has generated considerable controversy over the timing and magnitude of reproductive change in South Asia. Genuine uncertainty still exists over the details, particularly in Pakistan, where the quality of data is especially poor, but the broad contours of demographic change are well established.

Fertility Trends

The historic level of fertility in South Asia was probably similar to that of Europe and China, namely an average of 5.5–6.0 live births per woman, with fluctuations in the range of 4.5–6.5. Fertility rose during the 1950s and 1960s in some parts of the region, a reflection of declines in widowhood, changes in breastfeeding patterns and sexual behavior within marriage. Fertility peaked at around seven births per woman in Bangladesh and Pakistan in the 1970s but remained about one birth lower in the other three countries.

Sri Lanka was the forerunner of fertility transition (see Figure 1). The onset of decline dates back to the 1950s, and began largely in response to a rising age at marriage (Fernando, 1975). The first appreciable fall in national fertility in India occurred in the 1970s, in Pakistan and Bangladesh in the 1980s, and in Nepal in the early 1990s. Comparing the highest post-1950 fertility level with estimated levels between 2005 and 2010, childbearing has fallen most steeply in Bangladesh and Sri Lanka (66 and 59%, respectively), moderately in India and Nepal (about 50%), and least in Pakistan (about 45%). Undoubtedly, the most remarkable development has been the steep fall in fertility in Bangladesh, which remains one of the world's poorest countries.

These national trends mask geographical, ethnic, and socioeconomic variations in the timing of the onset and speed of reproductive change. In Bangladesh, fertility is and highest in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31079-0

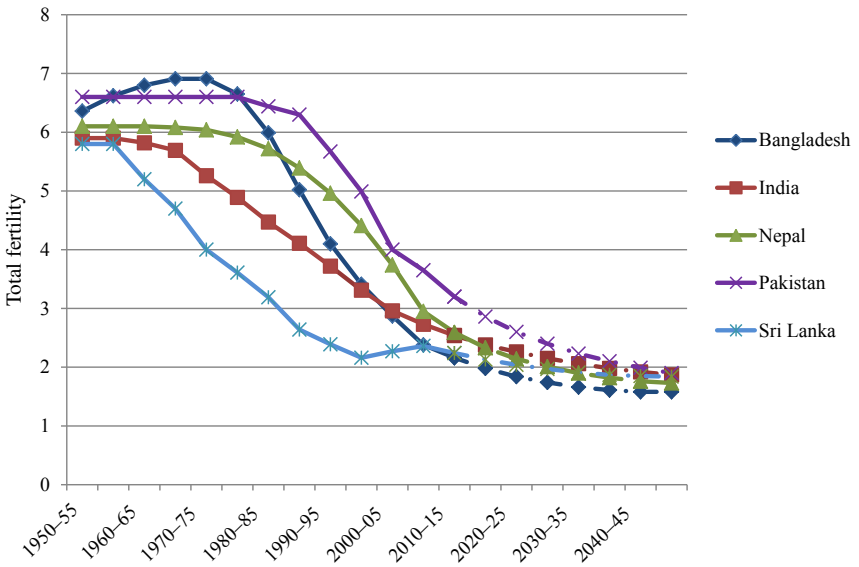

Figure 1 Past and projected trends in total fertility rates in South Asian countries. Source: United Nations, 2011. World Population Prospects: The 2010 Revision. Population Division, New York.

the eastern divisions of Sylhet and Chittagong Division, generally considered to be the most conservative but least poor parts of the country. In Nepal, fertility transition is more advanced in the hill zone than in the Terai (adjacent to India) or in the high-altitude mountain zone. The highest fertility in Pakistan is recorded in the least developed province, Baluchistan.

The diversity of reproductive trends in India is much greater (see Figure 2). In the northerly states of Bihar and Uttar Pradesh, which comprise nearly one quarter of the total population, demographic modernization has lagged far behind the rest of the country. Women's literacy, age at marriage, and life expectancy are all low; fertility has declined moderately but the level of childbearing was about four births per woman in 2005–06. By contrast, in the southern states of Kerala and Tamil Nadu fertility transition began early, probably in the late 1950s (Zachariah and Rajan 1997). According to the SRS, the level of childbearing had already fallen to about four births per woman by the early 1970s. Kerala reached replacement level fertility (i.e., the level required to maintain stability of population size in the long term) of a little above two births by 1988, eventually stabilizing at 1.8 births per woman. Replacement level was reached in Tamil Nadu in 1993. In the other two southern states, Andhra Pradesh and Karnataka, fertility is now at about replacement level.

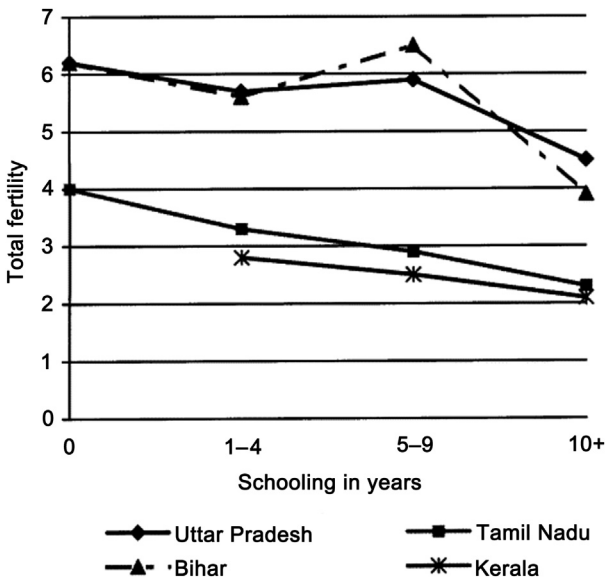

Figure 2 Rural total fertility rates by educational level of women, selected states of India in 1984. Note: Kerala estimate for 0 years of schooling not shown, because too few cases exist. Source: Office of the Registar-General 1990 *Fertility Differentials in India*. Vital Statistics Division, New Delhi.

Determinants of Transition

Direct Causes

South Asian transitions to lower fertility have been driven by two direct causes: a rise in marriage ages, thus reducing teenage childbearing; and increased family size limitation by contraception and abortion. There is less evidence of postponement of live births following marriage, or of changes in birth spacing. Contraception and abortion have been used to limit family size but not to delay first births, nor for birth spacing, except in Bangladesh. Reproduction is increasingly compressed within the age span of 20–30 years.

Throughout most of South Asia, early and universal marriage for women – with a wide gap of 5 to 8 years in the ages of brides and grooms – has been the historic norm. With the exceptions of Sri Lanka and parts of India (e.g., Kerala), the singulate mean age at marriage for women at the start of the twentieth century was about 13 years. (The singulate mean age

is calculated from census or survey data on proportions single by current age: for details see Hajnal, 1953; also Ní Bhrolcháin.) Marriage has remained universal but women's age at marriage has steadily increased. By the 1980s, the singulate mean ages were Bangladesh 16.6, India 18.3, Nepal 18.9, Pakistan 20.7, and Sri Lanka 24.4.

Marriage postponement probably accounts for 20 to 30% of the decline in total fertility in Pakistan and Sri Lanka but rather less in the other three countries. However, delayed marriage is not consciously motivated by a desire for smaller family sizes. The fundamental cause is a change in the marriage market. In a rapidly growing population, each age group is larger than the preceding older one and, in societies where women traditionally marry older men, this creates a surplus of marriageable women. It thus becomes more difficult and expensive for parents to find suitable bridegrooms for their daughters. Increasing secondary schooling for girls accentuates this upward pressure on marriage ages.

The major direct cause of the South Asia fertility transition is increased use of contraception and abortion. Abortion is believed to be commonly practiced in all five countries but its precise contribution to fertility decline is impossible to gauge with any confidence because in Sri Lanka, Pakistan and until recently in Nepal it is illegal, while in Bangladesh and India most terminations are unrecorded.

In contrast, trends in contraceptive practice have been monitored reliably. A very pronounced link exists between reported contraception and fertility. Around 2010, the percentage of married women in the reproductive age span who reported current use of a modern contraceptive method was in the narrow range of 44–53% with the exception of Pakistan at 19%. Highly effective methods are dominant. In India and Nepal, there is a heavy reliance on sterilization, mainly of women. In 2005, sterilization accounted for 68% of contraceptive protection in India, and the median age of sterilization of women was 25 years. In contrast, the great surge in contraceptive practice in Bangladesh since the late 1980s was dominated by increased use of hormonal methods.

Socioeconomic, Cultural, and Policy Influences

During the past 50 years, a fertility transition has spread across South Asia despite the variation in material, social, and cultural circumstances. Presumably some common underlying factor, or interrelated cluster of factors, accounts for the onset of transition over the same relatively short timescale. Such factors might include preceding mortality declines that have increased numbers of surviving children, or monetization of economies and growing aspirations for educating children and for material goods, all of which increase the cost of childbearing. The variations in the timing and speed of transition demand more context-specific explanations.

Population scientists disagree about the causes of change (*see* Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations), and rather few widely accepted generalizations can be made. One of the few is that the fertility transition in the region is not primarily a response to rising living standards. Bangladesh remains one of the poorest countries in the world and the southern states of India have income levels that are below the national average. Indeed, some commentators have proposed that reproductive change has been driven by poverty rather than prosperity, though the microlevel evidence does not support this view: poorer families have more children than richer families.

Progressive social policies, resulting in universal schooling, high life expectancy, relative equality of incomes, and state welfare provision, offer a compelling explanation for the early transitions in Sri Lanka and Kerala. But this thesis applies less obviously to Tamil Nadu, where literacy and life expectancy are not exceptionally high, and is implausible in the case of Bangladesh.

Gender inequality has been a major explanatory theme (*see* Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective). In an influential paper, Dyson and Moore (1983) interpreted the north–south gradient in Indian fertility in terms of the kinship structure. North India, Pakistan, and the Indo-Aryan majority in Nepal are part of a wider West Asian kinship system dominated by patrilineage. Its key characteristics include early exogamous marriage for women, a joint family structure, and an emphasis on purdah, which result in very limited autonomy for young wives and considerable pressure to produce sons. In northern Indian cities, the imperative to have a son is increasingly expressed in sex-selective abortion, an illegal practice but one that is difficult to suppress.

In contrast, southern India and Sri Lanka are part of a South and East Asian kinship system, where endogenous marriage and nuclear family are more prevalent, women's lives are less rigidly controlled, and preference for sons over daughters is less pronounced. According to Dyson and Moore this system, with its greater autonomy for women, is more amenable to demographic innovation, including the adoption of modern birth control. Bhat (1996) concluded that these elements of social structure and kinship account for some 30 to 50% of the north–south fertility difference in India. Possibly even more important is exposure to mass media, a result interpreted by Bhat as testimony to the powerful influence of Western ideas on reproductive behavior.

Considerable stress has also been laid on the role of education, particularly for women, as a facilitator of fertility transition. At the individual level, the reproductive behavior of unschooled women and those who have received primary schooling differs little, whereas those with secondary schooling invariably have fewer children. Yet the contextual influence of schooling may be more important than the impact of individual attainment. Women in Kerala and Tamil Nadu have lower fertility than those in Bihar and Uttar Pradesh, regardless of their length of schooling (see **Figure 2**). While it is widely accepted that high adult literacy has been conducive to reproductive change, it is also clearly not a precondition. In 1991, when Bangladesh was in the midst of rapid fertility decline, only 23% of adult women and 36% of men were literate.

The contribution of state family planning policies to the fertility transition is also controversial. All five countries have family planning policies with explicit demographic goals. All have offered financial incentives (or compensation payments) to individuals who agree to sterilization, and all have been criticized for coercive elements. In Sri Lanka, India, and Nepal, vigorous government promotion of family planning has probably accelerated fertility decline (see, for example,

Caldwell et al., 1982). In Bangladesh it may have been even more important. A well-known experiment in the Matlab district demonstrated that the appropriate provision of services could reduce fertility (Phillips et al., 1988) and it has been claimed that, while a preceding mortality decline was the underlying cause of the fertility transition, the family planning program was largely responsible for the timing and speed of change (Cleland et al., 1994). In Pakistan, the program has been largely ineffective because of weak political support.

Prospects for the Future

Bangladesh's fertility is projected to reach replacement level by 2015, Sri Lanka by 2020, India and Nepal by 2030, and Pakistan by 2040 (see **Figure 1**). The prospects for the continuation of past downward trends are good. Recent surveys indicate that at least 20% of births are unintended in Pakistan, Nepal, Bangladesh, and India. The United Nations Population Division expects fertility in the region to stabilize at 1.8 births. The crucial uncertainty about the future is whether or not fertility will fall further to the very low levels experienced in East Asia (*see* Fertility Transition: China; Fertility Transition: East Asia). In India, an appreciable proportion of men and women state a preference for one child only and the figure would be considerably higher in the absence of a desire for a son. However, sterilization, the dominant method, is a risky option for such couples and there is an obvious need for promotion of long-acting but reversible methods to facilitate this desire.

See also: Family Size Preferences; Fertility Transition: Cultural Explanations; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective.

Bibliography

Bhat, P.N.M., 1996. Contours of fertility decline in India: a district level study based on the 1991 Census. In: Srinivasan, K. (Ed.), Population Policy and Reproductive Health. Hindustan Publishing Corporation (India), New Delhi, India, pp. 97–177.

Bose, A., Premi, A.K. (Eds.), 1992. Population Transition in South Asia. B.R. Publishing Corporation, New Delhi, India.

Cain, M.T., 1983. Fertility as an adjustment to risk. Population and Development Review 9, 688–702.

Caldwell, J.C., Reddy, P.H., Caldwell, P., 1982. The causes of demographic change in rural south India. Population and Development Review 8, 689–727.

Cleland, J., Phillips, J.F., Amin, S., Kamal, G.M., 1994. The Determinants of Reproductive Change in Bangladesh: Success in a Challenging Environment. World Bank Regional and Sectional Studies, Washington, DC.

Dyson, T., Moore, M., 1983. On kinship structure, female autonomy, and demographic behavior in India. Population and Development Review 9, 35–60.

Fernando, D.F.S., 1975. Changing nuptiality patterns in Sri Lanka 1901–1971. Population Studies 29, 179–190.

Hajnal, J., 1953. Age at marriage and proportion marrying. Population Studies 7, 110–136.

Jeffery, R., Basu, A.M. (Eds.), 1996. Girls Schooling, Women's Autonomy and Fertility Change in South Asia. Sage, New Delhi, India.

Leete, R., Alam, I. (Eds.), 1993. The Revolution in Asian Fertility. Clarendon Press, Oxford, UK.

Phillips, J.F., Simmons, R., Koenig, M.A., Chakaborty, J., 1988. The determinants of reproductive change in a traditional society: evidence from Matlab, Bangladesh. Studies in Family Planning 19, 313–334.

Vlassoff, M., Vlassoff, C., 1980. Old age security and the utility of children in rural India. Population Studies 34, 487–499.

Zachariah, K.C., Rajan, S.I. (Eds.), 1997. Kerala's Demographic Transition: Determinants and Consequences. Sage, New Delhi, India.

Fertility Transition: Southeast Asia

Patrick Heuveline, University of California, Los Angeles, CA, USA
Charles Hirschman, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C. Hirschman, volume 8, pp. 5597–5602, © 2001, Elsevier Ltd.

Abstract

On average, fertility in Southeast Asia is barely above replacement level and declining. However, the region includes countries with some of the highest and some of the lowest fertility levels in the world. Different political contexts have affected the timing of the fertility transition across the region. Also notable were differences in the pace of the fertility transition, between the predominantly Buddhist mainland and the predominantly Muslim or Catholic nations in insular Southeast Asia. While some nations are completing their fertility transition, others are facing the prospects of population aging and may soon undergo the second demographic transition.

Located east of the Indian subcontinent, south of China, and west of Melanesia (Figure 1), the 11 nations of Southeast Asia (SEA thereafter) have a combined population size estimated to be 619 million in 2013 – a more than threefold increase from the region's population of 168 million in 1950 (United Nations, 2013a). The average annual growth rate of 2.07% over this period outpaced the world population's growth rate and reflected a substantial difference between birth and death rates. Population growth has slowed dramatically in recent years and only remains positive due to population momentum – a relatively youthful age structure. The average period total fertility rate (TFR) for the region is barely above replacement level (2.13 in 2010–15, United Nations, 2013a).

However, the overall level and decline of SEA's fertility masks significant variations from country to country. The region includes countries with some of the highest and some of the lowest fertility levels in the world (Timor-Leste, third highest with 6.53, and Singapore, fourth lowest with 1.25 in 2005–10, United Nations, 2013b; Table S.12). SEA also includes nations with some of the world's most rapid fertility decline, for example, Vietnam from 1975–80 to 2005–10 (United Nations, 2013b; Table S.13) whereas others have experienced fertility lulls and even temporary increases during that same period (Cambodia, the Lao PDR, and Timor-Leste).

Demographic diversity is not entirely surprising in a region, which shares some common geographical, historical, religious,

Figure 1 Countries of Southeast Asia.

and cultural features, but also experienced quite varied patterns of colonial rule, political and economic systems, and development (Reid, 1988a, 1993; Wolters, 1999). Created originally in 1967 with only five member states, the Association of Southeast Asian Nations (ASEAN) today includes all of SEA except but the recently independent nation of Timor-Leste. Participation in ASEAN may eventually contribute to greater political and economic integration in a region divided by decades of cold war politics.

The main line of geographical and cultural division is typically drawn between the nations of mainland and insular SEA. (Although most of its population lives on the mainland, Malaysia also includes part of the island of Borneo, shared with Indonesia and Brunei Darussalam. Together with these two nations, Malaysia is generally classified in insular SEA with Singapore (the two only became separate nations in 1965), as well as Timor-Leste and the Philippines.) However, substantial diversity is also found within nations, such as between the different islands of large archipelagoes (Indonesia and the Philippines), or on the mainland, between densely populated lowlands and the relatively isolated uplands home to small ethnic minorities (Scott, 2009). Numerically larger groups of Chinese and Indian-origin have also been present throughout the region since the late nineteenth and early twentieth century (Cushman and Wang, 1988; Reid, 1996, 2001), especially in the major metropolitan areas of the region. Driving the new regional economies, these metropolitan areas are increasingly distinct – ethnically, economically, culturally, and demographically – from their (mostly rural) national populations. In 2000, the TFR in Bangkok was estimated at 1.16 children per woman, closer to Singapore's than to the national figure for Thailand at the time (1.81, Jones, 2009).

The extent of between- and within-nation diversity cannot be fully described here, but key features are highlighted because of their potential influence on the fertility transition. The principal religions present in SEA are (1) Buddhism, predominant on the mainland (Myanmar, Thailand, the Lao PDR, Cambodia, and Vietnam), (2) Catholicism, influential in the Philippines and Timor-Leste, and (3) Islam, the majority religion in Malaysia, Indonesia, and Brunei Darussalam, with significant minorities on the mainland (Cambodia and Thailand), in the Philippines, as well as in Singapore. Hinduism is the major religion in Bali (Indonesia) and among the ethnic Indian populations of Malaysia and Singapore, whereas Confucianism remains influence among ethnic Chinese communities found in metropolitan areas throughout the region. These different religious influences, and colonial histories have pulled Southeast peoples in different directions during the modern era, but they have not erased the distinctive cultural feature of SEA: women's relatively high status and economic autonomy (Reid, 1988b; Van Esterik, 1982). Age differences between spouses tend to be comparatively small (Casterline et al., 1986), currently between 2 and 4 years (United Nations, 2012). Although polygamy is allowed under Islam, it was never very common pattern in SEA (Reid, 1988b; Jones, 1994). Matrilocality is not rare in any Southeast Asian country and is often as common as patrilocality among the ethnic Malay populations in Malaysia and Indonesia, the Philippines, and in most mainland nations (Reid, 1988a; Jones, 1995). The major exceptions are Vietnam and ethnic Chinese populations throughout the region which express strong gender preference (Wongboonsin and Ruffolo, 1995; Guilmoto, 2012). Another common cultural pattern, especially among Buddhist societies is the moral obligation to repay parents for one's care in early life. The cultural expectation that children will support their elderly parents is widely shared in throughout the region (Croll, 2006).

A Demographic Overview

Five of SEA's 11 nations are now among the world's 25 largest countries, compared to only 2 in 1950. Indonesia, the fourth largest country in the world with 250 million people, is the region's "demographic giant," accounting for 40% of the region's population (**Table 1**). Close to 100 million people also live in the Philippines, and more than half of the population of SEA lives on these two archipelagoes (56.3%). The other three nations are in mainland SEA, Vietnam at the eastern

Table 1 Total population size, growth, density, and distribution by country

		Annual growth rate				
	Total population size, 2013	*1950–2013*	*2010–15*	*% Urban, 2011*	*Density km^{-2}, 2010*	*Median age, 2010*
Brunei Darussalam	418	3.43%	1.35%	75.20	69.5	29.5
Cambodia	15 135	1.95%	1.75%	19.80	79.3	23.5
Indonesia	249 866	1.96%	1.21%	44.00	126.4	26.9
The Lao PDR	6770	2.21%	1.86%	32.00	27.0	20.3
Malaysia	29 717	2.51%	1.61%	71.30	85.7	26.1
Myanmar	53 259	1.76%	0.84%	33.00	76.8	27.8
The Philippines	98 394	2.65%	1.71%	48.70	311.5	22.3
Singapore	5412	2.65%	2.02%	100.00	7436.3	37.3
Thailand	67 011	1.87%	0.30%	33.60	129.4	35.4
Timor-Leste	1133	1.53%	1.66%	27.70	72.6	16.1
Vietnam	91 680	2.07%	0.95%	29.80	268.5	28.5

Source: United Nations, 2011a. World Urbanization Prospects, the 2011 Revision. United Nations, New York; United Nations, 2013a. World Population Prospects, the 2012 Revision. Comprehensive Tables, vol. 1. United Nations, New York; United Nations, 2013b. World Population Prospects, the 2012 Revision. Demographic Profiles, vol. 2. United Nations, New York.

end and Thailand and Myanmar at the western end of the region. Together, these five nations account for more than 90% of the region's population.

SEA is densely populated. Compared to the world average of 51 people per squared kilometer, the regional average is 133 people per squared kilometer (**Table 1**), and five nations have density twice the world average or more. Besides the city-state of Singapore (7436), the Philippines (311), Vietnam, (268), Thailand (129), and Indonesia (126) are still predominantly rural countries. Dense rural settlements suggest demographic pressure on arable land that likely contributed to the growth of the large urban agglomerations that have developed in each of these nations. As of 2011, the population size of the urban agglomeration of Ho Chi Minh City (Vietnam) has reached 6.4 million inhabitants, Krung Thep (*Bangkok*, Thailand) 8.4 million, Jakarta (Indonesia) 9.8 million, and Manila (Philippines) 11.9 million (United Nations, 2011a). In 1975, Manila was already the region's largest urban agglomeration but with a population size of just 5.0 million.

As in many nations in the less developed regions (LDR) of the world, mortality has declined tremendously in the second half of the twentieth century (Heuveline, 1999). Three SEA nations (Cambodia, Myanmar, and Timor-Leste) have life expectancies at birth in the mid-60s, slightly below the contemporary LDR average of 67.0 (in 2010–15) (United Nations, 2013a) (**Table 1**). The estimates of life expectancy for the wealthiest nations, Singapore (81.3 years) and Brunei Darussalam (78.2 years), exceed the average longevity in the more developed regions (MDR, 76.9 years). Only slightly lower are estimates of life expectancy for Malaysia (74.6 years), Thailand (74.4 years), and Vietnam (75.5 years). These countries have experienced very rapid economic progress in recent decades. The diversity of mortality transitions in SEA nations exemplifies the varied "routes to low mortality" (Caldwell, 1986; Kuhn, 2010).

As for the transition to low fertility, Singapore was the region's forerunner, with a decline that started in the mid-twentieth century and proceeded with unprecedented speed – fertility had dropped below the replacement level by the late 1970s (Saw, 2005). In the rest of SEA, fertility rates seem stable or even increasing slightly in the 1950s. Such limited increases are fairly common across LDRs and usually indicate a decline in sexually transmitted diseases and subsequent infertility, rather than a fertility increase among fecund, married women. Fertility might have been increasing during the nineteenth century as well, as the colonial-period record, imperfect though it is, indicates an acceleration of population growth that may have preceded any significant improvement in health. While this fertility increase and its causes are still a matter of scholarly debate (Henley, 2011; Reid, 1992), the demographic picture becomes much clearer in the 1950s and reveals fertility rates in the range 5.5–7.4 children per woman that are fairly typical of LDR nations at the time (**Table 1**).

In mainland SEA, fertility levels did not decline significantly until the late 1960s (Thailand), the late 1970s (Myanmar and Vietnam), or even the late 1980s (Cambodia and the Lao PDR). When fertility rates begin to fall, however, after centuries of stability or slow increase, they declined rapidly. Replacement level was reached in less than 30 years after the onset of the decline in Thailand, Myanmar, and Vietnam (**Table 2**). While Thailand has come to exemplify a fast fertility decline in a relatively poor nation (e.g., Hirschman et al., 1994; Knodel et al., 1996), the decline in Myanmar and even more so in Vietnam from over seven children per woman in the early 1970s to replacement level before the century's end are equally impressive (Goodkind, 1995; Haughton, 1997). In Cambodia and the Lao PDR, civil unrest and warfare might have temporarily depressed fertility levels, but the actual transition did not start until after peace had returned, sometimes followed by

Table 2 Mortality, nuptiality, contraceptive use and fertility indicators, by country

	Mortality indicators		*Marriage indicators*		*Contraception*	*Fertility transition*			*Total fertility rate*	
	Life expectancy at birth	*Under-5 mortality rate*	*Ever married*	*SMAM*	*Any modern method*[a]	*Peak value*	*Onset year*	*Years to replacement*[b]	*2010–15*	*2030–35*
Brunei Darussalam	78.2	6	88.6	25.0	n/a	7.00	1960–65	45.4	1.98	1.68
Cambodia	63.7	69	93.9	22.0	34.9	7.00	1985–90	*26.7*	2.42	1.75
Indonesia	70.0	31	97.9	22.3	57.9	5.62	1970–75	38.5	2.06	1.74
Lao PDR	67.9	46	96.5	21.7	42.1	6.36	1990–95	*22.3*	2.54	1.76
Malaysia	74.6	9	94.6	25.7	29.8	6.23	1965–70	*50.8*	2.57	2.15
Myanmar	66.0	57	85.2	26.1	45.7	6.10	1975–80	29.4	1.94	1.63
The Philippines	69.2	27	92.8	24.4	36.3	7.42	1960–65	*60.9*	3.05	2.43
Singapore	81.3	2	87.2	27.9	55.0	6.61	1955–60	18.7	1.37	1.69
Thailand	74.4	13	92.0	24.1	79.8	6.14	1965–70	23.2	1.53	1.53
Timor-Leste	63.1	76	96.8	23.1	20.8	7.00	2005–10	*68.1*	5.92	3.73
Vietnam	75.5	23	94.4	22.7	59.8	7.33	1975–80	21.6	1.75	1.58

[a]Most recent year available, typically in the last 10 years except for Malaysia (1995).
[b]For countries that have not yet reached replacement level fertility, the figures are estimated based on extrapolating the pace of the fertility decline to date (figures in italics).
Source: United Nations, 2011b. World Contraceptive Use 2010. United Nations, New York; United Nations, 2012. World Marriage Data. http://www.un.org/en/development/desa/population/ (accessed 15.08.13.); United Nations, 2013a. World Population Prospects, the 2012 Revision. Comprehensive Tables, vol. 1. United Nations, New York; United Nations, 2013b. World Population Prospects, the 2012 Revision. Demographic Profiles, vol. 2. United Nations, New York; Ministry of Health and Lao Statistics Bureau, 2012. Lao Social Indicator Survey (LSIS), 2011–12. ICF International, Calverton, Maryland; National Institute of Statistics and Ministry of Health Cambodia, 2011. Cambodia Demographic and Health Survey (CDHS), 2010. ICF Macro, Calverton, Maryland; and authors' calculations.

postcrisis "baby-boom" (Heuveline and Poch, 2007). However, at the current pace, these two nations are expected to reach replacement level fertility in less than 30 years after the onset of fertility decline. While different political contexts are likely to have affected the onset of fertility declines, all mainland SEA nations will likely share a fast transition to replacement fertility. Moreover, fertility declines have not stopped at replacement levels, and there is still no indication that the current estimates of the number of children per woman have stopped declining in Thailand (1.53), Myanmar (1.94), or Vietnam (1.75).

With the exception of Timor-Leste, which only recently won its independence, fertility began to decline in most of insular SEA from the early 1960s to the early 1970s. On the mainland at the time, only Thailand showed any sign of a fertility transition. However, no insular SEA nations had reached replacement fertility by 2010 (except Singapore of course). In spite of a population density almost three times the regional average, average fertility in the Philippines is still above three children per woman. At the current pace, the Philippine fertility transition may take over 60 years, more than twice the actual or projected duration in any of the mainland SEA nations. In spite of being wealthier, markedly more urbanized (71.3% vs 41.4%), and enjoying child mortality rates three times lower than average for the region, Malaysia's total fertility is still above 2.5 children per woman (G.W. Jones, 1990). As had been observed in Malaysia in the 1980s (Hirschman, 1986), the most recent data for Indonesia indicates a stall or even an increase in fertility from 2.27 in 2000 to 2.41 in 2010 (Badan Pusat Statistik, 2013). This casts doubts on the current projection that the fertility transition might be completed in 2010–15, which would have been the fastest in insular SEA besides Singapore, but would still have taken about 40 years from onset to replacement fertility.

Explanations of Fertility Transition in Southeast Asian Nations

Fertility transitions are typically driven by declines in marital fertility, and this has clearly been the dominant pattern in SEA (Retherford and Cho, 1973; Hirschman and Guest, 1990a). However, there have also been significant changes in marriage patterns in several Southeast Asian countries. The erosion of arranged marriages has led to declines in the proportion of married females at young ages and increases in the mean age at marriage among Muslims in Indonesia and Malaysia in particular (Leete, 1994; Jones, 1995; Hull, 2012; Jones and Gubhaju, 2012). The postponement of marriage was, however, a modest factor in shaping total fertility changes in SEA compared to changes in the behavior of couples after marriage. The gradual adoption of modern contraceptives appears to have played the crucial role in this decline of marital fertility (Hull et al., 1977; Freedman et al., 1981; Ross and Poedjastoeti, 1983; Knodel et al., 1987; Entwisle et al., 1997) in the 1970s through the 1990s. The recent changes in marital patterns, including significant proportions of never married women in their 30s may be a more consequential determinant of fertility in the coming years.

A growing consensus assigns to mortality change a central role in explaining the onset of the fertility transition (Hirschman, 1994; Mason, 1997; Cleland, 2001). The spectacular declines in child mortality, in particular, added strains on the family (Davis, 1963), but more generally, contributed to an acceleration of population growth in already densely populated areas. Pre-transitional fertility levels were lower in higher density areas in Indonesia (Hugo et al., 1987), higher in lower density frontier areas in Thailand (Vanlandingham and Hirschman, 2001) and increased after mortality-driven reduction in population density in Cambodia (Heuveline and Poch, 2007). Consistent with this interpretation, pre-1975 fertility declines are visible in the SEA nations with the lowest child mortality levels (Brunei Darussalam, Malaysia, Singapore) or with the highest population densities (Indonesia, Thailand, Philippines). The only exception to this generalization is Vietnam, but the transition likely delayed by the war there eventually outpaced transitions in the rest of the region.

In the classic formulation of the demographic transition theory (Notestein, 1953), mortality decline was assigned less of a central role, being held rather as one of the several dimensions of modernization, or even possibly a consequence thereof. The multidimensional development process, with rising levels of urbanization and education accompanying lower mortality rates may explain the early onset of the transition in Singapore. In the rest of the region, however, as in much of the LDR, mortality declines were achieved rapidly after World War II, ahead of any other sign of substantial modernization (in particular in Myanmar or Vietnam). In some countries, subsequent modernization likely played an accelerating role after the onset of fertility decline (Hirschman and Guest, 1990b; Gertler and Molyneaux, 1994; Hirschman et al., 1998). With the development of public education systems, parents cite the associated increase in the costs of childrearing as one of the factors influencing their decisions to have fewer children (e.g., Knodel et al., 1984). Some of the increase, however, originates less in the educational system than in the job market and the perception of an intensifying competition that leads to higher parental investments and eventually so-called "quantify/quality" trade-offs (Becker, 1991). Another crucial factor in the increasing costs of childrearing has been the opportunity costs of women's wages. Traditionally engaged in the formal economy, women have provided almost the entire workforce of delocalized factories in fast-expanding sectors like the garment industry.

While mortality, with or without other modernization factors, contributed to reduce the demand for children, family planning programs facilitated access to more efficient means to achieve smaller families. In Thailand, Indonesia, or Vietnam, the onset of fertility decline seems to occur very soon after family planning programs were set up (Ross and Poedjastoeti, 1983; Robinson and Rachapaetayakom, 1993; Haughton, 1997). These three nations still have the highest modern contraceptive prevalence rates in the region today (**Table 2**), and the effectiveness of their national programs was clearly a factor in their fertility reaching replacement levels or lower – in Vietnam in particular, with the introduction of a one–two child policy (Goodkind, 1995). Signs of a fertility decline were visible before the establishment of a national family

planning program in some countries, for example, in the Philippines, though by only a few years, and even more clearly in Myanmar (Herrin, 2007; Myint, 1991). In Cambodia, the Lao PDR, and Myanmar, modern contraceptive prevalence has increased in the first decade of the twenty-first century and now ranges from 34.9% to 45.7%. These contraceptive prevalence rates, though high by historical patterns, are actually lower than the levels usually associated with their fertility: Laos has a prevalence rate of 55.2% and a TFR of 2.57 children per woman. Other than contraceptive use, potential factors contributing to this low fertility includes spousal separation due to temporary work-related migration and voluntary abstinence.

Overall, the fertility transition is relatively advanced in SEA and the region's cultural context has typically been considered favorable to the fertility transition (e.g., relatively high-female autonomy). Contrasting nations within SEA, however, one can hardly overlook the religious dimension. In the predominantly Catholic Philippines, the transition has been very gradual, arguably more typical of Latin American nations. As noted earlier, the transition in predominantly Muslim Indonesia and Malaysia have proceeded unevenly, with some temporary lulls or possibly reversals. The transitions in predominantly Buddhist mainland SEA started later, proceeded faster, and to date, without interruption until well below replacement level. The uneven transition in multi-ethnic Malaysia and Indonesia might reflect fertility transitions proceeding at different paces among different communities – differences that might be related to government policies (Govindasamy and DaVanzo, 1992). Although their small population sizes often makes them invisible even in disaggregated, subnational data, the limited evidence on the various highland populations in the region suggests that their fertility remains high and stands out from the general transition in the region (Gray et al., 2005).

Completing the First, Starting a Second Demographic Transition?

Fertility is currently below replacement in 6 of the 11 SEA nations. Recent trends suggest that fertility in these nations is very likely to remain below replacement in the next couple of decades (**Table 2**). Concerns over high density and fast growing populations may soon be replaced by concerns over rapidly aging populations that may even experience absolute decline in the coming years. The median age in Thailand is projected to reach 51.1 years by 2050, which would make it one of the 10 oldest in the world (United Nations, 2013a).

Singapore has tried to increase fertility (Saw, 2005: Chapters 11–13), but as elsewhere in the world, their pro-marriage and pro-natalist policies only had a limited impact (Yap, 1995) and failed to durably and substantially raise fertility (Jones, 2012). In contrast to Europe, changes in the timing of fertility have had little impact on period fertility rates in SEA (Chuan, 2010). In Europe, the postponement of fertility has been a major reason for the ultra-low levels of fertility in recent years. Since fertility postponement cannot continue indefinitely, this downward trend appears to have run its course and period fertility measures are inching back up (Goldstein et al., 2009). By contrast, in SEA, mean ages at marriage remain below 25 in most nations (SMAM (Singulate Mean Age at Marriage), **Table 2**), and fertility rates have fallen even more after age 35 than below age 35 (e.g., Thailand, Vietnam: United Nations, 2012). Like Singapore, nations faced with durable below-replacement fertility may also attempt to counter population decline and aging through immigration, but this might prove politically challenging (Jones, 2012; National Population and Talent Division, 2012).

Fertility rates are expected to continue to decline in the five SEA nations still above replacement, albeit at the different paces noted above. By 2030–35, only Malaysia and the Philippines (due to a slow pace) and Timor-Leste (due to a late onset) are projected with fertility levels above replacement. These fertility projections are mere extrapolations of past trends, and reversal cannot be entirely ruled out. As noted in Indonesia, fertility seems to have stabilized and possibly increased slightly in the past 10 years, while the proportion of married women among 15–24 year olds increased slightly (United Nations, 2012) after a precipitous decline in the previous decades (Jones and Gubhaju, 2012). This inflection of these demographic trends might result from a reaction to secularization and globalization trends leading to an increase in religiosity (Hull, 2012).

The record from around the world seems to indicate that the completion of ongoing fertility transitions is all but certain, and that lulls and slight reversals are only temporary. A more speculative question is whether the region may experience the onset of a 'second' demographic transition, alongside the 'first' one still ongoing in some other parts of the region. As argued above, the fertility transitions of SEA have been driven by declining marital fertility rates, without necessarily involving major transitions in marriage, divorce, unmarried cohabitation, and childrearing or having children out of wedlock toward rates that have been associated with low fertility in Western nations.

To date, if by international standards, the proportions of women marrying are still relatively high across SEA nations, the trend away from universal and young marriages is clear. Surprisingly, the proportion ever marrying is lowest in Myanmar (85.2%, **Table 2**). Rather than a new trend, this might be linked to contemporary political uncertainties and to the traditionally accepted form of celibacy in order to look after aged parents (Jones and Gubhaju, 2012: 86). With smaller family sizes, older-parent care falls on an increasing proportion of the children's generation, and with economic restructuring, may now require work-related emigration and remittances from urban settings, increasingly conflicting with the traditional marriage process (Barbieri, 2009; Alipio, 2013). However, the clearer signs of decreasing proportions ever marrying are in Singapore (87.2%, **Table 2**), among ethnic Chinese communities in the rest of the region, and in some of the other large metropolitan areas (Jones, 2005). Given the current weight of the rural populations in most SEA nations still these changes are not too visible on national averages yet, but might soon become so with the likely continuation of rural-to-urban migration and possible diffusion to rural areas of new attitudes regarding marriage (Tey, 2007; Situmorang, 2007; Jayakody and Vu, 2009). Observed all over the world, the trend is not surprising and is

frequently associated with the expansion of educational and professional opportunities for women. While a major reason for marital postponement in SEA as well, rising educational levels are not the only factor as Jones and Gubhaju (2009) finds increasing proportions never marrying across all educational groups. A modest counterbalancing factor – visible, but still relatively small numerically – is the predicted increase in international marriages to males from countries confronted to the "marriage squeeze" generated by increasing male-to-female birth ratios (Jones and Shen, 2008; Yang and Lu, 2010).

Unmarried cohabitation seems to remain extremely rare in SEA, at least based on the standardized data that demographic surveys collect across many countries. In the region's national censuses, unmarried cohabitation only appears as a distinct marital status in the Philippines (Jones and Gubhaju, 2012). When unmarried couples openly live together, however, they are typically considered as married *de facto*. The census definition of marriage in Indonesia, which includes unmarried cohabitation (Hull, 2012), seems to reflect the general perception also prevailing in predominantly Buddhist nations. Elaborate and often expensive wedding ceremonies are socially desirable in the community, but Buddhism does not prescribe any particular ritual (Cherlin and Chamratrithirong, 1988). Unmarried cohabitation is likely more prevalent than data that suggest among the lower classes and in urban settings (Xenos and Kabamalan, 2007; Esara, 2012). Nonetheless, when married and unmarried partnerships are considered jointly, the prevalence of actual celibacy appears higher in SEA than in Europe (Jones, 2007).

In Western nations, increasing divorce rates are another feature of the second demographic transition (Lesthaeghe, 1983), consistent with the economic theory of marital instability (Becker et al., 1977). In SEA, however, some traditional societies were already accommodating high levels of marital disruptions (Jones, 1997; Hirschman and Teerawichitchainan, 2003). In Thailand, divorce rates were averaging about 15% after 20 years of marriage (Smith, 1981; Knodel et al., 1984), but were highest among Muslims, as also observed in Malaysia and Indonesia. Divorce there is relatively simple to obtain, at least for the husband, and was experienced by as many as 40% of men and women when arranged, early marriage was the norm (Jones, 1994), until the declining incidence of this marriage pattern brought a concomitant decline in divorce rates. In Cambodia, where marriage rates had remained relatively rare, however, the likelihood of divorce is now higher among couples who chose each other as marriage partners, and divorce rates have actually increased with the decline in arranged marriages (Heuveline and Poch, 2006). Slow increases in this 'modern' type of divorce are visible elsewhere in the region, including in populations where it was disapproved of before, such as those influenced by Confucian traditions (Jones and Gubhaju, 2012). Secular divorce still does not exist in the Philippines, where marriages can only be annulled by the Catholic Church.

Finally, out-of-wedlock birth ratios remain low, trivial even in some of the region's nations. As elsewhere in the LDR, there is evidence that premarital sexual intimacy is on the rise, in particular in the region's metropolitan areas (Brown et al., 2013; Ghuman et al., 2006; Gipson et al., 2012). A gradual change in extra-marital sexual behavior notwithstanding, there is little evidence to date to suggest that SEA will soon exhibit the high nonmarital fertility ratios of the United States or some North European nations. In these nations, out-of-wedlock fertility has maintained overall fertility fairly close to replacement level. In SEA, a more likely demographic path would seem to include the expectation of childbearing mostly within marriage, marital decisions still embedded in child–parent relationships, but increasingly the prerogative of individuals with changing aspirations, some of which may conflict with marriage and childbearing. In such a scenario, fewer and later marriages are typically paired with fertility substantially and durably below replacement (Jones, 2007).

Acknowledgments

While preparing this article, the first author benefited from the facilities and resources at the California Center for Population Research, UCLA, which is supported by infrastructure grant R24HD041022 from the Eunice Kennedy Shriver National Institute of Child Health & Human Development. He also wishes to thank Gavin Jones, Terry Hull and Jenna Nobles for bringing to his attention some of the work reviewed for this article. The authors remain solely responsible for the content of this article.

See also: Demographic Measurement: General Issues and Measures of Fertility; Family Size Preferences; Fertility Transition: Cultural Explanations; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Southeast Asia: Sociocultural Aspects.

Bibliography

Alipio, C., 2013. Young men in the Philippines mapping the costs and debts of work, marriage, and family life. The ANNALS of the American Academy of Political and Social Science 646 (1), 214–232.

Badan Pusat Statistik, 2013. Total Fertility Rate (TFR) by Province 1971, 1980, 1985, 1990, 1991, 1994, 1997, 1998, 1999, 2000, 2002, 2007 and 2010. http://www.bps.go.id/eng/tab_sub/view.php?tabel=1&daftar=1&id_subyek=12&notab=7 (accessed 08.08.13.).

Barbieri, M., 2009. Doi Moi and older Adults: intergenerational support under the constraints of reform. In: Barbieri, M., Bélanger, D. (Eds.), Reconfiguring Families in Contemporary Vietnam. Stanford University Press, Stanford, CA, pp. 133–165.

Becker, G., 1991. A Treatise on the Family. Harvard University Press, Cambridge, MA.

Becker, G., Landes, E.M., Michael, R.T., 1977. An economic analysis of marital instability. The Journal of Political Economy 85 (6), 1141–1187.

Brown, A.D., Jejeebhoy, S.J., Shah, I., Yount, K.M., 2013. Sexual Relations among Young People in Developing Countries: Evidence from WHO Case Studies. WHO. http://www.who.int/reproductivehealth/publications/adolescence/RHR_01.8/en/index.html (accessed 05.09.13.).

Caldwell, John, 1986. Routes to low mortality in poor countries. Population and Development Review 12 (2), 171–220.

Casterline, J.B., Williams, L., McDonald, P., 1986. The age difference between spouses: variations among developing countries. Population Studies 40 (3), 353–374.

Cherlin, A., Chamratrithirong, A., 1988. Variations in marriage patterns in Central Thailand. Demography 25 (3), 337–353.

Chuan, K.E., 2010. Will Singapore's fertility rise in the near future? Asian Population Studies 6 (1), 69–82.

Cleland, J., 2001. The effects of improved survival on fertility: a reassessment. Population and Development Review 27 (Suppl.: Global Fertility Transition), 60–92.

Croll, E.J., 2006. The intergenerational contract in the changing Asian family. Oxford Development Studies 34 (4), 473–491.

Cushman, J., Wang, G. (Eds.), 1988. Changing Identities of the Southeast Asian Chinese since World War II. Hong Kong University Press, Hong Kong.
Davis, K., 1963. The theory of change and response in modern demographic history. Population Index 29 (4), 345–366.
Entwisle, B., Rindfuss, R.R., Walsh, S.J., Evans, T.P., Curran, S.R., 1997. Geographic information systems, spatial network analysis, and contraceptive choice. Demography 34 (2), 171–187.
Esara, P., 2012. Moral scrutiny, marriage inequality: cohabitation in Bangkok, Thailand. The Asia Pacific Journal of Anthropology 13 (3), 211–227.
Freedman, R., Khoo, S.-E., Supraptilah, B., 1981. Use of modern contraceptives in Indonesia: a challenge to the conventional wisdom. International Family Planning Perspectives 7 (1), 3–15.
Gertler, P.J., Molyneaux, J.W., 1994. How economic development and family planning programs combined to reduce Indonesian fertility. Demography 31 (1), 33–63.
Ghuman, S., Vu, M.L., Vu, T.H., Knodel, J., 2006. Continuity and change in premarital sex in Vietnam. International Family Planning Perspectives 32 (4), 166–174.
Gipson, J.D., Gultiano, S.A., Avila, J.L., Hindin, M.J., 2012. Old ideals and new realities: the changing context of young people's partnerships in Cebu, Philippines. Culture, Health & Sexuality 14 (6), 613–627.
Goldstein, J.R., Sobotka, T., Jasilioniene, A., 2009. The end of 'Lowest-Low' fertility? Population and Development Review 35 (4), 663–699.
Goodkind, D.M., 1995. Vietnam's one-or-two-child policy in action. Population and Development Review 21 (1), 85–111.
Govindasamy, P., DaVanzo, J., 1992. Ethnicity and fertility differentials in Peninsular Malaysia: do policies matter? Population and Development Review 18 (2), 243–267.
Gray, R.,C., Podhisita, P. Vapattanawong, Varangrat, A., 2005. Fertility Decline among the Karen and the Hmong, Hill Tribe Minorities in Northern Thailand. Paper presented at the XXV International Population Conference, Tours, France, 18–23 July 2005.
Guilmoto, C.Z., 2012. Son preference, sex selection, and kinship in Vietnam. Population and Development Review 38 (1), 31–54.
Haughton, J., 1997. Falling fertility in Vietnam. Population Studies 51 (2), 203–211.
Henley, D., 2011. Forced labour and rising fertility in colonial Indonesia. Asian Population Studies 7 (1), 3–13.
Herrin, A.N., 2007. Development of the Philippines' family planning program: the early years, 1967–80. In: Robinson, W.C., Ross, J.A. (Eds.), The Global Family Planning Revolution: Three Decades of Population Policies and Programs. The World Bank, Washington, DC, pp. 277–297.
Heuveline, P., 1999. The global and regional impact of mortality and fertility transitions, 1950–2000. Population and Development Review 25 (4), 681–702.
Heuveline, P., Poch, B., 2006. Do marriages forget their past? Marital stability in post-Khmer Rouge Cambodia. Demography 43 (1), 99–125.
Heuveline, P., Poch, B., 2007. The phoenix population: demographic crisis and rebound in Cambodia. Demography 44 (2), 405–426.
Hirschman, C., 1986. The recent rise in Malay fertility: a new trend or a temporary lull in a fertility transition? Demography 23 (2), 161–184.
Hirschman, C., 1994. Why fertility changes. Annual Review of Sociology 20, 203–233.
Hirschman, C., Guest, P., 1990a. The emerging demographic transitions of Southeast Asia. Population and Development Review 16 (1), 121–152.
Hirschman, C., Guest, P., 1990b. Multilevel models of fertility determination in four Southeast Asian countries: 1970 and 1980. Demography 27 (3), 369–396.
Hirschman, C., Tan, J.E., Chamratrithirong, A., Guest, P., 1994. The path to below replacement-level fertility in Thailand. International Family Planning Perspectives 20 (3), 82–107.
Hirschman, C., Tan, J.E., Chamratrithirong, A., Guest, P., 1998. Explaining the Rapid Fertility Decline in Thailand. Center for Studies in Demography and Ecology. University of Washington, Seattle, WA, unpublished paper.
Hirschman, C., Teerawichitchainan, B., 2003. Cultural and socioeconomic influences on divorce during modernization: Southeast Asia, 1940s to 1960s. Population and Development Review 29 (2), 215–253.
Hugo, G.J., Hull, T.H., Hull, V.J., et al., 1987. The Demographic Dimension of Indonesian Development. Oxford University Press, Singapore.
Hull, T.H., Hull, V.J., Singarimbun, M., 1977. Indonesia's family planning story: success and challenge. Population Bulletin 32 (6), 1–51.
Hull, T.H., 2012. Fertility in Southeast Asia. In: Williams, L., Guest, M.P. (Eds.), Demographic Change in Southeast Asia: Recent Histories and Future Directions. Cornell University, Ithaca, NY, pp. 43–64.
Jayakody, R., Vu, T.H., 2009. Family change in Vietnam's Red River Delta: from war, to reunification, to renovation. In: Barbieri, M., Bélanger, D. (Eds.), Reconfiguring Families in Contemporary Vietnam. Stanford University Press, Stanford, CA, pp. 203–236.
Jones, G.W., 1990. Fertility transitions among Malay populations of Southeast Asia: puzzles of interpretation. Population and Development Review 16 (3), 507–537.
Jones, G.W., 1994. Marriage and Divorce in Islamic Southeast Asia. Oxford University Press, Singapore.
Jones, G.W., 1995. Population and the family in Southeast Asia. Journal of Southeast Asian Studies 26 (1), 184–195.
Jones, G.W., 1997. Modernization and divorce: contrasting trends in Islamic Southeast Asia and the West. Population and Development Review 23 (1), 95–114.
Jones, G.W., 2005. The "Flight From Marriage" in South-East and East Asia. Journal of Comparative Family Studies 36 (1), 93–119.
Jones, G.W., 2007. Delayed marriage and very low fertility in Pacific Asia. Population and Development Review 33 (3), 453–478.
Jones, G.W., 2009. Recent fertility trends, policy responses and fertility prospects in low fertility countries of East and Southeast Asia, United Nations Expert Group Meeting on Recent and Future Trends in Fertility. New York, 2–4, December 2009.
Jones, G.W., 2012. Population policy in a prosperous city-state: dilemmas for Singapore. Population and Development Review 38 (2), 311–336.
Jones, G.W., Gubhaju, B., 2009. Factors influencing changes in mean age at first marriage and proportions never marrying in the low-fertility countries of East and Southeast Asia. Asian Population Studies 5 (3), 237–265.
Jones, G.W., Gubhaju, B., 2012. Marriage trends in Southeast Asia. In: Williams, L., Guest, M.P. (Eds.), Demographic Change in Southeast Asia: Recent Histories and Future Directions. Cornell University, Ithaca, NY, pp. 65–91.
Jones, G.W., Shen, H.-H., 2008. International marriage in East and Southeast Asia: trends and research emphases. Citizenship Studies 12 (1), 9–25.
Kuhn, R., 2010. Routes to low mortality in poor countries revisited. Population and Development Review 36 (4), 655–692.
Knodel, J., Havanon, N., Pramualratana, A., 1984. Fertility transition in Thailand: a qualitative analysis. Population and Development Review 10 (2), 297–328.
Knodel, J., Chamratrithirong, A., Debavalya, N., 1987. Thailand's Reproductive Revolution: Rapid Fertility Decline in a Third-World Setting. University of Wisconsin Press, Madison, WI.
Knodel, J., Ruffolo, V.P., Ratanalangkarn, P., Wongboonsin, K., 1996. Reproductive preferences and fertility trends in post-transition Thailand. Studies in Family Planning 27 (6), 307–318.
Leete, R., 1994. The continuing flight from marriage and parenthood among the overseas Chinese in East and Southeast Asia: dimensions and implications. Population and Development Review 20 (4), 811–829.
Lesthaeghe, R., 1983. A century of demographic and cultural change in Western Europe: an exploration of the underlying dimensions. Population and Development Review 9 (3), 411–435.
Mason, K.O., 1997. Explaining fertility transitions. Demography 34 (4), 443–454.
Ministry of Health and Lao Statistics Bureau, 2012. Lao Social Indicator Survey (LSIS), 2011–12. ICF International, Calverton, Maryland.
Myint, N., 1991. Recent levels and trends of fertility and mortality in Myanmar. Asia-Pacific Population Journal 6 (1), 3–20.
National Institute of Statistics and Ministry of Health Cambodia, 2011. Cambodia Demographic and Health Survey (CDHS), 2010. ICF Macro, Calverton, Maryland.
National Population and Talent Division, 2012. A Sustainable Population for a Dynamic Singapore, Population White Paper. Prime Minister's Office, Government of Singapore, Singapore. http://population.sg/whitepaper/resource-files/population-white-paper.pdf (accessed 09.09.13.).
Notestein, F.W., 1953. Economic problems of population change. In: Proceedings of the Eighth International Conference of Agricultural Economists. Oxford University Press, London, pp. 13–31.
Reid, A., 1988a. Southeast Asia in the Age of Commerce, 1450–1680, Vol. 1: The Lands Below the Winds. Yale University Press, New Haven, CT.
Reid, A., 1988b. Female roles in pre-colonial Southeast Asia. Modern Asian Studies 22 (3), 629–645.
Reid, A., 1992. Economic and social change, c. 1400–1800. In: Tarling, N. (Ed.), The Cambridge History of Southeast Asia, Vol. 1, Part 2, from c. 1500 to c. 1800. Cambridge University Press, Cambridge, pp. 116–162.
Reid, A., 1993. Southeast Asia in the Age of Commerce, 1450–1680. In: Expansion and Crisis, vol. 2. Yale University Press, New Haven, CT.
Reid, A. (Ed.), 1996. Sojourners and Settlers: Histories of Southeast Asia and the Chinese. Allen and Unwin, St. Leonards, Australia.
Reid, A., 2001. Sojourners and Settlers: Histories of Southeast Asia and the Chinese. University of Hawaii Press, Honolulu.
Retherford, R., Cho, L.-J., 1973. Comparative analysis of recent fertility trends in East Asia. In: International Population Conference, vol. 2. International Union for the Scientific Study of Population, Liege, pp. 163–181.

Robinson, W.C., Rachapaetayakom, J., 1993. The role of government planning in Thailand's fertility decline. In: Leete, R., Alam, I. (Eds.), The Revolution in Asian Fertility: Dimensions, Causes, and Implications. Clarendon Press, Oxford, UK, pp. 54–66.

Ross, J.A., Poedjastoeti, S., 1983. Contraceptive use and program development: new information from Indonesia. International Family Planning Perspectives 9 (3), 68–77.

Saw, S.H., 1999. The Population of Singapore. Singapore Institute of Southeast Asian Studies, Singapore.

Saw, S.H., 2005. Population Policies and Programmes in Singapore. Institute of Southeast Asian Studies, Singapore. http://muse.jhu.edu/books/9789812305480 (accessed 01.09.13.).

Scott, J.C., 2009. The Art of Not Being Governed: An Anarchist History of Upland Southeast Asia. Yale University Press, New Haven, CT.

Situmorang, A., 2007. Staying single in a married world. Asian Population Studies 3 (3), 287–304.

Smith, D.P., 1981. Illustrative Analysis: Marriage Dissolution and Remarriage in Sri Lanka and Thailand. World Fertility Survey Scientific Report No.7. International Statistical Institute, Voorburg, the Netherlands.

Tey, N.-P., 2007. Trends in delayed and non-marriage in Peninsular Malaysia. Asian Population Studies 3 (3), 243–261.

United Nations, 2011a. World Urbanization Prospects, the 2011 Revision. United Nations, New York.

United Nations, 2011b. World Contraceptive Use 2010. United Nations, New York.

United Nations, 2012. World Marriage Data. http://www.un.org/en/development/desa/population/ (accessed 15.08.13.).

United Nations, 2013a. World Population Prospects, the 2012 Revision. In: Comprehensive Tables, vol. 1. United Nations, New York.

United Nations, 2013b. World Population Prospects, the 2012 Revision. In: Demographic Profiles, vol. 2. United Nations, New York.

United Nations, 2013c. 2013 Update for the MDG Database. http://www.un.org/en/development/desa/population/ (accessed 15.08.13.).

Vanlandingham, M., Hirschman, C., 2001. Population pressure and fertility in pre-transition Thailand. Population Studies 55 (3), 233–248.

Van Esterik, P., 1982. Women of Southeast Asia. Center for Southeast Asian Studies, Northern Illinois University, Dekalb, IL.

Wolters, O.W., 1999. History, Culture, and Region in Southeast Asian Perspectives. Southeast Asia Program, Cornell University, Ithaca, NY.

Wongboonsin, K., Ruffolo, V.P., 1995. Sex preference for children in Thailand and other Southeast Asian countries. Asia-Pacific Population Journal 10 (1), 43–62.

Xenos, P., Kabamalan, M., 2007. Emerging forms of union formation in the Philippines. Asian Population Studies 3 (3), 263–286.

Yang, W.S., Lu, M.C., 2010. Asian Cross-border Marriage Migration: Demographic Patterns and Social Issues. Amsterdam University Press, Amsterdam.

Yap, M.t., 1995. Singapore's 'three or more' policy: the first five years. Asia-Pacific Population Journal 10 (1), 39–52.

Relevant Websites

http://www.bps.go.id/eng/index.php – Badan Pusat Statistik (Statistics Indonesia).
http://www.statistics.gov.my/portal/index.php?lang=en – Department of Statistics (Malaysia).
http://www.singstat.gov.sg/ – Department of Statistics (Singapore).
http://www.gso.gov.vn/default_en.aspx?tabid=491 – General Statistics Office (Vietnam).
http://www.nsc.gov.la/ – Lao Statistics Bureau.
http://www.measuredhs.com/ – Measure DHS (Demographic and health surveys).
http://www.nis.gov.kh/ – National Institute of Statistics (Cambodia).
http://www.census.gov.ph/ – National Statistics Office (Philippines).
http://web.nso.go.th/ – National Statistics Office (Thailand).
http://www.un.org/en/development/desa/population/ – Population Division (United Nations).

Fertility Transition: Sub-Saharan Africa

David Shapiro, The Pennsylvania State University, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Fertility transition – the decline of fertility from high levels – emerged in the Third World following World War II. In sub-Saharan Africa the onset of fertility transition was delayed, and the pace of transition has also been distinctly slower than elsewhere in the developing world. Limited economic growth and development, low educational attainment, high infant and child mortality, and a strong cultural emphasis on the extended family all appear to have contributed to the delayed and slow fertility decline in the region. There is considerable diversity of experience with fertility decline among countries in the region, and stalling of fertility decline has been common.

Introduction

Fertility transition is part of the broader concept of the demographic transition, which describes the shift from the high rates of fertility and mortality that characterized most of human history to low rates of fertility and mortality that prevail in industrialized countries and increasingly in the developing world. The demographic transition unfolded during much of the nineteenth and early twentieth centuries in Europe, the United States, and elsewhere in what are now high-income countries. By contrast, it was not until after World War II that this transition emerged in the developing world (much of Asia, Africa, and Latin America).

Fertility Transition in the Developing World – a Comparative Perspective

Figure 1, based on the most recent United Nations estimates and projections, shows period total fertility rates and provides a comparative perspective on fertility transition in sub-Saharan Africa and other major regions in the developing world, dating back to 1950. The period total fertility rates (TFRs) indicate the number of children women would have during their lifetimes if they were to bear children at the rates prevailing in each particular period.

It is clear from the figure that fertility transition began to emerge in the Third World in the early decades of the second half of the twentieth century. Fertility declines are apparent for Latin America, Asia, and Northern Africa during the 1960s and 1970s. At the same time, sub-Saharan Africa stands out for the delay in initiation of fertility decline; overall decline is apparent only after 1980. In addition, following initiation of decline, the pace of that decline has been distinctly slower in sub-Saharan Africa than elsewhere.

As a consequence of both the later initiation and slower pace of decline, estimated current total fertility (for the period 2010–15) is substantially higher for sub-Saharan Africa as compared to the other regions. While Asia and Latin America and the Caribbean have estimated total fertility rates of 2.2 and the estimated rate for Northern Africa is 3, the corresponding figure for sub-Saharan Africa is 5.1. Hence, sub-Saharan Africa was the last major world region to initiate fertility decline, and its fertility remains substantially higher than fertility in the other major regions of the developing world.

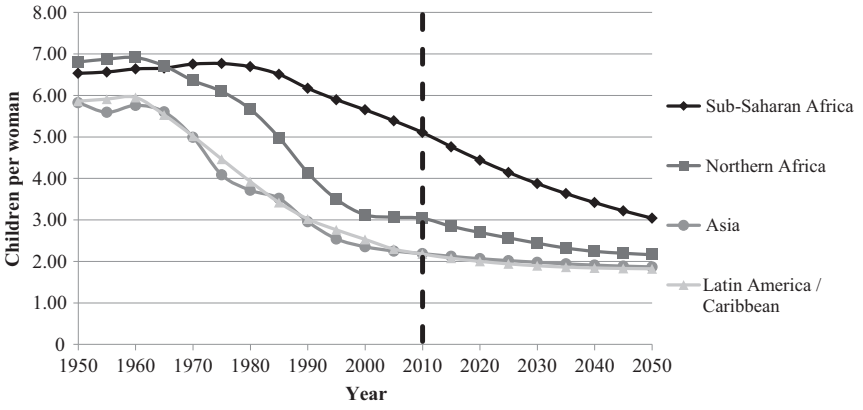

Figure 1 Total fertility, by region, 1950–2050. Source: Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat, World Population Prospects: The 2012 Revision, http//esa.un.org/unpd/wpp/index.htm.

Fertility Transition in Sub-Saharan Africa, by Region

Within sub-Saharan Africa, there is a notable difference in fertility between Southern Africa and the other sub-regions, as is evident from Figure 2 (also from United Nations data). Fertility estimates for Southern Africa (dominated by South Africa) look much like those for Asia and Latin America and the Caribbean, declining from early on in the post-World War II period and falling to well below three children per woman for the period from 2010 to 2015. However, the fertility rates for the other three sub-regions show delayed declines, beginning in the 1980s in Eastern and Western Africa and in the 1990s in Middle Africa. In addition, Middle Africa has had the highest estimated fertility for the past 25 years. For all three sub-regions there is sustained high fertility at present. For the period 2010–15, the estimated TFR for Middle Africa is 5.7, while for Western Africa, Eastern Africa, and Southern Africa the corresponding figures are 5.6, 4.9, and 2.5, respectively.

Fertility Transition in Sub-Saharan Africa, by Country

Further disaggregation, to the country level, shows substantial variation in fertility behavior. United Nations medium estimates of total fertility rates for individual countries for the period 2010–15 indicate that the total fertility rate is less than 4 in all five countries from Southern Africa, in Zimbabwe and Ghana, and in a few small island nations. These are the countries where fertility transition is most advanced. By contrast, half a dozen countries have estimated TFRs in excess of 6, including Niger (with an estimate in excess of 7.5), Mali, Somalia, Chad, Burundi, and Nigeria, the region's most populous country. In between, numerous countries are in various stages of fertility transition (Table 1).

In the literature on fertility transition, it is customary to consider the transition to have been initiated when the total fertility rate falls by at least 10% from some preexisting peak level. The United Nations Population Division reports estimates of the TFR for each country for 5-year periods beginning in 1950. There are 42 countries in sub-Saharan Africa with population of at least one million. Peak fertility in these countries typically is around seven children per woman, and varies from just under six in the Central African Republic, Gabon, and Lesotho to more than eight children per woman in Kenya and Rwanda.

In five of the 42 countries, estimated fertility in 2005–10 is greater than 90% of peak fertility. This is the case for Chad, Gambia, Mali, Niger, and Somalia. These, then, are the countries in which fertility transition has not yet begun. The (unweighted) average total fertility rate in these five countries is 6.8, and the range is from 5.8 (Gambia) to 7.6 (Niger). The average TFR constitutes 94% of the average peak fertility of 7.3 in these five pre-transitional countries.

There are 17 countries, distributed almost equally among Eastern, Middle, and Western Africa, which are in the early stages of fertility transition. Operationally, this group is defined here as consisting of countries in which the estimated fertility in 2005–10 represents more than 75 to 90% of the peak fertility. That is, in these countries the TFR has fallen by at least 10% from its peak, but not (quite) by as much as 25%. This is an arbitrary classification scheme, but it helps distinguish the countries in the region with respect to their position in the ongoing fertility transition in the region. The average total fertility rate in this group of countries is 5.7, and the range is from 4.9 (Central African Republic) to 6.5 (Angola and Burundi). The average TFR represents 82% of the average level of peak fertility.

Another dozen countries may be classified as mid-transitional. This group is operationalized here as countries in which the estimated TFR represents 60–75% of the peak fertility. Seven of these countries are in Western Africa and four others are in Eastern Africa. For this group, the average TFR is 4.9, with the range being from 4.2 (Ghana) to 5.3 (Ethiopia and Guinea-Bissau). Average peak fertility for this group is 7.2, so the average estimated fertility for 2005–10 represents 68% of the average peak fertility.

Finally, there are eight countries in which fertility has declined by more than 40% from its peak level. These are the countries in sub-Saharan Africa in which fertility transition is advanced. They include three countries from Eastern Africa, namely Kenya, Mauritius, and Zimbabwe, and all five countries

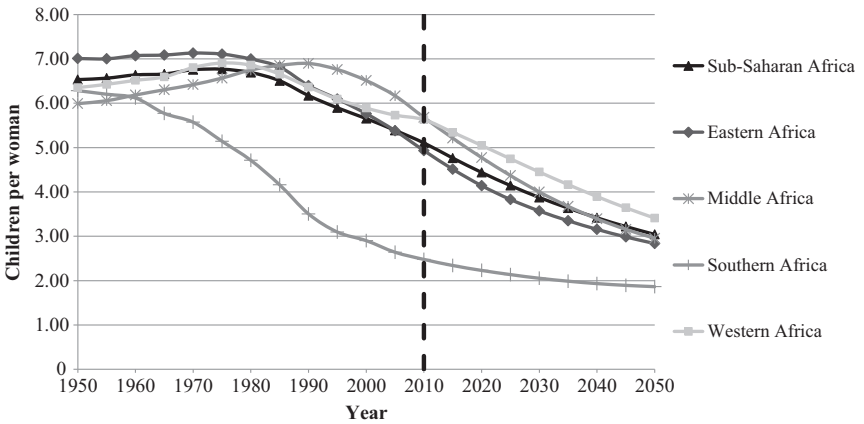

Figure 2 Total fertility, sub-Saharan Africa and its sub-regions, 1950–2050. Source: Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat, World Population Prospects: The 2012 Revision, http://esa.un.org/unpd/wpp/indexhtm.

Table 1 Fertility in sub-Saharan Africa by stage of fertility transition: estimated fertility as of 2005–10 and peak fertility

Country	*Peak TFR*	*2005–10 TFR*	*Last/peak*
Pre-transitional			
Chad	7.41	6.85	92
Gambia	6.34	5.79	91
Mali	7.09	6.8	96
Niger	7.77	7.58	98
Somalia	7.7	7.1	92
Average	7.3	6.8	94
Early transition (last is 75+ to 90% of peak)			
Burundi	7.59	6.52	86
Malawi	7.6	5.83	77
Mozambique	6.6	5.57	84
South Sudan	6.92	5.43	78
Uganda	7.12	6.38	90
United Republic of Tanzania	6.8	5.58	82
Zambia	7.43	5.9	79
Angola	7.4	6.5	88
Cameroon	6.7	5.21	78
Central African Republic	5.95	4.85	82
Congo Republic	6.29	5.1	81
Democratic Republic of Congo	7.24	6.5	90
Benin	7.01	5.31	76
Burkina Faso	7.17	6.08	85
Guinea	6.63	5.39	81
Liberia	6.96	5.23	75.1
Nigeria	6.76	6.01	89
Average	7.0	5.7	82
Mid-transition (last is 60–75% of peak)			
Eritrea	6.97	5.2	74.6
Ethiopia	7.43	5.26	71
Madagascar	7.3	4.83	66
Rwanda	8.43	5.13	61
Gabon	5.72	4.29	75.00
Cote d'Ivoire	7.93	4.89	62
Ghana	6.95	4.22	61
Guinea-Bissau	7.41	5.28	71
Mauritania	6.79	4.96	73
Senegal	7.46	5.11	68
Sierra Leone	7.03	5.16	73
Togo	7.28	4.89	67
Average	7.2	4.9	68
Advanced transition (last is <60% of peak)			
Kenya	8.11	4.8	59
Mauritius	6.2	1.58	25
Zimbabwe	7.4	3.9	53
Botswana	6.7	2.9	43
Lesotho	5.86	3.37	58
Namibia	6.6	3.4	52
South Africa	6.3	2.55	40
Swaziland	6.87	3.75	55
Average	6.8	3.3	49

Note: First column shows the peak estimated TFR since 1950; second column gives the estimated TFR for the period 2005–10; and third column gives the latter as a percentage of the former.
Source: United Nations, Department of Economic and Social Affairs, Population Division, 2013. World Population Prospects: The 2012 Revision, CD-Rom Edition. New York. – data on estimated fertility from 1950 until 2010 for 42 countries with population in excess of one million.

in Southern Africa: Botswana, Lesotho, Namibia, South Africa, and Swaziland. The average total fertility rate in 2005–10 for this group of countries was 3.3. Mauritius and Kenya were the outliers, with TFRs of 1.6 and 4.8, respectively. The remaining countries had TFRs ranging from 2.55 (South Africa) to 3.9 (Zimbabwe). The average TFR of 3.3 for the group represents 49% of the peak estimated fertility.

From this detailed review of experience in individual countries, it is apparent that the vast majority of countries in sub-Saharan Africa have initiated fertility transition, although several countries have not done so and, using the criterion of a 10% decline from peak fertility, two other countries (the Democratic Republic of the Congo and Uganda) just barely qualify as having begun their fertility transitions. Among those countries in which fertility decline has been seen, there is clearly substantial variation in the pace of the observed declines, and Southern Africa stands out for more substantial declines. In the subsequent sections we address the question of which factors seem to be most relevant in influencing the pace of fertility decline. Before addressing that question and other more detailed aspects of the fertility transition in the region, it is useful to consider why sub-Saharan Africa was so slow to initiate fertility decline, as compared to other parts of the developing world.

Why Was Fertility Transition So Slow to Emerge in Sub-Saharan Africa?

The delay in initiation of fertility transition in sub-Saharan Africa was the subject of considerable speculation, particularly during the 1970 and 1980s (see, e.g., Lesthaeghe, 1989). Several different explanations were set forth, reflecting economic, social, and cultural factors. Before examination of these explanations, we consider a conceptual framework that is useful for examining fertility transition.

The Easterlin Framework for Fertility Analysis

Demographic transition, as it unfolded in Europe, the United States, and elsewhere, was broadly associated with economic growth and development. Early work by Gary Becker (1960) and by Becker and Lewis (1973) emphasized that economic development was likely to contribute to decreased demand for numbers of children, and greater expenditures on education and health per child, characterized as higher quality of children. Richard Easterlin (1975; Easterlin and Crimmins, 1985) elaborated an economic approach to fertility that incorporated Becker's approach focused on demand for children and added elements pertaining to supply of children and costs of fertility regulation.

In brief, the Easterlin model emphasizes three broad categories through which the basic determinants of fertility operate and which influence, in turn, the proximate determinants of fertility. These three categories are the demand for children (the number of surviving children parents would want if fertility regulation were costless); the supply of children (the number of surviving children parents would have if they did not deliberately limit fertility); and the costs (subjective and objective) of fertility regulation.

Following Becker, a couple's demand for children is treated as analogous to the demand for goods and services. In particular, demand depends on household income, on the

cost (price) of children, and on parents' tastes or preferences for children relative to other goods and services that provide satisfaction (utility) to the couple. Other things being equal, higher income is expected to be associated with a greater demand for children (i.e., children are assumed to be a normal good). However, greater demand for children may be realized at least in part by greater resource endowments per child rather than simply by an increase in the number of children. In this respect, the demand for children or child services may be seen as comparable to the demand for consumer durable goods more generally, where higher income often translates into increased demand for quality rather than simply increased quantity.

Greater resource endowments (i.e., higher expenditures) per child are typically described in the economics literature on fertility as expenditures for child quality, and there is considerable discussion in the literature of quality–quantity tradeoffs. In empirical work, the two areas most commonly studied where resources are expended to enhance child quality are education and health.

The greater the cost of children, the lower is the quantity demanded. The cost of children includes not simply the direct costs of goods and services that are complementary to children, but also the indirect or opportunity cost of the mother's time spent in child care (often measured using estimates of the woman's earning power, or potential wage rate, in the labor market). Indeed, for children of given quality, it is typically differences among women in the opportunity cost of time that result in differences across households in the cost of children.

The stronger are a couple's (relative) preferences for children, the greater the demand for children, other things being equal. In considering this aspect of the demand for children, it is necessary to take into consideration the tastes relating to child quality. More generally, economics does not have a lot to say about tastes, but presumably this factor may be related to cultural factors such as ethnicity or religion, and it may also be related to individual factors such as educational attainment (e.g., greater mother's education has been found to be related to stronger preferences for greater child's education).

The supply of children reflects two factors: a couple's natural fertility and the chances of child survival. Natural fertility refers to the number of births a couple would have if they took no action aimed at limiting fertility behavior (i.e., in societies in which deliberate fertility control is not practiced). Cultural differences in behaviors that influence the likelihood of a birth (e.g., in duration of breast feeding or in the observance of periods of postpartum abstinence) can lead to differences in natural fertility between different natural fertility populations. Since the potential supply of children in the Easterlin framework refers to the number of children surviving to adulthood, it is clear that supply also varies inversely with the level of mortality. Hence, reductions in mortality increase the supply of children.

The costs of fertility regulation incorporate a couple's attitudes toward and access to fertility control methods and supplies. There are two types of costs of fertility regulation: psychic costs and market costs. Psychic costs refer to the displeasure associated with the practice or idea of fertility control, while market costs are the money and time costs necessary to learn about and use specific contraceptive techniques.

Couples have a motivation for fertility regulation if the potential supply exceeds the quantity of children demanded. This does not necessarily translate into efforts to control fertility – that depends also on the costs of fertility regulation. Given the extent of the motivation to limit fertility, the lower the costs of fertility regulation the more likely a couple is to opt for contraception. In this framework, then, family planning programs can lead to fertility reduction via reducing both the market costs and the psychic costs of contraception.

The basic determinants of fertility behavior include underlying socioeconomic conditions, or what Easterlin and Crimmins describe as modernization variables such as education, urbanization, and modern sector employment, as well as cultural factors such as ethnicity and religion, and other determinants such as genetic factors. These basic determinants influence fertility through their impact on the demand for children, the supply of children, and/or the costs of fertility regulation.

Between the basic determinants of fertility and realized fertility behavior are the proximate determinants of fertility. That is, the basic determinants influence fertility only indirectly, through their influence on the proximate determinants. It is these proximate determinants which are seen as determining fertility directly. Following Davis and Blake (1956) and Bongaarts (1978), the proximate determinants include factors such as extent of exposure to intercourse (heavily influenced by age at marriage, but note that cultural practices regarding intercourse outside of marriage will also be relevant here), fecundability (including frequency of intercourse), duration of postpartum infecundability (related especially to breast feeding durations, an important factor keeping fertility in sub-Saharan Africa below its biological maximum), sterility, and the use of deliberate fertility control including contraception and induced abortion.

From the perspective of this conceptual framework, the onset of fertility transition is likely to reflect declining demand for numbers of children along with increasing supply, eventually resulting in excess supply and hence a motivation for fertility control. As fertility regulation is adopted, fertility begins to decline. Consider now the explanations offered for the delayed emergence of fertility transition in sub-Saharan Africa.

Economic Development and Fertility

With respect to the association between economic development and fertility, it was apparent from historical evidence that while there was no indication that a specific level of development or industrialization would trigger the onset of fertility transition, in the wake of economic growth and development fertility decline would typically occur. This was evident in the developing world as well (Bryant, 2007). From this perspective, then, it was argued that sub-Saharan Africa's comparatively slow economic advancement was delaying the onset of fertility decline.

The region has remained predominantly agricultural throughout the years: as of 2010, it is estimated that more than 60% of employment in the region is in agriculture, higher than

anywhere else in the developing world. Becker's early work on fertility (1960) emphasized the importance of the net cost of children as an influence on the demand for children. To the extent that children in rural areas are able to contribute to agricultural production – as is typically the case in the land-abundant, low-technology slash-and-burn agriculture that typifies much of the region – this will lower their net cost and hence, other things being equal, be associated with greater numbers of children demanded and hence higher fertility.

Education, Infant and Child Mortality, and Fertility Transition

In addition to the direct economic influences on fertility, there are socioeconomic factors that contribute to high fertility in the region. Most notable in this regard are the comparatively low level of education among women in much of the region (Shapiro, 2012), and the relatively high level of infant and child mortality. Women's education and fertility are inversely related, for a variety of reasons. Women with more education typically desire more education for their children, reflecting a quality–quantity trade-off with respect to fertility, resulting in lower numbers of children demanded. Such women are also more likely than their lesser-educated counterparts to have some say in couples' decisions about childbearing. Better-educated women are also more likely to use contraception and to use more effective contraceptive methods than their lesser-educated counterparts. Hence, the low levels of education of women in sub-Saharan Africa, with some countries having high proportions of women with no schooling, tended to inhibit fertility decline.

High rates of infant and child mortality also helped perpetuate high fertility. Current United Nations estimates of mortality for sub-Saharan Africa, Asia, and Latin America show that infant mortality rates are highest by far for sub-Saharan Africa, being more than twice as high as those in Asia and nearly four times as high as in Latin America. Under-five mortality is almost three times higher in sub-Saharan Africa than in Asia and nearly five times higher than in Latin America. In the Easterlin framework for fertility analysis, infant and child mortality serve as a check on the supply of children. Hence, the likelihood of excess supply and hence a motivation for fertility control is low. With adult children serving as the primary means of old-age support for parents, high infant and child mortality translate into high demand for numbers of births, to assure adequate old-age support.

Cultural Factors and Fertility Transition

In addition to these socioeconomic factors, there were a number of cultural factors unique to sub-Saharan Africa that promoted or facilitated ongoing high fertility (cf Lesthaeghe, 1989; Romaniuk, 2011). Probably the most important in this regard was the emphasis on the extended family and the lineage. Children were viewed not simply as the offspring of their biological parents, but also as reinforcing the strength of the extended family and the clan. Hence, the birth of a child was not simply a matter for the parents, but because it expanded the lineage, it was a concern for the larger web of relations as well. Further, the widespread practice of child fostering within the extended family, by providing opportunities to foster out children, served to reduce the costs of children to their parents and thus contributed to high demand for numbers of children.

Another cultural aspect particularly relevant in sub-Saharan Africa is polygyny. While over time the incidence of polygyny has declined, it remains prevalent in a number of countries. Romaniuk (2011: p. 9), while noting that polygynous women often have lower fertility than monogamous women, points out that this is largely a selection effect, and argues that polygyny contributes to "a more complete mobilization of women into marriage and reproduction."

Onset, Current State, and Specificities of Fertility Transition in Sub-Saharan Africa

The World Fertility Survey, carried out during the 1970s and early 1980s in 42 countries, including 11 in sub-Saharan Africa, confirmed that at the level of national populations the region had still not experienced the onset of fertility transition. However, by the early 1990s, surveys carried out in the late 1980s as part of the Demographic and Health Surveys (DHS) program funded by the United States Agency for International Development provided clear evidence of declining fertility. Botswana, Ghana, Kenya, and Zimbabwe were the first countries to show this evidence, and during the 1990s additional DHSs provided indications of fertility transition in numerous other countries, albeit at different paces. As noted earlier, there remains a small group of countries in which fertility transition has not yet been initiated.

Fertility Transition and Economic Conditions

While experience elsewhere in both the industrialized world and the developing world suggests that economic growth and development will ultimately encourage fertility decline, there is also literature suggesting that economic adversity may contribute to initiation of fertility transition. This notion of crisis-led transition was discussed early on by Lesthaeghe (1989), and a study by the National Research Council (1993) found evidence of economic adversity contributing to lower fertility, primarily via delays in initiation of marriage and first births. Country-specific studies, such as that by Eloundou-Enyegue et al. (2000) for Cameroon, also found evidence that economic adversity could contribute to the onset of fertility transition.

Onset of Fertility Transition by Place of Residence

While the onset of fertility decline at the national level was not apparent prior to the end of the 1980s, subsequent work has made it clear that fertility transition in sub-Saharan Africa had begun in a number of urban places well before then (Garenne and Joseph, 2002; Garenne, 2008). This initiation of transition in urban areas was similar to the fertility transitions that had taken place earlier in Asia and Latin America. Shapiro and Tambashe (2002), in assessing fertility transition in the region, suggested a three-phase transition: initially, fertility

began to decline in urban places while remaining stable in rural areas; subsequently, fertility declined in both urban and rural places, but more rapidly in urban areas; and finally (with fertility having reached comparatively low levels in urban places) fertility decline in rural areas outpaced that in urban places. They also noted that for the most part, urban fertility transition was especially pronounced in the large capital cities.

The differences by place of residence in the timing of the onset of fertility transition are consistent with theoretical emphasis on the net costs of children: both children's prospective contribution to agricultural production in rural areas (noted above) and the higher cost of living in urban places make the net cost of children higher in urban settings. In addition, since infant and child mortality rates are typically greater in rural areas, this means that the supply of children in the Easterlin framework for fertility analysis (defined as the number of births surviving to adulthood) is greater in urban areas, and this in turn increases the likelihood of excess fertility and hence a demand for fertility control in urban places.

The especially low fertility that is often apparent in capital cities also appears in some cases to reflect the effects of economic adversity. Antoine (2006), in a study of four capital cities in the region, found evidence that economic adversity had contributed to slowing of rates of entry into marital unions and consequently to lower levels of fertility. Gurmu and Mace (2008) reported similar findings for Addis Ababa in Ethiopia.

Stalling of Fertility Transitions and Factors Influencing the Pace of Fertility Decline

Various countries in sub-Saharan Africa are at different stages of fertility transition, as noted above. In addition, as Bongaarts (2008), Shapiro and Gebreselassie (2008), and others have emphasized, stalling of fertility transition – i.e., a halt in the decline of fertility for some period of time – has been observed in a number of countries in the region. Observations of stalling have most commonly been made using data from the DHSs, since multiple surveys for the same country allow for a direct examination of stalling.

For example, there are 27 countries in sub-Saharan Africa that have had multiple DHSs. Three of these countries have not yet begun fertility transition. But of the remaining 24 countries, just over half have reported a TFR in one survey that was not lower than the TFR reported in the previous survey. This simple measure of stalling is but one of multiple possible measures. Of the 13 countries with stalls in the reported TFRs, five subsequently showed a resumption of fertility decline. At present, with eight countries in a situation of fertility stall, the phenomenon of stalling appears to be more common than was the case earlier in the African fertility transition.

Bongaarts (2008) has suggested that the slower pace of socioeconomic progress in sub-Saharan Africa as compared to elsewhere in the developing world may be pertinent in accounting for the slower pace of fertility decline in the region and the presence of stalling of fertility transitions. Shapiro and Gebreselassie (2008) and Shapiro (2012) have analyzed data from successive pairs of DHSs carried out in sub-Saharan Africa, examining the relationships between the magnitudes of fertility decline between surveys and changes in various factors that are presumed to influence the pace of fertility decline. Those studies have found that women's educational attainment and infant and child mortality are the key factors that appear to be related to the pace of fertility decline within the region. Over all, the longer-term trends show increases in women's educational attainment and declines in infant and child mortality occurring as fertility tends to decline. Multivariate regression analyses of data from DHSs, at national, urban–rural, and regional levels within countries, show that more rapid increases in women's education are associated with faster declines in fertility, and that more rapid declines in infant and child mortality are also associated with faster declines in fertility.

The use of modern contraception was a major factor in contributing to fertility transition elsewhere in the developing world. However, it appears that the relationship between contraceptive use and fertility is weaker in sub-Saharan Africa than elsewhere (Westoff and Bankole, 2001), and the analyses referred to in the preceding paragraph found only limited evidence of increased contraceptive use as a factor influencing the pace of fertility decline. One factor that may be relevant here is recourse to abortion. While there are difficulties gathering information on abortion (such information is not available in the DHSs, for example), there are a number of studies that suggest that recourse to abortion, particularly among younger women, is not uncommon in many places in the region, and especially larger urban places. Rossier (2010), in one such study, suggested that in the early stages of fertility transition, there is likely to be growing recourse to abortion as a means of limiting fertility, but that as transition proceeds, there will be greater reliance on modern contraception to control fertility, and a corresponding reduction in the incidence of abortion.

Is the African Fertility Transition Different from the Transitions Elsewhere?

A question of some interest is whether fertility transition in sub-Saharan Africa is qualitatively different from the earlier fertility transitions in Asia and Latin America (as opposed to simply unfolding more slowly). This issue was first raised by Caldwell et al. (1992), who suggested that fertility transition in sub-Saharan Africa would be characterized by similar declines across all age groups, in contrast to the earlier transitions in which fertility declines were initially concentrated among older women. A recent paper by Bongaarts and Casterline (2013) failed to find evidence of similar declines in fertility across age groups. However, comparing the current transition in sub-Saharan Africa with the earlier transitions elsewhere when they were at a comparable overall level of fertility, Bongaarts and Casterline did find some clear differences. Ideal family size is distinctly higher in sub-Saharan Africa than it was in Asia and Latin America when those regions were at a similar stage of the transition as reflected in the overall total fertility rate. Bongaarts and Casterline argue that this comparatively high ideal family size contributes to the slower pace of fertility transition in the region as compared to elsewhere in the developing world. Africa also is characterized by longer birth intervals.

Moultrie et al. (2012) have also addressed the question of whether fertility transition in sub-Saharan Africa is different

from earlier transitions. They focus on postponement of fertility and the corresponding birth intervals, using data from numerous DHSs from two dozen countries in the region. Their analyses indicate that there has been a lengthening of birth intervals for each country in the region that they analyze, but that the lengthening of these intervals does not show systematic variation either by woman's age or parity. Noting that this appears to differ from earlier fertility transitions elsewhere, they argue that "the fact that, across the 24 countries studied, median birth intervals have lengthened largely independently of either age or parity provides empirical support for our contention that the African fertility transition is distinctively different" (Moultrie et al., 2012: p. 251).

Future Prospects

The United Nations Population Division projects that by the middle of this century, the TFR for sub-Saharan Africa will have fallen from its current level of 5.1 to a level of 3.2 (Asia and Latin America are projected to have fertility below 2 by mid-century). sub-Saharan Africa is already the most rapidly growing region of the world, and United Nations projections anticipate that the region's population will more than double between 2015 and 2050, to exceed two billion people. The region's share of world population is projected to increase from about 13 to almost 22%. At the same time, however, the prevalence of stalling of fertility transition in the region calls into question these projections.

Whether this projected continuation of fertility decline will be realized will clearly depend on future socioeconomic development in the region. If the longer-term trends toward increasing educational attainment and declining infant and child mortality continue, prospects for ongoing fertility transition will be enhanced. Shapiro (2012) found that increased schooling of women to and beyond the secondary level appeared to be associated with an acceleration of the pace of fertility decline; hence, increased educational attainment of women may lead to a faster pace of fertility transition in the coming decades.

Likewise, economic improvement that would enhance the payoffs to investments in children's schooling and health would also be expected to contribute to ongoing fertility transition. This would come about via increased incentives to invest in child quality, which would likely result in declines in ideal family size and hence reduced demand for number of children. But given the weak state of many economies in the region, it is not at all clear whether such improvement will be realized.

See also: Children, Value of; Family Size Preferences; Family Theory: Economics of Childbearing; Fertility Theory; Fertility Transition: China; Fertility Transition: Cultural Explanations; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Fertility Transition: Latin America and the Caribbean; Fertility Transition: Middle East and North Africa; Fertility Transition: South Asia; Fertility Transition: Southeast Asia; Fertility Trends in the Formerly Socialist Countries of Europe; Fertility and Culture: Anthropological Insights; Second Demographic Transition.

Bibliography

Antoine, P., 2006. Analyse biographique de la transformation des modèles matrimoniaux dans quatre capitales africaines: Antananarivo, Dakar, Lomé et Yaoundé. Cahiers québecois de démographie 35, 5–38.

Becker, G.S., 1960. An economic analysis of fertility. In: Demographic and Economic Change in Developed Countries, a Conference of the Universities–National Bureau Committee for Economic Research. Princeton University Press, Princeton, NJ, pp. 209–231.

Becker, G.S., Lewis, H.G., 1973. On the interaction between the quantity and quality of children. The Journal of Political Economy 81, S279–S288.

Bongaarts, J., 1978. A framework for analyzing the proximate determinants of fertility. Population and Development Review 4, 105–132.

Bongaarts, J., 2008. Fertility transitions in developing countries: progress or stagnation? Studies in Family Planning 39, 105–110.

Bongaarts, J., Casterline, J., 2013. Fertility Transition: Is Sub-Saharan Africa Different? Paper Presented at the Annual Meeting of the Population Association of America (April), New Orleans, LA.

Bryant, J., 2007. Theories of fertility decline and the evidence from development indicators. Population and Development Review 33, 101–127.

Caldwell, J., Orubuloye, I.O., Caldwell, P., 1992. Fertility decline in Africa: a new type of transition? Population and Development Review 18, 211–242.

Davis, K., Blake, J., 1956. Social structure and fertility: an analytic framework. Economic Development and Cultural Change 4, 211–235.

Easterlin, R.A., 1975. An economic framework for fertility analysis. Studies in Family Planning 6, 54–63.

Easterlin, R.A., Crimmins, E.M., 1985. Theoretical framework. In: Easterlin, R.A., Crimmins, E.M. (Eds.), The Fertility Revolution: A Supply–Demand Analysis. University of Chicago Press, Chicago, IL.

Eloundou-Enyegue, P., Stokes, C.S., Cornwell, G., 2000. Are there crisis-led fertility, declines? evidence from central Cameroon. Population Research and Policy Review 19, 47–72.

Garenne, M., 2008. Fertility Changes in Sub-Saharan Africa. DHS Comparative Reports No. 18. Macro International Inc., Calverton, Maryland, USA.

Garenne, M., Joseph, V., 2002. The timing of the fertility transition in sub-Saharan Africa. World Development 30, 1835–1843.

Gurmu, E., Mace, R., 2008. Fertility decline driven by poverty: the case of Addis Ababa, Ethiopia. Journal of Biosocial Science 40, 339–358.

Lesthaeghe, R. (Ed.), 1989. Reproduction and Social Organization in Sub-Saharan Africa. University of California Press, Berkeley, CA.

Moultrie, T., Takudzwa, S., Timaeus, I., 2012. Birth intervals, postponement, and fertility decline in Africa: a new type of transition? Population Studies 66, 241–258.

National Research Council. Working Group on Demographic Effects of Economic and Social Reversals, 1993. Demographic Effects of Economic Reversals in Sub-Saharan Africa. National Academy Press, Washington, DC.

Romaniuk, A., 2011. Persistence of high fertility in Tropical Africa: the case of the Democratic Republic of the Congo. Population and Development Review 37, 1–28.

Rossier, C., 2010. Avortement provoqué et baisse de la fécondité en Afrique de l'Ouest [Induced abortion and fertility decline in West Africa]. In: Fassassi, R., Vignikin, K., Vimard, P. (Eds.), La régulation de la fécondité en Afrique: Transformations et différenciations au tournant du XXIe siècle. Academia Bruylant Louvain-la-Neuve, Belgium, pp. 155–183.

Shapiro, D., 2012. Women's education and fertility transition in sub-Saharan Africa. Vienna Yearbook of Population Research 2012 (10), 9–30.

Shapiro, D., Gebreselassie, T., 2008. Fertility transition in sub-Saharan Africa: falling and stalling. African Population Studies 23, 3–23.

Shapiro, D., Tambashe, B.O., 2002. Fertility transition in urban and rural sub-Saharan Africa: preliminary evidence of a three-stage process. Journal of African Policy Studies 8, 103–127.

Tabutin, D., Schoumaker, B., 2004. The demography of sub-Saharan Africa from the 1950s to the 2000s: a survey of changes and a statistical assessment. Population 59, 457–556.

United Nations, Department of Economic and Social Affairs, Population Division, 2013. World Population Prospects: The 2012 Revision. CD-Rom Edition, New York.

Westoff, C., Bankole, A., 2001. The Contraception – Fertility Link in Sub-Saharan Africa and in Other Developing Countries. DHS Analytical Studies No. 4. ORC Macro, Calverton, Maryland.

Fertility Trends in the Formerly Socialist Countries of Europe

Elizabeth Brainerd, Brandeis University, Waltham, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Avdeev, volume 8, pp. 5602–5605, © 2001, Elsevier Ltd.

Abstract

The formerly socialist countries of Europe have experienced a remarkable demographic transformation since the fall of the Berlin Wall in 1989. On many dimensions of fertility and family formation, much of the region now looks like Western Europe: below-replacement fertility rates, rising age at first marriage and first birth, and high and increasing nonmarital births characterize many countries formerly distinguished by replacement-level fertility and early, near-universal marriage and childbearing. Some of these changes appear to be driven by increased opportunities and rising returns to education for women, but much research on these issues remains to be done.

Under socialism, most countries of Eastern Europe and the former Soviet Union exhibited a distinctive pattern of fertility and family formation: marriage and first birth occurred at relatively young ages; births occurred predominantly within registered marriages; total fertility rates were at or near the replacement level of fertility (approximately 2.1 children per woman); and abortion rates were much higher than in other developed countries. After the collapse of socialism in 1989–91, these features of the demographic landscape in the region changed rapidly and profoundly. Across virtually all formerly socialist countries of Europe, fertility rates declined sharply in the 1990s and remain well below replacement levels. First births occur at significantly older ages and are much more likely to be registered as nonmarital births. Abortion rates have declined notably as well. In all dimensions of fertility, the formerly socialist countries of Europe appear to have converged with those of Western Europe and fertility patterns are now similar across almost all European countries, both East and West.

For the purposes of this article, the 'formerly socialist countries of Europe' refers to Albania, Bulgaria, the Czech Republic, Hungary, Poland, Romania, the Slovak Republic, the countries of the former Yugoslavia (Bosnia and Herzegovina, Croatia, Kosovo, Macedonia, Montenegro, Serbia and Slovenia; however demographic data are lacking for many of these successor states), as well as the former Soviet republics of Belarus, Estonia, Latvia, Lithuania, Moldova, Russia and Ukraine. The fertility data in this article are from the most recent national statistical or demographic yearbooks published by each country. When such data are missing, data from Eurostat (http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database) or Demoscope (http://demoscope.ru/weekly/pril.php) are used.

Fertility Trends under Socialism: A Distinctive Pattern

Fertility Trends 1945–89

Like most countries in Western Europe, the countries of Eastern Europe and the Soviet Union began their fertility transitions in the early twentieth century and recorded sustained fertility declines for several decades prior to World War II. In Russia, for example, it is estimated that the total fertility rate fell from 6.7 children per woman in 1927 to 4.3 children per woman in 1940, only 13 years later (Andreev et al., 1998). (The total fertility rate is calculated based on current age-specific birth rates (ASBRs) and represents the average number of children a woman can be expected to have during her lifetime.) In the early postwar period, however, fertility trends began to diverge markedly between East and West: Western Europe experienced a mild postwar baby boom, while fertility continued to decline across the socialist countries. By the mid-1960s, the total fertility rate had fallen below the replacement level of fertility in many socialist countries, including Hungary, the Czech Republic, Romania, and the Soviet republics of Russia, Ukraine, Estonia, and Latvia. Some analysts attribute this decline in fertility to the wave of abortion liberalization that occurred across the region in this period, beginning with the re-legalization of abortion in the Soviet Union in 1955 (David, 1982). Possibly in response to pronatalist measures undertaken in many countries in the 1970s, the total fertility rate stabilized in Eastern Europe and the western Soviet republics at around 2.0–2.5 and remained at this level until the mid-1980s (see **Figure 1**) (The sharp increase in the total fertility rate in Romania in 1967 was due to the imposition of Decree 770 in late 1966, which replaced Romania's liberal abortion policy with severe restrictions on the availability of abortion. See David (1999) for further details on abortion policy in Eastern Europe and the Soviet Union and its successor states.). In contrast, total fertility rates declined in Western Europe during this period, and as a result the socialist countries emerged as the highest fertility region in Europe by 1980.

The Socialist Fertility Pattern

The transformation of patterns of fertility and family formation that began in Western Europe in the 1970s and 1980s virtually bypassed Eastern Europe and the Soviet Union in these years. The transformation in Western Europe was marked by rising age at first marriage and at first birth, rising out of wedlock birth rates, and total fertility rates below the replacement level in many countries. In the socialist countries, in contrast, age at first birth remained stable or trended downward in the 1970s and 1980s (**Figure 2**). As an example of this divergence of fertility patterns, the mean age at first birth fell from 24.1 to 22.8 in Estonia between 1970 and 1989, while it increased in France from 23.9 to 25.9 in the same years.

Marriage and first birth both took place at relatively young ages in Eastern Europe and the Soviet Union, with women

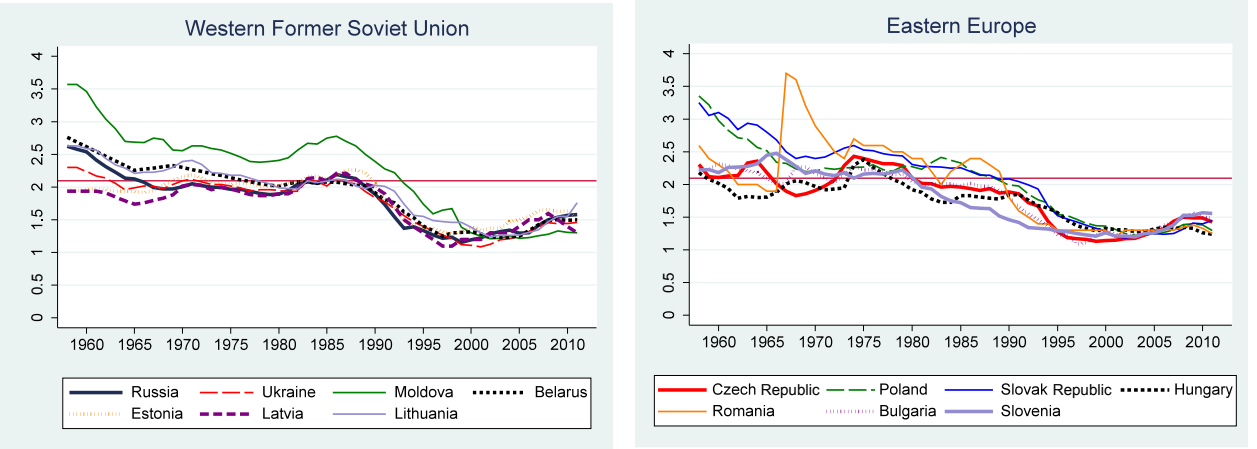

Figure 1 Total fertility rate, 1960–2011.

entering marriage at an average age of 21–22 years, soon followed by first birth at an average age of 22–23 years in most countries. Any subsequent births were closely spaced, so that most women completed their childbearing at relatively young ages. By far the highest rates of childbearing occurred among women aged 20–24. In Western Europe, marriage and fertility postponement increasingly became the norm, leading to rising age at first birth and the highest fertility rates occurring among women aged 25–29. Marriage and childbearing were near-universal in the socialist countries; the rate of childlessness was near the biological limit of 5–10%, well below the rates of childlessness in West European countries (Sobotka, 2011). As is well known, abortion rates were high in the socialist countries, with abortion being used as the primary means of birth control due to limited alternative contraceptive options. The highest abortion rates occurred in the western republics of the Soviet Union (Russia, Ukraine, Belarus) and in Romania prior to 1967; for instance in 1980 the number of abortions per 100 births in Russia was 205. In the United States in the same year, the ratio was 43 abortions per 100 births.

Influence of the Socialist System on Fertility Patterns

This 'East European reproductive pattern' – distinguished by young age at first marriage and first birth, stable fertility rates near the replacement level of fertility, and high abortion rates – likely reflects the incentives created by the institutional features of the socialist system as implemented in Eastern Europe and the Soviet Union, as well as the pronatalist policies pursued by many socialist governments in the 1970s and 1980s. The features of the socialist labor market that likely influenced fertility choices include the high job security and low unemployment rates, as well as the compressed wage differentials and relatively low rates of return to higher education that characterized the wage structure in most countries. In these circumstances, women had little economic incentive to delay

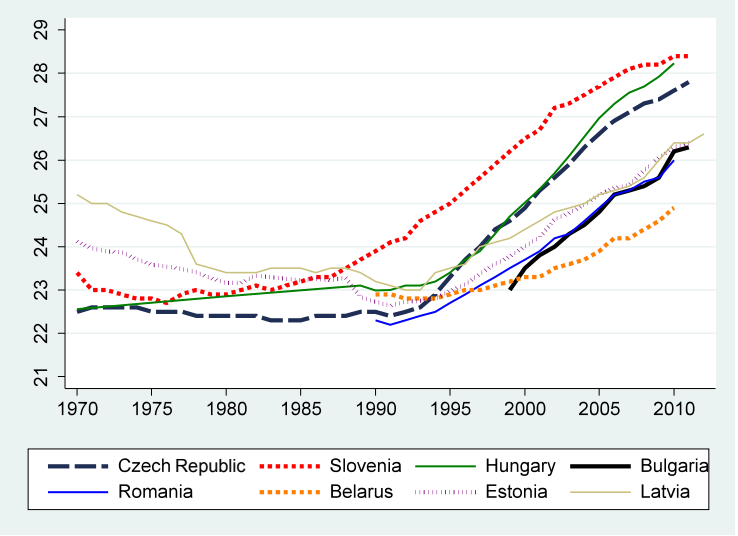

Figure 2 Age at first birth, selected countries, 1970–2011.

marriage and childbearing in order to invest in education: due to the relatively low payoff to additional years of education and the limited opportunities for career advancement (particularly for women), the opportunity cost of early childbearing was relatively low. Housing shortages in many countries added to the incentive to marry and have a child at a relatively young age, as priority in the allocation of apartments was given to married couples with children (Sobotka, 2011). Without a means to advance in the housing allocation, children often continued to live with their parents well into adulthood, often for years after marriage.

In addition to the priority in housing allocation given to families with children, many countries of Eastern Europe and the Soviet Union adopted explicitly pronatalist policies in order to promote higher birth rates. These policies included generous maternal leave policies, birth payments, child benefits, and the widespread provision of free- or low-cost child care; most of these policies were targeted toward increasing second- and third-order births. The effectiveness of these policies in increasing completed fertility has been widely debated by demographers and policymakers, and most analysts argue that the pronatalist policies of the socialist countries only temporarily increased birth rates while having a modest (if any) impact on completed fertility (David, 1982; Sobotka, 2011; Frejka and Zakharov, 2012).

Fertility in Postsocialist Countries: Convergence of East and West

The socialist system collapsed in Eastern Europe beginning in 1989, soon followed by the dissolution of the Soviet Union into its 15 successor states in 1991. Most countries subsequently embarked on an arduous process of transition to market economies, a process unevenly implemented across the region and resulting in mixed outcomes, from the rapid and more successful reformers (Poland, Czech Republic, Hungary) to the more gradual and less successful reformers (Belarus, Ukraine). Despite the diversity of reform experiences, however, the sweeping demographic changes that have occurred since 1990 have been remarkably consistent across almost all formerly socialist countries, resulting in a new pattern of fertility and family formation that is highly similar to that of Western Europe.

Declining Total Fertility Rates

In the immediate aftermath of the fall of communism in the early 1990s, fertility rates in the formerly socialist countries of Europe plummeted to often unprecedented levels. The lowest levels were recorded in Bulgaria, with a total fertility rate of 1.09 in 1997, and in Ukraine, with a total fertility rate of 1.085 in 2001. These appear to be the lowest fertility rates ever recorded in a European country during peacetime. The fertility decline occurred in all countries of the region and was similar in magnitude, with the average total fertility rate falling from about 2.0–1.4 in all countries despite the wide variety of transition strategies and varying degrees of success in restoring the macroeconomies to positive economic growth. For example, the decline of the total fertility rate in Belarus, one of the most gradual reformers, is nearly identical to the decline of the total fertility rate in the Czech Republic, one of the most rapid and successful reformers (see Figure 1). The fertility decline reached its nadir in 1999–2000, then reversed slightly after 2000, but fell back again in 2009–11 following the 2008 financial crisis. In all countries the total fertility rate remains well below the replacement level of fertility. The number of births in the region is now so low that in many countries the rate of natural increase of the population (births minus deaths) is negative, implying a declining population size in the absence of substantial in-migration. Negative natural increase occurred in 2011 in Belarus, Bosnia and Herzegovina, Bulgaria, Croatia, Estonia, Hungary, Latvia, Lithuania, Romania, Russia, Serbia, and Ukraine. Only in Russia immigration was large enough to offset the negative natural increase so that positive population growth was recorded in that year; all of these other countries registered negative population growth in that year.

Postponement or Permanent Decline?

A key question is whether this fertility decline represents a permanent transition to below-replacement fertility levels or a temporary decline driven by women deciding to delay rather than forego childbearing. As demographers have noted, it is difficult to separate the 'tempo' effect of a change in fertility timing from the 'quantum' effect of a change in completed fertility, and this is especially true during periods of rapidly decreasing fertility rates: when fertility rates are declining quickly, fertility indicators such as the total fertility rate can be highly inaccurate indicators of completed cohort fertility. In particular, when the decline in fertility is due to postponement of childbearing, the total fertility rate is biased downward (Billari, 2008).

There are indications that at least some of the decrease in fertility represents a postponement effect rather than a 'quantum' effect; the dramatic increase in age at first birth across countries, as illustrated in Figure 2, is perhaps the strongest evidence of this. For example, the average age at first birth in Hungary increased from 22.99 in 1990 to 28.23 in 2010, an increase of 5.24 years in just 20 years. In the Czech Republic, the average age at first birth increased by 5.1 years over the same period, from 22.5 to 27.6 years. This increase in age at first childbearing is astonishingly rapid and occurred in roughly half the time it took for a similar increase in West European countries. In Norway, for example, the average age at first childbearing increased steadily from 1970 to 2010, from 23.2 to 28.2 – an increase of 5.0 years over a 40-year period. As of 2011, the average age at first birth in many East European countries had nearly converged with West European levels of about 28–29 years.

The postponement of childbearing is also illustrated by the changes in ASBRs in these countries. As previously noted, across all of the formerly socialist countries of Europe, by far the highest birth rates occurred among women aged 20–24; this pattern was remarkably stable for several decades prior to 1989. Due to the postponement of fertility, however, in nearly all post-socialist countries women in the 25–29 age group now have the highest birth rates. The trends in ASBRs are illustrated in Figure 3 for two representative countries, Russia and Hungary. Russia's ASBR pattern is characteristic of many

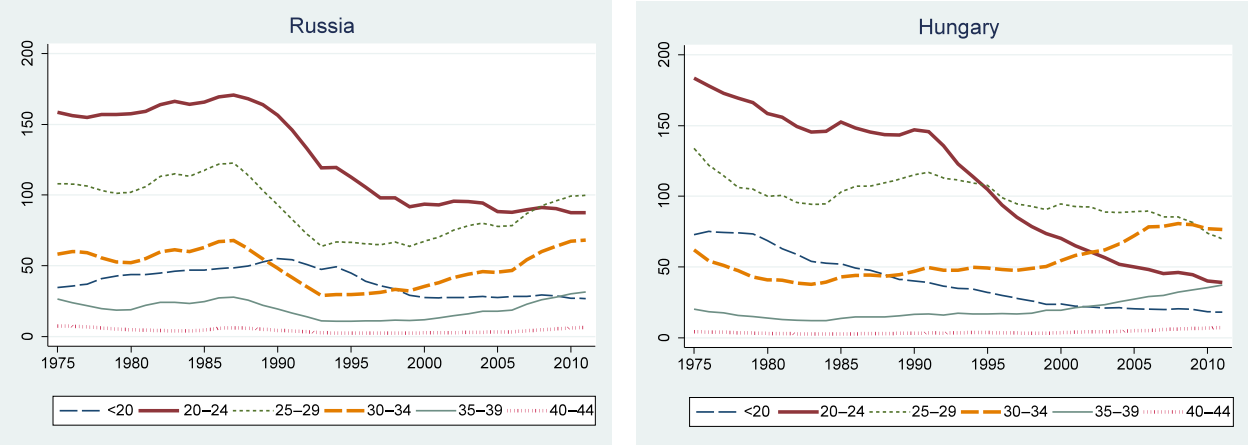

Figure 3 Age-specific birth rates, 1975–2011.

countries of the former Soviet Union, such as Ukraine and Belarus, while Hungary's pattern is typical of many Central European countries. Both figures illustrate that, until the 1990s, women aged 20–24 had the highest birth rates in both countries, followed by women aged 25–29. The decline in fertility rates in the 1990s was especially steep for women aged 20–24 in both countries as well as across the region overall. In Hungary birth rates have increased among all women aged 30 to 44, and as of 2010 the birth rate of 30- to 34-year-olds actually exceeded that of 25- to 29-year-olds. This 'crossing' of the 30–34 ASBR over that of the 25–29 ASBR also occurred in the Czech Republic in 2011. This 'crossing' has also occurred in many West European countries, beginning with Spain (1995) and Italy (1997), followed by Switzerland (2001), Sweden (2002), United Kingdom (2004), Greece and Portugal (2005), Germany and Finland (2006), and Norway (2010). In all of these countries, women in the 30- to 34-year-old age group continue to have the highest birth rates of all women age 15–49. Later childbearing is also occurring in Russia, with women aged 25–29 now having the highest birth rates as well as higher birth rates among women age 30–44. These changes in ASBRs and age at first birth represent fundamental shifts in the long-established patterns of the age structure of fertility, indicating that the region is becoming much like Western Europe in which the majority of childbearing is postponed until the early thirties.

Despite the evidence of fertility postponement, it is likely that the fertility decline in formerly socialist countries also reflects a reduction in completed cohort fertility. The extent of this change will not be known with certainty until the cohorts of women of childbearing age in the 1990s have reached the end of their childbearing years, but some early evidence is suggestive of a real decline in completed fertility. Results of the 2010 population census of Russia, for example, indicate that women born in 1946–55, whose reproductive years were largely completed prior to the transition, had an average of 1.85 live births in their lifetime. Women born in 1966–70, whose reproductive years coincided with the most turbulent period of transition in the 1990s, had an average of 1.64 live births, and women born in 1971–75 had 1.51 live births (fertility is not complete for this latter group) (Goskomstat of Russia, 2012).

Causes of the Fertility Decline?

Assuming that the fertility decline is at least in part real, rather than due to postponement, a number of possible explanations for the decline of fertility across the formerly socialist countries of Europe have been offered by observers. These include the loss of income associated with the collapse of GDP in the 1990s, the increased economic uncertainty caused by the transition from socialism to capitalism, and the decrease in the number of state-supported nurseries and preschool facilities as well as the near-disappearance of day care facilities provided by enterprises. In addition, increased returns to education have strengthened the incentive for young adults to obtain higher education, leading to the postponement of marriage and first births for many women. Analysts also argue that the new economic and political freedoms that accompanied the collapse of the socialist system created new opportunities for individual fulfillment, opportunities previously unavailable under socialism, which in turn raises the opportunity cost of childbearing and may further reduce fertility (Sobotka, 2011). However, credible evidence identifying the causes of the fertility decline in the transition economies is limited, so much of this discussion is speculative.

Other Fertility Trends

Other demographic changes across the region occurred with similar swiftness and further underscore the convergence of the Eastern pattern of fertility and family formation with that of West European countries. In particular, nonmarital birth rates have increased significantly in most of the formerly socialist countries, and have even surpassed the nonmarital birth rates of some West European countries. As shown in **Figure 4**, the share of births occurring outside of marriage increased in the region beginning in the late 1980s and early 1990s and has accelerated since those years. The highest nonmarital birth proportion occurs in Estonia, with 60% of all births occurring outside of registered marriage, followed closely by both Slovenia and Bulgaria. West European countries with high nonmarital birth shares include France (55.8% of births in 2011), Norway (55.0%), and Sweden (54.3%). Marriage rates have also fallen across the region while divorce rates have increased (see

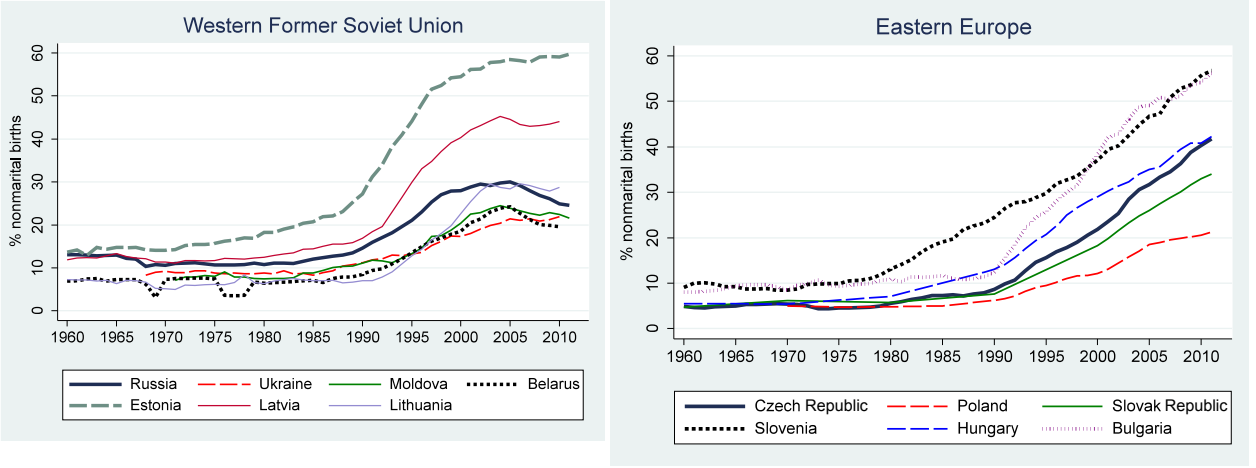

Figure 4 Nonmarital births, 1960–2011.

Heleniak, 2010; Brainerd, 2012, for more detailed discussions of demographic trends in formerly socialist countries).

Figure 5 illustrates that the decline in birth rates was not achieved through increasing abortion rates. On the contrary, abortion rates have plummeted in most countries in the decades since 1989. For instance, in 1989 the number of abortions per 100 births in Belarus was 167, but by 2010 this ratio had fallen to just under 31. This level is similar in magnitude to that of the United Kingdom (29 abortions per 100 births in 2009) and the United States (23 abortions per 100 births in 2009). While abortion rates remain high in a number of postsocialist countries (Russia, Estonia, Bulgaria, Hungary, Romania) compared with other Western countries, the decrease in abortion rates is substantial in all countries and appears to be due at least in part to the increased availability of modern contraceptive methods in the 1990s and a resulting substitution of contraception for abortion. (The decline of abortion in Poland is due to stringent restrictions imposed on abortion in that country in the early 1990s.)

Concluding Remarks

The demographic transformation of the formerly socialist countries of Europe has been remarkable in its speed and scope. Many East European countries now resemble West European countries in their patterns of later marriage and childbearing, rising nonmarital births, and below-replacement fertility rates. These changes in fertility and family formation will likely have far-reaching consequences for each country's population size and age structure, as well as for the well-being of children in the region. The changes in fertility patterns in postsocialist countries also demonstrate the remarkable degree to which major individual life choices – when to marry, when to have a child, and how many children to have – are influenced by the broader social and economic circumstances in which we live, and how rapidly these choices can shift under circumstances of profound societal change. As has long been the case even before the fall of communism, the unique demography of Eastern Europe

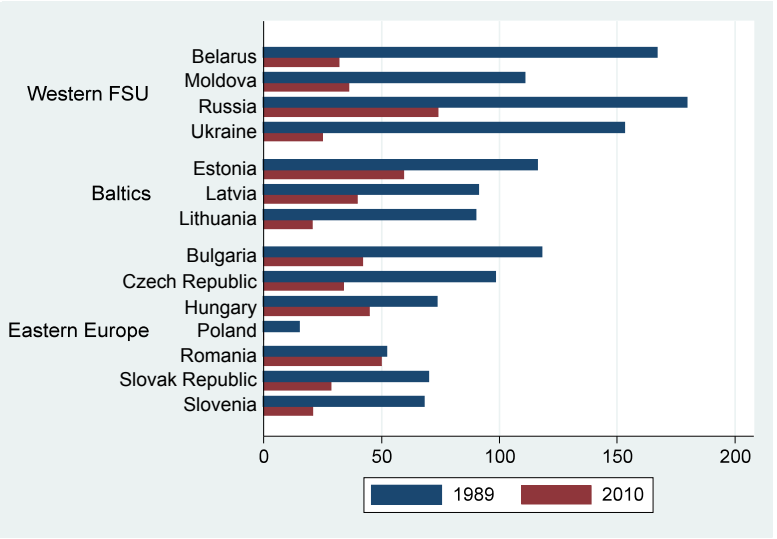

Figure 5 Abortions per 100 births, 1989 and 2010.

and the former Soviet Union presents a myriad of questions for researchers to investigate.

See also: Communism, History of; Economic Transformation: From Central Planning to Market Economy; Fertility Change: Quantum and Tempo; Fertility Theory; Fertility Transition: China; Fertility Transition: Cultural Explanations; Fertility Transition: East Asia; Fertility Transition: Economic Explanations; Fertility Transition: Latin America and the Caribbean; Fertility Transition: Middle East and North Africa; Fertility Transition: South Asia; Fertility Transition: Southeast Asia; Fertility Transition: Sub-Saharan Africa; Fertility and Culture: Anthropological Insights; Population Change in the Former Communist States of Europe and Asia; Post-Socialism: Political Economy; Revolutions of 1989–90 in Eastern Central Europe; Russian and Post-Soviet Studies: Gender; Second Demographic Transition; Socialism: Historical Aspects.

Bibliography

Andreev, E.M., Darskii, L.E., Khar'kova, T.L., 1998. Demograficheskaya istoriia Rossii 1927–1959 (Demographic History of Russia 1927–1959). Informatikia, Moscow.

Billari, Francesco C., March 2008. Lowest-low fertility in Europe: exploring the causes and finding some surprises. The Japanese Journal of Population 6 (1), 2–18.

Brainerd, Elizabeth, 2012. The demographic transformation of post-socialist countries: causes, consequences, and questions. In: Roland, Gérard (Ed.), Economies in Transition: The Long-Run View. Palgrave Macmillan, pp. 57–83. Also circulated as United Nations University-World Institute for Development Economics Research (UNU-WIDER) Working Paper No. 2010/15.

David, Henry P., February 1982. Eastern Europe: pronatalist policies and private behavior. Population Bulletin 36 (6), 1–48.

David, Henry P. (Ed.), 1999. From Abortion to Contraception: A Resource to Public Policies and Reproductive Behavior in Central and Eastern Europe from 1917 to the Present. Greenwood Press, Westport, CT, and London, UK.

Frejka, T., Zakharov, S., September 2012. Comprehensive Analyses of Fertility Trends in the Russian Federation during the Past Half Century. Max Planck Institute for Demographic Research MPIDR Working Paper WP 2012-027.

Goskomstat of Russia, 2012. Demograficheskii Ezhegodnik Rossii (Demographic Yearbook of Russia). Moscow.

Heleniak, Timothy, 2010. Causes and demographic consequences of fertility decline in the former Soviet Union and Eastern Europe. Marriage & Family Review 46 (1/2), 79–106.

Sobotka, Tomáš, 2011. Fertility in Central and Eastern Europe after 1989: collapse and gradual recovery. Historical Social Research 36 (2), 246–296.

Fetishism

Roger Sansi Roca, Universitat de Barcelona, Barcelona, Spain; and Goldsmiths, University of London, Greater London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concepts of 'fetish' and 'fetishism' (the adoration of objects and natural events) are some of the most controversial and contradictory in the history of anthropology and Western thought in general. In this article, we will present a brief history of these terms, from their origins in the encounters between West Africans and European traders in the seventeenth and eighteenth centuries, to current anthropological discussions on the nature–culture divide and the agency of things.

'Fetishism' could be defined, in very wide terms, as the adoration of objects and natural events. The 'fetish' would be this very object of adoration: not just the representation or image of a divinity, but also a material object or natural event that is worshiped as a divinity itself. These terms originally make reference to religious worship, but they have also been applied to other domains in which things acquire the qualities of persons – like in Marxist commodity fetishism or Freudian sexual fetishism.

Fetish and fetishism are two of the most controversial and contradictory terms in the social sciences. They have been explicitly rejected as anthropological concepts since the late nineteenth century; for Marcel Mauss, they revealed an 'enormous misunderstanding' (Iacono, 1992: 116). And yet, in recent decades, there has been a growing interest in this very misunderstanding, what William Pietz called "the problem of the fetish" (Pietz, 1985, 1987, 1988). The 'problem' of the fetish touches upon some of the basic distinctions of Western thought: people and things, action and event, subject and object, culture and nature, religion and economy. The questioning of these very distinctions in recent anthropology has inevitably brought the problem of the fetish back to the center of attention.

Origins of the Term

To address the 'problem of the fetish,' we should start by the original 'misunderstanding' that generated it. As Pietz says (1985: 5), the 'fetish' is a cross-cultural and radically historical event: it was born in the writings of European travelers to West Africa in the seventeenth century, to identify a supposedly African phenomenon. The *Encyclopédie Française* of Diderot and D'Alembert defined *fétiche* in the following terms: "Name that the peoples of Guiney in Africa give to their divinities. They have a fetish for each province, and a particular fetish for each family. This can be a tree, an ape's head, a bird, or something like that, following their imagination."("Fétiche, f, F. (Hist. mod) Nom que les peuples de Guinée en Afrique donnent à leurs divinités. Ils on une fétiche pour toute une province, & des fétiches particuliers pour chaque famille. Cette ici o est un arbre, une tête de singe, un oiseau, une quelque chose de semblable, suivant leur fantaisie" (D'Alembert and Diderot, 1751–1765: 598).) Their main source was the seventeenth-century Dutch geographer Dapper's *Description of Africa* (1686).

Interestingly enough, in this definition there is no reference to the fact that the term 'fetish' does not come from any African language, but from the Portuguese term, *feitiço*. This is the first misunderstanding: this is not an African term, but a European one. And yet, there is a significant shift between the *feitiço* and the *fetish*. The Portuguese *feitiço* is a magical event – a spell or charm, an act of sorcery; *feitiço* then is not exactly a divinity, but a magical action. The term itself has a long and interesting etymology: it comes from the Latin *factitius*, which means made up or artificial; from there comes the Medieval Latin *maleficium*, something said or done for or with magical effect. In fact, in Portuguese *coisa feita* ('something done') is synonymous with *feitiço* (Sansi, 2007). The *feitiço* was persecuted not only because it was evil, but also because it was a trick, a falsification. Originally, *feitiçaria* was more an object of Christian law than of Christian theology (Pietz, 1987: 31). In 1385 and 1403, King João I of Portugal made laws against it, in which he forbade his subjects to use or effect charms (*obre de feitiços*), bonds (*ligamentos*), or summoning up of devils (*chamar os diabos*), since *feitiçaria*, like idolatry, is inspired by the Devil. *Feitiçaria* is always an accusation; no one would admit to being a *feiticeiro*. But this official discourse on sorcery could be bypassed by the practitioner of magic affirming the contrary, that one was practicing magic against the Devil, not for it, and against the evil spirits that haunted living people. They could do this since it was not their techniques and instruments that constituted the *feitiçaria* but the inspiration from the Devil. This introduced a considerable ambiguity in its persecution.

The discourse of the *feitiço* arrived to the coasts of West Africa with the Portuguese in the fifteenth century. The Portuguese addressed the pagan practices of the Africans as 'idolatries and sorceries' (*Idolatrias e feitiçarias*) (Sansi, 2007). Idolatry would be a form of organized religion, but with false images of gods or 'idols' (Bernand and Gruzinski, 1988); *feitiçaria*, on the other hand, would not be organized religion, but individual practices of sorcery.

Where does 'fetishism' fit into this picture? The discourse of the 'fetish' and 'fetishism' emerges in the seventeenth and eighteenth centuries not in the writings of the Portuguese, but in those of other European travelers, essentially Dutch and Protestant ones like De Marees (1602), Dapper (1686), and

Bosman (1705). For these travelers, the *fetissos* were the gods of the Africans, and the *fetisseiros* were their priests. The reference to the Portuguese term is in most cases forgotten or, better, misrecognized. Misrecognition and disavowal are in fact central to the 'problem' of the fetish. As opposed to a discourse of idolatry, these travelers understood that Africans worshiped the objects themselves as gods, not as images (idols) of the gods. This was the radical novelty of the discourse of fetishism in relationship to idolatry: its 'untranscended materiality,' in Pietz's terms; the object does not represent a god, but it is a god itself.

Bosman made one of the more accurate accounts of the 'fetish.' According to this Dutch tradesman, Africans worshiped objects, like stones, trees, and even animals and artifacts. What was more surprising was that Africans recognized that they were 'found' by chance, that they made them with their own hands, and that they had as many as they wished (they were unaccountable). In other terms, they recognized that their gods were artifices made by them. The priests of these *fetissos* were the *fetisseiros*, inevitably defined as greedy, powermongering, and treacherous, manipulating the fear of the people for their own benefit. Bosman not only was scandalized by what he identified as the hypocrisy of the priests but also, in fact, thought that most people were quite aware of the fatuousness of their beliefs, but still they believed in the *fetissos*.

In Bosman's argument, we can identify some of the core issues of the problem of the fetish. First is the 'confusion' of creation and the creator: the Africans, for Bosman, worship things of their own making, instead of worshiping the creator who made them. Second, African religion is described as a form of 'trade' and exploitation by a priestly elite who manipulates the masses. Still, Bosman is very adamant to make clear that these masses are fully aware of their exploitation, but they accept it – again, the theme of misrecognition. Ultimately, Bosman's reading is also full of misrecognition; behind the construction of an 'African' other, it seems clear that the model he is attacking is much closer in space and time: the 'Papism' of the Portuguese who preceded Bosman and the Dutch in the African trade, and who still were their main enemies and competitors in the Atlantic world.

In fact, the problem of the fetish was not just one of religion but also one of commerce. For Bosman, the difficulty of trading with Africans was that they valued things irrationally, according to 'caprice.' The issue, then, was how to establish the terms of exchange between two radically different systems of value. Protestant European traders were particularly annoyed when they were obliged to make 'fetish-oaths' on the objects that Africans worshipped. The objects of trade themselves were also subject to this ambiguity and artificiality: European traders were particularly wary of 'fetish-gold' (Atkins, 1735: 183), or gold mixed with other metals.

Fetishism as a Countermodel of the Enlightenment

In the end, this is more a story of misrecognitions than misunderstandings. A theory of 'fetishism' emerged, built on these misrecognitions, but it was going to have crucial importance in the history of social thought. In *Du Culte des Dieux Fétiches*, Charles De Brosses (1760) finally developed a full-fledged theory of the *fetish* and invented the term *fetishism*. For De Brosses, the African fetish is not just a curiosity but also a general model at the origin of all religions – starting with Egypt. For De Brosses, the origins of religion are fear and need: 'savages' would worship unexpected events and beings that they couldn't understand and control, addressing them as persons with a will and power superior to those of humans – as gods. This is not a 'naturalism,' or a religion of nature, since fetishism doesn't have a notion of nature as opposed to the social or the supernatural. Fetishism confuses people with things, events with actions, the given with the made, and immanence with transcendence.

With his theory of fetishism, De Brosses organized the most powerful argument against religion in a century of skepticism. By identifying the worship of things as the first step of religion, it was finally possible to integrate the criticism of religion to a general theory of progress, which would see the Enlightenment as the last, logical end of human achievement: in modern Europe, men had finally managed to relate to things just in the terms of science, as objects of natural law, or in the terms of the economy, as commodities with a market value, separate from any religious prejudices and fears. Religion, science, and economy would be finally separated, in what Bruno Latour (1993) has called a process of 'purification' separating their religious beliefs from material things altogether.

By the same token, the African fetishists would be seen as the more primitive and abject of men. For Kant, the cult of fetishes "is so disgraceful that it seems to contradict human nature. A feather, a horn, a shellfish or any common thing, just being consecrated with some words, becomes an object of veneration invoked in oaths. The Blacks are extremely vain, and so loud that they have to be dispersed with a stick" (Kant, 1960: 111). Hegel thought that Africans 'adored the first thing that comes their way' arbitrarily; this was the reason why, for Hegel, fetishism was not even a form of religion, since the fetish was not even an object of religious worship: fetishist Africans lived in a moment prior to history, since they lacked "the principle which naturally accompanies all our ideas – the category of Universality" (Hegel, 1956: 93). Africans were slaves of chaotic contingency, or, better, of the false priests who mediated this chaos.

The 'purification' of religion from science, and of people from things, was also a geopolitical process of the separation of 'modern' Europe from fetishist Africa. As a matter of fact, the development of a theory of the fetish was essential for the very construction of this very model of European modernity, since it offered a perfect model of what an 'impure' humanity would look like. In these terms, the theory of *fetishism* in the social sciences as an elementary, and mistaken, form of materialism persisted during most of the nineteenth century. The main example is Auguste Comte (2001), who still defined as 'fetishism' the first stage of religious worship, before polytheism and monotheism. But by the end of the nineteenth century, with the emergence of social anthropology, the use of the term started to be questioned and replaced by other theories. Tylor (1874) subsumed 'fetishism' under his own theory of 'animism,' stating that the object of worship of the fetish was a spirit contained in the object, not the object itself. In other terms, he basically questioned the very notion of the

fetish as a radically material entity. More than based on specific empirical evidence, we should understand Tylor's questioning of the notion as an attempt to redeem the 'primitive' from their banishment outside history and human nature by Kant and Hegel. For the Western tradition, the very distinction between body and spirit, and between object and subject, is at the basis of human reason, and Tylor, as the founding father of anthropology, was keen to demonstrate the psychic unity of humankind, and the basic evolutionary capacities of all humans on these grounds. Fetishism, in these terms, is not just irrational but also outside of reason itself. To recognize the irredeemable materiality of the fetish would question the whole enterprise of a unified description of humanity based on shared reason.

Durkheim, the other founding father of both anthropology and sociology, took a very similar stance. From a Durkheimian perspective (1912), the problem of elementary religion starts with the very issue of representation – the capacity of understanding something beyond immediate sensory perception and individual existence; and this capacity is common to humanity as a whole. But the answer to this question for Durkheim is not animism but totemism: the basic issue is not the distinction between body and spirit, but the representation of a collective entity beyond immediate experience.

Fetishism as Cultural Critique

Mauss's rejection of the fetish as an enormous misunderstanding follows Tylor's and Durkheim's humanistic approaches – denying the very possibility of fetishism, and the accusations of irrationality and even inhumanity addressed to non-Western peoples. Paradoxically, by the time that anthropology was rejecting fetishism, the very notion had also been used critically against the West by some of its major contemporary thinkers, Marx and Freud. Marx's brief but powerful parable of 'commodity fetishism' reverts to the basic parallelisms between religion and commerce that were at the center of the Dutch traders' questioning of the African fetish. For Marx, instead, it was the very Western form of commodity value that could be understood as fetishism (Marx, 1992). Marx noticed that same mechanism the European travelers attributed to African fetishists: people in capitalism make commodities with their own hands, but then misrecognized that the value of these commodities was the result of their own making; in contrast, it is thought that commodities have value in themselves. Freud's use of the notion of the fetish (Freud, 1927) would also impinge on the question of misrecognition, or, in more psychoanalytical terms, disavowal. For Freud, the sexual fetish is a thing that becomes a surrogate object of desire, in replacement of the actual object of desire, which is repressed. This object emerges out of the event of an encounter, which the fetishist tries to recreate in his or her fantasies.

Marx's and Freud's approaches involve a clear emphasis on misrecognition and the repression or hiding of the truth of interpersonal relations behind objects, which are falsely endowed with a power that they don't have. The combined influence of these authors has been crucial in the twentieth century, and the critical theory of the fetish has been a tool for many modern authors. In the social sciences, Pierre Bourdieu is probably the most well-known proponent of this critical theory of fetishism, as applied to the field of art, for example (Bourdieu, 1992). Bourdieu describes the field of art as sustained in a collectively held *illusio* (an embodied belief) of the autonomous value of art. This *illusio* involves a misrecognition not only of the interdependency between art and other forms of social value, but also, essentially, of the artificiality of this value itself. Like the African fetishists of Bosman, Bourdieu's artists are perfectly aware that they have made their fetishes with their own hands, and yet they adore them as if they had an autonomous power.

Michael Taussig's *The Devil and Commodity Fetishism in South America* (1980) is probably one of the more original contributions to the literature on the fetish. Taussig described how the people of a Colombian valley read the arrival of modern capitalism and the money economy through traditional discourses on magic. In Taussig's terms, this was not a misunderstanding of capitalist production, but rather the opposite: an explicit description of the mechanisms of commodity fetishism that were misrecognized in Western ideology. Taussig's argument was a precedent to a growing ethnographic interest in the 1990s in what came to be called 'the magic of modernity' (Comaroff and Comaroff, 1993; Geischere, 1997; Meyer and Pels, 2003): many anthropologists, particularly in Sub-Saharan Africa, were witnessing an upsurge in magic accusations and beliefs. Rather than interpreting this phenomenon as a 'return to tradition,' many of these authors perceived it as a local adaptation to the global commodity fetishism of capitalism at the turn of the twentieth century. Only recently, though, have some authors started to think about this phenomenon in a longer historical and conceptual framework, showing how this 'encounter' with globalization is part of a very long process of exchanges, starting with the very origins of the term 'fetish' (Sansi, 2011).

The Untranscended Materiality of the Fetish

Independent scholar William Pietz made a major contribution to the literature of the fetish in a series of articles published in the late 1980s with the title "The Problem of the Fetish" (Pietz, 1985, 1987, 1988). Pietz offered an encompassing historical account of the term, emphasizing that it was the result of a very specific space and time: the coasts of West Africa in the seventeenth and eighteenth centuries. According to this author, 'the problem' could be defined around four fundamental questions – *personification*, *territoriality*, *historicity*, and *reification* – which we should address with some detail.

First is personification. Fetishes are intensely personal objects (Pietz, 1985: 11): they add to the person, as for example in the case of amulets worn on the body to protect it from attacks. By the same token, parts of the person, when stolen or found by enemies, can also be used in sorcery. This is a form of distributed personhood, as Gell has noted (1998), although it is based on negative reciprocity: a distributed person has to avoid leaving dispersed pieces of his or her self, to avoid the risk of sorcery, and has to add to the body elements that can act as shields. But of course, to make these

shields work one has to give a part of one's self to them, as an offering in a mediated exchange between person and thing: the fetishes have to become a part of the person, not just instruments or technological devices. That is why other people cannot use a person's own fetish: because they are a part of the self.

The life and the value of the fetish, however, cannot be solely understood as an extension of the personhood of humans: it is also a result of its territoriality and historicity. On the one hand, the 'untranscended' materiality of the fetish, which is not just a symbol or icon of a divinity but also a 'self-contained entity' with an active force, introduces the question of the position of the object in space and time as a 'territorialized' (Pietz, 1985: 12) object. The 'life' of the fetish is conditioned by material constraints, in space and in time: its inability to move physically, which makes it strictly dependent on its human associates, and its inscription in a concrete and specific place where it is protected, like an altar or a body.

But these are not the only ways in which materiality influences processes of value: in contrast, Pietz notes that "the fetish is always a meaningful fixation of a singular event; it is above all a 'historical' object, the enduring material form and force of an unrepeatable event" Pietz (1985: 12). The radical historicity of the fetish, which emerges from a unique event, is what most disturbed Hegel, and what convinced him to put Africa out of history. What was interpreted as 'caprice' or 'arbitrary' choice is, in fact, the recognition of the singular values generated by events, which cannot be explained by the list of elements that make a part of the situation before they happen (Latour, 2001: 131). Finding a fetish is an event that cannot be predicted: it is a single happening in which people find something unforeseen that they recognize as a part of themselves, something that becomes personified. Pietz explains this point wonderfully by making reference to Leiris and the surrealist notion of the *objet trouvé*: "these crisis moments of singular encounter and indefinable transaction between the life of the self and that of the world become fixed, in both places and things, as personal memories that retain a peculiar power to move one profoundly" (Pietz, 1985: 12).

This historicity founded in events is radically different from that of Hegelian history, as teleology. The event in which the fetish is found is not perceived by the person as arbitrary, but necessary. The value that the person finds in the object is not randomly attributed by the person, but it is seen as an *immanent* value of the object, something inchoate that was always there, waiting for this particular person, something that he or she recognizes. It is as if the thing was offering itself to the person – as if they always belonged together. In this sense, this is a process of mediated exchange, between the person and a hidden value that is giving itself to the person. Thus, contextual values and historical changes are perceived as immanent values and mediated exchanges.

The last point that Pietz raises is 'reification.' The fetish, for Protestant European travelers and Enlightenment philosophers, raised the question of the relative value of objects: values were not universal, but dependent on a "particular order of social relations, that in turn, they reinforce" (Pietz, 1987: 24). This leads to a questioning of contextualizaton, or, as Pietz puts it, of 'reification': how values attributed to objects are in fact constructed by humans. Of course, 'reification' is always an accusation: it is the others who reify, who think, through a process of disavowal, that objects have intrinsic values. Thus, the identification of a system of value with fetishism is an accusation of a false ideological projection, a false consciousness. Stemming from the critical and skeptical definition of De Brosses, the accusation of fetishism became an extreme form of criticism of ideology.

The first three questions that Pietz identifies – territoriality, historicity, and personification – are relatively independent from the problem of reification; they are central to the formation of the notion of the fetish in early colonial West Africa, while the problem of reification was raised later, in the radical criticism of the fetish by Enlightenment philosophers, in the formation of a theory of fetishism. There is a fundamental difference between the two moments: the first is a moment of intercultural mediation, in which the notion of the fetish emerges in a sphere of exchange; the second is a moment of differentiation and alienation, when enlightened Europeans try to make clear what makes them different from and superior to Africans (and, by the same token, the Portuguese).

Pietz's work inspired a revival of literature on the fetish, developing some of the different lines of his argument: some were closer to a more traditional reading of fetishism as cultural critique (Apter and Pietz, 1993), and some more interested in the radical historicity of the fetish as an object of cultural exchange (Spyer, 1998). The possible yet contradictory ways of approaching the 'problem of the fetish' could be exemplified by two radically different readings of Pietz by two contemporary authors: David Graeber (2005) and Bruno Latour (2010).

Graeber: Fetishism as Social Creativity

The central argument in Graeber's reading of Pietz is a defense of fetishism as social creativity. If fetishism, at root, is our tendency to see our own actions and creations as having power over us, how can we treat it as an intellectual mistake? The bad reputation of fetishism is a result of a misunderstanding. Africans would 'make' fetishes as a means of creating social relations. They wouldn't hide the fact that the objects, like the social relations that depended on them, were 'made up' and thus were, in a way, arbitrary. The Europeans, however, would not recognize that the value they gave to objects was the result of a social relation: it would be the Europeans who would misrecognize (or disavow) the social construction of reality, as it were, giving it a bad name: fetishism. In Graeber's reading, Europeans and Africans shared a certain existential pessimism: both see society and its institutions as necessary evils that have to be institutionalized to prevent chaos. The Europeans and the Africans were not that different, but in fact were all too similar. For Graeber, "it was the underlying affinity, I suspect, which accounted for the common European reaction of shocked revulsion and dismay on being exposed to so many aspects of African ritual: a desperate denial or recognition" (Graeber, 2005: 415).

Graeber is underscoring one of the four problems identified by Pietz: the problem of reification – how social values are

a social construction. Graeber insists upon the fact that there is nothing abnormal or stupid in using objects as the objectification of social relations, and that this process of objectification is ultimately arbitrary. It is not the object, but the action made upon it, that gives it social power. For Graeber, objects are "only the medium. Hence what they are is ultimately somewhat arbitrary: one case valuable objects from faraway lands, or one can, in fact, use pretty much any random object one lays one's hands on, 'a Lion's Tail… a Bird's Feather… a Pebble, a Bit of Rag.' In this, Pietz's sources had a point, because this is exactly the moment where the arbitrariness of value comes fully into focus. Because really, creativity is not an aspect of the objects at all, it's a dimension of action. In this sense the new does in fact emerge from the old, and the numinous, alien nature of the object is really the degree to which it reflects on that aspect of our won actions that is, in a sense, alien to ourselves" (Graeber, 2005: 425).

The first question to ask here is if these 'Africans,' who Graeber, Pietz, and their sources invoke, would really agree with this idea of creativity. Would they say that, really, the power they recognize in objects is of their own making? It seems like Graeber falls back into the European *understanding* of the problem of the fetish. The power of objects is not just the result of a process of reification, of imposition of human value. The power of the fetish, or these objects that the Europeans identified as 'fetishes,' is not *arbitrary* but *necessary*. The power of the fetish is not arbitrarily attributed by the person who finds it, but is recognized by this person as inherent to the object. This power, however, is only revealed in the encounter of the object and that person. In that sense, we could say, it emerges in and by this encounter. It is not "any random object one lays one's hands on, 'a Lion's Tail… a Bird's Feather … a Pebble, a Bit of Rag,'" but it *must* be, and it *could only* be, that Lion's Tail, that Bird's feather, that pebble, or that bit of rag that is found at that specific moment and at that specific place, which unveils the uncanny power of the fetish over the subject. It is not irrelevant, in this sense, that this object has an 'alien nature.' It is precisely its alien nature – its radical alterity – that awakes the shock of recognition. That connection of subject and object, and of space and time, is not just an action by which the subject infuses the object with his or her creativity, but is an event of encounter that brings about something else, something that was not present before the encounter. This encounter, in other terms, has a historicity. Historicity, as we have seen, is one of the central questions that the problem of the fetish creates for Pietz, and, in many ways, it is irreducible to reification.

Latour and the Factish

Latour's *On the Modern Cult of the Factish Gods* (2010) was also written as a follow-up to Pietz's 'The Problem of the Fetish' articles. But we should also read it as a follow-up to his now classical pamphlet, *We Have Never Been Modern* (1993), where he takes on the question of the great divides of modernity – the 'purification' of society from nature, and of people from things. The emergence of the problem of the fetish is crucial in this process of purification, as a first instance when European traders and slave mongers explicitly accused Africans of confusing nature with culture, cause with consequence, people with things, and the product with the producer. Latour argues that it was the construction of this African counterexample that actually helped the Enlightenment objectify these distinctions, essentially the distinctions between nature and culture, between natural law and human convention, and between 'fact' and 'fetish.' According to the European slave mongers, Africans worshiped the things they made with their own hands. That wouldn't be such a surprise for a cynical, Western, freethinking slave monger who wouldn't necessarily believe in God or religion: all religions were made by men, since gods don't really exist in 'nature' – they are not facts. But still, all religions would hide the construction of the gods: priests would keep the masses in ignorance of this construction. Africans, on the other hand, as Graeber says, were explicit about this construction. In these terms, religions would be described basically as mechanisms of social power, using fetishes as puppets through which they could control the ignorant masses. This, in the long run, became a general theory of society as opposed to nature. Nature would be based on facts, given things that happen in the world, and would respond to natural laws independent of the intentions of humans. Society, on the other hand, would not be based on facts but on fetishes, made-up things that happened only within humans responding to human intentions – their will to power. Fetishes, as opposed to facts, would not exist autonomously from humans; they would be just intermediary entities used to veil the agency of humans. This is an extremely powerful theory, and it has survived for 300 years: as we have seen, Bourdieu's use of the term 'fetish' in his writings on art is quite consistent with these notions. This is what Latour (1993) calls the 'antifetishist' tradition, which sees social institutions as fictions hiding the 'facts': in general terms, the social domination of some humans over others. Latour wonders if it is possible to overcome this distinction between social fictions and real facts, going back to the roots of the problem of the fetish, to overcome theories that reduce processes of objectification to fetishism, alienation, disavowal, and misrecognition. This is an important point: saying that values are attributed to objects through processes of objectification, and that these processes are culturally relative, is not the same as saying that all attribution of value to objects is a false projection that hides a will to power. This is a point on which Pietz is not clear when he reduces objectification to reification.

After all, the so-called African fetishes inspired the foundations of a great deal of modern European social theory, anti-fetishism. Still, what was shocking for Europeans about Africans was not so much that they believed in fetishes, since this was the very essence of social life everywhere. What was shocking is that they didn't hide their fetishism! They acknowledged they made their fetishes with their own hands, and they recognized they were artifacts. But for Latour, as opposed to Graeber, this doesn't mean that Africans were explicit 'fetishists.' In contrast, it means that they didn't really make a distinction between facts and fetishes, 'real' and 'invented' things: they really thought that they were autonomous entities with their own agency, not

just because they had made them. Latour called this hybrid of fact and fetish the *factish*.

Latour proposed to reverse social theory from this point: what if 'Africans' were right? What if it doesn't make sense to distinguish between fact and fetish, things given and things made, natural reproduction and human production, and nature and culture? What if what is counterintuitive, strange, and irrational are all these distinctions, and not the confusion between these distinctions? Starting from here, Latour's program for a social science is not to reveal the hidden truth of the artifice of social life dissecting fact from fetish, looking for the dark secrets of human nature, but to follow the *factish*, as it were, to see how things are made into autonomous entities with their own agency.

Is Latour, then, rehabilitating the much criticized notion of the fetish? His position is quite ambiguous because of his development of a third, hybrid term, *factish*. The basic point that Latour was making is that we have to acknowledge the possibility of thinking about agency beyond the human, beyond classical Western distinctions between nature and society and between people and things, and beyond all the oppositions we have inherited from the tradition of Western thought since the Enlightenment. In any case, we could say that if Graeber stresses one of Pietz's arguments, the one on reification, Latour stresses the argument on historicity, the recognition of the emergence of new agencies generated by events, which cannot be reduced to the list of elements that make up situations before they happen (Latour, 2001: 131).

Conclusions: From Fetish to Fetishism and Back

To conclude, it is important to make clear that the question of fetishism is slightly different from the problem of the fetish. As we have pointed out in this article, the discourse of fetishism is basically built upon the idea of reification – the confusion of human values with the values of objects, and the subsuming of relations between people to relations between people and things. The discourse of fetishism as misrecognition, disavowal, ideology, and false consciousness has been central to social thought from De Brosses to Bourdieu and beyond. The problem of the fetish is slightly different, since its starting point is not a general ideology that attributes power to objects (fetishism), but it starts from a radical singularity, the 'historicity' and 'untrascended materiality' of the particular event of the encounter in which a specific agent is formed. The 'fetish' is not a token of a general type; it does not *represent* anything else, but it is always a radically new entity, grounded on a specific place and time, and is inredeemably material. One can think of 'fetishes without fetishism' (Pouillon, 1975): there are particular instances and events of the fetish that do not require being subsumed under a general ideology of fetishism, such as African religion, modern capitalism, or modern art. In general terms, the very notion of the fetish is also radically historical and territorial: it is the result of the encounter of West Africans and Europeans in the wake of colonialism. In the future, it may be interesting to explore a bit further these dimensions of the problem of the fetish, besides the discussion of fetishism and reification, which has been at the center of attention for so long.

See also: Africa, Sociocultural Overviews: West Africa; Central Africa: Sociocultural Aspects; Colonialism, Anthropology of; Magic, Anthropology of; Materiality and Culture; Religion, Anthropology of; Southern Africa: Sociocultural Aspects; Totemism; Witchcraft.

Bibliography

Apter, Emily, Pietz, William (Eds.), 1993. Fetishism as Cultural Discourse. Cornell University Press, Ithaca, New York.
Atkins, John, 1735. A Voyage to Guinea, Brasil and the West Indies. London.
Bernand, Carmen, Gruzinski, Serge, 1988. De l'idôlatrie: une archéologie des sciences religieuses. Seuil, Paris.
Bosman, Willem, 1705. A New and Accurate Description of the Coast of Guinea. J. Knapton, London.
Bourdieu, Pierre, 1992. Les Règles de l' Art. Gènese et Structure du Camp Literaire. Seuil, Paris.
Brosses, Charles de, 1988 [1760]. Du Culte des Dieux Fétiches ou Parallèle de l' ancienne Religion de l' Egypte avec la Religion actuelle de Nigritie. Fayard, Paris.
Comte, Auguste, 2001. Cours de philosophie positive. Book Surge Publishing, Paris.
Comaroff, Jean, Comaroff, John (Eds.), 1993. Modernity and Its Malcontents: Ritual and Power in Postcolonial Africa. University of Chicago Press, Chicago, IL.
D' Alembert, Jean Le Rond, Diderot, Denis (Eds.), 1751–1765. Encyclopédie française. Paris.
Dapper, Olfert, 1686. Description de l' Afrique. W. Waesberge, Boom et Van Someren, Amsterdam.
De Marees, Pieter, 1987 [1602]. Description and Historical Account of the Gold Kingdom of Guinea. Oxford University Press, Oxford.
Durkheim, Emile, 1968[1912]. Les formes élémentaires de la vie religieuse. Les Presses Universitaires de France, Paris.
Freud, Sigmund, 1927. Fetishism. In: Complete Works (SE), vol. 21. pp. 147–157.
Geischere, Peter, 1997. The Modernity of Witchcraft: Politics and the Occult in Postcolonial Africa. University of Virginia Press, Charlottesville.
Gell, Alfred, 1998. Art and Agency. Carendon Press, London.
Graeber, David, 2005. Fetishism as social creativity, or, fetishes are gods in the process of construction. Anthropological Theory, 5 (4), 407–438.
Hegel, Friederich, 1956. The Philosophy of History. Dover, New York.
Iacono, Alfonso M., 1992. Le fétichisme. Historie d' un concept. P.U.F, Paris.
Kant, Immanuel, 1960. Observations on the Nature of the Beautiful and the Sublime. University of California Press, Berkeley, CA.
Latour, Bruno, 1993. We Have Never Been Modern. Harvard University Press, Cambridge, MA.
Latour, Bruno, 2001. L' Espoir de Pandore. Éditions de La Découverte, Paris.
Latour, Bruno, 2010. On the Modern Cult of the Factish Gods. Duke University Press, Durham.
Marx, Carl, 1992. Capital: Volume 1: A Critique of Political Economy. Penguin, London.
Meyer, Birgit, Pels, Peter, 2003. Magic and Modernity: Interfaces of Revelation and Concealment. Stanford University Press, Palo Alto.
Pietz, Willam, 1985. The problem of the fetish, I. In: Res, Anthropology and Esthetics 9 (Spring), 5–17.
Pietz, Willam, 1987. The problem of the fetish, II: the origin of the fetish. Res, Anthropology and Esthetics 13 (Spring), 23–45.
Pietz, Willam, 1988. The problem of the fetish, IIIa: Bosman's Guinaea and the enlightenment theory of fetishism. Res, Anthropology and Esthetics 16 (Fall), 106–123.
Pouillon, Jean, 1975. Fétiches sans fetichisme. Maspero, Paris.
Sansi, Roger, 2007. The fetish in the Lusophone Atlantic. In: Naro, N., Sansi-Roca, R., Treece, D. (Eds.), Cultures of the Lusophone Atlantic. Palgrave MacMillan, London, pp. 19–40.
Sansi, Roger, 2011. Sorcery and fetishism in the modern Atlantic. In: Pares, L.N., Sansi, R. (Eds.), Sorcery in the Black Atlantic. University of Chicago Press, Chicago, pp. 19–40.
Spyer, Patricia (Ed.), 1998. Border Fetishisms. Routledge, London.
Taussig, Michael, 1980. The Devil and Commodity Fetishism in South America. University of North Carolina Press, Chapel Hill.
Tylor, Edward, 1874. Primitive Culture. H. Holt and Company, London.

Feudalism

Levi Roach, University of Exeter, Exeter, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Feudalism has a wide variety of meanings. Traditionally it refers to the sociopolitical order of Western Europe in the central and later Middle Ages (*c*.900–1500), though it has also been used to characterize premodern Asian and African societies displaying similar features. The nature and relevance of feudalism as a model has been a subject of much debate and it is important to distinguish between three distinct (though broadly related) understandings of the term: the Marxist conception of feudalism as an economic system based on the 'feudal mode of production', the social historical model of 'feudal society' as something approximating a Weberian ideal type, and the legal definition of feudalism as a form of service tenement based on the fief.

Feudalism is a subject of enduring interest to historians, sociologists, and economists. Although definitions vary widely, the term generally designates the socioeconomic order of Western Europe in the central and later Middle Ages (*c*.900–1500). Feudalism has been the subject of intense debate and it is important to distinguish broadly between three models of the phenomenon (which amount, in effect, to different definitions of feudalism): the Marxist or historical materialist, the social historical/sociological, and the legal-tenurial (see Wickham 2001). Each of these is based on different premises, though all purport to describe aspects of European society in the Middle Ages.

Feudalism: The Noun and Its Derivation

Feudalism derives from the Medieval Latin *feudum* (plural: *feuda*), referring to a form of conditional land-tenure. *Feudum* is generally rendered into Modern English as 'fief' and is believed to derive from the Proto-Germanic **fehu*, which originally meant cattle (cf Modern German *Vieh*), though it also came to have the broader connotation of 'wealth' or 'property' (cf Modern English 'fee') (Arcamone, 2001). As an abstract noun feudalism makes its first appearance in English in the nineteenth century, when it began to be employed as an equivalent to the French *féodlité*, which is attested about a century earlier. As originally employed, feudalism refers to a tenurial system based around fief-holding. It continues to be used in this (strict) legal sense, as we shall see, but has also taken on broader connotations.

The Legal Model: Feudal Tenure and the 'Feudo-Vassalic Contract'

The pioneering work on feudalism was undertaken in eighteenth-century France. Within this context, the term (French: *féodalité*) was used to characterize the tenurial structure of the *Ancien régime*; feudalism was seen as the system of land holding, which evolved during in the Middle Ages. This legal-tenurial model of feudalism continues to be employed by historians and is sometimes termed feudalism in the 'strict' sense, since its places a strong emphasis on fief-holding as its defining characteristic (see Ganshof, 1961: pp. xv–xviii). In some languages this model is even distinguished from its other socioeconomic counterparts. Thus in German *Lehnswesen* denotes feudalism in a legal sense, while *Feudalismus* is preferred for other models, and in Dutch a similar distinction is made between *leenstelsel* and *feodalisme*.

According to the legal model, feudalism is characterized above all by feudal tenure: it is a system of land holding based on fiefs or benefices (*beneficia* (singular: *beneficium*): the latter term predominates in sources before the eleventh century), a form of conditional tenure, generally for one or more lifetimes. For tenure to qualify as truly feudal, however, the tenant must not only hold his land as a fief or benefice, but also be the vassal of his landlord. Thus feudalism combines a tenurial agreement with a social bond (in the technical language of feudal law these are referred to as 'enfeoffment' and 'vassalage' respectively). Traditional teaching holds that enfeoffment and vassalage were originally distinct, but came to be associated in the course of the seventh and eighth centuries, when the powerful Carolingian dynasty came to prominence in mainland Europe. This period saw much infighting within Francia (a region encompassing much of modern France as well as parts of Germany, the Low Countries, Switzerland, Austria, and northern Spain) as well as the rapid expansion of the Frankish realm from 751 onward. As a consequence – so the traditional wisdom – vassalage and fief-holding became associated: in order to attract ever larger numbers of warriors, the Carolingian rulers rewarded their followers with temporary landholdings in the form of benefices (generally taken from the Church). The result was that benefice-holding (or enfeoffment, as it would later be known) became related closely with vassalic status, eventually leading to the unification of the two phenomena (Ganshof, 1961: pp. 16–50, esp. 16–19; Mitteis 1933: pp. 521–524). This 'union of fief and vassalage', of a tenurial agreement and a personal bond of lordship, was the foundation of European feudalism. Although this first emerged in the Frankish heartlands, it was soon exported across the rest of Western Europe by means of conquest and settlement.

This form of tenure in exchange for service was defined by two rituals, homage and investiture, relating respectively to the

two elements of the feudal relationship (the personal and the tenurial). The first forged the bond of service: by placing his hands within those of his lord and swearing fealty a man became a vassal. The second, which generally followed on from the first, enacted the property transaction: by handing over objects symbolizing tenure and service (often swords, spears, or banners), the lord invested his vassal in his fief or benefice. Together these acts created the so-called feudo-vassalic contract (Mitteis, 1933: pp. 481–518; Ganshof, 1961: pp. 70–72 and 133–142; Le Goff, 1980). There were important consequences of this for both sides: henceforth the lord could expect loyalty and service – the 'advice and assistance' (*consilium et auxilium*) mentioned in many contemporary texts – from his man; the vassal, on the other hand, could anticipate receiving support and protection from his lord. In theory, if either party failed in its obligations, then the contract was dissolved and the property reverted to the lord. The contract also came to an end upon the death of either party. In the early days of feudalism, the lord might then enfeoff a new vassal; however, fiefs had a tendency to become inheritable and from the late ninth century what tended to happen was that homage and investiture would be repeated whenever lord or man died (what in German is termed *Herrenfall* and *Mannfall*). In this manner the fiction of the fief's reversion to the lord was retained, while the rights of the vassal's heirs to inherit were respected. In some parts of Europe heirs also had to pay a special due in order to inherit (a 'feudal relief'); nevertheless, the expectation of inheritance was generally strong (Mitteis 1933: esp. 135–146; Ganshof, 1961: esp. 133–139; cf Bonnassie, 1991: pp. 172–174).

The strength of this model of feudalism lies in its precision: it is possible to test for the presence of fiefs and vassalage empirically. Its greatest danger, however, lies in its prescriptiveness. The scholars responsible for it, above all the Belgian historian François Louis Ganshof and his German counterpart Heinrich Mitteis, tended to see feudal tenure in very legalistic terms. As a consequence, they frequently imposed modern tenurial and legal clarity onto affairs which were less formalized (indeed, in their work both scholars were forced to acknowledge many 'anomalies' and 'regional variations'). More problematic, however, is the premise on which this model rests: namely, that benefice-holding and vassalic status coalesced in the eighth and ninth centuries. As we shall see, there is good reason to doubt that this took place (at least so early). Rejecting such an early unification of fief and vassalage has significant implications. Whereas Ganshof and Mitteis treated feudalism as an early medieval phenomenon, treating the development of rights of inheritance and the multiplication of vassalic bonds as signs of the weakening of this system in the central Middle Ages, it now seems that these latter phenomena an integral part of feudalism from the start. As a consequence historians increasingly see the central and later Middle Ages as the heyday of feudal tenure, rather than as a period of decadence and decline (Carpenter, 2000; Spieß, 2009, 2013).

The Marxist Model: The 'Feudal Mode of Production'

As noted, beside the legal there are two other distinct models of feudalism: the historical materialist (or Marxist) and the social historical. These are broadly similar, insofar as they both treat feudalism as more than a form of tenure. Nevertheless, they are based on very different premises: the former on a Marxist understanding of society and economy as being shaped by distinctive 'modes of production', and the latter on a more sociologically inspired understanding of societal 'ideal types'.

The defining feature of feudalism in the Marxist sense is the so-called feudal mode of production. According to Marx, a mode of production (German: *Productionweise*) is defined by productive forces (encompassing both human labor and tools and technology) and relations of production (the societal relations governing production). Marx postulated a number of different modes, the most important of which within this context are the slave (or antique) mode, the feudal mode and the (early) capitalist mode, each of which succeeded one another in Western Europe between Roman antiquity and the Early Modern period. In all of these modes agriculture is the main source of surplus; what distinguishes them is the means by which this is generated: in the antique mode it is slave labor which underpins production, in the feudal mode it is peasant tenant labor, and in the capitalist mode it is wage labor. Any society can therefore be considered 'feudal' in a Marxist sense if production within it is defined by the work of peasant tenants. One of the salient features of societies dominated by the feudal mode is a tendency for competition to emerge between central authority and the ruling classes (often in the form of a struggle between center and periphery); since the latter derive their wealth from the same sources and in the same manner as the state, centrifugal tendencies are common, though by no means inevitable (see, with different emphasis, Anderson, 1974; Haldon, 1993).

Although Marx was relatively uninterested in precapitalist modes, a rich literature has developed on the theme. The greatest problem here has been distinguishing between different modes, and these debates have important implications for our understanding of the feudal mode. Hence, though the feudal mode is often presented as a preeminently European and medieval phenomenon (e.g., Anderson, 1974), some now prefer to define feudalism more broadly, treating any society in which surplus wealth is generated primarily by peasant labor as feudal (Haldon, 1993; Wickham, 2005: pp. 259–261; 2008). Seen in these terms, the feudal mode is the most common precapitalist mode, which is attested not only in medieval Europe, but also across much of Asia and Africa before the modern era. Indeed, it has been suggested that we would do better to speak in terms of a 'tributary' rather than 'feudal' mode for this reason (Haldon, 1993: pp. 63–69 and 91–92). Though many societies may qualify as 'feudal' in this sense, important distinctions can be discerned among them. Chris Wickham points to the extraction of surplus in the form of rent and services rather than taxation as the defining feature of medieval European feudalism (Wickham 1984; Wickham, 2005: pp. 56–62; cf Blank, 2013). However, it is precisely the variety witnessed within the feudal mode thus conceived which has led some commentators to call for a more restricted usage of feudalism within Marxist analysis. Jairus Banaji, in particular, argues that feudalism is in danger of losing its explanatory force when defined so broadly, suggesting that such work underestimates the differences between medieval

European feudalism and other tributary forms of society (Banaji 2010: esp. 181–250).

Sociological Models: Feudal Society and the *Annales* School

Perhaps the most widely employed model of feudalism is the sociological or social historical. Feudalism in this sense has been studied most diligently by historians of the *Annales* School, whose approach takes inspiration from sociology, treating feudalism something akin to a Weberian 'ideal type' (Wickham 2001: pp. 34–41). The starting point for this work was Marc Bloch's monumental *La société féodale* (Bloch, 1962 (first published in French in 1939)). Within this work, Bloch aimed to present rather than define feudalism and it was only after having completed his two-volume survey that he felt in a position to summarize its salient features: a subject peasantry; the prevalence of the service tenements (fiefs); the dominance of a warrior aristocratic class; the prominence of personal ties of obedience (vassalage); the fragmentation of public authority; and the survival of other (earlier) forms of association (Bloch, 1962: p. 446). Another important feature of feudal society was manorialism, though this was not a defining characteristic thereof. Like historians of feudal tenure, Bloch saw this society as a preeminently Frankish phenomenon; however, unlike his more legally minded counterparts he did not believe that the history of feudal society could be reduced to the story of one form of service tenement (viz, the 'feudo-vassalic contract'). According to Bloch, feudal society developed in two stages, neither of which could be defined by tenurial developments alone: a first feudal age, from the late Carolingian period up until the eleventh century, and a second, from the mid-eleventh to the thirteenth century. The first age witnessed the development of a feudal society: in the face of invasions by Vikings, Muslims, and Magyars, centralized authority began to decompose, giving way to a more localized and fragmented society, in which interpersonal bonds and regional power structures were predominant. The second feudal age saw the maturation and eclipse of this society: in this period a true hereditary nobility established itself and chivalric values developed; however, at the same time multiple vassalage and rights of inheritance began to threaten the centrality of the feudal relationships and mentalities. Bloch's treatment of feudal society set the parameters from future work, particularly within the Francophone world. Particularly influential was his treatment of feudalism as phenomenon that embraced all of society, from the greatest lords down to the lowliest peasants.

The most influential subsequent contributions to the field have come from Georges Duby, who was a product of Bloch's own *Annales* School. Duby developed and challenged Bloch's arguments about feudal society in a number of respects. In particular, he placed a greater emphasis on regional studies than had his illustrious predecessor: whereas Bloch sought to generalize about feudalism across all of Western Europe, Duby's most influential publication was his detailed study of the society of the Mâcon in Burgundy from the ninth to the twelfth centuries (Duby, 1971 (first published 1953)). Duby also parted ways with Bloch in other regards. He saw the development of feudalism in very different terms. Rather than postulating a first and second feudal age, Duby argued that feudal society emerged both later and more rapidly than Bloch had appreciated. His study of the Mâcon convinced Duby that public justice and authority, represented above all by the comital court or *placitum*, continued to operate well up until 980, suggesting that the decomposition of central authority diagnosed by Bloch had yet to set in at this point. Then, however, during a brief and tumultuous period between 980 and 1030 this system gave way to the (privatized) seigneurial system of justice, which was characteristically feudal (Duby, 1971: pp. 137–148). Other important changes could also be detected in these years: the disappearance of traditional forms of slavery and emergence of a servile peasantry, the formation of the nuclear family, and the development of a knightly class, to name but a few. Numerous scholars of the 1970s and 1980s followed Duby's lead, tracing how and when similar transformations took place across other regions of France, Spain, and Italy. The result of was a relatively coherent model of sudden transformation, around the year 1000, which Duby termed a 'feudal revolution' and others referred to as the 'feudal mutation' (*mutation féodale*). Therein was sought the explanation for numerous developments, from the end of slavery, to the Peace of God (a movement aimed at curbing aristocratic violence), to the growth of heresy (Duby, 1980; Bonnassie, 1991; Poly and Bournazel, 1991; Bois, 1992; cf Toubert, 1973; Bonnassie, 1975–1976).

At their best, these models avoid the pitfalls of their legal and Marxist counterparts: they neither limit feudalism to cases of a certain form of tenure, nor do they generalize it to such an extent as to deprive it of all distinctiveness. Nevertheless, these models present their own dangers. In particular, the 'feudal revolution/mutation' paradigm risks reducing complex socio-economic developments to what often amount to monocausal terms, as we shall see. Moreover, the use of the term feudal to describe this society raises the semantic issue of what is distinctively 'feudal' about it, if not the usage of service tenements in the form of fiefs. Nevertheless, despite these and other objections, it remains common to designate the society of the central Middle Ages feudal, particularly within Francophone scholarship (e.g., Mazel, 2010).

Modern Criticism I: 'The Tyranny of a Concept'

The concept of feudalism has been subjected to much criticism, particularly since the 1990s. Within the Anglophone tradition this has tended to focus on two related issues: the theoretical problem of how (or whether) it is helpful to employ feudalism as a sociohistorical model; and the empirical issue of whether feudalism in the legal sense ever existed. It was the first of these that Elizabeth Brown first raised in her justly famous article of 1974. There she argued that the concept of feudalism had grown too nebulous to be a useful tool of analysis; 'feudal', she pointed out, was becoming little more than a synonym for 'medieval'. Feudalism was, therefore, at risk of becoming a 'tyrannous construct', which inhibited rather than enhanced historical research. Brown's core objection to the term was thus semantic, but her arguments also went much further. She pointed out that feudalism was often employed in a dangerously reified fashion within scholarship: historians wrote of it as if it had an

independent historical existence, rather than being one model (among many) for understanding medieval society. Her ultimate conclusion was that historians should avoid the term feudalism altogether and restrict 'feudal' as an adjective to cases involving fiefs (Brown, 1974: p. 1086).

Criticisms of an altogether more fundamental nature were raised by Susan Reynolds in her 1994 monograph *Fiefs and Vassals*. Although Reynolds framed her arguments as a development of Brown's earlier critique, her target was quite different. Whereas Brown's criticisms were directed above all at historians who used feudalism as a socioeconomic model (who had tended to be most guilty of expanding the term's semantic range), Reynolds' were leveled at those who used feudalism in a legal-tenurial sense. In a nutshell, Reynolds' thesis was that feudalism in this form had never existed. Her argument proceeded from the belief that the union of fief and vassalage, which as we have seen was the lynchpin of tenurial models of feudalism, never took place; although fiefs and vassals existed during the Middle Ages, they were never associated in the formalized and systematic manner postulated by feudal theory. As such, the 'feudal system' was not a medieval creation at all, but a chimera, invented by lawyers and historians between the later Middle Ages and Early Modern era. Reynolds argued that this model was born in two distinct stages. First, jurists of the central and later Middle Ages (especially those responsible for the *Libri feudorum*) began to systematize existing property relations into a more formal guise. Then, during the Early Modern period Francophone legal experts picked up these terms, further systematizing the evidence to hand in the process. Feudalism is therefore a product of two stages of modeling and systematizing and cannot be expected to conform to (or illuminate) property relations in the early and central Middle Ages.

Reactions to these lines of criticism have been predictably mixed. Brown's arguments have been taken as a salutary warning about the semantic growth of the term feudalism and subsequent scholarship has employed it in a more cautious and self-conscious fashion (e.g., Wickham 2001). Nevertheless, her more strident calls to abandon the concept have found few supporters (Reynolds being perhaps the most prominent). Reynolds' work, on the other hand, has occasioned stronger feelings. Elements of her thesis have been warmly received. In particular, her call for a more flexible approach to the legal culture of the early and central Middle Ages has met with wide approval. Her arguments about the unification of fief and vassalage have also achieved a degree of acceptance and most historians would agree that earlier scholarship erred in placing this development in the Carolingian period. Nevertheless, Reynolds' more bold claims have met with resistance. In particular, her belief that fiefs and vassalage were never closely associated has convinced few (Barthélemy, 1997; Deutinger, 2002). Similarly, her insistence that feudalism *tout court* was an Early Modern invention has not been widely accepted (Wickham 2001; White, 2005b; nos. XI–XII; Patzold, 2012: pp. 43–94).

Modern Criticism II: Feudal Transformation

Criticism of a different nature has developed regarding the shift variously called the feudal revolution or mutation. As noted, it was Duby's belief that feudal society emerged not in two stages, as Bloch had suggested, but in one rapid transition around the year 1000. In this he was joined by many prominent French historians of the 1970s and 1980s and in 1994 these voices were joined by that of the American scholar Thomas Bisson, who diagnosed what he saw as a sudden and revolutionary rise in lordly violence across Europe in the early eleventh century (Bisson, 1994). Since then, however, voices of dissent have been raised. Thus, in a series of provocative articles and monographs starting in 1992 Dominique Barthélemy has sought to dismantle almost all aspects of the 'Duby thesis'. Like Brown, one of Barthélemy's concerns was that fidelity to a particular model (in his case, of feudal revolution) was straightjacketing scholarship: work produced within this framework tended to ask when and how transformation took place, without questioning the very mutationist paradigm on which it was built. In response, Barthélemy argued for a return to something closer to Bloch's original model of feudalism: he suggested that the social and political landscape of France changed slowly and subtly between the early and central Middle Ages, starting in the ninth century and continuing through to the twelfth and thirteenth (Barthélemy, 1992, 2009). In this respect, he argued that many developments, which first come fully to light around the year 1000, were already taking place much earlier. The turning of the first millennium did not so much witness a sudden societal transformation, as an important shift in the nature of the historical evidence, a 'documentary mutation' (*mutation documentaire*), as Barthélemy has neatly termed it (Barthélemy, 1993: esp. 19–127; Barthélemy, 2009: pp. 12–36).

Barthélemy has not been alone. Concerns of a similar nature have been raised by a number of American historians since the 1990s. Foremost among these are Stephen White, Frederic Cheyette, and Jeffrey Bowman, all of whom have argued against the mutationist model. One of their main criticisms is that it depends heavily on an anachronistic distinction between 'public' and 'private', treating early medieval comital courts as public institutions, but viewing their seigneurial successors as purely private. Yet, when examined closely, the operation of these courts reveals many similarities. Therefore, while there may have been developments in judicial practice around the turning of the millennium, to many lords and litigants such changes were probably neither so self-evident nor so fundamental as the mutationist paradigm proposes (White, 2005a: no. IX; Cheyette, 2002; Bowman, 2004). Indeed, it has been pointed out that the model of rapid transformation around the year 1000 runs the risk of oversimplifying change, presuming an a priori relationship between disparate religious, social, and economic developments (Bowman, 2004: pp. 226–227). More specific criticisms have been made of the feudal revolution as postulated by Bisson. Here it has been pointed out that the evidence for a sudden rise in violence in the eleventh century is probably more apparent than real (like the changes in documentary form diagnosed by Barthélemy), reflecting not only the increasing volume of source material, but also developments in approaches to historical writing. In this respect, it is important to note that most of our sources were written by monks and clerics, who had good reason to exaggerate (if not invent) the novelty of the forms violence they experienced at the hands of laymen (White, 2005a: nos. II–III; 2005b: no. XIII; Mazel, 2005).

Such criticisms have been well received and the proposition that all of Europe (or France) underwent a sudden feudal mutation around the year 1000 is no longer accepted. Nevertheless, the dangers of taking criticism too far have also been noted (Wickham, 1997; Fouracre, 2005). Indeed, there is no denying the changes which took place between the early and central Middle Ages and some recent work risks presenting a rather too static a picture of society in these years (Moore, 2000). In this respect, the antimutationist model is in danger of becoming just as monolithic as the revolutionary model it replaced.

Overview

Although the criticisms of feudalism – be they of the concept *tout court* or of the models used to frame it – have often sought to renounce the term altogether, in practice they have led to renewed interest in the subject. In particular, the work of Reynolds and Barthélemy has inspired scholars to return to first principles, revisiting the evidence for fiefs, vassals, and the complex interplay between these in the early and central Middle Ages. There is an emerging consensus that it is unhelpful to speak of feudalism in anything other than a Marxist sense before the central Middle Ages. Historians working within a legal-tenurial framework tend to see feudalism a development of the eleventh and twelfth centuries (Deutinger, 2002; Hyams, 2002; Dendorfer, 2004; Dendorfer and Deutinger, 2010; Roach, 2012), whereas those employing more sociologically informed models continue to see the eleventh century as the period in which classically 'feudal' forms of lordship emerged, though they emphasize the gradual and complex nature of developments (Kosto, 2001; Mazel, 2010; West, 2013). It would, therefore, seem that while 'textbook feudalism' never existed as such, something closely approximating it did. The relationship fiefs and vassalage is a case in point here: although it seems unlikely that these were ever as systematically associated as traditional teaching holds, it is equally clear that a close connection between the two developed during the eleventh and twelfth centuries. In this sense, there is something of an emerging reaction against what is perceived to have been the hypercriticism of the 1990s. However, above all recent work underlines the flexibility and adaptability of feudalism – as Steffen Patzold puts it, there was not so much one 'feudalism' as many forms of feudalism in Europe during the central and later Middle Ages (Patzold, 2012: pp. 120–121).

See also: Annales School; Bloch, Mark Léopold Benjamin (1886–1944); Land Tenure; Marx, Karl (1818–83); Middle Ages, The.

Bibliography

Anderson, P., 1974. Passages from Antiquity to Feudalism. NLB, London.

Arcamone, M.G., 2001. Germanico "fehu-patrimonio" e germanico "laihwna-prestito": contributo allo studio della terminologia feudale. Settimane di studio del Centro italiano di studi sull'alto medioevo 47, 915–943.

Banaji, J., 2010. Theory as History: Essays on Modes of Production and Exploitation. Brill, Leiden.

Barthélemy, D., 1992. La mutation féodale a-t-elle eu lieu? (Note critique). Annales: Economies, Sociétés, Civilisations 47, 767–777.

Barthélemy, D., 1993. La société dans le comté de Vendôme: de l'an mil au XIVe siecle. Fayard, Paris.

Barthélemy, D., 1997. La théorie féodale à l'épreuve de l'anthropologie (note critique). Annales: Histoire, Sciences Sociales 52, 321–341.

Barthélemy, D., 2009. The Serf, the Knight and the Historian (G.R. Edwards, Trans.). Cornell University Press, Ithaca, NY.

Bisson, T.N., 1994. The "feudal revolution". Past and Present 142, 6–42.

Blank, G.R., 2013. Rethinking the "other transition": towards an alternative Marxist explanation. Science & Society 77, 153–178.

Bloch, M., 1962. Feudal Society, vol. 2. (L.A. Manyon, Trans.). Routledge and Kegan Paul, London.

Bois, G., 1992. The Transformation of the Year One Thousand: The Village of Lournand from Antiquity to Feudalism (J. Birrell, Trans.). Manchester University Press, Manchester.

Bonnassie, P., 1975–1976. La Catalogne du milieu du Xe à la fin du XIe siècle, vol. 2. Publications de l'Université de Toulouse-Le Mirail, Toulouse.

Bonnassie, P., 1991. From Slavery to Feudalism in South-western Europe. (J. Birrell, Trans.). Cambridge University Press, Cambridge, UK.

Bowman, J.A., 2004. Shifting Landmarks: Property, Proof, and Dispute in Catalonia Around the Year 1000. Cornell University Press, Ithaca, NY.

Brown, E.A.R., 1974. The tyranny of a construct: feudalism and historians of medieval Europe. American Historical Review 79, 1063–1088.

Carpenter, D.A., 2000. The second century of English feudalism. Past & Present 168, 30–71.

Cheyette, F.L., 2002. Georges Duby's Mâconnais after fifty years: reading it then and now. Journal of Medieval History 28, 291–317.

Dendorfer, J., 2004. Was war das Lehnswesen? Zur politischen Bedeutung der Lehnsbindung im Hochmittelalter. In: Schlotheuber, E., Schuh, M. (Eds.), Denkweisen und Lebenswelten des Mittelalters. Hebert Utz Verlag, Munich, pp. 43–64.

Dendorfer, J., Deutinger, R. (Eds.), 2010. Das Lehnswesen im Hochmittelalter. Forschungskonstrukte – Quellenbefunde – Deutungsrelevanz. Jan Thorbecke Verlag, Ostfildern.

Deutinger, R., 2002. Seit wann gibt es Mehrfachvasallität? Zeitschrift der Savigny-Stiftung für Rechtsgeschichte: Germanistische Abteilung 119, 78–105.

Duby, G., 1971. La société aux XIe et XIIe siècles dans la région mâconnaise, second ed. SEVPEN, Paris.

Duby, G., 1980. The Three Orders: Feudal Society Imagined (A. Goldhammer, Trans.). University of Chicago Press, Chicago, IL.

Fouracre, P., 2005. Marmoutier and its serfs in the eleventh century. Transactions of the Royal Historical Society 6th Series 15, 29–49.

Ganshof, F.L., 1961. Feudalism (P. Grierson, Trans.), third ed. Harper Torchbooks, New York.

Haldon, J., 1993. The State and the Tributary Mode of Production. Verso, London.

Hyams, P.R., 2002. Homage and feudalism: a judicious separation. In: Fyrde, N.M., Mollat du Jourdin, M., Oexle, O.-G. (Eds.), Die Gegenwart des Feudalismus/Présence du féodalisme et présent de la féodalité/The Presence of Feudalism. Vandehoeck & Ruprecht, Göttingen, pp. 13–50.

Kosto, A.J., 2001. Making Agreements in Medieval Catalonia: Power, Order, and the Written Word, 1000–1200. Cambridge University Press, Cambridge, UK.

Le Goff, J., 1980. The Symbolic Ritual of Vassalage (A. Goldhammer, Trans.). Time, Work, and Culture in the Middle Ages. University of Chicago Press, Chicago, IL, pp. 237–287.

Mazel, F., 2005. Amitié et rupture de l'amitié: moines et grands laïcs provençaux au temps de la crise grégorienne (milieu XI–milieu XII siècle). Revue Historique 307, 53–96.

Mazel, F., 2010. Féodalités, 888–1180. Belin, Paris.

Mitteis, H., 1933. Lehnrecht und Staatsgewalt. Untersuchungen zur mittelalterlichen Verfassungsgeschichte. Böhlau, Weimar.

Moore, R., 2000. The birth of popular heresy: a millennial phenomenon? Journal of Religious History 24, 8–25.

Patzold, S., 2012. Das Lehnswesen. C. H. Beck, Munich.

Poly, J.-P., Bournazel, E., 1991. The Feudal Transformation (C. Higgitt, Trans.). Holmes and Meier, New York.

Reynolds, S., 1994. Fiefs and Vassals: The Medieval Evidence Reinterpreted. Oxford University Press, Oxford, UK.

Roach, L., 2012. Submission and homage: feudo-vassalic bonds and the settlement of disputes in Ottonian Germany. History 97, 355–379.

Spieß, K.-H., 2009. Das Lehnswesen in Deutschland im hohen und späten Mittelalter, second ed. Franz Steiner Verlag, Stuttgart.

Spieß, K.-H. (Ed.), 2013. Ausbildung und Verbreitung des Lehnswesens im Reich und in Italien im 12. und 13. Jahrhundert. Jan Thorbecke Verlag, Ostfildern.

Toubert, P., 1973. Les structures du Latium médiéval. Le Latium méridional et la Sabine du IXe siècle à la fin du XIIe siècle, 2 vols. École française de Rome, Rome.

West, C., 2013. Reframing the Feudal Revolution: Political and Social Transformation between Marne and Moselle, *C.* 800 to *C.* 1100. Cambridge University Press, Cambridge, UK.

White, S.D., 2005a. Feuding and Peace-making in Eleventh-Century France. Ashgate Publishing, Aldershot, UK.

White, S.D., 2005b. Re-thinking Kinship and Feudalism in Early Medieval Europe. Ashgate Publishing, Aldershot, UK.

Wickham, C., 1984. The other transition: from the ancient world to feudalism. Past & Present 103, 3–36.

Wickham, C., 1997. The "feudal revolution": a comment. Past & Present 155, 196–208.

Wickham, C., 2001. Le forme del feudalesimo. Settimane di studio del Centro italiano di studi sull'alto medioevo 47, 15–46.

Wickham, C., 2005. Framing the Early Middle Ages: Europe and the Mediterranean, 400–800. Oxford University Press, Oxford, UK.

Wickham, C., 2008. Productive forces and the economic logic of the feudal mode of production. Historical Materialism 16, 3–22.

Feyerabend, Paul (1924–94)

Daniel Kuby, University of Vienna, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Paul Feyerabend (1924–94) was a philosopher who gained recognition for his views on science and the role of science in society. Feyerabend is best known for his epistemological anarchism, the critique of the privileged position of science in society, and his endorsement of relativism. He made several contributions to the history and philosophy of science by studying the incommensurability of scientific theories and pluralistic nature of science, offering a new interpretation of the Scientific Revolution. His critique of classical empiricism and formal accounts of rationality remains his legacy in social studies of science and contemporary philosophy of scientific practice.

Paul Karl Feyerabend (born in 1924 in Vienna, died in 1994 in Genolier, Switzerland) was an Austrian-American philosopher who gained recognition for his views on science and the role of science in society. Feyerabend is best known for his epistemological anarchism, the critique of the privileged position of science in society, and his endorsement of relativism. Within the history and philosophy of science he is regarded as one of the preeminent critics of the 'received view' of Logical Empiricism and, later, Critical Rationalism.

Feyerabend enrolled at the University of Vienna studying mainly physics and philosophy, receiving his PhD in philosophy in 1951. He studied with Karl Popper at the London School of Economics the following year. Because of dim prospective academic chances in Austria, he accepted a position in Bristol, UK in 1955. From there, he accepted a permanent position at UC Berkeley in 1959, which he maintained till 1990. While in California, he accepted temporary positions at universities around the world, including the 1980 tenure at ETH Zürich, where he resigned in 1991. He died of a brain tumor only a few years later, in 1994, near Geneva.

His academic contributions range over a wide array of areas and topics: In general philosophy of science, he introduced the concept of incommensurability of scientific theories and, building on Karl Popper's falsificationism, developed several arguments for theoretical pluralism in the methodology of science; he later severely criticized mainstream philosophy of science for its shortcomings. In epistemology, he developed a behavioristic account of basic sentences and constructively discussed several problems of empiricism. In philosophy of mind, he is known for proposing eliminative materialism as a possible solution to the mind–body problem. In analytic philosophy of language, he delivered an early interpretation of Wittgenstein's *Philosophical Investigations* and a metaphilosophical critique of linguistic analysis. In the philosophy of physics, he first criticized the so-called Copenhagen interpretation of quantum mechanics, but soon came to appreciate aspects of Niels Bohr's scientific work and was influenced by David Bohm. Closely related to his interest in physics, in history of science, he contributed studies on the early development of quantum mechanics, put forward a controversial interpretation of the establishment of Copernicanism in the sixteenth century, and wrote a series of studies tracking the rise of rationalism in the Western history of ideas. He also highlighted the similarities between the history of science and art history as historically interwoven crafts. Additionally, he wrote about dramatics, commenting on Bertolt Brecht and Eugène Ionesco, and in political philosophy developed liberal ideas founded on John Stuart Mill and libertarian thought.

Although his patchwork-like philosophical works and writing style defy any easy systematization, there have been a number of proposals identifying common themes in Feyerabend's work. For one, his early allegiance to Popper's Critical Rationalism has been interpreted as an attempt to develop a systematic model of knowledge acquisition, furthering falsificationism, while also taking his cue from Wittgensteinian ideas. His later philosophy is, then, the result of relinquishing such a model, while developing early arguments to unwarranted extremes and adopting postmodern views (Preston, 1997). On another account, disparate works of his can be coherently grouped as sophisticated critiques of cases he viewed as manifest or lingering examples of conceptual conservatism in philosophy and the sciences. On this view Feyerabend – in true skeptical fashion – lacked allegiance to particular philosophical positions, employing them only instrumentally as *ad hominem* arguments against specific cases of conservatism (Oberheim, 2006). A third account highlights the roots of his views in the Viennese tradition of scientific philosophy, giving particular prominence to his early philosophy and later defense of scientist-philosopher Ernst Mach. On this view Feyerabend remained a distinctive 'Viennese philosopher' throughout his life, who furthered views pertaining to the intellectual and scholarly tradition of Central Europe (Stadler, 2008). Related to this interpretative scholarship, a more evaluative perspective has advanced charges of irrationalism (Stove, 2006) and postmodernism (Preston, 1998) against his later work, as well as defenses against these charges (Farrell, 2003, 2001).

From a sociological point of view, his influence within academe has been twofold. Staying in close contact with Karl Popper and collaborators allowed him to interact with like-minded scholars, though increasingly pernicious group dynamics led him to distance himself from the Popperian school, thus breaking with foremost academic interlocutors based at the London School of Economics (Collodel, forthcoming). Feyerabend also felt intellectually at home with the Minnesota Center for Philosophy of Science, led by Herbert Feigl (Stadler, 2008). Starting in 1957, he submitted several of his ideas to interdisciplinary scrutiny at the Center,

contributing to his fame as a rising star in the philosophy of science. However, with the broad exposure of *Against Method* in 1975 his more provocative ideas were received overwhelmingly negatively in the field of analytic philosophy of science, leading to vitriolic responses by Feyerabend and to a separation from the field. All the same, Feyerabend continued to receive job offers in philosophy of science from prestigious universities, permitting him to work unconstrained from particular allegiances, though, lacking a definite community. His legacy in philosophy has been similarly limited by the fact that he supervised very few students. Curiously enough, while the discussion of his work in philosophy decreased, central ideas found in Feyerabend's work – such as the pluralism and disunity of science – became commonplace in contemporary history and philosophy of science and scientific practice (Cartwright, 1999; Dupré, 1995). With similar dynamics, debates about expertise, democratic decision-making, and epistemic justice play a central role in contemporary philosophy of democracy and public policy (Kitcher, 2011, 2003) and epistemology (Fricker, 2009), but rarely acknowledge Feyerabend.

The exposure of his ideas in disciplines other than philosophy grew all the same, leading to notoriety outside the academe. Feyerabend has been widely read and referenced as one of the main theoretical resources in the post-Mertonian sociology of scientific knowledge (SSK). Backing away from Merton's 'weak program,' the emergence of social studies of science found the epistemological critique of 'postempiricist' philosophy of science by the likes of Feyerabend, Thomas Kuhn, and Richard Rorty to be a congenial framework for their approaches to science. While Feyerabend did not take active part in this reception, both the later Feyerabend and SSK shared an interest in anthropological approaches and conceptualized the ethnologist's view as a model of inquiry for their own studies.

Scientific Philosophy and Methodology

Feyerabend's early philosophical efforts were directed at reworking empiricism into an antifoundationalist enterprise, in which rock-bottom experiential certainty is given up in favor of overall testability and comprehensive corrigibility. His philosophical proposals grew out of scientific philosophy, aided by empirical findings of the sciences, logical analysis, and a deflationary strategy applied to philosophical concepts, which prompted to reinterpret major concepts of traditional philosophy (such as 'observability' and 'realism') as public dispositions of cognitive agents. Feyerabend used this behaviorist strategy time and again to make philosophical terms accessible to empirical analysis, to subject them to pragmatic reasoning about means–end relations and, more specifically, to include the cognitive agent in a third-person perspective as a necessary component of a scientific account of empiricism.

The behaviorist strategy opens the way for a distinctive conception where methodological rules can be interpreted as behavioral norms for attaining given epistemic aims in science. This conception of philosophy of science presented a line of action to approach an issue in his philosophy: the distinctive role of decisions in science. Perhaps surprisingly, the early Feyerabend cherished the distinction between 'facts' and 'conventions' or 'conventional decisions' in science. The distinction may be understood as discerning matters determined by fact and matters underdetermined by fact, i.e., not to be settled by reference to matters of fact alone, and thus open to decision. While the distinction is usually employed to underline the factual, stable content of science, Feyerabend used it to stress how many instances are a matter of decision and, therefore, show the pliability of science by human volition. This, in turn, was mirrored in his philosophy of science, which, as scientific philosophy, was after all conceived as being modeled after science. Methodology both exploits this freedom and tames its potential arbitrariness by devising means to end relations between freely available epistemic aims and the constrained possibilities to attain them: which avenues lead to certain epistemic aims? The behaviorist account further fleshes out the means: how should scientists act in order to achieve a given epistemic aim?

Feyerabend's early methodological efforts are focused on empiricism and the testability of scientific theories as a fundamental epistemic aim of science. The earliest analysis of empiricist accounts of science, found in his dissertation (1951) and later developed in 1958 and 1962, targets foundationalist accounts in Logical Empiricism. His notion of testability crucially builds on Popper's notion of empirical content where the empirical content of a universal statement is given by the class of its potential falsifiers. The larger the class of potential falsifiers, the greater its empirical content and thus testability (or, in Popper's terminology, falsifiability). Several distinct topics discussed by Feyerabend, like realism and instrumentalism, theoretical monism and pluralism, incommensurability and conceptual conservatism, are tied together by a simple methodological question: which avenues directly or indirectly increase or decrease the testability of scientific theories? Even though these issues may not be settled by appeal to matters of fact, they can be settled methodologically: if one is an empiricist, he argues, exactly those avenues which increase testability should be adopted, while those views which decrease testability should be dismissed. (The important presupposition here is that these issues can be sharply determined by appeal to epistemic aims.)

Feyerabend's early criticism targets: theoretical monism, the view that the reduction (or elimination) of concurring scientific theories is an indication of scientific progress; instrumentalism, the view that scientific theories are only predictive devices; and conceptual conservatism, an umbrella term introduced by Oberheim (2006) to collect several arguments devised to constrain future conceptual development by appeal to allegedly stable structures of components of our knowledge and language. Feyerabend goes to great length to show that the adoption of these views indirectly or directly decreases the testability of scientific theories and should therefore be opposed by 'good empiricists' (Feyerabend, 1963). Conversely, Feyerabend presents distinct arguments to show that the adoption of theoretical pluralism, realism and allowing for radical reconceptualizations increase the testability of scientific theories. The following sections examine these topics in more detail.

Pragmatic Theory of Observation

In the late 1940s, within the Kraft Circle, Feyerabend became interested in the Vienna Circle discussion of the early 1930s concerning the question of the empirical basis of science. (This discussion, known today as the protocol sentence debate, appeared as written articles in the journal *Erkenntnis*.) The discussion revolved around the following problem: Supposing that empirical sciences can be understood as a system of sentences, how should the subset of basic sentences (usually observational statements) that are used to test further sentences (usually theoretical statements) be demarcated and characterized?

Feyerabend's contribution, which he dubbed a causal or pragmatic theory of observation, was first presented in print in 1958 and then in 1962, yet it was already developed in detail in his unpublished dissertation (1951). His contribution is a behavioristic account of basic sentences according to which sentences uttered by an observer under certain circumstances are public dispositions of that observer. Instead of relying on phenomenal certainty of experience by appealing to phenomenalist or sense-data epistemology, which he finds empirically and logically flawed, Feyerabend's account is explicitly cast to end the 'special' treatment of human observers, by putting them on a par with other scientific instruments, whose capacities have to be checked and calibrated like any other part of the scientific process. According to the pragmatic theory of observation, in order to assess whether the sentence uttered by an observer is an observation sentence, it is necessary to assess how reliably this linguistic disposition is correlated with the presence of some state of affairs that is being observed. Feyerabend's behavioristic proposal models epistemic agents as sensible measuring instruments, which under certain conditions react in a certain way – the utterance of sentences being one such reaction. "We arrive at an empirical theory," Feyerabend writes, "which resembles in many respects the theory of thermometers. […] The theory establishes how reliably we can infer that 'something' exists from the fact that this 'something' is 'immediately given' to an observer" (Feyerabend, 1951: p. 58). The theory tells us which conditions need to be fulfilled in order to establish a causal relationship between the utterance of an observer and some event.

It is important to notice that within Feyerabend's account sentences uttered by an observer are simple physical facts, not interpreted linguistic statements. Feyerabend's analogy to thermometers, or measuring devices more generally, is meant quite literally. According to the pragmatic theory of observation, human observers are measuring devices just like thermometers. Like a scale signaling a certain value in correlation with a certain event, observation sentences are an observer's reaction (signal) to a certain stimulus (event). Whether the observer was reliably trained to raise her hand, simply growl, or utter the sentence "There's a blue flower" in reaction to such a stimulus is unimportant at this stage. Thus these utterances can be reliable indicators that something exists, but they alone do not give any indication of what exists – they are not yet descriptions of some state of affairs. They become descriptions only when being used in a theory, which provides their interpretation.

This account of basic sentences was extensively employed in Feyerabend's later papers under the label 'pragmatic theory of observation': "An observation sentence is distinguished from other sentences of a theory not by its content but by the cause of its production, by the fact that its production conforms to certain behavioral patterns" (Feyerabend, 1962: p. 36).

Theoretical Pluralism

Theoretical pluralism is the view that the proliferation of concurrent scientific theories is conducive to scientific progress and a sign of good scientific practice. Feyerabend thus counters the view that an ever growing limitation of concurrent theories is, instead, an indication and a desirable feature of scientific progress. This view, theoretical monism, figures in many accounts that view scientific progress as an advancing approximation to the one true theory. But Feyerabend also found the same view to be implicit in less idealized and more historical accounts, like Kuhn's *Structure of Scientific Revolution*, where theoretical monism is inscribed in the notion of paradigm, the necessary context in which normal science occurs. This aspect prompted Feyerabend to severely criticize Kuhn's model (Feyerabend, 1970a; Hoyningen-Huene, 2006).

The basic tenet of Feyerabend's theoretical pluralism is the *principle of proliferation* that scientists willing to increase the testability of established theories ought to "invent, and elaborate theories which are inconsistent with the accepted point of view even if the latter should happen to be highly confirmed and generally accepted" (Feyerabend, 1965: pp. 223–224). Feyerabend's adduces many arguments to support this principle. The earliest and most technical is Feyerabend's contention that a specific class of observations may become actual falsifiers of an established theory only if they are part of a concurrent theory. In such cases, only the development of a concurrent theory can produce instances that may falsify the established theory. This means that the empirical content of a given theory can be augmented by developing and improving alternative theories. Feyerabend elaborated on this possibility by discussing e.g., the historical case of Brownian motion in physics. He contented that Brownian motion, while having been already a well-known phenomenon, became supporting evidence for statistical thermodynamics and a refutation of classical thermodynamics only with the advent of kinetic theory.

Feyerabend introduced a further principle of theoretical pluralism: "The principle of proliferation not only recommends invention of new alternatives, it also *prevents the elimination of older theories which have been refuted*" (Feyerabend, 1965: pp. 224–225). The latter he called the *principle of tenacity*. From early on Feyerabend's interpreted falsificationism to extend to falsifying instances as well; and "testing a falsifying instance in all possible ways is just the same as trying out all possibilities of making it compatible with the proposed theory" (Hoyningen-Huene, 2006: p. 627). A fact inconsistent with the theory under scrutiny is not *ipso facto* a falsifying instance; it ought to be challenged by the very theory it is inconsistent with and it has to earn the right to become an actual falsifier. Yet "once a falsifying instance has been produced the theory falsified will have to go"

(Hoyningen-Huene, 2006: p. 627). The principle of tenacity goes further and invites to consider refuted theories as part of the realm of relevant theories: "The principle of tenacity is reasonable because theories are capable of development, because they can be improved, and because they may eventually be able to accommodate the very same difficulties which in their original form they were quite incapable of explaining" (Feyerabend, 1970a: p. 204).

Realism

Feyerabend has produced several arguments for a realistic interpretation of scientific theories. His realist position has puzzled many contemporary commentators, since his conception of realism does not resemble today's scientific realism; indeed, it denies central features of it. The disagreement in the early realism debate concerned almost exclusively the question of whether scientific theories should be taken at face value as truth-apt descriptions of the world. Proponents of a (strict or relaxed) verificationist approach would reject it ('positivism'), while empiricists repudiating verificationism would accept in various ways the claim that statements about theoretical entities could be truth-apt ('realism'). Thus the discussion was about semantic realism and both sides mostly agreed that epistemological and ontological considerations in support of realism – unlike today – were metaphysics to be rejected. From the perspective of today's scientific realism one can appreciate the latter position at best as a 'realist faction' in a principled empiricist framework. At that time, however, proponents in this tradition would self-identify as realists. Feyerabend's self-professed early realism must be positioned in this early context and interpreted in light of this early philosophical terminology.

In Feyerabend's realism his theoretical pluralism, critique of sense-data epistemology, the methodological nature of philosophy, come together. The earliest negative argument for Feyerabend's realism is that the alternative, namely positivism, is an untenable groundwork for empiricism. He simply shows that verificationist accounts of experience all rely on untenable sense-data epistemology (or equivalent accounts) thus blocking the main motivation to deny the meaningfulness of theoretical concepts. Theoretical concepts may be truth-apt after all. His positive argument for realism draws on his methodological understanding of the realism debate. Feyerabend's main arguments are pragmatic in character: "Is it useful to believe that scientific theories are true?" In this framework, the first premise of his argument for realism is that a realistic interpretation of scientific theories is conducive to maximize their testability (while instrumentalist interpretations reduce their testability). Since testability is a fundamental epistemic goal of the scientific enterprise (second premise), it follows that scientific theories ought to be interpreted realistically.

Feyerabend contends that allowing for a putative truth-value of statements about theoretical entities expands the class of potential falsifiers of the theory containing said statement, thus increasing its testability: simply put, if we take statements about theoretical entities to be truth-apt, those statements may turn out to be false. Under an instrumentalist interpretation, by contrast, the theory's statements about unobservables are not truth-apt and, therefore, not capable of being refuted. And this is, indeed, Feyerabend's contention: under a realistic interpretation a theory's empirical content increases, while under an instrumentalistic one it decreases. (As new observations alone cannot refute unobservables, this demand crucially extends to the proliferation principle: only realistically interpreted alternative theories may posit theoretical entities contradicting the theory under scrutiny.)

Feyerabend's disavowal of this general realism advanced throughout the 1960s with his increasing skepticism toward general methodological rules. His studies made him aware of cases in which scientists would contravene general methodological rules with good scientific reasons, as dictated by the specific situation of their research. While Feyerabend would still take exceptions on general methodological grounds, he came to appreciate instances of local instrumentalism. A scientist by training, Feyerabend appreciated how such instances made him doubt the viability of his previous conclusion that scientific theories ought to be interpreted in a certain way no matter what the particular research situation is. Abandoning this general realism, he came to appreciate both locally justified interpretations and methodological rules tied to specific theories and research situations.

In the 1980s and early 1990s, Feyerabend's position regarding realism has been variously interpreted, ranging from a strong idealism cast in social-constructionist terms (Preston, 1998), to 'process-realism' (Farrell, 2001), to a sort of neoplatonism (Kidd, 2012). Having renounced his earlier scientism, i.e., the view that science gives us the only, or, at least, better means to attain knowledge about the world, Feyerabend extended his pluralism beyond science. For a time Feyerabend seems to have entertained the idea of an ontological disunity: different ways of knowing the world may hint at different realities being discovered. This definitely changed when he later introduced a strong metaphysical view about the unity of reality in the form of a 'Being' or simply 'Nature,' as he variously called it: The different realities that are discovered through radically diverse ways of knowing the world are, in fact, different ways in which the one ineffable Nature reacts to these discovery attempts. Further, Feyerabend questioned the assumption of a principled separability between the act of discovery and resulting entities or processes being discovered, arguing that the entities or processes being discovered may not exist independently of the circumstances of their discovery (Feyerabend, 1999).

Incommensurability

The term 'incommensurable' ('without a common measure') was first used in 1962 by both Paul Feyerabend and Thomas Kuhn to describe a specific relation – or lack thereof – between pairs of universal theories (Feyerabend) or paradigms (Kuhn) occurring in the history of science. Though Feyerabend's and Kuhn's accounts share some features and are directed against earlier accounts of scientific progress in philosophy of science, they were developed independently and are not easily reconcilable (Hoyningen-Huene, 2004).

Feyerabend's notion of incommensurability can be divided into different stages, starting from a logico-linguistic conception in the 1950s and ending with a hermeneutic-ontological conception in the 1990s. The possibility of incommensurability appears first as a counterfactual thought experiment in his dissertation (1951), but he only started to think about it as an actual phenomenon a few years later, finally leading to the best known and most discussed version of the incommensurability thesis of 1962.

The best known version of the incommensurability thesis states that mutually logically disjoint pairs of scientific theories present themselves as an actual phenomenon in the history of science and that this phenomenon is necessary for the ongoing development of science. The thesis was directed against accounts of intertheoretic reduction and explanation put forward by Logical Empiricism. It was argued that famous transitions in physics, from Galileo's law of free fall to Newtonian mechanics, and from the latter to Einstein's special theory of relativity, would adhere to this schema. The incommensurability thesis maintains that this account fails in exactly these paradigmatic situations. The main argument presented by Feyerabend in support of his refutation contended that the reduction relation presupposes a principle of meaning invariance across consecutive theories, such that the meaning of central theoretical terms in the former theory, as well as the meaning of observational terms, would be retained in the newer theory. This principle, however, does not obtain when considering pairs of realistically interpreted universal theories, because the meaning of central theoretical terms and the meaning of at least some observational term of the older theory is not retained by the successive theory. Instead of reducing or explaining the older theory, the newer theory replaces it. While Feyerabend conceded the later theory would often retain the previous theory as a limiting case, the former would just retain a numerical surrogate of the latter, not its conceptual apparatus. Still, Feyerabend maintained that the proliferation of incommensurable theories is essential for the progress of science as they instantiate the strongest case of alternative theories that may mutually contribute to the increase of their respective empirical content, aiding theoretical pluralism.

The logico-linguistic version of the incommensurability thesis sparked a strong debate that targeted the tenability of the very concept of incommensurability as well as its application to specific cases in the history of science. Among the objections were the rivalry objection (How can two theories be both logically disjoint and yet contradict each other?) and the relevance objection (How can two incommensurable theories be said to apply to the same empirical domain?). Feyerabend replied to these objections throughout the 1960s, trying to correct what were to become long-lasting misreadings like the contention that incommensurability implies incomparability between scientific theories, a misreading far from Feyerabend's view. In fact he maintained that incommensurable alternatives, while being problematic for formal philosophical accounts of scientific development, give a variety and *better* means of comparing the merits of theories. However, under pressure and becoming increasingly dissatisfied with his own precisations, he finally gave up trying to devise a precise account, declaring to be unable to formulate such an account.

Nevertheless, he did not gave up the concept itself, maintaining that incommensurability has an important explanatory and descriptive role to play in the history of ideas and cultural practices. Starting with the introduction of epistemological anarchism, he recast incommensurability as a historical–hermeneutical thesis to be shown – rather than proved – by means of historical–anthropological analysis in subsequent frameworks of thought, in changes of perception frameworks as well as in styles in the figurative arts and in poetry.

Epistemological Anarchism

Feyerabend's epistemological anarchism (or Dadaism) was first advanced in 1970. Its central thesis is that methodological rules have only limited validity and application in science or, more generally, that science is governed by a liberal practice in which methodological rules are, at best, used as rules of thumb. This thesis is explicitly cast both in a descriptive and normative form: For one, historically, "[w]e find […] that there is not a single rule, however plausible, and however firmly grounded in epistemology, that is not violated at some time or other" (Feyerabend, 1993: p. 14). Second, "[t]his liberal practice […] is not just a fact of the history of science. It is not merely a manifestation of human inconstancy and ignorance. It is reasonable and absolutely necessary for the growth of knowledge" (Feyerabend, 1993: p. 14). A third, often ignored, but not less important motive is the ethical grounding of his proposal: "Such an anarchistic epistemology […] is not only a better means for improving knowledge, or of understanding history. It is also more appropriate for a free man to use than are its rigorous and 'scientific' alternatives" (Feyerabend, 1970b: p. 21).

Feyerabend supports his thesis with a detailed study of the Scientific Revolution and the rise of Copernicanism in the sixteenth and seventeenth century. The study purports to show that Galilei's defense of the heliocentric hypothesis violated basic tenets of modern scientific methodology and that it followed through exactly because of this violation. Feyerabend also purports to show that the opposing party, personified by Cardinal Bellarmine, argued against the heliocentric hypothesis on the basis of sound empiricist tenets. Concurrently, and even more controversially, the Church is interpreted as recommending the evaluation of the view of experts in the light of social and ethical values, whereas Galileo is seen as advocating *ante litteram* the acceptance of a privileged role of experts in society. The conclusion Feyerabend draws is twofold: Galilei could not have successfully defended the heliocentric hypothesis at the time without violating basic tenets of empiricism; and the opposition to Galilei, the Roman Catholic Church in particular, had good scientific and ethical reasons on their side to reject Galilei's scientific venture. Feyerabend's choice to use Galilei as a case study is crucial for his argument. If Galilei, the epitome of the modern scientist fighting against obscurantism, did not act according to the standard scientific methodology, who could possibly have? Because the Scientific Revolution has been viewed as the paradigmatic example of scientific progress, a revisionist historical–philosophical account, if correct, casts doubt on the usefulness of conventional accounts of scientific progress. In this way Feyerabend wants to generalize from a single but paradigmatic case to the scientific enterprise as a whole.

Because of its unabashed provocative claims, several charges have been leveled against Feyerabend's theses. Some have been unanimously identified as relying on misconceptions or misunderstandings of his work; other have been substantiated. For one, Feyerabend's conclusion has been read as against Galilei's scientific achievement or, alternatively, against the notion that Galilei's defense of heliocentrism was justified. Quite the contrary, Feyerabend's goal is to uphold Galilei's scientific achievement and to show the inadequacy of standard accounts of scientific rationality. Second, Feyerabend's intention has been interpreted as proposing a new anarchistic methodology to replace the empiricist one. This is arguably not the case: "My intention is not to replace one set of general rules by another such set: my intention is, rather, to convince the reader that *all methodologies, even the most obvious ones, have their limits*. The best way to show this is to demonstrate the limits and even the irrationality of some rules which she, or he, is likely to regard as basic" (Feyerabend, 1970b: p. 21).

The latter misconception arose because Feyerabend did not just show cases of basic empiricist rules being violated, but he affirmatively formulated and defended the use of increasingly heterodox methodological rules in science. Starting with the proliferation principle, later complemented by the tenacity principle, Feyerabend liberalized methodology, leading to the formulation of counterinduction as an additional rule of thumb stating that it may be useful to introduce hypotheses that are inconsistent not just with established *theories* – as the proliferation principle suggests – but with well-established *facts*. But counterinduction is not meant to be a new methodology. Indeed, Feyerabend's analysis of Galilei's scientific practice purports to show that Galilei proceeded counterinductively on several occasions. And it is not introduced to supplant existing methodological rules: proceeding counterinductively is suggested "*in addition* to proceeding inductively" (Feyerabend, 1970b: p. 26).

The catch phrase associated with epistemological anarchism is 'anything goes.' Though it appears only once in the book, it has been used ever since as purportedly expressing the main principle of Feyerabend's epistemological anarchism. This is arguably mistaken. Feyerabend introduces 'anything goes' as a rhetorical device, distilling the only principle an advocate of methodological monism could possibly uphold in the face of the pluralism to be found in science and its history. But his point is to show that such a principle would be empty, and thus implies to abandon monism and an invitation to accept pluralism – 'many things go' – not to adopt an empty methodological rule.

More substantive objections have been made to Feyerabend's historical interpretation of Galilei's undertaking and his argument with the Church. Since Feyerabend did not produce new historical material, these objections pertained Feyerabend's handling of the available historical material and the cogency of the conclusions he drew from it (Machamer, 1973; McMullin, 1970). Whether this invalidates Feyerabend pluralist conclusion as well is open to debate.

While epistemological anarchism has been viewed as a radical departure from his earlier views, the actual change is rather subtle, at least when compared to the strong consequences Feyerabend drew from it. The increasing skepticism toward general methodological rules was prompted by the raising awareness that philosophical issues in the sciences may at times not be resolved by appealing to general methodological rules, but be forced upon by the circumstances of specific research contexts. Nevertheless, methodological rules do not become useless in Feyerabend's analysis, they merely loose generic means of justification and become subject to context-dependent justification. Feyerabend's methodological pluralism embraces the different contexts of scientific research, analyzing which epistemic values were adopted and why they were adopted in given contexts.

Philosophy of Nature and the Rise of Rationalism

Against Method's main claim that the rationality of science is incompatible with its 'rational reconstruction' leaves open at least two options. The failure of standard accounts of rationality to account for science and its progress "means either that there cannot be any discoveries unless one leaves the house of reason, or that the house of reason is very different from what philosophers and other idea-mongers make it out to be" (Feyerabend, 2011: p. 112). The success of *Against Method* and its vitriolic rhetoric lent credibility to the idea that Feyerabend went with the first option, instead of hinting at his true intentions of giving a detailed account in which way(s) 'the house of reason' might differ from formal philosophical accounts, a project which was published only posthumously (Feyerabend, 2009).

In his final writings, he relented the acrimony of his earlier works, giving way to an unsystematic resumption of his project to describe 'the rise of rationalism' and to give an alternative account of scientific and nonscientific rationality and the relationship between the two (Feyerabend, 2011, 1999). In these works, he invites the reader to rethink the relationship between the empirical and theoretical dimensions of human agency, science being one such activity among many. While an abstract notion of knowledge, namely theoretical knowledge, is necessary to science, it is not sufficient, as science only succeeds embedded in specific practices. Indeed, his claim is that the theoretical dimension is but a kind of particular practice. Science is, therefore, best understood as a theoretical and practical craft.

Democratic Relativism

Feyerabend's strong ethical commitment and his considerations about the close relationship between morality and knowledge, between ethics and epistemology, are documented long before his more controversial pronouncements in *Science in a Free Society* (see Feyerabend, 1961). Yet they turned into more global considerations about science and society only with the advent of *Against Method*, borrowing from political philosopher John Stuart Mill and other thinkers. In particular, he applied Mill's insights in *On Liberty* (1859) to extend theoretical pluralism beyond science to other forms of knowledge, turning theoretical pluralism into global considerations about the necessity of a plurality of points of view and of traditions and the required conditions under which these points of view could develop and thrive (Feyerabend, 1987, 1978).

Feyerabend is best known for his positive appraisal of several social and philosophical theses which he subsumed

under the heading of relativism. 'Relativism' has long been used as a negative epithet in philosophy, rarely as a philosopher's positive self-description. (For a recent philosophical reappraisal of Feyerabend's relativism see Kusch, forthcoming.) Not so in fields like anthropology and sociology of knowledge, where relativism has a long standing pedigree as a central methodological premise. Likewise, his target audience is clearly not the philosophical community anymore, but a more general learned audience.

At the core of Feyerabend's relativism lies the ideal of a free society (a terminological choice polemically opposed to Karl Popper's open society) in which all traditions ought to enjoy equal rights and opportunities. (This he calls 'democratic' or 'political' relativism.) Feyerabend's central move is twofold: To conceptualize science as one tradition among many and to renounce even a moderate form of scientism, the view that science gives us better means to attain knowledge than other traditions, which lies at the basis of the scientific worldview. Yet this does not imply that all traditions' beliefs are equally true or false (a view which Feyerabend calls 'philosophical relativism' and rejects). His main target is the predominant *political* and *social* role of science within today's societies. While science is the only one among many traditions, the scientific worldview stipulates that science mediates and is introduced as an arbiter into the discourse among traditions, i.e., has a privileged position in collective social–political decision-making process. According to democratic relativism, this privileged position is incompatible with the flourishing of a plurality of traditions. Also, science is given the power to stipulate the very rules that govern the interchange between traditions. To exemplify his proposal of a free society and the role of experts therein, Feyerabend draws an analogy to courts of law in which citizens are judged by a jury of peers (Feyerabend, 1978: p. 97). While experts may be called to testify, they do not necessarily have an exceptional standing and are one factor among many leading to the final verdict that is deliberated and agreed upon by a group of peers. In a similar way, a free society urges that citizens should have the final say in social–political decision-making.

See also: Conventions and Norms: Philosophical Aspects; Critical Rationalism; Democracy: Normative Theory; Knowledge (Explicit, Implicit and Tacit): Philosophical Aspects; Kuhn, Thomas S. (1922–96); Liberalism: Historical Aspects; Logical Positivism and Logical Empiricism; Mill, John Stuart (1806–73); Objectivity: Philosophical Aspects; Physicalism and Alternatives; Postmodernism: Philosophical Aspects; Relativism: Philosophical Aspects; Science, History of; Social Constructivism; Vienna Circle: Logical Empiricism; Wittgenstein, Ludwig (1889–1951).

Bibliography

Cartwright, N., 1999. The Dappled World: A Study of the Boundaries of Science. Cambridge University Press, Cambridge, UK; New York.

Collodel, M. Was Feyerabend a Popperian? Methodological issues in the history of the philosophy of science. Studies in History and Philosophy of Science, Special Issue: Reappraising Feyerabend, forthcoming.

Dupré, J., 1995. The Disorder of Things: Metaphysical Foundations of the Disunity of Science. Harvard University Press.

Farrell, R.P., 2001. Feyerabend's metaphysics: process-realism, or voluntarist-idealism? Journal for General Philosophy of Science 32, 351–369.

Farrell, R.P., 2003. Feyerabend and Scientific Values: Tightrope-Walking Rationality, Boston Studies in the Philosophy of Science. Kluwer, Dordrecht.

Feyerabend, P., 2009. Naturphilosophie. Suhrkamp Verlag.

Feyerabend, P.K., 1951. Zur Theorie der Basissätze. Dissertation zur Erlangung der Doktorgrades an der philosophischen Fakultät der Universität Wien. Universität Wien, Wien.

Feyerabend, P.K., 1958. An attempt at a realistic interpretation of experience. Proceedings of the Aristotelian Society 58, 143–170.

Feyerabend, P.K., 1961. Knowledge without Foundations: Two Lectures Delivered on the Nellie Heldt Lecture Fund. Oberlin College, Oberlin.

Feyerabend, P.K., 1962. Explanation, reduction, and empiricism. In: Feigl, H. (Ed.), Scientific Explanation, Space, and Time, Minnesota Studies in the Philosophy of Science. University of Minnesota Press, Minneapolis, pp. 28–97.

Feyerabend, P.K., 1963. How to be a good empiricist: a plea for tolerance in matters epistemological. In: Baumrin, B. (Ed.), Philosophy of Science: The Delaware Seminar. Interscience Publishers, New York, pp. 3–39.

Feyerabend, P.K., 1965. Reply to criticism: Comments on Smart, Sellars and Putnam. In: Cohen, R.S., Wartofsky, M.W. (Eds.), Proceedings of the Boston Colloquium for the Philosophy of Science 1962–64: In Honor of Philipp Frank, Boston Studies in the Philosophy of Science. Humanities Press, New York, pp. 223–261.

Feyerabend, P.K., 1970a. Consolations for the specialist. In: Lakatos, I., Musgrave, A. (Eds.), Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965. Cambridge University Press, Cambridge, pp. 197–230.

Feyerabend, P.K., 1970b. Against method: outline of an anarchistic theory of knowledge. In: Radner, M., Winokur, S. (Eds.), Analyses of Theories and Methods of Physics and Psychology, Minnesota Studies in the Philosophy of Science. University of Minnesota Press, Minneapolis, pp. 17–130.

Feyerabend, P.K., 1978. Science in a Free Society. New Left Books, London.

Feyerabend, P.K., 1987. Farewell to Reason. Verso, London; New York.

Feyerabend, P.K., 1993. Against Method: Outline of an Anarchistic Theory of Knowledge, third ed. Verso, London.

Feyerabend, P.K., 1999. Conquest of Abundance: A Tale of Abstraction versus the Richness of Being. University of Chicago Press.

Feyerabend, P.K., 2011. The Tyranny of Science. Polity, Cambridge, UK.

Fricker, M., 2009. Epistemic Injustice: Power and the Ethics of Knowing. Oxford University Press, Oxford; New York.

Hoyningen-Huene, P., 2004. Three biographies: Kuhn, Feyerabend and incommensurability. In: Harris, R. (Ed.), Rhetoric and Incommensurability. Parlor Press, West Lafayette, IN, pp. 150–175.

Hoyningen-Huene, P., 2006. More letters by Paul Feyerabend to Thomas S. Kuhn on proto-structure. Studies in History and Philosophy of Science 37, 610–632.

Kidd, I.J., 2012. Feyerabend, Pseudo-Dionysius, and the ineffability of reality. Philosophia 40, 365–377.

Kitcher, P., 2003. Science, Truth, and Democracy, Hardcover ed. Oxford University Press, Oxford; New York.

Kitcher, P., 2011. Science in a Democratic Society. Prometheus Books, Amherst, NY.

Kusch, M. Relativism in Feyerabend's later writings. Studies in History and Philosophy of Science, Special Issue: Reappraising Feyerabend, forthcoming.

Machamer, P.K., 1973. Feyerabend and Galileo: the interaction of theories, and the reinterpretation of experience. Studies in History and Philosophy of Science Part A 4, 1–46.

McMullin, E., 1970. The history and philosophy of science: a taxonomy. Minnesota Studies in the Philosophy of Science 5, 12–67.

Munévar, G. (Ed.), 1991. Beyond Reason: Essays on the Philosophy of Paul Feyerabend, Boston Studies in the Philosophy of Science. Kluwer Academic Publishers, Dordrecht; Boston.

Oberheim, E., 2006. Feyerabend's Philosophy. Walter de Gruyter, Berlin.

Preston, J., 1997. Feyerabend: Philosophy, Science, and Society. Polity Press, Cambridge.

Preston, J., 1998. Science as supermarket: Post-modern themes in Paul Feyerabend's later philosophy of science. Studies in History and Philosophy of Science Part A 29, 425–447.

Stadler, F., 2008. Paul Feyerabend and the forgotten "Third Vienna Circle". In: Scazzieri, R., Simili, R. (Eds.), The Migration of Ideas. Watson Publishing International, MA, pp. 203–224.

Stove, D.C., 2006. Scientific Irrationalism: Origins of a Postmodern Cult. Transaction Publishers.

Relevant Website

http://www.collodel.org/feyerabend/ – The Works of Paul K. Feyerabend by Matteo Collodel (accessed 11.07.14.).

Fiducial and Structural Statistical Inference

Donald AS Fraser, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 8, pp. 5616–5620, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Fiducial inference provides probabilities for parameters on the basis of a statistical model and data; structural inference also provides such probabilities but requires a specialized transformation structure in the model. A comparison of the fiducial and structural methodologies with default Bayesian analysis is given and shows that both can be viewed as specializations of the Bayesian. Examples are given and comparisons made.

Fiducial inference is a statistical approach to interval estimation first advocated by R. A. Fisher as an alternative to the then dominant method of inverse probability, i.e., using Bayes' Theorem. Considerable effort has gone into formalizing Fisher's notions using such concepts as statistical invariance and pivotal quantities. This article describes elements of the fiducial approach and relates them to other currently more widely used statistical approaches to inference. Section A Simple Example introduces some basic inferential ideas via a simple example.

A Simple Example

Consider a variable y that is directly available or has arisen by some preliminary reduction process and suppose that y measures θ in an unbiased manner and has error that is normal with known variance σ_0^2. Then we can say for example that $P(y \leq \theta + 1.64\sigma_0;\ \theta) = 95\ \%$.

Parenthetically, we note that the preliminary reduction could have occurred as part of the underlying investigation or as part of some subsequent simplification of the statistical model by one of the common statistical reduction methods, sufficiency or conditionality. If the reduction is to a sufficient statistic, then the conditional distribution describing possible antecedent data has no dependence on the parameter and the model for the sufficient statistic is used in place of the original model. If the reduction is by conditionality then there is typically an ancillary variable with a distribution free of the parameter and the given model for the possible original data is replaced by the conditional model given the ancillary (supportive) variable; *see also* Statistical Sufficiency.

With a data value y^0 the fiducial methodology would take the above probability expression and substitute the value y^0 and then treat θ as the variable for the probability statement, thus giving a 95% fiducial probability statement $P(\theta \geq y^0 - 1.64\sigma_0;\ y^0) = 95\%$. The structural approach would consider the normal $(0;\ \sigma_0^2)$ distribution for the error $e = y - \theta$ and might record for example the probability $P(e \leq 1.64\sigma_0) = 95\%$; then with data value y^0 the probability statement would be applied to the error $e = y^0 - \theta$ giving the structural probability $P(y^0 - \theta \leq 1.64\sigma_0;\ y^0) = 95\%$ or equivalently $P(\theta \geq y^0 - 1.64\sigma_0;\ y^0) = 95\%$. In a somewhat related manner the Bayesian methodology might use a uniform prior $cd\theta$ and obtain a posterior distribution for θ that would have $P(\theta \geq y^0 - 1.64\sigma_0;\ y^0) = 95\%$. For this simple example the three methods give the same result at the 95 percent level and also at other levels, thus saying essentially that with data value y^0 the probability distribution describing the unknown θ is normal $(y^0; \sigma_0^2)$.

With more complicated models the results from the three methodologies can differ and philosophical arguments concerning substance and relative merits arise. However, for one straightforward generalization the methods remain in agreement: the normal distribution can be replaced by some alternative distribution form; this is discussed in some detail in Sect. 2.

Fiducial Probability

Fisher (1922, 1925; see also Fisher, Ronald A. (1890–1962)) had already introduced most of the fundamental concepts of statistical theory, such as sufficiency, likelihood, efficiency, exhaustiveness (minimal sufficiency), when he chose (Fisher, 1930) to address directly the aspiration mentioned above. He took Laplace and Gauss to task for 'fall(ing) into error on a question of prime theoretical importance' by adopting the Bayesian approach that 'Bayes (had) tentatively wished to postulate in a special case' and which was published posthumously (Bayes, ibid). Be then proposed in a restricted context the fiducial method, as discussed.

Neyman and Pearson (1933) then gave a mathematical formulation of fiducial probability that became known as confidence intervals. Fisher (1956) however treated Neyman and Pearson's formulation as a 'misconception having some troubling consequences …'; logical and philosophical arguments between the two sides were intense for many years. In particular the slight to Laplace and Gauss may well have affected the views of the more mathematical participants.

Fisher (1930) entitled his paper 'Inverse Probability' and examined a statistic $t(y)$ whose distribution depended on a single parameter θ. Let $P = F(t;\theta)$ be the distribution function of t, and let P itself be what we might now call a p-value for assessing θ; of course in the usual continuous case P has the uniform distribution on (0, 1). 'If now we give P any particular

value such as 0.95, we have … the perfectly objective fact that in 5% of samples' t 'will exceed the 95% value corresponding to the actual value of θ ….' Then to 'any value of' t 'there will moreover be usually a particular value of θ to which it bears this relationship; we may call this the 'fiducial 5% value of θ' corresponding to' the given t. This led to Neyman and Pearson's (ibid) confidence methodology but Fisher treated this as a misconception and he followed different directions and interpretations for the fiducial methodology; for a view on related approaches see Estimation: Point and Interval. For this present simple case with scalar t and scalar θ, there seems little difference between the fiducial and the confidence approaches and interpretations.

This discussion effectively ascribes a distribution to θ based on an observed t; this is called the fiducial distribution for θ. Just as the density of t for given θ is obtained as $(\partial/\partial t)$ $F(t;\ \theta)$ so also the fiducial density is obtained as $(-\partial/\partial\theta)$ $F(t;\ \theta)$; the negative sign is inconsequential and is merely the result of $F(t;\ \theta)$ being examined typically for the case that is increasing with θ. The fiducial distribution for cases like the present can also be called the confidence distribution, which recently has a variety of uses in statistical inference. This alternative name emphasizes its more conventional role and attempts to avoid the inappropriate stigma that surrounds the fiducial concept.

For the fiducial approach in more general contexts, Fisher (1935) recommended a maximum reduction to a statistic $t(y)$ by the use of sufficiency. For example, with a sample $y_1,\ldots,y_n$ from the normal $(\mu;\sigma_0^2)$ distribution, the reduction would be to $t(y) = (\bar{y}, s_2^y)$. The methodology then suggests the use of a pivotal quantity $p = p(t;\ \theta)$ with a fixed distribution and a one-one relationship between any two of p, t, θ; recall more generally that a pivotal quantity is a function of the variable and parameter that provides a measure of departure of variable value from parameter value, and has a fixed distribution which allows an assessment of an observed departure. Thus for the example a natural pivotal is

$$\left\{z = \frac{\bar{y}-\mu}{\sigma/\sqrt{n}}, \chi^2_{n-1} = \frac{(n-1)s_Y^2}{\sigma^2}\right\}$$

which has independent components, normal (0, 1) and chi-square with $n-1$ degrees of freedom. Fisher (1956, p. 172), however, rather deviously rejected this as a legitimate part of the fiducial methodology, but he was somewhat less explicit about what would be legitimate. This pivotal reduction procedure is now however a rather familiar component of standard inference theory and in particular of confidence theory.

The final step is to invert the pivotal quantity, that is, to insert the observed values for the variables and then transform the distribution of the pivotal quantity to the parameter. For the example this gives

$$\mu = \bar{y} - \left\{z/(n-1)^{-1/2}\chi_{n-1}\right\}s_Y/\sqrt{n}$$
$$\sigma^2 = (n-1)s_Y^{22}/\chi^2_{n-1}$$

where the $\bar{y}, s_Y^2$ have their observed values. We can then write

$$\mu = \bar{y} - ts_Y/\sqrt{n}$$

where t is Student with $n-1$ degrees of freedom. This fiducial calculation closely parallels that for the confidence approach, except that the limits here are calculated from the Student distribution for μ rather than from the Student distribution of the pivotal t. In quite wide generality fiducial regions can correspond to confidence regions; it is just a matter of whether the limits are calculated before or after the data are observed, a non-issue from the Bayesian viewpoint.

There are however cases where routinely obtained confidence and fiducial regions can differ. For example, consider a sample $(y_{11},y_{12}),\ldots,(y_{n1},y_{n2})$ from a bivariate normal distribution with means μ_1, μ_2, variances σ_1^2, σ_2^2, and correlation ρ, and suppose primary interest centers on the regression coefficient $\beta = p\sigma_2/\sigma_1$ of the second variable on the first variable. Standard regression analysis of the second variable on the first variable can produce an interval that differs from that obtained from the joint fiducial or structural distribution for all the parameters by integrating out the unwanted parameters; see for example, Fraser (1979, pp. 189, 204, 280, and 293). Also there are other familiar cases where fiducial, structural and Bayesian methods can differ; see for example, the Behrens-Fisher and the Fieller-Creasy problems, and for some discussion see Wallace (1980).

Now suppose we generalize and examine a sample $y_1,\ldots,y_n$ from some distribution with location μ and scale σ. For example the basic form $f(z)$ of the distribution might be, say, the longer-tailed logistic, or the Student with 6 degrees of freedom often cited as a realistic error pattern. The distribution for the underlying errors is then $f(z_1)\ldots f(z_n)$ and for the sample is

$$\sigma^{-n}\prod f\{\sigma^{-1}(y_i-\mu)\}$$

With a typical general error shape we do not have a simple minimal sufficient statistic as with the normal error pattern above. Fisher (1934) then recommended conditioning on an ancillary statistic and suggested the statistic

$$d_1 = \frac{y_1-\hat{\mu}}{\hat{\sigma}}, \ldots, d_n = \frac{y_n-\hat{\mu}}{\hat{\sigma}}$$

called the configuration statistic, whose coordinates are standardized residuals, standardized here with respect to maximum likelihood values. The conditional distribution of $(\hat{\mu}, \hat{\sigma})$ is fairly easily expressed in terms of likelihood. Let

$$L^0(\mu,\sigma) = c\sigma^{-n}\prod f\{\sigma^{-1}(y_i^0-\mu)\}$$

be the observed likelihood function from data $(y_1^0,\ldots,y_n^0)$, and $(\hat{\mu}^0, \hat{\sigma}^0)$ be the observed maximum likelihood values. Then the conditional distribution given the observed configuration statistic $(d_1^0,\ldots,d_n^0)$ is

$$g(\hat{\mu},\hat{\sigma}|d^0;\mu,\sigma)\frac{d\hat{\mu}d\hat{\sigma}}{\hat{\sigma}^2} = cL^0(\hat{\mu}^0 + (\mu-\hat{\mu})\hat{\sigma}^0/\hat{\sigma}, \sigma\hat{\sigma}^0/\hat{\sigma})\frac{d\hat{\mu}d\hat{\sigma}}{\hat{\sigma}^2}$$

which can be used to give confidence intervals; also a straightforward fiducial argument following that for the normal case gives

$$L^0(\mu,\sigma)\frac{d\mu d\sigma}{\sigma}$$

for the fiducial distribution. If f is replaced by the standard normal we get the result discussed earlier and we get it without

any reference to sufficiency. This with some other results suggests (Fraser and Reid, 2000) that the primary concept sufficiency can be replaced quite widely by appropriate conditioning, and with a major gain in generality of viewpoint.

Fisher (1935) did not use maximum likelihood notation for this location scale model. The notation here adapts to more general contexts (Fraser, 1968; 1979) but needs to be modified if the error distribution involves a shape parameter which would then make the maximum likelihood estimates parameter-dependent.

For more general problems the fiducial can have non-uniqueness difficulties, and these tend to arise from the choice of what pivotal to use. For example with the bivariate normal discussed, different pivotals can produce different fiducial distributions (Fisher, 1956, p. 172, where he elusively denigrates the pivotal approach).

There are also issues connected with finding a marginal fiducial for a component parameter. For a general discussion of marginalization paradoxes connected with methodologies concerning distributions for parameters, including the Bayesian methodology, see David et al. (1973).

For a mathematical analysis that avoids some of the difficulties with the fiducial methodology, see the transformation group framework in Fraser (1961a,b); this mathematical analysis led to the development of the structural approach described below. The marginalization issues still remain, however, and this is related to the common assessment technique of examining a procedure with many repetitions from the same distribution with fixed parameter value. Alternatives to this are suggested in Fraser and Reid (2000).

Structural Probability

In many applied problems it is possible to ascribe an underlying error distribution to the model. For example with measurements that are normal with unknown error scaling we might write

$$y_i = \mu + \sigma z_i \quad i = 1, \ldots, n$$

where the $z_1,\ldots,z_n$, are independent standard normal variables. If we then examine (Fraser, 1966) these expressions with observed data we can note rather easily that many characteristics of the underlying realized errors can be numerically evaluated. Specifically, let

$$d_1 = \frac{y_1 - \bar{y}}{s_Y}, \ldots, d_n = \frac{y_n - \bar{y}}{s_Y}$$

be standardized residuals; then the observed $y_1,\ldots,y_n$ and the underlying realized $(z_1,\ldots,z_n)$ have the same standardized residuals. In fact, $n-2$ characteristics as presented by the residuals are available concerning the underlying errors. Most theory concerning applications of probability would then say the analysis should be conditional on these observed characteristics, that is, on what you know concerning the realized error (Fraser, 1976, Chap. 4, pp. 161–2 and Chap. 11, pp. 456–66). This for quite general problems automatically gives the conditioning used by Fisher and avoids the need to invoke a conditioning principle and seek an ancillary such as the configuration statistic.

This present type of analysis carries through equally (Fraser, 1966) if the normal distribution is replaced by some other error pattern. The conditional error distribution can be viewed from an invariance argument as being valid after the full data are observed. This distribution then gives the structural distribution for μ and σ. The mechanics follow closely what is presented for the fiducial case and the resulting structural distribution agrees with the fiducial distribution, $L^0(\mu,\sigma)d\mu d\sigma/\sigma$. The difference here is that a rather strict invariance structure is assumed as part of the argument. This is summarized in general notation.

The structural approach is possible in cases where the underlying error can be directly modeled and the observed response is obtained as some transformation of this. Let Z designate a general error variable with known density $f(Z)$ on say R^n. Let Y designate the response as obtained by a transformation or reexpression θ in some space Ω. Then we have

$$Y = \theta Z, f(Z)dZ,$$

For the invariance we assume that a transformation θ carries R^n into R^n and that the collection Ω of such transformations is closed under composition and inversion. Also for details here we assume that the transformations are smooth and at most one transformation carries a point Z into a point Y. For some general background on transformations and invariance, *see* Causal Inference and Statistical Fallacies.

If we consider all the transformation Ω and apply them to a data point Y^0 we obtain a set ΩY^0 of images, of possible antecedent error values. If Z^0 is the unknown realization of error that produced Y^0 then we have $\Omega Z^0 = \Omega Y^0$ or equivalently we have that Z^0 lies in the set ΩY_0. This means that a probability assessment should be made conditional on the observed set ΩY^0. Let $D(Z)$ be a reference point in the set ΩZ and let $[Z]$ in Ω be the transformation in Ω that carries $D(Z)$ to Z. Then we have that the model $Y = \theta Z$ with data Y^0 can be rewritten as

$$[Y] = \theta[Z], \quad D(Y) = D^0,$$

giving the conditioning $D(Y) = D(Y) = D^0$ that appears as an observed ancillary in the fiducial framework. The conditional distribution of $[Z]$ given $D(Z) = D^0$ then describes the reduced model $[Y] = \theta[Z]$ with data $[Y^0]$. The structural distribution of θ is then given as $\theta = [Y^0][Z]^{-1}$ using the conditional distribution $[Z]$ just described.

In most contexts the support measure dZ can be replaced by an invariant measure $dM(Z)$

$$dM(Z) = \frac{dZ}{J(Z)}$$

where $J(Z)$ is an appropriate Jacobian (Fraser, 1979). Expressions are simpler if we use density $f(Z)$ with respect to the measure $dM(Z) : \bar{f}(Z) = f(Z)J(Z)$.

The distributions of $[Y]$ and $[Z]$ conditional on $D(Y) = D(Z) = D^0$ are

$$c\bar{f}(\theta^{-1}[Y]D^0)d\mu[Y], \quad c\bar{f}([Z]D^0)d\mu[Z]$$

where $d\mu[Z]$ is the corresponding invariant measure on the group Ω. The inverted distribution, called the structural distribution for θ, is

$$c\bar{f}(\theta^{-1}[Y^0]D^0)d\nu(\theta) = cL^0(\theta)d\nu(\theta)$$

using the right invariant measure ν on the group Ω. In the Bayesian context the typically preferred noninformative prior is the right invariant measure $d\nu(\theta)$. Thus the structural distribution coincides with this preferred Bayesian posterior.

This group type model covers a wide range of regression multivariate and spherical distribution models but not in the simplified notation used here; for further details see Fraser (1968, 1979).

Recent likelihood asymptotics has been able to obtain many aspects of these results in a general asymptotic context; see Fraser et al., (1999a,b).

Comment

Fiducial inference led to confidence intervals, with both having evolved from the Bayesian procedures initiated in the eighteenth century. In terms of freedom of usage, we have Bayesian inference as the more liberal. The foundational issues center on validity and interpretation of the probabilities for the parameters. Here the conditions are strictest for the structural and minimal for the Bayesian. In the nice cases they agree; and in the less standard cases they pressure the proponents to clarify their assumptions.

For some recent discussion of confidence and fiducial interconnections plus references see Barnard (1995), with complementing views in Fraser (1996). Also see Fraser and Reid (2000) for a discussion of the interplay with objective Bayesian methods.

Bibliography

Barnard, G.A., 1995. Pivotal models and the fiducial argument. International Statistical Review 63, 309–324.

Bayes, T., 1763. An essay towards solving a problem in the doctrine of chance. Philosophical Transactions of the Royal Society 53, 370–418 [Reprinted (1958) in Biometrika 45: 293–315].

Dawid, A.P., Stone, M., Zidek, J.V., 1973. Journal of the Royal Statistical Society 35, 189–233.

Fisher, R.A., 1922. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London A222, 309–368.

Fisher, R.A., 1925. Theory of statistical estimation. Proceedings of the Cambridge Philosophical Society 22, 700–725.

Fisher, R.A., 1930. Inverse probability. Proceedings of the Cambridge Philosophical Society 26, 528–535.

Fisher, R.A., 1934. Two new properties of maximum likelihood. Proceedings of the Royal Society of London A144, 285–307.

Fisher, R.A., 1935. The fiducial argument in statistical inference. Annals Eugenics 6 (4), 391–398.

Fisher, R.A., 1956. Statistical Methods and Scientific Inference. Oliver and Boyd, Edinburgh, UK.

Fraser, D.A.S., Reid, N., 2000. Strong matching of frequentist and Bayesian inference. Journal of Statistical Planning and inference, to appear.

Fraser, D.A.S., Reid, N., Wu, J., 1999a. A simple general formula for tail probabilities for frequentist and Bayesian inference. Biometrica 86, 249–264.

Fraser, D.A.S., Wong, A., Wu, J., 1999b. Regression analysis, nonlinear or nonnormal simple and accurate p values for likelihood analysis. Journal of the American Statistical Association 94, 1286–1295.

Fraser, D.A.S., 1961a. The fiducial method and invariance. Biometrika 48, 261–280.

Fraser, D.A.S., 1961b. On fiducial inference. Ann. Math. Statist 32, 661–676.

Fraser, D.A.S., 1966. Structural probabilities and a generalization. Biometrica 53, 1–9.

Fraser, D.A.S., 1968. The Structure of Inference. Wiley, New York.

Fraser, D.A.S., 1976. Probability and Statistics, Theory and Applications. ITS, Toronto.

Fraser, D.A.S., 1979. Inference and Linear Models. McGraw Hill, New York.

Fraser, D.A.S., 1996. Some remarks on pivotal models and the financial argument in relation to structural models. International Statistical Review 64, 23–25.

Hannig, J.L., Thomas, C.M., 2009. Generalized Fiducial Inference for Wavelet Regression. Biometrika 96 (4), 847–860.

Hannig, J., Iyer, H., Patterson, P., 2006. Fiducial Generalized Confidence Intervals. Journal of the American Statistical Association 101 (473), 254–269.

Hannig, J., 2009. On Generalized Fiducial Inference. Statistica Sinica 19 (2), 491.

Neyman, J., Pearson, E.S., 1933. The testing of statistical hypotheses in relation to probabilities a priori. Proceedings of the Cambridge Philosophical Society 29, 492–510.

Rubin, D.B., 2005. Causal Inference Using Potential Outcomes. Journal of the American Statistical Association 100 (469).

Wallace, D.L., 1980. The Behrers–Fisher and Creasy–Fieller problems. In: Fienberg, S.E., Henlele, D.V. (Eds.), R. A. Fisher: An Appreciation. Springer Verlag, New York, pp. 139–145.

Field Experiments

Ruth Ditlmann, WZB Berlin Social Science Center, Berlin, Germany
Elizabeth Levy Paluck, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Field experiments are experiments in settings with high degrees of naturalism. This article describes different types of field experiments, including randomized field trials, randomized rollout designs, encouragement designs, downstream field experiments, hybrid lab-field experiments, and covert population experiments, and discusses their intellectual background and benefits. It also lists methodological challenges researchers can encounter when conducting field experiments, including failure to treat, selective attrition, spillover, difficulty of replication, and black box causality, and discusses available solutions. Finally, it provides an overview over current and emerging directions in field experimentation and concludes with a brief history of field experiments.

Definitions

Field experiments are experiments in settings with high degrees of naturalism. *Experiments* are studies using some type of random procedure, such as a coin flip or a random number generator, which determines for each participant whether they receive a treatment versus no treatment or a comparison treatment. *Random assignment* in expectation ensures balance on observable and unobservable confounding variables between the treatment groups and thus allows for causal claims about the effect of the treatment on the outcomes of the group. In other words, through random assignment the researcher can causally attribute recorded outcomes to treatments or manipulations. For example, assume that participants are randomly assigned to two different weight loss programs, A and B. If, at the completion of the two programs, the average weight of participants in program A is significantly less than participants in program B according to a standard statistical test of difference in means, the researcher can infer that program A is more effective at causing weight loss compared to program B. In the absence of random assignment, it is possible that people with lower body fat, a higher motivation to lose weight, or another unobservable characteristic disproportionately participated in program A due to the biased choices of participants or the experimenter. These characteristics could explain the difference in weight loss that has nothing to do with the program itself.

In the weight loss example, individuals are the *experimental units* that are randomly assigned to different treatments. Experimental units may be people but can also be other entities, such as communities, households, or school classes. When groups of people, such as groups of friends, rather than individuals, are assigned to a treatment, this is referred to as *clustered random assignment*. *Blocked random assignment* occurs when experimental units are first grouped into blocks according to similarities in any particular observed characteristic, and then randomly assigned to treatment within each block. For example, researchers can first block individuals into six groups defined by previous academic achievement and sociodemographic background, and then randomly assign individuals within each block to an experimental tutoring intervention versus a no-tutoring control. In field experiments, random assignment usually occurs *between subjects*: each participant receives only one of several treatments. For example, researchers may randomly assign a sample of 300 companies to receive one of two resumes that are identical except that one is from a black applicant (treatment 1) and the other from a white applicant (treatment 2). The measured outcome is whether each company contacts the fictional applicant for an interview. In this case, each company only receives one resume, either from a white *or* from a black candidate. An alternative approach is a *within-subjects* design in which experimental units are exposed to multiple treatments. For example, researchers may send comparable resumes from both a white and a black applicant to the same company. An advantage of within-subjects designs is that they generally require smaller samples than between-subjects designs. A disadvantage is that it remains unclear if the outcome was caused by any single treatment or by the sequence of all available treatments. This requires at least a counterbalanced design in which half of the companies are randomly assigned to hear from a black then a white applicant, and the other half hear from a white and then a black applicant.

The term experiment has a precise definition: the term *field* is more open to interpretation. Generally, a field experiment is an experiment in a setting with a high degree of naturalism. Researchers can determine the *naturalism* of an experiment based on the following four considerations (Gerber and Green, 2012): (1) Does the treatment or manipulation in the study resemble the intervention of interest in the world? (2) Do participants in the study resemble the actors who usually encounter these interventions? (3) Does the context within which participants receive the treatment resemble the context of interest or is it, for example, more obtrusive? and (4) Does the outcome measure resembles the actual outcome of theoretical or practical interest? For example, a field experiment would test a real door-to-door campaign rather than a simulated doorstep conversation, would target citizens who are eligible to vote rather than a convenience sample of students in a class, would locate the door-to-door campaign in a neighborhood where it is meant to take place instead of a college campus, and would assess actual voting behavior directly rather than asking participants about their intention to vote in a survey.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10542-2

Conventional laboratory experiments with a nonnaturalistic intervention but with a nonstandard subject pool (i.e., members of the community who are the real-world targets of the simulated intervention) are sometimes called *artifactual field experiments*. An example for an artifactual field experiment recruited small business man of Turkish versus Belgium origin as research participants, and instructed them to play behavioral economics trust games (Bouckaert and Dhaene, 2004).

Experiments with a nonstandard subject pool and a naturalistic treatment or manipulation are sometimes called *framed field experiments*. In an example for this type of experiment, Love et al. (2005) randomly assigned low-income pregnant women and families with infants and toddlers to the 'early head start' child development program or to a control group. Early head start is an existing program that serves precisely these communities. Researchers gathered the outcomes of interest, child development, and parenting, mostly through interviews and test batteries, making participants aware of the research activities.

Quasi-experiments (see Shadish et al., 2002) where at least one treatment is not randomly assigned, as well as *natural experiments* (see Dunning, 2012) where random assignment occurs through an exogenous force that is not under the control of the researcher are distinct from field experiments and will not be covered in this encyclopedic article.

Intellectual Background

A considerable number of scholars across the social sciences advocate for field experiments. Which benefits they emphasize often depends on their disciplinary background. Pro-field experiment arguments made by social psychologists (Cialdini, 2009; Paluck and Cialdini, 2014) highlight the advantages of field experiments relative to laboratory experiments. In contrast, political scientists (Gerber et al., 2004; Gerber and Green, 2012) and sociologists (Pager, 2007) often focus on the advantages of field experiments relative to observational research. Economists stress both advantages when making the case for field experiments (Banerjee and Duflo, 2009).

The most frequent argument for field versus laboratory experiments is that *field experiments often have greater external validity than laboratory experiments*. Because treatments, participants, settings, and outcomes closely resemble those as they naturally occur in the world, lessons gained through field experiments may be more relevant to the social and political questions the experiment was designed to test.

In contrast to observational studies, field experiments can establish *causality* and provide *unbiased treatment effects* even in the absence of strong priors. Going beyond most laboratory experiments, field experiments can also *test the robustness of causal relations*. Because in their everyday realities people often encounter a myriad of stimuli and distractions, it is important to test the robustness of causal relations in such messy settings. For example, Gerber et al. (2008) found that high-social-pressure mailings that were randomly assigned to a subset of households increase voter turnout. These findings are remarkable considering how many steps lie in between receiving mail and casting one's ballot in a voting booth. In addition to providing useful knowledge about increasing voter turnout, the findings from this study powerfully demonstrate the strength and therefore theoretical importance of social pressure in civic behavior.

Although field experiments are often equated with program evaluation, the social pressure and voting study is one of many demonstrations that this method can be used to *advance scientific theory*. Convention in experimental social science had often assigned the role of advancing basic theoretical science to laboratory experiments, which were presumed to be free of confounds and unnecessary 'noise' or error variance. Field experiments, however, can produce highly accurate treatment effect estimates free of common laboratory artifacts such as experimenter bias and participant reactivity to the perceived demands of the experimental situation or instructions (see Adams and Stocks, 2008). For example, members of a dominant racial group often carefully monitor their behavior in laboratory experiments so as not to appear 'racist.' In field experiments, however, participants are more likely to be unaware of being observed, and real outcomes, for example, purchase transactions, are at stake. For instance, Ditlmann and Lagunes (2014) randomly assigned Latino and Anglo actors to pay for purchases with checks in retail stores, and, found that cashiers more frequently asked Latino actors to present ID when making a check purchase than Anglo actors. The fact that these cashiers were unaware that they were being studied allows us to be more confident that the experiment-captured average rates of discrimination as they occur in everyday life.

Finally, field experiments have a *great potential for innovation and discovery* because researchers do not control all elements of the setting, and thus can make unanticipated discoveries. For example, in a field experiment on the hiring process, Barron et al. (Barron et al., 2011) found that store personnel exhibited greater positivity and engaged in longer interactions when they were randomly assigned to minority applicants who displayed prominently their ethnic identification. This observation contradicts an earlier laboratory finding that high identification leads to negative evaluation of ethnic minorities (Kaiser and Pratt-Hyatt, 2009). The authors speculate that store owners fear discrimination lawsuits if they reject qualified minority candidates whom they suspect to be strong advocates for their groups' rights. The finding thus reveals the power of institutional structures in guiding human behavior – an important discovery for research inside and outside the laboratory.

Current Knowledge

Types of Studies

Field experimentation is a burgeoning field and new types of studies emerge regularly. This encyclopedic article describes some of the most commonly used types of field experiments and provides illustrative examples.

Randomized controlled field trials are the canonical design for field experiments. They compare the effects of two or more interventions or manipulations on outcomes of interest. Individuals from a pool of potential participants are randomly assigned to one or more interventions of interest vis-à-vis

a control or placebo condition. For example, Paluck (2009) randomly assigned Rwandan citizens, clustered in communities, to listen to a radio soap opera program either with prejudice reduction or with health messages. She found that the prejudice-reduction radio program improved norms for intergroup behavior, measured by the way listeners in the prejudice reduction program divided a community resource given to communities in each condition. In another randomized controlled trial, retail customers were offered one of four different pricing programs: 'Pay-what-you-want' or 'fixed-price,' each combined with information that either half of the money or no money would go to charity. Interestingly, participants paid the highest amounts in the program that offered the pay-what-you-want option combined with positive information about charity (Gneezy et al., 2010).

Randomized rollout designs (also called *waiting list designs* or *stepped wedge designs*) are similar to randomized controlled trials. Instead of being assigned to different treatments, all participants receive the same treatment but at randomly assigned varying points in time (time 1 or time 2). Researchers randomly assign half of all experimental units to the treatment at time 1, after which they can compare outcomes, and then researchers give control participants access to the treatment at time 2, after which they can test to see whether this group reacts similarly to the first group. This design is often used when there are ethical or political concerns about withholding the treatment from a portion of the sample. For example, researchers working with a bank in the Philippines offered an attractive savings plan to a randomly assigned selection of customers. Customers in a marketing treatment control condition received access to the plan after a 'product trial period,' i.e., after the research was completed (Ashraf et al., 2006).

Encouragement designs recognize that in some cases, researchers cannot guarantee participants' compliance with their assigned treatment. Instead, researchers randomly assign a subset of potential participants to an 'encouragement' to join the treatment intervention. For example, in the famous New York City school experiment, low-income children were randomly assigned to receive school vouchers, thus giving them the option to switch from public to private school (Peterson et al., 2003).

Downstream field experiments follow up on completed field experiments. They involve the collection of new data from participants in completed field experiments. Once random assignment is successful, researchers can assess the impact of the treatment on any number of outcomes at later points in time. For example, Sondheimer and Green (2010) gathered the voting turnout records of participants who were randomly assigned to participate in programs aimed at increasing high school graduation rates when they were children, to establish the causal effect of education on voting behavior.

Hybrid lab-field experiments are field experiments with elements of artificiality. This type of study is useful for researchers who want to have a high level of control over treatment administration or measurement, while simultaneously reaping some of the field experiment benefits discussed above. For example, Walton and Cohen (2007) brought black and white undergraduates into the laboratory for a brief intervention in which their fears about not belonging at the university were addressed or not. The success of this 'belongingness' intervention was demonstrated by measuring end of year grades that students earned in their college courses, which were higher for black and not white students in the treatment condition.

Covert population experiments use random assignment in a naturalistic setting to measure responses to members of subgroups within that setting that are identifiable by some sociodemographic (e.g., ethnic identity) or ideological (e.g., political affiliation) dimension. Importantly, participants who respond to the measurement paradigm are not aware that they are partaking in an experiment. As a result, covert population experiments often produce more accurate treatment effect estimates than more overt methods, such as self-report questionnaires. *Audit studies*, *correspondence studies*, and the *lost-letter paradigm* are three prominent examples for covert population experiments.

In *audit studies*, individual actors who vary on one or more important demographic dimensions are randomly assigned to engage in the same interpersonal transaction with a sample of people or institutions, for example, purchasing goods at a sample of stores or interviewing for an open position at a sample of companies. Because actors are trained to behave in an identical manner and because the experimental units (the interaction partners) are randomly assigned to one or the other actor, difference in responses is interpreted as evidence for discrimination. For example, in Ayres' (1991) famous 'fair driving' experiment, actors who varied in terms of their ethnicity and gender bargained for a retail car; car salespeople offered white men significantly better prices than white women and blacks.

In *correspondence studies*, researchers record responses to identical pieces of correspondence from (randomly assigned) 'senders' who vary only in one dimension of their identity, such as gender or age. In one such study, researchers randomly assigned identical resumes to be sent to prospective employers, sent from either Emily or Greg (whose names communicate a white racial identity) or from either Lakisha or Jamal (whose names communicate a black racial identity). The white applicants received 50% more callbacks for interviews than blacks (Bertrand and Mullainathan, 2004). Using the same randomized assignment of letters to a sample of recipients, state legislators were more likely to respond to requests for help with registering to vote if it came from a white rather than a black sender, even if the constituent shared partisanship with the legislator (Butler and Broockman, 2011).

In *Milgram's lost-letter paradigm*, researchers drop stamped and addressed but clearly unmailed letters in public places, in a randomized pattern. The addresses on the letters vary systematically to communicate demographic or ideological characteristics about the ostensible recipients. Researchers record how many letters reach the address (which typically belongs to the research team), or more specifically, whether community members display more helpfulness in picking up the letters and dropping them in the mailbox when the letter is addressed to one type of recipient versus another. In the original lost-letter study (Milgram et al., 1965), the ostensible recipients were "Friends of the communist party," "Friends of the Nazi party," "Medical research associates," or "Mr. Walter Carnap." Passersby picked up and mailed letters addressed to medical research associates and to the neutral

individual more frequently than letters addressed to either the Communist or Nazi party, between which there were no differences in helpful mailing. Recently, in an interesting adaptation of the original paradigm, Koopmans and Veit (2013) found that regardless of the letter recipient's ethnic identity, return rates of letters were lower in more ethnically diverse neighborhoods than in more homogenous neighborhoods.

Common Methodological Problems and Solutions

Several methodological problems require special attention from researchers who embark on field experiments. Most of these problems are not exclusive to field experiments, but afflict field experiments more frequently than laboratory experiments.

Failure to treat occurs when participants assigned to a treatment never receive the treatment. For example, during a get out the vote campaign not all individuals assigned to receive a specific canvassing message are at home. To handle failure to treat, researchers can identify what is called the *Complier Average Causal Effect* instead of the *Average Treatment Effect* and exercise caution when applying their findings to society as a whole (see Gerber and Green, 2012).

Selective attrition occurs when data are missing and attrition is systematically related to potential outcomes. For example, when in a study that compares the effect of two different health insurance plans on well-being, participants with poor health drop out at a greater rate from one of the plans (e.g., if one plan requires greater cost sharing and thus places a greater financial burden on individuals with more severe health problems). To address these problems, researcher can fill in potential missing values to estimate the largest and smallest possible treatment effects, thus placing bounds around the treatment effect (see Gerber and Green, 2012). Another option is gathering additional data from missing participants.

Spillover occurs when the treatment or treated participants influence untreated participants. For example, some participants in an intervention may discuss their experience with participants in the control group. To prevent spillover, researchers can space treatments out geographically or temporally, or assign clusters of interacting participants (e.g., networks of friends) rather than single individuals to treatment and control condition. In other cases, researchers may be interested in spillover effects as an outcome. Researchers may directly study the outcomes attributable to spillover by estimating average potential outcomes under different conditions of exposure to the treatment (Aronow and Samii, 2013).

Difficulty of replication is a final challenge for field experiments due to the way in which each study must be adapted to the particularities of the local setting. This is unfortunate because replication is an important tool to establish the robustness of effects in experimentation. Some researchers have replicated their experiments in different locations, and find close to identical results in some instances (Bobonis et al., 2006; Miguel and Kremer, 2004), while being unable to replicate the original effect in others (Banerjee et al., 2010; Duflo et al., 2007).

Limitations of Field Experiments

In addition to many benefits, field experiments also have several shortcomings. It is often difficult to obtain evidence for mediating processes in the field, especially when processes are believed to occur at a microlevel, involving individual cognition or emotion. To tackle this challenge of what some call *black box causality*, it can be fruitful to combine field and laboratory experimentation. Nonetheless, it is important to note that exploring mediation is a challenge for laboratory experiments as well (Bullock et al., 2010). A related problem is that manipulating and measuring precise and narrow constructs, such as distinct emotions like shame versus guilt, can be difficult in a field setting.

While black box causality and lack of precision are obstacles at the microlevel, researchers interested in the macrolevel struggle with a different problem: it is often difficult and sometimes impossible to randomize macrolevel variables, such as institutions or elites (e.g., the gender composition of a country's congress, or aid flow). While these barriers are serious, researchers have found creative ways to address both micro- and macrolevel variables. For a study that overcame the microlevel problem by assessing nonverbal displays of emotion in a field setting (see Ditlmann and Lagunes, 2014). Another study addressed the macrolevel problem of gaining insight into institutional corruption processes by studying requests for drivers' licenses in India (Bertrand et al., 2007).

Current and Emerging Directions

As field experiments become an increasingly popular methodology, several directions for future theory and research are emerging. *Collaborations* between researchers and governmental and nongovernmental organizations provide exciting opportunities for advancing science and conducting program evaluations at the same time (Banerjee and Duflo, 2009). International aid organizations have been sympathetic to field experiments for a while (Blum, 2011), whereas collaborations with the private sector are a more recent phenomenon (Levitt and List, 2009). For example, the Abdul Latif Jameel Poverty Action Lab promotes random assignment as an evaluation tool and fosters researcher–practitioner relationships.

Combining qualitative research and field experiments is another promising direction for future research. Qualitative research methods can illuminate processes of change and thus overcome the black box problem. Qualitative efforts can range from conducting qualitative interviews with selected participants in different treatment groups to *experimental ethnography* (Paluck, 2010; Sherman and Strang, 2004).

Comparative or contextual field experiments replicate the same experiment in two different contexts. Replications can test the robustness and often improve the precisions of average treatment effect estimates. It can also advance theory. For example, Carlsson and Rooth (2011) found that the effect of discrimination in a correspondence study was higher in municipalities with aggregate level more negative than national average attitudes as opposed to municipalities with aggregate level more positive than national average attitudes. These findings advance theory by demonstrating that attitudes are at least

one important process through which discrimination effects unfold.

Combining field experiments with social networks analysis takes advantage of spillover, originally identified as a methodological problem. This approach takes into account that individuals seldom act in isolation but are part of many, often highly complex networks. Combining field experiments and social networks shows how treatment effects spread throughout different types of networks. For example, Centola (2010) randomly manipulated the topology of a network in an online health community. He found that positive health behavior spread farther and faster across clustered-lattice networks (high level of clustering) than random networks (same number of neighbors per individual but less clustering). In a different application of this approach, Paluck and Shepherd (2012) first identified all highly connected individuals in an analysis of a high school's complete social network. Next, they randomly assigned a subset of these social referents to participate in an antiharassment program. Results revealed that the exogenously altered anticonflict behavior among treated individuals influenced the perceived norms and behavior of students with ties to these treated individuals (vs the control individuals) in the network.

Internet experiments are increasingly popular, especially in social psychology (Buhrmester et al., 2011). The Internet experiments can be, though are not necessarily, field experiments. For an Internet experiment to be considered a field experiment, the intervention of interest should be a treatment that is normally carried out online, such as participation in an online health community (Centola, 2010), or in an online video game. Online participants are not automatically more representative than convenience samples such as student participants; this depends on the research question and nature of the online subject pool. In a good example of an Internet field experiment, Facebook messages were randomly suggested to Facebook users to post in their newsfeed. Of three different information messages, one was designed to put social pressure on potential voters. This message increased clicking on a public 'I voted' button and online information seeking about polling locations, relative to the two control messages (Bond et al., 2012). This intervention represents a naturalistic online field experiment, since the message treatments and the outcome variables reflect typical online behavior among Facebook users.

History of Field Experiments

Birth of Field Experiments

Field experimentation literally originated in the field, when Ronald Fisher and Jerzy Neyman conducted a series of agricultural experiments that resulted in a book entitled *The design of field experiments* (Fisher, 1935). One of the first field experiments in the social sciences tested if providing information about the election stimulates voter registration (Gosnell, 1927).

Second Half of Twentieth Century

The second half of the twentieth century was the period of large-scale, so-called social experiments such as the British Pricing experiment and the New Jersey Income Maintenance experiment (Levitt and List, 2009). More than 235 social experiments were completed at that time (for a review see Greenberg and Shroder, 2004), with the primary purpose of informing policy.

In political science, an early wave of 'get out the vote' field experiments (Adams and Smith, 1980; Eldersveld, 1956; Miller et al., 1981) followed up on Gosnell's original experiment on this topic. Unlike social experiments, the get out the vote experiments were aimed at the scientific community and are cited widely up to present. In 1963, Campbell and Stanley's famous book *Experimental and Quasi Experimental Designs for Research* (Campbell and Stanley, 1963) further advanced field experimentation. Campbell advocated a vision for an 'experimenting society' where policy decisions are based on experimentation and follow scientific principles (Campbell, 1998).

In social psychology, some early and notable field experiments include the 'stranded motorist' studies on behavior regarding helping persons in need (Gaertner, 1973), the 'lost-letter' paradigm to study cooperation (Milgram et al., 1965), the 'billboard' experiments (Freedman and Fraser, 1966), and the 'parking lot' studies (Kallgren et al., 2000) to study social influence and persuasion. The advent of the cognitive revolution, the prioritization of mediation analysis to understand psychological processes, and the requirement for multiple study packages in top journals subsequently relegated field experimentation to a secondary role in social psychology (Cialdini, 2009).

Late-Twentieth and Twenty-First Century

In 2000, Gerber and Green published what later became a landmark get out the vote field experiment that overcame many methodological limitations of prior experiments (Gerber and Green, 2000) and inspired a great number of experimental studies on voter turnout (for a review see Green and Gerber, 2008). This wave of get out the vote experiments raised the profile of field experiments in political science (John, 2013). Field experiments with explicitly theoretical as well as policy aims also blossomed in economics (Levitt and List, 2009). They became especially popular in the area of development economics, in part because international aid agencies became sympathetic to the use of field experiments (Humphreys and Weinstein, 2009). The twenty-first century is also seeing a renewed interest in field experiments in social psychology (Cialdini, 2009; Paluck and Cialdini, 2014).

See also: Experimental Design: Randomization and Social Experiments; Quasi-Experimental Designs; Social Experiments; Social Network Analysis; Societal Impact Assessment.

Bibliography

Adams, W.C., Smith, D.J., 1980. Effects of telephone canvassing on turnout and preferences: a field experiment. Public Opinion Quarterly 389–395.

Adams, G., Stocks, E., 2008. A cultural analysis of the experiment and an experimental analysis of culture. Social and Personality Psychology Compass 2 (5), 1895–1912. http://dx.doi.org/10.1111/j.1751-9004.2008.00137.x.

Aronow, P.M., Samii, C., 2013. Estimating Average Causal Effects under Interference between Units. Retrieved from: http://arxiv.org/abs/1305.6156.

Ashraf, N., Karlan, D., Yin, W., 2006. Tying Odysseus to the mast: evidence from a commitment savings product in the Philippines. The Quarterly Journal of Economics 121 (2), 635–672. http://dx.doi.org/10.1162/qjec.2006.121.2.635.

Ayres, I., 1991. Fair driving: gender and race discrimination in retail car negotiations. Harvard Law Review 104 (4), 817–872. http://dx.doi.org/10.2307/1341506.

Banerjee, A., Banerji, R., Duflo, E., Glennerster, R., Khemani, S., 2010. Pitfalls of participatory programs: evidence from a randomized evaluation in education in India. American Economic Journal: Economic Policy 2 (1), 1–30. http://dx.doi.org/10.1257/pol.2.1.1.

Banerjee, A.V., Duflo, E., 2009. The experimental approach to development economics. Annual Review of Economics 1 (1), 151–178. http://dx.doi.org/10.1146/annurev.economics.050708.143235.

Barron, L.G., Hebl, M., King, E.B., 2011. Effects of manifest ethnic identification on employment discrimination. Cultural Diversity and Ethnic Minority Psychology 17 (1), 23–30. http://dx.doi.org/10.1037/a0021439.

Bertrand, M., Djankov, S., Hanna, R., Mullainathan, S., 2007. Obtaining a driver's license in India: an experimental approach to studying corruption. The Quarterly Journal of Economics 1639–1676. http://dx.doi.org/10.1162/qjec.2007.122.4.1639.

Bertrand, M., Mullainathan, S., 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. American Economic Review 94 (4), 991–1013. http://dx.doi.org/10.1257/0002828042002561.

Blum, A., 2011. Improving Peacebuilding Evaluation: A Whole-of-Field Approach. United States Institute of Peace special report. Retrieved from: http://www.usip.org/sites/default/files/SR-Improving-Peace-Building-Evaluation.pdf.

Bobonis, G.J., Miguel, E., Puri-Sharma, C., 2006. Anemia and school participation. The Journal of Human Resources 41 (4), 692–721.

Bond, R.M., Fariss, C.J., Jones, J.J., Kramer, A.D.I., Marlow, C., Settle, J.E., Fowler, J.H., 2012. A 61-million-person experiment in social influence and political mobilization. Nature 489 (7415), 295–298. http://dx.doi.org/10.1038/nature11421.

Bouckaert, J., Dhaene, G., 2004. Inter-ethnic trust and reciprocity: results of an experiment with small businessman. European Journal of Political Economy 20, 869–886.

Buhrmester, M., Kwang, T., Gosling, S.D., 2011. Amazon's mechanical Turk: a new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science 6 (1), 3–5. http://dx.doi.org/10.1177/1745691610393980.

Bullock, J.G., Green, D.P., Ha, S.E., 2010. Yes, but what's the mechanism? (Don't expect an easy answer). Journal of Personality and Social Psychology 98, 550–558.

Butler, D.M., Broockman, D.E., 2011. Do politicians racially discriminate against constituents? A field experiment on state legislators. American Journal of Political Science 55 (3), 463–477. http://dx.doi.org/10.1111/j.1540-5907.2011.00515.x.

Campbell, D.T., 1998. The experimenting society. In: Dunn, W.N. (Ed.), The Experimenting Society: Essays in Honor of Donald T. Campbell. Transaction Publishers, New Brunswick, NJ.

Campbell, D.T., Stanley, J.C., 1963. Experimental and Quasi-Experimental Research Design for Research. Rand McNally, Chicago, IL.

Carlsson, M., Rooth, D.-O., 2011. Revealing Taste-Based Discrimination in Hiring: A Correspondence Testing Experiment with Geographic Variation. Discussion paper No. 6153, Forschungsinstitut zur Zukunft der Arbeit. Retrieved from: http://nbnresolving.de/urn:nbn:de:101:1–2012010913732.

Cialdini, R.B., 2009. We have to break up. Perspectives on Psychological Science 4 (1), 5–6. http://dx.doi.org/10.1111/j.1745-6924.2009.01091.x.

Centola, D., 2010. The spread of behavior in an online social network experiment. Science 329 (5996), 1194–1197. http://dx.doi.org/10.1126/science.1185231.

Ditlmann, R.K., Lagunes, P., 2014. The (identification) cards you are dealt: biased treatment of Anglos and Latinos using municipal versus unofficial ID cards. Political Psychology 5 (4), 539–555.

Duflo, E., Dupas, P., Kremer, M., 2007. Peer Effects, Pupil-Teacher Ratios, and Teacher Incentives: Evidence from a Randomized Evaluation in Kenya (Unpublished manuscript). Retrieved from: http://isites.harvard.edu/fs/docs/icb.topic436657.files/ETP_Kenya_09.14.07.pdf.

Dunning, 2012. Natural Experiments in the Social Sciences: A Design-Based Approach. Cambridge University Press, Cambridge, England.

Eldersveld, S.J., 1956. Experimental propaganda techniques and voting behavior. The American Political Science Review 50 (1), 154–165. http://dx.doi.org/10.2307/1951603.

Fisher, R.A., 1935. The Design of Experiments. Oliver and Boyd, Edinburgh, Scotland.

Freedman, J.L., Fraser, S.C., 1966. Compliance without pressure: the foot-in-the-door technique. Journal of Personality and Social Psychology 4 (2), 195–202. http://dx.doi.org/10.1037/h0023552.

Gaertner, S.L., 1973. Helping behavior and racial discrimination among liberals and conservatives. Journal of Personality and Social Psychology 25 (3), 335–341. http://dx.doi.org/10.1037/h0034221.

Gerber, A.S., Green, D.P., 2000. The effects of canvassing, telephone calls, and direct mail on voter turnout: a field experiment. American Political Science Review 94 (3), 653–663.

Gerber, A.S., Green, D.P., 2012. Field Experiments: Design, Analysis, and Interpretation. W.W. Norton & Company, New York.

Gerber, A.S., Green, D.P., Kaplan, E.H., 2004. The illusion of learning from observational research. In: Shapiro, I., Smith, R.M., Masoud, T.E. (Eds.), Problems and Methods in the Study of Politics. Cambridge University Press, Cambridge, England, pp. 251–273.

Gerber, A.S., Green, D.P., Larimer, C.W., 2008. Social pressure and voter turnout: evidence from a large-scale field experiment. American Political Science Review 102 (01), 33–48. http://dx.doi.org/10.1017/S000305540808009X.

Glennester, R., Takavarasha, K., 2013. Running Randomized Evaluations: A Practical Guide. Princeton University Press, Princeton, NJ.

Gneezy, A., Gneezy, U., Nelson, L.D., Brown, A., 2010. Shared social responsibility: a field experiment in pay-what-you-want pricing and charitable giving. Science 329 (5989), 325–327. http://dx.doi.org/10.1126/science.1186744.

Gosnell, H.F., 1927. Getting Out the Vote: An Experiment in the Stimulation of Voting. The University of Chicago Press, Chicago, IL.

Green, D.P., Gerber, A.S., 2008. Get Out the Vote: How to Increase Voter Turnout. Brookings Institution Press, Washington, DC.

Greenberg, D.H., Shroder, M., 2004. The Digest of Social Experiments. Urban Institute Press, Washington, DC.

Hebl, M.R., Foster, J.B., Mannix, L.M., Dovidio, J.F., 2002. Formal and interpersonal discrimination: a field study of bias toward homosexual applicants. Personality and Social Psychology Bulletin 28 (6), 815–825. http://dx.doi.org/10.1177/0146167202289010.

Humphreys, M., Weinstein, J.M., 2009. Field experiments and the political economy of development. Annual Review of Political Science 12 (1), 367–378. http://dx.doi.org/10.1146/annurev.polisci.12.060107.155922.

John, P., 2013. Field Experiments in Political Science Research. Retrieved from: http://ssrn.com/abstract=2207877.

Kaiser, C.R., Pratt-Hyatt, J.S., 2009. Distributing prejudice unequally: do whites direct their prejudice toward strongly identified minorities? Journal of Personality and Social Psychology 96 (2), 432–445. http://dx.doi.org/10.1037/a0012877.

Kallgren, C.A., Reno, R.R., Cialdini, R.B., 2000. A focus theory of normative conduct: when norms do and do not affect behavior. Personality and Social Psychology Bulletin 26 (8), 1002–1012. http://dx.doi.org/10.1177/01461672002610009.

Koopmans, R., Veit, S., 2013. Cooperation in ethnically diverse neighborhoods: a lost-letter experiment. Political Psychology. http://dx.doi.org/10.1111/pops.12037.

Levitt, S.D., List, J.A., 2009. Field experiments in economics: the past, the present, and the future. European Economic Review 53 (1), 1–18. http://dx.doi.org/10.1016/j.euroecorev.2008.12.001.

Lewin, G.W., 1997. Problems of research in social psychology (1943–1944). In: Lewin, K. (Ed.), Resolving Social Conflicts & Field Theory in Social Science. American Psychological Association, Washington, DC, pp. 279–288.

Love, J.M., Kisker, E.E., Ross, C., Raikes, H., Constantine, J., Boller, K., Vogel, C., 2005. The effectiveness of early head start for 3-year-old children and their parents: lessons for policy and programs. Developmental Psychology 41 (6), 885–901. http://dx.doi.org/10.1037/0012-1649.41.6.885.

Miguel, E., Kremer, M., 2004. Worms: identifying on education and health in the presence of treatment externalities. Econometrica 72 (1), 159–217.

Milgram, S., Mann, L., Harter, S., 1965. The lost-letter technique: a tool of social research. Public Opinion Quarterly 29 (3), 437–438. http://dx.doi.org/10.1086/267344.

Miller, R.E., Bositis, D.A., Baer, D.L., 1981. Stimulating voter turnout in a primary field experiment with a precinct committeeman. International Political Science Review 2 (4), 445–459. http://dx.doi.org/10.1177/019251218100200405.

Pager, D., 2007. The use of field experiments for studies of employment discrimination: contributions, critiques, and directions for the future. The ANNALS of the American Academy of Political and Social Science 609 (1), 104–133. http://dx.doi.org/10.1177/0002716206294796.

Paluck, E.L., 2009. Reducing intergroup prejudice and conflict using the media: a field experiment in Rwanda. Journal of Personality and Social Psychology 96 (3), 574–587. http://dx.doi.org/10.1037/a0011989.

Paluck, E.L., 2010. The promising integration of field experimentation and qualitative methods. Annals of the American Academy of Political and Social Science 628, 59–71.

Paluck, E.L., Cialdini, R., 2014. Field research methods. In: Reis, H.T., Judd, C.M. (Eds.), Handbook of Research Methods in Social and Personality Psychology. Cambridge University Press, New York, NY, pp. 81–97.

Paluck, E.L., Shepherd, H., 2012. The salience of social referents: a field experiment on collective norms and harassment behavior in a school social network. Journal of Personality and Social Psychology 103 (6), 899–915. http://dx.doi.org/10.1037/a0030015.

Peterson, P., Howell, W., Wolf, P.J., Campbell, D., 2003. School vouchers: results from randomized experiments. In: Hoxby, C.M. (Ed.), The Economics of School Choice. University of Chicago Press, Chicago, IL, pp. 107–144.

Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin Boston, MA.

Sherman, L.W., Strang, H., 2004. Experimental ethnography: the marriage of qualitative and quantitative research. The ANNALS of the American Academy of Political and Social Science 595 (1), 204–222. http://dx.doi.org/10.1177/0002716204267481.

Sondheimer, R.M., Green, D.P., 2010. Using experiments to estimate the effects of education on voter turnout. American Journal of Political Science 54 (1), 174–189. http://dx.doi.org/10.1111/j.1540-5907.2009.00425.x.

Walton, G.M., Cohen, G.L., 2007. A question of belonging: race, social fit, and achievement. Journal of Personality and Social Psychology 92, 82–96.

Field Observational Research in Anthropology and Sociology

Roger Sanjek, City University of New York, Flushing, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 8, pp. 5620–5625, © 2001, Elsevier Ltd.

Abstract

Find observational research, or ethnography, has a 150-year history in anthropology and sociology. As a form of 'I-witnessing' ongoing social life, it triangulates fieldwork data preserved in notes and records with comparative social theory and documentary contextualization. The research process begins with problem and site selection considerations, access to natural behavior-stream events, and development of rapport with research subjects. As the fieldwork 'funnel' narrows, discovery procedures progress from situated listening and observation, to interviews. Fieldnotes, combined with headnotes, are later employed in constructing written ethnographies that may be assessed by canons of validity rather than reliability, which attaches more centrally to other research modalities.

Field observational research in anthropology and sociology – or the process of doing ethnography – brings the social science investigator directly to the scene of the human behavior she or he describes. It involves 'being there,' or 'I-witnessing.' So also, of course, do other forms of documentation: journalism, film, travel writing, explorer or missionary's report, even background research by a novelist. What makes ethnography distinctive as a research practice is that it forms one point of a triangular process of knowledge construction which also includes comparative social theory and documentary contextualization.

The fieldnotes that result from ethnographic discovery procedures are not merely 'written up,' but are filtered and interpreted against theoretical propositions and the comparative record of other field observational studies; they are also grounded and enhanced by historical, ecological, demographic, economic, and other documentary sources that provide background and context. As the products of this mode of research are read – monographs and articles also referred to as ethnography – they stimulate new comparative theoretical thinking, which in turn suggests further problems and interpretations to be resolved through more field observational research. Ethnographies also regularly lead to new demands and rising standards for documentary contextualization – more history, more ecological or demographic backgrounding, more attention to state policy, economic trends or the world system. This triangle of ethnography, comparison, and contextualization is, in essence, the process by which field observational research in anthropology and sociology is utilized to explain and interpret human cultures and social life.

History and Scope in Anthropology and Sociology

In 1851 Morgan published his *League of the Ho-de-no-sau-nee, or Iroquois*; combining fieldwork and comparative and theoretical interests in political organization, this book was the first anthropological ethnography. It depicted the structure and operation of Iroquois society, detailing matrilineal kinship, political and ceremonial life, material culture and religion. After formation of the US government Bureau of American Ethnology in 1879, a stream of ethnographic accounts of native American societies began, including landmark field observational studies of the Zuni by Cushing, the Baffin Island Inuit by Boas, and, after 1900, ethnographies of Plains, Eastern Woodland, and California groups by Boas' students trained at Columbia University. During the 1930s the scope broadened to include rural and urban studies in Latin America, the Caribbean, Africa, and Europe, and also in contemporary US settings.

In Great Britain's colonies, local government official or missionary fieldworkers followed the topical guidebook *Notes and Queries on Anthropology*, drafted by Oxford University theorist Tylor and others, and first published in 1874. This armchair scholar/man-on-the-spot division of labor ended two decades later when Oxford-trained Spencer collaborated in field observational study with local expert Gillen, producing their *Central Tribes of Native Australia* in 1899, and providing detailed coverage of cosmology, ritual, and social organization which revealed unheralded complexity, and which sparked theoretical work by Durkheim and Freud. In 1898 a team from Cambridge University conducted fieldwork in Torres Strait north of Australia, and member Rivers used genealogical data to probe community history, migration trajectories, marriage patterns, demography, and inheritance. Rivers and his students conducted further ethnographic studies in India and the Pacific, but their work was eclipsed by the more thorough Trobriand Islands field observational research of Malinowski, resulting in *Argonauts of the Western Pacific* in 1922 and its famous first chapter on method. Malinowskian ethnography dominated British social anthropology thereafter, and by the 1940s reshaped American cultural anthropology as well.

In sociology, there is less agreement about the disciplinary origin of field observational research methods. Conventionally, the monographs of the University of Chicago department during the 1920s and 1930s are cited (Madge, 1962): studies of itinerant 'hobo' workers, youth gangs, dancehall employees and patrons; Italians, Poles, Jews, French Canadians, Mexicans, and Chinese; and central city 'gold coast' apartment, artist colony, and rooming house dwellers. Two decades before this research, however, Du Bois used similar methods (both he and the Chicago school were inspired by British and American urban survey and settlement house movements) to produce *The Philadelphia Negro* (1899), which depicted life in an inner

city neighborhood comprised of southern migrant and northern-born African-Americans. Even earlier, Engels in 1844 studied new classes, Irish labor migration, urban spatial arrangements, symbolic attachments, and modes of observable behavior in Manchester and other rapidly industrializing British cities, although his *Condition of the Working Class in England* was not published in English until 1892.

Since the 1920s there has been much cross-fertilization between anthropological and sociological ethnography. A few of the Chicago studies were conducted by researchers identifiable as anthropologists, and the Lynds, authors of *Middletown* (1929) which credits anthropologists Rivers and Wissler for methodological inspiration, and Whyte, author of *Street Corner Society* (1955 [1943]) and influenced by Harvard anthropologists, are presented as major figures in the development of American sociology (Madge, 1962). Gluckman, a Manchester social anthropologist whose fieldwork was conducted in Africa, was influential in the development of ethnographic fieldwork by British sociologists working in schools, gangs, and factories during the 1950s and later (Burgess, 1984). In recent decades, both anthropologists and sociologists have undertaken field observational research in a wide range of overlapping settings, at home and abroad; so increasingly do scholars from folklore, education, environmental psychology, political science, and other disciplines. Ethnography, however, remains central to credentialing, teaching, publishing, and theory-construction in anthropology, while it occupies a more marginal position in contemporary sociology, or in other fields.

Selection, Access, Rapport

The selection of a particular population or site for field observational research is ordinarily related to some unanswered question or outstanding problem in social theory. Personal connections or predilections of researchers also shape this selection, but the fieldworker must justify his or her choice in terms of some significant theory to which the project is addressed. Usually this is made explicit in a written proposal for funds to underwrite the fieldwork.

While ethnography is thus lodged from conception in social theory, a comparative viewpoint also molds the fieldwork process. First, investigators acquire a cross-cultural or cross-site perspective by training and reading; when addressing any aspect of social life – marriage, leadership, ethnicity, etc. – field observational researchers recall examples of similarities and differences elsewhere. Second, the comparative perspective focuses attention on trends and transitions, not just on similarities and differences at random (which are infinite). Rather than treating each fieldwork case as unique, ethnographers place the social phenomena they study within comparative frames (foraging, horticultural, agricultural, pastoral, industrial, colon ial, neocolonial economies; cooperative, competitive, individualistic societies; gender subordination, complementarity or equality; etc.) Ethnographic results in turn provoke debate, revision, and innovation in such theorizing.

While significant theories bring fieldworkers to particular locations, actors, and activities, once they arrive they begin to observe and listen. Often they must learn to listen – by learning the language, the local vocabulary, and current verbal conventions. Field observational research now turns away from theoretical discourse to the viewpoints and concepts of the people (informants, subjects, actors, consultants) themselves. Ethnographers aim to document how the people see and talk about their everyday social activities and groupings, and about the wider worlds they live in. It is their normal scenes of activity, topics of conversation, and standards of evaluation that are the objects of field observational research.

This does not begin by announcing: 'I'm your ethnographer; when can I interview you?' Investigators must be honest about their role and sponsorship (covert field observational research remains controversial and rare), but their paramount aim is to listen and to move as quickly as possible into the natural settings of social life, the places people would be, doing what they would be doing, if the ethnographer was not there. This involves a continuing process of requesting and obtaining access to the many locations, activities, and groups that comprise the overall fieldwork arena. As ethnographers observe and listen in a wide-ranging manner, within parameters set by the significant theories that bring them there, they begin to understand culturally meaningful behavioral conventions, and to formulate culturally appropriate utterances. This indexes a second continuing process of building interpersonal rapport with informants, a topic that preoccupies many autobiographical accounts of field observational research – one of the most instructive being Whyte's, 1955 [1943] Appendix to *Street Corner Society*.

As the ethnographic process unfolds, the fieldworker must constantly make decisions about where to be, whom to listen to, what events to follow, and what to safely ignore and leave out. These decisions are guided both by the significant theories prefiguring fieldwork, and by new theories of significance that arise in the field. These latter theories – hunches, hypotheses, ideas about connections and relationships – emerge as listening and observation proceed. They suggest which persons and activities to focus upon, what places and events to attend to, and which objects and their circulation to follow. As this occurs, the fieldwork 'funnel' narrows, to use Agar's (1980) apt metaphor. The early phase of securing access and building rapport is wide, open, and nearly all-encompassing; as theories of significance emerge, pan out, or are discarded, the funnel of informants, events, and activities narrows. Goals crystallize. Research design sharpens.

Discovery Procedures: Listening

Ethnography's greatest strength is situated listening. Interviews become useful at later stages of fieldwork, but field observational research begins by listening to what British anthropologist Richards (1939) called 'speech in action.' Here the actors control topicality – talking to each other about what they usually do – and the researcher secures access to their turf – the locations they usually occupy. Early on, as rapport begins to be established, fieldworkers deliberately place themselves in a wide sampling of such situations. Then, as the research funnel narrows and other methods are deployed, the ethnographer becomes more selective about where to listen.

The importance of situated listening was emphasized memorably by Whyte. After he asked an intrusive question

during illicit gambling activities, and stopping the ongoing flow of behavior, Whyte's key informant Doc told him, 'Go easy on that "who," "what," "why," "when," "where" stuff, Bill. You ask those questions and people will clam up on you. If people accept you, you can just hang around, and you'll learn the answers in the long run without even having to ask the questions.' 'I found that this was true. As I sat and listened, I learned the answers to question that I would not even have had the sense to ask if I had been getting my information solely on an interview basis' (1955 [1943], p. 303).

Situated listening becomes the basis for understanding both what can be talked about in later conversations and interviews, and what categories and evaluations the actors use to observe and interpret ongoing behavior.

Discovery Procedures: Observing

Many veteran field observational researchers bristle when they are told that their methods are not quantitative. Using both emic categories articulated by the actors, and etic categories formulated by the community of social science observers, ethnographers count, sample, and calculate ratios, proportions, and probabilities as part of their fieldwork activities (Johnson, 1978). These countable quantities include numbers of people in events, their positions, their comings and goings; objects, inventories; and exchanges, movements, orderings, sequences, associations, assemblages, and arrangements of many sorts.

Malinowski's classic *Argonauts of the Western Pacific* (1922) incorporated his numerous observations of the interisland *kula* exchange of seashell valuables – necklaces traded for armlets – as canoes from one community arrived at another, exchanges were conducted, and then additional exchanges of items including foodstuffs, timber, greenstone, and pottery followed. Who exchanged what, the order, etiquette, volume, frequency, and other variables were all counted and analyzed. In Whyte's *Street Corner Society* (1955 [1943]) the size, composition, duration of activity, and initiation of interaction were recorded and analyzed for numerous group events; Whyte even demonstrated quantitatively a relationship between hierarchical group leadership structure and bowling scores.

Harris (1964) furnishes field observational researchers with an etic model of 'the great stream of uninterrupted behavior' that extends backward in time and unfolds continuously. He provides a technical vocabulary to operationalize recording of isolatable events and linked segments of the behavior stream – from individual body movements and their environmental effects at the most basic level, through sequences of multiactor scenes connected by actor, place, or object. Sanjek (1978) applied this approach in a multiethnic Ghanaian urban neighborhood to collect (through interviews) sequences of multiactor scenes for 40 adults, each over four-day periods, in order to test whether their social networks were structured more by common ethnicity or by class position, and in which behavioral settings ethnicity and class were stronger or weaker.

Johnson (1975, 1978) developed an observational fieldwork approach to formulate statistically valid time allocation patterns in his research among the Machiguenga of Peru. By randomly selecting actors and times for observation at their most likely location, and recording their ongoing behavior at the time of each random visit, a powerful quantitative record of how a given population allocates its activities among production, reproduction, leisure, and other categories resulted. This and similar methods have been employed in a wide range of fieldwork settings (Gross, 1984).

Discovery Procedures: Interviewing

Barth recounts of his highland New Guinea research, 'My field procedure was … to rely very heavily on observation of spontaneous, unelicited word and act. I paid particular attention to this in Baktaman questions and explanations *to each other* – in men's houses, in the fellowship of hunting and working, during initiations and other rites. Only secondarily, and with much reticence, did I ask my own questions, and then only when they seemed cast in a native pattern and received easy and natural response' (1975, p. 226).

Interviewing is problematic (Briggs, 1986). Human beings are apt to reinterpret or reformulate the past to make it conform with their ongoing sense of the present. Moreover, the ever present danger of an observer effect in fieldwork – that the presence of the observer affects and reshapes an actor's behavior and response – is especially difficult to control in an artificial one-to-one interview situation where normal social activity is suspended. Still, interviews are essential to field observational research. They are best left for later stages of the ethnographic encounter when, to use Agar's funnel metaphor, the speech events of fieldwork move from wide and open to narrow and more focused.

After the investigator has become at ease with the language and local codes and references, and has successfully gained access to natural conversations and amassed a stock of cultural knowledge via situated listening, deliberate intervention into speech events may begin. These steps may be graded from least to most intrusive and controlled.

1. Still on the informants' turf, and still in the accustomed activities of daily life, the field observational researcher begins to enter natural conversations as a speaker, and begins to shift topicality to her or his own interests. This process starts gently, by moving appropriately into rounds of chatting, gossiping, and ordinary comment. As cultural competence increases, and as theories of significance emerge, the fieldworker attempts to direct conversations by introducing questions and suggesting topics for responses from informants.
2. After an initial period of fieldwork – a few months perhaps – arranged interviews may begin. This class of speech events is disruptive – the informant is removed from her or his turf, either to the ethnographer's household or office, or by transforming an everyday location into a scene of ethnographer–informant dialogue – an activity that would otherwise not be occurring there. Typically the earliest of these deliberate breaks in time–place behavioral flow reserve topicality for the actor. In such open-ended or discovery interviews, the informant is encouraged to move the conversation according to his or her own interests.
3. In later and more productive interviews, the ethnographer begins to assert control. General topics are introduced, allowing informants to expand freely upon their own points

of view and knowledge. In more structured ethnographic interviews, topicality is more firmly shaped and directed by the fieldworker; informant responses now move away from orations and free commentaries and to more specific responses to questions.

4. In the most focused forms of interviewing, the ethnographer controls both turf and topicality as fully as possible. Questionnaires and interview schedules may be used, and the objective is to obtain particular types and pieces of data. These may include household interviews, psychological tests, or reports of disputes, or may encompass repeated interview sessions to secure lengthy life histories, with the fieldworker now guiding the subject according to preset standards of scope and comprehensiveness.
5. Though not a major part of ethnographic practice, in some instances, and when still on the informants' turf, the fieldworker may ask direct and pointed questions, and attempt to secure precise pieces of data for his or her records. Interventions of this sort are especially dangerous – the inappropriateness of such seizures of topicality in everyday settings may be jarring to the actors. Typically speech events of this sort occur in the final days of fieldwork, when local rapport is at its peak, research goals are most pressing, and the fieldwork funnel approaches its vanishing point.

Notes and Records

The production of notes and records (Sanjek, 1990) begins to move the field observational research process toward its ultimate written products. Focused interview sessions with seated informants often permit direct transcription of verbal statements, but in open-ended and ethnographic interviews, brief written notes – what Ottenberg (in Sanjek, 1990) terms scratch notes – are taken during the session, and these form the basis for the construction later of fuller written fieldnotes. Fieldworkers often go through this two-step process even when interviews are tape recorded, both as a backup to and index of the taped session, and because of the analytic gains many ethnographers note in transforming their scratch notes into fuller descriptive fieldnotes.

In field observational research in natural behavior stream settings, similar brief jottings may be inscribed, but major attention is directed to the event in progress. Often it is not even possible to record scratch notes, and both they and fuller fieldnote description occur later. Mead (1977) wrote about the nagging pressure to type up fieldnotes from scratch notes, and about the danger of scratch notes growing 'cold' when this was delayed, even by one day. But she also wrote of the satisfaction of being caught up with this work, and of the importance for later ethnographic writing of the insights gained in moving from scratch notes to descriptive fieldnotes.

Ottenberg sees this process as the interaction of scratch notes and headnotes – the stored memories and interpretations that arise from direct participant observation as filtered by the ethnographer's overall theoretical stance. Headnotes form an essential complement to fieldnotes (and to more formal fieldwork datasets or records). Headnotes are employed to make sense of one's fieldnotes when they are re-read later for ethnographic writing projects. The importance of headnotes is particularly evident when one ethnographer attempts to use another's fieldnotes, and quickly realizes how difficult it is to understand them without headnotes of one's own (Sanjek, 1990).

Fieldnotes typically are kept in running chronological order, and may minimally be indexed separately by actor and topic. Additionally, field observational researchers often generate records, or well-defined sets of data organized by topic, person, or other category. These may include records of household composition, land holdings, ritual performances, life histories, folktales, etc. Sometimes decisions to collect sets of records are part of the initial research design, and preset forms to enter such data may be brought with the investigator to the field setting. In other cases, a decision to collect a systematic body of records develops during fieldwork, often emerging within general running fieldnotes and then being separated out in new record files.

Fieldnotes and records present ethnographers with great masses of information – hundreds, even thousands of pages – that while being accumulated may be arranged only by chronological order or by topic. Malinowski (1922) urged that ethnographers constantly re-read and begin to organize and analyze their notes while still conducting field research, but commonly the more focused work of indexing them occurs when fieldwork is over. As ethnographers then turn to ethnographic writing, they readdress the theoretical discourse they turned away from in fieldwork. Fieldnotes and records must now be related directly to the comparative and contextual points of the knowledge-construction triangle.

Canons of Validity

Field observational research aims to maximize validity – 'the degree to which scientific observations actually measure or record what they purport to measure' – rather than reliability – 'the repeatability, including interpersonal replicability, of scientific observations' (Pelto and Pelto, 1978, p. 33). Reliability is valued in survey, medical, and product safety research. In ethnography, restudies to assess change may be conducted, but independent investigation of the same locale or population is unlikely or impossible.

Validity depends on the intensity and depth of field observational research, but it always remains a goal rather than a property that can be independently measured (Wolcott, 1995, pp. 167–70). A quantitative, and highly reliable, randomized sample study that investigates the gender, ethnicity, class, and other characteristics of the five best friends of members of a given population would be less valid than one in which cultural meanings and categories of friendship were first discovered through listening to natural conversations or eliciting such information in interviews. In the case of Indian immigrants in New York City, for example, the 'five best friends' study might miss entirely the category of *rakhi* ritual friendships between adult males and females which are affirmed in an annual festival where the man ties a flower bracelet on the woman's wrist.

If validity remains important, but elusive, how can it be assessed in field observational research results? Sanjek (1990, pp. 393–404) proposed three canons of validity by which

ethnographic writing may be evaluated. The first is theoretical candor, the openness with which the ethnographer addresses the significant theories, and even more, the local theories of significance that structured the fieldwork process. A second canon calls for explicit depiction of the ethnographer's fieldwork path – the number of informants from whom information was obtained, in what ways, and their relationship both to the wider population the field observational study concerns and to each other. A third canon concerns information about the fieldnote evidence itself: not simply 'how much' and its basis in listening, observing, or interviewing, but more significantly the precise relationship of notes and records to the written ethnography. Some ethnographies utilize fieldnotes directly, even masses of them; others, for rhetorical or narrative purposes, do not, and need not. What matters in the end is that readers of field observational research have a clear picture of what the ethnographer did and why, who they talked to and learned from, and what they brought back to document it.

Field observational research produces results that can be obtained in no other way. This ethnographic tradition entails that 'the description of people's activities, their interactions with each other, [and] their verbal behavior' be 'copious and detailed'; 'cumulative development … consists above all in the fact that such data, provided they are adequate, never become obsolete and can be used by analysts … in researching questions not visualized by the researchers who collected them' (Anglin, 1979, p. 49).

See also: Fieldwork in Social and Cultural Anthropology; Observational Studies: Overview; Participant Observation.

Bibliography

Agar, M., 1980. The Professional Stranger: An Informal Introduction to Ethnography. Academic Press, New York.

Anglin, A., 1979. Analytic and folk models: The Tallensi case. In: Holy, L. (Ed.), Segmentary Lineage Systems Reconsidered. Department of Social Anthropology, The Queen's University of Belfast, Ireland.

Barth, F., 1975. Ritual and Knowledge Among the Baktaman of New Guinea. Yale University Press, New Haven, CT.

Briggs, C.L., 1986. Learning How to Ask: A Sociolinguistic Appraisal of the Role of the Interview in Social Science Research. Cambridge University Press, New York.

Burgess, R.G., 1984. In the Field: An Introduction to Field Research. George Allen and Unwin, London.

Emerson, R.M., Fretz, R.I., Shaw, L.L., 1995. Writing Ethnographic Fieldnotes. University of Chicago Press, Chicago.

Gross, D.R., 1984. Time allocation: A tool for the study of cultural behavior. Annual Review of Anthropology 13, 519–558.

Harris, M., 1964. The Nature of Cultural Things. Random House, New York.

Johnson, A., 1975. Time allocation in a Machiguenga community. Ethnology 14, 301–310.

Johnson, A.W., 1978. Quantification in Cultural Anthropology: An Introduction to Research Design. Stanford University Press, Stanford, CA.

Madge, J., 1962. The Origins of Scientific Sociology. Tavistock, London.

Malinowski, B., 1922. Argonauts of the Western Pacific. Routledge, London.

Mead, M., 1977. Letters From the Field, 1925–1975. Harper and Row, New York.

Pelto, P.J., Pelto, G.H., 1978. Anthropological Research: The Structure of Inquiry. Cambridge University Press, New York.

Richards, A., 1939. The development of field work methods in social anthropology. In: Barlett, F.C., et al. (Eds.), The Study of Society. Routledge and Kegan Paul, London, pp. 272–316.

Sanjek, R., 1978. A network method and its uses in urban ethnography. Human Organization 37, 257–268.

Sanjek, R. (Ed.), 1990. Fieldnotes: The Makings of Anthropology. Cornell University Press, Ithaca, NY.

Whyte, W.F., 1955 [1943]. Street Corner Society: The Social Structure of an Italian Slum, second ed. University of Chicago Press, Chicago.

Wolcott, H.F., 1995. The Art of Fieldwork. AltaMira, Walnut Creek, CA.

Field Theory

Gisèle Sapiro, École des hautes études en sciences sociales, Paris, France; and CNRS (Centre Européen de Sociologie et de Science Politique), Paris, France

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Fine, volume 14, pp. 9311–9317, © 2001, Elsevier Ltd.

Abstract

Developed in sociology by Pierre Bourdieu, the concept of field is a heuristic tool for exploring the differentiation of social spheres of activity. The article first traces the origins and genesis of the concept in Bourdieu's theory. It then turns to the general theory of fields that Bourdieu has elaborated, based on empirical research on the literary, the scientific, the religious, the academic, and the political fields. The theory of field has become a research program for a group of scholars who examined issues such as the relations between fields, the use of field theory for comparative purpose, and the geographic borders (national or transnational) of fields. The last section presents other field theories more or less inspired by Bourdieu.

The concept of the field makes it possible to circumscribe individuals' action at the mesolevel, in differentiated social spheres with their own rules of the game and specific interests. Field theory was thoroughly developed by Pierre Bourdieu beginning in the 1960s in order to make sense of the process of social differentiation of activity spheres that comes with the division of labor, while at the same time implementing a relational and topographical approach of the social world. Field theory served as a research program (in Imre Lakatos' sense) and resulted in a great deal of research on the fields of politics, economics, religion, the academy, law, philosophy, literature, art, intellectuals, publishing, sport, etc. In the United States, other uses of the concept of field have appeared, the first of which was the notion of 'organizational field,' theorized by Paul DiMaggio and Walter Powell during the 1980s within a neoinstitutionalist approach. Later, the notion of 'strategic action field' was devised by Neil Fligstein and Doug McAdam in order to combine the neoinstitutionalist approach with collective action theories.

The Origins and Genesis of the Concept of Field

The concept of field is borrowed from theoretical physics: it apprehends relations between elements within a space seen as a force field, according to the laws of attraction and repulsion. It was transposed in psychology by the Gestalt theorists, especially Wolfgang Köhler, who underscored the interdependency of elements in perceptual experience and developed a topographic approach (Levi Martin, 2003). Kurt Lewin adapted it to social psychology in order to consider the interactions between an individual and his or her environment.

It is in this sense that Bourdieu introduced it in sociology, where the abstract concept allows for the methodological autonomization of a space of activity, provided that – as Bourdieu specified for the fields of cultural production – the historical conditions of its autonomization have been studied (Bourdieu, 1971c). The use of the concept of field responds to a double problem. First of all, how does one consider the historical process of the differentiation of social activities that accompanies the division of labor without lapsing into a functionalist approach? There is indeed nothing ineluctable, and nothing mechanical, about this double process. The autonomization of an area of activity generally results from the struggle led by a group of specialists (legal practitioners, for example) in order to obtain social recognition for their authority and jurisdiction in a given area, introducing a rupture between professionals and nonspecialists (for example, between clerics and laypeople). The theory of fields systematizes the analysis of this process, identified by Max Weber, and draws from it methodological consequences – that is, it is possible to autonomize a field as an object of study.

However, this autonomy is always relative. Bourdieu borrows the concept of autonomy from Marxist theory, where it was introduced as part of the debate on reflection theory, which considered works as a superstructure reflecting social conditions. Certain theorists indeed suggested that literary and artistic works were not a simple reflection of the social world, but rather that they expressed a vision of the world that could echo the contradictions that were simmering within the conditions of production. However, Bourdieu criticized these thinkers' failure to consider the mediation that fields of cultural production exert with regard to economic and social conditions: it is there, in what Bourdieu calls a 'field effect,' that he situates the relative autonomy of these spheres. Works indeed bear the marks of their own conditions of production. Cultural producers are engaged in a competitive struggle that abides by rules and specific interests irreducible to economic, political, and social interests. Regarding social space, the 'field effect' produces a 'refraction' effect, another concept Bourdieu borrows from physics, wherein exterior constraints are retranslated according to the particular logic of the field in question. This particular logic derives partly from the structure of the field, and partly from its history. Such an approach makes it possible to avoid, on the one hand, sociological reductionism and, on the other, approaches that examine only the internal contents of a single text (this is just as true for cultural production as it is for science, philosophy, and law).

The structure of the field is defined according to the (unequal) distribution of specific capital in the field: agents endowed with the highest amount of specific capital occupy dominant positions, whereas those who possess less, often newcomers, occupy dominated positions. Such a topographical approach breaks with substantialism (drawing on Ernst

Cassirer's book *Substance and Function*) and with interactionism (namely with symbolic interactionism, but also with spontaneous interactionism, which prevails in literary and art history); in contrast, it is structural (relational) and objectivist: united by their competition for the same issue at stake, that is, acquiring specific capital, agents are defined objectively in relation to one another independently of the interactions between them – interactions that an objectivist analysis makes possible, moreover, to explain. In this regard, network analysis, which is purely descriptive, becomes meaningful in light of the objective properties of individuals.

Though the theory of field and its conception of social space are indebted to the structuralist method, which Bourdieu uses not only to a group's world vision but also to analyze social relations themselves, he conserves from Marxism a dynamic conception of these relations, which derives from their agonistic dimension. The competitive struggle indeed determines the principles of opposition that structure the field and determines antagonisms and alliances according to the laws of attraction and repulsion (Bourdieu, 1984: pp. 113–120). These principles of opposition also create structural homology among fields. The first factor opposes the 'dominated' to the 'dominants': the dominants usually tend to preserve the dominant definition of the activity in question ('orthodoxy'), while the dominated tend to contest it ('heterodoxy'). Drawing on Max Weber's sociology of religion, Bourdieu (1971a,b) systematizes the opposition between priest and prophet, deeming it paradigmatic, and transposes it to the fields of artistic production, where it is transplanted to the opposition between established and avant-garde cultural producers (see Bourdieu, 1971c). A second opposition places the upholders of the autonomy of field, which is based on peer judgments formulated according to specific criteria for determining the symbolic value of products, against those who tend to import into it heteronomous – ideological, or economic – constraints.

These positions evolve based on various factors, some of which, such as crises or politicization, are exogenous (May 1968, for example) and some of which are endogenous, such as internal struggles and social aging. Whereas exogenous factors may contribute to the synchronization of fields (as with May 1968; see Bourdieu, 1992a), endogenous factors impose a temporality that is specific to each field, and this constitutes another sign of fields' relative autonomy. Moreover, relatively autonomous fields are characterized by their autotelic nature, or their self-referentiality – that is, by the reference to their own history. Such is the case for fields of cultural production (literary, artistic, musical) as well as scientific and legal fields, where past problems and solutions cannot be ignored by agents, at the risk of their exclusion from the field.

Field therefore consists in a 'space of possibilities' ('*espace des possibles*') where 'position-taking' ('*prise de position*') – choices between different options more or less constituted as such – are defined by significant gaps, according to the model of structural linguistics, gaps that assume their meaning in relation to the history of the field (for example, the opposition between tonal and atonal music, or between analytic and 'continental' philosophy). For this reason, in Bourdieu's eyes the subject of the work is neither the individual nor the class, as the Marxist perspective suggests, but rather the field as a whole.

The notion of a space of possibilities is in this sense linked to Foucault's concept of the 'field of strategic possibilities' (Foucault, 1969), but it is set apart by the very specific use Bourdieu makes of the concept of 'strategy' in order to apprehend agents' differentiated forms of investment in this space of possibilities. Far from supposing rational, reflexive, or even cynical action, the concept of strategy refers, in Bourdieu's theory of practice, to the extent to which agents are able to improvise according to the exterior constraints they must contend with and to their dispositions: in other words, the latitude they have to maneuver. The concept of strategy aims to describe the encounter between a social trajectory and a space of possibilities. This trajectory is largely determined by an individual's habitus, that is, by the social structures he or she has internalized throughout the process of socialization, in the form of dispositions that in turn structure his or her schemes of perception, action, and evaluation (taste). As such, the question of the relation between the social properties of individuals and their stands or position-taking is not a given but rather a central object in the study of field functioning.

Toward a General Theory of Fields

From the 1970s onward, Bourdieu developed a general theory of fields. The project was not realized in his lifetime, but many traces can be found, from the seminars and talks he dedicated to it during this period (Bourdieu, 1984, 2013a) to the research that he and the members of his team carried out on particular fields: literary, religious, scientific, political, legal, academic, philosophical, artistic, economic, editorial, not to mention the field of *haute couture*. Because it is impossible to take account of all of the research that puts the concept to work, we will limit ourselves here to discussing, on the one hand, the contribution field theory has made in the study of different areas of activity and, on the other hand, the way these fieldworks have enriched or redefined the general theory.

Bourdieu began investigating the fields of cultural production as early as the 1960s (Bourdieu, 1966). He aimed to establish a 'science of (creative) works' ('*une science des oeuvres*') that would overcome the alternative between internal analysis, represented at the time by New Criticism and structuralism in particular, and external analysis, which was split between the singularizing, biographical approach (see Sartre's biography of Flaubert, entitled *The Family Idiot*, for the most masterful example) and the reductionism of Marxist reflection theory. Against the myth of the 'uncreated creator' ('*créateur incrée*') of Romantic ideology, the theory of fields reminds us that cultural producers are not free from social determination, and that they do not produce works in isolation, but, against the backdrop as propounded by the reflection theory, it emphasizes the fact that these determinations are mediated, refracted by the field, i.e., by a preexisting space of possibilities, which needs to be reconstructed in order to account for the principles of their esthetic choice.

These considerations raise two historical questions. The first concerns the autonomization process of these fields which, according to Bourdieu, derives from three conditions: the appearance of a group of producers specializing in a given activity (literature, painting, music, sport); the existence of

specific authorities playing a major role in the recognition and consecration process; and the formation of a market of symbolic goods, which reverses the temporal order that patronage established between demand and supply (Bourdieu, 1971c). Though the emergence of such a market was observed as early as the end of the eighteenth century, the autonomization process did not occur at the same time or in the same fashion in literature and painting: while the rise of publishing and the liberalization of printing in the early nineteenth century gave writers up to the ruthless law of the book market, sanctioned by sales figures, the art market was regulated strongly until the middle of the century by a single body, the Académie des Beaux-Arts, which assumed the monopoly on consecration through control of access to the Salon, the only public gathering where works were exhibited and sold (Bourdieu, 2013). The development of societies of artists, which resembled a kind of sales cooperative, and the bohemian style of life adopted by the cultural producers who were excluded from the academic system – and who multiplied in number as schooling became more and more common – nevertheless contributed to the emergence of a market on the margins of this system.

While the artistic field was being formed against the State, the guarantor of the existence of the academic system, the autonomy of the literary field needed to assert itself with regards to the market, which had challenged the monopoly of the Académie française over literary consecration. As Bourdieu describes in *The Rules of Art* (1992a), against the pole of large-scale production and circulation ruled by the logic of the market, a pole of small-scale production formed in the middle of the nineteenth century, imposing the primacy of the symbolic value of works, established on the basis of peers' specific criteria for assessing its aesthetic quality. The economy of symbolic goods is a 'reversed economy' that opposes to short-term profitability the slow process of recognition of works on the long term. The process is liable to lead to the 'canonization' of certain authors or artists through the integration of their works, turned into 'classics,' as part of the national and universal cultural heritage, namely through the intervention of the educational system (Bourdieu, 1993).

In France, this reversal was carried out under the Second Empire by the proponents of 'art for art's sake,' who were as opposed to 'social art,' that is, to engaged art, promoted by left-wing writers, as to the École du bon sens (School of common sense), which gathered 'bourgeois' authors, who were close to the dominant fractions of the dominant class. Bourdieu analyzes the conditions of the conquest of autonomy via this double break, carried out by Flaubert and Baudelaire. He qualifies the latter as a 'nomothete' in that he established a new *nomos* in the literary field, characterized by its independence from external powers – both economic or political and by anomie. In liberating themselves from the bourgeois in order to affirm the prevalence of pure aesthetics, and, concurrently, an ethos combining rigor at work and anticonformism, Baudelaire and Flaubert achieved a 'symbolic revolution' that had everlasting consequences for the literary field (Bourdieu, 1992b).

How do such 'symbolic revolutions' come about? This is the second historical question raised by the study of fields of artistic production, which is also relevant to the fields of science and even *haute couture* (Bourdieu, 1984: p. 200). The most in-depth formulation of this question is found in the lecture courses on Manet Bourdieu (2013b) delivered at the Collège de France. These symbolic revolutions are difficult to apprehend, due to the fact that our own categories of aesthetic perception result from them. It is therefore necessary to reconstitute the vision of the world that predated the revolution in order to bring to light the principles that previously drove academic aesthetics. This academic code, backed by the state monopoly, was destroyed by Manet. In opposition to the importance granted to technical mastery (perspective, relief, clair-obscur) and the 'finish,' Manet worked the canvas in its bidimensionality – which meant he would transgress the laws of perspective and relief – and ennobled the sketch, seen by the masters of the Académie as a mere first step before 'invention,' or the finishing process. In addition, Manet challenged the classical, socially driven, hierarchy of objects. He also replaced the omnipresent reference to history by a reference to the history of art, recognizable only by insiders. Such a preference led him, if not to empty the work of all of its meaning, then at least to render it ambiguous, therefore breaking with the principle of readability (or the message) and going so far as to engender bewilderment and contradictory interpretation – and, sometimes, scoffing and scandalized scolding.

A thorough analysis of the transformations in the conditions of production – at once technical (the invention of tubes of color that make it possible to paint outdoors), morphological (the growth of the artist population), and economic (the formation of a market running parallel to the academic system) – helps to avoid the biases of idealism and individualism, but does not suffice to explain the symbolic revolution achieved by Édouard Manet. A sociological analysis must take into account Manet's dispositions: the generous economic, cultural, and social resources he possessed as the son of a judge and descendant from a line of legal professionals, his training within the academic system, his vast pictorial culture, the freedom his private income gave him, and his dense network of relations, high society on one hand, and cafés and the bohemian crowd on the other. This extraordinary concentration of assets, linked to what Bourdieu calls a 'cleft habitus' split between the two poles (economic and cultural) of the field of power, predisposed Manet to accomplish this revolution. Like all 'symbolic revolutions,' this one was defined not by destruction but by the integration of that which preceded it: the break occurs in continuity, as illustrated by the practice of pastiche. As such, even in spheres where, in contrast to the bureaucratic world, positions are yet to be created, the most innovative practices fit in a dialectic relationship with the space of possibilities.

The field of *haute couture*, or high fashion, offered an equally rich ground for observing the logic of fields. Producing belief in brand fetishes according to a process Bourdieu compares to Mauss' conception of the way magic functions, the field was structured at the time by an opposition between 'rive droite' (Balmain) and 'rive gauche' (Scherer). Courrèges achieved a symbolic revolution by replacing discourse on fashion with a reflection on the lifestyle of the modern woman, which in his view was meant to be carefree, comfortable, and relaxed. Courrèges was thus synchronizing internal revolution and social

transformation. Such a synchronizing process is typical of symbolic revolutions according to Bourdieu (1984: pp. 196–206).

This model of analysis also serves to reflect both upon the conditions of existence needed for a relatively autonomous scientific field where peer judgments prevail over heteronomous logic (ideological or economic) and upon scientific revolutions (Bourdieu, 1976, 1991, 1997, 2001). The scientific field is different from cultural fields of production in that its public is made up primarily of peers. This characteristic exacerbates the competition for acquiring specific capital. Emphasizing this particularity enables Bourdieu to refute relativism and instead propose a 'historicist rationalism' founded on the theory of fields. Introducing social, political, and economic interests in the analysis of the conditions of scientific knowledge may indeed lead to relativism. Thomas Kuhn's model of scientific revolutions has paved the way to such a relativism (though Kuhn himself did not subscribe to this view). The sociology of science underwent a shift from rationalism founded on the professional norms that dictate the scholarly ethos (universalism, communalism, disinterest, skepticism) according to Robert Merton's functionalist model, to the relativism of the sociology of interests and of constructivist approaches. Though Bourdieu breaks with the enchanted vision of a scientific community acting in harmonious cooperation in the universal interest of science, he contends that fields' relative autonomy and their imposed rules guarantee the universal scope of scientific knowledge – all the while reintroducing the power relations and extrascientific interests that occur in these spheres. The scientific field produces an interest in disinterest that, linked to the collective control of peers, is one of the conditions of its autonomy, even though this autonomy is relative. Analyzing the way of the scientific field functions has an added value in that it contributes the reflexivity that scholars have with regards to their own activity.

The 'field effect' can be observed particularly in the highly conceptualized field of philosophy, where mastering references from the past is the necessary condition for access to recognition by peers. The analysis of Martin Heidegger's political ontology (1975, 1988), resituated in the field of the ideological production of his time, thus reveals the conservative, even reactionary ethicopolitical schemes which, like most of the important figures experiencing a decline in their position in the field of power, he shared with the 'völkisch' atmosphere and led to his adherence to Nazism. However, Bourdieu also insists on the transubstantiation that the process of philosophical conceptualization forced these figures to go through, euphemizing them to the point of unrecognizability: the field's specific censorship is at the origin of this operation of 'philosophical sublimation.' As such, the field effect provides 'an objective foundation to the illusion of absolute autonomy' (1988: p. 10), all the while making it possible to overcome the alternative between an ideological reading of a work and a purely philosophical reading – an alternative that commenters of Heidegger's work still debate today.

The study on Heidegger followed a large empirical study on the academic field, published in 1984 under the title *Homo Academicus*, which allowed Bourdieu to develop his research program on the scientific field (Bourdieu, 1992a). Focusing on the French academic field of the 1960s, the study is based on a prosopographic survey combining individuals' social properties and indicators of the positions they occupy in the field. The indicators distinguish between various kinds of capital within temporal (heteronomous) forms of social recognition and symbolic (autonomous) ones: academic power, scientific power, scientific prestige, intellectual notoriety. Institutions, mainly the University, research organizations, and the Collège de France are conceived in this model as variables (affiliation with a given institution as an indicator of an individual's position) and as agents in and of themselves in the field. A Multiple Correspondence Analysis, the best-suited method for field analysis, was applied to the prosopographic data. The academic field appears to be structured according to the opposition between a temporally dominant pole of faculties of law and medicine, and a temporally dominated pole of science faculties, with faculties of letters and the human sciences located between the two. Advantageous sites for observing the struggles between these two forms of academic power, temporal and symbolic, given their intermediary position, humanities faculties are divided between a pole of university academics more concerned with the reproduction of the institution, and therefore toward the exercise of temporal power within the cultural order, and a pole of agents oriented toward research, affiliated most often with prestigious but marginal institutions such as the Collège de France. The agents in this second group tend to be 'producers' rather than 'reproducers,' and are also more often internationally oriented than the first group. The study therefore made it possible to differentiate between two types of consecration, temporal and symbolic, which have equivalents in the fields of cultural production: public success and institutional recognition (prizes, Academies) on the one hand, and on the other, acknowledgment by peers and specialized critics (criticism being in and of itself a field, structured by homologous oppositions).

A field's degree of centralization varies depending on an institution's capacity to monopolize the power of reproduction within the limits of the institution such as the Catholic Church, the University, or the academic school of Beaux-Arts in France. When an institution holds a significant monopoly in a given field, one speaks of a 'body' ('corps') – in the sense of the *Corporatio* of Medieval Canonistas – rather than of a 'field' (Bourdieu, 1985a, 1999, 2013b). Competitive examinations, numerus clausus, are means of controlling access to a field and may potentially lead to the field's being transformed into a body, via the homogenization of social recruiting and the instilling of an 'esprit de corps.' However, it is rare that the social recruiting in a field is so homogenous, and differences in status often generate principles of structural opposition, such as those that arose between the Académie des Beaux-Arts and bohemian painters, or between theoretical law professionals and law practitioners, according to an opposition Bourdieu borrows from Weber.

One of the advantages of field theory as compared with the sociology of professions is that it considers activities – even when they have reached a certain degree of autonomy – as still being relatively heteronomous (for example, the way in which class relations are refracted in the field through the habituses) and as being more or less heterogeneous. This heterogeneity may result from work conditions and from status (for example, independent, salaried, and civil servant statuses may

coexist within the same area of activity) or from social recruiting (social origins or training that set Oxbridge alumni apart from others, for example). Masked by professional ideology, such splits are often inherent in the power relations that structure fields and the internal struggles at the root of their transformations. For instance, the field of law in France during the Ancien Régime differed among the State law practitioners, who formed the bureaucratic pole of law, law officers, who claimed a certain autonomy from the king's authority, and the lower clergymen of the law, who managed during the French Revolution to reverse the hierarchy of power relations with the Nobles of the Gown – to whom the first two groups belonged – and succeeded in imposing the notion of the Nation-state (Bourdieu, 2012).

Another characteristic that results from the process of field autonomization is the break between professionals and nonspecialists by which laypeople are excluded from activities because their judgment is delegitimized. The dispossession of the nonspecialists, which has its roots in the religious model wherein clergy are distinguished from laypeople, may be observed in numerous professional areas (law, medicine, architecture, science, sport) and particularly in the political field. In his studies on the latter, Bourdieu analyzes the field's phenomenon of representation and attendant delegation, a phenomenon that is at the core of the monopolization of politics by professionals (Bourdieu, 1977, 1981, 2000a).

Together with the economic field, the field of politics is a dominant field within the field of power, which, for its part, is structured by the opposition between holders of economic and political capital and those who possess mainly cultural capital (Bourdieu, 1979). With the help of neoliberal ideology, the economic field tends more and more to subordinate other fields, namely the political field (through New Public Management), the journalistic field (Bourdieu, 1996), and fields of cultural production (through the logic of concentration and merging acquisition of the culture industry, the effects of which Bourdieu studied in the French publishing industry during the 1990s; Bourdieu, 1999), by imposing its logic of profitability. Subjection to this logic threatens the autonomy of these fields in reinforcing the heteronomous pole to the disadvantage of the autonomous pole. It also raises the question of the relationships between fields, which will be examined in the following section.

Uses of the Concept of Field and Methodological Issues

Having provoked much empirical research, the concept of field brought to light methodological and theoretical questions we will briefly expose here. A good amount of this research dealt with the literary field, but significant research has also been carried out on other fields (the field of academics, politics, law, journalism, art, cinema, as well as on disciplinary fields such as philosophy, economy, sociology), and though much of this research has focused on the French case, several significant inquiries have also been conducted on other countries, such as Sergio Miceli's (1981) pioneering research on the conditions of emergence of the Brazilian intellectual field.

Analyzing a field requires three operations: the study of specific authorities; the distribution of the social properties of individuals according to the positions they occupy within the field; the reconstitution of the space of possibilities (and the space of effective positions taken). A series of prosopographic surveys were dedicated to French writers in the nineteenth and twentieth centuries, and included studies on authorities playing a role in literary recognition, and on schools or literary movements (Ponton, 1977; Thiesse, 1991; Sapiro, 1999). In the wake of Bourdieu's analyses of Flaubert and Baudelaire, other studies have focused on a central figure: Sartre (Boschetti, 1985), Beckett (Casanova, 1997), Apollinaire (Boschetti, 2001), Mallarmé (Durand, 2008), and Kafka (Casanova, 2011). Parallel research has been carried out on the German literary field at different periods (Joch and Wolf, 2005; Tommek and Bodgal, 2012).

Substantial research has also been dedicated to the French academic field. A comparison of the social recruitment of the literary and academic fields at the beginning of the twentieth century shows that writers are generally better endowed in terms of economic and cultural inherited capital than Parisian university academics (Charle, 1990). The articulation between the processes of autonomization, professionalization, and specialization during the Third Republic coincided with the emergence of a body of philosophy professors and a philosophy book market separate from the remainder of literary production (Fabiani, 1988). The contemporary field of economics is structured around the opposition between, on the one hand, 'pure' economists (neoclassical and regulationalist) endowed with strong scientific capital and, on the other hand, those with temporal resources provided by the 'political' positions they occupy as consultants, analysts, or within firms (Lebaron, 2000).

Bourdieu's reflection on the problem of political representation paved the way for a series of studies in political science on the social properties of political professionals and their variations in time and space (Gaxie, 1980), and on the larger group of individuals involved in political work: consultants, activists, pollsters, communication specialists, etc., whose growing importance stems from the increasing dependence of the political field with regard to the field of media. In addition, Bourdieu's conception of the State as a *metafield* provides new perspectives for renewing the study of public policy (Dubois, 2014).

More broadly, the question of the relationships between fields has been raised in a number of studies: subordination (for example, intellectual production did not begin to gain autonomy from the religious field until the eighteenth century), dependency, hierarchy (between disciplines within the academic field, for example), and exchange (between the literary and artistic fields). Studies on the politicization of the intellectual field at times of crisis and the resultant loss of autonomy, such as in the case of the Dreyfus Affair (Charle, 1990) and furthermore during the German Occupation (Sapiro, 1999), show that political stands ('position-takings') are strongly linked to the positions agents occupy in their field of reference, which confirms the refraction effect the field exerts on external constraints, all the while revealing the heteronomous role played by certain authorities such as the Académie française, located at the intersection of the intellectual field and the field of power. This structural homology between positions in the literary or intellectual fields and political stands is also manifested in the various forms of politicization at different poles of the literary and intellectual fields, depending on the nature of the position

(dominant or dominated), the degree of autonomy, and the degree of specialization insofar as the intellectual field is concerned (Sapiro, 2003a, 2009a). The heteronomous logic resulting from subordination to the religious and political fields were dealt with in the case of French Catholic intellectuals (Serry, 2004) and French Communist intellectuals (Matonti, 2005). The relations between the journalistic and economic fields were examined namely within the prism of the field of economic press (Duval, 2004; on the journalistic field, see also Benson and Neveu, 2005).

The historicity of fields raises the question of their genesis and temporality. Some scholars trace the start of the literary field's process of autonomization back to seventeenth century upon the appearance of lists of best writers, claims for royalties, and the officialization of the Académie française, granted by the king the power to legislate on linguistic and literary matters (Viala, 1985). The periodization proposed by Bourdieu in *The Rules of Art*, placing the conquest of literary autonomy in the mid-nineteenth Century, is therefore questioned (Saint-Jacques and Viala, 1994). Additionally, it may appear paradoxical that the autonomization of the literary field goes through a reinforced dependency on the State symbolized by the officialization of the Académie française (Jouhaud, 2000). The same is true for the artistic field, which the State raised to the status of liberal art upon creating the Académie des Beaux-Arts. However, in both cases, though there exists a specific authority, as well as more or less specialized producers in the field, the third condition – the existence of the market, which makes it possible to escape state control – is lacking.

The fields of cultural production are thus situated between the state and the market, with more or fewer constraints depending on the regime and economic conditions. Though the appearance of the market allowed for autonomy from the State, the State is capable of providing protection from economic constraints in the form of subvention policy (Sapiro, 2003b). Research on Communist regimes, wherein the power of consecration was monopolized by a body controlled by the Communist Party, the Union des écrivains, has nevertheless shown that forms of relative autonomy could manifest themselves even in contexts of strong heteronomy and state dependency (Dragomir, 2007). Moreover, autonomous logic exists even in the most market-dependent fields of cultural production such as the field of cinema, also structured upon the opposition between the poles of large and small-scale circulation (Duval, 2006). The small-scale circuit in France benefits from heavy state support; as in the small-scale publishing field. Bourdieu (1992b) analyzed the process of autonomization of a small-scale pole of production with regards to the market; the study of literary trials provides rich grounds for understanding the heteronomous logic that continued to weigh on literature and the social expectations from writers, as well as the processes of progressive recognition of the field's autonomy from the State from the twentieth century onward (Sapiro, 2011). The process of autonomization in the literary field should, furthermore, be situated within the context of the division of intellectual labor that was on the rise in the nineteenth century (Sapiro, 2012a).

As mentioned, the relatively autonomous temporality of fields becomes blurred during moments of crisis or politicization, which may produce synchronization effects. These moments raise the question of how changes occur. Do changes obey previously instated logic, as certain research suggests (Sapiro, 1999; Gobille, 2005)? By the fluidity their circumstances engender, do crises create their own dynamics undetermined by the field's previous states (Dobry, 1992)? This rationale, particular to crises, is observed in the reworking of alliances and the ensuing reconfiguration of power relations, but the logic by which these reconfigurations occur are not random. Rather they may be explained according, on one hand, to the issues at stake in the field, and, on the other, to the dispositions and resources of its agents (Sapiro, 2012b).

The concept of field is a powerful heuristic tool for comparison, whether in diachronically comparing different states of the same field (for example, the French field of power in the 1970s and in the 1990s, when economic power took precedence over cultural power; Denord et al., 2011), or in synchronically comparing the principles of structuring for one field in two countries (for example, the place of translations in the French and American fields of publishing; Sapiro, 2010). However, in order not to fall into the trap of methodological nationalism, international comparison must take into account, on one hand, the exchanges between national fields, and, on the other, the power relations in which they are embedded (Sapiro, 2012c). These international power relations indeed condition the circulation of symbolic goods and the models between national cultures (Heilbron, 1999; Sapiro, 2008). The phenomena of importation and reception must also be brought into the perspective of the target culture (Bourdieu, 2002), as illustrated by research on the importing into France of legal theories (Hauchecorne, 2009) or literature from Eastern Europe during the Communist period (Popa, 2010). The study of Latin America's introduction of neoliberal economics and the philosophy of human rights has brought to light the outcomes of the process of globalization (Dezalay and Garth, 2002).

Transnational approaches have nevertheless raised the question of the geographical limitations of fields. These problems have come to the fore namely in connection with linguistic areas, and particularly the Francophone literary space (Bourdieu, 1985b; Aron, 2005). Later, other transnational spaces were taken into account. Though much research on fields has been situated within a national framework, Bourdieu never specified that fields were necessarily circumscribed by a Nation-state's borders. Borders are not given; rather they evolve with time and are constantly being reexamined. As a result, it is up to the individual scholar to construct them. The nationalization of professional fields and fields of cultural production is a historical fact connected to the monopolization of education and to state control of access to professions (varying from country to country) as well as to the construction of national identities (Sapiro, 2013). However, these national cultures were formed with regard to one another, and they rapidly made up an international space ruled by bodies such as the League of Nations, then the United Nations, and, for culture and science, the International Committee on Intellectual Cooperation, taken over by UNESCO after the Second World War – all bodies that worked in favor of the intercountry circulation of people and organizational models (Sapiro, 2009b). At the same time, markets emerged, with borders going far beyond those of the Nation-state (for example, linguistic areas constituted spaces wherein printed works in the lingua franca, English, French,

Spanish, Portuguese, or Arabic could circulate). Bourdieu (2000b) himself analyzed the formation process of a world economic field dominated by multinational companies and characterized, among other things, by delocalizations and capital flows. Moreover, the legal field, strongly linked to the Nation-state, became highly internationalized when its heteronomous pole was put at the service of the market (Dezalay, 1990; on the use of Bourdieuan theory for studying the legal field, see Dezalay and Madsen, 2012). Though the political field remains strongly national, European construction favors the emergence of a European bureaucratic field (Georgakakis, 2012) and of a European legal field, though both are still highly dependent on the Nation-states (Vauchez, 2008). The emergence of more or less autonomized transnational spaces depends on the existence of places of exchange and specific authorities. These entities differentiate fields from markets; examples are international scientific Congresses or the Nobel Prize in literature (Casanova, 1999). In parallel way, the space in which symbolic revolutions are received may circumscribe a transnational field. Temporal lags and the logics of retranslation of issues at stake from one field of reception to another can be regarded as indicators of the autonomy of national fields.

Apart from Brazil, the international reception of the concept of field, beginning in the 1980s, was initially due to literature specialists, in Belgium (Dubois and Durand, 1988), Germany (Jurt, 1992, 1995), Italy (Boschetti, 1985), and Israel (Even-Zohar, 1990). At the same time, the notion was also introduced in Greece and Russia. Used much less frequently in the United States than the concepts of cultural capital and habitus, the concept of field – as Bourdieu theorized it – did not begin to be appropriated by American scholars as a research program until after the sociologist's death in 2002. The concept contributed to the renewal of studies on imperialism and colonialism (Steinmetz, 2008; Go, 2008). In this perspective, colonial states are conceived as *metafields* including state fields within colonies, in which relations of power and competition between different fractions of the field of colonial power coincide in a struggle to define indigenous politics (Steinmetz, 2008). The concept of field is also more and more frequently introduced in the field of international relations (Bigo and Madsen, 2011), where it is used namely in order to think diplomatic relationships such as a *metafield* (Adler-Nissen, 2011). In the sociology of law, a study was carried out on the field of international penal justice, located at the intersection between three transnational fields–interstate relations, defense of Human Rights, and penal justice, which is undergoing internationalization (Dixon and Tenove, 2013).

Organizational Fields and Strategic Action Fields

The concept of field has met with two main theoretical redevelopments more or less inspired by Bourdieu's theory, around the notions of 'organizational fields' and of 'strategic action fields.'

The neoinstitutionalist theory of 'organization fields' developed by Paul DiMaggio and Walter Powell (1983, 1991) aims to show the phenomenon of institutional isomorphism between organizations within the same field. According to DiMaggio and Powell, the processes of rationalization such as Weber described is less dependent on competition and the search for efficiency as on specifically institutional factors. Constraint (via authoritarian means, the legal system, or orders) imitation in situations of uncertainty and normative pressure linked to professionalization are the three mechanisms that favor these fields' homogenization. It is therefore possible to predict this tendency for resemblance based on the type and degree of dependence between organizations within the same field, the degree of uncertainty as regards ties between means and objectives, and the degree of social recruitment of professionals with the same academic training and/or the same engagement with professional bodies. The empirical studies upon which sociologists have formed their analysis are the organizational models used for the production of high-end cultural services, models that appeared in the United States at the end of the nineteenth century (DiMaggio, 1981), and the progressive homogenization of American academic publishing (Coser et al., 1982). DiMaggio also created the notion of 'institutional entrepreneur,' which refers to the foundational figures in emerging fields.

Largely fueled by Bourdieu's field theory, the concept of 'strategic action field' developed by Neil Fligstein and Doug McAdam (2012) combines the neoinstitutionalist approach with collective action theories in order to consider reproduction and change on the mesolevel of social order. From the Bourdieuan conception, Fligstein and McAdam retain the ideas of a field's specific issues at stake, rules of the game, the unequal positions of agents and the link between these unequal positions and agents' vision of the world. However, these strategic action fields are less stable than the fields in Bourdieu's theory; their boundaries shift depending on the definition of the situation and the issues at stake in the field. The fields may be ordered along a continuum based on their degree of consensus. In the case of declared conflicts, struggles are liable to bring about a new order. These struggles are primarily fought between 'incumbents' and 'challengers,' and are regulated by 'governance units' that control a field's functioning and generally tend to conserve the existing power relations in the field. It is possible to recognize here the equivalent of the authorities, the crucial role of which Bourdieu brings to light in fields of cultural production or other types of regulating bodies. Sharing the agonistic conception developed by Bourdieu, the model of strategic action fields more clearly classifies the foundations of collective action between coercion, competition and cooperation, and thus more clearly delineates fields organized hierarchically and wherein different forms of coalition prevail. The question of the relation between individuals and the field, problematized by Bourdieu by means of the concepts of habitus, strategy, and the sense of the game, is considered here in terms of the 'skills' actors possess in order to act in a field. Far from being autarkic, these strategic action fields maintain relations with other fields that differ according to distance/proximity, independence/dependence (in certain cases, interdependence), and the distinction between state and nonstate fields, with the state being considered, as in Bourdieu's theory, as a set of fields (or a *metafield* in Bourdieu's terms). This field environment often plays a major role in change, through, for example, the importation of models from another field. Transformations are frequently provoked by exogenous shocks that provide challengers with new

opportunities, but they require a mobilization process along with use of organizational resources and efficient means of contest. For instance, Rosa Parks' refusal to give up her seat to a white man in a bus in 1955 and her subsequent arrest were not new events in themselves, but civil rights leaders succeeded in turning the episode into an opportunity for mobilization against injustice by convincing black ministers in Montgomery, Alabama, to urge congregants to protest against her arrest. During 'episodes of contention,' one often observes recourse to innovative forms of collective action, as well as a '(re)framing' of the established vision of the world, liable to lead to a new 'institutional settlement.'

The concept of field thus appears to be a heuristic tool for exploring the mesolevel of social activities. While different uses of the concept have been developed in relation with organizational and collective action theories, Bourdieu's field theory continues to function as a research program in different domains and disciplines, raising new questions about colonial and transnational fields, as well as relations between fields.

See also: Division of Labor: History of the Concept; Global History: Universal and World; Habitus: History of a Concept; Institutionalism and Neo-institutionalism: History of the Concepts; Intellectuals: History of the Concept; Laicization, History of; Nations and Nation-States in History; Prosopography (Collective Biography); Relativism: Philosophical Aspects; Scientific Revolution, History and Sociology of; Social Inequality in History (Stratification and Classes); Social Science Professions and Professionalization; Social Structure: History of the Concept; Weber, Max (1864–1920).

Bibliography

Adler-Nissen, R., 2011. Inter- and transnational field(s) of power. On a field trip with Bourdieu. International Political Sociology 5, 327–345.

Aron, P., 2005. La littérature en Belgique francophone de 1930–1960: débats et problèmes autour d'un 'sous-champ'. In: Einfalt, M., Erzgräber, U., Ette, O., Sick, F. (Eds.), Intégrité intellectuelle/Intellektuelle Redlichkeit. Mélanges en l'honneur de Joseph Jurt. Memmingen, Universitätsverlag, Winter Heidelberg.

Bigo, D., Madsen, M.R. (Eds.), 2011. A Different Reading of the International: Pierre Bourdieu and International Studies. International Political Sociology, vol. 5 (3), pp. 219–224.

Benson, R., Neveu, E. (Eds.), 2005. Bourdieu and the Journalistic Field. Polity Press, Cambridge, UK.

Boschetti, A., 1985. Sartre et "Les Temps Modernes". Une entreprise intellectuelle. Minuit, Paris.

Boschetti, A., 2001. La Poésie partout. Apollinaire, homme-époque (1898–1918). Seuil, Paris.

Bourdieu, P., 1966. Champ intellectuel et projet créateur. Les Temps modernes (246), 865–906.

Bourdieu, P., 1971a. Genèse et structure du champ religieux. Revue française de Sociologie XII, 295–334. English trans.: 1991. Genesis and structure of the religious field. Comparative Social Research (13), 1–44.

Bourdieu, P., 1971b. Une interprétation de la théorie de la religion selon Max Weber. Archives européennes de sociologie XII (1), 3–21. English trans. (revised version): 1987. Legitimation and Structure Interests in Weber's Sociology or Religion. In: Max Weber, Rationality and Modernity. Whimster, S., Lash, S. (Eds.), Allen & Unwin, London, pp. 119–136.

Bourdieu, P., 1971c. Le marché des biens symboliques. L'Année sociologique (22), 49–126. English trans.: 1985. The market of symbolic goods (R. Swyer, Trans.). Poetics 14 (1–2), 13–44.

Bourdieu, P., 1976. Le champ scientifique. Actes de la recherche en sciences sociales (2–3), 88–104. English trans.: 1975. The specificity of the scientific field and the social conditions of the progress of reason. Social Science Information XIV (6), 19–47.

Bourdieu, P., 1977. Questions de politique. Actes de la recherche en sciences sociales (16), 55–89.

Bourdieu, P., 1979. La Distinction. Minuit, Paris. English trans: 1984. Distinction. (R. Nice, Trans.). Harvard UP, Cambridge, MA.

Bourdieu, P., 1981. La représentation politique. Actes de la recherche en sciences sociales (36–37), 3–24. English trans.: 1991. Political representation: elements for a theory of the political field. In: Language & Symbolic Power. Polity Press, Cambridge, pp. 171–202.

Bourdieu, P., 1984. Questions de sociologie, 1980, first ed. Minuit, Paris. pp. 113–120.

Bourdieu, P., 1985a. Effet de champ et effet de corps. Actes de la recherche en sciences sociales (59), 73.

Bourdieu, P., 1985b. Existe-t-il une littérature belge ? Limites d'un champ et frontières politiques. Études de lettres III, 3–6.

Bourdieu, P., 1988. L 'Ontologie politique de Martin Heidegger. Minuit, Paris. English trans.: 1991. The Political Ontology of Martin Heidegger (P. Collier, Trans.). Polity Press, Cambridge.

Bourdieu, P., 1991. The peculiar history of scientific reason. Sociological Forum 6 (1), 3–26.

Bourdieu, P., 1992a. Homo academicus, 1984, first ed. Minuit, Paris. English trans.: 1988. Homo Academicus (P. Collier, Trans.). Polity Press, Cambridge.

Bourdieu, P., 1992b. Les Règles de l'art. In: Genèse et structure du champ littéraire, Seuil, Paris, ed. Paperback 1998. Seuil 'Points', Paris. English trans.: 1996. The Rules of Art (S. Emanuel, Trans.). Polity Press-Stanford University Press, Cambridge-Stanford.

Bourdieu, P., 1993. The Field of Cultural Production. Edited and Introduced by Randal Johnson. Polity Press, Cambridge.

Bourdieu, P., 1996. Sur la télévision, Liber-Raisons d'agir, Paris. English trans.: 1998, On Television & Journalism (P. Parkhurst Ferguson, Trans.). The New Press/Pluto Press, New York/London.

Bourdieu, P., 1997. Les Usages sociaux de la science. INRA, Paris.

Bourdieu, P., 1999. Le fonctionnement du champ intellectuel. Regards sociologiques (17–18), 5–27.

Bourdieu, P., 2000a. Propos sur le champ politique. Presses universitaires de Lyon, Lyon.

Bourdieu, P., 2000b. Les Structures sociales de l'économie. Seuil, Paris, 273–280.

Bourdieu, P., 2001. Science de la science et réflexivité. Liber-Raisons d'agir, Paris. English trans.: 2004. Science of Science and Reflexivity (R. Nice, Trans.). Polity Press, Cambridge.

Bourdieu, P., 2002. Les conditions sociales de la circulation internationale des idées (1st published 1989). Actes de la recherche en sciences sociales (145), 3–8. English trans.: 1999. The social conditions of the international circulation of ideas. In: Shusterman, R. (Ed.), Bourdieu: A Critical Reader. Blackwell Publishers, Oxford-Malden, MA, pp. 220–228.

Bourdieu, P., 2012. Sur l'État. Cours au Collège de France (1989–1992). Seuil/Raisons d'agir, Paris.

Bourdieu, P., 2013a. Séminaires sur le concept de champ, 1972–1975. Actes de la recherche en sciences sociales (200), 4–37.

Bourdieu, P., 2013b. Manet. Une révolution symbolique. Seuil-Raisons d'agir, Paris.

Casanova, P., 1997. Beckett, l'abstracteur. Seuil, Paris.

Casanova, P., 1999. La République mondiale des lettres. Seuil, Paris. English trans.: 2005. The World Republic of Letters. Harvard University Press, Cambridge, MA.

Casanova, P., 2011. Kafka en colère. Seuil, Paris.

Charle, C., 1990. Naissance des "intellectuels", 1880–1990. Minuit, Paris.

Coser, L.A., Kadushin, C., Powell, W.W., 1982. Books: the culture and commerce of publishing. Basic Books, New York.

Denord, F., Lagneau-Ymonet, P., Thine, S., 2011. Le champ du pouvoir en France. Actes de la recherche en sciences sociales (190), 24–57.

Dezalay, Y., 1990. The big bang and the law: the internationalization and restructuration of the legal field. Theory, Culture & Society 7, 279–293.

Dezalay, Y., Garth, B., 2002. The Internationalization of Palace Wars: Lawyers, Economists, and the Contest to Transform Latin American States. University of Chicago Press, Chicago.

Dezalay, Y., Madsen, M.R., 2012. The force of law and lawyers: Pierre Bourdieu and the reflexive sociology of law. Annual Review of Law and Social Science (8), 433–452.

Dixon, P., Tenove, C., 2013. International criminal justice as a transnational field: rules, authority and victims. International Journal of Transitional Justice 1–20.

DiMaggio, P.J., Powell, W.W., 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. American Sociological Review 48, 147–160.

DiMaggio, P.J., 1981. Cultural entrepreneurship in nineteenth century. part 1: the creation of an organizational base for high culture in America. Media, Culture and Society 4, 33–50.

Dobry, M., 1992. Sociologie des crises politiques. Presses de la Fondation nationale des sciences politiques, Paris.

Dragomir, L., 2007. Une institution littéraire transnationale à l'Est: l'exemple roumain. Belin, Paris.

Dubois, J., Durand, P., 1988. Champ littéraire et classes de textes. Littérature (70), 5–23.

Dubois, V., 2014. L'action de l'État, produit et enjeu des rapports entre espaces sociaux. Actes de la recherche en sciences sociales (201–202), 38–43.

Durand, P., 2008. Mallarmé. Seuil, Paris.

Duval, J., 2004. Critique de la raison journalistique. Les transformations de la presse économique en France, Seuil, Paris.

Duval, J., 2006. L'art du réalisme. Le champ du cinéma français au début des années 2000. Actes de la recherche en sciences sociales (161–162), 96–115.

Even-Zohar, I., 1990. Polysystem studies. Poetics Today 11 (1).

Fabiani, J.-L., 1988. Les Philosophes de la République. Minuit, Paris.

Fligstein, N., McAdam, D., 2012. A Theory of Fields. Oxford University Press, Oxford-New York.

Foucault, M., 1969. L'Archéologie du savoir. Gallimard, Paris. English trans.: 2002. The Archaeology of Knowledge. Routledge, London and New York (1st translated 1972).

Gaxie, D., 1980. Les logiques de recrutement du personnel politique. Revue française de science politique (1), 5–45.

Georgakakis, D. (Ed.), 2012. Le Champ de l'eurocratie. Une sociologie politique du personnel de l'UE. Economica, Paris.

Go, J., 2008. Global fields and imperial forms. Sociological Theory 6 (3), 201–229.

Gobille, B., 2005. Les mobilisations de l'avant-garde littéraire française en Mai 1968. Capital politique, capital littéraire et conjoncture de crise, Actes de la recherche en sciences sociales (158), 30–53.

Hauchecorne, M., 2009. Le "professeur Rawls" et "le Nobel des pauvres". La politisation différenciée des théories de la justice de John Rawls et Amartya Sen. Actes de la Recherche en Sciences Sociales (176–177), 94–113.

Heilbron, J., 1999. Towards a sociology of translation: book translations as a cultural world-system. European Journal of Social Theory 2 (4), 429–444.

Joch, M., Wolf, N. (Eds.), 2005. Text und Feld. Bourdieu in der literaturwissenchaftlichen Praxis. Max Niemeyer Verlag, Tubingen.

Jouhaud, C., 2000. Les Pouvoirs de la littérature. Histoire d'un paradoxe. Gallimard, Paris.

Jurt, J., 1992. Autonomie ou hétéronomie: le champ littéraire en France et en Allemagne. Regards sociologiques (4), 3–16.

Jurt, J., 1995. Das literarische Feld. Das Konzept Pierre Bourdieus in Theorie und Praxis. Wissenschaftliche Buchgesellschaft, Darmstadt.

Lebaron, F., 2000. La croyance économique. Les économistes entre science et politique. Seuil, Paris.

Levi Martin, J., 2003. What is field theory? American Journal of Sociology (109), 1–49.

Matonti, F., 2005. Intellectuels communistes. Essai sur l'obéissance politique. La Nouvelle Critique (1967–1980). La Découverte, Paris.

Miceli, S., 1995. Sergio Miceli, Les Intellectuels et le pouvoir au Brésil (1920–1945). Editions de la Maison des sciences de l'homme (1st published 1981), Paris.

Ponton, R., 1977. Le Champ littéraire de 1865 à 1906 (recrutement des écrivains, structures des carrières et production des æuvres), thèse de doctorat de 3e cycle. Université Paris V, Paris.

Popa, I., 2010. Traduire sous contraintes. Littérature et communisme (1947–1989). CNRS Éditions, Paris.

Saint-Jacques, D., Viala, A., 1994. À propos du champ littéraire. Histoire, géographie, histoire littéraire. Annales. Histoire, Sciences Sociales 49 (2), 395–406.

Sapiro, G., 1999. La Guerre des écrivains, 1940–1953. Fayard, Paris. English Trans.: 2014. French Writers' War. Duke UP, Durham.

Sapiro, G., 2003a. Forms of politicization in the French literary field. Theory and Society (32), 633–652.

Sapiro, G., 2003b. The literary field between the state and the market, Poetics. Journal of Empirical Research on Culture, the Media and the Arts 31 (5–6), 441–461.

Sapiro, G. (Ed.), 2008. Translatio. Le marché de la traduction en France à l'heure de la mondialisation. CNRS Éditions, Paris.

Sapiro, G., 2009a. Modéles d'intervention politique des intellectuels. Le cas français. Actes de la recherche en sciences sociales (176–177), 8–31. Spanish Trans.: 2011. "Modelos de intervención política de los intelectuales: El caso francés", Prismas. Revista de historia intelectual (Argentina) 15, 129–154. Portuguese trans.: 2012. Modelos de intervenção política dos intellectuais: o caso francês. Revista Pós Ciências Sociais (Brazil) 17, 19–50; German Trans.: 2012. Formen intellektuellen Engagements: Der Fall Frankreich. In: Mejstrik, A., Hübel, T., Wadauer, S. (Eds.), Die Krise des Sozialstaats und die Intellektuellen. Campus Verlag, Frankfurt/New York, pp. 83–102.

Sapiro, G. (Ed.), 2009b. L 'Espace intellectuel en Europe, XIXe–XXIe siècles: de la formation des Etats-nations à la mondialisation. La Découverte, Paris.

Sapiro, G., 2010. Globalization and cultural diversity in the book market: the case of translations in the US and in France. Poetics 38 (4), 419–439.

Sapiro, G., 2011. La Responsabilité de l'écrivain. Littérature, droit et morale en France (XIXe–XXIe siècle). Seuil, Paris.

Sapiro, G., 2012a. Autonomy revisited: the question of mediations and its methodological implications. Paragraph 35, 30–48. http://www.euppublishing.com/doi/abs/10.3366/para.2012.0040.

Sapiro, G., 2012b. Structural history and crisis analysis: the literary field during WWII. In: Gorski, P. (Ed.), Bourdieu and Historical Analysis. Duke UP, Durham.

Sapiro, G., 2012c. Comparaison et échanges culturels: le cas des traductions. In: coll., Faire des sciences sociales. Comparer, vol. 2. Editions de l'EHESS, pp. 193–221.

Sapiro, G., 2013. Le champ est-il national ? La théorie de la différenciation sociale au prisme de l'histoire globale. Actes de la recherche en sciences sociales (200), 70–85.

Serry, H., 2004. Naissance de l'intellectuel catholique. La Découverte, Paris.

Steinmetz, G., 2008. The colonial state as a social field: ethnographic capital and native policy in the German overseas empire before 1914. American Sociological Review 73 (4), 589–612.

Thiesse, A.-M., 1991. Ecrire la France. La littérature régionaliste, de la Belle Epoque à la Libération. Presses universitaires de France, Paris.

Tommek, H., Bodgal, K.-M., 2012. Transformationen des literarischen Feldes in der Gegenwart. Sozialstruktur, Medien-Okonomien, Autorpositionen. Synchron, Heidelberg.

Vauchez, A., 2008. The force of a weak field: law and lawyer in the government of the European union. International Political Sociology (2), 128–144.

Viala, A., 1985. Naissance de l'écrivain. Sociologie de la littérature à l'âge classique. Minuit, Paris.

Fieldwork in Social and Cultural Anthropology

Moshe Shokeid, Tel Aviv University, Tel Aviv, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article presents the origins of the method of fieldwork in European and American anthropology. It introduces the practice of fieldwork – 'participant observation' – as developed in the study of small, remote societies, during the era of colonialism. The article proceeds to discuss later transformations in ethnographic research: moving away from the colonial territories to fieldwork sites in Western industrial societies. The choice of new types of communal and institutional settings in urban and globalized spaces of social life also instigated major changes of methodology and theory in ethnographic projects. These include also a growing awareness of the impact of the researcher's presence during fieldwork, his/her identity, as well as the expansion of the variety of the types of data considered in ethnographic projects.

Origins of a Method

Fieldwork is a term that has been employed for nearly a century by social–cultural anthropologists as a major methodological tool and a profound professional experience that leaves its mark on their lives throughout their careers. It designates a complete concentration on observations in their chosen site of research, for a considerable period of time, usually of no less than 6 months and up to a few years. Originally, this engagement involved a total separation from the researcher's ordinary life and his/her immersion into a remote and an alien social environment. Anthropological fieldwork is known also as the 'ethnographic method.'

The idea of fieldwork owes its origin to the interests of scientists at the turn of the nineteenth century concerning the evolution of the human species, as much as to the expansion of European colonialism that brought under its rule vast territories inhabited by unknown 'exotic' peoples. But no less, that movement coincided with a growing awareness in the United States about the vanishing cultures of Native Americans. The new science of anthropology that intended to satisfy these interests was at its first stages mostly speculative, based on secondary sources of data collected by missioners, traders, colonial administrators, travelers, adventurers, etc. The practice of prefieldwork anthropologists was nicknamed 'armchair anthropology.' Its most celebrated as well as discredited representative remains James Frazer whose monumental work was all based on library material.

The invention of fieldwork as the characteristic method of anthropology, as well as of its corollary aspect of 'participant observations,' has often been attributed to Bronislaw Malinowski, the British–Polish scholar, who is also considered among the founders of the school of functionalism in anthropology. His claim to fame as a fieldworker is based on his long stay, of more than 2 years, during World War I, among the Trobriand islanders of Melanesia. Although he was not the first to employ the method engaging participant observation, nevertheless, his linguistic skills, the meticulous ethnographies he produced, and the introductory chapter on his method of inquiry that opened his most famous work (*Argonauts of the Western Pacific*, 1922) have become the hallmark and quintessential model of fieldwork for almost half a century.

The fieldwork projects carried out by Malinowski, his colleagues, and the later generations of younger anthropologists entailed observations of the minutiae of everyday life in the various domains of social behavior and cultural traditions. In sum, it was supposed to record the totality of the social experience in the studied society. The essence of fieldwork was encapsulated, for example, in the following statements by representatives of three generations in British anthropology:

> This goal is, briefly, to grasp the native's point of view, his relation to life, to realize his vision of his world. (Malinowski, 1922: p. 25)
>
> First, in my estimation, is the point that fieldwork in the empirical mode remains the *sine qua non* both for the testing of theory and, what is more important, for making new discoveries. (Fortes, 1978: p. 24)
>
> 'Good' fieldworkers are those who are prepared to have good 'trips,' that is, suspend as far as possible their own social conditioning in order to have sensory and mental knowledge of what is really happening around and to them. (Turner, 1985: pp. 205–206)

In the United States, the method of ethnographic fieldwork was initiated by Franz Boas and his students most of whom had until World War II concentrated their work in the not too far away reservations of Indian tribes. Their trips to the fields of research had rarely involved a long departure from their ordinary life and work. They were also far more dependent on key informants who provided them with data on specific subjects. Consequently, they were not as deeply engaged in the routine of everyday life of the people they studied. The ethnographies by Alfred Kroeber and Robert Lowie are good examples of the achievements and deficiencies of the tradition of short fieldwork trips and reliance on informants. Margaret Mead, probably the most celebrated among Boas' students, who went to Samoa in a style of research closer to the British tradition (1928), was later seriously criticized for her poor work as a fieldworker. It was mostly since the mid-1940s that American anthropologists have been equally engaged in intensive fieldwork projects away from home. Compared with the British tradition, American anthropologists, since the early stages, have developed a distinctive orientation in their fieldwork projects, more broadly directed toward inquiries of culture and personality (transformed in later years to domains of

experiential, interpretative, and symbolic anthropology, for example, in the works by Clifford Geertz (1973), Turner (1967), and Turner and Bruner (1986)).

The method of fieldwork was adopted in most Western European countries, though in France the interests of leading anthropologists (Lévi-Strauss in particular) were far more concentrated in oral and textual traditions such as tribal myths of creation. They attributed little importance to the ethnographer's experiences *in vivo*. The ethnographic tradition was also adopted outside Europe and the United States (in India in particular).

The Practice of Fieldwork

As a first step in their project, fieldworkers in the Malinowskian tradition have always been acutely aware that they must develop *rapport* and friendly relationships with the people in their chosen field. They are expected to listen attentively, and witness whatever happens around them in the field, without disturbing the natural flow of events. As they consolidate their position with locals, they usually add information through more active modes of communication (such as initiating conversations, interviews, collecting statistical data, taking photographs, etc.), but they are careful not to instigate a flow of information into preconceived patterns of their own making. They invest much of their time recording this material in their notebooks or computers. The bulk of that store accumulates into their most treasured fieldwork product – 'fieldnotes' (to be later transformed into an ethnography). The hallmark of anthropology, fieldwork had also come to differentiate anthropology from sociology. That major difference became more prominent a few decades later when anthropologists were no longer confined to the study of remote Third World societies.

The practice of fieldwork has raised many, as yet not fully resolved, problems of methodology and ethics. Considered a 'method,' it is assumed to maintain a standardized guide for professional behavior, as well as technical rules concerning the collection and recording of data. But throughout the history of anthropology, the craft of fieldwork was practiced mostly by individuals who went on their own to explore new research sites where they confronted conditions and constraints not envisioned by their teachers and colleagues. For many years, the teaching of anthropology in many universities did not include specific courses in methodology. The apprenticeship in anthropology was gained through more informal methods such as a close familiarity with the relevant corpus of ethnographies (and particularly those associated with the candidates' departments, for example, 'The Manchester School'), attending presentations and seminars by colleagues who returned from the field, communication with those who had already gone through the experience, and finally, through the good or bad fortunes of the novices on their arrival in the field and their success in establishing rapport with their native hosts. No surprise, fieldwork had been often described as the major *rite de passage* for students of anthropology. Many anthropologists have been engaged in one major fieldwork project only. But it seems that more ethnographers have gradually become involved in a few fieldwork ventures throughout their career.

Ethnographic work in the various sites of fieldwork, in the early days and today, has been based on an ethos of trust. Anthropologists, mostly engaged in a lone journey, have never been able to implement a procedure of testing the accuracy and truthfulness of fieldnotes reported by colleagues and students. Rarely did fieldworkers choose to return to a field studied already by another ethnographer. The few cases of contested reports have gained much attention, the controversy between Oscar Lewis and Robert Redfield who studied the same community of Tepotzlan, and in particular, Derek Freeman's (1983) refutation of Margaret Mead's observations in Samoa. But long-term research, carried out by anthropologists whose continuing interest and commitment to a particular field has been supported by sufficient funds or easy access, also offered a strategy of verification of observations made at earlier trips.

Although anthropologists have been expected to master the language of the society they studied, that prerequisite was rarely fully achieved. Constraints of time and linguistic skills often impelled them to rely on 'informants' as interpreters until they gained a better command of communication in the native languages. The length of stay in the field was usually determined by budgetary as much as by personal circumstances. But an informal norm of 1 year seems to have dominated mainstream fieldwork projects.

How fieldworkers present their role, and to what extent their subjects comprehend the purpose and consequences of their work, remains an unsolved ethical problem. However, some norms that pertain to a universal code of behavior have been institutionalized at an early stage, such as the disguise of the community studied. More rigorously maintained is the tradition of disguising the identity of individual participants. But no forcing regulations were designed in regard to the rights of privacy in recording and disclosing of intimate information collected during fieldwork. Anthropologists were actually left to make their own decisions concerning ethical issues under the assumption of a deeply felt responsibility toward 'their people.' This dilemma was addressed, for example, by Geertz (1968: p. 157): "A professional commitment to view human affairs analytically is not in opposition to a personal commitment to view them in terms of a particular moral perspective." In any case, fieldworkers were expected to record in their fieldnotes, without censoring, as much as they could remember from their daily observations and conversations.

During many years of practice, however, anthropologists informed their colleagues, also in a more public fashion, about the circumstances they encountered during fieldwork. The early publications had mostly revealed the personal tribulations of individuals in their respective fields (Powdermaker, 1966; Freilich, 1970). But certain specific subjects of methodology had engaged a more rigorous professional treatment. For example, 'the extended case method' (Gluckman, 1958), though mostly a method of ethnographic presentation, suggested a follow-up strategy that stimulated the fieldworker to concentrate his/her observations, thereby revealing the hidden connection between disjoined events. Another theme that gained much attention concerned the various types of informants one might meet at entrance to a new field, and the strategies and precautions one might consider working with them (Casagrande, 1960; Shokeid, 1988). Detailed manuals and reference materials for the various aspects of fieldwork

have gradually developed into books of methodology so long absent (Epstein, 1967; Wax, 1971; Agar, 1996; Ellen, 1984; Shaffir and Stebbins, 1991; and Wolcott, 1999). A more meticulous approach came to fore examining the methods of recording and safekeeping of fieldnotes, the most valuable product of fieldwork (Sanjek, 1990). Also, conferences dedicated to specific methodological issues have become frequent events (such as 'fieldwork at home,' 'long-term research').

Transformations in Fieldwork Research

The idea of fieldwork had gone through immense changes since the days of Malinowski and the generations that established the ethnographic tradition of participant observations in remote, apparently homogeneous societies, clearly identified within territorial and cultural boundaries. The consequences of World War II, the decline of colonialism on the one hand, but the expansion of anthropology in academia on the other, made both the entry and the budgeting of long trips to remote continents an impractical goal for the growing number of students. Instead, many candidates turned their interest to various attractive fields within the borders of their own societies. No doubt, there had been previous successful projects in the United States and Europe of ethnographic studies in fields 'at home' (e.g., the most celebrated work in an Italian lower class Boston neighborhood by Whyte, 1970 [1943]). But these few research endeavors were unusual cases in a tradition that was directed to confront 'other cultures' away from home.

The new search for fields close to home seemed at first to involve only a transfer of sites, such as the study of the apparently clearly defined urban ethnic neighborhoods in the United States (Lewis, 1966; Hannerz, 1969). However, that novelty demanded a change in the style and intensity of participant observation (Plotnicov, 1973). But of more profound consequence, the next change of spatial sites of fieldwork came to also include institutions that offered specific services to permanent or transient populations (for example, Myerhoff's (1978) work in a community center for aged people). That transformation opened the door to a choice of fields that, hitherto, have not been considered suitable for observation (e.g., fieldwork in the scene of gay life by Newton, 1972; Read, 1980; and Shokeid, 1995). In these fields, the fieldworker is often unable to observe the participants throughout a wide spectrum of daily activities. Moreover, in the fields of anonymous participants, the observer's ability to record behavior and communicate with his subjects of research had been very limited in scope (Leap, 1999). These new fields were instrumental in the development of the method and theory of *symbolic interaction* (associated with Erving Goffman's work).

The transformation of methods could also be identified in the choice of problem-oriented ethnographic projects. Even when the relevant site was actually a bounded territorial unit, it engaged a modified strategy of fieldwork (Ginsburg, 1989). But another major change of sites entailed a complete eradication of the bounded spatial characteristics of fieldwork. These unbounded sites included observations of participants in particular occupations (Agar, 1986; Hannerz, 2004), observing cultural performances (Heilman, 1983), as well as observing the members of dispersed social categories (Ortner, 1997). These fields, no longer designed to record the totality of personal and communal life experiences, have been variously defined: *delocalized*, *multisited*, *postcommunity*, *transnational* spaces, and *polymorphous engagements* (Marcus, 1998; Amit, 2000; Shokeid, 2007; Hannerz, 2010).

The ongoing change of the spatial research entities has not necessarily transformed some of the major characteristics of fieldwork. Anthropologists in the new fields continue to initiate their observations mostly as individuals, men or women, engaged in their lone projects. It seems, however, there are less couples involved in these projects, compared with the tendency of married anthropologists to take along their families to the more remote field sites.

In recent years, however, some issues not considered before, concerning the fieldworker's feelings and demeanor, have been openly raised in conferences and edited volumes. Anthropologists preparing for a fieldwork project have never been requested to consult about their personality traits and social capabilities to accommodate what might become stressful field circumstances. But since Malinowski's widow published his personal diaries (1967), revealing the prejudices and the emotional stress experienced in the field by the mythical founder of fieldwork, a more relaxed attitude was gaining place in the professional discourse about fieldworkers' personal problems. This attitude has been closely related to the genre of reflexivity that since the late 1970s gained popularity in ethnographic texts (Rabinow, 1977; Ruby, 1980). The lone fieldworker was no longer perceived through the image of the infallible man or woman but expected to withstand successfully all obstacles in a valiant test of personality that proves he/she is capable of penetrating the social codes of alien cultures in far away territories.

The new fieldwork sites in Western urban societies contributed to that growing awareness and doubts about some long-established taboos that separated fieldworkers from the subjects of their observations (Kulick and Willson, 1995; Lewin and Leap, 1996). The border between the observer and the observed that was naturally upheld in the old fields, where the alien researcher had been conspicuously distinguished (by race and other visible marks of personal characteristics), is no longer easily maintained in many of the new sites of fieldwork. The photographs displayed in many ethnographies of the founding generation, showing them in a landscape of exotic trees and simple huts, posing in the company of minimally clothed natives, have become a symbolic emblem for the changing circumstances of fieldwork.

It is probably an irony that by the end of the twentieth century, a century that gave rise to modern anthropology, fieldwork, its most inventive method and major signifier of professional identity, seems to have lost its incontestable paramount position. Critical views of the validity of positivist methodology, as well as the growing influence of post-modernist currents in cultural studies, are challenging the authority of ethnographic texts based on participant observations (Marcus and Fischer, 1986). These intellectual movements, popular in recent years, inspire and legitimize anthropological projects that are based on other types of data (such as textual, mass media creations, and Internet

connections), which can be collected without the need for fieldwork. It is too early to predict whether the new genre of armchair anthropology will have a more lasting impact on the role of fieldwork and its practitioners.

See also: Anthropology and History; Anthropology: Overview; Cultural Critique, Anthropological; Cultural Relativism, Anthropology of; Ethnography; Ethnology; Field Observational Research in Anthropology and Sociology; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Psychological Anthropology; Qualitative Methods, History of; Symbolic Interaction: Methodology; Thick Description: Methodology.

Bibliography

Agar, M.H., 1996 [1980]. The Professional Stranger, second ed. Academic Press, San Diego, CA.

Agar, M.H., 1986. Independents Declared: The Dilemmas of Independent Trucking. Smithsonian Institute Press, Washington, DC.

Amit, V. (Ed.), 2000. Constructing the Field: Ethnographic Fieldwork in the Contemporary World. Routledge, London.

Casagrande, J.B. (Ed.), 1960. In the Company of Man: Twenty Portraits by Anthropologists. Harper, New York.

Ellen, R.F., 1984. Ethnographic Research: A Guide to General Conduct. Academic Press, London.

Epstein, A.L. (Ed.), 1967. The Craft of Social Anthropology. Tavistock, London.

Fortes, M., 1978. An anthropologist's apprenticeship. Annual Review of Anthropology 7, 1–30.

Freeman, D.J., 1983. Margaret Mead and Samoa: The Making and Unmaking of an Anthropological Myth. Harvard University Press, Cambridge, MA.

Freilich, M. (Ed.), 1970. Marginal Natives: Anthropologists at Work. Harper & Row, New York.

Geertz, C., 1968. Thinking as a moral act: ethical dimensions of anthropological fieldwork in the new states. The Antioch Review 28, 139–158.

Geertz, C., 1973. The Interpretation of Culture. Basic Books, New York.

Ginsburg, F.D., 1989. Contested Lives: The Abortion Debate in an American Community. University of California Press, Berkeley, CA.

Gluckman, M., 1958 [1940]. Analysis of a Social Situation in Modern Zululand. Rhodes–Livingstone paper no. 28. Manchester University Press, Manchester, UK.

Hannerz, U., 1969. Soulside: Inquiries into Ghetto Culture and Community. Columbia University Press, New York.

Hannerz, U., 2004. Foreign News: Exploring the World of Foreign Correspondents. The University of Chicago Press, Chicago.

Hannerz, U., 2010. Anthropology's World: Life in a Twenty-First-Century Discipline. Pluto Press, London and New York.

Heilman, S.C., 1983. The People of the Book: Drama, Fellowship and Religion. University of Chicago Press, Chicago.

Kulick, D., Willson, M. (Eds.), 1995. Taboo: Sex, Identity and Erotic Subjectivity in Anthropological Fieldwork. Routledge, London.

Leap, W.L., 1999. Public Sex/Gay Sex. Columbia University Press, New York.

Lewin, E., Leap, W.L. (Eds.), 1996. Out in the Field: Reflections of Lesbian and Gay Anthropologists. University of Illinois Press, Urbana, IL and Chicago.

Lewis, O., 1966. La Vida: A Puerto Rican Family in the Culture of Poverty – San Juan and New York. Martin Secker and Warburg, London.

Malinowski, B., 1922. Argonauts of the Western Pacific. Routledge & Sons, London.

Malinowski, B., 1967. A Diary in the Strict Sense of the Term. Routledge & Kegan Paul, London.

Marcus, G.E., 1998. Ethnography through Thick & Thin. Princeton University Press, Princeton.

Marcus, G.E., Fischer, M.M.J. (Eds.), 1986. Anthropology as Cultural Critique. University of Chicago Press, Chicago.

Mead, M., 1928. Coming of Age in Samoa. Morrow, New York.

Myerhoff, B., 1978. Number Our Days, first ed. Dutton, New York.

Newton, E., 1972. Mother Camp: Female Impersonators in America. Prentice-Hall, Englewood Cliffs, NJ.

Ortner, S., 1997. Fieldwork in the postcommunity. Anthropology and Humanism Quarterly 22, 61–80.

Plotnicov, L., 1973. Anthropological fieldwork in modern and local urban contexts. Urban Anthropology 2, 248–264.

Powdermaker, H., 1966. Stranger and Friend: The Way of an Anthropologist, first ed. W.W. Norton, New York.

Rabinow, P., 1977. Reflections on Fieldwork in Morocco. University of California Press, Berkeley, CA.

Read, K.E., 1980. Other Voices: The Style of a Male Homosexual Tavern. Chandler & Sharp, Novato, CA.

Ruby, J., 1980. Exposing yourself: reflexivity, anthropology, and film. Semiotica 30, 153–179.

Sanjek, R. (Ed.), 1990. Fieldnotes: The Makings of Anthropology. Cornell University Press, Ithaca, NY.

Shaffir, W.B., Stebbins, R.A., 1991. Experiencing Fieldwork: An Inside View of Qualitative Research. Sage, London.

Shokeid, M., 1988. Anthropologists and their informants: marginality reconsidered. Archives Européennes De Sociologie 29, 31–50.

Shokeid, M., 1995. A Gay Synagogue in New York. Columbia University Press, New York (2003, augmented edition, University of Pennsylvania Press).

Shokeid, M., 2007. From the Tikopia to polymorphous engagements: ethnographic writing under changing fieldwork circumstances. Social Anthropology 15, 305–319.

Turner, V.W., 1967. The Forest of Symbols: Aspects of Ndembu Ritual. Cornell University Press, Ithaca, NY.

Turner, V.W., 1985. On the Edge of the Bush: Anthropology as Experience. The University of Arizona Press, Tucson, AZ.

Turner, V.W., Bruner, E.M. (Eds.), 1986. The Anthropology of Experience. University of Illinois Press, Urbana and Chicago.

Wax, R.H., 1971. Doing Fieldwork: Warnings and Advice. University of Chicago Press, Chicago.

Whyte, W.F., 1970 [1943]. Street Corner Society. University of Chicago Press, Chicago.

Wolcott, H.F., 1999. Ethnography: A Way of Seeing. Altamira Press, London.

Fieldwork: Ethical Aspects

Patricia Marshall, Loyola University of Chicago, Maywood, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 8, pp. 5625–5628, © 2001, Elsevier Ltd.

Abstract

Fieldwork involves a range of methodologies including ethnographic participant observation, in-depth interviews, direct observation of individual or group behavior, or a combination of both qualitative and quantitative approaches. Social scientists conducting fieldwork confront unique ethical dilemmas in the course of implementing their investigations, particularly when researchers are actively engaged with participants in their natural environments. Ethical dilemmas occur when there are conflicts about what is considered to be an appropriate response to a situation involving the investigators, research participants, or the production and use of research materials. A full treatment of these issues is beyond the scope of this article. This article briefly explores ethical issues associated with informed consent, the protection of confidentiality, the development of personal relationships with research participants, involvement in illegal activities, and intervention in ongoing social activities. Finally, a framework for resolving ethical dilemmas in fieldwork is outlined.

Fieldwork involves a range of methodologies including ethnographic participant observation, in-depth interviews, direct observation of individual or group behavior, or a combination of both qualitative and quantitative approaches (Denzin and Lincoln, 1998). Social scientists conducting fieldwork confront unique ethical dilemmas in the course of implementing their investigations, particularly when researchers are actively engaged with participants in their natural environments. Ethical dilemmas occur when there are conflicts about what is considered to be an appropriate response to a situation involving the investigators, research participants, or the production and use of research materials. A full treatment of these issues is beyond the scope of this article. This article briefly explores ethical issues associated with informed consent, the protection of confidentiality, the development of personal relationships with research participants, involvement in illegal activities, and intervention in on-going social activities. Finally, a framework for resolving ethical dilemmas in fieldwork is outlined.

Informed Consent

National and international guidelines for ethical conduct in research with human subjects identify specific requirements for informed consent: the provision of information, comprehension of information, and voluntary participation (Faden and Beauchamp, 1986; Weisstub, 1998). Voluntary participation in research depends upon the individual's ability to understand not just the meaning of the research, but the impact it may have on his or her life. Excessive financial compensation or unrealistic promises may constitute coercion, especially if someone is vulnerable because of social factors such as poverty or ethnic background. Voluntary participation may be inhibited when there are social status and social class differences between fieldworkers and research participants—differences based on gender, age, education, income level, religious affiliation, and ethnic identity. The institutional affiliation of the field researcher may also pose a threat to participants. Additionally, the fluid and unbounded nature of many social science field studies complicates the straightforward approach to informed consent associated with other forms of social science methodologies.

A number of factors influence informed consent in fieldwork—the specific goals of the research, the cultural context of the research project, and communication issues that influence comprehension of information. In international settings, the challenges associated with implementing informed consent in fieldwork are heightened because of language barriers that may inhibit effective communication, particularly regarding the translation of scientific concepts. Misunderstandings and miscommunication about social science research are more likely to occur when fieldworkers and participants speak different languages. Language barriers may be reduced through the use of an interpreter. However, translators must negotiate not only language, but also cultural and contextual factors. Some of the problems associated with interpretation in social science fieldwork include: the inability to translate equivalent expressions easily across languages; paraphrasing that results in omissions or erroneous substitutions of terms; varying levels of comprehension; and the influence of conflicting cultural beliefs and values concerning the meaning and purpose of research.

Differences in beliefs about who has the authority to give permission to participate in research also influence the consent process. Beliefs about personhood, individual autonomy, and decisional capacity are embedded within the social and cultural patterns of family ties and community obligations. In Western industrialized nations, where personal autonomy is emphasized, research participants are expected to make decisions for themselves or through designated surrogates. However, in many non-Western settings, religious or tribal leaders or a person's extended family may play a significant role in decisions concerning research participation. Cultural differences in beliefs about the nature of personhood and the location of decisional authority for consent have been problematic for social science investigators conducting fieldwork in international settings (Levine, 1991). Negotiating informed consent

with designated authorities in non-Western populations requires fieldworkers to move beyond narrow definitions of personhood, autonomy and self-determination. While respect for persons demands cultural sensitivity to local customs, social scientists conducting fieldwork are never authorized to conduct research without regard to potential harm and without attempts to be informative throughout the duration of the study (Levine, 1996). Moreover, although a tribal chief, village elder, or community leader may express approval of a research agenda, sensitivity to cultural customs is secondary to honoring individual choice regarding participation (World Health Organization and the Council for International Organizations of Medical Sciences, 1993).

Informed Consent in Observations of Group Interactions

In fieldwork involving direct observation of group activities, arrangements usually are made prior to initiating the study to inform group members that the fieldworker will be present in the course of routine activities. In closed systems such as a hospital unit, a community center, or other institutional setting, informed consent should be obtained from all those who are at the facility on a regular basis. Individuals who are not present on a regular basis but whose behavior may be observed in public group activities should be alerted to the presence of the fieldworker if it is feasible to do so. In some group observations conducted in the course of fieldwork, it may not be possible or necessary to obtain informed consent from every person present. For example, at informal gatherings such as family or community events, or in public places where individuals interact informally, it would be intrusive for the fieldworker to explain the study to everyone present.

Protection of Confidentiality

Protection of confidentiality and privacy for research participants is an important aspect of fieldwork (Cassell and Jacobs, 1987). Fieldwork in social science investigations often involves the collection of sensitive information concerning an individual's personal life, including information on physical conditions, mental illness, or behavior regarding illegal activities. Individuals or communities could be harmed if information was inadvertently or purposely disclosed. In some cases, information could be used to ostracize or marginalize research participants, particularly if research participants suffer from a stigmatizing illness such as HIV/AIDS, or if they are involved in activities that may be viewed as socially deviant.

Data collection strategies associated with fieldwork require careful attention to the representation of individuals and communities in written transcripts derived from interviews and observations (Fluehr-Lobban, 1991). Descriptive data recorded in field notes, daily logs, or diaries should be entered using a coding system that protects the confidentiality of the research participants. Additionally, field data should also be secured outside of the research setting in order to safeguard the identity of participants. Pseudonyms, rather than the names of participants, must be used in published manuscripts that include case narratives. This gesture, however, will not necessarily protect confidentiality because it may be possible to identify the individuals involved based on the details of the case and the research setting. Nevertheless, at each stage of the research—data collection, data analysis, and data reporting—every effort should be made to protect the privacy of the study participants. In obtaining informed consent, investigators are obliged to advise participants in fieldwork about the methods designed to protect their privacy.

Relationships with Research Participants

Social scientists involved in fieldwork have the opportunity to form close personal relationships with research participants because of the collaborative nature of ethnographic approaches such as participant observation. The development of trust between fieldworkers and the individuals involved in the research enhances the potential for collecting important information relevant to the study. However, in some circumstances, investigators may find themselves in compromising situations because of their personal relationships in the field.

Intimate Relationships

Although it is rarely addressed in the literature (for an exception, see Kulick and Willson, 1995), fieldworkers may become intimately involved with research participants. Sexual relationships with field participants are more likely to occur in the context of extended fieldwork. Traditional anthropological fieldwork—where the investigator may spend months or years in an environment far removed from familiar social networks (often in a different country)—represents the type of setting that contributes to the development of intimate personal relationships with research participants. Despite what some might view as the natural inclination to form a sexual attachment with a close companion, the prevailing opinion among social and behavioral scientists is that sex with participants in the field should be avoided. Sexual relationships between researchers and study participants may diminish the relative objectivity of the investigator.

More problematic is the potential for coercion of socially vulnerable individuals. Moreover, attitudes concerning sexual behavior are socially constructed and the meanings attached to sexual intimacy may vary significantly between the fieldworker's home community and the research setting (Singer et al., 1999). The social complications that can result from a sexual relationship (e.g., jealousy, accusations of infidelity, gossip) jeopardize the integrity of the research and place the investigator at risk for problematic interactions with the broader community.

Illegal Activities

Investigators conducting fieldwork with populations such as illegal immigrants, or with individuals involved in activities such as drug use or the sex trade, may be asked to participate in behavior that is legally questionable. For example, fieldworkers investigating the use of illegal drugs may be asked to share drugs with study participants or to assist in procuring or holding drugs or drug paraphernalia. While some researchers may believe that participation in these types of activities

strengthens the bonds of trust between a fieldworker and research participants and thus increases the likelihood of securing information that is meaningful and necessary to achieve the study goals, there are strong reasons for avoiding participation in illicit or illegal behavior. As with sexual involvement with study participants, engaging in illegal activities seriously threatens the integrity of study findings and places the researcher at risk socially, professionally, and in some cases, physically. Fieldworkers may find themselves targeted for blackmail or other forms of coercion. Additionally, fieldworkers who participate in illegal behavior diminish the potential to secure funding for future research. A full consideration of the ethical dimensions of involvement in illegal activities cannot be explored in this article.

Fieldworker Intervention in Ongoing Social Activities

Questions concerning the intervention of a fieldworker in problematic situations that occur in the course of conducting research are challenging. The fieldworker's decision should be a judgment based on the purpose of the study, the context, and the specific event that occurs. When investigators continually intervene in field situations that are directly related to the issue being investigated, the possibility that their research will have an impact is jeopardized since they would be unable to document the events being studied. For example, fieldwork with injection drug users to investigate HIV risk behavior would be inconclusive if researchers attempted to stop practices known to facilitate the transmission of the virus. The fundamental problem is one of advocacy—when to be an advocate for research participants and when to maintain the more circumscribed role of the investigator (Bosk, 1985). In general, nonintervention is the accepted mode for fieldwork.

Recommendations for Resolving Ethical Dilemmas

Investigators who confront moral dilemmas in the course of conducting fieldwork must consider carefully the full range of issues involved in the problem. A systematic approach to ethical decision-making begins with a robust description of the research dilemma. This would include details about the purpose of the study, the research design, the sponsors of the study, and the individual participants and community involved in the research. The researcher then must consider the cultural and social values represented in the ethical problem. The relevant values of the study participants, the study community, the professional community, and the study sponsors should be outlined. The researcher must determine the principal value conflict. At this point in the process of decision-making, the fieldworker should consider whose values are threatened and who are the most vulnerable to potential harms. This process of reflection facilitates the identification of the key issues involved in the ethical dilemma. When the primary ethical issues are determined, the investigator should outline a full range of strategies and consider the potential risks and benefits associated with each solution. The decision regarding a course of action should maximize respect for the individual and group values identified. The vulnerability of research participants and the communities within which they live should be of paramount importance in resolving the ethical dilemma encountered.

See also: Ethical Codes, Professional: Business Codes; Fieldwork in Social and Cultural Anthropology; Informed Consent of Research Subjects; Linguistic Fieldwork; Research Ethics, Cross-Cultural Dimensions of; Research Ethics: Research.

Bibliography

Bosk, C., 1985. The fieldworker as watcher and witness. Hastings Center Report 15, 10–14.

Cassell, J., Jacobs, S. (Eds.), 1987. Handbook on Ethical Issues in Anthropology Special Publication of the American Anthropological Association. Number 23, Washington, DC.

Denzin, N.K., Lincoln, Y.S. (Eds.), 1998. Collecting and Interpreting Qualitative Materials. Sage Publications, Thousand Oaks, CA.

Faden, R.R., Beauchamp, T.L., 1986. A History and Theory of Informed Consent. Oxford University Press, New York.

Fluehr-Lobban, C. (Ed.), 1991. Ethics and The Profession of Anthropology. University of Pennsylvania Press, Philadelphia PA.

Kulick, D., Willson, M., 1995. Taboo: Sex, Identity and Erotic Subjectivity in Anthropological Fieldwork. Routledge, New York.

Levine, R.J., 1991. Informed consent: Some challenges to the universal validity of the Western model. Law, Medicine, and Health Care 19, 3–4.

Levine, R.J., 1996. International codes and guidelines for research ethics: A critical appraisal. In: Vanderpool, H. (Ed.), The Ethics of Research Involving Human Subjects: Facing the 21st Century. University Publishing Group, Frederick, MD, pp. 253–259.

Singer, M., Marshall, P., Trotter, R.T., Schensul, J.J., Weeks, M.R., Simmons, J.E., Radda, K.E., 1999. Ethics, ethnography, drug use, and AIDS: Dilemmas in federally funded research. In: Singer, M., Marshall, P., Clatts, M. (Eds.), Integrating Cultural, Observational, and Epidemiological Approaches in The Prevention of Drug Abuse and HIV/AIDS. US Department of Health and Human Services National Institutes of Health Publication, Washington, DC, 99–4565.

Weisstub, D.N. (Ed.), 1998. Research on Human Subjects: Ethics, Law and Social Policy. Pergamon, Oxford, UK.

World Health Organization and Council for International Organizations of Medical Sciences (WHO-CIOMS), 1993. International Ethical Guidelines for Biomedical Research Involving Human Subjects.

Film and Video Industry

Douglas Gomery, University of Maryland, College Park, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 8, pp. 5637–5640, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The production and distribution of motion pictures occurs throughout the world but since the 1920s has been dominated by giant Hollywood studios. These vast multinational enterprises create films that are presented in virtually every country in the world, frequently capturing a major share of the business despite competition from locally made and distributed motion pictures. Almost continuous technological change has never lessened Hollywood's power, and indeed the innovation of home video extended Hollywood's influence. Marxist critics have long seen this influential cultural industry as a classic example of monopoly capitalism, focusing on how these long-lived corporations devised ways to maintain their power. Neoclassical free market economists, on the other hand, have stressed that Hollywood has long faced entertainment alternatives. Industrial organization economists note that basic forces, such as economies of scale of international distribution, have always enabled Hollywood not only to dominate the world market, but also to set the standard of narrative structured entertainment. Whatever the proper explanation of the maintenance of economic power, filmmakers outside Hollywood long have struggled to find their place in their national cinemas, often requiring the assistance of their governments. Only sporadically have other nation-states—first France, then Italy, Japan, and India—been able to mount a studio system as an alternative.

The production, and distribution of cinema occurs throughout the world but since the 1920s has been dominated by eight or less giant Hollywood studios. These vast multinational enterprises create films that are presented in virtually every country in the world, frequently capturing a major share of the business despite competition from locally made and distributed motion pictures. Almost continuous technological change – led by the Hollywood oligopolists – did not lessen Hollywood's industrial power, and so all other film industries sought to ape Hollywood's continuous success. The coming of sound and color, and the innovations of pay-TV and home video (played on VCRs) simply extended Hollywood's influence, and increased its profits. Analysis of the Hollywood studio system has led to praise or damnation of this omnipresent, inescapable, and highly influential industry.

The Film and Video Industry

A handful of companies have defined the Hollywood film industry, dominating the US and world markets. They have weathered a world war, and a Great Depression, innovated wide screen and color technologies, made peace with television, learned to exploit home video, and were more powerful at the end of the twentieth century than ever before.

The coming of sound solidified control by five studios over the world cinema market: Paramount Pictures, MGM, Twentieth Century Fox, Warner Bros., and RKO. During the 1930s and 1940s by controlling the key cinemas in the downtowns of nearly all major US cities, the 'Big Five' took in nearly all box-office take. Only after they granted their own theaters first-runs, did they then permit smaller, independently owned theaters to scramble for the remaining bookings months, or even years, after a film's premiere.

Paramount represented the most profitable, powerful, and traditional business-like Hollywood company. More than any other member of the Big Five, Paramount relied on its chain of more than 1,000 theaters to maintain its corporate might. Not surprisingly, a former theater operator, Barney Balaban, stood at the top of this corporate colossus, hiring more lawyers and MBAs than movie stars.

Metro-Goldwyn-Mayer (MGM) technically functioned as a successful unit of Loew's, a theater chain concentrated in the Eastern USA. MGM, more than any other studio, relied on the star system, producing films from its massive Culver City, California-based 'factory' of 27 sound stages (for interior shots), and a 168-acre back lot (for outdoor scenes).

Twentieth Century Fox ranked behind Paramount and MGM with neither as large a theater chain or studio facility. Studio boss Darryl F. Zanuck used what he had to authoritatively produce films for world distribution. His decisions were final, frequently bellowing to all subordinates: 'Don't say yes until I've finished talking!'

Warner Bros. represented the only family-run operation among the major movie studios, with eldest brother Harry as President, middle brother Abe supervising worldwide distribution, and baby brother Jack heading studio production. The innovation of sound pushed Warners into the Big Five, yet financially the company struggled through the 1930s and 1940s.

Radio-Keith-Orpheum (RKO) was formed so that radio's RCA could sell its sound equipment to Keith–Orpheum theaters. While the studio created many memorable films – such as *Citizen Kane* and the Astaire–Rogers musicals – RKO's only regular profitability came from the distribution of Disney animation. TV killed RKO.

The Big Five tolerated some competition – from Universal Pictures, Columbia Pictures, and United Artists – as the 'Little Three' depended on the Big Five for access to the top theaters in the USA. The suburbanization of the US, and the coming of television allowed Universal, Columbia, and United Artists to prosper because the advantage of owning theaters disappeared

because of antitrust action by the US government. The former Big Five, save RKO, tried to re-invent themselves.

Warner Bros. began its transition into what became Time Warner when in 1956 the founding brothers sold out, and a series of new owners embraced television production. In 1969, Kinney National Services, a New York conglomerate, purchased Warners, and Steven Ross, son-in-law of Kinney's founder, began to fashion the modern Hollywood-based prototypical multinational media conglomerate. In 1989 Ross merged Warner with Time, and the largest media company in the world was created. A decade later AOL took control.

Paramount stubbornly hung on until 1966 when Charles Bluhdorn's Gulf+Western Industries conglomerate, after a 2-year battle, took over. Through the 1970s and 1980s the 'new' Paramount thrived as part of a classic conglomerate, but in another 2-year struggle, was taken over in 1993 by another media conglomerate, Sumner Redstone's Viacom, Inc.

Twentieth Century Fox began its reinvention in 1956 when longtime studio boss Darryl F. Zanuck resigned. His replacements struggled, but did lead the way to licensing feature films to major television networks. NBC's 'Saturday Night at the Movies' began TV's showcasing of feature films. The surprise hit *Star Wars* (1977) moved Fox to the top of the Hollywood studio hierarchy, and within a decade led to the studio's purchase by Australian media mogul Rupert Murdoch. From the Fox studio base, Murdoch created the Fox TV network.

MGM never was able to reinvent itself, even with its merger with United Artists. The two other members of the Little Three did far better. Columbia went through several ownership changes (including during the 1980s when Coca-Cola owned it), but it was not until 1989 that stability came when Japanese electronics giant Sony purchased Columbia. During the 1990s Sony never was able to exploit its Hollywood studio to sell more VCRs, but did learn to create vast profits.

Universal also struggled until 1962 when the MCA talent agency acquired it. Thereafter, under Lew Wasserman, Universal ascended to the top of Hollywood movie and television studio hierarchy as Wasserman innovated the made-for-TV movie, and the modern film blockbuster – *Jaws* (1975). Wasserman retired in the early 1990s, and in 1995 Canadian liquor giant Seagram took control.

Disney joined Warners, Paramount, Fox, Sony, and Universal to form the 'Big Six' of the 1990s. Beginning in 1984, Michael Eisner leveraged Disney's weak position into major studio status, a significant change in an industry with so much long-term stability. Indeed, how had this small number of companies survived and thrived for so long – with no real newcomers joining their ranks?

First, even with all the social technological and industrial changes, the major companies alone continued to be able to distribute films around the world. At considerable expense, all maintained offices not only in the USA, but in every region of the world. In most countries around the world, owners of theater chains knew the Hollywood majors alone could continuously deliver popular films, and so even theater owners outside the USA preferred Hollywood films to native productions. The Hollywood studios' record at the box office satisfied theater owners – a conservative lot with most of their assets invested in real estate. Economies of scale prevented rivals from building and maintaining equal world distribution networks.

Second, modern Hollywood fashioned the blockbuster strategy to turn television into a friend, not a foe. Through the final quarter of the twentieth century, the Hollywood oligopoly learned to generate profits from creations that premiered with blitz TV advertising, which in turn was married to a multirevenue generating machine to sell musical sound tracks, novelizations of the stories, 'action figures' from movie characters, and endless sales on video. Each movie aspired theatrically to become a smash hit because if it succeeded then the feature turned into a 'product line,' designed to fill all entertainment needs and desires from toys to theme park rides to campaigns to sell more McDonald hamburgers.

The blockbuster strategy in turn led to ever escalating costs of production so it became more difficult for rivals to join the exclusive Hollywood club. By the close of the twentieth century the average Hollywood feature cost in excess of $50 million to produce, and half again that amount to publicize. *Titanic* doubled those average cost figures, but then earned in excess of $2 billion in revenues. Potential filmmakers knew that if they ever aspired to have a world audience see their creation and to attain undreamed of wealth, they needed to be 'green lighted' by one of the Hollywood major studios.

Alone, Hollywood could keep the revenues flowing year after year – particularly from home video. While theaters still premiered new films, it was 'downstream' on video where the bulk of the monies were generated. The theater functioned as the 'voting booth' where hits were made, and necessary publicity generated. But once the film had fully milked the box-office it went into release on pay-TV and home video that created more than two-thirds – in the long run – of the average film's revenues. Hollywood executives carefully planned – in classic price discrimination fashion – to drain all possible revenues from each of these release venues, only releasing it to home video when they reasoned all the revenues from theatrical and then pay-TV had been procured. Each window in this sequence was an exclusive; a new window opened only when all value of the previous window had been captured. This exploitation of video represented Hollywood's newest advantage in the film industry.

The Rest of the World

While the Hollywood majors maintained their economic power and influence, filmmakers outside Hollywood long struggled to find their place in their national cinemas, often requiring the assistance of their governments. Only sporadically have large nation-states – France, Italy, Japan, and India in particular – been able to mount a studio system as an alternative and serious rival to Hollywood in their own countries, let alone elsewhere.

Globally, only in the rare nation does Hollywood not capture more than half the business. The Hollywood majors maintain offices around the globe, where their representatives maintain constant contact with the heads of the dominant theater chains, pay-TV outlets, and home video retailers. Hollywood's regular production of hit films provides a strong incentive for foreign theater, pay-TV, and home video

companies to consistently deal with them rather than take a flyer with some new company – even a native company.

But other nations did try. During the 1920s European nation-states – led by the United Kingdom, France, and Germany – sought to build their own studio systems, and were successful for short periods of time, until their leading filmmakers left for Hollywood. During the 1920s, the new Soviet Union tried to build an alternative national cinema industry under Communism. The 1930s ended these experiments. At the close of World War II, the focus shifted to state sponsored television, but still in Japan and India studio systems did succeed, even while conceding Hollywood a vast share of the monies paid for cinema attendance.

After the advent of the blockbuster, it became harder for foreign governments to subsidize a native cinema industry. Most chose to focus on television instead. But Germany and Australia, for example, did for a time in the 1980s create enough different types of popular productions long enough to be discovered as the latest 'new wave.' Japan had long had a studio system, but the coming of television and the takeover of Hollywood studios by Sony, and temporarily, Matsushita, ended the importance of that native industry. But India, for example, was still able to maintain native popular cinema, complete with a star system and defined genres like Hollywood. But these films were intended for domestic consumption, with minimal export possibilities. Community viewing in cities and poor areas alike still flourished.

Hollywood has long maintained an influential trade association – the Motion Picture Association of America noted for its ratings of films, but far more influential around the world for lobbying to smooth the way for international distribution. The MPPA lobbied the US State Department to keep cinema trade 'free' to permit the Hollywood oligopoly to continue pulling billions in revenue from the foreign [read: non-US] market. Foreign governments, on the other hand, resented US pressure, and formally and informally sought to protect the national film industries, and their cultural integrity and identity.

Industrial Analysis

The common thread of Hollywood power and influence, and its growing presence around the world, have led to much analysis of the Hollywood film industry. Three different interpretations have dominated.

Marxist Industry Analysis

Marxist critics have long seen this influential cultural industry as a classic example of monopoly capitalism, focusing on how these long lived corporations colluded to devise ways to maintain their power and cultural imperialism. While Wall Street celebrated Twentieth Century Fox and Paramount's co-financing of *Titanic*, Marxist critics would see this as yet another example of studios working together – not competing – to maximize profit. Since there were at most eight players, they did not have to worry about going out of business, but simply seeking how much their exploit and keep in profits.

The long-term propensity toward concentration of ownership, endemic to monopoly capitalism, was simply exemplified by the Hollywood motion picture industry. That is what history should teach us, and that is how Hollywood has behaved even in the face of new technologies. Marxists focus on the expansion of giant corporations and the spread of their influence to television, music, toys, theme parks, and other entertainments. Market concentration and anticompetitive behavior describe the operation and ownership of the Hollywood movie industry.

Marxist industry analysis usually focuses on international distribution, seeing the trend toward globalization as evidence of cultural and economic imperialism. While Hollywood publicity focuses on the production of films – its stars, stories, and special effects – international distribution has always been a key to corporate longevity. No other national film industry has been ever been so far-reaching, endangering indigenous, nonprofit seeking culture. They correctly emphasize that Hollywood taught the mass culture businesses to focus on and exploit the advantages of globalizing as seen in the work of Wasko (1994) and Guback (1969).

Neoclassical Industry Analysis

Neoclassical economists emphasize that movies face competition from all forms of mass culture. While there are and have been a few very powerful studios, consumer sovereignty means that no Hollywood company can ever force a customer to pay to see a film – even *Titanic*. These analysts focus on the forces and interplay of supply and demand, seeing the movie theater as a classic voting booth, where top grossers win and unpopular films are pulled almost immediately. They stress that there is a great deal of product diversity that permeates the entire landscape of the mass media. This is why genres – from science fiction to westerns, from dramas to comedies – ebb and flow in relative popularity. With the rise of the VCR, for example, the choice of titles to rent or buy needs to measured in the thousands. Like the Marxists who come to their analysis with a predetermined anticapitalist assumption, these neoclassical analysts enter the debate with a procapitalism assumption as seen in the work of Litman (1998).

Industrial Organization Analysis

Industrial organization economists start with neither pro- or anticapitalistic assumptions, and instead seek first to define 'who owns the media,' then analyze what economic conduct emerges from that industrial structure, then detail how the industry performs given specified criteria, and finally recommend possible public policy corrections to make the industry do 'better.' Industrial organization economists have long recognized the host of advantages that enabled them to maintain their considerable economic power and keep out the competition.

For example, cross-subsidization enabled a Hollywood corporation – with interests in a number of media markets – to extract profits from a thriving area to prop up another less financially successful area. Single-line corporations do not have this luxury and so aspiring single-line Hollywood operations invariably fail as Gomery (1986) discusses.

Or, to take another example, reciprocity enables a diversified member of the Hollywood major studios to choose to

whom they will sell and then only deal with those companies that cooperate with other units of the media conglomerate. For example, Universal Pictures might not sell movies to Time Warner's HBO unless Time Warner's cable franchises book Universal's (partially owned) USA television network.

Horizontally, year after year, the Hollywood oligopoly has generated considerable profits from a wide spectrum of mass-media enterprises, including theme parks, music, and television production. There is no such thing as a single-line major movie company. The major Hollywood companies also control the process of production and distribution – as they long have – but also more and more of the increasing avenues of presentation, from pay-per-view channels to home video to digital home viewings.

See also: Film History; Film: Genres and Genre Theory; Video Cultures: Television.

Bibliography

Alexander, A., Owers, J., Carveth, R., 1968. Media Economics. Erlbaum, Mehwah, NJ.
Balio, T., 1985. The American Film Industry. University of Wisconsin Press, Madison, WI.
Bruck, C., 1992. Master of the Game: Steve Ross and the Creation of Time Warner. Simon & Schuster, New York.
Conant, M., 1960. Antitrust and the Motion Picture Industry. University of California Press, Berkeley, CA.
Gomery, D., 1986. The Hollywood Studio System. St. Martin's, New York.
Gomery, D., 1992. Shared Pleasures: A History of Movie Presentation in the United States. University of Wisconsin Press, Madison, WI.
Guback, T., 1969. The International Film Industry. Indiana University Press, Bloomington, IN.
Kindem, G., 1982. The American Movie Industry. Southern Illinois University Press, Carbondale, IL.
Lardner, J., 1987. Fast Forward: Hollywood, the Japanese, and the Onslaught of the VCR. Norton, New York.
Litman, B., 1998. The Motion Picture Mega-industry. Allyn and Bacon, New York.
Morin, A., 1996. Film Policy: International, National, and Regional Perspectives. Routledge, London.
Neale, S., Smith, M., 1998. Contemporary Hollywood Cinema. Routledge, London.
Wasko, J., 1994. Hollywood in the Information Age. University of Texas Press, Austin, TX.
Wyatt, J., 1994. High Concept: Movies and Marketing in Hollywood. University of Texas Press, Austin, TX.

Film: Genres and Genre Theory

Ib Bondebjerg, University of Copenhagen, Copenhagen S, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Genre is a concept used in film studies and film theory to describe similarities between groups of films based on aesthetic or broader social, institutional, cultural, and psychological aspects. Film genre shares similarities in form and style, theme, and communicative function. A film genre is thus based on a set of conventions that influence both the production of individual works within that genre and audience expectations and experiences. Genres are used by industry in the production and marketing of films, by film analysts and critics in historic analysis of film, and as a framework for audiences in the selection and experience of films.

Main Tendencies in the Study of Film Genres

The concept of genre has been important since the beginning of the history of film, as a category for the industry and thus for the production, marketing, and distribution of films; for film critics and the academic, aesthetic, historical, and theoretical analysis of film; and for audiences as a framework for the selection and experience of films. The study of film genres is therefore placed in a triangular structure relating to the institutional, social, and cultural context of films; to the aesthetic, formal, thematic, and stylistic dimensions; and to audiences and reception understood both in an empirical social and psychological sense.

The development of the study of film genre from classical to modern film theory have shifted between focus on different aspects of this triangular structure. Studies of particular film genres, e.g., the Western or the family melodrama, may stress the social and cultural aspects of genre, as ideological or ritual expressions of American history, values, and mythology, or as part of the Hollywood studio system and the institutional development of production norms (Schatz, 1981). Or studies may stress the formal, narrative, visual, iconographic, and thematic structure of particular groups of films in order to find similarities and variations and historical changes and developments (Staiger, 1995; Bordwell et al., 1985). But studies of genre have also focused on psychological aspects of genre such as gender and the imaginary, using Freudian, Lacanian, and feminist terminology (Mulvey, 1975) or, as it has been the case in recent thinking about genre, from a cognitive and emotional theory of basic prototypes of film genres (Grodal, 2009). Many theories of genre do not stress just one part of the triangle but clearly indicate that genres are unstable and fuzzy categories. Rather than defining film genres as sharp, distinctive, and essential categories, modern genre theory tends to define genres as prototypes, and to see the establishing of genres as processes related to institution, text, and audience and historical changes.

Auteur and Genre

Although film genres are found as categories from the beginning of film history, used by both production companies and film critics, the systematic theoretical and historical reflections on film genres were developed in the 1950s as European auteur theory met US genre films (Fowler, 2002). The early theories of film were often very strongly tied to a more individualized and aesthetic understanding of film as art, and therefore often viewed mainstream cinema and genres as commercially dominated formulas, that did not allow artistic freedom and innovation for the auteur. But although this discussion is often repeated even today, film genre theory in general has explicitly questioned this normative division between an industrialized product and mainstream cinema, on the one hand, and film art and the auteur concept, on the other.

In the 1950s, French critics centered around *Cahiers du Cinéma* and the New Wave stressed the variations and auteurism of classical genre films in their writings on classical Hollywood cinema. Bazin, despite his central role in the formulating of a realist film theory and an auteur theory, was also engaged in describing the special qualities of classical US genre films, such as the Western. In his two seminal articles (Bazin, 1971a,b) published in French in 1953 and 1954, he discussed the Western and especially John Ford's Westerns as US film art *par excellence*. In these articles, Bazin saw the Western as a visual and narrative embodiment of American history and myths, and in his definition of what constitutes a genre, he focused on both thematic elements and elements of cinematic technique and style, narrative strucand *mise-en-scène*. He also initiated later discussions on the historical transformation of the Western, defining both its pure classical forms and newer forms, which he called 'superwesterns' and 'novelistic Westerns.'

Bazin's articles mark a transition from a normative aesthetic notion of genre films as just degraded, commercialized, and industrialized art. He combined auteur perspectives and genre perspectives into a more general aesthetic and cultural perspective on all forms of film. This was also the intention behind a number of later US contributions to the general understanding and reevaluation of US film genres. Braudy (1976) not only explicitly criticized the division between film as art and film genres, but also claimed that mainstream genres can express broader cultural, social, and aesthetic values than classical works of art can. Buscombe ([1970]1995) takes a broad discussion of literary genre theory as his starting point and specifically uses Wellek and Warren's distinction in *Theory of Literature* ([1949]1956) between 'outer form' and 'inner

form' and their cinematic equivalents, thus opening a more structural and aesthetic specification of genre definition.

Semiotics and the Ritual and Ideological Turn

The move from auteur to genre from the late 1960s, was sustained and further developed through the influence of semiotics, structural anthropology, and literary genre theory to a broader ritual and ideological approach to the study of film and film genres. The ritual approach involved both a more systematic analysis of cinematic texts and a form of institutionalized practice expressing and negotiating fundamental cultural values. The ideological perspective involved a critical study of cinema and cinematic genres and the reproduction of dominant forms of ideology from sociological, Marxist, Freudian, and feminist perspectives. The two approaches do not necessarily exclude each other but in the ritual approach there is a tendency to see the relation between genre and audience as one of negotiation and exchange, whereas the ideological perspective more often stresses the principle of dominance.

In Cawelti's broad ritual study of popular literature and film (Cawelti, 1976) and in Wright's classical ritual and structural analysis of the Western (Wright, 1975), genres are seen as fundamental story types that represent and embody cultural images and myths with both universal and historical elements, and which are recognized by the audience. Genres and the act of genre reading balance between mimesis and escapism. Cawelti's idea is that genre formulas work because they are able to express and negotiate historical conflicts and mental and cultural oppositions and contradictions in a way that is accepted, understood, and appreciated by a large audience, not just dominated by ideology. He rejects deterministic theories in favor of symbolic or reflective theories, where assimilation, negotiation, and change take place but without denying that formula stories reflect and affirm dominant forms of thought and attitudes.

Cawelti defines five universal story types that can be used and combined in different ways: adventure, romance, mystery, melodrama, and stories of alien beings and states. In Schatz's (1981) study of classical Hollywood genres he talks about genres in a fundamental semiotic way in which he sees "film genre as a specific grammar or system of rules of expression and construction and the individual genre film as a manifestation of these rules" (Schatz, 1981: p. 19). Basic film genres form the *langue* and individual genre films form the *parole* of genre filmmaking and film language, and he defines the grammar in terms of a genre's iconography, characters, settings, and plot structure. Shatz's analysis is a study of some of the most stable film genres in the classical Hollywood system, the individual films, and the directors, but he also points to two very basic types of supergenres: genres of order (Western, gangster, detective) and genres of integration (musical, screwball comedy, family melodrama) with two very different narrative strategies, settings, types of hero, and solution models of conflicts that either express violent elimination and male dominance or emotional, familial integration, and couple or female orientation.

Shatz's approach points to forms of genre analysis also addressed in the more ideological and feminist genre theories developed at the same time, but with a background in French film semiotics or the influential UK journal *Screen*. One of the seminal articles articulating the ideological approach is Comolli et al. ([1969]1992) article, printed originally in *Cahiers du Cinema* and in English in *Screen* (1972). Every film is political and ideological, they state, in so far as no film or filmmaker can escape the commodity form and the capitalist economy. According to Comolli and Narboni, filmmaking at large is nothing but an expression of the prevailing ideology, which is working behind the back of both the filmmaker and the audience. But the article also defines a kind of ideological genre taxonomy, in which films are defined in relation to whether and how they try to oppose the dominant ideology. Among the seven genres defined are: the majority of films, which are just unconscious instruments of ideology; films that have an explicit critical content and or/form; films that have an ambiguous secret message and form; and films like documentary films, that are seen as only illusionary breaks with the dominant ideology.

Ideological genre analysis found dominance of ideology disguised behind classical realism or inscribed in the apparatus of cinema and therefore viewed the audience as rather passive. However, ideological analysis did look for both a counter-cinema and counterstrategies to the dominance of ideology in genre film and mainstream cinema. But whereas the ritual approach saw qualities in the mainstream film and its relation to the audience on a more direct level, the ideological approach often investigated ambiguities and contradictions in genre films dominated by ideology. Writers belonging to the ideological and semiotic project (e.g., Wood, [1977]1995; Klinger, [1984]1995) therefore also found elements of artistic personality still visible in genre films and often described subversion of the dominant tendencies in mainstream film.

Gender and Genre

The story of genre theory and gender starts with the feminist project based on the ideological and Freudian–Lacanian-inspired genre theory that sought to find the dominant. Mulvey's ([1975]1992) much quoted essay combined elements from all aspects of the ideological genre theory, but with the aim of demonstrating the male dominance of the feminine in mainstream Hollywood. She pointed to the male gaze and the feminine characters as the looked upon object and she also described the asymmetric relation of active/passive roles. But Mulvey's article did not just open the feminist critique of male dominance, but also the demand for another aesthetic, a feminist counter-cinema. Eventually, it also led to the investigation of subversive female readings of classical genre films, most notably 'women's films,' e.g., melodrama or film noir as it is very broadly demonstrated in Gledhill's (1987) anthology on melodrama and women's films and Kaplan's ([1978]1998) on women and film noir, in which prominent feminist film genre critics, e.g., Mary Ann Doane, Tania Modleski, and Annette Kuhn, figure. Later feminist writers have directly entered more action-oriented and excessive 'body genres,' such as Williams studying pornography, horror, and melodrama (Williams, 1989) or Clover's (1992) thorough study of the horror genre. And just as feminist criticism has opened toward seemingly more

'masculine genres,' gender studies of film genres have also included studies of gay and lesbian films or gay and lesbian aspects of mainstream genre films (Davies, 2008; Griffith, 2009).

Genre, Reception, and History

In literary genre theory Todorov (Todorov, [1970]1975) made a distinction between theoretical and historical genres. Theoretical genres are defined in more general terms by criticism and theory and historical genres are defined as genres actively used and recognized by critics and readers. This distinction is a starting point for Neale's (1980, [1990]1995) and Altman's ([1984]1995, 1999) contributions to a genre theory. Neale (1980) represents screen theory that combines ideological and semiotic–aesthetic aspects. But he expands the concept of genre to a very multifaceted process. Genres are both social and textual institutions and, according to Neale, they work to create a coherence and narrative closure, which positions the viewer safely inside a well-known structure: "a genre is, simultaneously, a coherent and systematic body of film texts, and a coherent and systematic set of expectations" (Neale, 1980: p. 37). Altman ([1984]1995) also clearly continues the semiotic project of genre definition, but at the same time seeks a merging of the ritual and ideological approach and the more general description of genres as a grammar and language. His suggestion in the article is thus to combine the semantic approach to genres, defining its 'building blocks,' and a syntactic approach defining the actual structures in which the semantic elements are placed. But at the same time he argues for a stronger historical complexity in the study of genres and for a stronger emphasis on the study of genres in historical contexts and to actual interpretative communities.

This line of argument focusing on reception and the historical process of 'genrification,' and genre as a kind of 'game' between critics, producers, and audiences in a concrete historical process, is strongly underlined in Neale ([1990] 1995) and Altman (1999). Taking some historical examples of how different early film genres were defined in industrial and critical terms and how long it took for well-established genres to stabilize as both institutional and audience categories, Neale argues for genre studies with more real historical integrity, actually studying the discourses on genre used in a historical context, rather than only constituting general, theoretical genre concepts. Along the same lines, Altman focuses very much on the unstable and fuzzy nature of genres and the differences in audiences' readings of individual genres and individual films. He adds a pragmatic dimension to his earlier semantic–syntactic dimensions to include the user-contexts aspect. So key terms in this reception-oriented historicizing of film genre theory are different readings, multiple users, and conflicting understandings of genres, instead of genres understood as relatively stable categories based on both general textual similarities and a social and cultural contract or common understanding.

Cognition, Emotion, and Genre

A move in the opposite direction back to a more general definition of genres can be found in the strong new trend toward cognitive and pragmatic studies of genres from the 1990s and on. The earlier semiotic and ritual sociocultural approach to genres defined dominant structures based on more general patterns of mental, generic procedures, that partly explain why mainstream genres are so strong and dominant. This point of view was based on structural arguments that were not very firmly anchored in a psychological dimension. Cognitive-pragmatic genre studies also define very basic patterns and schemata in our active construction of narrative forms and basic generic formulas. But the cognitive-emotional theories focus on more internal psychological dimensions, but do not exclude a pragmatic-historical dimension, since all such mental aspects of communication are situated and embodied in a specific historical and communicative situation. The cognitive-emotional genre theory also opposes the dominant Freudian, Lacanian, or Jungian-inspired psychology in the Screen tradition of genre studies. Important contributions to this new cognitive-emotional genre theory is Bordwell's (1985) description of different narrative prototypes and the way they cue the spectator and Carroll's studies of horror (1990) and his later article on emotion and genre (Carroll, 1999), in which melodrama, horror, and suspense are described from a cognitive-emotional perspective.

But the most ambitious attempt to formulate a cognitive and emotional theory of genre is that of Grodal (1997, 2009). In his book *Moving Pictures* (1997) Grodal defines a new typology of very basic genres based on the viewer's active–passive relation with the film, the kind of distanciation or identification involved, and the typical moods and emotional responses activated during viewing: associative lyricism, canonical narratives of action, obsessional paratelic fictions, melodramas, horror, schizoid fictions, comedy, and metafiction. Behind these basic forms we then find a number of more concrete, historical subgenres. The basic assumption of the book is that genres of visual fiction are mental structures that to a large degree are modeled on the basis of a simulated reality of actions and emotions to which we react with much the same experiences and capacities as in real-life experiences. The book thus argues that many of the fundamental structures of film experience are based on embodied emotional and cognitive patterns that interact with the concrete social and historical context of audiences.

In his first book Grodal describes both the typical embodied flow, during the experience of visual fiction with reference to cognitive-emotional psychology, and the reason for the strength and popularity of the prototypical genres of visual fiction and their relation to different mental and emotional activities and experiences. Grodal's book is, of course, not strongly oriented toward the cultural and historical aspects of moving images, but there is no essential opposition between a cultural, historical, and stylistic approach to visual fictions and Grodal's position, as long as one realizes that certain aspects of the experience and processing of moving images cannot be described as culturally constructed in any short-term sense.

Further cognitive studies of film genres have been developed by Murray Smith (1995) in a study of how we engage with characters in typical film genres and a broad representation of theories and analysis of film genres, cognition, and emotions can be found in Plantinga and Smiths (1999). Carl Plantinga (2009) takes up the aspect of audience in cognitive

film theory in an attempt to define the universality of the global experience of film genres. Finally, Grodal (2009) develops his basic theory into a more developed model of film experience (PECMA model) and situates it in both modern neurology and evolution theory.

Film and Genres

How many film genres are there and how do we define them and distinguish one genre from another? This is one of the basic questions raised by film genre theory, and this is also the question raised in more concrete, historical analysis of particular film genres and their development. Theoretically there are at one end theories that try to define genres as relatively stable and accepted narrative, stylistic, and thematic structures that transcend individual films and clearly influence and determine both production and audience response; at the other there are theories that stress the unstable nature of genres and the fuzziness of genre categories. Some theories argue for few and very basic generic prototypes with a variety of possible subgenres, that may vary and change historically, whereas practical criticism and the history of film genre categories often show that you can construct genres on the basis of many criteria, for instance, on content and dominant type of narrative action (war films, porno films), on specific stylistic features (film noir, musicals, animated films), on the basis of the expected audience (women's films, children's films), on the basis of emotional response (comedy, melodrama), etc.

Although genres are often related very clearly to the dominance of US cinema worldwide, national variations of genres are also very important and often have more universal influence, e.g., the Asian martial arts films or the British heritage film or costume drama. But although historical, national, and other variations are very common and the number of actual genres defined in some critical context is probably very large, then some basic narrative modes and generic fomulas are stronger and more permanent and universal than others.

Basic Film Forms and Universal Formulas

Kracauer (1960) defined two basic types of film: narrative films (story films), and nonnarrative films (nonstory films), which he divided into 'films of fact' and experimental/avant-garde films. He considered narrative film to be the most important and basic form. In a similar way Williams (1984) has suggested that before we talk about specific subgenres, we should perhaps talk about only three, very basic or principal film forms: narrative film, avant-garde/experimental film, and documentary film. This division is also supported by Bordwell's (1985) division of films belonging to the more canonical, classical narration, which dominates both Hollywood mainstream formats and national European mainstream films; art cinema narration with its more fragmented and less linear models of narration found in, for instance, European neorealism, New Wave cinema, and modernist and postmodern films; and finally nonnarrative forms such as documentary. All three forms have very distinct institutional backgrounds and are normally very easily recognized by the general audience, because distribution and indexing of art cinema, for example, is linked to special cinemas, festivals, and a special critical discourse, and documentary films have a totally separate distribution and are indexed much more for their special reference to reality.

Classical Hollywood and Beyond

Very often the classical Hollywood genres form the basis of what is generally understood as the fundamental genres and these would most typically include melodrama, comedy, action–adventure, science fiction, horror, Western, film noir, crime genres, and musical. But each of these categories has subcategories, such as the crime genre, which is often further divided into gangster films, thrillers, hard-boiled detective films, spy films, police films, etc. A popular Internet site like the Internet movie database lists no less than 21 fiction film genres, mostly based on what the films are about, but also with some style-oriented or audience-defined categories: action and adventure. Among the first are action, adventure, mystery, war films, or historical films. More style-oriented categories are animation, film noir, or musical and among the more audience-based categories we find family films. Neither the historical, critical, nor the popular taxonomy of genres are defined in very rigid and stable forms, although some basic forms are clearly present. As indicated in the ritual approach, the cognitive approach, and in most studies of the classical Hollywood system, it is however possible to define a number of more general generic forms and prototypes, which then often have different subforms and furthermore can change and vary historically.

Several models of typical historical developments of genres have been suggested in film genre theory. Neale ([1990]1995), for instance, discusses the evolutionary model, used by Schatz (1981). In this model genres move from an experimental stage to a classical stage and to stages of more opaque, baroque, ironic, or even metatextual modes in which the basic formula elements are so familiar that directors and audiences have a more playful and reflexive relation with them. Neale points to another model in which genres occupy shifting positions as dominant, and run through phases of canonization, automation, and reshuffling. In recent theories on genre mixing and intertextuality it is furthermore often indicated that it is an increasingly common strategy by both the producer and the director to try to mix genres, to have maximum audience effect, and that furthermore the general audience today is much more used to genre mixing and intertextual and metatextual elements even in mainstream films.

Nonfiction Films and Genre

Many theories and discussions of film genres relate mainly to fictional film; however, both nonfiction film genres and the distinction between fiction and nonfiction are very important for the understanding of the fundamental film genres. Historically, the term documentary was defined by John Grierson in 1926 as "creative treatment of actuality." This definition points to the fact that documentary films, like all nonfiction genres, have a special and more direct reference to reality than even the most realistic fictional film, but that on the other hand all documentary films are also a product of rhetorical and aesthetic

choices. Plantinga (1997) gives a cognitive and pragmatic definition of the distinction between nonfiction and fiction: Nonfiction is related to those discourses that make explicit claims about reality and fiction to discourses that do not, a distinction between an assertive and a fictional mode, in which the spectator treats the reference from text to reality in two fundamentally different ways. This distinction between two modes is imbedded in our cognitive framework and is related to the distinction between a play mode and a reality mode.

Historically, aesthetically, and rhetorically the documentary film has developed many different ways of representing and talking about reality. Barnouw (1993) uses different communicative roles to define different rhetorical purposes in documentary: explorer, reporter, advocate, poet, promoter, and observer. Renov (1993) in his theorizing about documentary forms makes a distinction between four fundamental functions of documentary films: to record, reveal, and preserve; to persuade or promote; to analyze and interrogate; and to express. Combinations of each of these communicative roles or purposes of course result in different subgenres of documentary.

Documentary genre theory offers several ways of defining the basic genres, but one of the most widely cited is Nichols's (1991) definition of four fundamental modes of representation: *the expository documentary* (with direct address to the viewer, often through authoritative voice-over and based on explicit arguments); *the observational documentary* (often based on simply edited visual observation and people expressing views in their observed reality, but without explicit commentary); *the interactive documentary* (based on a more active intervention and interviewing into reality, but still without an authoritative voice and comment); and *the reflexive documentary* (with a dominance of more self-reflexive and poetic strategies foregrounding the film's own strategy and relation to reality). Bondebjerg (2014a,b) defines four basic modes of documentary using both cognitive and pragmatic theory: the authoritative mode, the observational mode, the dramatized mode, and the poetic reflexive mode. If these modes can be taken as basic prototypes, then of course different subgenres can mix these modes and rhetorical strategies and it should also be added that mixing of fictional forms and documentary forms also has a long history. Docudrama and drama documentary or even more complicated fact–fiction formats have developed strongly in recent film and TV culture adding to the aesthetic and rhetorical variations of documentary forms or 'faked' documentary forms (Rosenthal, 1999).

See also: Art and Culture, Economics of; Art and Socialisation; Culture, Sociology of; Documentary and Ethnographic Film; Film History; Film and Video Industry; Semiotics; Television: General; Television: History.

Bibliography

Altman, R., [1984]1995. A semantic_syntactic approach to film genre. In: Grant, B.K. (Ed.), Film Genre Reader. University of Texas Press, Austin, TX, pp. 26–41.
Altman, R., 1999. Film/Genre. British Film Institute, London.
Barnouw, E., 1993. Documentary. A History of the Non-fiction Film. Oxford University Press, New York.
Bazin, A., 1971a. The evolution of the Western. In: Bazin, A. (Ed.), What Is Cinema? vol. II. University of California Press, Berkeley, CA, pp. 149–158.
Bazin, A., 1971b. The Western: or the American film par excellence. In: Bazin, A. (Ed.), What Is Cinema?, vol. II. University of California Press, Berkeley, CA, pp. 140–149.
Bondebjerg, Ib, 2014a. Engaging with Reality. Documentary and Globalization. Intellect, Bristol.
Bondebjerg, Ib, 2014b. Documentary and cognitive theory: narrative, emotion, memory. Media and Communication 2 (1), 13–22.
Bordwell, D., 1985. Narration in the Fiction Film. University of Wisconsin Press, Madison, WI.
Bordwell, D., Staiger, J., Thompson, K., 1985. The Classical Hollywood Cinema. Film, Style and Mode of Production to 1969. Routledge, New York, London.
Braudy, L., 1976. The World in a Frame: What We See in Films. Anchor, New York.
Buscombe, E., [1970]1995. The idea of genre in the American cinema. In: Grant, B.K. (Ed.), Film Genre Reader. University of Texas Press, Austin, TX, pp. 11–26.
Carroll, N., 1990. The Philosophy of Horror, or Paradoxes of the Heart. Routledge, New York.
Carroll, N., 1999. Film, emotion and genre. In: Plantinga, C., Smith, G.M. (Eds.), Passionate Views. Film, Cognition, and Emotion. Johns Hopkins University Press, Baltimore, pp. 21–48.
Cawelti, J., 1976. Adventure, Mystery and Romance. University of Chicago Press, Chicago.
Clover, C., 1992. Men, Women and Chain Saws: Gender in the Modern Horror Film. Princeton University Press, Princeton, NJ.
Comolli, J.-L., Narboni, J., et al., [1969]1992. Cinema/ideology/criticism. In: Mast, A., Cohen, M., Braudy, L. (Eds.), Film Theory and Criticism. Oxford University Press, New York, pp. 682–689.
Cook, P., Bernink, M. (Eds.), 1999. The Cinema Book, second ed. British Film Institute, London.
Davies, P.S., 2008. Out of the Movies: A History of Gay Cinema. Kamera Books, New York.
Fowler, C. (Ed.), 2002. The European Cinema Reader. Routledge, London.
Gledhill, C., 1987. Home Is Where the Heart Is. Studies in Melodrama and the Woman's Film. British Film Institute, London.
Griffith, R., 2009. Queer Cinema in Europe. Intellect, Bristol.
Grodal, T., 1997. Moving Pictures. A New Theory of Film Genres, Feelings and Cognition. Oxford University Press, Oxford, UK.
Grodal, T., 2009. Embodied Visions. Evolution, Emotion, Culture and Film. Oxford University Press, Oxford, UK.
Kaplan, A., [1978]1998. Women in Film Noir, rev. ed. British Film Institute, London.
Klinger, B., [1984]1995. Cinema/ideology/criticism revisited. In: Grant, B.K. (Ed.), Film Genre Reader. University of Texas Press, Austin, TX, pp. 74–91.
Kracauer, S., 1960. Theory of Film. Oxford University Press, London.
Mulvey, L., [1975]1992. Visual pleasure and narrative cinema. In: Mast, A., Cohen, M., Braudy, L. (Eds.), Film Theory and Criticism. Oxford University Press, New York, pp. 746–758.
Nichols, B., 1991. Representing Reality. Indiana University Press, Bloomington, IN.
Neale, S., 1980. Genre. British Film Institute, London.
Neale, S., [1990]1995. Questions of genre. In: Grant, B.K. (Ed.), Film Genre Reader. University of Texas Press, Austin, TX, pp. 159–187.
Plantinga, C., 1997. Rhetoric and Representation in Nonfiction Film. Cambridge University Press, Cambridge, UK.
Plantinga, C., 2009. Moving Viewers. American Film and the Spectator's Experience. University of California Press, LA.
Plantinga, C., Smiths, G.M. (Eds.), 1999. Passionate Views. Film, Cognition, and Emotion. John Hopkins University Press, Baltimore and London.
Renov, M. (Ed.), 1993. Theorizing Documentary. Routledge, New York.
Rosenthal, A. (Ed.), 1999. Why Docudrama? Fact-Fiction on Film and TV. Southern Illinois University Press, Carbondale, IL.
Schatz, T., 1981. Hollywood Genres: Formulas, Filmmaking and the Studio System. Random House, New York.
Smith, M., 1995. Engaging Characters. Fiction, Emotion, and the Cinema. Oxford University Press, Oxford.
Staiger, Janet (Ed.), 1995. The Studio System. Rutgers University Press, New Brunswick.
Todorov, T., [1970]1975. The Fantastic. Cornell University Press, Ithaca, NY.
Wellek, R., Warren, R., [1949]1956. Theory of Literature. Harcourt, Brace & World, New York.
Williams, A., 1984. Is a radical genre criticism possible? Quarterly Review of Film Studies 9 (2), 121–125.
Williams, L., 1989. Hard Core: Power, Pleasure, and the 'Frenzy or the Visible. University of Berkeley Press, Berkeley, CA.
Wright, W., 1975. Sixguns and Society: A Structural Study of the Western. University of California Press, Berkeley, CA.
Wood, R., [1977]1995. Ideology, genre, auteur. In: Grant, B.K. (Ed.), Film Genre Reader. University of Texas Press, Austin, TX, pp. 59–74.

Film History

Charles O'Brien, Film Studies/School for Studies in Art and Culture, Carleton University, Ottawa, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Film history has been an established field of study within higher education since the 1980s at least, defined by specialized journals, professional organizations, and conferences. Indicating film history's evolved condition is the extent to which film scholars have begun writing the history of their field of study. Each year, new discoveries stimulate revision of established scholarship just as they underline the incompleteness of the historical record, and hence the certainty that the study of film history will remain a work-in-progress.

As the art of film enters its second century, film historians not only encounter a longer and more heterogeneous film-historical past but do so from within contexts unimaginable during the 1930s, when the first wave of book-length film histories appeared. Film history has been an established field of study within higher education since the 1980s at least, with specialized journals, professional organizations, and conferences. Indicating film history's evolved condition is the extent to which film scholars have begun writing the history of their field of study (Polan, 2007; Grieveson and Wasson, 2008).

Academia's perpetual search for new objects of study, together with the ongoing discovery and restoration of lost films, has enabled a fundamental rethinking of key periods of film history. Foremost is the early cinema that had preceded the formation of Hollywood cinema, and which has been the focus of considerable research since the late 1970s (Abel, 2010; Grieveson and Kramer, 2004). Much of this research has concerned not only the production of films but their exhibition and critical reception, too, and it has entailed a far-reaching reinterpretation of the period. More recently, films produced outside of film history's familiar research domains of Western Europe and the United States have become a focus of study (Hill and Church Gibson, 2000), as in the studies of Bollywood cinema that have proliferated since 2000. Moreover, scholars are investigating not only live-action feature films but also various shots, including films screened in hospitals, schools, prisons, and other noncommercial venues. Animation, long a marginal area of study for film historians, is now receiving new attention in light of current changes in cinema linked to digital image making. New interpretations of film history are certain to occur as films continue to be found, identified, reconstructed, and made available to scholars. As new case studies accumulate, it is questionable whether world cinema can be said to have a history, if the term history is understood to imply a single narrative that might encompass the heterogeneity of the world's films and cinema practices.

Film as Art and Industry

Critics have often distinguished film from traditional arts such as painting, music, and literature by insisting that film is simultaneously an art and an industry. Cinema's industrialized processes differ from the artisanal, craft-based traditional arts in ways that separate the historiography of film from that of the other arts in three respects. First, films, as products of an industrial system, invite a variety of humanistic, psychological, and social scientific forms of analysis, ranging from traditional art-historical approaches centered on aesthetics and filmmakers' biographies to diverse economic, technological, sociopolitical, and quasianthropological investigations (Allen and Gomery, 1985; Lagny, 1992). Film history's object of study is rarely defined exclusively in aesthetic terms, as is evident in the long-standing practice of demarcating periods in film history by nonaesthetic events, such as the beginning of the First World War or the end of the Second World War. Even film histories concerned mainly with aesthetic questions typically also situate films in economic, technological, and social contexts; and certain film histories conceptualize film primarily as a technological, economic, and/or social phenomenon, to the point of excluding aesthetic considerations altogether, as in studies of the economics of national film industries.

Second, the industrial nature of much film production has encouraged modernist conceptions of historical change according to which film history moves forward, in a single direction, with new films counting as a technical if not aesthetic advance over old. This newer-is-better conception of change is promulgated by the entertainment industry itself, and it differs from the cyclical patterns of the histories of the established fine arts, where artistic movements are seen to rise, fall, and emerge again in the manner of the life cycle of a natural organism. Filmmaking, so evidently a function of technological, industrial, and commercial conditions and forces, invites conceptualization as an ongoing, forward-moving process, forever open to additional improvement. This conception of historical change as a modernizing process informs both the film industry's celebration of the latest advances in film technique and critics' nostalgia for the lost artistry of the filmmaking of earlier periods.

A third peculiarity of film as an object of historical research concerns the relatively short duration of cinema history, which has led film historians to reexamine the cinema's past in light of present-day developments (Bordwell, 1997). The focus on what the art form's present suggests of its past – and vice versa, what the art's past may reveal of its current possibilities – is unusual in the much longer histories of painting, poetry, and music, which typically are studied irrespective of present-day artistic developments (Belting, 1987). Film

historians, however, are concerned with an art whose past coincides with a recent and relatively short span of historical time, to the point that new understandings of film history have long been motivated by current developments in film practice. In the 1930s, for instance, the conversion to sound cinema had made it seem as if 'silent' films belonged to the historical past, thereby providing justification for the creation of film archives in Germany, Britain, the United States, France, and elsewhere. Something similar is happening in the early twenty-first century, as the introduction of digital film production and exhibition pushes historians to see the cinema's past in new ways. For instance, as computer-generated digital images supersede photochemical images, cartoons, and other nonphotographic animated films increasingly appear central rather than peripheral to film history. In any case, film-historical reasoning often displays a retrospective orientation in which the cinema's past comes up as a prelude to the cinema of the historian's present, which, for better or worse, provides a context for making sense of the historical past.

A Chronology under Revision

What follows below is a chronology of world cinema that discusses historiographical developments pertaining to specific periods. Keep in mind that the chronology in question concerns developments primarily in the United States and Europe, and may well require revision insofar as the focus of inquiry shifts toward cinema in other parts of the world, where analogous events happened, but often later.

Early Cinema, 1895–1906

Research on the so-called early cinema of 1895–1906 exemplifies contemporary film historiography's concern with films as embedded within contemporaneous contexts of production, exhibition, and reception. Early cinema is taken to begin in the mid-1890s with the first screenings of films by inventors of motion picture cameras and projectors, such as the Lumière brothers in France and the Skladonowsky brothers in Germany in late 1895, and, in early 1896, by Birt Acres in London and Thomas Edison in the United States. The period continues until roughly 1906, when narrative films began to dominate the film industry's output and the exhibition sites were no longer mainly fairgrounds, vaudeville houses, and rented theaters but nickelodeons and other permanent venues intended specifically for motion picture screenings.

Accounts of early cinema often center on the development of techniques of film editing, whose rapid evolution can be traced in films at the turn of the twentieth century. The films of inventors such as Edison and the Lumières were made up of only a single shot, usually lasting the length of a standard reel of film. But by 1899, in response to a greatly increased demand for motion pictures, film companies in France, England, and the United States began producing narrative films made up of multiple shots. Historians have analyzed these films as attempts to solve the problem of how separate shots might be staged and connected together to create the impression of continuous story action.

Film historians initially treated early multiple-shot films such as *The Great Train Robbery* (Edwin Porter, 1903) as landmarks in the development of the system of stylistic conventions associated with studio-era Hollywood (Jacobs, 1939). Since the late-1970s, however, a very different history of the cinema of 1895–1906 has been written (see Elsaesser and Barker, 1990). Instead of a forerunner of the commercial narrative cinema of the late 1910s, early films are examined in the context of their function relative to vaudeville routines, illustrated song slides, public lectures, and other forms of popular spectacle entertainment. Examining how musical performance, spoken narration, and other exhibition practices affected the viewing experience, recent studies of early cinema bring forth dimensions of early films occluded within traditional developmental accounts. An example is the theatrical or presentational aspect of early cinema, which became a focus of historical study in the 1980s. Investigating early cinema's practices of direct address, adapted from vaudeville, the fairground, the magic show, and other forms of popular theater, historians have come to understand early films less as primitive attempts at classical storytelling and more as performances within the context of live, spectacle entertainment at the turn of the century (Gaudreault, 2010). Innovations in early cinema research continue to inspire research into other periods of film and media history.

The Transition to Classicism

The cinema of the period 1906–17 has traditionally been conceptualized as a transition from early cinema to the feature films produced at the Hollywood studios of the late-teens. Key industrial developments include: a shift in editorial control away from exhibitors and toward producers; the creation of exchanges that rented film prints to exhibitors rather than sold them outright; the organization of film production according to a hierarchical division of labor that gave producers authority over directors and camera operators; the establishment of Hollywood as a production center; and the cinema's further expansion worldwide, marked, for instance, by the emergence of Denmark and Italy as major film-producing countries. Style-related developments of 1906–17 include: the dominance of the fiction film over the vaudeville-like attractions and documentary 'actualities' of early cinema; the emergence of celebrity performers or 'stars'; and a continuing increase in film length, leading, by 1914, to the establishment of the feature-length multireel film as the default format for commercial moviemaking.

Accounts of the transition to classicism often stress the development of devices of film editing (Salt, 2009). As early as the late-aughts, critics noted that editing enabled forms of acting unique to cinema, with actors' performances constructed through a planned assembly of numerous shots rather than captured in a lengthy, uninterrupted run of the camera. Editing allowed the filmmaker to direct the viewer's attention in a manner unimaginable in traditional theatrical entertainments, just as close placement of the actors relative to the camera allowed for a new, naturalistic acting technique. During the 1920s, filmmakers and theorists in France, the Soviet Union, and elsewhere argued that editing had differentiated film from theater to the point of establishing film as a singular art.

Accounts of the cinema of the 1910s have emphasized the work of director D.W. Griffith, who has been credited with having conceived of a motion picture less as a performance staged for the camera than as a composite of shots, staged, filmed, and assembled so as to construct a unified dramatic world. In some 480 short one- and two-reel films directed by Griffith between 1908 and 1913, devices such as the close-up, crosscutting, and the eyeline match are employed as elements within what film theorists have described as a storytelling 'language' unique to cinema. In films such as *The Girl and Her Trust* (Griffith, 1912), a one-reel film made up of nearly 130 shots, elaborate last-minute rescues include events occurring in three separate locales, cut together in rhythms that accelerate or slowdown relative to the intensity of the dramatic action. Since 1990, historians have challenged the idea that Griffith's editing provided the basis for the classical style that would take form in Hollywood circa 1917 by proposing instead that Griffith was a transitional figure whose style was peculiar to him in important respects (Gunning, 1991).

Also in the 1990s, the rediscovery of early films from Russia and Germany led historians to question the traditional emphasis in film historiography on editing practices and argue that a proper appreciation of cinema requires careful attention to mise-en-scène as well, to the staging of the dramatic action through the arrangement of actors, settings, and props within the camera's visual field (Brewster and Jacobs, 1997). Such attention is especially warranted for films made outside the United States, whose stories, in crucial respects, were told through techniques of staging rather than cutting. The film industries of Italy, Russia, France, Denmark, Sweden, and Germany, for instance, produced hundreds of narrative films with scenes playing out in extended long takes. Using far fewer shots than American films, European films of the 1910s often delineate story events through the careful choreography of multiple actors' movements within the film's frame. Ensemble performances in which the dramatic action involves several actors within a common framing situate character interactions within a shared environment; European films of the 1910s might thus be said to exhibit a social dimension less available to the contemporaneous American cinema's stress on linear, physical action.

National Cinemas

By 1918, with the disruption of Europe's film industries by the First World War, Hollywood had established itself as the world's foremost film exporter, with distribution branches in South America, Australia, the Far East, and Europe (Thompson and Bordwell, 2010). In response to Hollywood's new preeminence, Europe's film industries imposed quotas and other trade restrictions on Hollywood films, and drew inspiration from modernist trends in the arts to invent distinctive national film styles. From Expressionism and the Neue Sachlichkeit ('New Objectivity') in Germany to impressionism and surrealism in France and constructivism in the Soviet Union, the major national art cinemas of the 1920s inspired much reflection on film as an art, manifest, for one, in a large and sophisticated body of film-theoretical writing. In France, where there was no clear industrial division between commercial cinema and art cinema, avant-garde techniques were integral to many mainstream films.

The experimental film movements evolved within a new international film culture, organized around specialized journals, film clubs, and art-cinema theaters. An unprecedented intellectual interest in film stimulated theoretical writings and manifestos stressing the need to overcome the apparent theatricality of established film styles; and numerous filmmakers in Europe and elsewhere undertook radical explorations of film's artistic potential. As a consequence, the major European art films of the 1920s frequently exhibit many more shots than had earlier European films, along with an array of experimental cinematographic devices, from filtered lenses to superimposed images, double-exposures, and split-screen compositions.

Certain national film movements of the 1920s became known for exploring film's potential for radical alternatives to conventional scenography. Prominent examples include: the montage films of the Soviet Union, such as *The Battleship Potemkin* (Sergei Eisenstein, 1925) and *Man with a Movie Camera* (Dziga Vertov, 1929), famous for their aggressive experimentation in film editing; and French Impressionist films, from the pathbreaking *Fièvre* (Louis Delluc, 1921) to diverse films by directors such as Germaine Dulac, Jean Epstein, and Abel Gance, who employed multiple superimpositions and other optical effects to simulate the flux of diverse moods, feelings, and emotions. German films such as *The Cabinet of Dr. Caligari* (Robert Weine, 1920) and *Metropolis* (Fritz Lang, 1927) won international acclaim for their spectacular, studio-created settings, while *The Last Laugh* (F.W. Murnau, 1924) was celebrated for optical effects and elaborate moving camera shots that suggest a character's psychological 'point of view.' By the mid-1920s, with increasing numbers of films, artists, and critics crossing national borders, an international avant-garde had taken form, encompassing filmmakers in Europe, North America, and Japan, and exemplified by 'international' films, such as *Page of Madness* (Teinosuke Kinugasa, 1926) and *La Passion de Jeanne d'Arc* (Carl Dreyer, 1928), which confounded national labels with their unusual mix of styles (Thompson and Bordwell, 2010).

Questions remain open concerning how new research might alter the historiography of world cinema in the 1920s. One topic concerns the popular cinemas of the major Western countries, which has long been neglected in favor of the export-oriented art films. A further topic concerns cinema outside the United States, France, and Germany. In the 1920s, for instance, Japan and India were among the world's largest film-producing countries; yet, as with many other national cinemas during the years between the World Wars, the films were rarely distributed beyond the national and regional film markets, and extremely few have been preserved by distributors and archives. Although Japan's large film industry dates from the 1920s, research on the history of Japanese cinema began to flourish only in the 1970s, with the rediscovery of the pre-World War II films of Yasujiro Ozu, Kenji Mizoguchi, and other prominent art-cinema directors. Assessments of world cinema during the 1920s are certain to undergo revision, pending further research into the variety of film industries of the period, and their sometimes asynchronous chronologies. Pending the availability of studies of popular cinemas from around the world,

historical assessments of world cinema in the 1920s remain incomplete and provisional.

The Conversion to Sound Cinema

The introduction of recorded sound to cinema has attracted much analysis and commentary, which began as early as the late 1920s with the international success of films such as *The Singing Fool* (Lloyd Bacon, 1928), which provoked debate worldwide concerning the future of cinema. In the summer of 1928, as the film industry in the United States began its conversion to sound, observers described synchronous sound as nothing less than an artistic, technological, and economic revolution. The sense of crisis went on to inform subsequent historical accounts of the conversion period, which have traditionally emphasized traumatic changes such as: the sudden collapse of the careers of established actors and other talent; the flood of new stage, musical, and technical personnel into the film industry; the technical and aesthetic limitations of early sound films; the refinancing of the film industry by large banks; and the curtailment of avant-garde experimentation brought about by the high costs of sound-film production. The condition of crisis encouraged partisan assessments, whether in the form of Hollywood's celebration of sound cinema as a great democratic advance or critics' nostalgia for the vanished art of silent film. In either case, film history had seemed to start over again with the introduction of recorded sound.

Concerning film aesthetics, sound conversion is said to have produced two main consequences. First, the first sound films sometimes displayed relatively long takes, stationary framings, and immobile actors who recited dialogue in styles familiar to the popular stage, and thus counted for numerous critics as an artistic regression from the ambitious art films of the mid-1920s. Challenging the idea that sound cinema amounted to an artistic catastrophe, established histories of the early sound years have focused on exceptional films such as *Applause* (Rouben Mamoulian, 1929), *Sous les toits de Paris* (René Clair, 1930), and *M* (Fritz Lang, 1931), whose innovations in multitrack recording, postsynchronization, playback, and other techniques are credited with having restored movement to the movies. More recently, historians have expanded their purview to encompass vast numbers of ordinary films of the time (Crafton, 1997).

Second, the technical demands of sound-filmmaking appeared to push filmmakers everywhere toward adoption of important aspects of Hollywood's approach to style and narration. For critics familiar with the distinctive national film styles and transnational film movements of the 1920s, sound films often seemed to belong to a common family, regardless of where they were made. Even the Soviet Union, famous during the 1920s for its counter-Hollywood film culture, was, by 1934, producing films that looked and sounded much like commercial narrative films made elsewhere. By the late 1930s, filmmakers, in countries throughout the world, including Japan and China, were making sound movies in imitation of Hollywood's approach to storytelling.

In assessing sound's impact on film style, it is important to distinguish between filmmaking in Hollywood, which had already undergone full industrialization prior to sound conversion, and filmmaking in other countries, which was still largely artisanal when sound was introduced. Whatever the personal, technical, and administrative upheaval, Hollywood's processes of serial manufacture remained largely unchanged, which enabled a continuation of important aspects of silent-era film technique. Most notably, Hollywood's sound-era editing style remained essentially consistent with established methods for constructing a unified, linear narrative (Bazin, 1967). After a brief period of experimentation, the new sound techniques were brought into line with artistic precedent (Lastra, 2000). For instance, in 'backstage musicals' such as *The Broadway Melody* (Harry Beaumont, 1929), song-and-dance sequences came forward less as recordings of self-contained performances than contributions to narratives about the offstage lives of show-business personalities. By the summer of 1931, with the industry-wide adoption of multitrack recording and mixing, the foregrounding of sound technique as an artistic novelty had been abandoned in favor of the emphasis on storytelling characteristic of Hollywood films of the late-silent period.

In many countries, however, sound conversion had entailed fundamental industrial change, and hence consequences for film aesthetics that were more radical than in Hollywood. In India and Mexico, for instance, where filmmaking had been minimally industrialized, complete restructuring of the industry occurred, and new, sound-era genres, such as musicals that drew upon indigenous and regional entertainment traditions, came to define the national film output. Research into the worldwide conversion to sound remains in its early stages. As with cinema during the preceding decade, relatively little is known about the output of many national film industries of the 1930s.

Postwar Internationalism

The decade after the Second World War is known for the emergence of a revived international art cinema. Prominent institutional developments include: the creation of international film festivals in cities such as Cannes, Venice, Berlin, Edinburgh, and Tokyo; increased government funding of film archives, thereby enabling expansion worldwide of the International Federation of Film Archives; new film journals devoted to international coverage; and the opening of special art-house movie theaters in cities throughout the world. These developments had prewar antecedents that led observers to note parallels between the new art cinema and the international avant-garde of the 1920s. Where the new film culture differed was in the extent to which its global network of festivals functioned independently from nation-state institutions. An example can be found in the worldwide reception of Italian neorealism, whose status as the preeminent postwar film movement depended crucially on critical acclaim from outside Italy. More popular abroad than at home, films such as *Rome, Open City* (Roberto Rossellini, 1945), and *Bicycle Thieves* (Vittorio De Sica, 1948) drew admiration from filmmakers from every continent. Essential here were innovations in narrative. In contrast to the commercial cinema's linear and formulaic, genre-based plots, the neorealist films emphasized the acts and events of ordinary life, including chance occurrences, routine, nondramatic actions, and inconclusive chains of events. Neorealist films were also acclaimed for the documentary-like 'realism' manifest in their use of nonprofessional

actors, improvised performances, actual locations, and natural lighting.

Neorealism is said to have affected filmmaking in Asia, Africa, and Latin America more so than any other film movement of the sound era, and it has commanded attention from historians as a beginning for prominent subsequent movements and tendencies in world cinema. Neorealism offered a coherent, low-budget alternative to commercial cinema, with great potential both for naturalistic social representation and for sociopolitical allegory. Filmmakers who have cited Italian neorealism as a personal inspiration include: Fernando Birra and Tomás Gutiérrez Alea, leading directors during the 1960s in the film industries of Argentina and Cuba, respectively; Satyajit Ray, the famous Bengali director from the Indian subcontinent; Bollywood pioneer Raj Kapoor; and Nelson Pereira dos Santos, founder of the Cinema Novo movement in Brazil.

The emergence of innovative film movements in Latin America, Asia, and, by the late 1960s, in Africa, with the rise to prominence of the Senegalese director Ousmane Sembene, had been conditioned by two basic developments. First, in 1947, movie attendance in North America had begun to fall, and would continue to do so for the next two decades. The decline became apparent just as Hollywood faced the landmark 1948 decision by the U.S. Supreme Court that required that the five major studios sell off their theater chains, and that both major and minor studios cease a variety of monopolistic practices. In response to the economic and political adversity, the major Hollywood studios produced fewer films in the 1950s, and they began targeting smaller, more specific markets. Second, during this same period, in many countries around the world, booms in film attendance occurred. Attendance increases were especially pronounced in the so-called developing world, where television had not yet emerged as an entertainment alternative. Responding to the new demand for motion pictures, major film industries emerged in Egypt and Hong Kong, while the new nations of Burma, Pakistan, South Korea, and the Philippines also began producing films on an industrial scale. In the 1950s, India became one of the world's largest film producers, making hundreds of films exclusively for the South Asian film market. By the mid-1960s, some 60% of all of the world's feature films came from Asia, Latin America, and Africa (Thompson and Bordwell, 2010).

Coinciding with the worldwide boom in film production, young filmmakers worked to distinguish themselves from the previous generation by emulating new currents in world cinema. 'New waves' and 'young cinemas' took form in diverse countries, beginning in the late 1950s and continuing through the 1960s. Most prominent was the French New Wave, known for films such as *Les quatre cent coups* (François Truffaut, 1959), *Breathless* (Jean-Luc Godard, 1960), and *Hiroshima mon amour* (Alain Resnais, 1959), which featured inventive adaptations of conventions drawn from Hollywood and European genre films and also from the international art cinema. Further important new wave movements emerged in the 1960s in Japan, Czechoslovakia, Portugal, and, indeed, in virtually all countries, in one form or another. A famous example is the Young German Cinema, whose self-conscious opposition to Germany's established film culture opened the way for the international prominence in the 1970s of directors such as R.W. Fassbinder, Werner Herzog, and Wim Wenders.

Film Historiography since the 1970s

The new film culture of the 1960s had a profound impact on the writing of film history. An example can be found in the notion of the film director as artist or 'auteur.' Formulated in Paris in the 1950s by a group of young film critics, some of whom went to become key directors of the French New Wave, the notion of the auteur attributed to certain film directors the kind of creative role traditionally accorded to novelists, playwrights, painters, and composers. A precedent for the filmmaker-as-artist concept can be found in certain currents in film critical writing in the 1920s, which had celebrated the artistry of D.W. Griffith, Charlie Chaplin, F.W. Murnau, Erich von Stroheim, and others. But the auteur critics extended the sphere of art to encompass a remarkable diversity of films, including, in some cases, low-budget sound-era genre pictures. Provocatively comparing commercial movie directors to Mozart and Shakespeare, the auteur critics expanded the canon of great film artists to include current commercial directors such as John Ford, Orson Welles, Alfred Hitchcock, and Howard Hawks.

The 1970s marked the integration of film study into university curricula in Western countries, which opened a new, professionalized era in film historiography marked by an ongoing critical reexamination of established film history. A case in point is the new historiography of silent cinema. While previous film historians had often relied on other historians' accounts, together with personal memories of films, the new generation of archivists and professional scholars studied large collections of film prints, comprised of recognized classics as well as many lesser known or forgotten works. These historians also examined a wide variety of print documents, from film reviews to trade publications, film company catalogs, and paper prints of films from the U.S. Library of Congress. Silent-film festivals, conferences, and journal publications facilitated debate and discussion among historians. The result of these developments was a thorough reconceptualization of the early silent period, which, in turn, has served to inspire historians working on later periods.

As cinema now enters its second century as a thriving art and entertainment, new technological changes provoke new ways of seeing the cinema's past. At annual festivals, such as those held in Bologna and Pordenone, large bodies of films, representing the output of diverse directors, actors, studios, film movements, and national cinemas, are made accessible to critics and historians. Some of this material has found it way to DVD and other formats for digital distribution, which has vastly increased the availability of these films to researchers. With so much research material now available online, the nature of film-historical research is changing (Loiperdinger, 2003). Films virtually impossible to see a few decades ago are now readily available in beautiful restored copies. As old films continue to be restored and made available to scholars, significant changes in film historiography continue to occur. Online archives, chat groups, and blogs are making possible further interchange between scholars as well as new sorts of research projects – as are computer-based tools such as the cinemetrics interface, which allows the critic to perform easily

various statistical measurements of film editing (Tsivian, 2009). Each year, new discoveries stimulate revision of established scholarship just as they underline the incompleteness of the historical record, and hence the certainty that the study of film history will remain a work-in-progress.

See also: Documentary and Ethnographic Film; Film and Video Industry; Film: Genres and Genre Theory.

Bibliography

Abel, R., 2010. Encyclopedia of Early Cinema. Routledge, London and NY.

Allen, R., Gomery, D., 1985. Film History: Theory and Practice. Random House, New York.

Bacon, L. (Director), 1928. The Singing Fool [Motion Picture]. Warner Brothers, United States.

Bazin, A., 1967. The evolution of the language of cinema (Hugh Gray, Trans.). In: What Is Cinema? University of California Press, Berkeley and LA, pp. 23–40.

Beaumont, H. (Director), 1929. The Broadway Melody [Motion Picture]. Metro Goldwyn Mayer, United States.

Belting, H., 1987. The End of the History of Art? University of Chicago Press, Chicago and London.

Bordwell, D., 1997. On The History of Film Style. Harvard University Press, Cambridge, Massachusetts, and London.

Bordwell, D., Staiger, J., Thompson, K., 1985. The Classical Hollywood Cinema: Film Style and Mode of Production to 1960. Columbia University Press, New York.

Brewster, B., Jacobs, L., 1997. Theatre to Cinema: Stage Pictorialism and the Early Feature Film. Oxford University Press, Oxford.

Clair, R. (Director), 1930. Sous les toits de Paris [Motion Picture]. Tobis Films Sonores, France.

Crafton, D., 1997. The Talkies: America's Transition to Sound, 1926–1931. University of California Press, Berkeley and LA.

De Sica, V. (Director), 1948. Bicycle Thieves [Motion Picture]. Produzioni De Sica, Italy.

Delluc, L. (Director), 1921. Fièvre [Motion Picture]. Alhambra Film, France.

Dreyer, C. (Director), 1928. The Passion of Joan of Arc [Motion Picture]. Société Générale des Films, France.

Eisenstein, S. (Director), 1925. The Battleship Potemkin [Motion Picture]. Goskino, Soviet Union.

Elsaesser, T., Barker, A. (Eds.), 1990. Early Cinema: Space, Frame, Narrative. British Film Institute, London.

Gaudreault, A., 2010. Film and Attraction: From Kinematography to Cinema. University of Illinois Press, Urbana and Chicago.

Godard, J.-L. (Director), 1960. Breathless [Motion Picture]. Films Impéria/Georges de Beauregard/Société Nouvelle de Cinématographie (SNC), France.

Grieveson, L., Kramer, P. (Eds.), 2004. The Silent Cinema Reader. Routledge, London and NY.

Grieveson, L., Wasson, H. (Eds.), 2008. Inventing Film Studies. Duke University Press, Durham and London.

Griffith, D. (Director), 1912. The Girl and Her Trust [Motion Picture]. Biograph, United States.

Gunning, T., 1991. D.W. Griffith and the Origins of American Narrative Film. University of Illinois Press, Urbana and Chicago.

Hill, J., Church Gibson, P., 2000. World Cinema: Critical Approaches. Oxford University Press, Oxford.

Houston, P., 1994. Keepers of the Frame: The Film Archives. British Film Institute, London.

Jacobs, L., 1939. The Rise of the American Film. Columbia University's Teachers College, New York.

Kinugasa, T. (Director), 1926. Page of Madness [Motion Picture]. Kinugasa Productions/National Film Art/Shin Kankaku-ha Eiga Renmei Productions, Japan.

Lagny, M., 1992. De l'histoire du cinéma: Méthode historique et histoire du cinéma. Clin, Paris.

Lang, F. (Director), 1927. Metropolis. [Motion Picture]. Universum Film (UFA), Germany.

Lang, F. (Director), 1931. M [Motion Picture]. Nero Film, Germany.

Lastra, J., 2000. Sound Technology and the American Cinema: Technology, Perception, and Modernity. Columbia University Press, New York.

Loiperdinger, M. (Ed.), 2003. Celluloid Goes Digital. Historical-critical Editions of Films on DVD and the Internet. Wissenschaftlicher Verlag, Trier.

Mamoulian, R. (Director), 1929. Applause [Motion Picture]. Paramount, United States.

Murnau, F. (Director), 1924. The Last Laugh [Motion Picture]. Universum Film (UFA), Germany.

O'Brien, C., 2005. Cinema's Conversion to Sound: Technology and Film Style in France and the U.S. Indiana University Press, Bloomington.

Polan, D., 2007. Scenes of Instruction: The Beginnings of the U.S. Study of Film. University of California, Berkeley and LA.

Porter, E. (Director), 1903. The Great Train Robbery [Motion Picture]. Edison, United States.

Resnais, A. (Director), 1959. Hiroshima mon amour [Motion Picture]. Argos Films, France.

Rossellini, R. (Director), 1945. Rome, Open City [Motion Picture]. Excelsa Film, Italy.

Salt, B., 2009. Film Style and Technology: History and Analysis, third ed. Starword, London.

Thompson, K., Bordwell, D., 2010. Film History: An Introduction, third ed. McGraw Hill, New York.

Truffaut, F. (Director), 1959. Les quatre cents coups [Motion Picture]. Films du Carosse/Productions Sédif, France.

Tsivian, Y., 2009. Cinemetrics, part of the Humanities' Cyberinfrastructure. In: Ross, Michael, Grauer, Manfred, Freisleben, Bernd (Eds.), Digital Tools in Media Studies. Transcript Verlag, Bielfeld, pp. 93–100.

Vertov, D. (Director), 1929. Man with a Movie Camera [Motion Picture]. All-Ukranian Photo and Cinema Administration (VUFKA). Soviet Union.

Wiene, R. (Director), 1920. The Cabinet of Dr. Caligari [Motion Picture]. Decla-Bioscop, Germany.

Finance, Anthropology of

Karen Ho, University of Minnesota, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The anthropology of finance is composed of an interdisciplinary set of approaches for analyzing the mechanisms, worldviews, networks, and far-reaching socioeconomic effects of finance. Given the influence of financial values and practices, not to mention the tendency for academic and public approaches to finance to further obscure its workings, anthropology's ethnographic toolkit has a crucial role to play in unpacking the cultural specificities and underpinnings of finance. Indebted to and in conversation with many subfields, from economic sociology to political economy, the anthropology of finance navigates the theoretical and methodological challenges of abstraction and old binaries of economy and society to investigate the contours and consequences of finance and financialization.

Introduction

The anthropology of finance has engaged broadly with the social dynamic of financialization, as well as the importance and centrality of financial markets, institutions, values, and practices in the global social economy and everyday life. Given the pervasive influence of finance in our world, coupled with the production of highly specialized knowledge among financial actors and networks, the anthropology of finance has made important contributions toward a deeper understanding of the products, practices, ideologies, and socioeconomic consequences of finance and financialization. In particular, anthropology has brought to the study of finance its fine-grained ethnographic methods of participant observation, long-term immersion, and polymorphous engagements with natives' points of view, practices, and everyday lives.

These anthropological approaches are especially useful given that finance as an ethnographic site is not only characterized by inaccessibility to outsiders, specialized knowledge, and continual change and innovation, but is also framed by the larger culture as both mystifying and acultural. While certainly all sites of ethnographic investigation should be understood as possessing nuance and complexity, finance has perpetuated its rarified air through both 'inside' and 'outside' representations of its opacity. Anthropology thus bears a greater responsibility to use its toolkit to unpack the mechanisms, relations, and effects of finance and financialization in a way that resists conflation with broader cultural assumptions.

By finance, I refer to a constellation of priorities, practices, and ideologies that, in the context of a market economy, are mobilized to search for and convert already existing assets (including both tangible possessions, such as corporations and real estate, and personal attributes and status markers, such as educational background and credentials) into income streams, investments, or immediate profits. In order to 'realize' these assets as liquid monetary forms, intersecting cultural, material, political, and legal transformations need to occur such that these often nonfinancial assets are emplaced into commoditized spaces of exchange that allow for the conversion of these assets. For example, when relocated and reimagined as financial assets within spaces of exchange, sites such as large corporations, which were previously understood to be long-term social institutions, are 'converted' into their stock prices and thus become easier to buy, sell, and/or disassemble. Similarly, in this commoditized space, assets such as homes and mortgages are bundled, repackaged, and converted into 'collateralized debt obligations,' which are financial instruments sold as high-yield investments using the homes themselves as collateral and mortgage holders' payments as income streams. A social economy characterized by financial dominance thus actively searches for and converts assets into prices or securitized debt so that they can be bought and sold, and it is through these acts of exchange that financial institutions, actors, and interests amass influence and generate income.

It is important to emphasize that 'financialization' refers to the growing influence of finance, specifically the translations and interactions between financial modes of apprehending the world and other social domains. Institutions, natural resources, governmental entities, education, retirements, and so on are all increasingly dependent on financial products, measurements, and values. The processes of financialization, moreover, are uneven, specific, and contingent, creating major structural changes in the social landscape. This 'scaling up' does not only refer to the far-reaching influence of finance institutionally, regionally, and globally, but also to its remaking of 'individual' subjectivities and selves. Financialization, as Randy Martin (2002) shows in *Financialization of Everyday Life*, is concomitantly about the restructuring of the self through the alteration of intimate beliefs and practices. While the expanding sphere of finance might seem recent as well as deeply intentional, it is crucial to recognize that greater financialization is also a story of unintended consequences. For different but also intersecting reasons, policymakers, business advocates, and reformers alike turned to finance over the course of the twentieth century as an alternative site 'outside' of the state and its sedimented exclusions to attempt to both solve credit discrimination and fund increasing middle-class consumption, as well as obscure (and temper) growing sociopolitical unrest, declining wages, and struggles for inclusion (Hyman, 2012; Krippner, 2012; Ott, 2011; van der Zwan, 2013). That finance heightened socioeconomic inequalities and produced new kinds of desires, dependences, and precarity without solving poverty and disenfranchisement is precisely one of the central problems of the twenty-first century.

Specifically, in the US corporate context, financialization has meant that the growth of financial activities has become a key source of profits for financial and nonfinancial corporations alike. By the 1980s, American financial institutions had shifted away from funding the production of goods and services toward activities and financial transactions such as consumer lending, mergers and acquisitions, and trading securities and derivatives. Meanwhile, nonfinancial corporations increasingly shied away from long-term fixed investments in plant, equipment, and research and development. Instead, they too diverted capital to finance. The result was a redistribution of wealth toward certain investors and financial advisors, an explosion in the compensation of financiers and in stock-related incentives for corporate CEOs, increasing socioeconomic inequality, and a far more volatile economy periodically besieged by restructuring and financial crisis – all validated by the discourse of shareholder maximization and an increasingly predatory and short-term banking culture.

While anthropology has made important strides in studying both finance (especially from sites of expertise) and financialization (especially from 'below'), we have not been as effective in analyzing the concrete linkages between them, between the *mechanisms* of finance and the *effects* of financialization. Moreover, given that we have historically studied both capitalism and financialization from the vantage point of the less powerful, *and* given that recent ethnographic studies of finance have focused on financial products, networks, structures, knowledges, and ethics *without* necessarily examining how they *financialize*, anthropology may be running the risk of reproducing the point of view and epistemology of its more powerful subjects, as well as their idea as to what constitutes the proper subjects and objects of finance.

Central Assumptions and Challenges

To better understand and assess the challenges faced by the anthropology of finance and financialization and the potential directions it might take, an examination of our contemporary socioeconomic landscape as well as the dominant academic assumptions and analyses of these transformations would now be instructive. In the larger public imaginary, not to mention in the assumptions of critical scholars, finance is often simultaneously and negatively framed as separate from or antithetical to the 'real' economy, to production and productivity, while in the same breath it is automatically associated with inequality, 'fat cats,' and speculation. As such, many social analysts and critics, especially in the wake of the financial practices that helped to engender the Great Recession of 2008, have followed dominant interpretative paths for apprehending and critiquing finance, which are certainly understandable, yet also potentially misleading.

For example, mainstream assumptions associate finance with abstraction (i.e., distance and disconnection from 'the real' economy) and inequality, theorizing that the former produces the latter. Multiple assumptions are at work here that are crucial to unpack. First, the simplistic association of finance with abstraction discourages analysis of the workings and cultural processes of finance in that it further naturalizes the notion that finance is removed from everyday life and difficult to understand. Such an approach not only lets finance off the hook, it also falls prey to the rule of expertise whereupon the architects of unsustainable financial leverage are precisely those charged with diagnosing, explaining, and fixing its failures – because they are 'the smartest guys in the room.' Furthermore, abstraction also conveys and perpetuates the notion that financial crisis is the regrettable but unavoidable byproduct of the complexity of finance itself (La Berge, 2014). These rhetorics and tropes, again, cede authority to the experts themselves.

Abstraction and its twin complexity are also implicated as the central causes of the growing inequality in our social economy. In other words, the taken-for-granted understanding is that finance leads to inequality because its actions, focused as they are on short-term profiteering and making money out of money, have been divorced from 'real' production, the 'true' engine of creating jobs and manufacturing wealth. It was this sentiment, coupled with widespread public discontent, that fueled the ubiquitous and effective 'vulture capitalism' critiques of 2012 Republican presidential nominee Mitt Romney, in which he was accused of stripping and dismembering corporations and jobs for financial profit. Such an explanatory narrative of increasing abstraction as causing inequality is certainly understandable given the ever-widening wealth gap, and the inaccessibility and seeming remoteness of finance in the public imagination. Yet, if the culprit of growing inequality is abstraction itself – not the actions, practices, structures, and ideologies of financial actors and institutions – then it is not surprising that nebulousness characterizes our thinking about, critiques of, and general approach to finance.

Critical Approaches to Finance

This thread of addressing abstraction is an instructive one to think through in relation to anthropological and broader social scientific accounts of finance. While critical approaches to finance have overall been quite diverse and multidisciplinary, to the extent that certain orthodoxies have been reproduced in this scholarship, they cohere around the redichotomizing of the domains of economy and society as well as around perpetuating the taken-for-granted 'characteristics' of finance as abstract and corrosive in contradistinction to more 'authentic' and moral forms of economy and social relations. Because the practices, relationships, and values that are presumed to belong to the constructed domains of 'economy' and 'society,' not to mention the very manner in which these categories are thought to relate to each other, are so deeply engrained in US and western cosmologies (and presumed to be self-evident), critical scholarship is not exempt from these normative assumptions.

As such, it is not surprising that at one time or another, much of existing social scientific scholarship on finance reproduces a certain economism, or formalistic understanding of money and finance. For example, groundbreaking subfields such as economic sociology – which have contributed immensely to broader social scientific debates and understandings about the social construction of markets, the role of institutions and social networks, as well as the meanings and moralities of market forms and products – can, at times, fall prey to the reproduction of artificial binaries between economy

and society through its use of the narrative of the market 'disembedding from society.' The problem with this embeddedness paradigm is that demonstrating that 'the market' is, in fact, 'embedded' in society inadvertently perpetuates the assumption that 'market' and 'society' are stable categories and separable domains in the first place. While this re-embedding argument made sense as an intervention against the narrow mythologies of neoclassical economics, it is not a "great organizing principle for economy sociology" (Krippner and Alvarez, 2007).

Similarly, approaches from cultural economy to political economy, while demonstrating the mutual constitution of 'economy' and 'culture,' not to mention the structures of power that imbue capitalist social relations, can also reproduce economy/society binaries by framing 'finance' as a corrosive force against society. In this framing, finance is not only external to society, but also positioned as the immoral antithesis of a moral society; the solution is thus to re-embed and regulate finance. At issue is not the fact that the dominance of finance is leading to increased inequality (that is certainly the case); rather, it is that in highlighting the effects of inequality, we are assuming preformed 'market logics' for culturally specific and historically contingent financial values, priorities, and practices. In other words, the linear narrative of finance as a distant, autonomous, 'bottom line' economic force that colonizes society and catalyzes inequality sets the stage for the assumptions of corrosiveness and immorality. And, while the notion that finance is morally suspect and destructive is certainly an important contemporary sentiment, the analytical consequences of this frustration are that we absolve ourselves from investigating the mechanisms, categories, and assumptions of finance that are culturally and institutionally conditioned. Representing finance as disconnected and depraved depends on and reproduces a problematic narrative of authentic and inauthentic modes of capitalism, where finance is considered less authentic because it subverts normative productive capitalism due to its ability to self-reproduce from a distance without real productive labor or *longue durée* investments.

Yet another compelling approach to apprehending finance is what might be dubbed the social studies of finance, which itself derives loosely from interdisciplinary scholarship in science and technology studies and actor-network theories. This conceptual and theoretical trajectory has focused on the technology, networks, and architectures of markets – systems, models, objects, products, knowledges, and their relationships with each other and with the larger world. Such an approach has been crucial to the cultural study of markets and finance precisely because of its fine-grained analysis and tracking of the operations, formulas, and procedures that produce and structure finance (Knorr Cetina, 2007; Knorr Cetina and Preda, 2004; Lépinay, 2011). Its major contributions have included the training of attention on the interactions and interfaces between human and nonhuman actors that make financial transactions and knowledge production possible, as well as a focus on the interior architecture of finance almost akin to 'laboratory studies' that has helped to unravel the mechanics of the making of financial expertise. Its attention to the 'interiorities' of financial markets explores specifically how financial formulas and codes, interactions with prices, screens, and fellow analysts, and embodied understandings of numbers, volatilities, time, and space all participate in the construction of market architecture and the flows that permeate it. However, this literature's occasional framing of finance as a second-order market that is separate from actual firms and industrial production, or 'primary producer markets,' reproduces the very binaries that have been challenged earlier. Finance, it is important to emphasize, is intimately linked to production; it *does* depend on taking advantage of production, on converting a variety of assets into financial profits. Since finance is not so much separate from production as it is dependent on (and exploitative of) it, presuming disconnection or abstraction misses the point. Focusing too much on the interior workings of finance can make it difficult to simultaneously pay attention to the consequences of financialization, especially the production of socioeconomic inequality.

Perhaps the most influential contribution from this approach has been what is often termed 'performativity,' specifically the examination of how and to what extent formal economic models – the technologies and market devices described earlier – structure financial markets through the process of enacting and attempting to 'realize' these models (Fligstein and Dauter, 2007). This body of work has demonstrated how the implementation of reductive economic models and ideas (and the setting of these theories as goals to strive toward) has helped to construct socioeconomic realities that are framed by and akin to these representations (Callon, 1998; Callon et al., 2007; Carrier and Miller, 1998). Such work is crucial because it portrays the field of economics as a discipline involved in the active production of our world, not simply modeling it. Parallel to and building on the performativity of economics approach, sociologists of markets have not only demonstrated how mathematical and financial models help to 'make' markets (i.e., help to actualize some of what they model), but also shown that these models are mediated, influenced, and interpreted through organizational and institutional cultures that are also crucial to market-making (Beunza and David, 2008, 2012; Mackenzie, 2006; Mackenzie et al., 2008). Interestingly, this literature, while impactful in a multiplicity of ways, has also recently been critiqued for reproducing economic essentialism and a kind of inadvertent economism. Given its demonstration of the power of formal economic models, this subfield also runs the risk of including only certain sets of practices under the rubric of 'the economic' and overprivileging the capacity of certain market devices to remake the world.

In general, anthropological engagements with finance draw from and build on the many contributions of the aforementioned subfields. Given that finance continues to exercise an outsized influence over our social economy – its practices and sensibilities are utilized to govern and discipline a multitude of social domains and resources – a variety of cross-cutting and interdisciplinary research on finance is crucial. Paralleling these subfields, the anthropology of finance, forged by the larger discipline's history of engaging with capitalism and money, sheds light on 'the financial turn' by innovatively using anthropology's 'hallmark' methods of participant observation, of training and adoption, while it also reproduces some of our recurring analytical pitfalls

regarding the abstraction and economism of market logics (Maurer, 2012).

As the anthropologist Bill Maurer has concluded, using 'the broadest definition of finance,' the anthropology of finance 'would include all aspects of the management of money,' and in this general sense, finance has long been a core part of the anthropological tradition. The anthropologies of economy, capitalism, and colonialism, for example, have tackled the cultural construction of wealth and desires; shed light on multiple, alternative, and competing systems and values of exchange and reciprocity; interrogated the moral values, concepts, and cosmologies of western capitalism through the examination of exploitation at the margins; and demonstrated that imposed colonial systems of currency, credit, and debt, though influential and hierarchical, "were rarely more stable than those they sought to replace, and local systems were often durable and capable of sustaining long-distance and long-term financial relationships" (Maurer, 2012: p. 186).

In the anthropology of finance, then, the central tensions of challenging western universality through the figure and perspective of the other while simultaneously either eliding or homogenizing western capitalism are at work. For example, interrogating western assumptions of rationality and efficiency has been a central contribution of economic anthropology. Chris Hann and Keith Hart (2011) explain that in challenging the taken-for-grantedness of western economic categories and domains, economic anthropology importantly delved into alternative, culturally 'different,' and substantive ways of organizing and understanding affluence, household, reciprocity, and exchange. And yet, in their excavation of other ways of knowing, they also "removed world history from…[their] repertoire" (Hart, 2013: p. 1; Hann and Hart, 2011). In other words, in their attempts to focus on 'the other' and 'the alternative,' globalization, colonialism, and western capitalism were inadvertently bracketed off.

In a related development, some of the most trenchant critiques of western capitalism have come from anthropologists studying those living on the peripheries of global capitalism, where researchers have parlayed the cultural idioms and insights of 'the other' – as critical thinkers in their own right – to speak back to western capitalist culture (see, for example, Taussig, 1980; an exemplar of such work). Of course, such an approach was undoubtedly catalyzed by and intended to address the removal of world historical events from the purview of economic anthropology. And yet, a similar pitfall could be found here: although western capitalism was named and critiqued, it was not ethnographically explored nor rendered specific. These ensuing critiques held western capitalism itself to be a stable, homogenous, global penetrating logic that attempted to reframe the world in its image; 'the rest' of the world is thus charged with resisting the totalizing onslaught through 'local' cultural forms, values, and traditions.

Similarly, the anthropological scholarship on capitalism and neoliberalism has participated in the (re)production of universalizing discourses (such as the narrative of increasing abstraction), while it has also recognized and critiqued its own role in rendering neoliberalism overly totalizing, homogenizing, or top–down (see Ferguson, 2010; Gibson-Graham, 1996; Shever, 2012 for important interventions). Especially from the standpoints of the marginalized, critical studies of capitalism have been incisive and groundbreaking in their attention to both 'big picture' market developments and structures, and concrete strategies of negotiation in the face of growing inequality (Comaroff and Comaroff, 2000) and yet, in evoking this new world order, finance and financialization often get reified or homogenously evoked. Of course, given that the fine-grained study of financial worlds demands attention to the details of financial practice, ethnographic and cultural studies of finance and financial markets across multiple disciplines also serve as a crucial intervention against their taken-for-grantedness or presumed totalization.

Ethnographic Examples

Four relatively recent anthropological monographs of finance illustrate some of the insights, interventions, and perspectives that long-term ethnographic research on finance can offer. A brief tour of Caitlin Zaloom's *Out of the Pits: Traders and Technology from Chicago to London*, Hirokazu Miyazaki's *Arbitraging Japan: Dreams of Capitalism at the End of Finance*, Karen Ho's *Liquidated: An Ethnography of Wall Street*, and Alexandra Ourousoff's *Wall Street at War: The Secret Struggle for the Global Economy* will serve to demonstrate the range of current ethnographic work and the particular insights that can derive from such labors.

Caitlin Zaloom (2006) accesses traders' culturally and historically specific embodiments and approaches to financial futures markets as trading moved from the open-outcry pits to the screen, from Chicago to London, and from divergent understandings of who and what constituted the proper economic subject to represent and engage in the new global financial markets. In particular, she finds that an ethnographic approach that includes traders' cultivation of a 'feel' for the rhythms of the market, along with their techniques of speculation and understandings of networks and competitors, affords more crucial insights into their daily market practices than simply an interrogation of their knowledge of the financial products themselves. In fact, it is through Zaloom's fine-grained research among traders that she then recognizes why dominant critiques of 'greed' and 'client responsibility' are oftentimes dismissed or ignored by these market actors. Given that the traders conceive their own central virtue or 'ethics' to derive from their ability to 'be' and to 'make' markets through aggressively facilitating and exploiting buyers and sellers, it comes to no surprise that they would register the act of habitually taking positions opposed to the interests of their own clients as part of their market-making responsibility, not simply being 'greedy' (Zaloom, 2012). To the extent that the anthropology of finance can better illuminate the motivations and specificities of such value registers, perhaps additional and more effective interventions can be made.

Hirokazu Miyazaki's ethnography (2013) delves into the ethics, knowledge practices, and dreams of Japanese traders in Tokyo as a window into the deep interiorities of the mindset and dreams of professional market actors. One of the ethnography's key contributions and methodologies is

that instead of framing the anthropologist as analyzing subject and approaching these actors as objects of critique, Miyazaki's work approaches and represents derivative traders as "thinking subjects capable of reflecting on and reorienting the place of their expertise in economy and society" (p. 23). In other words, financial professionals learn, observe, and theorize their worlds, techniques, and subjectivities in such a way as to recognize their limitations, failures, hopes, and dislocations. As such, especially in the context of the ruptures and uneven fortunes of some in finance (such as the Japanese arbitrageurs that Miyazaki studies in contradistinction to more privileged New York traders), it is crucial not only to understand how their 'ends' and 'means' have radically shifted, but also to grasp why calls to 'reform' and regulate finance need to take into consideration the multiplicity of experiences and goals that populate diverse financial worlds.

In Karen Ho's *Liquidated: An Ethnography of Wall Street* (2009), she accesses and examines the cosmologies of Wall Street investment bankers. The purpose of the study is to access their worldviews and analyze the processes through which such views are forged, as well as to understand how Wall Street values came to be imposed onto and emulated by other industries, thus setting the stage for the financialization of corporate America at large. According to Ho, the ethical outlook and professional practices of investment bankers are often obscured or rendered invisible by taken-for-granted 'marketspeak.' In other words, although 'financial culture' is constructed through institutional contexts, university pedigrees, and relationships with other financial institutions and corporate America, it is often mistaken for 'the market' itself, a presumption which in turn automatically rehearses neoclassical assumptions rather than revealing the cultural contingencies that underpin Wall Street. As such, one of the key contributions of *Liquidated* is to demonstrate that 'the market,' and other economic models and logics such as 'shareholder value,' cannot be taken at face value: the set of actions that is commonly held to result in rising stock prices is not predetermined by 'objective,' top–down formulas, but is instead constructed through the particular viewpoints and experiences of Wall Street bankers.

For example, it is precisely investment bankers' own experiences of continual restructuring, downsizing, and turnover, their short-term incentivization through year-end bonuses, and their understanding of front-office, 'revenue-generating' work as the only valuable labor (vs 'costly' back-office work that can easily be outsourced) that shape their ideas of how to create shareholder value. As a demonstration of the phenomena of financialization, *Liquidated* shows how Wall Street's local cultural models of business advice and practices have become so influential over the past 30 years. Other industries and sectors have had to take Wall Street's particularities so seriously that they have now become normative. The cultivation of a hierarchical superiority buttressed by a discourse of smartness, the revival of a particular ideology of shareholder primacy, the corporate takeover movement, and the massive transfer of wealth and assets into financial markets elevated Wall Street's role and influence. It was in this context that Wall Street financial institutions leveraged the supposed ability of bankers to act as market avatars to trigger the financialization of corporate America. Such ethnographic work unpacks and problematizes finance and financial markets by laying bare their cultural and historically contingent building blocks.

In *Wall Street at War*, Alexandra Ouroussoff (2010) further bridges the gap between finance and industry, between Wall Street and corporate America, in an ethnography that focuses on the tensions between the two models of economy and entrepreneurship. Through fieldwork with executives in the corporate and industrial sector on the one hand, and with financial analysts in the investment and financial sector, especially credit rating agencies, on the other, Ourousoff documents how the two models for conducting business have become paradoxically linked. In her analysis, the movement to discipline and measure corporate America according to the rules and norms of finance has helped to undermine the contemporary economy. Specifically, Ourousoff demonstrates that when Wall Street financial institutions and credit ratings agencies strive for investment predictability, the norm they enforce privileges the certainty of returns – which *prevents* corporations from being able to engage in the entrepreneurship and risk taking necessary to create new products and business opportunities. In other words, normal business and entrepreneurial activities that produce uncertainty are ironically devalued by Wall Street as threatening losses, and are thus seen as deserving of poor credit ratings. Ourousoff concludes that financialization has made it more likely for corporate America to dissemble and subvert its own goals in order to abide by (or resist) the authority of financial analysts.

These ethnographies, and the anthropology of finance, more generally, intersect with the broad set of diverse concerns and questions that informs economic anthropology and the anthropologies of capitalism, money, class, neoliberalism, work, corporations, technology, knowledge, feminism, power and inequality, gender, race, and ethnicity. Moreover, the anthropology of finance also draws from a vast interdisciplinary literature, which includes economic sociology and the sociology of markets, the social studies of finance, and critical management studies. While this compelling subfield has substantively contributed to understanding the 'nuts and bolts' of finance – especially given the methodological (and theoretical) necessity of approaching finance as a local site – the dangers of analytical abstraction still loom large.

Blaming financial abstraction for the ills of contemporary capitalist inequality and change depends on presuming the context and directionality of our social economy, which, I would argue, is contrary to the contributions and strengths of ethnographic specificity. Such arguments, of course, stem in part from longstanding interdisciplinary understandings of capitalism as governed by inevitable, evolutionary, and totalizing logics, demands, and laws – an approach that anthropologists and many social scientists have challenged as well as reproduced. Assuming that finance exhibits these characteristics not only prevents us from examining the particularities and contingencies of finance on the ground, but also reproduces formalistic and dominant ideas about money, markets, society, and morality.

See also: Brokers and Brokerage, Anthropology of; Business Anthropology; Economic Anthropology; Elites, Anthropology of; Exchange in Anthropology; Expertise, Anthropological Approaches to; Globalization, Anthropology of; Markets: Anthropological Aspects; Money: Anthropological Aspects; Political Economy in Anthropology; Postindustrial Societies; Work, Anthropology of.

Bibliography

Beunza, D., David, S., 2008. Tools of the trade: the socio-technology of arbitrage in a wall street trading room. In: Pinch, T., Swedberg, R. (Eds.), Living in a Material World: Economic Sociology Meets Science and Technology Studies. MIT Press, Cambridge, pp. 253–290.

Beunza, D., David, S., 2012. From dissonance to resonance: cognitive interdependence in quantitative finance. Economy and Society 41 (3), 383–417.

Callon, M. (Ed.), 1998. The Laws of the Market. Blackwell Publishers, Oxford.

Callon, M., Millo, Y., Muniesa, F. (Eds.), 2007. Market Devices. Blackwell Publishers, Oxford.

Comaroff, J., Comaroff, J., 2000. Millennial capitalism: first thoughts on the second coming. Public Culture 12 (2), 291–343.

Carrier, J., Miller, F. (Eds.), 1998. Virtualism: A New Political Economy. Berg, Oxford.

Ferguson, J., 2010. The uses of neoliberalism. Antipode 41 (1), 166–184.

Fligstein, N., Dauter, L., 2007. The sociology of markets. Annual Review of Sociology 33, 105–128.

Gibson-Graham, J.K., 1996. The End of Capitalism (As We Knew It): A Feminist Critique of Political Economy. Blackwell Publishers, Oxford.

Hann, C., Hart, K., 2011. Economic Anthropology. Polity Press, Cambridge.

Hart, K., 2013. The Anthropology of Money and Finance: From Ethnography to World History. http://thememorybank.co.uk/2013/09/04/money-and-finance-from-ethnography-to-world-history/ (accessed 12.01.13.).

Ho, K., 2009. Liquidated: An Ethnography of Wall Street. Duke University Press, Durham.

Hyman, L., 2012. Debtor Nation: The History of America in Red Ink. Princeton University Press, Princeton.

Knorr Cetina, K., 2007. Economic sociology and the sociology of finance. Four distinctions, two developments, one field? Economic Sociology: The European Electronic Newsletter 8, 4–10.

Knorr Cetina, K., Preda, A. (Eds.), 2004. The Sociology of Financial Markets. Oxford University Press, Oxford.

Krippner, G., 2012. Capitalizing on Crisis: The Political Origins of the Rise of Finance. Harvard University Press, Cambridge.

Krippner, G., Alvarez, A., 2007. Embeddedness and the intellectual projects of economic sociology. Annual Review of Sociology 33, 219–240.

La Berge, C., 2014. The rules of abstraction: methods and discourses of finance. Radical History Review 118, 93–112.

Lépinay, V., 2011. Codes of Finance: Engineering Derivatives in a Global Bank. Princeton University Press, Princeton.

Mackenzie, D., 2006. An Engine, Not a Camera: How Financial Models Shape Markets. MIT Press, Cambridge.

Mackenzie, D., Muniesa, F., Siu, L. (Eds.), 2008. Do Economists Make Markets? On the Performativity of Economics. Princeton University Press, Princeton.

Martin, R., 2002. Financialization of Daily Life. Temple University Press, Philadelphia.

Maurer, B., 2012. Finance 2.0. In: Carrier, J. (Ed.), A Handbook of Economic Anthropology, second ed. Edward Elgar, Cheltenham, UK, pp. 183–201.

Miyazaki, H., 2013. Arbitraging Japan: Dreams of Capitalism at the End of Finance. Stanford University Press, Stanford.

Ott, J., 2011. When Wall Street Met Main Street: The Quest for an Investors' Democracy. Harvard University Press, Cambridge.

Ouroussoff, A., 2010. Wall Street at War: The Secret Struggle for the Global Economy. Polity Press, Cambridge, UK.

Shever, E., 2012. Resources for Reform: Oil and Neoliberalism in Argentina. Stanford University Press, Stanford.

Taussig, M., 1980. The Devil and Commodity Fetishism in South America. University of North Carolina Press, Chapel Hill.

van der Zwan, N., 2013. Of Unintended Consequence: The History of Finance Capitalism in the United States. H-Business, H-Net Reviews: http://h-net.org/reviews/showrev.php?id=37978 (accessed 05.05.14.).

Zaloom, C., 2006. Out of the Pits: Traders and Technology from Chicago to London. University of Chicago Press, Chicago.

Zaloom, C., 2012. The Ethics of Wall Street. Cultural Anthropology. Online: http://www.culanth.org/fieldsights/363-the-ethics-of-wall-street (accessed 15.01.14.).

Finance, Geography of

Gordon L Clark, Smith School of Enterprise and the Environment, University of Oxford, Oxford, UK; and Monash University, Caulfield, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by N. Thrift, volume 23, pp. 5655–5657, © 2001, Elsevier Ltd.

Abstract

Money has been with us since time immemorial. It serves many functions, underwriting exchange and transactions as well as the spatial and temporal integration of human activity. It comes in many forms and is more than a component part of market capitalism. Its modern form is represented by finance, the financial services industry, and what many commentators refer to as the financialization of economy and society. For many writers, modern finance is distinct and autonomous from industries that produce and trade in commodities. But there is little doubt that twenty-first-century globalization would not have been possible without the global financial services industry. Research on the topic is to be found across the social sciences, including accounting, economics, human geography, political science, and sociology. In this article, the geographical building blocks of the global financial services industry are identified, emphasizing the spatial structure of gross flows as well as the role and significance of financial centers. Understanding the geography of finance also requires an appreciation of the various scales, at which finance appears in everyday life. So, for example, the article also focuses upon the relevant literature in geography on financial literacy, financial exclusion, and financial crisis. The article opens and closes with reference to the global financial crisis and, most importantly, the Euro crisis. The Euro crisis has been accompanied by a landscape of despair that is likely to profoundly mark the prospects of whole regions and countries across Europe over the coming decades.

The geography of finance has been a topic of interest for centuries. Shakespeare's plays are replete with commentary on money changers and disreputable posers; some of the most tragic figures in English literature are caught up in financial leverage and obligation. The Globe Theater was just across the River Thames from the public houses and betting shops that were to form the basis of the City of London as the world's most important financial center in the eighteenth century. It remains a vital cog in the global economy, being the switch point for financial assets flowing east to west and north to south on a 24-h, 365-day basis (Clark and Thrift, 2005). Financial centers are more than trading posts. They fuel the real economy, facilitate the trade of commodities and services, and sustain economic growth and development (Wójcik, 2011). If the alchemists of folklore like Merlin the Magician could turn lead into gold, it is all too obvious that the modern alchemists of finance brought the global economy to its knees (Engelen et al., 2012).

It is arguable that money, its functions, and its various physical manifestations have been with us since time immemorial. On the other hand, its institutional structure and significance for society and economy vary over time and space. Some analysts make gross distinctions, comparing the role of money in different regimes of accumulation over enormous stretches of time (feudalism vs. capitalism, etc.). While sympathetic to this logic, too often it imposes a blanket over the subtleties and continuities apparent within and between historical and geographical formations (see Landreth, 2012). So, for instance, it has been argued that the rise of modern financial capitalism over the second half of the twentieth century was accompanied by a profound schism between corporate capitalism and the new institutions that have mobilized financial assets, setting these interests against inherited industry and national formations (Clark, 2000; Walker, 1989). The geography of finance opens up the problematics of finance to scrutiny – taking aim at convention while suggesting that finance can be seen in fresh ways otherwise lost in totalizing conceptions (see Christophers, 2012; Coval and Moskowitz, 2001; compare Merton and Bodie, 2005).

The intellectual foundations of modern financial theory were formalized over the past 50 years: notably Markowitz (1952) on modern portfolio theory, Fama (1970) on the efficient markets hypothesis, and Merton (1990) on modeling risk over time. Those following in their wake developed frameworks for conceptualizing and measuring risk, closing the gap between financial theory and practice with a 'universal' formula for strategy and behavior (as evident in the significance attributed to the Black–Scholes theorem; see MacKenzie, 2006). By the turn of the twenty-first century, it was possible to talk about 'financialization' as a new form of capitalism albeit centered in Anglo-American markets. Whereas it was convenient to suppose that inherited 'varieties of capitalism,' with attendant differences in state–society–economy relations (Hall and Soskice, 2001) would reasonably coexist with one another, financialization was shown to be a threat to persistent difference (Clark and Wójcik, 2007). The euro crisis can be seen as a political response to the power of global financial markets – a test of will between political elites and market sentiment.

In this account of the core themes and issues driving the research program on the geography of finance, we provide a map of finance from the global to the local, from gross flows through to individual decision-making in context. The survey focuses on key elements in the global financial system while providing spatial references for thinking about the forms and functions of finance (Dixon, 2012). By way of conclusion, the survey returns to the euro crisis and its distinctive geography. The role and status of finance in modern societies are highly contested. Nonetheless, understanding the organizing logic of

finance and financial markets has never been so important for the welfare of coming generations.

Global Flows

One way of accounting for the global flows of finance is to map onto commodity trade and exchange; in effect, mapping onto geographical patterns of national and international integration. It is widely observed that trade and exchange among Western developed economies is accompanied by massive financial flows just as low volumes of trade and exchange between developed economies and less developed economies is accompanied by a paucity of financial intermediation and integration (Mishkin, 2006). Even so, periodic resource shortages have tended to enrich the resource-dependent margins of the developed world. The flows of finance accompanying these surges in resource incomes are, however, spatially symmetrical: as the West pays for resources, these payments flow back to Western financial markets for investment.

Global financial integration over the twentieth century can be mapped onto (1) trade and exchange between the UK, its residual empire, and North America, thereafter (2) incorporating Europe and Latin America into systems of exchange, (3) providing Japan and then the newly industrialized Asian economies with a means of financing development, and ultimately (4) placing China at the very center of the global financial system. Each phase of incorporation has benefited the newly industrializing economies, raising living standards and providing financial resources for development. When switching financial assets into the global economy, these assets have been passed through the handful of financial centers that dominate the global financial system. It is not surprising that international trade and development is accompanied by the deepening of domestic financial markets and the mobilization of financial assets through financial centers, national and global.

Whereas commentary on twenty-first-century globalization tends to suggest that financialization has crept into every nook and cranny of the world, we should acknowledge that the global map of finance is highly differentiated. Some countries and some regions are awash with financial assets, while other countries and regions are financially impoverished. Whereas there are huge incentives for financial institutions to seek out opportunities at the margins of the global economy, 'home bias' is a common refrain (Coval and Moskowitz, 1999; Huberman, 2001).

Pricing Tradition

A remarkable feature of Western development since 1970 has been the enormous growth of the financial services industry. Even countries outside the Anglo-American system of financial markets and regulation have experienced high levels of accumulation in financial assets and burgeoning growth in employment in the financial services sector (Wójcik, 2013). Some countries like Germany and the Netherlands, as well as the countries of Scandinavia, are hosts to large financial institutions that have evident global footprints. The classification of national economic systems into 'varieties of capitalism' by reason of their reliance on banks as opposed to markets, corporatism as opposed to the market for corporate governance, and the premium on commodity production and export over financial services, is harder to sustain than is often acknowledged (Clark and Wójcik, 2007). There is considerable evidence that the management of many of Europe's larger companies is very sensitive indeed to the incentives and rewards available in global financial markets (see Bauer et al., 2008). Large corporations are more likely 'global' than 'local' (compare Doremus et al., 1998).

Western countries accumulated vast stores of financial assets through high savings rates whether workplace-based (as in funded pension and retirement schemes) or in individual saving schemes (insurance and otherwise). This is referred to, by some, as the fifth stage of capitalism or pension fund capitalism (Clark, 2000; Clark and Hebb, 2004). Most importantly, the nexus between the production and sale of commodities in local markets and the accumulation and management of financial assets was fractured such that financial institutions became autonomous from commodity-producing industries and sectors. Accompanying this development was the formation of a shared intellectual framework for the management of financial assets, which effectively legitimated the institutional and legal division of responsibilities between the production of savings and the investment of savings (Hawley and Williams, 2000). If a distinctly Anglo-American phenomenon (see La Porta et al., 1998), its boundaries have been systematically extended by the globalization of the financial services industry.

In the space of about 25 years, Western financial markets became drenched in 'surplus' financial assets managed by 'autonomous' financial institutions. The search for premium rates of return in financial markets encourages financial innovation and leverage, which were routinely rewarded (in the short term) even if transitory (in the long term). As markets became more efficient and as transaction costs declined, real rates of return above market performance became harder to find. Whereas opportunities were evident at the margins of the global economy, the transaction costs and the risks incurred by being at the margin were deemed too significant relative to the opportunities for arbitrage in developed financial markets. Indeed, through to the last decade of the twentieth century and the first years of the twenty-first century, it was widely believed that emerging markets lacked the transparency and depth of liquidity needed to sustain large-scale investment programs.

The fact that emerging markets were the source of global financial crises through much of the twentieth century reinforced an unjustified belief in the stability of developed economies' financial markets (Barro, 2006). It was inconceivable that Western financial markets could be 'hosts' to systemic patterns of behavior and institutional irrationalities that would produce a global financial crisis (Clark, 2011).

The Hierarchy of Financial Centers

Why is there a pronounced hierarchy of financial centers? This question can be answered with reference to history, geography, and economics. It is easy enough to explain the contemporary significance of London for Europe and the international financial system by invoking its historical significance as the

switch point for agricultural wealth-seeking opportunities through the Industrial Revolution (at home) and opportunities presented by the expansion of the British Empire (abroad). It is not just history that counts – explaining the resilience of London in relation to its competitors, past and present, is also important (Beaverstock and Hall, 2012). Likewise, it is easy enough to 'explain' London by the fact that it is at the center point in time and space between Asia and North America. If London did not exist, it would have to be invented. Amsterdam, Frankfurt, or Paris could provide the functions that London currently provides given their similar location in time and space. However, economics could also be invoked, referencing the virtues of London as a site for the production of financial services against the alternatives (Faulconbridge et al., 2007).

From a theoretical standpoint, there are three possible answers to the question posed above. Before doing so, it is important to acknowledge a set of economic factors that underpins these explanations. First, collecting, processing, and distributing financial assets on a global basis require enormous investment in the electronic infrastructure of information processing, so much so that just a handful of companies dominate the industry. Second, reinforcing the concentration of the industry is the fact that there are economies of scale in managing and processing financial assets such that just a few providers dominate national financial markets and act as the gateways or portals to the global financial system. Third, notwithstanding commonplace perceptions of a uniform global landscape of financial institutions and markets, there remain persistent differences between nation-states as regards the status attributed to domiciled banks, insurance companies, and related financial institutions. In many cases, nation-states have sought to protect and maintain the national importance of these institutions. They have an interest in mutual survival.

One explanation of the persistence of national and global financial centers has to do with access to information regarding the price of the traded securities, and the formation of expectations about future prices and the risks attending those expectations. At one level, information is virtually free and instantly accessible for market participants, whatever their locations in time and space. Information providers like Bloomberg and Thomson Reuters have transformed the trading desks of financial institutions, providing traders with a panoptic view of national and global financial markets, separately and together. By this account, it would seem that financial markets could easily fall into a single electronic platform floating free of its participants' jurisdictions. Attempts have been made to create such platforms, but ambitions have not been realized. In large part, this is because the electronic processing of publicly available information rarely provides market participants with an edge when seeking to anticipate shifts in market sentiment. In a world of shared information, accessible at a click of a mouse, there is an enormous premium on information gleaned from personal networks and unique vantage points in economic and financial decision-making (Hall and Appleyard, 2008).

A second explanation of the persistence of global financial centers has to do with the importance of market liquidity for investment groups and traders. One reason investors located in national financial centers converge on the relevant macroregional global financial center (such as Tokyo, London, and New York) is that price discovery in national financial centers is often hampered by the limited number of traders and clients. As a consequence, local liquidity may be shallow when compared to the depth of liquidity available in the adjoining global financial center. At the same time, traders located in other national financial centers but overlapping with traders in adjoining financial centers can, in effect, trade through the global center to take advantage of buy-and-sell orders in related stocks across the macroregion. In this manner, liquidity is pooled virtually if not always predictably. What is remarkable about the London market is the diversity of traders by region-of-origin and the buy-and-sell orders they bring with them reflecting very different time horizons, expectations, and desired payoffs. A well-functioning financial center combines liquidity with a diversity of expectations.

A third explanation of the persistence of global financial centers relies upon a quite conventional idea, found in textbooks on economic geography. This explanation rests on a set of basic premises or suppositions: (1) the more specialized a product or service, the larger the market has to be to support provision of that product or service; (2) specialized products and services are, inevitably, linked together in chains of complementary products and services such that provision of one or a small set of these products and services requires the availability of other related products and services; and (3) national and regional financial centers may provide these types of products and services, but are dominated by large financial institutions. In effect, market density is so shallow that there is only space for one provider, thereby limiting the effect of market competition on the nature, cost, and value of these products and services. Global financial centers have both the requisite range of products and services, and the requisite density of skilled market participants consistent with competition for the provision of those products and services (Faulconbridge, 2008).

By this account, London's place in the global financial system is underwritten by the infrastructure supporting the depth and scope of market information, liquidity, and the diversity of expectations. Reinforcing these advantages is the density of London-based intermediaries competing for mandates. Not surprisingly, financial institutions, not domiciled in these types of centers often leapfrog regional and national financial centers to gain access to the most important global markets for financial securities, products, and services. Leapfrogging has been made a more effective strategy over the past 20 years by the development of formal information channels, which have facilitated access to markets on a virtually instantaneous basis, if not always in the real-time flux and flow of changing expectations. In this context, there is also an intimate and thoroughly reinforcing relationship between London and off-shore financial centers (Wainwright, 2011).

Regional Growth and Development

There is, quite obviously, extraordinary interest in the structure and performance of financial markets, not least their coordination over time and space. However, care should be taken not to ignore other salient issues, including the significance of

financial innovation for economic growth and development. In a series of papers, King and Levine (1993a, 1993b) showed that developing countries' economic growth is, in the first instance, dependent upon the robustness of systems of exchange and transaction, including the availability of money that serves as a recognized and reliable store of value. They also show that economic development is accompanied by increasing financial sophistication when trade and exchange become a viable mechanism for accumulating surplus value. They do suggest, however, that financial sophistication can, at some point in the development pathway, outstrip the demand for financial products and services. Just as some countries may be hamstrung by the Dutch disease (resource wealth), others that specialize in financial products and services may find that other sectors are crowded out in the national economy.

Importantly, financial intermediation has played a major role in facilitating the growth and development of subnational regions of innovation such as Silicon Valley and Rt128/495 Boston, the Oxford–London–Cambridge complex, and the Third Italy (Babcock-Lumish, 2005). It would be misleading to suggest that these regions of innovation have relied upon the same types of financial institutions to fuel new start-ups, the investment in ideas, and the realization of value. Further, regions of innovation have faltered or failed to take off when they are unable to access investors knowledgeable about the innovation process and the different types of financing needed to transform an idea into a marketable product (Backher et al., 2014). Some regions rely upon dispersed and uncoordinated small investors, whereas other regions rely upon venture capital firms that are large enough to mobilize investors across the world through portfolios of investment opportunities. Other regions rely upon banking institutions and their lending practices, sustained by long-term relationships governed by trust rather than contract. In still other cases, intermediation is highly formalized and often remote from local conditions.

At the other end of the spectrum, large multinational companies rely upon a combination of internal financial resources and external investors to sustain their businesses. While notionally domiciled in a specific country, with related firms managed and reporting through to the global headquarters of these companies, the evidence suggests that these types of companies have become quite responsive to global investors (Bauer et al., 2008). In many cases, these types of companies carry many hundreds of investors from across the world. Few of these investors have the size or the significance necessitating 'active management'; in fact, these types of companies deliberately dampen the power of shareholders by cross-listing in other jurisdictions so as to attract investors who have neither the interest nor the capacity to organize themselves into voting coalitions. One or two steps removed from their host countries' governance procedures and rules and regulations, these types of companies have used the relative anonymity of global financial markets to shield themselves from intervention.

Nonetheless, financial institutions like pension funds and investment managers have become more active in holding large corporations to account, even if governments of their home jurisdictions are reluctant to affect the powers of their national champions. In some cases, these financial institutions have sought to 'name and shame' corporate executives, thereby demonstrating that reputation must be earned, not assumed. This occurs when institutions mount campaigns at annual general meetings to influence corporate governance policies as well as global and local environmental and social standards. Out of the limelight, financial institutions have also sought to engage with senior executives to affect corporate management on issues such as reputation, license to operate, and long-term sustainability. For some, such issues are rather ephemeral, given the map of corporate power. However, these issues are increasingly on the agenda for governments of advanced economies seeking to ensure the long-term success of their leading corporations.

Mapping Financial Literacy

By this account, the map of global finance is a map of financial stocks and flows mediated by institutions and markets. In play is the structure and performance of financial institutions and related markets – if we have the power to redesign and regulate the system. At another level, however, financial institutions and markets provide an action space for individuals and communities seeking to realize their own aspirations in terms of personal and collective welfare. Put slightly differently, financial institutions and markets can be treated as the 'environment' in which decisions are taken and plans are made about the future. Obviously, this environment can be characterized in various ways, including whether it is benign, poor in resources like knowledge and information, or rich in these resources (Leyshon and French, 2008). Since financial markets are subject to a risk and uncertainty, this may harm those who must rely upon institutions and markets for their own prospects and the prospects of others.

It is arguable that these issues have become more important over the past couple of decades than hitherto acknowledged (Leyshon and Thrift, 1997). For some, the increasing need for individuals to plan for the future in the context of financial risk and uncertainty reflects the hegemony of neoliberalism (the discounting of nation-state capacity, the increasing role of markets in pricing inherited traditions, and the advent of globalization; see Langley, 2008). Whether neoliberalism is a systematic political project or the artifact of juxtaposed and overlapping forces is subject to debate. It is apparent, however, that the past certainties associated with Anglo-American liberal capitalism, welfare states, and European social democracies have been systematically discounted over the past 25 years. This is especially obvious in areas of social policy such as pension and retirement benefits, the provision of health care, and the provision and pricing of higher education. It would be misleading to suggest that these trends are everywhere the same; European pension reform takes many shapes and sizes. None represent an accepted model to which nation-states are converging (Clark, 2003).

By convention, it is assumed that consumers bear the risks of their decisions. However, as the nature and scope of the decisions to be made have changed and the long-term consequences of making poor decisions have become apparent, the doctrine of *caveat emptor* has been subject to searching criticism and analysis. At one level, it is apparent that the market for financial products and services is not nearly as competitive as

theorists suppose, leading, in some instances, to the systematic discounting of individual and collective welfare. Here, the vulnerability of consumers of financial products can be explained by persistent information asymmetries, which shelter providers deploying the so-called sharp practices. The vulnerability of consumers can also be explained by the complex nature of financial products, which nonetheless seem to promise an entirely obvious result: a liveable retirement income. The evidence suggests that certain segments of the population, particularly in certain regions and city neighborhoods, are more vulnerable than others to 'mis-selling' (Leyshon et al., 2004, 2006).

The average consumer is bound to struggle when attempting to understand the possible consequences associated with certain types of financial products. For some, the issue is simply one of a lack of knowledge and understanding of finance (in general) and the products and services (in particular) that require domain-specific judgment and experience. For others, more confident in their ability and the scope of their experience, there remain pitfalls for the unwary. It was not uncommon through much of the 1990s for the professional classes of southern England to sign up to endowment mortgages, invest in with-profit insurance contracts, and take on interest-only mortgages for buy-to-let housing in some of the less fashionable districts of London. Each type of product carries with it considerable risks and uncertainties. Together these products represent a profound threat to the solvency of households that unknowingly carry highly correlated downside financial risks. Not surprisingly, both the global financial crisis and the euro crisis revealed that many households had assumed a level of financial leverage antithetical to their long-term interests.

One response to the costs associated with poor individual financial decision-making has been to improve the quality and quantity of information that vendors must make available when offering financial products for sale. Governments have established all kinds of information portals, as well as internet-based decision cues and decision trees that may prompt better, if not entirely optimal, decision-making. Public uptake of these advisory services has been limited and largely ignored by those who could benefit the most from their use. Governments have also sought to better understand the nature and scope of financial literacy in the community, paying particular attention to systematic trends by sociodemographic characteristics and geographical reference points. It has been shown that financial literacy is systematically distributed throughout the community with troubling implications for long-term income inequality and well-being. Here, again, government programs have been launched to 'improve' financial literacy especially among secondary school students who, it is assumed, would carry this training with them throughout their lives.

Most importantly, it is apparent that there are maps of financial literacy within countries and between their regions, just as there are significant differences between countries in their levels of financial literacy. Accounting for these maps of financial literacy is of considerable importance to academics, policymakers, and even multilateral organizations (Clark, 2013, 2014a). However, it has been recognized that the testing regimes upon which surveys of financial literacy are based often contain questions on financial concepts that, for many people, may appear rather abstract and lacking in saliency. Equally troubling, it is not always obvious that the supposed 'correct' answers to questions that seek to represent the foundations of financial literacy are unambiguous or even plausible to people in specific situations (time and space). That is, the testing regime underpinning the financial literacy project may not be as relevant as assumed nor indicative of the decision competencies of people given their particular circumstances.

Future Research

Much of the literature on the geography of finance is about institutions and markets and the tensions between local, national, and global imperatives. As such, this literature overlaps with and intersects with related work in political science on 'varieties of capitalism.' Recall that this school of thought has focused on the formal or institutional differences between countries, even blocks of countries, so as to explain how and why the imperatives of global finance may not prompt convergence upon one model of capitalism (Hall and Soskice, 2001). One criticism of this approach is that, for 'difference' to carry sufficient weight so as to counter the imperatives of global finance, nation-states are treated as organic wholes impervious to the attractions of financial rewards internal and external to these economies. Empirical evidence suggests that national models may fracture internally as those that benefit from financial markets bypass local custom and convention (Clark and Wójcik, 2007).

Related criticism, launched by Dixon (2012), is that explanations of persistent differences between countries in the face of global financial imperatives rely upon differences in the formal and informal institutions of nation-states. In effect, form is assumed determinant of function in that the latter can be read off the former. However, this type of analytical logic often misses the fact that institutional form is more malleable than many believe, being responsive to changes in the external environment on an incremental basis if not always on a structural basis. More often than not, those that hold fast to the difference thesis must look back in history for distinctive differences between nation-states and their associated regimes of accumulation. Part and parcel of this strategy of idealization is the discounting of current developments and apparent changes in institutional performance. This is not to say that global finance is hegemonic. But it does suggest that economic and political institutions may be forced to accommodate the imperatives of global finance (Dixon, 2014).

This issue is being played out in Europe via the euro crisis. Simply stated, the euro crisis pits the commitment of Eurozone members to an idealized conception of the European community against the expectations of global financial institutions. However imperfect the Eurozone is in terms of its structure, composition, and governance, Germany has led a coalition of the faithful in an attempt to stare down bond traders who believe that the Eurozone is not functionally viable in its current form. In part, this has involved holding to ransom countries such as Greece, Spain, and Portugal in the Eurozone as a calculated gesture of defiance, which implicitly signals that Europe is willing to bear the price of solidarity. Yet, the price of solidarity is unevenly distributed: throughout Europe, many young people face high and increasing rates of unemployment

and the prospect of long-term impoverishment. Throughout Europe, low or negative rates of economic growth are accompanied by decreasing real incomes with long-term implications for those who rely upon waged work for their well-being. Furthermore, there seems no way out for the residents of Greece, Spain, and Portugal (and perhaps southern Italy). These countries are experiencing economic depressions at least as significant as the Great Depressions of the 1890s and 1930s. Political extremism is on the horizon.

Is the price of maintaining this idealized community worth paying? Given the obvious and quite profound shortcomings in the design and construction of the Eurozone, is the price of European solidarity legitimate? At present, these questions are answered in the affirmative – at least by the political leaders of the dominant Eurozone countries. However, instead of seeing the issue through the eyes of Europe's political leaders, a worthy project that could join many social science disciplines, especially human geography with economic geography, would be to study the emerging map of European unemployment, poverty, and social crisis. As indicated above, this map is likely to spread across Europe and to be increasingly differentiated in terms of those regions that demonstrate resilience in the face of this disaster and those regions – even whole countries – that simply fail under the weight of history (Clark, 2014b).

See also: Cities and the Financial Crisis; Corporate Finance: Financial Control; Economic Geography; Financial Crises, History of; Financial Regulation; Networks, Geography of; Service Economy, Geography of; Urban System in Geography.

Bibliography

Babcock-Lumish, T., 2005. Venture capital decision-making and the cultures of risk: an application of Q methodology to US and UK innovation clusters. Competition and Change 9, 329–356.

Backher, J., Clark, G.L., Monk, A.H.B, Sridhar, K, 2014. The 'valley of opportunity': rethinking venture capital for long-term institutional investors. Alternative Analyst Investment Review 3 (1), 63–75.

Barro, R.J., 2006. Rare disasters and asset markets in the twentieth century. Quarterly Journal of Economics 121, 823–866.

Bauer, R., Braun, R., Clark, G.L., 2008. The emerging market for European corporate governance: the relationship between governance and capital expenditures, 1997–2005. Journal of Economic Geography 8 (4), 441–469.

Beaverstock, J.V., Hall, S., 2012. Competing for talent: global mobility, immigration, and the city of London's labour market. Cambridge Journal of Regions, Economy and Society 5, 271–287.

Christophers, B., 2012. Banking across Boundaries: Placing Finance in Capitalism. Wiley-Blackwell, Oxford.

Clark, G.L., 2000. Pension Fund Capitalism. Oxford University Press, Oxford.

Clark, G.L., 2003. European Pensions & Global Finance. Oxford University Press, Oxford.

Clark, G.L., 2011. Myopia and the global financial crisis: context-specific reasoning, market structure, and institutional governance. Dialogues in Human Geography 1, 4–27.

Clark, G.L., 2013. Mapping financial literacy. Geografiska Annaler B: Human Geography 95, 111–130.

Clark, G.L., 2014a. Financial literacy in context. Economic Geography 90, 1–23.

Clark, G.L. 2014b. The geography of the Euro crisis: the ECB, its institutional form, functions, and performance. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2456154.

Clark, G.L., Hebb, T., 2004. Corporate engagement: the fifth stage of capitalism. Relations Industrielles/Industrial Relations 59, 142–169.

Clark, G.L., Thrift, N., 2005. The return of bureaucracy: managing dispersed knowledge in global finance. In: Knorr Cetina, K., Preda, A. (Eds.), The Sociology of Financial Markets. Oxford University Press, Oxford, pp. 229–249.

Clark, G.L., Wójcik, D., 2007. The Geography of Finance: Corporate Governance in the Global Marketplace. Oxford University Press, Oxford.

Coval, J.D., Moskowitz, T.J., 1999. Home bias at home: local equity preference in domestic portfolios. Journal of Finance 54, 2045–2073.

Coval, J.D., Moskowitz, T.J., 2001. The geography of investment: informed trading and asset prices. Journal of Political Economy 109, 811–841.

Dixon, A.D., 2012. Function before form: macro-institutional comparison and the geography of finance. Journal of Economic Geography 12, 579–600.

Dixon, A.D., 2014. The New Geography of Capitalism: Firms, Finance, and Society. Oxford University Press, Oxford.

Doremus, P.N., Keller, W.W., Pauly, L.W., Reich, S., 1998. The Myth of the Global Corporation. Princeton University Press, Princeton.

Engelen, E.R., Ertürk, I., Froud, J., Sukhdev, J., Leaver, A., Moran, M., Williams, K., 2012. Misrule of experts? The financial crisis as elite debacle. Economy & Society 41 (3), 360–382.

Fama, E., 1970. Efficient capital markets: a review of theory and empirical work. Journal of Finance 25, 383–417.

Faulconbridge, J., 2008. Negotiating cultures of work in transnational law firms. Journal of Economic Geography 8, 497–517.

Faulconbridge, J., Engelen, E., Hoyer, M., Beaverstock, J., 2007. Analysing the changing landscape of European financial centres: the role of financial products and the case of Amsterdam. Growth and Change 38, 279–303.

Hall, P.A., Soskice, D.W., 2001. Varieties of Capitalism: The Institutional Framework of Comparative Advantage. Oxford University Press, Oxford.

Hall, S., Appleyard, L., 2008. 'City of London, city of learning'? Placing business education within the geographies of finance. Journal of Economic Geography 9, 597–617.

Hawley, J.P., Williams, A.T., 2000. The Rise of Fiduciary Capitalism: How Institutional Investors Can Make Corporate America More Democratic. University of Pennsylvania Press, Philadelphia, PA.

Huberman, G., 2001. Familiarity breeds investment. Review of Financial Studies 14, 659–680.

King, R., Levine, R., 1993a. Finance and growth: Schumpeter might be right. Quarterly Journal of Economics 108, 717–737.

King, R., Levine, R., 1993b. Finance, entrepreneurship, and growth: theory and evidence. Journal of Monetary Economics 32, 513–543.

Landreth, D., 2012. The Face of Mammon: The Matter of Money in English Renaissance Literature. Oxford University Press, Oxford.

Langley, P., 2008. The Everyday Life of Global Finance: Saving and Borrowing in Anglo-America. Oxford University Press, Oxford.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R., 1998. Law and finance. Journal of Political Economy 106, 1113–1155.

Leyshon, A., Burton, D., Knights, D., Alferoff, C., Signoretta, P., 2004. Towards an ecology of retail financial services: understanding the persistence of door-to-door credit and insurance providers. Environment and Planning A 36, 625–645.

Leyshon, A., Burton, D., Knights, D., Alferoff, C., Signoretta, P., 2006. Walking with moneylenders: the ecology of the UK home-collected credit industry. Urban Studies 43, 161–186.

Leyshon, A., French, S., 2008. Financial exclusion and the geography of bank and building society branch closures in Britain. Transactions, Institute of British Geographers NS33, 447–465.

Leyshon, A., Thrift, N.J., 1997. Geographies of financial exclusion: financial abandonment in Britain and the United States. Transactions, Institute of British Geographers NS20, 312–341.

MacKenzie, D., 2006. An Engine, Not a Camera: How Financial Models Shape Markets. MIT Press, Cambridge, MA.

Markowitz, H., 1952. Portfolio selection. Journal of Finance 7, 77–91.

Merton, R.C., 1990. Continuous-Time Finance. Blackwell, Oxford.

Merton, R.C., Bodie, Z., 2005. The design of financial systems: towards a synthesis of function and structure. Journal of Investment Management 3, 1–23.

Mishkin, F., 2006. The Next Great Globalization. Princeton University Press, Princeton.

Wainwright, T., 2011. Tax doesn't have to be taxing: London's 'on shore' finance industry and the fiscal spaces of a global crisis. Environment and Planning A 43, 1287–1304.

Walker, R.A., 1989. A requiem for corporate geography: new directions in industrial organization, the production of place and the uneven development. Geografiska Annaler B: Human Geography 71, 43–68.

Wójcik, D., 2011. The Global Stock Market: Investors and Intermediaries in an Uneven World. Oxford University Press, Oxford.

Wójcik, D., 2013. Where governance fails: advanced business services and the offshore world. Progress in Human Geography 37 (3), 330–347.

Financial Crises, History of

Giovanni Zanalda, Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the history of financial crises on the basis of recent historiography on the subject. First, it presents the main definitions of crises, banking, debt, currency, and other forms such as inflation and sudden stops. Then after reviewing the issue of contagion it describes a general pattern of boom and bust, from the initial euphoria to the distress stage. The last section reviews historical examples with a particular focus on the frequency of different types of crises in advanced and emerging economies.

Introduction

Financial crises are an old problem. They range from debt and inflation crises in ancient Greece and Rome to banking, debt, and currency crises that have been affecting countries in modern times. Aside from the historical value of the study of past crises, it is the lessons that can be learned and applied to prevent the formation of credit bubbles and the handling of crises that makes the study of past crises an important subject for social scientists, practitioners of finance, and policymakers.

Before the collapse of Lehman Brothers in 2008 and the following financial crisis in the United States and other advanced economies there was a widespread belief that financial crises had become mainly an emerging economies' problem. Indeed, monetary authorities and financial institutions in advanced economies with the exception of Japan had managed to contain market disruptions that followed financial crises over the course of the 1990s and early 2000s. The Great Recession that followed the 2008 crisis quickly reversed this opinion and triggered a renewed interest in the causes, consequences, and policy responses to financial crises.

The study of financial crises also improves our understanding of credit markets and financial institutions at national and international levels. History shows that governments and credit institutions have responded to financial crises by rushing to design and create institutions, agencies, and new regulations during and in the aftermath of crises. Hence, learning from past blunders as much as from successful actions can improve the design of institutions, policies, and market regulations which in turn could reduce the incidence and virulence of financial crises. For instance, the Bank of England started to perform the role of lender of last resort, at least *de facto* if not *de jure*, during the South Sea bubble in the early 1720s and acted in that capacity to manage other crises onwards; the precursor of the New York Stock Exchange was established after a major financial crisis in 1792; the Federal Reserve (FED) was funded in part as a response to the Wall Street crisis of 1907; agencies such as the FDIC and major banking regulations such as the Glass–Steagall Act were introduced in the 1930s during the Great Depression; reforms of financial systems in countries such as South Korea, Mexico, and other emerging markets were introduced in the second half of the 1990s in response to financial crises in those countries and contributed to make their banking systems more resilient to financial shocks; and the financial crisis that followed the collapse of Lehman Brothers in 2008 forced the US congress and other governments in Europe to introduce sweeping reforms of domestic and international financial markets as well as banks to implement stricter risk management procedures.

Furthermore, the study of crises improves our understanding of the work of an economy, its structure, and responses to domestic and external shocks. Analyses of banking, currency, debt, and inflation crises have, among other things, enhanced our understanding of monetary circulation, the crucial role of interest rates, exchange rate regimes, and macro linkages and contributed to the development of models which in turn have influenced central banks and governments' policies – Irving Fisher, Milton Friedman, John Maynard Keynes, Paul Krugman, and John B. Taylor are famous examples. The existence of a correlation between quantity of money and price, the underpinning of what became known as the quantity theory of money, is for instance the result of a long genealogy of research, initially circumstantial and then more theoretical, on inflation crises caused first by debasement, then by increases in the supply of silver and gold, and eventually by the excessive printing of paper currency in fiat money monetary systems.

The vast literature on the subject ranges from historical overviews to analyses of financial crises caused by macroeconomic instability (including the mishandling of monetary, fiscal, and exchange rate policies) or works on disruptions of financial systems caused by the mismanagement of credit instruments, banks, and other financial institutions. The most influential contributions include a combination of these approaches, as, for instance, in the now classic Charles Kindleberger's *Manias, Panics, and Crashes: A History of Financial Crises (1978, and with Aliber 2011)* and the more recent Carmen Reinhart and Kenneth Rogoff's *This Time Is Different: Eight Centuries of Financial Folly* (2009). New contributions such as *Financial Crises: Causes, Consequences, and Policy Responses* (Claessens et al., 2014) provide an excellent overview of where we stand in terms of knowledge of determinants, management, impact, and lessons of financial crises.

The analysis of a financial crisis follows a typical scheme, like in a pathology report as well captured by Kindleberger's definition of 'anatomy' of a crisis. First, it looks at the causes of the boom and bust, the formation of a bubble whether it started in the stock market, banking, real estate, or other sectors of the economy. Second, it looks at the unfolding whether it is has been circumscribed to an institution, sector, or systemic, and what type of policy tools have been used – monetary, fiscal, exchange rate, and regulatory interventions. Third, it

gauges the impact and implications for the financial sector and for the overall economy, and whether the crisis has been circumscribed to the domestic economy or spread to foreign markets and institutions – international contagion. Measures of output losses, employment, private and public debt levels, number of bankruptcies, sector-specific indicators (e.g., number of foreclosures in the real estate sector), stock market indicators, and capital and trade flows are commonly used to determine the impact of a crisis, its magnitude, and length. Examples of these indicators and recovery time can be found in Reinhart and Rogoff (2009).

Financial crises can be identified by a series of typical features: unusual variation in the level of asset prices and the volume of credit; marked disruptions in credit markets (credit crunch); liquidity and solvency problems of large/systemic financial institutions; balance sheet problems of firms and households and of sovereigns (debt crisis); exceptional monetary authority's intervention to tackle liquidity and solvency problems; and exceptional fiscal measures to lessen the impact of economic downturns caused by a major crisis (Claessens and Kose, 2014).

Types of Financial Crises

There are different types of financial crises. The classical typology of crises encompasses banking, currency, and debt crises. Crises occurring simultaneously are labeled twin crises in the case of a combination of two crises (banking and currency, banking and sovereign debt, currency and sovereign debt) and triplet crisis when a country experiences banking, currency, and sovereign debt within a short period of time. Argentina, for instance, experienced a triplet crisis in 2001 when the country defaulted on its external obligations (debt crisis), abandoned its exchange rate peg to the US dollar (currency crisis), and experienced a banking crisis. In response to a series of capital account or balance of payments crises in emerging economies in the 1980s and 1990s, a further classification, 'sudden stop,' has been added to the traditional ones (Calvo et al., 2004; Edwards, 2004).

Within this widely accepted classification of financial crises, scholars have adopted ad hoc definitions in particular when constrained by data availability. In their work on the history of financial crises Reinhart and Rogoff (2009) differentiate crises on the basis of whether they can be defined by quantitative thresholds or by qualitative evidence. In the first group (quantitative) they include currency crashes, debasement, inflation, and sudden stop. A *currency crisis* entails a speculative attack on the exchange value of a currency which results in a devaluation/depreciation of the currency and disruption of external payments relations. Authorities could defend the currency by selling international reserves or using monetary tools such as raising interest rates to prevent capital outflow or legal barriers such as the imposition of capital controls (Claessens and Kose, 2014). Causes of currency crises include exchange rate misalignment, tension between domestic policies and the peg, and expectations based on balance of payments imbalances. The increase in the occurrence and incidence of these crises that followed the breakup of the Bretton Woods system of fixed exchange rate in the early 1970s stimulated new lines of research. Economists identified and grouped currency crises into three generation models. The first generation focuses on speculative attacks triggered by the anticipation of an imminent change in a country's exchange rate regime or parity. The second generation deals with the existence of multiple equilibria when investors attack a currency on the expectations that other market participants will do the same. Third-generation models look at how the realization of financial and corporate balance sheet mismatches leads to currency crises (Krugman, 1998; Claessens and Kose, 2014).

Debasement, an important category for scholars interested in the history of ancient and early modern crises when sovereigns used it as an easy and effective tool to raise revenues to finance wars and state expenditures, is a form of default on obligations as much as inflation, currency, and debt crises that have plagued emerging markets over the course of the last three decades. *Inflation* crises are defined by some authors as such when a country experiences an annual inflation rate of 20% or higher with extreme cases (mild and acute cases of hyperinflation) when inflation rate exceeds the 40% per year threshold (Reinhart and Rogoff, 2009). A *sudden stop* (also called balance of payments crisis) occurs when a country experiences a sudden reversal of capital inflows accompanied by a sharp rise in its credit spreads and followed by exchange rate depreciation, and (often) stock market declines. Emerging markets in Latin America and East Asia in the second half of the 1990s and Central and Eastern Europe after 2008 experienced a sudden reversal after having recorded large capital inflows in the years before the beginning of the crisis.

The second group contains banking crises, external and domestic debt crises. A *banking crisis* is defined on the basis of two distinct occurrences. In the first type bank runs mark the beginning of a crisis that leads to closure, merging, or takeover of one or more financial institutions. In the second type, no bank run marks the beginning of a crisis but the closure, merging, or takeover by public authorities of a systemically important financial institution leads to a similar outcome for other financial institutions. A banking crisis could become 'systemic' when it generates severe financial distress including disruptions in credit markets, domestic and/or international, and is followed by defaults in the corporate and financial sectors. Triggers of banking crises include runs on banks (frequent in the past before central banks became effective lender of last resort and the creation of insurance on deposits); abrupt correction in asset prices (frequent in the presence of asset bubbles); loss of confidence in the banking system (frequent in countries with weak banking systems as in emerging markets before the 1990s crises); poor supervision; and solvency problems in systemically relevant financial institutions (an extremely dangerous situation) (Reinhart and Rogoff, 2009; Laeven and Valencia, 2014).

A *debt crisis* takes place when a debtor, sovereign in the case of a state or private, is unable to service interest and principal as scheduled. Debt crises could end up in default and repudiation. A foreign debt crisis occurs when a government defaults on its own external or publicly guaranteed private-sector obligations held by foreign creditors and often denominated in foreign currency. Argentina's default on $95

billion in external debt in 2001 is an example of this type of default. Episodes of rescheduling in which governments repay their debt at less favorable terms than in the original loan are also classified as debt crises in several studies on the subject. Likewise in terms of domestic debt crises in addition to outright default it is now common practice to include episodes of debt restructuring. They include the freezing of bank deposits and conversions of deposits from foreign to local currency; the debasement or depreciation of the currency; strict regulations on capital movements; and preferential tax treatments for government debt among other forms of what is often referred to as financial repression (Reinhart and Rogoff, 2009, 2014).

Contagion

The study of financial crisis has increasingly focused on the contagion effect, on the channels of international transmission of a crisis (Neal and Weidenmier, 2003). Financial crises propagate across countries either when a crisis has a simultaneous impact on various economies or when it spreads from the initial center to other economies (Kindleberger and Aliber, 2011). Channels of transmission range from trade and capital flows, security and asset markets, to currency fluctuations and institutional features – links among banks and financial institutions of different countries, level of cooperation among monetary authorities, exchange rate regime.

The first type of contagion – simultaneous impact – generally occurs when a common shock such as sudden increases in commodity prices, international interest rates, or simultaneous collapse in housing prices in different countries ignites major financial distress and eventually economic crises in economies sharing common fundamentals. For instance, oil shocks of the 1970s and early 1980s had a negative impact on most advanced economies and non-oil-exporting developing countries; a sudden rise in international interest rates in the early 1980s triggered a debt crisis in highly indebted countries in Latin America; and a rapid decline in housing prices in the United States and other economies in 2007 and 2008 had dramatic consequences on the banking sector of countries with highly leveraged housing markets like the United States, the United Kingdom, Ireland, Iceland, and Spain.

The second type of contagion occurs when a domestic crisis spreads to other countries through cross-border linkages. For instance, the downfall of Lehman Brother in 2008 had a profound impact on foreign banks and financial systems exposed to the US subprime market (via mortgage-based securities and credit default swaps); or the exposure of German and French banks to EU peripheral nations' sovereign debt crisis has contributed to raise the threat of international contagion had one of the EU members defaulted. On the other hand, countries with less internationally integrated financial and banking sectors or closed economies have traditionally been less exposed to contagion. The soviet economy of the USSR, an autarkical system, was not affected by the Great Depression and indeed outperformed most Western economies in terms of economic growth throughout the 1930s. The fact that China, India, and Vietnam had rather closed domestic financial systems contributed among other factors to shield them from the East Asian financial crisis of the 1990s.

The two channels are not exclusive. The existence of both common fundamentals and cross-border linkages contributed to the spread of major crises around the world. The Great Depression affected the United States and European nations because of common domestic fundamentals – e.g., solvency of banks, the gold standard – as well as of macro linkages – e.g., integration of capital markets, geopolitical factors (the issue of Germany's war repayments), trade, and the gold standard adjustment mechanism. The East Asian crisis of the 1990s spread across countries with common fundamentals like overvalued currencies, real estate and equity bubbles, and weak credit markets. Macro linkages like rapid decline in capital inflows also contributed to exacerbate the impact of the crisis. The late 2000s crisis with "the bunching of banking crises and sovereign debt difficulties across countries" features both common shocks and cross-country linkages (Reinhart and Rogoff, 2011: p. 18).

In studying the international implications of crises, scholars have also identified two other important features of transmission on the basis of whether contagion takes place in a matter of hours or days, 'fast and furious,' or over time, 'slow-burn spillover' (Kaminsky et al., 2003). Fast and furious type of contagion tends to be characterized by an element of surprise (unanticipated shock) while the slow type by anticipated changes (e.g., an expected rise in interest rate). The country experiencing fast contagion features a rapid surge in capital inflows, with short-maturity debt. This is typical of countries running current account deficits and experiencing large capital inflows which in turn sustain credit and asset bubbles. Following the abrupt explosion of financial crisis countries in this category often experience a 'sudden stop' crisis, a quick reversal in capital flows with highly leveraged investors scrambling to exit the market. The existence of a common lender for countries exposed to the same crisis is worse in the case of unanticipated crisis since the lender has very little time to reduce or hedge its exposure (Reinhart and Rogoff, 2011).

Bubbles, Crises, and Financial Cycles

One of the most famous interpretations of the causes of financial crises goes back to the work of the Swedish economist Knut Wicksell at the end of the nineteenth century. According to Wicksell, a view later developed with different nuances by John Maynard Keynes, Joseph Schumpeter, and Irving Fisher and more recently by Hyman Minsky, crises are intrinsic to the financial cycle (or part of the financial cycle). In the first stage of a financial cycle, loose monetary policy contributes to expand credit and increase asset prices. The rise in asset prices will continue to the point of becoming a bubble when the level of price for that specific asset category is not justified by underlying fundamentals. The bubble eventually bursts and will be followed by a rapid decline in asset prices and in the case of major asset markets such as real estate by a recession. The depth and length of the recession will depend on the scale of the bubble and on the management of the fallout. Within this tradition,

Minsky developed one of the most convincing theoretical frameworks for understanding the recurrence of financial crises over time – the financial instability hypothesis. According to this hypothesis, financial systems are unstable, fragile, and prone to crisis because of the instability of the supply of credit. Charles Kindleberger using Minsky's framework shows, on the basis of a comprehensive analysis of historical accounts of crises from the seventeenth century onwards, how investors' irrational behavior, policy mismanagement, asymmetry of information, and the unreliability of credit systems have contributed to cycles of booms and busts. The financial instability hypothesis, according to Kindleberger "has great explanatory power for earlier crises in the United States and in Western Europe, for the asset price bubbles in Japan in the second half of the 1980s and in Thailand and Malaysia and the other countries in Southeast Asia in the mid-1990s, and for the bubble in US stocks, especially those traded on the NASDAQ, at the end of the 1990s" (Kindleberger and Aliber, 2011: p. 15).

The general pattern of a crisis identified by Minsky and Kindleberger can be dissected into five distinct stages of the financial cycle. It starts with a 'displacement,' when new technologies, financial innovations, new markets, or new policies create unusually profitable investment opportunities in a relevant sector of an economy or in a particular country or group of countries. The belief that expected profits will keep rising in that sector of the economy or country leads to rapid growth in the price of the underlying assets; this is the 'euphoria' or overtrading stage. Then first-time investors attracted by easy profits will get into the market: the 'mania' or bubble phase. Also, in this phase investors become more willing to take on debt because of the wealth effect and because, in several instances, the value of their collateral has risen – they become more leveraged. Market professionals and insiders realize that high prices of the assets are not justified by fundamentals and start selling those assets; the 'distress' stage. Finally, there is a 'revulsion' phase when asset prices fall; investors dump their shares, exiting the market of the distressed assets and any other market of similar assets (Kindleberger, 2011). On the latter issue, for example, when investors scrambled to exit the real estate market during the 2008–09 subprime mortgage crisis they also walked out of all sectors connected to it, including financial services. Or when Thailand and other East Asian countries experienced a major crisis in the second half of the 1990s, investors withdrew funds not only from those countries but from all emerging and developing markets. For stock markets, as remarked by Niall Ferguson, three further features can be added to this general pattern. The first occurs when insiders take advantage of the existence of asymmetric information – they use reserved information to cheat outsiders who do not have access to the same information. The second depends on the existence of cross-border capital flows: investors in more developed financial centers are better at timing the market before the burst of the bubble. The third is related to the existence of an easy credit environment. The availability of easy money increases the likelihood of a bubble. When markets are flooded with capital, investors chase high returns in riskier investments and financial firms become less diligent in risk assessment – it is a typical environment for moral hazard (Ferguson, 2008).

The three most studied bubbles and crashes of the early modern period – the Tulipmania in Holland (1634–37), the Mississippi and South Sea bubbles in France and Great Britain (1719–20) – featured the pattern described above. In all three cases a sharp increase in the price of particular assets in a short period of time – future contracts of tulip bulbs in Holland, state-backed securities in France, and trading of company shares and government debt in both France and England – was followed by an even faster contraction in price after the burst of the bubble. From a historical point of view it is worth noting how these episodes took place in two of the most advanced financial centers of the time – Amsterdam and London – and the largest continental economy – France. It shows how financial markets began to experience instability already in their infancy and at the same time that countries with a more developed financial system, such as Holland and England, fared much better than France, a country with a less sophisticated financial system and no lender of last resort. It was the swift action of a *de facto* lender of last resort, the Bank of England, the presence of an advanced financial sector, and the introduction of new regulations such as the Bubble Act of 1720 that enabled Great Britain to contain the negative spillover from the South Sea Company fiasco. (Atack and Neal, 2009; Garber, 2000).

A recent example of the different recovery time span experienced by countries affected by the same type of financial crisis can be found in the bubble in real estate and in stocks that occurred in Japan and Sweden in the second half of the 1980s. In both cases financial liberalization, easy credit, and inflow of capital lead to increases in consumption and to a lesser extent in investments. A boom in real housing prices and stock prices was followed by a sharp decline in both categories after the bubble bursts. The Swedish government nationalized the largest banks, created a good bank/bad bank to clean banks' balance sheets, and increased deposit insurance ceilings, all measures to prevent the collapse of the financial system and restore confidence in the banking sector. It also devaluated the currency by unpegging the Swedish krona from the European Monetary System. A sustained and lasting rebound of Sweden's economy began in 1994 only a few years after the burst of the bubble. Japan experienced a similar bubble buildup in the second half of the 1980s when both the stock market index, Nikkei, and land prices featured a threefold increase. Then at the beginning of the 1990s the situation reversed, the 'distress' and 'revulsion' phases indicated the burst of the bubble, stock and real estate prices sharply declined, and despite Japanese authorities' aggressive monetary policy and attempts to rescue the banking and insurance industries the country never fully recovered. Japan in contrast to Sweden failed to deal with the root causes of the banking and insurance sector problems. The legacy of this crisis has been that despite extremely accommodative monetary policy (interest rate at or close to the zero lower bound for a protracted period of time) Japan's annual rate of growth has been on average close to 1% throughout the 1990s and 2000s. Furthermore, large public deficits and accumulation of debt over the same time span contributed to bring the debt/GDP ratio above 200 percent, the highest level in the world.

Frequency

Financial crises of various types have erupted in most countries at least once over the course of the last five centuries. In some instances as isolated crises, in many other cases as part of waves of crises that affect regions – e.g., Latin America debt crisis in the 1980s or the East Asian Crisis of 1997 – or the world as in the case of the Great Depression in the 1930s and the Great Recession in 2008–12. The following list reflects the vast range of typology in terms of origin, time, and geographical coverage of the recent historiography of financial crises.

Time	*Name*	*Region*	*Main features*
1636	The Dutch tulip bubble	Holland	Speculation in future contract of tulip bulbs
1720	The South Sea bubble	Great Britain	Speculation in government debt;
1720	The Mississippi bubble	France	Speculation in government debt;
1792	Panic of 1792	The United States	Speculation in bank stocks and government debt;
1825	Latin American crisis	Latin America	Speculation on future growth of newly independent Latin American countries
1837	Cotton crisis	The United States	Fall in cotton prices, bubble in land prices, credit crunch
1857	Railroad crisis	The United States	Speculation in future earnings of railway companies
1873	Long depression	Europe and the United States	Depression in Europe after Franco-Prussian war; speculation in railway companies in the United States
1890	Baring crisis	Argentina and England	Speculation in Argentine securities
1907	Wall Street Panic	The United States	Speculation in trust companies shares
1929	1929 crash; Great Depression	The United States; world	Wall Street crash and beginning of the Great Depression
1970s	Oil crisis	Advanced economies	Collapse of the Bretton Woods system and oil crisis
1980s	Latin America debt crisis	Latin America	Surge in bank loans to Mexico and other Latin American countries in the 1970s and early 1980s
1987	Black Monday	The United States and world	Stock market crash
1991	Japan crisis	Japan	Bubble in real estate and stocks in 1985–89
1991	Scandinavian crisis	Finland, Norway, and Sweden	Bubble in real estate and stocks in 1985–89 followed by currency and banking crises
1994	Tequila crisis	Mexico	The surge in foreign investment in Mexico 1990–93
1997	Asian crisis	Thailand, Malaysia, Indonesia, South Korea, and other Asian countries	Bubble in real estate and stocks followed by currency and banking crises
2001	Argentina triple crisis	Argentina	Banking, currency, and debt crises
2001	Dotcom crash	The United States	Bubble in over-the-counter (OTC) telecom and internet stocks in the United States 1995–2000
2008	Subprime crisis	The United States	Bubble in housing market and mortgage backed securities
2009/13	Debt crisis	Europe	Debt crisis in several European Union member states

Sources: Kindleberger, Charles P., Robert Z. Aliber, 2011. Manias, Panics, and Crashes: A History of Financial Crises. Basingstoke. Palgrave Macmillan, UK; Reinhart, Carmen M., Kenneth S. Rogoff, 2009. This Time is Different: Eight Centuries of Financial Folly. Princeton University Press, Princeton, NJ; Claessens, Stijn, Ayhan, Kose, Luc, Laeven, Fabian Valencia. (Eds.), 2014. Financial Crises: Causes, Consequences, and Policy Responses. International Monetary Fund, Washington, DC.

Frequency and severity of financial crises have increased over time (Bordo et al., 2001). History shows that all countries have been afflicted by different forms of crises during their emerging phase. Emerging markets whether in the nineteenth or late twentieth century (e.g., several countries in Asia, Latin America, Central and Eastern Europe) have experienced waves of financial crises (Bordo and Eichengreen, 2002). This is explained in part by the fact that they experience credit booms which in turn create bubbles in some sectors and by the increased exposure to the risk of international contagion via two main transmission channels, capital and trade flows. Furthermore, while emerging these countries are developing their own banking sector, financial markets, and institutions (including independent central banks) which once established tend to improve financial stability and capability to prevent and manage financial turbulences. The table below presents a stylized summary of the incidence of crises in advanced and emerging economies – emerging at the time under consideration – of sovereign debt, banking and inflation crises over the course of the last five centuries.

	Type of crisis		
Time	*Banking*	*External debt*	*Inflation*
1550s–1810s	Rare	Frequent in advanced economies of the time; serial in some cases, e.g., Spain and France	Rare
1820s–1850	Serial in advanced economies Rare in emerging markets of the time	Frequent in peripheral advanced economies and most emerging markets	Rare
1850s–1913	Serial in advanced economies Frequent in emerging markets	Frequent in peripheral advanced economies and most emerging markets	Rare
1913–45	Serial in advanced economies Frequent in emerging markets	Frequent in peripheral advanced economies and most emerging markets	Frequent in advanced economies and emerging markets
Bretton Woods period	Rare	Frequent in peripheral advanced economies and most emerging markets	Rare
1973–1980s	Frequent in advanced; serial in emerging	Serial in some emerging markets	Frequent in advanced; serial in emerging
1990s	Frequent in advanced; serial in emerging	Serial in some emerging markets	Frequent in emerging
2000–09	Frequent in advanced	Serial in some emerging markets	Rare

Source: Qian, Rong, Carmen M. Reinhart, Kenneth S. Rogoff, 2010. On Graduation from Default, Inflation and Banking Crisis: Elusive or Illusion? NBER Working Paper 16168. National Bureau of Economic Research, Cambridge, MA.

These tables provide some clues to guide us through the maze of financial crises in a long-term perspective. In the first period, sixteenth to early nineteenth century, several European states (the advanced economies of the time) defaulted on their external debt – serial defaults in some cases like Spain (8 times), France (9 times), Prussia (3 times), and Austria-Hungary (5 times) (Reinhart and Rogoff, 2009). Occasional banking crises occurred in some European states but most of them were circumscribed to a few institutions. Bubbles were characterized by a major surge in the price of particular assets or financial flows (e.g., loans) but not all of them were followed by deep and long financial crises (Qian et al., 2010). Episodes of inflation during the early modern period were short lived and relatively low in comparison to those recorded during the interwar period and again after the collapse of Bretton Woods in the 1970s.

The period between the end of the Napoleonic Wars and the beginning of World War I featured a series of bubbles followed by stock market crashes, banking, and debt crises. The spread of the industrial revolution from England to continental Europe, the United States, and Japan followed by large surges in capital, trade, and migration flows had indeed contributed with ups and downs to the internationalization of the world economy. This process had been made possible among other things by the introduction and development of new credit instruments, banking activities, the adoption of the gold standard in almost all advanced economies, the increase in the speed of communication and information across the world, the rise of international financial centers, and readiness to move capital around the world (O'Rourke and Williamson, 1999). On the other hand, it also meant that economic problems and financial crises in one country could spread to other countries and regions. Hence, financial crises and bankruptcies that occurred in major financial centers of the time – London, Paris, Hamburg, and New York – were transmitted across the world affecting banks in other advanced economies and emerging markets. One of these crises which started in the United States in the fall of 1857 featured liquidity problems in New York which in turn affected London, Germany, and Latin America, and was analyzed by Walter Bagehot in his famous study, Lombard Street (1873). In his work, Bagehot formulated one of the most famous rules of central banks' intervention to provide credit to an illiquid bank. He wrote that a central bank in his role of lender of last resort should lend freely but at a penalty interest rate to prevent the liquidity problem from escalating into a solvency problem and in turn spreading to the banking system (Marichal, 2014; Kindleberger and Aliber, 2011).

The most famous of the late nineteenth-century international crises began in 1873 with stock market crashes and banking crises in Austria, Germany, and the United States and then spread to other countries in Europe, Latin America, and the Near East. Given its impact in terms of bankruptcies (mainly in advanced economies), debt crises (in emerging economies), and in terms of deep, widespread, and protracted economic recessions this crisis has been labeled as the first great depression. Other speculative manias in Paris, London, and New York ended up in banking crises in the 1880s and 1890s. Another crisis which started in Argentina in 1890 and involved one of the most famous London banks, Baring, had repercussions throughout other financial centers in Latin America, Lisbon, and Madrid (Mitchener and Weidenmier, 2008).

In terms of nineteenth-century debt crises, 'peripheral' nonindustrialized economies such as Spain, Portugal, and Greece in Europe; Egypt, Tunisia, and the Ottoman Empire in the Mediterranean region; and eight countries in Latin America defaulted on their debt – in some cases multiple times over the 1830–1913 period. Over the same period, improvements in the management of public finance together

with an increased availability of domestic capital enabled the largest advanced economies to avoid external default (Qian et al., 2010; Marichal, 2014).

At the beginning of the twentieth century the United States suffered one of the most damaging financial crises before the Great Depression. It started with a bank panic in New York in November 1907 and lasted more than a year. During that time, the stock market dropped 40% and almost 30 banks and 17 trust companies failed. The violent contraction felt throughout all sectors of the economy led to a severe depression in 1908. This crisis exposed the fragility of the financial system without an effective lender of last resort – it was indeed a group of financiers led by J.P. Morgan who intervened to prevent a complete collapse of Wall Street – and became a catalyst for establishing the Federal Reserve (1913) and the US banking system as it still operates today (Bruner and Carr, 2007).

The stock price bubble of 1927–29 in the United States which marked the beginning of the Great Depression had major domestic and international implications in terms of financial crises. Banking, debt, and currency crises became a worldwide phenomenon throughout the 1930s (Kindleberger, 1986; Temin, 1989; Bernanke, 2000; Eichengreen, 2008; Feinstein et al., 2008; Crafts and Fearon, 2010). The literature on the Great Depression is so extensive that it would require an entry in its own right. Among the numerous lessons from the 1930s, we recall how lack of cooperation among governments and their central banks generated further deflationary pressure on the world economy and exposed weak currencies to speculative attack. Central banks and governments believed in the monetary orthodoxy of the gold standard, which prevented countries from embarking on countercyclical policies (Temin, 1989; Eichengreen, 1995). Overall, the best solution for the world economy would have been a coordinated effort among countries to dismantle the gold standard in an ordinate manner. This did not happen, and countries that began to use exchange rate devaluation early in the decade together with other strategies such as expansionary policies (the United Kingdom, Sweden, and Japan), protection and import substitution (Brazil and Colombia), and capital controls and domestic expansion (Germany) outperformed those that abandoned the gold standard a few years later (the United States, France, and the Gold Bloc countries) (Feinstein et al., 2008).

During the period from the end of World War II until the early 1970s in part for the stability of the international monetary system banking crises were rather infrequent. After the collapse of Bretton Woods in the early 1970s the world economy entered an extremely unstable period during which both advanced and emerging economies experienced severe financial crises. The following figure based on Laeven and Valencia's (IMF) database provides an overview of the number of episodes by type of financial crises for the period 1970–2012. According to this set of data the world experienced 147 banking crises, 211 currency crises, and 67 sovereign debt crises. Several countries experienced overlapping crises which resulted in 41 twin crises – those involving currency crises (banking/currency and currency/debt) the most frequent – and 10 triplet crises (Laeven and Valencia, 2013, 2014).

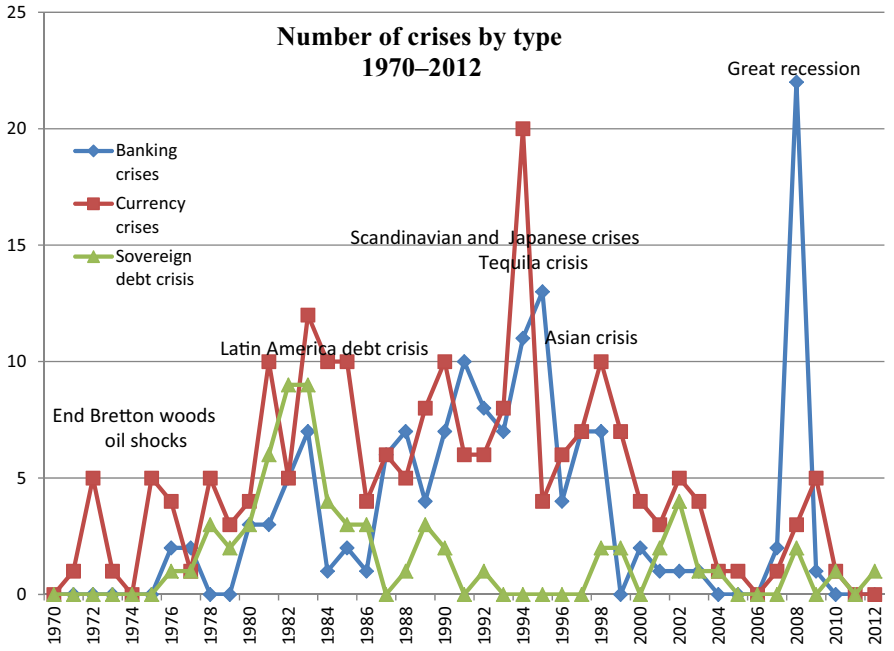

Source: Laevene and Valencia (2013, 2014).

See also: Cities and the Financial Crisis; Economic Growth; Economic History, Quantitative: United States; Economics, History of; Financial Regulation; Global Governance; Global Public Policy; Macroeconomics, History of From 1933 to Present; Market Structure and Performance; Monetary Policy Since the 2007–2009 Financial Crisis; Public Policy; Stock Market Predictability.

Bibliography

Atack, Jeremy, Neal, Larry (Eds.), 2009. The Origins and Development of Financial Markets and Institutions: From the Seventeenth Century to the Present. Cambridge University Press, New York.

Bagehot, Walter, 1873. Lombard Street: A Description of the Money Market. Scribner, Armstrong & Co, New York.

Bruner, Robert F., Carr, Sean D., 2007. The Panic of 1907: Lessons Learned from the Market's Perfect Storm. John Wiley & Sons, Hoboken, NJ.

Bernanke, Ben, 2000. Essays on the Great Depression. Princeton University Press, Princeton, NJ.

Bordo, Michael, Eichengreen, Barry J., Klingebiel, D., Martinez-Peria, M.S., April 2001. Is the crisis problem growing more severe? Economic Policy.

Bordo, Michael, Eichengreen, Barry J., 2002. Crises Now and Then: What Lessons from the Last Era of Financial Globalization. NBER Working Paper 8716. National Bureau of Economic Research, Cambridge, MA.

Calvo, Guillermo A., Izquierdo, Alejandro, Mejía, Luis-Fernando, 2004. On the Empirics of Sudden Stop: The Relevance of Balance-sheet Effects. NBER Working Paper 10520. National Bureau of Economic Research, Cambridge, MA.

Claessens, Stijn, Kose, Ayhan, Laeven, Luc, Valencia, Fabian (Eds.), 2014. Financial Crises: Causes, Consequences, and Policy Responses. International Monetary Fund, Washington, DC.

Claessens, Stijn, Kose, Ayhan, 2014. Financial crises: explanations, types, and implications. In: Claessens, S., Kose, A., Laeven, L., Valencia, F. (Eds.), Financial Crises: Causes, Consequences, and Policy Responses. International Monetary Fund, Washington, DC.

Crafts, Nicholas, Fearon, Peter, (Eds.), 2010. Lessons from the 1930s. Oxford Review of Economic Policy 26 (3), 285–570.

Edwards, Sebastian, 2004. Financial openness, sudden stops, and current account reversals. American Economic Review 94 (2), 59–64.

Eichengreen, Barry J., 1995. Golden Fetters: The Gold Standard and the Great Depression 1919–1939. Oxford University Press, Oxford, NY.

Eichengreen, Barry J., 2008. Globalizing Capital: A History of the International Monetary System. Princeton University Press, Princeton, NJ.

Feinstein, C.H., Temin, Peter, Toniolo, Gianni (Eds.), 2008. The World Economy between the World Wars. Oxford University Press, New York.

Ferguson, Niall, 2008. The Ascent of Money: A Financial History of the World. Penguin Press, New York.

Garber, Peter M., 2000. Famous First Bubbles: The Fundamentals of Early Manias. MIT Press, Cambridge, MA.

Kaminsky, Graciela L., Reinhart, Carmen M., Végh, Carlos A., 2003. The unholy trinity of financial contagion. Journal of Economic Perspectives 17 (4), 51–74.

Kindleberger, Charles P., 1986. The World in Depression. Allen Lane, London.

Kindleberger, Charles P., Aliber, Robert Z., 2011. Manias, Panics, and Crashes: A History of Financial Crises. Palgrave Macmillan, Basingstoke, UK.

Krugman, Paul R., 1998. Currencies and Crises. MIT Press, Cambridge, MA.

Laeven, Luc, Valencia, Fabian, 2013. Systemic banking crises database. IMF Economic Review 61, 225–270.

Laeven, Luc, Valencia, Fabian, 2014. Systemic banking crises. In: Claessens, S., Kose, A., Laeven, L., Valencia, F. (Eds.), Financial Crises: Causes, Consequences, and Policy Responses. International Monetary Fund, Washington, DC.

Marichal, Carlos, 2014. Historical reflections on the causes of financial crises: official Investigations, past and present, 1873–2011. Economic History Research 10, 81–91.

Mitchener, Kris J., Weidenmier, Marc D., 2008. The baring crisis and the great latin American meltdown of the 1890s. Journal of Economic History 68, 462–500.

Neal, Larry, Weidenmier, Marc, 2003. Crises in the global economy from tulips to today: contagion and consequences. In: Bordo, Michael, Taylor, Alan, Williamson, Jeffrey G. (Eds.), Globalization in History. NBER and University of Chicago Press, Chicago.

O'Rourke, Kevin H., Williamson, Jeffrey G., 1999. Globalization and History : The Evolution of a Nineteenth-century Atlantic Economy. MIT Press, Cambridge, MA.

Qian, Rong, Reinhart, Carmen M., Rogoff, Kenneth S., 2010. On Graduation from Default, Inflation and Banking Crisis: Elusive or Illusion?. NBER Working Paper 16168 National Bureau of Economic Research, Cambridge, MA.

Reinhart, Carmen M., Rogoff, Kenneth S., 2009. This Time Is Different: Eight Centuries of Financial Folly. Princeton University Press, Princeton, NJ.

Reinhart, Carmen M., Rogoff, Kenneth S., 2011. A Decade of Debt. Peterson Institute For International Economics, Washington, DC.

Reinhart, Carmen M., Rogoff, Kenneth S., 2014. Financial and sovereign debt crises: some lessons learned and those forgotten. In: Claessens, S., Kose, A., Laeven, L., Valencia, F. (Eds.), Financial Crises: Causes, Consequences, and Policy Responses. International Monetary Fund, Washington, DC.

Temin, Peter, 1989. Lessons from the Great Depression. MIT Press, Cambridge, MA.

Financial Regulation

Lucia Quaglia, Derwent College, University of York, York, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article begins by defining some key concepts in financial regulation. It subsequently discusses the main objectives of financial regulation and supervision, as well as some of the most problematic issues in this area. After reviewing the main perspectives developed in the literature on this topic, the final part of the article outlines the main causes and consequences of the global financial crisis.

Introduction

The global financial crisis, which began in 2007 with the collapse of the US subprime mortgage market, delivered a major shock to the existing architecture for financial services regulation and supervision worldwide. It also brought into the spotlight the political salience of financial regulation, which has become a matter of great public interest, outside the restricted circle of policy makers, stakeholders, and experts involved.

The crisis highlighted the fact that policy failures in this sector directly affect the citizens in their position as bank account holders, small investors, insurance holders, pensioners, and so on. They are also indirectly affected as taxpayers, whenever the public authorities have to intervene to bail out or prop up ailing financial institutions, as happened during the financial crisis. Moreover, there is increasing public concern about the transparency and accountability of institutions and policy-making processes in this area.

This article begins by defining some key concepts, such as financial regulation, financial supervision, financial services, etc. It subsequently discusses the main objectives of financial regulation and supervision as well as some of the most problematic issues. After reviewing the main perspectives developed in the literature on this topic, the final part of the article outlines the main causes of the global financial crisis and the principal trends in the reform of financial regulation after the financial crisis.

Defining Key Concepts in Financial Regulation

'Financial regulation' concerns rule setting, that is, establishing the legal framework governing market entry and market activities in the financial sector. 'Financial supervision' concerns rule implementation, that is, the monitoring and the enforcement of regulation as well as the practical cooperation between supervisory authorities across borders and across various segments of the financial sector.

Two main types of regulation and supervision can be identified: 'prudential' and 'conduct of business.' Prudential regulation and supervision are concerned with ensuring the adherence of individual financial entities, such as banks, to prudential regulatory standards with a view to maintain financial stability. The economic literature generally distinguish between 'macro-' and 'micro'prudential regulation (Borio, 2011), which have two specific objectives, namely, macro- and microfinancial stabilities. Macrostability concerns the stability of the entire financial system and the stability of the key institutions and markets that are part of the financial system. Microstability concerns the stability (i.e., solvency) of intermediaries, viewed from the prospective of the users of the system. 'Conduct of business' regulation and supervision focuses on how financial firms conduct business with their customers. It deals with mandatory information disclosure, the integrity of firms and their employees, fair business practices, the marketing of financial products etc.

'Financial services' is a term used to refer to the services provided by the finance industry. A common classification of financial services includes banking, securities, and insurance. In the recent decades, market segmentation has tended to disappear, following a growing interpenetration between the various segments of the market. In other words, there has been an increasing blurring of boundaries between banking, securities, and insurance. Indeed, most of the main banks are financial conglomerates, which are financial institutions engaged in at least two of the following activities: banking, insurance, and securities trading. Nonetheless, the important difference across the main sectors remains and certain policy dynamics are sector specific.

'Financial integration' refers to the removal of barriers of various types to the free circulation of financial services. Often, this is accompanied by 'mutual recognition' of national legislation, which means that a country recognizes the regulations and administrative practices of other countries as equivalent to its own, thereby precluding the use of differences in national regulations to restrict access to domestic markets. It goes beyond the concept of 'national treatment' because countries agree to offer a treatment that is more favorable than national treatment to financial firms from other countries. A corollary of mutual recognition is 'home country control,' which means that financial supervision is carried out by the supervisory authorities of the country where a financial firm is headquartered. Unlike mutual recognition of national legislation, 'international harmonization' involves the approximation of regulation across jurisdictions through the deliberate creation of international standards. In the European Union, which is a regional jurisdiction, a distinction is often made between 'minimum harmonization' and 'maximum harmonization.'

Other concepts often used with reference to financial regulation are 'deregulation,' 'reregulation,' and 'liberalization.'

Deregulation means the downsizing or the elimination of public authorities' rules, reducing their control in the economy, in this case in the financial sector. Reregulation means the reformulation of old rules and the introduction of new rules. Reregulation can be procompetitive, with emphasis on liberalization, or it can place emphasis on complementing the liberalization process. Liberalization means the injection of more competition in the economy, in this case the financial sector.

The Objectives and Rationales of Financial Regulation

The main objectives of financial regulation are (1) to safeguard systemic stability, (2) to maintain the safety and soundness of financial institutions, (3) to protect consumers, and (4) to promote competition and efficiency. Internationally, regulation can also constitute a means of trade protectionism by exploiting specific comparative advantages and reducing adjustment costs to the new rules set in place. Indeed, the content of regulation can produce different adjustment costs for different countries and financial operators, substantially affecting their competitiveness, as discussed below.

There are three main rationales for financial regulation, which to a large extent mirror the objectives listed above: (1) to prevent potential systemic problems related to negative externalities, (2) to correct market imperfections and failures, and (3) to reduce moral hazard related to safety next arrangements, such as lender of last resort and deposit protection schemes.

The first rationale for financial regulation is to deal with systemic risks and negative externalities (Davies and Green, 2008). Certain financial institutions, first and foremost banks, have a pivotal role in the financial system and their failures can cause systemic problems. Banks are highly interconnected to other financial institutions, not only other banks (e.g., in their gross positions with each other in clearing systems and in interbank deposits), but also insurance companies and securities firms. Furthermore, many banks are financial conglomerates that combine under one roof banking activities, securities trading, and insurance. Banks also perform a crucial role in payment systems, which are vital for the activities of nonfinancial firms as well as the wider public.

The special tasks performed by banks mean that there are several externalities to be considered in regulating finance, especially in the banking sector. The failure of a bank can directly cause losses to other interconnected banks within and across borders. It can also cause a gridlock in payment systems. Finally, bank runs may have contagious systemic effects because of information symmetries. In the absence of reliable information, depositors have an incentive to withdraw their deposits from a bank facing (or allegedly facing) solvency problems. This turns into a self-fulfilling prophecy, which may have contagious effects, spreading to other banks.

The second rational for financial regulation is to protect consumers, correcting market imperfections or market failures that would compromise consumer welfare in a regulation-free environment. Market imperfections and failures stem from the fact that consumers are not in a position to make informed decisions about the security of their banks: they lack relevant information, they do not have the ability to assess the information available, and the transactions costs involved in doing so are high (Llewellyn, 1999).

The third rational for financial regulation is to deal with the moral hazard associated with safety-net arrangements, such as lender of last resort and deposit insurance, which are set in place by the public authorities. The lender of last resort function is performed by central banks that have the authority and financial resources to act as the ultimate source of credit (Goodhart, 1999). In emergency situations, central banks extend their loans to solvent but illiquid financial institutions, often to avoid a run on the banks accompanied by financial panic. Deposit insurance is a formal scheme normally established by law that is designed to limit the losses of depositors in the event of banks failures.

These safety net arrangements pose four potential moral hazards. First, consumers may be less careful in the selection of banks and may even seek high-risk institutions that offer high interest rates on the grounds that if the bank fails, they will receive compensation. Second, the financial firm may be induced to take more risk because depositors are protected in the event that the institution fails. Third, risk is subsidized because depositors do not demand an appropriate risk premium in their deposit interest rates, and financial counterparties are not careful in pricing in risk in their financial interactions with other financial institutions (Llewellyn, 1999). In this scenario, banks adapt their strategies to maximize expected profits.

The Main Open Issues in Financial Regulation

Financial regulation raises a variety of public policy issues (for an in depth discussion see Brunnermeier et al., 2009). Some of them are common to all types of regulation, others are specific to the financial sector. The first issue is the need to balance the costs and benefits of regulation, so as to prevent 'regulatory arbitrage' across jurisdictions, which takes place when financial operators and their activities move toward jurisdictions that regulate them less tightly. Financial services are inherently borderless activities: they can move relatively easily across borders, hence countries have an incentive not to regulate them too tightly. Yet, policy failures in this area can be costly and can cause substantial negative externalities because financial instability does not stop at national borders (Herring and Litan, 1994).

Therefore, effective financial regulation requires to some extent harmonized rules across jurisdictions to prevent regulatory gaps and arbitrage. International regulatory harmonization has transaction costs when it is negotiated and it has implementation costs when domestic rules are changed to comply with either international rules or rules from third countries (Simmons, 2001). All this raises the problem of the underprovision of the public (or 'semipublic') good of financial stability, meaning that everybody benefits from financial stability, but the costs of providing this public good tend to be specific (Schinasi, 2006).

The second issue in financial regulation concerns the potential for 'regulatory capture,' meaning that the financial industry or part of it can have overwhelming influence in the shaping of regulation and the carrying out of supervision by the

competent authorities. The financial industry has powerful resources at its disposal, such as expertise and technical knowledge; human resources, engaged in lobbying activities; financial resources, often deployed to exert influence in policy making (Baker, 2010). Financial actors, especially the main transnational players, are often well-organized lobbyists active at the national and international levels. Prior to the financial crisis, the augmented complexity of governance in this area had led to an increased reliance on the private sector for the provision of information and expertise, the issuing of private (soft) regulation, and the self-monitoring and self-enforcement of agreed norms, at least for some financial activities or entities, such as hedge funds and rating agencies (Porter, 2005). The crisis partly reversed this trend (at the least in the short and medium term), as explained in the penultimate section.

Third, there is a disjuncture between globalized financial markets and national systems of supervision (Padoa-Schioppa, 2004). This fragmentation of powers and responsibilities severely constrains the public authorities' ability to regulate and supervise financial markets and financial entities active in their jurisdictions. A specific problem is posed by ailing cross-border financial institutions. To begin with, there is the distribution of burden sharing (i.e., the costs of rescue plans) between the home and host countries and across the several host jurisdictions in which cross-border financial institutions operate. Furthermore, sometimes financial institutions are too big to be rescued by their home countries (ultimately by the treasury and taxpayers of the country in which they are headquartered).

The fourth problematic issue in financial regulation concerns the separation of central banking functions from supervisory functions (e.g., Goodhart and Schoenmaker, 1995; Oosterloo and de Haan, 2004). This in turn raises the issue of which authorities should be responsible for macrostability (i.e., stability of the system), microstability (solvency of individual financial entities), and consumers protection. It also feeds into the ongoing debate concerning the 'best' model for financial supervision, whereby the main choice is between the establishment of single financial services authorities responsible for the oversight of all financial services and the allocation of supervisory functions to a variety of sectoral authorities. Worldwide there are different national models in place concerning the institutional arrangements for financial supervision (Masciandaro and Quintyn, 2007; Masciandaro, 2005). Several countries had reformed their domestic frameworks for financial supervision in the decade prior to the crisis, and changed it again in light of the crisis.

Fifth, the global financial crisis brought into sharp relief the interconnection between monetary policy, macroeconomic imbalances, and financial stability, in particular the effects that the former can have on the latter, as explained in the penultimate section.

Perspectives on Financial Regulation in the Academic Literature

The literature on financial regulation has been developed from a variety of disciplinary perspectives: law, economics, and political science. Of particular interest is the latter, which has focused on the public policy aspects of financial regulation. Broadly, three or four main bodies of scholarly works can be identified, which tend to focus their analysis at different levels of governance.

First, there is the international political economy literature on international financial regulatory harmonization and cooperation. The assumption is that international regulatory harmonization and cooperation benefits some countries more than others, affecting the competitiveness of national firms. Consequently, international harmonization and cooperation will take place only when they are in line with the preferences of the main jurisdictions (Drezner, 2008), which in the past was the United States because of the size of its domestic market. Recently, international political economy scholars have considered the European Union as an increasingly influential actor in the making of rules for global finance because of the sheer size of its internal market. In this literature, national policy makers' preferences are shaped by concerns for the competitiveness of national banks in international markets, and by the need to ensure financial stability domestically (Singer, 2007). According to this 'regulatory arbitrage' explanation, national policy makers seek to set in place at the international level rules that would promote financial stability, without creating serious competitive disadvantages to their national financial industry.

A second stream of literature focuses on the pervasive power of banks, which are influential lobbyists with plenty of financial, human, and technical resources at their disposal (Baker, 2010; Underhill et al., 2010). Hence, big banks are those most likely to yield influence in the making of financial regulation: they can lobby directly at the international level, and/or their national policy makers. The 'bank power' explanation argues that the most influential players whose preferences prevail in shaping international and domestic financial rules are the main stakeholders in the financial industry. It links to the economics literature on regulatory capture, which posits that financial regulators are captured by the industry they are supposed to regulate.

A third stream of literature argues that financial governance arrangements fall within the responsibility of a transnational policy community of experts actors (Kapstein, 1992; Porter, 2005). These actors have common professional and educational backgrounds and share similar epistemological views. The argument is that small groupings of experts facilitate information exchange and processes of deliberation. In this context, actors develop homogenous preferences and share common views on what problems need solutions and with what instruments. Over the 1980s and 1990s, a dominating paradigm for financial services regulation emerged internationally and in the main jurisdictions, including Europe. It has been referred as the 'neoliberal' paradigm (Gamble, 2009) or as a 'governance light' approach, which favored 'market-based governance,' based on a benign view of financial markets and grounded in efficient market theories (Underhill et al., 2010: 10). This regulatory paradigm was sponsored by the United States and the United Kingdom, which prior to the crisis exercised considerable influence in international regulatory fora.

The literature on varieties of capitalism (for a review, see Jackson and Deeg, 2008) has examined the main components of varieties of capitalism, namely, industrial relations

institutions; education and training systems; corporate governance and systems of corporate financing; product markets; social protection systems and welfare states; and public intervention in the economy. The financial system has also been considered a component of the variety of capitalism, but attention has centered on sources of corporate finance (Zysman, 1983; Deeg, 2010). A literature on national varieties of financial capitalism as such is largely missing. However, some works rooted in comparative political economy and inspired by historical institutionalism have examined how the configuration of national financial system (Allen and Gale, 2000; Rajan and Zingales, 2003) and domestic regulatory institutions affect the preferences of policy makers on financial regulation domestically and internationally (Fioretos, 2010; Posner, 2010).

Finally, there is a set of works that focus on national supervisory arrangements, analyzing different models of supervision with a view to assess which works best (Masciandaro and Quintyn, 2007; Masciandaro, 2005; Oosterloo and de Haan, 2004). Some economists and political economists who in the past had worked on indicators of central bank independence and its relation to national macroeconomic performances, first and foremost, inflation, have extended this kind of analysis to financial supervisory authorities. However, unlike in the case of central banks, there seems to be no 'winning' or 'best' model of financial supervision. Other economists have tried to identify to measure the stringency (or otherwise) of financial regulation across countries with a view to determine what type of regulation works best, using various indicators (Barth et al., 2006).

The Main Causes of the Global Financial Crisis

The financial industry expanded substantially over the last decades. In 2011, banking assets in the United States were $12 441 (data from the US Federal Reserve), in the European Union they totaled €35 901, distributed as follows: Germany €7995, France €6674, and the United Kingdom €11 142 (data from the European Central Bank). Securities trading including derivatives in the United States was $849 791 billion, in the European Union it was €655 259 billion (data from the World Federation of Exchanges). Securities trading, excluding derivatives, in the United States was $39 225 billion, in the European Union was €35 701 billion, distributed as follows: Germany €1302 billion, France €928 billion, and the United Kingdom €7711 billion (data from the ECB). Insurance premia in the United States totaled $2154 billion, in the European Union they were €1072 billion, distributed as follows: Germany €178 billion, France €190 billion, and the United Kingdom €206 billion.

The global financial crisis that began in late 2007 and reached its peak in the autumn of 2008 had three main interrelated causes, namely, a relatively loose monetary policy, international macroeconomic imbalances, and financial regulation that was not fit for purpose. The building up to the crisis took place in an international context of relatively loose monetary policy, especially in the United States and Europe, where central banks mainly focused on keeping inflation under control (which they did) without paying sufficient attention to other economic indicators, such as external and internal imbalances (such as property bubble or consumers indebtedness). It was a period of 'irrational exuberance.'

International macroeconomic imbalances were caused by increasing surpluses in the balance of payments of many Asian economies and the persistent deficit of the US balance of payments, compensated by capital inflows into the United States from surplus economies. Substantial macroeconomic imbalances existed also within the European Union. Capital inflows from Asian economies and from surplus to deficit countries in the European Union were mainly intermediated by the banking system. They contributed to the availability of credit, which fueled, for example, real estate bubbles and facilitated a loose monetary policy.

Financial regulation created the wrong incentive structures and promoted moral hazard. It facilitated excessive leverage by banks and the growth of a 'shadow' unregulated banking sector linked to regulated banks. Other activities, such as rating agencies, hedge funds, and over-the-counter derivatives were mainly subject to industry self-regulation that was not fit for purpose. For example, credit rating agencies were far too generous in their ratings of structured financial products.

The crisis started off as an 'old fashioned bubble' in the United States and spread quickly worldwide, first and foremost to the United Kingdom, whose financial industry is highly interconnected to that of the United States. Subsequently, the financial crisis hit the real economy because banks began to deleverage, cutting credit to industry and services, with direct negative effects on economic growth and employment. Moreover, in several countries, such as the United Kingdom and Germany, the state had to intervene to recapitalize banks. In some cases, such as Ireland and Spain, the huge losses incurred by banks and taken over by the state threatened the very solvency of the state, triggering a sovereign debt crisis. In other countries, such as Greece and Portugal, the sovereign debt crisis began as a traditional fiscal crisis, due to persistent fiscal imbalances between revenues and expenses and, consequently, the building of the public debt.

All these countries at the periphery of Europe were also affected by external and internal imbalances, and eventually had to resort to external financial support from the European Union in some cases from the International Monetary Fund to stave off the sovereign debt crisis. In turn, the sovereign debt crisis risked turning into a banking crisis, showing the interconnection between the two types of crises in certain countries, such as Greece and Italy, whose (rather 'traditional') banking system had coped rather well with the global financial crisis of 2007–08. However, once the Greek and Italian states faced default, their banking systems were penalized and the sovereign debt crisis risked turning into a banking crisis.

The Regulatory Response to the Global Financial Crisis

In the aftermath of the crisis, several jurisdictions revised their framework for financial regulation and supervision. Although the measures adopted are specific to each jurisdiction, three general trends can be detected. First, the reforms enacted should be seen as incremental changes tinkering with the

existing regulatory framework, rather than path-breaking reforms changing the fundamentals of the system. For example, the main regulatory measures that focused on banks, the so-called Basel III Accord (2010), negotiated by the Basel Committee on Banking Supervision and eventually endorsed by the G20, build on an existing set of rules, the Basel II Accord, rather than reinventing them altogether. These international standards are in the process of being implemented domestically in the jurisdictions that signed the accord.

Second, the vast majority of the measures adopted regulated activities or financial institutions that were previously not subject to public regulation, such as over-the-counter derivatives, or were subject to industry self-regulation. In other instances, they imposed heavier, more prescriptive, and more burdensome requirements on financial entities that were already regulated prior to the crisis, as in the case of higher capital requirements and new liquidity management rules for banks. For example, the European Union issued new (legally binding) rules on rating agencies, alternative investment funds managers, over-the-counter derivatives, etc. Analog measures were taken in the United States with the Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010, which increased the power of financial regulatory agencies, reduced regulatory gaps, and consolidated the regulation of systemically important institutions.

Third, the global financial crisis drew attention to macroprudential supervision (also referred to as systemic supervision) designed to safeguard the stability of the entire financial system and the interconnection between micro- and macrofinancial supervision. Prior to the crisis, the attention had very much focused on microprudential supervision designed to maintain the solvency of individual financial institutions. The global financial crisis demonstrated that this narrow focus was not enough to ensure financial stability. It also emphasized the need to adequately regulate and supervise all systemic institutions, that is, the financial institutions whose failure might endanger the stability of the system. In the aftermath of the crisis, several jurisdictions set up new supervisory bodies to deal with microprudential risk. For example, the European Union established the European Systemic Risk Board, its chair to be elected by and from the members of the General Council of the European Central Bank and in charge of monitoring macroprudential risk.

Finally, prior to the crisis, financial regulation was mainly a 'technical' policy area in which there was a very limited involvement of politicians and that was of marginal interest to the wider public. It was, however, an arena where the interests of the financial industry played out. After the crisis, heads of state and government, such as Barack Obama, Nicolas Sarkozy, and Angela Merkel, became interested in financial regulation, at times adopting populist stances to appease public opinion.

Conclusion

Financial regulation has become a matter of great public interest after the financial crisis, though it has always been a key (but rather 'technical') public policy. It pursues a variety of objectives, the most important of which is to preserve financial stability. The global financial crisis highlighted some problematic issue, which have only partly been addressed by the regulatory reforms that followed the crisis. Among the main open issues, there are the disjuncture between globalized markets and national supervision; the problem of ailing cross-border financial institutions or of financial institutions too big to fail; and the link between the financial crisis and the sovereign debt crisis, which have been left unaddressed, both internationally and at the national level. Financial regulation is therefore not well equipped to prevent or deal with the next financial crisis.

See also: Financial Crises, History of; Regulation and Administration; Regulation, Economic Theory of; Regulation, Political Economy of.

Bibliography

Allen, F., Gale, D., 2000. Comparing Financial Systems. MIT, Cambridge.
Baker, A., 2010. Restraining regulatory capture? Anglo-America, crisis politics and trajectories of change in global financial governance. International Affairs 86 (3), 647–663.
Barth, J.R., Caprio Jr., G., Levine, R., 2006. Rethinking Bank Regulation: Till Angels Govern. Cambridge University Press, New York.
Borio, C., 2011. Implementing the macroeconomic approach to financial regulation and supervision. In: Green, C.J., Pentecost, E.J., Weyman-Jones, T.G. (Eds.), The Financial Crisis and the Regulation of Finance. Edward Elgar, Cheltenham.
Brunnermeier, M., Crockett, A., Goodhart, C., Persaud, A.D., Shin, H.S., 2009. The fundamental principles of financial regulation. Geneva Reports on the World Economy 11.
Davies, H., Green, D., 2008. Global Financial Regulation: The Essential Guide. Polity, Cambridge.
Deeg, R., 2010. Institutional change in financial systems. In: Morgan, G., et al. (Eds.), The Oxford Handbook of Comparative Institutional Analysis. Oxford University Press, Oxford, pp. 312–351.
Drezner, D.W., 2008. All Politics Is Global: Explaining International Regulatory Regimes. Princeton University Press, Princeton.
Fioretos, O., 2010. Capitalist diversity and the international regulation of hedge funds. Review of International Political Economy 17 (4), 696–723.
Gamble, A., 2009. The Spectre at the Feast: Capitalist Crisis and the Politics of Recession. Palgrave, Basingstoke.
Goodhart, C., 1999. Myths about the lender of last resort. Journal of International Finance 2, 339–360.
Goodhart, C., Schoenmaker, D., 1995. Should the functions of monetary policy and banking supervision be separated? Oxford Economic Papers 47, 539–560.
Hardie, I., Howarth, D., 2013. Market-Based Banking and the International Financial Crisis. Oxford University Press, Oxford.
Herring, R., Litan, R.E., 1994. Financial Regulation in the Global Economy. Brookings Institution.
Jackson, G., Deeg, R., 2008. How Many varieties of capitalism? From institutional diversity to the politics of change. Review of International Political Economy 15 (4), 679–708.
Kapstein, E., 1992. Between power and purpose: central bankers and the politics of international regulation. International Organization 46 (1), 265–287.
Llewellyn, D., 1999. The Economic Rationale for Financial Regulation. FSA Occasional Papers in Financial Regulation, 1.
Masciandaro, D. (Ed.), 2005. Handbook of Central Banking and Financial Authorities in Europe: New Architectures in the Supervision of Financial Markets. Edward Elgar, Cheltenham.
Masciandaro, D., Quintyn, M. (Eds.), 2007. Designing Financial Supervision Institutions. Edward Elgar, Cheltenham.
Oosterloo, S., de Haan, J., 2004. Central banks and financial stability: a survey. Journal of Financial Stability 1 (1), 257–273.
Padoa-Schioppa, T., 2004. Regulating Finance: Balancing Freedom and Risk. Oxford University Press, Oxford.
Porter, T., 2005. Globalization and Finance. Polity, Cambridge.
Posner, E., 2010. Sequence as explanation: the international politics of accounting standards. Review of International Political Economy 14 (4), 639–664.

Rajan, G., Zingales, L., 2003. The great reversals: the politics of financial development in the twentieth century. Journal of Financial Economics 69 (1), 5–50.
Schinasi, G.J., 2006. Safeguarding Financial Stability: Theory and Practice. IMF, Washington.
Simmons, B., 2001. The international politics of harmonization: the case of capital market regulation. International Organization 55, 589–620.
Singer, D.A., 2007. Regulating Capital: Setting Standards for the International Financial System. Cornell University Press, Ithaca.
Underhill, G.R.D., Blom, J., Mügge, D. (Eds.), 2010. Global Financial Integration Thirty Years on. Cambridge University, Cambridge.
Zysman, J., 1983. Governments, Markets and Growth. Cornell University Press, Ithaca.

Relevant Websites

http://www.bankofengland.co.uk/Pages/home.aspx – Bank of England.
www.bis.org/ – Bank for International Settlements.
http://www.bis.org/bcbs/ – Basel Committee on Banking Supervision.
http://ec.europa.eu/internal_market/top_layer/financial_capital/index_en.htm – European Commission, Financial Services.
http://www.federalreserve.gov/ – Federal Reserve.
http://www.financialstabilityboard.org/ – Financial Stability Board.
http://www.group30.org/ – Group of Thirty G30.
http://www.g20.org/ – Group of Twenty G20.
http://www.ifrs.org – International Accounting Standards Board (IASB).
www.iaisweb.org – International Association of Insurance Supervisors (IAIS).
www.iosco.org – International Organization of Securities Commissions (IOSCO).
www.hm-treasury.gov.uk – UK Treasury.
http://www.treasury.gov – US Treasury.

Fire, Human Use, and Consequences

Johan Goudsblom, University of Amsterdam, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The ability to handle fire and exploit fuel is, like language and the use of tools, a universal human attainment: no human society in historical times is known to have lived without it. In addition, the ability to handle fire and exploit fuel is also exclusively human; while other animals have developed rudimentary signals and tools, only humans have learned to handle fire. Today fire is continuously present in all human societies. It is used in many different guises, some highly visible, others largely hidden from public view and consciousness. Access to, and control over, fire and fuel have played an important part in human history and social evolution, more important than is generally acknowledged in the literature of the social sciences.

Introduction

The ability to handle fire and exploit fuel is, like language and the use of tools, a universal human attainment: no human society in historical times is known to have lived without it. In addition, the ability to handle fire and exploit fuel is also exclusively human; while other animals have developed rudimentary signals and tools, only humans have learned to handle fire. Today fire is continuously present in all human societies. It is used in many different guises, some highly visible, others largely hidden from public view and consciousness. Access to, and control over fire, and fuel have played an important part in human history and social evolution, more important than is generally acknowledged in the literature of the social sciences.

The Original Domestication of Fire

Fire is, according to dictionary definitions, a process of combustion, manifested by heat and light. It occurs relatively rarely in nature; when it does, it does not last long, with a few exceptions such as the burning coal-fields in Northern China, which have been alight for many centuries. Most natural fires have rapid and drastic effects. They cause a quick disintegration of organic matter, reducing it to ash and smoke in a process of transformation that is almost the prototype of 'destruction.'

During the Paleolithic, human groups have learned to turn this destructive and essentially aimless natural force into a regularly available asset that could be made useful for a variety of 'productive' purposes. When and how exactly this transition occurred is still unknown (cf Stringer, 2011). It stands to reason to assume that the actual 'domestication' of fire was preceded by a prolonged phase of 'association' with fire (Burton, 2009), when groups of hominins gradually familiarized themselves with the presence of fire and learned the trick of keeping it burning by 'feeding' it with fuel. The domestication of fire was not a single event, but a series of incidents, endeavors, failures, and occasional successes, resulting eventually in the unique bond of humans and fire.

The key toward understanding the importance of this bond is fuel. All animals derive practically all of their energy from one single source – food, either in the form of vegetation or, in the case of predators, of other animals. Humans alone have added fuel to food as a second major source of energy. Originally, of course, fuel is also derived from plant vegetation; but whereas food is almost always consumed while it is still 'fresh,' in the sense of being still alive or almost alive, the matter that is used as fuel is usually dead, and has been dead for some time. It can serve as a source of energy because it still contains vegetable matter and thus preserves solar energy, which is as it were locked up in the dead organic matter and can be released by burning.

Of course, none of these ideas about the transformations of energy were known to our early ancestors. Still, the very process of domesticating fire is likely to have stimulated the further development of their general capacity for culture – for learning new forms of behavior, sharing those with others, and transmitting them to following generations. Adjusting their behavior in such a way that a fire could be kept burning continuously, they generated the uniquely human 'fire regime,' which enabled them to expand their control over their natural environment, while at the same time affecting their own social arrangements and their personal habits and feelings (cf Goudsblom, 1992: pp. 12–41). Fire thus became a part of the 'anthroposphere'; and most of the fires burning on the earth today are 'anthropogenetic' (De Vries and Goudsblom, 2002).

The uses of fire were diverse, ranging from the enjoyment of heat and light to clearing land for hunting and gathering and to cooking (Wrangham, 2009). Their overall tendency was to make life in human groups more secure and comfortable but, by implication, also more vulnerable and dangerous. Fire increased people's productive as well as their destructive potential. Its control thus reinforced a basic trend in human history and social evolution: *the increasing differentiation in behavior and power between humans and all biologically related species*.

Fire and Agrarianization

Toward the end of the Paleolithic, between 15 000 and 10 000 years ago, some human groups began to extend their care and control over selected plants and animals. This, the rise and spread of agrarian production, may be called 'agrarianization,' just as 'industrialization' is used to refer the subsequent rise and spread of industrial production in the modern era.

Agrarianization was not, as is sometimes assumed, the first, but the second great ecological transformation brought about by humans. That transformation could not have taken place without the firmly established control over fire. Once it was under way, it in turn deeply affected the further development of the part played by fire in human society.

The domestication of fire was a precondition for agrarianization first of all, because in most areas using fire ('slash and burn') was the most efficient technique of opening up fields for cultivation by removing the existing vegetation. Moreover, most of the crops selected for cultivation needed cooking before consumption as a staple food. As a further, less tangible precondition, the solidly established 'species monopoly' over fire gave humans a dominance over other animals, which enabled them not only to bring certain species under direct control but also helped to keep most of the remaining 'wild' animals at a distance from their crops and herds. Finally, the long-standing familiarity with a fire regime probably helped to prepare people for the strains of an agrarian regime demanding self-imposed 'deferred gratification.'

With agriculture and livestock, fire was no longer the only nonhuman source of energy under human control. Gradually it ceased to be the prominent focus of group life that it had been for many thousands of generations, becoming increasingly more dispersed over various sorts of specific 'containers,' such as hearths, ovens, and lamps, beyond which it was not allowed to spread. Its use was subjected to stricter regulation. Whereas during the initial stages of domestication, the main concern had been to keep the one fire burning, the main concern now became to prevent the many fires from causing a conflagration.

As the more advanced agrarian societies grew more populous, the number of fires in use increased proportionally. Those fires were increasingly concentrated within the 'fire-protected zones' of towns and villages, in special containers. The uses of fire became more specialized, with various crafts and professions such as smiths and potters developing their own pyrotechnic skills. With proliferation, concentration, and specialization came new forms of organization, if only to meet the growing need for fuel. Unavoidably, the use of fire also reflected the process of stratification, as some members of privileged groups came to command huge supplies of fuel, and could use fire for impressive displays of power, whereas many poor people might never call any fire their own.

A common problem in all advanced agrarian societies was the prevention of uncontrolled fire in cities. For the growing urban populations the proliferation of fires increased the risks of blazes, while the accumulation of property meant that people had more to lose. Of course, it was fire as a natural force with its capricious and self-generating destructiveness that was feared. In case of a military raid people had to reckon with the common practice of 'sack and burn.' Sometimes a rural region might be caught in a vicious circle of mutual arson (cf Frierson, 2002). But peasants and farmers ordinarily encountered fire only in its domesticated settings, where the greatest danger lay in carelessness and where they had to rely on their own and their neighbors' incessant caution.

In preindustrial towns, there was little one could do once a fire broke out. Usually the only way of stopping it was by creating a 'fire break' and tearing down the buildings toward which the fire was advancing; deliberate destruction by human hands had to put an end to the blind destruction by fire. By far the most effective way of restricting the risk of conflagrations was then, as it still is, prevention (cf Bankoff et al., 2012).

Today we are perhaps inclined to conceive of problems of fire prevention primarily in terms of technology: building materials and technical equipment. However, the problems have always been at least as much civilizational. The crucial issue was human behavior in its social context of informal and formal controls; of habits and regulations.

Little is known about the organization of fire brigades in preindustrial cities. The available evidence suggests that their effectiveness, and their very presence or absence, reflected important aspects of the social structure. In ancient Rome, fire brigades at first operated as private bands, and were later brought under imperial military command. A glimpse of the situation in the Roman provinces is shown in a letter from the emperor Trajan, written in 112 CE, in which he refused permission to the governor of Nicomedia to establish a fire brigade of citizens on the grounds that such an organization could easily "turn into a political club" (Mann, 1986: p. 324). One of the great social innovations in the towns of medieval Europe was that they had civil fire brigades.

Fire in Industrial Society

After the original domestication of fire and the rise of agriculture and animal husbandry, industrialization was the third major ecological transformation set in motion by humans. Once more, natural forces originally outside the human realm were brought under human control. The forces that now began to be exploited were sources of energy that, in their practical applications, were radically new, but as to their origins extremely old. They consisted of large, untapped sources of solar energy stored up in dead organic matter that, in secular geological processes, had been accumulated and transformed into various kinds of fuel. When, around 1750, means were invented to exploit these energy sources with the aid of steam power and other technological innovations, humankind entered, in the words of the environmental historian Rolf Peter Sieferle (2001), a 'subterranean forest' of huge dimensions that, in the course of time, turned out far to exceed the initial expectations. The successful tapping of the enormous supplies of coal, and later also of oil and gas, had the additional advantage of greatly facilitating human access to a great variety of other mineral resources, from iron to plutonium.

From the start, industrialization had contradictory and seemingly paradoxical effects. Being essentially a fuel revolution, it led to an enormous increase in the number, variety, and intensity of artificial fires. Just as the first stages of agrarianization were marked by extensive and conspicuous use of fire in the clearing of land by slash and burn, early industrialization dotted the landscape with smoke stacks and clouds of steam. In the development of agriculture, the extensive use of fire gradually made way to more intensive forms of tillage with less conspicuous use of fire. Something similar happened in the process of industrialization. As technology developed, fire and fuel became increasingly indispensible to maintain the rising standards of living, while at the same time, as part of the same process, their presence was made less cumbersome. Fire was

kept at bay as much as possible, out of hearing, seeing, or smelling distance.

Consequently, in highly industrialized societies, the fire and smoke involved in industrial production are now mostly, in the image of Norbert Elias (2012), 'hidden behind the scenes' of public as well as private life. Electricity in particular has helped to create the illusion that modern comfort and luxury can be provided without fire and fuel; it is too easily forgotten that most electricity (like applicable nuclear energy) is itself generated by the combustion of fossil fuels. In spite of appearances, industrial as well as agricultural production continue to be highly fuel intensive.

The same is true of transportation. Typical products of modern fuel-intensive industry are cars, with engines designed to be propelled by finely tuned and minutely controllable combustion processes. Cars are made of steel, plastic, and glass – materials produced and processed at great heat. Yet when we get into our vehicles and turn on the electrical ignition to start the engine we are usually hardly aware of using fire and the products of fire. Nor are most of us continuously alert to the fact that by driving we are producing CO_2 emissions that may be contributing to global warming.

The fuel-intensive character of modern technology is matched by advanced standards of fire prevention. Every branch of industrial production, from the building trade to the manufacture of clothes and furniture, is subject to strict measures with respect to fire. In such branches as the aircraft or the hotel industry, regulation has become international. Fire prevention is thus contributing to worldwide standardization in architecture and building.

The continuing growth in the sheer quantity of available energy has led to an enormous increase not only in productivity but also in the social potential to destroy, in 'destructivity.' Just as the means of preventing and fighting fires became more effective, so did the means of starting fires. In the agrarian era, walled cities were relatively safe havens of protection from military attacks. In the twentieth century, by contrast, cities have not only increased in numbers and size, they have also become easy targets for military arson from the air. The bombings of World War II caused some of the largest urban fires in history, as in Hamburg, Dresden, and Tokyo; even after the droppings of the atomic bombs on Hiroshima and Nagasaki, most casualties and the worst material damage were caused by the subsequent firestorms (cf Eden, 2004).

The comforts and discomforts of fire are distributed over the world population in a way that directly reflects the stratification in terms of wealth and power. In large parts of Asia and Africa there is a pressing shortage of firewood; by cutting down trees for fuel, poor country folk play a part in worldwide deforestation. At the same time, in many regions, from Indonesia to Latin America, large tracts of forest are burnt down in order to clear land for growing crops that can be used as food or, ironically, fuel. These contradictory practices have the similar effect of furthering deforestation (cf Pyne, 2001).

Worldwide inequality is also reflected in the uneven exposure to blazes. An increasingly large portion of the world population is now living in slums in the mega-cities of 'the developing world.' In densely crowded areas that are almost inaccessible to fire-fighting vehicles, the combination of inflammable dwellings and dependence upon open fires for heat and cooking serves as "a superlative recipe for spontaneous combustion" (Davis, 2006: p. 127). Fires break out so often that they only reach the international press if the number of casualties is written in at least three digits.

Fossil fuels rank today among the world's most highly valued natural resources and economic commodities. They form a major item in international trade, both in actual material exchange and in the virtual world of speculative 'international markets.' Some of the main loci of extraction are among the world's most violently contested areas, where large international companies are tense relations with each other as well as with the local population and the states that claim sovereignty over the territories from where the fuels are extracted. Some of the world's most volatile hot spots are in areas rich in fossil fuels, first of all the Middle East, but also such countries as Libya, Nigeria, Venezuela, and Azerbaidzhan. Pipelines and shipping routes to and from these centers are equally subject to geopolitical pressures. All this turmoil is accompanied by rising concerns over ecological degradation and problems of sustainability (cf Yergin, 2011).

Symbolic and Religious Aspects

From all over the world, myths have been collected relating how humanity first came to master fire, often in a feat of cunning and deceit (Frazer, 1930). Those myths convey a sense of the preciousness of fire, which is represented as a unique possession, separating humanity from the animal world. Many people used to tell that neighboring groups living further away in the bush did not have fire and were unable to cook their food; those stories have all been proven slanderous.

Fire is generally believed to have purifying effects. The domain of culture is seen as coinciding with the range of domesticated fire. Thus, cooked meat is regarded as edible for humans, raw meat as only fit for beasts. Similarly, land cleared by burning is trusted as belonging to the human domain, whereas uncleared bush and forest are feared as a dangerous hideout for predators, serpents, insects, and evil spirits.

The sense of preciousness has also found expression in rites. The communal fire would be the center of group life, and people would treat it as sacred, surrounding it with ceremony and taking care that it would never cease burning. The Roman cults of the hearth and of Vesta, the fire goddess, show how such ancient rites could persist in an urban setting (cf Pyne, 1997).

Yet, with urbanization, the sense of the preciousness of fire was often superseded by a general anxiety of the dangers it posed. Thus while Hindu and Zoroaster fire cults kept alive the age-old worship of fire, they were surrounded by ceremonial care. In Christianity and Islam, some old fire traditions persist, as in the burning of candles and incense. Fire no longer plays any part in beliefs about the origin of human society, however. Instead, it is strongly associated with evil. Sinners and unbelievers are threatened with a life after death in which they will be tormented by eternal burning.

Fire festivals, such as those held all over Europe on Saint John's day, when things considered impure would be committed to the flames, were occasions of combined joy and horror. Modern literature and art abound with evocations of

the terrors of fire. Most novelists who have described a blaze, attribute the disaster, quite realistically, not to natural forces such as lightning but to human action – either to personal negligence, madness, or malice, or to collective aggression, in war.

A more dispassionate way of regarding fire, not in terms of its good or bad effects upon human beings but as a natural force in its own right, was first developed by the philosophers of ancient Greece. Out of various antecedents a generally accepted doctrine emerged according to which the universe consisted of four elements: earth, water, air, and fire. In China and India similar systems of thought developed. With some variations, the view in which fire constituted one of the major elements remained the dominant cosmology for natural scientists and physicians well into the modern era.

Alchemists and chemists, as the direct successors to the metallurgists and 'fire masters' of a previous age, made the study of fire into their central concern. Just as modern mechanical industry in its initial stages owed a great deal to the steam engine, the rise of modern physical science would have been inconceivable without the laborious efforts of earlier investigators experimenting with, and theorizing about, fire.

Conclusions

The control over fire has been a human species monopoly since the Paleolithic. Access to fuel constitutes an integral part of the apparatus with which humans in general have established and continue to maintain dominance over other species, while, at the same time, it also serves as an important resource in the struggle for power and privilege among human groups.

Ever since the earliest stages of its domestication, the same sequence of events has recurred again and again in human relations with fire: deliberately sought advances in control were followed by unintended increases in dependency, and as the dependency sank deeper into the social and economic infrastructure, it tended to become increasingly less clearly perceptible. Like agrarianization, industrialization began with conspicuously huge applications of fire. As industrial production became more specialized and highly organized, so did the use of fire. In everyday life in highly industrialized contemporary societies, flames are now visibly present only in such highly domesticated guises as cigarette lighters, candles for ceremonial use, or wood fires lit to create a sphere of comfort and relaxation. When, on the other hand, a fire is shown on television or in the newspapers, it almost always spells war and disaster. The regularly controlled combustion processes upon which industrial production largely rests are mostly relegated away from public view.

In accordance with this apparent eclipse, fire has also ceased to be regarded in the natural sciences and philosophy as one of the elements out of which the world is composed. The very concept of fire has disappeared from scientific discourse, to be replaced by the more abstract concept of 'energy' – something that cannot be directly seen, heard, smelled, or felt. Similarly, many a day may pass in the life of the average citizen of a modern industrial state during which he or she does not witness any physically observable fire. In the social sciences, too, fire has become almost completely unnoticed. It can be concluded from the foregoing that this is unwarranted.

See also: Agricultural Sciences and Technology; Climate, History of; Food Production, Origins of; Land Change: The Merger of Land Cover and Land use Dynamics.

Bibliography

Bankoff, G., Lübken, U., Sand, J. (Eds.), 2012. Flammable Cities: Urban Conflagration and the Making of the Modern World. University of Wisconsin Press, Madison, WI.
Burton, F., 2009. Fire: The Spark that Ignited Human Evolution. University of New Mexico Press, Albuquerque, NM.
Davis, M., 2006. Planet of Slums. Verso, London.
De Vries, B., Goudsblom, J. (Eds.), 2002. Mappae Mundi: Humans and Their Habitats in a Long-Term Socio-Ecological Perspective. Amsterdam University Press, Amsterdam.
Eden, L., 2004. Whole World on Fire: Organizations, Knowledge, and Nuclear Weapons Devastation. Cornell University Press, Ithaca.
Elias, N., 2012. On the Process of Civilization: Sociogenetic and Psychogenetic Investigations. In: Collected Works, vol. 3. University College Dublin Press, Dublin.
Frazer, J.G., 1930. Myths of the Origin of Fire. Macmillan, London.
Frierson, C.A., 2002. All Russia Is Burning! a Cultural History of Fire and Arson in Imperial Russia. University of Washington Press, Seattle, WA.
Goudsblom, J., 1992. Fire and Civilization. Allen Lane, London.
Mann, M., 1986. The Sources of Social Power, vol. 1. Cambridge University Press, Cambridge.
Pyne, S.J., 1997. Vestal Fire: An Environmental History, Told through Fire, of Europe and Europe's Encounter with the World. University of Washington Press, Seattle, WA.
Pyne, S.J., 2001. Fire: A Brief History. University of Washington Press, Seattle, WA.
Sieferle, R.P., 2001. The Subterranean Forest: Energy Systems and the Industrial Revolution. The White Horse Press, Cambridge.
Stringer, C., 2011. The Origin of Our Species. Allen Lane, London.
Wrangham, R., 2009. Catching Fire: How Cooking Made Us Human. Basic Books, New York.
Yergin, D., 2011. The Quest: Energy, Security, and the Remaking of the Modern World. The Penguin Press, New York.

First Language Acquisition, Developmental Psychology of

RA Berman, Tel Aviv University, Israel

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 8, pp. 5681–5686, © 2011, Elsevier Ltd., with revisions made by the Editor.

Abstract

In this article, motivations for cross-linguistic comparisons of how children acquire their native language are noted, and major approaches to the field are reviewed, from traditional diary-studies of bilingual children, via the Anglocentric orientation of early developmental psycholinguistics, to extensive data collection and research in a wide range of languages and language-types over the last two decades of the twentieth century. Claims for universal underpinnings of language acquisition are discussed, including Chomskyan-based nativist claims for formal principles of Universal Grammar (UG) on the one hand, and psychologically motivated claims for extralinguistic perceptual and cognitive constraints shared by children acquiring different languages, on the other. In production of language, children across the world proceed from prelinguistic babbling to one-word utterances and on to early word combinations, grammatical clause structure, and clause combining. Such shared trends are compared with current research which emphasizes the marked impact of caretaker input and the effect of target-language typology on very young children in a range of domains, including early speech perception and babbling; the noun vs. verb bias in children's early lexicons; spatial distinctions; null subjects; word formation strategies; expression of reference, temporality, and locative motion in narratives; and differential deployment of structural devices such as passive voice for expressing impersonal perspectives in discourse. The article concludes by noting the need for further research designed for comparability between both those languages which are typologically closely related and those which differ in critical respects.

In this article, motivations for cross-linguistic research into children's acquisition of their first language(s) are noted with a brief historical outline of the field (Section Approaches to the Field). This is followed by a review of trends which appear to be universally shared across children acquiring different languages (Section Universal Trends in Language Development), compared with the impact of typological properties of particular target languages on processes of acquisition (Section Impact of Target-language Properties). The article concludes with a note on desiderata (Section Conclusion).

Approaches to the Field

The major motivation for the cross-linguistic study of language acquisition, as of cross-linguistic research in general, is to compare what is universally shared across children irrespective of the specific target language or speech community to which they belong, on the one hand, with interlanguage variation and the impact of properties peculiar to a particular language or language type on the other (Slobin, 1997b). A related, psycholinguistically motivated question is how the specifics of target-language typology interact with acquisitional principles and strategies deriving from more general extralinguistic factors of perception, cognition, and social interaction (Slobin, 1985b).

The first studies of children's acquisition of their native language(s) which can be characterized as cross-linguistic in orientation date from the nineteenth and early twentieth centuries. These were generally diary studies of childhood bilingualism, among children raised in more than one language, mainly in Europe. Such studies provided potentially rich data for cross-linguistic comparisons, as did a range of studies of children acquiring languages other than English (Slobin, 2014). However, these were by and large not systematically directed to the testing of psycholinguistically motivated hypotheses, and they tended to focus on phonology, inflectional morphology, and the lexicon, with little concern for syntax or semantics.

The rise of developmental psycholinguistics in the USA in the 1960s created a field of inquiry which aimed to integrate insights from linguistic theory and description with developmental patternings in first language acquisition. Such studies tended to be Anglocentric, but provided a basis for research in other languages. For example, the general developmental stages delineated in Brown (1973) for the early years of grammatical development appear to apply across languages (see Section Universal Trends in Language Development below).

This Anglocentricism was countered by the Berkeley project on four typologically distinct languages (Slobin, 1982). Since then, data have been collected from a wide range of languages, both across and outside Europe. These include surveys of several dozen typologically distinct languages in the five volumes edited by Slobin (1985–1997). Other collections focus on particular domains including: early clause structure (Friedmann and Novogrodsky, 2004), the lexicon–syntax interface in verb acquisition (Lieven, 1998), and narrative construction (Berman and Slobin, 1994; Berman and Slobin, 1994). Generative grammars gave rise to a parameter-based analysis of acquisition in different languages and domains, for example, Hyams (1986) and Weissenborn (1992) on null subjects, Meisel (1992) on functional categories and the ordering of constituents (*see* Generative Grammar).

Research in child language since the 1980s has thus yielded a large body of data on acquisition of numerous languages, including ones with non-European typology. Many of the latter studies are anchored in ethnographic-type fieldwork, for example, for Inuktitut by Shanley Allen, for Mayan Quiché by

Clifton Pye, and for Sesotho by Katherine Demuth; or in research conducted by native speakers, as in Soonja Choi's work on Korean. This work combines to demonstrate that (1) children early on adapt to the particular constraints of their native-language typology, as verb-initial or verb-final, accusative or ergative, as synthetic or agglutinating; (2) they do so on the basis of the particular cultural norms for caretaker–child interaction in their speech community; (3) they do so with equal rapidity and ease across languages; and (4) within languages, the acquisition of certain subsystems may be delayed for language-specific reasons of relative structural and/or semantic complexity, on the one hand, and because they belong to restricted registers of usage not accessible to young children, on the other.

Universal Trends in Language Development

Different points of departure underlie claims about universals in language acquisition. A nativist, Chomskyan-based view will specify these in terms of the formal principles characterizing a generative model of Universal Grammar (UG) with cross-linguistic variation accounted for by differential setting of structural parameters according to the target language. A Piagetian orientation will explain them in terms of general cognitive principles and shared conceptual underpinnings which provide a universal epistemological basis for acquisition of linguistic knowledge (Slobin, 1985). The view espoused by this author is that children start out the task of acquisition equipped with categorical and semantic distinctions of the type termed by Chomsky 'substantive' as against formal universals, akin to Keenan's (1979) idea of 'naive' or primitive universals (Reilly et al., 2002). For example, children do not need to learn that there are two main kinds of sound segments, consonants and vowels; that propositions distinguish between predicates and arguments; or that there are intransitive and transitive clauses, with one-, two-, or three-place predicates.

Whatever the explanation they propose, researchers agree that there is a clear biological underpinning to the overall developmental stages manifested by children in command of language production irrespective of mother tongue. Just as children sit before they stand, and walk before they run, so children across the world babble before they produce single words, rely on single words before they start combining words, and produce isolated clauses before they join these together by increasingly specific lexical means and complex syntactic devices.

Overall progression thus involves an evolution in construction types. Language production typically starts out with utterances, as behavioral elements which provide the initial basis for expressing relations between semantic content and phonetic form. These typically take the form of isolated lexical items or unanalyzed strings of words, starting around the end of the first year of life. Structure dependence emerges when these lexical units are combined in simple-clause structure, realized by morphosyntactic markings of verb–argument relations. Next, adjacent clauses are combined, first by coordination, subsequently by various kinds of subordinate constructions – initially mainly finite and subsequently nonfinite – leading up to complex clause-linkage by derived nominalizations (Berman, 1993). This needs to be qualified for a verb-final language like Turkish, where the unmarked means of clause linkage is by nonfinite, nominalized constructions. These emerge relatively early in Turkish compared with European or Semitic languages.

With respect to acquisition of morphology, the domain in which languages tend to differ most, children generally acquire the bulk of the inflectional system of their language before they master its derivational morphology. This is understandable since inflection is more regular, productive, and grammatically obligatory than derivation. Besides, derivational processes depend on a level of vocabulary expansion not achieved by young preschool age. In fact, in some instances, like derived nominals or the high-register Latinate lexicon of English, children will master derivational processes only well into school age. In contrast, they by and large command the inflectional systems of their language by the age of three, irrespective of how morphologically opaque or synthetic these happen to be. Cross-linguistic differences affect acquisition of inflection along two main dimensions. Psychological processing principles like perceptual saliency coupled with regularity and transparency of form-meaning mappings account for the fact that, for example, Turkish-speaking children gain command of the bulk of their rich system of agglutinating, inflectional suffixes by a much earlier age than children acquiring languages with more synthetic or syncretic morphologies like Hebrew or Icelandic. And experiential factors of frequency of use in the everyday colloquial speech that constitutes the linguistic input and output of young children explain why certain grammatical categories are acquired relatively late in languages where they constitute high-register or literate constructions.

In the domain of syntax, children typically gain command of simple clause structure in their language between ages 2 to 3 years. Basic word order emerges very early indeed, first between phrasal constituents, only later within phrases. Constituent ordering is followed by grammatical marking of verb–argument relations, around the same time or somewhat before inflectional marking of grammatical agreement. The overall progression is thus much as Brown (1973) defined for English: from single words to early word combinations, thence to simple clause structure, followed by a range of devices for clause linkage.

Another commonality is the general development observed across children from rote-learning of unanalyzed items to partial productivity followed by command of structure-dependent rules and generalizations, which are subsequently constrained by conventions of lexical and discourse usage. This has been demonstrated, for example, for how children acquire the semantics, morphology, and syntax of causative verbs in three typologically distinct languages: English at the more analytic or syntactic end of the scale, Hebrew with relatively synthetic bound morphology, and Inuktitut as a polysynthetic language which marks causativity by both morphological and lexical means (Allen, in Lieven, 1998).

General language development thus proceeds across different phases, from pregrammatical to structure-dependent to appropriate use. And it becomes both more elaborated and varied in structure and use, on the one hand, and more restricted or constrained by target-language specifics, on the other. Thus, in acquiring the sound system, babies proceed

from universal crying to uttering a range of speech sounds, followed by babbling, which is increasingly restricted to elements of the native language to a point where eventually the system will be so constrained as to impose a nonnative pronunciation in other languages. Similarly, in morphology, across languages children proceed from nonalternation between forms to immature strategies for initial markings of grammatical categories, thence to 'creative errors' of overregularization, until eventually the system becomes fully constrained by lexical convention. And in the acquisition of the lexicon, again across languages, children move from a restricted vocabulary of nursery terms and over or underextensions of conventional items to a period of creative innovation; eventually the conventional lexicon takes over, while at the same time vocabulary growth is both constrained and elaborated by socially and culturally appropriate distinctions of register and usage, lasting into adolescence.

Impact of Target-Language Properties

As against these cross-language commonalities and claims for universal constraints, whether formal-linguistic or perceptual-cognitive, current research reveals a marked effect of target-language typology from a very early age. Such more language-specific claims have been explained in cross-linguistic perspective as reflecting a modified, neo-Whorfian characterization of the relation between language and thought (Slobin, 1996) and as demonstrating the critical impact of caretaker input from the very start (McDonough et al., 2003). This very early sensitivity to target-language properties has been observed in a variety of linguistic domains. For example, language-specific effects exert an influence on speech perception and babbling during the first year of life (Jusczyk, 1997). And well before they have command of grammatical inflection, children will adjust their speech output to the prosodic character of their language, as intonation or tone-based (Demuth, 1993), and they will show rapid command of vowel-harmony where required, as in Turkish (Aksu-Koç and Slobin, 1985).

Children's encoding of the semantics of space and which distinctions they mark linguistically (e.g., in shape or location of figure and ground respectively) is also markedly influenced by target-language specifics. Thus, the universality of the early preference among children acquiring European languages (and also Hebrew) for the prepositions *in*, *on*, and then *under*, which was convincingly explained in terms of perceptual properties of space and shared cognitive preferences, has now been queried, since it turns out that quite different factors govern the encoding of spatial distinctions in languages like Korean or Tzeltal compared with English and Dutch (McDonough et al., 2003). Children appear to relate to the categories relevant to their own native language without initially manifesting universal patternings in the expression of spatial distinctions.

Null subjects are another domain in which Universalist claims have been made. Thus, Hyams (1986) suggested that children first set the prodrop parameter to the unmarked value of plus, irrespective of target language, since English-speaking children in the early stage of grammar acquisition produce numerous subjectless sentences, as do their Italian peers. Weissenborn (1992) showed, however, that target language input has a marked influence from very early on in this respect, since children acquiring French produce fewer subjectless sentences than those acquiring German, and the latter do so less than children acquiring strongly prodrop languages like Italian and Spanish.

The very early impact of target language is evident not only in which categories are distinguished, but also how. This is clearly shown in Clark's (1993) cross-linguistic comparisons in the domain of word formation. When young children express the notion of 'undoing' or reversal of actions, they do so from early on in ways peculiar to their native tongue: by the prefixes *un*-in English and French *dé*-, and by particles like *ab* and *aus* in German, *af* and *uit* in Dutch. Hebrew-speaking children, in contrast, must do this lexically or periphrastically, since their language lacks the morphological means for this semantic category. Similarly, in order to express causativity, children will start by using syntactic means with verbs like *make* in English and *faire* in French, a construction type used only rarely by Hebrew-speaking children for this purpose. Subsequently, English-speaking children will overextend verbs like *fall* or *sing* to transitive contexts; in Inuktitut, children will first use morphological and then lexical means of causativity while in Hebrew, they will first neutralize and then alternate morphological patterns to mark verb transitivity (Allen in Lieven, 1998; Reilly et al., 2002).

This research converges to suggest that apparently universal acquisitional principles of structural simplicity and semantic transparency in form-meaning mappings are affected and perhaps even superseded by typologically pervasive properties of the target language. For example, it has been argued that children will find it easier to juxtapose words than to affix bound elements to a stem or root, and that these processes are perceptually more salient and so more accessible to young children than stem-internal vowel alternation. However, in a language like Hebrew, which uses vowel alternation to express basic grammatical distinctions such as verb-tense, children mark past vs. present tense at much the same time as their counterparts acquiring languages which use more linear morphological means. Moreover, although their language allows both noun compounding and external, linear affixation to a stem for new word formation, Hebrew-speaking children typically innovate nouns, verbs, and adjectives by the more typically Semitic device of attaching syllabic affixes to a consonantal root: and they rarely use the option favored by English-speaking children for creating novel verbs by zero derivation or syntactic conversion (Berman, 1999).

Children's acquisition of word classes and the composition of their early lexicon are other areas which suggest that claims for universal underpinnings of acquisition may need to be modified. Prior research on children's lexical repertoire at the one-word stage in several languages showed that more nouns were acquired earlier than verbs, while the vocabulary spurt in the period before early word combinations also consisted mainly of noun like terms. Gentner (1982) explained this in terms of general cognitive principles, since verbs as relational terms are holistic rather than coherently packaged entities, while nouns (at least those in the input and output of one-to two-year-olds) tend to encode semantically 'natural' categories that refer to concrete objects which are perceptually identifiable. Recent research has shown, however, that this noun bias

may be an artifact of the nature of the lexicon and lexical input in the target languages, and that in some cases, verbs may be acquired earlier and in greater quantities than nouns (for example, the studies by Tomasello 2003 and by Choi in Lieven, 1998).

At more advanced stages, developments in form-meaning mappings in extended discourse provide further evidence for the impact of 'typological imperatives.' A five-language cross-linguistic sample of picturebook narratives produced by children aged three to five and nine-year-olds compared with adults (Berman and Slobin, 1994) showed that the expression of textual temporality was critically affected by native-language tense/aspect systems even among our youngest subjects. In Hebrew, children marked only grammatical tense, in Spanish they distinguished both simple from progressive and perfective from imperfective aspect, while in Turkish they gave grammatical expression to modality. Moreover, there was no evidence for 'compensation' in the sense that children attempted to give lexical expression to notions not grammatically encoded in their language (e.g., by using periphrastic adverbs to mark aspectual distinctions in languages like German or Hebrew). On the contrary, not only 'do speakers tend to limit their attention to those semantic distinctions marked in the grammar, they may even overattend to such distinctions in early phases of development' (Berman and Slobin, 1994, p. 619). Slobin (1996, 1985–97) has since extended these observations to the expression of locative trajectories. Proceeding from Talmy's (1985) distinction between verb-framed languages (Hebrew, Spanish, and Turkish in our sample) and satellite-framed languages like English and German that rely heavily on locative particles, he shows that the way young children – like adult speakers of the same language – talk about motion, paths, and directionality depends critically on how these notions are grammatically and lexically encoded in their language. Here, too, the underpinnings of such distinctions are evident very early: at the one-word stage, Germanic-speaking children use particles like *off* or *aus* to express direction of motion in ways encoded by verb-stems by children acquiring French, Spanish, or Hebrew. For example, at the age of 18 months, my Hebrew-English bilingual daughter would say either English *down* or a truncated stem form of the Hebrew verb *la-redet* 'to get/go down,' when she wanted to be taken out of her high chair.

In another cross-linguistic project comparing children's narrative texts produced by children aged from 5 to 15 years on the basis of two picture-series, Hickmann et al. (1996, p. 391) focused on the devices for introducing, maintaining, and shifting reference to characters in a story. They found that the English, French, German, and Chinese 'rely differentially on local vs. global devices to mark newness: postverbal position is obligatory in Chinese …, indefinite determines in the other languages; across languages: newness markings emerge late (age 7), local markings emerge before global, and local and global markings are strongly related.' Their findings point to an important interaction between shared developmental trends and linguistic specificity. Relatedly, in the Berman and Slobin (1994) study, across languages, three-year-olds tended to relate to each picture in isolation, four- and five-year-olds were able to encode linearly expressed local relations between events in terms of local temporal relations of sequentiality and subsequently of causality, school-age children constructed a globally organized hierarchical narrative, and adults tended to adopt individual rhetorical styles with rich interpretation and a personal narrative stance.

Ongoing cross-linguistic research on the production of narrative and expository texts by schoolchildren and adults reaffirms such interaction between generally shared developmental trends and typologically determined expression of semantic distinctions and selection of rhetorical options. Nine-year-olds can all construct well-formed personal experience narratives, but are unable to produce coherent expository texts. Further, only from high-school age will one find consistent, deliberate, and appropriate differentiation between language usage across spoken and written texts, marking highly language-particular register distinctions. In contrast, overall genre distinctions appear similar across languages. For example, narrative texts are couched in past tense and mainly perfective aspect, and they have light subjects (mainly pronominal in English and French or null subjects in Hebrew and Spanish), while expository texts rely more on generic or extended present tense and have more lexical and heavier, more noun-phrase subjects and objects. This, too, interacts with target-language options and preferences. For example, in order to express a detached, impersonal perspective in English, passive is quite widely used even by grade-school children, increasingly by adults; and passive is also quite common in the Icelandic texts. In contrast, passive is used far less in other languages, even though it occurs as a structural option: Swedish subjects rely heavily on the impersonal pronominal *man* corresponding to the very common use of *on* in French; Spanish and Hebrew achieve the same rhetorical effect by means of subjectless or second person impersonal constructions. Similar patterns emerged for differential use of pronominal compared with null subjects and also for relative clauses as devices for syntactic packaging in the picturebook narratives.

Discourse-based cross-linguistic comparisons thus suggest that reliance on certain linguistic forms may depend not only on structural factors of relative morpho-syntactic complexity or productivity, but also on the extent to which a given rhetorical option is favored for specific discourse purposes in current language usage.

Conclusion

In sum, on the basis of shared cognitive and linguistic universals, even very young children are attentive to key typological properties of their language. From an early age, they recognize 'where the action is at' in the input language, not only which categories are formally distinguished, but also how. *When* this type of sensitivity finds expression will depend on general developmental, largely extralinguistic factors. For example, the spatial distinctions studied by Bowerman (1995) and her colleagues will precede command of derivational marking of linguistic subcategories, and these will emerge earlier than rhetorical mastery of linguistic forms in the context of extended narratives. But in each case, the specific ways in

which children encode form-meaning relations accord with how this is done by adult speakers of the same target language rather than by children of the same age in other languages.

As noted in the introduction, a large body of data is currently available for acquisition of different domains in a rich range of languages from Europe and elsewhere. What now seems necessary is to extend systematic and deliberately comparable research designs across a range of languages and linguistic domains – along lines conducted for morpho–syntax by Slobin (1982), for word formation and derivational morphology by the structured elicitation and diary studies reported in Clark (1993), for semantic encoding of space by Bowerman (1995), for temporality by Weist and coworkers (Weist et al., 1993), as well as for narratives by Hickmann et al. (1996) and by Berman and Slobin (1994). Such studies might profitably focus on two contrasting directions of cross-linguistic comparison: (1) languages which are typologically very distinct, for example, ergative or accusative, verb-initial or verb-final, analytic vs. synthetic or agglutinating; and (2) languages which are typologically close, like Dutch and German, or Arabic and Hebrew, but which differ along certain restricted dimensions. Such work should aid in specifying the relative weight of the various factors touched on here. It would help to delineate the interaction between language-universal and language-particular, between intralinguistic, cross-linguistic, and extralinguistic factors in acquisition. And it could provide psycholinguistic motivation for the distinction between language-specific, typologically shared, and cross-linguistic or universal features of language structure and use in developmental perspective.

See also: Language Acquisition; Linguistic Typology; Neural Basis of Language Development; Second Language Acquisition; Vocabulary Acquisition.

Bibliography

Aksu-Koç, A.A., Slobin, D.I., 1985. The acquisition of Turkish. In: Slobin, D.I. (Ed.), The Cross-linguistic Study of Language Acquisition, vol. 1. Erlbaum, Hillsdale, NJ, pp. 839–878.

Berman, R.A., 1993. Crosslinguistic perspectives on native language acquisition. In: Hyltenstam, K., Viberg, A. (Eds.), Progression and Regression in Language. Cambridge University Press, Cambridge, UK, pp. 245–266.

Berman, R.A., 1999. Children's innovative verbs versus nouns. In: Menn, L., Bernstein-Ratner, N. (Eds.), Methods for Studying Language Production. Erlbaum, Mahwah, NJ, pp. 69–93.

Berman, R.A., Slobin, D.I., 1994. Relating Events in Narrative: A Crosslinguistic Developmental Study. Erlbaum, Hillsdale, NJ.

Bowerman, M., 1995. Learning how to structure space for language: a cross-linguistic perspective. In: Bloom, P., Peterson, M., Nadel, L., Garrett, M. (Eds.), Language and Space. MIT Press, Cambridge, MA, pp. 385–486.

Brown, R., 1973. A First Language: The Early Stages. Harvard University Press, Cambridge, MA.

Brown, P., 1998. Children's first verbs in Tzeltal: evidence for an early verb category. In: Lieven, E. (Ed.), Linguistics Special Issue: Developing a Verb-category: Cross-linguistic Perspectives, pp. 713–754.

Clark, E.V., 1993. The Lexicon in Acquisition. UK: Cambridge University Press, Cambridge.

Demuth, K., 1993. Issues in the acquisition of the Sesotho tonal system. Journal of Child Language 20, 275–301.

Friedmann, N., Novogrodsky, R., March 31, 2004. The acquisition of relative clause comprehension in Hebrew: a study of SLI and normal development. Journal of child language 661–681.

Gentner, D., 1982. Why nouns are learned before verbs: linguistic relativity versus natural partitioning. In: A.Kuczaj II, S. (Ed.), Language Development, Language, Thought, and Culture, vol. II. Erlbaum, Hillsdale, NJ, pp. 301–334.

Hickmann, M., Hendriks, H., Roland, F., Liang, J., 1996. The marking of new information in children's narratives: a comparison of English, French, German, and Mandarin Chinese. Journal of Child Language 23, 591–620.

Hyams, N.M., 1986. Language Acquisition and the Theory of Parameters. Reidel, Dordrecht, The Netherlands.

Jusczyk, P.W., 1997. The Discovery of Spoken Language. Bradford Books, Cambridge, MA.

Keenan, E.L., 1979. On surface form and logical form. Studies in Linguistic Sciences 8, 2.

Lieven E. (Ed.), 1998. Linguistics special issue: developing a verb-category. Cross-Linguistic Perspectives 36 (4).

McDonough, L., Choi, S., Mandler, J.M., 2003. Understanding spatial relations: flexible infants, lexical adults. Cognitive psychology 46 (3), 229–259.

Meisel, J.M. (Ed.), 1992. The Acquisition of Verb Placement. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Pye, C., 1994. A crosslinguistic approach to the causative alternation. In: Levy, Y. (Ed.), Other Children, Other Languages: Issues in the Theory of Language Acquisition. Erlbaum, Hillsdale, NJ, pp. 243–264.

Reilly, J.S., et al., 2002. Propositional attitudes: development of modal expression. Written Language and Literacy 5 (2), 183–218.

Slobin, D.I. (Ed.), 2014. The Crosslinguistic Study of Language Acquisition, vol. 4. Psychology Press.

Slobin, D.I., 1982. Universal and particular in the acquisition of language. In: Wanner, E., Gleitman, L.R. (Eds.), Language Acquisition: The State of the Art. Cambridge University Press, Cambridge, UK, pp. 128–170.

Slobin, D.I. (Ed.), 1985–1997. The Crosslinguistic Study of Language Acquisition, vol. 1 (1985a), vol. 2 (1985b), vol. 3(1993), vol. 4 (1997a), vol. 5 (1997b). Erlbaum, Hillsdale, NJ.

Slobin, D.I., 1996. From 'thought and language' to 'thinking for speaking. In: Gumperz, J.J., Levinson, S.C. (Eds.), Rethinking Linguistic Relativity. Cambridge University Press, Cambridge, UK, pp. 70–96.

Talmy, L., 1985. Lexicalization patterns: semantic structure in lexical forms. In: Shopen, T. (Ed.), Language Typology and Syntactic Description, Grammatical Categories and the Lexicon, vol. 3. Cambridge University Press, Cambridge, UK, pp. 36–149.

Weissenborn, J., 1992. Null subjects in early grammars: implications for parameter-setting theory. In: Weissenborn, J., Goodluck, H., Roeper, T. (Eds.), Theoretical Issues in Language Acquisition: Continuity and Change in Development. Erlbaum, Hillsdale, NJ, pp. 269–299.

Weist, R.M., Kaczmark, A., Wysocka, J., 1993. The function of aspectual configurations in the conversational and narrative discourse of Finnish, Polish, and American children. Papers and Studies in Contrastive Linguistics 27, 76–109.

First Language Acquisition, Linguistic Theory of

Rosalind Thornton, ARC Centre of Excellence in Cognition & Its Disorders (CCD), Australian Hearing Hub, Macquarie University, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Chomsky's theory of Universal Grammar (UG) is introduced to explain children's rapid convergence on the grammar of their local language. A system of innate principles and parameters guides children's acquisition of language. Children are shown to misset certain parameters causing them to hypothesize properties not observed in the language of the community. Children's temporary missteps are put aright with input from the surrounding language. The misset parameters are predicted by the theory of UG but not anticipated by usage-based theories of language acquisition.

Introduction

Every parent is fascinated by their child's rapid and seemingly effortless acquisition of their first language. The rapid strides children take in acquiring language seem particularly impressive when considered alongside other cognitive skills such as getting dressed, hopping, counting, and so on – skills that appear to take much more conscious practice. This article examines how children's acquisition of language is accounted for by Chomsky's theory of Universal Grammar (UG). Chomsky's theory of UG is a linguistic theory that posits an innate language faculty and, for this reason, is often referred to as a 'nativist' account (e.g., Chomsky, 1986, 1995). On this account, UG has a role to play in guiding children's rapid path to acquisition of their local language. Using UG as a guide, children have to identify various syntactic and semantic properties of their target language in the linguistic input and incorporate these into their grammar. At the same time, children must learn the lexicon of their language from their caretakers, and map this onto the structures made available by UG.

With UG in place, any typically developing child across the globe is ready to tackle acquiring the language of their community. As we will demonstrate, however, there are times when UG appears to initially steer children to flaunt the linguistic input that surrounds them. It is worth exploring these data, since a path that deviates from the surrounding linguistic input is not anticipated by alternative 'usage-based' accounts of language acquisition. Usage-based accounts of language acquisition expect children to attempt to follow the model provided by adults (Ambridge and Lieven, 2011; Goldberg, 2003, 2006; Tomasello, 2000, 2003). Usage-based linguists place a much heavier burden on the linguistic input, since the child has no advance linguistic knowledge. Not only do children learn the lexicon from the input, they must rely on the input to learn syntactic categories (such as 'noun phrase' or 'verb phrase'), to learn sentence structure, and to learn which structures and meanings for sentences are permissible and which are excluded from the language. On this view, children draw on the general learning mechanisms used to gain other kinds of cognitive knowledge and apply them to the task of acquiring language. We will demonstrate that the UG approach to language acquisition can explain data that is a challenge for usage-based approaches.

Chomsky's Theory of UG

Chomsky has argued that linguistic knowledge is part of our genetic endowment as human beings. The proposal is that children are born with a 'UG,' a domain-specific computational mechanism that is part of the language faculty. This mechanism supports language acquisition to ensure that it takes place in a rapid, effortless, and relatively error-free manner, despite considerable latitude in children's experience. In the following sections, we illustrate the predictions of the theory of UG for cross-linguistic acquisition. We first address the system of parameters that comprise part of the language acquisition device. Parameters are designed to address the rich diversity observed in the languages of the world.

Parameters

According to Baker (2001), "A parameter is simply a choice point in the general recipe for a human language." Continuing with the analogy to baking, Baker (2001) states: "A parameter is an ingredient that can be added in order to make one kind of language or left out in order to make another kind. A parameter could also be a combining procedure that can be done in two or three ways to give two or three different kinds of languages. If you take the generic ingredients of language, add spice B, and shake, you get English. If you take the same basic ingredients of language, but instead of spice B you add flavorings D and E and stir, you get Navajo" (Baker, 2001, p. 57). The human language recipe makes clear that languages that have most of the same basic elements, but one different property could make the language look very different on the surface.

There are a number of models of parameters currently under debate (cf Thornton and Crain, 2013). One point of debate is the scale of the proposed parameters. Some researchers propose that there are 'macroparameters,' large-scale typological parameters that have wide-reaching effects, while recent research tends to favor 'microparameters,' which account for small-scale variation. Baker (2001) envisages a large-scale hierarchy of parameters. The first parameter encountered by the learner at the top of the hierarchy designates the language as polysynthetic or not. Polysynthetic languages are ones which have very complex long words that, if translated into English, might be expressed as a full sentence. If the learner

 http://dx.doi.org/10.1016/B978-0-08-097086-8.53008-6

chooses the setting for a polysynthetic language, decisions relevant only for polysynthetic languages follow. If the learner chose the nonpolysynthetic parameter setting, a host of different decisions follow. Current models of parameters also can be distinguished by whether learners encounter parameters in a specific order, or whether order is unimportant. Another point of debate is the amount of input required to set a parameter. Does abundant input result in swifter setting of the parameter? Do children need a fixed amount of input before a parameter is set, or is just witnessing a particular type of data in the positive input sufficient? Does the learner initially begin with one particular setting of the parameter, and are they free to choose between two options, or do they engage both settings of the parameter until the input data resolve which one correctly matches the linguistic input? All of these issues make for lively debate. Despite the debate, there is, nevertheless, agreement in the generative linguistics field that parameters sit at the heart of cross-linguistic variation.

From the perspective of first language acquisition, parameters accelerate the learner's path to the adult grammar of the language of the environment, whatever it happens to be. All the learner need to do is fill out the menu of choices in consultation with the local language. The benefit is that the learner is kept 'on track' to converge on the adult grammar without mistaken hypotheses that would slow acquisition. The next section illustrates how parameters work with some examples from the empirical literature.

Parameters and Cross-Linguistic Acquisition

Parameters assist in guiding children's acquisition of syntactic and semantic properties. Those parameters that guide very basic properties of language such as word order are likely to be set early in the course of acquisition (Wexler, 1998). Indeed, there is little evidence from children's productions that they make word order errors. Other properties of languages can take more time to settle on. We will see one example from semantics where children are slower to map their initial hypothesis for a parameter onto the local language.

Let us begin with examination of the parameter known as the *wh*-parameter, originally proposed by Huang (1982). This parameter captures a two-pronged cross-linguistic difference in the position in which question words (*wh*-words, e.g., who, where, what) are pronounced. In languages like Chinese and Japanese, question words are produced 'in situ,' in the position of the phrase that is being questioned. This can be illustrated in (1), where it can be seen that in Mandarin Chinese, the question word for what 'shenme' is positioned after 'eat' as in 'The alpaca is eating what?'

1. Yangtuo zai chi **shenme?** Mandarin Chinese
   Alpaca be eat what
   'What is the alpaca eating?'

(2) the alpaca is eating what English Movement
what is the alpaca __ eating ____

In English and other Romance and Germanic languages, the question word originates in the base position in which it is pronounced in Chinese or Japanese. It is then assumed to move to the highest position in the hierarchical sentence representation, where it is pronounced. In English, the auxiliary verb (*is* in (2)) also moves to achieve the correct word order. Children use the language input to decide if their language is the Chinese type or the English type, and set the parameter accordingly. Assuming that the data that children need to set the parameter is abundant, children can be predicted to set the parameter early. In this case, word order is also implicated, which is perhaps another reason to expect early mastery of the position of the question word. At any rate, cross-linguistic reports suggest children do not misplace the question word. Of course, one might be tempted to say there is no parameter – children simply copy what they hear in the language input. This would be a reasonable hypothesis for the position of question words, but it will not work for some other cases we will consider.

Let us turn to another kind of question, in which the question word originates in an embedded clause; these are known as 'long-distance questions.' In English long-distance questions, the question word is pronounced, as before, in initial position, as shown in (3). Syntactic theory proposes that the question word must move in a local fashion, passing through the position between clauses on its way to the highest position in the main clause. In (3), it can be seen that the question word originates in the embedded clause after 'eating.' It then moves between clauses, which is shown by 'what' in small print, and finally to sentence initial position.

(3) you think the alpaca is eating what English Movement
what do you think what the alpaca is eating ___

'What do you think the alpaca is eating?'

It is interesting that some languages and language dialects including German and Dutch dialects, Frisian, Afrikaans, Romani, and Passamaquoddy, record the passage of movement of the question word in a position between clauses as well as sentence-initially. In some German dialects, for example 'who' (*wen*, which is the accusative form of the question word) can be copied, as in (4):

4. Wen glaubst du, wen das Alpaka küsst?
   Who think you who the alpaca kisses
   'Who do you think the alpaca is kissing?'

Turning to children's productions of long-distance questions, children acquiring English (Thornton, 1990; Crain and Thornton, 1998) and also Dutch (Van Kampen, 1997), French (Jakubowicz and Strik, 2008), Basque, and Spanish (Gutierrez, 2004, 2006) have been found to pronounce the question word between clauses, even though the adult grammar of these languages does not permit the copy of the question word between clauses. Thornton's (1990) finding for English using elicited production techniques was that just some children produce these '*wh*-copying' structures. *Wh*-copying questions in children's productions are shown for a range of languages in (5). (It is important to distinguish questions with a copy of the question word (*wh*-copying structures) from questions in which a scope marker is in the highest position and the true question word is positioned between clauses. In English, this would amount to a question like 'What do you think who is under there?' to mean 'Who do you think is under there?' This is known as 'partial movement.' It is a very productive structure in many languages, and also appears as a nonadult structure in child language in a number of languages, but it is not the structure under discussion here.)

5. a. English, Thornton 1990
   Who do you think *who*'s under there?
   Who do you think *who* jumped over the can?
   b. Dutch, Van Kampen, 1997
   Wat denk je *wat* ik zie
   what think you what I see
   'What do you think I see?'
   c. Basque, Gutiérrez, 2004
   Nor uste duzu nor bizi dela etxe horretan?
   who think aux-2SG who lives aux-comp house that-in
   'Who do you think lives in that house?'
   d. Spanish, Gutiérrez Mangado, 2006
   Dónde crees dónde ha ido el niño?
   where think-2sg where has gone the child
   'Where do you think the child has gone?'
   e. French, Jakubowicz and Strik, 2008, p. 120
   Où Lala a dit où le poisson nage?
   where Lala has said where the fish is swimming?
   'Where did Lala say that the fish is swimming?'

The possibility of pronouncing the copy of the question word between clauses varies cross-linguistically, so this can be termed a parameter. As Rizzi (2009) proposes, some parameters focus on such differences in pronunciation, while others target structural properties.

The empirical facts from languages studied to date show that children from a variety of languages produce the *wh*-copying structure, despite the fact that it is not part of their language input. These facts would be puzzling on a usage-based theory of language acquisition. There is no reason to expect children from diverse language groups to produce questions with a *wh*-copy between clauses in the absence of any evidence in the language input. The facts are unsurprising on the theory of UG, however, since the parameter simply presents the option of pronouncing the intermediate copy of the question word or not. There still remains the question of why children do not immediately set the parameter in accordance with the input produced by adults in their environment, however. Most likely, children have little experience with long-distance questions. In most circumstances, a simple question is sufficient to fulfill communicative requirements. Once placed in a context that specifically calls for a long-distance question (such as asking a puppet what he thinks or believes in an experimental context), it could be that children simply guess which parameter setting is correct for their language. This explanation would explain why some children produce long-distance questions with a copy and others produce adultlike question from the start. For those children who guess the non-English parameter setting, positive evidence from the language will provide evidence for the alternative 'menu' choice.

Language acquisition researchers working within a usage-based framework have responded to the empirical findings from English. The proposal is that children attempt to be creative by merging familiar question structures (Dabrowska et al., 2009). Children are claimed to overlay the templates *What do you think?* and *What the alpaca is eating?* The first question is a direct question and the second one, *What the alpaca is eating?* is 'cut' from an indirect question such *as I know what the alpaca is eating* and 'pasted' together with *What do you think?* to form the child's copying structure *What do you think what the alpaca is eating?* The problem with the 'cut and paste' strategy is that without advance knowledge of which combinations of questions yield grammatical complex questions, children potentially could create many ungrammatical combinations. For example, indirect questions can be nonfinite as in *I know what to eat*, so the child could paste *what to eat* together with *What do you think?* to form *What do you think what to eat?* Such questions are not attested in languages that allow *wh*-copying and neither have children been found to use a *wh*-copy in questions with infinitival clauses (see Thornton, 1990; Crain and Thornton, 1998). The same is true for complex question phrases such as 'which boy.' Questions with a copy of a complex *wh*-phrase are not attested in languages that allow a wh-copy, and neither are they found in children's copying productions. It is not clear how these would be excluded by Dabrowska et al.'s proposal that children overlay templates. These ungrammatical question forms would be difficult to recover from without 'negative evidence.' For further discussion of these data and further facts that align with this parameter, see Crain and Thornton (2011).

Children may also initially hypothesize a parameter setting that is not reflected in the local language for learnability reasons. This can be illustrated with a parameter that encodes scope relations between logical expressions. This parameter was proposed originally in work by Goro (2004) and explored in subsequent research by Crain and colleagues (see e.g., Crain et al., 2006; Crain, 2012).

Scope ambiguities typically arise when two logical operators appear in the same clause. One example is sentences with negation and disjunction. In negated disjunctions in English, negation takes scope over disjunction, so disjunction yields a conjunctive entailment. For example, (6) entails that Ted didn't order pasta and Ted didn't order sushi. This shows that English negative statements with disjunction correspond to one of de Morgan's Laws of propositional logic: $(A \vee B) \Rightarrow A \wedge B$.

6. Ted didn't order pasta or sushi.

Other languages, including Japanese and Mandarin Chinese, assign the opposite scope relations. In (7), the Mandarin disjunction word for 'or,' *huozhe*, appears with negation, *meiyou*. Although the surface order of negation and disjunction are the same in Mandarin as in English, adult speakers of Mandarin judge (7) to mean that Ted didn't order sushi *or* Ted didn't order pasta. Adult speakers of Mandarin accept (7) in three circumstances, where (i) Ted ordered pasta, but not sushi, (ii) Ted ordered sushi, but not pasta, and (iii) Ted ordered neither pasta nor sushi. In logic, the corresponding formula for the interpretation of (7) in Mandarin is $(A \vee B)$, which does not entail $(A \wedge B)$.

7. (*Wo cai*) *Ted meiyou dian yidalimianshi huozhe shousi*.
   (I guess) Ted not order pasta or sushi
   'It's either pasta or sushi that Ted did not order'

The different interpretations across languages are attributed to a parameter, called the disjunction parameter. In one class of languages, disjunction takes scope over negation in simple negative sentences. Let us call this the (OR > NEG) value of the parameter. In the second class of languages, negation takes scope over disjunction, so this is the (NEG > OR) value. Another difference across languages is the interpretation that is assigned to negated conjunctions, as in *Ted didn't order both sushi and pasta*, but we will restrict our attention to disjunction here.

It has been proposed that children's initial settings of scope parameters are constrained by a learning principle called the semantic subset principle (Crain et al., 1994) that is part of UG. If we apply this to the disjunction parameter, it is easily verified that the circumstances in which sentences are true on the (OR > NEG) value comprise a superset of those circumstances that make sentences true on the (NEG > OR) value. In other words, the binary values of the parameter are in a subset/superset relation. To see this, note that the NEG > OR value yields a 'neither' reading, whereas the OR > NEG value yields a 'not both' reading. Clearly, the statement *neither A nor B* is true in a subset of the circumstances corresponding to the statement *not both A and B*. When parameter values are in a subset/superset relation, a principle of learning dictates that children acquiring all languages must initially select the subset value. In this case, they can use positive evidence to add the extra interpretations. If they chose the superset value, there would not be positive evidence available for them to reset the parameter to the correct value.

Adopting the semantic subset principle, Goro (2004) predicted that children learning languages like Mandarin and Japanese (which Goro studied) would initially interpret (6) in the same way as English-speaking children and adults interpret *Ted did not order pasta or sushi*. That is, children acquiring Mandarin were predicted to initially set the parameter to (NEG > OR), which would be different from adults. If so, then children should generate a conjunctive entailment for (6), whereas adult speakers do not. In short, the commonality that is predicted is between Mandarin child language and English rather than between Mandarin child language and Mandarin adult language.

The prediction, then, is that children learning Mandarin and Japanese would initially speak a fragment of a foreign language, in this case a class of languages that includes English. This prediction was first confirmed in an experimental study of Japanese-speaking children, reported by Goro and Akiba (2004), and has been replicated in Mandarin-speaking children by Jing et al. (2005). Children acquiring both languages generate a conjunctive entailment for disjunction when it appears in the scope of negation, whereas adult speakers of both languages do not. See Crain (2012) for further contexts where similar cross-linguistic differences emerge.

The interpretations assigned by children are evidence that children acquiring both English, Japanese and Mandarin favor the subset value of the relevant parameters. Children hear the further interpretations that are possible in the language of adults, and use these data to reset the parameter to the correct value. This guarantees that children will converge on a grammar that is equivalent to those of adults in the same linguistic community. So, this view explains why some children deviate from the adult language of their community, while at the same time, children learning other languages conform. On usage-based accounts of language acquisition, however, the initial mismatch with adults that we have observed is not anticipated, and would be difficult to explain, since children are not acting in concert with the linguistic data in their environment. Another puzzle for usage-based accounts would be why the nonadult interpretations that children hypothesize are observed in children acquiring a range of typologically unrelated languages. In the next section, we consider how universal principles of UG put their mark on cross-linguistic language acquisition.

Universal Principles and Acquisition

Children harness UG as they engage with linguistic input from parents and caretakers. The principles that are part of UG constrain children's hypothesis space, eliminating hypotheses that would permit sentence forms and sentence interpretations that are not part of the adult grammar. The theory posits universal principles, in part, as a solution to empirical findings from Brown and Hanlon (1970) among others, that parents and caretakers tend not to correct children's productions for ungrammaticality. Instead, parents respond to the truthfulness of their children's utterances. For this reason, parents are said not to provide 'negative evidence' – that is, correction for ill-formed strings of words or for meanings that are not permissible. Without parental guidance about what sentences are not grammatical or what is not a legitimate meaning for a sentence, children would have difficulty learning the bounds of the adult grammar from positive evidence alone.

One universal principle that has been widely discussed in the literature is 'structure dependence.' This principle limits children to operations or rules that are based on hierarchical

representations for sentences. In effect, children should not hypothesize rules that are based on the linear order of words. This is true, no matter what language is being acquired. One characteristic of cross-linguistic acquisition, then, will be the absence of rules in children's language that could be said to violate structure dependence. More broadly, children's productions will never feature properties ruled out by principles, and children will never give sentence interpretations that are prohibited by universal principles. As far as cross-linguistic acquisition is concerned, then, universal principles are identifiable by the absence of the phenomena they prohibit. Here we will examine a particular example from English that has been debated at length in the literature. This concerns children's hypotheses about the relationship between ordinary sentences and their question counterparts.

In generative linguistic theory, ordinary declarative sentences are related to their yes/no question counterparts via a movement rule. Children need to observe declaratives and yes/no questions in the linguistic input and formulate a movement rule that works more generally. Imagine the child is faced with the following mini-set of data in (8). All of these sentences and questions are 'simple' sentences, and we can assume these make up the majority of the input data.

8. a. Alpacas are cute — Are alpacas cute?
   b. The alpaca is eating — Is the alpaca eating?
   c. The cute alpaca can run fast — Can the cute alpaca run fast?

Children could, in principle, come up with various hypotheses about how yes/no questions are derived. If children were able to entertain linear hypotheses, based on the order of words in the sentences, they could potentially come up with the rule in (9). This rule refers to notions like 'first,' so this would entail counting from left to right to identify the auxiliary verb that is to be moved.

9. Move the first auxiliary verb/modal to sentence initial position

This linear hypothesis would, nevertheless, correctly predict all the grammatical yes/no questions in the right hand column in (8). It would incorrectly predict the form of a more complex question, as in (10), in which the noun phrase that is the subject of the sentence is modified by a relative clause. That is, 'alpaca' is modified by 'that is eating.' In (10), the first auxiliary verb that you come to when encountering the string of words from left to right is 'is,' inside the relative clause.

10. The alpaca that is eating can run fast — *Is the alpaca that eating can run fast?
    ✓Can the alpaca that is eating run fast?

The universal principle of structure dependence should not permit children to hypothesize a rule like (9). Instead, their knowledge that movement rules are formulated in terms of elements in a hierarchical structure would guide them to use notions like 'auxiliary verb in the main clause,' as in (11) which would correctly target 'can' for movement:

11. Move the auxiliary verb in the main clause to sentence initial position

The example from yes/no question formation applies to English but structure dependence has broad application to all languages. No child learning any language should test linear hypotheses.

Usage-based theories of linguistics deny that our hypotheses are guided by linguistic principles specific to the language faculty. Ambridge et al. (in press) for example, claim that more general cognitive conceptual structure maps onto relevant linguistic units. The idea is that a noun phrase modified by a relative clause as in *The alpaca that is eating* is the same kind of functional unit as *the alpaca*, or *the cute alpaca* and therefore these functional units will map onto linguistic units in children's questions. One might question whether this is a tendency or an absolute prohibition like a universal principle. Either way, the empirical findings from child language reported in Crain and Nakayama (1987) are that children are not found to consider hypotheses that would cause them to ask structure dependence violating questions like *Is the alpaca that eating can run fast?*

This concludes our short tour of the theory of UG and its explanation of children's acquisition of language. We have seen that biological endowment of principles and parameters, for the most part, speeds children's path to convergence on the adult grammar. Occasionally, children misset parameters. This may be because their initial guess is incorrect, or because they are guided by a learnability principle. The result is that children temporarily hypothesize properties that are not matched in the grammars of adults around them. Such cases are important for two reasons. They cause us to question the common sense notion that children learn everything about language from their parents. Second, they present empirical data that can adjudicate between current theories of language acquisition.

See also: Dependency in Language; Language Acquisition; Language Development, Theories of; Language Development: Emergentist Theories; Movement Theory and Constraints in Syntax; Second Language Acquisition.

Bibliography

Ambridge, B., Lieven, E.V.M., 2011. Child Language Acquisition: Contrasting Theoretical Approaches. Cambridge University Press, Cambridge.

Ambridge, B., Pine, J.M., Lieven, E.V.M., 2014. Child language acquisition: why universal grammar doesn't help. Language 90 (3), e53–e90. http://dx.doi.org/10.1353/lan.2014.0051.

Baker, M., 2001. The Atoms of Language. Basic Books, New York.

Brown, R., Hanlon, C., 1970. Derivational complexity and the order of acquisition in child speech. In: Hayes, J. (Ed.), Cognition and the Development of Language. Wiley, New York, pp. 11–53.

Chomsky, N., 1986. Knowledge of Language: Its Nature, Origin and Use. Praeger, New York.

Chomsky, N., 1995. The Minimalist Program. MIT press, Cambridge, MA.

Crain, S., 2012. The Emergence of Meaning. Cambridge University Press, Cambridge.

Crain, S., Goro, T., Thornton, R., 2006. Language acquisition is language change. Journal of Psycholinguistic Research 35, 31–49.

Crain, S., Nakayama, M., 1987. Structure dependence in grammar formation. Language 63, 522–543.

Crain, S., Ni, W., Conway, L., 1994. Learning, parsing and modularity. In: Clifton Jr., C., Frazier, L., Rayner, K. (Eds.), Language Processing and Language Acquisition. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 443–467.

Crain, S., Thornton, R., 1998. Investigations in Universal Grammar: A Guide to Experiments on the Acquisition of Syntax and Semantics. MIT Press, Cambridge, MA.

Crain, S., Thornton, R., 2011. Syntax acquisition. WIREs Cognitive Science. http://dx.doi.org/10.1002/wcs.1158.

Dabrowska, E., Rowland, C., Theakston, A., 2009. The acquisition of questions with long-distance dependencies. Cognitive Linguistics 20, 571–597.

Goldberg, A., 2003. Constructions: a new theoretical approach to language. Trends in Cognitive Science 7, 219–224.

Goldberg, A., 2006. Constructions at Work: The Nature of Generalizations in Language. Oxford University Press, New York.

Goro, T., 2004. The emergence of universal grammar in the emergence of language: the acquisition of Japanese logical connectives and positive polarity (Ph.D. thesis). University of Maryland.

Goro, T., Akiba, S., 2004. The acquisition of disjunction and positive polarity in Japanese. In: Chand, V., Kelleher, A., Rodriguez, A.J., Schmeiser, B. (Eds.), Proceedings of the 23rd West Coast Conference on Linguistics. Cascadilla Press, Somerville, MA, pp. 251–264.

Gutierrez, M.J., July 2004. The acquisition of L1 Basque long-distance wh-questions: a longitudinal investigation, paper presented at the first Bilbao-Deusto student conference in linguistics, Bilbao. Spain 8–10.

Gutierrez, M.J., 2006. Non-adult questions in the acquisition of L1 Spanish long-distance wh-questions: a longitudinal investigation. In: Torrens, V., Escobar, L. (Eds.), The Acquisition of Syntax in Romance Languages. John Benjamins, Amsterdam, pp. 251–287.

Huang, C.T.J., 1982. Logical relations in Chinese and the theory of grammar (Ph.D. dissertation). MIT, Cambridge, MA.

Jing, C., Crain, S., Hsu, C.F., 2005. Interpretation of focus in Chinese: child and adult language. In: Otsu, Y. (Ed.), Proceedings of the 5th Tokyo Conference on Psycholinguistics. Hituzi Publishing Company, Tokyo, pp. 165–190.

Rizzi, L., 2009. Some elements of syntactic computations. In: Bickerton, D., Szathmáry, E. (Eds.), Biological Foundations and Origin of Syntax. MIT Press and Frankfurt Institute for Advanced Studies, Cambridge, MA, pp. 63–88.

Jakubowicz, C., Strik, N., 2008. Scope marking strategies in the acquisition of long distance question in French ans Dutch. Language and Speech 51 (1&2), 101–132.

Tomasello, M., 2000. Do children have adult syntactic competence? Cognition 74, 209–253.

Tomasello, M., 2003. Constructing a Language: A Usage-Based Theory of Language Acquisition. Harvard University Press, Cambridge, MA.

Thornton, R., 1990. Adventures in long-distance moving: the acquisition of complex wh-questions (Ph.D. dissertation). University of Connecticut, Storrs.

Thornton, R., Crain, S., 2013. Parameters: The Pluses and Minuses, Chapter 27: The Cambridge Handbook of Generative Syntax. Cambridge University Press, Cambridge, 927–970.

Van Kampen, J., 1997. First Steps in Wh-Movement. Ponsen & Looijen, Wageningen.

Wexler, K., 1998. Very early parameter setting and the unique checking constraint: a new explanation of the optional infinitive stage. Lingua 106, 23–79.

First World War, The

Jay Winter and Charles J Stille, Yale University, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Winter, volume 8, pp. 5688–5694, © 2001, Elsevier Ltd.

Abstract

The First World War was the first 'total war' in history, in the sense that it pitted industrialized imperial powers in a conflict that lasted for 50 months, drawing in the resources of the whole world and killing more soldiers in combat than any previous war to date. Its causes were built into the imperial and economic antagonisms of world powers for control of Northwestern Europe, and thereby for control of the empires they commanded. Its consequences included the construction of state capitalism, a mixed form of private ownership and public management, the emergence of a communist revolution in Russia, threatening to build a 'red bridge' across Europe, and a disgruntled and militarized minority in Germany and Italy prepared to overthrow the postwar order when the time was right. The Great War solved none of the conflicts that had brought it about; its demographic and economic effects were devastating, and it merely rearranged the explosive materials of international politics so that they could be set alight again in 1939.

The First World War was the first 'total war' in history. Its dimensions were greater, its casualty lists longer, its propensity to draw into its orbit all social and economic forces more powerful than any previous war in history. The conflict was not so much 'total' as 'totalizing,' tending to the full mobilization of all the resources of the combatant powers. Seventy million men put on uniforms; 9 million died on active service. What turned these differences in the scale of warfare into differences of kind was the fact that the conflict was fought between fully industrialized powers. Thus the stalemate; thus the lethality of warfare; thus the terminal effect of defeat on the imperial structures of Germany, Russia, Austria–Hungary, and Turkey. The Allies won because of their greater resources as imperial powers and because of their greater sophistication in distributing goods and services between military and civilian claimants. Democracies waged 'total war' more efficiently than dictatorships.

Causes

The outbreak of the First World War in the summer of 1914 has been attributed to accident, design, and confusion. The first position suggests that no one intended that armed conflict would break out in 1914; the second, that German and Austrian military elites planned the conversion of a Balkan diplomatic crisis into an armed confrontation; the third, that political and military leaders throughout Europe began to maneuver with some notion that war was possible, but found that they could not resist the momentum of confrontation (Joll, 1992).

What was the sequence of events subject to these interpretations, and which has commanded the support of most historians who have addressed the issue of the descent into war in 1914? On 28 June 1914, the Archduke Franz Ferdinand of Austria–Hungary, heir to the throne, was assassinated in Sarajevo. The crime was carried out by a group of Serb student patriots from neighboring Bosnia, then a province of the Austro–Hungarian empire. Their aim was to demonstrate the force of Slavic nationalism, and to challenge the domination of these south Slavic lands by imperialists in Vienna and Budapest.

In this aim, they succeeded completely. The government in Vienna decided to make this crime a matter of state, and to lay the blame for this political murder on the independent Kingdom of Serbia, and its intelligence services. On 5–6 July, Austrian officials went to Germany to seek support for their policy of making Serbia 'pay' for the assassination. German officials gave their approval, offering a 'blank check' to their Austrian allies. With this backing, on 23 July, the Austrians presented Serbia with a list of demands that constituted a challenge to the status quo in the Balkans.

This is where Russia entered the conflict. A humiliation for Serbia would humiliate her Slavic ally, Russia. With Russian backing, Serbia met most of the Austrian demands, but refused to capitulate completely. In support of Serbia, Russia ordered partial mobilization of her armies on 29 July. In support of Austria, and in response to the numerical superiority of Russian forces, the German army mobilized too.

Once German mobilization was in progress, the crucial link appears between a conflict in the Balkans and an all-European war. German military planning had for years established a set of priorities to safeguard the security of the German empire. The primary point here was the need to avoid fighting a two-front war. Thus, it was necessary for Germany to eliminate a Western threat on the part of Russia's ally, France, before facing the huge armies Russia could put in the field. To do so, a military plan – the Schlieffen plan – was developed through which German forces could move in a huge arc from Cologne in southwestern Germany westward through Belgium and then southward through France. The destination, after 42 days, was in the vicinity of Paris, where the French army would be destroyed, just as had occurred in 1870.

The Schlieffen plan was thus a German response to the threat of a two-front war. That threat had materialized out of the confrontation of Germany's ally Austria–Hungary with Russia's ally Serbia. But the German plan to destroy the French army by an arcing move through Belgium threatened to bring in the United Kingdom, guarantor of Belgian independence (Ritter, 1958).

And that is precisely what happened. The German invasion of Belgium precipitated a state of war among the five great powers: Germany and Austria–Hungary on one side, and France, Russia, and the United Kingdom on the other. By 4 August 1914, the First World War had begun.

Had anyone intended this moment to occur? Probably not. While the chief catalyst of the war crisis was the decision by Austria and Germany to press Serbia to pay a political price for the assassination, the outbreak of war was less a conspiracy than a complex mixture of arrogance, a sense of national 'honor,' ignorance, and confusion. All the actors in this drama misjudged each other. Each move forward in the crisis was followed by a further heightening of tension. First Austrian honor was at stake, then Russian honor; then German fears of encirclement led her to respond to a perceived threat from Russia by threatening Russia's ally, France, and when confronted by a likely invasion of France, the United Kingdom hesitated, and then in defense of her honor, she declared war on Germany.

Two elements in the crisis were crucial in preventing a diplomatic solution. The first was temporal; the second, structural. First, timing. The war crisis lasted for 1 month. But a closer view of these events presents us with an even shorter timetable in the slide toward war. One week separated the Austrian ultimatum to Serbia – on 23 July – from Russian and German mobilization – 29–31 July. In that time period, it was simply impossible for alternatives to be explored, or for the weight of antiwar and pacifist opinion, by no means negligible, to be brought to bear on political leaders. The crisis moved too fast to be stopped.

The second determinant of a general European war was, broadly speaking, structural. That is, the chronic instability of southeastern Europe was insufficient in and of itself to bring the Great Powers to war. What turned a minor conflict into a major crisis was the linkage between the Balkan conflict and the longer term balance of power between Germany and the United Kingdom. Once the United Kingdom and Germany were aligned on opposite sides, then the crisis of 1914 became a contest for control over northwestern Europe. Should Germany defeat France in 1914, as she had done in 1870, the German navy would occupy the Channel ports. British shipping lanes, necessary to feed her home population, would be open at the pleasure of the German navy. No British government could accept that prospect (Kennedy, 1980).

It is true that the British Foreign Secretary, Lord Grey, did not make this issue clear in his diplomatic handling of the war crisis. The German Chancellor Bethmann-Hollweg hoped that the United Kingdom would stay out of the war. But geographical and strategic imperatives overshadowed diplomatic miscalculations. An invasion of Belgium and France posed a real and present danger to British vital interests. And when the United Kingdom entered the war, so did the British empire. What had started in Sarajevo, echoed in Sydney, Capetown, and Calcutta. By 4 August 1914, the world was indeed at war.

Conduct

The 1914–18 war spanned the globe, but its outcome was determined by what happened on the battlefields of Western Europe. This article therefore offers an outline of the major military encounters of the war, encounters that left lasting traces both on the landscape of Europe and on the contours of the rest of the twentieth century.

The battlefield in 1914 was Victorian in character. Artillery provided more fire power than ever before, but most of it was horse-drawn. Four years later, the battlefield was transformed. By the end of the war, it took on contours recognizable to most soldiers who would come to fight in later conflicts. Infantry, artillery, tanks, and air power were all coordinated through communications systems and supply systems of great complexity and sophistication. Supplying animal power meant relatively little; supplying machines meant everything (van Creveld, 1977). The 'storm of steel' in the words of the German infantryman Ernst Jünger had come to stay (Jünger, 1929).

1914

The two major military encounters of 1914 produced very different outcomes. In Belgium and France, the German army, 1 million strong, pushed forward in a swing westward and then southward. It met stiff resistance from Belgian forts, and then was harried by stubborn French and British defensive action. As the German army moved west into Belgium, the French army moved east toward Alsace and Lorraine, provinces lost to Germany in 1871. German defenses held and inflicted massive casualties on the French. On one day, 24 August 1914, over 24 000 French soldiers died in action. This 'Battle of the Frontiers' was a massive defeat for the French. But the strength of their east–west rail links enabled the bulk of French forces to move west to meet the real threat of the German advance toward Paris. There, on the river Marne, the German thrust was stopped. The German army retreated northward and dug in on the river Aisne. When French and British forces pursuing the German army reached the Aisne, they were unable to move the invaders. There the 'Western Front' was born. It soon stretched from Belgium to Switzerland.

There is little doubt that the Battle of the Marne was a German catastrophe. The whole purpose of German strategy had failed. There would be a two-front war. By the end of 1914, the chief of staff of the German army, von Moltke, had been replaced by Erich von Falkenhayn. But what redressed the strategic balance was a massive German military victory on the Eastern front. Two columns of Russian troops advanced into East Prussia at the same time as the German army moved through Belgium and France. These columns were separated by a string of lakes. A German staff officer, Max Hoffmann, took advantage of the geography and of the existence of a north–south rail link. Under the cover of darkness, the German army in the north moved south. The Russian column in the north had no idea they were facing an empty landscape; the Russian column in the south had no idea they were facing a numerically superior force, that is, until they were annihilated at the Battle of Tannenberg. The commander of Russian forces in the south, Samsonov, committed suicide. The two commanders of the German forces, who had not planned the operation, but who reaped its rewards, became national heroes. They were Paul von Hindenburg and Eric Ludendorff (Showalter, 1991).

1915

The war spread in 1915, but the rough stalemate of 1914 remained unbroken. One major attempt to break it was in Turkey, then an ally of Germany. To "stop our troops from chewing on barbed wire" on the Western front, as Winston Churchill, Britain's First Lord of the Admiralty, put it, a plan was hatched to knock Turkey out of the war by a naval operation. The aim was to use British naval power to force the straits that connected the Mediterranean Sea to the Black Sea. Once done, the Turkish capital of Constantinople would be indefensible. The defeat of Germany's ally Turkey would serve many purposes. It would open supply lines to Russia; it would convince all and sundry, especially in the Islamic world, of the error of siding with Germany in the world war; it might even convince Germany to withdraw from France and Belgium.

Nothing of the sort happened. Here too, a mixture of arrogance, ignorance, and confusion ruled. The enemy was Oriental, and all the distortions attending Western attitudes toward the Orient flourished. The enemy was underestimated. Attempts to push a naval force through the Dardanelles failed in March 1915. The Allies decided to mount a joint land–sea operation, but no one had a clear idea of the terrain on which soldiers would land. When they went ashore on 24–25 April, they faced steep cliffs and stiff Turkish resistance. They never got off the beaches. A 6-month stalemate ensued, after which Allied forces withdrew (Prior, 2009).

1916

Two further attempts were made to shift the balance of power on the Western front. The first was at Verdun, in eastern France. There, on 21 February 1916, more than 1 million shells fell on French positions in the hills north of the garrison city of Verdun. The idea was to force the French into total defense of this relatively unimportant position. The French, Falkenhayn reasoned, valued Verdun as a symbol of French freedom and power (Afflerbach, 1994). He was right. The French committed 259 out of its full complement of 330 infantry regiments to the struggle for Verdun. They moved into Verdun along a vast conveyor belt, which the French called *la voie sacrée*, the sacred way. It certainly was a field of sacrifice. Falkenhayn succeeded in bleeding the French army white, but he did so at a massive cost to his own troops. Perhaps half a million men died at Verdun, where, by November, the French army had recaptured all the ground lost in the early days of the battle. Lasting 10 months, the Battle of Verdun was the longest battle in history. The front remained about where it had been before the bloodbath (Horne, 1962).

The same failure to break the stalemate occurred further to the north, where the British and French armies opened a major offensive on 1 July 1916. Here, the fault lay in an underestimate of the enemy's fortifications and an overestimate of the power of artillery to destroy them. A 1-week barrage by British guns was supposed to obliterate German defenses between Amiens and St Quentin in the lowlands of the Somme. But the deep dugouts under the German lines withstood the barrage, and when British forces moved out early in the morning of 1 July, they were mown down by German machine guns. Of the 100 000 men who went over the top, 60 000 were casualties, of whom 20 000 died in that single day. Thereafter, British and German forces regrouped. A further advance on 14 July was somewhat more successful, but by September, it was apparent that no breakthrough would take place. German lines held, but at a high price. German casualties numbered approximately 450 000; together French and British casualties surpassed 600 000. One million men fell, and the lines remained roughly where they had been at the beginning of 1916 (Keegan, 1976).

Stalemate was also the rule at sea, where the British fleet kept the German fleet penned up to the east of Denmark. One attempt to shift the war to the west at Jutland resulted in mixed fortunes for both navies. But the ultimate result was that the German high seas fleet was kept away from Britain.

1917

In 1917, the balance of power remained unchanged on the Western front. But in the east and in Italy, Germany and Austria were in the ascendant. First came the crisis in Russia. After the February revolution that toppled the Czar and put a provisional government in power, that government vowed to carry on the war. This was its worst mistake, made evident by a catastrophic offensive launched on 1 July 1917. When the Russian offensive failed, a counteroffensive produced a massive Russian retreat, which effectively destroyed the Russian army as a fighting force. When the Bolsheviks came to power in October, they saw what had to be done. They sued for peace, and signed an armistice on 3 December 1917. Germany and her allies had won the war on the Eastern front.

In Italy, the Allied cause suffered a series of alarming reverses. At Caporetto, the Italian army collapsed, and fled westward 100 km where a new defensive line was hastily formed. In France, a spring attempt to pierce German lines at the Chemin des Dames failed, and when the French commander Nivelle persisted in trying to do the 'undoable,' his army mutinied (Smith, 1994). Nivelle was replaced by Pétain; the offensive was called off; the anger of frontline troops abated. British attempts to shift the balance of power in Flanders were also unsuccessful, in part through the unhappy coincidence of their attack at Ypres occurring in the wettest summer on record. The British offensive, which ended in the rubble of the village of Passchendaele on 10 November 1917, literally drowned in mud (Prior and Wilson, 1996).

The German navy attempted to shift the balance of power by launching unrestricted submarine warfare in February 1917. This was a colossal miscalculation. It did not starve the United Kingdom into negotiating an armistice. Despite massive losses to British shipping, domestic food supplies remained adequate (Offer, 1991). What was worse from the German point of view was the provocation posed by naval attacks to the United States, which entered the war in April 1917.

1918

In January 1918, it would have been difficult for a neutral observer to predict Allied victory by the end of that year. How did it happen? First, the political and economic balance of power shifted with US entry into the war in 1917. Given the time needed to raise and then move an army of 2 million men across the Atlantic, it was apparent that US troops and US

supplies would directly enter the balance of power only in 1918. This gave the Allies a new and deep reservoir of men and materiel, which the Central Powers could not match. Thus, it was not US firepower that turned the tide, but US reserves.

This became apparent after the German army launched its last major military gamble of the war. On 21 March, 47 divisions of the German army broke through at the point where British and French forces met on the Western front. Finally, the Western front moved. Within a week, German forces had moved 40 miles (65 km) to the west. In April, they attacked at Ypres; in May and June to the south, reaching the Marne. But there the offensive broke down. The gamble had failed. The German army had lost 1 million men in a last, desperate, and futile effort to win the war (Keegan, 1999).

When German forces saw that they could not win, the balance of power shifted radically toward the Allies. Starting in early August, German troops began to surrender *en masse*. High Command was getting reports of low morale among troops, who were convinced that further sacrifices were in vain. At this point, unrest on the home front merged with despair over the military situation. The High Command saw that the game was up (Deist, 1991). At the same time, Austrian and Bulgarian positions crumbled. Ludendorff urged the kaiser to turn to the US President, Woodrow Wilson, to broker an Armistice. Ludendorff resigned. The kaiser fled to the Netherlands, leaving it to others to face a disaster he and his military had engineered. The Armistice came on 11 November 1918. The Great War was over.

Consequences

Demographic

Total casualties and losses as a proportion of those who served passed a threshold beyond previous experience. Wherever the threshold is, the total of roughly 9 million dead soldiers (according to varying estimates) is beyond it: this constitutes roughly one in eight of the men who served. Adding statistics on other casualties, it is apparent that roughly 50% of the men who served were either prisoners of war, wounded, or killed.

The most murderous theater of operations was the Eastern front, where disease and enemy action described the course of a nineteenth-century war waged with twentieth-century weapons. Of all Serbs who served in the war, 37% were killed; roughly one in four Rumanians, Turks, and Bulgarians also perished. On the Western Front, where the war was won and lost, combat was about half as lethal: German and French losses were about one in six of those who served; British losses were one in eight.

Initially, casualties among social elites were higher than among the rest of the population. The longer the war lasted, the greater was the democratization of loss. The reason is that officer casualties were higher than those in the ranks, and the social selection of the officer corps mirrored inequalities in prewar life. Consequently, in its initial phases, the higher up in the social scale a man was, the greater were his chances of becoming a casualty of war. By 1917, elites were sufficiently decimated to require the armies to draw junior officers from wider social groups, which in their turn suffered disproportionately higher casualties in the last 2 years of the war.

Among the poor and the underprivileged, the story is different. Prewar deprivation saved the lives of millions of working class men and poor peasants, whose stunted stature and diseases made it impossible for them to pass even the rudimentary standards of medical fitness for military service during the war. In the British case, roughly 35% of the men examined for military service were either unfit for combat or unfit to wear a uniform at all (Winter, 1985).

Numerically, workers and peasants formed the vast majority of the army of the dead. The same was true of disabled men, visible reminders of the war on street corners and in farming villages throughout the world.

More lethal than combat was a visitation of epidemic influenza, colloquially known as 'the Spanish flu.' In 1918 and 1919, this mutant virus killed young, healthy adults throughout the world. The war may have helped spread the disease, although it did not cause it.

Political

The Treaty of Paris, the peace settlement ending the Great War, was signed on 28 June 1919, 5 years to the day after the assassination of Franz Ferdinand in Sarajevo. Its major effects were to ratify the breakup of the Austro–Hungarian and Turkish empires, and to establish a new state system in Europe, wherein Germany was stripped of Alsace and Lorraine in the west and much of East Prussia in the east. Primary responsibility for the war and for making material reparations for it were laid at the feet of the new German Republic (Macmillan, 2002).

Through a League of Nations, international conflict would be regulated, if not resolved. An International Labour Organization was established to help address problems of working class life, a clear response to the Bolshevik revolution. The new Soviet Union was not a party to the peace negotiations or to the settlement. Instead, a kind of quarantine was applied to the Soviet Union, and tacit support was given by the victorious powers to modest military intervention in Russia on the side of the anti-Bolshevik forces. This intervention was a total failure.

While the principle of self-determination for all was established, colonial and imperial issues were resolved through bilateral talks among imperial powers. Thus, spheres of influence in the Middle East were divided between France and the United Kingdom, and contradictory promises were given to Arabs and Jews as to the future postcolonial disposition of the area. Anticolonial movements grew rapidly, and violent repression followed in Egypt, India, Korea, and China (Manela, 2007).

In short, the Great War set the terms of the international order for the rest of the twentieth century (Boemeke et al., 1998). Domestically, the political consequences of war were mixed. Socialist and labor parties had participated in government. Right wing nationalist groups also grew. Among them were fascist parties, whose elevation of the virtues of armed struggle was mixed with a hatred of the political left. The center of politics weakened, evidenced by the waning of support for Liberal parties. The polarized left–right divide took on violent form in Italy, Germany, and throughout Eastern Europe, mixing ideological and territorial conflicts in such a way as to ensure the outbreak of a new and more lethal world war in just 20 years.

The war opened a phase of domestic strife in which the upheaval of war was mixed with revolutionary struggles. Civil war in Russia lasted for 3 years after the Armistice; armed conflict continued in the new states of Poland and Hungary. The victorious Allies quarantined the new Soviet Union, and began a phase of cold warfare that lasted for 70 years. For this reason alone, some historians see the war as inaugurating the 'short twentieth century,' from 1914 to 1991 (Hobsbawm, 1995).

Economic and Social

The Great War was both the apogee and the beginning of the end of European domination of the world economy. The big loser was the United Kingdom, whose export economy suffered from import substitution in countries throughout the world who developed finished goods during the conflict, when British exports were unavailable. Chronic financial instability weakened London as a banking center. The era of free trade was over (Milward, 1984).

The losers paid for the war in a host of ways. Reparations were forced on Germany, although she managed to avoid paying most of the bill. Worse was the continuation and acceleration of wartime inflation into the postwar years, wiping out savings and deepening domestic social and political conflict in Germany and Austria (Feldman, 1993). This drastic reduction in the value of money had serious effects on the social structure and political outlook of middle class groups, threatened by proletarianization (Kocka, 1984).

Domestically, economies of scale were introduced during the war, which proved important after it. Hence, military conflict helped accelerate the movement away from the small family firm toward corporate economic life. State interest in and intervention in scientific research and education also expanded during and after war, in the light of evident strategic considerations (Hardach, 1987).

After the war, the state continued to play a greater role in economic life than ever before. This was not through the ownership of industry, but through the management of debt. In most industrialized countries, taxation levels doubled over the war decade, and much of it was spent servicing war debt. The United States was the exception here, turning because of the war from a debtor to a creditor nation, although the chronic weakness of countries indebted to it after the war, like the United Kingdom, helped turn the Wall Street crash of 1929 into the world depression.

Cultural

The cultural consequences of the Great War were evident in the politics of hatred, displacing violent conflict from the national enemy in the trenches to the class or racial enemy in the homeland itself. After 1918, Germany, Poland, Hungary, the Ukraine, Byellorussia, Turkey, Ireland, and Russia plunged into civil war (Audoin-Rouzeau and Becker, 2002).

In the 1920s, the League of Nations and pacifist veterans' organizations worked toward cultural demobilization. Their aim was to place the resort to war outside of the law. Their effort failed, and by the mid-1930s, Germany, Italy, and Japan geared up for war again.

Table 1 Some estimates of military losses among combatant countries in the 1914–18 war

					Total killed		
Country	*Total killed or died (in thousands)*	*Total mobilized*	*Prewar male population (aged 15–49 years)*	*Total prewar population*	*Per 1000 mobilized*	*Per 1000 males (aged 15–49 years)*	*Per 1000 people*
Britain and Ireland	723	6147	11 540	45 221	118	63	16
Canada	61	629	2320	8100	97	26	8
Australia	60	413	1370	4900	145	44	12
New Zealand	16	129	320	1100	124	50	15
South Africa	7	136	1700	6300	51	4	1
India	54	953	82 600	321 800	57	1	0
France	1327	7891	9981	39 600	168	133	34
French colonies	71	449	13 200	52 700	158	5	1
Belgium	38	365	1924	7600	104	20	5
Italy	578	5615	7767	35 900	103	75	16
Portugal	7	100	1315	6100	70	5	1
Greece	26	353	1235	4900	73	21	5
Serbia	278	750	1225	4900	371	227	57
Rumania	250	1000	1900	7600	250	132	33
Russia	1811	15 798	40 080	167 000	115	45	11
USA	114	4273	25 541	98 800	27	4	1
Allied total	5421	45 001	204 018	812 521	120	27	7
Germany	2037	13 200	16 316	67 800	154	125	30
Australia–Hungary	1100	9000	12 176	58 600	122	90	19
Turkey	804	2998	5425	21 700	268	148	37
Bulgaria	88	400	1100	4700	220	80	19
Central powers' total	4029	25 598	35 017	152 800	157	115	26
Grand total	9450	70 599	239 035	965 321	134	40	10

Source: Winter, J.M., 1985. The Great War and the British People. Macmillan, Basingstoke, UK.

Through its unprecedented human cost, the consequences of the Great War extended beyond the political sphere. Disabled veterans were visible on the streets of all combatant countries.

Every combatant country, every town, every village, mourned its war dead in public. The outcome is a vast array of war cemeteries near the fields of battle and of war memorials dotted in villages and towns throughout Europe, North America, Africa, Asia, and the Antipodes. Annual commemorative ceremonies were (and still are) held on 11 November (Winter, 1995) (Table 1).

The monumental character of the war indelibly marked the literature, poetry, and visual arts of the twentieth century (Cork, 1994; Fussell, 1977; Hynes, 1990). The Great War is the point of reference for much subsequent writing about catastrophe, and has introduced terminology – like shell shock – which has entered colloquial language. In many countries, although not in the former Soviet Union or the United States, its shadow is longer than that of the Second World War (Winter, 2009).

See also: Conflict and War: Anthropological Aspects; Contemporary History; Imperialism, History of; Military History; Russian Revolution, The; Second World War, The; Warfare in History; Wars among Nation-States: Patterns and Causes.

Bibliography

Afflerbach, H., 1994. Falkenhayn: Politisches Denken und Handeln im Kaiserreich. Oldenbourg, Munich, Germany.
Audoin-Rouzeau, S., Becker, A., 2002. Understanding the Great War. Hill & Wang, New York, pp. 14–18.
Boemeke, M.F., Feldman, G.D., Glaser, E. (Eds.), 1998. The Treaty of Versailles: A Reassessment after 75 Years. Cambridge University Press, Cambridge, UK.
Cork, R., 1994. A Bitter Truth: Avant Garde Art and the Great War. Yale University Press, New Haven, CT.
Deist, W., 1991. Militär, Staat und Gesellschaft: Studien zur Preussisch—Deutschen Militärgeschichte. R. Oldenbourg, Munich, Germany.
Feldman, G.D., 1993. The Great Disorder: Politics, Economics, and Society in the German Inflation, 1914–1924. Oxford University Press, New York.
Fussell, P., 1977. The Great War and Modern Memory. Oxford University Press, London.
Hardach, G., 1987. The First World War, 1914–1918. Penguin, Harmondsworth, UK.
Hobsbawm, E.J., 1995. Age of Extremes: The Short Twentieth Century, 1914–91. Abacus, London.
Horne, A., 1962. The Price of Glory: Verdun 1916. St. Martin's Press, New York.
Horne, J., Kramer, A., 2001. German Atrocities 1914. A History of Denial. Yale University Press, New Haven, U.S (2002. Understanding the Great War. Hill & Wang, New York, pp. 14–18.).
Hynes, S., 1990. A War Imagined: The First World War and English Culture. Bodley Head, London.
Joll, J., 1992. The Origins of the First World War, second ed. Longman, New York.
Jünger, E. (Ed.), 1929. Storm of Steel: From the Diary of a Storm Troop Officer on the Western Front. Chatto and Windus, London.
Keegan, J., 1976. The Face of Battle. Cape, London.
Keegan, J., 1999. The First World War. A Knopf, New York.
Kennedy, P.M., 1980. The Rise of the Anglo–German Antagonism 1860–1914. Allen and Unwin, London.
Kocka, J., 1984. Facing Total War: German Society 1914–1918. Berg, Leamington Spa, UK.
Macmillan, M., 2002. Paris 1919. Six Months that Changed the World. Random House, New York.
Manela, E., 2007. The Wilsonian Moment: Self-determination and the Origins of Anti-colonial Nationalism. Oxford University Press, New York.
Milward, A.S., 1984. The Economic Effects of the Two World Wars on Britain, second ed. Macmillan, London.
Offer, A., 1991. The First World War: An Agrarian Interpretation. Clarendon Press, Oxford, UK.
Prior, R., 2009. The End of the Myth. Yale University Press, New Haven, US.
Prior, R., Wilson, T., 1992. Command on the Western Front: The Military Career of Sir Henry Rawlinson, 1914–18. Blackwell, Oxford, UK.
Prior, R., Wilson, T., 1996. Passchendaele: The Untold Story. Yale University Press, New Haven, CT.
Ritter, G., 1958. The Schlieffen Plan: Critique of a Myth (A. Wilson, E. Wilson, Trans.). Praeger, New York.
Showalter, D.E., 1991. Tannenberg, Clash of Empires. Archon Books, Hamden, CT.
Smith, L.V., 1994. Between Mutiny and Obedience: The Case of the French Fifth Infantry Division during World War I. Princeton University Press, Princeton, NJ.
van Creveld, M., 1977. Supplying War. Cambridge University Press, Cambridge, UK.
Winter, J.M., 1985. The Great War and the British People. Macmillan, Basingstoke, UK.
Winter, J.M., 1995. Sites of Memory, Sites of Mourning: The Great War in European Cultural History. Cambridge University Press, Cambridge, UK.
Winter, J.M. (Ed.), 2009. The Legacy of the Great War. 90 Years on. University of Missouri Press, Columbia, MO.

Fisher, Irving (1867–1947)

William J Barber, Wesleyan University, Middletown, CT, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 8, pp. 5694–5698, © 2001, Elsevier Ltd.

Abstract

Irving Fisher has been widely recognized as one of the greatest American scientific economists. He has been credited with being a primary architect of the 'pillars and arches' of modern economics – to borrow a phrase from J. A. Schumpeter – through his contributions to the theory of capital and interest, to monetary theory, and to the construction of index numbers. He was also an activist in promoting economic and social reforms. His reputation suffered during the 1930s when his approach to the discipline was eclipsed by Keynesianism. His accomplishments were then neglected and had to be rediscovered.

Irving Fisher is widely recognized as one of the greatest American scientific economists. He has been credited with being a primary architect of the 'pillars and arches' of modern economics – to borrow a phrase from J. A. Schumpeter – through his contributions to the theory of capital and interest, to monetary theory, and to the construction of index numbers. He was also an activist in the promotion of economic and social reforms. His reputation suffered in the 1930s when his approach to the discipline was eclipsed by Keynesianism. His accomplishments were then neglected and had to be rediscovered.

Beginnings

Irving Fisher was born on 27 February 1867 in Saugerties, New York and was the elder son of a Congregational clergyman. His was a home grown talent, nurtured in a single American institution: Yale University. Thus, unlike most American economists reaching professional maturity in the last decades of the nineteenth century, he had no direct exposure to the teachings of the German historical school. His intellectual development was shaped by two forceful members of Yale's faculty: Willard Gibbs, a mathematical physicist, and William Graham Sumner, a political economist who was also a champion of Social Darwinism. Their joint influence was observable in his choice of topic for a Ph.D. dissertation – *Mathematical Investigations in the Theory of Value and Prices*. With the publication of this exercise in general equilibrium theorizing in 1892, Fisher was hailed as a path breaker in mathematical economics. Fisher joined the Yale faculty in 1892 as an instructor in mathematics. In the following year, he married Margaret Hazard, a daughter of a wealthy Rhode Island textile manufacturer; their union was to produce two daughters and a son. In 1895, his appointment at Yale was shifted to the Department of Political Economy, an affiliation he retained until retirement in 1935. In 1898, he was promoted to full professor, but was soon diagnosed thereafter with a potentially fatal case of tuberculosis that idled him for the next 3 years. Upon recovery, he decided that the life of an ivory tower academician was no longer sufficient to satisfy him. Indeed he was convinced that scientific professionals had an obligation to engage with the world by promoting economic and social uplift.

Contributions to the Theory of Capital, Income, and Interest

Fisher had undertaken preliminary inquiries into the nature of capital before the interruption occasioned by illness. He returned to this theme in 1904 when his energies approached full restoration. Results emerged quickly in two major treatises.

In *The Nature of Capital and Income* (1906), Fisher set out a rigorous formulation of key concepts that had formerly been clouded by confusion. Capital (including human capital) was identified as the stock of assets that yield services, whereas income referred to the flow of services that generated satisfactions. He further demonstrated that the value of capital was determined by discounting the future income stream. This finding marked a distinct analytic advance. Though now generally accepted, Fisher's insistence that asset values were oriented toward the future, rather than to historical cost, was controversial when first presented. One feature of the book's argument has continued to be controversial, i.e., his claim that the correct concept of income should identify it with consumption spending and should exclude saving. *The Rate of Interest* (1907) was a logical sequel to *The Nature of Capital and Income*.

The earlier book had established the pivotal role of the interest rate in capitalizing the value of an anticipated income stream. It thus remained to demonstrate how the rate of interest was itself determined. When Fisher first addressed this issue, debate was polarized between two rival positions: one explanation of interest focused on increments in output arising from the use of capital (the technological school); a second emphasized the necessity to reward savers for deferring consumption (the psychological school). Fisher's initial analysis was centered on 'impatience' – or an unwillingness to forgo today's consumption without the prospect of greater consumption possibilities tomorrow – as the principal explanatory factor. When he returned to this topic in *The Theory of Interest* (1930), he argued that the technological and the psychological arguments were complementary, that is, that the interest rate was determined by the interaction between

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61033-4

'impatience to spend income and the opportunity to invest it.' The real interest rate was thus unambiguously a real phenomenon, not a monetary one. This line of analysis would stand in sharp contrast to the monetary theory of interest that Keynes was to offer in his *General Theory of Employment, Interest and Money* (1936). For Fisher, money still mattered because of its impact on the price level. It could thus affect the nominal interest rate, which Fisher treated as the real interest rate plus the inflation rate.

Early Work in Monetary Theory

The Purchasing Power of Money (1911) was conceived as an exercise in establishing the validity and usefulness of the quantity theory of money, a doctrine that had been politically contaminated in the polemics over 'free silver' in the 1890s. In Fisher's formulation, 'the equation of exchange' was written as $MV + M'V' = PT$. (In this expression, M represented the quantity of cash and V the velocity of its circulation; M' stood for total deposits subject to check and V' for the average velocity of their circulation; P was defined as an average price and T as the volume of trade.) This book was a vigorously monetarist document in which Fisher maintained that changes in the general price level were linked to proportionate changes in the money supply. This was true in the 'normal period,' defined as a state of equilibrium in which the values of T and the V′s were constant. His choice of terminology on this point bred misunderstandings: he recognized readily that economic reality was typified by 'transition periods' – when adjustments to disturbances were being worked out – in which the constancies did not hold.

Fisher was not content to offer solely an analytic explanation for variations in the purchasing power of money. Convinced as he was that monetary instability produced distributive injustices by distorting the relative positions of creditors and debtors, he felt obliged to suggest a corrective. His proposed remedy prescribed that changes in the general price level should be offset by variations in the gold content of the dollar (i.e., the 'compensated dollar' plan). But this policy recommendation did not mesh with the analysis presented in *The Purchasing Power of Money*. The 'compensated dollar' scheme rested essentially on a 'commodity theory of money' in which the real value of money was determined by the value of its gold equivalent. This conclusion did not follow from the 'equation of exchange' that Fisher had set out as the basis for his 'quantity theory of money.' Schumpeter (1948) and Patinkin (1993) have suggested that this oddity can be explained because Fisher, the reformer, got the better of Fisher, the scholar.

Analytic Innovations in the 1920s

Given Fisher's concern for stabilization of the general price level, his interest in perfecting the instrument for measuring its behavior – namely, the index number – was readily comprehensible. He addressed this in *The Making of Index Numbers* (1922), a work which has continued to be influential. Shortly thereafter, he created an Index Number Institute (INI) that provided weekly reports on leading economic indicators, along with commentary on current events. These analyses were distributed nationally via a syndicated news service. At its peak, Fisher's INI column had a potential to reach about 7 million readers.

By the mid-1920s, Fisher's studies had convinced him that the stakes in price level instability were higher than he had earlier thought them to be. When writing *The Purchasing Power of Money*, he regarded distributive injustices as the major evil associated with unstable prices. His later analyses indicated that instability of the general price level was responsible for another major problem – fluctuation in the volume of trade and, with it, variations in aggregate income and employment. This conclusion, he maintained, was supported by statistical evidence indicating that the rate of change in the general price level was highly correlated with variation in the volume of trade. This relationship, however, was not instantaneous.

His empirical investigations suggested a lag of approximately 7 months between the turning point in the rate of change in prices and change in aggregate economic activity. In this research, Fisher introduced a novel statistical technique – the distributed lag – which has become a standard tool in econometrics. With its aid, his results appeared to show conclusively that a rising price level stimulated trade, while a falling price level depressed it. It thus followed that the way to eliminate macroeconomic fluctuations was to maintain a stable price level. In that spirit, the Fisher of the mid-1920s wrote that the 'so-called business cycle' was really 'a dance of the dollar' and that there was no reason why such disturbances should be tolerated.

Crusading for Causes

Fisher, the recovered TB patient, was an indefatigable champion of causes. This was evident in his tireless promotion of monetary reforms: in the first instance, in lobbying for a 'compensated dollar' that promised to stabilize the dollar's purchasing power. In the 1920s, he lobbied vigorously for legislation that would compel the Federal Reserve to maintain price level stability. His views by then had changed on how this should be accomplished. In 1923, with the discovery of the potency of 'open market operations' – a technique for altering commercial bank reserves through the purchase or sale of government securities by the Federal Reserve system – Fisher readily embraced this procedure as the major instrument for central bank control of the money supply. He was a frequent visitor to Washington as a star witness before Congressional committees with jurisdiction over monetary affairs. He also poured considerable energy into mobilizing public opinion behind reform proposals through the creation of organizations such as the Stable Money League.

Fisher's reformist agenda was not restricted to matters directly linked to his expertise as an economist. He was an enthusiast for measures to improve the health and longevity of the population. On his initiative, a group of leading citizens was organized in 1907 to lobby – unsuccessfully as it turned out – for the creation of a Federal Department of Health that would establish national standards for public health practices and monitor their enforcement. He enjoyed greater success in

the promotion of a Life Extension Institute to persuade life insurance companies that it would be profitable to require and pay for annual physical examinations for policy holders. To reinforce its activities, he co-authored a first-of-its-kind manual of good health practices, entitled *How to Live*, that went through some 21 editions.

Fisher was one of the nation's staunchest advocates of national prohibition of the consumption and sale of alcoholic beverages. Once the 18th amendment to the Constitution (1920) had been enacted for that purpose, he was a leading voice in arguing against its repeal. Even though enforcement of prohibition left much to be desired, he maintained that this legal ban had still contributed positively to American economic expansion in the 'new era' of the 1920s through its effect on improved worker efficiency and on reduced workplace absenteeism. This line of thinking was responsible, in part, for his disastrous misreading of economic conditions in the autumn of 1929 when he predicted that the American economy was on the threshold of permanent prosperity. In the 1920s, he was also an inveterate campaigner for an international organization to promote world peace and for membership of the US in the League of Nations. He was a prominent figure as well in the eugenics movement that he hoped would encourage state governments to enact legislation permitting sterilization of the mentally defective and the criminally insane.

Fisher contributed significantly from his own pocket to support these causes. His ability to do so was enhanced by the sale of one of his inventions – an Index Visible filing system – to Rand-Kardex (shortly to become Remington Rand) for a handsome block of stock and for a position on the board of directors. When informing his son about this turn of fortune, he wrote: 'This added income will enable me to further the four chief causes which we have at heart, the abolition of war, disease, degeneracy, and instability of money (Fisher to Irving Norton Fisher, 17 June 1925).

Coping with the Disaster of Depression in the 1930s

Fisher was totally caught off-guard by the collapse of the stock market in the autumn of 1929 and by the events that followed. His personal financial circumstances were left in shambles: he managed to avert personal bankruptcy only by borrowing heavily from his wife's unmarried sister. The Great Depression was also a professional embarrassment. It did stimulate him to rethink where his earlier line of analysis had failed him. The result – reported in *Booms and Depressions* (1932) – was the debt-deflation theory of Great Depressions. The problem, as he then saw it, stemmed from a situation of overindebtedness in the late 1920s. Once a break in the price level had occurred, the real burden of debt increased. This led to distress selling, to contraction in deposit currency, and to slackening in the velocity of monetary circulation. Deflation – far from being self-correcting – simply bred more deflation. The remedy was implicit in the diagnosis: 'reflation' of the price level to its elevation at the time when most outstanding debts were contracted. When that point had been reached, the general price level should be stabilized.

To accomplish this, Fisher urged a number of proposals on the Roosevelt administration. For example, he endorsed 'stamped scrip' – a scheme which called for a special currency issue that would lose 2% of its value each week unless a two-cent stamp were affixed to it. This device (which he had borrowed from Gesell) was intended to accelerate the velocity of monetary circulation. Though the idea was not original with him, he later put his full weight behind advocacy of a plan that would require commercial banks to hold '100 percent reserves' against demand deposits, thereby eliminating their capacity to create (or destroy) check-book money by extending (or calling) loans. The money supply would instead be controlled by a public authority (e.g., a Currency Commission) that would expand the money supply as needed to achieve a prescribed 'reflation' target and thereafter to stabilize. Fisher's conception of a 'monetary rule' was tied to outcomes, as opposed to the 'rule' recommended in Milton Friedman's brand of monetarism that stipulates a fixed rate of growth in the money supply.

Fisher's thinking about recovery from the Great Depression always had a monetary orientation. He was unsympathetic toward fiscal stimulants – such as Keynesian-style programs of public works – and he made no secret about that. He was also unsympathetic to the analytic message of Keynes's *General Theory*. Even so, he and Keynes came close to convergence in their recommendations for tax policy in wartime: both championed a tax targeting consumption. The analytic props to their conclusions, however, had nothing in common. Fisher's views – which he spelled out in his final book, *Constructive Income Taxation* (1942) – rested on the analysis he had worked out in 1906 in *The Nature of Capital and Income*. This argument maintained that income should properly be defined as expenditure on consumption: thus, a sound tax system, in all circumstances, should take aim at consumption spending. Keynes's position, on the other hand, was informed by his conception of the management of the components of aggregate demand, as formulated in *The General Theory*.

Despite the fact that the Roosevelt administration took little heed of his unsolicited advice, Fisher still believed that something hopeful could be salvaged from the wreckage of depression in the form of lasting improvement in scientific investigation in economics. It had long been his aspiration to build an international organization to bring economists interested in linking economic theory with mathematics and statistics into closer collaboration. Circumstances of the early 1930s permitted this dream to become a reality. The presence at Yale in 1930 of Ragnar Frisch of the University of Oslo made it easier for three who shared the dream – Fisher, Frisch, and C. F. Roos – to lay the groundwork for the Econometric Society. The obvious shortcomings of economic forecasting in the early depression years – and an apparent need for improvement – convinced a well-to-do businessman to underwrite this new organization.

A Word about the Legacy

Fisher was accorded numerous honors by professional colleagues: among them, election to the presidencies of the American Economic Association, the American Statistical

Association, the Econometric Society, the American Association for Labor Legislation. But he lacked a committed cadre of disciples: there was no 'Yale circus' to match Keynes's 'Cambridge circus.' In the last decade of his life, his approach was largely overshadowed by Keynesianism. To younger members of the profession, he appeared to be out-of-step with the times, with the consequence that his work tended to be ignored.

With the waning of the Keynesian ascendancy, Fisher's achievements have been revisited. His pioneering analyses of capital and interest have now approached the status of an orthodoxy in the graduate training of prospective professionals. A 'rediscovery' in the 1970s documented that an article he wrote in 1926 – entitled *A Statistical Relation between Unemployment and Price Changes* – anticipated the essential insight of the Phillips Curve by more than three decades. It is interesting to speculate on how the principals in an intense controversy that divided the profession in the 1960s and 1970s would have behaved had they known that Fisher had identified the 'reswitching problem' in 1907.

In the 1990s, some of Fisher's practical proposals have enjoyed a renewed lease on life. In 1997, the US Treasury issued a 'stabilized bond,' the return on which would be adjusted to maintain constancy in its purchasing power. Fisher had advocated the creation of this type of financial instrument some seven decades earlier. In 1995, the US Department of Commerce resurrected another Fisher conception, dating from 1922, when adopting his formula for an 'ideal index number' to calculate the GNP deflator.

Fisher died on 29 April 1947 in New York City.

See also: Economic History; Health Economics; Keynes, John Maynard (1883–1946); Monetary Policy Since the 2007–2009 Financial Crisis.

Bibliography

Allen, R.L., 1993. Irving Fisher: A Biography. Blackwell, Cambridge, MA.

Fisher, I., 1997. The Works of Irving Fisher. vols. 14. Pickering and Chatto, London.

Patinkin, D., 1993. Irving Fisher and his compensated dollar plan. Federal Reserve Bank of Richmond Economic Quarterly 79, 1–33.

Schumpeter, J.A., 1948. Irving Fisher's econometrics. Econometrica 16, 219–231.

Fisher, Ronald A (1890–1962)

AWF Edwards, Gonville and Caius College, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

R.A. Fisher was one of the great intellects of the twentieth century. His contributions to evolutionary theory and mathematical genetics ensure him a place in the Darwinian pantheon. He is the undisputed father of modern statistical theory, standing alongside his great predecessors Gauss and Laplace. His statistical methodology has had a far-reaching impact whether one considers the mathematical theory of distributions, the design and analysis of experiments, the theory of estimation, the theory of significance testing, or the logic of statistical inference. In genetics he was the principal architect of the application of statistics, much of his general theory having been inspired by practical problems, especially in human genetics.

Sir Ronald Fisher, the father of modern statistics, was for most of his working life a professor of genetics, first in London and then at Cambridge. He made lasting contributions to mathematical and evolutionary genetics as well as to statistical theory. In statistics his contributions stand alongside those of his great predecessors Gauss and Laplace, while in evolutionary theory he is acknowledged as the greatest of Darwin's successors.

Fisher was born in London on 17 February 1890, the son of George Fisher, a fine-art auctioneer, and his wife Katie. His twin brother was stillborn. At Harrow School, which he entered in 1904 as a scholar, he distinguished himself in mathematics despite being handicapped by poor eyesight, which prevented him working by artificial light. His teachers used to instruct him by ear, and Fisher developed a remarkable capacity for pursuing complex mathematical arguments in his head. This manifested itself later in life in an ability to reach a conclusion while forgetting the argument, to handle complex geometrical trains of thought, and to develop and report essentially mathematical arguments in English (only for students to have to reconstruct the mathematics later). Fisher's early interest in natural history was reflected in the books chosen for special school prizes at Harrow, culminating in his last year in the choice of the complete works of Charles Darwin in 13 volumes. Fisher entered Gonville and Caius College, Cambridge, as a scholar in 1909, graduating BA in mathematics in 1912. At college he instigated the formation of a Cambridge University Eugenics Society through which he met Major Leonard Darwin, Charles's fourth son and president of the Eugenics Education Society of London, who was to become his mentor and friend. He spent a postgraduate year in the Cavendish Laboratory, Cambridge, studying the theory of errors under F.J.M. Stratton and statistical mechanics and quantum theory under J.H. Jeans (later Sir James Jeans).

Prevented from entering war service in 1914 by his poor eyesight, Fisher taught physics and mathematics in schools for the duration of the war and in 1919 was appointed a Statistician to Rothamsted Experimental Station, an agricultural station at Harpenden north of London. In 1933 he was elected to succeed Karl Pearson as Galton Professor of Eugenics (i.e., of Human Genetics, as it later became) at University College, London, and in 1943 he was elected as Arthur Balfour Professor of Genetics at Cambridge and a Fellow of Gonville and Caius College, succeeding R.C. Punnett. After he retired in 1957, Fisher traveled widely, spending his last few years in Adelaide, Australia, as an honorary research fellow of the Commonwealth Scientific and Industrial Research Organisation (CSIRO) Division of Mathematical Statistics. He died there of a postoperative embolism on 29 July 1962. His ashes lie under a plaque in a side aisle of Adelaide Cathedral.

Fisher married Ruth Eileen Guinness in 1917 and they had two sons and six daughters, and a baby girl who died young. He was elected a Fellow of the Royal Society of London in 1929 (as a mathematician) and was created Knight Bachelor by Queen Elizabeth II in 1952 for services to science. He was the founding President of the Biometric Society (now the International Biometric Society) in 1947, and served as President of the Royal Statistical Society (of the UK), of the International Statistical Institute, of the Genetical Society of Great Britain, and of his Cambridge college, Gonville and Caius. He received many honorary degrees and accepted the honorary membership of many academies at home and abroad, and was awarded all the principal medals of the Royal Society, the Royal (1938), Darwin (1948), and Copley (1956).

Fisher's first publication, while still an undergraduate, proposed the method of maximum likelihood for statistical estimation (though not in those words) and before graduating he had already given an article remarking on the surprisingly good fit of Mendel's data. In 1916, without any academic appointment but encouraged by Leonard Darwin, he completed the founding article of biometrical genetics "The correlation between relatives on the supposition of Mendelian inheritance" (1918), eventually published in 1918 after it had been withdrawn following unfavorable reviews for the *Philosophical Transactions of the Royal Society* by Karl Pearson and R.C. Punnett (both of whom he later succeeded, as he was fond of remarking). In this article, Fisher introduced the word 'variance' and established the basis of the analysis of variance.

From his post of statistician at Rothamsted, Fisher made advances which revolutionized statistics, while his advances in genetics and evolution were hardly less path breaking. In a single publication in 1922 he proved that heterozygotic advantage gives rise to a stable gene-frequency equilibrium, introduced the first stochastic model into genetics (a branching process), and initiated the study of gene-frequency distributions by means of the diffusion approximation, and in another

article he applied the method of maximum likelihood for the first time, to the estimation of linkage.

Also in 1922, building on his method of maximum likelihood, Fisher published 'On the mathematical foundations of theoretical statistics' in the *Philosophical Transactions* (1922a), a revolution in statistical thought. This time the Royal Society's referees, A.S. Eddington and G.U. Yule, were more complimentary. Among the new words and phrases Fisher introduced in the article in order to clarify the basis of his approaches to statistical inference are likelihood, consistency, degrees of freedom, efficiency, sufficiency, information, significance level, test of significance, sampling distribution, percentage point, statistic, and others of a more technical nature still. Information in Fisher's statistical sense (1925a) antedated Shannon's (1948) use of the word for a somewhat different purpose: Fisher's information measures the content of a message, whereas Shannon's measures the ability of a channel to transmit a message, independently of its meaning. In 1930, Fisher published the first satisfactory account of what J. Neyman was later to call the 'confidence' property of interval estimation.

In distribution theory Fisher built on his 1915 article deriving the exact distribution of the correlation coefficient (for whose solution he introduced the concept of a sample-space) and used his considerable mathematical ability to solve many further distributional problems, including that of the multiple correlation coefficient and the variance ratio (which G.W. Snedecor of Iowa State University, where Fisher was visiting professor in 1931 and 1936, called 'F' in his honor). His work at Rothamsted on agricultural statistics soon led to the creation of the science of experimental design with the appreciation of the fundamental role of randomization, the development of the principles of blocking and confounding, and the introduction of Latin squares and factorial analysis and similar designs chosen for their statistical efficiency. The methods which Fisher invented or improved were brought together in 1925 in his book *Statistical Methods for Research Workers* (1925b), which he updated continually over the next 23 years until the thirteenth edition was published in 1958 (a posthumous fourteenth edition was prepared from notes written before his death).

In addition to the statistical work associated with his employment at Rothamsted, Fisher pursued his interest in evolutionary theory, guided and encouraged by Leonard Darwin. Putting his deep knowledge of Mendelian theory and Charles Darwin's writings to good use he built on the success of his earlier investigations of 1918 and 1922 to write *The Genetical Theory of Natural Selection* (1930), published in 1930. This book is now seen as the chief document of the 'Neo-Darwinian revolution,' that revival of Darwinism in the 1930s and 1940s from the neglect it had suffered in the early years of the century following the rediscovery of Mendel's article in 1900. The book contains a wealth of new evolutionary arguments, from the fundamental theorem of natural selection to ideas about sexual selection, inclusive fitness and parental expenditure. More than any other work *The Genetical Theory* established a firm basis for the modern view that evolution by natural selection is primarily a within-species phenomenon.

Taking up his appointment as Galton Professor of Eugenics at University College London in 1933, Fisher's pace did not slacken. Now able to pursue experimental work, which previously he had only undertaken in a private capacity, he demonstrated his ability as a natural historian. Experimental organisms included mice, poultry, and the purple loosestrife, and even dogs (under the auspices of the Genetical Society). In a long collaboration with E.B. Ford of Oxford he studied natural populations for the purpose of estimating the power of natural selection, thus contributing to the foundation of a school of ecological genetics. But it is in human genetics that he made his most lasting contribution of this period. In 1935 he secured funds from the Rockefeller Foundation to establish a Blood-Group Serum Unit at the Galton Laboratory with the express purpose of initiating the construction of a linkage map of man, for he had already seen the connection between 'Linkage studies and the prognosis of hereditary ailments' (to use the title of his lecture to the International Congress on Life Assurance Medicine in that year). Here is to be found the intellectual origin of the Human Genome Project. At the same time Fisher, with J.B.S. Haldane and L.S. Penrose, was advancing the special statistical theory required in the estimation of human linkage.

Fisher (1935) also saw the publication of *The Design of Experiments* in which his pioneering work in that subject at Rothamsted was expounded fully for the benefit of workers in agriculture and many other fields. Like *Statistical Methods for Research Workers*, the book was an outstanding success, running to seven editions by 1960 (an eighth was later prepared from notes left by Fisher). It opened with a famous description, often quoted, of an experimental design for the problem of testing the assertion of a lady who maintained that she could tell by tasting a cup of tea whether the milk or the tea had been the first to be poured into the cup. In the course of this discussion Fisher coined the phrase "null hypothesis" for the hypothesis under test.

In 1936, Fisher subjected Mendel's published data on segregation in the garden pea to close statistical scrutiny, concluding that in some respects the data seemed to follow too closely the expectations on Mendel's genetic hypothesis, a discovery that worried him greatly, and left him speculating that perhaps Mendel's assistant, knowing full well what his master hoped to find, was tempted to adjust some of the counts accordingly.

In 1943, Fisher moved to Cambridge as Arthur Balfour Professor of Genetics, where he was reunited with his colleagues from the Blood-Group Unit which had been evacuated there during the war. An immediate consequence was his brilliant solution of the Rhesus blood-group puzzle, involving three closely linked loci which between them explained the array of serological reactions which to everyone else had appeared chaotic: Fisher did for Rhesus what Mendel did for round and wrinkled.

After the Second World War ended in 1945 Fisher attempted to establish bacterial genetics in his Cambridge department and to retain for Cambridge the Blood-Group Unit, but without success. Work in his small department revolved around linkage in the mouse and studies on purple loosestrife, wood sorrel, and primroses, always with a strong background of mathematics and statistics. His *Theory of Inbreeding* (1949) was published in 1949, and in 1950 he published the first article applying a computer to a biological problem.

In the subject of statistics, developments in the United States during and after the war had tended to obscure the great advances which Fisher had made in the subject, and he found himself having to defend many of his views against critics brought up in the more mathematical and operational climate than dominant. His repeated insistence that statistical inference was a mode of reasoning separate from, though related to, approaches based on decision theory and other behavioral considerations won him few friends, and his continued (if somewhat puzzled) defense of his troublesome 1930 concept of fiducial probability was cited as evidence that his judgment was fallible. He defended his position in a fourth book, *Statistical Methods and Scientific Inference* (1956), published in 1956 with a second edition in 1959 and a posthumous edition with many additions and clarifications. In this work he particularly emphasized the direct role of likelihood in statistical inference, in addition to the method of maximum likelihood, and continued his life-long criticism of the improper use, as he saw it, of prior probabilities and Bayes's theorem in statistical inference, a methodology previously called 'inverse probability' for which he had himself coined the word "Bayesian" in 1950.

As well as writing his books, Fisher collaborated with Frank Yates, also of Rothamsted, to produce the most widely used compilation of statistical tables (before computers rendered such volumes redundant). First published in 1938, *Statistical Tables for Biological, Agricultural, and Medical Research* (1938) went through many editions.

In retirement Fisher continued his academic work, but also became embroiled in a controversy as to the extent to which the published data supported the hypothesis, then novel, that cigarette-smoking was a cause of lung cancer. Perhaps not fully appreciating the strong feelings which the issue was to generate, he focused on the finer points of the statistical analysis which interested him, on the need always to disentangle causation and correlation, and on the fact that the evidence did not seem to him to disallow the possibility that genetical variability might be a contributory factor in the observed association.

Fisher's articles number nearly 300, and he also wrote many reviews, particularly in *The Eugenics Review* between 1915 and 1935, and letters to journals. He was an accomplished formal lecturer as may be seen from his many presidential and similar addresses, and an occasional broadcaster on scientific topics. His interests were wide, as may be instanced from two examples from his publications: in 1934 he discovered and published the idea of a randomized or 'mixed' strategy independently of von Neumann, and in 1958 he published the first suggestion that the theory of games might be applied profitably to problems in evolutionary biology, the possibility having been drawn to his attention by Dr Cavalli (later Professor L.L. Cavalli-Sforza).

Small of stature, with thick glasses and a beard, Fisher did not suffer fools gladly. He was a skilled controversialist in conversation, but his quick temper sometimes rendered further discussion impossible. His contemporaries divided rather cleanly into those who regarded him with awe and affection and gratitude for the generosity with which he offered his ideas to them, and those who found him tetchy, difficult, and remote. Especially as a professor at Cambridge, he showed great interest in the few students who passed through his small department, and he always enjoyed the company of young people. Of the many thumb-nail sketches which his greatness has inspired, perhaps the following comment on Fisher by the Cambridge cosmologist Sir Fred Hoyle contains the closest likeness in the smallest span: "I am genuinely sorry for scientists of the younger generation who never knew Fisher personally. So long as you avoided a handful of subjects like inverse probability that would turn Fisher in the briefest possible moment from extreme urbanity into a boiling cauldron of wrath, you got by with little worse than a thick head from the port which he, like the Cambridge mathematician J.E. Littlewood, loved to drink in the evening. And on the credit side you gained a cherished memory of English spoken in a Shakespearean style and delivered in the manner of a Spanish grandee" (Hoyle, 1999).

Fisher was one of the great intellects of the twentieth century. His contributions to evolutionary theory and mathematical genetics ensure him a place in the Darwinian pantheon while he is the undisputed father of modern statistical theory.

The statistical methodology he developed has had a far-reaching impact in every field of science and social science. Whether one considers the mathematical theory of distributions, the design and analysis of experiments, the theory of estimation, the theory of significance testing, or the logic of statistical inference, Fisher's is the primary influence. As a teacher of research workers through his statistical books he had no peer.

In genetics he was the principal architect of the application of statistics, much of his theory having been inspired by practical problems, especially in human genetics. With J.B.S. Haldane and others he initiated the study of linkage in man specifically for the purpose of throwing light on hereditary disease risks, thus foreshadowing the human genome project, and he encouraged the study of associations between blood groups and disease. His studies on natural and artificial populations of both plants and animals would alone have given him an honored place in the development of genetics in the twentieth century. His '1918 paper' is accepted as the starting point for the entire development of biometrical genetics, with its ramifications in plant and animal breeding.

In evolutionary biology, his book *The Genetical Theory of Natural Selection* (1930) hides a brilliant series of evolutionary insights behind a density of writing that requires more concentration than most readers have been prepared to give it, but is now acclaimed as the cornerstone of the neo-Darwinian revival of the mid-twentieth century.

See also: Distributions, Statistical: Approximations; Estimation: Point and Interval; Hotelling, Harold (1895–1973); Significance, Tests of.

Bibliography

Bennett, J.H. (Ed.), 1971–1974. Collected Papers of R.A. Fisher, 5 vols. University of Adelaide, Adelaide, Australia.

Bennett, J.H. (Ed.), 1983. Natural Selection, Heredity, and Eugenics, Including Selected Correspondence of R.A. Fisher with Leonard Darwin and Others. Clarendon Press, Oxford, UK.

Bennett, J.H. (Ed.), 1990. Statistical Inference and Analysis: Selected Correspondence of R.A. Fisher. Clarendon Press, Oxford, UK.
Box, J.F., 1978. R.A. Fisher: The Life of a Scientist. Wiley, New York.
Fisher, R.A., 1918. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh 52, 399–433.
Fisher, R.A., 1922a. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society A 222, 309–368.
Fisher, R.A., 1922b. On the dominance ratio. Proceedings of the Royal Society of Edinburgh 42, 321–341.
Fisher, R.A., 1925a. Theory of statistical estimation. Proceedings of the Cambridge Philosophical Society 22, 700–725.
Fisher, R.A., 1925b. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.
Fisher, R.A., 1930. The Genetical Theory of Natural Selection. Clarendon Press, Oxford, UK.
Fisher, R.A., 1935. The Design of Experiments. Oliver and Boyd, Edinburgh.
Fisher, R.A., 1949. The Theory of Inbreeding. Oliver and Boyd, Edinburgh.
Fisher, R.A., 1950. Contributions to Mathematical Statistics. Wiley, New York.
Fisher, R.A., 1956. Statistical Methods and Scientific Inference. Oliver and Boyd, Edinburgh.
Fisher, R.A., Yates, F., 1938. Statistical Tables for Biological, Agricultural and Medical Research. Oliver and Boyd, Edinburgh.
Hoyle, F., 1999. Mathematics of Evolution. Acorn Enterprises, Memphis, TN.

Fishing Societies

Gisli Palsson, University of Iceland, Reykjavik, Iceland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the place of fishing activities and coastal economies in social scientific thought, some of the major issues developed in the relevant literature, and recent trends in maritime anthropology; the category of fishing; natural and social models of fishing activities and coastal communities; gender relations; ecological knowledge; resource management; and the revitalized ocean.

The Category of 'Fishing'

According to some of the evolutionary theories of the nineteenth century, fishing represented a separate and early stage in the history of humans. Such an idea was underlined in many contemporary accounts of particular groups of people largely dependent on fish. The evidence on the role of early fishing, however, is fairly limited. Some scholars suggest that, from the point of view of early humans, aquatic environments must have been both hostile and an inadequate source of food. Others suggest that coastal zones must have been quite attractive, providing a worthwhile challenge, and that people may have exploited them for a long time without leaving modern groundlings any evidence. While the archaeological evidence does not indicate that fishing predated the hunting of terrestrial animals, nonetheless among hunter-gatherers fishing may have played quite an important role.

The abundance of resources in coastal zones may, indeed, have provided an opportunity for the development of complex civilizations. There is evidence for a transformation of hunter-gatherer social relations in coastal regimes, although many of the important issues involved are far from settled. Rather than seeing aquatic resources themselves as determinants of complexity, however, one should regard coastal niches as just one possible avenue for intensification (Kelly, 1992). With the advent of fieldwork and participant observation, evolutionary speculations tended to be replaced by descriptive accounts. The category of 'fishing' became much less visible than before – sometimes being subsumed under the label 'hunting and gathering.'

While the category of fishing may seem straight forward, it need not be taken for granted. The distinctiveness of fishing activities, it has been argued (Hewes, 1948), has two aspects. First, objects behave in a particular manner while in an aquatic substance, due to special conditions of buoyancy, turbulence, solubility, and refraction of light. Second, fishers and their prey occupy different media. An emphasis on differences in habitat yields a definition of fishing as "that category of human activity which is connected with the capture or gathering of animals (or plants) that regularly dwell in the water" (Hewes, 1948). The concept of fishing as a particular kind of hunting, which happens to yield fish is one element of a common anthropological scheme for classifying types of subsistence: gathering, collecting, hunting (including trapping), husbandry (including fish farming), and plant cultivation. Much like medieval Europeans who often distinguished between fishing, fowling, and hunting, modern anthropology tends to operate with three concepts of foraging – fishing, gathering, and hunting.

Some linguistic evidence suggests that fishing and hunting are perceived as similar strategies, differing only with respect to the focus of subsistence efforts. Indigenous terminologies, however, do not necessarily distinguish between hunting and other subsistence activities, including trapping, collecting, and gathering. Also, the classification of aquatic organisms, including 'fish,' varies from one society to another. Although in many societies, fish are classified according to an elaborate taxonomy, depending on their shape and usefulness to humans, the category of 'fish' represents one of the largest and most diverse categories found in most environments; at times fish terms are extended to all creatures inhabiting aquatic environments, including turtles, frogs, crocodiles, crustacea, and fish-shaped mammals. In some languages, there is no term for fish; fish are included in a heterogeneous category along with snakes, worms, lizards, and other elongated creatures.

Natural and Social Models

Many approaches to fisheries, both of the present and the past, suggest a 'natural' model of fishing, emphasizing material context and ecological relations. Social scientists have often drawn attention to the fact that fishing takes place in a relatively uncertain environment (Palsson, 1991). People who adapt to earning a living by exploiting marine resources, it is argued, often seem to manage their lives in similar ways. Thus, crew organization tends to be flexible and based more on voluntary ties than kinship obligations, to ensure cooperation and the right combination of skills, to reduce uncertainty, and spread the risks of production. Fishing societies, as a result, seem to have a range of characteristics in common due to the fact that their members have to adapt to the corresponding environments and cope with identical problems. Such approaches are reminiscent of the method of cultural ecology, according to which social life is rather mechanistically adapted to the material world.

In line with the assumptions of the natural model of fishing, there has been a strong tendency to look for parallels between trawling (industrial hunting) and small-scale fishing on the basis of a common link to water. While focusing on the

boundary between land and water is helpful for drawing contrasts between economic and social systems that are organized in similar ways, between aquatic and terrestrial industrial systems, much recent research has emphasized that a unitary category of fishing diverts attention from important issues, in particular, the social relations in which production is necessarily to be found. Although there are real empirical differences between fishing and other modes of subsistence, the use of such differences gives misplaced importance to ecology.

One way to understand similarities and differences among fishing systems, to emphasize the social context of production, is to distinguish between societies in terms of mode of circulation – the motivation of the producers and the destination of the products. Production may be primarily for exchange or use. Many fisheries conflicts derive from the different rationales of production of the household economy of local small-scale fishing and the market economy of industrial fleets (McGoodwin, 1990). Another way to distinguish fishing systems is to underline different forms of managing access to fish. While knowledge of ecology and fishing techniques is important for the understanding of different forms of managing access, one should not ignore the social space in which they occur. The informal exclusion of outsiders by means of secrecy and formal ownership of local territories should not be regarded as functionally equivalent devices. Differences among fishing economies in terms of modes of circulation of products and access to resources insufficiently represent the variety of production systems that exist. On the other hand, they help to illustrate the fundamental point that fisheries are embedded in social life. Any study of what happens on board fishing vessels must consider variations in fishing systems and the larger contexts of fisheries and fishing communities.

Gender Roles and Identities

In removing 'fishing' from the context of social relations, social scientists have often failed to recognize the importance of the relations of men and women and the significant economic role of women in fishing economies. The issue of gender tends to be suppressed or distorted. Nevertheless, in recent years an extensive literature on the role of gender in fishing and coastal communities has been developed (Nadel-Klein and Davis, 1988). In fishing communities, it has sometimes been suggested, one can expect a sharp division of labor along gender lines in which men do the fishing while women are responsible for the processing of the catch. A rigid sexual division of labor should not, however, be regarded as inevitable. In many cases, women's participation in extraction is no less important than that of men. The ethnography of fishing is biased with respect to gender. This observer's bias is partly a reflection of the popular Western image of fishing as a particularly 'manly' activity. Thus, the pursuit of small aquatic animals has sometimes been regarded as mere 'gathering' or 'collecting' if performed by women, but as 'fishing' if performed by men.

Many indigenous accounts of fishing economies are no less gender specific than observers' accounts, treating women's labor as not only different but also somehow secondary to that of men. While the devaluation of women's work is not restricted to market economies, the change from a subsistence economy to markets often has important implications for the social construction of gender. The 'domestic' labor of women tends to be presented as private labor, on the grounds that it takes place internal to the household and that its consumption products are 'perishable' use values, requiring immediate use. Conversely, only the nondomestic work of men is regarded as socialized since it alone involves cooperation between households, the production of exchange values, and durable commodities. Domestic laborers are presented as natural beings outside the social life of the market economy. And women only become social or visible insiders to the extent that they engage in nondomestic production, outside the sphere of reproduction. Women, then, tend to be presented as unproductive (Cole, 1991).

Just as extraction is a social activity whether or not it takes place in a group or by isolated individuals, labor is necessarily a social activity whether or not it takes place in a domestic or nondomestic context. The household is not a private place devoid of social content but a social institution constituted by a particular set of relations – the social relations of domestic production involved in the creation of use values, particularly cooking and child care.

Ecological Knowledge

Since fish are relatively incapable of coping with human predators, their chances of using evasive tactics are small. The difference in medium, however, has contradictory implications and some observers have emphasized the disadvantages to fishers, pointing out that if humans are invisible to the fish so are fish to humans. Accordingly, it is argued, knowledge of marine organisms is likely to be relatively restricted. Nevertheless, as several ethnographic accounts have shown, in fishing communities, people accumulate a vast amount of knowledge of fish and aquatic habitats. The social distribution and nature of ecological knowledge and its importance for resource management, on the other hand, are matters of some debate.

In fisheries generally, catches vary from one boat to another. Accounts of and interest in such differences vary from one context to another depending on the nature of the social system involved and the nature of the fishery. Some explanations emphasize differences in luck, others underline personal differences in fishing skill. Anthropological studies of fishing success have drawn attention to the continuity of the social world to which the skipper belongs, focusing on the importance of the crew for decision making and, by extension, relative fishing success. In a very real sense, the practical skills of fishing skippers are the result of mutual enskillment, the collective property of a community of practice (Palsson, 1994).

Not only has there been a growing theoretical interest in the practical learning of fishers and the general divide between indigenous and scientific knowledge, but also fisheries science itself is increasingly seen as an embedded enterprise (Smith, 1996). Some scholars argue that learning, whether in science or everyday life, primarily takes place with the internalization of a cognitive scheme. Others have emphasized that learning

involves the whole person interacting with the social and natural environment (Lave, 1990).

In many fishing societies, there is little attempt to draw upon the knowledge that practitioners have achieved in the course of their work for the purpose of marine ecological research. Analyses of human–environmental relations are the privileged task of people formally trained in public institutions. A greater reliance on practical knowledge, however, is a practical necessity, given the randomness and complexity of many marine ecosystems. Multispecies marine ecosystems, in particular, are highly unpredictable, with constant fluctuations in interactions among species and between species and their habitat (Acheson and Wilson, 1996). This does not mean that governance is impossible; it suggests, however, increasing reliance on a finer spatial and temporal scale of observation and engagement – a scale that only the skillful practitioner is able to apply. There are good grounds, therefore, for exploring how fishers acquire knowledge about the ecosystem within which they operate, how their knowledge differs from that of professional biologists, and to what extent the former could be brought more systematically into the process of management for the purpose of ensuring sustainability.

Resource Management

Many fisheries are currently in a state of crisis. Increasingly social scientists are addressing the problems involved, both internationally and in particular ethnographic contexts (McCay and Acheson, 1987). Two main positions can be identified, the market approach, which gained momentum in the early 1970s, and the approach of comanagement, which has received increased attention since the 1980s. Comparative studies of the implications and consequences of different management regimes represent an important and growing forum for social research.

The market approach assumes that the crises in fishing are due to the open access nature of fisheries; in the absence of both government control and the institution of private property, it is assumed, people will inevitably drive the ecosystem beyond its sustainable limits. The paradigm further suggests that only the 'hidden' forces of the market can ensure efficiency and sustainable use of resources. Finally, the issues of ethics and equity typically are presented as irrelevant theoretical distractions. In several fisheries in different parts of the world, fishing stocks are being turned into private property. First, the resource is appropriated by regional or national authorities and later on the total allowable catch for a season is divided into quota holdings and allocated to producers, often the owners of boats. At a still later stage, the temporary privileges of quota holders are turned into a marketable commodity as individual transferable quotas.

In some fisheries, the crew's contribution and its relevance for quota allocation has been a pressing political issue. Since fishing quotas are often allocated with reference to 'fishing history,' fishers, boat owners, and managers debate on what exactly makes fishing history, focusing on the relative importance of the vessel itself and its technology (favoring the owners of capital) and the crew (favoring labor power, in particular that of the skipper). A related contested issue involves the practical and theoretical merits of effort quotas (number of fishing trips) vis-à-vis catch quotas (size of catch).

The other main approach to fisheries management, comanagement (Jentoft, 1989), emphasizes democratic decision making and the importance of practical knowledge. A number of anthropological studies have shown that in many fishing societies people have developed indigenous means of regulating access to fishing grounds. In some cases, local groups of users successfully control the reproduction of renewable resources without external intervention, effectively comanaging local resources. In some parts of the world, especially Asia and the Pacific, comanagement has a long tradition (Ruddle and Johannes, 1985). Recently, such examples have attracted much attention among both researchers and managers, informed by renewed interest in 'indigenous' or 'traditional' ecological knowledge on the international scene (Dyer and McGoodwin, 1994) as well as growing doubts about the modernist concept of 'resource management,' with its assumptions of human mastery, control, and predictability.

Some of the growing body of works in this vein draws upon the 'institutional' scholarship of Elinor Ostrom. Emphasizing the context of scarce natural resources, including fisheries, Ostrom (2005) presents a framework for discussing the development and use of collective action with respect to environmental problems and common-pool resources, an alternative to both the governance of the nation state and the neoliberal solution of private property and the market. Drawing upon the 'new institutionalism' developed by Douglas North and some others, she underlines the impact and interrelations of social institutions (what anthropologists and sociologists would be likely to refer to as cultural or social context), comparing cases and contexts in order to generalize about the 'building blocks' of institutions.

The oceans have often been presented as a virtual showcase for radical experimental schemes in the terrestrial domain, for fleshing out central issues relating to 'tragedies' of the commons and the inevitability of privatization and neoliberal politics. For many economists, in particular, the fisheries have provided illuminating demonstrations of the need for 'final' market-based solutions to problems of access, sustainability, and governance, much like the solutions of state socialism decades earlier. For others, however, the neoliberalization of fishing stocks has paved the way for virtual economies and, eventually, the current global financial crisis. Several anthropologists and sociologists have critically examined particular management schemes emphasizing contrasting promises and broad community implications (Pinkerton and Edward, 2009).

Maritime Anthropology Coming of Age

Traditionally, anthropologists have focused on isolated communities. Modern anthropology, however, provides useful perspectives on cultural flows and the link between micro- and macrolevels. Recently, anthropological students of fishing societies have increasingly moved from the narrow context of face-to-face encounters and local communities to the nation state and beyond. Thus, several anthropological studies have drawn attention to the ways in which local and professional identities are constructed within larger contexts. Maritime

anthropology developed in the 1970s onward, along with anthropology at home, applied anthropology, and a series of other developments that redefined the sites, methods, politics, and subject matter of anthropological research.

Early on, maritime anthropology was heavily focused on the community at sea, navigation and wayfinding, fishing skills, competition for space, crews, and fleets. Later on the discussion broadened to include fishing communities, fish processing and marketing, property rights, and resource management. In the process, anthropologists carved an important professional niche for themselves in administration, development, and governance. Anthropological analyses and ethnographic insights proved to be quite useful for charting alternative regimes of resource use, for mitigating change, and for settling disputes relating to rights and access.

To the extent that Helmreich's (2009) work on the 'alien' ocean belongs to maritime anthropology, it takes a radical course back to the sea, into the depths of the ocean and the microworlds of cellular life. On the one hand, he explains, the ocean has been domesticated through immense human effort, brought into the human domain, charted, and analyzed with respect to chemistry, food webs, life forms, etc.; an interesting recent development on this score is the addition of the ocean floor, more than half of the globe's surface, to the area covered by Google Earth – an achievement underlining the conquest of the last cartographic frontier on the globe, allowing Internet explorers wherever they are based to virtually navigate the ocean bed. But on the other hand, the ocean remains a wild domain, saturated with life at the extremes in terms of heat, pressure, fleeting boundaries, and genomic configuration, full of strange species and alien genes, traveling between microbial lineages.

In recent years, the notion of the tree of life, a central notion for Darwin, Linnaeus, and much of biological thought since their time, has been seriously challenged by a strange mix of scholars. For one thing, it has been reasoned, the tree of life might turn out to be a net or a rhizome, with endless reconnections rather than treelike bifurcation. Also, this makes the oceans centrally relevant for biological theorizing, and the fleeting nature of biological processes in the aquatic domain may facilitate lateral gene transfer between organisms and species. 'Oceans,' Helmreich (2009: p. 280) suggests, 'are good to think with about life.' While it is possible that theoretical challenges to the tree model and the empirical evidence available only minimally blur the main picture, simply shaking the tree a bit, they raise fundamental questions as to what should count as relatedness. Maritime anthropology has come of age.

See also: Apprenticeship, Anthropological Aspects; Body: Anthropological Aspects; Cognitive Anthropology; Economic Anthropology; Food in Anthropology; Gender: Anthropological Aspects; Hunting and Gathering Societies: Anthropology; Modes of Production; Practice: An Anthropological Approach; Sociology and the New Institutionalism.

Bibliography

Acheson, J.M., Wilson, J.A., 1996. Order out of chaos: the case for parametric fisheries management. American Anthropologist 98, 579–594.

Cole, S., 1991. Women of the Praia: Work and Lives in a Portuguese Coastal Community. Princeton University Press, Princeton, NJ.

Dyer, C.L., McGoodwin, J.R. (Eds.), 1994. Folk Management in the World's Fisheries: Lessons for Modern Fisheries Management. University Press of Colorado, Niwot, CO.

Helmreich, S., 2009. Alien Ocean: Anthropological Voyages in Microbial Seas. University of California Press, Berkeley, CA.

Hewes, G., 1948. The rubric 'fishing and fisheries'. American Anthropologist 50, 238–246.

Jentoft, S., 1989. Fisheries co-management: delegating government responsibility to fishermen's organizations. Marine Policy 13 (2), 137–154.

Kelly, R.L., 1992. Mobility/sedentism: concepts, archaeological measures, and effects. Annual Reviews of Anthropology 21, 43–66.

Lave, J., 1990. The culture of acquisition and the practice of understanding. In: Stigler, J.W., Shweder, R.A., Herdt, G. (Eds.), Cultural Psychology: Essays on Comparative Human Development. Cambridge University Press, Cambridge, UK.

McCay, B.M., Acheson, J.M. (Eds.), 1987. The Question of the Commons: The Culture and Ecology of Communal Resources. University of Arizona Press, Tucson, AZ.

McGoodwin, J.R., 1990. Crisis in the World's Fisheries: People, Problems, and Policies. Stanford University Press, Stanford, CA.

Nadel-Klein, J., Davis, D.L. (Eds.), 1988. To Work and to Weep: Women in Fishing Economies. Institute of Social and Economic Research, St. John's, Newfoundland.

Ostrom, E., 2005. Understanding Institutional Diversity. Princeton University Press, Princeton, NJ.

Palsson, G., 1991. Coastal Economies, Cultural Accounts: Human Ecology and Icelandic Discourse. Manchester University Press, Manchester, UK.

Palsson, G., 1994. Enskilment at sea. Man. The Journal of the Royal Anthropological Institute 29 (4), 901–927.

Pinkerton, E., Edward, D.N., 2009. The elephant in the room: the hidden costs of leasing individual transferable fishing quotas. Marine Policy 33, 707–713.

Ruddle, K., Johannes, R.E., 1985. The Traditional Ecological Knowledge and Management of Coastal Systems in Asia and the Pacific. UNESCO, Jakarta, Indonesia.

Smith, M.E., 1996. Public policy, sciencing, and managing the future. In: Nader, L. (Ed.), Naked Science: Anthropological Inquiry into Boundaries, Power, and Knowledge. Routledge, London.

Five Factor Model of Personality, Assessment of

Michael J Boudreaux and Daniel J Ozer, University of California Riverside, Riverside, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Five Factor Model (FFM) is a hierarchical taxonomy of personality traits. At the superordinate level are five factors labeled Extraversion, Agreeableness, Conscientiousness, Emotional Stability (vs Neuroticism), and Intellect (or Openness). Below this level are several more specific personality traits that are summarized by these five higher-order dimensions. In this article, we briefly describe the lexical and questionnaire approaches that led to the development of the FFM, and review measures that derive from these two traditions, measuring both higher- and lower-order traits. In addition, we discuss abbreviated instruments, measures that assess variants of the FFM, an open source personality item pool, and a structured interview. Future research will improve upon FFM assessment by examining the five factor structure across languages and developmental periods, and detailing the structure of personality traits at lower levels of the trait hierarchy.

The Five Factor Model (FFM) is a hierarchical taxonomy of personality traits organized around five broad dimensions. While the lower levels of this taxonomy are only partially explored, the highest order level is composed of the superordinate factors labeled Extraversion, Agreeableness, Conscientiousness, Emotional Stability (vs Neuroticism), and Intellect (or Openness). Given the comprehensiveness of the FFM, one could understand any assessment of personality traits as an assessment of at least a portion of the FFM. For example, the Woodworth Personal Data Sheet, published in 1920, perhaps the first modern standardized self-report measure of personality, is an assessment of Neuroticism. However, our focus here is on more contemporary measures of personality traits that explicitly embrace or are close variants of the FFM.

Evidence supporting the FFM is necessarily linked to assessment efforts. The model arose out of the assessment tradition (see Wiggins, 1973, for a description of that tradition just as the FFM was beginning to emerge), and measures of the five factors were developed to improve, elaborate, supplant, or shorten prior trait measures. The consensus that emerged among personality trait assessors, that the FFM might indeed possess substantial generality, arose from two largely separate but converging research programs: the lexical and questionnaire approaches. The goal of the lexical approach was to identify the structure of self- and observer-rated trait descriptive adjectives, whereas the questionnaire approach sought to identify the common trait dimensions underlying extant multiscale personality inventories.

The term 'Five Factor Model' derives from the questionnaire tradition and is exemplified by the work of Paul Costa and Robert McCrae. Those working in the lexical tradition, including Warren Norman and Lewis Goldberg, have preferred the label 'Big Five.' One superficial difference between these traditions is in the naming of the fourth factor, 'Emotional Stability' (Big Five) or 'Neuroticism' (FFM). A more telling difference is in the conceptualization of the fifth factor as 'Intellect' (Big Five) or the broader 'Openness to Experience' (FFM). There is some movement toward a preference for the shorter label of 'Openness,' which suggests not only intellectual but also imaginative and affective components as well.

Below, we describe various English language measures of the five factors that evolved out of these two traditions, as well as efforts to create abbreviated Big Five/FFM measures. While there has been considerable consensus surrounding the five factors, there are close variants of six- and seven-factor measures that merit attention. Finally, we discuss an open source personality item pool, a structured interview for an FFM assessment, and studies that have examined the comparative validity of Big Five/FFM measures.

Goldberg and the Lexical Approach

The lexical hypothesis asserts that the most salient and socially relevant ways in which people differ from one another will eventually become encoded as words. Allport and Odbert's catalog of English language trait descriptive terms, published in 1936, provided a starting point for studying the structure of personality trait adjectives. In subsequent years, various investigators examined the factor structure of these terms in both self- and observer reports. As early as 1963, Warren Norman suggested that the reappearance of the same factors across different studies and samples suggested that five factors might provide an adequate taxonomy of personality traits, though this solution was not named the 'Big Five' until 1981, by Lewis Goldberg.

In the early 1980s, Goldberg initiated a series of studies designed to verify the five-factor structure. He returned to the lexicon of personality and constructed an inventory of 1710 trait descriptive adjectives. He replicated the Big Five factor structure across a variety of samples and factor analytic methods, and subsequently developed several abbreviated marker scales for the Big Five. In one of these (Goldberg, 1990; Study 2), he reduced the terms to 479, grouped into 133 clusters. Then, in study 3, he removed the least homogenous items from each cluster (i.e., those with the lowest item-total correlations), as well as single-item categories, resulting in a set of 100 clusters based on 339 trait adjectives. The 100 clusters and adjectives defining each cluster are available in Goldberg (1990; Table 3).

The 100 synonym clusters have several merits. Most notable is the balance between bandwidth and fidelity. The clusters

 http://dx.doi.org/10.1016/B978-0-08-097086-8.25061-7

sample a broad range of personality characteristics, and each provides specific information that can be used for predictive purposes. However, the use of 100 synonym clusters is still time-consuming, so Goldberg (1992) developed two shorter measures: 50 bipolar rating scales (e.g., timid vs bold, rude vs polite, angry vs calm) and 100 unipolar markers (e.g., energetic, cooperative, relaxed). The rationale for using paired terms is that they can help clarify the ambiguity of some adjectives; however, Goldberg (1992) found that they are less robust across samples than factor markers presented in unipolar format. The bipolar rating scales are now less often used in research, but have been used in educational settings to demonstrate the nature of the Big Five (see Goldberg, 1992; Table 1).

Goldberg's (1992) 100 unipolar markers is an efficient and widely used measure of the Big Five. It contains 20 items per factor, with 10 items representing each pole of each factor (with the exception of Emotional Stability, which has 6 positive and 14 negative terms). From an initial pool of 566 terms, items were selected with high loadings on targeted factors and low loadings on other factors. Then, those items that maximized internal consistency and had a replicable factor structure were selected into the final list. Goldberg showed each of the five 20-item scales to be highly reliable (alpha coefficients above 0.80) and to converge with larger and more representative sets of trait adjectives, as well as with other inventories. The 100 unipolar markers are available in Goldberg (1992; Appendix A), and a 40-item, abbreviated version, the Big Five Mini-Markers, is available in Saucier (1994; Appendix). More recently, Saucier (2002) developed three sets of marker scales designed to reduce the interscale correlations often observed within Big Five measures, called the Ortho-Markers (Ortho-40), Modular Markers, and Mini-Modular Markers-40 (3M40).

Costa and McCrae and the Questionnaire Tradition

Numerous structured questionnaires to assess personality have been developed, with highly varied content and scales. Among these instruments was the NEO, developed by Paul Costa and Robert McCrae in 1980, which measured the three domains of Neuroticism, Extraversion, and Openness to Experience, and included 18 facet scales (6 per domain). Observing the findings reported by studies of the lexical structure noted above, they added (in 1985) scales to assess Agreeableness and Conscientious, and later developed (in 1992) facet scales to assess these domains. In a series of studies, Costa and McCrae found that the scales of some of the most widely known and frequently used personality measures could be placed within a five-factor framework. To be clear, their claim was not that all inventories measured all five factors (this was clearly not the case), but rather that inventory scales were largely related to the five (or a subset of the five); and scales that were outside the five-factor space were largely idiosyncratic to each inventory.

The Revised NEO Personality Inventory (NEO-PI-R; Costa and McCrae, 1992) is a 240-item questionnaire designed to measure the five major personality trait domains. Within each domain are six facet scales, with eight items per facet. The items are simple sentences that describe behaviors, preferences, and attitudes, and are responded to on a 5-point, Likert-type scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). The NEO-PI-R comes in two forms: Form S, for self-reports, and Form R, for observer ratings. Form R contains the same items as Form S, but is written in the third person. A more recent version of the NEO (NEO-PI-3, McCrae and Costa, 2010) was made to be applicable to a wider range of individuals, including school-aged children and respondents with lower levels of education. The NEO-PI-3 is thus appropriate for individuals who are 12 years of age or older.

The NEO-PI-R enjoys more than 30 years of research to support its reliability and validity. The scales are internally consistent, with alpha reliabilities for the domain scales in the range of 0.86–0.92 ($Mdn. = 0.89$), and for the facet scales in the range of 0.56–0.81 ($Mdn. = 0.76$). Scores show long-term stability, with retest values for the domain scales ranging 0.63–0.83 over 3–6 years. Six-year retest coefficients in spouse ratings, using Form R, are similar to those found in self-reports. The scales have also shown evidence of convergent and discriminant validity across instruments and observers. The NEO-PI-R has a clear internal structure, though some of the facet scales have large secondary loadings (i.e., >0.40) which contribute to modest correlations among the five NEO-PI-R domain scores. To obtain truly independent domain scores, factor score weights can be applied to each of the 30 facets (provided in the NEO-PI-R professional manual).

The Revised NEO Five-Factor Inventory (NEO-FFI-R; McCrae and Costa, 2004) is an abbreviated version of the NEO-PI-R. It consists of 60 items to measure the five global dimensions of the FFM, with 12 items per scale. In the original version of the instrument, items were selected based on their correlations with factor scores from the NEO-PI. The revised version replaced 14 of the original items, thus improving its factor structure, reliability, and validity. In an even more recent version, the NEO Five-Factor Inventory-3 (NEO-FFI-3; McCrae and Costa, 2010), one additional item was replaced. All items correspond with the NEO-PI-3. McCrae and Costa (2007) reported excellent psychometric characteristics of the NEO-FFI-3 in adolescent, adult, and middle-school samples.

McCrae and Costa (2007) also developed a brief measure of all 30 facet scales, corresponding to the first 120 items of the NEO-PI-3, the NEO-PI-3 First Half (NEO-PI-3FH). For those who desire a differentiated yet brief measure of the FFM, the NEO-PI-3FH provides a flexible alternative to both global and longer multitrait inventories.

The NEO Inventories have been used extensively in research and applied contexts as diverse as counseling and clinical psychology, behavioral medicine, industrial-organizational psychology, and educational psychology (Costa and McCrae, 1992), and are commercially available through their publisher, Personality Assessment Resources.

Brief Measures of the Big Five

As personality assessors increasingly regarded the five superordinate factors as an adequate summary view of personality traits, there arose an interest to include measures of these traits in various research contexts – contexts where the inclusion of a number of other measures led to a desire for a brief measure

of the Big Five. These short measures are now widely used in a number of research settings.

Big Five Inventory – 44-Item Version

The Big Five Inventory (BFI; John et al., 1991) is a brief, 44-item self-report inventory designed to measure the Big Five personality trait dimensions. The items are in the form of short phrases, written in the third person, and based on prototypical trait markers of the Big Five with elaborative or clarifying information added (e.g., "is original, comes up with new ideas," "is relaxed, handles stress well"). Each statement uses the stem, "I see myself as someone who …" Respondents indicate the extent to which they agree or disagree with each statement on a 5-point, Likert-type scale ranging from 1 (*disagree strongly*) to 5 (*agree strongly*).

The BFI scales include 8 to 10 items, and have very good psychometric properties: The scales are highly reliable (alpha coefficients range from 0.75 to 0.90, with an average above 0.80), stable over time (3-month retest coefficients range from 0.80 to 0.90, with a mean of 0.85), and possess convergent and discriminant validity with respect to other Big Five instruments (John et al., 2008). The BFI is widely used, and is reprinted in John et al. (2008; see Appendix 4.1).

Big Five Inventory – 10-Item Version

The Big Five Inventory-10 (BFI-10; Rammstedt and John, 2007) is an abbreviated version of the BFI-44. It consists of two items per scale, representing both the high and low poles of each factor. The response format is identical to the full BFI, although the instructions have been shortened. Items were selected based on their item-total correlations with the full BFI scales, as well as their factor loadings in analyses of all 44 items. Given the brevity of this measure, it holds up remarkably well across US and German samples: The two-item scales have an average correlation of 0.83 with the full scales, are stable over 6–8 weeks (average retest stability coefficient is 0.75), and have an average self-peer convergent validity correlation of 0.44. The BFI-10 scales possess substantial convergent and discriminant validity with the domain and facet scales of the NEO-PI-R, and the items show a clear five-factor structure. The BFI-10 takes about 1 min to complete. The items and instructions are provided in Rammstedt and John (2007; see Appendix A).

Ten-Item Personality Inventory

The Ten-Item Personality Inventory (TIPI; Gosling et al., 2003) is another short measure of the Big Five. Each item consists of paired descriptors (e.g., extraverted, enthusiastic) that use the common stem, "I see myself as:" Respondents indicate the extent to which they agree or disagree with each item on a 7-point, Likert-type scale ranging from 1 (*disagree strongly*) to 7 (*agree strongly*). In a large student sample, the psychometric properties of the TIPI reached satisfactory levels across three criteria: Scores converged with widely used Big Five measures, evinced substantial 60 week test–retest correlations (mean $r = 0.72$), and exhibited patterns of correlations with a broad range of other constructs that were comparable in size to those exhibited by the BFI in the same sample. The TIPI takes about 1 min to complete, and is provided in Gosling et al. (2003; see Appendix A).

Alternative Representations of Trait Structure

The questionnaires discussed to this point are widely understood to be different measures of the same five broad traits. The measures discussed below each embrace several features of the Big Five/FFM framework, while at the same time introduce important variations.

Abridged Big Five Dimensional Circumplex

The Abridged Big Five Dimensional Circumplex (AB5C; Hofstee et al., 1992) represents an integration of Big Five and Circumplex approaches to trait structure. The model recognizes that trait descriptors rarely show simple structure, but tend to show cross-loadings on two or more factors. It 'abridges' a full five-dimensional model, where each trait would load on all five factors, to a two-dimensional model, where each trait loads on two factors. Consisting of 10 two-dimensional circumplexes, with two of the five factors serving as the axes of each circle, the AB5C model has 45 facet scales (based on 540 trait terms) representing 'blends' of the two underlying factors (e.g., *sociable* (I+II+) corresponds to high Extraversion and high Agreeableness, and *dominant* (I+II−) corresponds to high Extraversion and low Agreeableness).

Big Seven Factor Model

The Big Seven Factor Model, and its associated measure, the Inventory of Personal Characteristics (IPC-7; Tellegen et al., 1991), specify seven higher-order factors: Positive Valence, Negative Valence, Positive Emotionality, Negative Emotionality, Conscientiousness, Agreeableness, and Conventionality. This model departs from the traditional Big Five in two important ways: First, by including in its item pool evaluative and state descriptors of personality, two primary dimensions of self-evaluation, Positive Valence (e.g., excellent, special, impressive) and Negative Valence (e.g., evil, wicked, awful), emerge in factor analyses in addition to the Big Five. Second, the factors labeled Positive Emotionality (Extraversion) and Negative Emotionality (Neuroticism) emphasize the emotional aspects of their Big Five counterparts, and Openness is reinterpreted as (low) Conventionality.

Waller (1999) reviewed Big Five and Big Seven models, and described the psychometric properties of the IPC-7. Benet and Waller (1995) demonstrated the cross-cultural generalizability of the IPC-7 in a Spanish sample using an abbreviated 70-item version. American and Spanish versions of the abbreviated IPC-7 are available in Benet and Waller (1995; Appendix).

HEXACO Personality Inventory – Revised

The Revised HEXACO Personality Inventory (HEXACO-PI-R; Lee and Ashton, 2004) is a 200-item questionnaire designed to measure six factors found in lexical studies of personality trait structure: Honesty-Humility (H), Emotionality (E),

Extraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O). The HEXACO factors are alternative rotations of the Big Five, and are defined by 24 facets (four per factor) within the HEXACO-PI-R. The major difference between the models is the separate dimension of Honesty-Humility within the HEXACO model, defined by sincerity, fairness, lack of greed, and modesty. Additionally, Emotionality (i.e., Emotional Stability) and Agreeableness (as well as Extraversion to a slight degree) have been reinterpreted. The HEXACO-PI-R (100- and 60-item versions) and a list of references are available on the HEXACO web site (http://hexaco.org).

Hogan Personality Inventory

The Hogan Personality Inventory (HPI; Hogan and Hogan, 1995) was designed to assess the FFM from the perspective of socioanalytic theory. According to socioanalytic theory, the FFM represents personality from the view of the observer in terms of a person's ability to 'get along' with other group members and to 'get ahead' to earn social status. The HPI assesses individual differences in a person's ability to attain these goals. It contains seven primary scales called, along with their FFM counterpart, Adjustment (Emotional Stability), Ambition and Sociability (Extraversion), Likeability (Agreeableness), Prudence (Conscientiousness), and Intellectance and School Success (Openness).

The HPI consists of 206 true-false items, grouped into 41 'homogenous item clusters' that define the seven primary scales. Hogan and Hogan (1995) reported the scales to be internally consistent (alpha reliabilities for the primary scales range from 0.71 to 0.86), to converge with other established Big Five measures, and to predict a variety of job performance criteria. The HPI has been used primarily in occupational settings, and is commercially available through Hogan Assessment Systems.

Revised Interpersonal Adjective Scales: Big Five

The Revised Interpersonal Adjective Scales: Big Five (IASR-B5; Trapnell and Wiggins, 1990) is an extension of the Revised Interpersonal Adjective Scales, a measure of the interpersonal circumplex. In addition to the interpersonal dimensions of Dominance (or Agency) and Nurturance (or Communion), the IASR-B5 includes marker scales for Neuroticism, Conscientiousness, and Openness to Experience. Although Dominance and Nurturance generally correspond to Extraversion and Agreeableness, the latter two dimensions are broader in content to include non-interpersonal aspects of personality.

The IASR-B5 consists of 124 adjectives rated on an 8-point, Likert-type scale ranging from 1 (*extremely inaccurate*) to 8 (*extremely accurate*). The first 64 items measure the Dominance and Nurturance dimensions of the interpersonal circle, and can be used to score the eight octants of the Interpersonal Adjective Scales. The remaining 60 items measure Neuroticism, Conscientiousness, and Openness to Experience (i.e., 20 items per factor, with 10 items marking the positive and 10 items marking the negative poles). Trapnell and Wiggins (1990) reported the IASR-B5 to have a clear internal structure at the item level, and discriminant and convergent validity when compared with the NEO and Hogan personality inventories.

Measures Derived from the International Personality Item Pool

The International Personality Item Pool (IPIP; Goldberg, 1999) is a repository of over 2000 personality items that have been used to develop a number of personality inventories, including measures of the FFM. Freely available in the public domain (http://ipip.ori.org), many of these measures serve as proxies for constructs measured by commercially available inventories, such as the NEO-PI-R and HPI, as well as nonproprietary instruments, such as the HEXACO-PI, the AB5C, and 100 unipolar markers. Tables comparing the psychometric characteristics of the original scales with the IPIP proxies can be found on the IPIP web site. As Goldberg et al. (2006) noted, alpha reliabilities of these scales generally match or exceed those of the original scales, and the IPIP scales correlate highly with their parent scales.

A noteworthy measure derived from the IPIP item pool is the Big Five Aspect Scales (BFAS; DeYoung et al., 2007), a 100-item inventory designed to assess two distinct (but correlated) 'aspects' of each factor (i.e., trait variables that occupy an intermediate level of personality structure between facets and domains). Item selection was based on correlating all of the IPIP items with scores derived from factor analyses of all 75 NEO-PI-R and AB5C-IPIP (Goldberg, 1999) facet scales. Then, in an independent sample, the 20 items showing the clearest two-factor solution within each domain were selected. The ten 10-item aspect scales demonstrated good psychometric characteristics, and were associated with factors derived from studies in behavioral genetics.

Structured Interview

Self-reports are the predominant source of personality data. However, alternative modes of assessment are sometimes desired. Although several observer rating forms are available for FFM measures, the instruments are largely the same as those used in the self-report versions with minor changes in wording and instructions to reflect the stance of the observer. One novel procedure for assessing the five factors utilizes a structured interview approach using a standard set of questions and scoring procedure. Mental health practitioners often prefer interview methods because they allow for open-ended, follow-up questions and an opportunity to observe the respondent's behavior during an assessment.

The Structured Interview for the Five-Factor Model of Personality (SIFFM; Trull and Widiger, 1997) is a 120-item semistructured interview that assesses both adaptive and maladaptive aspects of the FFM. It assesses the five higher-order dimensions (with 24 items per dimension) and 30 lower-order facet traits (with 4 items per facet), corresponding to the NEO-PI-R. The SIFFM places relatively greater emphasis on potentially maladaptive aspects of personality, and was explicitly designed to be used in clinical settings. The SIFFM items (i.e., interview questions) are scored as either 0 (*absent*), 1 (*present and does not result in significant dysfunction*), or 2

(*present and may result in significant dysfunction*). Follow-up questions may or may not be asked depending on the response to the initial question. The SIFFM takes about 1 h to administer.

In samples of undergraduates and outpatients receiving treatment at a community clinic, Trull et al. (1998) showed scores on the SIFFM to be internally consistent, stable across 2 weeks, and to converge with scores on the NEO-PI-R domain and facet scales and peer-ratings on the NEO-FFI. Trull et al. also demonstrated incremental validity of the SIFFM over the NEO-PI-R in predicting 10 of 13 personality disorders. Given its relative emphasis on maladaptive aspects of traits associated with the FFM, the SIFFM is primarily used by clinical researchers and practitioners to assess personality and personality dysfunction. The SIFFM is commercially available through its publisher, Personality Assessment Resources.

Comparative Validity of FFM Personality Inventories

Relatively few studies have compared the psychometric characteristics of FFM personality inventories, but some studies are beginning to appear. John and Soto (2007) compared the reliability and validity of the NEO-FFI, the BFI, and Saucier's (1994) 40-item version of Goldberg's (1992) 100 Trait Descriptive Adjectives (TDA). All three instruments showed relatively similar internal consistency coefficients, though some differences in convergent validity were noted for the NEO Extraversion and Openness scales. Profiles of the Big Five domains showed differences in terms of how the three measures define each factor with respect to the NEO-PI-R facet scales. As a result, patterns of intercorrelations vary somewhat across instruments, although the average discriminant within-instrument correlation is quite low, with an absolute value of 0.19 overall.

Grucza and Goldberg (2007) compared the validity of 11 personality inventories in predicting three criteria: frequencies of self-reported behavioral acts, informant reports, and clinical indicators. Among the instruments they reviewed, five were designed to assess the FFM – the NEO-PI-R, HPI, HEXACO-PI, TDA, and the IPIP-AB5C. Across the three criteria, the average multiple correlation coefficient fell between a narrow range of 0.42 (for the TDA) and 0.45 (for the NEO-PI-R), suggesting little difference in overall predictability for the instruments, though there were important differences for predicting particular criteria. Similarly, Saucier (2002) reported no overall predictive differences when comparing questionnaire measures (NEO-PI-R, NEO-FFI, BFI) and a variety of adjective-based marker scales (Mini-Markers, Ortho-40, Modular Markers, 3M40). Other studies have shown that narrower trait measures, such as the facet scales included in some of these instruments, can improve the prediction of behavior above and beyond the global domains (e.g., Paunonen and Ashton, 2001).

Conclusion

As reviewed above, numerous English-language measures of the five factors are available, varying in length and specificity. In addition, Big Five/FFM adjective- and phrase-based measures have been developed within other cultures and languages, several of which are reviewed in De Raad and Perugini (2002). Furthermore, measures designed to assess the Big Five/FFM in children have been developed. The adequacy of the five-factor framework to account for personality trait differences across languages and developmental periods remains to be fully explored, as does a taxonomy of trait structure that fully describes more specific personality traits. Research that extends the limits and explores the details of this hierarchical taxonomy will inevitably lead to new and presumably improved personality measures.

See also: Agreeableness; Big Five Factor Model, Theory and Structure; Conscientiousness; Extraversion; Factor Analysis and Latent Variable Models in Personality Psychology; Five Factor Model of Personality, Facets of; Five Factor Model of Personality, Personality Disorder; Five Factor Model of Personality, Universality of; Honesty and Humility; Neuroticism; Openness to Experience; Personality Assessment: Overview; Personality Assessment; Personality, Trait Models of; Personality: Historical and Conceptual Perspectives.

Bibliography

Benet, V., Waller, N.G., 1995. The Big Seven factor model of personality description: evidence for its cross-cultural generality in a Spanish sample. Journal of Personality and Social Psychology 69, 701–718. http://dx.doi.org/10.1037//0022-3514.69.4.701.

Costa Jr., P.T., McCrae, R.R., 1992. Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual. Psychological Assessment Resources, Odessa, FL.

De Raad, B., Perugini, M. (Eds.), 2002. Big Five Assessment. Hogrefe & Huber, Seattle, WA.

DeYoung, C.G., Quilty, L.C., Peterson, J.B., 2007. Between facets and domains: 10 aspects of the Big Five. Journal of Personality and Social Psychology 93, 880–896. http://dx.doi.org/10.1037/0022-3514.93.5.880.

Goldberg, L.R., 1990. An alternative "description of personality": the Big-Five factor structure. Journal of Personality and Social Psychology 59, 1216–1229. http://dx.doi.org/10.1037//0022-3514.59.6.1216.

Goldberg, L.R., 1992. The development of markers for the Big-Five factor structure. Psychological Assessment 4, 26–42. http://dx.doi.org/10.1037//1040-3590.4.1.26.

Goldberg, L.R., 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In: Mervielde, I., Deary, I., Fruyt, F., Ostendorf, F. (Eds.), Personality Psychology in Europe, vol. 7. Tilburg University Press, Tilburg, The Netherlands, pp. 7–28.

Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R., Gough, H.G., 2006. The international personality item pool and the future of public-domain personality measures. Journal of Research in Personality 40, 84–96. http://dx.doi.org/10.1016/j.jrp.2005.08.007.

Grucza, R.A., Goldberg, L.R., 2007. The comparative validity of 11 modern personality inventories: predictions of behavioral acts, informant reports, and clinical indicators. Journal of Personality Assessment 89, 167–187. http://dx.doi.org/10.1080/00223890701468568.

Gosling, S.D., Rentfrow, P.J., Swann Jr., W.B., 2003. A very brief measure of the Big-Five personality domains. Journal of Research in Personality 37, 504–528. http://dx.doi.org/10.1016/S0092-6566(03)00046-1.

Hofstee, W.K.B., de Raad, B., Goldberg, L.R., 1992. Integration of the Big Five and circumplex approaches to trait structure. Journal of Personality and Social Psychology 63, 146–163. http://dx.doi.org/10.1037//0022-3514.63.1.146.

Hogan, R., Hogan, J., 1995. The Hogan Personality Inventory Manual. Hogan Assessment Systems, Tulsa, OK.

John, O.P., Donahue, E.M., Kentle, R.L., 1991. The Big Five Inventory – Versions 4a and 54. University of California, Berkeley, Institute of Personality and Social Research, Berkeley, CA.

John, O.P., Naumann, L.P., Soto, C.J., 2008. Paradigm shift to the integrative Big Five trait taxonomy. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research, third ed. Guilford Press, New York, pp. 114–158.

John, O.P., Soto, C.J., 2007. The importance of being valid: reliability and the process of construct validation. In: Robins, R.W., Fraley, R.C., Krueger, R.F. (Eds.), Handbook of Research Methods in Personality Psychology. Guilford Press, New York, pp. 461–494.

Lee, K., Ashton, M.C., 2004. Psychometric properties of the HEXACO personality inventory. Multivariate Behavioral Research 39, 329–358. http://dx.doi.org/10.1207/s15327906mbr3902_8.

McCrae, R.R., Costa Jr., P.T., 2004. A contemplated revision of the NEO Five-Factor Inventory. Personality and Individual Differences 36, 587–596. http://dx.doi.org/10.1016/S0191-8869(03)00118-1.

McCrae, R.R., Costa Jr., P.T., 2007. Brief versions of the NEO-PI-3. Journal of Individual Differences 28, 116–128. http://dx.doi.org/10.1027/1614-0001.28.3.116.

McCrae, R.R., Costa Jr., P.T., 2010. NEO Inventories Professional Manual. Psychological Assessment Resources, Lutz, FL.

Paunonen, S.V., Ashton, M.C., 2001. Big Five factors and facets and the prediction of behavior. Journal of Personality and Social Psychology 81, 524–539. http://dx.doi.org/10.1037/0022-3514.81.3.524.

Rammstedt, B., John, O.P., 2007. Measuring personality in one minute or less: a 10-item short version of the Big Five Inventory in English and German. Journal of Research in Personality 41, 203–212. http://dx.doi.org/10.1016/j.jrp.2006.02.001.

Saucier, G., 1994. Mini-Markers: a brief version of Goldberg's unipolar Big-Five markers. Journal of Personality Assessment 63, 506–516. http://dx.doi.org/10.1207/s15327752jpa6303_8.

Saucier, G., 2002. Orthogonal markers for orthogonal factors: the case of the Big Five. Journal of Research in Personality 36, 1–31. http://dx.doi.org/10.1006/jrpe.2001.2335.

Tellegen, A., Grove, W.M., Waller, N.G., 1991. Inventory of Personal Characteristics #7 (IPC-7). Unpublished materials. University of Minnesota.

Trapnell, P.D., Wiggins, J.S., 1990. Extension of the Interpersonal Adjective Scales to include the Big Five dimensions of personality. Journal of Personality and Social Psychology 59, 781–790. http://dx.doi.org/10.1037//0022-3514.59.4.781.

Trull, T.J., Widiger, T.A., 1997. Structured Interview for the Five-Factor Model of Personality (SIFFM): Professional Manual. Psychological Assessment Resources, Odessa, FL.

Trull, T.J., Widiger, T.A., Useda, J.D., Holcomb, J., Doan, B.T., Axelrod, S.R., Stern, B.L., Gershuny, B.S., 1998. A structured interview for the assessment of the five-factor model of personality. Psychological Assessment 10, 229–240. http://dx.doi.org/10.1037//1040-3590.10.3.229.

Waller, N.G., 1999. Evaluating the structure of personality. In: Cloninger, C.R. (Ed.), Personality and Psychopathology. American Psychiatric Press, Washington, DC.

Wiggins, J.S., 1973. Personality and Prediction: Principles of Personality Assessment. Addison-Wesley, Reading, MA.

Five Factor Model of Personality, Facets of

Ryan Y Hong, National University of Singapore, Singapore
Sampo V Paunonen, University of Western Ontario, London, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In a hierarchical model of personality trait structure, a facet is defined as a narrow constituent trait that is subsumed under a broad factor. Although there are debates about the number and nature of these facets, they should not diminish the usefulness of such traits in predicting and explaining behavior. Compared to the broad factors of personality, not only do facets provide greater predictive validity of outcome criteria, they are also more helpful in identifying the precise mechanisms involved in explaining psychological phenomena. A few examples on how analyzing personality data at the facet level can yield novel and important findings are presented.

Personality psychologists have long espoused a hierarchical view of personality structure in which trait variables are arranged into several levels representing varying breadth of sampled behavior domains. Eysenck (1947) was probably the first theorist to formalize such a structure to personality, as shown in **Figure 1**. The lowest level of the hierarchy is represented by behavioral responses enacted in specific instances. Regular occurrence of a behavioral response across situations and time combines to form a habitual response, which represents the next level up in the hierarchy. Several habitual response propensities constitute a personality facet or trait at the third level. Finally, the top of the personality hierarchy represents several correlated personality facets that are connected to form the broad factors or dimensions of personality (Goldberg, 1993; Paunonen and Hong, 2015).

On Factors and Facets

The five factor (or the Big Five) model of personality (see John et al., 2008, for a review; and the following articles: Five Factor Model of Personality, Assessment of; Big Five Factor Model, Theory and Structure; Five Factor Model of Personality, Universality of) specifies a set of exactly five broad personality factors at the apex of the hierarchy. These factors are thought to capture most of the essential human variation in characteristic patterns of thinking, feeling, and behaving. The factors are commonly referred to as neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. Each of these factors is presumed to subsume several lower-level personality facets or traits. Because facets are seen as occupying a conceptual space between broad factors and specific patterns of behavior, they are regarded as the basic building blocks of personality. Cattell (1946) called them *primary traits* as they represent the most basic characterization of a consistent pattern of thought, affect, and behavior. In contrast to the broad factors, the facets are conceived to be narrower in definition and thus to correspond to more circumscribed domains of behavior.

The personality trait hierarchy serves a crucial function as a taxonomy of stable individual differences characteristics. In biology, a taxonomy maps out the broad classification of living

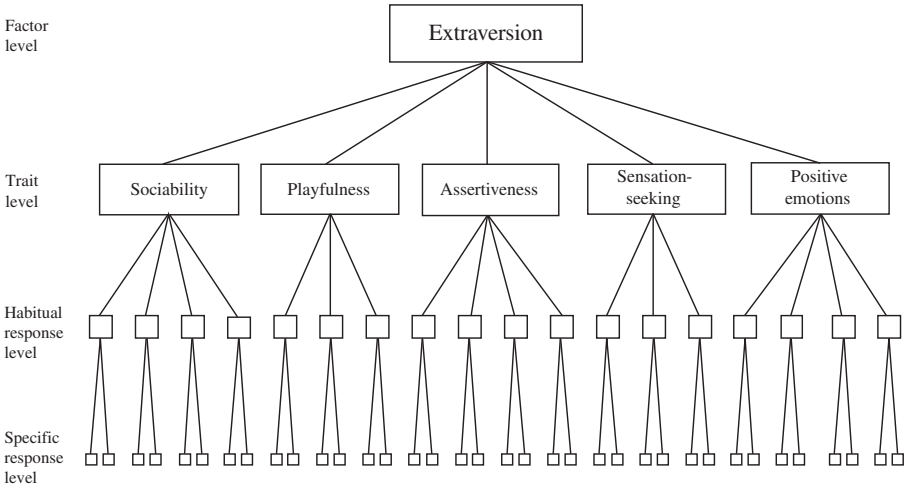

Figure 1 A hierarchical model of personality organization (after Eysenck, 1947). From "Hierarchical organization of personality and prediction of behavior," by S.V. Paunonen, 1998, Journal of Personality and Social Psychology, 74, p. 539. Copyright 1998 by the American Psychological Association. Adapted with permission.

things based on shared characteristics (e.g., mammals), which then branches out with more differentiation that highlights differences between (but similarities within) subclasses (e.g., rodents, primates). In a similar vein, a personality taxonomy organizes traits with different levels of abstraction into a hierarchical fashion that allows for the description of distinct versus overlapping features among them. For instance, lower-level traits like achievement-striving and self-discipline are conceptually distinct, yet they share common features (e.g., propelling an individual to work hard in order to obtain mastery in task-related endeavors). Their commonality means they might both be classified under the higher-level dimension of Conscientiousness. Conversely, a trait like hostility, which shares hardly any commonality with the traits of achievement-striving or self-discipline, would probably not be subsumed under Conscientiousness and, instead, would define a different factor. The hierarchy thus enables researchers to get a sense of how individual traits might be clustered according to their functional similarities and differences.

An alternative view conceptualizes personality facets not as narrow traits that are subsumed under a broad Big Five factor but as varying blends of two Big Five factors (Hofstee et al., 1992). This view capitalizes on a circumplex model (i.e., a circular layout of traits) to organize the placement of personality factors and facets, and is best represented by the Abridged Big Five Circumplex scales from the International Personality Item Pool (AB5C-IPIP; Goldberg, 1999). Two Big Five factors (e.g., extraversion and agreeableness) can be paired to form a circumplex where they make up two orthogonal axes at 90° to each other. Ten such circumplexes with unique pairwise combinations of the five factors can thus be derived. Each of these circumplexes is then partitioned into 12 segments of 30 each and trait-descriptive items that fall within a segment and its polar opposite are thought to define a personality facet. Across the 10 circumplexes, each Big Five factor yields one facet that comprises items defining the factor itself and eight other facets that have secondary loadings with other Big Five factors.

The nature of the personality facet is fundamentally very different between the hierarchical and the circumplex models. In the former, a facet is considered to represent a rather narrow unidimensional psychological construct that is conceptually homogeneous and yet, at the same time, shares some common variance with other conceptually close facets to form a broad factor. Under the latter model, however, a facet is thought to intersperse between two broad factors of personality; hence it is conceptually heterogeneous as it takes in elements from the two different factors. In other words, facets in the hierarchical model represent relatively pure conceptual units of personality whereas facets in the circumplex model constitute derivatives from a process of amalgamation. It would seem that the way facets in the circumplex model are derived actually expands the conceptual space of the Big Five factors, giving detailed information on the intersections among the factors. However, the circumplex model facets do not help in elucidating the specific features within the Big Five factors themselves.

The hierarchical view takes precedence over the circumplex model in this article. Discussions in the following sections adhere to the view that personality facets should represent conceptually homogeneous entities that inform the specific aspects underlying the broad personality factors.

How Many and What Facets Are Important?

The definition of the five broad dimensions of personality as exemplified in the five factor model has reached wide consensus among researchers. The same cannot be said of the facet-level traits. One issue of contention has to do with what might be the optimal number of facets needed to measure the universe of individual differences adequately and comprehensively. Going by the range of the number of facet-level scales proposed by the developers of established personality inventories, it seems that nobody has the final say on this issue. The Revised NEO Personality Inventory (NEO-PI-R; Costa and McCrae, 1992) is probably considered by many to be the gold standard in measuring the five broad personality factors and their constituent facets. The NEO-PI-R specifies six facets to underlie each factor; hence a total of 30 facets make up the domain of personality in that model. But other assessment models exist. The AB5C-IPIP (Goldberg, 1999) measures a total 45 facet-level traits, nine for each of the five factors. A lexical analysis on trait-related words in natural language yielded 18 facets in at least one study (Saucier and Ostendorf, 1999). Several other inventories, though not developed originally to assess the five factor model, have been shown to yield (more or less) the same five broad factors via factor analysis, even with different numbers of facet-level traits. Cattell's Sixteen Personality Factor Questionnaire (16 PF; Cattell et al., 1993), as the name suggests, proposes 16 personality facets whereas Jackson's (1984) Personality Research Form-E has 20 facet-level trait scales. Simply put, there is hardly any consensus when it comes to the optimal number of lower-level personality facets in the five factor model.

A second debatable issue concerns the assignment of personality facets to the Big Five factors. Take the Extraversion factor as an example. The NEO-PI-R facets underlying Extraversion are represented by scales called Warmth, Gregariousness, Assertiveness, Activity, Excitement-Seeking, and Positive Emotions. Conversely, the 16 PF Extraversion facets include Warm–Reserved, Lively–Serious, Bold–Shy, Private–Forthright, and Self-Reliant–Group-Oriented. Although facets similar in content appear across the two inventories (e.g., NEO-PI-R Warmth and 16 PF Warm–Reserved), there are also facets that are unique to one inventory but not to the other. The NEO-PI-R does not have a facet that is conceptually close to the 16 PF Self-Reliant–Group-Oriented scale, for instance.

The glaring lack of agreement among theorists on the nature and number of facets that underlie each of the five broad dimensions of personality stems from differences in theoretical traditions and test construction strategies (e.g., theoretical-rational, empirical, and factor-analytic). This state of affairs for the facet-level traits is reminiscent of the debate two decades ago about the nature and number of the broad dimensions of personality (Block, 1995). Given the current consensus among personality psychologists on the five broad factors, it is expected that the exposition of personality facets should become clearer with future research. How a Big Five factor might be defined and what subcomponents such a factor

should entail are major questions waiting to be addressed. Some preliminary effort has been made to identify the important facets of Conscientiousness through joint factor analysis of Conscientiousness-related facet scales obtained from several established personality inventories (Roberts et al., 2005). The composition of facets underlying the other Big Five factors can be clarified using this strategy. Note that, as we will describe more fully later, not all researchers agree that every personality trait has a place under the umbrella of the Big Five factors (see Paunonen and Jackson, 2000).

Some Properties of Personality Facets

Are people accurate when judging the personality of others (*see* Person Perception, Accuracy of)? To what extent are personality traits heritable (*see* Personality Differences and Development: Genetic and Environmental Contributions)? Are personality traits stable or malleable across the life span (*see* Personality Development: Systems Theories)? These are but a small sample of questions that have intrigued personality psychologists for many years. Very often, researchers have relied on measuring the Big Five factors when addressing these questions. For example, with regard to the question about person perception, much research has been conducted at the level of the five factors. Hence, researchers would normally correlate a target's (i.e., the person being judged) self-ratings and an observer's ratings about the target on the Big Five factors. Similarly, if researchers are interested in the degree of genetic (vs environmental) contribution to personality characteristics, the typical unit of analysis has been at the factor level. However, there is nothing to prevent researchers from examining these same questions at the level of the facets or traits. The complexity of the research design and analysis would no doubt increase as they would have to deal with a larger number of variables (e.g., 30 facet scales, in the case of NEO-PI-R, vs just five factor scales). Still, the richness of information obtained may outweigh the cost of the additional complexity. In the sections below, several related observations associated with the facets of the five factor model are briefly reviewed.

Self-Observer Agreement

Numerous studies have shown that people can rate an acquaintance's personality characteristics with reasonably good accuracy (or agreement). For example, the self-observer agreement coefficients of the Big Five factors, as assessed by measures like the NEO-PI-R and the Big Five Inventory (BFI; see John et al., 2008), average around 0.40–0.50 (Costa and McCrae, 1992; John et al., 2008) among acquaintance ratings. For the facet scales, the self-observer agreement coefficients average around 0.35 for the NEO-PI-R (Costa and McCrae, 1992) and 0.45 for the BFI (Soto and John, 2009). The agreement coefficients for spouses are somewhat higher for both Big Five factors (mean = 0.60) and facets (mean = 0.50) than for acquaintances, as expected (Costa and McCrae, 1992).

The general picture suggests that observers' judgment accuracy drops a notch when rating the targets' personality facets compared to their personality factors. A possible explanation is that, compared to the factor scales, facet scales are usually less reliable due to a smaller item set. As a result, accuracy coefficients suffer because of less measurement reliability among the facets. Nonetheless, the self-observer agreement coefficients for the Big Five facets are still quite respectable, especially when the dyads are in intimate relationships. It seems that people can accurately rate each other on numerous specific domains of behavior, not just on a few global personality factors.

Heritability

A well-established finding in behavior genetic studies of personality is that approximately half of the variability in individual differences on the Big Five factors can be attributed to genetic factors and the other half mostly to nonshared environmental influences (Bouchard and Loehlin, 2001). If personality facets are considered to be narrow domains of behavior within the broad factors, then it should not be surprising that they show similar estimates of heritability. Indeed, research suggests that the facets have, on average, heritabilities hovering around 0.40 (Briley and Tucker-Drob, 2012; Jang et al., 1998). Jang et al. (1998) demonstrated that with the Big Five factors partialed out, the residual variance in facet scales still showed substantial heritability. This finding led those authors to conclude that personality facet scales are "measures of discrete constructs with a heritable and thus biological basis" (p. 1563).

A few studies have examined the genetic and environmental structure of facets using multivariate methods. The emerging picture is that a simple latent trait model cannot account for all the common variance among facets thought to define a single factor, at least for some of the Big Five factors (Briley and Tucker-Drob, 2012; Jang et al., 2002; Johnson and Krueger, 2004). These findings call into question whether higher-level factors like Agreeableness mediate most of the genetic (and environmental) influences on individual facets. Instead, the observation that a substantial amount of the genetic variability is specific to the facet-level traits suggests that they should not be ignored in favor of broader factor measures in personality research.

Personality Development

Personality stability or change over the life span can be quantified by two independent indices: (1) rank-order stability and (2) mean-level change. Rank-order stability refers to the stability of the relative positions of individuals on a personality characteristic over time whereas mean-level change refers to how a trait level, averaged across a sample of individuals, might change over time.

Personality facets in adults appear to exhibit strong rank-order stabilities over long periods of time. For instance, Terracciano et al. (2006) reported that the median stability of NEO-PI-R facets over an average of 10 years was 0.70 (range = 0.57–0.82), which was slightly lower than that of the NEO-PI-R factors (median = 0.81; range = 0.78–0.85). This finding indicates that people's relative standings on individual traits do not change much over long time intervals during adulthood.

The analysis of mean-level changes in personality facets has produced very interesting insights. In both cross-sectional and

longitudinal studies, there is evidence that facets belonging to the same factor can show divergent developmental trajectories (Soto and John, 2012; Soto et al., 2011). From age 20 to about 60, conscientiousness exhibited an upward age trend (i.e., older adults being more conscientious than younger adults; Soto et al., 2011), and this was supported by longitudinal data as well (Soto and John, 2012). Facet-level analysis indicated that this trajectory for broad conscientiousness was driven specifically by facets of industriousness and self-discipline but not orderliness. Orderliness did not change much as a function of age. It seems that the additional life responsibilities for older adults elicit greater levels of achievement-related traits compared to younger adults, but these life tasks do not modulate the level of orderliness across ages. Extraversion facets also showed divergent trajectories: assertiveness and social confidence increased whereas gregariousness decreased over time (Soto and John, 2012). This is in line with the hypothesis that the dominance component of extraversion increases due to the enlarged occupational and social role responsibilities older adults take on, whereas the gregariousness aspect decreases as older adults become more selective in terms of who they socialize with (Roberts et al., 2006). Together, these preliminary findings highlight the advances in knowledge when developmental trends are tracked at the facet level, and not just at the factor level.

Prediction of Outcome Criteria

An appealing feature of a hierarchical organization of personality variables is the flexibility afforded to use either the broad factors or the narrow facets (or both) in the prediction of outcome criteria. A broad personality factor provides bandwidth in prediction through its ability to predict across a variety of criteria, though the strength of those associations may be modest. Conversely, a narrow personality facet offers fidelity by ensuring strong prediction of a smaller number of specific criterion variables (Cronbach and Gleser, 1957). The vast majority of empirical research in psychology has relied on the Big Five factors as predictors of consequential outcomes, with great success in many cases. The use of narrow personality traits as predictors, unfortunately, has not enjoyed the widespread acceptance that it should. We believe that the potential for facets to increase predictive power and explanatory value of personality constructs is immense, and we outline our reasons in the following sections (Paunonen and Hong, 2015).

Specific Variance in Facets

As the basic building blocks of personality, each facet-level trait represents a consistent pattern of thinking, feeling, and behaving that is psychologically coherent (e.g., gregariousness, modesty, orderliness). Several facets thought to be subsumed under a single global factor are expected to correlate with one another. Still, most of these correlations are not close to perfect, suggesting that each facet retains some unique aspects not found in other related facets (i.e., nonredundancy).

In most personality inventories, a score on a broad factor measure is derived by aggregating separate scores on several facet-level scales. This can be done in two ways. The first is to take the sum or average across the constituent facet scores, each unit-weighted so that they contribute equally to the broad factor measure. By emphasizing the common core that represents the factor, the unique variance associated with each facet is diluted. The second method is to estimate the broader factor from its constituent facets via a common factor analytic model. In such a model, facets are considered to be indicators of an underlying factor construct, and their unique variances are regarded as error and thus discarded. The factor loadings form the weights by which the facets are aggregated to derive the factor's score. Doing so means that any nonrandom variance specific to a particular facet is thrown out in favor of the common factor variance underlying those facets. In other words, individuating information provided by each of the facets has been diluted or omitted when they are aggregated into a single broad factor via these two methods. It could very well be that the lost trait-specific variance that exists at the facet level is highly predictive of some criterion of interest.

It is not hard to imagine a hypothetical scenario where two lower-level facets that underlie a higher-level factor exhibit opposite associations with an external criterion. One facet shows a positive association with the criterion whereas the other shows a negative relation; the net effect is that the factor (aggregated from the facets) would yield no relation with the criterion (see O'Neill and Paunonen, 2013, for examples). The more common situation in empirical data is that a facet might predict a criterion, but its predictive capability is masked when it gets aggregated with other facets (which are not related to the criterion). It would be short-sighted if one were to conclude that no relation exists between the factor and the criterion, when even one specific aspect of that factor (i.e., a facet) shows predictive validity. Therefore, researchers should be cognizant about the possibility of losing criterion-valid but facet-specific variance when only broad personality factors are used for prediction.

Predictive Validity of the Facets

Some have argued that facet-specific variance is not generally useful for criterion prediction and that little statistical utility can be obtained in behavior prediction by using personality predictors other than the five broad factors (Ones and Viswesvaran, 1996). If that is true, then adding narrow facets of personality to the broad personality factors in a prediction context should not increase the predictive accuracy of outcome criteria. This question of whether facet-level personality variables are useful to the prediction of criteria relative to factor-level personality variables can be addressed empirically.

In a series of studies, Paunonen and colleagues have demonstrated that facets provide incremental validity beyond the factors in the prediction of a wide variety of behavior of some social importance, such as alcohol/tobacco consumption, grade-point average, and dating variety (Paunonen, 1998; Paunonen and Ashton, 2001). Take predicting grade-point average as an example in Paunonen and Ashton's (2001) study. The amount of criterion variance in grade-point average accounted for by the five NEO-PI-R factors combined was 11.1%. On the other hand, using five NEO-PI-R facets in the prediction equation, rationally selected by judges as traits likely to be related to grades, accounted for 15.3% of criterion

variance (Paunonen and Ashton, 2013). More important, personality facets afforded such predictive validity superiority over the personality factors across many criteria.

The predictive advantage of the facet-level traits over their factor-level counterparts has been demonstrated in the prediction of a wide array of other outcomes, including job performance (Vinchur et al., 1998), externalizing behavior (Miller et al., 2008), prejudice (Ekehammar and Akrami, 2007), and mortality (Turiano et al., 2012). Collectively, these studies point to the importance of personality facets, because the gain in predictive accuracy over the personality factors can be substantial.

Correspondence in Complexity of Predictor and Criterion

Ones and Viswesvaran (1996) have put forth an argument that optimal prediction occurs if there is a match in terms of the factorial complexity of the predictor and criterion variables. That is, if a criterion is a multidimensional construct (e.g., job performance), then accuracy in prediction would be best achieved by using a multidimensional predictor. Such a predictor might be a linear composite of several unidimensional variables derived through aggregation and then used as a single predictor in a standard correlational analysis. However, Paunonen (1998; Paunonen and Ashton, 2001) noted that there is no mathematical requirement that the predictor and criterion complexities must correspond to each other for successful prediction. Nunnally's (1978) advice with regard to maximizing prediction of the multidimensional criterion is worth reiterating here:

> As some will argue, the criterion usually is factorially complex and consequently can be predicted best by a factorially complex predictor test. Instead of building the factorial complexity into a particular test, it is far better to meet the factorial complexity by combining tests in a battery by multiple regression, in which case tests would be selected to measure the different factors that are thought to be important.
>
> p. 268

There is no inherent advantage in using a factorially complex predictor such as a broad personality dimension to predict a similar factorially complex criterion such as job performance. A more defensible strategy is to include a number of unidimensional facet-level traits in a multiple regression context to predict a criterion variable, be it factorially complex or not (Christiansen and Robie, 2011). Optimal prediction should occur when personality facets are judiciously selected based on theoretical reasoning or prior empirical evidence (O'Neill and Paunonen, 2013).

Contribution to Behavior Understanding

The advantage of facet-level assessment over factor-level assessments applies not just to behavior prediction but also to behavior understanding as well. When an association is found between a personality factor and a criterion, a logical next step is to examine if any of the facets that are supposedly subsumed under this factor might be driving this association. Engaging in such fine-grained analysis has the benefit of identifying specific 'active ingredients' – facets that are responsible for the behavior outcome. For example, Hong and Paunonen (2009) isolated the NEO-PI-R facet scale of Deliberation (a Conscientiousness facet) as a common and robust predictor of several health-risk behaviors. This suggested that a failure in careful deliberation (i.e., not thinking of the consequences of one's actions and acting hastily) can lead one to engage in activities that are potentially health-damaging. By the same token, it would also mean that the engagement in such risky behaviors has nothing to do with other Conscientiousness-related facets such as a sense of mastery (i.e., NEO-PI-R facet scale of Competence) or being methodical and organized (i.e., NEO-PI-R facet scale of Order). This sort of precise information would not only be very useful for theory development but also for guiding good psychological practices (e.g., designing intervention programs).

Facets beyond the Five Factor Model

The focus of this article has been on the facets associated with the five factor model and how they can be exploited for better prediction and explanation of behavior. But it is important to note in this context the existence of facet-level traits that have little variance in common with the Big Five factors, but are nonetheless important individual differences variables that have behavioral consequences. Paunonen and Jackson (2000), for example, identified 10 such traits, including manipulativeness, femininity, and egotism, thought to be not well represented within the Big Five space. Although these facets are omitted from the five factor model, the fact that they represent stable personality characteristics makes them no less important than the Big Five facets. Exclusion of those traits means risking the possibility that the predictability of certain criterion outcomes might be compromised (e.g., Hong and Paunonen, 2009).

Summary

The assertion that one should consider and employ narrow personality facets in predicting and explaining behavior is not a call to abandon the use of broad personality factors altogether. Without a doubt, using personality factors as predictors can yield informative findings especially when the area of inquiry is new. However, when a sizable literature has accrued with respect to factor–criterion associations, it would be timely and appropriate to apply more fine-grained analysis using personality facets. From this perspective, analysis of personality at the facet level supplements, rather than replaces, the data provided by the factors. In fact, it can offer a particularly rich and detailed picture into the psychological mechanisms responsible for crucial outcomes. The factors at the top of the personality hierarchy might be too broad to provide the specificity of information needed. At the other extreme, the specific responses at the bottom of the hierarchy can be overly narrow, and the large number of these responses can make prediction cumbersome and inefficient. Assessment at the level of the unitary personality facet or trait seems about right, yielding sufficient precise information for accurate behavior prediction and understanding.

Conclusion

Much is known about the Big Five factors – their nature, origins, and associations with other constructs. Advances in current knowledge on personality traits now demand exploiting the numerous facets that underlie those five factors. As maintained throughout this article, the advantages of examining personality facets for behavior prediction and explanation are numerous and their inclusion in research agendas should be routinely considered.

See also: Big Five Factor Model, Theory and Structure; Five Factor Model of Personality, Assessment of; Five Factor Model of Personality, Universality of; Person Perception, Accuracy of; Personality Development: Systems Theories; Personality Differences and Development: Genetic and Environmental Contributions.

Bibliography

Block, J., 1995. A contrarian view of the five-factor approach to personality description. Psychological Bulletin 117, 187–215.

Bouchard, T.J., Loehlin, J.C., 2001. Genes, evolution, and personality. Behavior Genetics 31, 243–273.

Briley, D.A., Tucker-Drob, E.M., 2012. Broad bandwidth or high fidelity? Evidence from the structure of genetic and environmental effects on the facets of the five factor model. Behavior Genetics 42, 743–763.

Christiansen, N.D., Robie, C., 2011. Further consideration of the use of narrow trait scales. Canadian Journal of Behavioural Science 43, 183–194.

Cattell, R.B., 1946. The Description and Measurement of Personality. Harcourt, Brace & World, New York.

Cattell, R.B., Cattell, A.K., Cattell, H.E.P., 1993. 16PF Fifth Edition Questionnaire. Institute for Personality and Ability Testing, Champaign, IL.

Cronbach, L.J., Gleser, G.C., 1957. Psychological Tests and Personnel Decisions. University of Illinois Press, Urbana, IL.

Costa Jr., P.T., McCrae, R.R., 1992. Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual. Psychological Assessment Resources Inc, Odessa, FL.

Ekehammar, B., Akrami, N., 2007. Personality and prejudice: from Big Five personality factors to facets. Journal of Personality 75, 899–925.

Eysenck, H.J., 1947. Dimensions of Personality. Routledge, London.

Goldberg, L.R., 1993. The structure of personality traits: vertical and horizontal aspects. In: Funder, D.C., Parke, R.D., Tomlinson-Keasey, C., Widaman, K. (Eds.), Studying Lives through Time: Personality and Development. American Psychological Association, Washington, DC, pp. 169–188.

Goldberg, L.R., 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In: Mervielde, I., Deary, I., De Fruyt, F., Ostendorf, F. (Eds.), Personality Psychology in Europe, vol. 7. Tilburg University Press, Tilburg, the Netherlands, pp. 7–28.

Hofstee, W.K.B., De Raad, B., Goldberg, L.R., 1992. Integration of the Big Five and circumplex approaches to trait structure. Journal of Personality and Social Psychology 63, 146–163.

Hong, R.Y., Paunonen, S.V., 2009. Personality traits and health-risk behaviors in university students. European Journal of Personality 23, 675–696.

Jackson, D.N., 1984. Personality Research Form Manual. Research Psychologists Press, Port Huron, MI.

Jang, K.L., McCrae, R.R., Angleitner, A., Riemann, R., Livesley, W.J., 1998. Heritability of facet-level traits in a cross-cultural twin sample: support for a hierarchical model of personality. Journal of Personality and Social Psychology 74, 1556–1565.

Jang, K.L., Livesley, W.J., Angleitner, A., Riemann, R., Vernon, P.A., 2002. Genetic and environmental influences on the covariance of facets defining the domains of the Five-Factor model of personality. Personality and Individual Differences 33, 83–101.

John, O.P., Naumann, L.P., Soto, C.J., 2008. Paradigm shift to the integrative Big Five trait taxonomy: history, measurement, and conceptual issues. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research, third ed. Guilford Press, New York, pp. 114–158.

Johnson, W., Krueger, R.F., 2004. Genetic and environmental structure of adjectives describing the domains of the Big Five model of personality: a nationwide US twin study. Journal of Research in Personality 38, 448–472.

Miller, J.D., Lynam, D.R., Jones, S., 2008. Externalizing behavior through the lens of the five-factor model: a focus on agreeableness and conscientiousness. Journal of Personality Assessment 90, 159–164.

Nunnally, J., 1978. Psychometric Theory. McGraw-Hill, New York.

O'Neill, T.A., Paunonen, S.V., 2013. Breadth in personality assessment: Implications for the understanding and prediction of work behavior. In: Christiansen, N.D., Tett, R.P. (Eds.), Handbook of Personality at Work. Routledge, New York, pp. 299–333.

Ones, D.S., Viswesvaran, C., 1996. Bandwidth-fidelity dilemma in personality measurement for personnel selection. Journal of Organizational Behavior 17, 609–626.

Paunonen, S.V., 1998. Hierarchical organization of personality and prediction of behavior. Journal of Personality and Social Psychology 74, 538–556.

Paunonen, S.V., Ashton, M.C., 2001. Big Five factors and facets and the prediction of behavior. Journal of Personality and Social Psychology 81, 524–539.

Paunonen, S.V., Ashton, M.C., 2013. On the prediction of academic performance with personality traits: a replication study. Journal of Research in Personality 47, 778–781.

Paunonen, S.V., Hong, R.Y., 2015. On the properties of personality traits. In: Mikulincer, M., Shaver, P.R. (Eds.), The APA Handbook of Personality and Social Psychology, Personality Processes and Individual Differences, vol. 4. American Psychological Association, Washington, DC, pp. 233–259.

Paunonen, S.V., Jackson, D.N., 2000. What is beyond the Big Five? Plenty! Journal of Personality 68, 821–835.

Roberts, B.W., Chernyshenko, O.S., Stark, S., Goldberg, L.R., 2005. The structure of conscientiousness: an empirical investigation based on seven major personality questionnaires. Personnel Psychology 58, 103–139.

Roberts, B.W., Walton, K.E., Viechtbaur, W., 2006. Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. Psychological Bulletin 132, 1–25.

Saucier, G., Ostendorf, F., 1999. Hierarchical subcomponents of the Big Five personality factors: a cross-language replication. Journal of Personality and Social Psychology 76, 613–627.

Soto, C.J., John, O.P., 2009. Ten facet scales for the Big Five inventory: convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. Journal of Research in Personality 43, 84–90.

Soto, C.J., John, O.P., 2012. Development of Big Five domains and facets in adulthood: mean-level age trends and broadly versus narrowly acting mechanisms. Journal of Personality 80, 881–914.

Soto, C.J., John, O.P., Gosling, S.D., Potter, J., 2011. Age differences in personality traits from 10 to 65: Big-Five domains and facets in a large cross-sectional sample. Journal of Personality and Social Psychology 100, 330–348.

Terracciano, A., Costa, P.T., McCrae, R.R., 2006. Personality plasticity after age 30. Personality and Social Psychology Bulletin 32, 999–1009.

Turiano, N.A., Spiro III, A., Mroczek, D.K., 2012. Openness to experience and mortality in men: analysis of traits and facets. Journal of Aging and Health 24, 654–672.

Vinchur, A.J., Schippmann, J.S., Switzer, F.S., Roth, P.J., 1998. A meta-analytic review of predictors of job performance for salespeople. Journal of Applied Psychology 83, 586–597.

Five Factor Model of Personality, Personality Disorder

Cristina Crego and Thomas A Widiger, University of Kentucky, Lexington, KY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

It is evident that the conceptualization, diagnosis, and classification of personality disorder are shifting toward a dimensional trait model and, more specifically, the five-factor model (FFM) in particular. The purpose of this article is to provide an overview of the FFM of personality disorder. It will begin with a description of this dimensional model of normal and abnormal personality functioning, along with a brief overview of its empirical support. This will be followed by a discussion of its potential advantages, as well as a comparison with Section III of DSM-5 for emerging models and measures.

The diagnosis and classification of psychopathology is shifting to a dimensional model (Widiger and Gore, in press). The primary authors of the current, fifth edition of the American Psychiatric Association's (APA) Diagnostic and Statistical Manual of Mental Disorders (DSM-5; APA, 2013) stated that the primary goal of this latest edition was to begin this shift. "We have decided that one, if not the major, difference between DSM-IV and DSM-V will be the more prominent use of dimensional measures" (Regier et al., 2009: p. 649).

Most (if not all) mental disorders appear to be the result of an array of biological vulnerabilities and dispositions interacting in complex ways with a variety of environmental, psychosocial events that often exert their effects over a progressively developing period of time (Widiger and Gore, in press). The symptoms and pathologies of mental disorders are very responsive to a wide variety of neurobiological, interpersonal, cognitive, and other mediating and moderating variables that help to develop, shape, and form a particular individual's psychopathology profile. This complex etiological history is unlikely to be well described by single diagnostic categories that attempt to make distinctions at nonexistent discrete joints. Patients rarely meet diagnostic criteria for just one disorder; there is no clear boundary between the presence versus absence of any mental disorder; specific etiologies, pathologies, and treatments are not being discovered; and patients who meet the criteria for the same disorder are often different in very important ways from one another (Regier et al., 2009). As acknowledged in the introduction to DSM-5, "the once plausible goal of identifying homogeneous populations for treatment and research resulted in narrow diagnostic categories that did not capture clinical reality, symptom heterogeneity within disorders, and significant sharing of symptoms across multiple disorders" (APA, 2013: 12). It is further asserted that dimensional approaches will "supersede current categorical approaches in coming years" (p. 13).

The domain of psychopathology for which this shift is most readily suitable is the personality disorders (Widiger and Trull, 2007). The purpose of this article is to describe and discuss a particular dimensional trait model of personality disorder from the perspective of the five-factor model (FFM).

FFM of Personality

Most models of personality (and personality disorder) have been developed through the speculations and insights of prominent theorists. The development of the FFM was instead empirical; specifically, through studies of trait terms within existing languages. This lexical paradigm is guided by the hypothesis that what is of most importance, interest, or meaning will be encoded within the language. Language can be understood as a sedimentary deposit of people's observations over thousands of years of the respective language's development, usage, and transformation. The most important domains of personality are those with the greatest number of terms to describe and differentiate their various manifestations and nuances, and the structure of personality will be evident in the empirical relationship among these trait terms (Goldberg, 1993).

The initial lexical studies were conducted on the English language, and these investigations converged onto a five-factor structure (Goldberg, 1993), consisting of extraversion, agreeableness, conscientiousness, emotional instability, and openness (or unconventionality). Subsequent lexical studies have been conducted on the German, Dutch, Czech, Polish, Russian, Italian, Spanish, Hebrew, Hungarian, Turkish, Korean, Filipino, and other languages, and the findings have supported well the universal existence of the five domains, albeit replication is the strongest for the three domains that invariably appear first (i.e., extraversion, agreeableness, and conscientiousness; De Raad et al., 2010).

The universality of the FFM domains is not terribly surprising when one considers their content. The two relatively largest domains (extraversion and agreeableness) concern all manner of interpersonal relatedness. It is to be expected that the domains of personality functioning considered to be most important to people across all cultures and languages when describing themselves and others would concern how people relate to one another. The third domain of the FFM is conscientiousness (otherwise known as constraint). This domain concerns the control and regulation of behavior, and contrasts being disciplined, compulsive, dutiful, conscientious, deliberate, workaholic, and achievement-oriented, with being carefree, irresponsible, lax, impulsive, spontaneous, disinhibited,

negligent, and hedonistic. It is again perhaps self-evident that all cultures would find it important to describe the likelihood that a person will be responsible, conscientious, competent, and diligent as a mate, parent, friend, employee, or colleague. The fourth domain, emotional instability, is of considerable importance in the fields of clinical psychology and psychiatry, saturating most measures of personality disorder. It is again not surprising that most, and perhaps all, cultures consider the emotional stability (anxiousness, depressiveness, irritability, volatility, anger, and vulnerability) of its partners, children, friends, workers, laborers, and employees to be of considerable importance. The fifth domain, openness, intellect, or unconventionality, reflects a culture or society's interest in creativity, intellect, and imagination, contrasting being open-minded, unusual, odd, weird, creative, peculiar, and unconventional versus being closed-minded, practical, conventional, and rigid. After that, there is really very little that is left, if anything. It is in fact difficult to identify a trait term that lies outside of the five domains of extraversion (vs introversion), agreeableness (vs antagonism), conscientiousness (vs disinhibition), emotional instability, and unconventionality (vs conventionality).

There are both adaptive and maladaptive components for each of the 10 poles of the five FFM domains. The language is not proportional in the extent to which there are adaptive and maladaptive trait terms within each of the 10 poles. There are substantially more undesirable (and fewer desirable) trait terms for introversion, antagonism, disinhibition, emotional instability, and conventionality, than there are for extraversion, agreeableness, conscientiousness, emotional stability, and unconventionality (respectively). Nevertheless, there are undesirable trait terms for both poles of all five domains, including extraversion (e.g., flaunty, showy, and long-winded), agreeable (e.g., ingratiating and dependent), conscientious (e.g., leisureless, perfectionistic, workaholic, and stringent), unconventionality (e.g., strange, peculiar, and weird), and even emotional stability (e.g., emotionless, glib, and fearless). Table 1 provides illustrative traits within the normal range (largely adaptive) and the more abnormal extreme range for both poles of the six facets of agreeableness versus antagonism.

Since its original conceptualization through the lexical research, the FFM has amassed a considerable body of empirical support, including multivariate behavior genetics with respect to its structure (and even molecular genetic support for neuroticism), cortical functioning and structure, childhood antecedents, temporal stability across the life span, and cross-cultural validity, both through emic studies considering the structures indigenous to such languages and etic studies that address whether the FFM replicates well across major regions of the world, including North America, South America, Western Europe, Eastern Europe, Southern Europe, the Middle East, Africa, Oceania, South–Southeast Asia, and East Asia (Widiger et al., 2012). The FFM has also been shown across a remarkably vast empirical literature to be useful in predicting a substantial number of important life outcomes, both positive and negative, such as subjective well-being, social acceptance, relationship conflict, marital status, academic success, criminality, unemployment, physical health, mental health, job satisfaction, and mortality (Ozer and Benet-Martinez, 2006). The FFM is understandably the predominant dimensional model of general personality structure within psychology (John et al., 2008).

FFM and Personality Disorder

One of the strengths of the FFM is its robustness (John et al., 2008). Consistent with its comprehensiveness in the coverage of traits within the language, the FFM has been used effectively in many literature reviews as a basis for comparing, contrasting, and integrating diverse sets of personality scales and constructs (Widiger et al., 2012). The FFM is comparably robust in its coverage of abnormal as well as normal personality functioning. A substantial body of research now indicates that the FFM successfully accounts for the symptoms and traits of the DSM-IV (now DSM-5) personality disorders (Samuel and Widiger, 2008; Widiger et al., 2013b). Consider, for example, obsessive-compulsive personality disorder (OCPD). It is not difficult to understand that compulsivity, and much of the traits of OCPD, are maladaptive variants of FFM conscientiousness. The essential feature of OCPD is 'a preoccupation with orderliness, perfectionism, and mental and interpersonal control' (APA, 2013: 678), including within its diagnostic criteria such traits as perfectionism, preoccupation with order and organization, workaholism, and, quite explicitly, overconscientiousness. Similarly, FFM conscientiousness includes such facets as order, discipline, achievement-striving, and deliberation. It would seem apparent that maladaptive and/or extreme variants of order, discipline, achievement-striving, and deliberation would be the OCPD traits of perfectionism, preoccupation with order and organization, workaholism, and overconscientiousness.

Existing measures of the FFM can in fact be used to assess for the presence of a DSM-IV or DSM-5 personality disorder. A clinician or researcher can correlate an individual's scores on the 30 facets of the FFM with the FFM profile of a prototypic case of a respective personality disorder. This correlation can then serve as an index of the extent to which the individual's personality structure matches a respective DSM-5 personality

Table 1 Normal (adaptive) and abnormal (maladaptive) variants of facets of agreeableness

Facet	*Maladaptively low*	*Low*	*High*	*Maladaptively high*
Trust	Gullible	Trusting	Cautious, skeptical	Cynical, suspicious
Straightforwardness	Guileless	Honest, forthright	Savvy, cunning, shrewd	Deceptive, dishonest, manipulative
Altruism	Self-sacrificial, selfless	Giving, generous	Frugal, withholding	Greedy, self-centered, exploitative
Compliance	Yielding, subservient, meek	Cooperative, obedient, deferential	Critical, contrary	Combative, aggressive
Modesty	Self-effacing, self-denigrating	Humble, modest, unassuming	Confident, self-assured	Boastful, vain, pretentious, arrogant
Tender-mindedness	Overly soft-hearted	Empathic, sympathetic, gentle	Strong, tough	Callous, merciless, ruthless

disorder. Studies have consistently indicated that this correlation provides as a valid measure of a respective personality disorder, such as borderline, antisocial, or avoidant, as would be provided by the more traditional inventories for their assessment (Miller, 2012).

These quantitative prototype matching indices, however, have not worked as well for the assessment of the dependent, schizotypal, and OCPD (Miller, 2012). This probably reflects the fact that the traditional measures of the FFM lack adequate fidelity for the assessment of maladaptive variants of some of the FFM domains, particularly agreeableness, conscientiousness, and openness; and it is precisely these domains that are especially important for the assessment of the dependent, OCPD, and schizotypal personality disorders. However, researchers are now developing new measures of the FFM that are focused in particular on the maladaptive variants of its domains and facets (e.g., De Clercq et al., 2006). For example, there are self-report measures for eight of the 10 DSM-5 personality disorders from the perspective of the FFM, including (for instance) the dependent, obsessive-compulsive, and schizotypal, as well as the borderline, narcissistic, avoidant, histrionic, and antisocial (Widiger et al., 2012).

Clinicians and researchers can use these FFM personality disorder scales in two different ways. One approach, consistent with the more traditional conceptualization of a DSM-5 personality disorder, is to sum the scores for the scales for a respective syndrome to assess that particular personality disorder from the perspective of the FFM. For example, one can combine the 12 scales of the Five Factor Borderline Inventory (FFBI; Mullins-Sweatt et al., 2012) into one score to assess borderline personality disorder (BPD) from the perspective of the FFM. This approach is consistent with existing measures of BPD. An important distinction though between the FFBI and traditional measures of BPD is that the FFBI items and scales were coordinated conceptually and empirically with the FFM of general personality structure. Considering personality disorders from the perspective of the FFM will be useful in the development of a more integrative understanding of normal and abnormal personality (Clark, 2007; Widiger and Trull, 2007). To the extent that disorders of personality are understood as maladaptive variants of the FFM, one can bring to these personality disorders the extensive construct validity research concerning the genetics, childhood antecedents, course, universality, and positive and negative life outcomes identified for the FFM (Widiger et al., 2012).

Another approach to these instruments is to use only a subset of the scales. For example, a clinician or researcher might be interested in only assessing for maladaptive variants of agreeableness. This person might then use only the Gullibility, Selflessness, Subservience, and Self-Effacing scales from the Five-Factor Dependency Inventory. Alternatively, the clinician or researcher might wish to consider subscales from different personality disorders, such as all the various maladaptive variants of extraversion, including, for instance, the Dominance and Thrill Seeking scales from the Elemental Psychopathy Assessment, the Exhibitionism and Authoritativeness scales from the Five Factor Narcissism Inventory, and the Attention-Seeking, Social Butterfly, and Flirtatiousness scales from the Five-Factor Histrionic Inventory.

FFM Diagnosis of Personality Disorder

The purpose of the FFM of personality disorder is not simply to provide another means with which to diagnose the DSM-5 personality disorders. The purpose is, instead, to replace the diagnostic categories with a more valid and credible dimensional model of personality disorder classification. Widiger et al. (2013a) provide a four-step procedure for the diagnosis of a personality disorder from the perspective of the FFM (the fourth step, although, is optional).

The first step is to obtain an FFM description. There are a wide variety of alternative measures of the FFM that are readily available for this assessment, such as the NEO Personality Inventory-Revised, the Five-Factor Model Rating Form, and the International Personality Item Pool (Samuel, 2013). The second step is to identify problems in living and/or maladaptive traits that are associated with elevations on a respective facet of the FFM. The two outside columns of **Table 1** provide illustrative examples of the maladaptive variants of the FFM facets for agreeableness versus antagonism. As noted earlier, there are also now a number of alternative measures to assess for these maladaptive variants (e.g., De Clercq et al., 2006; Widiger et al., 2012).

The third step is to determine whether the impairments are at a clinically significant level warranting a diagnosis of a personality disorder. The diagnostic thresholds for most of the DSM-5 personality disorders were not based on any explicit or published rationale. The FFM of personality disorder, in contrast, proposes a uniform and consistent basis for determining when a disorder is present; more specifically, the level of impairment in social and occupational functioning as indicated in the global assessment of functioning scale (Widiger and Trull, 2007). A score of 71 or above on this scale indicates a normal range (e.g., problems are transient and expectable reactions to stressors); a score of 60 or below represents a clinically significant level of impairment (moderate difficulty in social or occupational functioning, such as having few friends or significant conflicts with coworkers).

The fourth step is the prototype matching described earlier, in which one correlates the FFM trait profile of a patient with the profile for a prototypic case of a respective personality disorder. This step is optional, provided for those who still wish to continue to use single diagnostic terms to describe a patient's personality profile. The FFM prototype matching provides a cross-walk between the FFM trait model and the DSM-5 syndromal approach. In fact, clinicians and researchers interested in studying diagnostic constructs that are outside of the existing nomenclature can use the FFM to provide a reasonably specific description of a clinical construct that is not currently recognized within the DSM-5 diagnostic manual, such as depressive personality disorder, vulnerable narcissism, alexithymia, or successful psychopathy (Mullins-Sweatt et al., 2010).

Advantages of an FFM of Personality Disorder

Shifting from the diagnostic categories of *DSM-5* to the dimensional traits of the FFM provides a number of

improvements and advantages over the existing nomenclature (Widiger et al., 2012). A few of these are discussed briefly in turn.

Individualized and Precise Description

One advantage of the FFM of personality disorder is simply a reflection of its dimensional format. Rather than force an individual into a category that fails to provide a fully accurate description, fails to capture important personality traits, and includes traits that the person does not have, the FFM allows the clinician to provide an individualized profile of precisely the traits that are present. This dimensional trait description would be considerably more precise and accurate than the existing DSM-5 categories, and, as a result, have obvious benefits for treatment, research, insurance, and other social, clinical decisions.

For example, a trait-specific description will be very helpful for treatment decisions. It is evident that treatment does not address or focus on the entire personality gestalt. Instead, clinicians treat, for instance, the affective instability, the behavioral dyscontrol, the despondence, or the self-mutilation of persons diagnosed with BPD, which are specifically identified within the FFBI (Mullins-Sweatt et al., 2012). Effective therapeutic change occurs with respect to these particular components of the disorder rather than with the entire, global construct, and it would then make sense that these specific facets of the syndrome be assessed individually.

The FFM conceptualization of each personality disorder enables researchers as well to disambiguate the DSM-5 diagnostic constructs and determine which particular component of the disorder best explains any particular finding. For example, rather than attribute a finding to the broad construct of schizotypal personality disorder, one can determine more precisely whether it reflects the social withdrawal, suspiciousness, or cognitive-perceptual aberrations. Lynam and Widiger (2007) demonstrated this point with respect to psychopathy, indicating how alternative theories for its core pathology (e.g., response modulation, lack of empathy, or fearlessness) reflect a differential emphasis on its different FFM components (e.g., low conscientiousness, the callousness of antagonism, or the fearlessness of low neuroticism, respectively).

Homogeneous Trait Constructs

The facet and domain constructs of the FFM are considerably more homogeneous than the DSM-5 diagnostic categories. The value of homogeneous diagnostic constructs has long been recognized within psychiatry but not well appreciated within the field of personality disorder. Two persons can meet the diagnostic criteria for the DSM-5 antisocial, borderline, schizoid, schizotypal, narcissistic, and avoidant personality disorders and yet in each case have only one diagnostic feature in common. This heterogeneity handicaps tremendously the effort to identity a specific etiology, pathology, or treatment for a respective personality disorder as there is tremendous variation within any particular group of patients sharing the same diagnosis.

The primary purpose of the APA diagnostic manual is to facilitate treatment planning (APA, 2013). However, it has been over 20 years since the APA has been developing practice guidelines for each of the disorders included within the DSM-5 and to date guidelines have been published for only one personality disorder: borderline. One possible reason for the absence of empirically based manualized treatment plans for personality disorders might be their complex heterogeneity. Given the substantial variability of defining features within each diagnostic category, it would be understandably difficult to develop a uniform treatment program for persons sharing the same personality disorder diagnosis.

The factor analytically based domains of the FFM might be better suited for treatment planning than the DSM-5 personality disorder diagnoses because they are considerably more distinct and homogeneous. Extraversion and agreeableness are concerned specifically with social, interpersonal dysfunction, an area of functioning that is relevant to relationship quality both outside and within the therapy office. Interpersonal models of therapy, marital-family therapy, and group therapy might be particularly suitable for these two domains. In contrast, neuroticism provides information with respect to mood, anxiety, and emotional dyscontrol, often targets for pharmacologic interventions, as well as cognitive, behavioral, and/or psychodynamic therapy. There are very clear pharmacologic implications for mood and anxiety dysregulation and emotional instability (e.g., anxiolytics, antidepressants, and/or mood stabilizers), but little to none for maladaptive antagonism or introversion, the interpersonal domains of the FFM. Maladaptively high openness implies cognitive-perceptual aberrations, and so would likely have pharmacologic implications (i.e., neuroleptics) that are quite different from those for neuroticism. The domain of conscientiousness, in contrast to agreeableness and extraversion, is the domain of most specific relevance to occupational dysfunction, or impairments concerning work and career. Maladaptively high levels involve workaholism, perfectionism, and compulsivity; low levels involve laxness, negligence, and irresponsibility. There might be specific pharmacologic treatment implications for low conscientiousness (e.g., methylphenidates) although, as yet, none for maladaptively high conscientiousness. Perhaps there never will be a pharmacotherapy for high conscientiousness, but the point is that the structure of the FFM could very well be more commensurate with more specific treatment implications than the existing diagnostic categories.

Inclusion of Normal, Adaptive Traits

An additional advantage of the FFM of personality disorder is the inclusion of normal, adaptive traits. The personality disorders are among the more stigmatizing labels within the diagnostic manual. They are relatively unique in concerning 'ego-syntonic' aspects of the self, or one's characteristic manner of thinking, feeling, behaving, and relating to others throughout much of one's entire life. Other mental disorders are events or experiences that happen to the person during the course of his or her life, whereas a personality disorder is who that person is. A personality disorder suggests that who you are and always have been is, itself, disordered.

The FFM of personality disorder provides a more complete description of each person's self that recognizes and appreciates that the person is more than just the personality disorder and

that other aspects of the self can be adaptive, even commendable, despite the presence of some maladaptive personality traits. Some of the personality strengths may be quite relevant to treatment, such as openness to experience indicating an interest in exploratory psychotherapy, agreeableness indicating an engagement in group therapy, and conscientiousness indicating a willingness and ability to adhere to the demands and rigor of dialectical behavior therapy.

Improved Construct Validity

An additional advantage of integrating the classification of personality disorder with general personality research is bringing a considerable body of scientific knowledge concerning the assessment, etiology, course, temporal stability, important life outcomes, and other matters of construct validity to the understanding of personality disorders (Ozer and Benet-Martinez, 2006; Widiger and Trull, 2007). The weak empirical support for many of the APA personality disorder categories contributed to an initial proposal for DSM-5 to delete virtually half of them (Skodol, 2012). This was not a strong endorsement for the validity of the nomenclature.

Many of the construct validity problems for the DSM-5 personality disorders are well addressed to the extent that they are understood as maladaptive variants of the domains and facets of the FFM. For example, one of the major difficulties for the categories is their excessive diagnostic co-occurrence (Skodol, 2012). Lynam and Widiger (2001) demonstrated empirically that the extent to which the personality disorders share FFM traits explains the co-occurrence among them. An additional issue for the DSM-5 personality disorders has been their differential sex prevalence rates, which some have suggested reflects a gender bias in a respective disorder's conceptualization, diagnosis, and/or assessment (Oltmanns and Powers, 2012). The differential sex prevalence rates obtained for the personality disorders are difficult to justify in the absence of any theoretical basis for the differential sex prevalence that should be obtained. In contrast, the FFM has proved useful in explaining gender differences in personality, including the personality disorders. The differential sex prevalence rates obtained for the DSM-5 personality disorders are in fact consistent with gender differences in general personality structure (Widiger et al., 2013).

FFM and DSM-5

Section III of DSM-5, for emerging models and measures, includes a five-domain dimensional trait model that aligns closely with the FFM. As stated in DSM-5, the five domains are "maladaptive variants of the five domains of the extensively validated and replicated personality model known as the 'Big Five,' or the Five-Factor Model of personality" (APA, 2013: 773). DSM-5 *Negative Affectivity* aligns with FFM neuroticism, *Detachment* with FFM introversion, *Antagonism* with FFM antagonism, *Disinhibition* within FFM low conscientiousness, and *Psychoticism* or *Peculiarity* with FFM openness. Within these five broad domains are 25 relatively more specific traits. For example, included within the domain of negative affectivity are emotional liability, anxiousness, separation anxiety, hostility, and depressivity. The assessment of these 25 traits is essentially equivalent to Step Two of the FFM of personality disorder four-step procedure for personality disorder diagnosis.

The diagnostic criterion sets for the personality disorder types included in Section III of DSM-5 consist of a combination of level of personality functioning with maladaptive personality traits (APA, 2013). For example, the diagnosis of DSM-5 Section III BPD includes an assessment of the traits of anxiousness, separation insecurity, hostility, depressivity, impulsivity, emotional liability, and risk taking. These traits align closely with those included with the FFBI assessment of BPD (Mullins-Sweatt et al., 2012), including anxious uncertainty (a maladaptive variant of FFM anxiousness), dysregulated anger (the BPD variant of FFM angry hostility), despondence (FFM depressiveness), behavioral dysregulation (FFM impulsivity), affective dysregulation (FFM vulnerability), and rashness (low FFM deliberation). As noted below, however, the FFM conceptualization of BPD includes additional traits (i.e., self-disturbance, fragility, dissociative tendencies, distrustfulness, manipulativeness, and oppositional).

There are, however, some differences between FFM of personality disorder diagnosis and DSM-5 Section III. These include the absence of normal traits, a limited recognition of bipolarity within the trait structure, and a relatively narrow list of abnormal traits. Each of these differences will be discussed briefly in turn.

Absence of Normal Traits

Normal personality traits were considered for inclusion within DSM-5, consistent with the original name of the work group as the Personality and Personality Disorders Work Group (Skodol, 2012). However, it eventually became apparent that no such traits would be included. All of the 25 DSM-5 traits refer to abnormal personality, as do the five domains (APA, 2013). Each DSM-5 Section III domain is scored on a four-point scale, where 0 = very little or not at all, 1 = mildly descriptive, 2 = moderately descriptive, and 3 = extremely descriptive. Having very little antagonism does not imply being agreeable. It just means being only a little bit antagonistic. There is no ability within the DSM-5 dimensional trait model to assess personality strengths.

Limited Recognition of the Bipolarity of Maladaptive Personality Structure

The five domains of the DSM-5 dimensional trait model are unipolar; that is, low scores on each domain suggest the absence of the maladaptive traits. Low scores do not indicate the presence of traits that are opposite to those at the high end. This is in stark contrast to the FFM, as the latter includes maladaptive variants at both ends of each pole. For example, in the FFM, opposite to antagonism is agreeableness, which includes such maladaptive traits as gullibility, meekness, subservience, self-denigration, and selflessness (see **Table 1**); opposite to disinhibition is constraint, or conscientiousness, including such maladaptive variants as perfectionism, workaholism,

rigidity, and doggedness. The DSM-5 dimensional model, however, does not include traits for maladaptive extraversion, agreeableness, low neuroticism, or low openness (and only one for conscientiousness: rigid perfectionism).

There is a considerable body of research to support the bipolarity in the structure of abnormal personality traits (Widiger et al., 2012). Equally important, the absence of the bipolarity in DSM-5 contributes to a number of problems, including the failure to include some important traits and the misplacement of others. For example, because the model does not include maladaptively low neuroticism (low negative emotionality), DSM-5 Section III is unable to acknowledge the existence of psychopathic fearlessness and glib charm (Lynam and Widiger, 2007). Because the model doesn't include maladaptive agreeableness, there is no ability to recognize dependent self-denigration, gullibility, and meekness.

The absence of bipolarity also results in some of the trait misplacements. For example, submissiveness is a manner of interpersonal relatedness that is associated with agreeableness, but in the absence of acknowledging any maladaptive agreeableness it was placed instead within negative affectivity. Attention-seeking is classified in DSM-5 as a facet of antagonism, inconsistent with the FFM research that places it within extraversion. Perseveration is a trait of compulsivity, but in the absence of conscientiousness it is placed within negative affectivity.

There is some effort within the DSM-5 to acknowledge the existence of the bipolarity. For example, DSM-5 includes a specifier for psychopathic fearlessness, to be indicated by low scores on anxiousness (APA, 2013). However, low maladaptive anxiousness does not necessarily suggest the presence of its opposite, maladaptive fearlessness. Maladaptive anxiousness is said to involve intense feelings of nervousness, tenseness, feeling fearful, apprehensive, and fears of falling apart, losing control, or embarrassment. Measures of anxiousness will correlate significantly with measures of fearlessness, but it is evident that the absence of maladaptive anxiousness does not necessarily suggest the presence of fearlessness.

Limited Number of Traits

Even for the FFM domains that are included, such as antagonism and neuroticism, the proposed model for DSM-5 is limited in its coverage. This is due in part to the fact that the original selection of traits was not based on a systematic survey of the clinical or research literature, nor on the basis of an effort to represent alternative dimensional trait models. The original list of 37 traits was a result of nominations by members of the DSM-5 Work Group, and it was subsequently pared down to just 25 through a factor analysis (Krueger et al., 2011).

Missing from the DSM-5 Section III dimensional trait model are such traits as workaholism, aggressiveness, shamefulness, rumination, vanity, entitlement, fearlessness, meekness, glib charm, and selflessness. The lack of full coverage is particularly evident when one considers the traits that are available for the diagnosis of respective DSM-5 personality disorders. For example, narcissistic personality disorder is diagnosed by just the two traits of grandiosity and attention-seeking; obsessive-compulsive by just rigid perfectionism, perseveration, restricted affectivity, and intimacy avoidance; and dependency by just submissiveness, anxiousness, and separation insecurity.

The limited number of traits and the lack of recognition of the bipolarity of maladaptive personality structure within DSM-5 Section III are perhaps understandable. There has been strong opposition to the dimensional trait model approach to the conceptualization and diagnosis of personality disorder (e.g., Shedler et al., 2010), and a common criticism of the FFM is that it will be too complex for clinicians to use. Opponents to the dimensional trait model would have had a ready target if the authors of the DSM-5 trait model had proposed the entire FFM model. Any model that provides a comprehensive description of normal and abnormal personality structure will be necessarily complex, at least at the facet level. An initial shift to a dimensional trait model should perhaps first propose a simplified version that would be more readily acceptable.

Conclusions

The FFM of personality disorder provides a reasonably comprehensive integration of normal and abnormal personality within a common hierarchical structure. Advantages of the FFM of personality disorder include the provision of precise, individualized descriptions of the personality structure of each patient, the inclusion of homogeneous trait constructs that will have more specific treatment implications, and the inclusion of normal, adaptive personality traits that will provide a richer, fuller, and more appreciative description of each patient. The FFM of personality disorder would address the many fundamental limitations of the categorical model (e.g., heterogeneity within diagnoses, inadequate coverage, lack of consistent diagnostic thresholds, and excessive diagnostic co-occurrence), and bring to the nomenclature a wealth of knowledge concerning the origins, childhood antecedents, stability, and universality of the dispositions that underlie personality disorder. It is apparent that DSM-5 is shifting closer to the FFM through the inclusion within Section III for emerging models and measures a five-domain dimensional model that aligns closely with the five factors of the FFM.

See also: Agreeableness; Big Five Factor Model, Theory and Structure; Conscientiousness; Extraversion; Five Factor Model of Personality, Assessment of; Five Factor Model of Personality, Facets of; Five Factor Model of Personality, Universality of; Neuroticism; Openness to Experience; Personality Theory and Psychopathology; Personality, Trait Models of.

Bibliography

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders. Text Revision, fifth ed. Author, Washington, DC.

Clark, L.A., 2007. Assessment and diagnosis of personality disorder: perennial issues and an emerging reconceptualization. Annual Review of Psychology 57, 227–257.

De Clercq, B., De Fruyt, F., Van Leeuwen, K., Mervielde, I., 2006. The structure of maladaptive personality traits in childhood: a first step toward an integrative developmental perspective for DSM-V. Journal of Abnormal Psychology 115, 639–657.

De Raad, B., Barelds, D.P.H., Levert, E., Ostendorf, F., Mlacic, B., Blais, L.D., Hrebickova, M., Szirmak, Z., Szarota, P., Perugini, M., Church, A.T., Katigbak, M.S., 2010. Only three factors of personality description are fully replicable across languages: a comparison of 14 trait taxonomies. Journal of Personality and Social Psychology 98, 160–173.

Goldberg, L.R., 1993. The structure of phenotypic personality traits. American Psychologist 48, 26–34.

John, O.P., Naumann, L.P., Soto, C.J., 2008. Paradigm shift to the integrative Big Five trait taxonomy: history, measurement, and conceptual issues. In: John, O.P., Robins, R.R., Pervin, L.A. (Eds.), Handbook of Personality. Theory and Research, third ed. Guilford, NY, pp. 114–158.

Krueger, R.F., Eaton, N.R., Clark, L.A., Watson, D., Markon, K.E., Derringer, J., Skodol, A., Livesley, W.J., 2011. Deriving an empirical structure of personality pathology for DSM-5. Journal of Personality Disorders 25, 170–191.

Lynam, D.R., Widiger, T.A., 2001. Using the five factor model to represent the DSM-IV personality disorders: an expert consensus approach. Journal of Abnormal Psychology 110, 401–412.

Lynam, D.R., Widiger, T.A., 2007. Using a general model of personality to identify the basic elements of psychopathy. Journal of Personality Disorders 21, 160–178.

Miller, J.D., 2012. Five-factor model personality disorder prototypes: a review of their development, validity, and comparison to alternative approaches. Journal of Personality 80, 1565–1591.

Mullins-Sweatt, S.N., Edmundson, M., Sauer, S.E., Lynam, D.R., Miller, J.D., Widiger, T.A., 2012. Five-factor measure of borderline personality traits. Journal of Personality Assessment 94, 475–487.

Mullins-Sweatt, S.N., Glover, N., Derefinko, K.J., Miller, J.D., Widiger, T.A., 2010. The search for the successful psychopath. Journal of Research in Personality 44, 559–563.

Oltmanns, T.F., Powers, A.D., 2012. Gender and personality disorders. In: Widiger, T.A. (Ed.), The Oxford Handbook of Personality Disorders. Oxford University Press, NY, pp. 206–218.

Ozer, D.J., Benet-Martinez, V., 2006. Personality and the prediction of consequential outcomes. Annual Review of Psychology 57, 401–421.

Regier, D.A., Narrow, W.E., Kuhl, E.A., Kupfer, D.J., 2009. The conceptual development of DSM-V. American Journal of Psychiatry 166, 645–650.

Samuel, D.B., 2013. Assessing the five-factor model of personality disorder. In: Widiger, T.A., Costa, P.T. (Eds.), Personality Disorders and the Five-Factor Model of Personality, third ed. American Psychological Association, Washington, DC, pp. 221–232.

Samuel, D.B., Widiger, T.A., 2008. A meta-analytic review of the relationships between the five-factor model and DSM-IV-TR personality disorders: a facet level analysis. Clinical Psychology Review 28, 1326–1342.

Shedler, J., Beck, A., Fonagy, P., Gabbard, G.O., Gunderson, J.G., Kernberg, O., Michels, R., Westen, D., 2010. Personality disorders in DSM-5. American Journal of Psychiatry 167, 1027–1028.

Skodol, A.E., 2012. Personality disorders in DSM-5. Annual Review of Clinical Psychology 8, 317–344.

Widiger, T.A., Costa, P.T., McCrae, R.R., 2013a. Diagnosis of personality disorder using the five-factor model and the proposed DSM-5. In: Widiger, T.A., Costa, P.T. (Eds.), Personality Disorders and the Five Factor Model, third ed. American Psychological Association, Washington, DC, pp. 285–310.

Widiger, T.A., Costa, P.T., Gore, W.L., Crego, C., 2013b. Five factor model personality disorder research. In: Widiger, T.A., Costa, P.T. (Eds.), Personality Disorders and the Five-factor Model of Personality, third ed. American Psychological Association, Washington, DC, pp. 75–100.

Widiger, T.A., Gore, W.L. Mental disorders as discrete clinical conditions: dimensional versus categorical classification. In: Beidel, D., Frueh, C. (Eds.), Adult Psychopathology and Diagnosis, seventh ed. Wiley, NY, in press.

Widiger, T.A., Samuel, D.B., Mullins-Sweatt, S., Gore, W., Crego, C., 2012. An integration of normal and abnormal personality structure: the five-factor model. In: Widiger, T.A. (Ed.), The Oxford Handbook of Personality Disorders. Oxford University Press, NY, pp. 82–107.

Widiger, T.A., Trull, T.J., 2007. Plate tectonics in the classification of personality disorder: shifting to a dimensional model. American Psychologist 62, 71–83.

Five Factor Model of Personality, Universality of

Velichko H Fetvadjiev, University of Pretoria, Pretoria, South Africa
Fons JR van de Vijver, Tilburg University, Tilburg, The Netherlands; North-West University, Potchefstroom, South Africa; and University of Queensland, St Lucia, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Five-Factor Model (FFM) is a language-based model of the structure of personality. Evidence from three lines of research on the FFM's universality across cultures is reviewed: (1) questionnaire studies that sought to replicate the FFM across cultures; (2) psycholexical studies that examined implicit models of personality in different languages using lexica; and (3) indigenous and emic–etic studies that examined personality structures in specific cultures using a variety of methods, and subsequently assessed the overlap between these models and the FFM. Support for the universality of the FFM is strongest when FFM-based questionnaires are used and tends to be stronger in Western than in non-Western cultures. Four factors are usually well replicated (Neuroticism, Extraversion, Agreeableness, and Conscientiousness), whereas the fifth factor, Openness, is not always found. From a psycholexical perspective, the replicability of three broad dimensions is best established: Extraversion, Agreeableness, and Conscientiousness. Emic–etic studies tend to converge on a structure similar to the FFM, although sometimes expanding it in the domain of social-relational functioning. The nonreplication of the FFM cannot be easily attributed to a common pattern of causes; data quality, availability of abstract terms in the language, and cultural salience of the concepts are often involved.

The Five-Factor Model (FFM) of personality is a model of the structure of personality firmly based in language, according to which human personality can be described using five broad dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. Each of these dimensions subsumes a number of lower level traits on both the positive and the negative pole, for example, *talkative* and *reserved* in Extraversion, *friendly* and *harsh* in Agreeableness, *orderly* and *messy* in Conscientiousness, *calm* and *fearful* in Neuroticism, and *perceptive* and *ignorant* in Openness (Goldberg, 1990). The terms *Five-Factor Model* and *Big Five* are often used interchangeably, although in the narrow sense, the former refers to a stricter version of the model with 30 definite lower level facets (Costa and McCrae, 1992), whereas the latter encompasses a broader array of models with different lower level components; we follow the habit of using the two terms as synonymous.

The FFM was initially developed to provide a parsimonious account of variation in personality descriptions in the English language. The FFM dates back to work by Allport and Odbert (1936) and Stagner (1937) in the 1930s, revived in the 1960s by Norman (1963) and Tupes and Christal (1961), and widely established in the last decades of the twentieth century by various researchers, notably Goldberg (1990) and McCrae and Costa (1997). This renewed interest spurred cross-cultural studies of the FFM, turning the attention toward the cross-cultural replicability of the model. If the FFM were found to replicate across cultures and languages, this would provide support for its validity as a panhuman model of personality and would allow it to be used as a basis for theorizing and empirical research transcending cultural and regional boundaries. In parallel developments, research on animal personality has identified personality components that could be interpreted in terms of (subsets of) the FFM in animals, whereby most overlap was found in humans' closest relatives, chimpanzees (Gosling and John, 1999).

To what extent is there a single, species-wide model of personality in humans? The positions taken on this question vary. Historically, the discussion goes back to the emic–etic distinction, introduced by Pike (1967). 'Etic' refers to research that studies cross-cultural differences from a comparative perspective (studying each individual culture less profoundly), whereas 'emic' refers to research that examines a specific culture in more detail with no (or only a secondary) cross-cultural focus. An etic, universalistic, perspective in personality can be found in five-factor theory (McCrae and Costa, 2008), where the dimensions of personality are assumed to be biologically based and universal, whereas specific manifestations of the model's components are shaped by culture. From a more relativistic perspective, personality can only be understood in cultural and historical context (Piekkola, 2011). There are methodological approaches, such as the psycholexical approach (discussed in a following section), which have no a priori position on the universality of the FFM or other models, but provide important tools for addressing the topic. Finally, there are theoretical and methodological approaches, labeled *emic–etic* (Cheung et al., 2011), that seek to combine universalistic and relativistic approaches. Emic–etic approaches seek to integrate evidence on more universal and more culture-specific elements in personality models.

We review the empirical evidence on the FFM accumulated from three lines of research: (1) questionnaire studies that sought to replicate the FFM across cultures; (2) psycholexical studies that sought to establish implicit models of personality in different languages using lexica; and (3) indigenous and emic–etic studies that sought to establish personality structures in specific cultures using a variety of methods, and subsequently assess the overlap between these models and the FFM. In the first two sections, we start from the evidence in support of FFM universality and progress toward divergent evidence, while also commenting on methodological limitations and

shortcomings. In the final section, we present findings on indigenous concepts and their degree of overlap with the FFM in several non-Western cultures.

Questionnaire Studies

In the questionnaire study design, participants respond to a list of questions about their typical characteristics, behaviors, feelings, and thoughts. Items include abstract trait terms (e.g., 'I am talkative') as well as behavioral terms (e.g., 'I make friends easily'). Responses are factor-analyzed, and the resulting structure is compared with an existing factor structure, typically derived from US data. Statistical criteria are used to evaluate agreement by using congruence coefficients that indicate the degree of equivalence between the two structures (interpreted as supportive evidence for the identity of factors across the cultures studied). Congruence coefficients of .95 and higher indicate equivalent structures (Lorenzo-Seva and Ten Berge, 2006; Van de Vijver and Leung, 1997), although values of .85 or higher are sometimes also taken as evidence of similarity. A similar procedure can be used to compare individual-level to country-level factor structures (Van de Vijver and Poortinga, 2002).

The most popular inventory for measuring the FFM is the Revised NEO Personality Inventory (NEO-PI-R; Costa and McCrae, 1992), which measures 5 broad domains and 30 lower level facets, 6 per domain (i.e., the FFM in the narrow sense, mentioned in the Introduction). Another popular instrument is the Big Five Inventory (BFI; John et al., 1991). Both instruments have been translated into a large number of languages from different families, using back-translation procedures. The self-report form of the NEO-PI-R has been administered in 36 cultures (McCrae, 2002), and the peer-report in 50 cultures (McCrae, Terracciano, and 78 Members of the Personality Profiles of Cultures Project, 2005a). The BFI has been administered in 56 cultures (Schmitt et al., 2007). All these studies tended to find fair to very good replicability of the factors of the FFM in individual cultures and for the total set of cultures compared with the US structure. The general replicability between the individual- and culture-level factor structure, together with evidence for the external validity of culture-level scores (e.g., their associations with cultural values and economic indices), has enabled researchers to construct 'personality profiles' of cultures (McCrae, Terracciano, and 79 Members of the Personality Profiles of Cultures Project, 2005b) and to explore the geographical distribution of Big Five traits (Schmitt et al., 2007). For example, Schmitt et al. (2007) found that average levels of Extraversion were lower in East Asia than in other regions. Finally, broad support for the replicability of the FFM has also been found in studies with less extensive culture coverage, both using other instruments designed to measure the FFM (Caprara et al., 2000; De Fruyt et al., 2009; Hendriks et al., 2003) and using instruments that had not been initially devised as a measure of the FFM, including nonverbal instruments (Paunonen et al., 1996).

Several limitations on the support for the universality of the FFM from questionnaire studies require attention. First, although the five factors generally tend to replicate, not all replications pass accepted thresholds for construct equivalence (Lorenzo-Seva and Ten Berge, 2006; Van de Vijver and Leung, 1997). More discrepancies tend to be found at lower levels of aggregation (i.e., for individual countries and country regions) than for global solutions. In McCrae et al. (2005a), there were a number of non-Western, especially African, countries with congruence coefficients below .80, especially for Openness. The authors attributed the finding to shortcomings in data quality rather than insufficiency of the model for these countries. In Schmitt et al. (2007), several coefficients especially in Africa and South and Southeast Asia were below .90, with the lowest one of .84 for Openness in Africa. Second, although the culture-level five factors map fairly closely on the individual-level factors, the two structures are not identical. On culture level, Extraversion was codefined by a number of facets from other factors (McCrae et al., 2005b), and the correlations between the NEO-PI-R and BFI factors were less supportive of the convergent and discriminant validity of the factors than on individual level (Schmitt et al., 2007). The implication is that interpretations of individual differences cannot be transferred simply to interpretations of country mean differences. Third, although the structural equivalence of the five factors on individual level has broadly been established, there is no comparable support for measurement equivalence but there is some evidence for inequivalence (Church et al., 2011). The implication is that direct numerical comparisons of trait scores across cultures are not yet warranted: Individual differences in the five factors do not have the same meaning across countries. Fourth, the questionnaire study design has the inherent limitation that it can only be used in a straightforward manner in literate populations. Most of the studies have been conducted on student samples. In one of the rare studies on preliterate societies, Gurven et al. (2013) failed to replicate the FFM using the BFI. The authors interpreted their finding as evidence that the FFM may not be applicable to the Tsimane ethnic group in the Bolivian Amazon and likely to other small-scale, preliterate societies; a more conservative interpretation would involve issues of method bias and inadequacy of questionnaire adaptation. Finally, an inherent limitation of questionnaires is that they tend to impose structure on the data to a certain extent (Church, 2000). Support for cross-cultural replicability has been obtained also for other models, measured with respective questionnaires, such as the Eysenck Personality Questionnaire (EPQ; Barrett et al., 1998). While questionnaires offer a powerful tool for replicating factor structures, they are less efficient as tools for discovering factors implicit in different cultures and languages. Such instruments cannot address the question whether the items are the best possible rendering of personality in that culture and whether culture-specific aspects were not left out.

Psycholexical Studies

Research in the psycholexical (or *lexical*) tradition is based on the assumption that the most important features of personality become encoded as single terms in language ('psycholexical hypothesis'; Goldberg, 1981). Single personality-descriptive terms can be derived from lexica and subsequently reduced through content and factor analysis to arrive at a few broad dimensions of personality. A lexical study begins with an exhaustive extraction of personality terms from current

dictionaries. In the stricter and more common version of the approach, only terms referring to internal psychological characteristics are selected, whereas terms that are purely evaluative, refer to physical appearance, beliefs, temporary states, or do not describe personality are excluded (Saucier, 2008). The list is further shortened by selecting the terms that appear most frequently in natural language usage. Participants are presented with the selected trait list and asked to assess to what extent each term is applicable to them (or to a target person they know well). The scores of the participants are factor-analyzed, resulting in models with a few dimensions that account for the covariation of trait terms. Because of the comprehensive selection of trait terms, this approach is less restrictive than the questionnaire approach in the dimensions it can identify and is thought to be more representative of the implicit structure of personality in each language.

The Big Five model was first identified in the outcomes of lexical research in English (Goldberg, 1990). Lexical studies have since been conducted in about 15 other Indo-European as well as non-Indo-European languages. Direct empirical comparisons and conceptual overviews of the lexical data across languages have not converged on a single model replicable across languages (De Raad et al., 2010; Saucier, 2008). Although factors similar to those of the Big Five have been identified across languages, the Big Five as a coherent structure defined by common elements has only been replicated in the Germanic languages (Saucier and Goldberg, 2001). In an analysis on data from 12 languages obtained according to the common premises (e.g., full rather than random extraction of terms, exclusion of purely evaluative terms), with explicit definition of common terms defining factors, and direct psychometric assessment of structure equivalence (with a liberal cutoff value for congruence coefficients at .80), De Raad et al. (2010) found that only three factors were fully replicable across languages: Extraversion, Agreeableness, and Conscientiousness. These factors were more broadly defined than the corresponding Big Five factors because they included terms from the other factors. These three factors tend to be the three largest factors in lexical studies. There is less agreement across languages on factors beyond these three.

Some limitations of the psycholexical approach are effected from the opposing aims of arriving at a common model for implicit personality concepts across languages while providing a comprehensive representation of these concepts. The first aim drives research toward the identification of fewer personality factors, whereas the second drives research toward more factors (De Raad et al., 2010). Comparability of factors across languages is challenged by procedural differences between studies with respect to the selection of data sets and of terms to be analyzed, item translation, and decisions regarding factor extraction and rotation. In addition, there is the more important limitation that different sets of items provide the input for the analyses in different languages. The factors are usually defined by a limited number of highly loading items, called *markers*. These markers tend not to overlap across languages, and decisions have to be made as to which markers to use for factor identification and subsequent comparisons. These decisions can be biased by prior expectations, for example, of retrieving the Big Five factors. A classic example is that of Assertiveness terms. In the English Big Five structure, Assertiveness is a facet of the first factor, Extraversion. When, in other languages, Emotional Stability fails to emerge as a fourth factor, Assertiveness terms loading on the fourth factor can be taken to indicate Emotional Stability because this is the expected factor. Such interpretations amount to circular reasoning in assuming what has to be shown (De Raad and Peabody, 2005).

A different set of limitations apply to the conceptual premises of the psycholexical approach. Although it is reasonable to expect that trait terms are used universally, it has not been established that they are equally frequent across languages and perceived as equally informative of personality across cultures. There is evidence that abstract trait terms are used more often in personality descriptions in Western groups, whereas specific behavioral descriptions are preferred in non-Western groups (Del Prado et al., 2007; Valchev et al., 2013). It may also be the case that some personality domains are less saturated with single abstract trait terms, which could explain the weak replicability of the smaller factors, Neuroticism and Openness, in lexical studies as opposed to questionnaire studies. In more practical terms, the psycholexical approach is not directly applicable to languages where there are wide variations between a common written language and spoken dialects, as in Arabic, and to languages that have no developed lexicography.

Indigenous and Emic–Etic Studies

In contrast to the etic studies that seek to replicate personality structure across cultures, such as the NEO-PI-R replicability studies, indigenous or emic studies focus on the identification of personality concepts perceived as relevant in specific, typically non-Western cultures. Emic studies employ an array of methods including psycholexical methods, ethnographic interviews, and reviews of folklore and literature for personality characteristics. The strength of this kind of research is in the comprehensive representation of the implicit personality models, not bound by questionnaire item selections or exclusive reliance on lexical records.

There is some work on individual, purely indigenous (emic) personality concepts, such as the East-Asian, religion-based selfless self (Ho, 1995) and the Japanese Amae (Doi, 1973). However, this work has never developed into full-fledged personality models. When the personality concepts derived in indigenous research are compared with those from other cultures or with supposed universal models, more confident conclusions on the generalizability and universality of personality concepts can be drawn. Some of the major findings from indigenous research in different parts of the world are presented below.

In the Philippines, indigenous research has identified a number of implicit personality dimensions represented in different models (Guanzon-Lapeña et al., 1998). Katigbak and colleagues (Katigbak et al., 1996, 2002) examined the indigenous dimensions jointly with the FFM and found that most indigenous concepts could be subsumed in the FFM. There were only a few concepts that were not empirically encompassed by the FFM, such as social curiosity, risk-taking, and religiosity (Katigbak et al., 2002). Importantly, the indigenous

concepts offered only modest incremental criterion validity above the FFM.

In Mexico, Ortiz et al. (2007) reviewed previous literature on indigenous concepts and examined the relations of a large number of indigenous measures with the FFM. They found that most of the Mexican concepts could be well subsumed within the FFM. Concepts of warmth and affection appeared to be more salient in the indigenous measures, whereas Openness could not be identified in the indigenous measures but only in the imported FFM measures. Overall, both in the Philippines and in Mexico, researchers were led to conclude that cultural differences may not be expressed in the clear identification of dimensions beyond the Big Five, but rather in the salience and composition of elements that can all be subsumed in the Big Five.

In China, several independent lines of research have explored the indigenous structure of personality (reviewed in Yang, 2006). The best established questionnaire that resulted from one of these lines of research is the Cross-Cultural (initially Chinese) Personality Assessment Inventory (CPAI-2; Cheung et al., 2003). Developed with input from interviews and reviews of wide-ranging literature sources, this inventory identified one dimension that could not be subsumed within the FFM, Interpersonal Relatedness, which involves aspects of functioning in social context and adherence to tradition. Subsequent studies have replicated the Interpersonal Relatedness factor in non-Chinese populations, including the US (Lin and Church, 2004), suggesting that although Interpersonal Relatedness is most salient in China, it is not unique to China but is recognizable in other cultures. Openness, on the other hand, has not been identified in China using indigenous measures but only using imported measures (Cheung et al., 2008). So, although single concepts referring to Openness are used in the Chinese context, they are not implicitly recognized as constituting a coherent personality dimension. This pattern is opposite to that of Interpersonal Relatedness. Finally, it is noteworthy that the Interpersonal Relatedness dimension has been found to offer incremental validity in the prediction of culturally relevant outcomes in China (Cheung, Cheung, and Fan, 2013; Zhang and Bond, 1998).

In South Africa, a culturally heterogeneous country with 11 official languages, recent indigenous research has identified nine clusters of personality concepts derived from free personality descriptions (Nel et al., 2012). All of the Big Five factors were recognizable within the nine clusters, but the nine clusters included a large number of concepts dealing with interpersonal relations. This finding is consistent with findings from lexical and other indigenous studies, where Agreeableness-related terms normally form the largest factor. The quantitative validation of the conceptual model is currently ongoing, but there are some initial indications that concepts related to social-relational functioning in a fairly norm-regulated context are salient in South Africa, yet not well covered by the FFM.

Conclusion

Support for the universality of the FFM is strongest when FFM-based questionnaires are used and tends to be stronger in Western than in non-Western cultures. Four out of the five factors are usually well replicated (Neuroticism, Extraversion, Agreeableness, and Conscientiousness), whereas the fifth factor, Openness, is not always found. From a psycholexical perspective, the replicability of three broad implicit dimensions is best established: Extraversion, Agreeableness, and Conscientiousness. The nonreplication of the FFM cannot be easily attributed to a common pattern of causes; data quality, availability of abstract terms in the language, and cultural salience of the concepts are often involved. Despite some strong proposals (e.g., Buss, 1991; McCrae and Costa, 2008), there is no commonly accepted theory of the Big Five or FFM that would account for the universal occurrence of the five factors or their causal role in behavior. In fact, the role of the Big Five factors as causal entities is periodically challenged (e.g., Cramer et al., 2012). The lack of such a commonly accepted theory can be perceived as contributing to the dangers of circular reasoning (De Raad and Peabody, 2005). Nonetheless, research on the universality of the FFM has produced invaluable data and rich insights into the patterns of similarities and differences in personality across cultures.

See also: Big Five Factor Model, Theory and Structure; Cross-Cultural Psychology; Cross-Cultural Research Methods in Sociology; Cultural Psychology; Cultural Relativism, Anthropology of; Five Factor Model of Personality, Assessment of; Five Factor Model of Personality, Facets of; Indigenous Psychology.

Bibliography

Allport, G.W., Odbert, H.S., 1936. Trait-names: a psycho-lexical study. Psychological Monographs 47 (211), 1–171.

Barrett, P.T., Petrides, K., Eysenck, S.B.G., Eysenck, H.J., 1998. The Eysenck Personality Questionnaire: an examination of the factorial similarity of P, E, N, and L across 34 countries. Personality and Individual Differences 25, 805–819. http://dx.doi.org/10.1016/S0191-8869(98)00026-9.

Buss, D.M., 1991. Evolutionary personality psychology. Annual Review of Psychology 42, 459–491. http://dx.doi.org/10.1146/annurev.ps.42.020191.002331.

Caprara, G.V., Barbaranelli, C., Bermúdez, J., Maslach, C., Ruch, W., 2000. Multivariate methods for the comparison of factor structures in cross-cultural research: an illustration with the Big Five questionnaire. Journal of Cross-Cultural Psychology 31. http://dx.doi.org/10.1177/0022022100031004002.

Cheung, F.M., Cheung, S.F., Fan, W., 2013. From Chinese to cross-cultural personality inventory: A combined emic–etic approach to the study of personality in culture. In: Gelfand, M., Chiu, C.-Y., Hong, Y.-Y. (Eds.), Advances in culture and psychology, Vol. III. Oxford University Press, New York, pp. 117–179.

Cheung, F.M., Cheung, S.F., Leung, K., Ward, C., Leong, F., 2003. The English version of the Chinese personality assessment inventory. Journal of Cross-Cultural Psychology 34, 433–452. http://dx.doi.org/10.1177/0022022103034004004.

Cheung, F.M., Cheung, S.F., Zhang, J., Leung, K., Leong, F., Yeh, K.-H., 2008. Relevance of openness as a personality dimension in Chinese culture: aspects of its cultural relevance. Journal of Cross-Cultural Psychology 39, 81–108. http://dx.doi.org/10.1177/0022022107311968.

Cheung, F.M., Van de Vijver, F.J.R., Leong, F.T.L., 2011. Toward a new approach to the assessment of personality in culture. American Psychologist 66, 593–603. http://dx.doi.org/10.1037/a0022389.

Church, A.T., 2000. Culture and personality: toward an integrated cultural trait psychology. Journal of Personality 68, 651–703. http://dx.doi.org/10.1111/1467-6494.00112.

Church, A.T., Alvarez, J.M., Mai, N.T.Q., French, B.F., Katigbak, M.S., Ortiz, F.A., 2011. Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the revised NEO personality inventory. Journal of Personality and Social Psychology 101, 1068–1089. http://dx.doi.org/10.1037/a0025290.

Costa Jr., P.T., McCrae, R.R., 1992. Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual. Psychological Assessment Resources, Odessa, FL.

Cramer, A.O.J., Van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S.H., Kendler, K.S., Borsboom, D., 2012. Dimensions of normal personality as networks in search of equilibrium: you can't like parties if you don't like people. European Journal of Personality 26, 414–431. http://dx.doi.org/10.1002/per.1866.

De Fruyt, F., De Bolle, M., McCrae, R.R., Terracciano, A., Costa Jr., P.T., Collaborators of the Adolescent Personality Profiles of Cultures Project, 2009. Assessing the universal structure of personality in early adolescence. Assessment 16, 301–311. http://dx.doi.org/10.1177/1073191109333760.

De Raad, B., Barelds, D.P., Levert, E., Ostendorf, F., Mlačić, B., Di Blas, L., Katigbak, M.S., 2010. Only three factors of personality description are fully replicable across languages: a comparison of 14 trait taxonomies. Journal of Personality and Social Psychology 98, 160–173. http://dx.doi.org/10.1037/a0017184.

De Raad, B., Peabody, D., 2005. Cross-culturally recurrent personality factors: analyses of three factors. European Journal of Personality 19, 451–474. http://dx.doi.org/10.1002/per.550.

Del Prado, A.M., Church, A.T., Katigbak, M.S., Miramontes, L.G., Whitty, M.T., Curtis, G.J., Reyes, J.A.S., 2007. Culture, method, and the content of self-concepts: testing trait, individual-self-primacy, and cultural psychology perspectives. Journal of Research in Personality 41, 1119–1160. http://dx.doi.org/10.1016/j.jrp.2007.02.002.

Doi, T.L., 1973. The Anatomy of Dependence. Kodansha International, Tokyo, Japan.

Goldberg, L.R., 1981. Language and individual differences: the search for universals in personality lexicons. In: Wheeler, L. (Ed.), Review of Personality and Social Psychology, vol. 2. Sage, Beverly Hills, CA, pp. 141–165.

Goldberg, L.R., 1990. An alternative "description of personality": the big-five factor structure. Journal of Personality and Social Psychology 59, 1216–1229. http://dx.doi.org/10.1037/0022-3514.59.6.1216.

Gosling, S.D., John, O.P., 1999. Personality dimensions in nonhuman animals: a cross-species review. Current Directions in Psychological Science 8, 69–75. http://dx.doi.org/10.1111/1467-8721.00017.

Guanzon-Lapeña, M.A., Church, A.T., Carlota, A.J., Katigbak, M.S., 1998. Indigenous personality measures: Philippine examples. Journal of Cross-Cultural Psychology 29, 249–270. http://dx.doi.org/10.1177/0022022198291013.

Gurven, M., Von Rueden, C., Massenkoff, M., Kaplan, H., Lero Vie, M., 2013. How universal is the Big Five? Testing the five-factor model of personality variation among forager–farmers in the Bolivian Amazon. Journal of Personality and Social Psychology 104, 354–370. http://dx.doi.org/10.1037/a0030841.

Hendriks, A.A.J., Perugini, M., Angleitner, A., Ostendorf, F., Johnson, J.A., De Fruyt, F., Ruisel, I., 2003. The five-factor personality inventory: cross-cultural generalizability across 13 countries. European Journal of Personality 17, 347–373. http://dx.doi.org/10.1002/per.491.

Ho, D.Y., 1995. Selfhood and identity in Confucianism, Taoism, Buddhism, and Hinduism: contrasts with the West. Journal for the Theory of Social Behaviour 25, 115–139. http://dx.doi.org/10.1111/j.1468-5914.1995.tb00269.x.

John, O.P., Donahue, E.M., Kentle, R.L., 1991. The "Big Five" Inventory – Versions 4a and 54. University of California, Berkeley, CA. Institute of Personality and Social Research, Berkeley.

Katigbak, M.S., Church, A.T., Akamine, T.X., 1996. Cross-cultural generalizability of personality dimensions: Relating indigenous and imported dimensions in two cultures. Journal of Personality and Social Psychology 70, 99–114. http://dx.doi.org/10.1037/0022-3514.70.1.99.

Katigbak, M.S., Church, A.T., Guanzon-Lapeña, M.A., Carlota, A.J., Del Pilar, G.H., 2002. Are indigenous dimensions culture-specific? Philippine inventories and the five-factor model. Journal of Personality and Social Psychology 82, 89–101. http://dx.doi.org/10.1037/0022-3514.82.1.89.

Lin, E.J.-L., Church, A.T., 2004. Are indigenous Chinese personality dimensions culture-specific? An investigation of the Chinese Personality Assessment Inventory in Chinese American and European American samples. Journal of Cross-Cultural Psychology 35, 586–605. http://dx.doi.org/10.1177/0022022104268390.

Lorenzo-Seva, U., Ten Berge, J.M.F., 2006. Tucker's congruence coefficient as a meaningful index of factor similarity. Methodology 2, 57–64. http://dx.doi.org/10.1027/1614-1881.2.2.57.

McCrae, R.R., 2002. NEO-PI-R data from 36 cultures: further intercultural comparisons. In: McCrae, R.R., Allik, J. (Eds.), The Five-Factor Model of Personality across Cultures. Kluwer Academic/Plenum, New York, pp. 105–125.

McCrae, R.R., Costa Jr., P.T., 1997. Personality trait structure as a human universal. American Psychologist 52, 509–516. http://dx.doi.org/10.1037/0003-066X.52.5.509.

McCrae, R.R., Costa Jr., P.T., 2008. The five-factor theory of personality. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research, third ed. Guilford Press, New York, pp. 159–181.

McCrae, R.R., Terracciano, A., 78 Members of the Personality Profiles of Cultures Project, 2005a. Universal features of personality traits from the observer's perspective: data from 50 cultures. Journal of Personality and Social Psychology 88, 547–561. http://dx.doi.org/10.1037/0022-3514.88.3.547.

McCrae, R.R., Terracciano, A., 79 Members of the Personality Profiles of Cultures Project, 2005b. Personality profiles of cultures: aggregate personality traits. Journal of Personality and Social Psychology 89, 407–425. http://dx.doi.org/10.1037/0022-3514.89.3.407.

Nel, J.A., Valchev, V.H., Rothmann, S., Van de Vijver, F.J.R., Meiring, D., De Bruin, G.P., 2012. Exploring the personality structure in the 11 languages of South Africa. Journal of Personality 80, 915–948. http://dx.doi.org/10.1111/j.1467-6494.2011.00751.x.

Norman, W.T., 1963. Toward an adequate taxonomy of personality attributes: replicated factor structure in peer nomination personality ratings. Journal of Abnormal and Social Psychology 66, 574–583. http://dx.doi.org/10.1037/h0040291.

Ortiz, F.A., Church, A.T., Vargas-Flores, J.J., Ibáñez-Reyes, J., Flores-Galaz, M., Iuit-Briceño, J.I., Escamilla, J.M., 2007. Are indigenous personality dimensions culture-specific? Mexican inventories and the five-factor model. Journal of Research in Personality 41, 618–649. http://dx.doi.org/10.1016/j.jrp.2006.07.002.

Paunonen, S.V., Keinonen, M., Trzebinski, J., Forsterling, F., Grishenkoroze, N., Kouznetsova, L., Chan, D.W., 1996. The structure of personality in six cultures. Journal of Cross-Cultural Psychology 27, 339–353. http://dx.doi.org/10.1177/0022022196273006.

Piekkola, B., 2011. Traits across cultures: a neo-Allportian perspective. Journal of Theoretical and Philosophical Psychology 31, 2–24. http://dx.doi.org/10.1037/a0022478.

Pike, K.L., 1967. Language in Relation to a Unified Theory of Structure of Human Behavior, second ed. Mouton, The Hague, the Netherlands.

Saucier, G., 2008. Measures of the personality factors found recurrently in human lexicons. In: Boyle, G.J., Matthews, G., Saklofske, D.H. (Eds.), The SAGE Handbook of Personality Theory and Assessment, vol. 2. SAGE, Los Angeles, CA, pp. 29–54.

Saucier, G., Goldberg, L.R., 2001. Lexical studies of indigenous personality factors: premises, products, and prospects. Journal of Personality 69, 847–879. http://dx.doi.org/10.1111/1467-6494.696167.

Schmitt, D.P., Allik, J., McCrae, R.R., Benet-Martínez, V., et al., 2007. The geographic distribution of Big Five personality traits: patterns and profiles of human self-description across 56 nations. Journal of Cross-Cultural Psychology 38, 173–212. http://dx.doi.org/10.1177/0022022106297299.

Stagner, R., 1937. Psychology of Personality. McGraw-Hill, New York.

Tupes, E.C., Christal, R.E., 1961. Recurrent Personality Factors Based on Trait Ratings (USAF Tech. Rep. No. 61–97). Lackland Air Force Base, San Antonio, TX.

Valchev, V.H., Van de Vijver, F.J.R., Nel, J.A., Rothmann, S., Meiring, D., 2013. The use of traits and contextual information in free personality descriptions across ethnocultural groups in South Africa. Journal of Personality and Social Psychology 104, 1077–1091. http://dx.doi.org/10.1037/a0032276.

Van de Vijver, F.J.R., Leung, K., 1997. Methods and Data Analysis for Cross-Cultural Research. SAGE, Thousand Oaks, CA.

Van de Vijver, F.J.R., Poortinga, Y.H., 2002. Structural equivalence in multilevel research. Journal of Cross-Cultural Psychology 33, 141–156. http://dx.doi.org/10.1177/0022022102033002002.

Yang, K.-S., 2006. Indigenous personality research: the Chinese case. In: Kim, U., Yang, K.-S., Hwang, K.-K. (Eds.), Indigenous and Cultural Psychology: Understanding People in Context. Springer, New York, pp. 285–314.

Zhang, J., Bond, M.H., 1998. Personality and filial piety among college students in two Chinese societies: the added value of indigenous constructs. Journal of Cross-Cultural Psychology 29, 402–417. http://dx.doi.org/10.1177/0022022198293002.

Flashbulb Memories

Daniel L Greenberg, College of Charleston, Charleston, SC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A *flashbulb memory* is a vivid and enduring memory for the circumstances in which one learned of a momentous public event (Brown and Kulik, 1977). For example, many people have vivid memories of the way in which they heard about the assassination of John F. Kennedy, the resignation of Margaret Thatcher, the explosion of the space shuttle *Challenger*, or the September 11th attacks. The idea has substantial appeal, and several articles in the popular media have discussed it (Berendt, 1973; Pew Research Center, 2002). Other studies, however, have shown that flashbulb memories are not as immutable as they seem. More recently, researchers have begun to focus on the phenomenological properties of these memories as well as the social and cultural factors that influence this kind of recall.

An Early Study

Colegrove (1900) conducted the first reported study of these memories. He did not refer to these as flashbulb memories (the photographic flash was not in widespread use at the time), but his methods and data clearly show that he was investigating the same phenomenon. In the course of a wide-ranging survey of memories and memory types, he asked people if they had vivid memories of how they heard about the assassination of President Lincoln over 30 years before. Here is the first example he provides:

> … When we were driving down a steep hill into the city we felt that something was wrong. Everybody looked so sad and there was such terrible excitement that my father stopped his horse, and, leaning from the carriage, called 'What is it, my friends? What has happened?' 'Haven't you heard?' was the reply – 'Lincoln has been assassinated.' The lines fell from my father's limp hands, and with tears streaming from his eyes he sat as one bereft of motion … (Colegrove, 1900: p. 263)

Colegrove (1900) acknowledges that most of the memories were not this detailed, and in fact the rest of his examples contain no more than the bare particulars. Nonetheless, he found these particulars noteworthy. He defined a vivid memory as one that contained information about the time, the location, and the informant; he found that 71% of participants (127/179) could report such a memory. Some of the remainder (Colegrove does not say how many) had a vivid memory for the assassination of President James Garfield in 1881. While this percentage might seem a bit low – after all, a flashbulb memory is of an event that 'everybody remembers' – Colegrove's (1900) data nonetheless show that many people can retain clear and detailed memories a full three decades after the original event.

The First Theory

After Colegrove (1900), the phenomenon seems to have gone unstudied (at least in the psychological literature) until 1977, when Brown and Kulik published the article that named the phenomenon. Their experiments investigated memories for a series of shocking and consequential events, including the assassination of President John F. Kennedy and Martin Luther King, Jr. In their words:

> [T]here is something strange about this recall… . Almost everyone testifies that his recall of his circumstances … has a primary, "live" quality that is almost perceptual. Indeed, it is very like a photograph that indiscriminately preserves the scene in which each of us found himself when the flashbulb was fired. (Brown and Kulik, 1977)

Brown and Kulik (1977) quickly tempered this description by noting that the memory was not quite as indiscriminate as a photograph, since many extraneous details were not in fact preserved. Rather, they suggested that flashbulb memories tended to focus on several canonical features. These included the place, the informant, and the activities that were going on at the time (as Colegrove (1900) had suggested), but also the participants' reaction, others' reactions, and the immediate aftermath. In spite of this, a memory could be classified as a flashbulb if it contained information from only one of these categories. That is, if a participant only remembered where they were, but not what they were doing, who told them, how anyone reacted, or what they did afterward, they would still be treated as having a flashbulb memory.

Brown and Kulik (1977) also attempted to identify the kinds of events that led to the formation of a flashbulb memory. First came *novelty* or surprise: the event had to be unexpected, like an assassination or a sudden eruption of scandal. Second, the event had to be seen as *consequential* for the person's life. According to their definition, consequentiality meant something distinct from personal or historical importance. Rather, consequentiality takes on something close to a biological significance; an event was defined as consequential to a person if his or her life was changed in some noteworthy way. (Brown and Kulik (1977) operationalized this property by asking people to think about how their lives would have been different if the event had not happened.) Finally, they observed that both of these factors could lead to an increased *rehearsal*: people tend to talk and think about consequential and novel events, and rehearsal can of course improve memory.

To tie all of these features together, Brown and Kulik (1977) proposed a theoretical explanation for flashbulb memories. In particular, they drew an analogy with Livingston's (1967) "Now Print!" hypothesis. They speculated that the human brain has evolved a 'special mechanism' to fix key events in memory. According to this view, under some circumstances – specifically, when an event is surprising and consequential – the special mechanism is activated and the 'flashbulb' fires; under other circumstances, it is not and does not. As Wright (2009) has suggested, this hypothesis implicitly assumes that flashbulb memories are qualitatively different from everyday memories (see Curci, 2009, for other methodological issues). That is, they are not just points on the extreme end of the vividness scale; they are a different type of memory, perhaps involving different cognitive and neural mechanisms.

Brown and Kulik (1977) maintained that this theory provided a good explanation for their data. People tended to find the assassinations and political scandals surprising, and they tended to have flashbulb memories for how they heard the news. (Note, however, that as with Colegrove's (1900) participants, not everyone had flashbulb memories for these events. In fact, fewer than half of Brown and Kulik's (1977) participants had flashbulb memories for the deaths of Malcolm X, Medgar Evers, and George Wallace.) The deaths of Martin Luther King and other civil rights leaders were presumed to be more consequential for black Americans than for white Americans, and, in at least some cases, black Americans were more likely to have flashbulb memories for these events than white people were (see Wright, 2009 for some criticisms of their data). People tended to report rehearsing these memories. Overall, Brown and Kulik (1977) seemed to have presented a plausible explanation of an intriguing phenomenon.

The 'flashbulb' metaphor is compelling for several reasons. First, from a general evolutionary perspective, it seems at least plausible that we might have a special mechanism for the creation of a vivid memory; it might help us avoid similar dangerous situations in the future. Second, while Livingston's (1967) proposal was speculative at the time, it does seem to be consistent with a substantial body of later research showing that emotional material (particularly negatively valenced material) is remembered better than neutral material. Third, the idea is consistent with our own intuitions: these memories really do seem different from other memories, we really do think and talk about them a lot, and we can rarely remember other events with such apparent clarity. In short, the idea made a great deal of sense.

Accuracy, Consistency, and Challenges to the 'Special Mechanism'

This apparently sensible proposal soon received a great deal of criticism. Specifically, contrary to the plain meaning of the metaphor, flashbulb memories are not extraordinarily accurate or consistent. Note that accuracy and consistency are not the same, nor are they necessarily correlated. A memory can be consistent without being accurate, as when we remember the same inaccurate information over and over again. Counterintuitively, a memory can also be accurate but not consistent, depending on how one measures accuracy (see Julian et al., 2009 for further discussion). Suppose that a student heard about the September 11th attacks while sitting in a psychology class and sending a text message to a friend. Upon first recalling the event, the student only reports sitting in class. At the second recall, only the text message is reported. The memories are entirely inconsistent, but they are accurate (albeit incomplete). We know relatively little about the accuracy of flashbulb memories, since researchers are rarely collecting data when the original event happens. At best, we can identify a few cases in which the memory could not have happened as remembered (Linton, 1975; Neisser, 1982; Greenberg, 2004).

Consistency is somewhat easier to measure. Neisser and Harsch (1992) reported inconsistencies in flashbulb memories for the Space Shuttle *Challenger* disaster, as did McCloskey et al. (1988). Talarico and Rubin (2003) compared flashbulb memories for 9/11 to memories of noteworthy personal events that had taken place the week before; they found that flashbulb memories were no more consistent than the personal memories, though participants did have greater confidence in the accuracy of the flashbulbs. Schmolck et al. (2000) showed that memories for the verdict in the O.J. Simpson trial were inconsistent over time. Crombag et al. (1999) showed that participants could falsely remember nonexistent footage of a plane crash. Greenberg (2004) reported that President Bush had done the same for the September 11th attacks. Neisser (1982) observed that his own recollection of Pearl Harbor was inaccurate (see Thompson and Cowan, 1986; Neisser, 1986, for discussion). Researchers do not agree about the degree of inconsistency in flashbulb memories, in part because almost every study uses a different measurement method. Nevertheless, the majority of the evidence suggests that some degree of inconsistency exists.

Furthermore, flashbulb memories seem to be susceptible to the same kinds of errors as other memories. Neisser and Harsch (1992) found that many people came to believe that they had learned about the *Challenger* disaster on television, even though they originally stated that they had heard about it another way. For example, one participant originally reported that she overheard people talking about it in her religion class; 2 years later, she claimed that she learned about it while watching TV with her roommate. Neisser and Harsch (1992) called this phenomenon 'TV priority' and suggested several possible reasons for it. For one, most if not all of the participants did see the news on television at some point. The TV coverage was particularly compelling; the image of the exploding shuttle was played over and over, creating a memory that was far more coherent and vivid than the original circumstances (see Figure 1 for the iconic image of the event).

This is an example of a 'wrong time slice' error (Brewer, 1988); participants remembered something that actually happened, but mislocated it in time. Similarly, Neisser and Harsch (1992) suggested that the participants might have been relying on a script or schema for learning about shocking events. That is, people typically learn events by learning about them on television (or, in today's world, the Internet), so they look for such memories when trying to reconstruct how they themselves heard the news. In short, flashbulb memories seem to be susceptible to the same sorts of errors as everyday memories.

Figure 1 Many people have a flashbulb memory for the explosion of the space shuttle *Challenger*. Source: National Aeronautics and Space Administration (http://grin.hq.nasa.gov/ABSTRACTS/GPN-2004-00012.html).

These findings have led many researchers to reject the notion of a special mechanism. McCloskey et al. (1988), for instance, suggested that there was no reason to believe in a unique flashbulb-memory mechanism that differed from the mechanisms involved in everyday recall. Rather, the qualities of the event (the shock, the disruption of the regular routine, the personal significance, etc.) probably make the event more novel and less susceptible to interference, and these properties are a well-established part of normal memory (Cohen et al., 1988, 1990; Pillemer, 1990; Schmidt and Bohannon, 1988; Talarico and Rubin, 2009; Wright, 2009). In accordance with this view, flashbulb memories are no more vivid than the most vivid of autobiographical memories (Rubin and Kozin, 1984), and they therefore belong on the continuum of autobiographical memories, not in a category of their own.

Beyond the Special Mechanism

For many researchers (though not all), flashbulb memories still merit investigation even if no special mechanism exists. Talarico and Rubin (2007, 2009) have suggested that the phenomenology of flashbulb memories is unusual, even if the consistency is not (Rubin and Kozin, 1984). In fact, flashbulb memories are interesting in part precisely because the phenomenology and the consistency diverge (Talarico and Rubin, 2003). In addition, one can examine the factors that lead to the development of flashbulb memories, whether or not they are accurate or consistent. Many studies have been conducted along these lines, and several of them have further challenged the proposals that Brown and Kulik (1977) originally made.

Novelty and Surprise

Several models of flashbulb memory have suggested that novelty and surprise are important features (Brown and Kulik, 1977; Conway et al., 1994; Finkenauer et al., 1998; Er, 2003; see Luminet, 2009, for a review). As Talarico and Rubin (2009) have argued, novelty and surprise are not the same – surprise is an emotional reaction, whereas novelty is a property of the event itself – though not all models draw a clear distinction between the two. Overall, the evidence clearly shows that flashbulb memories still occur for long-anticipated and therefore unsurprising events, such as the moon landing (Winograd and Killinger, 1983), the death of former French president François Mitterand (Curci et al., 2001), the US invasion of Iraq in 1991 (Weaver, 1993), and the death of Pope John Paul II (Tinti et al., 2009). Surprise may increase the likelihood of the formation of a flashbulb (Pillemer, 1984; Christianson, 1989; Davidson and Glisky, 2002), though not always (Otani et al., 2005; Berntsen and Thomsen, 2005). In short, surprise is not required for the formation of a flashbulb memory, though in some cases it does seem to help.

Emotional Valence

Almost by definition, most flashbulb-memory events are emotionally intense. Emotional valence, however, is a different matter. Most flashbulb events are negative – assassinations, accidents, natural disasters, and the like. Yet a small body of evidence suggests that positive events can lead to flashbulb memories as well. Bohn and Berntsen (2007) examined flashbulb memories for the fall of the Berlin Wall. This event was certainly momentous, but it was interpreted in different ways by different people: some saw it as a great moment for freedom, whereas others were worried about what the future might hold. They found that both groups of people had flashbulb memories for the fall, but flashbulbs were more common in the group who viewed it as positive. Similarly, Berntsen and Thomsen (2005) found that many Danish participants had flashbulb memories for the liberation of Denmark in World War II, and that people who were part of the resistance movement were more likely to have flashbulbs. Thus, both positive and negative events can lead to the development of flashbulb memories.

Consequentiality

As noted above, Brown and Kulik (1977) defined consequentiality in a very specific way: a consequential event was one that had a notable impact upon a person's life. Traumatic events, such as an assault or the death of a loved one, are consequential almost by definition; these events had direct effects upon the individual and likely changed the course of his or her life. Flashbulb memories are not necessarily so consequential: the explosion of the space shuttle *Challenger*, while shocking and upsetting, did not change the lives of most of the people who were alive at the time (with the obvious exception of the relatives and acquaintances of the deceased, including the students of the schoolteacher who died in the explosion). Similarly, the deaths of political figures, whether surprising or not, do not necessarily have a significant effect on an individual's life (see Neisser, 1982; Berntsen, 2009, for discussion). This is a first strike against the consequentiality hypothesis: like surprise, it seems to be unnecessary for the formation of a flashbulb.

Does consequentiality have any effect at all? Evidence for its importance is mixed, in part because different researchers have defined the property in different ways. For instance, Neisser et al. (1996) examined memories for the 1989 earthquake in Loma Prieta, California. They compared three groups of participants: residents of Santa Cruz, California (for whom the effects were severe), residents of Berkeley, California (who felt the earthquake, but were at a greater distance from the epicenter), and residents of Atlanta, Georgia (who were on the other side of North America and did not feel the earthquake at all). They found that the residents of Santa Cruz had more detailed memories than the residents of Berkeley, who in turn had better memories than the residents of Atlanta. Moreover, Atlanta residents who had some connection to California (relatives, friends, and the like) had more detailed memories than Atlanta residents who did not.

By itself, this evidence seems consistent with the consequentiality hypothesis: people whose lives were influenced – whether by experiencing the earthquake or knowing someone who might have been affected – had better memories than those who did not. Yet Neisser et al. (1996) disagree. In part, their disagreement seems to be with the whole idea of consequentiality: they observe that "[al]though the fact that an assassination or resignation [or, presumably, an earthquake] has occurred may have real consequences for citizens of a given country … the occasion on which they heard about it surely does not" (Neisser et al., 1996: p. 353). While this is true enough, the objection seems to miss the point of Brown and Kulik's (1977) original claim. They argued that if the event itself was consequential, it would trigger the retention of the reception circumstances; the circumstances themselves did not have to have any significant consequences at all. Thus, while Neisser et al. (1996) suggested that consequentiality was unimportant, it is not clear that this argument speaks to the specific claim that Brown and Kulik (1977) were making. More convincing is their finding that their participants did not find the earthquake to be maximally frightening or worrisome, and that people were more worried by catastrophes happening elsewhere (such as the collapse of the Bay Bridge).

Other studies have suggested that consequentiality has little or no effect on flashbulb-memory formation. Er (2003) conducted a study of flashbulb memories for the Marmara earthquake in Turkey. The results showed that the consequentiality of the earthquake (defined as how much it changed the participants' lives) had an indirect and inconsistent association with the formation of a flashbulb memory. Similarly, Berntsen and Thomsen (2005) found that consequentiality had no effect on the accuracy of responses to questions about World War II.

Significance

An event can be personally significant or personally relevant even if it is not consequential. The death of a prominent politician might have no particular objective consequences for an individual's life, but it could still be personally relevant if it affects one's national or cultural identity (see Berntsen, 2009, for discussion). Several studies have shown that significance has a key role in the development of flashbulb memories. Conway et al. (1994) studied memories for the resignation of UK Prime Minister Margaret Thatcher and found that residents of the United Kingdom were more likely to have flashbulb memories than nonresidents (see Pezdek, 2003, for a similar study of the September 11th attacks). Similarly, even if Neisser et al.'s (1996) study shows nothing about consequentiality, it still suggests that personal relevance has an important role to play. Though there have been a few exceptions (Davidson and Glisky, 2002), the majority of studies have confirmed the fundamental – and fundamentally unsurprising – finding that people are more likely to remember events that they find important.

Narrative and Rehearsal

Several authors, including Brown and Kulik (1977) have noted that flashbulb memories are often remembered as narratives or coherent stories, and that these narratives are regularly rehearsed. Neisser et al. (1996) argued that there were several reasons for rehearsal, including social demands and a desire for self-promotion. They noted that participants reported that they had frequently thought or talked about their memories (Pennebaker and Harber, 1993). Several other studies have found a link between rehearsal and flashbulb-memory formation (Bohannon, 1988), though as Neisser et al. (1996) observed, it is difficult to know whether participants can accurately report the frequency of rehearsal.

Hirst and Meksin (2009) expanded on this view by observing that the news media provide a substantial impetus for rehearsal. In fact, in many ways, the media virtually compel rehearsal: after a flashbulb event, one can hardly watch the news or walk into a store without encountering a reminder of what has just happened. They presented data showing that the maintenance of a flashbulb memory was correlated with the frequency of news coverage of the event: declines in media coverage for both the *Challenger* disaster and the September 11th attacks were associated with declines in performance.

Sociocultural Factors

Why do we rehearse and relate these narratives? Several authors have observed that flashbulb memories seem to serve a social purpose: people seem to want to tell these stories, and a number of articles have been devoted to chronicling the ways in which famous people learned the news (Berendt, 1973). Brown and Kulik (1977) mentioned this phenomenon in their paper, suggesting that their readers were "primed with an account of [their] own, which [they] would rather like to tell" (Brown and Kulik, 1977).

Neisser (1982) has suggested that flashbulbs are significant because they help us place ourselves in history:

> The flashbulb recalls an occasion when two narratives that we ordinarily keep separate – the course of history and the course of our own lives – were momentarily put into alignment … . We remember the details of a flashbulb occasion because these details are the links between our own histories and 'History.' (Neisser, 1982: pp. 72–73)

In his terms, flashbulbs serve as 'benchmarks' that signal our place in history, and they are compelling for that reason.

Berntsen (2009) suggests that social identity plays a fundamental role in the development of flashbulb memories. According to her view, flashbulb memories serve as a way to connect us with a particular social group. Thus, her flashbulb memory of the death of the Danish king plays a role in her identity as a Dane, and my flashbulb memory of hearing about the *Challenger* disaster plays a role in my identity as an American. This identification occurs even though neither event had any significant effect on our everyday lives. By contrast, the 2003 Iranian earthquake had no connection to either of our social identities, and neither of us has a flashbulb memory for it. (Danes or Americans of Iranian descent might, of course, have a different experience.)

Furthermore, these memories allow us to connect with multiple social groups (or subsets of social groups). Americans old enough to have Pearl Harbor memories do not just identify as Americans; they can also identify as members of the World War II generation, whereas younger Americans cannot. To put this another way, people who do not share your flashbulb memories are excluded from one of the social groups to which you belong, which might account for the (admittedly anecdotal) observation that people find it jarring to encounter people who are too young to have a flashbulb memory of a particular event.

As Berntsen (2009) notes, this idea could help explain several of the results described above. Several researchers found that physical proximity or personal significance were associated with stronger flashbulb memories (Conway et al., 1994; Neisser et al., 1996; Pezdek, 2003; Yarmey and Bull, 1978). Yet social identity might be the more important factor. If so, then expatriate British people should have flashbulb memories for the death of Princess Diana (provided that they are still identified as British). Moreover, the strength of the flashbulb memory should in some way depend on the strength and number of social ties to the affected group: an American who also identified as a New Yorker and a Manhattanite should have stronger flashbulb memories than Americans who did not, even if the Manhattanite no longer lived there and did not know anyone affected by the attacks. In Berntsen's (2009) view, incorporating social identity into the flashbulb-memory model could help resolve the disparate results of previous studies. Specifically, she suggests that flashbulb memories could be formed through a number of routes – emotional reaction and/or subsequent rehearsal – and social identity serves as the underlying impetus for each of these.

Conclusions

What can we conclude from all of these studies? We know that people often say that they can remember the circumstances in which they learned of a momentous event, such as the September 11th attacks. These memories are vivid, and they are also enduring – people claim to be able to recall them decades later (Colegrove, 1900; Berntsen and Thomsen, 2005). People tend to be very confident in their memories; nevertheless, they are not necessarily accurate or consistent (though sometimes they are). The consequentiality of an event does not seem to matter very much, but the personal significance (usually) does. They can cover both positive and negative events. They are best seen as extreme instances of normal memory processes, not as the products of a special mechanism.

These conclusions might seem inconsistent or frustrating, but they reflect the significant difficulties involved in working with real-world events. There is no way to obtain control over the events themselves, and even two similar events can differ in many significant ways. As an example, the explosion of the space shuttle *Challenger* seems to be a prototypical flashbulb event (Bohannon, 1988; Bohannon and Symons, 1992; Neisser and Harsch, 1992; Terr et al., 1996; see Figure 1). The explosion of the space shuttle *Columbia* is not – or at least only a few published studies seem to have addressed it (Julian et al., 2009; Kershaw et al., 2009). Why is this so? Perhaps the second explosion was less novel or surprising given that one shuttle had already exploded. Perhaps people felt more of an emotional connection to the first explosion because a schoolteacher was on board. Perhaps it was overshadowed (for researchers as well as participants) by the September 11th attacks. Perhaps there were social-identity difference as well – perhaps the space program was no longer as important to people because the Cold War 'space race' was effectively over. Through further examinations of all of these factors, we may be able to gain more insights into this ever-intriguing phenomenon.

See also: Autobiographical Memory; Childhood Memories; Episodic Memory; Eyewitness Testimony; False Memories; Reconstructive Memory, Psychology of; Visual Memory, Psychology of.

Bibliography

Berendt, J., 1973. Where Were You? Esquire.

Berntsen, D., 2009. Flashbulb memories and social identity. In: Luminet, O., Curci, A. (Eds.), Flashbulb Memories: New Issues and New Perspectives. Taylor and Francis, New York, pp. 187–205.

Berntsen, D., Thomsen, D.K., 2005. Personal memories for remote historical events: accuracy and clarity of flashbulb memories related to World War II. Journal of Experimental Psychology: General 134, 242–257.

Bohannon, J.N., 1988. Flashbulb memories for the space shuttle disaster: a tale of two theories. Cognition 29, 179–186.

Bohannon, J.N., Symons, V.L., 1992. Flashbulb memories: confidence, consistency, and quantity. In: Neisser, U., Winograd, E. (Eds.), Affect and Accuracy in Recall. Cambridge University Press, New York, pp. 65–91.

Bohn, A., Berntsen, D., 2007. Pleasantness bias in flashbulb memories: positive and negative flashbulb memories of the fall of the Berlin Wall among East and West Germans. Memory and Cognition 35, 565–577.

Brewer, W.F., 1988. Memory for randomly sampled autobiographical events. In: Neisser, U., Winograd, E. (Eds.), Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory. Cambridge University Press, New York.

Brown, R.W., Kulik, J., 1977. Flashbulb memories. Cognition 5, 73–99.

Christianson, S.Å., 1989. Flashbulb memories: special, but not so special. Cognition 17, 435–443.

Cohen, N.J., McCloskey, M., Wible, C.G., 1988. There is still no case for a flashbulb-memory mechanism: reply to Schmidt and Bohannon. Journal of Experimental Psychology: General 117, 336–338.

Cohen, N.J., McCloskey, M., Wible, C.G., 1990. Flashbulb memories and underlying cognitive mechanisms: reply to Pillemer. Journal of Experimental Psychology: General 119, 97–100.

Colegrove, F.W., 1900. Memory: An Inductive Study. Henry Holt and Company, New York. Reprinted by Nabu Public Domain Reprints (no date).

Conway, M.A., Anderson, S.J., Larsen, S.F., Donnelly, C.M., McDaniel, M.A., McClelland, A.G.R., Rawles, R.E., Logie, R.H., 1994. The formation of flashbulb memories. Memory and Cognition 22, 326–343.

Crombag, H.F.M., Wagenaar, W.A., van Koppen, P.J., 1999. Crashing memories and the problem of 'source monitoring'. Applied Cognitive Psychology 10, 95–104.

Curci, A., 2009. Measurement issues in the study of flashbulb memory. In: Luminet, O., Curci, A. (Eds.), Flashbulb Memories: New Issues and New Perspectives. Taylor and Francis, New York, pp. 13–32.

Curci, A., Luminet, O., Finkenauer, C., Gisle, L., 2001. Flashbulb memories in social groups: a comparative study of the memory of French President Mitterand's death in a French and Belgian group. Memory 9, 81–101.

Davidson, P.S.R., Glisky, E.L., 2002. Is flashbulb memory a special instance of source memory? Evidence from older adults. Memory 10, 99–111.

Er, N., 2003. A new flashbulb memory model applied to the Marmara earthquake. Applied Cognitive Psychology 17, 503–517.

Finkenauer, C., Luminet, O., Gisle, L., El-Ahmadi, A., van der Linden, M., Philippot, P., 1998. Flashbulb memories and the underlying mechanisms of their formation: toward an emotional-integrative model. Memory and Cognition 26, 516–531.

Greenberg, D.L., 2004. President Bush's false 'flashbulb' memory of 9/11/01. Applied Cognitive Psychology 18, 363–370.

Hirst, W., Meksin, R., 2009. A social–interactional approach to the retention of collective memories of flashbulb events. In: Luminet, O., Curci, A. (Eds.), Flashbulb Memories: New Issues and New Perspectives. Taylor and Francis, New York, pp. 207–225.

Julian, M., Bohannon, J.N., Aue, W., 2009. Measures of flashbulb memory: are elaborate memories consistently accurate? In: Luminet, O., Curci, A. (Eds.), Flashbulb Memories: New Issues and New Perspectives. Taylor and Francis, New York, pp. 99–122.

Kershaw, T.C., Hemmerich, J.A., Ahmed, S., 2009. Flashbulb memory for September 11 and the Columbia space shuttle disaster. In: Kelley, M.R. (Ed.), Applied Memory. Nova Science Publishers, pp. 129–146.

Linton, M., 1975. Memory for real-world events. In: Norman, D.A., Rumelhart, D.E. (Eds.), Explorations in Cognition. Freeman, San Francisco, pp. 376–404.

Livingston, R.B., 1967. Reinforcement. In: Quinton, G.C., Melnechuck, T., Schmitt, E.O. (Eds.), The Neurosciences: A Study Program. Rockefeller University Press, New York, pp. 568–576.

Luminet, O., 2009. Models for the formation of flashbulb memories. In: Luminet, O., Curci, A. (Eds.), Flashbulb Memories: New Issues and New Perspectives. Taylor and Francis, New York, pp. 51–76.

McCloskey, M., Wible, C.G., Cohen, N.J., 1988. Is there a special flashbulb-memory mechanism? Journal of Experimental Psychology: General 117, 171–181.

Neisser, U., 1982. Snapshots or benchmarks? In: Neisser, U., Hyman, I.E. (Eds.), Memory Observed: Remembering in Natural Contexts. Worth Publishers, San Francisco, pp. 68–74.

Neisser, U., 1986. Remembering Pearl Harbor: reply to Thompson and Cowan. Cognition 23, 285–286.

Neisser, U., Harsch, N., 1992. Phantom flashbulbs: false recollections of hearing the news about *Challenger*. In: Winograd, E., Neisser, U. (Eds.), Affect and Accuracy in Recall. Cambridge University Press, New York, pp. 9–31.

Neisser, U., Winograd, E., Bergman, E.T., Schreiber, C.A., Palmer, S.E., Weldon, M.S., 1996. Remembering the earthquake: direct experience vs. hearing the news. Memory 4, 337–357.

Otani, H., Kusumi, T., Kato, K., Matsuda, K., Kern, R.P., Widner, R., Ohta, N., 2005. Remembering a nuclear accident in Japan: did it trigger flashbulb memories? Memory 13, 6–20.

Pennebaker, J.W., Harber, K.D., 1993. A social stage model of collective coping: The Loma Prieta Earthquake and the Persian Gulf War. Journal of Social Issues 49, 125–145.

Pew Research Center for the People and the Press, September 5, 2002. One year later: New Yorkers more troubled, Washingtonians more on edge. Pew Research Center.

Pezdek, K., 2003. Event memory and autobiographical memory for the events of September 11, 2001. Applied Cognitive Psychology 17, 1033–1045.

Pillemer, D.B., 1984. Flashbulb memories of the assassination attempt on President Reagan. Cognition 16, 63–80.

Pillemer, D.B., 1990. Clarifying flashbulb memory concept: Comment on McCloskey, Wible, and Cohen (1988). Journal of Experimental Psychology: General 119, 92–96.

Rubin, D.C., Kozin, M., 1984. Vivid memories. Cognition 16, 81–95.

Schmidt, S.R., Bohannon, J.N., 1988. In defense of the flashbulb memory hypothesis: a comment on McCloskey, Wible, and Cohen (1988). Journal of Experimental Psychology: General 117, 332–335.

Schmolck, H., Buffalo, E.A., Squire, L.R., 2000. Memory distortions develop over time: recollections of the O.J. Simpson trial verdict after 15 and 32 months. Psychological Science 11, 39–45.

Talarico, J.M., Rubin, D.C., 2003. Confidence, not consistency, characterizes flashbulb memories. Psychological Science 14, 455–461.

Talarico, J.M., Rubin, D.C., 2007. Flashbulb memories are special after all; in phenomenology, not accuracy. Applied Cognitive Psychology 21, 557–578.

Talarico, J.M., Rubin, D.C., 2009. Flashbulb memories result from ordinary memory processes and extraordinary event characteristics. In: Luminet, O., Curci, A. (Eds.), Flashbulb Memories: New Issues and New Perspectives. Taylor and Francis, New York, pp. 79–98.

Terr, L.C., Bloch, D.A., Michel, B.A., Shi, H., Reinhardt, J.A., Metayer, S., 1996. Children's memories in the wake of *Challenger*. American Journal of Psychiatry 153, 618–625.

Thompson, C.P., Cowan, T., 1986. Flashbulb memories: a nicer interpretation of a Neisser recollection. Cognition 22, 199–200.

Tinti, C., Schmidt, S., Sotgiu, I., Testa, S., Curci, A., 2009. The role of importance/consequentiality appraisal in flashbulb memory formation: the death of Pope John Paul II. Applied Cognitive Psychology 23, 236–253.

Weaver, C.A., 1993. Do you need a "flash" to form a flashbulb memory? Journal of Experimental Psychology: General 122, 39–46.

Winograd, E., Killinger, W.A., 1983. Relating age at encoding in early childhood to adult recall: development of flashbulb memories. Journal of Experimental Psychology: General 112, 413–422.

Wright, D.B., 2009. Flashbulb memory methods. In: Luminet, O., Curci, A. (Eds.), Flashbulb Memories: New Issues and New Perspectives. Taylor and Francis, New York, pp. 33–50.

Yarmey, A.D., Bull, M.P., 1978. Where were you when President Kennedy was assassinated? Bulletin of the Psychonomic Society 11, 133–135.

Flow in Motivational Psychology

Jeanne Nakamura and Matthew Dubin, Claremont Graduate University, Claremont, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes the phenomenon of flow, the subjective experience of being completely absorbed in a task or activity. We provide an overview of the flow model of optimal experience and affirm the flow state's close association with intrinsic and emergent motivation. We also describe the various ways that flow has been measured, discuss the antecedents and consequences of flow within the domains of work and leisure, and identify emerging areas of flow research.

Introduction

Among theories of motivation, the flow model (Csikszentmihalyi, 1975/2000, 1990) is distinguished by its focus on subjective experience. In the 1960s, Csikszentmihalyi set out to study intrinsically motivated, or autotelic ('auto' = self, 'telos' = goal), activity – activity that is valued for its own sake separate from any results or rewards deriving from it. While there had been research on intrinsic motivation, no systematic empirical work had been undertaken to elucidate the *phenomenology* of intrinsically motivated activity or to situate it within the broader experiential landscape of daily life. Callois's and Huizinga's theories of play and Csikszentmihalyi's own observational research on artistic creativity and study of Continental phenomenology informed his 1975 work, which was written against the then-prevailing deterministic theories of motivation ("Beyond Boredom and Anxiety" played off of Skinner's influential title). This article describes the resulting flow model and explicates the current state of knowledge about flow.

In the initial interviews conducted in the United States by Csikszentmihalyi and colleagues, people pursuing very different potentially autotelic activities used very similar terms to describe what they experienced when things were going well. The subjective state was dubbed 'flow' because some individuals used that term in describing the fluid, focused engagement they experienced (see example, **Table 1**). The flow state's hallmark is intense involvement in which attention is complete and the person feels he or she is functioning at full capacity. In subsequent research, the experience has been described in much the same way across cultures regardless of an individual's gender, age, or social class (Delle Fave et al., 2011).

Flow research has been pursued since the 1970s, at first primarily in the laboratories of Csikszentmihalyi in the United States and Massimini in Italy (e.g., Csikszentmihalyi and Csikszentmihalyi, 1988; Inghilleri, 1999; Massimini and Carli, in Csikszentmihalyi and Csikszentmihalyi, 1988; Massimini and Delle Fave, 2000). Since then, researchers studying leisure (play, art, sport, games), intrinsic motivation, and interest have taken up the flow concept. Cultivating the flow experience has also drawn considerable interest from researchers and practitioners in schools and workplaces – that is, contexts where encouraging engagement is vitally important because of links to quality of life, learning, and development, as well as performance. Additionally, the flow concept has had an impact on professional sports, popular culture, and government, and it has been studied within positive psychology as a factor in well-being.

Table 1 The essential components of the flow state

Key dimensions of the flow state (Csikszentmihalyi, 1975/2000, 1990)	*Original descriptions from rock climbers (Csikszentmihalyi, 1975/2000)*
Total concentration on what one is doing; past, future, and stimuli outside of the immediate interaction recede	"It's a centering thing, being absolutely in the here and now, in the present" (p. 81)
A merging of action and awareness	"You're so involved in what you're doing [that] you aren't thinking about yourself as separate from the immediate activity. You're no longer a participant-observer, only a participant" (p. 86)
A sense of control; one feels that one will be able to respond to whatever occurs	"Once you're into the situation … you're very much in charge of it" (p. 81)
Loss of self-consciousness; one does not step outside of the stream of experience, watching and judging the self	"When you first start climbing you're very aware of your capabilities. But after a while you just do it without reflecting on it at the time" (p. 87)
Distorted sense of time; time may seem to dilate (slow down or stand still) or one may lose all track of time and later feel it passed very quickly	An "eternal moment" (p. 87)
'Autotelic' experience; the experience itself is intrinsically rewarding	"The purpose [of climbing is] not looking for a peak or utopia but staying in the flow …. you move up only to keep the flow going. There is no possible reason for climbing except the climbing itself" (pp. 47–48)

The Flow Model

The initial understanding of the nature and conditions of flow (Csikszentmihalyi, 1975/2000) was based on interviews with rock climbers, chess players, surgeons, and others pursuing activities thought to "maximize immediate, intrinsic rewards to the participant" (p. 21). In flow, the experience seamlessly unfolds from moment to moment and one enters a subjective state with the characteristics shown in **Table 1**. The central role of attention should be clear (for discussions of the dynamics of attention in flow, see Csikszentmihalyi and Csikszentmihalyi, 1988; Csikszentmihalyi and Nakamura, in Bruya, 2010).

In addition to these dimensions of the individual's subjective state, the initial research identified three proximal conditions conducive to entering flow. First, the flow state is more likely to occur when the person perceives *clear goals* along the way as the activity proceeds. Second, experiencing flow depends on perceiving *clear, immediate feedback* – unambiguous information about the progress being made and whether the course of action should be adjusted. Third, there needs to be a dynamic *balance between perceived opportunities for action (challenges) and perceived capacities for action (skills)*.

That is, a person operates at the limits of perceived capacities when in flow (cf deCharms, 1968; Deci, 1975; White, 1959). The state is one of dynamic equilibrium. Entering flow depends on establishing this balance (cf Berlyne's optimal arousal); staying in flow requires maintaining it. Importantly, the relation between perceived challenges and skills was employed in 1975 to sketch the broader experiential landscape associated with any potentially autotelic activity. Csikszentmihalyi (1975/2000) differentiated three conditions of experience in an activity: a flow channel along which perceived challenges and skills match at any level; a region of boredom, as opportunities for action relative to skills drop off; and a region of anxiety, as challenges increasingly exceed capacities for action.

Massimini and colleagues made important contributions in the 1980s to clarifying the role of challenges and skills in promoting flow. They observed that if operating at the limits of one's capacities is a condition for entering flow, then interactions perceived as offering few opportunities for action will not lead to experiencing flow even when one thinks that one has correspondingly little relevant skill. Employing the Experience Sampling Method (ESM) that had been developed by Csikszentmihalyi and colleagues to study daily life (see Section Measuring Flow), the Milan group thus began describing the flow condition as the balance of perceived challenges and skills specifically when both are above an individual's average levels (Massimini and Carli, in Csikszentmihalyi and Csikszentmihalyi, 1988). That is, flow is expected to occur when individuals perceive greater opportunities for action than they encounter on average, and they feel they have adequate skills to engage them. The research revealed a fourth subjective state, apathy, associated with low challenges and correspondingly low skills. Experientially, this is a state of stagnation and diffuse attention – the opposite of the flow state. In addition, the quality of experience when perceived skills exceed perceived challenges proved to be more positive than originally theorized.

The resolution of this phenomenological map can be made finer by differentiating the challenge/skill terrain into eight experiential channels rather than four quadrants (see **Figure 1**). Doing so reveals a gradient from flow to control to relaxation and then to boredom, as skills exceed challenges. Similarly it is possible to show empirically a gradient from flow to arousal and then to anxiety, as challenges exceed skills. Finally, the quality of experience (whether positive or negative) intensifies within a channel or quadrant as perceived challenges and skills move away from a person's average levels. Conceptual and methodological issues related to the role of perceived challenges and skills in flow continue to be active areas of theory and research. A key underlying distinction is between these conditions for entering flow and the subjective state itself (for summaries, see Delle Fave et al., 2011; Engeser, 2012).

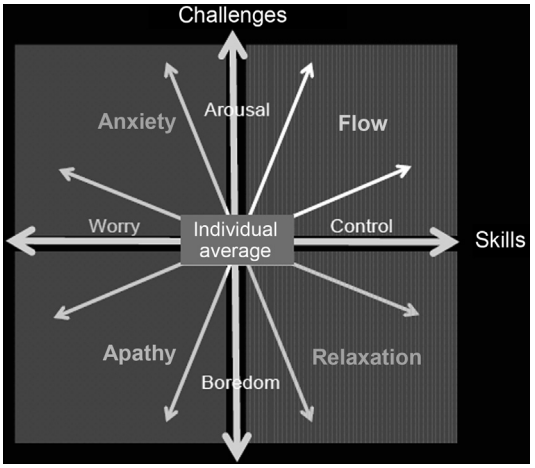

Figure 1 Flow within the experiential landscape defined by a person's subjective challenges and skills, at two levels of resolution: the quadrant model (flow, relaxation, apathy, anxiety) and the 8-channel model (arousal, flow, control, etc.). Source: Massimini and Carli, 1988. In: Csikszentmihalyi, M., Csikszentmihalyi, I. (Eds.), Optimal Experience. Cambridge University Press, Cambridge.

Emergent Motivation, Learning, and Development

It should be evident that the view of human functioning associated with the flow model (see Inghilleri, 1999; Rathunde and Csikszentmihalyi, 2006) is consistent with active-organism, interactionist perspectives in psychology. It focuses on the dynamic system comprised of person and context, and the specific subjective states that are experienced in person–environment interactions. In two respects, the flow model describes motivation as arising out of these person–environment interactions; that is, as 'emergent motivation.'

First, in flow, what happens at any given moment is responsive to what happened the moment before within the interaction (Csikszentmihalyi, 1985), rather than being predetermined by characteristics of either the person (e.g., a drive) or the environment (e.g., a social role). Shifts in subjective state provide feedback about the person's changing relationship to the environment. In the unfolding interaction, a person may respond to experiencing anxiety or boredom by seeking to adjust their level of skill and/or challenge, in order to escape the aversive state and reenter flow. In this first sense, proximal goals may be emergent.

Second, flow experiences may lead to expansion or transformation of a person's goals and interests (Csikszentmihalyi and

Nakamura, 1999). That is, an individual may enter an activity by chance or in response to extrinsic demands rather than by choice, neither seeking nor embracing the opportunities for action it might hold. However, it is possible to come to experience the new or previously unengaging activity as intrinsically rewarding if the proximal conditions of the interaction lead the person to find flow in it. The motivation to persist in or return to the activity arises out of the absorbing experience itself. In this second sense, long-term goals or interests may be emergent.

The flow state is intrinsically rewarding; it encourages people to seek to repeat experiences in which they have previously found flow. However, the flow model suggests that as people master the challenges of an activity, they develop greater levels of skill and the activity ceases to be as involving as before. To keep experiencing flow, one must engage in increasingly complex challenges. This factor distinguishes the flow model from motivational theories that define optimal challenge in terms of either a homeostatic equilibrium point to be returned to, or a maximum level of challenge to be reached (Moneta and Csikszentmihalyi, 1996). The optimal level of challenge stretches existing skills, resulting in more complex capacities for action. Both in the moment and in the direction of a life, this introduces into psychological functioning a selective mechanism (Massimini and Delle Fave, 2000) that fosters development – the acquisition of skill within an existing activity and the identification of new paths to growth.

In the context of individual lives, a personal 'flow activity' is an activity of this kind, into which an individual selectively invests attention and energy because he or she has tended to find flow in it before (cf 'individual interests'). It provides the individual with opportunities for flow and skill growth, thereby contributing to the quality of life. Of course, one person's cherished flow activity may hold no appeal for another individual. In every society, however, pursuits have evolved that seem to comprise culturally provided affordances for optimal experience because they embody the conditions of flow so well. A sport or game is a paradigmatic example, presenting clearly defined goals, rules, feedback, and a set of graduated challenges. It helps focus the participant's attention by circumscribing the 'stimulus field' (Csikszentmihalyi, 1975/2000). This is a cultural flow activity. The individual–culture relationship mediated by flow is held to be reciprocal. When people selectively participate in the culture's flow activities, they are more likely to experience flow and the activities are more likely to persist in the culture (for fuller discussions, see Csikszentmihalyi and Massimini, 1985; Delle Fave et al., 2011; Inghilleri, 1999).

The Autotelic Personality

Selectively investing attention and energy in a specific flow activity is one way to experience flow more frequently. Another is through a general proneness to flow. While flow seems to be a virtually universal experience, there are marked differences in the frequency and intensity of individuals' experience of flow. A potential explanation for this relates to Csikszentmihalyi's (1975/2000) concept of the 'autotelic personality,' which suggests that some individuals are more prone to experiencing and staying in flow than others. Csikszentmihalyi defined the autotelic personality as someone who pursues activities for their own sake instead of an external goal. Autotelics have a low level of self-centeredness and neuroticism. The inherent sense of curiosity and interest that they have in the world leads them to seek opportunities where they can build their skills by testing them against sizable challenges, and their endurance and conscientiousness lead them to practice and persist (Baumann, 2012; Ullén et al., 2012). They are able to better manage the balance between challenges and skills that marks the flow experience. Those with autotelic personalities should possess 'metaskills' that allow them to not only enter and experience flow, but also maintain it for an extended period of time (Nakamura and Csikszentmihalyi, 2009).

Measuring Flow

Researchers have developed self-report and indirect methods of studying flow (see Nakamura and Csikszentmihalyi, 2009; Delle Fave et al., 2011). In-depth interviews have provided qualitative descriptions of the characteristics and conditions of flow, beginning with Csikszentmihalyi's (1975) study. Individual differences in the frequency of flow experiences and population incidence of different flow activities have been measured with survey items and the Flow Questionnaire. A number of validated scales measure either disposition to experience flow or situational (state) flow, either in one specific activity or in life more generally. The goal of studying the occurrence of situational flow across people's daily lives stimulated development of the ESM, which uses *in situ* reports of subjective experience at randomly sampled moments to anchor flow experiences in their real-life contexts. Currently, ESM researchers still operationalize both the flow state and flow conditions in multiple ways. In addition to these self-report methods, researchers more recently have developed experimental paradigms that induce flow in the laboratory through controlled manipulation of its theorized proximal conditions (e.g., the balance of challenges and skills provided by video games). There is growing interest in measuring flow's physiological correlates in and beyond the laboratory (see below). Some researchers have sought visible behavioral markers of the flow state. All these latter methods promise indirect rather than self-report measures of flow, holding out the possibility of studying flow experiences 'online' without disrupting their unfolding.

Measuring the autotelic personality has mainly focused on gathering ESM data and assessing how much time individuals spend in high-challenge, high-skill situations. Another strategy has been to measure how intrinsically motivated individuals are during high-challenge, high-skill activities. Non-ESM measures ask individuals how often they experience flow in general or tap underlying traits (see Baumann, 2012; Nakamura and Csikszentmihalyi, 2009; Ullén et al., 2012).

Flow in Work and Leisure

Cross-culturally, the two prominent contexts in which adults report experiencing flow are work and leisure. Leisure is mentioned more often by Western respondents and work by non-Western ones (Delle Fave et al., 2011).

Leisure

Leisure activities (i.e., how individuals spend free time outside of work) are often intrinsically motivating and, therefore, should produce flow if other conditions of the experience are met (Csikszentmihalyi and LeFevre, 1989). Leisure time can be devoted to a myriad of different activities, such as hobbies (e.g., gardening), sports, and video games. Based on ESM research, Csikszentmihalyi and others separate leisure activities into two categories: Active leisure and passive leisure. Passive leisure activities, such as watching television, do not require substantial energy and are generally relaxing (Csikszentmihalyi, 1990). Despite occupying large amounts of free time in some cultures, watching television and similar activities do not usually lead to flow and can decrease positive affect (Csikszentmihalyi and LeFevre, 1989; Kubey and Csikszentmihalyi, 2002).

Active leisure pursuits may be structured and require substantial concentration toward achieving a certain goal (see Delle Fave et al., 2011). They include serious leisure activities that allow individuals to develop long-term skills or knowledge (e.g., reading a book). As previously mentioned, the active leisure activities of sports and games often exemplify flow's proximal conditions, containing clear goals, unambiguous feedback, and high challenges. Flow is more likely to occur during active leisure activities when the perceived ability of individuals is high and, in turn, encourages them to pursue high-challenge situations (Perkins and Nakamura, 2013). High-skill players during video games experience greater levels of flow when engaged in high-challenge situations, and chess players who are more intrinsically motivated enjoy competing against more challenging opponents than players who are focused on winning (Abuhamdeh and Csikszentmihalyi, 2009).

Work

While the original conceptualization of flow mainly focused on leisure activities, the integral flow condition of balanced challenge and skill has been found to occur more often in the workplace (Csikszentmihalyi, 1990; Csikszentmihalyi and LeFevre, 1989). An ESM sample of managers, engineers, clerical workers, and blue-collar workers spent 54% of their time at work in high-challenge, high-skill situations compared to only 17% during leisure activities (Csikszentmihalyi and LeFevre, 1989). The frequent occurrence of flow in the workplace has been called a paradox; workers report feeling focused and creative at work but say that they would rather be engaged in leisure activities (Csikszentmihalyi, 1990). This may reflect cultural attitudes toward work versus leisure or the stress and strain that accompanies flow at work (Nakamura and Csikszentmihalyi, 2009). In either case, it is a testament to the multidimensionality of the flow experience and the challenges researchers have in measuring it: While workers may often encounter high-skill, high-challenge situations at work, this alone may not capture whether they are experiencing the flow state. A more in-depth examination of flow at work illuminates why the experience often occurs in this setting. Flow at work is defined as short-term peak experiences where an individual is completely engaged, absorbed, and enjoys the task at hand. Many jobs have built-in characteristics that would be conducive to experiencing flow, specifically clear goals, immediate feedback, and high-challenge tasks (Csikszentmihalyi, 1990; Fullagar and Kelloway, 2013).

What Facilitates Flow at Work?

Skill variety and autonomy are two job characteristics that predict flow (see Fullagar and Kelloway, 2013). Research also shows both personal resources (e.g., self-efficacy) and organizational resources (e.g., social support, innovation, and clear goals) increase flow's occurrence in the workplace (Salanova et al., 2006; Fullagar and Kelloway, 2013). Flow at work is associated with subjective well-being (see Fullagar and Kelloway, 2013), and the relationship between flow and organizational performance has been found to be partially mediated by positive mood (Eisenberger et al., 2005).

To summarize the research investigating flow in the context of work and leisure, individuals are more likely to encounter high-challenge, high-skill situations at work than during leisure time, which may indicate that individuals are not spending their leisure time in a flow-conducive manner. Work is often stressful and exhausting for individuals, which may increase the appeal of pursuing passive leisure activities. More research needs to be done to discover why individuals are not more inclined to pursue active leisure activities in which they find flow. Furthermore, it is essential that researchers understand the full scope of flow's occurrence in the workplace since high-challenge, high-skill situations may not be a sufficient condition for flow in this context due to the stresses and extrinsic rewards that work holds.

The Potential for OverFlow

Although flow has mainly been associated with positive outcomes, it is important to acknowledge its potentially negative effects (see Schüler, 2012). Several of flow's characteristics, specifically deep concentration, time distortion, and a sense of control, can impede individuals from living a balanced life, due to single-minded devotion to one activity.

Flow activities are not necessarily health-promoting. The intrinsic rewards that flow provides can cause individuals to become addicted to certain flow activities. Flow has been associated with exercise addiction, online game addiction, and Internet addiction. Indeed, flow is amoral. Thus, another potential dark side of the flow experience is that individuals could potentially enter flow through harmful – for example, aggressive or violent – behavior. For example, soldiers have reported having optimal experiences while killing others in the heat of combat.

Flow also can lead to risky behaviors, as some individuals constantly seek greater challenges as their skills improve to maintain the challenge–skill equilibrium that allows them to experience flow. This can result in risky behaviors (e.g., climbing a treacherous mountain with minimal support) when novice flow-seekers misread performance feedback and do not recognize their vulnerabilities.

Future Directions

Multiple emerging areas of flow research have the potential to substantially expand the overall understanding of the flow experience. This section focuses on several of the most burgeoning, rich areas of new flow research.

Social Flow

As individual flow experience becomes better understood, a new topic of flow research has begun to emerge: flow in social contexts, including dyads, groups, and teams. Walker (2010) has identified three types of flow experiences: Individual solitary flow (e.g., playing the guitar), coactive social flow (e.g., going on a jog with friends), and interactive social flow (e.g., playing on a basketball team). At least for college students, flow in both social contexts (coactive social flow and interactive social flow) was more enjoyable than individual flow. However, many questions remain, such as whether social flow is a distinct construct from individual flow, and how best to specifically assess social flow.

Along these lines, the presence of flow in the context of interpersonal relationships needs further examination. Pursuing close connections with others is one of the core motivations of human existence, and close relationships are associated with happiness and life satisfaction. Although close relationships would seemingly result in optimal experiences, little is known about the specific role that flow plays in romantic and other relationships.

The Psychophysiology of Flow

Another emerging topic of research is the psychophysiology of the flow experience. A pioneer in this area, Jean Hamilton, conducted a positron emission tomography study in 1984. Consistent with the view that concentration is effortless in flow (Csikszentmihalyi, 1975/2000, 1990), she found that the more autotelic individuals took less effort to regulate their attention on a laboratory task.

Two decades elapsed before theoretical work on the neuroscience of flow and effortless attention resumed (for a summary, see Delle Fave et al., 2011), with Dietrich and Weber producing conflicting theorizations regarding the role of the prefrontal cortex (the brain region associated with higher cognition). While Dietrich suggested that there is a transient decrease of neural activity in the prefrontal cortex during flow (allowing the fluent execution of highly practiced skill), Weber and colleagues objected that prefrontal activity has been shown to increase in focused attention and that flow has a different neural basis: synchronization of the brain's attentional and reward networks.

Empirical research on the brain basis of flow and its correlates is germinal (for a recent summary, see Peifer, 2012). Both Keller and colleagues and Peifer have reported an inverted-U relationship between flow and markers of effort (lowered heart rate variability, cortisol). Also of note, Ullén and colleagues (in Bruya, 2010) hypothesized that effortless attention is associated with the *interaction* of high attention and positive affect and found that the physiological correlates of both (sympathetic autonomic nervous system activation, deep breathing, and activation of the 'smile' muscle) characterized pianists in flow. Firm claims about the psychophysiological correlates of flow have yet to be established but this is an active research front warranting further investigation.

Applications of the Flow Concept

As flow research continues to flourish, determining the most effective applications of the theory should be a priority. Flow interventions can either design activities and environments so they are conducive to experiencing flow, or they can directly help individuals enter a flow state (Nakamura and Csikszentmihalyi, 2009). Several flow interventions have been implemented with considerable success, including Green Cargo (a Swedish state-owned transportation company) implementing practices based on flow principles to help the company become profitable for the first time in 115 years. Flow principles also informed exhibit design at Los Angeles' Getty Museum and product design at Nissan USA (Nakamura and Csikszentmihalyi, 2009).

The flow model has also been applied in the educational context. Research shows that learning is often a flow experience but schooling too often is not. The practices of effective schools and programs have been studied from a flow perspective and programs that are directly based on the flow model have been implemented in some private and public educational settings. For example, the Key School, an Indianapolis public school, developed a Flow Activities Center, which gives students regular opportunities to identify personal flow activities by choosing what activities they would like to pursue. For reviews of educational applications, as well as the underlying research on the flow–performance relationship and on differences in student flow experience by school type, subject matter, and pedagogy, see Delle Fave et al. (2011) and Shernoff and Csikszentmihalyi (2011). More work is needed to understand the antecedents and consequences of flow in the classroom in order to design empirically based interventions that benefit both teachers and students.

Conclusion

As demonstrated by the growing body of research investigating the flow experience, flow is a theory with broad relevance. Flow plays a significant role in personal, organizational, and social contexts, and current research continues to expand flow's range of application. In recent years, the flow experience has been situated as a key theory within the field of positive psychology, which goes beyond human deficits to focus on how individuals, organizations, and communities flourish and thrive. In addition to its contribution to motivational psychology, the concept of flow has played an integral role in positive psychology's mission of understanding how individuals can lead happier, more fulfilling lives.

See also: Affect-Regulation Motivation; Avoidance and Approach Motivation: A Brief History; Burnout, Psychology of; Eudaemonism; Gender and Academic Motivation; Grit; Interest, Psychology of; Mastery Learning; Motivation, Learning, and Instruction; Motivation: Life Course and Sociological Perspectives; Passion and Motivation; Personal Projects; School Burnout and Engagement: Lessons from a Longitudinal Study in Finland; Self-Determination Theory; Self-Regulated Learning: Theories, Measures, and Outcomes; Self-Regulation in Adulthood; Sociodigital Revolution: Digital Natives vs Digital Immigrants; Test Anxiety and Academic Achievement; Vocational Interests, Values, and Preferences, Psychology of.

Bibliography

Abuhamdeh, S., Csikszentmihalyi, M., 2009. Intrinsic and extrinsic motivational orientations in the competitive context: an examination of person–situation interactions. Journal of Personality 77 (5), 1615–1635.

Baumann, N., 2012. Autotelic personality. In: Engeser, S. (Ed.), Advances in Flow Research. Springer, New York, pp. 165–186.

Bruya, B. (Ed.), 2010. Effortless Attention: A New Perspective in the Cognitive Science of Attention and Action. MIT Press, Cambridge, MA.

Csikszentmihalyi, M., 1975. Beyond Boredom and Anxiety. Jossey-Bass, San Francisco, CA.

Csikszentmihalyi, M., 1985. Emergent motivation and the evolution of the self. Advances in Motivation and Achievement 4, 93–119.

Csikszentmihalyi, M., 1990. Flow: The Psychology of Optimal Experience. HarperCollins, New York.

Csikszentmihalyi, M., Csikszentmihalyi, I. (Eds.), 1988. Optimal Experience. Cambridge University Press, Cambridge.

Csikszentmihalyi, M., LeFevre, J., 1989. Optimal experience in work and leisure. Journal of Personality and Social Psychology 56, 815–822.

Csikszentmihalyi, M., Massimini, F., 1985. On the psychological selection of biocultural information. New Ideas in Psychology 3, 115–138.

Csikszentmihalyi, M., Nakamura, J., 1999. Emerging goals and the self-regulation of behavior. Advances in Social Cognition 12, 107–118.

deCharms, R., 1968. Personal Causation. Academic Press, New York.

Deci, E., 1975. Intrinsic Motivation. Plenum, New York.

Eisenberger, R., Jones, J.R., Stinglhamber, F., Shanock, L., Randall, A.T., 2005. Flow experiences at work: for high need achievers alone? Journal of Organizational Behavior 26, 755–775.

Engeser, S. (Ed.), 2012. Advances in Flow Research. Springer, New York.

Fave, A.D., Massimini, F., Bassi, M., 2011. Psychological Selection and Optimal Experience across Cultures: Social Empowerment through Personal Growth. Springer, Dordrecht, Netherlands.

Fullagar, C., Kelloway, E.K., 2013. Work-related flow. In: Bakker, A., Daniels, K. (Eds.), A Day in the Life of a Happy Worker. Psychology Press, London, pp. 41–57.

Inghilleri, P., 1999. From Subjective Experience to Cultural Change. Cambridge University Press, Cambridge.

Kubey, R., Csikszentmihalyi, M., 2002. Television addiction is no mere metaphor. Scientific American 286, 74–80.

Massimini, F., Delle Fave, A., 2000. Individual development in a bio-cultural perspective. American Psychologist 55, 24–33.

Moneta, G., Csikszentmihalyi, M., 1996. The effect of perceived challenges and skills on the quality of subjective experience. Journal of Personality 64 (2), 275–310.

Nakamura, J., Csikszentmihalyi, M., 2009. Flow theory and research. In: Lopez, S.J., Snyder, C.R. (Eds.), Oxford Handbook of Positive Psychology, second ed. Oxford University Press, New York, pp. 195–206.

Peifer, C., 2012. Psychophysiological correlates of flow-experience. In: Engeser, S. (Ed.), Advances in Flow Research. Springer, New York, pp. 1–22.

Perkins, K., Nakamura, J., 2013. Flow and leisure. In: Freire, T. (Ed.), Positive Leisure Science. Springer, Dordrecht, Netherlands, pp. 141–157.

Rathunde, K., Csikszentmihalyi, M., 2006. The development of the person: an experiential perspective. In: Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology, vol. 1. Wiley, Hoboken, NJ, pp. 465–515.

Salanova, M., Bakker, A.B., Llorens, S., 2006. Flow at work: evidence for an upward spiral of personal and organizational resources. Journal of Happiness Studies 7, 1–22.

Schüler, J., 2012. The dark side of the moon. In: Engeser, S. (Ed.), Advances in Flow Research. Springer, New York, pp. 123–137.

Shernoff, D.J., Csikszentmihalyi, M., 2011. Flow in schools. In: Gilman, R., Huebner, E.S., Furlong, M. (Eds.), Handbook of Positive Psychology in Schools. Routledge, New York, pp. 131–145.

Ullén, F., de Manzano, O., Almeida, R., Magnusson, P., Nakamura, J., Csikszentmihalyi, M., Madison, G., 2012. Proneness for psychological flow in everyday life: associations with personality and intelligence. Personality and Individual Differences 52, 167–172.

Walker, C.J., 2010. Experiencing flow: is doing it together better than doing it alone? The Journal of Positive Psychology 5 (1), 3–11.

White, R., 1959. Motivation reconsidered: the concept of competence. Psychological Review 66, 297–333.

Fluency and Voice, Disorders of

David B Rosenfield, Neurological Institute, Houston Methodist Hospital, TX, USA and Weill Medical College, Cornell University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Stuttering is a global, pancultural disturbance that has afflicted human beings throughout time. An increasing corpus of research has identified the role of genetics, physiology, and aberrant neuromotor control in the genesis of this disorder. Further, evolving animal models of stuttering (e.g., birdsong) and aspects of 'acquired stuttering' lend information to understanding stuttering. Hopefully, the expansion of these multiple research venues will identify the causation and possible cure for this disorder.

Stuttering is a global, pancultural disturbance of speech that has been with us throughout time. Individuals who have stuttered since childhood are sometimes referred to as developmental stutterers (DS) and persons who stutter (PWS). These individuals represent over 1% of the adult population (Porfert and Rosenfield, 1978) and endure considerable emotional pain and social stigma, as well as compromise in the marketplace and workforce. Heretofore fluent adult individuals rendered dysfluent by neurogenic compromise are referred to as acquired stutterers (AS) and have speech-motor characteristics different from DS.

There are multiple ancient references to stuttering, including ancient Mesopotamian clay tablets, Egyptian hieroglyphics ('nit-nit'), the Old Testament (Moses stuttered), and the Holy Koran. Recent scientific investigations of dysfluent speech-motor output substantiate organic factors in the etiology of stuttering and highlight stuttering as a window into understanding normal as well as abnormal production of speech-motor control.

Definition

Bloodstein and Bernstein-Ratner (2007) review the multiple definitions of stuttering. The definition of stuttering significantly alters clinical as well as research paradigms. What is a 'stutter'? If a child bounces on a sound, or someone repeats a whole word, is that a stuttered dysfluency? Are they equivalent in etiology and in genesis? Should investigators consider dysfluency in adult stutterers to be isomorphic to effortless dysfluency (e.g., 'normal fluency of childhood') seen frequently among children? Answers to these questions significantly impact our inquiries into the etiology of dysfluency.

Children who stutter (DS) have difficulty producing intended words and sounds. They know what they want to say but are unable to do so. Thus, their language is normal but their speech is not. Further, the location of stuttered dysfluencies in DS is not random. These are often located at the beginning of sentences and phrases and are associated with part-word repetitions, as opposed to repetition of whole words or phrases. Sometimes, the child will back up in producing the sentence and repeat whole phrases or portions of words in an effort to overcome the dysfluent output. Regardless, the core problem of dysfluency remains repetition of a single sound or a fragment of a word (Bloodstein and Bernstein-Ratner, 2007).

In an effort to achieve fluent production and overcome these dysfluencies, the stuttering child often employs circumlocutions, substituting one word for the other. Thus, a child who stutters, might say "g-g-g … see you later," instead of continuing to stutter on the /g/ of 'good-bye.' Acoustic analysis suggests when that child attempts to say "g-g-good-bye," the initial /g/ sound in the stuttered dysfluency acoustically differs from his final, fluently produced target word. This is in addition to what are identified as 'secondary characteristics' of stuttering, referring to facial distortions associated with the speaker's effort to force out the appropriate sound. Whatever stuttering behavior one produces, that behavior reflects the underlying difficulty producing the sound, with a superimposed maneuver to achieve fluent speech (Bloodstein, 2008; Rosenfield, 2001).

It is not a simple task to identify the locus of dysfluency. If a stutterer says "s-s-sound," where is the actual dysfluency? Many investigators previously maintained that the dysfluency was on the /s/, but now most would contend the deficit is on the *transition* from one sound to the next. The stutterer is able to say the /s/ but not the /*ound*/. The stutterer's strategy to achieve fluent output can result in repeating the /s/ until the transition into the ensuing /*ound*/ is secured (Rosenfield, 1997, 2001).

Clinical Features of Stuttering

As previously noted, stuttered dysfluencies are not randomly located within a sentence. Instead, they occur where fluent speakers' occasional dysfluencies also occur – at the beginning of sentences and phrases. Thus, one seldom hears a stutterer say, "Go to the building-g-g-g." Rather, a stutterer says "G-g-go to the building" (Bloodstein and Bernstein-Ratner, 2007; **Figure 1**).

Emotional stress can make stuttering worse, but is not the cause of stuttering. A 10-year-old stutterer may be nervous regarding presenting a book report in front of his class due to fear of embarrassment due to stuttering, yet is fluent when singing before the class despite emotional angst regarding singing in front of others. The fact that emotional stress

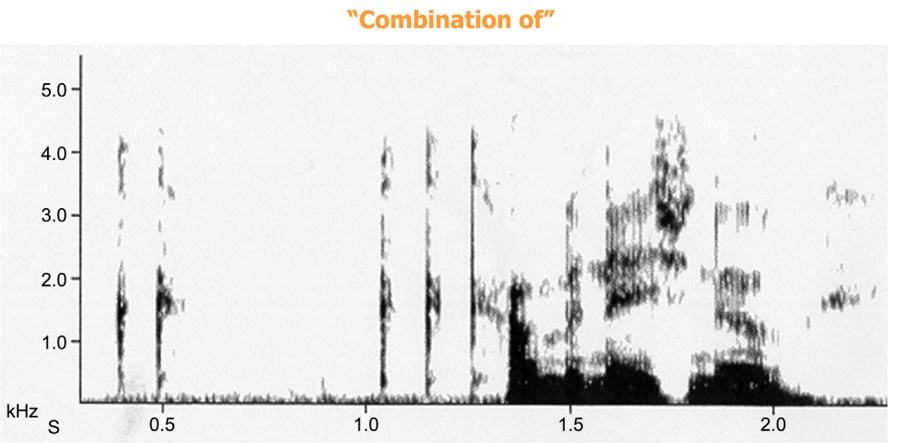

Figure 1 This is a spectrogram of a stutterer saying, 'domestic.' He repeats the /d/ sound several times before fluently achieving the target word, 'domestic' (spectrogram of stuttered speech).

exacerbates stuttering is consistent with stuttering being a disturbance of neuromotor control. Motor systems are affect-sensitive systems and worsen with stress.

Stutterers seldom repeat whole words. Their dysfluent output usually contains word fragments. The sounds (fragments, part-words) differ from the target sounds the stutterers are trying fluently to produce (Bloodstein and Bernstein-Ratner, 2007; Rosenfield, 1997, 2001).

Several speech-related maneuvers increase fluent output in stutterers, the most potent of which is singing. Stutterers' speech also considerably improves with choral reading, speaking in cadence with a metronome, or when loud, broadband noise prohibits their hearing their own speech. Delayed auditory feedback improves stuttering, as does speaking during inhalation, or repeating the same passage (adaptation effect) (Bloodstein, 2008).

Genetics

Stuttering is not randomly dispersed between males and females. Stuttering is significantly more prevalent among males than females. Eighty percent of children 'outgrow' their stutter but 20% do not. This confounding variable necessitates the importance of identifying current stutterers as well as past stutterers in any investigation of genetic causation of stuttering.

Genetic investigations among twins who stutter often focus upon concordance rates among monozygotic versus dizygotic twin pairs, as well as analyses of pedigrees in different families. Many of these investigations are hampered by small sample size, difficulty ascertaining twin status, and difficulty assessing who has what type of speech disruption. Howie (1981)directly examined twin pairs of stutterers, noting that 63% of 16 monozygotic twin pairs and 19% of 13 dizygotic twin pairs were concordant for stuttering. Andrews et al. (1991) investigated 3800 twins through self-employed reporting and ascertained 20% concordance among 50 monozygotic twin pairs, and 3% concordance among 85 dizygotic twin pairs. These investigations suggest heredity is an important factor in the genesis of stuttering, but may not be a sufficient condition, since the concordance among homozygous (identical, same genetic code) twins was not 100%. Therefore, environmental factors may also be a factor in the causation of stuttering.

Other investigators analyzed aggregates of stuttering within families, rather than primarily focusing upon twin pairs. Their findings suggest relatives of PWS have an increased risk of stuttering, ranging from a 3- to 10-fold increase. Segregation patterns of stuttering in these families do not suggest a single-gene, fully penetrable transmission model of inheritance, such as autosomal dominant, autosomal recessive, or X-linked. Rather, these studies suggest a diallelic model, with penetrance affected by whether a parent stutters and the sex of that parent who stutterers (higher among male-stuttering parents) (Kang et al., 2010; Viswanath et al., 2004; Suresh et al., 2006).

There is no singular genetic marker for people who stutter. Genetic predisposition may be a necessary condition for people to stutter, but this predisposition may not be sufficient. Similarly, investigations employing multiple avenues of brain imaging fail to identify a singular brain-related abnormality in people who stutter, despite several investigators identifying multiple cerebral (and cerebral-related) differences between those who stutter and their fluent counterparts.

Brain Imaging

There are many obfuscating factors in studying brain imaging in people who stutter, What is a stutter and how does one ascertain the severity of a stuttered dysfluency? For instance, is "b-b-b-ook" more severely dysfluent when repeating the 'b' 10 times, since both represent difficulty in achieving production of the correct sound? The repetition of the 'b' can merely represent accommodating to the 'error' in the speech-motor control system and not be isomorphic with the error itself. Is struggling with the sound once and then changing the word to a more fluently produced word, such as 'text,' a similar deficit? How should investigators incorporate previously directly or indirectly learned therapeutic strategies that ameliorate the stuttered output? Should people who stutter be evaluated in

groups and, if so, how should the groups be composed? Should individual stutterers, as opposed to groups of stutterers, be studied in the same paradigm in different trials and on different days? How does one incorporate strong handedness with less strong handedness, since handedness is a marker of language cerebral laterality and language laterality has been queried as a causal factor in stuttering? When one has stuttered for years, how does the investigator know which results relate to the stuttering itself or to the accommodation over years to the stuttering deficit, or even to the therapy itself (Rosenfield, 2001; Bloodstein and Bernstein-Ratner, 2007; Rosenfield, 2012)? Most investigators are cognizant of these constraints, as their ongoing research employs varying analyses of different brain imaging paradigms.

Sommer et al. (2002) employed diffusion tensor imaging (DTI) of the brain, a technique which assesses white matter neural fibers. These researchers observed that stutterers had fewer white matter fibers in the left brain, immediately below the representation of the tongue and laryngeal areas, and queried disconnection between the representation of the oropharynx (Brodmann's area 43) within sensorimotor cortex. Other investigators found DTI white matter abnormalities in the superior longitudinal fasciculus and corpus callosum (Chang et al., 2011; Cykowski et al., 2010). Still others, using different brain imaging paradigms, observed increased activations in the anterior insula, premotor cortex, and cerebellum, with decreased activation in superior temporal cortex (Brown et al., 2005; Ingham et al., 2012).

Other studies assess cerebral volumetric differences between stutterers and nonstutterers, as opposed to brain imaging assessment of cerebral physiological function or structural deficits. Foundas et al. (2001) observed that stutterers had qualitative differences in frontal lobe gyral patterns involving Broca's area and quantitative differences in the posterior temporal lobe regions, involving Wernicke's area. Given that most right-handed fluent speakers have a larger left superior posterior temporal area (Brodmann's area 22, planum temporale), primarily consisting of Wernicke's area, their most salient finding was that Wernicke's area on the left and its corresponding area on the right were bilaterally larger but less asymmetric in stutterers than in normal speakers.

Foundas et al. (2001) hypothesized that these brain abnormalities permit normal development of language but later contribute to abnormalities in the motor output of language, primarily speech. They query whether perisylvian (including Broca's and Wernicke's areas) speech–language cortex compromise contributes to instability in interaction between an outer 'linguistic' loop and an inner 'phonatory' loop, resulting in stuttering (discussed below in greater detail; Nudelman et al., 1992; Anderson et al., 1999). They note that abnormalities in the anatomic substrates of language can be different among stutterers, some having difficulty processing of speech-motor control due to aberrancies in frontal opercular areas, others having compromise near Wernicke's Area.

Other investigators employ PET and fMRI experiments in stuttering. PET experiments, including reading aloud, reading silently, and choral reading, suggest different activation patterns in stutterers versus nonstutterers (Fox et al., 1996; Braun et al., 1997; DeNil et al., 2000). Despite differences in experimental design, these studies highlight that right-handed stutterers lateralize activations to the right hemisphere in contrast to fluent speakers, who lateralize activations to the left hemisphere. Fox et al. (1996) observed that stuttering during solo reading induced widespread activation of motor systems in both the cerebrum and the cerebellum, with a right dominance.

Consistent with this finding, Braun et al. (1997) noted in stutterers that regional responses were absent, bilateral, or lateralized to the right hemisphere when the PWS engaged in speech production during stuttering episodes. DeNil et al. (2000) support this conclusion by observing a proportionately greater right hemisphere activation in stutterers reading individual words overtly or covertly in contrast to proportionately greater activation in the left hemisphere in nonstutterer controls. Fox et al. (1996) and Braun et al. (1997) conclude similar results pertaining to fluency evoked by chorus reading and evoked by prolongation of speech sounds. These studies suggest that brain activation patterns of stutterers differ from that of nonstutterers, not only during stuttered speech but also when PWS produce speech fluently.

Some scientists attempt to combine the excellent spatial resolution capability of fMRI with the excellent temporal resolution of magnetoencephalography (MEG), a technique permitting investigation of timing of cortical activation sequences in various tasks. MEG provides superb temporal resolution with good accuracy of localization of active cortical areas, and is useful for characterizing activation sequence from visual perception to oral output and for identifying cortical correlates of disorders of these processes. Salmelin et al. (2000) studied MEG paradigms in PWS and fluent controls. He had both groups read isolated nouns aloud in a delayed reading paradigm. The stutterers were fairly fluent in this task. Despite similar overt performance between the two groups, the cortical evoked patterns revealed clear differences, both in the evoked responses, time-locked to word presentation and mouth movement onset, and in task-related suppression of 20-Hz oscillations. Within the first 400 ms after seeing the word, physiologic processing in fluent speakers advanced from the left inferior frontal cortex (articulatory programming) to the left lateral central sulcus and dorsal premotor cortex (motor preparation). This sequence was reversed in stutterers; they had early left motor cortex activation followed by delayed left inferior frontal signals. Stutterers seem to initiate motor programming prior to preparation of the articulatory code. Suppression of motor cortical 20-Hz rhythms, caused by neuronal-related processing of tasks, occurred bilaterally in stutterers and fluent speakers, but the suppression was dominant in the right hemisphere in the stutterers whereas it was dominant in the left hemisphere in fluent speakers. Accordingly, the right frontal cortex was very active during speech in DS but did not generate synchronous time-locked responses. The authors postulate a dysfunctional network of the left inferior frontal cortex and right motor/premotor cortex network in stutterers.

Theories of Stuttering

Regardless of genetic or brain imaging input pertaining to why people stutter, or what definition of stuttering one chooses,

stutterers have difficulty controlling their laryngeal sound source and concurrent articulation. Investigators have demonstrated adductor as well as the abductor laryngeal muscles co-contract during dysfluencies, making it impossible to produce normal speech output. When stutterers try to 'fake' these dysfluencies, or if fluent speakers try to produce these, they do not produce these co-contractions. Electromyographic relationships between laryngeal abductor and adductor muscles are abnormal during moments of the stutterer's dysfluency (reviewed in Bloodstein and Bernstein-Ratner, 2007). Further, studies demonstrate that stutterers with laryngeal cancer who subsequently underwent laryngectomy and spoke with an electronic voicer, all became totally fluent with the electronic voicing apparatus (Rosenfield and Freeman, 1983).

Some theories contend aberrant cerebral laterality of language causes stuttering. Within this context, these focus on handedness and correlated cerebral laterality as a cause for stuttering, since right-handedness reflects left cerebral hemisphere laterality for language, and left-handedness reflects mixed or right cerebral hemisphere laterality for language. Investigations exploring cerebral hemisphere laterality for language evaluate handedness, dichotic listening (e.g., competing binaural linguistic inputs to measure laterality), and visual stimuli to competing hemispheres, all making this theory enticing.

Investigations of dichotic listening suggest PWS have disrupted auditory processing, compromising meaningful linguistic stimuli; many stutterers lack the normal left brain (i.e., right ear) advantage (Rosenfield and Goodglass, 1980; reviewed in Bloodstein and Bernstein-Ratner, 2007). However, being left- or right-handed has never been demonstrated to be a necessary or sufficient condition for stuttering (reviewed in Rosenfield, 1997; in Bloodstein and Bernstein-Ratner, 2007).

Some maintain that stuttering relates to abnormal auditory self-monitoring of speech. When loud, broadband noise prohibits a stutterer from hearing his speech, fluency markedly improves. This has also been observed with delayed auditory feedback, a setting in which a stutterer hears what he says 250 ms after he speaks. Frequency-altered feedback, in which the PWS hears what he/she just spoke but with an altered frequency, also promotes fluency (Kalinowski et al., 1993).

Other investigators contend stuttering reflects periodic irregularities in timing of muscle movements within the speech-motor control system (Zimmermann, 1980). When background muscle tension is elevated, as it can be in DS (Freeman and Ushijima, 1978), the high-precision adjustments needed to produce normal speech output are difficult to perform, and speech-motor movements are not smooth. Not only do stutterers have poor coordination of antagonist laryngeal muscles (Freeman and Ushijima, 1978) but also they are systematically slower in initiating phonation than are fluent speakers (Bloodstein and Bernstein-Ratner, 2007). Caruso (1991) adds that the supplementary motor area is dysfunctional in stuttering, resulting in poor motor planning of the speech output.

The above theories primarily note differences between groups of stutterers and groups of nonstutterers, positing mechanisms not common to all the PWS, since some of the PWS do not have the abnormalities noted. One theory, which has gained some popularity, describes the necessary and sufficient conditions for a dysfluency to occur. It is complex, involves mathematical modeling, and also incorporates control theory. Below, we present a basic discussion of control theory and its subsequent modeling of the necessary and sufficient conditions for stuttered dysfluencies to occur.

Employing control theory, researchers (Nudelman et al., 1989; Rosenfield et al., 1991; Anderson et al., 1999; Foundas et al., 2001) model the speech-motor control system as consisting of two nested loops, an inner phonatory loop producing sound and an outer linguistic loop that selects the sounds to be produced. Stuttering occurs when the timing between these 'functional loops' is disrupted. Evidence suggests that stutterers have temporal slowing within the outer loop. Some data (Anderson et al., 1999; Foundas et al., 2001) suggest that perisylvian speech–language cortex mediates the outer loop, whereas corticostriatal–thalamocortical circuits mediate the inner loop. Anderson et al. (1999) further corroborate this hypothesis by noting that dopamine blockade and Parkinson disease slow the inner cortical–basal ganglia loop, reducing the dysfluency in PWS.

Some investigators believe altered sensory (auditory) feedback in PWS causes unstable/dysfluent speech (reviewed in Civier et al., 2010). Others subscribe to the 'covert repair hypothesis' (Postma and Kolk, 1993), positing PWS make errors in phonological planning, erroneously formulated prior to the articulation of sound, resulting in the internal self-monitoring system stopping speech output. Others (e.g., Howell, 2004) suggest lack of readiness itself in the motor plan. Stromsta (1986) believes stuttering results from insufficient coarticulation – the PWS fails the preparatory planning of articulation, resulting in the articulators not being correctly positioned to proceed, causing halting of speech. Peters et al. (2000) propose the nexus of stuttering lies within the lower processing stages of the speech process, involving abnormal parameter settings for muscular force and muscular command execution.

Modeling of any disease is facilitated with an animal model of that disease. Our laboratory has developed an animal model of stuttering, hoping to delineate the neurophysiology and neuropharmacology underlying fluent and dysfluent production of birdsong (Helekar et al., 2000, 2003; Rosenfield et al., 2001; Helekar et al., 2013), possibly relating that to the aberrant speech in PWS.

Animal Modeling of Stuttering

Human beings have a unique system of language, whereas nonhuman animals have a system of communication. Animals can signal danger or food to another animal, but are not capable of producing rich language akin to ours, which incorporates semantics, phonology and, especially, a generative grammar. However, some nonhuman species do share with us mechanisms of sound production. For instance, zebra finches (*Taenopygia guttata*) have many features pertaining to communication and sound production similar to humans. Both zebra finches and human beings have critical periods during development during which they must hear the sounds of their respective communication/language they are to produce, and there are specific periods during which they must hear these sounds and practice them. For instance, a zebra

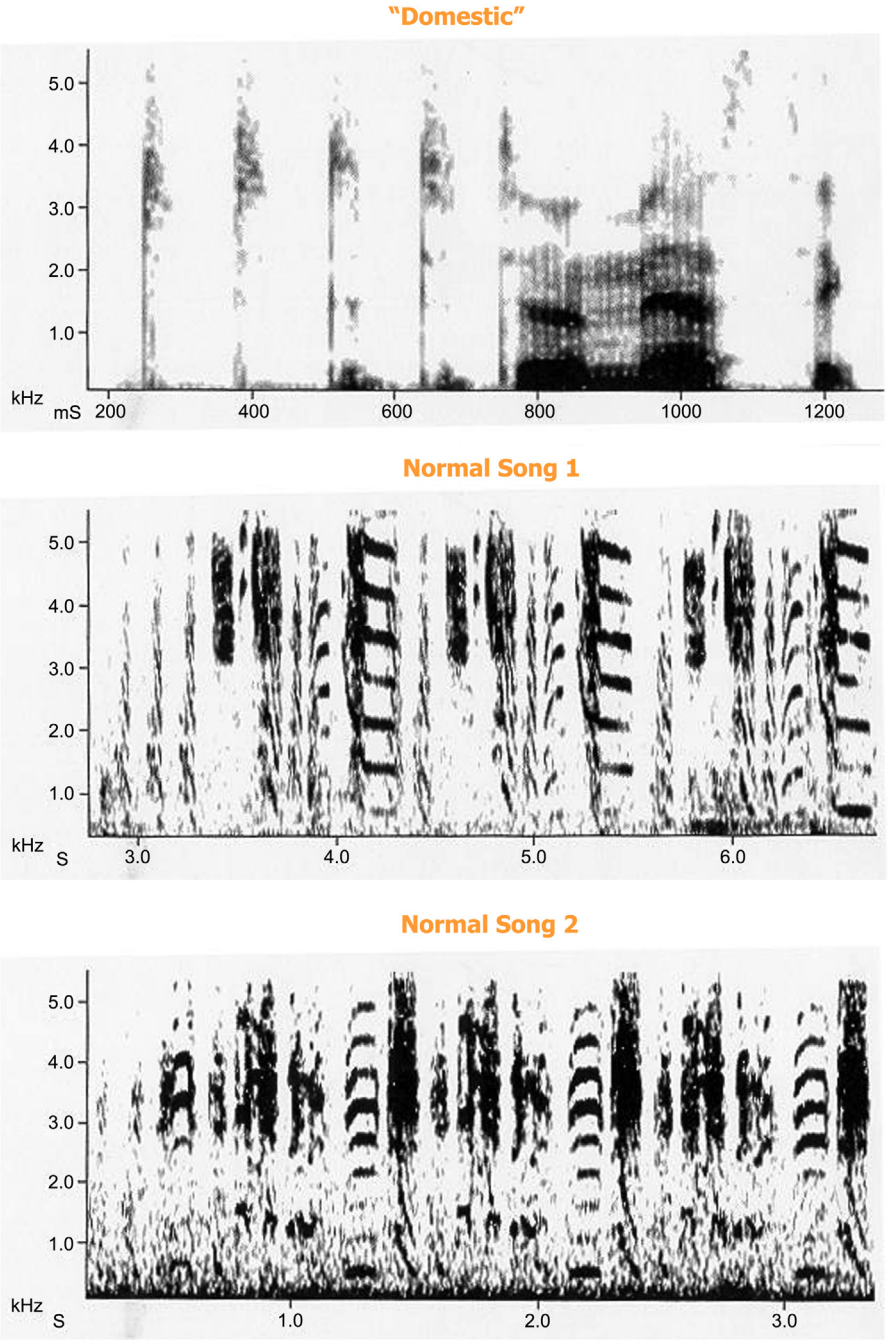

Figure 2 Spectograms of normal Zebra finch birdsong. The horizontal axis again reflects time; vertical axis reflects frequencies. Note that the Zebra finch birdsong is spectrally complex and changes over time, sharing these properties with humans (normal birdsong).

finch must hear songs from its adult tutors by day 65 and practice it by day 90. If disruption in hearing or social isolation prohibits this time course, the bird will not learn normal adult birdsong. Similarly, young children must hear and practice sounds they are to make if they are to acquire the language of their parents.

Zebra finches and humans have a spectrally and temporally diverse sound output (i.e., multiple frequencies and changes in frequencies over time). Both the finches and humans have brains that control these sound outputs and both have an established hierarchy (Rosenfield et al., 2001; Helekar et al., 2000, 2003, 2013).

Our laboratory has demonstrated that ~5% of Zebra finches born to normal parents and raised among normal birds producing normal birdsong have phonatory iterations similar to stuttering. When these birds raise other birds, born to normals, ~60% of these birds stutter. However, although the latter have multiple phonatory iterations, they do not repeat the spectral content of the birds from which they learned to stutter. In other words, they learn to make an error in timing,

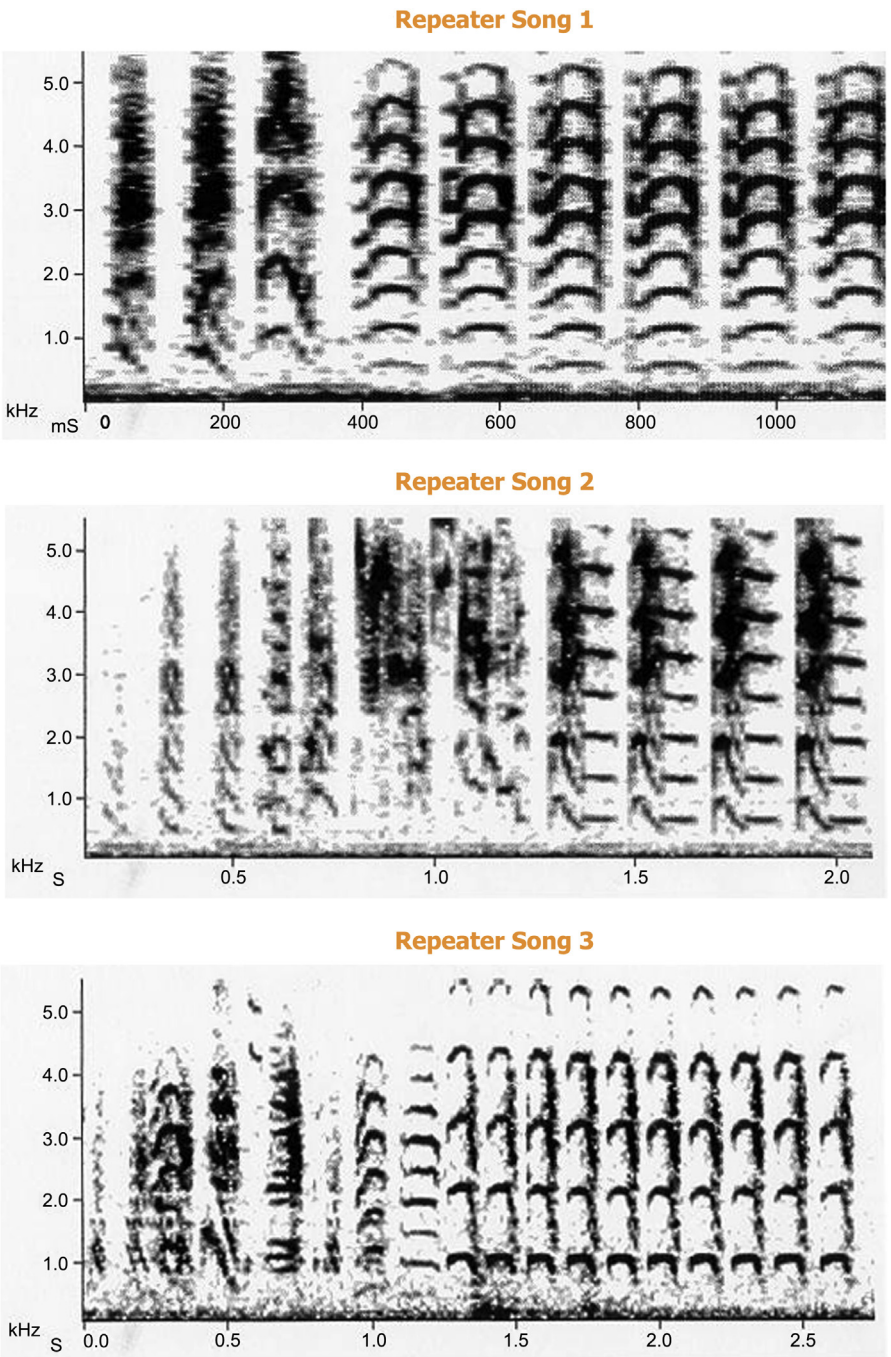

Figure 3 These spectograms describe phonatory iterations in a bird. Note the repetition of the 'syllable' at the end of the motif (stuttering in birds).

but not in the frequency domain. They have considerable improvement in these abnormal repetitions when they are placed in isolation or among normal singers (reviewed in Rosenfield et al., 2001; Helekar et al., 2013).

Thus, there appears to be a 'rule to repeat' in birdsong. Some birds (~5%) are born repeaters, perhaps reflecting a genetic mandate. Others birds learn to repeat abnormally and can improve toward normalcy with time, but still have the stuttering (reviewed in Rosenfield et al., 2001; Helekar et al., 2013).

Zebra finches provide a good animal model of stuttering. Similar to human stutterers, the birds' stuttering is position-specific, although it is at the end of the motif (e.g., major division of song production), whereas humans stutter at the beginning of sentences and phrases (see **Figures 2** and **3**).

The birds' and humans' errors are repetitive and nonvoluntary, as suggested by the regularity of the iterations (Helekar et al., 2013). One presumes the birds, similar to humans, prefer not to stutter, given that the tutored repeaters improve

Table 1 Acquired stuttering

	Developmental stuttering	*Acquired stuttering*	*Cluttering*
Locus of lesion	Unknown	Usually cortical, but can be subcortical	Unknown
Cause	Unknown	Vascular, metabolic, tumor	Unknown
Duration	80% of children outgrow: Adults have a worse prognosis	Unilateral: Good prognosis Bilateral: Poor prognosis	Varies
Locus of dysfluency	Beginning of sentence or phrase	Frequently scattered throughout sentence	Varies
Singing	Fluent	May improve somewhat, but not totally fluent	Varies
Onset	Subacute	Subacute or acute	Gradual
Reaction to dysfluencies	Anxious	Not anxious	Usually not concerned

significantly when they are placed among normal singing adults (Helekar et al., 2013). It is not known whether the birds that stutter are distraught by their abnormal iterations or whether their colleagues perceive them as different.

If one disregards the abstract linguistic and cognitive elements of speech and focuses, instead, on vocal motor components, songbird vocalization provides substantial insight into human speech-motor control disorders. Similarities between birdsong and speech can help understand the pathophysiology of human dysfluencies, and possibly in testing neuromotor theories of stuttering.

Acquired Stuttering

Acquired stuttering (AS) was initially reported by Rosenfield in 1972 and in ensuing years has been substantiated by many other investigators. As noted in that report, AS differs from DS in that the former is dysfluent throughout the sentence, stutters when singing, and is oftentimes not emotionally disturbed by their dysfluency. AS can be caused by cerebral injury, involving either hemisphere and multiple locations. The injury is usually mild. Prognosis is good if the compromise is unilateral but less so if brain damage is bilateral. In some AS patients, there may be psychogenic factors (Rosenfield and Barroso, 2000).

Whereas DS have difficulty achieving the target sound (e.g., ba-ba-ba-book), the AS achieves the target sound, repeating that sound (e.g., boo-boo-book). This difference is often clinically perceptible at the bedside.

Cluttering is sometimes confused with stuttering. Clutterers produce abnormal speech, characterized by excessive speed, repetition, interjections, drawling, disturbed prosody (i.e., altered rhythm of speech), and monotonous sound. Sometimes, these patients have inconsistent articulatory disturbances. Some investigators contend clutterers also make many grammatical errors, are hyperactive, have compromised concentration and poorly integrated thought processes.

Many clutterers omit sounds, syllables, and whole words, and invert the orders of sound, repeat the initial sound, and prolong several syllables of the word. Their rate of speech is usually very rapid, and the listener often complains about the rapidity of the speech. Opposed to DS, clutterers frequently are not concerned regarding their speech deficits (see **Table 1**; Rosenfield and Barroso, 2000; Bloodstein and Bernstein-Ratner, 2007).

Conclusion

Stuttering is a complex disturbance reflecting abnormal cerebral neuromotor processing of speech-related signals. There is probably a strong genetic component to stuttering, based at a minimum on studies of twin pairs and the marked prevalence of stuttering among males, but a genetic predisposition may not be sufficient to cause stuttering. Current investigations, employing clinical as well as genetic and brain imaging studies, coupled with studies in animals, may provide more insight into this pancultural, global disturbance.

Appendix A. Supplementary Data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/B978-0-08-097086-8.54042-2.

Acknowledgement

This work was sponsored by the M.R. Bauer Foundation.

Bibliography

Anderson, J.M., Hughes, J.D., Rothi, L.J., Crucian, G.P., Heilman, K.M., 1999. Developmental stuttering and Parkinson's disease: the effects of levodopa. Journal of Neurology, Neurosurgery, and Psychiatry 66, 776–778.

Andrews, G., Morris-Yates, A., Howie, P., Martin, N.G., 1991. Genetic factors in stuttering confirmed. Archives of General Psychiatry 48, 1034–1035.

Bloodstein, O., Bernstein-Ratner, N.A., 2007. Handbook on Stuttering, sixth ed. Delmare Cengage Learning, Clifton Park, New York.

Braun, A.R., Varge, M., Stager, S., et al., 1997. Alerted patterns of cerebral activity during speech and language production in developmental stuttering: a PET study. Brain 120, 762–784.

Brown, S., Ingham, R.J., Inhgam, J.C., Laird, A.R., Fox, P.T., 2005. Stuttered and fluent speech production: AN ALE meta-analysis of functional neuroimaging studies. Human Brain Mapping 25, 105–117.

Caruso, A.J., 1991. Neuromotor processes underlying stuttering. In: Peters, H.F.M., Halstijn, W., Starkweather, C.W. (Eds.), Speech Motor Control in Stuttering. Elsevier, Amsterdam, pp. 101–116.

Chang, S.E., Horwitz, B., Ostuni, J., Reynolds, R., Ludlow, C.L., 2011. Evidence of left inferior frontal-premotor structural and functional connectivity deficits in adults who stutter. Cerebral Cortex 21, 2507–2518.

Civier, O., Tasko, S.M., Guenther, F.H., 2010. Overreliance on auditory feedback may lead to sound/syllable repetitions: simulations of stuttering and fluency-inducing conditions with a neural model of speech production. Journal of Fluency Disorders 35, 246–279.

Cykowski, M., Fox, P.T., Ingham, R.J., Ingham, J.C., Robin, D.A., 2010. A study of the reproducibility and etiology of diffusion anisotropy differences in developmental stuttering: a potential role for impaired myelination. Neuroimage 52, 1495–1504.

DeNil, L.F., Kroll, R.M., Kapur, S., Houle, S., 2000. A positron emission tomography study of silent and oral reading of single words in stuttering and nonstuttering adults. Journal of Speech Language and Hearing Research 43, 1038–1053.

Foundas, A.L., Bollich, A.M., Corey, D.M., Hurley, M., Heilman, K.M., 2001. Anomalous anatomy of speech-language areas in adults with persistent developmental stuttering. Neurology 5, 207–215.

Fox, P.T., Ingham, R.J., Ingham, J.C., et al., 1996. A PET study of the neural systems of stuttering. Nature 382, 158–161.

Freeman, F.J., Ushijima, T., 1978. Laryngeal muscle activity during stuttering. Journal of Speech and Hearing Research 21, 538–562.

Helekar, S.A., Marsh, S., Viswanath, N.S., Rosenfield, D.B., 2000. Acoustic pattern variations in the female-directed birdsongs of a colony of laboratory-bred zebra finches. Behavioural Processes 49, 99–110.

Helekar, S.A., Espino, G.G., Botas, A., Rosenfield, D.B., 2003. Development and adult phase plasticity of syllable repetitions in the birdsong of captive zebra finches (Taeniopygia guttata). Behavioral Neuroscience 117, 939–951.

Helekar, S.A., Voss, H., Rosenfield, D.B., 2013. Stuttered birdsong. In: Helekar, S.A. (Ed.), Animal Models of Speech and Language Disorders. Springer Publications, Philadelphia, pp. 185–208 (Chapter 7).

Howell, P., 2004. Assessment of some contemporary theories of stuttering that apply to spontaneous speech. Contemporary Issues in Communicative Science Disorders 31, 123–140.

Howie, P.M., 1981. Concordance for stuttering in monozygotic and dizygotic twin pairs. Journal of Speech and Hearing Research 24, 317–321.

Ingham, R.J., Grafton, S.T., Bothe, A.K., Ingham, J.C., 2012. Brain activity in adults who stutter: similarities across speaking tasks and correlations with stuttering frequency and speaking rate. Brain and Language 122, 11–24.

Kalinowski, J., Armson, J., Mieszkowski, M., Stuart, A., Gracco, V., 1993. Effects of alterations in auditory feedback and speech rate on stuttering frequency. Language and Speech 36, 1–16.

Kang, C., Riazuddin, S., Mundorff, J., et al., 2010. Mutations in the lysosomal enzyme-targeting pathway and persistent stuttering. New England Journal of Medicine 362, 677–685.

Nudelman, H.B., Hoyt, B., Herbrich, K.E., Rosenfield, D.B., 1989. A neuroscience model of stuttering. Journal of Fluency Disorders 14, 399–427.

Nudelman, H.B., Herbrich, K.E., Hoyt, B.D., Rosenfield, D.B., 1992. Phonatory response times of stutterers and fluent speakers to frequency-modulated tones. Journal of the Acoustical Society of America 92, 1882–1888.

Peters, H.F., Hulstijn, W., van Lieshout, P.H., 2000. Recent developments in speech motor research into stuttering. Folila Phoniatrica et Logopedica 52, 103–119.

Porfert, A.R., Rosenfield, D.B., 1978. Prevalence of stuttering. Journal of Neurology, Neurosurgery, and Psychiatry 41, 954–956.

Postma, A., Kolk, H., 1993. The covert repair hypothesis: prearticulatory repair processes in normal and stuttered disfluencies. Journal of Speech and Hearing Research 36, 472–487.

Rosenfield, D.B., 1972. Stuttering and cerebral ischemia. New England Journal of Medicine 287, 991.

Rosenfield, D.B., 1997. Stuttering. In: Schachter, S., Devinsky, O. (Eds.), Behavioral Neurology and the Legacy of Norman Geschwind. Lippincott-Raven, Philadelphia, pp. 101–111.

Rosenfield, D.B., 2001. Do stutterers have different brains? (Editorial). Neurology 57, 171–172.

Rosenfield, D.B., Freeman, F.S., 1983. Stuttering onset after laryngectomy. Journal of Fluency Disorders 8, 265–268.

Rosenfield, D.B., 2012. What you see is not always what you get: a response to "Neural anomaly and reorganization in speakers who stutter," by Lu et al. Neurology 79, 625–632. Write Click, Neurology, September 7, 2012.

Rosenfield, D.B., Goodglass, H., 1980. Dichotic testing of cerebral dominance in stutterers. Brain and Language 11, 170–180.

Rosenfield, D.B., Viswanath, N.S., Callis-Landrum, L., DiDanato, R., Nudelman, H.B., 1991. Patients with acquired dysfluencies: what they tell us about developmental stuttering. In: Peters, H.F.M., Halstijn, W., Starkweather, C.W. (Eds.), Stuttering. Elsevier, pp. 277–284.

Rosenfield, D.B., Barroso, A.O., 2000. Difficulties with speech and swallowing. In: Bradley, W.G., Darroff, R.B., Fenichel, G.M., Marsden, C.D. (Eds.), Neurology in Clinical Practice: Principles of Diagnosis and Management, third ed. Butterworth-Heinemann, Boston, pp. 171–186. Chapter 13.

Rosenfield, D.B., Viswanath, N.S., Helekar, S.A., 2001. An animal model of stuttering. In: Bosshardt, H.G., Yaruss, J.S., Peters, H.F.M. (Eds.), Fluency Disorders: Theory, Research, Treatment and Self-Help, Proceedings of the Third World Congress on Fluency Disorders in Nyborg. Nijmegen University Press, Nijmegen, Denmark, pp. 119–122.

Salmelin, R., Schnitzler, A., Schmitz, F., Freund, H.J., 2000. Single work reading in developmental stutterers and fluent speakers. Brain 123, 1184–1202.

Sommer, M., Koch, M.A., Paulus, W., Weiller, C., Buchel, C., 2002. Disconnection of speech-relevant brain areas in persistent developmental stuttering. Lancet 360, 380–383.

Stromsta, C., 1986. Elements of Stuttering. Atsmonts Publishing, Oshtemo, MI.

Suresh, R., Roe, C., Pluzhnikov, A., Wittke-Thompson, J., C-Y Ng, M., Cook, E., Lundstrom, C., Garsten, M., Ezrati, R., Yairi, E., Ambrose, N., Cox, N., 2006. New complexities in the genetics of stuttering: significant sex-specific linkage signals. American Journal of Human Genetics 78, 554–563.

Viswanath, N., Lee, H., Chakraborty, R., 2004. Evidence for a major gene influence on persistent developmental stuttering. Human Biology 76, 401–412.

Zimmermann, G., 1980. Stuttering: a disorder of movement. Journal of Speech and Hearing Research 23, 122–126.

Focus Groups

Martha Ann Carey, Kells Consulting, Media, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Focus groups are popular in social science research as well as in marketing and other business areas. When participants feel comfortable and the group facilitator is well prepared, the conversation usually leads to rich stories that probably would not be told in such detail in another type of setting. This technique can be very engaging for participants, especially persons from communities who often do not have a voice in matters that affect them. Focus groups are best suited to environments and groups where the members are knowledgeable, willing, and capable of communicating; the topic and the group setting are compatible with group interaction; and the group facilitator has adequate skills.

This article presents the basics of this technique so as to assist in the understanding of the character of focus groups and their potentials. For additional information and examples of the needed documents (invitation letter, introduction to the session, guideline questions, checklist, format for reporting), please see Carey and Asbury (2012) and Krueger and Casey (2009).

Definition and Development

The definitions of a focus group vary but usually include a semi-structured session with multiple participants, an informal setting, moderation by a facilitator and often a co-facilitator, the use of general guideline questions and/or other data elicitation stimuli such as photos, and a means to record the information generated through group interactions (Carey and Asbury, 2012; Krueger and Casey, 2009). Basically, focus groups are a data collection technique that combines principles of qualitative research and knowledge of small group dynamics. This approach capitalizes on the interaction among the group members to enhance the collection of deep, strongly held beliefs and perspectives. This approach, as well as most qualitative methods, is especially useful to explore new topics and to examine complex issues involving values and beliefs that underlie behavior. Some sessions may provide helpful social support and education, but this is not the goal of the focus group. Focus group sessions generally consist of a one-time meeting of persons who have a common experience, such as a new diagnosis of cancer or experience as a caregiver for an elderly spouse.

The development of this approach is generally associated with the research programs of Paul Lazarsfeld, Robert Merton, and colleagues (Merton et al., 1990). Marketing studies have extensively used this approach, and politicians use focus group data to help shape their campaigns. Education, nursing, psychology, social work, and sociology are increasingly using focus groups in programs of research.

Similar to other qualitative techniques, focus groups are not useful when the researcher needs to assess the magnitude of a problem. Also, when rapport and trust cannot be established in the group setting, the data will be too comprised to be of value. Trust involves both the group members and the world outside of the session, and it can be affected by ongoing relationships among members of the group. Other issues may arise due to the sensitive nature of the topic, vulnerability of the population, and concerns with the confidentiality of the data. In some instances, a taboo subject may be better explored in an individual interview or via a computer interview than in a group setting. However, there is some evidence that groups may facilitate the discussion of sensitive topics (Kvale and Brinkmann, 2009).

Virtual Focus Groups

Researchers increasingly are exploring the potential of using telephone, the Internet, and computer-mediated approaches. Possible advantages include access to participants at a distance, access to persons who may be reluctant to join a face-to-face meeting due to physical problems, reduced cost of arranging sessions, potentially increased disclosure due to perceived anonymity, and more open responses (this varies with topic and population). Asynchronous online methods allow participants to engage at their convenience and their own pace.

Disadvantages for online sessions may include limited depth of comments, limited opportunity for the facilitator to guide and follow up on questions, a potential of online exploitation particularly for young participants, more complex issues of data security, and a possible misunderstanding of comments due to absence of nonverbal cues to assist in understanding comments. Also, it is generally harder to establish trust and rapport, and having to type thoughts is a more demanding task than speaking and may lead to a reduction in participation. In a synchronous session, a delay between responses may lead to less interaction and less synergy. Technology may, but not necessarily will, somewhat limit free communication. Abbreviations of emotion using punctuation symbols can help somewhat in interpreting text. Compared to online focus groups, phone-based focus groups have fewer disadvantages due to the perception of increased social cohesion because participants are able to use information from the characteristics (e.g., intonation, pacing) of actual voices.

In general, data collection may be somewhat constrained by the online type of medium; group interaction will be less and data will be less rich and deep. Adaptations could be made to enhance group interaction such as small size, previous contact

to establish rapport, and extra effort to describe the study and how data will be stored. However, a sensitive or moderately sensitive topic may be more openly discussed online due to a possible comfort level resulting from some degree of anonymity. If the study's purpose can be met, this approach could be preferred.

Mixed-Integrated Methods

Focus groups, as well as other qualitative approaches, have often been used to identify and explore the major concerns in a topic that is new or not well understood. The natural vocabulary of participants can be used in the development of items for a questionnaire. In addition, when planning to use a well-established questionnaire in a population different from the population where it was developed, the meaning of items can be explored in a focus group. Planning and research design can balance the strengths and weaknesses of each of the qualitative and quantitative approaches. Qualitative approaches have been used in combination with quantitative approaches in sequence or concurrently to reinforce, explain, or expand the data. To help explain quantitative results, focus groups have been used to understand the unanticipated or unexpected results of quantitative studies when the psychosocial instruments were used with a new population (Carey and Smith, 1992). Qualitative methods have also been used to explore the processes that underlie quantitative outcomes.

Criticisms

Criticisms common to many qualitative studies include the limited representativeness of samples, questionable validity of results, limited generalizability, lack of synthesis or under-analyzed findings, and inadequate documentation of processes. Concerns unique to focus group include data quality, analysis, and ethics. Data quality will be limited if the session is dominated by one or two people. When this occurs, it usually involves variation in power or status among members, such as gender or professional status, or possibly the facilitator's skill is not well enough developed to intervene effectively. Planning to have compatible members and group-facilitator training, with feedback on performance, will ameliorate most occasions of this problem.

Analytical problems can occur when there is insufficient preparation for the collection and recording of data during a focus group. Collecting only brief comments – snippets – can occur if the size of the group is large and the time is too limited to collect rich stories. When the goal is the collection of rich, detailed data, planning needs to include adequate time for each member to speak especially for people who are not used to being in groups, and the number of questions needs to be limited in order to have time for discussion.

Ethics

Although discussion of some topics in a focus group could lead to discomfort, there is no indication that the level of distress is greater than in everyday life, and there are many reports of therapeutic effects of being part of a focus group. For very sensitive topics and for some studies with people with special needs, the facilitator needs skills to monitor the level of discomfort or distress. In addition, to ameliorate discomfort from discussing some topics, arrangements for referral for readily available support services may need to be established. The inadvertent self-disclosure of more information than a member intended is also a concern that the facilitator needs to monitor.

Ethical issues include recognizing the participants' perceptions of risks and benefits to them, to their communities, and to science. Planning includes addressing issues of confidentiality, required reporting, monitoring level of discomfort, and ethics board approval. Gaining informed consent to participate in a group is a process, not only a signed document. The reading level of the informed consent document should generally not be beyond fifth grade. When appropriate, the informed consent should include the possible future use of data, such as secondary research. Gaining informed consent verbally may be appropriate for illiterate participants and is permitted in some situations.

Confidentiality presents unique ethical challenges in focus group studies because, while the researchers can protect their use of the data, they cannot guarantee that the group members will do the same. It is unethical for researchers to promise participants that no one will disclose what is said in a session. In contrast, required reporting, such as suspected child abuse or criminal behavior, needs to be made clear to group members if the topic could involve such information. Members need to understand the consequences of disclosing such information.

Monitoring the level of discomfort, especially related to sensitive topics, and the need to have ready referral to support services should be part of planning. Debriefing of participants regarding their level of comfort after the session is recommended.

Psychosocial Foundations

By understanding the psychosocial foundations of focus groups, the researcher can appreciate the complexity of data that are collected in a group setting and can plan for optimal use of these data. The major strength, and potential problems, arises from the group dynamics in the group session. Members in the group are interactive in the sense that each comment exists in a social context and is affected by previous comments. A person adjusts the public self that others see to match what he or she understands as appropriate or in best interests of self (Goffman, 1959). *Group think* can arise in focus groups where persons whose opinions are not well formed may be swayed by the common opinion of the group (Janis, 1972). To address these concerns, the group facilitator carefully follows the comments with the goal of grounding the opinions in the key characteristics as recommended by Merton et al. (1990): depth, personal context, range, and specificity.

In a focus group session, the members and the facilitator 'co-construct' the data, developing an understanding of the experiences in the comments by responding and shaping comments. This is similar to this process in an individual interview as described by Kvale and Brinkmann (2009).

Planning

A very important component of a research study is planning. Most limitations arising from a poorly planned design cannot be compensated for with more attention later, like a stricture in a pipeline that limits the subsequent quality of the study. Planning components involved in pre-session work include facilitator and co-facilitator preparation, sampling, group size and segmenting, recruitment, introduction, guideline questions and probes, consent, transcription, translation, and a checklist.

Preparation for the facilitator and co-facilitator begins with becoming familiar with the topic and the target population. In order to plan for recruitment, conduct the session, and explore the comments, it is necessary for the facilitator to understand the comments that are likely to be made in the group, and the local context for those comments. Having logistics settled avoids distractions and the facilitator can give full attention to the task of leading the session. Another model uses co-facilitators; in this case, the roles need to be carefully identified and reviewed after each session for success and problems. Usually the co-facilitator assists with recording and room arrangements, takes notes with attention to the nonverbal communication, and summarizes at the end of the session.

Sampling in qualitative studies is purposeful and focused on sampling for the topic, not for representative sampling of the population. This may involve a broad range or a narrow and deep range, depending on the study's purpose. In many qualitative studies, preliminary analysis is performed after each data collection in order to guide future data collection. This emergent design is in accord with the grounded theory method (Corbin and Strauss, 2008), as well as other qualitative methods. As the study progresses and themes are identified, data collection will focus on aspects that are not fully developed. When new data provide no new information and themes (or categories as described in a later section), then *data saturation* is reached and the collection is considered completed. (Data saturation is the state of data collection when new data do not add new information and the theory or themes are fully developed.)

Generally, it is advisable for members of the group to be homogeneous in status, such as occupation, social class, age, and education. People are more likely to share information with others whom they see as similar. If the group is too heterogeneous, the lower-status members will likely defer to higher-status members and not contribute fully. When status is expected to impede sharing, persons may be divided into similar status groups, a process called *segmenting*. However, Carey (1994) found, in working with a military population, the common experience of having HIV or AIDS overcame the demographic differences in rank.

Previously, focus groups usually consisted of strangers who expected not to have future contact. It is now common for group members to know each other and to expect to have future contact. This has important consequences for the data quality. Members may be more reticent to share, which could lead to self-censored data. On the other hand, knowing fellow members may lead to easier conversation with a shared language and experiences.

Planning for the size of groups considers the experience of the participants with being in groups, the culture of sharing, and the sensitivity of the topic. Persons not used to sharing in groups will feel more comfortable with only three or four others, rather than eight to ten others. People need an unhurried chance to talk in order for the researcher to collect stories instead of brief snippets, which are quick, superficial comments. Five to seven people is usually a workable size, although a greater or smaller number could be optimal.

For recruitment, the researcher may use the Internet, flyers at appropriate gathering places, or work through an organization such as a clinic or school. In the latter case, this is a two-step process with the organization sending the initial invitation that includes contact information for the researcher, the purpose of the study, and how the data will be handled. A consent form may be included at this stage, or may be at a later stage. The researcher will ask if there are special needs such as food requirements (if food is to be provided during the session) or accommodations for hearing or other physical limitations. A reminder phone call the day or evening before the session is scheduled leads to fewer people who do not attend.

Most focus group sessions use broad questions as a guide for the discussion. Guideline questions provide some structure for the session. Questions are developed based on what is known and the purpose of the study, often guided by previous research. It is important to spend adequate time trying out and refining these questions, preferably with pilot sessions. Documenting changes in guideline questions as the study progresses, and explaining why the decisions were made are an important part of documenting the process of the study.

The guideline questions provide a general structure for the discussion by easing group members into a mind-set to help them describe details of their opinions or experiences. One approach is to ask a very general question to get the group oriented to the topic, and then narrow the focus to a more specific aspect of the study. However, sessions often do not proceed as the facilitator anticipated in terms of the order of the questions, usually due to the interest of the members. When possible, it is usually wise to recognize the interest of the group and follow their lead as they likely are following what is most important to them.

An alternative to using guideline questions is the watching of videos, or requesting participants to draw a picture or take a photo. This level of engagement may be particularly useful for children and others who are not used to sharing in a group setting. Also, somewhat different data may be elicited because there is evidence that words and pictures are processed in differing ways in the brain (Padgett et al., 2013). These data may be more related to emotions and may help with articulating deeply held beliefs. Potential drawbacks to using drawings in a focus group include less engagement with the group due to participants' focus on their own drawing and the societal norm for preferring positive pictures.

Analysis of the session may be done with transcripts that are note based, selected excerpts of the session, or verbatim documents. Verbatim transcripts are not synonymous with all that occurred during the session. Transcription necessarily removes some of the richness of the data in the session. In addition to possibly not being an accurate record and difficulty in identifying who is speaking, transcripts generally do not include

voice, gestures, or other nonverbal communication. Capturing the emotional context is necessary. Field notes, especially notes on nonverbal communication and group interaction, are essential to understanding the data. Video recordings also help to address this concern.

Translation of sessions conducted in a language other than one understood well by the researcher is complicated in that the translation process is not merely passive and mechanical, but rather involves the co-construction of knowledge, especially in translating for meaning as contrasted with the word-for-word approach. In this sense, the translator is participating in the research. This role needs to be made explicit and that person should be included in the analysis and interpretation of the findings.

Logistics

Details important to the successful use of focus groups include the room setting, recording procedures and equipment, and food. The room should be comfortable and afford privacy so that few interruptions will occur. Consideration may be given to holding the focus group off the premises of the organization, which is away from the physical setting associated with the research topic, to provide a psychological break and possibly encourage sharing of information.

Food is a surprisingly important aspect, and often has a role in arranging seating. In my experience, brownies were a hit with military personnel, bagels and coffee were appreciated by older citizens for a morning session in a senior center, and pizza went well with high school students. Food facilitates pre-session conversation and provides group members with something to do. This pre-session chatting, often around the food table, helps break the ice socially, but more importantly allows the facilitator to observe the members' characteristics to arrange optimal seating. The placement of individuals around a table can affect the dynamics of the group. For example, the facilitator could place an outspoken person close to her chair, so that the facilitator can better guide the level of participation of that member by turning toward or away from that person. Similarly, a shy or reticent person can be seated directly across from the facilitator allowing for eye contact and other nonverbal means of encouragement to participate. The co-facilitator may be seated back from the group in order to better observe.

Recording details, such as extension cords, duct tape to cover cords and wires, and extra batteries, should be arranged as unobtrusively as possible. After an introduction outlining the purpose of the research, the facilitator then asks permission to record the session, via audiotape or videotape. Recent improvements in technology using videotaping allow the researcher to capture more of the nonverbal aspects of members' participation, which can be helpful in placing comments in context.

Rigor

The aspect of rigor that is unique to focus groups, in contrast to other qualitative methods, is the attention needed to ensuring that the data are not negatively influenced by group dynamics. This can be addressed by attending to the nonverbal communication and following the advice of Merton et al. (1990) to collect the range, depth, personal context, and specificity of personal experiences. This is in addition to the usual recommendations of strategies employed during the study: employing an audit trail, member checks of data, adequate analysis, peer debriefing with colleagues, and negative case analysis (Morse et al., 2002). Member checks are done in the session as the facilitator checks with the participants that the facilitator has captured the main ideas. Peer debriefing with colleagues is an important and ongoing process to check that the process and insights are grounded in the data – both verbal and nonverbal data.

The goal of rigor does not vary with different research purposes; rather, the level of detail in analysis varies. For readers to evaluate the worth of the study findings, they need to know not only what was found but also how it was found. Adequate details need to be provided of the process of moving from raw data to codes, to categories, and to themes, depending on the purpose of the study. Verification strategies that are used to enhance reliability and validity of the research process include methodological coherence, concurrent data collection and analysis, and theory development as appropriate to the research purpose.

Conducting the Session

Topics to consider in preparing to conduct focus group sessions include group management skills, starting the session, session structure and facilitator role, probing (follow-up questions for more detail), and handling problems. Facilitator skills include monitoring discomfort, guiding the discussion, and enhancing data quality. It is the responsibility of the facilitator to monitor the level of distress and refocus the group; and in unusual cases, to stop the session when necessary if a member becomes distressed. Listening to 'what is behind the words' spoken – active listening – can be learned through practice and feedback. The facilitator encourages participation while maintaining a nonjudgmental style. Avoiding nodding after a comment will prevent the perception of the comment as 'correct.'

When one or two members dominate the conversion, participation of others is limited. In addition to thanking the talkative person and turning to the group to ask for other comments, the facilitator can indicate that other comments are welcome. If the talkative person has intentionally been seated next to the facilitator based on the pre-session mingling, the facilitator can slightly turn her/his body away from the talkative person to face others, thereby nonverbally reinforcing the norm of other contributions. It is important to not be too brief with a chatty member, or the rest of the group will feel somewhat chastised also and be more reticent to speak. When the talkative person has been recognized for his/her contribution, no offense will be taken. On the other hand, by seating a quiet person directly across from the facilitator, the facilitator can encourage participation by making eye contact and restating the goal of listening to a range of experiences.

A major concern for data quality is the potential for a 'group think,' the phenomenon of participants being carried along by the group interaction and agreeing with the overall discussion

(Carey and Smith, 1994). It is usually apparent when this is occurring and is a challenge for the facilitator to explore. Going back to Merton and colleague's (Merton et al., 1990) advice to explore the depth and specific details can help to address this challenge. It is important not to directly challenge, but to ask for more information.

The facilitator begins with an introduction to the session. The introduction includes information about the project and the sponsoring organization that was previously provided in the recruitment letter or phone call, and answers any questions. Agreement for recording will have been obtained earlier as part of the consent process and it is repeated here before the video- or audiotaping is begun. Depending on the purpose, the members are told that a range of experiences rather than agreement is the goal. Members are encouraged to share comments that are not necessarily in agreement with each other, and members can feel comfortable to describe a range of opinions because it is made clear that there are no correct responses. Some knowledge of the topic and of the participants is needed to understand the comments well enough to guide the discussion and ensure quality of the data in terms of rich, detailed comments. Two simple, helpful 'rules' are (1) no disparaging remarks about another member's comments, and (2) only one person talking at a time. The request for one person at a time to speak prevents comments from overlapping. Cell phone and pagers are turned off, or if necessary for participants to be reachable, turned to quiet mode.

The purpose of the study guides the decisions on the amount of detail in the guideline questions and in the flexibility in following the questions. For an exploratory topic, less direction from the facilitator, more flexibility in following the questions, and more openness of the questions can allow unanticipated findings to arise that involve issues of importance to the members. The facilitator needs to have ongoing engagement with data to appreciate how the stories are related and how the synergy in the group is influencing the comments. Asking members to 'think back' to the experience can help put them in the setting and bring out some of the richness of their experiences. The 'ice breaker' question is an easy way to open the discussion and gets members comfortable with speaking in the group. A brief, factual question, such as "when did you receive your diagnosis and how did you learn of it," is easily answered as the participants take turns going around the group.

When the discussion is ready to move on to the next question, it is helpful to briefly summarize the main ideas (not characterized as a consensus), at the end of each question. The flow of the discussion may not permit this to be done comfortably, and it is not necessary for each question. A review of the main ideas at the end of the session, 'member checking,' is checking of main ideas for accuracy of understanding of the data collected. It is important to leave time for the last question – "If you could tell the organization just one thing, what would it be?" This last question often elicits the most useful data. Members have been discussing the topic and are usually eager to state their ideas.

Many topics in focus group studies will not lead to any stress or embarrassment, but some could. It has not been my experience, or the experiences of my colleagues, that conflict or embarrassment occurs. Appropriate planning and facilitating skills should lessen the probability of these occurrences. The facilitator monitors stress levels and intercedes to guide the discussion in a safe direction, or in extreme and very unusual cases, the session may be stopped and referral services arranged. When facilitating sessions on sensitive topics, it may be helpful to allow members to describe a friend's experience. The researcher may suspect that the 'friend's' story is actually the member's experience, but the member chooses to tell the story in another's voice. The detachment may arise from a concern that the group will think poorly of the member's comments. The researcher will judge how much to further question and how much to accept without inquiring further into the details.

Probing

The most important aspect of guiding the session to elicit good-quality data is knowing when and how to follow up, or probe. A preliminary type of analysis while conducting the sessions guides the facilitator in seeking further information, more details based on personal experience. Lack of agreement between verbal and nonverbal communications, and inconsistent comments can be bases for probing. Personal and specific details improve the credibility and general quality of the data. It is important to note a change in a participant's opinion during the session and try to determine what led to the change.

Analysis

The goal of analysis is the development of a clear and convincing picture of findings. Because there are few guidelines available, the researcher needs to thoroughly understand the psychosocial context of the data; the various analytic methods available; and the importance of methodological coherence of purpose, process, and product. The goal is to understand the participants' experiences and to communicate the findings so that they may be readily used.

A study's findings can be data rich and analysis poor; that is, they are underanalyzed to address the study's purpose. There are few, if any, agreed-on rules for qualitative analysis of focus group data. This is mostly because this technique, as a data collection method, can be used with different methodologies. From a general review of the literature, it appears focus groups are used most often with grounded theory.

Because each study is unique, guidelines need to be used with creativity and judgment. Charmaz (2006) recommends "imaginatively engaging" with the data and then solidly grounding the interpretation directly back in the data. This recommendation encourages the researcher to explore the possibilities that are not immediately obvious, thereby examining patterns that at first are not apparent. This approach can produce new and important theoretical development.

Some qualitative research methods are not distinct from each other. Approaches range from staying quite close to the data and identifying broad themes to developing explanatory theory through interpretation and synthesis. Most qualitative methods used with focus groups' studies have several components in common: identification of a topic, selection of a population, collection of data, inductive coding and further

iterative grouping of concepts, prioritizing of key ideas, and the use of the researcher as instrument in data collection and analysis. The most commonly used methods in published studies of focus group research include thematic analysis (Braun and Clarke, 2006), discourse analysis (Fairclough, 1993), content analysis (Elo and Kyngas, 2007), narrative analysis (Reissman, 1993), and grounded theory (Corbin and Strauss, 2008), and there are many variations of each of these. It is beyond the scope of this article to further explore the potential for each method.

Software can be a helpful tool to organize the large amounts of data that are often involved in a project. It does take time to master a software program in order for it to be a helpful tool and not a nuisance. Newer programs are capable of using videotapes and some have voice recognition capability to provide rapid transcripts. Some software programs promise to describe group dynamics beyond what researchers can do with other software. This automatic coding and grouping will likely be met with skepticism by some researchers who value the in-depth involvement of the researcher in understanding the text. It will be interesting to keep an open mind and see what enhancements will be helpful as technology continues to advance.

Reporting

This component should be included in the planning phase as an overall structure can guide the study. Consider who will use the information and how they could use it. In addition to the elements common to research reporting, the only element unique to focus group research reporting is the group context of the data. Emphasizing the verification strategies and rigor will assist the readers in evaluating the worth of the study. Please see Carey and Asbury (2012) for a helpful format.

Looking Ahead

The goal of this article is to present the rich potential of focus group research and provide some guidance in planning. Most of the literature on recommendations, although based on psychosocial frameworks, is based on personal experience. What is needed to improve is empirical research, such as the relative impact of group size, segmenting, stranger versus familiarity among members, relative structure in a session, matching demographics of facilitator and members. Technology will increasingly be used and the effects on data need to be explored.

Overall, focus group data collection is a technique that yields rich data that could not be obtained through other means. It is well worth the time and effort to plan and implement it in a study.

See also: Interpretive Methods: Micromethods; Problem Selection in the Social Sciences: Methodology; Qualitative Methods, History of; Qualitative Research; Small Group Research, History of; Symbolic Interaction: Methodology.

Bibliography

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qualitative Research in Psychology 3, 77–101.

Carey, M.A., Asbury, J.-E., 2012. Focus Group Research. Left Coast Press, Walnut Creek, CA.

Carey, M.A., 1994. The group effect in focus groups: planning, implementing, and interpreting focus group research. In: Morse, J. (Ed.), Critical Issues in Qualitative Research. Sage Publications, Thousand Oaks, CA.

Carey, M.A., Smith, M.W., 1992. Enhancement of validity through qualitative approaches: incorporating the patient's perspective. Evaluation & the Health Professions 15 (4), 107–114.

Carey, M.A., Smith, M.W., 1994. Capturing the group effect in focus groups: a special concern in analysis. Qualitative Health Research 4 (1), 123–127.

Charmaz, K., 2006. Constructing Grounded Theory: A Practical Guide through Qualitative Analysis. Sage, Thousand Oaks, CA.

Corbin, J., Strauss, A., 2008. Basics of Qualitative Research, third ed. Sage, Thousand Oaks, CA.

Elo, S., Kyngas, H., 2007. The qualitative content analysis process. Journal of Advanced Nursing 62, 107–115.

Fairclough, N., 1993. Discourse and Social Change. Polity, Cambridge.

Glaser, B.G., 1995. Grounded Theory: 1984–1994. Mill Valley Press, CA. Sociology Press.

Goffman, E., 1959. Presentation of Self in Everyday Life. Anchor Books, New York, NY.

Janis, L., 1972. Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes. Houghton Mifflin, Boston.

Krueger, R.A., Casey, M.A., 2009. Focus Groups: A Practical Guide for Applied Research, fourth ed. Sage, Thousand Oaks, CA.

Kvale, S., Brinkmann, S., 2009. InterViews: Learning the Craft of Qualitative Research Interviewing, second ed. Sage, Thousand Oaks.

Merton, R., Fiske, M., Kendall, P., 1990. The Focused Interview: A Manual of Problems and Procedures, second ed. Free Press, New York.

Morse, J., Barrett, M., Mayan, M., Olson, K., Spiers, J., 2002. Verification strategies for establishing reliability in qualitative research. International Journal of Qualitative Methodology 1 (2). Article 2. Retrieved January 2006 from: www.ualberta.edu.ca/~ijqm.

Padgett, D.K., Smith, B.T., Derejko, K.-S., Henwood, B.F., Tiderington, E., 2013. A picture is worth …? Photo elicitation interviewing with formerly homeless adults. Qualitative Health Research 23, 1435–1444. A.

Reissman, C., 1993. Narrative Analysis. Sage Publishing, Newbury Park, CA.

Folklore

Barbro Klein, Swedish Collegium for Advanced Study (SCAS), Uppsala, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Coined in 1846 to describe the customs, superstitions, and arts "of the olden time," folklore was in the 1990s defined as "expressive forms, processes, and behaviors" that are learned and utilized in face-to-face interaction and are judged to be traditional. This article addresses different understandings of the term along with theories proposed to account for the origin, spread, and functions of epics, games, rituals, and other folklore phenomena. In special focus are six keywords: folk, tradition, genre, performance, text, and art. Also addressed are the role of folklore study around the world today and its role in global heritage politics.

Folklore can be said to have four basic meanings. First, it denotes oral narration, rituals, crafts, music making, and other forms of vernacular expressive culture. Second, folklore or folkloristics names an academic discipline devoted to the study of such phenomena. Third, in everyday usage, folklore sometimes describes colorful 'folkloric' phenomena linked to the music, tourist, and fashion industries. Fourth, like myth, folklore can mean falsehood. Although in this article, focus is placed on the first two meanings, the other two are often present, implicitly or explicitly.

History

A Term Is Invented

The word folklore (spelled 'folk-lore') was proposed by British antiquarian William John Thoms who, in 1846, suggested that this "good Saxon compound" be used to describe "the manners, customs, observances, superstitions, ballads, proverbs," and other materials "of the olden time." Thoms was inspired by the works of the Brothers Grimm, and 'folklore' is apparently a translation of the German word *Volkskunde*, which was introduced in 1806.

Although the word was new, interest in the phenomena themselves was not (Cocchiara, 1981 (1952)). European and Asian rulers had long ordered that traditions that would glorify them be written down, and religious leaders had compiled records of dangerous beliefs in order to weed them out. However, during the national romantic currents of the late eighteenth and early nineteenth centuries, European intellectuals added something new when they insisted that the customs and arts of illiterate peasants were remnants from a remote human past and should be collected and preserved before they were bound to disappear as civilization advanced. Like Johann Gottfried von Herder, intellectuals saw a reflection of the true spirit of nations in the traditions of the anonymous folk living close to the national soil.

A Discipline Is Founded

During the nineteenth century, many people were fascinated with folklore, among them scholars with interests in literature, philology, religion, psychology, history, archeology, anthropology, and other fields. In some countries, enthusiasts formed societies: the British Folklore Society was founded in 1878, and the equivalent in the United States 10 years later. In other countries, primarily in Northern, Central, and Eastern Europe, the idea of folklore was directly linked to nation formation. In Finland, which was in the process of freeing itself, first from Swedish and then from Russian domination, Elias Lönnrot pieced together the *Kalevala Epic* from songs collected in the oral tradition. Published in 1835, the *Kalevala* became the first example of literature written in Finnish and a national symbol. In addition, the songs formed the core of a national folklore archive, and, in 1898, the world's first professorship in folklore was established: the chair in Finnish and Comparative Folklore at the University of Helsinki.

Other university departments were eventually established in some countries in Europe, East Asia, South America, and North America, usually (but not always) in conjunction with the founding of national museums and archives. These institutions could have different designations (e.g., folklore, *Volkskunde*, laographia, folk memory research, folklife research, or ethnology) or could emphasize different empirical traditions. The varieties in political and cultural histories and national styles were significant, even though both the fields of study and the academic institutions grew out of a similar spirit of nation building.

At the same time, anthropologists concentrated on the folklore of colonized (or tribal or indigenous) peoples, whether or not these peoples were in their own country or in colonies far away. It has sometimes been observed that "folklore grew out of nationalism while anthropology grew out of colonialism." There is some truth to this in so far that, to this day, anthropologists (nowadays often indigenous anthropologists) tend to study the folklore of indigenous, illiterate, or 'oral' populations, whereas folklorists tend to focus on literate (albeit often subaltern) people in their own countries. The United States constitutes a noteworthy case in so far that the discipline of folklore developed hand in hand with a protracted "dialogue of dissent" between anthropologists, such as Franz Boas who worked with Native American folklore, and literary scholars, such as Francis James Child, who concentrated on Euro-American traditions (Zumwalt, 1988). In the United States and elsewhere, early research traditions set the stage for political and ideological entanglements that have actuality to this day (see Section Folklore and the Politics of Culture and Heritage).

It should be emphasized that many (if not most) countries in the world lack academic institutions of folklore and that the field tends to be studied within other disciplines or outside formal academic institutions. England, for example, has never established university departments devoted to folkloristics, whereas Scotland, Ireland, and Northern Ireland have deemed it politically and linguistically important to do so.

A History of Approaches

On the whole, the ideas employed to explain and interpret folklore coincide with the major schools of thought in anthropology, beginning with evolutionism and diffusionism and moving on toward functionalism, structuralism, and later tendencies. Yet even more than in anthropology as a whole, the study of folklore long focused on questions of origin; an array of explanations were proposed to account for the origin of such ancient and widely disseminated tales as "The Dragon Slayer" and other traditions. The Indianist school headed by the German orientalist Theodor Benfey sought to establish that all European folktales had originated in India and spread from there. The mythological school, championed by the German-born orientalist Max Müller at the University of Oxford, maintained that all folktales were survivals of ancient nature-myths, whereas ritualists, such as the French folklorist Pierre Saintyves (Émile Nourry), tried to demonstrate that all folktales had originated in rituals.

While these explanations now have primarily historical interest, other early trends of thought remain intellectually influential. Among these is, first, the diffusionist 'historical–geographical' method developed by Finnish scholars at the end of the nineteenth century. In this cumbersome procedure, all recorded versions and variants of a tale type (or a ballad or a riddle) are studied so that the location of the earliest traits, and theoretically also the tale's place of origin, can be determined. The gathering of comparative evidence resulted in massive indexes and catalogs that are still often consulted (Aarne and Thompson, 1961; Uther, 2004). Another intellectual movement that remains important to the understanding of folklore is psychoanalysis (Holbek, 1987). Sigmund Freud, Carl G. Jung, Otto Rank, and others came to regard folklore as essential expressions of the individual or collective human unconscious. A third essential strand of thought is the insight that forms of folklore are not only widely spread but also firmly localized and therefore must be studied in their immediate social contexts. Franz Boas' ethnographies of the stories of native peoples on the northwestern coast of the United States initiated an influential format for the study of folklore in local contexts.

Functionalism, not least Bronislaw Malinowski's analyses of storytelling among the Trobrianders, is a fourth development that has had a lasting impact on the study of folklore. Particularly influential is the psycho-functionalist premise that folklore is "socially sanctioned" fantasy; as such, it permits action that is otherwise not approved (Dundes, 1965: p. 277). Finally, structuralist thought has also endured among students of folklore. One structuralist approach derives from Russian formalism and emphasizes features of plot sequence; in an influential study, Vladimir Propp demonstrated that all Russian fairy tales share the same plot structure (Propp, 1968). A second approach derives from the work of Claude Lévi-Strauss and points to the ways in which logical relationships within myths and tales, such as binary oppositions, reflect universal processes of the human mind. Lévi-Strauss' thinking has been central to the analysis of a wide range of traditional forms, not least Henry Glassie's monumental investigations of vernacular material culture (1982, 1993).

Keywords

Despite the many decades of inquiry, it remains difficult to define folklore. In one attempt proposed in the 1990s, it was suggested that folklore embraces "expressive forms, processes, and behaviors" that are learned, taught, and utilized "during face-to-face interactions"; it was also noted that people judge these phenomena to be traditional "because they serve as evidence of continuities … through time and space in human … thought, belief, and feeling" (Georges and Jones, 1995: p. 1). However, increasingly scholars avoid such generalized statements. In their search for "a common ground for a disciplinary discourse" (Bendix and Hasan-Rokem, 2012: p. 2), they prefer instead to look at folklore as perspectives and processes that might be illuminated through such keywords as "folk," "tradition," "genre," "performance," "text," and "art" (Feintuch, 1995, 2003).

Folk

Influenced by the typologies of early social thought and by such conceptualizations as Robert Redfield's "ideal folk society," students of folklore long took it for granted that the folk (i.e., the carriers and transmitters of folklore) were illiterate peasants or tribal peoples. However, during the last few decades, the word 'folk' is increasingly regarded as a cumbersome inheritance from earlier thinking, and most researchers agree that all human beings in some way or another are involved in folklore processes, as performers, participants, watchers, viewers, readers, or audiences – sometimes indifferent or disdainful audiences (see Section Performance, Text, and Art).

But it has been difficult to abolish the word 'folk.' For one thing, it remains important in studies of folklore in feudal systems or other class hierarchies. Second, it remains alive in such compounds as 'folk art,' 'folk music,' and 'folk dance'; attempts to replace 'folk' with 'people,' 'vernacular,' or other words have not been successful (see Section Folklore and the Politics of Culture and Heritage). Furthermore, despite their conviction that all human beings partake in folklore processes, contemporary folklore researchers resemble their predecessors in concentrating on marginal or subaltern peoples. But their reasons for doing so may differ from those of nineteenth-century scholars; many students now argue that the role of folklorists is to assist peoples who lack political power in their quest to gain a public voice.

Tradition

Used both to designate a process of transmission and to name specific repeatable and retrievable cultural elements (i.e., traditions), the concept of tradition is basic to the understanding of

folklore and folkloristics (Bauman and Briggs, 2003). It has long been recognized that tradition involves not only historical continuity but also change. Because folklore forms are often transmitted orally or in other informal ways, they are extremely adaptable to fleeting moments of social interaction. One could even say that it is because of their adaptability that simple string figures or guessing games have remained in circulation through millennia.

The materials deposited in archives, listed in comparative indexes, and appearing in popular culture demonstrate both the adaptability of some themes and their continuity through time and space. For example, comic book hero Superman travels through space in a womblike vessel only to be found and raised by a poor couple in the middle of North America. His story has parallels with the stories of Moses, Cyros, and other heroes who are similarly found as children floating around in small containers. Furthermore, identity changes are characteristic not only of Superman but also of countless other heroes and heroines in world traditions. The transformation of an orphaned young girl into a rich and beautiful queen is reenacted over and over in oral tales, films, and real life. Folklore themes have a psychological force that makes people recreate them century after century in different communicative modes and media (Bacchilega, 1997).

During recent decades, students of folklore increasingly utilize the verb 'traditionalize' to name processes in which the past is interpreted or reinvented in a given present. Many students of folklore now emphasize that they do not investigate given sets of traditions, but all sorts of practices of traditionalization in social life. Furthermore, by focusing on active traditionalization processes, scholars avoid cementing an opposition between tradition and creativity. They see tradition as a source of creativity, not as a constraint to it (Briggs, 1988; Bauman, 2004).

Genre

Genre means 'kind,' and a genre may be seen as a distinguishable form characterized by a specific combination of stylistic traits and elements of content. Classification into genres was for a long time so central that scholars thought that enumerating genres was the easiest way to describe what folklore is (Dundes, 1965).

When analyzing genres, students of folklore often find themselves caught in dilemmas familiar to all students of culture. On the one hand, all peoples make local, culture-specific, or 'emic' distinctions between genres in their repertoires: coyote tales, superstitions, dirty jokes, little girls' games, gossip, songs while making baskets, and so on. On the other hand, scholars may observe more overriding features in the same materials and may note how elements of content, style, or structure are shared across cultural boundaries. The extent to which local classifications are emphasized and the extent to which broader ones come into play vary with different analytic tasks.

One distinction that was long taken for granted but has proven problematic is that between 'fixed' genres, such as riddles and proverbs, and 'unfixed' genres, such as legends or anecdotes. While some genres, magic formulas for example, are fixed in the sense that they must be recited exactly as learned in order to have power, other genres that were long called 'fixed' may change too. For example, people often do not cite proverbs in full but merely hint at them through their tone of voice. Thanks to performance-centered approaches (see Section Performance), scholars have recently gained new insights into the ways in which oral genres are malleable social phenomena. In social interaction, people mix or blur materials so that artful hybrids emerge. At the same time, genres are needed as 'social contracts' that aid memory and establish structure.

Performance

Since the 1960s, the concept of performance has transformed the study of folklore. Thanks to increasing refinements in recording techniques, scholars have been able to analyze folklore as part of communicative processes in ways that had not been possible before. Two performance-centered approaches will be recognized here; in both, careful ethnographic fieldwork is essential.

One approach is rooted in ethnography of speaking and other sociolinguistic methods. A basic idea is that verbal art and other forms of folklore emerge in social interaction as aesthetically marked forms. They are recognized as something special thanks to cues made through tone of voice, mimicry, gestures, and other paralinguistic devices. Arguing that performance is "constitutive" of verbal art, Dell Hymes, Richard Bauman (1977), and others developed ways of studying folklore as artistic communication in everyday life, communication in which people shape "deeply felt values into meaningful, apposite form" (Hymes, 1975: p. 346). If epic singers, storytellers, weavers, carvers, and carnival participants in the past were seen as transmitters of handed-down traditions, they were now regarded as active artists with power to transform social life (Mills, 1991).

A second performance-centered approach is oral formulaic theory, which was first developed by Albert Lord (1960, 2000) and others who, in studies of South Slavic oral epic poetry, show how songs that are many thousands of lines long can live in tradition, although they have not been written down. They persist because singers learn to draw on a store of formulas and formulaic expressions and learn to adjust their singing to different social events. Several scholars have continued along these paths, and oral formulaic studies are now enriching the understanding of both ancient (in particular, Homeric) epic singing and contemporary performances of epic songs and other forms of oral poetry in India, China, Africa, Eastern Europe, and the United States (Finnegan, 1992; Honko, 1998; Foley, 2002).

Text

It has long been recognized that the distinction between oral and written texts is far from firm and that many kinds of texts can be of interest to students of folklore (Hasan-Rokem, 2000). Some folklore texts have led their primary existence in writing, for example as court records, religious documents, or newspaper articles, and today we find countless texts and expressions on the internet that are of interest to students of folklore. Other folklore texts have circulated primarily as visual representations, for example fables painted inside European medieval churches and so-called xeroxlore. Furthermore, literary authors have continuously made use of traditional themes, and folklore texts can be located in the works of celebrated writers. For example,

contemporary African authors, such as Ayi Kwei Armah, acknowledge their profound debt to oral narrative art, and Shakespeare often built his plays on traditional materials. Actually, a great many art forms can profitably be studied as folklore texts. Classical ballet, for example, shares some of its stylized and repetitive character with the European folktale.

During the last few decades, performance-centered approaches have opened up oral texts in unprecedented ways in that scholars have invented new methods to translate oral performances into writing. In procedures that have sometimes been named "laying out texts ethnopoetically," students transcribe not only words but also pauses, cadences, gestures, mimicry, and other oral features (Briggs, 1988). These procedures have led to many insights. One is that oral narration in many cultures tends to be closer to drama than to prose literature. Another is that repetitions and other features of oral narration (or visual folk art) are not to be dismissed as poor artistry as was often the case in the past. Rather, repetitions can be seen as ways to control complex materials. A repetitive structure simultaneously offers restrictions and creative freedom; it aids memory at the same time as it frees fantasy and improvisation.

During the last few years, the term 'textualization' is increasingly used to describe modes of translating oral performances onto the written page (Honko, 1998). Long regarded as routine exercises, textualization processes are now seen as central analytic tasks.

Art

On some level, scholars always seem to have agreed that folklore has to do with art, creativity, and imagination. Even at times when the study of folklore was reduced to enumerations of genres and tale types, there was a sense that folklore touches upon the aesthetic dimensions of life. This sense has become heightened during the last few decades, in part due to the interest in folklore as performance. This is particularly the case in the United States, where all kinds of vernacular expressions – from singing to cooking and quilting – are often called 'the folk arts.'

The focus on art is linked to several perspectives. First, students make a distinction between less articulate folklore performers and artists who are celebrated in their communities and have the power to move others to action. Second, students of folklore have made important discoveries about social life by focusing, not on finished products, but on moments when material objects or community events are being planned or created. Third, an interest in art has led to new analyses of the languages of folklore. Whether verbal, visual, musical, kinesthetic, or tactile, these languages are bold in their use of intense and multivalent metaphors and other tropes.

It is because of this expressive intensity that folklore, like other forms of art, can be a vehicle through which people can examine or critique the lives they lead (Scheub, 1998; Noyes, 2003). For example, to people who have fled from one part of the world to another, a few songs, stories, or dishes can become embodiments of a precious inheritance that must be preserved and reconstituted in new lands. Conversely, folklore can incarnate some of the worst aspects of life: detested symbols of racism or sexism, humiliating jokes that outsiders direct at one's own group, and rumors and legends that demonize people(s) (Goldstein, 2004). Because of its power of presence, folklore can be invoked to hold communities together or to separate them from one another.

Folklore and the Politics of Culture and Heritage

Among contemporary scholars, there is both confidence and unease about the future of folklore and folklore study. There is confidence, for example, about the supply of material. The idea that folklore is 'dying out' is no truer today than it was 200 years ago (Bausinger, 1990 (1961)). People shape rituals and tell stories, perhaps more than ever before, but they do so not only in face-to-face interaction but also on electronic media. There is also confidence in the sense that interest in the field continues to expand. For many, the study of folklore constitutes a crucial approach to understanding human life (Briggs and Shuman, 1993). Some scholars are convinced that, regardless of the discipline to which students of folklore belong, today these students play a special role as advocates of the artistic expressions and traditions of all those groups and individuals who are caught in the cracks of modern life. Most notably, there is a heightened interest in folklore and the study of folklore outside of Europe and North America, both among native scholars and among scholars from other countries (Yankah, 1989). This is particularly true in Asia, primarily in China and India, where the number of folklore departments is increasing; but there is new interest also in Africa (Peek and Yankah, 2004). In part, these developments are a result of critical reflections on the discipline's long Eurocentric and colonial history (Naithani, 2010). Folklore researchers on all continents now recognize the need for a thorough revision of the history of the field, a revision in which Asian, Latin American, African, and Oceanic perspectives are highlighted (Bendix and Hasan-Rokem, 2012).

But at the same time, there is uncertainty about the directions of folklore study. The prevalence of folklore on the internet has led folklorists to wonder how understandings of folklore will have to be modified. Does it still make sense to speak about 'face-to-face interaction' or even about 'the social base' of folklore? Many uncertainties and critical reflections concern the discipline's often naïve, and sometimes disastrous, ideological and political entanglements, not least in Europe (Bendix, 1997). Among the indications of these uncertainties are the seemingly never-ending controversies surrounding the designation of the field. In Germany, the term *Volkskunde* is now often avoided, because some practitioners were deeply involved with the creation of symbols for the Third Reich, and in France the word 'folklore' still resonates with memories of fascist scholarship. In most European countries, 'ethnology' is now preferred to compounds containing 'folk' as an overarching disciplinary designation (see Section Folk).

Many uncertainties concern the paradoxical situation that the study of folklore is almost everywhere a construction in the-service of nation building and related local endeavors at the same time as it regards itself an international field. As was the case in the nineteenth century, national concerns often continue to prevail over international and transnational ones. For example, in

nations reconstituted after the fall of the Soviet Union, folklore and folklore archives play a critical role in efforts to select and (re-) construct visions of exclusive national heritages. In other countries, the links to nation building can take other expressions. In the United States, for instance, folklore and folklorists have become important in affirming a vision of a country in which a multitude of diverse cultures can coexist. This is particularly true of so-called public folklorists (Baron and Spitzer, 2007). Whether they are employed in museums; at the American Folklife Center in Washington, DC; as state folk arts coordinators; or in privately funded endeavors, these professionals have special skills in collecting, preserving, and presenting the folklore of (and with) an immense diversity of groups.

On a global level, folklorists are now entangled in cultural politics and heritage politics, perhaps more than ever before (Kirshenblatt-Gimblett, 1998). In the last few decades of the twentieth century, folklorists worked with UNESCO in efforts to safeguard folk traditions around the globe from the onslaughts of modernization. These efforts assumed new dimensions in 2003, when (31 years after adopting a convention concerning material heritage) UNESCO adopted the Intangible Cultural Heritage Convention (ICHC), which, in 2012, was ratified by 148 countries. To a great extent, the phenomena selected as intangible cultural heritage "in need of urgent safeguarding" (such as Iranian Naqqali storytelling) or selected as "representative" (such as the celebration of *nowruz* in several Asian countries) have long been studied within the purviews of folkloristics.

The 2003 UNESCO convention has encouraged national or local community endeavors to increase their efforts to appoint certain forms of folklore and other cultural materials as worthy of preservation. These efforts raise many questions concerning ethics, intellectual property rights, identity politics, and local ecologies and economies, not least those involving tourism. Is the protection of selected phenomena always morally good, valuable, and harmless? In whose interest are the efforts being carried out? Whose traditions should be safeguarded, and whose should be deleted? To what extent will protection freeze certain forms and ideas and appoint them as the correct and authentic ones to the detriment of other traditions?

The preservationist impulses that have guided the field of folklore since Herder's days (and before) are now assuming bureaucratic proportions that were once unthinkable. Undoubtedly, different kinds of heritage politics and policies and a variety of ideological mazes will become important foci of attention for folklore researchers in years to come. In the meantime, people will go on making music, telling stories, playing games, performing rituals, and whittling toys – in their families, on the social media, or to attract visitors to festivals and performance sites. To scholars, it is important to recognize and investigate the multitude of levels and perspectives that are involved in the study of folklore.

See also: Cultural Heritage; Cultural Resource Management: Conservation of Cultural Heritage; Ethnology; Everyday Life, Anthropology of; Geist (Spirit): History of the Concept; History from Below, the History of Everyday Life, and Microhistory; Oral History; Performance: Anthropological Aspects; Ritual and Symbolism, Archaeology of; Ritual; Tradition, Anthropology of; Tradition: Social.

Bibliography

Aarne, Antti, Thompson, Stith, 1961. The Types of the Folktale. A Classification and Bibliography. FF Communications, Helsinki. No. 184.
Bacchilega, Cristina, 1997. Postmodern Fairy Tales: Gender and Narrative Strategies. University of Pennsylvania Press, Philadelphia.
Baron, Robert, Spitzer, Nicholas R. (Eds.), 2007 (1992). Public Folklore. Smithsonian Institution Press, Washington, DC.
Bauman, Richard, 1977. Verbal Art as Performance. Waveland Press, Prospect Heights, IL.
Bauman, Richard, 2004. A World of Others' Words: Cross-Cultural Perspectives on Intertextuality. Blackwell Publishing, Oxford.
Bauman, Richard, Briggs, Charles L., 2003. Voices of Modernity: Language Ideologies and the Politics of Inequality. Cambridge University Press, Cambridge.
Bausinger, Hermann, 1990 (1961). Folk Culture in a World of Technology. (Elke Dettmer, Trans.). Indiana University Press, Bloomington.
Bendix, Regina, 1997. In Search of Authenticity: The Formation of Folklore Studies. University of Wisconsin Press, Madison.
Bendix, Regina, Hasan-Rokem, Galit (Eds.), 2012. A Companion to Folklore. Wiley-Blackwell, Chichester.
Briggs, Charles L., 1988. Competence in Performance: The Creativity of Tradition in Mexicano Verbal Art. University of Pennsylvania Press, Philadelphia, PA.
Briggs, Charles L., Shuman, Amy (Eds.), 1993. Theorizing folklore: toward new perspectives on the politics of culture. Western Folklore 52, 109–400 (special issue).
Cocchiara, Guiseppe, 1981 (1952). The history of folklore in Europe (John N. McDaniel, Trans.). Institute for the Study of Human Issues (ISHI), Philadelphia.
Dundes, Alan (Ed.), 1965. The Study of Folklore. Prentice-Hall, Englewood Cliffs, NJ.
Feintuch, Burt (Ed.), 2003. Eight Words for the Study of Expressive Culture. The University of Illinois Press, Urbanaand Chicago (Revision of 1995. Common Ground: Keywords for the Study of Expressive Culture. Journal of American Folklore 108 (430), 391–550).
Finnegan, Ruth, 1992. Oral Traditions and the Verbal Arts: A Guide to Research Practices. Routledge, London.
Foley, John M., 2002. How to Read an Oral Poem. University of Illinois Press, Urbana.
Georges, Robert A., Jones, Michael Owen, 1995. Folkloristics: An Introduction. Indiana University Press, Bloomington.
Glassie, Henry, 1982. Passing the Time in Ballymenone: Culture and History of an Ulster Community. University of Pennsylvania Press, Philadelphia.
Glassie, Henry, 1993. Turkish Traditional Art Today. Indiana University Press, Bloomington.
Goldstein, Diane E., 2004. Once upon a Virus: AIDS Legends and Vernacular Risk Perception. Utah State University Press, Logan.
Hasan-Rokem, Galit, 2000. Web of Life: Folklore and Midrash in Rabbinic Literature. Stanford University Press, Palo Alto.
Holbek, Bengt, 1987. Interpretation of Fairy Tales: Danish Folklore in a European Perspective. FF Communications, Helsinki. No. 239.
Honko, Lauri, 1998. Textualizing the Siri Epic. FF Communications, Helsinki. No. 264.
Hymes, Dell, 1975. Folklore's Nature and the Sun's Myth. Journal of American Folklore 88 (350), 345–369.
Kirshenblatt-Gimblett, Barbara, 1998. Destination Culture: Tourism, Museums, Heritage. University of California Press, Berkeley.
Lord, Albert B., 1960. The Singer of Tales. Harvard University Press, Cambridge, MA (2000. second ed. by Stephen A. Mitchell and Gregory Nagy, with audio and video CD).
Mills, Margaret A., 1991. Rhetorics and Politics in Afghan Traditional Storytelling. University of Pennsylvania Press, Philadelphia.
Naithani, Sadhana, 2010. The Story-Time of the British Empire: Colonial and Post-colonial Folkloristics. Mississippi University Press, Jackson.
Noyes, Dorothy, 2003. Fire in the Plaza: Catalan Festival Politics after Franco. University of Pennsylvania Press, Philadelphia.
Peek, Philip M., Yankah, Kwesi (Eds.), 2004. African Folklore: An Encyclopedia. Routledge, New York and London.
Propp, Vladimir, 1968 (1928). Morphology of the Folktale. Trans. Lawrence Scott. University of Texas Press, Austin.
Scheub, Harold, 1998. Story. The University of Illinois Press, Madison.
Uther, Hans-Jörg, 2004. The Types of International Folktales: A Classification and Bibliography Based on the System of Antti Aarne and Stith Thompson. Three parts. FF Communications, Helsinki. No. 284–286.
Yankah, Kwesi, 1989. The Proverb in the Context of Akan Rhetoric: A Theory of Proverb Praxis. Peter Lang, Bern. (2012. Reprinted by Diasporic Africa Press, New York).
Zumwalt, Rosemary Lévy, 1988. American Folklore Scholarship: A Dialogue of Dissent. Indiana University Press, Bloomington.

Food Deserts in US Cities

Amy M Donley and Marie C Gualtieri, University of Central Florida, Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Food insecurity is a worldwide problem experienced by millions of people. A specific manifestation of food insecurity has been documented in cities of the United States. It is related to the so-called food deserts, urban areas where individuals make up for the lack of supermarkets located in their vicinity by eating at nearby fast food restaurants and shopping at local convenience stores where they buy generally unhealthy food. The number of individuals living in food deserts in the United States, or in other countries, is unknown because there is no standard definition for the term 'food desert.' Food deserts are viewed as a form of constructed inequality because their distribution is not an accidental, naturally occurring phenomenon. For local officials to address the issue of food deserts is of the utmost importance, as there is a need for alternative ways to supply urban residents with nutritious, affordable foods.

Global Hunger

In 2013, 842 million people in the world are experiencing chronic hunger, meaning they do not have enough to eat daily (FAO, IFAD, and WFP, 2013). The majority of individuals facing chronic hunger and malnourishment reside in developing countries. More specifically, Asia has approximately 552 million individuals facing hunger, which is the largest concentration of individuals experiencing food insecurity worldwide (FAO, IFAD, and WFP, 2013). On average, 1 in 6 children residing in developing countries is underweight, and annually, poor nutrition is the cause of death for approximately 3.1 million children (Black et al., 2013). Even while malnourishment and hunger exist in these places, developing countries are progressing toward the first goal of the Millennium Development Goals set forth by the United Nations, which is to "eradicate extreme poverty and hunger by 2015" (FAO, IFAD, and WFP, 2013). While the 2015 goal has already been met in some areas of the world, other areas have either made no progress or will not meet their goal if their current trend of ending hunger continues. Therefore, despite efforts, hunger is still a serious problem that people continue to experience globally (Millennium Development Goals, 2013).

According to the Food and Agriculture Organization of the United Nations, policies seeking to improve agricultural production and increase food accessibility would ultimately reduce hunger in areas where food insecurity is prevalent (FAO, IFAD, and WFP, 2013). More specifically, these policies could support markets in these areas, provide employment opportunities, and thus, lead to economic growth, while reducing food insecurity (FAO, IFAD, and WFP, 2013). Public policies toward gender equality have also been suggested as a way to reduce food insecurity; in many countries, women farmers have less access to resources (land, capital, credit, markets) compared to their male counterparts. If women had the same amount of resources to use, then it is estimated that the number of individuals facing hunger would decrease by 150 million (Food and Agriculture Organization of the United Nations, 2011). Remittances are also discussed as a way to reduce hunger and boost economic well-being (Bassett, 2010; FAO, IFAD, and WFP, 2013). Remittances are when individuals send some of their paycheck back to family members in their home country, typically a less developed country. This source of income provides family members with money for food and to cover expenses. Additionally, in the bigger picture, the money gets put into the country's economy in which it is spent. By giving people more power in terms of food systems, known as food sovereignty, ongoing relationships can be created between food providers and consumers, as well as a relationship between local land and the residents that use it.

While hunger occurs globally and mainly in the developing economies, this article examines the issue of hunger in relation to *food deserts* experienced in the United States. The phrase 'food desert' does not translate universally. By the US standard, a large portion of the world would qualify as a 'food desert.' But contrary to world opinion, there are still hungry and food-insecure people in the United States and a major contributing factor are the large number of 'food deserts' in US cities.

This article addresses the multiple definitions of food deserts, theories of how these areas developed in the United States, problems and controversies, and alternative food sources that could be used as solutions to food deserts.

Hunger in the United States

Food insecurity in the United States has no boundaries. The issue of hunger impacts individuals of different socioeconomic backgrounds, geographical locations, ages, genders, and races (Hunger Facts, 2012). In a recent report, Coleman-Jensen et al. (2013) found that 49 million Americans experienced food insecurity, with 33.1 million of those individuals being adults and 15.9 million of them being children. Additionally, 4.8 million seniors aged 60 and older are experiencing food insecurity (Ziliak and Gundersen, 2013). The issue of food insecurity transcends just hunger, it also impacts childhood development and exacerbates health issues among many adults and seniors (Bhargava et al., 2012; Bhattacharya et al., 2004; Cook et al., 2004; Holben, 2010; Rose-Jacobs et al., 2008; Skalicky et al., 2006; Stuff et al., 2004).

Food insecurity is measured on different levels: least severe, moderately severe, and more severe (Cook et al., 2004).

The difference among these levels lies with how food is rationed within households. Least severe is described as making adjustments to one's diet and variety of food. Moderately severe is described as cutting the sizes of meals and the amount of food intake. More severe food insecurity takes place in households typically with children. Usually, adults will reduce their food intake before the children. However, in a 'more severe' household, the child's intake is reduced as well because there are no other options.

In the standard United States Department of Agriculture (USDA) questions used to measure food insecurity, the problem is *defined* as lacking enough money to buy the amount and variety of food one needs or wants. It is therefore worth noting here that food insecurity can also result from mobility and disability issues (unable to get to food outlets), cultural preferences (nearby stores do not carry the food products a household would normally eat), or insufficient social capital (no one to help in procuring the food one needs.)

What Is a Food Desert?

The issue of food insecurity is centered on the '3As': Accessibility, Adequacy, and Affordability. While the term was first developed in the United Kingdom (Cummins and Macintyre, 1999), according to the USDA 'food deserts' are defined as "urban neighborhoods and rural towns without ready access to fresh, healthy, and affordable food. Instead of supermarkets and grocery stores, these communities may have no food access or are served only by fast food restaurants and convenience stores that offer few healthy, affordable food options" (Food Deserts, 2013). This definition specifically describes low-income neighborhoods. While this definition is used nationwide, it may miss the differences found in the food deserts of urban and rural areas. In urban areas, a location is deemed a food desert if there is no accessibility to food within a one-mile radius. In rural areas, a location is considered as a food desert if there is no accessibility to food within a ten-mile radius (Berg and Murdoch, 2008; Sparks et al., 2010; USDA, 2009; Wrigley et al., 2002). Food deserts, typically low-income communities, are areas that experience a special type of food insecurity. Food deserts are 'insecure' because there is no adequate food, meaning deemed appropriate to consume by the USDA's food pyramid guide, to purchase within a manageable distance from an individual's residence. The establishments where food can be found typically lack variety and nutritious food. However, these foods are not necessarily affordable and the lack of nutrition has caused an abundance of health issues among residents of these food desert areas (Andrews et al., 2013).

However, the USDA definition of a food desert is not the only one. Other definitions include, in addition to the geographic proximity, the elusive nature of accessing nutritious foods due to physical and economic barriers (Reisig and Hobbiss, 2000). Others focus on the quality of food available to residents and the affordability of those foods based on the socioeconomic condition of residents in the community (Cummins and Macintyre, 2002; Reisig and Hobbiss, 2000; Walker et al., 2010). One study operationalized the term 'food desert' as "urban areas with 10 or fewer grocery stores and no stores with more than 20 employees" (Hendrickson et al., 2006). These alternative definitions are used because rather than being concerned with just the factor of nutrition, these researchers were more concerned with the type, size, and number of stores available to residents and not necessarily the quality of food (Hendrickson et al., 2006).

Theories on How Food Deserts Formed in the United States

Researchers propose different theories to explain how food deserts emerged in the United States. One theory is that chain supermarket stores have put smaller, neighborhood 'mom and pop' grocery stores out of business (Cliff et al., 2004). These local 'mom and pop' stores are forced to close from the enticing visions created by these large supermarkets; they have large parking lots, longer hours, and some even offer to bag your groceries and bring them out to your car, something small stores cannot offer. However, these types of 'mega stores' are only available for those who have cars or can access and afford public transportation (Cliff et al., 2004). Without transportation to these locations, many are left with doing their grocery shopping at convenience stores where nutritious food is less available (Hendrickson et al., 2006).

Other researchers argue that food deserts arose in inner-city areas because the median income dropped when affluent residents migrated to suburban areas in the 1960s through the 1980s (Bianchi et al., 1982; Flournoy, 2006; Nyden et al., 1998; Wienk, 1979; Walker et al., 2010). Studies show that this migration to suburbs from urban areas resulted in closing of one-half of the supermarkets in the three largest cities of the United States (Alwitt and Donley, 1997; Diesenhouse, 1993; Miller, 1994; Walker et al., 2010). The same migration caused a significant decrease in the median income of the inner-city areas, transforming many aspects of these cities. The migration to suburban areas also resulted in fewer places to purchase groceries. Shaffer (2002) explored the gap of supermarkets between lower-income, inner-city areas and middle-income, suburban communities in Los Angeles, California. She found that middle- and upper-income communities have twice as many supermarkets as low-income communities. Specifically, white residential areas had three times the supermarkets than black residential communities. Other studies analyzing access to grocery stores in communities by racial composition and poverty produced similar findings (Morland et al., 2002; Powell et al., 2007a; Zenk et al., 2005). Morland et al. (2002) collected data on the dietary intake of residents participating in the Atherosclerosis Risk in Communities Study to examine the local food environment's impact on nutrition. These communities were located in Maryland, North Carolina, Mississippi, and Minnesota. Findings suggested that census tracts in which white Americans resided had 5 times the number of supermarkets that black American neighborhoods had. Findings were similar in Zenk et al.,'s (2005) study examining the accessibility of food stores by the income level and racial composition of neighborhoods. Using GIS (Geographic Information Systems) analysis of Detroit, Michigan, findings suggest that residents of high-income neighborhoods containing a high population of white residents had greater access

to supermarkets and large-chain food stores. Residents of low-income neighborhoods containing a high population of black and minority residents had greater access to small grocery stores.

Compared to suburban areas, land prices in urban areas are high and often only small plots are available. This makes it more expensive or sometimes impossible for retailers to build supermarkets in urban areas (Alwitt and Donley, 1997; Walker et al., 2010). In order to profit, supermarkets must move large quantities of goods. This proves to be difficult in urban communities where population density is high, but poverty is also high. The space issue is different in the development of supermarkets in rural areas. In this case, plenty of relatively cheap land is available but population density is low, so moving large volume of product through rural stores is problematic. The optimal combination of land, affluence, and people is found in the suburbs, and that is where the supermarkets tend to be (Grocery Store Development, 2012). Because it is difficult to successfully fit supermarkets into heavily urbanized areas, alternative must be created to ensure a healthy lifestyle for residents. However, some communities are faster than others in developing these alternatives.

Problems and Controversies

Because of the different definitions of what constitutes a food desert, it is difficult to have an accurate count of how many individuals reside in such areas (Cummins and Macintyre, 2002; Cummins, 2003; Reisig and Hobbiss, 2000; Shaw, 2006). Additionally, some people do not accept the phrasing and definitions of food deserts because they do not suggest a solution; moreover, the designation of a neighborhood as a 'food desert' creates stigma that makes these neighborhoods even less attractive (Bare, 2013). Thus, the term, 'food desert' is problematic for many advocates, one of whom notes: "A desert [meaning the type of climate] is a natural phenomenon. Lack of access to fresh, healthy food is not natural. It is not accidental" (Cook, 2010). Differential access to healthy affordable food, it is argued, is a socially constructed inequality; lack of access to affordable and healthy food in the areas where food deserts are located results from intentional decisions made by profit-maximizing businesses (Cook, 2010). The further implication is that the problem will not be solved simply by placing supermarkets in these areas; rather, local officials, service providers, and community advocates need to develop alternatives that will bring affordable and adequate food into these areas.

Residents' nutrition is often at stake because they make decisions on their food intake based on the food available to them in their neighborhood (Furey et al., 2001). In food deserts, it is typical for residents to purchase their food at the local convenience store. According to Drewnowski and Specter (2004), most food available at these locations is considered 'empty calorie' food. Hendrickson et al. (2006) state that "low-income neighborhoods have a higher concentration of fast-food restaurants and corner stores that offer prepared foods when compared to higher income communities." The American Dietetic Association believes that this kind of food insecurity can lead to negative consequences for residents, including "physical impairments, disease, psychological suffering, and sociofamilial disturbances" (Position of the American Dietetic Association, 2006). Researchers have found this to be true: individuals exposed to processed foods full of sugar and sodium experience negative health outcomes when compared to individuals who have a nutritionally balanced diet (Block and Kouba, 2006; Lewis et al., 2005; Swinburn et al., 2004). Research has found that neighborhood food environments, such as the number and types of restaurants and types of grocery stores, are associated with obesity (Powell et al., 2007b).

Another controversy surrounding food deserts is the question, 'Do they actually exist?' The idea that established the paradigm of food deserts was that people were not eating healthily because their location limited their access to full-service supermarkets. However, some studies have found that access and geography may not play a large role in healthy eating (Cummins et al., 2005; Drewnowski et al., 2010; Aggarwal et al., 2014; Wrigley et al., 2003). More specifically, Wrigley et al. (2003), Cummins et al. (2005), and Aggarwal et al. (2014) found that access to healthy foods does *not* mean that people develop healthy eating habits. Drewnowski et al. (2010) and Aggarwal et al. (2014) found by using census tract data that individuals do not necessarily shop and purchase food at supermarkets close to their residential areas. In fact, only 15% of respondents shopped within their home census tract, while others traveled nearly twice the distance to their primary food store, which was not necessarily their closest food store. The choice of supermarket was based on demographic (Aggarwal et al., 2014) and socioeconomic factors (Drewnowski et al., 2010), not convenience or location.

Possible Solutions and Alternatives

So what can help to eradicate food deserts (assuming for the moment that they do in fact exist)? While there are ideas on what could potentially be successful alternatives for food acquisition, little research has evaluated effective approaches related to food acquisition barriers (Andrews et al., 2013). In attempts to build upon the literature, Andrews et al. (2013) examined how an increase in Supplemental Nutrition Assistance Program (SNAP; aka 'food stamp') benefits levels affected SNAP redemptions at supermarkets and supercenters. More specifically, would people be more apt to travel the distance to supercenters, where prices are lower, if the SNAP program created a bigger incentive to do so? Their findings suggest that individuals would travel the distance to supercenters if the SNAP benefits were increased. However, if gas prices increased, then those who lived in 'severe' food deserts were less likely to shop at these locations (Andrews et al., 2013).

What can be changed locally? A local food system can help those who live in food deserts through Public Market Distribution, Regional (Cooperative) Food Hubs, and Community Supported Agriculture (Alternative, 2012). Public Market Distribution, meaning a resource that sells goods within local communities, serves a community a few times a week, often in the form of temporary farmer's market stands but sometimes through permanent structures. Regional (Cooperative) Food Hubs link producers to new markets. The USDA describes

a food hub as "a centrally located facility with a business management structure facilitating the aggregation, storage, processing, distribution, and/or marketing of locally/regionally produced food products" (United States Department of Agriculture, 2013). This allows for farmers to drop off their goods to distributors and customers who want to purchase locally grown food. This helps farmers because it allows their goods to be distributed across a larger area. Lastly, Community Supported Agriculture can help ameliorate food deserts. Also known as 'community gardens,' these areas of supported agriculture allow individuals to have access to fresh foods right in their neighborhood yards.

A standard definition would assist in food desert mapping and research. There is a lot of controversy surrounding not only the term, but also what qualifies as a food desert. By developing a concrete, formal definition agreed upon by all, we can map locations and count individuals that experience food deserts more accurately. Such data could then be used to educate the public and local officials on the topic of food deserts, and therefore promote plausible solutions to help reduce the food insecurities and hardship, individuals in these areas face.

See also: Alternative Food Movements; Climate Change and Global Poverty; Culture of Poverty: Critique; Food in Anthropology; Geographies of Food: Politics, Place, Power Reprised; Poverty, Culture of; Poverty, Geography of; Poverty, Sociology of; Slow Food as a Social Movement; The History and Sociology of Food; Urban Poverty in Neighborhoods; Urban Studies: Overview.

Bibliography

Alternative Food Distribution Networks May Alleviate Food Deserts, 2012. http://www.triplepundit.com/2012/05/alternative-food-distribution-networks-alleviate-food-deserts/ (retrieved 02.12.13.).

Aggarwal, Anju, Andrea, J. Cook, Jiao, Junfeng, Seguin, Rebecca A., Moudon, Anne Vernez, Hurvitz, Philip M., Drewnowski, Adam, 2014. Access to supermarkets and fruit and vegetable consumption. American Journal of Public Health 104, 917–923.

Alwitt, Linda F., Donley, Thomas D., 1997. Retail stores in poor urban neighborhoods. The Journal of Consumer Affairs 31 (1), 139–164.

Andrews, Margaret, Bhatta, Rhea, Ploeg, Michele Ver, 2013. An alternative to developing stores in food deserts: can changes in SNAP benefits make a difference? Applied Economic Perspectives and Policy 35 (1), 150–170.

Bare, John, 6 November 2013. Get Rid of the 'Food Desert' Label. CNN.com. http://www.cnn.com/2013/11/06/opinion/bare-food-desert/ (retrieved on 29.11.13.).

Bassett, Thomas J., 2010. Reducing hunger vulnerability through sustainable development. Proceedings of the National Academy of Sciences of the United States of America 107 (3), 5697–5698.

Berg, Nathan, Murdoch, James, 2008. Access to grocery stores in Dallas. International Journal of Behavioural and Healthcare Research 1, 22–37.

Bhargava, Vibha, Sun Lee, Jung, Jain, Rahul, Ann Johnson, Mary, Brown, Arvine, 2012. Food insecurity is negatively associated with Home health and out-of-pocket expenditures in older adults. Journal of Nutrition 142 (10), 1888–1895.

Bhattacharya, Jayanta, Currie, Janet, Haider, Steven, 2004. Poverty, food insecurity, and nutritional outcomes in children and adults. Journal of Health Economics 23 (4), 839–862.

Bianchi, Suzanne M., Farley, Reynolds, Spain, Daphne, 1982. Racial inequalities in housing: an examination of recent trends. Demography 19 (1), 37–51.

Black, Robert E., Victoria, Cesar G., Walker, Susan P., Bhutta, Zulfigar A., Christian, Parul, Onis, Mercedes de, Ezzati, Majid, Granthan-McGregor, Sally, Katz, Joane, Martorell, Reynaldo, Uauy, Ricardo, The Maternal and Child Nutrition Study Group, 2013. Maternal and child undernutrition and overweight in low-income and middle-income countries. The Lancet 382 (9890), 427–451.

Block, Daniel, Kouba, Joanne, 2006. A comparison of the availability and affordability of a market basket in two communities in the Chicago area. Public Health Nutrition 9 (7), 837–845.

Coleman-Jensen, Alisha, Christian, Gregory, Anita, Singh, 2013. Household Food Security in the United States in 2013. ERR-173, U.S. Department of Agriculture, Economic Research Service, September 2014.

Cook, Christopher, 2010. Covering food deserts: tips for bringing context to a complex story. http://www.foodfirst.org/en/node/3209 (retrieved 02.12.13.).

Cook, John T., Frank, Deborah A., Berkowitz, Carol, Black, Maureen M., Casey, Patrick H., Cutts, Diana B., Meyers, Alan F., Zaldivar, Nieves, Skalicky, Anne, Levenson, Suzette, Heeren, Time, Nord, Mark, 2004. Food insecurity is associated with adverse health outcomes among human infants and toddlers. The Journal of Nutrition 134 (6), 1432–1438.

Cummins, Steven, Macintyre, Sally, 1999. The location of food stores in urban areas: a case study in Glasgow. British Food Journal 101 (7), 545–553.

Cummins, Steven, 2003. The local food environment and health: some reflections from the United Kingdom. American Journal of Public Health 93 (4), 521.

Cummins, Steven, Macintyre, Sally, 2002. Food deserts—evidence and assumption in health policy making. British Medical journal 325, 436–438.

Cummins, Steven, Patticrew, Mark, Higgins, Cassie, Findlay, Anne, Sparks, Leigh, 2005. Large scale food retailing as an intervention for diet and health: quasi-experimental evaluation of a natural experiment. Journal of Epidemiological Community Health 59, 1035–1040.

Cliff, Guy, Clarke, Graham, Eyre, Heather, 2004. Food retail change and the growth of food deserts: a case study of Cardiff. International Journal of Retail & Distribution Management 32 (2), 72–88.

Diesenhouse, Susan, June 1993. As suburbs slow, supermarkets return to the cities. The New York Times 27, F5.

Drewnowski, Adam, Specter, S.E., 2004. Poverty and obesity: the role of energy density and energy costs. American Journal of Clinical Nutrition 79 (1), 6–16.

Drewnowski, Adam, Aggarwal, Anju, Moudon, Anne Vemez, 2010. The Supermarket Gap: How to Ensure Equitable Access to Affordable, Healthy Foods. A publication of the University of Washington Center of Public Health Nutrition. http://depts.washington.edu/uwcphn/pubs/reports.shtml (retrieved 19.12.).

Food Deserts, 2013. http://apps.ams.usda.gov/fooddeserts/foodDeserts.aspx (retrieved 29.11.13.).

FAO, IFAD, and WFP, 2013. The State of Food Insecurity in the World 2013. The Multiple Dimensions of Food Security. FAO, Rome.

Flournoy, Rebecca, 2006. Healthy Foods, Strong Communities. National Housing Institute. http://www.nhi.org/online/issues/147/healthyfoods.html (retrieved on 29.11.13.).

Food and Agriculture Organization of the United Nations, 2011. The State of Food and Agriculture 2011. Women in Agriculture: Closing the Gender Gap for Development. FAO, Rome.

Furey, Sinéad, Strugnell, Christopher, McIlveen, Heather, 2001. An investigation of the potential existence of "food deserts" in rural and urban areas of Northern Ireland. Agriculture and Human Values 18, 447–457.

Grocery Store Development, 2012. http://www.policylink.org/site/c.lklXLbMNJrE/b.7962289/k.8730/Grocery_Store_Development/apps/nl/newsletter2.asp (retrieved 30.11.13.).

Guy, Cliff, Clarke, Graham, Heather, Eyre, 2004. Food retail change and the growth of food deserts: a case study of Cardiff. International Journal of Retail & Distribution Management 32 (2), 72–88.

Hunger Facts, 2012. http://feedingamerica.org/hunger-in- america/hunger-facts.aspx (retrieved 29.11.13.).

Hunger & Poverty Statistics, 2012. http://feedingamerica.org/hunger-in-america/hunger-facts/hunger-and-poverty-statistics.aspx (retrieved 29.11.13.).

Hendrickson, Deja, Smith, Chery, Eikenberry, Nicole, 2006. "Fruit and vegetable access in four low-income food deserts communities in Minnesota. Agriculture and Human Values 23, 371–383.

Holben, David, 2010. Position of the american dietetic association: food insecurity and hunger in the United States. Journal of the American Dietetic Association 110 (9), 1368–1377.

Hurvitz, Philip M., Moudon, Anne V., Aggarwal, Anju, Drewnowski, Adam, 2010. Proximity to food stores may not predict use. The Journal of the Federation of American Societies for Experimental Biology lb331.

Lewis, LaVonna Blair, Sloane, David C., Nascimento, Lori Miller, Diamant, Allison L., Guinyard, Joyce Jones, Yancey, Antronette K., Flynn, Gwendolyn, 2005. African Americans' access to healthy food options in South Los Angeles restaurants. American Journal of Public Health 95 (4), 668–673.

Millennium Development Goals, 2013. Progress Chart. http://www.un.org/millenniumgoals/pdf/report-2013/2013_progress_english.pdf (retrieved 01.05.13.).

Miller, Cyndee, 1994. Rediscovering the inner city. Marketing News 28 (2), 1.

Morland, Kimberly, Wing, Steve, Roux, Ana Diez, 2002. The contextual effect of the local food environment on residents' diets: the atherosclerosis risk in communities study. American Journal of Public Health 92 (11), 1761–1767.

Morton, Lois, Blanchard, Troy C., 2007. Starved for access: life in rural America's food deserts. Rural Realities 1 (4), 1–10.

Nyden, Philip, Lukehart, John, Maly, Michael T., Peterman, William, 1998. Chapter 1: neighborhood racial and ethnic diversity in U.S. cities. Cityscape: A Journal of Policy Development and Research 4 (2), 1–17.

Position of the American Dietetic Association, 2006. Food insecurity and hunger in the United States. Journal of the American Dietetic Association 106 (3), 446–458.

Powell, Lisa M., Auld, M. Christopher, Chaloupka, Frank J., O'Malley, Patrick M., Johnston, Lloyd D., 2007a. Access to fast food and food prices: the relationship with fruit and vegetable consumption and overweight status among adolescents. Advanced Health and Economic Health Services Research 17, 23–48.

Powell, Lisa M., Auld, M. Christopher, Chaloupka, Frank J., O'Malley, Patrick M., Johnston, Lloyd D., 2007b. Associations between access to food stores and adolescent body mass index. American Journal of Preventive Medicine 33 (4 Suppl.), S301–S307.

Reisig, Vmt, Hobbiss, A., 2000. Food deserts and how to tackle them: a study of one city's approach. Health Education Journal 59, 137–149.

Rose-Jacobs, Ruth, Black, Maureen M., Casey, Patrick H., Cook, John T., Cutts, Diana B., Chilton, Mariana, Heeren, Timothy, Levenson, Suzette, Meyers, Alan F., Frank, Deborah A., 2008. Household food insecurity: associations with at-risk infant and toddler development. Pediatrics 121 (1), 65–72.

Shaffer, Amanda, 2002. The Persistence of L.A.'s Grocery Gap: The Need for a New Food Policy and Approach to Market Development." UEP Faculty & UEPI Staff Scholarship. http://scholar.oxy.edu/uep_faculty/16 (retrieved on 30.11.13.).

Shaw, Hillary J., 2006. Food deserts: towards the development of a classification. Geografiska Annaler: Series B, Human Geography 88 (2), 231–247.

Skalicky, Anne, Meyers, Alan F., Adams, William G., Yang, Zhaoyan, Cook, John T., Frank, Deborah A., 2006. Child food insecurity and iron deficiency anemia in low-income infants and toddlers in the United States. Maternal and Child Health Journal 10 (2), 177–185.

Sparks, Andrea, Bania, Neil, Leete, Laura, 2010. Comparative approaches to measuring food access in urban areas: the case of Portland, Oregon. Urban Studies 48, 1715–1737.

Stuff, Janice E., Casey, Patrick H., Szeto, Kitty L., Gossett, Jeffrey M., Robbins, James M., Simpson, Pippa M., Connell, Carol, Bogle, Margaret L., 2004. Household food insecurity is associated with adult health status. Journal of Nutrition 134 (9), 2330–2335.

Swinburn, Boyd, Caterson, Ian, Seidell, Japp, James, W.P.T., 2004. Diet, nutrition, and the prevention of excess weight gain and obesity. Public Health Nutrition 7 (1A), 123–146.

United States Department of Agriculture, 2009. Access to Affordable and Nutritious Food Measuring and Understanding Food Deserts and Their Consequences: Report to Congress. U.S. Government Printing Office, Washington, DC.

United States Department of Agriculture, 2012. Definitions of Food Security. http://www.ers.usda.gov/topics/food-nutrition-assistance/food-security-in-the-us.aspx#.UpjXkCj_7-A (retrieved on 29.11.13.).

United States Department of Agriculture, 2013. The Role of Food Hubs in Local Food Marketing. http://www.ams.usda.gov/AMSv1.0/getfile?dDocName=STELPRDC5102827 (retrieved 10.12.13.).

Walker, Renee E., Keane, Christopher R., Burke, Jessica G., 2010. Disparities and access to healthy food in the United States: a review of food deserts literature. Health & Place 16, 876–884.

Wienk, Ronald E., 1979. Measuring Racial Discrimination in American Housing Markets: The Housing Market Practices Survey. U.S. Department of Housing and Urban Development, Washington, DC.

Wrigley, Neil, Warm, Daniel, Margetts, Barrie, 2002. Assessing the impact of improved retail access on diet in a 'food desert': a preliminary report. Urban Studies 39, 2061–2082.

Wrigley, Neil, Warm, Daniel, Margetts, Barrie, 2003. Deprivation, diet and food retail access: findings from the leeds 'food deserts' study. Environment and Planning a 35, 151–188.

Zenk, Shannon N., Schulz, Amy J., Israel, Barbara A., James, Sherman A., Bao, Shuming, Wilson, Mark L., 2005. Neighborhood racial composition, neighborhood poverty, and the spatial accessibility of supermarkets in Metropolitan detroit. American Journal of Public Health 95 (4), 660–667.

Ziliak, James P., Gundersen, Craig C., 2013. Spotlight on Food Insecurity Among "Senior Americans: 2011.". National Foundation to End Senior Hunger (NFESH) Report.

Food in Anthropology

Steffan I Ayora-Diaz, Universidad Autónoma de Yucatán, Mérida, Yucatán, Mexico

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C.M. Counihan, volume 8, pp. 5715–5719, © 2001, Elsevier Ltd.

Abstract

The anthropology of food has adopted a broad range of methodological and theoretical approaches to the study of diets, food habits, and culinary practices. Traditionally, it has examined the social, economic, and cultural differences that explain inequalities in access to food, and the problems derived from these differences. In recent years, anthropologists have begun to examine the origins and consequences of emerging culinary and gastronomic traditions linked to various forms of identity: gender, ethnic, local, regional, and national. Against this complexity and fluidity, anthropologists of food are engaging new theories to better understand contemporary problems.

Generally speaking, the anthropology of food is a broad and growing field. As the chapters in Counihan and Van Esterik (2013) illustrate, it encompasses a diversity of methodological and theoretical approaches covering different and, sometimes, opposed understandings and explanations of food and related phenomena, from multi- and interdisciplinary perspectives. Some studies privilege the biological, medical, and chemical definition of foods and taste; others privilege the cultural, symbolic, and interpretive analysis of food preferences; and still others take the preference for certain foods as a 'natural' disposition, while others focus on the political structures and relations that define who cooks and who eats what, and what is edible or unpalatable where. Also, political analyses of food and cooking have been conducted from Marxist and related political economy viewpoints, as well as from the perspective of new social movements analysis and postcolonial theory. Across most anthropological approaches to food, gender, and ethnicity are often considered important dimensions. In the past, food issues were often studied as localized phenomena, but now food and cooking have begun to be approached from broadly regional or global perspectives. In particular, the analysis of culinary practices and preferences, and the emergence of formalized cuisines, have directed the anthropologists' attention toward the relationship between food and identity from the angle of ethnic, local, regional, and global-cosmopolitan affiliations.

Prevailing Ethnographic and Anthropological Approaches

During the last two decades, the anthropology of food, and especially of food cultures and culinary practices, has come of age. In the early stages of the discipline, anthropologists looked at food as a minor feature, even if present in many dimensions of human life. For example, some anthropologists examined the part food played in religious rituals or in age or gender hierarchies in simpler, small societies. Others saw it as a product crafted from resources obtained from the immediate environment. Still others discussed food as the unprocessed product of specific environmental restrictions that shaped the ways in which a group lived.

The ethnographic record is rich in cases that illustrate how food can be used to define an array of social ties. Food offerings redefine the relationship between humans, transforming them in relatives or friends – or denying them recognition as equal human beings. Sharing food and drinks often has strong cultural meaning and symbolic significance. The anthropology of hospitality practices, for example, has demonstrated how the positions of guest and host are defined by the various manners in which food and drinks are exchanged. In general, food and drinks are ways to include the stranger into the host's society and to recreate reciprocity obligations (Ayora-Diaz, 2000). The meaning and logic of food exchange, in turn, are transferred to the domestic space. The contemporary home is a compartmentalized space; the rooms the guests are allowed to enter, along with the types of food and drinks they are offered, express the level of their affective integration into the space of the family. Food can also help maintain the affective relations between the living and the dead, as when the relatives of deceased persons set a table of his or her favorite food and drinks during the Days of the Dead in Mexico. In many places in Asia, Africa, the Mediterranean, and Latin America, food gifts maintain kinship ties among the living, and between the living and other spiritual beings, including various deities. In fact, in some cases, interrupting food offerings to deities can be seen as leading to illness and misfortune, as Laurel Kendall (1985) shows for Korean families. When families fail to perform their periodical food offerings of rice cakes to their domestic spirits they are punished with illness and other misfortunes.

Since Audrey Richards's (1939) early study on the agricultural and domestic organization of food practices and its impact on the diet of a group, anthropologists have often looked at food as a result of local production, circulation, and consumption practices; or to the different political, economic, and sociocultural reasons why food crops sometimes become an abundant or, sometimes, a scarce nutritional resource (think of the 'fast-and-feast' characterization of nomadic hunters). Against the pressing need to focus on, and find solutions to, such life-and-death issues, it is no wonder that culinary practices and the emergence of cuisines had a difficult time in gaining prominence in anthropological work.

Early in the second half of the twentieth century, Claude Levi-Strauss (1970 [1968], p. 428) suggested an ingenious approach to food codes through his now well-known 'culinary triangle.' The structure of this triangle shows at the apex the term 'raw,' which is associated with roasted; at the base of the triangle one finds on the one hand the term 'cooked' associated with smoked and, on the other, the term 'rotted' associated with boiled. Shortly before that, Mary Douglas (1984 [1966]) had examined the social meaning of food and of the organization of the meal in British society. They both paved the road for the sociocultural study of food-related practices, but, generally speaking, anthropologists were slow to take up the challenge. The recognition that the availability or shortage of foodstuffs was tied to broad social inequalities and to unequal access to the market, and that the objective quality of food is biomedically and biochemically defined, led anthropological research to study the effects of poverty on the human diet, and to medically and nutritionally defined problems, focusing on phenomena such as obesity, alcoholism, diabetes, and undernourishment – some of them defined as diseases of advanced societies, others of impoverished regions of the world. In some studies, culture was defined as the cause of ill-nutritional practices, while in others the explanation for food preferences and avoidances was found in the biological needs of human beings who would then choose their food based on rational choices (de Garine and Pollock, 1995).

Anthropologists found that the unequal access to food was tied to structural imbalances. Gender difference became an important explanatory variable in this regard (Counihan and Kaplan, 1998). The focus on gender, food, and cooking was then inevitable: With a few exceptions, in most societies anthropologists found a tendency toward the subordination of women to men, which led to the confinement of women to everyday chores, including cooking. Although cooking is necessary for the biological reproduction of the species, women's cooking, even in the so-called 'advanced' societies, was trivialized as an everyday chore performed within the private domain of the kitchen. Men, on the other hand, directed and controlled the cooking and display of food in public celebrations. In North Atlantic cultures, the professionalization of cooking became a male-dominated field, and women had to learn from male chefs the 'correct' cooking procedures. These inequalities, which led anthropologists to see food in its more public roles as an economic and symbolic good to be exchanged in public rituals meant to cement broad social ties, is now compensated by explorations of the oral and written strategies displayed by women to transmit cooking knowledge from one generation to another, by the analysis of women's sociality during cooking events, and by studies of breast feeding, diets, and of women's role as the care givers of the family. In any case, anthropologists now see the explanatory power of gender as necessarily entwined with other dimensions; looking at the convergence of gender and poverty one can better understand the sociocultural reasons why the economic organization of food production and distribution takes certain shapes. Relating gender to food can help us understand why and how in some societies women may be excluded from certain productive activities, while in others they are assigned the responsibility over their production and circulation.

Gendered food practices can become tightly intertwined with ethnicity. Thus, for example, in central Mexico, women become responsible for both nurturing the family and for maintaining the cultural ties between central Mexican indigenous groups and their past. In some groups, women are assigned the task to watch over the maintenance of religious values and practices, and food plays a very important symbolic part, sometimes as offerings to gods, saints, or ancestors and sometimes as the vehicle of spiritual powers, but in every case as a means to found gender roles, to cement family ties, and to preserve the social status quo that give ethnicity its naturalized qualities. Thus, in contemporary anthropology, the analysis of gender and food becomes inseparable from the study and analysis of power relations among ethnic groups. However, even beyond gender, food preferences can become tied, for different reasons, to ethnic identities. A common practice to draw the boundary between a group and another lies in invidious differences of what each groups considers edible, or in the foods they prefer. Mock names used to refer to members of a different society often tend to ridicule or to signal their group preferences as contemptible: 'Frogs' and 'Krauts' are broadly recognizable as terms intending to ridicule French and German people. In Mexico, outside Mexico City, 'chilangada' is often used to despise central Mexican food preferences and cooking. The author has also found, in Chiapas, Mexico, that Mayans who speak, for example, Tzeltal, sometimes equate the food of the Tzotzil-speaking people with animal feed and claim they have never eaten some foods favored by the Tzotzil-speaking Mayas. Also, anthropologists have also looked at the ways in which foods perceived as 'ethnic' may become acceptable to members of dominant society (Gabaccia, 1998). For example, Chinese, Italian, and Mexican foods were tied in the US imagination to low-class immigrants and took some time in becoming an everyday choice in most cities and towns of the country.

Anthropologists have also taken an interest in the ways in which food can be a vehicle or an instrument in religious practices and discourses, or both. From Mary Douglas's *Purity and Danger* (1984 [1966]) to more recent work, anthropologists of different persuasions have debated the meaning of food taboos and preferences as informed by religious worldviews. Different religions have their own understanding of ritual cleanliness and pollution and their choice of food is frequently informed by those religious definitions of the edible and inedible. In places where several religions coexist, anthropologists have taken advantage of this plurality of worldviews to look at the intersection of religion and, according to the case, more or less rigid hierarchical social structures, examining some of the ways in which food articulates discourses and exclusionary practices. A related approach has focused on popular beliefs and discourses, legends, and puns that highlight the cultural importance of foodstuffs in articulating religious differences, local identities, gender relations, and various mystical/mythical attributions, as seen in the work of Fabre-Vassa (1997 [1994]). She describes and discusses the stories surrounding the Jews and their imagined relations with pigs in French rural society, the ways in which the pig became a marker of cultural difference and distance, and how it turned into a justification for the social discrimination and exclusion of Jews in Catholic France.

Globalization and the Anthropology of Food

The development of critical and reflexive anthropology led anthropologists to understand local practices and identities as tied to broader processes. Regarding food, globalization theories and analyses helped to take into account different levels of complexity that went simply unnoticed in previous work. The seminal work by Sidney Mintz (1985), *Sweetness and Power*, demonstrated the necessity for anthropologists to take a deeper historical perspective to understand contemporary food production and consumption practices. These are practices and processes that irremediably tie different places of the world to, and different social classes within, any single society. Mintz unveiled the historical reasons that transformed sugar from a luxury commodity proper of the aristocracy into an everyday good of the working classes in England, while at the same time disclosing the shifting power position of different European countries and the sugar-producing Caribbean region over time. His work inspired further research into the differential inscription of edible commodities in the global market. Nonetheless, other theories have also informed anthropologists' study of local–global interactions. For example, Gabriela Vargas-Cetina (2000) has described and analyzed how a local Sardinian product, *pecorino* cheese, was the result of the appropriation of nonlocal technologies and techniques, and the process whereby it became a valued global commodity consumed elsewhere, from the times of the Roman empire into the present. Other products have had rather uncertain fates. Some agricultural products traveled from one place of the world to another, first despised by the mass and elite consumers of Europe, but later they became accepted, sometimes as essential ingredients of national foods. In the opposite direction, colonizers despised local foods and enforced the consumption of imported European and Asian products in the American continent. It is known that in Europe tomatoes were initially thought to be toxic, but later they were integrated into European high cuisines. Chili peppers originated in the American continent, but they are today essential ingredients of European, Asian, African, and other cuisines. In some societies, a product may be valued by a segment of the population, while another social segment despises it as animal fodder – or food for the poor.

In recent times, industrially, mass-produced goods have become prime examples of the homogenizing process of globalization. However, true to the social sciences' recent interest in the processes through which the global becomes localized and the local globalized, many studies have shown the diversity and heterogeneity of outcomes resulting from the local appropriation, adaptation, and resignification of global commodities. Robert Foster (2008), for example, has followed one of the paradigmatic industrial manufactured goods, cola drinks, in their commercial displacements and described the dissimilar forms of appropriation and resignification that this commodity undergoes in its travel from the United States to Papua New Guinea. Foster traces the material and symbolic transformations of these drinks, the images that become associated with it during marketing campaigns (e.g., modernity and youth), and the different understanding that people have of the same images both in the United States and in Papua New Guinea. His study illustrates how local adaptations and appropriations explain contrasting codes that signify the use of a product and its consumers' self-representations.

In the contemporary context of globalization, not only single commodities, such as ingredients, travel. Through a long history of intercultural and international contacts, different food traditions have been adapted to the new environment and food markets in which large numbers of migrants found themselves. Economic migrants, diasporic subjects, political refugees, and exiles develop different relations to the food of the homeland. Despite fears of global homogenization, the anthropological literature, perhaps because of its privileged focus on local practices and meanings, has continuously shown that foods are regularly appropriated and resignified, even if this local redefinition leads to attempts to defend its purity or authenticity. Indentured workers, slaves, economic migrants, refugees, exiles, and diaspora subjects, despite the suffering involved in their displacement, or sometimes because of it, seek to replicate and transfer their food taste and preferences to their new homes. So, they adapt and transform their homeland food traditions to the ingredients and environment they moved into. Often, cooking is an occupation that can help a group to physically survive. At the same time, migrants replicate and adapt the foods of their homelands and globalize their local cuisines. Chinese, Indian, Japanese, Korean, Thai, Vietnamese, Finnish, German, Italian, Spanish, Argentinean, Brazilian, and Mexican (to name those numerically important groups of migrants) carry and adapt their culinary traditions to new environments. Less concerned with 'authenticity,' these groups seek to negotiate the meaning and importance of their culture in different host societies. They are more willing to transform their foods both to satisfy the migrants' nostalgia for the homeland, and the host's acceptance of their food ways as manifestation of cultural 'authenticity.' Migrant foods and populations are often constrained to adapt the food to the ingredients they find in their new homes, and to the taste preference of the members of the host societies who also become consumers of imported foods. So, not only the social demarcation of identity, but also the necessarily shifting reproduction of home cultures from elsewhere, is expressed through food and dietary adaptations.

Another aspect of the globalization of food relates to the expansion of North Atlantic food businesses. Hamburgers, hotdogs, and other fast food sometimes become associated with images of youth and freedom, and temporarily displace local foods in the preference of Asian, European, African, and Latin American consumers. Anthropologists have looked at how these newly introduced food products can contribute to changes in local food habits and dislocating age and gender hierarchies, as when McDonald's in Asia helped reverse the symbolic place of children in Asian families, which were increasingly focused on fulfilling the every wish of small infants (Watson, 1997).

Very few anthropologists have looked into the globalization of European cuisines, leaving the field mainly to historians and scholars from cultural studies. Instead, they have chosen to study the ways in which cooking may become recognized as a particular set of local practices. Trubeck (2008), for example, looked at French cuisine from a historical angle, and showed the influence of chefs and restaurants in establishing French

gastronomy as the universal gold standard to measure all other cuisines. She documents how, after the French Revolution, chefs lost their place in the homes of the aristocracy and opened restaurants in France and in many other countries, and on their own authority established the principles of French cooking as universally applicable rules. Today, one can see that Italian and Spanish traditions have followed suit, putting efforts into establishing theirs as national 'high' cuisines. Similarly, as cultural brokers from different local societies, cooks seek to reposition their ethnic or regional foods as equal to established high cuisines.

Anthropology, Modernity, and Food

Modernity is today, like 'tradition' was in the past, a primary focus of anthropological studies. No longer constrained to study local, bounded communities, anthropologists are looking at global–local and translocal relations, and at the cultural consequences of late capitalism, in particular, to the way in which emerging cultural formations – be they called 'modernity' or 'postmodernity' – create new meanings for the different types of food that we consume, the culinary practices that produce 'authentic' and hybrid cuisines, and the ways in which food becomes tied to fluid, changing social structures.

Contemporary tourism has played a significant role in creating a demand for authentic, local cuisines. When migrants moved into new countries, they produced meals for themselves, to satisfy their longing for the homeland. However, gradually, consumers from the host society became acquainted with immigrants' culinary traditions and accepted local adaptation to their foods. As travel became more accessible for the consumers from affluent societies, so grew the desire for more 'authentic' renditions of ethnic and national foods. Many travelers favored regional foods as mementos to take back home from a trip. However, as several authors in Lucy Long's volume (2004) suggest, tourists and foods providers have been forced to negotiate their understanding of authenticity, and cooks are often consciously seeking to satisfy both the tourists' and their own imagination of what they perceive to be the longing for exoticism attracting visitors in tourist destinations. Both within the United States and elsewhere in the world, local people recognize the variations in taste displayed by tourists and sometimes they modify their foods to fit the visitors palate, or like in Sardinia, shepherds stage 'traditionally authentic' cooking in the desolate landscape of the mountain plateaus of the island, seeking to satisfy the ecotourist imagination of those visiting the region.

Still, in the contemporary world, the search for the authentic and the pure is not restricted to the tourist imagination. Distrustful of the food fed by corporations, by the growing perception of risk associated with epizootic crises such as the aviary flu, bovine spongiform encephalopathy, food and mouth disease, and the growing production of genetically modified organisms (GMOs) and transgenic foods, as well as the bacterial contamination of vegetables and other produce, consumers are developing a strong resistance to industrially processed and packaged foodstuffs (Lien and Nerlich, 2004). In response, different movements have emerged that take a philosophically and morally critical view of the dominant contemporary food-provisioning system, and demand for alternatives. The Slow Food, and the 'Going Local' movements are models for a range of responses articulated by consumers who increasingly and in growing numbers come to value locally produced foodstuffs, organic food, free-range animals, and food free of transgenic manipulation and GMOs, valuing 'natural' over industrially produced, 'artificial' foods. Today, many consumers attribute their favored foods a better taste, a higher cultural value related to artisanal production, and a heightened purity and authenticity deriving from organic agriculture. Although many of these movements originated and gained strength in Europe and North America, the values and worldview they uphold are now changing the perception of food, and its political–economic and moral values, everywhere else. Hence, consumers recognize a higher value in wine, cheeses, vegetables, and animal meats for which producers can guarantee and certify as local, organic, free-range, artisanal, free of chemicals, antibiotics, and hormones, that is, free of artificial manipulation. These values have driven in recent years the commercial success of organic food supermarkets in North America and Europe, and contributed to the resurgence of farmers' markets everywhere.

National and Regional Cuisines and Identities

As anthropologists have focused their attention on the relationship between food and identity, it has become increasingly evident that just as gender, religious, and ethnic identities are entwined with the meaning of foods, and culinary practices, so are national and regional identities. Since rice, corn, and wheat constitute the basic, everyday ingredient in the diet of different groups, they tend to be equated with life itself and are often attributed great religious, social, political, and cultural significance. The culinary preference for one among the different varieties of rice grown in different Asian countries is sufficient to found national and moral differences between one group and another. In Mesoamerica, corn has been taken as the life-giving ingredient. One of the most religiously important texts, the *Popol Vuh*, defines corn as the substance from which humans were made during the last creation of the world. In Europe and the Middle East, wheat receives religious, mystical, symbolic significance, and different stories affirm its divine origin, sometimes even granting a higher status to the people who consume them. Anthropologists have argued, on these grounds, that these food staples become intimately tied to the group's cultural notion of self in these societies and constitute the symbolic core of their ethnic identities. This notion of ethnicity has been extended to argue, in some instances, that a national cuisine, such as the Mexican, is tied to a single ethnic indigenous society encompassed by the nation-state. Cultural elites in Mexico have promoted the view that there is a single Mexican national cuisine, which is based on pre-Colombian ingredients such as corn, beans, tomatoes, squash, and chili peppers (Ayora-Diaz, 2012a).

The relationship between food and ethnic, regional, and national identities is a very complex one. One artifact of the invention of the modern nation-state is the need to visualize the coextensiveness of the political and cultural forms. With the exception of a few societies, most political regions that emerged

as nation-states encompassed a diversity of groups that had developed their own cultural practices, including those pertaining to the realm of the kitchen. England, France, Germany, Italy, and Spain, countries in which the modern nation-state was conceived, encompassed different regions where different produce and different cooking styles had developed. The emergence of national cuisines partially concealed, sometimes on purpose, the regional diversity of each country. It is only by looking at the history of a nation-state that one can understand, anthropologically and ethnographically, the emotions and identities associated sometimes with a single ingredient, and sometimes with full-fledged gastronomic traditions.

Arjun Appadurai's (1988) groundbreaking paper led anthropologists' attention to the study of the ways in which a cuisine could be used to stand for the nation. Appadurai explains the growth of the printing industry of cookbooks of Indian cuisine in the context of the colonial relations between India and England, and later the process of decolonization. Despite the great diversity of worldviews and related food preferences and proscriptions, members of the Indian upper castes, at home and abroad, collected recipes into cookbooks, creating the illusion of a national cuisine. Thus, anthropologists have been studying how, from a minor literary genre, cookbooks around the world were turned into the foundations of gender, ethnic, regional, and national identities.

The preference for ingredients, culinary techniques, and technologies are generally based on discursive references to authenticity, but anthropologists increasingly understand national cuisines as modern inventions. The creators of national cuisines have been forced to negotiate between local and imported culinary practices, and most national cuisines, as those exemplified by anthropological studies on French, Japanese, Chinese, Italian, Belizean, Brazilian, and Mexican cuisines, are the product of local–global connections that have stimulated the appropriation and adaptation of ingredients and culinary techniques to the local culinary aesthetic code. People and their cuisines travel, and in recent decades gastronomic tourism has become an important stimulus in the invention and affirmation of national cuisines, as service providers vie for a position of privilege within the market of ethnic and national foods. In the cases listed, these adaptations have led to the creation of something different from the cuisines of origin, and attributed to the ingenuity of national cooks. In all these cases, anthropologists have shown that cooks/chefs, restaurants, and cookbooks have played an important part in establishing national gastronomic traditions.

Regional cuisines, on the other hand, have received a different treatment. In some cases, local pride and identity have been described as tied to a single product of the land or water: garlic, camembert cheese, wine grapes, champagne, caviar, and crayfish, to mention just a few, are taken as markers of the land and the people who produce them, and as foundations of a local or regional identity. In other cases, in Europe, Latin America, and the United States, anthropologists have taken different regions within a country as characterized and circumscribed by a particular bioecological environment and inhabited by people with distinct cultural backgrounds. That is, some regions have become the chosen territory for descendants of African slaves, of indigenous people, mestizos, or Italians, French, Spaniards, Germans, Russians, Lithuanians, Chinese, Japanese, Koreans, Algerians, Iranians, Indonesians, etc., and understand their food preferences as supported by a strong regional identity (Counihan and Kaplan, 1998; Counihan and Van Esterik, 2013; Long, 2009; Pazos Barrera, 2010). However, in other instances, local groups with a history of secessionism or with strong regional identities, use food as an instrument to distinguish their taste, moral worldview, and social status vis-à-vis strong central governments (as it is the case, for example, of the Basques and Catalans in Spain, of Sardinians in Italy, Corsicans in France, and Yucatecans in Mexico) (see Ayora-Diaz, 2012a, b). These cases allow us to see how, in its increasing diversity, the anthropology of food is finally becoming a major subfield of cultural anthropology.

Future Directions in the Anthropology of Food

The diversity of topics and problems encompassed by the anthropology of food provide a wide field in which anthropological theories and concepts, and field methods can be tried to enrich our understanding of the contemporary world. Since multiple aspects of everyday life, in rural and urban societies, will continue to inspire anthropological studies, practitioners of the field can broaden their multidisciplinary and interdisciplinary approaches to the study of changing diets, food habits, food-related bodily images, gender and structural inequalities, the effects of other organisms on the health of groups and individuals, and the advantages and nutritional properties of a vast array of foodstuffs that contemporary agricultural and industrial production practices allow.

Anthropologists are also finding a wealth of issues to study and analyze regarding the different social movements, their ideologies, moral understandings, political, and economic motivations around issues such as GMOs and transgenic foods in the contemporary diet. Also, worldwide growing movements such as Slow Food, 'Going Local,' permaculture, and City Gardens deserve closer ethnographic attention. Food corporations compete with these movements for the consumers' favor and both groups unfold commercial, political, and moral practices and discourses that can be productively analyzed from an anthropological perspective.

Also, the continuous displacement of groups of people from one place of the world to another, forced by war, political, economic, religious, and other reasons, generates multicultural contexts in which members from different cultures must negotiate the meaning of their own culinary practices. Tourism, another form of mass displacement of people, fosters the emergence of ethnic and national cuisines that at the same time procure a niche in the global market of edible commodities and constitute the basis for affirming local, ethnic, regional, and national identities. In all these cases, the contemporary world imposes on anthropology the need to take into account the blurring of cultural boundaries; the fluidity and friction in the movements of people, ideas, commodities, and culinary codes; the instability of contemporary social and cultural formations; and the reactive attempts to affirm local against foreign foods. These different issues demand the creative appropriations of new concepts and theories, and post-structural and postcolonial theories could help explain many of

these transformations as well as emerging food and culinary practices and ideologies. These theoretical approaches are beginning to supplement other, well-established ones, such as political economy theory, the different types of structuralism, cultural relativism and neopositivism. It is possible to say that in the near future, food anthropologists will continue to engage in different manners with the complexity of today's contemporary culinary and dietary issues, and will develop increasing ties with other already established disciplines surrounding food problems.

See also: Anthropology and History; Authenticity, Anthropology of; Body: Anthropological Aspects; Creolization: Sociocultural Aspects; Cultural Critique, Anthropological; Cultural Heritage; Diaspora; Drinking, Anthropology of; Friendship, Anthropology of; Gender: Anthropological Aspects; Globalization, Anthropology of; Hybridity; Identity in Anthropology; Inequality: Comparative Aspects; Migration: Anthropological Perspectives; Modernity: Anthropological Aspects; National Character; Nationalism: General; Nostalgia, Anthropology of; Personhood, Anthropology of; Postcoloniality; Sociality In Anthropology.

Bibliography

Appadurai, A., 1988. How to make a national cuisine: cookbooks in contemporary India. Contemporary Studies in Society and History 30, 3–24.

Ayora-Diaz, S.I., 2000. Hospitality in Sardinia: the moral Construction of identities. Europaea VI, 229–257.

Ayora-Diaz, S.I., 2012a. Foodscapes, Foodfields and Identities in Yucatán. Berghahn, New York.

Ayora-Diaz, S.I., 2012b. The naturalization of Yucatecan taste: gastronomic inventions and the aesthetics of regional food. Etnofoor 24 (2), 57–76 (special issue on "taste").

Counihan, C., Van Esterik, P. (Eds.), 2013. Food and Culture. A Reader, third ed. Routledge, London.

Counihan, C.M., Kaplan, S.L. (Eds.), 1998. Food and Gender. Identity and Power. Hardwood, Amsterdam.

de Garine, I., Pollock, N.J. (Eds.), 1995. Social Aspects of Obesity. Taylor and Francis, London.

Douglas, M., 1984 [1966]. Purity and Danger. An Analysis of the Concepts of Pollution and Taboo. ARK, London.

Fabre-Vassa, C., 1997 [1994]. The Singular Beast. Jews, Christians, and the Pig. Columbia University Press, New York.

Foster, R.J., 2008. Coca-Globalization: Following Soft Drinks from New York to New Guinea. Palgrave, New York.

Gabaccia, D., 1998. We Are What We Eat. Ethnic Foods and the Making of Americans. Harvard University Press, Cambridge.

Kendall, L., 1985. Shamans, Housewives, and Other Restless Spirits. Women in Korean Ritual Life. University of Hawaii Press, Honolulu.

Levi-Strauss, C., 1970 [1968]. Mitológicas III: El origen de las maneras de mesa. Siglo XXI, Mexico City.

Lien, M.E., Nerlich, B. (Eds.), 2004. The Politics of Food. Berg, Oxford.

Long, L. (Ed.), 2004. Culinary Tourism. The University Press of Kentucky, Kentucky.

Long, L.M. (Ed.), 2009. Regional American Food Culture. Greenwood Press, Santa Barbara.

Mintz, S.W., 1985. Sweetness and Power. The Place of Sugar in Modern History. Penguin Books, New York.

Pazos Barrera, J. (Ed.), 2010. Cocinas regionales andinas. Corporación Editora Nacional, Quito.

Richards, A.I., 1995 [1939]. Land, Labour and Diet in Northern Rhodesia. An Economic Study of the Bemba Tribe. LIT Verlag, Münster–Hamburg.

Trubeck, A.B., 2008. The Taste of Place. A Cultural Journey into Terroir. University of California Press, Berkeley.

Vargas-Cetina, G., 2000. From handicraft to monocrop: the production of pecorino cheese in highland Sardinia. In: Haugerud, A., Stone, M.P., Little, P. (Eds.), Commodities and Globalization. Anthropological Perspectives. Rowman and Littlefield, Lanham, pp. 219–238.

Watson, J.L., 1997. McDonald's in Hong Kong: consumerism, dietary change, and the rise of a children's culture. In: Watson, J.L. (Ed.), Golden Arches East: McDonald's in Asia. Stanford University Press, Stanford, pp. 77–109.

Food Preferences, Psychology and Physiology of

Paul Rozin, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Human food choice is influenced by biological, psychological, and cultural factors. Because humans are food generalists, and because the food environment in the first world is so different from the ancestral human environment, innate predispositions play a minimal role. Humans principally learn to eat and like certain foods, and avoid others, with traditions being the main influence. Preferences can be based on the instrumental consequences of eating (e.g., nutritive or toxic properties of foods) or on liking for the taste. Mere exposure, pairing of foods with positive or negative consequences, or a variety of social influences foster liking or disliking. In the extreme, these and other forces turn an innate dislike, as for chili pepper, into a strong liking. Surprisingly, family influences on food preferences are very modest. Food functions in many ways, for example, as a social or moral entity; food preferences are therefore influenced by things such as the nature or origin of a food, and moral/religious attitudes, such as concerns about killing animals. In the developed world, with a surplus amount of foods, a wide variety of foods, and the conquest of most infectious diseases, concerns about food have focused more on obesity and the long-term consequences of diet. As a result, the pleasures of eating, especially in the United States, have become tempered by a fear of fat and food. The situation and person are both major determinants of choice, including the social setting, the time of day (e.g., breakfast vs dinner), eating at home or away from home, and the stage of life, from infancy to old age.

Introduction

Food and food choice constitute a major part of waking human activity, and, in the developing world, the major expense. Furthermore, food choice may be the single most important selective force in animal evolution, as suggested by the fact that many animal groups are named for their food habits (e.g., carnivores, insectivores). In human evolution, adaptations to obtaining large animal foods played a significant role in the evolution of the human brain and human social organization. Early advances in food procurement, agriculture and domestication, set the stage for major advances in technology, and the development of urban centers. Food has served as a foundation for the cultural evolution of nonfood systems, such as religion, ritual, and social exchange, through the cultural evolutionary process of preadaptation. Through many elaborations and the development of distinctive cuisines, food has become an important component of group identity, and a major source of pleasure, which sometimes can be classified as aesthetic. Thus, an understanding of human relationships to food promises to both enlighten a major area of human concern, and to enlighten other areas. The elaboration of eating into a refined and civilized activity in Western culture is beautifully described in Leon Kass' (1994) book, *The Hungry Soul*.

The Basic Relationship between Humans and Their Food

For any person, the world can be divided into the self and everything else. Eating involves taking matter from outside the self and putting it inside the self. This is a very intimate act. It is not surprising that people feel strongly about what they eat. The costs in terms of toxins, microorganisms, or imbalanced nutrients are high, but the benefits are even higher: survival is at stake.

Humans (and rats and cockroaches) are food generalists. They eat a wide range of foods; virtually anything that can fit into the mouth is potential food. Generalists have few innate determinants of food choice, simply because it is not easy to predict the nutritive and toxic properties of a potential food on sensory grounds. There are only a few documented innate biases in human (and rat) food choice. First, there are innate predispositions to like or dislike certain tastes. There is an innate preference for sweet tastes, which, in nature, are predictive of calorie sources. The long history of sweetness in human culture, from fruit preferences to cultivation of fruits, to extraction of sugar from fruits, to colonization of the Americas partly to get a source of sugar, to the development of sugar substitutes, is all driven by the innate liking for sweet tastes (sugar). There is an innate tendency to dislike bitter tastes, and there are probably innate predispositions to reject very strong tastes including oral irritation (as from peppers) and to like fatty textures. Second, there is a suspicion about trying new foods called neophobia (on account of potential toxicity), but also a conflicting interest in them (on account of their potential as a new nutrient source). This has been described as the generalist's (omnivore's) dilemma. Third, there is a special learning mechanism, which allows learning about the consequences of ingestion, even when these may occur hours after ingestion. Fourth, there are a few specific internal states, which seem to signal the need for specific nutrients. Examples are thirst promoting procurement of water and hunger, promoting procurement of sources of energy. There may be a specific internal signal associated with sodium deficiency.

Humans are mammals, and hence have a unique first food, milk. Humans and other mammalian generalists make a transition from this single food to a very wide range of foods. This transition is made under the guidance of parents and other caretakers, and usually consists of a scheduled and graduated introduction of new foods, accompanied by mental maturation of the child and increased ability to chew food. For all mammal species except a subset of modern humans, full digestion and utilization of milk is limited to infancy. The development of dairying afforded the opportunity for humans to exploit milk and its products as adult foods.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.55047-8

Determinants of Human Food Choice

For humans, food choice is accounted for by some combination of biological, psychological, and cultural factors. The role of biological factors is limited for two reasons. One, as discussed above, is the human generalist orientation to food. Second, the modern developed culture world presents a food world to its members, which is vastly different from the ancestral world in which the biological predispositions evolved. A shortage of food has been replaced by a surplus of food. A modest range of natural, local foods has been replaced by an enormous range of foods, representing most of the world's cuisines at a neighborhood supermarket. 'Super foods' like chocolate with an extraordinary appeal to our innate liking for sweet and fatty tastes have been invented and widely distributed. Food choice, for some in the developed world, represents concerns about child labor, species at risk, and the appropriation of much of the food system by large corporations. Many of these themes are well expressed in Michael Pollan's (2006) book, *The Omnivore's Dilemma*. A major risk of food poisoning from toxins and microorganisms has been sharply reduced; the longer lifetime that has resulted from this and other medical advances establishes a new and more subtle link between dietary patterns over decades and degenerative diseases. Direct experience of food and the consequences of ingestion have been largely replaced by exposure through the media, schools, and other social contacts. Massive amounts of information (some accurate) relating dietary pattern to wellness and disease is widely disseminated. And finally, the linkage between energy expenditure and finding food (or finding anything else) has been almost eliminated. Thus, while the ancestral food environment focused on shortage and short-term consequences of eating, the modern environment stresses overabundance and long-term consequences. However, we must remember that more than 80% of people on Earth are not in the developed world. For them, getting enough food, both calories and nutrients, remains a major concern. Most of their food choices are dictated by strong culinary traditions and economic constraints. These people are not yet inundated by an abundance of different foods, although colas and candies and fast foods have already entered their lives. As globalization continues, and agriculture and distribution systems improve, more and more of the world will face the inversions of our ancestral environment. Culture is the most powerful determinant of human food choice. More can be learned about an individual's food attitudes and preferences from his or her cultural identity than from any other single piece of information. Food choice is highly constrained by availability and cost, and both of these are largely determined by culture, as well as local climate. Exposure to a food is a necessary, and sometimes sufficient, condition to produce a liking for a food, so that cost and availability indirectly but powerfully influence the development of food preferences.

Cost and availability aside, human preferences and food attitudes can be framed by a psychological (as opposed to nutritional) taxonomy of foods. There may be three types of reasons for rejecting or accepting a food: *sensory affective*, that is, feelings (pleasure or displeasure) based on the sensory experience (largely taste and smell); *anticipated consequences*, what the expected consequences are of eating a food; and *ideational*, what is known about a food (e.g., where it comes from, what the nature of it is).

Food rejections can be understood in terms of the selection and interplay of these reasons. One category of rejections is called *distaste*. These are entities rejected because of negative sensory affective properties, such as lima beans, broccoli, beer, or chili pepper, for those who find these foods distasteful. A second category is *danger*. These are things rejected primarily because they are believed to be harmful, because of acute or long-term consequences. The emotion of fear is often associated with their consumption. A third category of rejected things – the largest category – is called *inappropriate*. These are things that the culture labels as inedible, such as pencils, grass, paper, or cloth. They might taste good, and might be harmless, but they are rejected for ideational reasons. The fourth category is *disgust*. Disgusting food rejection is also based on ideational, culturally transmitted information, but unlike inappropriates, there is a strong belief that disgusts taste bad and are harmful. Unlike the affectively neutral response to inappropriates, the response to disgusts is strongly negative and emotional. Disgust is the most powerful reaction people have toward food. Disgusting entities are so powerful that if they touch an otherwise acceptable food, they render it undesirable, disgusting, and inedible (the principle of contamination or contagion).

On the positive side, there are four comparable categories. Good taste (acceptance principally because of sensory properties), beneficial (acceptance largely because of consequences), appropriate (acceptance because it is culturally designated as food, or food for a particular occasion), and transvalued (food enchanced because of its prior history). The transvalued category is much weaker and smaller than the disgust category, in most cultures. In Hindu India, food that has been 'shared' with the Gods (via donation to the priests in the temple, and then returned, in part), called 'prasad,' is an example of transvaluation.

The Acquisition of Food Preferences

For newborn infants, the only functioning categories are good taste (e.g., sweet) and distaste (e.g., bitter). Generally, infants will place in their mouth anything that might fit in, including feces and potentially toxic foods. Gradually, they acquire more distastes and good tastes, and learn about dangerous and beneficial foods. Disgust appears later. The full adult categorization is in place by roughly 5–8 years of age.

Most is known about the distinction between distaste and danger. When ingestion of a food is followed by nausea, it tends to become disliked, that is a distaste. However, if ingestion of a food is followed by most other negative symptoms (e.g., lower gut pain, skin rash, respiratory distress), the food typically becomes a danger. With respect to affect and emotion, it is notable that dislikes (distastes) and dangers have very different properties, although the outcome (rejection) is the same. The nausea-based acquired distaste (often called a conditioned taste aversion), unlike dangers, is not based on a legitimate sense of danger. Even if a person knows that the nausea/upper gastrointestinal illness was not produced by the food, the aversion remains.

The acquisition of good tastes is more complex and less understood. Mere exposure to a food, in itself, often seems sufficient to produce an acquired like. In addition, the association of a food with an already positive event (an already liked food mixed with it, positive regard by a respected person, a pleasant environment, nutritive postingestional consequences), by a process called evaluative conditioning, can lead to acquired likes (or acquired dislikes, if the paired events are negative, as in conditioned taste aversions). Indications of liking by a significant other (peer, older child, teacher, parent) may cause acquisition of liking. The process at work here is not understood. Efforts by adults to promote liking by emphasizing the beneficial consequences (better health, a specific reward for eating) seem to block the acquisition of liking for the food. When a child observes respected others enjoying a food, this promotes liking; when she is rewarded for consuming it, this seems to block the acquisition of liking.

The acquisition of knowledge of appropriate and inappropriate foods seems to be largely cognitive and affect free. The affect-laden acquisition of disgust contrasts with this, and may originate with the process of toilet training, which produces the universal disgust, feces. Some communication of affect (facial and other) is almost certainly involved, but there are also important cognitive aspects; it is the nature of the foods and their history that is central to this category.

Individuals in specific cultures develop likings (the good taste category) for some foods that are innately unpalatable (e.g., bitter, very strong tasting, irritating). Innately unpalatable foods are typically among the favorite foods: chili pepper, black pepper, ginger and other irritant spices, coffee, bitter chocolate, tobacco, alcohol, burnt foods, and highly salted foods. Such reversals of innate aversions are common in humans, and absent or rare in animals. We do not know how these preference reversals occur. In part, it may be through the same processes that produce normal likes: mere exposure, evaluative conditioning, and social approval. But there may also be special mechanisms that are involved in preference reversals. One possible mechanism places liking for innately unpalatable foods in the same category as thrill-seeking activities such as roller coaster riding. That is, humans get pleasure out of situations in which innate aversions or fears are stimulated, but in which they realize that there is no real danger. Roller coasters are safe, and so is chili pepper. This may be the case of pleasure derived from 'mind-over-matter.'

Family Influences and the Family Paradox

There are strong arguments for high parent–child resemblance in food habits. There is the common genetic heritage (including genetically based bitter sensitivity), mother's milk transmits some of the flavors of the mother's diet, parents control access (and hence exposure) to foods, and the principal affective signals about foods in early life come from the parents. Yet parent–child correlations in liking for foods are very low, in the range of 0–0.3, in comparison to parent–child correlations in values (such as attitudes to abortion). These low parent–child correlations appear whether the children studied are 4 year olds or college students. This family paradox cannot presently be explained. Of course, our first food, milk, was not available as a food postweaning in our ancestral environment, so from an evolutionary perspective, we should not develop strong lifelong attachments to our first food. Although all experiences with food, from the womb, onward have some influence on adult preferences, there is no reason to privilege the first 6 years of life. There is reason to believe that peer influence is very important, throughout the school years, although this influence has received relatively little attention.

Preferences and Values: Moralization and Vegetarianism

Selecting and consuming a food has social and moral implications, as well as physiological and sensory consequences. The source of a food (e.g., who prepared it, animal or vegetable origin) can powerfully influence acceptance. Particular food or food product boycotts in recent decades testify to this fact. Hindu Indian concerns about food fall largely within the moral domain. In the United States, there is a long history linking particular foods with the good life or immorality. Through the process of moralization, choice of a particular food may move from the domain of preferences to the domain of values. This has clearly happened with cigarette smoking in the United States. For some vegetarians, eating of animals is an immoral act. When ingestion of a food (including alcohol) becomes moralized, censure of consumers becomes licensed and governments and institutions may take limiting or prohibiting actions.

The Nature and Context of Food Choice

Food choices are sometime made on cognitive/rational grounds, whereas at other times the choice is made more rapidly and intuitively. First exposures to a food (at the point of purchase or ingestion) are more likely to engage rational concerns, whereas subsequent choices of the same food tend to be made in a more automatic, habitual manner. Repeated consumption of the same food on different occasions is likely to lead to increased preference, perhaps because of mere exposure, but repeated consumption over a short period of time can lead to boredom and reduced pleasure (called sensory-specific satiety). Food choices are often made well in advance of actual consumption, as in ordering in restaurants or purchase in a food market. Under the latter conditions, important determinants include branding and salience of the location of foods in the food display. The individual context is also important, including the time of day (breakfast vs dinner foods), the time of year and special annual celebrations like the American Thanksgiving, and stage of life, from infancy to old age.

Across the life course, one goes from intake determined largely by others in infancy to more and more agency in food choice with age, until older ages when the power of choice may become more circumscribed. Important areas of recent concern include social class and gender as determinants of food choice, and cultural differences in concern about genetically modified foods. Finally, the social context of eating affects both the amount eaten and food choice. Social comparison and social

norms have substantial effects. There are intense and often very pleasant social interactions, sometimes called commensality, that take place during meals. In short, understanding a single food choice engages a wide range of influences.

Food, Fear, and Pleasure: Cultural Perspectives

In the nonhuman animal world, and among traditional humans, food is basically a source of pleasure. Even in the developed world, cost and availability aside, the principal factor influencing food choice is taste, that is to say, pleasure, although convenience, health, and processing history of a food are also important determinants. However, with the major changes in the human environment in the developed world, obesity and the consequences of long-term exposure to particular diets have become more salient. As a result, the earlier simple pleasure of eating has become, for many people, an ambivalence, or even an outright fear. This is most well developed in American women, who often fear food as much as they savor it. In some, there is now embarrassment and guilt about consuming high-fat foods, great concern about thinness, and great concern about the healthfulness of every bite of food. This is not a necessary consequence of the modern environment, since in France, the pleasures of eating still dominate.

See also: Emotion, Neural Basis of; Hunger and Eating, Neural Basis of; Olfactory System.

Bibliography

Barker, L.M., 1982. The Psychobiology of Human Food Selection. AVI, Bridgeport, CT.

Birch, L.L., Fisher, J.O., Grimm-Thomas, K., 1996. The development of children's eating habits. In: Meiselman, H., MacFie, H. (Eds.), Food Choice, Acceptance and Consumption. Blackie Academic & Professional, Glasgow, UK, pp. 161–206.

Chandon, P., Wansink, B., 2012. Does food marketing need to make us fat? A review and solutions. Nutrition Review 70, 571–593.

Diamond, J., 1999. Guns, Germs, and Steel. W.W. Norton, New York.

Frewer, L., van Trijp, H., 2007. Understanding Consumers of Food Products. Woodhead, Cambridge, England.

Kass, L., 1994. The Hungry Soul. Free Press, New York.

Koster, E.P., Mojet, J., 2007. Theories of food choice development. In: Frewer, L., van Trijp, H. (Eds.), Understanding Consumers of Food Products. Woodhead, Cambridge, England, pp. 93–124.

Logue, A., 1991. The Psychology of Eating and Drinking, third ed. W. H. Freeman, New York.

Meiselman, H., MacFie, H. (Eds.), 1996. Food Choice, Acceptance and Consumption. Blackie Academic & Professional, Glasgow, UK.

Nicklaus, S., Issanchou, S., 2007. Children and food choice. In: Frewer, L., van Trijp, H. (Eds.), Understanding Consumers of Food Products. Woodhead, Cambridge, England, pp. 329–358.

Pollan, M., 2006. The Omnivore's Dilemma. Penguin, New York.

Prescott, J., 2012. Taste Matters. Reaktion Books, London.

Rozin, P., Fallon, A.E., 1987. A perspective on disgust. Psychological Review 94, 23–41.

Rozin, P., 2007a. Food and eating. In: Kitayama, S., Cohen, D. (Eds.), Handbook of Cultural Psychology. Guilford, New York, pp. 391–416.

Rozin, P., 2007b. Food choice: an introduction. In: Frewer, L., van Trijp, H. (Eds.), Understanding Consumers of Food Products. Woodhead, Cambridge, UK, pp. 3–29.

Shepherd, R., Raats, M. (Eds.), 2006. The Psychology of Food Choice. CABI, Wallingford, UK.

Shepherd, R., Raats, M., 1996. Attitudes and beliefs in food habits. In: Meiselman, H., MacFie, H. (Eds.), Food Choice, Acceptance and Consumption. Blackie Academic & Professional, Glasgow, UK, pp. 346–364.

Sobal, J., Bisogni, C.A., Devine, C.M., Jastran, M., 2006. A conceptual model of the food choice process over the life course. In: Shepherd, R., Raats, M. (Eds.), The Psychology of Food Choice. CABI, Wallingford, UK, pp. 1–18.

Yeomans, M.R., 2006. The role of learning in development of food preferences. In: Shepherd, R., Raats, M. (Eds.), The Psychology of Food Choice. CABI, Wallingford, UK, pp. 93–112.

Food Production, Origins of

Gary W Crawford, University of Toronto Mississauga, Mississauga, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P.J. Watson, volume 8, pp. 5722–5728, © 2001, Elsevier Ltd.

Abstract

The intensified intervention of people with the lives of plants and animals that ultimately led to their domestication is arguably the most important cultural revolution since the beginning of the Upper Paleolithic (the appearance of modern human behavior *c.*120 000–50 000 BP). Food production is considered here to be synonymous with agriculture. Food production references a range of adaptations from pastoralism to low through higher degrees of dependence on domesticated plants and animals. Archaeological evidence indicates that food production and agriculture developed independently in more than 10 locations around the globe during the Holocene. The first evidence of ecological management that leads to agriculture dates to between 13 000 and 10 000 years ago.

The Origins of Food Production in World Areas

Systematic research on the origins of food production demonstrates that independent origins were widespread around the globe. The main exceptions are Australia and Arctic regions where indigenous plant and animal domestication did not occur. Yet even in Australia aboriginal peoples developed sophisticated plant and animal management strategies, usually focused around burning. Human-set fires in the Western Desert have been crucial in maintaining biodiversity and habitat mosaics and were an important aspect of landscape management (Bird et al., 2005). Interdisciplinary archaeological projects are recovering evidence for early food production in Africa, Southwest, South and East Asia, New Guinea, and the Americas. A brief global survey follows.

The Origins of Food Production in Southwestern Asia

The first domesticated food plants were a tree fruit (the fig), large-grained grasses (rye, wheat, barley), and legumes (lentils, peas, vetches) (Zeder, 2011). The earliest regular use of these crops dates to 11 000–10 000 BP. Tilling the soil and tending plants may have begun even earlier, by 13 000 BP. Evidence of breeding control of animals (goats and sheep) also comes from roughly the same period, 12 000–11 500 BP. Daniel Zohary (1996) thinks that most of the founder crops (wheat, barley, and legumes) were domesticated only once or twice, and then spread both east and west. The first domesticated animals were sheep and goats at, for example, Nevali Cori at the headwaters of the Euphrates River (10 500–10 200 BP); pigs and cattle appear to have been domesticated by 10 300 BP and 9000 BP, respectively.

An extraordinary discovery in the late 1990s exposed a 20 000-year-old occupation in the shallow waters near the shores of Lake Galilee (Kislev et al., 1992; Weiss et al., 2008). Not only were small, fragile houses evidenced in the well-preserved remains at Ohalo but also were the remnants of gathering and hunting. Particularly informative were the vast array of grass seeds and tree fruits the site's inhabitants collected. The grasses included the wild ancestors of plants that, over 10 000 years later, would be domesticated. As time passed, the descendants of Ohalo's inhabitants began to focus on a narrower range of grasses and turned their attention to growing barley and wheat. Gathering evidence from between 12 000 and 8000 years ago tells us that there were distinctive pathways to food production not only in the hilly flanks of the Zagros to the east but also within the Levant as well (Goring-Morris and Belfer-Cohen, 2011). These distinctive pathways seem to correlate with the mosaic of habitats and climates that characterized the Levant in the early Holocene. Ecotones within this mosaic are sensitive to both short-term and long-term climatic fluctuations.

The Natufian culture (15 000–11 600 BP) is particularly relevant because it was at the end of this period, and the beginning of the subsequent Pre-Pottery Neolithic (PPN; 11 600–8400 BP) that populations began to invest in food production. The Natufian, or Late Epipaleolithic, is characterized by a significant increase in community size, marked sedentism, and much longer duration of settlements than in the preceding Upper Paleolithic. Although it is tempting to refer to the Natufian as transitional between the Upper Paleolithic and Neolithic, they were not; rather, their lifeway represented a long-lived adaptation in the region marked by increased packing of settlements and strong territoriality marked by local variations in material culture (Goring-Morris and Belfer-Cohen, 2011). Unfortunately, few plant remains have been found from Natufian and early PPN occupations although animal bones are well represented. Hunting intensified with people focusing on gazelle (Legge and Rowley-Conwy, 1987). The first hint that plants were being domesticated come from the PPNA occupation (11 600–10 500 BP) of the Abu Hureyra and Gigal sites. Seedless fig that requires human intervention to reproduce is the earliest evidence (11 400–11 200 BP) of a domesticated plant in SW Asia (Kislev et al., 2006). Large, plump rye grains dating to *c.*11 000 BP are more like domesticated than wild rye (Moore et al., 2000), but not everyone agrees on the meaning of the evidence. A brief hiatus is recognized between the PPNA and PPNB (10 500–10 100 BP) when resident populations were small or non-existent (for reasons not understood). By the beginning of the second period (PPNB), small farming villages had coalesced

(for further details see Goring-Morris and Belfer-Cohen, 2011; Zeder, 2011).

Africa

Insights from Africa provide an important counterpoint to the common narratives of the evolution of food production. Crops such as wheat, barley, lentil, and pea were introduced to the northern coast of the continent between 7600 and 7000 BP, likely from Spain (Morales et al., 2013), but indigenous resource management and domestication developed elsewhere in Africa too. One hallmark of African agriculture is pastoralism, which became an important aspect of human adaptations in northeastern Africa by 6000 BP (Marshall and Weissbrod, 2011). During the last glacial maximum, *c*.18 000 BP, much of northern Africa was too arid to support people, but by the Early Holocene it was more humid than today. Hunter-gatherers were widespread in what is now the Sahara desert where people lived near permanent water sources. Animal domestication appears to have developed in the context of these climate changes. Animal management and their subsequent domestication may have been a response to risk inherent in the Middle Holocene development of the desert regions of northeastern Africa. Early pastoralism may not have developed strictly for food production because the economy at the time was based on early codependence of people and animals, particularly the donkey and cattle. Whether these animals were domesticated is not clear because morphological changes are difficult to see. Reduced size is not necessarily an indication of animal domestication in this instance (Marshall and Weissbrod, 2011). Human diets do not appear to have been dependent on these animals but the close relationship between people and animals involved logistical arrangements of movement to pastures. These animals held important positions in the symbolic life of these northeastern African peoples, as evidenced by the prominence of cattle in rock art. Close relationships between people and domesticated animals, and possibly animal management, appear to have begun in the early Holocene between 10 000 and 8000 years ago. Domesticated cattle may have been present *c*.9000–8000 BP at Nabta Playa, a site that is well outside the natural range of cattle (Jórdeczka et al., 2013). By 5000 or 4000 BP, cattle remains are common at archaeological sites although the evidence for reliance on domesticated plants is minimal.

Sorghum, pearl millet, cowpea, African rice, and t'ef were domesticated in Africa but not particularly early. Pearl millet was being grown by Kintampo culture people (3600–3200 BP) and further evidence of landscape management comes from the extensive use of oil palm at this time, reflecting what is likely early arboriculture (D'Andrea and Casey, 2002; Logan and D'Andrea, 2012). T'ef, an extremely small-grained grass, has an archeological record that is difficult to interpret because seeds of the domesticated form are not particularly different from those of wild t'ef; however, it was likely under cultivation in Ethiopia at least 2000 years ago (D'Andrea, 2008). Sorghum was a significant component of the plant assemblage at Nabta Playa 8000–7300 BP so it may have been cultivated at the time (Wasylikowa et al., 2001), but its domestication may not have occurred until about 2000 years ago. Domesticated sorghum and African rice are present at the site of Jenne-Jeno in Mali about 2000 years ago (McIntosh, 1995). In South Africa, agriculture and pastoralism are thought to have arrived in the early Iron Age about 2000 years ago with the spread of food producers into this part of the world (for further details see Breunig and Neumann, 2002; Harlan, 1992; Morales et al., 2013).

South Asia

The origins of food production in the Indian subcontinent may involve as many as five independent areas or as few as three (Fuller, 2011). Understanding the circumstances of these origins is clouded by ambiguous plant remains, the lack of early hunting and gathering sites, as well as a scarcity of sites that provide data about the people who first developed food production. Available evidence points to relatively small groups of mobile hunters and gatherers making the first steps to food production. An exception is found in northwestern Pakistan where the earliest documented food production in South Asia is found at the site of Merhgarh, the earliest levels of which date to about 9000 BP. The crops here, as well as goats, were introduced from Southwest Asia. To what extent food production in Pakistan influenced the origins of food production elsewhere in South Asia is not known. The widespread development of food production in the Indian subcontinent in disparate regions and the fact that the wild progenitors of the earliest millets and legumes are found here strongly suggests that food production began independently, at least in some areas. Significant crops include native millets (small-grained grasses) and a dozen legumes (in particular, mungbean and horsegram). Wheat, barley, chickpea, lentil, and pea were all introduced from Southwest Asia. The latter three crops, all legumes, appear to be limited to the Indus valley in Pakistan. Domesticated rice appears many millennia later in India than in China. The circumstances under which rice became an important resource in the Indian subcontinent remain a significant outstanding issue in the study of food production in the region. The wild progenitor of rice grows in India but a sequence of remains documenting the evolution of domesticated rice here has not been found in the region. Rice, in association with other crops, is present by about 3500 BP (for further details see Fuller and Madella, 2001; Fuller, 2011).

East Asia

The origins of food production in East Asia are complex and did not follow a single pattern (for further details see Cohen, 2011; Crawford, 2011; Liu and Chen, 2012; Zhao, 2011). The last 15 years have seen significant developments in understanding the origins and development of food production in China, South Korea, Japan, and the Russian Far East. Among these regions, only China and Japan have evidence of independent origins of food production. The situation in Japan is still debated (Crawford, 2011), but in China at least three independent origins have been documented: North China, Central-East China, and tropical South China. Rice was an important technological resource, its

chaff being used to temper clay used for pottery between 10 000 and 8000 years ago in the highlands to the south of Hangzhou Bay in South China (Jiang and Liu, 2006). The extent to which it was used as food at this time is not known. Significant levels of rice management are evident by 8000 BP at the Kuahuqiao site in Zhejiang Province. From 8000 to 7000 BP, rice spread northward and was a significant resource at the Jiahu and Yuezhuang sites. Many plants important to the modern Chinese economy such as the peach, foxnut, lotus root, and water chestnut are evidenced in the archaeological remains between 8000 and 5000 BP in the lower Yangzi valley. People were also raising pigs, dogs, and possibly cattle throughout much of North and Central-Eastern China.

Initially, the most important crops in North China were foxtail millet and broomcorn millet. Broomcorn millet was widely used 8000–7500 years ago from the West through central Yellow River Valley to Inner Mongolia. Starch grains recovered from grinding stones dating between 10 000 and 8000 years ago suggest that at the very least the wild ancestor of broomcorn millet was being harvested in North China. The soybean was also being exploited at this time. Research on tropical South China is just beginning but preliminary evidence indicates that tubers and root crops such as yams were being harvested at Zengpiyan (12 000 BP) while rice production was introduced, likely from the Yangzi valley, *c*.6000 BP according to studies at sites such as Xiaojing in Guangxi province (Zhao, 2011).

Millet production spread to the Korean Peninsula sometime during the Middle Holocene. How this happened is poorly understood. By 3500–3000 BP, wet rice and dryland food production were well established in South Korea and from there moved into southwestern Japan by 2800 BP. Before food production from the mainland was introduced to Japan, local people belonging to the Jomon cultures had developed a variety of resource management strategies. Although the Jomon did not develop intensive grain-based food production, they managed nut orchards, groves of lacquer trees, and domesticated Japanese millet and a type of soybean. A close relationship with the wild boar developed too, but these animals were not domesticated.

Southeast Asia and the Islands of the Southwest Pacific

An array of tuber, root, and tree crops were domesticated in Southeast Asia, New Guinea, and the islands of the southwestern Pacific (Denham, 2011). Evidence of these is difficult to retrieve archaeologically because of the fragile nature of starchy tissue in particular and the difficulty in assessing where along the continuum from wild through domesticated forms that tree seed or nuts recovered from sites belong. Much of the discussion about plant domesticates from the region revolves around ethnographic, ethnobotanical, and linguistic data as well as delineation of prehistoric agricultural terraces, ditches, and other landscape features. For now, it appears that crops spread from New Guinea to the rest of Southeast Asia. Banana, taro, a yam, and a nut produced by an evergreen tree (karuka; *Pandanus julianettii*) are the only archaeologically represented crops. Our understanding comes primarily from the Wahgi valley, New Guinea. Mounded cultivation is evident by 7000–6500 BP followed by ditch fields *c*.4000 BP at the Kuk site and presumably banana, yam, and taro at least were under production before the landscape was engineered. Investigators believe that vegetative reproduction methods such as grafting and transplanting were favored over seed planting. Other crops that may have been domesticated in the lowlands include sugarcane, sago, breadfruit, and *Canarium*.

A strong case has been made for landscape modification in the Wahgi Valley, New Guinea, especially the creation of ditched field systems. Such systems presumably were constructed in the interests of growing taro and native yam species, long preceding the *c*.3500 BP expansion of people from Southeast Asia onto the Pacific islands. Those colonizing populations were accompanied by domestic animals (probably dogs, chickens, and pigs) and an array of domesticated tuber and root species (*see* Pacific Islands, Archaeology of; for further details see Denham, 2011).

The Americas: Mexico

Archaeologists are not in agreement on when people first arrived in the New World but it appears to have been before 15 000 BP. By 10 000–9000 BP, the first steps to food production were beginning in Mexico. These steps involved the domestication of a few plants between 10 000 and 6000 BP by relatively small groups of mobile hunters and gatherers during the Archaic period (*c*.10 000–4000 BP). Villages that relied to some extent on food production did not develop across the Mexican landscape until perhaps 4000–3500 years ago. What we know comes from only a few rock shelter and cave sites, mainly Guila Naquitz (Flannery, 1986) and Coxcatlan and San Marcos (Byers, 1967; MacNeish, 1992) in the dry interior of Oaxaca and Puebla, respectively, and the Xihuatoxtla site in the lowlands of the Balsas valley of Guerrero. At least a dozen plants of worldwide economic importance owe their domestication to the ancient inhabitants of Mexico. Among them is maize or corn (*Zea mays*), a grass native to Mexico. Yet, despite its well-known genetics, its archaeological record is not nearly as well documented as many of the old world grasses. Its wild ancestor (teosinte) is indigenous to the Balsas River drainage in western Mexico and the only archaeological record for maize in the Balsas valley consists of starch grains and phytoliths recovered from a grinding stone in contexts that appear to be older than 9000–8600 BP at the Xihuatoxtla site. Squash phytoliths are also evidenced in this context. Excavations by Kent Flannery in 1966 at Guila Naquitz (Flannery, 1986), a rockshelter in Oaxaca, have resulted in information about several other cultigens: squashes, bottle gourd, beans, and chili peppers. *Amaranthus* (amaranth), *Chenopodium* (goosefoot), *Agave* (agave, maguey). Several root crops and tree fruits thought to have been encouraged and harvested if not actually domesticated have been recovered too. Large squash seeds provide evidence that the plant was being domesticated by *c*.9000 BP at Guila Naquitz. The oldest preserved maize cobs so far discovered in Mexico date to about 6000 BP at Guila Naquitz where there is no evidence that the wild ancestor of maize grew. It was likely imported, probably from the Balsas valley. No fieldwork on early Mexican agriculture comparable to

MacNeish's and Flannery's has taken place since the 1960s (for further details see Piperno, 2011; Smith, 2005).

The Americas: Central and South America

Food production in Central and South America appears to have involved a complex mosaic of plants and animals over a broad geographic range and elevations, as well as many different regions. Given this variation it should come as no surprise that agriculture began in many different places in these regions (Piperno, 2011). Many herbaceous plants, trees, root crops, grains, medicinal plants, and legumes were domesticated in the region, and some of the more economically important ones were potato, tomato, sweet potato, manioc, peanut, chile pepper, quinoa, coca, pineapple, avocado, several squashes and beans, and cotton. Unlike the Old World, only a few animals were domesticated here. These include the llama, alpaca, and guinea pig that were domesticated in Andean South America.

The exploitation of plants that would eventually be domesticated is widespread early on. According to Dolores Piperno (2011), several species of *Cucurbita* (squash) have been recovered from Panama, Colombia, Ecuador, and Peru, from 11 750 to 8000 BP; bottle gourd was exploited as much as 11 000 BP; peanut and manioc make their first appearance in the record *c*.8500 BP in the Zaña valley of northern Peru. Bottle gourd is intriguing because the botanical understanding is that it is an African plant that spread to many parts of the world at a relatively early date (Kistler et al., 2014). Evidence for the first dispersals of manioc and peanut northward dates to about 8500–7600 BP, while based on phytolith evidence maize appears to have diffused into South America from Mexico by 8000 BP. In the western Andean foothills of northern Peru, small-scale irrigation evidences ecological engineering as early as 6500–5400 BP and by 4000 BP agriculture was well established (Dillehay et al., 2005). Similar to Mexico, the first people to heavily exploit these and other domesticated plants lived in small, seasonally dispersed settlements.

The only herd animals domesticated in the New World are the llama and alpaca. Their distribution is limited to the Andes. Intensified use, changes in patterns of mortality, and morphological changes in bones indicate that they were being domesticated between 4500 and 4000 BP (Mengoni Goñalons, 2008). Guinea pig may have been domesticated around the same time. Quinoa may also have been domesticated in the highlands by 5000 to 4000 years ago too, but the best evidence for the crop is from the Early Formative period Chiripa site *c*.3500 BP in the south-central Andes (Bruno and Whitehead, 2003).

The arid west coast of South America in Peru and Chile has its own food-producing story to tell, related to that of the highlands in some ways, but with significant differences. Preservation of even very delicate plant parts as well as textiles and human remains makes archaeological remains in the region attractive to professionals and to looters. Evidence for plant use begins early at some sites, as far back as 8800 BP for gourds, and includes many of the same species referred to above for other areas of South America. Whether these plants were wild and their natural distribution included all of these areas or domesticated in one or more regions still needs investigation. The coastal valleys of Peru currently provide the oldest directly dated evidence for beans (one lima bean pod) at *c*.5600 BP; Mexican common and tepary beans are no older than 2500 BP (for further details see Piperno, 2011).

The Americas: Eastern North America, with Comparative Reference to the Southwestern United States

Indigenous food production was being established by 4500–4000 BP during the Late Archaic period in Eastern North America. The people who began plant cultivation were focusing their subsistence activities on rich river valleys and floodplains while exploiting a range of resources from a variety of other habitats (Smith and Yarnell, 2009). The occupants of Archaic shell mound sites along the Green River in Kentucky, for instance, were managing nut trees and likely encouraging anthropogenic animal communities by burning away competing plants. Archaic peoples' understanding of ecological management techniques set the stage for subsequent food production developments. Sunflower (*Helianthus annuus*), sumpweed (*Iva annua* var. *macrocarpa*), goosefoot (*Chenopodium berlandieri*), little barley (*Hordeum pusilum*), and squash (*Cucurbita pepo*) were independently domesticated here, and by 3800 BP they formed a crop complex that became the focus of food production for cultures throughout the U.S. Midwest. Eventually maygrass (*Phalaris caroliniana*) and erect knotweed (*Polygonum erectum*) joined the crop complex. Bottle gourd was also in the region quite early but it is not native, apparently having arrived from Africa by floating on ocean currents some time before 10 000 BP (Kistler et al., 2014). When and where morphologically distinct domesticated forms appeared is not yet known.

Maize was introduced from Mexico into the U.S. Southwest shortly after 4500 BP and appears to have been used by hunter–gatherers to make their annual hunting rounds more secure rather than shifting people into becoming food producers (Wills, 1995). In fact it was not until slightly over 2000 years later (around 2000 BP) that communities of food producers appeared. In all cases, considerable use was made of game animals and indigenous southwestern plants, some domesticated, others semidomesticated or encouraged, and others harvested from wild stands (Wills, 1995). In some areas, even before crops were introduced, people managed local resources to the extent that the people may well have been involved with an agricultural ecology before crops became important resources. Maize then appeared in Eastern North America, likely introduced from the Southwest, some time between 3000 and 2000 years ago. It was a minor component of local economies until *c*.1000 BP. After that it was a critically important resource to societies in much of Eastern North America including southern Ontario and Quebec, Canada (for further details see Crawford et al., 2006; Minnis, 2004; Smith, 2011; Wills, 1995).

Discussion and Conclusions

This article interweaves domestication and agriculture in a way that implies that they are the most significant elements related to food production. This is certainly the way it is

treated by archaeology, the discipline that is primarily concerned with the topic. However, food production encompasses an array of strategies involving animal management, transplanting and encouraging desirable plants, grafting, burning to create habitats favorable to certain types of plants and animals, and the like. These strategies form an important part of modern human behavior that lies within the realm of ecological engineering or niche construction (Laland et al., 2001). Ecological management or engineering behaviors likely developed along with, and contributed to, the evolution of modern humans. The difficulty for archaeologists is determining how to recognize these behaviors especially when human populations were quite low. Future research on Upper Paleolithic (Old World) and Paleoindian (New World) human ecology will be critical to our understanding of food production developments as habitats and environments transformed at the end of the Pleistocene.

Cultures diversified around the world at the same time. The research outlined for the Ohalo II site in Israel and the latest Upper Paleolithic sites in North China attest to this. The tendency for these Late Pleistocene cultures to be viewed as primarily big-game hunters is changing because of this new research. Archaeologists are also coming to terms with the active role people have played in their local environments.

Agriculture is probably best considered as a range of human behaviors that involves people intervening throughout the life spans of specific plants and animals. Domestication involves genetic changes to plants and animals (Fuller et al., 2010; Larson, 2011; Doebley et al., 2006) that not only increases the dependence of these organisms on people but also emphasizes changes in plant parts that people exploit and in how these plants reproduce. Phenotypic or physical changes in domesticated animals in addition to behavioral changes that make them tamer are also important but archaeologists are learning that initially, at least, management of certain animals is easier to recognize by looking at their population structures. The first evidence for animal management comes in the form of a high proportion of young males, a higher proportion of older females, and, conversely, very few older males and very few younger females in archaeologically recognized populations. Domestication was an unintended consequence of strategies practiced by humans long before they made any commitment to agriculture. The difficulty for archaeologists is to sort out these behaviors in the record before intensive agriculture developed.

For now the best evidence comes from Southwest Asia, Eastern North America, and North and East-Central China but not South China. Even in these regions, significant gaps in our knowledge need to be filled. The record for South and Southeast Asia, Africa, as well as from South and Central America is significantly stronger than it was two decades ago, but archaeologists are still struggling to outline the details of late Pleistocene and early Holocene developments there.

Be that as it may, certain patterns are emerging. The evidence suggests that there were 10 or more areas where indigenous and distinctive food-producing economies developed: Southwest Asia (more than one area), Saharan and sub-Saharan Africa (perhaps with West African and East African foci), North, East-Central, and South China, Japan, Indian subcontinent (as many as six areas), Eastern North America, Mexico, Central and South America (with neotropical and highland foci), and New Guinea. Discussion of early crops in these locations centers heavily upon those that dominate modern world economies (e.g., wheat, rice, maize, soybean), sometimes to the detriment of understanding the nature and function of ancient subsistence systems, whether food producing or not. Given the present concerns about the grave dangers of monocropping on a regional, let alone a global, scale, the focus on primary crops is doubly unfortunate. Lost crops and lost varieties of current crops and domestic animals are vital resources for modern agropastoralism, but can be retrieved only through the interdisciplinary work of archaeologists, paleoethnobotanists, and zooarchaeologists, on the one hand, and molecular biologists and plant and animal breeders on the other.

Two more issues central to any consideration of how ancient food-producing economies began are chronology and explanation. Is the timing of the events in the disparate regions informative? Were the causes and processes of independent, indigenous transitions to food production essentially the same? How one answers these questions depends almost entirely on how detailed and how secure regional chronologies and databases may be. On the evidence of present chronologies, developed with a significant dependence on AMS radiocarbon dating, it appears that only in Southwest Asia and East Asia were significant developments underway by the Late Pleistocene. Significant short-term fluctuations in temperature and rainfall occurred in the late Pleistocene. No one is clear whether these changes encouraged food production. The tropics were not impacted significantly by these climate change episodes. Highly localized environmental changes related to global-scale changes may still have been crucial. Significant developments were occurring shortly after 12 000 years ago in some areas but much later in other areas such as Eastern North America and Japan.

Prime movers used to aid explanation of transitions to food production range from the older favorites, climatic or environmental change (see earlier), and population pressure to newer models drawn from economics and behavioral ecology, to competition among lineage or clan heads, or to the necessity of intercommunity trade for forging alliances guaranteeing mutual assistance in hard times (e.g., Hayden, 1996). Some of these models are quite ingenious, and plausible for specific origins cases, but they do not fit others at all, making them less than satisfactory as general explanations for the origin of food production in the human past. That abstraction is very useful for certain purposes, but self-destructs when placed in contact with all the evidential bases now available, meager though many of them are. The shift to greater dependence on food production was happening between 13 000 and 8000 years ago in most areas although it was much later in other regions such as Eastern North America, Africa, and South Asia.

Through the years scholars often cited the natural tendency for human populations to grow. A crisis point would be reached and people would be compelled to produce food rather than extract food from the wild. A number of observations negate this proposal. Human demographics are complex and dependent on culture and environment. Archaeologists are also becoming increasingly aware of the active human role

impacting the productivity of resources throughout human existence. Furthermore, the process of domesticating plants and animals is known in hindsight, but hunters and gatherers in the Late Pleistocene would not have known about domestication. The development of food-producing economies relying on these domesticated organisms appears to have been an unintended consequence of decisions people were making at specific times and places to solve immediate problems. These problems likely had little to do with paucity of resources because in most areas where we have a detailed archaeological record food production was developing in relatively resource-rich areas.

High-level, theoretical formulations must be devised and tested against independent data. Hypotheses developed from what we already know need to be examined in the context of their implications. This means collecting new data, possibly using new biomolecular and archaeological recovery techniques, and including regions that have not been investigated in any detail such as southernmost China. This way we can expand our knowledge about what, in the abstract, is indeed the most important transition in human prehistory (for further details see Bocquet-Appel, 2011; Price and Bar-Yosef, 2011).

See also: Agroecology and Agricultural Change; Cultural Evolution: Overview; Food Security and 'Green Revolution'; Food in Anthropology; Hunter-Gatherer Societies, Archaeology of; Hunting and Gathering Societies: Anthropology; Pacific Islands, Archaeology of; Pastoralism in Anthropology.

Bibliography

Bird, D.W., Bird, R.B., Parker, C.H., 2005. Aboriginal burning regimes and hunting strategies in Australia's Western Desert. Human Ecology 33, 443–464.

Bocquet-Appel, J.-P., 2011. The agricultural demographic transition during and after the agriculture inventions. Current Anthropology 52 (S4), S497–S510.

Breunig, P., Neumann, K., 2002. From hunters and gatherers to food producers: new archaeological and archaeobotanical evidence from the West African Sahel. In: Hassan, F. (Ed.), Droughts, Food and Culture. Springer, US, pp. 123–155.

Bruno, M.C., Whitehead, W.T., 2003. *Chenopodium* cultivation and formative period agriculture at Chiripa, Bolivia. Latin American Antiquity 14, 339–355.

Byers, D.S., 1967. The Prehistory of the Tehuacan Valley. Robert S. Peabody Foundation, Phillips Academy, Andover.

Cohen, D.J., 2011. The beginnings of agriculture in China. Current Anthropology 52 (S4), S273–S293.

Crawford, G.W., 2011. Advances in understanding early agriculture in Japan. Current Anthropology 52 (S4), S331–S345.

Crawford, G.W., Saunders, D., Smith, D.G., 2006. Pre-contact maize from Ontario, Canada: context, chronology, variation, and plant association. In: Staller, J., Tykot, R., Benz (Eds.), Histories of Maize: Multidisciplinary Approaches to the Prehistory, Linguistics, Biogeography, Domestication, and Evolution of Maize. Elsevier, Amsterdam, pp. 549–559.

D'Andrea, A.C., 2008. T'ef (*Eragrostis tef*) in ancient agricultural systems of highland Ethiopia. Economic Botany 62, 547–566.

D'Andrea, A.C., Casey, J., 2002. Pearl Millet and Kintampo Subsistence. African Archaeological Review 19, 147–173.

Denham, T., 2011. Early agriculture and plant domestication in New Guinea and island Southeast Asia. Current Anthropology 52 (S4), S379–S395.

Dillehay, T.D., Eling, H.H., Rossen, J., 2005. Preceramic irrigation canals in the Peruvian Andes. Proceedings of the National Academy of Sciences of the United States of America 102, 17241–17244.

Doebley, J.F., Gaut, B.S., Smith, B.D., 2006. The molecular genetics of crop domestication. Cell 127, 1309–1321.

Flannery, K., 1986. Guila Naquitz. Academic Press, New York.

Fuller, D., Allaby, R., Stevens, C., 2010. Domestication as innovation: the entanglement of techniques, technology and chance in the domestication of cereal crops. World Archaeology 42, 13–28.

Fuller, D., Madella, M., 2001. Issues in Harappan archaeobotany: retrospect and prospect. In: Settar, S., Korisettar, R. (Eds.), Indian Archaeology in Retrospect, Protohistory. Indian Council of Historical Research, vol. II. Manohar Publishers and Distributors, New Delhi, pp. 317–390.

Fuller, D.Q., 2011. Finding plant domestication in the Indian subcontinent. Current Anthropology 52 (S4), S347–S362.

Goring-Morris, A.N., Belfer-Cohen, A., 2011. Neolithization processes in the Levant. Current Anthropology 52 (S4), S195–S208.

Harlan, J.R., 1992. Crops and Man, second ed. American Society of Agronomy/Crop Science Society of America, Madison, Wisconsin.

Hayden, B., 1996. Pathways to power: principles of creating socioeconomic inequalities. In: Price, D.T., Feinman, G. (Eds.), Foundations of Social Inequality. Plenum, New York, pp. 15–86.

Jiang, L., Liu, L., 2006. New evidence for the origins of sedentism and rice domestication in the Lower Yangzi River, China. Antiquity 80 (308), 355–361.

Jórdeczka, M., Królik, H., Masojć, M., Schild, R., 2013. Hunter–gatherer cattle-keepers of early Neolithic El Adam type from Nabta Playa: latest discoveries from Site E-06-1. African Archaeological Review 30 (3), 253–284.

Kislev, M.E., Hartmann, A., Bar-Yosef, O., 2006. Early domesticated fig in the Jordan Valley. Science 312, 1372–1374.

Kislev, M.E., Nadel, D., Carmi, I., 1992. Epipalaeolithic (19,000 BP) cereal and fruit diet at Ohalo II, Sea of Galilee, Israel. Review of Palaeobotany and Palynology 73, 161–166.

Kistler, L., et al., 2014. Transoceanic drift and the domestication of African bottle gourds in the Americas. Proceedings of the National Academy of Sciences of the United States of America. Published online 02.02.14.

Laland, K.N., Odling-Smee, J., Feldman, M.W., 2001. Cultural niche construction and human evolution. Journal of Evolutionary Biology 14, 22–33.

Larson, G., 2011. Genetics and domestication. Current Anthropology 52 (S4), S485–S495.

Legge, A.J., Rowley-Conwy, P.A., 1987. Gazelle killing in stone age Syria. Scientific American 255 (8), 88–95.

Liu, L., Chen, X., 2012. The Archaeology of China: From the Late Paleolithic to the Early Bronze Age. Cambridge University Press, Cambridge.

Logan, A.L., D'Andrea, A.C., 2012. Oil palm, arboriculture, and changing subsistence practices during Kintampo times (3600–3200 BP, Ghana). Quaternary International 249, 63–71.

MacNeish, R.S., 1992. The Origins of Agriculture and Settled Life. University of Oklahoma Press, Norman.

Marshall, F., Weissbrod, L., 2011. Domestication processes and morphological change. Current Anthropology 52 (S4), S397–S413.

McIntosh, S.K., 1995. Excavations at Jenne-Jeno, Hambarketolo, and Kaniana (Inland Niger Delta, Mali): The 1981 Season. University of California Press, Berkeley.

Mengoni Goñalons, G.L., 2008. Camelids in ancient Andean societies: a review of the zooarchaeological evidence. Quaternary International 185, 59–68.

Minnis, P.E., 2004. People and Plants in Ancient Western North America. Smithsonian Books, Washington.

Moore, A.M.T., Hillman, G.C., Legge, A.J., 2000. Village on the Euphrates: from Foraging to Farming at Abu Hureyra. Oxford University Press, Oxford.

Morales, J., et al., 2013. The origins of agriculture in North-West Africa: macro-botanical remains from Epipalaeolithic and Early Neolithic levels of Ifri Oudadane (Morocco). Journal of Archaeological Science 40, 2659–2669.

Piperno, D.R., 2011. The origins of plant cultivation and domestication in the new World tropics. Current Anthropology 52 (S4), S453–S470.

Price, T.D., Bar-Yosef, O., 2011. The origins of agriculture: new data, new ideas. Current Anthropology 52 (S4), S163–S174.

Smith, B.D., 2005. Reassessing Coxcatlan Cave and the early history of domesticated plants in Mesoamerica. Proceedings of the National Academy of Sciences of the United States of America 102, 9438–9445.

Smith, B.D., 2011. The cultural context of plant domestication in eastern North America. Current Anthropology 52 (S4), S471–S484.

Smith, B.D., Yarnell, R.A., 2009. Initial formation of an indigenous crop complex in eastern North America at 3800 B.P. Proceedings of the National Academy of Sciences of the United States of America 106, 6561–6566.

Wasylikowa, K., Barakat, H.N., Lityńska-Zając, M., 2001. Part IV: Nabta Playa Sites E-75-8, E-91-1, E-92-7, E-94-1, E-94-2, and El Gebel El Beid Playa Site E-77-7: seeds and fruits. In: Wendorf, F., Schild, R., Associates (Eds.), Holocene Settlement of the Egyptian Sahara, The Archaeology of Nabta Playa, vol. 1. Kluwer Academic/Plenum, New York, pp. 544–591.

Weiss, E., et al., 2008. Plant-food preparation area on an upper paleolithic brush hut floor at Ohalo II, Israel. Journal of Archaeological Science 35, 2400–2414.

Wills, W.H., 1995. Archaic foraging and the beginning of food production in the American Southwest. In: Price, T.D., Gebauer, A.B. (Eds.), Last Hunters, First Farmers: New Perspectives on the Prehistoric Transition to Agriculture. School of American Research Advanced Seminar Series. School of American Research Press, Santa Fe, NM, pp. 215–242.

Zeder, M.A., 2011. The origins of agriculture in the Near East. Current Anthropology 52 (S4), S221–S235.

Zhao, Z., 2011. New archaeobotanic data for the study of the origins of agriculture in China. Current Anthropology 52 (S4), S295–S306.

Zohary, D., 1996. The mode of domestication of the founder crops of southwestern Asian agriculture. In: Harris, D.R. (Ed.), The Origins and Spread of Agriculture and Pastoralism in Eurasia. UCL Press., London, pp. 142–158.

Food Security and 'Green Revolution'

William G Moseley, Macalester College, Saint Paul, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H.-G. Bohle, volume 8, pp. 5728–5730, © 2001, Elsevier Ltd.

Abstract

Food security refers to enough food at all times for a healthy and active life. In other words, countries, households, or individuals who are food secure are no longer at risk for hunger on a chronic, seasonal, or sporadically recurring basis. While some promote food security by making sure all households have adequate access to food, others mainly focus on increased food production as the best strategy. The 'Green Revolution' was such a production-based approach and is a term used to describe the changes that took place in farming in Asia and Latin America after the introduction of semidwarf wheat and rice varieties in the mid-1960s and 1970s. The term and approach have more recently been revived and reborn as the New Green Revolution in much of Africa since approximately 2005.

Food Security

Food security is commonly defined as "access by all people at all times to enough food for an active, healthy life" (World Bank, 1986: 2). While this definition makes no explicit reference to the spatial or social unit of analysis, others have focused on a particular scale. For example, Frankenberger and Goldstein (1990: 21) define the term to be a condition in which "[t]he viability of the household as a productive and reproductive unit is (not) threatened by food shortage." In these terms, food insecurity may be thought of as inadequate access to sufficient amounts of food during all or part of the year for a healthy and active life either through insufficient subsistence production or limited ability to acquire food through other means, for example, purchase, barter, or informal reciprocity arrangements. Increasingly, since the seminal work of Sen (1981), the various elements of food security are encapsulated in the idea of household food entitlements (defined as legitimate claims on food resources). Sen's insight reveals that national food audits, using the food balance sheet approach, do not necessarily correlate with household food security. The typical audit is couched in terms of an aggregate supply–demand problem in which national supply (production + stocks + imports) is compared to national demand (total population × caloric need per capita). Although not a useless exercise, national audits (where sufficient supply is often defined as national demand + 10%) (Geier, 1995) might satisfy national goals without accounting for the differing abilities of households to access food, even when it is available on the market. To phrase the problem more baldly, strategies designed to achieve national food goals do not necessarily and concomitantly achieve household and community food security. By contrast, it is reasonable to suppose that when household food entitlements are obtained, national food security goals would, of necessity, also be achieved.

Until the early 1980s, international experts had labored under the misperception that sufficient food supply – a function of homegrown production and net imports – was equivalent to food security. As such, it was argued, the best way to fight hunger was to boost agricultural production, as exemplified by the Green Revolution. United Nations experts monitored food supply in relation to the caloric needs of a country – known as the food balance sheet approach – and then assumed that all was well if the two sides were at least equal. History is riddled with examples of the poor dying of hunger when food was plentiful. Classic among these is the famine that wracked the West African Sahel during the early 1970s. While people were dying of hunger in Senegal, Mali, and Niger, peanuts – a key sauce ingredient and source of protein across the region – were being exported to Europe.

Sen's definition of food security, with its attention to access, dominated international food policy circles for about 15 years, from the late 1980s through the early 2000s. Then, after a 20-year hiatus, the global community's attention has subsequently shifted back to agriculture. The international consensus on 'food security as access' shifted to a technocratic focus on food production as the best way to solve global, and especially African food insecurity.

Early History of the Green Revolution

The term 'Green Revolution' is most commonly used in the social sciences to describe the changes that took place in farming in Asia and Latin America after the introduction of semidwarf wheat and rice varieties in the mid-1960s. If planted on irrigated land with the use of chemical fertilizers and pesticides, these crop varieties gave higher yields than traditional ones. These innovations led to a substantial increase in the yield of wheat and rice, which was sustained until the 1990s when the rate of increase of yield and output declined. The term was also applied to land reform in Eastern Europe in the 1930s and more recently to the attempts to persuade governments and business to adopt policies and practices that are not inimical to the environment.

The period immediately following World War II in South Asia and Latin America was characterized by rapid population growth, food shortages, and poverty (which were connected in the minds of the Green Revolution architects). A great variety of wheat and rice were in use, each adapted to their local microenvironment; yields were low with limited fertilizer and pesticide use. In 1966 a semidwarf variety of rice known as IR8, bred at the International Rice Research Institute in the

Philippines, was distributed among farmers in parts of India, while a semidwarf wheat originally bred in Mexico was also introduced. Where these high-yielding varieties (HYVs) were adopted, yields and output increased dramatically. This led W.S Gaud to state in a lecture in the United States in 1968: "this and other developments in the field of agriculture contain the makings of a new revolution. It is not a violent Red Revolution like that of the Soviets … I call it the Green Revolution" (quoted in Dalrymple, 1979). It is also clear that the Green Revolution was taking place in the ideological context of the post WWII Cold War. As such, part of the motivation for the Green Revolution was ideological – that is driven by Western fears that inadequate food supplies would lead to socialist political revolutions across the Global South (Patel, 2013).

The breeding of higher-yielding or disease-resistant varieties had been going on in Europe, North America, and parts of Asia since the rediscovery of Mendel's work at the beginning of the twentieth century, while the use of chemical fertilizers and applications of pesticides had been practiced since the 1920s in Western Europe, although adopted at far greater rate after 1945. More relevant were the developments in Japan. Much Japanese rice had long been irrigated, but the area irrigated was increased in the 1890s when the most successful existing rice varieties were selected and their seed distributed to farmers. Chemical fertilizers were also applied. By the 1930s, rice yields in Japan and its colonies, Korea and Formosa, were substantially above those elsewhere in Asia. Even in South Asia there had been advances in technology before 1966. Indian grain output increased by 62% before 1950 and 1964, and in Pakistan by 59% (Farmer, 1981). Although rice yields in China in the 1950s were low compared with the rest of East Asia, semidwarf rices were bred there independently of the research in the Philippines and distributed to Chinese farmers before IR8 was released.

The Nature of HYVs in the Green Revolution

The principal characteristic of the HYVs were their high response to chemical fertilizers. They were also short with a stiff straw, and so, unlike the tall traditional varieties, did not lodge with higher yields. They were insensitive to photoperiodicity and matured in about 110 days rather than 180 days; it was thus possible to grow two or even three crops in a year. The yield potential of these varieties was greater in the temperate regions of Asia and in the dry season in the monsoon region than in the humid tropics, because of the longer hours of sunshine and hence the greater potential photosynthesis available to the plant. The highest yields are achieved in areas of well-managed irrigation and with adequate amounts of chemical fertilizers. At first the new varieties were highly susceptible to disease, and pesticides were necessary to obtain the high yield potential, but later varieties have been bred with a greater resistance to pests. Breeding of new varieties of rice and wheat has continued since the 1960s, increasingly in research stations in the countries where they are to be grown. HYVs of maize have also proved successful and to a lesser extent new varieties of sorghum. Since the 1970s Chinese farmers have used hybrid rice seed. The new varieties introduced since the first HYVs are now often referred to as MVs (modern varieties).

The Diffusion of the Modern Varieties in the Green Revolution

The inputs of the Green Revolution are divisible: seed, fertilizer, and pesticides can be bought in small amounts and so in theory could be obtained by farmers with only very small holdings. In practice, however, the first adopters were generally occupiers of larger farms with better credit facilities and a greater awareness of the new technology. By the 1980s, however, there was near universal adoption in irrigated areas with no differentiation of inputs by farm size or tenure (although this may be an artifact of an unknown number of small farmers going under and leaving agriculture altogether). The principal incentive to adopt was the greater profit per hectare, in spite of the increased unit costs. At first the use of fertilizers and seeds were subsidized, and the improvement and extension of irrigation was undertaken by government. In the 1980s, however, subsidies began to be withdrawn, and the further extension of irrigation had become very expensive. The introduction of modern varieties began at different times in different countries and different regions. Rice was introduced into India, for example, in 1965 but not in South Korea until 1972. The rate of adoption also differed; by 1976 four-fifths of all the rice in China was sown with modern varieties, but only 7% in Burma, 11% in Thailand, and 14% in Bangladesh. National differences persisted in the 1990s.

Within countries there have also been major regional differences in the rate of adoption; thus in India in 1990, 95% of the rice in Punjab and Tamil Nadu was sown with modern varieties, but only one-third in Bihar and Kerala (Singh, 1997). By the 1990s some of the major international differences in the rate of adoption had been reduced, but a greater proportion of Asia and Latin America's wheat was sown with modern varieties than Africa or the Near East, while more of China's was in new varieties than the rest of Asia. But of course there is a difference in the proportion of each continent sown in wheat, rice, maize, and other crops. Little wheat or irrigated rice is grown in tropical Africa, so the proportion of these crops in modern varieties is of little significance. This distribution explains why, in the mid-1980s, modern varieties were planted on less than one-third of the total cereal area in the developing countries: only 1% in Africa, 22% in Latin America, and 36% in Asia (Tangley, 1987).

The Consequences of the Green Revolution

In the 1950s and 1960s it was feared that future population growth would outrun cereal output in Asia and increase malnutrition and hunger. This has not occurred. The rate of increase in output of both rice and wheat in Asia exceeded population growth rates between 1966 and 1995. Population increased 76% but rice output doubled; and the real price of rice fell steadily since the 1950s, halving between 1970 and 1990. Cereal consumption per capita has increased by 25%, imports have been greatly reduced, and grain reserves built up. The increase has varied between countries, and has been greater in wheat than rice output, but there have been dramatic increases in cereal harvests throughout the region. In the 1950s and 1960s increases in cereal output were largely due to an extension of the

cultivated area; since the 1960s, overwhelmingly due to the increases in crop yields following the adoption of the new HYVs.

The shorter growing season for HYVs has also allowed more double cropping in some parts of Asia. Yields have stagnated in the 1990s, however, particularly in Japan and Korea, which have the highest rice yields. The yield potential of the original HYVs has not been increased since the 1960s; the many new varieties bred since then have had greater disease resistance or better quality grain, but not greater yield potential. Additionally, in the high-producing irrigated rice areas, waterlogging, salinization, and disease have led to a decline in yields in some places. The fall in rice prices and the continued increase in import prices have squeezed cereal producers, and many have left farming or turned to more rewarding crops. In the last few years there have been renewed fears for the future of cereal output in Asia.

Criticisms of the Green Revolution

Initially the Green Revolution was greeted with extravagant praise and high expectations. These were excessive, for it was unlikely that many farmers could reach the yields obtained on experimental farms, or that all farmers could profitably use the new techniques. A barrage of criticism followed to the present, although with changing targets. Controversy remains over the social and economic effects of the Green Revolution. There are three technical reasons why interpretations may differ. First, national and state statistics in India and other parts of Asia are not reliable and so different interpretations of changes are possible, particularly with reference to land tenure, farm size, and labor. Second, it would be surprising if there were identical responses to new technology over such a vast area and population with greatly varying physical and socioeconomic environments. Third, research produces different results at different stages of the adoption cycle of agricultural innovations. It is unlikely that the consequences of technological change in the 1960s would be the same in the 1990s.

But there are other criticisms that are a result of the unequal distribution of farm sizes and land ownership. Many critics argue that the adoption of HYVs benefited the larger farmers most, and often led to increased landlessness and declining real wages. While those with larger farms (and in most of Asia more than 20 acres is a large farm and many farmers have less than 5 acres) were able to rapidly adopt the new technology and make considerable profits, smaller farmers were unable to do so and often sold their holdings to larger farmers. Many tenants were evicted to allow landowners to farm larger units. The number of landless households increased and so forced down real wages. In some areas the larger farmers adopted machinery to replace labor, and real incomes of the landless fell. Consequently income inequality increased.

Other critics pointed out that the HYVs were only profitable in irrigated areas, and that only improved varieties of wheat and rice had been bred. Furthermore, in many parts of India farmers grew other crops so the benefits of the 'Green Revolution' were limited to certain regions. The early HYV rices had an inferior taste, and although subsequent varieties were of higher quality, richer consumers still prefer the taste of traditional varieties. Other critics were alarmed that the inputs for the new farms were imported, while some believed that monoculture was replacing the diversity of cropping in traditional Asian farming, particularly the replacing of protein-rich pulses with cereals. There was rising problem of pesticide resistance among insect pest populations. This meant that increasingly large doses of pesticides needed to be applied, or new pesticide formulas developed altogether.

By the 1990s new fears replaced these traditional concerns. Yields were – and still are – stagnating and even declining as a result of waterlogging and poor drainage in irrigated areas, while chemical pollution was becoming a problem. Muldavin's (1997, 2000) work in China suggests that soil degradation related to overuse of fertilizers and pesticides is a growing problem.

Global Food Insecurity and the New Green Revolution for Africa

The problem of food insecurity remains one of the most serious challenges facing humankind. In 2012, by the UN Food and Agriculture Organization's conservative estimate, 868 million people experienced extreme chronic undernourishment, and micronutrient deficiencies affected around two billion worldwide (FAO, 2012). Sub-Saharan Africa suffers the highest prevalence of macro- and micronutrient malnutrition, and lags behind other regions in reducing hunger (Nature, 2010; IMF, 2012). In response to this problem, philanthropic foundations, industrialized country governments and aid agencies, international agricultural research centers, and African governments have come together in an unprecedented and highly coordinated effort to improve food security in Africa. Through initiatives such as the Rockefeller and Gates Foundations' Alliance for a Green Revolution for Africa (AGRA), the G8's New Alliance for Food Security and Nutrition, USAID's Feed the Future Program, and the African Union's Comprehensive African Agricultural Development Program (CAADP), these actors have committed billions of dollars to the goal of raising productivity among African smallholder farmers, on the assumption that increased yields are the key to food security, poverty reduction, and economic growth (Toenniessen et al., 2008). Equally important to this donor-led development strategy, popularly known as a 'New Green Revolution for Africa' (GR4A), is tighter integration of African smallholders into formal markets, both to improve their access to yield-boosting inputs and to encourage the sale of crop surpluses for cash (Annan, 2007). A more reliable and nutritious food supply and diminished hunger rates will, according to this approach, inevitably follow.

This new concerted effort to construct a market-based model for African agriculture constitutes a substantial shift in the way that international development actors organize their responses to hunger at the global level. Inspired by the technological revolution that transformed food production in Asia and Latin America in the 1950s to 1960s, but seeking to incorporate lessons learned from that experience, the GR4A focuses on 'scaling up' African smallholder participation in commercial input and output markets through the development of national, regional, and global agricultural value chains. This model, which relies heavily on public–private partnerships, is expected to have a profound impact on the continent's national and regional agricultural systems, local agroecologies,

and rural livelihoods (Gómez and Ricketts, 2013). However, its specific consequences for food and nutrition security – at the level of the household, and especially for women and children – remain unclear. Despite the centrality of food security goals to GR4A projects and discourses, much uncertainty remains around the policy, market, and social mechanisms that will translate increased crop yields and incomes into improved access to a nutritious food supply for all. Moreover, by conceptualizing food insecurity principally in terms of crop yields and market access, this model overlooks the many political, socioeconomic, and biophysical dynamics that contribute to hunger and malnutrition – or 'hunger vulnerability' (Bassett and Winter-Nelson, 2010).

Moreover, according to Sen (discussed earlier), the problem is that improving household food security involves more than just increasing food production. A food security approach that is sensitive to issues of 'access' targets the poorest of the poor who are habitually the most food insecure. Helping poor households in rural Africa feed themselves in an affordable manner means introducing low-cost, sustainable enhancements to farming. These improvements include intercropping or the mixing of more than one crop in a field, agroforestry, the blending of trees and crops, composting, and soil conservation measures. These reforms are critical because they enhance soil fertility and control pest problems without cash outlays for expensive chemicals and fertilizers (Snapp et al., 2010).

Agroecological alternatives to the GR4A approach may be seen in a variety of contexts, such as in West Africa where traditional subsistence farmers using plant associations and organic inputs have often fared better than those farmers heavily engaged in cash cropping with the use of chemical inputs (Moseley, 2005). Similarly, such farmers also did better than market-integrated farmers and consumers during the 2007–08 global food crisis (Moseley et al., 2010; Moseley, 2011). Other laboratories for agroecological alternatives have been Cuba following an economic crisis that forced it to become more food self-sufficient (Miller, 2007). Finally, Bezner-Kerr's (2012) work in northern Malawi (a country known for embracing inorganic fertilizer subsidies as a way to improve food self-sufficiency) shows the viability and advantage of such approaches – even when situated in a country dominated by a strong maize monocropping ethos.

Sadly, advocates of the GR4A are not pushing these types of alternative interventions. Rather, they envision a more capital-intensive approach to agriculture involving supply chains, increasingly large producers, agro-processors, expanding international markets, and farming with intensive, and often expensive, inorganic fertilizers, pesticides, and seeds. This view sometimes legitimizes long-term land leases by foreign entities, so-called land grabs, which are seen as a showcase for new production strategies. While this approach will likely improve agricultural production in sub-Saharan Africa, it will do little to improve household food security for the poorest.

See also: Agricultural Sciences and Technology; Agriculture, Economics of; Agroecology and Agricultural Change; Environment and Development; Environmental Planning; Food Security and 'Green Revolution'.

Bibliography

Annan, K.A., 2007. Remarks on the Launch of the Alliance for a Green Revolution in Africa at the World Economic Forum. Cape Town, South Africa.

Bassett, T., Winter-Nelson, A., 2010. The Atlas of World Hunger. University of Chicago Press, Chicago.

Bezner Kerr, R., 2012. Lessons from the old green revolution for the new: social, environmental and nutritional issues for agricultural change in Africa. Progress in Development Studies 12, 213–229.

Burmeister, L., 1987. The South Korean green revolution: induced or directed innovation. Economic Development and Cultural Change 35, 767–790.

Byerlee, D., 1996. Modern varieties, productivity and sustainability: recent experience and emerging challenges. World Development 24, 697–718.

Dalrymple, D.G., 1979. The adoption of high-yielding grain varieties in developing nations. Agricultural History 53, 704–726.

Evenson, R.E., Herdt, R.W., Hossain, M., 1996. Recent Developments in the Asian Rice Economy: Challenges for Rice Research. CAB International, Wallingford, UK.

FAO, 2012. The State of Food Insecurity in the World. Rome: Food and Agriculture Organization. http://www.fao.org/docrep/016/i3027e/i3027e.pdf.

Farmer, B.H., 1977. Green Revolution? Technology and Change in Rice-Growing Areas of Tamil Nadu. Westview, Boulder, CO.

Farmer, B.H., 1981. The green revolution in South Asia. Geography 66, 202–207.

Frankenberger, T., Goldstein, D., 1990. Food security, coping strategies, and environmental degradation. Arid Lands Newsletter 30, 21–27. Office of Arid Lands Studies, University of Arizona.

Geier, G., 1995. Food Security Policy in Africa between Disaster Relief and Structural Adjustment. Frank Cass, London.

Gómez, M.I., Ricketts, K.D., 2013. Food value chain transformations in developing countries: selected hypotheses on nutritional implications. Food Policy 42, 139–150. http://www.sciencedirect.com/science/article/pii/S0306919213000766.

Griffin, K., 1974. The Political Economy of Agricultural Change: An Essay on the Green Revolution. Macmillan, London.

International Monetary Fund, 2012. World Economic and Financial Surveys, Regional Economic Outlook: Sub-Saharan Africa, Sustaining Growth Amid Global Uncertainty. International Monetary Fund, Washington, DC.

Karim, M.B., 1986. The Green Revolution: An International Bibliography. Greenwood Press, New York.

Lipton, M., Longhurst, R., 1989. New Seeds and Poor People. Unwin Hyman, London.

Miller, S., 2007. An Environmental History of Latin America. Cambridge University Press (Epilogue: Cuba's Latest Revolution, pp. 229–336).

Moseley, W.G., Carney, J., Becker, L., 2010. Neoliberal policy, rural livelihoods and urban food security in west Africa: a comparative study of the Gambia, Côte d'Ivoire and Mali. Proceedings of the National Academy of Sciences of the United States of America 107 (13), 5774–5779.

Moseley, W.G., 2011. Lessons from the 2008 global food crisis: agro-food dynamics in Mali. Development in Practice 21 (4–5), 604–612.

Moseley, W.G., 2005. Global cotton and local environmental management: the political ecology of rich and poor small-hold farmers in southern Mali. Geographical Journal 171 (1), 36–55.

Muldavin, J., 1997. Environmental degradation in Heilongjiang: policy reform and agrarian dynamics in China's new hybrid economy. Annals of the Association of American Geographers 87 (4), 579.

Muldavin, J., 2000. The paradoxes of environmental policy and resource management in reform-era China. Economic Geography 76 (3), 244–271.

Nature, 2010. The growing problem. Nature (Online) 466, 546–547.

Patel, R., 2013. The long green revolution. Journal of Peasant Studies 40 (1), 1–63.

Pingali, P.L., Hossain, M., Gerpacio, R.V., 1997. Asian Rice Bowls. The Returning Crisis? CAB International, Wallingford, UK.

Sen, A., 1981. Poverty and Famines. Clarendon, Oxford.

Shiva, V., 1989. The Violence of the Green Revolution. Natray Publishers, Dehra Dun, India.

Singh, J., 1997. Agricultural Development in South Asia. A Comparative Study in the Green Revolution Experiences. National Book Organisation, New Delhi, India.

Snapp, S., Blackie, M., Gilbert, R., Bezner-Kerr, R., Kanyama-Phirie, G., 2010. Biodiversity can support a greener revolution in Africa. Proceedings of the National Academy of Sciences of the United States of America 107 (48), 20840–20845.

Tangley, L., 1987. Beyond the green revolution. Bioscience 37, 176–180.

Toenniessen, G., Adesina, A., DeVries, J., 2008. Building an alliance for a green revolution in Africa. Annals of the New York Academy of Sciences 1136 (1), 233–242.

World Bank, 1986. Poverty and Hunger: Issues and Options for Food Security in Developing Countries. World Bank Policy Study. Washington, DC.

Food, Sociology of

Alain Drouard, National Center for Scientific Research, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Food is a subject that seems to both slip away and to expand when one tries to understand it. We cannot simply analyze the production and consumption of food, of food diets, and of table manners, and their evolution in time and space. As food has been cooked for thousands of years, we need to integrate the notion of *cuisine*, a polysemous term with multiple significations, into the study of food. We also need to include in this domain of study the representations, discourses, and collective beliefs associated with the notion of food.

Recent advances in the social sciences – notably history, sociology, and anthropology – that deal with food reinforce this observation. While this article emphasizes the advances as well as the limits of knowledge acquired in this area in recent decades, it also seeks to take into account epistemological and methodological problems encountered in social science research in the domain of food, whether it be with the definition of food itself, with interdisciplinary difficulties, or with the synthesizing of the approaches of different disciplines dealing with a complex subject that has all the characteristics of what Marcel Mauss called a 'total social fact.'

Considered as a trivial, frivolous topic unworthy of interest for established science, the study of food long remained a domain of study limited to individual scholars, to literati, and to amateur enthusiasts.

However, we must nuance this statement of a general state of indifference to and ignorance of the subject of food, since in France – and to a certain extent also in other countries – some authors have, in their own way, been the early 'historians' and the first 'sociologists' of food, and have assigned to it an important function and a particular place in the lives of their characters and of their heroes. Balzac (1838) even went further: he did not simply describe the social and cultural role of the table place in the bourgeois society of the early nineteenth century, he defined the political aspect of food, as follows: "Food as a whole is one of the most important and most neglected parts of politics. I even dare say that it is in its infancy."

When we speak of food, what exactly are we talking about: a social fact? In this case, food would only be a subject of interest for the social sciences and the humanities. A biological fact? Then it would be in the domain of nutrition and dietetics. And if food was both a biological and a social fact, how could we articulate the biological and the sociological approaches? Finally, how could we ignore the political dimension of food inherent in the vocabulary itself? The word *régime* (food habits, diet) comes from the Latin '*regere*: govern, rule.' In ancient societies the king was the foster father of his people, and organized their food diet, i.e., the way food was produced and consumed. The word *régime* thus applies both to its political and nutritional aspects.

Much more than its ambiguity, it is the complexity of food that has long been a problem for specialists who sought to define it, whatever their discipline of origin. Replying to those who thought that its study fell within the scope of biology, Dr. Jean Trémolières, a biologist himself, pointed out (1975) that:

> "Our relationship with food, the most basic of all our relationships, remains on a completely subconscious level …" and he stressed that in consequence we need to combine different and complementary disciplinary approaches to understanding man's eating behavior. A behavior that is "very curiously established …. We do not sit down to table to rebuild depleted reserves. We sit down to table and we choose our food driven by psycho-sensory and symbolic motivations that are only linked to biochemical motivations by all that we think and feel."

For long the social sciences did not see nutrition as an object worthy of their attention. Thus at the end of the nineteenth century, as Louis Bourdeau pointed out (1894), historians still largely ignored the subject:

> […] Disdainful of the conditions of common life they (historians) dislike the unhealthy curiosity attached to this low aspect of life and remain in an ideal world, foreign to the needs of prosaic necessity. People eat little in their writings. Their heroes – like the novelists' – seem to be ethereal heroes, freed from vulgar necessities, who live off glory and passion…

And what is true for historians is true as well for sociologists. Indeed, one only has to refer to the writings of Emile Durkheim (1894) and the founding principles of this new discipline that he wanted to establish ("consider social facts as things" and "the causes of a social fact should be found in another social fact") to understand the reluctance and hesitation of the emerging discipline of sociology when dealing with nutrition.

Indeed, Durkheim considered initially that nutrition, drinking, and eating are too organic in nature to be considered as social facts: "each individual drinks, sleeps, eats and reasons, and it is in society's interest that these functions be carried out regularly. So if these facts were social, sociology would not have an object of its own and its scope would be confused with that of biology and psychology."

However, he then underlined that what characterizes social facts is that "they are ways of acting, thinking and feeling external to the individual, and which are endowed with a coercive power that they impose on him."

He proceeded to give drinking and eating as examples of social facts:

> We can also confirm by a typical experiment this definition of a social fact: it suffices to observe how children are raised … It is obvious that all education consists of a continuous effort to impose on the child ways of seeing, feeling, and acting that he would not have adopted on his own. From the start, he is forced to eat and drink at regular times … he is forced to be clean, quiet, and obedient; and later he is forced to take others into account and to be respectful of customs and established practice.

It is not surprising that sociologists considered nutrition as a minor topic or at least not as worthy as religion or art.

As Claude Grignon observed (1995):

> This lack of academic prestige undoubtedly comes partly from the fact that the social dignity of subjects tends to reflect on the value of objects, including, it seems, in the minds of some sociologists. With this point of view, sociology is lagging behind art: everyone now knows that a painting of an apple can be a masterpiece.

Another obstacle that hampered research for many years related to the plurality and the diversity of the scientific disciplines that studied nutrition. Each discipline having a specific object and methodology (ethnology considers primitive peoples, sociology studies modern developed societies) addressed only one aspect of such a complex problem, but nevertheless was not particularly disposed to cooperating with other disciplines. In spite of declarations of principle in favor of the synthesizing of disciplines or in favor of interdisciplinarity!

In fact, nutrition has only recently become a legitimate object of research for the social sciences and the humanities.

In France, the first investigations into the economic and social history of nutrition were carried out after World War I by the School of Annals. Lucien Febvre, Marc Bloch, and then Fernand Braudel (1967) and their disciples studied both the production and the consumption of food and also diverse food régimes and their variations in space and time. Marc Bloch and Lucien Febvre (1954) were thus the originators of research into the 'types of cuisine,' which delineate in France the frontiers between regions, which use butter, oil, and fat for cooking.

The publication in 1970 of a 'Cahier des Annales' (Casebook of the School of Annals) marked the beginning of the development of a history of nutrition. Food was perceived as a new domain of historical research to be explored. In addition to an original article by Marc Bloch on food used in ancient France, the volume included several articles on medieval food régimes, thereby highlighting the importance of economic and quantitative history in this domain.

The first French food historians primarily addressed issues of quantity, of supply, of famines and shortages, or of the price of food: the essential notion was then what 'food ration' could be determined for each social class in the time and space in question.

In France in the final decades of the twentieth century, the cultural history of food began to take precedence over its economic and social history, while opening up to other social sciences such as sociology, anthropology, psychology, and economics. Jean Paul Aron (1967, 1973) was one of the pioneers of this new approach even before the rediscovery of the work of Norbert Elias, a German sociologist and historian. In his book *Uber den Prozess der Zivilisation* (On the Process of Civilisation, 1939), which has lastingly influenced food historians, Elias had analyzed the processes by which a society gets 'civilized': i.e., by which it adopts practices and techniques to control corporal impulses. He was particularly interested in practices of politeness and courtesy, but also studied the formation of table manners in Western civilization. After the publication of the translation of his book into French in 1973 and 1975, historians – Jean-Louis Flandrin and his disciples, Léo Moulin (1988) – analyzed the history of tastes and dietary practices. The same orientation in other European countries (Mennell, 1985; Flandrin and Montanari, 1996) has resulted in a significant overall development of historical research (Becker, 2012) and in the publication of a first general synthesis (Flandrin and Montanari, 1996).

If food is a subject of research that is both hard to pin down and which expands as soon as one tries to identify it precisely, it is because it is a *total social fact*, to use the concept defined by Marcel Mauss (2007), which thus designated every social fact "which involves in some cases the entire society and its institutions and in other cases only a very large number of institutions […]. All these phenomena are both legal, economic, religious and even aesthetical and morphological," and which thereby constituted a major research challenge, namely the construction of a social science of nutrition.

Nutrition

In the word *alimentation* (nutrition), there is *aliment*, which does not have the same meaning in French as in English, and whose definition is problematic.

Aliment in French is what feeds and nurtures both body and spirit. It is a very old word, which first appeared in 1120, coming from the Latin noun *alimentum* and the Latin verb *alere*: to nourish.

In the plural form, *aliments* (which first appeared in 1690) designate the means of existence necessary for a human being to survive.

As for *alimentation* (nutrition), which first appeared in 1412: it designates the action of nutrition, whereas *aliment* designates a nutritional object.

Initially, as the *Littré* dictionary reminds us, the French used the term *viande*, which comes from the Latin *vivenda*, to designate the things necessary to sustain life.

It is worthy of note that one of the first cookbooks, written by Taillevent in 1375, was called *Viandier* (Laurioux, 1997).

In English, even though the word 'aliment' exists, it has long been replaced by 'food.' And Americans say 'food is fuel.' So there are two different concepts involved: food as an object, the product of science and technology, and food as a symbol of what helps man to survive. There are also two different cultural attitudes involved: on the one hand, that of catholic countries and, on the other hand, that of protestant countries.

Over and above the study of the production and consumption of foodstuffs, of their usages in many diverse and various food régimes, the history and the sociology of food are faced with many changes – in time and space – in the social and

symbolic significance of food: changes of names and changes in the foodstuffs themselves. Ancient authors did not write about the same food that we do. What were the tastes of those foodstuffs? With the industrialization and the globalization of modern food production, innovations have multiplied and new meat-based diets have replaced grain-based ones, which had remained stable for centuries.

In old times or new, food has been consumed by respecting rules and rituals, if only because of the sacredness of food in traditional civilizations (Varagnac, 1941). The activities that relate to the acquisition, preparation, or consumption of food have been considered as sacraments. Thus, in the era of Christian civilization, bread and wine were at the heart of the celebration of the worship of the cult and mystery of the Lord's Supper.

Paul Valéry (1933) wrote:

> Bread and wine are at the core of Christianity.
>
> Catholicism requires them. Bread, wine and the concept of substance.
>
> The basic operation that defines Catholicism is the change in substance of two products developed by the industry of man. Or bread and wine are wheat and vines and processes of baking and winemaking….

The same applies to the Last Supper, the last meal of Christ and the archetype of the meal represented hundreds of times in Western painting.

Food and how it is consumed – meals – are thus powerful cultural identity markers. Suffice it to recall the place of rice in Asian civilizations or of corn in pre-Colombian America.

Do we choose our food? When the individual chooses his food, he takes into account certain constraints of income, of socioeconomic status, of availability, and of preferences depending on his belief system (i.e., what is allowed by his culture) and on his reference group.

The question then merges with that of food régimes, of their evolution over time, of their diversity in space, and of their variations according to the socioeconomic conditions and cultures being considered. Since the research done by Le Play (1878) and Maurice Halwachs (1913), we know that the share of food in family budgets not only varies according to income but is also a marker of social status.

With regard to the production and consumption of food, very numerous sociological analyses of inequalities and their impact on health and well-being (such as the obesity pandemic in the world) have been undertaken. But however accurate they are, these empirical studies and statistics do not enable us to fully understand the relationships and interactions that individuals have with their food, because they do not sufficiently integrate analyses of dietary behavior.

This question is more important today than ever because the distinction between food production and food consumption has been put into question both by consumer movements like Slow Food, which view consumers as coproducers, and by studies on institutions such as the restaurant, which show that chefs are both producers and mediators of taste, and even themselves *gastronomes* ('gourmets').

Cooking

Food is also *cuisine* (the art of cooking). However, words that refer to cooking are not the same in different languages. In English, one distinguishes between *the kitchen* and *(the art of) cooking*; in German between *die Küche* and *die Kochkunst;* in French *la cuisine* means both the kitchen space and the preparation of food. The concept itself is polysemous, with several meanings.

Before sociologists began to analyze this domain, ethnologists and anthropologists tried to define the meaning of this activity that came into existence with the invention of fire and the cooking of foodstuffs.

Claude Lévi-Strauss (1965) saw in the passage from raw to cooked food and in the appearance of cooking the birth of "a language in which each society encodes messages that allow it to signify a part of what it is" or as "the manifestation of its structure unless, without knowing it, it is resigned to reveal its contradictions." He used a linguistic approach to draw up an inventory of the distinctive characteristics of different types of cooking that he called *gustèmes*, and to define binary oppositions. Moreover, he emphasized the prominence of fire, which transforms raw food into cooked food, establishing a separation between nature and culture. Considering cooking as a universal activity, he built a 'culinary triangle' underpinned by a double opposition between culture and nature and between processed and unprocessed food. Raw food, unprocessed, is opposed to cooked food, which is culturally processed, and to rotten food, which is naturally processed. Cooking methods are part of this triangle: something roasted relates to the raw category (it is put in direct contact with fire without intermediaries, and its cooking is rarely complete); something boiled relates to the rotten category (it deconstructs food in a similar way to decay, by the double mediation of water and container); and smoked food relates to the cooked category (it is a 'deep and slow' way of cooking, without any other intermediary than air). Prepared dishes, depending on their method of cooking, have different degrees of prestige in a constant relation from one society to another. Thus, the roast would in most kitchens be considered as a ceremonial dish to be served to guests, while boiled food, cooked in a pot, would be an intimate family dish for a closed circle of people. Because of the high level of abstraction of the 'culinary triangle' approach, analysis of actual practices sometimes reveals counterexamples, but we need nevertheless to acknowledge the opposition between different types of prepared foods that consistently appears within every society.

For the British anthropologist Mary Douglas, food is a code. She therefore seeks to 'decode,' to decipher the specificity and variability of cultural facts and worldviews of cultures, groups, or social classes. Assuming a universal need for taxonomic consistency, she presented in her famous book *Purity and Danger* (1971) an interpretation of the prohibitions of the Book of Leviticus in terms of classification anomalies. She showed that each natural milieu relates to a specific animal category: thus "in the firmament one finds winged two-legged poultry, in the water scaled fish swimming with fins, on earth four-legged animals that jump or walk." The sullied animals are the hybrids, those who do not fit into a category. Conversely, "only the animals that fully fit in their class are pure." She showed that in the Jewish culture, sanctity is associated with unity, integrity, and perfection, so

much so that "dietary prescriptions simply develop the metaphor in the same direction." Thus, for Mary Douglas, in general: "food taboos are rooted in the way we structure our overall experience of life."

On the other hand, cultural anthropology – particularly after the work of Margaret Mead (1943) – defines eating habits as "ways in which individuals or groups of individuals, in response to social and cultural pressures, choose, consume and exploit certain portions of available food supplies."

Closer to us, Claude Fischler (1990) formulated 'the principle of incorporation,' which declares that the absorption of food involves incorporating its properties on both a real and on an imaginary level. We become what we eat.

In the domain of cooking, history and sociology have progressively identified three distinct areas of study:

- The space called the kitchen. The kitchen as we know it today has not always existed: in the West it became a separate entity only in modern times. This space, separated from other rooms of the house or apartment, is also socially determined by the functions performed by its occupants, traditionally servants and women. It follows that the sociology of cooking is at the crossroads between history and the sociology of women and of the sexual division of labor.
- The techniques for preparing, cooking, and preserving food as well as their evolution over time: this covers the history and sociology of technology and of innovations.
- The men and women who work in kitchens: chefs, cooks, housewives, and amateurs. How can their activity be defined? Is cooking a science or an art? In reality, the history and sociology of culinary creations, or simply of culinary production, inevitably involves analysis of the status of cooks and its evolution throughout time (Drouard, 2004).

Representations, Discourse, and Symbols

Finally, food is the matter of representations and beliefs. It is symbolic. We do not only eat nutrients, i.e., proteins, lipids, or carbohydrates, but symbols as Jean Trémolières already declared (1975):

> Food, above and beyond its biochemical composition, becomes a source of pleasure and well-being by revealing the kind of sensations that we love and choose. Thus it becomes 'symbol'.

Hence, the central question of the choice of foodstuffs and our relationship with them. In every society, rankings and choices are made: there are 'pure' food choices and 'unclean' food choices, food that is allowed and food that is forbidden, food that we like and food that we dislike.

As Claude Lévi-Strauss said, food should not only be good to eat, but also good to think about (Lévi-Strauss, 1962). This passion for classification that produces the differentiation between what is clean and what is dirty, between what is pure and what is impure, between what is beautiful and what is ugly, and between what is masculine and what is feminine, is at work all over the world because it is at the root of food practices, of food prohibitions, and ultimately of food diets.

Hence also the discourse on the art of good eating. Gastronomy is one of the most important examples.

Based on *nomos* (law in Greek) and *gaster* (belly), *la gastronomie* (gastronomy) is the law or rule of the belly.

The word appeared for the first time in the four line poem by Joseph Berchoux "*La gastronomie ou l'Homme des champs à table*" ("Gastronomy or the Man of the Fields at Table") published in 1800.

It replaced the words *gourmand*, *glouton*, *goulu*, and *gourmet*, all synonyms for greedy or glutton, that had previously been used to qualify the art of eating or the sin of gluttony.

The term *gastronome* (gourmet in English) began to be used in 1803 and *gastronomie* (gastronomy) first appeared in the Dictionary of the French Academy in 1835.

Historians and sociologists have developed approaches to gastronomy that, while considerably overlapping, remain distinct.

What is gastronomy? A 'French passion' among others, to borrow the title of a book by Jean-Robert Pitte (1991). Is it *La Grande cuisine* or *La Haute Cuisine* as Priscilla Parkhurst Ferguson (1998) repeatedly (but incorrectly) claimed when she defined gastronomy on numerous occasions as "the pursuit of culinary excellence"?

In fact, even if they are often confused with each other and used interchangeably, the two terms *grande/haute cuisine* (fine cooking) and gastronomy do not have the same meaning.

The notion of a gourmet as someone in need of *bonne chère* (good food) can only be properly appreciated, as Pascal Ory wrote in 1998, in that he lives in an environment of gastronomic discourse and literature (completely distinct from culinary literature such as recipe books written by chefs and cooks) written for and supporting him.

Without minimizing the importance of the gastronomic discourse and of gastronomic literature, Alain Drouard (2010) sees in gastronomy a myth: a narrative based on collective beliefs and representations of the excellence of French cuisine and its centuries-long preeminence over other national cuisines.

Born after the French Revolution, this myth has accompanied the establishment of what he calls the system of contemporary French cuisine: a set of relations of dependence and interdependence between the three main groups involved: food critics, chefs, and lovers of fine cuisine.

It is a fact that there is no gastronomy without food critics; i.e., without both a discourse on the art of good food and a gastronomic literature composed of guidebooks and gourmet magazines.

There is also no gastronomy without 'gourmet' restaurants which serve *Grande Cuisine* or *Haute Cuisine* (Fine Dining) to a wealthy clientele: a luxurious and sumptuous cuisine, which was codified in the early nineteenth century by Antonin Carême and reviewed by Auguste Escoffier at the beginning of the twentieth century.

And there is no gastronomy without culinary artists. After the famous artists of the nineteenth century (Carême, Viard, Beauvilliers, Urbain Dubois, Jules Gouffé, and Escoffier), Michelin-starred chefs now carry forward the torch of French gastronomy.

Initially expounded by gourmets and cooks, the myth of French gastronomy grew and flourished in the nineteenth century before spreading over time into the general population and being disseminated abroad by French chefs.

Sociologists for their part focus on the production and the dissemination of both the culinary and the gastronomic discourses.

Thus the sociologist Priscilla Parkhurst Ferguson considers texts devoted to gastronomy as having themselves a function in the culinary system.

In line with the analyses of Pierre Bourdieu (1979), she studies the formation of an autonomous gastronomic domain in France in the nineteenth century as the product of culinary texts.

In her book *Accounting for Taste; The Triumph of French Cuisine* (2004), she argues that "the centrality of French cuisine in the Western world for the past 300 years is intrinsically linked to the extension and depth as well as the prestige and authority of the culinary discourse that it supports and thanks to which it is amplified, magnified, glorified and disseminated."

Conclusion

The history of food has developed with widespread input from other social sciences: anthropology, ethnology, and sociology, which stress the fundamental role of food and drink in the social process.

Today it is a field in full expansion in all of Europe and in North America: it seems to have acquired a real legitimacy in historical studies, being considered less systematically than before as a 'minor' topic. The creation in 2002 of the European Institute of the History and Culture of Food, located in Tours in France, has enabled the bringing together of many researchers from various disciplines – historians, sociologists, and anthropologists – for regular seminars and various publications, especially the journal *Food & History*, launched in 2003.

In Italy, the Slow Food movement has founded a University of Gastronomic Sciences, a private institution whose diplomas are now recognized by the Italian state.

The table place has now become a vantage point for the many-faceted scrutiny of human societies, as aptly described by Jean-Pierre Poulain in his book *Sociologies de l'alimentation* (Sociologies of Food) (2002).

If this field now sometimes seems to take on the aspect of a vast construction site, the current proliferation of research studies would, however, tend to rule out the prospect of a synthesis of knowledge or of a general sociology or history of food being achieved any time soon. The multiplicity of disciplinary approaches and techniques being used are in the framework of an increasing internationalization, even globalization, of research activities. In English-speaking countries, this whole field is now aptly referred to as simply 'Food Studies.'

See also: Alternative Food Movements; Altruism and Prosocial Behavior, Sociology of; Collective Memory, Sociology of; Consumption, Sociology of; Culture, Sociology of; Evolutionary Sociology; Food Deserts in US Cities; Food Preferences, Psychology and Physiology of; Food Production, Origins of; Food Security and 'Green Revolution'; Food in Anthropology; Geographies of Food: Politics, Place, Power Reprised; Human Ecology; Interactionism, Symbolic; Positivism, Sociological; Slow Food as a Social Movement; Sociobiology and Sociology: A New Synthesis.

Bibliography

Aron, J.P., 1967. Essai sur la sensibilité alimentaire au XIX e siècle. Cahier des Annales 25. Armand Colin, Paris.
Aron, J.P., 1973. Le mangeur du XIX e siècle. Laffont, Paris.
Aymard, M., Grignon, C., 1995. Le temps de manger. Editions de la Maison des sciences de l'homme, Paris.
Balzac, H. de, 1838. Traité des excitants modernes in Balzac, H. de, 1902 Œuvres diverses 2. Ollendorf, Paris.
Becker, K. (Ed.), 2012. Studia alimentorum 2003–2013. Une décennie de recherche. A Decade of Research. Food & History 10 (2).
Bourdeau, L., 1894. Histoire de l'alimentation. Félix Alcan, Paris.
Bourdieu, P., 1979. La distinction Critique sociale du jugement. Les Editions de Minuit, Paris.
Braudel, F., 1967. Civilisation matérielle, économie et capitalisme (XVe–XVIIIe siècle). Armand Colin, Paris.
Claflin, K.W., Scholliers, P. (Eds.), 2012. Writing Food History a Global Perspective. Berg Publishers, Oxford.
Douglas, M., 1971. De la souillure: essai sur les notions de pollution et de tabou Trad de (Purity and Danger: An Analysis of Concepts of Pollution and Taboo 1966) par Anne Guérin. Maspéro, Paris.
Drouard, A., 2004. Histoire des cuisiniers en France XIX XX e siècles. CNRS Editions, Paris.
Drouard, A., 2010. Le mythe gastronomique français. CNRS Editions, Paris.
Durkheim, E., 1894. Les règles de la méthode sociologique. Payot, Paris.
Elias, N., 1973. La civilisation des mœurs, tome 1 ; La dynamique de l'Occident tome 2. Calmann-Lévy, Paris. Edition originale, Über den Prozess der Zivilization (1939), Haus zum Falken, Basel.
Febvre, L., 1954. Enquêtes et problèmes Les graisses de cuisine usuelle. In: L'Encyclopédie française, Tome XIV. Société d'édition de l'Encyclopédie française, Paris, pp. 14–40.
Ferguson, P.P., 1998. A cultural field in the making: gastronomy in 19th-century France. American Journal of Sociology 104 (3), 597–641.
Ferguson, P.P., 2004. Accounting for Taste: The Triumph of French Cuisine. The University of Chicago Press, Chicago and London.
Fischler, C., 1990. L'homnivore. Odile Jacob, Paris.
Flandrin, J.L., Montanari, M. (Eds.), 1996. Histoire de l'alimentation. Fayard, Paris.
Halbwachs, M., 1913. La classe ouvrière et les niveaux de vie Recherches sur la hiérarchie des besoins dans les sociétés industrielles contemporaines. Félix Alcan, Paris.
Hémardinquer, J.J., 1970. Pour une histoire de l'alimentation Recueil de travaux présentés par Jean-Jacques Hémardinquer. Cahier des Annales 28. Armand Colin, Paris.
Kiple, K.F., Ornelas, K.C. (Eds.), 2000. The Cambridge World History of Food. Cambridge University Press, Cambridge, UK.
Laurioux, B., 1997. Le règne de Taillevent Livres et pratiques culinaires à la fin du Moyen-Age. Publications de la Sorbonne, Paris.
Le Play, F., 1878. Les ouvriers européens, 5 tomes, A. Mame et fils, Tours.
Lévi-Strauss, C., 1962. La pensée sauvage. Plon, Paris.
Lévi Strauss, C., 1965. Le triangle culinaire l'Arc, 26, Paris.
Mauss, M., 2007. Essai sur le don: forme et raison de l'échange dans les sociétés archaïques (1 e édition Année Sociologique seconde série 1923–1924). Presses Universitaires de France, Paris.
Mead, M., 1943. The Problem of Changing Food Habits. National Research Council National Academy of Sciences, Washington, DC.
Mennell, S., 1985. All Manners of Food: Eating and Taste in England and in France from the Middle Ages to the Present. Blackwell, Oxford.
Moulin, L., 1988. Les liturgies de la table Une histoire culturelle du manger et du boire. Albin Michel, Paris.
Ory, P., 1998. Le discours gastronomique français des origines à nos jours. Gallimard (Archives), Paris.
Pitte, J.R., 1991. Gastronomie française. Histoire et géographie d'une passion. Fayard, Paris.
Poulain, J.P., 2002. Sociologies de l'alimentation. Presses Universitaires de France, Paris.
Trémolières, J., 1975. Diététique et art de vivre. Editions Seghers, Paris.
Valéry, P., 1933. Rhumbs. Gallimard, Paris.
Varagnac, A., 1941. La nourriture substance sacrée. Annales d'histoire sociale 3, 22–30.

Forecasting

Kenneth C Land, Department of Sociology, Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Forecasting comprises of the formation of expectations about future states or processes of social entities, such as populations or subpopulations thereof, and associated entities, such as economies and political and social institutions. This article commences by describing and giving examples of empirical applications of four methods of forecasting that have been developed, and are widely used, in the social sciences: leading indicators, quantitative trend analysis, formal forecasting models, and qualitative (theoretical–institutional–scenario) analysis and forecasts. This is followed by a section on limitations of forecasting. All social science forecasts are subject to deterioration of their accuracy with increasing temporal distance from the initiation point of the forecasts. This is due to the complexity of social systems to which forecasting models and the data from which they are estimated are only an approximation, the evolving nature of such systems, and the fact that the forecasts sometimes are just elements in the social system to which some actors respond, thus changing their behavior. It is concluded that forecasts should incorporate prediction intervals that grow with increases in the forecast horizon and the forecasts should be updated frequently.

Introduction

Forecasting refers to the formation of expectations about future states or processes of specific historical entities. Forecasting involves the estimation or calculation of expected future events or developments from a conceptual model. At the most informal level, the conceptual model can be a heuristic model based on individuals' life experiences and the particular social, cultural, and economic circumstances in which they live, as in the formation of expectations about the routine activities and rituals to take place on a given day. At the other end of the spectrum, the conceptual model can be analytic, formal, and complex based on numerical measurements, equations, simulations, and projections.

The most widely known forecasts by social scientists are economic and demographic forecasts, whereas natural scientists are associated with weather, climate, earthquake, and ecological forecasts (Land and Schneider, 1987). But these disciplines produce forecasts of many other social and natural phenomena and there are interstitial topics, such as energy, transportation, and technological forecasting, to which both have contributed.

The subject of forecasting, encompassing both methods of forecast construction and substantive applications thereof, is very large and would require a large volume for adequate description. Accordingly, this article will be highly focused. The next section describes four widely used methods of forecast construction. This is followed by a section on limitations of forecasting.

For purposes of this article, forecasting will refer to the formation of expectations on outcomes of study in the social sciences for social entities, such as populations or subpopulations thereof, and associated entities, such as economies and political and social institutions. This excludes applications of forecasting tools in clinical settings wherein the focus is on expected trajectories of health or behavioral outcomes for individuals.

Within this specified context, forecasts can be made for single-outcome variables or multidimensional distributions of outcomes at a specific time or time horizon into the future from the date of the forecast. The methods of construction of forecasts range from combining simple extrapolations of past performance with human judgment to elaborate methodological formalisms based on universal laws or accounting identities, or some combination of lawful behavior, extrapolation, and judgment. The 'future' forecasted can be near or distant. The scope of forecasts can encompass the world's physical/social/economic systems or a single local phenomenon. The uses of forecasts range from their fulfillment of a statutory requirement for practical policy uses to an assessment of their explanatory value in a scientific sense.

Methods of Forecasting

To focus the exposition, four methods of forecasting that have been developed, and are widely used, in the social sciences and business will now be described: leading indicators, quantitative trend analysis and extrapolation, formal forecasting models, and qualitative (theoretical–institutional–scenario) analysis and forecasts. Examples of empirical applications of each method will be included in the descriptions.

Leading Indicators

The *leading indicator approach to forecasting* in the social sciences has been most extensively developed and applied to economic and business forecasting, due very much to the work of the National Bureau of Economic Research (NBER) and one of its founders Wesley Mitchell (1927). It is based on the view that market-oriented economies experience business cycles of expansion and contraction within which repetitive sequences occur and that these sequences underlie the generation of the business cycle itself (Lahiri and Moore, 1991). The *leading economic indicator* (*LEI*) *approach* focuses on finding repetitive sequences, explaining them, and using them to forecast emerging stages of the current business cycle.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10544-6

The LEI approach differs from efforts to construct structural models of the economy – econometric models (described in more detail below) – which do not differentiate business cycles from other economic fluctuations, except possibly seasonal variation (Lahiri and Moore, 1991). The LEI construct has no fixed requirements on the duration of expansions or contractions, their amplitude, or their scope because the observed sequences – leads and lags – can be quite variable over time and are often systematically different at peaks and troughs, and because cycle phases vary widely in duration. The LEI emphasis on cycles has contributed to its longevity and helped it to spread around the world. Early signals of recession or recovery are of high salience to the actions of business officers, policy makers, job seekers, and investors. In the 1930s and 1940s, a group of NBER economists, including Moses Abramovitz, Arthur Burns, Milton Friedman, Gottfried Haberler, Wesley Mitchell, and Geoffrey Moore, developed a system of leading, coincident, and lagging indicators of business cycles.

For the United States, the system is maintained and published monthly by the Conference Board, a global independent business membership and research association. The components of the Conference Board's Leading Economic Indicators Index are the following statistics:

1. *Average weekly hours* (*manufacturing*) – Average weekly hours is a leading statistical indicator for changes in unemployment, as adjustments to the working hours of existing employees are usually made in advance of new hires or layoffs.
2. *Average weekly jobless claims for unemployment insurance* – The Conference Board reverses the value of this component statistic from positive to negative because a positive reading indicates a loss in jobs.
3. *Manufacturers' new orders for consumer goods/materials* – Increases (decreases) in new orders for consumer goods and materials usually mean positive (negative) changes in actual production, which makes this a useful leading indicator.
4. *Vendor performance* (*slower deliveries diffusion index*) – Vendor performance is estimated from a monthly survey conducted by the National Association of Purchasing Managers (NAPM). This statistics measures the time it takes to deliver orders to industrial companies. Vendor performance leads the business cycle because an increase in delivery time can indicate rising demand for manufacturing supplies.
5. *Manufacturers' new orders for non-defense capital goods* – As stated above, new orders lead the business cycle because increases (decreases) in orders usually mean positive (negative) changes in actual production and perhaps rising demand. This statistic is the producer's counterpart of new orders for consumer goods/materials component.
6. *Building permits for new private housing units*. An increase (decrease) in this statistic is indicative of more (less) activity in housing construction.
7. The *Standard & Poor's 500* (*S&P 500*) *stock index* – The S&P 500 is considered a leading indicator because increases (decreases) in stock prices reflect investor's expectations for the future of the economy and interest rates.
8. *Money Supply* (*M2*) – The money supply statistic measures demand deposits, traveler's checks, savings deposits, currency, money market accounts, and small-denomination time deposits. M2 is adjusted for inflation by means of the deflator published by the federal government in the periodic Gross Domestic Product (GDP) report. Bank lending, a factor contributing to account deposits, usually declines when inflation increases faster than the money supply, which can make economic expansion more difficult. Thus, an increase in demand deposits is indicative of expectations that inflation will rise, resulting in a decrease in bank lending and an increase in savings.
9. *Interest rate spread* (10-year Treasury vs Federal Funds target) – The interest rate spread (often referred to as the yield curve) indicates the expected direction of short-, medium-, and long-term interest rates. Changes in the yield curve have been the most accurate predictors of downturns in the economic cycle. This is particularly true when the curve becomes inverted, that is, when the long-term returns are expected to be less than the short rates.
10. *Index of consumer expectations* – This is the only component of the leading indicators that is based solely on expectations. This statistic leads the business cycle because consumer expectations can indicate future consumer spending or tightening. The data for this component come from sample surveys conducted by the University of Michigan Survey Research Center and is released once a month.

Figure 1 contains graphical representations of the Conference Board's Leading and Coincident Economic Indices from January 1999 through May 2014. Periods of economic recessions are shaded in gray. It can be seen that the Leading Economic Index turns down, signaling the oncoming 2008–09 Great Recession well before the Coincident Economic Index, the downturn of which marks the official beginning of the recession. The Leading Economic Index also turns up, signaling an oncoming economic expansion well before the Coincident Economic Index, an upturn of which marks the official beginning of an expansionary period. At the international comparative level, the Organisation for Economic Cooperation and Development (OECD) in Paris, France, publishes leading indices monthly for 34 OECD member countries plus Brazil, China, India, Indonesia, Russian Federation, and South Africa.

The LEI approach to economic forecasting persists because of its ability to survive repeated evaluations of its performance (Lahiri and Moore, 1991, pp. 2, 3). This does not mean that all of the indicators remain unchanged forever. Indeed, the need to change the list of indicators in the LEI occurs every few years because new indicators and time series become available, because research has produced some new results, or because the economy has changed. In brief, two important aspects of the LEI are that they are easy to understand and every component indicator needs periodic review.

Quantitative Trend Analysis and Extrapolation

Quantitative trend analysis and extrapolation is a second major approach to forecasting. Among other empirical applications,

Figure 1 The Conference Board's Leading Economic Index, June 2014. Source: The Conference Board.

quantitative trend analysis and extrapolation is a part of the methodology of the *technical analysis* of stock market trends. In technical analysis, the developmental characteristics (e.g., gaps, reversals, trend lines of highs and lows, and trend channels) of prices of corporate stock equities traded in stock exchange markets, using both visual and quantitative properties, have been developed and applied for several decades (Edwards et al., 2013). There also is a tradition of *application of logistic and other S-curve models* to time trends of data in order to extrapolate/forecast population growth, life expectancy, the spread of a new disease such as AIDS, energy prices, the diffusion of innovations and new products/devices, and so forth (see, e.g., Modis, 1992).

In the social sciences more generally, *population projections* are a prototypical empirical application of quantitative trend analysis and extrapolation. For instance, the U.S. Census Bureau periodically produces projections of the United States resident population by age, sex, race, and Hispanic origin. These projections are produced using a *cohort-component method* that decomposes changes in the age–sex–race/ethnicity counts of the population obtained from a decennial census year into three components of change from year to year: *births* into the population at age zero, age-specific *deaths* to the population, and age-specific net international *migrants* (immigrants into the population minus emigrants from the population) into the population (U.S. Bureau of the Census, 2012). Application of this method to the production of annual age–sex–race/ethnic-specific population projections from a base year thus requires assumptions about yearly changes in these three demographic components of change (future births, deaths, and net international migration).

As a specific example, the Census Bureau National Projections for the years 2012–60 were produced using a cohort-component method beginning with an estimated base population for July 1, 2011 as follows (U.S. Bureau of the Census, 2012). First, components of population change (mortality, fertility, and net international migration) were projected. Next, for each passing year, the population was advanced 1 year of age and the new age categories were updated using the projected survival rates and levels of net international migration for that year. A new birth cohort was then added to form the population under 1 year of age by applying projected age-specific fertility rates to the average female population aged 10–54 years and updating the new cohort for the effects of mortality and net international migration. The assumptions for the components of change were based on time-series analysis of historical trends as follows.

Age-specific fertility rates were calculated and projected for women aged 10–54 years from birth registration data for 1989–2009, which were compiled by the U.S. National Center for Health Statistics (NCHS). The birth registration data were used in conjunction with the Census Bureau's Intercensal Estimates to produce a series of age-specific fertility rates by mother's race and Hispanic origin for five race- and Hispanic-origin groups: (1) non-Hispanic white, (2) non-Hispanic black, (3) non-Hispanic American Indian or Alaska Native (AIAN), (4) non-Hispanic Asian or Pacific Islander (API), and (5) Hispanic (of any race). Race and Hispanic origin was assigned

to projected births based on the race of the mother, the racial composition of men in the projected population, and the 2010 Census distribution of race and ethnicity of women and men with children less than 18 years of age in the household. Sex was assigned to projected births within each race- and Hispanic-origin group. The sex ratios (males per 100 females) of future births were set to equal the average of the sex ratios of births for the period from 1989 to 2009, within each of the five race- and Hispanic-origin groups. The age-specific fertility rates then were projected to 2060 by assuming convergence by 2100 of the age-specific fertility rates of all five race- and Hispanic-origin groups to the average age-specific fertility rates of the non-Hispanic white group for the years 1989–2009 (1.83 births per woman).

Only one series of mortality rates was projected for the 2012 National Projections. Mortality rates were calculated from NCHS-compiled death registration data for 1989 to 2009. These rates were used in conjunction with the Population Estimates Program's Intercensal Estimates to produce a series of mortality rates by age and sex for three race- and Hispanic-origin groupings. Mortality was projected based on projections of the life expectancy at birth (e_0) by sex. Changes in life expectancy at birth by sex were modeled assuming that the complement of the life expectancy (difference between an upper bound value, A, and life expectancy values) would decline exponentially:

$$C(t) = A - e_0(t) \qquad [1]$$

where $C(t)$ = the observed complement of life expectancy at birth at time t, A = the upper asymptote of life expectancy, and $e_0(t)$ = the life expectancy at birth at time t. The complement of life expectancy was then projected for future dates as:

$$\widehat{C}(t) = \widehat{C}(t_0)^{er(t-t_0)} \qquad [2]$$

where $\widehat{C}(t)$ = the observed complement of life expectancy at birth at time t, r = the rate of change in the complement of life expectancy at birth, and $\widehat{C}(t_0)$ = the model complement of life expectancy at time t_0. The parameters r, $\widehat{C}(t_0)$, and A were estimated simultaneously by minimizing the sum of squared errors (SSE) between the model and the observed values of life expectancy, by sex, for the years 1999 through 2009. It was assumed that the complement of life expectancy for each of the three race- and Hispanic-origin groups would change at the same rate as for the total country for each sex.

Projected values for the complement of life expectancy for each group for selected years from 2010 through 2060 were produced by assuming that the rate of change in the complement of e_0 is the same for each subpopulation as it is for the total country. Mortality rates by age were then produced using the most recent observed rates by sex- and race-origin group, the trajectory of life expectancy values, and an ultimate life table. To get an ultimate age pattern of mortality by sex, the United Nations' single age versions of the extended Coale and Demeny model life tables were used (United Nations, 2010). The West model mortality rates with life expectancy values of 87 for males and 91 for females were selected. Using the Coale–Demeny West model, age-specific central death rates were projected for each of the three race-origin groups by sex using a Census Bureau algorithm, which creates life tables for years that have intermediate life expectancy estimates by finding the interpolation factors for the most recent and next death rate inputs that would result in the desired life expectancy at birth value (Arriaga and Associates, 2003). The interpolation is done on the logarithms of the death rate values.

The 2012 National Projections include a Middle/Main series and three alternative series. These four series of projections provide results for differing assumptions for net international migration. The alternative series were based on assumptions of Low, High, and Constant levels of net international migration. The Constant series was produced by holding the level of net international migration from the Middle series for 2012 constant from 2012 to 2060. The High and Low series were produced by varying the level of net international migration of the foreign-born from the projection used in the Middle series, ±30%, respectively. All other methodology and assumptions used in the Low, High, and Constant series are the same as those used in the Middle series. The three alternative series are useful for analyzing potential outcomes of different levels of net international migration relative to the Middle series.

Table 1 presents the resulting projections of the resident population for 2012 through 2060 for the Middle series and the three alternative series based on Constant, High, and Low projections of international migration. In the Middle series, the population is projected to increase from 314 million in 2012 to 420 million in 2060. The Constant Migration series lowers the 2060 projected count to 393 million, the High Migration series raises this to 442 million, and the Low Migration series yields 398 million.

The 2012 Census Bureau National Population projections have been described and presented in some detail in order to illustrate an important application of quantitative trend forecasting in the social sciences. In addition to the National Population projections, the Census Bureau produces subnational projections for each of the US states. The cohort-component method recently has been generalized and extended to the projection of both national and regional populations by household structures/composition (Zeng et al., 2013, 2014). Key features of these demographic and other applications of quantitative trend analysis and projections are the application of relatively simple statistical models (e.g., minimum least-squares estimates of time trends) and the presence of an accounting identity that coordinates the various component projected series combined with considerable expert substantive knowledge of the subject matter of the forecast.

Formal Forecasting Models

Formal forecasting models, based on mathematical and statistical equations with empirically estimated parameters, are another widely used tool for forecasting. As noted by Land and Schneider (1987, pp. 9–12), forecasts in the social and natural sciences using formal models often are based on one or more universal laws or identities. An example is the classical accounting identity used on econometric models:

$$C_t + I_t \equiv Y_t \qquad [3]$$

which states that, by definition, aggregated consumption (C) for some collectivity (such as a nation) plus aggregate investment (I) is identically equal to aggregate income (Y) in

Table 1 Projections of population for the United States, 2015–60

Year	Middle series population estimates	Constant net international migration estimates	High international migration series	Low international migration series
2015	321 363	321 219	321 595	321 130
2016	323 849	323 606	324 200	323 497
2017	326 348	325 979	326 844	325 851
2018	328 857	328 335	329 524	328 191
2019	331 375	330 671	332 238	330 511
2020	333 896	332 981	334 983	332 808
2021	336 416	335 260	337 754	335 077
2022	338 930	337 504	340 548	337 312
2023	341 436	339 707	343 361	339 511
2024	343 929	341 866	346 190	341 669
2025	346 407	343 977	349 032	343 782
2026	348 867	346 036	351 885	345 848
2027	351 304	348 041	354 745	347 863
2028	353 718	349 987	357 611	349 824
2029	356 107	351 875	360 482	351 731
2030	358 471	353 704	363 358	353 584
2031	360 792	355 473	366 194	355 390
2032	363 070	357 185	368 991	357 149
2033	365 307	358 841	371 749	358 864
2034	367 503	360 443	374 471	360 536
2035	369 662	361 992	377 158	362 166
2036	371 788	363 494	379 816	363 759
2037	373 883	364 950	382 447	365 318
2038	375 950	366 365	385 054	366 846
2039	377 993	367 741	387 640	368 346
2040	380 016	369 081	390 210	369 821
2041	382 021	370 390	392 766	371 274
2042	384 012	371 670	395 313	372 710
2043	385 992	372 925	397 853	374 129
2044	387 965	374 158	400 391	375 538
2045	389 934	375 374	402 929	376 939
2046	391 902	376 574	405 471	378 333
2047	393 869	377 759	408 016	379 722
2048	395 841	378 934	410 570	381 110
2049	397 818	380 101	413 136	382 500
2050	399 803	381 262	415 714	383 892
2051	401 796	382 416	418 305	385 287
2052	403 798	383 566	420 910	386 686
2053	405 811	384 712	423 531	388 091
2054	407 835	385 857	426 168	389 502
2055	409 873	387 001	428 823	390 922
2056	411 923	388 146	431 497	392 350
2057	413 989	389 293	434 189	393 788
2058	416 068	390 442	436 900	395 236
2059	418 161	391 593	439 629	396 694
2060	420 268	392 746	442 374	398 160

Source: U.S. Census Bureau, Population Division, Release Date: December 2012.

any time period (t). The National Income and Product Accounts contain this and other identities.

Another example: Systems of national and regional demographic accounts – counts of stocks and flows of people among demographic and/or social statuses such as age and occupational categories (Long and McMillen, 1987) – lead to the population transformation identity:

$$I_{t-1,t} \equiv P_{t-1,t} O_{t-1,t} + B_t \qquad [4]$$

This identity says that a column vector of population inflows into a fixed geographical unit (such as a nation) from time period $t-1$ to period t ($I_{t-1,t}$) is identically equal to the product of an outflow transition proportions matrix from $t-1$ to t ($P_{t-1,t}$) times a column vector of population outflows (survivors) from $t-1$ to t ($O_{t-1,t}$) plus a column vector of births and net migrants in period t(B_t). In scalar form for total population aggregates, eqn [4] is the population accounting identity that lies beneath the population trends/projection analyses reviewed in the previous section. When members of the population are classified by demographic and/or social statuses (e.g., age, location of residence, employment status, occupational category, marital/union status, etc.), the full vector–matrix representation of eqn [4] is necessary.

From a physical science point of view, identities such as those in eqns [3] and [4] are specific instances of the law of conservation of mass. More generally, conservation laws are the natural science analog of social science accounting identities (Land and Schneider, 1987, p. 10). Both in natural science and in social science forecasting, formal models proceed by specifying parametric models/equations for one or more elements of the accounting identities. An econometric example is the classical Keynesian parameterization of the C_t in eqn. [3] as a linear function of current national income (Y_t):

$$C_t = a + cY_t \qquad [5]$$

where a and c are constants that can be estimated by simultaneous equation econometric methods. This parameterization, termed a behavioral equation, expresses the Keynesian specification that, in the short run such as a year, aggregate consumption is the sum of autonomous consumption (a) plus the current national income scaled by the marginal propensity to consume (cY_t). When combined with the specification that aggregate net investment (I_t) is fixed or autonomous in the short run, eqns [3] and [5] can be solved to determine C_t and Y_t (see, e.g., Brems, 1968).

The foregoing exposition illustrates *two key properties of the structural models approach to forecasting* (termed econometric models when applied to model the economy) to the development of formal forecasting models, be they of social or natural phenomena. First, the models often include one or more accounting identities/conservation laws. Second, the models incorporate parametric representations of one or more of the component elements of the accounting identities.

Illustrative of this, Pescatori and Zaman (2011) describe three broad categories of models that have been developed and used by the U.S. Federal Reserve Board to forecast the US and world economies over the past 50 years, each with its own strengths and weaknesses: structural, nonstructural, and large-scale models. S*tructural macroeconomic models* are built using the fundamental principles of economic theory, often at the expense of the model's ability to predict key macroeconomic variables like GDP, prices, or employment. *Nonstructural models* are *primarily statistical time-series models* (taking the form of Auto-Regressive Integrated Moving Average (ARIMA) for single time series or Vector Auto-Regressive (VAR) models for multiple time series or variations thereon; see, e.g., Granger and Newbold 1986; Woodward et al. 2012) – that is, they represent correlations of historical data that incorporate very little economic

structure, and this fact gives them enough flexibility to capture the force of history in the forecasts they generate. Indeed, it has been shown that the accuracy in various substantive contexts that the accuracy of single-period-ahead forecasts from ARIMA models is extremely difficult to beat (Nelson, 1973). In macroeconomic forecasting, these statistical time-series models are used to produce unconditional forecasts that generate the expected future paths of economic variables and provide the most accurate forecasts if the overall monetary policy regime does not change. The third category, *large-scale models*, is a kind of middle ground between the structural and nonstructural models. They are like nonstructural models in that they are built from many equations, which describe relationships derived from empirical data, and they are like structural models in that they also use economic theory, namely to limit the complexity of the equations. Their large size has the advantage that relationships can be selected from a huge variety of data series, making it possible to provide a thorough description of the economic condition of interest. For example, structural models rarely feature highly specific variables such as 'car sales,' while large-scale models often do.

This characterization of macroeconomic forecasting models, especially the distinction between structural models and nonstructural/statistical time series models, applies generally across forecasting in the social sciences. Briefly, structural models seek to represent behavioral and other structural relationships between exogenous variables determined outside the system being modeled and as well as the internal relationships among the endogenous or outcome variables of the system. By contrast, statistical time series models seek to model one or more time series of observations on the outcome variables in terms of their systematic properties purely as temporal phenomena. Articles on formal models of forecasting, both structural and time series, with applications to business, economic, demographic, and social phenomena are published in the *International Journal of Forecasting* and *Foresight: The International Journal of Applied Forecasting* by the International Institute of Forecasters (http://forecasters.org/), which was founded in 1981 for the objective of developing and furthering the generation, distribution, and use of knowledge on forecasting.

Qualitative Trend Analysis and Forecasts

A fourth general class of approaches to forecasting can be termed *qualitative trend analysis methods*. These methods pertain to the forecasting of general features of society and take the form of theoretical–institutional–scenario analyses. Daniel Bell's (1973) treatise, *The Coming of Post-Industrial Society*, on the transition in the late-twentieth century from an industrial societal form to a postindustrial form is a leading example. Bell posited that the United States in the late-1960s and 1970s was transitioning from an industrial form of society to a postindustrial form characterized by three key elements: (1) a shift from economy based on manufacturing to one based on services, especially the creation, processing, and distribution of information; (2) the increasing importance of theoretical knowledge and centrality of new science-based industries; and (3) the rise of new technical elites and the advent of a new principle of stratification that emphasize the dominance of the professional and technical classes – and that other developed societies would follow a similar transition. Bell's analysis was based, in part, on quantitative data on employment trends in the mid-twentieth century United States, but it largely was a social forecast about a change in the *social framework* of society "… the structures of the major social institutions that order the lives of individuals in a society" (Bell, 1973, p. 8), especially the distribution of persons by occupation, the education of the young, and the regulation of political conflict.

A decade later, John Naisbitt (1982) published *Megatrends*, in which he identified 10 major trends shaping the evolution of American society in the last quarter of the twentieth century. These included transitions from a national economy to a world economy, from hierarchical forms of organization to networking, and from either/or personal choices (e.g., either we get married in a male/female union or we do not) to multiple options (many forms of domestic unions from traditional heterosexual marriage to cohabitation to serial marriages/unions to same sex marriage).

The qualitative, institutional–structural transitions forecasted by Bell and Naisbitt have been highly evident in societal evolution over the decades since 1970. These forecasts are part of the innovation and development of a field of scholarly inquiry termed futurology or futurism. Just as the International Institute of Forecasters was established to further the development of formal forecasting models and methods, the World Future Society (http://www.wfs.org/) and its journal, *The Futurist*, was established in 1966 to foster the development of ideas about the future, including forecasts (especially qualitative forecasts), recommendations, scenarios, and alternatives.

The body of qualitative trend analyses and forecasts produced by futurists over the past four or five decades is huge, of which the examples cited here are only illustrative. The methods of analysis used in futurist studies also are quite diverse. Bell (1973), for example, based his projections of the postindustrial society very much on a theoretical–conceptual analysis of changes in the nature of economic production and the implications this would have for occupational structures and for the educational system and its relation to occupational success and stratification systems. This is illustrative of one class of approaches to futurist studies, namely, a *theoretical–conceptual–institutional analysis of implications of technological, demographic, or political changes*. Related to this is the use of *possible scenarios of futures* that might develop as a consequence of scientific technological advances, demographic, or political changes; an example is Mayes (2014) in which possible scenarios for recent developments in synthetic genomics/biology (= biology + engineering/editing of genomic structures) are sketched. A third approach is that on which Naisbitt's (1982) research was based, namely the *content analysis* of the verbal content of more than 2 million articles published in some 6000 local US newspapers during a 12-year period. Content analysis is a method of identifying specified characteristics of messages that has its roots in US intelligence efforts in World War II, led by the sociologist Paul Lazarsfeld and the political scientist Harold Lasswell, to obtain information on enemy nations from their published newspapers. Since the 1970s, qualitative trend followers such as Naisbitt have made extensive use of this method. With the

development and widespread use of the Internet, similar content-based trend-identification tools have also been developed by providers of major search engines. An example is *Google Trends* (http://www.google.com/trends/), which tracks how often a particular search term is entered relative to the total search volume across various regions of the world and in various languages.

Limitations of Forecasting

All demographic, economic, and social forecasts have limitations with respect to accuracy. Demographic forecasts of national and other large populations, as noted above, deal with movements of demographic processes of fertility, mortality, net migration, and multistate status changes from time period to time period. These processes change relatively slowly. Even so, demographers consider population projections for up to a decade or so beyond the initiating time period to have informative forecasting information, whereas projections beyond this time horizon are considered to be scenarios that may or may not develop (Zeng et al., 2014).

While the time horizon for macroeconomic forecasts typically is shorter (monthly or quarterly) than that for annual population projections, their limitations are even stronger. As Pescatori and Zaman (2011) noted, it is unlikely that macroeconomic models will ever provide perfectly accurate forecasts, because forecasts are ultimately just another variable in a system in which there are many rational actors (e.g., managers of production and distribution corporate units, bank managers, investors) for whom an economic forecast may change their behavior. In fact, the economic actors who attend most closely to forecasts are the people whose behavior is most likely to affect the future course of the variables forecasted. Both the structural and nonstructural varieties of economic forecasting models have been extensively developed and improved since the 1970s. However, most models, including the Federal Reserve Board's models (Greenspan, 2013, p. 7), failed to predict the financial crisis of September 2008 that led to the Great Recession of 2008–2009. This failure may be partly attributed to the models' failure to fully incorporate the growing role of the financial sector or the worldwide financial and trade linkages that globalization has generated. A key point is that economic systems evolve and change over time and it is difficult to incorporate these changes into the macroeconomic models until a forecasting crisis has raised their salience. Even so, the lack of forecasting ability does not prevent models from being useful devices that can help policy makers in making decisions.

Similar considerations apply to the accuracy of forecasts based on qualitative trend analyses. That is, while the accuracy of qualitative forecasts is more difficult to assess because of their more ephemeral nature (the author knows of no existing large-scale systematic assessment), it likely is the case that some such forecasts hit the mark, some do not, and some work temporarily but lose their accuracy as the system being forecasted evolves. For instance, of the 10 major trends shaping the evolution of American society in the last quarter of the twentieth century identified by Naisbitt (1982), a fair assessment is that 5 have evolved and lasted in the long-run (Industrial Society > Information Society, National Economy > World Economy, Hierarchies > Networking, More Development in the north United States > More Development in the south United States, and Either/Or > Multiple Options), while 5 may have had some staying power in the 1980s but have not been as widespread or lasted as strongly or long (Forced Technology > High Tech/High Touch, Short-Term > Long-Term, Centralization > Decentralization, Institutional Help > Self-Help, Representative Democracy > Participatory Democracy).

To understand these limits to forecasting accuracy, it is helpful to recall Land and Schneider's (1987, pp. 17–19) analysis of the accuracy of weather and social science forecasting. Briefly, weather forecasts are based on nonlinear partial differential equations of hydrodynamic flow that depend continuously on both boundary and initial conditions. To produce accurate forecasts of indefinitely long range from such equations would require perfect boundary specifications and an observational net of weather stations finer than the radius of the smallest atmospheric eddy; otherwise, small weather events 'hidden beneath the net' eventually become amplified, with time into the forecast horizon from the initiation point, into larger uncertainties by inherent instabilities in the nonlinear system. In brief, the finite resolution of observations on weather variables (such as temperature, barometric pressure, relative humidity, wind speed) is due to the fact that atmospheric measurements are available at only a finite number of grid points at which there are weather stations. Forecasts then are based on initial conditions that assume that the weather variables for points between stations are equal to the average or interpolated values between the stations. Because the interpolated values of the initial conditions on the grid often deviate from atmospheric microevents (e.g., thunderstorms) that lie within the grid, the resulting 'errors in initial conditions' in weather forecasting models ultimately places limits on their accuracy. Observational and theoretical studies have found that the limit of weather predictability is about 2 weeks (Somerville, 1987).

In the social sciences, corresponding limits to forecasting accuracy are due to the fact that, no matter how much disaggregation and cross-classification is introduced into the forecasting models and data, at some point a statistical average value must be estimated for each subcategory or subpopulation. But such an average, even if it is not subject to measurement or sampling error, likely ignores some population heterogeneity that is 'hidden' relative to the specified resolution of the model. This, combined with the property that social science models often are based on linear approximations to underlying nonlinear, multilevel interactive mechanisms, produces forecast errors that, again, become amplified with time into the forecast horizon from the initiation point. Accordingly, it is not surprising that demographic, economic, and social forecasts, whether quantitative or qualitative, have errors that deteriorate with an increasing forecast horizon from the point of initiation.

In brief, limits to forecasting accuracy are the reason that forecasters emphasize the need to incorporate uncertainty into forecasts (Goodwin, 2014). Uncertainty can be due to several sources, including uncertainty in the structure of models, errors in measurement of variables, and errors in estimation of model parameters. To the extent possible, therefore,

forecasts should incorporate prediction intervals that grow with increases in the forecast horizon, such as the high and low bounds on the U.S. Census Bureau National Population projections described above, if possible, these intervals should be based on probabilistic analyses, and the forecasts should be updated frequently.

See also: Data Bases and Statistical Systems: Applied Social Research; Demography: History Since 1900; Economics, History of; Population Forecasts; Social Changes: Models; Structural Equation Modeling; Time Series: General.

Bibliography

Arriaga, E.E., Associates, 2003. Population analysis with microcomputers. In: (Extract B) Software and Documentation. Rural–Urban Projection (RUP) Program, vol. II. U.S. Census Bureau, U.S. Agency for International Development, and United Nations Population Fund, Washington, DC.

Bell, D., 1973. The Coming of Post-Industrial Society. Basic Books, New York.

Brems, H., 1968. Quantitative Economic Theory: A Synthetic Approach. Wiley, New York.

Edwards, R.D., Magee, J., Bassetti, W.H.C., 2013. Technical Analysis of Stock Trends, tenth ed. CRC Press, New York.

Goodwin, P., 2014. Getting real about uncertainty. Foresight: The International Journal of Applied Forecasting 33, 4–7.

Granger, C.W.J., Newbold, P., 1986. Forecasting Economic Time Series, second ed. Academic Press, New York.

Greenspan, A., 2013. The Map and the Territory: Risk, Human Nature, and the Future of Forecasting. The Penguin Press, New York.

Lahiri, K., Moore, G.H., 1991. Introduction. In: Lahiri, K., Moore, G.H. (Eds.), Leading Economic Indicators: New Approaches and Forecasting Record. Cambridge University Press, New York.

Land, K.C., Schneider, S.H., 1987. Forecasting in the social and natural sciences: some isomorphisms. In: Land, K.C., Schneider, S.H. (Eds.), Forecasting in the Social and Natural Sciences. D. Reidel Publishing Company, Boston, pp. 7–31.

Long, J.F., McMillen, D.B., 1987. A survey of census bureau population projection methods. In: Land, K.C., Schneider, S.H. (Eds.), Forecasting in the Social and Natural Sciences. D. Reidel Publishing Company, Boston, pp. 141–177.

Mayes, R., 2014. Where will the century of biology lead us? The Futurist 48, 26–30.

Mitchell, W., 1927. Business cycles: the problem and its setting. National Bureau of Economic Research Studies in Business Cycles (1).

Modis, T., 1992. Predictions: Society's Telltale Signature Reveals the Past and Forecasts the Future. Simon & Schuster, New York.

Naisbitt, J., 1982. Megatrends: Ten New Directions Transforming Our Lives. Warner Books, New York.

Nelson, C.R., 1973. Applied Time Series Analysis. Holden-Day, San Francisco.

Pescatori, A., Zaman, S., 2011. Macroeconomic Models, Forecasting, and Policy-making. Federal Reserve Bank of Cleveland, Cleveland, OH.

Somerville, R.C.J., 1987. The predictability of weather and climate. In: Land, K.C., Schneider, S.H. (Eds.), Forecasting in the Social and Natural Sciences. D. Reidel Publishing Company, Boston, pp. 239–246.

United Nations Population Division, 2010. World Population Prospects: Extended Model Life Tables. United Nations, New York.

U.S. Bureau of the Census, 2012. Methodology and Assumption for the 2012 National Projections. U.S. Census Bureau, Population Division, Washington, DC.

Woodward, W.A., Gray, H.L., Elliott, A.C., 2012. Applied Time Series Analysis. CRC Press, New York.

Zeng, Y., Land, K.C., Wang, Z., Gu, D., 2013. Population and household projections at sub-national levels: an extended cohort-component approach. Demography 50, 827–852.

Zeng, Y., Land, K.C., Gu, D., Wang, Z., 2014. Household and Living Arrangement Projections: The Extended Cohort-component Method and Applications to the U.S. and China. Springer, New York.

Foreign Investment: Direct

Klaus E Meyer, China Europe International Business School, Pudong, Shanghai, The People's Republic of China

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S. Bagchi-Sen, volume 9, pp. 5731–5733, © 2001, Elsevier Ltd.

Abstract

Foreign direct investment (FDI) is defined as the ownership of assets in one country by the resident of another country. Studies explaining the upsurge of FDI in the post-World War II era focused on US multinationals and their worldwide operations in manufacturing industries. As a result, initial theoretical discussions emphasized the behavior of firms belonging to oligopolistic industries. The growth of FDI from Japan and the newly industrializing nations, as well as FDI in service industries, called for an eclectic approach toward explaining the dynamics of FDI. Other studies on FDI emphasize the location patterns of FDI within a host country and the nature of the impact of FDI on financial markets, technology, employment, competitiveness, and trade.

Foreign direct investment (FDI) is defined as investment into business units in another country with an equity stake sufficient to influence the strategy of the foreign business. It is distinguished from foreign portfolio investment (FPI), which is conducted solely with financial aims as investors take small equity stake in firms abroad without direct involvement in the business. Following OECD guidelines, most national statistics and scholarly studies use 10% equity as a threshold whether a foreign investment is classified as FDI or FPI (OECD, 1996). Firms undertaking FDI and hence operating business in multiple countries are known as multinational enterprises (MNEs).

Firms can undertake FDI in many forms (Peng and Meyer, 2011: chapters 6 and 12). The most common distinctions are between acquisitions and greenfield, that is whether a local firm has been taken over or a local operation has been built from scratch 'on a greenfield' (or nowadays more likely by renting few offices). The other key distinction is by the level of ownership. Foreign investors can take full ownership to establish wholly owned subsidiaries, or they can share ownership with other investors. Among shared equity arrangements, the most common are firstly the creation of a joint venture with a local partner, that is a new entity is created in which a foreign and a local partner hold equity, and secondly partial acquisitions, where a foreign investor acquires a stake in an existing local firm.

Trends of FDI

The volume of FDI is hard to measure because of definitional and measurement issues. The most commonly cited data are based on balance of payments statistics, and capture the financial flows between associated firms, with some corrections for issues such as retained profits. There are a number of methodological limitations to these data; for example, they do not capture projects for which a foreign investor raises capital in capital markets of the host country; but they do include capital flight into firms based in tax havens. However, at high levels of aggregation, these data outline the broad trends in the recent history of FDI (UNCTAD, various years). The annual flows of FDI show a highly volatile pattern that is strongly associated with business cycles. This is not only because in boom periods firms have more capital to invest and more ambitious expansion plans, but the values of acquired firms are inflated by high stock market valuations.

Since the 1970s, there has been a massive shift in the countries of origin of FDI. In the 1970s, the USA accounted for about 54% of global FDI flows, followed by the UK with 12%. At the time, developing or emerging economies accounted for less than half of a percentage of global FDI firms. In the next decades, Japanese and West European multinationals gained prominence while the US–UK dominance declined. For example, Japanese firms surged in the late 1980s and accounted for over 20% of new FDI. German and French companies have in the 1990s exceeded 10% of global flows in several years in the year 1990. In the mid-2000s, a new trend emerged with multinationals from emerging economies, especially China and Russia entering the global stage. By the year 2012, their share had increased to over 20% while the US accounted for 23% and the UK for 6%.

With the diversity of countries of origin, the diversity of FDI also increased in other dimensions. Early investment was often in the form of market-seeking investment by manufacturing businesses between industrialized economies. By the early 1990s, FDI was also frequently used to establish global value chains exploiting variations in human capital and production costs. By the 2000s, emerging economies had also become a major destination not only for market-seeking FDI but for businesses aiming to reduce their production costs. Following liberalization of service sectors in many countries, service sector businesses such as banks and insurances, construction firms, or e-commerce businesses joined manufacturing and primary sector firms as major investors. With the diversity of origins, their ownership pattern also changed: while the typical investor of the 1970s was listed on a stock exchange in New York or London, investors in the 2000s included many firms that are controlled by family business groups or are owned by entities of the state.

Motivations of FDI

The motives inducing firms to engage in FDI are commonly categorized into four types: market seeking, efficiency seeking,

resource seeking, and natural resource seeking (Dunning and Lundan, 2009). The vast majority of foreign investors are market seeking and aim to sell their products or services beyond their home market. Some market-seeking investors establish local production sites, acquire distribution channels, or service infrastructure (for example, in telecommunications). However, most market-seeking operations are very small as they may entail only a sales office or representation. These only account for a very small part of the financial flows reported for FDI. Firms typically start their international operations with such market-seeking investment in neighboring countries, and at later stages of their internationalization establish operations in more distant countries.

Efficiency-seeking FDI aims to enhance the operational efficiency of a business operation, for example, by lowering production costs, in particular labor costs, or by exploiting economies of scale. Efficiency-seeking FDI thus frequently uses lower cost production sites to serve international markets, for example, by locating a production base in Eastern Europe to serve Western Europe, or by using East Asia as a global production hub. Natural resource-seeking investment aims to explore natural resources such as oil, gas, coal, or minerals for sale in global markets or in the investor's home country. For example, oil majors have invested in emerging economies around the world seeking scarce resource deposits. Because of the value of such natural resources, such investments tend to involve relatively large capital investments (and hence a large entry in the balance of payments), but they are also more likely than other types of projects to trigger political objections.

Strategic asset-seeking investment is defined as the acquisition of assets such as technologies and brands overseas with the aim to not only compete in the host country, but to use the acquired assets to enhance competitive positions in global markets, and in the home market. Such strategic investments are typical for firms from countries aiming to accelerate catch-up with the global technological edge, such as Japanese in the 1970s, Koreans in the 1990s, and Chinese in the 2010s (Meyer and Thaijongrak, 2013).

Location of FDI

FDI enters virtually in all countries of the world; some developing countries and socialist countries used to prohibit most forms of FDI inflow, but such restrictions have largely been phased out since the early 1990s (with a minor trend reversal after the 2008 financial crisis). However, different types of investors seek different locations, dependent on their investment motives and their abilities to manage overseas operations. Market-seeking FDI tends to focus on large and fast growing markets. Traditionally, the most sought after markets were in industrialized economies, but the fast growth of emerging economies attracted an increasing share of investors in search of new markets. Efficiency-seeking investment traditionally had a North–South dimension to it (with West–East added in Europe since the 1990s), but even that has become more diverse. For example, Indian or Chinese companies may invest in selected activities such as sourcing of high-tech components in Europe or North America to enhance the quality of products they sell both globally and in their home markets.

Natural resource-seeking investment is focused on specific resources available only at select locations around the world. In consequence, natural resource seekers account for many of the largest FDI projects in countries with unattractive local markets or high economic or political risks. For example, many Chinese investments in Africa in the 2000s and 2010s are motivated by the quest for resources ranging from oil and gas, to minerals, and to farm produce; in fact the Chinese government provides ample support for such investments because securing access to resource is a prime objective of national economic development policy. In contrast, strategic asset-seeking investment tends to focus on technology clusters in the developed economies such as California or South-West Germany, where catch-up companies aim to embed themselves in local business communities to learn and to enhance their own capabilities.

While globalization opened opportunities to engage in FDI in distant parts of the world, the overall location pattern of FDI remains strongly influenced by economic, geographic, and cultural distance. Especially in early stages of their internationalization, firms tend to invest in countries close to their country of origin where they face fewer obstacles to doing business. Once they build experiential knowledge to operate in the global economy, they become more capable of managing larger commitments and operate in more distant countries (Meyer and Thaijongrak, 2013). Even among the largest multinationals, most have the bulk of their operations in one particular region of the world, with sales to other regions accounting for only a small part of their revenues (Rugman and Verbeke, 2004).

Impact

FDI influences its host economies in multiple ways, both positively and negatively. While these influences have been widely debated in the scholarly literature (for a review, see Meyer, 2004), they offer no simple answers as to whether host countries benefit from their presence. Many FDI projects generate positive effects, while some generate negative effects. Yet, the conditions that would attract positive or negative effects remain an issue of extensive political discussions and scholarly analysis.

FDI can benefit the local economy, for example, through creating employment in areas of high unemployment, by taking over bankrupt local firms, or by introducing new technologies that ultimately also benefit local business partners – or even unrelated firms observing technology in action. However, foreign investors may also crowd out local firms when they introduce more efficient and labor-saving production processes that lead local firms to surrender. Problematic in terms of benefits to the host economy can be projects where foreign investors attain a quasi-monopoly position in the host economy and can extract rents, or when they are protected from competition by poorly designed or enforced regulations.

Foreign investors also introduce new standardization practices on issues such as environment, labor, or transparency. Since multinationals tend to prefer standardized practices across their global operations and they are subject to more scrutiny by NGOs and by the media, their standards tend to be higher than those of local firms. However, the standards employed in

developing countries are often lower than those in the home country, which leads to critique from NGOs. Elsewhere, labor-intensive operations may not respect relevant legal standards, which triggers critique by NGOs; though these cases are typically associated with local firms or with outsourced manufacturing rather than with FDI. The ethics of international operations, as FDI or in other organizational forms, is complex and cannot be discussed here comprehensively.

Summary

Firms establish FDI projects overseas to expand the scope of their operations, to reach new markets, to enhance the efficiency of their overall operation, reduce their production costs, and to tap knowledge pools or natural resources overseas. FDI thus is a major vehicle through which MNEs contribute to the economic integration of the world economy. Foreign investors make important contributions to host economies, which may normally benefit the host economy, but in some incidences may also be harmful.

Bibliography

Dunning, J.H., Lundan, S.M., 2009. Multinational Enterprises and the Global Economy, second ed. Elgar, Cheltenham.

Meyer, K.E., 2004. Perspectives on multinational enterprises in emerging economies. Journal of International Business Studies 34 (4), 259–277.

Meyer, K.E., Thaijongrak, Ornjira, 2013. The dynamics of emerging economy MNEs: how the internationalization process model can guide future research. Asia Pacific Journal of Management 30 (4), 1125–1153.

OECD, 1996. OECD Benchmark Definition of Foreign Direct Investment. OECD, Paris.

Peng, M.W., Meyer, K.E., 2011. International Business. Cengage Learning, London.

Rugman, A.M., Verbeke, A., 2004. A perspective on regional and global strategies of multinational enterprises. Journal of International Business Studies 34 (1), 3–18.

UNCTAD. World Investment Report. UNCTAD, Geneva, various years.

Foreign Language Teaching and Learning

Aleidine J Moeller and Theresa Catalano, Teaching, Learning and Teacher Education, University of Nebraska-Lincoln, Lincoln, NE, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Foreign language teaching and learning have changed from teacher-centered to learner/learning-centered environments. Relying on language theories, research findings, and experiences, educators developed teaching strategies and learning environments that engaged learners in interactive communicative language tasks. A shift in foreign language pedagogy from a specific foreign language method to the measurement of language performance/competency has resulted in a change in the role of the teacher from one of authority/expert to that of facilitator/guide and agent of change. Current developments point to public pedagogy, social media, and action research as additional ways to foster intercultural competence and language learning.

Foreign language learning and teaching refer to the teaching or learning of a nonnative language outside of the environment where it is commonly spoken. A distinction is often made between 'foreign' and 'second' language learning. A second language implies that the learner resides in an environment where the acquired language is spoken. In the area of research, the term second language acquisition (SLA) is a general term that embraces foreign language learning and investigates the human capacity to learn languages other than the first language once it has been acquired. Scholarly inquiry into the acquisition of a nonnative language includes the disciplines of psychology, linguistics, language pedagogy, education, neurobiology, sociology, and anthropology. Inquiries of learning and teaching innovations have provided new insights into successful language learning strategies and environments designed to increase language achievement and proficiency.

Definition

A language is considered foreign if it is learned largely in the classroom and is not spoken in the society where the teaching occurs. Study of another language allows the individual to communicate effectively and creatively and to participate in real-life situations through the language of the authentic culture itself. Learning another language provides access into a perspective other than one's own, increases the ability to see connections across content areas, and promotes an interdisciplinary perspective while gaining intercultural understandings. Language is the vehicle required for effective human-to-human interactions and yields a better understanding of one's own language and culture. Studying a language provides the learner with the opportunity to gain linguistic and social knowledge and to know when, how, and why to say what to whom National Standards in Foreign Language Education Project (NSFLEP) (2014).

Language scholars distinguish between the terms acquisition and learning: 'acquisition' refers to the process of learning first and second languages naturally, without formal instruction, whereas 'learning' is reserved for the formal study of second or foreign languages in classroom settings. One usually distinguishes between the relatively effortless process of SLA by children and the more formal and difficult SLA by adults. Foreign language education refers to the teaching of a modern language that is neither an official language nor the mother tongue of a significant part of the population.

Theories of Language Learning

Foreign language learning and teaching have undergone a significant paradigm shift as a result of the research and experiences that have expanded the scientific and theoretical knowledge base on how students learn and acquire a foreign language. Traditionally, learning a foreign language was thought to be a 'mimetic' activity, a process that involved students repeating or imitating new information. Grounded in behaviorist theories of learning and structural linguistics, the quality and quantity of language and feedback were regarded as the major determinants of language learning success. A popular method of teaching in the 1950s, called the audio-lingual approach (ALM), promoted an imitation and practice approach to language development. The major figure in the ALM classroom was the instructor who was cast into the role of drill sergeant, expert, and authority figure. Students were relegated to practicing and imitating patterns to a point of automatic response in the belief that the learner would then merely have to slot in lexical items appropriate to the conversational situation. It was believed that the first language interfered with the acquisition of the second language and that a transfer would take place from the first to the second language, resulting in errors. In 1959, Noam Chomsky's review (Chomsky, 1959) of B.F. Skinner's (1957) Verbal Behavior dramatically changed the way of looking at language by arguing that language was a rule-governed activity, not a set of habits. Chomsky argued that stimulus–response psychology could not adequately account for creativity involved in generating novel utterances using internalized rules. The creative aspect of language behavior implies that the human mind is involved in deep processing of meaning rather than in memorized responses to environmental stimuli. Chomsky's view of language and cognitive psychology, dubbed generative transformational grammar, regarded language acquisition as an internal thinking–learning process. Chomsky claimed that children are biologically programmed for language and have an

innate ability to discover for themselves the underlying rules of a language system. Chomsky's ideas led to the demise of structural linguistics, behaviorist psychology, and the ALM approach to language learning.

An alternative theoretical position emerged centered on the role of the linguistic environment in combination with the child's innate capacities in acquiring language. This position (interactionist) viewed language development as the result of a complex interplay between innate language capacities of the learner and the learner's environment. Unlike the innatist position (e.g., Chomsky, 1959), the interactionists claimed that language had to be modified to the ability of the learner. According to Long (1985), language input was made comprehensible by simplifying the input, by using linguistic and extralinguistic cues, and by modifying the interactional structure of the conversation. Long maintained that speakers adjust their language as they interact or negotiate meaning with others. Through negotiation of meaning, interactions are changed and redirected, leading to enhanced comprehensibility. Long proposed that learners, in order to acquire language, cannot simply listen to input, rather they must be active co-constructive participants who interact and negotiate the type of input they receive.

Each of these theories of language acquisition addresses a different aspect of a learner's ability to acquire a language. Behaviorist explanations explain systematic aspects, whereas innatist explanations explain the acquisition of complex grammar. Interactionist explanations assist in understanding how learners relate form and meaning in language, how they interact in conversation, and how they use language appropriately.

More recently, researchers have identified nine contemporary language learning theories: Universal Grammar, Autonomous Induction, Associative-Cognitive CREED, Skill Acquisition, Input Processing, Processability, Concept-Oriented Approach, Interaction Framework, and Vygotskian Sociocultural Theory (VanPatten and Williams, 2008). Some of these theories share a linguistic view of language cognition, others view it from a psychological point of view and in the case of Sociocultural Theory, a social approach is taken. The Universal Grammar (UG) and Autonomous Induction theory share the linguistic view that learners have innate knowledge of grammatical structures that is not learned through mere exposure to input. They believe that linguistic knowledge is predetermined and is independent from experience. Learning is believed to occur incidentally by deduction from innate abstract knowledge.

The psychological view of language cognition is represented by the following theories: Associative-Cognitive CREED, Skill Acquisition theory, Input Process theory, Processability theory, Concept-Oriented Approach, and the Interaction Framework. While these approaches share a psychological view of cognition, there are some distinct differences. The Associative -Cognitive CREED, Input Processing, Processability, and Concept-Oriented theories view language acquisition as implicit and language learning is presented as an incidental and a subconscious learning process. However, according to the Skill Acquisition theory there is a conscious processing in language acquisition that requires explicit instruction in order for deliberate learning to occur.

The most prevalent and most widely held theory, the Sociocultural Theory (SCT) proposed by Vygotsky, views cognition as a social faculty. According to this theory, participation in culturally organized activities is essential for learning to occur. Active engagement in social dialogue is important. Learning is regarded as intentional, goal-directed, and meaningful and is not a passive or incidental process but is always conscious and intentional. According to Ellis and Larsen-Freeman (2006) learning from exposure comes about "as part of a communicatively rich human social environment" (p. 577). This is discussed in more detail later in this article.

Emphases in Second Language Research and Teaching

Research has revealed that knowledge of language structures demonstrated on discrete-point tests does not ensure communicative ability when the measure of language knowledge is one of more spontaneous language use. Further studies have shown that there is little correlation between the rules learners are taught and their developing knowledge of the second language. Language scholars have demonstrated that certain aspects of second language learning cannot be altered through instruction, and that intermediate, nonnative-like second language competencies, known as stages of interlanguage, characterize the progression of SLA. Selinker (1974) viewed interlanguage as an intermediate system located on a continuum stretching from the native language to the target language. Corder (1978) stated that, in the interlanguage process, the learner constantly and progressively adjusts the native language system to approximate the target language system more closely (restructuring continuum). Corder noted that not all learners showed evidence of transfer from native language to target language and suggested that there was a uniformity about the way second language learners progress and that they follow approximately the same sequence of development regardless of their native language (developmental continuum). More recent studies in the area of interlanguage such as Vidaković's study of Serbian learners of English support Corder's findings that not only is a learner's interlanguage a continuing developmental process, but that it is also systematic in its development. However, new findings contain evidence that the acquisition paths of the two linguistic systems of the learner are influenced by a rich interplay of mostly universal (as opposed to language-specific) factors and show similarities unrelated to the first or second language (2010). According to this view of SLA, the controlling factor is the innate ability for learning language that all human beings possess. Pica (1983) determined that all language learners progressed through a fixed series of stages, known as developmental sequences, in learning particular linguistic subsystems, such as word order, negation, or relative clauses. In English negation, for example, when communicative samples were examined, it was revealed that both foreign language and second language learners progressed through the same four-stage sequence, defined in terms of placement of negation. Ellis (1986) reviewed several studies that involved Japanese, Spanish, German, and Norwegian children, adolescents, and adult learners. He concluded that all English-as-a-second

language learners pass through the following prescribed set of stages: (1) 'no' phrase, for example, 'No drink'; (2) negator moves inside the phrase, for example, 'I no can swim'; (3) negator is attached to modals, for example, 'I can't play this one'; and (4) auxiliary system is developed and learner acquires correct use of not and contractions, for example, 'He doesn't know anything.' This suggests that learners make particular kinds of errors at particular stages in the acquisition of a structure. Each stage marks some kind of restructuring in the mind of the learner regarding that particular structure. Structure evolves over time.

Is L2 learning possible without rules? In the absence of rules, low-level associative learning that draws on information driven processes supported by memory is possible but does not lead to knowledge of a systematic rule. Future research should investigate whether all aspects of a second language are equally learnable by implicit means or whether more complex aspects of the second language may require more conceptually driven processing in order for associations to be formed (Ellis, 2002).

Recent trends in foreign language research have increasingly focused on multilingualism and the interplay of multiple linguistic systems in the language learner. One area of multilingualism that has been much examined is cross-linguistic influence (also known as language transfer, linguistic interference, the role of the mother tongue, native language influence, and language mixing) (Odlin, 2003). Studies point to the complexity and dynamic nature of the multilingual system and have identified a number of factors involved in cross-linguistic influence in the acquisition of a foreign language, particularly of a third language. Some of these factors include (psycho) typological distance (e.g., the similarity of the languages or perceived similarity), foreign language effect (a coping strategy used as a type of 'foreign language cognitive mode'), proficiency level, and recency of use or context of the interaction. Studies also provide evidence for stronger language transfer between L2 and L3 rather than L3 and L1 (De Angelis, 2007; Wrembel, 2010). Moreover, current studies of cross-linguistic influence tend to treat each aspect of language acquisition separately (e.g., phonological transfer and transfer of literacy skills) and reveal that not each type of transfer works in exactly the same way or is influenced by the same factors.

Learner-Centered Instruction

Two communicative approaches, the input model and the input interaction model, represent two models of foreign language theory and teaching that investigate the language acquisition process from the perspective of the learner. Krashen (1982) is the principal advocate of the input model of foreign language teaching. His theory is grounded in (1) Chomsky's generative linguistics; (2) research on the effectiveness of different second/foreign teaching methods; and (3) research on affective factors (such as motivation, anxiety, and personality). Krashen posited that SLA occurs when the learner comprehends the language input in a low-anxiety, high-motivation situation, and proposed that the teacher's role is to create such a learning environment. Krashen further claimed that conscious grammar teaching/learning is effective only in a monitoring capacity to check for grammatical accuracy, not in the acquisition of the second language itself.

Because classrooms remained a major setting for language learning, the pursuit to determine those elements that enhanced classroom language achievement became particularly important. Why do two learners who seemingly have the same instructional opportunity achieve varying levels of language proficiency? Investigations focused on individual skills or abilities and environmental factors that may impact foreign language achievement and proficiency.

Individual cognitive (e.g., intelligence, aptitude, or ability) and affective (e.g., attitude and personality variables) factors were analyzed. Skehan (1986) noted a fairly strong relationship between cognitive variables such as aptitude, intelligence, and language achievement for learners in foreign language classrooms. Other factors analyzed include the age of the learner. Researchers have typically aimed at understanding how early versus late learning affects successful acquisition, and discussed this issue in terms of a critical period of acquisition in which language acquisition seemed to depend on appropriate input during this time frame (Hernandez and Ping, 2007). Although critical period effects in L2 learning are still being debated, researchers generally agree that early learning of L2 is associated with higher ultimate proficiency, and age of acquisition is reliably the strongest predictor of ultimate attainment in the language (Birdsong, 2006). Recent developments in the fields of neurolinguistics and neurobiology provide evidence that L2 grammatical processing is carried out through the same brain computational devices as those in L1. Furthermore, proficiency, age of acquisition, and amount of exposure to the L2 has been found to interact in complex ways with the different types of language performance (Perani and Abutalebi, 2005). Interestingly, not only is this true for L2 acquisition but also brain imaging research in neurobiology has revealed a general tendency that early learning (of any type) leads to dedicated neural circuitry that affects the form of cognitive and neural structures at later stages of development (Hernandez and Ping, 2007: p. 646). Moreover, studies have suggested that the attainment of broad native-likeness among late L2 learners is in fact possible (Marinova-Todd, 2003; Hernandez and Ping, 2007; Perani and Abutalebi, 2005). Future research in L2 acquisition must account not only for the typical decline in L2 attainment with age but also for the native-like achievement levels of which some late learners are capable (Birdsong, 2006: p. 37).

The predictive power of the above-mentioned traits, however, has been shown to decrease as the criteria for language proficiency became more communicative and the learning setting became more natural (versus formal and instructional). The most avid pursuit in research occurred in investigations of the role of motivation in learning language and the learner's attitude toward the target language and culture. Using Gardner and Lambert's (1972) differentiation between integrative and instrumental motivation, researchers reported no significant advantage for an integrative (intrinsic) motive and others reported an advantage when the learner was driven by instrumental (extrinsic) motives. Integrative motivation was defined as one in which the target language was being learned by an individual in order to be accepted by the native speaker community. Instrumental motivation was one

in which the language was being learned for external benefits, such as securing a better job.

Results of studies investigating environmental factors reported on the effect on achievement scores. Carroll (1975) conducted a survey of French instruction in eight countries and noted effects on achievement by gender, school type, and teacher gender, and mixed effects according to parental interest. Social factors outside the school were determined to have a significant impact on the development of language proficiency. Both cognitive and affective factors were investigated to explain the variance in foreign language achievement. Motivation, attitudes, anxiety, self-esteem, tolerance of ambiguity, risk-taking, cooperation, and competition proved to be key variables that explained individual differences in foreign language learning (Ellis, 1994). Successful language learning was determined to be largely dependent on who was learning the language, under what circumstances, and for what purposes. Foreign language acquisition was revealed to be a complex, multidimensional process influenced by both learner and environment variables. The questions generated by these theories and research studies began to focus on significant new responsibilities on the part of the teacher in the design and support of individual and personalized learning tasks.

Learning and Measurement

Language teaching has experienced numerous curricular innovations in response to the importance of providing students with opportunities to acquire and practice the foreign language in contextualized and meaningful language communicative tasks at all stages of the second or foreign language acquisition process. Communicative language teaching (CLT), the term most associated with current discussion of method, emerged as a significant approach that found universal resonance and support in theory and application in many contexts and across disciplines (linguists, methodologists, and curriculum developers). Central to the rise of CLT was the realization that linguistic competence does not on its own achieve communicative competence (Canale and Swain, 1980) and that language used in meaningful, authentic contexts is more readily acquired.

Pair work, group work, cooperative/collaborative learning settings, authentic materials, culturally integrated lesson content, and interactive tasks focused on the cognitive and affective domains were integrated into foreign language classrooms. In addition, there has been a call for the reconceptualization of theoretical underpinnings related to use of the target language for language instruction.

Past instructional policies have been dominated by monolingual instructional principles largely unsupported by empirical evidence. In today's multilingual classrooms there is a need to revisit the common assumptions that translation from L2 to L1 (or L3 to L2 for that matter) has no place in the teaching of language or literacy, that instruction should be carried out exclusively in the target language without recourse to students' L1, and that L1 and L2 should be kept rigidly separate (Cummins, 2010). In contrast to these assumptions, recent research has shed light on the fact that the L1 should be seen as a cognitive and linguistic resource that can function as a stepping stone to support more effective performance in the L2 (p. 238).

Furthermore, constructivist teaching practices, influenced by Vygotsky's emphasis on social interaction in learning and development, helped learners to internalize and reshape new information. The theoretical underpinnings of Vygotsky's (1978) view of language learning that maintained contextualized input in cooperative, meaningful interactions with others formed a basis for Sociocultural Theory (SCT), which has enhanced language acquisition and taken hold in classrooms around the globe. According to Lantolf and Pavlenko (1995), the goal of SCT is to understand how people organize and use their mind in the daily process of living. From a sociocultural stance, acquiring language amounts to more than just mastery of the linguistic properties of the L2. It involves the "dialectic interaction of two ways of creating meaning in the world" (p. 110). The interaction between an expert (teacher) and novice (learner) in a problem-solving task (scaffolding) in which the expert's role was to provide the novice with instructional support then became the model for communicative tasks in the foreign language classrooms. Based on Vygotsky's concept of a Zone of Proximal Development (the distance between the actual developmental level and the level of potential development), the expert's and teacher's role was to gain the learner's interest in the task, simplify the task, keep the learner motivated, point out important features, reduce anxiety and frustration during problem solving, and model appropriate form. In accordance with the new responsibilities, the role of the classroom teacher shifted to that of an architect, creating meaningful, interactive, and cooperative learning tasks designed to engage the learner actively in negotiating language meaning in authentic contexts that are co-constructed.

The focus on student language proficiency as measured through performance-based tasks made itself felt both in language learning research and in teaching. Questions emerged regarding how language proficiency could be enhanced and how best to measure the level of language proficiency.

As the proficiency movement has gained momentum in the US and most recently in Europe, consensus was sought about describing and measuring language abilities. The development of the Proficiency Guidelines by the American Council on the Teaching of Foreign Languages (ACTFL) defined what language users are able to do with language in speaking, listening, reading, and writing at various levels of performance. These Guidelines marked a major shift in language pedagogy from methodology to measurement and a focus on learner outcomes. In 1996, content standards were published and subsequently revised (National Standards in Foreign Language Education Project, 1996, 2006, 2014) that delineated what learners should know and be able to do with language. The ACTFL Performance Guidelines for K-12 Learners (ACTFL, 2006) described language performance within three modes of communication (interpersonal, interpretive, and presentational) to assist teachers in understanding how well students demonstrate language ability at various points along the language-learning continuum. A similar effort by the International Association for the Evaluation of Educational Achievement's Language Education Study is seeking a comparison and evaluation of the outcomes of different educational systems across Europe. The Standards Movement, seeking to promote the establishment of guidelines for the teaching of foreign languages for all learners, indicates the growing concern with learner outcomes and accountability.

In a standards-driven environment, the shift to student performance requires that teachers have a repertoire of approaches that target specific goal areas or standards.

Intercultural Competence

Increasingly, language educators contend that foreign language learning should increase students' intercultural competence (IC) which would allow them to see relationships between different cultures, mediate across these cultures, and critically analyze cultures including their own (Chapelle, 2010). Language teachers have now recognized their role in eliciting culture learning in their classrooms and ways to access that learning (Moloney and Harbon, 2010). One such way proposed by Schulz (2007) is through utilization of culture-learning portfolios. According to Schulz, the teaching of intercultural competence should include developing awareness of variables that affect communicative interactions, recognizing stereotypes and evaluating them, and developing awareness of types of causes for cultural misunderstandings between members of different cultures. The use of a culture-learning portfolio allows teachers to assess students' progress over time based on specific objectives that can be related to individual student interest. These portfolios encourage critical reflection and self-evaluation and, especially important in the area of cultural learning, the use of multiple sources of evidence (Schulz, 2007: p. 18). Despite much research into effective strategies and approaches to teaching and assessing intercultural competency in foreign language classrooms (particularly in the United States), several challenges have been put forward. One such challenge is that of sensitizing students to the value of seeing the world through the language/culture of another and creating a more affective climate for developing intercultural competency in an environment where a monolingual monocultural national linguistic identity rules at home and global English rules abroad (Fonseca-Greber, 2010: p. 117).

Foreign Languages: Future Directions

One area that remains controversial in the world of foreign and second language teaching today is the question: Is native-like attainment a necessary or desirable goal in the global world we live in today? In the field of English as a Foreign Language (EFL), the question of whether speakers should conform to native speaker norms of English in light of its increasing use in international contexts has been widely debated in recent years (Timmis, 2002). In light of this issue many scholars in the field have raised the question of why native speaker communities are most often a model for learners of English as an international language. In reaction to this, a deluge of terms have been developed (e.g., Global English, International English, International Standard English, World English, or World Englishes) some of which challenge the idea that only native speaker community varieties are valued (McArthur, 2001). Proponents of the term 'Global Englishes,' for example, promote the idea that English belongs to all who use it, however they use it (p. 4).

Another important direction in research that requires more attention is use and effect of computer technology on foreign language learning. As classroom tasks become more focused on real-world issues, texts, or events, and problem-solving-based tasks, technology introduces a new dimension to the teaching and learning process that incorporates the use of social media such as Facebook, Twitter, Skype, Voice Thread, and others. Digital media allows students to manipulate learning materials and language at their own pace and according to individual needs. Students examine reports, authentic documents, and web pages to find information that can be synthesized and discussed later and can collaborate electronically with youth from around the world. In such a learning environment the role of the teacher changes from one of authority figure or expert who delivers knowledge to one who facilitates, guides, and supports student learning. The teacher assumes greater responsibilities in designing and supporting individual and personalized learning tasks. This has tremendous implications for teacher educators and teacher trainers to act as agents of change as they foster language learning through the use of public pedagogy and critical media literacy. One of the most effective research methodologies that emerged in the last few years has been action research. Inquiring into one's own instructional practices through classroom-based investigations, teachers actively contribute to the research endeavor and change practices based on findings. Such research promises to improve teaching practices that are of interest to both researchers and teachers.

Methodologically classroom-oriented research has been largely conducted within the framework of correlational approaches, case studies, survey research, ethnographic research, experiments, and discourse analysis (Johnson, 1992). While the choice of research method is largely determined by the nature of the research question to be investigated, or by the hypothesis to be tested, thoughtful combinations of qualitative and quantitative research on foreign/second language learning conditions will provide valuable insights into language acquisition processes. Greater use of qualitative and mixed methods investigating students in their classrooms with special attention to cultural, situational, and longitudinal contexts is needed and recommended. As foreign language research draws on related disciplines (psychology, psycholinguistics, neurobiology, neurolinguistics, sociology, and linguistics) to better explain conditions that lead to greater language proficiency and differential success among foreign language learners, a deeper understanding of how languages are acquired and consequently how they should be taught will be gained. Furthermore, as learning and teaching innovations continue to be tested and researched, new insights will be gained that will influence teaching practices globally.

See also: Chomsky, Noam (1928–); Communicative Competence: Linguistic Aspects; Language Acquisition; Multilingualism; Second Language Acquisition; Teacher Education; Vygotsky's Sociocultural Theory.

Bibliography

ACTFL, 2006. ACTFL Performance Guidelines for K-12 Learners. American Council on the Teaching of Foreign Languages, Yonkers, NY.

Birdsong, D., 2006. Age and second language acquisition and processing: a selective overview. Language Learning 56, 9–49.

Canale, M., Swain, M., 1980. Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics 1, 1–47.

Carroll, J.B., 1975. The Teaching of French as a Foreign Language in Eight Countries. Wiley, New York.

Chapelle, C., 2010. If intercultural competence is the goal, what are the materials?. In: Proceedings of CERCLL Intercultural Competence Conference. CERCLL, Tucson, Arizona, pp. 27–50.

Chomsky, N., 1959. Review of verbal behavior by B.F. Skinner. Language 35, 26–58.

Corder, S.P., 1978. Language-learning language. In: Richards, J.C. (Ed.), Understanding Second and Foreign Language Learning: Issues and Approaches. Newbury House, Rowley, MA, pp. 71–93.

Cummins, J., 2010. Rethinking monolingual instructional strategies in multilingual classrooms. Canadian Journal of Applied Linguistics (CJAL) 10, 221–241.

De Angelis, G., 2007. Third or Additional Language Acquisition. Clevedon: Multilingual Matters.

Ellis, R., 1986. Understanding Second Language Acquisition. Oxford University Press, Oxford, UK.

Ellis, R., 1994. The Study of Second Language Acquisition. Oxford University Press, Oxford, UK.

Ellis, N.C., 2002. Frequency effects in language processing. Studies in Second Language Acquisition 24, 143–188.

Ellis, N.C., Larsen-Freeman, D., 2006. Language emergence; implications of applied linguistics-introduction to the special issue. Applied Linguistics 27, 558–589.

Fonseca-Greber, B., 2010. Social obstacles to intercultural competence in American language classrooms. In: Proceedings of CERCLL Intercultural Competence Conference. CERCLL, Tucson, Arizona, pp. 102–123.

Gardner, R.C., Lambert, W.E., 1972. Attitude and Motivation in Second Language Learning. Newbury House, Rowley, MA.

Hernandez, A.E., Ping, L., 2007. Age of acquisition: its neural and computational mechanisms. Psychological Bulletin 133 (4), 638–650.

Johnson, D.M., 1992. Approaches to Research in Second Language Learning. Longman, White Plains, NY.

Krashen, S., 1982. Principles and Practice in Second Language Acquisition. Pergamon, Oxford, UK.

Long, M.H., 1985. Input and second language acquisition theory. In: Gass, S., Madden, C. (Eds.), Input in Second Language Acquisition. Newbury House, Rowley, MA, pp. 377–393.

Lantolf, J., Pavlenko, A., 1995. Sociocultural theory and second language acquisition. Annual Review of Applied Linguistics 15, 108–124.

Marinova-Todd, S., 2003. Comprehensive Analysis of Ultimate Attainment in Adult Second Language Acquisition (Unpublished doctoral dissertation). Harvard University.

McArthur, T., 2001. World English and world Englishes: trends, tensions, varieties and standards. Language Teaching 34, 1–20.

Moloney, R., Harbon, L., 2010. Making intercultural language learning visible and assessable. In: Proceedings of CERCLL Intercultural Competence Conference. CERCLL, Tucson, Arizona, pp. 281–303.

National Standards in Foreign Language Education Project (NSFLEP), 2014. World Readiness Standards for Learning Languages (WRSLL). Alexandria. VA: Author Retrieved from. http://actfl/org/publications/all/world-readiness-standards-learning-languages.

Odlin, T., 2003. Cross-linguistic influence. In: Doughty, C., Long, M. (Eds.), The Handbook of Second Language Acquisition. Blackwell, Malden, MA, pp. 436–486.

Perani, D., Abutalebi, J., 2005. The neural basis of first and second language processing. Current Opinion in Neurobiology 15 (2), 202–206.

Pica, T., 1983. Adult acquisition of English as a second language under different conditions of exposure. Language Learning 33, 465–497.

Skehan, P., 1986. The role of foreign language aptitude in a model of school learning. Language Testing 3, 188–221.

Skinner, B.F., 1957. Verbal Behavior. Methuen, London.

Schulz, R.A., 2007. The challenge of assessing cultural understanding in the context of foreign language instruction. Foreign Language Annals 40 (1), 9–26.

Selinker, L., 1974. Interlanguage. In: Schumann, J.H., Stenson, N. (Eds.), New Frontiers in Second Language Learning. Newbury House, Rowley, MA.

Timmis, I., 2002. Native speaker norms and international English: a classroom view. ELT Journal 56 (3), 240–249.

VanPatten, B., Williams, J. (Eds.), 2008. Theories in Second Language Acquisition: An Introduction. Routledge Taylor & Francis Group, New York.

Vidaković, I., 2010. Second language acquisition of dynamic spatial relations. Research and Validity Group, Cambridge ESOL 42, 23–33.

Vygotsky, L.S., 1978. Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA.

Wrembel, M., 2010. L2 accented speech in L3 production. International Journal of Multilingualism 7 (1), 75–90.

Forensic Archaeology

John J Schultz, University of Central Florida, Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Forensic archaeology involves the utilization and adaptation of traditional archaeological methods for locating, documenting, and recovering human remains and associated evidence from a full range of medicolegal scene contexts. While forensic archaeology developed within the field of forensic anthropology, the application of field methods lagged behind the development of forensic anthropology laboratory methods for skeletal identification. However, proper field methods are now routinely used for various contexts involving recoveries of human remains. At the same time, forensic archaeologists routinely improve field methods through research and provide training to law enforcement personnel and various death investigators.

Forensic archaeologists routinely use archaeological field methods when recovering human remains from various medicolegal scene contexts including large mass graves and mass fatality scenes. While this discipline developed within the field of forensic anthropology, the development of laboratory methods for skeletal identification occurred before the development and integration of field methods. However, numerous forensic anthropologists realized that a more comprehensive field recovery was needed as it provided added benefits to the laboratory analysis. Forensic archaeology is now a robust discipline that utilizes and adapts traditional archaeological methods for locating, documenting, and recovering human remains and associated evidence. Furthermore, forensic archaeologists are often involved in improving this discipline through training of law enforcement personnel and research involving improved field methods.

Brief History of Forensic Archaeology

While forensic anthropology has traditionally been a lab-based forensic science, the majority of practicing forensic anthropologists now stress that archaeological methods should be incorporated when recovering human remains. More recently, forensic archaeology has been described as a key development within forensic anthropology (Dirkmaat et al., 2008). However, forensic archaeology methods did not develop at the same pace as laboratory methods for skeletal analysis. The emphasis of incorporating archaeological methods when processing scenes with skeletal remains originated, and was continued, by forensic anthropologists with field experience because they recognized the invaluable contributions that the field component could provide to the skeletal analysis. This initial momentum in applying forensic archaeology methods to process crime scenes involving human skeletal remains and decomposing bodies appears to have begun in the 1970s when a number of publications, including those authored by prominent forensic anthropologists, stressed the important contribution that archeological field methods could provide when processing these scenes (e.g., Morse et al., 1976; Bass and Birkby, 1978; Kerley, 1978; Ubelaker, 1978; Rhine, 1979). These early publications stressed that scene reconstruction involved proper field methods with proper collection and documentation of remains.

While not universally accepted by all forensic anthropologists, forensic archaeology methods were becoming more accepted in the 1980s. These publications included forensic anthropologists and archaeologists emphasizing the application of forensic archaeology field methods as a service that could, and should, be provided to law enforcement and death investigators (e.g., Snow, 1982; Berryman and Lahren, 1984; Wolf, 1986; Brooks and Brooks, 1984; Levine et al., 1984; Morse et al., 1983; Sigler-Eisenberg, 1985). At the same time, the first forensic archaeology manuals were published, which provided step-by-step approaches to proper field methods (Morse et al., 1983; Skinner and Lazenby, 1983). While there were many forensic anthropologists that advocated the use of forensic archaeology methods, Dirkmaat and Cabo (2012) assert that forensic archaeology had remained a peripheral endeavor as the majority of forensic cases received by forensic anthropologists were in a box after the remains were collected by law enforcement, which had not utilized proper archaeological methods.

During the 1990s and 2000s, a large number of important publications (e.g., Connor and Scott, 2001; Dirkmaat and Adovasio, 1997; Haglund, 2001; Haglund et al., 2001; Hoshower, 1998; Scott and Connor, 1997; Skinner et al., 2003; Stover and Ryan, 2001) including additional field manuals in the US and Great Britain (Connor, 2007; Dupras et al., 2006; Hunter et al., 1996; Killam, 1990) discussed why forensic archaeology methods should be implemented for various contexts involving human remains, including human rights investigations. At the same time, forensic anthropology textbooks and manuals started including a more detailed forensic archeology chapter (e.g., Byers, 2002; Burns, 1999). This proliferation of publications is a clear indication that archaeological methods had become accepted within the forensic community. At the same time, investigators also realized the additional benefits of incorporating forensic archaeology methods which included establishing the chain of custody at the beginning of the investigation, providing a more comprehensive trauma analysis, and the reconstruction of past events at the scene (Dirkmaat and Cabo, 2012).

Working in a Medicolegal Context

While traditional archaeology and forensic archaeology both "seek to protect the physical and spatial integrity of potential evidence and remains" (Haglund, 2001: 28), it is important to recognize the differences between the two disciplines. Most importantly, forensic archaeology differs from traditional archaeology because the discipline involves working in a medicolegal context. The forensic archaeologist must be concerned with issues such as rules of evidence, chain of custody, possible court testimony, and their work can be subjected to legal scrutiny (Haglund, 2001). The forensic archaeologist may have a number of constraints when performing recoveries such as foul weather, the media, limited time to perform an excavation for various reasons, and decomposing soft tissues (Schultz and Dupras, 2008; Dupras et al., 2011). For example, the forensic archaeologist may have to perform a recovery regardless of rainy, hot, or cold weather conditions.

Although the field techniques are generally similar between forensic archaeology and traditional archaeology, the recovery goals normally differ (Connor and Scott, 2001). Traditional archaeology is undertaken to answer specific research questions. Conversely, the goal of forensic archaeology is to document the scene and all evidence because it may be useful at a later time. In addition, since forensic archaeologists work under a medicolegal paradigm, rather than an anthropological paradigm, evidence is used to reconstruct single events, and not broad patterns of human behavior (Connor and Scott, 2001). However, the application of proper scene methods allows the forensic archaeologist to understand and interpret the prolonged history of crime scene transformation or change (Skinner et al., 2003). At the same time, it is important to stress that forensic archaeologists must normally adapt and have flexible strategies with their archaeological approach by tailoring archaeological methods for each type of scene (Hoshower, 1998).

Forensic Archaeology Methods

Forensic archaeology methods begin with the search for surface depositions and clandestine graves, and if needed, searches will incorporate multiple methods (Dupras et al., 2011; Schultz and Dupras, 2008). Investigators generally begin with nondestructive methods when searching for surface depositions and clandestine graves to preserve evidence and context of the scene. After nondestructive methods have been exhausted without the success of locating a grave, personnel can then employ a number of destructive methods. At the same time, while nondestructive methods are initially used to locate an area containing a possible grave, a destructive follow-up method is needed to confirm that a grave is actually detected. In addition, while the primary goal of a search is to locate skeletal remains on the ground surface, a clandestine grave, or associated evidence, searches are equally important to clear an area so investigations can be directed to other areas. After the scattered skeleton or grave is located, it is important for the scene to be processed and documented using proper archaeological methods.

Searching for Skeletal Remains Deposited on the Ground Surface

When searching for scattered skeletal remains on the ground surface, searches should be performed in a controlled manner rather than investigators walking over the ground surface without following a particular pattern. At the same time, proper search strategies are used to delineate site boundaries, to locate the primary depositional site, to locate skeletal remains that have been scattered, and to determine the original body position (Dupras et al., 2011; Schultz and Dupras, 2008). Searches normally begin with nondestructive methods such as pedestrian searches and the use of cadaver dogs. For example, a pedestrian search involves personnel searching specific areas while walking side-by-side in a straight line. Larger areas will be divided into more manageable smaller areas that will be searched separately. This search method provides 100% coverage of an area when performed by properly trained personnel. Evidence of a possible body deposited on the ground surface can include a number of characteristics: skeletal remains and soft tissue, clothing and personal objects, decomposition odor, brush or loose trash placed over remains for concealment, animal activity and scavenging, and materials used for wrapping a body (Dupras et al., 2011; Schultz and Dupras, 2008).

Skeletal remains that have been deposited in outdoor environments are most often disarticulated and dispersed due to a number of environmental variables that can include carnivore activity, gravity, human activity, and fluvial (water) transport (Dupras et al., 2011; Schultz and Dupras, 2008). Proper search strategies for locating scattered skeletal remains should consider the method of dispersal. For example, if skeletal elements were scattered because a body was deposited on an incline and skeletal elements traveled downhill because of gravity, investigators should search along the incline and at the bottom of the incline. It is also common for carnivores to disarticulate and scatter skeletal elements. After cleaning the area containing the primary deposit and mapping the visible elements, the forensic archaeologist can perform an inventory of the skeleton to determine the elements that are missing. It may then be possible to locate additional skeletal elements and teeth by determining the pattern of dispersal, which may indicate a direction that carnivores scattered the elements (Dupras et al., 2011; Schultz and Dupras, 2008). Investigators can then begin searching for additional bones along the paths of dispersal and expanding out from those areas.

Searching for a Clandestine Grave

Searches for clandestine graves can be more difficult to locate because a body would have been buried to conceal its location. Personnel should begin searches with nondestructive methods, such as a pedestrian search, by looking for surface features of a grave that are consistent with digging and concealing a body (Dupras et al., 2011; Schultz and Dupras, 2008). A number of common surface features associated with the burial process include excess mounds of soil, or a mounded area, resulting from displacement of soil due to the body. Digging around the grave may be evident because additional soil was needed to cover the body. Soil color changes, with an absence of

vegetation growth, can be noted due to the intermixing of soil horizons comprising different colors. There can also be an increase of vegetation growth in the area of the grave. A primary burial depression, with or without a smaller secondary depression, may occur. It may also be possible to locate a burial from evidence of animal digging and the subsequent disturbance of the remains. In addition, brush piles should also be investigated to determine if brush and tree branches were purposely used to conceal the location of a burial.

Additional nondestructive methods used for detecting a burial include cadaver dogs and geophysical methods. Geophysical surveys are commonly conducted to search for buried bodies and evidence. The majority of geophysical tools used for forensic and archaeological contexts are nondestructive methods that are used to locate and map buried features or objects, such as a grave, without disturbing the ground and scene context. When using a geophysical tool for a search, the application of the method would be used to highlight smaller areas within a much larger search area. These smaller areas would then require limited invasive testing of specific areas, rather than the entire site.

Ground-penetrating radar (GPR) has become the most common geophysical tool used for detection of clandestine graves (Dupras et al., 2011; Schultz, 2007; Schultz and Dupras, 2008) as controlled research has determined it to be the most successful tool used to delineate clandestine graves (France et al., 1997). All of the components of a GPR unit can be cart-mounted and pushed throughout a search area using a grid. The antenna is an interchangeable component, and antenna selection is generally based on soil type and size of the target. The GPR operates by emitting electromagnetic pulses into the ground from the antenna (Dupras et al., 2011; Schultz, 2007; Schultz and Dupras, 2008). As the electromagnetic wave travels through the subsurface, it will be refracted and reflected as it encounters contrasting items and surfaces in the subsurface. The returning wave is then recorded on the monitor of the unit and buried features in the subsurface are recorded as anomalies or reflections. However, these are nonspecific features noted on the imagery that are recognized as a localized area of contrasting properties in the soil. As a result, investigators can sometimes use scene context to determine the origin of a reflection, such as proximity to a tree may indicate detection of tree roots. However, the only reliable method to determine the specific item producing the feature is to perform a destructive follow-up method such as testing the area with a T-bar probe or a shovel test pit. There are drawbacks when using a geophysical tool as scene characteristics such as soil type are a consideration, and clear, flat, and open areas are more desirable for a survey. While the success of a GPR survey is somewhat dependent on scene characteristics, GPR is an optimal search tool when performing a search over a hard-packed surface, such as concrete or pavement, without destruction of the surface as reported by a number of successful case studies (Davenport, 2001; Dupras et al., 2011; Schultz, 2007). When performing this type of search, suspicious areas located under the hard surface by the GPR can be inspected with only minimal destruction.

When nondestructive methods are exhausted during a clandestine grave search without success at locating a grave, investigators may continue searching with a number of destructive methods (Dupras et al., 2011; Schultz and Dupras, 2008). Common methods include a grid search using either a T-bar probe or shovel test pits. For example, a T-bar probe is pushed into the ground to detect areas comprising less compact soil associated with digging a grave. Investigators will incorporate a T-bar probe to highlight smaller areas for follow-up testing. Once these smaller areas are located, investigators can further investigate the follow-up areas using shovel shining. This method involves scraping away the soil to discern soil changes associated with digging that could indicate a grave. Finally, when all of these search methods are exhausted, a forensic backhoe, with a flat-toothed blade, may be the last option to clear an area or locate a grave. When operated properly, the backhoe is used in the same fashion as shovel shining to discern evidence of disturbed soil that could indicate a possible grave.

Recovering and Excavating Human Remains

Since processing a scene is a destructive endeavor, the ultimate goals are to document scene context, to reconstruct prior events, or behavior, and to recover evidence properly (Dupras et al., 2011; Schultz and Dupras, 2008). To achieve these goals, the forensic archaeologist must perform a recovery or excavation that utilizes proper methods. Scene documentation first includes note taking that is continued throughout the entire recovery process. While scene photographs should be taken to document the entire process, photographs should also be taken to document the terrain of the surrounding area. In addition, all evidence and bones must be mapped to record the spatial location of this material.

When presented with a surface recovery of a skeleton in a wooded area, investigators should not begin by immediately removing bones without any documentation method. The visible bones and surrounding area should first be cleaned of leaf litter to expose the ground surface along with the skeletal elements and evidence. Prior to physically removing the bones, investigators should record the scene using notes, photographs, and a mapping method that involves recording the special location of all bones and evidence. The bones and evidence should be removed carefully to limit any postmortem damage, and then the material should be placed in labeled bags using a notation system that allows investigators to reconstruct the exact scene location for all of the evidence and bones. The forensic archaeologist would then scrape the ground surface with a trowel and then screen the removed soil to make sure all of the evidence and bones were recovered. In addition, if the skeleton was scattered, they would formulate a strategy to search for additional bones. The recovery will also consist of collecting additional scene evidence, such as entomological and botanical, that can be used to assist with time since death. The detailed recording of scene information can then be used to reconstruct the scene, which will include detailed maps.

Applications of Forensic Archaeology

Forensic archaeology methods are routinely used in various contexts when recovering human remains. While forensic archaeologists are commonly involved in the recovery of

remains, usually involving single individuals, for law enforcement agencies, medical examiners, and coroners, they are also involved in other recoveries that can pose additional challenges. A number of the noteworthy examples include the recovery of the US serviceman listed as missing in action (MIA) or prisoners of war (POW) from previous conflicts, processing mass fatality incidents (MFIs), and the excavation of mass graves involving victims of human rights abuses.

Joint POW/MIA Accounting Command

An important military organization that utilizes forensic archaeology is the Joint POW/MIA Accounting Command (JPAC). This organization is operated through the Department of Defense and is the largest forensic skeletal identification laboratory in the world (Holland et al., 2008). The headquarters and the Central Identification Laboratory (CIL) are located near Honolulu, Hawaii at the Joint Base Pearl Harbor-Hickam. The mission of JPAC is "to achieve the fullest possible accounting of the US service personnel missing from past wars and conflicts" (Holland et al., 2008: 47). In order to achieve this mission, JPAC regularly conducts missions to areas of prior conflicts throughout the world utilizing a mix of archaeological principles and crime scene investigation procedures to locate and recover the remains of missing US service personnel. Since 1976, JPAC and its predecessor organization Central Identification Laboratory Hawaii (CILHI) have performed recoveries in over 40 countries representing different types of terrain such as the jungles of South America and Southeast Asia, the Himalayan Mountains of Tibet, the Mediterranean Sea, the frozen tundra of Siberia, the beaches of the Solomon Islands, and the desserts of Iraq (Holland et al., 2008).

According to Holland et al. (2008), the identification process begins with the field recovery that includes a team of 12–14 personnel. Team members also include military personnel, a medic, a linguist, as well as specialists in explosive-ordinance disposal, photography, aircraft evidence analysis, and communications. The recovery leader of the team is normally a civilian anthropologist who is trained in human osteology, archaeological techniques, and evidence-handling techniques. The field recovery involves three stages: research and analysis, investigation, and recovery (Emanovsky and Belcher, 2012). Research and analysis include a historical overview of records to identify possible cases for field investigation. The investigation stage involves trying to correlate the exact location of a specific loss incident and traveling to the area to perform an investigation. Cases are then selected for recovery based on the probability of locating remains. Upon completion of a mission, all recovered remains of service personnel and material evidence are brought to the JPAC-CIL for analysis and identification by forensic anthropologists.

Mass Fatality Incident

In recent years, forensic archaeology methods have become standard practice when processing MFIs. Since September 2001, many jurisdictions have been preparing for a wide range of disaster scenarios (Dirkmaat, 2012). At the same time, investigators may revisit prior mass disaster scenes to implement continued search strategies for remains of victims. For example, when human remains were unexpectedly located near the World Trade Center (WTC) site in New York City in 2006, the Office of Chief Medical Examiner renewed efforts to locate additional remains of potential victims (Rainwater et al., 2008). The goals of the renewed efforts were to "identify the boundaries of the WTC debris deposits, excavate and document all material that may contain WTC victims' remains and personal effects in a continued effort to identify victims" (Rainwater et al., 2008: 562).

One of the most important forensic archaeology contributions for MFIs occurred in the field recovery protocol of airplane crashes. Mass fatality scenes had historically involved random searches over a few days, usually along paths of least resistance, to flag the largest items that would be collected after a few days without noting the provenience of the recovered items (Dirkmaat, 2012). The search for human remains would continue again to locate any remains not discovered during the initial search, while not noting provenience of the remains. According to Dirkmaat (2012), this inefficient method of processing the scene of an aircraft crash was the result of not considering the scene as a crime scene because all of the individuals died of a catastrophic event. Eventually, processing of the scene would end after a couple of weeks, because it was believed that additional human tissue could not be used to identify the individual.

It was finally recognized in the early 2000s that this inefficient processing method needed to change because DNA was now being used to regularly identify fragmented remains, including small fragments, and that criminal intent, particularly through terrorism, could result in MFIs (Dirkmaat, 2012). In the late 1990s, Dirkmaat and colleagues developed and implemented large-scale scene processing recovery protocols to be used at MFI scenes such as airplane crashes (Dirkmaat, 2012). These protocols involved straight-line searches through the crash site, where personnel would flag all significant evidence (human remains, personal effects, and identifiable plane parts). A second team would then document and map the location of the evidence prior to the collection phase. Overall, advantages of implementing these forensic archaeology protocols when processing a mass disaster scene include 100% search coverage of the ground surface, provenience and chain of custody is maintained for all recovered evidence, and the proper reconstruction of crash dynamics (Dirkmaat, 2012).

Human Rights and Mass Graves

Forensic archaeologists are important team members in the search, recovery, and identification of victims of human rights violations around the world (e.g., Latin America, Croatia, Bosnia, East Timor, Iraq, Rwanda, and the former Yugoslavia). Archaeologists have been recruited, along with numerous other forensic experts, by the United Nations and human rights organizations (e.g., Physicians for Human Rights) to excavate the individual and mass graves associated with investigations of war crimes, genocide, and political killings (Stover and Ryan, 2001). The history of properly recovering and analyzing human rights victims began when noted forensic anthropologist Clyde Snow was asked to recruit a qualified forensic team to assist with exhumations and identifications of individuals that died during the Argentinian Dirty War (Doretti and Snow,

2009). Applying techniques from forensic archaeology and anthropology, Snow's team conducted exhumations using proper field and laboratory methods. Snow retuned to Argentina many times between 1985 and 1990 to train and work with the team that was formally organized as the Argentine Forensic Anthropology Team in 1986.

The majority of methods used to excavate mass graves were developed in standard archaeological practices and applied to contemporary mass graves (Tuller, 2012). The forensic archaeologist is involved in determining the location of a grave, the delineation of the grave, the excavation of the grave, and body removal (Haglund et al., 2001). While witness testimony is the most successful method for locating graves (Haglund et al., 2001), the forensic anthropologist working in human rights investigations may also be involved in interviewing witnesses (Tuller, 2012). Once a grave is located, the archaeologists will need to confirm or deny the presence of a grave using standard archaeological methods, including geophysics (Tuller, 2012). At the same time, it may be necessary to delineate the horizontal extent and amount of overburden of large mass graves with tens or hundreds of bodies prior to determining an appropriate excavation strategy (Haglund et al., 2001).

More recently, mass grave excavations have focused on collection of evidence to assist with cause of death and an understanding of site formation processes (Skinner et al., 2003). In order to reconstruct the events at a site before, during, and after the creation of a mass grave, the excavation should be conducted to determine the order, or depositional events, that made the graves, as each depositional event should be present as distinct strata (Tuller, 2012). Furthermore, proper documentation of each article of evidence is essential during the excavation of a mass grave as this information can be important in linking evidence from multiple sites (Tuller, 2012). For example, linking evidence can be used to confirm that perpetrators were acting in an organized manner on specific orders from superiors, or acting independently.

Forensic Archaeology Training

Personnel trained in forensic archaeology methods may also be involved in training and research associated with this endeavor. Many forensic archaeologists routinely provide training lectures to death investigation personnel. They will lecture on various topics that can range from a general overview of forensic archaeology or anthropology, more specific archaeological methods such as the application of geophysical methods, nontraditional searches, and they will normally include case studies. This generally occurs when they are invited by a death investigation organization (i.e., law enforcement agency) to provide a lecture about forensic archaeology that normally is 1 or 2 h in duration. In addition, there are a variety of forensic archaeology short courses available in the United States for death investigation personnel to learn basic forensic archaeology techniques. More recently, forensic archaeology training courses are available for learning protocols for processing an MFI scene (Dirkmaat, 2012). Death investigation personnel normally attend these lectures as continuing education training required for their employment. These courses are designed to provide participants with proper search, excavation, mapping, and documentation techniques, as well as collection and proper packaging methods (Dupras et al., 2011; Schultz, 2007; Schultz and Dupras, 2008). In addition, participants are taught how to recognize, collect, and preserve different classes of associated evidence such as botanical, entomological, and soil samples.

The participants learn the various archaeological methods through a number of mock scenarios devised by the instructors. For example, one common scenario includes scattering a skeleton and associated evidence over a moderately large area. Participants would then learn the proper methods that are first needed to locate all of the skeletal elements such as performing a pedestrian line search and visually looking for the individual bones. After they locate and mark the location of all of the bones and associated evidence, they would then learn how to map the scattered skeleton and properly document and collect the bones and associated evidence. Another common mock exercise for the participants is to learn the methods for excavating a grave. The graves will commonly contain a plastic human skeleton or animal carcass such as a pig (*Sus scrofa*). Part of the exercise may first include trying to locate the grave by recognizing ground surface characteristics. Once the grave is located, participants will learn how to set up an archaeological grid that is used to maintain spacial controls before performing the actual excavation.

Forensic Archaeology Research

Research has been instrumental in improving and integrating traditional archaeology methods to forensic contexts. In particular, research focusing on various technologies used to search for graves and mapping scenes must be undertaken to understand how these technologies can be integrated as part of the multidisciplinary protocol. For example, controlled research has been instrumental in understanding how geophysical technologies such as GPR could be implemented for clandestine grave searches (France et al., 1997; Schultz, 2008; Schultz et al., 2006; Schultz and Martin, 2011, 2012). Researchers commonly devise a number of grave scenarios using pig carcasses as proxies for human bodies. They then control a number of variables such as carcass size, grave depth, soil type, length of internment, and possibly scenario type. While the research provides experience to the personnel, it also allows them to make predictions concerning grave detection during actual forensic GPR searches.

More recently, forensic archaeology research has explored the use of mapping different scenes using a differential global positioning system and analyzing the scatter with global information systems. This research involves devising a number of different surface deposition scenarios in open and obstructed environments (Walter and Schultz, 2013). The scenarios were then mapped and the degree of scatter was analyzed based on the type of scenario. Overall, recommendations were provided for integrating this mapping technique based on scene variables when recovering a scattered skeleton.

See also: Geoarchaeology; Human Rights, Anthropology of; Survey and Excavation (Field Methods) in Archaeology.

Bibliography

Bass, William, Birkby, Walt, 1978. Exhumation: the method could make the difference. FBI Law Enforcement Bulletin 47, 6–11.

Berryman, Hugh, Lahren, Craig, 1984. The advantages of simulated crime scenes in teaching forensic anthropology. Journal of Forensic Science 20, 699–700.

Brooks, Sheilagh, Brooks, Richard, 1984. Problems of burial exhumation, historical and forensic aspects. In: Rathbun, Ted, Buikstra, Jane (Eds.), Human Identification: Case Studies in Forensic Anthropology. Charles C. Thomas, Springfield, IL, pp. 64–86.

Byers, Steven, 2002. Introduction to Forensic Anthropology: A Textbook. Allyn & Bacon, Boston, MA.

Burns, Karen, 1999. Forensic Anthropology Training Manual. Pearson, Prentice Hall, Upper Saddle River, NJ.

Connor, Melissa, 2007. Forensic Methods; Excavation for the Archaeologist and Investigator. Altamira Press, New York.

Connor, Melissa, Scott, Doug, 2001. Paradigms and perpetrators. Historical Archaeology 35, 1–6.

Davenport, G.C., 2001. Remote-sensing applications in forensic investigations. Historical Archeology 35, 87–100.

Dirkmaat, Dennis, 2012. Forensic anthropology at the mass fatality incident (commercial airliner) crash scene. In: Dirkmaat, Dennis (Ed.), A Companion to Forensic Anthropology. Wiley-Blackwell, Malden, MA, pp. 136–156.

Dirkmaat, Dennis, Cabo, Luis, 2012. Forensic anthropology: embracing the new paradigm. In: Dirkmaat, Dennis (Ed.), A Companion to Forensic Anthropology. Wiley-Blackwell, Malden, MA, pp. 3–40.

Dirkmaat, Dennis, Cabo, Luis, Ousley, Stephen, Symes, Steven, 2008. New perspectives in forensic anthropology. Yearbook of Physical Anthropology 52, 33–52.

Dirkmaat, Dennis, Adovasio, James, 1997. The role of archaeology in the recovery and interpretation of human remains from an outdoor setting. In: Haglund, William, Sorg, Marcella (Eds.), Forensic Taphonomy: The Postmortem Fate of Human Remains. CRC Press, Boca Raton, FL, pp. 39–64.

Doretti, Mercedes, Snow, Clyde, 2009. Forensic anthropology and human rights: the Argentine experience. In: Steadman, Dawnie (Ed.), Hard Evidence: Case Studies in Forensic Anthropology. Pearson Education, Inc, Upper Saddle River, NJ, pp. 303–320.

Dupras, Tosha, Schultz, John, Wheeler, Sandra, Williams, Lana, 2011. Forensic Recovery of Human Remains: Archaeological Approaches, second ed. CRC Press, Boca Raton, FL.

Dupras, Tosha, Schultz, John, Wheeler, Sandra, Williams, Lana, 2006. Forensic Recovery of Human Remains: Archaeological Approaches. CRC Press, Boca Raton, FL.

Emanovsky, Paul, Belcher, William, 2012. The many hats of a recovery leader: perspectives on planning and executing worldwide forensic investigations and recoveries at the JPAC Central Identification Laboratory. In: Dirkmaat, Dennis (Ed.), A Companion to Forensic Anthropology. Wiley-Blackwell, Malden, MA, pp. 567–592.

France, Diane, Griffin, Tom, Swanburg, Jack, Lindemann, John, Davenport, G.C., Trammell, Vickey, Travis, Cecilia, Kondratieff, Boris, Nelson, Al, Castellano, Kim, Hopkins, Dick, Adair, Tom, 1997. NecroSearch revisited: further multidisciplinary approaches to the detection of clandestine graves. In: Haglund, William, Sorg, Marcella (Eds.), Forensic Taphonomy: The Postmortem Fate of Human Remains. CRC Press, Boca Raton, FL, pp. 497–509.

Haglund, William, 2001. Archaeology and forensic death investigations. Historical Archeology 35, 26–34.

Haglund, William, Connor, Melissa, Scott, Douglas, 2001. The archaeology of contemporary mass graves. Historical Archeology 35, 57–69.

Holland, Thomas, Byrd, John, Sava, Vincent, 2008. Joint POW/MIA accounting command's central identification laboratory. In: Warren, Michael, Walsh-Haney, Heather, Freas, Laurel (Eds.), The Forensic Anthropology Laboratory. CRC Press, Boca Raton, FL, pp. 47–63.

Hoshower, Lisa, 1998. Forensic archaeology and the need for flexible excavation strategies: a case study. Journal of Forensic Science 43, 53–56.

Hunter, John, Roberts, Charlotte, Martin, Anthony, 1996. Studies in Crime: An Introduction to Forensic Archaeology London. B.T. Batsford, England.

Kerley, Ellis, 1978. Recent developments in forensic anthropology. Yearbook of Physical Anthropology 21, 160–173.

Killam, Edward, 1990. The Detection of Human Remains. Charles C. Thomas, Springfield, IL.

Levine, Lowell, Campbell Jr., Homer, Rhine, J.S., 1984. Perpendicular forensic archaeology. In: Rathbun, Ted, Buikstra, Jane (Eds.), Human Identification: Case Studies in Forensic Anthropology. Charles C Thomas, Springfield, IL, pp. 87–95.

Morse, Dan, Duncan, Jack, Stoutamire, James, 1983. Handbook of Forensic Archaeology and Anthropology. Florida State University Foundation, Tallahassee, FL.

Morse, Dan, Crusoe, Donald, Smith, H.G., 1976. Forensic archaeology. Journal of Forensic Sciences 21, 323–332.

Rainwater, Christopher, Crowder, Christian, Hartnett, Kristen, Fridie, Jeannette, Figura, Benjamin, Godbold, Jennifer, Warnasch, Scott, Adams, Bradley, 2008. Forensic anthropology at the New York city office of the chief medical examiner. In: Dirkmaat, Dennis (Ed.), A Companion to Forensic Anthropology. Wiley-Blackwell, Malden, MA, pp. 549–566.

Rhine, J.S., 1979. Bones in the backyard: physical anthropology and medicolegal investigation. Pathologist 33, 300–306.

Schultz, John, 2007. Using ground penetrating radar to locate clandestine graves of homicide victims: forming forensic archaeology partnerships with law enforcement. Homicide Studies 11, 1–15.

Schultz, John, 2008. Sequential monitoring of burials containing small pig cadavers using ground penetrating radar. Journal of Forensic Sciences 53, 279–287.

Schultz, John, Martin, Michael, 2012. Monitoring controlled graves representing common burial scenarios with ground penetrating radar. Journal of Applied Geophysics 83, 74–89.

Schultz, John, Martin, Michael, 2011. Controlled GPR grave research: comparison of reflection profiles between 500 and 250 MHz antennae. Forensic Science International 209, 64–69.

Schultz, John, Dupras, Tosha, 2008. The contribution of forensic archaeology to homicide investigations. Homicide Studies 12, 399–413.

Schultz, John, Collins, Mary, Falsetti, Anthony, 2006. Sequential monitoring of burials containing large pig cadavers using ground penetrating radar. Journal of Forensic Sciences 51, 607–616.

Scott, Doug, Connor, Melissa, 1997. Context delicti: archaeological context in forensic work. In: Haglund, William, Sorg, Marcella (Eds.), Forensic Taphonomy: The Postmortem Fate of Human Remains. CRC Press, Boca Raton, FL, pp. 27–38.

Sigler-Eisenberg, Brenda, 1985. Forensic research: expanding the concept of applied archaeology. American Antiquity 50, 650–655.

Skinner, Mark, Alempijevic, Djordje, Djuric-Srejic, Marija, 2003. Guidelines for international forensic bioarchaeology monitors of mass grave exhumations. Forensic Science International 134, 81–92.

Skinner, Mark, Lazenby, Richard, 1983. Found! Human Remains: A Field Manual for the Recovery of the Recent Human Skeleton. Archaeology Press Simon Fraser University, Burnaby, BC.

Snow, Clyde, 1982. Forensic anthropology. Annual Reviews of Anthropology 11, 97–131.

Stover, Eric, Ryan, Molly, 2001. Breaking bread with the dead. Historical Archeology 35, 7–25.

Tuller, Hugh, 2012. Mass graves and human rights: latest development, methods, and lessons learned. In: Dirkmaat, Dennis (Ed.), A Companion to Forensic Anthropology. Wiley-Blackwell, Malden, MA, pp. 157–174.

Ubelaker, Douglas, 1978. Human Skeletal Remains. Taraxacum Press, Washington, DC.

Walter, Brittany, Schultz, John, 2013. Mapping simulated scenes with skeletal remains using differential GPS in open environments: an assessment of accuracy and practicality. Forensic Science International 228, e33–e46.

Wolf, David, 1986. Forensic anthropology scene investigations. In: Reichs, Kathleen (Ed.), Forensic Osteology: Advances in the Identification of Human Remains. Charles C. Thomas, Springfield, IL, pp. 3–23.

Forensic Genetic Databases: Ethical and Social Dimensions

Barbara Prainsack, King's College London, London, UK
Jay D Aronson, Carnegie Mellon University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article starts with a brief summary of the origins of forensic genetic profiling and databasing. It then provides a typology of forensic genetic databases and discusses key societal and ethical issues related to different configurations of forensic genetic databases across countries. It concludes with an overview of the most current ethical and societal concerns related to DNA databasing in the context of criminal investigation and criminal justice.

Background: The Beginning of Forensic Genetic Profiling

Forensic genetic databases are important tools for criminal investigation and criminal justice in many countries across the globe (Hindmarsh and Prainsack, 2010; Krimsky and Simoncelli, 2011). An answer to the question of when the history of forensic genetic databases began is not a straightforward one. Most official histories of forensic genetic *profiling* (not databasing as such) start with a discovery that geneticist Sir Alec Jeffreys made in his lab in Leicester, England, in 1984: After finding regions of the mammalian genome that had a particularly high degree of length variability (due to the presence of differing numbers of short, repetitive sequences of DNA), Jeffreys and his colleagues decided to find out whether similar variations could also be found in the human genome. Preliminary experiments revealed not only the existence of these polymorphisms (i.e., variations in length) in humans, but also that patterns tended to be inherited in a regular fashion. The more distant the genetic relationships among family members, the fewer similarities their DNA profiles displayed on average; the closer the genetic relationships, the more similarities in their DNA profiles. In cases of monozygotic twins and multiples (who share virtually 100% of their DNA), as well as in cases where two DNA profiles come from *the same person*, all the markers seen in the DNA profile are identical.

Jeffreys called his discovery DNA fingerprinting in order to borrow some of the legal and social credibility of its namesake 'regular' fingerprinting (dactyloscopy) (Lynch et al., 2008). This new technique, as well as refinements of the basic idea made by researchers around the world, rendered possible the comparison of people's genetic makeup to establish whether two (or more) DNA profiles came from the same person. This, in turn, enabled criminal investigators to link particular people to particular crime scenes (e.g., if, through comparison of DNA profiles, it was found that a person's DNA profile matched that of a biological trace found at a crime scene).

Despite this effort to establish DNA analysis as a reliable form of evidence, the use of DNA technologies in criminal investigations and in the courtroom was highly contested; experts and practitioners engaged in heated discussions about the scientific robustness and reliability of the technology, as well as about various issues concerning the probability that specific DNA variants or even entire profiles might be shared by two different individuals (Aronson, 2007; Lynch et al., 2008; Kaye, 2013). These controversies are now referred to as the 'DNA wars'; they were *de facto*, if not in principle, resolved in the course of the 1990s, when experts moved on from discussing whether DNA technologies were fit to be used as evidence at all to which tools and methodologies were most robust in the criminal justice context and what regulatory mechanisms needed to be put into place to ensure their validity and reliability.

How Does Forensic Genetic Profiling Work?

In contrast to fingerprints, which are recorded in the form of (annotated) visual data, information from modern polymerase chain reaction-based DNA profiling techniques is recorded as sets of discrete, numerical data (specifically, the size of a particular DNA fragment). When we look at a predefined set of specific places on the genome, we find that different people have different numbers of so-called short tandem repeats (STRs) in these places. Like the variant regions of the genome that Jeffreys focused on, STRs are repetitions of chains of nucleotides (nucleotides make up DNA). The greater the number of such places (loci) on the genome that are analyzed, the smaller the likelihood that two different people *coincidentally* have the same number of repetitions in all places. The chance that two people do have the same number of repetitions in the same places – the so-called random-match probability – increases not only when fewer loci are analyzed, but also when the number of repeats in a certain locus are relatively common in the respective population (in other words, when they are fairly typical for this population) or when individuals are biologically related. For example, full siblings share 50% of their DNA, on average. This means that on average, half of the genetic information (number of repeats) in the loci that forensic tests analyze will be the same. This phenomenon is utilized in a technique called familial searching. This method assumes profiles that match in many but not all loci are an indication that the real perpetrator could be a biological relative of the person whose profile produced the partial match (Murphy, 2010; Kaye, 2013; *see* Genetics and Forensics).

The basic underlying principle of STR-based genetic analysis is thus one of stratification – of assigning people to different

groups according to their genetic markers – and not of absolute individualization (Amorim, 2010). This is an important caveat with respect to this technology and an important difference from fingerprinting, for example, when we assume that fingerprints are indeed unique for every individual. (It should be noted that the reliability of fingerprint analysis is being challenged increasingly, partly as a result of the emergence of DNA profiling as the gold standard of good forensic science. The problem here, however, lies with methods and practices of comparison and not with the uniqueness – or lack thereof – of individual fingerprints; Cole, 2001; Dror and Mnookin, 2010; Mnookin et al., 2011.)

Despite the probabilistic nature of DNA profiles, they are often treated in criminal justice as if they could be unambiguously linked to one person. There are two main reasons for this. The first reason is that although the results of a DNA profile comparison are always probabilistic and never certain, random-match probabilities (the chance that a profile matches a randomly drawn profile) are often so high that they become forensically irrelevant. For example, in the context of the world population's currently being just under 7 billion, if we were faced with a random-match probability of a given profile of 1:10 trillion, it would be practically (albeit not statistically) impossible that the DNA profile had not come from the originator of the matching profile (provided that there were no issues with the preservation and analysis of the profile, such as contamination, etc.; see below). The second reason for DNA profiles' being treated as if they allowed for the establishment of a certain link between a person and a profile is a phenomenon that has been referred to as the Crime Scene Investigation (CSI) effect. This label captures, broadly speaking, the notion that the wide popularity of technocentric television drama – series in which the main heroes are not people, but technologies – has shaped our perception of what forensic technologies can and should do. The CSI effect has taken its name from the widely popular US series 'Crime Scene Investigation.' It does not apply only to members of the public but also to actors within the criminal justice system, such as jurors and judges (e.g., Cole and Dioso-Villa, 2007, 2009; Kim et al., 2009) and even prisoners (Machado and Prainsack, 2012).

Forensic Genetic Databases: The International Landscape

For much of the first decade after its introduction into criminal investigation in the 1980s, DNA profiling was used strictly on a case-by-case basis. This means that DNA samples were taken from crime scenes and suspects for specific investigations, but DNA profiles were typically not stored in a central database. The first central forensic genetic database was set up in England and Wales in 1994, with countries such as the Netherlands (1997), Austria (1997), and Germany (1998) following suit. Other countries in Europe (e.g., Italy, Greece, Malta, and Ireland) as well as many countries outside of Europe (e.g., many African countries) at this time early in 2013 still did not have centralized forensic DNA databases. In the United States, a system for the archiving and comparing DNA profiles was launched as a pilot program between the Federal Bureau of Investigation (FBI) and a subset of state and local laboratories in 1990 (Combined DNA Index System; CODIS), was enabled at the national level by the DNA Identification Act of 1994 and formally came into existence in October 1998 (Aronson, 2010). Unlike the databases in many other countries, the FBI is not responsible for physically uploading to or maintaining profiles on the database. Rather, it enables connectivity among a collection of databases controlled by state and local law-enforcement agencies. Decisions about what offenses are worthy of inclusion in the database are determined by state legislatures, not the U.S. federal government, and although the FBI issues requirements for the laboratories that participate in CODIS, it does not fully regulate or control them. CODIS allows investigators in one jurisdiction to look for matches in the databases in jurisdictions other than their own. The CODIS database started out as a repository of DNA profiles of convicted sex offenders; it became more expansive over time and now includes profiles of various types of convicted offenders and suspects, depending on provisions at state levels. Some states restrict inclusion criteria to serious felonies, but others allow inclusion of misdemeanors and even arrests. Today, CODIS stores profiles of more than 5 million people, that is, about 1.5% of the nation's total population. In terms of the absolute number of profiles stored, it is the largest DNA database in the world.

Within the European Union, the so-called Prüm Decision (2008) required EU member countries that had not already established centralized forensic databases to set them up by August 2011 in order to allow authorized actors in law enforcement in each country to search for DNA profile matches across DNA databases in the whole of the EU. The Prüm Decision, although supported by countries such Germany, Austria, and the Netherlands, faced considerable resistance from other countries such as the United Kingdom (closer to production date: update whether the UK has opted out of pre-Lisbon acquis); and some countries did not make the 2011 deadline due to financial or political struggles (Prainsack and Toom, 2013). Although the Prüm Decision is binding for EU member countries only, (unless they have opted out) some of the substantive provisions of the decision are mirrored in bilateral treaties signed between the United States and individual European countries (at the time of writing this article, these bilateral treaties focused mostly on the exchange of fingerprints; yet they were devised in such a way as to allow them to be extended to the exchange of DNA profiles by signing an additional protocol).

Infrastructures for the transnational exchange of forensic bioinformation (DNA and fingerprints), however, predate the Prüm regime. Interpol has been providing an infrastructure for forensic DNA data exchange internationally since 2003. Every member country of Interpol hosts a national Interpol contact point where dedicated matching software is installed on a separate computer system. No nominal data (names, addresses, etc., of existing persons) are included in the system, nor is it connected to any other forensic information network. Member countries can define which countries they do and do not want to compare data with. The rationale for excluding some countries from the range of those that a particular country compares profiles with can be underpinned by concerns regarding how that other country runs the database, or it can be underpinned by purely practical considerations

(e.g., European countries that have already ratified the Prüm regime compare their data only via the Prüm network and thus exclude each other in the list of countries with which they exchange data via the Interpol DNA Gateway). Whether individual member countries respond to queries from other countries by checking whether the profile matches any profiles in their own databases is up to them, and response rates vary considerably across countries. Although the Interpol DNA Gateway stores the DNA profiles that member countries upload to it in a central system, it was never meant to replace national DNA databases; its role is to provide a complementary structure dedicated to the exchange of profiles across borders. Although the Interpol DNA Gateway shares this latter goal with the Prüm regime, an important difference between the two – besides the international vs European remit and the degree of automatization of the comparison – is that the Prüm system is not a physical database at all; it is merely a mechanism allowing mutual access to the database of EU member countries (i.e., member countries technically search certain elements of each other's databases directly (Prainsack, 2010)).

A Typology of Forensic DNA Databases

There is no universal set of criteria according to which forensic genetic databases in different countries are organized. There are, however, certain dimensions that can help us to understand some core differences and similarities (see **Table 1**).

The first of such dimensions are the criteria for profile inclusion. Some countries (and American states) include in their forensic genetic databases only DNA profiles of convicted offenders or also of suspects of serious offenses (for the latter group, in some countries profiles are removed from the database by authorities if these suspects are not convicted later; for more details, see Hindmarsh and Prainsack, 2010). The underlying rationale for this restrictive approach toward profile inclusion is often a combination of two assumptions on the side of the authorities: The first assumption is that it is mostly a particular subset of the population that commits crimes and that including a wide range of people, or even the entire population, in the DNA database would increase the administrative burden without significantly increasing detection and conviction rates. The second assumption is that most members of the public do not like the idea of having their DNA profiles stored in any centralized database under governmental control and that, therefore, storing profiles of convicted offenders only or of suspects of serious crimes represents a good compromise. Other databases, such as the National DNA Database (NDNAD) in England and Wales, allow for the additional inclusion of profiles of people who have committed very minor offenses or who were arrested but never charged or were later acquitted. In the United States, the DNA Fingerprinting Act of 2005 allows the federal government to include the profiles of all persons arrested for federal crimes as well as all non-U.S. detainees. The last category is, of course, directly relevant to the American war on terror. Another major issue in the United States is that state and local authorities often keep a non-CODIS (offline) database with wider inclusion criteria that they use for local investigations.

Table 1 Central dimensions of a typology of forensic genetic databases

Classification criteria	*Features*
1. Profile inclusion criteria	Restrictive/expansive
2. Profile retention criteria	Restrictive/expansive
3. Sample retention	Yes/no
4. Genetic markers tested	Dependent on type of chip and software used (e.g., Next Generation Multiplex: NGM, CODIS)
5. Tendency toward expansion, function creep	Yes/no (and how exactly)
6. Responsibility for, and custody of, the database	Police/judiciary
7. Profiles come from	Labs run by the government/private labs/mixed systems

A second, related dimension is the criteria for the retention of DNA profiles. In some jurisdictions such as, again, England and Wales, once somebody's DNA profile has been included in the central forensic DNA database, it can stay there indefinitely. Such an expansive approach toward profile retention is typically underpinned by the assumption that people who start out by committing small offenses often 'mature' toward committing more serious crimes, and thus, having their DNA stored in the database can help to increase detection rates later on. Moreover, legal provisions stipulating that profiles must be removed from the database a certain number of years after inclusion – provided that the profile's originator has not been found to have committed any offenses or crimes in that period, or when the originator has reached a certain age – always imply a duty on the side of the relevant authorities to actually remove the DNA profiles. This is often logistically difficult, especially against the backdrop of scarce human and financial resources.

A result of the coexistence of restrictive and expansive profile inclusion and retention criteria is that the proportion of a country's (or state's) population that has DNA profiles stored in a database for police and forensic purposes varies widely. At the end of 2012, the NDNAD in England and Wales had stored DNA profiles from the greatest proportion of their total population – roughly 10%. (These percentages can change fast when the inclusion and retention criteria in a country change.) Also, although most countries and states include predominantly profiles of convicted offenders, others also include profiles of mere arrestees (who were not charged or not convicted) as well as profiles of other groups, such as victims and volunteers; the latter two groups are included to enable profiles from crime scenes to be excluded from the investigation. For example, if several DNA samples are taken at a crime scene, it can happen that some of the samples taken stem from people who are not suspects, i.e., the victims' family members, friends, or neighbors, etc. Profiles of these people are often collected – and in some countries, included in the database – to make sure that the police do not waste time investigating leads coming from DNA profile matches that are unlikely to have come from the suspect(s).

A third dimension in the classification of forensic genetic databases is sample retention. A DNA sample is the physical biological matter – e.g., a blood or semen stain, a trace of saliva,

etc. – that a DNA profile is derived from. DNA samples are never, as such, included in the dabatases (only profiles are), but practices vary greatly among countries regarding the storage of DNA samples. Some countries, such as Germany and at least one US state, Wisconsin, destroy samples as soon as the DNA profile has been derived from it; many others, including most US states, store the samples. Especially the retention of so-called subject samples – that is, samples that were not collected at crime scenes but that were taken by means of a buccal swabs from certain groups of people, such as convicts, suspects, arrestees and, in some countries, also volunteers and even victims – has attracted a great deal of concern. Critics of subject sample retention are worried that stored samples could be used for more wide-ranging genetic analysis; for example, the samples could be searched for markers that disclose information about genetic predispositions to diseases or personality traits (which is not possible to infer from the information contained in a traditional DNA profile stored in the DNA database). Many of those who support sample retention, in contrast, hold that this practice poses relatively few additional privacy risks and that it has considerable practical advantages, for example, for EU countries in the context of the aforementioned Prüm regime. Here, certain kinds of preliminary DNA matches need to be confirmed by an additional DNA analysis (because the Prüm regime treats certain kinds of mismatches as matches in the first instance). If the sample has not been retained, the police would need to find the originator of the DNA profile and obtain another sample, both of which may not always be feasible.

A fourth dimension in our typology pertains to the set of markers included in a forensic DNA profile. Not all countries include the same set of markers in the DNA profiles stored in their databases. In Europe, the so-called European Standard Set (ESS) of loci was endorsed by the European Council in 2001. The ESS is a minimal consensus of markers that all EU countries are encouraged to include in their DNA profiles to allow comparison across countries, regardless of the particular chip and software they use for sample analysis and profile storage. In the second half of the early 2000s, the European DNA Profiling Group, an association of forensic scientists, proposed the addition of five, new STR loci to the ESS (Gill et al., 2006) in order to enhance the discriminating power of DNA profile comparisons. The EU Council enforced this in a Council Resolution in 2009. In the United States, plans to increase the CODIS markers from 13 to 24 (Hares, 2012) were still under discussion at the time of this writing. Interpol, because of its role as facilitator of bioinformation exchange for police and forensic purposes internationally, has developed a minimal consensus set of seven loci that are included in most databases in the world (the so-called Interpol Standard Set of Loci (ISSOL)). The five new ESS markers are now being added to the existing seven ISSOL, so that Interpol effectively uses 12 loci for transnational DNA profile comparisons.

A fifth dimension pertains to whether a database is showing a tendency toward expansion. Such a tendency can manifest itself in a trend toward becoming more inclusive of the kinds of profiles that are stored and/or retained in the national DNA database, e.g., in the Netherlands (Toom, 2012) and in New York State's 2012 decision to include profiles of all convicts, irrespective of whether the crime was mass murder or a minor misdemeanor, in its state database (New York State Division of Criminal Justice Services, 2012). Expansion can also occur in the functions for which DNA samples or DNA profiles are being used. This latter phenomenon of expanding the functions and purposes of samples and profiles has been referred to as function creep (e.g., Dahl and Sætnan, 2009). In some instances, the functions and purposes of materials initially collected and analyzed in connection with a forensic genetic database become so widely extended that they are no longer limited to the original functions of the DNA database. The aforementioned method of familial searching is an example of function creep. Familial searching, also known as genetic proximity testing, can be performed if a DNA profile derived from a crime scene stain results in a partial match with a profile in the database. Police could then apprehend the person identified and investigate whether she or he has any biological relatives who could have committed the crime (Bieber et al., 2006; Greely et al., 2006; http://www.denverda.org/DNA/Familial_DNA_Database_Searches.htm). Genetic proximity testing can be seen as function creep because it adds an additional feature to the existing functions in forensic DNA databases and consequently broadens the scope of individuals who are likely to be targeted by police investigations (*see* Genetics and Forensics).

Function creep can also occur when databases become increasingly interlinked. In the European context, the aforementioned Prüm Decision is the most important example. A concern related to the large number of profiles stored in such database networks is that the chance of false-positive matches increases as well.

A sixth dimension relevant to classifying forensic genetic databases is this question: in whose jurisdiction and under whose control does a forensic DNA database exist? In some countries, the database lies within the authority of the judiciary, whereas in others it is run by police authorities.

The seventh and final dimension relates to the kinds of laboratories that carry out the genetic analyses (i.e., who derives the profiles from the DNA samples that are then uploaded to the database). In some countries, law enforcement authorities operate their own labs for this purpose. In other countries, private or university labs that are properly accredited also can provide DNA profiles for inclusion in the DNA database. In the United States, public and private labs coexist. In Great Britain, due to the recent closure of the Forensic Science Service, DNA profiling has been almost entirely privatized – in keeping with the notion of austerity promoted by that country's coalition government.

Ethical and Societal Concerns about Forensic Genetic Databasing

DNA databases are not foolproof producers of truth. Along with the fact that, as discussed above, DNA profiling per se does not produce certain but only probabilistic links between profiles, there are challenges to be met at every stage of the process, from securing a DNA trace at a crime scene to presenting DNA evidence in court.

At the stage of securing traces at a crime scene, insufficient resources and inadequate training of scene-of-crime officers

can cause problems, such as overlooking potentially usefully evidence or causing contamination. Contamination occurs when a DNA sample is accidentally mixed with somebody else's DNA – such as the scene-of-crime officer's, the perpetrator's, or a victim's – without there being a record that this has happened (sometimes crime labs analyze traces from crime scenes when it is known that the traces contain DNA from more than one subject, e.g., from the victim and the perpetrator; such mixed traces are not instances of contamination). The risk of contamination can be minimized but never eliminated entirely, and it can happen also in light of very high standards. Contamination can also take place in locations other than the crime scene, e.g., during transportation or storage: Samples could be swapped, misplaced, or deliberately tampered with.

Once a sample has arrived at an accredited crime lab, the risk for contamination should in theory be very low because all accredited labs are obliged to have mechanisms in place to avoid contamination. (It should be noted, however, that not all labs in which DNA analyses for police or forensic purposes are carried out have accreditation.) Such mechanisms include, depending on the lab, electronic tagging, videotaping, or using different facilities for the analysis of samples from victims and those from suspects. Once lab workers have derived a DNA profile, it is uploaded to the central DNA database.

We have already discussed issues involved in the calculation of random-match probabilities in the course of profile matching. A growing problem is that although forensic scientists and those running DNA databases are usually well aware of what a DNA profile match can tell us about a case and what it cannot, other actors in the criminal justice system are often less aware of such limitations. Because not only many members of the public, but also many judges and jurors, regard DNA science as infallible, numerical interpretations of DNA evidence are often taken as facts (related to this is also the aforementioned CSI effect). Another problem pertaining to using DNA evidence in court is the so-called prosecutor's fallacy in assessing the probabilities of guilt or innocence of a suspect based on the presentation of forensic evidence. For example, the expert witness in a case could state that the chance that a DNA obtained from the crime scene randomly matches somebody other than the defendant's is one in one billion. In that case, the prosecutor's fallacy would be to misinterpret this to mean that the chances that the defendant is innocent are one in one billion. This, of course, is incorrect, because even if it could be established with certainty that the DNA from the crime scene was indeed the defendant's, that would not prove that the defendant committed the crime: her or his DNA could have been at the crime scene for legitimate reasons, or it could have been planted, or the sample could have been contaminated at some stage with the defendant's DNA. Obviously, such misunderstandings could have disastrous consequences for the defendant. The term defense attorney's fallacy is used to describe the opposite scenario. In light of a random-match probability of, for example, one in one million, some defense attorneys argue that in a city of 8 million inhabitants (half of whom can be assumed to be of the same sex as the defendant), there are four other people who could be the originator of the DNA. This is false conclusion because it ignores all other circumstances of the crime (e.g., not all 4 million other people had access to the crime scene; some of these 4 million may be young children, etc.).

Overarching Societal and Ethical Issues

In sum, although there is wide agreement that DNA databases are a very helpful tool in criminal investigation and criminal justice more broadly, it is important to be aware of their limitations. Besides the limitations inherent in the (STR-based matching) technology itself, namely, that it produces only probabilistic results and that it is impossible to abolish entirely the risk for contamination, there are some overarching societal and ethical issues related to DNA databases for police and forensic purposes. An important one is the issue of the over-representation of certain population groups. On the basis of known issues concerning racial discrimination in arresting, charging, and sentencing, it is plausible to assume that this over-representation of some minorities is, at least in part, due to such discriminatory practices on the side of law enforcement and state authorities (Kaye, 2013). In the United States, for example, there are significant racial disparities at all levels of the criminal justice system, from the likelihood of being arrested and convicted of a crime, to sentencing, and to representation in the prison population (Duster, 2004). African Americans and Hispanics are the most highly at risk in these contexts (arrest rates are more than six times higher for blacks and Hispanics than for whites), suggesting that these minority populations will soon be over-represented in CODIS as well. It should also be kept in mind in this context that behaviors that are more prevalent among underprivileged groups are more prone to being criminalized (Washington, 2010). Also for this reason, forensic genetic databases are seen to exacerbate existing ethnic and socioeconomic biases, particularly in the context of familial searching. In this case, the high prevalence of minorities and the poor in the database may increase the likelihood that a poor person or a person of color who commits a crime is found through the DNA of their close relatives. In the context of the United States, Levine et al., (2008) worry that CODIS may soon become 'Jim Crow's database' (i.e., may represent an instance of *de facto* racial segregation). Some experts (including Alec Jeffreys) have argued in favor of universal DNA databases that include profiles of every citizen, or even every resident, of a country. No country has implemented this solution yet, although some have considered it (such as Portugal, for example).

A further issue is that we have no solid data concerning the effectiveness of DNA databases. DNA evidence is rarely used in isolation; in criminal investigations and in court, other kinds of evidence, such as witness testimonials and findings from other forensic analyses (finger- or footprints, toolmarks, etc.), are typically considered in conjunction with DNA evidence. Thus it practically impossible to measure the effects that DNA analysis and DNA evidence alone have had on an investigation or a trial. As a consequence, critics of (the expansion of) forensic DNA databases argue that our assumption that this is an effective tool for criminal justice is based on ideology rather than hard evidence. Critics also contend that the channeling of resources into DNA databases

contributes to the technologization of policing, where values such as human experience and judgment are increasingly replaced by machine-generated knowledge that receives excessive amounts of trust.

Another overarching issue related to DNA databases is that although legal texts and technical protocols often speak of DNA samples that are given voluntarily, the voluntary aspect of such an act is dubious in light of the consequences of refusing to volunteer a sample in most cases. Volunteers fall mostly into one of two groups. The first group comprises people who are potential suspects, for the purpose of comparing their DNA with traces from the crime scene. The second group includes those who may have left their DNA at a crime scene for legitimate reasons (friends, employees, etc.). Such individuals are often requested to volunteer DNA samples to eliminate their DNA profiles from the crime-scene stains. If individuals in the first group refuse to provide samples, they typically attract suspicion, the assumption being that if people have nothing to fear, they have nothing to hide and that, in turn, refusing to volunteer a sample is a *de facto* admission of guilt. This is a very problematic conclusion, of course, because there may be many good reasons for people wanting to protect their privacy other than the need to hide criminal deeds (much like people who draw curtains). In certain jurisdictions and contexts, people refusing to volunteer samples even render themselves official suspects. If people in the second group refuse to provide samples, they either attract suspicion as well – for example, a neighbor who was not among the suspects could suddenly become one – or it would be explained to them that providing samples is in their best interests. In both cases, although informed consent forms would normally be signed, it remains questionable how voluntary such consents can be considered to be.

With the increased ability to extract DNA from even minuscule biological samples, many law-enforcement agencies have turned to abandoned DNA samples when suspects refuse to comply with requests for samples. In this procedure, police can retrieve DNA from saliva left on cigarette butts, glasses, cans and bottles, pizza crusts, or even spit on the sidewalk. American law generally allows such uses of biological material because obtaining them requires neither access to the body nor a search warrant (because the sample is taken from a public, rather than a private, space (Winickoff, 2004; Genetics and Public Policy Center, 2008)).

Finally, with the technological tools advancing fast, the scenario of using the full sequence of a person's genome, rather than a selected set of genetic markers, for identification purposes may soon be technically and financially feasible. Despite the significant costs that changing DNA databases from STR profiles to whole-genome information would incur, it is unclear for how much longer forensic genetic databases will stick to the present format. Many critics are concerned about governmental authorities' holding whole-genome information of individuals because of the comparative wealth of information regarding disease predispositions and traits that can be obtained from a full sequence in contrast to an STR profile. It should be noted, however, that whole-genome sequences do not predict the health or other phenotypic traits of a person entirely, with few exceptions. One of the major lessons from the Human Genome Project so far has been that genetic information alone typically tells us very little. Moreover, it is possible that once people's genome sequences are routinely held in the clinical domain, or even on a person's own computer, it is possible that genomic information will be seen as less sensitive than is the case today.

See also: Biobanking: Ethical Issues; Bioethics in the Post-genomic Era; Bioethics: Genetics and Genomics; Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute; Expert Testimony; Genetic Admixture; Genetic Ancestry Testing; Genetics and Disaster Victim Identification; Genetics and Forensics; Genetics and Social Justice; Genetics and Society; Genetics and Sociology; Genetics and the Media; Genetics: Legal Aspects; Genetics: The New Genetics; Genomics, Ethical Issues in; Genotype and Phenotype; Human Genome Project: History and Assessment; Law and Society: Development of the Field; New Genetics and Race; Privacy: Theoretical and Legal Issues; Race: Genetic Aspects; Science and Law; Science and Technology Studies: Experts and Expertise; Scientific Controversies; Scientific Evidence: Legal Aspects; Truth and Credibility in Science.

Bibliography

Amorim, A., 2010. Opening the DNA black box: demythologizing forensic genetics. New Genetics and Society 31 (3), 259–270.

Aronson, J.D., 2007. Genetic Witness: Science, Law, and Controversy in the Making of DNA Profiling. Rutgers University Press, New Brunswick, NJ.

Aronson, J.D., 2010. On trial! Governing forensic DNA technologies in the United States. In: Hindmarsh, R., Prainsack, B. (Eds.), Genetic Suspects: Global Governance of Forensic DNA Profiling and Databasing. Cambridge University Press, Cambridge, UK, pp. 240–261.

Bieber, F.R., Brenner, C.H., Lazer, D., 2006. Human genetics: finding criminals through DNA of their relatives. Science 312, 1315–1316.

Cole, S.A., 2001. Suspect Identities: A History of Fingerprinting and Criminal Identification. Harvard University Press, Cambridge, MA.

Cole, S.A., Dioso-Villa, R., 2007. CSI and its effects: media, juries, and the burden of proof. New England Law Review 41, 435–470.

Cole, S.A., Dioso-Villa, R., 2009. Investigating the 'CSI effect': media and litigation crisis in criminal law. Stanford Law Review 61 (6), 1335–1373.

Dahl, J.Y., Sætnan, A.R., 2009. 'It all happened so slowly': on controlling function creep in forensic DNA databases. International Journal of Law, Crime and Justice 37 (3), 83–103.

Dror, I.E., Mnookin, J.L., 2010. The use of technology in human expert domains: challenges and risks arising from the use of automated fingerprint identification systems in forensic science. Law, Probability and Risk 9 (1), 47–67.

Duster, T., 2004. Selective arrests, an ever-expanding DNA forensic database, and the specter of an early-twenty-first-century equivalent of phrenology. In: Lazer, D. (Ed.), The Technology of Justice: DNA and the Criminal Justice System. MIT Press, Cambridge, MA, pp. 315–334.

Genetics and Public Policy Center, 2008. DNA Forensics and the Law: Issue Brief. http://www.dnapolicy.org/policy.issue.php?action=detail&issuebrief_id=42 (accessed 15.04.13.).

Gill, P., et al., 2006. The evolution of DNA databases – recommendations for new European STR loci. Forensic Science International 156 (2–3), 242–244.

Greely, H.T., et al., 2006. Family ties: the use of DNA offender databases to catch offenders' kin. The Journal of Law, Medicine & Ethics 34 (2), 248–262.

Hares, D.R., 2012. Expanding the CODIS core loci in the United States. Forensic Science International: Genetics 6 (1), e52–e54.

Hindmarsh, R., Prainsack, B. (Eds.), 2010. Genetic Suspects: Global Governance of DNA Profiling and Databasing. Cambridge University Press, Cambridge, UK.

Kaye, D., 2013. The genealogy detectives: a constitutional analysis of "familial searching". American Criminal Law Review 51 (1).

Kim, Y.S., et al., 2009. Examining the "CSI-effect" in the cases of circumstantial evidence and eyewitness testimony: multivariate path analyses. Journal of Criminal Justice 37, 452–460.

Krimsky, S., Simoncelli, T., 2011. Genetic Justice: DNA Data Banks, Criminal Investigations, and Civil Liberties. Columbia University Press, New York.

Levine, H.G., et al., 2008. Drug arrests and DNA: building Jim Crow's Database. GeneWatch 21 (3–4), 9–11.

Lynch, M., et al., 2008. Truth Machine: The Contentious History of DNA Fingerprinting. University of Chicago Press, Chicago.

Machado, H., Prainsack, B., 2012. Tracing Technologies: Prisoners' Views in the Era of CSI. Ashgate, Farnham.

Mnookin, J.L., et al., 2011. The need for a research culture in the forensic sciences. UCLA Law Review 58, 725–779.

Murphy, E., 2010. Relative doubt: familial searches of DNA databases. Michigan Law Review 109 (3), 291–348.

New York State Division of Criminal Justice Services, 2012. The NYS DNA Databank and CODIS. Available at: http://www.criminaljustice.ny.gov/forensic/dnabrochure.htm.

Prainsack, B., 2010. Key issues in DNA profiling and databasing: implications for governance. In: Hindmarsh, R., Prainsack, B. (Eds.), Genetic Suspects: Global Governance of Forensic DNA Profiling and Databasing. Cambridge University Press, Cambridge, UK, pp. 15–39.

Prainsack, B., Toom, V., 2013. Performing the union: the Prüm decision and the European Dream. Studies in the History and Philosophy of Science 44 (1), 71–79.

Toom, V., 2012. Bodies of science and law: forensic DNA profiling, biological bodies and biopower. Journal of Law and Society 39 (1), 150–166.

Washington, H., 2010. Base assumptions? Racial aspects of US DNA forensics. In: Hindmarsh, R., Prainsack, B. (Eds.), Genetic Suspects: Global Governance of Forensic DNA Profiling and Databasing. Cambridge University Press, Cambridge, UK, pp. 63–84.

Winickoff, D., 2004. The Constitutionality of Forensic DNA Databanks: 4th Amendment Issues. updated 2005. http://www.aslme.org/dna_04/reports/winickoff_update.pdf (accessed 16.01.13.).

Forensic Psychiatry

John Bradford and Graham Glancy, University of Ottawa, Ottawa, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 9, pp. 5740–5745, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

A number of definitions of forensic psychiatry are considered, looking at the development of the concept and its practice in the USA, Canada, Britain, and Europe. The relationship between criminal law, civil law, and forensic psychiatry is traced. The author also considers the legal regulation of psychiatry, the special clinical skills needed (particularly in evaluating and treating violent patients), and the likely future organization and direction of the field.

A number of definitions of forensic psychiatry must be considered. The American Academy of Psychiatry and the Law (AAPL) (1985) endorses the definition of forensic psychiatry adopted by the American Board of Forensic Psychiatry Incorporated which is:

> Forensic Psychiatry is a subspecialty of psychiatry in which scientific and clinical expertise is applied to legal issues in legal contexts embracing civil, criminal, correctional or legislative matters: forensic psychiatry should be practiced in accordance with the guidelines and ethical principles enunciated by the profession of psychiatry.

In the preface to the British textbook of forensic psychiatry, its field of operation is defined as

> the overlap, interface, and interaction of psychiatry and the law in all its aspects; criminal behaviour, civil litigation, family law, the diagnosis, care and treatment of psychiatric patients where the disorder is associated with abnormalities of behaviour, legislation and numerous other problems such as the management of violence in the study of sexual deviation.

It adds that

> forensic psychiatry continues to be concerned primarily with mentally disordered offenders, but now encompasses a wide range of offences and gives much more prominence to diagnosis, management and treatment in prison, hospital or the community using the skills of psychologists, nurses, social workers, and probation officers in addition to psychiatrists.

The Canadian approach to forensic psychiatry mirrors the British definition as is seen in the recent application for subspecialty status by forensic psychiatry under the auspices of the Canadian Academy of Psychiatry and the Law. The goal of forensic psychiatry is defined as

> promoting the mental health of Canadians involved in legal matters. We aim to provide a functional interface between psychiatry and the law and to improve Canadian society as a whole through early intervention programs, better diagnosis and treatment and ultimately a reduction in offending.

In twentieth-century Europe, forensic psychiatry has generally centered on advances in the fields of neurology, psychiatry, and sciences related to these disciplines. In general terms, although there is no single definition of forensic psychiatry in Europe, it would tend to follow the British and Canadian approaches. In summary therefore, as we move into the twenty-first century there appears to be a convergence of the American definition of forensic psychiatry with the British, Canadian, and European definitions which relate to the evaluation and treatment of the mentally abnormal offender while at the same time being involved in the psychiatric assessment (for legal purposes) of fitness to stand trial, criminal responsibility in forensic psychiatry dealing with criminal matters, and similar assessments involving the civil aspects of the law. As we approach the twenty-first century, forensic psychiatry in all parts of the world appears to be moving and establishing its roots as a medical specialty and a subspecialty of psychiatry. Even though the medical roots are emphasized, it is quite clear that the forensic psychiatrists must be familiar with basic issues in law.

The Concept

The forensic psychiatrist needs to be an expert clinician as well as being well versed in law and psychiatry as it applies in the jurisdiction where he or she lives. Law is understood to encompass criminal, civil, and mental health law. In medicine as well as the social sciences there is an increasing number of attempts to complete research studies involving large samples in order better to define the characteristics of disease and aberrant as well as normal behavior. This has led to progress in many areas including forensic psychiatry. This allows the individual to be classified according to whether they have certain clinical and other characteristics in common with a certain reference group. An estimate of the individual's personality characteristics can be measured against a reference sample. The difficulty in forensic psychiatry is that the legal system looks at unique characteristics of a given individual before the courts, rather than any theoretical concept of how he or she may fit a given reference group. This can potentially create a conflict where the scientific aspect of forensic psychiatry could be seen as more prejudicial than probative in the context of the legal system.

The development of forensic psychiatry in Britain, Europe, and the United States was based on historical events in the eighteenth and nineteenth centuries. In Britain the writings of Sir Mathew Hale and the trials of Arnold, Ferris, Hadfield, Oxford, and McNaughton were major landmarks in the recognition of forensic psychiatry. The development of the Bethlem Hospital and later Broadmoor as special facilities for the criminally insane were also fundamental to the growth and development of forensic psychiatry. This was extended to other facilities such as the special psychiatric facilities at Rampton and Parkhurst in the early part of the twentieth century.

The development of forensic psychiatry in Europe followed a similar path. Forensic psychiatry emerged from events in the eighteenth and nineteenth centuries. The French Revolution brought to light documents about prisons, workhouses, and civil hospitals being used to detain members of the public. This was directed specifically to the way and rationale for the detention of persons who had been found insane. Concerns over the method of detention led to pressure for scientists of the day to define insanity more precisely. The influence of religion declined and this allowed the scientific roots of modern clinical psychiatry to develop. Philippe Pinel had an enormous influence in his *Medico-Philosophical Treatise on Mental Alienation or Mania*, which was rapidly translated into English and German. This led to a system of classifying mental disorders that was soon adopted in Europe and England. This helped in the development of forensic psychiatry as it brought more consistency to the testimony in insanity hearings before the courts. Esquirol, Morel, Lombroso, Krafft-Ebing, and others contributed to these important advances. As in Britain, special institutions for the criminally insane were developed and further research occurred in these institutions.

The United States saw a similar development with Benjamin Rush delivering the first lecture on forensic psychiatry under the guise of medical jurisprudence. The lecture covered topics ranging from civil incompetence caused by psychiatric disorder, and financial incompetence, to mental disorder and criminal responsibility. Rush was followed by Beck who in 1823 published *Elements of Medical Jurisprudence*. He referred to British and French psychiatry as well as Benjamin Rush in his book. Isaac Ray followed Rush and Beck and published the *A Treatise on the Medical Jurisprudence of Insanity* in 1838. This was used to defend Daniel McNaughton. Ray's approach in which he carefully studied the law and the judicial opinions of the day – in particular, anything in the law that obstructed the psychiatrically ill (the insane) being able to receive treatment. He became the superintendent of the Maine Insane Asylum and later the Butler Hospital. Perhaps most importantly he was one of the original founders of the American Psychiatric Association, and was its President from 1855 to 1859. In 1869 Ray went on to write a paper *Confinement of the Insane* that has remained fundamental to forensic psychiatry throughout the twentieth century.

> In the first place, the law should put no hindrance in the way of the prompt use of those instrumentalities which are regarded as the most effectual in promoting the comfort and restoration of the patient. Secondly, it should spare all unnecessary exposure of private troubles and all unnecessary conflict with popular prejudices. Thirdly, it should protect individuals from wrongful imprisonment. It would be objection enough to any legal provision, that it failed to secure these objects, in the completest manner.
>
> (Ray, 1869).

Highly significantly, Quen writing in *The Principles and Practices of Forensic Psychiatry* (1990) reports 'These elements remain necessary, valid, and unachieved in America today.'

As outlined above forensic psychiatry was fundamental to the development of psychiatry in the eighteenth and nineteenth centuries and was adopted by the founders of British, European, and American psychiatry as part of their role to promote psychiatry as a medical specialty.

Criminal Law and Forensic Psychiatry

This involves fitness-to-stand-trial evaluations and assessments for criminal responsibility. It involves the forensic psychiatric assessment of mentally abnormal offenders. The forensic psychiatrist is, by virtue of his or her professional training, an expert in psychiatry and specifically in mentally abnormal offenders. This expertise is applied to legal matters such as fitness to stand trial and criminal responsibility. The challenge to the forensic psychiatrist is that the traditional doctor–patient relationship does not exist under these circumstances. The traditional aims of psychiatry of treating or helping the patient are also not obvious and may even be contradicted. The definition of forensic psychiatry by the well-known American forensic psychiatrist Dr Seymour Pollack (1974) addresses these contradictions. Pollack defined forensic psychiatry thus:

> forensic psychiatry is limited to the application of psychiatry to evaluations for legal purposes. Psychiatric evaluation of the patient is directed primarily to the legal issues in which he is involved, and consultation is concerned primarily with the ends of the legal system, rather than the therapeutic objectives of the medical system.

Pollack is correct in terms of evaluations for fitness to stand trial and criminal responsibility. In the former, the initial role of psychiatric evaluation is to establish the diagnosis, from the standpoint of either a medical illness or a psychiatric disorder affecting the person's mental functioning. Once this has been established the next task for the forensic psychiatrist is to assess whether the degree of disorder present affects the person's capacity to proceed to trial. The legal basis may vary slightly between jurisdictions but all are similar and would be analogous to Section 2 of the Criminal Code of Canada, which defines 'unfit to stand trial' as follows:

> unfit to stand trial means unable on account of mental disorder to conduct a defence at any stage of the proceedings before a verdict is rendered or to instruct counsel to do so, and, in particular, unable on account of mental disorder to

a. understand the nature and object of the proceedings
b. understand the possible consequences of the proceedings, or
c. communicate with counsel (*Section 2 Criminal Code of Canada 2000*).

The process is clearly conducted for a legal purpose and, arguably, not necessarily in the person's best interest, as it may mean finding them fit for trial which may in turn result in a guilty verdict and possibly lengthy incarceration in a correctional setting. Alternatively, if they were found unfit and remained unfit, while they would still be incarcerated, they would probably be detained in a hospital setting. In Canada and most other jurisdictions, a finding of unfitness means that a judicial hearing would follow as to whether psychiatric treatment would render the individual fit to stand trial. In Canada this could result in a treatment order issued by a judge, ordering psychiatric treatment (most commonly antipsychotic medication) to render the person fit to stand trial. This can be done independently of the person's capacity to consent or refuse psychiatric treatment.

An assessment of criminal responsibility involves judging whether the person's mental status at the time the crime took place rendered him or her incapable of appreciating the nature of their actions as well as knowing whether their actions were wrong. It is beyond the scope of this article to go into criminal responsibility in detail as the legal basis varies considerably in different jurisdictions. This particularly true in the United States where the criminal law is under state jurisdiction and therefore each state varies in terms of what it uses as the legal basis for a finding of not criminally responsible on the basis of mental disorder. The most commonly used standard in Western countries is the McNaughton Test or some modification of the test. In recent years a number of American states have either abolished or severely limited the application of 'insanity' (mental disorder) to reduce the *mens rea* component of criminal responsibility. In order to be criminally responsible for a crime the state must prove the accused committed the unlawful act (*actus reus*) and that they also had the intent to commit the crime (*mens rea*). Mental disorder if present at the time the crime was committed can negate *mens rea*, leading to a finding of 'not guilty by reason of insanity,' not criminally responsible on the basis of mental disorder, or guilty but mentally ill. Such a verdict usually results in confinement in a forensic psychiatric facility until the acquitee is regarded as having recovered and no longer being dangerous to the community at large. In other cases, a mental disorder may result in a diminishing of criminal responsibility, allowing the accused to be found guilty of a lesser offence carrying a less severe penalty. This is seen in situations where a homicide has been committed and the individual suffered from a mental disorder at the time of the homicide. The degree of mental disorder is regarded as not sufficient to negate *mens rea* but to reduce it. This could result in a finding of manslaughter as opposed to murder where the intent to kill was reduced although the accused was still criminally responsible for the death of the victim. The original McNaughton Test was strictly a cognitive test. In its original form it was: *To*

establish a defence on the grounds of insanity, it must be clearly proved that at the time of committing the act, the party accused was labouring under such a defect of reason, from disease of the mind, as not to know the nature and quality of the act he was doing, or if he did know it, that he did not know he was doing what was wrong.
(Quoted in Hunter and MacAlpine, 1963.)

Various modifications of the original McNaughton Test have occurred in the United States and elsewhere. The most important of these has been the adoption, in the United States, of the American Law Institute Test (ALI Test), which includes both cognitive and volitional aspects. To complicate things even further the actual legal interpretation of the meanings of the various tests of criminal responsibility is subject to case law. This means the interpretation may change, depending on a specific interpretation by a particular judge in a specific court case that is upheld by various appeal courts. Forensic psychiatrists who act as expert witnesses in cases of criminal responsibility based on mental disorder are therefore becoming involved in a very complicated process where they have to use their psychiatric skills initially to establish a psychiatric or medical diagnosis. Drawing on the observations of others as well as the medical history of the accused, they must establish, by a process of reconstruction, whether the disorder was present at the time of the offence. If it was present, they must assess its degree or severity and how that would affect the operating mind of the accused both cognitively and possibly volitionally. Then they must form an opinion with reference to the legal test for criminal responsibility in the particular jurisdiction where the trial is being held. He or she will need a working knowledge of the case law interpretations as they apply in the specific jurisdiction and in the specific case. Finally, as expert witness he or she must understand the facts of the case as they emerge at the time of the trial and be able to incorporate them into the opinion being rendered.

In addition to criminal responsibility and fitness to proceed with trial, forensic psychiatrists are also involved in the sentencing of mentally abnormal offenders. Here the role of an expert is to address the third principle of sentencing, which is rehabilitation (the first two being retribution and deterrence). In certain jurisdictions this may be the most prominent role for the forensic psychiatrist as it is used to mitigate the sentence while at the same time allowing the judge to direct treatment in a correctional facility for a psychiatric disorder. It therefore provides a link between forensic psychiatry and correctional psychiatry. This linkage varies in different jurisdictions and countries according the degree of development of forensic psychiatry and the availability of treatment resources in the correctional system. A highly contentious issue related to sentencing is the participation of forensic psychiatrists in the death penalty process in the United States, including pretrial assessment, the sentencing phase, the determination of competence to be executed, and most controversially the assessment of restoration of competence to be executed as a result of psychiatric treatment. Most psychiatrists but perhaps more forensic psychiatrists have to give expert evidence in court (Rix, 2011).

Civil Law and Forensic Psychiatry

This branch of forensic psychiatry involves psychiatric malpractice; psychiatric disability determinations; personal injury evaluations; testamentary capacity and guardianship;

and family law matters, including parental competence and child custody evaluations. Included here is the specialty of the child and adolescent forensic psychiatrist. These individuals must combine the skills of two psychiatric subspecialties, namely, child and adolescent psychiatry, and forensic psychiatry. In some university divisions of forensic psychiatry these skills are available through a family court clinic. This clinic would receive consultations from the family court system to evaluate parental competence in child abuse and neglect cases and also assist in determining the best interests of the child in divorce custody cases. Family court clinics have also traditionally provided evaluation of young offenders. The evaluation of young offenders, principally adolescents, requires the clinical skills and experience necessary for evaluating children and adolescents. The importance of these dual skills cannot be underestimated and at times has called for the recognition of a sub-subspecialty of child and adolescent forensic psychiatry in some countries.

Testamentary capacity may be a relatively small component of forensic psychiatry, but there is no doubt as to its complexity and the special skills necessary for these retrospective evaluations and being able to communicate the evaluation to the legal authorities.

The Legal Regulation of Psychiatry and the Forensic Psychiatrist

The legal regulation of psychiatry involves informed consent and competence; the right to refuse psychiatric treatment and the right to psychiatric treatment; psychiatric hospitalization and involuntary civil commitment; outpatient civil commitment; confidentiality; the duty to warn or protect against professional misconduct, including the impaired physician, competence to practice, and professional sexual misconduct. These are areas of psychiatry and the law where the average psychiatrist also needs to develop skills and competence and are, therefore, not an exclusive area of practice for the forensic psychiatrist. The forensic psychiatrist, however, contributes special skills and training in the legal system, and the skill to be a competent expert witness. This means their role is to assist the non-forensic psychiatrist by consultation and advice and, in the more complex cases, forensic psychiatric expertise may be desirable.

Special Clinical Skills of the Forensic Psychiatrist

A wide range of special clinical skills is unique to the forensic psychiatrist. Special skills in evaluating and treating violent psychiatric patients are a fundamental requirement for all forensic psychiatrists. This is also expanded to include the assessment of the danger posed by an offender and risk assessment, as well as the risk management of violent mentally abnormal offenders. Associated with these clinical skills are skills in assessing antisocial personality disorder as well as impulse control disorders. Because of the association between organic brain dysfunction and impulsivity and violence, forensic psychiatrists need to develop skills in assessing the neuropsychiatric aspects of psychiatric conditions that can result in violence or can affect competence. They also need to be able to evaluate and treat the most severe and violent major mental disorders. Forensic psychiatrists must develop special skills in evaluating malingering and the use of hypnosis. The forensic psychiatric aspects of certain psychiatric disorders (e.g., multiple personality disorder) are also within the skill set of the average forensic psychiatrist. The assessment and treatment of sexual deviation is also a required skill. In addition assessing capacity to consent becomes important (Zimond, 2013) although all psychiatrists are expected to do this.

Correctional Psychiatry

The strict definition of correctional psychiatry is 'psychiatry practiced within a correctional system.' Broader definitions would include any agency of social control that is involved in the treatment and rehabilitation of adult or juvenile offenders. These populations have significant psychiatric morbidity and clearly would benefit from psychiatric intervention and this had led to a growth of correctional psychiatry in the last few decades. There is a close relationship between correctional psychiatrists and forensic psychiatrists with correctional psychiatric programs often being outreach programs of university departments of forensic psychiatry (as, for example, in Canada).

Organization and Future of Forensic Psychiatry

This is a subject of debate in North America at the beginning of the twenty-first century. There is a consensus that the future growth and development of forensic psychiatry must lie within the field of medicine as a subspecialty of psychiatry. This is already well established in general terms in Canada and Europe. If we accept this consensus, some comment as to the ideal forensic program is warranted. This author strongly endorses the subspecialty status of forensic psychiatry as a critical component of its future development. This implies that the ideal forensic program would be an academic division within a university department of psychiatry. In turn, this would mean that education, clinical training, and research would be fundamental components of the program. General psychiatric residents should be required to do a rotation as part of their residency requirement. As the largest interface between medicine and the law is between psychiatry and the law, it should be part of the basic training of all psychiatric residents. It is almost impossible for the average psychiatrist to go through his career without some interaction with law and psychiatry tribunals, the criminal justice system, or the civil legal system.

An ideal program would provide fellowship training in forensic psychiatry in an accredited program for a minimum of one year. The forensic clinical program should have a medium or maximum inpatient assessment unit to provide pretrial evaluations of fitness and criminal responsibility. In addition, inpatient and outpatient facilities are required for the treatment and rehabilitation of individuals who have been found unfit to stand trial or not criminally responsible on the basis of mental disorder. The length of stay should allow for both acute and chronic care so a cautious approach to treatment and

rehabilitation would prevail. There also need to be outreach programs to correctional facilities and an active participation in pretrial diversion programs. In an ideal situation, specialized clinics for the assessment and treatment of sexual deviation, and aggression and impulse control disorders should also be present. A family court clinic with child and adolescent forensic psychiatrists on the staff should also be included. A multidisciplinary team approach to assessment and treatment is essential. Research into all aspects of forensic psychiatry is critical for its future growth and the development of its profile as a medical subspecialty.

See also: Civil Law; Criminal Liability and Criminal Defenses.

Bibliography

American Academy of Psychiatry and the Law (AAPL), 1985. Definition of Forensic Psychiatry.

Barras, V., Bernheim, J., 1990. The history of law and psychiatry in Europe. In: Blueglass, R., Bowden, P. (Eds.), Principles and Practice of Forensic Psychiatry. Churchill Livingstone, London, pp. 103–109.

Blueglass, R., Bowden, P. (Eds.), 1990. Principles and Practice of Forensic Psychiatry. Churchill Livingstone, London.

Bradford, J.M.W., Glancy, G., 2000. Commentary on 'Ensuring that forensic psychiatry thrives as a medical specialty in the 21st century.'. The Journal of the American Academy of Psychiatry and the Law 28, 20–22.

Brindle, N., Branton, T., Stanfield, A., Zigmond, A., 2013. A clinician's brief guide to Mental Capacity Act. RCPsych Press, London.

Forshaw, D., Rollin, H., 1990. The history of forensic psychiatry in England. In: Blueglas, R., Bowden, P. (Eds.), Principles and Practice of Forensic Psychiatry. Churchill Livingstone, London, pp. 61–101.

Glancy, G., Bradford, J., Gagne, P., 2000. Psychiatry and the law in Canada: A proposal for subspecialization: Part I. Bulletin of the Canadian Psychiatric Association, 20–26.

Hunter, R., MacAlpine, I., 1963. Three Hundred Years of Psychiatry 1535–1860. A History Presented in Selected English Texts. Oxford University Press, London.

Pollack, S., 1974. Psychiatry in Criminal Law. University of Southern California, Los Angeles.

Quen, J.M., 1990. The history of law and psychiatry in America. In: Blueglass, R., Bowden, P. (Eds.), Principles and Practice of Forensic Psychiatry. Churchill Livingstone, London, pp. 111–116.

Ray, I., 1869. Confinement of the insane. American Law Review 3, 193–217.

Rix, K., 2011. Expert Psychiatric Evidence. RCPsych Press, London.

Rosner, R. (Ed.), 1994. Principles and Practice of Forensic Psychiatry. Chapman and Hall, New York.

Watt, D., Fuerst, M., 1998. The 1998 Annotated Tremeear's Criminal Code. Carswell, Toronto, ON.

Weinstock, R., Leong, D.B., Silva, A.J., 1994. Defining forensic psychiatry: Roles and responsibilities. In: Rosner, R. (Ed.), Principles and Practice of Forensic Psychiatry. Chapman and Hall, New York.

Forensic Psychology

Dap Louw, University of the Orange Free State, Bloemfontein, South Africa

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 9, pp. 5745–5750, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Forensic psychology is the branch of psychology concerned with the production and application of psychological knowledge and principles within the legal process. Although forensic psychology is a relatively young discipline, psychology and law have been intertwined since the dawn of recorded history. Today there is hardly an area of the law where psychological expert testimony is not being implemented and there can be no doubt that forensic psychology has made significant contributions to the execution of justice. Consequently, the forensic psychologist often serves as an expert witness in court. The activities of forensic psychologists include the assessment of competency to stand trial and criminal responsibility, determining the validity of defenses (such as intoxication, amnesia, dissociative identity disorder, and post-traumatic stress disorder), predicting violent behavior, assessing eyewitness testimony and personal injury, providing testimony on appropriate sentencing, profiling, and contributing to child custody evaluations. The most important forensic assessment procedures are the forensic interview and psychological testing, while forensic hypnosis is also sometimes used. Because reliability is questioned, the polygraph is seldom employed. The status of forensic psychology can be increased by quality research, presenting objective and critical evidence, and realizing that there are areas in which the contribution of forensic psychologists is questionable.

Forensic psychology is the branch of psychology concerned with the production and application of psychological knowledge and principles within the legal process. (The term 'forensic,' meaning 'legal' or 'judicial,' is derived from the Latin word *forensis*, which literally means 'relating to the forum or market.' In ancient Rome, a forum was an open square, which served as a market, as well as a place where legal matters were disposed of.)

Historical Background

Since both psychology and law are concerned with human behavior, they have been intertwined since the dawn of recorded history (Gudjonsson and Haward, 1998). Indications of interest shown in the law by philosophers, from whose ranks psychology would later develop, are found in the earliest Asian, Egyptian, Greek, and Latin writings. Later Greek philosophers continued to express themselves on practical questions of law and topics that today constitute the subject matter of psychology. For example, Plato (ca. 427–347 BC) in his *Republic* considered the question whether mentally disordered people were responsible for their behavior (Allan, 1995). However, it took more than 2000 years for psychological knowledge to mature to a level where it was officially allowed in the courtroom for the first time. This happened in 1896 when Albert Von Schrenk-Notzing, student of Wilhelm Wundt who founded the first psychology laboratory in Germany in 1879, testified in a murder trial concerning the effect of pretrial publicity on memory and suggestibility (Blau, 1998). Since then, forensic psychology has made great strides. As we enter the twenty-first century, there is hardly an area of the law where psychological expert testimony is not implemented. Ewing (1985) believes that the modern justice system could not function without significant participation by behavioral scientists such as psychologists.

Roles of the Forensic Psychologist

As both psychology and law essentially cover all aspects of human behavior, it is understandable that the forensic psychologist serves as an expert on a variety of matters including those considered in the following sections.

Competency to Stand Trial

Also known as fitness to plead, competency to stand trial refers to the defendant's mental capacity to make a proper defense, i.e., to understand the legal proceedings and the charges against him or her.

Contrary to popular opinion, the mere existence of a mental disorder *per se* does not render an individual incompetent to stand trial. Many mental disorders are of such a nature and intensity that the defendant's competency is not significantly influenced. The crucial question is whether the mental disorder affects the defendant's current functional abilities within the context of relevant legal criteria for determining competency (Gudjonsson and Haward, 1998).

As laws and 'legal cultures' differ from country to country, it is understandable that statistics concerning competence to stand trial will vary considerably. For example, proportionally speaking, about 100 times more defendants in the USA are found incompetent than in England and Wales. Even within one country, the percentage of defendants found to be incompetent often varies between 1% and 80% across jurisdictions.

It does seem that individuals adjudicated incompetent to stand trial often have certain characteristics in common, such as

a history of mental health treatment (especially for schizophrenia and depression), lower intellectual functioning, and a present charge of violent crime.

The evaluation of competency is usually very reliable and the agreement rate between psychologists often exceeds 80%. Next to the traditional assessment techniques, particularly the clinical interview, several competency assessment instruments are also used.

Criminal Responsibility

Criminal responsibility (liability/accountability) should not be confused with competency to stand trial. Competency to stand trial is specifically concerned with the defendant's mental capacity during the legal proceedings. Criminal responsibility, also known as the insanity defense, refers to the time of the alleged offense and more specifically asks the question whether the defendant had the mental capacity to distinguish between right and wrong at the time of the alleged offense and could control his or her conduct accordingly. Should the defendant be found not criminally responsible, he or she rarely goes free. The person is usually referred to a mental health facility where he or she will remain as long as the commitment criteria are met.

The insanity defense is one of the most controversial issues in criminal law. One reason for this are the many misconceptions surrounding this plea. For example, although the public believes (probably as a result of the publicity given to such cases) that the insanity plea is very often entered, such cases are rare: often fewer than 1 in every 1,000 criminal cases. The success rate of the defense varies from country to country and from jurisdiction to jurisdiction, with 25–50% as a broad average.

Most defendants whose insanity pleas are successful are diagnosed with major psychotic disorders and have extensive mental health histories, often accompanied by prior civil commitments or findings of incompetence (Golding et al., 1999).

Insanity evaluations are usually regarded as more complex than competency evaluations. It is, after all, exceedingly difficult to determine what another person was thinking, feeling, and doing at the time of an offense that usually happened weeks, months, or even years earlier (Bartol and Bartol, 1994). However, it seems that specialized training and experience can lead to an agreement rate between mental health professionals of more than 80%. Several measuring instruments have been developed to assist mental health professionals in their assessment of the insanity plea.

Other Defenses

Next to competency to stand trial and criminal responsibility, the expertise of psychologists is employed in several other defenses in both criminal and civil cases.

a. *Intoxication* by alcohol or other drugs, because it is self-induced and temporary, is seldom given complete exculpatory effect (Melton et al., 1997). The major exception is when the substance abuse has been prolonged to a point where it has produced a serious mental disorder. However, in most countries voluntary intoxication can be used as a diminished capacity defense claim. In such cases mental health professionals have to produce evidence that the intoxication did indeed affect the defendant's mental capacity significantly.
b. *Amnesia* (the total or partial loss of memory) at the time of the offense is one of the most common defenses heard in the criminal court, especially concerning violent crimes (*see* Amnesia: General). It is also one of the most difficult defenses to assess. It is not only difficult to determine the weight that should be given to the claimed amnesia, but also to distinguish between genuine and simulated amnesia. An even more complex task is to determine whether the defendant was really amnesic when committing the crime, or whether he or she only developed the amnesia afterwards. Although psychologists frequently give expert testimony on amnesia, their ability to assist the court with scientific evidence is questioned. Courts are therefore often skeptical about accepting their testimony as a viable defense element.
c. *Dissociative identity disorder*, formerly called *multiple personality disorder*, is primarily characterized by the presence of two or more distinct personalities that in turn take control of the person's behavior (American Psychiatric Association, 1994). However, the disorder is highly controversial and sometimes called the 'UFO of mental health.' Many mental health professionals believe that the disorder does not exist and amounts to no more than role playing. Although screening instruments and structural interviews have been developed and hypnosis is used to diagnose dissociative identity disorder, experts admit that it is very difficult to distinguish between real cases and malingering (Alloy et al., 1999). It is therefore understandable that legal professionals tend to be very cautious about making use of this defense. However, in rare cases in several countries, dissociative identity disorder has been accepted as a defense, ranging from an insanity plea to extenuating circumstances.
d. *Post-traumatic stress disorder* (PTSD) is characterized by intensely distressing symptoms (e.g., irritability and sleeplessness) caused by exposure to a traumatic event such as rape, death, combat, accident, and even harassment and difficult work conditions (*see* Post-Traumatic Stress Disorder). PTSD can therefore be a factor in both criminal and civil cases. However, establishing the existence of PTSD poses several difficulties for the forensic psychologist. For example, the validity of the diagnoses is often questioned. Although psychological measuring instruments and techniques to assess PTSD are available, they are usually based on self-report. The result is that malingering cannot be ruled out easily.

Prediction of Violent Behavior

As one of the primary duties of the court is to protect society, assessing the offender's potential for future violence often forms an integral part of the judicial process. This is especially true if the offender is charged with a violent crime. Determining the propensity for future violence is therefore an important part of the forensic psychologist's functions.

However, the role of the mental health professional in violence prediction is controversial. The main reason for this is that past research showed that mental health professionals have a bad track record when it comes to predicting the probability of a person committing future violence, and even more so when it comes to predicting the type and severity of such potential violence. As such, professionals can pay a heavy social, professional, and legal price for failing to predict violent behavior, they tend to 'play it safe' and therefore often overpredict. The result is that research has shown that more than 50% of persons predicted likely to commit future violence, do not do so.

Ongoing research in this field is fortunately beginning to bear fruit. For example, several variables have been identified which can be regarded as risk factors for violence, e.g., a history of violence, substance abuse, symptoms of a psychosis or antisocial personality disorder (psychopathy), weapon availability, early onset of violent behavior, and absence of social support. Measuring instruments to assist psychologists have also been developed.

Eyewitness Testimony

'It is … inevitable that what a witness says in court will be a subjective and incomplete version of 'the truth' interspersed with small portions of fiction.' Although most psychologists will agree with this statement by Ainsworth (1998, p. 3), the testimony of an eyewitness still remains a very important part of evidence presented in both criminal and civil cases. However, as a result of a rapidly growing data bank in this widely researched area, especially in the 1980s and 1990s, the courts are currently examining some of their assumptions about the reliability and thus the value of eyewitness testimony. It is therefore understandable that courts increasingly rely on psychologists to provide expert testimony in this regard.

The following are some of the more important research findings that have been used in court by psychologists: (a) eyewitnesses tend to be less accurate in stressful situations; (b) people experience difficulty making cross-racial identifications; (c) eyewitnesses focus on weapons rather than faces; (d) memory decay starts immediately after a perception; (e) memory gaps are often and easily replaced by unconscious fabrication or by external suggestions; (f) there is not necessarily a correlation between the degree of certainty evinced by the eyewitness and accuracy (Melton et al., 1997); and (g) although younger children are often more suggestible and remember less detail than older children and adults, they do not necessarily have poorer memories (Saywitz and Goodman, 1996).

Personal Injury

The world is progressively becoming a litigious society in which monetary compensation is sought for almost any maloccurrence (Tsushima and Anderson, 1996). As many such claims involve 'psychological pain and suffering,' psychologists are increasingly becoming involved in so-called mental injury cases. Although determining the presence and cause of mental injury constitutes the bulk of psychologists' work in this field, they are also involved in injury claims of a more physical nature, e.g., to assess brain damage or determine the psychological effect of losing a limb.

To assess whether the plaintiff does indeed suffer from a mental injury is not always easy. First, it has to be proved that there is a causal relationship between the specific incident and the plaintiffs symptoms, i.e., that the symptoms did not already exist before the incident or were not caused by other factors after the accident. Second, as possible financial remuneration is involved, plaintiffs often tend to exaggerate or even fake their symptoms. No foolproof method to determine such relationships and distinctions is available. However, by using the most objective, valid, and reliable tests appropriate to the complaints, conducting thorough interviews, and carefully recording case notes and test reports, psychologists can make a valuable contribution towards clarifying the extent to which psychological factors may be involved in personal injury cases (Butcher and Miller, 1999).

Sentencing

If a defendant is found guilty of the charge against him or her, the court has to impose an appropriate punishment. In most countries the aim of such punishment consists of a combination or all of the following: deterrence, retribution, prevention, and rehabilitation. Since these goals have a strong psychological element and psychologists are supposed to be well equipped to provide the court with a psychological profile of the defendant, they often testify in this regard. For example, the psychologist can provide the court with information concerning the defendant's potential for future violence, the sincerity of the repentance shown, the possibility of rehabilitation, a variety of mitigating factors, and the effect of different sentences on the defendant as a unique individual.

Although there seems to be a high correlation between psychologists' recommendations and the final sentencing decision of the courts, it remains an open question as to what extent the courts do indeed take the testimony of the psychologist into account or merely use it to 'scientifically' justify decisions they have already made (Melton et al., 1997).

Profiling

Criminal profiling refers to the process in which the nature of a crime is used to make inferences about the personality and other characteristics of the likely offender.

The sensational and dramatic elements of profiling portrayed in various well-known movies, television series, and books (usually those depicting a serial killer) seem to be in sharp contrast with reality. Profiling is not widely accepted in the psychological and legal community, and some courts have even ruled profiling testimony inadmissible. There are two main reasons for this (Gudjonsson and Haward, 1998). First, a criminal profile only gives a broad indication of the type of person who may have committed the crime. It does not indicate a specific individual who happens to fit the profile. The profiler is therefore unable to say whether it is more probable than not that a specific offender did, in fact, commit the crime. Second, there is no scientific evidence to support the reliability and validity of criminal profiling in solving crimes. Indeed, it

seems that when profiling does assist the police in solving a case or in opening up new lines of inquiry, it is the exception rather than the rule.

On the other hand, it cannot be denied that criminal profiling has proven helpful in some, albeit exceptional, cases. However, much more research needs to be done before criminal profiling will earn its place as a valuable forensic tool.

A variation of profiling that has evoked some interest in the news media is psychological autopsy, which comprises the compilation of a psychological profile of especially well-known deceased individuals. It is also used in suicide cases, for example, to determine whether the deceased could indeed have committed suicide. However, as in the case of criminal profiling, its reliability and validity are also questioned.

Child Custody

With the worldwide increase in divorce rates and the resulting child custody battles between divorcing parents, mental health professionals have come to play an increasingly important role in such disputes.

Two principles play a major role in courts' decisions in child custody cases: the tender-years principle and the best-interest-of-the-child principle. The tender-years principle reflects the traditional belief that the mother is inherently better equipped to care for children of a 'tender age' (which particularly refers to children around or under the age of seven). However, several factors have led to the questioning and even abolishment of this principle in some courts. For example, research has shown that fathers and mothers are equally capable of fulfilling the psychological and other needs of children in their 'tender years.' The principle is also seen as a violation of equal gender rights. Most courts base their decisions in child custody cases on the best-interest-of-the-child principle, which involves criteria such as (a) the love, affection, and other emotional ties between parent and child, as well as between the child and his or her siblings or any other significant person; (b) the ability of the parent to provide for the child's psychological, educational, cultural, and environmental development; (c) the ability of the parent to provide the basic physical needs of the child, e.g., food, clothing, and housing; (d) the mental and physical health and moral fitness of the parent; and (e) the child's preference as to his or her custodian. However, this principle has also come under fire from critics who, for example, claim that many of the criteria are vague, difficult to assess, and open to subjective interpretation.

Assessment in child custody disputes usually includes interviewing the parents and children, as well as any other third-party source that may provide valuable information, observation of the interaction between especially parents and child, and the use of traditional psychological and specialized tests. However, the interview is usually recommended as the best method. The reasons are the following. The observation technique is regarded as artificial and thus unreliable (the individuals know that they are being observed). The traditional psychological tests are often irrelevant (e.g., there is no correlation between good parenthood and IQ). The specialized tests are also not very reliable and valid. It is generally accepted that a psychologist, acting according to professional guidelines, can make a valuable and often decisive contribution in child custody cases.

Forensic Assessment

Forensic assessment refers to the procedures, methods, and techniques that psychologists use to gather information in order to assist the court in coming to a final decision. The nature of the assessment will depend on the theoretical orientation, training, experience, and preference of the psychologist, as well as the nature of the specific case. Usually a combination of procedures, methods, and techniques are used.

The forensic interview can be considered the central element of the assessment process. The interview can be structured (the questions being determined and systematized in advance), unstructured (the interview being allowed to develop spontaneously), or a combination of the two. The forensic interview is much more than just an 'ordinary' interview. It is a scientific method for gathering information and requires specialized training and experience. As mentioned, it is usually necessary to interview not only the specific individual, but also family members, friends, and other third-party sources. Such collateral information may be helpful not only in acquiring a broader picture, but also in determining the reliability of the information provided by the different sources.

In addition to this basic assessment method, several others are employed, and examples are given below.

Psychological Testing

Virtually all forensic psychologists incorporate psychological tests in their assessment armament. Psychological test results can be powerfully persuasive instruments because of society's long-standing high regard and respect for these measuring instruments (Tsushima and Anderson, 1996). Many courts also seem to expect that such tests will be used.

However, it has also become clear that psychological tests have not always conformed to the high qualities attributed to them. Not only are the reliability and validity of many of the tests used by psychologists generally disputed, but also they have not been developed specifically for forensic issues. The result is that their usefulness in forensic settings is questioned. Attempts to develop specialized forensic assessment instruments have also not yet been very successful, mainly owing to psychometric flaws. Another problem especially encountered in underdeveloped and developing countries is that tests are used on groups for which they have not been developed. An example is the worldwide use of various American tests developed for the US population on groups for which they were not standardized. Cultural differences cast doubt on the reliability of such results.

On the other hand, can there be little doubt that, in the hands of a well-trained and experienced examiner who is well aware of the pitfalls, psychological tests can be very valuable tools and can provide valuable information.

Forensic Hypnosis

There are few topics in psychology more shrouded in mystery and misconception than hypnosis. Mainly responsible for this is the sensational representation of this technique by the movie industry, news media, and stage hypnotists. Even in the

scientific world controversy surrounds the hypnosis phenomenon, with some regarding it as a valuable tool in many areas, while others do not believe in the existence of such a state, and mainly contribute it to suggestion and role-playing (*see* Hypnosis, Psychiatry of).

The primary use of forensic hypnosis centers on enhancing the memory of a witness, victim, or even defendant, especially with a view to helping him or her remember forgotten or vague details of a crime. Most research shows that, compared with people who have not been hypnotized, people under hypnosis tend to recall more detail about an event. They also usually feel confident that their recollections are correct. However, their stories contain a great deal of inaccurate information, even more so than in the case of nonhypnotized subjects. It is generally accepted that forensic hypnosis is seldom more efficient in improving memory than any other standard psychoforensic procedure. Courts therefore hesitate to accept testimony acquired through hypnosis. In some countries such testimony is even inadmissible. Others do allow it, but usually only if certain safeguards and guidelines are applied. It is therefore important than forensic psychologists meet these preconditions if they wish to employ this controversial forensic tool.

Polygraph

The polygraph, also called 'lie detector,' is an apparatus that measures physiological variables such as heart rate, blood pressure, and respiration rate. Polygraphers believe that when people are asked threatening questions to which they respond by lying, physiological changes will occur which will show up on the polygraph record. This makes the polygraph a potentially valuable forensic tool.

Unfortunately, the validity of the polygraph leaves much to be desired. Polygraph results seem to be wrong in at least 30% of all cases. This concerns innocent people who are wrongly accused, as well as guilty persons who plead innocence. For this reason, most courts do not admit evidence based on polygraph results, or admit it only if it adheres to certain stipulations (e.g., that all parties agree to the test before it is administered).

Ironically, it seems that one of the polygraph's biggest strengths lies in people's belief that it is reliable. It seems that about 30% of suspects confess after they have failed a polygraph test (Memon et al., 1998). However, it does not necessarily mean that these suspects had been lying. It could also be that innocent suspects tend to confess because, having failed the test, they see little possibility of convincing the court of their innocence thereafter. By confessing they hope to get a lighter sentence.

The Future

Only a few decades ago the relationship between psychology and law was compared to an unhappy marriage. In the twenty-first century, it is rather regarded as a 'normal' marriage with its ups and downs, but definitely not heading for a divorce. There can be no doubt that both professions find the interaction and collaboration enriching.

However, in order to make full use of this interdependent relationship, much more needs to be done. The key lies in research, aimed not only at developing new procedures, methods, and techniques, but also at determining weaknesses in the general professional approach of forensic psychologists. Forensic psychologists should especially try to move away from the 'hired guns' stigma attached to them. This can only be achieved by telling 'the truth and nothing but the truth.' This principle can be best served by presenting objective and critical both-sides-of-the-story evidence without subjectively favoring one party. The saying, 'It is not a question whether forensic psychology is good, but what it is good at,' should also serve as a principle guideline for acknowledging that there are areas where little or no contribution can made and sticking to those areas where it is clear that psychologists are held in high esteem by courts. With such an approach and quality research, there can be no doubt that forensic psychology will flourish in the future.

Editor's Note

Forensics has been a strong growth area in anthropology, psychology, finance, and many other fields since 2000, due largely (it seems) to the popularity of television shows such as CSI or Bones; and the literature in these fields has grown rapidly. For recent overview, treatments of forensic psychology specifically, see Adler and Gray (2010). An important emerging field, not mentioned in Louw's original entry, is forensic neuropsychology. On this, see Larrabee (2012). Forensic psychology probably finds more application in criminal law and criminal behavior than in any other field; indeed, the popular image of the forensic psychologist has him or her on the stand testifying to the fitness of accused to stand trial. On crime and forensic psychology, see Bartol and Bartol (2012). The leading journals in the field are the American Journal of Forensic Psychology, founded in 1983 and published quarterly; and the Taylor and Francis Journal of Forensic Psychology Practice, published five times per year.

See also: Criminal Liability and Criminal Defenses; Expert Testimony; Forensic Psychiatry.

Bibliography

Ainsworth, P.B., 1998. Psychology, Law and Eyewitness Testimony. Wiley, New York.
Allan, A., 1995. The role of the psychologist in South African courts. Bloemfontein: University of the Orange Free State Ph.D. dissertation.
Alloy, L.B., Jacobson, N.S., Acocella, J., 1999. Abnormal Psychology. Current Perspectives, eighth ed. McGraw-Hill, New York.
American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders: DSM-IV, fourth ed. American Psychiatric Association, Washington, DC.
Adler, Joanna R., Gray, Jacqueline M. (Eds.), 2010. Forensic Psychology: Concepts, debates and practice, second ed. Taylor and Francis, London.
Bartol, C.R., Bartol, A.M., 1994. Psychology and Law. Research and Application, second ed. Brooks/Cole, Pacific Grove, CA.
Bartol, C.R., Bartol, A.M. (Eds.), 2012. Current Perspectives in Forensic Psychology and Criminal Behavior, third ed. Sage, Thousand Oaks, CA.
Blau, T.H., 1998. The Psychologist as Expert Witness, second ed. Wiley, New York.

Butcher, J.N., Miller, K.B., 1999. Personality assessment in personal injury litigation. In: Hess, A.K., Weiner, I.B. (Eds.), The Handbook of Forensic Psychology, second ed. Wiley, New York, pp. 104–126.

Ewing, C.P., 1985. Psychology Psychiatry and the Law: a Clinical and Forensic Handbook. Professional Resource Exchange, Sarasota, FL.

Golding, S.L., Skeen, J.L., Roesch, R., Zapf, P.A., 1999. The assessment of criminal responsibility: current controversies. In: Hess, A.K., Weiner, I.B. (Eds.), The Handbook of Forensic Psychology, second ed. Wiley, New York, pp. 379–408.

Gudjonsson, G.H., Haward, L.C.R., 1998. Forensic Psychology. A Guide to Practice. Routledge, London.

Larrabee, Glenn J. (Ed.), 2012. Forensic Neuropsychology: A Scientific Approach. Oxford University Press, New York.

Melton, G.B., Petrila, J., Poythress, N.G., Slobogin, C., 1997. Psychological Evaluations for the Courts, second ed. Guilford Press, New York.

Memon, A., Vrij, A., Bull, R., 1998. Psychology and Law. Truthfulness, Accuracy and Credibility. McGraw-Hill, New York.

Saywitz, K.J., Goodman, G.S., 1996. Interviewing children in and out of court. In: Briere, J., Berliner, L., Bulkley, J.A., Jenny, C. (Eds.), The APSAC Handbook of Child Maltreatment. Sage, Thousand Oaks, CA, pp. 217–318.

Tsushima, W.T., Anderson Jr, R.M., 1996. Mastering Expert Testimony. Erlbaum, Mahwah, NJ.

Formal Methods of Cultural Analysis

John W Mohr and Craig Rawlings, University of California, Santa Barbara, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Our focus is on methods of cultural analysis, and specifically on those methods that are formal in the sense that they rely upon the purposeful gathering (or simulating) of cultural data and a systematic analysis that involves at least some mathematically based technique. The meaning of culture is more complex, as scholars have understood cultural phenomena in very different ways across the decades and across the disciplines. These different goals in cultural analysis give rise to different types or styles of formal methods of cultural analysis. We briefly review the history of these methods in the social sciences and conclude with a description of some of the main arenas within which formal methods of cultural analysis are being deployed and actively developed today.

What Is a Formal Method of Cultural Analysis?

We describe developments in the use of formal methods of cultural analysis in the social sciences, focusing on approaches that rely upon the gathering (or simulating) of data concerning matters of culture. By this classification, it is precisely the use (either ultimately or potentially) of formal measurement practices and the construction of a data set that can be analyzed according to logical, relational, or other quantitative methods that define an approach to cultural analysis as formal in the sense of the term used here. We begin our article by describing different types or styles of formal analysis of cultural data. We then review some of the history of how these methods have been used for analyzing culture in the social sciences. We conclude with a description of a few of the main arenas in which formal methods of cultural analysis are currently being developed and deployed by research scientists.

Our title indicates a focus on formal *method*s of cultural analysis suggesting a survey of techniques but a more accurate description would be to say that we focus here on alternative research *programs* for studying culture each of which includes a shared tradition, organized communities with hierarchal social structures, shared ways of scientific knowing linked with (articulated into) shared systems and styles of practice which implies immediately that these are also technological ways of knowing and technological ways of doing (see Bourdieu, 1990; Pinch, 1986). A focus on methods alone strikes us as less productive since methods are never really neutral. For one thing, there are many examples where the same styles of data collection (e.g., interviewing or surveying) or the same styles of analyzing data (network analysis, linear modeling, or Boolean algebras) have been deployed in the service of quite different types of research goals.

The meaning of culture is inherently difficult to pin down, as scholars have understood cultural phenomena in quite different ways across the decades and across the disciplines, a topic that we take up here as well. As a start, recall that Raymond Williams (1977) traced the origins of the concept of culture from its early roots referring to "the growth and tending of crops and animals, and by extension the growth and tending of human faculties" (p. 11) through a series of changes and developments. With the rise of Romanticism, Williams notes that culture takes on a new meaning as something set against the character of modern life "as distinct from or actually opposed to 'civilization' or 'society' in its new abstract and general sense … 'culture' as a general process of 'inner' development was extended to include a descriptive sense of the means and works of that development: that is, 'culture' as a general classification of 'the arts,' religion, and the institutions and practices of meanings and values" (p. 14–15). Williams identifies a second break in the concept of culture that he traces back to the original (eighteenth century) insights of Giambattista Vico and Johann Gottfried Herder on culture as styles of life, which then emerge full-blown in the nineteenth century associated with the idea that "a fundamental social process which shapes specific and distinct 'ways of life' is the effective origin of the comparative social sense of culture and its now necessary plural 'cultures'" (p. 17). While the theory of culture has changed, in some ways this age old split in thinking about culture still resonates today with how formal methods are applied to the study of culture today. On the one hand, a focus on culture as primarily concerned with domains of creativity and the arts, and on the other, those who would instead emphasize the study of culture as a 'way of life' founded perhaps on core patterns of understanding, believing, valuing, or moralizing. In this case, as Clifford Geertz offered a definition of culture as "an historically transmitted pattern of meanings embodied in symbols, a system of inherited conceptions expressed in symbolic forms by means of which men communicate, perpetuate, and develop their knowledge about and attitudes toward life" (1973: p. 89).

For analytic purpose we distinguish between two types of formal research programs for studying culture. The first type, which we classify as interpretative or hermeneutic styles of cultural methods have as their goal advancing our ability to know or to understand a shared cultural meaning or form of collective understanding (or to simply assist in providing a reading of a cultural text). These are methods that are often grounded in views of culture that see systems of shared meaning (or core logics of understanding) as being what most deeply anchors an effective cultural analysis. Methods of content analysis are one example of a style of cultural research programs that are frequently associated with hermeneutic research goals. Hermeneutic style projects tend to focus on

speech activity (including texts) and their formalisms are generally geared toward helping the analyst to make sense of the speech data, system of discourse, or other forms of cultural expression. Here, cultural meanings can include larger types of shared cultural patterns (like ideologies, fields, and hierarchies of artistic tastes or institutional logics) (Friedland and Alford, 1991; Geertz, 1964; Swidler, 2003; Thornton et al., 2012; Wuthnow, 1989). But they might just as well be directed at smaller groups or organization level meaning systems, including those that Gary Fine calls idiocultures (Fine, 1979) or what Harrison White describes as netdoms – conjunctures of identities, networked together in a time and a place, with a foundation in shared collective meanings, ways of talking together and repertoires of stories. White says that this "presupposes the mixture of relation and topic, plus understanding" (White, 2008).

We classify other forms of measuring culture as non-hermeneutic. Here the goal is to measure aspects or traces of cultural forms, and then to use those measures as indicators (or factors) for understanding or explaining other social phenomena. One might also say that the goal of a non-hermeneutic program of cultural analysis is at least potentially a simpler approach to measurement since it comes without the added ambition of needing to also decipher the hermeneutic meaning of the cultural data that has been collected. DiMaggio sums it up, "Nothing is more central to the study of culture than the interpretation of meaning, yet nothing is more difficult" (DiMaggio, 2011: p. 294).

Some History

Early Uses of Data to Measure Culture

As states and state apparatuses grew, so too did their capacities and appetites for the gathering of systematic data on national, regional, and local populations (for taxation, conscription, tariff collection, and other administrative functions). They also developed measures of 'economic' indicators – the money supply, forms of wealth, the circulation and availability of commodities, etc. (Porter, 1986; Spengler, 1961) These endeavors were constitutive of the basic discursive, professional, and organizational cultures and informational infrastructures that facilitated the rise of modern states, political economies, and systems of surveillance. As this data was collected, so too were social scientists developing new mechanisms for measuring the types and characteristic qualities of the people who made up these nations (Foucault, 1970; Fourcade, 2010; Hacking, 1975, 1990; Lazarsfeld, 1961; Porter, 1986; Stigler, 1986). And as soon as these official statistics were being gathered, researchers began to make creative use of those measures to describe and analyze matters of culture.

In his classic review of the origins of quantification practices in sociology, Lazarsfeld (1961) recounts numerous efforts by pioneering social scientists to take the measure of norms, values, moral sentiments, and systems of shared meanings. Lazarsfeld (1961) dates the origins of modern demography to the work of Graunt in 1662. Spengler (1961) tells us that estimates of the supply of money were being made as early as 1694. State data efforts in this era were, however, spotty, unreliable, or altogether unavailable (in no small part because local populations were properly suspicious of state enumeration efforts). For many years, clever estimation strategies were relied on. As Lazarsfeld explains, "(T)he ingenuity of early scholars was directed mainly toward obtaining estimates of population size and age and sex distributions from meager and indirect evidence. Multiplying the number of chimneys by an assumed average family size or inferring the age structure of the population from registered information regarding age at the time of death were typical procedures in what was then called political arithmetic" (1961, pp. 149–150). But it was not only demography that was being estimated. The specialists in these arts, the political arithmeticians, prided themselves on showing the relationship between national populations (which were thought to express very specific characters), the nature of particular state administrations (which were often viewed in highly personalized terms), and the principles of good governance (which were a curious mix of theistic moralities and Aristotelian inspired protoscientific rationalities). In the process, topics regarding culture were mixed in with demographic facts, mercantilist tallies, and pseudosociological folk theories (often built around rhetorics of national historical dramas) (Lazarsfeld, 1961; Porter, 1986).

Elsewhere, Lazarsfeld sums up what he sees as the main trend in the early years of culture modeling by explaining that in the mid-nineteenth century, "Le Play assessed the religious feelings of the families he studied by examining their budgets and noting how much money they spent on candles for church masses. It was objected that church attendance might be a better indicator of the religious feelings Le Play sought to measure, and thus slowly the notion of time budgets – in addition to money budgets – developed. It was then further argued, that perhaps church attendance is indicative only of conformity to social customs; it is the attitude toward religion which really matters. And so attitude measures began their triumphant course" (Lazarsfeld, 1970: p. 65).

These early efforts at measuring culture have a distinctly Durkheimian flavor. The importance of culture is its role in the collective conscious – or the social glue that provides cohesion through shared beliefs and sentiments. Largely absent are the conscious deliberations over various interpretations of cultural objects, as well as more complex *systems* of meaning in which cultural processes unfold. Culture is subsumed by social structure, while the structures of cultural systems are not an empirical concern. It would take some time for culture to escape the confines of this larger functional edifice.

Kroeber's Analysis of Fashion Trends

An early example of a social scientist who sought to develop a precise set of formal measures of cultural forms was Alfred Kroeber (1919) who studied changes in women's dress fashions. Kroeber systematically coded a sample of images from a collection of women's fashion magazines (measured annually from 1844 to 1918). By taking basic measurements of the models and their gowns (the length of the figure, length of dress, width of dress at hem, width of dress at waist, etc.) and analyzing these data across time, Kroeber found some clearly visible and apparently cyclic trends in women's dress fashions. In one graph, for example, he shows the quick and unidirectional growth in the width of the skirt, neatly capturing the

fashion in crinoline hoop skirts out to an apex, followed by a steep decline (save for a short sartorial diversion by way of a widening of trains on the dresses). The data that are used for these analyses are very simple measures of size and ratio in women's evening fashions. This is cultural data but it is not hermeneutic data. In other words, whether meanings are ultimately at work in determining what goes on here or not, what was actually being recorded as data by Kroeber are just a simple set of measures assessing the geometric dimensions of the cultural forms (evening dresses), measures that turn out to be useful for illustrating increments of change in cultural forms over time. Kroeber's work is, in this sense, a classic and early demonstration of how nonhermeneutic measures of culture could be collected and usefully analyzed.

Stanley Lieberson (2000) finds a similar type of effect (80 years later) when he studies baby names. Lieberson calls this the ratchet effect. The cycle is explained by the idea that tastes involve establishing a distinctiveness, a difference from less-valued taste categories and that in a system such as women's fashion, part of what is valued is *the new* (as distinct from the old) and thus an important marker of the new is the length of the dress. If hem lines do not change, then the newness of the fashion is harder to perceive. So they must be different than last year's hem line. But if they go backward, to the length of the previous year, then again there will be a danger that the new will be confused with the old, with the really old – that is, with the *passé*. As a result the hem has to continue moving in the direction it has been moving in. And this continues until the hem hits an obstacle (the floor or some measure of moral propriety) at which point it has to reverse itself.

Lieberson (2000) shows that given names also exhibit the ratchet effect. Names turn out to be a good source for cultural data because they are collected very systematically (e.g., in birth records) and because they have increasingly become defined as free choices and thus they are subject to the vagaries of style. Names become appealing to a lot of people at the same time. As this happens, the name becomes more and more popular until such point that a given name will become too popular. As an explanation, Lieberson offers a basic thought experiment. Suppose you are about to name your child and you were searching for a name that appealed to you. And you hit upon a certain name. But you looked about you and you saw lots and lots of people with that name. So many people that you wanted your child to be different, to be special, to be given their own more unique name. And so you pick something else.

Lieberson describes both endogenous and exogenous effects. The former focuses on the means according to which taste moves. This depends on a largely tacit set of cultural affinity processes. Choices happen because within the sea of available names, some are in one's mind, others are not. Some sound nice to one's ear and others do not. Some seem like good things to choose and others do not. Culture in the end is about how the self is subjectively tuned into a complex, flowing field of cultural meaning systems. The choice of a child's name reflects how some social constellation of influences, embedded within an interlocking set of family, friend, and significant other networks, usually including a fantasized network of media personalities, are affected by the rise and fall of aesthetic appeals, and the ebb and flow of identity dilemmas and associations, all of which weigh down upon the moment when a child is announced to have a given name. Compare this with the theory of culture held by Kroeber. Like Durkheim, Kroeber was trying to demonstrate institutional effects, seeking to show that there are things that occur in cultural life that operate at a level, which is not just at the experience of a particular individual. But unlike Lieberson, he has little to turn to as a means of explanation. And, in the end, when he ponders about what might account for the regularities of culture he suggests, "The super-organic or superpsychic or super-individual that we call civilization appears to have an existence, an order, and a causality as objective and as determinable as those of the subpsychic or inorganic" (p. 263). In this sense, going beyond Durkheim, Kroeber creates a semiautonomous realm in which cultural processes unfold according to a guiding internal logic. The value placed upon 'newness' may derive from broader institutional contexts, but the cultural manifestation of this value is guided by the logic of an interrelated set of symbols and their behavioral correlates.

Lévi-Strauss on Formal Methods for Interpreting Myths

Contrast Kroeber's ambitions with those of the French anthropologist, Claude Lévi-Strauss. Lévi-Strauss is famous for his contributions to the original development of structuralism. Coming to the field a generation after Kroeber, Lévi-Strauss made a huge impact on anthropology by applying the principles of structural linguistics (and formal semiotics) to nonlinguistic matters of culture. Following in the path originally pioneered by Swiss linguist Ferdinand de Saussure and later by structural linguists such as Roman Jakobson (whom Lévi-Strauss met in New York during World War II while teaching at The New School), Lévi-Strauss used structuralist principles to analyze anthropological data on kinship structures and myths (Lévi-Strauss, 1963). Kinship structures are the single most important institution in traditional societies and Lévi-Strauss began his career by showing that structuralism provided a new way of interpreting the logic of kinship systems. Later he turned to the study of myths in traditional societies and it was especially here that we can see how Lévi-Strauss provides an early example of a social scientist seeking to use formal methods of analysis to approach what was an explicitly hermeneutic (or interpretive) research goal.

For Lévi-Strauss, the goal was to understand how the interpretation of a myth can be facilitated by analyzing the deeper relational structures (the bundles of relations linking concepts and actions within the myth). Lévi-Strauss saw that these structures provided an order for the content and the meaning of the myth, and ultimately it was the structural logic that made the whole greater than the sum of the parts. Specifically, Lévi-Strauss looked to develop analogues to the way that linguistic structures worked in language. Rather than investigating the 'morphemes' of language structure, however, Lévi-Strauss gathered up and analyzed the 'mythemes' (or 'gross constituent units' of meaning) that make up the content elements of myths.

In his essay on "The Structural Study of Myth," Lévi-Strauss illustrated this approach by deconstructing the Oedipus myth. Lévi-Strauss' technique depends on breaking the myth into simple sentences (or 'statements') and then looking for

commonalities in the semantic functions that each sentence performs. In the case of the Oedipus myth, the statements are semantically (and thus structurally) divided into common bundles of relations. Thus Lévi-Strauss' interpretation of Oedipus: "The myth has to do with the inability, for a culture which holds the belief that mankind is autochthonous [born from the earth], to find a satisfactory transition between this theory and the knowledge that human beings are actually born from the union of man and woman. Although the problem cannot be solved, the Oedipus myth provides a kind of logical tool which relates the original problem – born from one or born from two? – to the derivative problem: born from different or born from the same? By a correlation of this type, the overrating of blood relations is to the underrating of blood relations as the attempt to escape autochthony is to the impossibility to succeed in it" (p. 216).

Lévi-Strauss' approach is hermeneutic in the sense that his goal is to employ this style of cultural analysis to facilitate his interpretation of the meaning of the myth. It is formal because he gathers up data about the content of myths, he measures the distribution of the mythemes and the relations between them, and he provides us with an analysis (based on a system of Boolean transformations) that provides us with a plausible reading of one meaning of the myth. It is explicitly through his linking of the analysis of patterns in his collected data to his research goal of interpreting the meaning of the myth that Lévi-Strauss provides what is an especially useful example of a hermeneutically oriented approach to formally measuring cultural data. But, as Pierre Bourdieu would later show, Lévi-Strauss' structural approach to culture leads to a largely disembodied mapping of meaning, overlooking the important lived territories and material conditions that shape meanings and give rise to various symbolic struggles.

Developments of Cultural Measurements in Sociology through Community Studies and social Work Theory

In the early years of professional social science, there were a variety of quite different programs of research that sought to develop formal measures of culture. One stream of work developed out of the efforts to conduct neighborhood surveys such as those that were being used by Charles Booth to study the London slums in the late-1890s. These ideas were later applied (in 1896) by W.E.B. Du Bois in his survey of a black neighborhood of Philadelphia, which was followed by the Pittsburgh survey of 1907 (and dozens of other city surveys that came after that). These were intensive collections of data, even though the procedures were not the same as what we might imagine when we think of a modern, systematic, door-to-door canvassing with a survey questionnaire. "Rather it was an effort to provide an inventory and an overview of the state of the city, for which the investigators were omnivorous in their methods of data collection" (Converse, 2009[1987]: p. 24). Much of the data came in the form of the schedule, which "in the hands of the social surveyors was an instrument for making observations or for conducting interviews with respondents, or a mixture of both" (Converse, 2009[1987]: p. 34). With funding from the Russell Sage Foundation, the Pittsburgh initiative had 74 field staff compiling a mass of detailed information about all kinds of social phenomena. For example, the checklist for observing clothing worn by inhabitants included categories for, "spotted … dusty … torn … worn … patched … mussed … wrinkled …" (p. 34). Researchers compiled information by collecting "case-history interviews, which were gathered and then counted and compared among some dimensions, thus providing a 'casemounting' that represented a merger of the case study and statistical methods" (p. 34). But, as Bulmer points out, "the Survey used quantitative data in more of an exploratory and descriptive than analytic way" (1996: p. 18). Here and elsewhere, "It was as if there was an intuitive sense of the value of collecting extensive data about individuals in the population being studied, without the necessary knowledge either about sampling or how to handle the data once collected other than to compute simple counts of characteristics and then treat respondents on a case-by-case basis" (p. 26).

A different force for developing measures of culture came from another wing of the professional social work community. While community workers (including settlement house workers) had been visiting and surveying slum residents since the mid-1880s, by the turn of the century social workers were becoming more professionalized, and this included a strong push to rationalize and quantify social workers' field observations and to use those data in productive analytic ways. From her post as director of the Charity Organization Department at the Russell Sage Foundation, Mary Richmond (1917) sought to promote a style of professionalization in the field of social work that was explicitly modeled after the medical profession. Social workers were encouraged to gather systematic data from their cases in order to create a coherent system of data about behavior, character, and culture that could be used for diagnosing and treating the social pathologies that lead to poverty and ruin.

Meanwhile, professional sociologists were also studying communities. One of the most important of the early studies was the research collaboration between W.I. Thomas and Florian Znaniecki (1918) concerning the lives of Polish immigrants. Drawing particularly on letters written to and from family members in Poland, they emphasized the utility of using these types of documents for sociological research. "We are safe in saying that personal life-records, as complete as possible, constitute the perfect type of sociological material, and that if social science has to use other materials at all it is only because of the practical difficulty of obtaining at the moment a sufficient number of such records to cover the totality of sociological problems, and of the enormous amount of work demanded for an adequate analysis of all the personal materials necessary" (Thomas and Znaniecki 1958: 1832, as cited by Łuczewski, 2009).

Using these materials, Thomas and Znaniecki combined their intellectual interests. Thomas was mostly focused on attitudes while Znaniecki was concerned with values, but together they produced a solid grounding and advance for social science's understanding of how to analyze cultural forms. Łuczewski (2009) points out that for Znaniecki "The great mistake […] was to treat the problems of values […] as ultimate and self-sufficient, instead of taking them only as starting-points of future investigations." Thus he described values on two levels: content (ability to become an object of experience either sensual or imaginative) and meaning … In his relational perspective, meaning of a given value depended on its relations

with other values. Simply put, the meaning is the relation (2009, p. 8). These studies helped early on to situate the formal analysis of culture within the lived experience of groups seeking to use culture to achieve practical ends. This more Pragmatic view of culture, and thus the reasons for measuring cultural artifacts, resurfaces much later in the work of Pierre Bourdieu, which we discuss below.

Surveying Values, Attitudes, and Opinions

The invention of attitude measurements provided the means to use sample surveys to measure cultural content systematically. Not long after Thomas and Znaniecki published their work (1918), both Floyd Allport and Emory Bogardus published independent methods for formalizing the study of attitudes through the use of 'attitude scaling.' Later, Thurstone, Likert, and others produced rapid advances in the use of social surveys for measuring subjective experience, including cultural orientation, values, systems of meanings and beliefs, and subjective understandings of social situations (Converse, 2009[1987]). These developments coincided with the emergence of independent polling bureaus and university centers that conducted independent surveys, as well as the growing sophistication of population sampling and statistical methods for gauging significance and for interpreting scales measuring subjective experience. This includes cultural orientation, values, systems of meanings and beliefs, and subjective understandings of social situations (Converse, 2009[1987]).

World War II was a critical juncture. It brought academics together into large, multidisciplinary, state-funded research teams. In the case of survey research, Samuel Stouffer led a team of social scientists who surveyed over a half million U.S. Military personnel through the war years. Surveys asked about complicated matters like troop morale, racial integration, and officer performance. By the end of the war, survey methodologies had been improved and their legitimacy had been secured. Over the course of the next two decades, social scientists like Paul Lazarsfeld (Director of the Columbia Bureau for Social research) helped to build up the conceptual, theoretical, and practical foundations of modern social survey science, writing essays, and commissioning papers about how to construct effective survey questions, how to use the data to build variables, how to use cross-tabulations to compare variables, how to assess statistical significance and to run regression models, eventually establishing survey research as the new "language of empirical social research" (Lazarsfeld, 1968: p. vii).

Survey research has continued to improve since those years, developing into what is the social sciences' most mature and essential means of collecting data about culture and social life (Schuman, 2008). In recent years, survey analysis methodologies have been undergoing a renaissance as new scholars have come to the field and challenged traditional assumptions. Steve Vaisey (2009) helped to bring the dual-process model of information processing over from cognitive science, arguing that much of how people operate is indeed nonreflexive and driven by habit, but that other parts of human cognition is active and deliberative and that we should use our surveys more strategically to unearth these effects. Vaisey and Lizardo (2010) applied the dual-process model to look at how people's world views impact the character of the types of social networks that they form. Amir Goldberg's (2011) development of relational class analysis is another example of recent innovations in survey research. Instead of comparing respondents on the basis of having similar attitudes, relational class analysis emphasizes comparing respondents who have similar beliefs about the relations between attitudes.

Some Contemporary Approaches to the Formal Analysis of Culture

Measuring Meanings with Cognitive Science

Other areas of social science have also focused on the formal analysis of culture and meaning. In 1957, Charles Osgood (and his first two graduate students, George Succi and Percy Tannenbaum) published *The Measurement of Meaning*. The monograph introduced and demonstrated their efforts to use 'the semantic differential' as a tool for formally measuring the connotative aspects of word meanings. Their basic idea was to present respondents with a key word (such as 'father') followed by a series of paired opposites presented as seven point scales, happy/sad, hard/soft, slow/fast, and the respondent is asked to rank their association of the key word to each of the scales. Osgood and colleagues explain, "The semantic differential is essentially a combination of controlled association and scaling procedures. We provide the subject with a concept to be differentiated and a set of bipolar adjective scales against which to do it" (1957, p. 20). Factor analysis was used for simplifying the data. Of course a critical issue was identifying the right paired oppositions and these types of matters occupied Osgood and his colleagues for many years. Nonetheless, by the time the Snider and Osgood volume, *Semantic Differential Technique: A Sourcebook*, was published in 1969, the editors could report that "the semantic differential technique has gradually captured the imagination of psychologists…in recent years, hundreds of articles using the technique have appeared in professional journals" (1969, p. v). One direction these studies moved in was in the application of semantic differential tools to comparative studies of cultural meaning.

Much of the research by cognitive scientists tends to focus on the more cognitive and/or biological sides of cognitive processes. But these types of tools were quickly embraced by one group of anthropologists who had an explicit focus on culture, a story that is effectively told by one of the field's leaders, Roy D'Andrade (1995). About the time that Osgood and his colleagues were preparing to publish their book on meaning measurement, a pair of anthropologists, Floyd Lounsbury and Ward Goodenough, were publishing two new papers on the use of componential analysis, a method that was used to breakdown and analyze the core meanings of kinship systems (1995, 21). As D'Andrade explains, "The reason that the kinship terminology papers by Lounsbury and Goodenough had such a large impact was that they presented a rigorous method for identifying 'idea units' and analyzing the organization or structure of these units. While developed specifically for the analysis of kinship terms, it appeared that the general principles involved in identification and analysis could be extended to other domains. It was this body of methods and goals which was to become the agenda of cognitive anthropology" (1995, p. 17). By the mid-1960s this

field was growing. D'Andrade recounts the move from formal studies of kinship terms to a variety of studies of language terms (such as words for firewood) and from the use of simple schema toward more elaborate models for measuring cognition. DiMaggio (1997) provides an account of how studies in cognitive science intersect with research on culture.

Content Analysis

While the careful and ordered reading of texts goes back many centuries through various traditions of hermeneutics, the systematic treatment of texts as data in the more formal sense that we have employed in this article is relatively new. Harold Lasswell was an important contributor to these developments. After having studied the use of propaganda during World War I for his doctoral dissertation at the University of Chicago, Lasswell was well prepared to step into the position as director of the 'experimental division for the study of wartime communications, established at the Library of Congress during World War II' (Lasswell et al., 1949, 1952: p. 40). In this position he helped devise rigorous sets of procedures for coding textual material in a way that was intended to measure cultural and group-level discourse processes. After the war, Lasswell used these methods to coordinate a study of a corpus of 20 000 newspaper editorials sampled from 10 major newspapers (in five countries) between the years of 1890–1945 during which many of these conventions and strategies were published (Lasswell et al., 1952).

Content analysis tools continued to improve. Philip Stone and colleagues at Harvard (1966) developed the General Inquirer, which was a general purpose computer-based content analysis program that relied upon the building of specific dictionaries for various focused investigations. By the 1980s sociologists like Roberto Franzosi (1989) were using semantic grammars as a way to code the meaning of texts. These formal methods and project styles come together especially well in John Markoff's (1996[2004]) analysis of the origins of the French revolution (which builds on collaborative work done with Gilbert Shapiro). Drawing upon rich textual material preserved in the *cahiers de doléances* (the official lists of grievances drawn up in 1789 by rural communities, and others), Markoff is able to follow the flow of discourse back and forth between the legislators and the peasants across both time and space. Here the texts are interpreted with the help of formal tools and the lessons that are learned imply changes in how we now understand the role of the peasantry within the system of discourse and political maneuver around the years of the revolution.

One of the most productive offshoots of the content analysis tradition has been in the area of scientometrics – that is, a branch of formal analysis concerned with the sociology of knowledge and its measurement in research publications. A key component of this lineage is citation studies as represented, for example, in the many achievements of Loet Leydesdorff (2001) and his colleagues. This perspective also reflects an early blending of text analysis and network analysis, as authors, articles, and even journals are analyzed through the networks they form in coauthorships and cocitations. Still, in a number of ways, research projects in this field are being supplanted by new problems and new methods coming from the developments in computational analysis that are now being incorporated into fields like sociology, linguistics, and the humanities (see below).

Cultural Capital Studies

In a sense the French sociologist Pierre Bourdieu was foundational to the modern use of formal methods of cultural analysis in several important ways. One way was in the stream of research concerning the role of culture in social stratification and especially research surrounding the concept of cultural capital (Bourdieu and Passeron, 1977). The original insight was linked to the argument that participation in forms of high culture was a mechanism by which elite households were able to pass along hidden advantages to their children. Children from privileged households are inculcated with elite styles of cultural knowledge and forms of practice that enable them to display the required levels of what Bourdieu and his colleagues call cultural capital. The key insights here – that culture was a resource, one that was associated with mechanisms of class reproduction, and that these forms of cultural mastery could be identified using standard survey techniques proved to be a powerful catalyst for new streams of measuring culture. Paul DiMaggio borrowed the notion of cultural capital from Bourdieu and used a factor analysis to construct a set of measures of cultural capital that proved to be effective in predicting success in both educational and marital attainment (DiMaggio, 1982; DiMaggio and Mohr, 1985). Both the DiMaggio research and the ongoing stream of work by Bourdieu himself led to the production of a wide stream of research concerning the effects of cultural capital on stratification outcomes in different countries and across different institutional systems (e.g., DeGraaf et al., 2000; Kane, 2000; Roscigno and Ainsworth-Darnell, 1999; Van Rees et al., 1999). One especially influential reformulation grew out of a classic paper on the phenomena of 'cultural omnivores' by Peterson and Kern (1996) showing that increasingly, high-status culture consumers focused less on specializing entirely on exclusive cultural tastes and instead began to acquire a broad set of cultural engagements.

More recently, this field of work has morphed again as a range of new projects on measuring the relationship between culture and social structure have begun to take center stage. Omar Lizardo (2006), for example, using data from the General Social Survey, finds that measures of cultural tastes predict an individual's network style. Specifically, individuals with higher levels of culture participation are found to be more likely to have a greater number of strong network ties, whereas engagement with popular culture tends to be associated with larger weak-tie nets. The general sense here, as articulated in an important essay by Pachucki and Breiger on *cultural holes* (2010), is that scholars are increasingly coming to pay attention to how it is that culture acts as a general social resource that leads to the creation of things like social networks, thereby inverting the more traditional assumption that it is networks, which are primary and culture that is epiphenomenal.

Field Theory

A second important legacy of Bourdieu's research program is the recent growth in what is known as field theory. The idea of

describing social situations as being defined by 'field effects' goes back to several sources. The Berlin Gestalt psychology made famous the notion of fields in structuring perceptions. In social psychology, Kurt Lewin was instrumental in developing models of how to formally model social fields. Lewin himself drew upon the time he had spent with the Gestalt scholars and also the German philosopher Ernst Cassirer during his years at the University of Berlin. Borrowing also from physics and especially Einstein's conception of fields, Lewin (1951) sought to measure cultural phenomena by applying the idea of a field to an understanding of socially constructed interactional space. Much of his work drew upon studies of children's behavior conducted during his years as director of the Child Welfare Research Station at the University of Iowa.

Lewin began to use the field concept to construct models of an individual's 'life space,' which he described as "the person and the psychological environment as it exists for him" (1951: 57). These were essentially cognitive maps of individual's choice situations. They included "specific items as particular goals, stimuli, needs, social relations, as well as more general characteristics of the field as the atmosphere (for instance, the friendly, tense, or hostile atmosphere) or the amount of freedom." The definition was pragmatic. The life space included within it "everything that affects behavior at a given time" (1951: 241). The key goal for Lewin was to bring a level of formal analysis to the study of social psychology. Field theory worked well because of the type of mathematics that defined it. Inspired by the work of the German philosopher Ernst Cassirer on relationalism versus substantialism, Lewin was convinced that life spaces were essentially relational systems. Objects in the life space stood in particular relation to one another but their location could not be defined in precise metric terms. It was impossible to specify precise dimensions, linear measurements, or definable coordinate systems within which objects could be located inside an individual's life space, but field theory provided a way to measure these things in the purely relational sense of topology theory. Extensions of Lewin's ideas about measurement theory found their way into modern social network analysis, which is like Lewin's field space in that it is based on purely relational measures (Mohr, 2013).

Later in his career, Pierre Bourdieu began to focus on the study of fields, a term he used to designate "relatively autonomous social microcosms" corresponding to regions of institutional life. For example, Bourdieu studied the fields of art, academia, religion, and law. Each field is understood to be defined by a set of social relationships and a shared collective understanding about the meaning of what goes on inside the field or, as Bourdieu puts it, each field consists of "spaces of objective relations that are the site of a logic and a necessity" (Bourdieu and Wacquant, 1992, p. 97). Like Lewin, a primary reason that Bourdieu adopts a field approach is that it made it easier to theorize and measure social life and especially matters of culture. As Bourdieu explains with respect to his preference for correspondence analysis models, "if I make extensive use of correspondence analysis, in preference to multivariate regression for instance, it is because correspondence analysis is a relational technique of data analysis whose philosophy corresponds exactly to what, in my view, the reality of the social world is. It is a technique which 'thinks' in terms of relation, as I try to do precisely with the notion of field" (p. 96). Bourdieu (1984) famously employed correspondence analysis techniques in his studies of the association between social class and cultural tastes in France.

Field theory therefore brings together the embodied agency of actors and the meaningfulness of relational systems that enable and constrain action. The formal analysis of culture – the arraying of categories and identities in some relational space defined by those participating in these cultural forms – therefore reveals a state of the struggle for positions and dominance according to the underlying logic of a field. Inspired by Bourdieu's research, a number of other areas of social science have spawned studies of field. Institutional theorists in organizational studies have focused on studying fields (Powell and DiMaggio, 1991; Fligstein, 2001; Fligstein and McAdam, 2012; Martin, 2003; Padgett and Powell 2012). Within this group, a strong emphasis on cultural analysis has emerged among those who study organizational fields as being fundamentally grounded in a system of meanings described as institutional logics (Friedland and Alford, 1991; Friedland et al., 2013; Thornton et al., 2012).

Networks and Culture Perspective

Paul DiMaggio argues that "network analysis is the natural methodological framework for empirically developing insights from leading theoretical approaches to cultural analysis" (2011, p. 286). He goes on to suggest that "Social scientific work on culture has three analytic foci: formally organized systems that produce and distribute cultural products; expressive symbols that facilitate the production of individual and group identities and intergroup boundaries; and the symbolic organization of meaning. Relational theories are central to each of these topics" (DiMaggio, 2011, p. 286).

The interest in relationality comes up again and again in these research projects and this is partly a reflection of the common lineages of the work. Jacob Moreno observed network effects in his study of small groups during the 1930s, providing the first real example of modern network analysis. But Kurt Lewin's work on field theory was also connected to the emergence of modern network science. One link occurred through the work of Lewin's student Alex Bavelas (1950) who applied ideas from Lewin's hodological analysis of field space to assess the properties of networks. Later, Lewin's ideas were recast in graph theoretic terms by Dorwin Cartwright, Lewin's successor as Director of the University of Michigan Research Center for Group Dynamics (Harary et al., 1965; Mohr, 2013). Network theory also has deep connections with formal traditions of kinship analysis that were being conducted by anthropologists (Breiger, 2004). Indeed, Harrison White who went on to become the leading figure in social network analysis in American sociology during these years started out his career in sociology in part due to his fascination with Lévi-Strauss' work on kinship algebras (White, 1963).

In its early years (1960s–1970s) the sociological field of network analysis was explicitly anticultural in its approach, the feeling being that network ties were material, concrete, and measurable, while culture was none of those things. But by the early 1990s that sentiment had begun to change. Harrison White made a pivot toward culture when he published the first edition of *Identity and Control* (1992) in which he lays out an

expansive theory of social life that relied quite explicitly on theorizing the intersection of culture and networks. Karen Cerulo (1988) had already been working on matching semiotic theory with the empirical analysis of data from cultural artifacts that included visual symbols and written music. A number of other scholars had begun merging social network methodologies with the analysis of cultural data. Carley (1994) used network models to analyze the meaning structure of science fiction novels, Bearman and Stovel (2000) used network graphs to analyze narrative structures in Nazi party members' life stories, and Smith (2007) did something similar with people's narratives about conflicts they have lived through in Istria. Network analysis reveals how some narrative elements can link disparate narrative structures together as a kind of cultural bridge.

Much of this work makes use of two mode data, meaning that relationships are examined that span across two categories or domains of things – relations that link individuals and groups, members and committees, or categories of persons and categories of treatments. Breiger (1974) did the original work on this in his development of a formal model for analyzing the linkages between individuals and groups as a structural duality. These types of cultural models were then used by Mohr and Duquenne (1997) when they analyzed the implicit institutional logic of nineteenth-century poverty directories with Galois lattices to assess how organizations matched categories of poor with poverty treatments regimes. Mische and Pattison (2000) also used lattices as a way to analyze the duality between Brazilian youth organizations and their projects, which they define as "evolving, imaginatively constructed configurations of desired social possibility, accompanied by an implicit or explicit theorization of personal and∼or collective capacity to act to achieve that possibility." Yueng (2005) uses lattices to compare members of different communes according to the types of ways that commune members characterize their relationships with one another. Martin (2000) compares the kinds of animals and the types of occupations that they hold in children's books of Richard Scarry.

The New Computational Approach to Culture

Computational Linguistics

Linguistics has long been more scientifically advanced than the other social sciences in the sense that more progress has generally been made in understanding the basic rules and principles of linguistic operations, and much of that is due to the elaboration of sophisticated theories and methods of structural analysis of language systems. But with the rise of the digital society, linguistics has also become both a data science (dedicated to processing large amounts of textual data) and an engineering science focused on trying to create useful technologies.

Whether linguistic projects of cultural analysis are also grounded in hermeneutic styles of culture analysis depends. In some sense, all of linguistic science has an interpretive goal since it is all geared toward analyzing and understanding language and speech activity. But, on the other hand, there are also many different levels and styles of understanding in linguistic science. Syntax tries to account for language use at the level of the sentence, while semantics and pragmatics deal with the relation between these lower-level language processes and the broader problems of meaning and language use. Moreover, modern linguistic science has moved well beyond purely structural analysis. As Hajic (2004) notes, "Today we can say that one of the turning points for linguistics was Claude Shannon's work (1948)" (p. 80). Shannon provided the research that lays out the grounding for modern communication theory. It was built upon a long tradition of mathematical equations that had been developed to manage the engineering of telegraph signals. Shannon and Weaver introduce their formulation by writing, "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages" (1949, p. 3). Shannon's approach to communication theory has come to provide one foundation for how engineering researchers such as computer scientists approach the measuring of culture, through the use of stochastic based models, an issue that turns out to have relevance in other fields of computational social science today.

Applied computational linguistics today is divided into a number of subspecialties, "on the analysis side it deals with phonetics and phonology (sounds and phonological structure of words), morphological analysis (discovering the structure of words and their functions), tagging (disambiguation of part-of-speech and/or morphological function in sentential context), and parsing (discovering the structure of sentences; parsing can be purely structure-oriented or deep, trying to discover the linguistic meaning of the sentence in question). Word-sense disambiguation tries to solve polysemy in sentential context, and it is closely related to lexicon creation…Language modeling (probabilistic formulation of language correctness) is used primarily in systems based on stochastic methods" (Hajic, 2004, p. 82).

All of these technologies provide powerful tools for social scientists seeking to analyze meanings in texts. One of the technologies that is already finding fast adoption among social scientists are topic models. The most important of this class of stochastic models is latent Dirichlet allocation (LDA) introduced by Blei and colleagues (Blei et al., 2003). This is a method that provides an automated procedure for coding the content of a corpus of texts (including very large corpora) into a set of substantively meaningful coding categories called 'topics.' The algorithms work with little or no human intervention. Instead of beginning with predefined codes, the researcher begins by specifying the number of topics for the algorithm to find. The program then identifies that number of topics and returns the probabilities of words being used in a topic as well as an accounting of the distribution of those topics across the corpus of texts. While not infallible, when used thoughtfully and applied carefully, the method seems to consistently yield very plausible readings of the texts, demonstrating what DiMaggio et al. (2013) describe as high levels of 'substantive interpretability.' Natural Language Processing tools are another class of text analysis tools that should be of interest (for an example,

see Mohr et al., 2013). Bail (2014) provides a useful overview of the range of tools from computational linguistics that are of use to social scientists who want to study culture.

Information Sciences

The digitalization of numerous types of texts and archival records has led to the increasing demand for quick and efficient information search and retrieval – and, thus, revitalized the interdisciplinary field of information science. With respect to the formal analysis of culture, information science scholars most closely resemble their colleagues in computer science – that is, the research program is one focused upon developing automated tools for quickly (and efficiently) processing and summarizing the content of a large amount of textual data. The purpose is not to understand culture within a broader theoretical framework of human behavior, but to create rapid-fire queries that extend scaling and clustering techniques to more readily handle very large corpora. Salient examples include the work of Katy Börner and colleagues who have developed a number of appealing visualization technologies applied to texts (e.g., Börner et al., 2005). Social scientists have picked up on some of these tools, and developed their own, in ways that are more theoretically informed. Kathleen Carley and colleagues, for example, have worked closely across technical and social science disciplines in producing a suite of tools for optimizing organizational communication and performance. This work is grounded in network and organizational theories, and can therefore be considered a line of work that is consistent with paradigms at the center of operations research or management science – that is, scholars may create more optimal and rational complex organizations. This is done in part by seeing an organization as a type of machine or system in which an integrated circuitry of contact and communication is essential for proper goal attainment. To the extent that culture and *meanings* (i.e., interpretations) are essential to the day-to-day workings of and long-term learning within a complex organization, information science scholars play a vital role by developing tools for quickly retrieving and summarizing such qualitative data through a number of reduction techniques. Of course, once created, these tools are available for adapting to projects seeking to build upon theories of culture as it pertains to understanding and predicting human behavior.

Computational Humanities

In a sense, the humanities were a natural place to bring computing together with the study of culture. Early work by Roberto Bursa, for example, who, beginning in 1949, began creating the first computer-based concordance of the writings of Thomas Aquinas was carrying out a style of work, which was in many ways a logical extension of the very sort of systematic archival analysis and inking that biblical scholars had been doing for centuries. Similarly, there was a natural adaptation of classics studies to the computer analysis of texts, precisely because scholars in the field were already used to thinking in terms of carefully classified and categorized bits of information (Crane, 2004). However, the digital humanities movement has really accelerated over the last decade or so as massive quantities of digital texts began to become available.

An article published in *Science* by Michel et al. (2011) captured something of the excitement about this new state of affairs. The essay, *Quantitative analysis of culture using millions of digitized books* drew upon the availability of digitized texts in the Google Books project. In this case, over 5 million books (in multiple languages) were included in the database, a number that according to the authors, represents about 4% of all the books ever published. The corpus contains over 500 billion words and the authors use this material as a means for marking off what they call the new science of Culturomics, defined as "the application of high-throughput data collection and analysis to the study of human culture" (p. 181). Although the corpus itself is extremely large, the actual analysis of these 5 million texts was fairly simple, the authors focused on identifying the most frequent n-grams (clusters of words) and then tracking their distributions across time.

There are other contributors to this field who are themselves accomplished humanist scholars who have begun using the availability of digitalized texts as an opportunity to continue a more traditional humanistic style of research, but to do so through a decidedly different observational lens. Franco Moretti, for example, Director of the Stanford Literary Lab, has recently published a new book entitled *Distant Reading*, an obvious nod to the traditional style of literary methodology known as 'close reading.' In this book (and in the series of downloadable SLL pamphlets), Moretti recounts a number of the ways in which he and his colleagues have effectively transformed the kinds and styles of research questions that they pursue as a response to the new opportunities of pursuing 'distant readings.' This includes, to cite just one example, the development of automated procedures that can be used to transform a corpus of thousands of dramas and plays into network graphs (in which speakers are connected in these graphs to other speakers with whom they share direct dialogue). Describing his analysis of Shakespeare's play *Hamlet* using these techniques, Moretti writes, "You see the possibility here: different uses of language emerging in different network regions. Style, integrated with plot as a *function* of plot. It would be a breakthrough, and not just for literary analysis – but for the analysis of culture more broadly" (p. 229).

Computational Sociology

The rise of digital information also is relevant for a variety of emerging new projects among sociologists interested in measuring culture. Bail (2014) tells us, "More data were accumulated in 2002 than all previous years of human history combined. By 2011, the amount of data collected prior to 2002 was being collected every 2 days" (p. 465).

To take one example, Brian Uzzi and his colleagues (Saavedraa et al., 2011) collected information from commercial day traders, a task that they were able to do remarkably easily given the fact that all of the relevant information about every transaction, including all e-mail and text conversations with clients, was automatically recorded and stored as a part of the modern way of conducting business. By measuring the meaning of texts written (using automated sentiment analysis algorithms) and correlating those

measures with real-time performance measures, every trade made, the value of the trade obtained, and the possible value of that trade, Uzzi and his colleagues are able to directly connect the style and the tone of a trader's conversations with the rationality (as measured by money earned) of that trader's performance.

It is the ready availability of digital data that makes all the difference here. Social scientists have traditionally been severely hampered by the difficulties of designing, gathering, and analyzing social data. Now clever researchers have the ability to turn toward the simplicity of our new digital lives to their advantage. For another example, McFarland et al. (2013) take advantage of this in their study of speed dating encounters by using digital microphones and other types of real-time data gathering devices whose output can be formally measured and analyzed in order to provide a far richer record of (in this case) the many small cues, intonations, and emotional markers that characterize everyday interactions.

Bail (2014) provides an overview of the types of digital data sources that are becoming available for use as well as an assessment of the many kinds of algorithmic tools and measures that computational researchers have developed to facilitate the exploration and analysis of large-scale text corpora and the Internet network systems. He reports that there are a great many opportunities as well as a number of definite drawbacks to simply making use of computational tools that have been designed and crafted by scientists and engineers with very different ambitions and goals. Here, we are reminded again of the wisdom of Ernst Cassirer who warns us that any attempt to measure culture will be dependent upon our conceptual framework for theorizing culture. Cassirer writes, "Everywhere physical thought must determine for itself its own standards of measurement before it proceeds to observation …. It is not clocks and physical measuring rods, but principles and postulates that are the real instruments of measurement (Cassirer, 1923: p. 364–365).

See also: Classifications in Popular Music; Cultural Capital and Education; Cultural Production in Networks; Culture and Institutional Logics; Culture and Networks; Culture, Cognition and Embodiment; Culture, Production of; Embodiment and Culture; Narrative Networks; Networks and Cultural Consumption; Networks and Meaning.

Bibliography

Bail, Christopher A., July 2014. The cultural environment: measuring culture with big data. Theory and Society 43 (3–4), 465–482.
Bavelas, A., 1950. Communication Patterns in Task Oriented Groups. Journal of the Acoustical Society of America 22, 271–282.
Bearman, P.S., Stovel, K., 2000. Becoming a nazi: a model for narrative networks. Poetics 27, 69–90.
Blau, P., Duncan, O.D., 1967. The American Occupational Structure. Freeman Press, New York.
Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.
Börner, K., Chen, C., Boyack, K., 2005. Visualizing knowledge domains. Annual Review of Information Science and Technology 37, 179–255.
Bourdieu, P., 1990. Homo Academicus. Stanford University Press, Stanford, CA.
Bourdieu, P., 1984. Distinction. Harvard University Press, Cambridge, MA.
Bourdieu, P., Passeron, Jean-Claude, 1977. Reproduction in Education, Society, and Culture. Sage, Thousand Oaks, CA.
Bourdieu, P., Wacquant, L., 1992. An Invitation to Reflexive Sociology. University of Chicago Press, Chicago.
Breiger, R.L., 1974. The duality of persons and groups. Social Forces 53, 181–190.
Breiger, R.L., 2000. A tool kit for practice theory. Poetics 27, 91–115.
Breiger, R., 2004[2009]. "The analysis of social networks." Chapter 22. In: The Handbook of Data Analysis. Sage, Los Angeles, pp. 505–526.
Bulmer, M., 1996. The social survey movement and early Twentieth Century sociological methodology. In: Maurine, W. Greenwald, Anderson, Margo (Eds.), Pittsburgh Surveyed: Social Science and Social Reform in the Early Twentieth Century. University of Pittsburgh Press, Pittsburgh, pp. 15–34.
Carley, K., 1994. Extracting culture through textual analysis. Poetics 22, 291–312.
Cassirer, E., 1923. Substance and Function and Einstein's Theory of Creativity. The Open Court Publishing Co., Chicago.
Cerulo, K., 1988. Analyzing cultural products: a new method of measurement. Social Science Research 17, 317–352.
Converse, J.M., 2009(1987). Survey Research in the United States: Roots and Emergence 1890–1960. Transaction Publishers, New Brunswick, NJ.
Crane, G., 2004. Classics and the computer: an end of the history. In: Schreibman, Susan, Siemens, Ray, Unsworth, John (Eds.), A Companion to Digital Humanities. Blackwell Publishing, Oxwell, pp. 46–55.
D'Andrade, R., 1995. The Development of Cognitive Anthropology. Cambridge University Press, New York.
DeGraaf, Nan D., De Graaf, P.M., Kraaykamp, Gerbert, 2000. Parental cultural capital and educational attainment in the Netherlands: a refinement of the cultural capital perspective. Sociology of Education 73, 92–111.
DiMaggio, P.J., 1997. Culture and cognition. Annual Review of Sociology 23, 263–287.
DiMaggio, P.J., 1982. Cultural capital and school success: the impact of status-culture participation on the grades of U.S. high school students. American Sociological Review 47, 189–201.
DiMaggio, P.J., 2011. Cultural networks. Chapter 20. In: Scott, John, Carrington, Peter J. (Eds.), The Sage Handbook of Social Network Analysis. Sage, Los Angeles, pp. 286–300.
DiMaggio, P., Mohr, J., 1985. Cultural Capital, Educational Attainment, and Marital Selection. American Journal of Sociology 9, 1231–1261.
DiMaggio, P.J., Manish, N., David, B., December 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of U.S. government arts funding. Poetics 41 (6), 570–606.
Fine, G.A., 1979. Small groups and culture creation: the idioculture of little league baseball teams. American Sociological Review 44, 733–745.
Fligstein, N., 2001. Social skill and the theory of fields. Sociological Theory 19, 105–125.
Fligstein, N., McAdam, D., 2012. A Theory of Fields. Oxford University Press, New York.
Foucault, M., 1970. The Order of Things: An Archaeology of the Human Sciences. Pantheon, New York.
Fourcade, M., 2010. Economists and Societies: Discipline and Profession in the United States, Britain, and France, 1890s to 1990s. Princeton University Press, Princeton.
Franzosi, R., 1989. From words to numbers: a generalized and linguistic-based coding procedure for collecting textual data. In: Clifford, C. (Ed.), Sociological Methodology. American Sociological Association, Washington, DC, pp. 263–298.
Friedland, R.O., Alford, Robert, 1991. Bringing society back in: symbols, practices and institutional contradictions. In: W.Powell, W., DiMaggio, P.J. (Eds.), The New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago, pp. 232–263.
Friedland, R., Mohr, J.W., Roose, H., Gardinali, P., 2013. An institutional logic for love: measuring intimate life. Theory and Society 43 (3–4), 333–370.
Geertz, C., 1964. Ideology as a cultural system. In: Apter, D.E. (Ed.), Ideology and Discontent. Free Press, Glencoe, IL.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Gibson, D.R., 2003. Participation shifts: order and differentiation in group conversation. Social Forces 81, 1135–1181.
Hacking, I., 1975. The Emergence of Probability. Cambridge University Press, Cambridge, UK.
Hacking, I., 1990. The Taming of Chance. Cambridge University Press, Cambridge, UK.
Hajic, January 2004. Linguistics meets exact sciences. In: Schreibman, Susan, Siemens, Ray, Unsworth, John (Eds.), A Companion to Digital Humanities. Blackwell Publishing, Oxford, pp. 79–87.
Harary, F., Norman, R.Z., Cartwright, Dorwin, 1965. Structural Models: An Introduction to the Theory of Directed Graphs. Wiley, New York, NY.

Kane, D., 2000. Distinction worldwide: Bourdieu's theory of taste in international context. Poetics 31, 403–421.
Kroeber, A., 1919. On the principle of order in civilization as exemplified by changes in fashion. American Anthropologist 21, 235–263.
Laswell, H.D., Nathan, L., Associates, 1949. Language and Politics: Studies in Quantitative Semantics. George W. Stewart, Publisher, Inc., New York.
Lasswell, H.D., Lerner, D., Pool, I.d.S., 1952. The Comparative Study of Symbols: An Introduction. Stanford University Press, Palo Alto, CA.
Lazarsfeld, P.F., 1968. Foreword. In: The Logic of Survey Analysis, by Morris Rosenberg. Basic Books, NY, pp. vii–x.
Lazarsfeld, P.F., 1961. Notes on the history of quantification in Sociology—Trends, sources and problems. In: Woolf, Harry (Ed.), Quantification: A History of the Meaning of Measurement in the Natural and Social Sciences. Bobbs-Merrill, New York, pp. 147–203.
Lazarsfeld, P.F., 1968. Foreword. In: Rosenberg, Morris (Ed.), The Logic of Survey Analysis. Basic Books, New York, pp. vii–x.
Lazarsfeld, P.F., Barton, Alan H., 1982[1955]. Some functions of qualitative analysis in social research. In: Kendall, Patricia (Ed.), The Varied Sociology of Paul Lazarsfeld. Columbia University Press, New York, pp. 239–285.
Lazarsfeld, P.F., 1970. Sociology. In: Main Trends of Research in the Social and Human Sciences: Part One: Social Sciences. UNESCO, Paris, pp. 61–165.
Lévi-Strauss, C., 1963. Structural Anthropology. Basic Books, New York.
Lewin, K., 1951. Field Theory in Social Science. In: Dorwin, C. (Ed.), Selected Theoretical Papers. Harper and Row, New York.
Lieberson, S., 2000. A Matter of Taste: How Names, Fashions, and Culture Change. Yale University Press, New Haven, CT.
Lizardo, O., 2006. How cultural tastes shape personal networks. American Sociological Review 71, 778–807.
Leydesdorff, L., 2001. The Challenge of Scientometrics: The Development, Measurement, and Self-organization of Scientific Communications. DSWO Press, Leiden.
Łuczewski, M., 2009. Revisiting Boudon, Znaniecki, and Polish Peasants. In: Raymond, Boudon, Cherkaoui, Mohamed, Hamilton, Peter (Eds.), A Life in Sociology. Essays in Honour of Raymond Boudon. Bardwell Press, Oxford, pp. 1–21.
Markoff, J., 2004. The Abolition of Feudalism: Peasants, Lords and Legislators in the French Revolution. Penn State University Press.
Martin, J.L., 2003. What is field theory? American Journal of Sociology 109, 1–49.
Martin, J., 2000. What do animals do all day? Poetics 27, 195–231.
McFarland, Daniel, Jurafsky, Dan, Rawlings, Craig M., 2013. Making the connection: social bonding in courtship situations. American Journal of Sociology 118, 1596–1649.
McLean, P., 2007. The Art of the Network. Duke University Press, Durham, NC.
Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Team, T.G.B., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., Aiden, E.L., 2011. Quantitative analysis of culture using millions of digitized books. Science 331 (6014), 176–182.
Mische, A., Pattison, P., 2000. Composing a civic arena: publics, projects, and social settings. Poetics 27, 163–194.
Mohr, J.W., 1994. Soldiers, mothers, tramps and others: discourse roles in the 1907 New York city charity directory. Poetics: Journal of Empirical Research on Literature, the Media, and the Arts 22, 327–357.
Mohr, J.W., 2013. Bourdieu's relational method in theory and in practice: from Fields and capitals to networks and institutions (and back again). In: Dépelteau, François, Powell., Christopher (Eds.), Applying Relational Sociology. Palgrave, NY, pp. 101–135.
Mohr, J.W., Duquenne, V., April/June, 1997. The duality of culture and practice: poverty relief in New York city, 1888–1917. Theory and Society 26 (2–3), 305–356.
Mohr, J.W., Wagner-Pacifici, R., Breiger, R., Bogdanov, P., 2013. Graphing the grammar of motives in U.S. National Security Strategies: Cultural interpretation, automated text analysis and the drama of global politics. Poetics: Journal of Empirical Research on Literature, the Media, and the Arts 41 (6), 670–700.
Mohr, J.W., White, H.C., 2008. How to model an institution. Theory and Society 37, 485–512.
Pachucki, M.A., Breiger, R.L., 2010. Cultural holes: beyond relationality in social networks and culture. Annual Review of Sociology 36.
Patricia, H.T., William, O., Michael, L., 2012. The Institutional Logics Perspective. Oxford University Press, Oxford.
Padgett, J.F., Powell, W.W., 2012. The Emergence of Organizations and Markets. Princeton University Press, Princeton.
Peterson, R.A., Berger, D.G., 1975. Cycles in symbol production: the case of popular music. American Sociological Review 40, 158–173.
Peterson, R.A., Kern, R.M., 1996. Changing Highbrow Taste: From Snob to Omnivore. American Sociological Review 61, 900–907.
Pinch, T., 1986. Confronting Nature: The Sociology of Solar-neutrino Detection. Kluwer, Dordrecht, Holland.
Platt, J., 1996. A History of Sociological Research Methods in America: 1920 – 1960. Cambridge University press, Cambridge.
Porter, T.M., 1986. The Rise of Statistical Thinking 1820–1900. Princeton University Press, Princeton.
Powell, W.W., DiMaggio, P.J. (Eds.), 1991. The New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago.
Rawlings, C.M., Bourgeois, M.D., 2004. The complexity of institutional niches: credentials and organizational differentiation in a field of U.S. higher education. Poetics 32, 411–437.
Rawlings, C.M., McFarland, D., 2011. Influence flows in the academy: using affiliation networks to assess peer effects among researchers. Social Science Research 40, 1001–1017.
Richmond, M., 1917. Social Diagnosis. Russell Sage Foundation, New York.
Ross, D., 1991. The Origins of American Social Science. Cambridge University Press, New York.
Roscigno, V.J., Ainsworth-Darnell, J.W., 1999. Race, cultural capital, and educational resources: persistent inequalities and achievement returns. Sociology of Education 72, 158–178.
Ruef, M., 1999. Social ontology and the dynamics of organizational forms: creating market actors in the healthcare field, 1966–1994. Social Forces 77, 1403–1432.
Saavedraa, S., Hagerty, K., Uzzi, B., 2011. Synchronicity, instant messaging, and performance among financial traders. Proceeding of the National Academy of Sciences (PNAS). http://dx.doi.org/10.1073/pnas.1018462108 v1-201018462.
Schuman, H., 2008. Methods and Meaning in Polls and Surveys. Harvard University Press, Cambridge, MA.
Shannon, C.E., Weaver, W., 1949. The Mathematical Theory of Information. The University of Illinois Press, Urbana.
Smith, T., 2007. Narrative boundaries and the dynamics of ethnic conflict and conciliation. Poetics 35, 22–46.
Spengler, J.J., 1961. Theory, ideology, non-economic values, and politico-economic development. In: Braibanti, Ralph, Spengler, J.J. (Eds.), Tradition, Values, and Socioeconomic Development. Duke University Press, Durham, NC.
Stigler, S.M., 1986. The History of Statistics: The Measurement of Uncertainty before 1900. Harvard University Press, Cambridge.
Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A., Williams Jr., R.M., 1949. Studies in Social Psychology in World War II: The American Soldier. In: Adjustment during Army Life, vol. 1. Princeton University Press, Princeton, NJ.
Swidler, Ann, 2003. Talk of Love. University of Chicago Press, Chicago.
Thomas, W.I., 1918. The Polish Peasant in Europe and America. In: Znaniecki, F. (Ed.). University of Chicago Press, Chicago.
Thomas, W.I., Florian, Z., 1918. The Polish Peasant in Europe and America, Vol. 1. The Gorham Press, Boston.
Thornton, P.H., William, O., Michael, L., 2012. The Institutional Logics Perspective. Oxford University Press, Oxford.
Turner, C.F., Martin, E., 1984. Introduction. In: Turner, Charles F., Martin, Elizabeth (Eds.), Surveying Subjective Phenomena, vol. 1. Russell Sage Foundation, New York, pp. 3–21.
Van Rees, Kees, Vermunt, J.K., Verboord, M., 1999. Cultural classifications under discussion: latent class analysis of highbrow and lowbrow reading. Poetics 26, 349–365.
Vaisey, S., 2009. Motivation and justification: a dual-process model of culture in action. American Journal of Sociology 114, 1675–1715.
Vaisey, S., Lizardo, O., 2010. Can cultural worldviews influence network composition? Social Forces 88, 1595–1618.
Watts, D., 2011. Everything Is Obvious: How Common Sense Fails Us. Crown, New York.
White, H.C., 1963. An Anatomy of Kinship: Mathematical Models for Structures of Cumulated Roles. Prentice Hall, Englewood Cliff, NJ.
White, H.C., 1992. Identity and Control: A Structural Theory of Social Action. Princeton, NJ.
White, H.C., 1997. Can mathematics be social? Flexible representations for interaction process and its socio-cultural constructions. Sociological Forum 12 (1), 53–71.
White, H.C., 2008. Identity and Control: How Social Formations Emerge. Princeton, NJ.
Williams, R., 1977. Marxism and Literature. Oxford University Press, Oxford.
Wuthnow, R., 1989. Communities of Discourse: Ideology and Social Structure in the Reformation, the Enlightenment, and European Socialism. Harvard University Press, Cambridge, MA.
Yeung, King-To, 2005. What does love mean? Exploring network culture in two network settings. Social Forces 84 (1), 391–420.

Foucault, Michel (1926–84)

Arpad Szakolczai, University College, Cork, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Michel Foucault is one of the most influential, and also most controversial, social thinkers of the century. His work represents a sustained effort to reconstruct the way in which the specifically modern Western form of subjectivity has become what it is through the institutionalization of techniques of self and practices of truth-telling, and the will to truth and knowledge.

Foucault's Intellectual Trajectory

Foucault was born on 15 October 1926 in Poitiers, France. His upbringing was organized around knowledge since his early childhood (he started to attend school at the age of three), and major political events, especially wars (the Ethiopian War, the Spanish Civil War, and World War II), also had a major impact on his formation. In September 1945, he moved to Paris where he listened to the lectures of Jean Hyppolite during a preparatory high school year, and was then admitted to the prestigious Ecole Normale Supérieur. These were troubled years for Foucault, intellectually, personally, and politically alike. He had a short spell with the Communist Party and had difficulties in managing his homosexuality (Eribon, 1991).

After graduating in philosophy in 1951, he received further qualifications in 1952 and 1953 in psychopathology and experimental psychology. He then started teaching, working on his thesis, and also attending the St. Anne psychiatric hospital. However, under the effect of an intensive reading of Heidegger and then especially Nietzsche, his interests gained focus, and he decided to leave the standard academic career, even the country.

Between 1955 and 1960, Foucault worked as a cultural diplomat, first in Uppsala, Sweden (at the invitation of Georges Dumézil), then in Warsaw, Poland, and Hamburg, Germany. He thus experienced personally the different economic and political regimes of the period, noticing their latent similarities, and also worked on his doctoral thesis. This was defended in 1961, under the supervision of Georges Canguilhem, and published as *Folie et déraison* (Foucault, 1961) (*Madness and Unreason*, republished in 1972 with a new preface as *Histoire de la folie*, or *History of Madness*) with the support of Philippe Ariès.

After completing the book, Foucault continued his work on the reconstruction of the joint emergence of modern forms of scientific (especially medical) knowledge and its institutional context, and published in 1963 *The Birth of the Clinic* (Foucault, 1963). At the same time, he also became increasingly involved in literary criticism, publishing various essays and a book on Raymond Roussel. As a result, he came to 'bracket' institutional practices temporarily in his work and turned toward language and discourse – a shift he would later much regret. The result was *The Order of Things* (Foucault, 1966). Published in 1966, at the height of structuralism, the book immediately became a huge success, identifying Foucault as a main protagonist of the field.

Becoming increasingly uneasy with this label, Foucault soon embarked on a major self-reflexive exercise, attempting to clarify his project and its underlying methods for others and for himself. This eventually led to *The Archaeology of Knowledge* (Foucault, 1969) and a series of major methodological papers. At the same time, he moved from Paris to Tunis, where he lived during the events of May 1968. However, instead of a systematic and final clarification, the reflexive exercises rather brought about a major reorganization of Foucault's entire work. Appointed as Chair of the History of Systems of Thought at the prestigious Collège de France, a few months after his inaugural lecture, Foucault suddenly shifted his intellectual interest from knowledge to power, and at the same time he became engaged in political struggles as the founder of GIP (group for information on prisons).

His new interests led to the publication of *Discipline and Punish* (Foucault, 1975), easily his most widely read and influential work. However, around the time of its publication, Foucault was again already reflecting on his work. This yielded a book-length study published as the first of six planned volumes on the *History of Sexuality* and a series of significant and quite influential interviews and shorter pieces (see Foucault, 1980b). But perhaps most important, it led to another major reorientation of his project. When trying to specify the main characteristics of his approach to the study of power, he eventually came to realize that the question of power was not central to his work. This change – often interpreted as Foucault's 'crisis' – is visible in successive displacements of the time horizon of the work, first to the fifteenth and sixteenth centuries, then to early Christianity, and finally to Greco-Roman antiquity; and also in a reorganization at the level of central concepts: from power-knowledge to government by truth; from the genealogy of power to the hermeneutic of the subject; from technologies of power to techniques of self. It was strongly influenced by Foucault's encounter with the works of major scholars of antiquity, such as Pierre Hadot and Peter Brown. At the same time, he also increasingly recognized affinities with the works of social scientists like Norbert Elias and especially Max Weber. These encounters, together with a systematic reflection on the trajectory of his past work, helped Foucault identify the underlying problem of his work as being concerned neither with power nor with discourse, rather with the link between truth and subjectivity (Foucault, 1997: 281).

The new orientation is clear in the published volumes two and three of the *History of Sexuality* project (Foucault, 1984a, 1984b), which discuss the way practices related to pleasure were problematized in ancient Greece and Rome, respectively.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61035-8

Volume four, covering early Christianity, and written before the published volumes two or three, exists only in the form of an uncorrected manuscript, as Foucault was hospitalized, terminally ill, just days after he started to revise it for publication.

The best guides for the last, arguably most significant period of Foucault's work are the courses he gave at the Collège de France between 1978 and 1984, especially the last two courses on parrhesia or frankness of speech and truth-telling. Foucault died in Paris on 25 June 1984.

The Work and Its Context

Throughout his writings, covering over three decades, Foucault's work focused on the reasons and manner in which the individual subject, especially its 'true being,' has become the center of attention in the pursuit of knowledge and the search for social order, considered as an evident starting point and foundation. Thus proceeding, Foucault posed questions on what has been taken for granted by the main available approaches in philosophy and the social sciences, without dismissing them as mere ideology or illusion. This explains much of both the dynamics of Foucault's intellectual trajectory (discussed in this section) and its reception (covered in the next).

The trajectory has three major peculiarities. First of all, far from following a straightforward path of development, it was quite bumpy and curvy, progressing through crises (see Szakolczai, 1998). This was probably unavoidable, as Foucault's work did not simply represent an addition to the existing stock of knowledge nor a general attack on it; rather it was a problematization of the taken for granted. Furthermore, and for much the same reasons, the trajectory cannot be separated from the substance of the ideas. Given its detours, it would seem particularly desirable to get over such details and focus on the main results. However, this would undermine the effectivity of Foucault's undertaking, as doing so one would fail to problematize one's own hidden related preconceptions. Finally, Foucault was quite aware of the pitfalls of his project, and throughout his career kept reflecting on its directions and stakes. These reflections had a significant effect on the trajectory and contain a concise introduction to Foucault by Foucault, a starting point for any interpretative effort.

For these reasons, the exact mapping of Foucault's early steps is of crucial relevance, as this not simply defines the starting point of the investigation that would later be assumed as given, but also indicates the way in which Foucault plunged into the questioning and mapping of the aspect of the taken for granted he singled out for attention as problematic. The first anchor point to be questioned was academic philosophy that assumed, from Descartes through Kant to Husserl and Merleau-Ponty, that the 'subject of knowledge' or the 'transcendental ego' was the starting point of any philosophical undertaking. Foucault recognized that behind the epistemological assumptions of this approach there lay an entire anthropology, and was particularly interested in Kant's related work, which he translated and introduced as his minor thesis.

Given his problems with academic philosophy, he turned to psychology, or an empirical approach to the Kantian question: What is man? At the same time, he experimented with the two main alternative theoretical frameworks to academic philosophy, Marxism and psychoanalysis, finding them unsatisfactory. It is this combination that underlies the profound significance and originality of Foucault's work. His background incorporates both theoretical and empirical work: the philosophy of the subject in its various kinds and positivism on the one hand, and their main critiques, the works of Marx and Freud on the other, while being orthogonal to this entire epistemological field.

In his attempts to go beyond the available approaches, Foucault found inspiration in the works of Heidegger and Nietzsche. Instead of starting from the experiences of the subject, gained through a combination of sense perception and the application of universally valid categories of mind, Foucault rather posed the question of how this subject has been actually historically formed through experience.

In his last writings, Foucault would argue that there are three main axes of experience: knowledge, power, and subjectivity. This conceptualization, however, was only the fruit of late reflection on his own work. In his first major book, *History of Madness*, all three axes were dealt with, and their separation was not at all clear-cut. Furthermore, instead of directly reconstructing the subject of knowledge, he started at the opposite end, as if following a reflexive strategy, by reconstructing the opposite or the mirror, the subject of nonknowledge or the nonsubject, the mad, the carrier of 'unreason.' The aim was to show that mental illness was not simply an objective aspect of the human condition that was there, waiting to be discovered finally by the methods of true science, but a complex form of behavior that was itself shaped by these very methods and their institutional context.

In this book and the follow-up work, *The Birth of the Clinic*, Foucault tied closely together the reconstruction of the views concerning the aim and methods of collecting knowledge about individual human beings and its context of closed or total institutions, or the knowledge and the power axis. This changed with his third main book, *The Order of Things*, where the institutional dimension or actual scientific and social practice was left out of the picture. On the one hand, this enabled Foucault to sharpen his tools here and in the following, reflexive work, defining his archaeological method as "an attempt to introduce the diversity of systems and the play of discontinuities into the history of discourses" (Foucault, 1991). On the other hand, however, this also meant a certain one-sidedness, the suspension of the study of the link between displacements in thought and broader historical processes.

As a result, in the 1970s, Foucault reversed his approach and turned to the power axis, developing conceptual tools like power-knowledge (which never meant a simple identification between knowledge and power, only the assertion of the impossibility of a complete separation) or the microphysics of power. This meant a shift of disciplinary allegiances from philosophy and the life sciences to sociology and the social sciences, and a reorientation of focus from the subject of knowledge of phenomenology to the liberal individual as a subject of rights, the taken for granted starting point of modern political and social theory. Here as before, Foucault chose an indirect approach, focusing his reconstructive historical work on the formation of the subject deprived of rights: the criminal.

Discipline and Punish not only returned Foucault's interest to institutional practices but also sharpened his analytical tools

concerning the power axis. However, if his work in the 1960s was somewhat affected by flirtations with literary criticism and structuralism, his new work overcompensated by showing too much affinity with the ascendant leftist, neo-Marxist vogue, most visibly at its often excessive rhetoric. Both the achievements and the problems of the work can be shown through its central term and most striking metaphor, the Panopticon, as symbol and embodiment of modern power. It is a circular building structure, where inmates are closed into individualized cells, with no possibility of horizontal communication, while the cells are fully transparent toward the central tower, occupied by the guards. Inmates cannot see whether they are under supervision, while everything they do is fully visible to the guards, resulting a trap of visibility. The idea is to achieve the internalization of supervisory gaze, thus instituting an efficiently functioning, automatic mechanism of power. While Foucault was often criticized for painting an excessively negative vision of an inescapable, omnipresent, and omnipotent power, more applicable to totalitarian states than liberal democratic societies, the full prevalence and significance of panoptical mechanisms can only be realized by recognizing that the Panopticon has the same spatial arrangement as the circus. This helps one to understand the nature of the power exerted by modern propaganda machines and the media, for which Foucault's work opened a path: such technologies of power perpetually switch between asphyxiating control and relaxing circus modes of operation, both being based on the same fixating of the gaze on the center, where time and again all-powerful guard-politicians reveal themselves as clowns, while the jovial antics of the buffoons in the center of media attention imperceptibly transmogrify themselves into demonic machinations.

Foucault's next series of works on sexuality were intended to continue and complement the studies of power-knowledge as started in *Discipline and Punish*. An account of the continuities and discontinuities between the project as announced when its first, reflexive, and methodological volume of the series, *La volonté de savoir* (unfortunately omitted from the English title), is fundamental for understanding the last and arguably most important period of Foucault's work. At one level, the differences are marked. In the years around 1980, Foucault again stepped back from his previous approach and undertook another major reorientation. This meant that while previously, whether focusing on knowledge or institutional practice, Foucault studied how certain human beings shaped other human beings as objects of power and knowledge, now he set out to analyze directly how individual human beings were driven to reflect on their own being as subjects.

This was accompanied by a marked change in style, even a refocusing of his entire project. While Foucault's aim so far was to render visible the singularity of the idea that the individual subject should be the foundation of knowledge and social order, now he recognized the seriousness of the problem to which this came to be the answer. The shifting focus of interest on the individual was an answer to the manner in which being has become problematized at a certain point, the way in which human beings came to pose question marks to 'what they are, what they do, and the world in which they live' (Foucault, 1984a: 16). While this turn on the one hand rendered visible the Heideggerian inspiration behind his work, on the other it also brought to the surface the close affinities between Foucault and Max Weber.

Yet, continuities were also fundamental, and these concern the persistent driving force underlining Foucault's interest in the 'problematization' – a crucial concept discovered in this period, in order to capture the manner in which a vague sense of unease is transformed into the concrete formulation of a problem – of sexuality. The understanding of the exact nature of such continuities and discontinuities is now possible, given that publication of Foucault's Collège de France lectures is under way (for studies using the material that was previously only available in the Foucault Archives, see McGushin, 2007; Szakolczai, 1998, 2003).

Foucault's (1976) book was motivated by a distancing from an increasingly dominant and seemingly self-evident discourse in the social sciences, and in public life in general, which on the one hand laments the presumed pervasiveness of sexual repression characteristic of the rising modern world, explained by referring to Victorian morality and capitalism, and on the other proclaims sexual liberation as a program, in particular through the free realization of one's true sexual identity. In contrast, Foucault argues that the problem lies rather with this discourse itself, and the will that animates it. This will is a form of the will to knowledge, which is best visible in modern legal and medical discourses, traced provisionally to the Christian practice of the confessional. Foucault's central claim, however, is that there is a basically positive connection between these two discourses, instead of a radical difference; and that both discourses hide, even mask, the more fundamental issue, which concerns the incitement rather than repression of desires. What is uniquely specific to the modern world is thus not the repression of sexuality, rather a continuous playing with scandalizing and seducing, with attractions, evasions, and 'circular incitements,' that constitute *perpetual spirals of power and pleasure* (Foucault, 1980a: 49; emphasis in original). The inciting of desires results in the constitutive investment and formation of sexuality, thus fixes human beings by their own purposefully incited passions into certain prefabricated categories of sexual identity, thus giving them "an analytical, visible and permanent reality" (Foucault, 1980a: 44). The sexuality that supposedly captures the heart and truth of one's being is therefore nothing else than the correlative of these discursive practices (Foucault, 1980a); with the "irony of this deployment (being) in having us believe that our 'liberation' is in the balance" (Foucault, 1980a: 159; last sentence of the book).

Foucault never retracted his position; the introduction to the second volume of the *History of Sexuality* project, a central summary of the work completed thus far and of his methodological approach, restates the same position about the invested and constructed nature of sexuality. However, as it can be reconstructed through the last five lecture courses of the Collège de France, the single most important trace of the movements in Foucault's thinking over his last period, he moved beyond the earlier position not by discarding it as erroneous, rather in the sense of gaining a further, magnitude depth.

The 1979–80 lectures dealt with obligation assigned to individuals concerning the manifestation of truth and the 'regimes of truth' that emerged in this regard. The title of the course, 'The government of the living,' was irrelevant for its content, as that title was provided before Foucault reoriented

his work. Such reorientation was marked by a long discussion of *Oedipus King* by Sophocles, revisiting the lectures given during the 1970–71 course. Regimes of truth were defined as the types of relations that exist between manifestations of truth and the accepted procedures to ascertain truth, and also the subjects that are the operators, witnesses, and eventually objects of such procedures. The course focused on three such procedures, central for early Christianity: baptism, penitence, and the examination of conscience. The next year's course, entitled 'Subjectivity of Truth,' was the only course explicitly discussing sexuality, situated now on the broader horizon of the links between individual subjects and truth. The course of 1981–82, entitled 'The Hermeneutics of the Subject,' continued to study the links between subjectivity and truth, but did so from the perspective of a next central concept, the 'care of the self' (in fact, the title again lags behind this significant semantic shift from 'subject' to 'self'). The course started with a long discussion of Plato's *Alcibiades*, introducing the central concern of the course, the contrast between the care of the self and the Delphi injunction to know yourself, understood as the two central principles defining ancient philosophy. The last two courses both bear the general title 'The Government of Self and Others,' and both are centrally about parrhesia, or the courageous practice of telling the truth. The course of 1983 systematically introduced the term, starting with a long analysis of Euripides's play *Ion*, and then discussing the relevant passages in Plato's dialogues, including *The Republic*, *Laws*, *Gorgias*, the *Apology*, and *Phaedrus*, and especially a long and important analysis of Plato's *Seventh Letter*. The 1984 course, also carrying the title 'The Courage of Truth,' continues the study of parrhesia. After a crucial initial discussion of the last words of Socrates, using Dumézil's just-then published interpretation, and Plato's *Laches*, most of the course focuses on the Cynics. While occasionally recognizing certain kudos in their acts, Foucault's discussion presents the activities of the Cynics as an exaggerated and excessive, thus a bad form of, parrhesia. The course closes with a very cursory discussion of Christian parrhesia.

All of these courses were introduced by lectures in which Foucault was reflecting on the main stakes of his work, its underlying unity, and the central problem it was exploring, providing a crucial running commentary on Foucault by Foucault, which eventually culminated in the published introduction and draft preface to the *History of Sexuality* (see in *The Foucault Reader*), or the essays discussing Kant's reflections on Enlightenment and the French Revolution. This running commentary not only adds crucial further details to these conclusive pieces, but also is necessary for understanding a thought in its movements, in its vitality, helping avoid a dogmatic and often out of context canonization, relying on a few simplified ideas, which can then be routinized and force-fed.

The directions opened up in these works can be resumed through three interconnected concepts. The first is the importance attributed to the term care, in particular in expressions like care of the self and the care for truth, present in prominent lecture, article, and interview titles of the past period. While previously Foucault had a more or less exclusively negative, critical angle on the will to truth and the formation of the self through true discourse, now he perceived in addition a significant legitimacy of a care of the self, in terms of an ethic and esthetic of existence that helps give a shape to and control over one's being, in order to resist the invasive influences of certain aspects of the social and political environment, and started to trace the historical source of such practices, so central to European culture, back to antiquity. This focus on care was part of a return to the thinking of Heidegger, one of his most important formative influences. The second implied a rethinking of the connection between the self and truth, through his new interest in the practice of truth-telling, traced back to the Greek concern with parrhesia, characteristic in particular of Plato's presentation of Socrates, and his political philosophy. The discovery of this practice and its significance is connected to a systematic rereading of Nietzsche (see in particular Nietzsche's discussion of the truth-teller in *Gay Science*, no. 343–345). Parrhesia (a Greek word meaning telling everything) is closely connected to the care of the self and its ethical stance. It means a courageous act of truth-telling, involving serious risks and danger to one's identity and life, the revealing of a truth that is based on personal conviction, concerning both the content of the statement and its significance. Foucault also discussed the ambivalence of the practice, as telling everything is, strictly speaking, impossible, and thus can be used to promote various agendas and masked games. Beyond ethical questions, this practice also has fundamental epistemological implications, brought out in particular in Foucault's very cursory discussion of parrhesia in Christianity (this was the last section of the course, with little time left; for details, see Szakolczai, 2003: 229–240). Modern science restricts truth to matters that can be repeated and confirmed through experiments. However, the most important aspects of human and social life are unrepeatable events, where the only source of evidence is buried in the mind and heart of human beings. The care for truth implies a need to restore a concern with the nature of such personal evidence at the heart of philosophy, uprooted from there since Descartes and Kant. Foucault's recurrent discussions of the shortcomings of Descartes are among the most precious gems in these lectures.

The third major concept, technologies of the self, concerns the manner in which these discoveries enabled Foucault to return to his lifelong preoccupation with the way in which the very identities of human beings are being tampered with and fabricated in the modern world. The care of the self relies on techniques by which individuals could master their own behavior in order to find a way to live harmoniously, and with dignity, under conditions of external adversity. However, these very techniques could themselves be taken up, organized, and manipulated for opposite purposes as well, thus becoming part of a technological machinery of power, aiming at a mechanical and quasichemical transformation of human beings. Many of Foucault's late lectures and interviews were concerned with the attempt to clarify this extremely ambivalent and complicated relationship between the care of the self and for truth on the one hand, and the technologies of self and individualization, the will to knowledge or truth, and apparatuses and machineries of power on the other. This line of research, just as the previous two, remained fundamentally incomplete, cut short by Foucault's death in 1984. Yet, the very presence of this project makes Foucault one of the most important thinkers of the last century, and arguably the most faithful heir to both Nietzsche and Heidegger.

The Foucault Effect

Foucault has been one of the most influential social theorists of the past decades. His works are widely read in all branches of the humanities and the social sciences, including philosophy, sociology, history, anthropology, cultural studies, and literary criticism. However, they have always been controversial and most probably will remain so in the future.

Both characteristics can be traced to the same source. Foucault's work developed not simply as a dissent from the then-ruling paradigms in philosophy and psychology, phenomenology and positivism, but was similarly critical, after some initial flirtation, of the undertakings of Marx and Freud. Based on his reading of Kierkegaard, Heidegger, and especially Nietzsche, Foucault went into directions that did not easily fit in the contemporary intellectual landscape. This led to a persistent problem of understanding, reception, classification, and (mis)labeling. It was only deepened by some strategic concessions Foucault made when trying to establish bridges between his lonely undertaking and the work of his contemporaries.

Foucault's work first gained wide currency in the mid-1960s when, after the publication of *The Order of Things*, he was considered a prime exponent of structuralism, together with Barthes, Lacan, and Lévi-Strauss. His archaeological method, aiming at the establishment of historical discontinuities at the level of discourses and considering texts written between two consecutive historical breakpoints as belonging to a discursive formation, governed by certain autonomous rules, is still influential in wide circles within philosophy and literary criticism. The work largely followed the footsteps of Foucault's earlier works, relying more on Dumézil, Canguilhem, and Bachelard than on Lacan or Lévi-Strauss. It was also different from the structuralist project and mentality, as it made no attempt at developing formal schemes with claims of general validity. Still, the label structuralist stuck due to the gestures Foucault made in its last chapter to the then-rising structuralist tide.

Though Foucault's work of the 1970s on power-knowledge came to be inserted in a different debate and exerted its impact on a different field, the modalities were remarkably similar. As a conscious break with his earlier work, *Discipline and Punish* aimed at direct political relevance and impact. While *The Order of Things* was critical of Marxism and took no side in the political landscape, the new work was clearly aiming at making contact with the far left and started by using a language heavily indebted to Marxism. As the work progressed, Foucault steadily retrieved his central interest in the formation of the self through true discourse, in this case the shaping of the criminal. However, in its style and rhetoric, the ballast was still present, and the work came to be read and regarded as a prime example of the post-structuralist and postmodern alternative to old-fashioned Marxism. While in the 1960s Sartre considered Foucault's work as the last rampart of the bourgeoisie, his work of the 1970s served as a lifebuoy for many on the sinking boat of the radical left.

Soon Foucault became just as deeply disturbed by this reception of his work as with the structuralist label in the 1960s. His self-reflexive exercises of the late 1970s and the early 1980s not only helped him to regain focus in his own work but also to find, at the end, after some flirtation with the thought of the Frankfurt school, the most fruitful intellectual connections with the works of philosophers such as Gadamer and Hadot, and social scientists like Elias and Goffman. Most important of all, however, was his late in-depth encounter with the work of Max Weber, in whom he recognized a kindred spirit and a selected predecessor.

Though this was not lost in some segments of the Foucault reception (Dreyfus and Rabinow, 1982; Gordon, 1991; Szakolczai, 1998), by then the damage had already been done. Foucault thus came to be regarded as a prime exponent of the very intellectual currents whose presuppositions he wanted to question and map in his work. In this, his fate can again be best compared to Max Weber's. As a result, in the burgeoning literature for or against Foucault, the problem he started to pose almost half a century ago concerning the taken for granted, allegedly universal but quite peculiar and questionable character of the way modern Western subjects form their subjectivity (selfhood or identity) – a problem very close to the Weberian interest in inner-worldly asceticism and in the methodical conduct of life, or the interest of Elias in the formation of habitus through the civilizing process – became blurred. This is demonstrated most paradoxically in the way the new literature on sexual or gender identities takes Foucault as one of its sources, although his work should rather be seen as a most effective way of problematizing this 'politics of identity.'

A number of important recent publication projects are rendering Foucault's work more accessible, in the French original as well as in English. Based on the 1994 Gallimard publication of all of Foucault published shorter writings (Foucault, 1994), Allen Lane published a three-volume selection in 1997–2001, edited by Paul Rabinow, Nikolas Rose, and James Faubion. Furthermore, and most importantly, the publication of Foucault Collège de France lectures is now well under way, both in French, edited by François Ewald and Alessandro Fontana, and in English, edited by Arnold Davidson and carefully translated by Graham Burchell. Starting in 1997, so far 9 of the 13 lecture courses delivered between 1970 and 1984 have been published in French, and 8 of these already appeared in English as well. Once this project is completed, and its contents carefully analyzed, the full significance of Foucault's work, and the paths opened up by him, will become fully visible.

While Foucault, just like Weber, is one of the most widely read social theorists of the past century, his work is similarly still more of an unfulfilled promise than a closed, exhausted chapter in the history of thought.

See also: Comparative History; Control, Social; Crime, Sociology of; Elias, Norbert (1897–1990); Oversharing: The Eclipse of Privacy in the Internet Age; Postmodernism: Philosophical Aspects; Publishing: Academic and University Presses; Self: Philosophical Aspects; Weber, Max (1864–1920).

Bibliography

Dreyfus, H.L., Rabinow, P., 1982. Michel Foucault: Beyond Structuralism and Hermeneutics. University of Chicago Press, Chicago.

Eribon, D., 1991. Michel Foucault. Harvard University Press, Cambridge, MA.

Foucault, M., 1961. Folie et déraison: Histoire de la folie à l'age classique. Plon, Paris [2006 History of Madness. Routledge, London].
Foucault, M., 1963. Naissance de la clinique: Une archéologie du regard médical. Presses Universitaires de France, Paris [1975 The Birth of the Clinic. Vintage, New York].
Foucault, M., 1966. Les mots et les choses: Une archéologie des sciences humaines. Gallimard, Paris [1973 The Order of Things. Vintage, New York].
Foucault, M., 1969. L'archéologie du savoir. Gallimard, Paris [1976 The Archaeology of Knowledge. Harper, New York].
Foucault, M., 1975. Surveiller et punir: Naissance de la prison. Gallimard, Paris [1979 Discipline and Punish. Vintage, New York].
Foucault, M., 1976. La volonté de savoir: Histoire de la sexualité, vol. 1. Gallimard, Paris.
Foucault, M., 1980a. The History of Sexuality. In: An Introduction, vol. 1. Vintage, New York.
Foucault, M., 1980b. In: Gordon, C. (Ed.), Power/Knowledge: Selected Interviews and Other Writings by Michel Foucault, 1972–1977. Harvester Press, Brighton, UK.
Foucault, M., 1984a. L'usage des plaisirs: Histoire de la sexualité, vol. 2. Gallimard, Paris [1986 The Use of Pleasure. Vintage, New York].
Foucault, M., 1984b. Le souci de soi: Histoire de la sexualité, vol. 3. Gallimard, Paris [1987 The Care of the Self. Vintage, New York].
Foucault, M., 1984c. In: Rabinow, P. (Ed.), The Foucault Reader. Pantheon, New York.
Foucault, M., 1991. Politics and the study of discourse. In: Burchell, G., Gordon, C., Miller, P. (Eds.), The Foucault Effect: Studies in Governmentality. Harvester Wheatsheaf, London.
Foucault, M., 1994. In: Defert, D., Eward, F. (Eds.), Dits et écrits, 4 vols. Gallimard, Paris.
Foucault, M., 1997. The ethics of the concern of the self as a practice of freedom. In: Rabinow, P. (Ed.), Ethics: Subjectivity and Truth, vol. 1. The New Press, New York.
Gordon, C., 1991. Introduction. In: Burchell, G., Gordon, C., Miller, P. (Eds.), The Foucault Effect: Studies in Governmentality. Harvester Wheatsheaf, London.
McGushin, E., 2007. Foucault's Askésis: An Introduction to the Philosophical Life. Northwestern University Press, Evanston, IL.
Szakolczai, A., 1998. Max Weber and Michel Foucault: Parallel Life-Works. Routledge, London.
Szakolczai, A., 2003. The Genesis of Modernity. Routledge, London.

Fourier Analysis

Rainer Picard, Technische Universität Dresden, Dresden, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The core concepts of Fourier analysis and a variety of applications are briefly discussed (for simplicity mainly in the one-dimensional case). The connections to analyzing translation (or shift) invariant linear input–output systems, such as those described by differential-algebraic equations, ordinary differential equations, or partial differential equations and systems, are emphasized. The connection to various applications such as holography, image processing, visual perception, edge detection, motion analysis, and sampling is indicated.

What Is Fourier Analysis?

The term Fourier analysis commemorates the role that the French baron Jean-Baptiste-Joseph Fourier (21 March 1768 to 16 May 1830) played in utilizing methods involving a particular so-called integral transform, which now is known as the Fourier transform, to solve partial differential equations. Fourier's influential work has its precursors in early studies of the vibrating string by the Swiss mathematicians Daniel Bernoulli (8 February 1700 to 8 March 1782) and Leonhard Euler (15 April 1707 to 18 September 1783). We do not aim to give an introduction to Fourier analysis for the uninitiated, nor can we provide hands-on instruction on how to apply the concepts of Fourier analysis. There are fortunately numerous text books on various levels, which cover such needs (for a more accessible text, see, e.g., Howell, 2001). Here, in this rather brief exposition, we shall instead attempt to reveal the essence of why Fourier analysis is so powerful and of widespread usefulness. Since this may be – even to the practitioners of Fourier analysis – somewhat hidden, we shall need to elaborate on the central ideas in a little more detail.

Fourier analysis is the mathematical analysis of translation (or shift) invariant linear input–output systems, such as differential-algebraic equations, ordinary differential equations, or partial differential equations and systems arising thereof (unbeknownst to most) in the analysis of many if not the majority of real-world phenomena. Translation is a mathematical operation simply shifting the argument by a prescribed amount. For a given function f on the real line $\mathbb{R}$ the translation operation τ_h – translation of the argument by an amount $h\in\mathbb{R}$ – results in a new function $\tau_h f$, which is given by

$$(\tau_h f)(x) := f(x+h) \qquad [1]$$

for every $x\in\mathbb{R}$. (Here, '…∈…' should be read as '… element of… .' The colon marks the side of an equation that is defined by the other side; thus, the symbol ':=' could be read as 'defined as.') Shift invariance for an input–output system S means that applying a shift to the input results simply in the shifted output associated with the original input, that is, for every suitable function f on $\mathbb{R}$

$$S\tau_h f = \tau_h S f.$$

In other words, the input–output system S commutes with the shift τ_h. Such systems are well suited to be analyzed using the Fourier transform.

The Fourier Transformation

The Fourier transform can be described by its action on a suitable class of (in general) complex-valued functions f in terms of an integral (i denotes the imaginary unit):

$$\hat{f}(p) := \int_{-\infty}^{\infty} \exp(-2\pi \mathrm{i} p\gamma) f(\gamma)\mathrm{d}\gamma \quad \text{for every } p\in\mathbb{R}. \qquad [2]$$

A simple accessible class of functions would be piecewise continuous functions, i.e., functions that are continuous up to finitely many jumps, which in addition vanish outside of a bounded interval.

The mapping $\mathscr{F}$ suitably established via the association between the function f and the new function $\hat{f}$, i.e., $\hat{f} = \mathscr{F}f$, is called the Fourier transformation (see Goodman, 1968; Gaskill, 1978 as general references). The function $\hat{f}$ itself is referred to as the Fourier transform (or the spectrum) of f.

If f is real-valued it may often be convenient to consider the real part $\Re\hat{f}$ and the negative of the imaginary part $\Im\hat{f}$ of the Fourier transform $\hat{f}$ of f separately:

$$(\mathscr{F}_{\cos} f)(p) := \Re\hat{f}(p) = \int_{-\infty}^{\infty} \cos(2\pi p\gamma) f(\gamma)\mathrm{d}\gamma \text{ for every } p\in\mathbb{R}, \qquad [3]$$

$$(\mathscr{F}_{\sin} f)(p) := -\Im\hat{f}(p) = \int_{-\infty}^{\infty} \sin(2\pi p\gamma) f(\gamma)\mathrm{d}\gamma \text{ for every } p\in\mathbb{R}, \qquad [4]$$

which are known as the Fourier cosine transform and Fourier sine transform of f, respectively. By Euler's equation

$$\exp(2\pi \mathrm{i} x) = \cos(2\pi x) + \mathrm{i}\sin(2\pi x)$$

for every $x\in\mathbb{R}$, we see that

$$\mathscr{F} = \mathscr{F}_{\cos} - \mathrm{i}\mathscr{F}_{\sin}.$$

The operation of rescaling, which we define by

$$\sigma_L f := \sqrt{|L|} f(L\cdot), L\in\mathbb{R},$$

interacts with the Fourier transform in the following way:

$$\begin{aligned}\mathscr{F}(\sigma_L f) &= \sqrt{|L|}\mathscr{F}(f(L\cdot)) = \frac{1}{\sqrt{|L|}}(\mathscr{F}f)(\cdot/L)\\ &= \sigma_{1/L}(\mathscr{F}f) \text{ for } L\in\mathbb{R}, L\neq 0. \end{aligned} \qquad [5]$$

 http://dx.doi.org/10.1016/B978-0-08-097086-8.43042-4

Clearly a smaller scale is turned into a larger scale and vice versa.

Using rescaling with -1, which is of course simply reflection at the origin, we get $\mathscr{F}^* := \sigma_{-1}\mathscr{F}$, which turns out to be the inverse of the Fourier transformation, appropriately called inverse Fourier transformation, and so we are led to the so-called Fourier inversion formula

$$f = \int_{-\infty}^{\infty} \exp(2\pi \mathrm{i} p\cdot)\hat{f}(p)\mathrm{d}p =: \mathscr{F}^*\hat{f}. \qquad [6]$$

Defining the norm of a suitable function f by

$$||f|| := \sqrt{\int_{-\infty}^{\infty} |f(s)|^2 ds},$$

it can be shown that $\mathscr{F}$ satisfies

$$||\mathscr{F}f|| = ||f||, \qquad [7]$$

i.e. $\mathscr{F}$ is norm-preserving, which in turn allows the extension of the Fourier transformation (as well as equations [6], [7], [5]) by continuity to all elements of a space denoted by $L^2(\mathbb{R})$. If in particular $||f|| = 1$ we say f is normalized. Such a norm provides a metaphorical concept of "length" or "size" for functions analogous to the absolute value for numbers. In this sense equality [7] states that the Fourier transform of f has the same "length" or "size" as f itself. The elements of $L^2(\mathbb{R})$ may be understood as being represented by certain square integrable functions and can be characterized by having piecewise continuous functions vanishing outside a bounded interval as approximations of every desired precision. This extension step generalizes the Fourier transformation to such functions, but the original integral expressions [2] and [6] do not in general make sense any more since square integrable functions need not be integrable. As a convenient structural trade-off for this inconvenience one finds, however, the remarkable benefit that every element $f \in L^2(\mathbb{R})$ is the Fourier transform of an element in $L^2(\mathbb{R})$ (given by $\mathscr{F}^*f$). In particular,

$$\mathscr{F}^*\mathscr{F}f = \mathscr{F}\mathscr{F}^*f = f \qquad [8]$$

for all $f \in L^2(\mathbb{R})$, i.e., we can recover f from $\hat{f} = \mathscr{F}f$ by simply applying $\mathscr{F}^*$ to $\hat{f}$. The rescaling operation σ_L introduced above has been defined to ensure that it shares this property with the Fourier transformation: $\sigma_L : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ is also norm-preserving and has $\sigma_{1/L} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ as its (norm-preserving) inverse. Sometimes a rescaled version of $\mathscr{F}$ also goes by the name of Fourier transformation: $\sigma_{1/2\pi}\mathscr{F}$ with $\sigma_{1/2\pi}\mathscr{F}^*$ as its inverse. Indeed, we also have

$$\begin{aligned}\sigma_{1/2\pi}\mathscr{F}\left(\sigma_{1/2\pi}\mathscr{F}^*\right)f &= \mathscr{F}\mathscr{F}^*f = f = \mathscr{F}\mathscr{F}^*f \\ &= \sigma_{1/2\pi}\mathscr{F}^*\left(\sigma_{1/2\pi}\mathscr{F}\right)f\end{aligned}$$

for all $f \in L^2(\mathbb{R})$. In some application it is occasionally also found profitable to call $\sqrt{2\pi}\sigma_{1/2\pi}\mathscr{F}$ Fourier transformation with $\frac{1}{\sqrt{2\pi}}\sigma_{1/2\pi}\mathscr{F}^*$ as inverse. The reader should be aware of the possibility of these variants occurring. A variety of transformations derived by combination of various other norm-preserving transformations with the Fourier transformation are also of interest in this context. An example is the so-called Mellin transformation, which is scale invariant in contrast – as we have just seen – to the Fourier transformation itself. For image analysis (in polar coordinates) a radial Mellin transformation can be combined with a Fourier series expansion in the angular variable to produce the so-called Fourier-Mellin transformation, which shows advantages in analyzing scale and rotation motions of an image.

It follows from [7] that the inverse Fourier transformation $\mathscr{F}^*$ is also norm preserving. This norm, which makes $L^2(\mathbb{R})$ what is known a normed linear space, is induced by a so-called inner product given here by

$$\langle f|g\rangle := \int_{-\infty}^{\infty} \overline{f(x)}g(x)\mathrm{d}x$$

for $f, g \in L^2(\mathbb{R})$ (the bar denotes complex conjugation). Inner products allow the transfer of the geometric concept of orthogonality to other objects, here to $L^2(\mathbb{R})$: $f, g \in L^2(\mathbb{R})$ are called orthogonal if

$$\langle f|g\rangle = 0.$$

We note the following important relation between $\mathscr{F}$ and its inverse $\mathscr{F}^{-1} := \mathscr{F}^*$:

$$\langle \mathscr{F}f|g\rangle = \langle f|\mathscr{F}^*g\rangle \qquad [9]$$

for all $f, g \in L^2(\mathbb{R})$. Relation [8] shows that by transforming with $\mathscr{F}$ no information is lost. In particular, orthogonality and normalization are preserved. This information is, however, encoded in such a form that local information in $\hat{f}$ represents global information in the original f. This feature is utilized (for the two-dimensional case, see the brief discussion in section Higher dimensional Fourier transformation) in image processing where local information is removed from or added to the Fourier transform $\hat{f}$ of an image f (modeled as an illumination distribution function f) to actually improve the global visual appearance of the original f by, for example, removing grid structures or sharpening edges. The visual system itself appears to perform image processing, at least approximately, in this manner (see Cornsweet et al., 1984; Georgeson, 1979; Graham, 1981). Another application of the Fourier transformation can be found in holography (see Stroke, 1969). Holographic encoding has been utilized as a model for particular features in memory and perception (Pribram et al., 1974). The stability of the original data f despite 'noise' affecting the Fourier transformed signal $\hat{f}$ is also utilized in data transmission security.

In applications one has to deal with real quantities (the complex notation just being a matter of mathematical convenience). In this context, the so-called power spectrum $|\hat{f}(\cdot)|^2$ is of particular interest (the inverse Fourier transform of the power spectrum is known as the autocorrelation function). The power spectrum is usually the only part of the Fourier spectrum $\hat{f}$ that can actually be measured in experiments. The power spectrum (usually in its two-dimensional form, see later) is utilized to distinguish so-called Fourier motion from non-Fourier motion in the analysis of moving images (Chubb and Sperling, 1988; Fleet and Langley, 1995).

Note that if f is even, i.e., $f(-x) = f(x)$, then $\mathscr{F}f = \mathscr{F}_{\cos}f$ (the Fourier cosine transform), and if f is odd, i.e., $f(-x) = -f(x)$, then $\mathrm{i}\,\mathscr{F}f = \mathscr{F}_{\sin}f$ (the Fourier sine transform) all of which are then real quantities if f is real-valued. Even for the complex-valued case, the Fourier cosine transformation would recover the even part $E(f)(x) := \dfrac{f(x)+f(-x)}{2}$ of f from $\hat{f}$ and the Fourier sine transformation its odd part $O(f)(x) := \dfrac{f(x)-f(-x)}{2}$. Readers who are uncomfortable with complex numbers (the less mathematically inclined reader may also find in Weisstein (1980) an interesting and entertaining survey on Fourier transforms) may restrict their attention to odd and even real functions separately. The information contained in a real even or odd function is then fully preserved by its (real) Fourier cosine and its (real) Fourier sine transform, respectively.

To give a flavor of how the Fourier transformation acts, we introduce two interesting examples (for a more visual approach to Fourier transforms consult Harburn et al. (1975)). The Gauss distribution function γ – well-known from statistics – is given by $\gamma(x) := \exp(-\pi x^2)$ for $x \in \mathbb{R}$. It has the remarkable property that

$$\mathscr{F}_{\cos}\gamma = \mathscr{F}\gamma = \gamma. \qquad [10]$$

For the second example, we introduce the notation χ_M for the characteristic function of a subset $M \subseteq \mathbb{R}$ defined by

$$\chi_M(x) := \begin{cases} 1 & \text{for} \quad x \in M, \\ 0 & \text{for} \quad x \notin M, \end{cases}$$

for $x \in \mathbb{R}$. The characteristic function $\Pi := \chi_{[-1/2,1/2]}$ of the unit interval $[-1/2,1/2]$ (box function) has the so-called sinc function as Fourier transform:

$$\mathscr{F}_{\cos}\Pi = \mathscr{F}\Pi = \mathrm{sinc},$$

where $\mathrm{sinc}(x)$ is defined for all real x by

$$\mathrm{sinc}(x) := \begin{cases} \dfrac{\sin(\pi x)}{\pi x} & \text{for} \quad x \neq 0, \\ 1 & \text{for} \quad x = 0. \end{cases}$$

The sinc function is important in connection with the Shannon sampling theorem (see below).

The Fourier Transformation as a Spectral Representation

Derivatives

The significance of representing a function in terms of its Fourier transform becomes clear if we consider how that transform interacts with differentiation, symbolized by ∂:

$$\partial f = f'.$$

We find that

$$\partial = \mathscr{F}^{-1} 2\pi\mathrm{i}m\mathscr{F} \ \text{ in } \ L^2(\mathbb{R}), \qquad [11]$$

where m stands for the multiplication-by-argument operation, i.e.,

$$(mf)(p) := pf(p) \text{ for all } f \in L^2(\mathbb{R}),\ p \in \mathbb{R}.$$

It should be noted that defining the differential operator ∂ by the right-hand side of [11] implies a generalized concept of differentiability.

The elements in the domain of definition $\{f \in L^2(\mathbb{R}) | m\hat{f} \in L^2(\mathbb{R})\}$ of m as an operator in $L^2(\mathbb{R})$ are continuous functions but need not be differentiable in the classical sense. Equation [11] – called a spectral representation for $\dfrac{1}{2\pi\mathrm{i}}\partial$ – shows the equivalence of applying $\dfrac{1}{2\pi\mathrm{i}}\partial$ with multiplication-by-argument m or in other words of differentiation ∂ with multiplication by $2\pi\mathrm{i}m$. The obvious symmetry of m, i.e.,

$$\langle mf|g\rangle = \langle f|mg\rangle \text{ in } L^2(\mathbb{R}) \qquad [12]$$

implies, via the spectral representation of $\dfrac{1}{2\pi\mathrm{i}}\partial$, the skew symmetry of ∂, i.e.,

$$\langle \partial f|g\rangle + \langle f|\partial g\rangle = 0 \text{ in } L^2(\mathbb{R}), \qquad [13]$$

known as integration by parts.

Functions of ∂

The Fourier transform spectral representation in [11] generalizes to polynomials in ∂ and $2\pi\mathrm{i}m$. That is, for any polynomial P we have

$$P(\partial) = \mathscr{F}^{-1}P(2\pi\mathrm{i}m)\mathscr{F}.$$

This suggests that we define more general functions of ∂ by letting

$$F(\partial) := \mathscr{F}^{-1}F(2\pi\mathrm{i}m)\mathscr{F} \text{ in } L^2(\mathbb{R}), \qquad [14]$$

where F can be fairly general (and even matrix-valued, in which case the Fourier transformation is to be carried out component-wise). This concept is extremely powerful, since a large class of models discussed in so-called linear systems theory, (see Gaskill, 1978; Norman, 1981), can be understood as such functions $F(\partial)$ acting on an input u resulting in an output $F(\partial)u$. Characterizing properties of such systems are that their input–output relations are linear and translation-invariant, i.e., sums, multiples, and shifts of input result in sums, in (up to scaling) the same multiple and equally shifted output. Indeed, translation τ_h defined in [1] is itself a function of ∂ in the above sense:

$$\tau_h = \exp(h\partial) = \mathscr{F}^{-1}\exp(2\pi\mathrm{i}hm)\mathscr{F}.$$

It may be interesting to note that by substituting the series expansion for the exponential function, this formula (if applied to suitable functions) is easily recognized as nothing but the Taylor expansion. Translation-invariant, linear systems occur (usually in a higher dimensional form, compare section Higher Dimensional Fourier Transform) as good approximations of many aspects of visual and auditory perception.

Convolution

The action of a function $\hat{S}\left(\dfrac{1}{2\pi\mathrm{i}}\partial\right)$, where $\hat{S}$ denotes the Fourier transform of S, defines what is called a convolution (of S with g):

$$S * g := \hat{S}\left(\frac{1}{2\pi\mathrm{i}}\partial\right)g, \qquad [15]$$

which is defined for any g in the domain of $\hat{S}\left(\frac{1}{2\pi i}\partial\right)$. In many cases, $S * g$ is actually given as an integral expression, which is a so-called convolution integral:

$$(S * g)(x) = \int_{-\infty}^{\infty} S(x-y)g(y)\mathrm{d}y \text{ for } x \in \mathbb{R}. \qquad [16]$$

In accordance with [15] for such convolution integrals, the so-called convolution theorem holds

$$\mathscr{F}(S * g) = \hat{S}(m)\hat{g} = \hat{S}\hat{g}. \qquad [17]$$

Particularly, the above-mentioned relationship between the autocorrelation and the power spectrum reveals that the autocorrelation is a convolution.

Generalized Fourier Transformation and Some Applications

We extend the meaning of $\langle\cdot|\cdot\rangle$ by introducing even more generalized functions f (also called distributions) as linear functionals:

$$\langle f|\varphi\rangle := f(\varphi),$$

where φ is a smooth function, i.e., infinitely often differentiable, vanishing outside a bounded interval in $\mathbb{R}$ – we shall write $\varphi \in C_0^\infty(\mathbb{R})$ – such that f is linear, i.e.,

$$\langle f|\varphi + \alpha\psi\rangle = \langle f|\varphi\rangle + \alpha\langle f|\psi\rangle$$

for all complex numbers α and all $\varphi, \psi \in C_0^\infty(\mathbb{R})$. Clearly, for any $f \in L^2(\mathbb{R})$ the inner product $\langle f|\varphi\rangle$ defines a generalized function (i.e., a complex-valued linear function defined for $\varphi \in C_0^\infty(\mathbb{R})$), but $\exp(2\pi i p\cdot)$, although not an element of $L^2(\mathbb{R})$, also defines a generalized function via

$$\begin{aligned}\langle \exp(2\pi i p\cdot)|\varphi\rangle &:= \int_{\mathbb{R}} \overline{\exp(2\pi i p x)}\varphi(x)\,\mathrm{d}x \\ &= \hat{\varphi}(p)\end{aligned} \qquad [18]$$

for all $\varphi \in C_0^\infty(\mathbb{R})$. Further, a generalized function, which is not given by a classical function at all, is the distribution $\delta_{\{\omega\}}$ defined by

$$\delta_{\{\omega\}}(\varphi) = \langle \delta_{\{\omega\}}|\varphi\rangle := \varphi(\omega) \text{ for all } \varphi \in C_0^\infty(\mathbb{R}).$$

(This distribution samples evaluate the 'test function' φ at point $\omega \in \mathbb{R}$.) The special choice $\delta := \delta_{\{0\}}$ is the so-called Dirac-δ-distribution. Since it can only be distinguished from 0 by testing with $\varphi \in C_0^\infty(\mathbb{R})$ satisfying $\varphi(0) \neq 0$ we say $\delta = 0$ on $\mathbb{R} \setminus \{0\}$. However, δ is (by definition) not identically zero as a functional. It is particularly useful in many contexts to have $\delta_{\{\omega\}}(=\tau_{-\omega}\delta)$ as a mathematical model for a physical phenomenon that is vanishingly small in space or time such as a mass point, an elementary charge, a light point, a short impulse, or a molecule at location ω (see section Higher Dimensional Fourier Transform for the particularly interesting higher dimensional case).

With sufficient care, a large class of such distributions f (e.g., the so-called tempered distributions) can be established for which the concept of derivative, multiplication by the argument, and Fourier transform can (guided by [9] and [13]) be generalized by defining

$$\begin{aligned}\langle \partial f|\varphi\rangle &:= -\langle f|\partial\varphi\rangle, \\ \langle \mathscr{F}f|\varphi\rangle &:= \langle f|\mathscr{F}^*\varphi\rangle, \\ \langle \mathscr{F}^*f|\varphi\rangle &:= \langle f|\mathscr{F}\varphi\rangle,\end{aligned} \qquad [19]$$

for all $\varphi \in C_0^\infty(\mathbb{R})$. As an example for this generalized differentiation, we find for the characteristic function $\chi_{]0,\infty[}$ of the interval $]0, \infty[$ of all positive real numbers

$$\partial\chi_{]0,\infty[} = \delta.$$

The Fourier transform of the distribution $\delta_{\{\omega\}}$ is, according to [18] and [19],

$$\mathscr{F}\delta_{\{\omega\}} = \exp(-2\pi i\omega\cdot) \qquad [20]$$

and conversely we have

$$\mathscr{F}\exp(2\pi i\omega\cdot) = \delta_{\{\omega\}}. \qquad [21]$$

The latter fact can be used to detect oscillatory behavior of frequency ω, which may be hiding in seemingly random data. The oscillation would show up in the Fourier transformed data as a 'peak' or 'spike' at point ω.

The concept of functions of ∂ and in particular the concept of convolution can also be carried over to tempered distributions. This will be illustrated by a few applications.

Impulse Response

The importance of the Dirac-δ-distribution is that a translation-invariant, linear system $S(\partial)$ is completely described by its impulse response $S(\partial)\delta$. Indeed, the general response $S(\partial)f$ is given by convolution with the impulse response

$$S(\partial)F = (S(\partial)\delta) * F$$

In the case that $S(\partial)$ is the solution operator $S(\partial) = P(\partial)^{-1}$ of the differential equation $P(\partial)u = f$, P a polynomial, the impulse response is also known as fundamental solution (or Green's function).

Shannon's Sampling Theorem

Another distribution of particular interest is the (sampling or) comb distribution $Ш := \sum_{n=-\infty}^{+\infty} \delta_{\{n\}}$, which takes samples at every integer point. This distribution is reproduced by the Fourier transformation, i.e.,

$$\mathscr{F}Ш = Ш, \qquad [22]$$

which is another way of stating the Poisson summation formula:

$$\mathscr{F}\sum_{n=-\infty}^{+\infty} \delta_{\{n\}} = \sum_{n=-\infty}^{+\infty} \exp(-2\pi i n\cdot) = \sum_{n=-\infty}^{+\infty} \delta_{\{n\}}.$$

Note that for the first equality we have used [18] and [20].

If $f \in L^2(\mathbb{R})$ and $f = 0$ outside of $[-1/2, 1/2]$, then $Ш * f$ is simply the periodic extension of f from the unit interval $[-1/2, 1/2]$, to all of $\mathbb{R}$. Therefore, by 'cutting off' with the characteristic function $\Pi := \chi_{[-1/2,1/2]}$ we recover

$$f = \Pi(Ш * f). \qquad [23]$$

As we now demonstrate, the so-called Shannon sampling theorem follows from [22]. In fact, applying [23] to $\hat{f} \in L^2(\mathbb{R})$ with $\hat{f} = 0$ outside of $[-1/2, 1/2]$ we get

after the inverse Fourier transformation has been applied that

$$\begin{aligned} f &= F^*(\Pi(\text{Ш} * \hat{f})) \\ &= \text{sinc} * \left(\sum_{n=-\infty}^{+\infty} f(n)\delta_{\{n\}} \right), \\ &= \sum_{n=-\infty}^{+\infty} f(n)\text{sinc}(\cdot - n), \end{aligned} \qquad [24]$$

which is indeed the statement of Shannon's sampling theorem.

Any function f with $\hat{f}$ vanishing outside a bounded interval is called band-limited. Since any such function can be easily rescaled to have $\hat{f} = 0$ outside of $[-1/2,1/2]$, [24] shows that any band-limited function f can be completely recovered from equidistant sampling. The sampling theorem has many applications, in particular in the analysis of auditory and visual perception (see, e.g., Yellot, 1990).

By interchanging the role of f and $\hat{f}$ in the Shannon sampling theorem we have

$$\hat{f} = \text{sinc} * \left(\sum_{n=-\infty}^{+\infty} \hat{f}(n)\delta_{\{n\}} \right) \qquad [25]$$

and after applying the inverse Fourier transformation we obtain the so-called Fourier series

$$f = \sum_{n=-\infty}^{+\infty} \hat{f}(n)\Pi\exp(2\pi\text{i}n\cdot) \qquad [26]$$

of a function $f \in L^2(\mathbb{R})$ with $f = 0$ outside of $[-1/2,1/2]$. Note that $\{\Pi\exp(2\pi\text{i}n\cdot) | n \in \mathbb{Z}\}$ is an orthonormal set, i.e., all elements are normalized and pairwise orthogonal, which allows a Fourier series expansion in the sense of [26] for every $f \in L^2(\mathbb{R})$ with $f = 0$ outside of $[-1/2,1/2]$.

Wavelets

It is interesting to note that the set of functions $\{\tau_{-n}\text{sinc} | n \in \mathbb{Z}\}$ occurring in Shannon's sampling theorem is also an orthonormal set in $L^2(\mathbb{R})$. Indeed, this is a complete orthonormal set in $L^2(\mathbb{R})$ in the sense that every $f \in L^2(\mathbb{R})$ has a series expansion with respect to these functions. By virtue of this we get

$$g = \sum_{n=-\infty}^{+\infty} \langle \tau_{-n}\text{sinc} | g \rangle \tau_{-n}\text{sinc}$$

for every $g \in L^2(\mathbb{R})$. According to [24] we have that for functions with bandwidth less than or equal to 1, i.e., with $\hat{f} \in L^2(\mathbb{R})$ such that $\hat{f} = 0$ outside of $[-1/2,1/2]$, g is actually represented by this sum. In particular,

$$\langle \tau_{-n}\text{sinc} | g \rangle = g(n)$$

for $n \in \mathbb{Z}$. To cover all band-limited functions one would have to allow for sufficiently large scaling factors L. Letting, e.g., $L = 2^p, p \in \mathbb{Z}$, then $(\sigma_{2^p}(\tau_{-n}\text{sinc}))_{p,n\in\mathbb{Z}}$ can be used to approximate any element $f \in L^2(\mathbb{R})$. All elements of $(\sigma_{2^p}\tau_{-n}\text{sinc})_{p,n\in\mathbb{Z}}$ are generated by the combination $\sigma_{2^p}\tau_{-n}$ of scaling and translation from the single function sinc, the so-called scaling function of this particular construction. Unfortunately, $(\sigma_{2^p}\tau_{-n}\text{sinc})_{n\in\mathbb{Z}}$ is only orthonormal for fixed $p \in \mathbb{Z}$. By a suitable orthonormalization procedure (multiresolution analysis) one can obtain, however, that $(\sigma_{2^p}\tau_{-n}(\text{sinc} - \sqrt{2}\sigma_2(\text{sinc})))_{p,n\in\mathbb{Z}}$ is a complete orthonormal system in $L^2(\mathbb{R})$. Any orthonormal system generated by $\sigma_{2^p}\tau_{-n}, p, n \in \mathbb{Z}$, from a so-called mother wavelet – here the function $(1 - \sqrt{2}\sigma_2)\text{sinc}$ – is called an orthonormal wavelet basis. Wavelets derive their utility largely from the fact that they do not have a fixed base scale like, e.g., the Fourier system $(\exp(2\pi\text{i}n\cdot)\Pi)_{n\in\mathbb{Z}}$, which by allowing for translations would also yield an orthonormal system $(\tau_r(\exp(2\pi\text{i}n\cdot)\Pi))_{n,r\in\mathbb{Z}}$ for $L^2(\mathbb{R})$. Wavelets naturally allow for the 'right' scale in $L = 2^p, p \in \mathbb{Z}$, to emerge for any given f according to their weight in its orthonormal expansion. As the Fourier transform may be considered to be a continuously scaled Fourier series, there are also continuously scaled versions of wavelet expansions known as wavelet transforms.

Higher Dimensional Fourier Transformation

Much of what has been explored above can be generalized to higher dimensions. This is due to the particular product structure of the higher dimensional exponential

$$\begin{aligned} &\exp(2\pi\text{i}(x_1p_1 + x_2p_2 + \ldots + x_np_n)) \\ &\quad = \exp(2\pi\text{i}x_1p_1)\exp(2\pi\text{i}x_2p_2)\ldots\exp(2\pi\text{i}x_np_n) \end{aligned}$$

for $x_i, p_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$, i.e., $(x_1, x_2, \ldots, x_n), (p_1, p_2, \ldots, p_n) \in \mathbb{R}^n$. Thus, the n-dimensional Fourier transformation, $n \in \mathbb{N}, n \geq 2$, can be understood as a repeated one-dimensional Fourier transform, i.e.,

$$\begin{aligned} (\mathscr{F}f)(p) &= \int_{\mathbb{R}^n} \exp(-2\pi\text{i}xp)f(x)\text{d}x \\ &= \int_{-\infty}^{\infty} \exp(-2\pi\text{i}x_1p_1) \int_{-\infty}^{\infty} \exp(-2\pi\text{i}x_2p_2) \\ &\quad \ldots \int_{-\infty}^{\infty} \exp(-2\pi\text{i}x_np_n)f(x)\,\text{d}x_n \ldots \text{d}x_2\,\text{d}x_1, \end{aligned}$$

where $x = (x_1, x_2, \ldots, x_n)$, $p = (p_1, p_2, \ldots, p_n)$, and $xp = \sum_{i=1}^n x_ip_i$. As a consequence, many of the above considerations can be carried over to the higher dimensional case. For example, with $x^2 := xx$, $x \in \mathbb{R}^n$, the definition of the Gauss distribution function carries over to the n-dimensional case and the property of [10] remains true. The n-dimensional Dirac-δ-distribution $\delta = \delta_{\{0\}}$ simply takes a single sample at 0 and the shifted version $\delta_{\{x\}}$ correspondingly at $x \in \mathbb{R}^n$. The latter still has $\exp(-2\pi\text{i}x\cdot)$ as its Fourier transform. In addition, $\text{Ш} := \sum_{x\in\mathbb{Z}^n} \delta_{\{x\}}$, which takes samples at every point of $\mathbb{R}^n$ with integer coordinates, still satisfies

$$\mathscr{F}\text{Ш} = \text{Ш}. \qquad [27]$$

Thus, in particular, corresponding variants of the Shannon sampling theorem and of the Fourier series expansion hold. Note, however, that the scaling has to be generalized to $\sigma_L f := |L|^{n/2} f(L\cdot)$, but then [5] also carries over to the n-dimensional case. The higher dimensional Fourier transformation finds its applications in higher dimensional, translation-invariant, linear systems (Norman, 1981). A particularly prominent application is in optics (see, e.g., Gaskill, 1978; Goodman, 1968; Taylor, 1978). As pointed out earlier, the visual

system may be considered at least approximately to be a two-dimensional (or three-dimensional) example of such a system and may be analyzed with Fourier techniques (see, e.g., Cornsweet et al., 1984; Georgeson, 1979; Graham, 1981; Julesz, 1980; Yellot, 1990). Higher dimensional Fourier methods also play a central role in the discussion and analysis of problems associated with the perception of moving images. For a survey of recent developments see Burr and Thompson (2011).

See also: Signal Detection Theory; Vision, High-Level Theory of; Vision, Low-Level Theory of.

Bibliography

Burr, D., Thompson, P., 2011. Motion psychophysics: 1985–2010. Vision Research 51 (13), 1431–1456.

Chubb, C., Sperling, G., 1988. Drift-balanced random stimuli – a general basis for studying non-Fourier motion perception. Journal of the Optical Society of America A: Optics, Image Science and Vision 5 (11), 1986–2007.

Cornsweet, T.N., Wandell, B., Yellot, J.I., 1984. The beginnings of visual perception: the retinal image and its initial encoding. In: Darian-Smith, I. (Ed.), Handbook of Physiology, Sec. 1. The Nervous System. Sensory Processes, vol. 3. American Physiological Society, Bethesda, pp. 257–316.

Derrode, St, Ghorbel, F., 2001. Robust and efficient Fourier-Mellin transform approximations for gray-level image reconstruction and complete invariant description. Computer Vision and Image Understanding 83, 57–78.

Fleet, D.J., Langley, K., 1995. Computational analysis of non-Fourier motion. Vision Research 34, 3057–3079.

Gaskill, J.D., 1978. Linear Systems, Fourier Transforms and Optics. John Wiley & Sons, New York, Chichester, Brisbane, Toronto.

Georgeson, M., 1979. Spatial Fourier analysis and human vision. In: Sutherland, N.S. (Ed.), Tutorial Essays in Psychology, vol. 2. Erlbaum, Hillsdale, New Jersey, pp. 39–88.

Goodman, J.W., 1968. Introduction to Fourier Optics. McGraw-Hill, New York.

Graham, N., 1981. The visual system does a crude Fourier analysis of patterns. SIAM-AMS Proceedings 13.

Harburn, G., Taylor, C.A., Welberry, T.R., 1975. Atlas of Optical Transforms. Bell & Sons, London.

Ho, H.T., Goecke, R., 2008. Optical flow estimation using Fourier-Mellin transform. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008 Anchorage (AL), USA, pp. 1–8.

Howell, K.B., 2001. Principles of Fourier Analysis. Studies in Advanced Mathematics. Chapman & Hall/CRC, USA.

Julesz, B., 1980. Spatial-frequency channels in one-, two-, and three-dimensional visions: variations on an auditory theme by Bekesy. In: Harris, C.S. (Ed.), Visual Coding and Adaptability. Erlbaum, Hillsdale, New Jersey.

Norman, F.M., 1981. Lectures on linear systems theory. Journal of Mathematical Psychology 23, 1–89.

Pribram, K., Nuwer, M., Baron, R., 1974. The holographic hypothesis of memory structure in brain function and perception. In: Krantz, D.H., Atkinson, R.C., Luce, R.D., Suppes, P. (Eds.), Contemporary Developments in Mathematical Psychology, vol. 2. Freeman, San Francisco, pp. 416–457.

Stroke, G.W., 1969. An Introduction to Coherent Optics and Holography. Academic Press, New York, London.

Taylor, C.A., 1978. Images. A Unified View of Diffraction and Image Formation with All Kinds of Radiation. Wykeham Publ., London, Basingstoke.

Weisstein, N., 1980. The joy of Fourier analysis. In: Harris, C.S. (Ed.), Visual Coding and Adaptability. Erlbaum, Hillsdale, New Jersey.

Yellot, J.I., 1990. The photoreceptor mosaic as an image sampling device. In: Advances in Perception. Proceedings of a Symposium on Frontiers of Visual Science. National Research Council. National Academy Press, Washington, pp. 117–134.

Frankfurt School: Institute for Social Research

Dustin B Garlitz, University of South Florida, Tampa, FL, USA
Hans-Herbert Kögler, University of North Florida, Jacksonville, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Institute for Social Research, or Frankfurt School, is an interdisciplinary research center associated with the University of Frankfurt in Germany and responsible for the founding and various trajectories of Critical Theory in the contemporary humanities and social sciences. Three generations of critical theorists have emerged from the Institute. The first generation was most prominently represented in the twentieth century by Max Horkheimer, Herbert Marcuse, Theodor W. Adorno, Walter Benjamin, Leo Löwenthal, and also for some time Erich Fromm. The so-called 'second generation' of the Institute is centrally represented by Jürgen Habermas, whose work has functioned as the focal point of a wide range of critical theorists. The third generation of the Frankfurt School is represented by Axel Honneth who emerged as a new center, with different strands or readings of who else belongs to the third generation, some in Germany, some internationally, and some more in sociology and social and political theory than philosophy.

Introduction

The Institute for Social Research (*Institut für Sozialforschung*), an academic institution of the University of Frankfurt in Germany opened in 1924 and commonly referred to as the Frankfurt School, comprises a group of scholars who promote a critical theory of society, or simply critical theory, and approach its study from interdisciplinary and multidisciplinary perspectives. The critical theorists who made up the first generation of the Frankfurt School include social philosopher Max Horkheimer (1895–1973), philosopher and political theorist Herbert Marcuse (1898–1979), social psychologist and psychoanalyst Erich Fromm (1900–80), philosopher, sociologist, and musicologist Theodor W. Adorno (1903–69), social scientist and philosopher Friedrich Pollock (1894–1970), sociologist Leo Löwenthal (1900–93), as well as political and legal theorist Otto Kirchheimer (1905–65) and political theorist Franz Neumann (1900–54). These scholars shared a common Jewish ancestry and a Marxist political and scholarly orientation. Cultural theorist and literary critic Walter Benjamin (1892–1940) was a loose affiliate of the Institute in its first generation, as were cultural critic and sociologist Siegfried Kracauer (1889–1996) and philosopher Ernst Bloch (1885–1977). Other lesser known full-members of the Institute in its first generation included historian Karl August Wittfogel (1896–1988) and economist and historian Henryk Grossman (1881–1950). The so-called second generation of the Frankfurt School is most prominently represented by philosopher and sociologist Jürgen Habermas (1929–present), whose work, however, has been flanked and aided by Albrecht Wellmer (1933–present) and Claus Offe (1940–present). The third generation is represented by the social philosopher Axel Honneth (1949–present) at the center, the latter of whom is currently director of the Institute. There exist different strands or readings of who else belongs to the third generation, some in Germany, some internationally, some more in sociology and social theory. Marking the internationalizing of Critical Theory, social philosophers Nancy Fraser (1947–present) and Seyla Benhabib (1950–present) are in some readings considered representatives of the third generation of the Frankfurt School. Other readings regard the work of political theorist Claus Offe and philosopher and social theorist Oskar Negt (1934–present) as representative of the third generation, while others also consider the work of social theorist Hans Joas (1948–present) as representative of this generation.

The Institute was founded on an endowment made possible by Felix Weil (1898–1975), an Argentinean-born German Marxist scholar and son of a millionaire businessman. Philosopher Carl Grünberg (1861–1940) was the Institute's first director. In October 1930, Max Horkheimer succeeded Grünberg as the Institute's director. In 1934, the Institute relocated to the United States, finding a home at Columbia University in New York City, where it stayed until 1950, at which time it was then moved back to the University of Frankfurt in Germany.

History of the Frankfurt School I: From 1923 to 1969

Historical Background and Institutional Beginnings

The Institute for Social Research opened in 1924, founded on money given from Felix Weil, an Argentinean-born German Marxist scholar and son of a millionaire businessman. Sociologist Kurt Albert Gerlach (1886–1922), a left-wing socialist professor, was an important figure throughout the creation process, despite dying before the Institute's founding. Weil intended to found an Institute that served first and foremost for the study and broadening of scientific Marxism. Weil, with Gerlach's assistance, created an institute connected to the University of Frankfurt but independent from it in that is was directly answerable to the Prussian Ministry of Culture. Weil and his father financed the building and equipment of the Institute, paid it a yearly grant of 120 000 marks, and offered the lower floor to the university's Faculty of Economics and Social Science, in addition to later on funding the professorial chair which the Institute's director held in that faculty. Authorization from the ministry for the foundation of an Institute for Social Research as an academic institution of the University of Frankfurt which will also serve the University for

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61221-7

teaching purposes' came through at the beginning of 1923. Construction began in March of that year. Marxist philosopher Carl Grünberg became the Institute's first director in 1924. In Grünberg, Weil had found a director of the new Institute who was both a full Marxist and a recognized scholar. The Faculty of Economics and Social Sciences in Frankfurt agreed to Grünberg as director at once and voted unanimously at the beginning of January 1923 to nominate him to the ministry for appointment to the professorial chair funded by the Institute for Social Research. On Sunday, 22 June 1924, at 11 a.m., the academic ceremony celebrating the opening of the Institute for Social Research was held in the hall of the University of Frankfurt.

Grünberg's first two assistants were social scientist and philosopher Friedrich Pollock and economist and historian Henryk Grossman. Pollock had taken his doctorate in Economics at the University of Frankfurt in 1923, and had temporarily been director of the Institute until Grünberg arrived. Grünberg offered him a position as a research associate at the Institute, and he accepted it immediately. Grossman joined the Institute in 1926 at Grünberg's invitation, also as a research associate. Another of the Institute's very first assistants was Rose Wittfogel, who was a librarian. In 1925, her husband, historian Karl August Wittfogel, became a full-time research associate of the Institute. Sociologist Leo Löwenthal started receiving a scholarship from the Institute in 1926.

The Turn toward a 'Critical Theory of Society'

The decisive turn for the Institute for Social research came when Max Horkheimer assumed the position of its second director. In January 1928, Grünberg had to stop work and resign as director of the Institute due to a stroke. There was an interim period when Pollock served as acting director. In October 1930, Pollock (who had been Weil's executive agent since 1925), representing the board of the Institute of Social Research, and social philosopher Max Horkheimer, who had been appointed 2 months before by the Faculty of Philosophy at the University of Frankfurt to Professor of Social Philosophy, signed a contract making Horkheimer the Institute's new director. On 24 January 1931, Horkheimer gave his inaugural lecture on taking up the Chair of Social Philosophy and the directorship of the Institute for Social Research. Horkheimer launched a new research program which intended to overcome the crisis of Marxism by fusing social philosophy and empirical social science. Horkheimer proved a highly integrative and formidable research organizer, who understood to be highly influential less by fully defined ideas and theories and more by suggesting the broad outlines within which inter- and multidisciplinary work of original scholars could unfold. The perspective which he developed in the early writings of the Institute, including his inaugural address, was enabling rather than constraining, gave direction without determining every detail.

In a nutshell, the first and founding generation of the Frankfurt School attempted to actualize Marx's critique of capitalism. This project not only involved the need to account for capitalism's continued existence, but also for the rise of fascist as well as state-capitalist systems of totalitarian domination. Marx's critique thus needed to be expanded to explain the susceptibility of the masses, including the working class, to nationalist propaganda, the fascist cult of leadership, or the (equally disabling) complacency with the existing order of injustice via consumerism and the culture industry. The approach, as we saw, was to combine Marx's analysis of economic macrostructures with a Freudian social psychology such that the lack of revolutionary agency could be explained: Capitalism, so their argument, undermines the familial structures necessary for autonomous agency, and thereby opens the door for cultural-political formations such as fascism and consumerism. Here the subject identifies with surrogate pleasures, its own inability to change anything, or the submission under a strong leader who falsely represents one's own fake powers. The subtlety and comprehensiveness with which this research program was played out in all possible domains of social and cultural experience still today looks for an equal. From Horkheimer's, Adorno's and Marcuse's critique of ideology in philosophy; Adorno's, Löwenthal's and Benjamin's analysis of music, literature, and film; Pollock's and Kirchheimer's analysis of the economy and the state; as well as numerous other studies; the reconstruction of how economic and political powers shape agents such that the normative overcoming of an unjust social order remains impossible still today defines a practically unmatched standard for the fusion of empirical social research and philosophical conceptualization.

The Historical Unfolding of a Collaborative Social Theory

The Institute began a new era with Horkheimer and with his new journal *Zeitschrift für Sozialforschung* (Journal for Social Research), launched in 1932, as its public face. An important collaborator from the beginning of the Institute's Journal for Social Research was philosopher, sociologist, and musicologist Theodor W. Adorno. The first issue of the Institute's *Zeitschrift für Socialforschung* included contributions by Erich Fromm and Leo Löwenthal. Löwenthal had initial discussions with philosopher and political theorist Herbert Marcuse in Frankfurt in 1932, which led to his admission to the Institute in 1933. Thus Horkheimer's assistants at the Institute included Fromm, Pollock, Löwenthal, Adorno, and Marcuse. Looser connections were also established to scholars who at a later time would become more closely tied to the Institute, including political theorist Franz Neumann and political and legal theorist Otto Kirchheimer (1905–65).

By the beginning of 1933, the Institute and its members offered competence in a number of scholarly areas, with Marcuse and Horkheimer representative of the Institute's strength in philosophy, Fromm and Löwenthal covering psychology, Pollock, Karl August Wittfogel, and Grossman showing expertise in economy and economic history, Marcuse and Löwenthal in aesthetics and literature, and Grossman in statistics. From the winter semester of 1930–31 onwards, psychoanalyst and social psychologist Erich Fromm was listed as a member of the Institute's teaching staff, alongside Horkheimer, Grossman, and Pollock.

When Hitler became Chancellor in January that year, Horkheimer was ready with plans for a move from Germany. Pollock and Wittfogel were safe in Switzerland at the time, and the Institute's other members were soon able to leave Germany as well. Horkheimer had pondered the idea of relocating the Institute to another location in Europe, and branches had

already been established in Switzerland, Holland, and France. Horkheimer soon realized the only feasible alternative for relocation of the Institute would be to the United States. The Institute was offered an office building in New York City, and was given a contract that loosely tied them to Columbia University there. Therefore the Institute relocated to the United States in 1934, with their new home at Columbia University, where it would eventually stayed until 1950, at which time it was relocated back to the University of Frankfurt in Germany. After the initial relocation to the United States, the Institute's work slowly resumed there by the summer of 1934. By the autumn of that same year, nearly all of the Institute's scholars were reunited in the United States. However, Adorno stayed a bit longer in Germany and did not join the Institute in the United States until 1938. When the Institute first relocated to New York in 1934, Adorno went to Oxford University in England, where he unsuccessfully studied philosophy under the British analytic philosopher Gilbert Ryle (1900–76).

During its first years in the United States, the Institute worked on projects that had been initiated in Germany. Among the most important of these was an empirical social psychology project in which material had been collected before the Institute's relocation. The result of this project was published as *Studien über Autorität und Familie*. It represented the first, but unfortunately also the last, large-scale empirical study of the Institute displaying the original aims and scientific visions of Horkheimer's group of scholars. A forerunner to the study was almost completed by Fromm but never published. This fact, coupled with other internal difficulties of the Institute, led to the Institute's parting ways with Fromm in 1939. Fromm's importance at the Institute had been on the increase since the early 1930s. His social-philosophical interpretation of Freudian psychoanalysis, paired with Horkheimer's high regards for social-psychological explanations was integral for the *Studien*. Since the beginning of Horkheimer's reign, psychoanalysis had been a central method in the dialectical materialism of the Frankfurt School. In addition to Fromm's psychoanalysis and social psychology, Otto Kirchheimer pursed a path-breaking study in criminology while a member of the Institute in the 1930s. His coauthored study in criminology was titled *Punishment and Social Structure* (1939). The study became known as an early example of scholarship in the field of critical criminology.

Fromm eventually decided to stay permanently in the United States, becoming a major name in social psychology and psychoanalysis. Marcuse and Löwenthal also decided to stay in the United States after the Institute had relocated back to Frankfurt in 1950 with Marcuse, who taught at UC San Diego, becoming a major figure of the American New Left in the 1960s, and Löwenthal assuming a sociology professorship at UC Berkeley.

The Philosophical Turn and the Dialectic of Enlightenment

Horkheimer and Adorno coauthored *Dialectic of Enlightenment* in 1944 during their stay in Los Angeles, California. Horkheimer and Adorno lived in the Pacific Palisades neighborhood of Los Angeles, among a group of notable émigré intellectuals and artists who had fled Europe in the rise of National Socialism, including Thomas Mann (1875–1955), Bertold Brecht (1898–1956), and Arnold Schoenberg (1874–1951). *Dialectic of Enlightenment* became the most well known of the works in social philosophy from the Frankfurt School. While it certainly was, in some way, an immediate continuation of the problems discussed by earlier essays from the 1941 issue of the Institute's *Zeitschrift*, it was nevertheless also a complete redefinition of the project. In *Dialectic of Enlightenment*, a broad theoretical framework develops the decline, or rather the complete reversal, of Enlightenment reason. It is true, there was also a more specialized discussion of phenomena such as the 'Culture Industry' and anti-Semitism, and so a certain connection to the meta-theoretical framework and thus a continuation of the interdisciplinary vision from the earlier versions of the Institute's critical theory was still present. The *Dialectic of Enlightenment* was, in fact, a philosophical work, but the critical methodologies employed by Horkheimer and Adorno throughout the book ultimately made it multidisciplinary, and therefore still a valid example of the Institute's critical theory. *Dialectic of Enlightenment* was an exemplar of the Institute's project of interdisciplinary and multidisciplinary research to the extent that it entailed the discussion of philosophy, myth, enlightenment, morality (including the work of philosopher Marquis de Sade), anti-Semitism, and the Culture Industry. However, the overall context of these discussions was now a pessimistic account of the history of enlightened reason. The *Dialectic of Enlightenment* is ultimately a philosophical text that reconstructs, in the methodology of philosophy of history, how reason turned into its opposite. In such fashion, Critical Theory of this magnitude has become a negative philosophy of history.

Adorno's Rise and the Frankfurt School's Rise to Public Prominence

Adorno had continued to pursue multidisciplinary research in the United States after the publication of *Dialectic of Enlightenment*. His *Philosophy of New Music* was published in 1949, and he also focused on the sociology of music in that time period. He collaborated with Daniel J. Levinson, Max Horkheimer, and Else Frenkel-Brunswik in the writing of *The Authoritarian Personality* (1950). Adorno had collaborated with sociologist Paul Lazarsfeld (1901–76) in the Princeton Radio Research Project in New Jersey, as well as with Hanns Eisler (1898–1962) in *Composing for the Films* (1947). The Princeton Radio Research Project was made possible by funding from the Rockefeller Foundation in 1937, and the Institute for Social Research also had a hand in it. Horkheimer suggested to Lazarsfeld that he bring in Adorno to work on the project. When Adorno first moved from London to New York in 1938, he spent half of his time working for the Institute for Social Research and the other half working for the Princeton Radio Research Project. His stay at the Radio Research Project ended in 1941.

Eventually, Adorno emerged as a leading member of the Institute, and such a fact had consequences for the further development of critical theory at the Frankfurt School. Adorno's full entrance into the inner circle of the Institute meant that Horkheimer turned his attention to him and his enormous working capacity. In Adorno, Horkheimer saw a possibility to realize the project which he was dreaming of,

which was a project on dialectical philosophy. Marcuse up to this point had been the natural partner for Horkheimer in philosophical concerns, but when Adorno fully entered into the theoretical work of the Institute, he, rather than Marcuse, became Horkheimer's chief ally. The idea behind the critique of positivistic science, which was strongly influenced by Adorno, was introduced in the articles of the 1937 issue of the *Zeitschrift*. It was in the years following those articles that Horkheimer's close collaboration with Adorno developed. Such phase of critical theory at the Frankfurt School was a continuation of the arguments from its preceding phase, but took on a form which transcended the arena of science and was extended to a total critique of civilized reason. Horkheimer wrote *Eclipse of Reason* (1947) and *Critique of Instrumental Reason* (1967) on his own after collaborating with Adorno in the writing of *Dialectic of Enlightenment*.

By 1950, Horkheimer had reestablished the Institute in Frankfurt. By that time, Adorno and Horkheimer were the Institute's two undisputed leaders in critical theory. In the course of time, Horkheimer's productivity declined while Adorno had his most productive years in the 1950s and 1960s. In 1955, without giving up his influence, Horkheimer decided to share his directorship of the Institute with Adorno in a cooperative directorship. Adorno was highly visible and vocal as a public intellectual in Europe, engaging in the positivist dispute in German sociology and a critical dialogue with philosopher Karl Popper (1902–94). In addition, he wrote a polemic of the existentialism of philosopher Martin Heidegger (1889–1976), *The Jargon of Authenticity* (1964), and wrote *Negative Dialectics* in 1966, where he both continued his critique of Heidegger and extended *Dialectic of Enlightenment*'s critical theory. His *Aesthetic Theory* was posthumously published in 1970. In the postwar Germany after 1945, in the context of the return of 'the Frankfurt School' to Frankfurt itself, Adorno rose to become the public face of the Institute, which in turn rose to public prominence.

History of the Frankfurt School II: From 1970 to the Present

The New Beginning with Habermas

However, despite the comprehensiveness of the first program of 'critical social theory,' the initial founding paradigm – centered around the combination of Marx and Freud via an explanation of current forms of capitalist power – also entailed major problems. It is largely to the credit of one major social theorist, Jürgen Habermas, to establish his work as well as person as a new turning table, as a discursive and organizational medium in which these problems could be redefined and addressed. Habermas thereby managed to preserve the spirit of critical theory by advancing it through entirely new conceptual and methodological means. Habermas ascent to the leadership position of the so-called second generation of the Frankfurt School is prepared through his early influential work in the 1960s. Here, Habermas, a generation younger than the founders, established himself as an equal member of the now increasingly recognized 'Frankfurt School' by basically two practices. First, Habermas made himself a name in a set of important debates. He comes to Adorno's aid in the famous positivism debate with Popper, coining influential terms like the one about Popper's positivistically reduced concept of reason. Similarly, he engages in well-known debates with the most established figures of postwar Germany in philosophy and social theory, such as Hans-Georg Gadamer (1900–2002) and his philosophical hermeneutics, and Niklas Luhmann (1927–98) and his social systems theory. In these debates Habermas establishes himself as an apt critique of gaps and issues in both paradigms. Second, Habermas is able to present his own epistemological version of critical theory as based in human species-specific 'knowledge interests.' By suggesting that the natural sciences are grounded in the need and interest to dominate nature, the humanities are grounded in the need and interest for cultural and historical self-understanding, and the social sciences are grounded in the need and interest for emancipation from unjust social orders, Habermas situates critical social theory as a legitimate human endeavor next to other established scientific projects.

Yet the second generation of the Frankfurt School is borne only when Habermas abandons his own early theory for the sake of a 'paradigm change of critical social theory,' which leads to the 'theory of communicative action.' In a well orchestrated and not modest self-installation as the savior of critical theory via its complete conceptual overhaul, Habermas engages in a new beginning, triggered by the widely diagnosed problems of his earlier conception, which among other issues conflates the epistemological-transcendental critique of grounds of knowledge with a social-normative critique of society. In fact, the issue of the lack of an adequate conceptualization of the normativity of social criticism becomes the hallmark of the second generation and Habermas' lasting contribution to critical theory.

The Theoretical Core of Communicative Reason

Habermas comes to define the second generation of the Frankfurt School. He renews this tradition of social critique, of critical social theory by breaking with its conceptual framework. Instead of the Hegel–Marxian dialectic of enlightenment, in which 'reason' – which once meant to liberate and emancipate the subject from its objective determinants – becomes transformed into its opposite, *reason* for Habermas is now reconstructed as a linguistically mediated potential of a historically evolving human society. This potential for a rationalization of our linguistic practices can be unleashed in two diametrically opposed ways: either as the instrumental thinking that permeates functional systems like the capitalist economy and modern administrative power, or as the intersubjective orientation at communicative processes in which all agents are respected as equal, valuable, and knowledgeable contributors. Habermas now grounds the emancipatory potential of *modernity* – as he prefers to designate his indebtedness to the legacy of the Enlightenment – in the in-built presupposition of the validity of speech in social communication. This allows one to see that the reflexive distanciation from tradition is defined by a dialectic that does not turn reason per se into its dominating opposite. It rather defines rationalization as enabling two different processes, either as functional or as communicative reason. The modernization and rationalization of society since the Renaissance now

enables a liberation from tradition that can lead either to functionalist thought *or* to communicative action. For Habermas, this dialectic defines an open, contested, and challenging choice for modern societies. Instead of being doomed by a dialectic of Enlightenment according to which reason turned into its opposite, we are now called to actively participate in the reconstruction of the full scope of reason so as to avoid succumbing to functionalism and other objectifying and anticommunicative forms of power.

Habermas' communicative turn in critical theory proved to be successful as the new conceptual face of critical theory. Four reasons largely account for this success. First, the theory of communicative action can justify its own critical normative orientation. Horkheimer and Adorno's position had left unclear how a totally dominated agency can still rebel and revolt, and how a completely destroyed enlightenment can still speak of enlightenment values like autonomy, truth, and equality. Second, Habermas' theory represents an unparalleled achievement of theoretical synthesis. The over 1000 pages of the two-volume *Theory of Communicative Action* (German 1981, English 1983/87) integrate the major sociological classics, argumentation theory, critical theory, and speech act theory to a new conception of validity-based communicative action. Third, Habermas' theoretical product appears at a time when the globalization of theory production takes off. Habermas' theory draws as much on American as on German sources, including Charles Sanders Peirce (1839–1914), George Herbert Mead (1863–1931), and Talcott Parsons (1902–79), and generally speaks to the new cosmopolitan class of social and cultural theorists who emphasize language, discourse, and a politics of recognition in the public sphere, the latter being one of Habermas' central concepts and research topics. Fourth, Habermas has been supported by several influential scholars, such as Albrecht Wellmer, Karl-Otto Apel, and Thomas McCarthy. Besides those, the second generation also includes thinkers like the Adorno's student Alfred Schmidt (1931–2012). Wellmer and Schmidt, along with others, some who would in later certain readings of the school find themselves regarded as members of the third generation at the Institute, contributed to an increasing academic as well as public recognition of the Frankfurt School as one of the most important German philosophical movements of the twentieth century.

Establishing the Communicative Paradigm and the Historization of the Frankfurt School

The brief conceptual sketch allows us to see how Habermas could indeed establish his communicative proposal at the center of a renewed critical theory. Yet as Bertolt Brecht rhetorically asked the historian who wrote that Caesar crossed the Alps: "Did he not at least have a cook with him?" we should note that Habermas did not do this on his own. Indeed, Habermas is an authorial center of discourse, as the actual author of the most influential books on the communicative turn; he is also the organizational center as the director of the Max Planck Institute for Social Research in Starnberg, which he codirected from 1969 to 1983; and he becomes, after the deaths of most the old guard in the late 1960s to mid-1970s, the increasingly famous living representative of the Frankfurt School. Yet on every account, the communicative turn is the work of many colleagues and contributors; it is impossible to conceive without such context and support. Theoretically, the transcendental pragmatics of philosopher Karl-Otto Apel (1922–present) gave essential insights, as did the analytic rigor and pragmatist acumen of American philosopher Thomas McCarthy (1940–present). Institutionally, the Starnberg Institute gave Habermas the resources to assemble some of the world's most relevant philosophers and social scientists in a uniquely productive interdisciplinary atmosphere. And finally, Habermas was a public persona who drew attention to his rather academically articulated theory. His important public standing emerged from his constant historico-political interventions which related him, at times antagonistically, to the student revolt of 1968 and the new democratic Germany; also important for Habermas' reputation were his at-times heated engagements with some of the most prominent post-structuralists and postmodernists, which include his highly visible critique of post-structuralism, especially with regard to Michel Foucault (1926–84) and Jacques Derrida (1930–2004).

With the death of Adorno in 1969, the withdrawal of Horkheimer to Switzerland around that time, Habermas' move to Starnberg in 1969, and with Marcuse and Löwenthal faraway in sunny California, the actual Institute for Social Research in Frankfurt became a rather modest center for social-empirical research. Under the able but unassuming directorship of Ludwig von Friedeburg (1924–2010), also minister of culture for the state of Hesse, solid social research projects were carried out. When the real action had moved elsewhere, the Institute for Social Research became an historical site of the imaginary entity of the Frankfurt School. What set in during the 1970s, besides the Habermasian paradigm shift and renewal, was the historization of the Frankfurt School as one of the most important twentieth-century movements in German philosophy and social theory. Rather marginalized as it was before the Second World War, and still long after, the Frankfurt School now became an object of historical desire, a piece in the mosaic of a new Germany that desperately searched for a new identity beyond Hitler. Works by Helmut Dubiel, the Californian Martin Jay, and the path-breaking volume by Rolf Wiggershaus defined this historical turn. Despite the dangers that were associated with such a historization, this move brought the first generation of Horkheimer, Adorno, Benjamin, etc., back into academic and public consciousness. It enables up until today to see critical theory as what it really is a complex and diversified project in which a multiplicity of approaches and methodologies come to bear to analyze the dialectic of human freedom and equality against the real backdrop of social power and domination.

Habermas returns to the Johann Wolfgang von Goethe Universitat in Frankfurt in 1983. The move comes after turning down offers from the University of California, Berkeley and Northwestern University and is well calculated: while Habermas does not formally assume the directorship of the still existing *Institute for Social Research*, he inhabits the University Philosophicum right next to it. The return of the master reinvigorates the historical awareness of the Frankfurt School, albeit Habermas aims to build his reputation as the undisputed king of critical theory by three essential and independently pursued projects. First, he writes *The Philosophical Discourse of*

Modernity, a book of questionable value as it is written in a largely polemical tone similar to Georg Lukacs' equally one-sided *Destruction of Reason*. Habermas blames the Nietzsche/Heidegger turn in contemporary French philosophy for the flaws to be found in Foucault, Derrida, Jean-François Lyotard (1924–98), and the like. Second and much more constructive is his work on a 'discourse ethics' which achieves a strong worldwide reception. Finally, and of perhaps the most sustaining influence, is his work on a new 'Rechtsphilosophie,' a 'philosophy of right,' culminating in *Facticity and Norms*, where he develops a democratic-theoretical account of the complex intertwinement of morality and law. Habermas also leads during this decade the famed Monday Colloquium where many of the best German and international, young and established, critical theorists presented and discussed their work in progress. The years from the mid-1980s to the mid-1990s, when Habermas retired from his Frankfurt post to teach on occasion at Northwestern, New York University, and elsewhere, define the coming of age of critical theory and its ascent to one of the world's most known traditions on social theory.

The Third Generation: Recognition, Feminism, Justice

The sustained productivity of the Frankfurt School can be measured by the fact that a third generation of influential and prolific authors is currently carving out new intellectual space for critical theory. While critical theory has justly become a term with much wider application than the initial Frankfurt School tradition, representatives that directly draw from it nevertheless exist and include social theorists such as Axel Honneth. Honneth is unique since with him, for the first time since the demise of Adorno in 1969, a major social philosopher both represents the philosophical heritage and organizationally heads the actual *Institute for Social Research*. Honneth succeeded Habermas on his chair for social philosophy in Frankfurt, and shortly afterward also assumed the directorship of the Institute. Honneth's increasingly influential theory of recognition builds on insights from the early Hegel and G.H. Mead in order to construct a threefold understanding of recognition: love or emotional recognition, respect or moral recognition, and esteem or sociocultural recognition represent for Honneth, three anthropologically deep-seated needs the fulfillment of which agents require to fully develop as autonomous selves; conditions that withhold such fulfillment can be understood to fuel social struggles for recognition. Emotional recognition builds self-confidence, moral recognition self-respect, and cultural recognition self-esteem – all three of which together define an ethically acceptable and good life for which agents fight if the conditions for their realization are withheld from them. The communicative recognition that Habermas grounded in argument-based validity claims thus receives a deeper and more social-philosophical and social-historical basis. Honneth advances a truly critical-theoretical approach, as such a perspective – since the early Frankfurt School – has always consisted in combining a normative orientation toward more justice and freedom with an analysis of the social-historical and empirical grounds that would enable its realization practically. As Institute director as much as social philosopher, Honneth searches for constant dialogue and expansion with related critical-theoretical approaches. This led to a much noted exchange with Nancy Fraser in *Redistribution or Recognition?* in which the focus on moral and cultural recognition is challenged by the continued structural need for economic justice. Nancy Fraser and Seyla Benhabib represent third generation critical theorists who in a distinctively feminist vein approach issues of identity, power, gender roles and justice, as well as globalization and cosmopolitanism – besides also internationalizing critical theory in person. Closely affiliated with Axel Honneth, and important because of his work on G.H. Mead, Hans Joas may also be seen to make an important contribution. All these thinkers represent potent voices in an increasingly transnational theory discussion over the right approach to the social-political challenges from a normatively informed view.

There are in fact different strands or readings of who else belongs to the third generation, both in Germany and internationally, and some in sociology and social and political theory more than philosophy. Joas is at times read into the tradition of the third generation in social theory. With the increasing internationalization of critical theory and with a look at theoretical developments of both Habermas' discourse ethics and Axel Honneth's theory of recognition, Nancy Fraser's and Seyla Benhabib's work has become a constitutive part of a third generation's critical theory. Claus Offe and Oskar Negt are in certain readings, especially but not only in Germany, also regarded as third generational. Offe's monographs first started to appear in 1965. Since then, he has engaged contemporary conceptions of the welfare state and its implicit contradictions, and has characterized late capitalism as disorganized. Negt, who trained under Adorno and then went on to serve as an assistant to Habermas, has further developed Habermas' notion of the public sphere and has embarked on a theory of history in his more philosophical writings.

Honneth and with him Joas, Offe and Negt, and Fraser and Benhabib are among the most prominent contributors of the third generation of a school the influence of which is, however, much more widespread and sustained than the mentioning of these names could capture. Noteworthy are also recent attempts to transgress and cross-over between different paradigms of critical theory, which includes building bridges between the Frankfurt School in all its versions and British Cultural Studies, French Post-Structuralism, and contemporary social theory in all of its different forms. Besides the original Institute of Social Research, which continues to exist under its current director Axel Honneth, several important venues of exchange and research grounded in the Frankfurt School critical theory have sprung up in recent decades: The American Critical Theory Roundtable, the Prague Conference of the Social Sciences, and the Frankfurt University Research Cluster Normative Orders. Critical social theory, as conceived by its original founders, did care little about discursive boundaries or school designations; its aims has been to advance critical reflexivity with regard to the most pressing and least understood currents of our contemporary and continuously unjust social orders. Its motto was and is the critique of the social status quo by any theoretical means necessary, with the aim to be able to conceive its practical overcoming. Unfortunately, due to the continued pervasive influence of power, oppression, and ignorance, critical social theory and the project of

the Frankfurt School has lost nothing of its urgency in today's world.

See also: Benhabib, Seyla (1950–); Critical Theory; Democracy: Normative Theory; Foucault, Michel (1926–84); Fromm, Erich (1900–80); Habermas, Jürgen (1929–); Hebb, Donald Olding (1904–1985); Kant, Immanuel (1724–1804); Law and Morality: A Continental European Perspective; Lazarsfeld, Paul Felix (1901–76); Luhmann, Niklas (1927–98); Marx, Karl (1818–83); Mead, George Herbert (1863–1931); Nietzsche, Friedrich (1844–1900).

Bibliography

Abromeit, John, 2011. Max Horkheimer and the Foundations of the Frankfurt School. Cambridge University Press, Cambridge.

Benhabib, Seyla, 1992. Situation the Self: Gender, Community, and Postmodernism in Contemporary Ethics. Routledge, New York.

Benjamin, Walter, 1968. The work of art in the age of mechanical reproduction (Harry Zohn, Trans.). In: Arendt, Hannah (Ed.), Illuminations: Essays and Reflections. Schocken Books, New York, pp. 217–251.

Bottomore, Tom, 1984. The Frankfurt School. Ellis Harwood Limited, Chichester.

Dubiel, Helmut, 1985. Theory and Politics: Studies in the Development of Critical Theory. The MIT Press, Cambridge, MA.

Fraser, Nancy, 1989. Unruly Practices: Power, Discourse, and Gender in Contemporary Social Theory. University of Minnesota Press, Minneapolis, MN.

Habermas, Jürgen, 1984. The Theory of Communicative Action (Thomas McCarthy, Trans.). In: Reason and the Rationalization of Society, vol. 1. Beacon Press, Boston.

Habermas, Jürgen, 1987. The Theory of Communicative Action (Thomas McCarthy, Trans.). In: Lifeworld and System: A Critique of Functionalist Reason, vol. 2. Beacon Press, Boston.

Held, David, 1980. Introduction to Critical Theory: Horkheimer to Habermas. University of California Press, Berkeley, CA.

Honneth, Axel, 1995. The Struggle for Recognition: The Moral Grammar of Social Conflicts (Joel Anderson, Trans.). Polity Press, Cambridge.

Honneth, Axel, Fraser, Nancy, 2004. Redistribution or Recognition? A Political-philosophical Exchange. Verso, London and New York.

Horkheimer, Max, Adorno, Theodor W., 2002 (Edmund Jephcott, Trans.). In: Schmid Noerr, Gunzelin (Ed.), Dialectic of Enlightenment: Philosophical Fragments. Stanford University Press, Stanford, CA.

Jay, Martin, 1996. The Dialectical Imagination: A History of the Frankfurt School and the Institute of Social Research, 1923–1950. University of California Press, Berkeley, CA.

Marcus, Judith, Tar, Zoltán (Eds.), 1984. Foundations of the Frankfurt School of Social Research. Transaction Books, New Brunswick, NJ; London.

Marcuse, Hebert, 1991. One-dimensional Man: Studies in the Ideology of Advanced Industrial Society. Beacon Press, Boston.

Tar, Zoltán, 1977. The Frankfurt School: The Critical Theories of Max Horkheimer and Theodor W. Adorno. Wiley, New York.

Wheatland, Thomas, 2009. The Frankfurt School in Exile. University of Minnesota Press, Minneapolis, MN.

Wiggershaus, Rolf, 1994. The Frankfurt School: Its History, Theories, and Political Significance (Michael Robertson, Trans.). The MIT Press, Cambridge, MA.

Freedom, History of

Frank Lovett, Washington University in St. Louis, St. Louis, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the history of the ideal of political freedom from ancient times to the present day. Prior to the seventeenth century, political freedom was widely understood to mean not having a master, but the value and feasibility of this good was contested. Hobbes introduced the 'negative' conception of freedom as noninterference in the seventeenth century, and Rousseau introduced the 'positive' conception of freedom as effective self-mastery in the eighteenth century. The debate between these two conceptions has dominated discussions ever since, obscuring the older and once widely shared view.

This article discusses the history of the political ideal of freedom or liberty, that is, what it means for a person or group to enjoy freedom or liberty in the political sense. (Freedom and liberty are simply two words for the same concept, the first of Germanic and the second of Latin origin; they will here be used interchangeably.) This article will thus not discuss freedom of the will and determinism, or what was at one time debated under the heading of 'liberty and necessity'.

As is well known, the meaning of political freedom is presently the subject of much debate among political theorists and philosophers. Remarkably, this was not always so. Indeed, prior to the late eighteenth century, its meaning was debated relatively little, and prior to the mid-seventeenth century, scarcely at all. This is not to say that there were no vigorous debates about freedom or liberty before that time; quite the contrary, as we shall see. But these debates generally did not concern what it *meant* for a person or group to enjoy freedom in the political sense.

Historical Origins

For a person or group to enjoy freedom or liberty in the political sense, on the long-standing standard view, was simply for them to enjoy a certain sort of independence. Roughly speaking, freedom was understood as the sort of independence enjoyed by an individual who, not being a slave, serf, bondsman, or servant, does not have a master, and so is a 'free person', or, analogously, the sort of independence enjoyed by a community that does not live under the rule of an autocratic king, tyrant, oligarchy, or foreign empire, and so is a 'free people'. In part because the meaning of political freedom was not itself the subject of controversy, its precise scope remained vague and open-ended. It was a negative ideal, in the straightforward sense that to enjoy freedom was simply *not* to have a master of one sort or another.

This traditional conception of political freedom or liberty had its roots in the ancient world. In Greece, freedom (*eleuthera*) was viewed as the independence men and women feared to loose if they were captured in war or succumbed to debt and thus became slaves. In the *Iliad*, Hektor tells his wife that he is troubled less by the prospect of Troy's fall than by

> …the thought of you, when some bronze-armored Achaian leads you off, taking away your day of liberty, in tears; and in Argos you must work at the loom of another…unwilling, but strong will be the necessity upon you.
> (p. 165)

Freedom was analogously understood as the self-rule of independent city-states that Athens and her allies defended at Marathon and Salamis against Persian imperialism. Thus, according to Herodotus, a Spartan explained to a Persian that

> You understand well enough what slavery is, but freedom you have never experienced, so you do not know if it tastes sweet or bitter. If you ever did come to experience it, you would advise us to fight for it not with spears only, but with axes too.
> (p. 413)

At Rome, liberty (*libertas*) was thought to have been established when the Tarquins were expelled, briefly endangered when the decemviri tried to seize permanent authority for themselves, and finally lost in the dissolution of the Republic by the Caesars. Cicero, for example, observes that

> …if the people hold to their own rights, they deny that there is anything more outstanding, more free, more blessed …. [A]nd so it is usual for the "concern of the people" to be liberated from the domination of kings and aristocrats, and not for kings or the power and wealth of an aristocracy to be sought by a free people.
> (p. 21)

Historically speaking, philosophical interest in freedom or liberty was greatest when it was most felt to be in danger. It is most energetically extolled in Herodotus' and Aeschylus' writings on the Persian wars, or by Aristotle and Demosthenes as the independent city-states fell under Macedonian rule, for instance. Similarly, most Latin reflections on liberty come from around the time of the collapse of the Republic, especially in the works of Cicero, Sallust, and Livy. After centuries of relative disinterest, a similar pattern can be observed in the Renaissance: although independent city-state republics populated northern Italy for much of the Middle Ages, only in the fifteenth century – as their self-rule faced increasing domestic and foreign challenges – did Bruni, Rinuccini, and Machiavelli pen their panegyrics to freedom and liberty.

It was this traditional conception that the early-modern English political writers inherited. While they were certainly well read in the classical and Renaissance sources that we have mentioned, the text that – directly or indirectly – informed them most profoundly was not produced by a political theorist at all. It was rather the *Corpus Iuris* of Justinian that for centuries fixed the meaning of liberty in Western political thought. Early on, the *Digest* states flatly that all law concerns persons, things, or actions, and that the main distinction among persons is that between those who are free on the one hand, and those who are slaves on the other. This sharp contrast between freedom and servitude is reiterated again and again in the *Corpus*, and constitutes one of its fundamental conceptual legacies. It is hardly surprising, then, that it was above all to this notion of freedom that writers like Milton, Harrington, Sidney, and Locke unconsciously and repeatedly turned during England's constitutional struggles of the seventeenth century.

Now while there was little dispute concerning what it *meant* to enjoy freedom, there were many significant debates *about* freedom. Two controversies in particular frequently recurred. The first was normative, and concerned how highly freedom should be valued relative to other important social or political goods. While perhaps a majority of the ancients regarded freedom very highly indeed, there were always some dissenting opinions: Plato, for instance, argued vigorously in the *Republic* that justice was far more important than the freedom enjoyed by the citizens of a democracy. During the Middle Ages, advocates for the priority of freedom were decidedly in the minority. Far more important than freedom, argued Augustine, Aquinas, and many other writers, was peace and order. Debates concerning the relative value of freedom continued in the Renaissance and after. Castiglione illustrates the debate clearly in the fourth book of his *Book of the Courtier*, where one character argues that a free republic is the supreme gift of God, while another replies that a tranquil monarchy is more consistent with living a dignified life in peace and quiet.

The second recurring controversy was empirical. It concerned how much freedom could be safely enjoyed without undermining social order. Especially after the defeat of Athens in the Peloponnesian War, it was widely assumed that too much unstructured independence could be self-defeating, and thus self-government needed to be carefully balanced and contained within the framework of what Polybius termed a 'mixed constitution'. Subsequent advocates for political freedom, from Cicero to Machiavelli to Montesquieu, were all insistent that its preservation depended on maintaining a balanced political system and a strict rule of law. Critics of freedom, such as William Barclay and Robert Filmer, remained unconvinced that such a system could be self-sustaining; a society of 'masterless men', they believed, must necessarily slide into chaos.

Both these debates and others, however, revolved around a conception of freedom or liberty that was largely uncontroversial. While the relative value or practical feasibility of political freedom might be contested, it was largely taken for granted that everyone understood the idea under discussion: to be free was, plain and simple, to have no master.

Hobbes and Rousseau

Two great political thinkers, from the seventeenth and eighteenth centuries, were primarily responsible for upending the traditional understanding of political freedom and throwing its meaning into doubt. These thinkers were, respectively, Thomas Hobbes and J.-J. Rousseau.

Perhaps to a greater extent than any previous thinker, Hobbes recognized both the power, and also the mutability, of definitions. Writing during the English Civil War, and fervently believing that only the general recognition of an absolute sovereign could deliver an end to the conflicts of his day, Hobbes faced a peculiar challenge. On the long-standing traditional view, political freedom and absolute sovereignty are strictly incompatible. The advocates of limited sovereignty and commonwealth government accordingly leaned heavily on the powerful rhetorical appeal of freedom. Machiavelli had indeed observed the force of this appeal some time before:

> Whoever becomes the master of a city accustomed to freedom, and does not destroy it, may expect to be destroyed himself; because, when there is a rebellion, such a city justifies itself by calling on the name of liberty …never forgotten despite the passing of time and the benefits received from the new ruler.
> (p. 18)

In order to diffuse this rhetorical appeal, Hobbes sought a conception of freedom that, contrary to all previous understanding, could be shown compatible with absolute sovereignty. He found his answer in what is now widely called the 'negative' conception of freedom.

On the negative conception of freedom, we are free to the extent that we are merely not interfered with by others. There have been many variations on this conception since Hobbes, depending on how exactly one chooses to define 'interference' in the relevant sense, but they all have in common the basic intuition that to be free is, more or less, to be left alone to do whatever one pleases. Now of course the value in being left alone had long been recognized; contrary to popular opinion, even the ancients were perfectly well attuned to this value. Thus Thucydides recounts that Pericles approvingly remarked of Athens that

> …just as our political system is free and open, so is our day-to-day life in our relations with each other. We do not get into a state with our next-door neighbor if he enjoys himself in his own way ….
> (p. 145)

But it was always recognized that an absence of interference – sometimes termed 'license' – is not at all the same as freedom or liberty in the political sense. (Indeed, it was pointed out by Machiavelli and others that too much of the former could actually undermine the latter by eroding the rule of law.) The two ideas are distinct because, while the traditional view of freedom is indeed 'negative' in the simple sense described earlier of consisting in the absence of something, having a master might easily be consistent with enjoying a considerable degree of noninterference. An imperial power might, for

example, refrain from interfering much with the local administration of its colonies, and yet those colonies do not, on the traditional view, thereby enjoy political freedom.

This last point is especially significant because it answers precisely to Hobbes's aims. On the new negative conception, as he says, political freedom is merely the "silence of the laws" (p. 146), and there can be as many or as few laws under any one form of government as under any other. Thus, he infers, there can be no logical grounds for objecting to absolute sovereignty on the basis of political freedom:

> There is written on the turrets of the city of Lucca in great characters at this day, the word libertas; yet no man can thence infer, that a particular man has more liberty, or more immunity from the service of the commonwealth there, than in Constantinople. Whether a commonwealth be monarchical, or popular, the freedom is still the same.
> (p. 143)

By redefining political freedom in this way, he hoped to defang the most potent objection to absolute sovereignty.

Not surprisingly, Hobbes's new conception did not immediately take hold. James Harrington among others quickly recognized the sleight of hand and pointed it out:

> The mountain has brought forth, and we have a little equivocation! For to say that a Lucchese has no more liberty or immunity from the laws of Lucca than a Turk has from those of Constantinople, and to say that a Lucchese has no more liberty or immunity by the laws of Lucca than a Turk has by those of Constantinople, are pretty different speeches. The first may be said of all governments alike, the second scarce of any two; much less of these, seeing it is known that whereas the greatest bashaw is a tenant, as well of his head as of his estate, at the will of his lord, [while] the meanest Lucchese … is a freeholder of both ….
> (p. 20)

For some time thereafter, the negative conception of freedom remained no more than an intellectual curiosity. A generation later, we find an especially clear statement of the traditional view reiterated in John Locke: freedom is not "a liberty for every one to do what he lists", he writes, but rather "to have a standing rule to live by" and "not to be subject to the inconstant, uncertain, unknown, arbitrary will of another man" (p. 110). Indeed, the traditional conception remained dominant well into the eighteenth century, as is clear from a reading of *Cato's Letters*, Blackstone's *Commentaries*, or the pro-American pamphlets of Priestly, Price, or Paine. Only much later, as we shall see below, did Hobbes's negative conception of liberty catch on.

In the meantime, we must consider the second major challenge to the traditional conception of freedom, this one raised by Rousseau. Like Hobbes, Rousseau was interested in showing that freedom is consistent with absolute sovereignty – provided, on his version of argument, that the sovereign is the whole body of citizens. This he does by drawing a distinction between what he terms the "natural liberty" enjoyed in the state of nature on the one hand (roughly, the negative liberty of Hobbes), and what he terms the "civil liberty" enjoyed through community on the other. The latter he defines as "obedience to the law one has prescribed for oneself" (p. 151). This has come to be known as the 'positive' conception of freedom.

The positive conception of freedom is more difficult to characterize with analytic precision than the negative. Roughly speaking, persons or groups are free in the positive sense to the extent that they exercise some degree of self-control or self-mastery. It is not agreed, however, what exactly constitutes self-mastery in the relevant sense. According to one particularly influential account, offered by Harry Frankfurt, to be free in the positive sense is to be able to act on one's second-order desires. For example, the addicted gambler may be free in the negative sense not to gamble (no one forces him to gamble), but he is not free in the positive sense unless he can actually succeed in acting on his second-order desire not to desire gambling. Applied to the political sphere, we can see that something like this thought animates Rousseau's conception: "to be driven by appetite alone", he says, "is slavery" (p. 151). It follows that, when the community as a whole compels one of its members to realize some second-order desire, he is merely "forced to be free" (p. 150).

Rousseau no more invented his distinctive conception of freedom out of whole cloth than did Hobbes. From the time of the Stoics, it had been commonplace to use the language of freedom and slavery when discussing the struggle to master one's own passions and desires. This language was adopted by early Christian writers, who frequently contrasted the 'slavery of sin' with the 'freedom of righteousness', a popular formula down to the time of Martin Luther's famous tract on "The Freedom of a Christian" and beyond. This usage was always clearly metaphorical, however. It was never suggested that freedom or liberty in the *political* sense should be understood in this manner. Indeed, the force of the metaphor in discussions of personal ethics was parasitic on the standard view of political freedom as consisting in the absence of a master. It was Rousseau (anticipated perhaps by Spinoza, but with little consequence for respectable opinion) who first advanced the positive conception as an account of *political* freedom, with remarkable consequences for the history of political thought, as we shall see.

III

In June 1789, representatives of the Third Estate abandoned the Estates-General and declared themselves the National Assembly of France. Together with the storming of the Bastille by the people of Paris in July, these events initiated what would become a quarter century of revolution, international war, and social turmoil in Europe – an experience that naturally had a great many consequences for the history of political thought generally, and for the concept of freedom or liberty specifically. When during the darkest days of the revolution the Committee of Public Safety subjected France to a reign of terror, it did so in the name of *Liberté, égalité, fraternité*. When Napoleon's armies spread across Europe, they did so in the name of revolutionary liberation. What were people to make of these proclamations?

Reflecting on such events a few years after peace had finally been restored, Benjamin Constant delivered a famous and influential speech on 'The Liberty of the Ancients Compared

with that of the Moderns'. The societies of ancient Greece and Rome, he argued, were very different from our modern societies. They were small, tight-knit communities; their relative vulnerability meant that military affairs necessarily loomed larger than commercial affairs; and their employment of slave labor provided citizens with ample leisure. It is only to be expected, he said, that the ancients would understand political freedom in a manner reflecting their circumstances. For them, freedom meant "exercising collectively, but directly, several parts of the complete sovereignty" of the community as a whole. This they were able to do only because their societies were relatively small and because they enjoyed sufficient leisure time in which to dedicate themselves to public affairs. However, "the complete subjection of the individual to the authority of the community" is "compatible with this collective freedom". In the ancient societies, Constant claimed, no "importance was given to individual independence" (p. 311).

Modern societies are entirely different from ancient ones. They are large, diverse, and impersonal; commercial interests predominate over military interests; and of course, the practice of slavery has been abolished. The moderns, Constant said, have a very different conception of political freedom, better suited to their own times. On this modern view, freedom is simply

> …the right of everyone to express their opinion, choose a profession and practice it, to dispose of property, and even to abuse it; [and] to come and go without permission, and without having to account for their motives or undertakings.
> (pp. 310–311)

It is immediately apparent that what is being described here is the negative liberty of Hobbes – the freedom to be left alone, without outside interference. Negative liberty, Constant claims, is the only sort of freedom suitable to modern conditions. The excesses of the French Revolution arose precisely from the futile dream, inspired by Rousseau, of reviving positive liberty of the ancients in modern times.

His diagnosis proved extremely influential. Many writers soon embraced enthusiastically the negative conception of freedom as noninterference, among them Jeremy Bentham and William Paley. Their combined influence was such that by the time J.S. Mill wrote his essay *On Liberty* in the 1850s, he could write without pausing for debate that "the only freedom which deserves the name, is that of pursuing our own good in our own way, so long as we do not attempt to deprive others of theirs" (p. 17). The negative conception of freedom is clearly the dominant view today, particularly among Anglo-American political theorists and philosophers. Of course, there has been a dissenting minority who did not give up on Rousseau's dream – Hegel, for instance, and some other continental philosophers. Later supporters of the positive conception of liberty include the English Hegelians like T.H. Green, and more recently, some communitarians like Charles Taylor, but these writers have never been able to overturn the dominance of the negative view.

Nearly a century and a half after Constant, Isaiah Berlin delivered an even more famous lecture reviewing the debate between positive and negative freedom, 'Two Concepts of Liberty'. Berlin reiterates Constant's warning against the dangers of positive liberty. "It is one thing to say that I may be coerced for my own good which I am too blind to see", he says.

> [T]his may, on occasion, be for my benefit …. [But] it is another to say that if it is my good, then I am not being coerced, for I have willed it, whether I know this or not, and am free (or 'truly' free) even when my poor earthly body and foolish mind bitterly reject it, and struggle against those who seek however benevolently to impose it ….
> (p. 134)

This warning is, no doubt, well heeded. What is peculiar about Berlin's lecture, however, and about the earlier one by Constant on which it was (consciously or unconsciously) based, is that the positive conception of liberty attributed to the ancients was not in fact the understanding of political freedom to which they – nor indeed anyone, prior to Rousseau – subscribed. Negative liberty is held up as the political ideal much safer than its positive alternative, without apparently recognizing that both are modern inventions.

What explains this peculiar turn of events? One possibility, of course, is simply that the magnetic power of Rousseau's political writings sewed the confusion. For subsequent writers, 'the ancients' simply became the ancients as Rousseau imagined them, and the 'freedom' they enjoyed the conception of freedom Rousseau extolled: both those attracted to his vision and those repelled by it could no longer see past it clearly. Another possibility, however, is that deeper interests were at work. The early nineteenth century was a period in which the working classes, servants, women, and national minorities were steadily forcing themselves onto the political stage. Combining the traditional view of freedom or liberty with a more inclusive political ethic would have potentially revolutionary consequences, as a few writers astutely recognized: it would mean granting independence from mastery to women, to servants, and so forth. Thus William Paley suggests that

> …those definitions of liberty ought to be rejected, which, by making that essential to civil freedom which is unattainable in experience, inflame expectations that can never be gratified, and disturb the public content with complaints, which no wisdom or benevolence of government can remove.
> (p. 315)

Far less disruptive to the social order is the negative conception of political freedom which, as we have seen, is perfectly consistent with mastery so long as that master lets one alone. On this second reading of intellectual history, the positive conception of political freedom was little more than a convenient straw man.

Although the negative freedom–positive freedom dichotomy continues to shape discussion among contemporary political theorists and philosophers, there has recently been a revival of interest in the traditional view. Beginning primarily with historical work on Machiavelli and the classical republican tradition, and increasingly branching out into areas of contemporary normative political philosophy, Quentin Skinner, Philip Pettit, and others have challenged the conceptual topography laid

down by Constant and Berlin, and it is difficult to foresee where the debate will go from here.

See also: Berlin, Isaiah (1909–97); Free Will and Action; French Revolution, The; Liberalism: Historical Aspects; Republicanism: Philosophical Aspects; Rousseau, Jean-Jacques (1712–78).

Bibliography

Arendt, Hannah, 1993. What is freedom? In: Between Past and Future: Eight Exercises in Political Thought. Penguin Books.

Baron, Hans, 1966. The Crisis of the Early Italian Renaissance. Princeton University Press.

Berlin, Isaiah, 1969a. Two concepts of liberty. In: Four Essays on Liberty. Oxford University Press.

Berlin, Isaiah, 1969b. Four Essays on Liberty. Oxford University Press.

Cicero, 1999. On the Commonwealth and On the Laws (E.G.Z. James, Trans.). Cambridge University Press.

Constant, Benjamin, 1988. The liberty of the ancients compared with that of the moderns. In: Fontana, Biancamaria (Ed.), Political Writings. Cambridge University Press.

Frankfurt, Harry G., 1982. Freedom of the will and the concept of a person. In: Watson, Gary (Ed.), Free Will. Oxford University Press.

Harrington, James, 1992. In: Pocock, J.G.A. (Ed.), The Commonwealth of Oceana and a System of Politics. Cambridge University Press.

Hayek, Friedrick A., 1960. The Constitution of Liberty. University of Chicago Press.

Herodotus, 1954. The Histories (Aubrey de Sélincourt, Trans.). Penguin Classics.

Hobbes, Thomas, 1996. In: Gaskin, J.C.A. (Ed.), Leviathan. Oxford University Press.

Homer, 1951. The Iliad (R. Lattimore, Trans.). University of Chicago Press.

Laborde, Cécile, Maynor, John (Eds.), 2008. Republicanism and Political Theory. Blackwell Publishing.

Locke, John, 2003. In: Shapiro, Ian (Ed.), Two Treatises of Government and a Letter Concerning Toleration. Yale University Press.

MacCallum, Gerald C., 1991. Negative and positive freedom. In: Miller, David (Ed.), Liberty. Oxford University Press.

Machiavelli, Noccolò, 1961. The Prince (G. Bull, Trans.). Penguin Classics.

Mill, John Stuart, 1991. In: Gray, John (Ed.), On Liberty and Other Essays. Oxford University Press.

Miller, David (Ed.), 1991. Liberty. Oxford University Press.

Paley, William, 2002. The Principles of Moral and Political Philosophy. Liberty Fund.

Patterson, Orlando, 1991. Freedom in the Making of Western Culture. Basic Books.

Pettit, Philip, 1989. The freedom of the city: a republican ideal. In: Hamlin, Alan, Pettit, Philip (Eds.), The Good Polity. Blackwell Publishers.

Pettit, Philip, 1991. Reprinted in liberty. In: Miller, David (Ed.), The Paradoxes of Political Liberty. Oxford University Press.

Pettit, Philip, 1997. Republicanism: A Theory of Freedom and Government. Clarendon Press.

Pettit, Philip, 1998. Liberty before Liberalism. Cambridge University Press.

Pettit, Philip, 2001. A Theory of Freedom: From the Psychology to the Politics of Agency. Oxford University Press.

Pettit, Philip, 2003. Agency-freedom and option-freedom. Journal of Theoretical Politics 15, 387–403.

Pohlenz, Max, 1966. Freedom in Greek Life and Thought: The History of an Ideal (C. Lofnark, Trans.). Humanities Press.

Rousseau, Jean-Jacques, 1987. The Basic Political Writings (D.A. Cress, Trans.). Hackett Publishing.

Ryan, Alan (Ed.), 1979. The Idea of Freedom: Essays in Honour of Isaiah Berlin. Oxford University Press.

Skinner, Quentin, 1984. The idea of negative liberty. In: Rorty, Richard, Schneewind, J.B., Skinner (Eds.), Philosophy of History: Essays on the Historiography of Philosophy. Cambridge University Press.

Taylor, Charles, 1991. What's wrong with negative liberty? In: Miller, David (Ed.), Liberty. Oxford University Press.

Thucydides, 1954. History of the Peloponnesian War (R. Warner, Trans.). Penguin Classics.

Watson, Gary (Ed.), 1982. Free Will. Oxford University Press.

Wirszubski, Chaim, 1950. Libertas as a Political Idea at Rome during the Late Republic and Early Principate. Cambridge University Press.

Relevant Website

http://plato.stanford.edu/entries/liberty-positive-negative/ – Stanford Encyclopedia of Philosophy.

Freedom of the Press

Kyu Ho Youm, University of Oregon, Eugene, OR, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article defines press freedom-related concepts and theories. In addition, it examines various key issues relating to freedom of the press in the United States and abroad. Among the topics covered are censorship, international law on censorship, self-censorship, right of reply, press subsidies, licensing of journalists, defamation law, criminal defamation, transnational defamation litigation, Internet defamation, preventive counseling, freedom of information, free press vs fair trial, journalist's privilege, and US press freedom in the global twenty-first century.

No press system is absolutely free from various social, political, economic, or legal controls of its society. Most discussions of freedom of the press center on the relationship of government and the press, and here the basic question is not whether government controls the press but how and to what extent. Freedom of press refers to negative freedom – freedom from governmental constraint, or it may refer to positive freedom – the capacity to do or achieve certain ends. In liberal rule-of-law democracies, most issues concerning press freedom have revolved around negative freedom understood as the independence of the news media from governmental restraint. Hence, private censorship by the corporate news media has rarely been a legal issue. But it has taken a heightened urgency. In September 2012, for example, Google's handling of the controversial anti-Islam video 'The Innocence of Muslims' in the United States and abroad showcased the American Internet company's global role in arbitrating free speech issues in the twenty-first century (Benesch and MacKinnon, 2012).

The Term 'Press'

The 'press' in 'freedom of the press' is elusive. The general public considers the word press as the conventional institutional news media, including newspapers, magazines, and radio and television broadcasters. However, the popular image of the term press is too limiting to reflect the press in general, especially in the age of 'citizen journalism,' a new form of journalism that has no connection with the institutional media. Regardless, the US Supreme Court noted in the late 1930s: The press goes beyond mainstream newspapers and magazines to 'every sort of publication which accords a vehicle of information and opinion' (*Lovell v. Griffin*, 1938). Can the term press be distinguished from 'speech' in 'freedom of speech'? Thus far, the answer to the press–speech dichotomy question is 'no' in US law. The now well-established US rejection of the distinction between press and nonpress communicators presents a striking contrast to the clear-cut statutory recognition of the press in continental Europe and some Asian countries. It should be emphasized, however, that the presence (or absence) of a direct press law does not necessarily ensure a privileged (or unprivileged) press in a country.

The breathtaking technological advances are challenging the traditional concept of the press. In the Internet era, more people say, 'We're *all* journalists' (Gant, 2007). The transformative communication technology belies the structure-centric notion of the press. The function and practice of the press in information gathering and dissemination should be more relevant than the institutional elements of the press. The Web 2.0-based 'citizen journalism' is a case in point. It is changing traditional journalism in so many ways, including 'crowdsourcing' and intended audience (Friedland and Kim, 2009). The influence of the blogosphere as part of citizen journalism is growing, and press freedom theories need to be refined or developed to explain why.

Theories of Press Freedom

In the sense of 'a reasoned explanation of why particular actions lead to certain outcomes' (Christians et al., 2009), various press freedom theories help to answer why press freedom exists. But "one's views on *why* [freedom of speech and the press] exists will tend to influence heavily one's views on *what* it means and *how* it should be implemented" (Smolla, 2014a).

The key values sought by society in protecting freedom of expression are that freedom of expression assures individual self-fulfillment; helps to discover the truth; secures participation by citizens in social, political, and economic decision making; and promotes a stable change in society (Emerson, 1970). The press is protected as the surrogate of the public when performing its informing and checking functions as a Fourth Branch supplementing the executive, legislative, and judicial branches of government.

The foremost value of press freedom comes from the premise that a press free from government control is essential to a political system based on the will of the people. The 'checking value' of a free press posits that the press should be guaranteed freedom to report abuses of power by public officials (Blasi, 1977). This compels a logical conclusion that the government should not prohibit the people from receiving news critical of the government.

Hence, when journalists publish classified information that whistleblowers have obtained illegally, they are not criminally prosecuted in American law. Under the source-reporter double standard in the United States, former National Security Agency (NSA) contractor Edward Snowden, who obviously violated laws in 2013 in leaking to the reporters of the *Washington Post* and other newspapers classified documents about the NSA's surveillance programs, is not protected by the First Amendment to the US Constitution on freedom of the press. But the

reporters who received Snowden's leaks for news reporting as their source are protected insofar as they did nothing illegal in accessing the classified NSA documents from him.

The checking value theory on press freedom has not been explicitly embraced by the US Supreme Court, especially in the reporter's right to refuse to identify news sources. Nonetheless, the European Court of Human Rights (ECtHR) expounded on freedom of expression as an 'essential foundation of a democratic society' and more specifically on the 'safeguards that a free press needs in serving its crucial role as a watchdog' (*Goodwin v. United Kingdom*, 1996).

The libertarian concept of press freedom most fully developed and most ardently supported in the United States focuses exclusively on negative freedom. Since the 1940s, it has been challenged by views that emphasize 'freedom for' as much as 'freedom from.' Concerned about the inexorably growing influence of the mass media on society, so-called social responsibility theories of the press have emerged as a significant counterpoise to the laissez-faire libertarian view. The increasing significance of broadcasting regulated by government 'in the public interest' from early on has no doubt contributed to this emphasis.

The criticism of a libertarian concept of press freedom stems from an innovative proposition that freedom of the press carries little meaningful value unless it is supported by an independence of editorial decision making from both internal and external pressures. In this context, 'journalistic freedom' also known as 'internal freedom' is distinguished from press freedom. Journalistic freedom is concerned with a relationship between the working journalists of a news medium and the managing executives and editors of the news medium. It is an individualized, not institutional, type of press freedom.

The News International phone-hacking scandal of the 2000s in England that centered on the now defunct *News of the World* illustrated the egregious lack of journalistic freedom *within* the news media organization. The UK newspaper's phone hacking, checkbook journalism, and other types of reporting malpractice had more to do with commercial abuse of press freedom as an institution than with its supposed pursuit of the public interest as a trusted watchdog. The scandal led to the 'Leveson Inquiry' (chaired by Lord Justice Leveson) in 2011 into the ethics and practices of the UK press and its future regulation. The resulting Leveson Report of 2012 proposed that the existing self-regulatory body, Press Complaints Commission, be replaced with a more independent body that would be empowered with a range of sanctions. The report has generated a great deal of controversy in and outside the UK press. But its value is associated with its various press-related proposals and also with its 'wealth of information' on the ethics, practices, self-regulation, and structure of the British news media industry (Special issue: Media law after Leveson, 2013).

As in the United Kingdom, press freedom in the United States is to a large degree guaranteed and practiced institutionally, but journalistic freedom is rarely accepted. By contrast, several free press countries in Europe and Asia try to accommodate the conflict between the institutional press freedom and the internal journalistic freedom. A case in point is South Korea, a formerly authoritarian polity but now a vociferous democracy. To ensure separation between the editorial department and management, the *Hankyoreh Shinmun* (One Nation Newspaper) and several Korean newspapers have reporters directly involved in deciding on the overall editorial policy of their newspapers with managing editors elected by reporters or selected from those recommended by them.

Censorship of the Press

In connection with press freedom, both institutional and journalistic censorship is an unending concern to the print and broadcast media around the world. Censorship, particularly if it is exercised by an administrative official with few procedural constraints, is fraught with so many vices against the press. Thus, judicial censorship via injunctions is viewed as 'more tolerable' than administrative censorship (Barendt, 2005).

Censorship is intertwined with power so inextricably that it is referred to as the 'handmaiden of power.' Censorship in the sense of prior restraint on the press is more effective than subsequent punishment because of its broader scope and lack of countervailing mechanisms, such as public appraisal of its rationale. It also has a propensity to decide against publication because "[t]he function of the censor is to censor. He has a professional interest in finding things to suppress" (Emerson, 1955).

Censorship can be total when all the media are systematically controlled. In the United Kingdom, courts may issue *contra mundum* orders against the news media not to publish the information covered. The orders bind any journalist or media organization that is aware of them. Also, 'super-injunctions' are granted by English courts to bar news organizations from reporting the very fact that they are subject to the judicial gag orders.

Administrative vs Judicial Censorship

The legality and scope of press censorship is sometimes examined by focusing on who primarily dictates the censorship mechanism. Is censorship entirely in the hands of executive officers with no judicial oversight? On this issue, the United States differs from Japan. The US law does not allow even court-issued prior restraint under the First Amendment. In Japan, however, the anticensorship principle does not extend to judicial restrictions because censorship is limited to administrative imposition of prior restraint on the press.

International Law on Prior Censorship

Prepublication censorship as an *administrative* action is incompatible with the free press principle in a democracy. This is explicitly recognized by the American Convention on Human Rights:

> The exercise of the right [to freedom of thought and expression] … shall not be subject to *prior censorship* but shall be subject to subsequent imposition of liability which shall be expressly established by law to the extent necessary to ensure:
> 1. respect for the rights or reputations of others; or
> 2. the protection of national security, public order, or public health or morals.

Self-Censorship of the Press

Self-censorship is viewed as a widespread and insidious problem among reporters and editors. Journalists resort to self-censorship when they decide to avoid legitimate stories, not because of their professional judgments, but on nonprofessional grounds. In some cases, self-censorship can be intimately intertwined with state censorship. This is the case in Israel: "When acting as concerned members of society, editors and journalists are of one mind with the censors concerning the prohibition of public exposure of issues considered crucial to the physical welfare of Israel or to its ideological raison d'etre (even though the public may not share that view)" (Liebes, 1997).

Intrusion of commercial interests into editorial decision making is another major cause of self-censorship. "Market forces are seen as the primary reason why worthwhile stories are not pursued and this factor is especially prevalent in the broadcast arena" reports a recent survey of American journalists on self-censorship (Kohut, 2000).

Self-censorship is closely related to threat of lawsuit for defamation, invasion of privacy, breach of confidentiality agreements, and related issues. Nearly 30 state anti-SLAPP (Strategic Lawsuit Against Public Participation) laws in the United States (no federal law) aim to prevent individuals or news media from being intimidated into self-censorship by litigious, well-financed business people or organizations. Most SLAPP lawsuits are meritless and lose, but their chilling effect on the news reporting on controversial issues is considerable. When sued, media defendants have to reallocate their limited money, time, and resources in defending the SLAPP lawsuits.

Right of Reply

Requiring news media outlets to allow the subjects of their reporting to reply or correct their stories does not necessarily transgress the free press principle if the obligations are not sweepingly broad or intrusive of the media's editorial independence. In the United States, however, the mandatory right of reply statutes governing the news media as a whole and the print press in particular are seen as violations of freedom of the press.

The US Supreme Court in 1974 struck down a right of reply statute on the ground that it would infringe upon the editorial rights of the press: "A newspaper is more than a passive receptacle or conduit for news comment and advertising. The choice of material to go into a newspaper and the decisions made as to limitations on the size and content of the paper and treatment of public issues and public officials – whether fair or unfair – constitute the exercise of editorial control and judgment. It has yet to be demonstrated how governmental regulation of this crucial process can be exercised consistent with First Amendment guarantees of a free press as they have evolved to this time" (*Miami Herald Publishing Co. v. Tornillo*, 1974).

In foreign and international law, however, the right of reply is increasingly accepted. More than a dozen countries in Europe, including Germany and France, recognize the right of reply. The right of reply in German law is derived from French law. The policy justifications for the right of reply in Germany are not so much to provide the public with access to the media as to protect individuals from false defamation. This differentiates German law from South Korean law on the right of reply. The right of reply in Korea is justified partly because it can contribute to the public's discovering truth and forming diverse public opinions by creating a more open marketplace of ideas.

Especially noteworthy is the international acceptance of the right of reply. The International Convention on the Right of Correction, the European Convention on Human Rights, and the American Convention on Human Rights provide for the right of reply.

The explicit recognition of the right of reply in several civil law countries in Africa, Asia, Europe, and Latin America, and under international law signifies that there are alternative visions of press freedom. This stands in sharp contrast with the US approach to the negative freedom of the press; that is, access to the press is not entirely antithetical to freedom of the press, but it can be viewed as a plausible 'positive' notion of press freedom in ensuring more opportunity for freedom of expression.

Press Subsidies

Amid the ongoing financial difficulties facing the legacy news media in the United States, which follows a minimalist public support model, governmental financial support for media operations and content has been a topic of considerable policy debate in and outside the media industry. Those who are wary of state subsidies for the press are concerned about the government's possible interference with journalistic work. However, their worry is likely misplaced, given the established First Amendment jurisprudence.

Also, evidence from the Scandinavian press subsidies disproves an assumption underlying the adversarial press–state relations in the United States that newspapers receiving direct public subsidies will lose their editorial independence and autonomy from the government. The Swedish subsidy program 'has had little effect on the willingness of subsidized publications to criticize government' (Gothberg, 1983; Gustafsson, 1980).

Licensing of Journalists

Government licensing of journalists is incompatible with freedom of the press as a basic human right. As the Inter-American Court of Human Rights (IACHR) ruled in 1985: "Such a law (on licensing of journalists) would contain restrictions to freedom of expression … and would consequently be in violation of not only the right of each individual to seek and impart information and ideas through any means of his choice but also the right of the public at large to receive information without any interference" (IACHR, 1985).

Defamation Law

Defamation was one of the earliest legal actions available against the press, and it is still the most common legal danger

to the media. Defamation is an occupational hazard for media professionals. About 70% of all American defamation actions are filed against the mass media; at least two-thirds of these actions are brought against newspapers in the United States. Defamation law protects the reputational interests of individuals because it involves the expression of opinions about them by others. It also attempts to compensate for economic and emotional injury and promotes human dignity by providing a civilized forum in which a dispute is settled. Most importantly, however, defamation law acts as a deterrent on the publication of false and injurious speech through the award of damages. In addition, defamation law serves a vital social interest by providing a check on media power. It opens the newsgathering and decision-making process of the media to public scrutiny and accountability.

Criminal Defamation

There is growing evidence globally, whether statutory or decisional, that criminalizing defamatory statements is contrary to a right to freedom of speech and the press. In the United States, however, 15 states and one US territory still have laws on the books that allow the government to prosecute people for criminal defamation. American criminal defamation law, which the US Supreme Court has not repudiated, stands in contrast with foreign law, in which libel is increasingly decriminalized. For example, criminal defamation has been abolished in England, Ireland, Mexico, and other countries. An in-depth empirical study of criminal defamation in the United States has stated: "Criminal libel can be especially useful for people of limited means, those who would have difficulty finding lawyers to pursue civil lawsuits for libel. Criminal libel may also be useful in curbing defamatory gossip on the Internet" (Pritchard, 2009).

Internet Defamation

The 'information superhighway' epitomized by the Internet exerts an enormous impact on defamation law all over the world. This is especially the case with jurisdiction and enforcement of defamation law. While other countries continue to sort out a vast array of jurisdictional and other related issues, the US Congress in 1996 passed a law immunizing Internet service providers (ISPs) from liability for third-party postings of defamatory material.

An oft-cited Internet defamation expert, Matthew Collins, noted the relevance of American law to other countries:

- (P)ersons in other common law countries are defamed by material written or edited in the Untied States, or hosted on the computer system of an Internet intermediary located in the Untied States;
- (S)uccessful claimants in foreign defamation actions seek to enforce their judgments in the Untied States, or where defendants apply to American courts to seek to prevent cases from proceeding in other countries or to have foreign defamation judgments declared unenforceable in the United States;
- (T)he American cases provide useful examples to illustrate the operation of the principles of defamation law discussed (in *The Law of Defamation and the Internet*);
- (T)he extent to which principles derived from the American Internet defamation cases might be capable of being adapted or applied in analogous circumstances in other common law countries cannot be assessed without an appreciation of the uniquely American approach toward, in particular, the elements of the cause of action for defamation, intermediary liability, jurisdiction over foreign publishers, and choice of law in multi-State defamation cases (Collins, 2010).

Transnational Defamation Litigation

As off-shore litigation against US and other transnational news media is growing, the globalizing media confronts a new type of defamation threat. The High Court of Australia's 2002 decision in *Dow Jones & Co. v. Gutnick* (2004) is a good illustration. Australian businessman Joseph Gutnick sued the American publisher of *Barron's* Online magazine in Australia for defamation. The Australian court held: When a defamatory statement is accessible to and read by ISP subscribers in an Australian state (Victoria in this case), a court of that state has jurisdiction to hear an action for defamation.

American courts have consistently refused to recognize and enforce foreign defamation judgments, holding that the judgments were based on defamation law repugnant to the US Constitution on press freedom. In 2010, in an effort to preempt various press freedom issues arising from 'defamation tourism' abroad, US President Barack Obama signed the SPEECH (Securing the Protection of our Enduring and Established Constitutional Heritage) Act, a federal law that bars the domestic enforcement of foreign defamation judgments against Americans if they are inconsistent with the First Amendment.

The gap between the United States and the rest of the world remains noticeable. But free speech defenses such as the American 'actual malice' (i.e., knowledge of falsity or reckless disregard for the truth) rule for the news media against liability for public defamation are increasingly accepted in other countries. For example, the UK Defamation Act 2013 and the Canadian Supreme Court have adopted 'public interest' and 'responsible communication,' respectively. This new media-friendly defamation defense precludes liability for responsible news reporting on a matter of public interest, although it resulted in false defamatory stories.

Privacy

Privacy never stops evolving as a dynamic value. It reflects keenly on how a society balances the 'right to be let alone' with its competing needs for information. Thus, it varies from society to society, which makes it challenging to generalize what privacy is as an individual right and why it should be so cross-culturally. Strikingly illustrative is the difference between Europe and the United States in weighing informational privacy against the freedom to publish.

In the United States, whose privacy tort law is one of the most highly developed and influential in the world, four types of invasion of privacy are recognized: intrusion, public

disclosure of private facts, false light, and appropriation. Especially, the 'private facts' tort is recognized widely, but its practical impact is severely limited by the First Amendment protection of the media's right to publish truth. News media are not liable for publicizing private facts if the disclosure is not highly offensive to a reasonable person and it is not a matter of public concern. American judges rarely second guess the media professionals' news judgments. Few American media organizations lose in private facts cases, although their stories have little to do with political expression on public issues.

The ECtHR could not be more different from American courts on the private facts tort. In 2004, for example, the ECtHR, ruling on a privacy complaint by Princess Caroline of Monaco, held:

> [A] fundamental distinction needs to be made between reporting facts – even controversial ones – capable of contributing to a debate in a democratic society relating to politicians in the exercise of their functions, for example, and reporting details of the private life of an individual who … as in this case, does not exercise official functions. While in the former case the press exercises its vital role of "watchdog" in a democracy by contributing to "impart[ing] information and ideas on matters of public interest … it does not do so in the latter case.
>
> (*Von Hannover v. Germany*, 2004)

When balancing freedom of the press with the right of privacy, the European Court has defined press freedom by focusing on the public's interest vs the public interest.

Access to Information and Free Press vs Fair Trial

The press's role of serving as a 'public watchdog' on the government and of imparting information on matters of public interest is difficult, if not impossible, to fulfill without access to public records. Without some protection for seeking out the news freedom of the press could be a farce. Now freedom of information (FOI) is accepted more widely around the world than ever. As of March 2014, a total of 100 countries recognized FOI as a right for the press and others. This is a profound development in freedom of the press as a positive concept. In this connection, it should come as no surprise that the Scandinavian countries that embrace a 'proactive' concept of press freedom have the most comprehensive FOI laws. For example, Sweden, whose government information statute is considered the oldest in the world, has adopted the policy on access to official records as a fundamental foundation of its democracy since 1766.

More importantly, the ECtHR in *Hungarian Civil Liberties Union v. Hungary* (2009) and the IACHR in *Claude Reyes v. Chile* (2006) have ruled that access to government information is a human right. Also, the Council of Europe has adopted an FOI treaty, the first of its kind in the world, and the United Nations Human Rights Committee has confirmed FOI as a fundamental human right.

Traditionally, freedom of the press in the United States has been confined to the right of news reporters and editors to disseminate any information in their possession, which does not necessarily extend to the legal right for the media to gather information. But newsgathering has now emerged as a vital issue for most journalists and news media organizations. In the United States, how to deal with access to government records, government agency meetings, and the like constitutes by far the largest body of mass media law.

Freedom of the press is often in conflict with the accused's right to a fair trial because news media wish to report on court proceedings as an important source of news information, while trials can be jeopardized by irresponsible news reporting. Meanwhile, cameras in the courtroom as a media right remain a contentious issue around the world. The US Supreme Court strictly prohibits television cameras. On the other hand, the Canadian Supreme Court has extensive experience with its hearings being broadcast. The Supreme Court of the United Kingdom is more camera-friendly than its US counterpart. The UK Supreme Court has allowed its hearings to be broadcast since its opening in October 2009 to replace the House of Lords. Brazil is one of the most open judicial systems when it comes to the broadcasting media's access to its highest court.

Journalist's Privilege to Shield News Sources

Equally important is the issue of whether journalists should be privileged to protect their confidential sources from forced disclosure. In the United States, a journalist's privilege is not recognized as such under the First Amendment. Indeed, it lags far behind Argentina, Japan, Norway, and Sweden, where journalistic privilege is the rule, not the exception.

The journalist's privilege has been read by the ECtHR into freedom of expression under the European Convention on Human Rights (*Goodwin v. United Kingdom*, 1996). More recently, the International Criminal Tribunal for the former Yugoslavia (ICTY) accepted the journalist's privilege as a right in international law. In endorsing the journalist's privilege, the ECtHR and the ICTY have embraced the watchdog press concept and the free flow of information notion, respectively.

Preventive Counseling for News Media

Because freedom of the press is litigated in increasing frequency in the United States, England, and other countries, media organizations involve lawyers in prebroadcasting and prepublication reviews. Lawyers in newsrooms have now become a fact of life for many, if not all, reporters and editors in dealing with a host of legal problems ranging from defamation to access to information. Some journalists are concerned about the chilling effect of lawyers' presence in and outside the American newsroom. Nonetheless, the journalist–lawyer relationship is not necessarily harmful to freedom of the press insofar as lawyers are sufficiently 'sensitive to the sometimes murky line between legal advice and journalistic or artistic judgment' (Smolla, 2014b).

No less important is the fact that journalists as the integral part of the media system are more often responsible than lawyers for determining what to broadcast or publish or what not to broadcast or publish. Likewise, a thoughtful UK study of the impact of defamation law on the media both in England

and in Scotland concluded "(A) solicitor's advice is just that and … on occasion editors do decide to ignore it" (Barendt et al., 1997).

Freedom of the US Press in the Globalizing Era

Freedom of speech and the press has been the international buzzword since the end of the Cold War in the late 1980s. The positive concept of press freedom has crystallized the undisputed paradigm throughout the world, and it inspires the standard for quality journalism. The growing impact of the modified libertarian press theory with an emphasis on the state's affirmative role in press freedom on Asia, former eastern Europe, and Latin Africa is credited partly to the American experiment of 200-plus years with a free press. At the end of his highly acclaimed book of 2010, *Uninhibited Robust and Wide-Open: A Free Press for a New Century*, meanwhile, Columbia University President Lee Bollinger, a leading First Amendment scholar, noted:

> This book has looked at the problem of what must be done for a global free press from a U.S. perspective. But we need to do this from the vantage point of other countries as well, especially those that generally share the same philosophy about the importance of a free and independent press. A collective effort toward the same end is a necessity.
> (Bollinger, 2010)

See also: Censorship in Research and Scholarship; Control, Social; Freedom: Political; Media and History; Political Communication.

Bibliography

Barendt, E., Lustgarten, L., Norrie, K., Stephenson, H., 1997. Defamation and the Media: The Chilling Effect. Clarendon Press, Oxford UK.

Barendt, E., 2005. Freedom of Speech, second ed. Oxford University Press, Oxford.

Benesch, S., MacKinnon, R., October 5, 2012. The innocence of YouTube: it's time for Internet giants to explain when censorship is and isn't OK. Foreign Policy. http://www.foreignpolicy.com/articles/2012/10/05/the_innocence_of_youtube?page=0,0.

Blasi, V., 1977. The checking value in First Amendment Theory. American Bar Foundation Research Journal 2, 522–649.

Bollinger, L., 2010. Uninhibited Robust and Wide-Open: A Free Press for a New Century. Oxford University Press, New York.

Christians, C.G., Glasser, T.L., McQuail, D., White, R.H., 2009. Normative Theories of the Media: Journalism in Democratic Societies. University of Illinois Press, Urbana, IL.

Claude Reyes v. Chile, 2006. Inter-American Court of Human Rights Judgment of 19 September 2006. http://www.corteidh.or.cr/docs/casos/articulos/seriec 151 ing.pdf.

Collins, M., 2010. The Law of Defamation and the Internet, third ed. Oxford University Press, Oxford, UK.

Dow Jones & Co. v. Gutnick, H.C.A. 59 (2004).

Emerson, T.I., 1955. The doctrine of prior restraint. Law & Contemporary Problems 20, 648–671.

Emerson, T.I., 1970. The System of Freedom of Expression. Vintage Books, New York.

Friedland, L., Kim, N., 2009. Citizen journalism. In: Sterling, C. (Ed.), Encyclopedia of Journalism. Sage, Los Angeles.

Gant, S., 2007. We're All Journalists Now: The Transformation of the Press and Reshaping of the Law in the Internet Age. Free Press, New York.

Glasser, C.J., 2013. International Defamation & Privacy Handbook, third ed. Bloomberg Press, New York.

Goodwin v. United Kingdom, 22 E.H.R.R. 123 (1996).

Gothberg, J.A., 1983. Newspaper subsidies in Sweden pose no dangers. Journalism Quarterly 60, 629–634.

Gustafsson, K., 1980. The press subsidies of Sweden: a decade of experiment. In: Smith, A. (Ed.), Newspapers and Democracy. MIT Press, Cambridge MA.

Hungarian Civil Liberties Union v. Hungary, 2009. European Court of Human Judgment of 19 April 2000. http://hudoc.echr.coe.int/sites/eng/pages/search.aspx#{%22dmdocnumber%22:[%22849278%22],%22itemid%22:[%22001–92171%22]}.

Inter-American Court of Human Rights (IACHR). San Jose Advisory Opinion of 13 November 1985. No. OC-5 85 – Compulsory Membership in an Association Prescribed by Law for the Practice of Journalism (Arts. 13 and 29 American Convention on Human rights).

Kohut, A., May/June 2000. Self-censorship: counting the ways. Columbia Journalism Review 42.

Liebes, T., 1997. Reporting the Arab-Israeli Conflict: How Hegemony Works. Routledge, London.

Lovell v. Griffith, 303 U.S. 444 (1938).

Merrill, J.C., 1989. The Dialectic in Journalism. Louisiana State University Press, Baton Rouge, LA.

Miami Herald Publishing Co. v. Tornillo, 418 U.S. 241 (1974).

Mullis, A., Doley, C., 2010. Carter-Ruck on Defamation and Slander, sixth ed. LexisNexi, London.

Nielsen, R., Linnebank, G., 2011. Public Support for the Media: A Six-Country Overview of Direct and Indirect Subsidies. Reuters Institute for the Study of Journalism. University of Oxford, Oxford.

Price, M.E., Krug, P., 2000. The Enabling Environment for Free and Independent Media. Programme in Comparative Media Law & Policy. Oxford University, Oxford UK.

Pritchard, D., 2009. Rethinking criminal libel: an empirical study. Communication Law & Policy 14 (1), 303–339.

Smolla, R.A., 2014a. Smolla and Nimmer on Freedom of Speech. Thomson Reuters/West, St. Paul, MN.

Smolla, R.A., 2014b. Law of Defamation, second ed. Thomson West, St. Paul.

Special issue: media law after Leveson, 2013. Journal of Media Law 5 (2), 167–296.

Von Hannover v. Germany, E.M.C.R., 39 (2004).

Warrby, M., Moreham, N., Christie, I., 2011. The Law of Privacy and the Media, second ed. Oxford University Press, Oxford.

Youm, K.H., 2006. International and comparative law on the journalist's privilege: the Randal case as a lesson from the American press. Journal of International Media & Entertainment Law 1, 1–56.

Youm, K.H., 2008. The right of reply and freedom of the press: an international and comparative perspective. George Washington Law Review 76, 1017–1064.

Youm, K.H., 2012. Cameras in the courtroom in the twenty-first century: the U.S. Supreme Court learning from abroad? Brigham Young University Law Review, 1989–2012.

Freedom: Political

Timothy O'Hagan, University of East Anglia, Norwich, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Political liberty or freedom is an essentially contested concept, as Berlin emphasized. He distinguished positive liberty (the 'liberty of the ancients,' exercised in political participation, as well as in Kantian moral autonomy) from negative liberty (the 'liberty of the moderns,' the 'right to be let alone' in one's private life). That distinction has been refined, criticized, rejected, and resurrected in the subsequent debate. In the light of it, five rival approaches are distinguished within the English language literature. These are (1) welfare liberalism (Rawls and Dworkin), (2) libertarianism (Nozick), (3) freedom, economics, and power (the neo-Marxism of Cohen, the post-Marxism of van Parijs, and the neo-Aristotelianism of Sen), (4) freedom as attaining autonomy through community (Raz), and (5) civic republicanism, i.e., freedom as involving political activity (the instrumental version of Skinner; the constitutive version of Pettit).

Liberty, Equality, Fraternity, and Two Concepts of Liberty

Discussions of political freedom since the mid-twentieth century have rotated around two separate but related themes: on the one hand the slogan of the French Revolution 'liberty, equality, fraternity' and on the other hand Benjamin Constant's distinction between 'the liberty of the ancients and the liberty of the moderns,' transformed and transmitted to the English-speaking world by Isaiah Berlin in *Two Concepts of Liberty*.

Liberty, Equality, and Fraternity

Since the French and American Revolution modern states have been committed, at least formally, to the ideals of liberty, equality, and fraternity. However, from the start it became evident that those three aspirations are by no means automatically compatible. Attempts to balance them have taken different forms. Libertarian advocates of the *laissez-faire* market have argued that any attempt to promote greater equality of resources involves the illegitimate restriction of the right of individuals freely to pursue their own interests, whatever they may be. For some Marxists formal 'bourgeois' liberties are valueless until the inequality between social classes is abolished. Many others, who may be labeled social democrats in Europe or liberals in the United States, have sought to show that the three aspirations, far from being incompatible, are linked by ties of mutual entailment.

Two Concepts of Liberty

The most influential contribution to the topic of political freedom in the English language literature of the twentieth century remains Isaiah Berlin's *Two Concepts of Liberty*, delivered as an inaugural lecture at Oxford University in 1958. In it he distinguished 'positive liberty' (PL), exercised in political participation, as well as in Kantian moral autonomy, from 'negative liberty' (NL), the 'right to be let alone' in one's private life. The author made some changes to the later edition, but the core distinction remained intact. Berlin's lecture is an unsystematic mixture of different elements: conceptual analysis, political engagement in the postwar antitotalitarian struggle, and a rich, if eclectic, essay in intellectual history. But for all its shortcomings, it made the opposition between PL and NL central to much subsequent discussion.

Berlin's political task was to unmask the perversion of language by Stalinist ideologues, who presented the worst forms of oppression as 'real (socialist)' freedom, as opposed to 'mere bourgeois (formal)' freedom. Berlin aligned the former with 'positive' liberty and the latter with 'negative' liberty and reasserted the value of the latter against its critics.

Historically, Berlin identified NL with liberalism, the tradition that defended the individual against 'interference,' particularly when exercised by the state. However, the idea of 'interference' is hard to tie down. Berlin himself was ambivalent as to whether the interference must be deliberate or not, but he was adamant that an inability caused by the social structure (e.g., the inability of the worker to control the means of production in traditional Marxist theory) is not a case of unfreedom. He agreed that the 'classical English political theorists' were 'prepared to curtail freedom in the interests of other values and indeed of freedom itself.' They held that without law, 'freedom for the pike is death for the minnow.' A natural interpretation of that insight is that the law positively creates freedom for the minnow, allowing it to do things it would otherwise be unable to do. In the words of Locke (a quintessential negative libertarian for Berlin): "the end of Law is not to abolish or restrain, but to preserve and enlarge freedom." Berlin could agree that the pike's freedom is restricted for the sake of the minnow's freedom ('curtailed in the interests of liberty itself'), but his polemical drive was against partisans of PL who would maintain that the pike's freedom too is (really) enhanced by the restriction.

Another ambivalence within NL concerns the connection between freedom and wishes. Berlin began by tying the two together directly, defining oppression as "the part I believe to be played by other human beings … in frustrating my wishes." Later he rejected the definition of freedom as "the ability to do what one wants" because it is compatible with a situation in which all one's wants are restricted by brainwashing or the removal of desirable objects. At the heart of NL lies 'the right to be let alone,' but that right turns out to be remarkably elusive.

If NL is hard to tie down, PL is even more so. PL contains both (1) political elements (PL-pol) and (2) moral elements

(PL-mor). (1) is essentially the freedom to participate in the political process; (2) includes the ability to be the autonomous master of one's own drives, which is associated with the romantic notion of the divided self. Rousseau says that both (1) and (2) are the outcome of entering civil society, but he is careful not to identify them. When Constant (1988) introduced the distinction between the 'liberty of the ancients' and the 'liberty of the moderns,' he was concerned with (1). His polemical claim was that the 'liberty of the ancients' nowadays meant the unbridled exercise of popular sovereignty (PL-pol) to the detriment of NL. For Berlin, the danger of modern totalitarian ideology is that it harnesses (1) to (2). It suggests that, through laws endorsed by 'the people,' irrational empirical selves should be forced to coincide with rational ends (their 'real selves').

In response to the two themes identified earlier, five broad approaches to political freedom can now be distinguished. In each of these, the three fundamental values of liberty, equality, and fraternity are addressed. Writers inspired by Berlin focus on incompatibilities between those values. Others seek to show that when they are more clearly understood, it will be seen either that they are compatible or that the very idea of balancing one against the other must be abandoned. In their analyses, the terminology of PL and NL frequently recurs, but the distinction made above between dimensions (1) and (2) of PL frequently is ignored.

The approaches to be considered are (1) welfare liberalism; (2) libertarianism; (3) freedom, economics, and power; (4) freedom and community; and (5) civic republicanism.

Welfare Liberalism

For theorists of welfare liberalism, political freedom is exercised in a context of representative democracy constrained by constitutional checks to protect individual rights and by a fluctuating sense of obligation to redress the inequalities thrown up by the operation of the market. The two outstanding proponents of this approach are John Rawls and Ronald Dworkin.

Rawls

Rawls's *A Theory of Justice* (1971) is the most comprehensive, systematic work of modern political philosophy. In it the author expounds the moral basis of pluralist welfarism, combining the three elements of liberty, equality, and fraternity into a coherent whole.

At the heart of the book lie the two principles of justice. The first principle states that "Each person is to have an equal right to the most extensive total system of basic liberties compatible with a similar system of liberty for all." 'Basic liberties' are specified as "roughly speaking, political liberty (the right to vote and to be eligible for public office) together with freedom of speech and assembly; liberty of conscience and freedom of thought; freedom of the person along with the right to hold (personal) property; and freedom from arbitrary arrest and seizure as defined by the concept of the rule of law." These are what Marxists called 'bourgeois liberties.' The list is not equivalent to NL-pol, since they include certain 'positive' political elements, even if the 'negative' liberties of privacy may be more important than 'positive' ones to modern citizens.

In subsequent work, Rawls (1993) has responded to Hart's criticism that the idea of 'maximizing freedom' is incoherent. He has rewritten the first principle to read: "Each person has a claim to a fully adequate scheme of basic rights and liberties, which scheme is compatible with the same scheme for all; and in this scheme the equal political liberties, and only those liberties, are to be guaranteed their fair value." He has also added the 'integrity' (physical and psychological security) of the person to the list of basic liberties to be protected by the first principle. However, he has not modified the central claims of *A Theory of Justice* that the modern liberal state must (1) guarantee 'basic liberties' and (2) strive to maximize the welfare of the worst-off members of society by redistributive mechanisms, and that the first principle must not be sacrificed to the second, since 'liberty can be restricted only for the sake of liberty.'

Most of the other important revisions have concerned structure and methodology rather than content. Rawls has anchored the two principles more explicitly in a conception of citizens of a modern pluralist state, as it has developed since the Reformation. They are 'regarded as free and equal persons,' endowed with "two moral powers … namely a capacity for a sense of justice and a capacity for a conception of the good." Such citizens display their freedom in the following ways. They have 'the moral power to have a conception of the good' and are 'capable of revising and changing this conception.' They 'regard themselves as self-authenticating sources of moral claims.' They are capable of taking responsibility for their 'ends.'

The two principles of justice remain unaltered and are still subject to attack from both left and right. Left-wing critics charge that first principle 'bourgeois' liberties are hollow and worthless for the mass of the population. To them Rawls had already responded in *A Theory of Justice* by distinguishing equal liberty from equal worth of liberty, acknowledging that it is the latter that is important, and predicting, perhaps optimistically, that it would be enhanced by the application of the difference principle. Right-wing 'libertarian' critics reject all redistributive mechanisms, because they infringe the individual's unrestricted right to accumulate property. Rawls' response to them is also contained in *A Theory of Justice*, namely, that individuals have a basic right to own personal property, but that right does not entail any particular property system, whether capitalist or socialist. In each particular case the choice of system will be determined by its ability to maximize the welfare of the worst-off.

Dworkin

Dworkin's political philosophy is an attempt to 'fine-tune' both the methodology and the political recommendations of Rawls's welfare liberalism. Methodologically, Dworkin denies that liberalism must balance liberty against equality. That image wrongly presupposes that liberty can be quantified, so that 'fundamental' liberties weigh more heavily than trivial ones. Instead, 'the fundamental liberties are important because we value something else that they protect.' That 'something else' is, unexpectedly, a form of equality, the principle that a government should treat all its citizens with 'equal concern and respect.' The 'fundamental' liberties, then, are entailed by the equality principle itself, according to which 'the government (should) treat all its citizens as free, or as independent, or with equal dignity.' Dworkin accepts the Rawlsian requirement that "political decisions must be … independent of any particular conception of the good life …." Thus governments are bound by the negative requirement to be neutral between

lifestyles, and also by the positive requirement to 'respect' the 'dignity and independence' of its citizens.

In terms of political recommendations, Dworkin proceeds from a tension at the heart of Rawls's two principles. While the first principle defends the conditions of basic freedoms, and addresses people as autonomous agents, the second principle might undermine people's motivation to exercise their freedom to 'better themselves,' addressing them as passive recipients of benefits. In response to this tension, Dworkin introduces a key modification. He follows Rawls in compensating people for 'irrelevant differences, like differences in talent.' However, he argues that people should be 'responsible for the costs of their choices.' The system, in short, must be both 'ambition sensitive' and 'endowment insensitive.' However, Dworkin's modification may have introduced a new tension of its own, since, however finely tuned it is, its two requirements must work against one another: "the more we try to make the distribution sensitive to people's ambitions, the more likely it is that some people disadvantaged by circumstances will be undeservedly penalized, and vice versa" (Guest, 1997). Everything now hinges on the possibility of making a realistic distinction between ambition and endowment, between 'personality' and 'circumstances.' The fact that this is so difficult suggests that the tension between freedom and equality may still be unresolved.

Libertarianism, Freedom as the Unrestricted Right to Own Private Property: Nozick

Nozick's *Anarchy, State and Utopia* (1974) was a polemical response to Rawls' *A Theory of Justice*. At the time of its publication, there was a backlash against welfarism and state intervention in the economy. That backlash was driven by the economic theory that managed markets are inefficient, since redistribution through taxation takes away incentives. However, although Nozick's book was congenial to that theory, his philosophical position is independent of it. Nozick would defend his position on the grounds that it was the correct one, even if it were inefficient. He would in fact reject appeals to economic efficiency as improperly utilitarian.

Although labeled a 'libertarian,' Nozick's starting point is not the idea of 'liberty' or 'freedom,' but rather the Lockean one of 'self-ownership.' This has polemical value because it allows Nozick to proceed against Rawls' (and other welfarists') programs of redistribution by appealing to a hatred of slavery shared by all across the political spectrum since the 'bourgeois revolutions,' and suggesting, by a rhetorical sleight of hand, that "taxation of earnings from labor is on a par with forced labor." The slide has some source in Locke's 'labor theory of property,' the idea that one has a property right to something if one has 'mixed one's labor with it.' However, Locke did not suggest that that right is absolute in the same sense that one's right to bodily integrity is.

The sleight of hand at the heart of libertarianism is this. We all agree that if A has the right to make use of B's body, of which B is the rightful owner, without B's consent, then B is A's slave. By parity of reason, it is suggested, if A has the right to make use of B's property, of which B is the rightful owner, without B's consent, by taxing B, then B is A's slave. The sleight of hand depends on accepting the extension of self-ownership from ownership of one's own body to ownership of whatever one has acquired legally, an extension that is not evidently justified. Nonetheless, libertarians may be able to deploy a 'slippery slope' argument against welfarism without making this particular sleight of hand. Rawls's two principles of justice are intended 'to balance the twin demands of (1) respecting choices and (2) rectifying circumstances.' Rawls intends (1) to be applied to individuals, (2) to groups. However, according to the libertarian, there is nothing within Rawls's theory to prevent the radical welfarist from extending rectificatory treatment to individuals too. How is one to define what intervention is legitimate in order to rectify circumstances? If the line is to be drawn around the body, it would seem to mean that any nonbodily interference is at least in principle legitimate in order to rectify circumstances. That sounds sinister even to those who are not extreme 'libertarians.'

Freedom, Economics, and Power

In the past, egalitarians, Marxists, and non-Marxists alike derided 'bourgeois freedoms' for masking from the workers their real enslavement. The modern egalitarians considered in this section all take seriously the demand for real equality of freedom. However, unlike their predecessors, they all also take seriously the importance of the 'basic' freedoms that lie at the heart of liberalism. The three positions considered in this section are linked by the insight that freedom and equality are linked intimately, and that the conditions of equality and inequality are zero-sum relations of economic power. In each of the three, real freedom is 'positive' in the sense that it requires more than just the 'right to be let alone.' On the other hand, it is not the 'liberty of the ancients,' since it recognizes the importance of private life in the modern world.

Neo-Marxism: Cohen

In the most sophisticated version of neo-Marxism, that of Cohen, "there is … a conflict between social equality and the liberty of some people," but that should not stand in the way of "the pursuit of social equality, since a humane concern for liberty must first of all direct itself to the condition of those who enjoy hardly any of it."

Neo-Marxists thus agree with liberals that freedom of the individual is the most important political value, and that modern capitalism delivers it to all members of society, including proletarians, in several important ways. First, all members of a capitalist society (worker and capitalist alike) are legally free. Second, any particular worker is always legally free, and often economically free, to leave the proletariat and become a petty bourgeois or even a capitalist. However, within a capitalist society, 'although most proletarians are free to escape the proletariat, indeed even if all are, the proletariat is an imprisoned class.' Third, capitalism has delivered 'important freedoms beyond that of buying and selling.' These include 'freedom of speech, assembly, worship, publication, movement, and political participation.' The neo-Marxist Cohen is agnostic about 'how accidental the connection between capitalism and those freedoms has been and is,' but is committed to the view that only 'freedom to buy and sell belongs to

capitalism's inmost nature.' However, Marxists must still acknowledge that 'bourgeois freedoms' really are freedoms: "… when socialists suggest that there is no real liberty under capitalism, or that socialism promises liberty of a higher and unprecedented kind … their line is theoretically incorrect and politically disastrous. For liberty under capitalism is, where it exists, just that, liberty; and if socialism will not give us plenty of it, we shall rightly be disappointed."

In recent writings, Cohen (1995) has begun to investigate the idea of real freedom, understood as 'autonomy, the circumstances of genuine control over one's own life.' With this turn, neo-Marxism returns to the old site of conflict, since we must 'ask what kind and degree of control over external things a person must have to enjoy autonomy, and then to ask whether such control is compatible with socialist equality.' Cohen's project now is to translate the idea of real freedom into terms that are applicable to the real world, and, where possible, quantifiable. His model of 'equal access to advantage' now stands alongside Van Parijs's (1995) and Sen's (1992) models. It thus forms part of a rich developing research program. Whether it should still be called Marxist is another question.

Post-Marxism: van Parijs

In the post-Marxist program of van Parijs, the normative goal of equalizing access to freedom is maintained. At the same time, the Marxian theory of exploitation is abandoned, and the capitalist market is now seen as the only vehicle that can deliver 'real freedom for all' in practice. The goal is to 'empower' the greatest number of people. The method is to guarantee them an unconditional 'basic income' that will prevent personal dependence and allow individuals to make meaningful choices between life styles. The 'basic income' project has now entered the real political agenda in Western Europe.

The goal of the post-Marxists is to reconcile freedom and equality by the principle to 'maximization of an unconditional income, subject to formal freedom being protected and to no one's endowment being unanimously found worse than someone else's.' That would guarantee the widest, and fairest, distribution of 'real' freedom that could also be delivered by a 'sustainable' economic system. Van Parijs maintains that that system would be a highly regulated, interventionist form of capitalism. His reasons are fourfold. First, it is (probably) more efficient than socialism. Second, it can (probably) guarantee a higher basic income than socialism can. Third, it can (certainly) guarantee formal freedoms, whereas socialist régimes have still to prove that they can. Last, and this is a decisive consideration, even if there are reforms of international economic affairs, it will be impossible to promote an 'autarkic' socialist régime in one country without imposing unacceptable restrictions on the movement of its people to other (higher paying) countries.

Neo-Aristotelian Freedom as Capability: Sen

The third egalitarian model combines sophisticated economic theory with a deep philosophical anthropology derived from Aristotle. The leading proponent of this model is Sen.

The aim here is to construct a model of freedom that is sufficiently objective to be used in cross-cultural comparisons, but that is also sensitive to the different weightings given to the elements of a flourishing life in different cultures. Sen (1992) draws up a list of 'functionings' that include 'being adequately nourished' and 'having self-respect,' and defines 'well-being freedom' as the capability to achieve those functionings. However, there is also 'agency freedom,' defined as the capability to achieve whatever one's objectives are. Since many people have objectives other than their own well-being, these two freedoms may conflict.

It is agency freedom that has been at the center of most liberal discussions. Sen acknowledges its importance, but weighs it alongside well-being freedom in evaluating the totality of a person's life. Even within agency freedom, Sen advises liberals not to exaggerate the importance of the agent's direct control of an outcome. Instead, he proposes a counterfactual formula, "I am free when control is exercised in line with what I would choose." That formula can be generalized to cover all forms of freedom, even when the term 'freedom' is extended to cover 'freedom from hunger,' 'freedom from malaria,' etc.: 'One values living without malaria, desires such a life, and would have chosen it, given the choice.'

Society has different obligations to its members concerning their well-being and their agency. It has an evident 'responsibility to a person's well-being, especially when it is in some danger of being particularly low,' which it does not have for that person's particular agency objectives. In advanced liberal societies at least, 'in which the focus is on the freedom that people have,' no patterned outcome can be imposed. For 'even when complete equality of well-being freedom is … entirely realized, this need not lead to an equality of well-being achieved, since different individuals may give different priority to the pursuit of their own well-being.' In rejecting patterned outcomes, Sen has shown that the capability model does not conflict with the central values of liberalism expounded by Rawls and Dworkin.

The capability model inherits from Aristotle the idea that human beings are endowed with specific capacities that can be more or less fully realized in a flourishing life. In his presentation of the model, Sen is concerned that Aristotle's particular 'view of human nature (with a unique list of functionings for a good human life) may be tremendously overspecified.' He prefers to leave open 'both … the exact grounds underlying the determination of relative weights, and the actual relative weights chosen.' This gives Sen's version of the 'capability approach' not only a degree of incompleteness relative to Aristotle's but also, he thinks, a "considerable 'cutting power'" for use in the real world of cultural diversity.

Freedom as Attaining Autonomy through Community: Raz

Communitarian critics including Taylor, MacIntyre, and Sandel brought two charges against liberalism. At the level of theory, they claimed that it is based on a methodological individualism that in turn derives from an impoverished, atomistic conception of human nature. At the level of practice, they argued that because the liberal order fails to provide a satisfactory sense of identity for its members, 'intolerance flourishes most where forms of life are dislocated, roots unsettled, and traditions undone.'

Along with the communitarians, Raz (1986) rejects methodological individualism, but against them he argues that this

does not entail a rejection of liberal values. Raz's liberalism, 'the doctrine of limited government,' is rooted in a 'non-individualist conception of morality.' In opposition to Rawls's antifoundationalism, Raz propounds a "doctrine of political authority that rests on a perfectionist political defense and promotion of liberty and autonomy." However, this is not a brutish libertarianism, since freedom is 'a distinct value, but one that is intimately intertwined with others, and cannot exist by itself.' Raz in short is a foundationalist, who seeks to establish 'continuity' between a moral base and a political–legal superstructure.

For someone to attain 'the ideal of autonomy' and live an autonomous life, certain conditions must be satisfied. These are the possession of 'appropriate mental abilities,' of an 'adequate range of options,' and of 'independence.' The ideal is positive and comprehensive. It constitutes the moral base of the system, but is not 'moralistic.' In particular, it does not involve a teleological conception of self-realization, since, as an autonomous person, one is free to choose whether or not to realize oneself. However, without teleology it is difficult to make autonomy foundational. If autonomy involves integrity ('identification with one's life'), self-awareness and commitment to a project, it can still be asked why it is valuable. It is insufficient to answer that 'many people desire to be autonomous,' since people's desires are many, conflicting, and mutable. Rather, says Raz, "the value of autonomy does not depend on choice … the conditions of autonomy concern a central aspect of a whole system of values of a society …." The kind of society in question is a pluralistic society, which both supports and is supported by this central value 'valuing autonomy leads to the endorsement of moral pluralism.'

Other modern liberals, notably Rawls, are cautious about seeking to found political legitimacy on a particular understanding of freedom. They hold that particular individuals and groups can value freedom for different reasons, some welcoming it because it is a purely negative guarantee against interference, others because it embodies their highest aspirations. Against them Raz holds that most people in a liberal state must subscribe to the positive value of freedom as autonomy, otherwise its tolerant pluralism will perish.

Civic Republicanism

In the fifth model of political freedom, the 'liberty of the ancients' reenters the modern debate. In both versions of 'civic republicanism' considered below, political participation plays a key role, a purely instrumental role in the first, and a constitutive role in the second.

The Instrumental Notion of Civic Republicanism: Skinner

Skinner's model, known variously as 'civic republicanism,' 'classical republicanism,' and 'neo-Romanism' is rooted in the Renaissance, in particular in the idea derived from Machiavelli that a state that is 'free from all external slavery' guarantees 'personal liberty, understood in the ordinary sense … each citizen remains free from any elements of constraint, especially those that arise from personal dependence and servitude, free to pursue his own chosen ends … To be free … is simply to be unconstrained from pursuing whatever goals we may happen to set ourselves.' The independence of the state is linked to the freedom of the individual insofar as 'a self-governing republic is the only type of régime under which a community can hope to attain greatness at the same time as guaranteeing its citizens their individual liberty.' Such a state requires two capacities or virtues from it citizens, courage and determination to defend their community against external attack, and practical reason that qualifies them to play an active part in public life. This means that the state can, legitimately, force the citizen to be free. Despite appearances, this 'strenuous view of citizenship' is, according to Skinner (1990), consistent with NL in two respects. First, it does not presuppose any specific teleological theory of human nature. Second, when we are 'forced to be free,' we are being forced not to 'reason about ends,' but to 'recognize the range of actions we have good reason to perform in order to bring about the ends we actually desire.' Skinner illustrates the latter point by a contrast between 'liberals' like Hobbes and Locke on the one hand and Machiavelli on the other. The 'liberals' assert only that the law preserves A's freedom by coercing B (the potential law-breaker). Machiavelli asserts this. But also asserts that the law also preserves A's freedom by coercing A "into acting in a particular way … forcing us out of our habitual patterns of self-interested behavior … into discharging the full range of our civic duties and thereby ensuring that the free state on which our own liberty depends is itself maintained free of servitude." In short, because "freedom depends on service and so on our willingness to cultivate the civic virtues … we may have to be coerced into virtue and thereby constrained into upholding a liberty that, left to ourselves, we would have undermined." Skinner comments that "contemporary liberalism … is in danger of sweeping the public arena bare of any concepts save self-interest and individual rights." Ideologies suspicious of the whole modern agenda of individual freedom have stepped in to fill that vacuum. Skinner (1998) thinks that 'republican freedom' provides a robust defense against such reactionary positions.

The Constitutive Notion of Civic Republicanism: Pettit

In the second version of civic republicanism, espoused by Pettit (1997), Berlin's dichotomy of the positive freedom of self-mastery and the negative freedom of noninterference is not exhaustive. Beyond those two lies republican freedom, which means absence of domination, not absence of interference. Domination differs from interference in important respects. Domination is a function of the relationship of unequal power between persons or groups of persons. It relates to the capacity of one person or group of persons to interfere in the choices of other persons or groups, rather than to actual instances of interference. However, that would not be enough to pick out the undesirable character of domination, since any non-anarchist social order will empower certain persons or groups, giving them authority over others in given circumstances. So domination is not just the possession of power, but the possession of arbitrary power. The goal of republican freedom, then, is that 'no one is able to interfere on an arbitrary basis – at their pleasure – in the choices of the free person.' With this formula Pettit seeks to extract the rational kernel of both negative and positive freedom, while showing that neither is

sufficient in itself. On the one hand, freedom as non-domination 'will facilitate the achievement of autonomy,' as positive freedom requires. On the other hand, it will respect people's empirical choices, as negative freedom requires, but it will go beyond by showing that dominated people are not really free even if they have the good fortune to suffer no interference when their dominators choose to refrain from exercising their arbitrary power.

Insofar as republican freedom is tied to power, it is essentially egalitarian. It is held to protect each individual against arbitrary power, and also to be a 'communitarian good,' allowing people to identify with a state that protects their freedom. This version of republican freedom is heavily influenced by Rousseau, purged of totalitarian accretions, and updated to the advanced capitalist societies of the late twentieth century. They are now explicitly inclusive, bestowing their benefits on all members of society, and also multicultural, displaying liberal neutrality toward different substantive conceptions of the good. How far such societies can provide a stable balance between the participatory core of republican freedom and the centrifugal drives of modern pluralism remains to be seen.

See also: Analytical Marxism; Autonomy, Philosophy of; Economics and Ethics; Justice (Philosophical Aspects); Libertarianism; Republicanism: Philosophical Aspects.

Bibliography

Berlin, I., 1958. Two Concepts of Liberty. Clarendon Press, Oxford.
Cohen, G.A., 1995. Self-ownership, Freedom and Equality. Cambridge University Press, Cambridge, UK.
Constant, B., 1988. The liberty of the ancients compared with that of the moderns. In: Fontana, B. (Ed.), Political Writings. Cambridge University Press, Cambridge, UK.
Dworkin, R., 1981. What is equality parts 1 and 2. Philosophy and Public Affairs 10, 185–246, 283–345.
Dworkin, R.M., 1977. Taking Rights Seriously. Duckworth, London.
Guest, R., 1997. Ronald Dworkin. Edinburgh University Press, Edinburgh, UK.
Honohan, I., 2002. Civic Republicanism. Routledge, London.
Nozick, R., 1974. Anarchy, State, and Utopia. Blackwell, Oxford, UK.
Pettit, P., 1997. Republicanism: A Theory of Freedom and Government. Clarendon Press, Oxford, UK.
Rawls, J., 1971. A Theory of Justice. Belknap Press of Harvard University Press, Cambridge, MA.
Rawls, J., 1993. Political Liberalism. Columbia University Press, New York.
Raz, J., 1986. The Morality of Freedom. Clarendon Press, Oxford, UK.
Sandel, M.J., 1982. Liberalism and the Limits of Justice. Cambridge University Press, Cambridge, UK.
Sen, A., 1992. Inequality Reexamined. Clarendon Press, Oxford, UK.
Skinner, Q., 1990. The republican idea of political liberty. In: Bock, G., Skinner, Q., Viroli, M. (Eds.), Machiavelli and Republicanism. Cambridge University Press, Cambridge, UK.
Skinner, Q., 1998. Liberty before Liberalism. Cambridge University Press, Cambridge, UK.
Taylor, C., 1979. What's wrong with negative liberty? In: Ryan, A. (Ed.), Essays in Honour of Isaiah Berlin. Oxford University Press, Oxford, UK.
Van Parijs, P., 1995. Real Freedom for All. Clarendon Press, Oxford, UK.

Free Will and Action

Michael Smith, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

When agents are in circumstances that do not constrain their choices in any way they enjoy *freedom of action*. They are able to do whatever they want to. However freedom of action, so understood, is neither necessary nor sufficient for moral responsibility. It is not sufficient because, at least when placed in suitably unconstraining circumstances, spiders, mice, infants, and the like, all enjoy freedom of action, but they are not morally responsible for anything that they do. It is not necessary because people who falsely believe that they enjoy freedom of action may be morally responsible. What is both necessary and sufficient for moral responsibility is *freedom of the will*. Agents enjoy freedom of the will when they possess a dual capacity. On the one hand, they must have the capacity to form well-justified beliefs about what they have reason to do. On the other hand, they must have self-control: the capacity to bring their desires in line with their reflectively formed beliefs about their reasons. Many urge that moral responsibility is impossible in a deterministic world. But whether or not this is so depends on whether rational capacities, quite generally, are possible in a deterministic world. It is a question for further investigation whether this is so.

According to the standard account of action, an agent's acts are those of his bodily movements that spring in an appropriate way from his will. Bodily movements that are caused in some other way – the bodily movement of someone who gets blown over by the wind, say – are not among the actions agents perform. More precisely, an agent's actions are those of his bodily movements that occur because he wills them to, where an agent's will is, in turn, identified with his desires and means-end beliefs: actions are bodily movements that occur because an agent desires some outcome and believes that that outcome can be produced by his moving his body in the relevant way, a way he can move it. Bodily movements that trace their causal history to other states or events in an agent's psychology, or to nonpsychological states or events are not actions (Davidson, 1963).

The standard account of action suggests that all animals that are capable of desire and belief, whether those animals are human or nonhuman, are capable of action. The bodily movement of a mouse occurs because the mouse wants some cheese and believes that it can get some by running in the direction of the crumbs on the floor is an action of the mouse, just as much as the intricate performance of a gymnast or a ballet dancer. By contrast, the movement that the mouse makes when it is picked up by a cat and flung across the room is not an action, any more than the movement of a ballet dancer who is unceremoniously dropped to the floor by her partner.

Not all of an agent's actions, whether the agent is human or nonhuman, are free, however, notwithstanding the fact that they all spring from his will. When a mouse is caught in a trap by its tail, it may give up the struggle and sit down near the trap. Its body remains under its control, not under the control of some alien force, but staying near the trap is not something that the mouse does freely. Similarly, when a person is locked in a prison cell the individual may give up the struggle and sit on the chair behind bars. But staying behind bars is not something that the individual does freely. What, then, within the class of actions, is the mark of those of an agent's actions that are free, as opposed to those that are unfree?

Everyone agrees that free actions are those that are under the agent's control, and some suggest that this means free actions are all and only those that are done intentionally: that is, they are those of an agent's actions which, suitably described, can be seen to have been rationalized by the individual's desires and beliefs (Davidson, 1973). This idea might seem supported by paradigm cases of unfree action, such as that of the human and the mouse. For although the mouse stays near the trap and the human stays behind bars, neither does what they do intentionally under these descriptions. The human, for example, intentionally sits on a chair, as opposed to sitting on the ground – that is something that is rationalized by the individual's desires and beliefs – and one consequence of thus sitting on the chair is that the individual stays behind bars. But the individual does not have any desires and beliefs that rationalize the individual's staying behind bars. That is merely a consequence of what the individual does intentionally, not itself something that the individual does intentionally.

But though this certainly captures one sense in which we might say of actions that they are under an agent's control, it does not capture all there is to the idea. It entails, for example, that whenever someone who acts in a certain way knowingly, but not intentionally, he does not act freely, and this is surely quite wrong. Suppose that someone, A, desires to convey to another, B, the low opinion he or she has of a third person, C. But let us suppose further that he or she goes on to tell B in full knowledge of the fact that C is standing within earshot, and so causes C to become embarrassed. A could have waited until C left, but did not because A did not care. A then knowingly causes C embarrassment, but A does not do so intentionally, as there is no desire of A's such that A believes he or she can satisfy that desire by embarrassing C. C's embarrassment is merely a foreseen consequence of what A does do intentionally, which is to tell B his or her low opinion of C. It would, however, be extremely implausible to suppose that the fact that C's embarrassment was merely foreseen, rather than intentionally produced, makes it the case that A did not cause C's embarrassment freely. It was, after all, entirely up to A whether or not

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63026-X

C would be embarrassed by what A said to B about C. A was not forced to embarrass C. Embarrassing C or desisting from doing so was under A's control.

Or suppose that someone, D, desires to convey to another, B, the low opinion he has of a third person, C. But this time let us suppose that, eventhough D does not know that C is standing within earshot, in view of the fact that C is nearby where D and B are, D could have ascertained the fact that C was in close proximity, and then waited until C left, but D did not do so because D did not care enough. In this case, D both unintentionally and unknowingly causes C to become embarrassed; D has neither any specific desire or satisfaction by embarrassing C, nor does D even foresee that he or she is embarrassing C. C's embarrassment is in this case a negligently unforeseen consequence of what D does do intentionally, which is to tell B his or her low opinion of C. It would, however, once again be implausible to suppose that the fact that C's embarrassment was neither intentional nor foreseen makes it the case that A did not cause C's embarrassment freely. It was, after all, entirely up to A whether or not C would be embarrassed by what he said. He was not forced to embarrass C. He could easily have availed himself of C's presence. Embarrassing C (or desisting from doing so) was therefore under A's control.

The more standard idea has thus been that an agent's acts are free just in case they are under the agent's control in the more mundane sense that the agent could have done otherwise than what he in fact did when the agent acted (Wiggins, 1973). The question 'Did the agent act freely?' is in this way supposed to reduce to the question 'Did the agent have any alternative option?' A's causing C to become embarrassed is therefore indeed a free action, on this way of seeing things, because A had the option of acting in some other way instead. He could have waited until C was out of earshot. This is the relevant sense in which free actions are under an agent's control. The reason that a mouse caught in a trap does not stay in the vicinity of the trap freely is thus that it has no other option: it cannot do otherwise but stay near the trap. And the reason a prisoner who sits on a chair in a locked prison cell does not stay behind bars freely is that this individual, too, has no other option: the individual cannot do otherwise than stay behind bars.

However this more standard account of free action immediately raises a difficult issue, for without a special explanation of what the 'could' means, it seems to imply that freedom of action is impossible in a deterministic world. Every action performed by an agent in a deterministic world is fixed by the way the world was long before the agent was born together with the operation of causal laws. In a very familiar sense of 'can,' no agent can ever do otherwise than he in fact does. Agents are therefore always forced to do what they do by their past and the forces of nature.

Incompatibilists – those who think that freedom is impossible in a deterministic world – happily accept this conclusion (van Inwagen, 1975). But they in turn divide over the substantive question whether the world is or is not deterministic. Those who think that it is – the so called 'hard' determinists – draw the conclusion that there are no free actions. Those who think that there are free actions – the 'libertarians' – draw the conclusion that the world is not deterministic. But not everyone is an incompatibilist. Compatibilists find it difficult to see how the mere fact of causal determinism is supposed to undermine freedom of action. If anything, they insist, things are the other way around. Events that cannot be explained causally are completely random: that is what it means to say that an event eludes causal explanation. But free actions can hardly be randomly produced. Free actions, like all actions, are produced in a very reliable fashion by the desires and beliefs of their agents. They therefore conclude that free actions, if there are any, are caused by desires and beliefs, and hence the sense of 'could' at issue must be different from the one that so impresses the incompatibilists (Ayer, 1954).

Compatibilists have therefore coupled the suggestion that an agent's acts are free just in case he could have done otherwise with a special explanation of what this 'could' means. To say of an agent that he could have acted otherwise, in the sense relevant to establish freedom of action is, they suggest – or, at any rate, this is the way the suggestion was originally put – to say of him that he would have done otherwise if he had chosen to (Moore, 1911). In other words, freedom of action turns on whether or not, in the nearby possible worlds in which agents choose to do otherwise than they do in actuality – these are possible worlds in which the past is ever so slightly different – they succeed in doing what they choose to do. The truth or falsity of determinism is irrelevant.

The intuitive idea here should be clear enough, for the suggestion amounts to the claim that agents enjoys freedom of action if the external world presents no barrier to their efforts to act as they will: freedom is a matter of the absence of such external barriers or constraints. To see how this allows us to sort cases of free from cases of unfree action, consider the difference between two men who stay in a prison cell, one of whom is locked in, and one of whom is not. The difference is that the one, but not the other, would have walked out of his prison cell if he had chosen to do so: in the nearby possible worlds in which the agent who is locked in his prison cell chooses to walk out, he fails to do so, as he is still locked in, whereas in the nearby possible worlds in which the agent who is not locked in his prison cell chooses to walk out, he succeeds in doing so, as his cell presents no barrier to him. The one has the option to walk out, whereas the other does not. Moreover this account is, apparently, fully general. Of two mice that stay nearby a trap, the one whose tail is not caught in the trap, and so moves away in the nearby possible worlds in which it chooses to do so, stays nearby the trap freely, whereas the one whose tail is caught in the trap, and so does not move away from the trap in the nearby possible worlds in which it chooses to do so, does not stay nearby the trap freely. The external world presents a barrier to the movements of the latter, it serves as a constraint, but it presents no such barrier or constraint to those of the former.

Intuitive though this idea might be, however, the suggestion in the form in which it was originally made is problematic. The problem is that, since choosing is itself an action, it is something that people can do freely or unfreely (Chisholm, 1964). But now suppose we ask whether a particular action is free. This turns on whether the agent could have done otherwise than what he in fact did, and, according to the suggested analysis of 'could,' this turns on what happens in those nearby possible worlds in which the agent chooses to act otherwise. But whatever answer we give to this question, we have been forced

to admit that the status of the original action as free or unfree turns on what happens when another merely possible act is performed, an act whose status as free or unfree is left up in the air. Moreover the status of this merely possible act turns on whether or not, in those nearby possible worlds in which the agent chooses otherwise than to choose to act otherwise, he succeeds in so choosing. And, whatever answer we give to this question, it follows that we have therefore been forced to concede the possibility of yet another act whose status as free or unfree needs to be determined. And so we must go on.

The problem is thus that in order to raise and answer any particular question about the freedom of an action we are required to posit further acts of choosing to act otherwise, and choosing otherwise than to choose to act otherwise, and so on, about which we can raise the question whether they are free or unfree. The problem is not just that questions about freedom of action are never answered, once and for all – every question answered leaves another posed but unanswered, and so on *ad infinitum* – the problem is rather that it is implausible to suppose that every being capable of free action is capable of such ever higher and higher order acts of choosing. In order to raise the mundane question whether a mouse, caught in a trap, sits nearby the trap freely or unfreely it is surely quite incredible to suppose that we must thereby concede the possibility of its choosing to do otherwise than to sit near the trap, and choosing otherwise than to choose to do otherwise than to sit near the trap, and so on. Mice simply are not cognitively sophisticated enough to make such choices.

How are we to solve this problem? The suggested analysis of 'could have done otherwise' clearly needs to be rephrased so that the possibility of the freedom of an agent's actions does not entail the possibility of his making ever higher and higher order acts of choice. We should therefore suppose that for it to be true that an agent could have done otherwise it has to be the case that he would have done otherwise, not if he had chosen, but rather *if he had wanted*. This not only removes the problem we have just encountered – it is indeed possible for agents whose acts are free, humans and mice alike, to want other than they actually want, and that possibility does not commit us to the possibility of their making higher order acts of choice – it also better captures the intuitive idea presented at the outset, the idea that an agent's actions are free to the extent that they are dependent on his will. For, as we saw, an agent's will is to be identified with his wants and means-end beliefs. To say of an agent, be he human or mouse, that the agent acts freely is thus, according to the rephrased suggestion, to say of him that, in the nearby possible worlds in which he wants something different, he acts in that different way. This is what it is for him to have alternative options.

Note two consequences of this account of free action. The first is that certain acts which we would have ordinarily supposed to be unfree are not really unfree. When someone says 'Your money or your life!' and I give him my money, we would ordinarily say that I do not act freely because I have been forced to hand over my money. But, according to the account on offer, I do act freely. I have another option: I refuse in those nearby possible worlds in which I want not to give him my money more than I want him not to kill me. Many think that this is implausible. They insist that there must be a sense in which my freedom is compromised by such a threat, as even though I refuse in the nearby possible worlds in which I want to refuse, wanting to refuse would not be at all reasonable, in those circumstances. Indeed, not only would it not be reasonable, it would be completely crazy. There is, they suggest, thus the following sense in which the person who threatens me forces me to act: he puts me in circumstances in which there is no other reasonable option but to comply with his threat. But if this is right – that is, if this is the sense in which threats compromise my freedom – then it seems to follow that people who make me irresistible offers equally compromise my freedom, for they too put me in circumstances in which the only reasonable option is to accept their offer. Whether it is acceptable to suppose that the only sense in which threats compromise my freedom is a sense in which offers also compromise my freedom clearly needs much more thought (Frankfurt, 1973).

The second consequence concerns the relationship between free action and moral responsibility. According to this account of free action, freedom is not sufficient for an allocation of moral responsibility. Mice often act freely, but they are not morally responsible for any of their acts. Something else is therefore required. Other more complicated cases tell in favor of the same conclusion. Consider, for example, the plight of a reluctant alcoholic who wants to have a drink, and so has one, knowing he should not. He presumably drinks freely, as he does not drink in those nearby possible worlds in which he wants not to have a drink. Yet it is surely an open question whether or not he is morally responsible for his drinking. If, for example, his desire to drink is literally irresistible, then he is not responsible. Indeed, we would even think that he is not free, in the important sense of being unable, in the circumstances, to want to do anything else except to drink. This more complicated case thus suggests that in order to understand the conditions of moral responsibility we must look not so much to an account of free action, but rather to an account of freedom of the will. The idea is that, though the alcoholic's actions may be free, his will is not, and that this is therefore the source of our judgments about moral responsibility (Frankfurt, 1971).

Is the performance of a free action a necessary condition of moral responsibility? The answer seems to be that we can and do hold agents responsible for acting when they have no alternative, and that acting freely would thus seem not to be a necessary condition either. Those who act in the false belief that their acts are free are often held morally responsible for what they do, notwithstanding the fact that they don't act freely. Imagine, for example, that when a family visits an old prison they disperse so that they can explore the place individually, but that they agree to meet at a certain time to go home. When the agreed time to leave arrives, suppose that one of the children fails to turn up. His parents look for him and find him in a cell. Because he does not want to leave, he has hidden from them. But now suppose that the cell door is closed, and that, unbeknown to the child, it is not just closed but also locked. It then follows that the child had no alternative but to be in the prison cell when it came time to leave. He could not have met the rest of his family. Is the child off the hook in terms of responsibility? Hardly. His parents would surely quite rightly continue to hold him morally responsible for failing to turn up. The fact that he could not have done otherwise is

irrelevant because it does not figure, in the right kind of way, in the explanation of the child's failure to meet them. In order to see what is relevant, though, we must once again look at what does figure. This case, too, suggests that we need to look more carefully to an account of freedom of the will (Frankfurt, 1969).

It might be thought that we should simply model our account of freedom of the will on the account we give of freedom of action. Just as freedom of action is a matter of being in circumstances in which one can do whatever one wants to do, it might be thought that freedom of the will is a matter of being in circumstances in which one can want whatever one wants to want (Frankfurt, 1971). Where freedom of action is a matter of the absence of external constraints on action, so freedom of the will is a matter of the absence of internal constraints on wanting. An agent's will is free just in case, in those nearby possible worlds in which he wants to want to act in a certain way, he wants to act in that way.

There is clearly something right about this idea. Those who enjoy freedom of the will do seem to have the capacity to control their will, where this includes the capacity to control their first order desires. The problem with the reluctant alcoholic seems to be that he has no such ability: his desire to drink is beyond his control. What is less clear, however, is that this idea is well captured by the suggestion just made. The desires an agent has about his first order desires are, after all, just more desires he has, and are hence just another part of his will, a part which may or may not be under his control (Watson, 1975). Thus, even supposing it to be true that in those nearby possible worlds in which an agent wants to want to act otherwise, he wants to act otherwise, we can always imagine that the agent is unable to have wanted to want to act otherwise. And if the agents were unable to want to act otherwise then we must surely ask in what sense he enjoys freedom of the will. Certain of the desires such an agent has in certain possible worlds are controlled by certain others in those worlds, but the agent himself does not seem to be in control of any desires at all.

Some theorists have argued that in order to solve this problem we need to postulate a special kind of higher order desire: a desire which captures the agent's own sense of his identity (Frankfurt, 1987). When an agent has this sort of higher order desire, they suggest, he identifies himself with the lower order desires that are this higher order desire's objects. Freedom of an agent's will is, the suggestion goes, a matter of his being able to have desires that match this sort of higher-order desire: in other words, it is a matter of having desires with which he identifies.

The difficulty with this suggestion, however, is to come up with a suitably independent account of identification. There certainly is a sense in which an agent who acts of his own free will identifies with the desires on which he acts, as he is himself the originating source of such desires. But this is not an independent sense of identification. It is a sense that we define in terms of the idea of desires originating in a free will. There is also at least the following independently characterizable sense of identification. Some people who take drugs against their will, and who struggle to overcome their addiction, ultimately give in, and so come to identify with their addiction in the sense of building all of their expectations around it. In other words, they come see themselves as hopeless drug addicts: this becomes their self-conception. But this is hardly a sense of identification that will help us understand the idea of those desires that originate in a free will.

There is something right about connecting freedom of the will with the possession of desires with which an agent identifies, however. Since agents are the originating source of such desires it does seem that, when they act of their own free will they express their true nature. The desires on which they act are, in some sense, essential to who they are. But though this seems right, it in fact suggests a quite different account of free will, one which does not depend crucially on our having an independent idea of what it is for an agent to have desires with which he identifies.

Since the Enlightenment we have come to conceive of people as being essentially truth-oriented reasoners. They are beings who have the capacity to recognize the truth and to respond appropriately to it. In so far as we are deliberators, what this suggests is that we are essentially beings who have the capacity to recognize and respond to the facts about what we have reason to do. Thus, according to this account, an agent whose will is free is someone who asks himself or herself what there is reason to do and, in answering this question, he or she is sensitive to the available evidence. This is not to say that he or she will inevitably acquire true beliefs about what there is reason to do, but it is to say that when he or she fails to do so, and when the evidence was available to him or her, one's failure will be traceable to a failure to exercise his or her belief-forming capacities appropriately: he could and should have formed the correct belief, but he failed to do so. Moreover, an agent whose will is free is someone who, in the formation of desires, is sensitive to the beliefs he has about what he has reason to do. Again, this is not to say that he will inevitably form desires that accord with his beliefs about his reasons, but it is to say that when he fails to do so, his failure will be traceable to a failure properly to exercise self-control: he could and should have formed the correct desire, but he failed to do so (Wolf, 1990).

As is perhaps clear, this conception of free will differ crucially from the one suggested earlier. What is important for freedom is indeed a capacity to desire otherwise, but the possession of this capacity is not to be analyzed in terms of what happens in nearby possible worlds in which the agent wants to want otherwise. Instead it is to be analyzed in the same way, whatever that way is, in which we analyze the ability rational thinkers possess to respond appropriately to evidence in the formation of their beliefs. Just as a rational thinker who fails to adjust his beliefs to the available evidence makes a mistake he should not have made, and so, we suppose, must have had the capacity to make the adjustment and believe rightly instead, so a rational deliberator who fails to form beliefs about what he has reason to do in accordance with the available evidence, or who fails to desire in accordance with his beliefs about what he has reason to do, makes a mistake he should not have made, and so is supposed to have had the capacity to control himself and acquire the correct belief or desire instead (Pettit and Smith, 1996).

Note two consequences of this account of freedom of the will. The first is that it looks tailor-made to provide an account of the conditions of moral responsibility. Those who have no capacity to believe that they have reason to do anything other

than they in fact believe they have reason to do, or to desire to do anything other than they in fact desire to do, can hardly be praised for having the right beliefs and desires, if it so happens that they are right, or blamed for having the wrong ones, if it so happens that they are wrong. This explains why we don't hold the alcoholic responsible. The irresistible nature of his desire to drink makes him incapable of an exercise of self-control, despite the fact that he believes, correctly, that he has reason to abstain. It also explains why we don't hold mice morally responsible for what they do. Their lack of cognitive sophistication deprives them of the capacity even to form beliefs about their reasons for doing what they do, let alone the capacity to desire accordingly.

Moreover, nothing other than the capacity to form the right beliefs and desires would seem to be required for people to be responsible either. This is why we hold the child described earlier morally responsible for failing to meet his parents at the appropriate time, notwithstanding the fact that he was locked in a prison cell and so could not have met them. We hold him responsible because we assume that his behavior was in fact the product of a desire he had which accorded with his beliefs about what he had reason to do, a belief that we take to be culpably mistaken. It is this that attracts our censure.

The second consequence concerns the possibility of freedom of the will in a deterministic world. A proof that freedom of the will is possible within a deterministic world would seem now to require no more and no less than a demonstration that individuals in a deterministic world have the wherewithal to possess just such a rational capacity. There have recently been attempts to provide such a demonstration (Smith, 2003; Vihvelin, 2013). But whether these particular attempts succeed or not, even on the face of it, there does not look to be any particular difficulty in principle involved in providing such a demonstration. We regularly discriminate between people who do and people who do not have rational capacities, and, in so doing, we seem to pay no regard whatsoever to issues of determinism. This is because, even if determinism is true, that only shows that people will make the mistakes that they make. It does not show that they could not help making those mistakes, in the sense of having no rational capacity to get things right.

But, whatever we ultimately come to think about the possibility of rational capacities in a deterministic world, the crucial lesson of this account of freedom of the will is that, if there do turn out to be such difficulties, then it is not just the idea of freedom of the will that becomes problematic, it is the very idea of a rational agent as someone who has the capacity to recognize and respond appropriately to the truth. Whether freedom of the will is possible in a deterministic world or not, its possibility goes hand in hand with the possibility of agents who have the capacity to form beliefs and desires in a rational manner.

See also: Autonomy, Philosophy of; Causation: Physical, Mental, and Social; Identity and Identification, Social Psychology of; Mental Causation; Practical Reasoning: Philosophical aspects; Responsibility: Philosophical Aspects.

Bibliography

Ayer, A.J., 1954. 'Freedom and Necessity' in Watson 1982. Original Publication.
Chisholm, Roderick M., 1964. 'Human Freedom and the Self' in Watson 1982. Original Publication.
Davidson, Donald, 1963. 'Actions, Reasons and Causes' in Davidson 1980. Original Publication.
Davidson, Donald, 1973. 'Freedom to Act' in Davidson 1980. Original Publication.
Davidson, Donald, 1980. Essays on Actions and Events. Clarendon Press, Oxford.
Frankfurt, Harry G., 1969. 'Alternate Possibilities and Moral Responsibility' in Frankfurt 1988. Original Publication.
Frankfurt, Harry G., 1971. 'Freedom of the Will and the Concept of a Person' in Frankfurt 1988. Original Publication.
Frankfurt, Harry G., 1973. 'Coercion and Moral Responsibility' in Frankfurt 1988. Original Publication.
Frankfurt, Harry G., 1987. 'Identification and Wholeheartedness' in Frankfurt 1988. Original Publication.
Frankfurt, Harry G., 1988. The Importance of What We Care about: Philosophical Essays. Cambridge University Press, Cambridge.
Honderich, Ted, 1973. Essays on Freedom of Action. Routledge and Kegan Paul, London.
Moore, G.E., 1911. Ethics. London.
Pettit, Philip, Smith, Michael, 1996. Freedom in belief and desire. Journal of Philosophy 93 (9), 429–449.
Smith, Michael, 2003. Rational capacities. In: Stroud, Sarah, Tappolet, Christine (Eds.), Weakness of Will and Varieties of Practical Irrationality. Oxford University Press, Oxford, pp. 17–38.
van Inwagen, Peter, 1975. 'The Incompatibility of Free Will and Determinism' in Watson 1982. Original Publication.
Vihvelin, Kadri, 2013. Causes, Laws, and Free Will: Why Determinism Doesn't Matter. Oxford University Press, New York.
Watson, Gary (Ed.), 1982. Free Will. Oxford University Press, Oxford.
Watson, Gary, 1975. 'Free Agency' in Watson 1982. Original Publication.
Wiggins, David. Towards a Reasonable Libertarianism in Honderich 1973.
Wolf, Susan, 1990. Freedom within Reason. Oxford University Press, New York.

French Revolution, The

L Hunt, University of California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 9, pp. 5785–5789, © 2001, Elsevier Ltd.

Abstract

The French Revolution was a momentous historical event that set enduring patterns for modern revolutionary movements and for much of modern politics in general. Contemporary reactions to it fed into the development of many ideologies, from conservatism to communism. Scholarly controversies about the significance and interpretation of the French Revolution have been animated in particular by Karl Marx and Friedrich Engels' use of it as the prime example of a 'bourgeois revolution' that prepared the way for the triumph of capitalism. Although the controversies over the Marxist interpretation continue, recent scholarship has turned in new directions: gender relations; the notion of citizenship and its connection to religion, ethnicity, and race; and the intellectual history of the Revolution and its impact as an event in global history.

Introduction

The French Revolution was a momentous historical event that set enduring patterns for modern revolutionary movements and for much of modern politics in general. Contemporaries experienced the French Revolution as a set of interlocking changes or stages that seemed driven by some kind of mechanism or impetus. Analysis of this mechanism has preoccupied revolutionaries and scholars ever since. The conflicting political reactions to the French Revolution fed into the development of ideologies from conservatism to communism. Scholarly controversies about the significance and interpretation of the French Revolution have been animated in particular by Karl Marx and Friedrich Engels' use of it as the prime example of a 'bourgeois revolution' that prepared the way for the triumph of capitalism. Although the controversies over the Marxist interpretation continue, recent scholarship has turned in new directions: gender relations; the notion of citizenship and its connection to religion, ethnicity, and race; and the Revolution's impact on the colonies.

The year 1789 marks the clear beginning of the French Revolution, but opinion is much more divided about its ending: did it end in 1794 with the fall of Maximilien Robespierre, in 1799 with the rise of Napoleon Bonaparte, in 1804 when Napoleon established his empire, or in 1815 with the final defeat of Napoleon at the Battle of Waterloo? Debate on the question of endings cannot be resolved definitively because it is so closely entwined with interpretive questions. The choice of ending depends on the determination of the central significance of the French Revolution.

What most struck contemporaries and has fascinated commentators ever since was the unpredictability and yet seeming inexorability of the unfolding of events. Even in the middle of this unfolding, political actors and witnesses spoke of 'the system' of 'the French Revolution,' as if revolutionary events had an internal logic or mechanism pushing in a consistent direction. After spending years studying this logic, and having published his enduringly influential study of the French Revolution, Alexis de Tocqueville still complained of his difficulty in grasping 'this malady of the French Revolution,' this '*virus* of a new and unknown kind' (de Tocqueville, 1977).

Revolutionary Stages

As a series of political events, the French Revolution is usually divided into five periods or stages: the constitutional monarchy or liberal revolution of 1789–91; the second revolution and the Terror of 1792–94; the Directorial regime or liberal republic of 1795–99; Napoleon Bonaparte's consulate of 1799–1804; and the Napoleonic empire of 1804–15. This vertiginous succession of regimes established many of the prototypes of modern political authority and even much of modern political language, including the key distinction between Left and Right (determined by the location of deputies vis-à-vis the speaker in the National Assembly). The very possibility of ascertaining periods or stages – already apparent to contemporaries – raised questions about the natural history or morphology of revolution: Was it like a sickness or fever from which a polity eventually recovered? Or was it rather a kind of speeding up of historical development, a bootstrap up to the next stage of capitalism or modernity?

As the French Revolution moved from stage to stage, the revolutionaries generated at least three different models for modern (secular, nonmonarchical) political authority. By abolishing 'feudalism' (eliminating serfdom and seigneurial dues), suppressing noble titles, declaring the 'rights of man,' and founding the first French republic with universal manhood suffrage (1789–92), the French revolutionaries laid the foundations for democratic, representative government. By arresting dissidents, executing opponents, and carrying out a cultural revolution in the form of de-Christianization and republican political reeducation (the Terror of 1793–94), the revolutionaries set the mold for 'totalitarian' government with its repressive insistence on political conformity in the name of a future social good. Napoleon Bonaparte developed a third model of paternalist, authoritarian government by a state with a strong military caste. In other words, the French Revolution adumbrated virtually all the possibilities of modern political authority.

The Origins of Ideology

The French Revolution gave birth both to the term 'ideology' and to many different ideologies that have continued to

organize political discussion right up to the present day. Under the impact of revolutionary events, tradition lost its givenness, hierarchy no longer seemed natural, and monarchy no longer went without justification. The revolutionaries showed that society could be reorganized by an act of self-conscious political will (the definition of revolution itself). The attempt to reshape hearts and minds through festivals, paintings, engravings, operas, plays, and even ordinary items such as snuffboxes and playing cards drew dramatic attention to revolution as a mode of action. This attempt did not produce unanimity, however; rather it produced social conflict. As a consequence, revolution brought 'the social' as a category into sharper relief. As early as 1792, Antoine Barnave wrote his *Introduction à la Révolution française*, arguing that new forms of wealth produced ineluctably new forms of political authority (Barnave, 1988). As the social determinants of politics became more visible, i.e., more of a problem to be investigated, ideologies took shape as formal doctrines about the proper relationship between politics and society.

The first ideology to emerge from the French Revolution was conservatism. Its founding text, Edmund Burke's *Reflections on the Revolution in France* (1790), was written in response to the events of 1789. The National Assembly's efforts to remake the French constitution led Burke to denounce what he termed a new 'political metaphysics' based on natural rights and abstract reason. Unlike many reactionary royalists, however, Burke did not just excoriate the French Revolution as the incarnation of evil and impiousness. He understood that the revolutionaries were in the process of erecting an entirely new foundation for politics, and he undertook to counter them point-by-point. By arguing for a conservatism that accommodated change through practical experience and gradual amelioration, he admitted the necessity of confronting demands for change and the parallel demand for an elaborated doctrine to support the status quo. Even for conservatives, then, the French Revolution changed the rules of the game. The status quo now required explicit defense.

After 1789 liberal supporters of the Revolution also had to clarify their position. Most liberals supported the constitutional revolution of 1789 and rejected the Terror of 1793–94. This distinction was not always easy to uphold, for conservatives routinely accused liberals of supporting policies and positions that might open the way to more radical experiments, just as those of the revolutionaries of 1789 had opened the way to the Terror of 1793–94. The unfolding of stages in the revolutionary process thus shaped ideological debate, either explicitly or implicitly, for many generations after the French Revolution.

Socialists, and especially Marxists, traced their lineage directly back to the French Revolution. The early socialists – Claude Henri de Saint-Simon and Charles Fourier, for example – concluded from the French Revolution that society could be reorganized according to a technocratic project (thus 'utopian socialism' according to Marx). Marxists believed that only class struggle could produce a new social order. According to Marx and Engels, the French Revolution marked the overthrow of the aristocracy by the bourgeoisie. The bourgeoisie would develop capitalism on the ruins of the feudal order it dismantled, thus preparing the way for a subsequent revolution in which the proletariat formed by capitalist development would overturn the bourgeoisie and replace capitalism with communism. The French Revolution was therefore a crucial stage in the class struggle and an essential forerunner to the future communist order.

Marx's emphasis on the need for self-conscious revolutionary organization had been at least partially inspired by Gracchus Babeuf and his 'Conspiracy of Equals' of 1796. Although Babeuf had few followers and his plans were foiled by police infiltration, his dream of a revolutionary seizure of political power by a secretly organized conspiratorial group lived on through the efforts of the Italian revolutionary Filippo (or Philippe) Buonarroti. Buonarroti, a member of Babeuf's circle in 1796, published a history of the Conspiracy of Equals in 1828. It brought the message of revolutionary seizure of power to secret societies throughout Europe. One of Buonarroti's followers, the French revolutionary socialist Auguste Blanqui, combined Buonarroti's appeal for a secret revolutionary brotherhood with the idea of a 'dictatorship of the proletariat,' a term coined by Blanqui to define the political goal for socialism. Marx and Engels incorporated both Babeuf's and Blanqui's notions into their program for proletarian revolution.

The Foundations of Modern Social Science

Because the French Revolution brought the relationship between society and politics into sharper focus, it also accelerated the development of the social sciences, first in France and then more broadly in Europe. The Directory regime created the first research institute for the social sciences with its department of political and moral sciences in the Institut National des Sciences et des Arts. Bonaparte, himself a member of the Institute, took more than 150 scientific experts with him on his expedition to Egypt in 1798. They set up an Institut d'Egypte in Cairo and helped inaugurate 'oriental studies' in Europe. They also discovered the Rosetta Stone, which enabled experts to decipher Egyptian hieroglyphics.

The French Revolution had an especially dramatic impact on historical studies in France. The first full-length histories of the Revolution were published in the 1820s by François-Auguste Mignet and Adolphe Thiers, journalists and editors in the political opposition. They used the history of the French Revolution to defend the notion of a constitutional monarchy and to criticize the restored Bourbons. In the 1840s, Louis Blanc, Alphonse Lamartine, and Jules Michelet wrote histories of the Revolution as part of their campaign for a republic. Although Marx did not write a history of the French Revolution of 1789, he did read the French historians and made constant reference in his work to the French Revolution of 1789.

Marx was not alone in considering the French Revolution an important stage in historical development. As the notion of historical evolution spread in the nineteenth century, the French Revolution almost inevitably became an important benchmark. Auguste Comte, one of the founders of sociology, took as his point of departure the last writings of the Marquis de Condorcet, a French nobleman and revolutionary who committed suicide rather than face the guillotine in 1794. Comte proposed three stages of historical development: the theological, the metaphysical, and the positive or scientific. The French Revolution helped bring about the ongoing transition

to the third, positive stage by liberating modern social elements from the dead hand of the Old Regime. The task of social science was to complete this transition, which could only be accomplished finally in ideas and opinions, according to Comte, not by revolutionary action. The ultimate aim of his system was the reorganization of society by means of scientific sociology.

Alexis de Tocqueville, who helped lay the foundations for political sociology and political science with his *L'ancien régime et la Révolution* (1856), considered the French Revolution not as a stage so much as an accelerator of the general modern trend toward democracy and equality. Tocqueville had read the previous historians of the French Revolution, though he rarely cited them, and absorbed the view that social changes had inevitable political repercussions. He wove his analysis of social changes into a subtle examination of French political culture and especially the dominance of the central state. In so doing he created an alternative to the Marxist emphasis on class struggle. Tocqueville considered the French Revolution pivotal in the advance of modernity but largely negative in its effects. Tocqueville saw in the French Revolution the tendency of democratic revolution to reproduce the despotism of monarchical authority by leveling all institutions that might have provided a check on state power. Although the French made their Revolution for liberty and equality, they gave up liberty in order to be assured of equality and ended up with an equality of servitude to the Napoleonic state. This line of analysis had a long posterity – Hannah Arendt's analysis of revolution has clear affiliations to Tocqueville's – and continues to influence political and social theory right down to the present.

Scholarly Controversies Concerning the Revolution

Since analysis of the French Revolution has played such a prominent role in the development of competing political ideologies, it is hardly surprising that it has generated disputes among scholars as well. In the nineteenth century, writers used the French Revolution to argue for and against monarchy, empire, or republicanism as a form of government. After the eruption of the Russian Revolution in 1917, however, debate shifted to the terrain defined by Marxism and communism, and in particular, the French Revolution as the exemplar of bourgeois revolution.

Within France, Marxist historians of the French Revolution enjoyed the advantage of having captured the chair of the history of the French Revolution at the Sorbonne (University of Paris). For most of the twentieth century, the chair was occupied by a succession of talented socialists and communists. They pioneered the study of social tensions and social classes, of 'history from below.' The most remarkable of them was Georges Lefebvre, who wrote wide-ranging studies of the peasantry during the Revolution. Lefebvre's combination of narrative skills and analytical studies gave him an unrivaled reputation; even those who objected to his Marxist views regarded him as the most accomplished academic historian of the Revolution.

The Right did not have the same initial foothold within the French university system. Hippolyte Taine, for example, wrote his four volumes on the Old Regime and the French Revolution as a general literary and cultural critic, not as a professional historian. Taine's views remained fundamental to the Right wing criticism of the Revolution. He had nothing but contempt for 'the people' so beloved by Michelet. He frequently likened the lower classes to beasts or insects; at worst, they were a crazed, frenzied rabble. Augustin Cochin toned down this virulence, arguing that the lower classes were manipulated by professional politicians who came to power through the activities of secret societies such as Masonic lodges. The Right took up some of the claims of the Left – that the French Revolution foreshadowed the Russian Revolution of 1917 – but gave them a negative rather than a positive spin.

Before the 1970s, the Marxist Left enjoyed the dominant position in French academic debates, in part because of its strength within the university, and in part because the Right was associated with antidemocratic politics. The balance began to shift, however, as critics of the Revolution finally accepted the republican form of government. A more moderate republican Right got support from a growing body of work produced in the Anglophone countries, especially the United States and the United Kingdom, which attacked the central Marxist thesis that the French Revolution was a bourgeois revolution. Critics argued that the capitalist bourgeoisie had not made the Revolution (first aristocrats and then middle-class officials and lawyers had led it, not manufacturers); that the Revolution had not overthrown feudalism (because feudalism no longer existed); and that the Revolution did not pave the way for capitalism (since industrialization did not take place until the 1840s).

François Furet, a former communist, brought together the new Anglophone work in a frontal assault on the Marxist interpretation. This 'revolutionary catechism,' 'Lenino-populist vulgate,' he argued, fundamentally distorted the history of the French Revolution by depicting it as a class struggle that anticipated communism. Furet claimed as his ancestors Tocqueville and Cochin; he did not reject democracy, but rather wanted to understand why democracy in the French Revolution developed into Terror or totalitarianism. The social underpinnings of the French Revolution could never explain this skidding out of control into Terror, he maintained; revolution was an autonomous political and ideological movement, not the expression of a social class (Furet, 1978). By 1989, the year of the bicentennial of the French Revolution, Furet's version of the French Revolution had become the new orthodoxy. The political balance had shifted, not only in France but almost everywhere else as well. The French Revolution no longer seemed a communist preserve.

Since 1989, as the polemics over Marxism have subsided, the work of historians from other countries has attracted more attention. American, British, German, and Italian historians now offer alternative views of their own. No one school, whether Left wing or Right wing, French or Anglophone, currently dominates the historiography of the French Revolution. The shift of political interest away from Marxism has opened up new areas of research, including gender relations, definitions of citizenship, and the role of the colonies. These new perspectives reflect concerns with human rights and globalization, and with the relationship between politics and political economy, and between the Revolutions and international relations, themes emphasized by scholars emphasizing the importance of the intellectual history of the 1790s.

Since the French Revolution began by enshrining the notion of 'the rights of man,' it has been a prime site for work on the dilemmas of universalism. Feminists have pinpointed the Enlightenment's masculinist bias as the chief cause of the failure of the revolutionaries to include women in their promise of citizenship; so-called Enlightenment writers, such as Montesquieu and Rousseau, used nature and reason to justify women's exclusion from the public sphere, and the revolutionaries simply drew the logical conclusion from these assumptions. Other feminists, however, have insisted that the Revolution opened up new opportunities for women, legalizing divorce, assuring equal inheritance to girls, and affording women the opportunity to establish political clubs. Unlike women, men who were Protestants, Jews, free blacks, mulattos, or even slaves gained political rights during the French Revolution. Yet even these enfranchisements revealed telling fault lines in the revolutionary conception of citizenship; the National Assembly, for instance, insisted that the Jews must give up their particular identity and corporate privileges as Jews in order to become French citizens. Since discussions of free black and slave rights inevitably involved the slave colonies, and since France's colony of Saint-Domingue (San Domingo) was home to the first successful slave revolt in history beginning in 1791, the interest in citizenship has also stimulated study of France's slave colonies. Because the French Revolution made revolution an acceptable element of politics, and because so many have seen the Revolution as the origin of the modern world, it remains of singular significance.

See also: American and French Revolutions: Impact on the Social Sciences; Bourgeoisie and Middle Classes, History of; Citizenship, Historical Development of; Democratic Theory; Enlightenment; Nationalism: General; Political Representation; Rousseau, Jean-Jacques (1712–78); Social Science, The Idea of; Socialism: Historical Aspects; Tocqueville, Alexis de (1805–59).

Bibliography

Arendt, H., 1963. On Revolution. Viking Press, New York.
Barnave, A., 1988. Introduction à la Révolution française. De la Révolution et de la Constitution. Presses Universitaires de Grenoble, Grenoble, France.
Buonarotti, P., 1828. In Conspiration pour l'égalité dite de Babeuf suivie du procès auquel elle donna lieu, et des pièces justificatives, 2 vols. A la Libraire Romantique, Brussels.
Burke, E., 1790. Reflections on the Revolution in France, and on the Proceedings in Certain Societies in London Relative to that Event, third ed. J. Dodsley, London.
Censer, J.R., Hunt, L., 2001. Liberty, Equality, Fraternity: Exploring the French Revolution. Pennsylvania State University Press, University Park, PA.
de Tocqueville, A., 1856. L'ancien régime et la Révolution, second ed. Michel Lévy frères, Paris.
de Tocqueville, A., 1977. Oeuvres complètes. In: Jardin, A., Lesourd, J.-A. (Eds.), Correspondance Tocqueville-Louis de Kergolay, vol. 13. Gallimard, Paris.
Desan, S., Hunt, L., Nelson, W.M., 2013. The French Revolution in Global Perspective. Cornell University Press, Ithaca.
Doyle, W., 2001. The French Revolution. Oxford University Press, Oxford.
Furet, F., 1978. Penser la Révolution française. Gallimard, Paris.
Furet, F., 1986. Marx et la Révolution française. Flammarion, Paris.
Godineau, D., 1988. Citoyennes tricoteuses: les femmes du peuple à Paris pendant la Révolution française. Alinéa, Aix-en-Provence, France.
Hunt, L.A., 1984. Politics, Culture, and Class in the French Revolution. University of California Press, Berkeley, CA.
James, C.L.R., 1938. The Black Jacobins: Toussaint L'Ouverture and the San Domingo Revolution. Secker & Warburg, London.
Kates, G., 2002. The French Revolution Recent Debates and New Controversies. Routledge, London.
Landes, J.B., 1988. Women and the Public Sphere in the Age of the French Revolution. Cornell University Press, Ithaca, NY.
Sonenscher, M., 2007. Before the Deluge Public Debt, Inequality, and the Intellectual Origins of the French Revolution. Princeton University Press, Princeton, NJ.
Stone, B., 2002. Reinterpreting the French Revolution a Global-historical Perspective. Cambridge University Press, Cambridge, UK.
Taine, H., 1885–1888. La Révolution, 3 vols. Hachette, Paris.

Freud, Sigmund (1856–1939)

Clifford Yorke, South Moreton, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 9, pp. 5798–5805, © 2001, Elsevier Ltd.

Abstract

Sigmund Freud (1856–1939) the founder of psychoanalysis, was born in Freiberg, Moravia. He qualified in medicine at the University of Vienna and published almost 200 neuroscientific papers. But a visit to Charcot in Paris aroused his interest in hysteria and hence psychological development, both normative and abnormal. His first major psychoanalytic work, *The Interpretation of Dreams* (1900), was partly founded on his self-analysis, using the technique of free association. Intensive clinical investigation led to his elaboration of the concept of a mental organization inaccessible to introspection, and to formulations of preconscious mental functioning frequently, though not invariably, accessible to consciousness. Freud was always prepared to modify his views in the light of further experience: the most radical changes were described in the *Ego and the Id* (1923) and included, *inter alia*, the concept of unconscious guilt. He wrote extensively on the application of psychoanalysis to nonclinical fields, including sociology, anthropology, group psychology, and religion. When the Nazis occupied Vienna in 1938, he left for London with members of his family, warmly welcomed by the British Psychoanalytical Society. Within 15 months he died from a longstanding and painful cancer of the jaw. He wrote until the end.

Family Background and Early Life

Sigmund Freud, neuroscientist turned founder of psychoanalysis, was born on May 6, 1856, in Freiberg, Moravia, now part of the Czech Republic. His father, Jakob (1815–96), a wool merchant, came from Galicia, and married Amalie (1835–1930, neé Nathansohn), 20 years younger, when he was 40. Jakob had two sons by a previous marriage, Emanuel and Philipp. Jakob was warm and affectionate, with a strong sense of humor, and Freud evidently admired and respected him. Amalie was a lively and attractive woman who was proud of Sigmund, her first-born.

The family multiplied quickly: five daughters and two more sons were born within 10 years. The first of these, Julius, died when 8 months old and Sigmund was aged 19 months. Both parents were Jews, with no religious affiliations, though a devout Catholic Nanny took Sigmund to Church. She left when he was two and a half. A nephew, John, a year older than Freud was very close to him: the two fought and played together. That childhood ambivalence, Freud reflected years later, must have had a profound effect on his character and on his ability to defend himself (Freud, 1900).

Financially, times were hard. In 1859 the family moved to Leipzig, and in the following year to Vienna. Freud missed the beauty of the countryside round Freiberg. Although the children were well looked after, poverty was not assuaged by rapid family growth. By 1875 family fortunes improved: Amelie's family were helping, making possible a move to a larger house. Freud had a room of his own that served as a study.

Freud's memories of his childhood as revealed in his writings are without early detail, but one unforgettable recollection was of urinating in his parents' bedroom while they were present. Reprimanding him, his father said bluntly: 'The boy will come to nothing!' Freud's wounded ambitions were reflected recurrently in dreams in which he listed his achievements as if to say: 'You see! I have come to something!' (Freud, 1900; see also Shengold, 1993).

Education and Early Interests

Freud's schooling began with his mother, until his father took over before sending him to a private school. Freud learned rapidly, and was reading Shakespeare (whom he loved all his life) from the age of eight years. When nine years of age, he won a place at the Sperl Gymnasium, where he was a distinguished pupil, passing out, at 17, with the distinction *'summa cum laude.'* He became an accomplished linguist, with a firm foundation in Latin and Greek, a sound knowledge of Hebrew, fluency in English and French, and good Italian. He corresponded with an adolescent friend in Spanish. He became one of the greatest stylists in the German language (Kaufmann, 1980). He studied at every opportunity. He loved the Arts, but took an early dislike to music, which he never felt able to appreciate. He was fond of walking, swimming, and skating. He liked travel, and made his first visit to England at the age of 19.

Early Career and Scientific Outlook

Freud (1925) said that curiosity and human concerns motivated his interest in medicine, but it was 'hearing Goethe's beautiful essay on Nature read aloud ... that decided me to become a medical student.' He was attracted strongly by Darwin – a necessary step towards his own evolutionary approach to the history of the individual's mental development. His neurological work was affected profoundly by the evolutionary approach of John Hughlings Jackson, the founder of modern neurology. Furthermore, Freud's admiration of Goethe as scientist as well as writer was life long. Goethe's belief in the importance of observation and reflection was profound, and his meticulous examination of a single sheep's skull in 1790 led to his deduction that the vertebrate's skull developed from vertebrae. This type of detailed study – applied by Freud to the individual case – is the essence of poetic (or romantic) science that stands in contrast with (though not

in opposition to) Newtonian (or classical) science. (Goethe's unfortunate opposition to Newton's theory of color does not vitiate this point.) In psychology, the corresponding distinction is between the ideographic (romantic) approach and the nomothetic (classical). Both methods have drawbacks and advantages; for Freud's psychology, in particular, the nomothetic approach has serious limitations. The distinctions have been discussed extensively by Kaufmann (1980) and by the founder of modern neuropsychology, Luria (1979). It is impossible fully to appreciate Freud's work outside this context.

Freud began his medical training in October 1873 and qualified MD in March 1881. The years were crowded: Freud dealt with the tedium of a year's compulsory military service by translating a book by John Stewart Mill (Jones, 1953). He was influenced by many illustrious figures, none more important than Ernst Brücke (1819–92) head of the Institute of Physiology. Freud, as a research scholar, studied the histology of nerve cells, and his work led to a number of significant conclusions. After qualification, Freud continued working at the Institute, where his love for research had full scope. He had already found 'rest and full satisfaction' there (Freud, 1925).

Neuroscience

But in 1882 Freud was still poor and, reluctantly, accepted Brücke's advice to practice medicine. He left the laboratory to start work in the General Hospital. He spent 6 months under Theodor Meynert (1833–92), the Professor of Psychiatry, whose work and personality Freud greatly admired. His research at the Institute of Cerebral Anatomy led to important publications. He began to study nervous diseases, and quickly established a reputation as an outstanding diagnostician. He was still urgently short of money: he was by now engaged to Martha Bernays (1861–1951), and the need to earn enough to keep a wife and family was compelling.

In the spring of 1885 Freud was appointed Lecturer (*Dozent*) in Neuropathology. Brcke's influence resulted in the award of a valuable Traveling Bursary, and Freud fulfilled an old ambition to visit Paris and learn from Charcot at the Salpêtrière. His offer to translate a new volume of Charcot's lectures into German was accepted by the great man, and Freud was soon admitted to Charcot's inner circle. Charcot's lectures and demonstrations were outstanding, but Freud was astonished by his investigations into hysteria. Charcot demonstrated that hysterical symptoms were genuine; that hysteria was common in men as well as women; and that hysterical paralyses produced by hypnosis were identical with spontaneous attacks. Charcot met skepticism with patience and firmness: once, faced with an objection made on theoretical grounds, he replied: 'Theory is all very well, but it doesn't stop facts from existing.' Freud never forgot that remark.

The Paris visit changed the direction of Freud's career. Charcot had broadened his grasp of neurology, but the neuroses had captured his imagination. He left Paris at the end of February 1886, intent on further investigations of hysteria. But before he returned to Vienna he stayed for some weeks in Berlin to study nervous disorders in children, and this important detour led in time to the publication of a number of monographs. Freud's report on his visits, to his sponsors in Vienna, is a model of clarity and concise exposition (Freud, 1956).

Indeed, Freud's interest in neuropathology by no means stopped abruptly. Important papers lay ahead: observations on hemianopsia in two very young children (Freud, 1888) and his monograph on *Aphasia* (Freud, 1891b) occupy a permanent place in the history of aphasiology. Freud had already abandoned the strict localisation theory of brain functioning – to which Meynert strongly adhered – in favor of the hierarchical model, with its concomitant theory of evolution and dissolution of the nervous system, advanced by Hughlings Jackson (e.g., 1884). Freud recognized the importance of these major advances, which were otherwise scarcely appreciated for another quarter of a century. With his young friend Oscar Rie he wrote a work on the unilateral cerebral palsies of children – a book still of repute in neuroscience (Freud, 1891a). Freud's overall contributions to the field were massive, and he published around 200 papers.

On his return to Vienna, however, Freud was trying to build up a practice to support a family: he married Martha in September 1886, starting a life long partnership of unusual harmony and mutual support. Both were devoted to their six children: in order, Mathilde, Martin, Oliver, Ernst, Sophie, and Anna – the only one to follow Freud into psychoanalysis. The first 5 years of marriage covered the closing phase of neurological work, and saw the growth of a friendship with the distinguished physician Joseph Breuer (1842–1925), with profound effects on Freud's career.

The Birth of Psychoanalysis

Between 1880 and 1882 Breuer had treated the now famous Frau Anna O. for severe hysterical symptoms, including multiple limb paralyses, disturbances of sight and speech, contractures, an intense cough, and anorexia. Accounts of the illness, its treatment, outcome, and significance for the foundations of psychoanalysis, and its effect on the Breuer–Freud friendship vary (e.g., Breuer, 1882; Breuer and Freud, 1895; Freud, 1914, 1925, 1932; Jones, 1953; Gay, 1988; Ellenberger, 1972; Hirschmüller, 1989), but, at the risk of distortion by brevity, the present summary seems safe. Anna O. suffered from states of autohypnosis alternating with near-normality. When, under self-hypnosis, she recounted the history of one of her symptoms, recalling the details of its first appearance, Breuer was astonished to find that it disappeared. Under induced hypnosis other symptoms were treated in similar fashion, and the method was christened the 'talking cure.' Breuer told Freud about it in 1982, and during the next two years they discussed it repeatedly. But, in Paris, Charcot showed little interest in the story, and Freud allowed the matter to lapse.

Back in Vienna, Freud treated a number of hysterical patients by hypnotic suggestion. But 'recovery' was too often transient, and Freud's interest in Breuer's treatment of Anna O. revived. When her first symptom appeared, the patient was nursing her dying father, and her recall resulted in 'abreaction' (the release of pent-up emotion attached to the event).

Adopting this 'cathartic method' over several years, Freud repeatedly confirmed Breuer's findings. The links with traumata were hitherto unconscious.

A preliminary communication by Breuer and Freud (1893) preceded the *Studies on Hysteria* (Freud, 1895d). Breuer contributed a theoretical chapter and an account of Anna O; Freud, four case histories. Differences of opinion became a strain in 1894, and the collaboration ceased in 1896. Freud valued Breuer's achievement, but was thinking on different lines. Breuer believed that pathological processes arose during 'hypnoid states' (e.g., Anna O's autohypnosis). For Freud, this begged the question of how 'hypnoid states' arose. He considered that psychoneurosis involved psychic defences against unacceptable internal states of 'emotional excitation,' habitually sexual in kind. The actual neuroses, however, were states of anxiety arising from unrelieved sexual excitation linked with abstinence or *coitus interruptus*. Breuer could not accept these views; he had failed to find any sexual component in Anna O's disturbance. But Breuer had never recognized Anna O's strong attachment to him ('transference') as sexual; alarmed by its intensity, he brought the treatment to an end. Anna O. responded to this loss with a phantom pregnancy (Freud, 1932). Breuer continued to see mentally ill patients, but never again used the cathartic method.

In time, Freud abandoned the use of hypnosis altogether. The power of transference (from the patient's unrecognized re-experience, in the treatment process, of infantile relationships) meant that excellent clinical progress could still be vitiated by disturbances in the doctor/patient relationship. That recognition profoundly affected the future development of psychoanalysis.

For a time, Freud used encouragement, persuasion, and similar measures as aids to treatment, but increasingly found them unnecessary. They were replaced slowly by the developing technique of free association, which took over completely by 1897–98. The patient lay on a couch (as before) with the analyst sitting behind him, but was now expected to try to say everything that entered his mind, irrespective of apparent irrelevance, distastefulness, or indiscretion. The analyst listened impartially, without giving more weight to any one utterance than he did to another. This major shift in clinical method had surprising results. Conflicting mental forces, undetected by hypnosis and related devices, were now revealed. Memories of striking internal and external mental experiences in early life had been denied access to consciousness because they had been 'alarming or painful or shameful by the standards of the subject's personality' (Freud, 1925). Memory was actively opposed by the force of repression, experienced by the clinician as resistance to unwelcome material. Forces striving to break through this barrier stemmed from the drives, founded substantially in infantile sexuality.

Psychoanalytic Career

Until the spring of 1897 (Jones, 1953) Freud believed his female adult patients whose associations pointed to seduction in childhood, usually by the father, as the traumatic root of their illness. But what at first presented as memory was found increasingly to be fantasy. Seduction in childhood undoubtedly occurred, but could no longer be invoked as the root of all neurosis. As it stood, Freud's theory was wrong. This devastating conclusion nonetheless pointed to the importance of fantasies in mental life, and to the realization that fantasy that is dynamically unconscious, that is, inaccessible to introspection, had all the force of reality since it could not be tested in the light of objective experience. A second, fateful, event, did more to foster psychoanalytic thinking than anything else: Freud courageously (and painfully) undertook his self-analysis, beginning in 1897 and continuing systematically through and beyond the writing of *The Interpretation of Dreams* (1900). With interruptions, it was life long.

Freud's letters to his friend Wilhelm Fliess (1858–1928), an ear, nose, and throat specialist living in Berlin, provide a unique source of information about these events. Never intended for publication, they were rescued by Princess Marie Bonaparte of Greece (the letters from Fliess have not survived). Starting in 1887, they set out in astonishing detail Freud's reflections, doubts, and formulations as they arose, and can be followed until after the turn of the nineteenth century. An edited version (Freud, 1954) was followed by the complete letters (Masson, 1985). The demise of the seduction theory and Freud's comments on his self-analysis are described in both.

The self-analysis was crucial for the development of basic psychoanalytic concepts. The richness and complexity of infantile life, including its sexuality, was not fully appreciated until well after *The Interpretation of Dreams* saw the light of day. The Oedipus complex, for example, is described in that book, but its fuller significance was underlined only in footnotes added much later. But Freud regarded the dream book as his single most important work. With copious illustrations from his self-analysis and clinical experience with patients, the study goes beyond the meaning of dreams and approaches the baffling problem of how the mind works.

The theory is involved, but a few key points may be mentioned. The function of dreaming is the preservation of sleep. Stimuli arising from outside the organism (noises, the pressure of bedclothes) or from within (pain, excitation), could often be contained when switched into nonmotor channels (dreaming). The formation of the dream is more complex. In normal waking life, accessible mental content is either conscious at any given moment, or is preconscious – capable of reaching awareness with varying degrees of difficulty. But some mental content is dynamically unconscious – barred from reaching the preconscious by repression. During sleep, this barrier is relaxed, and a partial regression to earlier forms of thinking allows normally unacceptable childhood wishes to find expression, though only in disguise. Even in sleep a censorship is partially operative, and the result is a distorted representation of infantile wishes that gives the dream its seemingly bizarre character. Any semblance of coherence is imposed during waking and recollection: the repression barrier is restored and the dream is (unwittingly) clothed in a measure of rationality.

The recollected (manifest) dream content arises from several sources. Stimuli that threaten sleep are the most obvious, but events and experiences of the day before

(day residues) are also used as building blocks. These have no intrinsic importance: they simply provide convenient representations for forbidden wishes. But emerging wishes are themselves distorted by primitive thinking before they are disguised further by the censorship. For Freud, the dream was a hallucinatory wish fulfillment, though generally far removed from the wishes of everyday appearance.

Do dreams provoking anxiety or portraying punishment also conceal wishes? Freud argued that, even in sleep, a wish can be so strong that its forbidden character can still be recognized and occasion distressing anxiety. In punishment dreams, the transgression portrayed brings with it a wish for punishment. This became clearer still when Freud, later, had conceptualized an unconscious sense of guilt.

The model of the mind outlined in the dream book was the foundation for all Freud's later theoretical formulations. Since it was written during his self-analysis, his pride in it also reflected a new self-confidence. Jones (1955) claims that, at the age of 45, Freud had reached full maturity: intellectual development, always precocious, had not hitherto been matched emotionally. For many years there had been 'continual periods of restlessness, uncertainty, instability and even more definitely neurotic disturbance.' A deep self-assurance had been hidden behind 'strange feelings of inferiority, even in the intellectual sphere.' He had over-idealized men who were, certainly, important to him; of these, Breuer had already withdrawn his support, and the relationship with Fliess was loosening. Self-analysis had revealed the neurotic nature of the tie.

Intellectual loneliness did not diminish productivity. *The Psychopathology of Everyday Life* (Freud, 1901) dealt with unconscious motivations in forgetting, slips of the tongue and pen, as well as some 'accidents.' The *Fragment* of Freud's analysis of 'Dora' (1905) described her treatment in 1900. *Three Essays on the Theory of Sexuality* (Freud, 1905) met with ferocity hard to imagine today. Freud had widened the concept of sexuality: all existing definitions failed to fit the facts and could not, for example, account for the sexual perversions. Sexuality was now understood within a much wider, developmental perspective. It was considered diphasic: development from infancy to roughly the age of five, centered around bodily erotogenic zones, was followed by a period of comparative quiescence (latency) before a second stage of sexual efflorescence started around puberty. *Jokes and their Relation to the Unconscious* (Freud, 1905) arose out of the work on the dream book.

The Growth of the Discipline

By the time these works appeared Freud was no longer intellectually isolated. Already, in 1902, two young Viennese doctors, Kahane and Reitler, and, more famously, Stekel and Adler were intrigued by psychoanalysis. Freud invited them to meet him for discussions; they gathered together on Wednesdays as 'The Psychological Wednesday Society.' This was the nucleus of an expanding group renamed, in 1908, the Vienna Psychoanalytical Society. By then much had happened. In 1904 Eugene Bleuler, Professor of Psychiatry in Zurich, wrote to inform Freud that he and his staff, most notably his Chief Assistant, C. G. Jung, had been making use of psychoanalytic ideas for the past two years. (In particular, Jung had applied them in his word-association tests.) By 1906 Freud and Jung were corresponding, and in 1907 Jung visited Freud in Vienna. The friendship grew rapidly and when, in 1910, the International Psychoanalytic Association was formed, Freud asked Jung to be President. Jung's bearing, intellect, qualities of leadership, psychiatric background, and dedication seemed suitable qualifications. Freud was attracted by his personality and regarded him as his 'son and heir.' Jung was greatly impressed by Freud's penetrating intellect. The rise and fall of their relationship is chronicled in their correspondence (McGuire, 1974).

Otto Rank had joined the group in 1906; and Karl Abraham, a major figure in the psychoanalytic movement, had long been fascinated by Freud's work when he visited him in 1907. Again, the correspondence between the two gives a vivid picture of their relationship; it continues until Abraham's premature death in 1925 (Abraham and Freud, E. 1965). In 1907, too, Sandor Ferenczi visited from Budapest, establishing himself the following year as a major figure in the discipline. With the visit of Ernest Jones in 1908, and Hans Sachs in 1910, Freud had a close circle of friends and collaborators.

In 1908, Jung organized an international conference in Salzburg. Freud's work was now discussed widely, in Europe, the USA, and Australia. In 1909 Freud lectured at Clark University, Massachusetts, accompanied by Ferenczi, Jones, and Jung. In 1910 a second international congress was held at Nuremberg, organized by Jung. But internal dissent soon surfaced. In 1911 Adler defected: to Freud's dismay, he refused to try to resolve their conceptual differences. Stekel, gifted, charming, but scientifically unreliable (Jones, 1955) left about the same time. Although Freud disliked leadership, he was a fatherly figure for his young colleagues, and bickering and rivalry, especially among the Viennese (cf. Jones, 1955) threatened stability. Freud's obvious preference for Jung was resented. But by 1912 Jung was showing signs of disaffection. Many Swiss psychiatrists denounced Freud's theory of sexuality, and, during a course of lectures in the United States Jung omitted any reference to it. Personal relations between the two men cooled, but it was not until 1914 that Jung left the movement on the grounds of incompatible views. The story of the rupture is complex: the Freud/Jung letters and the account by Jones (1955) repay study.

Freud's continuing productivity defied these dissensions. He addressed the problem of the sexual enlightenment of children (Freud, 1907), discussed children's sexual theories (Freud, 1908), and considered the relationship of the character traits of orderliness, parsimony, and obstinacy to anal eroticism – a paper greeted with indignation when it appeared (Freud, 1908). He tried to show that psychoanalysis could contribute to the understanding of art in discussions of Jensen's *Gradiva* (Freud, 1907) and of the relationship between creative writing and day dreaming (Freud, 1908). He wrote many other papers in these industrious days.

Two extensive clinical studies appeared in 1909. The first (Freud, 1909) describes the very first analysis of a child, though one undertaken by proxy. The successful treatment of 'Little Hans,' who had developed a phobia of horses, was carried out by his father, under Freud's supervision. Hitherto, the

childhood roots of neurosis in adults had been understood through a process of reconstruction; with Little Hans those sources could be identified at the time they were operative. Freud believed that the analysis could not be conducted by a stranger, and that the affection and authority of the father (in this case, a physician, too) made treatment possible. Child analysis as a speciality was yet unborn. It was many years before the possibility of directly analysing young children, under favorable conditions, was recognized.

The second patient suffered from an obsessional neurosis (Freud, 1909). The symptoms included fears that harm would befall his father and a woman he admired; compelling impulses, one of which was to cut his throat with a razor; and self-imposed irrational prohibitions. He was horrified but attracted by an account of a punishment in which rats ate their way up the victim's rectum. The unconscious sadism underlying his 'fascination' with the torture was among the features on which the analysis cast light. A deeper psychoanalytic formulation of obsessional neurosis emerged from the treatment. Later critics of this case report, and that of Dora, have suggested that Freud neglected the role of the mother in the genesis of the disorders. But Dora's treatment was brief, unfinished, and took place when clinical analysis was still in its infancy. In the case of the Rat Man, the criticism, when sound, is largely based on hindsight.

Another case study (Freud, 1911) concerned a subject whom Freud had never seen. Dr Schreber suffered from a paranoid state, and was under care for 6 years. He later wrote his *Memoirs of My Nervous Illness*, though Freud read them only in 1910. The illness fell into two phases: distressing delusions of persecution preceded delusions of grandiosity. Freud made an ingenious and fascinating reconstruction of the course of the illness. His work stimulated much clinical research. Psychoanalytic theories of psychotic illness fall into two broad groups: either it is part of a neurosis–psychosis continuum, or, unlike neurosis, it involves the loss of higher psychic functions and the re-emergence of more primitive ones, hitherto kept in check by the lost capacities. Many claim that Freud supported the first (unitary) theory; others the second (discontinuity theory). The controversy deserves close attention and is summarized elsewhere (Yorke et al., 1989).

In 1910 Freud (1910) published a study of a childhood memory of Leonardo da Vinci, and the first of three essays on the psychology of love. In 1911, the first of a series on psychoanalytic technique, and from 1910 to 1912 some short papers on theory that paved the way for later and larger works. An excursion into social anthropology, *Totem and Taboo*, was written between 1912 and 1913. Before war broke out in the summer of 1914, Freud's works included a history of the psychoanalytic movement (Freud, 1914), giving his version of the disagreements with Adler and Jung; a fascinating paper on the *Moses* of Michelangelo (Freud, 1914); a discussion of the scientific status of psychoanalysis (Freud, 1913); and a wide ranging paper on narcissism from which so many ideas were later developed (Freud, 1914). These were productive years indeed.

During the war Freud was anxious about two of his sons involved in it. Martin fought in Galicia and Russia; Ernst against Italy after that country entered the conflict. Martin was decorated for gallantry. Medical reasons prevented Oliver's acceptance by the army. Optimism about the outcome diminished after 1915. In 1916 the food shortage began to bite and got worse as the war continued. Freud had few analytic patients. Contact between analysts faced obstacles, some severe, even in the central powers.

For all the difficulties, Freud continued to write. He wrote 12 papers on the theoretical basis of psychoanalysis between 1914 and 1917: only five were published, and a sixth manuscript was found in 1983. There is no trace of the others. *Thoughts for the Times on War and Death* (Freud, 1915) anticipated the correspondence with Einstein (Freud, 1933) *Why War*? At the University, in the winter terms of 1915–16, and 1916–17, Freud gave a total of 28 lectures on psychoanalysis, and wrote them down, from memory, later (Freud 1916–17). Important changes were made in the *New Introductory Lectures* (Freud, 1933), but, taken together, the two series offer an excellent entry into Freud's psychological thinking. The most extensive of Freud's case histories (the *Wolf Man)*, written in 1914, was published much later (Freud, 1918). In 1919 the baffling problem of masochism (sexual pleasure in pain), and a psychoanalytic look at 'the uncanny' (Freud, 1919) were among the many subjects Freud addressed.

The return of many old colleagues to Vienna in 1919 ended wartime isolation. For some time life was hard, but psychoanalysis continued to grow worldwide. In 1919 Martha Freud contracted influenza leading to pneumonia and was ill for many months; the following year Sophie Freud died from a similar cause. Freud grieved deeply.

In 1923 Freud's heavy cigar smoking seemed responsible for a growth affecting the jaw and palate, removed surgically. This was the first of many operations for cancer that Freud survived over the next 16 years, bearing discomfort and pain, often severe, with fortitude. But he continued to work. In that same year, he set out major changes in psychoanalytic theory (Freud, 1923). He had already felt obliged to give aggression (more accurately, destructiveness) the same status as libido. For clinical and theoretical reasons the structure of the mental apparatus was redefined. The ego was now the executive mental agency, holding a balance between the conflicting demands of reality, the id, and the superego. The id strove for immediate drive gratification; its mentation followed the primitive primary process; and its content was dynamically unconscious. But it intruded into an unconscious part of the ego where it encountered checks and controls against inappropriate discharge. The superego was thought of as a greatly extended conscience, based on internalized but modified parental aims and proscriptions, but a substantial part of its operations were unconscious. Its functions included that of the earlier censor. With the aggressive drives of the id partly at its disposal, it put pressure on the ego to withstand those inner urges that ran counter to its prohibitions, aims, and ideals. It was responsible for the sense of guilt, conscious, and unconscious. It must be added that higher mental functions such as conscious thinking and reflection (secondary process), reality testing, and judgment were functions of the ego.

Freud continued to elaborate and modify this model (it should not be reified). He made radical changes in his theory of anxiety (Freud, 1926); extended his work on masochism (Freud, 1924); and in the field of social psychology added to his earlier study of the psychology of groups (Freud,

1921) the far reaching *Civilization and its Discontents* (Freud, 1930). The work of his female colleagues and his own clinical experience led him to re-examine the many problems of female sexuality (Freud, 1931, 1933). This work aroused opposition, particularly among feminists: Eissler (1977) countered some of these criticisms, but there have been many subsequent developments. It is worth remembering that Freud's close friendships included many women (see, e.g., Freud and Andreas-Salome, 1972); and many found his work did much to reduce guilt about sexuality. In *Analysis Terminable and Interminable* Freud (1937) emphasized the therapeutic limitations of psychoanalysis, but he was apprehensive about the possible reception of *Moses and Monotheism* (Freud, 1939), which extended his inquiries into early social organization *(Totem and Taboo* 1912–13, *Group Psychology* 1921).

Exile: England

In 1938 the Nazis entered Vienna. Freud's books had already been burned in Germany (Freud thought this more advanced than burning him). Through the good offices of Princess Marie and Ernest Jones Freud, together with Martha, their daughter Anna, and others, were allowed to leave for London. Freud was quite ill and tired by the journey, but was pleased by the warm reception that awaited him. The British Psychoanalytical Society gave him a house at 21 Maresfield Gardens in Hampstead. Anna Freud was his devoted nurse. His remarkable *Outline of Psycho-Analysis* (Freud, 1940) was written there. It was there, too, that that death ended his considerable pain and suffering on September 23, 1939. His ashes remain at Golder's Green Crematorium.

Of his many honors, he particularly valued the award of the Goethe Prize for Literature (1930) and his Corresponding Membership of the Royal Society (1936).

Freud's Impact on Twentieth-Century Thought and Culture

There is still a widespread view that Freud's impact on twentieth-century thought, on the arts, sciences, social attitudes, and beliefs, has been far-reaching, even immense. Almost every intellectual discipline, it is said, has been influenced by his ideas. The importance of unconscious motivation is acknowledged widely. Furthermore, many of his concepts – such as the 'Freudian slip' of pen or tongue – have passed into the common language of everyday social intercourse; and his views on sexuality have had a liberating impact on social attitudes and broadened sexual mores.

This view is by no means universal. Freud is undeniably a household name, but more than any other major figure, is the subject of widely varying and conflicting claims. These differences may arise in part from the nature of his subject, for it may be said, with a certain truth, that every man is his own psychologist. They may also arise from an investigative method difficult to understand without personal acquaintance. Freud's neuroscience aside, his principle field of inquiry was subjective and internal, and not readily open to methods appropriate to objective and external research. Many academic psychiatrists and psychologists consider that both Freud's clinical method and his theoretical constructs lack empirical support: in this, they generally equate empiricism with mensuration.

This fact stands in the way of dispassionate assessment even for those actively and intelligently involved in debate. The casual bystander, faced with controversy on a substantial scale, may be simply bemused or follow an opinion in line with personal appeal. And if the intelligent and industrious layman turns to Freud's own writings, he will still encounter significant obstacles. Freud's ideas cannot be picked up piecemeal, but comprehensive reading is difficult without informed guidance. Moreover, reliable introductory works are hard to come by. So it seems unlikely that the forces behind the 'Freudian slip,' and the distinctions between the preconscious and the unconscious, are understood widely. (The term 'subconscious' is more likely to be heard – a word Freud never used except by way of criticism.) Finally, while the liberalization of sexual attitudes may owe something to Freud, social, economic, and political forces are surely of greater importance. These include the availability of reliable contraception.

Indeed, Freud's views on sexuality, when they are not bowdlerized, are misunderstood widely. But the wider point to be made is this. Most people, in all sections of society, who have any interest in Freud, do not know what he really said. They only know what they think he said. And that is just as true for most critics and commentators who seemingly write with authority, including psychologists, 'hard' scientists, sociologists and art critics, whether the views they express are positive or negative. The viewpoint expressed here may at first seem novel. It was presented in extended form at a meeting of the *Psychoanalytic Heritage Foundation* on October 6, 1992, when it became evident that many Freud scholars were thinking on similar lines. At the recent Conference in London (see below), Douglas Watts stated from the platform that psychoanalysts needed to do more to make the general public more familiar with Freud. 'At present all they have is a caricature.'

For reasons that are not dissimilar, though highly complex, even psychoanalysts have widely diverse approaches to both theory and clinical practice and to the weight they attach to basic Freudian principles (Yorke, 1995). Psychoanalysis is avowedly pluralistic (Wallerstein, 1988, 1990). And although psychotherapy is often said to derive from the 'talking cure,' similar considerations apply.

In a few instances, Freud's influence, directly or indirectly, on other disciplines is evident, though localized. Disciplines not already mentioned include: education (e.g., many members of the Montessori movement); criminology (The Institute for the Study and Treatment of Delinquency in London); family law in relation to children (e.g., liaison between the Child Study Center and the Law School at Yale University). Properly understood, Freud's work and its true derivatives is capable of fostering advances in many intellectual and sociological fields.

It is pleasing to end on a positive note – in relation to Freud's first love, neuroscience. Freud always hoped that the underlying basis for mental activity in terms of the functioning brain would, one day, yield to investigation. For the greater part

of a century that hope seemed slim indeed. But the last few years have brought startling developments. Solms (1997), for example, in a fascinating series of experiments, showed convincingly that the neuropsychology of dreaming was very much in line with the argument set out in *The Interpretation of Dreams*. Work on the neuropsychology of affects, for example, attracted interest from an increasing number of psychoanalysts, and in 1999 a new scholarly journal, *Neuro-psychoanalysis,* was founded. In the first issue, Solms and Nersessian (1999) demonstrated, in a wide-ranging survey, the extensive nature of Freud's contributions to affect theory. On July 21–23, 2000, the first annual Neuro-psychoanalysis Conference was held in London entitled *Neuropsychology and Psychoanalytic Persepectives on Emotions,* addressed by the world authorities Jaak Panksepp, Antonio Damasio, and Mark Solms, and introduced by Oliver Sacks. The meeting was a heartening success, and the compatibility between the findings of the two disciplines was seen to be remarkable.

All references to Freud, other than those in the bibliography, are to the *Standard Edition of the Complete Psychological Works of Sigmund Freud* (1953–74) Strachey J (ed.) 24 Vols. Hogarth, London. For specific volumes and page numbers see Vol. 24.

See also: Arendt, Hannah (1906–75); Bleuler, Eugen (1857–1939); Bowlby, John (1907–90); Dreaming, Neural Basis of; Erikson, Erik Homburger (1902–94); Foucault, Michel (1926–84); Hypnosis, Psychiatry of; Hysteria: History and Critiques; Psychoanalysis, History of; Psychological Treatment, Effectiveness of; Transference in Psychoanalysis: Classical, Contemporary, and Cultural Contexts.

Bibliography

Abraham, H.C., Freud, E.L. (Eds.), 1965. A Psycho-analytic Dialogue: The Letters of Sigmund Freud and Karl Abraham. Hogarth, London.

Breuer, J., 1882. Report to Bellevue Sanatorium, Kreuzlingen, on Bertha Pappenheim. In: Hirschmüller, A. (Ed.), 1989. The Life and Work of Joseph Breuer. New York University Press, New York.

Breuer, J., Freud, S., 1893. On the Psychical Mechanism of Hysterical Phenomena: A Preliminary Communication. Hogarth, London.

Breuer, J., Freud, S., 1895. Studies on Hysteria. Hogarth, London.

Eissler, K.R., 1977. Comments on penis envy and orgasm in women. Psychoanalytic Study of the Child 32, 29–83.

Ellenberger, H.F., 1972. The story of 'Anna O': a critical review with new data. Journal of the History of the Behavioral Sciences 8, 267–279.

Freud, S., 1891b. Zur Auffassung der Aphasien [trans. Stengel E as *On Aphasia* 1953]. Imago, London.

Freud, S., 1932. Letter to Stefan Zweig, June 2, 1932. In: Freud, E.L. (Ed.), Letters of Sigmund Freud, 1873–1939. Hogarth, London.

Freud, S., 1954. The Origins of Psycho-analysis: Letters to Wilhelm Fliess, drafts and notes: 1887–1902. In: Bonaparte, M., Freud, A., Kris, E. (Eds.). Imago, London.

Freud, S., 1987. A Phylogenetic Fantasy. In: Grubrich-Simitis, I. (Ed.). Belknapp, Cambridge, MA.

Freud, S., Andreas-Salome, L., 1972. Letters [trans. Robson-Scott W, Robson-Scott E]. Hogarth, London.

Gay, P., 1988. Freud: A Life for Our Time. Dent, London and Melbourne.

Hirschmuller, A., 1989. The Life and Work of Joseph Breuer: Physiology and Psychoanalysis. New York University Press, New York.

Jackson, J.H., 1884. Evolution and dissolution of the nervous system. In: Taylor, J. (Ed.), Selected Writings of John Hughlings Jackson. Basic Books, New York.

Jones, E., 1953. Sigmund Freud: Life and Work, vol. 1. Hogarth, London.

Jones, E., 1955. Sigmund Freud: Life and Work, vol. 2. Hogarth, London.

Jones, E., 1957. Sigmund Freud: Life and Work, vol. 3. Hogarth, London.

Kaufmann, W., 1980. Discovering the Mind, vol. 3. McGraw-Hill, New York.

Luria, A.R., 1979. The Making of Mind: A Personal Account of Soviet Psychology. Harvard University Press, Cambridge, MA.

McGuire, W. (Ed.), 1974. The Freud/ Jung Letters [trans. Manheim R, Hull R E C]. Hogarth & Routledge, London.

Masson, J.M. (Ed.), 1985. The Complete Letters of Sigmund Freud to Wilhelm Fliess: 1887–1904. Belknap Press of Harvard University Press, Cambridge, MA.

Shengold, L., 1993. 'The Boy Will Come to Nothing!' Freud's Ego Ideal and Freud as Ego Ideal. Yale University Press, New Haven, CT.

Solms, M., 1997. The Neuropsychology of Dreaming: A Clinico-Anatomical Study. Erlbaum, Mahwah, NJ.

Solms, M., Nercessian, E., 1999. Freud's theory of affect: questions for neuroscience. Neuro-Psychoanalysis 1, 5–14.

Wallerstein, R.S., 1988. One psychoanalysis or many? International Journal of Psycho-analysis 69, 5–21.

Wallerstein, R.S., 1990. Psychoanalysis: the common ground. International Journal of Psycho-Analysis 71, 3–20.

Yorke, C., 1992. Freud's impact on the twentieth century. Paper read to the Freud Literary Heritage Foundation, New York, October 6.

Yorke, C., 1995. Freud's psychology: Can it survive? Psychoanalytic Study of the Child 50, 3–31.

Yorke, C., Wiseberg, S., Freeman, T., 1989. Development and Psychopathology: Studies in Psychoanalytic Psychiatry. Yale University Press, New Haven, CT.

Freyre, Gilberto (1900–87)

Peter Burke, Emmanuel College, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Gilberto Freyre (1900–87) was not a professional academic, although he held appointments as a visiting professor of sociology and anthropology. However, he made contributions to both disciplines, as well as drawing upon economics, geography, and psychology in his classic studies of the social and cultural history of Brazil, notably the trilogy he published on the subject between 1933 and 1959.

Gilberto Freyre (1900–87) was born in Recife, in the state of Pernambuco, in northeastern Brazil, into a family of the professional middle class. His father, Alfredo, taught economics, while his mother came from the plantation-owning elite. He went to the American Baptist College in Recife, and was converted to Protestantism – for a short time. To complete his studies his family sent him to the United States, first to Baylor University in Waco (1918–20), and then to the Faculty of Political Science at Columbia (1921–22). Returning to Recife in 1923, he remained there for the rest of his long life, apart from a brief exile, in 1930, and visits to Europe, the United States, Latin America, and, in 1951–22, at the invitation of António Salazar, to Goa and the Portuguese colonies in Africa.

Freyre would have hated to be called a 'social scientist.' In his own eyes he was a man of letters, who published poems and two 'seminovels,' although he usually wrote nonfiction. He was not a professional academic, although he held several appointments as a visiting professor both in Brazil and in the United States, at Stanford in 1931, at the University of Michigan in 1939, and at the University of Indiana in 1944, where he offered a general interpretation of Brazil (Freyre, 1945a). He preferred freedom and leisure to a permanent post in a university and supported himself and his family mainly by writing for (and sometimes editing) newspapers, especially the *Díario de Pernambuco* and *A Provincia* (both published in Recife) and *O Cruzeiro*, the most important Brazilian weekly in the 1950s.

Freyre was a public intellectual, involved in politics. From 1926 to 1930, he was cabinet secretary to the governor of Pernambuco, Estácio Coimbra, and went into exile with him after the revolution of 1930 that brought Getúlio Vargas to power. From 1946 to 1950, he was a deputy in the Constituent Assembly, his party being the National Democratic Union that opposed the Vargas regime. He was denounced as a Communist and described himself as opposed to anti-Communism. However, in 1964, Freyre came out in support of the military coup of 31 March, thus losing some old friends. Thus, his political position, difficult to define, veered between radical (or as he would say, 'anarchist') and conservative.

Freyre was a polymath who wrote about art, literature, current affairs, the social sciences, and, above all, social and cultural history. He was an eclectic thinker, soaking up ideas like a sponge, and an omnivorous reader, both of literature (especially English literature) and the social sciences, beginning when he was young and keeping up with new developments until his 70s. He was also an active reader, turning down pages, underlining and writing in the margins of his books. The books that he discussed and annotated most carefully provide valuable clues to his development and to the approaches to the social sciences that interested him the most (Burke and Pallares-Burke, 2008).

The books he read on the subject might be divided into three archaeological strata. In the first place, old books published before he became an adult. In sociology, these works included Auguste Comte, Herbert Spencer, Frédéric Le Play, Gabriel Tarde, Emile Durkheim, Georg Simmel (whose essayistic approach he found appealing), Thorstein Veblen (whom he described as a social scientist 'with a scientific and even a poetic imagination') on the leisure class, and Max Weber on the Protestant ethic. In geography, Friedrich Ratzel, Paul Vidal de la Blache, Jean Brunhes, Ellen Semple, and Ellsworth Huntington. In anthropology, Edward Tylor and Alfred Haddon, Bronisław Malinowski, Marcel Mauss, and the early work of Franz Boas. In psychology, Wilhelm Wundt, William James, Sigmund Freud, and Havelock Ellis. More important for his thinking than most of these social scientists, however, were the ideas of John Ruskin, William Morris, and (on race mixture) the writer Lafcadio Hearn.

The second of the three strata was the work of Freyre's contemporaries: Robert Park, William Thomas, Gabriel Le Bras, and Georges Gurvitch, 'one of the most important sociologists of our time,' in sociology; Edward Evans-Pritchard in anthropology; Carl Sauer and Pierre Gourou in geography; and Gordon Allport in social psychology. In the last of the strata, one finds books by a younger generation: Clifford Geertz, Marshall Sahlins, R.D. Laing, Liam Hudson, Jean Duvignaud, and Alain Touraine.

In what follows, Freyre's studies of and his contributions to geography, psychology, anthropology, and, above all, sociology and history will be discussed discipline by discipline, followed by a more general assessment of his contribution to social and cultural theory.

Geography

Freyre was well read in geography. He discussed climatology and soil science on occasion but gave most attention to human geography, especially cultural regions and what Boas called their 'ecotypes' (Freyre was a Pernambucan patriot as well as interested in regionalism in general). The geographer who was most important for his work was Carl Sauer, whom he knew,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61199-6

and whose cultural and historical approaches resonated with his interest in the environment and its history, growing out of a concern, already expressed in the 1920s, for the 'devastation of the forest' in Brazil. Already visible in his best known work, *The Masters and the Slaves* (1933), this interest took center stage in *Nordeste* (1937), an essay in what the author called 'social ecology' that examined the effects of the concentration on the cultivation of sugarcane on the landscape, the animals and the people of his native region. What Freyre called 'the drama of monoculture' is emplotted as tragedy or as 'social pathology': the pollution of rivers, the poisoning of fish, and so on. *Nordeste* replaced what the author called the system of 'patriarchal economy' in its social and cultural contexts. This work, which has never been translated into English, is a small masterpiece.

Psychology

Freyre's eclecticism is perhaps most visible in his interest in and use of a variety of approaches to psychology, from William James via Gordon Allport to R.D. Laing. In the case of Allport, whom he met, it was the use of personal documents for the analysis of personality that most attracted Freyre, allowing him to engage in what would later be known as 'psychohistory.' Freyre often referred to psychoanalysis but was somewhat ambivalent about it, attracted to the explanatory system yet skeptical of its claims, whether made by Freud or Jung, Karen Horney or Erik Erikson. He criticized colleagues such as Artur Ramos for excessive fondness for psychoanalytical explanations and he gradually retreated from this approach himself. All the same, in the 1930s, Freyre interpreted the relations between masters (and mistresses) and their slaves in terms of sadomasochism. He emphasized what he called the 'psychoanalytical function' of confession and carnival, and noted 'the Brazilian inferiority complex.' He not infrequently remarked on split personalities, citing Robert Louis Stevenson's Dr Jekyll and Mr Hyde rather than the psychological literature. He described the rebel António Conselheiro as paranoid, and the bishop of Olinda (near Recife), Frei Vital, as 'sublimating' his love for his mother into his love for the Church.

Anthropology

Freyre's work was closer to anthropology than it was to either geography or psychology. He taught anthropology at the Escola Normal in Recife in the 1920s, and sometimes described himself as more of an anthropologist than a sociologist. He carried out something like fieldwork on Afro-Brazilians in his native Recife in the 1920s, paying particular attention to their material culture (the women's turbans, the men's pipes, the children's toys, and so on). He called Franz Boas his 'master' on the strength of having attended some of his lectures at Columbia, and he was familiar with the work of students of Boas, notably Ruth Benedict, Robert Lowie, and Melville Herskovits. In the case of Brazilian anthropology, he was an admirer of Edgar Roquette-Pinto and a friend of Heloísa Alberto Torres, René Ribeiro, Darcy Ribeiro, and the expatriate Frenchman Roger Bastide.

What difference did anthropology make to Freyre's achievement? He devoted only one of his many books to the subject, *Problemas brasileiros de antropologia* (1943), a collection of essays discussing both culture and race (while claiming that 'there are no races in an absolute sense'). However, the most important theory with which Freyre is associated, that of hybridity or mixture, discussed below, depends on his anthropological studies. Again, Freyre could hardly have written *The Masters and the Slaves*, especially the chapters on the culture of the Indians and the Africans, without a wide reading in anthropology, leading him to revalue Afro-Brazilian culture, generally dismissed by white intellectuals at that time, and to organize the Congress on Afro-Brazilian studies in Recife in 1934 (Levine, 1973). As for his shift from a concern with race to a concern with culture, discussed below, it owed much to Freyre's friend Rüdiger Bilden, who had studied with Boas; to Boas himself (especially to the famous article on 'Race' in the *Encyclopaedia of Social Sciences*), and to Roquette-Pinto. Again, the interest in material culture and everyday life that Freyre revealed in his historical studies, long before these topics became fashionable, like his concern for what he called 'microscopic history,' was inspired by both anthropology and by sociology.

Sociology

Of all the social sciences (unless they include history), the one to which Freyre was closest in practice was sociology. As a student, he took two sociology courses at Columbia with Franklin H. Giddings, and 6 years later he was teaching the same subject in Recife, and again, from 1935 to 1937, in Rio de Janeiro, in the new University of the Federal District, adopting a microsociological approach and recommending his students to begin by studying everyday life in the street in which they lived. He published a general textbook on sociology (1945b); an introduction to the sociology of medicine (1967); a *contribution to the sociology of biography* (1968); and, with characteristic ambivalence and egocentrism, an essay entitled *How and Why I am and am not a Sociologist* (also 1968).

In this domain as in others, Freyre's eclecticism is apparent, but his strong likes and dislikes are more visible than usual. He expressed enthusiasm for a wide range of sociologists. Among North Americans, he particularly appreciated the work of the Chicago school. Ernest Burgess was praised for his ecological approach to the city, along with the near-forgotten Indian sociologist Radhakamal Mukherjee, another 'master.' On the other hand, Freyre referred on occasion to the oversimplifications of American sociology, failed to mention Talcott Parsons (doubtless deliberately) and mocked Pitirim Sorokin for his ambition to be the 'Thomas Aquinas' of the discipline.

Among the Germans, one might speak of Freyre's elective affinity with Max Weber, going back to his brief flirtation with Protestantism. He was fascinated by what he sometimes called 'Protestant civilization,' as exemplified by punctuality, for instance, by diary keeping (a substitute for Catholic confession) and by an ascetic style of cooking. Freyre did not always agree with Weber (he criticized the Protestant ethic, for instance), but he treated Weber as an interlocutor even when

he did not mention his name, sharing his interest in patrimonialism, the relation between the public and the private, and the idea of time as money.

Freyre's extremely individual approach to the discipline is at its clearest in *Sociology* (1945b), which was planned as a textbook, thus giving the author both the need and the opportunity to comment on the development of the discipline. The text may even have been used as a textbook, since it went through four editions, but it is better described as a long essay (originally 772 pages, expanded later), criticizing claims made for sociology and discussing its relations with neighboring disciplines rather than introducing students to the subject. Freyre denounced what he called sociological 'purity,' declared that Proust had portrayed the French aristocracy more successfully than all the disciples of Durkheim, and called for a triple approach that would combine sociology with psychology and with history. His discussion was supported and enlivened by many Brazilian examples, for the most part repeating points that the author had already made in two historical studies, *The Masters and the Slaves* and *The Mansions and the Shanties* (1936) or would make in a third, *Order and Progress* (1959).

History

Freyre was and is most famous as a social historian (Hanke, 1939; Briggs, 1981). From the early 1920s, he declared his ambition to write the kind of history that would include the people as well as political or military leaders, women as well as men and children as well as adults (he planned to write a history of childhood in Brazil). It is no surprise that he was drawn to the American 'New History,' to the *Annales* School in France (describing Marc Bloch and Lucien Febvre as 'masters') and to the *English Social History* of George M. Trevelyan (he filled the margins of his copy with annotations).

Freyre practiced what he liked to call 'intimate history,' a study of ordinary people, humble objects, and apparently trivial but significant details. This project has of course enlisted many recruits since the 1980s under the banner of the 'new cultural history,' but in Freyre's time it was still relatively unusual, if not eccentric. He studied what he called 'culture,' or 'civilization' in a broad, anthropological sense of those terms, writing, for instance, about the 'sugar civilization' of Pernambuco or the 'carboniferous civilization' of Britain in the age of the Industrial Revolution.

Freyre's interest in cultural encounters and exchanges, including 'transculturation' (which he described as "the excellent neologism of Professor Fernando Ortiz"), can be found in many of his writings but it is particularly clear in *The English in Brazil* (1948), where, discussing the fashion for English goods among the upper classes of early nineteenth-century Brazil, he speaks not only of cultural imperialism but also of active reception, domestication, and on occasion of reverse acculturation, of Englishmen who adopted a Brazilian style of life.

However, Freyre is best known for the trilogy in which he traced the social and cultural history of Brazil from 1500 to 1900, with special emphasis on the patriarchal family (Tannenbaum, 1963). The first volume, covering the first three centuries, focused on the countryside, and in particular on the sugar plantations of the Northeast, arguing that the particular form of culture and society that developed in that region exercised a decisive influence on the rest of Brazil and also on later centuries. The second volume shifted to the cities of Brazil in the early nineteenth century, especially Rio, Salvador and Recife, while the third volume was concerned with the late nineteenth and early twentieth centuries.

The trilogy is remarkable in a number of respects. It was brilliantly written, especially the first two volumes. Indeed, Freyre was one of the most important Brazilian writers of the first half of the twentieth century, one of the first to abandon formality and to write in a manner that was close to colloquial Portuguese in both vocabulary and syntax. He was a master of vivid metaphors and he appealed to his readers with what he called a 'sensuous history' that reconstructed not only the look of the past but also its sounds (church bells, street cries), its smells (including urine and semen), its tastes (such as water drunk from an earthenware jar), and even touch (the practice of *cafuné*, stroking someone's hair and removing the lice). The three volumes were 'impressionist' in the sense of producing their effect by a succession of vivid vignettes and memorable anecdotes, describing rather than analyzing and using personal documents more often than statistics.

The first volume of the trilogy caused something of a sensation for its choice of topics; one of the chapters of *The Masters and the Slaves* bears the title, in its English translation, of 'The negro in the sexual and family life of the Brazilian' (it is difficult to know whether it was the reference to race or sex that readers found more shocking in 1933). To organize the material of the first volume, Freyre had the brilliant idea of focusing on the house as a microcosm of society; literally translated, the title of the book would be *The Big House and the Slave Quarters*. In similar fashion, the second volume focused on the *sobrado*, the town house of the rich, and the *mucambo* or shack, the ancestor of today's *favelas*, although it also had much to say about the street as an arena for work and play alike. It is in this second volume that Freyre made most use of what he had learned from Veblen, essentially the idea of conspicuous consumption as a way of affirming or maintaining social status. Like its predecessors, the third volume discusses types of building (notably the hotel) as symbols of society, but its most original feature is its method. The main source that Freyre used in *Order and Progress* was a questionnaire about social attitudes (including 'the heroes of your childhood') and memories of everyday life (toys, dances, fashions, and so on) addressed to survivors from the period, a 1000 people born between 1850 and 1900. About 20% of them responded (for a critique of this volume, see Skidmore, 1963–1964). This original and ambitious work of history was also the principal vehicle for Freyre's social and cultural theory.

Theory

Freyre's late essays on futurology, urbanization, and other topics of contemporary interest lack the richness and density of his earlier work. It was in his studies of social history that he made his most important contribution to theory. A central concept in Freyre's work was that of equilibrium. However, he imagined this equilibrium not as static but dynamic, an

'equilibrium of antagonisms' (an idea that he learned from Herbert Spencer and Franklin Giddings). He has often been accused of having too rosy a picture of Brazil, of over-emphasizing harmony. This may well be the case, but it should be added that he did not deny the existence of conflict. What he suggested was that conflicts could be and sometimes were harmonized or to use one of his favorite words, 'softened.'

Related to this process of softening is the best known theory associated with Freyre: the theory of hybridity, in the sense of miscegenation and also of what he often called the 'interpenetration' of cultures (Needell, 1995; Cleary, no date). In *The Masters and the Slaves*, he argued that this interpenetration was beneficent and that it had made Brazil what it was, writing that "Every Brazilian, even the light-skinned fair one, carries about with him on his soul, when not on soul and body alike … the shadow, or at least the birthmark of the Indian or the African" (1933: p. 278). It was this argument in particular that made him famous. In the 1930s, the people, or at least the intellectuals, were anxious about the Brazilian's apparent lack of identity. Since they were descended from a mixture of the indigenous inhabitants of the country, European immigrants (mainly Portuguese), and African slaves, what could they be?

Educated Brazilians tended to believe that interbreeding led to degeneration and that the only hope for the country lay in its 'whitening,' thanks to a new wave of immigration from Europe. Freyre himself believed this for a time, but he reversed his position to claim that mixture was not an obstacle to Brazilian identity, it was the identity. He owed this conversion to his friend Rüdiger Bilden, who had studied with Franz Boas; to Boas himself, whom he adopted as a 'master' (displaying a photograph of Boas in his study); and to the Brazilian anthropologist Roquette-Pinto (Pallares-Burke, 2005, 2012).

Freyre saw miscegenation as a key to Brazilian history, associating it with a regime of slavery that was, in its patriarchal, paternalist way, more gentle than slave regimes elsewhere. This point was a relative one; he was well aware of the violence built into the system. On the other hand, he paid little attention to the fact that miscegenation was generally asymmetrical, the result of relations between upper class white males and black female slaves. He has also been criticized for his rosy picture of what is often called 'racial democracy,' in other words the lack of racism in Brazil (Guimarães, 2005). Once again, his point was a relative one; Brazilians, he argued, came nearer to this ideal than most if not all other cultures. Later studies (including one by Fernando Henrique Cardoso in his days as a sociologist) focused on discrimination in cities in the south of Brazil and undermined Freyre's argument, although it appeared and indeed still appears more plausible in the region he knew best, the Northeast, than it does elsewhere.

The Masters and the Slaves also discussed cultural mixing or hybridization (in Portuguese, *mestiçagem*) in domains that range from vernacular architecture to language, clothes, and cuisine. At a time when sociolinguistics was in its infancy, and had not yet been christened, Freyre wrote brilliant pages describing the influence of Africa, and especially of African slave nannies, on the vocabulary and even the accent of upper class Brazilians (another striking example of reverse acculturation).

In later works, Freyre returned to the concept of *mestiçagem*, to play with it or as he might say to dance with it, pushing it to see how far it would go. Thus, he extended the concept to sport, writing about 'mulatto football,' for instance, not only as a game played by mulattos but also as a sport that mixed English football with the Brazilian tradition of *capoeira*. He extended it to literature, describing the novels of his friends João Lins do Rego and Jorge Amado as hybrids of fiction, social history, and folklore. He extended the idea to the practice of sociology, which he viewed as a mixture of art and science, a discipline that was 'amphibious' or 'polytheist.' He extended the concept of *mestiçagem* to politics, especially to Brazilian politics during the transition from empire to republic, or more exactly, so he suggested, from a republican form of empire to an imperial form of republic. He even extended the concept to sex, allowing a character in his 'seminovel' *Mother and Son* to suggest that people of hybrid gender sometimes realize themselves more fully than the pure do.

The idea of *mestiçagem* underlies what may well be Freyre's most important contribution to the social sciences, his frequently reiterated suggestion that their evidential base, essentially Western Europe and the United States was too narrow, and that generalizations about human culture and society needed to draw on the experiences of other parts of the world. In a word, a word that Freyre liked to use, the social sciences needed to be 'tropicalized.' For example, Freyre wanted to see a tropical sociology on the analogy of tropical medicine, or even an interdisciplinary 'tropicology.' In short, he advocated what is now called 'Southern Theory' long before it was christened.

Linked to the idea of tropicalization is Freyre's least successful contribution to social theory, the concept of 'Luso-Tropicalism,' according to which the Portuguese were better adapted than other Europeans to building an empire in the tropics because they were accustomed to miscegenation and cultural mixture at home, thanks to the Arab invasion of Iberia in the Middle Ages. This argument appealed to the Salazar regime in Portugal at a time when their colonies were becoming restive, and Freyre was invited to pay an official visit to 'Ultramar,' in other words the Portuguese possessions in Africa and Asia, in 1951.

One criticism of the idea of tropicalization is that it offers a binary opposition, between tropical and temperate, at the expense of many possible nuances. As the titles of some of his books suggest – *The Masters and the Slaves*, *The Mansions and the Shanties*, *Adventure and Routine*, and so on – Freyre was fascinated by polarities, although he also liked to undermine them by examining connections, exchanges, and hybridization. Two of these polarities reverberate throughout much of his work: Jekyll and Hyde, and Apollo and Dionysus. He used Robert Louis Stevenson's 'Dr Jekyll' and 'Mr Hyde' as a means to move away from a vision of history as a conflict between heroes and villains, arguing that in the case of Getúlio Vargas, for instance, this conflict took place within the President's personality and policies. As for the opposition between the Apollonian and the Dionysian, Freyre borrowed it, with acknowledgments, from Ruth Benedict (although he had read Nietzsche for himself) and made use of it again and again. English football, for instance, was described as Apollonian, while Brazilian football was Dionysian. Intellectuals, and more specifically social scientists, were divided in similar fashion, sociological purists being labeled as Apollonian while Freyre himself was Dionysian.

Reception

The reception of Freyre's books in Brazil has passed through three main phases. Despite criticisms of *Casa Grande* in some quarters, it quickly became both a sociological and a literary classic, while its author was treated as a 'master' (among others, by the novelist Jorge Amado). However, the criticisms made of Freyre's work by the sociologists of São Paulo diminished Freyre's reputation, which was also undermined, at least among intellectuals of the left, by his accepting the official invitation to Ultramar in 1951, thus implicitly accepting the use of his work to legitimate a decaying colonial regime. His support for the military coup in Brazil in 1964 blackened his image still more, and in some universities, students were advised by their teachers not to read his work. Following the reestablishment of democracy in Brazil in the 1980s, Freyre's star began to rise again. Some leading figures in the social sciences in Brazil, notably the anthropologist Roberto DaMatta, work in what might be called a Freyrean tradition. Freyre also enjoyed success at a popular level of a kind rarely attained by a social scientist, since some of his books were translated into plays, films for the cinema and television, and even a comic strip. More recently, when Rio was selected to host the 2016 Olympics, President Lula's speech, describing Brazilians as "a mixed people and a people that like to be mixed," sounded like a summary of the central thesis of *Casa Grande*.

Abroad, Freyre was launched by the North Americans and the French. His lectures at Bloomington in 1944 helped make him known in academic circles in the United States, and his trilogy on Brazilian history was published in translation between 1946 and 1970. In the case of France, the historian Fernand Braudel discovered Freyre's work while he was teaching in Brazil in the late 1930s, and later published an enthusiastic review of it (Braudel, 1943). Braudel's colleague Lucien Febvre wrote a preface to the French translation of *Casa Grande*, and Braudel did the same for the Italian translation. Praise from Roger Bastide, Roland Barthes, Jean Duvignaud, and Georges Balandier also helped Freyre's work to circulate in France. His reputation in Europe was probably at its height in the 1950s, but translations of his books still appear from time to time, into Hungarian, for instance, and Japanese.

Conclusion

Among the social scientists of his day, Freyre stands out as a southern theorist, as a man of letters turned anthropologist and sociologist, and as a scholar who insisted on the need for students of culture and society to learn from both history and literature. He lived long enough to see the humanistic approach to the social sciences that he always advocated come back into fashion. Preferring ethnography or 'microsociology' to quantification, and insisting that social scientists avoid specialized jargon, it might be said that Freyre was closer to the generations before and after him than he was to his own and even that his attachment to tradition helped him to become an innovator.

Bibliography

Braudel, F., 1943. A travers un continent d'histoire: le brésil et l'oeuvre de Gilberto Freyre. Mélanges D'histoire Sociale 4, 3–20.

Briggs, A., 1981. Gilberto Freyre and the Study of Social History. In: Rpr Collected Essays, vols. 2. Harvester Press, Brighton, 1985, vol. 2, pp. 272–287.

Burke, P., Pallares-Burke, M.L., 2008. Gilberto Freyre: Social Theory in the Tropics. Peter Lang, Oxford.

Cleary, D., no date. Race, Nationalism and Social Theory in Brazil: Rethinking Gilberto Freyre. www.transcomm.ox.ac.uk/working%20papers/cleary.pdf.

Freyre, G., 1933. Casa Grande e Senzala. English translation, The Masters and the Slaves. Knopf, New York, 1946.

Freyre, G., 1936. Sobrados e Mucambos. English translation, The Mansions and the Shanties. Knopf, New York, 1963.

Freyre, G., 1937. Nordeste, seventh ed. Global, São Paulo, 2004.

Freyre, G., 1945a. Brazil: An Interpretation. Knopf, New York.

Freyre, G., 1945b. Sociología. Olympio, Rio de Janeiro.

Freyre, G., 1948. Ingleses No Brasil. English translation, The English in Brazil. Boulevard, Oxford, 2011.

Freyre, G., 1959. Ordem e Progresso. Abridged English translation, Order and Progress. Knopf, New York, 1970.

Guimarães, A.S., 2005. Racial democracy. In: Souza, J., Sinder, V. (Eds.), Imagining Brazil. Lexington Books, Lanham, MD, pp. 119–140.

Hanke, L., 1939. Gilberto Freyre: Brazilian social historian. Quarterly Journal of Inter-American Relations 1, 24–44.

Levine, R.M., 1973. The first Afro-Brazilian Congress. Race 15, 185–193.

Needell, J., 1995. Identity, race, gender and modernity in the origins of Gilberto Freyre's oeuvre. American Historical Review 100, 51–77.

Pallares-Burke, M.L., 2005. Gilberto Freyre: Um Vitoriano dos Trópicos. UNESP, São Paulo.

Pallares-Burke, M.L., 2012. O triunfo do Fracasso: Rüdiger Bilden, o amigo esquecido de Gilberto Freyre. UNESP, São Paulo.

Skidmore, T.E., 1963–1964. Gilberto Freyre and the early Brazilian republic: some notes on methodology. Comparative Studies in Society and History 6, 490–505.

Tannenbaum, F., 1963. Introduction. Freyre, Mansions, pp. vii–xii.

Friendship, Anthropology of

Bettina Beer, Universität Luzern, Switzerland
Don Gardner, Universität Luzern, Switzerland; and CASS, Australian National University, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Friendship has frequently been juxtaposed to the relationships that constitute the 'structuring structures' of social life, be they the kin and descent relations of small-scale societies or the institutionalized forms of modernity. But interest in friendship and related concepts (mutuality, trust, intimacy, love) has increased dramatically in recent decades. The processes of globalization and postclassical trends in social theory are equally implicated in the increasing relevance of friendship to both ethnographic inquiry and theoretical concerns. Nevertheless, for anthropologists, friendship and kinship remain connected in ethnographically, theoretically, and conceptually unsettling ways.

Introduction

It seems obvious that the significance of friendship to general social life varies across time and space. However, the classification of the empirical characteristics of friendship has not been simple, even for scholars of Western society (Pahl, 2000); anthropologists, who consider friendship in a comparative perspective in which empirical and conceptual issues intersect, also note the refractory nature of the topic (Paine, 1974; Firth, 1999; Coleman, 2010).

The anthropology of any topic must avoid the twin dangers of a specificity that inhibits generalization or comparison and an analytically useless generality, but there are additional factors that make the anthropology of friendship an exacting area of inquiry. First, the philosophical roots of the social sciences include influential treatises on friendship by (among others) Aristotle, whose writings on the topic continue to inform many discussions, notwithstanding the conflict between his orientation and the dominant traditions in social theory, which have found the constitutive role of biographical particularity in friendship hard to integrate. More conspicuously, the very concept of friendship implicates one or more ethically rich, affect-laden notions (such as love, intimacy, trust, loyalty, sincerity, and equality) as do certain more experience-distant counterparts of those notions ('social capital,' 'ontological security,' for example). Moreover, anthropologists interested in friendship have worried that the very identification of a friendship involves the observer in the inner motivational states of social agents. So, while friends may be crucial to the agent's lived experience, scholars of social structure appear to have reason to ignore friendship, despite its empirical entanglement with significant politico-economic phenomena (exchange and alliance). Accordingly, complex issues can arise in ethnographic investigations of friendship no less than in cross-cultural comparisons or consideration of its relations to cognate phenomena (kinship, bond friendship, blood brotherhood, *compadrazgo*, and other patron–client relations).

Additionally, friendship has remained a shadowy region of the theoretical landscape because of the light shed by anthropologists on kinship and familial relations – the classical 'structuring structures' of so many ways of life. More recently, however, the conceptualization of friendship has been complicated by three interconnected factors: first, the globalization of cultural elements of Western modernity, in association with the consolidation of capitalism's global economic power, has produced complex hybrid localities whose ethnographic descriptions are difficult and theoretically contested; second, these same currents have increased the tendency of Western subjects to understand themselves as post-traditional, self-fashioning beings, sensitized to the work of gender, class, and racial stereotypes; third, the last 40 years have seen the confrontation/complication of structural approaches through an emphasis on agency and its specificities with respect to power, gender, and historical process (including factors that are contingent with respect to all preexisting local characteristics).

These trends do not, however, imply that there are great discontinuities in scholarly approaches to friendship. It must still be defined by contrast with other relations in a multidimensional space that sets up the possibilities for the various dyadic and polyadic links between ego and others, including those that form the developmental context of persons. The axes that best define the personal and institutional possibilities of that space are, in truth, essentially contested aspects of social theory; but key features of friendship (its 'private nature,' 'voluntarism' and 'individualism') mean that it has in large part been constituted by its juxtaposition with, on the one hand, the affective but 'obligatory' relations of the kin domain, and, on the other, those impersonal relations that constitute the broader institutional order of modernity.

A remark concerning the scope of this contribution: recent work in the broad area of the anthropology of friendship has come from two quite different directions: evolutionary theory, most conspicuously in Hruschka's (2010) large-scale comparative investigation of friendship (augmented by test data and a broad sampling of ethnographic literature); and from a socio–cultural anthropology enlivened, to various degrees, by the advent of 'new kinship studies' and the anthropology of personhood, as well as by the challenge of dealing analytically with the entanglement of local socio–cultural settings and global flows. While these two strands of scholarship are inspired by very different commitments, there is much about friendship upon which they agree: the ubiquity and variability of friendship; its significance in organizing important socio–economic relationships within and between culturally

defined networks; the role of positive affect in paradigm cases of friendship, even those that are organized with respect to explicit norms or are referable to instrumental values; the centrality to paradigm friendships of generalized reciprocity (sharing or mutuality) rather than balanced reciprocity. These two streams also converge because conceptual issues concerning altruism, egoism, and what is involved – in several different senses – in a person's doing something for the sake of another, arise in evolutionary and social theory alike. This entry must nevertheless concern itself with the reconfigured interest in friendship in socio–cultural anthropology.

Finally, we can only mention in passing the extensive literature on friendship in sociology, psychology, and gender studies. It must suffice for us to stress that the configuration of empirical, theoretical, and conceptual issues that characterizes anthropological discussions of friendship are quite general (Adams and Allan, 1998; Pahl, 2002).

Terms

The word 'friendship' needs some preliminary attention: it can refer either to the relationship between persons ("Their friendship is of long standing"), or to the inclinations, dispositions, and actions of a single person ("She assured him of her friendship"), in which case it is roughly synonymous with 'friendliness'; of course, these are intimately connected, but we focus on the relationship between friends, for friendly sentiments can be unreciprocated, while a friendship necessarily involves mutuality.

Although anthropologists can rarely ignore issues of translation, their usual concerns about lexical and conceptual equivalents across languages are, in the case of 'friendship,' compounded by the continuing influence of classical sources – most notably, Aristotle, in whose treatises on ethics much of the academic discussion on the subject still finds its source. The Greek term (*philia*) usually translated as 'friendship' approximates the English notion of (nonerotic) love and encompasses family members. Indeed, the term 'friend' itself derives from an old Germanic verb meaning 'to love,' and the noun's cognates in certain northern European languages continue to designate kin rather than friends. Yet the semantics of contemporary English, at least, offer no impediment to a description of, say, a sibling as a mortal enemy; similarly, one's jailer or bank manager may be counted a friend. In many small-scale settings, by contrast, social life is such that ego's total social field and ego's kindred are isomorphic, so that those an English speaker would count as ego's 'friends' necessarily represent a subset of his or her relatives.

The general point here is not just that attention to linguistic or conceptual detail is advisable, but also that difference and change with respect to configurations of friendship involve ethical, linguistic, and social norms and the relationships between them. For example, a principal issue in the analysis of friendship (and its cognates) has always been the extent to which relations that involve very different configurations of affect, degrees of 'utilitarian' motivation or 'choice' and power differentials, or which subsist in highly contrasting social formations, can be brought under a single, analytically significant category. The conceptual issues here seem tailor-made for a scholar like Derrida (1997), with his nose for 'undecidability' and 'aporetic structures'; as he evidently appreciates, these issues are connected to taxing issues in other areas of anthropology (like the gift).

These issues notwithstanding, we follow Aristotle's line in taking paradigmatic friends to be those who enjoy one another's goodwill, trust, and affection; they share these things and – crucially – each not only knows this of the relationship, but also knows that the other knows it; this suite of interconnected characteristics means that friends can count on or trust one another, which thereby becomes relevant to their actions and strategies. Friendship, then, is sensitive to time and the interpersonal experiences of those involved, which helps to explain the organic images and metaphors often used to characterize it.

Friendship and Anthropology

Many discussions of this topic begin by pointing out that friendship has not loomed large within anthropology. In a discipline that grew to maturity through its engagement with the different ways in which kin relations permeated, organized, or regulated social structures, friendship has been easy to overlook, even if its universal presence and general significance has been acknowledged (Leyton, 1974; Firth, 1999). More recently, anthropologists (and sociologists) have grown more interested in the specificities of friendship; but by the time its claims to attention became less residual, the emphasis on personhood and its cultural construction that had come to dominate post-Schneiderian kinship theory ensured that friendship once again had to accommodate itself to an environment already occupied. Accordingly, Coleman's view that classical anthropological studies of friendship are "few and far between" (Coleman, 2010: p. 197) is apt.

Yet, a picture of general neglect is a distortion: certain topics and the organization of certain classical small-scale societies necessitated a consideration of interpersonal relations forged, and often formally recognized, between friendly individuals unrelated by primary identities (for example, Evans-Pritchard, 1933; Firth, 1936). Moreover, students of peasant communities, proletarian urban networks, and other politically encompassed social settings made sure that friendship and other informal, private social relationships were given analytic space (for example, Whyte, 1943; Mintz and Wolf, 1950; Hart, 1973). In addition, for much of their history, anthropology and sociology shared sources of inspiration, research agendas, and range of theoretical orientations. Accordingly, a background awareness that the significance of friendship in social networks was an empirical cultural variable was common among anthropologists too; indeed, it was sometimes stressed by those interested in developing transactionalist alternatives to structuralism, or in refining functionalism (Eisenstadt, 1956; Paine, 1969; Boissevain, 1974).

When friendship was considered in a comparative perspective, it was usually in an institutionalized, often ritualized, form, or as it shaded into patron–client and comparable inegalitarian relations. This is not very surprising, given the predominant twentieth century views on interpersonal

relations and social order. General theoretical commitments, which stressed the externality and coercive power of social facts, could hardly avoid marginalizing informal, personal relationships. So, while friendship was never ignored, it was often considered to constitute a residual category, comprising phenomena seen as peripheral or interstitial, especially with respect to kinship and descent and other institutional orders. Friendship, it seemed, challenged even the most muscular of social theories: neither the invocation of rules or custom nor the elucidation of useful effects sufficed to account for actual behavior between friends (Paine, 1969, 1974; Pitt-Rivers, 1968). On the other hand, many scholars underlined the 'always already' cultural nature of human agents, and stressed the social homophily of friendship (with respect to class, gender, age, and other aspects of identity).

Studies of friendship were no less affected by interests in agency and by feminist theory and critique than other aspects of the social sciences. Students of friendship, in fact, may have been especially susceptible to feminist critique, for gendered stereotypes about the capacity for making friends, commitment to domestic spaces and to familial values were an important part of many cultures, including that of most social scientists. In anthropology, this predicament was compounded by the prevalence of the Mediterranean and Latin America in the ethnography of friendship, regions where gender stereotypes are conspicuously relevant to relationship patterns (Uhl, 1991).

Volumes dedicated to friendship that have been published recently (Bell and Coleman, 1999; Schmitt et al., 2007; Desai and Killick, 2010), and several theoretically motivated articles (e.g., Killick, 2009; Fausto, 2012) seek to bring friendship closer to the center stage of anthropology. They take account of the scope of global cultural flows as well as the transformation of social theory, and show how, once questions about the effects of globalization on local patterns of sociality became an explicit research topic, the question of friendship became salient. Interestingly, in such settings, both kinship and friendship contrast with the impersonal relations that constitute state and international institutions: although they are personal, relatively private aspects of social life, they are nevertheless crucial to the networks that define local political economies (including those that are part of diasporic networks). If kinship and friendship are similar in these respects, the constructionist and 'performative' orientations inspiring the 'new kinship studies,' also weakens the contrast between them. However, this has been subject to considerable discussion and controversy.

Friendship also evokes those Maussian lessons on the role of exchange in forging links between individuals and groups that were central to twentieth-century anthropology's understanding of social relations. Sahlins' oft-repeated chiasmus generalizes the point: "If friends make gifts, gifts make friends." These Maussian insights have been challenged by more recent anthropological models of relatedness as radically generalized: if relationships are performed in accord with local constructions of personhood, practices predicated of interactions between folks related as kin and those related as friends become less important. This trend finds clear expression in a recent exposition from Sahlins, which endorses Aristotle on the *friend* as 'another self,' but as part of a wide-ranging case for seeing *kin* as "persons who are members of one another, who participate intrinsically in each other's existence" (Sahlins, 2011a: p. 2). This is not so remarkable, since Aristotle saw relations between close kin as paradigms of the love (*philia*) that characterized true friendship. Yet Sahlins' conflation, while consistent with currently dominant views on relatedness in the anthropology of kinship, does raise concerns about the advisability of eliding distinctions between kinds of relatedness, and not only because such distinction are often crucial to those we study; after all, these differences are, *prima facie*, of existential no less than sociological importance. Recent objections by specialists to the annexation of friendship to an expanded notion of 'relatedness' (discussed below) raise the problem explicitly, but these relate to matters that have been crucial to the theoretical characterization of relationships since the advent of the social sciences.

Friendship, Kinship, and Difference in Cultures of Relatedness

At least since "the first book-length study of friendship in modern anthropology" (Leyton, 1974), scholars have spoken of the need to maintain an open, 'exploratory' orientation to the topic. This need arose because of the crosscutting articulation of 'psychological' and 'social situational' levels on which friendship operated. In Paine's terms ([1970]1974: p. 1), the stark juxtaposition friendship made apparent between the 'moral utilitarian and the social functional' approaches to social relations tended to raise doubts about how comprehensive the latter could be as social theory. Paine's use of categories that evoke Talcott Parsons' theoretical edifice (in particular, its 'pattern variables') is typical of the theoretical currents of the times; Paine (1974: p. 12) deemed functionalist theory appropriate to 'particularistic, solidary' groups classically studied by anthropologists, but wondered whether, in contemporary society, the prevalence of 'universalistic' relations and values meant that Simmel and Goffman would be better sources of understanding. The basic issue here is the difference between friendship and kinship as the latter was depicted in the structural-functionalist readings of 'traditional societies,' especially those to which descent theory had been applied so extensively and apparently fruitfully.

Space precludes a comprehensive genealogy of anthropological conceptions of kinship as these are relevant to friendship, but it would entail a consideration of, on the one hand, the characterization of those socio-political structures whose reproduction underpins the transgenerational identity of culturally defined populations (the "political-jural" realm, as Fortes (1969) famously designated it) and, on the other, the affective, morally charged arena of primary interpersonal relations within which ego, as a cultural being, acquires specificity, which – again, according to Fortes – is governed by 'the axiom of prescriptive altruism.' We can, though, sketch the interconnections here with a brief look at a text that has been much referred to (not always fairly), even in the most recent anthropological discussions of friendship: Pitt-Rivers' (1973) 'The Kith and the Kin,' written for 'The Character of Kinship,' a *festschrift* for Meyer Fortes. This and other works by Pitt-Rivers engage Fortes on kinship and its

fundamental character, but from the perspective of an Andalusian community, in which friendship is both socially fundamental and existentially problematic.

Pitt-Rivers paper begins by remarking the oddness of Fortes having chosen 'amity,' derived from the French term for 'friendship,' to designate 'the axiom of prescriptive altruism,' which Fortes claimed as central to kinship. It seemed that Fortes had "chosen to define the essence of kinship by appealing to the very concept of what it is not" (1973: p. 90). For, Pitt-Rivers continues, most scholars oppose kinship, as "inflexible, involuntary, immutable, established by birth and subject to the pressures of 'the political-jural domain' … [to] friendship, … which is its contrary in each of these respects" (p. 90). He delicately analyses his teacher's Aristotelian inclination to see amity as intrinsic to the child's experience of familial relations, as having their own moral values (beyond self-interest and egoistic calculation), while retaining the structural-functionalist's commitment to the prescriptions of the social order that bind behavior and sentiment. Pitt-Rivers seems to suggest (but not state) that Fortes' refusal to deal with the tension that his bifurcation of kinship (into a realm governed by primordial altruistic sentiments and a broader political–jural order grounded in rules) is unsustainable. For his own purposes, though, Pitt-Rivers is clear that friendship should be assimilated to the primordial/familial order of kin relations: as both involve the submersion of self-interest "*for the sake of* someone else" (p. 90; – his italics), and both are dimensions of the realm of sentiment (he says several times that 'the heart' must rule in these relationships), truly moral values, and of "the extension of the self" (p. 91).

Pitt-Rivers remains equivocal (or excessively delicate) about revising his teacher's perspective. He sometimes seems to endorse Fortes' suggestion that it is the behavior that matters, not whether it is motivated by nonjural sentiments rather than jural norms; for example, he suggests – by contrast to his stress elsewhere on the heart and its uncoerced inclinations – that amity can result from moral *obligation*. He underlines the difficulties caused by the fact that "the state of the heart … cannot be known for sure" (p. 98). At other times, by contrast, he stresses the "antithesis between the notions of jural and moral relations" (p. 96), and the requirement that amity remain "*purely* moral … untrammelled by the jural domain …" (p. 102; – his emphasis).

Yet the careful analysis that Pitt-Rivers offers in this rumination on Fortes' work gently but definitely underlines the tensions in Fortes' conception of kin relations and their relation to altruism. Indeed, the reader is sometimes struck by the thought that Pitt-Rivers aims to suggest that his esteemed teacher's key concept of 'prescriptive altruism' embodies a contradiction: an act that is performed because it is required by a social rule is not altruistic, since the heart's inclinations are trumped by jural requirements no matter what.

Pitt-Rivers (1954), in his earlier ethnographic analysis of an Andalusian village (for which Evans-Pritchard wrote the foreword), had already had occasion to ponder the epistemic difficulties of friendship – of telling when it truly sprang from the heart. He worried that the criterion that "distinguished true from false friendship flees from the anthropologist into the realms of motive" (p. 139). Interestingly, Paine cites this passage in each of his three pieces on friendship.

The stresses generated by the question of the role of motivations in intimate social relations have appeared elsewhere: Marshall (1977), worrying along similar lines to Pitt-Rivers, but under provocation from David Schneider's writings on kinship, also concluded that motivations, in the form of commitments (to kin *and* friends) to share and thereby maintain what he calls "intensive interpersonal relationships," had to be counted as central to the relations sustaining Trukese society.

Paine, like Pitt-Rivers, underlines the trust and affectivity that friendship involves, and the dangers of miscalculation or betrayal that these entail in many social settings. But, for the Western middle class, friendship choices are rendered less hazardous – on the whole – by their interstitial nature relative to market and bureaucratic relationships; friendships carry less political and economic weight than other relationships, so that friends are free to explore the possibility of attaining 'ideal' relations (1969). His later essay (1999) moderates this picture somewhat: the ideal values associated with modern friendship presuppose relations between persons, each of whom has 'verifiable self'; but this makes for vulnerability too, either through the friend's intent to deceive or because of the friend's self-deception. Here, Paine seems to return to Aristotle's view about that personal virtues of friends determines the degree of perfection of the relationship between them (Pakaluk, 1998); but it is evident that he still sees these personal virtues as somehow tangential to the "rights and obligations that are developmentally prior to … friendship and are conceived as indispensable to the proper functioning of the society" (1969: p. 508). Paine's later reflections offer a picture of Western friendship similar to that which he gave 30 years earlier: "a kind of institutionalized non-institution" (1969: p. 514) the understanding of which depended upon blending the structuralist's concern with rules with the Goffmanesque transactionalist's "discernment of ploys or strategies" (1974: p. 13).

Reference to Paine and Pitt-Rivers's work readily evokes the influential Allan Silver (1989, 1990, 1997, 2003) and his careful, historically informed work on models of friendship in various strands of social science theory. Silver endorses Paine's (1969) observations about the specificity of contemporary friendship and draws many threads together to demonstrate how historically specific were the nineteenth-century oppositions to which most later social theory was heir. Although, he argues, the juxtaposition of personal, private, moral, and intimate modes of sociality to impersonal, public, commercial, and formal modes runs together different and independent sociological parameters, it fits with certain orientations to nineteenth-century Europe's past, common to radicals no less than conservatives. This dichotomous picture came to represent the historical (and teleological) transition to modernity, thereby (and in addition) serving as the basis for societal typologies.

On this view, friendships under modernity are stereotypically "voluntary, unspecialized, informal and private … grounded in open-ended commitments without explicit provision for their termination" (1989: p. 274). As such, friendship is exemplary of "the private domain of personal life" (1997: p. 46) and diametrically opposed to the contractual, public, and impersonal foundations of modernity's diacritical institutions. But, Silver argues, with a wealth of evidence from literature, "that this domain of the private, however suffused with

historical imagery, is less a historical survival than a distinctive creation of the impersonal order central to modern economies and polities" (1997: p. 44). Indeed, the leading figures of the eighteenth-century Scottish Enlightenment (Smith, Hume, Ferguson) confound the presentism of later social theorists in espousing a liberalism that celebrates commercial society's ability to partition self-interested relations from those grounded in moral sentiments (1990); commerce actually made room for purer, more sincere, friendships.

In turn, James Carrier's (1999) provocative essay on friendship and personhood, in Bell & Coleman's important volume, extends Silver's line of argument: he begins by contrasting contemporary Western views about friendship, as "based on spontaneous and unconstrained sentiment" (1999: p. 21), with relations grounded in "the demands or expectations placed upon them by the ties of kinship, trade, propinquity, interest or the like" (pp. 22, 23). This notion of friendship "entails a distinct conception of what people are like, of the self" (p. 23). Then, through an account of the history of that conception of self, and a contrast between it and the Melanesian self as discussed in seminal anthropological accounts, Carrier suggests that "the terms by which Westerners distinguish friendship from other sorts of relationships" are not available to Melanesians who sort relations by reference to kinship and to social situatedness (p. 31). Finally, Carrier draws attention to a point repeatedly underlined by sociologists of friendship (Adams and Allan, 1998; Pahl, 2000), that Westerners' pretheoretical intuitions about friendship, its moral qualities, and what it entails or permits in the way of behavior, is very much a function of time and place (p. 36), even controlling for such factors as class or age. In the context of the debates occasioned by the juxtaposition of Melanesian dividuality and Western individuality, which has been prominent in discussions of the relationship between kinship and personhood for two decades, Carrier's analysis seems to have implications that he forbears to state: that Melanesian views of the person contrast with those of some Westerners much more than with others. He discusses sociological research showing that commitment to individualism in the West varies in important ways with gender (pp. 31, 32), class (pp. 32–34), and other aspects of "the broader social, political and economic context" (p. 36).

Carrier's article poses clearly the sort of questions anthropologists have felt constrained to ask about using a term like 'friendship' in settings that contrast fundamentally with the contemporary West. His piece is challenging because it pushes hard at a certain take on Silver's line, not only to suggest that the anthropologist is always involved in a double hermeneutic, so that interpretation of the practices and concepts of ethnographic subjects always presuppose an analysis of the practices and concepts of the ethnographic observer's culture, but also (polemically) that this sets up a boundary that separates 'us and them' with respect to friendship, but which cuts across the dividual/individual divide many have accepted as distinguishing Melanesian and Western cultures.

Desai and Killick (2010) find they have reason to critique Carrier's position. Indeed, they appeal to the 'pre-industrial' Aristotle in urging that 'the autonomous self' is not exclusively a product of the capitalist West, but can be produced by other histories (2010: p. 10). For example, the indigenous Mapuche category of the person, as described in the volume by Course, involves an "autonomous individual self, together with the capacity to make friends"; and while sentiment and its revelation is not as central to Mapuche friendship as it is to contemporary Western conceptions of paradigmatic friendship, it does entail the 'ontological' individual (2010: pp. 9, 10). Killick's research with the Asheninka of the Peruvian Amazon likewise underlines "the importance of self-sufficiency and of personal independence and autonomy" in the formation of crucial social relationships (2009: p. 705).

Although key terms in these characterizations evoke ancient controversies in Western philosophy and the foundations of the social sciences, Desai and Killick's aim is to moderate the sort of point made by Paine, Silver, and Carrier, rather than to oppose it entirely. They, and others in their volume, argue that it is important to oppose the tendency of 'new kinship studies' to annex friendship, yet to conceptualize it flexibly enough to explore contemporary transformations of local fields of sociality (Desai and Killick, 2010). For them, patron–client relations no less than those between childhood playmates can count as forms of friendship; they cite and endorse Firth (1999: xiv) on the significant cross-cultural variability in the 'intensity' of friendship. They do not repudiate constructionist notions of personhood but worry that without attention to the specificities of friendship's forms we will miss how these articulate with ideas about kinship and other key relationships, as well as with ideas about 'being a person' (2010: p. 15). They are concerned that if the category of kinship is to engulf that of friendship, we will lose the capacity to track and explain empirical variation in the forms of each and in the relations between them. So while they seek to make connections between different kinds of social relations, they see it as important to retain the distinctions that are present in the empirical material.

Coleman's appreciative afterword to Desai and Killick's volume endorses this line, but adds that being 'residual' can give friendship 'its own power' and (in an allusion to Giddens on modernity's concern with 'pure relationships') render it "deeply and satisfyingly impure" (2010: p. 205). We need, he goes on, to understand that there is "voluntarism as well as constraint in the formation of relationships" (2010: p. 205) a point made by many contributors to the earlier volume (Bell and Coleman, 1999). Much will depend upon the extent to which we can avoid approaching 'voluntarism' and 'constraint' as though they were independent variables in the generation of social outcomes, which would precisely return us to mid-twentieth-century social theory. The continued appeal to the distinction between the ascribed (kin relations) and the achieved (relations between friends) also suggests that such theory remains influential.

Killick observes that "while various attempts have been made to stabilize the category of friendship using ideas of autonomy, sentiment, lack of ritual, and lack of instrumentality, … a general consensus is still lacking." Whether we should therefore "side-step the search for a final definition and instead … base analysis on the indigenous forms encountered" (2009: p. 702) is not entirely clear, for without a specification of the open category 'indigenous forms' Killick's plea might be thought to beg the question. Perhaps, as Aguilar (1999) seems to claim, a commitment to the view that "friendship as any other social phenomenon is culturally constructed … [and] its

manifestations are influenced by localized ways of being human and of being social" (pp. 170–171) will give as much basis for an answer as we need, for it directs our attention to the local context and the concepts and relations that constitute it. In this way, studies of friendship will remain part of the broader dialogues on personhood and ontology that have become topical (especially in the South American and Melanesian literature). However, both Killick (in the works discussed) and Coleman (2010) seem to remain equivocal about constructionist accounts of friendship (and kinship). This is not surprising: what makes a vernacular construction of a relationship one that we are justified in saying is a construction of *friendship*, as opposed to some contrasting species of relationship, does not seem to be a straightforward empirical issue.

As noted earlier, anthropology always faces questions of translation and commensurability. Certainly, what Aristotle wrote about *philia*, which most of his translators render as 'friendship' (its inclusion of primary kin relations notwithstanding), does not make Aristotle party to our interest in whether "different configurations of the relationship between history, personhood, sentiment, and friendship [show] that comparable, if not identical, ideas of friendship may occur in nonwestern cultures" (Killick, 2009: p. 702). And for those interested in friendship in cross-cultural comparison, the question whether the self or 'the autonomous individual,' is constructed, more or less, by this or that social formation, is anything but straightforward. How, relative to a given set of issues and questions, the theoretical picture is most fruitfully put together remains a substantive question not least because it is all too easy for the conceptual bump in the carpet simply to move as we put pressure now on this notion (achievement), now on that (voluntariness), or on another (privacy).

Interestingly, the 'Western view' of friendship that Carrier proposes (which is conjunctive, – "spontaneous and unconstrained" – and features at least one gradable conjunct) must underdetermine the question that Silver's work shows to have been salient in strands of Western history, and which also thematised Aristotle's writings on the subject: how much constraint (by circumstantial need or desire) on the part of those so related is compatible with friendship? That this question is a refraction of the practical, epistemic difficulty that besets agents dealing with trust, goodwill, and mutuality in the course of life, as many ethnographic studies have reported, and which was stressed by Pitt-Rivers, seems to underline a constitutive dimension of friendship.

If this line is correct, it suggests that the anthropology of friendship connects directly to issues that go back to the founders of the social sciences. It is not just that thinking about friendship and values brings out the antinomies involved in the distinction between the jural and non-jural (and its cognates), which was so important to anthropological efforts to find the functional equivalent of Roman Law in small-scale polities (Roman Law was, via Radcliffe-Brown, the source of the concept of the jural that Fortes set in the foundations of his account of kinship's role in the social order (1969: p. 90)), something thought of as a *sine qua non* of social order itself. Rather, the vision of the individual social being that classical theoretical formulations presuppose, and which the problem of order was thought to require for its solution, makes the non-substitutability of the friend or beloved something incomprehensible.

After characterizing the obligatory and external source of "my duties as a brother, a husband or a citizen" Durkheim (1982) grounds sociology in the significance of "types of behavior and thinking external to the individual, but [which] are endowed with a compelling and coercive power by virtue of which, whether he wishes it or not, they impose themselves upon him" (pp. 50, 52). This line hardened later:

> Human beings as individuals are objects of study for physiologists and psychologists. The human being as a person is a complex of social relationships. He is a citizen of England, a husband and a father, a brick-layer, a member of a particular Methodist congregation, a voter in a certain constituency, a member of his trade union, an adherent of the Labour Party, and so on. Note that each of these descriptions refers to a social relationship, or to a place in a social structure. Note also that a social personality is something that changes during the course of the life of the person. As a person, the human being is the object of study for the social anthropologist."
>
> (Radcliffe-Brown, 1952: p. 194)

In Radcliffe-Brown's formulation, becoming friends, (like getting married or divorced), are aspects of the projects of 'individuals' – human beings, in "actually existing social relations" with one another (1952: p. 192) – not of the 'persons' that the social anthropologist is concerned to theorize over, which are complex relations the individual is compelled to enter into that are defined by norms and organized into institutions.

It is not coincidental that the professions and other institutionally defined aspects of Western modernity loom so large in these attempts to define 'place in a social structure.' (In passing, we might note that Durkheim, unlike some of his intellectual descendants, avoided equating what it is for someone to *be*, for example, a brother, and what it is to *act* in accordance with whatever behavioral norms apply to brothers.) In the more highly articulated forms of structural-functionalism developed by Talcott Parsons, the forms peculiar to the twentieth-century nation-state were even more conspicuous in accounts of social structure. There too the friction evident in Fortes' (1969) efforts to differentiate between familial morality and the "kinship rules and concepts" that determine "citizenship in the political community" (p. 87) is generated again.

Without laboring the point with some of the more exotic solutions to the empirically unavoidable problem of 'role conflicts' that this perspective projects upon folks, we note that the difficulties of this sort of approach became clear; but that did not prevent their recurrence in alternative theoretical visions. Today, there is more diversity within the social theory than at any point in its history. Yet, the problem posed by the anthropology of friendship remains: the separation of the motivational bases of interaction and the behavioral evidence available to an observer, in the context of empiricist assumptions that the latter suffices for the specification of action and social norms. The diremption of the intentionality of agency and the behavior of bodies (common to psychological behaviorism and classical social theory and derived from positivistic pictures of science) has proved fateful for social science as a whole, not just in respect of friendship.

The conceptions of social facticity that attracted twentieth-century social scientists is routinely seen as grounded in Hobbesian visions of humans as psychological egoists; Spencer

and Comte (the latter coined the word 'altruism') drew on this vision in forging a sociology that understands human social life to be predicated on the triumph of altruism over our inherent egoism; Durkheim retained a perspective in which the 'dualism of human nature' is basic. Just what that dualism entails, and what significance it has for social theory, is still a fundamental issue for several disciplines, but there is no doubt that Durkheim, at different periods, and with different degrees of clarity, presented the human individual sometimes as that which socialization *controls*, and sometimes as that which socialization *produces*, through a postpartum process that is no less crucial to the product itself than genetic mechanisms and prepartum epigenesis. On the first view, social embeddedness constrains or molds, and, on the second, it is a crucial part of what constitutes, a human person/ego: on the first view the ego is, so to speak, motivationally stratified (and sometimes conflicted), while on the second, the ego is merely motivationally variegated (and sometimes conflicted). There are different possibilities, of course, within this division: for example, on the first perspective there is room for deep differences on the nature of that upon which socialization has its effects, from the ontotheological 'bearer of freedom' to the product of Pleistocene selection pressures. That these visions were differentially taken up subsequently and that they have different implications for the constitution of the 'domains' or 'realms' of social life, is obvious. They also tend to be more or less congenial to those who entertain the thought that social science is relevant to cultural critique and hope that an improvement in future prospects is possible. Finally, these visions have somewhat different implications for ways of thinking about the relationship between friendship and kinship as analytical foci for anthropology.

The point of making these remarks is not simply to bring this piece to closure through a gesture toward its beginnings. We have already cited Coleman's summary views – so evocative of Parsons – about the combination of 'voluntarism' and 'constraint' at work in the formation of social relationships. There is also some contemporary interest in notions like 'doing' family or friendship, 'performing' relatedness or kinship, and 'kinning' as a social practice. Consideration of the anthropology of friendship and its struggles suggests that such formulations preempt long-standing puzzles, or presuppose, without stating, contestable answers to some of the fundamental questions, and reminds us of the possibility – indeed, the standing possibility – of the presence of the theoretical past in the present.

See also: Altruism and Prosocial Behavior, Sociology of; Family and Kinship, History of; Friendship During Adolescence and Cultural Variations; Friendship During Infancy and Early Childhood and Cultural Variations; Kinship in Anthropology; Kinship, Evolution of; Trust, Sociology of.

Bibliography

Adams, R., Allan, G., 1998. Reflections on context. In: Adams, R., Allan, G. (Eds.), Placing Friendship in Context. Cambridge University Press, Cambridge, pp. 183–194.

Aguilar, M.I., 1999. Localized kin and globalized friends: religious modernity and the 'educated self'' in East Africa. In: Bell, S., Coleman, S. (Eds.), The Anthropology of Friendship. Berg, Oxford, pp. 169–184.

Bell, S., Coleman, S. (Eds.), 1999. The Anthropology of Friendship. Berg, Oxford.

Boissevain, J., 1974. Friends of Friends: Networks, Manipulators and Coalitions. Basil Blackwell, Oxford.

Carrier, J., 1999. People who can be friends: selves and social relationships. In: Bell, S., Coleman, S. (Eds.), The Anthropology of Friendship. Berg, Oxford, pp. 21–38.

Coleman, S., 2010. Making friendship impure: some reflections on a (still) neglected topic. In: Desai, A., Killick, E. (Eds.), The Ways of Friendship: Anthropological Perspectives. Berghahn, Oxford, pp. 197–205.

Derrida, J., 1997. Politics of Friendship (G. Collins, Trans.). Verso, New York.

Desai, A., Killick, E. (Eds.), 2010. The Ways of Friendship: Anthropological Perspectives. Berghahn, Oxford.

Durkheim, E., 1982. In: Lukes, S. (Ed.), The Rules of the Sociological Method (W.D. Halls, Trans.). Free Press, New York.

Eisenstadt, S., 1956. Ritualized personal relations: blood brotherhood, best friends, compadre, etc.: some comparative hypotheses and suggestions. Man 56, 90–95.

Evans-Pritchard, E.E., 1933. Zande blood-brotherhood. Africa 6, 369–401.

Fausto, C., 2012. The friend, the enemy, and the anthropologist: hostility and hospitality among the Parakanã (Amazonia, Brazil). Journal of the Royal Anthropological Institute 18 (1), 196–209.

Firth, R., 1936. Bond friendship in Tikopia. In: Burton, L.H.D. (Ed.), Custom is King. Hutchinson, London, pp. 259–269.

Firth, R., 1999. Preface. In: Bell, S., Coleman, S. (Eds.), The Anthropology of Friendship. Berg, Oxford, pp. xi–xv.

Fortes, M., 1969. Kinship and the Social Order. Routledge, London.

Hart, K., 1973. Informal income opportunities and urban employment in Ghana. The Journal of Modern African Studies 11, 61–89.

Hruschka, D.J., 2010. Friendship: Development, Ecology, and Evolution of a Relationship. University of California Press, Berkeley, Los Angeles and London.

Killick, E., 2009. Ashéninka amity: a study of social relations in an Amazonian society. Journal of the Royal Anthropological Institute 15 (4), 701–718.

Leyton, E. (Ed.), 1974. The Compact: Selected Dimensions of Friendship. Institute of Social and Economic Research, Memorial University of Newfoundland, Canada (Orig, 1970. Humanities 6 (2)).

Marshall, M., 1977. The nature of nurture. American Ethnologist 4 (4), 643–662.

Mintz, S., Wolf, E., 1950. An analysis of ritual co-parenthood (Compadrazgo). Southwestern Journal of Anthropology 6, 341–368.

Pahl, R., 2000. On Friendship. Polity, Cambridge.

Pahl, R., 2002. Towards a more significant sociology of friendship. European Journal of Sociology 43, 410–423.

Pahl, R., Spencer, L., 2010. Family, friends and personal communities: changing models-in-the-mind. Journal of Family Theory & Review 2, 197–210.

Paine, R., 1969. Search of friendship: an exploratory analysis in 'Middle-Class' culture. Man (NS) 4, 504–524.

Paine, R., 1974. In: Layton, E. (Ed.), The Compact: Selected Dimensions of Friendship. Institute of Social and Economic Research, Memorial University of Newfoundland, Canada.

Paine, R., 1999. Friendship: the hazards of an ideal relationship. In: Bell, S., Coleman, S. (Eds.), The Anthropology of Friendship. Berg, Oxford.

Pakaluk, M., 1998. Aristotle. Nicomachean Ethics Books VIII and XI. Translation with Commentary. Clarendon Press, Oxford.

Pitt-Rivers, J., 1954. People of the Sierra. Criterion Books, New York.

Pitt-Rivers, J., 1968. Kinship: III. Pseudo-kinship. International Encyclopedia of the Social Sciences 8, 408–413.

Pitt-Rivers, J., 1973. The kith and the kin. In: Goody, J. (Ed.), The Character of Kinship. Cambridge University Press, Cambridge.

Radcliffe-Brown, A., 1952. Structure and Function in Primitive Society. The Free Press, Glencoe, IL.

Sahlins, M., 2011. What kinship is (part one). Journal of the Royal Anthropological Institute (N.S.) 17 (1), 2–19.

Schmidt, J.F.K., Guichard, M., Schuster, P., Trillmich, F. (Eds.), 2007. Freundschaft und Verwandtschaft. Zur Unterscheidung und Verflechtung Zweier Beziehungssysteme. Universitätsverlag, Konstanz.

Silver, A., 1989. Friendship and trust as moral ideals: an historical approach. European Journal of Sociology 30 (2), 274–297.

Silver, A., 1990. Friendship in commercial society: eighteenth-century social theory and modern sociology. American Journal of Sociology 95 (6), 1474–1504.

Silver, A., 1997. Two different sorts of commerce – friendship and strangership in civil society. In: Weintraub, J., Kumar, K. (Eds.), Public and Private in Thought and Practice: Perspectives on a Grand Dichotomy. University of Chicago Press, Chicago, pp. 43–74.

Silver, A., 2003. Friendship and sincerity. Sozialer sinn. Zeitschrift für Hermeneutische Sozialforschung 4, 123–130.

Uhl, S., 1991. Forbidden friends: cultural Veils of female friendship in Andalusia. American Ethnologist 18 (1), 90–105.

Whyte, W.F., 1943. Street Corner Society. University of Chicago Press, Chicago.

Friendship During Adolescence and Cultural Variations

Hildegunn Fandrem, University of Stavanger, Stavanger, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

First, in this article previous research on the issue of friendship during adolescence is reviewed regarding cultural variations in the understanding of differentiating between individualistic and collectivistic cultures. The main theme is then to raise issues that may question the importance and appropriateness of differentiating cultural variations in these two categories. How the identity consists of different dimensions is focused. Furthermore, the impact of social and cultural changes on peer interaction and friendships is paid attention to. This article concludes with a discussion of future directions for friendship in research on adolescence and cultural variations.

It is argued that theories of friendship have, by and large, been developed with little regard for notions of culture (e.g., Howes, 2009). Rogoff (2003) uses the concept of cultural community to include the aspect of a grouping of people who participate in a shared set of practices and traditions. In this regard, peer group represent a form of cultural community. However, the focus in this article, when it comes to culture, is more cultural variations meaning different national cultures, or characteristics of cultures based on geographic location in the world. The idea of differentiating between individualistic and collectivistic cultures, and research regarding this differentiation related to friendship are presented. The main theme of this article is then to raise issues that may question the importance and appropriateness of differentiating cultural variations in the categories of collectivistic and individualistic cultures.

Generally, friendship formation may be understood by attachment theory (Bowlby, 1969), as the processes of relationship formation embedded in Bowlby's attachment theory are applicable for formation of friendship relationship (Howes, 1996). According to this interpretation, relationships develop through multiple and repeatedly interactive experiences. These experiences, and furthermore, making and keeping friends requires social and emotional regulation skills and social cognition. Thus, theories of social and emotional socialization may also be relevant to the understanding of friendships. Howes (2009) has argued that to form friendships, individuals must recognize the potential friend as a social actor and be aware of the reciprocal nature of social behaviors in social interaction. Harris (1995) claims that cultural norms, either adopted from existing cultural systems or created by children themselves, serve as a basis for children's peer interactions and mutual socialization. Similarly, sociocultural theory (e.g., Cole, 1996; Vygotsky, 1978) emphasizes that culture influence human activities. Thus, the individual internalizes external signs, along with their cultural meanings, during collaborative or guided learning.

While friendships during early childhood may be said to make an important contribution to the development of social cognitive – and social skills beyond experiences with adults (Dunn and Hughes, 2001), friendships during years of youth is however, characterized, or influenced, by also the notion of who one is, or want to be, ones identity. A person's identity is partly influenced by the person's cultural belonging, or maybe also whether the person have several cultural belongings. Perspectives regarding formation of identity are the first issue to be discussed in this article regarding questioning the idea of differentiating between individualistic and collectivistic cultures. Then how the motive for establishing friendships may vary between different groups is focused. Cultural variations will especially in this discussion include the focus on immigrant adolescents, as representatives for multiple cultures. Special attention is paid to aggression as a strategy to gain friendships. The issues will be discussed with regard to recent research. Conditions for social networks and friendship establishment are then discussed, with ethnicity as one of several important factors. Last, the impact of social and cultural changes on peer interaction and friendships is paid attention to. Thus, the discussion of whether differentiating in individualistic vs collectivistic cultures is appropriate is related also to the aspect of modernizing of societies all over the world.

Adolescents Friendships in Individualistic vs Collectivistic-Oriented Cultures

The Role of Adolescence and Peer Experience across Cultures

The most classic and fundamental question in cross-cultural adolescent research is whether all cultures in the world view adolescence as a unique life period. Even if time and duration of adolescence may differ between countries, all societies might be said to have the notion of adolescence. All over the world the time of adolescence is normally relatively brief, about 2 years for girls and 2–4 years for boys (Berry et al., 2002). Exactly how long the adolescent period lasts is, however, dependent of the cultures' demands according to what training is needed for adult roles. For instance, in India children have to fulfill adult tasks from an early age, thus not as much time and attention can be spent on adolescent activities as in the Western world (Saraswathi, 1999). As a consequence, friendship formation in adolescence may play a secondary role in countries like India. Peer interaction is, however, a fundamental part of human development in most societies. Even in India, or other cultures where children are required to spend much of their time helping with family duties, such as caring for their

younger siblings, children and adolescents often do this in company of peers (Tietjen, 2006).

Although research on peer relationships and adolescents friendships has traditionally been conducted in Western, particularly North American, cultures, there has been increased interest in peer relationships in different national cultures in the past two decades. Most of these studies consistently support the view that individual's peer experiences are likely to be affected by cultural norms and values (Chen et al., 2009).

Conditions for Friendship Formation in Individualistic vs Collectivistic Cultures

More specifically, different cultures may place different values on, e.g., social initiative and norm-based behavioral control in adolescents. In Western cultures, which is often characterized as self-oriented or individualistic cultures (e.g., Hofstede, 2001) acquiring autonomy and assertive skills is an important socialization goal, and also social initiative is viewed as a major index of social maturity. In addition, self-regulation and control are encouraged. The cultural emphasis on individual decision making, however, allows individuals to maintain a balance between the needs of the self and those of others. Thus, behavior control is often regarded as less important, especially if it is in conflict with the attainment of individual goals (Triandis, 1995). On the contrary, in non-Western cultures that are characterized as more collective or group-oriented societies, social initiative may not be highly valued because it may not facilitate harmony and cohesiveness in the group. In these cultures to maintain group harmony is encourage, thus, individuals need to restrain personal desires in an effort to address the needs and interest of others. Self-control may be emphasized in a more consistent and absolute manner, and the lack of control may be regarded as a serious problem during childhood and adolescence (Triandis, 1995).

How active and prosocial–cooperative behavior vs shyness is valued and perceived is also found to differ between individualistic and collectivistic cultures. Peers and adults in Western cultures regard children and adolescents who are active and assertive in their peer interactions as socially competent (e.g., Rubin et al., 2002). Thus, sociability is valued and may be seen as an important aspect to be accepted by peers and to achieve success in friendship formation. Opposite, adolescents who display shy and wary behaviors are often viewed as socially incompetent and immature. Research shows that shy-inhibited students are likely to experience low peer acceptance and poor social adjustment in Western cultures (e.g., Coplan et al., 2004). Especially shy-inhibited boys in North America have been found to experience difficulties in peer relationships (e.g., Caspi et al., 1988). Result from another study showed that when shy-inhibited children in Canada made passive and low-power social interactions, they were more likely than others to receive fewer positive responses and more rejection from peers (Chen et al., 2006), while shy-inhibited Chinese children who displayed the same behavior (made passive and low-power interactions) were, on the other hand, more likely to receive positive responses and being accepted by peers (Chen et al., 1995). Thus, shyness-inhibition is not necessarily perceived as social incompetence or associated with social adjustment problems in collectivistic-oriented cultures, where social initiative and assertiveness are less valued. As Indonesian children and adolescents are found to more often display shyness-inhibition compared to Americans (Farver and Howes, 1988), how behavior is perceived, or to which degree it is valued, may also influence the prevalence of the actual behavior.

When it comes to prosocial–cooperative behavior individual in Western societies attempt to maintain a balance between prosocial concerns and individual freedom of choice. Thus, interpersonal cooperation and prosocial behavior, and thus friendship establishments, are seen as a personal decision based on factors such as how much one like the target person (Greenfield et al., 2006). In collectivistic-oriented cultures, responsiveness to the need of others may, on the other hand, be seen as a fundamental commitment (Miller, 1994). Thus, there might be a considerable pressure on children in these cultures to view prosocial cooperativeness as obligatory.

Researchers have, however, found little cultural variations in the linkage between prosocial–cooperative behavior and the quality of friendships, this suggests that social behavior based on the consideration of the interests of others or the group may be a universal antecedent of peer acceptance, and thus a universal factor for friendship establishment. This suggests that aspects of culture in friendship formation might be overestimated.

In most cultures, children and adolescents usually have reciprocal friends (e.g., French et al., 2003; Schneider et al., 1997). One may, however, argue that there might be cultural variations in the extent to which group integration or development of friendships is encouraged (e.g., Sharabany and Wiseman, 1993). As involvement in peer relationships increases from childhood to adolescents in North America, because of the desire of youth to become autonomous from the family (Rothbaum et al., 2000), more collectivistic-oriented cultures emphasize conformity to others and commitment to social relationships. In group-oriented cultures, children and adolescents are expected and encourage to maintain strong social affiliations and to identify with the group (Sharabany, 2006). Thus, involvement with age mates and friendship formation typically occurs within extended family networks of siblings and cousins, leaving little opportunity for relationships with nonkin. There may also be less dyadic friendships in group-oriented cultures (Sharabany and Wiseman, 1993). However, schools may be an arena that, independent of cultural orientation, bring together large numbers of nonkin, same-age adolescents, thus providing opportunities to form friendships also across family bonds.

The Role of Friendship in Individualistic vs Collectivistic Cultures

Friendship intimacy is associated with constructs as sensitivity, mutual understanding, exclusivity, attachment, emotional closeness, and trust, and may therefore be difficult to understand across cultures (Chen et al., 2009). Chen and French (2008) argue, however, that the extent to which these different aspects are emphasized may vary across cultures. Although it is claimed that friendship in collectivistic-oriented cultures seem to be more intimate than those in

individualistic-oriented cultures (Triandis et al., 1988), Sharabany (2006) suggests, based on his research among Arab and Jewish children in Israeli Kibbutz, that in some collectivistic societies, friendships are characterized as low in intimacy because of the availability of other sources of emotional support. This is also supported by the studies of French et al. (2005, 2006a) finding Indonesian to report lower intimacy than their North American counterparts. Indonesian students have also been found to be less likely to develop close friendships, because they are encouraged to focus on integration into general peer networks and larger community (Jay, 1969). However, friendships among South Korean adolescents have, e.g., been found to be highly intimate (French et al., 2006a,b). In addition, Schneider's research shows that more close and harmonious friendships are reported among Italian (Central and Southern), than among Canadians. Also, within the United States, e.g., Latino adolescents perceive their friendships as more intimate and closer than adolescents in European American and other individualistic-oriented groups (Gonzales et al., 2004; Way, 2006).

While in Western cultures the support of friends may be an important mechanism to accomplish the socialization goal of developing confidence and positive views about the self (Sullivan, 1953), Chen and French (2008) argue that the role of friendship in promoting self-worth may be less salient in many non-Western cultures, where the development of self is not considered an important developmental task. This is supported by research on Chinese (Chen et al. 2004), Indonesian (French et al., 2005), and Arab and Caribbean (Dayan et al., 2001) children, as none in these studies reported the enhancement of self-worth as an important function of friendship. Instead of self-validation, research shows that instrumental aid and assistance, like e.g., sharing money, protection from harm, helping others solve problems and learn social skills appears to be important aspects for friendship in Asian nations, such as China (Chen et al., 2004), Indonesia (French et al., 2005), South Korea (French et al., 2006a), the Philippines (Hollsteiner, 1979), and in Latinos societies, such as Cuba (Gonzales et al., 2004) and Costa Rica (DeRosier and Kupersmidt, 1991) and also in low-income black and Hispanic adolescents in America (Way, 2006).

Research in North America shows that as interaction within small groups decline from middle childhood to adolescence, affiliation with multiple groups and larger crowds becomes more salient during adolescence (e.g., Brown, 1990). The various peer groups than can be identified; e.g., 'jocks' 'brains,' 'nerds' etc. may be said to reflect individualistic values as they are more individual characteristics than they focus on relations (Hofstede, 2001). Thus, adolescents in Western cultures may be said to form mainly peer groups on their particular individual interest. As adolescents increasingly seek independence and attempt to avoid group restrictions, there is furthermore a general loosening of group ties, consequently, adolescent's sense of belongingness decline steadily with age (e.g., Rubin et al., 2006). In general, the role of peer groups in adolescents is often highlighted as a major social resource for fulfilling individual psychological needs such as formation of self-identity and the development of positive self-perception and self-feelings (Chen et al., 2009). In collectivistic-oriented cultures, the balance between the pursuit of independence and personal identity, and the commitment to group undertaking tend to be less evident compared to individualistic-oriented cultures. Research shows that adolescents from e.g., China usually describe group activities in terms of how they are in accordance to adult's social requirements and standards, such as maintaining interpersonal cooperation and collectivistic well-being (Chen et al., 2004). Thus, we may conclude that the culture may pay particular attention to the nature of peer groups and friendships, in terms of whether social relationships or group activities are guided by the 'right' social goals.

The Role of Culture in Formation of Identity in Adolescence

All over the world identity formation may be defined as a lifelong process characterized by cycles of exploration and consolidation, as well as experiences of competence and vulnerability (Grotevant and Cooper, 1998). Although the process is lifelong, it takes new dimensions during years of youth because of the confluence of physical, cognitive, and relational changes during this period. Thus, there are reasons to believe that adolescence, especially early adolescence, may be a turbulent period (Chen and Ferruggia, 2002). However, the time of adolescence is not necessarily a period of 'storm and stress,' as claimed by Western developmental and clinical psychologists throughout most of the twentieth century (Hall, 1904). Berry et al. (2002) claims that the time of adolescence is most of all a time for learning new social roles, with attendant psychological tensions. However, deciding who one wants to be may not necessarily be an easy task, and difficulties in identity development have been associated with serious problematic behavior, as e.g., substance abuse (Jones et al., 1989). Generally, adolescence may be said to be the most difficult time of life. This claim is supported by research, which generally finds depression to be rare among children (Kessler et al., 2001) while among adolescents depressive mood is found to be higher than among their parents (Kandel and Davies, 1982).

Descriptively, a range of aspects is traditionally held to concern identity, independent of where one is from. Dimensions like nationality, ethnicity, gender, social memberships and group characteristics, individual character, personality, and social preferences are all part of the identity (Verkuyten, 2005). Hence, identity consists of both collectivistic (e.g., ethnic) and personal (e.g., individual characteristics) dimensions. *Ethnic identity* may be said to refer to the idea of common origin (Kanstad, 1994) as it comprises identification with one's ethnic group or culture of origin. This dimension is, furthermore, an example of a stable dimension. The *national identity*, on the other hand, refers to the country or nation, in which one lives. Hence, it comprises identification with the larger society surrounding the person and thus it changes with resettlements. The concept of *cultural identity* covers both the ethnic and national dimensions. Triandis et al. (1990) define culture as those constructs shared by speakers of a particular dialect, living in a proximal location, during the same historical period. Thus, culture may not necessarily refer to only a country. Important in this sense is, however, that cultural identity refers to more than the group category that one chooses, it involves a sense of belonging to one or more cultural groups and the

feelings associated with a group membership (Phinney, 1990). If culture refers to countries it is important to have in mind, regarding adolescents that have migrated, that national and ethnic identity differ for the same individual. Furthermore, the relationship between the two dimensions needs to be in focus as it might have an influence on psychological and sociocultural adaptation, and the possibility for friendship formation for this group of adolescents. *Social identity* may also be said to occur when one discover that one have something in common to someone, at the same time as one discover that there are people one have little in common with. This type of collectivistic identity comprises the notion of a social group (e.g., gender, status, age). Thus it is also less stable. Friendships play an important role in the formation of especially social identity. At the end all the collectivistic dimensions of the identity (ethnic, national cultural, and social) are given a personal expression through the *personal identity*. The personal dimension of the identity is established on the basis of each individual's personal experiences. Thus, personal identity is never the same among different individuals. It is also claimed that because of increased *global* and *social* mobility the personal, or subjective, aspect of the identity becomes more and more important (Fandrem, 1996).

From this we may claim that identity is constructed and maintained through social interactions, it changes over time and context, and across generations. It is, however, through the time of adolescents that the formation of the identity is an especially critical developmental task (Marcia, 2002; Phinney et al., 2006). How the identity process develops is highly dependent on the peers, the family, the community, and the context to which the adolescent is exposed. As mentioned earlier, independent of culture, parents (vertical relations) are the most significant others to the individual in the earliest years of life. With increasing age, especially during years of youth, peers (horizontal relationships), however, become increasingly significant. Research have shown that during years of youth, peers become more important to the individual, and the amount of time spent with family members decreases (Laible et al., 2000; Larson et al., 1996).

Considering that cultural variation in this article also includes variations regarding individuals belonging to more than one culture, immigrant adolescents should be left a comment. In immigrant families, especially the mother is the one who represents the culture of origin. In addition to generally being more interested in keeping traditions, mothers often spend more time with people from the same origin as themselves, especially if they are not employed and thus through work gain social relationships with people from the dominant culture. Adolescents, on the other side, gain social relationships with peers from the dominant culture at school. Thus, for immigrant adolescents, parents and peers may represent more distinct values, compared to what is the case for native adolescents. Consequently, immigrant adolescents may struggle more than natives during development of their identity. For adolescents in immigrant families, formation of ethnic identity involves examining also the ethnic attitude, values, and practices learned at home from their immigrant parents and considering them in relation to those of their peers (Phinney, 1989).

Regarding the different dimensions of the identity, for adolescents with immigrant background also the concept of *ethnic minority identity* (Phinney, 1990) may be used, besides *majority identity* or *host national identity*, which refers to the individual's identification with the majority or host society. It has been suggested that it is possible to have high or low identification on both constructs, or a high identification on one construct, and a low one on the other (Phinney, 1990; Sanchez and Fernandez, 1993). Furthermore, a strong sense of ethnic identity devoid of positive identification with the larger society may aggravate psychological conflicts; this may in turn influence the opportunity to establish friendships for this group.

The Role of Aggression in Gaining Friendships

Western cultures, where competitiveness and the pursuit of personal goals are in focus, seem to allow, or also encourage, aggressive behavior to some degree. Despite the general discouragement of this kind of behavior, research shows that aggressive adolescents may receive social support from peers, and are sometimes even perceived as 'stars' in their peer groups (Cairns and Cairns, 1994). Opposite, non-Western cultures, where group harmony and personal control are important, tend to inhibit aggressive behavior (e.g., Bergeron and Schneider, 2005). This may influence the prevalence of aggressive behavior in different cultures.

There may be cultural variations also in the expectations of how behavior is to be performed. For example, bullying is an externalizing behavior problem, in which there are three key elements: (1) it is a type of aggressive behavior; carried out toward a victim, (2) it is repeated over time, and (3) there is an imbalance of power in favor of the perpetrator(s) when the bullying occurs (e.g., Olweus and Roland, 1983; Smith et al., 2002). Although this definition is widely accepted across different cultures, bullying may be conceptualized slightly different according to language and culture (Smith et al., 2002). In, e.g., Japan the concept of bullying, which is 'ijime,' includes an aggressive behavior, which is performed together with somebody. The collectivistic dimension (Hofstede, 2001) is automatically included in the use of this concept. Bullying may appear in different ways, for instance, by teasing or direct verbal aggression (such as name calling), active exclusion from a social group, also called indirect aggression, and/or by physical assaults or direct physical aggression (e.g., Björkqvist et al., 1992; Roland, 1998). These different forms are however identified across different cultures.

Prevalence rates of bullying vary according to methods used and samples investigated, and as mentioned this may vary in different countries as, e.g., different concepts are used. A systematic comparison of bullying among girls and boys between different countries is done in a study of Craig and Harel (2004). For the 35 other countries in this study the percentage was ranging between 2.3% (Sweden) and 29.5% (Lithuania) for girls and 5.1% (Sweden) and 49.8% (Lithuania) for boys. Generally, independent of which country the bullying is taking place, the figures of being victimized decrease with increasing age through primary and secondary school, while the figures for bullying others seem to be more stable, or there is a small increase, during the same period (Craig and

Harel, 2004). In addition, gender differences do not seem to differ between countries. Bullying is found to be more frequent among boys compared to girls, the gender difference is however smaller concerning victimization (Craig and Harel, 2004; Olweus, 1978).

Regarding characteristics of bullies and victims, there is solid evidence for concluding that victims as a group are different from other individuals not victimized (Hawker and Boulton, 2000; Isaacs et al., 2001). Focusing on bullies, most researchers now consider aggressiveness as a strong predictor for bullying others (e.g., Berkowitz, 1993; Dodge, 1991; Olweus, 1978; Olweus and Roland, 1983; Roland, 2002). Aggressiveness may be defined as the trait, or tendency, to wanting to hurt or attack someone. In the research literature, *reactive* and *proactive* aggressiveness are the concepts often used for the two different main types of aggressiveness (e.g., Card and Little, 2006; Dodge, 1991; Vitaro et al., 2006). While *reactive* aggressiveness is theoretically grounded in the frustration-aggression model and the behavior occurs as an angry reaction to a perceived frustration (Berkowitz, 1989), *proactive* aggressiveness has its roots in social cognitive theory and is related to a planned behavior controlled by external rewards and reinforcement (Bandura, 1986; Vitaro et al., 2006). Thus, there are different drives in the two different types of aggressiveness; while anger is the central emotion in reactive aggressiveness, pleasure through social rewards is the dominant emotion in proactive aggressiveness. Only a few studies have empirically investigated reactive and proactive aggressiveness among bullies and nonbullies (e.g., Salmivalli and Nieminen, 2002). Roland and Idsøe (2001) distinguished between power-related proactive aggressiveness and affiliation-related proactive aggressiveness while investigating the associations between different types of aggressiveness and bullying behavior. In power-related proactive aggressiveness, it is a goal for the individual to act aggressively to feel power over the victim. In affiliation-related proactive aggressiveness, the goal is on the other hand to gain affiliation with another individual with whom he or she bullies others. Thus, affiliation-related proactive aggressiveness might to some degree be related to the drive of establishing friendships. In the study of Roland and Idsøe (2001), power-related proactive aggressiveness was found to be associated with bullying behavior among boys in eighth grade. While for girls in grade eight the results indicated higher associations between bullying others and affiliation-related proactive aggressiveness. Thus, adolescent girls seem to use bullying as a strategy to gain friendships more than boys.

In the last years, the focus on bullying as a group phenomenon has increased. Affiliation and peer acceptance as underlying motives of aggressive behavior make it relevant to focus on peer groups when studying aggressive behavior and bullying. When viewing bullying as a group phenomenon, the bullying behavior is determined not only by characteristics of bullies and victims but also by social relationships or roles within the group (e.g., Salmivalli et al., 1996). Many of the studies on bullying and peer groups focus on social status and view bullying as an attempt to gain social status and friends, or to be accepted, by peers (e.g., Olthof and Goossens, 2008; Salmivalli and Peets, 2008; Veenstra et al., 2007). However, studies applying these ideas to better understand peer relations across different cultures are still relatively scarce. Those found emphasize, however, the affiliation motive to be strongly related to bullying for immigrant boys, but not to the same degree for native boys (Fandrem et al., 2010; Fandrem et al., 2009; Strohmeier et al., 2012a,b). Thus, even if research that investigates level differences between natives and immigrants in different countries are not conclusive (e.g., Boulton, 1995; Eslea and Mukhtar, 2000; Moran et al., 1993; Strohmeier and Spiel, 2003), there seem to be consistence across cultures regarding motives for bullying. Furthermore, when immigrants and natives in, e.g., Austria and Norway are compared the studies show similar results (Strohmeier et al., 2012a). Thus, despite different cultures, e.g., different migration history, different national policy for migration and immigrants, differences in the composition of the immigrant groups in the country, there are some psychological mechanisms across different nations that are the same when it comes to need for friendships and strategies for achieving friendships. Independent of which culture one originates from, or live in, it seems like being multicultural increase the chance for using aggression as a strategy to gain friendships. Thus, it is not the *kind of culture* that necessary matters, but *being in the situation of feeling loss of cultural belonging* may influence how friendships are established.

Other Contextual Conditions for Social Networks and Friendship Establishment

As mentioned, time spent with peers increase during years of youth, usually on the expense of time spent with family members (Larson et al., 1996). Furthermore, as peer attachment is found to be relatively more influential on adolescent adjustment than parent attachment (e.g., Laible et al., 2000), peers play the most crucial role in adolescent's social networks. A social network may be defined as a set of connections between a limited set of persons, and in addition, characteristics of the connections can be used to understand the social behavior of each person (Mitchell, 1969). Networks are also formed and sustained by the cohesion or attachment, which is to be found between the persons in the network (Bøe, 1993). From this, we may also conclude that affiliation seems to be a central mechanism in a peer network.

In all cultures, all over the world, most adolescents affiliate with a peer group. In the first years of youth peer groups often comprise same-sex members, with more mixed-sex groups appearing in late adolescence. In both individualistic and collectivistic-oriented cultures more girls than boys form mutual friendships, and generally girl's friendships tend to be more stable over time (e.g., Beneson et al., 1997; Schneider et al., 1997). The size of peer groups seems to increase with age. Boys' groups tend to be larger than girls' groups. Research in many countries, including the United States, China, Japan, Singapore, and Senegal, suggest, furthermore, that boys more often than girls belong to groups and are more susceptible to peer pressure than girls (e.g., Beneson et al., 1997; Dekovic et al., 2002; D'Hondt and Vandewiele, 1980; Salmivalli et al., 1997; Sim and Koh,

2003). However, in some other studies in the United States and China, differences between boys and girls are not reported (Cairns et al., 1995; Chen et al., 2003).

Adolescents form several smaller networks or subgroups, inside a class at school or on different leisure arenas. Within these smaller peer networks group mechanisms connected with aggressive behavior might be even more powerful than at the class level, because social influence can be especially effective within small friendships groups (Salmivalli et al., 1997). Research shows that both ethnicity and aggressive behavior make young people affiliate with one another (Cairns et al., 1988; Espelage et al., 2003; Farmer et al. 2003; Glueck and Glueck, 1950; McPherson et al., 2001; Salmivalli et al., 1997). More specifically, e.g., the study of Salmivalli et al. (1997) found that Finnish students who behaved in similar or complementary roles in bullying situations (bullies, assistants/reinforces, defenders, victims, and outsiders) established networks with each other. Also, among Chinese children homogeneity has been found regarding leadership status, aggression, and learning problems (Chen et al., 2003). Thus, peer groups are likely to be homogenous on social and behavioral attributes across different cultures. But, homogeneity is not necessarily the rule; inequality according to both ethnicity and aggressiveness are found, especially proactive aggressive adolescents, and nonaggressive adolescents are found to interact with each other (Estell et al., 2007; McPherson et al., 2001).

Regarding immigrant adolescents, social contact with peers from the national vs ethnic group is a central factor in the acculturation process, i.e., the process of psychological and sociocultural adaptation one face when moving to another culture (e.g., Berry, 1997). Social interactions, or friendships, are widely used in assessing acculturation (e.g., Berry et al., 2006). An immigrant may be involved with members from the larger, or host, society, or with others from ones own ethnic group. Time is one aspect that may influence degree of contact with members from the larger society vs culture of origin. Immigrants usually extend their contact with the host society over time. Titzmann et al. (2012) found that using the new host language more often appeared to be a crucial variable for lower levels of *friendship homophily* (i.e., the preference for intraethnic over interethnic friends). However, ethnic ties typically remain strong. Several studies suggest the case of in-group preferences; i.e., contact with peers of ones own ethnic group (Hallinan and Williams, 1989; Hamm, 2000; Schofield and Whitley, 1983), especially regarding pupils in small ethnic minorities (Quillian and Campbell, 2003). Another aspect of importance that has been suggested is that the number of intra- and interethnic relationships reflects the opportunity for establishing these relationships (Baerveldt et al., 2004; Moody, 2001). Consequently, the number of available adolescents of different ethnicities influences ones affiliations, and thus the ethnic composition of peer groups and the possibility to form friendships across ethnicity. The study of Titzmann et al. (2012) that was based on information from Russian Jewish and ethnic German Diaspora migrant adolescents (i.e., migrants returning to the country of their ancestors from the former Soviet Union) in Israel and Germany, respectively, also showed higher levels and a less pronounced decrease of friendship homophily in Israel as compared to Germany. Especially, acculturation-related variables were in this study found to be best at predicting intercept and slope of friendship homophily.

Research on adolescents shows that interaction with peers from one's own ethnic group is strongly related to ethnic identity (Phinney et al., 2001), thus contact with ethnic alike peers seems necessary in order to keep ones ethnic identity. Although there are wide individual variations in immigrants, almost all research on ethnic identity carried out in the United States suggest high scores on ethnic identity in this group. Compared to ethnic identity, there has been far less attention paid to the conceptualizing of national identity and to studying immigrants identification with the host society. Phinney and Devich-Navarro (1997) found adolescents to vary according to if they have a strong positive sense of belonging to the host, or larger, society, or if they feel they are excluded or rejected by the larger society. The school is an important arena, in which adolescents form friendships. Studies in schools including adolescents from many different ethnic groups show that the pupils vary to which extent they relate with people with the same ethnicity as themselves or mix with others (e.g., DuBois and Hirsch, 1990; Semons, 1991).

Focusing specifically on the issue of bullying and friendship, research is in line with the research on peer groups and aggressive behavior generally; bullies tend to be friends with other bullies, but are not segregated from peers who are not bullies (Estell et al., 2007; Pellegrini et al., 1999; Rodkin and Hodges, 2003; Salmivalli et al., 1997). However, few studies have investigated which factors, besides having migration experience, influence whether bullies affiliate with other bullies or with their nonaggressive peers.

In the study of Roland and Idsøe (2001), where a distinction of proactive aggressiveness is made (affiliation-related proactive aggressiveness vs power-related proactive aggressiveness) the findings at least shed some light on this aspect. The results shows that in grade five the bullies (both girls and boys) seem to be motivated by affiliation, and not by power at all. In grade eight the pattern changes, as among these older pupils power is found to be associated with bullying behavior as well as affiliation, though this is applicable only for the boys. Among girls in grade eight affiliation-related proactive aggressiveness is still the only type of proactive aggressiveness that is associated with bullying. Thus, age and gender may be factors that influence if bullies affiliate with other bullies. The above-mentioned results suggest that girls use aggression to gain friendships more than boys do, at least in early adolescence. Regarding gender, another study shows that antisocial girls, more than antisocial boys, find it important to be accepted by boys in general, while for boys it was important to be accepted by boys only if they also where bullies (Olthof and Goossens, 2008). Hence, girls who bully are likely to establish friendships also with nonbullies, especially if these are boys.

Bullying is a way to interact with peers. In affiliation-motivated bullying, there is an interaction not only between the bully and the victim only, but also between bullies that bully together, or the group of bullies. If the bullying is mainly associated with affiliation-related aggressiveness, the bully may carry out the aggressive act on order

to feel affiliation with cobullies. Having a common enemy makes the relation between two parts stronger. From this, we may conclude that bullying might be a strategy to gain friendships.

To be affiliated with peers includes being affiliated with what the peers represent. For example, an immigrant boy this means that it is crucial with whom he bullies regarding what he wants to feel affiliated to. If the immigrant has mainly other immigrants as cobullies, they affiliate with their culture of origin, thus the bullying behavior may be interpreted also as a separation strategy (Fandrem et al., 2010). On the other hand, if immigrants bully with both immigrants and natives, they also interact with people in the host culture, the outcome then could be interpreted as integration. This shows how bullying behavior among immigrants also may be related to acculturation strategies, i.e., how one choose to adapt into a new culture (see Berry, 1997).

The Impact of Social Change in Societies

Besides that migration is characteristic of modern society, today's urbanization of traditional rural societies and modernization generally should also be taken into account. Research has shown that, e.g., industrialization of rural areas has resulted in adolescents having fewer interactions with their extended kin group and a decrease on cooperative and obedient behaviors (e.g., Graves and Graves, 1983). Urban lifestyles may be said to increase affluence and change the values in the culture to, e.g., more individual autonomy, which in turn may influence on the character of friendships. Social relationships tend to serve a more utilitarian function in societies where relying on others for basic needs is essential (Guthrie, 2006). Chen et al. (2009) argue that increased value of personal independence and social initiative during modernization may undermine the utilitarian function of relationships and enhance their role in providing emotional support.

More specifically, in China the traditional utilitarian and instrumental type of friendship is losing significance, as the country moves toward a market economy and stronger legal infrastructure (Tamis-LeMonda et al., 2008). Also, studies in Berlin support this notion (Little et al., 1999): When adolescents in former East Berlin was compared to adolescents in West Berlin, the Eastern adolescents reported less fun and enjoyments within their friendships then the Western, largely due to the adult control of their interactions with friends. However, with the changes of the education system after the German unification, friendships and group activities among children in East Berlin seem to be based more on personal choices and are less controlled by adults, thus friendships in East Berlin today are characterized of more play and intimacy than earlier.

On the basis of these studies we may argue that as the whole world get more modernized, adolescents all over the world increasingly appreciate the function of friendships in fulfilling individual psychological needs. Thus, the differentiation between individualistic and collectivistic-oriented cultures will in the future lose it's meaning.

Conclusions and Future Directions

The research reviewed in this article show inconsistency in findings related to collectivistic and individualistic-oriented cultures. This may make us conclude that the nature and quality of adolescents' friendships may depend on complex social, historical, and ecological factors. Thus, it is not necessarily the culture itself that matters. Generally, there seem to be a need for focus on specific socioecological circumstances and factors. Future research needs to take into account specific social circumstances and other historical and ecological factors, as well as emphasizing individuality. Although one may claim that culture may to some degree affect not only overall social engagement but also the nature or characteristics of peer interaction, learning new social roles also has to do with many individual characteristics, e.g., learning ability. External factors as the family, the school, the neighborhood and peers may play an important role in regulation of behavior, psychological autonomy, and connection with significant others during adolescence (Barber and Olsen, 1997). This emphasize that the focus on *national* culture might be overestimated. Taken into account also all the different dimensions of ones identity, future research should to a higher degree pay attention to that why and how people interact and form friendships is due to a multifaceted and complex set of factors.

Also *global* and *social* mobility of today make the individual aspect of the identity increasingly important (Fandrem, 1996). Thus, there might also be large individual differences within the same country or type of culture, i.e., Western countries. Research shows that, e.g., shyness has been linked with fewer negative outcomes in Swedish youth compared to American youth (Kerr et al., 1996; Caspi et al., 1988). This aspect, in addition to how the undergoing dramatic reforms toward a market economy in China have made, e.g., the adaptive value of shyness-inhibition decline in urban children (Chen et al., 2005) emphasizes the need to question whether the differentiation between Western and non-Western cultures should be paid as much attention as it has until recently.

This article has showed us that cultural norms and values may affect the display of significance of adolescent's social behaviors in peer interaction and friendships, and also the structural and functional features of friendships and organization of peer groups. Based on the inconsistency of research results there is, however, a need to emphasize that macrolevel, social (e.g., class, religion, etc.), economic and individual factors also need to be taken into account when studying adolescent relationship and friendships. Thus, within-culture factors should receive greater attention. With the more use of sophisticated methodology, future studies will be able to take into account different levels of social complexity, and thus may reduce the importance that until now has been attributed culture.

In addition, cross-cultural issues in peer relationships and friendships should be examined also from a developmental perspective in future studies. By the use of longitudinal studies antecedents, concomitants, and outcomes, as well as developmental patterns can be identified.

See also: Acculturation; Adolescence, Sociology of; Adolescent Sexual Risk; Behavioral Inhibition and Social Withdrawal across Cultures; Childhood and Adolescence: Developmental Assets; Collectivism and Individualism: Cultural and Psychological Concerns; Counseling and Psychotherapy: Ethnic and Cultural Differences; Cross-Cultural Psychology; Cultural Influences on Interpersonal Relationships; Defiant Behavior During Adolescence across Cultures; Early Emotional Development and Cultural Variability; Ethnic Identity, Psychology of; Ethnicity and Migration in Europe; Friendship During Childhood and Cultural Variations; Friendship During Infancy and Early Childhood and Cultural Variations; Friendship Experiences of Lesbian, Gay, Bisexual, and Transgender Individuals; Friendship, Anthropology of; Immigration: Social Psychological aspects; Intergroup Relations; Personality Changes During Adolescence Across Cultures; Prosocial Behavior During Adolescence; Racism: Social Psychological Perspectives; Self and Identity Development During Adolescence across Cultures; Self-Regulation During Adolescence: Variations Associated with Individual–Context Relations; Social Competence During Adolescence Across Cultures; Social Identity in Social Psychology; Social Networks; Social Psychology.

Bibliography

Baerveldt, C., Van Duijn, M.A.J., Vermeij, L., Van Hemert, D.A., 2004. Ethnic boundaries and personal choice. Assessing the influence of individual inclinations to choose intra-ethnic relationships on pupils' networks. Social Networks 26 (1), 55–74.

Bandura, A., 1986. Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice-Hall, Englewood Cliffs, NJ.

Barber, B.K., Olsen, J.A., 1997. Socialization in context: connection, regulation, and autonomy in the family, school, and neighbourhood, and with peers. Journal of Adolescent Research 12 (2), 287–315.

Beneson, J.F., Apostoloris, N.H., Parnass, J., 1997. Age and sex differences in dyadic and group interaction. Developmental Psychology 33, 538–543.

Bergeron, N., Schneider, B.H., 2005. Explaining cross-national differences in peer-directed aggression: a quantitative synthesis. Aggressive Behaviour 31, 116–137.

Berkowitz, L., 1989. Frustration-aggression hypothesis: examination and reformulation. Psychological Bulletin 106 (1), 59–73.

Berkowitz, L., 1993. Aggression: Its Causes, Consequences, and Control. Mc Graw-Hill, New York.

Berry, J.W., 1997. Immigration, acculturation, and adaptation. Applied Psychology: An International Review 46 (1), 5–34.

Berry, J.W., Poortinga, Y.H., Segall, M.H., Dasen, P.R., 2002. Cross-Cultural Psychology: Research and Applications. Cambridge University Press, Cambridge.

Berry, J.W., Phinney, J.S., Sam, D.L., Vedder, P., 2006. Immigrant youth: acculturation, identity, and adaptation. Applied Psychology: An International Review 55 (3), 303–332.

Björkqvist, K., Lagerspetz, K.M., Kaukiainen, A., 1992. Do girls manipulate and boys fight? Developmental trends in regard to direct and indirect aggression. Aggressive Behavior 18 (2), 117–127.

Bøe, I., 1993. Folks Sosiale Landskaper. En Innføring I Sosiale Nettverk. Tano, Oslo.

Boulton, M., 1995. Patterns of bully/victim problems in mixed race groups of children. Social Development 4 (3), 277–293.

Bowlby, J., 1969. Attachment and Loss: Attachment, vol. 1. The Hogarth Press and the Institute of Psycho-Analysis, London.

Brown, B.B., 1990. Peer groups and peer cultures. In: Feldman, S.S., Elliott, G.R. (Eds.), At the Threshold: The Developing Adolescent. Harvard University Press, Cambridge, MA, pp. 171–196.

Cairns, R.B., Cairns, B.D., 1994. Lifelines and Risk: Pathways of Youth in Our Time. Cambridge University Press, New York.

Cairns, R.B., Cairns, B.D., Neckerman, H.J., Gest, S.D., Gariepy, J.-L., 1988. Social networks and aggressive behavior: peer support or peer rejection? Developmental Psychology 24 (6), 815–823.

Cairns, R.B., Leung, M.C., Buchanan, L., Cairns, B.D., 1995. Friendship and social networks in childhood and adolescence: fluidity, reliability and interrelations. Child Development 66 (5), 1330–1335.

Card, N.A., Little, T.D., 2006. Proactive and reactive aggression in childhood and adolescence: a meta-analysis of differential relations with psychosocial adjustment. International Journal of Behavioral Development 30 (5), 466–480.

Caspi, A., Elder, G.H., Bem, D.J., 1988. Moving away from the world: life-course patterns of shy children. Developmental Psychology 24, 824–831.

Chen, X., Kaspar, V., Zhang, Y., Wang, L., Zheng, S., 2004. Peer relationships among Chinese and North American boys: a cross-cultural perspective. In: Way, N., Chu, J. (Eds.), Adolescent Boys in Context. New York University Press, New York, pp. 197–218.

Chen, C., Ferruggia, S., 2002. Culture and Adolescent Development. Online readings in psychology and culture, Retrieved from: http://www.wwu.edu∼culture.

Chen, X., French, D., 2008. Children's social competence in cultural context. Annual Review of Psychology 59, 591–616.

Chen, X., Rubin, K.H., Li, Z., 1995. Social functioning and adjustment in Chinese children: a longitudinal study. Developmental Psychology 32, 352–539.

Chen, X., Chang, L., He, Y., 2003. The peer group as a context: mediating and moderating effects on the relation between academic achievement and social functioning in Chinese children. Child Development 74, 710–727.

Chen, X., Cen, G., Li, D., He, Y., 2005. Social functioning and adjustment in Chinese children: the imprint of historical time. Child Development 76, 182–195.

Chen, X., DeSouza, A., Chen, H., Wang, L., 2006. Reticent behaviour and experiences in peer interactions in Canadian and Chinese children. Developmental Psychology 42, 654–665.

Chen, X., Chung, J., Hsaio, C., 2009. Peer interactions and relationships from a cross-cultural Perspective. In: Rubin, K.H., Bukowski, W.M., Laursen, B. (Eds.), Handbook of Peer Interactions, Relationships and Groups. The Guilford Press, New York, pp. 432–451.

Cole, M., 1996. Cultural Psychology. Harvard University Press, Cambridge, MA.

Coplan, R.J., Prakash, K., O'Neil, K., Armer, M., 2004. Do you "want" to play? Distinguishing between conflicted-shyness and social disinterest in early childhood. Developmental Psychology 40, 244–258.

Craig, W., Harel, Y., 2004. Bullying, physical fighting and victimization. In: Currie, C. (Ed.), Young People's Health in Context: Health Behaviour in School-Aged Children (HBSC) Study: International Report from the 2001/2002 Survey. World Health Organization, Regional Office for Europe, Copenhagen.

Dayan, J., Doyle, A.B., Marciewicz, D., 2001. Social support networks and self-esteem of idiocentric and allocentric children and adolescents. Journal of Social and Personal Relations 18, 767–784.

Dekovic, M., Engels, R.C.M.E., Shirai, T., De Kort, G., Anker, A.L., 2002. The role of peer relation in adolescent development in two cultures: The Netherlands and Japan. Journal of Cross-Cultural Psychology 33, 577–595.

DeRosier, M.E., Kupersmidt, J.B., 1991. Costa Rican children's perception of their social networks. Developmental Psychology 27, 656–662.

Dodge, K.A., 1991. The structure and function of reactive and proactive aggression. In: Pepler, D.J., Rubin, K.H. (Eds.), The Development and Treatment of Childhood Aggression. Erlbaum, Hillsdale, NJ, pp. 201–218.

DuBois, D.L., Hirsch, B.J., 1990. School and neighborhood friendship patterns of blacks and whites in early adolescence. Child Development 61 (2), 524–536.

Dunn, J., Hughes, C., 2001. "I got some swords and you're dead!": violent fantasy, antisocial behaviour and friendship, and moral sensibility in young children. Child Development 72, 491–505.

D'Hondt, W., Vandewiele, M., 1980. Adolescents' groups in Senegal. Psychological Reports 47, 795–802.

Eslea, M., Mukhtar, K., 2000. Bullying and racism among Asian school children in Britain. Educational Research 42 (2), 207–217.

Espelage, D.L., Holt, M.K., Henkel, R.R., 2003. Examination of peer-group contextual effects on aggression during early adolescence. Child Development 74, 205–220.

Estell, D.B., Farmer, T.W., Cairns, B.D., 2007. Bullies and victims in rural African American youth: behavioral characteristics and social network placement. Aggressive Behavior 33 (2), 145–159.

Fandrem, H., 1996. The role of the subjective identity in the intercultural contact situation. Paper presented at the NIC Symposium on Intercultural Communication and National Identity. Aalborg University, Aalborg, Denmark. November 20–23.

Fandrem, H., Strohmeier, D., Roland, E., 2009. Bullying and victimization among Norwegian and immigrant adolescents in Norway: the role of pro-active and reactive aggressiveness. Journal of Early Adolescence 29 (6), 898–923.

Fandrem, H., Ertesvåg, S., Strohmeier, D., Roland, E., 2010. Bullying and affiliation: a study of peer groups in non-immigrant Norwegian and immigrant adolescents in Norway. European Journal of Developmental Psychology 7 (4), 401–418.

Farmer, T.W., Estell, D.B., Bishop, J.L., O'Neal, K.K., Cairns, B.D., 2003. Rejected bullies or popular leaders? The social relations of aggressive subtypes of rural African American early adolescents. Developmental Psychology 39 (6), 992–1004.

Farver, J.M., Howes, C., 1988. Cultural differences in social interaction: a comparison of American and Indonesian children. Journal of Cross-Cultural Psychology 19, 203–315.

French, D.C., Jansen, E.A., Riansari, M., Setiono, K., 2003. Friendships of Indonesian children: adjustment in children who differ in friendship presence and similarity between mutual friends. Social Development 12, 605–621.

French, D.C., Pidada, S., Victor, A., 2005. Friendships of Indonesian and United States youth. International Journal of Behavioural Development 29, 304–313.

French, D.C., Lee, O., Pidada, S., 2006a. Friendships of Indonesian, South Korean and United States youth: exclusivity, intimacy, enhancement of worth, and conflict. In: Chen, X., French, D., Schneider, B. (Eds.), Peer Relationships in Cultural Context. Cambridge University press, New York, pp. 379–402.

French, D.C., Bae, A., Pidada, S., Lee, O., 2006b. Friendships of Indonesian, South Korean and United States' college students. Personal Relationships 13, 69–81.

Glueck, S., Glueck, E., 1950. Unraveling Juvenile Delinquency. Commonwealth Fund, New York.

Gonzales, Y., Moreno, D.S., Schneider, B.H., 2004. Friendship expectations of early adolescents in Cuba and Canada. Journal of Cross-Cultural Psychology 35, 436–445.

Graves, N.B., Graves, T.D., 1983. The cultural context of prosocial development: an ecological model. In: Bridgeman, D.L. (Ed.), The Nature of Prosocial Development. Academic Press, San Diego, pp. 795–824.

Greenfield, P.M., Suzuki, L.K., Rothstein-Fish, C., 2006. Cultural pathways through human development. In: Renninger, K.A., Sigel, I.E. (Eds.), Handbook of Child Psychology: Vol. 4. Child Psychology in Practice. Wiley, New York, pp. 655–699.

Grotevant, H.D., Cooper, C., 1998. Individuality and connectedness in adolescent development: review and prospects for research on identity, relationships and context. In: Skoe, E.E.A., Lippe, A. L. v. d (Eds.), Personality Development in Adolescence: A Cross National and Life Span Perspective. Routledge, London.

Guthrie, D., 2006. China and Globalization: The Social, Economic and Political Transformation of Chinese Society. Routledge, New York.

Hall, G.S., 1904. Adolescence: Its Psychology and its Relation to Physiology, Anthropology, Sociology, Sex, Crime, Religion, and Education, vols. I and II. Prentice-Hall, Englewood Cliffs, NJ.

Hallinan, M.T., Williams, R.A., 1989. Interracial friendship choices in secondary schools. American Sociological Review 54, 67–78.

Hamm, J.V., 2000. Do birds of a feather flock together? the variable bases for African American, Asian American, and European American adolescents' selection of similar friends. Developmental Psychology 36 (2), 209–219.

Harris, J.R., 1995. Where is the child's environment?: a group socialization theory of development. Psychological Review 102, 458–489.

Hawker, D.S.J., Boulton, M.J., 2000. Twenty years' research on peer victimization and psychosocial maladjustment: a meta-analytic review of cross-sectional studies. Journal of Child Psychology and Psychiatry 41 (4), 441–455.

Hofstede, G.H., 2001. Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations across Nations, second ed. Sage Publications, Thousand Oaks.

Hollnsteiner, M.R., 1979. Reciprocity as a Filipino in value. In: Hollsteiner, M.R. (Ed.), Culture and the Filipino. Atteneo de Manila University, Quezon City, Phillipines, pp. 38–43.

Howes, Carollee, 1996. The earliest friendships. In: Bukowski, W.M., Newcomb, A.F., Hartup, W.W. (Eds.), The Company They Keep: Friendships in Childhood and Adolescence. Cambridge University Press, New York, pp. 66–86.

Howes, Carollee, 2009. Friendship in early childhood. In: Rubin, K.H., Bukowski, W.M., Laursen, B. (Eds.), Handbook of Peer Interactions, Relationships and Groups. The Guilford Press, New York, pp. 180–194.

Isaacs, J., Card, N.A., Hodges, E.V.E., 2001. Victimization by peers in the school context. NYS Psychologist 13 (1), 21–24.

Jay, R.R., 1969. Javanese Villagers: Social Relations in Rural Modjokuto. MIT Press, Camebridge, MA.

Jones, R.M., Hartmann, B.R., Grochowski, C.O., Glider, P., 1989. Ego identity and substance abuse: a comparison of adolescents in residential treatment with adolescents in school. Personality and Individual Differences 10 (6), 625–631.

Kandel, D.B., Davies, M., 1982. Epidemiology of depressed mood in adolescents: an empirical study. Archives of General Psychiatry 39 (10), 1205–1212.

Kanstad, M., 1994. Sosiale relasjoner på tvers av kulturelle grenser: ungdom, etnisitet og identitet i det moderne samfunnet. Ad notam Gyldendal: I samarbeid med NAVFs program for utdanningsforskning, Oslo.

Kerr, M., Lambert, W.W., Bem, D.J., 1996. Life course sequelae of childhood shyness in Sweden: comparison with the United States. Developmental Psychology 32, 1100–1105.

Kessler, R.C., Avenevoli, S., Merikangas, K.R., 2001. Mood disorders in children and adolescents: an epidemiologic perspective. Biological Psychiatry 49 (12), 1002–1014.

Laible, D.J., Carlo, G., Raffaeli, M., 2000. The different relations of parent and peer attachment to adolescent adjustment. Journal of Youth and Adolescence 29 (1), 45–59.

Larson, R.W., Richards, M.H., Moneta, G., Holmbeck, G., Duckett, E., 1996. Changes in adolescents' daily interactions with their families from ages 10 to 18: disengagement and transformation. Developmental Psychology 32 (4), 744–754.

Little, T.D., Brendgen, M., Wanner, B., Krappmann, L., 1999. Children's reciprocal perception of friendship quality in the sociocultural context of East and West Berlin. International Journal of Behavioural Development 23, 63–89.

Marcia, J.E., 2002. Adolescence, identity, and the Bernardone family. Identity 2 (3), 199–209.

McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a feather: homophily in social networks. Annual Review of Sociology 27, 415–444.

Miller, J.G., 1994. Cultural diversity in the morality of caring: Individual oriented versus duty-based interpersonal moral codes. Cross-Cultural Research 28, 3–39.

Mitchell, J. Clyde (Ed.), 1969. Social Networks in Urban Settings. University of Manchester, Manchester.

Moody, J., 2001. Race, school integration, and friendship segregation in America. American Journal of Sociology 107 (3), 679–716.

Moran, S., Smith, P., Thompson, D., Whitney, I., 1993. Ethnic differences in experiences of bullying: Asian and white children. British Journal of Educational Psychology 63 (3), 431–440.

Olthof, T., Goossens, F.A., 2008. Bullying and the need to belong: early adolescents' bullying-related behavior and the acceptance they desire and receive from particular classmates. Social Development 17 (1), 24–46.

Olweus, D., 1978. Aggression in the Schools: Bullies and Whipping Boys. Hemisphere Pub. Corp.: New York, Washington.

Olweus, D., Roland, E., 1983. Mobbing - Bakgrunn Og Tiltak. Kirke- undervisnings- og forskningsdepartementet, Oslo.

Pellegrini, A.D., Bartini, M., Brooks, F., 1999. School bullies, victims, and aggressive victims: factors relating to group affiliation and victimization in early adolescence. Journal of Educational Psychology 91 (2), 216–224.

Phinney, J.S., 1989. Stages of ethnic identity development in minority group adolescents. Journal of Early Adolescence 9 (1), 34–49.

Phinney, J.S., 1990. Ethnic identity in adolescents and adults: review of research. Psychological Bulletin 108 (3), 499–514.

Phinney, J.S., Devich-Navarro, M., 1997. Variations in bicultural identification among African American and Mexican American adolescents. Journal of Research on Adolescence 7 (1), 3–32.

Phinney, J.S., Romero, I., Nava, M., Huang, D., 2001. The Role of Language, Parents, and Peers in Ethnic Identity Among Adolescents in Immigrant Families, vol. 30. Springer, Germany.

Phinney, J.S., Berry, J.W., Vedder, P., Liebkind, K., 2006. The Acculturation experience: attitudes, identities, and behaviours of immigrant youth. In: Berry, J.W., Phinney, J.S., Sam, D.L., Vedder, P. (Eds.), Immigrant Youth in Cultural Transition: Acculturation, Identity and Adaptation across National Context. Lawrence Erlbaum, New Jersey, pp. 71–116.

Quillian, L., Campbell, M.E., 2003. Beyond black and white: the present and future of multiracial friendship segregation. American Sociological Review 68, 540–566.

Rodkin, P.C., Hodges, E.V.E., 2003. Bullies and victims in the peer ecology: four questions for psychologists and school professionals. School Psychology Review 32 (3), 384–400.

Rogoff, B., 2003. The Cultural Nature of Human Development. Oxford University Press, New York.

Roland, E., 1998. School Influences on Bullying. University of Durham, Durham.

Roland, E., 2002. Bullying, depressive symptoms and suicidal thoughts. Educational Research 44, 55–67.

Roland, E., Idsøe, T., 2001. Aggression and bullying. Aggressive Behavior 27 (6), 446–462.

Rothbaum, F., Pott, M., Azuma, H., Miyake, K., Weisz, J., 2000. The development of close relationships in Japan and the United States: paths of symbiotic harmony and generative tension. Child Development 71, 1121–1142.

Rubin, K.H., Burges, K.B., Coplan, R.J., 2002. Social withdrawal and shyness. In: Smith, P.K., Hart, C.H. (Eds.), Blackwell Handbook of Childhood Social Development. Blackwell, Malden, MA, pp. 330–3529.

Rubin, K.H., Bukowski, W., Parker, J.G., 2006. Peer interactions, relationships and groups. In: Eisenberg, N. (Ed.), Handbook of Child Psychology: Vol. 3. Social, Emotional and Personality Development. Wiley, New York, pp. 571–645.

Salmivalli, C., Nieminen, E., 2002. Proactive and reactive aggression among school bullies, victims, and bully-victims. Aggressive Behavior 28 (1), 30–44.

Salmivalli, C., Peets, K., 2008. Bullies, victims and bully-victim relationships in middle childhood and early adolescents. In: Rubin, K.H., Bukowski, W.M., Laursen, B.P. (Eds.), Handbook of Peer Interactions, Relationships and Groups. Guilford Press, New York, pp. 322–340.

Salmivalli, C., Lagerspetz, K., Björkqvist, K., Österman, K., Kaukainen, A., 1996. Bullying as a group process: Participant roles and their relations to social status within the group. Aggressive Behavior 22 (1), 1–15.

Salmivalli, C., Huttunen, A., Lagerspetz, K.M.J., 1997. Peer networks and bullying in schools. Scandinavian Journal of Psychology 38 (4), 305–312.

Sanchez, J., Fernandez, D.M., 1993. Acculturative stress among Hispanics: a bidimensional model of ethnic identity. Journal of Applied Social Psychology 23, 654–658.

Saraswathi, T.S., 1999. Culture, Socialization and Human Development: Theory, Research, and Applications in India. Sage publications, Thousand Oaks.

Schneider, B.H., Fonzi, A., Tani, F., 1997. A cross-national exploration of the stability of children's friendships and predictors of their continuation. Social Development 6, 322–333.

Schofield, J.W., Whitley, B.E., 1983. Peer nomination vs. rating scale measurement of children's peer preferences. Social Psychology Quarterly 46 (3), 242–251.

Semons, M., 1991. Ethnicity in the urban high school: a naturalistic study of student experiences. Urban Review 23 (3), 137–158.

Sharabany, R., 2006. The cultural context of children and adolescents: peer relationship and intimate friendships among Arab and Jewish children in Israel. In: Chen, X., French, D., Schneider, B. (Eds.), Peer Relationships in Cultural Context. Cambridge University Press, New York, pp. 452–478.

Sharabany, R., Wiseman, H., 1993. Close relationship in adolescence: the case of kibbutz. Journal of Youth and Adolescence 22, 671–695.

Sim, T.N., Koh, S.F., 2003. Domain conceptualization of adolescent susceptibility to peer pressure. Journal of Research on Adolescence 13, 58–80.

Smith, P.K., Cowie, H., Olafsson, R.F., Liefooghe, A.P.D., Almeida, A., Araki, H., et al., 2002. Definitions of bullying: a comparison of terms used, and age and gender differences, in a fourteen-country international comparison. Child Development 73 (4), 1119–1133.

Strohmeier, D., Spiel, C., 2003. Immigrant children in Austria: aggressive behavior and friendship patterns in multicultural school classes. Journal of Applied School Psychology 19 (2), 99–116.

Strohmeier, D., Fandrem, H., Spiel, C., 2012a. The need for peer acceptance and affiliation as underlying motive for aggressive behaviour and bullying others among immigrant youth living in Austria and Norway. Anales de psicología 28 (3), 695–704.

Strohmeier, D., Fandrem, H., Stefanek, E., Spiel, C., 2012b. Acceptance by friends as underlying function of aggressive behaviour in immigrant adolescents. Scandinavian Journal of Psychology 53, 80–88.

Sullivan, H.S., 1953. The Interpersonal Theory of Psychiatry. Norton, New York.

Tamis-LeMonda, C.S., Way, N., Hughes, D., Yoshikawa, H., Kalman, R.K., Niwa, E., 2008. Parents' goals for children: the dynamic co-existence of collectivism and individualism in cultures and individuals. Social Development 17, 183–209.

Tietjen, A., 2006. Cultural influences on peer relations: an ecological perspective. In: Chen, X., French, D., Scneider, B. (Eds.), Peer Relationships in Cultural Context. Cambridge University Press, New York, pp. 52–74.

Titzmann, P.F., Silbereisen, R.K., Mesch, G.S., 2012. Change in friendship homophily: a German Israeli comparison of adolescent immigrants. Journal of Cross-Cultural Psychology 43, 410–428.

Triandis, H.C., 1995. Individualism and Collectivism. Westview Press, Boulder, CO.

Triandis, H.C., Bontempo, R., Leung, K., Hui, C.H., 1990. A method for determining cultural, demographic, and personal constructs. Journal of Cross-Cultural Psychology 21 (3), 302–318.

Triandis, H.C., Bontempo, R., Villareal, M.J., Asai, M., Lucca, N., 1988. Individualism and collectivism: cross-cultural perspectives on self-ingroup relationships. Journal of Personality and Social Psychology 54, 323–333.

Veenstra, R., Lindenberg, S., Zijlstra, B.J.H., De Winter, A.F., Verhulst, F.C., Ormel, J., 2007. The dyadic nature of bullying and victimization: testing a dual-perspective theory. Child Development 78, 1843–1854.

Verkuyten, M., 2005. The Social Psychology of Ethnic Identity. Hove, UK: New York.

Vitaro, F., Brendgen, M., Barker, E.D., 2006. Subtypes of aggressive behaviors: a developmental perspective. International Journal of Behavioral Development 30 (1), 12–19.

Vygotsky, L.S., 1978. Interaction between learning and development (M. Lopez-Morillas, Trans.). In: Cole, M., John-Steiner, V., Scribner, S., Souberman, E. (Eds.), Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA, pp. 79–91.

Way, N., 2006. The cultural practice of close friendships among urban adolescents in the United States. In: Chen, X., French, D., Schneider, B. (Eds.), Peer Relationships in Cultural Context. Cambridge University Press, New York, pp. 403–425.

Friendship During Childhood and Cultural Variations

Kelsey B Borner, Laurie A Gayes, and Jeffrey A Hall, University of Kansas, Lawrence, KS, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Throughout the world, children seek out friendships and friendships develop naturally over time. Despite cultural differences, friendships are founded on several core features that transcend cultural boundaries, including reciprocity, mutual liking, and similarity. Theories of childhood friendship began with the work of Harry Stack Sullivan in the 1950s and have developed more broadly since the 1970s. Childhood friendships are typically evaluated based on prevalence, characteristics, stability, and quality. Friendships are beneficial for children's social and emotional development, and may buffer children against the negative life events.

Friendship during Childhood

Friendship is a type of relationship between two people based on mutual liking, wherein each member affirms the friendship, does not feel obligated to continue the relationship, and is not primarily motivated by instrumental need (Rubin et al., 2008). Distinct from popularity, which is the degree of unilateral acceptance by peers, a key feature of friendship is mutuality. Friendships are 'horizontal' relationships; they are equal in nature.

Although there is emerging evidence on cultural variation in childhood friendship, the vast majority of research focuses on children from Western cultures and the exact nature of cultural differences is poorly understood. Nonetheless, certain features of friendship appear to be universal and therefore less likely to vary by culture. Reciprocal exchange, similarity, and mutual positive regard and liking appear to be definitive features of friendship and are sometimes called deep structures of friendship (Hartup and Stevens, 1997). This article will explore theories and perspectives, characteristics and correlates, and cultural variation and similarity in childhood friendship.

Prevalence and Assessment

Although prevalence rates differ depending on whether or not children are asked to identify a friend or a 'best' friend, between 60 and 80% of children have at least one friend (Rubin et al., 2008). The most common method for assessing childhood friendship is sociometric. Typically, children are asked to nominate a specified number of classmates who they consider 'friends.' Sometimes, children are also asked to rank the closeness of their friends, differentiating between their 'very best' friend, and their second, third, fourth, and so on, best friends. The responses of the children within the classroom are compared, and reciprocal nominations are treated as evidence of friendship.

Theories of Childhood Friendship – Historical Perspectives

Childhood friendship has gained attention in several disciplines, including developmental psychology, sociology, and anthropology. Interest into the social development of children gained popularity during the 1920s and 1930s, yet a particular focus on friendship did not appear until several decades later. In the early 1950s, Harry Stack Sullivan brought the importance of friendship in child development research into focus. Previous literature in child psychology had investigated peer interactions and social development to the exclusion of friendship. Sullivan (1953) interwove social development theory with interpersonal theory and posited that social relationships satisfy interpersonal needs at different developmental stages. He concluded that childhood friendship becomes particularly important during preadolescence when the need for acceptance shifts to the need for interpersonal intimacy. According to Sullivan (1953), one of the primary outcomes of preadolescent friendship is the development of self-worth. Self-worth develops as children learn that the ideas/opinions/values they hold and share with others are important and that another person uniquely values their self-worth.

Continued focus on the importance of childhood friendship is due to the work of Willard W. Hartup and colleagues (e.g., Hartup and Sancilio, 1986), who stressed the importance of research on childhood peer relations and friendship relations. Throughout the 1970s and 1980s, there was a significant increase in research devoted to how children conceptualize friendship, behave with friends, and differ by gender and age. Hartup helped move research focus away from peer interactions toward a deeper understanding of the essence of friendship, by integrating theories of development, cognition, interpersonal relations, and attachment. Emphasizing the developmental nature of friendship over the lifespan, Hartup and Stevens (1997) differentiated between the deep structure and surface structure of friendship. While the deep structure can be thought of as the foundation or essence of friendship, the surface structure implies the observable features of friendship, including the actual interactions, activities, and exchanges between friends. The deep structure tends to be developmentally stable, whereas the surface structure of friendship is more subject to change.

The observable changes in friendship that occur as children mature must be understood in relation to children's cognitive and social development. Drawing from the work of Sullivan, Selman and colleagues (Selman, 1981) introduced a stage theory of friendship that tied closely with Jean Piaget's theories of cognitive development (*see* Piaget's Theory of

Cognitive Development). According to Selman, children move from a superficial level of friendship (e.g., based on physical and geographic factors, such as propinquity) to levels that include the deeper characteristics of friends (e.g., reciprocity, trust, and empathy).

Research confirms many of the theories of Sullivan and Selman. Young children first become friends based on propinquity or proximity (e.g., classroom, homeroom, or neighborhood). Children are more likely to be friends with those who behave as they do or are from similar racial or ethnic backgrounds, otherwise referred to as homophily. However, in a groundbreaking study, Bigelow and La Gaipa (1975) empirically demonstrated developmental changes in children's perceptions of their friendships. Validating the theories of Sullivan and Selman, Bigelow and La Gaipa (1975) found that second and third grade children expect that their friends will provide material goods, have similar demographics, and engage in common activities. As they age, children's perceptions of friendships become less superficial. As their cognitive development matures through 4th–6th grades, children incorporate the dimensions of acceptance, loyalty, commitment, and genuineness when choosing friends. By preadolescence (e.g., 7th grade), children consider intimacy and shared values important to their friendships, and this also marks the transition into adolescence.

Youniss (1980) also expanded on Sullivan's theory by incorporating Piaget. Youniss criticized traditional models of child development for relying on the parent–child relationship as the sole vehicle for children's social and cognitive development. Termed the Sullivan-Piaget thesis, Youniss (1980) called attention to the importance of peer relations and declared that children's relations with other children and adults serve equally important, but distinct, functions in social development. Whereas the parent–child relationship requires the child to learn and accept rules and values from or through the adult, friendships allow children to socially develop with other children in ways that would be impossible to develop alone or with adults (e.g., interpersonal sensitivity, reciprocity, mutual understanding, cooperation, and intimacy). Similarly, Shantz (1983) outlined three developmental features of childhood friendship: (1) a shift from strict behavioral (e.g., activity-based) interactions to emotional relationships, (2) progression from egocentric to mutually beneficial relationships, and (3) evolution from unstable but conflict-free relationships to those that are stable and able to withstand conflict.

Psychosocial Associations of Friendship and the Lack Thereof

Friendship development during childhood is necessary for a child's successful social, cognitive, and emotional development. Children develop core social skills through the experience of companionship, intimacy, and conflict resolution with friends. Such skills may be enhanced by a qualitatively stronger relationship with a best friend. Although the literature is mixed, friendships may also contribute to stronger emotional development by providing a context for learning emotion regulation, emotional intelligence (i.e., understanding the emotions of others), and appropriate emotional expression. Cognitively, friendship enhances the child's development of verbal and nonverbal sharing, problem solving, and advanced communication skills.

The presence and stability of a friendship have social and emotional implications for children. Children who are friendless are more likely to be lonely, as are children who have lower quality friendships. Furthermore, children who have no, few, or low-quality friendships are more likely to develop depression if they experience significant feelings of loneliness. By comparison, children who have stable best friends are more likely to exhibit prosocial behaviors (*see* Prosocial Behavior during Childhood and Cultural Variations), use more effective conflict management strategies, and be more popular with peers. For example, children with stable best friendships are more likely to use effective problem-solving strategies (e.g., talking about the problem) when presented with stressful social situations. More stable best friendships also predict lower levels of loneliness, bullying, and victimization. Although it is likely that children with the requisite social skills to maintain a best friendship and resolve conflict are also more prosocial and less lonely in general, there is some evidence that having a best friendship offers some protection against victimization. Research shows that children who lose a best friend over the course of a year are more likely to become victimized, but children who gain a best friend over a year are less likely to become victimized.

Although friendship is important for all children, friendship may be especially important for children who experience both normal stress and family- or peer-related stress. For most children, transitioning to a new school can be a stressful experience and friends are emotionally helpful during those times. In fact, successful friendships are as important as home and environmental factors during the transition to school. Other stressors may come in the form of parental changes (e.g., divorce, rejection from parents, harsh parental discipline) or severe rejection from peers. In both scenarios, friendship appears to serve as a buffer against the negative consequences that otherwise might accompany stressful experiences.

Characteristics of Friends and Quality of Friendships

Summarizing decades of research on the characteristics of childhood friendships through meta-analysis, Newcomb and Bagwell (1995) completed the most comprehensive investigation of the components of childhood friendships by comparing friends with nonfriends (i.e., acquaintances, strangers, disliked peers). They concluded that friends spend more time together and engage in more conversation than nonfriends, mostly due to their similar play behaviors and shared activities. Friends and acquaintances look similar in many aspects, particularly in the areas of similarity, equality, and dominance. However, friendships stand out from acquaintanceships because of the unique characteristics of mutual liking, closeness, and loyalty. Although friendships are regarded as generally positive relationships, they are not without disagreement and conflict. While some disagreement is normative, friends, unlike nonfriends, make greater effort to resolve conflicts.

Research interest in the quality of friendship increased during the 1990s. Berndt was one of the first researchers to

identify friendship quality as an individual construct. The construct implies an evaluation and assumes that some friendships are better or worse than others. Friendship quality is typically evaluated as the ratio of positive to negative features. As expected, friendships with more positive features (i.e., closeness, companionship) and fewer negative features (i.e., conflict) are considered higher in quality. Notably, early researchers focused solely on the positive features of friendship. Not until researchers began asking children about the negative features of friendship did they discover that they did indeed exist. While friendships contain positive and negative features, there does not appear to be a strong negative relationship between negative and positive features. In fact, childhood friendships that are high on positive features may be high or low on negative features, such that the two constructs are not mutually exclusive, and often friendships with many positive features also may have many negative features.

Friendship Stability

The stability of friendship refers to the maintenance of the friendship over time. Because friendships are voluntary relationships of choice, they are more likely to end than permanent relationships, like relationships with siblings or cousins. Until recently, most studies assumed that childhood friendships were static and researchers only measured friendships at a singular point in time. As children age, friendship networks expand and evolve to accommodate changing cognitive, social, and emotional development (Poulin and Chan, 2010). This means that some children end a friendship after a few months while others can last over years. Some friendships are easy to replace, while others are much more difficult to replace, if at all. Developmentally, early adolescence appears to represent a unique period when friendship instability peaks, which then stabilizes again during the transition to adolescence. This is likely due to the transition to high school and significant and ongoing biological and emotional development. (More information on adolescent friendship development may be found here (*see* Friendship during Adolescence and Cultural Variations).)

According to Poulin and Chan (2010), several factors affect the stability of a child's friendships. Friendships are more likely to be stable when children form friendships with others who are similar to themselves. Better quality friendships are more stable because a pattern of positive interactions promotes stability. Friendships with more negative features (e.g., higher conflict, rivalry, poor conflict resolution) are less stable. The context of a friendship also impacts stability, such that friendships that overlap across several contexts (e.g., school, extracurricular activities, neighborhood) are more stable over time. Changes in a child's environment disrupt the stability of friendships. For example, when a child transitions to a new school, the stability of existing friendships suffers.

Characteristics Negatively Affecting Childhood Friendship

Some personal characteristics of children inhibit their ability to make friends, keep friends, or maintain high-quality friendships. Children with internalizing problems such as depression are less likely to maintain stable friendships, especially if they exhibit high levels of loneliness. Children with behavior problems, such as aggression, impulsivity, and hyperactivity, also have friendship difficulties. For example, children with attention-deficit/hyperactivity disorder are at high risk for friendship problems due to poor problem-solving strategies (i.e., impulsive, less friendly, and more assertive strategies). Similarly, children with oppositional defiant disorder exhibit problem behaviors among peers, behaving more aggressively when provoked. Finally, children with autism spectrum disorders are at increased risk for social deficits and may experience problems developing and maintaining friendships due to difficulties in approaching same-age peers, engaging in similar conversations, engaging in social reciprocity, and sharing interests.

Cultural Variations

Culture has been defined as a set of values, norms, and practices that shape the individual's behavior. Culture influences children's friendships by structuring the parameters of interactions between peers throughout development. Childhood friendships occur within a dynamic environment where children both learn behavioral expectations dictated by their culture and practice playing out adult cultural roles. Many cultural factors shape a child's life and the nature of his/her friendships, including school structure and extracurricular activities, economic opportunities and socioeconomic status, as well as ideological cultural values, such as collectivism and individualism. These factors strongly influence the nature of friendship and the relative importance of peer relationships in a child's life. While some features of friendship appear to be stable across child development and across cultures, other features differ. Three primary dimensions of friendship (i.e., time spent with friends, reciprocity, and similarity) will be discussed as they relate to similarities and differences across cultures.

Cultural Similarity

Despite the wide range of cultural dimensions that shape childhood friendships, many features of childhood friendship are consistent across the globe (French et al., 2003). Three such stable features of friendships are proximity, reciprocity, and similarity. These qualities influence childhood friendship in every culture, but how each culture emphasizes these characteristics differs.

Friendship typically develops because children are proximal to one another (e.g., same community, school, classroom) and this trend is generally consistent across the globe. However, the type of potential friends who share the same space as a child is influenced by cultural variations. For example, how a child's time is divided between family and peers influences the relative importance that the child places on those relationships. For example, in Kenya most children spend the bulk of their time on family lands, so large families become the primary social sphere for the child (i.e., they are most proximal to the child). Other social opportunities are limited, as parents teach their children to be wary of nonfamily trespassers. In North America, by contrast, children are typically encouraged to be outgoing and social

with nonfamily peers and this behavior is associated with being well-liked and behaving in a culturally acceptable manner. Cultures that place importance on a child spending time with peers and learning to be social with peers send the message to the child that relationships with peers are an important part of childhood. Whatever the cultural context, whether the child spends time with peers or family, close friendships form with the individuals with whom the child spends the most time.

Symmetrical reciprocity has been conceptualized as one of the deep structures of friendship. In many cultures, reciprocity is seen in children sharing with one another and behaving cooperatively. There appears to be a great deal of cross-cultural similarity in this aspect of friendship. For example, both children from Columbia and children from the United States shared more candies with friends than with nonfriends. In every culture, friends have mutual responsibilities to one another, but the exact nature of those responsibilities differs according to the values of the society.

In all cultures, children choose friends who are similar, and friends become more like each other as they spend more time together. Across the globe, childhood friends resemble one another in terms of physical characteristics, such as age, gender, and race, and personal characteristics, such as interests, preferences, and behaviors. Youth who are friends even display similar levels of internalizing symptoms, academic achievement, and aggression (French et al., 2003). Across cultures, children seem to have roughly the same number of friendships, suggesting that children are fairly consistent in the number of peers with whom they can build relationships. This similarity in the number of friends provides some evidence for developmental perspectives on children that emphasize the importance of emotional and psychological development on children's friendship. Furthermore, there appears to be consistency worldwide in the features of children without strong peer relationships. Children without friends seem to consistently display heightened aggression, social withdrawal, and academic struggles (French et al., 2003). Although there are important similarities between cultures in childhood friendship, several factors that vary by culture are examined below.

Classrooms and Extracurricular Activities

For many children across the world, schools are the hub for peer socialization and friendship formation. Children who live in industrialized parts of the world are most likely to go to school and have their friendships shaped by the school structure. The age-segregated style of classrooms and the often age- and gender-segregated style of extracurricular activities strongly influences childhood friendships. Because children tend to form friendships with those with whom they spend the most time, they tend to make friendships with others who are enrolled in the same class, participate in similar activities, and engage in similar style of play. The degree that classrooms and activities are segregated by gender may influence the gender differentiation of friendship. However, there is evidence that youth tend to choose best friends of the same gender as a general rule, no matter the level of industrialization, of the culture, and participation in school of the children. Indeed, a recent meta-analysis on friendship expectations demonstrated that gender differences remained stable across cultures, despite cultural variation in friendship expectations. This suggests that some gender differences in childhood friendship may exist across cultures.

Time with Peers versus Family

Not all children spend significant periods of time at school. Due to economic demands, some children need to work to support the family. In these circumstances, children either become employed or take care of tasks at home so that the parents can work. Compared to children in cultures where school for children is the norm, children who work at home or elsewhere do not experience strict age segregation or separation from immediate family. For example, children who do not attend school often have responsibilities such as caring for younger siblings or working with older siblings in the home, thus primarily spending time being socialized by siblings rather than by nonkin peers. In contrast, in more industrialized countries where both parents tend to work long hours outside the home, such as Taiwan, children tend to spend most of the day at daycare, school, and after-school activities, which may carry on into the evening hours. These activities cause the child to spend considerably more time with peers than with family members, heightening the importance of peer relationships in the child's life.

Children Who Work

Children who must work outside the home to support their families or themselves relate to peers somewhat differently than children whose peer relationships are purely social, because they interact with peers in a work context. The nature of the work itself shapes the setting wherein friends interact, the types of interactions that take place, and the qualities of a friend that are valued. In the Himalayas, friendship responsibilities for leaf-collecting girls include helping each other with physical labor, making work fun for the other, and looking out for the other girl's reputation (Dyson, 2010). The demands of life on the street in Ghana demand that friends are sources of more tangible reciprocity, such as sharing food and resources when they are available (Mizen and Ofosu-Kusi, 2010).

Providing an example of one cultural context where children spend their days assisting the family and the community instead of going to school, Dyson (2010) describes young girls' friendship and work behavior in a small society in the rural Himalayan region. Girls in this remote village spend their days gathering large bushels of leaves to serve as bedding for the livestock. The girls must climb high into the mountains, collect leaves, and then fashion backpacks out of branches to transport the leaves back down the mountain. This job is extremely important for the survival of the community and a girl who can gather and successfully carry a large quantity of leaves is a highly valued member of society. Friends support each other and help each other look good in the eyes of the community. If your friend sees you struggling or believes that you have not yet met the expectation set for you by the community for leaf collecting, she will come to your

aid. In the words of one of the girls, a friend is someone who "doesn't leave me alone in the forest. We collect leaves and wood for each other" (Dyson, 2010).

A second example examines friendships between street children in Accra, Ghana (Mizen and Ofosu-Kusi, 2010). For these children, the cooperative dimensions of friendship are most salient as friends are valued in part for their ability to provide one another with tangible instrumental aid in times of need. Friends are necessary for survival, and fortunately nearly all street-children were able to identify a friend. Being without a friend can mean that you go to sleep on an empty stomach or are left to deal with trouble by yourself, significantly weakening your ability to survive in such a harsh environment. When children were asked to define friendship, there was near consensus that being friends meant helping each other. Friends provide for one another when they are unable to manage on their own and friends can rely on each other when they are in need.

In this way, the cultural context of friendship shapes the salience of the qualities of a friendship. When children are faced with the need to be earners to survive, friends become valued for their helpfulness and hard work, rather than their style of play or interests. In this circumstance, the practical value of friendship (i.e., that we can accomplish more together than alone) becomes essential, but the importance of mutual regard and reciprocity remain. For all children, the nature of friendship is shaped in the context of culture, and friendship is also the avenue through which the values of a culture become reinforced and perpetuated over generations.

Cultural Ideology

The ideological values of a culture add another important layer of cultural influence on childhood friendship. The dimension of collectivism versus individualism is one of the defining constructs of culture. According to Hofstede (1980), traditional individualistic societies place priority on the needs of the individual, whereas traditional collectivistic societies tend to place emphasis on the community as a whole and on the needs of a group. This cultural ideology shapes children's experiences through the dialectal tensions of independence and individualism versus harmony and duty. Not surprisingly, childhood friendships mirror the values of this dynamic. These values are evident in friendship factors such as the degree to which the family is seen as the primary source of companionship and support for a child as compared to peers; and acceptable ways for friends to behave toward one another.

In Costa Rica and South Korea, which are more collectivistic nations, family is considered the most important social realm for a child (Oh, 2008). Support and intimacy are nurtured in the parent–child relationship and sought to a lesser degree from peer relationships. The child from an individualistic culture, as exemplified by traditional Western values, tends to perceive higher social reward (or at least equal social reward) from peers as compared with the family, (Oh, 2008). Children from the United States report receiving significant support, companionship, and intimacy from best friends as compared with parents. Similarly, children in individualistic societies tend to display more independence and less influence from parents in choosing friends as compared with children from collectivistic cultures. Another study compared the reasoning of children from Iceland, traditionally a more individualistic nation, and China, a more traditionally collectivistic nation, when faced with an interpersonal moral dilemma. Findings suggested that Icelandic children tended to make moral decisions based more on self-interest and contractual concerns, as compared to Chinese participants who valued altruistic and relationship-related concerns more highly.

Conclusions

As interest into childhood friendship has advanced over the past several decades, research points to the unique contributions that friendships provide to a child's emotional and social development. The specific components, contributors, and consequences of friendship are multifaceted and fluid and create challenges for researchers to document the theories proposed by early theorists. From early theories, decades of research confirm that childhood friendships should be interpreted from a developmental perspective. From childhood to adolescence, children's friendships mature from a superficial level (e.g., friends chosen based on shared play) to a deeper level (e.g., loyalty and intimacy). Having a best friend is important for a child and the particular characteristics of friendships, such as quality and stability, also have implications for healthy development. Unlike adolescent or adult relationships, childhood friendships are unstable and children may change friends several times in a given year. Neighborhoods, schools, and personal characteristics of the child impact his or her ability to make and keep friends, and some children are better equipped than others to make friends. Despite the intricacies of friendships and the differences between individuals, friendships are beneficial for all children and can be especially important for buffering against the effects of negative life events.

The similarities of friendship in youth across the globe are striking. Children worldwide seek out peers with whom they can become close. Despite the tremendous range of life circumstances in childhood, such as economic demands on the family and the child, the settings where youth spend time, and role of the family in the child's life, childhood friendships are remarkable in their universal significance across cultures and the similarity of their features. Worldwide, children seek peers who resemble themselves, provide mutual support, and spend time together. The differences in friendship across cultures seem to lie in the circumstances of where friendships form, the types of support that are expected from peers, and the qualities of the children and of the relationships that are encouraged. In this way, the children learn the proper societal expectations for relationships with others and practice embodying these characteristics in preparation for becoming successful and social adult members of their community.

See also: Friendship During Adolescence and Cultural Variations; Piaget's Theory of Cognitive Development; Prosocial Behavior During Infancy and Early Childhood: Developmental Patterns and Cultural Variations.

Bibliography

Bagwell, C., Schmidt, M., 2011. Friendships in Childhood and Adolescence. Guildford Press, New York.

Bigelow, B.J., La Gaipa, J., 1975. Children's written descriptions of friendship: a multidimensional analysis. Developmental Psychology 11, 857–858.

Bukowski, W.M., Newcomb, A.F., Hartup, W.W., 1996. The Company They Keep: Friendship in Childhood and Adolescence. Cambridge University Press, New York.

Cosaro, W.A., 2003. We're Friends, Right?: Inside Kids' Culture. Joseph Henry Press, Washington, DC.

Dyson, J., 2010. Friendship in practice: girls' work in the Indian Himalayas. American Ethnologist 37, 482–498.

French, D.C., Jansen, E.A., Riansari, M., Setiono, K., 2003. Friendships of Indonesian children: adjustment of children who differ in friendship presence and similarity between mutual friends. Social Development 12, 605–621.

Hartup, W.W., 1983. Peer relations. In: Hetherington, E.M. (Ed.), Handbook of Child Psychology, Socialization, Personality, and Social Development, vol. 4. Wiley, New York, pp. 103–198.

Hartup, W.W., Sancilio, M.F., 1986. Children's friendships. In: Schopler, E., Mesibov, G.B. (Eds.), Social Behavior in Autism. Plenum Press, New York, pp. 61–79.

Hartup, W.W., Stevens, N., 1997. Friendships and adaptation in the life course. Psychological Bulletin 121, 355–370.

Hofstede, G., 1980. Culture's Consequences: International Differences in Work-Related Values. Sage, Beverly Hills, CA.

Howes, C., 1983. Patterns of friendship. Child Development 54, 1041–1053.

Mizen, P., Ofosu-Kusi, Y., 2010. Asking, giving, receiving: friendship as survival strategy among Accra's street children. Childhood 17, 441–454.

Newcomb, A.F., Bagwell, C.L., 1995. Children's friendship relations: a meta-analytic review. Psychological Bulletin 117, 306–347.

Oh, W., 2008. Children's perceptions of their relationships with mothers, fathers, and friends: a cross-cultural study (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses.

Poulin, F., Chan, A., 2010. Friendship stability and change in childhood and adolescence. Developmental Review 30, 257–272.

Rubin, K., Fredstrom, B., Bowker, J., 2008. Future directions in friendship in childhood and early adolescence. Social Development 17, 1085–1096.

Selman, R.L., 1981. The child as a friendship philosopher. In: Asher, S.R., Gottman, J.M. (Eds.), The Development of Friendship. Cambridge University Press, New York, pp. 242–272.

Shantz, C.U., 1983. Social cognition. In: Mussen, P.H. (Ed.), Carmichael's Manual of Child Psychology. Wiley, New York, pp. 495–555.

Sullivan, H.S., 1953. The Interpersonal Theory of Psychiatry. Norton, New York.

Youniss, J., 1980. Parents and Peers in Social Development: A Sullivan-Piaget Perspective. University of Chicago Press, Chicago, IL.

Friendship During Infancy and Early Childhood and Cultural Variations

Kenneth H Rubin, Maureen A Wimsatt, Sara Heverly-Fitt, and Matthew G Barstead, University of Maryland, College Park, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The functions and prevalence of friendship in infancy and early childhood are described. Methodologies used to study early friendship as well as limitations of extant research are discussed. Also, several studies that use etic and emic approaches to study cultural variations in friendship during early stages of development are noted.

Introduction

In the literature on children's peer relationships, friendship has garnered more attention than all others. The issue of what constitutes a friendship, especially during the earliest years of childhood, has spurred debate that is well beyond the scope of this article. However, we note some relevant aspects of this debate herein.

First, there is widespread agreement that friendship is a *reciprocal* relationship that must be affirmed or recognized by both parties (Bagwell and Schmidt, 2011; Rubin et al., 2006). Reciprocity is the factor that distinguishes between friendship and the nonreciprocal attraction of only one partner to another. A second defining characteristic of friendship in early childhood (and beyond) is mutual *affection* (Rubin et al., 2006). The interdependence of the two partners derives primarily from socioemotional rather than instrumental motives. Third, friendships are *voluntary*, not obligatory or prescribed. In some cultures, children may be assigned their 'friends,' sometimes even at birth (Krappman, 1996). Although these relationships may take on some of the characteristics and goals of voluntary relationships, most scholars would agree that by not meeting the criteria of reciprocal choice and mutual affection, their involuntary nature argues against confusing them with friendship.

The mutuality of friendship and the autonomy children have in choosing their friends has long intrigued developmental scientists. For example, Sullivan (1953) believed that the understanding of mutual respect, equality, and reciprocity developed from social experiences of chumships. Sullivan's theory has proved influential in the contemporary study of children's friendships as well as in the understanding of the negative psychological consequences of not having close dyadic relationships with friends (e.g., Wojslawowicz et al., 2006).

The aims of this article are twofold. First, we trace the early development of friendship from infancy through early childhood. Second, we discuss the importance of cultural influences on the processes and outcomes associated with friendships in early childhood. Before engaging in either pursuit, it is important to highlight the developmental significance of friendship and to examine how it is that friendship is assessed in the earliest years of childhood.

The Developmental Significance of Friendship

Human beings are innately social creatures, driven by a fundamental need to belong and form lasting attachments. This drive first manifests in the attachment relationship formed between an infant and his or her primary caregiver (Bowlby, 1969). As children age, they are able to interact socially with an increasing array of individuals, including peers. Given the opportunity, almost all children form friendships. Even a large majority of socially withdrawn children have at least one mutual friend (Rubin et al., 2009). The ubiquitous nature of friendship suggests that it may be evolutionarily adaptive. Friends can provide both instrumental and emotional support (Rubin et al., 2006), and indeed friendship analogs have been found in other social species (see Seyfarth and Cheyney, 2012; for a review).

One significant aspect of friendship is *prevalence*; that is, how many friends does a given child have? Another important aspect of friendship is involvement; does the child have at least one mutual friendship or do none exist? Generally, children who have friendships fare better than those who do not. For example, as early as the period of transition into elementary school, children who are lacking in friendship are at greater risk for the development of adolescent loneliness and depression than those who have a friend (Pedersen et al., 2007). There are at least two reasons that having a friend is associated with more positive outcomes. First, as noted above friends can be sources of support and comfort, thereby buffering the effects of stressors. Second, as interaction partners, friends provide opportunities to improve social and emotional competence and to avoid psychological maladaptation (Rubin et al., 2006).

The Assessment of Friendship in Infancy and Early Childhood

From the time children enter preschool (age of 3 years), researchers can rely on several techniques to identify friendship dyads. For example, parents and teachers can be asked with whom it is that children play with most often. These reports can be especially useful in confirming friendship with children who do not attend the same preschool or child care center, but reside in the same neighborhood. However, a noted caveat to using parent-report is that it may be more indicative of who the parent prefers their child befriend and spend time with (Howes, 2009).

Ofttimes, preschoolers are provided with pictures of all their classmates and are asked to identify the peers they like the most or with whom they most often play. Typically, researchers may use the child's top three choices of preferred playmates or friends (Rubin et al., 2006). When using such sociometric procedures, it is necessary to assess *mutual preference*; thus

a friendship is one in which two children choose each other as the most preferred playmate.

Of course, during infancy and toddlerhood, when communication skills are rather limited, sociometric procedures cannot be used. Instead, in addition to parent and teacher reports, researchers have come to rely on observational procedures to identify preferred dyadic relationships. Observational methods are particularly useful because they allow researchers to watch for more nuanced behaviors that may not be measurable with standard questionnaires or interview techniques. Thus, observations have been used to identify peers who frequently seek each other out, spend significant time together (in close proximity), and share positive affect at levels greater than chance probability (Howes, 2009).

Friendships in Infancy (0–2 years)

Infants do not have friendships as traditionally defined and measured later in childhood. In many cases, infants lack the opportunities to interact regularly with peers, and even when they do, their limited communication skills hamper friendship formation. Still, some researchers contend that infants have the capacity for rudimentary friendships. By 6 months, infants begin to more consistently orient toward peers and engage in longer interactions, often centered on toys and other simulating objects in the environment (Hay et al., 1983). During the second year, infants increasingly play with a partner rather than alone, and they appear to prefer peers as interaction partners over adults (Eckerman et al., 1975). Although infants are less likely to demonstrate positive affect or engage in complementary play compared to toddlers and preschoolers, they nevertheless show clear preferences for specific social partners.

For example, by observing infants and toddlers during play, Howes (2009) has found evidence that mutual friendships can be identified. In observational studies of very young children, behavioral indices of mutual affection, companionship (assessed by the frequency of observed proximity to specific peers), complementary play (e.g., rolling a ball back and forth), and relative frequency of initiations directed to and received from specific peers are used as markers of friendship.

Significantly, although the friendships of infants and toddlers do not involve such features as intimate disclosure, validation, and instrumental aid, it is nevertheless the case that infants and toddlers who have opportunities to regularly interact with peers are more socially competent at the age of 4 years than their counterparts who have few such experiences (Howes, 1991).

Summary

During the infant and toddler periods, friendships, as construed by the constructs noted above (mutuality, reciprocity, positive affect), do appear to exist; however, it is doubtful that they carry the same strength of psychological meaning as the friendships of older children. Nevertheless, these early relationships may lay the groundwork for the establishment and maintenance of friendships throughout the childhood years (Dunn, 2004).

Friendships in Early Childhood (2–6 Years)

The function and prevalence of friendship in early childhood is distinctly different from the social interactions of infants. During early childhood, friendships are primarily rooted in the principle of *homophily*. Homophily refers to the notion that individuals are attracted to others who are similar to them in some noticeable regard; in effect, homophily is the "birds of a feather flock together" principle. For example, similarities in age, race, and sex draw young children together (Aboud and Mendelson, 1996; Howes, 2009). Homophily also extends to behavioral characteristics; preschoolers are drawn to others who engage in similar play behaviors and prefer similar activities and materials. Furthermore, it has been shown that aggressive children tend to befriend other aggressive children, and withdrawn children tend to befriend other withdrawn children (Rubin et al., 2006).

The Understanding of Friendship

Unlike infants, young children are able to communicate what they know about the construct of friendship. Young children tend to perceive friendship as comprising the spending of significant time with a preferred other who engages in activities that are pleasant, and who have access to materials that are fun to play with (Bigelow, 1977). Friends help each other, have similar interests, share preferred possessions, cooperate on tasks, and play together (Dunn, 2004). Gottman (1986) audio recorded children aged 3–6 as they played with a best friend. When transcribing the conversations, it was confirmed that the friendships were strongly based on mutual likes and dislikes. The recordings also provided evidence of early instances of self-disclosure. Thus, as early as the preschool years, a friend represented someone who could be trusted with knowledge of one's personal beliefs, desires, or fears.

Play and Friendship

Play is an essential component of friendships in early childhood. During this period of development, most friendship activities consist of play. However, play is not a frivolous activity to be overlooked or dismissed. By 24 months, children are able to engage in *intersubjectivity* during play – the ability to share meaning during pretense (Trevarthen, 1979). That is, children are able to take on imaginary roles, while doing so in a rule-governed context. Sociodramatic play becomes more frequent during the preschool years, and is purposeful in that it allows children to challenge and take on different social roles and identities (Dunn, 2004). Children who have friends are able to more seamlessly enter playgroups, and have more complex social pretend play. The increasing frequency of play aids in the maturation of such social cognitive abilities as theory of mind (perspective-taking or being able to understand the thoughts, feelings, and visual viewpoints of others) and hypothetical-reflective reasoning (e.g., social information processing), which provide a basis for the development of social and emotional competence (Rubin et al., 2011).

Preschoolers' social interactions with friends are distinct from their behavior with other children who are familiar but not friends. For example, children as young as 3.5 years

direct more social overtures, engage in more social interactions (including sociodramatic play), and interact in more cooperative and prosocial ways with friends than nonfriends (e.g., Dunn, 2004).

Conflict and Friendship

Conflict occurs more frequently among preschool friends than among familiar or unfamiliar nonfriend peers. For example, researchers have noted that preschool friends demonstrate more quarreling and more active (assaults and threats) and reactive (refusals and resistance) hostility than nonfriends. These seemingly curious findings are best understood by recognizing that friends spend much more time actually interacting with each other than do nonfriends (Rubin et al., 2006). However, it is also the case that children are more likely to negotiate and compromise until a solution is reached with friends, compared to nonfriends (Howes, 2009). Furthermore, children are more likely to reach equitable resolutions and have mutual agreements with friends than nonfriends. In conflict outcomes, friends are more likely to have equal resolutions, relative to win or lose occurrences. Also, following conflict resolution, friends are more likely than neutral associates to stay in physical proximity and continue to engage in interaction (Hartup, 1996). These increased instances of compromise, negotiation, conflict resolution, and positive affect in early friendships have been proposed to influence the development of self-regulation and affect (Bagwell and Schmidt, 2011).

The Prevalence of Friendship in Early Childhood

Approximately 75% of preschoolers have reciprocally nominated best friendships (Bagwell and Schmidt, 2011). However, not all friendships in early childhood are equally stable and not all young children have a best friend (Ladd et al., 1996). Those friendships that involve higher levels of positive friendship qualities (e.g., validation) and lower levels of negative friendship qualities (e.g., low conflict) are most likely to be stable.

Importantly, during early childhood, the ability to make friends, friendship quality, and stability of young children's friendships are associated with, and predicted by, social cognitive and emotional maturity. For example, the abilities to perspective-take are associated with friendship formation, friendship maintenance, and friendship quality (Dunn, 2004). Furthermore, the young child's ability to regulate emotions is associated with, and predictive of, both the number of mutual friends and friendship quality (Walden et al., 1999).

Adjustment and Friendship

Friendships in early childhood have been found to promote positive adjustment. For example, children who start the academic year with a friend have more positive feelings regarding school. Ladd and colleagues (e.g., Ladd et al., 1996) found that children with many friends at the time of elementary school entry developed more favorable attitudes toward school in the early months than children with fewer friends. Those who maintained their friendships liked school better as the year went by. Findings also revealed positive associations between children's perceptions of best friendship quality in kindergarten and later indices of scholastic adjustment (school-related affect, perceptions, involvement, and performance) in grade school. It is clear that the benefits gained from friendships in early childhood are extensive.

Research Gaps: Early Friendships

First, and perhaps most notably, extant research has generally failed to examine the development or maintenance of early close friendships. Although peer relationships are vital to early social development, close friendships can provide children with a sense of companionship and security that facilitates enhanced social competence and emotional growth, particularly during early childhood (Howes, 2009). Friendships also buffer children from loneliness and social dissatisfaction and provide a social context to help children develop loyalty and commitment toward other social partners. Early friendship *quality*, often predicts children's long-term psychosocial adjustment (Rubin et al., 2006). Thus, in the future, researchers should study friend-specific relationships in addition to peer acceptance, rejection, and children's social behavioral proclivities.

A second limitation of extant research involves the failure to assess reported and observed friendship *quality* among *mutually* nominated friends or best friends. Assessing perceived friendship quality from the perspective of a single member of a putative friendship dyad poses a number of methodological limitations. For one, relying on the perspective of a single member of the dyad provides a one-dimensional view of the relationship. It could be the case that children may inadvertently overestimate or underestimate the quality of their friendships. Thus, it would behoove researchers who are interested in examining the significance of friendship quality to rely on information provided by both members of the reciprocally nominated friendship pair. In addition, researchers should increasingly rely on observations of friendship quality during naturalistic interaction, especially during the early years of development.

Cultural Variations in Early Friendships

Defining Culture

Culture may be defined as the set of attitudes, behaviors, beliefs, and values shared by a group of people. Members of a cultural group communicate shared perspectives from one generation to the next, with younger children learning about cultural norms from older members of a society (Matsumoto, 1997). Since most individuals in a cultural group share beliefs and customs, scholars view the development of 'typical' cultural behaviors as adaptive to children who are fitting into a larger group with established norms. In comparison, children who do not develop the normative identity of their cultural group are thought to demonstrate maladaptive behaviors that may hinder the child's acceptance by other members of the culture.

Cross-Cultural Research Approaches

Culture is primarily studied using two research approaches. The first, an 'etic approach,' assumes that behavioral constructs

(i.e., concepts, methods, measures) studied in one culture have significance in all other cultures. Alternatively, an 'emic approach' describes the study of cultural norms that are specific to one group of people or within one culture. Studies using an etic approach require researchers to develop, test, and analyze constructs in the same manner across groups. As a result, findings may overlook specific cultural perspectives or culturally specific behavioral nuances. For example, researchers using an etic approach may assume that friendship 'looks' the same in all cultural groups, and would likely assess friendship processes by asking the same research questions and methodologies in all groups. Scholars using an emic approach assume that their cultures may not share the same attitudes, behaviors, beliefs, or values about friendship and may tailor empirical questions or methodologies to specific cultural groups. One example of emic research could involve the comparison of friendship processes and provisions in Western (or relatively independent, individualistic) cultures versus more interdependent, collectivist Eastern or Southern cultures. In this case, an emic approach would entail the use of culturally adapted instruments and observational coding schemes to ensure the collection and analysis of culture-specific information about children's friendships.

Using both etic and emic approaches, scholars can help distinguish those research constructs that are universally meaningful and those that are culture specific. It is likely that aspects of many constructs, such as friendship, may support the etic and emic perspectives. That is, some elements of friendship may be viewed as universally held while others may be unique to particular cultural groups.

Friendship versus Familial or Community Relationships

Young children from Western cultures in North America and Europe are often encouraged to develop and nurture friendships in addition to caring for and respecting family members; indeed, children from Eastern and Southern countries such as South Korea and Cuba have been found to form close, dyadic friendships that last into adulthood (Chen and French, 2008).

In comparison, some cultures place a stronger emphasis on early family or community relationships rather than on extrafamilial friendships. For example, young children from Yucatec Mayan communities rarely form close friendships with children outside of the immediate and extended family. In fact, Mayan 'friendships' are typically close relationships between children and their same-age siblings or cousins (Gaskins, 2008).

Community-level relationships are a priority among children in Israel. Rather than spending time with close friends, children from kibbutz communities are immersed in group activities and community events at a very early age (Sharaban, 2006). Due to cultural emphases on group relationships in Indonesia, Javanese children also have multiple important relationships with nonfriend members of the community rather than dyadic friendships (French et al., 2005).

Shared Activities

There is cultural variation in the types of activities in which young children participate. In Western contexts, play is a primary focus of peer interactions and friendship during early childhood. Young children from Western cultures often use toys, games, or imaginative role-play activities to guide interactions with friends. In some cultures, however, children's play is not viewed as important as observing others or learning about work responsibilities (Gosso and Carvalho, 2013). These cultures are usually ones in which dyadic friendship is not valued above relationships within the family unit or community. For example, young Yucatec Mayan children who typically do not form relationships outside of immediate and extended family networks spend most time engaged in maintenance activities (i.e., eating, sleeping, grooming) and the observation of activities occurring within the family compound (Gaskins, 2008). In Mayan culture, childhood play is a tertiary activity since it is believed to compete with adult work goals. When play does occur, young Mayan children rarely participate in symbolic or 'pretend' play with same-age children, instead engaging in play through the manipulation of objects or production of large motor movements.

The Giriama of Kenya believe it is important to teach children about shared work responsibilities. Giriama children typically 'peak' in peer interactions at around the age of 5 years, with children's social time gradually giving way to errands, infant caregiving assistance, and adult-focused work tasks (Wenger, 2008). Throughout their lives, however, Giriama males prepare for the role of '*mwenye mudzi*,' a person who has learned to balance dominance and sociability in the Kenyan patriarchal society. Therefore, among Giriama children, young boys are provided with more unstructured social time than girls, and this divergent allotment of social time lasts through adolescence and into adulthood. For this reason, Giriama males have more opportunities to develop and maintain close familial, peer, and community-level relationships than females.

Intimacy

In many cultures, friendship is defined by mutual trust and feeling sufficiently secure to share personal information with a peer. Around the world, however, not all friendships take on the same characteristics of intimacy. In some Alaska Native villages, it is culturally inappropriate to discuss sensitive topics, regardless of whether a child wishes to share this information with a friend or nonfriend. In these cases, silence may be more indicative of mutual understanding and/or respect than the sharing of information (Gudykunst, 1998). In contrast, young children from Eastern and Southern cultures such as South Korea and Cuba are often described as having cultural values that exemplify group harmony and intimate relationships. Individuals from South Korea and Cuba describe the exchange of intimate information exclusively within a dyad as the most important characteristic of friendship and suggest that friendship is crucial for societal success (Triandis et al., 1988). Researchers have shown that young children from these two countries value intimacy more than children in North America (Gonzalez et al., 2004; French et al., 2006).

Conflict

As previously noted, since friends spend more time together than nonfriends, there are more opportunities for conflict

(Newcomb and Bagwell, 1995). In North America where assertiveness and competition are prominent cultural values, parents encourage young children to negotiate or 'work out' problems with friends (Rubin and Menzer, 2010). Young children are viewed as developmentally skilled if they are able to resolve interpersonal conflicts because individuals from Western cultures often view the 'friendly' disagreements in early childhood as practice for navigating discord later in life.

In Italy, friendships are wrought with what others may perceive as extreme conflict and disagreement. At a very young age, Italian children begin to engage in '*discussione*,' or lengthy verbal exchanges with same-age peers (Corsaro and Rizzo, 1988). *Discussione* includes intense 'debates and points of contention about various aspects of children's social worlds' as well as the accuracy of previous conversations between the children. To some, this may appear as if one child is 'picking on' another, but *discussione* is viewed as normative in many Italian families. Further, *discussione* is often perceived as a process by which young, preschool-aged friends use conflict to help each other learn about the adult world and advance through development.

In contrast to young Italian children or those from North American and other Western European countries, children from East Asian cultures tend to disengage from conflict instead of mediating or debating about disagreements (Xu et al., 2006). In China, for example, young children practice '*ren*,' also known as forbearance, during times of animosity. The self-control and restraint associated with *ren* helps promote group harmony and tolerance of others with conflicting viewpoints, a social skill that is important in East Asian societies. Thus, in some areas of the world, the active withholding of disagreement may be more salient in close friendship than conflict.

Psychosocial Outcomes

Several researchers have demonstrated that children with positive quality friendships also have higher self-esteem and lower levels of childhood internalizing and externalizing problems, such as depression, anxiety, aggression, and/or other delinquent behaviors (Rubin et al., 2006). However, much of the extant research about the positive outcomes of early friendship involves children from Westernized countries.

In other areas of the world, close friendships are not always predictive of higher self-esteem. For instance, having close friends is not directly associated with self-worth among Chinese, Indonesian, Arab, or Caribbean children (Chen and French, 2008; Dayan et al., 2001). Scholars suggest that this cultural disparity is related to overarching socialization values and ideals. If children are tasked with making friends and being socially competent members of society (as in Western cultures), the failure to achieve such cultural goals will contribute to lower levels of self-esteem. In cultures where fitting into social situations is not a primary cultural requirement, children without friends will not develop a poor sense of self (Chen and French, 2008).

Social scientists suggest that young children may exhibit more internalizing and externalizing behaviors in societies that value independence, competitiveness, and assertiveness because it is acceptable to acknowledge individual concerns in individualistic cultures (Cole and Tan, 2002). In contrast, for children from countries that value cooperation and self-control, such as India or other Eastern Asian countries, it may be culturally inappropriate to disrupt group processes by focusing on personal sentiments (Wilson et al., 2012). Therefore, it is argued that children without friends from Western cultures may display poorer psychosocial outcomes than children from Eastern or Southern cultures because it is more acceptable to have or report such behaviors in regions of the world that value expression over self-restraint (Cole and Tan, 2002; Wilson et al., 2012).

Research Gaps: Early Friendship and Culture

In addition to the aforementioned limitations of research about friendship during infancy and early childhood, there are gaps in the literature about early friendship and culture. Most studies have used an etic approach and empirical instruments developed in Western countries to study cross-cultural variations in friendship. In order to most accurately study culture and friendship, however, researchers should consider using an emic perspective in their cross-cultural or cultural studies. In particular, scholars should analyze reported and observed indicators of friendship using measures and methodologies coconstructed with local stakeholders and cultural experts.

Another limitation of extant research involves the generalization of studies within a society. In the United States, for instance, children grow up in rural, suburban, and urban areas, and such contexts may provide within-culture variations in friendship values and ideals. Thus, researchers would do well to investigate within country 'subcultures' in an effort to appropriately depict the nature of early childhood friendships.

Conclusion

In this article, we have focused on the function and prevalence of friendship in infancy and early childhood. We discussed the methodologies used to study early friendship as well as limitations of extant research. Also, we highlighted several studies that used etic and emic approaches to study cultural variations in friendship during early stages of development. Despite the information currently available about early friendships, additional studies are required to explore characteristics of young children's friendships. Furthermore, cross-cultural researchers are tasked with using dyadic, observational, and emic methodologies to comprehensively explore within- and across-culture variations in friendship during infancy and early childhood.

See also: Attachment During Infancy and Early Childhood: Cultural Variations, Development of; Cross-Cultural Research Methods in Psychology; Emotional Development, Effects of Parenting and Family Structure On; Friendship During Childhood and Cultural Variations; Integrated Attachment Theory; Play During Infancy and Early Childhood: Cultural Similarities and Variations; Prosocial Behavior During Infancy and Early Childhood: Developmental Patterns and Cultural Variations; Self-Development in Childhood and Adolescence; Social and Emotional Development in the Context of the Family.

Bibliography

Aboud, F.E., Mendelson, M.J., 1996. Determinants of friendship selection and quality: developmental perspectives. In: Bukowski, W.M., Newcomb, A.F., Hartup, W.W. (Eds.), The Company They Keep: Friendship in Childhood and Adolescence. Cambridge University Press, New York, pp. 366–405.

Bagwell, C.L., Schmidt, M.E., 2011. Friendships in Childhood and Adolescence. Guilford Press, New York.

Bigelow, B.J., 1977. Children's friendship expectations: a cognitive-developmental study. Child Development 48, 246–253.

Bowlby, J., 1969. Attachment, second ed. Basic Books, New York.

Chen, X., French, D.C., 2008. Children's social competence in cultural context. Annual Review of Psychology 59, 591–616.

Cole, P.M., Tan, P.Z., 2002. Cultural differences in children's emotional reactions to difficult situations. Child Development 73, 983–996.

Corsaro, W.A., Rizzo, T.A., 1988. Discussione and friendship: socialization processes in the peer cultural of Italian nursery school children. American Sociological Review 53, 879–894.

Dayan, J., Doyle, A.B., Markiewicz, D., 2001. Social support networks and self-esteem of idiocentric and allocentric children and adolescents. Journal of Social and Personal Relationships 18, 767–784.

Dunn, J., 2004. Children's Friendships. Blackwell Publishing, Malden, MA.

Eckerman, C.O., Whatley, J.L., Kutz, S., 1975. Growth of social play with peers during the second year of life. Developmental Psychology 11, 42–49.

French, D.C., Lee, O., Pidada, S., 2006. Friendships of Indonesian, South Korean, and U.S. youth: exclusivity, intimacy, enhancement of worth, and conflict. In: Chen, X., French, D.C., Schneider, B. (Eds.), Peer Relationships in Cultural Context. Cambridge University Press, New York, pp. 379–402.

French, D.C., Pidada, S., Victor, A., 2005. Friendships of Indonesian and United States youth. International Journal of Behavioral Development 29, 304–313.

Gaskins, S., 2008. Children's daily lives among the Yucatec Maya. In: LeVine, R.A., New, R.S. (Eds.), Anthropology and Child Development. Blackwell, Malden, MA, pp. 280–288.

González, Y.S., Moreno, D.S., Schneider, B.H., 2004. Friendship expectations of early adolescents in Cuba and Canada. Journal of Cross-Cultural Psychology 35, 436–445.

Gosso, Y., Carvalho, A.M.A., 2013. Play and cultural context. In: Encyclopedia on Early Childhood Development. http://www.child-encyclopedia.com/pages/PDF/Gosso-CarvalhoANGxp1.pdf (accessed 13.08.01.).

Gottman, J.M., 1986. The observation of social process. In: Gottman, J.M., Parker, J.G. (Eds.), Conversations of Friends: Speculation on Affective Development. Cambridge University Press, Cambridge, pp. 51–100.

Gudykunst, W., 1998. Bridging Differences: Effective Intergroup Communication, third ed. Sage University Press, Thousand Oaks, CA.

Hartup, W.W., 1996. The company they keep: friendships and their developmental significance. Child Development 67, 1–13.

Hay, D.F., Nash, A., Pedersen, J., 1983. Interaction between six-month-old peers. Child Development 54, 557–562.

Howes, C., 1991. A comparison of preschool behaviors with peers when children enroll in child-care as infants or older children. Journal of Reproductive and Infant Psychology 9, 105–115.

Howes, C., 2009. Friendship in early childhood. In: Rubin, K.H., Bukowski, W.M., Laursen, B. (Eds.), The Handbook of Peer Interactions, Relationships, and Groups. The Guilford Press, New York, pp. 180–194.

Krappman, L., 1996. Amicitia, drujba, shin-yu, philia, freundschaft, friendship: on the cultural diversity of a human relationship. In: Bukowski, W.M., Newcomb, A.F., Hartup, W.W. (Eds.), The Company They Keep: Friendship in Childhood and Adolescence. Cambridge University Press, New York, pp. 19–40.

Ladd, G.W., Kochenderfer, B.J., Coleman, C.C., 1996. Friendship quality as a predictor of young children's early school adjustment. Child Development 67, 1103–1118.

Matsumoto, D., 1997. Culture and Modern Life. Thomson Brooks/Cole Publishing Co., Belmont, CA.

Newcomb, A.F., Bagwell, C., 1995. Children's friendship relations: a meta-analytic review. Psychological Bulletin 117, 306–347.

Pedersen, S., Vitaro, F., Barker, E.D., Borge, A.I., 2007. The timing of middle-childhood peer rejection and friendship: linking early behavior to early-adolescent adjustment. Child Development 78, 1037–1051.

Piaget, J., 1932. The Moral Judgment of the Child. Harcourt, Oxford.

Rubin, K.H., Bukowski, W., Parker, J.G., 2006. Peer interactions, relationships, and groups. In: Eisenberg, N. (Ed.), Handbook of Child Psychology, Social, Emotional, and Personality Development, vol. 3. Wiley, New York, pp. 571–645.

Rubin, K.H., Coplan, R.J., Bowker, J.C., 2009. Social withdrawal in childhood. Annual Review of Psychology 60, 141–171.

Rubin, K.H., Coplan, R., Chen, X., Bowker, J., McDonald, K.L., 2011. Peer relationships in childhood. In: Bornstein, M.H., Lamb, M.E. (Eds.), Developmental Science: An Advanced Textbook. Psychology Press, New York, pp. 519–570.

Rubin, K.H., Menzer, M., 2010. Culture and social development. In: Encyclopedia on Early Childhood Development. http://www.child-encyclopedia.com/documents/rubin-menzerangxp.pdf (accessed 13.08.01.).

Seyfarth, R.M., Cheney, D.L., 2012. The evolutionary origins of friendship. Annual Review of Psychology 63, 153–177.

Sharabany, R., 2006. Peer relationships and intimate friendships among Arab and kibbutz children in Israel. In: Chen, X., French, D.C., Schneider, B. (Eds.), Peer Relationships in Cultural Context. Cambridge University Press, New York, pp. 452–478.

Sullivan, H.S., 1953. The Interpersonal Theory of Psychiatry. W.W. Norton & Co., New York.

Trevarthen, C., 1979. Communication and cooperation in early infancy: a description of primary intersubjectivity. In: Bullowa, M. (Ed.), Before Speech: The Beginning of Interpersonal Communication. Cambridge University Press, London, pp. 321–347.

Triandis, H., Bontempo, R., Villareal, M.J., Asai, M., Lucca, N., 1988. Individualism and collectivism: cross-cultural perspectives on self-ingroup relationships. Journal of Personality and Social Psychology 54, 323–338.

Walden, T., Lemerise, E., Smith, M.C., 1999. Friendship and popularity in preschool classrooms. Early Education and Development 10, 351–371.

Wenger, M., 2008. Children's work, play, and relationships among the Giriama of Kenya. In: LeVine, R.A., New, R.S. (Eds.), Anthropology and Child Development. Blackwell, Malden, MA, pp. 289–306.

Wilson, S.L., Raval, V.V., Salvina, J., Raval, P.H., 2012. Emotional expression and control in school-age children in India and the United States. Merrill-Palmer Quarterly 58, 50–76.

Wojslawowicz, J.C., Rubin, K.H., Burgess, K.B., Rose-Krasnor, L., Booth-LaForce, C.L., 2006. Behavioral characteristics associated with stable and fluid best friendship patterns in middle childhood. Merrill-Palmer Quarterly 52, 671–693.

Xu, Y., Farver, J.M., Chang, L., Yu, L., Zhang, Z., 2006. Culture, family contexts, and children's coping strategies in peer interactions. In: Chen, X., French, D.C., Schneider, B. (Eds.), Peer Relationships in Cultural Context. Cambridge University Press, New York, pp. 264–280.

Friendship Experiences of Lesbian, Gay, Bisexual, and Transgender Individuals

MP Galupo, Towson University, Towson, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Friendships can play an important role in the lives of lesbian, gay, bisexual, and transgender (LGBT) individuals. LGBT friendship patterns and experience have been shown to vary across sexual orientation, gender, gender identity, and race. Friendships also differ based on whether or not friends identify as part of the LGBT community. LGBT individuals often think of their friends as 'families of choice' as friendships take on increased significance when traditional support from family and community may not be consistently offered. In addition to having their own unique benefits, friendships with individuals outside of the LGBT community have been seen as important in the development of heterosexual allies to the LGBT community.

Historically, research on the relationship experiences of lesbian, gay, bisexual, and transgender individuals (LGBT) has deemphasized friendships in favor of focusing on 'primary' or romantic relationships (Weinstock, 1998). Friendships, however, can play an important role in the lives of LGBT individuals. In many ways, the friendship experiences of LGBT individuals are similar to the friendships of their heterosexual counterparts where LGBT and heterosexual individuals report a similar number of close friends (Galupo, 2009). Despite similarities to heterosexual friendships, friendship patterns and experience have been shown to vary across both sexual orientation (Galupo, 2007a, 2009; Ueno et al., 2012) and gender identity (Hines, 2007; Galupo et al., 2014). Among sexual minorities, friendship patterns have also been shown to differ across gender (Baiocco et al., 2012; Galupo, 2007a) and race (Galupo, 2007a).

Friendships are considered to have increased importance for LGBT individuals as friendship is emphasized during times of social change and is particularly salient for individuals when their identity is at odds with social norms and during times of transition (Weeks, 1995). The importance of friendship is underscored by the fact that when LGBT individuals first 'come out,' the majority make that first disclosure to a friend. Coming out to friends can be associated with positive outcomes such as strengthening feelings of trust and acceptance. It can also be met with negative reactions and outcomes, leading to loss of relationships or the creation of emotional distance within the relationship. Either way LGBT individuals experience friendship as a means of social support, which may take on increased significance when traditional support from family and community may not be consistently offered; LGBT individuals report less discrimination in the context of friendship than with their families of origin (Takács, 2006). Friendships, may thus take on additional importance by fulfilling a supportive role in relation to minority stress and status.

Scholars consider friendships to be distinct from other types of social relationships in that they are voluntary relationships constructed around reciprocity, mutuality, and equality. Research on friendships between LGBT individuals emphasizes that these friendships are particularly important because they represent a social relationship in which individuals can experience relative equality. For example, in his research on friendships between gay men, Nardi (1999) describes how these friendships can provide opportunities for gay men to experience equity not easily achieved in other friendships where social identity is being bridged and where differences in sexual orientation or gender require constant negotiation. Nardi's research underscores the importance of understanding the way in which friendships of LGBT individuals are impacted by broader cultural attitudes and how these broader social context impacts are experienced at a social and personal level. In particular, it is important to understand intersections of identity among social groups when considering the social and personal relationships of LGBT individuals. Friendships have been shown to be simultaneously impacted by sexism, racism, heterosexism, biphobia, and transphobia (Galupo, 2006; Galupo et al., 2014); not all LGBT individuals have the same experiences in friendship.

Looking at LGBT friendships in more depth, LGBT friendships have been examined using multiple perspectives including classifying friendships based on closeness with distinctions made between casual, close, and best friends (Baiocco et al., 2012; Galupo, 2009) and addressing friendship in the context of friendship dyads (e.g., Galupo, 2007b; Price, 1999), groups or networks (e.g., Tillman-Healy, 2001), and larger communities (e.g., Esterberg, 1997; Hines, 2007; Nardi, 1999). Although not always described as friendship analysis, these perspectives have been included in the research on most interpersonal relationships of LGBT individuals and are thus illustrative of LGBT friendships.

Friendships within the LGBT Community

Friends as Family

LGBT friendship research has mostly focused on understanding friendships that form between individuals who identify within the LGBT community. Consistent with the notion of homophily (the idea that individuals connect with others who are similar to themselves) lesbians and gay men form the majority of their friendships with individuals of the same sex who also identify with the LGBT community (Galupo, 2007b, 2009) with lesbian women befriending other lesbians (Stanley, 1996) and gay men befriending other gay men (Nardi, 1992, 1999). LGBT friendship

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10214-4

is often characterized as providing a unique type of familial support (Nardi, 1992; Weinstock, 1998; Hines, 2007) where friendships function as 'families of choice' (Weston, 1991) and serve to buffer LGBT individuals from social isolation or rejection associated with homophobia, biphobia, and transphobia. That LGBT individuals consider others within the community in familial terms is evidenced by the use of the colloquial term 'family' to refer to other individuals in the community.

Within LGBT communities, friendships are often characterized as social networks that make up the larger LGBT (Esterberg, 1997; Tillman-Healy, 2001) and transgender (Hines, 2007) communities and thus allow individuals to positively experience their lesbian, gay, and bisexual (LGB) or transgender identity (Riggle and Rostosky, 2011). These within-community friendships are seen as having unique benefits as they may provide a sense of shared experience and an avenue for processing minority status. Within-community friendships can provide a buffer against being socially devalued as a sexual minority and can provide an outlet for sharing daily aspects of life, especially for individuals who are not generally open about their sexual orientation identity (Stanley, 1996). LGBT friends can also serve as role models in a culture where LGBT experience is not widely visible.

While lesbians and gay men may find a unique type of support through same-orientation friendships, bisexual women and men are less likely to do so. Bisexual women and men are less likely to have a friend with the same (bisexual) identity as themselves (Galupo, 2007a) and report that their bisexual identity is often not acknowledged within their cross-orientation friendships with heterosexual (Galupo, 2007b) and lesbian/gay (Rust, 1995) individuals. Similarly, LGBT individuals of color are faced with simultaneously negotiating sexual orientation and race in their friendships (Galupo, 2007b; Hall and Rose, 1996) and their pressure to choose one identity over the other is likely to impact their experience of friendship.

Transgender Friendships: Friends as 'Family' and as 'Community Resource'

In many ways, lesbian and gay within-community friendships generally serve as a proxy to family relationships. They can provide a sense of belongingness and shared culture that may not be fulfilled by families of origin. Friends can serve as 'families of choice' by providing emotional support not provided by lost or strained familial relationships due to differences in sexual orientation. Transgender friendships, likewise, fulfill a similar familial role. However, the stigma and marginalization experienced by transgender individuals is even more pronounced than for LGB individuals. Even though LGB and transgender individuals are generally conceptualized as a unified group, transgender persons often experience additional stigmatization and discrimination, are pathologized within the medical community, and do not enjoy the same legal protections afforded to sexual minorities (Lev, 2004). Even within the LGBT community, transgender individuals and issues related to gender identity have a history of being overshadowed by the LGB community. Taken together, transgender individuals are more likely to be subject to social isolation and invisibility.

It makes sense, then, that friendships take on an additional level of significance for transgender individuals. Transgender friends are generally characterized as having an understanding of nonnormative experience; are knowledgeable on issues of gender, sex, and privilege; are nonjudgmental/open-minded; and offer a relationship in which transgender individuals can feel comfortable being themselves and talk about transgender issues (Galupo et al., 2014). In short, these friendships provide a shared sense of belonging, family, and community.

While providing support and having similar experiences are regarded as characteristics of friendship in general, transgender individuals describe these benefits primarily in relation to decisions and experiences surrounding transitioning and in ways that are specific to transgender experience (Hines, 2007). Transgender friends are also seen as providing needed counseling unavailable from the traditional health care system. Benefits of friendships with other transgender individuals are largely seen as filling in the gaps of support and services not otherwise provided by traditional family, friends, and institutions. In particular, transgender–transgender friendships have been shown to have unique benefits not experienced to the same degree with LGB individuals. When compared to their friendships with LGB individuals, transgender individuals find that with their transgender friends they are more likely to have shared experiences, be able to talk about transgender issues, receive support via mentoring and shared resources, and are more comfortable being themselves (Galupo et al., 2014).

Friendships, Romance, and Sexuality

In the general literature, friendships are often explicitly distinguished from romantic relationships and sexual interest. The platonic nature of friendships, then, is assumed to provide a unique type of relationship. However, the LGBT friendship literature from the beginning has suggested that the same distinctions may not hold for LGBT individuals. Rather, LGBT friendships may defy traditional heterosexist definitions of friendships in important ways (Rose, 2000). Friendships of LGBT individuals may not be void of romantic feelings and have sometimes even been described as 'love relationships' (Kitzinger and Perkins, 1993). For LGBT individuals, friendships have been seen as the most common relational context in which romantic relationships develop (Herdt and Boxer, 1996) where many LGB individuals report having first been friends with their partner before becoming romantically involved (Rose and Zand, 2002). Likewise, LGB individuals when asked about their friendships often count lovers and exlovers among their close friends (Clunis and Green, 2000; Weinstock, 1998).

Passionate friendships have also been described in the literature as developing most often out of best same-sex friendships where intense emotional relationships develop in ways similar to romantic relationships but lack a sexual component (Diamond, 2002). Although more likely to be studied among young lesbian and bisexual women, friendships can lead to intense emotional interactions that incorporate elements of both friendships and romantic relationships for LGB women or men. It is also important to note that passionate friendships are not necessarily exclusively experienced by individuals who identify as sexual minorities (Diamond, 2002).

Clearly, the research on friendships of LGBT individuals emphasizes that friendships can include elements similar to romantic relationships. Likewise, friendships may also include elements of sexual interest (Diamond, 2002; Herdt and Boxer, 1996; Rose and Zand, 2002). For example, friendships may begin first with sexual interest and activity and then later develop into friendship (Nardi, 1999; Weinstock, 1998). In addition, casual friendships may encompass sexual activity without a romantic component and friendships may also serve as a site for identity development as it relates to sexual exploration (Morgan and Thompson, 2006).

Friendships Outside the LGBT Community

Friendships outside the LGBT community are tenuously constructed around a number of barriers. Barriers include the stigma of having an LGB friend, sexual tension, and reduced comfort for both the LGB (Price, 1999) and heterosexual (Mohr and Sedlacek, 2000) individuals in the friendship. Although bisexual individuals are more likely to have cross-orientation friendships with heterosexual individuals (Galupo, 2007a), these friendships exist at the cost of bisexual identity where issues related to bisexual identity are not always acknowledged within the friendship (Galupo, 2007b).

Benefits of friendships of LGBT individuals with others outside the LGBT community have also been considered. For example, lesbian and bisexual women report that through their friendships with heterosexual women they gain an understanding that acceptance from heterosexuals is possible, gain an objective perspective in their lives, are able to break down stereotypes, and experience increased closeness and trust within the friendship accompanying sexual orientation disclosure (Galupo and St. John, 2001). Muraco (2012) discusses ways in which friendships with heterosexual individuals can function as 'intentional families' for both LGB and heterosexual men and women.

The literature on the friendships of transgender individuals also details a variety of benefits to friendships outside of the LGBT community. These include fostering a feeling of normalcy, offering more diverse perspectives and interactions, helping transgender individuals to present as their identified gender, as well as validating transgender identity and experience. Additionally, these friendships provide an opportunity to educate others about transgender experience (Galupo et al., 2014).

Friends as Allies

Friendships outside the LGBT community provide both LGB and transgender individuals the opportunity to educate individuals about LGBT experience and discrimination. Friendships, then, may be an important avenue for developing allies to the LGBT community. The research literature has long connected interpersonal contact – inclusive of friendship – as being related to positive attitudes toward LGBT individuals. More recently, friendships with LGBT individuals has been more directly conceptualized as important in the development of heterosexual allies to the LGBT community (Duhigg et al., 2010; Fingerhut, 2011). LGBT allies can serve to both personally support LGBT individuals and advocate for larger acceptance and rights of LGBT individuals.

Conclusion

Friendship experience of LGBT individuals challenges the notion of friendship as secondary to romantic relationships. The emphasis on friends as 'family' (chosen family) illustrates that for LGBT individuals friendships both within and outside the LGBT community serve important roles. Research on the experiences of individuals within the transgender communities indicate that friendship fulfills even a broader role by providing support and services not otherwise provided by traditional family, friends, and institutions. Likewise, friendships with individuals outside the LGBT community provide a sense of acceptance and an opportunity for developing allies to the LGBT community. Although the research emphasizes a unique and potentially broader role of friendship for LGBT individuals when compared to their heterosexual counterparts, it is important to note that this is, in part, due to the fact that friendships are experienced within a larger system of discrimination and inequality still present in family and other societal institutions.

See also: Gay and Lesbian Movements; Heterosexism and Homophobia; Lesbian, Gay, Bisexual and Trans-sexual Butch-Femme Subcultures; Lesbian, Gay, Bisexual and Trans-sexual Drag Culture; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Neighborhood Effects; Prosocial Behavior and Empathy; Queer Geography; Sexual Minority Families of Choice and Family Networks: We Are Family; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Baiocco, R., Laghi, F., Di Pomponio, I., Nigito, C.S., 2012. Self-disclosure to the best friend: friendship quality and internalized sexual stigma in Italian lesbian and gay adolescents. Journal of Adolescence 35, 381–387.

Clunis, D.M., Green, S.A., 2000. Lesbian Couples. Seal Press, Seattle, WA.

Diamond, L.M., 2002. "Having a girlfriend without knowing it": intimate friendships among adolescent sexual-minority women. Lesbian Studies 6 (1), 5–16.

Duhigg, J.M., Rostosky, S.S., Gray, B.E., Wimsatt, M.K., 2010. Development of heterosexuals into sexual-minority allies: a qualitative exploration. Sexual Research and Social Policy 7, 2–14.

Esterberg, K.G., 1997. Lesbian and Bisexual Identities: Constructing Communities, Constructing Selves. Temple University Press, Philadelphia.

Fingerhut, A., 2011. Straight allies: what predicts heterosexuals' alliance with the LGBT community. Journal of Applied Social Psychology 41, 2230–2248.

Galupo, M.P., 2006. Sexism, racism, and homophobia: the framing of bisexual women's friendships. Journal of Bisexuality 6, 35–45.

Galupo, M.P., 2007a. Friendship patterns of sexual minority individuals in adulthood. Journal of Social and Personal Relationships 24 (1), 139–151.

Galupo, M.P., 2007b. Women's close friendships across sexual orientation: an analysis of lesbian-heterosexual and bisexual-heterosexual women's friendships. Sex Roles 56, 473–482.

Galupo, M.P., St. John, S., 2001. Benefits of cross-sexual orientation friendships among adolescent females. Journal of Adolescence 24, 83–93. http://dx.doi.org/10.1006/jado.2000.0364.

Galupo, M.P., 2009. Cross-category friendship patterns: comparison of heterosexual and sexual minority adults. Journal of Social and Personal Relationships 26, 811–831.

Galupo, M.P., Bauerband, L.A., Gonzalez, K.A., Hagen, D.B., Hether, S.D., Krum, T.E., 2014. Transgender friendship experiences: benefits and barriers of friendships across gender identity and sexual orientation. Feminism & Psychology 24 (2), 193–215. http://dx.doi.org/10.1177/0959353514526218.

Hall, R., Rose, S., 1996. Friendships between African-American and White lesbians. In: Weinstock, J.S., Rothblum, E.D. (Eds.), Lesbian Friendships. New York University Press, New York, pp. 165–191.

Herdt, G., Boxer, A., 1996. Children of Horizons. Beacon, Boston.

Hines, S., 2007. TransForming Gender: Transgender Practices of Identity, Intimacy, and Care. Policy Press, Bristol, UK.

Kitzinger, C., Perkins, R., 1993. Changing Our Minds: Lesbian Feminism and Psychology. New York University Press, New York.

Lev, A.I., 2004. Transgender Emergence: Therapeutic Guidelines for Working with Gender-variant Individuals and Their Families. Haworth Press, New York.

Mohr, J.J., Sedlacek, W.E., 2000. Perceived barriers to friendship with lesbians and gay men among university students. Journal of College Student Development 41, 70–80.

Morgan, E.M., Thompson, E.M., 2006. Young women's sexual experiences within same-sex friendships: discovering and defining bisexual and bi-curious identity. Journal of Bisexuality 6, 7–34.

Muraco, A., 2012. Odd Couples: Friendships at the Intersection of Gender and Sexual Orientation. Duke University Press.

Nardi, P.M., 1992. That's what friends are for: friends as family in the gay and lesbian community. In: Plummer, K. (Ed.), Modern Homosexualities: Fragments of Lesbian and Gay Experience. Routledge, New York, pp. 108–120.

Nardi, P.M., 1999. Gay Men's Friendships: Invincible Communities. Chicago University Press, Chicago.

Price, J., 1999. Navigating Differences: Friendships between Gay and Straight Men. Harrington, New York.

Riggle, E.D.B., Rostosky, S.S., 2011. A Positive View for LGBTQ: Embracing Identity and Cultivating Well-being. Rowman & Littlefield, New York.

Rose, S., 2000. Heterosexism and the study of women's romantic and friend relationships. Journal of Social Issues 56 (2), 315–328.

Rose, S.M., Zand, D., 2002. Lesbian dating and courtship from young adulthood to midlife. Journal of Lesbian Studies 6 (10), 85–109.

Rust, P.C., 1995. Bisexuality and the Challenge to Lesbian Politics: Sex, Loyalty, and Revolution. New York University Press, New York.

Stanley, J.L., 1996. The lesbian's experience of friendship. In: Weinstock, J.S., Rothblum, E.D. (Eds.), Lesbian Friendships. New York University Press, New York, pp. 39–59.

Takács, J., 2006. Social Exclusion of Young Lesbian, Gay, Bisexual, Transgender (LGBT) People in Europe. ILGA Europe, Brussels, Belgium.

Tillman-Healy, L.M., 2001. Between Gay and Straight: Understanding Friendship across Sexual Orientation. Rowman and Littlefield Press, Lanham, MD.

Ueno, K., Wright, E.R., Gayman, M.D., McCabe, J.M., 2012. Segregation in gay, lesbian and bisexual youth's personal networks: testing structural constraint, choice homophily and compartmentalization hypotheses. Social Forces 90, 971–991.

Weeks, J., 1995. Invented Moralities: Sexual Values in an Age of Uncertainty. Polity Press, Cambridge.

Weinstock, J.S., 1998. Lesbian, gay, bisexual and transgender friendships in adulthood. In: Patterson, C.J., D'Augelli, A.R. (Eds.), Lesbian, Gay, and Bisexual Identities in Families: Psychological Perspectives. Oxford University Press, London, pp. 122–155.

Weston, K., 1991. Families We Choose: Lesbians, Gay, Kinship. Columbia University Press, New York.

Fromm, Erich (1900–80)

Iga Mergler, Neil McLaughlin, and Ismaël Traoré, McMaster University, Hamilton, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents an analytic history of the legacy of Erich Fromm, the German psychoanalyst, social psychologist, critical thinker, best-selling author, and an early member of the Frankfurt School. Fromm's intellectual insights flowed from aspects of his life and optimal marginality in relationship to various intellectual social movements and his complex relationship with the Frankfurt School. But Fromm also became a 'forgotten intellectual' for some of the same biographical and sociological reasons. Ideas and scholarly influence can rise again as well as decline, so with his contemporary relevance in mind, we will discuss some of the underappreciated contributions Fromm made to the study of authoritarianism, race, and ethnicity as well as his influence on theories of gender and the practice of public sociology. There is new interest in Fromm's work, suggesting that he might find again a place in the history and contemporary scholarship in social psychology, sociology, critical theory, and psychoanalysis.

Introduction

Erich Fromm (1900–80) was a German psychoanalyst, social psychologist, critical social thinker, social activists, best-selling author, and a public intellectual. He was first recognized for his original analysis of the rise of the Nazi movement in the 1930s presented in *Escape from Freedom* (1941), a book where he creatively connected Marxist, Freudian, Weberian, and existentialist insights. In subsequent years, he wrote such best sellers as *The Sane Society* (1955), *The Art of Loving* (1956), and *To Have or To Be?* (1976) and returned to academic scholarship in *Social Character in a Mexican Village* (1970) (with Michael Maccoby) and *Anatomy of Human Destructiveness* (1973). Popularity among wider audiences, however, gradually undermined Fromm's recognition among academics, especially after he was ostracized from psychoanalysis and air brushed out of the 'origin myths' of early Frankfurt School 'critical theory,' rendering him a forgotten intellectual (McLaughlin, 1998, 1999).

Fromm's position of 'optimal marginality' (McLaughlin, 2001b), in his professional as well as private life, allowed him to become well acquainted with numerous theoretical perspectives and to acquire skills for effortlessly maneuvering between different intellectual networks, movements, and social groups. Fromm was trained in Germany, bringing to America European Marxist, Freudian, and Weberian theories, but he also integrated into and learned from American culture and intellectual life more than most other émigré intellectuals of his generation. What resulted from the combination of Fromm's training and wide reading, his sociological 'optimal marginality,' and his cultural border crossing, was an eclectic theoretical stand, a compelling ethical vision, and a jargon-free language that reached different publics around the world.

We will first discuss Fromm's personal life and career and then we will highlight the relevance of his ideas to contemporary social science scholarship. We will present the largely unacknowledged intellectual contributions he made to the three fields of study (authoritarian personality research, race and racism, and gender studies) and discuss the pioneering role he played in mid-twentieth-century development of the role of the public intellectual and public sociologist, two overlapping models for intellectuals that are being widely debated today.

Fromm's Personal Life and Career

Fromm was born in 1900 to a middle-class Orthodox Jewish family. The loveless marriage his parents were trapped in created an environment whereby Fromm became a neurotic child (Burston, 1991). Overcoming obstacles, young Erich studied under prominent rabbis and acquired intellectual tools he used to challenge his parents' worldview. After abandoning a plan to become a rabbi himself, Fromm enrolled at the age of 17 years at the university in Heidelberg, where he was mentored by Max Weber's younger brother, Alfred Weber. Fromm's doctoral thesis (1922) analyzing three Jewish sects ability to follow their religious traditions without receiving any institutional support (Ortmeyer, 1998) was essentially a social psychological work focusing on the influence of social forces on the lives of individuals, a reappearing theme in Fromm's books. Impressed by his thesis, Weber encouraged Fromm to pursue an academic career, a plan he rejected because he wanted to make a social and political difference (Friedman, 2013). Instead, he became an editor of a Jewish newspaper.

The decisions taken by Fromm in those early years signaled the type of intellectual that he wanted to be. Firstly, he was not afraid to take an unpopular stance, as he moved beyond the Judaism and later the Zionism of his family. Secondly, in the Marxist vein that increasingly came to influence him, Fromm did not want to simply interpret the world, but aspired to change it. Realizing that such ambitions might be possible by reaching a broader public, he aspired to be what we now call a public intellectual.

His public intellectual career, however, would have to wait. After graduating and then quitting the Jewish newspaper, Fromm became a trained psychoanalyst working in Berlin. In that period he resumed his academic and theoretical work when he accepted the invitation from Max Horkheimer in 1929

to become a member of *Institut für Sozialforschung* (the Frankfurt Institute for Social Research), a network of interdisciplinary scholars we now know of as the Frankfurt School. Fromm was director of social psychology and worked with the critical theorists as a researcher (Friedman, 2013: 26). The rise of National Socialism and growing anti-Semitism pushed Fromm to leave Germany first to Switzerland then to New York, where he worked with the now exiled immigrant members of the Frankfurt School (Wheatland, 2009) at Columbia University. The Frankfurt School was brought to Columbia largely due to the efforts of the sociology department, particularly Robert Lynd, because they were interested in Fromm's research on the working class in Weimar Germany (Wheatland, 2009).

Fromm's work in the institute dealt with family, religion, anti-Semitism, the class-biased punitive justice system, scapegoating, and much more (Friedman, 2013: 34). Fromm participated in and codesigned collaborative projects, wrote for the critical theorist's journal, and received a significant wage (Friedman, 2013: 31). During the 1930s, Fromm had conducted a major empirical study of the social psychology of Weimar German workers, work that created the concept of 'authoritarian character.' Fromm's position in the Frankfurt School evolved together with changing social and personal life circumstances. When they started cooperation, Fromm had just finished his psychoanalytic training and could maintain a working relationship with Horkheimer, who was influenced by Freudian orthodoxy. With time, however, this cooperation would unravel. Fromm started to distance himself theoretically from the Frankfurt School at the time when he started a relationship, in the US, with Freudian revisionist Karen Horney. As Lawrence Friedman argued, Fromm adapted to life in the U.S. rather quickly. He was the only member of the Frankfurt School who quickly moved from 'proficiency to mastery of English' and started publishing in this language (2013: 78). Additionally, Fromm put in extra effort to become familiar with and interested in American popular culture, and dabbled with commenting on relevant American matters, activities that Adorno criticized. Soon, he joined a network of thinkers called the 'Culture and Personality' movement with his friends Edward Sapir, Ruth Benedict, and Margaret Mead (Friedman, 2013: 59) and began attending weekly meetings of the Zodiac Group which consisted of artists, scholars, and poets exchanging ideas (Greenberg and Mitchell, 1983). Thus, Fromm's social networks went beyond Horkheimer's circle to include relationships with other social and political scientists from Columbia University. Finally, Fromm acquired independence from Frankfurt School in yet another way – by being a successful psychoanalyst, with a stable source of income. All this intensified the growing conflict between Fromm and the members of Frankfurt School. Both Horkheimer and Adorno were suspicious of Fromm's ability and willingness to connect to American scholars and political and cultural trends (Wiggerhaus, 1986). Over time, they would respond by trying to diminish his role in the critical school of thought, helping render him a 'forgotten intellectual' (McLaughlin, 1998, 1999).

Fromm always rejected orthodoxies, and he did not hesitate to publish a series of critiques of psychoanalytic theory, throughout the 1930s (Burston, 1991), as his emerging ideas represent a significant break from Freudian orthodoxy. For Horkheimer and his colleagues, Fromm's work deviated far too much from their theoretical agenda and eventually they would pressure him to leave (Friedman, 2013: 37). Conflicts over Fromm's unorthodox reading of Freud and Marx, financial matters (Friedman, 2013: 57), and disagreements over methodological and practical matters relating to the proposed publication of his study of the working class in Weimar Germany led to him leave Horkheimer's group in 1939 (Burston, 1991; Wiggerhaus, 1986). Fromm's research from this period, however, created the basis for *Escape from Freedom* (1941). Published with a commercial press and written in engaging nonacademic prose, *Escape from Freedom* argued that popular psychological explanations overemphasized Hitler's pathological personality and the 'madness' of the sociopolitical movement supporting him. Furthermore, Fromm opposed Marxist deterministic interpretations suggesting that Nazis simply served the interest of authoritarian German capitalists. Instead, Fromm offered a more complex sociological explanation of Nazism, linking this movement with anxieties arising out of rapid social changes brought about by modernity (McLaughlin, 1996). He suggested that humiliation experienced by Germany after World War I and the subsequent economic crisis undermined the legitimacy of its democratic institutions, making its citizens more anxious. The millions of Germans were seduced by Hitler's 'evangelism of self-annihilation' (Fromm, 1969[1941]: 259) that was offering a way to reverse the effects of cultural and economic collapse. One path to escape from freedom, in Fromm's view, was offered through following obediently a leader and his racist, nationalistic, and militaristic Nazi Party (Fromm, 1969[1941]: 235). As the noted political theorist and historian of psychoanalysis Paul Roazen observed, the publication of this book "was a notable moment in intellectual history" (Roazen, 1977). Not only was it enthusiastically reviewed in elite newspapers and magazines (McLaughlin, 1996) but it also influenced generations of social scientists such as Robert K. Merton's graduate students at Columbia University.

Ironically, the success of *Escape from Freedom* helped erase from our historical memory Fromm's influence on Horkheimer's circle. Fromm became known as a theorist of freedom not authoritarianism (Rickert, 1983), and gradually came to be ignored by contemporary critical theorists who accepted Marcuse and Adorno's account of the conflict within the Frankfurt School circle (McLaughlin, 1999). This erasure of Fromm from this history is unjustified, since his study of the character of Weimar German workers laid the foundation for the concept of 'authoritarian personality' that Theodor Adorno used to write *The Authoritarian Personality*, making authoritarianism almost a brand-name concept for the Frankfurt School (McLaughlin, 1998; Brunner, 1994). Fromm's work underpinned the development of the authoritarian personality tradition, which allowed social scientists to combine psychological explanations of scapegoating occurring in political life with an empirical 'F' scale measure of the level of authoritarianism. The involvement of Fromm in developing an earlier version of the 'F' scale was never adequately recognized until the publication of the manuscript of *The Working Class in Weimar Germany* (Bonns, 1984).

The Sociology of Race and Racism

The concept of authoritarian character introduced in *Escape from Freedom* was relevant for the development of literature on xenophobia and prejudice in the early twentieth century (Chancer, 1991). Fromm's studies of prejudice concentrated on historical circumstances conducive to its emergence, but also attempted to reveal more universal aspects of oppression. He identified a mechanism in which the dominating authoritarian person feels freer, more secure, and powerful knowing that there is someone whom they can or they have already managed to dominate. In this sense, the term 'social character' coined by Fromm to describe common traits of people belonging to the same group can be seen as providing means for the expression of group narcissism (superior 'us' as opposed to inferior 'them') and of nonegalitarian or authoritarian personality that feed unconscious racism (Fromm, 1964), key elements in contemporary scholarship. In short, Fromm's contribution to the study of prejudice and racism comes down to highlighting social psychological aspects of these phenomena.

Revisiting Fromm's contributions (especially the concept of 'social character') seems fruitful in the context of racism studies' shift of their scope from explicit conscious racism to automatic, unconscious racism (see Wheeler and Fiske, 2006[I1]). Few race scholars implied that subscribing to egalitarian values that bypass the cognitive process of forming stereotypes can also inhibit the development of unconscious biased, prejudiced, or racist attitudes (Pearson et al., 2009). Pearson et al. tested this contention by utilizing Fromm's Marxist-influenced concept of the 'marketing character,' a personality interested in maintaining economic differences among people. Pearson et al. concluded that people demonstrating traits of 'marketing character' are more prone to be racially prejudiced than those believing in an equal distribution of material resources. Fromm's use of Freudian insights into unconscious drives and emotions combined with Marxist and Weberian sociological perspectives on class and status provides productive intellectual resources for contemporary prejudice and race scholars.

The strengths as well as some of the limitations of Fromm's theorizing of race emerged both from his positioning between various networks and from his personal life. The combination of Fromm's socialist politics with his prophetic Jewish ethical thinking was creatively translated into his works (e.g., *The Heart of Man* or *The Anatomy of Human Destructiveness*) which, while discussing more ethical concepts such as the human roots of evil and destructiveness, criticized at the same time racism, ethnocentrism, and nationalist chauvinism (Fromm, 1964, 1973). On a more personal level, a large part of Fromm's energies in the 1930s were spent on trying to rescue Jewish family and friends from Nazi prosecution (Friedman, 2013). But also, after his break up with German psychoanalyst Karen Horney, he started seeing Katherine Dunham – a talented African-American dancer, choreographer, graduate of social anthropology, and researcher at the University of Chicago (Friedman, 2013: 91). This rare interracial relationship in 1930s America, as Lawrence Friedman argued, opened Fromm's eyes to discrimination and constraints that African-Americans experienced, even though, like many German émigré scholars, Fromm never fully understood the nuances of American racial politics. His connection to Dunham, his involvement in the American socialist party in the 1950s and early 1960s, and his intellectual engagement with the American pragmatism of John Dewey (a foreign territory to Marcuse and Adorno) meant that *Escape from Freedom*, which he wrote while dating Dunham, (1941) and his later book *The Sane Society* (1955) contained an emotionally compelling defense of antiracist universalism, even though he failed to theorize racism and anti-Semitism adequately. This universalism led him to gender theories to which we now turn.

Gender, Freud, and Psychoanalytic Feminist Sociology

Fromm's contributions to theories of gender are considerable yet controversial. Fromm's role in transforming of Freud's patriarchal psychological theory into the object relation psychology, self-psychology and feminist psychoanalytic sociology was obvious from the 1930s till today. By synthesizing Marx's and Freud's insights into the dynamics of character, emotions, and the unconscious, in the context of an empirical sociology, Fromm arrived at completely different conclusions than Freud about the role of the mother in the life of child. Freud's views of woman as essentially a castrated man were replaced by a theory that saw the mother as a powerful figure in child's life and their relationship as the strongest connection humans can have.

Another element of Freud's theory that Fromm questioned was the assumption of the universal character of femininity and masculinity; Fromm's revision of psychoanalysis attempted to show the socially constructed nature of these categories (Davis, 2003). This theoretically based criticism was later built on by sociologist Nancy Chodorow, who used empirical clinical data to develop the critique of the adequacy of Freudian theories of Oedipus complex and penis envy (1978). The revision of Freudian theory that Fromm offered (Burston, 1991; Rasmussen and Salhani, 2008) made him one of the most visible, articulate, and effective critics of psychoanalytic orthodoxy in the 1940s through the early 1960s.

Fromm's contribution to the emergence of feminist psychoanalysis can, once more, be explained by his personal life and sociological position on the optimal margins of psychoanalysis. Fromm's ideas can be recognized as contributing to the development of object relation theory (Chodorow, 1989; Greenberg and Mitchell, 1983; Chodorow, 1978: 50). Fromm's sociological vision insisted on seeing each patient's individual relational history as a sociologically shaped trajectory, which indirectly led to a crystallization of object relations theory in psychoanalysis and in social science more generally.

Fromm's ties and relationships with several remarkable women impacted the direction he took in his work (Davis, 2003). These women were either his mentors, like his first wife Frieda Fromm-Reichmann, or his intellectual partners who helped to broaden his horizons, like Katherine Dunham or Karen Horney, or his colleagues or fellows, in this case, female intellectuals, like Margaret Mead, with whom he worked within various intellectual circles (Burston, 1991; Rasmussen and Salhani, 2008).

Fromm's fascination with psychoanalysis started at the university, but was accelerated by his encounter with his

therapist and then first wife Frieda Fromm-Reichmann (Friedman, 2013: 21). Thus, in his late 20s, Fromm underwent psychoanalytic training and went on to work successfully as a clinician for 44 years (Ortmeyer, 1998; Burston, 1991). Frieda at the time helped him transition from Jewish Orthodoxy and Freudian orthodoxy as he moved beyond both. Fromm-Reichmann's ideas shaped Fromm's emerging intellectual agenda in the 1920s and certainly she helped him establish himself professionally. She was not, however, the last woman to have a transformative and generative influence on his life and ideas.

Protofeminist psychoanalytic theorist (Westkott, 1986) and cultural critic Karen Horney also had a pivotal influence on Fromm's revised psychoanalysis and on his career as a public intellectual. Fromm learned a lot from her while they dated in the United States in the 1930s. Horney's writing that made her books best sellers was arguably the inspiration for Fromm's strategy of producing easily accessible works, such as *Escape from Freedom* (1941), *The Sane Society* (1955), *To Have or to Be?* (1976), that could be published by commercial presses. But what is more important, Fromm owed Horney for her insights into the 'womb' envy theory that she offered as an alternative to the Freudian patriarchal concept of 'penis envy.'

Fromm's success as the major critique of Freudian orthodoxy in the 1940s and 1950s had much to do with his position on the optimal margins of the field and his connection to various collaborative circles and intellectual social movements. Fromm and Horney shared a critique of Freudian orthodoxy, but his connection to the Frankfurt School network made him more independent than Horney, who was essentially a Freudian psychoanalyst who wrote books on the side. As a consequence, Fromm's critiques of Freudian orthodoxy were more direct and perhaps harsher. However, after his relationship with Horney dissolved, she limited his power in Association for the Advancement of Psychoanalysis (APA), which was dependant on the financial support from institutions working in the Freudian orthodox tradition (Friedman, 2013). In short, allowing Fromm a central position in APA could jeopardize Horney's professional project (Friedman, 2013). As a consequence, Fromm was assigned to a place ever more on the fringes of networks among Northern American Freudian orthodox psychoanalysts than he had occupied in the early 1930s.

Fromm's personal life took two big turns when the poor health of his second wife (Henny Gurland) pushed the couple to move to Mexico in 1950 (Friedman, 2013: 139), where, after a few years of fighting depression and chronic pain, Henny committed suicide. A year later life offered a last romantic surprise when Fromm met his third wife (Annis Freeman) and ended up staying in Mexico. His new companion was fascinated with Buddhism, meditation, and international politics resulting (Friedman, 2013: 170) in a happy home in Mexico and their remaining married for 27 years until his death in 1980. The deep affection between Erich and Annis found expression in Fromm's most popular book *The Art of Loving* (1956) as well as in a general increase in Fromm's productivity. Between 1955 and 1968, when he had an almost fatal heart attack, Fromm published many books, including *The Sane Society* (1955), *Marx's Concept of Man* (1961), *May Man Prevail?* (1962), *Beyond the Chains of Illusion* (1962), *The Heart of Man* (1964), and *The Revolution of Hope* (1968).

Over these years in Mexico, Fromm founded the Mexican Institute of Psychoanalysis, trained early generations of psychoanalysts, and became a central figure in Latin American psychoanalysis. Once again, Fromm's marginality in one place was optimal in another because his connections transcended various networks. The freedom that Fromm gained by distancing himself from the New York psychoanalytic establishment combined with the importance of the mother in Catholic Latin American culture, reinforced his revisionist tendencies. Even though, as sociologist Lynn Chancer argued, Fromm's work anticipated new trends in psychoanalytic sociology and emphasized pioneering feminist and existentialist themes beyond libido theory (1992), his Mexican exile meant that he was on the margins and not in the center of transformations that psychoanalysis was undergoing; Fromm's pioneering role in the development of psychoanalytic sociology hardly made him a feminist in contemporary terms. Fromm's writing, filled with exaggerated male and female difference and often marred by the homophobia deeply inscribed with much psychoanalytic theory, limits the usefulness of his gender analysis without significant reformulation. Ironically, nonetheless, the ways that Fromm drew on and developed critiques of orthodox Freudian theory from his very unique marginalized position helped create contemporary psychoanalytic feminism – though also ensuring that Fromm himself would be marginalized in psychoanalysis.

Fromm's critique of Freud had its foundations in the largely ignored works of Bachofen, a nineteenth-century Swiss aristocrat, who anticipated some of Freud's views but whose work analyzing ancient myths suggested that the matriarchal social structure preceded the patriarchal one. Bachofen's *Mother Right – Das Mutterrecht* (1861) inspired Fromm to write an article on matriarchy back in 1934 (Burston, 1991; Fromm, 1934; Fromm and Maccoby, 1970; Ortmeyer, 1998). Later, in *The Anatomy of Human Destructiveness*, Fromm extensively discussed the excavations at Çatal Hüyük (in present day Turkey) and the existence of a Neolithic culture characterized by feminine values (Bacciagaluppi, 2001; Fromm, 1973), making the argument that matriarchal social organization could be seen as a legitimate alternative to patriarchy. Various feminist thinkers, including Mary Daly and Riane Eiasler, would build from Fromm's discussion of matriarchy (Bacciagaluppi, 2001). Fromm thus contributed to the emergence of a set of ideas within radical feminist circles in the 1970s.

Fromm's critique of Freudian orthodoxy did not flow simply from his reading of Bachofen but also emerged from many discussions in the 1930s with the New York City–based Zodiac club, an informal social circle of thinkers that existed on the margins of psychoanalysis and anthropology and psychiatry. This group consisted of such brilliant intellectuals as Horney, Margaret Mead, Ruth Benedict, and American psychiatrist Harry Stack Sullivan. The mental health professionals in the group founded the White Institute in Washington DC, a nonorthodox Freudian center of innovative theorizing, where the position of the director of clinical training was given to Fromm. This new institutional affiliation allowed him to even more openly reject drive theory, since his colleagues were also Freudian revisionists, social psychologists, or neo-Freudians (Greenberg and Mitchell, 1983).

Fromm's revisionist ideas, forged in marginal networks but institutionalized into a revisionist training institute, went on to contribute greatly to the development of object relation theory (Chodorow, 1989; Greenberg and Mitchell, 1983). A version of Fromm's theory of social character and object relations theory is applied to the phenomenon of mothering in *The Reproduction of Mothering* by Nancy Chodorow and developed further in her extensive writings. Chodorow's sociological project was rooted in an attempt to understand how people create a social environment that exclusively recruits mothers as the only ones doing the mothering or caretaking (1978). Chodorow approached this theoretical task by linking psychoanalysis with a feminist sociology of gender while paying attention to finding an empirical basis in clinical evidence, an approach pioneered by Fromm in opposition to the purely philosophical Freud defended by Adorno and Marcuse (Chodorow, 1978: 74; Rickert, 1983).

Fromm did not receive an adequate credit for his role in the development of the rich object relations tradition in psychoanalysis for a variety of reasons. Bachofen's work is problematic in light of contemporary archeological research, and thus Fromm's use of his ideas hurts his scholarly reputation. Fromm's followers among radical feminist circles were highly controversial themselves in academic and intellectual circles, and they were not likely to become followers of Fromm given some of his problematic writings on gender. The personal and political-intellectual differences between the various members of the Zodiac club, and their very different career trajectories, ensured that no coherent schools of thought emerged out of various neo-Freudian networks (McLaughlin, 1998). And some of Fromm's personal sexual politics leave much to be desired (Friedman, 2013).

Public Intellectual and Public Sociologist

Over and above the theoretical contributions Fromm made to psychoanalysis and sociology, his legacy is worth remembering and building on because of his role in creating the very genres of modern 'public intellectual' and 'public sociologist' that have become widely debated in recent decades (Jacoby, 1987; Burawoy, 2005). Even though citations of Fromm in sociology journals declined dramatically from the height of his influence in the 1940s and 1950s, he had an unacknowledged influence on major public sociologists.

Fromm's clear way of communicating with readers was emulated by Fromm's psychoanalytic patient and close friend, Harvard professor, David Riesman, whose *The Lonely Crowd* published in 1950 became the best-selling sociological book of all time (McLaughlin, 2001b). Fromm's *Escape from Freedom* and *Man for Himself* (1947) were a major inspiration for the analysis of 'inner-' and 'other-' directed social characters outlined in Riesman's sociological classic (McLaughlin, 2001b). In many ways Riesman's work improved on Fromm's by providing a political balance, a scholarly style, and a liberal philosophical commitment while avoiding the excessively prophetic tone of some of Fromm's socialist polemic and moralistic writings (Maccoby, 1995). But Fromm's influence on Riesman is significant, and Riesman was arguably the major public sociologist of the twentieth century, at least in America.

There is also considerable similarity between the style and the content of the work of Erich Fromm and Zygmunt Bauman (Davis, 2008). Bauman, a Polish sociologist, now an émigré UK professor at Leeds University, known for theorizing 'liquid modernity,' shares with Fromm roots in humanistic Marxism. Both thinkers discussed the importance of capitalist ideology; both described life in late modernity as marked by individuation, alienation, an emergence of freedom that becomes an 'unbearable psychological burden,' and the disappearance of people's ability to love (Catlaw, 2009); both stressed the negative consequences of obsessive-compulsive consumption and market-oriented personalities (Ziółkowski, 1998). Despite numerous differences (Bauman's multiple worlds in liquid modernity as opposed to Fromm's commoditized and internalized world (Smith, 1999; Best, 2013)), the similarities between their works are undeniable. And if Riesman was the dominant American public sociology in the 1950s, Bauman has been a very influential public sociologist in Europe from the 1990s through the first decade of the twenty-first century.

Fromm's legacy as a public intellectual and public sociologist should not be ignored. As Lawrence Friedman has documented, Fromm played a direct role in President Kennedy's disarmament policy (Friedman, 2013); Fromm was a devoted antiwar activist whose condemnation of the Vietnam War and American imperialistic ambitions was articulated in work and in his testimony before the U.S. Congress, and he was a philanthropist who gave enormous amounts of money to Amnesty International (Friedman, 2013). In short, being faithful to his humanistic views, Fromm considered the survival of the human race as a way to "realize … human solidarity" (Fromm in Wilde, 2000). Fromm's politically engaged work paved the way for other sociological books to appear on the shelves at homes of millions of readers, including policy makers, around the world.

Conclusion

Erich Fromm's legacy can be summarized in two ways. Firstly, his ideas shaped scholarship on the sociology of authoritarianism, race and racism, gender, and historical social psychology. Secondly, through his life and work as a public intellectual and public sociologist he created a template for future generations of engaged intellectuals. Fromm's work had an enormous influence on twentieth-century social science and his legacy remains useful today.

The revival of interest in his theories, activism, and engagement in public affairs cuts across all disciplines and genres of social science writing. His theory of love is being used by education scholars (Wesh in Hay, 2011; Fleming, 2012). Furthermore, Fromm's work has inspired criminologists (Anderson, 1999; Lowy, 2013), peace scholars (Housden, 2013), political theorists (Durkin, 2014), and academics theorizing social work (Rasmussen and Salhani, 2008; Houston, 2010). Political theorists have returned to Fromm's radical humanism, suggesting new ways forward for his political philosophy, perhaps now framed in less prophetic and moralistic terms (Wilde, 2000). Critical theorists seem far less interested in refighting sectarian intellectual battles from the 1930s, as the Frankfurt School retools itself to engage empirical

social science anew with insights that we might place more modestly now in a tradition that Burawoy calls 'critical sociology' (2005). The dominance in sociological debates of Bourdieu's reflexive sociology and concept of habitus calls for a revisiting of Fromm's theoretically complementary concept of social character (Cheliotos, 2011a). And after decades of the dominance of rational choice theory in the social sciences, there is much space to reinsert Fromm's psychoanalytic revision of emotions, albeit now more centrally attuned to gender and race. We live in a time where authoritarianism and violence are breaking out everywhere, and Fromm's early ecologically oriented vision is more relevant than ever (Fromm, 1976). Fromm's legacy is likely to be built on creatively in the twenty-first century in various ways, in the optimal margins of contemporary social science, as was the case in the 1930s, 1940s, 1950s, and 1960s.

See also: Attitudes, Political: Authoritarianism and Tolerance; Critical Theory; Feminist Psychotherapy; Freud, Sigmund (1856–1939); Gender and Women's Studies, Applied Research on; Marx, Karl (1818–83); National Socialism and Fascism; Prejudice and Discrimination; Psychoanalysis in Sociology; Racism, Sociology of; Social Psychology; Women and Psychiatry.

Bibliography

Anderson, K., 1999. Erich Fromm and Critical Criminology: Beyond the Punitive Society. University of Illinois Press.

Bacciagaluppi, M., 2001. Fromm's concern with feminine values. The Journal of the American Academy Psychoanalysis 29 (4), 617–624.

Best, S., 2013. Zygmunt Bauman: Why Good People Do Bad Things. Ashgate, United Kingdom.

Bonns, W., 1984. The Working Class in Weimar Germany. A Psychological and Sociological Study. Berg Publishers, London.

Brunner, J., 1994. Looking into the hears of the workers, or: how Erich Fromm turned critical theory into empirical research. Political Psychology 15 (4), 631–654.

Burawoy, M., 2005. American sociological association presidential address: for public sociology. American Sociological Review 70 (1), 4–28.

Burston, D., 1991. The Legacy of Erich Fromm. Harvard University Press, Cambridge, MA.

Catlaw, T.J., 2009. Kill the king, love your neighbour. Public Administration Quarterly 33, 318–332.

Chancer, L.S., 1992. Sadomasochism in Everyday Life: The Dynamics of Power and Powerlessness. Rutgers University Press, New Brunswick.

Cheliotos, L.K., 2011a. For a Freudo-Marxist critique of social domination: rediscovering Erich Fromm through the mirror of Pierre Bourdieu. Journal of Classical Sociology 11 (4), 438–461.

Cheliotos, L.K., 2011b. Violence and narcissism: a Frommian perspective on destructiveness under authoritarianism. Canadian Journal of Sociology XXXVI (4), 337–360.

Chodorow, N., 1989. Feminism and Psychoanalytic Theory. Yale University Press.

Chodorow, N., 1978. The Reproduction of Mothering: Psychoanalysis and the Sociology of Gender. University of California Press.

Davis, H.B., 2003. Erich Fromm and postmodernism. The Psychoanalytic Review 90 (6), 839–853.

Davis, M., 2008. Freedom and Consumerism: A Critique of Zygmunt Bauman's Sociology. Ashgate, London.

Daly, M., 1985. Beyond God the Father. Towards a Philosophy of Women's Liberation. Beacon Press.

Durkin, K., 2014. The Radical Humanism of Erich Fromm. Palgrave Macmillan.

Fleming, T., 2012. Fromm and Habermas: allies for adult education and democracy. Studies in Philosophy and Education 31, 123–136.

Friedman, L.J., 2013. The Lives of Erich Fromm's Love's Prophet. Columbia University Press, New York.

Frie, R., 2003. Erich Fromm and contemporary psychoanalysis. From modernism to postmodernism. Psychoanalytic Review 90 (6), 855–868.

Frie, R., 2009. Foreword. In: Funk, R. (Ed.), The Clinical Erich Fromm. Personal Accounts and Papers on Therapeutic Technique. Rodopi, Amsterdam.

Fromm, E., 1934. On the theory of mother right and its relevance for social psychology. In: Funk, R. (Ed.), Love, Sexuality, and Matriarchy. International Publishing Corporation (1997), New York, pp. 21–37.

Fromm, E., 1969[1941]. Escape from Freedom. Holt, Reinhart and Winston, New York.

Fromm, E., 1955. The Sane Society. Holt, Reinhart and Winston, New York.

Fromm, E., 1961. May Man Prevail? Doubleday, New York.

Fromm, E., 1964. The Heart of Man. Harper and Row, New York.

Fromm, E., 1973. The Anatomy of Human Destructiveness. Henry Holt, New York.

Fromm, E., Maccoby, M., 1970. Social Character in a Mexican Village, a Socio-psychoanalytic Study. Prentice-Hall, New Jersey.

Fromm, E., 1976. To Have or To Be? Harper & Row Publishing.

Funk, R., 2009. The Clinical Erich Fromm. Personal Accounts and Papers on Therapeutic Technique. Rodopi, Amsterdam.

Greenberg, S., Mitchell, S., 1983. Object Relations in Psychoanalytic Theory. Cambridge University Press, Cambridge.

Hiel, A.V., Cornelis, I., Roets, A., 2010. To have or to be? A comparison of materialism-based theories and self-determination theory as explanatory frameworks of prejudice. Journal of Personality 783, 1037–1070.

Hinitz, B.F., 2004. Step-by-step guidance for learning and teaching peace. Theory and Research in Social Education, Winter, 123–126.

Houston, S., 2010. Beyond homo economicus: recognition, self-realization and social work. British Journal of Social Work 40, 841–857.

Housden, M., 2013. Psychoanalysis and peace: Erich Fromm on history, politics and the nation. In: Nationalism and the Body Politic. Karnac Books Ltd, London.

Jacoby, R., 1987. The Last Intellectuals. American Culture in the Age of Academe. Basic Books.

Löwy, M., 2013. Erich Fromm: from Messianic Utopia to critical criminology. In: Malloch, M., Munro, B. (Eds.), Crime, Critique and Utopia. Palgrave Macmillan, London, pp. 62–71.

Maccoby, M., 1995. The two voices of Erich Fromm: the prophetic and the analytic. Society 32, 72–82.

McLaughlin, N., 1996. Nazism, nationalism and the sociology of emotions: *escape from freedom* revisited. Sociological Theory 14 (3), 241–261.

McLaughlin, N., 1998. How to become a forgotten intellectual: intellectual movements and the rise and fall of Erich Fromm. Sociological Forum 13, 215–246.

McLaughlin, N., 1999. Origin myths in the social sciences: Fromm, the frankfurt school, and the emergence of critical theory. Canadian Journal of Sociology 24 (1), 109–139.

McLaughlin, N., 2001a. Optimal marginality: innovation and orthodoxy in Fromm's revision of psychoanalysis. The Sociological Quarterly 42 (2), 271–288.

McLaughlin, N., 2001b. Critical theory meets america: Riesman, Fromm, and the Lonely Crowd. American Sociologist 2 (1), 5–26.

Ortmeyer, D.H., 1998. Revisiting Erich Fromm. International Forum of Psychoanalysis 7, 25–33.

Pearson, A.R., Dovidio, J.F., Samuel, L.G., 2009. The nature of contemporary prejudice: insights from aversive racism. Social and Personality Psychology Compass 3 (3), 1–25.

Rasmussen, B., Salhani, D., 2008. Resurrecting Fromm. Smith College Studies in Social Work 78 (2–3), 201–226.

Ralston, R.J., 1983. The Fromm-Marcuse Debate Revisited. University of Texas, Austin.

Roazen, P., 1977. A stranger to narrow fashion. The Nation 5 (2), 151–154.

Shils, E.A., 1954. Authoritarianism: right and left. In: Christie, R., Jahoda, M. (Eds.), Studies in the Scope and Method of "The Authoritarian Personality", pp. 24–49.

Smith, D., 1999. Zygmunt Bauman: Prophet of Postmodernity. Polity Press, London.

Wesch, M., 2011. The art of loving and learning. In: Hay, Iain (Ed.), Learning with the World's Great University Teachers. Open University Press, Maiden, UK, pp. 23–29.

Westkott, M., 1986. The Feminist Legacy of Karen Horney. Yale University Press.

Wheatland, T., 2009. Frankfurt School in Exile. University of Minnesota Press.

Wheeler, M.E., Fiske, S.T., 2006. Controlling racial prejudice: social-cognitive goals affect amygdala and stereotypes activation. Psychological Science 16 (1), 56–63.

Wiggerhaus, R., 1986. The Frankfurt School: Its History, Theories and Political Significance. Polity Press, Cambridge.

Wilde, L., 2000. In search of solidarity: the ethical politics of Erich Fromm (1900–1980). Contemporary Politics 6 (1), 37–54.

Ziółkowski, M., 1998. On the diversity of the present: suspended between tradition, the legacy of socialism, modernity and postmodernity author. Polish Sociological Review 121, 21–43.

Frontiers in History

Alfred J Rieber, The Central European University, Budapest, Hungary

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Studies on the frontier in history have undergone important methodological and thematic changes over the past two decades, becoming more interdisciplinary and reinterpreting the main features of frontiers in the fields of North American, Russian, Ottoman, and Chinese history. This article seeks to illustrate the themes common to all these fields and to examine separately each one emphasizing the revisionist interpretations and presenting a comparative dimension where appropriate.

The Changing Context

Three shifts in perception over the past two decades have stimulated new conceptual and interpretive approaches to the study of frontiers in history. The first was driven by the spatial turn, the second by changes in global geopolitics, and the third by the expansion of scholarship in non-European areas of specialization. The spatial turn reflected the qualitative and quantitative increase in the borrowing and adaptation of the social sciences by historians, as well as a greater awareness of historical processes in fields such as geography and anthropology (Osterhammel, 1998). The interaction has not always been an equal exchange. Witness the terminological and conceptual difference that has emerged between frontiers and borders. Although the distinction between the two is often blurred in the literature, historians prefer the former and geographers, sociologists, and anthropologists prefer the latter. In a recent compendium on borders, only one of the 40 contributors was a historian (Wastl-Walter, 2010). For most historians, a frontier signifies a shifting and contested intermediate zone of conflict and exchange on the periphery of a territorial core with an ethnically mixed population, exposed to internal and external change over a long duration and engaged in mutual interactions with the institutions and cultural practices of the core. Border scholars share some of these views, but they treat borders as a clearly defined line separating organized spatial units while reserving the concept of borderland for similar ethnocultural areas on opposite sides of state boundaries. Their consideration of borders tends to be chronologically limited to the twentieth and twenty-first centuries, referring to earlier periods by analyzing collective memory. An attempt to reconcile these differences argues that "a basic distinction between frontiers which are zones of settlement and frontiers which constitute political barriers is apparent in most frontier historiography" (Power and Standen, 1999). The second impetus to the growing number of frontier studies came with the dissolution of the Soviet Union, the redrawing of Europe's borders, the deterritorializing effects of globalism, and the reterritorializing of state boundaries after 9/11, all factors in the displacement of familiar spaces. The third impetus came from regional specialists in the historical profession. They have made more extensive use of the comparative approach and revived interest in the role of frontiers in the processes of state formation and ethnic identification. The proliferation of these studies, challenging established theories, has been particularly significant in North American, Ottoman, Russian, and Chinese histories.

Among the most recurrent themes in recent histories of frontiers four stand out: the tension between hybridity, transfer, and exchange on the one hand and military, sanitary, and cultural barriers on the other hand; the reciprocal influences of the frontier (periphery) and the centers of power (core) on constructing and deconstructing imperial state systems; the multiplicity of both external and internal frontiers; and symbolic or discursive as opposed to geopolitical frontiers.

Frontiers and the Social Sciences

Anthropologists came later than historians to frontier studies but then arrived in considerable force. In the 1950s, they began to engage in extensive field work on the margins of sedentary and nomadic societies. Their insights were rapidly absorbed by historians who had already been seriously exploring frontiers for over half a century. Geographers based in local history and influenced by cultural studies began to redefine the ideas of space and frontiers in terms of linguistic and social contexts (Godlewka and Smith, 1994; Driver and Samuel, 1995; and Lewis and Wigen, 2000). The so-called textual approach appealed strongly to non-European postcolonialist scholars who coined the term 'textualizing the world,' meaning the mental construction of the globe and its discursive subdivisions to fit the European vision (Spivak, 1990). Both sociologists and anthropologists explored ways in which ethnic identity and concepts of citizenship corresponded to the drawing of territorial boundaries (Brubaker, 1992). Semioticians perceived frontiers as 'zones of cultural bilingualism' (Lotman and Uspenskii, 1984, 3–15). Despite changing contexts and multidisciplinary approaches, most American and comparative historians of the frontier keep returning to the original work of Frederick Jackson Turner, if only to refute him. Only a few comparative studies have sought to develop systematic frameworks for analysis that have not taken the Turner thesis as a point of departure or a foil for revision (Rieber, 2004).

The Turner Thesis and Its Critics

It is well over 100 years ago that Frederick Jackson Turner published his now world famous influential essay on the frontier in

American history (Turner, 1920, 1893). For American historiography, the great explanatory power of his vision resided in its claim to explain the uniqueness of American civilization. Since then his concept of the frontier has undergone extensive revision, which in the middle years of the twentieth century reached the point where a body of scholarship banished the word frontiers from historical discourse. However, the complexity of his approach guaranteed its longevity and survival even as a foil. Multidisciplinary in spirit if not in word, he combined elements of economic thought, geography, social history, and the history of ideas even though he was not always consistent in establishing the relationship among them. His brilliant essays added up to something less than a full-blown theory and left the way open for multiple interpretations, extensive revision, and critical refutation. Like many great theses in historiography, its importance lies as much in the literature it generated as its original propositions. Given the large number of American historians and the academic culture of methodological restlessness in the United States, it is not surprising that the main changes in conceptualizing the history of frontiers has been dominated by revisions of Turner's original thesis (Adelman and Aron, 1999).

Revisionism in the United States

In American historiography, the revisions and reformulations of the Turner thesis radiate around two conceptual poles: new methodologies, particularly anthropology and environmental studies, and fresh interpretive perspectives, reflecting the shifting concerns of a society in search of itself. The task of charting the currents of the debate has itself become a minor cottage industry (Nash, 1991). Turner's frontier as an advancing line of settlement has given way, in the work of the regionalists, to a multiplicity of locations. However, Turner's tendency to conflate the frontier as a line and 'the Great West' as a region continues to be a source of confusion. There has also been a trend in American historiography away from spatial to cultural definitions of the frontier, although Turner himself made room for both place and process in his own work.

Turner's concept of the frontier as dividing 'savagery from civilization' or alternatively as a forward movement into 'empty lands' was challenged, under the influence of anthropologists, by the idea of the frontier as a 'middle ground' of intermingling of Native Americans and European settlers engaged in a variety of cultural and commercial 'frontier exchanges' (White, 1991). Turner's vision of the West as an environment that bred rugged individualism and privileged self-reliance also encountered criticism from historians and sociologists who gave equal prominence to social collectivities. Critics also uncovered a darker side of the frontiersman in the legacy of a gun culture and the persistent cult of violence (Graham and Gurr, 1969).

Environmentalist historians exploded Turner's arcadian picture of natural harmony on the frontier by exposing the depredations, manipulations, and 'species shifting' impact of the Western advance (Cronon et al., 1992). Turner's insistence on the defining role of the frontier experience in building American democratic institutions failed to account for the contributions of urban life and the antecedent values that the pioneers brought with them from the eastern seaboard or Europe. At one end of the historiographical spectrum, Turner's thesis lies in ruins. At the other end, attempts have been made to synthesize the newer ideas on race, class and gender, and region or to revive older ideas of conflict and conquest in order to rescue once again the place of the frontier in American history (Nobles, 1997; Wooster, 2009). However, a multicultural interpretation of the frontier requires a radical reversal of the original myth. Not only does the advancing frontier signify the virtual extermination of the Native American population, but it also takes on a plurality of meanings for other ethnic groups. Their frontiers are formed and advanced after the official closing of the frontier in 1892 that caused Turner to bemoan the loss of what he regarded as a unique part of American life. In the twentieth century, the territorial and social mobility of the Hispanic-, Afro-, and Asian-Americans constituted in their eyes a formative experience that *mutatis mutandis* enthralled the pioneers of the nineteenth century (Takaki, 1993).

The commercialization of American history found the perfect vehicle in the myth of the frontier. Frontier images penetrated the world of mass entertainment, advertising, and journalism. The 'Western' film as a genre enjoys enduring popularity because it is able to incorporate changes in social values while retaining the centrality of the myth. New social and environmental concerns have left deep marks on the Turnerian stereotypes of the Old West. In films and television, the heroic cowboy in the white hat, the intrepid pioneer fighting off the Indians, and the US cavalry riding to the rescue at the last moment have become more ambiguous figures. Black cowboys and troopers, environmentally sensitive Native Americans, and even women gunslingers have challenged the dominant white male image.

The official closure of the land frontier in 1892 did not deter Turner from speculating on the further advance of an American oceanic frontier that would extend to the shores of the Asian mainland. In this quest for new frontiers to conquer, he was joined by several other influential American publicists who were even more specific in their advocacy of an expansionist American empire overseas. Their ideas fit in well with a cluster of ideas in the 1890s that merged geopolitics, social Darwinism, the Anglo-American mission or Manifest Destiny, and the Open Door Policy, all with strong anti-Russian implications. They influenced the world view of Woodrow Wilson and later politicians. The use of the word frontiers became so attenuated as to lose much of its original meaning. Surfacing in different contexts, it took its most politicized form in countering the perceived threat of postwar Soviet expansion of its frontiers.

Revisionism in the British Commonwealth

The Turner thesis had implications beyond American history insofar as he established the frontier as a historical problem. Paradoxically, it also became a reference point for scholars seeking to adopt or refute its implications for a whole series of different societies that shared, however, similar features of a frontier of moving settlements. The applicability of the Turner thesis to other British colonial experiences leading to the establishment of white settler communities has given rise to discussion over the degree to which the American frontier was unique. There were certain common features of the frontier experience and its incorporation into the national narrative that were shared

by Canada, Australia, and New Zealand, even to a lesser extent South Africa: the belief in a manifest destiny, the struggle to tame the wilderness, the origins of a democratic society. Some comparative studies of the frontier experience in Canada, Australia, and New Zealand with Turnerian overtones argue that despite superficial similarities with the United States, geographical factors and the prolonged tutorship of the metropolitan center in Great Britain discouraged the mythology of frontier exceptionalism. In Canada, there were close parallels between the farming frontiers in Ontario and Saskatchewan and in North Dakota and Montana. However, the expansion toward west was delayed by the physical barrier of the Canadian Shield. When it came it was promoted by Royal Governors who established a national police (the Royal Mounties) before mass immigration and who negotiated settlements with the Blackfoot tribes thus heading off the local violence and white–Native American wars that characterized the American west. To an even greater extent, the geography of Australia, dominated by a vast arid region in the center, 'a hinterland but no heartland,' relegated the bulk of the population to the cities on the east and south coast while royal land policy did not favor the small farmer in competition with the large sheep ranges. Even here there are regional similarities between the semiarid fringes of California and Victoria. Australia had its frontier folk hero in the form of the bushranger, although he was connected to the convict life rather than the freebooter born in an open society.

Some Canadian historians interpreted the differences from the United States in order to propose their own myth of a 'northern vision' to set against Turner's 'West' as a foundation block of Canadian national character. The thesis drew heavily on British and continental writers and social Darwinists who romanticized northern peoples as hardy, industrious, and democratic. They combined this version of geographical transcendence with the 'staple thesis' that tied Canada's economy to the global metropolitan community through the production of raw materials, thus departing from the American concept that separated its West from the Old World. However, the combination of the northern vision and the stables thesis failed to generate the same mythic power as the American frontier (Katerberg, 2003).

Recent research has forced a reevaluation of Turner's preoccupation with the experience of Anglophone settlers and a reorientation of his East to West movement by portraying the role played by French merchant families along the Creole corridor stretching from the Great Lakes through the Mississippi Valley to the Gulf of Mexico. Instead of Turner's frontiersman as a rugged individual, the legendary mountain man, fur trader, and Indian hunter, an alternative portrait emerges of an urban, cosmopolitan, 'bourgeois' who negotiated rather than drove the Indians out of their lands even after the French lost Canada (Gitlin, 2010). Like the Hispanic frontier of the southwest United States, the Creole frontier adds to the complexity and multiplicity of American frontiers.

Russian Frontiers

As an interpretive point of departure, historians of the Russian frontiers have been most strongly influenced by the frontier thesis proposed by Turner and the subsequent revisions. For the classic Russian historians S.M. Solov'ev, V.O. Kliuchevsky, and M.K. Liubavskii, the latitudinal zones of forest and steppe provided the natural setting, facilitating the advance of the frontier through colonization and conquest. Although they did not develop a frontier thesis, they drew on many of the same intellectual foundations as Turner (Bassin, 1993). By contrast, however, their interpretations emphasized the negative effects: The draining of resources from the center, the unfavorable conditions of climate and soil, and the threat of nomadic incursions. Historians in the early Soviet period added to this somber list the exploitation of conquered people, exemplified by M.N. Pokrovskii's well-known image of the empire as a 'prison of nations.' A strong reversal set in only after World War II when the new state school of Soviet historians extolled the peasant colonizer as a heroic figure and revived the prerevolutionary term *sblizhenie* (drawing together in friendship) to signify the civilizing mission of the Russian people. A third turn has been taken only recently by the emerging regional (Siberian) school of historians, who, in readjusting their focus to the specific features and values of societies on the periphery at odds with those of the center, have nonetheless also acknowledged a debt to Turner. In the most recent general reassessment, Boris Mironov has interpreted colonization as primarily a demographic phenomena. In his view, Russia's territorial expansion was the optimal solution for a society burdened by over population, poor in capital but rich in human resources and land available for colonization. In making comparisons with the United States, he used the Turner thesis and its critics to emphasize the differences and the greater benefits brought to American civilization by frontier expansion (Mironov, 1999; Mironov and Ekloff, 2000).

Since the 1980s, a new generation of American and British specialists in Russian colonization has been incorporating fresh insights borrowed in part from revisionist American historians of the western United States and in part from anthropology and cultural studies while mostly avoiding direct comparisons with the American experience. The emphasis has shifted from conquest and resistance on the frontier to cultural interaction, an interrogation of geographical visions, multiple locus, and images of frontier utopias. The multiplicity of Russia's frontiers has inspired scholars to propose functional categories such as military, extractive, and colonized (Kappeler, 2003). The rejection of Turner's idea of 'empty lands' and dismissal of the Native American experience as factors in forming the frontier have their parallel in the restoration of the steppe people in the history of the Russian frontier. The view of the frontier as a locus of exchange and hybridity has eroded the traditional emphasis on conflict and hostility (Breyfogle, 2005). White's concept of 'the middle ground' has been employed to analyze the dynamic interaction between the state and the migrating peasants in settling the frontier (Moon, 1997).

The closing of the Russian frontiers was no sooner completed with the delimitation of the Afghan boundary in 1885 than the Bolshevik Revolution burst them open again. The subsequent Civil War confronted the early Bolshevik leaders with two irreconcilable problems. Externally, they were committed to abolishing frontiers, but practical necessity dictated establishing new interstate boundaries on ethnically

mixed frontiers. Internally, they were committed to creating boundaries for newly created national units on equally fluid frontiers. The attempt to reconcile these inherent contradictions ultimately failed. The external expansion of the frontiers through the medium of popular democracies in a protective belt around the Soviet homeland began to break down almost immediately after World War II with the Yugoslav case and was then repeatedly challenged by revolts in Hungary and Czechoslovakia; disaffection in Poland; and defections by Albania, Romania, and China. The internal consolidation of national Soviet republics spurred centrifugal forces that together with the external centripetal forces destroyed the entire system, opening up new, multiple, and contested frontiers in the twenty-first century.

Chinese Frontiers

The pioneering figure in Chinese frontier studies was Owen Lattimore. His central thesis was to establish that frontiers as distinct from boundaries were the margins of socioeconomic systems defined by their 'optimal limit of growth.' He was the first to stress the dynamics of frontier exchange in Chinese nomadic relations. The interaction of the nomadic and sedentary societies produced a unique form of frontier culture, which he called 'frontier feudalism.' The key to the system was the shift from a clan to a territorial-based organization of the nomads on the margins. This was the result of a Chinese policy of creating a patron–client relationship with the frontier nomads. While admitting that Frederick Jackson Turner was an acute observer, Lattimore added, "what he saw so clearly he saw standing on his head. In large measure, when he thought he saw what the frontier had done to society, he was really seeing what society did to the frontier" (Lattimore, 1962).

For Lattimore, Manchuria occupied a unique niche on China's frontiers because of its inner facing frontier. Under the Qing (Manchu) dynasty, the population beyond the Great Wall was composed of tribal elements "who remained outside of the conquered territory but were identified with the alien dynasty within the Wall." In periods of alternating barbarian and Chinese ascendancy, Manchuria served as a reservoir of officials and troops. The indigenous population and the colonists from the south looked back to China rather than forward to settling new territories (Lattimore, 1940, 1962). The Manchu policy of restricting immigration into their homeland broke down in the late nineteenth century, and a process of sinicization gradually erased the ethnic frontier between the Han Chinese core and Manchuria.

On the frontier with the Mongol tribes, the intermingling of the Chinese and the steppe peoples shaped an ethnically and economically diverse ecology. Here the expansion of the Qing and Romanov Empires separated the Mongol tribes into three parts, those living under Chinese rule in Inner Mongolia, under Russian rule in Buryat Mongolia, and in a state of independence in Outer Mongolia. However, all continue to cherish memories of unity under the legendary rule of Chingghis Khan.

Beginning in the 1980s, Western scholars began to move away from the focus on the maritime frontier and follow up on Lattimore by arguing that the interaction of the sedentary Han Chinese population with the nomads along the Inner Asian frontier established precedents for subsequent dealings with the Western colonial powers on China's coastal frontier. The process was guided by bargaining for mutual benefit. Its main features were the regulation of trade, religious toleration, and the establishment of a tribute system (Fletcher, 1986; Barfield, 1989; Jagchid and Symonds, 1989). From these studies, the nomads emerge as more dependent and hence more committed than the imperial power to the maintenance of an exchange culture on the frontiers. They clearly preferred trading to raiding. As long as the nomads accepted Chinese cultural superiority and tributary status peace was assured. Throughout the eighteenth century, the Qing dynasty pursued an active frontier policy in the mixed Muslim and Turkic populations of the northwest (Gansu and Xinjiang). Preoccupied by the Inner Asian frontiers, which were traditionally regarded as posing the greater threat, its leaders failed to judge accurately the new challenge on its maritime frontiers beginning in the early nineteenth century. Belatedly, they responded by employing the same tactics of conciliation with the Western maritime powers that they employed in Inner Asian frontiers with disastrous results. From the Opium War in the 1840s to the Boxer Rebellion in 1900, the Chinese steadily lost control over their maritime frontier to the Western naval powers and Japan. By the end of the century, they confronted the prospect of also losing control over the Inner Asian frontiers to the Japanese or the Russians.

In the last decade of the Qing dynasty, Chinese officials in Manchuria strengthened their military defense of the frontier. They invoked the time-honored strategy of using barbarians against barbarians. This time it was not the nomads but the great powers who assumed the role of barbarians. The Chinese sought to balance the Russians and Japanese with the United States as a part of their frontier defense in Manchuria. The effort failed because the United States was unwilling to transform its Open Door Policy from an abstraction to an active participation in economic development backed up by a political commitment. Under similar circumstances, the Nationalist Government of Chiang Kai-shek attempted to exploit the renewed American endorsement of the Open Door in its defense of the Manchurian frontier against the Japanese in the 1930s and the Russians in 1945 with equal lack of success.

The next important stage in the reconceptualization of the Chinese frontier policy came in the 1990s from a group of scholars who insisted that the Qing was an Inner Asian empire rather than a Chinese dynasty. They argued that a Manchu ethnic identity did not diminish but grew stronger throughout the nineteenth century as the new ruling elite maintained the frontier between their homeland and China. For them sinicization was based on the mistaken view of the Han people as a homogenous ethnic group (Crossley, 1997, 1999; Rawski, 1998; Milward, 1998). Their revisionist view of the frontier in Chinese history has attached new and unprecedented importance to the complex interaction between the core provinces of China and the Inner Asian frontiers. This perspective has stimulated comparisons between the Chinese, Russian, and Ottoman frontiers (Perdue, 2005). The Chinese Revolution and Civil War ending in 1949 revived the process of sinicizing the frontier provinces of Tibet, Xinjiang, Inner Mongolia, and Manchuria and created a new frontier along the state borders with the Russian Republic.

Ottoman Frontiers

The Ottoman concept of frontiers emerged from the fusion of three traditions that shaped its self-identification as a state: the nomadic warrior, the Islamic religious, and the Byzantine imperial. The founders of the Ottoman frontier thesis, Paul Wittek and Mehmet Fuad Köprülü, made a distinction in the early period of Ottoman expansion between the core and frontier in terms of social structure and cultural characteristics. Köprülü, influenced by the Annales School, took a broader view of the frontier embracing its distinctive religious, legal, economic, and artistic institutions. Wittek stressed the gazi warrior milieu rooted in Islamic religious zeal. By the thirteenth century, he argued, warrior cultures appeared on both sides of the porous Turkic-Byzantine frontier. They were originally composed of Islamic *gazis*, and Greek *akritai* increasingly replaced by Turkmen tribesmen recruited from the other side. In this intermediate zone, war and trade often alternated in a pattern similar to that on Roman and Chinese frontiers and facilitated the penetration and conquest of the Byzantine Empire by the Ottoman Turks (Wittek, 1938).

The concept of the early Ottoman Empire as a gazi state, one emanating from an Islamic warrior culture, has run into as much scholarly opposition as the Turner thesis in the United States. It is now clear that gazi meant different things to different people as reflected in the frontier narratives and subsequent Islamic religious texts. Over the centuries the various interests – rulers, border warriors, and ulema – vigorously promoted their own views of what gazi activity meant. The current view is that like so many terms employed in the attempt to construct a founding myth gazi has always been highly contested. Recent interpretations substitute for the 'Ghazi Thesis' an Islamo-Christian syncretism and expand the time frame, broaden the context to include the larger Islamic world, reexamine the sources critically, and factor in the spatial distance of the frontier from the distant central state (Darling, 2000; Lowery, 2003). One thing is certain: no longer can the early Ottoman frontiers be equated with the idea of jihad or holy war. However, there is no denying that it represented the frontier spirit in Ottoman state building.

Jihad, a term with both military and spiritual aspects derived from the precepts of Islam, was employed by the Ottoman ruling elite to represent the division of the world into two cultural spheres, *dār ul-Islām*, the abode of Islam, and the *dār ul-harb*, the abode of war. Between them lay a frontier zone a contested territory where warriors fought the just war consecrated by Islam. This provided the ruling elites with a justification for the expansion of the frontiers in all directions, but this rigid duality could not be strictly maintained. The Ottoman rulers created frontier marches (*uc*) under the leadership of frontier lords who enjoyed considerable autonomy. In return, they were obliged to furnish armed men, both Muslims and Christians, as frontier troops. The image of the Islamic warrior tradition eroded over the following centuries, bringing change to the concept of frontiers.

The expansion of the Ottoman frontiers was an almost a continuous process from the fourteenth century to the late seventeenth century. The Sultans adopted pragmatic and flexible strategies to deal with the great variety of frontiers, particularly those in Hungary, Kurdistan, and Eastern Anatolia. In the first case, joint administration by Hungarian and Ottoman officials did not check the border raiding by the frontier troops on both sides, often joined by hayduks. The peasantry formed local self-defense forces called *paraszt vármegye* (a peasant county) by the Hungarian and *zapis* ('the enlistment') by the Slavic population, but they too often behaved more like insurrectionists than forces of order. Similar experiments were tried on other frontiers, often certified by the Ottoman authorities. At the height of its power from the fourteenth century to the end of the seventeenth century, Ottoman rule preserved and in some cases expanded the role of nomadic tribes, creating 'a moveable empire' (Kasaba, 2009). The nomadic tribes performed important functions throughout the empire in providing networks of trade and communications but were especially valued on the expanding frontiers. They were used to occupy conquered areas where the political structures were weak and local communities disrupted or dispersed. They operated as a powerful military force in the early centuries when cavalry was the dominant branch of armies. In the fourteenth century in the Balkans, the spontaneous migration of 10 000 nomads linking up with Vlachs and Albanians prepared the way for subsequent conquest by the regular army. The rulers devised special rules for regulating tribal affairs. They classified the tribes and appointed tribal officials to administer and collect taxes. They governed relations between the nomads and sedentary population and protected the migratory routes. The end of the period of rapid expansion and the first big loss of territories following the long war with the Hapsburgs in 1699 began the process of restricting the movement of nomads and attempting to settle them on vacant or underpopulated land. This coincided with a reevaluation of the symbolic frontier that in theory separated the Islamic Ottoman polity from its Christian neighbors. The closing of the frontier in 1699 was followed by an almost steady retreat of the frontiers as the result of wars with the Habsburg and Russian Empires and the rebellions of the orthodox population in the Balkans. By the early twentieth century, the Ottoman Empire had become a largely Muslim state with frontiers in Europe reduced to a small strip of territory shielding Istanbul. The collapse of the Empire meant a further shrinking of the frontiers in Asia and the reconstitution of the state boundaries along more generally Turkish national lines, with the important exception of the Kurds whose demands for autonomy kept the eastern frontiers fluid.

See also: Colonization and Colonialism, History of; Comparative History; Global History: Universal and World; Nomads and Nomadism in History.

Bibliography

Adelman, J., Aron, S., 1999. From borderlands to borders: empires, nation-states and the peoples in between in North American history. American Historical Review 3, 814–841.
Barfield, T., 1989. The Perilous Frontier. Harvard University Press, Cambridge, MA.
Bassin, M., 1993. Turner, Solov'ev and the "frontier hypothesis": the significance of open spaces. Journal of Modern History 65, 473–511.
Breyfogle, N., 2005. Heretics and Colonizers: Forging Russia's Empire in the South Caucasus. Cornell University Press, Ithaca, NY.
Brubaker, R., 1992. Citizenship and Nationhood in France and Germany. Cambridge, University Press, Cambridge, MA.

Cronon, W., Miles, G., Gitlin, J., 1992. Under an Open Sky: Rethinking America's Western Past. Norton, New York.
Crossley, P.K., 1997. The Manchus. Oxford University Press, Oxford.
Crossley, P.K., 1999. A Translucent Mirror: History and Identity in Qing Imperial Ideology. University of California Press, Berkeley, CA.
Darling, L.T., 2000. Contested territory: Ottoman holy war in comparative context. Studia Islamica 91, 133–163.
Driver, F., Samuel, R., 1995. Rethinking the idea of place. History Workshop Journal 39, v–viii.
Fletcher, J., 1986. The Mongols: ecological and social perspectives. Harvard Journal of Asiatic Studies 6, 11–50.
Gitlin, J., 2010. The Bourgeois Frontier: French Towns, French Traders and American Expansion. Yale University Press, New Haven, CT.
Godlewka, A., Smith, N., 1994. Geography and Empire. Oxford University Press, Oxford.
Graham, H.D., Gurr, T.R., 1969. Violence in America: Historical and Comparative Perspectives. F.A. Praeger, Washington, DC.
Jagchid, S., Symonds, V.J., 1989. Peace and Trade along the Great Wall: Nomadic-Chinese Interaction through Two Millennia. Indiana University Press, Bloomington, IN.
Kappeler, A., 2003. Iuzhnyi i vostochnyi frontier Rossii v XVI-XVIII vekakh. Ab Imperio 1, 47–64.
Kasaba, R., 2009. A Moveable Empire. Ottoman Nomads, Migrants and Refugees. University of Washington Press, Seattle, WA.
Katerberg, W.H., 2003. A northern vision: frontiers and the west in Canadian and American imagination. The American Review of Canadian Studies 33, 543–563.
Lattimore, O., 1940 [1962]. Inner Asian Frontiers. Beacon Press, Boston, MA.
Lattimore, O., 1962. Studies in Frontier History: Collected Papers. Oxford University Press, New York/London, p. 257.
Lewis, M., Wigen, K.E., 2000. The Myth of Continents: A Critique of Metageography. University of California Press, Berkeley, MA.
Lotman, J.M., Uspenskii, B.A., 1984. The Semiotics of Russian Culture. University of Michigan Press, Ann Arbor, MI.
Lowry, H.W., 2003. The Nature of the Early Ottoman State. State University of New York, Press, Albany, 142–143.
Millward, J.A., 1998. Beyond the Pass: Economy, Ethnicity and Empire in Qing Trans Caspia, 1759–1864. Stanford University Press, Stanford, CA.
Mironov, B.N., 1999. Sotsial'naia istoriia Rossii. Period Imperii (XVIII-nachalo XX V.) Genezis lichnosti, demokraticheskoi sem'i, grazhdanskogo obshchesvta i pravovogo gosudarstva, vols 2. St. Petersburg. 1999.
Mironov, B.N., Ekloff, B.A., 2000. Social History of Imperial Russia. Westview Press, Boulder, CO, 1700–1917.
Moon, D., 1997. Peasant migration and the settlement of Russia's frontiers, 1550–1897. The Historical Journal 40 (4), 859–893.
Nash, G.D., 1991. Creating the West: Historical Interpretations, 1880–1990. University of New Mexico Press, Albuquerque, NM.
Nobles, G.H., 1997. American Frontiers: Cultural Encounters and Continental Conquest. Hill and Wang, New York.
Osterhammel, J., 1998. Die Widerkehr des Raumes: Geopolitik, Geohistorie und historische Geographie. Neue Politische Literatur 43, 374–397.
Perdue, P.C., 2005. China Marches West: The Qing Conquest of Central Eurasia. Harvard University Press, Cambridge, MA.
Power, D., Standen, N. (Eds.), 1999. Frontiers in Question. Eurasian Borderlands, 700–1700. Macmillan, London.
Rawski, E.S., 1998. The Last Emperors: A Social History of Qing Institutions. University of California Press, Berkeley, CA.
Rieber, A.J., 2004. The comparative ecology of complex frontiers. In: Miller, Alexei, Rieber, Alfred J. (Eds.), Imperial Rule. Central European University Press, Budapest.
Spivak, G.C., 1990. The Post-Colonial Critic: Interviews, Strategies, Dialogues. Routledge, London.
Takaki, R., 1993. A Different Mirror: A History of Multicultural America. Little, Brown and Co, Boston, MA.
Turner, F.J., 1920 [1893]. The Significance of the Frontier in American History. The Frontier in American History. Henry Holt, New York.
Wastl-Walter, D. (Ed.), 2010. The Ashgate Research Companion to Border Studies. Ashgate, Farnham Surrey/Burlington, VT.
White, R., 1991. The Middle Ground: Indians, Empires and Republics in the Great Lakes Region, 1650–1815. Cambridge University Press, Cambridge, MA.
Wittek, P., 1938. The Rise of the Ottoman Empire. Royal Asiatic Society, London.
Wooster, R., 2009. The American Military Frontiers: The United States Army in the West, 1783–1900. University of New Mexico Press, Albuquerque, NM.

Functional Approaches to Grammar

William Croft, University of New Mexico, Albuquerque, NM, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W. Croft, volume 9, pp. 6323–6330, © 2001, Elsevier Ltd.

Abstract

The term 'functional' or 'functionalist' has been applied to any approach to grammar that puts the primary emphasis on explanation in terms of factors outside the formal structure of language. The four most significant functional approaches can be characterized in terms of the types of functional explanations that are put forward and the position taken vis-à-vis formalist explanations of grammar. More recently, a dynamic, usage-based approach to the functionalist analysis of grammar has been developed.

Introduction

The term *functional* or *functionalist* has a meaning in linguistics distinct from the use of the term found in anthropology, sociology, or philosophy. In linguistics, a functional analysis is an analysis of linguistic structure that puts the primary emphasis on explanation in terms of factors outside the structural form, such as meaning or the function of language in discourse. Functional analysis is in contrast to *formal* analysis, which is an analysis of linguistic structure that puts the primary emphasis of explanation on principles defined purely in terms of structural form. The term *functionalist* has been applied to any school of grammatical analysis which argues for the primacy or centrality of functional explanations over formal explanations.

Such a broad characterization of functionalism in linguistics, of course, requires further elaboration. We begin by illustrating the commonest types of functional explanations, and then turn to different schools of linguistic theory that call themselves functionalist.

Types of Functional Explanations

There are three types of functional explanations that usually are distinguished in the functionalist linguistic literature: processing, cognitive, and discourse or information structure explanations.

Processing Explanations

Processing explanations make reference to models of the production and comprehension of utterances. A processing explanation argues that some aspect of grammatical structure is due to limitations on the human ability to produce or comprehend the structure.

Processing explanations typically are employed to account for two sorts of linguistic facts. One is the relative frequency of a particular grammatical structure across the world's languages. For example, it has been observed that suffixes are on the whole more frequent than prefixes (or more complex alterations of word forms) in the world's languages. A processing explanation has been proposed for the suffixing preference. It is hypothesized that speakers process word stems before their inflections. Psycholinguistic experiments imply that the beginning of a word is most salient, hence the word stem should ideally occupy the initial position; the ends of words are the next most salient, hence inflections should ideally occupy the final position.

The second sort of phenomenon for which processing explanations are employed is to account for degrees of acceptability of a construction in a single language. For example, in English, subordinate clauses embedded in the middle of a clause are relatively less acceptable than subordinate clauses embedded at the end of a clause: compare *I know the boy that owned the cat that caught your parakeet* to *The parakeet that the cat that the boy I know caught died*. This fact has been explained as a consequence of the difficulty in interpreting the latter sentence's center-embedded clauses. In fact, it is grammatical in English to move a relative clause from the middle of a clause to the end under certain circumstances: *A woman contacted me who I had shared an apartment with during my last year at college*.

Processing explanations have been proposed for constraints on certain types of syntactic structures, constraints on patterns of morphemes within words, and on sound patterns. The latter explanations typically are described in terms of ease of articulation or ease in discrimination (in speech perception). Processing explanations for syntax and morphology, especially in explaining differences in cross-linguistic frequency of word order and morpheme order, are particularly associated with the work of J. Hawkins (e.g., Hawkins, 2004).

Cognitive Explanations

Cognitive explanations generally make reference to the conventional meaning of a word, inflection, or syntactic construction (e.g., Bybee, 2010; Tomasello, 1998/2003). A cognitive (semantic) explanation of a grammatical phenomenon essentially hypothesizes that some or all aspects of the analysis of the grammatical phenomenon make reference to conventional semantic properties. For example, some languages have plural inflections for all count nouns, while other languages restrict plural inflections to only nouns referring to animate beings, still others to only nouns referring to humans. This variation is built on a general ranking or hierarchy of nouns according to the semantic class of their

referents: human < animate < inanimate. The description of the constraint on the occurrence of the plural inflection in each language, and the cross-linguistic pattern of constraints, requires reference to the semantic classes of entities denoted by the noun, described in terms of animacy.

Cognitive explanations, like processing explanations, have been used to describe both grammatical constructions in particular languages and also patterns of cross-linguistic variation. The example of the animacy hierarchy in the preceding paragraph illustrates both sorts of explanations. It hypothesizes that the distribution of the plural inflection in any particular language will make reference to the semantic class of the entity denoted by the noun. It also hypothesizes that the cross-linguistic variation in the distribution of the plural inflection is subject to the hierarchical ranking of animacy categories given in the preceding paragraph. In recent cross-linguistic research (see Croft, 2001, 2003 and references cited therein), explanations of cross-linguistic patterns have been offered in which the patterns of occurrence of a construction are mapped onto a single semantic or conceptual space applicable to all languages.

Some cognitively oriented functionalists make a stronger claim about the relationship between grammatical form and semantic function, namely that certain grammatical categories are better described in semantic terms. This hypothesis will be discussed below on the contribution of cognitive linguistics to functionalism.

Discourse (Information Structure) Explanations

Discourse explanations make reference to the organization and presentation of information in discourse, that is, the structuring of information for the purpose of communication. Such explanations are called discourse explanations by some functionalists, and information structure explanations by other functionalists. For example, the English active and passive voice constructions (*The car hit the tree* vs *The tree was hit by the car*) both describe the same event and participants, but differ in which participant is encoded as the subject of the construction. It has been argued (based on naturally occurring examples) that the participant that is more topical is encoded as subject. Another example is the choice of construction in response to a question such as *Wasn't that incredible when Mary called the boss a pig?* One can respond with *Yeah, it really shocked me that she called him that*, but not with **Yeah, what really shocked me was that she called him that*. The unacceptable response uses the pseudocleft construction, and the pseudocleft construction is acceptable only if the information 'something really shocked me' is already in the hearer's consciousness. It is not in this case, and hence is unacceptable.

As with processing and cognitive explanations, discourse functional explanations have been offered both to explain restrictions on the acceptability of specific constructions in particular languages and to explain the frequency of distribution of particular linguistic features across languages. Examples of the former were given in the preceding paragraph. An example of the latter is certain patterns of word order. In the overwhelming number of languages, the subject precedes the object in the most common order of elements in a clause. It has been proposed that the reason for this asymmetry is that the subject tends to be the topic, and the most common discourse pattern is for the topic to precede the 'comment' or at least other less topical participants.

Schools of Functionalist Analysis

One of the most fundamental issues that divides schools of linguistic theory is: what is the relationship between linguistic form and linguistic function? The relationship is not at all straightforward. Also, depending on the position that a linguist takes on the relationship between form and function, certain additional hypotheses have to be made in order to accommodate linguistic facts. These hypotheses have to do with aspects of language not directly connected to the form–function relationship, such as the organization of grammatical knowledge in a speaker's mind, the nature of semantics, and the role of cross-linguistic evidence. Since these additional hypotheses are closely associated with various schools of functionalist analysis, they will be discussed here.

Functionalism and Formalism

Functionalist schools generally set themselves up as contrasting with the formalist approach, associated most strongly with N. Chomsky, but also advocated by other linguists who reject some of Chomsky's more specific claims about syntax. A purely formal explanation of a grammatical phenomenon makes reference exclusively to properties of grammatical form or structure. With respect to grammatical structure, a formal explanation of some pattern of sentence structure will make reference only to syntactic categories and structural relationships between the elements of a sentence, and not to the meaning of the words, inflections or constructions of that sentence, their discourse function, or their production and comprehension in language use. With respect to word structure, a formal explanation will make reference only to morphological categories and relations between roots and affixes, and not to aspects of the meaning and use of word forms. With respect to sound structure, a formal explanation will make reference only to the structure of the sound system independent of its articulatory and auditory realization.

The concept of a formal explanation is closely associated with the theoretical position that the organization of grammatical knowledge in a speaker's mind is divided into components which separate a linguistic form from its conventional meaning and use in discourse. A typical model describes grammatical knowledge as being compartmentalized into phonological, syntactic, and semantic components. The phonological component is separated from its phonetic realization, and the syntactic component is separated from its semantic interpretation. Linguistic form is linked up to its phonetic realization or semantic interpretation, eventually: formal models have rules linking syntax to semantics and phonology to phonetics. The formalist hypothesis does however imply that linguistic form is organized in a purely formal way to a high enough degree that it can be described best as existing in separate components. That is, the formal analysis of grammatical structure is self-contained relative to functional factors.

Functionalist approaches to language challenge or reject the formalist model of the relationship between form and function to varying degrees. However, all functionalists accept that not all aspects of grammatical structure can be accounted for by meaning and use. There will always be some degree of arbitrariness in language structure, where arbitrariness is intended to mean not wholly explainable in terms of language function. The question is, how much arbitrariness is there in language structure, and to what extent does the arbitrary aspect of language structure form a self-contained system? Different functionalist schools give different answers to this question.

Autonomous Functionalism

The closest approach to formalism which calls itself functionalist can be called *autonomous* functionalism. Autonomous functionalism argues that the number of properties of grammatical form that are explainable solely in grammatical terms are considerably fewer than most formalist theories assume. In particular, the restrictions on the use of particular constructions are sought in functional properties, whether cognitive, processing, or discourse-functional. To take a simplified example: in the *There*-construction (as in *There is a bird in the bushes*), a noun phrase with an indefinite article (*a bird*) is acceptable but one with a definite article usually is not (compare **There is the bird in the bushes*). A plausible explanation for this restriction is that the *There*-construction is used to introduce a new referent into the discourse, but definite noun phrases describe referents already known to both speaker and hearer. This would be an autonomous functionalist analysis of the restriction.

However, even in an autonomous functionalist approach, the rules governing the formation of the construction itself, such as the fact that *There* is the first element followed by a form of *be*, are considered to be purely formal. Autonomist functionalists focus their research efforts on accounting for as many of the constraints on the construction as possible in terms of functional properties, while leaving the basic structure of the construction for formalist analysis. The autonomist functionalist research program is found in the work of S. Kuno, E. Prince, G. Ward, and their students and associates. Their explanations generally appeal to properties of information structure rather than semantic properties. Because they accept the basic structural analyses of formalist approaches, autonomous functionalists generally do not reject the view that linguistic knowledge in a speaker's mind is organized into components. Instead, they have proposed that there is another component, the information structure component, which should be included along with the syntactic and semantic components.

In fact, many functional analyses of specific constructions in more radically functionalist approaches to language are actually autonomous functionalist analyses, that is, they account for the restrictions on syntactic constructions in terms of the meaning and use in discourse of the construction. We may call these autonomous functional analyses in whatever sort of functionalist approach they are found in. That is, autonomist functionalist analyses of particular constructions are not incompatible with more radical functionalist approaches.

Mixed Formalism/Functionalism

Another sort of approach, not clearly identifiable with any particular school of linguistic theory, but sometimes called functional, challenges the hypothesis of a sharp separation between the use of formal or functional properties to describe grammatical constructions. In these theories, the description of both the structure and the constraints on a construction can mix formal and functional properties. Two schools, *Functional Discourse Grammar* of S. Dik and K. Hengeveld and associates and *Role and Reference Grammar* of R. Van Valin and associates, both use formal and functional properties (typically semantics, but also information structure) to describe the grammatical structure of sentences as well as the constraints on them. In addition, many grammatical descriptions of languages informally describe both syntactic and semantic properties of constructions without implying any sharp division between the two.

Typological Functionalism

As was noted above, cognitive, processing, and discourse explanations have been offered for patterns of cross-linguistic variation as well as in the analysis of the grammars of individual languages. A school of linguistic theory known as typology has arisen which focuses on cross-linguistic variation and its limits (see, for instance, Comrie, 1989; Croft, 2003; Givón, 2001; Greenberg, 1990). Much cross-linguistic variation appears to be built on the base of general functional patterns, as illustrated above for cognitive, processing, and discourse explanations. This observation has led some typologists to argue that all cross-linguistically verified universals of language are susceptible to functional explanation in this way. This approach to analysis is called typological functionalism or the functional–typological approach.

Typological functionalism makes a stronger claim than the schools described in the preceding sections. The autonomous functionalist approach assumes that there still remains a significant amount of syntactic structure that can be described in terms of a self-contained set of rules (which are further assumed to be in part language-universal). The mixed formal–functional theories do not assume that formal principles governing syntactic structure are self-contained, but those theories allow for the possibility that there are formal properties of languages that are universal for all languages. The typological functionalist approach hypothesizes that any universal generalizations about linguistic structure are ultimately explained in functional terms.

The typological functionalist approach does not, in fact cannot, assume that the structure of a particular language's grammar is entirely explainable in functional terms. For instance, one example given above was that the occurrence of a plural inflection on nouns could be accounted for by the hierarchy human < animate < inanimate. The semantic hierarchy is universal, but each language applies the hierarchy to the distribution of the plural inflection in its own way, that is, one cannot predict what semantic class a language will extend (or restrict) its plural inflection to. That is, the exact class of nouns allowing a plural inflection is partly arbitrary, although

it is partly explainable in terms of the semantic hierarchy: if the language allows plural inflection for animate nouns, it will allow plural inflection for human nouns. This sort of explanation of a linguistic phenomenon is sometimes described as motivation, in contrast to a set of rules that exhaustively predicts the forms that would be found.

A typological functionalist analysis must allow for some degree of arbitrariness in accounting for cross-linguistic patterns. If there were no arbitrariness, all languages would be alike in their grammatical structure; and they are not. But the typological functionalist approach hypothesizes that the arbitrary aspects of language structure are language specific, and that any universal generalizations about linguistic structure are ultimately explained in functional terms.

An important element of typological functional explanations is the hypothesis of iconicity or iconic motivation, in which grammatical structure is hypothesized to reflect semantic or conceptual structure to some degree. For example, the closer grammatical relationship between verb and object as opposed to verb and subject in many languages is attributed to the closer semantic relationship between states of affairs and the participant usually encoded as the object. Iconic motivation can be used to account for grammatical structure itself, not just constraints on grammatical structures. In this respect, typological functionalism represents a more radical functionalist approach than autonomous functionalism.

Cognitive Linguistics and Functionalism

Cognitive linguistics (e.g., Lakoff, 1987; Langacker, 2008; Talmy, 2000) is a broad school of linguistic theory that is usually called functionalist. Cognitive linguistics, as its name indicates, focuses on cognitive (semantic) explanations for grammatical structure. Cognitive linguistics has contributed a number of ideas to other functionalist approaches that are relevant to assessing functional approaches.

Cognitive linguistics has proposed a model of grammatical knowledge, construction grammar, that offers an alternative to the formalist model in which form and function (semantics and information structure) are separated in distinct components (Croft, 2001; Goldberg, 2006). Construction grammar hypothesizes that knowledge of syntax involves the knowledge of individual constructions which combine formal properties (syntactic structures, morphological inflections) and functional ones (semantics and discourse function). For example, the ditransitive construction [X Verb Z Y], as in *Marilyn sent Gina a book*, specifies not only the syntactic structure of the construction but also the semantic relationships among the participants, such as the fact that Z comes to possess Y, by virtue of the form of the construction (in this case, two object phrases, referring to Z and Y, respectively, immediately following the verb). Constructions are organized in a network of grammatical knowledge in a speaker's mind. More broadly, grammatical knowledge is organized as a system of signs or symbols, consisting of form (the signifier) and function (the signified). A model organizing grammatical knowledge in terms of constructions/symbols/signs allows for a direct statement of functional properties that are relevant to particular grammatical constructions and other grammatical units.

Cognitive linguistics argues that semantics involves conceptualization or construal of an experience by a speaker for the purposes of linguistic communication. For example, an English count noun can be used in a mass noun grammatical context, as in *There was a huge Buick there; just acres of car* (attested example). The noun *car*, normally denoting an individuatable object, is construed to denote the size dimension, which is unbounded, in the noun phrase construction lacking a determiner. These and other examples indicate that the English count noun–mass noun distinction has at least some semantic basis. Cognitive linguists also accept that some alternative construals are conventionalized, such as *chicken* (meat, mass noun) vs. *chicken* (live bird, count noun). Nevertheless, the conventionalized grammatical distinction conforms to the semantics of count noun vs. mass noun: a live bird is an individuatable entity, while meat is an unbounded substance.

Conventionalized construals are of course partly arbitrary (see below); but the semantic regularity of both conventional and novel construals has led cognitive linguists to argue that even grammatical categories such as 'noun' and 'subject' are susceptible to a cognitive semantic explanation. That is, a grammatical category such as 'noun' encodes a construal of an entity in a certain way, which can be discerned by the occurrence of particular word roots directly as nouns, and in the semantic changes that occur when a non-noun root is used as a noun (for example, *a dance* reifies the act of performance and defines it as a particular sequence of steps), or can be morphologically derived as a noun (for example, *a musical composition* refers to the product of the act of composing).

Finally, cognitive linguistics has developed a model of categorization, which has in turn been exploited by functionalist linguists. A classical view of meaning is that the meaning of a word, inflection, or grammatical category must be a single definition that will cover all and only the appropriate uses of the word or category. Cognitive linguists have argued that many word and category meanings can be described in terms of polysemy, that is, distinct but semantically related meanings. For example, the meaning of *over* in *The bird flew over the hill* (path of motion) is distinct from its meaning in *The house is over the hill* (location), but they are related systematically. More specifically, cognitive linguists (drawing on research by psychologists) have suggested that many polysemous categories have one meaning that can be considered central or basic, called the category's prototype. Cognitive and other functionally oriented linguists have argued that many if not all grammatical categories are in fact polysemous, and many of these categories have a semantic/discourse-functional prototype which is found to be valid across languages (*see* Prototype Theory in Linguistics).

Functionalism and the Dynamic Approach to Grammar

Grammatical structure is commonly assumed to exist in speakers' minds. However, grammatical structure is also directly involved in social interaction in language use, and language use is central to accounting for language acquisition, language variation, and language change. In the more dynamic process of language acquisition and language change, functional factors have been argued to play a role.

In both language acquisition and language change, it has been argued that competing motivations among functional principles play a major role. The idea of competing motivations is that (functional) principles may come into conflict such that there is no grammatical system that satisfies all of the functional principles. As a result, change occurs over time in acquisition and in the history of a language, and the languages of the world exhibit structural diversity even though their speakers' linguistic behavior conforms to the same functional principles. (Competing motivation models do not presuppose that the competing principles are functional, and in fact competing motivation models are now used by many formalists; see below.)

The most general and commonly offered example of competing motivations is that between economy and iconicity (*see* Iconicity in Linguistics). *Economy* is the principle that a speaker uses the least effort necessary to express himself or herself. For example, English leaves the singular number of nouns unexpressed: *book* vs *book-s*. Economy is considered to be a speaker-oriented functional principle: its influence on language is for the benefit of the speaker. *Iconicity* is, in part, the principle that all of the relevant parts of the meaning conveyed are in fact conveyed by grammatical elements in the utterance (words, inflections, etc.). For example, in some languages, both the singular and the plural of a noun are expressed by overt suffixes. This aspect of iconicity is considered to be hearer oriented: any aspect of meaning left out by the speaker may not be recoverable by the hearer. Economy and iconicity compete with each other: a linguistic expression that is economical will not be iconic (since it leaves some elements of meaning unexpressed), and an expression that is iconic will not be economical (since certain elements of meaning are not left unexpressed). Hence, across languages, there is diversity in the expression of the category of number, and languages change from one form of expression of number to another over time.

Competing motivations models have been proposed in language acquisition (Bates and MacWhinney, 1989) and in language change (Haiman, 1985). More recently, some functionalist linguists (P. Hopper, R. Langacker, J. Bybee, W. Croft) have emphasized the dynamic character of language in ordinary use, and have argued that a speaker's grammatical knowledge should not be considered to be as static and immutable as is usually believed. They argue that a speaker's grammatical knowledge is not a tightly integrated system, but rather a more loosely structured inventory of conventionalized routines that have emerged through language use. The empirical research of the functionalist linguists who advocate this view has focused on the role of frequency of use on the entrenchment of grammatical knowledge, and on inductive models of abstracting grammatical knowledge from exposure to language use, both in acquisition and in adult usage.

These functionalists have argued for a usage-based or exemplar-based view of grammatical representation, in part following recent proposals in phonology. In the exemplar view, all tokens of grammatical units (words, constructions, etc.) to which a speaker has been exposed are at least initially stored as part of the mental representation of that grammatical unit, as a cloud of exemplars. The exemplars are structured qualitatively by dimensions of conceptual space (analogous to phonetic space), and processes in memory lead to their consolidation, entrenchment, and decay. These processes are dynamic processes that lead to changes in grammatical representation in the lifetime of an individual speaker.

Directions and Prospects for Functional Approaches to Grammar

Taken as a whole, functional approaches to grammar are widely practiced in linguistics. Nevertheless, there are important divisions between different schools of functionalist theory and between functionalists and formalists.

A common misunderstanding (among formalists in particular) is the belief that functionalists do not accept the existence of *arbitrariness* in language, that is, that not all aspects of grammatical knowledge are explainable by functional principles. There is no functionalist that holds that view. However, many functionalist theories have not developed a model of the representation of grammatical form, instead focusing their attention on the analysis of functional domains and their role in accounting for grammatical structure. There are significant exceptions to this observation, notably the mixed formal/functional theories, Bybee's model of lexical and morphological representation, and construction grammar in cognitive linguistics. Construction grammar is gaining a wider audience, and some functionalists are beginning to develop more detailed models of grammatical structure that are compatible with functionalist theories.

A significant split in functionalism is between those who are more cognitively oriented, focusing their attention on cognitive explanations, and those who are more discourse oriented. These two functionalist approaches have largely gone their own ways, developing models of language meaning (cognitive linguists) and language in use (the discourse functionalists). However, there are some functionalists such as T. Givón and W. Chafe whose work draws on both cognitive and discourse models. Usage-based models integrate cognition and discourse, in that there is a feedback relationship between the mental representation of linguistic knowledge and language use. The two approaches are inherently compatible: language must be understood from both a psychological and a social-interactional perspective.

Functionalist approaches to language generally have also neglected research in sociolinguistics (*see* Sociolinguistics). Sociolinguistics begins with the hypothesis that language is a dynamic, variable phenomenon, and argues that factors outside linguistic form, namely social parameters, account for variation and change in language. In both of these respects, sociolinguistics shares theoretical assumptions in common with functionalism. The increase of research on social-interactional factors and on relative frequency of use in functionalist research may lead to an integration of social with other functional factors in a social-functional theory of linguistic variation and change.

A rapprochement between functionalism and formalism is less likely, because the two differ in fundamental assumptions about the nature of language and the relation between form and function. There are some respects in which practical research is converging among formalists and functionalists,

however. For instance, competing motivations as a mechanism of analysis is now widely used in formalist theories under the name of *Optimality Theory*. The rise of computer-aided corpus research has led to a convergence in interests on naturally occurring language and the role of usage in grammatical structure. Some formal semanticists have made reference to prototypes and polysemy, and some semanticists working in generative grammar also give conceptualization a significant role in analysis.

See also: Functionalism in Anthropology; Generative Grammar; Grammaticalization; Iconicity in Linguistics; Language Acquisition; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Linguistic Typology; Optimal Control Theory; Phonology; Prototype Theory in Linguistics.

Bibliography

Bates, E., MacWhinney, B., 1989. Functionalism and the competition model. In: Bates, E., MacWhinney, B. (Eds.), The Cross-linguistic Study of Sentence Processing. Cambridge University Press, Cambridge, UK.

Birner, B.J., Ward, G., 1998. Information Status and Noncanonical Word Order in English. Benjamins, Amsterdam.

Bybee, J.L., 2007. Frequency of Use and the Organization of Language. Oxford University Press, Oxford.

Bybee, J.L., 2010. Language, Usage and Cognition. Cambridge University Press, Cambridge, UK.

Chafe, W., 1994. Discourse, Consciousness and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing. University of Chicago Press, Chicago.

Comrie, B., 1989. Language Universals and Linguistic Typology, second ed. University of Chicago Press, Chicago.

Croft, W., 2001. Radical Construction Grammar. Oxford University Press, Oxford.

Croft, W., 2003. Typology and Universals, second ed. Cambridge University Press, Cambridge, UK.

Dik, S.C., 1997 The Theory of Functional Grammar, second rev. ed., 2 vols. Mouton de Gruyter, Berlin.

Givón, T., 2001. Syntax: A Functional-Typological Introduction, second ed. John Benjamins, Amsterdam.

Goldberg, A.E., 2006. Constructions at Work: The Nature of Generalization in Language. Oxford University Press, Oxford.

Greenberg, J.H., 1990. In: Denning, K., Kemmer, S. (Eds.), On Language: Selected Writings of Joseph H. Greenberg. Stanford University Press, Stanford, CA.

Haiman, J., 1985. Natural Syntax. Cambridge University Press, Cambridge, UK.

Hawkins, J.A., 2004. Efficiency and Complexity in Grammar. Oxford University Press, Oxford.

Hengeveld, K., Mackenzie, J.L., 2008. Functional Discourse Grammar: A Typologically-Based Theory of Language Structure. Oxford University Press, Oxford.

Kuno, S., 1987. Functional Syntax: Anaphora, Discourse and Empathy. University of Chicago Press, Chicago.

Lakoff, G., 1987. Women, Fire and Dangerous Things: What Categories Reveal about the Mind. University of Chicago Press, Chicago.

Langacker, R.W., 2008. Cognitive Grammar: A Basic Introduction. Stanford University Press, Stanford, CA.

Prince, E.F., 1988. Discourse analysis: a part of the study of linguistic competence. In: Newmeyer, F.J. (Ed.), Linguistic Theory: Extensions and Implications, Linguistics: The Cambridge Survey, vol. II. Cambridge University Press, Cambridge, UK.

Talmy, L., 2000. Toward a Cognitive Semantics, 2 vols. MIT Press, Cambridge, MA.

Tomasello, M. (Ed.), 1998/2003. The New Psychology of Language: Cognitive and Functional Approaches to Language Structure, 2 vols. Erlbaum, Mahwah, NJ.

Van Valin Jr., R.D., LaPolla, R.J., 1997. Syntax: Structure, Meaning and Function. Cambridge University Press, Cambridge, UK.

Relevant Websites

http://www.linguistic-typology.org/ – Association for Linguistic Typology.
http://www.cognitivelinguistics.org/ – International Cognitive Linguistics Association.

Functional Brain Imaging of Language processes

Brigitte Stemmer, Université de Montréal, Montreal, QC, Canada; and Brock University, St. Catherines, ON, Canada
Frank A Rodden, Abteilung für Persönlichkeitspsychologie und Diagnostik, University of Zurich, Psychologisches Institut, Zürich, Switzerland

Published by Elsevier Ltd.

Abstract

Brain imaging and brain mapping techniques are major tools in language research and language rehabilitation. While this section aims to introduce the reader to the major techniques used in the study of language, it also wishes to raise the reader's awareness of the limitations of each of these methods. This discussion includes electroencephalography, magnetoencephalography, positron emission tomography, functional magnetic resonance imaging, density tensor imaging, functional near-infrared spectroscopy, transcranial magnetic stimulation, and transcranial direct current stimulation. The strengths and weaknesses of each of these methods – as well as examples of how these techniques have been applied to the study of language – follow explanations of how they function.

Introduction

Over the past 30 years, functional neuroimaging and brain mapping techniques have emerged that would have exceeded the wildest dreams of scientists 50 years ago. Findings derived from these techniques have given rise to cognitive neuroscience, a field in which questions in cognitive psychology are investigated with brain imaging techniques. Brain mapping attempts to relate the brain's structure to its function (or vice versa) in health or disease, or investigates how the brain changes physically in response to external or internal influences such as the aging process or learning and so on. Functional neuroimaging and brain stimulation techniques can be used to map brain functions. Functional neuroimaging techniques attempt to measure neuronal activity in the brain and relate this activity to behavior. There are basically two ways of doing so. One is by measuring changes in electrical or magnetic activity when neurons become active (electroencephalography (EEG) and magnetoencephalography (MEG)), and the other is by measuring the metabolic changes due to energy consumption of the active neurons (positron emission tomography (PET), functional magnetic resonance imaging (fMRI), and functional near-infrared spectroscopy (fNIRS)). Brain stimulation techniques such as transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS) influence brain structure and function through magnetic or electrical fields.

Regardless of the technique used, the research findings obtained are usually presented as colorful brain images that exert a curious fascination in academics, lay people, and media alike.

'Seeing is believing' is both a promise and a threat for those interested in brain science. Evolution has gifted us humans with an exquisite sense of sight that allows us to 'see how the world around us works' and part of that world is our own brains. In general, vision serves us well, but the ancient threat of mirages, of 'seeing things that aren't there,' is a constant companion when we read titles such as 'the anatomy of language.' Without language, understanding complex phenomena would be impossible, but language can not only elucidate, it can also obfuscate and confuse – and the tendency to confuse image with reality, to inappropriately reify images into something material is nowhere more treacherous than in neuroimaging (Bennett et al., 2007).

Caution must be applied when discussing how a piece of living meat, the brain, processes information. The images to be discussed in this chapter are perhaps best considered as parallels with their etymological counterparts, 'imaginations.' These 'pictures' are not technical diagrams to help understand the workings of a machine whose purpose is known, but rather they are highly abstract syntheses of thousands of infinitesimal phenomena framed in a theory that attempts to demonstrate how they may be directly associated with linguistic meaning. Further, images of brain activity only have meaning when the acquisition of the data on which they are based are acquired using the appropriate experimental design, the correct analyses, and when they are interpreted with care.

Each of the methods described below has its own strength and weakness; what they have in common is that they are limited by the physical properties of the system itself and by the way the brain works, or, in other words, the physiology of the brain.

Every image based on neurophysiological data is characterized by the degree of its resolution in space and time. The spatial resolution (or 'sharpness') of fMRI, for example, is on the order of a few millimeters but this relative 'sharpness' comes at the cost of long 'exposure times': fMRI images demonstrate events that unfolded in the brain over a period of several seconds. Images derived from EEG or MEG, on the other hand, demand only a few thousandths of a second (milliseconds) 'exposure time' but, compared with fMRI, these images are 'blurred' over large areas of the brain.

Compounded with basic limitations inherent in all neuroimages, one must add further possible distortions that can arise from the individual measuring devices themselves. Each measuring instrument is susceptible to artifacts generated by its own internal programming or by interference from nonneural tissue (the pulsation of the blood, for instance and, to a lesser extent, that of the cerebrospinal fluid). Further artifacts can arise at interfaces between/among various physical phases (solid, liquid, gas) of the tissues depicted in the images: scalp,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54009-4

skull, air-filled sinuses, membranes, blood, spinal fluid, and brain.

A satisfactory, simple explanation of how some of the methods discussed in this chapter work is not always possible. To understand, for example, how images from fMRI or MEG are generated demands a basic knowledge of Maxwell's laws of electromagnetism and elementary quantum mechanics, prerequisites that some readers may not possess. The good news, however, is that a working knowledge of these methods is within the reach of anyone with high school knowledge of electricity, magnetism, and simplified quantum theory. Part of this working knowledge includes knowing the benefits and limitations of each method. A discussion of these factors is found at the end of each of the following sections.

And, finally, the images produced by these methods are susceptible, not only to distortions produced by the measuring instrument itself and from interactions between the measuring instrument and the brain, but the brain-being-measured is also influencing *itself* by sustaining its vacillating psychological states of wakefulness/sleep/high-arousal/interest/fear or pain, and so on.

Trying to measure isolated neural activity associated with a particular brain task is a bit like trying to record the sound of a single flute during a musical festival in which several huge bands are loudly playing different songs simultaneously under the same roof. Whatever the object of the investigator's interest is, most of the brain will be actively doing something else at the same time: controlling the heartbeat, maintaining the sitting position, looking out for danger, and so on. With respect to energy demand, the brain is a very active organ: it has been calculated that although the brain represents only 2% of the body weight, it accounts for about 20% of the body's total energy consumption, and most of this energy ('dark energy') is unaccounted for by its responses to external stimuli (Raichle, 2006; see **Figure 1**).

Even given an adequate understanding of the imaging methods described below, it seems certain that new possibilities of misinterpreting neuroimages will continue to emerge and skepticism in science will always remain a cardinal virtue.

Functional Neuroimaging Techniques Based on Electromagnetism

The human brain contains various types of nerve cells (neurons) that communicate with each other via chemical and electrical signals. Information transfer from the cell that sends out information (presynaptic neuron) to the cell that receives the information (postsynaptic neuron) occurs at the synapse. Coherent postsynaptic activity of thousands of co-aligned pyramidal cortical neurons gives rise to electric and magnetic field changes that can be measured on or near the scalp. A technique that measures noninvasively *electric* fields is EEG and a technique that measures *magnetic* fields is MEG. EEG records electric field changes that arise from components of the postsynaptic currents that are oriented tangentially *and* radially with respect to the scalp, whereas the MEG measures the tangential component of these currents exclusively and is thus most sensitive to activity in the cortical sulci.

Electroencephalography (EEG)

The first measurement of human brain activity – in the form of minute changes in electrical potential measured between two locations on the scalp – was reported in 1924 by the German physiologist and psychiatrist, Hans Berger. A graphic recording of these voltage changes over time is called an 'electroencephalogram' (EEG) (see **Figure 2**). An EEG is the blurred result of huge numbers of neurons discharging at roughly the same time (in synchrony) compared with a huge number of other

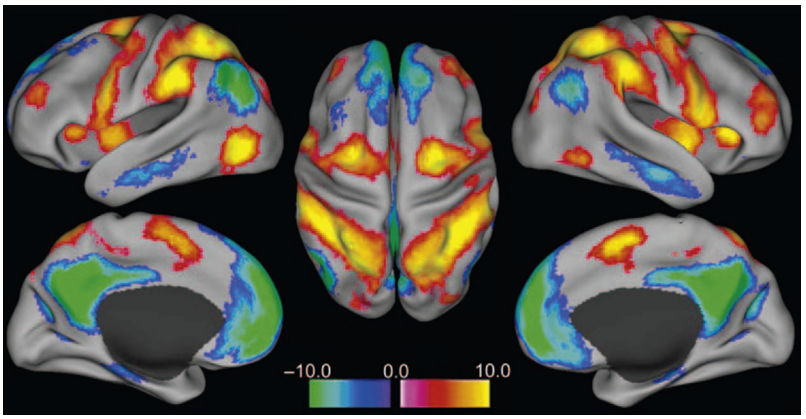

Figure 1 fMRI images of spontaneous brain activity in the absence of a task while the brain is at 'rest' (also called resting state). This, of course, does not mean that the brain is not doing anything – it only means that we do not know what the brain is doing and what goes on in a person's mind. It is assumed that the activity we 'see' when the brain is at 'rest' represents intrinsic brain activity, represented by correlated spontaneous fluctuations in the fMRI signal. The color scale is an arbitrary visualization of the correlations. In this image, activations (red to yellow colors) and deactivations (blue to green colors) routinely observed in response to attention-demanding tasks are also seen in the resting human brain, that is in the absence of any overt task or behavior. (Left) Lateral and medial views of the left hemisphere; (center) dorsal view; and (right) lateral and medial views of the right hemisphere. Taken from Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., Raichle, M.E. 2005. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proceedings of the National Academy of Sciences of the United States of America 102 (27), 9675. Legend adapted by B. Stemmer.

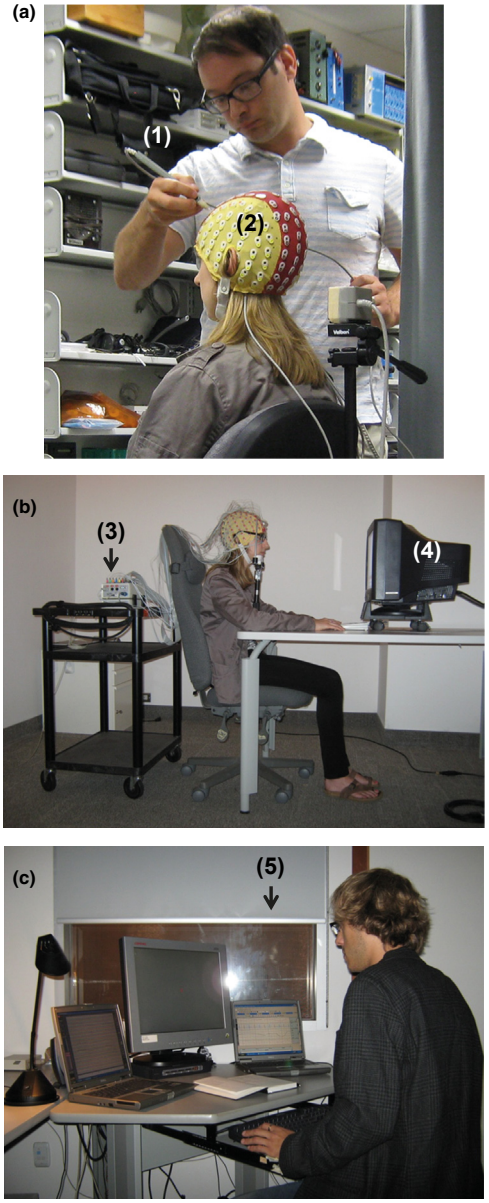

Figure 2 The EEG lab. (a) The subject is being prepared for the recording session. A 128-channel EEG cap has been placed on the head (2). The digitizer (2) is used for marking landmarks on the head and the position of the electrodes in 3D space. This will later allow specification of the anatomical location of the collected EEG signals. To also place a model of EEG generators into the brain, a structural MRI scan of the subject's head and brain has to be obtained in a separate session. The MRI scan is later overlaid ('coregistered') with the data obtained from the EEG. (b) The subject is inside the EEG recording room and watching stimuli on a computer screen (4). The electrode cables from the EEG cap go to the small amplifier box (3). From the amplifier cables go to the data acquisition computer placed outside the recording room. Underneath the amplifier box (not visible in the picture) is a small battery supplying the EEG system with electricity. (c) Outside the recording room a person supervises the EEG recording session. The person can watch the subject inside the recording room through the window (5). On one computer the EEG data are recorded and another computer shows the stimuli the subject sees inside the recording room. Reprinted with thanks to and permission from the Cognitive and Affective Neuroscience Laboratory of Brock University (http://www.brocku.ca/social-sciences/undergraduate-programs/psychology/research-labs/cognitive-affective-neuroscience). Copyright J. Desjardin and S.J. Segalowitz.

neurons discharging at roughly the same time somewhere else in the brain. EEG's have been compared with sound recordings of roaring crowds during sporting events. The voltage changes generated by single nerve cells are vanishingly small when compared with the potentials recorded with an EEG. Therefore, amplifiers are used to enhance the signals and analysis techniques have been developed to extract the signals we are interested in.

In conventional EEG, electrodes are fixed to the scalp (manually or prearranged in an electrode cap or net) and electrically conducting gel, paste, or liquid is applied. EEG signals represent the potential difference measured between two electrodes, an active electrode and the so-called reference electrode. The total number of electrodes used varies depending on the objective of the measurement and can go up to 256 electrodes (so-called high-density EEG). During neurosurgical operations, somewhat more spatially exact EEGs can be recorded by electrodes placed either on or under the leathery dura mater covering the brain. Even better spatial resolution can be obtained by sinking electrodes directly into the tissue of the brain itself (intracortical recordings). No matter where the electrodes are placed, however, the results of an EEG are due to the combined activities of huge numbers of pyramidal cells firing simultaneously and must be seen as a mass phenomenon. Voltage changes recorded from scalp EEGs range between 20 and 100 μV but reach the millivolt range (i.e., are roughly a hundred times stronger) when measured under the dura mater ('subdurally').

In order to plot the voltage changes over time, the recorded voltages must be amplified roughly a thousand-fold and separated ('filtered' in EEG jargon) from the much larger voltage changes that occur, for example, due to muscle artifacts or heartbeat. The amplified and filtered results are then digitally recorded as a function of changing voltage over time and viewed on a computer monitor.

EEG is a direct measure of brain activity in the millisecond range, i.e., EEG has a very good temporal resolution, and is thus particularly suited to investigate 'when something happens' such as specific processes underlying language. A drawback, however, is its spatial resolution, that is 'where something happens.' Since the size of the electrical potential diminishes exponentially with increasing distance from the source, neural activities deep in the brain are difficult to capture. In addition, the electric signals 'spread out' due to volume conduction as they traverse brain tissue, liquids, and the scalp, thus making it difficult to determine from a given pattern of electrical charge observed on the surface of the scalp where in the brain the neural generators are located (so-called inverse problem). To improve spatial resolution of EEG, the number of electrodes on the head can be increased (so-called high-density recording using 60–256 electrodes). Source localization techniques have been developed to identify the loci of the neural activities that give rise to the voltage distribution on the surface of the scalp. With these approaches spatial resolution of EEG has been estimated to be at about 1 cm. Once the sources of the neural

activities are identified in the brain, the exact anatomical location within the brain needs to be determined. This can be done by 'overlaying' the source image with an anatomical brain scan.

The information we can derive from an EEG depends on the analysis technique we use. For clinical purposes, for example, it is often sufficient to decompose the EEG signals into specific 'brain waves' as they are popularly called, such as alpha waves (7–12 oscillations per second, i.e., 7–12 Hz) associated with a relaxed but awake state, beta waves (13–20 Hz) associated with a more aroused mental state or very high-frequency oscillations such as gamma waves (>30 Hz) associated with higher cognitive functions. Although neglected for a long time, decomposing EEG signals through spectral decomposition, so-called time-frequency analyses, has – in the meantime – also attracted the attention of researchers interested in language. In combination with other analysis approaches, one can, for example, isolate components one is especially interested in and determine their sources in the brain and how they are connected to other components (see Figure 3). Special analyses techniques have also been developed to study how individual subjects perform on single trials in an experiment, this is particularly useful in children and patient populations. Although there may be a shift in the future to include other types of analyses, the analysis technique most frequently used in language research until today is the event-related potential (ERP) technique which will be discussed in the next section.

Event-Related Potentials (ERPs)

As mentioned above, the EEG has been likened to a sound recording of a crowd's roaring during a sporting event. From such a recording, information about, say, when a goal was made, might be possible, but probably not much else. Not much could be said about how the roars were related to particular events during the game.

If, however, one wanted to see if there was a connection between, say, the entrance of a particular player into the game and the crowd noise, one might consider analyzing the sound spectrum of the crowd, exactly when that player entered the game. If the player of interest was particularly popular with women, one might propose that the pitch of the crowd's roar whenever that player entered the game might be higher due to the cries of women. ERPs are a little like that.

Changes in the EEG due to such trivial events as 'seeing an X or an O on the screen' or 'hearing a particular word' are tiny and are generally swamped out by the 'background noise' of the EEG (see Figure 4). If, however, the event is repeated a large number of times (in linguistic research, typically 50–100 times but obviously the more, the better), and the focus of the analysis is on the time window during which the event took place, there is a good chance that one can tease out the response of the EEG at the time the critical stimulus (i.e., the word) was presented. To define the time window, one usually takes the time at which the stimulus is presented and marks it (in EEG jargon, one 'time-locks' the event to the onset of the stimulus) in the EEG recording. To extract the ERP from the background EEG a series of specific analysis steps are necessary, but the most important one is to average across all waveforms at the time of the occurrence of the critical stimulus. An example of a very common ERP experiment is presented in Figure 4. The ERP is that part of the EEG associated with the particular recurring event.

ERPs arising within roughly between 10 and 100 ms after a stimulus (in ERP jargon, 'with a latency of 10–100 ms') are generally considered to reflect early sensory (visual, acoustic, tactile) processes. Longer latency ERPs, i.e., those arising after 100 ms, are usually considered to be associated with perceptual processes and those after 200 ms with cognitive processes such as language or memory.

ERP waveforms are named according to whether the major component is positive ('P') or negative ('N'), according to the length of time they occur after the event, or according to the ordinal position after stimulus onset. One commonly observed ERP is the 'N400' which occurs as a negative ('N') wave roughly 400 ms after onset of the stimulus. Commonly, 'P3' OR 'p300' is the third positive peak arising about 300 ms after the onset of the stimulus. ERP waveforms have meaning, of course, only within the context of the specific experiment in which they are measured.

Advantages and Disadvantages. EEG and ERPs are direct measures of brain activity in the millisecond range. EEG is low-cost, noninvasive, portable and the measurement is silent (a distinct advantage in some linguistic studies). Compared to MEG and fMRI, EEG is generally more movement but its degree of spatial resolution is rather coarse (about 1 cm).

Magnetoencephalography (MEG)

According to Maxwell's Laws, wherever there is a change in an electric field, there is a corresponding change in a magnetic field; 'MEG' is the magnetic correlate to the EEG. (see Figure 5) The MEG is a graph of variations in minute magnetic fields that are measured just above the surface of the scalp and which reflect brain activity. Because the magnetic signals emitted by the brain are extremely weak, an MEG is housed in a specially shielded room so that external magnetic signals, including the earth's magnetic field, cannot interfere. Compared to electric fields, magnetic fields are, however, less prone to distortion due to intervening brain tissue between the source of the magnetic oscillations in the brain and the scalp. MEG thus yields better spatial resolution than does EEG. In comparison to EEG, MEG detects only changes in the components of the field tangential to the detectors, i.e., from the sulci of the brain rather than from both the sulci and the gyri as is the case with EEG. As with EEG, the changes measured by MEG are due to the simultaneous firing of hundreds of thousands of roughly parallel nerve cells.

The detectors of MEG are quite different from the scalp electrodes used for EEG measurement. MEG detectors are called 'squids' (for a superconducting quantum interference device) and are very sensitive instruments used to detect tiny changes in magnetic fields. Rather than have electrodes attached to the scalp, in MEG experiments, the subject sits or lies with the head inside a fixed helmet that contains the squids.

As is the case with EEG, MEG can also be used to detect changes in brain activity in the millisecond range but with a much finer spatial resolution: 0.1–1 cm. And similar to EEG, event-related phenomena can also be obtained by MEG – so-called event-related fields (ERFs). The sources of the neural activity identified with MEG can be projected on to individual brain scans obtained from MRI to visualize the anatomical location of the sources.

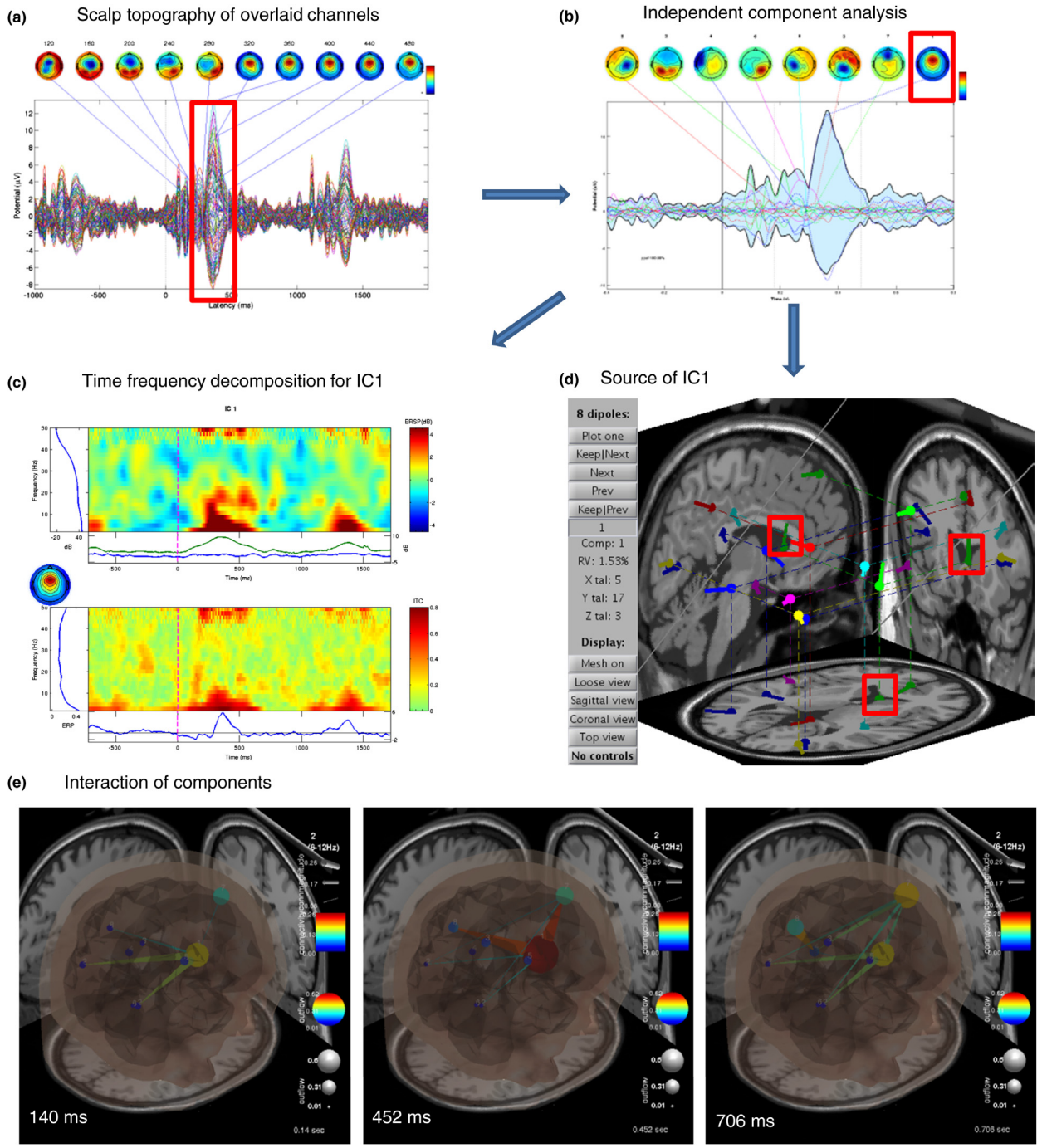

Figure 3 The figure depicts the ERPs in response to a go-nogo task. (a) Overlaid voltage distribution of 128 scalp channels. The largest field power occurs at around 360 ms (red box). (b) To reveal the hidden factors (latent variables) accounting for the large field power, an independent component analysis was performed, identifying eight components that account for almost all of the scalp ERP. The last component shown on the right (red box) accounts for nearly all the variance of the nogo P300 as can be seen in the waveform below. (c) On this component (here called IC1) a time-frequency decomposition was performed. This allows us to see when this component has power changes across time and frequency, independent of the phase consistency from trial to trial. This activity can then be examined on single trials to investigate whether there are patterns that show up from trial to trial (not shown here) for specific individuals. (d) To identify the sources of the components (i.e., where in the brain they occur) a dipole model is created from the identified eight independent components. Each dipole has a positive and a negative pole in its electric field. The tail points to the positive pole. The source identified for the (dark green) IC1 component is in the red box. Note how the tail of this component points to the peak positivity at the scalp, represented by the red circular region in the topographical map of IC1 in (b). (e) The Source Information Flow Toolbox analysis allows us to investigate in which way the individual components interact with the other components in different regions of the brain. Each figure corresponds to information flow at a specific point in time, where information flow refers to the Granger causation of one source's activation on another. The numbered dipoles correspond to the components identified. The size of the dipole head (i.e., the round 'node') reflects the strength of the outflow (causal activation coming from this source) and the lines joining the nodes show the end location and asymmetry of the information flow. The more asymmetrical the lines, the more asymmetrical the information flow between the adjoined sources. Reprinted with thanks to and permission from the Cognitive and Affective Neuroscience Laboratory of Brock University (http://www.brocku.ca/social-sciences/undergraduate-programs/psychology/research-labs/cognitive-affective-neuroscience). Copyright J. Desjardins and S.J. Segalowitz.

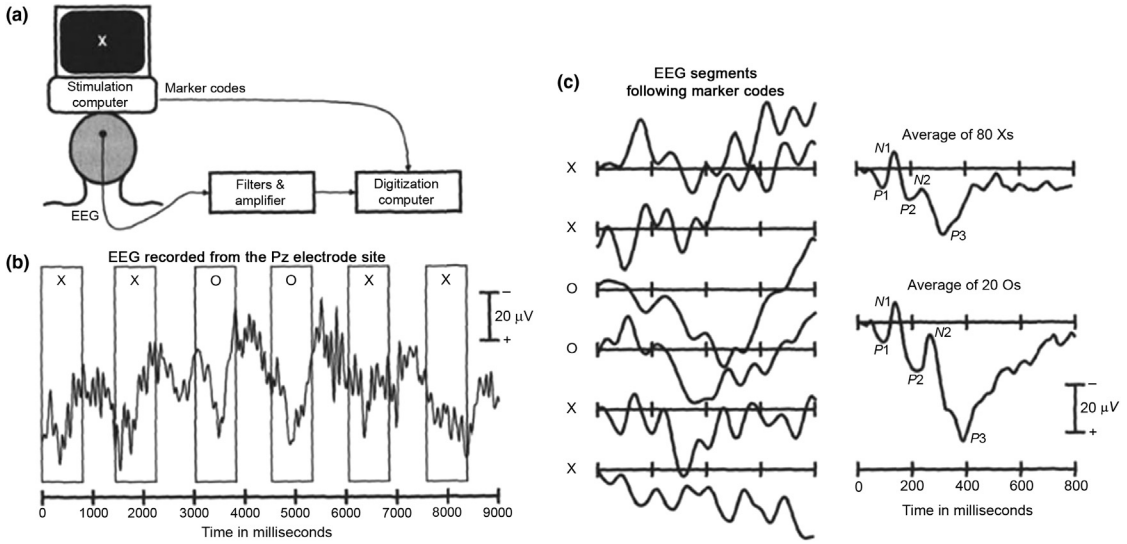

Figure 4 From EEG to ERP: the oddball paradigm. The oddball paradigm is a very commonly used experiment that reliably produced a specific ERP component, the so-called P3. (a) The subject views the letter X or O presented on a computer screen. Whenever an X or O appears on the screen, a computer program sends this information (via so-called marker codes or triggers) to the program that records the EEG. These can then be 'seen' in the EEG recording. The X is presented more often than the O, here a total of 80 times and the O 20 times, in random order and while the EEG is continuously recorded from a midline parietal electrode site. The recorded signals are filtered and amplified, making it possible to see the EEG waveforms. (b) Based on the marker codes signaling the onset of the stimulus (X or O), the EEG is now segmented into 800 ms time windows (or epochs). This is shown by the rectangles. A great deal of trial-to-trial variability can be seen in the EEG, but experienced viewers can already see a downward (positive) deflection following the infrequent O stimuli. (c) If one now averages across all waveforms in the O time window and all waveforms in the X time window, the P3 component (on the right side) becomes apparent. The P3 is much stronger for the infrequent O compared to the frequent X stimuli. Taken from Steven Luck, 2005. The Event-Related Potential Technique in Neuroscience. MIT Press, Cambridge, MA. (The same Figure was used in Rodden and Stemmer, in Stemmer and Whitaker, 2008. Handbook of the Neuroscience of Language, Elsevier, p. 59.) Legend by B. Stemmer (also in Handbook of the Neuroscience of Language).

Figure 5 MEG machine. The subject lies (as in the image) or sits with the head inside a fixed helmet that contains the superconducting quantum interference device (SQUID).

Advantages and Disadvantages. The major benefits of MEG are its speed (as with EEG in the millisecond range) and its good spatial resolution. The major disadvantages of MEG are price (several million dollars), size (several tons and several scores of cubic meters), and operating/maintenance demands. A further limitation in linguistic studies is its demand for strict head immobility during measurement as well as the exclusion of subjects with any ferromagnetic implants in their bodies.

Investigating Language with the EEG and MEG Techniques

It was previously mentioned that EEG and MEG are particularly suited to study *when* something happens in the brain and *how* linguistic processes unfold over time. There are a number of characteristic brain potentials recorded at the scalp that are assumed to reflect specific neurocognitive processes. These 'ERP components' will be described below.

ERP Responses Reflecting Aspects of Language Processing

An ERP component is usually characterized by the type of stimulus by which it is elicited, its polarity, the time window in which it occurs, and its location (distribution) on the scalp. Note that the location on the scalp is not identical with the source from which the ERP is generated! For example, the highest amplitude of an ERP may have been recorded over electrode F3 ('F' for frontal; uneven numbers refer to the left hemisphere and even numbers to the right hemisphere). This, however, does not mean that the generator of the ERP is in the left frontal cortex. Source localization analyses are used to indicate where the generator is located.

The following ERP responses ('components') are frequently used to address questions of language processing (for details see Steinhauer and Connolly, 2008).

N100. The N100 is a negative component peaking around 100 ms and has classically been associated with physical features such as the loudness of an auditory stimulus or brightness of a visual stimulus. This component has also been considered as a reflection of word segmentation processes.

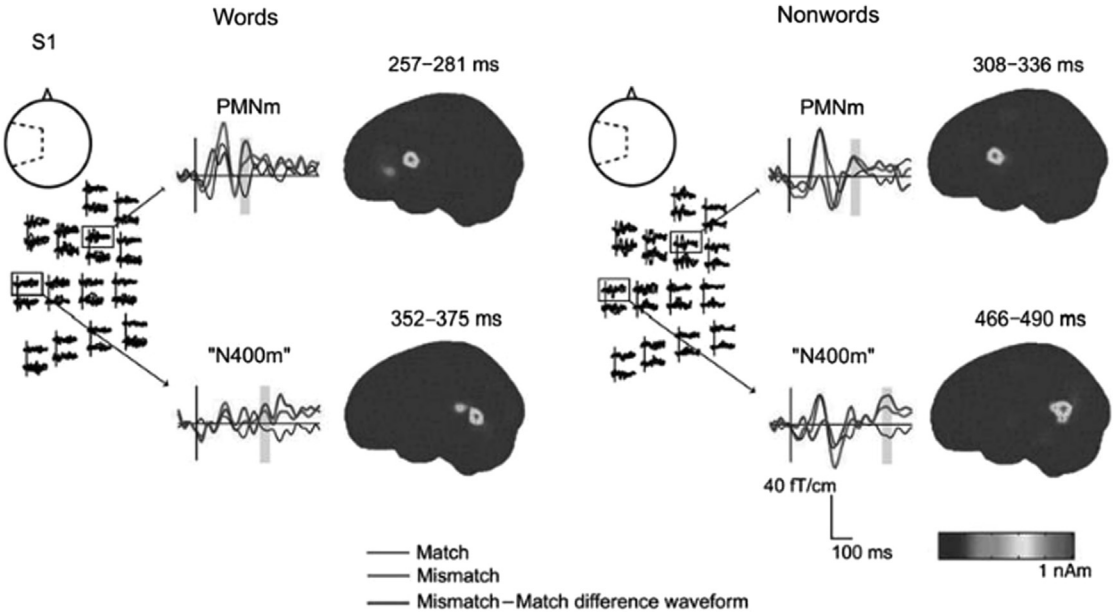

Figure 6 Phonological mapping negativity (PMN) and semantic N400. MEG responses to words (left) and nonwords (right) for one participant for those left-hemisphere channels showing maximum amplitude for the magnetic PMN (PMNm) and the N400-like response. The corresponding estimates of the PMN- and N400m-like response sources (over a 25-ms time window centered at the peak of the response) are depicted in the brain images. The gray vertical bars indicate the 50-ms time periods within which significant PMNm- and N400m-like responses occurred. Source: Modified after Kujala, A., Alho, K., Service, E., Ilmoniemi, R.J., Connolly, J.F., 2004. Activation in the anterior left auditory cortex associated with phonological analysis of speech input: localization of the phonological mismatch negativity response with MEG. Cognitive Brain Research 21 (1), 106–113. Figure and legend taken from Steinhauer, Connolly, 2008. Event-related potentials in the study of language. In: Stemmer, Whitaker (Eds.), Handbook of the Neuroscience of Language, Elsevier, San Diego.

PMN. The *phonological mapping negativity (PMN)* is elicited when the last word of a sentence violates the contextually developed *phonological* expectation, such as:

The pigs wallowed in the pen (mud).

The PMN occurs after auditory stimuli and is prelexical, i.e., occurs in response to words and nonwords. It peaks in the late 200 ms range (270–310 m) and has a fronto-central scalp distribution. It is assumed that the PMN reflects phoneme awareness and subsequent phonological processing activity (see Figure 6).

MMN. The mismatch negativity (MMN) is elicited by discriminable auditory differences between a frequently occurring 'standard' stimuli and a rare 'deviant' stimulus (the so-called 'oddball paradigm'). The negative component occurs between 150 and 250 ms and shows a frontal scalp distribution (see Figure 7).

N400. The N400 is the first language component that was described (in 1980). It peaks around 400 ms but its latency can range from about 300 to 500 ms. The N400 has a centro-parietal scalp distribution and is slightly right lateralized. The component was first elicited in response to words at the end of a sentence when these words violated the semantic context of the sentence in which they occurred (e.g., *The mother boiled the tires)*. It has been suggested that the N400 amplitude reflects costs of lexical activation and/or semantic integration. The N400 is sensitive to word frequency and it is larger in response to concrete than abstract words (see Figures 8 and 9).

Kutas and Federmeier (2011: 642) summarized the way that ERPs, and the N400 in particular, have been used in 30 years of research as follows:

> 1. ERPs provide temporally exquisite brain measures that have proven especially powerful tools for getting at aspects of cognition largely impenetrable with behavioral measures—for distinguishing among qualitatively different processes and tracking their time courses, often without any overt task and sometimes even in the absence of awareness.
>
> 2. ERP data in general—and the N400 in particular—underscore the crucial role for time (and timing) in comprehension (as in cognition more generally).
>
> 3. Furthermore, N400 data:
>
> (a) point to a distributed, multimodal, bihemispheric comprehension system that is simultaneously open to linguistic and nonlinguistic influences, which often interact;
>
> (b) suggest that access to meaning is a natural part of the stimulus-elicited processing stream, not dependent on an information state such as "recognition," and thus open to all stimuli in all task conditions (unless the feedforward stream has been suppressed by selective attention);
>
> (c) have shown that comprehension is largely nonserial, at least partially incremental, predictive, flexible, and context-dependent; and
>
> (d) have been instrumental in blurring the line between long-honored dichotomies such as prelexical versus postlexical, automatic versus controlled, and literal versus nonliteral processing, among others.

LAN. The left anterior negativities (LANs) are elicited by structural/syntactic violations and have been linked to

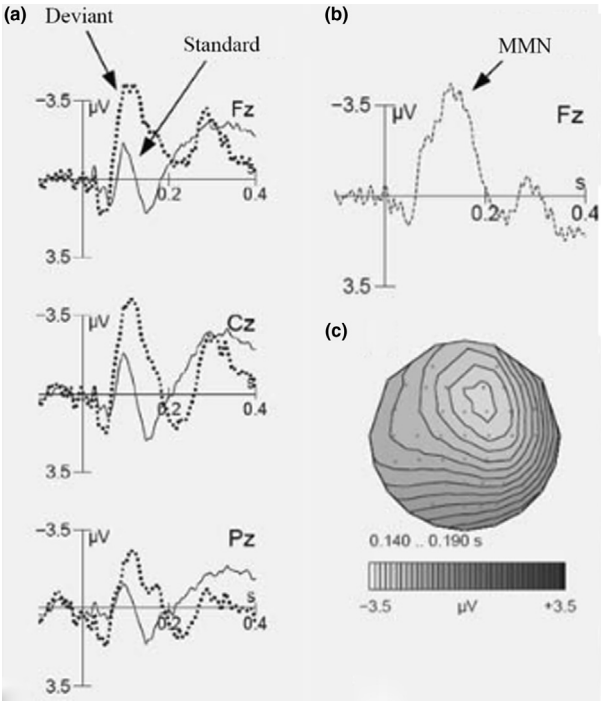

Figure 7 The mismatch negativity (MMN) is a preattentive response to discriminate auditory differences between a frequently occurring 'standard' stimulus and a rare 'deviant' stimulus – the so-called oddball paradigm (Näätänen et al., 2001). Panel (a) shows ERPs at midline electrodes for standard stimuli (85%, solid line) and deviant stimuli (15%, dotted line) in an auditory oddball. In panel (b) the difference wave (deviant minus standard) at frontal electrode FZ displays a large MMN and in panel (c) the voltage map between 140 and 190 ms post stimulus onset illustrates the frontal distribution of the MMN (Steinhauer, unpublished data). Within language research the MMN has been used to study categorical phoneme perception, most recently also probing the audio–visual McGurk– MacDonald effect (Massaro, 1998). This effect is achieved by having a subject watch a speaker articulate a syllable that differs from what is recorded on the audio channel the subject hears. The effect occurs when the subject sees the speaker articulate /ka/ but hears /pa/–this combination results in the perception of /ta/. Examination of the McGurk–MacDonald effect has led some researchers to suggest the existence of an 'audio–visual' MMN that has generators different to or overlapping with those that produce the purely acoustic MMN (Colin et al., 2002). They argue that, since their audio–visual stimuli had no acoustic deviants and the visual stimuli per se were incapable of eliciting MMNs, the visual stimuli required an appropriate auditory context within which an MMN was generated to an illusory deviant auditory percept. Also, they suggest that the MMN must result from an automatic, precognitive comparison between phonetic traces held in short-term memory. Figure and legend taken from Steinhauer, Connolly, 2008. Event-related potentials in the study of language. In: Stemmer, Whitaker (Eds.), Handbook of the Neuroscience of Language, Elsevier, San Diego, p. 94.

automatic first pass parsing of a sentence (i.e., automatically analyzing the grammatical relations between words and their hierarchical structure) (e.g., *The child wanted to *car the truck*). The LAN occurs between 100 and 500 ms and shows a left anterior distribution although the distribution is variable (see Figure 10).

A particularly early LAN is only elicited by word category violations and occurs between 100 and 300 ms (*She observed Frank's *on weight lifting sessions)*. However, the functional significance of this component has been seriously questioned by some researchers (Steinhauer and Drury, 2012).

P600. Another component that responds to syntactic violations is the *P600* (also called *syntactic positive shift, SPS*), a positivity that occurs between 500 and 1000 ms and which is distributed over the centro-parietal scalp. It is assumed to reflect relatively controlled attempts to reanalyze and fix the violation. Unlike the LAN, the P600 can be modulated by non-syntactic factors such as processing strategies or experimental tasks (see Figure 10).

CPS. The *closure positive shift (CPS)* is elicited at prosodic boundaries and is assumed to reflect the closure of intonational phrases. This component is among the first brain responses that can be observed in first and second language acquisition (see Figure 11).

In the remainder of this section, we will provide some examples of research questions addressed by the EEG and MEG techniques. It should be pointed out that these illustrations are only a selection of the large number of studies that have used the EEG or MEG method.

As mentioned previously, although the ERP technique is currently the most frequently used approach to address linguistic research questions, other analytical approaches have been developed but have not yet been widely used in language studies.

Brain Oscillations Investigated with EEG and MEG

Rhythmic changes in brain activity can be detected by EEG and MEG and extracted by time-frequency techniques. Combining such techniques with ERP approaches may deepen our insight into language processes in the brain (see also Figure 3 demonstrating the combination of these techniques with the ERP approach).

In language comprehension, our stored knowledge about a word or sentence must be retrieved from memory and integrated with other language input to form an overall representation of the utterance. A large network of widely distributed brain regions is involved in this process. The question how information from these different parts of the brain is integrated or 'unified' is referred to as the 'binding problem.' Assemblies of neurons that fire together simultaneously (i.e., are synchronized) are indicative of linking ('coupling') brain areas into functionally coherent networks. If neural synchronization occurs at a particular frequency, it is assumed that the various components belong to the same functional network. Memory retrieval processes have been related to oscillatory synchrony in the theta frequency range (4–7 Hz) and lexical retrieval processes with synchrony in the alpha frequency range (around 10 Hz); sentence-level semantic unification seems to be reflected in increasing oscillatory synchrony in the gamma frequency range (30–80 Hz), and syntactic unification seems to be reflected by neuronal synchronization in the lower-beta frequency band (13–18 Hz) (for a summary on oscillations in relation to aspects of language comprehension see Bastiaansen and Hagoort, 2006; Bastiaansen et al., 2009; Bastiaansen et al., 2012).

While numerous studies have investigated brain oscillations in language comprehension, applying this technique to language in pathological populations is still rare. An example

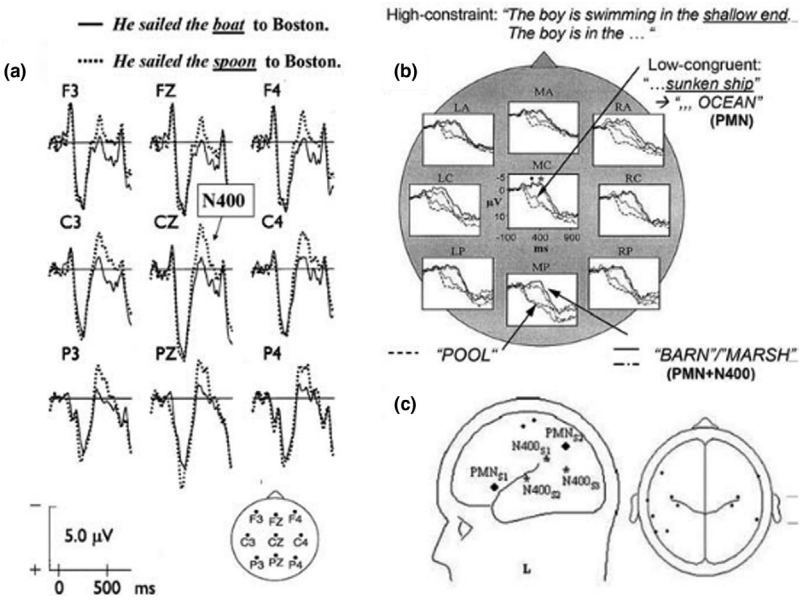

Figure 8 ERPs to target words illustrating N400 effects. Negative polarity is plotted upwards. (a) Semantic anomaly in sentences. Semantically implausible content words in sentence contexts elicit larger centro-parietal N400s (dotted lines) than plausible words (solid lines) between 300 and 600 ms. Source: Modified after Steinhauer et al. (2001). (b) PMN and N400 effects in a cross-modal priming study in which written sentences that ended in subordinate words primed superordinate words that ended spoken sentences. Exemplar-subordinate probabilities for congruent endings determined high-constraint primes (shallow end) or a low-constraint primes (sunken ship) which primed auditory target words in the paired sentences (in capitals) that were either congruent (*Pool/Ocean*) or not (*Barn/Marsh*). While incongruent targets (e.g., *Barn/Marsh*) always elicited PMNs and N400s, congruent targets following low-constraint contexts (*Ocean*) yielded PMNs only (L left, R right, M middle, A anterior, C central, P parietal). Source: Modified after D'Arcy, R.C.N., Connolly, J.F., Service, E., Hawco, C.S., Houlihan, M.E., 2004. Separating phonological and semantic processing in auditory sentence processing: a high-resolution event-related brain potential study. Human Brain Mapping 22 (1), 40–51. (c) Source localization revealed distinct neural generators for PMN and N400 components, primarily in the left hemisphere. Figure and legend taken from Steinhauer, Connolly, 2008. Event-related potentials in the study of language. In: Stemmer, Whitaker (Eds.), Handbook of the Neuroscience of Language, Elsevier, San Diego, p. 93.

is a study with dyslexic individuals. Based on the observation that dyslexic individuals exhibit phonological deficits, it has been hypothesized that the timing of auditory processing in these patients is impaired. An MEG study investigated this hypothesis by applying time–frequency analyses to the data obtained from dyslexic individuals. The findings indicated that phonemic processing in the left auditory cortex of dyslexic individuals seemed impaired as reflected in a decreased responsiveness to acoustic modulations at specific frequencies that are optimal for phonemic analysis (30 Hz) (Lehongre et al., 2011).

ERPs in the Study of Figurative Language

The temporal precision of electromagnetic approaches has been exploited to investigate the processes that underlie the interpretation of nonliteral language (e.g., hinting: *I took my car to the garage today* meaning 'Can you give me a lift home?') or figurative language (metaphors, irony, and so on). ERP studies using the N400 have indicated that literal language (*I think that surgeon is a woman*) and figurative language (*I think that surgeon is a butcher*) follow the same time course – thus suggesting that literal and metaphorical types of comprehension do not involve qualitatively distinct processing modes. One ERP study reported that pragmatic cues influenced sentence processing already at a very early stage. These (and similar) findings argue against the views of the standard pragmatic model.

Neuropsychological investigations have indicated a dominant involvement of the right hemisphere in idiom and metaphor processing. Most ERP studies have not supported strict right hemisphere dominance. Joke comprehension, however, seemed more reliant on the right hemisphere than idiom or metaphor processing.

An ERP study on nonliteral language addressing literal statements and indirect requests reported different brain activities for literal statements and indirect requests in response to target sentences that had exactly the same wording but whose interpretation differed depending on the context. The findings suggested greater memory retrieval demands for indirect requests (for a summary of ERP studies of figurative language see Coulson, 2012; Kutas and Federmeier, 2011).

ERPs in Language Development

In the immature brain, the EEG looks different from that of adults. Similarly, ERPs develop as the brain matures and can thus be used to reflect the emergence of cognitive and linguistic development stages (for a summary see Steinhauer and Connolly, 2008; see **Figure 12**).

The MMN can already be identified in newborns thus reflecting their ability to discriminate sounds in their native language from nonnative languages. At around 10 months, the infant shows a specialization for the discrimination of sounds in her/his native language which is reflected in a larger

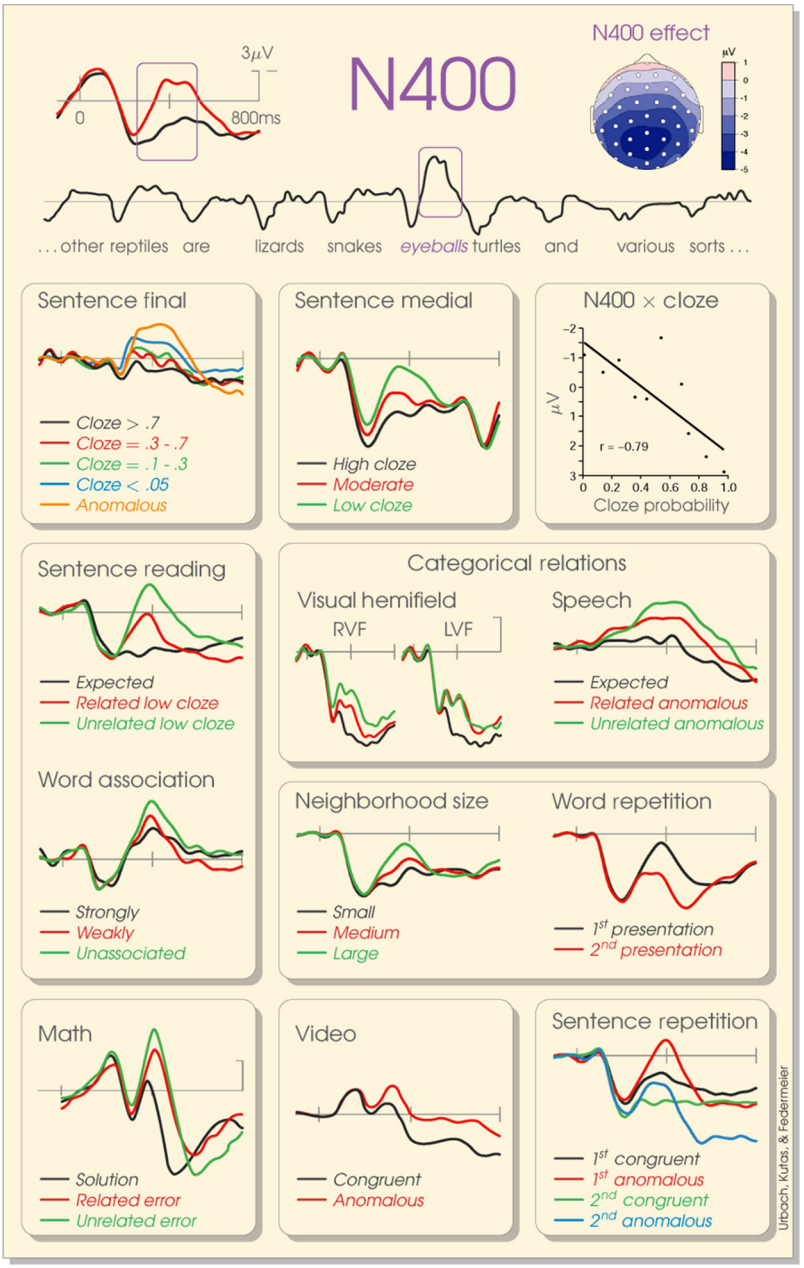

Figure 9 Over-sentence average ERP to words in written text, including large N400 effect to a semantically anomalous word. Above, on left, is canonical visual ERP comparison of congruent versus anomalous sentence final words, on a 3-μV scale (which applies to all but two waveforms, with 5-μV scales instead); above, on right, is a topographic map of the ERP difference (anomalous minus congruent, N400 effect) from 300 to 500 m postfinal word onset. (*Row 1*) All data sets from central midline site. *Sentence Final:* N400 to semantic anomaly and other sentence endings as a function of average cloze probability (Kutas et al. 1984); *Sentence Medial*: sentence medial noun at three levels of cloze probability; *N400 x Cloze*: correlation between mean N400 (300–500 ms) amplitude and cloze probability (DeLong et al., 2005). (*Row 2) Sentence Reading*: midline parietal ERPs to expected or best completion (BC) ending given sentence context, low cloze probability endings that were and were not related to the BC (Kutas, 1993); *Categorical Relations*: ERPs to sentence final word that is an expected category exemplar or an unexpected, implausible exemplar from the same category as the expected one (related anomalous) or from a different category (unrelated anomalous); *Visual Hemifield:* at the right medial central site in response to final word briefly flashed in the left visual field (LVF) – i.e., initially to the right hemisphere – or right visual field (RVF); *Speech*: at the frontal midline site in response to natural speech (see Federmeier, 2007). (*Row 3*) *Word Association*: ERPs to second word of a word pair, strongly related, moderately related, or unrelated to the first word from same participants as *Sentence Reading* above (Kutas, 1993); *Neighborhood Size*: midline parietal ERP to letter strings including words, pseudowords, and acronyms in a list as a function of orthographic neighborhood size (Laszlo and Federmeier, 2010); *Word Repetition*: ERPs to just the words upon initial presentation and their repetition, from these same experimental runs and participants. (*Row 4*) *Math*: midline central ERPs to correct equation solutions and errors that were or were not related to the solution (Niedeggen et al., 1999); *Video*: midline frontal ERPs to last video frame, congruent or anomalous with prior video context (Sitnikova et al., 2008); *Sentence Repetition*: midline central ERP to congruent or anomalous sentence final words on initial presentation and one repetition. Figure and legend taken from Kutas, M., and Federmeier, K.D., 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annual Review of Psychology 62, 621–647.

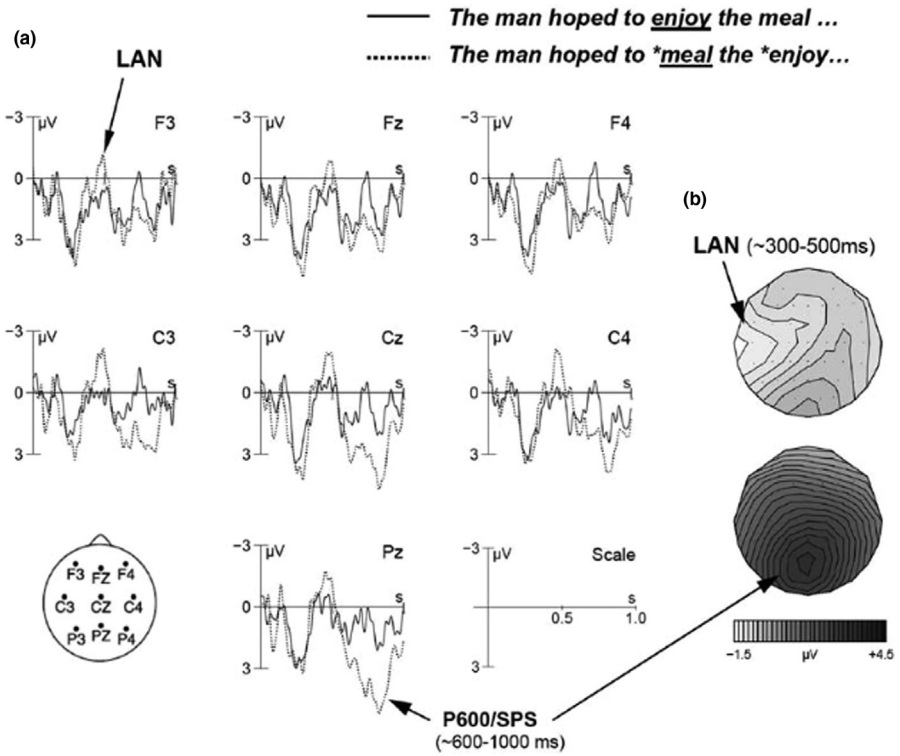

Figure 10 Biphasic LAN and P600 pattern elicited by syntactic word category violations in an English reading study. Negative polarity is plotted upwards. (a) ERP plots of the ungrammatical target words (dotted lines) show a LAN effect between 300 and 500 ms, that is in a similar time interval as the N400, which was followed by a late posterior 600/SPS between 600 and 1000 ms. (b) Voltage maps of the difference waves (violation minus correct control) illustrate the scalp distribution of both ERP components (Steinhauer, unpublished data). Figure and legend taken from Steinhauer, Connolly, 2008. Event-related potentials in the study of language. In: Stemmer, Whitaker (Eds.), Handbook of the Neuroscience of Language, Elsevier, San Diego, p. 97.

and more robust MMN. This discriminatory ability for nonnative sounds decreases and the MMN disappears with age. When infants are able to distinguish prosodic boundaries at around 6–8 months of age, the closure positive shift (CPS) can be identified. Just after infants start to babble, the lexico-semantic N400 can be elicited at around 12–14 months of age. ERP components reflecting syntactic processing (LANs, P600) are the latest to appear. P600 is seen first at around 24 months and LAN at around 32 months. (for summaries of ERPs in the study of language see Steinhauer and Connolly, 2008 and for the N400 see Kutas and Federmeier, 2011).

ERPs in Bilingualism and Second Language Research

The use of EEG to study learning or training programs is still rare and only recently has it begun to be employed. Questions that might be addressed using EEG methods could, for example, relate to beneficial teaching and learning strategies, maturational effects, or influencing variables (cognitive abilities, other language knowledge, cultural background). The question as to whether there is a 'critical/sensitive period' for acquiring a second language has been addressed by several studies. While early ERP studies supported the critical period hypothesis, subsequent studies provided mixed results (for a brief review see Steinhauer and Connolly, 2008).

Most EEG/ERP studies on second language acquisition have been performed outside of formal settings (such as classroom teaching). An important finding has been that ERPs reflecting phonological, semantic, or syntactic processing are modulated by the age at which the language was acquired and the language proficiency of the bilingual individual.

Numerous studies have investigated the 'bilingual advantage' (see Section Investigating Neural Substrates Underlying Bilingualism). Studies using nonimaging methods have reported that bilinguals have an advantage over monolinguals on tasks measuring 'executive functions' (for a summary see Hilchey and Klein, 2011; Kousaie and Phillips, 2012). Executive functions is an umbrella term for cognitive processes that control and manage other cognitive processes or systems such as organizing, planning, attention, working memory, reasoning and problem solving, mental flexibility, as well as initiation, monitoring, and inhibition of actions.

One MEG study reported that despite similar behavioral performance, monolinguals and two bilingual groups differed in activity patterns in specific brain regions (left prefrontal cortex and anterior cingulate cortex) when the groups performed a task that has been related to executive control. Whereas the two bilingual groups activated similar brain regions (compared to the monolingual group), only one of the bilingual groups (Cantonese) outperformed the monolingual group. Questions have been raised as to whether it was the *involvement* of the identified brain regions, or *how* these regions were used that contributed to the general behavioral advantages.

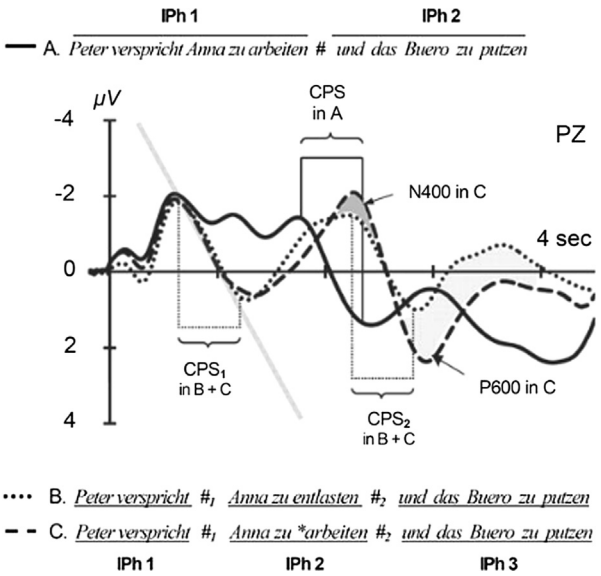

Figure 11 Illustration of the closure positive shift (CPS) at prosodic boundaries of spoken German sentences and a prosody-induced syntax violation negative polarity is plotted upwards. Boundary positions and intonational phrases (IPh) are aligned to the time axis. Sentence A (solid line) has only one prosodic boundary (#) after the verb 'arbeiten,' while B (dotted line) and violation condition C (dashed line) have two such boundaries. At each boundary position a large CPS component was elicited in each condition. Magnitude and slope of the CPS (illustrated by the thick gray line at CPS1) are very similar at all boundary positions. In C, the syntax-prosody mismatch on the verb 'arbeiten' additionally elicited an N400/P600 pattern which superimposes the second CPS. Waveforms represent a grand average ERP at PZ across 40 subjects and approximately 5000 trials per condition. Prosodic boundary information is not only important during language learning but also guides the listener's syntactic analysis and sentence comprehension. Source: Modified after Steinhauer, K., 2003. Electrophysiological correlates of prosody and punctuation. Brain and Language 86, 142–164 and Steinhauer, K., Alter, K., Friederici, A.D., 1999. Brain potentials indicate immediate use of prosodic cues in natural speech processing. Nature Neuroscience 2 (2), 191–196. Translation of sentences: A. Peter promises Anna to work # and to clean the office. B. Peter promises # to support Anna # and to clean the office. C. Peter promises # *[to work Anna] # and to clean the office. (Conditions A and C are lexically identical and differ only prosodically.) Figure and legend taken from Steinhauer, Connolly, 2008. Event-related potentials in the study of language. In: Stemmer, Whitaker (Eds.), Handbook of the Neuroscience of Language, Elsevier, San Diego, p. 100.

The relationship between cognitive control and bilingualism was also investigated with EEG using three different tasks related to various aspects of cognitive control (Stroop, Simon, and Eriksen flanker task). Behavioral measures did not show any bilingual advantage for these tasks. Whereas fMRI studies had shown the activation of similar brain regions in response to the three tasks, ERP responses to the three tasks were different. A comparison of the ERPs of bilinguals and monolinguals demonstrated differences in specific processes related to each task. These differences, however, were not consistent across the three tasks, thus suggesting that bilinguals and monolinguals do not respond uniformly to the tasks (for a summary offv neuroimaging studies investigating bilingualism and second language acquisition see Hilchey and Klein, 2011; Kousaie and Phillips, 2012; van Heuven and Dijkstra, 2010).

Language-Related ERPs in Pathological and Special Populations

ERPs have been used in studies of a large range of pathological conditions and special populations (patients with, for example, aphasia, Alzheimer disease, autism, head injury, dyslexia, mood disorders, schizophrenia, and so on). The focus in these studies has been on cognitive processes and some of the studies have also included investigations of language. The ERPs technique has, for example, been used with aphasic patients to specify impaired language processes, to predict outcome, or to understand how the brain recovers during the acute and chronic stages of aphasia. In dyslexia research, ERPs have been used to try to predict language development in children with a familial risk for dyslexia, or to determine brain responses in brain regions assumed to be implicated in dyslexia. Language-related ERPs have also been used to investigate the states of consciousness in patients in a coma.

Summary of Functional Neuroimaging Techniques Based on Electromagnetism

EEG and MEG are noninvasive techniques that measure electric or magnetic fields, respectively, in the brain. Due to their good temporal resolution in the milliseconds range they are particularly well suited to investigate language processes. With high-density EEG and source localization techniques, a spatial resolution of about 1 cm can be obtained. Scalp-recorded ERPs and ERFs are based on EEG or MEG, respectively, and are assumed to reflect specific neurocognitive processes. A number of ERP responses have been identified and associated with specific language processes. Applying special analysis techniques, the topographical brain sources of an ERP can be identified and its relationship to other brain regions investigated. Another analysis approach makes use of time–frequency analyses to characterize brain oscillations. These have been associated with various aspects of language processing.

The EEG/MEG and ERP/ERF techniques have been applied in a large number of language studies with healthy individuals, individuals with pathological conditions, and special populations.

Functional Neuroimaging Techniques Based on Blood–Brain Interactions

It has been less than 50 years since the first images of localized brain activity associated with brain functions based on blood–brain interactions were taken (for a historical summary see Figure 13). Researchers and lay people alike have embraced these new techniques with enthusiasm (especially PET and fMRI). Reasons are manifold such as the (false) belief of some laypeople that scientists can now watch *what* a person is thinking as witnessed in the vivid and colorful brain images produced by these techniques. Another reason is that these techniques have been used to study the relationship between the 'working' brain and behaviors that are of interest to many people such as sex, emotions, moral values, memory and language, and so on. Some of the philosophical

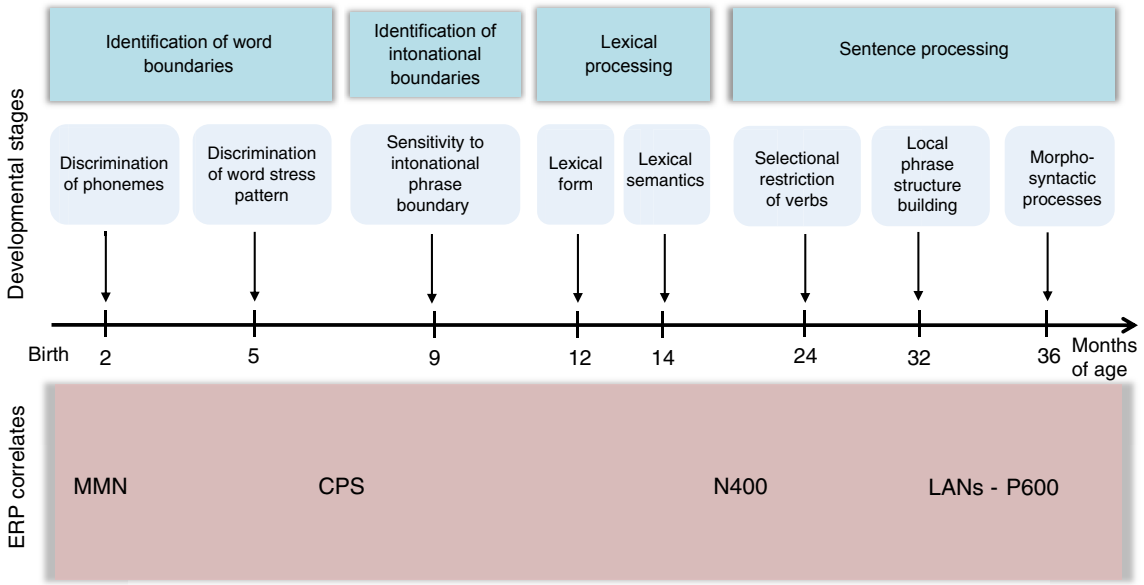

Figure 12 Emerging ERP components in language development. Adapted from Steinhauer, K., Connolly, J.F., 2008. Event-related potentials in the study of language. In: Stemmer, B., Whitaker, H.A. (Eds.), Handbook of the Neuroscience of Language. Academic Press/Elsevier, San Diego, pp. 91–104.

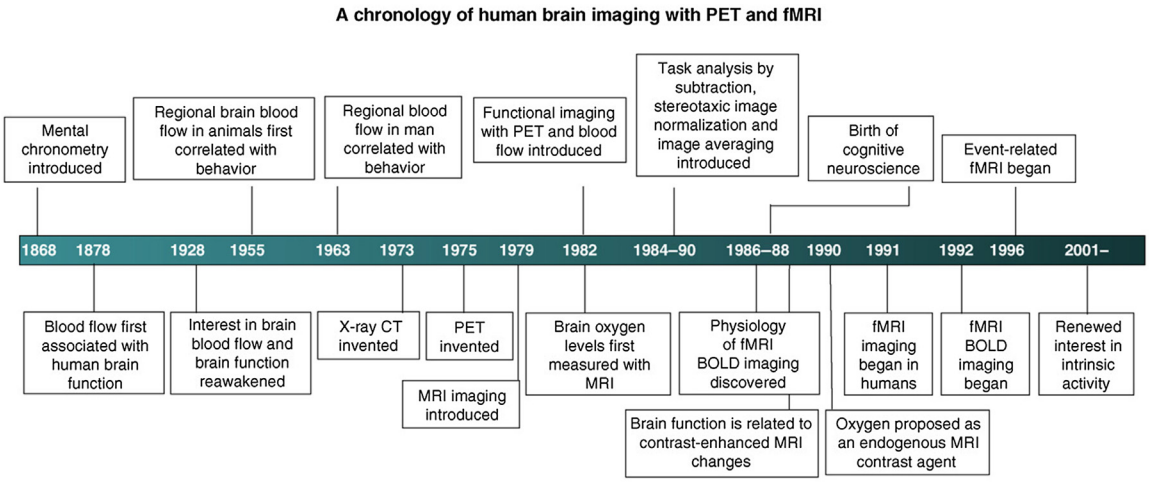

Figure 13 A chronology of major events associated with the development of human brain imaging with PET and fMR. Taken from Raichle, M.E., 2009. A brief history of human brain mapping. Trends in Neurosciences, 32 (2), 118–126 (Figure 1, p. 119).

pitfalls in these associations were already mentioned in the Introduction. Critical voices have also come from brain scientists calling for a more cautious use of the techniques, with more care being given to the analyses and interpretation of the data. In 2008 a paper from one of the leading fMRI researchers noted that "fundamental questions concerning the interpretation of fMRI data abound, as the conclusions drawn often ignore the actual limitations of the methodology" (Logothetis, 2008). In 2009 a vivid (and probably long overdue) discussion ensued after the publication of two critical papers. A paper entitled 'Voodoo in Social Neuroscience' questioning the reliability of social brain imaging studies circled the internet and caused hot discussions amongst scientists even before it was later published under a different title (Vul et al., 2009). The paper inspired Craig Bennett and colleagues to perform an unusual experiment. The findings were published at the Human Brain Mapping conference in Boston demonstrating *activated* brain regions of a *dead* Atlantic salmon that was scanned with fMRI while (supposedly) performing a decision-making task (see Figure 14).

Every method has its caveats and limitations but within carefully constructed paradigms, applying appropriate analyses, and using wise interpretation, functional neuroimaging techniques can exhibit changes taking place in the brain when cognitive processes such as language are investigated.

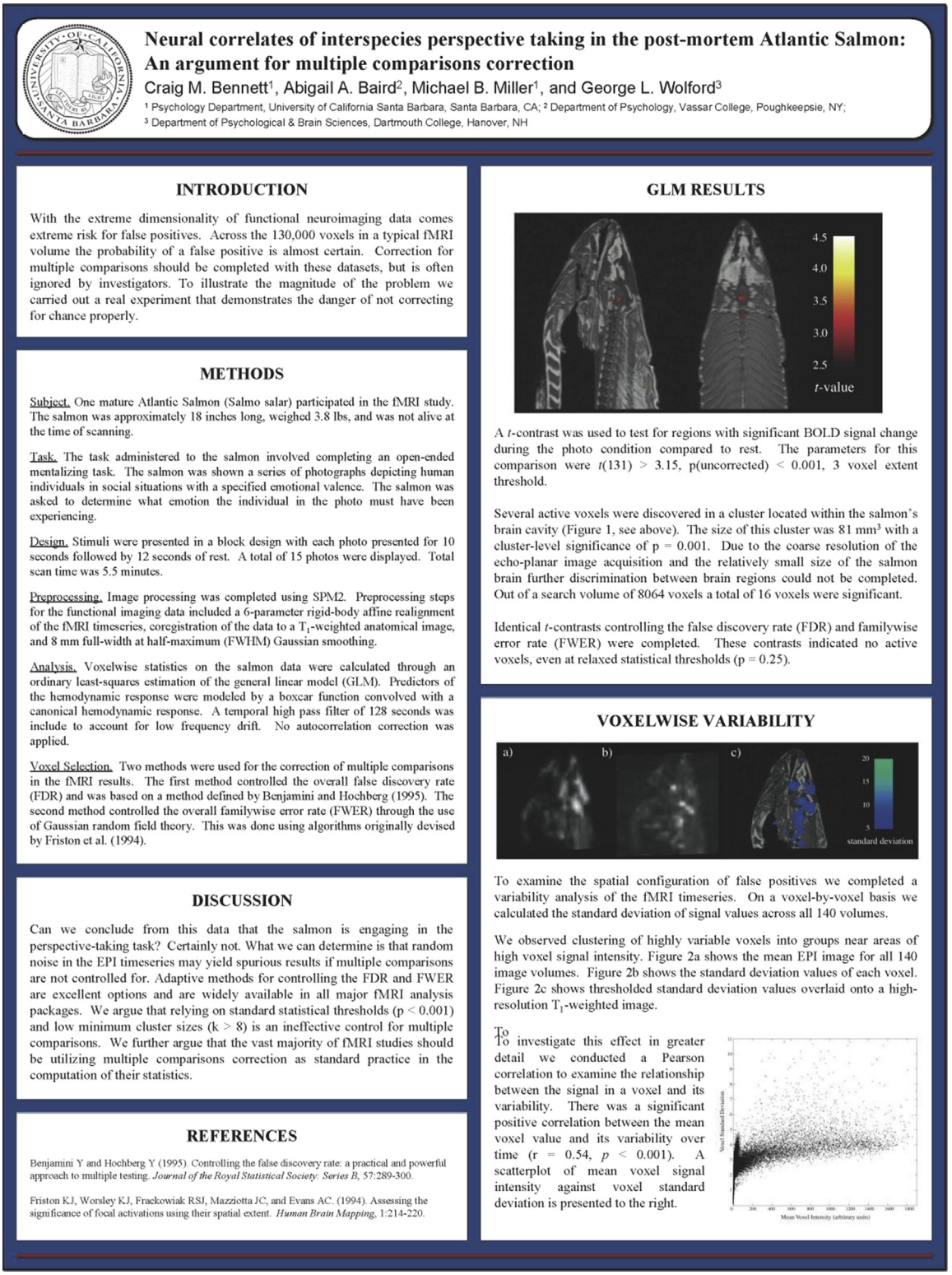

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY; [3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

Preprocessing. Image processing was completed using SPM2. Preprocessing steps for the functional imaging data included a 6-parameter rigid-body affine realignment of the fMRI timeseries, coregistration of the data to a T_1-weighted anatomical image, and 8 mm full-width at half-maximum (FWHM) Gaussian smoothing.

Analysis. Voxelwise statistics on the salmon data were calculated through an ordinary least-squares estimation of the general linear model (GLM). Predictors of the hemodynamic response were modeled by a boxcar function convolved with a canonical hemodynamic response. A temporal high pass filter of 128 seconds was include to account for low frequency drift. No autocorrelation correction was applied.

Voxel Selection. Two methods were used for the correction of multiple comparisons in the fMRI results. The first method controlled the overall false discovery rate (FDR) and was based on a method defined by Benjamini and Hochberg (1995). The second method controlled the overall familywise error rate (FWER) through the use of Gaussian random field theory. This was done using algorithms originally devised by Friston et al. (1994).

DISCUSSION

Can we conclude from this data that the salmon is engaging in the perspective-taking task? Certainly not. What we can determine is that random noise in the EPI timeseries may yield spurious results if multiple comparisons are not controlled for. Adaptive methods for controlling the FDR and FWER are excellent options and are widely available in all major fMRI analysis packages. We argue that relying on standard statistical thresholds ($p < 0.001$) and low minimum cluster sizes ($k > 8$) is an ineffective control for multiple comparisons. We further argue that the vast majority of fMRI studies should be utilizing multiple comparisons correction as standard practice in the computation of their statistics.

REFERENCES

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57:289-300.

Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, and Evans AC. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214-220.

GLM RESULTS

A t-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

Identical t-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ($p = 0.25$).

VOXELWISE VARIABILITY

To examine the spatial configuration of false positives we completed a variability analysis of the fMRI timeseries. On a voxel-by-voxel basis we calculated the standard deviation of signal values across all 140 volumes.

We observed clustering of highly variable voxels into groups near areas of high voxel signal intensity. Figure 2a shows the mean EPI image for all 140 image volumes. Figure 2b shows the standard deviation values of each voxel. Figure 2c shows thresholded standard deviation values overlaid onto a high-resolution T_1-weighted image.

To

To investigate this effect in greater detail we conducted a Pearson correlation to examine the relationship between the signal in a voxel and its variability. There was a significant positive correlation between the mean voxel value and its variability over time ($r = 0.54$, $p < 0.001$). A scatterplot of mean voxel signal intensity against voxel standard deviation is presented to the right.

Figure 14 Poster presented at the Human Brain Mapping conference in San Francisco, 2009. Reprinted with thanks to and permission from Craig Bennett Copyright C. Bennett. For a more detailed discussion see Bennett C.M., Wolford G.L., Miller, M.B., 2009. The principled control of false positives in neuroimaging. Social Cognitive and Affective Neuroscience 4 (4), 417–422.

Common to all techniques discussed below is that they are based on the assumption that those regions of the brain that are particularly active (say, for example, when judging the grammaticality of a sentence) show an increase in blood flow and energy consumption such as glucose and oxygen. The increase in cerebral blood flow and energy consumption that occurs in specific regions of the brain is then, in some way, measured (albeit differently for all three techniques) and associated with the processes underlying the task (see Figure 15).

As can be seen in Figure 13, PET was invented in 1975 and 7 years later, in 1982, functional imaging with PET and blood flow was introduced. Ten years later, fMRI was first performed in humans. Although today fMRI is the dominant technology used for cognitive studies, the first functional imaging study with language was performed with PET. We will first introduce PET followed by the fMRI and fNIRS techniques.

Positron Emission Tomography (PET)

PET can be used in three different measuring modalities. In one modality, PET can measure pharmaceutical parameters such as the uptake, binding, accumulation, or release of molecules of interest in the brain. These 'binding kinetic studies' are most commonly carried out to determine pharmacological parameters of pharmaceuticals (and their derivatives) such as antidepressants, neurotransmitters, or analogues of neurotransmitters.

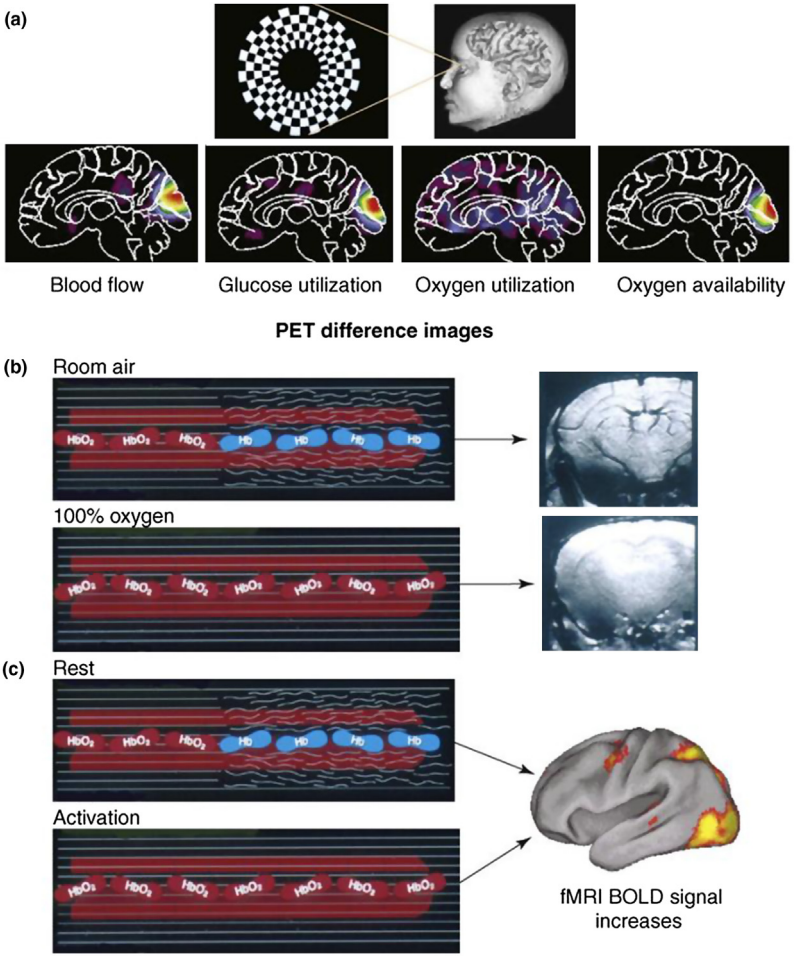

Figure 15 Functional brain imaging with PET or fMRI is based on changes in brain circulation and metabolism that are associated with activity changes in both neurons and astrocytes. The profile of these changes was first detailed by PET (a), which showed that activity increases, in this case evoked by stimulation of the visual cortex by a reversing annular checkerboard, consist of increases in blood flow (49%) and glucose utilization (21%) that far exceed that of oxygen utilization (5%) (images adapted from Fox et al., 1988). As a result, the amount of oxygen available in the brain increases (supply increases relative to demand) causing the relative percentage of deoxyhemoglobin (a paramagnetic substance; Box 1) to decrease. The fMRI signal arises because of this change in the relative amount of deoxyhemoglobin (Hb). As shown in part (b), veins containing Hb (dark areas on the upper image) disappear when the animal is breathed on 100% oxygen (lower image) (original experimental data from Ogawa, S. et al. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proceedings of the National Academic of Sciences of the USA 87, 9868–9872. In the human fMRI experiment (c), Hb is decreased because blood flow increases more than oxygen consumption as in part (a) leading to an enhanced signal portrayed here as a localized increase in the fMRI blood oxygen level density (BOLD) signal. Taken from Raichle, M.E., 2009. A brief history of human brain mapping. Trends in Neurosciences 32 (2), 118–126 (Figure 2, p. 120).

In a second modality, PET can be used to measure the regional utilization of glucose (the biochemical 'fuel' of the brain) or to measure regional oxygen consumption as the glucose is metabolized. Most often, studies of glucose metabolism have utilized fluorodeoxyglucose (FDG) labeled with Fluorine-18 as a substance whose accumulation in brain tissue is proportional to glucose metabolism. Such measurement of local brain metabolism by PET takes between 30 minutes and an hour and yields spatial resolutions on the order of 1–3 cm.

In the third modality, information exchange between neurons is indirectly measured by measuring the influx of fresh blood into areas of the brain in which trans-synaptic transmission has taken place. The assumption (which has been largely confirmed) is that, just as in muscle tissue, local areas of the brain that 'work harder' get more fresh blood.

Common to all modalities is the procedure of injecting a radioactive solution into the experimental subject's veins while she/he rests in the narrow PET tunnel which is surrounded by an array of gamma-ray detectors (see Figure 16). As soon as the substance is injected, it is transported by the circulation of the blood throughout the body and its distribution, binding, metabolism, and/or release from locations in the brain can be monitored as the subject performs the experimental tasks.

Advantages and Disadvantages. The advantages of PET lie in its ability to demonstrate the location over time of radioactively tagged molecules in the living human brain. Up to now, this has been used to show regional changes in the metabolism of the brain correlated with linguistically relevant pathologies (aphasias) and with blood-flow changes due to regional neuronal activity in response to linguistic tasks. The spatial resolution of PET is on the order of one half to a few centimeters and the temporal resolution on the order of 1 or 2 min. Compared to fMRI, acoustic the noise of the PET scanner is very low.

The disadvantages of PET are its high cost, its physically invasive nature (the intravenous injection of relatively harmless radioactive compounds), and the demand that the subject's head be immobilized during the experiment. The short time available during which stimuli can be presented (depending on the half-life of the radioactive compound) is a limiting factor for the design of linguistic experiments.

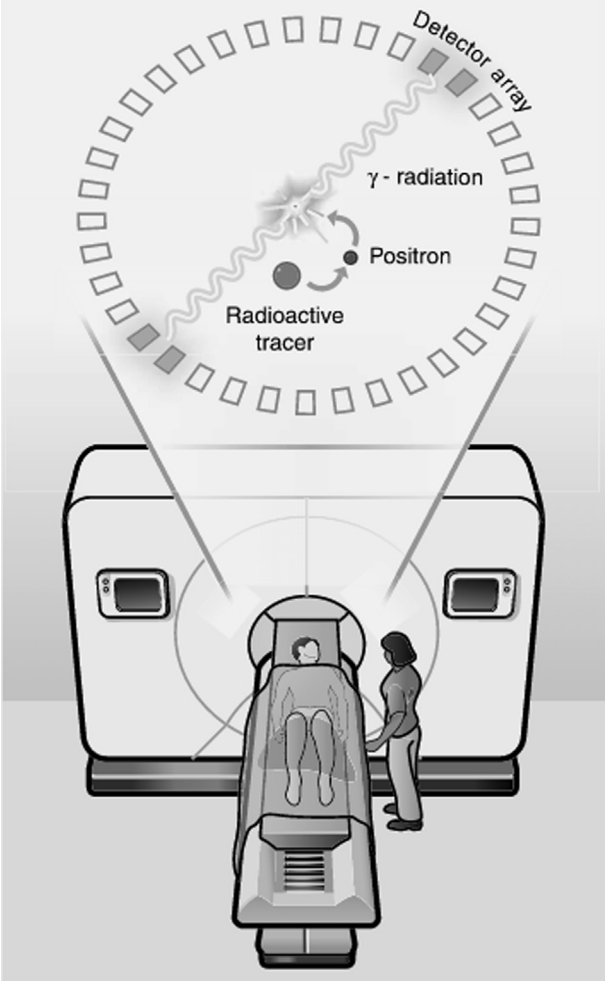

Figure 16 How does PET work? A tracer compound labeled with a positron emitting radionuclide is injected into the participant's circulatory system. She/he is placed within the field of view of an array of detectors capable of registering incident gamma rays. As the radioisotope decays, it emits a positron that subsequently annihilates on contact with an electron. Annihilation leads to liberation of energy in the form of two gamma rays traveling in opposite directions. The gamma rays interact with the scintillating material of the detectors producing a flash of visible light. A photomultiplier tube enhances the scintillations of light so that the signal can be detected. Individual detector units are linked such that two detection events unambiguously occurring within a certain time window may be called 'coincident' and thus be determined to have come from the same radioactive annihilation. Using statistics collected from tens of thousands of coincidence events, a set of simultaneous equations for the total activity of each parcel of tissue can be solved, and a map of radioactivities as a function of location can be plotted. The resulting map shows where the molecular probe has been. Most PET scans in neurolinguistic studies have been 'coregistered' with structural (MRI) scans, to yield the most accurate anatomic information possible. Radionuclides used in PET scanning are typically isotopes with short half-lives such as oxygen-15 (about 2 min). Due to its short half-life, the radionuclides must be produced in a cyclotron which must be not too far away from the PET scanner. The radionuclide is incorporated into compounds normally found in the body such as glucose, water, or ammonia and then injected into the body in order to visualize traces of where they become distributed. The 'neuroimage' is based on an assumption that areas of high radioactivity are associated with brain activity. When oxygen-15 water is used, what is actually measured indirectly is, as is the case with fMRI, the flow of blood – enriched with oxygen-15 containing water – to different parts of the brain. Figure and legend taken from Rodden, Stemmer, 2008. In: Stemmer, Whitaker (Eds.), Handbook of the Neuroscience of Language, p. 63.

Functional Magnetic Resonance Imaging (fMRI)

Similar to PET, fMRI allows three-dimensional representations of locations of altered brain states associated with some task that the brain is carrying out. Another similarity with PET is that identifying brain regions or networks that are involved in a particular task requires comparing the brain activity in response to the particular task (for example, reading words) with one or more baseline/control conditions (for example, the 'resting state'). Inasmuch as it is, however, difficult to control what actually happens 'when the brain is at rest,' the target condition (reading words) can also be compared with a better controlled condition such as reading nonwords. Although this may sound simple, in order to obtain valid results much relies on the ingenuity of the experimenter to find appropriate control conditions and exclude confounding

variables (see Figure 17 for an illustration of an (f)MRI machine).

As previously mentioned, fMRI relies on increased regional blood flow to areas of the brain that are involved in the activity of interest (for example, reading words). Fresh blood carries oxygen bound to hemoglobin, so-called oxyhemoglobin. When the oxygen is extracted from the hemoglobin we talk about deoxyhemoglobin. Due to the increased metabolism in the area of interest, oxygen is extracted resulting in a decrease in oxyhemoglobin and an increase in deoxygenated hemoglobin (deoxyhemoglobin). Things are, however, a bit more complicated. Although deoxyhemoglobin increases, it is proportionally underrepresented. The reason is that more blood than is needed is sent to the area of interest resulting in an oversupply of oxyhemoglobin and thus a decrease in the relative concentration of deoxyhemoglobin per unit volume in the active region. The fMRI signal arises because of this change in the relative amount of deoxyhemoglobin (see Figure 15).

fMRI is not a direct reflection of neuronal activity per se, but rather a reflection of the influx of a particular component of fresh blood (oxyhemoglobin) to sites where nerve cells have very recently fired. The coupling between blood supply and demand in nerve cells is fast but not infinitely fast – it takes time for fresh, oxygenated blood to be re-routed to wash out 'spent blood.'

In fMRI parlance, this indirect but reliable metric of nervous activity is called the 'Blood-Oxygen-Level-Dependent' or 'BOLD' effect. In brief, areas of increases in nervous activity related to changes in brain function can be visualized as areas of increased 'BOLD' activity by the use of fMRI (see Figure 18 for a typical graphical representation of the results of an fMRI experiment).

The images generated are (ideally) depictions of brain activity exclusively associated with some particular cognitive task. This task must be carefully constructed so that it has a distinct beginning and an end. Each task may contain several stimuli such as words. Periods of the target activity (e.g., reading words) are alternated with a control activity (e.g., 'resting condition' or 'reading nonwords condition'). Most commonly, PET experiments use the 'block-design' paradigm whereas fMRI experiments are framed either within the 'block-design' paradigm or the 'event-related' paradigm. Each activity ('block') lasts a fixed amount of time (e.g., 30 s) and is repeated several times. In event-related designs the individual stimuli of the target and the control activity are not summarized in 'blocks,' but each stimulus is presented individually such that the stimuli of the target condition alternate randomly with the control activity. The individual stimuli are presented for a brief period of time such as milliseconds and occur repetitively (see Figure 19).

Figure 17 (f)MRI machine. The radio frequency head coil is being placed on the subject's head. These coils are needed to transmit and receive radio-frequency waves used in MRI scanners. Note that the same machine is used to produce structural (anatomic) MRI and functional MRI (fMRI). Figure courtesy of Siemens Healthcare.

It should be noted that besides BOLD contrast fMRI, there are other forms of fMRI (e.g., arterial spin labeling or blood oxygenation sensitive steady-state) that are, however, rarely used in language research or language rehabilitation and are thus not discussed here.

Advantages and Disadvantages. As with PET, the advantages of fMRI lie in its ability to demonstrate regional nervous activity in the living, behaving human brain. In contrast to PET, fMRI does not require radioactive substances to be injected. The spatial resolution of fMRI is on the order of a few millimeters. Currently, the first machines combing PET and (structural) MRI in *one* machine have become commercially available. They have, however, not been applied to linguistic research yet.

The disadvantages of fMRI are its high cost for the machine (millions of dollars), maintenance, and expert staff, the condition that no magnetically relevant objects are allowed to be present in the subject's body and the demand that the subject's head be rigidly fixed in place during the experiment. fMRI is very loud, which may be a disadvantage for some linguistic experiments. Another disadvantage is its relatively slow temporal resolution on the order of a half second to 5 s.

Diffusion Tensor Imaging (DTI/dMRI) and Tractography

Whereas fMRI allows us to monitor the activity of brain cells, another magnetic resonance imaging technique, diffusion tensor imaging (DTI), provides us with a method to image the fiber tracks/pathways (so-called white matter of the brain) that connect brain cells and through which brain cells communicate. Although a relatively new technique, DTI is rapidly evolving and already a commonly used tool in studies on brain development, disease processes, aging, cognitive neuroscience, and more recently also in language research. DTI provides information on the *structural* (and not necessarily functional) integrity of white matter, and, using tract tracing techniques (tractography), allows one to trace and visualize fiber pathways connecting brain systems. Researchers can thus obtain information on the structural organization of white matter pathways that was previously only possible through brain dissection in the human cadaver. DTI has been used to make brain maps of fiber pathways and investigate their integrity in healthy and patient populations, or correlate them with behavioral and cognitive measures.

DTI measurements depend on the directionally selective (anisotropic) diffusion of water molecules in biological tissues such as fiber tracts. The uninterrupted distance that a water molecule can travel along the length of a nerve fiber is obviously much longer than the distance that it can travel in

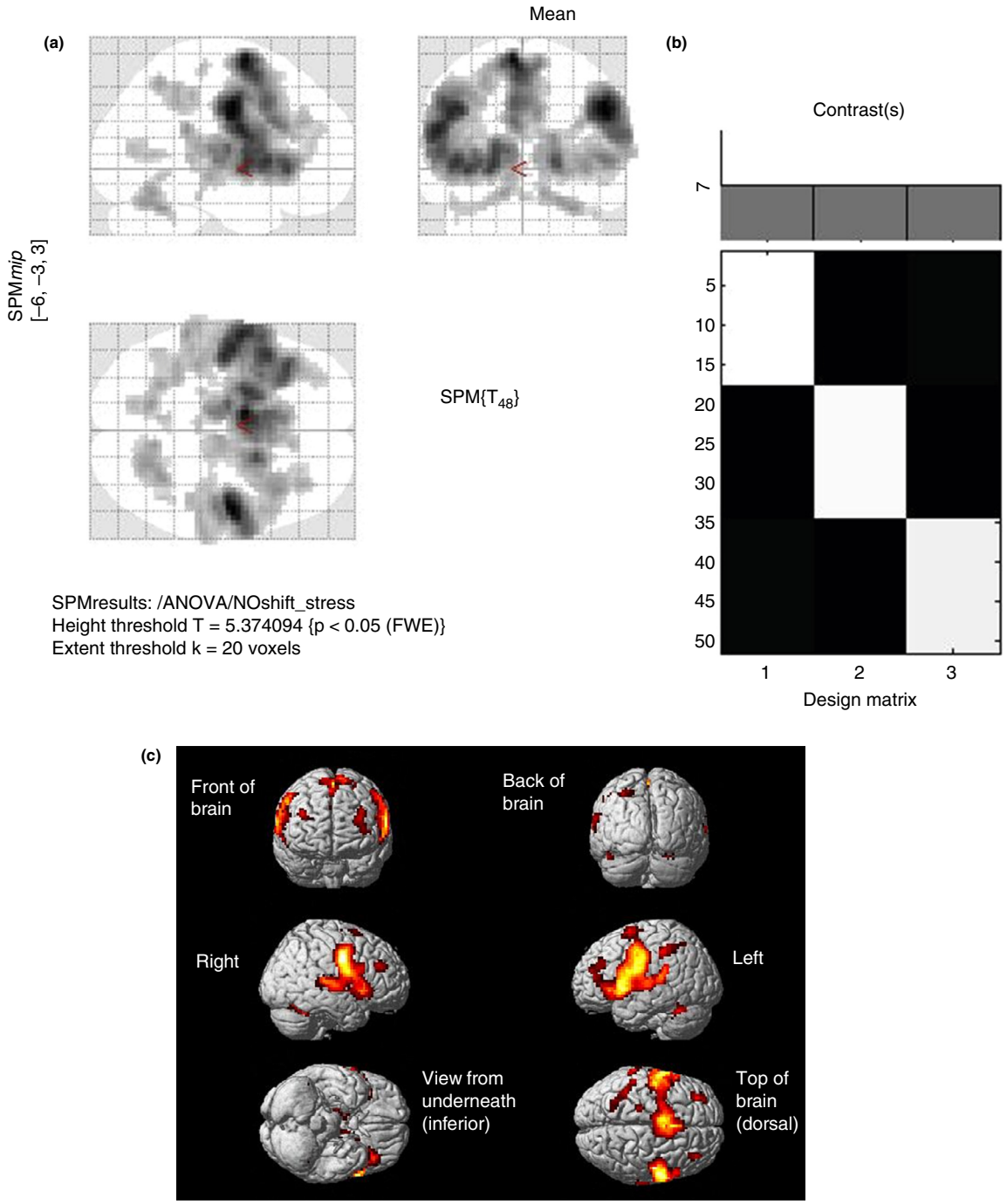

Figure 18 Graphical representation of brain activity measured with fMRI. The overall goal of this research was to investigate brain activity in response to stress patterns in German nouns that were manipulated in various ways. In one condition the participants had to read three-syllable German nouns with varying regular stress patterns (e.g., stress on the first, or second, or third syllable). Reading of these stress patterns was compared to 'reading' non-letter signs, the so-called baseline condition. Other comparisons were also performed but are not shown here. The data were analyzed with the Statistical Parametric Mapping (SPM) software, a common tool used in fMRI analysis. One finding easily seen in the brain image was that activation patterns in the left hemisphere were more widely distributed than in the right hemisphere. (a) Top figure on the left: differences in brain activity are represented as if the brain were transparent ('glass brain') in three different cuts (sagital, coronal, horizontal). (b) Top figure on the right: the SPM-design matrix specifies the statistical model for the group of 18 speakers. Each 'box' specifies a stimulus category (nouns with three different stress patterns) for the 18 speakers. (c) Figure at the bottom: the colorful patches represent the location of voxels that have yielded statistically significant differences when contrasting all overt noun productions with regular stress patterns with the baseline condition. The lighter the color, the higher the T value. The analysis is FWE corrected for multiple comparisons at $p < .05$. Reprinted with thanks to and permission from Bettina Brendel, Universität Tübingen, Center of Neurology, Germany. Copyright Bettina Brendel.

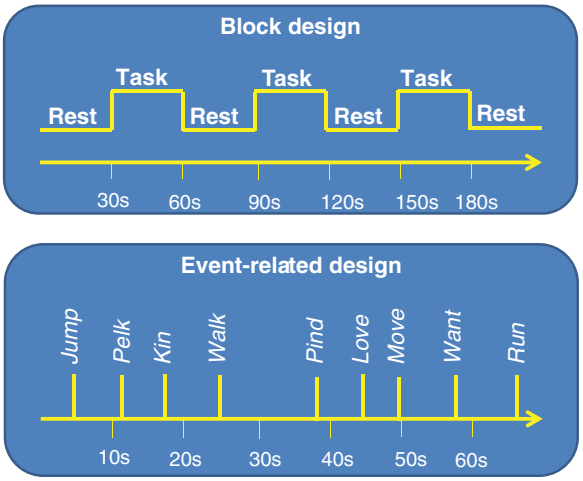

Figure 19 Block design and event-related design. Schematic representation of a blocked and an event-related paradigm. Block design: periods of task performance (e.g., reading words) are alternated with a control task (e.g., rest), each lasting a fixed amount of time (e.g., 30 s). Event-related design: periods of task performance (e.g., reading movement verbs) are alternated with control tasks (e.g., reading nonwords; reading non-movement words). All stimuli are shown pseudorandomly during a limited amount of time (e.g., 500 ms). Copyright B. Stemmer.

a direction orthogonal to that length and DTI exploits this difference. With DTI, it is possible to measure the diffusion of water molecules along various axes and map regions of the brain and thus determine orientation and properties of nerve fibers.

A technique similar to DTI is 'diffusion-weighted imaging' (DWI). DWI portrays the differential rates of water diffusion in microareas of the brain. DWI is most effective in brain regions in which water molecules can move freely in any direction; such regions are found primarily in the gray matter, ganglia, and brain nuclei. Up to now, DWI has proven most useful in the diagnosis of vascular pathologies of the brain such as those found in stroke patients.

Advantages and Disadvantages. DTI and DWI can be used to uniquely investigate the organization and microstructure of white matter in the healthy brain and they can detect white matter abnormalities caused by disease. DTI has also been used in studies of brain development and maturation.

DTI can only trace the path of white matter tracts but not the direction of signal transduction. Its imaging resolution is about 2 or 3 mm in humans and DTI may thus not 'see' certain white matter tracts. There are also problems related to fibers crossing which DTI may not be able to distinguish (for a more technical description of the method see Mori and Zhang, 2009).

Investigating Speech and Language with PET, fMRI, or DTI

A large number of studies have been published investigating various aspects of speech and language with neuroimaging techniques. In 1988 Petersen et al. (1988) published the first language study using PET. They described brain areas which were active during auditory and visual word processing in healthy subjects and proposed a functional–anatomical model for word processing. These early PET studies had a tremendous impact as they showed that functional imaging could provide information about anatomical locations in relation to behavior in healthy individuals. Until then, researchers had had to rely on studies with brain-damaged patients (so-called brain lesion studies). The first fMRI studies on language were published in 1993 and PET (due to its radioactive component) was pushed into the background. Today, the majority of language studies use the fMRI technique with the number of DTI studies and the combination of the different imaging techniques on the rise. In the following section we will provide examples to illustrate the application of the previously discussed technologies in language research and language rehabilitation.

Investigating Lexico-Semantic Processing

Most functional imaging studies with fMRI have reflected an effort to elucidate the brain regions or networks implicated in specific aspects of language processing in healthy and pathological populations. Early PET and fMRI studies essentially replicated and extended the findings reported by Petersen and colleagues (Petersen et al., 1988; Petersen and Fiez, 1993). Since then the functional-anatomical word processing model has been extended and modified by the large number of neuroimaging studies that followed. What has been accomplished after 20 years of neuroimaging language? Cathy Price at the University College of London has undertaken the enormous task of describing these accomplishments and she has summarized the major findings related to speech, spoken language, and reading at the word level (Price, 2010, 2012; see **Figures 20** and **21**).

When we produce the sounds of speech not only those regions of the brain that are implicated in movement (sensorimotor cortex) become active but also regions that are involved in control and timing processes (such as the anterior cingulate, subcortical areas, and the cerebellum). Hearing sounds and recognizing them as speech has been associated with the left and right superior temporal gyri.

Reading words implicates lexico-semantic processing and a lexico-semantic reading route has been described that involves neural substrates different from those detected in a nonsemantic phonological decoding route. In addition, skilled reading requires the extraction of visual features and thus regions in the occipitotemporal cortex are activated.

When we assign meaning to a word we need to access and retrieve knowledge we have about the world. The neural systems underlying storage and retrieval of semantic knowledge are widespread in the human brain and include large regions of the cortex in the parietal, prefrontal, and subcortical regions of the brain (more specifically, posterior heteromodal assocation cortex, specific subregions of heteromodal prefrontal cortex, and subcortical areas such as medial paralimbic regions) (see **Figure 22**). Heteromodal (also called supramodal) cortex refers to cortex that receives information from several modalities (e.g., visual, auditory, gustatory) and not just a single modality.

Investigating Syntactic Processing

Although it is possible to communicate with single words or utterances only, most of the time we string together words to form sentences and larger text or discourse units. The

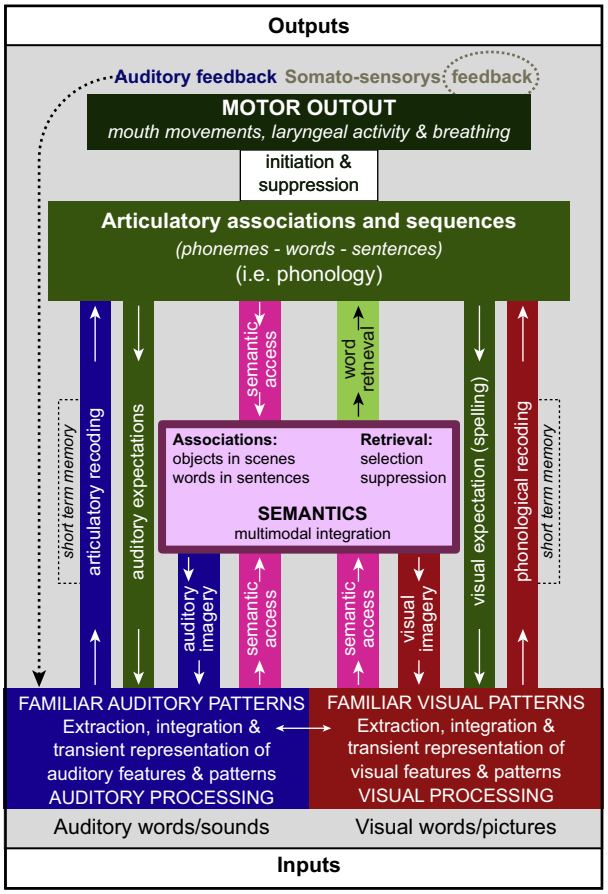

Figure 20 Functional model based on neuroimaging studies of language. Figure taken from Price, C.J., 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage 62, 816–847 (figure on p. 820).

question we can thus ask is which brain areas or networks are implicated in sentence processing. Lesion studies have pointed to Broca's area in the inferior frontal gyrus of the left hemisphere. Generally, most neuroimaging studies seem to support this view, however, in a somewhat modified version. Numerous neuroimaging studies have associated syntactically *complex* sentences with regions that include Broca's area but that also extend to regions surrounding the left inferior frontal gyrus. Some studies have suggested that different regions *within* the left inferior frontal gyrus are associated with various aspects of syntactic processing. Other studies have also implicated regions in the left temporal lobe in syntactic processing. Again others have associated bilateral frontal and posterior temporal regions with processes that integrate information into a sentence structure. Due to these generally diverse findings or interpretations, one possibility is that several different and distributed networks are implicated in syntactic processing. Another possibility is that the distributed neural architecture may be obscured by the way neuroimaging data are analyzed. It has, for example, been suggested that instead of relying on group-based analysis methods it may be more beneficial to choose an 'individual-subject's functional localization approach' to account for anatomical variability seen in individual subjects (Fedorenko et al., 2012).

It is noteworthy that the regions related to syntactic processing have also been shown to be activated in other linguistic as well as nonlinguistic cognitive tasks thus raising the question whether or to what degree other cognitive functions overlap with or support linguistic functions.

Another issue that has raised concern is the difficulty of clearly distinguishing syntactic and semantic processing components inasmuch as fMRI studies always contrast different task conditions and at least one of them contains not only structural but also has semantic elements. For example, neuroimaging studies often compare meaningful complex sentences (e.g., *Sophia, who had a broken leg, was given a lift home by Peter*) with syntactically simpler sentences (*He gave her a lift home*) or with nonmeaningful utterances (*Mo soodled lorn en saweek*).

It has been mentioned previously that when assigning meaning to a word the semantic system (also called semantic memory) is activated. The same holds true for comprehending sentences, or, more globally, language. It is assumed that semantic memory stores our abstract, generalized conceptual knowledge without referring to any specific experience. This store is called upon when we try to make sense of words, sentences, or discourse. Generally, fMRI studies comparing brain regions that respond to syntactic tasks with brain regions that respond to semantic tasks showed a large overlap of brain regions that become active during syntactic as well as semantic processing. Functional neuroimaging studies suggest that large areas of the heteromodal association cortex are involved in semantic memory processes and that information is integrated in regions in the temporal and inferior parietal lobes. Another view is that the temporal pole is the principal site of information convergence. (For meta-analyses of neuroimaging studies that investigate language processing see Vigneau et al., 2006, 2011. For a summary on neuroimaging and syntactic processing see Fedorenko et al., 2012. For a summary of the semantic system see Binder et al., 2009; Binder and Desai, 2011).

For a summary on studies investigating language processing in the left and right hemispheres.

Investigating Fiber Tracts Connecting Brain Regions Associated with Language

Until recently, most neuroimaging studies of language have been rather localization oriented and focused on identifying specific brain regions that are activated during task performance. This, however, is only part of the story. Neurons communicate via fiber bundles and to obtain a more complete picture about language processing, it is necessary to identify and elucidate the functional significance of these fiber bundles. Fiber bundles connecting brain regions associated with language were discussed as early as the nineteenth century (for a review see C. Weiller et al., 2009, 2011). A special role was assigned to the arcuate fasciculus, a large fiber bundle running between Broca's and Wernicke's region. Due to advances in MRI technology, fiber bundles can now be made visible by DTI (see **Figures 23** and **24**).

The frontal cortex and the temporal cortex are connected via several fiber bundles. The fiber bundles run in two main

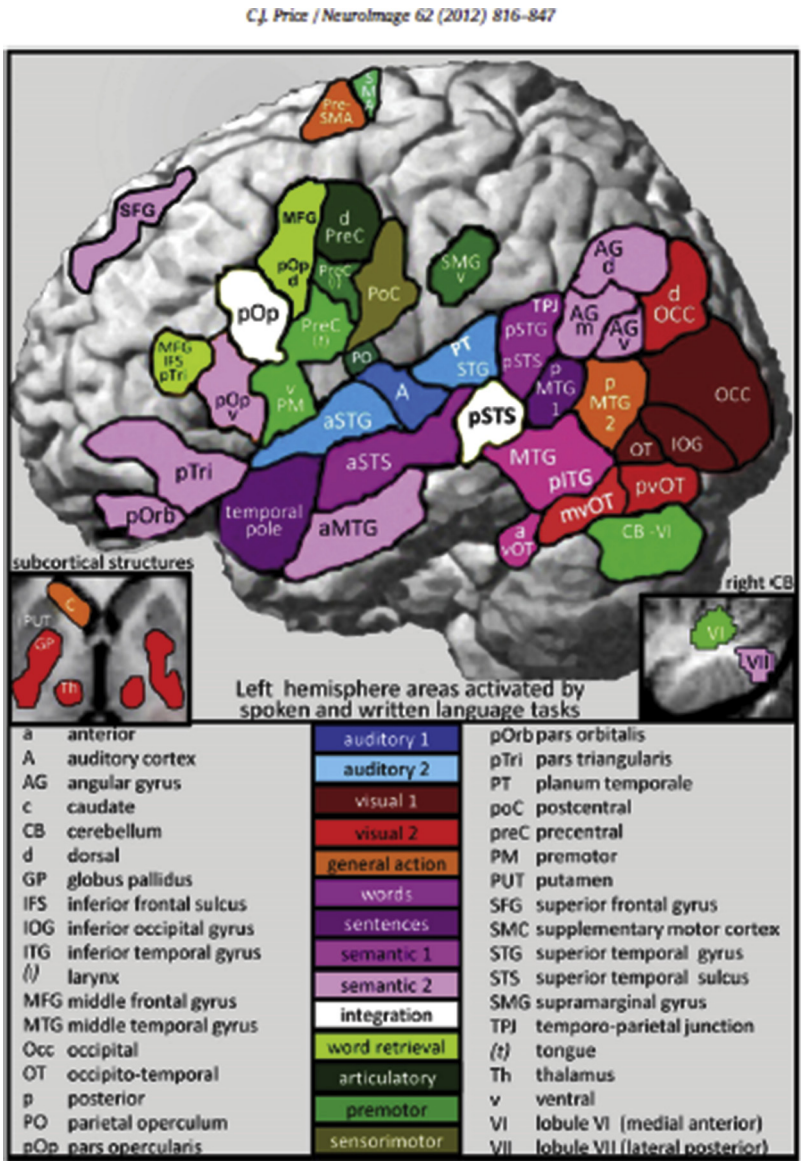

Figure 21 Left hemisphere areas activated by spoken and written language tasks. Left hemisphere areas activated by spoken and written language tasks. Illustrative sketch of the location of language-related activations, based on data from Price et al. This figure was created by overlaying images of activations from many different studies that C.J. Price has coauthored. The colors indicate the type of task or processing that caused the activation. The blue areas are activated by auditory stimuli (auditory 1 followed by auditory 2). The red/brown areas are activated by visual stimuli (visual 1 then visual 2). The orange areas are activated by general action selection (hand or speech). The pink and purple areas are involved in different levels of semantic and syntactic processing depending on the task demands. The dark pink areas are sensitive to the semantic content of the stimuli, the light pink areas are those that are more activated for semantic than for phonological decisions. The light purple areas are activated by spoken and written sentences with the dark purple areas most activated by meaningful and grammatical sentences. The green areas are involved in generating or rehearsing speech. The light green areas are involved in word retrieval; the dark green areas are involved in silent phonological decisions. The khaki green area and PreC/vPM areas are activated by mouth movements during speech. Finally the white areas, corresponding to Broca's area (pOp) and Wernicke's area (pSTS) are involved in both perception and production tasks with familiar stimuli. They may function as convergence zones that receive and send signals to all the other areas involved in perceiving and producing speech. The connections between areas are not shown because we do not yet know enough about how all the areas connect to one another. Figure and legend taken from Price, C.J., 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage 62, 816–847 (figure on p. 821).

pathways relevant for syntactic and semantic processing: the dorsal ('upper') pathway and the ventral ('lower') pathway. The dorsal pathway runs above and the ventral pathway below the Sylvian fissure which separates the frontal from the temporal lobe (see Figure 25).

Two large fiber bundles of the dorsal pathway have been identified that connect the left (posterior) temporal lobe with the left (posterior) frontal lobe: the arcuate fasciculus and the superior longitudinal fasciculus. Various suggestions have been made for the functional significance of the dorsal

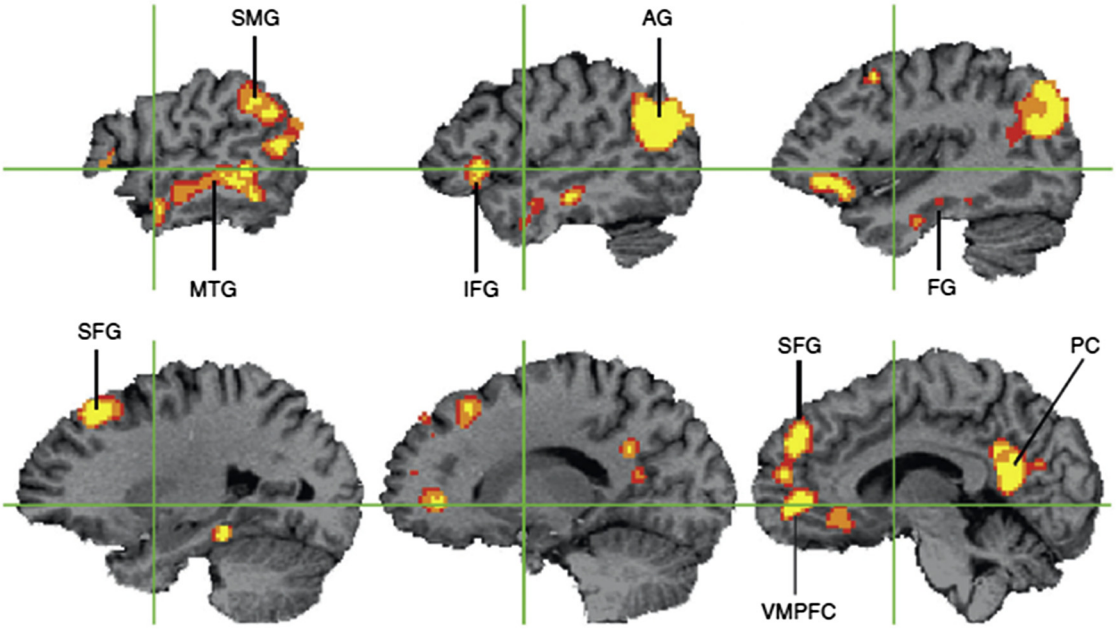

Figure 22 Neural substrates underlying semantic processing of words as identified by neuroimaging studies. Meta-analysis of functional imaging studies of semantic processing. This figure displays brain regions reliably activated by general semantic processes, based on reported activation peaks from 120 independent functional imaging studies ($p < .05$ corrected for family-wise error). The analysis method assigns a significance value to the degree of spatial overlap between the reported activation coordinates in a standard volume space. The figure shows selected sagittal sections in the left hemisphere; right hemisphere activations occurred in similar locations but were less extensive. AG = angular gyrus, FG = fusiform gyrus, IFG = inferior frontal gyrus, MTG = middle temporal gyrus, PC = posterior cingulate gyrus, SFG = superior frontal gyrus, SMG = supramarginal gyrus, VMPFC = ventromedial prefrontal cortex. Green lines indicate the *Y*- and *Z*-axes in standard space. Adapted from Binder, J.R. et al. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cerebral Cortex 19, 2767–2796. Taken from Binder et al., 2011. The neurobiology of semantic memory. Trends in Cognitive Sciences 15 (11), 530.

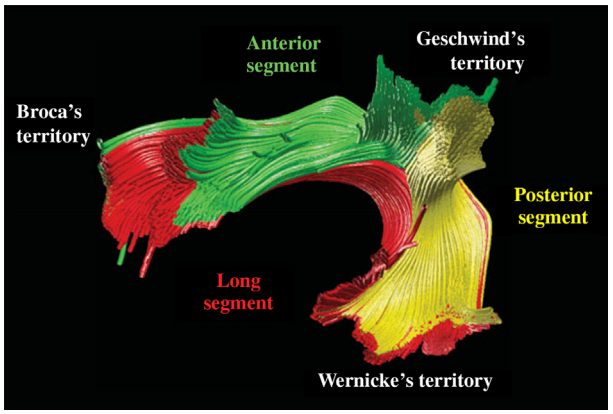

Figure 23 Tractography reconstruction of the arcuate fasciculus using the two region of interest approach. Broca's and Wernicke' territories are connected through direct and indirect pathways in the average brain. The direct pathway (long segment shown in red) runs medially and corresponds to classical descriptions of the arcuate fasciculus. The indirect pathway runs laterally and is composed of an anterior segment (green) connecting the inferior parietal cortex (Geschwind's territory) and Broca's territory and a posterior segment (yellow) connecting Geschwind's and Wernicke's territories. Figure and legend taken from Catani, M., Jones, D.K., ffytche, D.H., 2005. Perisylvian language networks of the human brain. Annals of Neurology 57 (1), 8–16 (Figure 3, page 11).

pathway: mapping sound onto articulatory based representations, being implicated in speech repetition, processing complex syntactic operations, and serving higher-level semantic and language function. Ventral pathways connect the left temporal cortex with the left prefrontal cortex via the extreme fiber capsule system (also called the inferior fronto-occipital fasciculus) and the uncinate fasciculus. It has been suggested that these pathways support the mapping of sound to meaning, the acquisition of natural grammar, and syntactic processing; the uncinate fasciculus has especially been related to semantic processing. Another left lateralized pathway connects posterior Broca's region with areas in the medial frontal cortex (presupplementary motor area): the frontal aslant tract. Alterations in the aslant tract, the uncinate fasciculus, and the arcuate fasciculus have been described in patients suffering from primary progressive aphasia, a progressive breakdown of language functions. In these patients semantic impairment was correlated with degeneration of the uncinate fasciculus, grammar deficits with degeneration of the arcuate fasciculus, and deficits in verbal fluency with the frontal aslant tract, thus suggesting a link between these fiber tracts and the observed language behavior (M. Catani et al., 2013) (for details on language-relevant fiber bundles see Catani et al., 2005, 2013; Catani and Mesulam, 2008; Friederici and Gierhan, 2013; Hickok and Poeppel, 2004; Saur et al., 2008; Weiller et al., 2009).

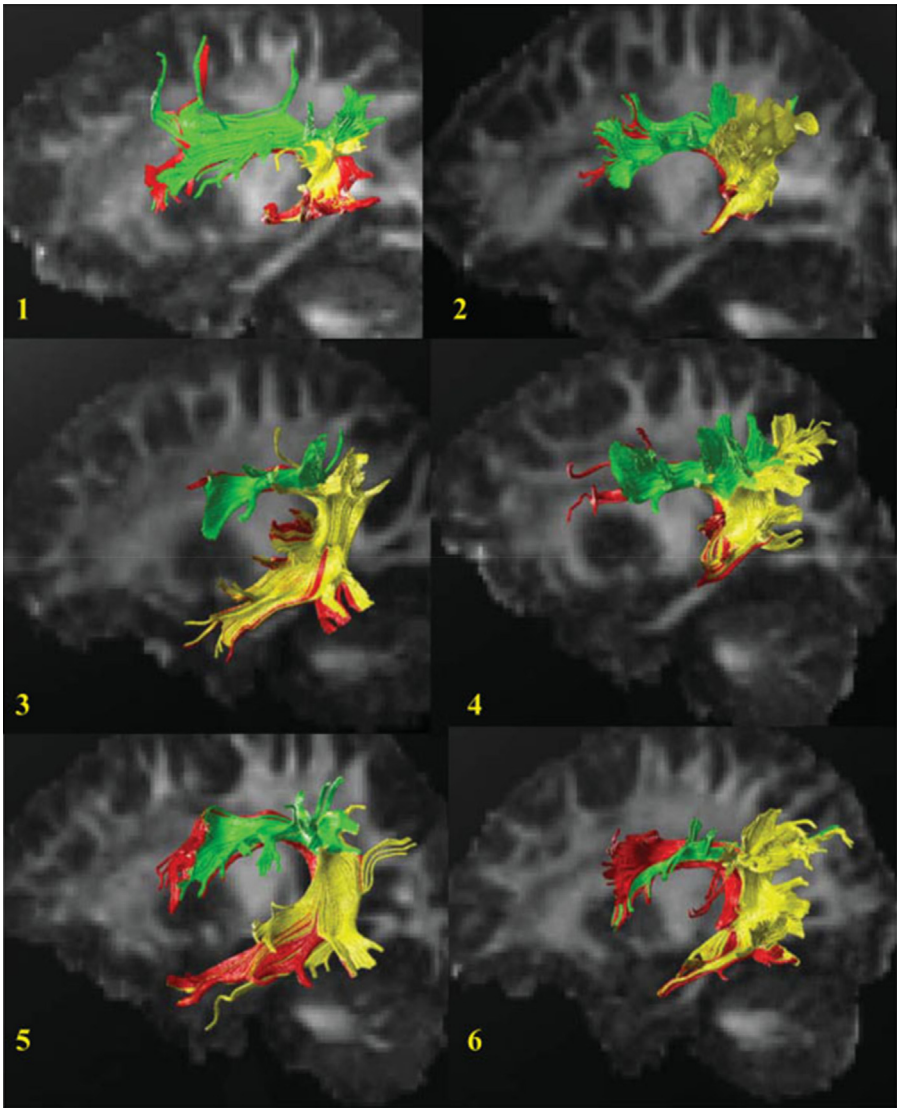

Figure 24 Tractography reconstructions of the direct and indirect pathways in six individual subjects superimposed on sagittal fractional anisotropy data. The three segments are color coded using the conventions of **Figure 23**. Both direct and indirect pathways are present in each subject, but an intersubject variance is evident. Figure and legend taken from Catani, M., Jones, D.K., ffytche, D.H., 2005. Perisylvian language networks of the human brain. Annals of Neurology 57 (1), 8–16 (Figure 4, p. 12).

The Default Language Network

Numerous fMRI studies using tasks targeting similar language functions at the word or sentence level have reported activation in similar brain regions. These brain regions have been referred to as the default language network. It is important to note that the default language network is not very well defined and represents only a rough approximation at a very general level of the neural substrates implicated in word- and sentence-level processing. Note that not all researchers adhere to the view that such a default network exists.

In infants as young as 2 days old the language-related neural substrates are already fully active in both hemispheres although they are still functionally and structurally immature (Perani, 2012). By the age of 6, maturation processes still continue as shown with a functional connectivity MRI approach (i.e., analyzing low-frequency fluctuations) that examined the default language network underlying sentence processing in adults and children (Friederici, 2011). In adults, the default language network showed a strong connectivity *within* one hemisphere, i.e., between left frontal and left temporal language regions. These connectivity patterns were not seen at the age of 6 years in children, although, by that age, the classical language regions are functional. Instead, children showed a strong functional connectivity *between* the hemispheres (especially between superior temporal regions). These findings suggest that the default language network differs in adults and children of this age. An alternative interpretation would be that the default language network does not yet exist in children.

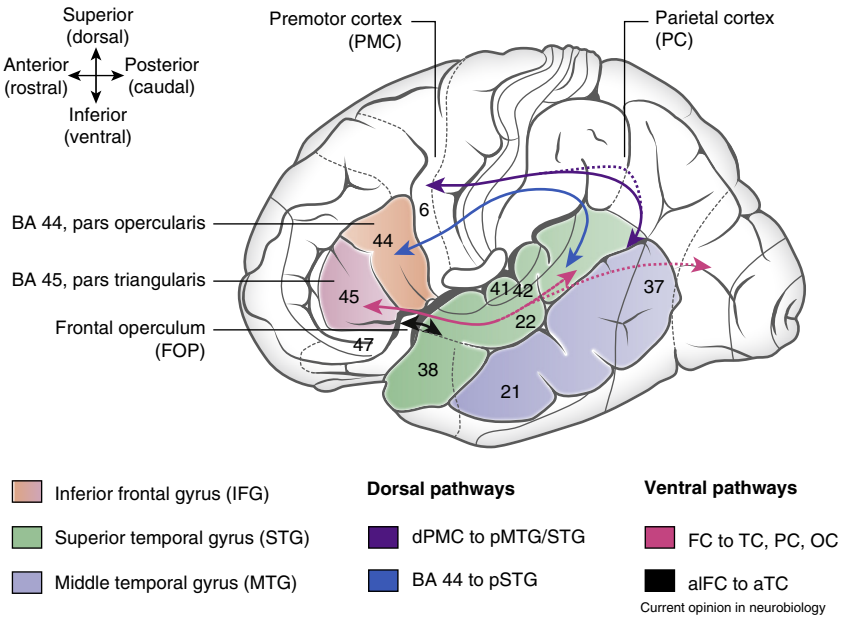

Figure 25 Language-relevant brain regions and fiber tracts (schematic and condensed view of the left hemisphere). The dorsal pathway connecting dorsal premotor with posterior temporal cortex (pMTG/STG) involves the superior longitudinal fasciculus (SLF) III and/or the SLF II and the SLF-temporo-parietal; the dorsal pathway connecting BA 44 with the posterior STG involves the AF. The ventral pathway connecting the frontal cortex (FC), that is, BA 45 and others, with the temporal (TC), the parietal (PC), and the occipital (OC) cortex, involves the IFOF (also called the ECFS); the ventral pathway connecting the anterior inferior FC (aIFC), that is, BA 47 and others, and the FOP, with the anterior TC (aTC), involves the uncinate fasciculus (UF). Figure and legend taken from Friederici, Gierhan, 2013. The language network. Current Opinion in Neurobiology 23, 250–254 (Figure 1, p. 251).

Investigating Neural Systems Underlying Linguistic Pragmatic Behavior

Linguistic pragmatics views language as action – interpreting utterances, texts, and discourse through its use in context. Considering that linguistic pragmatic behavior operates on words, it is not surprising that language networks active at the word and sentence level (possibly the default language network) are implicated. The question is, thus, whether pragmatics relies on these networks alone, or whether additional networks are implicated.

A meta-analysis of fMRI studies addressing text comprehension reported activation of brain regions that are typical for the default language network (e.g., left inferior frontal regions, anterior temporal lobe in both hemispheres, and left-dominant middle and posterior temporal regions). In addition, regions outside but still adjacent to the default network were activated and were termed the 'extended' language network (Ferstl, 2010; Ferstl et al., 2008). There is currently no agreement amongst researchers on the specific function of these 'extended' networks.

The situation is similarly unclear concerning the neural substrates underlying figurative language such as metaphors, idioms, irony, or sarcasm. Studies with patients who had damage to the right hemisphere and showed problems with figurative language gave rise to the classical right hemisphere hypothesis that stated that figurative language is processed in the right hemisphere. There is currently little evidence from fMRI studies supporting this hypothesis. What seems to emerge is that properties of the figurative language concept such as degree of figurativeness, familiarity, conventionality, as well as cognitive processing difficulty may determine the degree to which the right hemisphere is implicated, with cognitive processing difficulty possibly playing a particular role. Cognitive and psycholinguistic approaches have grounded figurative language in bodily experience and linked them to the sensory–motor system. The very few fMRI studies that have addressed this issue seem to support the idea, at least when the concepts investigated contain some sort of action components.

Finally, it has been suggested that pragmatics is an emergent phenomenon and – apart from the default language network – the neural substrates implicated may be determined newly every time figurative language occurs. This may happen because of the specific characteristics of the figurative language concept, the cognitive systems, mental state and experience specific to the individual, the environment in which these concepts occur, and the way these variables are weighted (for a summary on neuroimaging studies of pragmatics see Stemmer, 2008, 2013).

Investigating Neural Substrates Underlying Bilingualism

The question of how language is processed in people who master more than one language has raised considerable interest not only in academics but also in the general population and at a political level.

Due to the constraints of the imaging methods, studies in bilingualism usually use different experimental designs than behavioral methods thus making comparisons difficult. It has also repeatedly been pointed out that it is difficult to obtain a clear picture inasmuch as not all studies control important variables, or they are heterogeneous in those

variables such as age of acquisition of a second language; proficiency in the language; frequency and context of use of the language; and stimuli, tasks, and experimental design. These limitations should be kept in mind during the following discussion.

Numerous studies have addressed the question of whether the first and second language are structurally represented or functionally processed similarly or differently in the brain. The picture that emerges is heterogeneous. Some studies reported that brain regions activated by two or more languages overlapped while others reported that language tasks in bilinguals activated different or partially different brain regions. One meta-analysis of 24 fMRI studies with bilinguals reported stronger activation and recruitment of common first/second language regions during the processing of a second language when the second language was learned later in life, or second language proficiency was low, or the frequency of exposure was low (Indefrey, 2006).

Several fMRI studies have addressed the question of whether bilinguals have an advantage over monolinguals on tasks measuring executive functions (see Section ERPs in Bilingualism and Second Language Research).

An fMRI study comparing mono- and bilinguals reported that despite similar behavioral performance, bilinguals and monolinguals exhibited activity in different brain regions in tasks that required selecting a response and suppressing interfering material (Flanker task and No-go tasks) (Luk et al., 2010). The same brain regions were, however, recruited by both language groups when a response had to be inhibited. The findings were interpreted as indicating that bilingualism modified the neural networks responsible for achieving effective general cognitive control of nonverbal stimuli. Other fMRI studies reported better adaptation to the conflict task by bilinguals compared to monolinguals and associated this with less activity in a specific brain region (anterior cingulate cortex). fMRI studies investigating whether bilinguals have an advantage switching from one task to another reported that monolinguals demonstrated a switch cost (reflected as an increase in reaction time). Monolinguals and bilinguals differed in the recruitment of brain regions in response to the switch and nonswitch tasks. A task-switching paradigm was also used in a recent fMRI study that addressed the question of whether lifelong bilingualism offsets age-related declines in neural efficiency for cognitive control processes (Gold et al., 2013). Behavioral data showed that older adult bilinguals switched faster between perceptual tasks than monolinguals. The fMRI data revealed that compared to monolinguals, bilingual older adults outperformed the monolingual peers while showing a decrease in activation in the left lateral frontal and the cingulate cortex. This was interpreted as suggesting that lifelong bilingualism offsets age-related declines in neural efficiency during specific cognitive control processes (for a summary on neuroimaging studies investigating bilingualism and second language acquisition see, for example, Bialystok et al., 2012; Hilchey and Klein, 2011; Indefrey, 2006; Kousaie and Phillips, 2012; van Heuven and Dijkstra, 2010).

For a discussion of language dominance in each hemisphere (*see* Lateralization of Language as Demonstrated by Brain Imaging Procedures).

Generally, the few fMRI studies that have been carried out show that neuroimaging may reveal differences in response to cognitive control tasks that are not apparent in behavioral measures. Further, it seems that fMRI data support behavioral findings that describe a bilingual advantage for cognitive control. Similar to the MEG study reported in the Section ERPs in Bilingualism and Second Language Research, however, one may ask whether the mere findings of some studies that functionally distinct brain regions were involved in the task performance of bilinguals compared to monolinguals justifies the interpretation that these regions are associated with inhibitory control and are responsible for the bilingual advantage.

Investigating the Neurochemical Bases of Speech Production

The findings reported in the previous sections have used PET, fMRI (including special analysis techniques to obtain functional connectivity patterns), or DTI methods. Language studies that combine various techniques (such as fMRI and EEG) are on the rise although 'combining techniques' does not necessarily mean simultaneous application of the methods. A combination that has been used less frequently is fMRI and PET possibly due to the high costs and logistics involved. One study combining PET, fMRI, and DTI techniques investigated the relationship between speech production and dopamine release in healthy subjects (Simonyan et al., 2013). Investigating the neurotransmitter dopamine during speech production was motivated by the clinical observation that alterations in the dopamine system may lead to speech and language impairments (e.g., inability to control motor aspects of speech production; difficulties with phonological and syntactic processing; and so on). Data were acquired in separate PET and fMRI sessions while subjects listened to grammatically correct sentences or while they were resting. PET was used with a dopamine receptor radioligand to study dopamine release. fMRI was used to investigate activity patterns in a specific subcortical brain region (the striatum which is known to be implicated in movement control) and to study functional connectivity patterns. DTI was obtained to examine the structural network organization in the striatum. The authors reported predominantly left lateralized dopamine release within specific regions of the striatum during sentence production. Dopamine release seemed to influence the lateralization of functional speech networks but not of structural speech networks.

Although the study has limitations (e.g., choice of stimuli and experimental conditions: speech versus resting state; separate PET and fMRI sessions; method of analyses), the study nevertheless demonstrates what is currently possible. In this context it is noteworthy to mention that the first 'all-in-one' PET-(f)MRI machines are now commercially available.

Language Rehabilitation and Recovery

An important issue in language rehabilitation concerns the cerebral structures or networks implicated in recovery of function such as in individuals suffering from aphasia. Brain repair and reorganization depend, at least in part, on the site and extent of the brain lesion, and the recruitability of the partially

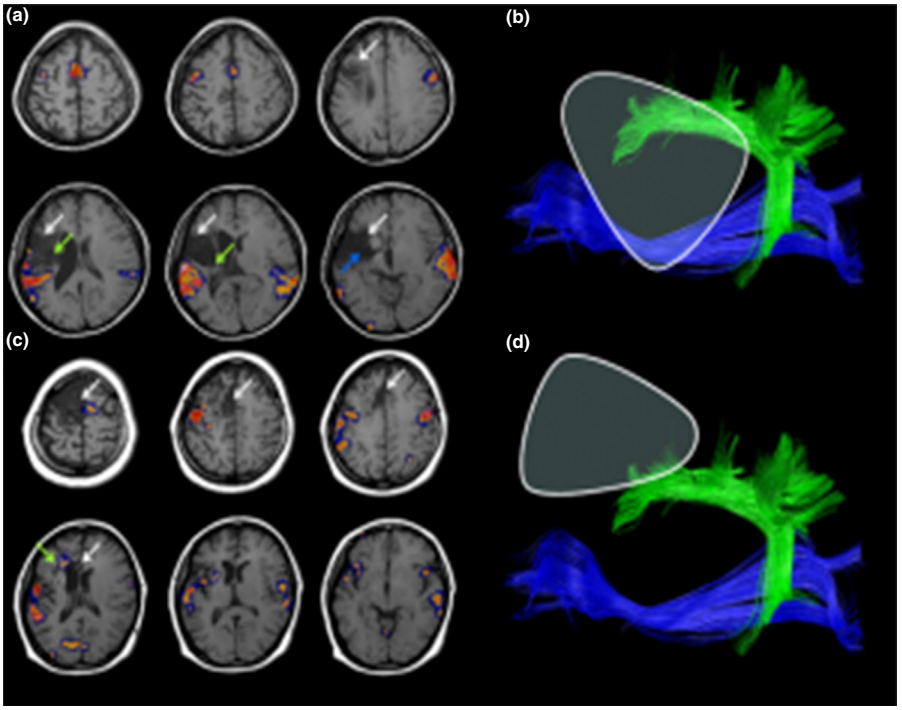

Figure 26 Dissociated performance on repetition performance of nonwords in chronic aphasia according to the site and extent of structural damage. (a) MRI of a patient with chronic nonfluent aphasia and *impaired* repetition (Broca's aphasia). A large area of infarction is seen involving the left frontotemporal cortex with extension into insula, basal ganglia (white arrows), and different segments of dorsal (arcuate fasciculus) (green arrow) and ventral pathways (blue arrows) crucial for language repetition. Despite significant bilateral activation of temporal and frontal lobes and supplementary motor areas on fMRI during nonword repetition, the patient's performance on this task was severely impaired (27/104). (b) Schematic diagram depicting the area of infarction (gray triangle) which involves both the dorsal (green) and ventral (blue) pathways underlying language repetition. (c) MRI of a patient with chronic nonfluent aphasia and *preserved* repetition (transcortical motor aphasia). A large area of infarction is seen on the left mesial frontal cortex (white arrows) with minimal compromise of the anterior portion of the dorsal pathway (green arrow). Patient's performance on nonword repetition was preserved (95/104) and the fMRI during nonword repetition shows significant activation of a bilateral phonological language network similar to the one activated in normal control subjects. (d) Schematic diagram depicting the area of infarction (gray triangle) which minimally involves the anterior part of the dorsal pathway (green) with sparing of the ventral pathway (blue). The left side is represented on the left side of the MRI. Figure and legend taken from Berthier M.L., García-Casares N., Walsh S.F., Nabrozidis A., Ruíz de Mier R.J., Green C., Dávila G., Gutiérrez A., Pulvermüller, F., 2011. Recovery from post-stroke aphasia: lessons from brain imaging and implications for rehabilitation and biological treatments. Discovery Medicine 12 (65), 275–289.

damaged and nondamaged structures to contribute to language processing (see Figure 26).

fMRI has been used to identify neural activity patterns that allow predictions about the recovery process. In individual patients it has been shown that increased activation foci in the intact right hemisphere during language task performance very early after the stroke predicted favorable long-term outcomes. As most studies rely on a small sample size, an international database is currently collecting recovery patterns of poststroke aphasic patients with the objective of allowing outcome prediction based on a large database.

Only few studies have used MRI/fMRI or DWI to predict the degree of recovery in the very early *acute* stage poststroke. These studies suggest that recovery of language function depends on three basic mechanisms: (1) reperfusion of incompletely damaged tissue around the lesion (perilesional); (2) resolution of edema at the site of the lesion and improvement of reduced metabolism in regions distant from the site of the lesion, and (3) reorganization of the relationship between structure and function (Berthier et al., 2011).

Most brain imaging studies have investigated the neural processes underlying recovery from *chronic* poststroke aphasia. Late recovery mechanisms seem to rely on compensation (i.e., recruitment of additional brain areas) and reorganization (i.e., displacement of a primary language area). The potential capability of the brain to compensate for lesions has been referred to as 'neural plasticity.'

The role of the right (nondamaged) hemisphere in the recovery process is still a debated issue. While some have reported that both hemispheres contribute to network reorganization in chronic poststroke aphasia (Turkeltaub et al., 2012), others have associated better recovery and less right hemisphere involvement with small- and medium-sized lesions in the left hemisphere (Berthier et al., 2011). An increase in right hemisphere activity during language tasks as measured by PET or fMRI has been viewed as maladaptive and impeding recovery, or as reflecting other cognitive strategies, or as inhibitory on the remaining left hemisphere functions.

fMRI has also been used to study the effect of medication in chronic poststroke aphasia. Treatment-induced decreased

activity in fMRI, for example, was interpreted as reflecting decreased effort and more efficient processing of the task. (For a summary on the role of the two hemispheres in the recovery process of aphasic patients; for a summary on neuroimaging studies in language rehabilitation see Berthier et al., 2011; Meinzer et al., 2011; Turkeltaub et al., 2011.)

Functional Near-Infrared Spectroscopy

The pioneering work on the application of near-infrared spectroscopy (NIRS; formerly known as diffuse optical imaging) in humans goes back to 1977. While the early systems were confined to using a single channel (i.e., one near-infrared source and one detector), by the mid-1990s multichannel NIRS systems were available thus making it possible to explore the 'functional' working of the brain with NIRS – hence the name 'fNIRS.' Unlike methods such as PET or fMRI, NIRS is better suited to simultaneously provide information about the source of neural activity (*where* something happens) as well as its time course (*when* something happens). fNIRS measurements have been shown to be highly consistent with fMRI findings and fNIRS results have been reproduced reliably over time (for a summary and history of NIRS see Ferrari and Quaresima, 2012).

Amazing but true – simply by shining light on the surface of the brain and by recording and analyzing the reflected light patterns one can obtain maps of the functional architecture of the brain. Similar to other hemodynamic techniques (PET, fMRI), fNIRS depends on the assumption that those regions of the brain that are particularly active demand more freshly oxygenated blood than the surrounding brain tissue. Hence, oxygenated blood (more precisely hemoglobin in the red blood cells carrying oxygen) rushes to the areas associated with the task where the oxygen is used. While there is consensus that the concentration of oxygenated blood (HbO_2) can be viewed as an indicator of brain activity in fNIRS, there is an on-going discussion over which of the parameters, HbO_2, total hemoglobin (HBT), or deoxygenated hemoglobin (HbR), reflects activation best.

When shining light of an appropriate wavelength (in this case infrared light) on the head, this light penetrates tissues and bone and is absorbed by hemoglobin to various degrees. Light reflected after this absorption is measured by a set of probes (near-IR emitters and detectors) that are attached on the scalp (see **Figure 27**), and from these measurements changes in the brain's activity can be estimated. In more sophisticated terms, as a light source NIRS uses the wavelengths of light in the near-infrared range (700–1000 nm) to measure changes in the brain's oxygen absorption based on optical properties of hemoglobin (see **Figures 28** and **29**). Although there are three different types of NIRS techniques, the one that has been used in nearly all language studies is the continuous wave technique which simply measures the attenuation of light through the head. An important feature of NIRS not shared by fMRI is the capability to rapidly and *simultaneously* measure changes in local cerebral blood volume, and concentrations of hemoglobin carrying oxygen (oxyhemoglobin) and deoxygenized hemoglobin. fMRI provides a combined blood oxygen level density (BOLD) measure (that is, a ratio between oxygenated and deoxygenated hemoglobin).

Advantages and Disadvantages. NIRS is noninvasive, silent, low cost, and (depending on the number of channels) light weight and portable. It can be used with infants, children, and adults. NIRS is also relatively movement tolerant, and, compared to MEG, fMRI, or PET, the participant is more mobile and not confined to a narrow space. The main disadvantage of NIRS is the spatial resolution. Because of the nature of scattered light, the penetration depth of NRIS is limited to about 3 cm deep into the head (and 1–1.5 cm into the cortex) depending on the distance between IR emitters and detectors. The spatial resolution thus decreases with increasing depth and deep brain structures cannot be measured. Temporal resolution is good but lower than that of EEG. Another disadvantage is the difficulty of separating the NIRS signal of interest from interfering signals from other cerebral tissue. Further disadvantages are that near-infrared light is attenuated by dark hair color and that the placement of the array of optical fibers is complicated and time consuming. Furthermore, the optical fibers cannot cover the entire surface of the brain thus making it impossible to explore all possible neural

Figure 27 An example of functional near-infrared spectroscopy measurements over the human head. The illuminating optic fiber bundles (sources) and the collecting ones (detectors) are assembled into a flexible probe holder ensuring that the 4-cm distance, relative to each other, is fixed. The probe holder is positioned over the left hemisphere using the 10–20 frontal and temporal coordinates (International 10–20 system). The probe holder is fixed to the head by a velcro brand fastener, adapting it to the size and shape of the head. The probe holder and its position on the head allow the creation of a stable optical contact with the scalp for all optic fiber bundles. Figure and legend taken from Quaresima, V., Bisconti, S., Ferrari, M., 2012. A brief review on the use of functional near-infrared spectroscopy (fNIRS) for language imaging studies in human newborns and adults. Brain and Language 121 (2), 79–89.

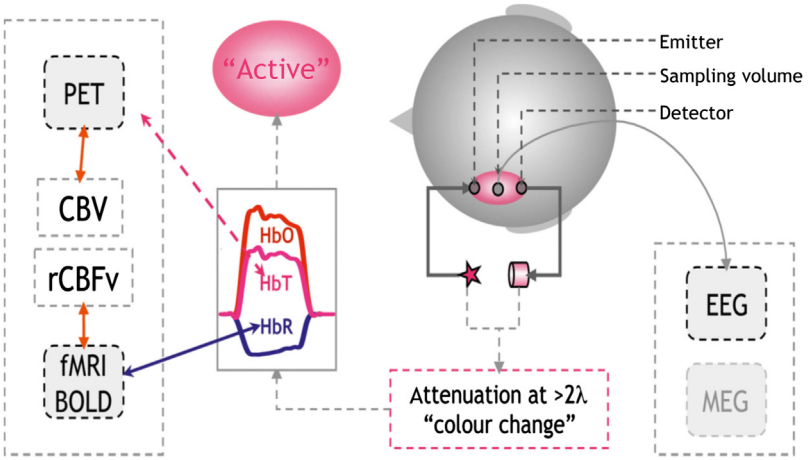

Figure 28 Principles of noninvasive optical imaging: optical probes are fixed to the head defining the sampling volume (pink ellipse). Measuring changes in attenuation at two or more wavelengths allows us to calculate changes in the concentration of HbO and HbR. An increase in HbO and/or a decrease in HbR indicate cerebral activation. Physiologically this pattern results from an increase in regional cerebral blood flow (rCBF) consisting of an increase in blood flow velocity (rCBFv) and volume (rCBV). The increase in rCBF overcompensates the demand in oxygen (Fox and Raichle, 1986). An increased 'washout' of HbR results from increased rCBFv. This decrease in HbR is the major source of fMRI BOLD-contrast increases (Steinbrink et al., 2006). The change in total hemoglobin concentration (HbT = HbR + HbO) correlates with blood volume (CBV), which can be assessed by PET. Notably the assessment of electrophysiological markers of activation by EEG or MEG (right side of graph) is undemanding. Figure and legend taken from Obrig, H., Rossi, S., Telkemeyer, S., Wartenburger, I., 2010. From acoustic segmentation to language processing: evidence from optical imaging. Frontiers in Neuroenergetics 2 (13), 1–12 (Figure 1, p. 3).

systems in an experiment. (for summaries of the method see Aslin and Mehler, 2005; Ferrari and Quaresima, 2012; Gervain et al., 2011; Quaresima et al., 2012).

fNIRS Application to Language Studies

Nearly all language studies with fNIRS have targeted the brain's classical language areas. About one-third of currently published studies have investigated newborns and infants addressing topics such as the development of language lateralization, prosodic processing, or structural properties of language.

fNIRS studies with newborns and infants. The first study to show that aspects of language are lateralized to the left hemisphere already in newborns used fNIRS. Left temporal brain regions showed increased activation when newborns listened to speech stimuli played forward but not when they were played backwards. Individual acoustic/phonological properties, however, seem to lateralize later as Japanese infants have shown left lateralization adult patterns for individual acoustic/phonological properties only after 25–28 months of age.

fNIRS studies have also addressed the development of prosodic processing. Generally, however, these data are difficult to reconcile. The findings indicated that prosodic information is used at an early age by infants, possibly as of 3 months, and it seems that there is a tendency for a stronger response in the right hemisphere (although not all studies have found a lateralization effect). For example, increased activation in frontal brain regions of both hemispheres (thus no lateralization) was reported in response to normally intonated prosody but not to monotonous speech with flat prosody in Japanese newborns. At 3 months, lateralization to the right hemisphere (temporoparietal brain area) was observed with greater activation in response to normally intonated speech as compared to flattened prosody. These findings were interpreted as indicating that over-development prosody is initially processed only acoustically (and not linguistically). Later in development, at 10 months, the frontal regions on both sides in addition to left temporal and temporoparietal regions were more activated in response to flattened than to normal speech. The latter findings are somewhat difficult to interpret. One suggestion was that it reflects the *effortful* attempt to integrate not-normal, flattened speech into the linguistic system thus activating larger language-associated regions. Other studies have observed that the inferior frontal cortex in the right hemisphere seemed to be associated with emotional prosody already in infancy while regions in the left temporal hemisphere seemed to process voices very early in development.

At birth infants are able to distinguish phonemes of a wide range of languages. During the process of native language acquisition, this broad discrimination ability is lost. Using fNIRS phonetic processing was investigated in two groups of bilingual and monolingual babies, that is babies younger than 4–6 months and older than 10–12 months. In both groups, regions in the left temporal brain (superior temporal gyrus) for phonetic processing and regions in the left frontal brain (left inferior frontal cortex) for meaning retrieval and syntactic and phonological patterning were implicated. In the latter case, activation became stronger the older the babies. It was also observed that older bilingual babies were still sensitive to nonnative phonetic contrasts when monolingual babies could no longer make these distinctions (Petitto et al., 2012; see **Figure 30**).

Different languages have different combinatorial rules for phonemes in their language, so-called phonotactic rules. For example /fl/ is a possible combination at the onset of English words (*flight*) while /tl/ is not. Studies using fNIRS

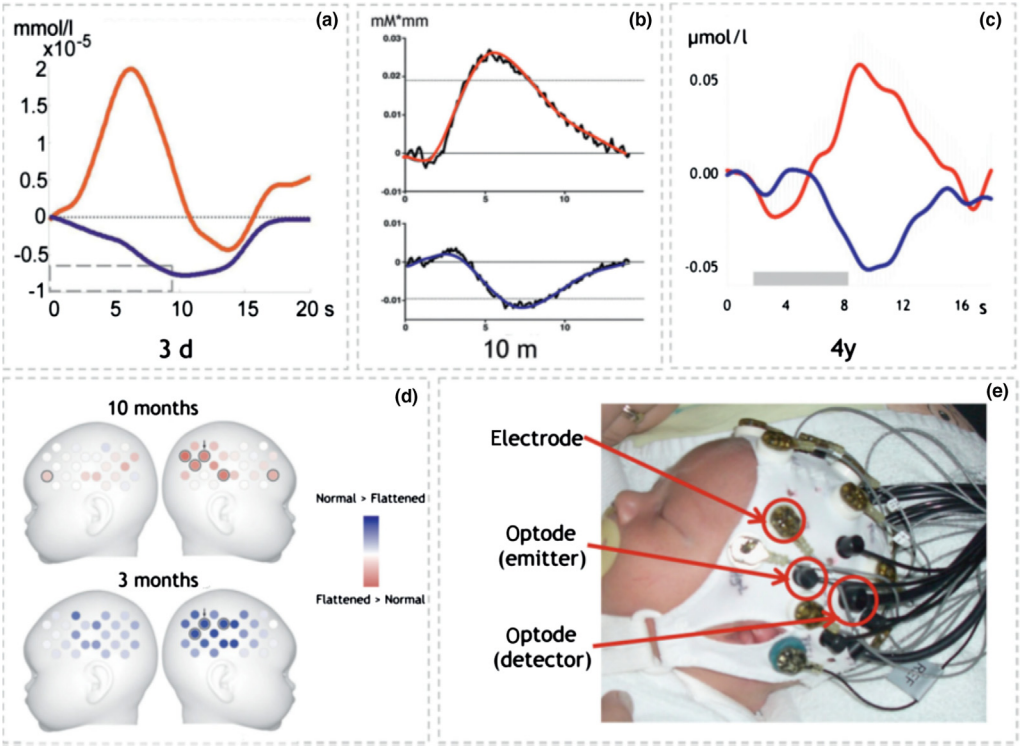

Figure 29 Examples of optical imaging results in infants and neonates. (a–c) Time courses of HbO and HbR in three different age groups. Note that the response pattern is similar in all three studies though magnitude differs. (a) Neonates' response to structured noise. Here the statistics on HbR↓ yielded a lateralized response (adapted with permission from Telkemeyer et al., 2009). (b) Response to sentences in 10-month-old infants. In this study HbO↑ yielded the effects shown in (d) (adapted with permission from Homae et al., 2007). (c) Response to hummed sentences in 4-year-old children. HbO↑ *and* HbR↓ showed statistically significant results (adapted with permission from Wartenburger et al., 2007). (d) Topographical information is available when using probe arrays (24 volumes over each hemisphere; adapted with permission from Homae et al., 2007). The results show the differential activation to normal vs. flattened speech in two age groups (10 and 3 months). Note that the lateralization changes sides (red: larger response amplitude for flattened speech than normal speech; blue: normal > flattened; results are based on HbO↑). (e) Simultaneous assessment of EEG and optical imaging in a neonate as used in (Telkemeyer et al., 2009). (Permission from the infant's parents to show the picture was obtained.) Figure and legend taken from Obrig, H., Rossi, S., Telkemeyer, S., Wartenburger, I., 2010. From acoustic segmentation to language processing: evidence from optical imaging. Frontiers in Neuroenergetics 2 (13), 1–12 (Figure 2, p. 4).

with infants have reported that phonotactic rules seem to operate already in infants as of the age of 3 months without clear lateralization patterns.

Another question addressed with NIRS was when newborns begin to process the structural properties of language. It was observed, for example, that newborns reacted with increased neural activity in the temporal regions of both hemispheres (with left temporal more pronounced than right temporal) and in left frontal areas when listening to speech that followed a structural rule as compared to unstructured speech. This seems to indicate that newborns are sensitive to structural patterns as of very early on.

Other rare applications of fNIRS have been to investigate language in epileptic children, in children with autism and in deaf children.

Generally, with regard to fNIRS studies with infants and children, concern has been expressed that the fNIRS findings in children may need to be interpreted differently from those of adults inasmuch as the infant brain is still immature. Vascular regulation is different from the adult brain and light scattering and absorption may also differ in the immature brain. (For a summary of language studies in infants and children see Dieler et al., 2012; Gervain et al., 2011; Obrig et al., 2010; Quaresima et al., 2012.)

fNIRS studies with adults. Numerous fNIRS studies have been performed with adults addressing topics such as language lateralization , syntactic processing, semantic processing, reading, or language organization in bilinguals. Generally, most fNIRS studies have corroborated and, to some extent, expanded findings from previous imaging studies with other techniques. There are, however, also inconsistencies with other brain imaging studies.

For example, an fNIRS study that investigated syntax processing reported involvement of the left inferior frontal gyrus (a region containing Broca's area) when participants had to judge whether a sentence was syntactically correct. Another study suggested that, in native Japanese subjects, phoneme discrimination was associated with hemodynamic changes of the left hemisphere. Several non-fNIRS neuroimaging studies, however, have shown that intonational pitch is usually processed in the right hemisphere. When, however, intonational pitch is used to discriminate words, as

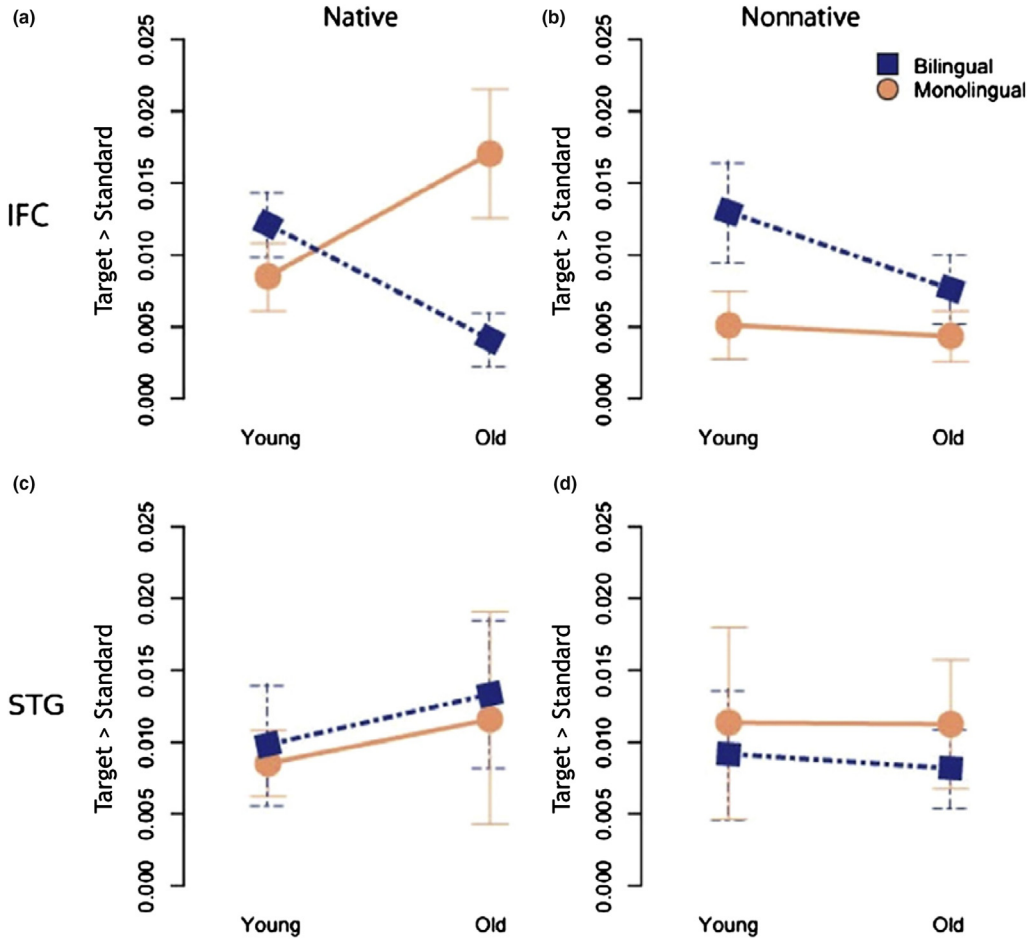

Figure 30 Average difference between standard and target peak activations for bilingual and monolingual babies by age in the left inferior frontal cortex (IFC) and left superior temporal gyrus (STG) for the native and nonnative language conditions. Note that the bilingual and monolingual babies show different patterns of activation in the left IFC. The bilingual babies had more similar activity levels for the native and nonnative phonetic contrasts than the monolingual babies, who showed greater left IFC activations to their native language and this corresponds to behavioral data showing that older bilingual babies' phonetic discrimination capacities remain 'open' at the time when older monolingual babies' phonetic discrimination capacities 'close' to only their native language phonetic contrasts. Figure and legend taken from Petitto, L.A., Berens, M.S., Kovelman, I., Dubins, M.H., Jasinska, K., Shalinsky, M., 2012. The 'Perceptual Wedge Hypothesis' as the basis for bilingual babies' phonetic processing advantage: new insights from fNIRS brain imaging. Brain and Language 121 (2) 130–143 (Figure 4, p. 138).

in Japanese, left frontal regions were activated in fNIRS studies.

An early PET study had indicated that listening to continuous speech activates regions in the left temporal lobe, a finding that was later also supported by an fNIRS study. A few studies have investigated word processing in adults. One study implicated the left inferior frontal brain regions with the statistical segmentation of tone sequences and this has been viewed as the basis of segmenting sentences into meaningful words.

Several studies have reported increased activity in Broca's and Wernicke's area in response to reading tasks.

Overt speech studies pose a problem for fMRI due to the noise of the MRI scanner as well as to the head and jaw movement artifacts during the task. fNIRS is better suited for these types of experiments and were used during face-to-face conversation in which topic and conversation time were predefined. Compared to reciting syllables, during conversation periods an increase in oxyhemoglobin and a decrease in deoxyhemoglobin were observed in frontal and temporal regions (Suda et al., 2010). It might be mentioned that the control condition (reciting syllables) may not have been an optimal choice against which to compare conversation.

Verbal fluency tasks reflecting lexical retrieval processes are a common task used in imaging and nonimaging methods. Numerous fNIRS studies have used these tasks (such as naming words beginning with a certain letter, i.e., phonological fluency, or naming animals, i.e., semantic fluency) and most have reported increases in oxyhemoglobin over bilateral frontal and temporal brain regions and a simultaneous decrease of deoxyhemoglobin. tDCS studies (see Section Transcranial Direct Current Stimulation), however, have suggested left lateralization for these tasks. It should be noted

that some fNIRS studies using verbal fluency tasks showed methodological flaws, which may have influenced the results and their interpretation. In addition, while the reliability of the fNIRS results for verbal fluency tasks has been established at the group level, some researchers have questioned its reliability at the single subject level.

fNIRS has been compared with the intracarotid amobarbital test (Wada test) which has been the gold standard to determine presurgically in which hemisphere language is predominantly represented. Results of tests relying on fNIRS have shown a good correlation with the Wada test concerning the prediction of language dominance. As mentioned previously, however, inasmuch as fNIRS may not be reliable at the single subject level it is not suitable as an alternative to the Wada test. While fNIRS may be a potential candidate to replace the Wada procedure, hardware, software, and data analysis need to be improved to make it a reliable tool (for studies of fNIRS in the context of presurgical evaluation, see, e.g., Gallagher et al., 2012; Watanabe et al., 1998; Watson et al., 2004).

fNIRS has been used to investigate poststroke aphasic individuals. These studies suggested that neuronal activities in lesioned brain tissue lead to altered changes in cerebral oxygen metabolism and regional cerebral blood flow during task performance. In poststroke patients, language tasks activated areas outside Broca's region. Oxygenation patterns were changed in the sense that there was less increase of oxyhemoglobin and in some studies also less decrease of deoxyhemoglobin (for a summary of language studies in adults see Dieler et al., 2012; Obrig et al., 2010; Quaresima et al., 2012).

Summary

Compared to other neuroimaging techniques such as PET, fMRI, EEG, and MEG, fNIRS still lacks standardization of hardware, software, and data analysis procedures; this lack makes the comparison of studies difficult. Currently, the main problems are the low penetration depth limiting studies to investigate only the surface of the cortex and exclude deep brain structures. Further, only certain parts of the head can be covered with probes thus limiting the regions of analysis.

Most fNIRS studies have shown that language tasks in healthy individuals are accompanied by an increase in oxyhemoglobin and a decrease or no increase in deoxyhemoglobin thus confirming the sensitivity of fNIRS to language tasks. Currently, using fNIRS has been most beneficial in studies on language development with infants and children where other techniques are less suitable. fNIRS seems particularly suitable to investigate language development over time in young children. In aphasic patients abnormal oxygenation patterns were found but more studies are needed to show the advantages of fNIRS compared to other techniques.

As all neuroimaging techniques have their advantages and disadvantages, researchers have started to combine methods to exploit their individual strengths and overcome or minimize disadvantages of one technique. While simultaneous EEG and fMRI recording is already used frequently, combining fNIRS with other techniques such as EEG is still in its infancy. The benefits of such a combination have been reviewed by Wallois et al. (2012).

Brain Stimulation Methods

Applying electrical currents to investigate brain functions or modify the brain in humans has a long history. In 1802 the Italian scientist Giovanni Aldini electrified the brains of decapitated prisoners in public to induce movement in the facial muscles. He observed that left cortical stimulation of the dead brain elicited right-sided muscular contractions (for a historical summary see Harris and Almerigi, 2009; Zago et al., 2008). Toward the end of the nineteenth century, Roberts Bartholow and later Ezio Sciamanna reported the effects of electric stimulation to the exposed brain of two patients: one suffering from an eroding cranial tumor of the skull and the other from a traumatic brain injury. While these experiments on humans raised ethical disapproval in the scientific community, Alberto Alberti received an award for similar work between 1883 and 1884. He performed a series of experiments involving the electrical stimulations of the exposed cortex of an epileptic patient over a period of 8 months. These experiments confirmed the electrical excitability of the cortex in humans and demonstrated somatosensory and motor functions contralateral to the side of stimulation. The next decades saw scientists use the technique intermittently in a variety of settings. Wilder Penfield, the founder of the Montreal Neurological Institute, systematically stimulated the patients' brains with intracerebral electrodes during neurosurgery and associated specific regions of stimulation with motor, sensory, and speech functions. Today intracortical brain stimulation is a routine procedure to establish functional brain maps of patients who are scheduled to undergo surgery mainly for epilepsy with the goal of sparing as much functional brain tissue as possible. Although this is obviously an invasive technique – after all the skull is cut open to expose the brain – the invasiveness in epileptic patients is medically justified. In normal healthy subjects, however, scientists have tried to electrically stimulate the brain through the intact skull using the TMS technique or the tDCS technique.

Transcranial Magnetic Stimulation (TMS)

In 1991, transcranial magnetic stimulation (TMS) was introduced as a tool to study language. Since then the technique has become an integral part in language research and has also been applied in language rehabilitation. Before providing illustrative examples of the use of the technique in language studies, the nature of TMS will be described.

TMS (like tDCS to be discussed below) is a method for noninvasive brain stimulation. It should be noted that 'noninvasive' in this context simply means that the skull is not opened to apply the stimulation. Unlike EEG or MEG, however, which *detect and measure* brain activity, TMS *interferes* with brain function and, in exceptional cases, may provoke seizures. Due to this infrequent complication, TMS is sometimes considered invasive (for example, by ethics committees).

Essential to TMS is a simple coil of wire (encased in plastic) through which an electric current is introduced usually with the help of a capacitor. The coil of wire (often round or in the form of a figure eight) is placed close to the head and area to be stimulated. The rapidly changing current flow in the coil of wire produces a magnetic field oriented orthogonally to the plane of the coil. The field transverses the skin, skull, meninges, and brain fluid without any deviations and induces an oppositely directed electric current in neurons lying parallel to the plane of the coil. The current flow leads to neural depolarization which, in turn, excites or inhibits the cortex thus affecting information processing in the brain (see **Figure 31**).

TMS comes in three varieties. In single-pulse TMS, stimulation is delivered once every few seconds and can be used, for example, to study the relationship between language processes and behavior. In paired-pulse TMS, two TMS pulses are fired very close together in time to see how they interact. Pairing can, for example, provide measures of intracortical facilitation and inhibition during language processing. Finally, in repetitive TMS (rTMS), single pulses are regularly repeated over time. Depending on a variety of physiological effects and degrees of risk, variations are fast or high-frequency rTMS during which more than one magnetic pulse is applied every second (such as 5 Hz, 10 Hz, and 20 Hz). Fast rTMS can induce a transient increase in cortical excitability. With slow or slow-frequency rTMS one magnetic pulse (1 Hz) or less is applied every second and this can induce a decrease in cortical excitability (for a summary see Rossi et al., 2009). rTMS is used, for example, in therapeutic settings to modulate affective (depression) or cognitive functions including language. When planning TMS studies it is important to define the intensity and frequency of the stimulation, train duration, intertrain intervals, and cumulative application (over days, weeks). In addition, to increase precision of the stimulation region, TMS is frequently used with a stereotactic frame and structural MRI scans of the participant.

TMS is relatively safe when the safety guidelines are followed (Rossi et al., 2009). Ferromagnetic material (e.g., cardiac pacemaker, metal in the body) may be dangerous and participants need to be carefully questioned. Discomfort such as pain and headaches have been reported in some participants. The most severe adverse event of rTMS is the induction of seizures; their occurrence, however, is very rare and does not increase the risk of seizures in the future (for details see Rossi et al., 2009).

Advantages and Disadvantages. Participants can act as their own controls, that is allowing comparisons before, during, or after stimulation. TMS allows for an examination of the

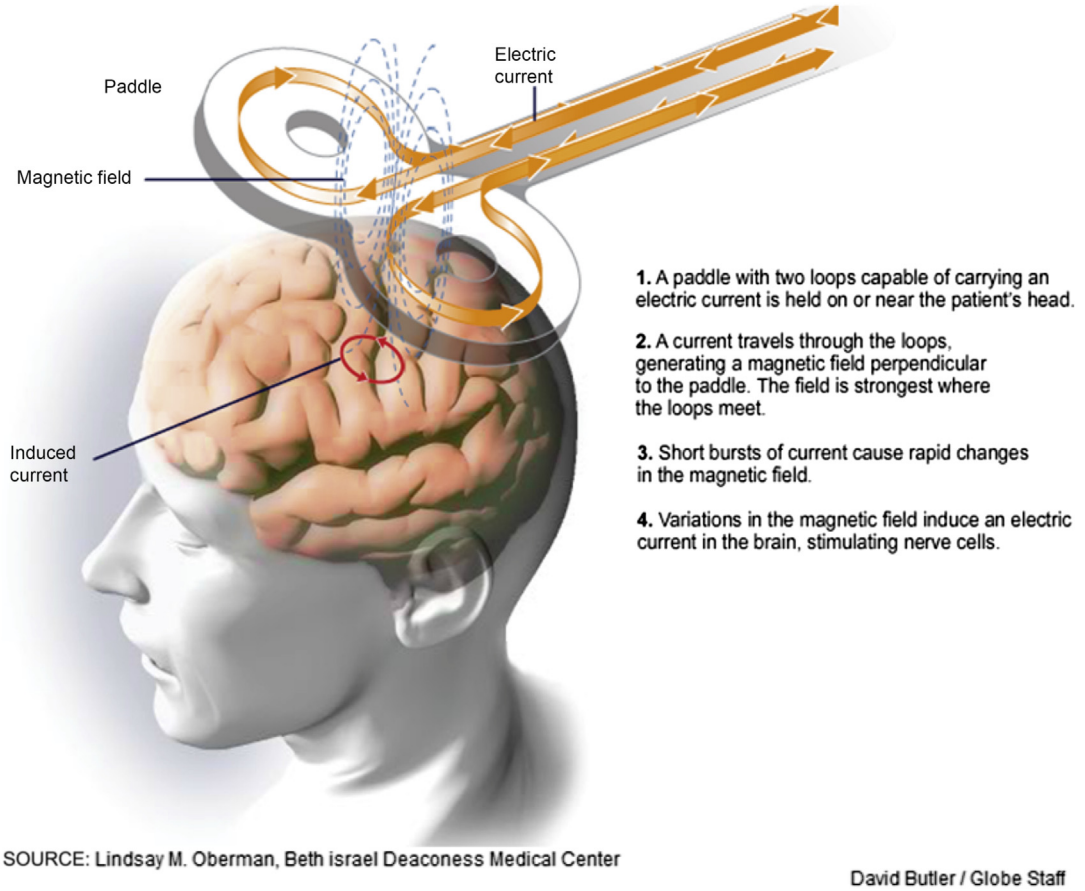

Figure 31 How does TMS work? Copyright by Boston Globe via gettyimages, costs 222 Euro and can be obtained from: http://www.gettyimages.com/detail/news-photo/an-illustration-published-in-the-boston-globe-on-june-8-news-photo/175697813.

timing of information processing with a temporal resolution of tens of milliseconds and can be combined with other techniques such as PET, fMRI, or EEG. Limitations of TMS include possible discomfort (headache, pain) and, in rare cases, seizures. TMS only reaches areas at or near the cortical surface. A loud click is heard when TMS is applied, a sensation on the scalp may be felt and muscles may twitch. These may interfere with the experiment and may be particularly disturbing in language studies (for a summary see Watkins and Devlin, 2008).

(For summaries describing the technique of TMS see, for example, Devlin and Watkins, 2007; Rossi et al., 2009; Rossini et al., 2010; Watkins and Devlin, 2008).

TMS in Language Studies with Healthy Individuals

For many years the Wada procedure, that is selectively anesthetizing one hemisphere has been the gold standard to determine on which side of the brain language is predominantly represented. Several studies have investigated the possibility of using TMS to replace the Wada test by, for example, using rTMS in an attempt to inhibit language output in the hemisphere that is dominant. Although some studies have reported promising results, other studies could not replicate the findings.

Lexical retrieval processes have also been investigated with TMS. Imaging studies have suggested that left posterior brain regions (more specifically the left posterior inferior-temporal cortex) are implicated in picture naming. Most rTMS studies have shown increased picture-naming latencies and thus support this view. There are, however, studies that reported *decreased* naming latencies with rTMS over Wernicke's area, and still others have reported faster reaction time on a word-picture verification task – regardless of where (out of five sites) stimulation was applied. Interpretation of these findings is thus difficult (for a summary see Watkins and Devlin, 2008).

TMS studies have supported and extended the insight from other neuroimaging techniques that the left inferior frontal brain regions are involved in semantic and phonological processing. TMS showed that the posterior left inferior frontal gyrus is necessary for phonological processing while the anterior left inferior frontal gyrus is necessary for semantic processing (Watkins and Devlin, 2008).

As early as the 1980s, some theories suggested that understanding speech involves sensory and motor representations. Studies using TMS indicate that speech perception and speech production are closely linked and that Broca's area seems to be the mediator between the two.

Another question that has been addressed is whether motor areas are involved in the neural representation of actions as, for example, expressed in verbs. Some studies have shown a decrease in response times during lexical decisions on action verbs (e.g., *folding arms*) when the motor cortex was stimulated with TMS. Interpretation of these results is, however, complex as one needs to tease apart general activation of the motor cortex and specific effects inherent in action verbs. Other findings are equally difficult to interpret. One TMS study associated a region anterior to the motor region (that is, the left dorso-lateral prefrontal cortex) with action verbs whereas another found it to be implicated in the processing of grammatical categories (for a discussion of these studies see Watkins and Devlin, 2008). (For summaries on the use of TMS in studies of language see, for example, Devlin and Watkins, 2007; Flöel, 2011; Naeser et al., 2012; Martin et al., 2009; Schlaug et al., 2011; Watkins and Devlin, 2008).

TMS in Language Rehabilitation

In patient populations TMS and tDCS (see next section) have been used to enhance the recovery process. The idea is based on studies investigating the motor system. Models of motor recovery after stroke assume that there is rivalry between the intact and nonintact hemisphere. For example, if the left hemisphere is lesioned, studies have shown that there is increased inhibition and thus reduced output in the motor regions of the lesioned hemisphere. This seems to go hand-in-hand with a disinhibition in the same regions of the nonlesioned (in this case right) hemisphere resulting in an imbalance in the interaction of the two hemispheres (for a summary and details see Schlaug et al., 2011). Balance may be achieved by either increasing excitation in cortical regions of the lesioned hemisphere, by decreasing excitation in regions of the nonlesioned hemisphere, or by a mixture of the two. Although the model may be somewhat simplistic, it has received some support from studies with stroke patients whose motor systems were impaired. A similar model has been used to explain favorable and unfavorable recovery in individuals with an impaired language system such as stroke patients with aphasia whose left hemisphere is lesioned (see **Figure 32**). It is assumed that the lesioned left hemisphere is inhibited thus reducing output whereas the intact right hemisphere is disinhibited and overexcited. This is where noninvasive brain stimulation methods such as TMS or tDCS come in. The idea is – in simple terms – to use cortical stimulation to decrease the inhibition of the left hemisphere and make it more excitable, or to decrease the over-excitability of the right hemisphere.

The model is supported by functional neuroimaging studies with aphasia patients who have frequently shown an increased activation in right hemisphere language homologs. These observations have led some researchers to apply slow rTMS over the right hemisphere of chronic aphasic individuals. According to the model this should decrease the activity of the overly excited right hemisphere and facilitate the task. Indeed, studies showed an improvement in naming accuracy and reaction times. If rTMS was performed daily for 2 weeks, this improvement lasted for at least 2 months (Naeser et al., 2005, 2012). Efforts have also been made to combine TMS with speech therapy to enhance the recovery process and additional improvement beyond TMS alone has been shown. Generally, most studies have reported beneficial effects to some degree when delivering rTMS to chronic aphasia patients with or without combined therapy that lasted well beyond the stimulation. There are, however, some studies, that reported that right hemisphere stimulation interfered negatively with performance as did left hemisphere stimulation (for a summary see Naeser et al., 2012; Watkins and Devlin, 2008). Another well-controlled study only found significant improvements when comparing real rTMS with sham applications in severe but not moderate aphasic individuals (Seniow et al., 2013).

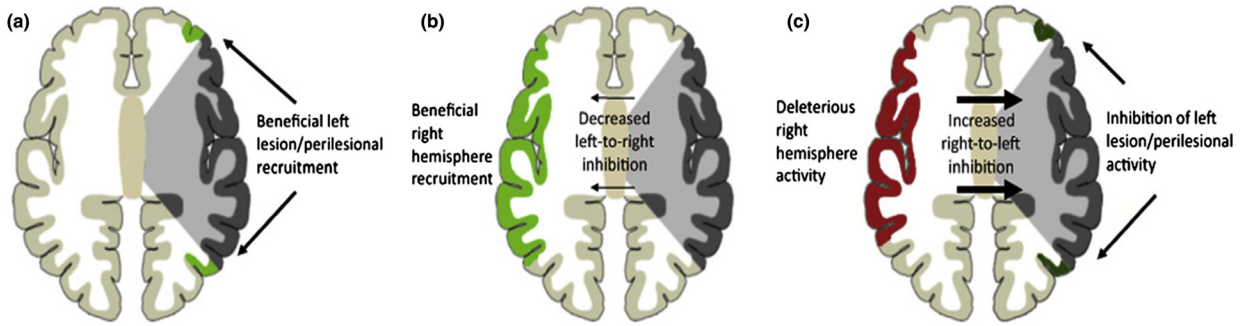

Figure 32 Differing accounts of plasticity in language systems in chronic aphasia. (a) After unilateral left hemisphere stroke (gray), some language functions may be subserved by recovered lesional areas or recruited perilesional areas (light green). (b) Right perisylvian areas (light green) may be recruited to subserve some language functions, a process facilitated by decreased transcallosal inhibition of the right hemisphere by the damaged left hemisphere. (c) By contrast, right hemisphere activity may be deleterious. Released from interhemispheric inhibition, right hemisphere structures (red) may exert increased inhibitory influence on left perisylvian areas, impeding functional recovery of lesional and perilesional areas in the left hemisphere (dark green). Figure and legend taken from Hamilton, R.H., Chrysikou, E.G., Coslett, B., 2011. Mechanisms of aphasia recovery after stroke and the role of noninvasive brain stimulation. Brain and Language 118 (1–2), 40–50 (Figure 2, p. 43).

The inconsistent findings may be due to the different types of stimulation and protocols used and the individual patients investigated. Most studies up to now have not been well controlled and have been based on small sample sizes. We do not know the pathological characteristics of patients that may predict success under these procedures. Better controlled studies with larger patient samples and common stimulation standards (as recently suggested by Naeser et al., 2012) are necessary to clarify the findings and the role each hemisphere plays in the recovery process.

Transcranial Direct Current Stimulation (tDCS)

With the knowledge gained from neuroimaging and mapping techniques, and particularly from TMS studies, in recent years tDCS has experienced a resurgence in the clinical and academic world and, due to its simple set-up, has also raised the interest of laypeople (websites such as *tDCS DIY* – do it yourself – bear witness).

A weak direct current is applied most frequently via two electrodes attached to the scalp. The current transverses the skull and creates an electric field in brain tissue leading to voltage changes across the neuronal membranes thus changing neuronal activity and modulating human cerebral cortical function. Depending on the polarity of the stimulation, tDCS results in depolarization of the resting membrane potential thus increasing neuronal excitability (anodal tDCS) or in hyperpolarization and decreasing neuron excitability (cathodal tDCS) (for a summary see Nitsche and Paulus, 2001; Nitsche et al., 2008). tDCS generally produces diffuse (and not focal) cortical stimulation. Size, polarity, and position of the electrodes as well as the applied current intensity and the properties of the tissue influence the strength of the current density in the cortex (Dmochowski et al.; Utz et al., 2010). Depending on the duration and intensity of the stimulation, changes in neuronal activity have been observed not only during but also after the stimulation has ended. The long-lasting aftereffect has been explained by various mechanisms such as the induction of long-term potentiation and depression, an effect on neurotransmitters or a change of local pH, blood flow, brain oxygenation, or calcium neuronal influx (Holland and Crinion, 2012; Monti et al., 2013).

Advantages and Disadvantages. Compared to TMS, tDCS is less expensive, easier to use, more portable, and does not produce noise or muscle twitching thus making fake or placebo trials more reliable. Whereas TMS induces neuronal action potentials, tDCS does not do so as it lacks the rapid depolarization required for the production of action potentials. TMS is both neurostimulatory and neuromodulatory while tDCS seems to be neuromodulatory only (Fregni and Pascual-Leone, 2007). tDCS is considered safe and without serious adverse effects if the safety standards are maintained. If the threshold for current and the charge density are exceeded, brain lesions will result (Utz et al., 2010). It is also important to note that certain medications modulate the effects of tDCS. Finally, it is difficult to determine exactly which brain region is stimulated with tDCS.

Application in Language Research and Language Rehabilitation

tDSC has been used in a wide range of applications in clinical and nonclinical studies investigating motor, sensory, mood, and cognitive functions (for a summary see Utz et al., 2010). In nonclinical language research tDCS has shown effects on word generation and retrieval and verbal learning. In clinical settings, tDCS has been applied to improve language performance in patients with aphasia.

tDCS in healthy individuals. Applying tDCS in a single session in healthy individuals has been shown to influence lexical retrieval processes. These processes are often tested with verbal fluency tasks or naming tasks. Verbal fluency tests have been used as a measure of phonemic fluency (generating words in response to a letter cue-word such as 'F') or as a measure of retrieval from semantic memory (generating words that belong to a particular semantic category, such as animals). Applying anodal tDCS over the left prefrontal cortex during verbal fluency tasks and over the left dorsolateral prefrontal cortex during naming tasks improved the performance on such tasks while cathodal or sham (fake) tDCS did not show any effect (for summaries see Cattaneo et al., 2011; Fertonani et al., 2010; Holland and Crinion, 2012; Iyer et al., 2005).

For example, when individuals named the pictures they saw on a screen while their left inferior frontal cortex (Broca's region) was being stimulated, their response time decreased compared to the nonstimulation (Holland et al., 2011). When participants were engaged in learning an artificial grammar, stimulation of the same region facilitated the detection of syntactic violations (de Vries et al., 2009). Instead of naming objects, in still another study, individuals had to name famous people. This time tDCS was applied over the right anterior temporal lobe and this led to improved accuracy in naming famous people (Ross et al., 2010).

tDSC has also been shown to have an effect on verbal acquisition and learning. Cathodal tDCS over the left motor cortex produced a negative effect on vocabulary acquisition while anodal or sham (i.e., fake) stimulation had no effect (Liuzzi et al., 2010). Learning speed and accuracy during the acquisition of new words increased when anodal tDCS was applied over Wernicke's area (Flöel et al., 2008; Valentina Fiori et al., 2010). Word learning over 5 days improved with tDCS and the learned material could still be retrieved a week later (Meinzer et al., in press).

tDCS in aphasic individuals. Compared to studies with healthy individuals, the number of studies with aphasic individuals is still rather limited. Applying tDCS to regions in the affected left hemisphere in individuals with fluent or nonfluent aphasia during naming tasks showed some improvement either in naming accuracy or exhibited an effect in the reaction time for trained items (Fiori et al., 2013; for summaries see Hamilton et al., 2011; Holland and Crinion, 2012; Schlaug et al., 2011). One study found naming improvement in only one out of several patients (Volpato et al., 2013). Another study applied anodal tDCS to the temporo-parietal junction of the intact right hemisphere of a chronic aphasic patient and reported positive effects in a naming task up to 2 weeks after the stimulation (Flöel, 2011). Multisession application of tDCS and stimulation of the left hemisphere (the site of stimulation varied in the studies) also showed naming improvement that lasted beyond the duration of the session (see Holland and Crinion, 2012, for a summary).

In sum, current evidence shows that tDCS induces an effect on word generation and retrieval and lexical learning in healthy individuals. The effect seems to differ depending on the region of application and on which language function is under consideration. As there are currently no methodological standards defined for tDCS, it is difficult to compare studies. Methodological details are not always reported, and the methods used differ, for example, in the duration and intensity of stimulation, in electrode size and placement, and in various other parameters used in the study protocol. In addition to the limited number of language studies, most extant studies have only investigated small populations of subjects. The number of studies investigating the application of tDSC in aphasia rehabilitation is still very limited and in its early stages. Although most studies report some effect, the findings are not consistent. The limitations mentioned previously also apply here. In addition, questions such as where to stimulate, when to stimulate, and how long to stimulate are even more pertinent in a rehabilitation context. For healthy and pathological populations, more systematic studies are needed with common methodological standards and larger subject pools to clarify the effect of tDCS on various aspects of language.

Summary Noninvasive Brain Stimulation

As methods of stimulation, TMS and tDCS have certain obvious advantages in neurolinguistic research over the purely passive observational methods mentioned in previous sections. By employing TMS and tDCS, neural circuits associated with language understanding and production can be both inhibited and facilitated. tDCS is valuable for its ability to induce not only motor functions but to accelerate some elements of language processing. TMS allows for the testing of causality in selected areas of the brain that neuroimaging methods have shown to be important in linguistic phenomena. The therapeutic possibilities of both TMS and tDCS in patients suffering from brain lesions involving language-related areas of the cerebral cortex are currently being explored.

Disadvantages of the brain stimulation methods include their applicability only to sites on or near the surface of the brain. Another disadvantage is the fact that the areas of stimulation accessible for TMS and tDCS are relatively large (centimeters) and at the present time, selective stimulation of discrete brain regions of interest is impossible.

Inasmuch as these methods are relatively new and have been carried out by many independent research groups, there is also at the present time a lack of standardization of the stimulation parameters used in TMS and tDCS.

Outlook

Each technique discussed above has its own weakness and its strength. Further development of hardware and software tools as well as the combination of techniques will diminish the limitations inherent in each individual method. Those techniques that provide the best combination of temporal and spatial resolution and are most cost efficient and portable will probably be those of the future. Taking the hardware out of the laboratory, applying it to 'real life situations' and linking it to modern communication tools might change the approach to the study of language in all its facets and will also have an impact on language rehabilitation. Attempts in this direction in cognitive domains are already under way. Despite all the technological advances, however, it should not be forgotten that 'classical' observational and behavioral techniques cannot always be replaced by even the most advanced technology. And, as mentioned in the introduction, all techniques are in some way limited by their own physical properties, by the physiology of the brain, and by those who try to make sense of both.

See also: Eye Movements in Reading; Intracarotid Sodium Amytal Test and Language; Language Processing, Functional Magnetic Resonance Imaging of; Lateralization of Language as Demonstrated by Brain Imaging Procedures; Pragmatic Disorders; Transcranial Direct Current Stimulation and Language.

Bibliography

Aslin, R.N., Mehler, J., 2005. Near-infrared spectroscopy for functional studies of brain activity in human infants: promise, prospects, and challenges. Journal of Biomedical Optics 10 (1), 11009. http://dx.doi.org/10.1117/1.1854672.

Bastiaansen, M.C.M., Magyari, L., Hagoort, P., 2009. Syntactic unification operations are reflected in oscillatory dynamics during online sentence comprehension. Journal of Cognitive Neuroscience 22 (7), 1333–1347.

Bastiaansen, M.C.M., Hagoort, P., 2006. Oscillatory neuronal dynamics during language comprehension. Progress in Brain Research 159, 179–196.

Bastiaansen, M.C.M., Mazaheri, A., Jensen, O., 2012. Beyond ERPs: oscillatory neuronal dynamics. In: Luck, S.J., Kappenman, E.S. (Eds.), The Oxford Handbook of Event-related Potential Components. Oxford University Press, New York, pp. 31–50.

Bennett, M., Dennett, D.C., Hacker, P., Searle, J., 2007. Neuroscience and Philosophy: Brain, Mind, and Language. Columbia University Press, New York.

Berthier, M.L., García-Casares, N., Walsh, S.F., Nabrozidis, A., Ruíz de Mier, R.J., Green, C., Dávila, G., Gutiérrez, A., Pulvermüller, F., 2011. Recovery from post-stroke aphasia: lessons from brain imaging and implications for rehabilitation and biological treatments. Discovery Medicine 12 (65), 275–289.

Bialystok, E., Craik, F.I.M., Luk, G., 2012. Bilingualism: consequences for mind and brain. Trends in Cognitive Sciences 16 (4), 240–250.

Binder, J.R., Desai, R.H., 2011. The neurobiology of semantic memory. Trends in Cognitive Sciences 15 (11), 527–536.

Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cerebral Cortex 19 (12), 2767–2796.

Catani, M., Jones, D.K., ffytche, D.H., 2005. Perisylvian language networks of the human brain. Annals of Neurology 57 (1), 8–16.

Catani, M., Mesulam, M.M., 2008. The arcuate fasciculus and the disconnection theme in language and aphasia: history and current state. Cortex 44 (8), 953–961.

Catani, M., Mesulam, M.M., Jakobsen, E., Malik, F., Martersteck, A., Wieneke, Thompson, C.K., Thiebaut de Schotten, M., Dell'acqua, F., Weintraub, S., Rogalski, E., 2013. A novel frontal pathway underlies verbal fluency in primary progressive aphasia. Brain 136 (Pt 8), 2619–2628.

Cattaneo, Z., Pisoni, A., Papagno, C., 2011. Transcranial direct current stimulation over Broca's region improves phonemic and semantic fluency in healthy individuals. Neuroscience 183, 64–70.

Coulson, S., 2012. Cognitive neuroscience of figurative language. In: Spivey, M.J., McRae, K., Joanisse, M.F. (Eds.), The Cambridge Handbook of Psycholinguistics. Cambridge University Press, Cambridge, pp. 523–538.

Colin, C., Radeaua, M., Soquetc, A., Demolinc, D., Colind, F., Deltenre, P., 2002. Mismatch negativity evoked by the McGurk–MacDonald effect: a phonetic representation within short-term memory. Clinical Neurophysiology 113, 495–506.

Devlin, J.T., Watkins, K.E., 2007. Stimulating language: insights from TMS. Brain 130 (Pt 3), 610–622.

Dieler, A.C., Tupak, S.V., Fallgatter, A.J., 2012. Functional near-infrared spectroscopy for the assessment of speech related tasks. Brain and Language 121 (2), 90–109.

Dmochowski, J.P., Bikson, M., Datta, A., Richardson, J., Fridriksson, J., Parra, L.C., 2012. On the role of electric field orientation in optimal design of transcranial current stimulation. Conference Proceedings – IEEE Engineering in Medicine and Biology Society 6426–6429.

DeLong, K.A., Urbach, T.P., Kutas, M., 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nature Neuroscience 8, 1117–1121.

Fedorenko, E., Nieto-Castañón, A., Kanwisher, N., 2012. Syntactic processing in the human brain: what we know, what we don't know, and a suggestion for how to proceed. Brain and Language 120 (2), 187–207.

Ferrari, M., Quaresima, V., 2012. A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. Neuroimage 63 (2), 921–935.

Ferstl, E.C., 2010. Neuroimaging of text comprehension: where are we now? Italian Journal of Linguistics 22 (1), 61–88.

Ferstl, E.C., Neumann, J., Bogler, C., von Cramon, D., 2008. The extended language network: a meta-analysis of neuroimaging studies on text comprehension. Human Brain Mapping 29 (5), 581–593.

Fertonani, A., Rosini, S., Cotelli, M., Rossini, P.M., Miniussi, C., 2010. Naming facilitation induced by transcranial direct current stimulation. Behavioural Brain Research 208 (2), 311–318.

Fiori, V., Cipollari, S., Di Paola, M., Razzano, C., Caltagirone, C., Marangolo, P., 2013. tDCS stimulation segregates words in the brain: evidence from aphasia. Frontiers in Human Neuroscience 7. http://dx.doi.org/10.3389/fnhum.2013.00269.

Fiori, V., Coccia, M., Marinelli, C.V., Vecchi, V., Bonifazi, S., Ceravolo, M.G., Provinciali, L., Tomaiuolo, F., Marangolo, P., 2010. Transcranial direct current stimulation improves word retrieval in healthy and nonfluent aphasic subjects. Journal of Cognitive Neuroscience 23 (9), 2309–2323.

Flöel, A., 2011. Non-invasive brain stimulation and language processing in the healthy brain. Aphasiology 26 (9), 1082–1102.

Flöel, A., Rösser, N., Michka, O., Knecht, S., Breitenstein, C., 2008. Noninvasive brain stimulation improves language learning. Journal of Cognitive Neuroscience 20 (8), 1415–1422.

Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., Raichle, M.E., 2005. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proceedings of the National Academy of Sciences of the United States of America 102 (27), 9673–9678.

Fregni, F., Pascual-Leone, A., 2007. Technology insight: noninvasive brain stimulation in neurology-perspectives on the therapeutic potential of rTMS and tDCS. Nature Clinical Practice Neurology 3 (7), 383–393.

Friederici, A.D., 2011. The brain basis of language processing: from structure to function. Physiological Reviews 91 (4), 1357–1392.

Friederici, A.D., Gierhan, S.M., 2013. The language network. Current Opinion in Neurobiology 23 (2), 250–254.

Federmeier, K.D., 2007. Thinking ahead: the role and roots of prediction in language comprehension. Psychophysiology 44, 491–505.

Fox, P.T., et al., 1988. Nonoxidative glucose consumption during focal physiologic neural activity. Science 241, 462–464.

Fox, P.T., Raichle, M.E., 1986. Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. Proceedings of National Academy of Sciences U.S.A. 83, 1140–1144.

Gallagher, A., Beland, R., Lassonde, M., 2012. The contribution of functional near-infrared spectroscopy (fNIRS) to the presurgical assessment of language function in children. Brain and Language 121 (2), 124–129.

Gervain, J., Mehler, J., Werker, J.F., Nelson, C.A., Csibra, G., Lloyd-Fox, S., Shukla, M., Aslin, R.N., 2011. Near-infrared spectroscopy: a report from the McDonnell infant methodology consortium. Developmental Cognitive Neuroscience 1 (1), 22–46.

Gold, B.T., Kim, C., Johnson, N.F., Kryscio, R.J., Smith, C.D., 2013. Lifelong bilingualism maintains neural efficiency for cognitive control in aging. Journal of Neuroscience 33 (2), 387–396.

Hamilton, R.H., Chrysikou, E.G., Coslett, B., 2011. Mechanisms of aphasia recovery after stroke and the role of noninvasive brain stimulation. Brain and Language 118 (1–2), 40–50.

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92 (1–2), 67–99.

Hilchey, M.D., Klein, R.M., 2011. Are there bilingual advantages on nonlinguistic interference tasks? Implications for the plasticity of executive control processes. Psychonomic Bulletin and Review 18 (4), 625–658.

Holland, R., Crinion, J., 2012. Can tDCS enhance treatment of aphasia after stroke? Aphasiology 26 (9), 1169–1191.

Holland, R., Leff, A.P., Josephs, O., Galea, J.M., Desikan, M., Price, C.J., Rothwell, J.C., Crinion, J., 2011. Speech facilitation by left inferior frontal cortex stimulation. Current Biology 21 (16), 1403–1407.

van Heuven, W.J.B., Dijkstra, T., 2010. Language comprehension in the bilingual brain: fMRI and ERP support for psycholinguistic models. Brain Research Reviews 64, 104–122.

Indefrey, P., 2006. A meta-analysis of hemodynamic studies on first and second language processing: which suggested differences can we trust and what do they mean? Language Learning 56, 279–304.

Iyer, M.B., Mattu, U., Grafman, J., Lomarev, M., Sato, S., Wassermann, E.M., 2005. Safety and cognitive effect of frontal DC brain polarization in healthy individuals. Neurology 64 (5), 872–875.

Kousaie, S., Phillips, N.A., 2012. Ageing and bilingualism: absence of a "bilingual advantage" in stroop interference in a nonimmigrant sample. Quarterly Journal of Experimental Psychology 65 (2), 356–369.

Kutas, M., Federmeier, K.D., 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annual Review of Psychology 62, 621–647.

Kutas, M., 1993. In the company of other words: electrophysiological evidence for single-word and sentence context effects. Language and Cognitive Processes 8, 533–572.

Lehongre, K., Ramus, F., Villiermet, N., Schwartz, D., Giraud, A.-L., 2011. Altered low-gamma sampling in auditory cortex accounts for the three main facets of dyslexia. Neuron 72 (6), 1080–1090.

Liuzzi, G., Freundlieb, N., Ridder, V., Hoppe, J., Heise, K., Zimerman, M., Hummel, F.C., 2010. The involvement of the left motor cortex in learning of a novel action word lexicon. Current Biology 20 (19), 1745–1751.

Logothetis, N.K., 2008. What we can do and what we cannot do with fMRI. Nature 453 (7197), 869–878.

Luk, G., Anderson, J.A.E., Craik, F.I.M., Grady, C., Bialystok, E., 2010. Distinct neural correlates for two types of inhibition in bilinguals: response inhibition versus interference suppression. Brain and Cognition 74 (3), 347–357.

Laszlo S., Federmeier K.D. The N400 as a snapshot of interactive processing: evidence from regression analyses of orthographic neighbor and lexical associate effects. Psychophysiology, in press.

Martin, P.I., Naeser, M.A., Ho, M., Treglia, E., Kaplan, E., Baker, E.H., Pascual-Leone, A., 2009. Research with transcranial magnetic stimulation in the treatment of aphasia. Current Neurology and Neuroscience Reports 9 (6), 451–458.

Meinzer, M., Harnish, S., Conway, T., Crosson, B., 2011. Recent developments in functional and structural imaging of aphasia recovery after stroke. Aphasiology 25 (3), 271–290.

Meinzer, M., Jähnigen, S., Copland, D.A., Darkow, R., Grittner, U., Avirame, K., Rodriguez, A.D., Lindenberg, R., Flöel, A. Transcranial direct current stimulation over multiple days improves learning and maintenance of a novel vocabulary. Cortex. http://dx.doi.org/10.1016/j.cortex.2013.07.013, in press.

Monti, A., Ferrucci, R., Fumagalli, M., Mameli, F., Cogiamanian, F., Ardolino, G., Priori, A., 2013. Transcranial direct current stimulation (tDCS) and language. Journal of Neurology, Neurosurgery and Psychiatry 84 (8), 832–842.

Mori, S., Zhang, J., 2009. Diffusion tensor imaging (DTI). In: Larry, R.S. (Ed.), Encyclopedia of Neuroscience. Academic Press, Oxford, pp. 531–538.

Massaro, D.W., 1998. Perceiving Talking Faces: From Speech Perception to a Behavioural Principle. MIT Press, Cambridge, MA.

Naeser, M.A., Martin, P.I., Ho, M., Treglia, E., Kaplan, E., Bashir, S., Pascual-Leone, A., 2012. Transcranial magnetic stimulation and aphasia rehabilitation. Archives of Physical Medicine and Rehabilitation 93 (1, Suppl.), S26–S34.

Naeser, M.A., Martin, P.I., Nicholas, M., Baker, E.H., Seekins, H., Kobayashi, M., Theoret, H., Fregni, F., Maria-Tormos, J., Kurland, J., Doron, K.W., Pascual-Leone, A., 2005. Improved picture naming in chronic aphasia after TMS to part of right Broca's area: an open-protocol study. Brain and Language 93 (1), 95–105.

Nitsche, M.A., Cohen, L.G., Wassermann, E.M., Priori, A., Lang, N., Antal, A., Paulus, W., Hummel, F., Boggio, P.S., Fregni, F., Pascual-Leone, A., 2008. Transcranial direct current stimulation: state of the art 2008. Brain Stimulation 1 (3), 206–223.

Nitsche, M.A., Paulus, W., 2001. Sustained excitability elevations induced by transcranial DC motor cortex stimulation in humans. Neurology 57 (10), 1899–1901.

Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., Winkler, I., 2001. "Primitive intelligence" in the auditory cortex. Trends in Neurosciences 24, 283–288.

Niedeggen, M., Roesler, F., 1999. N400 effects reflect activation spread during retrieval of arithmetic facts. Psychological Science 10, 271–276.

Obrig, H., Rossi, S., Telkemeyer, S., Wartenburger, I., 2010. From acoustic segmentation to language processing: evidence from optical imaging. Frontiers in Neuroenergetics 2 (13), 1–12.

Ogawa, S., et al., 1990. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proceedings of the National Academy of Sciences U.S.A. 87, 9868–9872.

Perani, D., 2012. The development of language neural networks. International Journal of Psychophysiology 85 (3), 336.

Petersen, S.E., Fiez, J.A., 1993. The processing of single words studied with positron emission tomography. Annual Review of Neuroscience 16, 509–530.

Petersen, S.E., Fox, P.T., Posner, M.I., Mintun, M., Raichle, M.E., 1988. Positron emission tomographic studies of the cortical anatomy of single-word processing. Nature 331 (6157), 585–589.

Petitto, L.A., Berens, M.S., Kovelman, I., Dubins, M.H., Jasinska, K., Shalinsky, M., 2012. The "perceptual wedge hypothesis" as the basis for bilingual babies' phonetic processing advantage: new insights from fNIRS brain imaging. Brain and Language 121 (2), 130–143.

Price, C.J., 2010. The anatomy of language: a review of 100 fMRI studies published in 2009. Annals of the New York Academy of Sciences 1191, 62–88.

Price, C.J., 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage 62 (2), 816–847.

Quaresima, V., Bisconti, S., Ferrari, M., 2012. A brief review on the use of functional near-infrared spectroscopy (fNIRS) for language imaging studies in human newborns and adults. Brain and Language 121 (2), 79–89.

Raichle, M.E., 2006. The Brain's dark energy. Science 314 (5803), 1249–1250.

Ross, L.A., McCoy, D., Wolk, D.A., Coslett, H.B., Olson, I.R., 2010. Improved proper name recall by electrical stimulation of the anterior temporal lobes. Neuropsychologia 48 (12), 3671–3674.

Rossi, S., Hallett, M., Rossini, P.M., Pascual-Leone, A., 2009. Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research. Clinical Neurophysiology 120 (12), 2008–2039.

Rossini, P.M., Rossini, L., Ferreri, F., 2010. Brain-behavior relations: transcranial magnetic stimulation: a review. IEEE Engineering in Medicine and Biology Magazine 29 (1), 84–95.

Saur, D., Kreher, B.W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M.-S., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, W., Rijntjes, M., Hennig, J., Weiller, C., 2008. Ventral and dorsal pathways for language. Proceedings of the National Academy of Sciences 105 (46), 18035–18040.

Schlaug, G., Marchina, S., Wan, C.Y., 2011. The use of non-invasive brain stimulation techniques to facilitate recovery from post-stroke aphasia. Neuropsychology Review 21 (3), 288–301.

Seniow, J., Waldowski, K., Lesniak, M., Iwanski, S., Czepiel, W., Czlonkowska, A., 2013. Transcranial magnetic stimulation combined with speech and language training in early aphasia rehabilitation: a randomized double-blind controlled pilot study. Topics in Stroke Rehabilitation 20 (3), 250–261.

Simonyan, K., Herscovitch, P., Horwitz, B., 2013. Speech-induced striatal dopamine release is left lateralized and coupled to functional striatal circuits in healthy humans: a combined PET, fMRI and DTI study. Neuroimage 70 (0), 21–32.

Steinhauer, K., Connolly, J.F., 2008. Event-related potentials in the study of language. In: Stemmer, B., Whitaker, H.A. (Eds.), Handbook of the Neuroscience of Language. Elsevier, San Diego, pp. 91–104.

Steinhauer, K., Drury, J.E., 2012. On the early left-anterior negativity (ELAN) in syntax studies. Brain and Language 120 (2), 135–162.

Stemmer, B., 2008. Neuropragmatics: disorders and neural systems. In: Stemmer, B., Whitaker, H.A. (Eds.), Handbook of the Neuroscience of Language. Elsevier, San Diego, pp. 175–188.

Stemmer, B. Neuropragmatics. In: Huang, T. (Ed.), Handbook of Pragmatics. Oxford University Press, Oxford, in press.

Suda, M., Takei, Y., Aoyama, Y., Narita, K., Sato, T., Fukuda, M., Mikuni, M., 2010. Frontopolar activation during face-to-face conversation: an in situ study using near-infrared spectroscopy. Neuropsychologia 48 (2), 441–447.

Sitnikova, T., Holcomb, P.J., Kiyonaga, K.A., Kuperberg, G.R., 2008. Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events. Journal of Cognitive Neuroscience 20, 1–21.

Steinhauer, K., 2003. Electrophysiological correlates of prosody and punctuation. Brain and Language 86, 142–164.

Steinhauer, K., Alter, K., Friederici, A.D., 1999. Brain potentials indicate immediate use of prosodic cues in natural speech processing. Nature Neuroscience 2 (2), 191–196.

Steinbrink, J., Villringer, A., Kempf, F., Haux, D., Boden, S., Obrig, H., 2006. Illuminating the BOLD signal: combined fMRI-fNIRS studies. Magnetice Resonance Imaging 24, 495–505.

Turkeltaub, P.E., Coslett, H.B., Thomas, A.L., Faseyitan, O., Benson, J., Norise, C., Hamilton, R.H., 2012. The right hemisphere is not unitary in its role in aphasia recovery. Cortex 48 (9), 1179–1186.

Turkeltaub, P.E., Messing, S., Norise, C., Hamilton, R.H., 2011. Are networks for residual language function and recovery consistent across aphasic patients? Neurology 76 (20), 1726–1734.

Utz, K.S., Dimova, V., Oppenlander, K., Kerkhoff, G., 2010. Electrified minds: transcranial direct current stimulation (tDCS) and galvanic vestibular stimulation (GVS) as methods of non-invasive brain stimulation in neuropsychology–a review of current data and future implications. Neuropsychologia 48 (10), 2789–2810.

de Vries, M.H., Barth, A.C.R., Maiworm, S., Knecht, S., Zwitserlood, P., Flöel, A., 2009. Electrical stimulation of Broca's area enhances implicit learning of an artificial grammar. Journal of Cognitive Neuroscience 22 (11), 2427–2436.

Vigneau, M., Beaucousin, V., Herve, P.-Y., Jobard, G., Petit, L., Crivello, F., Tzourio-Mazoyer, N., 2011. What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing? Insights from a meta-analysis. Neuroimage 54 (1), 577–593.

Vigneau, M., Beaucousin, V., Herve, P.Y., Duffau, H., Crivello, F., Houde, O., Mazoyer, B., Tzourio-Mazoyer, N., 2006. Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. Neuroimage 30 (4), 1414–1432.

Volpato, C., Cavinato, M., Piccione, F., Garzon, M., Meneghello, F., Birbaumer, N., 2013. Transcranial direct current stimulation (tDCS) of Broca's area in chronic aphasia: a controlled outcome study. Behavioural Brain Research 247 (0), 211–216.

Vul, E., Harris, C., Winkielman, P., Pashler, H., 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspectives on Psychological Science 4 (3), 274–290.

Wallois, F., Mahmoudzadeh, M., Patil, A., Grebe, R., 2012. Usefulness of simultaneous EEG-NIRS recording in language studies. Brain and Language 121 (2), 110–123.

Watanabe, T.K., Black, K.L., Zafonte, R.D., Millis, S.R., Mann, N.R., 1998. Do calendars enhance posttraumatic temporal orientation? A pilot study. Brain Injury 12 (1), 81–85.

Watkins, K.E., Devlin, J.T., 2008. Transcranial magnetic stimulation (TMS) as a tool for studying language. In: Stemmer, B., Whitaker, H.A. (Eds.), Handbook of the Neuroscience of Language. Elsevier, San Diego, pp. 115–124.

Watson, N.F., Dodrill, C., Farrell, D., Holmes, M.D., Miller, J.W., 2004. Determination of language dominance with near-infrared spectroscopy: comparison with the intracarotid amobarbital procedure. Seizure 13 (6), 399–402.

Weiller, C., Bormann, T., Saur, D., Musso, M., Rijntjes, M., 2011. How the ventral pathway got lost – and what its recovery might mean. Brain and Language 118 (1–2), 29–39.

Weiller, C., Musso, M., Rijntjes, M., Saur, D., 2009. Please don't underestimate the ventral pathway in language. Trends in Cognitive Sciences 13 (9), 369–370.

Zago, S., Ferrucci, R., Fregni, F., Priori, A., 2008. Bartholow, sciamanna, alberti: pioneers in the electrical stimulation of the exposed human cerebral cortex. The Neuroscientist 14 (5), 521–528.

Functional Data Analysis

James O. Ramsay, McGill University, Montreal, QC, Canada
Bernard W. Silverman, University of Bristol, Bristol, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 9, pp. 5822–5828, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Most statistical analyses involve one or more observations taken on each of a number of individuals in a sample, with the aim of making inferences about the general population from which the sample is drawn. In an increasing number of fields, these observations are curves or images. Curves and images are examples of functions, since an observed intensity is available at each point on a line segment, a portion of a plane, or a volume. For this reason, we call observed curves and images 'functional data,' and statistical methods for analyzing such data are described by the term 'functional data analysis.' It is the smoothness of the processes generating functional data that differentiates this type of data from more classical multivariate observations. This smoothness means that we can work with the information in the derivatives of functions or images. This article includes several illustrative examples.

Most statistical analyses involve one or more observations taken on each of a number of individuals in a sample, with the aim of making inferences about the general population from which the sample is drawn. In an increasing number of fields, these observations are curves or images. Curves and images are examples of functions, since an observed intensity is available at each point on a line segment, a portion of a plane, or a volume. For this reason, we call observed curves and images 'functional data,' and statistical methods for analyzing such data are described by the term 'functional data analysis' (FDA), coined by Ramsay and Dalzell (1991). Though the individual methods and techniques are important, FDA is also a general way of thinking, where the basic unit of information is the entire observed function rather than a string of numbers. A few illustrations will be offered in this article, but Ramsay and Silverman (1997) may be consulted for many more examples and methods, for consideration of the underlying philosophy, and for further references to the literature.

The goals of FDA are essentially the same as for other branches of statistics, and include the following:

1. to represent and transform the data in ways that aid further analysis,
2. to display the data so as to highlight various characteristics,
3. to study important sources of pattern and variation among the data, and
4. to explain variation in an outcome or dependent variable by using input or independent variable information.

We shall illustrate the nature of functional data, these goals, and FDA tools available or under development, through a series of examples.

Human Growth Data: Looking at Velocity and Acceleration

Figure 1 shows the growth patterns of 10 Swiss boys involved in the Zurich Longitudinal Growth Study (Falkner, 1960). Each boy's height is measured at 29 separate ages, but it is more instructive to think of this as a sample of 10 growth curves or functions than as a set of 290 numerical observations. Of particular interest in the study of growth is the way that growth speeds up and slows down at various stages of development, and this is much better illustrated by studying the velocity and acceleration of height. Figure 2 gives estimates of the acceleration or second derivative of these 10 height functions, and shows that around puberty there is a strong pulse of acceleration in growth, followed by sharp deceleration. This is the pubertal growth spurt (PGS), and there is also often a smaller spurt, hardly visible in Figure 1 but obvious in Figure 2, at around age 6.

The Mean Function and the Registration Problem

How would one estimate an overall acceleration curve to quantify the size of these spurts? The dashed curve in Figure 2 is the average acceleration over the sample at each time point, but the PGS for this average is smaller in size and longer in duration

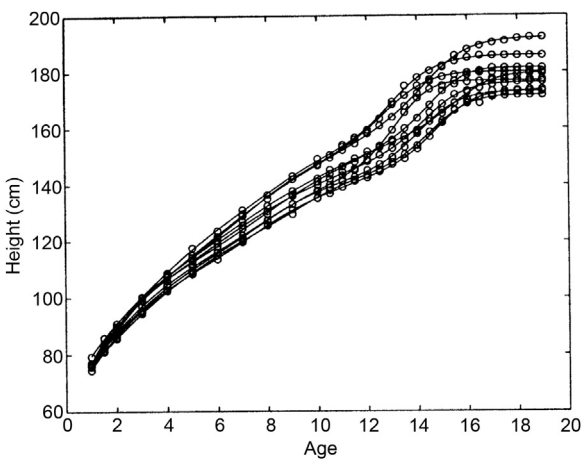

Figure 1 The heights of 10 Swiss boys measured at 29 ages. Notes: Points indicate unequally spaced ages of measurement; solid line is a function fitted to these data by monotone smoothing.

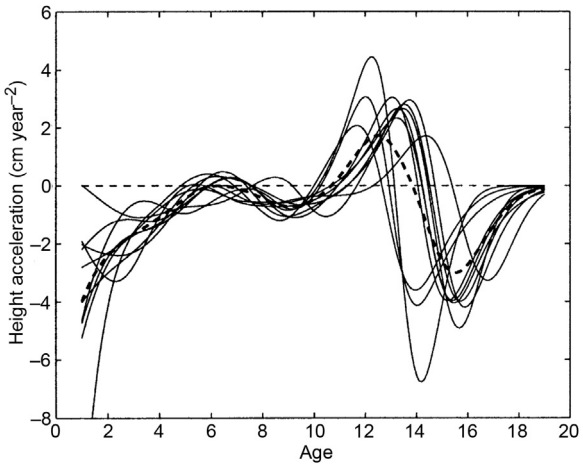

Figure 2 The estimated accelerations of height for 10 Swiss boys. Notes: Heavy dashed line is cross-sectional mean. This is a rather poor summary of the curves because of the presence of the timing of the pubertal growth spurt, called phase variation.

than that of any of the individual spurts. Averages are supposed to look like typical observations; so what went wrong?

The problem is that the curves vary in two ways, termed amplitude variation and phase variation. Alternatively, these are called vertical and horizontal variation, respectively, since amplitude concerns the height of a function at a given point, and phase the location in time of a feature such as an acceleration peak. In short, the actual acceleration peaks in the PGS vary in both size and timing. An artificial example demonstrating these two sorts of variation is presented in **Figure 3**. The phase shift displayed in this figure is constant across the timescale, but many data sets require more complicated and nonlinear transformations of the time axis in order to align curve features like the peaks and valleys in **Figure 3**. The presence of phase variation arises in FDA because many processes involve a system time, or in this case biological time, that does not unfold at the same rate as clock time. In this case, the biological clock of each boy marks off time at its own unique rate.

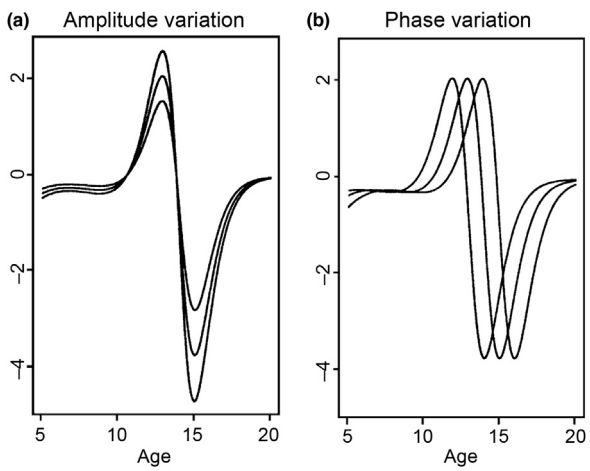

Figure 3 Panel (a): three height acceleration curves varying only in amplitude; Panel (b): three curves varying only in phase.

A key part of the FDA methodology is time-warping or registration of the data curves to align specific features or to minimize variability, since estimating the mean and most other analyses are thrown off by the presence of phase variation. The time-warping functions may be simple constants, linear functions, or more complicated, nonlinear functions. The amount or shape of warp may be of great interest (see Ramsay and Silverman, 1997: Chapters 5 and 8). **Figure 4** shows time-warping functions and registered height acceleration curves for the growth curves of **Figure 1** estimated using a method developed by Ramsay and Li (1998). The time-warping curve for each individual shows his maturation pattern. A warping function consistently above the diagonal dashed line implies that clock time lags behind development, and therefore development is late in clock terms. If height is measured relative to biological age, the curves in **Figure 2** are aligned as shown in **Figure 5**. Differences in height acceleration after registration correspond to pure amplitude variation in growth

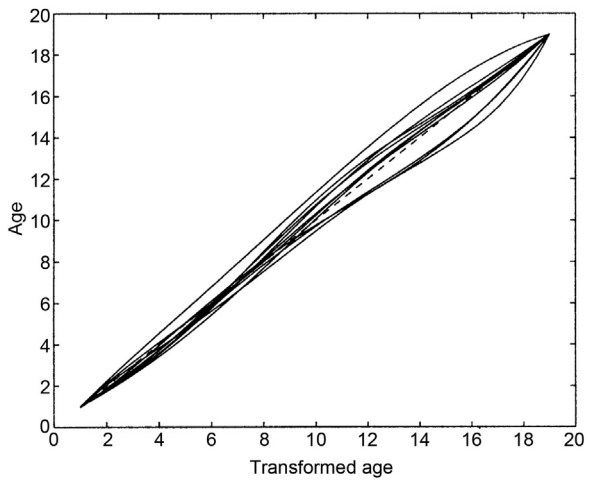

Figure 4 Time-warping functions taking chronological age on the horizontal axis into biological age on the vertical axis.

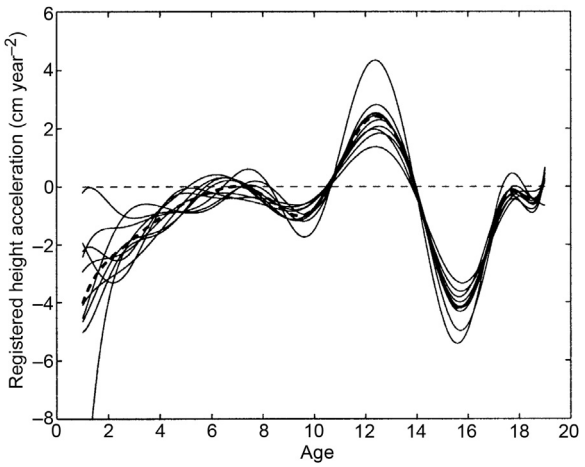

Figure 5 The height acceleration curves in **Figure 2** aligned or registered by measuring height with respect to the biological time shown in **Figure 4**. Notes: Heavy dashed line is the mean acceleration. This provides a much better summary of the actual curves than **Figure 2**.

patterns. The mean of the registered curves in Figure 5 is a much better summary than the mean of the unregistered data in Figure 1.

The Nondurable Goods Index and More Derivatives

Valuable information may be gained by plotting one derivative of a functional observation against another, a technique called phase/plane plotting. The US index of nondurable goods production is an important economic indicator, and considerable insight may be gained by plotting the acceleration of this index against its rate of increase. If the index had a smooth trend plus a sinusoidal seasonal variation, this plot would be approximately circular with suitable choice of scales. The typical year 1970 in Figure 6 has three main cycles, of which the spring cycle is rather smaller than the summer and fall/winter cycles. Comparisons of plots for different years is instructive. During the stable period 1964–66, there is little year-to-year change, but the period 1973–75 in Figure 6 shows considerable instability, due primarily to the end of the Vietnam War. A typical plot for a more recent year, 1994, shows that the whole pattern has changed, indicating structural changes in the methods and economics of production. For further details, see Ramsay and Ramsey (2002).

Functional Principal Components Analysis

Many FDA methods are adaptations of classical multivariate methods such as principal components analysis (PCA), linear modeling, and analysis of variance (see also Linear Hypothesis; Multivariate Analysis: Discrete Variables, Overview). Functional PCA demonstrates the way in which a set of functional data varies from its mean, and, in terms of these modes of variability, quantifies the discrepancy from the mean of each individual functional datum. Figure 7 (Figure 6.2 in Ramsay and Silverman, 1997) shows the first four principal component weight curves for temperature records for 35 Canadian weather stations. The first mode of variability is overall temperature, but with larger effects in the winter; the second corresponds to variation in annual range of temperature; the third partly to a time-shift of the overall cycle, and the fourth to short, hot summers and long, mild winters, or vice versa. Individual weather stations can then be characterized according to their scores on these four criteria. The interpretability of the components can often be improved by a suitable rotation of the original principal components; see Ramsay and Silverman (1997: Chapter 6).

A Functional Linear Model and Regularization

An example of functional linear modeling is provided by the study of the dependence of the acceleration of the lower lip (lipacc) in speech on neural activity, as measured by electromyographical (EMG) recording. It is reasonable to suppose that the neural activity at time s continues to affect lip movement for only a short period of time, which we denote by δ and this leads to a model of the form

$$\text{lipacc}(t) = \alpha(t) + \int_{t-\delta}^{t} \text{EMG}(s)\beta(s,t) + \varepsilon(t)$$

developed by Malfait et al., (2000). The bivariate function $\beta(s, t)$ plays the same role as the regression coefficient in standard linear regression, and quantifies the relationship between the

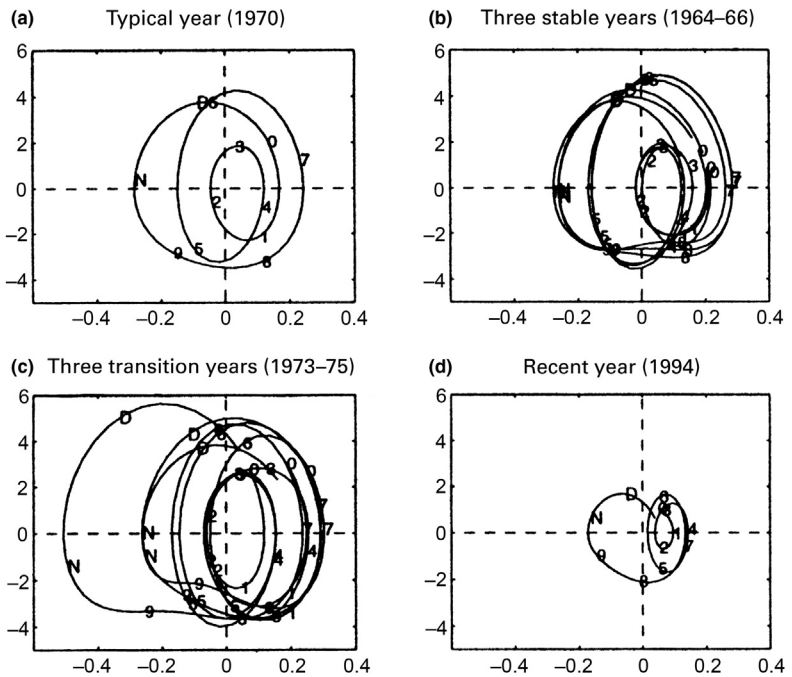

Figure 6 A phase/plane plot of *US index of nondurable goods production*. Each panel plots second derivative on vertical axis against first derivative on horizontal axis. Plots show structure of seasonal variation in index, and how structure varies with time.

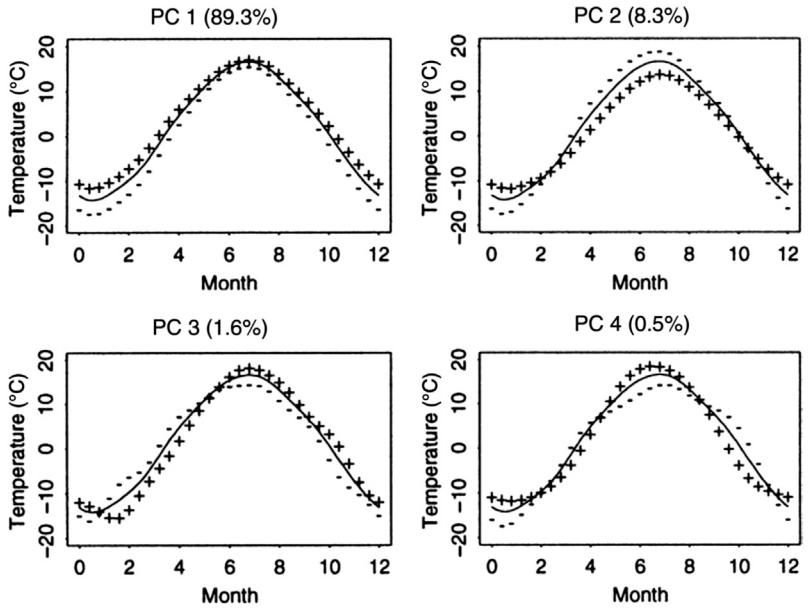

Figure 7 A principal components analysis of mean temperature curves for 35 Canadian weather stations. Notes: Solid line is the mean curve taken across all stations. The effects of adding (+) and subtracting (−) a suitable multiple of each component are also shown.

independent functional variable EMG, and the dependent functional variable lip. One particular issue is the appropriate choice of the time lag δ the length of time within the model that neural activity can continue to affect movement. Models for increasing values of δ correspond to nested linear models, and in this case a study of the amount of variation explained demonstrates that there is little effect beyond a lag of about half a second.

An important theoretical and practical issue is the necessity or otherwise of incorporating some form of smoothing or regularization into a functional analog of a classical multivariate procedure. In principal components analysis, the principal components, such as those displayed in Figure 7, are smoothed or regularized and in some procedures, such as canonical correlation analysis (see also Multivariate Analysis: Discrete Variables, Overview) or certain types of linear modeling for the lip/EMG data, regularization is necessary to give any meaningful results at all; see Leurgans et al., (1993), and Ramsay and Silverman (1997: Chapters 11 and 12).

Modeling with Derivatives: A Central Theme

It is the smoothness of the processes generating functional data that differentiates this type of data from more classical multivariate observations. This smoothness means that we can work with the information in the derivatives of functions or images. We looked at acceleration in the growth data, found something interesting in the joint variation of the velocity and acceleration of nondurable goods production, and used derivatives to define roughness penalties; see Ramsay and Silverman (1997: Chapter 4) for the weather data.

We may also use derivatives to construct models for data. Such models are called differential equations, and are part of the standard model-building toolkit in the natural sciences and engineering. Differential equation models are often called dynamic models, since they are especially effective at representing processes unfolding in time. Ramsay (2000) developed a linear differential equation for handwriting involving velocity, acceleration, and the third derivative of the tip of the pen. Figure 8 shows a sequence of pen positions during the writing of the word 'statistics' in Mandarin, and also shows as a solid line its approximation on the basis of the dynamic equation estimated from 50 such samples of Chinese script. The equation not only captured the essential features of this particular sample, but was also able to model a fair amount of the sample-to-sample variation as well. And, finally, it also provided an excellent account of the velocity, acceleration, and jerk functions at the same time. Ramsay and

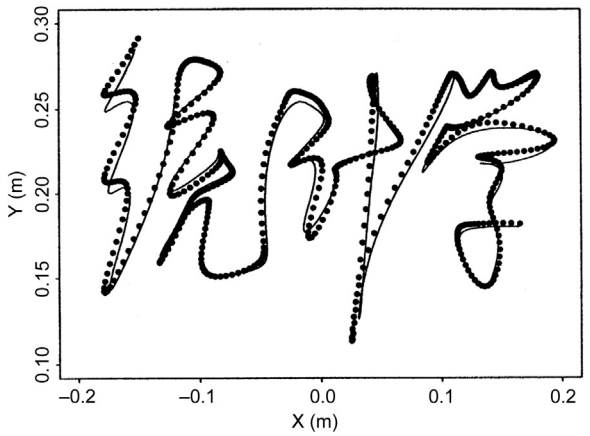

Figure 8 Observations of the tip of the pen while writing the word 'statistics' in Mandarin. Notes: Dots indicate 2400 observations; solid line is an approximation based on a second-order linear differential equation.

Silverman (1997: Chapters 13 and 14) take up this topic of dynamic modeling with functional data, including an account of a technique called 'principal differential analysis,' developed by Ramsay (1995).

Relationships with Other Branches of Statistics

Of course, statisticians have been exploring curves for a very long time, and there are many older statistical methodologies that can continue to inform and enrich an FDA. First of all, FDA usually begins with some kind of nonparametric regression (e.g., see also Exploratory Data Analysis: Univariate Methods) in order to replace discrete and possibly noisy curve values by smooth, continuous functions. Density estimation also involves data smoothing. Typically, time series analysis (Brillinger, 1981; Box and Jenkins, 1994) considers long sequences of curve values corresponding to equally spaced time values that are more or less stationary; that is, the data show little overall trend, and the covariance among neighboring values does not change much from one time region to another. However, methods for nonstationary processes are also under active development, and there are strong structural links between times series methods such as state-space modeling and FDA methods such as smoothing. Longitudinal data analysis or multilevel modeling (Diggle et al., 1994; Searle et al., 1992) usually looks at shorter sequences of curve values sampled over many cases, and studies the structure of between- and within-curve variation. It seems likely that this field and FDA will merge in a number of useful ways. Multivariate statistics has many methods developed specifically for functional data that have preceded FDA, but which have not explicitly exploited the smoothness of functional data or made other uses of derivatives (see also Probability Density Estimation; Longitudinal Data: Event–History Analysis in Discrete Time; Multivariate Analysis: Discrete Variables, Overview).

Conclusion

By the smoothness of a functional observation we have meant the possibility that one or more derivatives, such as velocity or acceleration, may be examined or used in the modeling process. Otherwise, FDA makes only minimal assumptions about the structure of variation within and between curves. We have seen, however, that teasing apart amplitude and phase variation is a central issue, and one that has largely been ignored until recently. Certain types of models considered to be fairly routine and well understood in multivariate statistics, such as linear regression, have many more variants in the functional domain, and are being researched actively. Software for FDA is also being developed, and it is likely that FDA will soon be considered as a routine item in a data analyst's black bag of instruments and nostrums. Further information on FDA and related issues may be obtained in Pinheiro and Bates (2000), Ramsay and Silverman (1997), and Simonoff (1996).

Bibliography

Box, G.E.P., Jenkins, G.M., 1994. Times Series Analysis: Forecasting and Control, third ed. Prentice-Hall, Englewood Cliffs, NJ.

Brillinger, D.R., 1981. Time Series: Data Analysis and Theory. Holden Day, San Francisco.

Diggle, P.J., Liang, K.-Y., Zeger, S.L., 1994. Analysis of Longitudinal Data. Clarendon Press, Oxford, UK.

Falkner, F. (Ed.), 1960. Child Development: An International Method of Study. Karger, Basle, Switzerland.

Ferraty, Frédéric, Vieu, Philippe, 2006. Nonparametric Functional Data Analysis: Theory and Practice. Springer, New York.

Leurgans, S.E., Moyeed, R.A., Silverman, B.W., 1993. Canonical correlation analysis when the data are curves. Journal of the Royal Statistical Society, Series B 55, 725–740.

Malfait, N., Ramsay, J.O., Froda, S., 2000. The Historical Functional Linear Model (Unpublished manuscript). McGill University, Montreal, PQ.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-PLUS. Springer, New York.

Ramsay, J.O., 1995. Principal differential analysis: data reduction by differential operators. Journal of the Royal Statistical Society 58, 495–508.

Ramsay, J.O., 1998. Estimating smooth monotone functions. Journal of the Royal Statistical Society, Series B 60, 365–375.

Ramsay, J.O., 2000. Functional components of variation in handwriting. Journal of the American Statistical Association 95, 9–15.

Ramsay, J.O., Dalzell, C., 1991. Some tools for functional data analysis (with discussion). Journal of the Royal Statistical Society 53, 539–572.

Ramsay, J.O., Li, X., 1998. Curve registration. Journal of the Royal Statistical Society, Series B 60, 351–363.

Ramsay, J.O., Ramsey, J.B., 2002. Functional data analysis of the dynamics of the monthly index of non-durable goods production. Journal of Econometrics 107, 327–344.

Ramsay, J.O., Silverman, B.W., 1997. Functional Data Analysis. Springer, New York.

Searle, S.R., Casella, G., McCulloch, C.E., 1992. Variance Components. Wiley, New York.

Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer, New York.

Yao, Fang, Müller, Hans-Georg, Wang, Jane-Ling, 2005. Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association 100 (470), 577–590.

Functional Explanation: Philosophical Aspects

Graham Frank Macdonald, University of Canterbury, Christchurch, New Zealand; and University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the philosophy of biology and philosophy of psychology and social science there has been considerable controversy over what the correct account of functional explanation is, with the contending accounts splitting into two main camps. Both camps agree that functional explanation is dependent on an account of functions, but disagree on the account they give. Within evolutionary biology, the favored view is that functions emerge over time as a consequence of evolutionary processes, a feature of an organism only having a function if it has a suitable history. This version is aptly called the historical, or etiological, view of functions. The opposed view is that the operation of a system can be analyzed to reveal how parts of the system enable the system to behave in a certain manner. This version is best known as a 'systems' or 'Cummins'' account of functions and functional explanation. This article examines both these views. In addition, the controversy over what it is that functional explanations explain will be outlined. Along the way, the notion of 'malfunction' will be defined, the criticisms of panadaptationism scrutinized, functional explanation of novel features or behavior defended, and the claims of dispositional or 'virtual' selection inspected.

Functional explanation has received a great deal of philosophical attention in recent years, partly as a consequence of a surge of interest in the philosophy of biology, and partly as a consequence of attempts to make functions and functional explanation central to issues in the philosophy of psychology. In both these areas, there has been some controversy over what the correct account of functional explanation is, with the contending accounts splitting into two main camps. Both camps agree that functional explanation is dependent on an account of functions, but disagree on what the correct account is. Within evolutionary biology, the favored view is that functions emerge over time as a consequence of evolutionary processes, so a certain feature of an organism will only have a function if it has a suitable history. The opposed view is that the operation of a system can be analyzed into the way in which parts of the system function to make the system behave in that manner. This version has many names, but is perhaps best known as a 'systems' or 'Cummins'' account of functions and functional explanation, but it has also been called 'causal-role' functionalism. In what follows, both these views will be examined; in addition, the controversy over what it is that functional explanations explain will be outlined.

Functions and Systems

The 'systems' account of functions takes as its starting point the fact of (biological) complexity and proposes to understand the complex functioning of a biological entity (or system) in terms of the working of its parts. The fact of complexity suggests that the systems are goal directed, and if this is taken for granted one can recover what is distinctive of many explanations in biology, that they invariably cite the effects of causes as explaining why the cause is there. It does this by specifying how a part of the system operates in the production of the aimed-at goal.

The best known account of how such an explanation operates is given by Cummins (1975, 1983). He distinguishes explanation of changes, which is done by 'transition' theories, from explanation of properties. Transition theories are suitable for explaining 'What caused S to acquire P?' whereas a property theory explains what it is for S to instantiate P. This is best done by constructing an analysis of S, such an analysis adverting to the properties of S's components and their organization. So the kinetic theory of gases explains what it is for a gas to have temperature via an account of the properties of the molecules contained within the gas. Cummins calls the analysis of such a system compositional analysis: the analysis of a dispositional property being called functional analysis. The general claim is that in science the ascription of a function to something is to specify a capacity, such a capacity being identified by its role in enabling a system to operate in a certain manner. Thus, in biology, the capacities of an organism are explained by analyzing them into a number of systems (circulatory, digestive, etc.), such systems having a particular task to perform. How these tasks get performed is in turn explained by functional analysis, such an analysis citing the capacities of parts of these systems. Such capacities are the functions of the parts, so we get the Cummins-style definition of function:

> X functions as an F in s (or: the function of X in s is to F) relative to an analytical account A of s's capacity to G just in case X is capable of F-ing in s and A appropriately and adequately accounts for s's capacity to G by, in part, appealing to the capacity of X to F in s (Cummins, 1975, p. 762).

This account clearly does not restrict the attribution of functions to biological organisms; one can recognize the attribution of this type of function to parts of societies within sociological and anthropological functionalism. Here, the system was seen as a society or culture, and the parts were the subsocietal structures or the recurrent patterns of behavior, the function of which was to maintain the structural continuity of the whole. The fact that this account of functionality is applicable to any system has given rise to two criticisms of this approach, one seeing a problem in how to specify non-arbitrarily which behavior or capacity of the containing system

is relevant for the functional analysis to analyze, the second seeing a difficulty in producing a sufficiently determinate notion of a 'system' for which a functional analysis is appropriate. The problem is that the choice of both a capacity and a system may seem to depend upon our decisions in a way that could make the functional analysis of any capacity or system seem justifiable. The pertinent question here concerns the explanatory gain made by the appeal to functional analysis. There seems to be no explanatory payoff in some cases where functional analysis may be applied. Take the 'system' that is my study, which includes as parts various items of furniture, books, and me. The property of the room that one wants explained is its temperature. Say my presence in the study raises the temperature of the room by 1/2 a degree. One does not gain any insight into the working of this system by attributing to me the function of enabling the room (the system) attain its present temperature.

Cummins was aware of this trivialization of functional analysis, and suggested constraints on the choice of a system and capacity that could be analyzed illuminatingly. One problem with the above analysis is that the capacity to be explained via functional analysis, the temperature of the room, is similar in kind and complexity to the property doing the explaining, my body temperature. One only gets an explanatory payoff when the analyzing property is both less complicated than, and different in kind from, the analysandum. This suggests the third constraint imposed by Cummins: one is not going to get a less sophisticated property doing all the explanatory work in showing how a complex system operates. The analysis will have to appeal to the operation of component parts and their properties, their combined action producing the complex behavior of the system. It is through the organization of interdependent parts that simplicity generates complexity, and functional analysis succeeds when it shows how this can be achieved.

The choice of system and capacity to be analyzed, then, is constrained: functional analysis will be applied fruitfully only to a complex system whose behavior is produced by the action of simpler but well-organized parts. It is unsurprising, then, that the systems most often favored as exemplars of systems suitable for functional analysis are those that have been designed. The interdependence of the parts of such systems is often attained by careful planning. The second theory of functions that will now be examined focuses on this 'design' aspect of the systems to whose parts functions can be aptly attributed.

Functions and Selection

It is the improbability of complex systems existing by chance that suggests that such systems are generated by design, and so have parts that function so as to produce the intended behavior of the system. With artifacts, it is easy to see how functionality can be ascribed: the complex system is designed intentionally to have the capacity to produce certain effects, and the parts are chosen for their ability to contribute toward that goal. They are selected for that task. This type of functionality is derived from intentionality. The achievement of Darwin and Wallace was to provide a nonintentional explanation of how such complexity could originate, and they did this via the theory of natural selection. The major strength of the historical account of functions is that it ties attribution of functions to 'design' by natural (nonintentional) selection. The essential features of this account are dependent upon the idea that to attribute a function to an item is to say that the item has that feature or behavior because it has a certain history: in the past that feature produced certain consequences, those consequences themselves having the effect of reproducing items with that feature. It is this latter aspect that makes the account teleological. One can say that the item has the relevant property in order to produce those consequences. It is because of (some of) its effects that the item with this capacity is selected for. What the theory of natural selection does is make respectable the apparent purposefulness of the 'in order to.' It does this by showing how environmental pressures on variable items can produce varied reproduction of those items; those items with certain features will become more prevalent because those features have typical effects on the reproductive capacity of the items that bear them.

The long neck of the giraffe, to take a standard example, is an adaptation and has the function of enabling giraffes to eat the leaves at the top of tall trees. On the historical account, this capacity is the function of the long neck because previous giraffes whose genes produced longer necks than their contemporaries were better producers of offspring, and were so just because of the advantage conferred on those giraffes in the food wars. This results, in the long run, in more long-necked giraffes, until the giraffe is a long-necked species. Put as briefly and generally as possible, the etiological account says that functionality arises because some variants within a group (a mutant gene, for example) have (or produce) novel features with capacities that are favorable to their possessors' ability to reproduce (i.e., more favorable than the nonpossession of those features by the other members of the group). Such features are transmitted to their descendants, proliferating within the group in the process. The feature of the organism will then, and only then, has as its function the exercise of the favorable capacity.

As with Cummins-style functions, this account of functionality is very general. It applies not just to biological entities, but to all items that have the required characteristics of

1. variation in the characteristics of members of the population,
2. which causes different rates of production, and
3. which is inherited. (Godfrey-Smith, 2007: p. 515)

Selection cannot take place if there is no initial variation; the same selection forces operating on a homogenous population will have no effect. For a cause to be there in order to produce certain effects it must also be the case that instances of properties with those causal powers have been selected in the past because they produced those effects. The selection process requires that the causal property interact with salient features of the environment in such a way that other items without that property are seen to 'suffer' some loss as a consequence of their nonpossession of the property. Finally, it is because the favored properties are transmitted to descendants of the items having those features that the present items have these properties. Then it can be said that the property is, or was functional, its

function being to produce those types of effects that lead to its ancestors reproducing items with the property in question.

As the giraffe example indicates, it is in its forging a connection between function and adaptation that the strength of this account resides. Adaptations are produced historically via the operation of natural selection, so they are functional. A creature may have a feature that renders it 'adapted' to its environment because it so happens that that feature will help it flourish in that environment, but such a feature, lacking the relevant selectionist history, will be adaptive for the creature, but not an adaptation. The historical (etiological) theory of functions claims that it is because adaptations are selected, not accidental, features of organisms that we are entitled to the use of functional explanation.

The historical and systems account of functions are not necessarily competitors. It has been seen that the systems for which functional analysis is appropriate are complex, with the analysis proceeding by showing how the operation of simpler, interdependent parts can produce the relevant capacity of the system. Such systems, it can be claimed, come into being either by intention or by selection, so functional analysis will be appropriately applied in just those cases where some design-mechanism is operating. On this view, functional analysis is an accessory to the historical theory of functions. There are, however, two aspects of the historical conception that render it significantly different from the systems version. The first is a consequence of what has been suggested above: it is probable that complex systems are generated by some design process, but it is not impossible for such a system to emerge by chance. For such a system, a functional analysis would still be appropriate, but the historical version of functions would disallow the attribution of any function to the system or its parts, these not having the appropriate history. Second, the historical theory has the ability to include malfunctions within its purview. On the systems account, if a part 'functions' so as to produce a system with a different capacity from that expected, that part cannot be said to be malfunctioning. A functional analysis will proceed as usual, with the new capacity being analyzed, and the analysis showing just how it is produced given the nature of the contained parts, and their causal powers. On the historical conception, given a suitable account of reproduction, one can allow for mistakes to be made in the reproducing process, so that the items that are considered to be descendants of some common ancestor need not be perfect copies of that ancestor, and in particular may not have the causal powers for which items of that type have been selected previously. The item may then not do what it is supposed to do – it may malfunction. To appreciate this better, and to make manifest the power of the explanations that the historical theory makes available, that theory needs to be fleshed out in more detail.

Elaborations: Malfunction, Novelty Proper and Actual Domains

Malfunction

An essential aspect of the historical account is that a function results from selection operating over a period of time on items that are descended from each other. What is it to be descended in the relevant sense? Initially one may think in terms of copying, one item being a descendant of another if it is an exact copy of the former, and is so as a result of some lawlike process that ensures that such a replica is forthcoming. This is too strict a definition: for one thing, it would be impossible to provide a decent conception of a malfunctioning device if all descendants had to be exact copies of their ancestors. A malfunction involves some mistake, so a device that malfunctions will not be an exact copy of a properly functioning device. One has to allow for an item being a descendant of another even though the copying process is not perfect. One way of making what is involved more precise is due to Millikan (1984, 1993), and it goes like this: individual A is a reproduction of individual B if and only if B has some determinate properties in common with A, and this correlation of properties can be explained by a natural law operating in that environment. These common properties are the reproductively established properties of B, and the devices sharing these properties form reproductively established families (refs). (Only first-order refs reproduce in this manner. Higher order refs (horefs) are produced by first-order refs. My and my ancestors' two-leggedness is the result of common genes, such genes being first-order refs. They have certain properties in common, this being ensured by the laws of genetic inheritance.) Direct proper functions are functions of devices that are members of refs or horefs.

So how do we get malfunctioning members of refs or horefs? Millikan suggests that a malfunctioning member of a ref can be produced provided it has been produced by a device the proper function of which is to produce members of the ref in question, that it is in some aspects similar to other members of the ref, and that its production has an explanation that approximates (in some undefined degree) to a normal explanation for the production of members of the ref. Because a particular reproductive process need only approximate the 'perfect' reproductive process, it can produce imperfect copies, ones that malfunction.

Novelty

The historical theory's emphasis on the role of reproduction and selection has given rise to the thought that functional explanation, construed in this way, cannot explain novel behavior. On the historical account, it is inheritance from past (selected for) patterns of behavior that gives rise to the present functional behavior. However, to suppose that novel behavior cannot be produced by something whose function it is to produce that behavior is to underestimate the variety of functions. Millikan notes that there can be a relational proper function, which is a function to do something bearing a certain relation to something else. The chameleon's skin-color mechanism has the relational function of making the chameleon's skin match that of its environment (whatever that color may be). Given a specific color to adapt to, the mechanism then acquires an adapted proper function. So when it is sitting on a brown plant, the adapted proper function of the device is to produce brown skin. Crucial here is that the brown skin may have never been produced before, so it is not a member of a ref, and the production of it is not a direct proper function of the device. That to which an adapted device adapts is its adaptor (the brown plant being the adaptor in this case). One can extend this picture to include the functions of those who may

be consumers of the output of an adapted function. Consider the dance of the bees, which acts as a signal to other bees, 'telling' them to go southeast, where the nectar is to be found. This dance is adapted to the location of the nectar, so has an adapted function. The response it calls forth, the flight of the bees to the southeast, is adapted to this adapted function. It is not a short step to extend this approach to the communicative linguistic and other cultural capacities of human beings. Complications abound, but the fundamentals for doing this are in place. For the extension to be seen as a realistic prospect, two important features of these extensions of the historical approach need to be noted. The first is that functional explanation includes a normative aspect: a functional property is there in order to produce an effect, the effect being one it ought to cause, given that the property is instantiated here because past instantiations have produced that effect. Malfunctioning items are not doing what they ought to do. The extent to which functional explanation can account for psychological and sociocultural properties is partly dependent upon whether the normativity intrinsic to these latter properties can be explicated by the norms of functional explanation.

The second is that the above extensions indicate the explanatory reach of that theory. Any items (including patterns of behavior (e.g., rituals), social structures, ideas, 'memes'), can be the bearer of functions, provided that the basic conditions for this are met: variety, varied reproduction due to that variety. The reproduction need not be gene-centered, nor even broadly 'biological.' The transmission of bodies of knowledge and ways of behaving across generations is a form of reproduction, and so these items are possible bearers of functional properties. Whether they are or not depends on whether selection operates on these reproduced items, and here it is important to note that selection can take many forms. As behaviorists emphasized, behavior can be molded by painful experience, so pain is one selector. Emotional experience can be another. These have been called 'vicarious' or 'secondary' selectors (Campbell, 1960), indicating that they themselves have been selected for by natural selection, their purpose being to be local, less severe eradicators of those types of behavior that are inimical to the flourishing of the organisms that produce the behavior.

Given that reproduction and selection can be ubiquitous, it would appear that functional explanation has a much wider reach than once thought. Recent attempts to explain cultural evolution have utilized the flexibility afforded by the recognition of different 'selection regimes,' and so some cultural evolutionists talk of cultural and biological systems containing their own, interacting, evolutionary trajectories, thus resurrecting functional explanations in sociology but doing so with a selectionist foundation. Recent attempts to extend the Darwinian paradigm to culture include Boyd and Richerson (1985), Richerson and Boyd (2005), Cziko (1995), Dennett (1995), Runciman (2009), Heinrich and McElreath (2003), Hodgson and Knudsen (2010), Mesoudi (2011). For a discussion of functional explanation in the context of a Marxist theory of history, see Cohen (1979).

Proper and Actual Domains

One task of functional explanation is to assist in the explanation of why an animal, say, is exhibiting snake-avoiding behavior. If the behavior has been selected for because ancestors of those animals who behaved in that way in appropriate snake-populated environments lived longer, and so reproduced more prolifically, than those who did not, then one explains snake avoidance via its functionality. This suggests that the functional explanation is limited to only those behaviors targeted at types of inputs past tokens of which are responsible for the selection of the behavior – in this case, the presence of snakes. Some snakelike branches, however, might trigger the same behavior, and it is plausible to explain that behavior in the same way, as the triggering of a response that has been selected for. The snakes are said to form the proper domain of the function; the merely snakelike shapes are the actual domain (see Sperber and Hirschfeld, 2004 for a discussion of the proper domain–actual domain distinction).

There have, however, been suggestions that the explanatory scope of functional explanation has become too broad, that it is being abused by those who tend to see functionality everywhere, and that it is too narrow, not encompassing explanations of items that have no past history of selection, but are nevertheless beneficial to the organism. In the next section, these criticisms of the historical theory are reviewed.

Difficulties and Additions

Panglossianism

The etiological theory is supposed to explain the persistence and spread of a trait in terms of its adaptive history, such explanations involving historical claims as to why the feature being explained was adaptive, and then tying this adaptation to its present function. At times, it is extremely difficult to trace the relevant history, and the temptation is to ascribe functionality anyway, with the hope that the historical detail will be forthcoming. This has led to the accusation that such explanations are 'Panglossian,' presuming that this world is the 'biological best' world, one in which all features of organisms have some function or other (see Gould and Lewontin, 1978). This accusation is particularly relevant to the sociobiological explanations of psychological and social behaviors, as the relevant history supporting these explanations is often unknown. The complaint is a welcome caution against epistemologically unconstrained attributions of function, but does not attack the heart of the historical theory. It simply warns against its abuse.

Exaptations

A related worry about attributing functions incorrectly, also voiced by Gould (Gould and Vrba, 1982), concerns a trait that may be adaptive for one reason, but then change function in midstream, either doubling up in its functions or (because of environmental change may be) sloughing off the old function in favor of the new, without any change of structure. Now the present persistence of the trait cannot be explained via its previous adaptive history, as that would give a misleading account of its present function. Such a change in function has been called an exaptation. An example of such an exaptation is the wing feathers of birds, the function of which we now think is to enable birds to fly. A

selectionist story, however, would have to say that their function was as a thermal regulator, as that is what they were selected for originally. Clearly, some complication of the original selectionist theory is called for, but the complications are hardly unexpected. Causal explanation and accounts of causality in general have difficulties dealing with causal overdetermination, and what one has in the above examples is functional overdetermination. What one needs to say, in the functional case, is that the functions of an item are those fitness-enhancing effects that advantage (or advantaged) the bearer of the item relative to rivals, and that have been reproduced differentially in the relevant population. With the change of function of feathers from thermal regulation to flight, the stage at which some feathers produced a flight response would be critical in noting the acquisition of the new function. As soon as the feathered and flighted birds became preponderant, then one can say that the feathers had acquired a new function.

Dispositional and Virtual Selection

Once one notes that selection can be recent, a question arises as to how recent selection can be for functionality to be correctly attributed. To some extent, the answer will be vague; there seems no exact stage at which one can say, before now there was no function, but now there is. But given that selection can be recent, a natural suggestion will be that present selection suffices for functionality. The radical proposal is that history can be irrelevant, that the essential component in the historical theory is the contribution of a feature to the reproductive success of the item possessing that feature, and that a trait without any history of selection can be seen to be useful for its possessor in the precise sense of making a (future, or possible) positive contribution to the survival or reproductive capacity of the organism. This 'dispositional' or 'virtual' selectionism involves the recognition that selection does play some role in the attribution of functionality, but it need not be actual selection. Differential reproduction need not have taken place. On this account, a trait is functional because of its potential contribution to the fitness of the organism possessing the trait, given the environment in which the organism operates.

A weak version of this view can be developed by requiring the environment to engage actively with the trait, so that the trait can be seen to contribute positively to the organism's fitness. On this view, suitably developed, the trait is seen as being currently under a selection pressure. That is, it is seen as potentially contributing to the differential reproduction of those organisms with the trait, where 'potentially' only means 'barring accidents, it will contribute to the differential reproduction of those organisms.' So the weak view agrees with the historical view in all respects except that it allows for functionality prior to differential reproduction. The stronger interpretation does not require actual engagement with the selection-relevant environment; it only requires that there are factors present in the environment that could influence the organism's chances of reproduction and survival.

The crucial question facing these amendments to the historical theory is: what is the advantage in using the notion of function here? What does it add to a simple nonfunctional causal explanation of what is happening? On the historical view, part of the distinctive force of functional explanations derives from the normative elements implicated in attributions of function, the idea that the functional trait is there in order to do something. The selectionist story, cashed out in terms of differential reproduction, makes this normativity naturalistically respectable. Without the selection occurring, this distinctive feature of functional attribution is threatened.

It is clear that these accounts cannot give any functional explanation for the existence, proliferation, or shape of the trait. Take the first generation of long-necked giraffes. The existence of the long necks (their proliferation) cannot be given an explanation relating to the reproductively beneficial effect, which the long necks have had in the past. The only explanation seems to be a straightforward causal explanation in terms of mutations. A suggestion is that one change the explanandum, which is what Pettit (1996) does in his advocacy of virtual selection. The idea is that there can be an explanation of a fact that is not an explanation of its existence or actual persistence, but an explanation of why it would persist in certain nonactual conditions, an explanation of its resilience. This would be an explanation of the robustness of the fact under various hypothesized circumstances. An example given is that of equilibrium explanation, where the explanation of, say, a certain ratio of prey to predator in a given environment need not explain why this ratio (which is in fact an equilibrium) has come about – why it exists – but why, if disturbing influences were brought to bear upon it, it would return eventually to this ratio. It is in this sense that the resilience of the pattern is being explained. With functional explanation, the thought is that the absence of the operation of the function, the absence of the 'good' effect, may itself initiate certain mechanisms or processes that will reinstate the function's effect. So independently of how the trait performing the function emerged or proliferated, so long as it has this modal property of being reinstatable (in certain conditions) when disturbed, then the attribution of function will be in order. There will be something for that attribution to explain, namely, the resilience of the pattern.

The problem for the advocate of functions here is that a normal causal story, perhaps one with feedback built into it, seems adequate to the task of explaining what is going on. The notion of function adds nothing, or at least, the plausibility it has derives only from the systems idea of functions. There is a system here that works so as to maintain an item, or a pattern of behavior, and any subsystemic processes that work so as to maintain the item or behavior 'functions' to that end. There is nothing more that functional explanation does here, and, so far, this notion of functional explanation does not have the distinctive normative character that the historical account imparts to functionality.

One could, however, construe the notion of feedback as playing something like the role of selection, legitimizing the thought that when the continued presence of a feature (or of an organism possessing the feature) is dependent on the (past) effects that it has caused, then that feature has acquired the function to produce those effects. This would still be an historical account of functionality, depending as it does on past effects, but it extends the application of functionality from features of variably reproduced items to features of variably persisting items (see McLaughlin, 2001: chapter 8).

The Explanandum

There is some debate about what it is that a functional explanation explains. One suggestion is that, given selection only 'chooses' among the traits that are in existence, favoring some at the expense of others, all it can explain is why there are (so many) items with that feature. The proposal is that selection cannot explain how individuals come to have the traits that they do have. The explanation for the traits possessed by any individual would be simply the causal story instantiated in the genetic processes that produced the trait in that individual. Sober (1984) called this latter type of explanation developmental explanation, and the good point he was making was that selection operates on traits already in existence.

One can accept this point, however, and still deny that all that functional explanation can explain is the very general fact concerning the proliferation rates of the causally explained traits. Neander has pointed out that each such trait will have an evolutionary history, the elaboration of which will show how the trait came to have the form it now has (Neander, 1995). Given the falsity of saltationist doctrines, traits such as opposable thumbs or long necks did not emerge fully fledged; there were many mutations, and those that increased, say, opposable thumbness were favored. A restriction to looking at 'selection of preexisting traits' has almost an air of saltationism about it. It is as though the traits came into existence formed as they now are, and were then selected for on that basis. Ignoring the more gradual, at times almost imperceptible, changes that took place in order to arrive at the present trait is to ignore a part of the explanation as to why the trait is what it is. It is because functions can be performed by different mechanisms that attending to the historical detail is important in showing how and why this mechanism evolved into its present form. The central point is that, although selection does only distribute characteristics, the history of successive distributions shows selection as also molding those characteristics. And, as Forber points out (Forber, 2005), sometimes a novel trait only emerges in an environment where selection has produced relevant precursors to that trait, such precursors being essential to the viability of the new trait. Thus, an explanation of the evolution of the human eye, for example, needs to pay attention to how selection operated on protoeyes, forming the necessary background for the emergence of adaptive fully formed eyes.

See also: Adaptation, Fitness, and Evolution; Causation: Physical, Mental, and Social; Causes and Laws: Philosophical Aspects; Evolutionary Theory, Structure of; Explanation: Conceptions in the Social Sciences; Functionalism, History of; Methodological Individualism in Sociology.

Bibliography

Boyd, R., Richerson, P., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago.
Campbell, D.T., 1960. Blind variation and selective retention in creative thought and in other knowledge processes. Psychological Review 67, 380–400.
Cohen, G.A., 1979. Karl Marx's Theory of History: a Defence. Oxford University Press, Oxford.
Cummins, R., 1975. Functional analysis. Journal of Philosophy 72, 741–765.
Cummins, R., 1983. The Nature of Psychological Explanation. MIT Press, Cambridge, MA.
Cziko, G., 1995. Without Miracles: Universal Selection Theory and the Second Darwinian Revolution. MIT Press, Cambridge, MA.
Dennett, D., 1995. Darwin's Dangerous Idea. Simon and Schuster, New York.
Forber, Patrick, 2005. On the explanatory roles of natural selection. Biology and Philosophy 20, 329–342.
Godfrey-Smith, P., 2007. Conditions for evolution by natural selection. Journal of Philosophy 104, 489–516.
Godfrey-Smith, P., 2009. Darwinian Populations and Natural Selection. Oxford University Press, Oxford.
Gould, S., Lewontin, R., 1978. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proceedings of the Royal Society of London, Series B: Biological Sciences 205, 581–598.
Gould, S.J., Vrba, E.S., 1982. Exaptation – a missing term in the science of form. Paleobiology 8 (1), 4–15.
Hodgson, G., Knudsen, T., 2010. Darwin's Conjecture: The Search for General Principles of Social and Economic Evolution. University of Chicago Press, Chicago.
Hull, D., Ruse, M. (Eds.), 1998. The Philosophy of Biology. Part IV. Oxford University Press, Oxford, UK.
McLaughlin, P., 2001. What Functions Explain. Cambridge University Press, Cambridge.
Mesoudi, A., 2011. Cultural Evolution. University of Chicago Press, Chicago.
Millikan, R.G., 1984. Language, Thought, and Other Biological Categories: New Foundations for Realism. MIT Press, Cambridge, MA.
Millikan, R.G., 1993. White Queen Psychology and Other Essays for Alice. MIT Press, Cambridge, MA.
Neander, K., 1995. Pruning the tree of life. British Journal for the Philosophy of Science 46, 59–80.
Pettit, P., 1996. Functional explanation and virtual selection. British Journal for the Philosophy of Science 47, 291–302.
Richerson, P., Boyd, R., 2005. Not Be Genes Alone. University of Chicago Press, Chicago.
Runciman, W.G., 2009. The Theory of Cultural and Social Evolution. Cambridge University Press, Cambridge, UK.
Sperber, Dan, Hirschfeld, Lawrence A., 2004. The cognitive foundations of cultural stability and diversity. Trends in Cognitive Sciences 8 (1), 40–46.
Sober, E., 1984. The Nature of Selection: Evolution Theory in Philosophical Focus. MIT Press, Cambridge, MA.

Functional Magnetic Resonance Imagery, Analysis of

Martin A Lindquist, Johns Hopkins University, Baltimore, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of human brain function using functional magnetic resonance imaging (fMRI) has exploded in the past decade. A standard fMRI study results in enormous amounts of data. In addition, the signal of interest is relatively weak and the data exhibit a complicated temporal and spatial noise structure. In this article, we provide an overview of the analysis of fMRI data, from the initial data acquisition to its use in locating brain activity, making inference about brain connectivity and predictions about psychological or disease states.

Introduction

Functional magnetic resonance imaging (fMRI) is a noninvasive technique for studying brain activity. During the course of an fMRI experiment, a series of brain images are acquired while the subject engages in a set of tasks. Changes in the measured signal between images are used to make inferences regarding possible task-related activations in the brain. There are several common objectives in the analysis of fMRI data, including localizing regions of the brain activated by a task, determining distributed networks that correspond to brain function, and making predictions about psychological or disease states.

Together with other neuroimaging modalities, such as positron emission tomography, electroencephalography (EEG), and magnetoencephalography, fMRI has provided researchers with unprecedented access to the brain in action and numerous insights into human brain function. In recent years, fMRI, in particular, has experienced an explosive growth and found applications in a wide variety of fields, such as psychology, economics, political science, computer science, and statistics. In addition, the use of fMRI data is central to the emerging fields of cognitive neuroscience, affective neuroscience, neuroeconomics, and other related neurobehavioral disciplines. In these areas, fMRI data are being combined with data on performance and psychophysiology to yield new models of human thought, emotion, and behavior, as well as to confirm existing theories.

Any given fMRI experiment gives rise to massive amounts of data. In addition, the signal of interest is relatively weak and the data exhibit a complicated temporal and spatial noise structure. The area of research surrounding fMRI has given rise to a number of interesting and important statistical problems that cover a variety of topics, including the acquisition of raw data, image reconstruction, experimental design, preprocessing, and data analysis (Lindquist, 2008).

Data Structure

During the course of an fMRI experiment, a series of three-dimensional brain images are acquired while the subject performs a set of tasks. Each image consists of roughly 100 000 uniformly spaced volume elements, or voxels, that partition the brain into equally sized boxes. Each voxel corresponds to a spatial location and has a number associated with it that represents its intensity, which is related to the spatial distribution of the nuclear spin density in that area. During the course of an fMRI experiment, between 100 and 2000 images of this type are acquired. The time between each acquisition is referred to as the TR of the experiment, and is typically around 2 s. Intensity values from each individual voxel can be extracted to create a time series of length T, where T corresponds to the number of acquired images. See Figure 1 for an illustration.

A standard fMRI study is performed on multiple subjects (e.g., 10–50) brought in for multiple scanning sessions, each consisting of a number of runs (replications of a certain task). On top of that, multiple high-resolution structural scans are collected for preprocessing and presentation purposes, and often diffusion tensor imaging (DTI) is also collected to inform subsequent network analyses. In sum, the total amount of data that needs to be analyzed is enormous.

Data Acquisition and Reconstruction

To collect fMRI data, the subject is placed into a magnetic resonance imaging scanner, which consists of an electromagnet with a very strong magnetic field (typically between 1.5 and 7.0 Tesla for human studies). This scanner is extremely versatile and can be used to study brain structure, brain function, and even the manner in which water diffusion takes place in the brain.

During data acquisition, the subject is exposed to a radio-frequency (RF) electromagnetic field pulse. This causes the nuclei of interest (e.g., hydrogen atoms) to absorb the energy at a particular frequency band, and become 'excited.' After the pulse is removed, the nuclei emit the energy at the same frequency as they 'relax,' and this quantity can be measured. Pulse sequences that implement particular RF pulses and manipulations of the magnetic field's shape are used to acquire data that can be reconstructed into a map of the underlying signal sources (the hydrogen atoms), thus providing an image of the brain. For a more in-depth discussion see Huettel et al. (2004) or Haacke et al. (1999).

The relaxation process can be described by three different values: T1, T2, and T2*. T1 and T2 are time constants that depend upon tissue type. T1 refers to the rate at which spins relax back to alignment with the main magnetic field, and T2

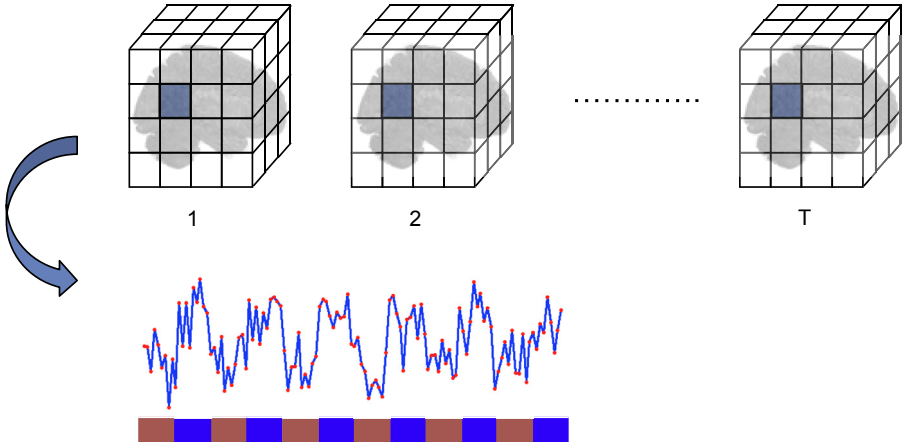

Figure 1 An image consists of roughly 100 000 brain voxels, each corresponding to a spatial location and an intensity value. During the course of an experiment T images are acquired. Extracting the data from a single voxel allows us to track its intensity over time and relates it to some experimental paradigm.

to the rate of attenuation of the magnetic field applied by the RF pulse. T2* is similar to T2; in addition, it depends on local inhomogeneities in the magnetic field caused by changes in blood flow and oxygenation. Different pulse sequences are used to produce images sensitive primarily to T1, T2, or T2*. Because T1 and T2 vary with tissue type, T1- and T2-weighted images can be used to provide detailed representations of the boundaries between gray matter, white matter, and cerebrospinal fluid. Because T2* is sensitive to flow and oxygenation, T2*-weighting is used to create images that can be used to study brain function.

The most popular approach toward fMRI, the blood oxygenation level-dependent (BOLD) contrast (Kwong et al., 1992; Ogawa et al., 1992), takes advantage of the difference in T2* between oxygenated and deoxygenated hemoglobin. As neural activity increases, so do metabolic demands for oxygen and nutrients. Neural firing leads to the extraction of oxygen from hemoglobin in the blood. As oxygen is extracted, the hemoglobin becomes paramagnetic, creating small distortions in the magnetic field that cause a T2* decrease (i.e., a faster decay of the signal). Thus, increases in deoxyhemoglobin lead to a decrease in BOLD signal that can be measured by the scanner.

Properties of the BOLD Response

The temporal resolution of fMRI is generally not limited by how quickly the data can be acquired, but instead by properties of the underlying evoked hemodynamic response to a certain neural event. The BOLD response is a complex, nonlinear function that depends on neuronal and vascular changes (Wager et al., 2005). The shape of the response depends both on the applied stimulus and the hemodynamic response to neuronal events; the latter is typically referred to as the hemodynamic response function (HRF), shown in Figure 2(a). The increased metabolic demands due to neuronal activity lead to increases in the inflow of oxygenated blood to active regions of the brain. Since more oxygen is supplied than consumed,

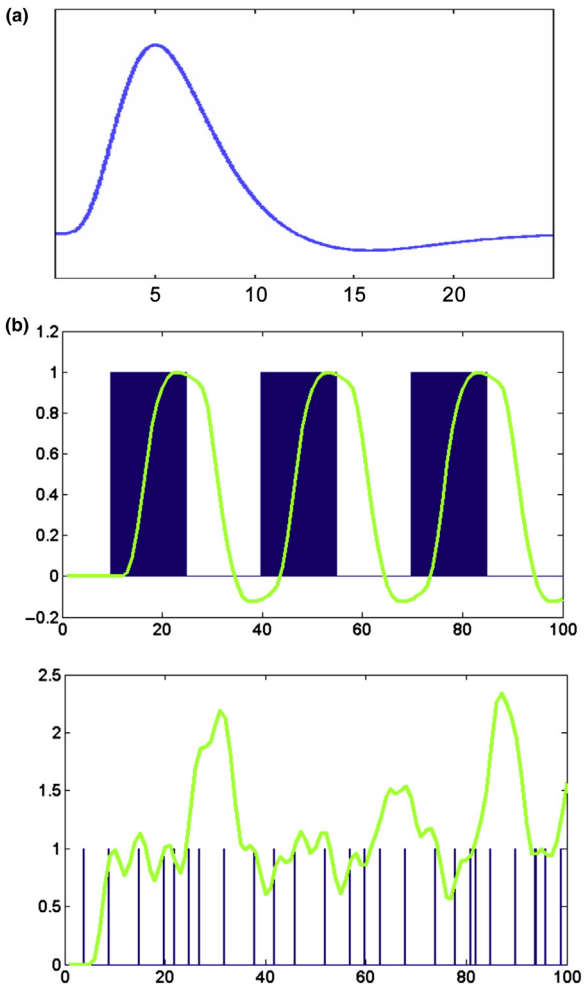

Figure 2 (a) The standard canonical hemodynamic response function (HRF) used in functional magnetic resonance imaging (fMRI) data analysis. (b) The blood oxygenation level dependent (BOLD) response is often modeled as the convolution of the stimulus function and the HRF. Here are sample responses for a blocked design (top) and an event-related design (bottom).

this leads to a decrease in the concentration of deoxyhemoglobin and an increase in signal. This rise in signal peaks 5–8 s after the onset of neural activity. Thereafter, the BOLD signal decreases to below a baseline level, which is sustained for roughly 10 s. This effect, the poststimulus undershoot, is due to the fact that blood flow decreases more rapidly than blood volume, allowing for a greater concentration of deoxyhemoglobin in previously active brain regions.

It is common to assume a linear relationship between neuronal activity and the BOLD response, where linearity implies that the magnitude and shape of the evoked HRF do not depend on the preceding stimuli. Studies have shown this to be a reasonable assumption (Boynton et al., 1996), as long as events are spaced at least 5 s apart (Miezin et al., 2000). The ability to assume linearity is important, as it allows the relationship between stimuli and BOLD response to be modeled using a linear time invariant system, where the stimulus acts as the input and the HRF acts as the impulse response function. This relationship is shown in **Figure 2(b)**.

The timing and shape of the HRF are known to vary across a single subject's brain and across individuals (Aguirre et al., 1998). Part of this variability is vascular, which may cause differences in the HRF across brain regions in the same task for purely physiological reasons (Vazquez et al., 2006). Another source of variability is differences in the pattern of evoked neural activity in regions performing different functions related to the same task. It is important that these regional variations are accounted for when modeling the BOLD signal.

The BOLD response is corrupted by random noise and various nuisance components that arise due to both hardware reasons and the subjects. For instance, fluctuations in the magnetic resonance signal intensity caused by thermal motion of electrons within the subject and the scanner give rise to noise that tends to be highly random and independent of the experimental task. This type of noise does not exhibit spatial structure and its effects can be minimized by averaging data. Another source of variability is due to scanner drift, caused by scanner instabilities, which result in slow changes in voxel intensity over time. The amount of drift varies across space, and it is important to include this source of variation in subsequent models. Finally, physiological noise due to patient motion, respiration, and heartbeat all cause fluctuations in signal across both space and time, which should be accounted for if possible.

Experimental Designs

Designing an fMRI study involves a trade-off between experimental power and the ability to make strong inferences based on the results. The two main types of experimental designs used in fMRI are blocked designs and event-related designs. In a blocked design, different conditions in the experiment are presented as separate blocks of trials, each for an extended period of time. In event-related designs, discrete events of short duration are presented in a randomized order. Examples of these two design types are shown in **Figure 2(b)**.

Blocked designs typically yield high experimental power, but do not provide precise information about the particular psychological processes that activated a brain region. Event-related designs, on the other hand, allow brain activation to be linked more precisely to the particular cognitive processes engaged in certain types of trials, but suffer from decreased power. These types of designs rely more strongly on assumptions about the time course of both evoked neural activity and the HRF, and it is common to assume a near-instantaneous neural response for brief events and a canonical HRF shape to generate linear models for statistical analyses. Blocked designs tend to be less sensitive to the variability of the HRF because they depend on the total activation caused by a train of stimulus events, which makes the overall predicted response less sensitive to variations in the shape of responses to a single event.

Preprocessing

Prior to analysis, fMRI data undergo a series of preprocessing steps. The goals of preprocessing include minimizing the influence of data acquisition and physiological artifacts, checking statistical assumptions and transforming the data to meet assumptions, and standardizing the locations of brain regions across subjects to achieve validity and sensitivity in group analysis. Below we outline the most common preprocessing steps.

Statistical analysis typically assumes that every voxel in an image is acquired simultaneously. In reality, fMRI data are collected in a slice-wise manner, so data from different slices are shifted in time relative to one another. Therefore, it is common to estimate the signal intensity in all voxels at the same moment in the acquisition period, in a process referred to as slice-timing correction. This can be done by interpolating the signal at the chosen time point using information from the same voxel measured in adjacent acquisitions. Some researchers do not use slice timing, as it adds potential interpolation error to the data, but instead use more flexible hemodynamic models to account for variations in acquisition time.

Subject motion during data acquisition causes a voxel's signal intensity to be 'contaminated' by signal from its neighbors. Therefore, it is critical to realign each individual image to compensate for subject movement. This is typically performed by choosing a reference image (e.g., the first image) and using a rigid body transformation, which allows images to be translated and rotated, to match each image to the reference. An iterative algorithm is used to search for the transformation parameters that provide the best match. While realignment is able to correct small movements of the head, it does not correct for the more complex 'spin-history artifacts' created by motion (Friston et al., 1996). Therefore, parameter estimates at each time point are often included in the later statistical analysis as covariates of no interest.

Coregistration is the process of aligning structural and functional images. This procedure is performed both for presentation purposes and to aid in later preprocessing steps. Here a single structural image is mapped onto the first, or mean, functional image. This is typically performed using an affine transformation matrix whose parameters can be estimated by maximizing the mutual information between the two images.

For group analysis, it is necessary for each voxel to lie within the same brain structure for each subject. However, it is well known that individual brains have different shapes and features. Normalization attempts to register each subject's anatomy with a standardized atlas space defined by a template brain, such as the Talairach or Montreal Neurological Institute brain (Evans et al., 1993; Talairach and Tournoux, 1988). Normalization can be linear, involving simple registration of the gross shape of the brain, or nonlinear, involving warping to match local features, and is typically performed by matching a subject's structural scan to the template brain. If coregistration has been performed, the same transformation can thereafter be applied to each functional image.

Finally, many investigators apply spatial smoothing to the functional data, thereby blurring their intensities in space. One reason is to improve intersubject registration and remove possible artifacts related to that procedure. A second reason is to validate certain modeling assumption regarding image smoothness and normality. Finally, if activation is occurring over several voxels, smoothing can help increase the signal-to-noise ratio. Smoothing is typically performed by convolving each image with a Gaussian kernel described by the full width of the kernel at half its maximum height, measured in millimeters.

Data Analysis

The statistical analysis of fMRI data has a variety of goals. These include localizing brain areas activated by a task, determining networks corresponding to brain function, and making predictions about psychological or disease states; see Figure 3 for illustrations. We briefly outline each of these goals below; for more information see Lindquist (2008).

Localizing Brain Activation

To date, the most common use of fMRI data has been to detect areas of the brain that activate in response to a specific task. Brain mapping studies are necessary for the development of biological markers and increasing our understanding of brain function. Typically, data analysis is performed using the so-called massive univariate approach, where a separate model is constructed and fit to data from each voxel.

The general linear model (GLM) approach of Worsley and Friston (1995) has arguably become the dominant way to analyze fMRI data. It models the fMRI time series in each voxel as a linear combination of several different signal components and tests whether it is systematically related to any of these known input functions. Using the GLM, the data at each voxel can be expressed as $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where $\varepsilon \sim N(0,\mathbf{V})$. Here $\mathbf{Y}$ represents the data (a vector of length T), $\mathbf{X}$ is a design matrix containing information about various signal components, and $\mathbf{V}$ is a covariance matrix that incorporates information about autocorrelations in the data, typically modeled using either an autoregressive or an autoregressive moving-average model.

Proper construction of the design matrix is critical when using the GLM. This is complicated by a number of factors, including the fact that the BOLD response contains low-frequency drift and artifacts related to head movement and

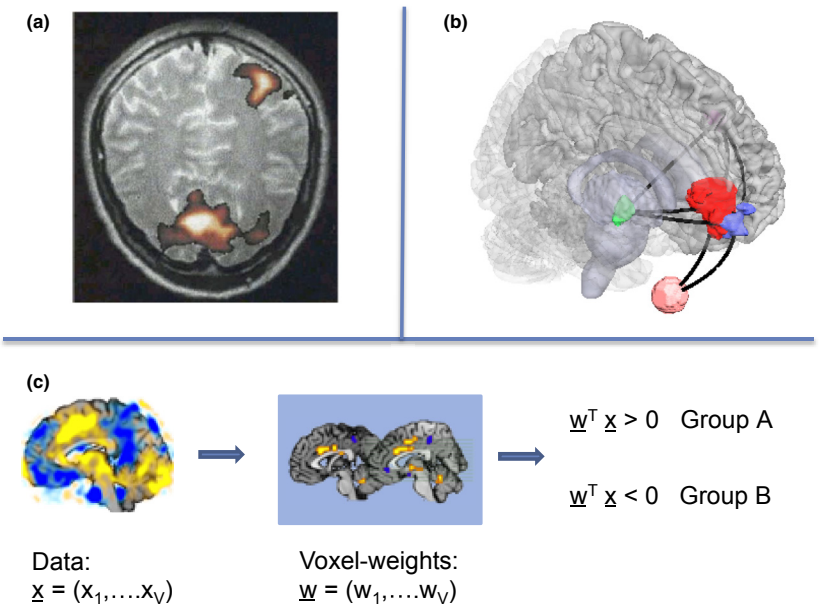

Figure 3 (a) An example of a thresholded statistical map shows activation in the visual and motor cortices (Lindquist, M., Zhang, C., Glover, G., Shepp, L., 2008b. Rapid three-dimensional functional magnetic resonance imaging of the negative BOLD response. Journal of Magnetic Resonance 191 (1), 100–111). (b) An undirected graph, obtained using an inverse covariance matrix estimation technique, shows the relationship between several brain regions and heart rate during a stress task (Cribben, I., Haraldsdottir, R., Atlas, L.Y., Wager, T.D., Lindquist, M.A., 2012. Dynamic connectivity regression: determining state-related changes in brain connectivity. NeuroImage 61 (4), 907–920). (c) An illustration of a pain biomarker (Wager, T.D., Atlas, L.Y., Lindquist, M.A., Roy, M., Woo, C.-W., Kross, E., 2013. An fMRI-based neurologic signature of physical pain. New England Journal of Medicine 368 (15), 1388–1397), where the cross product of the data **x** and voxel weights **w** provides scores that allows us to classify subjects.

cardiopulmonary-induced brain movement, the neural response shape may not be known, and the hemodynamic response varies in shape across the brain. For these reasons, the design matrix **X** usually consists of both nuisance parameters (corresponding to drift components and the estimated motion parameters) and signal of interest. The simplest version of the GLM assumes that both the stimulus function and the HRF are known. The stimulus is assumed to be equivalent to the experimental paradigm, while the HRF is modeled using a canonical HRF, typically either a gamma function or the difference between two gamma functions (see **Figure 2(a)**). Under these assumptions, the convolution of the stimulus and HRF is used to summarize the BOLD response and entered as a column of **X**; see **Figure 2(b)** for examples corresponding to a blocked design and an event-related design. If one is unwilling to assume a fixed canonical HRF, it is possible to model the shape by expressing the HRF as the linear combination of basis functions. Here each basis function is convolved with the stimulus and entered as separate columns in the design matrix (Lindquist and Wager, 2007; Lindquist et al., 2008a).

After fitting the model, one can test for an effect $\mathbf{c}^T\beta$ where **c** is a contrast vector. The contrast vector can, for example, be used to estimate signal amplitudes in response to a single condition, an average over multiple conditions or the difference in amplitude between two conditions. Significance is determined by testing individual model parameters using a t-test, and subsets of parameters using an F-test. Since the covariance matrix has to be estimated, a Satterthwaite approximation is used to calculate the effective degrees of freedom for the test statistics. This procedure is repeated across all brain voxels and the results are summarized in a statistical map, which is an image of each voxel's test statistic.

The next step is to threshold these maps, with the implication that significant voxels are activated by the experimental task (see **Figure 3(a)**). In a typical experiment, up to 100 000 hypothesis tests (one for each voxel) are performed simultaneously, and it is crucial to correct for multiple comparisons when determining the threshold. Several approaches toward controlling the false-positive rate have been used, with the fundamental difference between methods lying in whether they control the family-wise error rate (FWER) or the false discovery rate (FDR).

Random field theory (Worsley et al., 2004) is the most popular approach for controlling the FWER in the fMRI community. Here, the image of voxel-wise test statistic values are assumed to be a discrete sampling of a continuous smooth random field. One begins by estimating the smoothness of the image, expressed in terms of resolution elements, or resels. Next, using information about the number of resels and the shape of the search volume, mathematical theory exists for calculating the expected Euler characteristic. For large thresholds, this value is equal to the number of clusters of activity that one would expect by chance at a certain statistical threshold. Hence, it can be used to determine the appropriate threshold that controls the FWER at a certain level.

The FDR controls the proportion of false positives among all rejected tests (Genovese et al., 2002). The FDR controlling procedure is adaptive in the sense that the larger the signal, the lower the threshold. If all of the null hypotheses are true, the FDR will be equivalent to the FWER. In addition, any procedure that controls the FWER will also control the FDR. Hence, any procedure that controls the FDR can only be less stringent and lead to increased power.

So far, the discussion has centered on single-subject analysis. However, researchers typically want to make conclusions on population effects, and statistical analysis needs to be extended to incorporate information from groups of subjects. Multisubject fMRI data are intrinsically hierarchical in nature, with lower level observations (i.e., individual subjects) nested within higher levels (i.e., groups of subjects). Standard mixed effects models can be used to analyze multisubject fMRI data. However, it is most common to fit separate GLMs for each subject and perform a t-test on a particular contrast of interest at the second level (Mumford and Nichols, 2009).

Connectivity

Recently, there has been enormous interest in performing connectivity studies that describe how various brain regions interact and how these interactions depend on experimental conditions. It is common in the fMRI literature to distinguish between functional and effective connectivity (Friston, 1994). Functional connectivity is defined as the undirected association between two or more fMRI time series, while effective connectivity is the directed influence of one brain region on others.

The simplest approach toward functional connectivity analyses compares correlations between regions of interest, or between a 'seed region' and other voxels throughout the brain. Recently, inverse covariance estimation methods have been applied since, for multivariate normal data, conditional independence between variables (regions) corresponds to zero entries in the inverse covariance matrix (see **Figure 3(b)**). Other approaches include using multivariate decomposition methods, such as principal components analysis (Andersen et al., 1999) and independent components analysis (Calhoun et al., 2001; McKeown and Makeig, 1998), to identify task-related patterns of brain activation without making a priori assumptions about its form. These methods involve decomposing the time-by-voxel data matrix, **Y**, into a set of spatial and temporal components according to some criteria (e.g., independence between components). These types of approaches have been especially fruitful for analyzing so-called resting-state data (Biswal et al., 1995), where the subject lies in a scanner without performing a specific task. This is an area of intense research where standard methods for localizing brain activation are not applicable due to the lack of stimulus.

In effective connectivity, a small set of regions with a proposed set of directed connections are specified a priori, and tests are used to assess the statistical significance of individual connections. Most effective connectivity methods rely on two models: a neuroanatomical model that describes the areas of interest, and a mathematical model that describes how they are connected. Commonly used methods include structural equation modeling (McIntosh and Gonzalez-Lima, 1994) and dynamic causal modeling (Friston et al., 2003). Granger causality (Roebroeck et al., 2005) is another approach used to test effective connectivity. This approach does not necessarily rely on a structural model, but instead quantifies the usefulness of past values extracted from one brain region in predicting current values in another.

Prediction

There is a growing interest in using fMRI data for classification of mental disorder and predicting the early onset of disease. In addition, there is interest in developing methods for predicting stimuli directly from functional data. This opens up the possibility of inferring information about subjective human experience directly from brain activation patterns. (See Pereira et al. (2009) for an overview.)

Predicting brain states is challenging and requires the application of novel statistical and machine learning techniques. Various techniques have successfully been applied to fMRI data in which a classifier is trained to discriminate between different brain states and then used to predict the states in a new set of fMRI data. The application of machine learning methods to fMRI data is often referred to as multivoxel pattern analysis (MVPA). Instead of focusing on single voxels, MVPA uses pattern-classification algorithms applied to multiple voxels to decode patterns of activity. In MVPA the goal is to determine the model parameters that allow for the most accurate prediction of new observations (see **Figure 3(c)**).

In fMRI it is particularly important to make analysis choices that balance interpretability with predictive power. Certain methods may give good predictions but the resulting voxel weights may be difficult to interpret and may not generalize well to new subjects.

What the Future Holds

In neuroscience there is a general trend toward using multiple imaging methods in tandem to overcome limitations of each approach in isolation. For example, advances in engineering and signal processing allow EEG and fMRI data to be collected simultaneously (Goldman et al., 2000). EEG has an extremely high temporal resolution (on the order of milliseconds) but poor spatial resolution, while fMRI suffers from the opposite problem. By merging these two techniques, the hope is that one can get the best of both worlds. In another example, DTI, a technique for measuring directional diffusion and reconstructing the fiber tracts of the brain (Le Bihan et al., 2001), can be combined with fMRI to determine appropriate regions of the brain to include in subsequent connectivity models. In yet another example, neuroimaging data are being combined with transcranial magnetic stimulation (TMS) to integrate the ability of neuroimaging to observe brain activity with the ability of TMS to manipulate brain function (Bohning et al., 1997). Using this technique, one can stimulate or suppress activation in certain regions while the subject performs certain tasks. One can then attempt to infer causal relationships by studying differences in a brain network when a region is manipulated and when it is not. Finally, integrating genetics with brain imaging is seen as a way to study how a particular subset of polymorphisms may affect functional brain activity. In addition, quantitative indicators of brain function could facilitate the identification of the genetic determinants of complex brain-related disorders such as autism, dementia, and schizophrenia (Glahn et al., 2007).

Each of these multimodal approaches promise to be important topics of future research, and to fully realize their promise, novel statistical techniques will be needed. Ultimately, combining information from different modalities will be challenging to data analysts, if for no other reason than that the amount of data will significantly increase. In addition, since different modalities are measuring fundamentally different quantities, it is not immediately clear how to best combine the information. However, clearly, this is an extremely important problem that has already started to become a major area of research.

See also: Analysis of Variance and Generalized Linear Models; Functional Data Analysis; Graphical Markov Models: Overview; Hierarchical Models: Random and Fixed Effects; Hypothesis Testing in Statistics; Hypothesis Tests, Multiplicity of; Linear Mixed Models; Longitudinal Data: Event–History Analysis in Discrete Time; Meta-Analysis: Theory; Multilevel Models in Demography; Multiple Comparisons, Statistics of; Network Analysis; Spatial Data; Statistical Analysis: Multilevel Methods; Time Series: General.

Bibliography

Aguirre, G.K., Zarahn, E., D'Esposito, M., 1998. The variability of human, BOLD hemodynamic responses. NeuroImage 8 (4), 360–369.

Andersen, A., Gash, D., Avison, M.J., 1999. Principal component analysis of the dynamic response measured by fMRI: a generalized linear systems framework. Magnetic Resonance in Medicine 17, 785–815.

Biswal, B., Zerrin Yetkin, F., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magnetic Resonance in Medicine 34 (4), 537–541.

Bohning, D.E., Pecheny, A.P., Epstein, C.M., Speer, A.M., Vincent, D.J., Dannels, W., George, M., 1997. Mapping transcranial magnetic stimulation (TMS) fields in vivo with MRI. Neuroreport 8, 2535–2538.

Boynton, G.M., Engel, S.A., Glover, G.H., Heeger, D.J., 1996. Linear systems analysis of functional magnetic resonance imaging in human V1. Journal of Neuroscience 16, 4207–4221.

Calhoun, V.D., Adali, T., Pearlson, G., Pekar, J., 2001. Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. Human Brain Mapping 13, 43–53.

Cribben, I., Haraldsdottir, R., Atlas, L.Y., Wager, T.D., Lindquist, M.A., 2012. Dynamic connectivity regression: determining state-related changes in brain connectivity. NeuroImage 61 (4), 907–920.

Evans, A.C., Collins, D.L., Mills, S., Brown, E., Kelly, R., Peters, T.M., 1993. 3D statistical neuroanatomical models from 305 MRI volumes. In: Nuclear Science Symposium and Medical Imaging Conference, 1993., 1993 IEEE Conference Record. IEEE, pp. 1813–1817.

Friston, K., 1994. Functional and effective connectivity in neuroimaging: a synthesis. Human Brain Mapping 2, 56–78.

Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. NeuroImage 19, 1273–1302.

Friston, K.J., Williams, S., Howard, R., Frackowiak, R.S., Turner, R., 1996. Movement-related effects in fMRI time-series. Magnetic Resonance in Medicine 35 (3), 346–355.

Genovese, C., Lazar, N., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15, 870–878.

Glahn, D.C., Thompson, P.M., Blangero, J., 2007. Neuroimaging endophenotypes: strategies for finding genes influencing brain structure and function. Human Brain Mapping 28, 488–501.

Goldman, R.I., Stern, J.M., Engel Jr., J., Cohen, M.S., 2000. Acquiring simultaneous EEG and functional MRI. Clinical Neurophysiology 11, 1974–1980.

Haacke, M.E., Brown, R.W., Thompson, M.R., Venkatesan, R., 1999. Magnetic Resonance Imaging: Physical Principles and Sequence Design. Wiley-Liss, New York.

Huettel, S.A., Song, A.W., Mccarthy, G., 2004. Functional Magnetic Resonance Imaging. Sinauer Associates, Sunderland.

Kwong, K.K., Belliveau, J.W., Chesler, D.A., Goldberg, I.E., Weisskoff, R.M., Poncelet, B.P., Kennedy, D.N., Hoppel, B.E., Cohen, M.S., Turner, R., 1992. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. Proceedings of the National Academy of Sciences 89 (12), 5675–5679.

Le Bihan, D., Mangin, J.-F., Poupon, C., Clark, C.A., Pappata, S., Molko, N., Chabriat, H., 2001. Diffusion tensor imaging: concepts and applications. Journal of Magnetic Resonance Imaging 13 (4), 534–546.

Lindquist, M.A., Loh, J., Atlas, L., Wager, T., 2008a. Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. NeuroImage, 45, S187–S198.

Lindquist, M.A., Zhang, C., Glover, G., Shepp, L., 2008b. Rapid three-dimensional functional magnetic resonance imaging of the negative BOLD response. Journal of Magnetic Resonance 191 (1), 100–111.

Lindquist, M.A., 2008. The statistical analysis of fMRI data. Statistical Science 23 (4), 439–464.

Lindquist, M.A., Wager, T.D., 2007. Validity and power in hemodynamic response modeling: a comparison study and a new approach. Human Brain Mapping 28, 764–784.

McIntosh, A., Gonzalez-Lima, F., 1994. Structural equation modeling and its application to network analysis in functional brain imaging. Human Brain Mapping 2, 2–22.

McKeown, M.J., Makeig, S., 1998. Analysis of fMRI data by blind separation into independant spatial components. Human Brain Mapping 6, 160–188.

Miezin, F., Maccotta, L., Ollinger, J., Petersen, S., Buckner, R., 2000. Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. NeuroImage 11.

Mumford, J.A., Nichols, T., 2009. Simple group fMRI modeling and inference. Neuroimage 47 (4), 1469–1475.

Ogawa, S., Tank, D., Menon, R., Ellerman, J., Kim, S., Merkle, H., Ugurbil, K., 1992. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping and magnetic resonance imaging. Proceedings of the National Academy of Sciences 89, 5951–5955.

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45 (1), S199–S209.

Roebroeck, A., Formisano, E., Goebel, R., 2005. Mapping directed influence over the brain using granger causality and fMRI. NeuroImage 25, 230–242.

Talairach, J., Tournoux, P., 1988. Co-planar Stereotaxic Atlas of the Human Brain. 3-dimensional Proportional System: An Approach to Cerebral Imaging, Thieme Medical Publishers, New York.

Vazquez, A.L., Cohen, E.R., Gulani, V., Hernandez-Garcia, L., Zheng, Y., Lee, G.R., Kim, S.G., Grotberg, J.B., Noll, D.C., 2006. Vascular dynamics and bold fMRI: CBF level effects and analysis considerations. NeuroImage 32, 1642–1655.

Wager, T.D., Atlas, L.Y., Lindquist, M.A., Roy, M., Woo, C.-W., Kross, E., 2013. An fMRI-based neurologic signature of physical pain. New England Journal of Medicine 368 (15), 1388–1397.

Wager, T.D., Vazquez, A., Hernandez, L., Noll, D.C., 2005. Accounting for nonlinear BOLD effects in fMRI: parameter estimates and a model for prediction in rapid event-related studies. NeuroImage 25 (1), 206–218.

Worsley, K.J., Friston, K.J., 1995. Analysis of fMRI time-series revisited-again. NeuroImage 2, 173–181.

Worsley, K.J., Taylor, J.E., Tomaiuolo, F., Lerch, J., 2004. Unified univariate and multivariate random field theory. NeuroImage 23, 189–195.

Functionalism in Anthropology

Joan Vincent, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 9, pp. 5844–5847, © 2001, Elsevier Ltd.

Abstract

Functionalism is a sociological research program of historical and systematic significance. Sociological explanations in terms of functions are teleological. Social phenomena are to be explained by their effects and consequences – chiefly of a beneficial, morally desired or nonintended character – and/or by a circuit of equilibration (a homeostatic loop) which is maintaining a certain state of proper functioning. The notion of (mono-) functionally specialized subsystems of society can be seen as the most important precipitate of functionalist thinking in sociology.

In its simplest common-sense form functionalism in anthropology describes approaches which purport to explain cultural or social phenomena in terms of the purpose they serve.

Since the 1960s, specialized research into the sociological and historical framing of anthropology has revealed the extent to which discussions of functionalism contribute to analysis of knowledge as culture. As yet, however, it falls short of meeting the criticism that it tends to treat the discipline of anthropology as a closed discourse, despite calls that it be viewed as part of a larger discursive field in which anthropology operates (Asad, in Stocking, 1991; Trouillot, in Fox, 1991).

The Intellectual History of an Idea

The term 'function' was first introduced as a mathematical concept by Gottfried Wilhelm Leibniz in 1714, to refer to a variable considered in relation to one or more other variables in terms of which it might be expressed.

Two main bundles of ideas underlay the concept of 'function' that dominated academic anthropology in the first half of the twentieth century: (a) a relation of interdependence and (b) an orientation toward given ends. The first involved reciprocity and reversibility in terms of cause and effect. The ends might be the needs of the individual, as in the utilitarian functionalism of Malinowski (1884–1942) or the needs of society, as in the structural–functionalism of Radcliffe-Brown (1881–1955). Both adopted notions of culture or society in the modern pluralized usage as totalities with a certain unity and integrative character, a notion derived from the German philosopher Johann Gottfried Herder.

In the second half of the century, the premise remained that cultures and societies were structured wholes, now more consistently referred to as cultural or social systems. At this point functionalism's orthodoxy came under challenge in the aftermath to the Second World War which saw the withdrawal of sovereignty claims over a plethora of colonial territories within the British Empire and the establishment of United States 'trusteeship' in Oceania. This spurred British functionalists to address complex 'new nations' and provided American anthropologists *de novo* with a laboratory of functioning colonial cultures (Bashkow, in Stocking, 1991).

By the 1970s, functionalism (as an achieved, shared paradigm which defined problems and procedures) was seen to be outmoded. It was replaced, amidst repeated calls for anthropology to reinvent itself, by struggle among new modes of discourse, chief among them Levi–Straussian structuralism, Marxism, and symbolic and interpretive anthropology. Nevertheless, echoes of functionalism lingered on into the decentered, fragmented, pluralizing, flexible, refractive, postmodern world of the new millennium.

Functionalism's Multivocality

A distinction among (a) functionalism as doctrine (or theory); (b) functionalism as method; and (c) functionalism as interest or ideology was a critical feature in Bronislaw Malinowski's modernist legacy to functionalism (Vincent, 1991). This, along with the essential dualism between natural science and the humanities (particularly science and history) that entered modernist anthropology with Franz Boas's *The Study of Geography* (1897, reprinted in Stocking, 1996), goes some way towards ordering the cultural complexity of functionalist knowledge. The scientific tendencies of the physicist who tries theoretically to reduce reality to simple elements and the humanist tendencies of the cosmographer who seeks to comprehend a complex phenomenon as a whole are both contained within the functionalist moment in anthropology.

Functionalism as Doctrine or Theory

Three cross-cutting anthropological trajectories may be recognized within functionalist anthropology.

Utilitarian Functionalism (Malinowskian)

Recent research has exposed the extent to which Malinowski's utilitarian functionalism, which dominated British anthropology between the wars, derived from his engagement with the neopositivism of Ernst Mach in his doctoral dissertation – particularly Mach's view that knowledge is a response to biological human needs (Thornton, 1985; Thornton and Skalnik, 1993). In his last, posthumously published, position paper, *The Functional Theory*, Malinowski takes his stand on two axioms: first, that every culture must satisfy the

biological system of needs, and second, that every cultural achievement that implies the use of artifacts or symbols is an instrumental enhancement of human anatomy, and refers directly or indirectly to the satisfaction of a bodily need (Malinowski, 1944, p. 171).

Strictly followed, these axioms would restrict cultural anthropology to the economism of the rationalizing individual (human nature), or the ecologism of selective advantage (external nature) (Sahlins, 1976), but Malinowski's practice deviated considerably from his theorizing. As with Boas, his sentient functionalism lay immanent in his ethnography; neither man was deemed 'a theorist.'

Hybrid Cultural Functionalisms (Boasian/Parsonian)

This complex functionalist trajectory, triggered by Franz Boas's (1858–1942) 'vaguely functionalist ideas' (Sahlins, 1976, p. 71), drew explicitly on German humanist intellectual traditions. It comprised several contradictions which manifested themselves transgenerationally. Initially this took the form of Alfred Kroeber's (1876–1960) functionalist historicism and Robert Lowie's (1883–1957) neopositivist empiricism. A second generation was further divided; it was also developing in the context of the 'British' functionalist paradigm.

Thus, in 1934, Alexander Lesser (1902–82), recognizing functionalism as the most vigorous tendency extant in anthropology, set out to distinguish the particular interests and conceptions of individual American functionalists from the broader, foundational Boasian functionalist approach. He adopted a Leibnizian mathematical definition of a functional relation, established by observation, while arguing that contemporaneous events associated by the observer might be merely coexistences. Functional historicity recognized that 'institutions, customs, beliefs, artifacts have careers in time, and that their form and character is molded more by what has happened to them in the course of that history than by what particular things they occur associated with at any one time' (Lesser in Mintz, 1985, p. 60). Radcliffe-Brown's position paper 'On the concept of function in social science' (1935) was prompted by his taking Lesser's critique personally.

It was actually directed at Lesser's fellow Boasians at Columbia University in New York, Ruth Benedict (1887–1948) and Margaret Mead (1901–78). Romantic holistic sensibility figured most prominently in Benedict's cultural configurations or patterns of culture. Mead, on the other hand (after corresponding with Radcliffe-Brown), designed a framework for the comparative functional analysis of social systems. This was published in *Cooperation and Competition among Primitive Peoples* (1937).

At the same time a more scientific anthroposociological functionalism was being developed at Harvard where sociologist Talcott Parsons (1902–79) joined forces with anthropology to set up a Department of Social Relations. One of its earliest products was arguably the most thoroughgoing functionalist ethnography ever published, Conrad Arensberg and Solon Kimball's *Family and Community in Ireland* (1940). Under the direction of W. Lloyd Warner (formerly Radcliffe-Brown's student) this extended functionalist anthropology from simple/primitive to complex/modern cultures. Later critics faulted the work for presenting as eternally cultural what was no more than a well-observed historical moment in county Clare.

Structural-Functionalism (Radcliffe-Brown)

Structural-functionalism's core concepts are, in harness, structure and system. Under the auspices of Radcliffe-Brown (1881–1955) it emerged as a clear-cut alternative to hybrid Boasian functionalism and successor to Malinowski's particular brand of economism and radical individualism. It derived not from German and British precursors but mainly from Durkheimian French sociology with supplementary insights from the Russian geographer and anarchist, Peter Kropotkin (1842–1921).

Structural-functionalism emphasized the formal ordering of parts and their functional interrelations as contributing to the maintenance needs of a structured social system. The function of any institution (or 'recurrent social activity') was the part it played in the maintenance of the larger structural whole. This assumption attributed to social systems an internal integration of parts similar to that found in organisms.

Whereas Malinowskian functionalism was seen to have outworn its usefulness by 1945, and American cultural functionalism remained factionally divided, Radcliffe-Brown's successors, particularly Meyer Fortes (1906–83), Edward Evans-Pritchard (1902–73), and Max Gluckman (1911–75), expanded structural-functionalism to address situation, selection, and history while continuing to work implicitly with equilibrium models of processes internal to the sociocultural system. Marxisant anthropological analyses of contradictions in structure and function supplemented structural-functionalism with a compatible, historically oriented, approach to competition, conflict, and change.

Functionalism as Method

Malinowski claimed that the so-called functional method in modern anthropology constituted a specific and necessary method of knowledge production. The relationship between theory and practice was reciprocal. Practice consisted of the intensive fieldwork that had emerged in contradistinction to the scientific survey methods of the Cambridge University Torres Straits Expedition of 1898–9, and the writing up of research findings in monographic form. The 1912 edition of *Notes and Queries*, that Malinowski carried with him into the field, contained W. H. R. Rivers, (1864–1922) delineation of 'intensive' fieldwork in which 'the worker lives for a year or more among a community of perhaps four or five hundred people and studies every detail of their life and culture; in which he comes to know every member of the community personally; in which he is not content with generalized information, but studies every feature of life and custom in concrete detail and by means of the vernacular language' (Rivers 1913, p. 7, quoted in Urry, 1993, p. 28). This became functionalism's ideal mode of production.

Functionalism's methods of reproduction (knowledge as culture) included grantsmanship, correspondence, seminars, public events, unpublished and published lectures, oral histories, biographies, and the writing of monographs. Commemorations and conferences held in Cracow, London, and New York to memorialize Malinowski in 1984

(Gellner, 1987) and in America to memorialize Boas in 1992 exposed a more complex functionalist heritage than that commonly recognized.

Paradigmatic adjustment was essential to the production and reproduction of functionalism. Thus, the last volume of Malinowski's Trobriand ethnography, *Coral Gardens and their Magic* (1935), documents the extent to which he was beginning to question 'the wholeness of culture,' criticize his neglect of the colonial order and of the inadequacy of the concept of culture contact or acculturation. He noted the possibility of placing his own functional analyses of land tenure in the same frame as political controversy over land rights in Ireland and India (Vincent, 1991). But such paradigmatic adjustment was rare. More routinely, it was the thick, detailed ethnographies functionalism engendered that empowered the reader to re-evaluate functionalism long before the textualization of theory became emblematic of paradigmatic change. The perdurance of a romantic ethnographic tradition contributed significantly to the persistence of inherent functionalist notions in anthropology even when the theoretical adequacy of functionalism remained an issue (Gellner, 1987).

Functionalism as Interest: The 'Functionalist Movement'

Functionalism was both a concept and a scientific community contesting the constitution and distribution of cultural capital. Throughout the first three decades of the twentieth century in both Britain and America evolutionism, diffusionism and functionalism represented warring trajectories and interests and the paradigmatic dominance of functionalism was by no means secure (Vincent, 1986). Explanations of dominance, as George Stocking has consistently demonstrated, are to be found not in the valency of ideas but in academic power, control, and institutionalization – to which observation must be added strategies of reproduction. In anthropology departments where field materials were presented in seminars (as at the London School of Economics and Manchester) 'fieldnotes were treated as facts, as concrete things, as sources of authority, as weapons in the struggle to gain the high ground of functionalist debate' (Bond in Sanjek, 1990, p. 279).

Methodological Issues or Problems in Functionalism

As with other -isms, functionalism has been described most clearly and simply by those with most to gain from detracting interest from it. Yet anthropological functionalism has been characterized by constant immanent critique as well as external challenge. For critics it has been something of a moveable feast. What is noted depends on whose functionalism they are addressing and at what phase in that functionalist's thinking. Broad, generic charges laid against anthropological functionalism relate to positivism and pragmatism, on the one hand, and primitivism, on the other. Specific critiques have been directed against its holism along with its assumption of interdependence of parts, synchronism, homogeneity and the notion of collective consciousness, uniformity and its transgenerational replication through custom. Re-evaluations of structural-functional classics demonstrated that even so-called simple societies/cultures were internally differentiated and heterogeneity gave rise to conflicting concerns among different sectors of the population. As the paradigm expanded further, new charges were made that functionalism could not accommodate the analysis of power.

Conclusion

The radical antipositivism of the 1960s put functionalism on trial. Functional anthropologists who declined to commit anthropology to the scrap heap argued ethnography's value as social history. Others argued its continued relevance for specific subfields such as ecological and applied anthropology.

It is not likely that future theoretical and research problems will follow the trends set by functionalism except, perhaps, in re-evaluations of problems in Foucault's adoption of functional terminology in his conceptualization of 'discourse'; in the work of anthropologists who holistically find instances of cultural resistance to what they perceive to be Gramscian style 'hegemony'; and in the relationship of anthropological functionalism to other modernisms (Webster, in Manganaro, 1990).

The main thrust is likely to be the further undercutting of the two pluralized concepts integral to functionalism: 'cultures' and 'societies' (Wolf, 1988; Barth, 1992). This new emergent paradigm, 'New Critical Theory,' retains functionalism on its agenda, albeit in a literary and historical mode. It promises to further problematics addressed by anthropological historical realism (sometimes in the guise of political economy) and postinterpretive cultural critique. In both instances, functionalism, in anthropology, lingers on (just as evolutionism and diffusionism linger on) not just as situated knowledge ever in the making but as historical ethnography of the operations of culture.

See also: Anthropology, History of; Functionalism, History of; Malinowski, Bronislaw (1884–1942); Parsons, Talcott (1902–79).

Bibliography

Barth, F., 1992. Towards greater naturalism in conceptualizing societies. In: Kuper, A. (Ed.), Conceptualizing Society. Routledge, London and New York.

Fox, R.G. (Ed.), 1991. Recapturing Anthropology: Working in the Present. School of American Research Press, Santa Fe, NM.

Gellner, E., 1987. The Zeno of Cracow, or revolution at Nemi, or the Polish revenge. In: Culture, Identity and Politics. Cambridge University Press, Cambridge, UK.

Malinowski, B., 1944. A Scientific Theory of Culture, and Other Essays. University of North Carolina Press, Chapel Hill, NC.

Manganaro, M. (Ed.), 1990. Modernist Anthropology: From Fieldwork to Text. Princeton University Press, Princeton, NJ.

Mead, M. (Ed.), 1937. Cooperation and Competition among Primitive Peoples. McGraw-Hill, New York.

Mintz, S.W. (Ed.), 1985. History, Evolution, and the Concept of Culture: Selected Papers by Alexander Lesser. Cambridge University Press, Cambridge, UK.

Radcliffe-Brown, A.R., 1935. On the concept of function in social science. American Anthropologist 37, 394–402 (reprinted 1952 in Structure and Function in Primitive Society. Free Press, Glencoe, IL, pp. 178–87).

Sahlins, M.D., 1976. Culture and Practical Reason. University of Chicago Press, Chicago.

Sanjek, R. (Ed.), 1990. Fieldnotes: The Makings of Anthropology. Cornell University Press, Ithaca, NY.

Stocking, G.W. (Ed.), 1991. Colonial Situations: Essays on the Contextualization of Ethrographic Knowledge. History of Anthropology, Vol. 7. University of Wisconsin Press, Madison, WI.

Stocking Jr., G.W., 1992. The Ethnographer's Magic and Other Essays in the History of Anthropology. University of Wisconsin Press, Madison, WI.

Stocking Jr., G.W. (Ed.), 1996. Volksgeist as Method and Ethic: Essays on Boasian Ethnography and the German Anthropological Tradition. History of Anthropology, Vol. 8. University of Wisconsin Press, Madison, WI.

Thornton, R., 1985. Imagine yourself set down …: Mach, Frazer, Conrad, Malinowski and the role of imagination in ethnography. Anthropology Today 1 (5), 7–14.

Thornton, R.J., Skalnik, P. (Eds.), 1993. The Early Writings of Bronislaw Malinowski. Cambridge University Press, Cambridge, UK.

Urry, J., 1993. Before Social Anthropology. Essays on the History of British Anthropology. Harwood, Chur, Switzerland.

Vincent, J., 1986. Functionalism revisited: An unsettled science. Reviews in Anthropology 13, 331–339.

Vincent, J., 1991. Engaging historicism. In: Fox, R.G. (Ed.), Recapturing Anthropology: Working in the Present. School of American Research Press, Santa Fe, NM.

Wolf, E., 1988. Inventing society. American Ethnologist 15 (4), 752–761.

Functionalism, History of

Richard Münch, Otto-Friedrich-Universität Bamberg, Bamberg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Starting with a brief description of the essential features of functionalism in the social sciences, the article provides an overview of the major contributions to its development. Comte and Spencer shaped the origins of functionalism in the nineteenth century by conceptualizing society as an organism. Durkheim continued along this line in Sociology. Radcliff-Brown and Malinowski did so in Social Anthropology. Merton, Davis, Moore, and Parsons made functionalism the dominant paradigm of post–World War II sociology. Despite its decline in the 1960s, there has been a significant revival of the paradigm since the 1980s with neofunctionalism and Luhmann's systems theory.

Functionalism is a method of social science analysis, which has its roots in nineteenth-century social thought. It originated from the application of methods from biology, the rising science at that time, to the analysis of society. Three features are characteristic of this orientation toward biology: society is more or less strictly conceived as an organism; its development is understood as an evolutionary process; special attention is given to the relationship between the parts and the whole constituted by them, particularly to functions that have to be fulfilled by the parts for the working of the whole. The parts are institutions such as forms of family life, economic production, government, and religion as well as the organizational embodiments of such institutions: families, enterprises, governmental bodies, and churches. The institutions are binding rules, which regulate the activities and relationships between people. The functions are, for example, biological reproduction for the family, generating resources and allocating them to needs for economic production, binding decisions for the collectivity for the government, or the production of meaning for religion. It is assumed that, for its simple continuation, society needs the subserving of such functions by institutions and their organizational embodiments.

Origins of Functionalism: Auguste Comte and Herbert Spencer

It was the fundamental change from traditional agrarian to modern industrial society, which constituted the background for the rise of functionalism in nineteenth-century social thought, particularly in the work of Auguste Comte and Herbert Spencer. In his *Social Statics*, Comte conceived of society as a whole which, in its ideal state, is kept in perfect harmony by the balanced contribution of its constituent parts to its working in the process of fulfilling their specific functions. In his *Social Dynamics*, he conceptualized social change as an evolutionary process from lower to higher forms of societal organization. Because he expected science to bring about the perfect organization of social life in its final stage in modernity, he couched the evolutionary process in terms of the evolution of thinking from the theological to the metaphysical to the positive stage shaped by 'positive' science.

The other major source of functionalism in nineteenth-century social thought next to Comte is Herbert Spencer (1908). Though he shared the adoption of biological and evolutionary thought with Comte, he demarcated his approach explicitly from Comte's by conceiving society and social development not as a matter of ideas, but as a matter of objective things, namely, the behavior of human individuals. The major process in the development of society is differentiation; functions become more numerous and the tendency is that a function is fulfilled by one specific institution specialized in it. The more society grows in size, the more differentiated will be its functions and corresponding institutions. The first differentiation occurs between regulative, sustaining, and distributive functions. The corresponding institutions are government, economy, and the infrastructure of channels for transport and communication. With the differentiation of functions and institutions, society works like higher organisms, it becomes more heterogeneous, definite in the demarcation of institutions, and integrated; group life is transformed from the coexistence of groups with the same structure without coherence between them (families, clans) to greater heterogeneity but also greater coherence between groups. The evolution of social life goes from incoherent homogeneity to coherent heterogeneity in the process of specialization and growing division of labor.

The Classical Framing of Functionalism: Emile Durkheim

Among the great classical founders of sociology at the turn from the nineteenth to the twentieth century, it is Emile Durkheim (1893/1973a, 1895/1973b, 1912/1968) who made a distinctive contribution to functionalism as a method of sociological analysis. From Comte he borrowed the holistic view of society, from Spencer the idea that the growing division of labor is a core feature of social evolution. In his *Rules of the Sociological Method*, Durkheim made a clear distinction between functional and causal analysis and a clear statement that every functional analysis needs to be complemented by causal analysis. He argues that the utility of an institution in serving a need for individuals or in performing a function for society does not suffice to explain its existence.

In his study on *The Division of Labor in Society*, Durkheim points out that specialization and division of labor are tools for the self-preservation of the individual and the society and for integrating society. They help to overcome the struggle for scarce resources. This struggle is caused by the increasing material and dynamic density that results from population growth and the extending as well as denser weaving of ways and means of transport and communication. Nevertheless, serving self-preservation and social integration do not produce specialization and the division of labor by themselves. In Durkheim's eyes, this functional explanation of specialization and division of labor needs a complementation by causal analysis. In doing so, he looks for forces, which lead the selection of behavior and individuals in the direction of increasing labor division. He says that emigration, suicide, and crime would have been other possible solutions of the problem besides labor division. He explains the fact that they have not been selected to the same extent by the greater resistance called for by them. The bonds to one's home country work against emigration, the bonds to one's life against suicide, the sympathy for our fellows prevents us from committing a crime. They are on average stronger and more resistant feelings than the habits, which could prevent us from further specialization. The latter had to give way to the pressure toward specialization. In doing so, Durkheim points out the 'functional equivalents', which would serve self-preservation and then gives answers to the question as to the forces that lead human behavior in the direction of one of them. This is his complementation of functional analysis by causal explanation in his study on the division of labor.

In his later work on *The Elementary Forms of Religious Life*, Durkheim draws on anthropological studies on the Australian aborigines. Here he makes use of his idea of the circular link between cause and functional effect in the study of ritual practices in which the sacred is strengthened in distinction to the profane and in which the community is regularly becoming conscious of itself. We have a circular link here. People gather because they want to influence the demonic powers positively for practical reasons of a good harvest or hunt, their willingness to come together is, however, supported by their feeling of belongingness and solidarity, which is again strengthened by the ritual itself.

A Paradigm of Functional Analysis: Robert K. Merton

Durkheim's analysis of tribal societies was continued in the anthropological study of tribal societies, particularly by the leading British anthropologists, Alfred R. Raddcliffe-Brown and Bronislaw Malinowski. According to Malinowski (1948), every social phenomenon has to be explained by an indispensable function it fulfills not only for the whole of society, but also for the human individual: biological needs for food, shelter, and reproduction, psychological needs for support, needs for maintaining culture and social organization. Institutions of exchange, decision making, social control, and education should be explained by such functions. Radcliffe-Brown (1952) criticized Malinowski's emphatic functionalism and preferred to call his approach structuralism. He did not accept the presumption that every social phenomenon exists because of a necessary function it fulfills. In his view, the unity of society was not a postulate claimed for every society, but a hypothesis that allowed a functional explanation of a social phenomenon by its contribution to the persistence of society. One of his examples is the explanation of lineage systems by the fact that they help to avoid conflicts about property rights after the death of property holders and thus contribute to the maintenance of solidarity. It is obvious that such an explanation is nevertheless insufficient in the light of Durkheim's methodological objections against purely functional explanations.

At this point, it is quite reasonable to turn to the detailed methodological analysis of functionalism contributed by Robert K. Merton (1949/1968) in his paradigm of functional analysis. Merton points out three postulates of functionalism, which all have to be doubted in his eyes: (1) Functional unity postulates that society is one consistent whole and every phenomenon contributes to its working and that there is no division of society into parts with conflicting needs. (2) Universal functionalism postulates that every social phenomenon performs a positive function for the whole of society. (3) Functional indispensability postulates that every social phenomenon performs an indispensable function for the whole of society and cannot be replaced by some other phenomenon. Merton says that all three postulates are unfulfilled by concrete societies, particularly by modern societies, so that functional explanations of social phenomena cannot be based on them.

Instead, Merton introduced a more realistic paradigm of functional analysis with a set of 12 requirements: (1) Only standardized social phenomena like institutions should be analyzed with regard to their functions. (2) Subjective dispositions of actors and objective consequences of their action in social practices should be outlined and distinguished. (3) Positive, negative (dysfunctional), and neutral effects of standardized practices should be discerned. (4) The units for which an item fulfills a function have to be exactly determined. (5) The functional requirements for maintaining a system have to be denoted. (6) The mechanism of their fulfillment must be specified. (7) The functional alternatives have to be enumerated. (8) The structural context needs description regarding the narrowing down of space for functional alternatives. (9) Dysfunctions have to be detected and analyzed with regard to effecting stress and strain with resulting pressures for change. (10) The functional explanations should be validated in comparative research. (11) Functional analysis must be neutral in ideological terms. (12) Functional analysis of a social practice needs to include (a) the location of participants in the social structure; (b) alternative modes of behavior excluded by the pattern; (c) the emotive and cognitive meaning of the pattern for the participants; (d) a distinction between motivation and objective behavior; and (e) behavioral regularities that are part of the pattern but are not recognized by the participants.

A good example of Merton's idea is his functional explanation of the political machine by the various needs it fulfills for the whole of society and for a variety of groups. As an informal network with a boss at its head, it concentrates power and thus makes political decisions possible in the American political system of checks and balances; it also serves as a vehicle of helping the poor and weak, legitimate and

illegitimate business and provides channels for upward mobility. Thus the political machine becomes stabilized because – according to Merton's requirement – the positive functions outweigh the dysfunctions by far.

Functional Prerequisites and Social Stratification: Aberle et al. and Davis and Moore

In the 1940s and 1950s, functionalism achieved to the position of the dominant paradigm of sociological analysis under the name of "structural functionalism". The investigative strategy of structural functionalism was the discovery of the functions, which have to be fulfilled in order to secure the survival of society and the discovery of the corresponding structures serving these functions (Aberle et al., 1950/1967). Another representative contribution to structural functionalism is the functional theory of social stratification outlined by Kingsley Davis and Wilbert E. Moore (1945) in their much debated article 'Some Principles of Social Stratification' published in the *American Sociological Review*. Davis and Moore argue that every society needs a system of social stratification for the following reasons: (1) The members of the society share common views on the importance of a certain position needed for the provision of public and private goods and services. (2) To perform the corresponding roles, more or less talent, training and/or effort is required. (3) Thus, positions can be more or less important, the personnel needed for filling the positions more or less scarce. The greater the importance and scarcity, the greater the reward. The result is that positions are always stratified. The representative criticism of this functional theory of stratification was formulated by Melvin M. Tumin (1953) in an article published in the *American Sociological Review* 8 years later. It is couched in terms of conflict theory: The class structure is based on the power of the dominant class, which defines the importance of positions; its major effect is the differentiation of access to the higher positions according to socialization advantages and training of the offspring originating from the dominant class. Rather than contributing to the working and integration of society, the stratification system has disruptive effects so that society is not stabilized by value consensus and stratification, but by the application of power.

Functionalism as an Attempt at Synthesis: From Talcott Parsons' Structural Functionalism to Neofunctionalism

Five years earlier than Davis and Moore, in 1940, Talcott Parsons published another functional theory of social stratification, which he revised twice, in 1954 and 1975 (Parsons, 1954a,b, 1977). In the perspective developed by Parsons, social stratification has to be explained by the interaction of dynamic and controlling forces of action, interests and power on the one hand, and solidarities and cultural values on the other. Interests and power determine the access to positions and the enforcement of positional differences and outcomes, particularistic solidarities of family and lineage explain the inertia of class structures, because of the tendency of families to secure their offspring remaining in the same class position. The *legitimacy* of the stratification system in distinction to its factual production and reproduction is, however, an effect of the consensus on a grading scale according to which prestige is attributed to positions. Only in as much as such a consensus on a grading scale is achieved will the system of stratification be accepted as legitimate. Otherwise, the definition of the importance of positions and the access to them will be a matter of the war of all against all. In general terms, this is the problem of double contingency.

Other than Merton's deliberate plea for empirical middle range theories, Parsons' program was analytical grand theory. Merton stood for empirical functionalism, Parsons for analytical functionalism (Parsons, 1951). He introduced his four-function paradigm as an analytical instrument of analyzing structures as well as dynamic processes with regard to abstract functions (Parsons, 1966, 1967, 1978; Parsons and Smelser, 1956). By cross tabulating the instrumental/consummatory and the internal/external orientation of action, he defined four abstract functions: instrumental/external: adaptation (A), consummatory/external: goal attainment (G), consummatory/internal: integration (I), and instrumental/internal: latent pattern maintenance (L). For further differentiation, Parsons distinguishes three levels of applying the four-function scheme: (1) social system: economic (A), political (G), community (I), and fiduciary (L) system; (2) general action system: behavioral (A), personality (G), social (I), and cultural (L) system; and (3) human condition: physicochemical (A), organic (G), action (I), and telic (L) system. Further differentiation results from the application of the scheme to one of the subsystems. After having developed an analytical action theory in his first work of 1937, *The Structure of Social Action* (Parsons, 1937/1968), Parsons established structural functionalism as a link between analytical functionalism and systems theory in *The Social System* in the early 1950s, and extended that link with the introduction of his theory of generalized media of interchange (money (A), political power (G), influence (I), and value commitments (L) for the social system) and his theory of evolution as instruments for the analysis of dynamic processes of inflation and deflation and social change in the 1960s. With his turn to functionalism, systems, media and evolutionary theory he did, however, not give up the core position of his early action theory: the idea that social order emerges from the interaction between controlling and dynamizing forces. In his systems theory, he couched this idea in the double cybernetic hierarchy of conditions and controls. From (L) to (A) energy and dynamics in the system increase, while information and control decrease, and vice versa.

Looking at one analytical level, the social system, this system is composed of subsystems, which fulfill their specific functions (A G I L) with specific structures (markets (A), authority (G), association (I), and communication (L)) and specific media of interchange. They build an interrelated set of systems with in- and output processes of factors and products processed by the circulation of the media from one system to the other (interpenetration). Crises in the working of the systems can be analyzed as inflation–deflation spirals of the media involved.

In his theory of evolution, Parsons tries to point out the evolutionary advantage of specific institutions. Religion, communication by language, social organization through kinship and technology are the first conditions of building human societies. A breakthrough to higher forms of social organization is realized by social stratification according to prestige based on achievement instead of ascription by origin. The rigidity of kinship systems is pushed back in this way. A further step forward is the replacement of tradition by cultural legitimation through general values. The developmental stage of a modern society is realized by the establishment of bureaucratic organizations, markets and money, universalistic norms and democratic association. In two volumes, he outlines the evolutionary process in four dimensions according to his AGIL scheme: adaptive upgrading (A), structural differentiation (G), inclusion (I), and value generalization (L).

Evolutionary achievements require breakthroughs in all four dimensions. He points out such breakthroughs in his analysis of primitive societies, archaic societies (Egypt and Mesopotamia), intermediate historic empires (China, India, Islamic, and Roman empires), seedbed societies (Israel, Greece), and, finally, the system of modern societies from ancient cultural roots via the Renaissance and the Reformation to the Enlightenment and the building of modern citizenship, state and capitalism. The major modern breakthroughs are the industrial, the democratic, the educational, and eventually the 'expressive' revolution of the 1960s and 1970s (Parsons, 1966, 1967, 1971).

According to his position that social order always has to reside in some form of consensus and solidarity in order to avoid an unlimited power struggle, Parsons lays special emphasis on the development of citizenship as a core of a modern 'societal community'. It is a modern form of community, which has the greatest potential for inclusion, it associates people of the most varied ethnic and cultural origins. In this respect, it is the cornerstone of modern social organization. He sees the society of the United States farther developed in this respect than the ethnically more homogeneous European welfare states (Parsons, 2007). The modern community of citizens is the core, which is complemented by a capitalist economy; democratic government; and universalistic institutions of education, science, professional work, and voluntary religious association. This is the abstract path of sociocultural evolution, which has to be distinguished from concrete historical development that can more or less deviate from this path. There is particularly always the danger of fundamentalist movements against modernity, which originate from the severe tensions caused by the evolution of modern forms of social organization. For Parsons, fascism as well as communism were such fundamentalist reactions to modernity, which have, however, no chance of long-enduring stabilization in his eyes. Therefore, he expected that the Soviet system would crumble some time, an expectation grounded in 'abstract' evolutionary theory, which turned out to be very 'realistic' eventually.

Criticism of Functionalism

The dominant role of functionalism during the 1950s and into the 1960s provoked criticism pointing out its weaknesses and aiming at new approaches to study society. In the 1950s already, Lewis Coser (1956) and Ralf Dahrendorf (1958a,b) objected to functionalism with regard to its inadequate understanding of social conflict. Drawing on Georg Simmel, Coser emphasized that social conflict is not only disruptive in its effects, but – under certain specific conditions – also meets functions of maintaining social relations and integrating society. He developed a kind of conflict functionalism. There are conditions that mitigate conflicts such as chances of expressing frustration openly, a greater number of alternatives for action, little frustration before a conflict becomes evident, exclusion of emotional and ideological motives, and a focus on the substantial objects of conflict and stable social relations. External conflicts with other groups strengthen the groups' identity and cohesion, but also coerce nonconforming members of the group. Also, external conflicts contribute to stabilizing and strengthening relations between the conflicting individuals or groups.

Ralf Dahrendorf criticized functionalism's one-sided focus on consensus, social integration and social stability. He did not want to abandon the functionalist view of society, but wanted to complement it with a special emphasis on constraint, conflict, and change. As he puts it, functionalism considers society as a relatively stable and well integrated configuration of elements, all of which contribute to its functioning, while society is founded on consensus. From the complementary point of view of conflict theory, society is based on the constraint exercised by some of its members on the other members, and it is in continuous change entailing always conflict. All its elements contribute to its change. George Homans (1961, 1964) challenged functionalism's holistic approach to study society and to focus on institutions, structures, and functions of the society at large and the corresponding search for "grand theory" with a plea for methodological individualism. He argued for "bringing men back in" and for a sociological explanation based on psychological laws that he derived from behavioristic learning psychology. Any explanation of social institutions and social change is required to trace them back to the actions of human individuals. For example, the stability or change of an institution like the appearing of students in class at 8 a.m. every morning of a school day resides in the students' motivation to appear or not appear in class at that time every morning of a school day. Homans initiated a kind of micro-sociological revolution against functionalism (Alexander, 1987: pp. 156–280). Besides Homans' psychologically based theory of human behavior, other parts of this revolution included symbolic interactionism as it was advanced by Herbert Blumer (1969) based on George Herbert Mead's social behaviorism; Erving Goffman's dramaturgical theory of interpersonal communication; and Harold Garfinkel's studies in ethnomethodology. These approaches were addressing the construction of social order in social interaction. They all shared Denis Wrong's (1961) criticism of Parsons' "oversocialized conception of man". Wrong emphasized that members of social groups, communities, organizations or societies are not completely socialized to the extent that their action is nothing but role-taking and conformity to norms and role expectations. Socialization itself is

a conflictual process and is incomplete so that there is always space for deviation from established norms and for change of social order. From the point of view of symbolic interactionism, any social order is not given once and for all, but has to be negotiated and renegotiated time and again in everyday interaction (Strauss, 1978).

From the macro-sociological perspective, it was first of all C. Wright Mills (1956, 1959) who challenged what the critiques called functionalism's preoccupation with consensus, order and stability and the understanding of society as a widely harmonious whole composed of interdependent parts that all contribute to its functioning. From Mills' point of view, this preoccupation with consensus, order and stability made functionalists unable to address power, conflict and constraint in a realistic way. This was particularly true for the functionalist perspective on American society. As Mills elaborated in his widely received study *The Power Elite*, American society underwent a major change after World War II and in the wake of the Cold War between East and West. Up to that time, the business elite was most influential in shaping society and politics. With the Cold War, the federal government's leadership became more important along with the military leaders. Mills saw American society widely ruled by what he called the military–industrial complex, that is a close network of leaders from military and big business and their involvement in the federal government. In Parsons' (1960: pp. 206–247) eyes, Mills exaggerated the connections between military, business, and government and underestimated the functional differentiation of society. For Parsons, military, business, and government fulfill specific functions for society. Their leaders play a crucial role in meeting these functions, but do not unite in a closed elite that rules society in its particularistic interest.

Here is the crucial point at which Alvin Gouldner's (1970) critique of Parsonian functionalism set in. While the critiques mentioned so far were directed against functionalism in a still dominating position, Gouldner launched his criticism at a time at which the decline of the paradigm was well under way. His book *The Coming Crisis of Western Sociology* was written as an account of this decline and put forward a plea for a reflexive and a radical sociology. Gouldner elaborated the similarity of Parsonian functionalism to Plato's philosophy of a partially good world, which can be improved by moving toward an ideal order where every part of it is in equilibrium with all other parts and contributes to the functioning of the whole. Society is a whole composed of interdependent parts that all contribute to its functioning. By looking at society in this way, functionalism is conservative in the same way as Plato's philosophy was. That means, even if it does not directly aim at conserving a concretely existing society and its institutions, it has to look at it in a way so that practices and institutions are identified in their functional contribution to society and its order. Even if problematical phenomena can always be observed in society, they can only be conceived as deficient, which calls for improvement in order to meet the required function for society. Thus the functionalist perspective may see deviance, conflict, and disorganization in society but will never open the observer's eyes for the deeper causes of these phenomena and will always search for remedies that help to improve given institutions so that the existing society arrives a better shape and greater balance. This conservatism inherent in functionalism implies its adaptation to society as it exists. It may fulfill this conservative 'function' in a liberal capitalist society, in a society characterized by capitalism's embeddedness in democracy and welfare institutions and even a socialist society. Against this conservative character of functionalism, Gouldner argues for a reflexive sociology that reveals the deeper roots of contemporary tensions and conflicts in society and a radical sociology that opens up ways toward a more profound change of society and toward a better future.

Marxism became a major stream of radical sociology in the 1970s. The revitalization of Marxism at that time entailed, however, a remarkable bloom of functionalist thinking in its broader terms. Social theory inspired by Marx had to find an answer to the question as to why capitalism did not break down in the West as postulated by Marx in his study on the rise and future decline of capitalism. The widespread answer to this question was the rise of stabilizing institutions like science and technology, social partnership between industrial federations and trade unions and government interventions in the economy aiming at minimizing its disruptive effects. It was Jon Elster (1982) who addressed exactly this turn of Marxism to a kind of functionalist method of analysis. Elster referred to the logical incorrectness of functionalist explanations as already identified by Carl G. Hempel from a methodological point of view. As he emphasized, it is logically impossible to explain the existence or rise of an institution by the function it fulfills for society. Against the preoccupation of functionalism with methodological holism – the whole is more than its constituent parts – he argued in favor of methodological individualism. The emergence and existence of institutions have to be explained by the motives and actions of the major actors who have created them or are maintaining them even against resistance at a certain time in history. Elster identifies Christian theodicies as a major root of functionalism reaching their summit in Leibniz's teaching that everything happening in this world will end up as good for the human individual. For him, Bernhard de Mandeville's teaching that private vices produce public benefits constitutes a secular version of Christian theodicy. Adam Smith's statement that some invisible hand turns private utility maximization into public wealth on the market also belongs to this category of functionalist reasoning. Elster considers this type of reasoning as a weak functional paradigm which assumes that some behavioral pattern has beneficial consequences for some social unit, but is unintended and not recognized as such by the beneficiaries (Elster, 1982: p. 454). Elster also identifies a main functional paradigm that explains the existence of an institution or behavioral pattern by referring to their latent functions for a social unit, and a strong functional paradigm, which assumes that all institutions and behavioral patterns can be explained by a function they fulfill for a social unit. Elster supplies a variety of examples for these types of functional explanations in the social sciences, particularly in Marxism. They all share the assumption of some objective teleology in social processes such as Hegel's idea that history develops dialectically in the

direction of reality coming closer to reason, that means the real turns out to be the reasonable. Elster gives a variety of examples of this kind of functionalist reasoning and he shows that not only functionalism in mainstream sociology adheres to this type of reasoning, but also, very broadly, Marxism from its very beginning in Marx's writings themselves through to the renewal of Marxism in the 1970s in particular. One example is Marx's explanation of the Bonapartist regime that deprives the bourgeoisie of exerting political power directly in order to protect it from self-destruction so that it maintains its dominant position in society in the long run (Elster, 1982: p. 458). Elster demonstrates furthermore that many Marxist analyses explain what exists by its function for maintaining capitalism and serving the capitalist class. The state in whatever form it may exist is explained by its function for capitalism and the capitalist class. As Elster points out, all these functional explanations are deeply flawed, as they simply cannot explain what they want to explain. There are always functional alternatives and the attributed function for a social unit does not bring about an institution or behavioral pattern by itself. For Elster, there is no other way than looking at the actors and their motives in creating and maintaining institutions and behavioral patterns. This is what methodological individualism requires from sociological explanation.

Functionalism and Modern Systems Theory

The dominant role played by Parsonian functionalism in the 1950s provoked counter-movements, which concentrated mostly on some particular aspect of social organization, which was relativized within the very general Parsonian scheme. The need for revitalizing the Parsonian approach has indeed called for a new movement, which aims at continuing with the general program of Parsons on the contemporary level of development. This is the movement of neo-functionalism, which arose in the 1980s. Neofunctionalism tries to advance the Parsonian legacy by way of learning from new developments in social theory and social research, for example from advancements provided by Clifford Geertz or Jürgen Habermas. This new movement has contributed fresh interpretations of Parsons' work along with a new reading of the classics, studies on civil society and public discourse, on the evolution of the culture of modernity, on communication society, European integration and globalization (Alexander, 1982/83, 1998, 2006; Münch, 2010, 1987/2011, 2012).

A comprehensive renewal of functionalism has been undertaken by Niklas Luhmann (1970, 1984, 1997). He, too, combines functionalism with systems, media, and evolutionary theory. In his early phase of theory-building, Luhmann introduced a turn from structural-functional to functional-structural systems theory. To justify this turn, he claims that the question of function has to be posed in a more fundamental way. Before we ask for the structures of systems, which have to fulfill certain functions, we should raise the question as to which function is served by systems building as such. What is the problem solved by systems building? This is how he asks his question. His answer is: world complexity, which needs reduction so that it can be handled. Systems just perform this function of reducing world complexity.

In his second phase of theory building, Luhmann undertook another turn of his functionalist systems theory: the conceptualization of social systems as autopoietic, self-referential systems. Because there is always the danger of too much complexity, the major problem is systems building and the continued maintenance of systems boundaries, which are boundaries of meaning. Autopoiesis is a form of systemic operation, which guarantees the maintenance of boundaries and, at the same time, allows for maximum absorption of environmental complexity. An autopoietic system is operationally closed, but open to any kind of complexity in the environment. Operational closure is guaranteed by a binary code according to which any kind of environmental complexity is processed. The concrete processing is organized by a system's program. Special links to other systems in the environment are stabilized by structural coupling. Generalized media of communication allow for most flexible forms of systemic processing: money, power, law, and love, for example. Social systems are built up by communication, they originate from that and continue to exist as long as communication takes place according to the established code. Communication consists of three elements: information, message, and interpretation.

An autopoietic system produces and reproduces all its elements – communication – by itself, i.e., by systemic communication, but it makes use of environmental complexity. It does not take inputs from the environment to transform them into outputs, but receives impulses from the environment to which it shows more or less response. The system observes events in the environment and absorbs them according to its own code and program. Problems of maladaptation to the environment can be discovered by second-order observation, which is self-observation and self-description.

Luhmann distinguishes three levels of social systems: interaction systems, organizational systems, and societal systems. According to his theory of evolution, more complex societal systems evolve by variation, selection, and stabilization. There are three (later: four) stages of societal evolution, which are defined by the form of differentiation of systems: segmentary, estate (later: center/periphery), and functional differentiation.

Modern societies are functionally differentiated and no longer have a center. The economic, political, legal, science, educational, and arts systems are more and more differentiated on a world scale and thus transcend their centering around a societal community and/or a state. Thus the nation-state societies are transitional stages on the way to world society. Paying vs. not paying, having power vs. having no power, right vs. wrong, true vs. false are the binary codes of the economy, polity, law, and science systems. The evolutionary process works toward the establishment of functionally differentiated systems on a world scale, which are linked by structural coupling and by mutually stabilizing coevolution.

Niklas Luhmann's systems theory attracted great interest, particularly in the 1970 and 1980s, and above all in Germany. Many younger sociologists adopted his paradigm of autopoietically operating systems. Also, he provoked some

controversy as to the best approach to social theory. It was Jürgen Habermas, working in the tradition of critical theory, who launched the most thoroughgoing criticism against systems theory in his critique of functionalist reason. He borrowed from Parsons's and Luhmann's works to construct a twofold social theory, a theory of communicative action that focused on the interaction between systems and life-world (Habermas, 1981). In reference to Marx's concept of alienation in modern capitalism, he developed the thesis of the colonization of the life-world of communicative relationships between people by the systems of economy, politics, bureaucracy and law.

Future Need for Functionalism

The question finally is: What will be the future of functionalism after its renewal by neofunctionalism and by Luhmann's functionalist systems theory? Whatever the movements against functionalism have contributed to our knowledge, we cannot say that they can replace the insights of functionalism completely. They have merely emphasized some particular aspect of social reality and some particular question. But they did neither raise nor answer the questions posed by functionalism from Comte and Spencer to Parsons and Luhmann. Because their questions are still legitimate, there is still a place for functionalist theorizing, whatever form it might take in the future.

See also: AGIL, History of; Anthropology, History of; Community and Society: History of the Concepts; Discipline Formation in the Social Sciences; Division of Labor: History of the Concept; Modernity: History of the Concept; Progress: History of the Concept; Schools in the Social and Behavioral Sciences: Concepts and Historical Relevance; Society and People: History of the Concept; Sociology, History of; Solidarity: History of the Concept.

Bibliography

Aberle, D.F., et al., 1950/1967. The functional prerequisites of a society. In: Demerath, N., Peterson, R. (Eds.), System, Change and Conflict. Free Press, New York, pp. 317–331.
Alexander, J.C., 1982/83. Theoretical Logic in Sociology, 6 vols. University of California Press, Berkeley and Los Angeles, CA.
Alexander, J.C., 1987. Twenty Lectures. Sociological Theory since World War II. Columbia University Press, New York.
Alexander, J.C., 1998. Neofunctionalism and After. Blackwell, Oxford.
Alexander, J.C., 2006. The Civil Sphere. Oxford University Press, Oxford.
Blumer, H., 1969. Symbolic Interactionism. Prentice Hall, Englewood Cliffs, NJ.
Comte, A., 1830–42/1969. Cours de philosophie positive. In: Culture et Civilisation, 6 vols. Brussels (1855. The Positive Philosophy of Auguste Comte. Calvin Blanchard, New York).
Coser, L., 1956. The Functions of Social Conflict. Free Press, New York.
Dahrendorf, R., 1958a. Out of utopia. Toward a reorientation of sociological analysis. American Journal of Sociology 64, 115–127.
Dahrendorf, R., 1958b. Toward a theory of social conflict. Journal of Conflict Resolution 2, 170–183.
Davis, K., Moore, W.E., 1945. Some principles of stratification. American Sociological Review 10, 242–249.
Durkheim, E., 1893/1973a. De la division du travail social. Presses Universitaires de France, Paris (1964. The Division of Labor in Society (G. Simpson, Trans.). Free Press, New York).
Durkheim, E., 1895/1973b. Les règles de la méthode sociologique. Presses Universitaires de France, Paris (1982. The Rules of Sociological Method. In: Lukes, S. (Ed.) (W.D. Halls, Trans.). Macmillan, London).
Durkheim, E., 1912/1968. Les formes élémentaires de la vie religieuse. Presses Universitaires de France, Paris (1976. The Elementary Forms of Religious Life (J.W. Swain, Trans.). Allen & Unwin, London).
Elster, J., 1982. Marxism, functionalism and game theory. The case for methodological individualism. Theory and Society 11, 453–483.
Garfinkel, H., 1967. Studies in Ethnomethodology. Prentice Hall, Englewood Cliffs, NJ.
Goffman, E., 1959. The Presentation of Self in Everyday Life. Penguin Books, Harmondsworth.
Gouldner, A.W., 1970. The Coming Crisis of Western Sociology. Heinemann, London.
Habermas, J., 1981. Theorie des kommunikativen Handelns. Suhrkamp, Frankfurt a.M.
Homans, G.C., 1961. Social Behavior. Its Elementary Forms. Harcourt, Brace, New York.
Homans, G.C., 1964. Bringing men back in. American Sociological Review 29, 809–818.
Luhmann, N., 1970. Soziologische Aufklärung, vol. 1. Westdeutscher Verlag, Opladen (1982. The Differentiation of Society (S. Holmes, C. Larmore, Trans.). Columbia University Press, New York).
Luhmann, N., 1984. Soziale Systeme. Grundriß einer allgemeinen Theorie. Suhrkamp, Frankfurt a.M.
Luhmann, N., 1997. Die Gesellschaft der Gesellschaft, 2 vols. Suhrkamp, Frankfurt a.M.
Malinowski, B., 1948. Magic, Science and Religion. Doubleday, New York.
Merton, R.K., 1949/1968. Manifest and Latent Functions. Social Theory and Social Structure. Free Press, New York, 73–138.
Mills, C.W., 1956. The Power Elite. Oxford University Press, New York.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, New York.
Münch, R., 2010. European Governmentality. The Liberal Drift of Multilevel Governance. Routledge, London and New York.
Münch, R., 1987/2011. Theory of Action. Towards a New Synthesis Going beyond Parsons. Routledge, London and New York.
Münch, R., 2012. Inclusion and Exclusion in the Liberal Competition State. The Cult of the Individual. Routledge, London and New York.
Parsons, T., 1937/1968. The Structure of Social Action. Free Press, Glencoe, IL.
Parsons, T., 1951. The Social System. Free Press, Glencoe, IL.
Parsons, T., 1954a. Essays in Sociological Theory. Free Press, New York, pp. 69–88.
Parsons, T., 1954b. Essays in Sociological Theory. Free Press, New York, pp. 386–439.
Parsons, T., 1960. Structure and Process in Modern Societies. Free Press, Glencoe, IL.
Parsons, T., 1966. Societies. Evolutionary and Comparative Perspectives. Prentice Hall, Englewood Cliffs, NJ.
Parsons, T., 1967. Sociological Theory and Modern Society. Free Press, New York.
Parsons, T., 1971. The System of Modern Societies. Prentice Hall, Englewood Cliffs, NJ.
Parsons, T., 1977. Social Systems and the Evolution of Action Theory. Free Press, New York, 321–380.
Parsons, T., 1978. Action Theory and the Human Condition. Free Press, New York.
Parsons, T., 2007. American society. In: Sciortino, G. (Ed.), A Theory of the Societal Community. Paradigm Publishers, Boulder, CO.
Parsons, T., Smelser, N.J., 1956. Economy and Society. Free Press, New York.
Radcliffe-Brown, A.R., 1952. Structure and Function in Primitive Society. Macmillan, New York.
Spencer, H., 1908. The Principles of Sociology, 3 vols. Appleton, New York.
Strauss, A., 1978. Negotiations. Varieties, Contexts, Processes and Social Order. Jossey-Bass, San Francisco, CA.
Tumin, M.M., 1953. Some principles of stratification: a critical analysis. American Sociological Review 18, 387–394.
Wrong, D., 1961. The oversocialized conception of man. American Sociological Review 26, 183–193.

Functioning, Disability and Health, International Classification of

Alarcos Cieza, University of Southampton, Southampton, UK
Jerome E Bickenbach, Leiter Disability Policy Group, Schweizer Paraplegiker-Forschung, Nottwil, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The World Health Organization's *International Classification of Functioning, Disability and Health* (ICF) constitutes a shift in how we understand and as a society respond to disability. Conceptually, the ICF embodies an integrative understanding of the universal human experience of disability and as a practical tool it provides a multidimensional classification and standard language for health information. We focus here on the conceptual framework of the ICF as well as its classificatory structure and health applications before considering issues and applications particularly relevant to the social and behavioral sciences that are. We also contrast ICF George Engel's 'biopsychosocial' model and ICF's role in operationalizing health.

Introduction

The endorsement of World Health Organization's International Classification of Functioning, Disability and Health (ICF) (WHO, 2001) by the 54th World Health Assembly in 2001 represented both a deeply conceptual and practically significant advance for natural, social, behavioral sciences in the conceptualization, description, and assessment of human functioning and disability. Conceptually, the advance arises from a person–environment interaction model, termed as the 'biopsychosocial model,' upon which the ICF is based. This model avoids sociological and biomedical reductionisms and provides a starting point for a comprehensive and integrative understanding of the universal human experience of functioning and disability, and by extension the experience of health itself (Cieza et al., 2008). Practically, the ICF provides a multidimensional classification that constitutes both an international standard language of functioning and disability, and the ontology of a systematic description of the total health experience.

In this article, we will first focus on the conceptual framework, or model, of the ICF before moving on to details of its classificatory structure. That will be followed by a brief review of the potential uses of the ICF within the health sciences, broadly construed, and conclude with a discussion of the range of applications that are particularly relevant to the social and behavioral sciences. Along the way, and to avoid a natural confusion, we will contrast the ICF model with George Engel's influential, and similarly named biopsychosocial model (Engel, 1977). We will also explore the potential for the ICF – both as a conceptual model and a classification of functioning – to provide a much-needed operationalization of the contentious notion of health.

The ICF Biopsychosocial Integrative Model

For more than 40 years models of disability have proliferated, shedding more heat than light (for overviews see Shakespeare, 2006; Bickenbach, 2012). The debate, part science and part politics, has pitted the 'medical model' against 'the social model' – the first purportedly reducing disability to biomedical facts, the second reducing it to the social construction of a marginalized and stigmatized subpopulation. The developers of the ICF sought to avoid both extremes, without abandoning what is valuable in each position. The resulting model implemented the following interlocking principles:

First, disability should be understood as a universal feature of humanity, not the mark of a minority population. Disability is not some 'special' attribute that some people possess and others lack: at bottom disability arises from the blunt fact that human life arises out of a myriad of biological functions that are, in the end of the day, both fragile and susceptible to less-than-optimal levels of functioning – either congenitally, or because of disease, disorder or injury at some point in life. Aging itself is a matter of a progressive decline in biological functions – yet another reason to see disability as a normal feature of living. From a policy perspective, the universality of disability, as the disability sociologist Irving Zola recognized 25 years ago, 'demystifies the specialness of disability,' opening the door to mainstreaming disability across all sectors of society (Zola, 1989).

Disability is universal for another important reason. All biological functions, and the vast number of human actions and behaviors that they make possible, are continuous not dichotomous phenomenon. Each biological function operates in the domain of 'more or less' not 'all or nothing.' Although some manifestations of disability are total – blindness, for example – most are a matter of degree. Where on this continuum between 'complete' functioning and 'total lack' of functioning constitutes a problem (termed in the ICF, an impairment) is a matter of judgment and susceptible to a wide range of considerations, both personal and social. As a consequence, the ICF is principally a classification of functioning and only derivatively of problems or decrements of functioning, that is, disability.

Finally, the experience of disability depends, not merely on biological functions and their impact on the actions and behaviors that people can perform, but also and crucially on the total context in which people live. That is, functioning and disability are outcomes of an interaction between intrinsic features of people (usually expressed in language of biology and psychology) and the physical, human-built, attitudinal, social, political, economic, and cultural environment in which people

live. In particular, disability arises from both biology and the physical and social world. And again, the point at which the operation of a biological function becomes problematic for an individual will depend both on what the person wants to do or behave and what is possible given the person's environment, what is required or expected to do or behave.

To fully operationalize the principles of universality, continuity, and interaction, the ICF 'biopsychosocial model' is constructed in terms of three dimensions of functioning: body functions and structures, activities, and participation. Problems or difficulties in these dimensions are termed impairments, activity limitations, and participation restrictions. For terminological consistency, a decision was made to jointly call these three negative dimensions 'disability.' To complete the model, the two primary categories of determinants of disability are added. First, the underlying health condition (understood as a health state that yields problems of functioning, that is, diseases, disorders, and injuries), and secondly, contextual factors of environmental and personal factors (the latter only mentioned and left for further development). The environmental factors component reflect the long-accepted account of health as being, at least in part, determined by underlying physical and social determinants (cf Whitehead et al., 2003). Diagrammatically, the ICF model is represented in Figure 1.

As an example, an individual who has incurred a spinal cord injury will, as a result, experience a range of impairments of body function and structure, including lower body paralysis, loss of bowel and bladder control, and muscle wasting. These and other impairments will directly affect the person's physical capacity to perform a variety of simple and complex activities, including walking and managing excretion, and may well affect other very complex social activities such as making friends or pursuing education or employment. These activities and tasks, however, can be ameliorated to one degree or another by changes to the person's environment that can facilitate activities and participation. The most obvious is the provision of a wheelchair and other assistive devices. Less obvious, but often more consequential, are alterations to the physical environment of the school or place of employment (in the form of elevators, adjustable desks, personal assistance, and so on). These facilitators enable the individual to carry out the appropriate complex tasks. Finally, less obvious still, but also highly significant, are those social, cultural, and political environment changes that either remove attitudinal or cultural barriers such as stigma and discrimination, or else effect those political changes that ensure full social inclusion.

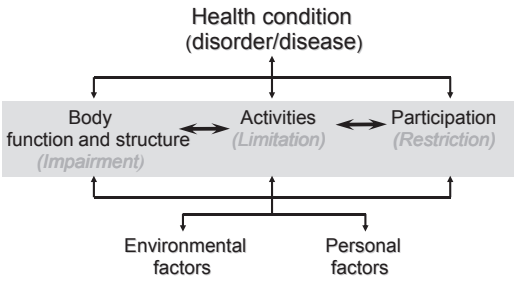

Figure 1 The model of functioning and disability on which the ICF is based.

Other health conditions and impairments will yield a very different portrait of the disability experience. For contrast, think of mental health conditions in which the impact of attitudinal environments, in the form of stigma, fear, and misunderstanding, can be far more determinative of limitations in social participation than the underlying health condition itself. People with noticeably aberrant or nonstandard behaviors – from simple facial tics and spasms to complex forms of antisocial behavior – may be more limited (and hence, more disabled) by their social environment than by their impairments. Yet, some mental health problems such as depression and extreme forms of dementia can rarely be ameliorated by environmental facilitators (other than, perhaps, medication) and are more dramatically determined by the underlying health condition and resulting impairments.

Significantly, each dimension of disability is associated with a range of appropriate social responses or interventions. At the level of impairment, medical or rehabilitative responses are the most appropriate, and in this sense, the 'medical model' is an accurate, though partial, representation of the disability experience. At the person level of activity limitations – operationally identified by assessing a person's capacity to perform activities in a given or standardize environment – the appropriate responses will either aim to increase inherent capacity levels, or else designed assistive technology that augments capacity suitable for actual contexts or environments appropriate to the individual's needs. Finally, at the level of the person-in-context, the paradigmatic interventions are those that either remove environmental barriers to full participation or else put into place environmental facilitators, or both. At this, the most complete level of experience, the appropriate interventions should be conceived as including, not merely the physical, organizational, and attitudinal environmental alteration, but also all of the social and political antecedents needed to bring about these environment change: research, political mobilization and lobbying, legal change, resource allocation, policy creation and implementation, and so on. This is true whether the environmental alteration involves putting in curb cuts for greater wheelchair mobility, or instituting a policy of positive action in public-sector hiring.

The ICF conceptualization of functioning and disability identifies both underlying impairments and the overall context or environment in which the person lives as coequal determinants, with the precise configuration of influences being identifiable on a case by case basis. Unlike both the social and the medical models, the ICF provides no account of how disability comes about in general, but offers a relatively theory-neutral descriptive model that is a template for explanation. If a researcher wishes to show that, say of an instance of employment restrictions for people with mobility problems, this is primarily the product of social prejudice, or unaccommodating workplace environments, or the underlying impairment, the researcher can rely on the ICF to collect relevant evidence to make the case. In short, the ICF's biopsychosocial model rejects both the medical and the social model are reductively distorting, while preserving the germ of truth contained in each.

Comparing the ICF Model with Engel's Biopsychosocial Model

Because of the similarity in the label, it is tempting to compare the ICF 'biopsychosocial model of functioning and disability' with the approach famously offered up in 1977 by psychiatrist George Engel. As it happens, the similarity between the two models ends with the common label, and it is instructive to see why this is so.

In his 1977 presentation of his model, Engel's concern was to "provide a basis for understanding the determinants of disease and arriving at rational treatments and patterns of health care" (Engel, 1977, p. 196). His critique of the 'biomedical model of disease' was in part conceptual – rejecting its reductivism and tendency to dismiss nonbiomedical explanations as irrelevant – but was for the most part clinical or therapeutic. His biopsychosocial model was a plea for the scientific legitimacy of investigating nonbiological, or what we now call social determinants of disease (see **Figure 2**); yet all of his arguments focus on the single, purely clinical phenomena that patients seek medical help not merely because of experienced biological dysfunction, but also because of expectations about what it means to be healthy, expectations that are socially, culturally, and even linguistically, determined.

At one point, he argues that, biomedically, the state of grief has all the hallmarks of a disease – bodily and psychological disturbances and a predictable symptomatology and prognosis. But, significantly, grief is rarely so perceived because, socially and culturally, it is understood as a 'problem of living' not a disease. In short, patients acknowledge diseases only if they perceive them as sickness, and they will only do so within the framework provided by psychological and cultural expectations. The physician, Engel concludes, must therefore supplement her or his biomedical knowledge, with skill in social and psychological perceptiveness in all therapeutic encounters.

By contrast, the ICF does not offer any account of the determinants of disease or other health conditions, and for this reason alone the two models are radically different. But in addition, the ICF does not make any distinction between health states and the perception or feeling of being 'sick.' The ICF is a wholly objective, descriptive tool. Of course, it is evident that a person can have a disease and not feel sick, or feel sick and be perfectly healthy: this is a crucially important clinical realization that, as Engel argued, a responsible physician needs to take on board. But the distinction between health and sickness is irrelevant for the description of health states and the dimensions of functioning that can, in interaction with environmental factors, led to disability. Description, as a precondition for assessment and measurement, is the raison d'etre of the ICF.

Finally, it is not entirely accurate to say that the ICF utilizes a biopsychosocial model at all, at least one that bears a family resemblance to Engel's model of determinants of disease. In fact, although many of the environmental factors are social phenomena, many if not most environmental factors are physical things in the world – weather, buildings, elevators, and other people. It might have been better to use the terminology now becoming more common in the disability literature and naming the ICF model the 'environmental' or 'interactive' model.

Understanding Health in Light of the ICF

Conceptually, how might the ICF assist in our understanding of health? Needless to say, health is a highly contested notion, even though we have an intuitive grasp of what health is or at least what it means to us. The technical, philosophical debate over the definition of health – fluctuating between purely descriptive and biological conceptions (e.g., Boorse, 1975, 1977) to normative accounts linking, or equating, health with well-being (e.g., Nordenfeld, 1987) – is inconclusive. WHO's famous 1948 definition of health as "a state of complete physical, mental and social well-being and not

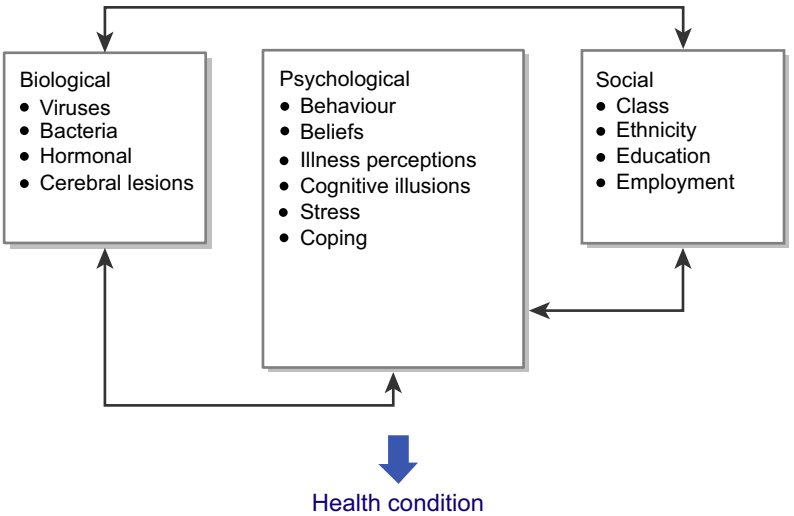

Figure 2 Engel's model as depicted by Albery and Munafò (2008) is presented in the upper part. The mentioned psychosocial factors represent the determinants of a health condition. The ICF biopsychosocial model is presented in the lower part.

merely the absence of disease or infirmity" has been roundly criticized (e.g., Calahan, 1973) for being utterly resistant to scientific application. Yet, though WHO has never abandon its aspirational definition, in more technical papers it has responded to the need for a basis for quantifying health, without which not even the first steps in clinical quality assurance, basic epidemiology, or the evaluation of public health strategies could be taken. This approach relies on the ICF notion of functioning.

In 2003, WHO offer a conceptual of health 'for measurement purposes' that built on the following intuitions (Salomon et al., 2003):

- Health is a separate concept from well-being, although it is of intrinsic value to human beings and instrumentally linked to other components of well-being.
- Health is comprised of states or conditions of the human body and mind, and therefore any attempt to measure health must include measures of body and mind functions.
- Health is an attribute of an individual person, although aggregate measures of health may be used to describe populations.

The second intuition is fundamental, and reflects the influence of the ICF: at bottom health is a matter of the operation of body functions and structures, simple and complex. It may be that we need to posit central human goals and values in order to characterize the state of 'good health' (e.g., as Nordenfelt, 1987 argues). And we may need a broader notion of health that includes the performance of highly complex and social-constructed activities such as are require for community life, education, or employment (Daniels, 2007). But, as at the end of the day if health is not rooted in functions and structures of the body and mind, then the notion arguably lacks scientific or practical usefulness. This, indeed, is the underlying implication of the ICF model, and in this sense the ICF can be viewed as providing an operationalization of the concept of health, if not in general, then at least for scientific purposes (see Cieza et al., 2008).

The ICF Classifications and Health Information

The ICF is more than its conceptual framework, of course. Essentially, it is an epidemiology standard instrument, a classification. Certainly, WHO's own interest in developing and promulgating the ICF is to fulfill its constitutional mandate to provide the world with comparable health information. WHO's more well-known information standard classification, International Statistical Classification of Diseases and Related Health Problems (ICD-10) (WHO, 1992), currently under revision, is used worldwide to report causes of death for mortality statistics. Though essential for measuring life expectancy, ICD data did not capture the overall health status of living populations. Missing was information about nonfatal health outcomes, that is levels of functioning and disability across all areas of life. The ICF was therefore developed in order for standardized information about living with a state health, or as it is sometimes put, the lived experience of health. This information is proving to be essential as studies indicate that diagnosis alone does not predict health service needs, length of hospitalization, or level of care required. Nor is the presence of a disease or disorder an accurate predictor of receipt of disability benefits, work performance, return to work potential, or likelihood of social integration. With the decrease in significance of infectious diseases and the burgeoning impact of population aging and the increasing prevalence of chronic health conditions, including those like mental health problems that do not shorten life, it was clear that for nearly all epidemiological, health systems and clinical care research, what is required was information about functioning and disability.

As a practical tool, therefore, the ICF is designed to capture the information needed, at the individual and population level, to fully comprehend the experience of health, including the ramifications for functioning and disability. To accomplish this, the ICF is composed of three classifications: a classification of body functions and structures, a classification of domains of activity and participation, and a classification of environmental factors. The 'words' of the ICF standard language are called ICF categories. The ICF contains a total of 1454 ICF categories: 493 body functions, 310 body structures, 393 activities and participation domains, and 258 environmental factors. These categories are hierarchically organized in chapters with progressively detailed specification. ICF chapters represent the first level of the hierarchy and the first frame of reference for capturing health information. Each chapter is then decomposed into more detailed second-level categories, which in turn are decomposed into third- and fourth-level categories. For example, the following is the structure of the first chapter of Body Functions:

'b1	Mental functions'	(first/chapter level)
'b114	Orientation functions'	(second level)
'b1142	Orientation to person'	(third level)
'b11420	Orientation to self'	(fourth level).

The overall structure of the components of the ICF classification is represented in **Figure 3**.

To capture the severity of an impairment, limitation, or restriction, the ICF includes a general or generic qualifier with a simple rating scale of five response options: 0 = no impairment (limitation or restriction), 1 = mild, 2 = moderate, 3 = severe, and 4 = complete impairment (limitation or restriction). In the case of environmental factors, the qualifier captures the extent to which an environmental factor is a facilitator or a barrier, with a total of nine response options ranging from +4 (complete facilitator) to −4 (complete barrier), with a 0 value indicating neither a facilitator nor a barrier. Although strictly speaking, the ICF is not an assessment tool but a descriptive classification – a classification of 'what to measure' not a guide to 'how to measure' – the use of qualifiers provider, a rough guide to the extent or severity of category being used. In practice, existing assessment instruments or patient-reported outcomes can be mapped onto the values of the ICF categorical coding so that information collected by these instruments can then be compared by means of the standardization that the ICF provides.

Given WHO's motivation for developing the ICF, most of the current uses of ICF are geared to serve the requirements

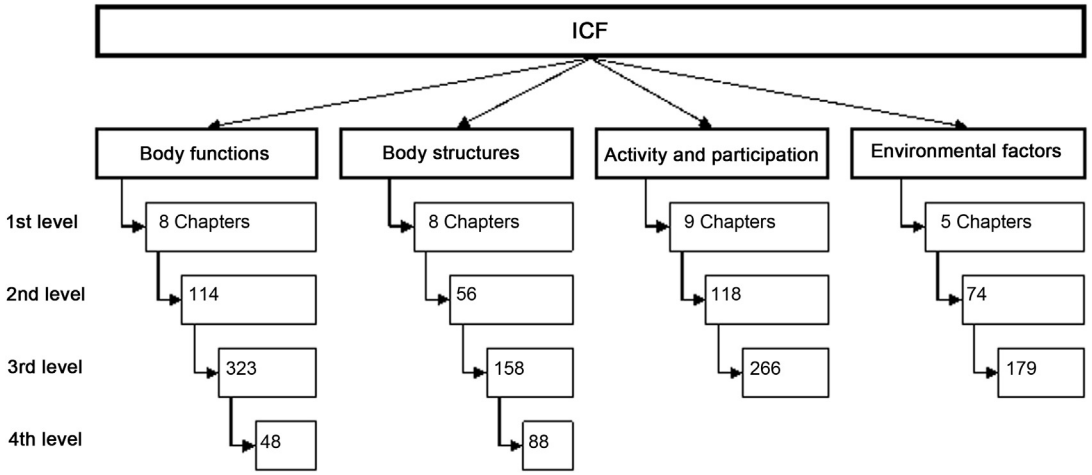

Figure 3 The structure of the ICF.

of epidemiology and health information systems. Given the enormous number of health and functioning assessment tools – both generic and health condition-specific – that are currently in use (McDowell, 2006), the ICF has become an essential comparative standard (a kind of Rosetta stone) for linking information collected using different instruments and in different health and rehabilitation settings. At the highest population level, ICF can be similarly used to compare data collected from population health or disability surveys at the national level, for, ultimately, more valid and reliable data on prevalence and incidence. Given the huge range of operationalization of disability used in such surveys around the globe, the ICF is only available standard language for international comparable statistics (WHO, 2011, Chapter 2).

At the level of clinical practice, the ICF is increasingly being used as a reference framework and classification for the development of clinical assessment and measurement instruments (Rauch et al., 2008). Further steps toward a wider clinical application of ICF is the development of condition-specific and generic 'core sets' of ICF categories that identify those categories that are necessary and sufficient for clinical description of condition and treatment setting (see, e.g., Biering-Sorensen et al., 2006), as well as 'linking rules' that enable a direct translation between existing clinical terms and ICF categories (Cieza et al., 2005). Although there is still work to be done to fulfill the measurement potential of ICF, which in part depends on the development of interval scales for particular ICF categories, based on existing clinical test batteries and patient-orient instruments, significant progress is being made in this direction (Stucki et al., 2008). Finally, ICF has also proven to be a useful basis for patient education programming (Neubert et al., 2011).

Applications of ICF to the Social and Behavioral Sciences

Because of WHO's original motivation for developing the ICF, and its health information applications, most of the applications of the ICF found in the literature focus on population health and clinical information applications, including assessment instrumentation, principally in fields of rehabilitation (see Mpofu and Oakland, 2010). Limit work has been done in the application of ICF in psychology (Peterson, 2011), but the full potential value of the ICF in this area has yet to be surveyed. Given that most of the literature on the 'models of disability' has been firmly rooted in central sociological and social psychological traditions seeking to theorize deviance, stigma, and minority status, it is somewhat surprising that the ICF has generally been received with indifference. It is undoubtedly true that, given its providence, the ICF can easily be dismissed as a classification by and for health scientists and epidemiologists. But that would be unfortunate since there are at least three themes embedded in the ICF that have significance far outside the health sphere, and would benefit from closer scrutiny by social scientists.

Understanding Environmental Factors

Environmental factors in the ICF are determinants of disabilities, coequal in significance with health conditions. Two individuals with precisely the same health condition, and impairments, may experience vastly different restrictions in the lives because of features of their physical and social environment: one may be able to be fully employed because of workplace accommodations; the other may face such significant social stigma that he or she is housebound. The presence of facilitators and barriers shape individual lives and the experience of disability. Yet, other than the easy cases of the provision of assistive technology and accommodations, or social exclusion and discrimination, we know very little about how environmental factors interact with health states to yield decrements in functioning. The ICF provides the vocabulary for describing these factors, but the science and instrumentation needed to assess and measure this impact is, at best, in its infancy.

To be sure, occupational and physical therapists have long appreciated the need for exact description of environmental impact since many of their interventions depend on it. Yet, while their knowledge about the direct or immediate physical

environment of assistive technology and physical accommodations may be well developed, the impact, positive or negative, of other domains of the environment are far less so. Social attitudes, both cultural and interpersonal, clearly affect how health conditions and impairments impact people's lives, but the mechanisms underlying this impact are not at all well-known. We are even more ignorant of how macrolevel environmental changes in the form of laws and regulations, policies, and programs impact people's lives. While political activism may successfully alter our laws and policies, astonishingly, we lack the tools and indicators for identifying, let alone measuring, the impact of policy change on the lived experience of disability.

Universalism

Although it did not originate with the ICF, universalism is at the heart of the paradigm shift in our understanding of disability that the ICF offers. Conceptually, universalism stands in contrast to strains of cultural or social relativism that suggests that the disability experience is incommensurable between members of different cultures, or indeed between those with and those without disabilities. Politically, universalism rejects the claim that people with disabilities are a discreet and insular social minority group – a position that, as a civil rights movement, the disability rights movement successfully used to highlight inequities in the treatment of persons with disabilities, including marginalization and discrimination. Universalism does not downplay or ignore these inequities; indeed it heightens the scrutiny by emphasizing the need to socially 'mainstream' disability across policy rather than emphasis differences and 'special' provisions.

The sociological and political implications of treating disability as a universal human situation, and more practically of using ICF broadly as a policy tool for employment, education, and other social policies, have not been adequately explored. Although the planning and architectural program of 'universal design' is a dominant force in disability advocacy, the prospect of transforming universalism as an abstraction into concrete policy recommendations requires a solid empirical foundation that can only be provided by sound sociological research.

Human Rights and Justice

The ICF is, or at least was intended to be, an apolitical, even atheoretical, tool for collecting and analyzing comparable health and disability data. But there was also an implicit message in the ICF model that is politically significant: the administrative, political, and social world needs to be designed for all people, whatever their functioning characteristics. Political equality demands no less than the creation of social institutions that maximize opportunities for all, given the manifold ways that people differ. Differences that compromise equality – either because that limit the intrinsic functional capacity of a person, or because the physical and social environment creates unnecessary and unfair and burdens for the individual – need to be accommodated for the sake of equality of concern and respect.

This intuition, supported by the conceptual framework of ICF, is often operationalized in the language of rights – legal, civil and more generally, human rights. The United Nations' Convention on the Rights of Persons with Disabilities (UN, 2006) is the most recent statement of human rights in which the universalistic message of equal concern and respect for persons with disabilities has been enunciated. It is noteworthy that, although the Convention does not explicitly define disability, it characterizes it in language that is exactly parallel to the ICF: "Persons with disabilities include those who have long-term physical, mental, intellectual or sensory impairments which in interaction with various barriers may hinder their full and effective participation in society on an equal basis with others" (UN, 2006, Article 3). As with any lofty statement of human rights, the Convention will remain mere words unless given effect and implemented into concrete policies at the national level. Increasingly, human rights advocates have seen the need to marshal the resources of the social sciences in order to provide the robust, empirical evidence required to give effect to the aspirations of human rights treaties, such as the Convention. Indeed, this Convention explicitly requires states to "undertake to collect appropriate information, including statistical and research data, to enable them to formulate and implement policies to give effect to the present Convention" 2 7 (UN, 2006, Article 31). Social science in service of human rights, and aligned with the model of the ICF, is clearly anticipated by this Convention as a human right that persons with disabilities also enjoy.

Conclusions

The World Health Organizations ICF is a multifaceted document: It is a classificatory instrument designed to ensure comparable international information about health and disability to serve the informational needs across the health sciences, from public health and epidemiology to clinical care. The ICF offers the prospect of operationalizing the contentious notion of health in a way that maximizes its usefulness across professions, disciplines, and social sectors. The ICF is also a structural framework for health and disability measurement, the basis for assessment instruments, both specific and generic, and a common language that different members of multidisciplinary teams can used to serve the common interests of their patients. The ICF also offers the prospect of a universal framework for comprehensive disability policy development, including but extending far beyond the scope of health. Both conceptually and practically, the ICF opens the door to a vast range of applications and developments that are of great potential significance to the social and behavioral scientist.

See also: Culture as a Determinant of Mental Health; Disability: Social and Psychological Aspects; Health Outcomes, Assessment of; Health and Illness: Mental Representations in Different Cultures; Health: Definitions; Illness and Dyadic Coping; Resilience.

Bibliography

Bickenbach, J.E., 2012. Ethics, Law, and Policy. In: The SAGE Reference Series on Disability: Key Issues and Future Directions, vol. 4. Sage Publications, Thousand Oaks, CA.

Biering-Sorensen, F., Scheuringer, M., Baumberger, M., Charlifue, S.W., Post, M.W., et al., 2006. Developing core sets for persons with spinal cord injuries based on the international classification of functioning, disability and health as a way to specify functioning. Spinal Cord 44, 541–546.

Boorse, Christopher, 1975. On the distinction between disease and illness. Philosophy and Public Affairs 5 (1), 49–68.

Boorse, Christopher, 1977. Health as a theoretical concept source. Philosophy of Science 44 (4), 542–573.

Calahan, Daniel, 1973. The WHO definition of 'health'. The Hastings Center Studies 1 (3), 77–87.

Cieza, A., Bickenbach, J., Chatterji, S., 2008. The ICF as a conceptual platform to specify and discuss health and health-related concepts. Gesundheitswesen 70, e47–56.

Cieza, A., Geyh, S., Chatterji, S., et al., 2005. ICF linking rules: an update based on lessons learned. Journal of Rehabilitation Medicine 37, 212–218.

Cieza, A., Stucki, G., 2008. The international classification of functioning, disability and health: its development process and content validity. European Journal of Physical and Rehabilitation Medicine 44, 303–313.

Daniels, N., 2007. Just Health. Cambridge University Press, New York.

Engel, G.L., 1977. The need for a new medical model: a challenge for biomedicine. Science 196 (4286), 129–136.

McDowell, Ian, 2006. Measuring Health: A Guide to Rating Scales and Questionnaires, third ed. Oxford University Press, New York.

Mpofu, Eilias, Oakland, Thomas (Eds.), 2010. Rehabilitaiton and Health Assessment: Applying ICF Guidelines. Springer Publishing Company, New York.

Neubert, S., Savariego, C., Stier-Jarmer, M., Cieza, A., 2011. Development of an ICF-based patient education program. Patient Education and Counseling 84 (2), e13–e17.

Nordenfelt, L., 1987. On the Nature of Health: An Action-theoretical Approach. Reidel, Dordrecht.

Peterson, David B., 2011. Psychological Aspects of Functioning, Disability and Health. Springer Publishing Company, New York.

Rauch, A., Cieza, A., Stucki, G., 2008. How to apply the international classification of Functioning, disability and health (ICF) for rehabilitation management in clinical practice. European Journal of Physical and Rehabilative Medicine 44, 329–342.

Salomon, Joshua A., Mathers, Colin D., Chatterji, Somnath, Sadana, Ritu, Üstün, T.Bedirhan, Murray, Christopher J.L., 2003. Quantifying individual levels of health: definitions, concepts, and measurement issues. In: Murray, Christopher J.L., Evans David, B. (Eds.), Health Systems Performance Assessment Debates, Methods and Empiricism. WHO, Geneva, pp. 301–318.

Shakespeare, T., 2006. Disability Rights and Wrongs. Routledge, London.

Stucki, G., Kostanjsek, N., Üstün, B., Cieza, A., 2008. ICF-based classification and measurement of functioning. European Journal of Physical and Rehabilitation Medicine 44, 315–328.

UN, 2006. Convention on the Rights of Persons with Disabilities. G.A. Res. 61/106. http://www.un.org/esa/socdev/enable/rights/convtexte.htm (accessed 15.01.14.).

Whitehead, M., Dahlgren, G., Gilson, L., 2001. Developing the policy response to inequities in health: a global perspective. In: Evans, T., Whitehead, M., Sisweixhawn, F., Bhuiya, A., Wirth, M. (Eds.), Challenging Inequities in Health Care: From Ethics to Action. Oxford University Press, New York, pp. 309–322.

WHO, 1992. International Statistical Classification of Diseases and Related Health Problems, 10th revision. World Health Organization, Geneva.

WHO, 2001. The International Classification of Functioning, Disability and Health. World Health Organization, Geneva.

WHO and the World Bank, 2011. World Report on Disability. WHO, Geneva.

Zola, I.K., 1989. Toward the necessary universalizing of a disability policy. Milbank Quarterly 67 (Suppl. 2, Pt. 2), 401–428.

Funding of Social Science, History of

Thomas König, Vienna Science and Technology Fund, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The history of funding in the social and behavioral sciences can best be described along different types of funding regimes. Starting with defining the term 'funding,' the article sets out to discuss those regimes and their emergence over the past two hundred years. It concludes with some remarks on recent discussions on the relationship of funding and research in the social and behavioral sciences, as well as some recent trends in their funding.

Introduction

To state the obvious right away: "Science costs money" (Stephan, 2012: p. 1). As in other fields, funding has been crucial for social and behavioral sciences from the very beginning. Surprisingly, however, so far its undeniable impact on the "social stratification of science" (Merton, 1988) has not induced comprehensive attempts to investigate the relationship of the disciplines toward funding. Reflection on the topic is restricted on case studies, which will be briefly discussed later.

Why are the social and behavioral sciences so timid when it comes to their endowment? One reason might be that there is a commonly shared uneasiness – to talk about funding might expose the disciplines as being compromised or manipulated. Any historical assessment of the social and behavioral sciences, however, cannot rely on the assumption "that there can be any pure form of social knowledge, uncontaminated by the situation in which it is created"; rather, it should be situated "in the context of the longer-term historical development of the relation between knowledge production and sociopolitical institutions" (Wagner, 2003: p. 546). The following sketch, as fragmented as it may be, is also an attempt to contribute to a history under such assumption.

Definition and Restrictions

A working definition of funding in the social and behavioral sciences encompasses the allocation of resources to allow trained personnel, for a time, conducting research on topics with relevance to social and behavioral sciences, and/or teaching those topics for training purposes. It would also include accompanying activities, such as advising and disseminating the results of the research, and fulfilling administrative requirements. Given this broad definition, it is tempting to sketch a simple, straight forward history of funding based on numbers: how many resources (valued in a standardized, aggregated currency) were allocated to the social and behavioral sciences? If it were possible to answer this question, it would provide a comprehensive overview about their relative worth, not only at a given time in a certain place, but also across regions and nations; in historical perspective, it would also allow to see allocation fluctuations over time. Of course, such a history cannot be written, at least not for now. First and foremost, the data are simply lacking. Current OECD data sets, for example, do not reach further back than to the early 1980s, since the Frascati Handbook, which is the internationally binding categorization system for science and technology statistics, included the social sciences only from 1976 (Godin, 2005a: p. 73).

Even for this rather short period, the figures displayed in Table 1 should be regarded with caution. First, the four countries were selected simply because they are the only ones where numbers are detailed along scientific fields, including the social sciences. Second, in the original table, every single data point contained a remark about specific methodological problems. Furthermore, it is not clear what 'all costs' really include, because the data collecting relies on national administrations and their standards are usually varying. "Statistics on science

Table 1 Expenditures in natural sciences and engineering and social sciences in four selected countries (in percent of total expenditures)

Australia	*1981*	*1986*	*1992*	*1996*	*2002*	*2006*
Natural sciences and engineering	69.64	69.55	68.52	72.72	73.22	73.94
Social sciences	NA	18.80	21.86	19.33	20.59	20.00
Germany	*1981*	*1985*	*1991*	*1996*	*2001*	*2006*
Natural sciences and engineering	75.53	76.10	78.08	77.96	78.23	NA
Social sciences	7.92	7.73	7.39	8.46	8.62	NA
Sweden	*1981*	*1985*	*1991*	*1997*	*2001*	*2006*
Natural sciences and engineering	85.48	85.59	83.62	80.85	79.37	80.51
Social sciences	9.31	9.18	10.86	11.10	12.51	12.07
The USA	*1981*	*1985*	*1991*	*1996*	*2001*	*2006*
Natural sciences and engineering	89.68	88.73	90.93	90.93	91.08	NA
Social sciences	6.96	5.25	5.68	6.23	5.98	NA

Sector of performance: higher education; units for expenditure: million 2005 dollars – constant prices and Prices and Purchasing Power Parties: type of costs: total (all types of costs).
Own calculation, based on OECD, 2014. Other National R-D Expenditure by Field of Science and by Type of Cost (Database). OECD Science, Technology and R&D Statistics, OECD.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03014-2

and innovation remain marked by significant methodological limitations" (Godin, 2005b: p. 15). That being said, thinking about the history of the social and behavioral sciences from the perspective of their funding might still yield some interesting, if not provocative insights.

At this point, it is helpful to distinguish also between different types of funding. The first type concerns direct, regular expenses, such as annual budgets for outfitting social science departments at universities, etc. The second type concerns direct, extraordinary funds, such as grants for social and behavioral research projects, or onetime investments to build up new infrastructure for teaching and research. The third type concerns indirect (mostly), regular expenses, including costs incurred through providing places at higher-education institutions for studying social sciences at different education levels. If the topic of funding has sparked interest at all among historians of social sciences, they have dealt with issues related to funding type number two. This is understandable, since this type of funding usually is the one most explicitly stated in historical records, as it is often the center point of concern in public discourses. For the remainder of this contribution, too, the main focus will be on the this type. For all its advantages that make it prone to historical analysis, however, it should be mentioned that, in relation to the two other types of funding, direct, extraordinary funding has to be considered rather small in size. If one were to discuss the public relevance of social and behavioral sciences in general, therefore, one would have to aim at an all-encompassing perspective.

Funding Regimes

From the point of their funding, the history of the social and behavioral sciences is to be written not only as one of remarkable growth in size. Over the past two hundred years, it also witnesses an increasing number of different procedures of allocating and accounting of funds. Those procedures can be described as funding regimes: As such, each of them implies a set of certain political purposes and addresses specific scientific practices, even though this may not be openly declared. Each regime's purpose is founded on ethical principles; the addressed scientific practices express specific epistemic expectations. Over time, more and more funding regimes evolve (with older ones usually being continued, albeit often significantly modified), offering more and more different sorts of funding and contributing to today's dense fabric of the social and behavioral sciences. Broadly, then, the history of funding can be described along different regimes of funding, each of which consisted of a unique mixture of ethical principles and the epistemic expectations.

Initially, there were basically two funding regimes: direct allocation of funds within state bureaucracies and philanthropic activities. Since the emergence of the social sciences is closely related to the rise of the modern state (Wittrock, 2000), the direct application of statistics became a crucial feature of the then still nascent state bureaucracy. Its purpose was to describe numerically the populace of a given territory, and it relied on the epistemic expectation to describe, rather than to analyze, the totality of the state's subjects (Desrosières, 1998). Census bureaus and statistical offices were usually directly connected to state apparatuses. In the beginning, at least, funding for this kind of activity was not yet on a regular basis.

Complementary, in all industrializing countries during the first half of the nineteenth century, individuals with a philanthropic background began addressing the need to respond to pressing social issues. Their curiosity was predominantly satisfied through social surveys, and it was clearly stimulated by political interest. For Friedrich Engels, "comprehending the present state of the English proletariat" was synonymous with his political sympathies for socialism (Engels, 1892). Others based their ambitions "to engineer the scientific management of the poor" on their religious beliefs, perceiving "disciplined ordering of the self and of the social world as the only possible bulwarks against the ever-surging sea of sin" (Yeo, 1991: p. 51). Soon enough, however, surveying also reached the policy arena: in England, Royal Commissions were established in order to investigate a range of issues, from the condition in coal mines to the traffic problems in London City. Here, surveying was no more the concern of affluent individuals but was intended to provide empirical evidence for public policy, particularly in its regulatory form.

When other forms of early funding emerged, they were based on a different set of core assumptions. As a sort of transitory vehicle, the efforts of the German Verein fur Socialpolitik contained the political interests of its members, but those were nonetheless intertwined with the social political program of the German Reich (Stremmel et al., 2006). As the social and behavioral sciences made their way into universities, they gained in autonomy (Shils, 1992); the relation between ethical principles underpinning the political purpose of funding changed, as did the epistemic expectations, but the strong connections were not lost. Nor did the older funding regimes lose traction. The statistical offices remained important centers of knowledge production and politically relevant, even though their practices were no longer considered the vanguards of research. With others, a division of labor was taking place. The philanthropic individual conducting research on his own (since its exponents were predominantly male) in the nineteenth century was replaced by the philanthropic foundation (based on the wealth of a philanthropic individual, nonetheless) that gave grants and fellowships to scholars (Bremner, 1988; Lingelbach, 2011). The financial means of the foundations were often enormous; their 'scientific philanthropy' relied on the idea that the "development of fair, objective, expert knowledge would enable all Americans [and eventually citizens of other nations, TK] voluntarily to coordinate and rationalize their actions" (Lagemann, 1992: p. 29).

As the last quote already indicates, the foundations were partly responsible for a major spatial shift of the center of social and behavioral sciences. From the 1920s on, that center was in Northern America, and no longer in Europe. In particular, the American foundations were critical in establishing what could be called 'big science' in the social and behavioral sciences through funding large projects (at least in comparison to other budgets in those disciplines), and in establishing new research institutions, within the US as outside (e.g., Hauptmann, 2012). Not surprisingly, thus, the processes of disciplining and professionalizing were strongly enforced by the foundations' funding activities (e.g., Capshew, 1999).

The US were also the first to see the advent of yet two other funding regimes that emerged during the first half of the twentieth century: independent funding agencies and research councils, and research commissioned by governmental institutions to (formally independent) research institutions. Before World War II, funding agencies and research councils relied mostly on privately raised money; only after the war, they were endowed by public funds. Despite the considerable differences that those funding agencies had in their national setting of science policy (Braun, 1997), they all were based on the idea that spending taxpayers' money through independent intermediary organizations would foster technological innovation and thus eventually stimulate economic growth.

The funding agencies established two important features of research in the social and behavioral sciences: The first feature was that research practices could be distinguished into basic and applied research. As they were supporting the former, the main purpose of funding agencies would become known as "policy for science" (Pulparampil, 1978: pp. 43–45). The second feature was that research should be done primarily in the form of projects. The social and behavioral sciences were often at the fringes of agencies' funding portfolio, since their main focus was on the natural sciences (although in the US and in the UK, research councils dedicated to the social and behavioral sciences also emerged). Still, agency funding probably had important effect on the developments particularly of psychological and economics research that successfully claimed to fulfill the epistemic expectation of scientific projects.

At the same time, the political agenda of governments was diversifying, and the 'applied' side of funding was taking off, too. Ministries and executive agencies commissioned research projects, either within their own administrative bodies, or to independent research institutions; sometimes they also supported entire institutions through block grants in order to facilitate long-term research on a certain subject. The most powerful funder became the U.S. Department of Defense, including its subbranches, such as the Air Force, which founded the RAND Corporation and its system analysis; other examples were intelligence analysis and counterinsurgency (Project Camelot). Yet another state-funded activity that became widespread in that period concerned academic exchange – not only a cheaper version than institution or project funding, but also differently framed under the key word of cultural diplomacy (Arndt, 2006). Commissioned research became a standard practice of governmental departments, and a major source of funding for social and behavioral sciences.

Obviously, the split between knowledge produced primarily for academia and knowledge produced primarily for political usage can be found in those two new funding regimes. Since then, no new funding regime has emerged sensu stricto. Rather, as the existing funding regimes were growing in number and size, they were increasingly subjugated under an all-encompassing metaregime of performance, with evaluation, assessment, and accounting as its basic means. In Great Britain, for example, the 'Research Assessment Exercise' was established with the hope to channel funding more effectively, essentially rewarding more productive researchers; in the US, a program called 'Star Metrics' now wants to verify how many jobs are created by grants of the National Science Foundation. In general, science policy has made a fundamental change in its focus, its principle now being 'science for policy.' Even basic research funding now must prove its impact, either in the academic world or through societal relevance.

A very popular recent categorization distinguishes research activities along two axes: whether or not the result is intended to be useful, and whether or not it aims at answering a fundamental scientific question (Stokes, 1997). From the four types of research that are constructed through this scheme, the one falling into 'Pasteur's Quadrant' is expected to be the most relevant, because it is supposedly advancing fundamental knowledge and at the same time expected to solve practical problems. Consequently, this kind of research is nowadays the most sought-for through funding schemes. For the social and behavioral sciences, this has on the one hand opened up new fields of activities – such as accompanying research, because it is supposed to help knowledge produced by research in the natural sciences to better relate to the public. On the other hand, It has increased doubts about the usefulness of research in some domains of social and behavioral sciences, as the recent debates in the US as well as in Europe bear witness (for Europe, see Mayer et al., 2013).

Controversies of Funding

The overview in the previous chapter treated the emergence of funding regimes as if they were unilaterally the same across the world. In some sense, there are indeed remarkable similarities across the globe, probably due to the internationalization of the academic world and the intrinsic power exerted by the US. That should not imply, of course, that funding regimes should be perceived as historical phenomena without links to political narratives and powers; quite the contrary. It would be interesting for more detailed historical analysis to find out the nuanced differences between similar funding modes in different regions of the world.

In historical terms, the different funding regimes (and their underlying ethical principles) have had impact on the shape of the social and behavioral sciences, at least in a very general sense. The regimes themselves came not into being without active shaping of social and behavioral scientists; one (in terms of funding involved) minor, but well-investigated case study that makes this very clear was the (failed) project of Max Weber's Presse Enquete (Weischenberg, 2012). The funding of 'big science' projects in the twentieth century would not have happened without crucial lobbying by representatives of economics, sociology, political science, psychology, and neither would have new institutions be founded. Similarly, the developments regarding the broader relationship of science and politics since the 1950s briefly outlined in the previous chapter are unthinkable without the theoretical concepts of economists and scholars in the field of science and technology studies, as they provided the basis for the narratives of science policy (Godin, 2009).

Often, the impact of funding is discussed within the social and behavioral sciences only on a polemical level. Two such controversies deserve to be looked at more closely. One concerns the (mostly historical) question whether research funding had any sort of impact on the shape of social and behavioral science disciplines; the other concerns the question

(focused primarily on the present) what the current forms of legitimation procedures of research funding decisions imply for research in those disciplines.

Has dedicated funding influenced the development of disciplines and fields in the social and behavioral sciences? Not surprisingly perhaps, this question has been posed in several disciplines in parallel and is always closely related to the issue of methodology. As is known today, there have been different phases of social research methods (Crothers and Platt, 2010). It seems natural to ask if those phases were connected to the emergence of new funding regimes. In sociology, the debate was sparked by the question whether the Rockefeller Foundation – for a long time the most affluent foundation active in funding social scientific research – had been preferred certain methodical approaches over others (Fisher, 1993). The allegation is difficult to get away with, although studies show that it is not vindicable (Platt, 1996). Similar allegations have made in political science, where the exclusion of political theory as one subfield of the discipline were supposedly the result of the work of foundations and political scientists alike (Hauptmann, 2006). Those historical debates, of course, have to be seen as one means in a broader context of redefining the disciplinary field in the present and in order to change them for the future (cf Engelmann, 2010). However, the larger issue at stake – reflecting the role of funding for social and behavioral research in face of historically shifting purposes and practices – remains underdeveloped.

Similarly, the current metaregime of performance begs the question how the shape of social and behavioral sciences is affected by it. While most scholars in the disciplines will not deny that meritocracy is an important feature in the academic world, there is often disgust displayed when the procedures established nowadays for legitimating the distribution of funding – projects and researchers are subject of many such procedures, from peer reviewing, evaluation, reporting, to accounting – are discussed. It is assumed that the research-funding apparatus established today has massive influence on language, style, scope, and behavior of researchers (Billig, 2013). Again, however, research is rather limited on that issue.

Changing Modes of Legitimacy

An important, though still underdeveloped question on the relation between funding and research activities in the social and behavioral sciences concerns the rhetoric of legitimacy of funding distribution over time. Lutz Raphael has recently suggested to divide the history of the social sciences into four configurations, in which not only the ideas and discourses and experts are taken into considerations, but also 'clients, institutions, and technologies' prevalent in each of those periods. According to Raphael, the period of 'social reform' lasted until 1910 and was superseded by the period of 'social engineering.' From the 1950s on, 'planned modernization' took place, which was replaced by 'the age of therapy,' from the 1980s on (Raphael, 2012: pp. 50–54). Broadly, those periods correspond with waves of new funding regimes as shown in the previous chapter. It is useful to look at those configurations also as a way to broadly understand the respectively accepted form of legitimacy for funding during each of those periods.

In the first configuration, when experts and clients were often close if not the same, and institutions as well as technologies not much advanced, funding required hardly any legitimation. The next period of social engineering, which saw the rise of the foundations, brought an important innovation. The large funds of the foundations required some sort of decision-making instrument. In the case of the Rockefeller Foundation, this was done by in-house and independent scholars who had gained the trust of the foundation's administrators (Fleck, 2011). Here, networks of trusted peers fulfilled the task of justifying the distribution of funds to scholars and researchers. When states entered the domain of research funding on large scale, the need to legitimate was further advanced. In the case of the agencies funding basic research, where the decision-making was basically left to (social) scientists, peer reviewing became the ultimate procedure of justification. In the case of commissioned research, the issue remained more in the dark, but it is obvious that it was left to a good part in the common interest between researchers and those commissioning the research. In the age of therapy, the metaregime of funding performance has given procedures such an important role that some of them, in particular decision-making, have become a topic of research themselves. Occasionally, a study in this field even makes it on the best seller list of social science literature (cf Lamont, 2009). Such studies bring great value to our understanding of legitimation procedures in the academia. What is missing, so far, is how those procedures influence the daily practice of scholars in the social and behavioral sciences.

See also: Academic Research and Employment: Recent Changes in Europe and the United States; Awards, Public Recognition, and Evaluation; Cold War: Impact on Social and Behavioral Sciences; Methodology of the History of the Social and Behavioral Sciences; Quantification in the History of the Social Sciences; Social Science Infrastructure: European Union (Resourcing and Funding); Social Science and Universities.

Bibliography

Arndt, R.T., 2006. The First Resort of Kings. American Cultural Diplomacy in the Twentieth Century. Dulles.

Billig, M., 2013. Learn to Write Badly: How to Succeed in the Social Sciences. Cambridge University Press, Cambridge.

Braun, D., 1997. Die politische Steuerung der Wissenschaft: Ein Beitrag zum "kooperativen Staat". Campus Verlag.

Bremner, R.H., 1988. American Philanthropy, second ed. University of Chicago Press.

Capshew, J.H., 1999. Psychologists on the March: Science, Practice, and Professional Identity in America, 1929–1969. Cambridge University Press, Cambridge; New York.

Crothers, C., Platt, J., 2010. The history and development of sociological research methods. In: Historical Developments and Theoretical Approaches in Sociology. EOLSS.

Desrosières, A., 1998. The Politics of Large Numbers: A History of Statistical Reasoning. Harvard University Press.

Engelmann, S., 2010. Theory trouble: the case of biopolitical science. Österreichische Zeitschrift für Politikwissenschaft 39, 55–71.

Engels, F., 1892. The Condition of the Working Class in England with Preface Written in 1892. Swan Sonnenschein & Co., London.

Fisher, D., 1993. Fundamental Development of the Social Sciences: Rockefeller Philanthropy and the United States Social Science Research Council. University of Michigan Press, Ann Arbor.
Fleck, C., 2011. A Transatlantic History of the Social Sciences: Robber Barons, the Third Reich and the Invention of Empirical Social Research. Bloomsbury Academic, London; New York.
Godin, B., 2005a. Measurement and Statistics on Science and Technology: 1920 to the Present. Routledge.
Godin, B., 2005b. The Who, What, Why and How of S&T Measurement (Working Paper No. 26), Project on the History and Sociology of S&T Statistics. Centre Urbanisation Culture Société, Montréal.
Godin, B., 2009. The Making of Science, Technology and Innovation Policy: Conceptual Frameworks as Narratives, 1945–2005. Centre Urbanisation Culture Société de l'INRS, Montréal.
Hauptmann, E., 2006. From opposition to accommodation: how Rockefeller Foundation grants redefined relations between political theory and social science in the 1950s. American Political Science Review 100, 643–649.
Hauptmann, E., 2012. The Ford Foundation and the rise of behavioralism in political science. Journal of the History of the Behavioral Sciences 48, 154–173.
Lagemann, E.C., 1992. The Politics of Knowledge: The Carnegie Corporation, Philanthropy, and Public Policy. University of Chicago Press, Chicago.
Lamont, M., 2009. How Professors Think: Inside the Curious World of Academic Judgment. Harvard University Press.
Lingelbach, G., 2011. Private Wissenschaftsförderung im 19. und 20. Jahrhundert. In: Kocka, J. (Ed.), Stiften, Schenken, Prägen: Zivilgesellschaftliche Wissenschaftsförderung im Wandel. Campus, Frankfurt am Main, pp. 43–57.
Mayer, K., König, T., Nowotny, H., 2013. Horizons for Social Sciences and Humanities (Conference Report). Mykolas Romeris University, Vilnius.
Merton, R.K., 1988. The Matthew effect in science, II: cumulative advantage and the symbolism of intellectual property. Isis 79, 606–623.
OECD, 2014. Other National R-D Expenditure By Field of Science and By Type of Cost (Database). OECD Science, Technology and R&D Statistics, OECD.
Platt, J., 1996. Has funding made a difference to research methods? Sociological Research Online 1.
Pulparampil, J., 1978. Science and Society: A Perspective on the Frontiers of Science Policy. Concept Pub. Co., Delhi.
Raphael, L., 2012. Embedding the human and social sciences in Western societies, 1880–1980: reflections on trends and methods of current research. In: Brückweh, K., Schumann, D., Wetzell, R.F., Ziemann, B. (Eds.), Engineering Society: The Role of the Human and Social Sciences in Modern Societies, 1880–1980. Palgrave Macmillan, pp. 41–56.
Shils, E., 1992. The universities, the social sciences, and liberal democracy. Interchange 23, 183–223. http://dx.doi.org/10.1007/BF01435232.
Stephan, P.E., 2012. How Economics Shapes Science. Harvard University Press, Cambridge, MA.
Stokes, D.E., 1997. Pasteur's Quadrant: Basic Science and Technological Innovation. Brookings Institution Press, Washington, DC.
Stremmel, R., Tennstedt, F., Fleckenstein, G., Peterle, M., Rust-Schmöle, G., Rassow, P., Born, Karl Erich, 2006. Kirchen, Parteien, Vereine und Verbände, Quellensammlung zur Geschichte der deutschen Sozialpolitik 1867 bis 1914. Wissenschaftliche Buchgesellschaft (WBG), Darmstadt.
Wagner, P., 2003. The uses of the social sciences. In: The Modern Social Sciences, The Cambridge History of Science. Cambridge University Press, Cambridge, pp. 535–552.
Weischenberg, S., 2012. Max Weber und die Entzauberung der Medienwelt: Theorien und Querelen – eine andere Fachgeschichte. VS Verlag für Sozialwissenschaften.
Wittrock, B., 2000. Modernity: one, none, or many? European origins and modernity as a global condition. Daedalus 129, 31–60.
Yeo, E.J., 1991. The social survey in social perspective, 1830–1930. In: Bulmer, M., Bales, K., Sklar, K.K. (Eds.), The Social Survey in Historical Perspective, 1880–1940. Cambridge University Press, Cambridge, pp. 49–65.

Relevant Website

https://www.starmetrics.nih.gov – Star Metrics.

Future Thinking

Stefania de Vito, University of Edinburgh, Edinburgh, UK; and Suor Orsola Benincasa University, Naples, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Thinking about the future can take numerous forms, varying from planning actions to foreseeing possible scenarios by means of knowledge and informed guesses, speculations and intuitions, or imagination and creativity. Different cognitive processes are needed for each of these different types of future thinking. The present article encompasses the relevant literature revolving around the issue of future thinking, that is, the cognitive ability, specifically of human, of envisaging one's own future. It aims at offering different interpretations about the relations between future thinking and a widespread range of other cognitive functions, like memory, executive functioning, and subjective time.

What actually occurs when we anticipate our personal future in thought is wonderfully expressed in the story of Apollo and Daphne. Apollo falls madly in love with Daphne. While running after the nymph, who desperately tries to flee from him, the god brags about being Jupiter's son and hence about his formidable power to clearly see the present, the past, and the future. However, Apollo, aspiring to Daphne's bed, hopes what he seeks and his hope becomes a misleading oracle. In fact, she begs her father's help to escape from the oracular god and obtains to be changed into a laurel tree (Ovid, "Metamorphoses," I book). More accurately than could any neuroscientific theory, the old myth shows what happens when people imagine their future. Even prophetic deities do regularly get it wrong. Nonetheless, we frequently act upon our fallacious predictions. Well aware that anticipations cannot be reliable, the Cumaean Sibyl paltered with a soldier, who consulted her about his fate in an upcoming battle, in a double sense: "*Ibis redibis non morieris in bello*," which sounds intentionally ambiguous unless one puts a comma before or after the word '*non*' (The sibylline statement might translate as follows: *You will go, will come back, will not die in the war*. In Latin, the word "*non*" could be yoked to "*redibis*" (i.e., You will go, will not come back, will die in the war), or with "*morieris*" (i.e., You will go, will come back, will not die in the war)) (de Vito and Della Sala, 2011). The present article aims at analyzing why our forecasts cannot be more precise than a sibylline prophecy without commas.

Conceptual Development and Definitions

Humans have the omnipresent faculty to think about their future. They do so by projecting themselves forward in time in richly detailed scenes and by imagining how different variations of an event will unfold. This ability is ubiquitous, given that, far from being idle in the absence of external task direction, our brain is spontaneously and constantly (also in the default – passive – cognitive state) engaged in devising simulations that might approximate real future events (Ingvar, 1979; Bar, 2007; Buckner and Carroll, 2007; Buckner et al., 2008). Thus, people might either deliberately cast themselves into hypothetical futures or unexpectedly think of their future during the unoccupied moments of daily life. According to Cicero, Cato used to say that "never is a man more active than when he does nothing, never is he less alone than when he is by himself."

This complex process reflects a composite of cognitive functions. An ongoing research aims at characterizing these components and assessing their specific roles. The gist of the debate mainly revolves around episodic/non-episodic contributions to future thinking.

Initially, researchers have emphasized the importance of the episodic contribution, considering episodic memory as a fundamental part of future thinking, since it provides the 'raw material' (Suddendorf, 2010: p. 100) to envisage situations likely to occur in the future.

The preferential focus on episodic memory, rather than on other memory systems, such as semantic memory, seems to stem from the unique relationship that episodic memory bears with time, a relationship of which semantic memory system is devoid of (Tulving, 2002). Episodic memory, indeed, differs from semantic memory in the diverse typologies and sources of to-be-remembered information. That is, individual experiences in the former case and general knowledge of the world in the latter (Tulving, 1972). Thus, only episodic memory has the special connection with time, which enables "people to consciously re-experience past experiences" (Tulving, 2002: p. 6).

The relations between the past and the future have been variously identified as "continuation of a repeating past" and "projection of past trends" (Haith, 1997): people are inclined to predict a reoccurrence in the future of what happened in the past (respectively, *verbatim* or through a detail recombination). The second type of relation is close to Tolman's "theory of expectancy" that illustrated how rats might flexibly recombine prior experiences to deal with *ex novo* situations. The use of memory may be instrumental in the development of adaptive responses that extend beyond the stereotyped recurrence of reinforced sequences (Tolman et al., 1946). However, there is another possibility that Haith (1997) labeled "imagination and invention," which contemplates the imagination of novel episodes that have never happened, like a hypothetical trip on Mars. In this case, "episodic memory (together with other memory systems) may hence provide the vocabulary for episodic foresight. Humans can recursively combine and recombine basic elements into novel scenarios and evaluate these in terms of their likelihood, desirability and so forth" (Suddendorf, 2010: p. 101).

Subsequent considerations concerning the ability of foreseeing by reshuffling different snippets to shape novel arrangements have led to the hypothesis that also processes

unrelated to episodic memory play a crucial role when imagining the future. Therefore, in the attempt of gaining greater control over the simulation of scenes, a more recent branch of research has been addressing the non-episodic contributions to future thinking (e.g., Rosenbaum et al., 2009; Gaesser et al., 2011; de Vito et al., 2012). Instead of considering episodic memory as a "vocabulary" of elements on which to draw to foresee, it has been argued that episodic memory and future thinking share the same neurocognitive resources. They might draw on a widespread range of specific abilities, like a particular type of consciousness (autonoetic consciousness; Tulving, 1985), the ability of binding details, executive functions, self-projection, and related capacities, like conceiving others' viewpoint (theory of mind) (Buckner and Carroll, 2007). That is, a number of cognitive functions other than episodic memory, unconnected with both recent and remote autobiographical episodes, might play an integral role when constructing scenes.

Szpunar (2010) noticed that even the heterogeneous nomenclature adopted to identify forethought depends on whether or not it was meant to highlight the tight linkage with episodic memory. Atance and O'Neill (2001) have built their definition of "episodic future thinking" upon Tulving's most recent conceptualization of episodic memory. Indeed, the two researchers formulated a specular distinction between semantic and episodic future thinking, which mirrors Tulving's categorization for past memories. Whereas semantic future thinking concerns the projection into the future of a general knowledge of the world and, consequently, of reasonably good scripts, episodic future thinking "refers to an ability to project the self forward in time to pre-experience an event" (p. 537). In-line with Tulving's theory, they maintained that the episodic memory system underlies the capacity to cast personal events into the future. A key account for this linkage has been provided by Suddendorf and Corballis (1997), who referred to the ability to mentally travel in time as "a prime mover in human cognitive evolution" (p. 165). For our survival, episodic memory informs us about potential futures (Suddendorf, 2010). The "constructive simulation hypothesis" arises from this evolutionary perspective and states that the mechanism lying behind the episodic memory evolved as constructive rather than reproductive to allow individual simulations. People can collect elements from different memories and combine them in an unpredicted manner to shape unique scenarios (Schacter and Addis, 2007). In order to plan the future, we need a system of representations rather than mere recordings of events. Thus, according to this view, memories are considered as "building blocks of predictions" (Bar, 2007: p. 280). That is, the linked stored elements activate associations that are relevant in a specific situation. Predictions are generated by these associations, which is why they so often go awry.

On the contrary, when researchers did not intend to emphasize the connection with episodic memory, referred to the capacity of formulating prospects as "projection" (Okuda et al., 2003) or "prospection" (Gilbert, 2006).

Although episodic future thinking is a unique mental faculty, originally, the concept has been extensively discussed in relation to episodic memory. To this discussion, we will turn next.

Episodic Contribution to Future Thought

It has long been posited that an intimate relation underpins episodic past and future thoughts. Immanuel Kant had this insight in 1798: "Recalling the past (remembering) occurs only with the intention of making it possible to foresee the future; we look about us from the standpoint of the present in order to determine something, or to be prepared for something. Empirical foreseeing is the anticipation of similar cases (exspectatio casuum similium) and requires no knowledge of causes and effects, but only the remembering of observed events, as they usually follow upon each other. Repeated experiences help to develop skill in empirical anticipation" (1996: p. 77). Past memories are thought to play an essential role in driving future simulations to such a point that when no parallel in the past is found to what is occurring, people get lost: "as the past has ceased to throw its light upon the future, the mind of man wanders in obscurity" (de Tocqueville, 1863: p. 408). The recurrent observation that "we have no grasp of the future without an equal and corresponding outlook over the past" (Bergson, 1913: pp. 69–70) led William Faulkner to state that "the past is never dead, it is not even past" (1951, Act 1, sc. 3), and René Char to complain about the lack of a "testament" that could show us how to use our "inheritance." Consistent with these precocious considerations, functional magnetic resonance imaging (fMRI) studies have shown that remembering and imagining the future engage a common core network, encompassing specifically the medial prefrontal cortex, posterior regions in the medial and lateral parietal cortex, the lateral temporal cortex, and the medial temporal lobe, including the hippocampus (Okuda et al., 2003; Addis et al., 2007; Szpunar et al., 2007). Future episodes simulation has been shown to activate the hippocampus, a region that is traditionally considered to play a fundamental role in memory encoding. In particular, preferential right hippocampal engagement has been observed when imagining the future relative to remembering past. There has been an increasing interest in interpreting the specific contribution that hippocampus makes to prospects and, consequently, in understanding the relation between remembering and future thinking. Critical lines of inquiry are to determine whether or not the hippocampus plays a crucial role in future thought, and which processes rely on the hippocampal activation.

Some amnesic patients are reported to be prisoners of a permanent present (O.S., Lidz, 1942; G.A., Dalla Barba et al., 1997; D.B., Klein et al., 2002). Five patients with bilateral hippocampal damage, encouraged by Hassabis et al. (2007) to generate "as many sensory and introspective details as they could" about a fictitious event, depicted very poor scenarios lacking in spatial coherence. Likewise, a group of patients suffering from mild Alzheimer's disease (Addis et al., 2009), a group of patients with amnestic mild cognitive impairment (Gamboz et al., 2010), and patients with medial temporal lobe damage (Race et al., 2011) uttered fewer details in a future thinking task than their matched controls. When asked about personal plans in the next future, most of these patients answered in the same way. At the question, "What are you going to do tomorrow?", the famous amnesic patient H.M. replied, "Whatever is beneficial" (Corkin, 2002); K.C., who suffered from a dense amnesia following frontal and medial lobes damage,

smiled faintly, then answered, "I don't know," and, when required to provide further explanations, said, "It's like swimming in the middle of a lake. There's nothing there to hold up or do anything with" (Tulving, 1985: p. 4); M.C., with bilateral hippocampal damage, replied, "I don't see past, today and tomorrow, and can't picture myself in anything beyond the immediate present" (Andelman et al., 2010: p. 432). Nonetheless, the role played by episodic memory in the construction of future prospects is still controversial.

The Debated Role of Episodic Contribution in Foresight

There is a substantial amount of similarities between remembering episodic memories and imagining future events. However, these similarities are to be treated with caution. Recently, it has been suggested that episodic memory may not be a *condicio sine qua non* in simulations of future scenarios. Such an intuition has also been corroborated by several criticisms to the studies mentioned above and consequent research. Some of the studies mentioned earlier do not detail the lesion location of the amnesiacs (O.S. and D.B.). In other cases (K.C. and Hassabis and colleagues' patients), the information provided indicates an extensive damage, including impairments encroaching upon areas beyond the hippocampus, as Squire et al. (2011) argued. Furthermore, a crucial caveat to the hypothesis that the two time-related functions might be closely interconnected is that a number of amnesiacs show preserved ability to perform constructive tasks. Spared future thinking capacities in a context of amnesia and hippocampal damage have been observed in five amnesic patients whose performance fell within, or even above, the control range (Squire et al., 2010), in the single case of Jon, with developmental amnesia (Maguire et al., 2010); in a group of 21 children with a bilateral hippocampal damage (Cooper et al., 2011); and in P2, one of the patients examined by Hassabis and colleagues (2007) (for a review, see Addis and Schacter, 2012). These observations downsize the importance of episodic memory and hippocampus in future imaging. The relationship between remembering the past and imagining the future appears to be more subtle than it has been postulated in various studies. Information about specific events that have already occurred might not be necessarily needed. A loss or a reduction of episodic memory might be compensated with an increased reliance on semantic information when remembering (Levine et al., 2002), when imagining (Gamboz et al., 2010), and even when more basically describing pictures (Gaesser et al., 2011). People can easily generalize and extrapolate information from their general knowledge, adding arbitrary particularities. "Without the ability to extrapolate from generalities, the amount of particularities the brain would have to store would be subject to a combinatorial explosion. For every prediction of a future event, the memory of a corresponding past event would be needed. Conversely, what one can predict about the future will be limited by one's episodic memory" (Hegdè, 2007: p. 324). If it was strictly true that episodic memory helps to pre-experience future episodes, this would imply the possibility of foreseeing only those events that we have episodic memory of. Furthermore, if it was true that one could predict his future only on the basis of his past, and to the extent that events never replicate themselves precisely, one could never apply any past memory to a future simulation (Hegdè, 2007). Therefore, without unnecessary limitations, it can be argued that different types of memory might cooperate with each other and with other cognitive domains to aid forecasting.

When participants attempt to build personal futures in response to verbal cues, they rely upon disparate elements from a number of sources, which cannot be solely identified in episodic memory. D'Argembeau and Mathy (2011) found that when attempting to think about a possible future episode, people most frequently activate general knowledge. There is also much evidence that when adults are required to complete sentences about their personal future (e.g., "Next year I… ."), they frequently report future thoughts concerning extended future life periods or personal semantic information (Anderson and Dewhurst, 2009). And, in fact, a considerable number of future simulations consist of abstract representations rather than specific events (D'Argembeau et al., 2011). These findings raise the possibility that access to episodic details might be not necessary. In the absence of episodic memories, one might adopt compensation strategies and rely on semantic information. Indeed, latest studies using fMRI have provided no evidence of a hippocampal involvement in subjective time travel. Nyberg et al. (2010) required well-trained participants to imagine walking in a familiar environment in the past, present, and future. Thus, they were not asked to create an episode content but solely to self-project in a standard situation recurring in different temporal frames. The task, which held constant the informational content, did not recruit the hippocampal region. The authors interpreted their findings, suggesting a closer relation between the hippocampus and the 'message,' rather than between the hippocampus and the "conscious temporality of the experience" (p. 22357). Likewise, Squire et al. (2010) reported on future thinking performances in five patients with a well-characterized bilateral hippocampal lesion and one patient with extensive medial temporal damage, presenting with a subsequent, respectively, moderate and marked impairment in recent personal memories. The intact ability of these patients in future thinking sharply contrasted with their striking impairment in recalling recent past episodes. Consequently, the authors suggested that a selective hippocampal damage, which spares the rest of the core network proposed, is not sufficient to affect future imaging. Maguire and Hassabis (2011) assumed that the patients could have supplied reports more semantic in nature, prompted by the single word cue, likely to cause a semantic bias. However, Squire et al. (2011) noticed that also one of the patients (P1) described by Hassabis et al. (2007), albeit the marked amnesia, was successful at constructing scenes. And in that case, the scenes construction was elicited by full sentence descriptors. This demonstrates that amnesia per se does not necessarily affect the ability of simulating scenes.

Episodic retrieval might not be sufficient for future events to be shaped. For instance, Romero and Moscovitch (2012) suggested that the re-encoding of retrieved snippets into novel events is another aspect of event construction that might change with age and medio-temporal lobe dysfunction. In their study, indeed, the relative coherence of imagined events decreased in older adults when the number of given items to be

integrated was higher. Therefore, in addition to the access to episodic details, there are at least three more functions that are essential in allowing future event simulations: the production of specific details not necessarily episodic in nature, the actual construction of the episode itself, and the awareness of the plausibility of the imagined event.

First, obviously, the ability of future thinking is related to the capacity to produce specific personal details, mediated by executive control. These details might also be produced in conditions that do not require an involvement of episodic memory (when, for instance, participants are simply required to describe an environment or a picture).

For what concerns the construction of the episode itself, it should be considered that as long as we have a jumble of snippets, we have to reshuffle them in novel patterns. This step is even more complex in future thinking than in episodic memory. Whereas past episodes require a search for an exclusive arrangement of information, future thinking is open-ended. The number of potential scenarios that one might envisage is infinite. Future events might be shaped with any arrangement of an unlimited array of information (Suddendorf and Corballis, 2007; Anderson et al., 2012).

At this point, during and after the construction of the event, we have to evaluate whether the event is actually likely to happen in the future, considering our background, our personal history, and our past and present situations. Therefore, the feasibility of our scenarios has to be constantly monitored.

Non-Episodic Contribution to Future Thinking

Our mind does not work like a video camera and we do not have at our disposal full-fledged scenes to be performed as in a movie ("The past, like the future, can only be imagined"; De Hériz, 2006: p. 21). Episodes are constructed by drawing from different snippets and by reshuffling them to form infinite and elaborated simulations of possible upcoming episodes. The construction of such simulations implies an interaction of different functions that play crucial roles in the production and the flexible combination and integration of the amount of the available details into a coherent whole. The generation of detail-rich descriptions requires reconstructive mechanisms for the binding of information, which are, in turn, mediated by executive control. Deficits or decline of these mechanisms might hamper the generation of details and, consequently, account for impairments in a wide range of conditions. At odds with what Levine et al. (2002) and Addis et al. (2008) maintained, the reduction in specificity observed in elderly people does not entirely reflect an impairment of the constructive episodic memory system underling both past and future thoughts. And indeed, such reduction also extends to tasks that do not require an access to episodic details. Gaesser et al., (2011) documented an age-related decline in reporting details also during a condition of picture description task, which does not necessitate episodic memory. These findings uncovered a non-episodic source of age differences in past and future thinking tasks, a source unrelated to memory itself. Such observations were instrumental in the development of other studies designed to explore the possible causal link between executive functioning and future prospects, in the attempt of gaining greater control over the simulation of scenes. This new branch of research takes into account that the extensive damages exhibited by the amnesic patients described in the previous paragraph also include areas that play a crucial role in executive functions (e.g., see Samarasekera et al., 2007). Rosenbaum et al., (2009), for instance, a more general construction deficit as a plausible and alternative root of K.C.'s impairment in future simulations. Aiming at testing this possibility, K.C. was required to produce fictional episodes and to retrieve or identify elements of semantic narratives (famous fairy tales and bible stories). Also, in these conditions, the patient's scores fell well below the control range. These results speak against the episodic constructive simulation hypothesis. If future simulations must rely on the reformulation of past personal memories, then K.C. should have presented with a normal recollection of semantic narratives, albeit his severely impaired autobiographical episodic memory. However, this was not the case. Indeed, he failed to construct detailed and integrate narratives even when the necessary information was available. Similar results have been obtained by Addis et al. (2010), who documented an age-related reduction in specificity also under a recombination paradigm, simply requiring healthy older adults to reformulate various elements into a cohesive episode. In-line with these evidences, Addis et al. (2008) noticed a close linkage between tasks involving executive performances and tasks requiring episodic elements generation. In their study, the amount of episodic details that were provided in the past and future reports correlated with the Backward Digit Span task, which measures the executive control and engages the dorsolateral prefrontal cortex (Hoshi et al., 2000). Accordingly, a number of neuroimaging studies have shown a greater frontopolar activation, including BA 10, during future thoughts than during past recollections (Okuda et al., 2003; Addis et al., 2007; Summerfield et al., 2010). Interestingly, the greater was the activation of BA 10, the larger was the amount of elements provided in future constructions (Addis et al., 2008). In another fMRI study, Summerfield et al. (2010) showed which components of the core network (Schacter et al., 2007) are variably engaged and disengaged during the various steps of a scene construction task. The participants were instructed to combine the single-scene elements they were provided within a cohesive scenario. The details were presented one at a time with no context. The fMRI documented a significant activation of the hippocampus and the dorsolateral prefrontal cortex during the presentation of the first element. However, the introduction of the second element did not recruit the hippocampus but required an increasing activation of the dorsolateral prefrontal cortex, perhaps to maintain two elements in working memory. The addition of the third element upregulated several regions (i.e., the medial parietal cortex), downregulated in both the previous phases, and increased the activation of the hippocampus. Thus, the hippocampus might not be invariably required in a construction scene process. On the contrary, frontally based executive functions are likely to be an integral part of the episodic elements reformulation mechanism.

Another frontally based function, of which future thinking is considered a 'prototypical example' (Buckner and Carroll,

2007: p. 50), is the self-projection. Self-projection is a striking feature of our cognitive system that allows us to shift personal perspectives from the immediate environment to alternative situations. The brain network engaged in self-projection embraces frontal lobes and medial temporal–parietal lobes, which are also upregulated in envisaging the future and episodic remembering, theory of mind, and navigation. Buckner and Carroll (2007) considered these abilities as single components of a broader number of functions. They maintained that the adaptive value of episodic memory lies in its contribution to conceiving mental models of what might happen in the future or other perspectives on the surrounding environment, such as what others are thinking. However, some answers provided by K.C. stand in stark contrast with this hypothesis. When asked, in more than one occasion, to describe what he felt when trying to imagine a future scenario, K.C., as previously mentioned, claimed, "It's like being in a room with nothing there and having a guy tell you go to find a chair, and there's nothing there" (Tulving, 1985: p. 4). This observation uncovered his ability to project himself in different situations. Although he did not provide much detail, he was able to shift his perspective from the immediate environment to alternative presents. And indeed, many years later, Rosenbaum et al. (2007) documented that K.C. preserved his ability to infer other people's thoughts, presenting no difficulties in tasks assessing theory of mind. Therefore, the authors argued that theory of mind's abilities, which are strictly related to self-projection, can be uncoupled by severe impairment of episodic memory and autonoetic consciousness. On the contrary, some researchers consider the concept of autonoetic (derived from the ancient Greek words αὐτός, which means 'own self,' and νοέω, which means 'I think') consciousness as closely related to the capacity of self-projecting into the future (Tulving, 1985; Dalla Barba et al., 1997). When a person remembers an event from his own past, this kind of consciousness confers the awareness that the event is actually a veridical part of his personal past experience (Tulving, 1985). The awareness of subjective time, or temporal consciousness, allows the cognizance of a unique sense of time extending from the past to the future, passing through the present moment (Wheeler et al., 1997). In fact, temporal consciousness usually interacts with less stable, and thus more vulnerable, modifications in the long-term storage system. This work allows quick and steady updates in present orientation and future planning. When temporal awareness is in the condition to make use solely of the more stable modifications of the long-term storage system, the result is that a patient might remember "another" past and plan "another" future, mistakenly involving his habits and personal semantic information in a personal temporal workspace (as in the case of G.A., described by Dalla Barba et al., 1997).

Conclusions

Overall, recent literature seems to suggest that general knowledge and frontally based functions, other than episodic memory, are fundamental when thinking about future. Future thinking is a costly process, which, far from being an encapsulated cognitive function, impinges upon a constellation of constituent abilities. The mental creation of future experiences requires the activation of much of cortical mass and it is likely to be disrupted at different stages, and potentially by "any kind of cortical brain damage" (Berryhill et al., 2010: p. 1391). The matter is intriguing, given that we are all prone to act into the present according to our predictions. Our behavior is determined by personal past experiences and by designs formulating the future. As in Kafka's parable, the individual, "He," gives battle in the present to two antagonists: his past and his future. "The first supports him in his fight with the second, for he wants to push him forward, and in the same way the second supports him in his fight with the first, since he drives him back" (1972: pp. 811–812). Past events never replicate themselves in the same manner. Nonetheless, they undeniably affect our future-oriented thoughts and, consequently, our present. This is the reason why it has been suggested that the best way to predict the future is to forget what we know and give free range to our imagination. "All the rest does not exist, yet. To invent it. That would be wonderful" (Baricco, 1993: p. 143, own translation). And, indeed, Tiresias, legendary prophet of Thebes, was blind. Metaphorically, his prophecies ignored the surrounding environment. Perhaps instead, the best way of thinking about the future is to use our imagination to re-create the past. In this way, we will avoid to be driven back by our past experiences and will rise up, as Kafka wished, to the position of umpire over the two antagonists in their fight with each other.

See also: Amnesia: General; Autobiographical Memory; Episodic Memory; Hippocampus and Related Structures; Reconstructive Memory, Psychology of.

Bibliography

Addis, Donna Rose, Schacter, Daniel L., 2012. The hippocampus and imagining the future: where do we stand? Frontiers in Human Neuroscience 5, 1–14.

Addis, Donna Rose, Wong, Alana T., Schacter, Daniel L., 2007. Remembering the past and imagining the future: common and distinct neural substrates during event construction and elaboration. Neuropsychologia 45, 1363–1377.

Addis, Donna Rose, Wong, Alana T., Schacter, Daniel L., 2008. Age-related changes in the episodic simulation of future events. Psychological Science 19, 33–41.

Addis, Donna Rose, Musicaro, Regina, Pan, Ling, Schacter, Daniel L., 2010. Episodic simulation of past and future events in older adults: evidence from an experimental recombination task. Psychology and Aging 25, 369–376.

Addis, Donna Rose, Sacchetti, Daniel C., Ally, Brandon A., Budson, Andrew E., Schacter, Daniel L., 2009. Episodic simulation of future events is impaired in mild Alzheimer's disease. Neuropsychologia 47, 2660–2671.

Andelman, Fani, Hoofien, Dan, Goldberg, Ilan, Aizenstein, Orna, Neufeld, Miri Y., 2010. Bilateral hippocampal lesion and a selective impairment of the ability for mental time travel. Neurocase 16, 426–435.

Anderson, Rachel J., Dewhurst, Stephen A., 2009. Remembering the past and imagining the future: differences in event specificity of spontaneously generated thought. Memory 17, 367–373.

Anderson, Rachel J., Dewhurst, Stephen A., Nash, Robert A., 2012. Shared cognitive processes underlying past and future thinking: the impact of imagery and concurrent task demands on event specificity. Journal of Experimental Psychology: Learning, Memory and Cognition 38, 38356–38365.

Atance, Cristina M., O'Neill, Daniela K., 2001. Episodic future thinking. Trends in Cognitive Sciences 5, 533–539.

Bar, Moshe, 2007. The proactive brain: using analogies and associations to generate predictions. Trends in Cognitive Sciences 11, 280–289.

Baricco, Alessandro, 1993. Oceano Mare. Rizzoli Libri S.p.A., Milano.

Bergson, Henri, 1913. Matter and Memory. George Allen & Co., Ltd, London.

Berryhill, Marian E., Picasso, Lauren, Arnold, Robert, Drowos, David, Olson, Ingrid R., 2010. Similarities and differences between parietal and frontal patients in autobiographical and constructed experience tasks. Neuropsychologia 48, 1385–1393.

Buckner, Randy L., Andrews-Hanna, Jessica R., Schacter, Daniel L., 2008. The brain's default network. Anatomy, function, and relevance to disease. Annals of the New York Academy of Sciences 1124, 1–38.

Buckner, Randy L., Carroll, Daniel C., 2007. Self-projection and the brain. Trends in Cognitive Sciences 11, 49–57.

Cooper, Janine M., Vargha-Khadem, Faraneh, Gadian, David G., Maguire, Eleanor A., 2011. The effect of hippocampal damage in children on recalling the past and imagining new experiences. Neuropsychologia 49, 1843–1850.

Corkin, Susan, 2002. What's new with the amnesic patient H.M.? Nature Reviews Neuroscience 3, 153–160.

D'Argembeau, Arnaud, Mathy, Arnaud, 2011. Tracking the construction of episodic future thoughts. Journal of Experimental Psychology: General 140, 258–271.

D'Argembeau, Arnaud, Renaud, Olivier, Van der Linden, Martial, 2011. Frequency, characteristics, and functions of future-oriented thoughts in daily life. Applied Cognitive Psychology 35, 96–103.

Dalla Barba, Gianfranco, Cappelletti, Jee Youn, Signorini, Matteo, Denes, Gianfranco, 1997. Confabulation: remembering "another" past, planning "another" future. Neurocase 3, 425–436.

De Hériz, Enrique, 2006. Bugia. Dalai Editore.

de Tocqueville, Alexis, 1863. Democracy in America. Sever and Francis, Cambridge (Original work published 1835–40).

de Vito, Stefania, Della Sala, Sergio, 2011. Predicting the future. Cortex 47, 1018–1022.

de Vito, Stefania, Gamboz, Nadia, Brandimonte, Maria Antonella, Amboni, Marianna, Barone, Paolo, Della Sala, Sergio, 2012. Future thinking in Parkinson's disease: an executive function? Neuropsychologia 50, 1494–1501.

Faulkner, William, 1951. Requiem for a Nun. Random House, New York.

Gaesser, Brendan, Sacchetti, Daniel, Addis, Donna Rose, Schacter, Daniel L., 2011. Characterizing age-related changes in remembering the past and imagining the future. Psychology and Aging 26, 80–84.

Gamboz, Nadia, de Vito, Stefania, Brandimonte, Maria Antonella, Pappalardo, Stella, Galeone, Filomena, Iavarone, Alessandro, Della Sala, Sergio, 2010. Episodic future thinking in amnesic mild cognitive impairment. Neuropsychologia 48, 2091–2097.

Gilbert, Daniel T., 2006. Stumbling on Happiness. Vintage Books, New York.

Haith, Marshall M., 1997. The development of future thinking as essential for the emergence of skill in planning. In: Friedman, Sarah L., Scholnick, Ellin Kofsky (Eds.), The Developmental Psychology of Planning: Why, How, and When Do We Plan? Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, pp. 25–42.

Hassabis, Demis, Kumaran, Dharshan, Vann, Seralynne D., Maguire, Eleanor A., 2007. Patients with hippocampal amnesia cannot imagine new experiences. Proceedings of the National Academy of Sciences of the United States of America 104, 1726–1731.

Hegdè, Jay, 2007. Mental time travel sickness and a Bayesian remedy. Behavioural and Brain Sciences 30, 323–324.

Hoshi, Yoko, Oda, Ichiro, Wada, Yukihisa, Ito, Yasunobu, Yamashita, Yutaka, Oda, Motoki, Ohta, Kazuyoshi, Yamada, Yoshifumi, Tamura, Mamoru, 2000. Visuospatial imagery is a fruitful strategy for the digit span backward task: a study with near-infrared optical tomography. Cognitive Brain Research 9, 339–342.

Ingvar, David H., 1979. Hyperfrontal distribution of the cerebral grey matter flow in resting wakefulness: on the functional anatomy of the conscious state. Acta Neurological Scandinavica 60, 12–25.

Kafka, Franz, 1972. Confessioni e diari. Mondadori.

Kant, Immanuel, 1996. Anthropology from a Pragmatic Point of View. Trans. Victor Lyle Dowdell. Southern University Press, Carbondale, Illinois. Paperback edition (original work published 1798).

Klein, Stanley B., Loftus, Judith, Kihlstrom, John F., 2002. Memory and temporal experience: the effects of episodic memory loss on an amnesic patient's ability to remember the past and imagine the future. Social Cognition 20, 353–379.

Levine, Brian, Svoboda, Eva, Hay, Janine F., Winocur, Gordon, Moscovitch, Morris, 2002. Aging and autobiographical memory: dissociating episodic from semantic retrieval. Psychology and Aging 17, 677–689.

Lidz, Theodore, 1942. The amnestic syndrome. Archives of Neurology and Psychiatry 47, 588–605.

Maguire, Eleanor A., Hassabis, Demis, 2011. Role of the hippocampus in imagination and future thinking. Proceedings of the National Academy of Sciences 108, 39.

Maguire, Eleanor A., Vargha-Khadem, Faraneh, Hassabis, Demis, 2010. Imagining fictitious and future experiences: evidence from developmental amnesia. Neuropsychologia 48, 3187–3192.

Nyberg, Lars, Kim, Alice S., Habib, Reza, Levine, Brian, Tulving, Endel, 2010. Consciousness of subjective time in the brain. Proceedings of the National Academy of Sciences of the United States of America 107, 22356–22359.

Okuda, Jiro, Fujii, Toshikatsu, Ohtake, Hiroya, Tsukiura, Takashi, Tanji, Kazuyo, Suzuki, Kyoko, Kawashima, Ryuta, Fukuda, Hiroshi, Itoh, Masatoshi, Yamadori, Atsushi, 2003. Thinking of the future and the past: the roles of the frontal pole and the medial temporal lobes. Neuroimage 19, 1369–1380.

Race, Elizabeth, Keane, Margaret, Verfaellie, Mieke, 2011. Medial temporal lobe damage causes deficits in episodic memory and episodic future thinking not attributable to deficits in narrative constructions. Journal of Neuroscience 31, 10262–10269.

Romero, Kristoffer, Moscovitch, Morris, 2012. Episodic memory and event construction in aging and amnesia. Journal of Memory and Language 67, 270–284.

Rosenbaum, R. Shayna, Stuss, Donald T., Levine, Brian, Tulving, Endel, 2007. Theory of mind is independent of episodic memory. Science 318, 1257.

Rosenbaum, R. Shayna, Gilboa, Asaf, Levine, Brian, Winocur, Gordon, Moscovitch, Morris, 2009. Amnesia as an impairment of detail generation and binding: evidence from personal, fictional, and semantic narratives in K.C. Neuropsychologia 47, 2181–2187. http://www.sciencedirect.com/science/article/pii/S0028393208004570-aff3.

Samarasekera, Supun R., Vincent, Angela, Welch, Janet L., Jackson, Michael, Nichols, Philip, Griffiths, Thomas D., 2007. Course and outcome of acute limbic encephalitis with negative voltage-gated potassium channel antibodies. Journal of Neurology, Neurosurgery and Psychiatry 78, 391–394.

Schacter, Daniel L., Addis, Donna Rose, 2007. The cognitive neuroscience of constructive memory: remembering the past and imagining the future. Philosophical Transactions of The Royal Society B 362, 773–778.

Schacter, Daniel L., Addis, Donna Rose, Buckner, Randy L., 2007. Remembering the past to imagine the future: the prospective brain. Nature Review Neuroscience 8, 657–661.

Squire, Larry R., McDuffb, Susan G., Frascino, Jennifer C., 2011. Reply to Maguire and Hassabis: autobiographical memory and future imagining. Proceedings of the National Academy of Sciences 108, 40.

Squire, Larry R., van der Host, Anna S., McDuff, Susan G., Frascino, Jennifer C., Hopkins, Ramona O., Mauldin, Kristin N., 2010. Role of the hippocampus in remembering the past and imagining the future. Proceedings of the National Academy of Sciences of the United States of America 107, 19044–19048.

Suddendorf, Thomas, 2010. Episodic memory versus episodic foresight: similarities and differences. Wiley Interdisciplinary Reviews Cognitive Science 1, 99–107.

Suddendorf, Thomas, Corballis, Michael C., 1997. Mental time travel and the evolution of the human mind. Genetic, Social & General Psychology Monographs 123, 133–167.

Suddendorf, Thomas, Corballis, Michael C., 2007. The evolution of foresight: what is mental time travel, and is it unique to humans? Behavioural and Brain Sciences 30, 299–351.

Summerfield, Jennifer J., Hassabis, Demis, Maguire, Eleanor A., 2010. Differential engagement of brain regions within a 'core' network during scene construction. Neuropsychologia 48, 1501–1509.

Szpunar, Karl K., 2010. Episodic future thought: an emerging concept. Perspectives on Psychological Science 5, 142–162.

Szpunar, Karl K., Watson, Jason M., McDermott, Kathleen B., 2007. Neural substrates of envisioning the future. Proceedings of the National Academy of Sciences 104, 642–647.

Tolman, Edward C., Ritchie, Benbow F., Kalish, Donald, 1946. Studies in spatial learning. I. Orientation and the short-cut. Journal of Experimental Psychology 36, 13–24.

Tulving, Endel, 1972. Episodic and semantic memory. In: Tulving, Endel, Donaldson, Wayne (Eds.), Organization of Memory. Academic Press, New York, pp. 381–403.

Tulving, Endel, 1985. Memory and consciousness. Canadian Psychology 26, 1–12.

Tulving, Endel, 2002. Episodic memory: from mind to brain. Annual Review of Psychology 53, 1–25.

Wheeler, Mark A., Stuss, Donald T., Tulving, Endel, 1997. Toward a theory of episodic memory: the frontal lobes and autonoetic consciousness. Psychological Bulletin 121, 331–354.

Future Time Perspective and Motivation

Willy Lens[†], University of Leuven, Leuven, Belgium; and University of the Free State, Bloemfontein, South Africa
Rachel Seginer, University of Haifa, Haifa, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article we discuss future time perspective (FTP) and future orientation (FO) as two cognitive–motivational conceptualizations of future thinking. We first delineate each conceptualization and conclude by highlighting some of its commonalities and differences. We then summarize research on the motivational antecedents and consequences of FTP and FO. Finally we relate our research to three issues: Why is it that not all future thinking prompt behavior? Does culture (as well as other aspects of the social milieu) matter? How can our research findings be implemented to improve individuals' prospects?

Introduction

People live of course in the present, in the here and now, but they also live to a certain degree with their past and they can foresee a more or less distant future. The personal past, present, and future are psychologically integrated in the psychological life space (Lewin, 1942) so that the psychological present encompasses all three time periods: the experienced present, remembered past, and anticipated future. Depending on which of the three time periods is the most dominant part of their present life space, people are either past-, present-, or future-oriented (Zimbardo and Boyd, 1999). While each orientation may have distinct effects on present behavior, our aim here is to focus on the future and describe its motivational properties.

Of the multiple conceptualizations psychologists have been using to describe future thinking (Seginer, 2009), in this article we address mainly two: the future time perspective (FTP) and future orientation (FO). The article consists of three parts. In the first we delineate each conceptualization and conclude by highlighting some of its commonalities and differences. In the second we summarize research carried out mainly in the last decade on the motivational consequences of FTP and FO. In the final part we summarize our research by relating to three issues: Why is it that not all future thinking prompt behavior? That is, what conditions enhance or decrease future thinking–behavior relations? Does culture (as well as other aspects of the social milieu) matter? How can our research findings be implemented to improve individuals' prospects?

Two Conceptualizations: FTP and FO

Future Time Perspective

FTP is the degree to which, and the way in which, the future is anticipated and integrated in the psychological present of an individual (Lewin, 1942). Two important characteristics of the individual FTP are its extension and its content: How far into the chronological future reaches one's psychological future and what does one anticipate in the future? Does it consist mostly of positive or negative events, hopes, or fears? Which life domains are predominantly anticipated: education, work, family life, and leisure (Seginer, 2009)? As we will document, both aspects – extension and content – are motivationally relevant.

For two reasons, we conceptualize FTP as a cognitive–motivational variable: It results from goal setting and has motivational consequences (Lens, 1986).

FTP Results from Goal Setting

Formulating distant motivational goals and developing long-range behavioral projects to achieve those goals, create a long or extended FTP. People with a rather short FTP set most of their goals in the near future and do not take into account what will come later. People with a longer or deeper FTP formulate relatively more long-term goals and their present temporal life space extends into the very distant future. They have no problem at all to reckon, already now, with or to be motivated by events or action outcomes in the rather distant future (Lens, 1986; Seginer, 2009).

People have a rather long FTP not only when they set goals for themselves in the rather distant future but also when they take into account in the present personally relevant events that may happen to them in the distant future, even in the afterlife. Boyd and Zimbardo (2005) developed a scale measuring the transcendental FTP with items that suppose the belief in life after death (e.g., My body is just a temporary home for the real me; Humans possess a soul). For a review of other techniques used to measure FTP, see Nuttin et al. (1979).

Other aspects of FTP than its extension are the degree of realism, density (i.e., how many goals or events are foreseen), structure (e.g., means-end relations) (Lens, 1986); and connectedness with the present (Husman and Shell, 2008) as endogenous and exogenous relationships between present and future goals (Husman and Lens, 1999). Especially the content of FTP is getting more and more research attention. Nuttin and Lens (1985) developed a coding system identifying the content of future goals (e.g., self-realization, knowledge, affiliation, power, achievement, etc.).

Based on the self-determination theory (SDT, Deci and Ryan, 2002) de Bilde et al. (2011) distinguished intrinsic and extrinsic future goals. Peetsma and Van der Veer (2013) and Seginer (2009) analyzed people's FTP – mostly adolescents and emerging adults – in different life domains such as

[†] Deceased.

education, work, leisure, and family life. People do indeed not anticipate an empty future, they anticipate positive and negative events, they formulate hopes and fears for the future, they pursue positive (i.e., things they want to achieve) and negative (i.e., things they want to avoid) future goals. Seginer (2009) described this as a thematic approach of FTP and distinguishes it from an athematic approach. In the athematic approach of FTP, one only considers to what extent people are future-oriented, how far into the future people anticipate, to what extent they link their present and future life periods, and whether they view the future as mainly positive or negative without considering any contents.

FTP became the object of predominantly correlational studies, relating individual differences in the length or extension of FTP (and its content) with other personal characteristics such as intelligence, age, gender, socioeconomic status, ethnicity, well-being, self-esteem, achievement motivation, delinquency, substance (ab)use, and health behavior (see Seginer, 2009). Given the correlational nature of these studies and absence of longitudinal approaches, the findings did not really allow inferring causal antecedents of differences in FTP. There was, and still is, a lack of research on how to develop a longer FTP or how to help elderly people to stay future-oriented, rather than turning too early in life to a less motivating past- or present orientation.

In summary, the concept of FTP as an individual psychological construct is a rather simple one: How far into the future do people project their goals and what type of goals, aspirations, and fears do they have? How far into the future do they anticipate personally relevant events and what is the content of those events, which life domain do they represent? Do individuals and social groups differ in the content of relevant events?

Starting from this unidimensional cognitive–motivational conceptualization of FTP and other similar conceptualizations Seginer (2005, 2009) developed a multidimensional conceptualizations of FTP – or FO according to their terminology – each consisting of three components. Here we describe the conceptualization initially developed by Seginer et al. (1991) and further elaborated by Seginer (2009; Seginer et al., 2004).

A Three-Component Model of FO

Although here we focus on the three-component model, we start its presentation by briefly describing its evolvement out of the unidimensional approach.

The Unidimensional Approach

A basic premise of the FO conceptualization has been that FO is not about predicting the future but about the subjective images individuals hold about it. These images are thematic (Nuttin and Lens, 1985), expressed in terms of the *hopes* individuals have about the future and the *fears* they would rather avoid (Trommsdorff, 1983), and can be subsumed under life domains such as education, work, and family. Work carried out in Israel showed that adolescents growing up in such diverse sociocultural milieus as Jewish urban and kibbutz, Muslim, and Druze rural communities described the future in terms of six common life domains: education, work and career, marriage and family, self-concerns, others, and the collective. Given the normative life course of Jewish adolescents and Druze boys, their FO includes also military service.

While the domains were similar, endorsement has been culture specific. Thus, Muslim and Druze girls construct an FO that has high-density (i.e., relative representation) higher education. Moreover, data collected for over 20 years show Muslim and Druze girls list more hopes and fears about higher education than do Muslim and Druze boys and Jewish girls and boys (Seginer, 2009).

The advantages of the unidimensional approach notwithstanding, its narrow conceptualization and methodological limitations led to the construction of the three-component model. This model draws on the FO unidimensional conceptualization in two ways. First, it takes the cognitive representation as its starting point, and asks two questions: What prompts the cognitive representation of the future and what are its outcomes. Second, given that time images are not void of content, the model applies to each of several future life domains, thus preserving the thematic approach to FO.

Expanding the Model

Drawing on a basic premise that cognitive representation is prompted by motivational forces and often results in behavior, the expanded model consists of motivational, cognitive representation, and behavioral components. Together they form a multiple-step model in which the motivational component affects the cognitive and the behavioral components, and the cognitive component affects the behavioral component (Seginer, 2009). The components are theoretical and indicated by empirical variables as next described.

The Motivational Component

Drawing on motivation theory and earlier conceptualizations of FTP (Nuttin and Lens, 1985) and FO (Trommsdorff, 1983), three empirical variables indicate this component: the *value* of the prospective life domain; *expectance* defined as the subjective probability that hopes and plans will be materialized, and its affective tone. The third is a sense of *internal control*: the person's responsibility for the materialization of future hopes and plans.

The Cognitive Representation Component

This component relates to the extent to which individuals think about the future, in terms of hopes and fears about each future life domain.

The Behavioral Component

Its conceptualization draws on the work of Lewin and Erikson. *Exploration* relates to seeking information and advice regarding future options and testing how well they fit personal abilities and values, social expectations, and environmental circumstances. Commitment consists of the decision to pursue one option and results in "a sense of knowing where one is going" (Erikson, 1968: p. 165). Both add to the instrumentality of FO and prompt the achievement of future hopes, goals, and plans.

The Importance of Content: Domain Specificity

The three-component model applies to each of several future life domains included in individuals' future life space. Its viability is supported by empirical findings indicating the domain-specific effect of FO on present behavior (Seginer, 2009).

Summary: FTP and FO Commonalities and Differences

A basic premise common to the FTP and FO is that time perspective is content-based. Thus, both approaches are thematic. Moreover, while both understand the motivational qualities of future thinking, they also emphasize conditions enhancing and reducing motivation for present behavior (see Summary and Conclusion section). The two approaches differ on three main issues: only FTP conceptualization emphasizes extension, the FO approach is more concerned with various life domains and cultural context. Finally, in the FTP theoretical framework, goals precede FTP whereas in the FO conceptualization, goal setting follows FO.

The Motivational Consequences of FTP and FO

Motivational Consequences of FTP: Extension and Content Matter

Extension of FTP Affects the Quantity of Motivation

Estimating the present motivational effects of future goals by using the 'expectancy × value' or the 'expectancy × instrumentality × value' theories of motivation (Eccles and Wigfield, 2002) has been proven useful. Extension of FTP is positively related to motivation: Students with a longer FTP are in general more motivated. How comes? Extension of FTP is negatively related to the psychological distance toward moments in the near or distant future. Differences in extension will probably not affect the psychological distance to moments in the very near future (e.g., tomorrow, next week) or the very distant future (e.g., 50 years from now for an adolescent). But, for example, 1 or 5 years from now into the future is psychologically much closer for people with a long FTP than for those with a short FTP. This has two motivationally important implications: anticipated value of the future goal and perceived instrumentality of present actions. In general, the anticipated rewarding value of a given goal decreases the more delayed it is. But this decrease is less steep for people with a longer FTP. This can explain why students with a short FTP may choose to become a nurse after 3 years of study rather than a physician after 7 years, or discount the possible unhealthy and even dangerous future consequences of present risky sexual behavior, smoking, drinking, and using drugs (Keough et al., 1999; Apostolidis et al., 2006). Second, the longer the FTP, the easier it is to see the instrumental or utility value (Eccles and Wigfield, 2002) of present actions. For example, students with a longer FTP have not only educational goals for the near future but also educational and vocational goals in the more distant future. Their present study activities are more instrumental, have a higher utility value than for students with a short FTP. As predicted, perceived instrumentality is positively related to motivation in education (Andriessen et al., 2006; Husman and Shell, 2008; Mikkonen et al., 2013). An important limitation of most of these studies is that they are correlational and cross-sectional so that there is no room for causal interpretations.

Content of FTP Affects the Quality of Motivation

The motivation to pursue future goals is by definition extrinsic motivation and especially in educational psychology extrinsic motivation was considered to be of lower quality than intrinsic motivation. More recent empirical research based on the SDT (Deci and Ryan, 2002) shows that not all types of extrinsic motivation are of low quality motivation. The intrinsic–extrinsic distinction is replaced by the more valid distinction between autonomous motivation and controlled motivation. An autonomous motivation is 'high-quality motivation'. When behavior is externally regulated (e.g., by rewards, punishments, supervision, etc.) or regulated by introjected motives (e.g., avoid shame and guilt feelings) its motivation is controlled and thus a nonoptimal type of motivation. When it is intrinsically motivated its motivation is autonomous, volitional. However, extrinsically motivated behavior regulated by future goals that are personally important (e.g., to become a psychologist; identified regulation) or perceived as totally congruent with one's true self (e.g., to become a physician with doctors without borders; integrated regulation) is also autonomously motivated.

Whether future goals create controlled or autonomous motivation depends on their content. SDT makes a distinction between intrinsic and extrinsic goals (Deci and Ryan, 2002). Intrinsic future goals create autonomous motivation and extrinsic future goals create controlled motivation. Examples of intrinsic goals are self-development, physical fitness, competence, community contribution, affiliation, altruism, etc. Examples of extrinsic goals are financial success, power, status, physical attractiveness, outperforming others, admiration, etc.

There is ample empirical evidence from correlational but especially from experimental research in real life conditions that experimentally inducing future intrinsic goals has more adaptive outcomes than inducing extrinsic goals (for a review, see Lens et al., 2012).

The SDT (Deci and Ryan, 2002) predicts that pursuing intrinsic goals is always, everywhere, and for everybody better than pursuing extrinsic goals. This difference is explained by the fact that intrinsic goals satisfy and extrinsic goals frustrate the three innate basic psychological human needs proposed by the SDT: autonomy, competence, and relatedness/belongingness. To examine its viability, Vansteenkiste et al. (2008) designed an experiment in which the effect of intrinsic vs extrinsic goals was contrasted with the effect of a match or fit between people's personal goal orientations and the goals that are emphasized or valued in the immediate environment (the match-perspective or fit hypothesis; Pervin, 1968). Results showed that the hypothesis that students pursuing intrinsic goals will score higher than students pursuing extrinsic goals on the dependent measures (i.e., autonomous motivation, conceptual learning, rote learning) was confirmed, and the hypothesis predicting that students whose goals match the

goals of the experimental goal condition would score higher on the dependent measures was rejected.

The Motivational Consequences of FO

To date, our research on the motivational effect of FO relates to three criteria, each relevant to the development of adolescents and emerging adults. The three are academic achievement, resilience, and achieved identity. Although conceptually different, they are all indicated by behaviors and beliefs and therefore the rationale underlying the effect of FO on each bears some similarities. Specifically, the effect of FO on each is explained by two principles. One is *perceived instrumentality* (Husman and Lens, 1999; Nuttin and Lens, 1985), i.e., understanding the instrumentality of present behavior for the fulfillment of future hopes, goals, and plans. The second pertains specifically to the behaviors and processes described by the three-component FO construct.

Thus, it is not just being concerned about the consequences of present behavior for one's future (Strathman et al., 1994), or having a positive attitude toward the future. Instead, these processes are domain specific and involve activities indicating each of the three component variables. For the motivational component, they are appraising the *value*, considering the probability of materialization (*expectance*) of hopes (and the goals ensuing form them), of a specific domain and one's *control* over it. For the cognitive representation component, these activities are developing *hopes* and *fears* associated with it. Finally, for the behavioral component they are actively exploring its suitability for oneself and going through a process of decision making. Jointly, they prompt behaviors leading to academic achievement and resilience, and beliefs and behaviors indicating achieved identity.

Findings supporting the effect of the three-component FO construct on each of the three criteria are presented below. FO is assessed by a self-report battery consisting of an open-ended protocol in which respondents list their hopes and fears for the future (the hopes and fears questionnaire), and two Likert-type questionnaires estimating all empirical variables for each life domain (prospective life course questionnaire, my future questionnaire).

The Effect of FO on Academic Achievement

For children, adolescents, and emerging adulthood students, academic achievement is an indicator of competence and adjustment. Particularly for adolescents and emerging adults, it is also an essential condition for successfully engaging in adult roles. Data collected in the last decade and analyzed by means of structural equation modeling (SEM) confirmed the hypothesis about the effect of the three-component model of FO on academic achievement for several age and cultural groups. These groups consist of Muslim, Druze, and Jewish 11th and 9th graders and emerging adults (Seginer, 2009). For each of these groups, the model is satisfactorily estimated. Nonetheless, the findings also indicate domain specificity and culture variations.

Thus, the FO competence domains (i.e., higher education, work, and career) explain a higher percentage of the variance of academic achievement (reported by the participants' schools) than does the relational domain (marriage and family). In fact, the effect of marriage and family on academic achievement is significant for only two groups: the Muslim girls in Israel, and ultraorthodox Chabad Jewish girls (Seginer, 2009). However, the effect is *negative* for the Muslim girls and *positive* for the Chabad Jewish girls. Underlying these findings is the cultural meaning of marriage and family for these two cultural groups.

In the Muslim tradition girls should marry early, devote themselves to the family, and stay close to home. Although with growing numbers of educated Muslim women and economic family needs, the number of women holding a paid job has increased and stands at 21% (vs 60% among Muslim men, Israel Central Bureau of Statistics, 2010), girls still feel the pressure. They worry that early marriage will interfere with their desire for higher education, which they regard as "a weapon in women's hands" (Seginer and Mahajna, 2012). The importance Muslim girls in Israel attribute to higher education must be considered in light of two conflicting forces: its high promise for emancipated life and the strong pressure exerted by their families and communities toward early marriage. This tension is voiced by the girls' hopes and fears narratives. "For me marriage is very important but not at the expense of the girl's education. Marriage is inescapable but it has its conditions and criteria. Therefore, my hope is first study and only later on get married" (Seginer and Mahajna, 2004: p. 132). "Regarding marriage, I don't think about it now because I want to complete my education first" (Seginer, 2009: p.170).

In the ultraorthodox Jewish tradition too marriage and family are of high value and girls (as well as boys) are expected to marry early. However, for two reasons marriage and education do not conflict but rather complement each other. Given that men devote themselves to religious studies, women are responsible for bringing up the children and – by holding a paid job – for the family income. With better education, they will give their children a better education and get a better job. The girls' awareness of the importance of academic achievement for fulfilling their traditional responsibilities is echoed in their future hopes and fears narratives. "With God's help I would like to have high grades in order to have a profession and so with God's help my future husband will devote himself to religious studies." "Graduate from high school with very high grades, get a high psychometric score, study at the Chabad Teachers' college, with God's help marry a Chasidic husband, become an emissary, and have many well-educated Chasidic children."

The Effect of FO on Resilience

Common to the different conceptualizations of resilience is the tendency to engage in behaviors that indicates successful adaptation to serious adversity (Masten, 2014). Its nature varies according to the type of adverse circumstances and the individuals' developmental period, gender, and sociocultural milieu. For adolescents diagnosed as at-risk and placed in institutional care ('youth villages') with adolescents similarly diagnosed, resilience is indicated by overcoming deviant behavior, accepting the authority of adult figures, particularly teachers and the youth village care team, and participating in social and educational activities in the youth village.

Research on the effect of FO on resilience is at its beginning. The findings presented here are the result of one study about

at-risk youth who for adverse family circumstances live and attend school in residential care known in Israel as youth villages. Resilient behavior was assessed by the response of the youth village care team to four scales pertaining to aggressive behavior, deviant behavior, acceptance of adult authority, and active participation in the youth village social and educational activities. SEM analyses show a satisfactory fit with the theoretical model in which the three-component FO affects resilience. However, the extent to which FO explains the variance of resilient behavior varies. It is highest (17%) for continued education and lowest (4%) for work and career, possibly because for these youths work is not a pathway to career. The marriage and family domain explains 8% of the variance of resilience.

The Effect of FO on Achieved Identity

Identity has multiple definitions, but a common theoretical origin. We owe the term and the richness of its conceptualization to Erikson. The analysis of the effect of FO on achieved identity draws on his proposition, "The young person, in order to experience wholeness, must feel a progressive continuity between that which he has come to be during the long years of childhood and that *which he promises to become in the anticipated future*…" (1968: p. 87, *italics* added). However, while Erikson contended that identity "comes to its ascendance" (1968: p. 95) during adolescence, social circumstances in industrial societies have resulted in a postponed entrance to adulthood via the period of emerging adulthood. Therefore, here we report of research carried out with emerging adults.

Of the three studies we describe here, two were carried out with Israeli Jewish and one with Belgian emerging adults; one is cross-sectional and two are longitudinal. Given the multiple conceptualizations of FO and FTP and identity formation, each study has used different conceptualizations. Nevertheless, their findings lead to common conclusions, indicating that while concurrently FO precedes identity development, across time FO and FTP and identity are reciprocally related.

The study (Seginer, 2009) examining the effect of FO on identity concurrently assessed the motivational and behavioral components of FO regarding the two universal adult roles: work and family. Findings showed that for both future life domains, the effect on identity was strongest for expectance (a motivational variable) and commitment (a behavioral variable).

Next, (Seginer and Ablin, in preparation) FO was conceptualized and assessed by all three FO components for the two adult role domains. Identity was assessed by the objective measure of ego identity in adolescence assessing two identity domains: ideological and relational. Drawing on Erikson's continuity proposition, the hypothesis was that across time (18 months), FO has a direct effect on achieved identity. Testing the hypothesis separately for the effect of the work and career FO on ideological identity (instrumental) and of marriage and family FO on relational identity (relational), called for a modified model. Showing satisfactory goodness of fit, this model indicated that while in both Time 1 and Time 2 FO had a direct effect on achieved identity, across time; across time, they are reciprocally related. Testing a similar question using different conceptualizations and research instruments, Luyckx et al. (2010) found that over a 4 months period, for Belgian college students FTP and identity had a reciprocal effect. SEM testing three alternative models in which FTP and two present time perspective (PTP) scales were included (FTP and PTP affect identity, identity affects FTP and PTP, FTP and PTP and identity are mutually related), showed the best fit was obtained for the mutual effect of FTP (and PTP) and identity. Moreover, the reciprocal model also shows that Time 1 FTP has the highest positive effect on commitment and the highest negative effect on the diffuse-avoidant style.

Summary: Conditions for Prompting the Motivational Consequences of FTP and FO

By way of summary it is important to note that not all future thinking motivates present behavior. Indefinite and non-committing general goals, self-enhancing possible selves (Oyserman et al., 1995), and hopes and fear pertaining to existential domains (Seginer, 2009) have low motivating power. Moreover, positive fantasies have a negative effect on academic achievement among lower class students (Kappes et al., 2012). These findings make it especially important to highlight condition that enhances the motivational qualities of FTP and FO.

Terminology and conceptual differences notwithstanding research emanating from both approaches highlights three conditions enhancing the motivating qualities of FTP and FO. One relates to temporal proximity: The closer the distance between the task to which FO relates and present behavior, the stronger is the effect of FTP and FO on present behavior. However, as indicated particularly by FTP research, individuals with extended FTP perceive distant goals as psychologically closer. Hence, they attribute higher value to distant goals and perceive the instrumentality of behavior leading to the fulfillment of these goals more clearly than do individuals with short FTP.

The second condition relates to content similarity between FTP or FO under consideration and present behavior. High similarity – or *endogenous instrumentality* – enhances the motivational power of FTP and FO, whereas low similarity – or *exogenous instrumentality* – does not. The third condition also relates to content. Specifically, it is the content of future goals as arousing autonomous (rather than controlling) motivation. Future goals arouse autonomous motivation if these goals are personally important or perceived as congruent with the person's true self, so that their pursuit is characterized by identified or integrated regulation.

While these conditions relate to individuals' characteristics and the nature of goals, a fourth condition relates to the sociocultural context. As indicated by FO findings, in comparison to middle class Jewish adolescents, the effect of higher education FO on academic achievement is higher for underprivileged adolescents. Aware of social and economic obstacles awaiting them in fulfilling higher education goals, they perceive the instrumentality of academic achievement for materializing their higher education hopes.

See also: Academic Achievement Motivation, Development of; Academic Self-Concept and Achievement Erikson's Theory of Psychosocial Development; Erikson, Erik Homburger (1902–94); Future Thinking; Lewin, Kurt (1890–1947).

Bibliography

Andriessen, I., Phalet, K., Lens, W., 2006. Future goal setting, task motivation and learning of minority and non-minority students in Dutch schools. British Journal of Educational Psychology 76, 827–850.

Apostolidis, T., Fieulaine, N., Soulé, F., 2006. Future time perspective as predictor of cannabis use: exploring the role of substance perception among French adolescents. Addictive Behaviors 31, 2339–2343.

Boyd, J.N., Zimbardo, P.G., 2005. Time perspective, health, and risk-taking. In: Strathman, A., Joireman, J. (Eds.), Understanding Behavior in the Context of Time: Theory, Research, and Applications. Erlbaum, Mahwah, NJ, pp. 85–107.

de Bilde, J., Vansteenkiste, M., Lens, W., 2011. Understanding the association between future time perspective and self-regulated learning through the lens of self-determination theory. Learning and Instruction 21, 332–344.

Deci, E.L., Ryan, R.M., 2002. Handbook of Self-Determination Research. University of Rochester Press, Rochester, NY.

Eccles, J.S., Wigfield, A., 2002. Motivational beliefs, values, and goals. Annual Review of Psychology 53, 109–132.

Erikson, E.H., 1968. Identity, Youth and Crisis. Norton, New York.

Husman, J., Lens, W., 1999. The role of the future in student motivation. Educational Psychologist 34, 113–125.

Husman, J., Shell, D.F., 2008. Beliefs and perceptions about the future: a measurement of future time perspective. Learning and Individual Differences 18, 166–175.

Israel Central Statistical Bureau, 2010. Israel Statistical Yearbook. Central Bureau of Statistics, Jerusalem.

Kappes, H.B., Oettingen, G., Mayer, D., 2012. Positive fantasies predict low academic achievement in disadvantaged students. European Journal of Social Psychology 42, 53–64.

Keough, K.A., Zimbardo, P.G., Boyd, J.N., 1999. Who's smoking, drinking, and using drugs? Time perspective as a predictor of substance abuse. Basic and Applied Social Psychology 21, 149–164.

Lens, W., 1986. Future time perspective: a cognitive-motivational concept. In: Brown, D.R., Veroff, J. (Eds.), Frontiers of Motivational Psychology. Springer-Verlag, New York, pp. 173–190.

Lens, W., Paixão, M.P., Herrera, D., Grobler, A., 2012. Future time perspective as a motivational variable: content and extension of future goals affect the quantity and quality of motivation. Japanese Psychological Research 54, 321–333.

Lewin, K., 1942/1948. Time perspective and morale. In: Lewin, K. (Ed.), Resolving Social Conflict. Harper & Brothers Publishers, New York, pp. 103–124.

Luyckx, K., Lens, W., Smits, I., Goossens, L., 2010. Time perspective and identity formation: short-term longitudinal dynamics in college students. International Journal for the Study of Behavioral Development 34, 238–247.

Masten, A.S., 2014. Global perspectives on resilience in children and youth. Child Development 85, 6–20.

Mikkonen, J., Ruohoniemi, M., Lindblom-Ylänne, S., 2013. The role of individual interest and future goals during the first years of university studies. Studies in Higher Education 38, 71–86.

Nuttin, J., Lens, W., 1985. Future Time Perspective and Motivation: Theory and Research Method. Leuven University Press & Erlbaum, Leuven, Belgium and Hillsdale, NJ.

Nuttin, J., Lens, W., Van Calster, K., De Volder, M., 1979. La perspective temporelle dans le comportement humain: étude théorique et revue de recherches. In: Fraisse, P., et al. (Eds.), Du temps biologique au temps psychologique. P.U.F, Paris, pp. 307–363.

Oyserman, D., Gant, L., Ager, J., 1995. A socially contextualized model of African American identity: school persistence and possible selves. Journal of Personality and Social Psychology 69, 1216–1232.

Peetsma, T., Van der Veer, 2013. Relations between the development of future time perspective in three life domains, investment in learning, and academic achievement. Learning and Instruction 21, 481–494.

Pervin, L.A., 1968. Performance and satisfaction as a function of individual-environment fit. Psychological Bulletin 69, 56–68.

Seginer, R., 2005. Adolescent future orientation: intergenerational transmission and intertwining tactics in culture and family settings. In: Friedelmeier, W., Chakkarath, P., Schwarz, B. (Eds.), Culture and Human Development: The Importance of Cross-Cultural Research to the Social Sciences. Psychology Press, Hove, UK, pp. 231–251.

Seginer, R., 2009. Future Orientation: Developmental and Ecological Perspectives. Springer, New York.

Seginer, R. Ablin, E. Future orientation and achieved identity in emerging adulthood: a longitudinal study, manuscript in preparation.

Seginer, R., Mahajna, S., 2004. How the future orientation of traditional Israeli Palestinian girls link beliefs about women's roles and academic achievement. Psychology of Women Quarterly 28, 122–135.

Seginer, R., Mahajna, S., 2012. With God's help: the future orientation of Palestinian girls in Israel growing up Muslim. In: Trommsdorff, G., Chen, X. (Eds.), Values, Religion, and Culture in Adolescent Development. New York and Cambridge, pp. 253–270.

Seginer, R., Nurmi, J.-E., Poole, M., 1991. Adolescent Future Orientation in Cross-Cultural Perspectives. Paper Presented in the 11th Meeting of International Society for the Study of Behavioral Development. Minneapolis, MN.

Seginer, R., Vermulst, A., Shoyer, S., 2004. The indirect link between perceived parenting and adolescent future orientation: a multiple-step analysis. International Journal of Behavioral Development 28, 365–378.

Strathman, A., Gleicher, F., Boninger, D.S., Edwards, C.S., 1994. The consideration of future consequences: weighing immediate and distant outcomes of behavior. Journal of Personality and Social Psychology 66, 742–752.

Trommsdorff, G., 1983. Future orientation and socialization. International Journal of Psychology 18, 381–406.

Vansteenkiste, M., Timmermans, T., Lens, W., Soenens, B., Van den Broeck, A., 2008. Does extrinsic goal framing enhance extrinsic goal-oriented individuals' learning and performance? An experimental test of the match-perspective vs. self-determination theory. Journal of Educational Psychology 100, 387–397.

Zimbardo, P., Boyd, J.N., 1999. Putting time in perspective: a valid, reliable, individual-differences metric. Journal of Personality and Social Psychology 77, 1271–1288.

G

Gaia

Stephan Harding, Schumacher College, Dartington, Devon, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Gaia theory suggests that the Earth displays emergent self-regulation of key surface parameters such as temperature and oxidation state due to feedbacks between living beings and rocks, atmosphere, and water. This article gives a scientific outline of the theory and considers some of its philosophical and societal implications.

In the early 1970s, the English scientist and inventor, James Lovelock (helped later on by the American evolutionary biologist Lynn Margulis) postulated the Gaia hypothesis, which stated that life on Earth is a global phenomenon which drives a planetary self-regulating system that binds the atmosphere to the continents and oceans in a quasiphysiological, self-regulating process (Lovelock, 1972, 1979; Lovelock and Margulis, 1975). Lovelock was inspired to conceive of the Gaia hypothesis while contributing to NASA's efforts to find life on Mars (Lovelock, 1965). As a theory of evolution, Gaian thinking in science holds that life forms have coevolved with their environment in such a way that environmental conditions have remained suitable for life over geological time despite a continually increasing influx of solar energy and a variable input of carbon dioxide (a greenhouse gas) from volcanic activity.

The Gaia theory, Lovelock's more recent, more sophisticated elaboration of his initial insight, posits that feedbacks between living organisms and rocks, atmosphere and water regulate the planetary surface environment, which would otherwise be rendered uninhabitable by the forces of planetary physics and chemistry (Lovelock, 1988). The evidence put forward for the Gaia theory draws attention to the fact that the chemical composition of the Earth's atmosphere has stayed remarkably constant over the last 545 million years or so, and is quite unlike the composition that would be expected in the abiological steady state of a lifeless planet. The geological record suggests that temperature has remained within the narrow limits that life can tolerate over geological time, since neither global freezing nor exceedingly hot conditions have prevailed for long on the planet, if at all. Mean surface temperature has probably never departed for geologically significant periods from the range 5–50 °C (Lovelock and Whitfield, 1982). The Gaian perspective has spawned new avenues of scientific inquiry aimed at furthering the understanding of biogeochemical cycles, especially in the atmospheric sciences. For example, the importance of methane in climate change came to be taken seriously with the recognition that organisms emitting this gas play a vital role in determining the surface temperature of the planet. The Gaia theory also led to the discovery that organisms influence planetary temperatures by accelerating the weathering of silicate rocks such as granite and basalt, and by emitting gasses that seed planet-cooling clouds over both land and ocean.

More recent scientific work inspired by the Gaia theory strongly suggests that there have been several hugely important but highly improbable global revolutions during the 3500 million years that have totally reconfigured our planetary environment in ways that paved the way for today's multicellular biosphere (Lenton and Watson, 2011). This suggests that planets with high atmospheric oxygen and multicellular organisms as seen on Earth might be very rare in the cosmos.

Lovelock (1988) coined the term geophysiology to denote an emerging systems approach to earth science that might assist in furthering understanding of the implications of human interventions into climate and biogeochemical cycles in ways that could threaten the current configuration of the planetary operating system. By using the term geophysiology, Lovelock distanced himself from teleological criticisms and the mysticism that has come to surround the notion of Gaia. For Lovelock, Gaia is a quasiliving entity that has the capacity for global level self-regulation operating according to cybernetic principles, which like a thermostat, does not require will or sentience.

Gaia caught the public imagination in the 1980s, along with an emerging conviction in the scientific community that the global environment was being seriously affected by human action (Ruse, 2013). For environmentalists, Gaia became a symbol of the need for promoting an ecocentric view of the world around us, in contrast to the dominant anthropocentric view. For many, Gaia also became an argument that cooperation rather than competitiveness was needed to unite humanity in the face of threats to the global environment and an argument for alternative ways of living. It is important to distinguish between the Gaia theory as mainstream science, and Gaia, the popular symbol of environmentalists, which underpins a call for an ecological egalitarianism to end all forms of human domination of

nature (Serafin, 1988). A considerable literature has developed exploring Gaia both as a scientific concept and as an ethical, philosophical, and sociological proposition. Debates in the literature about Gaia have been set increasingly in the historical context of an emerging science of the biosphere. Gaia is seen as an elaboration of ideas put forward by the Englishman James Hutton who first proposed Earth to be a living organism in 1785 and the Russian Vladimir Vernadsky who, in the early part of the twentieth century, realized the importance of life as a geological force.

See also: Environment and Development; Sustainable Development: An Economic Perspective.

Bibliography

Bunyard, P. (Ed.), 1996. Gaia in Action: Science of the Living Earth. Floris Books, Edinburgh, UK.
Crist, E., Rinkler, B. (Eds.), 2010. Gaia in Turmoil. MIT Press.
Harding, S.P., 2009. Animate Earth: Science, Intuition and Gaia. Green Books, Dartington, UK.
Lenton, T., Watson, A., 2011. Revolutions that Made the Earth. Oxford University Press, Oxford, UK.
Lovelock, J.E., 1965. A physical basis for life detection experiments. Nature 207, 568.
Lovelock, J.E., 1972. Gaia as seen through the atmosphere. Atmospheric Environment 6, 579–580.
Lovelock, J.E., 1979. Gaia: A New Look at Life on Earth. Oxford University Press, Oxford, UK.
Lovelock, J.E., 1988. The Ages of Gaia: A Biography of Our Living; Earth. Norton, New York.
Lovelock, J.E., Margulis, L., 1975. The atmosphere as circulatory; system of the biosphere – the Gaia hypothesis. CoEvolution Quarterly 6, 30–41.
Lovelock, J.E., Whitfield, M., 1982. Life span of the biosphere. Nature 296, 561.
Midgley, M., 2007. Earthly Realism. The Meaning of Gaia. Imprint Academic, Exeter, UK.
Ruse, M., 2013. The Gaia Hypothesis: Science on a Pagan Planet. University of Chicago Press.
Schneider, S.H., Boston, P. (Eds.), 1992. Scientists on Gaia. MIT Press, Cambridge, MA.
Schneider, S.H., et al. (Eds.), 2004. Scientists Debate Gaia. MIT Press, Cambridge, MA.
Serafin, R., 1988. Noosphere, Gaia and the science of the biosphere. Environmental Ethics 10, 121.
Volk, T., 1997. Gaia's Body. Springer-Verlag, New York.

Galton, Sir Francis (1822–1911)

John CW Waller, University College, London, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 9, pp. 5860–5863, © 2001, Elsevier Ltd.

Abstract

An overview of the life, scientific interests, and historical legacy of the Victorian scientific polymath, Sir Francis Galton (1822–1911). Galton's chief scientific contributions stemmed from a dual fascination with heredity and quantification. He made significant contributions to the sciences of heredity and evolutionary biology – devoting much of his career to the statistical analysis of physical and mental traits – and he played a critical role in the formulation of the statistical procedures of correlation and regression analysis. In addition, he was active in advocating 'eugenics,' the practice of selective human breeding, a term he coined in 1883. Finally, Galton was responsible for several advances in the field of meteorology, including discovering and naming the anticyclone in 1862.

Francis Galton was born in Birmingham on February 16, 1822, the third son of Samuel Tertius Galton, a wealthy banker whose fortune was mainly derived from his father's musket manufactories. Francis was also the grandson of Erasmus Darwin, in whose scientific and evolutionist footsteps he was later to follow. However, he was originally destined for a career in medicine. Aged 16, at his father's behest, Francis enrolled as a medical student at the Birmingham Free Hospital, from where he moved to King's College, London to study medicine in 1839, and to Trinity College, Cambridge in 1840 to read mathematics. Having achieved only a pass degree at Cambridge, largely due to the first of the mental breakdowns that punctuated his career, Galton continued with his medical instruction despite a growing aversion to it. Coming into a substantial fortune on the death of his father in 1844, he immediately abandoned medicine and spent several indolent years 'hunting, fishing and shooting.' Then, in 1850, he led and financed an exploratory expedition to South-West Africa (modern Namibia). On his return, in 1852, Galton was lionized by the scientific establishment. And having achieved this entrée into the scientific community, gradually he established himself as one of Victorian Britain's most respected men of science. He was knighted in 1909.

Francis Galton is now remembered chiefly for his innovations in the field of statistics and for advocating the practice of eugenics. But his interest in both of these areas initially arose from an attempt to provide scientific evidence for the inheritance of human mental attributes. In fact, Galton devoted most of his professional energies to investigating varied aspects of heredity, a predilection largely inspired by reading Charles Darwin's (his half-cousin) *The Origin of Species* in 1859. Throughout his career, Galton's views of mind and morality were unequivocally materialist and evolutionary. In 1865, he published an article in two parts, under the heading 'Hereditary talent and character,' that were later expanded into his first book on the subject of mental heredity, *Hereditary Genius* (1869). In each of these publications he sought to show that the pedigrees of geniuses contain considerably more forebears of high social distinction than the lineages of those with mediocre parentage. Galton was aware that any positive results might instead suggest the importance of nepotism and unequal access to education. But on the basis of several ingenious – albeit less than convincing – arguments he insisted that for the highly intelligent neither education nor preferment are necessary for the fulfillment of their inborn intellectual potential. This permitted him to interpret any instance of relatedness among the eminent as being due to shared heredity.

Having presented detailed pedigrees for several hundred eminent men Galton calculated that 50 percent of illustrious individuals have at least one distinguished relation. On this basis he concluded that heredity is the primary determinant of intellectual ability. He also devoted a section of *Hereditary Genius* to estimating, in terms of the overall stock of hereditary intelligence, the 'comparative worth' of different nations and races. Galton reproduced the typical mid-Victorian conception of a racial hierarchy stretching from the Caucasian to the Negroid, although he argued that the ancient Athenians were appreciably more intelligent on average than any modern civilized nation.

During the 1870s, Galton continued his researches into human mental heredity. In 1874 he issued questionnaires to leading men of mid-Victorian science, hoping to shed light on the relative importance of 'nature and nurture' in their choice of vocation. Predictably, his data interpretation was prejudiced in favor of a hereditarian explanation. In 1875, although he had no conception of the modern distinction between monozygotic and dizygotic twins, Galton turned to the phenomenon of twinning as a means of delineating between heredity and environment. He collected the life histories of pairs of like-sexed twins and sought to determine if those that were similar at birth and in childhood became unlike upon leaving their similar home environments. He also asked if like-sexed twins that were dissimilar at birth subsequently became more similar in the shared home environment. In both cases Galton concluded that character had been formed prior to birth and had thereafter shown only limited flexibility in response to environmental pressures.

On the basis of his studies of heredity, Galton proposed the peremptory regulation of human breeding so as to maximize the biological quality of the nation. In 1869 he recommended the confinement of the socially undesirable in 'celibate monasteries or sisterhoods,' and the creation of incentives for the intellectual elites to reproduce more

frequently. The term 'eugenics' was coined in 1883 by Galton from the Greek *eugenes* meaning 'good in stock.' Yet, eugenics was widely considered impracticable during the 1860s and it was only during Galton's last decade that he began championing the idea. By then, foreign challenges to Britain's industrial and mercantile hegemony, and fears of an increasingly volatile urban underclass had rendered eugenics an appealing notion to many *fin-de-siècle* Britons. In 1908, after initial reluctance occasioned by his advanced age, Galton became President of the newly formed Eugenics Education Society.

Political considerations bulked large in Galton's advocacy of eugenics. Running through his work on heredity and eugenics, and becoming more pronounced with time, was a strongly elitist conviction that the vast majority of the lower classes were sexually profligate and irremediably inferior to the higher social classes. Accordingly, he expressed the fear that the lower classes were out-breeding the 'respectable' classes. Moreover, Galton's predilection for studying intelligence and his commitment to enhancing the reproductive rate of the most intelligent are symptomatic of his upbringing as a member of the Victorian professional middle class. This was a social stratum that set great store by intellectual accomplishment and produced many of the century's foremost social reformers, commentators and scientists: Galton unconsciously celebrated the core values and accomplishments of this important social group. In advocating eugenics he was also influenced by the widely held belief in the possibility of human perfectibility – of which his grandfather Erasmus Darwin had been a leading exponent – and by the potent mid-Victorian ideologies of human and social progress advanced most forcefully by his friend and colleague Herbert Spencer. Lastly, Galton's eugenic thought reflects the growing impetus among Victorian life-scientists for making their discipline practically useful to the state.

In the context of studying the inheritance of human mental and bodily traits, Galton emerged as an innovative statistician. His maxim was 'Whenever you can, count,' and his researches suggest a man with an almost obsessive compulsion to do so. In *Hereditary Talent* and *Hereditary Genius*, Galton became the first person to employ the Gaussian error curve in classifying human *mental* attributes. Although his use of statistics was at this stage rudimentary, he accomplished the important step of measuring – and focusing almost entirely upon – statistical variation. This aspect of the normal curve had not received much attention from his contemporaries (with the exception of the Belgian astronomer and statistician Adolphe Quetelet), many of his predecessors regarding variation from the mean as 'error' rather than a natural phenomenon worthy of investigation. All of Galton's major statistical insights stemmed from this concern for statistical spread as opposed to averages alone. In particular, his work was pivotal in the development of both the correlation co-efficient and regression analysis. Galton formulated a graphical method to measure correlation in 1886 by plotting the distributions of both variables, and two years later he devised an 'index of co-relation.' In fact, his computational procedures for 'co-relation' involved measuring the slope of the regression line and producing a measure of the regression coefficient. The mathematical underpinnings of correlation were developed more fully by his admirer and first biographer, Karl Pearson. From Galton's idea of regression, Pearson also developed a statistical system for the linear production of two continuous variables.

Galton's role in the development of regression analysis is more ambiguous. As a means of studying the mathematics of hereditary variation, during the late 1870s Galton began to breed sweet peas. He found that the F_1 generation peas were, on average, only a third as extreme in any given quality as the parental generation. Thus, through his breeding experiments, Galton had stumbled upon the statistical principle of regression. However, he always assumed that the phenomenon of regression related to a fundamental physiological process, and his explanation for it was biological rather than statistical. Galton asserted that each individual is composed of distinct hereditary units derived both from parents and from earlier ancestors. In each generation a random selection of the ancestral and parental hereditary material is made, such that only some of the total complement of hereditary units are actually expressed in each offspring. Galton believed that, from a purely probabilistic point of view, the entire hereditary material – stretching back several generations – would on average be close to the population mean. However, because only a random selection of hereditary units gain expression, it is possible for more extreme characteristics to be produced. Regression occurs when, as is likely, the more mediocre hereditary units regain supremacy in the following generation. This conception of the nature of heredity led to Galton's formulation of the 'ancestral law of heredity,' a law stipulating that any two parents together contribute approximately one-half of the total hereditary material of the offspring, the four grandparents one-quarter, and so on until the proportions become vanishingly small. Galton believed this law to describe the actual hereditary composition of each individual and the model formed the basis of his statistical treatment of heredity in *Natural Inheritance* (Galton, 1889). During the twentieth century, 'Galton's law' would be eclipsed by the advent of Mendelian genetics.

'Galton's law' may not have long survived the rediscovery of Mendel's work, but Galton has been credited with anticipating August Weismann's concept of the continuity of germplasm. Unlike Darwin and the majority of his contemporary life-scientists, from as early as 1865 Galton was largely dismissive of the notion of the inheritance of acquired characteristics. During the early 1870s he experimentally tested Darwin's theory of 'Pangenesis' – according to which bodily tissues throw off the germ cells from which offspring are produced – by transfusing blood among different varieties of rabbit. His predictable conclusion brought no pleasure to Darwin, and Galton's 1871 rejection of Pangenesis in *Nature* elicited from him a swift rebuke. Subsequently, Galton's resistance to the inheritance of acquired characteristics hardened. In the models of hereditary transmission he published during the first half of the 1870s, he claimed that the 'patent' hereditary units that become bodily structures only rarely return to the reproductive organs to be transmitted to progeny. In contrast, the 'latent' elements remain in the testes and ovaries ready to be transmitted to the next generation. In this rather circumscribed sense his hereditarian model did foreshadow Weismannism.

However, aside from the many technical inconsistencies between Galton's and Weismann's concepts of heredity, it is important to note that after 1875 Galton departed significantly from the position outlined above. He gradually came to appreciate that his case for the continuity of traits between parents and offspring was weakened if it were assumed that the latter inherited only the former's latent elements and very few of their patent elements. As a result, by the late 1880s Galton was writing of patent and latent elements as being equally heritable. Whilst he remained only mildly sympathetic to the notion of the inheritance of acquired characteristics, he had abandoned his earlier emphasis upon the near-absolute separation of bodily from reproductive cells.

Galton's dual fascination with heredity and quantification also generated an interest in anthropometry, the measurement of bodily morphology. During the 1870s, appalled that so little was known about the physique of the British people and fearful that a biological decline was already in progress, Galton sought to obtain anthropometric information on children from different social classes. Subsequently, he became a member and later chairman of a British Association committee investigating occupational disparities in size and weight. Throughout his involvement with this committee Galton was convinced that heredity was the chief determinant of anthropometric variation. As a means of procuring raw data from which to study hereditary transmission in humans, in 1884 he set up an elaborate anthropometric laboratory at the International Health Exhibition in London. Equipment was installed to measure the height, weight, breathing-capacity and, *inter alia*, the strength of pull of members of the public. Also measured were such qualities as keenness of sight, hearing, and color discrimination.

These last tests reflect Galton's interest in what he dubbed 'psycho-physics,' the scientific investigation of mental phenomena. Generalizing his earlier conviction that individual nurture is a subordinate force to heredity, he emphasized aspects of unconscious cerebral function in which the 'will' seems to play little if any role. In his 1883 *Inquiries into Human Faculty*, Galton described and, where feasible, quantified a range of unconscious and sensory phenomena, including number forms, visual imagery, color associations, visions, and sensory awareness among the blind. He also documented the results of his introspective investigations into the power of the individual will. In the course of these studies, Galton devised the method of 'word-association' with which to demonstrate his claim that very few thoughts or actions are ever the spontaneous product of the will but relate to desires, ideas and associations of which we have little conscious awareness.

In addition, Galton was a pioneer in the design and employment of composite photography. He hoped that through superimposing numerous photographic portraits on top of each other he would be able to discern the common hereditary features shared by families or groups. His main work in this area involved an attempt to discover the physiognomical indices of different forms of criminality from a series of photographs supplied by the Inspector of Prisons. To his disappointment, Galton found no evidence of distinctive criminal types. However, this interest in hereditary variation, measurement and their relevance to criminal investigations also culminated in his extensive study of fingerprints. Galton was not an innovator in this field but his several publications on the subject drew attention to the enormous value of fingerprints as a means of personal identification.

Although it is clear that most of Galton's scientific work stemmed from his early fascination with heredity, his scientific curiosity was nonetheless polymathic in breadth. Over the course of half-a-century's involvement with the Royal Geographical Society, he was able to make significant contributions to the expeditions of such explorers as R. Burton, J. Speke, and J. Grant, and he worked hard to place geography on a more scientific footing. As an active meteorologist he was instrumental in the construction of weather maps and was responsible for discovering the 'anticyclone' in 1862. Finally, he was an enthusiastic advocate of establishing scientific activity as an occupation of dedicated, professional and, if necessary, salaried, researchers. Alongside T. H. Huxley, John Tyndall, Joseph Hooker, and others, he campaigned vigorously to free science from the limitations imposed by religious orthodoxy. To this end, many of Galton's works carried thinly veiled attacks on theology and the Church, most strikingly his 1872 attempt to prove statistically that prayer is inefficacious.

In retrospect, there can be little doubt that Galton gave an important impetus to the scientific study of heredity and the development of psychometric testing. Despite the shortcomings later revealed in his own work, he was the first to subject human intellectual attributes to systematic numerical analysis. With the aid of W. F. R. Weldon and Karl Pearson, he was also intimately involved with establishing a statistical approach to the study of evolution. Perhaps Galton's most lasting contributions to science arose from his statistical breakthroughs of the 1870s and 1880s. Statistical procedures derived from his initial insights are now indispensable to both the natural and the social sciences. More practically, his private wealth enabled him to set up a Eugenics Research Office at University College London in 1904. He honored and extended this commitment upon his death (on January 17, 1911) when he created the Galton Professorship of Eugenics and named Karl Pearson as the Chair's first holder.

Any summary of Galton's legacy must recognize that eugenic ideas were implemented after his demise with horrific consequences, and that the hereditarian ideologies he promoted have often been used as means of social, sexual and racial oppression. In the wake of the Nazi atrocities of the 1930s and 1940s, and of the civil rights movements of the 1960s and 1970s, Galton's stock as an important scientific figure has declined rapidly. However, in the interests of a proper understanding of the past, it should be pointed out that to implicate Galton in the rise of twentieth-century eugenic practices is to over-personalize a complex series of events and set of national circumstances. Eugenic measures were linked to a wide range of social, economic and political factors against which Galton's advocacy is scarcely more than trivial. Furthermore, even in 1865 the notion of selective human breeding was far from novel: Galton's originality lay

only in buttressing the idea with apparently rigorous scientific research.

See also: Eugenics, History of; Nature-Nurture Controversy, History of; Psychometrics.

Bibliography

Forrest, D.W., 1974. Francis Galton: The Life and Work of a Victorian Genius. Paul Elek, London.

Galton, F., 1863. Meteorgraphica, or Methods of Mapping the Weather. Macmillan, London and Cambridge.

Galton, F., 1869. Hereditary Genius. Macmillan, London.

Galton, F., 1874. English Men of Science: Their Nature and Nurture. Macmillan, London.

Galton, F., 1883. Inquiries into Human Faculty and its Development. Macmillan, London.

Galton, F., 1889. Natural Inheritance. Macmillan, London.

Galton, F., 1908. Memories of My Life, second ed. Methuen, London.

Kevles, D.J., 1986. In the Name of Eugenics: Genetics and the Uses of Human Heredity. Penguin, Harmondsworth, UK.

Keynes, M. (Ed.), 1993. Sir Francis Galton, FRS: The Legacy of His Ideas. Macmillan, Basingstoke and London.

Pearson, K., 1914/1930. The Life, Letters and Labours of Francis Galton. Cambridge University Press, Cambridge, UK.

Stigler, S.M., 1986. The History of Statistics: The Measurement of Uncertainty Before 1900. The Belknap Press of Harvard University Press, Cambridge, MA.

Game Theory

Guillermo Owen, Naval Postgraduate School, Monterey, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

We study situations of conflict of interest, which we call 'games.' Both noncooperative and cooperative games are considered, and several types of solution concept are defined and analyzed. For noncooperative games, optimal strategies and equilibrium *n*-tuples are defined. For cooperative games, such concepts as the core, the Shapley value, and the bargaining set are treated.

Obviously, game theory is applicable to parlor games and warfare. Other applications of game theory to economics, political science, and evolutionary biology are given and several examples are worked out in detail.

Game Theory is the mathematical study of situations of conflict of interest. As such it is applicable to parlor games (hence its name), but also to military and economic situations, and, to a lesser extent, to situations in other social sciences. Game theory studies three general phases in the process of interaction: the choice of *strategies*, the formation of *coalitions*, and *bargaining* within coalitions.

As concerns ethics, game theory is useful as an arbitration technique for bargaining problems (Section Two-Person Cooperative Games) and, in distributive justice, for allocating the gains from cooperation (Section *n*-Person Games). Conversely, it can be used to design rules (e.g., assignment of weights for voting in a parliament whose members represent constituencies of different sizes) so that normal play of a game will lead to a 'fair' outcome.

Representation of Games

Games are represented in several ways. Among these are the extensive form, the *normal* form, and the *characteristic* function.

The Extensive Form

The extensive form of a game is a tree (graph) which exhibits the logical sequence of moves in a game, showing the order in which players move, the information they have when making a move, and, for chance moves, the associated probabilities.

Example 1 – Matching Pennies

Player I chooses either H (heads) or T (tails). Then player II, in ignorance of I's choice, chooses either H or T. Now, if both choose H, I pays \$3 to II; if both choose T, I pays \$1 to II. Otherwise, II pays \$2 to I.

This game is represented by the tree in **Figure 1**. The game starts at vertex A, which 'belongs' to player I. From here, I's choice leads to either B or C, both of which belong to player II. The shaded area joining these two vertices is used to represent II's lack of information: when he moves, he knows he is at one of these two vertices, but cannot distinguish between them. Technically, vertices B and C belong to the same *information set*. From either of these, II has two alternatives, H and T. After this, the game terminates at one of the vertices, D, E, F, or G. A vector with two components is assigned to each of the terminal vertices: these are the payoffs. For example, the $(-3, +3)$ assigned to vertex D means that player I loses \$3 and II wins \$3.

Example 2

Consider the same game as in Example 1, with the sole change that player II is allowed to make his decision after seeing what player I has chosen. All else is as in Example 1.

This game is shown in **Figure 2**. The only difference lies in the fact that vertices B and C are not in the same information set; in this game, player II can distinguish between them.

In Example 2, note that all information sets consist of a single vertex. In such a case, the game is said to have *perfect information* (a situation in which a player knows all moves that have been made). The game in Example 1 does not have perfect information, nor does the one in the following Example 3.

Example 3 – Rudimentary Poker

Chance gives player I either a high card (H) or a low card (L), with probabilities of 0.2 and 0.8, respectively. Player I, seeing the card, can bet (B) or pass (P).

If player I passes, the game terminates immediately. Player II pays \$1 to I if I has H, while I pays \$1 to II if I has L. If player I bets, then II can call (C) or fold (F). If player II folds, then he pays \$1 to player I. If II calls, then II pays \$3 to I if I has H, and I pays \$3 to II if I has L. Recall that, in deciding to call or fold,

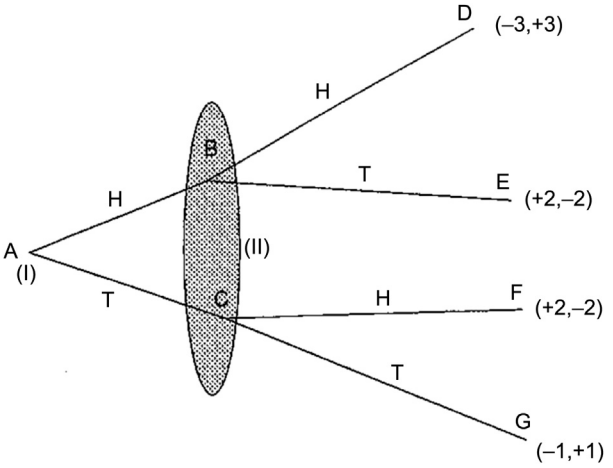

Figure 1 The extensive form of the game in Example 1.

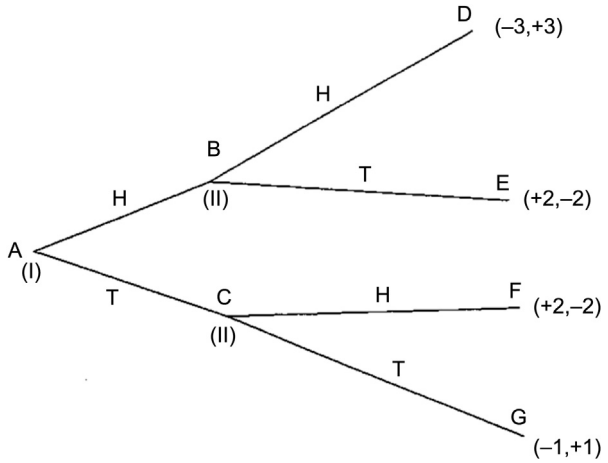

Figure 2 The extensive form of the game in Example 2.

player II knows whether I has bet or passed, but not whether she is holding H or L.

This game is represented in Figure 3. Note that at the chance move, the probabilities are specified; we assume that both players know the probabilities. Any attempt to change the probabilities (e.g., by stacking the deck) is considered cheating and is illegal. Note also that two vertices (E and F) are in one information set; hence player II's information is imperfect.

The Normal Form

While the extensive form representation of a game gives us a very good description of the actual play of a game, it is easier to analyze the game (as regards maximizing winnings) with the normal form. This is based on the notion of a *strategy* (a rule that tells a player what to do at each position).

A strategy for a given player is a comprehensive set of rules telling a player which alternative to choose in each information set. Note that a strategy may be good or bad, and that the rules that make up a strategy do not have the force of laws: a player may disregard such a rule if he/she wishes so.

To go back to Example 1, we see that player I has one information set (vertex A) with two alternatives. Thus he/she has two strategies, H and T. Similarly, player II, with one information set (B and C) and two alternatives there, has two strategies, also called H and T.

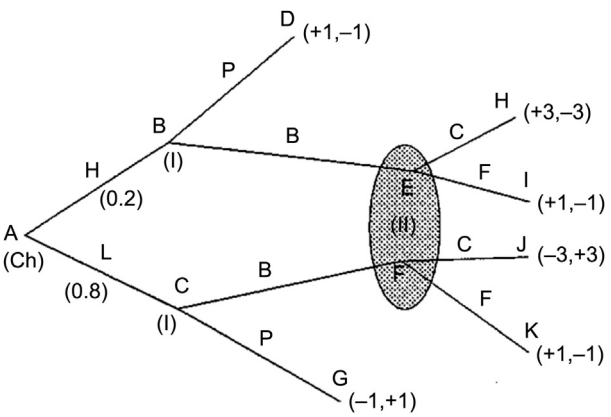

Figure 3 The extensive form of the game in Example 3.

In Example 2, player I has once again two strategies, H and T. Player II, however, has two information sets, with two alternatives in each. Thus player II has four strategies: HH (always call H), HT (call H at B, call T at C), TH (call T at B, call H at C), and TT (always call T).

In Example 3, player I has two information sets, with two alternatives at each, and thus four strategies: BB (always bet), BP (bet at B, pass at C), PB (pass at B, bet at C), and PP (always pass). Player II has only the two strategies, C and F.

In a given game, suppose each player has chosen a strategy. If the game has no chance moves, several players' strategies will entirely determine the course of play, and hence the outcome and payoff of the game. In a game with chance moves, the outcome will also depend on the chance moves. By hypothesis, however, the probabilities at the chance moves are well determined. Hence a choice of strategies will determine, up to a probability distribution, the outcome of the game. It is thus possible to calculate the mathematical expectation of the payoff. In other words, the expected payoff of the game is a function only of the strategies used by the players. This function, usually shown as a table or matrix, is the *normal form* of the game.

The normal form of the game in Example 1 is the matrix

$$\begin{array}{c} \\ \mathrm{H} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{cc} \mathrm{H} & \mathrm{T} \end{array} \\ \begin{pmatrix} (-3,\ 3) & (2,\ -2) \\ (2,\ -2) & (-1,\ 1) \end{pmatrix} \end{array}$$

By convention, in two-person games, the normal form is a matrix, with rows representing player I's strategies and columns representing those of player II.

The game of Example 2 is represented by

$$\begin{array}{c} \\ \mathrm{H} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{HH} & \mathrm{HT} & \mathrm{TH} & \mathrm{TT} \end{array} \\ \begin{pmatrix} (-3,\ 3) & (-3,\ 3) & (2,\ -2) & (2,\ -2) \\ (2,\ -2) & (-1,\ 1) & (2,\ -2) & (-1,\ 1) \end{pmatrix} \end{array}$$

It may be noted that, at each cell of the matrix, the sum of the two entries is equal to 0. The reason for this is that, in each of our examples, money is transferred directly from one player to the other without any external credits or debits for either player. Such games (where one player loses whatever the other one wins) are known as *zero-sum games* (one in which the sum of payoffs to all players is always zero). In a two-person zero-sum game, it suffices to show player I's payoff, since the payoff to player II is always its negative. Thus, the normal form of a finite two-person zero-sum game is a matrix whose entries are, by convention, the payoffs to player I.

In the game of Example 3, suppose player I chooses BB (always bet) while player II chooses C. In that case, player I has a 0.8 probability of losing \$3, and a 0.2 probability of winning \$3. Thus I's expected payoff is $0.8(-3) + 0.2(3) = -1.8$. Other payoffs are calculated similarly, giving us the normal form

$$\begin{array}{c} \\ \mathrm{BB} \\ \mathrm{BP} \\ \mathrm{PB} \\ \mathrm{PP} \end{array} \begin{array}{c} \begin{array}{cc} \mathrm{C} & \mathrm{F} \end{array} \\ \begin{pmatrix} -1.8 & 1 \\ -0.2 & -0.6 \\ -2.2 & 1 \\ -0.6 & -0.6 \end{pmatrix} \end{array}$$

We postpone discussion of the characteristic function until treatment of cooperative games with three or more players.

Two-Person Zero-Sum Games

As just discussed, these games represent a closed system: whatever one player wins, the other player loses. In the finite case, such a game is represented by a matrix. If player I has m strategies, while II has n strategies, then the normal form is an $m \times n$ matrix A, where a_{ij}, the entry in the ith row and jth column, is the payoff to player I if she uses her ith strategy while II uses his jth strategy. The game is a *matrix game* (a finite two-person zero-sum game in normal form).

Saddle Points

Now, player I wishes to maximize, and player II to minimize, the payoff. For Example 1, the normal form (abbreviated) is

$$\begin{array}{c c} & \begin{array}{cc} \text{H} & \text{T} \end{array} \\ \begin{array}{c} \text{H} \\ \text{T} \end{array} & \begin{pmatrix} -3 & 2 \\ 2 & -1 \end{pmatrix} \end{array}$$

and it may be seen that there is no strategy that can be considered truly 'good': whatever one player chooses, the other can win by simply 'guessing correctly.' Thus if I chooses H, then II can win by choosing H; but if so then it would be better for I to choose T; but then it would be better for II to choose T; but then it would be better for I to choose H; etc. Thus it is not clear what should be done in this game.

Contrast this with the game of Example 2, whose matrix is

$$\begin{array}{c c} & \begin{array}{cccc} \text{HH} & \text{HT} & \text{TH} & \text{TT} \end{array} \\ \begin{array}{c} \text{H} \\ \text{T} \end{array} & \begin{pmatrix} -3 & -3 & 2 & 2 \\ 2 & -1 & 2 & -1 \end{pmatrix} \end{array}$$

In this game player II has a 'sure thing' strategy in HT. Whatever player I chooses, II does at least as well with HT as with any other strategy. Thus II should always use HT. Given that II will choose HT, player I should choose T as this minimizes her losses.

Example 4

Consider the following matrix game:

$$\begin{array}{c c} & \begin{array}{ccc} \text{A} & \text{B} & \text{C} \end{array} \\ \begin{array}{c} \text{D} \\ \text{E} \\ \text{F} \end{array} & \begin{pmatrix} 4 & 1 & -3 \\ 5 & 2 & 4 \\ -1 & 0 & 6 \end{pmatrix} \end{array}$$

In this game, there is no 'sure thing,' in the sense of a strategy (for either player) which is always better than anything else. Rather, we analyze the game as follows.

If player I is cautious, then, in using any strategy, she will worry about the worst that can happen to her. This 'worst' is represented by the smallest entry in each row, namely -3 for D, -2 for E, and -1 for F. Thus caution suggests using row E.

Suppose, however, that player II figures out what player I is thinking (i.e., using row E). Then he would use column B as this minimizes the payoff in row E. But, if II is using B, this is all the more reason for I to use E (as this maximizes the payoff in column B). Thus, the pair of strategies (row E, column B) presents a very stable type of equilibrium. The value of 2 at their intersection, which is both the smallest entry in its row and the largest entry in its column, is a *saddle point* (an entry (in a matrix) which is the smallest in its row and the largest in its column).

Definition 1

In a matrix game A, the entry a_{kh}, is a saddle point if, for each $i = 1, \ldots, m$,

$$a_{kh} \geq a_{ih}$$

and, for each $j = 1, \ldots, n$,

$$a_{kh} \leq a_{kj}$$

If a matrix game has a saddle point a_{kh}, then this saddle point is a very satisfactory *solution* for the game. In fact, by using the kth row, player I expects to win at least a_{kh}; by using the hth column, player II expects to lose no more than this same amount. We say then that a_{kh} is the *value* (in a matrix game, a quantity which the maximizer can ensure herself of winning; the minimizer in turn can ensure himself against losing more than that quantity) of the game, and the two strategies (row k, column h) are *optimal strategies* (a strategy which guarantees the value). In case of more than one saddle point, the following theorem is helpful:

Theorem 1

Let the matrix game A have saddle points at a_{kh} and at a_{ij}. Then $a_{ij} = a_{kh}$, and, moreover, the entries a_{kh} and a_{ij} are also saddle points.

The importance of Theorem 1 lies in the fact that (1) players have no incentive for trying to change from one saddle point to another, and (2) if the two players try for two different saddle points, the outcome will nevertheless be a saddle point.

Mixed Strategies

While saddle points represent a very satisfactory solution for matrix games, the fact remains that, in general, such games need not have saddle points. (Note Examples 1 and 3 for games with no saddle points.) In such a case, it seems that the winner will be the one who can think one step further than the other player. ("He bluffed last time, so he will expect me to call this time, and therefore he will not bluff, so I should fold if he bets. But perhaps he knows that I am thinking this, and will bluff again, so I should call his bet. But perhaps he is thinking this, too, so….") Unfortunately, thinking two steps (instead of one) further ahead than the other might well lead to defeat. Rather than this attempt at outguessing, mathematicians have developed the idea of a *mixed strategy* (a randomization scheme for choosing among a player's pure strategies).

Consider once again Example 3. The first thing to notice is that there is no point in using either row 3 (PB) or row 4 (PP). In fact, BB is always at least as good (for player I) as PB, and sometimes better, while BP is always at least as good as PP, and sometimes better. This says quite simply that player I should always bet on a high card. The question is whether she should

bet (bluff) or pass on a low card. As mentioned above, it is important for player I to keep her opponent guessing.

To keep player II from outguessing her, player I might well decide to choose her strategy at random, i.e., in case she receives L, she might flip a coin and bluff if the coin comes up heads, or pass if the coin comes up tails. In fact this is a possibility, but it effectively seems to take away any 'free choice' that player I might have. Moreover, it is not clear why bluffing exactly half the time should be particularly good.

It turns out that, in this game, player I's best choice – in a way to be explained shortly – is to bluff one-eighth of the time, i.e., always bet if she has H, and, with L, bet with probability 0.125. Note that, in this way, she does not really give up her free choice – she chooses the probability, 0.125, in a rational way, and then lets a randomization device tell her what to do. More precisely, she chooses among her four strategies through a random mechanism which assigns probabilities 0.125, 0.875, 0, and 0 to them. This is a mixed strategy. We will use the term *pure strategy* for the rows and columns of the matrix (i.e., for what we have up to now called strategy).

Definition 2

In an $m \times n$ matrix game, a mixed strategy for player I is a randomization scheme over that player's pure strategies, or equivalently, a vector $\mathbf{x} = (x_1, \ldots, x_m)$ with nonnegative components which add up to 1. Similarly, a mixed strategy for player II is a vector $\mathbf{y} = (y_1, \ldots, y_n)$ with nonnegative components that add up to 1.

Essentially, the rationale for mixed strategies lies in the fact that player II might be able to reproduce (in some way) player I's thinking process and thus might be able to guess I's mixed strategy. However, II will not know what player I actually does since this depends on a random mechanism. Similarly, I can guess II's mixed strategy but not his actual choice of action.

Assuming player I uses the mixed strategy $\mathbf{x}$, while II uses $\mathbf{y}$, then under the assumption of independence (these two randomization schemes are carried out independently of each other), the expected payoff will be

$$\sum_{i=1}^{m} \sum_{j=1}^{n} x_i a_{ij} y_j$$

or, in matrix notation, $\mathbf{x}^{\mathrm{T}} A \mathbf{y}$.

Now, player I will try to choose $\mathbf{x}$ (among all her mixed strategies) so as to maximize $\mathbf{x}^{\mathrm{T}} A \mathbf{y}$, while II will try to choose $\mathbf{y}$ so as to minimize this same product. We then define saddle points in mixed strategies in terms of the expected payoff, i.e.,

Definition 3

A pair of mixed strategies $\mathbf{x}^*$ and $\mathbf{y}^*$ form a saddle point for game A if, for any mixed strategy $\mathbf{x}$ of player I,

$$\mathbf{x}^{*\mathrm{T}} A \mathbf{y}^* \geq \mathbf{x}^{\mathrm{T}} A \mathbf{y}^*$$

and, for any mixed strategy $\mathbf{y}$ of player II,

$$\mathbf{x}^{*\mathrm{T}} A \mathbf{y}^* \leq \mathbf{x}^{*\mathrm{T}} A \mathbf{y}$$

In Example 3, the pair of strategies $\mathbf{x}^* = (0.125, 0.875, 0, 0)$ and $\mathbf{y}^* = (0.5, 0.5)$ is a saddle point, with $\mathbf{x}^{*\mathrm{T}} A \mathbf{y}^* = -0.4$. To see this, note that, against either one of player II's pure strategies, $\mathbf{x}^*$ gives an expected payoff of -0.4. Since any mixed strategy of player II simply chooses one or the other of these two with certain probabilities, it follows that the expected payoff, if I uses $\mathbf{x}^*$, will be -0.4, no matter what mixed strategy II uses. Similarly, against player I's four pure strategies, $\mathbf{y}^*$ gives expected payoffs of -0.4, -0.4, -0.6, and -0.6, respectively. Since none of these gives I a better payoff than -0.4, none of her mixed strategies can give a better expectation. This quantity -0.4 is the *value* of the game.

The Minimax Theorem

The question obviously arises as to the existence of saddle points in mixed strategies. The following theorem, due to J. von Neumann, is probably the most important in game theory.

Theorem 2 (The Minimax Theorem)

Any matrix game has at least one saddle point in mixed strategies.

If $(\mathbf{x}^*, \mathbf{y}^*)$ is a saddle point for game A, then the strategies $\mathbf{x}^*$ and $\mathbf{y}^*$ are said to be *optimal strategies*, while the quantity

$$v = \mathbf{x}^{*\mathrm{T}} A \mathbf{y}^*$$

is the *value* of the game. There may be several saddle points, but a theorem similar to Theorem 1 guarantees that the value is unique. A *solution* of the game is a triple $(\mathbf{x}^*, \mathbf{y}^*, v)$.

It may be that the game has a saddle point in pure strategies. If so, Theorem 2 is still true because a pure strategy can always be thought of as a (degenerate) mixed strategy: one that assigns probability 1 to one of the pure strategies, and 0 to all others. Thus, the game of Example 2 has the optimal strategies $\mathbf{x}^* = (0, 1)$ and $\mathbf{y}^* = (0, 1, 0, 0)$ and value $v = -1$.

For Example 1, the reader can verify that the solution is $\mathbf{x}^* = (0.375, 0.625)$, $\mathbf{y}^* = (0.375, 0.625)$, and $v = 0.125$.

While the minimax theorem guarantees the existence of optimal strategies and a value, the problem of computing these remains. There are two well-known techniques for this. One is based on linear programming: the optimal strategies are solution vectors for a pair of mutually dual linear programs. The other technique, known as *fictitious play*, assumes that two players play the same game a large number of times; each keeps track of his or her opponent's past choices, and at each iteration, uses the pure strategy which will do best against the opponent's empirically observed strategy. A theorem due to J. Robinson shows that under such assumptions, the empirically used strategies of both players will converge to optimality.

Infinite Games

Where each player has an infinite number of pure strategies, it is still possible to define mixed strategies and optimality. Unfortunately, the minimax theorem will not always hold for such games, and where it holds, computation of the optimal strategies may be quite complicated. We will not discuss such games here.

General-Sum Noncooperative Games

As opposed to Section Two-Person Zero-Sum Games, we consider here games in which either the sum of the payoffs is

not zero or there are more than two players. For these games, cooperation may or may not be allowed. We discuss first the noncooperative case.

The number of players may be any integer $n \geq 2$; we will label them 1, 2,…, n, and let N be the set of these players.

Nash Equilibria

In normal form, each player, j, has a strategy set which we denote by $\mathbf{\Gamma}_j$. A game is represented by n payoff functions π_j, defined for n-tuples of strategies. Thus, if $\sigma_1,\ldots, \sigma_n$ are strategies for players 1,…, n, respectively, then

$$\pi_j(\sigma_1, \ldots, \sigma_n)$$

is the payoff to player j $(1 \leq j \leq n)$ for this choice of strategies. As usual, we assume each player wishes to maximize his payoff.

In this case, the analogue to the saddle point is known as a *Nash equilibrium* (after J. Nash) (an n-tuple of strategies such that no player can gain by a unilateral change). Specifically,

Definition 4

For an n-person game with payoff functions π_j, the n strategies $\sigma_1^*, \sigma_2^*, \ldots, \sigma_n^*$ form a Nash equilibrium if, for each player j $(1 \leq j \leq n)$, and each strategy $\tau \in \mathbf{\Gamma}_j$,

$$\pi_j\left(\sigma_1^*, \ldots, \sigma_j^*, \ldots, \sigma_n^*\right) \geq \pi_j(\sigma_1^*, \ldots, \tau, \ldots, \sigma_n^*)$$

(On the right-hand side of this last inequality, it is understood that only the single term σ_j^* has changed, to be replaced by τ.) Thus, no player can gain by a unilateral change of strategies.

Example 5 – The Battle of the Sexes

Player I (a woman) and player II (a man) wish to go to either a hockey game (G) or the ballet (B). Player I would prefer the game; II prefers the ballet. However, both wish to be together and will derive no utility from going to either event alone. We represent this game by

$$\begin{array}{cc} & \begin{array}{cc} \text{G} & \text{B} \end{array} \\ \begin{array}{c} \text{G} \\ \text{B} \end{array} & \begin{pmatrix} (3, 1) & (0, 0) \\ (0, 0) & (1, 4) \end{pmatrix} \end{array}$$

where, as usual, player 1 chooses the row and II the column. In each case, the first component is the payoff to I, and the second the payoff to II. It may be seen that this game has two Nash equilibria, consisting of the strategy pairs (G, G) and (B, B).

Example 6 – The Prisoner's Dilemma

In this game, due to A.W. Tucker, players I and II are prisoners, suspected of a particular felony. The prosecutor, having no hard evidence, must hope that they will confess. He therefore offers each of them a deal.

If neither confesses, both will go to jail on minor charges. If both confess, both will receive medium sentences. If only one of the two confesses, then the one that confesses will be given a suspended sentence, and the other will receive a severe sentence.

For this game, the prisoners both have the strategies confess (C) or do not confess (D). (The prosecutor is not considered a player.) The game can then be represented by

$$\begin{array}{cc} & \begin{array}{cc} \text{C} & \text{D} \end{array} \\ \begin{array}{c} \text{C} \\ \text{D} \end{array} & \begin{pmatrix} (-5, -5) & (0, -10) \\ (-10, 0) & (-1, -1) \end{pmatrix} \end{array}$$

This game has the unique equilibrium (C, C). The reason for this is that for each player, C is always better than D, *whatever the other player does*. Hence both will, presumably, confess. (This although both will be better off if neither confesses!)

Example 7 – Cournot Oligopoly

Assume there are n producers of a homogeneous good. Each firm, i, chooses the quantity q_i it will produce. The market price p of the good will depend on the total quantity produced by the n firms. Each firm wishes to maximize its profits, defined as revenue minus costs, where revenue is the product pq_i.

In the simplest case, we assume that the inverse demand function is linear, given by $p = A - Q$, and that each firm's costs are zero. Then firm i's profits are $\pi_i = (A - Q)q_i$, where $Q = \sum_j q_j$.

For this game, an equilibrium (here a Cournot–Nash equilibrium) is obtained by differentiation, maximizing each function π_i with respect to the corresponding q_i. This equilibrium gives us $q_i = A/(n + 1)$ for each firm. All firms produce the same amount, with total production $Q = An/(n + 1)$, and market price $p = A/(n + 1)$. Profits for each firm will be $\pi_i = A^2/(n + 1)^2$.

In all three of Examples 5, 6, and 7, Nash equilibria (in pure strategies) exist. In general this is not so. It turns out, however, that, when mixed strategies are allowed, equilibria do exist. The following theorem is due to Nash.

Theorem 3

Any finite n-person game will have at least one equilibrium n-tuple of mixed strategies.

This is an existence theorem, based on a fixed-point argument, and calculation can be quite complicated. For two players, a clever technique due to C.E. Lemke and A. Howson is frequently used; for three or more, an approximation technique due to H. Scarf is extremely complicated but is the only available method.

More troublesome than the problems with computation of Nash equilibria is the fact that these tend to exhibit pathologies which never occur with matrix games. For one thing, Example 5 shows that no analogue of Theorem 1 holds, even for two-person general-sum games. The given game has two equilibria, but payoffs at both equilibria are different; thus, there will be attempts by the two players to steer the game toward one or the other of the two equilibria. Moreover, if player I tries for one equilibrium, say (G, G), while II tries for the other equilibrium (B, B), the result will be the 'disaster' (G, B). There seems to be no way to choose between the two equilibria. A third equilibrium, in mixed strategies, is very unsatisfactory.

Even in cases where there is only one Nash equilibrium, as in Example 6, this equilibrium may be unsatisfactory. In fact, both players will do much better if they both employ their nonequilibrium strategy, D. There is no way that the theory of Nash equilibria can solve this problem.

Two-Person Cooperative Games

Where cooperation is permitted, much of the interest is centered on the bargaining process that takes place prior to cooperation. This is best studied in the two-person case, where the distraction of alternative coalitions is avoided.

Nash's Bargaining Model

In this model, also due to Nash, two players are to divide a 'pie' (some asset). If they can come to an agreement, this will take effect. Otherwise, neither will receive anything.

A game such as this is represented by a set S in two-dimensional space, the set of *feasible points*: if the point with coordinates (u, v) lies in set S, then players I and II can, if they so agree, obtain utilities of u for player I and v for II. It is usually assumed that S is closed and bounded above (so that infinite amounts of utility are impossible). It is also assumed that S is convex (this is a realistic assumption since, given any two outcomes, the players are allowed to randomize between them). Apart from this, a *conflict point* (u_0, v_0) is given: in case the two players cannot agree, they will receive u_0 and v_0, respectively. The conflict point will also lie in the set S, since the two players may, if they wish, 'agree to disagree.' We wish, now, a *general bargaining rule* which will assign, to each such problem, a feasible point (u^*, v^*) as a solution.

Nash lays down four axioms that the bargaining rule should satisfy. These axioms, explained further on, are Paretooptimality (N1), independence of irrelevant alternatives (N2), invariance under linear transformations of utility (N3), and symmetry (N4).

N1 states that if point (u^*, v^*) is the solution to a problem with feasible set S, then there can be no point (u, v) in S such that either $u > u^*$, $v \geq v^*$, or $u \geq u^*$, $v > v^*$.

N2 is a rather technical axiom which states, essentially, that if point (u', v') is chosen rather than (u'', v'') for a given problem, then enlarging the feasible set (by adjoining new alternatives), and keeping the same conflict point, can never lead us to choose (u'', v'') over (u', v').

N3 is based on the idea that each individual's utility is subjective, and that there is no interpersonal comparison of utility. Moreover, each individual's utility scale is subject to linear transformations (much in the way that temperature can be measured either in the Celsius or in the Fahrenheit scale). Thus the same real-life problem has many different mathematical representations. The axiom states that the solution point should have the same real-life interpretation for all representations of the same real-life problem.

N4 states, essentially, that if the problem is symmetric (in the sense that whenever a point (u, v) is feasible, then so is (v, u), and that $u_0 = v_0$), then the solution point must satisfy $u^* = v^*$. We then obtain the following.

Theorem 4

There is a unique bargaining rule which satisfies Nash's four axioms. This rule chooses the point (u^*, v^*) so as to maximize the product

$$g(u, v) = (u - u_0)(v - v_0)$$

of the increments of the utilities, subject of course to the constraint that the point must lie in S (and, in fact, in the part of S 'above and to the right' of the conflict point).

Example 8

Players I and II are to divide the sum of \$100. Player I's utility is linear with money, i.e., her utility for x dollars is equal to x. On the other hand, player II's utility for y dollars is only $y^{1/2}$. If they cannot come to an agreement, neither will receive anything.

In this case, the conflict point is (0, 0). If player I receives x dollars, then her utility is $u = x$, while II's utility is $v = (100 - x)^{1/2}$. Thus, in an agreement, u and v satisfy the equation

$$u + v^2 = 100$$

We look then for the point which maximizes the product uv. Some calculus tells us that the maximum is at $v^* = 33.3^{1/2}$ and $u^* = 66.7$. Thus player I should receive \$66.7, while II receives only \$33.3. While this division looks strange, the point is that, because of II's utility function, the \$33.3 which he gets represents more utility to him than the \$66.7 which he does not get. This gives a bargaining advantage to player I, and she can use it to get a larger share of the \$100.

Threats

One problem with the model above lies in the determination of the conflict point. In Example 8, the situation is deliberately kept simple: failure to agree reduces each player to the status quo. In general, of course, the situation can be much more complicated. Consider a situation in which player I is a law-abiding citizen, walking along a lonely street, while player II is an armed robber. The two will 'bargain' over I's wallet, and, if no agreement is reached, player I may find herself much worse off than at the status quo. Player II, by wielding a gun, has changed the conflict point from (0, 0) to, say, $(-500, -10)$. At the new conflict point player I loses a lot; II also loses something as he would prefer to get the wallet without need for violence.

Some further points should be made here: (1) player II may have several different threats; (2) player I may herself have some available threat(s); and (3) in any case, it is not clear that the players will carry out their threats.

To take threats into consideration, Nash modified his model. In this new model, each player has finitely many threats (threat strategies) (a strategy to be used in case of no agreement). If player I uses her strategy σ_i, while II uses his strategy τ_j, they receive utilities of a_{ij} and b_{ij}, respectively. Thus the game is represented by two matrices A and B (payoffs to the two players in case of conflict) and a set S of feasible points. It is assumed that all points (a_{ij}, b_{ij}) lie in S. S might, however, be much larger than the convex hull of these conflict points.

Nash now assumes that the bargaining will go as follows:

1a. Player I chooses a (mixed) threat strategy, **x**.
1b. Player II, in ignorance of **x**, chooses a threat strategy, **y**.
2. The two players, knowing both **x** and **y**, bargain (look for a point of S on which they might agree). If they cannot agree, then they *must* use their threat strategies, obtaining (expected) payoffs of $\mathbf{x}^{\mathrm{T}}A\mathbf{y}$ and $\mathbf{x}^{\mathrm{T}}B\mathbf{y}$, respectively.

Applying Nash's previous model to this last bargaining step, we come to the conclusion that players I and II will choose that point $P^* = (u^*, v^*)$ which maximizes the product,

$$\left(u - \mathbf{x}^{\mathrm{T}}A\mathbf{y}\right)\left(v - \mathbf{x}^{\mathrm{T}}B\mathbf{y}\right)$$

subject of course to the constraints that $(u, v) \in S$ and $u \geq \mathbf{x}^{\mathrm{T}}A\mathbf{y}$, $v \geq \mathbf{x}^{\mathrm{T}}B\mathbf{y}$.

Since P^* depends on the threat strategies **x** and **y**, it follows that player I should choose **x** so as to maximize u^*, while II chooses **y** so as to maximize v^*. Now (u^*, v^*) lies on the Paretooptimal ('northeast') boundary of S, which means that as u^* increases, v^* decreases, and vice versa. The situation then is very similar to that for matrix games: there exist *optimal threat strategies* and a *bargaining value* for such games. Computation of these is generally based on the solution of a sequence of related matrix games, each of the form $A - kB$.

It should be noted that this model assumes that players *will* use their threat strategies if no agreement is reached. It is not clear just how this can be enforced. No generally accepted theory is currently available to deal with unenforceable threats.

n-Person Games

We consider now cooperative n-person games, where $n \geq 3$. The point here is that players will typically have several alternative coalitions. Thus bargaining within a coalition must be made with these competing alternatives in mind.

The Characteristic Function

The theory developed by J. von Neuman and O. Morgenstern assumes a finite set $N = \{1, 2, \ldots, n\}$ of players. A *coalition* is any nonempty subset of N (including N itself and all the one-player sets). An assumption is now made that, if a coalition S forms (or even if its members merely discuss this idea), it must consider the possibility that the remaining players – the complementary coalition $N - S$ – will form. Thus S will find itself playing a two-party game against $N - S$. Still another assumption says that utility is directly transferable (at a one-for-one rate) between members of a coalition, so that all that matters at this point is how much total utility S will gain. (Side payments within S will then adjust the several members' shares.) This is generally known as the TU (*transferable utility*) case.

If the n-person game is zero sum (the sum of payoffs to all players is zero) or at least constant sum, then the game between S and $N - S$ can be solved as in Section Two-Person Zero-Sum Games, with S as maximizer and $N - S$ as minimizer. Let $v(S)$ be the value of this game.

If the n-person game is not constant sum, then the two-party game will be a bimatrix game as in Examples 5 and 6. In this case the situation is not quite as clear as before; nevertheless, we will define $v(S)$ as the largest amount (of utility) which S can guarantee for its members *no matter what* $N - S$ does.

We call $v(S)$ the *worth* of coalition S. The function which assigns to each subset S of N its worth $v(S)$ is the *characteristic function* of the game.

Example 9

In a game with three players, there is a trunk full of precious objects, valued at 1000 units. The trunk is too heavy for one person to lift, but any two of the three players can lift it together, and they will then be in possession of the trunk.

In this case, we have $N = \{1, 2, 3\}$, and

$$\begin{aligned} v(\{i\}) &= 0 \quad \text{for } i = 1, 2, \text{ or } 3 \\ v(S) &= 1000 \quad \text{for any } S \text{ with two or three players} \end{aligned}$$

By convention, we also have $v(\emptyset) = 0$ ($\emptyset$ being the empty set).

Example 10

A woman (player I) has a horse that she wishes to sell. She has no use for it. There are two possible buyers, players II and III. Player II values the horse at 1000 units, while player III values it at 500 units.

In this case, a single player can produce no utility. Neither can the coalition $\{2, 3\}$. On the other hand, the coalition $\{1, 2\}$ can produce a total of 1000 units of utility, to be divided among them as they wish. (If player I sells the horse to II, I's profit is then the price she receives for the horse, while II's profit is 1000 minus the price he pays. Their total profit is then 1000.) Similarly, $\{1, 3\}$ can produce 500 units of utility. This game has the characteristic function

$$\begin{aligned} v(\{i\}) &= 0 \quad \text{for } i = 1, 2, \text{ or } 3 \\ v(\{1,2\}) &= 1000 \\ v(\{1,3\}) &= 500 \\ v(\{2,3\}) &= 0 \\ v(\{1,2,3\}) &= 1000 \end{aligned}$$

Given the characteristic function of the game, we look for a way to distribute the amount $v(N)$ which is the total utility available to the set of n players. Assuming that no player will accept less than what he can guarantee in the worst of cases, we define an *imputation* as a vector **x** with n components such that

$$\begin{aligned} x_i &\geq v(\{i\}) \quad \text{for all } i \in N \\ \sum_{i \in N} x_i &= v(N) \end{aligned}$$

Thus, imputations are possible payoff vectors – allocations to the n players of the total amount $v(N)$.

In Examples 9 and 10, notice that the imputation set in each case is the set of all vectors with three nonnegative components adding to 1000. In this, the games are similar, but as we shall see, they are quite dissimilar in other ways.

The Core

Given a game in characteristic function form, **v**, we know its imputations. We still require some way to determine which imputations are in some way 'better' or 'more stable' than

others. We base this on *domination*. An imputation **x** dominates another, **y**, if there is some coalition S such that (1) all members of S prefer **x** to **y**, and (2) S is strong enough to obtain **x**. Mathematically this means that $x_i > y_i$ for all $i \in S$, and $\sum_{i \in S} x_i \leq v(S)$. Then, the *core* of a game is the set of all undominated imputations.

In Example 9, the core is empty. For, let $\mathbf{y} = (y_1, y_2, y_3)$ be an imputation. The three components of **y** add to 1000, and thus at least one of them is positive. Assume that $y_1 > 0$. Then $\mathbf{x} = (0, y_2 + y_1/2, y_3 + y_1/2)$ dominates **y** (with $S = \{2, 3\}$). Similarly, all imputations are dominated, i.e., the core is empty.

In Example 10, the core is nonempty; it consists of all imputations of the form $(x_1, 1000 - x_1, 0)$, where $500 \leq x_1 \leq 1000$. This says that player I will sell the horse to II; the price cannot be more than 1000 since the horse is only worth 1000 units to II. On the other hand, the price must be at least 500 units as, below 500, player III will bid so as to raise the price.

The Shapley Value

While the core is an intuitively appealing solution concept for n-person games, we note that in many cases it will be empty. It also fails to take certain possibilities into account. In Example 10, e.g., the two buyers might collude so as to bring the price of the horse below 500 units: player III realizes he will be outbid anyway, so he agrees to stay out of the bidding for a price (a side payment from II), and so an outcome such as (300, 600, 100), although not in the core, is not out of the question. For these reasons, theorists have looked at alternative solution concepts. Perhaps the best known one is the *value*, or *power index*, due to L.S. Shapley.

Given an n-person game **v**, Shapley looks for an outcome $\mathbf{\Phi[v]}$ – a vector with n components – which says what each player might reasonably expect to obtain in the game. His approach is axiomatic, based on the idea that two games (with the same players) can be added to give another game. He then gives four axioms that the value rule $\mathbf{\Phi}$ should satisfy: efficiency (S1), dummy exclusion (S2), symmetry (S3), and additivity (S4).

Of these, S1 says that all players together should receive the amount $v(N)$. S2 says that a player who contributes nothing – good or bad – to any coalition will receive nothing. S3 says in effect that if $v(S)$ depends only on the number of players in S, then all players in the game should receive the same amount. Finally, S4 says that the value $\mathbf{\Phi[v + w]}$ of the sum of two games is equal to the sum $\mathbf{\Phi[v]} + \mathbf{\Phi[w]}$ of the values of the two games.

We then obtain the following theorem:

Theorem 5

There is a unique value rule, defined over the space of all n-person games, satisfying the axioms S1–S4. This rule $\mathbf{\Phi}$ gives to player i in game **v** the amount

$$\mathbf{\Phi}_i[\mathbf{v}] = \sum \frac{(s-1)!(n-s)!}{n!}[v(S) - v(S - \{i\})]$$

where the summation is taken over all coalitions S with $i \in S$, and s is the number of elements in set S.

As may be seen, the value assigns to player i a weighted average of her *marginal contributions* $v(S) - v(S - \{i\})$ to the several coalitions she may join (cf the usual economic idea of paying factors of production according to their marginal product).

In Example 9, it is no great surprise that the value is (333.3, 333.3, 333.3). In Example 10, the value is (583.3, 333.3, 83.3). Note that player I – as the only seller, with two buyers – is in a position to obtain most of the profits. Note, however, that player III gets something – he is not quite so helpless as the core analysis above would have us believe.

Other Solution Concepts

The core represents a very strong type of stability, which need not exist; the Shapley value gives us a kind of expectation which does not take into account prior relations among the players. Thus, other solution concepts have been proposed, some looking for a weaker type of stability, and others perhaps looking for the outcome of a single play, or taking prior agreements among the players into account. We mention only a few of these.

Von Neumann and Morgenstern developed *stable sets* which represent 'standards of behavior.' A set V of imputations is stable if (a) no imputation in V dominates another, and (b) every imputation outside V is dominated by at least one in V.

R.J. Aumann and M.B. Maschler developed *bargaining sets*, based on the idea that, in a coalition, some players may *object* to a payoff if they feel that others in the coalition are receiving too much of the coalition's profits. The objected players may be able to show, by means of a *counterobjection*, that they can protect their share. A payoff vector is in the bargaining set if, for any objection, there is a counterobjection.

Other solution concepts include the *kernel*, due to M.B. Maschler and M. Davis, and the *nucleolus*, due to D. Schmeidler. Modifications of the Shapley value, using information about some players' relationships to others, have been suggested. Chief among these are one by R. Myerson and another by G. Owen.

Modifications of the Game Model

Games with Nontransferable Utility

As mentioned in Section The Characteristic Function, von Neumann and Morgenstern's theory is based on the assumption that utility is freely transferable (at a one-for-one rate) between members of a coalition. In real life this may not be possible, either because there is no perfect 'money' (to be identified with utility) or because rules (e.g., antitrust laws) forbid it. Theorists have therefore developed a model for games without side payments (the NTU case) (one in which utility is not freely transferable). In these games, $v(S)$, rather than a single number, is a set in s-dimensional space – a generalization of the two-dimensional sets discussed in Section Nash's Bargaining Model. It is possible to define imputations, dominations, and the core for these games. Attempts have been made to generalize other concepts – mainly the value and the bargaining sets – but no generalization seems as satisfactory as the original (TU case) (one in which utility can be freely transferred among coalition partners) concepts.

Games with Incomplete Information

Game theory usually assumes that the players have *complete information* (a situation in which a player knows all the laws of the game she is playing) as to the game they are playing – e.g., they know the strategies available, the probabilities at chance moves, and the payoff functions. In practice this need not be so. Thus theorists (J.C. Harsanyi, R.J. Aumann, and M.B. Maschler, among them) have studied situations in which the players have subjective probabilities as to the games they are playing. Aumann and Maschler have studied the problem faced by a player who wishes to use private information but fears that its use may reveal this information to an opponent, and the dual problem faced by a player who wishes to elicit information from an opponent's moves.

See also: Axiomatic Theories; Decision and Choice: Luce's Choice Axiom; Decision and Choice: Random Utility Models of Choice and Response Time; Deterrence Theory: Crime; Dynamic Decision Making; Optimal Control Theory.

Bibliography

Aumann, R.J., Maschler, M.B., Stearns, R.E., 1995. Repeated Games with Incomplete Information. MIT Press, Cambridge, MA.
Harsanyi, J.C., 1977. Rational Behavior and Bargaining Equilibria in Games and Social Situations. Cambridge University Press, Cambridge.
Hurwicz, L., Reiter, S., 2006. Designing Economic Mechanisms. Cambridge University Press, Cambridge.
Krishna, V., 2010. Auction Theory. Elsevier, Amsterdam.
Luce, R.D., Raiffa, H., 1957. Games and Decisions. Wiley, New York.
Owen, G., 2013. Game Theory. Emerald, London.
van Damme, E., 1987. Stability and Perfection of Nash Equilibria. Springer, Berlin.
Vohra, R.V., 2011. Mechanism Design. Cambridge University Press, Cambridge.
von Neumann, J., Morgenstern, O., 1944, 1947. The Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ.

Game Theory: Noncooperative Games

Eric van Damme, Tilburg University, Tilburg, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

We describe noncooperative game models and discuss game theoretic solution concepts. Some applications are also noted. Conventional theory focuses on the question 'how will rational players play?', and has the Nash equilibrium at its core. We discuss this concept and its interpretations, as well as refinements (perfect and stable equilibria) and relaxations (rationalizability and correlated equilibria). Motivated by experiments that show systematic theory violations, behavioral game theory aims to integrate insights from psychology to get better answers to the question 'how do humans play?'. We provide an overview of the observed regularities and briefly sketch (beginnings of) theories of boundedly rational play.

Introduction

Games are mathematical models of interactive decision situations, i.e., situations in which multiple decision makers, each one with its own objectives, jointly determine the outcome. Game theory aims to predict what players will do in such situations and what outcomes will result. The theory has been applied in economics, other social sciences, biology and computer science, among others. Aumann (1987) presents an overview of how the field developed in the twentieth century. The 3-volume *Handbook of Game Theory with Economic Applications* (Aumann and Hart, 1992/1994/2002) provides a fairly complete overview of rationalistic game theory in almost 2400 pages. Excellent textbooks at the graduate level are Myerson (1991) and Osborne and Rubinstein (1994). For more information and detailed references, the reader is advised to consult any of these sources. Behavioral game theory is more recent, less established and developing more quickly; Camerer (2003) provides a good starting point for this branch.

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel has been awarded to game theorists four times. In 1994, John Harsanyi, John Nash and Reinhard Selten shared the Prize for developing equilibrium theory in noncooperative games; in 2005 Robert Aumann and Thomas Schelling shared it for enhancing our understanding of conflict and cooperation by mean of game theoretic analysis; in 2007 Leonid Hurwicz, Eric Maskin, and Roger Myerson were praised for developing the theory of mechanism design, while in 2012 Alvin Roth and Lloyd Shapley received the Prize for the theory of stable allocations and the practice of market design. Researchers from closely related fields were honored in 1996 (James Mirrlees and William Vickrey; incentives under asymmetric information), in 2002 (Daniel Kahneman, Vernon L. Smith; behavioral economics and experimental economics), and in 2009 (Elinor Ostrom (who shared the Prize with Oliver Williamson); institutional economics). Excellent information on the contributions of the Prize winners is available on the official web site of the Nobel Prize, http://www.nobelprize.org/nobel_prizes/economics/laureates/index.html. In 2014, the Prize was awarded to Jean Tirole, who also made important contributions to noncooperative game theory, among others together with Drew Fudenberg and Eric Maskin.

The game theory literature distinguishes two main classes of models: cooperative games and noncooperative games. The terminology, which suggests that in one case the players cooperate and in the other they do not, is misleading. The difference is not in what players want, but rather in what they are allowed to do. Traditional game theory assumes that players are rational and strive to maximize their utility; in this respect, there is no difference between the two models. In noncooperative theory, however, it is assumed that the model is complete and that the players are bound by its rules. In particular, contracts or commitments are binding only if the formal rules explicitly allow this. By contrast, in cooperative theory, players are free to negotiate, form coalitions and possibly make side payments, and are assumed to have access to a costless external mechanism that enforces agreements. Noncooperative theory assumes such an external mechanism is absent, hence, focuses on self-enforcing agreements.

The distinction was coined by John Nash in his Ph.D. thesis (Nash, 1950), which also introduced the fundamental solution concept for noncooperative games (the Nash equilibrium concept). In the path-breaking Von Neumann and Morgenstern (1944), the founders of game theory had developed two distinct theories: one for two-person games in which the players have strictly opposite interests (two-person zero-sum games) and another for n-person games in which the players can form coalitions and make side payments. They had argued that, as soon as there are more than two players, choosing an ally and forming a coalition becomes crucial, with side payments being key in stabilizing cooperation; hence, that two-person zero-sum games, in which such possibilities are irrelevant, are the exception. Consequently, they assumed that a mechanism enforcing coalitions and contracts was available and focused on cooperative theory. Nash extended Von Neumann and Morgenstern's two-person zero-sum theory and developed the general noncooperative theory. In this theory, each player acts independently, without collaborating with any of the others, however, by making full use of all the possibilities for cooperation that the game allows.

A model should be rich enough to allow for the relevant possibilities, but also simple enough to allow in-depth analysis and yield insight. Nash argued that noncooperative models are more fundamental, as it should always be possible to model coalition negotiations as formal moves in a noncooperative game. Although correct, the resulting model may be too complicated and the attention for details may blur the general picture, hence, each type of model has its advantages. Possibilities for cooperation can be modeled either as part of the game, or as part of the solution concept. In this contribution,

we limit ourselves to the former approach. As we will see, there exist various deep links between the two approaches. Cooperative game theory is surveyed in William Thomson's contribution to this Encyclopedia.

The remainder of the material is structured as follows. We first describe the two main classes of noncooperative models (the extensive form and the strategic (or normal) form), and the concept of strategy that allows the reduction of one model to the other. Next, we turn to solution concepts that are based on the assumption that players are perfectly rational and have full understanding of the game. We discuss Nash equilibrium, some of its drawbacks, as well as extensions (correlated equilibria) and refinements of it, such as perfect and stable equilibria. The experimental literature has shown that human players may deviate from perfect rationality in systematic ways; hence, we next discuss recent results from the behavioral game theory literature, which aims to construct models of thinking and learning that are descriptively more accurate. We close by briefly discussing some applications, including the link between cooperative and noncooperative theory.

Noncooperative Game Models

An 'extensive game' is a very detailed model of a conflict situation; it specifies which players are involved and how the game evolves over time: which player moves when, what information does the player then have, what can he do, what are the possible consequences of his actions, and how do the players evaluate the outcomes? Von Neumann and Morgenstern (1944) already provided a set-theoretic description of this model, but Kuhn (1953) provided a graph theoretic formulation that is easier to work with and that has become the standard.

A special case is a game with 'perfect information,' in which the moves are sequential and each player, whenever he has to move, is fully informed about everything that has happened before. Chess is a game with perfect information. **Figure 1** gives a very simple example: player 1 (P_1) moves first and chooses between terminating the game, O, with payoff 1 to P_1 and X to P_2 or giving the move to P_2(action I), who then determines whether each player gets 0, or whether P_1 gets 2, with P_2 getting Y. We will return to this game below. Most of the literature has restricted attention to games with 'perfect recall' in which a player never forgets what he knew or what he has done before. Bridge, when modeled as a two-player game, has imperfect recall: if EW is the defending team, then W does not know (recall) the cards of E.

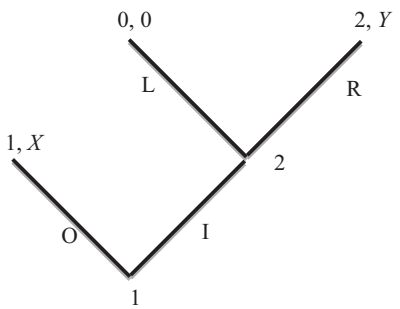

Figure 1 $X > Y > 0$

By means of the concept of strategy, already introduced in early work of John von Neumann, an extensive form game can be reduced to its strategic form. A 'strategy' for a player is a full plan of action for how to play the game, i.e., it specifies a unique action for each decision point of this player and each piece of information that this player might then have. Denote by N the set of players in the game and by S_i the set of all strategies for player $i \in N$. Assume that players evaluate outcomes by Von Neumann Morgenstern utility functions and let u_i be the utility function of player i. An n-tuple of strategies $s = (s_1, \ldots, s_n)$, one for each player, determines a probability distribution over the outcomes and, hence, implies a unique expected utility $u_i(s)$ for each player. Von Neumann argued that the 'strategic form' of the game, the tuple $<N, S_1, \ldots, S_n, u_1, \ldots, u_n>$ specifying players, strategies and payoffs, contains all the information that is needed for rational players to determine what to do. Consequently, it would suffice to develop theory for 'strategic games.' The literature has debated whether the details of the extensive form are indeed irrelevant (see Section Rationality and Equilibrium). **Table 1** gives the strategic form of the game from **Figure 1**: the rows are the strategies of P_1, the columns the strategies of P_2, and in each cell, the first number is the payoff (i.e., the utility) to P_1 and the second the payoff to P_2; a convention that will be followed throughout. Note that in a strategic game each player only moves once, with players choosing strategies simultaneously.

The interpretation of a 'strategic game' $<N, S, u>$ is that the data of the game are common knowledge: all players are fully informed who the players are, what strategies each of them has available and how all players evaluate the possible outcomes. Harsanyi (1968) showed how incomplete information can be incorporated into the model. In Harsanyi's model, the game starts with a chance move that distributes private information to each player. It is then common knowledge what pieces of information each player might have and with what probabilities, but the exact piece of information (also called the player's type) is only known to the player himself. Formally, a 'Bayesian game' is a tuple $<N, T, S, p, u>$, where N is the set of players, $T=T_1 \times \cdots \times T_n$ is the type space, p is a probability distribution on T, and u is the vector of utility functions, where each player i's utility may generally depend on all types and all actions taken, $u_i = u_i(t, s)$.

In a Bayesian game, information is distributed asymmetrically: P_i is informed about his type t_i and can choose his action $s_i \in S_i$ on the basis of it; P_j only knows his own type t_j and on this basis forms beliefs $p(t_i, t_j)$ about what P_i might know and might do, $s_i(t_i)$. As an example of a Bayesian game, think of a sealed bid auction: each bidder knows what the auctioned object is worth to him, but has only imprecise information

Table 1 A strategic game

	L	R
O	1, X	1, X
I	0, 0	2, Y

about the value of others. Also *see* the article on Auctions. A special class of Bayesian games, 'signaling games,' has proved fertile for theory development. In such a game, there is only one player with private information, but this player moves first; he has to decide how much information to reveal, while his opponents have to figure out what information his action might be signaling (*see* the article Information, Economics of).

It should be noted that above we referred to common knowledge in a loose sense: all players knowing something is different from that being common knowledge. The latter also implies that all players know that all players know it, and that all players know that all players know that all players know it, etc; see Aumann (1976) for a formal definition. Rubinstein (1989) shows that games with almost common knowledge are very different than games with common knowledge.

In the above discussion, we restricted attention to pure strategies; however, there is also the possibility of randomization. In a strategic form game, a 'mixed strategy' of P_i is a probability distribution σ_i over this player's set of pure strategies S_i. (We will write $\Delta(S_i)$ for the set of all such probability distributions.) It can be interpreted either as an act of deliberate randomization of this player, or as an expression of the uncertainty that the other players face about what P_i is going to do. In the latter case, σ_i represents the common beliefs held by the opponents of P_i. (The assumption that they have the same beliefs makes sense if they have the same information; in a Bayesian game, σ_i are the *ex-ante* beliefs, before the own type is known.)

In an extensive game, one can distinguish two types of randomization. When using a mixed strategy, the player randomizes over his pure strategies before the game starts. If the player randomizes locally over his actions at each of his information sets, this is called a 'behavior strategy.' A mixed strategy always induces a behavior strategy, but the converse only holds in games with prefect recall (Kuhn, 1953); hence, in these games, the restriction to behavior strategies is without loss of generality. Bridge provides an illustration that, without perfect recall, one may be able to play better with mixed strategies: to play optimally, the defending team needs to perfectly coordinate the actions of its members without revealing too much information; this is possible in a mixed strategy, but not when the team members randomize independently.

Rationality and Equilibrium

Nash Equilibrium

The fundamental solution concept for noncooperative games was introduced by John Nash in his 1950 PhD thesis. The mathematical core of the thesis was published as Nash (1951), but the chapter 'Motivation and Interpretation,' was not published, which may have led to misunderstandings and may have delayed development of the field.

Let $G = <N, S, u>$ be an n-person strategic game. A 'Nash equilibrium' of G is a strategy combination of σ(either mixed or pure) with the property that each player i is playing a best response against the strategies played by the others, hence $u_i(\sigma) = \max_{\tau_i} u_i(\tau_i, \sigma_{-i})$ for all $i \in N$. In other words, as long as the others do not deviate from their equilibrium strategies, player i cannot improve his payoff by deviating from his equilibrium strategy. An equilibrium of an extensive game is defined similarly; it simply is an equilibrium of the associated strategic game. An equilibrium of a Bayesian game is frequently called a Bayesian Nash equilibrium: each player plays a best response against the strategies of the others, whatever his type might be.

Nash provided two justifications for his concept. The first interpretation is rationalistic; Nash equilibrium is an answer to the question: what would perfectly rational players do? If we assume that a theory of rational play produces a unique solution and if the players know the solution, then rational (payoff maximizing) players will conform to this solution only if it is a Nash equilibrium. Any other (single-valued) theory is self-defeating. While assuming rational players to know the solution and to make use of their knowledge seems fine, the assumptions of existence and uniqueness are crucial. In fact, as a game may have multiple Nash equilibria (see below), this rationalistic justification seems incomplete at best. How can a player predict what another player will do if there are multiple Nash equilibria? This question motivated a fruitful, long-term research project of two Nobel Prize winners, who ultimately showed that additional, but not undisputable, assumptions allow one to come up with an answer; Harsanyi and Selten (1988).

The second 'mass action interpretation' assumes the game to be repeated, with each time players being newly drawn from certain populations and with each player accumulating empirical information on the relative advantages of their own strategies as well as on how often the opponents play their strategies. If the frequencies with which the various pure strategies are used converge, then "the mixed strategies representing the average behavior in each of the populations form an equilibrium point" (Nash, 1950: p. 22). In other words, under certain assumptions, learning leads to Nash equilibrium. Note that for this second interpretation, uniqueness is irrelevant; the initial conditions may determine at which equilibrium the process ends up. A large literature has investigated various types of learning processes and under which conditions convergence to equilibrium is indeed obtained; see Fudenberg and Levine (1998, 2009) and Young (2004) for overviews.

A third interpretation originates in the biological literature; see Smith (1982). Strategies are assumed to be randomly matched against each other with the payoff $u_i(\sigma)$ representing the fitness (expected number of offspring) of strategy σ_i when the state of the system is σ. The fittest strategies grow fastest, hence, if the system converges, it must be to a Nash equilibrium. The literature has studied various evolutionary processes, among which the replicator equation; see Weibull (1995).

By using a fixed point theorem (such as Brouwer's or Kakutani's) one can show that any strategic game has a Nash equilibrium, provided one allows equilibria in mixed strategies. 'Matching pennies' (two players simultaneously choose H or T with P_1 winning both pennies if the choices match and P_2 winning otherwise), shows that not all games have equilibria in pure strategies: in this game each player has to randomize.

Mixed strategies and mixed strategy equilibria are frequently viewed as being problematic, as players do not have

strict incentives to play them: any pure strategy in the support of a mixed equilibrium strategy is a best response as well. Mixed strategies, however, can also be interpreted as beliefs. Harsanyi (1973) showed that mixed equilibria arise naturally as beliefs associated with pure equilibria of a Bayesian game in which the uncertainty that each player faces about the other players' payoffs is explicitly taken into account. If one allows for the payoffs to be slightly uncertain, but each player being perfectly informed about his own payoff, then we have a larger Bayesian game $G(\varepsilon)$ in which each player i can play a pure strategy as for each payoff realization a specific pure action is optimal. However, as i's opponents do not know i's payoff realization, they will be uncertain about what action P_i will actually play. Given an equilibrium σ of a generic strategic game G one can find an equilibrium $\sigma(\varepsilon)$ of the Bayesian game $G(\varepsilon)$ such that, in the limit, as the uncertainty vanishes, for each $i \in N$ the beliefs of i's opponents associated with $\sigma(\varepsilon)$ converge to σ_i.

As noted, Nash's noncooperative model is a generalization of Von Neumann and Morgenstern's two-person zero-sum game. Indeed, Nash's equilibrium concept is a generalization of their minimax solution. The founders of game theory asked the question what is the highest payoff that a player can guarantee himself and they defined a minimax strategy as one that guarantees this value. For two-person zero-sum games, σ is a Nash equilibrium if and only if, for each player i, σ_i is a minimax strategy. If games are not strictly competitive, however, the two concepts differ; best responding against a player that pursues his own interests is different from optimally defending yourself against somebody that plays against you.

Table 2 lists three well-known games. In the Prisoners' Dilemma (**Table 2(a)**), (D, D) is the unique Nash equilibrium. This shows that an equilibrium may be (Pareto) inefficient; another outcome is preferred by both players. That constraints on cooperation can hurt players is unsurprising; they no doubt would agree on (C, C), if they could sign binding contracts. 'Battle of the Sexes' (**Table 2(b)**, with $X = 0$) is a game with two pure equilibria, and a mixed strategy one yielding each player 5/6, hence, a game may have multiple equilibria. In stag hunt (**Table 2(c)**, with $0 < X < 4$), (S, S) and (R, R) are Pareto ranked equilibria. Although (R, R) yields higher payoffs for both players, S is a safer strategy: it guarantees the payoff X, while, if only one player chooses R, he ends up with 0. In this game, there is a conflict between risk dominance and payoff dominance: if $2 < X < 4$, then (S, S) risk dominates (R, R) (Harsanyi and Selten, 1988). The strategic game from **Table 1** (with $Y > 0$) shows that some Nash equilibria may be unstable. (O, L) is a Nash equilibrium, but, for P_2, R is always at least as good as L and sometimes it is strictly better; hence, R weakly dominates L. We conclude that Nash equilibria always exist, that there may be multiple (nonequivalent) equilibria, and that some equilibria may be unstable.

Table 2 Three strategic form games

(a) Prisoners' dilemma			*(b) Battle of the sexes*			*(c) Stag hunt*		
	C	*D*		*L*	*R*		*S*	*R*
C	3, 3	0, 4	U	5, 1	0, 0	S	*X*, *X*	*X*, 0
D	4, 0	1, 1	D	*X*, *X*	1, 5	R	0, *X*	4, 4

By means of a model similar to Harsanyi's, Carlsson and Van Damme (1993) have shown that common payoff uncertainty can eliminate equilibria, hence, can serve as an equilibrium selection device. Consider a game as in **Table 2(c)**. If $X < 0$, then S is a strictly dominated strategy; R is strictly dominated if $X > 4$, while for intermediate values, there are multiple equilibria. Carlsson and Van Damme consider the situation where X can take any real value, with players facing uncertainty, but each player receiving a reasonably accurate independent signal X_i about the true value of X before making his decision. They coined the term 'global game' for the resulting Bayesian game. If P_i receives a very large signal ($X_i \gg 4$), he can be reasonably sure that R is dominated, hence he will play S. Similarly, each player will choose R if his signal is very negative. It is natural to focus on simple (switching point) equilibria of the global game in which each P_i plays R if and only if $X_i < x^*$, for some x^*; indeed the authors show that under certain assumptions only such equilibria exist. In such a Bayesian equilibrium, a player must be indifferent when receiving the signal x^*. However, as each player believes that the events $X_1 > X_2$ and $X_1 < X_2$ are approximately equally likely, indifference can hold only if, $x^* = 2$. In the limit, as uncertainty vanishes, players, hence, coordinate on the risk dominant equilibrium of the stag hunt. For further discussion on global games, on the conditions under which these have unique limit equilibria, the relation with common knowledge, and their applications, the reader is referred to Morris and Shin (2003).

Rationalizability, Iterated Dominance and Correlated Equilibria

Nash equilibrium assumes that players optimize and have correct beliefs about their opponents. The concept of rationalizability (Bernheim, 1984; Pearce, 1984) keeps the first assumption, but relaxes the second. It can be obtained as outcome of an interactive process. In the first step, all beliefs are allowed and any player i can choose any best response to these. In the second step, for any $i \in N$, the beliefs of i's opponents are only allowed to put positive weight on strategies that are best responses for P_i and any player j is only allowed to play strategies that are best responses to profiles of the resulting beliefs. The 'rationalizable strategies' are those that survive iterative application of this procedure. In games with a unique rationalizable outcome, weaker rationality assumptions suffice to obtain the outcome.

The above procedure is related to (but not fully equivalent with) the iterative elimination of strictly dominated strategies, where a pure strategy r_i of player i is 'strictly dominated' if there exists a mixed strategy σ_i such that $u_i(r_i, s_{-i}) < u_i(\sigma_i, s_{-i})$ for all s_{-i}. The difference is related to the question of whether beliefs about different players can be correlated or not. This issue does not arise in two-person games, for which the two procedures are equivalent. Note that any Nash equilibrium is rationalizable, however, Nash equilibria may vanish if weakly dominated strategies are eliminated (a pure strategy r_i is 'weakly

dominated' if there exists s_{-i} satisfying the above inequalities for some s_{-i} and corresponding weak inequalities for all s_{-i}).

Aumann (1974) introduced the concept of correlated equilibrium that generalizes Nash's concept to games in which communication between the players is possible. It assumes that players can conduct joint lotteries (play correlated strategies), but cannot make binding agreements or side payments. As an introduction, consider the game from **Table 2(b)** with $X = 0$ ('Battle of the Sexes'). If players can communicate, they can decide to throw a coin together and to play (U, L) when the outcome is H and (D, R), when the outcome is T. This correlated strategy yields each player the payoff 3, which is a good compromise. Furthermore, the agreement is self-enforcing: whatever the outcome of the coin toss, one player has a strong incentive to abide by the agreement and to follow up on it, so that it is in the best interest of the other to do so as well. We have a correlated equilibrium: a correlated strategy in which no player has an incentive to deviate from.

It will be clear that any convex combination of Nash equilibria is a correlated equilibrium. However, we can do more. Consider again the game from **Table 2(b)**, but now with $4 < X < 5$ and consider the scenario in which the players instruct a trustworthy mediator to randomize equally among the three cells of the matrix that have positive payoffs. Furthermore, they instruct him that, for each outcome of the lottery, P_1 shall only be informed about which row resulted while P_2 shall only be informed about the column. Viewing this information as a recommendation of what to play, one notices that, if one player always follows the recommendation, it is in the interest of the other player to do so as well. For example, if U is recommended, P_1 knows that P_2 has been recommended (and will play) R, hence, U yields the highest payoff. Similarly, if D is recommended, P_1 knows that P_2 will play L or R, each with probability 1/2, hence, again following the recommendation is best. The entire scheme is self-enforcing, hence, a correlated equilibrium, but it is not a convex combination of Nash equilibria. Formally, a 'correlated equilibrium' is a correlated strategy $\sigma \in \Delta(S)$ such that, for each player i, if $\sigma_i(s_i) > 0$, then s_i is a best response against $\sigma_{-i}(s_{-i} \mid s_i)$; in words: any recommendation s_i that any P_i might receive is a best response given $i's$ beliefs after hearing s_i.

Equilibrium Refinements and Stable Equilibria

We now focus attention on games in extensive form. Zermelo (1913) already showed that, in theory, chess can be solved by a backward induction procedure: starting at the end of the game, one works backwards replacing each decision point with the outcome that is obtained if a local best reply is taken there. It is straightforward to extend this procedure to any game with perfect information. Note that this procedure assumes persistent rationality: whatever happened before, each player assumes that all players will act rationally from then on. This assumption seems appropriate, but it is noteworthy that Von Neumann and Morgenstern (1944) already criticized it.

In the game of **Figure 1**, backward induction produces the Nash equilibrium (I, R). There is, however, a second equilibrium, (O, L), that is not consistent with this procedure. In effect, in this second equilibrium, P_2 threatens to choose L and, if believed, P_1 chooses O, so that the threat does not have to be executed. The question is whether the threat to play L really is credible. What is at issue here is that the strategic form of the game seems to assume that a player can commit himself to a strategy. However, in a noncooperative game such commitments are impossible, hence, when faced with the *fait accompli* that P_1 has chosen I, the best P_2 can do is to choose R. Strategies should not just be optimal at the beginning of the game, but also from each decision point onwards; in a noncooperative setting, only backward induction equilibria make sense. When a player is rational, the possibility to reoptimize should not lead him to deviate from his original strategy.

Following Selten (1965), a large literature on 'equilibrium refinements' has studied the question of how to generally eliminate Nash equilibria that rely on 'incredible threats.' Three related but conceptually different strands of literature can be distinguished.

The first line groups concepts that aim to extend the backward induction procedure beyond games with perfect information. Selten (1965) proposed 'subgame perfect equilibria': equilibria that induce Nash equilibria in all subgames; a subgame being a part of the game tree that constitutes a game of itself. Kreps and Wilson (1982) strengthened this idea and introduced 'sequential equilibria.' Such an equilibrium consists of a strategy profile σ, together with a system of beliefs μ that for each player specifies a probability distribution over the nodes in each information set. Two conditions are required:

1. Sequential rationality: at each information set, the player's strategy is optimal against the strategies of the others given the beliefs.
2. Consistency: the system of beliefs should be compatible with the strategy profile.

Various formalizations of consistency have been proposed. The main advantage of this framework is that it provides a natural language to discuss 'reasonableness' of beliefs and of the associated equilibria, hence, sequential equilibria can be further refined by imposing additional conditions on the beliefs. For example, in signaling games, one can insist that, upon observing action a_i, uninformed players assign beliefs $\mu(t_i, a_i) = 0$ to those types t_i of the informed player for which a_i is dominated or equilibrium dominated (*see* the article on Alternative Schools of Economic Thought).

In the second strand of literature, it is assumed that players will, with a small probability, make mistakes and it is required that equilibria be robust against this possibility. Hence, perfect rationality is viewed as a limiting case of slightly imperfect rationality. The seminal paper in this strand is Selten (1975), which introduced the concept of 'perfect equilibrium,' based on the idea that all players make independent mistakes at each of their information sets. When players tremble, all decision nodes are reached with positive probability, so that an optimizing player is forced to choose a best response everywhere. Consequently, in the perturbed game there cannot be incredible threats. An equilibrium is said to be perfect if it can be obtained as limit of Nash equilibria of a sequence of perturbed games when the mistakes vanish. As Kreps and Wilson (1982) define their main consistency requirement in terms of mistakes, there is a clear

link between the two strands: any perfect equilibrium is sequential. The perfectness concept can be refined by imposing further conditions on the mistakes. For example, Roger Myerson's proper equilibrium insists that more costly mistakes occur with much lower frequency.

Sequential equilibria and perfect equilibria rely essentially on the extensive game structure. As a result, two extensive games with the same strategic form may have different sequential or perfect equilibria. Kohlberg and Mertens (1986) argued that such dependency is undesirable: fully rational players are not misled by presentation details that are strategically irrelevant. Hence, they argued in favor of a solution that satisfies invariance, i.e., which only depends on the strategic form. The game from **Figure 1** illustrates that the backward induction property may be uncovered in the strategic form: (I, R), the only strategy pair to survive backward induction in the extensive game, is also the only one surviving interactive elimination of weakly dominated strategies in the strategic form. Hence, we have two different rationality principles that produce the same outcome. More generally, a proper equilibrium of a strategic game induces a sequential equilibrium in any extensive form game with that strategic form. Hence, it seems possible that 'robust' equilibrium outcomes can be identified in the strategic form.

Kohlberg and Mertens (1986) initiated the axiomatic approach to equilibrium refinement: they postulate several properties that a rational solution should satisfy and investigate whether a solution satisfying these properties exists. Examples of such properties are: (1) invariance (already discussed above), (2) consistency with one-person decision theory (admissibility), (3) independence of strategies that are dominated or that are suboptimal responses against the solution, and (4) a solution should remain whenever a game is embedded in a larger one (the small worlds property). Kohlberg and Mertens (1986) proposed to strengthen perfectness by insisting not just on stability against one particular sequence of trembles, but against all small trembles. As typically a single equilibrium will not have this property, they suggested looking at minimal closed and connected sets of equilibria that are stable in this sense. This initial attempt did not satisfy all properties that they considered desirable, but Mertens (1989) next proposed a concept that indeed satisfies all of them. The definition of Mertens stability is highly technical, insisting on certain homology properties of the best reply correspondence. Govindan and Wilson (2008) defined the related concept of metastable equilibria that is somewhat weaker than Mertens stability, but satisfies the same decision-theoretic properties.

We thus conclude that the question 'how to exclude Nash equilibria that rely on incredible threats?' has led to highly technical questions about the best reply correspondence of the game. Exactly why such sophisticated techniques appear necessary to solve such an intuitive question is still imperfectly understood. Nevertheless, that stability suffices for that purpose can be shown with a simple example, which also illustrates the concept of forward induction. Consider the game in which P_1 first chooses whether to take up an outside option yielding him payoff 2 or to play Battle of the Sexes (the game from **Table 2(b)** with $X = 0$). Taking up the outside option is part of the perfect equilibrium (OD, R): if P_1 thinks that (D, R) will be played in the subgame, he is better off taking his option. However, this outcome does not seem reasonable: P_1 not choosing the outside option and then playing D is strictly dominated. Being requested to play, it seems that P_2 should, therefore, conclude that P_1 will play U and should respond with L. Hence, only the outcome (5, 1) seems reasonable. Indeed, this is the only stable equilibrium outcome.

This same outside option game may, however, illustrate that in case communication is possible (hence, when the basic solution concept is correlated equilibrium), we cannot insist on the solution to just depend on the strategic form. Exactly when the communication takes place may matter. If players can only communicate before the start of the game, P_1 can never be induced to play ID, hence, communication is immaterial and the outcome is (5, 1). On the other hand, if players can communicate after P_1 has thrown away his option, then players can randomize between (5, 1) and (1, 5), hence, P_1 can be induced to give up his option. Extensive form correlated equilibria are different from strategic form correlated equilibria, and for good reasons; see Myerson (1991; Chapter 6), where one can also find some remarks on how to refine correlated equilibria.

Behavioral Game Theory

Conventional game theory, with its focus on the question 'how will rational players play?' frequently makes sharp predictions about the outcome or about how this outcome changes with a change in the data (comparative statics), hence, studying the theory's empirical relevance appears quite natural. Testing with field data, however, has its limitations, and although there are exceptions, serious experimental investigation of the descriptive relevance of rationality-based theory only started in the 1980s (Kagel and Roth, 1995). The first wave of experimental studies established that standard theory sometimes (or frequently, depending on one's viewpoint) provides poor predictions of how humans play, and that there are systematic patterns in the deviations, which has then led to revised theories (or at least models) of play incorporating these regularities. In the last 25 years, emphasis has thus shifted to the question 'how do humans play non-cooperative games?' leading to a strong interaction between theory and empirical work.

This section provides a brief overview of the results that have been achieved in this rapidly developing field. We start by describing how human players deviate from conventional rationality. Three aspects can be distinguished: (1) motivation (what drives people; how do players evaluate outcomes?); (2) cognition (how do people reason; what thinking processes do they use when they are confronted with a new game?); and (3) adaptation (how do people learn when they play the same game repeatedly?). After having described empirical regularities, we briefly discuss novel theories relating to each of these aspects.

Bounds on Human Rationality

Humans have emotions, face cognitive limitations, and may be under (time) pressure; they process information differently, or may be motivated differently than standard theory assumes. Conventional theory is silent on the decision-making processes

that are used. It assumes that players only differ in preferences, not in cognitive abilities. Models of bounded rationality take into account limits on human knowledge, informational processing capacity and computational ability. Obviously, in some games, these limits are more important than in others.

When observed behavior differs from the game theoretic prediction, one can point to one of two main causes:

1. The game, as perceived by the players, is different from the one analyzed by the theorist.
2. The solution concept is not applicable in this situation.

Conventional game theory starts with the model and assumes that it is common knowledge among the players. Real life conflict situations are less structured and have to be interpreted; a model has to be constructed, and different players may perceive the situation differently. At first, one does not necessarily see all aspects of the problem, leading to superficial decision making. Nevertheless, this may already produce a satisfactory solution, not inviting further reflection. Selten (1998) notes that such superficiality may explain the framing effect (Tversky and Kahneman, 1981): the way the situation is presented may have an important influence on the outcome. The reasoning process anchors at aspects that deeper inspection might reveal to be irrelevant. Alternatively, framing may provide clues to the solution that conventional theory mistakenly neglects; Schelling (1960).

Traditional models assume that players are rational in the sense of Von Neumann and Morgenstern (expected utility) and Savage (subjective expected utility). Mostly, the assumption of players having a common prior is added. Following the path-breaking Kahneman and Tversky (1979), behavioral economics has shown that humans deviate from these assumptions and the resulting behavior in systematic ways: (1) utility depends not just on the final state (outcome), but on changes in the state; (2) the change is measured with respect to a reference point; (3) losses loom larger than gains (loss aversion); (4) individuals use decision weights that are different from probabilities, with small probabilities being overweighted and large ones underweighted; and (5) information is processed differently than Bayes' rule describes (*see* the article on Behavioral Economics for more details). Furthermore, while the rational choice model allows general preferences, in empirical work it is frequently assumed that players are selfish and care only about own materialistic payoffs. Many humans are motivated differently: altruism and reciprocity play a role, with (social) norms influencing behavior as well.

Equilibrium concepts assume that (1) players form beliefs about what others will do, (2) these beliefs are correct, and (3) players best respond to them. In equilibrium, players never are surprised. A game may be too complicated to find a best response, or players may be insufficiently motivated to find one. Equilibrium concepts are based on circular reasoning (fixed points; solutions to systems of equations), but, as stressed by Selten (1998), humans have a tendency to avoid circular concepts. The natural way of problem solving is by using step-by-step reasoning processes. The rationalistic interpretation of equilibrium assumes, but leaves unexplained how the beliefs of players that are confronted with a new game come to be correct, an assumption that is especially problematic in games with multiple equilibria. Finally, natural learning processes need not converge to (Nash) equilibrium, or the learning may be too slow to be practically relevant.

Although for certain classes of games, nonequilibrium concepts, such as rationalizability or iterated dominance, are sufficient to predict the outcome, these rely on an unlimited number of iterations; humans seem to do only a very small number of rounds of iterated strategic thinking. Related, humans find processes like backward induction to be unnatural and do not always use these.

Empirical Regularities

Bounded rationality (the rationality displayed by humans in decision-making situations), hence, differs significantly from perfect rationality. As a result, it is not surprising that experiments have revealed that outcomes observed when humans play games differ systematically from the standard game theory predictions; see Camerer (2003), Goeree and Holt (2001), and Selten (1998). The following are some observed empirical regularities:

1. Framing effects can be very important, even in simple zero-sum games.
2. The outcome may depend on aspects of the game that conventional theory considers to be irrelevant; in contrast, aspects that the theory considers relevant need not matter; for example, two games with the same unique mixed strategy equilibrium may be played differently.
3. Not only the ordering of the payoffs matters, but also payoff differences: strategies that are not best responses may be played and small payoff differences may be ignored altogether.
4. Frequently, players care about other aspects than own (material) payoffs.
5. In games played once, the observed outcome may differ from the unique rationalizable one; in strategic games, people only do a small number of rounds of iterated elimination.
6. In perfect information games, monitoring of the decision making process shows that players may not do backward induction; this procedure does not come naturally, but it can be taught.
7. When players gain experience with a game, they adjust behavior; players learn in different ways and with different speeds, which may depend on the game; learning processes may be very slow.

Players' Motivations

The ultimatum game (Güth et al., 1982), along with its variants, the dictator game and the trust game, has spawned a large literature on social preferences. In the ultimatum game, P_1 proposes how to divide an amount of money between him and P_2; if P_2 accepts, the proposal is implemented; otherwise, each player gets 0. If both players are selfish (care only about own material payoffs), P_1 offers (close to 0) to P_2, which the latter accepts. Experiments, however, show that most offers are close to 50/50 and that the responding player may reject

considerable amounts. The dictator game (the variant in which P_2 is forced to accept all proposals) allows to test whether the behavior of P_1 is driven by altruism or by the fear that low offers will be rejected. In the trust game, P_1 can first transfer all or part of the amount to P_2, with that amount, T, being multiplied by a known constant $K > 1$, and with P_2 subsequently deciding how much to transfer back. In this game, responding players display positive reciprocity: the larger the transfer received, the more is transferred back. In public goods games, we see negative reciprocity: players that do not contribute sufficiently to the public good are punished. All these games show considerable individual heterogeneity, while culture matters as well; Henrich et al. (2004).

While the experimental results do not refute game theoretic analysis as such, they suggest that great care is needed in modeling a situation as a game; the frequently made assumption of selfishness does not describe most situations very well. Player 2 may be motivated to get a reasonable share and he may prefer conflict to an outcome in which he gets much less than the proposer. The game of **Figure 1** is a kind of mini-ultimatum game: P_2 may reject if Y is too small. On the other hand, if $X \cong 0$ and $Y = 1$, P_2 might realize that P_1 is forced to divide asymmetrically; he may accept in this case, while he might possibly reject if X is large and $Y \cong 0$. Hence, not only the outcomes matter, but also the context in which these arise; consequentialism is violated. In the literature, a large variety of models has been proposed and tested to incorporate these aspects, including models of pure distributional preferences, as well as models in which intent or procedural aspects matter. Given heterogeneity, incomplete information about preferences is natural, and there are also models in which players care about what others think of them.

Sobel (2005) is a recent overview of the literature on interdependent preferences.

In many models that fit under this heading, just the players' payoff functions are changed by incorporating behavioral aspects, while the solution concept remains conventional. Geanakoplos et al. (1989), however, argued that emotions cannot be captured by assuming that payoffs just depend on players' actions. Players' emotions, which influence how outcomes are valued (payoffs), will also depend on their expectations and, hence, on what they learn in the game. For example, in **Figure 1**, if P_2 expects P_1 to choose O, then he may be disappointed when asked to move and, therefore, choose R, while a P_2 in that expected I may simply play L. Geanakoplos et al. (1989) introduce 'psychological games' as games in which payoffs depend both on what players do and on what they think and they define a 'psychological equilibrium' as a profile in which each player best responds and beliefs are correct. Rabin (1993) builds on this idea to construct a model that incorporates considerations of fairness. Starting from an ordinary game and players' expectations, he first applies a kindness function to modify the payoffs to take into account the emotions and construct a psychological game, to which then the concept of psychological equilibrium is applied. The resulting outcome is called a fairness equilibrium. In the game from **Figure 1**, both the outcome (1, X) and the outcome (2, Y) can be supported by fairness equilibria.

Modeling Players' Thinking Processes

The p-beauty contest game ($0 < p < 1$) illustrates the difference between individual human rationality and rationalizability. In this game, n players are asked to simultaneously pick an integer from $\{1, 2, \ldots, 100\}$, with the person whose number is closest to p times the average being the winner (ties broken randomly). The unique Nash equilibrium is for all to choose 1; in fact, if $p < 3/4$, only 1 is rationalizable. Experiments show that players that are new to the game exhibit distinct, bounded levels of reasoning. For example, for $p = 2/3$, games played with a large population of players, show spikes at numbers such as 50 (superficial thinking), 33 (the best response against 50), 22, 15, 10, etc. Obviously, not all players are perfectly rational, and in such situations, a fully rational player should not necessarily pick 0.

If one knows that other players are not that sophisticated and engage only in limited thinking, a natural idea is to try to figure out what naïve players might do and to estimate the distribution of 'levels of thinking' in the population, and to best respond against the resulting beliefs. For example, if one thinks that naïve, level-1 and level-2 players are equally likely, one believes the average chosen by the others to be 35 and chooses 23. Theories of cognitive hierarchy (Camerer, 2003) or level-k thinking have been developed, and have been brought to the data with reasonable success; see Crawford et al. (2013) for a recent overview.

Some insight has also been gained on the question of how 'naïve play' is obtained: players come to the laboratory with prior ideas and may seek for analogies with known situations, or may seek for clues, such as labels and focal points; Schelling (1960). It has also been shown that backward induction reasoning does not come natural; for example, eye tracking studies show that many players only look for a limited time, or not at all, at the parts of the game tree that come later; hence, there is a tendency for myopic decision making.

Learning

When players get more experience with a game, they will typically adjust their behavior. As already suggested in Nash (1950), learning may produce outcomes that are more in line with equilibrium predictions. For example, in repetitions of p-beauty contest games, smaller numbers are chosen over time. We already noted that there is a very large theoretical literature on learning. In broad terms, two types of learning models can be distinguished. In belief learning, players adjust their beliefs about what others will do and best respond to these. In reinforcement learning, a player's own strategies have certain 'attractions,' which determine the probability of playing them and which are adjusted through time. Camerer (2003) describes a hybrid model, experience-weighted attraction learning that has successfully been fitted to the data of several games. The data and model show that learning is much quicker in some games than in others, hence, that players come to certain games with rather strong, but not necessarily correct, beliefs. Of course, not all people learn in the same way and some people may learn quicker than others. Selten (1998) stresses that learning may involve a combination of adaptive and analytical approaches.

Players that view the situation as routine and use adaptive methods may fail to acquire an overview of the situation and, hence, may not be able to reach equilibrium; for that, a more analytical approach may be needed.

Conclusion

The allotted space only allows making just a few remarks on the applications of noncooperative game theory; the interested reader is suggested to consult the Handbook of Game Theory or to follow any of the leads given below.

One application is to cooperative game theory. The 'Folk Theorem' shows that when a game is repeated sufficiently often, players are sufficiently patient and obtain enough information about past play (and when some other technical conditions are satisfied), the set of subgame perfect equilibrium outcomes coincides with those of the static cooperative game. Hence, in a context of repeated interaction, contracts may not be necessary to achieve efficient outcomes. Work of Nobel Prize winner Elinor Ostrom on the exploitation of common property resources combines this theory with rich institutional data. Avner Greif's contribution to the Handbook surveys the related literature that employs game theory for economic history analysis.

Following Nash's (1950) suggestion, the literature has pursued the 'Nash program' and has shown that several cooperative solution concepts can be implemented noncooperatively, i.e., there exists a noncooperative game of which the (refined) equilibrium outcomes correspond to those predicted by the cooperative solution concept. For example, the Nash-bargaining solution can not only be implemented by means of the simultaneous bargaining game proposed already in Nash (1953), but also by a very natural alternating offers procedure proposed in Rubinstein (1982). In a similar vein, solution concepts such as the Shapley value, the kernel, and the nucleolus can be obtained by noncooperative procedures; we refer to Thomson's article in this Encyclopedia for details.

Within economics, the field of industrial organization has proved a fertile ground for application. Probably the most successful applications thus far have been to auctions or more generally market design; see the article on Auctions. Following the seminal work of Vickrey, auction theory first focused on the questions of how to bid in a single-item auction and what is the revenue maximizing mechanism for selling such item. In implementing spectrum policy, governments were, however, confronted with the question of how to sell multiple heterogeneous items, given certain objectives, such as market efficiency and raising sufficient government revenue. Research, in which game theorists interacted with experimental economists and Operations Researchers, has led to highly innovative auction designs, such as the simultaneous multiround auction and the combinatorial clock auction, which have subsequently been successfully implemented by various governments around the world.

By using both cooperative and noncooperative approaches, and by combining theory with experimental and empirical studies, Nobel Prize winner Alvin Roth and coworkers showed that stability is important for understanding the success of particular market institutions. Building on this insight, they successfully reengineered several existing institutions, such as those for matching organ donors with patients. Importantly, this work shows that economics can take into account ethical restrictions, such as the prohibition of side payments.

As seen above, game theory has provided a strong stimulus for experimental economics. No doubt, the further development of behavioral game theory will give a further boost to applications.

See also: Auctions; Behavioral Economics; Cooperative Game Theory; Experimental Economics; Information, Economics of; Old and New Institutionalism in Economics; Transaction Costs and Property Rights.

Bibliography

Aumann, R.J., 1987. Game theory. In: Eatwell, J., Milgate, M., Newman, P. (Eds.), The New Palgrave Dictionary of Economics, vol. 2, pp. 460–482.
Aumann, R.J., 1974. Subjectivity and correlation in randomized strategies. Journal of Mathematical Economics 1, 67–96.
Aumann, R.J., 1976. Agreeing to disagree. Annals of Statistics 4, 1236–1239.
Aumann, R.J., Hart, S.,1992/1994/2002. Handbook of Game Theory with Economic Applications, vol. 1, pp. 1–731, vol. 2, pp. 735–1520, vol. 3, pp. 1521–2351, Elsevier, North Holland.
Bernheim, B.D., 1984. Rationalizable strategic behavior. Econometrica 52, 1007–1029.
Camerer, C., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press, Princeton, NJ.
Carlsson, H., Van Damme, E., 1993. Global games and equilibrium selection. Econometrica 61, 989–1018.
Crawford, V.P., Costa-Gomes, M.A., Iriberri, N., 2013. Structural models of nonequilibrium strategic thinking: theory, evidence, and applications. Journal of Economic Literature 51 (1), 5–62.
Fudenberg, D., Levine, D., 1998. The Theory of Learning in Games. MIT Press, Cambridge, MA.
Fudenberg, D., Levine, D.K., 2009. Learning and equilibrium. The Annual Review of Economics 1, 385–419.
Geanakoplos, J., Pearce, D., Stacchetti, E., 1989. Psychological games and sequential rationality. Games and Economic Behavior 1 (1), 60–79.
Goeree, J.K., Holt, C.A., 2001. Ten little treasures of game theory and ten intuitive contradictions. The American Economic Review 91 (5), 1402–1422.
Govindan, S., Wilson, R., 2008. Metastable equilibria. Mathematics of Operations Research 33 (4), 787–820.
Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization 3 (4), 367–388.
Harsanyi, J., 1967–1968. Games with incomplete information played by Bayesian players, parts I, II and III. Management Science 14, 159–182, 320–334, 486–502.
Harsanyi, J., 1973. Games with randomly disturbed payoffs: a new rationale for mixed-strategy equilibrium points. International Journal of Game Theory 2, 1–23.
Harsanyi, J., Selten, R., 1988. A General Theory of Equilibrium Selection in Games. MIT Press, Cambridge, MA.
Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., 2004. Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-scale Societies. Oxford University Press, Oxford, UK.
Kagel, J., Roth, A., 1995. Handbook of Experimental Economics. Princeton University Press, Princeton, NJ.
Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decisions under risk. Econometrica 47 (2), 263–291.
Kohlberg, E., Mertens, J.F., 1986. On the strategic stability of equilibria. Econometrica 54, 1003–1039.
Kreps, D., Wilson, R., 1982. Sequential equilibria. Econometrica 50, 863–894.

Kuhn, H., 1953. Extensive games and the problem of information. In: Kuhn, H., Tucker, A.W. (Eds.), Contributions to the Theory of Games II. Princeton University Press, Princeton, NJ, pp. 193–216.
Mertens, J.F., 1989. Stable equilibrium: a reformation, part I: definition and basic properties. Mathematics of Operations Research 14, 575–625.
Morris, S., Shin, H., 2003. Global games: theory and applications. In: Dewatripont, M., Hansen, L., Turnovsky, S. (Eds.), Advances in Economics and Econometrics (Proceedings of the Eighth World Congress of the Econometric Society). Cambridge University Press, Cambridge, UK.
Myerson, R.B., 1991. Game Theory. Harvard University Press, Cambridge, MA.
Nash, J., 1950. Non-cooperative Games. Ph.D. dissertation. Princeton University, Princeton, NJ.
Nash, J., 1951. Non-cooperative games. Annals of Mathematics 54, 298–395.
Nash, J., 1953. Two-person cooperative games. Econometrica 21, 128–140.
Osborne, M.J., Rubinstein, A., 1994. A Course in Game Theory. MIT Press, Cambridge, MA.
Pearce, D., 1984. Rationalizable strategic behavior and the problem of perfection. Econometrica 52, 1029–1051.
Rabin, M., 1993. Incorporating fairness into game theory and economics. American Economic Review 83 (5), 1281–1302.
Rubinstein, A., 1982. Perfect equilibrium in a bargaining model. Econometrica 47, 1353–1366.
Rubinstein, A., 1989. The electronic mail game: strategic behavior under 'almost common knowledge'. American Economic Review 79, 385–391.
Schelling, T., 1960. The Strategy of Conflict. Harvard University Press, Cambridge, MA.
Selten, R., 1965. Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit. Zeitschrift für die gesamte Staatswissenschaft 121, 301–324, 667–689.
Selten, R., 1975. Re-examination of the perfectness concept for extensive form games. International Journal of Game Theory 4, 25–55, 19.
Selten, R., 1998. Features of experimentally observed bounded rationality. European Economic Review 42, 413–436.
Smith, J., 1982. Evolution and the Theory of Games. Cambridge University Press, Cambridge, UK.
Sobel, J., 2005. Interdependent preferences and reciprocity. Journal of Economic Literature 43 (2), 392–436.
Tversky, A., Kahneman, D., 1981. The framing of decisions and the psychology of choice. Science 211 (4481), 453–458.
Von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior, third ed. Princeton University Press, Princeton, NJ.
Weibull, J., 1995. Evolutionary Game Theory. MIT Press, Cambridge, MA.
Young, H.P., 2004. Strategic Learning and Its Limits. Oxford University Press, Oxford, UK.
Zermelo, E., 1913. Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiel. Proceedings of the 5th International Congress of Mathematicians 2, 501–504.

Gangs, Sociology of

Lorine A Hughes, University of Nebraska at Omaha, Lincoln, NE, USA
James F Short, Jr., Washington State University, Pullman, WA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.F. Short Jr., volume 9, pp. 5880–5883, © 2001, Elsevier Ltd.

Abstract

This article summarizes theoretical and empirical developments and trends in the sociology of gangs. Definitional issues and the history of sociological interest in gangs are reviewed first, followed by discussion of major research findings concerning gangs and gang members. Attention then turns to continuity and change in youth and gang cultures and in their relationships with violence. We close with brief comments concerning the state of gang scholarship and future research needs.

Introduction

Beginning with Frederic Thrasher's classic, *The Gang: A Study of 1313 Gangs in Chicago* (1927), gangs have been a subject of continuous sociological scrutiny, law enforcement attention, and public interest. Although the term *gangs* has been applied to groups of adult criminals, especially those that control segments of organized and professional crime, the preponderance of scholarly attention has focused on youth. The ever-changing nature of gangs and disagreement among and between scholars and law enforcement agencies concerning what kinds of groups constitute youth gangs hampers growth of knowledge about gangs and gang control. Nevertheless, study of gangs provides opportunities for advancing knowledge of broader principles of human behavior at every level of explanation. Defining gangs broadly as groups of youth that are unsupervised by adults and self-determining with respect to membership criteria, group structure, and behavior, this article traces continuity and change in processes associated with gang formation and behavior.

Definitional Issues

The term *gang* is both a theoretical construct and an object of varying definitions in legal statutes, legal and social agency policy, and popular discourse. *Gang* and other terms (*mob*, *syndicate*, *outfit*, etc.) have been applied to many types of groups, including those associated with organized and professional crime and incarcerated felons. Scholarly attention, however, has focused primarily on youth gangs, although 'setting and style' are beginning to matter more as "age becomes less of a defining characteristic" of gangs (Klein, 1995: p. 21).

A major sticking point in the literature concerns the role of crime in defining gangs. While some scholars argue that gangs and crime must be defined and measured independently of one another in order to avoid tautology, others contend that the centrality of crime and violence in youth gangs is what most distinguishes them from other troublesome peer groups, such as tagger crews. Still others suggest compromising by focusing on the "tipping point beyond which we say, 'Aha – that sure sounds like a street gang to me'" (Klein, 1995: p. 30). Two major 'signposts,' Klein argues, are degree of commitment to a criminal orientation and self-recognition of gang status. In later writings, Klein articulates the Eurogang consensus nominal definition of street gangs, based on intense collaboration involving American and European gang scholars (see Klein, 2006): "A street gang (or a troublesome youth group corresponding to a street gang elsewhere) is any durable, street-oriented youth group whose own identity includes involvement in illegal activity."

With Cheryl Maxson, Klein has investigated the practical implications of differing official definitions of gang crime. In the United States, some jurisdictions require evidence of gang motivation before a crime is legally defined as 'gang-related,' whereas other jurisdictions consider a gang-related crime to have occurred simply if the offender, victim, and/or both are suspected or known gang members. Analyzing data from Los Angeles police and sheriff departments, Maxson and Klein (1990) showed that, while estimates of the characteristics of homicide incidents (e.g., demographics, setting, use of firearms, offender-victim relationship, retaliatory motive, etc.) are robust to definitional variations, estimates of the prevalence of gang-related homicide are substantially higher under the gang-member definition than the more restricted gang-motivation definition. Such definitional ambiguity leaves open the possibility of exploitation by law enforcement agencies and officials primarily for financial, political, and other organizational gain, as observed in a study of a 'moral panic' surrounding gangs in Nevada in the late 1980s (McCorkle and Miethe, 1998).

Historical Background of Scholarly Research on Gangs

Frederic Thrasher's (1927) *The Gang* was the first attempt to survey the extent of youth gang activity in a major city, and perhaps the first in any jurisdiction. The project involved census and court records, personal observations, and personal documents collected from gang boys and from persons who had observed gangs in various contexts. Studies by Clifford Shaw and his collaborators provided even more information about the collective nature of youthful delinquency, consistently demonstrating that most boys who were brought before the juvenile court committed their delinquent acts in

the company of others. Case studies documented patterns of friendship, the association of younger with older offenders, and the influence of organized crime and other forms of adult criminality in communities with high rates of juvenile delinquency (Shaw, 1930; Shaw et al., 1929; Shaw and McKay, 1931, 1969[1942]; Shaw and Moore, 1931).

Ecological studies located delinquency in space and in relation to urban development, documenting economic and institutional contexts within which urban lives were lived and identifying forces that shaped the ability of communities to aid in the socialization of children and exercise control over their misbehavior. Theoretical insights from these studies continue to be important in thinking about youth crime. Foreshadowing 'labeling theory,' for example, Shaw's Jack-Roller – reflecting on his confinement in a reformatory – noted that he was no longer "just a mischievous lad, a poor city waif, a petty thief, a habitual runaway," but "a criminal" (Shaw, 1930: p. 103). Conversely, both Shaw and Thrasher noted that incarceration and criminal notoriety often had positive value among gang members, as it does today: "… residents in the vicinity south of the stock yards were startled one morning by a number of placards bearing the inscription 'The Murderers, 10,000 Strong, 48th and Ada.' In this way attention was attracted to a gang of thirty Polish boys, who hang out in a district known as the Bush" (Thrasher, 1927: pp. 62–63).

Thrasher's Murderers were involved in a good deal of criminal activity, but their primary pastimes "were loafing, smoking, chewing, crap-shooting, card-playing, pool, and bowling." Their "rudeness and thievery" were "an awful nuisance" to local shop keepers and neighbors (Thrasher, 1927: pp. 62–63). The arcane language notwithstanding, these excerpts highlight patterns similar to those observed among contemporary and earlier gangs involving 'hanging out' and episodic minor, as well as more serious, criminal behavior (Decker and Van Winkle, 1996; see Klein, 1995).

Ironically, gang research and subsequent ethnographic community studies (Suttles, 1968; Whyte, 1943) challenged the major theoretical legacy of ecological research in the city. Social disorganization theory attributed high rates of crime and delinquency to the lack of effective social organization and institutions in communities characterized by poverty, population heterogeneity, and turnover – just the sort of communities in which gangs were located. Yet, later studies documented highly organized groups and institutions in communities with high rates of crime and delinquency (Suttles, 1968; Whyte, 1943; see also Patillo, 2007).

Contemporary community research takes into account networks of interpersonal relationships and density of organizations, institutional domains, and leadership connections within and among neighborhoods (Bursik, 2002). Sampson and colleagues (1999; Sampson, 2012) combine these intra- and inter-neighborhood and community processes using the concept of 'collective efficacy,' that is, intergenerational closure linking adults and children, reciprocal local interfamily and interaction among adults with respect to children, and expectations that community residents will share child-control responsibilities. Collective efficacy is associated with the ability of local communities to exert effective control over violent and other forms of criminal behavior. Though not as well established empirically, so, too, are cohesive leadership and organizational relationships within the institutional domains that characterize the city.

Midway between Thrasher's classic work and the turn of the century, his book was abridged and reissued. Social changes had vastly altered the gang landscape, and gang research, heretofore largely descriptive, was changing in response to seminal theoretical proposals by Albert Cohen (1955), Richard Cloward and Lloyd Ohlin (1960), and Walter Miller (1958). Competing explanations for the origins of the delinquent subculture (Cohen), variations in delinquent subcultures (Cloward and Ohlin), and the role of lower-class culture in producing gang delinquency (Miller) stimulated empirical research and subsequent theorizing about the nature and behavioral effect of 'group processes,' especially 'status management' among gang leaders (Short and Strodtbeck, 1965). Controversy also erupted over the role of group cohesiveness in gang violence and the value of aggressive group programing by 'detached' community outreach workers (Jansyn, 1966; Klein, 1969, 1971).

By the last decade of the twentieth century, events that could hardly have been anticipated by earlier researchers had overtaken both theory and research. Whereas mid-century theories had outstripped available data, modest theoretical advances, and related empirical research were overwhelmed by the rapid proliferation of gangs in the United States and their spread to many other countries (Klein, 1995; Klein et al., 2001). In the United States, sophisticated firearms became more readily available to young people, often turning what previously had been 'non-zero-sum' contests between gangs into lethal confrontations (Anderson, 1999; Short and Strodtbeck, 1965). Weakened conventional economic, political, and family institutions lost legitimacy, contributing to rising crime rates (LaFree, 1998); and technological revolutions in the automotive industry, electronic communication, and entertainment media altered youth and gang cultures, aided by the institutionalization of some gangs, and created heretofore unknown problems for institutions of social control (Hagedorn, 2007, 2008; Moore, 1991; Vigil, 1988).

Change and Continuity

It is possible, nevertheless, to note both continuity and change in what is known about gangs, gang members, and the collective forms of delinquent and criminal behavior engaged in by young people (Short and Hughes, 2006):

1. Youth gangs are more prevalent in many countries than previously. Law enforcement agencies in all large US cities (over 250 000) report the presence of youth gangs, as do many suburban counties, small cities, and rural counties (including "Indian Country"), where gang prevalence is quite recent (Howell, 2012). European and UK researchers have identified variations in youth gang organization and behavior, some of them patterned after US gangs (Decker and Weerman, 2005).
2. Gangs in many parts of the world are comprised predominantly of youth who are marginalized in mainstream society as a result of race, ethnicity (including religion), or economic status (Hagedorn, 2007, 2008; Vigil, 1988). In the United States, these are predominantly African American and Hispanic, with increasing numbers of new Asian

immigrant children and children of immigrant parents. In Continental Europe and the United Kingdom, they tend to be the children of refugees from other countries or migrant workers. Gangs comprised of majority population youth, nearly always white, sometimes form in opposition to these marginalized groups, thus becoming marginalized themselves. Historically, however, white gangs formed first, followed by defensive black and Hispanic gangs (Howell, 2012).

3. Youth gangs are comprised overwhelmingly of young males, and the experience of females who associate with male-dominated gangs, or in all-female gangs, is quite different from the gang experience of males (Miller, 2008; Chesney-Lind and Hagedorn, 1999). Victimization, by male gang and family members, is a prominent theme in studies of female gang members. However, the specific experiences of female gang members appear to be conditioned by the structure and internal dynamics of their gangs, including group sex composition (Peterson et al., 2006).
4. The distribution and patterning of collective youth crime reflect the ecology of the places where they are located. In the United States and many other cities around the world, these are inner-city areas, much as they were in Thrasher's day. Where public housing projects are located outside central cities, gangs of loosely structured, ethnically mixed groups often 'commute' to areas where they engage in carousing, 'smash-and-grab' hits on stores, rolling drunks, robberies, and other problem behaviors.
5. Large aggregations of gang 'nations' emerged in a few large US cities, most notably Chicago and Los Angeles, during the 1960s and subsequent decades. Local gangs – wherever they are found – are primarily 'home grown,' but the influence of gang nations is widespread through mimicry, migration (sometimes forced by deportation) of gang members, and social media contacts. As in the past, many (probably most) gangs begin as nondelinquent friendship groups, 'hanging' groups, or groups formed around institutional (schools, youth serving agencies) nondelinquent activities, such as sports, singing, dancing, and so forth (see, e.g., Schwendinger and Schwendinger, 1985; Sullivan, 1989; Schwartz, 1987; Short, 1997). Although gangs vary along a number of dimensions, such as size, age range, and duration (Klein and Maxson, 2006), the primary distinguishing characteristic of all youth gangs is that they are self-determining with respect to membership criteria. Moreover, contrary to popular myths and depictions, most gangs are only loosely organized and incapable of sophisticated legitimate or illegitimate operations (see Coughlin and Venkatesh, 2003).
6. Studies suggest that unsupervised youth groups become violent, and otherwise delinquent, as a result of one or more of the following processes:
   a. escalation of the natural rough-and-tumble punching and wrestling that occurs among most male groups, and the association of status with fighting prowess;
   b. competition with rival groups involving status-enhancing activities, including retaliation, or economic motivations;
   c. conflicts based on racial/ethnic differences;
   d. definitions of groups as gangs, by others (including law enforcement personnel), and behavior toward them that reinforces a violent group identity;
   e. group processes that create or reinforce group cohesion based on violent or other delinquent behavior.
7. Due to rising rates of incarceration throughout much of the latter part of the century, the line between prison and street gangs has begun to blur, "as gang members are released into the community and as particular street gangs, such as the Bloods and the Crips, grow to be part of prison life when street members enter the correctional system" (McGloin, 2005: p. 608).

Street gangs are diverse in behavior, organization, continuity across generations, and in their relationships with other gangs and their communities. They change as members age and move into adulthood. The age range of many gangs has expanded in many countries, however, in response to the lack of attractive legitimate jobs. This change often is accompanied by increased violence, as gang members have turned to selling drugs and other illegal activities in order to support themselves, often with family approval and/or assistance (see Fagan, 1996; Fagan and Wilkinson, 1998; Hagedorn, 2007, 2008; Spergel, 2007). Violence and other delinquent activities occur frequently in 'staging' areas in which young people gather with friends and others in order to escape adult supervision (Anderson, 1999; Schwartz, 1987; see also Tita and Ridgeway, 2007).

Diversity and change among street gangs tend to blur distinctions between them and a variety of types of youth and young adult collectivities, including 'skinheads' and perhaps even perpetrators of mass shootings that have occurred on school properties (Moore et al., 2003). Although differing among themselves, these groups (and occasional isolated individuals) have in common their hatred of 'outsiders,' based on racial, ethnic, local status, or other exclusionary criteria, similar to crowds that gather to attack community outsiders, usually members of minority groups.

Drug and other 'working' or 'specialty' gangs are organized primarily around revenue-producing activities, such as drug selling, stealing, fencing stolen goods – all activities requiring greater organizational skills and structure than is characteristic of traditional street gangs. Also unlike street gangs, 'bikers,' 'punkers,' and sports fanatics and 'hooligans' are organized around specialized interests. Except for the latter (most of whom appear to be disadvantaged socially, rather than economically), these groups are comprised primarily of youth who have been marginalized by minority ethnic and/or economic status and who have poor prospects for integration or social mobility within dominant majority societies. The excessive behavior of bikers and sports hooligans also appears to be motivated in some measure by resentment of minorities.

Youth Culture

Virtually all youth are heavily influenced by the development, rapid growth, and diffusion of youth culture. Peddled by media advertising, and augmented by large social and economic forces that fail to provide meaningful roles for young people but cater to their distinctive fads, appetites, and currency, youth culture and gang culture together underlie the excesses of consumption and violence perpetrated by gangs.

Youth culture developed rapidly during the 1960s, especially in the United States. Young people born during the baby boom following World War II shared in the general level of affluence in the country, thereby creating a youth market with great economic power. Moreover, the period of education prior to entering the labor market was extended; increasing numbers of women joined the workforce, further separating mothers from youth in the home and in the neighborhood; adults worked increasingly in large organizations away from young people; and the mass media greatly expanded and focused their attention on the youth market (see Coleman et al., 1974: pp. 114–119). These changes accelerated as the century wore on, expanding their influence throughout the world.

Field researchers document the pernicious seductions of media-advertised products among less-affluent young people, leading often to thefts and assaults. Anderson (1999) describes the "zero-sum quality" that pervades the "code of the streets," in the search for respect – often associated with trivial items, such as shoes and jackets – among young African American men in the ghetto. Sullivan identifies "mutually valorizing cultural symbols" (clothing, drugs, alcohol, and recreational artifacts and activities, including graffiti and tattoos) that cut across racial and ethnic boundaries, often resulting in invidious comparison, confrontation, theft, and assault (Sullivan, 1989: pp. 248–249).

Adolescence is a period of especially intense identity formation, intense relationships, and shared feelings of friendship, acceptance, and respect. Conversely, feelings of rejection and disrespect also are especially intense and often the basis for group and subcultural formation. Observers of youth groups everywhere note that status differences within and between them are extremely variable and highly refined. Status criteria within a group, or between one group and another, may be based on race or ethnicity, relative economic affluence, skills in valued activities, public appearance, school performance or, perhaps most importantly, lifestyle differences. These bases of stratification typically also become criteria of inclusiveness and exclusiveness, which in turn create opportunities for both friendship and rejection.

Next to families, schools are perhaps the most important contexts for adolescent friendship, achievement, and recognition. Because of this, school contexts (including the journey to and from schools) are the settings for much adolescent behavior, including gang activities. On occasion, they also have been the setting for the most extreme forms of violence, including mass killings by students or others alienated from their fellows or from mainstream institutions in general. Although the specific causes of such extreme alienation are complex, it appears that schools have been targeted precisely because of their importance in the lives of adolescents, as symbols of rejection by both peers and an adult world that seems far removed from adolescent concerns. The ready availability of guns at times transforms normal adolescent turmoil and conflict into deadly confrontation (Moore et al., 2003).

More than ever before in history, young people, targeted for commercial exploitation and isolated from mainstream adult roles and institutions, confront economic conditions beyond their control. Not only are economic decline, severe unemployment, and the unavailability of 'good jobs' associated with the presence of street gangs, but with their transformation into 'economic gangs' (including drug gangs) and with ethnic, racial, and class-related identities and antagonisms that lead to other types of collective violence (Hagedorn, 2008; Pitts, 2000). These same forces alter both intergang relationships and relationships between gangs and their communities.

Research Trends and Needs

Research and theory concerning gangs, and effective means of controlling the violence and crime associated with gangs, are hampered by the lack of a viable typology of youth groups and groups comprised of young adults that would situate the many observed variations of gangs and gang-like groups within broader collective forms of behavior. Comparative research in many countries suggests that distinctions among and between such groups vary a great deal, and that the meaning and function of behavioral variations lack universality.

Varying definitions, meanings, and behaviors have consequences. To the extent that gangs and gang membership confer 'master status' on groups and individuals – among law enforcement personnel, in media portrayals, and among the general public – they contribute to the deterioration of relationships between economic, racial, and ethnic groupings and between the generations. Reification and demonization of gangs, by media, law enforcement, the public, and occasionally by scholars, for example, may contribute to gang problems rather than to their solution.

Self-reported gang membership nonetheless appears to provide a robust measure of gang involvement (Esbensen et al., 2001), supporting continued use of survey research, the primary means of estimating the prevalence of gangs and gang membership, and for identifying the antecedents and consequences of gang involvement. Surveys of youth have shown, for example, that male and female gang members tend to attribute their involvement in gangs to similar factors, such as having gang-involved friends and family members or a desire for protection, a sense of belonging, status, and/or support in coping with boredom and the exigencies of growing up under harsh social and economic circumstances. Moreover, suggesting something more to gangs than "birds of the feather flocking together," longitudinal surveys that track youth over time indicate that delinquent behavior, by girls as well as boys, increases during periods of active gang membership, independently of peer delinquency and compared to both prior and subsequent periods of reported nonmembership in gangs. In addition, multisite prospective survey data from evaluations of two recent school-based law-related education programs, Gang Resistance Education and Training and Teens, Crime, and the Community and Community Works, reveal that onset of gang membership significantly predicts techniques of neutralizations and unstructured socializing with peers, which, in turn, significantly predict delinquency six months later. These data also show that higher rates of unstructured and unsupervised socializing among gang members increase their risk of general (but not serious violent) victimization (Taylor et al., 2007, 2008) and that, consistent with findings from the Mobile (Alabama) Youth Survey in high-poverty

neighborhoods (Spano et al., 2008), gang members are more likely than other youth to adopt a delinquent lifestyle, which places them at increased risk of victimization regardless of whether or not they joined the gang for protection or feel safer as a result of their gang affiliation (Melde, 2009; Melde et al., 2009; Peterson et al., 2004).

Despite such advantages, self-reports have been criticized for failing to provide a realistic appreciation of peer delinquency and the *group* properties of gangs and gang crime (see Kreager et al., 2011). Gang researchers thus are turning increasingly to social network analysis, demonstrating the utility of the method in studying gang boundaries and networks of resource acquisition, leadership, and influence (Fleisher, 2006; McGloin, 2005; Papachristos, 2006; see also Kreager et al., 2011), the structure of inter-gang murder networks (Papachristos, 2009), and relationships between network position within the gang and criminal involvement and desistance from gang membership (Pyrooz et al., 2013; Roman et al., 2012). Additional research is needed to understand how the nature of ties both between and within gangs influences, and is influenced by, the attitudes, resources, and behaviors of individual gang members. However, this will require specialized data on relational ties from a census of all, or at least all core, members of more than a single gang or two; and none of the standard methods of accessing and studying gangs is well suited for this purpose.

Perhaps the best opportunity to study the relationship between group variables and individual behavior is provided by comprehensive detached worker programs. These programs fell out of favor in the 1970s, however, having been replaced largely by law enforcement and legal suppression efforts (e.g., California's Street Terrorism and Prevention Act). Neither of the two most well-known exceptions – Irving Spergel's Little Village project in Chicago (adapted and applied as the 'Comprehensive Community-Wide Approach to Gang-Prevention' in Mesa and Tuscon, Arizona; Riverside, California; and San Antonio, Texas) and the Boston Gun Project (applied to gang violence as Operation Ceasefire in Boston, Chicago, Cincinnati, and Indianapolis) – included a research component consisting of systematic analysis of relational ties within and between individual members of differing gangs. The Boston Gun Project did, however, collect valuable information about networks of violence at the gang level.

A second option is to analyze 'old' data from studies emerging out of the wave of intervention efforts in the 1950s and 1960s, including the Group Guidance Project in Los Angeles (Klein, 1969, 1995), the Midcity Project in Roxbury, Boston (Miller, 1962), and the Program for Detached Workers of the YMCA Metropolitan Chicago (Short and Strodtbeck, 1965). Short and Strodtbeck's study was unique in that it was intended to test theoretical propositions rather than evaluate the program with which their research was associated, but all three studies relied on detached workers to gain access to multiple groups of youth defined as *gangs*. Fortunately, data from these studies have survived and are being extended and analyzed in new ways that can be expected to shed light on the nature and effect of group processes (Decker, 2012; Hughes and Short, 2005; see also Papachristos, 2006).

Although gangs and gang phenomena have changed a great deal in the decades since these early studies and Thrasher's work in Chicago, there is continuity as well as change in the contexts and processes associated with gang formation, gang involvement, and gang crime and delinquency. The theoretical and empirical challenge is to recognize and account for these influences at all levels of explanation (Short, 1985; McGloin and Decker, 2010). In terms of gang policy, comprehensive efforts guided by well-established principles stand the best chance of success, provided they are implemented properly and expected outcomes clearly communicated and agreed upon by all participants, something that has proven difficult to accomplish in practice. As noted by Klein and Maxson, although the history of gang policy fails to inspire much optimism, "[t]here *is* room for intervention, even successful intervention, if we are willing to diagnose the situation carefully before we act" (2006: p. 208, emphasis in original).

See also: Control, Social; Crime, Sociology of; Delinquency, Sociology of; Norms; Violence: Public; Youth Gangs.

Bibliography

Anderson, Elijah, 1999. Code of the Streets: Decency, Violence, and the Moral Life of the Inner City. W.W. Norton, New York.

Bursik, Robert J., 2002. The systemic model of gang behavior: a reconsideration. In: Ronald Huff, C. (Ed.), Gangs in America, third ed. Sage Publications, Thousand Oaks, CA, pp. 71–81.

Chesney-Lind, Meda, Hagedorn, John M., 1999. Female Gangs in America. Lake View Press, Chicago, IL.

Cloward, Richard A., Ohlin, Lloyd E., 1960. Delinquency and Opportunity. Free Press, New York.

Cohen, Albert K., 1955. Delinquent Boys. Free Press, New York.

Coleman, James S., Bremner, Robert H., Clark, Burton R., Davis, John B., Eichorn, Dorothy H., Griliches, Zvi, Kett, Joseph F., Ryder, Norman B., Doering, Zahava B., Mays, John M., 1974. Youth: Transition to Adulthood. Report of the Panel of the President's Science Advisory Committee. University of Chicago Press, Chicago, IL.

Coughlin, Brenda C., Venkatesh, Sudhir A., 2003. The urban street gang after 1970. Annual Review of Sociology 19, 41–64.

Decker, Scott H., 2012. Gang Research at ASU. Retrieved 1 February 2012 from: http://gangresearch.asu.edu/.

Decker, Scott H., Winkle, Barrik Van, 1996. Life in the Gang: Family, Friends, and Violence. Cambridge University Press, Cambridge.

Decker, Scott H., Weerman, Frank M., 2005. European Street Gangs and Troublesome Youth Groups. AltaMira, Lanham, MD.

Esbensen, Finn-Aage, Winfree Jr., L. Thomas, He, Ni, Taylor, Terrence J., 2001. Youth gangs and definitional issues: when is a gang a gang, and why does it matter? Crime and Delinquency 47, 105–130.

Fagan, Jeffrey, 1996. Gangs, drugs, and neighborhood change. In: Ronald Huff, C. (Ed.), Gangs in America, second ed. Sage, Thousand Oaks, CA.

Fagan, Jeffrey, Wilkinson, Deanna L., 1998. Guns, youth violence, and social identity in inner cities. In: Tonry, Michael, Moore, Michael (Eds.), Youth Violence, Crime and Justice, vol. 24. University of Chicago Press, Chicago, IL, pp. 105–188.

Fleisher, Mark S., 2006. Youth gang social dynamics and social network analysis: applying degree centrality measures to assess the nature of gang boundaries. In: Short Jr., James F., Hughes, Lorine A. (Eds.), Studying Youth Gangs. AltaMira Press, Lanham, MD, pp. 85–98.

Hagedorn, John M. (Ed.), 2007. Gangs in the Global City: Alternatives to Traditional Criminology. University of Illinois Press, Urbana and Chicago, IL.

Hagedorn, John M., 2008. A World of Gangs: Armed Young Men and Gangsta Culture. University of Minnesota Press, Minneapolis, MN.

Howell, James C., 2012. Gangs in America's Communities. Sage Publications, Thousand Oaks, CA.

Hughes, Lorine A., Short Jr., James F., 2005. Disputes involving youth street gang members: micro-social contexts. Criminology 43 (1), 43–76.

Jansyn Jr., Leon R., 1966. Solidarity and delinquency in a street corner group. American, Sociological Review 31, 600–614.

Klein, Malcolm W., 1969. Gang cohesiveness, delinquency, and a street-work program. Journal of Research in Crime and Delinquency 6, 135–166.
Klein, Malcolm W., 1971. Street Gangs and Street Workers. Prentice-Hall, Englewood Cliffs, NJ.
Klein, Malcolm W., 1995. The American Street Gang. Oxford University Press, UK.
Klein, Malcolm W., 2006. The value of comparisons in street gang research. In: Short Jr., James F., Hughes, Lorine A. (Eds.), Studying Youth Gangs. AltaMira Press, Lanham, MD, pp. 129–144.
Klein, Malcolm W., Maxson, Cheryl L., 2006. Street Gang Patterns and Policies. Oxford University Press, Oxford.
Klein, Malcolm W., Kerner, Hans-Juergen, Maxson, Cheryl L., Weitekamp, Elmar G.M. (Eds.), 2001. The Eurogang Paradox: Street Gangs and Youth Groups in the U.S. and Europe. Kluwer Academic Publishers, Dordrecht.
Kreager, Derek A., Rulison, Kelly, Moody, James, 2011. Delinquency and the structure of adolescent peer groups.
LaFree, Gary, 1998. Losing Legitimacy: Street Crime and the Decline of Social Institutions in America. Westview Press, Boulder, CO.
Maxson, Cheryl L., Klein, Malcolm W., 1990. Defining gang homicide: An updated look at member and motive approaches. In: Huff, C. Ronald (Ed.), Gangs in America, 2nd ed. Sage Publications, Newbury Park, CA, pp. 3–20.
McCorkle, Richard C., Miethe, Terance D., 1998. The political and organizational response to gangs: an examination of a "moral panic" in Nevada. Justice Quarterly 15 (1), 41–64.
McGloin, Jean M., 2005. Policy and intervention considerations of a network analysis of street gangs. Criminology and Public Policy 4 (3), 607–636.
McGloin, Jean M., Decker, Scott H., 2010. Theories of gang behavior and public policy. In: Barlow, Hugh, Decker, Scott (Eds.), Criminology and Public Policy: Putting Theory to Work, pp. 150–165.
Melde, Chris, 2009. Lifestyle, rational choice, and adolescent fear: a test of a risk assessment framework. Criminology 47, 781–812.
Melde, Chris, Taylor, Terrance J., Esbensen, Finn-Aage, 2009. 'I got your back': examining the protective function of gang membership in adolescence. Criminology 47, 565–594.
Miller, Jody, 2008. Getting Played: African American Girls, Urban Inequality, and Gendered Violence. New York University Press, New York.
Miller, Walter B., 1962. The impact of a "total-community" delinquency control project. Social Problems 10 (2), 168–191.
Miller, Walter B., 1958. Lower class culture as a generating milieu of gang delinquency. Journal of Social Issues 23 (3), 5–19.
Moore, Joan W., 1991. Going Down to the Barrio: Homeboys and Homegirls in Change. Temple University Press, Philadelphia, PA.
Moore, Mark H., Petrie, Carol V., Braga, Anthony A., McLaughlin, Brenda A. (Eds.), 2003. Deadly Lessons: Understanding Lethal School Violence. National Academies Press, Washington, DC.
Papachristos, Andrew V., 2006. Social network analysis and gang research: theory and methods. In: Short Jr., James F., Hughes, Lorine A. (Eds.), Studying Youth Gangs. AltaMira Press, Lanham, MD, pp. 99–116.
Papachristos, Andrew V., 2009. Murder by structure: dominance relations and the social structure of gang homicide. American Journal of Sociology 115, 74–128.
Patillo, Mary, 2007. Black on the Block: The Politics of Race and Class in the City. University of Chicago Press, Chicago, IL.
Peterson, Dana, Miller, Jody, Esbensen, Finn-Aage, 2006. The impact of sex composition on gangs and gang member delinquency. Criminology 39 (2), 411–440.
Peterson, Dana, Taylor, Terrance J., Esbensen, Finn-Aage, 2004. Gang membership and violent victimization. Justice Quarterly 21 (4), 793–815.
Pitts, John, 2000. Discipline or Solidarity: The New Politics of Youth Crime. Macmillan, Basingstoke, UK.
Pyrooz, David C., Sweeten, Gary, Piquero, Alex R., 2013. Continuity and change in gang membership and gang embeddedness. Journal of Research in Crime and Delinquency 50 (2), 239–271.
Roman, Caterina G., Cahill, Meagan, Lachman, Pamela, Lowry, Samantha, Orosco, Carlena, McCarty, Christopher, Denver, Megan, Pedroza, Juan, 2012. Social Networks, Delinquency, and Gang Membership: Using a Neighborhood Framework to Examine the Influence of Network Composition and Structure in a Latino Community. Report prepared for the Office of Juvenile Justice and Delinquency Prevention. The Urban Institute, Washington, DC.
Sampson, Robert J., Morenoff, Jeffrey D., Earls, Felton, 1999. Beyond social capital: spatial dynamics of collective efficacy for children. American Sociological Review 64 (5), 633–660.
Sampson, Robert J., 2012. Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago Press, Chicago, IL.
Schwartz, Gary, 1987. Beyond Conformity or Rebellion: Youth and Authority in America. University of Chicago Press, Chicago, IL.
Schwendinger, Herman, Schwendinger, Julia, 1985. Adolescent Subcultures and Delinquency. Praeger, New York.
Shaw, Clifford R., 1930. The Jack-Roller: A Delinquent Boy's Own Story. University of Chicago Press, Chicago, IL.
Shaw, Clifford R., McKay, Henry D., 1931. National Commission on Law Observance and Enforcement Report, vol. 2. Report on the causes of crime. USGPO, Washington, DC.
Shaw, Clifford R., McKay, Henry D., 1969[1942]. Juvenile Delinquency and Urban Areas. University of Chicago Press, Chicago, IL.
Shaw, Clifford R., Moore, Maurice E., 1931. The Natural History of a Delinquent Career. University of Chicago Press, Chicago, IL.
Shaw, Clifford R., Zorbaugh, Harvey, McKay, Henry D., Cottrell, Leonard S., 1929. Delinquency Areas. University of Chicago Press, Chicago, IL.
Short Jr., James F., 1985. The level of explanation problem. In: Meier, Robert F. (Ed.), Theoretical Methods in Criminology. Sage Publications, Beverly Hills, CA, pp. 51–72.
Short Jr., James F., 1997. Poverty, Ethnicity, and Violent Crime. Westview, Boulder, CO.
Short Jr., James F., Hughes, Lorine A. (Eds.), 2006. Studying Youth Gangs. AltaMira Press, Lanham, MD.
Short Jr., James F., Strodtbeck, Fred L., 1965. Group Process and Gang Delinquency. University of Chicago Press, Chicago, IL.
Spano, Richard, Freilich, Joshua D., Bolland, John, 2008. Gang membership, gun carrying, and employment: applying routine activities theory to explain violent victimization among inner city, minority youth living in extreme poverty. Justice Quarterly 25 (2), 381–410.
Spergel, Irving A., 2007. Reducing Youth Gang Violence. AltaMira, Lanham, MD.
Sullivan, Mercer L., 1989. "Getting Paid": Youth Crime and Work in the Inner City. Cornell University Press, Ithaca, NY.
Suttles, Gerald, 1968. The Social Order of the Slum: Ethnicity and Territory in the Inner City. University of Chicago Press, Chicago, IL.
Taylor, Terrance J., Peterson, Dana, Esbensen, Finn-Aage, Freng, Adrienne, 2007. Gang membership as a risk factor for adolescent violent victimization. Journal of Research in Crime and Delinquency 44 (4), 351–380.
Taylor, Terrence J., Freng, Adrienne, Esbensen, Finn-Aage, Peterson, Dana, 2008. Youth gang membership and serious violent victimization: the importance of lifestyles and routine activities. Journal of Interpersonal Violence 23 (10), 1441–1464.
Thrasher, Frederic M., 1927[1963]. The Gang: A Study of 1,313 Gangs in Chicago. University of Chicago Press, Chicago, IL.
Tita, George, Ridgeway, Greg, 2007. The impact of gang formation on local patterns of crime. Journal of Research on Crime and Delinquency 44 (2), 208–237.
Vigil, James D., 1988. Barrio Gangs: Street Life and Identity in Southern California. University of Texas Press, Austin, TX.
Whyte, William F., 1943. Street Corner Society. University of Chicago Press, Chicago, IL.

Garfinkel, Harold (1917–2011)

Stacy L Burns, Loyola Marymount University, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Harold Garfinkel made fundamental contributions to twentieth-century sociology and beyond. He created a new and unique field of inquiry, 'ethnomethodology,' to study the everyday methods of practical action and practical reasoning used by members of society to make sense and produce order in their daily lives. Garfinkel and his students dissolved the supposed chasm between common sense and scientific/professional reasoning, unpacked the taken-for-granted existence of society, developed detailed descriptions of commonplace and specialized organizations of practice, and specified how society members assemble and sustain the sense that they live, act, and mean in a shared orderly world.

Biographical Facts

Harold Garfinkel was one of the most original and influential scholars of the twentieth century and beyond. Garfinkel created a new subfield of sociology when he coined the term 'ethnomethodology' ('EM') to refer to the study of the people's methods, i.e., the everyday methods of practical action and reasoning that members of society use to make sense and produce social order in their daily lives. Rather than accepting the presumed existence of social order as a given and treating the orderly features of society as a priori, objective structures disengaged from real-worldly settings and actual contexts of activity, he recommended that the orderly features of social action be treated as the 'contingent accomplishment of socially organized common practices' and as work done by participants in various settings to interactionally produce and interactionally recognize those orderly features (Garfinkel, 1967: p. 33).

Garfinkel was born on 29 October 1917 in Newark, New Jersey, and passed away on 21 April 2011 in Pacific Palisades, California. At the time of his death at the age of 93 years, he was survived by his daughter, Leah, and son, Mark, and by his wife of 65 years, Arlene (who passed away in 2013). Garfinkel attended the University of Newark where he earned a degree in business and accounting in 1939. After graduation, he worked as a volunteer at a Quaker work camp in rural Georgia. This was a transformative experience for Garfinkel and motivated him to pursue graduate training in sociology at the University of North Carolina, Chapel Hill, in the fall of 1939.

While working as a research fellow at the University of North Carolina under the prominent sociologist, Howard Odum, Garfinkel authored his first publication, *Color Trouble* (1940), an insightful short story on the dynamics of segregation in the American south. The story involved a dispute that occurred when an African-American woman refused to sit at the back of a bus as it traveled through Virginia on transit from New Jersey to North Carolina. It was originally published in *Opportunity* magazine in 1940 and later included in a collection of the best short stories of 1941 (O'Brien, 1941). Garfinkel's thesis advisor at North Carolina was Guy Johnson, a sociologist with expertise in the study of race relations. His master's thesis addressed the relationship between race and homicide and was based on extensive fieldwork and review of case files in county courthouses across the State of North Carolina. It was an early and highly innovative examination of how workers provide for the accountability of their records and decisions and the consequentiality of formulating race for case-processing and the social construction of statistics. A version of this work entitled Research Note on Inter- and Intra-Racial Homicide was published in *Social Forces* (Garfinkel, 1949).

Garfinkel continued his studies at North Carolina until being drafted into the Army Air Corps in 1942.

After serving in World War II, he was admitted to the first postwar class of graduate students at the newly formed Harvard Department of Social Relations. Garfinkel started his doctoral studies with Talcott Parsons, and also worked with such luminaries as Jerome Bruner, Gordon Allport, Robert Bales, and Pitirim Sorokin. Parsons was a major figure of twentieth-century sociology who sought to revolutionize sociology by integrating the work of Durkheim, Weber, and others into a new and general 'theory of social action' that would not exclude consideration of 'subjective' phenomena. Parsons believed that patterned social action and society in general were structured by a relatively limited number of rules, norms, and values. The society and culture delineate the relevant rules, norms, and values and, once internalized by society members through proper socialization, people accept their priority and seek out culturally valued goals by selecting among socially approved means, thereby structuring their conduct in accordance with how persons are expected to act in specific situations. The norms and values of a society are treated by Parsons as causes of action that become institutionalized and attain continuity as stable, orderly, and constraining features of the society. Although Garfinkel accepted the 'problem of social order' as the fundamental question for the human sciences, he rejected Parsons' theoretical solution that accounted for social action and order in terms of internalization of collectively shared norms and values. For Garfinkel, Parsons' abstract formulation of social reality was not a solution to the problem of social order. Instead, the solution was to be found in a descriptive and practical orientation that would respecify the problem of social order and 'social facts' as members' practices.

From 1946 to 1951, while writing his dissertation in Cambridge, Massachusetts, Garfinkel was significantly influenced by phenomenology, and especially by two phenomenological philosophers, Aaron Gurwitsch and Alfred Schutz. He met regularly with Gurwitsch to discuss phenomenology

and specifically Gurwitsch's then untranslated book, *Field of Consciousness* (1964). He also exchanged letters with Schutz and sometimes traveled to the New School for Social Research in New York for tutorial sessions with him on various topics related to sociology and phenomenology. Garfinkel was intrigued with Gurwitsch's phenomenological respecification of the Gestalt theory of form, including how meaning is determined as functional significations in a context and in part–whole relationships, with the part and whole being mutually constitutive of each other (Gurwitsch, 1964). These ideas were precursors to Garfinkel's recovery of the 'documentary method of interpretation' as a sense-making procedure from Karl Mannheim, who first introduced it in 1952 (Garfinkel, 1967: Chapter 3). Garfinkel credited Gurwitsch with his own prevailing interest in the 'achieved coherence of objects' (Garfinkel and Weider, 1992: p. 206 ftn. 11). The phenomenological ideas of Merleau Ponty's phenomenology of perception and his investigations of embodied action in context and space (Merleau Ponty, 1962) were of particular interest to Garfinkel and later ethnomethodologists, notably David Sudnow (1978) and Douglas Macbeth (2012). Garfinkel explicitly misread phenomenological texts to empirically investigate phenomena of social order and ultimately to further the interests of EM. His dissertation entitled The Perception of the Other: A Study in Social Order used Schutz's phenomenology of the 'life world' and concept of the 'attitude of everyday life' to critically work out Parsons' theory of social action and was completed in 1952. Later that year, Garfinkel traveled to Ohio State University to work on a project researching leadership on airplanes and submarines. When funding for this project ran out in 1953, he went to Witchita, Kansas where he observed trials, conducted postverdict interviews with jurors in the federal district court, and analyzed tape-recorded jury deliberations as part of a University of Chicago Law School study which came to be known as the 'American Jury Project' (Garfinkel, 1967: Chapter 4).

In 1954, Garfinkel joined the sociology faculty at the University of California, Los Angeles ('UCLA') in its early decade. At UCLA, he impacted the views of generations of students and faculty colleagues and sometimes transformed them into cohorts committed to his ideas. He also received frequent invitations to deliver lecturers, conduct seminars, and give plenary talks at professional conferences worldwide, including a series of seminars at the University of Paris in 1995 and an invitation to deliver the Alfred Schutz Memorial Lecture at the meeting of the Society for Phenomenology and the Human Sciences in 2004. Garfinkel was Simon Visiting Professor of Sociology at the University of Manchester in the UK in 1973, an invited fellow at the Center for Advanced Study in the Behavioral Sciences at Stanford University from 1975 to 1976, and a visiting fellow at the Center for Socio-Legal Studies, Oxford University from 1979 to 1980.

Ethnomethodologists and conversation analysts have met regularly for international conferences beginning in 1973 at the University of Manchester, where William ('Wes') Sharrock and other ethnomethodologists have built a long and lasting legacy of research and teaching in EM. Subsequent meetings have been held in the United Kingdom (at Oxford University) and in the United States, including at Boston University and UCLA. The 'International Institute for Ethnomethodology and Conversation Analysis' was formally established in 1989 and has held international conferences in The Netherlands (Amsterdam), England (Manchester), Japan (Tokyo), Switzerland (Fribourg), and Canada (Waterloo). The American Sociological Association accorded official Section status to Ethnomethodology/Conversation Analysis in 2005.

Garfinkel was based at UCLA faculty for over 50 years and, for four decades, UCLA has been the world's leading center for EM and its offshoot, conversation analysis (including institutional talk, Drew and Heritage, 1992). In classrooms and seminars, Garfinkel was a masterful teacher who inspired students and changed lives on a regular basis and whose longevity and passion made him a local legend. Although he retired in 1987, he continued to give public lectures and teach seminars at UCLA and elsewhere for many years thereafter and remained professionally active until shortly before his death. Even after more than half a century of teaching and research in EM, Garfinkel never seemed to lose his enthusiasm or burn out.

He received the Cooley Mead award from the American Sociological Association for his lifetime contributions to social psychology in 1995, was awarded the degree of Doctor of Letters honoris causa by the University of Nottingham in 1996, and received the Liftetime Achievement Award from the Ethnomethodology/Conversation Analysis Section of the American Sociological Association in 2007. At the time of his passing, Garfinkel was working on an unfinished book tentatively entitled, *Lebenswelt Origins of the Sciences* (Garfinkel, 2006). He also left behind an extensive body of unpublished materials including papers, transcribed lectures, and tape-recorded lectures and conversations with students and colleagues. Garfinkel's landmark publication, *Studies in Ethnomethodology* (*Studies*, 1967, reprinted in 1984) remains his seminal work and is EM's primary text.

Garfinkel's Local Cohort of Students and Colleagues

EM represents a radically different understanding of the problem of social order and structured social action. Garfinkel's groundbreaking work provided the source and context for understanding the work of others who came later and whose work was encompassed under the banner of EM. Much of this work was done by Garfinkel's students and colleagues at UCLA and UC Irvine. His earliest cohort included Harvey Sacks, Emanuel Schegloff, and Gail Jefferson (who founded the subfield of conversation analysis ('CA'), Sacks et al., 1974). Garfinkel's idea that people in society organize their activities to be mutually intelligible and recognizable to each other on specific occasions and in actual contexts of activity inspired the development of conversation analysis. The 'indexicality' of meanings communicated in particular situations, how people make inferences, the real-time, contingent character of social action, and the consequentiality of breaching the expectancies of daily life were ethnomethodological insights underpinning CA. Conversational analysts study orderly phenomena in ordinary conversation and look at 'conversational structures,' like turn taking and sequence structure, adjacency pair organization (e.g., question/answer), story-telling structure, and the negotiation of preferences (e.g., preference to accept rather than decline an invitation). CA proposes that the turn-taking system

and other publicly available conversational structures comprise a relatively limited number of 'constitutive rules' that organize how members of a society do a social/speech act and interpret talk (e.g., how to do a 'greeting' and under what circumstances). From this perspective, turn taking in conversation is a normative system oriented to by interactional participants who tend to conform to its specified rules (e.g., 'one-speaker-at-a-time') and to treat departures from these rules as sanctionable violations. For Garfinkel, 'one-speaker-at-a-time' (or any other orderly phenomena in conversation) is premised on the production and recognition of just what talk is mutually oriented to as a 'conversation' in the first place, without resort to concepts of rules or rule systems.

At UCLA, Garfinkel also influenced the ideas of his colleague Melvin Pollner, especially Pollner's work on 'mundane reasoning' (1974, 1975, 1987). While Garfinkel's breaching experiments explored the practices of sense-making by manufacturing fractures in the sense of collectively shared reality, Pollner (1974, 1975) examined naturally occurring settings "that foster interpretive asymmetries … as a matter of course," especially organizational settings, such traffic court sessions or psychiatric observations of 'delusional' patients (Lynch, 2012b: pp. 70–71). Despite disagreements in these settings among experiencers and observers of events about what actually happened, Pollner found that what results is not an agreement that multiple worlds exist, but rather the use of several members' practices to reconcile the competing versions. His analysis of 'reality disjunctures' (Pollner, 1979) elaborated the presumed existence of a single, world-known-in-common that operates as an unyielding background expectancy and is sustained even "in the face of obstinate disagreement about it [or] … any momentary anomaly or disjuncture" (Lynch, 2012b: pp. 69–70).

Among Garfinkel's earliest graduate students at UCLA were Egon Bittner and D. Lawrence Wieder. The collaborative work of Garfinkel and Bittner (reported in Chapters 6 and 7 of *Studies in Ethnomethodology*, Garfinkel, 1967) represents a fundamental contribution to professional sociology's interest in 'organization,' as well as "innovations in the study of coding and classifying, and the agenda for and use of textual records in sociological inquiries" (Slack and Carlin, 2013: p. 2). Wieder's dissertation (published in 1974) was an ethnography conducted among residents, staff, and Wieder (as the researcher) in a half-way house for paroled drug offenders. The study described the relevance of 'telling the convict code' in that setting and developed the ethnomethodological idea of rules as methods for formulating settings. Wieder showed that reference to the maxims of the 'code' (e.g., do not snitch; do not interfere with other residents' interests; do not trust staff) was a consequential social practice that served to persuasively structure the sense-making and practical action in specific circumstances, but that its use in providing for the accountability of conduct could also be challenged by those present in the half-way house.

EM reconsiders "what it means to produce observations, descriptions, and explanations of something 'actual'" (Lynch, 1993: p. 113). It has no imperative or prohibited set of research procedures or methods, so long as the data obtained is adequate to the particular phenomena under investigation. The idea is "to identify the practices through which social order is performed in everyday life" (Lynch, 1984: p. 333). The widespread availability and increasing use of audio-video recording technology in the 1970s enabled ethnomethodologists to record and study the social world and its activities in real-time and contingent audio-visual detail, yielding unprecedented discoveries of the everyday methods for producing social order. EM details the sights and sounds of real-worldly affairs and can help the human sciences do better job, in a manner akin to using a more powerful microscope (Heath, 1986: p. 4): "The point was to recover a kind of technical 'oriented to' sense of spontaneous communicative actions" (Lynch, 1984: p. 334). When ethnomethodologists examined social activity in greater detail, they found that they would lose the phenomena, if they did not see it in operation. Currently, there are two active and productive branches of ethnomethodologically informed research: conversational analysis (e.g., Sacks et al., 1974), including 'institutional talk' (Drew and Heritage, 1992), and ethnographic studies of workplace activities. 'Studies-of-work' focus on the practical action and tasks of participants in different settings – on what people are trying to do and how they are doing it.

From the late 1970s onward, Garfinkel encouraged his students to investigate specialized domains of scientific, professional, or technical practice and to conduct 'hybrid studies of work' that would cross traditional academic boundaries, forming the basis for interdisciplinary cross-fertilization between EM and other disciplines (Burns, 2001, 2009; Livingston, 1986; Lynch, 1985, 1993; Macbeth, 1996, 2012). As part of Garfinkel's 'unique adequacy requirement of methods,' several of his students attempted to master the technical practices of particular fields of inquiry before explicating their local organization, raising the question of whether researchers must 'go native' to be able to understand and know how to produce/reproduce the materially detailed work constituting a specific domain of scientific, technical, or professional practice. Garfinkel's recommendation to 'hybridize' EM with other disciplines, such as mathematics, the natural sciences, or law, was part of his ambitious studies-of-work program in which "the long-term aim was not only to enrich the sociology of the sciences, arts, and other occupations and preoccupations, but [that] … ethnomethodology would take a distinctive form as it became integrated with the pedagogies and practices of music, law, medicine, mathematics, and so on" (Lynch, 2012a: p. 165).

The studies-of-work of Garfinkel and his students self-consciously set aside conventional theories of social action and order and "proposed a radically new approach to studies of workplace activities, including work in the sciences and technical professions, which would encompass the contents of practices, and not only the organizations that would trace back to sociology's toolbox of concepts and variables" (Lynch, 2012a: p. 165). These studies turned the theoretical problem of social order into a practical procedural problem: "In ethnomethodology, 'procedural' means labor of a certain incarnate methodological sort: at the workplace, progressively and developingly coming upon the phenomenon via the work in … the … directly observed phenomenal field detail of producing it" (Garfinkel, 1996: p. 10, ftn. 11). What Garfinkel called the 'corpus' of ethnomethodological studies-of-work includes Michael Lynch's research in scientific

laboratories (1985, 1993), Eric Livingston's investigation of mathematician's work (1986), Britt Robillard's study of medical practice and the 'lived experience' of disability (1991), Kenneth Liberman's examination of philosophical debates among Buddhist scholar-monks (2004), Douglas Macbeth's studies of educational settings (1996, 2001), and Stacy Burns's research in legal contexts (2001, 2009).

Short Description of Garfinkel's Major Works and Publications

Garfinkel's early work in the 1950s and 1960s was strongly influenced by Alfred Schutz and the practical questions he asked about 'verstehen,' or how people come to achieve inter-subjectivity and create a world of mutual understanding. For Schutz (1962), the maintenance of a world-known-in-common relies on society-members knowing and using 'common understandings' and typified 'recipe knowledge' to engage in practical courses of action on a moment-by-moment basis. His belief that the 'attitude of everyday life' had an organized structure governed by cognitive norms was based on the idea that members' knowledge and skills exist prior to actual situations and that social order and human action could be investigated independent of actual social contexts and contingencies. Garfinkel's publication on 'trust' (1963) relied on Schutz's notions of 'background expectancies' and the maintenance of the 'reciprocity of perspectives' to make social life work and what it is to ourselves and others. For Garfinkel, the notion of 'member' suggested that people who belong to a specific society possess knowledge of the "socially-sanctioned-facts-of-life-in-a-society-that-any-bona-fide-member-of-the-society-knows" and can trust that others in society also know (1967: p. 76). 'Trust' was "compliance with the expectancies of the attitude of daily life as a morality" (1967: p. 50, ftn. 5). It permeated not only the intelligibility of mundane affairs, but also their moral accountability. In his famous breaching experiments, Garfinkel would use a series of experiments to demonstrate the fragility of trust, along with "the suspension of indefinite possibilities of mistrust … that maintains normal social order in the absence of unequivocal rules" (Lynch, 2012a: p. 164; Garfinkel, 1963).

Chapter 1 of *Studies in Ethnomethodology* described EM's unique and distinctive subject matter as an essentially unexplored domain of phenomena: the situated production of social order. To investigate this domain, Garfinkel recommended that the stable and organized features of ordinary settings and activities be treated as the contingent achievements of the participants to those settings and actions. In presenting this new understanding of society and social action, he not only provided a revealing correction on what had previously been presented by virtually every other conventional theory of social order, but also set forth an incommensurably alternate research agenda.

In Chapter 2 of *Studies*, Garfinkel reported the findings from his so-called 'breaching experiments,' in which the experimental procedure was "to start with a system with stable features and ask what can be done to make for trouble" (Garfinkel, 1963: p. 187). The familiar, taken-for-granted routines and background understandings of households, retail businesses, or children's games were suspended in these exercises by student-researchers who undertook to act like a polite lodger in their own home ignorant of the informal routines, or to violate the 'institutionalized one price rule' at a retail store by bargaining over prices (Garfinkel, 1967: p. 69), or to disregard the rules of play in a game of tic-tac-toe. These disruptions produced much anxiety and indignation in the research subjects, demonstrated the unraveling of trust, and prompted "reflections through which the strangeness of the obstinently familiar world can be detected" (Garfinkel, 1967: p. 38). The breach provided a 'perspective by incongruity,' which made visible the assumptive and otherwise unnoticed orderliness of normal environments and offered first-hand access to their constitutive practices, such that "[p]aradoxically, the breach constituted the presumption of the prior agreement" (Lynch, 2012a: p. 164).

In Chapter 3 of *Studies*, Garfinkel explored how background knowledge is relied upon and used by society-members in everyday life in a manner akin to the 'documentary method of interpretation.' The 'method,' according to Garfinkel, consisted of treating actual appearances as standing on behalf of a presupposed pattern. But not only were the appearances used to infer the pattern, the pattern was also utilized to decide what the appearances were appearance of. Garfinkel designed several experiments to show the documentary method in operation. In one experiment, student research subjects attempted to receive counseling by asking questions they understood a 'counselor' would answer either 'yes' or 'no' (Garfinkel, 1967: pp. 79–80). The subjects asked a series of questions, received a response, and then asked another question, tied to the answer they had received to the prior question. Unknown to the subjects was the fact that the answers were predetermined according to a table of random numbers. Nonetheless, the subjects were able to get through the question-answer interchange and believed they had participated in a meaningful exchange in which they had received answers to their specific questions and "advice … for managing their affairs" (Garfinkel, 1967: p. 94). Garfinkel observed that the documentary method was a ubiquitous sense-making device in social life and exhibited the part–whole relationships of a Gestalt context (Gurwitsch, 1964). His exercise also raised important questions about "the validity of the 'underlying patterns' discerned through the use of this method" (Lynch, 2009: p. 89), questions that are broadly relevant to the work of using documents and other 'evidence' in historical reconstruction.

Chapter 4 of *Studies* reported on Garfinkel's earlier research on jury deliberations, to which he attributed the origins of EM. When Garfinkel examined the tape-recorded jury deliberations, he noted that despite the jurors' lack of formal legal training and credentials, they nonetheless diligently engaged in 'methodological' inquiries to assess evidence, decide upon disputed issues, and reach a 'just and fair' verdict in the case (Garfinkel, 1967: p. 105). With this insight, he began a series of lifelong investigations into the relationship between scientific/professional methods and reasoning and 'vulgar' or commonsense reason and methods.

In Chapter 5 of *Studies*, Garfinkel rejected received versions of gender that simply posit its relevance and

explicate a priori cultural assumptions and definitions. Instead, investigated how gender is a moment-by-moment accomplishment and specified the practices of managing one's taken-for-granted social (gender) identity. The subject of the study, 'Agnes' (pseudonym), was a transsexual person who was born with normal male genetalia and raised as a boy until age 17, when he adopted a female gender identity. Garfinkel met her at age 19, after she was referred to the UCLA Department of Psychiatry for counseling and ultimately a sex-change operation (i.e., sex reassignment surgery). Although he found Agnes' appearance to be 'convincingly female' (Garfinkel, 1967: p. 119), she regarded her efforts to 'pass' unquestionably as female as 'tests,' requiring her to become a quick study and to use constant diligence so her secret would not be revealed. Notwithstanding that gender is usually considered to be dichotomously determined at birth, as either male or female, Garfinkel's investigation made perspicuous the infinitely many, yet otherwise 'seen-but-not-necessarily-noticed,' ways that social (gender) identity is done, recognized, and made accountable in almost every facet of daily life.

In Chapter 6 of *Studies*, Good Organizational Reasons for 'Bad' Clinic Records, Garfinkel described a study conducted with Egon Bittner and aimed at developing a systematic empirical account of patient pathways through a medical clinic. Relying on clinic records as the primary data, he highlighted the many problems that were encountered by researchers in using the records (e.g., information recorded in files was not standardized, file folders were incomplete, and information of importance to the researchers was missing). Yet instead of abandoning the study, or reporting its findings directly, Garfinkel turned his attention to the 'good organizational reasons' for the deficient clinic records and described how clinic staff 'managed' the records and case files to document their activities. His analysis offered a powerful conceptual tool for investigating 'the substantive production of history,' by "transform[ing] the historian's problem into a substantive, constitutive phenomenon … a matter of producing documents, withholding details, and collecting them into files before any investigator gets hold of them … " (Lynch, 2009: p. 92).

Chapter 8 of *Studies* relied on Schutz's work to explore 'the rational properties of scientific and commonsense activities.' While Schutz "drew a strict demarcation between scientific theory and scientific practice and between scientific and commonsense rationality" (Lynch, 1993: p. 134), Garfinkel found that "'commonsense' rationalities were included in the principles and operative practices of science [and] provided a radical antidote to the idea that commonsense was the domain of 'prescientific' notions" (Lynch, 1993: p. 139). He thereby "pluralized 'rationality'" to become "the rationalities of science and commonsense" (Lynch, 2012b: p. 72) and deprivileged scientific/professional over lay knowledge and competencies (Lynch, 1993: p. 139). Unlike Schutz, Garfinkel asked how 'rationality' and intersubjectivity are achieved in actual contexts of activity. He worked out the 'properties of rationality' by using the phenomenological method of bracketing the taken-for-granted, objective existence of 'what is really going on,' instead treating it as 'empirically problematical material' (1967: p. 282) and as researchable things. Garfinkel also diverged from Schutz's treatment of science and other practical action as governed by distinctive cognitive norms (Lynch, 1993: pp. 140–141) and demonstrated the intertwining of scientific (or technical/professional) knowledge and expertise with commonsense knowledge and competencies (e.g., Burns, 2001, 2009).

In a later publication on 'Ethnomethodology's Program' (Garfinkel, 1996), Garfinkel explicitly linked EM to the work of Emile Durkheim and particularly to what he termed Durkheim's 'aphorism,' that "the objective reality of social facts is sociology's fundamental principle" (Durkheim, 1895/1982). Garfinkel subsequently clarified this connection, stating (as only he could put it) that "ninety years of Durkheim scholarship … missed from the beginning … that the objective reality of social facts was a stand in for every topic of logic, meaning, reason, method, and order in intellectual history specified as an achieved empirical phenomenon of immediate, unmediated…practices, production, recognition and witness" (Garfinkel, 1998, Course Announcement, 'Ethnomethodology,' UCLA Sociology Department). According to Garfinkel's intentional misreading of Durkheim, 'social facts' were to be reinvestigated and respecified by EM as situated practical achievements and work-site practices.

Perhaps not surprisingly, given this alternate reading and 'respecification' of Durkheim, recent writings have attempted to trace the origins of EM back to Durkheim's 'classic' scholarly program (Rawls, 1997), or to tie EM to the initiatives of American pragmatist theorists (Erylbaum and Maynard, 2011). The suggested link with pragmatism is apparently based on its call for the recovery of concrete practices and focus on language use in actual situations (Erylbaum and Maynard, 2011). However, such attempts to link EM back to 'classic' theories (Durkheimian or pragmatist) seem puzzling for some EM scholars because Garfinkel insisted throughout his career that EM makes a 'radical break' with formal theory and analysis and instead provides foundational analyses of society-members' everyday routines and practices for achieving social order. Both Durkheimian and pragmatist formulations leave essentially unexplored and unrecoverable the core phenomena of ethnomethodological interest: the situated organization of practical tasks and activities. To paraphrase Garfinkel, such formulations are not a solution to the problem of social order. Rather, the real-worldly affairs that sustain the professional interpretations are lost in the interpretations and neither Durkheimian nor pragmatist theorists have an interest in how social order is produced and achieved, or how it could be taken seriously, analytically, or descriptively.

For many years and until fairly recently, *Studies in Ethnomethodology* was Garfinkel's only book, so to a large extent his legacy and contribution are "embodied in the work of others on whom he had a profound influence" (Lynch and Sharrock, 2003: p. x) and who pursued his research agenda. EM's productivity as a research approach is evident in a wide range of studies-of-work conducted by Garfinkel and his students in a diverse variety of settings. This corpus of EM studies is represented in an early collection of definitive works (Turner, 1974) and in two more recent, multivolume

collections of ethnomethodological work (Lynch and Sharrock, 2003, 2011).

Lasting Impact of Garfinkel's Work

Garfinkel revolutionized the field of sociology and the human sciences more generally when he founded EM as a discipline for investigating "members' ordinary practices for assembling intersubjectively accountable worlds" (Macbeth, 2001: p. 48). Rather than providing causal explanations or predictions, Garfinkel and his students developed detailed descriptions of commonplace and specialized organizations of practical action and practical reason that went beyond sociology's existing toolbox of concepts and variables. His ground-breaking insights into how people in society 'make sense' dissolved the supposed chasm between common sense and scientific/professional reasoning, unpacked the taken-for-granted existence of society, and specified the practices by which people produce and sustain the orderliness of their everyday activities and the sense that they live and mean in a shared orderly world.

In seminars and classes, Garfinkel sometimes played a videotape for students to demonstrate the difference between the 'lived work' of a practice (Garfinkel, 1989) and theorized versions of it. The video showed a student typing at the keyboard while speaking a moment-by-moment commentary on exactly what she was doing as she typed. The tape made evident an influx of detail (e.g., a variety of pauses (to think of something to say, or stop from running off the page), backspacing to cross out a passage and restart it, etc.). Garfinkel proposed that the videotape of the 'lived work' of typing and the resulting typed document were 'asymmetric alternates' to each other, just as EM was to an 'asymmetric alternative' to virtually every other existing formal theory and analysis of social action and order. While "one could derive from the video how the final text was composed in a continuously unfolding series of actions, … one could not read the final typescript in isolation and recover the 'lived work' that produced it" (Lynch, 2012a: p. 166).

Garfinkel's exercises on 'instructions' and 'instructed action' (Garfinkel, 1989, 2002) further elaborated this ethnomethodological property of 'reflexivity,' i.e., treating the features of a setting and the action in that setting as situated accomplishments of the participants to that setting and activity. The exercises involved students following a variety of instructions, such as instructions for assembling parts into a finished chair, or making a way-finding journey to someone's home by using an 'occasioned map,' or complying with a simply request like 'sit over there' while wearing inverting lenses (which made problematic doing even the most routine things in seemingly familiar environments by literally turning the familiar world upside down (Garfinkel, 2002)). In each case, he showed that there was something more to the present sense, meaning, order, relevance, reason, and method of the instructions for the user than the written text or verbal instructions provided, and that this 'what more' was a reflexively embedded, *in vivo* detail of the constitutive practices that made up the course of following the instructions, or instructions-in-use (Garfinkel, 1989, 2002). EM addresses the 'instructable reproducibility' of a practice (Garfinkel, 1989) and in this sense its findings "can be likened to manuals of instructions that describe how the practices being studied are produced and performed … [and] delve into non-obvious features of situated conduct that elude rule-governed models and cognitive schemes" (Lynch, 1984: p. 334).

The corpus of EM studies demonstrates a radically unorthodox understanding of the 'objective reality of social facts' and 'immortal ordinary society': an independently existing world of specific occasions, practical tasks, embodied skills, and contexts of activity. As Garfinkel described EM, it "is concerned with 'What More' immortal, orderly society could possibly consist of, as the locus and the setting of every topic of order, every topic of logic, of meaning, of method, specified and respecifiable as the most ordinary Durkheimian things in the world" (Garfinkel, 1996: p. 6). Of course, just what that could mean is only to be found in the material examination of actual cases, as Garfinkel always insisted.

See also: Bureaucracy, Sociology of; Durkheim, Emile (1858–1917); Macrosociology-Microsociology; Parsons, Talcott (1902–79); Phenomenology in Sociology; Professions, Sociology of; Science and Technology Studies, Ethnomethodology of; Science and Technology Studies, History of; Situated Learning (Learning In Situ); Sociological Theory; Trust, Sociology of; Work and Industry, Sociology of.

Bibliography

Burns, S., 2001. 'Think your blackest thoughts and darken them': judicial mediation of large; money damage disputes. Human Studies 24, 227–249.

Burns, S., 2009. Doing justice and demonstrating fairness in small claims arbitration. Human Studies 32, 109–131.

Drew, P., Heritage, J., 1992. Talk at Work. Cambridge University Press, Cambridge, UK.

Durkheim, E., 1895/1982. Rules of the Sociological Method. The Free Press, New York.

Erylbaum, M., Maynard, D., 2011. Pragmatism and ethnomethodology. Qualitative Sociology 15, 221–245.

Garfinkel, H., 1941. Color trouble. In: O'Brien, E.J. (Ed.), The Best Short Stories of 1941: The Yearbook of the American Short Story. Houghton Mifflin, Boston, pp. 97–119.

Garfinkel, H., 1949. Research note on inter- and intra-racial homicide. Social Forces 47, 369–381.

Garfinkel, H., 1963. A conception of, and experiments with, 'trust' as a condition of stable, concerted action. In: Harvey, O.J. (Ed.), Motivation and Social Interaction. Ronald Press, New York, pp. 187–238.

Garfinkel, H., 1967. Studies in Ethnomethodology. Prentice Hall, Englewood Cliffs, NJ.

Garfinkel, H., 1989. Instructions and instructed actions. Lecture, Department of Sociology. University of Oklahoma.

Garfinkel, H., Wieder, D.L., 1992. Two incommensurable, asymmetrically alternate technologies of social action. In: Watson, G., Seiler, R.M. (Eds.), Text in Context: Contributions to Ethnomethodology. Sage, Newbury Park, California, pp. 175–206.

Garfinkel, H., 1996. Ethnomethodology's program. Social Psychology Quarterly 59, 5–21.

Garfinkel, H., 1998. Unpublished course announcement for Sociology 271, Ethnomethodology, Spring Quarter. Department of Sociology, UCLA.

Garfinkel, H., 2002. Ethnomethodology's Program: Working Out Durkheim's Aphorism. Rowman and Littlefield, Lanham, MD.

Garfinkel, H., 2006. The Lebenswelt Origins of the Sciences. Department of Sociology, UCLA.
Garfinkel, H., et al., 1981. The work of a discovering science construed with materials from the optically discovered pulsar. Philosophy of the Social Sciences 11, 131–158.
Gurwitsch, A., 1964. Field of Consciousness. Duquesne University Press, Pittsburg.
Garfinkel, H., Sacks, H., 1970. On formal structures of practical actions. In: McKinney, J.C., Tiryakian, E.C. (Eds.), Theoretical Sociology: Perspectives and Developments. Appleton-Century-Crofts, New York, pp. 337–366.
Livingston, E., 1986. The Ethnomethodological Foundations of Mathematics. Routledge and Kegan Paul, New York.
Lynch, M., 1984. Ethnomethodology. In: Lewis Beck, M., et al. (Eds.), Sage Encyclopedia of Social Science Research Methods. Sage Publications, Thousand Oaks, CA, pp. 333–334.
Lynch, M., 1985. Art and Artifact in Laboratory Science. Routledge and Kegan Paul, London.
Lynch, M., 1993. Scientific Practice and Ordinary Action: Ethnomethodology and the Social Studies of Science. Cambridge University Press, Cambridge, UK.
Lynch, M., 2009. Ethnomethodology and history: documents and the production of history. Ethnographic Studies 11, 87–106.
Lynch, M., 2012a. Garfinkel stories. Human Studies 35, 163–168.
Lynch, M., 2012b. Philosophy on the ground: an appreciation of a 'spirit master. American Sociologist 43, 67–75.
Lynch, M., Sharrock, W. (Eds.), 2003. Harold Garfinkel, vols. 1–4. Sage, London.
Lynch, M., Sharrock, W. (Eds.), 2011. Ethnomethodology, vols. 1–4. Sage, London.
Macbeth, D., 1996. The discovery of situated worlds: analytic commitments or moral orders? Human Studies 19, 267–287.
Macbeth, D., 2001. On 'reflexivity' in qualitative research: two readings, and a third. Qualitative Inquiry 7, 35–68.
Macbeth, D., 2012. Some notes on the play of basketball in its circumstantial detail. Human Studies 35, 193–208.
Merleau, P.M., 1962. The Phenomenology of Perception. Routledge and Kegan Paul, London.
Pollner, M., 1987. Mundane Reason: Reality in Everyday and Sociological Discourse. Cambridge University Press, Cambridge, UK.
Pollner, M., 1974. Mundane reasoning. Philosophy of the Social Sciences 4, 35–54.
Pollner, M., 1975. The very coinage of your brain: the anatomy of reality disjunctures. Philosophy of the Social Sciences 5, 411–430.
Pollner, M., 1979. Explicative transactions: making and managing meaning in traffic court. In: Psathas, G. (Ed.). Everyday Language. Irvington, New York, pp. 227–253.
Rawls, A., 1997. Durkheim and pragmatism: an old twist on a new problem. Sociological Theory 15, 5–29.
Sacks, H., et al., 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50, 696–735.
Schutz, A., 1962. Collected Papers I: The Problem of Social Reality. Martinus Nijhoff, The Hague.
Slack, R., Carlin, A., 2013. Egon Bittner: approaching explications. Ethnographic Studies 13, 1–17.
Sudnow, D., 1978. Ways of the Hand: The Organization of Improvised Conduct. Harvard University Press, Cambridge, MA.
Turner, R. (Ed.), 1974. Ethnomethodology: Selected Readings. Penguin, Harmondsworth, U.K.
Wieder, D.L., 1974. Language and Social Reality: The Case of Telling the Convict Code. Mouton, The Hague.

Gastrointestinal Diseases: Psycho-social Aspects

Antonina Mikocka-Walus, Mental Health and Addiction Research Group, Department of Health Sciences, University of York, York, UK
Lesley Graff, Department of Clinical Health Psychology, Faculty of Medicine, University of Manitoba, Winnipeg, MB, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

An interconnection between psychological and gastrointestinal (GI) functioning has long been recognized, incorporating multiple mechanisms and considering bidirectional processes. However, exciting discoveries regarding the role of stress and depression in etiology and disease course have shed new light on the understanding of biopsychosocial processes in chronic GI conditions. This article provides an introduction to GI tract functioning, GI disorders, and stress mechanisms in the gut, followed by an overview and discussion of the psychosocial impact of these disorders, the role of stress and mental comorbidity in GI disorders, and the current knowledge regarding psychological interventions for GI disorders.

Background

Historically, a link between psychological and gastrointestinal (GI) functioning has been well recognized. The term 'nervous diarrhea' started appearing in medical texts in the mid-1800s. In the early 1900s, physiologist Walter Cannon observed that emotional calmness was important to normal digestion, and speculated that mental distress may contribute to disturbed digestion. For several decades in the twentieth century, there was a predominant view that psychodynamic conflicts played a central role in the cause of ulcerative colitis, an inflammatory bowel disease (IBD), supported in part by Cecil Murray's published case reports in 1930. This link between psychology and gut functions has worked its way into folk wisdom and common language, with expressions such as "that took a lot of guts" and "butterflies in the stomach." The contemporary understanding of the interplay between psychological functioning and GI disease is much more complex, incorporating multiple mechanisms and considering bidirectional processes.

The next sections provide an introduction to GI tract functioning, GI disorders, and stress mechanisms in the gut, followed by an overview and discussion of the psychosocial impact of these disorders, the role of stress and mental comorbidity in GI disorders, and the current knowledge regarding psychological interventions for GI disorders.

The GI Tract and the Brain–Gut Axis

The human GI tract extends from the mouth to the anus, and includes the esophagus, stomach, approximately 13 feet of small intestine, and 7 feet of large intestine. The GI tract has a crucial role in survival, as it is responsible for breaking down food, extracting nutrients, exchanging fluids, and expelling waste. It is a dynamic organ that actively moves contents through the tract using coordinated ripple-type muscle contractions. A microworld of bacteria, known as the gut flora or gut microbiota, exists throughout the GI tract, and facilitates the breakdown and processing of foodstuffs. This is the most highly innervated organ in the body, and has the largest concentration of neurotransmitters outside the brain, including approximately 90% of the body's serotonin, supporting the view of a highly developed brain–gut communication pathway. The GI tract is regulated through its own internal nervous system, the enteric nervous system, and the central nervous system, via a complex interaction of the neuro–endocrine–immune systems, with the autonomic nervous system playing a central role in brain–gut communication (Bonaz and Bernstein, 2013).

This bidirectional connection system is known as the brain–gut axis. While gut action typically occurs without much awareness or focus, visceral afferent information from the GI tract has been shown to be more available to conscious awareness than other body systems (e.g., cardiovascular). Actions including gastric emptying, transit time, and small and large intestine contractions can be altered by psychological processes such as emotion and stress (Mayer et al., 2001). Certainly, gut upset (e.g., loose stool, nausea) is a common experience for many when stressed.

Organic and Functional GI Disorders

Many of the disorders of the GI tract are attributed to motor or sensory disturbances. The GI disorders are typically categorized as organic or functional; the organic disorders are those defined by structural or inflammatory abnormalities, and the functional disorders reflect chronic or recurrent GI symptoms in the absence of these abnormalities, suggesting a problem in the functioning of the gut. While the functional disorders are not typically associated with greater health morbidity or mortality, the organic disorders can be. Regardless, both types of GI disorders result in significantly poorer quality of life even compared to those with other chronic diseases, affecting social, personal, and work domains of functioning.

Organic GI disorders include the acid peptic diseases (e.g., gastroesophageal reflux disease) and IBD. IBD is characterized by recurring inflammation in the GI tract leading to abdominal pain, bloody diarrhea, and fatigue. IBD has been increasing in prevalence in recent decades, with current estimates as high as 500 per 100 000. The peak incidence is in young adulthood (second and third decades); however, 25% of new cases are diagnosed in children and adolescents. While IBD is clearly less

common than its functional counterpart, irritable bowel syndrome (IBS), it can be difficult to achieve and maintain remission for this condition, and it often requires costly medication and surgery for disease management. Although rare, it can also be associated with mortality. The pathogenesis of IBD is thought to involve an exaggerated immune response to normal microbes in the intestine, triggered by environmental factors in those with a genetic susceptibility.

Functional disorders are generally more prevalent than organic, with IBS the most common type. IBS is a chronic relapsing condition involving bowel habit disturbance and lower abdominal pain or discomfort. It affects approximately 10–20% of the adult population, has peak onset in the 30s and 40s, and is one of the 10 most common reasons to visit the primary care physician. Nonulcer dyspepsia, with an estimated prevalence of 12–15%, is an upper gut disorder with symptoms of pain, bloating, and nausea; the symptoms are similar to those of an ulcer, but investigation (upper endoscopy) does not show an ulcer to be present. It is thought that these conditions arise because of sensory nerve disturbances along the brain–gut axis.

Stress and GI Disorders

While many of the GI disorders have differing etiologies, clinical and experimental research suggests that stress may be a common risk factor affecting disease course, and comorbid psychological disturbances may be higher for those with GI disease than the base rate in the community, as detailed in the next sections. It is thought that the special link between the brain and gut, and in particular, disruptions in communication along the brain–gut axis, may help to explain these associations (Bonaz and Bernstein, 2013). Altered gut motility and visceral sensitivity have been seen as fundamental processes in the modern understanding of functional GI disorders. Regions in the brain that regulate emotional and cognitive functions, including the amygdala, hippocampus, and prefrontal cortex, are also involved in modulating gut function (Mayer et al., 2001).

Experimental research suggests that stress can prompt a range of GI tract disturbances such as increased intestinal permeability, mucosal inflammation, and gastric ulcers (Caso et al., 2008). At cell level, mucosal mast cells, and at endocrine level, corticotrophin-releasing factor have a central role in immune system changes and inflammation of the mucosa. The autonomic nervous system is highly sensitive to emotional and evaluative input (e.g., the 'fight or flight response'). Recent work suggests that there may be autonomic dysfunction in individuals with organic or functional GI disorders, although the specific nature of the sympathovagal balance, including parasympathetic hyporeactivity, may be different among disease types (Pellissier et al., 2010). Finally, the gut microbiome is just beginning to be explored as another pathway through which psychological processes such as mood disorders may interact with gut function. Converging research is suggesting that intestinal microbes might be involved centrally in modulating behavior and brain chemistry, and that environmental influences such as stress can affect the microbial balance in the gut (Bercik et al., 2012) (**Figure 1**).

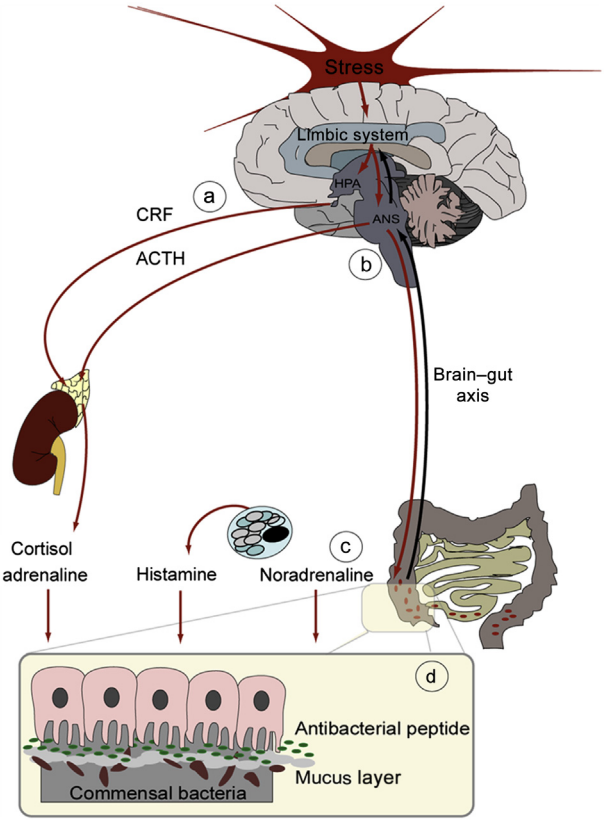

Figure 1 Psychoneuroimmunological pathways in inflammatory bowel disease. Stress, via the limbic system, stimulates the HPA (a) and the ANS (b) leading to the production of CRF, ACTH, and then cortisol and the catecholamines. Noradrenaline is also released directly into the gut mucosa from sympathetic nerve endings of the ENS (c). In addition, mast cells lying close to enteric nerve endings degranulate releasing tryptase, histamine, and other mediators (d). These neuromediators through incompletely characterized mechanisms, including activation of lymphocytes, stimulate epithelial chloride secretion and mucus production, increase paracellullar and transcellular epithelial permeability. Subsequent ingress of luminal bacteria into the mucosa triggers further immune and inflammatory events involving dendritic cells, macrophages, and other cell types. ACTH: adrenocorticotropic hormone; ANS: autonomic nervous system; CRF: corticotrophin releasing factor; ENS: enteric nervous system; HPA: hypothalamic–pituitary–adrenal axis. Originally published by Wahed, M., Rampton, D., 2013. Psychological stress and related mood disorders, and their therapeutic implications in IBD. IBD Monitor 13, 143–153. Reproduced with permission from Remedica Medical Education and Publishing.

Psychosocial Impact of GI Disorders

Health-Related Quality of Life

Health-related quality of life (HRQOL) in patients with chronic conditions encompasses the effects of these conditions and associated treatment on physical and psychosocial functioning of the person living with the illness. HRQOL frequently has been found to be impaired in those with chronic GI disorders. For example, a systematic review concluded that HRQOL is significantly reduced in patients with moderate-to-severe IBS, at a level comparable to other chronic disorders such as depression (El-Serag et al., 2002).

In organic GI conditions such as IBD, impaired HRQOL may result from its incurability, unpredictability, and severity of symptoms, surgery, and medication side effects, although psychosocial factors may also play a role. A meta-analysis found that disease-related factors only accounted for 37% of the variance in HRQOL (van der Have et al., 2014). The authors concluded that other variables such as illness perception, sociodemographic factors, and mental health are also contributory, and addressing those which are modifiable could improve the HRQOL. In the functional disorders, diminished HRQOL is attributed to disease-related factors including unclear diagnosis and unavailability of universal treatment, combined with debilitating symptoms and avoidance of potential symptom triggers.

When HRQOL was directly compared between patients with organic and functional GI conditions in a large ($n = 4441$) population-based study exploring self-perceived health status, IBD participants reported poorer physical health, and IBS participants reported poorer mental health and greater levels of stress (Tang et al., 2008), demonstrating the interrelatedness of somatic and psychosocial influences on HRQOL. In many chronic conditions, including GI disorders, disability and poorer HRQOL can be more strongly associated with underlying mental illness or psychological distress than the disease per se, perhaps as depression and/or comorbid anxiety can aggravate disease activity and negatively impact on disease management.

HRQOL can be compromised for young people with IBD as well (Greenley et al., 2010), as the impact of the disease and treatment is evident in the overall development, with common short stature and delayed puberty contributing to a visible difference from peers. Pediatric patients with IBS also report poorer HRQOL compared to healthy peers, although the impact of their disease on time missed from school is not as high as those with organic GI conditions (Varni et al., 2006). For both pediatric IBD and IBS patients, poorer HRQOL is often found to be related to embarrassment resulting from frequent use of toilets and not rarely becoming a target of the 'bathroom humor' by the peer group.

Coping

Coping with chronic illness involves managing demands imposed by the illness. These can include physical and mental symptoms, as well as psychological adjustment to changes in the functioning of the body and the social impact of illness on one's roles within the family and society. Comparing the coping approaches between patients with organic and functional GI disorders, studies disagree on whether IBD and IBS patients differ in their preferred coping strategies. While IBS patients have been found to catastrophize more often than IBD patients in some studies, other research does not demonstrate such interdisease differences, showing that both those with IBD and IBS differ from healthy controls, in that the former tend to rely on passive coping approaches. A comprehensive systematic review found that emotion-focused coping was associated with worse psychological outcomes, while the effect of problem-focused coping was, to some extent, associated with better psychological outcomes in IBD (McCombie et al., 2013).

There is a dearth of data with respect to pediatric and adolescent populations' coping with the disease, with only anecdotal reports of less well-developed coping skills than in the healthy counterparts.

Social Support

Social support involves positive assistance (actual and perceived) from a network of friends, family, and others. High perceived social support has been linked with good health outcomes; this is thought to relate to the practical assistance as well as the emotional support that caring individuals can provide. In patients with chronic GI disorders, social support has not had much research attention. Available studies suggest that those with IBD generally feel well-supported, although this may not be the case for individuals with moderate-to-severe functional symptoms. IBD patients reported more tangible, affective, or emotional support and were more likely to have experienced a positive social interaction compared to their community non-IBD counterparts ($n = 388$) (Rogala et al., 2008). A beneficial effect of social support on disease course has also been reported, with higher social support shown to decrease the likelihood of disease deterioration in IBD (Camara et al., 2011). For IBS patients, stress arising from conflict-laden relationships is reported to be a stronger predictor of IBS outcomes than the benefits arising from positive social support in some studies, suggesting a detrimental effect of poor quality support. Online social support groups, a relatively recent phenomenon, are increasingly being reported as important tools in providing social support for patients with GI conditions (Malik and Coulson, 2011).

Social Functioning and Stigma

People with chronic GI disorders frequently report challenges in social functioning. Preoccupation with toilet availability, and symptoms such as diarrhea, bloating, and gas can make social interactions difficult and may lead to isolation. Stigma associated with bowel function can prevent those affected by IBD and IBS from sharing the information about the diagnosis or their needs regarding symptom management. The nature of GI symptoms, the shame around potential loss of control, and the anatomy focus on the bowel and rectum are taboo in many cultures. Social gatherings often include or revolve around food, which can be an obstacle for those with chronic GI disorders due to self-imposed diet restrictions and avoidance of restaurants resulting from fear of problematic ingredients. Individuals who avoid certain social or professional situations (e.g., visit to a theater or a business trip) may have their actions misunderstood and interpreted as evidence of their lack of dedication or negligence, rather than as a consequence of their illness.

According to the largest study to date ($n = 269$ IBS; $n = 227$ IBD), IBS patients feel more stigmatized than their IBD counterparts, which may be explained by the perception of IBS as being a less 'real' disease given lack of pathological changes to the bowel and lack of effective medical treatment (Taft et al., 2011). Despite this, perception of stigma is more highly correlated with poorer patient outcomes in IBD than IBS. Having a GI disorder can adversely affect employability,

particularly for those with more severe illness or difficulty settling symptoms, resulting in loss of income, fewer career prospects, or unemployment.

Social functioning is also reported to be impaired in young people living with GI conditions. While parents report that their adolescents with IBD function comparably to healthy youth, the adolescents report the opposite; namely that their social functioning is poorer than their healthy counterparts and similar to that of other chronically ill young people (Greenley et al., 2010).

Self-Image and Sexuality

People living with GI disorders, particularly disorders that affect bowel habit, frequently report perceptions of being 'unclean' and concerns about managing bodily odors. While there can be problems maintaining hygiene due to fecal leakage, or in the case of IBD, as a consequence of colectomy, often the perception is only subjective and results from the belief that having a bowel problem makes a person 'unclean.'

Disease-affected body image and self-esteem concerns can fluctuate depending on disease status and life stage. In an Australian study of middle-aged IBD patients ($n = 347$ patients), 67% reported negative body image related to their disease, with both men and women endorsing this self-view (Muller et al., 2010). In addition, women and those who had had bowel surgery were more likely to report decreased frequency of sexual activity and decreased libido. While early studies of self-esteem in young people with IBD suggested that low self-esteem and distorted self-image were common, a recent meta-analysis found no difference in self-esteem levels between IBD patients, healthy controls, and young people with other chronic conditions (Greenley et al., 2010). No similar studies have been identified for functional adolescent or pediatric populations.

In functional GI disorders, and IBS in particular, the majority of studies on sexuality focus on childhood adversity and trauma. When this was first examined, patients with IBS were reported to have very high rates of childhood abuse, including sexual abuse (up to 50%), at a prevalence that was significantly greater than healthy controls. However, these findings have been reexamined and the base rates of childhood abuse in the general population have now been found to be high, so there is some controversy about whether IBS is uniquely associated with sexual abuse. In nonulcer dyspepsia, a higher rate of physical and emotional abuse has been observed when compared with healthy controls; however, very few studies have explored this issue.

Mental Comorbidity

Like other chronic health conditions, GI disorders have been associated with high levels of perceived stress and greater prevalence of mental disorders (e.g., anxiety and depression) as compared to the healthy populations. These psychological difficulties may result from disease burden, but may also be a consequence of common pathways related to dysfunction in the brain–gut axis and interconnection of neuro–endocrine–immune systems as discussed above. Despite some similarities, depression, anxiety, and stress may play a somewhat different role in etiology and disease progression in organic and functional disease and are thus discussed separately.

Depression, Anxiety, and Stress in IBD

Population-based studies have documented a higher lifetime prevalence of depression in IBD compared to the community, with estimated rates of 27 versus 12% (Walker et al., 2008). In pediatric and adolescent IBD populations, prevalence of depression was also found to be higher than for youth with other chronic illnesses (Greenley et al., 2010). There is more uncertainty about a higher prevalence of anxiety disorders in IBD patients. While studies have reported higher levels of anxiety symptoms for IBD patients, lifetime prevalence of anxiety disorders was not found to be greater than for matched controls in a population-based study; panic disorder showed a trend for higher rates (8 vs 4.7%), and social anxiety rates were significantly lower in IBD than controls (6 vs 11%) (Walker et al., 2008). Comparisons of rates with other national samples suggest that IBD patients may have higher rates of panic, generalized anxiety, and obsessive–compulsive disorders.

Anxiety and depression have been understood to be disease sequelae, with impact on disease course rather than causal factors for onset. However, studies have found that depression predated the onset of IBD in many cases, and new experimental work with animal models has demonstrated that inducing depression increases susceptibility to spontaneous intestinal inflammation in mice, raising the possibility of depression as a potential etiological risk factor. With regard to disease course, a review of 11 longitudinal studies of depression in IBD concluded that it negatively impacts IBD course, with effects such as a shorter time to relapse (Rampton, 2009).

There is a growing recognition of the role of stress in IBD disease course, independent of mental disorders. A review of nine longitudinal studies of stress identified a relationship with increased inflammation and IBD symptoms (Maunder and Levenstein, 2008), and recent prospective work confirmed higher stress levels in the periods preceding symptomatic disease flare (Bernstein et al., 2010). In addition, studies in animal models have demonstrated that stress can induce colitis and treating the animal with antidepressants restores vagal function and reduces the intestinal inflammation.

Regarding the role of personality factors in IBD, studies have examined personality traits such as neuroticism, perfectionism, and alexithymia and found them to be more common in IBD patients than their healthy counterparts. However, none of these studies have prospectively evaluated personality prior to disease onset, and these traits can be affected by current emotional state.

Depression, Anxiety, and Stress in Functional GI Disorders

While high rates of mental comorbidity, ranging from 40 to over 90%, have been reported for IBS patients presenting to tertiary care centers (Drossman et al., 2002), population-based studies have clarified that the majority of individuals with IBS do not seek medical care, and are similar to non-IBS

individuals with respect to their mental health. Those with moderate-to-severe GI symptoms tend to exhibit higher rates of anxiety and depression, but it cannot be said with certainty whether anxiety and depression precede or follow the GI symptoms. In contrast, individuals with nonulcer dyspepsia, who do not seek medical care, have been found to have poorer scores on psychological domains than healthy controls, suggesting that psychological morbidity might not be associated with health care–seeking behavior of this group (Van Oudenhove and Aziz, 2013).

Few studies have examined mental health in pediatric or adolescent functional GI patients, so rates of comorbid anxiety and depression in this population are largely unknown. There is currently some evidence that patients with recurrent abdominal pain in childhood have an increased risk of IBS during adolescence and young adulthood, and thus possibly a similar likelihood of developing a mental comorbidity at any stage of their life.

The bidirectional relationship between anxiety, depression, and functional GI conditions is well illustrated by the largest longitudinal population-based study to date (Koloski et al., 2012). Individuals with higher levels of anxiety or depression at baseline were more likely to develop IBS, and baseline depression was predictive of subsequent dyspepsia by the 12-year follow-up. Among individuals without elevated levels of anxiety and depression at baseline, those with a functional GI condition at baseline had significantly higher levels of anxiety and depression at follow-up.

Stress is currently viewed as a risk or triggering factor for the onset of IBS and nonulcer dyspepsia, with studies showing higher levels of stress in the months preceding onset. Stress has also been established as a contributor to ongoing symptoms (Drossman et al., 2002). In fact, individual stress-sensitivity appears to be a key factor in the severity and persistence of functional GI symptoms. People reporting these gut symptoms have been found to exhibit exaggerated motor and sensory responses to stressors, including visceral hypersensitivity.

Similar to the research findings in IBD, very few studies have explored personality traits in patients with functional GI conditions. The available ones reviewed by Romanian researchers (Surdea-Blaga et al., 2012) have found higher levels of neuroticism for patients with nonulcer dyspepsia and IBS, compared to healthy controls, with some also reporting higher levels of conscientiousness and alexithymia in IBS. The IBS patients were also found to have higher neuroticism levels than IBD patients. However, it should be noted that neuroticism is a trait which is very susceptible to state influences and has been commonly reported as elevated in many chronic conditions; thus it is unclear whether it should be considered a precursor or a consequence of the chronic illness.

Psychological Therapies in GI Disorders

The close relationship between the brain and the gut via the brain–gut axis, the high prevalence of mental comorbidity in GI disease, and the evident impact of psychological processes on GI symptoms, disease course, and overall quality of life all support that psychological health should be a clinical consideration in GI disease management. Psychological therapies have been used as part of the treatment approach for several of the GI diseases. These psychosocial interventions have been more extensively examined in the functional illnesses such as IBS and nonulcer dyspepsia, but the efficacy of various therapy approaches has also been investigated in IBD. Psychotherapy, and in particular, cognitive behavioral therapy (CBT), a type of psychological therapy that uses an active, time-limited, and structured approach, has been well-established as an effective treatment for depression and anxiety, and in that regard it can be an appropriate intervention for mental disorders in GI patients. However, psychological therapies also aim to enhance coping, decrease stress reactivity, and address problematic behaviors (e.g., medication nonadherence, poor health habits), which in turn may have a positive impact on GI disease outcomes. While psychological therapies can affect GI and immune functioning, there is some question about whether they directly improve GI symptoms and disease course, depending on which GI illness is being reviewed.

There has been growing interest in the efficacy of psychological interventions for organic GI disorders, in part because of the positive results for functional GI disorders (described below). Interventions that have been evaluated for IBD using randomized controlled trials have included stress management, CBT, psychodynamic therapy, and most recently, hypnotherapy, with outcomes relating to psychological functioning and impact on IBD symptoms and course. There have been several systematic reviews, with mixed conclusions, in part related to differing decisions on inclusion criteria and study groupings. Most reviews included any psychotherapy-type intervention, rather than considering only psychological therapies with empirical support for efficacy in the general psychotherapy literature. One meta-analysis of psychological interventions in IBD (Timmer et al., 2011) concluded that there was no significant impact on emotional problems or disease activity. However, this review pooled theoretically different approaches with potentially differing efficacy, thus decreasing the likelihood of demonstrating an effect.

Overall, there is some consensus that the empirically supported treatments such as CBT and stress management (based on CBT strategies) have more consistently demonstrated improvements in HRQOL for IBD patients as well as in the targeted comorbid depression and anxiety (Knowles et al., 2013). In contrast, there have been few positive findings of impact directly on IBD outcomes, although many studies included only patients in remission, contributing to potential floor effects. One well-designed large case–control study found decreased medication use, decreased physician visits, and fewer disease relapses following psychological treatment (Wahed et al., 2010). Also treatment studies using hypnotherapy reported reduction in IBD symptom severity. The hypnotherapy results should be considered promising but preliminary, given that they were based on a small number of patients who received treatment. Methodological issues need to be more systematically addressed to better ascertain efficacy for the disease outcomes in particular.

The research on psychotherapeutic interventions for functional GI disorders has significantly expanded over the last decade. Studies have evaluated multiple therapy modalities, including brief psychodynamic therapy, hypnotherapy, relaxation training, and CBT. Psychological treatments, as a class of

interventions, have been found to be effective in reducing symptoms in IBS, compared to pooled outcomes for controls (Lackner et al., 2004). CBT has been shown to have a direct effect on IBS symptoms independent of its effects on psychological distress. A systematic review and meta-analysis of IBS outcomes and psychological therapies, which accessed data for almost 1300 adult patients in randomized controlled trials, also concluded that psychological therapies were effective for IBS (numbers needed to treat = 4; Ford et al., 2009). A similar review of pediatric studies reached a parallel conclusion, describing that psychological interventions with multiple therapy components (e.g., relaxation, psycho-education, cognitive therapy) were superior to standard care for outcomes of pain and functioning (Brent et al., 2009). While much of the research has focused on IBS, psychotherapy studies assessing outcomes for nonulcer dyspepsia have almost uniformly reported significant improvements in symptoms posttreatment, and benefits have been found to last up to 1 year (Soo et al., 2005).

All these reviews acknowledge study limitations, which qualify the strength of the conclusions, reflecting common challenges in many psychotherapy studies such as small sample size and difficulty blinding to treatment allocation. Nevertheless, treatment guidelines for IBS include psychological therapies as recommended interventions, particularly emphasizing CBT, psychodynamic interpersonal therapy, and hypnosis. New directions for psychological intervention with IBS have explored interoceptive exposure to gut sensations, mindfulness, and minimal contact treatments such as Internet delivery of CBT interventions, with promising results.

Studies that directly compared behavioral therapies with antidepressant medication for functional disorders (Drossman et al., 2003) have reported a better response to the psychological interventions and lower dropout rates, the latter suggesting a greater tolerance for the psychological versus pharmacological therapy. Psychotherapy was also found to result in significantly lower health care costs than medication or usual treatment approaches, highlighting an important cost-offset of psychotherapy.

Many, but certainly not all, patients are receptive to psychological care as part of the GI disease management plan. Inclusion of a psychological therapy can depend on availability of the treatment, and is often recommended depending on symptom severity, evident maladaptive coping, patient awareness of a stress–symptom connection, and patient readiness to engage in this type of treatment. While GI specialty clinics in many countries have adopted a biopsychosocial approach in their disease management, this approach is not yet a standard or common practice.

Conclusion

The interconnection of psychological and GI functioning is evident in both organic and functional GI disorders, supporting the applicability of a biopsychosocial model for these chronic conditions. The relationship seems to be bidirectional, with each able to affect the other, although the etiological role of psychological distress in GI disease such as IBD remains to be elucidated. The findings from both animal and human studies in this area have contributed to new understandings of biological underpinnings in psychological processes. These interrelationships have opened up further directions for treatment of GI disorders, with positive findings from psychological interventions related to improvement of patient outcomes and reduction of health care utilization. Nevertheless, more rigorous treatment studies are needed to determine patients most apt to benefit, efficient modalities for treatment delivery, and protective effects for disease course. Continued research efforts regarding the role of stress and preexisting depression, particularly in inflammatory disease, are likely to increase the understanding of the mechanisms linking the psyche and GI disease, and perhaps will facilitate more widespread integration of biopsychosocial treatment in gastroenterology.

See also: Anxiety and Anxiety Disorders; Behavioral Medicine; Chronic Illness, Psychosocial Coping with; Co-morbidity of Mental and Physical Conditions; Depression; Evidence-Based Medicine; Health Outcomes, Assessment of; Health Psychology; Interoception; Mental Disorders, Epidemiology of; Mindfulness; Personality and Health; Psychological Treatment, Effectiveness of; Psychological Treatments: Randomized Controlled Trials; Psychoneuroimmunology; Psychosomatic Medicine; Stigma, Social Psychology of; Stress, Coping and Health.

Bibliography

Bercik, P., Collins, S.M., Verdu, E.F., 2012. Microbes and the gut-brain axis. Neurogastroenterology & Motility 24, 405–413.

Bernstein, C.N., Singh, S., Graff, L.A., Walker, J.R., Miller, N., Cheang, M., 2010. A prospective population-based study of triggers of symptomatic flares in IBD. American Journal of Gastroenterology 105, 1994–2002.

Bonaz, B.L., Bernstein, C.N., 2013. Brain-gut interactions in inflammatory bowel disease. Gastroenterology 144, 36–49.

Brent, M., Lobato, D., Leleiko, N., 2009. Psychological treatments for pediatric functional gastrointestinal disorders. Journal of Pediatric Gastroenterology and Nutrition 48, 13–21.

Camara, R.J., Lukas, P.S., Begre, S., Pittet, V., Von Kanel, R., Swiss Inflammatory Bowel Disease Cohort Study, G., 2011. Effects of social support on the clinical course of Crohn's disease. Inflammatory Bowel Diseases 17, 1277–1286.

Caso, J.R., Leza, J.C., Menchen, L., 2008. The effects of physical and psychological stress on the gastro-intestinal tract: lessons from animal models. Current Molecular Medicine 8, 299–312.

Drossman, D.A., Camilleri, M., Mayer, E.A., Whitehead, W.E., 2002. AGA technical review on irritable bowel syndrome. Gastroenterology 123, 2108–2131.

Drossman, D.A., Toner, B.B., Whitehead, W.E., Diamant, N.E., Dalton, C.B., Duncan, S., Emmott, S., Proffitt, V., Akman, D., Frusciante, K., Le, T., Meyer, K., Bradshaw, B., Mikula, K., Morris, C.B., Blackman, C.J., Hu, Y., Jia, H., Li, J.Z., Koch, G.G., Bangdiwala, S.I., 2003. Cognitive-behavioral therapy versus education and desipramine versus placebo for moderate to severe functional bowel disorders. Gastroenterology 125, 19–31.

El-Serag, H.B., Olden, K., Bjorkman, D., 2002. Health-related quality of life among persons with irritable bowel syndrome: a systematic review. Alimentary Pharmacology & Therapeutics 16, 1171–1185.

Ford, A.C., Talley, N.J., Schoenfeld, P.S., Quigley, E.M., Moayyedi, P., 2009. Efficacy of antidepressants and psychological therapies in irritable bowel syndrome: systematic review and meta-analysis. Gut 58, 367–378.

Greenley, R.N., Hommel, K.A., Nebel, J., Raboin, T., Li, S.H., Simpson, P., Mackner, L., 2010. A meta-analytic review of the psychosocial adjustment of youth with inflammatory bowel disease. Journal of Pediatric Psychology 35, 857–869.

van der Have, M., van der Aalst, K.S., Kaptein, A.A., Leenders, M., Siersema, P.D., Oldenburg, B., Fidder, H.H., February 2014. Determinants of health-related quality of life in Crohn's disease: a systematic review and meta-analysis. Journal of Crohn's and Colitis 8 (2), 93–106.

Knowles, S.R., Monshat, K., Castle, D.J., 2013. The efficacy and methodological challenges of psychotherapy for adults with inflammatory bowel disease: a review. Inflammatory Bowel Diseases 19, 2704–2715.

Koloski, N.A., Jones, M., Kalantar, J., Weltman, M., Zaguirre, J., Talley, N.J., 2012. The brain–gut pathway in functional gastrointestinal disorders is bidirectional: a 12-year prospective population-based study. Gut 61, 1284–1290.

Lackner, J.M., Mesmer, C., Morley, S., Dowzer, C., Hamilton, S., 2004. Psychological treatments for irritable bowel syndrome: a systematic review and meta-analysis. Journal of Consulting and Clinical Psychology 72, 1100–1113.

Malik, S., Coulson, N.S., 2011. The therapeutic potential of the internet: exploring self-help processes in an internet forum for young people with inflammatory bowel disease. Gastroenterology Nursing 34, 439–448.

Maunder, R.G., Levenstein, S., 2008. The role of stress in the development and clinical course of inflammatory bowel disease: epidemiological evidence. Current Molecular Medicine 8, 247–252.

Mayer, E.A., Naliboff, B.D., Chang, L., Coutinho, S.V., 2001. V. Stress and irritable bowel syndrome. American Journal of Physiology – Gastrointestinal and Liver Physiology 280, G519–G524.

McCombie, A.M., Mulder, R.T., Gearry, R.B., 2013. How IBD patients cope with IBD: a systematic review. Journal of Crohn's and Colitis 7, 89–106.

Muller, K.R., Prosser, R., Bampton, P., Mountifield, R., Andrews, J.M., 2010. Female gender and surgery impair relationships, body image, and sexuality in inflammatory bowel disease: patient perceptions. Inflammatory Bowel Diseases 16, 657–663.

Pellissier, S., Dantzer, C., Canini, F., Mathieu, N., Bonaz, B., 2010. Psychological adjustment and autonomic disturbances in inflammatory bowel diseases and irritable bowel syndrome. Psychoneuroendocrinology 35, 653–662.

Rampton, D., 2009. Does stress influence inflammatory bowel disease? The clinical data. Digestive Diseases 27 (Suppl. 1), 76–79.

Rogala, L., Miller, N., Graff, L.A., Rawsthorne, P., Clara, I., Walker, J.R., Lix, L., Ediger, J.P., Mcphail, C., Bernstein, C.N., 2008. Population-based controlled study of social support, self-perceived stress, activity and work issues, and access to health care in inflammatory bowel disease. Inflammatory Bowel Diseases 14, 526–535.

Soo, S., Moayyedi, P., Deeks, J., Delaney, B., Lewis, M., Forman, D., 2005. Psychological interventions for non-ulcer dyspepsia. Cochrane Database of Systematic Reviews, CD002301.

Surdea-Blaga, T., Baban, A., Dumitrascu, D.L., 2012. Psychosocial determinants of irritable bowel syndrome. World Journal of Gastroenterology 18, 616–626.

Taft, T.H., Keefer, L., Artz, C., Bratten, J., Jones, M.P., 2011. Perceptions of illness stigma in patients with inflammatory bowel disease and irritable bowel syndrome. Quality of Life Research 20, 1391–1399.

Tang, L.Y., Nabalamba, A., Graff, L.A., Bernstein, C.N., 2008. A comparison of self-perceived health status in inflammatory bowel disease and irritable bowel syndrome patients from a Canadian national population survey. Canadian Journal of Gastroenterology 22, 475–483.

Timmer, A., Preiss, J.C., Motschall, E., Rucker, G., Jantschek, G., Moser, G., 2011. Psychological interventions for treatment of inflammatory bowel disease. Cochrane Database of Systematic Reviews, CD006913.

Van Oudenhove, L., Aziz, Q., 2013. The role of psychosocial factors and psychiatric disorders in functional dyspepsia. Nature Reviews Gastroenterology & Hepatology 10, 158–167.

Varni, J.W., Lane, M.M., Burwinkle, T.M., Fontaine, E.N., Youssef, N.N., Schwimmer, J.B., Pardee, P.E., Pohl, J.F., Easley, D.J., 2006. Health-related quality of life in pediatric patients with irritable bowel syndrome: a comparative analysis. Journal of Developmental & Behavioral Pediatrics 27, 451–458.

Wahed, M., Corser, M., Goodhand, J.R., Rampton, D.S., 2010. Does psychological counseling alter the natural history of inflammatory bowel disease? Inflammatory Bowel Diseases 16, 664–669.

Wahed, M., Rampton, D., 2013. Psychological stress and related mood disorders, and their therapeutic implications in IBD. IBD Monitor 13, 143–153.

Walker, J.R., Ediger, J.P., Graff, L.A., Greenfeld, J.M., Clara, I., Lix, L., Rawsthorne, P., Miller, N., Rogala, L., Mcphail, C.M., Bernstein, C.N., 2008. The Manitoba IBD cohort study: a population-based study of the prevalence of lifetime and 12-month anxiety and mood disorders. American Journal of Gastroenterology 103, 1989–1997.

Relevant Websites

www.ibsclinic.org.au – IBSclinic.org
www.IBDclinic.org – IBDclinic.org
http://www.ibdmanitoba.org/ – IBD Clinical and Research Centre, University of Manitoba.
http://www.romecriteria.org/ – Rome Foundation

Gated Communities

Renaud Le Goix, University Paris 1 Panthéon-Sorbonne, UMR Géographie-cités, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Gated communities are enclosed residential subdivisions, privately operated under contractual agreements between homeowners, yielding an interlocking of access control, legal arrangements, and social interactions. Built as morphological enclaves with physical enclosures and security features, they are also designed to exclude others: property rights are implemented in property owners associations, the private governance structures regulating the selection of residents based on age, lifestyle, or social preferences. Gated neighborhoods are therefore viewed as securitization strategies for the subset of the public who seek comfort in social homogeneity.

'Gated communities' (GCs) is a euphemism. Gated estates are a subset of a wider genre of enclave-style developments, better defined as privately governed and securitized neighborhoods. GCs are residential schemes organizing the governance and social structure with an interlocking of spatial, legal, and social systems.

Predominantly spatial and morphological analyses usually view GCs as simply walled and gated residential neighborhoods. Security systems, gates, and around-the-clock surveillance mean that gating blocks public access (Blakely and Snyder, 1997). GCs therefore represent a form of urbanism where public space has been privatized: they differ from condominiums and secured apartment complexes because behind the gates they include shared areas and amenities that are elsewhere open to the public and part of the urban common grounds (streets, parks, sidewalks, playgrounds, etc.). There has since been wide consensus to describe the security logic as a nonnegotiable requirement in contemporary urbanism and architecture. Many agree that both the privatization of public space and the fortification of urban realm, in response to the fear of crime, have contributed significantly to spatial and social diffusion of the phenomenon in different national contexts (Bagaeen and Uduku, 2010).

But legally speaking, all forms of residential enclosures also rely on property rights arrangements, implemented in property owners associations (POAs) and private governance structures that are designed to exclude others (common interest developments (CIDs)). These governing bodies are formed from among residents tied to a common set of interests by contract (Kennedy, 1995; McKenzie, 2006b, 1994; Owens, 1997), and there has been a variety of them used in different local contexts. GCs and walled neighborhoods are therefore a family member of a more general class that includes master-planned communities (horizontal version) and condominiums (vertical version) governed by collective tenure and incorporated organizational arrangements (Glasze, 2005). Important considerations from this perspective include the nature of ownership, governance, and management and how it interacts with local bodies of governments and communities (Kennedy, 1995; Le Goix, 2006).

That is to say that GCs rely on securitization forms that are embedding social strategies to seek 'comfort' and social homogeneity (Low, 2003; Gordon, 2004; Kirby, 2008), with a special focus on the impact of GCs on crime, segregation, property values, citizenship, and related behavior.

Mass Growth of Private Urban Governments

In almost two decades of academic debate on GCs, the different versions of GCs have been analyzed usually elaborating on the same language that describes privatized neighborhoods, but gating a neighborhood accentuates the societal impact. Their rise was initially fastest in the United States and Latin America, where the media and academic commentators were quick to describe the phenomenon in terms of security-oriented privatized urbanism. A popular critique easily followed, warning of the social fragmentation of the city, out-of-control urban segregation, secession, and the end of civic order as we know it. GCs became for some both symbols and symptoms of a line that is being crossed from voice-based citizenship to exit-based citizenship (Tiebout, 1956; Foldvary, 1994), from politically organized to market-organized civic society.

The Diffusion of Gated Communities in Different Contexts

Commentators have recorded the phenomenon across national contexts, under a diversity of denominations, all with contextual references and an emphasis on historical patterns of enclosures (Low, 2006; Bagaeen and Uduku, 2010): in Western Europe (Le Goix and Callen, 2010; Blandy, 2006; Raposo, 2006), in postcommunist Europe (Stoyanov and Frantz, 2006; Blinnikov et al., 2006; Cséfalvay, 2009), in the Arabian peninsula (Glasze and Alkhayyal, 2002), in Israël (Rosen and Razin, 2009), in China (Webster et al., 2006; Wu, 2005), in Southeast Asia and Australia (Kenna and Dunn, 2009), etc. Main motives have been identified, mainly hypothetical (Kenna and Dunn, 2009), although recent researches bring new empirical results (Cséfalvay and Webster, 2012). A first strong thesis implies that crime drives the market, linking security and fear of others – sometimes distinguished from the desire for security of person and

property. This strong linkage between the segregation of the affluent and their fear of crime has been thoroughly discussed in Latin America (Carvalho et al., 1997; Caldeira, 2000; Thuillier, 2005) and in South Africa (Landman, 2006), more than in other contexts. Gating has often been associated with both fear and a lack of confidence in the public security enforcement. A second hypothesis has focused on how residential preferences and economic rationale prevail, and GCs are therefore understood as an exit-option from the public realm, from the overregulated and overcrowded cities, inefficient in providing community service, in a 'gating growth machine (Vesselinov, 2008)'. Indeed, the main task of private urban governance is to privately organize and regulate the allocation of land and the local planning decision making (McKenzie, 1994), rather than a radical privatization of public goods as initially suggested by Blakely and Snyder (1997). As in the United States or in Latin America, contenders of the 'secession of the successful' tend to link GCs with inequalities and fear of others, and public defiance against the State. Figure 1 shows that GCs might be found in different contexts, and are as likely to be found in countries with lower level of inequalities and higher level of taxations, as in other countries. In Europe, however, an analysis of gating trends in several countries shows that driving factors (fear or crime, secession of the successful, and flight from blight) play different roles, and some other factors such as the political local attitude against centralization might also play a role. Indeed, in Europe, GCs are more common in countries with a lower economic level, highly centralized financing of local governments, higher social inequalities, and moderate redistribution of wealth (Cséfalvay and Webster, 2012). Gating in different local contexts is undoubtedly influenced in many ways by global models but it has also developed according to local political, legal, and architectural traditions, and national and local contexts prevails in analyzing the reasons why GCs are a common feature or exception, as in the cases of the Netherlands, Finland, or Germany.

Gates and Walls for Security

GCs have been described as a physical expression of post-industrial social changes (fragmentation, individualism, and rise of communities). Accordingly, they have been viewed as representing a deep penetration in society of the ideologies of fear and security – ideologies developed by dominant economic and political actors including municipalities and the homebuilding and security industries (Marcuse, 1995).

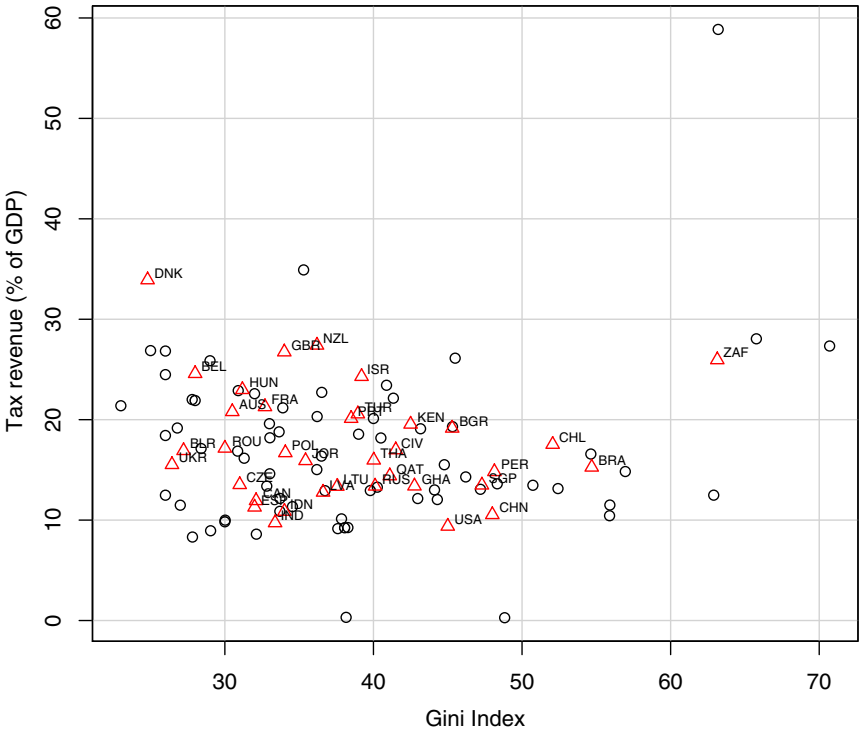

Figure 1 Countries with gated communities by Gini Index (2010*) and State Taxation Revenue (% GDP) (2010*).
Note: Triangle = countries with gated communities according to author's survey of academic literature; blank circles = countries without gated communities. AUS, Australia; BEL, Belgium; BGR, Bulgaria; BLR, Belarus; BRA, Brazil; CAN, Canada; CHL, Chile; CHN, China; CIV, Cote d'Ivoire; CZE, Czech Republic; DNK, Denmark; ESP, Spain; FRA, France; GBR, United Kingdom; GHA, Ghana; HUN, Hungary; IDN, Indonesia; IND, India; ISR, Israel; JOR, Jordan; KEN, Kenya; LTU, Lithuania; LVA, Latvia; NZL, New Zealand; PER, Peru; POL, Poland; PRT, Portugal; QAT, Qatar; ROU, Romania; RUS, Russian Federation; SGP, Singapore; THA, Thailand; TUR, Turkey; UKR, Ukraine; USA, United States; ZAF, South Africa. *Data as in 2010. For some countries, earlier reference data have been used (post. 2005). Sources: Author personal survey, 2012; World Bank, 2012; CIA, 2012.

The security argument drives a noticeable consensus among authors. Security features and guarded booths are a level of neighborhood security above the self-defensive 'armed response' placards posted on lawns, community programs of 'neighborhood watch' and community policing, and defensible space design approaches (Newman, 1972). At a more fundamental level, neighborhood gates simply extend the idea of the front door. An early theorization of gated streets as defensible spaces was developed by Newman and his Institute for Community Design Analysis. His studies have been widely publicized and incorporated in public policies through urban design guidelines aimed at preventing crime, which make an apology for gating as a device that prevents urban decay by giving social control over the environment to residents. This includes the erection of street barriers in retrofitted residential neighborhoods as a way of reintroducing public safety and controlling gang activities.

If there is a general understanding that GCs are places where the potential gain for stealing is higher than in other neighborhoods, empirical studies relating the erection of gates and crime are scarce. In fact, GCs frequently contribute to a relative decrease in burglaries and larcenies (Atlas and Leblanc, 1994), but also divert crime to other nongated communities in the close vicinity (Helsley and Strange, 1999). This diversion of crime often leads to a contagion of street barriers and security measures in neighborhoods close to GCs. In some places (Los Angeles and Mexico, for instance) cases of carjacking of GCs residents outside of the walls have also been reported.

While acknowledging that there are undoubtedly processes of market extension and cultural exportation going on, some authors have also stressed the local specificity of gated developments in specific national contexts and the importance of local emergence (Webster and Glasze, 2006). For example, Latin America has a rich history and modern culture in gating (Thuillier, 2005). Many appeared in the 1970s as a middle- and upper-income exodus to suburban areas where lower income groups had been settling since the 1940s. In those circumstances, gates were used to maintain a physical separation between classes when physical distances between rich and poor are short (Carvalho et al., 1997). Exploring the dual idea of spread and emergence in Lebanon, Glasze notes the similarity between some gated compounds and American developments, and also describes the links between US developments and Lebanese investors (Glasze and Alkhayyal, 2002) during the insecurity of the civil war and was for many a necessity in the face of public government not being able to provide personal and property security. The domination of the gated genre in South Africa clearly has very specific local causes. It has been linked to postapartheid postures, unilaterally by residents (Durington, 2006), by developers, or by more complex coalitions of local actors (developers and municipalities). The spread of enclosed enclaves (street closures and walls) is not only due to developers but also due to a strategy coordinated by local and national authorities and by very proactive residents associations. This has given rise to the extraordinary phenomenon of mass road closures, in which streets belonging to public governments are unilaterally enclosed by a boom gate and guard house and the enclave permitted to incorporate as an owners association ('Section 21 Company') (Landman, 2006).

Landscapes of Walls: An Historical Perspective

Gated cities have a long history. Socially and physically defined urban enclaves are as old as cities themselves. Private urban governance emerged in nineteenth-century industrial European cities such as London and Paris, where the new industrial bourgeoisie sought in privately operated and enclosed suburban neighborhoods, a quiet retreat from the busy city center. Le Parc de Montretout, in Saint-Cloud, France, developed in 1832, was probably the first of its kind (Le Goix and Callen, 2010). In the United States, the spread of GCs has roots in that country's long-standing ideology of suburban development, exclusionary zoning, and redlining (Ihlanfeldt, 2004; Kato, 2006). One early thread of influence is the romantic suburban utopias and utopian-influenced projects. Haskell's Llewellyn Park was probably the first modern GC built in the United States. It has continuously operated a gatehouse and a private police force since 1854 and introduced private governance of shared amenities based on deed restrictive covenants that protected the stability and homogeneity of the neighborhood. A second thread links America's modern GCs to the historical processes that brought CIDs and exclusionary restrictive covenants laws from Europe to the United States. McKenzie (1994) explores the long European history of restrictive covenants and residential associations (observable since 1743 in London). The first homeowners association (HOA) per se was created in the United States in 1844 in Boston's Louisburg Square. Llewellyn Park and Roland Park (1891) were the first large privately owned and operated luxury subdivisions, yielding exclusive neighborhoods. They established consumer and real estate developer expectations and legal and organizational approaches that helped shape contemporary private urban governance in the United States. In the first half of the twentieth century, this kind of high-end subdivision became quite common (Mission Hills, Missouri, in 1914; Kansas City Country Club District in 1930s; and Radburn in 1928). Along with landscaping and architectural requirements, the idea of social preferences as a commoditized attribute has become common in CIDs. Exclusive lifestyle developments became common by the turn of the 1960–70s, designed as mass-consumption real estate developments, financed by large corporations attracted by potential profits and backed by the Government through the Department of Housing and Urban Development.

Historical investigation of gated enclaves has produced a rich international comparative analysis that reveals the importance of local focus on the condition of appearance of gated developments. In China, GCs are nothing new, but simply a continuation of enclosed work-unit territories and the walled low-rise courtyard houses and residential neighborhoods found traditionally in villages and urban neighborhoods (Wu, 2005). In fact, China's strong control over land and housing has enabled municipal governments to engineer a faster rate of retreat to private (collective) territorial

government than has been possible in any other country. This is partly because of the strong legacy of collective organization left by the central-planning era.

Living by Contract

The economic theory of clubs (Buchanan, 1965) challenges the neoclassical notion of public goods. A club, in economics, is a method of supplying jointly consumed goods efficiently on the basis of controlled membership and fee. GCs can therefore be understood to be territorial clubs, the sole purpose of which is to supply club goods (Webster, 2002). This line of argument sees GCs as neither public nor private spaces. They are club spaces – spaces governed by 'small publics.' Compared to general purpose municipal governments, they are alleged to be more efficient organizations for allocating scarce resources for a number of reasons. They provide greater choice for one thing, and residents can vote with their feet to choose the bundle of shared goods and services they prefer at a given price. This leads to a comparison between the club community market and the imaginary spatial economy of geographical economist Charles Tiebout (1956). This is more efficient than public government in the sense that there are fewer unrealized gains from trade in the urban land market. In other words, more people have their specific preferences (for civic goods and services and for type of neighbor and neighborhood) met (Tiebout, 1956). The exclusionary mechanism of 'membership' means that there is a more precise relationship between payment and benefit received than there is in publicly managed neighborhoods. Significant investments made by POAs are capitalized in land rent. Not only do owners have an incentive to invest in their local environments, but also they can recover investment costs from rents (Foldvary, 1994). The sustainability is ensured by the existence of the HOA, which has three basic characteristics. Elected boards act as neighborhood decision makers, contracts govern resident behavior (through covenants, conditions, and restrictions), and monthly fees finance local amenities and services. By interpreting GCs as mechanisms for supplying environments, security, and other goods and services that are neither private nor public goods, the club economy model allows a more nuanced and specific debate about the roles of markets and governments in shaping, governing, and managing the city.

Regardless of local traditions and national legal contexts, there are different organizational types of private residential neighborhoods, differentiated by the way property rights are assigned, over-shared spaces, facilities, and exclusively used housing units: condominiums, stock cooperatives, corporations, and HOAs (Glasze, 2005). In *condominiums*, in addition to individual property of the housing units, the owners hold title to an undivided interest in the common property of streets, open spaces, amenities, and so on. The shared interest assets are governed by a POA run by an elected board of directors. In *stock cooperatives*, individuals purchase a share in an entire residential complex, but the co-op owns both the exclusively owned housing units and the subsequent shared amenities. The share gives the member voting rights at the co-op assembly. In *corporations*, the common spaces and facilities are the property of an incorporated body set up specifically for that purpose – the corporation or the HOA. A covenant is typically attached to the deed, making the owner a shareholder.

As of today, there has not been any empirical study making it possible to analyze the quantitative diffusion of gated developments in countries and cities. Some insights have been available in the United States, where by the year 2000 over 15% of the housing stock was in CIDs – and the number of units in these privately governed residential schemes rose from 701 000 in 1970 to 16.3 million in 1998 (McKenzie, 2006a). The Community Association of America estimated in 2002 that 47 million Americans were living in 231 000 community associations and that 50% of all new homes in major cities belonged to community associations (Sanchez and Lang, 2005). Only a proportion – varying from 12% to 30% in the region of Los Angeles (Le Goix, 2005) – of these private local government areas are also gated. Roughly 200 gated subdivisions in France and more than a 100 in the United Kingdom have been reported (Cséfalvay and Webster, 2012). But their spatial diffusion is also very much related to metropolitanization processes, and they are more commonly found in larger metropolitan areas, and more specifically in areas of faster growth within metropolitan areas, nearby innovative industrial clusters and newly developed urban edges. GCs are not specifically linked to wealthier categories of owners. In the United States and Europe at least, their diffusion to the middle class and upper middle class has been a turning point of the last 20 years, with products targeting specific public, such as Hispanics or Asian in US suburban areas, or niche markets based on lifestyle preferences (not only golf clubs, but also GCs with a private shooting range, or a private runway for general aviation home and plane owners).

The reshaping of metropolitan areas into cities of clubs can raise the total amount of revenue going into the management of the urban environment since people are clearly willing to submit to a double taxation regime (municipal government taxes plus GC fees). The flip side of this is that local governments can save money by off-loading responsibilities to HOAs. Seen in this light, the search for security is only a very partial explanation of the spread of GCs. Gated neighborhoods have given prospective buyers greater choice, lower risk, lower search costs, and greater investment security. They have also enabled public authorities to manage growth with greater fiscal sustainability. In fact, a strong hypothesis is that gating is highly correlated with the fiscal gaps increasingly experienced by local government bodies worldwide, and some empirical findings converge in Europe (Cséfalvay and Webster, 2012) or in the United States (Le Goix, 2006). Indeed, the diffusion of GCs in a metropolitan area is often related to suburban growth, in many countries an endemic antifiscal posture leading to municipal fragmentation and a Tibeoutian world of small competitive local governments. Urban sprawl in the United States and in many developing countries continues to generate demand for new infrastructure but political, legal, and technological limits on the amount of property tax that can be collected lead to undersupply. As a result, GCs, which bring wealthy taxpayers at minimal cost, have become the perfect cash cow for local municipalities.

Social Homogeneity, Property Prices, and Segregation

GCs are territories of exclusiveness, building up by design social homogeneity on security, snob values, fear of crimes, and symbolic and physical distance from others. But all these attributes are not truly independent, as they derive from the contractual agreement binding all property owners. Questions raised about their alleged effects usually address their efficiency on preservation of the tidiness and value of the neighborhood, and ultimately on segregation patterns.

Generally speaking, ownership in proprietary neighborhoods also encourages a kind of speculation around real estate prices. But gating such neighborhoods reinforces the proactive private governance effort toward property values preservation. The liberal hypothesis assumes that operating costs of private governance are paid for by the increase in property values.

First, the quasi-governmental regime previously discussed has a preeminent role in shielding property values. The short-term apparent cost-benefits market efficiency in providing collective services must be matched up to the risks of long-term spillover effects, inefficiency of the decision-making process, residents' lack of involvement, and the risks of obsolescence and inflating maintenance costs undermining the tidiness and reputation of a neighborhood and ultimately its property values. Many court cases and legal restrictions apply only to GCs and make a special case of their governmental regime, which cannot be extended to nongated private neighborhoods (Brower, 1992; Kennedy, 1995). At last, as public dedication cannot be obviously applied to gated streets, GCs need to live up to their promise and to be founded on a financial model that takes account of the rising costs due to the obsolescence of infrastructure and amenities managed behind the gates by the POAs. Gating a CID ultimately stresses the private realm, thus reinforces the selection of residents. This effort toward social control and homogeneity contributes to the overall effort of shielding property values and creating a price premium.

There are multiple and concurring evidences, based on diverse methodologies, of the price premium of GCs over nongated private neighborhood. On average, GCs are known to generate a price premium, and to better guaranty the homogeneity of property values within the neighborhood and to better protect values on the long run than other nongated private neighborhoods in the United States (Bible and Hsieh, 2001; Le Goix and Vesselinov, 2012). But the same studies also suggest that failure of POAs occurs when costs are rising above a sustainable level compared to rapidly decreasing property values. In principle, this is addressed by government legislation requiring adequate reinvestment funds, but there will undoubtedly be many failures in many countries before legislation has its effect. Some legislative adjustments of this sort already occurred in France in the 1920s, after numerous blue-collar and often-fenced private streets experienced rapid obsolescence and lack of maintenance detrimental to property values.

As mentioned earlier, private governance leads to an implicit selection of the owners. The effect of GCs on social homogeneity (Le Goix, 2005; Vesselinov et al., 2007; Wu, 2005) has been well established. Social homogeneity is achieved through design guidelines, age restrictions, or a selective club membership, and yields a measurable effect on local segregation: in the United States, GCs tend to segregate more by age (life cycle and age polarization) and by socioethnic status (White vs. Hispanics, correlated with wealth and age), and do not locally influence segregation patterns in terms of racial segregation (Table 1). On the other side of the spectrum, the integrative effects of GCs have also been emphasized, but it is all a matter of scale. A high-income GC has been built next to a long-established poor squatter community on the outskirts of Santiago, Chile, and lower income residents welcomed their new neighbors as sources of employment and utilities (water

Table 1 Social distance indices in southwestern metropolitan areas in the United States in 2010

	Principal components of socioeconomic segregation			
Metropolitan areas	*White vs. Hispanic status, correlated with wealth and age*	*Life cycle and ownership status*	*Life cycle and age polarization*	*Racial segregation*
Las Vegas, NV	1.3	1.4	1.6	0.8
Los Angeles – Long Beach, CA	1.4	1.4	3.1	1.5
Oakland, CA	1.3	1.1	3.6	0.8
Orange County, CA	1.2	1.5	3.0	1.2
Phoenix, AZ	1.6	1.0	3.0	1.5
Riverside – San Bernardino, CA	2.4	1.3	2.8	1.3
San Diego, CA	1.4	1.4	3.3	0.7
San Francisco, CA	1.7	1.8	1.5	4.1
San Jose, CA	1.0	1.0	0.9	1.3
Santa Cruz – Watsonville, CA	1.5	1.6	1.6	1.5
Ventura, CA	1.8	2.2	7.7	1.3

Note: Values indicated measure a normalized social distance (Social Distance Index (SDI)) between a gated community area and adjacent nongated neighborhoods, based on US Census geography (block groups) and factor analysis of social, economic, and racial variables.
For instance, on factor 1, the SDI between gated communities and abutting neighborhoods is 2.4 higher than between two adjacent nongated neighborhoods.
Methodology: Le Goix, R., 2005. Gated communities: sprawl and social segregation in Southern California. Housing Studies 20, 323–343.
Le Goix, R., Vesselinov, E., 2012. Inequality Shaping Processes and Gated Communities in US Western Metropolitan Areas, in: American Association of Geographers Annual Meeting - AAGs 2012. New York, NY (February 2012).

and sewerage) (Salcedo and Torres, 2004). The interesting question posed when new social geographies are established through GC-based gentrification – and the challenge to established understanding of segregation – is how well GCs and surrounding neighborhoods, especially of lower income, become functionally integrated and how this changes intergroup attitudes and wider complex systemic effects.

See also: New Urbanism; Suburbanization and Suburbanism.

Bibliography

Atlas, R., Leblanc, W.G., 1994. The impact on crime of streets closures and barricades: a Florida case study. Security Journal 5, 140–145.

Bagaeen, S., Uduku, O. (Eds.), 2010. Gated Communities: Social Sustainability in Contemporary and Historical Gated Developments. Earthscan, London.

Bible, D.S., Hsieh, C., 2001. Gated communities and residential property values. Appraisal Journal 69, 140–145.

Blakely, E.J., Snyder, M.G., 1997. Fortress America, Gated Communities in the United States. Brookings Institution Press & Lincoln Institute of Land Policy, Washington DC, Cambridge, MA.

Blandy, S., 2006. Gated communities in England: historical perspectives and current developments. GeoJournal 66, 15–26.

Blinnikov, M., Shanin, A., Sobolev, N., Volkova, L., 2006. Gated communities of the Moscow green belt: newly segregated landscapes and the suburban Russian environment. GeoJournal 66, 65–81.

Brower, T., 1992. Communities within the community: consent, constitutionalism, and other failures of legal theory in residential associations. Land Use and Environmental Law Journal 7, 203–273.

Buchanan, J.M., 1965. An economic theory of clubs. Economica 32, 1–14.

Caldeira, T.P.R., 2000. City of Walls: Crime, Segregation, and Citizenship in Sao Paulo. University of California Press, Berkeley, CA.

Carvalho, M., Varkki George, R., Anthony, K.H., 1997. Residential satisfaction in condominios exclusivos in Brazil. Environment and Behavior 29, 734–768.

Cséfalvay, Z., 2009. Demythologising gated communities in Budapest. In: Smigiel, C. (Ed.), Gated and Guarded Housing in Eastern Europe, Forum Ifl. Selbstverlag des Leibniz-Instituts für Länderkunde, Leipzig, pp. 35–47.

Cséfalvay, Z., Webster, C., 2012. Gates or no gates? A cross-European enquiry into the driving forces behind gated communities. Regional Studies 46, 293–308.

Durington, M., 2006. Race, space and place in suburban Durban: an ethnographic assessment of gated community environments and residents. GeoJournal 66, 147–160.

Foldvary, F., 1994. Public Goods and Private Communities: The Market Provision of Social Services. Edward Elgar, Aldershot.

Glasze, G., 2005. Some reflections on the economic and political organisation of private neighbourhoods. Housing Studies 20, 221–233.

Glasze, G., Alkhayyal, A., 2002. Gated housing estates in the Arab world: case studies in Lebanon and Riyadh, Saudi Arabia. Environment and Planning B: Planning and Design 29, 321–336.

Gordon, T.M., 2004. Moving up by moving out? Planned developments and residential segregation in California. Urban Studies 41, 441–461.

Helsley, R.W., Strange, W.C., 1999. Gated communities and the economic geography of crime. Journal of Urban Economics, 80–105.

Ihlanfeldt, K.R., 2004. Exclusionary land-use regulations within suburban communities: a review of the evidence and policy prescriptions. Urban Studies 41, 261–283.

Kato, Y., 2006. Planning and social diversity: residential segregation in American new towns. Urban Studies (Routledge) 43, 2285–2299.

Kenna, T.E., Dunn, K.M., 2009. The virtuous discourses of private communities. Geography Compass 3, 797–816.

Kennedy, D.J., 1995. Residential associations as state actors: regulating the impact of gated communities on nonmembers. Yale Law Journal 105, 761–793.

Kirby, A., 2008. The production of private space and its implications for urban social relations. Political Geography 27, 74–95.

Landman, K., 2006. Privatising public space in post-apartheid South African cities through neighbourhood enclosures. GeoJournal 66, 133–146.

Le Goix, R., 2005. Gated communities: sprawl and social segregation in Southern California. Housing Studies 20, 323–343.

Le Goix, R., 2006. Gated communities as predators of public resources: the outcomes of fading boundaries between private management and public authorities in Southern California. In: Glasze, G., Webster, C.J., Frantz, K. (Eds.), Private Cities: Global and Local Perspectives. Routledge, London, pp. 76–91.

Le Goix, R., Callen, D., 2010. Production and social sustainability of private enclaves in suburban landscapes. Local contexts and path dependency in French and US long-term emergence of gated communities and private streets. In: Bagaeen, S., Uduku, O. (Eds.), Gated Communities: Social Sustainability in Contemporary and Historical Gated Developments. Earthscan, London, UK, pp. 93–114 (Chapter 6).

Le Goix, R., Vesselinov, E., 2012. Gated communities and house prices: suburban change in Southern California, 1980–2008. International Journal of Urban and Regional Research. http://dx.doi.org/10.1111/j.1468-2427.2012.01139.x.

Low, S., 2003. Behind the Gates: Life, Security, and the Pursuit of Happiness in Fortress America. Routledge, New York.

Low, S., 2006. Towards a theory of urban fragmentation: a cross-cultural analysis of fear, privatization, and the state. Cybergeo.

Marcuse, P., 1995. Not chaos but walls: postmodernism and the partitioned city. In: Watson, S., Gibson, K. (Eds.), Postmodern Cities and Spaces. Basil Blackwell, Oxford, UK, pp. 243–253.

McKenzie, E., 1994. Privatopia: Homeowner Associations and the Rise of Residential Private Government. Yale University Press, New Haven, CO; London.

McKenzie, E., 2006a. The dynamics of privatopia: private residential governance in the USA. In: Glasze, G., Webster, C.J., Frantz, K. (Eds.), Private Cities: Local and Global Perspectives. Routledge, London, pp. 9–30.

McKenzie, E., 2006b. Emerging trends in state regulation of private communities in the US. GeoJournal 66, 89–102.

Newman, O., 1972. Defensible Space: Crime Prevention through Urban Design. MacMillan, New York.

Owens, J.B., 1997. Westec story: gated communities and the Fourth Amendment. American Criminal Law Review 34, 1127–1160.

Raposo, R., 2006. Gated communities, commodification and aestheticization: the case of the Lisbon metropolitan area. GeoJournal 66, 43–56.

Rosen, G., Razin, E., 2009. The rise of gated communities in Israel: reflections on changing urban governance in a neo-liberal era. Urban Studies 46, 1702–1722.

Salcedo, R., Torres, A., 2004. Gated communities in Santiago: wall or frontier? International Journal of Urban and Regional Research 28, 27–44.

Sanchez, T., Lang, R.E., 2005. Security vs. Status? A first look at the census' gated community data. Journal of Planning Education and Research 24, 281–291.

Stoyanov, P., Frantz, K., 2006. Gated communities in Bulgaria: interpreting a new trend in post-communist urban development. GeoJournal 66, 57–63.

Thuillier, G., 2005. Gated communities in the metropolitan area of Buenos Aires, Argentina: a challenge for town planning. Housing Studies 20, 255–271.

Tiebout, C.M., 1956. A pure theory of local expenditures. Journal of Political Economy 64, 416–424.

Vesselinov, E., 2008. Members Only: Gated Communities and Residential Segregation in the Metropolitan United States. Sociological Forum 23 (3), 536–555. http://dx.doi.org/10.1111/j.1573-7861.2008.00075.x.

Vesselinov, E., Cazessus, M., Falk, W., 2007. Gated communities and spatial inequality. Journal of Urban Affairs 29, 109–127.

Webster, C., Glasze, G., 2006. Dynamic urban order and the rise of residential clubs. In: Glasze, G., Webster, C., Frantz, K. (Eds.), Private Cities: Local and Global Perspectives. Routledge, London, pp. 222–236.

Webster, C.J., 2002. Property rights and the public realm: gates, green belts, and Gemeinschaft. Environment and Planning B: Planning and Design 29, 397–412.

Webster, C.J., Wu, F., Zhao, Y., 2006. China's modern walled cities. In: Glasze, G., Webster, C.J., Frantz, K. (Eds.), Private Cities: Local and Global Perspectives. Routledge, London, pp. 153–169.

Wu, F., 2005. Rediscovering the 'gate' under market transition: from work-unit compounds to commodity housing enclaves. Housing Studies 20, 235–254.

Gatekeepers in Social Science

Barbara Hoenig, Innsbruck University, Innsbruck, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A gatekeeper is an individual or collective actor who is in a position to control access to resources and rewards relevant in a particular social system. The term's usage in the sociology of science goes back to Robert K. Merton, who defines gatekeeping as one of the four complementary roles in the role set of scientists and scholars. Gate-keeping activities are relevant in different arenas of scholarship such as between scientific masters and apprentices, in collective mentoring programs, in publishing and in research funding.

A gatekeeper is an individual or collective actor who is in a position to control access to resources and regulate the allocation of rewards relevant in a particular social system. These resources might take the form of money or information, of reputation or social capital. In bureaucratic institutions, both formal and informal authorities can exercise a gate-keeping function. In the social institution of science, the normative role of a gatekeeper typically is twofold: "providing or denying access to opportunities" (Merton, 1973: p. 522), which includes giving advice, encouragement, and support on the one hand, and exercising control, regulation, and sanctions on the other. The concept of the gatekeeper has been introduced by Lewin (1943), who has investigated the food habits of five groups of Americans by tracing back various channels of buying and gardening, baking and canning food. Subsequent inquiries of gatekeeping have been most influential in communication research and in science and technology studies. In a panel study on voter decisions during a US presidential election campaign in Erie County, Ohio, Lazarsfeld et al. (1944) have investigated the hypothesis that mass media's influence on potential voters typically takes on a two-step flow of communication, insofar "ideas often flow from radio and print to opinion leaders and from them to the less active sections of the population" (Lazarsfeld et al., 1944: p. 151). In network analysis, recent research focuses on gatekeeping as a 'structural power condition' in order to discover structures that favor gatekeeping (for instance, Corra and Willer, 2002). In science, the gate-keeping role is important for institutionalized procedures of communication and decision making. Moreover, it contributes to the maintenance of scientific standards, evaluation, and quality assurance. Its institutionally central role has also stimulated research scrutinizing possible particularistic biases in decision making within a generally universalistic peer-review system (Zuckerman and Merton, 1971). Support and gate-keeping activities are relevant in different arenas of scholarship, as in the social ties between scientific masters and apprentices and in collective mentoring programs by affirmative action. It also bears upon the work done by journal editors in the publishing industry, by reviewers and foundation officers in research funding, and by departmental chairs and boards in career promotion of scientists and scholars. Apart from individual actors, institutions such as peer-review panels, funding organizations, scientific journals, and career track systems can take on a gate-keeping function in science and scholarship.

The Gatekeeper's Role in Science

The term's usage in the sociology of science goes back to its founding father, Robert K. Merton, who defines gatekeeping as one of the four complementary roles in the role set of scientists and scholars: Apart from teaching, research, and scientific administration, they are expected to fulfill a gate-keeping function in the institutionalized peer-review system (Zuckerman and Merton, 1972). Merton characterizes gate-keeping as "basic to the system of evaluation and the allocation of roles and resources in science" (Merton, 1973: p. 521), where gatekeepers mostly operate through panels of peers. Becoming a member in these panels is part of the scientific socialization, while gate-keeping activities "help shape the permanent record of scientific work" within "patterned sequences of role configurations" (Merton, 1973: p. 521). According to Merton, the gate-keeping role is clearly structured toward an older age bias, because scientists and scholars usually act as gatekeepers in later phases of their career, when their reputation as researcher in the field of question is already established or assumed. In dynamic aspects, the role of the gatekeeper effects the 'sponsored mobility' of scientists and scholars (Turner, 1960; Hargens and Hagstrom, 1967): In contrast to a mobility pattern similar to a sporting contest, where individual achievements are solely judged according to universal criteria, sponsored mobility involves a "controlled selection process" where members of the elite "*call* those individuals who have the appropriate qualities" (Turner, 1960: p. 857). The fact that, irrespective of their actual achievements, particular researchers and scientific institutions are structurally more successful than others in the cumulative process of accumulating reputation and reward is known as the 'Matthew effect' (Merton, 1968). Structurally seen, the gate-keeping role is crucial for understanding the operation of intellectual authority and the reproduction of social stratification and scientific elites (Cole and Cole, 1973; Zuckerman, 1977; Whitley et al., 2010). Gatekeepers control and regulate legitimate access to institutions as those intellectual authorities that define what counts as 'scientific excellence' or not. In this sense, gatekeeping has far reaching consequences on whether a particular person is allowed to continue research, be it on a particular topic or doing research at all. In the following, we take a look on different arenas of science as a social institution, where gate-keeping activities are involved: in the social ties of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03011-7

scientific masters and apprentices, in collective mentoring by affirmative action programs, in the publishing and research funding industry, and in sociocognitive networks or 'invisible colleges' in the scientific community.

Masters and Apprentices in Science

One of the most important, albeit highly personalized social ties in science, are those between the scientific master on the one hand and the apprentice on the other. Alternative terms often used to characterize this role relation are teacher–pupil, mentor–student, or senior–junior collaborators. In the process of scientific socialization-by-filiation, the master acts as teacher and supervisor, supporter and mentor, and as an individual role model for the apprentice. As Zuckerman and Merton (1972) have observed, complementary age roles of teachers and students that usually are members of different generations and cohorts often become integrated in research collaborations; the relation is also characterized by a remarkable amount of conflict and sociological ambivalence for both (Zuckerman and Merton, 1972). The mentor's relevance for the career prospect of any young scientist or scholar has also been an object of social research. For instance, Nobel laureates in interviews have frequently emphasized the important role of a reputed mentor for making a career within the scientific ultraelite (Zuckerman, 1977). In her study on Nobel laureates in the United States, Zuckerman has characterized these enduring, consequential social ties between scientific masters and apprentices as much more important than ties by kinship. More than half of the laureates in her sample have been working as graduate students, postdoctorates, or junior collaborators under older laureates or similar members of the ultraelite without Nobel prize (Zuckerman, 1977). Zuckerman has found a self-perpetuating pattern of the scientific ultraelite, where Nobel laureates, by acquiring experience and judgment and by acting as "skillful advocates of their candidates" (Zuckerman, 1977: p. 106), produced their own 'heirs to the prize' in a lineage sometimes stretching over half a century. As fundamental difference between biological and social heredity Zuckerman emphasizes the aspect of social choice in that relationship: At least to some extent, both masters and apprentices can mutually choose with whom they want to work. Moreover, it can be assumed that scientific masters possibly tend to prefer apprentices that are perceived to be similar to themselves at a younger age. Mostly informal, personalized patterns of recruitment and support, in the aggregate, then might favor structural majorities at the expense of 'minority' members in science.

Collective Mentoring by Affirmative Action

In the social stratification of science, the situation of women and other 'minorities' has always been rather peculiar. Irrespective of their actual individual achievements, they are more frequently working in organizational contexts disadvantageously affecting the cumulative allocation of reward during their scientific career. This bears upon appointments to institutions with low prestige, lacking eminent mentors, and access to scientific networks, publishing later and less than others, more seldom being held eligible for research funds and fellowships (Keith et al., 2002). This is one reason why members of these groups are much scarcer as gatekeepers of science as well and structurally encounter fewer opportunities to shape research policy issues. Respecting the normative ethos of universalism (Merton, 1942), which regards the validity of scientific knowledge claims as independent of irrelevant personal characteristics of scientists, often leads to support institutional measures for equal opportunities in science. In this sense, affirmative action for disadvantaged groups can be understood as a critique on a still lacking universalism in access to science.

Affirmative action designates institutional measures of support, for instance, by quotas, targets, and mentorship, in order to counteract individuals' exclusion from opportunities based on group membership such as by race, class, and gender. These institutionalized forms of support aim at integrating members of these groups in academic networks by mediating knowledge about scholarly practices and by generating opportunities for experiencing career support and collegiality. Moreover, by affirmative action it is expected to provide more equal access for 'minority' members to shaping research policy at all relevant levels. Within the last decades, there has been done much research on affirmative action in the educational and employment system (Harper and Reskin, 2005). However, experience shows that it very much depends on programs' actual implementation practices whether their outcomes really benefit 'minority' members or not (Aguirre, 2000). Both social disparities and institutional programs to combat discrimination in the academe significantly vary between countries; particularly Anglo-Saxon and European Scandinavian countries are perceived as more advanced in that regard. While in the United States, affirmative action in higher education has responded to the Civil Rights movement from the 1960s onward, European countries only recently have begun to learn from these experiences (Bacchi, 1996; Appelt and Jarosch, 2000). In 1997, the European Union has adopted a legal guideline for actively pursuing anti-discriminiation strategies as part of the 'Treaty of Amsterdam,' such as by implementing gender mainstreaming and strategies combating racism. Except for the United Kingdom (HESA, 2012), however, most European countries' higher education statistics do not provide information with respect to known ethnicity-status, which indicates that 'minority' issues still are not much on the science policy agenda there.

Gatekeeping in the Publishing Industry

The gatekeeper's role in science has mostly been investigated in fields of publishing, such as by research on journal or book editors or reviewers evaluating manuscripts submitted for publication. In science and technology studies, Price (1963) has coined the term 'invisible colleges' for informal groups of scientists and scholars that, by communicating, reading, and citing each other's work, contribute to scientific growth. While Price has estimated the size of such a scientific network as "a few hundred colleagues for every worker" (Price, 1963: p. 72), he has not explicitly addressed possible disciplinary differences in that

regard. Crane (1972) has built upon Price in her research on the role of invisible colleges in the diffusion of scientific knowledge and ideas. By bibliometrics and survey research, she investigated the social organization of informal collaboration networks among mathematicians and rural sociologists. In addition, she provided a dynamic model of scientific growth, both of scientific communities and their knowledge production. In an earlier study of comparing 50 journals from two disciplines, sociology and economics, Crane (1967) has scrutinized factors of the academic stratification system that highly affect editorial decisions in the selection of journal articles. She has found empirical evidence that an author's academic affiliation, doctoral origin, and professional age happened to be rather similar to the distribution of those characteristics among journal editors. Moreover, she suggested that these effects might be specific for each discipline. The publishing industry in an American context has been described by Coser, Kadushin, and Powell (Coser, 1965, 1975; Coser et al., 1982) as by three structural characteristics: its decentralization; its operating in a highly uncertain, unpredictable market; and its internal organization characterized by a predominance of craft over bureaucracy. According to the authors, decisions of publishers and editors are based on multiple factors such as estimations of prospective sales, but also on the tradition of a particular publishing house, the size of a firm, editors' previous education, career aspirations, and self-images. Moreover, they indicated that these points of reference at least partially depend on an editor's structural characteristics, to a less extent they are chosen at will. Since the turnover rate among editors is rather high, structurally induced uncertainty and ambiguity of that role are even enforced so that "the complicated network of particularistic relations between authors and editors runs counter to the universalistic contractual relations between the firm and its authors" (Coser, 1975: p. 19). As Price and others have shown, the scholarly communication between scientists through scientific papers goes back to the seventeenth century. However, despite their central role for scientific communication, social research on journals' editorial boards is still rather scarce. Recently, the role of gatekeepers in scholarly communication has been measured as interlocking editorial board membership, taking the number of editorial board members that two journals share as an indicator for the journals' proximity (Ni et al., 2013). For 58 journals from the disciplinary field of Library and Information Science, the authors could show that about 10% of editorial board members served on more than one journal. Their findings demonstrated a rather high concentration for this variable when compared with journal proximity in term of producers, scientific papers, and papers' topics (Ni et al., 2013). The approach of editorial board member coupling can generate interesting insights in networks of journal clusters within a given field or research area and in the structure of editorial gatekeeping as well.

Gatekeeping in Research Funding

Since the norm of universalism requires that the evaluation of scientific achievements is independent of individual attributes of any scientist (Merton, 1942), it is reasonable to ask whether the actual evaluation practice of peer reviewers reflects this norm. However, researchers studying the role of peer review in research funding have frequently emphasized that it is extremely hard to find any statistically significant evidence for particularistic effects. In addition, there might be no way of objectively evaluating new scientific work neither. Cole (1992) reported several empirical studies on peer review in the US National Science Foundation where the authors could not find statistically significant correlations of grant approval rates neither with individual characteristics of researchers nor with their institutional affiliations. Cole hinted to the fact that these findings not only contrast much theorizing in the field, but also those of qualitative studies and academic insiders' everyday-knowledge of what affects working at universities. Rather, he assumed that informal network ties can produce significant outcomes interpretable as Matthew effect, and he suggested to investigate them by applying social network analysis or qualitative research techniques. Lamont's (2009) qualitative study on scientific decision making has shown that external peer reviewers of funding programs in the United States practically apply some rules of thumb when talking about what constitutes a proposal's potential scientific excellence. Moreover, she unearthed significant variation in these 'customary rules of deliberation' (Lamont, 2009) across disciplinary cultures and corresponding epistemological styles in the social sciences and humanities.

Not only scientific actors, but also funding institutions and their representatives administering the granting process play an important gate-keeping role. Decisions consequential for grant approvals can be taken by individual philantrophs, science policy advisors, members of the scientific administrative elite, or by the program officer appointed to a research foundation. Although not acting as scientist, these actors often received academic training, which also equipped them with an insider-knowledge of the research field and with access to scientific networks. In the history of social scientific research funding, such as by the Rockefeller or by the Ford Foundation, professional foundation officers were often trained social scientists using their institutional position to promote social scientific research (Platt, 1996: pp. 164ff.). Therefore, a sharp distinction between foundation officers and researchers might be misleading, both because of the officers' scientific education and the multifold trajectory of their professional careers, occasionally moving back and forth between the administrative, the advisory, and the research sector (Platt, 1996: pp. 164ff.). At meso- and macrolevel, research foundations, state ministries, administrative elites, and employers' organizations can be conceived as gate-keeping agencies shaping the research policy of a country, for instance, by setting up academic evaluation systems or funding regulations operative in a national research landscape (Whitley et al., 2010). As Whitley and others have shown, different types of change in governance and state steering might then affect the reorganization of authority relations and innovation in public science systems as well.

Historical and Cultural Variation of the Gatekeeper's Role

Probably it is most hard to investigate the role of gatekeepers in career track decisions and scientific promotion, for instance

in the decision whether a scientist shall get tenure or not or whether a scholar shall be appointed for a prestigious professorship. Moreover, the gate-keeping role apparently is subject to historical change and cultural variation. For instance, concerning the historical chains between masters and apprentices in sociology, Merton and Riley (1980) have edited a small book in which eminent sociologists report about their teachers and their relation to them. When, as readers of today, we contrast their experiences with current conditions of doing research, we might recognize that paternalistic hierarchies of highly personalized, gendered, age-structured, and elitist roles at least partially have changed toward more entrepreneurial forms of science and scholarship. Structural transformations of public science have given rise to new actors and institutional forms of knowledge production, such as by industrial enterprises or science policy advisors. By extending beyond traditional boundaries of academic scholarship, the future role of gatekeepers might become more diverse and diffuse, possibly enforcing structural contradictions and ambivalences for scientists and scholars depending on these contexts of utilization. The cultural variance of the scientific role in different social contexts has been investigated by Ben-David (1971) who, inspired by Weber's and Merton's historical studies, observed successive shifts of scientific centers from England to France, then to Germany, and finally to the United States. While Ben-David has characterized the specific strength of French universities in their strongly centralized, elite-oriented system, this has been partially replaced by the Humboldt tradition of autonomous, hierarchical universities in Germany. Particularly after World War II, a steadily increasing number of independent universities in the United States with more democratically structured departments has established itself as a globally dominant academic system till today. Although Ben-David has not explicitly studied social variation in the gate-keeping role, his early studies demonstrated the context-dependent nature of knowledge production in different academic systems. Despite an increasingly globalized higher education system, the historical variance of academic institutions and national traditions of scholarship can be regarded as still of importance. The fact that the gatekeeper's role is subject to remarkable historical change and cultural diversity reminds us of the need to apply context-sensitive research techniques in empirically studying it.

Seen from the perspective of the sociology of science, the gatekeeper's role is of relevance for maintaining what is constitutive for scientific knowledge production: an institutionalized procedure of 'organized skepticism' (Merton, 1942) for assuring scientific quality standards, shaping the growing stock of knowledge, and contributing to scientific advance. Although by current science studies it is widely acknowledged that subjective judgments are an indispensable part of any evaluative activity in science (Polanyi, 1958), it remains an open question how and to which extent this is the case. The gatekeeper's role seems to be more important in counter-balancing subjective effects of scientific evaluations by implementing the norm of universalism in academic judgments. Even though the most universal evaluation systems in science might not always be the most efficient ones for advancing scientific knowledge, it can be assumed that "'Universalism' and 'Fairness' are 'good' ways to distribute rewards because they fit our general value system" (Cole, 1992: p. 203). The continuing relevance of the gatekeeper's role lies in the fact that it fundamentally forms both the procedures and outcomes of scientific knowledge production. In this sense, it contributes to what is constitutive for science as a social institution.

See also: Academic Careers in Comparative Perspective; Discrimination and the Law; Funding of Social Science, History of; Network Analysis; Peer Review: Organized Skepticism; Science and Technology Studies, History of.

Bibliography

Aguirre, A., 2000. Academic storytelling: a critical race theory story of affirmative action. Sociological Perspectives 43 (2), 319–339.

Appelt, E., Jarosch, M. (Eds.), 2000. Combating Racial Discrimination: Affirmative Action as a Model for Europe. Berg, Oxford and New York.

Bacchi, C.L., 1996. The Politics of Affirmative Action: 'Women', Equality and Category Politics. Sage, London.

Ben-David, J., 1971. The Scientist's Role in Society: A Comparative Study. Prentice-Hall, Englewood Cliffs, NJ.

Cole, J.R., Cole, S., 1973. Social Stratification in Science. University of Chicago Press, Chicago.

Cole, S., 1992. Making Science: Between Nature and Society. Harvard University Press, Cambridge.

Corra, M., Willer, D., 2002. The gatekeeper. Sociological Theory 20 (2), 180–207.

Coser, L.A., 1965. Men of Ideas: A Sociologist's View. Free Press, New York.

Coser, L.A., 1975. Publishers as gatekeepers of ideas. Annals of the American Academy of Political and Social Sciences 421, 14–22.

Coser, L.A., Kadushin, C., Powell, W.W., 1982. Books: The Culture and Commerce of Publishing. Basic Books, New York.

Crane, D., 1967. The gatekeepers of science: some factors affecting the selection of articles for scientific journals. The American Sociologist 2, 195–201.

Crane, D., 1972. Invisible Colleges: Diffusion of Knowledge in Scientific Communities. University of Chicago Press, New York and Chicago.

de Solla Price, D., 1963. Little Science, Big Science. Columbia University Press, New York.

Hargens, L.L., Hagstrom, W.A., 1967. Sponsored and contest mobility of American academic scientists. Sociology of Education 40, 24–38.

Harper, S., Reskin, B., 2005. Affirmative action at school and on the job. Annual Review of Sociology 31, 357–379.

HESA Higher Education Statistics Agency, 2012. Staff in Higher Education Institutions 2010/11 in the United Kingdom. www.hesa.ac.uk.

Keith, B., Layne, J., Babchuk, N., Johnson, K., 2002. The context of scientific achievement: sex status, organizational environments, and the timing of publication on scholarship outcomes. Social Forces 80 (4), 1253–1281.

Lamont, M., 2009. How Professors Think. Inside the Curious World of Academic Judgment. Harvard University Press, Cambridge, UK.

Lazarsfeld, P.F., Berelson, B., Gaudet, H., 1944. The People's Choice: How the Voter Makes up His Mind in a Presidential Campaign. Columbia University Press, New York and London.

Lewin, K., 1943. Forces behind food habits and methods of change. Bulletin of the National Research Council 108, 35–65.

Merton, R.K., 1942. A note on science and democracy. Journal of Legal and Political Sociology 1, 115–126.

Merton, R.K., 5 January 1968. The Matthew effect in science. Science 199 (3810), 55–63.

Merton, R.K., 1973. The Sociology of Science. Theoretical and Empirical Investigations. University of Chicago Press, Chicago.

Merton, R.K., Riley, M.W. (Eds.), 1980. Sociological Traditions from Generation to Generation: Glimpses of the American Experience. Ablex Publishing Corporation, Norwood, NJ.

Ni, C., Sugimoto, C.R., Cronin, B., 2013. Visualizing and comparing four facets of scholarly communication: producers, artifacts, concepts, and gatekeepers. Scientometrics 94, 1161–1173.

Platt, J., 1996. A History of Sociological Research Methods in America 1920–1960. Cambridge University Press, Cambridge, UK.

Polanyi, M., 1958. Personal Knowledge. Routledge and Kegan Paul, London, UK.

Turner, R., 1960. Sponsored and contest mobility and the school system. American Sociological Review 25, 855–867.

Whitley, R., Glaser, J., Engwall, L. (Eds.), 2010. Reconfiguring Knowledge Production: Changing Authority Relationships in the Sciences and Their Consequences for Intellectual Innovation. Oxford University Press, Oxford, UK.

Zuckerman, H., 1977. Scientific Elite: Studies of Nobel Laureates in the United States. Transaction Publishers, New Brunswick, NJ and London, UK.

Zuckerman, H., Merton, R.K., 1971. Patterns of evaluation in science: institutionalization, structure and functions of the referee system. Minerva 9 (1), 66–100.

Zuckerman, H., Merton, R.K., 1972. Age, aging, and age structure in science. In: Riley, M.W., Johnson, M., Foner, A. (Eds.), A Sociology of Age Stratification, Aging and Society, vol. 3. Russell Sage Foundation, New York, pp. 292–356.

Gauss, Carl Friedrich (1777–1855)

Richard W Farebrother, Victoria University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Carl Friedrich Gauss was born on 30 April 1777 in Brunswick (Germany) and died in Göttingen (Germany) on 23 February 1855. Gauss contributed to all areas of science known in his day, notably to astronomy, physics, and geodesy, but this article concentrates on his contributions to statistics and numerical methods. In 1805, Legendre published an algebraic procedure for the linear fitting problem that he named the method of least squares. Gauss's first (1809) statistical derivation of this procedure was based on a combination of the normal distribution with a Bayesian variant of the maximum likelihood optimality criterion. Later, Gauss (1823–28) gave a very full account of an alternative decision-theoretic approach, which employed the mean squared error of the estimator as optimality criterion. In particular, he identified the least squares estimator as the linear unbiased estimator of the parameters with minimum mean squared error. Gauss developed the Gaussian elimination technique for solving determinate systems of linear equations, and explained how nonlinear relationships may be fitted by means of an iterative variant of the method of least squares. He also developed a second iterative (relaxation) procedure related to the Gauss–Seidel procedure.

Introduction

The name of Gauss will already be known to many readers for its use in such phrases as Gaussian distribution, Gauss linear model, and the Gauss–Markov theorem, and for its use in the titles of the Gaussian elimination, Gauss–Jordan, Gauss–Doolittle, Gauss–Newton, and Gauss–Seidel procedures, if not for the *Système International* unit of magnetic induction and the associated terms Gaussage and degauss.

Gauss contributed to all areas of science known in his day, notably to astronomy, physics, and geodesy, but in this article we shall concentrate on his contributions to statistics and numerical methods. For more wide-ranging surveys of his work, see Eisenhart (1968) and May (1972).

Life of Gauss

Carl Friedrich Gauss (originally Johann Friedrich Carl Gauss) was born on 30 April 1777 in Brunswick (Germany). He was the son of a bricklayer and gardener who also kept the accounts of a local insurance company. Gauss received his elementary education at St. Katherine's School in Brunswick. However, his humble origins were not allowed to hamper his intellectual development, for, in 1791, his talent for mathematics was brought to the attention of Carl Wilhelm Ferdinand, Duke of Brunswick, who undertook to support his continuing education, at first at the *Collegium Carolinum* in Brunswick and later at the University of Göttingen (in the Hanoverian territory of King George III of Great Britain and Ireland). Gauss did not graduate from the University of Göttingen but was awarded a doctoral degree *in absentia* by the University of Helmstedt in 1799 for his work on the fundamental theorem of algebra.

In October 1806, the French land forces consolidated the command of German territory that they had established in 1805 in the battle of Austerlitz by defeating a Prussian army under the command of the Duke of Brunswick at Auerstedt (and mortally wounding the Duke). As a consequence, when in 1807 Gauss submitted the manuscript of his future *Theoria Motus Corporum Coelestium* for publication in German, the political situation was such that his publisher felt obliged to insist on its translation into Latin. Consequently, the manuscript that Gauss had completed in 1806 was not published until 1809. Fortunately, Gauss had an excellent command of Latin, so that this imposition may justly be compared with requiring a modern German scholar to translate his work into English before publication. On the other hand, Gauss and other German scholars continued to publish in Latin until the middle of the nineteenth century.

In 1807 Gauss was appointed director of the University of Göttingen observatory and professor of mathematics at the University of Göttingen. He continued to hold both of these positions until his death in Göttingen on 23 February 1855.

Elementary Fitting Procedures

Given any number of observations n on a lesser number q of variables, the linear fitting problem is concerned with establishing the linear relationship between one of these variables and the remaining $q - 1$ variables and a constant term. Before the middle of the eighteenth century, practitioners were obliged to employ a variety of more or less arbitrary procedures to determine the values of the parameters of such relationships. But between 1750 and 1800 four algebraic procedures were proposed for the linear fitting problem, namely the method of selected points which sets q of the deviations from the fitted relationship equal to zero; the method of averages which sets q simple linear combinations of the deviations equal to zero; the method of (constrained) least absolute deviations which chooses values for the parameters to minimize the sum of the absolute deviations subject to the condition that the signed deviations sum to zero; and the minimax absolute deviation procedure which chooses values for the parameters to minimize the largest absolute deviation.

By contrast with Laplace (*see* Laplace, Pierre-Simon, Marquis de (1749–1827)), who wrestled with various aspects

of the last three algebraic fitting procedures for more than 30 years (1786–1818), Gauss (1809) did not have much to say about any of them. He summarily rejected the method of selected points as the imposition of zero values on q of the deviations necessarily requires some of the remaining deviations to take larger absolute values than they need. Then, turning to the method of least absolute deviations, he defined the unconstrained variant of the problem (which Laplace subsequently named the method of situation) and characterized the solution of either variant of this fitting procedure by a set of zero deviations equal in number to the number of unknown constants minus the number of constraints. He also noted that once the set of zero deviations defining the optimal fit had been determined, the remaining observations had no influence on the values to be assigned to the unknown constants. Then, presumably recollecting his critical view of the method of selected points, he seems to have had nothing further to say about either variant of this procedure.

Method of Least Squares

In 1805 Adrien-Marie Legendre (1752–1833) published an alternative algebraic procedure, which he named the method of least squares. This procedure chooses values for the parameters of the relationship to minimize the sum of the squared deviations between the observed values and the fitted relationship. For a detailed history of the four algebraic procedures, of Section Elementary Fitting Procedures and the method of least squares, see Farebrother (1999). Also note that the relevant pages of Legendre's original statement have been reproduced by Stigler (1986: p. 56).

Legendre's approach to the fitting problem was entirely algebraic and had no statistical content. Further, he did not prove many of the optimality properties he claimed for the method of least squares. Indeed, he had not even shown that the values determined by setting the q first derivatives equal to zero actually minimize the sum of the squared deviations function (except in the unlikely case when the fitted relationship is exact). Such a proof was subsequently supplied by Gauss (1809). (Further, many readers will know that Legendre's solution represents a minimum of the sum of the squared deviations function as the corresponding $q \times q$ matrix of second derivatives is symmetric and positive definite.)

As recently as 2005, it was discovered that the portrait universally employed to illustrate articles about the mathematician Adrien-Marie Legendre was actually that of the butcher and revolutionary politician Louis Legendre (1752–97).

Normal Distribution

As a precursor to his first statistical derivation of the method of least squares, Gauss obtained the statistical distribution now known as the normal or Gaussian distribution from a preliminary analysis in which the arithmetic mean of a set of direct observations is associated with the maximum value of the Bayesian variant of the maximum likelihood or maximum probability optimality criterion.

Despite its obvious defect, Gauss's argument spawned a host of papers by various authors which were concerned with the derivation of the Principle of the Arithmetic Mean from simpler axioms; see Farebrother (1999) for details. In particular, in their textbook *The Calculus of Observations* (1924), Sir Edmund Taylor Whittaker (1878–1937) and G. Robinson showed that the arithmetic mean is the only twice differentiable function of the observations, which is equivariant to changes of location, equivariant to changes of scale, and invariant to permutations of the order of the observations.

Similarly, in his *A Treatise on Probability* (1921), Keynes (*see* Keynes, John Maynard (1883–1946)) used variants of Gauss's argument to identify continuous symmetric and nonsymmetric distributions associated with the arithmetic, geometric, and harmonic means and the median, and suggested that it is not possible to find such a distribution associated with the mode.

Most Probable Value Criterion

As noted above, Gauss's first statistical derivation of the method of least squares was based on a combination of the normal or Gaussian distribution with a Bayesian variant of the maximum likelihood or maximum probability optimality criterion. As Gauss was working in a Bayesian context, there was no need for him to estimate the disturbance variance, but he employed a simple algebraic procedure that we would now associate with the LDL′ decomposition of a symmetric matrix to obtain measures of relative precision of each of the parameter estimates.

Priority Disputes

In his 1809 book, Gauss referred to the method of least squares as 'our principle, which we have made use of since the year 1795' and thus sparked off a serious priority dispute with Legendre that ran on for several years. Although Legendre had certainly been the first to publish a description of the method of least squares, it is now generally agreed that Gauss had employed a similar procedure in his unpublished work, notably in his 1801 analysis of the geodetic measurements made in France, see Plackett (1972) and Stigler (1999: Chapter 17).

In this context, we should also mention that the (normal) probability distribution named for Gauss had previously been discussed by Abraham de Moivre (1667–1754) and Pierre-Simon Laplace (1749–1827) although neither of Gauss's predecessors seem to have considered the normal curve as a probability distribution in its own right; *see* Husserl, Edmund (1859–1938) and Stigler (1999: Chapters 14 and 22).

Following the rediscovery of his work in the 1870s, somewhat belated claims for priority were also made on behalf of the Irish-American Robert Adrain (1775–1843), who in 1809 had identified the normal distribution independently of Gauss (if not of Legendre) and applied the method of least squares to a variety of practical problems related to land surveying and dead reckoning at sea.

Asymptotic Results

In the main text of his *Théorie analytique des probabilités* of 1812, Laplace had made use of a nonrigorous asymptotic argument to establish the optimality of the method of least squares. Gauss (1816) employed a similar technique to establish the relative efficiency of the scaled sums of powers of the absolute disturbance terms when employed as estimates of the variance of their common distribution function. In particular, he found that the scaled sum of the squared disturbances is a more efficient estimator of the variance parameter than is the scaled sum of their absolute values and significantly better than the scaled sums of their third and higher powers.

The Most Plausible Value Criterion

In 1818–19 Gauss abandoned his earlier Bayesian derivation of the method of least squares and adopted an alternative approach, which he later published as Gauss (1823–1828) in three parts. Farebrother (1999) has suggested that the principal features of Gauss's second approach to the fitting problem may be traced to a careful recombination of certain elements in Laplace's earlier work. Gauss adopted Laplace's decision-theoretic approach to the fitting problem, and his restriction to what we would call a class of linear unbiased estimators, but employed finite sample theory rather than asymptotic analysis, and chose a value for the parameter to minimize the expected value of the squared error of estimation where Laplace had chosen to minimize its expected absolute value.

The principle result of this new theory, now known as the Gauss–Markov theorem, identifies the least squares estimator as the member of the class of linear unbiased estimators with minimum mean squared error (and thus minimum variance).

On the other hand, it came as something of a surprise to the present author to discover in 2002 that the result underlying the so-called Gauss–Markov theorem is algebraically identical to the result defining the method of least squares.

In this context, it is interesting to note that Gauss's restriction to the class of linear unbiased estimators was an accidental by-product of Laplace's earlier concern with estimators defined by the method of averages, see Stigler (1986) or Farebrother (1999). However, Gauss also showed that the scaled sum of squared deviations is an unbiased estimator of the disturbance variance, but without making any strong claims for the optimality of the result.

Updating Formulas and Recursive Residuals

Gauss (1823–1828) gave a very full account of his second approach to the subject. In particular, he extended his earlier analysis by considering the effect on the fitting problem of adding a further observation, or of imposing a linear constraint on the parameters of the model, see Plackett (1950). The resulting updating formulas now form the basis of the Kalman filter of times series analysis, see Farebrother (1990–1992).

In this connection, it is interesting to note that a particular use of Gauss's updating formulas results in a simple demonstration of Paul Anthony Samuelson's (1915–2009) formula for how deviant the square of a particular observation can be when expressed as a proportion of the sample variance.

The standard updating formulas also implicitly define the set of $n-q$ recursive residuals employed in some modern statistical tests.

Solution of Linear Equations I

The solution technique now known as Gaussian elimination is a development of the procedures known to Tobias Mayer (1723–62) and the ancient Chinese for solving systems of q linear equations in q unknowns. In his (1809) worked example, Gauss employed a solution technique, which may possibly be associated with the ancient method, see Farebrother (1999). But in 1811 Gauss outlined the steps of a solution technique that identify the elimination procedure which now goes by his name. This procedure has two phases: in the forward elimination phase the first equation is used to eliminate one of the unknown constants from the other $q-1$ equations; then the second equation is used to eliminate a second unknown constant from the remaining $q-2$ equations; and so on. This yields an upper triangular system of equations, which is then solved for the unknown constants in reverse order.

Gauss's early work in astronomy was concerned with the computation of the elements of the orbits of the asteroids Ceres and Pallas, which had been discovered in 1801 and 1803, respectively. It therefore comes as no surprise to find that Gauss (1811) illustrated his computational technique by a worked example concerned with the least squares determination of the six elements of the orbit of Pallas. (It is interesting to note in passing that the discovery of these asteroids is also remembered in the names cerium and palladium attached to the chemical elements with atomic numbers 58 and 46 discovered in the same years.)

In the standard elimination procedure, the elements of the upper triangular system are gradually adjusted until they take their final values. In 1881 Myrick Hascall Doolittle (1830–1913), a 'computer' at the US Coast and Geodetic Survey, published his account of a variant of Gauss's procedure in which all the calculations necessary to determine the final value of any given element of the upper triangular system are performed in a single operation. This slight adjustment of the standard computational procedure was extremely valuable in an era of pencil and paper computations and continues to be of value today when particularly accurate results are required, see Farebrother (1988).

Numerous minor variants of the Gauss–Doolittle procedure were proposed over the next 70 years, but it is perhaps fitting to single out that of Henry Agard Wallace (1888–1965) who independently proposed a method similar to Doolittle's in 1925, some 20 years before he served as US Vice-President in 1941–45.

Solution of Linear Equations II

The so-called Gauss–Jordan procedure differs from the original Gaussian elimination procedure in that each step of the forward elimination phase of the revised procedure eliminates

one of the unknown constants from all but one of the equations. Thus, the first equation is used to eliminate one of the unknown constants from all equations except the first; the second equation is used to eliminate a second unknown constant from all equations except the second; and so on until the original system of equations is reduced to diagonal form and this reduced system solved by simple division.

This minor modification of the standard elimination procedure will be familiar to many readers from textbook descriptions of the simplex procedure of linear programing. It leads to a considerable simplification of the second (back substitution) phase of the elimination procedure but usually at the cost of some loss of numerical accuracy, see Farebrother (1988) and Stewart (1995).

Matrix Algebra

As a by-product of his work in number theory and the calculus of observations, Gauss (1801, 1809, 1823–1828) made use of several concepts now recognized as belonging to the theory of matrices. These include linear dependence; linear transformations; the Lagrange LDL′ and Cholesky LL′ decompositions of a symmetric matrix; the rank-one adjustment of a symmetric matrix; the inverse of a triangular matrix; the inverse of a symmetric matrix; and, possibly, the generalized inverse of a matrix of full rank. He also explicitly proved that the inverse of a symmetric matrix is itself symmetric.

Moreover, in the *Premier Supplément* (1816) to his *Théorie analytique des probabilités*, Laplace employed from of the ideas implicit in the notation of Gauss's (1809, 1811) computational procedures to develop a prototype of the orthogonalization procedure which modern textbooks of numerical analysis identify as a 'modification' of the so-called classical Gram-Schmidt orthogonalization procedure first enunciated some 80 years after Laplace's death by Erhard Schmidt (1876–1959) in 1907. Farebrother (1988) has suggested that the current terminology may be simplified by naming these two orthogonalization procedures for Laplace and Schmidt, respectively.

Despite these early beginnings, the familiar matrix formulation of the least squares problem did not materialize until the 1930s, and it did not become a common feature of statistical textbooks until the second half of the twentieth century, see Plackett (1950) and Farebrother (1999). Indeed, the immediate reaction of two leading statisticians, Florence Nightingale David (1909–93) and Jerzy Neyman (*see* Neyman, Jerzy (1894–1981)), to a 1935 publication by Alexander Craig Aitken (1895–1967) was to reestablish the same result using the more familiar notation of determinants. Fortunately, the only permanent effect of this retrograde step was the superfluous addition of the name of Andrei Andreivich Markov (1856–1922) to that of Gauss in the conventional name of the Gauss–Markov theorem of Section The Most Plausible Value Criterion.

Iterative Fitting Procedures

Gauss (1809) also explained how nonlinear relationships may be fitted by means of an iterative variant of the method of least squares, now known as the Gauss–Newton procedure, in which the fitted relationship is linearized by a Taylor series expansion and fitted by the method of least squares. This process being repeated until a satisfactory level of convergence is obtained.

Despite its early origin, and the publication by Gauss of detailed worked examples in astronomy and geodesy, the Gauss–Newton procedure was not widely adopted for practical work in other areas until the necessary computing equipment became readily available in the second half of the twentieth century. Instead, practitioners avoided the computational burden imposed by the iterative procedure by transforming their nonlinear relationships to a linear or a more linear form before estimating the parameters of the resulting equations by a *single* application of the method of selected points, the method of averages, or the method of least squares.

Nor was the use of such preliminary transformations restricted to the work of applied scientists with little training in statistics as Hotelling (*see* Hotelling, Harold (1895–1973); Pearson, Karl (1857–1936)) among others employed variants of this method in their practical work. Indeed, as late as 1930, Henry Schultz (1893–1938) attempted to persuade his fellow practitioners to employ the 'true' method of least squares when fitting nonlinear statistical models as the alternative approximate procedures sacrificed variance to computational ease, see Farebrother (1998, 2006).

A second iterative procedure usually named for Gauss is the Gauss–Seidel procedure, although Gauss and Seidel were actually not concerned with the procedure that usually goes by this name, but with a distinct fitting procedure known as the method of relaxation, see Gauss (1823–1828) for details.

Concluding Remarks

The method of least squares and the related statistical procedures known as linear regression analysis and the analysis of variance are widely used in all areas of science. Thus there can be few numerate practitioners of the social and behavioral sciences who have not read research reports that employ these methods to fit linear or curvilinear relationships to data on the variables of interest.

The method of least squares was developed by Legendre, Gauss, and Laplace in the first quarter of the nineteenth century in the context of important practical problems in astronomy and geodesy. But it was not until later in the century that Quetelet (*see* Quetelet, Adolphe (1796–1874)), Galton (*see* Galton, Sir Francis (1822–1911)), Edgeworth (*see* Edgeworth, Francis Ysidro (1845–1926)), and Pearson (*see* Pearson, Karl (1857–1936)) among others adjusted this technique to the needs of the social and behavioral sciences. There was no great shortage of practitioners trained in the use of advanced statistical techniques, but such individuals as were active in this area were not able to treat their observations with the necessary level of detachment, so that the original source of the observations remained a significant feature of their subsequent analyses.

The method of least squares was absorbed relatively rapidly into psychology, but its rate of infusion into the economic, social, and behavioral sciences (other than psychology) was

more gradual. Indeed, applications of this technique were not at all common in these areas until well into the twentieth century. See Stigler (1986, 1999) for a detailed account of the gradual transformation of these subject areas (*see* Quetelet, Adolphe (1796–1874); Edgeworth, Francis Ysidro (1845–1926); Thorndike, Edward Lee (1874–1949)).

The history of the method of least squares clearly demonstrates that practical scientists with little or no interest in the social and behavioral sciences can make significant contributions to this subject area by developing general analytical procedures that can be adjusted to the particular needs of social and behavioral scientists.

Readers wishing to study Gauss's original work on the method of least squares will be reassured to learn that they do not have to read his texts in their original Latin as Bertrand (1855), Börsch and Simon (1887), and Trotter (1957) have translated Gauss's principal works on the method of least squares into French, German, and English, respectively.

See also: Algorithm; Bayesian Theory, History, Applications, and Contemporary Directions; Laplace, Pierre Simon, Marquis de (1749–1827); Stochastic Models.

Bibliography

Bertrand, J.L.F., 1855. Méthode des moindres carrés: mémoires sur la combinaison des observations par Ch.-Fr. Gauss. Mallet-Bachelier, Paris, France [1957 Trotter, H.F. Gauss's Work (1803–1826) on the Theory of Least Squares, Technical Report 5, Statistical Techniques Research Group. Princeton University, Princeton, NJ].

Börsch, A., Simon, P., 1887/1964. Abhandlungen zur Methode der kleinsten Quadrate von Carl Friedrich Gauss. Physica-Verlag, Würzburg, Germany.

Eisenhart, C., 1968. Carl Friedrich Gauss (1777–1855). In: Sills, D. (Ed.), International Encyclopedia of the Social Sciences, vol. 6. Macmillan, New York, pp. 75–81.

Farebrother, R.W., 1988. Linear Least Squares Computations. Marcel Dekker, New York.

Farebrother, R.W., 1990–1992. Mnemonics for the Kalman filter updating formulas (in two parts). Statistische Hefte 31: 281–284 and 33: 269–272. springerlink.com.

Farebrother, R.W., 1998. Nonlinear curve fitting and the true method of least squares. The Statistician 47, 137–147. jstor.org.

Farebrother, R.W., 1999. Fitting Linear Relationships: A History of the Calculus of Observations 1750–1900. Springer-Verlag, New York.

Farebrother, R.W., 2006. Early explorations in econometrics. In: Patterson, K., Mills, T.C. (Eds.), Palgrave Handbook of Econometrics, vol. 1. Palgrave Macmillan, Basingstoke, UK, pp. 88–116.

Gauss, C.F., 1801/1870. Disquisitiones Arithmeticae. In: Carl Friedrich Gauss Werke, vol. 1. Dieterichsche Universitäts-Druckerei, Göttingen, Germany. gdz.sub.uni-goettingen.de [1966 A.A. Clarke, English Trans. Disquisitiones Arithmeticae. Yale University press, New Haven, CT].

Gauss, C.F., 1809/1871. Theoria motus corporum coelestium in Sectionibus Conicis Solem Ambientium. In: Carl Friedrich Gauss Werke, vol. 7. F. Perthes, Gotha, Germany. gdz.sub.uni-goettingen.de [(1857) 1966 Davis, C.H. Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections. Dover, New York.] [1864 Dubois, E. Théorie des mouvements des corps célestes parcourant des sections coniques autour du soleil. Arthus Bertrand, Paris, France.] [1865, Haase, C. Theorie der Bewegung der Himmelskörper welche in Kegelschnitten die Sonne umlaufen. Meyer, Hannover, Germany].

Gauss, C.F., 1811/1874. Disquisitio de elementis ellipticis Palladis ex oppositionibus annorum 1803, 1804, 1805, 1807, 1808, 1809. In: Carl Friedrich Gauss Werke, vol. 6. Dieterichsche Universitäts-Druckerei, Göttingen, pp. 1–24. gdz.sub.uni-goettingen.de.

Gauss, C.F., 1816/1880. Bestimmung der Genauigkeit der Beobachtungen. In: Carl Friedrich Gauss Werke, vol. 4. Dieterichsche Universitäts-Druckerei, Göttingen, pp. 109–117. gdz.sub.uni-goettingen.de.

Gauss, C.F., 1823/1903. Letter to Gerling, Göttingen 26 December 1823. In: Carl Friedrich Gauss Werke, vol. 9. Dieterichsche Universitäts-Druckerei, Göttingen, pp. 278–281. gdz.sub.uni-goettingen.de [1951 Forsythe, G.E. Gauss to Gerling on relaxation, Mathematical Tables and Other Aids to Computation 5: 255–258].

Gauss, C.F. 1823–1828/1880. Theoria combinationis observationum erroribus minimis obnoxiae (in three parts). In: Carl Friedrich Gauss Werke, vol. 4. Dieterichsche Universitäts-Druckerei, Göttingen, pp. 3–93. gdz.sub.uni-goettingen.de. [1995 Stewart, G.W. Theory of the Combination of Observations Least Subject to Errors. Society for Industrial and Applied Mathematics Publications, Philadelphia, PA].

May, K.O., 1972. Carl Friedrich Gauss. In: Gillispie, C.C. (Ed.), Dictionary of Scientific Biography, vol. 5. Scribners, New York, pp. 298–315.

Plackett, R.L., 1950. Some theorems in least squares. Biometrika 37, 149–157. Online open access.

Plackett, R.L., 1972. The discovery of the method of least squares. Biometrika 59, 239–251. Online open access.

Stigler, S.M., 1986. The History of Statistics: The Measurement of Uncertainty before 1900. Harvard University Press, Cambridge, MA.

Stigler, S.M., 1999. Statistics on the Table: The History of Statistical Concepts and Methods. Harvard University Press, Cambridge, MA.

Gay and Bisexual Latinos in US Gay Culture

Xavier L Guadalupe-Diaz, Framingham State University, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Gay culture and community in the US are rooted in shared histories of struggle, celebration, pride, and identity. As a diverse community, Latinos comprise a significant population among gay and bisexual men (GBM) in the US. The distinct lived experiences of Latino GBM often result in marginalization from both their broader communities and gay culture. While Latino cultures vary, overarching themes of machismo and familismo contextualize experiences of homophobia and heterosexism within their broader communities. In addition to gay identity development, racism and stereotyping, and impacts on health and well-being, various factors contribute to the relegation of Latino GBM in mainstream US gay culture.

The lesbian, gay, bisexual, and transgender (LGBT) community encompasses a diverse range of intersecting social identities along the lines of race, ethnicity, nation of origin, religions, gender identities, and more. Among these, Latino gay and bisexual men (GBM) characterize unique realities facing many other racial and ethnic minority groups within the broader context of mainstream American gay culture.

A brief note on nomenclature: The terms Hispanic and Latino are often used interchangeably. In the 2010 Census, Hispanic or Latino was defined as a "person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin regardless of race" (US Census Bureau, 2011). However, Hispanic generally refers to a group's relationship to Hispania, or modern-day Spain, while Latino more specifically refers to geographic origin; i.e., Latin America, Central America, or the Caribbean. Recently, the Pew Research Center's Hispanic Trends Project explored preference in terminology among 5103 Latino adults. Among the findings, researchers found that half of the sample had no preference in terminology. Nevertheless, for purposes of this article, the term Latino was chosen to describe those GBM with ethnic ties to any country in Central or South America or the Caribbean.

Within the broader scope of gay culture, Latino GBM face many of the same issues regarding stereotyping, cultural and religious problems, and racial disparities that other nonwhite GBM may experience. While Latin culture is vibrant and diverse with a variety of nation-specific qualities globally, overall there remains a strong, patriarchal emphasis that constructs rigid expectations for Latin men. Commonly referred to as 'machismo', the social constructs of Latin masculinity place heavy expectations on men to be emotionally distant, authoritarian, aggressive, and be a hypersexual heterosexual (Torres, 1998). Further, 'familismo' or familism places a heavy emphasis on the centrality of family. For Latino families, family interests take priority over individual pursuits or concerns (Ingoldsby, 1991). For Latino GBM, this means the pressures of heteronormative family structures and strong legacies of predominantly conservative Catholic upbringings may intensify feelings of shame.

Beyond these issues, in a predominantly white and middle-class-oriented gay culture, Latino GBM may often encounter discrimination or a lack of inclusivity in the gay community. While racism and stereotyping affect the broader Latino community, Latino GBM face specific adversities within the gay community. Some scholars have denoted this difference as general versus gay-related racism (Ibañez et al., 2009). Whereas general racism may refer to the larger attitudes and institutional biases on the basis of race, gay-related racism refers to racial biases within the gay community itself. These experiences may include rejection from potential sexual or romantic partners on the basis of race, being made to feel uncomfortable in gay settings, or sexual stereotyping and more.

Latin Cultural Issues and Gay Identity

Strong patriarchal and machista histories have unfolded and continue to dominate Latin communities around the world. Cultural dynamics plays a crucial role in shaping sexual identities, behaviors, and perceptions. While machismo informs Latin masculinity in a strictly heterosexual expectation, Latino GBM internalize the same narratives as heterosexual Latino men. While Latino masculinities are diverse, several aspects of machismo are central and frequently found across Latin communities. These narratives inform expectations about sexual identity and relationships that place men in dominant, active roles while women are passive or receptive partners. Scholars have argued that one such narrative of machismo is the representation of Latin masculinity as accomplished through penetrative sexual intercourse (Gupta, 2002; Reisen et al., 2010b). As women are expected to be the only receptive partner, homosexuality is assumed to involve being the receptive partner in male same-sex intercourse. Some scholars have argued that among Latinos, regardless of the gender of the receptive partner in penetrative sex, the insertive partner may still be perceived as heterosexual (Carillo, 2002; Finlinson et al., 2006).

In describing male same-sex penetrative sex, the broader American gay community popularly refers to the insertive role as 'top', the receptive role as 'bottom', and those who may perform either role as 'versatile'. In the broader gay community these roles are not rigid and most GBM describe taking versatile roles (Wegesin and Meyer-Bahlburg, 2000). For Latino GBM, tops are referred to as activos and bottoms are referred to as pasivos. As no universal term exists to describe versatile roles, some scholars have suggested that Latino GBM make stricter distinctions between activos and pasivos indicating a stronger adherence to a heternormative pattern in sexual dynamics (Carballo-Diéguez et al., 2004). Carballo-Diéguez et al., (2004: p. 160) argued that "activo and pasivo roles, especially in anal

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10261-2

sex, have radically different meanings for Latino GBM". As traditional notions of masculinity are linked to penetrative roles, pasivo roles are characterized as submissive and feminine. Machismo relegates pasivo roles to that of less than what an ideal man should be, privileging activo roles. This may contribute to the stigmatization of pasivo roles while excusing the homosexual behaviors of activo Latin GBM.

Familismo in Latin communities also commonly plays a crucial role in the development of gay identities. In comparison to Anglo family structures which may be more individualistic, Latino culture prioritizes the desires of the family unit over those of individual members. Traditionally conservative Latino family backgrounds contribute to the silencing of many Latino GBM especially when considering the high expectations placed on Latino men. As the family unit takes priority, the continuity of family lineages are rooted in a heterosexist expectation. Further, family rules pertaining to dating, abstinence, and marriage are highly regarded by Latino children. This may be more the case for Latino children born outside of the US as opposed to those who are American-born (Ingoldsby, 1991). For Latino GBM, coming out may seem like a selfish endeavor that would bring shame to the family and extreme disapproval. In addition, religiosity and adherence to traditional Catholic or other conservative Christian denominations characterize many Latino family backgrounds. Beyond the humiliation engendered on the family, Latino GBM also face rejection from religious institutions that have typically been central to the Latino family.

While much of what is known about these traditional cultural conflicts address a broader Latino culture, some differences may exist across nations of origin. Although some Latino nations such as Bolivia and Paraguay have constitutional bans on marriage equality indicating much stronger cultural resistance toward homosexuality, many other nations in Central and South America have legalized same-sex marriage or other unions. Argentina became the first nation in South America to legalize same-sex marriage in 2010 and more recently Uruguay has followed suit. Other South American nations including Colombia and Ecuador have legalized some form of recognizable same-sex unions. In addition, Mexico City and the Mexican state of Quintana Roo have both legalized same-sex marriage. These variations may suggest that differences across nations in Latin America exist along the lines of tolerance and the political acceptability of same-sex relationships.

Among Latino cultures, Mexican machismo is often thought of as more rigid and masculinist than others. Traditional Mexican gender norms illustrate similar patterns across other Latin cultures that may, however, have a stronger emphasis on bravado and independence. Some scholars have discussed similar narratives through the notion of the Mexican-vaquero or vaquero masculinity in Mexican culture (Gutmann, 2000). Vaquero, or cowboy, is an image of masculinity perpetuated through cultural narratives of manhood in Mexican culture and other Latin American countries (Taggart, 1997). In mainstream American culture, Mexican masculinity is often villainized as hypermasculine, aggressive, and sexist or demeaning toward women often in comparison to white masculinity. While these aspects of masculinity in Latino culture exist, overgeneralizations of Latino men lead to false characterizations. Utilizing a sample of 148 Latino men, Torres et al., (2002) challenged the 'myth of sameness among Latino men and their machismo'. The sample included Mexican Americans, Puerto Ricans, Central Americans, and Cubans among other ethnic identities. Using the Survey on the Application of Masculinity and Gender Role Identity Among Latino Men, the researchers examined machismo utilizing the Macho Scale, the Gender Role Conflict Scale, and the Mirandé Sex-Role Inventory. These measures assessed various aspects of machismo including sensitivity, dominance, emotionality, and more. Among the findings, only approximately 10% of the sample could be characterized by a machismo that was authoritarian, emotionally restrictive, and controlling. Additionally, Mexican Americans made the majority of a group classified as more flexible with gender roles and masculinity. These findings indicate that traditional aspects of machismo may not necessarily be uniform across Latino men and that wider range of gender ideologies is represented.

Stereotyping and Racism

Within the predominantly white, mainstream American gay community, Latino GBM face stereotyping and gay-related racism. Mainstream gay culture may often be characterized as highly commoditized and dominated by white, middle-class expectations of beauty and overall liveliness. In comparison to Asian and Pacific Islander and African American GBM, Latino GBM may not be so rigidly stereotyped sexually by mainstream gay sexual identities. While Asian and Pacific Islander GBM may be characterized as effeminate and submissive, African American GBM are often characterized as dominant and aggressive. These racialized sexual stereotypes in the mainstream gay community may be somewhat different for Latino GBM. Latino GBM may either be characterized as submissive or assumed bottoms or as aggressive, assumed the tops posture. As the Latino community ranges in skin color and body types, these aspects of racialized sexual stereotyping may vary. For example, even when considering all Latino GBM samples, Carballo-Diéguez et al., (2004) examined sexual role behaviors among 294 respondents and found that darker skinned and taller Latino GBM were expected to be activos or tops while those who were lighter skinned or shorter were assumed pasivos or bottoms. In the mainstream gay community, racialized Latino GBM hypersexual stereotypes may construct the aggressive narrative of a dominant Macho similar to those stereotypes faced by African American GBM. Conversely, racialized sexual stereotypes may characterize Latino GBM as submissive or exotic bottoms, similar to the narratives of submissive nonwhite masculinities among Asian and Pacific Islander GBM.

Beyond the general racism experienced by Latino GBM, mainstream gay cultural ideals and standards may also perpetuate racism in the community. Gay-related racism manifests in a wide range of ways including rejection by potential friends or sexual partners on the basis of race or a lack of inclusivity among gay scenes. Commoditized, white gay culture constructs expectations of attractiveness linked with lighter skinned and middle-class appearances (Diaz, 1998). For

Latino GBM, skin color and physical characteristics vary across ethnicities. These varying aspects of physical characteristics result in divergent experiences for Latino GBM. In one example, Ibañez et al., (2009) examined contexts of racism in a sample of 911 Latino gay men. The authors examined variations of experiences on the basis of skin color and physical characteristics (i.e., European, Indian, African/mixed race) among other factors. Further, measures of gay-related racism included aspects of sexual attention given, being treated unfairly, being made to feel uncomfortable at white gay clubs, and being turned down for sex or relationships, all on the basis of race. Among the results were high rates of gay-related racism experienced in the sample but also those factors of skin darkness and non-European physical characteristics significantly predicted greater experiences of gay-related racism. These findings echo others in the literature that illustrates the extent of racial biases in the gay community against Latino GBM (Ayala and Diaz, 2001; Williams et al., 1997).

Health

Cultural barriers, racism, and stereotype have resulted in several disparate outcomes in terms of the health and well-being of Latino GBM. Machismo and family expectations may leave many Latino GBM with little social support within their own communities. Coming out processes are further complicated when considering how central the family unit is among Latinos and the risk of losing vital connections to the community. Beyond the psychological impacts self-repression and/or family rejection, serious physical health consequences have been linked to the lack of discussion on homosexuality in Latino communities as well as to gay-related racism. In comparison to their white peers, Latino GBM are at higher risk for contracting HIV/AIDS (CDC, 2012). While accessibility to sexual health resources is more limited for Latino GBM, cultural factors also play a role in risky behaviors (Rhodes et al., 2011). Insertive role-donning Latino GBM may feel less inclined to use condoms and perceive that to be the responsibility of receptive partners. Further, machismo may involve taking riskier decisions which also influence the lack of safe sex practices (Reisen et al., 2010). Within the context of gay-related racism, Latino GBM may feel they have less negotiating power in sexual relationships. In particular, darker skinned Latino GBM may feel less inclined to request condom usage to attract more sexual partners (Ayala and Diaz, 2001; Ibañez et al., 2009). More recently, Mizuno et al., (2012) examined the influence of experiences with homophobia and racism in a sample of 1081 Latino men who have sex with men. Measures for experiences with homophobia included being beaten, treated rudely, or been made fun of; these experiences were inclusive to homophobia experienced within Latino groups and the broader community. In assessing experiences with racism, the researchers included aspects of gay-related racism. In addition to being beaten or treated rudely on the basis of race, the researchers asked whether they had "felt uncomfortable in a crowd of white gay men, had trouble finding a male lover or boyfriend or had been turned down for sex because of their race or ethnicity" (p. 726). Regarding gay-related racism, 29% of the sample reported having been uncomfortable in a crowd of gay men, 25% reported having had trouble finding a male lover or boyfriend because of race, and 26% had been turned down for sex because of their race; over 40% of the sample experienced both homophobia and racism. While experiences with racism were found to increase the likelihood of risky sexual behavior, those who had experiences with homophobia and racism were more likely than men who reported not experiencing so to report sexually risky behaviors such as unprotected receptive anal intercourse with a casual sex partner.

In addition to the impact that homophobia and gay-related racism may have on sexual options and risk taking, sexual health issues related to internalized homophobia have also been cited as particularly problematic among Latino GBM (Herrick et al., 2012). The internalization of homophobic attitudes and misconceptions about homosexuality has been linked to sexual risk taking and disproportionately affects Latino GBM (Diaz et al., 2001). As an impact on mental health and well-being, internalizing feelings of lower self-worth may result in greater apathy toward self-protection.

Latino Response

One way in which Latino GBM express community solidarity in an often hostile mainstream gay culture is by engaging in organizations, events, or activities aimed specifically at celebrating Latino LGBT and advancing Latino issues. In one example not limited to Latino GBM, 'United We Dream', an immigration social justice organization led by young activists, has expanded an ongoing initiative called the 'Queer Undocumented Immigrant Project' (QUIP). QUIP works to advocate and organize around issues facing undocumented Lesbian, Gay, Bisexual, Transgender, and Queer (LGBTQ) immigrants in the US. Often issues specific to the Latino LGBTQ community go largely ignored by larger, mainstream organizations working to advance gay rights. In 2013, during a large marriage equality rally organized by mainstream organization such as the Human Rights Campaign and the National Gay and Lesbian Task Force, organizers asked a QUIP speaker not to mention that he was undocumented (Brydum, 2013).

Washington DC, Portland, and Long Beach are among the American cities that host annual Latino-specific gay pride events. These events may highlight Latino-specific aspects of gay culture by featuring music, food, celebrities, businesses, and organizations that reflect the interests of the community. In cities or areas in which one Latino ethnicity is prominent, gay pride events are often hosted with specific emphasis to their culture. For example, Latino gay pride events in New York City may be more focused on Puerto Rican culture given the demographics of the city. Likewise, Latino gay pride parades in Los Angeles may reflect more Mexican and/or Central American themes. Several differences exist across Latino ethnicities which are often observable by diverse approaches to these pride events. In the Latino-American Caribbean, Cuba, Dominican Republic, and Puerto Rico communities, salsa, merengue, and bachata are popular music genres while Mexican or Central American genres may include more cumbia and ranchera genres. Various approaches to music influence diverse traditional dances, art and expression, and more. Beyond events, the

gay Latino magazine *Adelante*, is a national publication that highlights news stories, health topics, music, travel, and more. Coverage includes immigration reform and other political issues, arts, and literature by Latino GBM.

Overall, Latino GBM experience unique challenges both within their own cultural contexts and from the mainstream gay community. Recent estimates from the Pew Hispanic Center illustrate a changing dynamic in the Latino community regarding perspectives on homosexuality that indicate progress in acceptance. In the report, 59% of US Latinos responded that homosexuality should be accepted by society; for second generation US Latinos, this increased to 68% (Taylor et al., 2012). While these findings demonstrate a shift toward inclusion, addressing issues of sexual identity, health, and racial discrimination in culturally specific ways remain priorities in improving the lives of Latino GBM.

See also: Friendship Experiences of Lesbian, Gay, Bisexual, and Transgender Individuals; Gay and Lesbian Movements; Gender: Gendering of Categories; Heterosexism and Homophobia; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Queer Theory and Intersectionality; Sexual Minority Families of Choice and Family Networks: We Are Family; Sexual Orientation: Historical and Social Construction.

Bibliography

Arreola, S.G., Ayala, G., Díaz, R.M., Kral, A.H., 2013. Structure, agency, and sexual development of Latino gay men. Journal of Sex Research 50 (3–4), 392–400.

Ayala, G., Diaz, R.M., 2001. Racism, poverty, and other truths about sex: race, class, and HIV risk among Latino gay men. Revista Interamericana de Psicologia 35, 59–77.

Brydum, S., 1 April 2013. HRC and coalition apologize for silencing undocumented, trans activists at supreme court rally. Advocate. http://www.advocate.com/politics/marriage-equality/2013/04/01/hrc-and-coalition-apologize-silencing-undocumented-trans.

Carballo-Diéguez, A., Dolezal, C., Nieves, L., Diaz, F., Decena, C., Balan, I., 2004. Looking for a tall, dark, macho man sexual-role behaviour variations in Latino gay and bisexual men. Culture, Health, & Sexuality 6 (2), 159–171.

Carillo, H., 2002. The Night is Young: Sexuality in Mexico in the Time of AIDS. Chicago University Press, Chicago, IL.

Centers for Disease Control and Prevention, 2012. HIV/AIDS Statistics Overview in the U.S. by Race/Ethnicity, 2011. http://www.cdc.gov/hiv/statistics/basics/ (accessed 02.06.13.).

Diaz, R.M., 1998. Latino Gay Men And HIV: Culture, Sexuality, and Risk Behavior. Routledge, New York.

Diaz, R.M., Ayala, G., Bein, E., Henne, J., Marin, B.V., 2001. The impact of homophobia, poverty, and racism on the mental health of gay and bisexual Latino men: findings from 3 US cities. American Journal of Public Health 91 (6), 927.

Ennis, S., Rios-Vargas, M., Albert, N., 2011. The Hispanic Population 2010: 2010 Census Briefs. United States Department of Commerce – United States Census, Washington, DC. C2010BR-04. http://www.census.gov/prod/cen2010/briefs/c2010br-04.pdf.

Finlinson, H.A., Colón, H.M., Robles, R.R., Soto, M., 2006. Sexual identity formation and AIDS prevention: an exploratory study of non-gay-identified Puerto Rican MSM from working class neighborhoods. AIDS & Behavior 10, 531–539.

Gupta, G.R., 2002. Vulnerability and resilience: gender and HIV/AIDS in Latin America and the Caribbean. http://www.iadb.org/sds/doc/GuptaGeetaRao.pdf.

Gutmann, M.C., 2000. A (short) cultural history of Mexican machos and hombres. In: Lugo, A., Maurer, B. (Eds.), Gender Matters: Rereading Michelle Z. Rosaldo. University of Michigan Press, Ann Arbor, MI, p. 160.

Herrick, A.L., Stall, R., Chmiel, J.S., Guadamuz, T.E., Penniman, T., Shoptaw, S., Plankey, M.W., 2012. It gets better: resolution of internalized homophobia over time and associations with positive health outcomes among MSM. AIDS and Behavior, 1–8.

Ibañez, G., Marin, B., Flores, S., Millett, G., Diaz, R., 2009. General and gay-related racism experiences by Latino gay men. Cultural Diversity and Ethnic Minority Psychology 15 (13), 215–222.

Ingoldsby, B., 1991. The Latin American family: familism vs. machismo. Journal of Comparative Family Studies 22 (1), 57–62.

Meyer, I.H., 2003. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. Psychological Bulletin 129 (5), 674.

Mizuno, Y., Borkowf, C., Millett, G.A., Bingham, T., Ayala, G., Stueve, A., 2012. Homophobia and racism experienced by Latino men who have sex with men in the United States: correlates of exposure and associations with HIV risk behaviors. AIDS and Behavior 16 (3), 724–735.

Reisen, C.A., Cecilia, Z., Bianchi, F., Poppen, P., Sheldin, M., Penha, M., 2010a. Latino gay and bisexual men's relationships with non-gay-identified men who have sex with men. Journal of Homosexuality 57 (8), 1004–1021.

Reisen, C.A., Iracheta, M.A., Zea, M.C., Bianchi, F.T., Poppen, P.J., 2010b. Sex in public and private settings among Latino MSM. AIDS Care 22 (6), 697–704.

Rhodes, S., Hergenrather, K., Vissman, A., Stowers, J., Davis, A., Hannah, A., Alonzo, J., Marsiglia, F., 2011. Boys must be men, and men must have sex with women: a qualitative CBPR study to explore sexual risk among African American, Latino, and White gay men and MSM. American Journal of Men's Health 5 (2), 140–151.

Taggart, J.M., 1997. The Bear and His Sons: Masculinity in Spanish and Mexican Folktales. University of Texas Press, Austin.

Taylor, P., Lopez, M., Martínez, J., Velasco, G., 2012. When labels don't fit: Hispanics and their views of identity. Pew Research Hispanic Center. www.pewhispanic.org/2012/04/04/when-labels-dont-fit-hispanics-and-their-views-of identity/1/ (accessed 30.05.13.).

Torres, J.B., 1998. Masculinity and gender roles among Puerto Rican men: a dilemma for Puerto Rican men's personal identity. American Journal of Orthopsychiatry 68, 16–26.

Torres, J.B., Solberg, V.S.H., Carlstrom, A.H., 2002. The myth of sameness among Latino men and their machismo. American Journal of Orthopsychiatry 72 (2), 163–181.

US Census, 2011. Overview of Race and Hispanic Origin: 2010. http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf (accessed 01.03.2011.).

Wegesin, D.J., Meyer-Bahlburg, H.F.L., 2000. Top/bottom self-label, anal sex practices, HIV risk and gender role identity in gay men in New York City. Journal of Psychology and Human Sexuality 12, 43–62.

Williams, D.R., Yu, Y., Jackson, J.S., Anderson, N.B., 1997. Racial differences in physical and mental health: socioeconomic status, stress, and discrimination. Journal of Health Psychology 2, 335–351.

Gay and Lesbian Movements

Amy L Stone, Trinity University, San Antonio, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Altman, volume 9, pp. 5895–5899, © 2001, Elsevier Ltd.

Abstract

Since World War II, the gay and lesbian movement in the United States has grown from a small network to a large movement. This article traces the origins of the US gay and lesbian movement from before World War II to the turn of the twenty-first century.

Introduction/Overview

The gay and lesbian movement in the United States developed in response to the growing public spaces for gay life, along with large-scale social changes initiated by World War II and urbanization processes. Although this movement has undergone many changes over time and includes many smaller movements, these movements have been fundamentally oriented toward outcomes of visibility and recognition, including recognition through the establishment of laws and policies. These movements have also become increasingly inclusive and expansive, including more attention to the role of class, race, and gender in gay and lesbian organizing, along with the inclusion of bisexual, transgender, queer, questioning, and intersex organizing at different points in time. This article divides the history of gay/lesbian movements into the closet era (pre–World War II), the beginning of the coming out era (World War II to Stonewall), gay liberation and lesbian feminism, human immunodeficiency virus infection/acquired immunodeficiency syndrome (HIV/AIDS), the Religious Right and culture wars, and same-sex marriage in the contemporary lesbian, gay, bisexual, and transgender (LGBT) movement.

The Closet Era (Pre–World War II)

The period before World War II is often referred to by sociologists as 'the closet era,' a time in which gay and lesbian identities were largely underground and unorganized. Although often referred to by scholars as 'the closet era,' historians suggest that 'the closet' was not a meaningful metaphor for gay and lesbian individuals during this time period (Chauncey, 1994).

The creation of a movement around gay and lesbian identities necessitates the development of the modern concept of the 'homosexual' as a certain kind of subject who is attracted to same-sex individuals. It was not until the mid-nineteenth century that the budding discipline of sexology and work by European sexologists like Richard von Krafft-Ebing, Havelock Ellis, and Magnus Hirschfeld developed a concept of the 'sexual invert,' an individual who defied gender norms with their behavior and attraction to individuals of the same sex.

However, the modern concept of homosexuality as a type of same-sex desire without being simultaneously marked by gender deviance was not developed until after the turn of the century. According to research by George Chauncey (1994) on the development of the gay world in pre–World War II New York City, in working class communities this modern concept of homosexuality did not develop until the 1930s.

According to scholars like John D'Emilio (1983), the development of industrialization and spread of wage labor created the conditions for the emergence of gay and lesbian culture. Capitalistic industrialization dramatically transformed family life, sexual relations, and heterosexuality, which in turn laid the conditions for public spaces such as gay bars. These developments also created different structural conditions for gay and lesbian lives, as white, middle-class men had more access to wage labor and urban public spaces. Before World War II, there was some early political organizing in the United States and abroad. This organizing included the German organization the Scientific-Humanitarian Committee created by Magnus Hirschfeld in 1897 and advocacy by gay men in New York City over police repression in the 1930s (Chauncey, 1994).

The Beginning of the Coming Out Era (World War II to Stonewall)

World War II dramatically transformed social life in the United States. It moved both men and women into the cities and into the military, creating more opportunities for homosocial connections independent of the family. In his groundbreaking book, *Coming Out Under Fire: The History of Gay Men and Women in World War Two*, Allan Bérubé (1990) analyzes the way military life during the war both constrained and created opportunities for gay men and lesbians to meet others and explore their sexuality. The war years facilitated entry into the gay world for many men and women at this time.

An urban gay subculture grew in the wake of World War II, in spite of growing public hostility toward homosexuality and the reinforcement of traditional gender roles in the 1950s. Exclusive gay bars were established in cities across the country (D'Emilio, 1983). The gay and lesbian bar was critical in creating the conditions for movement organizing due to the ways it created public spaces, solidarity, and collective identities (Kennedy and Davis, 1993). Although racial and ethnic minorities were frequently excluded from gay bars, black and

Hispanic gay men and lesbians may have congregated at house parties and urban drag balls often.

This growing lesbian and gay subculture was supported by Alfred Kinsey's work in the 1940s and his publication of the *Sexual Behavior in the Human Male*, which documented widespread experience with same-sex behavior among men. Kinsey's work disputed prevailing ideas that all adults were either homosexual or heterosexual and demonstrated instead that many adult men were sexually fluid (D'Emilio, 1983). By documenting the breadth of sexual variation, Kinsey's work aided gays in their efforts to describe themselves as something more than a very small minority with highly unusual behaviors.

This developing subculture provided a fertile ground for the first gay and lesbian activist organizations in the United States; homophile organizations devoted to educating homosexual men and women and engaging in small-scale social actions. In 1951, gay men founded the Mattachine Society in Los Angeles. The secret cell-like structure of the homophile organization was modeled after Communist organizations and allowed branches of the organization to be founded across the country yet operate in secrecy even from each other (D'Emilio, 1983). The lesbian homophile organization Daughter of Bilitis (DOB) was founded in the mid-1950s by Del Martin and Phyllis Lyon in San Francisco, and by the end of the 1950s, there were several DOB chapters across the country (Faderman, 1991). Both Mattachine and DOB published their own magazines (*ONE* magazine and *The Ladder*, respectively), and engaged in small-scale social actions, including protesting police entrapment in Los Angeles in 1952. During the more liberal 1960s, homophile organizations became more visible and confrontational. The visibility and growth of other movements such as the civil rights movement and student movement also buoyed homophile activism and enforced the viability of different tactics. For example, after the Homophile League of New York picketed a military induction center in the early 1960s, other homophile groups engaged in picketing in front of the White House and Pentagon (Faderman, 1991).

This increasing visibility included the involvement of gay men and lesbians in other social movements. For example, Bayard Rustin, a gay African-American man, was the lead organizer of the 1963 March on Washington for Jobs and Freedom. Lesbians were involved in the growing second-wave feminist movement, which gave them frames to interpret their relationship to gay homophile activists. The shift of lesbian organizing to feminist organizations led to the dissolution of the DOB in 1970 due to conflicts over feminist radicalism (D'Emilio, 1983).

The increasing militantism included riots and visible protest in US cities in the late 1960s, including rejecting police harassment at the New Year's Ball Raid in 1965 in San Francisco, at the Compton's Cafeteria in August 1966 in San Francisco, and at the Black Cat bar in Los Angeles in 1967 (Armstrong and Crage, 2006). The 'turning point' of this militantism is often described as the riot in response to the raid on the Stonewall Inn in Greenwich Village in the early hours of 29 June 1969. Gay bar patrons fought back against the police raid in an event that was commemorated and memorialized as the beginning of the contemporary gay and lesbian movement in the United States.

Gay Liberation and Lesbian Feminism

Although the growth of the gay liberation movement is often attributed to the riot at the Stonewall Inn, the strength of the movement probably resulted from the dramatic rise in radicalism in the late 1960s. Several movements including the civil rights movement and antiwar movement radicalized; this radicalism included growing discontent at institutions, expression of revolutionary sentiment, and cultural radicalism. Cultural radicalism included not only separatism but also the growth of the counterculture, which endorsed alternative gender and sexual norms for men and women.

The gay liberation movement developed quickly after Stonewall. The first gay pride march commemorated the riots in June 1970 and included thousands of men and women marching from Greenwich Village to Central Park (D'Emilio, 1983). Groups such as the Gay Liberation Front and Street Transvestite Activist Revolutionaries were established in New York City and San Francisco, but gay liberation organizations quickly spread across the country. This 'liberation moment' in the gay/lesbian movement included a new emphasis on 'coming out' to the public and on gay cultural pride, which was part of the influence of identity politics movements on gay liberation (Armstrong, 2002). Early gay liberationists were involved in other movements' actions along with taking to the street in more visible and radical forms of activism around issues like the criminalization and medicalization of homosexuality. This militantism was buoyed by the assassination of Harvey Milk, the first openly gay person to be elected to public office in California.

The politics of gay liberation in the United States was quickly beset by ideological differences, including a divide between revolutionary ideologies, redistributive political ideologies, and single-interest politics (Armstrong, 2002). These ideological differences materialized in group conflicts over whether or not to support Black Panther activities, which led to permanent splits and the founding of the more moderate single-issue group the Gay Activists Alliance. The movement also faced ongoing disputes about the inclusion of transsexuals and drag queens, along with the capacity of the groups to address sexism and racism. By the early 1970s, the modern gay movement struggled to include diversity within the community; although the ideological emphasis of the movement was 'unity through diversity,' this diversity was often limited to sexual and ideological differences between white gay men (Armstrong, 2002).

During this time, many lesbians became involved in lesbian feminism rather than gay liberation. Lesbians formed lesbian feminist organizations, including the Radicalesbians in New York, the Furies Collective in Washington, DC, and the Gay Women's Liberation in San Francisco (D'Emilio, 1983). In the essays 'The Woman-Identified Woman,' the New York Radicalesbians attested that "A lesbian is the rage of all women condensed to the point of explosion." Many lesbians focused on building separate women-only communities rather than engaging with male-dominated gay liberation, although by the end of the 1970s some lesbians had pushed for more inclusion in the gay liberation movement, which was increasingly referred to as the 'gay and lesbian movement' (Armstrong, 2002). Toward the end of the 1970s, there was also a sudden

growth in gay and lesbian organizations for racial minorities, particularly women's organizations (Armstrong, 2002).

Accomplishments of this burgeoning movement included the removal of homosexuality as a mental illness from the Diagnostic and Statistical Manual of Mental Disorders in 1973 by the American Psychiatric Association and inclusion of gay rights in the 1980 Democratic Party platform. Several national gay and lesbian organizations were also formed. Gay lawyer Bill Thom founded Lambda Legal Defense and Education Fund in 1972 and had to sue for the right of the organization to exist when New York courts denied the need for a legal voluntary organization for homosexuals (Anderson, 2005). Former GAA activists founded the National Gay Task Force in 1973. The Gay Rights National Lobby was created in 1978 and turned into the Human Rights Campaign Fund (now HRC or the Human Rights Campaign) in 1980.

In 1972, lesbian and gay college students in East Lansing, Michigan, helped pass the first local nondiscrimination ordinance that included sexual orientation. For the 14 local gay rights ordinances passed in cities and municipalities across the country between 1972 and 1974, there was little organized opposition, beyond routine objections from conservative community members or the Catholic Church (Button et al., 1997). These separate activities were organized into a national movement after the 1979 National March on Washington for Gay and Lesbian Rights (Ghaziani, 2008).

Although there were successes in the 1970s, there was also the beginning of an antigay backlash. When Dade County Commissioners passed a local nondiscrimination law that included protections for discrimination based on 'sexual preference,' it was the beginning of the first antigay backlash and national conversation about gay rights. The convergence of antigay, prolife, and antifeminist organizing led to the entry of evangelicals into the public life, creating a fierce countermovement in the form of the Religious Right that would fight with the gay and lesbian movement for several decades. In the mid-1970s, organized antigay activism grew out of small, localized social networks that ran through Christian churches and radio and television shows. Anita Bryant was a former Miss America contestant and spokesperson for Florida orange juice. She mobilized her followers in opposition to the newly passed nondiscrimination ordinance in Dade County, Florida, and her role in the opposition received national visibility. Bryant's campaign persuaded voters with language about religious rights and influence on children, implying that all gay men were pedophiles and looking to recruit children. The Dade County referendum in 1977 was the beginning of the first wave of antigay activism, a series of referendum campaigns that spread to cities across the country (Fetner, 2008; Stone, 2012).

HIV/AIDS

The growing momentum of the gay and lesbian movement was dramatically transformed by the discovery of the first diagnosed cases of HIV/AIDS in gay men in New York, Los Angeles, and San Francisco in 1981. Lacking a treatment for AIDS in those initial years, AIDS transformed the gay and lesbian movement organizationally and demographically. Movement resources, personnel, and leadership shifted into health-based organizations, which were often more bureaucratic and connected to the state than previous gay and lesbian activism had been (Armstrong, 2002). These organizations included early HIV/AIDS organizations, such as the Gay Men's Health Crisis in New York City. Some organizations and events were created to memorialize the growing number of people who had died from AIDS. Starting in 1985, longtime gay activist Cleve Jones also organized the NAMES Project AIDS Memorial Quilt, a patchwork of quilt squares in memorialization of over 90 000 people who died of complications due to AIDS. Events like candlelight vigils and Walk for Life memorialized and raised money for HIV/AIDS.

Growing anger toward the inaction of the Reagan administration on AIDS and the Supreme Court decision supporting sodomy laws (*Bowers v. Hardwick*, 1986) also led to a surge in gay militancy and the development of direct action politics, including the rise of AIDS Coalition to Unleash Power (ACT UP) in March 1987 (Gould, 2009). ACT UP targeted the funding of AIDS medical research, availability of drugs for treatment, and media depictions of AIDS.

According to some scholars, an additional consequence of HIV/AIDS was the growing focus on same-sex relationships and relationship recognition rights. The lack of relationship recognition for gay couples navigating the medical system due to HIV/AIDS and the growing lesbian baby boom in the 1990s led to greater attention to relationship recognition within the gay and lesbian movement and eventually the focus on same-sex marriage (Chauncey, 2004).

The Religious Right and Culture War 1990s

The gay 1990s included increasing visibility in the mass media, including the coming out of Ellen DeGeneres on her television show *Ellen* in 1997 and increasing inclusion of gay and lesbian television characters. Political recognition also grew, with the passage of new progay laws in state legislatures and Bill Clinton, the first presidential candidate to court the gay and lesbian vote (Ghaziani, 2008).

However, this visibility was coupled with a backlash against gay and lesbian rights. The 1992 Republican Convention sparked a culture war when the Republican profamily platform included a multipronged attack on LGBT rights in legislation, public education, and the arts. This backlash against the gay and lesbian movement played out in state legislatures and ballot boxes, with an increase in both progay and antigay state laws and ballot measures (Ghaziani, 2008). This convention was followed quickly by two of the most draconian antigay initiatives in Oregon and Colorado. In Oregon, the Oregon Citizens Alliance led by vocal antigay activist Lon Mabon sponsored a statewide initiative in November 1992, Ballot Measure 9. This was a broad, moralistic initiative that eliminated future and existing gay rights legislation, paralleled homosexuality with sadomasochism and pedophilia, and prohibited government promotion of homosexuality. In the same election, Colorado for Family Values sponsored Amendment 2, a broad legalistic initiative that eliminated future and existing gay rights laws in the state. They also developed a 'special rights' language that pitted African-Americans against gay rights and appealed to moderate voters. This

intensity continued until 1996. Between 1992 and 1996, antigay activists attempted antigay initiatives in 13 states and more than 30 cities and towns (Stone, 2012). In conjunction with this legal backlash, the exgay movement grew in the 1990s and became more public with its assertions that gay and lesbian individuals could be transformed into heterosexuals.

Some of this backlash was a result of LGBT legal gains. After the 1993 Hawaii Supreme Court case *Baehr v. Lewin* opened the door for same-sex marriage in Hawaii, legislators passed a federal Defense of Marriage Act (DOMA) in 1996 and a series of statewide statutory laws that reinforced DOMA. DOMA created a federal definition of marriage as between a man and woman, along with permitting states to ignore same-sex marriage performed in other states. In a few states, legislators referred these DOMAs to the ballot box to be overwhelmingly passed by voters and written into the state constitution (Stone, 2012). The 1993 "Don't Ask, Don't Tell" policy about gay and lesbian inclusion in the military was also a response to pressure from President Bill Clinton to allow gay and lesbian service people to openly serve in the military. However, some legal gains by the gay and lesbian movement were a result of ending antigay laws; specifically, the 1996 Supreme Court case *Romer v. Evans* overturned Colorado Amendment 2 (Fetner, 2008).

The 1990s was also a time period of political change for the lesbian and gay movement. National organizations grew larger and more involved in local and statewide activism (Stone, 2012). The movement grew increasingly institutionalized and focused on mainstream politics. Activism also became more specialized, as more groups developed for racial minorities, bisexual, and transgender individuals within the movement. The growth of bisexual and transgender organizations throughout the 1980s and 1990s led to pressure by bisexual and transgender organizers to be included in the mission and organization of the gay and lesbian movement, along with being included in the 1993 March on Washington (Ghaziani, 2008). The gay and lesbian movement turned into the LGBT movement during the 1990s, although bisexual and transgender individuals and issues continued to be marginalized within the movement.

The close of the 1990s saw two dramatic events: the legalization of same-sex civil unions in Vermont in 1999 and the murder of Matthew Shepard, a young gay man in Laramie, Wyoming.

Same-Sex Marriage and the Contemporary LGBT Movement

Although same-sex marriage was not a central goal of the LGBT movement in the 1970s, at the turn of the century it quickly came to the center of the movement agenda both in the United States and in many other industrialized nations. The movement for same-sex marriage is part of an international movement. The first nation to legalize same-sex marriage was the Netherlands in 2000, followed by over a dozen countries in Western Europe, North America, and South America, along with the nation of South Africa. By the start of 2013, in the United States, nine states and the District of Columbia had legalized same-sex marriage, and four additional states recognized same-sex marriages performed in other states.

Although contested by some activists within the LGBT movement, the push for same-sex marriage has been at the center of the LGBT movement agenda since the turn of the century. Some of this momentum has been motivated by organizers like Evan Wolfson, who was involved in Hawaii *Baehr v. Lewin* and went on to found The Freedom to Marry Coalition in 2003 to advocate for same-sex marriage. The 2003 Massachusetts Supreme Judicial Court ruling in *Goodridge v. Department of Public Health* that established the right of same-sex couples in Massachusetts to marry was the beginning of a backlash against same-sex marriage that diverted significant LGBT movement resources to the legalization of same-sex marriage.

Most of this movement energy was spent fighting antigay ballot measures that would ban same-sex marriage. Same-sex marriage has been voted on by the American public in a series of ballots measures across the country since 1998. Thirty of these ballot measures were legislative-referred initiatives to revise the state constitution to ban same-sex marriages. These marriage bans have passed overwhelmingly at the ballot box; they have been defeated only twice in Arizona (2006) and Minnesota (2012) (Stone, 2012). The most contentious ballot measure to date has been California Proposition 8, the only marriage ban to overturn legalized marriage. Proposition 8 countered a 2008 decision by the California Supreme Court to allow same-sex marriage. Between the California Supreme Court case and the campaign, more than 18 000 same-sex marriages were conducted in the state of California. The LGBT campaign to fight Proposition 8 and the Religious Right campaign raised over $40 million dollars, shattering campaign spending records. When Proposition 8 passed in the November 2012 election, there were protests across the country, and the marriage ban was contested in the Supreme Court case *Hollingsworth v. Perry*. Although these ballot measures have been a successful strategy for the Religious Right, the LGBT movement experienced its first winning streak in the November 2012 election, when voters in Minnesota rejected a marriage ban and voters in Maryland and Washington affirmed same-sex marriage laws that had been recently passed by their legislatures. Maine voters also passed same-sex marriage through an initiative that was proposed by LGBT organizers.

Marriage was not the only goal. There were many other LGBT movement accomplishments during this time period. For example, the overturn of sodomy laws in the Supreme Court case *Lawrence v. Texas* (2003) struck down existing sodomy laws in 14 US states. In October 2009, President Barack Obama signed the Matthew Shepard and James Byrd, Jr. Hate Crimes Prevention Act, which covered hate crime acts for both sexual orientation and gender identity. Later, President Obama repealed the ban on gay and lesbian service members, although transgender service members are still banned from the US military.

There were unsuccessful attempts to pass a federal Employment Non-Discrimination Act, which has been ongoing since the 1970s. However, there was a dramatic growth in nondiscrimination laws at the local level. As of 2007, over half of the US population was covered by a state, county, and/or city law against discrimination based on sexual orientation, whereas in 1990, less than a third of the population was covered by such laws. An increasing number of state and city laws also included

gender identity and expression, which offers more protection for transgender individuals (Stone, 2012). The passage of these local laws was a consequence of growing support in public opinion for LGBT rights, strength in local movement organizations, and assistance by national organizations.

The LGBT movement also increasingly focused attention on intersectional issues, such as immigration reform and social security. Transgender organizing and media visibility grew, along with the first national transgender rights organization, the National Center for Transgender Equality, which was established in 2003.

Future of LGBT Movement

Since the earliest homophile organizing at the end of World War II, the US LGBT movement has grown into a major social movement that has transformed homosexuality from a marginalized mental illness to a recognized minority group. The centrality of same-sex marriage to the LGBT movement has not only given the movement focus but also may diminish the significance of the movement after the legalization of same-sex marriage. It may be more difficult to sustain momentum in the LGBT movement once the goal of same-sex marriage has been reached. The centrality of same-sex marriage to the movement also has led to criticisms from scholars and activists alike that the LGBT movement has become more homonormative or focused on the integration of heteronormative ideals such as binary gender roles, marriage, and procreation into LGBT culture. This homonormativity may be contrary to the growing emphasis on diversity and inclusiveness in the LGBT movement, thus creating more divisions within the movement.

See also: Feminist Theory: Radical Lesbian; Gay Men's Emancipation Movement in Germany: 1897–1933; Gay and Bisexual Latinos in US Gay Culture; Heterosexism and Homophobia; Lesbian and Gay Activism in Latin America; Lesbian, Gay, Bisexual and Transgender Aging; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Lesbians: Social and Economic Situation; Queer Anarchism; Queer Geography; Sexual Minorities and Sexual Citizenship; Sexual Minorities in China; Sexual Minorities in Modern Russia; Sexual Minorities in South Africa; Sexual Minorities in the Maghreb; Sexual Minority Families of Choice and Family Networks: We Are Family; Sexual Orientation: Historical and Social Construction; Social Movements: A Social Psychological Perspective; Social Protest; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Adam, Barry, 1987. The Rise of a Gay and Lesbian Movement. Twayne Publishers, Boston.

Anderson, Ellen A., 2005. Out of the Closets and into the Courts: Legal Opportunity Structure and Gay Rights Legislation. University of Michigan, Ann Arbor.

Armstrong, Elizabeth A., 2002. Forging Gay Identities: Organizing Sexuality in San Francisco, 1950–1994. University of Chicago Press, Chicago.

Armstrong, Elizabeth A., Crage, Suzanna M., 2006. Movements and memory: the making of the Stonewall myth. American Sociological Review 71 (5), 724–751.

Bérubé, Allan, 1990. Coming Out Under Fire: The History of Gay Men and Women in World War Two. Free Press, New York.

Button, James W., Rienzo, Barbara A., Wald, Kenneth D., 1997. Private Lives, Public Conflicts. Congressional Quarterly Inc., Washington, DC.

Chauncey, George, 1994. Gay New York. Basic Books, New York.

Chauncey, George, 2004. Why Marriage? The History Shaping Today's Debate Over Gay Equality. Basic Books, New York.

D'Emilio, John, 1983. Sexual Politics, Sexual Community. University of Chicago Press, Chicago.

Faderman, Lillian, 1991. Odd Girls and Twilight Lovers: A History of Lesbian Life in Twentieth-Century America. Columbia University Press, New York.

Fetner, Tina, 2008. How the Religious Right Shaped Lesbian and Gay Activism. University of Minnesota Press, Minneapolis.

Ghaziani, Amin, 2008. The Dividends of Dissent: How Conflict and Culture Work in Lesbian and Gay Marches on Washington. University of Chicago Press, Chicago.

Gould, Deborah B., 2009. Moving Politics: Emotion and ACT UP's Fight against AIDS. University of Chicago Press, Chicago.

Kennedy, Elizabeth, Davis, Madeline, 1993. Boots of Leather, Slippers of Gold: The History of a Lesbian Community. Routledge, New York.

Stone, Amy L., 2012. Gay Rights at the Ballot Box. University of Minnesota Press, Minneapolis.

Relevant Websites

http://www.freedomtomarry.org/ – Freedom to Marry.
http://www.hrc.org/ – Human Rights Campaign.
http://www.lgbtmap.org/ – Movement Advancement Project.
http://transequality.org/ – National Center for Transgender Equality.
http://www.ngltf.org/ – The National Gay and Lesbian Task Force.
http://williamsinstitute.law.ucla.edu/ – Williams Institute.

Gay and Lesbian Tourism

Kevin Markwell, Southern Cross University, NSW, Australia
Gordon Waitt, University of Wollongong, Wollongong, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines gay and lesbian tourism and its relationship with gay and lesbian cultures and identities. First, an explanation is provided for why travel for some people became an integral part of their sexuality from the Victorian era through to the present day. Second, a critical overview is given of the role of the gay and lesbian tourism industry in producing and reproducing gay and lesbian identities with a special focus on events and festivals. Finally, the concept of risk (cultural risk and medical risk) is explored within the context of gay and lesbian travel and tourism.

Introduction

The gay and lesbian tourism industry sells travel products, such as tours, cruises, accommodation, or packages to attend events and festivals, specifically for same-sex attracted people. The niche lesbian, gay, bisexual, and transgender (LGBT) market is frequently positioned by industry commentators in economic terms as 'fast growing' and 'resilient.' Out Now Consulting, based in the Netherlands, estimated the income generated by the global gay and lesbian tourism industry at around US $165 billion per year (cited in UNWTO, 2012), while the United Nations World Tourism Organization (2012) estimated that the United States' gay and lesbian tourism market is alone worth US$65 billion. Community Marketing Incorporated (2009) noted the resilience of the gay and lesbian tourism market in the United States by highlighting that the number of holidays taken by American gay and lesbians remained unaffected by the global financial crisis in 2008.

Despite the financial strength of the gay and lesbian tourism market, a focus on market indicators alone masks a more nuanced understanding of this form of tourism. The starting point of market-led interpretations of the notion 'gay and lesbian tourism' is the assumption of stable, recognizable, and preconfigured 'gay' and 'lesbian' identities that comprise a homogenous market sector. Instead, our starting point is that participation in tourism contributes to the production and reproduction of gay and lesbian identities. In doing so, the social practices of purchasing and consuming products and services that comprise the gay and lesbian tourism industry is inevitably exclusionary as well as inclusionary (Waitt and Markwell, 2006; Casey, 2009; Binnie and Klesse, 2011). Statements about gay and lesbian tourism as a homogenous market category are therefore quite misleading. Indeed, many same-sex-attracted tourists may not identify particularly comfortably with a 'gay' or 'lesbian' identity for a range of reasons relating to ethnicity, culture, sociodemographic status, or religious affiliation.

To critically explore the gay and lesbian tourism industry, this article is divided into four sections. The first section charts the emergence of Western gay and lesbian tourism. This section points to how people with same-sex desires, and who had the financial and social capital, and the leisure time, traveled to destinations that offered some form of homosocial or homoerotic 'utopia,' long before the emergence of a fledgling, market-led gay and lesbian tourism industry in the late 1970s/early 1980s. The second investigates the relationship between tourism, consumption, and sexuality. Particular attention is drawn to how market relations inherent in gay and lesbian tourism enable particular understandings of sexuality to be produced and reproduced. The third section investigates the roles of festivals and events within gay and lesbian tourism. Here, attention is drawn to recent efforts by some festival organizers to unsettle the narrow definitions of 'gay' from market-led definitions or presentations of sexuality. The final section examines the notions of 'cultural' and 'medical' risk.

Emergence of Gay and Lesbian Tourism

In the nineteenth century, travel was integral to how affluent Western same-sex-attracted men and women explored their sexuality, often in spite of, rather than because of, an organized and integrated tourism industry. Why overseas travel became aligned with sexuality at this time can be understood in terms of how middle-class ideas of gender, race, religion, respectability, and morality became tied to reproductive sex and the nuclear family. The figure of the homosexual emerged in British Victorian society as a counter to the respectable, upstanding, Christian, heterosexual subject (Pronger, 1990). If the heterosexual subject was the keystone to a respectable, productive society underpinned by the nuclear family, the homosexual emerged as an 'illness,' an 'evil,' and a 'threat' to society. Laws were enacted to criminalize sodomy. Travel elsewhere for many same-sex-attracted men and women can thus be understood as a need to escape, albeit temporarily, from the heterosexist pressures and assumptions operating in Victorian society such as marriage and nuclear family life.

Due to there being more documented examples of the travel patterns of men in the nineteenth century, this section will focus on the growth of homosexual/gay male travel. Popular destinations at this time for same-sex-attracted, affluent British men included southern Italy (including Capri, Taormina, Palermo, and Naples), Spain, and northern Africa. In North America, affluent men in particular traveled to destinations such as Mexico, Haiti, and Cuba. Driving this 'north' to 'south' travel were ideas about the relationship between geography

and sexuality circulated by a form of neo-Darwinism and in the works of artists, writers, photographers, and other culture workers (Aldrich, 1993). Neo-Darwinist accounts of humanity described the people living in the tropics (and, for that matter, in Eastern Europe, Asia, and Africa) as sexually adventurous, promiscuous, with heightened libidos, and morally inferior to those living in Western Europe and North America (Waitt and Markwell, 2006). Such beliefs created and sustained a considerable level of demand among affluent Western men, regardless of their sexuality, to travel 'south' to seemingly less 'civilized' places and societies, where sex acts were not bound by European moral codes. Layered over this nineteenth century classed and racialized geography of 'sexual paradise' was the work of artists and writers that included August von Platen, Hans Christian Anderson, Andre Gide, and Wilhelm von Gloeden (Aldrich, 1993; Waitt and Markwell, 2006). These cultural workers were central to circulating ideas that configured places as 'homosexual paradises.' Imagined as utopias, these destinations were positioned as offering 'freedoms' from persecution and interference for men seeking sexual and social intimacy with other men. These quests had considerable resonance for affluent men whose daily lives involved considerable self-discipline to conceal their same-sex desires in order to avoid punishment by the law at the hands of their fellow citizens. For men whose lives were characterized by alienation, estrangement, or fear of being 'caught out,' these temporary sojourns may have been experienced in terms of personal freedom, and importantly, opportunities to become themselves. Helping creates the reputations of some destinations as a kind of utopia for same-sex-attracted men was the combination of less state surveillance in those locations, or a willingness to 'turn a blind eye,' and a marked disparity of wealth between the visitors and those visited, which may have facilitated sexual contact.

While, no doubt, there would have been knowledge passed by word of mouth about places safest to stay and socialize, these men did not have the benefit of a gay tourism industry to help facilitate their travel. Neither were there recognizable gay hotels nor were there gay bars or cafes. Men socialized at the homes of expatriates, inviting residents to join them, or frequented bars, restaurants, and cafes that were perhaps known to be tolerant of men who enjoyed the company of other men. More often than not, these men were alert to the possibility of being exposed, and subsequently disciplined their behavior and appearance to conform to and reproduce particular expectations of masculinity. These men visited and enjoyed these destinations facilitated through an informal network of relationships, information transmission, and situated knowledge provided by expatriates, rather than the bourgeoning tourism industry of the Victorian era. This was to emerge much later.

Following the First World War, greater acceptance of sexual diversity began to show in Western Europe, especially in the cities of Paris and Berlin, and, somewhat later, Amsterdam. While most European nations had laws that criminalized homosexual activity, specifically sodomy, in the nineteenth and the first half of the twentieth centuries, homosexuality began to be discussed, at least within intellectual discourses, in a more socially accepting way. The writings of Sigmund Freud and Magnus Hirschfield contributed to a more enlightened view of human sexuality, which embraced sexual difference such as homosexuality. During the 1920s and 1930s, Berlin became an important destination for homosexual men, and, as Vyras (1997: p. 24) has noted, "Berlin had become the destination, a kind of pilgrimage for homosexuals all over Europe." The British author, Christopher Ishwerwood, provides insights to how he negotiated his sexuality through the commercial venues that emerged in Berlin during the 1930s (see, for example, Isherwood, 1945). Homosexual 'subcultures' were able to flourish in Berlin and Paris through a combination of community-based political and social organizations together with commercial venues such as cafes and restaurants. Sibalis (1999: p. 25) commented that in Paris by the early 1900s, "a growing commercial world that catered to homosexual men comprising bathhouses, bars and restaurants had emerged." This mix of community-based structures and commercial venues had also begun to emerge in a number of North American centers as well, notably San Francisco and New York. Throughout the latter part of the nineteenth century and into the twentieth century, San Francisco, and particularly the Barbary Coast neighborhood, had a reputation for licentious entertainment and a growing tourist trade (Boyd, 2011; Waitt and Markwell, 2006). While European cities like Berlin and Paris became an integral travel destination for men seeking sexual relationship with men, in the early twentieth century, the imagined geography of the 'south' retained its pull. The Indonesian island of Bali, in the early twentieth century, became known as a 'homosexual paradise' (Vickers, 1989: p. 106), while central and South America, particularly the islands of Cuba and Haiti, and countries such as Brazil, attracted waves of homosexual men, who searched for, and found, sexual opportunities and freedoms that were largely unattainable in their homelands (Bleys, 1996).

From the 1960s to the 1970s, travel guidebooks such as *Damron's Men's Travel Guide* (first published as *Bob Damron's Address Book* in 1964) and *Spartacus International Gay Guide* (first published in 1971) targeted same-sex-attracted men. The publications chart the emergence of a commercial gay tourism industry because these guidebooks listed not only destinations but also the various components that comprise the gay tourism industry such as gay-operated or 'gay-friendly' hotels and resorts, cafes, restaurants, night clubs, discos, sex on premises venues, as well as 'gay beaches' and cruising areas. With the assistance of these guidebooks, men could now plan their holidays to destinations with the assistance of the hospitality infrastructure provided by the emerging gay tourism industry. These guidebooks are evidence of a shift from an unstructured and rather amorphous assemblage of operators and destinations to an integrated and commercial gay tourism industry (Waitt and Markwell, 2006). Magazines targeted at the gay male market such as *The Advocate* in the United States and *Gay Times* in the United Kingdom also began featuring travel sections and regular travel columns from the late 1960s onward.

By the early 1980s, the gay tourism industry started to become organized both internationally and nationally. In 1983, the International Gay and Lesbian Travel Association (IGLTA) was established (initially as the International Gay Travel Association) and today it is an association with more than 2000 member businesses from over 80 countries. There

are also nationally based gay and lesbian tourism organizations in countries such as Australia and Canada and regional organizations such as the Gay European Tourism Association. These organizations occupy an important position within the gay and lesbian tourism sector, acting as a vehicle to promote the interests of their members, lobbying government relating to policy matters, and raising the visibility of the gay and lesbian sector within the broader tourism industry. With the advent of the World Wide Web, a huge amount of gay and lesbian tourism and travel content can be found online. Indeed, a simple Google search using the term 'gay tourism' revealed 59 million results (compared with 'wildlife tourism,' 36 million; 'religious tourism,' 36 million; and 'cultural tourism,' 111 million).

Many businesses today describe themselves as 'gay friendly.' Indeed, many cities and towns also promote themselves using the same descriptor. Hodes et al. (2007) provide one of the few studies that have sought to understand tourists' conceptions of 'gay friendly,' in the context of Amsterdam. This work suggests that 'gay friendly' was not strongly attributed to material infrastructure such as the presence of hotels, bars, and nightclubs, but instead, their results indicated that 'gay friendly' is an amalgam of mostly intangible or embodied aspects of a destination, such as a feeling of being welcomed and accepted.

Nevertheless, as Boyd (2008: p. 226) suggests, the term has a rather 'amorphous quality.' At one level it seems apparent that 'gay friendly' means exhibiting genuine friendliness, acceptance, and hospitality to gay (and lesbian) people by an establishment that is owned and largely operated by 'nongay' people. This definition positions a 'gay-friendly' business as one that is not necessarily owned and operated directly by gay or lesbian people but rather by individuals or a company aiming to create a welcoming business where people do not have to feel afraid to disclose their sexuality. But to what depths does this friendliness extend? Is gay friendliness conditional on a reasonably narrow definition of 'gay'? Does gay friendly imply complete lack of discrimination or prejudice based on sexual identity? Within the context of gay and lesbian tourism, gay friendly suggests a welcoming approach to gay and lesbian tourists who are treated with the same dignity and respect afforded to heterosexual people. However, if a hotel proclaimed itself to be 'African-American friendly' or 'Chinese friendly,' it would surely be the subject of scorn, for a hotel would be expected to treat all cultures and ethnicities with equal respect. The notion of 'gay friendliness' then only really makes sense within a heteronormative framework, which positions nonheterosexual identities as 'other.' And how is a business' 'gay friendliness' to be measured or assessed? Should staff be trained in understanding the needs of gay and lesbian travelers, for example, before a hotel can call itself 'gay friendly'? Should there be some kind of accrediting process, similar to that which exists within the ecotourism sector, so that consumers can be reassured that a business has met certain standards of 'gay friendliness'? Gay friendly, as part of the smorgasbord of the tourism industry's marketing tools, is perhaps best understood as part of a neoliberal economic logic, where it is acceptable to be gay or lesbian so long as one remains within the boundaries of acceptable behavior and presentation of self that are proscribed by the heterosexual majority.

Tourism and Identity

Three strands of thought have emerged on the relationship between tourism and sexual identity. One strand of literature addressed this question by focusing on how travel for pleasure helps to make and remake a particular version of a 'gay' and 'lesbian' identity. Hughes (1997) positions recreational travel as a basic component of becoming gay and lesbian. Indeed, Hughes (2006: p. 83) suggests that "holidays have come to be one of the defining characteristics of what it means to be gay or lesbian." Most gay men do not have the responsibility of children (notwithstanding the increasing number of gay and lesbian couples who are indeed having children and also the relatively larger proportion of lesbians who have children as part of previous heterosexual relationships) and many (but certainly not all) gay men and lesbians have sufficient disposable income to allow them to enjoy more frequent recreational travel. It may well be that greater mobility is a characteristic of gay and lesbian cultures but this mobility depends on having sufficient disposable income and time. Rather than searching for difference, gay and lesbian tourists might, in part at least, be looking for sameness. Traveling to destinations with overt gay and lesbian precincts marked out with rainbow flags, retail outlets selling gay merchandise, and cultural expressions such as festivals and special events provide possibilities to validate a particular understanding of a gay and lesbian identity (Waitt and Markwell, 2006). Travel and tourism often facilitate the production and reproduction of a gay or lesbian identity that is constructed out of the interplay of the gay tourism economy with shared understandings of what it is to be gay or lesbian from ideas surrounding Western gay liberation history and politics.

A second strand of literature is critical of how the tourism industry narrows understandings of sexuality around particular versions of being 'gay' or 'lesbian.' The commodification of gay and lesbian identity through consumption practices tends to produce an exemplary gay and lesbian consumer who comes to be seen as representative of gay and lesbian identities. The emergence of a 'globalized gay consumer' (Waitt and Markwell, 2006: p. 248) excludes those who do not fit this subject identity because of such things as socioeconomic circumstances, ethnicity, or body shape. The images most frequently portrayed in gay tourism marketing collateral are normally male, young, white, and conventionally good looking. As Casey (2009: p. 159) notes, "The aged, disabled, fat or poor gay man has no place in travel imagery, advertisements or the destinations to be visited." Burns (2012: p. 316) makes the point that gay and lesbian tourism thus becomes a means for sustaining a 'global sexual citizenship' through a series of consumptive practices that privilege a very narrow conception of the gay or lesbian citizen – one that is "almost always a white western citizen-subject who has access to 'difference' through urban living, global travel and through his or her investment in the project of global queer world-making." Binnie and Klesse (2011: p. 159) advance this argument by noting that "while the right to consume and have one's identity recognized within tourist and city marketing can be seen as a marker of inclusion, this inclusion is often tied to a narrow construction of gay identity that is bought at the cost of exclusion of those who dis-identify with the commercial scene." Further, 'gay' or 'lesbian' identities

have emerged from within Western society, and although these identities have diffused to other non-Western nations, the degree to which they 'fit' with local understandings of sexuality and sexual identity is problematic, and can create difficulties for people trying to accommodate a (Western) gay or lesbian identity within their own cultures.

A third strand of literature has positioned the purchasing power of gay and lesbian tourists as leverage to achieve sexual equality. The intertwining of gay civil rights and the market is made very clear in the World Trade Organization's report on LGBT tourism, which states that "With a view to global tourism, many countries are beginning to realise that human rights considerations have also a positive effect on their economies and to actively promote LGBT tourism as a consequence" (Careaga and Sabbadini, 2012: p. 13). Because of the perceived economic impact of gay and lesbian tourism, organizations such as the IGLTA can use the purchasing power of its members (and by extension, the greater gay and lesbian market) to impact on government policies. Boycotts can be organized, either through industry organizations or by individuals using social media, through which gay and lesbian consumers can express their anger at governments or businesses that are perceived to discriminate against gay and lesbian customers. For instance, in 2004, activists in the Australian state of Tasmania urged gays and lesbians and their supporters to boycott Tasmanian gourmet food products because of that state's then antigay laws. However, gay activists today are urging Australian consumers to buy these very same products because of the current Tasmanian government's positive stance on gay marriage (Tasmanian Gay and Lesbian Legal Rights Group, 2012). Unfortunately, what is often overlooked in these arguments is the important question of whose sexual politics are advanced by the consumption practices of the gay and lesbian tourism industry.

Special Events and Festivals

Gay and lesbian tourism is powerfully driven by special events and festivals. Such events range in scale from international events hosted in 'world cities,' such as the International Gay Games and the World Pride event through to annual pride parades and festivals in regional centers – like Hamilton Pride, Waikato, New Zealand. Such events bring LGBT people together to collectively and publicly celebrate gay and lesbian cultures, reaffirm identity, and create enjoyment and, often, and importantly, public spectacle. It is the 'conspicuous spectacle' dimension of so many contemporary gay and lesbian events and festivals that differentiate them from earlier events that were often hidden and only known to 'those in the know.' Today, gay and lesbian communities create and host a diverse array of special events and festivals, which not only provide platforms for the production and reproduction of diverse sexualities but also frequently challenge heteronormative assumptions and expectations, and, at the same time, also attract a large nongay audience.

Perhaps, the most popularly known of all gay and lesbian events are the various pride parades and associated festivals. These events "challenge the production of everyday spaces as heterosexual" (Valentine, 1996: p. 152) through their use of public spaces (streets and roads) as a platform upon which to make visible gay, lesbian, and other nonheterosexual subjectivities. To give an indication of the scale of some of these events, it is worth noting that the 2011 Sao Paolo pride parade in Brazil was estimated to have attracted an astonishing "four million participants which included 650 000 tourists of which 15 000 were international tourists" (De Britto Lomanto, 2012: p. 19).

Many, if not most, of these events have become incorporated, to a greater or lesser extent, into the gay and lesbian tourism industry and indeed may serve as a gay and lesbian 'anchor attraction' for particular destinations. For example, the Sydney Gay and Lesbian Mardi Gras (SGLMG) has played a crucial role in helping to develop an identity for Sydney, Australia, as a welcoming and sophisticated gay and lesbian tourist destination (Markwell, 2002). Yet, its origins had little if anything to do with tourism, but instead was part of an ongoing political response to systemic discrimination that was being directed toward the lesbian, gay, bisexual, and transgendered citizens of New South Wales and indeed, Australia as a whole. Increased media coverage and a desire by LGBT citizens from other cities to participate in the parade or to witness the parade first hand led to the parade's growing intersection with the tourism economy. The parade, and the 2- to 3-week festival that developed, which is held prior to the parade, has become an important, and indeed, defining component of the gay and lesbian tourism industry operating not only within Sydney but also nationally. It was the international success of the SGLMG, which no doubt played a pivotal role in Sydney being awarded the International Gay Games that took place in 2002, 2 years after the Olympic Games were held in that city.

Gay and lesbian events and festivals are often critiqued as becoming somewhat "sanitized representations of queer identities and ways of living" (Binnie and Skeggs, 2006). This critique is mounted particularly for state- and corporate-sponsored events and festivals; the heterosexual 'mainstream.' This analysis of pride parades specifically, and gay and lesbian tourism more generally, posits festivals and events as sites of exclusion rather than inclusion. Furthermore, this line of argument suggests that pride parade organizers who collude with the interests of capital help to create an exemplary gay and lesbian tourist/consumer who is more interested in conspicuous consumption than in civil rights and fighting injustice. However, the annual global gay and lesbian events calendar also comprises events that radically challenge heteronormative assumptions around the appearance and behavior of bodies in public spaces.

The 'Queerball Strikes Back'! event (which was a street party that took place in New York in June 2012) is one such example. Organizers explicitly set about to 'trouble' the concept of the ubiquitous pride parade. While the event was small in numbers, the organizers raised the two general critiques of pride parades. First, that pride parades misrepresent the diversity of sexualities by privileging a white, middle-class, male parade participant. Second, that pride parades have 'sold out' because as the organizers of this event proclaimed, "we CAN'T BUY LIBERATION through corporate-sponsored Pride" (Anon., 2012). Another example is the Folsom Street Fair, held in September in San Francisco each year. The Folsom Street Fair is the penultimate event in that city's Leather Pride Week. In terms of its contribution to the tourism economy, the event draws something in the order of 400 000 participants and is

considered to be the third largest street event in California (Folsom Street Events, 2012). Billed as the world's largest leather event, this fair spreads out over a number of streets in the South of Market district, centered on Folsom Street. The event's origins lie within the city's gay and lesbian Bondage, Discipline, Sadomasochism (BDSM)/leather community. Norms surrounding sexuality are subverted through sexual acts of sadomasochism and bondage occurring on public streets during daylight hours and because many participants are either naked or seminaked, dressed in leather fetish gear. Importantly, bodies do not have to conform to conventional notions of sexual attractiveness – young, waxed, gym toned, and slim. Folsom Street Fair therefore broadens, rather than narrows, understandings of sexuality.

In contrast to the sexual freedom of Folsom Street Fair, gay and lesbian weddings and honeymoons, alongside pride events, are becoming increasingly integrated into the tourism industry in nations/states that have legalized same-sex marriage. Some cities now market themselves as gay and lesbian marriage destinations. For example, the American city of Baltimore offered wedding and honeymoons soon after legislation was passed legalizing gay marriage in Maryland (Yancey, 2012). Like pride events, evaluations of gay and lesbian wedding tourism is open to critique. On the one hand, gay wedding tourism offers possibilities to generate economic revenue, as well as to reconfigure destinations along the lines of social equality and 'gay friendliness.' This point is not lost on those advocating for gay marriage. For example, in New Zealand, the increased revenue generated from the tourism benefits of legalizing gay marriage are integral to arguments of proponents of changing the marriage act in that nation. On the other hand, as Boyd (2008: p. 228) suggests, "the commodification of gay marriage via the marketplace produces a new kind of queer citizen, one that participates in civic life via the social rituals of marriage and the commercial rituals of conspicuous consumption." Gay wedding tourism is another example of how expression of gay and lesbian sexuality and place is entwined within heteronormative assumptions – in this instance, monogamous couples and romantic love.

Risk

Tourist studies have conceptualized risk drawing on different schools of thought. In this section, attention is drawn to the insights offered to gay and lesbian tourism by two forms of risk: 'cultural risk' and 'medical risk.' Cultural risk draws on arguments that tourists are mindful of the social norms surrounding sexuality, ethnicity, religion, food, nudity, and so on when making destination choices. However, what we know about the social norms of sexuality in some nations suggests that cultural risks may play a more important role when people do not conform to the norms of heterosexuality. In some instances, the norms of heterosexuality remain institutionalized in law. Homosexuality or homosexual acts may be legislated as illegal, or homosexual citizens or visitors are positioned without any state protection from acts of prejudice or discrimination. For example, a small number of nations such as Iran and Saudi Arabia punish sodomy by the death penalty, while others such as Uganda, Jamaica, Belize, and Kenya may impose long-term imprisonment. A 2001 study by Amnesty International found instances where gay men who reported crimes at certain destinations such as Jamaica, the Caribbean, and Malaysia were "met with indifference and at worse face(d) further victimisation" (Brunt and Brophy, 2006: p. 290). People may choose to avoid nations in travel itineraries where they understand the cultural risks arising from sexuality as too high. Waitt and Markwell (2006) discussed the cultural risks that men opened themselves up to and the resulting compromises to their safety when visiting destinations such as Egypt and the islands of the Caribbean, where prejudice and discrimination toward homosexual people is coded within the law and where violence toward homosexuals is seemingly tolerated or even sanctioned by the state.

The concept of cultural risk is important to better understand why people who do not conform to dominant understandings of sexuality often find themselves as the object of victimization when configured as 'gay tourists.' Tomen and Markwell (2009) explored the cultural risks of visibility in places branded as 'gay friendly.' In their study of safety and hostility at a number of Australian gay and lesbian public events in the early 2000s, Tomen and Markwell (2009: p. ix) found what they called an "undercurrent of hostility, abuse and unreported violent attacks at these events, particularly in the aftermath of the Sydney Gay and Lesbian Mardi Gras parade." About 40% of the sample had experienced some form of hostility at one of these events and many participants employed a range of individual and group strategies to avoid hostility. These strategies included being careful about when and where to show overt signs of affection to their partner or friends, remaining within their friendship group traveling, concealing their appearance, or 'acting straight' on the way to and from the event (Tomen and Markwell, 2009). Likewise, the work of Brunt and Brophy (2006) found that victimization based on sexual identity was experienced either directly or indirectly by all the men in their sample of 20. The men in this study itemized physical assault, verbal abuse, threatening behavior, theft, and discrimination (Brunt and Brophy, 2006). Forms of discrimination were also considered by these men to be just as much a safety problem as victimization. The cultural risk identified in this study was the appearance of gay men who 'stood out from the crowd,' particularly those who did not conform to a conventional masculinity (Brunt and Brophy, 2006: p. 292). The authors concluded that the level of reporting of gay victimization and discrimination was underreported, suggesting a much greater level of incidence than official statistics might suggest. Discrimination, in all its various overt and covert, direct and indirect guises can sometimes be difficult to detect let alone measure.

The work of Hughes et al. (2010) illustrates that the cultural risks for gay and lesbian tourists may arise not from understandings of sexuality per se, but from notions of the respectful visitor. Hughes et al. (2010) in their study of resident attitudes toward men visiting a beach popular with gay men in Mexico found that residents rarely showed strong disapproval or dislike of homosexuality. Rather, residents expressed disapproval of overt sexual activity, and, to a lesser extent, nudism, in the 'public' spaces of the beach. However, the authors did suggest that there was "some indication of an underlying dislike of homosexuality, which was usually expressed in non-personal terms and probably a reflection of

deeper societal and cultural pressures" (Hughes et al., 2010: p. 783). The lack of respect shown by the number of visitors to the beach, rather than sexuality per se, was a source of contention. This points to how hostility toward increasing numbers of 'gay' visitors may be tied to understanding of respectable displays of affection and consideration for the integrity of the environment, which can frequently show evidence of damage from overuse or misuse at cruising areas.

Medical risks are the focus of health and behavioral scientists. Medical risks are defined by practices that increase the likelihood of long- or short-term harm by sustaining an injury or disease. For some people, becoming a tourist and engaging in relaxation is characterized by participation in a range of activities that involve practices associated with medical risks, such as excessive alcohol and recreational drug consumption, which might also lead to a greater risk of unsafe sexual activity. Indeed, some of the earliest research undertaken on gay tourism concerned aspects of sexual health, and, in particular, human immunodeficiency virus (HIV) (see Clift and Forrest, 1999, 2000; Luongo, 2000; Clift et al., 2002). Safer sex campaigns in affluent Western nations, built on scientific evidence of how HIV is transmitted, have been highly successful in reducing the infection rates of this virus. The fact that more people, including same-sex-attracted people, might be traveling to an ever-expanding range of destinations, does not necessarily mean this will lead to greater levels of HIV or other sexually transmitted infections (STIs). In other words, travel per se is not necessarily a risk factor in the acquisition of HIV or STIs. However, travel draws attention to the inequities in access to HIV therapies between the global south and north. Moreover, important questions are raised about how tourists from the affluent global north, who may be more knowledgeable about HIV, and its prevention, negotiate sex while traveling through the global south.

Results from a travel behavior study of British men who identified as gay suggested that 10% of the sample were "more likely to take sexual risks on holiday" (Clift et al., 2002: p. 187). However, the same study found that the proportion of these men who reported unprotected anal sex was around 15% (Clift et al., 2002: p. 188). The authors concluded that "while gay men are considerably more likely than heterosexual men to have sex with new partners on holiday, they are also are less likely to engage in unprotected penetrative sex" (Clift et al., 2002: p. 188). More recent research into gay men, sexual health, and travel (Benotsch et al., 2011: p. 402) found that a 'significant minority' of the men in their sample (which had been obtained from men attending a gay pride festival in Denver, Colorado) had engaged in medically risky sexual behavior, such as unprotected anal sex and substance use. Such results may point to the success of educational campaigns involving safer sex and condoms. However, as the authors noted, the participants in this study reported lower rates of both sexual and substance use behaviors associated with medical risk as well as 'lower motivations to seek sex' than reported in other studies (Benotsch et al., 2011: p. 403).

Conclusion

Travel is particularly important in fashioning Western gay and lesbian cultures. Traveling to other places, destinations at which sexual diversity is either imagined or legally recognized, continues to provide many people with possibilities to build and/or consolidate a gay or lesbian identity. The gay and lesbian tourism industry relies upon sustaining flows of visitors through marketing particular destinations (which may themselves wax and wane in popularity over time) as 'homelands' – Greenwich Village, New York, The Castro, San Francisco, Soho, London, and Darlinghurst, Sydney. As homelands, these locations offer opportunities to experience and learn about shared histories and cultures. Further, the gay and lesbian tourism industry relies on guidebooks, the World Wide Web, 'gay-friendly' businesses, and special events and festivals celebrating sexual diversity, to drive gay and lesbian tourism. However, gay and lesbian tourism is not without inherent contradictions and inequalities. The ways by which the gay and tourism industry is dovetailed with capitalist market forces through state and corporate sponsorship can lead it to be as exclusionary as it is inclusionary. Market-driven ideas of what it means to be gay or lesbian can create very narrow portrayals of gay and lesbian subjects. The profit-oriented imperatives of businesses mean that the gay and lesbian subject of the tourism industry is usually white, relatively affluent, middle class, conventionally attractive, and able-bodied – and also usually male. How sexuality intersects with the market forces too often compromises the social equality and social justice ambitions of gay and lesbian tourism. How sexuality intersects with travel, geography, and market forces of the tourism industry remains a crucial research agenda.

See also: Friendship Experiences of Lesbian, Gay, Bisexual, and Transgender Individuals; Gay and Lesbian Movements; Heterosexism and Homophobia; Lesbian, Gay, Bisexual and Trans-sexual Drag Culture; Lesbian, Gay, Bisexual and Trans-sexual: Commercial Sex; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Medicine and Health for Sexual Minorities; Queer Anarchism; Queer Geography; Same-Sex Sexualities in Turkey; Sexual Minorities and Sexual Citizenship; Sexual Minorities in the Maghreb; Sexual Orientation: Historical and Social Construction; Sexuality and Space; Tourism.

Bibliography

Aldrich, R., 1993. The Seduction of the Mediterranean: Writing, Art and Homosexual Fantasy. Routledge, London.

Anon., 2012. Queerball Strikes Back! A Radical Queer Street Party. http://newyork.gaycities.com/events/151805-queerball-strikes-back-a-radical-queer-street-party (accessed 12.12.12.).

Benotsch, E.G., Martin, A.M., Espil, F.M., Nettles, C.D., Seal, D.W., Pinkerton, S.D., 2011. Internet use, recreational travel and HIV risk behaviours in men who have sex with men. Journal of Community Health 36, 398–405.

Binnie, J., Klesse, C., 2011. 'Because it was a bit like going to an adventure park': the politics of hospitality in transnational lesbian, gay, bisexual, transgender and queer activist networks. Tourist Studies 11, 157–174.

Binnie, J., Skeggs, B., 2006. Cosmopolitan knowledge and the production and consumption of sexualised places: Manchester's gay village. In: Binnie, J., Holloway, J., Millington, S., Young, C. (Eds.), Cosmopolitan Urbanism. Routledge, London and New York, pp. 220–245.

Bleys, R.C., 1996. The Geography of Perversion: Male to Male Sexual Behaviour Outside the West and the Ethnographic Imagination 1750–1918. Cassell, New York.

Boyd, N.A., 2008. Sex and tourism: the economic implications of the gay marriage movement. Radical History Review 100, 222–235.

Boyd, N.A., 2011. San Francisco's Castro district: from gay liberation to tourist destination. Journal of Tourism and Cultural Change 9, 237–248.
Brunt, P., Brophy, K., 2006. Gay tourist victimisation. International Review of Victimology 13, 275–299.
Burns, K., 2012. Cosmopolitan sexual citizenship and the project of queer world-making at the Sydney 2002 gay games. Sexualities 15, 314–335.
Careaga, G., Sabbadini, R., 2012. Recent advances in LGBT rights. In: UNWTO Global Report on LGBT Tourism. World Tourism Organization, Madrid.
Casey, M.E., 2009. Tourist gay(ze) or transnational sex: Australian gay men's holiday desires. Leisure Studies 28, 157–172.
Clift, S., Forrest, S., 1999. Gay men and tourism: destinations and holiday motivations. Tourism Management 20, 615–625.
Clift, S., Forrest, S., 2000. Tourism and the sexual ecology of gay men. In: Clift, S., Carter, S. (Eds.), Tourism and Sex: Culture, Commerce and Coercion. Pinter, London, pp. 179–199.
Clift, S., Callister, C., Luongo, M., 2002. Gay men, holidays and sex: surveys of gay men visiting the London freedom fairs. In: Clift, S., Luongo, M., Callister, C. (Eds.), Gay Tourism: Culture, Identity and Sex. Continuum, London and New York, pp. 231–248.
Community Marketing Incorporated, 2003. 8th Annual Gay and Lesbian Travel Survey. Community Marketing Incorporated, San Francisco.
Community Marketing Incorporated, 2009. LGBT Tourism Demographic Profile. Cited in UNWTO 2012.
De Britto Lomanto, M., 2012. EMBRATUR: committed to putting Brazil on the map for LGBT visitors. In: Global Report on LGBT. UNWTO, Madrid, pp. 18–19.
Folsom Street Events, 2012. http://www.folsomstreetevents.org/ (accessed 12.12.12.).
Hodes, S., Vork, J., Gerritsma, R., Bras, K., 2007. Amsterdam as a gay tourism destination in the twenty-first century. In: Richards, G., Wilson, J. (Eds.), Tourism, Creativity and Development. Routledge, London and New York, pp. 178–188.
Hughes, H., 1997. Holidays and homosexual identity. Tourism Management 18, 3–7.
Hughes, H., 2006. Pink Tourism: Holidays of Gay Men and Lesbians. CAB, Wallingford.
Hughes, H., Monterrubio, J.C., Miller, A., 2010. 'Gay' tourists and host community attitudes. International Journal of Tourism Research 12, 774–786.
Isherwood, C., 1945. The Berlin Stories. New Directions, New York.
Luongo, M., 2000. Gay tourists in New York city. In: Clift, S., Carter, S. (Eds.), Tourism and Sex: Culture, Commerce and Coercion. Pinter, London, pp. 109–130.
Markwell, K., 2002. Mardi Gras tourism and the construction of Sydney as an international gay and lesbian city. GLQ: A Journal of Lesbian and Gay Studies 8, 81–99.
Pronger, B., 1990. The Arena of Masculinity, Sports Homosexuality and the Meaning of Sex. GMP Publishers, London.
Sibalis, M.D., 1999. Paris. In: Higgs, D. (Ed.), Queer Sites: Gay Urban Histories Since 1600. Routledge, London and New York.
Tasmanian Gay and Lesbian Rights Group, 2012. 'Buy Tasmanian' Call from Man Who Once Led Boycott against State's Former Anti-gay Laws. http://tglrg.org/index/C0_2_1/ (accessed 12.18.12.).
Tomen, S., Markwell, K., 2009. When the Glitter Settles: Safety and Hostility at and Around Gay and Lesbian Public Events. Australian Institute of Criminology Research and Public Policy Series, 100, Canberra.
United Nations World Tourism Organization, 2012. Global Report on LGBT Tourism. World Tourism Organization, Madrid.
Valentine, G., 1996. (Re)-negotiating the heterosexual street: lesbian productions of space. In: Duncan, N. (Ed.), Body Space: Destabilising Geographies of Gender and Sexuality. Routledge, London and New York, pp. 146–155.
Vickers, A., 1989. Bali, A Paradise Created. Penguin, Ringwood, Victoria.
Vyras, P., 1997. Magnus Hirschfeld in Greece. Journal of Homosexuality 34, 17–30.
Waitt, G., Markwell, K., 2006. Gay Tourism: Culture and Context. The Haworth Press, Binghamton, New York.
Yancey, K., 2012. Baltimore Tourist Bureau Puts up a Gay Wedding Site. http://www.usatoday.com/story/dispatches/2012/11/07/baltimore-ready-with-gay-wedding-website/1689573/ (accessed 11.23.12.).

Relevant Websites

http://gaygames.org/wp/ – Federation of Gay Games (accessed 12.17.12.).
http://www.folsomstreetevents.org/ – Folsom Street Fair (accessed 12.17.12.).
http://galta.com.au/ – Gay and Lesbian Tourism Australia (accessed 12.17.12.).
http://www.geta-europe.org/ – GETA, Gay European Tourism Association (accessed 12.17.12.).
http://www.iglta.org/ – International Gay and Lesbian Travel Association (accessed 12.17.12.).
http://interpride.org/ – InterPride, International Association of Pride Organizers (accessed 12.17.12.).
http://www.mardigras.org.au/ – Sydney Mardi Gras (accessed 12.17.12.).
http://www2.unwto.org/en/publication/am-reports-volume-3-global-report-lgbt-tourism – United Nations World Tourism Organization, AM Reports Volume 3: LGBT Tourism (accessed 12.17.12.).

Gay Men's Emancipation Movement in Germany: 1897–1933

Jason Crouthamel, Grand Valley State University, Allendale, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

As has been true in contemporary actions for gay rights, tensions over gender identity in the early German homosexual rights movement presented a difficult challenge. This article provides an overview of the diverse scientific, ideological, and cultural perspectives on homosexuality within the male homosexual community in imperial and Weimar Germany. It also examines the effects of the World War I on how homosexual men conceptualized masculinity and the nature of homosexuality, and the struggle for civil rights in the Weimar years before the catastrophic Nazi seizure of power. Implications of German history for contemporary gay rights are noted.

This article focuses on the origins of the homosexual rights movement in Germany, its diverse ideological trajectories, the effects of the World War I on perceptions of homosexuals, and the struggle for civil rights in the Weimar years before the catastrophic Nazi seizure of power. Germany's founding homosexual rights movement engaged in debates that still persist in contemporary cultural wars: Is homosexuality natural? Are homosexual men essentially feminine or masculine? For scholars, it is important to consider that, similar to mainstream culture's debates over the heterosexual ideal, there was not a hegemonic image of the homosexual male within the newly defined gay community. The fragmented homosexual rights movement in traumatized and divided German history must be analyzed in its particular historical framework, but it also reveals the degree to which contemporary debates over gay identity and civil rights have precedents in over a century of struggle for homosexual emancipation.

In nineteenth-century Germany, an intersection of medical, political, and cultural developments led to the identification of homosexuality as an innate characteristic and a distinct identity. Based on this interpretation of homosexuality as a natural condition, German pioneers of the homosexual rights movement organized and mobilized against the criminalization of homosexual behavior. However, the homosexual emancipation movement was not a monolithic entity. The homosexual rights organizations that emerged in imperial Germany shared the goal of dismantling Paragraph 175, the section of the Prussian law code that criminalized 'sodomy' between men, but the movement was ideologically fractured over questions about the nature of male homosexuality. For scholars dealing with the history of homosexuality, this era presents substantial challenges. Sources produced by homosexual men who experienced discrimination and war in imperial and Weimar Germany highlight a fierce debate over the image of the homosexual male as a 'feminine' or 'masculine' being.

The Origins of the Homosexual Rights Movement in Late Nineteenth-Century Germany

Historian Robert Beachy demonstrates that the homosexual rights movement originated in Germany in the mid-nineteenth century, where the notion of homosexuality as a congenital trait was first defined by scientists and advocates for this sexual minority. Armed with the belief that homosexuality was a natural, inborn tendency, activists critiqued the criminalization of homosexuality as unjust (Beachy, 2012). In the wake of this new scientific theory, there emerged the first homosexual rights organization – the *Wissenschaftlich humanitäres Komitee* (WhK, or Scientific Humanitarian Committee), founded in 1897 by Dr Magnus Hirschfeld and his colleagues. Hirschfeld, a doctor, publisher, and supporter of the Social Democratic Party (SPD), saw the WhK as part of other movements for social reform and civil rights in imperial Germany, and one of the group's main goals was the dismantling of Paragraph 175 of the constitution, which criminalized same-sex behavior between men. The early homosexual rights movement was particularly focused on male homosexuality, in part because Paragraph 175 specifically targeted male homosexuality, and it reflected the broader cultural perception that female homosexuality was not a real phenomenon. The WhK worked vigorously to educate the public about the scientific nature of homosexuality and provided support to victims of Paragraph 175 and blackmail. Hirschfeld and his supporters collected petitions and lobbied Social Democrats sympathetic to reforming the law, and they published pamphlets and gave lectures on the biological nature of homosexuality (Steakley, 1992). Hirschfeld also reached out to broader sexual reform movements, including Helene Stöcker's League for the Protection of Motherhood and Sex Reform (*Bund für Muttereschutz und Sexualreform*), founded in 1905. Stöcker's socialist political leanings and organization of sex counseling clinics, also spearheaded by Hirschfeld's Institute for Sex Research, created natural groundwork for organizational and ideological alliances between the homosexual and women's movements in imperial Germany, between 1871 and 1914 (Grossmann, 1997).

Within the context of the late-nineteenth century medicalization of homosexuality, in which scientists began to define homosexuals in biological terms, Hirschfeld's WhK was an integral organization in the push to characterize male homosexuals as a separate, unique minority, a sexual intermediary, or 'third sex' with a partially female constitution that deserved both further scientific research and protection under the law (Oosterhuis and Kennedy, 1991). Homosexuality, he argued, could be recognized by its mental and physical symptoms,

including effeminate characteristics in manner, voice, and body. The WhK's interpretation of homosexuality and its cultural, political, and scientific significance was challenged by a homosexual men's organization founded in 1903, the *Gemeinschaft der Eigenen* (GdE, or 'Community of the Unique'), cofounded by Adolf Brand. An iconoclastic former teacher in Berlin who expressed both anarchistic and nationalistic rhetoric, Brand began publishing the journal *Der Eigene* ('The Unique One'), which was designed as a forum for male culture that celebrated *Freundesliebe* ('love of friends') as interpreted through a revival of ancient Greek celebrations of the male body and mind. Brand held up the Greek ideal of *Eros* as a goal that could liberate men from the shackles of bourgeois repression. He called for young men to discover the value of 'friend-love' in nationalist youth groups like the *Wandervogel* and the *völkisch* nudist groups who were part of a broader assault on prevailing bourgeois and sexual norms in imperial Germany (Ramsey, 2008). In their interpretations of homosexual identity, Brand's GdE and Hirschfeld's WhK were polar opposites. Brand and his colleagues sharply attacked what they saw as Hirschfeld's 'effeminate' classification of homosexuals. The idea that homosexuals possessed a feminine constitution was considered repulsive by members of the GdE, who envisioned an all-male utopia, or at least a society in which women were kept in their place as mothers and servants. Further, Brand argued, the biologists and sexologists had drained beauty from eroticism, and turned male sexuality into a base, scientific category (Oosterhuis and Kennedy, 1991), In contrast to Hirschfeld's WhK, the GdE had no interest in forming alliances with the women's movement. Similar to contemporary tensions in gay male culture and mainstream society identified by cultural theorist Judith Halberstam, Brand was antagonistic toward men who did not fit into a clear 'masculine'/'feminine' binary. Brand and his acolytes perceived 'alternative masculinities,' to use Halberstam's term, including women performing traditionally masculine roles, as a threat to male dominance (Halberstam, 1998).

According to historian Jeffrey Schneider, police surveillance of homosexuals in *fin-de-sieclé* Germany forced the community underground, but persecution was also a key catalyst in the formation of the homosexual rights movement (Schneider, 2012). Sources documenting Germany's homosexual community are rare, but there are sources that reveal the existence of a homosexual subculture, which imperial Germany's police was willing to tolerate as long as it remained contained and controlled. As Robert Beachy documents, the Department of Blackmailers and Homosexuals, created with the Berlin police force in 1885, demonstrated a 'tacit forebearance' of consensual homosexual sex. While the police placed surveillance on the gay and transvestite bars, keeping extensive mugshots of individuals described by police as 'pederasts,' reflecting widely held views that homosexual men preyed on underage males, homosexuals were allowed to form a community (Beachy, 2012). However, the German military was tainted by a series of public scandals over the discovery of homosexual aristocratic officers, most famously Count von Eulenburg in 1907, who were in the close circle of Kaiser Wilhelm II. These scandals shook the aristocratic officer corps and reflected anxiety about an epidemic of 'effeminate' and degenerate men in an institution that was supposed to be the epitome of masculinity (Funck, 2002). The Prussian War Ministry also recorded its anxiety about mostly working-class soldiers who cruised the streets, especially in Berlin, exchanging sex with mostly upper class civilians for money. The War Ministry portrayed these soldiers as innately heterosexual men who were being taken advantage of, reinforcing prevailing assumptions about soldiers as innocent victims morally corrupted by homosexual civilians. Similar to debates over "Don't Ask Don't Tell" in the United States in 1993–2011, many opponents to the integration of gays perceived them as sexual predators who threatened heterosexual soldiers. Within Germany's army barracks, soldiers who prostituted themselves were generally not stigmatized, or even considered homosexual, especially if they took the active, 'masculine' role in the sexual encounter with a passive, 'effeminate' partner. In imperial Berlin, the police tacitly accepted the homosexual subculture, and Paragraph 175, the 1871 law that prohibited 'unnatural fornication' (*widernatürliche Unzucht*) between men, was only lightly enforced (Schneider, 2012). The homosexual community would be seriously challenged, however, by the experience of the Great War (World War I).

The Effects of the World War I (1914–18) on Germany's Homosexual Rights Movement

The Great War profoundly traumatized millions of men, shattering their bodies and minds in industrialized violence. However, for many homosexual veterans, the war, despite its traumatic effects, opened up the possibility of social and political integration. At the same time, the homosexual rights movement remained ideologically divided, and it is important to examine the war's impact on the two main camps in the movement, Hirschfeld's WhK and Brand's GdE. Two of the original four members of the WhK were called into the army, and Hirschfeld mobilized scientists and activists in the WhK to study the broader sexual effects of the war (Steakley, 1992). The WhK also collected letters from thousands of soldiers on their war experiences and sexual behavior. Based on these data, Hirschfeld argued that the war created circumstances of physical and psychological deprivation that drove otherwise heterosexual men to seek sex with other men. Hirschfeld identified what he called 'pseudohomosexuality,' which he defined as heterosexual men engaging in homosexual behavior, which caused the most anxiety in military authorities who feared a breakdown of masculinity (Hirschfeld, 1930). Hirschfeld also found that the front experience enabled men who were 'constitutionally homosexual' to find other homosexual men in an environment that was more tolerant of same-sex relations than prewar or mainstream culture. The war did not create homosexual behavior, in other words, rather facilitated relations between homosexual men who were otherwise sexually and socially repressed before the war. Hirscheld identified three forms of 'intimate comradeship': the consciously erotic, unconsciously erotic, and bonds between men that remained unerotic. Hirschfeld recorded numerous accounts of men who discovered love with other men, and these relationships were often tolerated by their heterosexual comrades as natural examples of tenderness and love experienced between men in combat. While he was convinced that

homosexual men could easily adapt to their 'comradely' roles in war, Hirschfeld also pointed to numerous reports that exemplified his theory that constitutional homosexuals were essentially feminine, including accounts of soldiers who practiced transvesticism (Herrn, 2003). While he acknowledged that some men with homoerotic inclinations adapted easily to the militaristic culture of war, he argued that 'feminine homosexuals' ('feminine Urnings' or members of the third sex) were repulsed by the war and preferred the nurturing work of caring for the wounded and other 'feminine' tasks (Hirschfeld, 1930).

Adolf Brand and the GdE largely suspended its publications when many of its members, including Brand, entered military service. When Brand's publications resumed after the war and Brand himself returned after serving in the army, it was clear that the trench experience provided the GdE with new images and rhetoric for celebrating the ultramasculine homosexual male. In one of the first editions of *Der Eigene* published after the war, the GdE-sponsored journal adopted a militarized language where words like 'sacrifice,' 'the front,' 'battle' all began to permeate the movement's way of thinking about the relationship between homosexual men and the prevailing culture. Although Brand had long worshipped ancient Greek warrior ideals, the front experience of 1914–18 gave his organization a new context and language for articulating homosexual emancipation. The war, Brand insisted, proved that homosexual men were exceptional individuals upon whom the nation depended. He argued that they possessed the emotional characteristic – male love – essential for military life and thus the defense of the nation, and that this placed them in the elite of German society. One of Brand's acolytes, G.P. Pfeiffer, argued that 'physiological friendship' was the foundation for heroism, courage, and sacrifice displayed in war. War did not necessarily cause men to engage in homosexual behavior, Pfeiffer emphasized, but it did awaken intimate emotional bonds between men – comradeship – that transcended sexual instincts (Crouthamel, 2011).

The GdE's conviction that homosexual men were ideally suited to the defense of the nation resonated among other right-wing activists who saw the homosexual warrior as the core of a new society. Hans Blüher, a leading figure in Germany's *Wandervogel* movement since 1902, believed that the war demonstrated the essential role of male–male relationships in preserving Germany's future. Male–male relationships experienced in *Männerbünde* (male associations) were superior to heterosexual relationships, he argued, because women weakened men with their alleged spiritual and intellectual inferiority and pressure to conform to effeminate, bourgeois family norms (Bruns, 2008). Even some of the politically moderate homosexual activists made that case that homosexual men were particularly suited to enduring the strain of modern war because they were accustomed to emotional stress from persecution. In the 1919 issue of *Die Freundschaft* (*Friendship*), a newspaper that published articles by homosexual men from diverse ideological perspectives, one poem celebrated the bravery of homosexual men who must remain manly and unshakeable as they endure scorn. These themes were more explored in the novel *Alf*, written by Bruno Vogel in 1929. Vogel, a survivor of the trenches in Flanders and founder of the Leipzig branch of the WhK, compared the emotional stress of living under antisodomy laws to the traumatic experience of combat. In the novel, Alf consoles his friend Felix, who is arrested on the home front under 'their stupid paragraphs.' Writing from the trenches, Alf denounces the architects of the war and of Paragraph 175 as part of the same 'evil and stupidity,' and he promises to fight against both war and homophobia to help future generations. Alf portrays himself as emboldened by his combat experience to fight on behalf of the less confident and passive Felix (Vogel, 1929). Homosexual activists thus suggested that homosexual men, long experienced in the trauma of repressed emotions, were uniquely equipped to cope with the psychological trauma of war.

Military defeat in 1918 brought widespread accusations, especially from the political right, that 'enemies' at home had betrayed the army. Although some had argued that homosexuality fits very well with military victory, homosexuals, along with Jews, were frequently the targets of this 'stab-in-the-back' legend. In his pamphlet *The Sexual Cruelties of Love-Crazy Men*, conservative journalist H. A. Preiss lamented that the war led "normal men," as well as "intrinsically degenerate men," to turn to homosexuality, fetishism and other "abnormal sexual practices in order to relieve their tense nerves." Preiss claimed that veterans had become so focused on their abnormal sexual needs that they were no longer willing to sacrifice for the nation. Homosexual activists were self-conscious about how society perceived these new warriors for social justice, and they sought to combat 'stab-in-the-back' accusations. Instead, homosexual veterans characterized homosexuals as responsible members of the national community whose war experience entitled them to be accepted members of the community. In 'Against What Do We Fight?', veteran and contributor to *Die Freundschaft* B. Eden depicted homosexual men as warriors dedicated to protecting the nation and German values. Within this context of defeat and intensifying political fragmentation, homosexual activists struggled to make their case that the war experience made homosexuals comparable to heterosexual citizens, and thus deserving of respect (Crouthamel, 2011).

Divisions in the Movement during the Weimar Republic and the Crisis of Hitler's Rise to Power

Defeat and revolution in 1918 highlighted the disparities between Germany's homosexual rights organizations. When Germany lost World War I, revolution broke out, the monarchy was overthrown, and, after a brief civil war, a coalition of prodemocratic parties formed a constitutional democracy called the Weimar Republic in 1919. Germany's defeat traumatized Brand's hypernationalistic GdE, which perceived the new Weimar Republic as a threat to its elitist ideals. In Brand's universe, defeat and democracy threatened to annihilate the superior warrior-male. In contrast, the politically progressive WhK and the politically moderate *Bund für Menschenrecht* (BfM – League for Human Rights), cofounded in 1923 by businessman Friedrich Radszuweit, saw Weimar as an ideal opportunity. The broader movements for social revolution and sexual reform that intensified in the 1920s gave the more progressive and moderate homosexual rights organizations, which endorsed the ascendant prodemocratic parties, a new context in which they could argue for the decriminalization of

homosexuality, their long-held priority. However, it was not only Weimar's new social and political context that gave homosexuals confidence to articulate goals for legal reform. Politically moderate and progressive veterans from the WhK and BfM also employed the militarized language of the trenches, and the new image of a homosexual warrior-activist, to justify the abolition of Paragraph 175. In contrast to the WhK's medical and scientific focus on homosexuality, and the GdE's concentration on the aesthetics of 'friend-love' and the nationalistic identities of homosexual men, Radszuweit's BfM focused on the democratic and legal concerns of integrating homosexuals into the social fabric. The BfM concentrated on championing the individual rights of homosexual citizens, and they hoped to unify homosexuals across the political spectrum to support the legal demand for homosexual rights (Ramsey, 2008).

Homosexual veterans from both moderate and progressive perspectives idealized the front experience, and they saw themselves as the ultimate models of the struggle and exceptional victims of oppression who tragically suffered under both war and Paragraph 175. Their newspaper articles focused on militaristic values of fighting, sacrifice, and vigilance. This rhetoric was especially aggressive in *Die Freundschaft*, the most widely circulated homosexual periodical at that time. Homosexual veterans called upon their colleagues in the movement to avoid wallowing in the pain of war-time suffering and instead channel their energies into the postwar battle to bring homosexuality into the light and take pride in one's identity. The fight for homosexual emancipation, writers for *Die Freundschaft* argued, was an ideal path for simultaneously healing the wounds of war and securing a place and purpose in postwar society. In his 1919 article 'Dead Sunday' (*Totensonntag*), for instance, veteran A. Lange compared combat in the trenches and postwar combat to gain psychological and political freedom. Lange suggested that the same spirit of sacrifice found at the front lived on in those fighting for homosexual emancipation (Crouthamel, 2011). Lange's postwar experience reflected that of many gay veterans around the world who after World War II and subsequent wars gained confidence to come out of the closet and organize activism in the fight for civil rights (Vickers, 2009).

The postwar battle fought by homosexual men was not just for political emancipation; it was also a battle for the image and perceptions that homosexual men held of themselves. The new postwar political movements had little room for 'effeminate' men. For Adolf Brand and the GdE, this was not a change, as Brand celebrated comradeship at the front as an ideal hypermasculine environment, an all-male utopia where 'real men' could explore ideals of eros without the feminine restraints imposed by bourgeois culture. In a 1930 GdE-supported publication, *Eros*, a supplement to *Der Eigene*, Brand tore apart the conservatives' image of homosexual men as antithetical to the warrior ideal. Specifically, Brand attacked the Nazis' claim that "man-man love emasculates our *Volk* and makes one the play-thing of one's enemies." Brand lamented that homophobes who denounced homosexuals as weak and effeminate symbolized society's complete failure to understand homosexuality. Despite all the scientific work done on homosexuality by his rival Magnus Hirschfeld, Brand claimed that little progress had been made to educate the public on the reality of 'friendship' and homosexual identity (Crouthamel, 2011).

Progressive reformers also embraced a warrior image for the homosexual activist. *Die Freundschaft* welcomed contributions from homosexual activists who represented warrior masculinity in a positive light. One writer for *Die Freundschaft* portrayed homosexual men as battle-hardened veterans, and he specifically targeted men "who clean themselves like young girls and go play and dance" as useless in the new fight. Effeminate homosexuals were weakening the struggle: "We need men, complete men (*ganze Männer*). Effeminate men (*weibliche Männer*) are no good for battle and conflict"(Crouthamel, 2011).

Ironically for the left-wing WhK and moderate BfM, the most public controversy over whether or not homosexual men were really 'masculine' and capable of warrior virtues was fought over the openly homosexual Nazi Stormtrooper, Ernst Röhm. Social Democrats and other left-wing parties used Röhm's homosexuality as a lightning rod to denounce his moral character. While the WhK and BfM shared the SPD's ideological disdain for Röhm and the Nazis, they defended Röhm's personal character, arguing that he was a victim of prejudice and hypocrisy, and that his sexual orientation should be a private matter that had no bearing on his moral qualifications for leadership. The BfM's Friedrich Radszuweit saw Röhm's commitment as a political fighter as evidence that he was entitled to be treated as any other member of society, regardless of his sexual orientation. Radszuweit chastised the SPD's homophobic motives and alleged hypocrisy when *Vorwärts* published a cartoon depicting a Nazi leader standing behind a row of Stormtroopers, Hitler's elite street fighters, and the caption: "As your leader I'm right behind you." BfM activist Paul Weber reminded the SPD of its long-standing commitment to decriminalizing homosexuality, and noted that among its members and leaders were numerous patriotic, upstanding homosexuals (Herzer 1995). Radszuweit stressed that the party should be consistent in its commitment to treating homosexuality as a private issue and he called on the left to confine its critiques to the Nazi party's worldview and politics. Referring to the role played by men like Röhm in the war, Radszuweit emphasized that homosexuals "have performed great things for their *Volk* and for their countries." The presence of loyal party comrades in the Nazi movement, BfM activists carefully pointed out, should be a lesson to the Nazis that homosexuals are productive members of society: "One should take note that the National Socialists have learned something from this scandal: that homosexuals are also capable and decent people who can be useful, and that these men are entitled to the same rights to exist as every so-called normal citizen"(Crouthamel, 2011).

Despite Radszuweit's call for greater tolerance, the Nazis targeted homosexuals as enemies of the 'racial state' once they came to power in 1933. The Nazis perceived homosexual men as effeminate, degenerate 'asocials' who polluted the nation's racial health and tarnished the image of the hypermasculine warrior ideal (Mosse, 1998; Burleigh and Wippermann, 1991). When Hitler ordered Röhm executed in 1934 after the latter's expressed intent to take over the army, Röhm's homosexuality was used as a pretense. The head of the Schutzstaffel, Heinrich Himmler, began a campaign of violence against homosexual

men and promised to purge them from German society. The Nazis quickly modified the Paragraph 175 law, expanding it to criminalize any perceived homosexual intimacy, including gestures, in addition to the prohibition on sexual intercourse between men. Virtually overnight, the homosexual subculture and community that flourished in Weimar Germany was eradicated. Hirschfeld's Sexual Research Institute was destroyed, gay venues were shut down, and homosexuals were sent to concentration camps. There was wide consensus for Nazis' antihomosexual measures, as the regime's perceptions of homosexuals as 'asocials' were shared by many ordinary Germans (Pine, 2007).

Conclusion

In order to analyze the experiences and perspectives of homosexual men in early twentieth-century Germany, it is important for historians to recognize the homosexual community's diverse perspectives on the nature of homosexuality. While Hirschfeld and many of his adherents argued that homosexual men were an essentially feminine, 'third' sex that should be accepted, many homosexual men who experienced the comradeship and militarization of the Great War asserted that they were masculine warriors who fulfilled the hegemonic gender ideal expected from heterosexual men. This new image of the hypermasculine, homosexual warrior-male came to permeate the movement, which identified these ex-front fighters as ideally suited to carrying out the battle for civil rights and social integration. While the gay community in the 1920s found its path for arguing that homosexuals were made into 'real men' through the war experience, others still perceived them as enemies of the 'national community,' as evidenced by the violence orchestrated by the Nazi regime against gay men after 1933. Embracing the hegemonic image of masculinity did not substantially alter the gay community's status as 'social outsiders' in the eyes of mainstream culture. Even after the collapse of Nazism in 1945, Paragraph 175 remained law until 1969, when, after an international movement for homosexual rights intensified in the wake of the Stonewall riots in New York, homosexuality was finally decriminalized in Germany. For contemporary gay politics, the early twentieth-century homosexual rights movement in Germany reveals that, since its origins, the homosexual rights movement has constructed diverse interpretations of homosexual identity, and the movement's attempts to convince mainstream society that homosexuality is natural has been a difficult challenge. While decriminalization has been followed by greater integration of homosexuals into mainstream society, perceptions of homosexual men as deviant persist, despite attempts by homosexual men to assert that their identity is consistent with, or even essential to, the interests of the nation.

See also: Clinical Psychology, Psychiatry and Homosexuality; First World War, The; Gay and Lesbian Movements; Heterosexism and Homophobia; Masculinity, Scripts of; Mass Media, Representations In; Second World War, The; Sexual Minorities and Sexual Citizenship; Sexual Orientation: Historical and Social Construction; War Propaganda; War and Nationalism.

Bibliography

Beachy, R., 2012. To police and protect: the surveillance of homosexuality in Imperial Berlin. In: Spector, S., Puff, H., Herzog, D. (Eds.), After the History of Sexuality: German Genealogies with and beyond Foucault. Berghahn, New York, pp. 109–123.

Bruns, C., 2008. Politik des Eros – Der Männerbund in Wissenschaft, Politik und Jugendkultur, 1880–1934. Böhlau Verlag, Köln.

Burleigh, M., Wippermann, W., 1991. The Racial State: Germany 1933-1945. Oxford, New York.

Crouthamel, J., 2011. 'Comradeship' and 'friendship': masculinity and militarization in Germany's homosexual emancipation movement after the first world war. Gender & History 23, 111–129.

Funck, M., 2002. Ready for war? Conceptions of military manliness in the Prusso-German officer corps before the first world war. In: Hagemann, K., Schüler-Springorum, S. (Eds.), Home/Front: The Military, War and Gender in Twentieth-Century Germany. Berg, New York, pp. 43–68.

Grossmann, A., 1997. Reforming Sex: The German Movement for Birth Control and Abortion Reform, 1920-1950. Oxford University Press, New York.

Halberstam, J., 1998. Female Masculinity. Duke University Press, Durham.

Herrn, R., 2003. Schnittmuster des Geschlechts – Transvestitismus und Transsexualität in der frühen Sexualwissenschaft. Psychosozial-Verlag, Gieβen.

Herzer, M., 1995. Communists, Social Democrats and the Homosexual Movement in the Weimar Republic. In: Hekma, G., Oosterhuis, H., Steakley, J. (Eds.), Gay Men and the Sexual History of the Political Left. Harrington Park Press, New York.

Hirschfeld, 1930. Sittengeschichte des Weltkrieges. Erster Band. Verlag für Sexualwissenschaft, Schneider & Co, Leipzig.

Mosse, G., 1998. The Image of Man: The Creation of Modern Masculinity. Oxford, New York.

Oosterhuis, H., Kennedy, H., 1991. Homosexuality and Male Bonding in Pre-Nazi Germany. Harrington Press, New York.

Pine, L., 2007. Hitler's 'National Community': Society and Culture in Nazi Germany. Bloomsbury, New York.

Ramsey, G., 2008. The rites of the Artgensossen: contesting homosexual political culture in Weimar Germany. Journal of the History of Sexuality 17, 87–102.

Schneider, J., 2012. Soliciting fantasies: knowing and not knowing about male prostitution by soldiers in Imperial Germany. In: Spector, S., Puff, H., Herzog, D. (Eds.), After the History of Sexuality: German Genealogies with and beyond Foucault. Berghahn, New York, pp. 124–138.

Steakley, J., 1992. The Homosexual Emancipation Movement in Germany. Salem, Ayer.

Vickers, E., 2009. The 'Good Fellow': negotiation, remembrance and recollection – homosexuality in the British armed forces, 1939-1945. In: Herzog, D. (Ed.), Brutality and Desire: War and Sexuality in Europe's Twentieth Century. Palgrave Macmillan, New York, pp. 109–134.

Vogel, B., 1929. (Original). Alf (B. Samuel, Trans.). Johnson, East Haven.

Gay, Lesbian, and Bisexual Youth

Ritch C Savin-Williams and Kenneth M Cohen, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.C. Savin-Williams, L.M. Diamond, volume 9, pp. 5892–5895, © 2001, Elsevier Ltd.

Abstract

Gay, lesbian, and bisexual youth have multiple developmental trajectories that consist of a number of milestones, the order and timing of which vary across individuals. After considering the deceptively simple constructs of who is gay and how many nonheterosexual youth exist, we review what is known about early feelings of being different from peers, the onset of same-sex attraction, sexual questioning, first sexual experience, recognition and self-labeling, disclosure to others, first romantic relationship, and self-acceptance.

Introduction

This article is based in part on Cohen and Savin-Williams, 2012. Same-sex attracted youth have always existed throughout history and across cultures. What is unique to contemporary society is that far more youth now label their sexuality as gay, lesbian, or bisexual. By constructing a sexual identity based on their sexual orientation, sexual behavior, romantic relationships, and, perhaps, a lifestyle, they are at the vanguard of making sexuality a critical component of personal identity. Although adolescents may experience their same-sex oriented self as newly emerged or energized with the onset of puberty, it nonetheless likely has been present from birth and will remain fundamentally unchanged on the deathbed.

During the past four decades, research on same-sex oriented youth has provided an increasingly sophisticated and nuanced understanding of their lives. Here, we focus on the development of the individual from first memories to self-acceptance. This process is an evolving and frequently lengthy one, consisting of multiple developmental milestones that vary in order, intensity, timing, and significance across individuals. This empirically validated perspective corrects previous conceptions of gay youth development as following universal, invariant, and steplike identity stages.

Evolution of Gay Youth

Although homoerotic youth have been described by artists, poets, religious writers, and historians for hundreds of years, it was not until the early 1970s and 1980s that medical, public health, and clinical professionals began investigating youth who they labeled as gay (and hence creating a new sexual category) as a unique group with its own exceptional developmental characteristics. This research cast doubt on earlier interpretations of adolescent same-sex sexual encounters as experimental, puberty-inspired outcomes unrelated to an inevitable heterosexuality.

Early research studies primarily comprised small samples of adolescents identified as gay whom investigators could identify and entice to participate – overwhelmingly youth in trouble and in need of social and mental health facilities (runaways, delinquents, prostitutes, suicide survivors). Some of these youth were being served by social service agencies such as New York City's Institute for the Protection of Lesbian and Gay Youth (later, the Hetrick-Martin Institute and its affiliated Harvey Milk School) that were established in large urban centers to address the physical, social, educational, and therapeutic needs of gay youth. Not surprisingly, the findings from these studies highlighted what was believed to be the ubiquitously difficult lives of gay adolescents who inevitably were in dire peril. As information accumulated, the portrait that emerged was both simplistic and monolithic. The representation of gay adolescents as necessarily suffering miserable and risk-filled lives, sometimes referred to as the suffering suicide script (Savin-Williams, 2005), was widely accepted as truth.

Few researchers at the time questioned whether these participants were representative of most youth with a same-sex sexuality. For example, the first empirical study included 60 young men, many of whom were urban hustlers living on the fringes of society (Roesler and Deisher, 1972). Predictably, many who visited gay bars recounted high numbers of sexual encounters (one reported over 3000 sex partners) and sought mental health counseling for depression and suicidality. Most felt rejected and disdained and few were hopeful about their destitute lives. A second study published over a decade later, again with a similarly exceptional youth population, explored the social stressors that placed gay (again, male) youth at high risk for physical and psychological distress (Remafedi, 1987). Many boys reported emotional difficulties, poor grades, school truancy, substance abuse, sexually transmitted infections, crimes, suicide attempts, and hospitalization for psychological reasons. Despite this morbid profile, Remafedi concluded that it was 'unlikely' that his adolescent sample was biased because the mere act of negotiating a gay or bisexual identity during adolescence placed one at risk for dysfunction.

The legacy of this early research has left an indelible impression on the minds of public policy makers, health care providers, educators, parents, and youth: being young and gay is intrinsically risky with little hope for a positive outcome. Although the research data were embraced to buttress arguments for the inclusion of gay youth in health care services and as a protected class, slow to emerge has been research that challenges this single trajectory characterization.

During the past two decades, scholars from other disciplines (particularly the biological sciences and developmental and

social psychology) have addressed normative gay development with representative samples of youth. As a result, an empirically based portrait has emerged, which is more nuanced and complex. It is these data that we present in this article.

Domains of Sexuality and Romance

Sexuality has various domains or expressions, including erotic attraction or fantasy (sexual orientation), behavior (genital contact), cognitive (sexual identity), and, perhaps, emotion (infatuation, romance). The first three are most frequently used by researchers when they define a sample as being of gay youth, but the four are imperfectly related to each other. For example, a young man might be sexually attracted to both sexes (bisexual sexual orientation), have sex only with females (heterosexual behavior), develop crushes on men (homosexual romance), and identify as heterosexual because he sees this pattern as the only option given by his family and religious community. A young woman with homoerotic desires (homosexual sexual orientation) might have sexual experiences with both sexes (bisexual behavior), romantic relationships with men (heterosexual romance), and a lesbian identity for ideological reasons. Depending on the domain assessed, radically different conclusions about a youth's sexuality can be reached.

Sexual orientation refers to the balance of erotic thoughts, attractions, feelings, and fantasies that one has for members of a particular sex or multiple sexes. Likely the consequence of genetic variation or the prenatal environment and its effects on brain development and sexual orientation are probably present from birth and as such is relatively immutable, stable, and resistant to conscious control. Most commonly, sexual orientation is assessed by asking participants about their sexual attractions or fantasies as represented by a scale ranging from exclusively male to exclusively female. Self-report measures, however, can be problematic because individuals are not always aware of their sexual orientation or are unwilling to disclose it.

One potential resolution is to use biometric measures that capture automatic or unconscious responses. These assessments can mediate the distortions of data from participants who are unaware of their sexuality or who are concerned about social stigma and hence report their sexuality in what they perceive to be a socially desirable manner. Psychophysiological measures of sexuality include genital arousal, functional magnetic resonance imaging (fMRI), electroencephalography, eye dilation, and eye tracking. The most common of these measures is genital arousal that assesses penile and vaginal response to sexual stimuli. More recently, fMRI brain imaging and eye tracking/dilation have been used to identify neural and attentional correlates of sexual arousal networks that differentially respond to preferred and nonpreferred sex stimuli. As opposed to traditional self-report measures of sexuality, the possible benefits of these two include the assessment of a wide range of participants across ages and cultures (e.g., non-English speaking).

Sexual behavior refers to the genital activities in which an individual engages. Definitions of what constitutes sex vary across individuals, cohorts, and societies and can refer to disparate activities such as vaginal or anal intercourse, oral/genital contact, and any manner of genital/genital contact between two or more people. The reference to sexual behavior includes proportion directed to each sex (i.e., ranging from exclusively females to exclusively males) and frequency of activity (e.g., 'daily,' 'within the past month,' 'once in the past year,' to 'never had'). A challenge to accurate measurement of sex is that cultural factors such as social class, geographic region, religious affiliation, and race/ethnicity influence the choice of sexual behaviors and partners and these might not be consistent with an underlying sexual orientation.

Sexual identity is a socially constructed label self-chosen to represent an array of potential sexual and romantic situations. Sexual identities are limited to those available within a given culture and historic time and as such the interpretation of particular labels, and even the specific identity terms used, varies. For some, a sexual identity is based on sexual feelings, attractions, behaviors, or some combination of these. Compounding these complex issues, some contemporary youth reject traditional sexual identity labels altogether, including the terms lesbian, gay, and bisexual, in favor of specific descriptions that better represent their sexuality and, frequently, gender expression (e.g., mostly straight, sexually fluid, pansexual, gender queer). Other youth reject sexual categories because they loathe being boxed in or simplistically designated by their sexuality.

Infatuation is an essential aspect of what might best be considered, *romantic orientation*. Infatuation or crush refers to an emotional, irrational state of being in which an individual feels blissfully passionate about another and subsequently believes she or he is 'in love' with the person. The feelings are generally considered to be short term, intense, and additive in quality as an individual becomes obsessed with increasing occurrences by the infatuation she or he feels. The object of desire can be unattainable (e.g., a rock star, a popular upperclassman, a straight friend), is unaware of the infatuation, or fails to reciprocate the feelings. The relationship between infatuation and romantic love is thought to be orthogonal or, perhaps, related to the extent that infatuation frequently precedes romantic love. The way in which a crush can grow into mature love and intimacy is unknown, especially among same-sex attracted youth. Although infatuation is believed to be derived from libidinal attraction, the relationship between sexual and romantic orientation is essentially an unchartered territory, in large part because the latter, especially as expressed in romantic behavior and romantic identity, has seldom been defined or assessed.

Recent research illustrates the intricacies of these various sexual and romantic indicators for the current generation of youth. For example, the majority of those who engages in same-sex behavior identifies as heterosexual and the majority of nonheterosexually identified individuals experience heterosexual behavior and romance. So, too, about 1 in 10 heterosexually identified youth report same-sex attraction or fantasy, and some youth fall in love with someone to whom they are not sexually attracted (as well as the reverse). Thus, it is most reasonable to conceptualize sexual orientation as existing along a continuum with exclusively heterosexual individuals and exclusively homosexual individuals at the poles and nonexclusive individuals – those who vary both in degree (e.g., 70% other-sex and 30% same-sex attracted) and domain (e.g., sexually attracted to girls and romantically attracted to boys) of their sexuality – somewhere in between.

These discrepancies have been attributed largely to the effects of external stigma and the resulting sexual prejudice that some

youth internalize. Although few aspects of the self are as essential as sexual and romantic longings, youth may struggle to self-recognize, disclose to others, and fully embrace their underlying same-sex sexuality. They may be unaware or distrusting of their sexual desires because they have limited exposure to nonheterosexual alternatives, conservative family attitudes, attenuated libido, or excessive self-monitoring of expressed behavior. As a result, they may vacillate between rejecting a socially stigmatized sexual identity and being true to self. These irregularities appear to be decreasing for the current cohort of youth, perhaps because their generation is far more comfortable with sexual and gender diversity during a time when society in general has dramatically shifted its views toward sexual prejudice and nonheterosexuality (Savin-Williams, 2005).

After considering the deceptively simple construct of how many gays exist, we review what is known about developmental milestones that together constitute the process of identifying oneself as a gay. These milestones include early feelings of being different from peers, the onset of same-sex attraction, sexual questioning, first sexual experience, recognition and self-labeling, disclosure to others, first romantic relationship, and self-acceptance.

How Many Gays?

Beyond verbal admission that one's sexual orientation is not heterosexual, the determination of sexual orientation is unexpectedly complex, depending on the way in which it is defined. Consistent with issues raised previously, it is necessary to distinguish among three related sexual domains: sexual orientation, sexual behavior, and sexual identity. Data from four waves (12 years) of the National Longitudinal Survey of Adolescent Health (Add Health) corroborate this point. Prevalence rates of nonheterosexual youth vary between 1 and 20%, and are highest among young women, those who identify as mostly heterosexual, and those who acknowledge same-sex attraction but do not necessarily identify as gay/lesbian (Savin-Williams et al., 2012). By the age of 28, 2 and 4% of Add Health men and women, respectively, identified their sexual orientation as gay, lesbian, or bisexual; however, more than 6% of men and nearly 20% of women claimed that they were not exclusively heterosexual (see table below). Thus, not all youth with same-sex sexuality necessarily identify as gay, act on their sexual desires, or disclose the nature of their sexuality to others, perhaps especially not to adults perceived by youth as having narrowly defined ideas about what constitutes sexual identity.

	Straight	*Mostly straight*	*Bisexual*	*Mostly gay/lesbian*	*Gay/lesbian*
Women	80%	16%	2%	1%	1%
Men	94%	4%	1%	1%	2%

One Trajectory or Many Trajectories?

Coming-out stage models, many of which were proposed during the first two decades of research on gay youth, characterize development as an invariant, universal progression from initial unawareness and confusion to eventual identity pride and synthesis (Cass, 1979; Troiden, 1979). Despite the widespread popularity of stage-based coming-out models, linear and idealized portrayals have not been empirically substantiated. For example, whereas identifying as gay usually precedes disclosure to others or same-sex dating, some youth have a committed same-sex romantic relationship prior to self-labeling. Youth may identify as gay and remain a virgin, although stage theories often posit engagement in sexual activities with others of their sex prior to self-identification. In contrast to previous generations, it is not unusual for contemporary youth to forego sexual confusion or to be without fear that they are mentally ill for being gay; almost from the start, they calmly believe that it is just who they are.

Because recent research has documented the variability of developmental milestones, most scientists are now wary of assumptions about what are normal or typical developmental trajectories. The coming-out process is best understood as an individual's achievement of developmental sexual and romantic milestones that increasingly occur without respect for tradition, timing, and order.

Feeling Different

The vast majority of same-sex attracted youth recalls growing up feeling different or manifesting behavior or interests discrepant from those of their peers. This sense of divergence from other children can begin as early as first memories and tends to intensify with age. The recalled aspect of differentness frequently involves gender expression – having the personality, interests, or behavior of the other sex. In one study, by the age of 8 years, two-thirds of youth who later identified as gay or lesbian were considered gender atypical by others and over half reported that their parents said that they were gender inappropriate; half of such parents attempted to change the youth's gender presentation (D'Augelli et al., 2008).

Feeling different because of perceived gender inadequacy often perseveres throughout adolescence and into adulthood and relates to deficiencies in gender conformity (e.g., low femininity in women), elevations in gender nonconformity (e.g., high masculinity in women), or both. In one study, men reported that their initial feelings of being different were determined less by inappropriate sexual longings than by the shame generated by their feelings of gender ineptitude (Troiden, 1979). By adolescence, however, another source of differentness emerges – feeling sexually different because of homoerotic desires, sexual interactions with the same sex, and negligible erotic attraction for the other sex.

Although the correlation between same-sex attraction (addressed below) and gender-atypical behavior is not perfect, many studies document a strong association (Cohen, 2002). Nonetheless, some heterosexual youth report gender-atypical interests (being a male dancer) and some homosexual youth report gender-typical pursuits (being a football quarterback). Many youth find it easier to suppress the behavioral expression of their homoeroticism (a lesbian eschewing sex with girls) than to conceal gender nonconformity (a lesbian walking in a masculine manner), possibly because the latter is less subject to conscious control.

Same-Sex Attraction

The age range during which youth first experience same-sex attraction is broad, extending from first memories onward, though it usually first occurs several years before the onset of puberty (during adrenarche, 7–9 years of age). Although supporting empirical evidence is limited, a decline in the recalled age of first attraction of about 4 years, particularly among girls, has been observed since the 1980s. Whereas previous cohorts of boys reported awareness of their sexual attraction earlier than girls, this is less true today, likely the result of young women's growing acceptance of all aspects of their sexuality (Diamond and Savin-Williams, 2009; Rosario et al., 1996).

Relative to boys, girls more frequently remember their first same-sex attraction as romantic rather than sexual. For example, in one study girls were more likely to recall their first same-sex attraction as an emotional attachment or crush; boys, as a sexual thought, arousal, or behavior (Savin-Williams and Diamond, 2000). Whether this reflects differences in biology, socialization, or both is unknown. However, it is frequently reported that female sexuality is more dependent on environmental, situational factors, such as passionate, interpersonal relationships with women and university courses with strong pro-feminist readings and lectures (Diamond, 2008; Peplau, 2003).

As a predictor of sexual identity, the presence of sexual attraction for one or the other sex might not be the best indicator. Many heterosexually identified individuals, especially women, report some degree of same-sex attraction, prompting several researchers to argue for a revised sexual identity taxonomy that would include the label *mostly straight* (Thompson and Morgan, 2008; Savin-Williams and Vrangalova, 2013). In addition, as many as 80% of lesbians and 60% of gay men experience opposite-sex attraction and fantasy (Rosario et al., 1996). Developmentally, sexual attractions consistent with one's sexual orientation emerge earlier than attractions that are not; regardless, the onset of sexual attraction awareness may be unrelated to the timing of other developmental milestones, such as coming out to self and others, dating, and accepting one's sexuality.

Overall, most youth report that they are not disturbed by their early homoeroticism because it feels natural, enjoyable, and a reflection of who they are (Savin-Williams, 2005). When these desires fail to dissipate with time, suspicion that one might not be heterosexual can prompt an interval of sexual questioning, uncertainty, or elaborate denial that may be fleeting or last a lifetime. Increasingly, however, contemporary gay/lesbian youth are less likely to question their sexuality because they know almost simultaneously that same-sex attraction signifies a sexual minority status. Sexual identity resolution may be more confusing and problematic for individuals who are attracted to both sexes or are primarily heterosexual yet experience same-sex attraction or engage in same-sex behavior.

Questioning Assumed Heterosexuality

When same-sex attraction becomes especially frequent or undeniably intense, youth who previously considered themselves heterosexual might wonder whether the attraction has meaning for their sexuality – that is, whether they might be less than totally heterosexual. Such questioning can intensify over the course of adolescence and lead to clarity or confusion, which in turn elicits divergent resolutions, not all of which include the assumption of a sexual minority identity label. Perhaps they believe it is a phase or experimentation, or that everyone has homoerotic feelings; these views allow individuals to minimize or discount their sexual desires. Others might consider their sexual fantasies – but not themselves – to be gay, or, if they are willing to assume more personal responsibility and acceptance, they may concede that they are slightly gay.

Having any degree of heterosexual interest and behavior might convince some youth that they are, or can be in the future, totally straight, as long as they try hard enough and maintain focus on the other sex. This belief allows them to preserve hope for an eventual heterosexuality even as they acknowledge and perhaps derive pleasure in their same-sex sexuality. Increasingly, youth who accept their same-sex attraction as memories previously considered innocuous come to assume greater meaning and support the naturalness and intractability of their homoeroticism. For youth with some heterosexual interest, the result might be a self-label as mostly straight, bisexual, or sexually fluid. In each of those three categories, more young women than men reside and, in most surveys, their prevalence exceeds those who identify as lesbian, usually by a factor of two or three (Savin-Williams et al., 2012).

Sexual Behavior

Although the average age of first same-sex encounter (specific types of sexual activity are seldom given) is around 14 years for both sexes, the onset of sexual activity can precede, co-occur, or follow self-recognition as a gay, lesbian, or bisexual person. That is, self-identified gay virgins do exist and sexual minority youth, similar to heterosexual youth, may delay sexual activity until they are in love or are in a committed relationship. Corresponding to the increasingly younger age of first awareness of a nonheterosexual identity, the majority of contemporary youth now embraces a same-sex identity prior to experiencing same-sex behavior (Savin-Williams, 2005).

Most same-sex oriented adolescents do not participate exclusively in same-sex behavior. Reasons are varied, including a desire to experiment or have fun, pressure while dating an opposite-sex peer, alcohol intoxication or substance use, and genuine heteroerotic desire. Although some straight-identified youth engage in same-sex encounters, those who are primarily homoerotic pursue such interactions with greater regularity and zeal and reportedly derive greater physical and emotional satisfaction from them (Vrangalova and Savin-Williams, 2010).

Same-sex oriented young men are more likely than same-sex oriented young women to report same-sex behavior, to have a greater number of sex partners, and to engage in such activities prior to heterosexual experiences. The first same-sex encounter among young women typically occurs with a friend or someone they are dating, but for young men it is more likely to take place in the context of a purely sexual encounter, despite stated preferences for relationship sex

(Savin-Williams, 2005). Although the documented sexual fluidity among same-sex oriented women may explain their greater participation in heterosexual encounters, they may also be exposed more frequently than young gay men to opposite-sex sexual invitations. Nonetheless, the majority of both sexes engages in heterosexual sexual activity, typically within a dating or friendship relationship (Diamond, 2008; Diamond and Savin-Williams, 2009).

Self-Identification

Coming out to oneself – that is, recognizing that one's same-sex attractions, crushes, fantasies, and behaviors are meaningful and signify a nonheterosexual sexuality – is a milestone that some youth vividly recall. This is especially true for those who experience the realization as either terrifying or relieving, and is less the case for youth who have always known what their homoeroticism means for their sexuality and personal identity. The latter trajectory is becoming increasingly common as contemporary cohorts are appreciating and naming their sexuality while still in middle school or high school or, in some instances, even while in elementary school (Savin-Williams, 2005).

Whereas young men may take months or even years to process the meaning of their same-sex attraction and progress from comprehension to self-identification as gay/bisexual, young women typically make the transition to lesbian/bisexual far more quickly. It is not uncommon for women to recognize, name, and disclose to others their homoeroticism within a few days or months, or even shortly after developing their first crush on another young woman (Savin-Williams and Diamond, 2000). Bisexuals may take longer to reach these developmental milestones because their sexual identity is somewhat more fluid over time, they have fewer sources of support, or they face disbelief from others about their sexuality (Rust, 2002).

Transcending traditional labels of same-sex sexuality, some youth are redefining sexual identity categories or rejecting them altogether. That is, a growing constituency of youth is refusing categorical sex-based labels, accepting notions of sexual and gender fluidity (e.g., pansexual, heteroflexible), and describing their sexuality as an interaction between their erotic preferences and their gender expression (e.g., transboi). Sexual self-definitions may be dynamic and unspecified (e.g., sexually fluid), a conduit for amalgamating multiple identities (e.g., black dyke who loves men), or a rejection of sexual taxonomy altogether (e.g., unlabeled, 'I love Robert').

Other youth eschew sexual labels because they fear negative family or social repercussions or they prefer to emphasize other aspects of their life that they consider to be more personally meaningful, such as their religion, race/ethnicity, or geographic region. Still other youth reject traditional sexual classifications because they associate particular stereotypes with those labels (lesbians as butch; gay men as sissies) that they find abhorrent. Consequently, youth might acknowledge their same-sex attraction and behavior but deny that they are gay, lesbian, or bisexual.

Diamond's (2008) prospective longitudinal investigation found that many same-sex attracted women considered their sexuality to be fluid and that most had experienced several changes in their sexual identity (but not sexual orientation) since childhood. Quite common was the belief that sexual labels are limiting and that individual characteristics trump gender in determining romantic interest. Although some young women with same-sex interest or activity eventually embraced a heterosexual label, they insisted that their same-sex attractions persisted.

Disclosure

Coming out to others can be a lifelong process during which same-sex attracted youth face decisions regarding whether, when, and how to reveal their sexuality to family, friends, coworkers, and new acquaintances. Disclosure to others can be a positive event that facilitates personal integrity, identity synthesis, and psychological health or a negative event that engenders ridicule and threats to physical safety. For young men, disclosure usually first occurs several years after coming out to self. As noted previously, the gap between the two for young women is somewhat less. The recipient of first disclosure for both young men and women is a best female friend, though it is no longer rare to first reveal one's sexuality to a favorite sibling, another same-sex attracted youth, or a supportive heterosexual male friend. When the object of first disclosure is carefully chosen, the reaction is almost always positive, even celebratory. Indeed, today's generation of MTV-viewing, Internet-consuming young adults routinely consider sexual diversity to be normal and positive and thus are typically nonplussed when peers come out (Savin-Williams, 2005). It is not uncommon for friends to progress quickly from mild surprise (or amused confirmation) to resolute determination that they will find a sexual or romantic partner for their friend.

Unique challenges, however, may confront youth who are foreign nationals temporarily residing in North America, members of conservative racial/ethnic or social class groups, or religious members of orthodox communities. For cultural reasons they may be unable or unwilling to live a same-sex active or oriented sexually open and integrated life, and thus their priority is to avoid real or feared widespread social admonishment. Although little is known about these mostly invisible youth, they likely face unique barriers in coming out to self and others as they attempt to reconcile their atypical, unacceptable sexuality through the lens of their primary family and social community (Greene, 1994). Beleaguered, they may be left alone to navigate their sexual development within the context of a mostly white gay subculture in which disclosure problems related to racism, classism, or religion are not acknowledged or challenged. For other youth, the considerable support they receive from their primary family or community outweighs the advantages of coming out, at least during adolescence and young adulthood.

Another subgroup of sexual minority youth who encounter particular hurdles disclosing to others are gender-nonconforming individuals whose cross-sex behavior and interests deviate significantly from those of their peers. They may be assumed to be gay because of their gender expression, perhaps long before they realize the connection themselves. The harassment they face from peers who are antagonistic

toward anyone who fails to act as a boy or girl 'should act' may delay their own self-awareness because such insight would acknowledge the truthfulness of the hated peer accusations.

Despite these barriers, disclosing one's sexuality to others can be an act of empowerment and self-affirmation, a means for obtaining personal authenticity, social support, and interpersonal closeness. It frees those with same-sex attraction from the trepidation and feared rejection they long anticipated. Reasons for not disclosing include fear of the unknown, avoidance of disappointing others (especially parents), and the belief that they will become irrevocably boxed in ('what if I later discover I am not gay?'). It is noteworthy that today's youth are increasingly worried about being chastised by friends who are insulted because they were not told earlier ('you did not trust me!').

Young people may delay disclosing to parents not because parents are deemed unimportant, but precisely because of their importance. Thus, youth want to be absolutely certain about their sexuality before they make this final, most highly significant, disclosure. Some delay coming out to parents because of the general discomfort youth of all sexual orientations experience when discussing sexual matters with parents; others, because they fear emotional devastation and financial ruin (e.g., lost academic financial support) should parents reject them. Surprisingly, even gay-affirming parents with gay friends and progressive politics may be among the last to know about their child's sexuality because their daughter or son wishes to avoid the anticipated parental enthusiasm and encouragement to bring home a romantic partner sooner than she or he would like.

Nevertheless, adolescents who come out to a parent (the mother is usually first) while living at home seldom face rejection or expulsion. Parents might not be initially thrilled with having a gay child because they fear for the child's safety, are concerned about what their neighbors or parents will think, or worry that their dreams of grandchildren are dashed. However, despite stereotypes and highly publicized instances of abuse and condemnation, youth seldom receive an ongoing, severely negative response. Parental reactions vary from celebration to rejection, but with time most eventually accept what they had long suspected (Savin-Williams, 2001). Similar to their children, parents have been helped by normalizing media portrayals and the coming out of friends and coworkers.

Romantic Relationships

A desire for a committed romantic relationship is articulated by many sexual minority youth. Until recently, such a relationship was considered unobtainable until high school graduation when living away from home would render them somewhat immune from the judgmental and punitive eyes of parents and hometown peers. Thus, the first same-sex date frequently occurred while a freshman in college or when entering the workforce. Although most youth continue to postpone dating until such times, the growing cultural acceptance of sexual diversity and expanding peer support (e.g., school-based Gay-Straight Alliance clubs) are increasingly motivating young people to come out while still in high school. Consequently, same-sex dating and taking a same-sex partner to the prom are no longer extremely rare occurrences.

This growing visibility of same-sex dating and romance was evident in a study documenting that sexual minority youth did not differ from heterosexuals in number of romantic relationships, especially for those who were most out and thus most able to identify and attract potential dating partners (Diamond and Lucas, 2004). Youth in romantic relationships were also less likely to be depressed or anxious (cause and effect were not addressed). Nonetheless, sexual minority youth were more likely than heterosexuals to fear not finding the type of relationship they wanted, and sexual minority young men, in particular, faced obstacles in developing highly intimate same-sex relationships during their teen years.

As with heterosexual dating, few of these early relationships endure, yet they afford important opportunities for psychosocial exploration, integration of emotional and sexual intimacy, and acquisition of a positive identity. Locations and strategies for finding partners mirror those of heterosexuals, although nonheterosexuals also face unique challenges. For example, sexual minority youth must search within a smaller pool of potential partners containing many who are closeted and perhaps unidentifiable. Consequently, they may convey interest toward individuals they erroneously believe to be romantically or sexually available, creating possibly awkward, if not dangerous, repercussions.

Self-Acceptance and Synthesis

Considered in many coming-out identity models to be among the final developmental stages, acceptance of one's sexuality and its synthesis with other aspects of personal identity can be achieved at any point in the life course. Those who attain sexual identity synthesis recognize the importance of their sexuality but no longer experience it as all-consuming or overarching; it is simply one of many aspects of self. Thus, a woman who earlier defined herself as a lesbian novelist would now see herself as a novelist who happens to be attracted to women.

Despite earlier stereotypical representations of the tragic lives of gay youth, contemporary sexual minority youth are as content with their lives as are heterosexuals (reviewed in Savin-Williams, 2005). For example, two-thirds of lesbian and gay adolescents reported feeling anywhere from indifferent to very good about their sexuality; less than one-tenth expressed a desire to not be gay. If given the option, most would not take a magic pill to make them straight. Some youth are content with their same-sex orientation without identifying as a sexual minority; they might never identify as gay yet nonetheless accept and integrate their same-sex attraction with other domains of their life. Whether youth identify themselves as a sexual minority is less important than acquiring the ability to integrate their sexuality with other essential elements of life. The result is authenticity, peace, and positive social interactions.

Conclusion

The twenty-first century has witnessed a prolific rise in gay youth research and social exposure. The media are increasingly celebrating their culture and characterizing them as they would

any other youth: the girl/boy next door. Charting their development by identifying and describing milestones is one way in which their presence is validated and normalized. It escorts them into the realm of human development rather than the fringes of child development gone awry.

See also: Adolescence, Sociology of; Adolescent Sexual Risk; Friendship During Adolescence and Cultural Variations; Heterosexism and Homophobia; Identity in Childhood and Adolescence; Self and Identity Development During Adolescence across Cultures; Sexual Orientation and the Law; Sexual Orientation: Biological Influences; Sexual Orientation: Genetic Aspects; Sexual Orientation: Historical and Social Construction.

Bibliography

Cass, V., 1979. Homosexual identity formation: a theoretical model. Journal of Homosexuality 4, 219–235.

Cohen, K.M., 2002. Relationships among childhood sex-atypical behavior, spatial ability, handedness, and sexual orientation in men. Archives of Sexual Behavior 31, 129–143.

Cohen, K.M., Savin-Williams, R.C., 2012. Coming out to self and others: developmental milestones. In: Levounis, P., Drescher, J., Barber, M. (Eds.), Working with Lesbian, Gay, Bisexual, and Transgender People: Basic Principles and Case Studies. American Psychiatric Publishing, Washington, DC, pp. 17–33.

D'Augelli, T.R., Grossman, A.H., Starks, M.T., 2008. Gender atypicality and sexual orientation development among lesbian, gay, and bisexual youth: prevalence, sex differences, and parental responses. Journal of Gay & Lesbian Mental Health 12, 121–143.

Diamond, L.M., 2008. Sexual Fluidity: Understanding Women's Love and Desire. Harvard University Press, Cambridge, MA.

Diamond, L.M., Lucas, S., 2004. Sexual-minority and heterosexual youths' peer relationships: experiences, expectations, and implications for well-being. Journal of Research on Adolescence 14, 313–340.

Diamond, L.M., Savin-Williams, R.C., 2009. Adolescent sexuality. In: Lerner, R.M., Steinberg, L. (Eds.), Handbook of Adolescent Psychology, third ed. Wiley, New York, pp. 479–523.

Greene, B., 1994. Ethnic-minority lesbians and gay men: mental health and treatment issues. Journal of Consulting Clinical Psychology 62, 243–251.

Peplau, L.A., 2003. Human sexuality: how do men and women differ? Current Directions in Psychological Science 12, 37–40.

Remafedi, G., 1987. Adolescent homosexuality: psychosocial and medical implications. Pediatrics 79, 331–337.

Roesler, T., Deisher, R.W., 1972. Youthful male homosexuality: homosexual experience and the process of developing homosexual identity in males aged 16 to 22 years. The Journal of the American Medical Association 219 (8), 1018–1023.

Rosario, M., Meyer-Bahlburg, H.F.L., Hunter, J., Exner, T.M., Gwadz, M., Keller, A.M., 1996. The psychosexual development of urban lesbian, gay, and bisexual youths. Journal of Sex Research 33, 113–126.

Rust, P.C.R., 2002. Bisexuality: the state of the union. Annual Review of Sex Research 13, 180–240.

Savin-Williams, R.C., 2001. "Mom, Dad. I'm gay." How Families Negotiate Coming Out. American Psychological Association, Washington, DC.

Savin-Williams, R.C., 2005. The New Gay Teenager. Harvard University Press, Cambridge, MA.

Savin-Williams, R.C., Diamond, L.M., 2000. Sexual identity trajectories among sexual-minority youths: gender comparisons. Archives of Sexual Behavior 29, 419–440.

Savin-Williams, R.C., Joyner, K., Rieger, G., 2012. Prevalence and stability of self-reported sexual orientation identity during young adulthood. Archives of Sexual Behavior 41, 103–110.

Savin-Williams, R.C., Vrangalova, Z., 2013. Mostly heterosexual as a distinct sexual orientation group: a systematic review of the empirical evidence. Developmental Review 33, 58–88. 10.1016/j.dr.2013.01.001.

Thompson, E., Morgan, E.M., 2008. "Mostly straight" young women: variations in sexual behavior and identity development. Developmental Psychology 44, 15–21.

Troiden, R.R., 1979. Becoming homosexual: a model of gay identity acquisition. Psychiatry 42, 362–373.

Vrangalova, Z., Savin-Williams, R.C., 2010. Correlates of same-sex sexuality in heterosexually-identified young adults. Journal of Sex Research 47, 92–102.

Gays, Lesbians, and Sports: No Longer on the Sideline

Eric Anderson, University of Winchester, Winchester, UK
Mark McCormack, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

High levels of cultural homophobia meant that early research on homophobia in sport did not collect data with sexual minority athletes who were public about their sexuality. However, as homophobia has decreased in Western cultures, there has been a proliferation in research on openly gay athletes. Accordingly, this article investigates the experiences of openly gay male and lesbian athletes in ostensibly heterosexual sports. The authors show that while sport was founded in exclusionary practices, changing cultural norms regarding sexual minorities have extended to sporting cultures.

Early research on homophobia in sport did not collect data with sexual minority athletes who were public about their sexuality (e.g., Messner, 1992; Pronger, 1990); this is for the reason that at the time, openly gay or lesbian athletes were unheard of at collegiate and professional levels of sport (Anderson, 2005). However, in more recent years, as homophobia has decreased in Western cultures (Loftus, 2001; Weeks, 2007), there has been a proliferation in research on openly gay athletes (e.g., Anderson, 2002, 2005). Accordingly, this article investigates the experiences of openly gay male and lesbian athletes in ostensibly heterosexual sports. The authors show that while sport was founded in exclusionary practices, changing cultural norms regarding sexual minorities (particularly for youth) have extended to sporting cultures. Matters are much improved for sexual minorities in sport today in the West, although one recognizes a dearth of research into the experiences of bisexual and transgendered athletes in the West or of sexual minorities more generally in sports in non-Western countries.

Foundation of Sport as a Heteromasculine Preserve

Although there are various purposes of organized, competitive team sport participation in the West, a consistent finding is that sport has served as a social institution organized around the political project of defining certain forms of masculinity as acceptable, while denigrating all others (Messner, 1992; Pronger, 1990). Contact team sports, like American football and rugby, have been recognized as enshrining this politically preferred form of masculinity, where participation is made near-compulsory through boyhood culture (Anderson, 2005; Connell, 1995).

Sport was not always held in such esteem, with a more *laissez-faire* approach before the turn of the twentieth century. In the past, sport was neither compulsory nor culturally valued. While some sport was played, few people had the time, desire, or equipment to play in organized, competitive leagues. The compulsory nature of sport for male youth is instead a product of the fear of feminization of young boys that arose during the turn of the twentieth century (Kimmel, 1996).

There are a number of reasons for this increased focus on sports, including that sport taught boys to become good soldiers as it socialized them into being docile to authority, to sacrifice for the sake of team, that physical fitness was a moral responsibility to society, and to value the greater good (Mangan, 2000; Sage, 1990). These characteristics were also expected behaviors of factory workers, and sport was thus also valued by the owners of industry. However, sport was also valued because it socialized boys into particular, esteemed gendered and sexual identities (Anderson, 2009; Bryson, 1987).

Concerns about gender identity arose partly out of social fears based on Freud's (1905) notion that homosexuality is the product of boyhood socialization into femininity and the absence of a father figure. Given that factory work, much expanded during the second industrial revolution, meant fathers were away from their children for extended periods of time (Cancian, 1987), it was feared that without this masculinizing influence, male youth would grow up to be less attuned to the heterosexuality necessary to produce future workers. Accordingly, in most Western nations, alongside muscular Christianity and the Boy Scouts movement, sport served as a vessel for the indoctrination of young boys into a desire to be associated with heterosexual masculinity (Anderson, 2009; Connell, 1995).

Competitive team sport culture was therefore constructed to reproduce society's gendered and sexual values, myths, and prejudices. Throughout the decades of the twentieth century, no other institution naturalized heterosexual more than competitive sport (Anderson, 2005). As a highly segregated, homophobic, sexist, and misogynistic gender-regime, sport not only contributed to the gender order of men dominating women, but also reproduced a conservative form of heterosexual masculinity that has considerable costs for gay men. This is why women were excluded from sport for so long; women who competed equally alongside men would disrupt the myth of men's athleticism and women's frailty (Burton-Nelson, 1994).

Although male-dominated for much of the twentieth century, women began accessing the institution in the late 1960s, and in the United States, women's sporting participation has rapidly increased since the 1972 passage of Title IX (McDonagh and Pappano, 2008). Today, women are well represented as participants at the lower levels of sport. However, women's representation dramatically decreases in the professional ranks; particularly among the sports that capture media interest (Burstyn, 1999).

Historically, women participating in sport have been subject to homosexual suspicion, in part because of the number of high-profile female athletes who came out as lesbian (Griffin, 1998). However, heterosexual women began also playing sport in large numbers in the 1980s, and their participation began to heterosexualize the sport for women, requiring its participants to boast their heterosexual capital, and silence their lesbian teammates. Even so, matters for openly gay male athletes during the 1980s were far worse. Compared to women, gay men have made very little progress in opening mainstream sports to gay participation (Anderson, 2005; Pronger, 1990) up through this time. Yet by this decade, sport had become institutionalized, even compulsory, within public school systems. It was also this decade that saw the apex of homohysteria in sport.

Homohysteria and Sport

With sport's traditional function of masculinizing and heterosexualizing boys – and subsequently marginalizing effeminacy – it is unsurprising that homophobia and antifemininity have been commonly employed among athletes in the construction of gender-normative performances (Bryson, 1987; Sabo and Panepinto, 1990). This was particularly true of sport (in the Western world) in the 1980s. Here, sport took on renewed importance for boys and young men as it was a central tool in heterosexualizing men in a culture that is labeled as 'homohysteric' (Anderson, 2009, 2011a).

In describing the construction of masculinities, homohysteria refers to a "homosexually-panicked culture in which suspicion (of homosexuality) permeates" (Anderson, 2011a: p. 7). In order for a culture of homohysteria to exist, three factors must coincide: (1) the mass cultural awareness that homosexuality exists as a static sexual orientation within a significant portion of the population; (2) a cultural zeitgeist of disapproval toward homosexuality; and (3) disapproval of men's femininity or women's masculinity, as they are associated with homosexuality.

There are three stages related to homohysteria that a culture (might) move through. The first is prehomohysteria, and the authors suggest this can also be called 'homoerasure.' Here, the culture is extremely homophobic but homosexuality as an identity is effectively erased. Importantly, because the citizenry do not believe that homosexuality exists in the general population, homophobia does not regulate gendered behaviors (although it of course polices sexual behaviors). Exemplifying this, much of the Islamic world today, as well as throughout much of Africa, homosexuality is thought to be 'only' a Western phenomenon (Frank et al., 2010). So while, for example, homophobia is intensely high in Iran, in 2007, President Mahmoud Ahmadinejad claimed that, "…in Iran we don't have homosexuals like in your country." Accordingly, boys in Iran will have less need to distance themselves from cultural suspicion of homosexuality. It is this mass denial that homosexuals exist in large numbers that permits Iranian men to walk together in public holding hands. It is also likely that this prehomohysteric culture will become homohysteric if Western discourses of gay rights are adopted by gay people in those countries.

In contrast, in a homohysteric culture, there is widespread awareness that a significant percentage of the population can be gay and homophobia regulates gendered behaviors. In the United States, homohysteria occurred in the 1980s as the growing visibility of gay people, and a strengthening gay politic contested high levels of homophobia. This was heightened by an increasingly noisy fundamentalist Christianity that was opposed to and demonized homosexuality, alongside the acquired immunodeficiency syndrome epidemic that brought closeted people out of the closet as symptoms of the illness became visible. In this homohysteric, culture boys and young men (particularly those who were unmarried) needed to establish and establish themselves as heterosexual by aligning their gendered behaviors with idealized notions of masculinity. It is especially between the years 1983 and 1993 that boys in Western cultures needed to use sport in order to prove their hetero-masculinity (Pronger, 1990). This is because homosexuality is not readily visible (like gender or race); ostensibly, anyone can be gay (see also Ibson, 2002; Floyd, 2000 for discussion of this).

Given that men's masculinity is/was associated with heterosexuality in Western, industrialized cultures, boys in a culture of homohysteria are/were required to elevate their display of masculinity to prove that they are/were not gay. In other words, they used culturally endorsed sports to distance themselves from being socially perceived as gay. Participation in organized team sports is less important for the construction of heterosexuality in a culture where homosexuality is not believed to exist as a significant demographic of the population.

Because sport has been defined as "the last bastion of male domination" (Burton-Nelson, 1994: p. 6), homohysteria has created a problem not only for effeminate and/or gay men, but also for women. Although women challenge traditional gender boundaries in sport by simply taking part, only certain sports are seen as feminine-appropriate; mainly those that do not require physical contact. Writing specifically about rugby, Wright and Clarke (1999) suggested that women's participation "could therefore be expected to challenge fundamentally what it means to be a male and female" (p. 299). Hence, women playing male-dominated team sports, such as football and rugby, face particular confrontation and questioning about their involvement.

Hargreaves (1994: p. 171) argued that when women participated in male-dominated sports they faced "the greatest criticism and exposure to ridicule." Lenskyj summarized that the central issues concerning women's participation in sport were that, "femininity and heterosexuality (were) seen as incompatible with sporting excellence: either sport made women masculine or sportswomen were masculine from the outset" (Lenskyj, 1986: p. 95).

Griffin (1998) argued that in order to limit controversy about their participation, women have been shown to promote heterosexual (heterosexy) images and use overt homophobia as a way to socially distance themselves from being thought lesbian. The erasure of lesbianism through the promotion of femininity and heterosexuality is known as the 'apologetic.' Felshin (1974) argues that "because women cannot be excluded from sport and have chosen not to reject sport, apologetics develop to account for their sport involvement in the face of its social unacceptability" (p. 36). Davis-Delano et al. (2009) explain that apologetic behavior occurs in numerous

ways: creating a feminine image, or apologizing for on-field behavior, such as aggression (Ezzell, 2009). Even in recent years, Davis-Delano et al. (2009) show that apologetic behavior is still evident; specifically, looking feminine, appearing heterosexual, and apologizing for aggressive behavior.

Even early in the twenty-first century, Cox and Thompson (2001) found that female footballers were assumed to be lesbian because of their choice to play a traditional male team sport. Shire, Brackenridge, et al. (2000) describe the behavior of heterosexual women within one hockey team, saying, "They joked about the lesbian women in order to reinforce their heterosexuality to others" (p. 49).

Accordingly, the conditions described of homohysteria for men in sport also existed for women in sport. Just as a 1980s and 1990s culture of homohysteria presented a problem for both gay and straight men, the homohysteria of women's sport also created a problem for both lesbian and straight women. Women were understood as being pressured into presenting an image of hyperfeminine heterosexuality in order to gain and maintain public support of their new found sporting freedoms (Lenskyj, 2003). The silence and denial of lesbianism in sport permit stereotypes and discrimination to continue unopposed (Krane and Barber, 2003).

Navigating Homohysteria

Since sport has been a homohysteric culture for both men and women at the end of the last century, what does one know about gay men and lesbians who participated in sports during that period? The sheer numbers of lesbians interested in sport did enable their study (e.g., Griffin, 1998). Conversely for men, there is a dearth of research concerning the relationship between sport, masculinities, and homosexuality before the 1980s (see Garner and Smith, 1977 as a notable exception). This is likely because gay athletes had not yet begun to emerge from their sporting closets, nor did they exist openly within the sport-related occupational industry (Anderson, 2005). For example, when Pronger (1990) studied closeted Canadian gay athletes in the late 1980s, he was unable to find men who were out to their teammates. Whether participating in individual sports (e.g., tennis, swimming, and running) or team sports (e.g., football, basketball, and rugby), there were few openly gay athletes in the Western world to design a study around.

Gay male athletes were (and remain) underrepresented in this period for a number of reasons. Anderson (2005) describes the selection hypothesis as gay men avoiding these forms of sport because of their homophobia (Zipp, 2011). However, because most male youth are indoctrinated into sport before awareness of their sexual orientation, it is likely that this hypothesis combines with the silence hypothesis to keep the numbers of openly gay men in sport low.

The silence hypothesis dictates that athletes of the time likely remained closeted because they assumed that the high degree of homophobic discourse, alongside their teammates vocalized opposition to homosexuality, and the high levels of societal homophobia, indicated that they would have a troubled experience coming out (Woog, 1998). Messner (1992) confirmed sport's homophobic disposition: "The extent of homophobia in the sports world is staggering," he wrote. "Boys (in sport) learn early that to be gay, to be suspected of being gay, or even to be unable to prove one's heterosexual status is not acceptable" (p. 34). These attitudes also extended into recreational level sporting leagues, even in liberal cultures. Discussing the Netherlands, Hekma (1998) wrote, "Gay men who are seen as queer and effeminate are granted no space whatsoever in what is generally considered to be a masculine preserve and a macho enterprise" (p. 2).

This paradigmatic view was supported by the quantitative research on university athletes in the United States, too. For example, in 2001, Wolf Wendel, Toma, and Morphew found that white male athletes exhibited disproportionate degrees of homophobia compared to their attitudes toward racial minorities. Hence, sport has been widely recognized as an institution that promotes heterosexuality over homosexuality.

Shifting Intersections of Youth and Homophobia

By the end of the first decade of the twenty-first century, however, studies began reporting a rapidly decreasing level of homophobia among youth (Keleher and Smith, 2012; Kozloski, 2010; Savin-Williams, 2005) even in men's team sports at the collegiate level (Anderson, 2009; McCormack and Anderson, 2010; Southall et al., 2011). Research on British high school students found that heterosexual boys not only espoused progay attitudes, but also condemned homophobia as unacceptable (McCormack, 2012), while this positive generational change had also permeated sport, with openly gay athletes having markedly better experiences than athletes a decade earlier (Anderson, 2011b).

For example, between 1999 and 2002, Anderson (2002) interviewed 26 openly gay high school and university athletes throughout a spectrum of sports in the United States. Due to the attitudes toward gays at the time, finding these athletes required the use of multiple channels. The sample, albeit mostly white, is large enough to say something meaningful about the experiences of openly gay athletes of the time.

The study provided the first examination of the experiences of openly gay male athletes on ostensibly all heterosexual teams. It showed that even without formal exclusion of openly gay athletes from sport, heterosexual athletes within team sports (both contact and noncontact) somewhat resisted the intrusion of openly gay athletes through the creation of a culture of silence around gay identities similar to the US military's expolicy of "don't ask, don't tell" (Britton and Williams, 1995).

Although publicly out, the athletes in this study were victimized by heterosexual hegemony and largely maintained a heteronormative framework by self-silencing their speech, and frequently engaged in heterosexual dialogue with their heterosexual teammates. This is to say that once an athlete came out of the closet, they were met by discourse that basically said "Okay, you're gay. Just keep it to yourself."

In that study, and perhaps unsurprisingly, there were more openly gay runners and swimmers than football or baseball players. Pronger (1990) theorized that competitive team sports that involve collision are more likely to be over representative of macho men, and that gay men might be likely to deselect out of them. Using data from the 1994–95 Longitudinal Adolescent Health Study, Zipp (2011) empirically validates this thesis,

showing that while gay youth played team sports equally with their heterosexual counterparts in middle school, they began to self-select out of team sports by high school.

Of course, it is possible that deeply closeted gay youth play contact sports because of the veneer it offers them against cultural suspicions of homosexuality. There is even evidence to suggest that (back then) many youth intentionally joined macho sports with the ill-informed aim of believing that it would somehow make them straight.

In 2005, Anderson expanded this work on gay male athletes to 40 openly gay and 20 closeted athletes. That study found that openly gay athletes were not physically harassed or bullied, although it was argued that their acceptance was partially attributable to the stigma of homosexuality being downplayed because these were mostly top-performing athletes. Even so, this documented that a split from the past was occurring in sport, with many high school athletes supportive of homosexuality.

Matters have improved significantly in the first decade of the twenty-first century. Supporting this statement, a 27 February, 2006, *Sports Illustrated* magazine poll of 1401 professional team sport athletes showed that the majority would welcome a gay teammate; this included 80% of those in the National Hockey League. Similarly, Cashmore and Cleland (2012) find that 93% of British football fans would be happy with an openly gay player on their team.

Throughout multiple contemporary studies of young men in sport, heterosexual men have been shown to maintain positive attitudes toward gay men in general (Adams, 2011; Cavalier, 2011; Dashper, 2012) and to the possibility of having an openly gay player on their team (Adams, 2011). The same research also showed that decreasing cultural homophobia has a positive impact on athletes' homosocial relations with their straight teammates. Here, heterosexual boys are permitted to engage in an increasing range of behaviors that once led to homosexual suspicion, all without threat to their publicly perceived heterosexual identities (Anderson, 2009; McCormack, 2012).

Collectively, these studies highlight that as cultural homophobia diminishes, it frees heterosexual men to act in more feminine ways without threat to their heterosexual identity. In other words, the homohysteria that so characterized men's sporting participation in the 1980s has dramatically waned (Anderson, 2011b); although homophobia used to be the chief policing mechanism of an orthodox form of masculinity, there no longer remains a strident cultural force to approximate the mandates of one type of homophobic masculinity (see also Floyd, 2000; McCormack and Anderson, 2010).

There is increasing evidence that as cultural homophobia continues to dissipate (particularly among male youth), team sport athletes are coming out in greater numbers. This is clear if one clicks on *Outsports.com*, where hundreds of articles related to openly gay athletes are available. More systematically, Anderson recently documented the positive experiences of 26 openly gay American high school and university athletes (Anderson, 2011b). Compared to the results in the 2002 study, these athletes (who represent the same class and racial demographics) did not fear coming out in the same way or to the same degree as the 2002 athletes. Unlike the men from the 2002 study, they did not fear that their coming out would result in physical hostility, marginalization, or social exclusion (either on or off the field). Athletes in the 2011 cohort were a more diverse group of athletes, too. Not only were team sport athletes represented equally with individual sport athletes, but also excellence on the field was no longer required for acceptance. The men were simply not as good a group of athletes, thus, they were not using sporting capital as a shield against homophobia. Still, these men were widely accepted by their teammates. This study also found that openly gay athletes evade the culture of "don't ask, don't tell" that characterized the experiences of athletes in the 2002 study. Instead, athletes in the 2011 cohort found their sexualities accepted among their teammates. These athletes talked about their sexualities frequently, and none reported that their teammates tried to publicly or privately heterosexualize them. While newspaper and other sources do continue to report high levels of fear concerning coming out in professional sports, it is argued that this is primarily attributable to the attitudes and advice given by key stakeholders in sport (managers and agents) rather than the players themselves.

There are two possible reasons for the improvement of experience of gay athletes. It could be that sport has 'learned' from pioneering openly gay athletes across America; or more likely, cultural homophobia has decreased in the local cultures of the 26 men of the 2011 sample. If the latter is the case, it speaks to a broader decrease in homophobia throughout the country (see Keleher and Smith, 2012). Accordingly, the existence of local cultures with great social inclusivity speaks at some level to inclusivity in the broader culture.

Unfortunately, despite a very large number of lesbian sport scholars, considerably more research needs to be conducted on the experience of openly lesbian athletes in school and community sports today. There has been comparatively very little on the topic since Griffin's (1998) book *Strong Women, Deep Closets*. In few studies that exist, Fink et al. (2012) showed improvements in educational sports for the dozen lesbian women studied, and Drury's (2011) study on lesbian football teams in England shows that, in some cases, sport teams can provide a 'safe zone' for athletes to be open about their sexuality. Melton and Cunningham (2012) also observed that many lesbian athletes of color experienced sexual prejudice and internalized stigma in American university sports; however, they also felt social support from various teammates and, to a limited degree, employees in the sport organization. Finally, Cunningham et al. (in press) observed that lesbian university athletes playing on inclusive, psychologically safe teams were more likely to express their sexual orientation personal identity than were their peers.

Although limited, the research on the experience of lesbian athletes reflects the aforementioned improving attitudes toward homosexuality more broadly. Furthermore, Cunningham (2010) has surveyed nearly 700 university athletic department members in nearly 200 institutions to show that while sexual diversity lags behind age or gender diversity, 54% of the universities studied maintained strong sexual orientation diversity, only 17% showed no diversity.

Future Directions

As cultural homophobia and therefore homohysteria continue to erode in Western countries, future research needs to examine

how this will impact not only upon the lives of self-identified gay and lesbian athletes, but also on that of other sexual and gender minorities. There is a dearth of academic research on the experiences of bisexual athletes (of both sexes) and limited research on the experience of transsexual athletes as well. The experiences of athletes in non-Western countries are particularly problematic, and requires significant research as well. Finally, intersections of sexual and gendered orientations need more examination as they intersect with race.

See also: Gay and Lesbian Movements; Gay, Lesbian, and Bisexual Youth; Gender: Gendering of Categories; Heterosexism and Homophobia; Masculinity, Scripts of; Oppression; School Environment for LGBTQ/Sexual Minority Youth; Sexism; Sexual Orientation: Historical and Social Construction; Sport, Sociology of.

Bibliography

Adams, A., 2011. Josh wears pink cleats. Journal of Homosexuality 58 (5), 579–596.

Anderson, E., 2002. Openly gay athletes: contesting hegemonic masculinity in a homophobic environment. Gender & Society 16 (6), 860–877.

Anderson, E., 2005. In the Game: Gay Athletes and the Cult of Masculinity. State University of New York Press, Albany, NY.

Anderson, E., 2009. Inclusive Masculinity: The Changing Nature of Masculinities. Routledge, London.

Anderson, E., 2011a. The rise and fall of western homohysteria. Journal of Feminist Scholarship 1 (1), 80–94.

Anderson, E., 2011b. Updating the outcome: gay athletes, straight teams, and coming out in educationally based sport teams. Gender & Society 25 (2), 250–268.

Britton, D.M., Williams, C.L., 1995. 'Don't ask, don't tell, don't pursue': military policy and the construction of heterosexual masculinity. Journal of Homosexuality 30 (1), 1–21.

Bryson, L., 1987. Sport and the maintenance of masculine hegemony. Women's Studies International Forum. 10, 349–360.

Burstyn, V., 1999. The Rites of Men. University of Toronto Press, Toronto.

Burton-Nelson, M., 1994. The Stronger Women Get the More Men Love Football. Avon Books, New York.

Cancian, F.M., 1987. Love in America. Cambridge University Press, Cambridge.

Cashmore, E., Cleland, J., 2012. Fans, homophobia and masculinities in association football. British Journal of Sociology 63 (2), 370–387.

Cavalier, E.S., 2011. Men at sport: gay men's experiences in the sport workplace. Journal of Homosexuality 58 (5), 626–646.

Connell, R.W., 1995. Masculinities. University of California Press, Berkley, CA.

Cox, B., Thompson, S., 2001. Facing the bogey: women, football and sexuality. Football Studies 4, 7–24.

Cunningham, G.B., Pickett, A., Melton, E.N., Lee, W., Miner, K. Free to be me: psychological safety and expression of sexual orientation personal identity. In: Hargreaves, J., Anderson, E. (Eds.), Routledge Handbook of Sport, Gender and Sexuality. Routledge, New York, in press.

Cunningham, G.B., 2010. Predictors of sexual orientation diversity in intercollegiate athletics. Journal of Intercollegiate Sport 3, 256–269.

Dashper, K., 2012. Dressage is full of queens: masculinity, sexuality and equestrian sport. Sociology 46 (6), 1109–1124.

Davis-Delano, L., Pollock, A., Ellsworth Vose, J., 2009. Apologetic behaviour among female athletes. International Review for the Sociology of Sport 2–3, 131–150.

Drury, S., 2011. It seems really inclusive in some ways, but … inclusive just for people who identify as lesbian: discourses of gender and sexuality in a lesbian-identified football club. Soccer & Society 12 (3), 421–442.

Ezzell, M.B., 2009. "Barbie dolls" on the pitch: Identity work, defensive othering, and inequality in women's rugby. Social Problems 56, 111–131.

Felshin, J., 1974. The triple option for women in sport. Quest 21, 36–40.

Fink, J., Burton, L.J., Farrell, A.O., Parker, H.M., 2012. Female intercollegiate athletes' experiences in revealing their sexual identities. Journal for the Study of Sports and Athletes in Education. 6, 83–106.

Floyd, K., 2000. Affectionate same-sex touch: the influence of homophobia on observers' perceptions. The Journal of Social Psychology 140 (6), 774–788.

Frank, D.J., Camp, B.J., Boutcher, S.A., 2010. Worldwide trends in the criminal regulation of sex, 1945 to 2005. American Sociological Review 75, 867–893.

Freud, S., 1905. Three Essays on the Theory of Sexuality. Complete Psychological Works, vol. 7. Hogarth, London.

Garner, B., Smith, R., 1977. "Are there really any gay male athletes?" An empirical study. The Journal of Sex Research 13 (1), 22–34.

Griffin, P., 1998. Strong Women, Deep Closets. Human Kinetics, Champaign, IL.

Hargreaves, J.A., 1994. Sporting Females. Routledge, London.

Hekma, G., 1998. "As long as they don't make an issue of it…": gay men and lesbians in organized sports in the Netherlands. Journal of Homosexuality 35 (1), 1–23.

Ibson, J., 2002. Picturing Men: A Century of Male Relationships in Everyday Life. Smithson Books, Washington, DC.

Keleher, A., Smith, E., 2012. Growing support for gay and lesbian equality since 1990. Journal of Homosexuality 59 (9), 1307–1326.

Kian, T., Anderson, E., 2009. John Amaechi: changing the way sport reporters examine gay athletes. Journal of Homosexuality 56 (7), 1–20.

Kimmel, M., 1994. Masculinity as homophobia. In: Brod., H., Kaufman, M. (Eds.), Theorising Masculinities. Sage, London, pp. 213–219.

Kimmel, M., 1996. Manhood in America. Free Press, New York.

Kozloski, M., 2010. Homosexual moral acceptance and social tolerance. Journal of Homosexuality 57 (10), 1370–1383.

Krane, V., Barber, H., 2003. Lesbian experiences in sport: a social identity perspective. Quest 55 (4), 328–346.

Lenskyj, H.J., 1986. Out of Bounds: Women, Sport and Sexuality. The Women's Press, Ontario.

Lenskyj, H.J., 2003. Out on the Field: Women, Sport and Sexualities. Women's Press, Toronto.

Loftus, J., 2001. America's liberalization in attitudes towards homosexuality, 1973-1998. American Sociological Review 66, 762–782.

Mangan, J.A., 2000. Athleticism in the Victorian and Edwardian Public School. Frank Cass, London.

McCormack, M., 2012. The Declining Significance of Homophobia: How Teenage Boys Are Redefining Masculinity and Heterosexuality. Oxford University Press, New York.

McCormack, M., Anderson, E., 2010. The re-production of homosexually-themed discourse in educationally-based organized sport. Culture, Health & Sexuality 12 (8), 913–927.

McDonagh, E., Pappano, L., 2008. Playing with the Boys. Oxford University Press, Oxford.

Melton, E.N., Cunningham, G.B., 2012. The effect of LGBT-inclusive policies, gender, and social dominance orientation on organizational attraction. International Journal of Sport Management 13, 444–462.

Messner, M., 1992. Power at Play. Beacon Press, Boston.

Pronger, B., 1990. The Arena of Masculinity. St. Martin's Press, New York.

Sabo, D., Panepinto, J., 1990. Football ritual and the social reproduction of masculinity. In: Messner, M., Sabo, D. (Eds.), Sport, Men and the Gender Order: Critical Feminist Perspectives. Human Kinetics, Champaign, IL, pp. 115–126.

Sage, G., 1990. Power and Ideology in American Sport: A Critical Perspective. Human Kinetics, Champaign, IL.

Savin-Williams, R.C., 2005. The New Gay Teenager. Harvard University Press, London.

Shire, J., Brackenridge, C.H., Fuller, M., 2000. Changing positions: The sexual politics of a women's field hockey team 1986–1996. Women in Sport & Physical Activity Journal 9 (1), 35–64.

Southall, R., Anderson, E., Southall, C., Nagel, M., Polite, F., 2011. An investigation of ethnicity as a variable related to US male college athletes' sexual-orientation behaviors and attitudes. Ethnic and Racial Studies 34 (2), 293–313.

Weeks, J., 2007. The World We Have Won. Routledge, London.

Wolf-Wendel, L., Toma, D., Morphew, C., 2001. How much difference is too much difference? Perceptions of gay men and lesbians in intercollegiate athletics. Journal of College Student Development 42 (5), 465–479.

Woog, D., 1998. Jocks: Stories of Real Gay Athletes. Alyson Press, Hollywood, CA.

Wright, J., Clarke, G., 1999. Sport, the media and construction of compulsory heterosexuality. International Review for the Sociology of Sport 34, 227–248.

Zipp, J.F., 2011. Sport and sexuality: athletic participation by sexual minority and sexual majority adolescents in the U.S. Sex Roles 64, 19–31.

Geertz, Clifford (1926–2006)

Fred Inglis, University of Sheffield, Sheffield, UK; University of Warwick England, UK; and School of Social Science, Institute of Advanced Study, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Clifford Geertz's early anthropological studies in Indonesia were conducted as a straightforward ethnographer, although always with a sense of humor. His career-changing sally into Balinese life and the cockfight created a quite new idiom and conceptual apparatus for the understanding of such extravagant excess. His implicit method was that of 'thick description' and he carried this banner into argumentative combat with the predominant schools of quantitative and sternly materialist anthropology. In his classic text *Negara: The Theatre-State in 19th Century Bali* he contends that our understanding of public life is necessarily an interpretation of value and meaning as expressed in the most commonplace of ordinary action. This action embodies 'a constellation of enshrined ideas' where all behavior is symbolic and the drama of everyday life is best grasped as texts to be read. These lessons he sustained by writing some of the best academic prose of his day, demonstrating his precept that humankind interprets as it makes, that human science is one of the humanities and that 'thinking is a moral act.'

I

Clifford Geertz (1926–2006) was not only the leading anthropologist of his generation but, in transforming the subject into his own interpretive idiom while insisting that same idiom be as economical and unaggressive as possible, he summoned the human sciences at large to the common pursuit of true judgment. He brought off this momentous task by giving what he called, in a mind-altering phrase peculiarly his own, 'the blurred genres' of 'the way we think now,' both precision and accessibility. It is fair to add that in so doing he restored more than a little of his country's good name, disfigured as it had become, in Vietnam, Salvador, and elsewhere, by rather too emphatic a view of what was good and bad for other peoples. Geertz taught, by lived example and by explicit doctrine, the best virtues of the American and English literature he knew by heart and which shaped his formation every bit as deeply as the disciplinary conversation of anthropology. It could readily be claimed that, for a season, he made anthropology queen of the sciences, and in so doing gave the intellectual life of his times his signature version of that 'sane, affirmative speech' (W.H. Auden's phrase) without which the social sciences lose sight of their public duty and become the stooges of managerial power.

II

Clifford Geertz's parents were divorced when he was only 3 years old, and he was dispatched into the care of a 60-year-old foster mother whom he called 'Nana,' living out the Great American Depression of the 1930s with her in the hills of northern California, while all the while seeing his parents (who lived into very ripe old age) for a day visit once or twice a year. He went off to a two-room, two-teacher village school with four dozen farmworkers' children and his own, inexplicably intellectual bent, "learning early that 'brightness' was my passport out of exile and into the great world."

At 17, the United States barely 2 years at war, Geertz volunteered for the navy and was allowed in as an electronic technician's mate, repairing radar gear on the USS *St Paul*, sister ship to the much-mourned and unlucky *Indianapolis*, torpedoed in the dark night of 4 August 1945 after delivering the most frightening cargo of the century, the first battle-usable atom bomb, Little Boy, to the US air base in Guam.

As Geertz's own ship steamed onward toward what its crew were braced for as the certain-to-be-terrible invasion of Japan, the bomb was dropped and one of the inventions of the twentieth century for which he and so many intellectuals of the epoch felt the most acute moral disapproval very probably saved his life. It was an early instance of the kind of moral ambivalence, a paid-on-both-sides queasiness, which, over the years, he was to raise into a sort of intellectual principle for his kind of anthropology and the good it might do to others.

On board ship in the empty Pacific, as a friend of his put it, you had all the advantages of suicide and none of the disadvantages. Geertz read all he could get hold of, including the venereal disease manuals ('but the plots tended to repeat'), and played bridge incessantly, warding off the wardroom's stupefying tedium, until he was very good at it, so never touched a card again after demobilization.

He had spent his first 20 years with the kind of oppositions he fashioned into his businesslike, usable, and always generous-hearted schemata: between the agora, the wide-open spaces of familylessness, upstate California in 1929, the hot empty reaches of the *St Paul*, the shimmering blankness of the ocean, and the noisy, crowded racket of a rural school, the close and sweaty din below decks, the packed bunks, and companionways. He took his chances by way of the Navy's natural and colossal wastefulness, his quick wits hitting it off with opportunity. The next, gaping opportunity was the magnificent GI Bill, which provided more or less free college education for all servicemen who wanted it. Geertz's IQ scores were on his naval file; they too had probably helped save his life by switching him into radar and out of direct combat (another useful bit of paying-on-both-sides). They took him to Antioch College, Ohio, to read literature.

Once at Antioch he continued the tumultuous reading with which he had filled the vast blanks of life in the Pacific, and found American literature of the twentieth century as crowded

and exhilarating as he could wish, taking particularly keen and comradely pleasure in the sumptuousness of Wallace Stevens, the heavy historicizing of the action in Faulkner, and, a little later, the hilarity of Saul Bellow, together with the wry, delightful difficulty of James Merrill. His guide at Antioch through this busy, noisy throng was one George Geiger, a great name in the subject and a teacher who not only cultivated Geertz's terrific eye for the comedy of human nature but, at the same time as introducing him to contemporary philosophy, then at the height of its proselytizing for minute attention to linguistic particularity, warned Geertz off such study when he looked forward to postgraduate work.

Geiger advised his brilliant pupil that this philosophy, however acute, was too far from the facts of life and the irresistible messiness of everyday human endeavor. With a shrewd sense of his student's character as well as an accurate judgment as to the then most thriving subject-matter in human inquiry, he directed Geertz to Harvard's Department of Social Relations to which the anthropologist Clyde Kluckhohn introduced him and to its boss, Talcott Parsons, whose ambition was to crown sociology, not anthropology, queen of the human sciences by way of a crazily inclusive Theory of Everything, in two volumes entitled *The Structure of Social Action*, (Glencoe: the Free Press, 1937 and 1951.) and *The Social System*.

Such ambitions had little place in the view of the world, Geertz contrived for himself out of a personal experience compounded of solitary self-sufficiency, an exceptionally lively sense of humor, and the pragmatic cast of mind, at once doctrinal and characteristic, of a quiet and talkative American. His ideal type came not to cajole other people into the American way of life, but rather to comprehend the sheer variety of humankind as no more and no less than the polyglot, technicolor materials out of which his own nation was made.

III

Building himself a method and a style out of a character and its dispositions, he turned back in a way then unexpected to the canons of contemporary philosophy. From J.L. Austin (1962) he took the contention that language is not an endless sequence of true, false, or meaningless propositions, as A.J. Ayer would have it (Ayer, 1936). Rather, it is an unstoppable business of making and doing and instituting; speech and writing was, in Austin's neologism, *performative* and actively worldmaking.

Beside Austin, Geertz placed Gilbert Ryle's *The Concept of Mind*, in which he found the bracing argument that there is no such thing as 'a ghost in the machine,' which is to say that each of us does not have a 'real' 'inner' person much truer and more important than the 'outer' creature other people see. Each of us is constituted by our daily actions, including our speech, for everyone else to see and hear. From Ryle, Geertz also took a key slogan, 'thick description,' which he turned, in one of his most gripping and inclusive essays, into the first principle of ethnographic practice: "Analysis … is sorting out the structures of signification … and determining their social ground and import" (Geertz, 1975). 'Thick description' is a matter of telling the truths observed, for sure, but telling them by way of all that is going on in human encounters, layers of it, subjunctive piled on indicative, hints hidden in flagrancies.

Lastly, Geertz's architecture, only fully constructed, in its finished form, by the time he became Professor (and Head) of Social Science at the Institute for Advanced Study, Princeton, rested upon an American literary critic of a sort-of-Marxist persuasion, Kenneth Burke along with, clinchingly, the great iconoclast of philosophy, Ludwig Wittgenstein in his second incarnation as philosopher of everyday language.

Psychological or emotional words do not, Wittgenstein told him, name private experiences which we alone can know. The language of feeling is what teaches us to differentiate emotions and makes it possible, from infancy forward, to distinguish them from raw sensations. We learn the rules of feeling according to what Wittgenstein famously described as the language-games of social practice, and we can only make ourselves understood (including to ourselves) if we follow the rules. "Ideas are not," as Geertz himself puts it, invoking Wittgenstein, "and have not been for some time, unobservable mental stuff. They are envehicled meanings …," (Geertz, 1980) and the vehicle may be the usual signs of language, as monologue, tête-à-tête, or tumult, or they may be stylized symbolic actions – rituals, ceremonies, legal disputes, religious meditation, scientific experiment, childbirth, deathbed.

Wittgenstein gave Geertz the key to unlock the significance of commonplace signification, and enabled him to say, in one of his grandest aphorisms (their occasional grandeur in no way affecting their careful domesticity and localism), made in this case about Balinese ceremony,

> The real is as imagined as the imaginary … The dramas of the theatre state, mimetic of themselves, were, in the end, neither illusions nor lies, neither sleight-of-hand nor make-believe. They were what there was.
>
> *Negara*: p. 136

This is a mature formulation of the idea found in Kenneth Burke 30 years previously. 'Dramatism' is the second of his key concepts (the first being Wittgensteinian), and Geertz found and endorsed it in Burke. If language is a "symbolic means of inducing cooperation in beings that by nature respond to symbols," then the "dancing of an attitude," which is a particular symbolic action, is held in the larger plot of the social drama. Nor is this use of drama, in Burke's eyes, metaphorical. "Drama is employed *not as a metaphor* but *as a fixed term* that helps us discover what the implications of the terms 'act' and 'person' really are" (my italics).

Geertz, as one would expect, fairly delighted in the phrase "the dancing of an attitude," and was glad to endorse Burke's fundamentally comic view of human action where 'comic' connotes not merely all there is in people's conduct to make you laugh, but an inclusiveness of vision taking in, according to Burke's definition, (This quotation from Kenneth Burke. In: Gusfield, J.R. (Ed.), On Symbols and Society. Chicago, University of Chicago Press, pp. 135–138.) "The maximum of forensic complexity." It encompasses human misery and cruelty by seeing their operation not as tragic but either as mistakes or as shameless acts of cheating. By this token, politics is all culture, and "ideology is to myth as rhetoric is to poetry" (Burke, *Ideology and Myth*', Kenneth Burke. In: Gusfield, J.R. (Ed.), On Symbols and Society. Chicago, University of Chicago Press, p. 303).

IV

Geertz, naturally, only shook these readings into theoretic order gradually, and as his career led him far east, to Java and Indonesia in 1952, where the Dutch were disencumbering themselves of the White Man's Burden, and the universities of Harvard and Chicago, under whose auspices Geertz and his wife Hildred were traveling, were busying themselves, with government aid, in devising forms of inquiry which would be good for new nations, and keep them away from dabbling in communism.

The Javanese themselves had more violent and disputatious plans for new nationhood, and Geertz spent his almost 20 years (1952–71) of regular but necessarily intermittent fieldwork with a background (mostly) and foreground (sometimes) of murderous civil war. He wrote his four books on Indonesia out of these bloodstained passages, finding, in particular while writing *The Social History of an Indonesian Town* (the town in question being Pare disguised as Modjukoto) and *Agricultural Involution* (The first published by MIT Press, Cambridge, MA, 1965; the second by University of California, Berkeley, 1963. *Peddlers and Princes* from Chicago in 1963.) ('involution' being for Geertz an unmodern variant of modernity in which an old ecology is forced to accommodate new and impossible overcrowdings of agricultural production) that hypothesis-testing was no sort of method, and that daily life went unrepentantly on even with tanks in the street and corpses hanging from telegraph poles. Out of these clashing disparities he fashioned his daring interpretations of symbolic action, *Peddlers and Princes* giving space to his old novel-writing ambitions by way of a boisterous cast of historical heroes each retaining busy actuality in the present, and *The Religion of Java*, the major work of the two decades and an early compendium of symbolic anthropologizing in the accents of Max Weber, leveling old Muslim theodicy at new capitalist economics.

During 1957 Geertz opened his long, loving, and horrified relationship with Bali, and, 'malarial and diffident,' attended the cockfight which provided the occasion for Geertz's most famous contribution to the literature of anthropology as well as being itself a striking addition to English literature *tout court*.

The cockfight was then illegal in Bali, but everywhere practiced. Two eagerly living, not insubstantial creatures rip each other to bits with razors fastened to their heels. The illegality, the smell of blood and violent death, the likelihood of a bust by the cops, the implicit, inescapable jokes about men's cocks and cockfighting (but, by God, the dead seriousness of both), the noisy, excited, dusty gregariousness of it all, gives the traveler's tale terrific vitality, a whiff of prurience, and an object lesson.

The cockfight is duly raided by the police, the Geertzes run for cover like everyone else, are shielded by complete strangers, and are thereafter one of the family, getting to see plenty of cockfighting, cocktending, cocknurture, cockworship.

Crucial to an understanding of the cockfight is how utterly unlike everyday Balinese decorum the crazily abandoned behavior of the fight itself is. Animality to the Balinese is obscene, cockfights dramatize beastliness with shocking immediacy. As with all sports, there are intricate divisions of labor and indispensable experts – spurriers, handlers, umpires, and of course holders of the betting ring. The passion of gambling, the smell of blood, the flutter of torn feathers, the commitment of the (male) self, the rivalry of owners and villages and of mutual standing, catch up in a single whirl of good luck and bad the same mighty themes as do.

> … for other peoples with other temperaments *King Lear* and *Crime and Punishment* – death, masculinity, rage, pride, loss, beneficence, chance – and, ordering them into an encompassing structure, presents them in such a way as to throw into relief a particular view of their essential nature.

It is a bit of a shock to see Shakespeare and Dostoevsky put so coolly beside the fighting cocks. It is meant to be. Geertz uses anthropology's licence to be cavalier with categories. Art is both systemic across cultures and utterly peculiar to cultures. Its first use is

> … to render ordinary everyday experience comprehensible by presenting it in terms of acts and objects which have had their practical consequences removed and been … raised to …, the level of sheer appearances.

Art forms are first of all a way of theorizing life without paying in blood or money (or, in the case of gambling, not too much money). You may pay in tears, at the cockfight or at *La Traviata*, but not for long. Art gives form to a story about ourselves in which we can try out how things might have been otherwise if only we had been or had met the heroine in the movie or the novel.

In the case of the cockfight as Geertz interprets it for us (and for the Balinese also), a people of unusual blandness and obliquity in ordinary dealings present a tale of themselves as wildly clamorous and murderously cruel. Like poems and symphonies, the cockfight is all passion and melodrama, and the gambler-spectators as well as the gambler-owners throw themselves into the fight with the grand gestures of Guignol, the masks of comedy and tragedy. What tragic actor, symphony conductor, and concerto pianist perform for us, they perform, with a little help from their cocks, for themselves.

Thus cockfights, puppet plays, renaissance frescos, football matches, television drama alike present aspects of a people's subjectivity to itself, divided according to the groups and classes who use the form and the contents in question. That particular corner of the subjectivity may be dusty and unvisited or familiar and well swept, but in *recognizing* it as a place for the moment to be, the audience not only view themselves, they remake themselves as well. The forms and contents of our cultural lives, heavy with feeling, are part of our constitution in all senses of the word.

In the case of the Balinese, as in all other cases, the game is played within the thick, deep and taken-for-granted habitations and habiliments which make the plays what they are. Cockfights are, of course, for men only and not any men but those with the status and the money to go in and play deep (lose from the depths of those same pockets also, if need be), as the eponym of Geertz's title (taken, surprisingly, from Bentham) indicates.

Geertz places the betting structure isomorphically upon the social structure and the evenness of the fight. That it fits both

so well is testimony to its significance. As a matter of making categorical mischief – the anthropologist's contrastive–comparative business being with understanding the cultural forms of others by way of a deliberate misprision of one's own (*King Lear* and the cockfight) – Geertz levels the works of high European literature with low Balinese birds.

Like all good ethnography it carries a sting in its tale, a tale told by the Balinese to themselves. It is in this essay that Geertz most plainly teaches his methodological lesson that the present-day human scientist is, in his or her puzzles and epiphanies, much more like a literary critic, an art historian or a jurisprudentialist than like a mechanic ('how does this work?'), an anatomical surgeon ('scalpel, please') or even a cryptographer ('E means P'). The shadow play, the funerary rite, the cockfight and, in a moment, the bazaar bargaining in Sefrou are ways of living-and-telling stories of ourselves and their selves to us and them. All social institutions and cultural forms are imaginatively embodied works, graspable by anybody who takes enough trouble. Towns have symbolic forms as much as marriage ceremonies, mortgage agreements, election campaigns, and poems. Each is turned to particular purposes; each "[puts] emotion to cognitive ends"; each is a thread woven into the dense tapestry of social narratives which is our definition of a culture.

V

The admission of the Balinese cockfight to the disciplinary canon was unstoppable and right. But it briefly obscures a disciplinary principle at the heart of Geertz's *oeuvre*. He quotes but then sidesteps a remark of George Santayana to the effect that "people compare when they can't get to the root of the matter," but comparing unalike conduct and its latent symbolism became for Geertz the very mode of his vision. A casual acquaintance had suggested Morocco for the other half of his bifocal method – 'it is safe, dry, Muslim, and the food is good' – and Geertz intended to dramatize his comparativism by using the optic he had cut to look at Java as a lens for looking at the market town of Sefrou tucked beneath the Atlas mountains. It was a jostling, argumentative, thronged, and sufficiently peaceable place packed with Berbers, Arabs, Jews, Quranic teachers; with shepherds, sheikhs, imams; the spacious market-place and the 600 small shops in the narrow streets round about jammed with sociable tea-and-coffee drinking, cheroot-smoking bargain-makers.

This *suq* was his subject, and with a small team of colleagues including his first wife, Hildred, Geertz pursued Meaning and Order in Moroccan Society (With Hildred Geertz and Lawrence Rosen and a photographic essay by Paul Hyman, Cambridge University Press, New York, 1979.) over many visits and 7 years, discovering those crucial values firmly established in a baffling whirl of linguistically defined genealogy (the *nisba*), the economic and physical immediacy of the church (*habus* and *zania*, tithe and college of worship), the density of market life (*suq*) and the completeness with which its frequenters (*suwwaq*) disappear into that life.

There *is* order in all this, and any amount of meaning. (The predicament for the modern world is that as the density of this kind of local life is dislimned, meaning begins to make itself scarce.) But there is no Lévi-Straussian bipolarity nor Leachian algebra nor Douglasian quadrant which will allocate and clarify it. Such sorting may *always* be attempted with the incorrigible messiness of human societies, but binary patterns or quadrant allocation are more likely than not to lift life up into the upper air of abstraction and to deprive the unbearable lightness of human beings of the heavy weighing down so blessedly provided by custom, practice, and locality.

The comparative, bifocal, astigmatic vision which is Geertz's prime contribution not just to his parent discipline but to our common understanding of (in the blank phrase) the modern world, everywhere insists upon the endless, contradictory minuteness of social conduct, coexistence of quite incompatible beliefs and practices in its members, the many contexts.

> … within which men [this is Morocco in the 70s, after all] are separated by their dissimilitudes and those … where, however warily and however conditionally, men are connected by their differences.
>
> *Meaning and Order*: p. 141

The moral lesson he was never tired of repeating – *had* to repeat so often because his listeners were deaf to it – is the Wittgensteinian one that there is no such thing as a private language. Inner life is visible and audible in outer action. The 'Moroccan mind,' like any other mind, "is an overt, observable feature of a well-lit public arena in Moroccan society" (*Meaning and Order*: p. 212).

To drive this lesson home, he then turned to a radically historical example once more to show (noting in a pungent dictum on the way, that 'examples instruct, they do not prove') that the public drama of life is eloquent of the deepest meanings of the society in question (not of the whole world, that is, but, in this case, of the Balinese people, which is quite enough to be going on with).

Negara: The Theatre-State in 19th Century Bali (Princeton University Press, Princeton, 1980.) is Geertz's single and most complete masterpiece. It describes, and interprets by describing, how a society mad about status, scrupulous in husbandry, frequent, brief and severe in battle, gorgeously elaborate in its many lavish forms of cultural expression, ordered its rulers, its pageants, and the cultivation of its rice paddies into its long-perpetuated 'theater state.'

This is political theory all right, but far away from what Geertz calls "those great beast views of the state that, from Hobbes … to de Jouvenal … locate its power in its threat to harm." He goes on to admonish those who, with Marx or Pareto in hand, see the show of state, its semiotics and public spectaculars as mere devices with which to adorn itself in legitimacy. There is a good deal of sanctimony in the zeal with which political theorists point out the presence of power, with "coercion its expression, violence its foundation, and domination its aim," (*Negara*: p. 134.) but "such tired commonplaces … put out of sight all that is important and interesting about politics." They are blind to its drama, to the forms, rituals, and conventions which hold a whole society in its place. The Negara did the job for the Balinese for seven whole centuries before the Dutch took slaughterously over. The question then for us, readers of Geertz's wholly exotic and very tall, detailed, and actualized tale, is what are the dramas which

command our own allegiance? "For the [Balinese] pageants were not mere aesthetic embellishments, celebrations of a domination independently existing: they were the thing itself." (*Negara*: p. 122.)

Geertz compresses the theater-state into a complex model, holding in a tense equipoise and without remainder the 'high gorgeousness' of pagoda and palace, processions and cremations (sumptuously described in his historical source) together with the practical arrangements for smallholding, irrigation, and ensuring the staff of life, the rice crops.

In the Bali of Geertz's figuration the model of the state is all-inclusive; neither culture nor society leaks out anywhere. The drastic, implied challenge to political scientists in (to go no further) the United States, is then to match Geertz's style of metaphor with a comparable (but of course unalike) dramatization of, over here, the top-heavy ceremonies of public statements outside the White House and, over there, the erratic disjunctions of political-life-as-televised and then responded-to in Tennessee and Montana.

VI

After *Negara*, which came out in 1980, Geertz more or less finished with fieldwork. Within the discipline, he was much beset by a new generation tormenting itself about the past complicity of American social science in continued imperialist adventures of a bloodthirsty, anticommunist kind, whether in Columbia, Chile, or Vietnam. This self-torment frequently declared itself as a sort of *angst-de-politesse* about whether the fieldworker condescended to, overrode the self-explications of, or otherwise abused the subject-and-object of inquiry.

The *angst* in question is most readily to be seen in the title of a revisionist study of Geertz's own territory, John Pemberton's *On the Subject of 'Java,'* praised at that date by another youngish Turk, Benedict Anderson, as 'a brilliantly Derridan ethnography' (Derrida being the all-purpose authority for putting everyone's understanding of other people in the inverted commas of self-delusion). Geertz would have none of this; it was 'moral hypochondria.' He set himself, in a series of gripping essays – each one, it should be added, a model of the best Anglo-American prose of his day – consistently to enumerate the principles of his very own 'blurred genre.' 'The way we think now' should be couched in a humane and commonly intelligible language with which to write (the title of another of his canonical essays) "the social history of the moral imagination." (In his second collection, *Local Knowledge: Further Essays in Interpretive Anthropology*. Basic Books, New York, 1983.)

From his spacious, egalitarian purview in the Chair at the Institute for Advanced Study, he assembled his doctrinal version of modern letters, fluent, talkative certainly, prompt with a succession of memorable aphorisms, unfailingly funny, wise also, one best face and manner of his nation. In his classic memoir, *After the Fact*, (*After the Fact: Two Countries, Four Decades, One Anthropologist* (Harvard University Press, Cambridge, MA, 1995); *Works and Lives: the anthropologist as author* (Polity Press, Cambridge, 1988).) with its James Thurber-punning title ("I'm just trying to creep up on it"), he surveys his own work; in *Works and Lives* he surveys that of his masters, mistresses, and equals (with particular affection, Margaret Mead); in *Available Light* (a typical title) (Geertz, 2000) he tours the terrain of the way we think now with a habitual emphasis on, as one essay title has it, "thinking as a moral act" – which goes on, relevantly, "ethical dimensions of anthropological fieldwork in the new states."

The collection, it is much to the point of this coda to add, was given to worldwide audiences, often in those many Institutes of Advanced Study made in the image of the great original at Princeton: in Florence, the Netherlands, Berlin, Vienna, and under the aegis of the tanner lectures.

The final essay betokens this large purview. It is called "The World in Pieces" and asks of us all, "what is a country if it is not a nation?" It commands by example Geertz's mode of thought as refusing "intensely generalised reflection on intensely generalised matters" commending instead an intellectual activity 'mobile, exact, and realistic' – well, yes – a kind of thought taking to heart all he has had to say about the necessity of local knowledge, the jarring coexistence of modernity and antiquity in all our lives, the utter unamenability of new nations to old power, all this and our "deepest and most central commitment; the moral obligation to hope" (*Available Light*: p. 260).

See also: Anthropology, History of; Anthropology: Overview; Ethnography in Applied Social Research; Everyday Life, Anthropology of; Fieldwork in Social and Cultural Anthropology; First Language Acquisition, Developmental Psychology of; Interpretation in Anthropology; Ritual; Symbolism in Anthropology; Thick Description: Methodology; Wittgenstein, Ludwig (1889–1951).

Bibliography

Austin, John Langshaw, 1962. How to Do Things with Words. Clarendon Press, Oxford.
Ayer, Alfred Jules, 1936. Language, Truth and Logic. Clarendon Press, Oxford.
Burke, Kenneth, 1989. On Symbols and Society. University of Chicago Press, Chicago.
Clifford, James, Marcus, George, 1986. Writing Culture: The Politics and Poetics of Ethnography. University of California, Berkeley.
Geertz, Clifford, 2000. Available Light: Anthropological Reflections on Philosophical Topics. Princeton University Press, Princeton.
Geertz, Clifford, 1975. Thick description: toward an interpretive theory of culture. In: The Interpretation of Cultures. Hutchinson, London, pp. 3–33.
Geertz, Clifford, 1980. Negara: The Theatre-State in 19th Century Bali. Princeton University Press, Princeton, p. 135.
Geertz, Clifford, 2010. Life among the Anthros. Princeton University Press, Princeton.
Inglis, Fred, 2000. Clifford Geertz: Culture, Custom, and Ethics. Polity, Cambridge.
Ryle, Gilbert, 1957. The Concept of Mind. Oxford.

Geist (Spirit): History of the Concept

Eric R Lybeck, Girton College, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the century before the 'Classical' period of social science, a number of modes of scholarship made use of the concept of *Geist* – translated in English as 'Spirit' or 'Mind.' The concept was relational, and connected the universal and particular in an organic unity. This article traces the history of the concept from its origins in Christian theology to Montesquieu's *Spirit of the Laws*. Later applications in German philosophy included Herder's nationalism and Hegel's dialectical idealism. The concept had direct influence on the historical legal sciences until falling into disrepute as a result of application in spiritualism and Nazi ideology.

Introduction

Early in his classic essay, Max Weber asked the reader's forgiveness for using 'the somewhat pretentious sounding' expression 'Spirit of Capitalism' (Weber, 2002: p. 8). He then explained that his object of study could not be defined, but only illustrated. To that end, he referenced Benjamin Franklin's words embodying this "ethically slanted maxim for the conduct of life" (Weber, 2002: p. 11). No one doubts the influence of Weber's *Protestant Ethic and the Spirit of Capitalism*'s influence on the social sciences. Yet, few today recognize that Weber's use of the term 'spirit,' or '*Geist*' in the original German was not of his own invention. Rather, within the legal sciences in which Weber was trained, the term was a technical concept characteristic of his Age. To grasp this usage, we must reconstruct the context in which the concept was situated.

Unfortunately, though widely used throughout the nineteenth century, '*Geist*' did not have a clear-cut conceptual definition, translatable across different disciplines or authors employing the term. The word reflects a disparate set of meanings – sometimes referring to a specific object of study in cultural studies, and sometime to the totality of the cosmos. The complex definitional constellation in the German language, where the term was employed most regularly, is reflected in the difficulty of translating '*Geist*' into English. The three most widely used translations are 'Spirit,' 'Mind,' or 'Ghost.' The last is etymologically closest, reflecting the Germanic origins of English. However, in academic texts, 'Spirit' and 'Mind' are more frequently employed. In other instances, as in, *Zeitgeist* – literally, the spirit of the age – the term is left as is.

This summary conceptual history will not provide a firm definition, but rather traces the historical development of a wide set of usages that the term has engendered during the modern period. Just as Weber could not define the particular spirit of capitalism, but only describe it, so too will this history of the conceptual usage of '*Geist*' provide illustration of the meaning for those who used it.

Origins of the Term

Theology: The Holy Ghost

Over the course of the modern era, particularly since the eighteenth century, the concept of spirit took on secular connotations. But these origins were prefigured by the use of the term in Theology. The notion of the Holy Spirit had substantive import within Christian religious conceptions of the relationship between God and humanity. A comprehensive history of theological interpretations of the Holy Spirit exceeds present purposes, but we might begin by noting the change that occurred during the Protestant Reformation. Traditional church doctrine had established that the Holy Spirit was the mediating force between humans and God. However, Protestantism grounded this relationship within the individual, rather than in the institution of the church and its hierarchy. In John Wesley's Methodism, for example, the Spirit reflected 'holiness' and 'love,' which was embodied in the individual before he or she noticed it. By recognizing this Spirit in themselves – not through scholarly exegesis, but through faith, love, and deeds – the faithful converted their lives into one in accord with God's will. The separation between God and Man was mediated by the Spirit, so the elected came to see herself as enacting the will of God through her worldly action.

This brief illustration of the meaning of Spirit in Christian theology highlights several aspects retained in the secularized concept. First, Spirit is relational. Though there are two realms of experience – human and divine – the Holy Spirit mediates between the two. Second, Spirit is particularistic, while remaining universal. No contradiction between part and whole adheres precisely because the individual and God are mutually related through Spirit. Finally, the Christian encounters the Holy Ghost through nonrational experience irreducible to logical form or cognitive effort. Spirit is felt, not thought.

Montesquieu's *The Spirit of the Laws*

Though some aspects of the theological notions of spirit were retained, the term was secularized and grounded further within the mundane, human realm during the eighteenth and nineteenth centuries. The most influential source of this development was certainly Baron de Montesquieu's *De l'esprit des lois (The Spirit of the Laws)* (1989). As Raymond Aron (1998) noted, Montesquieu can be considered the father of modern social science insofar as his aim was "to analyze and to comprehend the social as such, both as element and as entity" (Aron, 1998: p. 10). *The Spirit of the Laws* reflected a revolution in legal writing, providing the lay reader access to the logic of given modalities of law, while adding a novel demonstration of law's fundamentally relational character. Various environmental

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03120-2

particularities, including climate, size, population, wealth, trade, manners, and custom, were essential conditioning factors in the function of different laws. These related environments particularized the function of law within given territories and societies. On one hand, the spirit of the law refers to the sum totality of all the relations of law to other elements. On the other hand, the 'general spirit, moral, and customs of a Nation' represents one essential element itself (Montesquieu, 1989: bk. XIX). This general spirit prepares people's minds to receive the best laws, thereby delimiting the possibility of wholesale overhaul of the law. To change the law, one must first address the spirit in which the laws are to be situated.

As a foundational Enlightenment text, *The Spirit of the Laws* contributed to a *trans*-European conceptualization of a spirit which lay behind and through the laws and practices of given societies. The spirit was no longer that of God, but a quality and product of the people themselves. Though, of course, influential in his native France, Montesquieu's influence was felt most firmly in England, as in Edmund Burke's conservatism, or in Sir Henry Maine's comparative legal scholarship. Appreciating the Frenchman's valorization of English institutions, Burke also asserted the priority of historical tradition which limited the possibility (or ethical validity) of wholesale political revolution outside the peaceful evolution of the customs, practices, and beliefs of a people. Similarly, Maine's comparative investigation of the historical shift from status-based to contract-based societies would have been inconceivable without Montesquieu's revelation of an object, 'spirit,' which historically prefigured emergent, positive laws.

Nation and Spirit in Romantic Germany

The history of the concept of *Geist* in Germany is part and parcel of the development of idealist philosophy, and therefore can begin with Immanuel Kant's posing of the question: how is valid knowledge possible? However, Kant's program was primarily epistemological, and it was through the rejection of his formalistic, systemic account of knowledge that future idealist philosophers developed their various notions of Spirit. These early criticisms of Kantian epistemology moved, generally, from the level of cognition toward that of experience. Relinquishing the primacy of the transcendental subject encouraged substantive inclusion of history, which was traceable to the spirit of the time or people. Furthermore, the dualisms and formalism of Kant's schemas tended to be contrasted against the organic unity of reality and humanity. The resulting Romantic notion of *Geist*, in a sense, stood as a place holder for this relational, experiential, nonrational unity.

These Romantic counterpositions were not univocal and developed in various directions. Again, this multiplicity is reflected in the numerous connotations the term, *Geist*, obtains in German thought. One line of conceptual development can be traced within idealism itself, leading from Kant to Fichte to Schelling to Hegel. Another can be read in parallel, which includes the hermeneutics of Herder, whose influence on legal science and history was perhaps most significant. Following the initial flourishing of idealism in Germany, however, intellectual trends, including materialism, positivism, empiricism, utilitarianism, and Darwinism, led the term into temporary disrepute until Wilhelm Dilthey recovered the 'science of spirit' '*Geisteswissenschaft*' in the late nineteenth century. It is important to recall the overlapping discursive contexts of the term's usage.

To Hegel

Since the dense history of philosophical idealism cannot be condensed easily, a summary of the line from Fichte to Schelling to Hegel can serve as a point of arrival at the high watermark for the role of *Geist* in German thought.

Johann Gottlieb Fichte was the first to take the inaccessible 'thing-in-itself' from Kant's noumenal realm to human consciousness itself, asserting the knowing subject becomes truly and authentically *Geist* in the course of wilful activity. Rather than subjective consciousness contrasted against an objective 'real,' Fichte recognized the intersubjectivity of self-consciousness accessible through human action.

Friedrich Schelling further extended Kant's observation that the world must be organized in a way that allowed consciousness access to reality to develop a natural philosophy in which the ideal springs from the real. The world becomes one of the ineffable experience and feeling. In contrast to the scientific knowledge given primacy by Kant and the Enlightenment, the Romantic world of Nature becomes the ultimate site of truth.

Hegel's philosophical system represents the synthesis of the three positions of Kant, Fichte, and Schelling. In *The Phenomenology of Spirit*, he criticized Kant's formal system as a synoptic index – 'like a skeleton with tickets stuck all over it' – which loses connection to living reality (Hegel, 1967: p. 110). Schelling's Romantic alternative vision of an Absolute world of sensation and feeling proves no solution as Kant's bright noonday light becomes "a night in which all cows are black" (Hegel, 1967: p. 79). Mediating between the two extremes, Hegel extended Fichte's emphasis on self-consciousness by positioning the emergence of reflexivity within a historical trajectory of the Absolute Idea. Over time, the Absolute has developed the means through which it can finally become conscious of itself.

Hegel's synthesis extends Schelling's Idea as objectively encountered in Nature. And yet, reality is always being *realized* through an objective process moving teleologically toward higher levels of self-knowledge. Idealist philosophy has provided the basis upon which Spirit now recognizes itself as "the grandest conception of all, and one which is due to modern times and its religion. Spirit is alone Reality. It is the inner being of the world" (Hegel, 1967: p. 86).

Hegel introduced his conception of *Geist* in *The Phenomenology of Spirit* completed in 1806 at the very moment French Revolutionary forces entered his home town of Jena. Indeed, Hegel described Napoleon Bonaparte as the World-Spirit (*Weltgeist*) on horseback. As the embodiment of the concrete universal, Napoleon, like other great men of history, realized objectively the unfolding Absolute. This social philosophy of history is revisited within the *Philosophy of Right* written in 1821. The State is objectified spirit for Hegel, which is achieved to the extent social freedom is actualized in law. Coercive prohibitions allow individuals to go about their business without

interference, while the law reflects the morals and will of the people. The citizen confronts the State as an ordered, objectified Idea, which is rational, not in a formal logical sense, but rather as dialectic moving through the spiral of historical trajectories and higher planes of integration.

Hegel's political theory obtained an immense influence within the Prussian state during the *Vormärz* (pre-1848) period. His conception of the State as the objective embodiment of the World-Spirit provided conservative justification for the status-quo and the ultimate authority of civil servants, like himself as professor of philosophy at the University of Berlin.

To the German Historical School

Hegel's philosophical system certainly provided the most prominent use of the term, *Geist*. Yet, other related scholars and discourses developed the term in directions that limited the concept in such a way as to ground the object in society, language, and culture, thereby encouraging the treatment of spirit through social science. Indeed, scholarly attention to spirit can be seen to be the very source of the emergence and proliferation of social science in German academia, beginning especially within the historical legal sciences of Friedrich Carl von Savigny and Jacob Grimm.

Before encountering the German Historical School's recognizably social scientific, empirical, and historical treatment of societies, we must step back to revisit the hermeneutic inspiration provided especially by Johann Gottfried Herder's concept of '*Volksgeist*.' Indeed, Herder should be read as the most important theorist of Spirit since even Hegel's broader conception was, in part, inspired by the earlier theologian. In 1769, Herder coined the term '*Zeitgeist*,' defined as the general cultural, intellectual, ethical, spiritual, and political climate of a given era. This reflected Herder's essential role as midwife of history, which thoroughly penetrated German consciousness in years to come. And, just as *Zeitgeist* cast, the general spirit in diachronic time, the term, *Volksgeist*, defined as 'national culture,' grounded the overall climate of particular cultures and peoples in territorial space.

Later commentators have seen Herder as a source of racial thinking insofar as particular groups' national character and history are determined by their 'nature' and 'heredity.' There is certainly merit in this interpretation, especially in light of later inspiration drawn by National Socialists and other racist ideologies. However, Herder's conception of the *Volk* was a suprabiological category. Spirit was grounded, not in blood or physical attributes but in culture and language; a unique manifestation of groups and societies reflected in "dress, food, customs, domestic economy, arts, and amusements" (Herder, 1803: p. 8). Thus, Herder was among the first proponents of cultural studies and anthropology.

The *Volksgeist* is 'natural' in that it engenders a real and constraining force upon the development and character of nations and their citizens – connecting all classes and subgroups into one community. Herder provided an original theory of nationalism from which later modern forms derived – a vision of a unified German spirit nearly a century before a united German empire came into legal existence in 1871.

Herder's philosophy of history moved in the direction of social science when Friedrich Carl von Savigny produced an influential pamphlet titled *Vom Beruf unserer Zeit für Gesetzgebung und Rechtswissenschaft (On the Vocation of our Age for Legislation)* (Savigny, 2002) during the course of a debate over legal codification with the jurist Anton Friedrich Thibaut in 1814. The document would become the foundation of the German Historical School of Law. Savigny suggested that the law was derived from the *Volksgeist* (spirit) of the population; the result of custom, language, and the particularities of the 'nation' from which the law emerged. At the same time, an internal, organic drive compelled the evolution of law according to 'silently operating powers' (Reimann, 1989). In contrast to the formalism and positivism of Thibaud's projected codification project, Savigny insisted that too little was known of the actual cultural spirit of the German people. Without precluding the possibility of codification in the future, the task of a realistic historical jurisprudence managed by trained legal scientists was a necessary precondition for successful legislation.

Savigny separated the study of legal science into two modes – the systematic and the historical. With reference to the Roman *Institutiones* collated during the sixth century CE reign of Justinian I, Savigny argued for the systematic treatment of the abstract relations between laws and civilization. However, equally essential was historical analysis of the particularities of the *Geist* at different times and places. The historical school Savigny inaugurated became consolidated around the *Zeitschrift für geschichtliche Rechtswissenschaft* founded in 1815 with Karl Eichhorn and Johann Göschen. Though Savigny, like Hegel, was criticized in 1848 as a spokesman of the conservative forces of the *Vormärz* period, the *Zeitschrift* had been one of the primary vehicles for political criticism during a period of radical censorship in German lands as a result of Metternich's Carlsbad Decrees. Embedded within the arcana of Roman historical detail were indirect claims for ideal constitutional reform. Though presented as the objective description of historical fact, legal science was a Trojan horse for political science and normative political theory. Romanist scholars, like Savigny and Berthold Niebuhr, focused on the ideal freedom obtained in the early Roman Republic, contrasted against the conflict and chaos of the later Empire. In so many words, liberal Republicanism for Germany was projected on and through historical Roman images. All the while, the conservative influences of Montesquieu, Burke, Herder, and Hegel came through insofar as no revolution was projected; rather, careful thoughtful evolutionary reform according to tradition and national spirit.

A second wing of the legal science movement dedicated attention to scholarship into existing German culture itself. Through the work of Eichhorn and especially Jacob Grimm, who collected the folktales and customs of German villagers with his brother, an image of the German *Geist* was obtained and interpreted by the Germanists. Images of the *Gemeinwesen*, or, civil community, were constructed out of the etymological and philological interpretation of symbols in faerie tales. For example, Grimm traced the legal concept of *Lösegeld* (compensation in tort claims) to the archaic form of the word 'Otter,' thereby connecting to a story of a hunter who infringes upon the property rights of farmer whose son had transformed into an otter (Crosby, 2008: p. 112). Similarly, the familiar Cinderella story of the glass slipper should be understood in

terms of the custom of the groom fitting a shoe on his virgin bride as a symbol of property ownership exhibited before the community. Grimm and his liberal legal scientific colleagues sought enshrinement of these traditional customs in the rapidly developing German legal apparatus, which they often perceived as aristocratic, elitist, and overly rationalist, in contrast to the organic unity of the common German national spirit.

The intellectual debates between the Germanist and Romanist wings of legal scholars encouraged, on one hand, considerable investment in historical scholarship – especially in Classical texts. On the other hand, the application of linguistic study to their own cultures and peoples secured description of the object *Geist* in the present, and was the first iteration of anthropology, sociology, and ethnology as recognizable within modern disciplines.

The Decline of *Geist* and the Rise of *Geisteswissenschaft*

In large part due to the arrogance and totalization of Hegel and his followers, several anti-idealist reactions set in within German philosophical thought after 1830. Empiricism, positivism, and materialism – for example, in Ernst Haeckel's Darwinian evolutionism – replaced the metaphysics of *Geist* with that of concrete reality verifiable through scientific procedures. Politically, the Left Hegelians, including Marx and Engels, demanded the flipping of idealism upon its head, so that culture became interpreted as the product of economic and material interests rooted in class conflict. Similarly, in legal science, rejection of the perceived conservatism of Savigny and his followers contributed to a returned interest in codification and positive theories of law as outlined by Rudolf Ihering. This utilitarian analysis of law and the product of will and conflict led directly to the next phase of legal science and political economy, in which Ferdinand Tönnies, Max Weber, and Werner Sombart were inculcated.

Prominently, the German historical school of economics surrounding Gustav Schmoller, Lujo Brentano, and others rejected the abstraction of both the classical English political economy – Manchesterismus and Hegelian idealism. Focusing instead on empirical material, especially statistics and monographic surveys of urban conditions, the German historical school further extended the trend toward objective study of contemporary society. At the same time, historical interest in Prussian bureaucracy resulted in the *Acta Borussica* in 1887 – a team research project led by Schmoller and Otto Hinze. Thus, both the historicist particularism of the jurisprudential tradition were maintained, while the efforts away from grand, abstract theorizing brought researchers 'closer to the ground,' where they could address issues of social reform according to local circumstances.

This was a period of frequent visitation from American scholars, who would become the founders of many of the early social science departments in America, including John W. Burgess, Albion Small, Edward Ross, and others. The influence of both historical legal studies and historical economics cannot be underestimated. The German scholars' principles of empiricism and social reform found their way into historical institutionalist economics, sociology, and political science.

Meanwhile, within Germany, Wilhelm Dilthey struggled to rehabilitate and revise the science of Mind, which he called *Geisteswissenschaft* (Ermath, 1978). Unlike Hegel, Dilthey did not posit an Absolute teleological entity operating through the totality of the cosmos. Rather, Dilthey's historicism recognized that the concepts that actors, communities, and discourses use are constantly shifting and evolving, so the context in which action takes place is constantly changing and becoming something new. Like Fichte, Dilthey emphasized the role of self-consciousness in this process. Indeed, the human sciences must be distinguished from the natural sciences precisely because conscious reflection is built into human experience, so that our confrontation with human life is not a confrontation with passive external objects as in natural science. An adequate philosophy of knowledge would engage with the history of knowledge itself. Thus, the primary science should be that of 'Mind' – hence, the priority of *Geisteswissenschaft*, or humanities, as the basis for a science of science.

Dilthey's conception of psychology differed from the materialist and empiricist variants developing concurrently through the experimental work of Gustav Fechner, Hermann von Helmholz, and Wilhelm Wundt. In this sense, his work stands in closer relation to Friedrich Schleiermacher's social psychology, as well as the *Lebensphilosophie* (life philosophy) of Henri Bergson, Friedrich Nietzsche, Edmund Husserl, and, later, Martin Heidegger. Dilthey insisted that the object of human scientific study is 'life' – the processual energy flowing through the concrete course of conscious experience. Though constantly changing, life contains relatively stable coherence and structure which can be analyzed and studied scientifically.

Spirit's Tail: The Lost Legacy of the Concept of *Geist*

Despite Dilthey's best efforts and initial success in carving a space for *Geisteswissenschaft*, developments in Germany and elsewhere during the course of the twentieth century led to changes in the human sciences that relegated the concept of *Geist* to the dustbin of history. Most significant in this decline was the political use of the term *Volksgeist* during the Nazi era. Drawing on Herder's nationalism, Adolf Hitler developed the ideology of the Third Reich according to a racist image of the Aryan race and the German *Volk*. Ever since, the idea of positively studying an objective national spirit has become suspect within the social sciences.

Well before this nail in the coffin, however, the term was losing its salience, even as the approach engendered by its use was leaving final imprints on the emerging disciplines of sociology, legal realism, psychology, anthropology, and philosophy. During the late nineteenth century, popular mystical movements associated with 'spiritualism' grew in prominence within English-speaking countries especially. Spiritualist groups sought access to transcendental realms via Mediums, tarot, astrology, and other practices since deemed 'pseudoscientific' (Brandon, 1983). Broadly linking with novel European interest in Eastern religions, scientific advances in electromagnetism and totalizing metaphysical philosophies including Theosophy, Anthroposophy, and Occultism, these discourses were typically rejected by academia. This exclusion reflected not only nonconformity with prevailing scientific methods, but also the over-representation of women in these circles (Braude, 2001). Still members of spiritualist societies overlapped with

'public' sociological, nationalist, and political movements, including the social reconstruction efforts of Patrick Geddes in Scotland, women's suffrage movements, and, indeed, Comte's Positivist religion of humanity. In the long run, these associations between 'spirit' and spiritualism only made matters worse, as pre-World War II social philosophies of this kind were lumped together in a general outdated and dangerous 'irrationalism' – quite out of fashion in the postwar positivist mainstream.

Nonetheless, many social scientific disciplines had already been marked by the ideas and agendas of prior scholars' interest in *Geist*. Franz Boas' four-field anthropology, for example, is widely recognized as a foundational paradigm for modern anthropological study. And yet, Boas' nonracial, culturalist, and systematic study has been described by George Stocking (1996) as "*Volksgeist* as Method and Ethic." Just as the legal scholars, historical economists, and *Geisteswissenschaftlicher*, in Germany distinguished the 'historical,' 'positive,' and 'empirical' from the cultural, meaningful, and abstract – so, too, did Boas' separation of cultural anthropology from bioanthropology, archaeology, and linguistics reflect the organic unity of differentiated analytic spheres.

Similarly, the legal realism of Oliver Wendell Holmes, and, indeed, pragmatist philosophy more broadly would have appeared out of an historical vacuum without reference to Hegelian idealist influences (Menand, 2002). To this day, the social sciences struggle to connect the logical contradictions between the dualisms of individual/society, structure/agency, and so forth. In so doing, we follow the contours of problems that earlier scholars of *Geist* often found noncontradictory – precisely because the mediating, relational concept of spirit provided the organic connection between universal and particular. Contemporary terms – e.g., habitus, network relations, reflexivity, etc. – provide substitutes with greater technical precision, but, perhaps, less ethical force; and, certainly with less of the 'spirit' of the previous Age at the roots of our disciplinary history.

The 'Spirit of Capitalism' in the German *Geist*

We can thus return to the discussion of Weber's classic study of the 'spirit' of capitalism. Some scholars have rightly drawn attention to Weber's training in the historical jurisprudential tradition of Wolfgang Mommsen, and the homology between his categorical scheme and that of Rudolph Ihering (Turner and Factor, 1994). Others have noted the influence of the 'younger' German historical school of Schmoller et al. (Shionoya, 2006). Certainly, Weber was extending these traditions into new territory, and in so doing, he helped differentiate sociology from these older traditions.

In its most specific context, Weber's work was an extension of Werner Sombart's own description of capitalism as an 'objectified *Geist*' embodied in new forms of organizations and technology. Sombart's cultural source for this spirit of acquisitiveness, competition, and rationality came from the bourgeois mentality emerging from the Renaissance. Weber's contribution repositioned the same spirit, within the paradoxical source of ascetic Protestant sects. Without diminishing Weber's achievement, he was not necessarily the 'father' of all that he is given credit for in contemporary social science. The notion of 'objected Spirit,' the notion of *Weltanschauung*, and the *Verstehen* approach were recognizable in Dilthey's *Geisteswissenschaft* program, and Weber's 'ideal-typical' approach was only a few generational iterations removed from the early post-Kantian idealists, Hegal and Herder.

In recalling the admittedly archaic terminology in his foundational essay, Weber recovered the utility of addressing an object which was familiar to contemporary German scholars. Though positive usage had lapsed by 1905, the term *Geist* nonetheless retained a technical terminological thrust even in his own time. Though this 'object' has since been equated with culture, or religion, or rationalization depending on one's interpretation of the Protestant Ethic thesis, we are wise to position Weber's innovative analysis within the ideational context in which it historically emerged. Indeed, in so doing, we recover the original historicist spirit of these scholars' foundational contributions to the human sciences.

See also: Anthropology, History of; Community and Society: History of the Concepts; Cultural Sociology, History of; Enlightenment: Impact on the Social Sciences; National Character, History of; Political Theory, History of; Sociology, History of.

Bibliography

Aron, R., 1998. Main Currents in Sociological Thought. Transaction Publishers, New Brunswick, NJ.
Brandon, R., 1983. The Spiritualists: The Passion for the Occult in the Nineteenth and Twentieth Centuries. Weidenfeld and Nicolson, London.
Braude, A., 2001. Radical Spirits: Spiritualism and Women's Rights in Nineteenth-Century America, second ed. Indiana University Press, Bloomington.
Crosby, M.B., 2008. The Making of a German Constitution: A Slow Revolution. Berg, Oxford and New York.
Ermath, M., 1978. Wilhelm Dilthey: The Critique of Historical Reason. University of Chicago Press, Chicago.
Hegel, G.W.F., 1967. Phenomenology of Mind. Harper Torchbooks, New York.
Herder, J.G., 1803. Outlines of a Philosophy of the History of Man. Johann Gottfried von Herder, Trans. from the German of John Godfrey Herder by T. Churchill. Printed by Luke Hansard for J. Johnson.
Menand, L., 2002. The Metaphysical Club: A Story of Ideas in America, first ed. Farrar, Straus and Giroux.
Montesquieu, C. de S. baron de, 1989. Montesquieu: The Spirit of the Laws. Cambridge University Press.
Reimann, M., 1989. Nineteenth century German legal science. Boston College Law Review 31, 837.
Savigny, F.K.V., 2002[1831]. Of the Vocation of Our Age for Legislation and Jurisprudence. The Lawbook Exchange, Ltd.
Shionoya, Y., 2006. The Soul of the German Historical School: Methodological Essays on Schmoller. Weber and Schumpeter. Springer.
Stocking, G.W., 1996. Volksgeist as Method and Ethic: Essays on Boasian Ethnography and the German Anthropological Tradition. Univ of Wisconsin Press, Madison.
Turner, S.P., Factor, R.A., 1994. Max Weber: The Lawyer as Social Thinker. Taylor & Francis Group.
Weber, M., 2002. The Protestant Ethic and the "spirit" of Capitalism and Other Writings. Penguin Books, New York.

Relevant Websites

www.academia.edu/428817/BASIC_CONCEPT_OF_SAVIGNYS_VOLKSGEIST – Basic concept of Savigny's Volksgeist.
www.counter-currents.com/2011/05/herders-theory-of-the-volksgeist – Herder's Theory of the Volksgeist
www.counter-currents.com/2011/05/savigny-the-volksgeist-and-law – Savigny: The Volksgeist & Law.

Gellner, Ernest (1925–95)

Chris Hann, Max Planck Institute for Social Anthropology, Halle/Saale, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ernest Gellner's prolific writings were anchored in his early philosophical studies in Oxford, notably of Hume and Kant. As a social scientist he drew on Weber, Durkheim, and Malinowski, rejecting both the economic determinism of Marxism and all relativist, idealist currents. His models of nationalism, Islam, and civil society have been much debated, and he exercises a diffuse influence in historical sociology and social anthropology. Gellner's social philosophy of modernity has affinities to that of Popper and Hayek, but he is distinctive because he couples his insistence on the cognitive superiority of Western rationalist individualism with a deep appreciation of the diversity of human communities.

Introduction and Life

Ernest Gellner was a self-proclaimed 'Enlightenment fundamentalist' with a unique grasp of the world in which he lived. He ignored boundaries between philosophy and several branches of social science, working against the grain of his generation in all of them simultaneously. Though he spent almost all his adult life in Britain, where he became an influential public intellectual (to the limited extent that Britain is able to host this species), he was also a Central European and a much-traveled multilingual cosmopolitan. Gellner was born in Paris on 9 December 1925 and spent his childhood in multiethnic Prague. His family moved in perilous conditions to England in 1939 following the Nazi invasion; some close relatives did not survive the Holocaust. The significance of these events and of his Jewish middle class background for his work has been carefully explored by his biographer (Hall, 2010). At Oxford University, where his studies were interrupted by military service in the Czech Brigade, Gellner concentrated on. After teaching this subject for 2 years at the University of Edinburgh, he joined the Department of Sociology at the London School of Economics (LSE) in 1949. Here he concluded that British academia had little to say about what really mattered in the postwar world. However, positive stimulus came from his discovery of the cohesive intellectual community of social anthropologists in the department established by Bronislaw Malinowski. Gellner established his reputation with *Words and Things* (1959) a withering critique of the linguistic philosophy that was dominant in Oxford. In 1962, he was awarded a doctorate in social anthropology, based on fieldwork in Morocco which he had pursued alongside mountaineering expeditions in the High Atlas. *Thought and Change* (1964) was a vigorous declaration of intent toward the social sciences, anticipating many of the themes that he followed up in later years, including his celebrated theory of nationalism. After two more highly productive decades at the LSE, in 1984 Gellner was appointed to succeed Jack Goody as William Wyse Professor of Social Anthropology at Cambridge, where he was also elected to a Fellowship at King's College. After retiring, from 1991 he spent much of his time in Prague directing the Center for the Study of Nationalism at the newly established Central European University. He died in Prague on 5 November 1995.

Philosophy, Anthropology, and World History

Gellner's interests extended well beyond the conventional boundaries of the subjects in which he held academic posts. His prolific writing was always anchored in his early philosophical studies, particularly in Hume and Kant, augmented later by his readings of Weber, Durkheim, and Malinowski. In short, Gellner drew on classical sources in several disciplines. Negative influences were equally important. Marxism was always rejected, ostensibly because of its extreme economic determinism, though his encounters with Soviet power in Prague shortly before he left the city for good in 1945 must also have influenced his position. Much of Gellner's own work emphasized a materialist standpoint and he later cultivated an enthusiastic curiosity toward Soviet Marxist anthropology (Gellner, 1988a; Skalnik, in Malešević and Haugaard, 2007). He was even more emphatic in his rejection of what he termed the relativism of the later Wittgenstein, and of the postmodern trends that became prominent across the humanities and social sciences in the last decade of his life. For Gellner, the truths which mattered were independent of 'culture.'

Some critics attacked Gellner with the old bogey of positivism, and some of his more polemical formulations do indeed expose him to such charges. However, in the fuller statements of his philosophical position (Gellner, 1975, 1985) he shows that the central epistemological problems that confront the social scientist are indeed necessarily different from those of the natural sciences. Gellner's distinctive 'metaphysic' (Hall, in Malešević and Haugaard, 2007) is a reconciliation of the scientific with the humanistic, with the ultimate goals of 'legitimating belief' and resolving the puzzles of world history, notably the emergence of 'modernity.' Our modern condition, he argues, requires the rejection of all comforting ideologies in favor of an austere 'bloodless universalism.' At the end of his most concentrated engagement with the work of Karl Popper, from whom he drew much inspiration while maintaining a critical distance, he concluded that "… the positivists are right. For Hegelian reasons" (1985: p. 67).

Gellner's combination of philosophy with anthropology had nothing in common with the continental genre of philosophical anthropology. It was more like a new form of *Universalgeschichte*. Together with the philosopher Ian Jarvie, he

hailed Bronislaw Malinowski as the founder of a social science that offered a better handle on the real world than idealist philosophers in Oxford or evolutionist sociologists at the LSE. When Gellner realized much later that Malinowski had obtained his doctorate in 1908 with a study of Ernst Mach's *Empiriokritizismus*, it was clear to him that in making long-term fieldwork among flesh and blood individual actors the revolutionary basis of anthropological knowledge, Malinowski had simply reverted to the ultra-positivism he had imbibed in his youth. Gellner's fascination with Malinowski led him, late in his career, to further explorations of the founder's intellectual roots in late Habsburg Cracow and to an illuminating (posthumously published) contrast with another product of the same imperial milieu, Ludwig Wittgenstein, whose path was so different (Gellner, 1998). Gellner was greatly attracted by the Pole's combination of cultural nationalism with political internationalism. He was somewhat dismayed to realize that Malinowski himself, alongside all the hard facts he collected on the Trobriand Islands, had made far-reaching concessions to the power of magical words (and thus to linguistic philosophy, the specter of which never ceased to haunt Gellner).

The synchronic character of Malinowski's functionalism was rationalized by Gellner as the natural aversion of a Pole to the Goddess Clio, who had treated his own nation so unkindly. By breaking with the speculative evolutionism of James Frazer and Edward Tylor, twentieth century ethnography, as pioneered by Malinowski, had set new scientific standards. Nevertheless Gellner always argued for reopening the older conversations about diachronic development. From this point of view, he engaged in debates with hunter-gatherer specialists such as his LSE colleague James Woodburn, prehistorians such as Paul Mellars in Cambridge, and numerous Soviet colleagues whose anthropological paradigm had never confined them to the synchronic. Scholars such as Anatoly Khazanov (for pastoral nomads) and Viktor Kabo (for hunter-gatherers) could not match Malinowski when it came to analyses of micro-level, short-term social dynamics, but their long-term visions were equally important for Gellner's grasp of social anthropology as both contemporary social science and world history.

Malinowski's functionalist holism was not the only component of British social anthropology which appealed to Gellner. A.R. Radcliffe-Brown's insistence on the comparison of social structures was still very influential in the 1950s. Gellner implemented this in own ethnographic work, which prioritized pattern and system, rather than close-up analysis of the flux of events (Gellner, 1969). He also took something of the spirit of Radcliffe-Brown's 'natural science of society' agenda with him when he turned to history. This culminated in a three stage philosophy of history, the fruit of his decades of study in various anthropological traditions and in historical sociology (Gellner, 1988b). His aim was to explain long-term evolutionary changes both in terms of the distribution of economic and political power and in terms of ideas and ideologies. Gellner developed these themes most productively in a famous seminar at the LSE which he led jointly with Michael Mann and John A. Hall. His manner of explicating entire eras and societies (or cultures, or civilizations) by means of elegant structural models or ideal types was very much at odds with most of his anthropological contemporaries, who stressed the importance of local knowledge and hermeneutic interpretation. More controversially, Gellner insisted, with increasing vehemence as time went by, that *one* culture, namely that of the West, based ultimately upon premises of science, Cartesian rationality, and individualism, was technologically and 'cognitively' more powerful than all others. To pretend otherwise, in the manner of the relativist philosophers and their admirers in anthropology who insisted that other knowledge claims were equally valid, was romantic nonsense. The challenge was how to explain the dominant power of the West and how best to spread the benefits of that superiority in our modern world.

For many critics, of course, this is an unacceptably Western way to pose the question. For example, Jack Goody, Gellner's predecessor in Cambridge, argues for the distinctiveness of Eurasia, not Europe, and is skeptical of theories proclaiming just one recent breakthrough to a condition called modernity. In Goody's view, China and Europe alternated at the forefront of science and technology from the late Bronze Age onward. By contrast, Gellner agreed with sociologist John Hall and historian Patricia Crone in emphasizing the uniqueness of Europe (he was also stimulated by the work of historical anthropologist Alan Macfarlane, his closest intellectual ally in Cambridge, but judged that this colleague went much too far in stressing the peculiarities of the English). His arguments tend to be as Eurocentric as those of Weber himself. Hall (2010) points out that Gellner attaches no self-congratulatory moral significance to the 'European miracle'; but it is often unclear in his work as to what exactly he means by this term (and similar images such as the 'great escape'). In economic history it is usually taken to refer to the transition from *Agraria* to *Industria*, the causes of which must be sought in new forms of polity and production. But for Ernest Gellner the key variable sometimes seems to be not so much industrialism but a vaguely specified novel mode of cognition. His biographer covers over this tension in his work by opting for the broader blanket term 'modernity.'

Nationalism and Modernity

There can be little doubt that, at least since the publication in 1983 of *Nations and Nationalism*, Ernest Gellner is best known for his work on nationalism. It is important to place this component of his work in the wider context of his interests throughout his career, and to recognize its genesis and early formulation in *Thought and Change* (1964). Nationalism was a driving force in the anticolonial movement in Morocco to which Gellner was a witness. More importantly, it was also the force which had broken up the multicultural empires of Eastern Europe and caused the catastrophes that had clouded his family history. The benign, emancipatory potential of nationalism was evident in the successful nation-states of Britain and France, but how could such models be emulated in what scholars were beginning to term 'the Third World'? This was the era of modernization theory, led of course by the United States, the dominant global power. The fragmentary archival sources that we have for these years show that Gellner, too, shared a widespread postwar optimism that the deployment

of new technologies could eliminate hunger, poverty, and 'nastiness' from the human condition once and for all. Much later, in a remarkable letter to Perry Anderson in 1981, he professed to being a principled Philistine, ready to accept a dumbing-down of European high culture in the interests of social progress ("… I am prepared to pay the price of vulgarity for peace, reasonable diffused prosperity and equality" – cited in Hall, 2010: pp. 159–60).

According to Gellner, the most basic condition for this modernist utopia was a new form of polity, one in which the boundaries of the state would correspond to those of the culture. Whereas his LSE colleague Elie Kedourie had emphasized the dissemination of the nationalist *idea* from its origins in Europe, Gellner stresses the uneven spread of industrialism and the functional needs of an industrial society, which requires a highly mobile labor force and educational institutions that promote literacy in a single 'high culture.' The individual human being emerges from such training as a 'modular man,' capable of performing many tasks in the course of his career (Gellner was notoriously insensitive in his work to the dimension of gender). Consequently, the whole society was characterized by high levels of both occupational and social mobility. This society was radically egalitarian in comparison with the agrarian empires of earlier centuries. This model, elaborated in several variations over more than three decades, was supported not with fieldwork materials but with vivid suggestive sketches, notably that of 'Ruritania' (1983), to indicate the combination of economic, political and emotional factors that could lead the elites of peripheral cultural groups to mobilize the masses to pursue an independent state. The Ruritanian allegory was evidently colored by his Central European roots, and Gellner may also have been influenced by earlier scholars with a background resembling his own, such as Hans Kohn and Karl Deutsch. In his later works (Gellner, 1994, 1996), he elaborated four 'time zones' for the development of nationalism, moving eastward across Europe from the Atlantic seaboard.

Gellner's model was an elegant general theory which, with its stress on structural principles, neatly complemented Benedict Anderson's emphasis upon 'imagined communities,' in a well-known study published almost simultaneously with Gellner's definitive statement (Anderson, 1983). Gellner felt that Anderson's emphasis upon the new technologies of print capitalism failed to acknowledge the enormity of the transformation from agrarian empire to industrial nation-state. "It must be repeated that nationalism is a phenomenon of *Gesellschaft* using the idiom of *Gemeinschaft*: a mobile anonymous society simulating a closed cosy community" (Gellner, 1996: p. 74).

The Gellnerian model has attracted both admiration and criticism (Hall, 1998, 2010: Chapter 10). Some of the problems are those inherent to functionalist theories generally, though these have their defenders (see Mouzelis in Hall, 1998; in Malešević and Haugaard, 2007). Empirically, the Gellnerian model is not so easy to operationalise. Historians and above all historical sociologist Anthony Smith have pointed out that many, if not most modern nations demonstrate more complex continuities with the ethnic groups of the agrarian age than Gellner's model allows. His account of nationalism's geography is equally problematic. Even within Central Europe, one finds Ruritanias such as Slovakia (which Gellner visited as a Boy Scout in the 1930s, and probably supplied the inspiration for his parable) jostling alongside both newly industrialized regions such as Bohemia and old gentry *nacji* (such as that of Poland, one to which Malinowski was proud to claim allegiance, which had once possessed a state but contrived to lose it). The patterns, where there were any, could hardly be reduced to the uneven spread of industrialism. In short, the model does not work, not even in Bohemia (Hall, 2010: Chapter 10). It offers little help in predicting the circumstances in which national sentiment becomes 'hot,' partly because the emphasis is placed on economic development and there is no adequate engagement with (geo)politics, international relations, the growth of bureaucracy, and warfare; and partly because incalculable human subjectivities are a key element in the model, notably elite consciousness of humiliation and propensities to resist and/or to cross cultural boundaries.

Hall (Hall, 2010: Chapter 10) points to a major 'wobble' in Gellner's views following the collapse of the Soviet Union. He seems at this point to have changed his mind concerning the basic desirability of the nation-state. The structural need for social homogeneity, which had destroyed the cosmopolitan world in which he grew up, suddenly seemed less attractive in the 1990s. Repeated eulogies of Franz Joseph reflect the direction of his thought in these years, even though he remained reluctant in print to defend federations created by socialists. Nor did he engage closely at all with the new 'multicultural mess' (Eriksen, in Malešević and Haugaard, 2007) in the old-established, nation-states. These weaknesses limit the usefulness of Gellner's work on nationalism and cultural diversity, even for the European context which inspired it; but it remains the most fruitful ideal type in the social science study of nationalism.

Islam

When it comes to Islam, Gellner's abstract model was nourished not with childhood memories but with data from many years of fieldwork in the High Atlas mountains of Morocco. His monograph, *Saints of the Atlas* (1969), an analysis of how charismatic religious specialists contribute to the maintenance of political order in the absence of state power, is a classic of the British social anthropological school. Intellectual inspiration was drawn from the segmentary lineage theory of the French tradition (Durkheim, and behind him Masqueray and Fustel de Coulanges) as well as from E.E. Evans-Pritchard in Britain. But it was typical of Gellner that, in expanding upon the significance of his local study, he invoked the cyclical models of tribal–urban interaction developed centuries earlier for the Arab world by Ibn Khaldun, and managed to combine them with the religious sociology of David Hume, notably his concept of 'enthusiasm.' Gellner generalized his composite 'pendulum swing model' to the entire Moslem world (initially in the 1960s and later, greatly elaborated, in Gellner, 1981: Chapter 1). It followed from his schematic view that the strict text-based high culture of Islam made it the most modern of the great religions. It was deeply puritanical. Given Weber's famous thesis of the elective affinity between Calvinism and the 'spirit of capitalism,' it seemed to follow that Islamic

societies would be receptive to the modernizing forces of secularism and nationalism. In later years, the growing strength of 'fundamentalism' in the Islamic world was explained in terms of the availability of this modern puritanical faith as a compensatory resource to populations unfortunately prevented from gaining access to the Western institutions of the nation-state and civil society.

Gellner's penduum-swing model of Muslim societies has not been taken up to the same extent as his model of nationalism. Even in the local Moroccan cases, his application of segmentary theory has been heavily criticized by later fieldworkers (see Munson in Hall and Jarvie, 1996). Whereas Gellner showed flexibility in modifying his theories of nationalism to the end of his life, he defended his ideal type of Islam intransigently. Hall (2010: Chapter 9) joins many previous critics in judging Gellner guilty of 'essentializing' Islam, with models that are 'crude' and even 'a little facile.' His biographer is at the same time respectful of Gellner's extension of the Weberian thesis, seeing it as a vital component of his general theory of modernity. In effect Gellner postulates a more 'generic Protestantism,' one which posits a 'rigid and austere deity [who] had no cognitive favourites,' as a precondition, or at least a significant enabling factor, for the great escape from agrarian society. But if the *ulema* of the Islamic city epitomize the puritan emphasis on texts and direct communication with a jealous God, then why was Islam not at the heart of this breakthrough? Gellner has no answer. His arguments that Islam is well adapted to modernity and thus highly immune to secularization are not borne out in the social realities of contemporary Muslim communities. Yet Gellner was no crude Orientalist: he was certainly respected in Morocco, had numerous distinguished Muslim students, and arguably possessed a deeper, more sympathetic understanding of Muslim communities than the scholars with whom he tried to debate these matters, notably Edward Said and Clifford Geertz.

Socialism and Civil Society

The Marxist–Leninist regimes of twentieth-century Eurasia were another form of social organization – and of faith – that fascinated Gellner. As with nationalism the reasons had much to do with his family history. Although he visited the Soviet Union frequently from the early 1970s and spent a year in Moscow at the height of *Perestroika*, most of his research was textual. As noted above, he was deeply interested in the questions of social evolution addressed by Soviet ethnographers, who continued debates with colleagues in archaeology of a kind that had largely ceased in the West following the Malinowskian revolution. At the same time Gellner delighted in reading between the lines of ethnographers' adaptations of Marxist theories in order to see what light these might throw on Soviet society (1988a). He did not predict the demise of the socialist system. However, his first-hand observations of socialist disenchantment in the 'squalor' of the Brezhnev years confirmed his materialist intuition that no doctrine promising salvation here on this earth was likely to endure if the ideologues could neither satisfy accelerating consumer expectations nor control corruption and widening inequalities.

In his last years, Gellner embraced an ideal of 'civil society' as the alternative to the Marxist-Leninist dystopia. Strong mediating institutions between the state and its citizens were hailed as the necessary basis for a pluralist democracy that would avoid the disasters of socialist central planning and all other forms of totalitarianism. *Conditions of Liberty* (1994) has much in common with Samuel Huntington's contemporaneous *The Clash of Civilizations* (Huntington, 1996). We see here a significant shift in Gellner's values from the early 'philistine' endorsement of postwar development policies. By the 1990s, with the Cold War over, he is content to sacrifice living standards and accept the massive widening of inequalities as the price to be paid for 'liberty.' He stops short of endorsing the 'free market' policies of the neoliberals, but insists that the role of the state must be clearly differentiated from the economy. 'Our' civilization, epitomized by a 'modular man' who opts in and out of free associations within a nationally defined high culture, is contrasted not only with Soviet totalitarianism but also with the 'tyranny of cousins' that Gellner had known in Morocco. Neither 'traditional man' nor Homo Sovieticus have any room for individual self-expression, in the former case due to a 'demanding culture' and in the latter due to the fatal combination of modern technologies with the monist pathology of Marxism, which in Russia was a seamless extension of Byzantine Caesaro-Papism. Only a separation of hierarchies in an open, liberal society can save us, argues Gellner. However, it is still not possible to 'validate' this Western ideal, since that step would imply some new form of absolutized ideocracy, which is disqualified in his definition of modernity. This is Gellner's final position: "We have to live with this" (1994: p. 214).

With this late work, to an even greater extent than in his earlier studies of world history, of nationalism, and of Islam, Gellner laid himself once more open to charges of ethnocentrism, or perhaps more precisely, Eurocentrism. The shrill fetishism of civil society reflects both a harshening of his political stance and a coarsening of his rhetorical skills. Gellner seems to forget the healthy starting point of his critique of the anthropologist-philosophers of the 1950s. He himself had once shown that real flesh and blood individuals enjoyed considerable room to maneuver, despite the seemingly rigid concepts and rituals that structured tribal Morocco. In the course of his career, he had supervised and examined innumerable ethnographic studies documenting the concrete realities of diverse societies all over the globe, including the mediating institutions of Islamic and socialist societies. Yet in the last major book published in his lifetime he forgot all this and argued instead that a specifically Western civil society was the right way to fill the void or 'vacuum' left by Soviet socialism. He paid little attention to what anthropologists and sociologists were already reporting about the high social costs of the shift to new forms of politics and economics throughout the former Soviet bloc. In short, he argued in the manner of a politically conservative linguistic philosopher, rather than a critical, empiricist, social scientist. The slogan 'civil society' supplied the later Gellner with an ultimate ethical grounding, though he had spent most of his

life insisting that no such grounding was possible and a formal 'validation' to support its prescription to the world was even now out of reach.

Impact and Assessment

Testimony to Gellner's influence in his lifetime can be found in a remarkable multidisciplinary collection of critical essays, to which he delivered an eloquent reply shortly before his death (Hall and Jarvie, 1996). One obvious key to his fame lies in the topicality of the subjects he treated: nationalism, Islamic fundamentalism, socialism, and civil society were all vital problems of the age. Another is the appeal of a scholar who was not merely disrespectful of disciplinary barriers but who consistently opposed many of the most influential intellectual currents of the day. His critique of Oxford philosophy was followed by equally irreverent dismissals of many other 'churches,' which space has not allowed me to discuss in this essay: Parisian structuralism, Frankfurt neo-Marxism, psychoanalysis, the 'Orientalism' paradigm of Edward Said, and every variety of postmodernism. Even his relationship with social anthropology, the discipline in which he came closest to finding a home, was uneasy. He remained basically loyal to the canons of Malinowski's functionalism and Radcliffe-Brown's structural-functionalism long after these had been abandoned by the profession, somehow combining these preferences with an even less fashionable sympathy for evolutionist agendas. Gellner's anti-relativism led to countless sharp disagreements in anthropology, for example, with Rodney Needham over the nature of kinship, and with Clifford Geertz, who he recognized as the leading contemporary exponent of the erroneous idealistic concept of culture.

Neither anthropology nor any of the other disciplines to which he contributed, has a Gellnerian school. With the hindsight of two decades his positions in numerous areas seem decidedly old-fashioned: for example, his blindness toward gender issues; toward the ecological consequences of the economic growth which was to overcome 'nastiness' on our planet; and toward multiculturalism in countries such as Britain, with all its implications for his theory of nationalism. And yet his ideal types and allusive fables of modernity have proved endlessly stimulating. Gellner's considerable impact as a public intellectual owes much to his forthright defense and provocative reformulations of universalist enlightenment ideals in a century in which these foundations came to seem increasingly fragile to the great majority of philosophers and social scientists in the West. An iconoclastic temperament and a distinctive aphoristic literary style helped him in communicating his message. This stubborn defender of outmoded ideals of rationality, clarity, and above all truth, delighted in controversy and had a preference for striking images and analogies over footnotes and the more prosaic techniques of social science. An integral part of his appeal to some, his acerbic wit and affectation of anti-intellectual positions only intensified his bogeyman status for others.

Indeed, Gellner was not the most painstaking of scholars – neither in the field nor in the library. He is best seen not as a backward-looking defender of European traditions, but as a thinker who strove to adapt a European liberal heritage to reach a better understanding of the central problems of the world in which he lived. Caricatures such as that of enlightenment fundamentalist are wide of the mark (even if he himself regularly suggested such labels). Despite his criticism of Wittgenstein and of the relativist trend in anthropology, he was no more ready than Malinowski to dispense with the concept of culture. He insists on the uniqueness of modern Western cognition because it really is stronger than any other. He embraces it personally, along with what he takes to be its socio-political corollaries, above all civil society; but even at the end of his life, in the aftermath of the Cold War, when he applies all his rhetorical genius to convince us that liberal individualism is not just cognitively more powerful but also *better*, he concedes that he cannot validate this. Compared to Popper and Hayek, this Central European liberal has a nagging awareness that cohesive human communities require stronger social foundations than those which the modern West has been able to provide.

This lingering doubt, an implicit balancing act frequently disguised by the swagger of his prose, make Gellner's quest a continuing source of inspiration in the new century. He would be the first to applaud new ideal types that upstaged his own: original fables to replace his account of Ruritania in order to capture new empirical realities. Ernest Gellner was himself a remarkable 'modular man' – not of a national community but of twentieth century Western scholarship. In an age of specialization, a scholar who transcends the boundaries is easy to dismiss as a gadfly or a maverick. Gellner made the most of the outsider position thrust upon him by his early life. His work has a high degree of unity and consistency. But there are also significant modifications, 'wobbles,' and subtleties and ambiguities easily missed if one gets lost in his polemical pyrotechnics. Gellner repeatedly trivialized relativist positions as mere 'cultural wallpaper,' but he was not disrespectful of 'difference.' What he objected to was 'excessive charity' toward complex concepts, which led some anthropologists to assert radical alterity and fail to recognize the common sense knowledge of their subjects. He objected to the conceit which placed multifarious other cultures on a par with modern scientific knowledge. The latter had manifestly demonstrated its superiority through the changes it had wrought in the world, a 'brute' fact that stood beyond any ethical justification. This insistence on rational science as the basis of a new mode of 'cognition' leads John Hall to claim that Gellner is 'more Weberian than Weber himself' (2010: p. 229). Indeed, in the fullest outline of his philosophy of history (Gellner, 1988b), the man who had castigated idealist theories in philosophy and in the explanation of nationalism comes close to arguing that our modern world only became possible when we somehow began to *think* more rationally, in the spirit of a 'generic Protestantism.'

See also: Anthropology: Overview; Determinism: Social and Economic; Durkheim, Emile (1858–1917); Historical Explanation, Theories of: Philosophical Aspects; Kant,

Immanuel (1724–1804); Malinowski, Bronislaw (1884–1942); Nationalism, Historical Aspects of: Arab World; Nationalism, Historical Aspects of: The West; Nationalism: General; Popper, Karl Raimund (1902–94); Weber, Max (1864–1920); Wittgenstein, Ludwig (1889–1951); von Hayek, Friedrich A. (1899–1992).

Bibliography

Anderson, B., 1983. Imagined Communities. Reflections on the Spread of Nationalism. Verso, London.

Gellner, E., 1959. Words and Things. A Critical Account of Linguistic Philosophy and a Study in Ideology. Gollancz, London.

Gellner, E., 1964. Thought and Change. Weidenfeld and Nicolson, London.

Gellner, E., 1969. Saints of the Atlas. Weidenfeld and Nicolson, London.

Gellner, E., 1975. Legitimation of Belief. Cambridge University Press, Cambridge, UK.

Gellner, E., 1981. Muslim Society. Cambridge University Press, Cambridge, UK.

Gellner, E., 1983. Nations and Nationalism. Blackwell, Oxford, UK.

Gellner, E., 1985. Relativism and the Social Sciences. Cambridge University Press, Cambridge, UK.

Gellner, E., 1988a. State and Society in Soviet Thought. Blackwell, Oxford, UK.

Gellner, E., 1988b. Plough, Sword and Book. Collins, London.

Gellner, E., 1994. Conditions of Liberty. Civil Society and Its Rivals. Hamish Hamilton, London.

Gellner, E., 1996. Nationalism. Weidenfeld and Nicolson, London.

Gellner, E., 1998. Language and Solitude. Wittgenstein, Malinowski and the Habsburg Dilemma. Cambridge University Press, Cambridge.

Hall, J.A. (Ed.), 1998. The State of the Nation: Ernest Gellner and the Theory of Nationalism. Cambridge University Press, Cambridge, UK.

Hall, J.A., 2010. Ernest Gellner. An Intellectual Biography. Verso, London.

Hall, J.A., Jarvie, I. (Eds.), 1996. The Social Philosophy of Ernest Gellner. Rodopi, Amsterdam.

Huntington, S.P., 1996. The Clash of Civilizations and the Making of the New World Order. Simon and Schuster, New York.

Malešević, S., Haugaard, M. (Eds.), 2007. Ernest Gellner and Contemporary Social Thought. Cambridge University Press, Cambridge.

Gender and Academic Motivation

Charlotte A Agger and Judith L Meece, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Recent research on the motivation and achievement patterns of girls and boys has called attention to several significant trends. Notable developments, such as the favoring of girls on many reading-focused assessments and a plethora of new studies on boys, appear both in the United States and internationally. This article covers several prominent gender-related trends in motivation research on both national and international levels. Specifically, the article discusses (1) the increased focus on boys' motivation and achievement, (2) differences in motivation and achievement within gender, (3) cross-national trends in gender-related motivation and achievement patterns, and (4) the expanding of existing theoretical motivation frameworks.

The past decades have brought forth much research on trends related to the motivation and achievement patterns of girls and boys both in the United States and abroad. Most notable in the current literature is the fact that girls have, across academic domains, most countries studied, and cultures, caught up with boys in terms of their performance on school-based and national assessments. In addition, girls are more likely than boys to complete secondary school by either completing standard schooling or earning a high school/secondary school equivalency degree (Meece and Askew, 2012). Regarding postsecondary education, in most Western countries, more women than men graduate college with a bachelor's degree. However, women continue to be less involved in mathematics-intensive fields during and after college in all countries studied (NSF, 2013; Watt, 2008) despite boys' and girls' comparable performance on mathematics assessments (Lindberg et al., 2010). Relatedly, in the USA, women's participation in the social and biosciences has dramatically increased since 1991, but men are still more likely to enter the physical sciences (NSF, 2013). These trends, such as the favoring of girls on a variety of large-scale reading assessments and the largely equivalent performance of girls and boys in the domains of science and mathematics, are apparent across geographic boundaries.

In recent decades, educational researchers have begun to focus on the underlying social and psychological processes and mechanisms driving the remaining gaps in gender-differentiated achievement patterns. Much of this new literature in Western countries centers on an academic choice model, rather than a personality theories of motivation. As such, current work looks at how students make sense of their academic achievement experiences and come to value particular academic and career domains. More recently, the field has shifted again. This time, from utilizing theories strongly rooted in motivation beliefs to theories that include broader socio-cultural influences and their combined effect on motivation (e.g., how teacher characteristics interact with student ethnicity to influence motivation).

Gender-related motivational processes are key to understanding the educational trajectories of boys and girls in countries across the world. The expectancy-value framework enables researchers to model how social-cognitive processes (e.g., within classroom and familial contexts) shape student motivation. Motivation and achievement trends not only inform our understanding of students' beliefs, values, choices, and subsequent educational outcomes, they are also telling of the gender-related educational progress made in the past decades and the needed changes in educational programing moving forward (Watt et al., 2012). This article discusses several gender-related motivation and achievement trends occurring on both national and international levels. The article covers five key topics (1) the increased focus on boys' motivation and achievement, (2) differences in motivation and achievement within gender, (3) cross-national trends in gender-related motivation and achievement patterns, and (4) the expanding of existing theoretical motivation frameworks.

Increased Focus on Male Students in US Studies

For many years a strong concentration on girls' achievement and advancement dominated much of the academic literature on gender and motivation. Research on US students in the 1980s and 1990s focused on gender differences in mathematics and science achievement, particularly in regard to girls' motivation. However, in recent years, an abundance of research on male students has emerged. This surge of research has found several notable trends. For example, in comparison to girls, boys receive lower grades (Cornwell et al., 2013) and take fewer advanced classes (Taylor and Lorimer, 2002). Girls, on the other hand, in the US school system and abroad, have made substantial gains in academic performance and are now on par, if not outperforming, male students in a variety of subjects. Results from the National Assessment of Educational Progress since 2008 and through 2011 show tests of reading and writing favor girls (NCES, 2008; NCES, 2012).

Along with assessment and grade-related trends favoring girls in US studies, research also documents major discrepancies in secondary and postsecondary graduation rates and postsecondary academic pursuits of male and female students. For example, boys currently graduate high school and college at lower rates and do not demonstrate clear aspirations for higher education compared to their female counterparts (Meece and Askew, 2012). There are also clear discrepancies in regard to fields of study during postsecondary education and students'

subsequent occupational choices. Whereas the gap between men and women who choose to pursue the life sciences has significantly narrowed, women are still less likely to choose mathematics-intensive fields of science such as engineering, physics, and many technology-related areas (Ceci and Williams, 2010).

Despite the presence of these pronounced trends, scant research is available to explain modern gender-differentiated patterns in academic motivation, achievement, and post-secondary attainment of US students. Research in the US in the 1980s and 1990s, using an expectancy-value motivation framework, documented clear gender differences in youth's competency and value beliefs, which were shown to affect the achievement striving of girls and boys (Eccles et al., 1983; Wigfield and Eccles, 2000). This framework was utilized by Jacobs et al. (2002) in conducting a longitudinal study of 761 students from 10 Midwestern schools. This study documented significant declines in both girls' and boys' competency beliefs, with declines varying according to domain and gender. Specifically, girls' perceptions of their ability in mathematics were found to decline at a slower rate than boys. When motivational beliefs related to language arts were examined, boys' and girls' perceptions of ability were similar at the beginning of elementary school, but boys' ability beliefs declined more rapidly than girls beginning in the late elementary years. Jacobs et al. (2002) also documented declines in value-related beliefs across the school years for both mathematics and language arts, with boys demonstrating a more rapid decline in language arts than girls. Although patterns of declines in competency and value beliefs may explain gender differences in academic achievement, these linkages were not specifically examined.

In looking at US-based literature on student motivation over the last decades, there are poignant differences in the motivational beliefs and trajectories of boys and girls. Many recent research efforts have focused on the lowered academic motivation of boys in literacy-focused subjects. Longitudinal, national assessments have provided a rich resource for documenting trends in educational achievement and attainment. However, to date, national assessments of educational progress have generally not included measures to examine motivational processes underlying and explaining these gender-differentiated patterns.

Variations within Gender Groups in Studies in the USA

Throughout schooling, some children deal with harmful discriminatory experiences and intense social pressures, which can greatly affect their motivation. These achievement-related inequities are becoming a significant component of US studies of educational achievement and attainment. In fact, as girls and young women in the United States were making important achievement strides in the 1990s, researchers cautioned the need to disaggregate data by ethnicity, family income, geographic location, and other key demographic characteristics, as these variables can significantly influence the educational experiences of different children. However, although research has consistently found that gender differences in students' schooling experiences are moderated by race and ethnicity (Meece and Askew, 2012), researchers have just recently begun to investigate how gender differences in motivation vary by ethnicity, race, or socioeconomic status (Eccles et al., 2006; Meece and Askew, 2012). Studies that have investigated variations within gender have looked at how race, ethnicity, and contextual variables affect motivation.

Extant research has found that gender-related achievement patterns vary across ethnic groups and can oscillate depending on academic domain and student ethnicity. For instance, African-American girls' and boys' scores on mathematics assessments are relatively equal (Lindberg et al., 2010; Vanneman et al., 2009). Yet on assessments of reading and writing, girls score higher across ethnic and cultural groups (Meece and Askew, 2012). In US studies, the group that has shown the greatest increases in their academic achievement is African-American girls. Over the last 30 years, African-American girls have made significant gains in reading and mathematics compared to their male counterparts. Academic gains for white boys and girls, as well as African-American boys, are not as great (Vanneman et al., 2009).

These gender-related achievement patterns found among students belonging to different ethnic groups have led to increased research on the motivational processes underlying these achievement trends. Many studies focused on African-American students have shown gender differences favoring girls in terms of academic expectations, aspirations, and values. Graham et al. (1998) were among the first achievement motivation researchers to look at gender and ethnicity differences in academic valuing in relation to barriers to academic success. In their original work and in subsequent supporting work, using peer nomination procedures, they have found that ethnic minority girls in elementary and middle school were more likely to report valuing average and high academic achievers than their low-achieving peers. In contrast, although the ethnic minority boys preferentially nominated high achievers during the elementary school years, they increased their preference for low-achieving classmates as they progress through school (Graham et al., 1998; Taylor and Graham, 2007). Consistent with this pattern, Wood et al. (2007) have shown that ethnic minority girls report higher expectations for future educational attainment than their male peers.

Recent research on within-group motivation patterns has begun to examine racial socialization influences. For instance, for boys, racial centrality (i.e., the extent to which one's racial background is important to one's self-concept) is associated with lower risk for negative school-related attitudes and poor grades from experiencing classroom discrimination (Chavous et al., 2008). For girls, higher centrality is associated with protection from peer discrimination toward school importance and academic self-concept (Chavous et al., 2008). Other research has highlighted the importance of racial pride in the study of motivation. Butler-Barnes et al. (2012) found a positive relationship between educational utility beliefs (i.e., the perception that education bears utility for future success) and grades among boys with high racial pride.

In summary, current within-group research on ethnicity and motivation has illustrated several key findings; (1) African-American boys are at particular risk for lack of motivation and school failure, (2) positive racial identity and school utility beliefs are critical to the success of both girls and boys and can have buffering effects against discrimination, and (3) racial discrimination can hinder a student's school-related beliefs.

Cross-National Trends

Research on gender-related motivation patterns has generally stemmed from theory and empirical research generated in the United States. There is less information on the role of student motivation in schooling for other countries. Yet recently, research focused on gender-related patterns of student motivation in different parts of the world has surfaced. In comparing the academic attitudes, beliefs, and achievement of students cross-nationally, many researchers use the Trends in International Mathematics and Science Study (TIMSS) and the Program for International Student Assessment (PISA). TIMSS focuses on assessing learned items from curricular resources and PISA examines mathematics literacy, or the ability for students to apply and engage with mathematics. Several notable trends have emerged from this new focus on cross-national trends.

Trends found in the US regarding gender, achievement, and motivation are present around the globe. This article covers several of these trends, broken into the following four main themes: (1) the advantage of girls on literacy-focused assessments, grades, and college graduation rates, (2) gender-related differences in motivational processes, (3) differences in gender-related motivation and achievement patterns between developing and industrialized countries, and (4) the effects of gender-related stereotyping on achievement.

First, girls do better than boys on literacy-focused assessments, school marks, and rates of graduating from college in many countries around. Indeed, studies using the Organization for Economic Co-operation and Development's Program for International Student Assessment (OECD-PISA) have found that girls receive better grades, graduate high school at larger rates, and garner higher reading achievement scores in most countries (OECD, 2011). Even in countries lauded for their rigorous and exemplary school systems, such as Singapore, there are gender differences in scores on reading and writing assessments. For example, scores on the 2006 Progress in International Reading Literacy Study favored Singaporean girls (Mullis et al., 2007). Using PISA data with a sample of 193 841 fifteen-year-olds, Chiu and Chow (2010) found that girls outscored boys in 95% of the 41 countries included in the sample. This pattern was especially pronounced in Denmark, Norway, Portugal, Sweden, and Thailand (Chiu and Chow, 2010). In other domains, such as mathematics, boys are generally favored, although gender gaps are quite small (less than 5%) in countries such as Finland, Slovenia, Sweden, Indonesia, and Russia. There are no significant differences in performance in science (OECD, 2011).

Current gender-related cross-cultural trends in achievement reflect patterns in student motivation (e.g., beliefs and attitudes). In fact, PISA data at the aggregate level support the conclusion that gender differences in educational choices are more related to students' attitudes and interest in a particular subject than students' actual ability (OECD, 2011). This has led researchers to adopt the belief that many of the trends in secondary and postsecondary education are partially, if not largely, attributable to gendered beliefs and stereotyping, rather than ability (OECD, 2011).

Indeed, research stemming from around the world shows marked differences in boys' and girls' motivation and school engagement. For instance, in Australia, using the Student Motivation Scale, Martin (2004) found that girls were significantly higher than boys in planning, study management, learning focus, and persistence, whereas boys were higher than girls in their use of self-handicapping strategies. Similarly, Lam et al. (2012) examined gender, student engagement, and academic performance among students in Austria, Canada, China, Cyprus, Estonia, Greece, Malta, Portugal, Romania, South Korea, the United Kingdom, and the United States and found that girls maintained higher levels of school engagement and that student engagement explained the gender-related variance in academic performance (Lam et al., 2012). Similarly, research on achievement expectations in several different countries has found that boys rate their math abilities higher than girls. In fact, using TIMSS and PISA data from 69 countries, Else-Quest et al. (2010) found that boys report more positive mathematics-related attitudes and affect girls despite the overall similarities in mathematics achievement between boys and girls.

When gender-related trends in achievement and motivation are investigated in developing countries in contrast to industrialized countries, different patterns arise. In developing countries, girls have poorer educational attainments, particularly as students reach higher levels of schooling (OECD, 2011). Yet, although girls in developing countries have lower educational attainment and are less likely to be enrolled in school compared to boys, girls who have attended school at one point in time make equal or better schooling progress than boys throughout many regions and across age groups (Grant and Behrman, 2010). In addition, current research in developing countries is rife with examples of rampant gender biases in certain societies. Lower rates of social promotion, limited access to educational opportunities, and unequal treatment in society are issues that many girls and women face and are prime areas where continued support and research are needed.

Expanding of Theoretical Frameworks

It is quite apparent from the literature on children and schooling that students do not develop in a vacuum; they exist in a multilayered, complex social world. As such, students and their academic motivation should be studied with these complex social and environmental processes and contexts (e.g., peer processes, economic and political climates, cultural practices) in mind. Fortunately, the past decade has seen a rise in research that includes contextual variables as researchers are beginning to acknowledge the importance of these influences on student motivation. In fact, Perry et al. (2006) assert, "contextual understandings are more integral to research on motivation today, reflecting the general shift in educational research toward situated and social perspectives on learning" (p. 328). The Eccles et al. (1983) developmental model of achievement motivation is one framework that has continuously extended to study combinations of contextual variables (e.g., the effect of both teacher and ethnic influences) affecting motivation.

Since its introduction, the model proposed by Eccles et al. (1983) has examined the role of children's socialization experiences in creating and reinforcing gender-related differences in motivation and achievement. The strong socialization

component of this model examines processes by which socializers (particularly parents and teachers) (1) model sex-typed behaviors, (2) impart different expectations and goals for boys and girls, and (3) support boys and girls when they undertake different activities and skills. This model has continued to yield a rich body of research on children's socialization experiences at home and school (Wigfield et al., 2006).

Indeed, many researchers have applied the Eccles et al. (1983) model to studying how a child's home environment (e.g., interactions with his or her parents and siblings) affects their academic motivation. These familial socialization patterns and relationships are especially key in understanding gendered patterns of academic motivation, as early gendered beliefs are most often formed through interactions with parents. Research has found that gendered beliefs formed in early childhood influence long-term thinking about competency, school valuing, and the self (Meece et al., 2006). Parental beliefs about school valuing and educational aspirations and goals, and even different academic domains, are communicated to children through choice of leisure activities and everyday conversation (e.g., parents talking about their own skills and attitudes) (Eccles et al., 1983). In fact, research shows that parental beliefs about their children's gender and abilities are strong predictors of their children's own beliefs about ability and performance (Bleeker and Jacobs, 2004). Jacobs (1991) found that parents holding stereotypical beliefs regarding males' supposed expert mathematical abilities exhibited less confidence in their child's mathematics ability if they had daughters compared to sons.

In addition to factors in the home environment, school-related influences on an individuals' achievement motivation are critical to study, yet extremely complex. Students come to school with preconceived ideas and attitudes about gender-typed behaviors, and subsequently school often serves as a place where students fine-tune and verify their own beliefs. At school, students see older male and female teachers, staff, and administrators carry out specific roles. For example, women are more likely to take care of young children while men are more likely to appear in management roles within the school (Meece and Askew, 2012). The learning materials, such as textbooks, videos, and computer software, utilized in schools present images of gender-typed behaviors and actions. It is also important to consider the fact that schools significantly differ in terms of geographic location, access to resources, teacher quality, and socioeconomic statuses of families. These factors greatly influence the bidirectional relationship that the child and school experience, which result in varying academic motivational trajectories for individual students.

Recent studies have used the Eccles et al. (1983) framework to examine the complex interplay of African-American youth and their schooling experiences. In one study, Eccles et al. (2006) reported that daily racial discrimination during seventh and eighth grades predicted declines in grades, academic ability self-concepts, and academic task values. However, a strong connection to one's ethnic group acted as a buffer in assuaging some of the negative effects of racial discrimination. In another study, Graham et al. (1998) examined gender differences in the ways African-American youth respond to racial discrimination. Specifically African-American boys, when compared to girls, demonstrated a higher likelihood of devaluing academic success under conditions of racial discrimination. This strategy, they found, enabled boys to maintain positive self-perceptions of ability. African-American girls, on the other hand, placed more emphasis on relationships and approval from peers and teachers, when self-concepts are threatened by racial discrimination. Differences in the ways African-American boys and girls respond to racial discrimination in school setting may explain variations in achievement for this subgroup of US students.

A related line of research is examining the role of cultural stereotypes in explaining gender-related achievement patterns across countries. That is, variations in gender-related cultural attitudes and the valuing of education can trickle down to the student level and greatly influence student motivation. Recent research on the effect of how cultural stereotypes affect student motivation and achievement provides several interesting insights. In particular, using samples of students from around the world (Chiu and Chow, 2010; Nosek et al., 2009), numerous studies document the relation of cultural stereotypes to students' educational behaviors, choices, and performance. For example, Nosek et al. (2009) used a gender-science Implicit Association Test (i.e., a measure of associations between mental representations) with over half a million participants belonging to 34 countries. The researchers found a significant relationship between gender-science stereotyping and sex differences in mathematics and science achievement (Nosek et al., 2009). More specifically, the level of implicit stereotyping in a country was related to the level of inequality in science and mathematics achievement in that country. These findings surrounding implicit stereotypes and achievement patterns may be extended to show disparities in larger participation trends in science and mathematics (Nosek et al., 2009). Related to implicit beliefs and stereotypes, other studies have highlighted the influence of cultural stereotypes on achievement in particular domains (Chiu and Chow, 2010; Fredericks and Eccles, 2002). Chiu and Chow (2010) found that cultural values were linked to reading motivation and achievement. Girls who lived in countries that valued strict gender roles learned less than in other countries.

In summary, recent studies using both social-cognitive and sociocultural frameworks continue to broaden the knowledge base on gender and motivation. These studies highlight the critical role of socialization processes and mechanisms in the home, family, and cultural contexts. There is an abundance of research on these socialization patterns and current research findings are promising for understanding gender differences across and within cultures. Most of this research has been done in the USA. More work is needed in all other countries.

Conclusion

This article examined research focused on several national and cross-national trends regarding gender differences in motivation and achievement. The article focused on four major areas including (1) the increased focus on boys' motivation and achievement, (2) differences in motivation and achievement within gender, (3) cross-national trends in gender-related motivation and achievement patterns, and (4) the expanding of existing theoretical motivation frameworks.

It is clear from international data on student achievement that girls and boys perform equally well in mathematics and science and girls largely outperform boys on reading assessments. However, despite the relative parity of girls and boys in terms of performance, their motivational processes look quite different and may fuel persistent inequalities and gender gaps in participation in various domains. For example, the USA science and engineering workforce is still largely composed of white, male workers (NSF, 2013).

Not only is this kind of gender imbalance in mathematics (and literacy-related) participation present in the USA, it also appears in international contexts (Watt, 2008). Given the finding that motivational beliefs are predictive of certain forms of successful schooling (e.g., high school mathematics participation), it is critical to continue studying domain-specific motivational processes and mechanisms across a wide range of groups to help understand and address continued disparities (OECD, 2011; Watt, 2008; Watt et al., 2012). To help address these critical areas of study, motivation frameworks have recently broadened to examine motivation *in situ* moving from studying student perceptions to investigating interactions and activities (Perry et al., 2006). These new studies largely emanate from social-cognitive and sociocultural perspectives of motivation and marry such variables as family, teacher, and student characteristics, learning and engagement, and instructional tasks and academic domains. In order to promote greater equity in schools and in the workplace, continued research and support targeted at understanding the aforementioned gender-related motivation and achievement patterns is essential.

See also: Motivation, Learning, and Instruction; Race and Academic Motivation; School Achievement: Motivational Determinants and Processes; Schooling: Impact on Cognitive and Motivational Development.

Bibliography

Bleeker, M.M., Jacobs, J.E., 2004. Achievement in math and science: do mothers' beliefs matter 12 years later? Journal of Educational Psychology 96 (1), 97–109.

Butler-Barnes, S.T., Williams, T.T., Chavous, T.M., 2012. Racial pride and religiosity among African American boys: implications for academic motivation and achievement. Journal of Youth and Adolescence 41, 486–498.

Ceci, S.J., Williams, W.M., 2010. Sex differences in math-intensive fields. Current Directions in Psychological Science 19 (5), 275–279.

Chavous, T.M., Rivas-Drake, D., Smalls, C., Griffin, T., Cogburn, C., 2008. Gender matters, too: the influences of school racial discrimination and racial identity on academic engagement outcomes among African American adolescents. Developmental Psychology 44, 637–654.

Chiu, M.M., Chow, B.W.Y., 2010. Culture, motivation, and reading achievement: high school students in 41 countries. Learning and Individual Differences 20, 579–592.

Cornwell, C., Mustard, D.B., Van Parys, J., 2013. Noncognitive skills and the gender disparities in test scores and teacher assessments: evidence from primary school. Journal of Human Resources 48 (1), 236–264.

Eccles, J.S., Adler, T.F., Futterman, R., Goff, S.B., Kaczala, C.M., Meece, J.L., 1983. Expectancies, values and academic behaviors. In: Spence, J.T. (Ed.), Achievement and Achievement Motives. Freeman, San Francisco, pp. 75–146.

Eccles, J.S., Wong, C.A., Peck, S.C., 2006. Ethnicity as a social context for the development of African American adolescents. Journal of School Psychology 44, 407–426.

Else-Quest, N.M., Hyde, J.S., Linn, M.C., 2010. Cross-national patterns of gender differences in mathematics: a meta-analysis. Psychological Bulletin 136, 103–127.

Fredericks, J.A., Eccles, J.S., 2002. Children's competence and value beliefs from childhood through adolescence: growth trajectories in two male-sex typed domains. Developmental Psychology 38, 519–533.

Graham, S., Taylor, A., Hudley, C., 1998. Exploring achievement values among ethnic minority early adolescents. Journal of Educational Psychology 90, 606–620.

Grant, M.J., Behrman, R.J., 2010. Gender gaps in educational attainment in less developed countries. Population and Development Review 36 (1), 71–89.

Jacobs, J.E., 1991. Influence of gender stereotypes on parent and child mathematics attitudes. Journal of Educational Psychology 83 (4), 518–527.

Jacobs, J., Lanza, S., Osgood, D.W., Eccles, J., Wigfield, A., 2002. Changes in children's self-competence and values: gender and domain differences across grades one through twelve. Child Development 73, 509–527.

Lam, S.F., et al., 2012. Do girls and boys perceive themselves as equally engaged in school? The results of an international study from 12 countries. Journal of School Psychology 50 (1), 77–94.

Lindberg, S.M., Hyde, J.S., Petersen, J.L., Linn, M.C., 2010. New trends in gender and mathematics performance: a meta-analysis. Psychological Bulletin 136, 1123–1135.

Martin, A.J., 2004. School motivation of boys and girls: differences of degree, differences of kind, or both? Australian Journal of Psychology 56, 133–146.

Meece, J.L., Askew, K.J.S., 2012. Gender, motivation, and educational attainment. In: Harris, K.R., Graham, S., Urdan, T., Royer, J.M., Zeidner, M. (Eds.), APA Educational Psychology Handbook, Individual Differences and Cultural and Contextual Factors, vol. 2. American Psychological Association, Washington, DC, pp. 139–162.

Meece, J.L., Glienke, B.B., Burg, S., 2006. Gender and motivation. Journal of School Psychology 44 (5), 351–373.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., Foy, P., 2007. PIRLS 2006 International Report. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, Boston, MA.

National Center for Education Statistics, 2008. NAEP Data Explorer. Retrieved from: http://nces.ed.gov/nationsreportcard/naepdata/.

National Center for Education Statistics, 2012. The Nation's Report Card: Writing 2011 (NCES 2012-470). US Department of Education, Washington, DC.

National Science Foundation, 2013. Women, Minorities, and Persons with Disabilities in Science and Engineering: 2013. Retrieved from: http://www.nsf.gov/statistics/wmpd/2013/digest/theme2_1.cfm.

Nosek, B., et al., 2009. National differences in gender science stereotypes predict national sex differences in science and math achievement. Proceedings of the National Academy of Sciences of the United States of America 106, 10593–10597.

OECD, 2011. Report on Gender Initiative: Gender Equality in Education, Employment, and Entrepreneurship. Retrieved from: http://www.oecd.org/education/48111145.pdf.

Perry, N.E., Turner, J.C., Meyer, D.K., 2006. Classroom contexts for motivating learners. In: Alexander, P., Winnie, P. (Eds.), Handbook of Educational Psychology, second ed. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 327–348.

Taylor, A.Z., Graham, S., 2007. An examination of the relationship between achievement values and perceptions of barriers among low-SES African American and Latino students. Journal of Educational Psychology 99, 52–64.

Taylor, D., Lorimer, M., 2002. Helping boys succeed. Educational Leadership 60 (4), 68–71.

Vanneman, A., Hamilton, L., Baldwin Anderson, J., Rashman, T., 2009. Achievement Gaps: How Black and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress (NCES 2009 455). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Watt, H.M.G., 2008. What motivates females and males to pursue sex-stereotyped careers? In: Watt, H.M.G., Eccles, J.S. (Eds.), Gender and Occupational Outcomes: Longitudinal Assessments of Individual, Social and Cultural Influences. APA Books, Washington, DC, pp. 87–113.

Watt, H.M., Shapka, J.D., Morris, Z.A., Durik, A.M., Keating, D.P., Eccles, J.S., 2012. Gendered motivational processes affecting high school mathematics participation, educational aspirations, and career plans: a comparison of samples from Australia, Canada, and the United States. Developmental Psychology. Advance online publication. http://dx.doi.org/10.1037/a0027838.

Wigfield, A., Eccles, J.S., 2000. Expectancy–value theory of achievement motivation. Contemporary Educational Psychology 25 (1), 68–81.

Wigfield, A., Eccles, J.S., Schiefele, U., Roeser, R.W., Davis-Kean, P., 2006. Development of achievement motivation. In: Eisenberg, N. (Ed.), Handbook of Child Psychology, vol. 3. Wiley, New York, pp. 933–1002.

Wood, D., Kaplan, R., McLloyd, V.C., 2007. Gender differences in the educational expectations of urban, low-income African American youth: the role of parents and the school. Journal of Youth and Adolescence 36, 417–427.

Gender and Education

Erin Pahlke, Whitman College, Walla Walla, WA, USA
Priscilla Goble, School of Social and Family Dynamics, Arizona State University, Tempe, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Recent international emphasis has been placed on closing persistent gender gaps in education. In this article we explore current gender gaps in educational outcomes, focusing on students' experiences in developed countries. We then investigate potential causes of the gender gaps, focusing on differences in cognitive and socioemotional skills, socialization, gender stereotypes, and interest. Next, we discuss societal responses to the gender gaps, including movements focused on girls being shortchanged, boys being in crisis, gender differentiated and segregated learning approaches, and attempts to ameliorate gender gaps by reducing the focus on gender. Finally, we end with suggestions for future work.

Introduction

In 2000, the United Nations Educational, Scientific and Cultural Organization (UNESCO), as part of the Dakar Framework for Action, made the goal of "eliminating gender disparities in primary and secondary education by 2005, and achieving gender equality in education by 2015, with a focus on ensuring girls' full and equal access to and achievement in basic education of good quality" (UNESCO, 2012: p. 8). By many measures, significant progress has been made toward these goals. However, even in developed countries, discrepancies persist in terms of test scores and grades in some domains, attitudes and behavior in classrooms, and aspirations for educational and occupational attainment. In this article, we explore these gender differences in educational outcomes. We then investigate some of the proposed causes for these gender gaps and investigate societal responses to the persisting gaps. The article then ends with ideas about future research. Given our experiences and expertise, we focus on data related to children's, adolescents,' and college students' experiences in developed countries. For a more complete discussion of the current state of gender differences in educational outcomes in developing countries, see UNESCO (2012).

Gender Differences in Educational Outcomes

The following sections present recent data on the state of gender differences within education, focusing on test scores and grades, engagement and behaviors in classrooms, and educational attainment.

Test Scores and Grades

Historically, there have been persistent gender gaps in academic achievement. Males have generally outperformed females in mathematics, while females have generally outperformed males in literacy. Current statistics suggest that some of these patterns are slowly changing.

On many mathematics examinations, males slightly outperform females (e.g., Lee et al., 2007; Planty et al., 2009). However, this gap is not consistent across ages and contexts. The gender gap in mathematics performance often does not appear until early adolescence and only in some exams (Ceci et al., 2009). Examining scores on state-standardized exams of children in second- through eleventh-grade US public schools, Hyde et al. (2008) reported no mean level differences between boys' and girls' math performances. Looking at the mathematics exam scores by state and grade level, boys outperformed girls in 21 cases, girls outperformed boys in 36 cases, and there was absolutely no gender difference in nine cases. Looking across countries, in the 2009 Programme for International Student Assessment (PISA), a standardized test that 15-year-old students around the world completed, boys outperformed girls on the mathematics exam by an average of 12 points (or 0.12 of a standard deviation; OECD, 2010). There were, however, variations across countries and economies; while in 35 countries boys outperformed girls, in 25 countries there were no significant gender differences, and in five countries girls outperformed boys.

In comparison to the shrinking mathematics achievement gap, the gender gap favoring females on assessments of literacy are often present from the moment children enter formal schooling and persist across the life course (e.g., Perie et al., 2005; Rathbun and West, 2004; Robinson and Lubienski, 2011). For example, research examining literacy skills in a nationally representative sample of kindergartners in the US has shown that girls not only enter kindergarten with greater literacy skills than boys, but they also learn slightly more across the kindergarten year (Ready et al., 2005). In the 2009 PISA reading literacy exam, in every one of the 60 participating countries and economies, 15-year-old females significantly outperformed their male peers (OECD, 2010). Furthermore, the size of the literacy gap was over three times larger than the math gap favoring males.

Considering children's academic performance in the classroom as opposed to performance on standardized exams, girls have consistently outperformed boys throughout much of the developed world since at least the mid-1900s (Mickelson, 1989). Girls tend to earn higher grades than boys from the time they enter school through the end of college – even in traditionally-masculine subjects like science and math – and these differences are seen across cultures (Perkins et al., 2004). Italian female college students earned higher grades than the males in 11 of the 15 subjects (with males outperforming

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92151-2

females in only one area and no significant difference in three other areas; Castagnetti and Rosti, 2010).

Engagement and Behavior in Classrooms

From the moment they enter school, gaps appear between boys' and girls' behavior in the classroom. Research conducted in the United States suggests that in kindergarten, twice as many boys than girls struggle to pay attention (Zill and West, 2001), and throughout childhood and adolescence, boys consistently engage in more disruptive classroom behaviors than girls (Schaefer, 2004). Further, girls are often rated to be more dutiful and cooperative in classes than boys (Mickelson, 1989). Given these gender differences, it is not surprising that starting in prekindergarten and continuing throughout their educational careers, boys are about four times more likely than girls to be referred to school administrators, suspended, or expelled (Gilliam, 2005).

In addition to behaving more poorly in the classroom, boys also tend to have more negative attitudes about school and teachers than girls throughout their educational careers. In a recent study of US students of 8–18 years of age, Sullivan et al. (2008) found small but consistent differences suggesting boys have stronger negative attitudes about school than girls. In a similar vein, as early as kindergarten, gender differences appear in attitudes toward teachers; for example, boys are more likely than girls to think that their teachers do not like or care about them (Valeski and Stipek, 2001).

Unfortunately, boys' tendency to behave worse in classrooms and endorse more negative attitudes about school and teachers than girls may be related to their lower achievement levels. Inattentive and disruptive classroom behaviors predict poorer academic achievement (Finn et al., 1995). Similarly, negative attitudes toward school and teachers are associated with lower achievement and lower expectations of future success (Baker, 1999; Brier, 1995). Additionally, negative achievement leads students to have more negative attitudes about school, producing a cyclical effect (Valeski and Stipek, 2001). Slight gender differences in classroom behavior and attitudes toward school and teachers at the beginning of formal schooling may, over time, result in large gender gaps in achievement.

Educational Attainment

Historically, females have had less access to education than their male counterparts. Although gender parity persists in most countries in South and West Asia and sub-Saharan Africa, gender parity has been achieved in much of the rest of the world (UNESCO, 2012). In fact, in many developed countries, the persistent gender gaps favoring boys have flipped. In the US, for example, boys are more likely than girls to drop out of school (Bowlby, 2005). Girls, meanwhile, are now more likely than boys to complete high school (Horn et al., 2002), enroll in college (NSF, 2009), and graduate from college (Freeman, 2005). These trends stand in stark contrast to the data from just 50 years ago: in 1960 there were 1.6 males graduating from 4-year US colleges for each female; in 2003, in comparison, there were 1.35 females graduates for each male (Goldin et al., 2006). Other countries have experienced similar shifts – in East Asia and the Pacific, Latin America and the Caribbean, and North America and Western Europe, women went from being the minority of college students in 1970 to the majority of college students in 2009 (UNESCO, 2012). Notably, this flip in the gender gap in educational achievement in developed countries is not due to fewer men participating than in the past; instead, it is due to a dramatic increase in the percentage of women participating in postsecondary education.

Unfortunately, gender gaps in educational attainment are complicated by gender differences in attainment in different educational fields. In 2006, women earned 58% of the bachelor's degrees awarded in the US (NSF, 2009). However, there was considerable variation in the size and direction of the gender gap by field. Women earned 90% of the degrees in psychology and 62% of the degrees in biological sciences. Men, however, earned more of the degrees in mathematics and statistics (55%), physical sciences (58%), and engineering (80%; NSF, 2009).

Potential Explanations for Gender Gaps

Researchers have identified several factors that may play a role in the gender differences that emerge in educational achievement, attitudes, and attainment. In this section, we focus on several broad areas: cognitive and socioemotional skills, socialization, gender stereotypes, and gender-typed interests.

Cognitive and Socioemotional Skills

In the past, scholars argued that gender gaps in educational outcomes could primarily be attributed to differences in overall intelligence. This perspective is not supported by research: today, most experts agree that gender differences in overall intelligence do not exist (Halpern and LaMay, 2000). Indeed, the vast majority of research on gender differences in cognitive skills is consistent with the gender similarities hypothesis (Hyde, 2005): in most areas, males and females are similar.

There are, however, some cognitive differences that may help to explain persistent gender gaps in education. For example, results from a meta-analysis suggest that, on average, males outperform females on visual-spatial tasks starting as early as 3 years of age, suggesting that boys' advantage in this area is not due to differences in life experience alone (Cohen's $d = 0.37$; Robinson et al., 1996; Voyer et al., 1995). These slight differences in visual-spatial tasks may relate to performance on some math and science exams. Females, meanwhile, tend to outperform males in the area of memory, with effect sizes ranging from 0.20 to 0.60. These gender differences, which are consistent across developmental periods and cultural backgrounds, may help explain gender gaps in literacy achievement (see Halpern, 2000 for a review). For example, vocabulary size, which has a strong memory component, is typically larger among females than males. This difference appears as early as 16 months and exists even after controlling for mother–child interactions, suggesting a biological component in girls' advantage in the areas of reading and writing (Huttenlocher et al., 1991).

Gender gaps in self-regulation skills, which have cognitive and socioemotional underpinnings, may also contribute to

gender gaps in education, especially in areas related to classroom behavior and engagement. A recent meta-analysis of temperament in childhood and preadolescence reported a large difference between boys' and girls' effortful control (Cohen's $d = -1.01$; Else-Quest et al., 2006). The authors concluded that girls have a stronger ability than boys to regulate their attention and inhibit their impulses – skills that certainly serve girls well in the classroom (Blair and Razza, 2007). It may be that group level differences between boys' and girls' behavior in the classroom is driven in part by boys' relatively slower developing regulation skills. It should be noted, however, that boys and girls do not differ substantially in their ability to resist temptation or susceptibility to distraction (Else-Quest et al., 2006; Silverman, 2003). Boys may, on the group level, have a slightly harder time regulating their behavior in the classroom, but developmental differences cannot explain all of the differences between boys' and girls' classroom behavior or achievement.

Socialization

Children come to understand gender in part through their experiences with gender socialization, which encompass both explicit and implicit messages (Maccoby, 1998). Accordingly, it may be that gender gaps in achievement develop, in part, from differences in the education-related socialization of boys and girls. As researchers, we know surprisingly little about the explicit gender-differentiated messages children receive about academics and occupations. We do know, however, from a recent survey of US adolescents that 54% of girls had experienced explicit academic sexism (i.e., reports of discouraging or other negative statements from others about their abilities in math, science, or computers because they are female; Brown and Leaper, 2010). Yet, the extent to which these explicit gendered messages influence achievement and explain boys' advantage in the areas of math and science is not clear.

In addition to explicit, sexist messages, socializers often influence children's views of gender, education, and occupations through more circuitous or implicit routes (e.g., Gelman et al., 2004; Langlois and Downs, 1980). Consider the domain of children's play activities. A large-scale meta-analysis of parental treatment of boys and girls found that parents who encouraged children's gender-typed toys and activities and discouraged cross-gender-typed toys and activities had children with more gender-typical activity involvement (Lytton and Romney, 1991). With regard to academics, mothers have been shown to provide more explanations of scientific content in a museum setting when talking to boys than to girls (Crowley et al., 2001), perhaps because parents view their sons as more competent in math as opposed to their daughters who they view as better in language arts (Eccles et al., 1990). Thus, parents may shape children's views about academics through the encouragement they provide to children to pursue gender-typed academic areas.

Similarly, teachers may influence children through their use of reinforcements in the classroom. Prior research has documented that boys receive more teacher attention (albeit generally in response to negative behavior), are called on to answer more questions, are interrupted less frequently, and receive more 'thinking time' and positive comments from teachers than do girls (AAUW, 1992; Hendrick and Stange, 1991; She, 2000). Even if teachers never make explicit comments about which gender is better at math, their differential treatment of boys and girls may negatively affect girls' motivation and interest in math and science, especially if they receive messages from others that boys are better in these subject areas. In addition to their use of reinforcement, teachers may also impact children implicitly through modeling; 76% of US public school teachers are women (NCES, 2010), perhaps sending a strong message to children that school is a place for girls.

Finally, it is worth noting that socialization has different effects on different individuals; while many women have reported being deterred from the areas of math and science because of negative gender-typed messages they received at home or school, other prominent female scientists, mathematicians, and physicians have reported being emboldened by negative messages and gender discrimination (e.g., McGrayne, 2001).

Gender Stereotypes

The limited research available on children's and adolescents' academic gender stereotypes suggests that from a young age, individuals categorize math and science as 'boys' subjects' and language arts and spelling as 'girls' subjects' (e.g., Plante et al., 2009; Rowley et al., 2007). Researchers have just recently begun to explore how these gender stereotypes about subjects may influence students' achievement and explain the gender gap. However, prior work on gender stereotypes suggests that the effects are powerful (Bigler and Liben, 2006).

The influence of the endorsement of gender stereotypes on students' achievement has been explored in studies examining boys' endorsement of masculinity stereotypes. Many stereotypes about masculinity (e.g., aggression, physical toughness, and emotional stoicism) are in stark contrast to the compliance and responsiveness required of students in the classroom. Consistent with the hypothesis that high endorsement of these stereotypically masculine ideals is negatively related to achievement, US middle-school boys with average levels of achievement tend to endorse traditional masculine stereotypes at a higher rate than boys classified as 'gifted' (Shepard et al., 2011). Similarly, extremely masculine gender-typed US boys are less engaged in school and have lower achievement than their less extreme male peers (Ueno and McWilliams, 2010). Endorsing stereotypes about masculinity – and behaving in those stereotypically-masculine ways – appears to be related to lower academic performance.

Importantly, mere *awareness* of gender stereotypes can influence youths' academic performance. Stereotype threat refers to the fear of confirming a negative stereotype and tends to arise when students who are confronted with a difficult task are aware of negative stereotypes about their group in the task's domain (Steele, 1997). In cases in which individuals are aware of a culturally-held negative stereotype about their group, they often become anxious that they will conform to the negative stereotype, and that anxiety leads to poor performance. So, for example, when college students are told that the math test they are taking has shown gender differences favoring men in the past, women perform worse than men. In comparison, when

students are told that the test has never shown gender differences in the past, women perform as well as the men (Spencer et al., 1999). Importantly, over time stereotype threat also has the potential to undermine students' identification with the relevant academic domain (e.g., Steele, 1997), which may explain some of adolescent girls' dampened interest in science, technology, engineering, and math (STEM) fields and boys' dampened interest in traditionally-feminine fields.

Interests

Finally, it is important to consider the role gender differences in interests may play in explaining gender gaps in educational outcomes. Children's preferences for gender-typed activities (e.g., balls vs dolls) emerge during the first year of life and continue into late childhood (McHale et al., 2004). Starting in elementary school, gender differences also appear in boys' and girls' interests in different academic subjects; for example, elementary school-aged boys have higher levels of interest in math than do girls – and the difference holds even when controlling for prior math achievement (Simpkins et al., 2006). At the same time, boys are more likely than girls to be interested in jobs stereotyped as masculine by their culture, such as jobs in STEM fields (Bigler and Liben, 2006), and the girls who are interested in male-dominated career fields in early adolescence often move away from these preferences – and toward female-dominated career fields – by adulthood (Frome et al., 2006). More work is needed to understand how these gender differences in interests develop (see Goble et al., 2012). However, as long as the gender differences in interest persist, it is likely that gaps will remain in boys' and girls' achievement and persistence in various academic areas.

Societal Responses to Gender Gaps

Over the last several decades, researchers, educators, and policy makers have responded to gender gaps in educational outcomes in a variety of ways. Many of these responses have in effect pitted girls against boys, rather than focusing on improving the academic achievement of all students. In this section, we discuss the movements focused on girls being shortchanged, boys being in crisis, gender differentiated and segregated learning approaches, and attempts to ameliorate gender gaps by reducing the focus on gender.

Identifying Gender Bias Approaches

Over the last several decades, many scholars have proposed that issues related to gender and education be addressed by focusing on the ways in which school systems are biased against one gender or the other. In 1992, the American Association of University Women (AAUW) spearheaded an effort to focus on the challenges girls face in schools. In *How Schools Shortchange Girls*, the AAUW drew on research to argue that schools needed to do more to confront the supposed sexism girls face in school. Perhaps as a result of these scholars' efforts, over the last two decades, researchers have focused on better understanding girls' experiences with sexism and discrimination in the classroom (e.g., Brown and Leaper, 2010) and developing interventions that encourage girls to enter STEM fields (e.g., Little and Leon de la Barra, 2009).

Some have labeled the argument that schools shortchange girls as 'false political propaganda,' citing gender gaps that actually favor females (such as grades in school and literacy scores; Kleinfeld, 1998). Indeed, in a movement developed in the late 1990s to focus on the challenges boys face in schools, scholars argued schools had gone too far in support of girls and were now hostile toward boys. The 'Boy Crisis' was born with books such as *The War Against Boys* (Sommers, 2000) and *Real Boys* (Pollack, 1998). Scholars began to argue that boys are falling behind girls academically in large part because they have begun to see school as a context more suitable for girls.

School systems have responded to these concerns. In the United Kingdom, the government supported the development of the Raising Boys' Achievement Project (see Younger and Warrington, 2006). In Australia, meanwhile, the government commissioned the report *Boys: Getting It Right* and is now funding Boys' Education Lighthouse grants to address boys' educational experiences (see Weaver-Hightower, 2009). Opponents of the boy crisis, in comparison, question whether the boy crisis movement is driven merely by concerns about boys' relative performance compared to girls – and boys' loss of status and power (e.g., Rivers and Barnett, 2006). These scholars point out that boys' educational achievement is generally improving overall; girls' performance is just improving at a faster rate. For now, camps focused on the challenges girls vs boys face in schools remain divided.

Gender-Differentiated Approaches

In recent years, there has been a resurgence in interest in gender-differentiated instructional approaches as a means to addressing gender gaps in education. Dozens of books marketed toward educators and the general public have been published with titles such as *Boys and Girls Learn Differently!* (Gurian et al., 2001) and *Teaching the Male Brain* (James, 2007). The instructional practices these books advocate tend to be based on gender essentialist beliefs, or beliefs that members of the same gender share inherent, unobservable, and meaningful similarities that cause observed gender differences in behavior between girls and boys (Gelman et al., 2004). For example, some of the authors argue that girls need cooperative classroom environments to succeed because they are often too sensitive for competitive environments, whereas boys need competitive classroom environments because by nature boys are competitive (Gurian et al., 2001).

The popularity of these gender essentialist strategies seem to be fueled, in part, by recent innovations in the field of neuropsychology. Neuroimaging studies have yielded new findings of gender differences (e.g., Baron-Cohen et al., 2005). For example, recent work has highlighted slight gender differences in hearing and the perception of pitch, which has been used to support a recommendation that teachers speak more loudly when talking to boys than when talking to girls (Sax, 2010). Neurologists, however, caution educators from changing classroom policies based on individual study findings; neuroimaging study results tend to be based on small sample sizes or effect sizes and are often highly tentative (Fine, 2010). Furthermore, there is likely within gender variability in hearing

and the perception of pitch for girls and boys. Accordingly, changing classroom policies to accommodate a finding at the group level might still be ineffective for some students.

Building on calls for gender-differentiated instruction, educators and policy makers around the world (and in Australia, the United Kingdom, and the United States, in particular) have begun to return to single-sex education, often in the form of all-girl and all-boy classes within coeducational schools (Younger and Warrington, 2006). Proponents of single-sex schooling believe that gender-segregated classrooms can help to ameliorate gender gaps in education by allowing teachers to use gender-differentiated instructional strategies while removing the distractions of other-sex peers (see Bracey, 2006). Opponents of single-sex schooling, in contrast, question the effectiveness of gender-differentiated instruction and worry that segregated classrooms increase students' gender stereotypes (Halpern et al., 2011). The debate about single-sex schooling is complicated by the difficulty in researching the effects of single-sex schooling on student outcomes. Single-sex schools are often schools of choice, and so there are both student and school selection factors that make it difficult to compare the effects of single-sex vs coeducation schooling (Bracey, 2006).

Gender-Blind Approaches

While some educators and policy makers have responded to the persistent gender gaps in education by focusing on gender, still others have argued for the opposite approach: deemphasizing gender. This approach comes out of intergroup theories that highlight the potentially derisive effects of highlighting gender as a social category (e.g., Developmental Intergroup Theory; Bigler and Liben, 2006). Deemphasizing gender as a category, these scholars argue, will reduce gender stereotypes, which, in turn, will grant students more freedom to pursue their academic interests and strengths without concerns about the 'gender-appropriate' approach (e.g., Bigler and Liben, 2012). A preschool in Sweden has put this perspective into action: the staff avoid the words 'him' and 'her' and instead refer to all children as 'friends' (Soffel, 2011). In the US, another intervention program is being developed with the goal of bringing girls and boys together. Program developers have taken the approach of discussing gender issues (such as bias and segregation) explicitly with teachers but have deemphasized gender in all discussions with children (Martin et al., 2012).

Future Directions

In the coming years, policy makers and researchers will continue to grapple with issues related to gender and education. In particular, further empirical work is needed that investigates mechanisms and developmental changes that may account for the changing educational gender gaps. For example, future work needs to examine the role of young children's gender-typed interests, and how those interests relate to their long-term educational outcomes, because we currently know very little regarding the degree to which children's interests motivate their performance in particular subject areas. Work examining the effect of changes in preferences and attitudes longitudinally would be particularly helpful in terms of identifying protective factors, risk factors, and mechanisms that relate to gender differences in educational outcomes. Although social developmental models have proposed pathways through which gender attitudes, stereotypes, and interests relate to educational outcomes (see, for example, Bigler and Liben, 2006; Ceci et al., 2009; Eccles, 1994), more data are needed to empirically support these theories.

Finally, in the coming years it will be particularly important to place issues related to gender and education in the proper context. As discussed, in many educational domains boys and girls are performing similarly (Hyde, 2005). Literature on stereotypes and stereotype threat also suggests that highlighting gender differences may serve to exacerbate gender discrepancies. Thus, care should be taken not to exaggerate the extent of the problem. Furthermore, researchers will need to pay particular attention to variations by nationality, race/ethnicity, and family income level. The gender gaps in the US, for example, are considerably larger among African-Americans and Hispanics than among Whites and, furthermore, the racial/ethnic achievement gaps (between, for example, African-Americans and Whites) are larger than any of the gender gaps in achievement (Freeman, 2005; Lee et al., 2007). Understanding the causes of these issues will require gender-focused researchers and policy makers to collaborate with experts from other areas of psychology and education.

Conclusion

In this article we have highlighted some of the critical issues related to gender and education facing policy makers and researchers. Although many of the traditional gender gaps in education have closed in recent years, discrepancies persist in areas related to achievement, attitudes, behaviors, and attainment. Several possible causes of these gaps have been proposed, ranging from gender differences in cognitive skills to differences in boys' and girls' socialization experiences and interests. Recently, educational programs have been proposed to address these issues and ameliorate the persisting gaps. In the coming years, researchers and policy makers will continue to grapple with how to best address the existing gaps, while also working to identify the causal mechanisms that may be affecting boys' and girls' chances for academic success.

See also: Academic Self-Concept and Achievement; Classroom Climate; Coeducation and Single-Sex Schooling; Cognitive Development: Child Education; Cognitive Styles and Learning Styles; Education in Cross-National Perspective: Findings from the Program for International Student Assessment (PISA); Environments for Education; Ethnicity and Educational Achievement; Families as Educational Settings; Family and Schooling; Self-Concepts: Educational Aspects; Self-Efficacy: Education Aspects; Sex Role Development and Education; Social Inequality and Schooling; Socioeconomic Status and Academic Achievement; Teacher Behaviours and Student Outcomes.

Bibliography

AAUW, 1992. How Schools Shortchange Girls. National Education Association, Washington, DC.

Baker, J., 1999. Teacher-student interaction in at-risk classrooms: differential behavior, relationship quality, and student satisfaction with school. The Elementary School Journal 100, 57–70.

Baron-Cohen, S., Knickmeyer, R.C., Belmonte, M.K., 2005. Sex differences in the brain: implications for explaining autism. Science 310, 819–823.

Bigler, R.S., Liben, L.S., 2006. A developmental intergroup theory of social stereotypes and prejudice. Advances in Child Development and Behavior 34, 39–89.

Bigler, R.S., Liben, L.S., 2012. If, when, why, and how to address gender in the classroom. In: Paper Presented at the American Educational Research Association Annual Meeting. Vancouver, BC.

Blair, C., Razza, R.P., 2007. Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. Child Development 78, 647–663.

Bowlby, J., 2005. Provincial Drop-out Rates: Trends and Consequences, Education Matters. Statistics Canada.

Bracey, G.W., 2006. Separate but Superior? A Review of Issues and Data Bearing on Single-sex Education. EPRU, Tempe, Arizona.

Brier, N., 1995. Predicting anti-social behavior in youngsters displaying poor academic achievement: a review of risk factors. Journal of Developmental and Behavioral Pediatrics 16, 271–276.

Brown, C.S., Leaper, C., 2010. Latina and European American girls' experiences with academic sexism and their self-concepts in mathematics and science during adolescence. Sex Roles 63, 860–870.

Castagnetti, C., Rosti, L., 2010. The gender gap in academic achievements of Italian graduates. In: Davies, S.A. (Ed.), Gender Gap: Causes, Experiences, and Effects. Nova Science Publishers, New York.

Ceci, S.J., Williams, W.M., Barnett, S.M., 2009. Women's underrepresentation in science: sociocultural and biological considerations. Psychological Bulletin 135, 218–261.

Crowley, K., Callanan, M.A., Tenenbaum, H.R., Allen, E., 2001. Parents explain more often to boys than to girls during shared scientific thinking. Psychological Science 12, 258–261.

Eccles, J.S., 1994. Understanding women's educational and occupational choices: applying the Eccles et al. model of achievement-related choices. Psychology of Women Quarterly 18, 585–610.

Eccles, J.S., Jacobs, J.E., Harold, R.D., 1990. Gender role stereotypes, expectancy effects, and parents' socialization of gender differences. Journal of Social Issues 46, 183–201.

Else-Quest, N.M., Hyde, J.S., Goldsmith, H.H., Van Hulle, C.A., 2006. Gender differences in temperament: a meta-analysis. Psychological Bulletin 132, 33–72.

Fine, C., 2010. From scanner to sound bite: issues in interpreting and reporting sex difference in the brain. Current Directions in Psychological Science 19, 280–283.

Finn, J.D., Pannozzo, G.M., Voelkl, K.E., 1995. Disruptive and inattentive-withdrawn behavior and achievement among fourth graders. The Elementary School Journal 95, 421–434.

Freeman, C.E., 2005. Trends in Educational Equity of Girls and Women: 2004 (NCES 2005-016). U.S. Government Printing Office, Washington, DC.

Frome, P.M., Alfeld, C.J., Eccles, J.S., Barber, B.L., 2006. Why don't they want a male dominated job? An investigation of young women who changed their occupational aspirations. Educational Research and Evaluation 12, 359–372.

Gelman, S.A., Taylor, M., Nguyen, S.P., 2004. Mother-child conversations about gender. Monographs of the Society for Research in Children Development 69, 1–126.

Gilliam, W.S., 2005. Prekindergarteners Left behind: Expulsion Rates in State Prekindergarten Systems. Retrieved from: http://www.plan4preschool.org/documents/pk-expulsion.pdf.

Goble, P., Martin, C.L., Hanish, L.D., Fabes, R.A., 2012. Children's gender-typed activity choices across preschool social contexts. Sex Roles 67 (7–8), 435–451.

Goldin, C., Katz, L., Kuziemko, I., 2006. The Homecoming of American College Women: The Reversal of the Gender Gap. NBER working paper.

Gurian, M., Henley, P., Trueman, T., 2001. Boys and Girls Learn Differently! A Guide for Teachers and Parents. Jossey-Bass, New York.

Halpern, D.F., 2000. Sex Differences in Cognitive Abilities, third ed. Erlbaum, Mahwah, NJ.

Halpern, D.F., Eliot, L., Bigler, R.S., Fabes, R.A., Hanish, L.D., Hyde, J., Liben, L.S., Martin, C.L., 2011. The pseudoscience of single-sex schooling. Science 333, 1706–1707.

Halpern, D.F., LaMay, M.L., 2000. The smarter sex: a critical review of sex differences in intelligence. Educational Psychology Review 12, 229–246.

Hendrick, J., Stange, T., 1991. Do actions speak louder than words? An effect of the functional use of language on dominant sex role behavior in boys and girls. Early Childhood Research Quarterly 6, 565–576.

Horn, L., Peter, K., Rooney, K., 2002. Profiles of undergraduates in U.S. postsecondary education institutions: 1999-2000. Education Statistics Quarterly 4, 79–86.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., Lyons, T., 1991. Early vocabulary growth: relation to language input and gender. Developmental Psychology 27, 236–248.

Hyde, J.S., 2005. The gender similarities hypothesis. American Psychologist 60, 581–592.

Hyde, J.S., Lindberg, S.M., Linn, M.C., Ellis, A.B., Williams, C.C., 2008. Gender similarities characterize math performance. Science 321, 494–495.

James, A.N., 2007. Teaching the Male Brain: How Boys Think, Feel, and Learn in School. Corwin Press, Thousand Oaks, CA.

Kleinfeld, J., 1998. The Myth that Schools Shortchange Girls: Social Science in the Service of Deception. The Women's Freedom Network, Washington, DC.

Langlois, J.H., Downs, A.C., 1980. Mothers, fathers, and peers as socialization agents of sex-typed play behaviors in young children. Child Development 51, 1237–1247.

Lee, J., Grigg, W., Dion, G., 2007. The Nation's Report Card: Mathematics 2007. U.S. Government Printing Office, Washington, DC.

Little, A.J., Leon de la Barra, B.A., 2009. Attracting girls to science, engineering and technology: an Australian perspective. European Journal of Engineering Education 34, 439–445.

Lytton, H., Romney, D.M., 1991. Parents' differential socialization of boys and girls: a meta-analysis. Psychological Bulletin 10, 267–296.

Maccoby, E.E., 1998. The Two Sexes: Growing Apart and Coming Together. Harvard University Press, Cambridge, MA.

Martin, C.L., Sallquist, J., Miller, C., Foster, S., 2012. Improving relationships between boys and girls: preliminary findings from the Sanford project. In: Paper presented at the Gender Development Conference. San Francisco, CA.

McGrayne, S.B., 2001. Nobel Prize Women in Science: Their Lives, Struggles, and Momentous Discoveries. Birch Lane Press, New York.

McHale, S.M., Kim, J.-Y., Whiteman, S., Crouter, A.C., 2004. Links between sex-typed time use in middle childhood and gender development in early adolescence. Development Psychology 40, 868–881.

Mickelson, R.A., 1989. Why does Jane read and write so well? the anomaly of women's achievement. Sociology of Education 62, 47–63.

NCES, 2010. Teacher Attrition and Mobility: Results from the 2008-09 Teacher Follow-up Survey (NCES 2010-353). U.S. Government Printing Office, Washington, DC.

NSF, 2009. Women, Minorities, and Persons with Disabilities in Science and Engineering: 2009 (NSF 09–305). National Science Foundation, Arlington, VA.

OECD, 2010. PISA 2009 at a Glance. OECD Publishing, Paris.

Perie, M., Moran, R., Lutkus, A.D., 2005. NAEP 2004 Trends in Academic Progress: Three Decades of Student Performance in Reading and Mathematics (NCES 2005–464). U.S. Government Printing Office, Washington, DC.

Perkins, R., Kleiner, B., Roey, S., Brown, J., 2004. The High School Transcript Study: A Decade of Change in Curricula and Achievement, 1990-2000. U.S. Government Printing Office, Washington, DC.

Plante, I., Theoret, M., Favreau, O.E., 2009. Student gender stereotypes: contrasting the perceived maleness and femaleness of mathematics and language. Educational Psychology 29, 385–405.

Planty, M., Hussar, W., Snyder, T., Kena, G., KewalRamani, A., Kemp, J., Bianco, K., Dinkes, R., 2009. The Condition of Education 2009 (NCES 2009–081). U.S. Government Printing Office, Washington, DC.

Pollack, W.S., 1998. Real Boys: Rescuing Our Sons from the Myths of Boyhood. Random House, New York.

Ready, D.D., LoGerfo, L.F., Burkham, D.T., Lee, V.E., 2005. Explaining girls' advantage in kindergarten literacy learning: do classroom behaviors make a difference? The Elementary School Journal 106, 21–38.

Rivers, C., Barnett, R.C., 2006. The myth of 'the boy crisis.' Washington Post.

Robinson, J.P., Lubienski, S.T., 2011. The development of gender achievement gaps in mathematics and reading during elementary and middle school: examining direct cognitive assessments and teacher ratings. American Educational Research Journal 48, 268–302.

Robinson, N.M., Abbott, R.D., Berninger, V.W., Busse, J., 1996. The structure of abilities in math-precocious young children: gender similarities and differences. Journal of Educational Psychology 88, 341–352.

Rowley, S.J., Kurtz-Costes, B., Mistry, R., Feagans, L., 2007. Social status as a predictor of race and gender stereotypes in late childhood and early adolescence. Social Development 16, 150–168.

Sax, L., 2010. Sex differences in hearing: Implications for best practice in the classroom. Advances in Gender and Education 2, 13–21.

Schaefer, B., 2004. A demographic survey of learning behaviors among American students. School Psychology Review 33, 481–497.

She, H.C., 2000. The interplay of a biology teacher's beliefs, teaching practice and gender-based student-teacher classroom interaction. Educational Research 42, 100–111.

Shepard, S.J., Nicpon, M.F., Haley, J.T., Lind, M., Liu, W.M., 2011. Masculine norms, school attitudes, and psychosocial adjustment among gifted boys. Psychology of Men & Masculinity 12, 181–187.

Silverman, I.W., 2003. Gender differences in delay of gratification: a meta-analysis. Sex Roles 49, 451–463.

Simpkins, S.D., Davis-Kean, P.E., Eccles, J.S., 2006. Math and science motivation: a longitudinal examination of the links between choices and beliefs. Developmental Psychology 42, 70–83.

Soffel, J., 2011. No 'him' or 'her': Preschool Fights Gender Bias. Associated Press.

Sommers, C.H., 2000. The War against Boys: How Misguided Feminism Is Harming Our Young Men. Simon & Schuster, New York.

Spencer, S.J., Steele, C.M., Quinn, D.M., 1999. Stereotype threat and women's math performance. Journal of Experimental Social Psychology 35, 4–28.

Steele, C., 1997. A threat in the air: how stereotypes shape intellectual identity and performance. American Psychologist 52, 613–629.

Sullivan, J., Riccio, C., Reynolds, C., 2008. Variations in students' school- and teacher-related attitudes across gender, ethnicity, and age. Journal of Instructional Psychology 35, 296–305.

Ueno, K., McWilliams, S., 2010. Gender-typed behaviors and school adjustment. Sex Roles 63, 580–591.

UNESCO, 2012. World Atlas of Gender Equality in Education. UNESCO Publishing, Paris.

Valeski, T.N., Stipek, D.J., 2001. Young children's feelings about school. Child Development 72, 1198–1213.

Voyer, D., Voyer, S., Bryden, M.P., 1995. Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. Psychological Bulletin 117, 250–270.

Weaver-Hightower, M.B., 2009. Issues of boys' education in the United States: diffuse contexts and futures. In: Martino, W., Kehler, M., Weaver-Hightower, M.B. (Eds.), The Problem with Boys' Education: Beyond the Backlash. Routledge, New York, pp. 1–35.

Younger, M.R., Warrington, M., 2006. Would Harry and Hermione have done better in single-sex classes? A review of single-sex teaching in coeducational secondary schools in the United Kingdom. American Educational Research Journal 43, 579–620.

Zill, N., West, J., 2001. Entering Kindergarten: A Portrait of American Children When They Begin School: Findings from the Condition of Education, 2000. U.S. Government Printing Office, Washington, DC.

Gender and Feminist Studies in Geography

Janice Monk, University of Arizona, Tucson, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Gender and feminist studies in geography address the material and symbolic lives of women and men as they are shaped by aspects of place and social relations and how considerations of gender are implicated in the creation of space and place. Initially inspired in the 1970s by women's rights movements, geographers questioned the exclusion of women's lives from their disciplines' theories, methods, and practices. Since then feminist geography has introduced new themes by attending not only to binary distinctions of male and female but by addressing the intersections of gender with other sources of social difference such as age, class, race, sexuality, physical ability, and religion at multiple scales from the body to the global. Diverse methodologies are brought to bear as research, teaching, and engagement with community organizations have been influenced.

Gender and feminist studies in geography address the material and symbolic lives of women and men as they are shaped by aspects of place and social relations and how considerations of gender are implicated in the creation of space and place. Initially inspired in the 1970s by women's rights movements geographers questioned the exclusion of women's lives from disciplinary theories, methods, and practices (Monk and Hanson, 1982). Since then feminist geography has introduced new themes by attending not only to binary distinctions of male and female but by addressing the intersections of gender with other sources of social difference such as age, class, race, sexuality, physical ability, and religion at multiple scales from the body to the global. Diverse methodologies are brought to bear as research, teaching, and engagement with community organizations have been influenced.

Geography of Feminist and Gender Studies in Geography

In considering the nature and development of feminist and gender studies in geography it is important to examine where and by whom they are produced within and across national settings. Early work in Britain and the United States was central in initiating the field. Literature from those countries has continued to expand and to have high international visibility, but it is also striking to see the diversification of topical, conceptual, and methodological approaches. One indication of the scope is that in 2012 the Geographic Perspectives on Women (GPOW) Specialty Group (for further information on GPOW see the website www.gpow.org (last accessed 21 November 2012)) at the Annual Meeting of the Association of American Geographers sponsored 57 papers or panel sessions. Reflecting the setting, American geographers (mostly women) accounted for the largest number of presenters but international participation was also notable, especially from Canada, Britain, and several western European countries. Indicative of the wider interests in gender research, 20 of the sessions were cosponsored by other specialty groups, such as those on political geography, sexuality and space, qualitative research methods, and socialist and critical geography.

Gender research has also expanded substantially internationally. Involvement in the Commission on Gender and Geography of the International Geographical Union (IGU) (for further information see http://igugender.socsci.uva.nl (last accessed 27 November 2012)) reveals substantial interest in many parts of the world. The Commission has over 600 subscribers on its electronic mailing list and has regularly organized sessions at IGU Congresses, most recently in Germany (2012), Chile (2011), and Israel (2010). It holds or cosponsors focused events in other locations (for example, in Brazil and Singapore (2011), India and New Zealand (2010), Hungary/Romania (2009), and Taiwan (2008)). Twice a year the Commission publishes a newsletter with an extensive bibliography, archived on its website. Themes and approaches continue to offer points of contact between feminism and the three core traditions of the discipline identified by Susan Hanson (1992): finding significance in everyday life, appreciation of the importance of context, and thinking about difference, though practices reflect differences across places (Monk, 1994). Distinctions in theoretical orientations and methods of research were evident between early British and American work, for example, possibly more so than in later years as circulation of literature increased. British feminist geographers initially drew on socialist theorizing and the concept of patriarchy to examine women's position in labor markets, whereas work in the United States followed empirical and positivist traditions, documenting differences between women's and men's spatial behavior in, for example, daily travel patterns or migration. By comparison, Dutch initiatives emerged from student critiques of Marxist writing and relatively quickly turned to the applied and policy-related work that characterizes much Dutch geography. Joos Droogleever Fortuijn has since characterized the situation as balancing the margin with the mainstream in two respects: engagement with the dominant Anglophone work while working within Dutch traditions, and integrating gender content into mainstream teaching (rather than teaching separate gender courses in geography) while also collaborating in interdisciplinary programs such as international development studies (Droogleever Fortuijn, 2011). In Spain, early studies reflected a national tradition of rural geography and were empirically rather than theoretically oriented. Drawing

initially on census data, Spanish feminist geographers then turned to field research and qualitative methods in order to reveal women's linked roles in production and social reproduction. Subsequently Spanish interests have diversified to address gender in urban spaces, youth and immigrant communities, and aspects of the history of Spanish colonialism in North Africa as represented in women's travel writings (Diaz-Cortés et al., 2007). These trajectories partly reflected changes in the national context and partly the stimulus of Anglophone work, though recently there have been expressions of concern about Anglophone hegemonies and anxiety that these mask the situatedness of work in other languages and traditions. Spanish interests in gender are also unevenly located across academic spaces, centered at the Autonomous University of Barcelona.

The work of geographers associated with the IGU Commission on Gender and Geography reveals yet more diversity. In 2012, geographers in at least 45 countries were represented on its electronic mailing list, with strong representation from India, Spain, Italy, and increasingly from Brazil and Japan. In the latter case there has been more significant participation by male geographers than is common elsewhere (Murata, 2005). Representation is low in most Middle Eastern and African nations, China, and the former socialist nations of Eastern Europe and the Soviet Union. Economic, political, cultural, and linguistic diversity across contexts, scholarly traditions, academic structures, resources, and leadership are evident in the themes, approaches, and the extent to which feminist geography is taught. In Latin America feminist geography is more evident in Argentina and Brazil than elsewhere and is expanding rapidly in Brazil. Whereas their early studies focused on class and economic themes (Veleda da Silva and Lan, 2007), sexuality studies have recently emerged, accompanied by the creation of online multilingual (Portuguese, Spanish, English) publications. (See http://www.revistas2.uepg.br/index.php/rlagg (last accessed 27 November 2012).) As a consequence, work that previously circulated mostly to local audiences is now distributed internationally. Gender studies in India have faced barriers in school and higher education but are gaining greater international visibility partly as a consequence of a conference in Delhi in 2010. (Co-led by Anindita Datta (University of Delhi) and Kuntala Lahiri-Dutt (Australian National University) it attracted about 100 participants, primarily from India but also from southeast Asia, Europe, and the United States.) Publications from the conference include a special issue of the international journal *Gender, Technology, and Development* (Datta, 2011) while a recent book (Raju and Lahiri-Dutt, 2011) addresses the trajectories in gender studies in Indian geography as well as offering new research by local geographers. Empirical rather than theoretical studies predominate using both qualitative and quantitative methodologies, while attention to policy implications is especially evident following a national tradition.

With diverse international participation, gender studies in geography address many topical themes and employ multiple paradigms. Still, there are common strands, primarily the understanding of gender as a complex and varying social, cultural, historical, and geographic construction, not simply a binary biological distinction between women and men, that power relations are implicated in gender relations, and that space, place, and gender are interdependent. Additionally, research is not confined to studies on scholars' own national settings. There is a substantial record of research, especially on the global South, by scholars who are based outside the global South and undertake work in regions where feminist geography is not currently widely evident. (Examples of such geographers who have made sustained contributions are Sarah Radcliffe (UK) on political themes in Andean Latin America, Sylvia Chant (UK) on poverty in Latin America, southeast Asia, and west Africa, Janet Momsen (UK, Canada, and US), in the Caribbean and more generally on gender and rural development; Melissa Wright (US) on gender and violence the Mexico–US border and Ragnhild Lund (Norway) in development research and with consultancies especially in Sri Lanka, but also in India, China, southeast Asia, and Africa including collaboration with local colleagues.) Additionally, geographers in regions from which migrants originate are turning attention to gender and migration in receiving sites. (See, for example, the multiple studies by Nora Chiang (Taiwan) in Canada, New Zealand, Australia, and most recently collaboratively in Argentina. http://homepage.ntu.edu.tw/~cos/html/personalE_chiang.htm (last accessed 2 April 2013).)

Concepts and Themes

Since the 1990s, feminist geography has diversified not only as a function of its geographic expansion but also through engagement with interdisciplinary currents, especially with the linguistic turn in the social sciences that foregrounded discourse analysis. They have engaged with perspectives emanating from postmodernist, poststructural, and postcolonialist scholarship that questioned grand narratives and essentialism and that destabilized dichotomous categories such as man–woman, mind–body, public–private, and global–local. Much attention has focused on the concept of the body – as experienced, politicized, and portrayed symbolically in different spaces and times. Especially important has been awareness of intersectionality – the recognition that gender identities and experiences intersect with, among others, those of race, ethnicity, class, sexuality, religion, age–life stage, dis/ability, indigeneity, and citizenship. Attention to the construction of masculinities has also emerged and become a theme of growing interest. Further, it is recognized that knowledge and categories are situated in place and time, both for the researcher and for the people who are the subjects of research. While race/class/gender may be accorded primacy in some Western settings, some geographers such as Ghanaian Mariama Awumbila (2007), have noted that for African scholars foundational categories may include kinship, seniority, religion, and region. In India, caste and communalism as well as region are likely to be addressed. Concepts of place identification have also been complicated by the emergence of 'transnationals,' people whose lives are not bounded by a single place and time.

Some examples are helpful for illustrating the various ways in which these concepts have been taken up. Robyn Longhurst (2001), in her New Zealand context, has addressed multiple aspects of bodies ranging from responses to and restrictions on women's pregnant bodies in public spaces to reflections on the spaces of losing weight, the meaning of food for

migrant women and mothers' engagement with electronic technologies (for examples of such studies see http://www.waikato.ac.nz/fass/about/staff/robynl (last accessed August 10, 2013). Vidyamali Samarasinghe (2008) has contrasted the national contexts of Nepal, Cambodia, and the Philippines as sites of sexual trafficking of women. She highlights ways in which varying cultural values, economies, links to sexual markets, and national policies and relations shape different forms of exploitation of young women. Worlds away, Linda McDowell's (1997) multifaceted examination of the environment of merchant banking firms in London reveals how multiple aspects of discourses and bodily performances of masculinity and femininity (for example in speech and dress, class background, media representations of behavior, and assumptions of gender roles), limit women's opportunities for career advancement.

Awareness has also developed that identities and behavior can be fluid in their construction and expression with movement across different spaces, situations, and scales. Geraldine Pratt, for example, drawing on work by Alison Blunt, notes that the nineteenth-century British travel writer, Mary Kingsley, was constrained at home by her class and gender positions whereas in Africa, positioned by her white racial and class privileges, she seemed to abandon those restrictions and to move between roles identified as masculine and feminine. (For the many works by Pratt on intercultural experiences and collaborations see http://www.geog.ubc.ca/~gpratt/Site/Publications.html (last accessed 3 April 2013).) In a vastly different setting, Mary Thomas's (2009) research with teenage girls in a Los Angeles area high school also reveals the fluidities of situatedness. Identifying as Hispanic, Armenian, Filipina, and African American, the girls clearly recognized strong racial–ethnic territories at the school. Yet some also spoke of having friendships across those divides. Tellingly, however, expressions of friendship were spatially restricted, and could not be acknowledged in encounters such as in a shopping mall when the 'friend' was with a homogenous group of another ethnicity. Nor did these girls contemplate being present in the home of the friend. While expecting such limitations, their conversations nevertheless revealed hurt and discomfort with such situations.

Engagements with the complexities of gendered identities and inequalities across cultures, scales, contexts, time, generations, religions, and public, private, and emotional arenas of life have become a significant and recurring concern for researchers. They are very well illustrated by multiple studies of transnational processes' experiences in Singapore by Huang, Yeoh, and colleagues (see, for example, the multiple studies of diverse experiences of foreign-born women workers and their Singaporean women employers by Shirlena Huang (http://profile.nus.edu.sg/fass/geoslena/) and Brenda S.A. Yeoh http://profile.nus.edu.sg/fass/geoysa/ (last accessed 2 April 2013)) and by Japanese geographer Yui (2009). The national government policies have shifted from emphases on limiting population growth to ideas about eugenics and fostering childbearing by native-born educated and professional couples in order to stimulate population growth. Though local women may support larger family sizes in their responses to surveys, in personal interviews they discuss 'stopping at two.' Their participation in the workforce has prompted substantial reliance on foreign-born women, especially Indonesians, Filipinas, and Sri Lankans, for domestic help and child care. Those women's lives and choices are regulated by state policies as well as practices of employers. The long work hours required, together with control of and restrictions on spaces within the home, and incidences of verbal and physical violence from employers, disrupt the early feminist representations of separations between 'public' and 'private' spaces. Simultaneously, other groups of foreign-born women in Singapore, such as Chinese 'study-mothers' who accompany their children to Singapore, face limited employment choices, financial stress, difficulties in obtaining housing, and being stereotyped as sexual threats. In the Singapore offices of Japanese international businesses, yet another national group, young educated Japanese women, are valued as clerical workers for their language and culturally appropriate social skills but are restricted in their opportunities for upward mobility compared to their male compatriots. Their long-term opportunities for employment in Singapore are, however, limited as they approach their thirties when they also face prospects that they will no longer be considered employable if they return to Japan (Yui, 2009). Collectively these studies reveal how class, ethnicity, citizenship status, and cultural differences shape experiences in particular places.

Methodologies and Methods

From the beginning, feminist researchers critiqued assumptions exhibited in the methodologies and empirical content of mainstream geography with implications for its theories. Monk and Hanson (1982), for example, pointed out that research on migration with interviews of 'heads' of household generally presented men as key actors and decision makers and women as passive followers, despite an absence of studies confirming how decisions were made and the existence data that showed women as numerically dominant in migration streams at particular times and places. Urban travel models that dealt with the journey to work assumed full-time employment (implicitly male), rather than attending to the more complex patterns of part-time (usually female) workers, or of people combining responsibilities for household and paid work. To counter such limitations, feminists advocated attention to arenas more central to women's lives such as child care, housework, relations within work places, and women's fears in public spaces. To understand such themes they favored research methodologies, especially qualitative and in-depth investigations that would reveal how women saw their own worlds and that had the potential to modify hierarchical power relations between the researcher and those being studied.

As the research progressed a variety of methodological approaches emerged with much attention devoted to the politics of creating knowledge and recognition of its partiality in both senses of this word – that no story is considered complete or objective. Research aimed to contribute to the feminist agenda for social change. Challenges multiplied as studies crossed boundaries such as class, cultures, and national settings, and recognized the different priorities of the theorizing academic and of communities such as nongovernmental organizations. Reflections on such issues appeared quite early, for example in a series of articles in *The Canadian Geographer* in

1993 on the aspects of ethnographic methods and others on dilemmas in field work and qualitative research published in *The Professional Geographer* in 1994 and 1995. Concerns included approaches to collaborating with community groups across class and cultural boundaries and how and to whom research should be presented. Such themes were later explored in depth by Geraldine Pratt in the course of her sustained relationship with immigrant women at the Philippine Women's Center in Vancouver. She has complemented interview methods with participation in group activities such as role-playing and creating theater performances. Richa Nagar turned to collective writing and publication in both India and the United States in her projects with community women who were working against domestic violence in India. She also recognized that such collaboration involved local participants in managing differences among themselves and with nongovernmental organizations and that they faced social and economic risks and issues of backlash (Nagar, 2013).

As postmodernist and poststructuralist approaches became more widely integrated into feminist geography, methodologies expanded to include interpretation of discourses, including not only the language of interviews but also written and visual expressions in public documents, the press, fiction, travel writing, films, works of art, graffiti, and to experimental approaches such as autophotography and participatory video. Examples are well represented in the journal *Gender, Place and Culture*. Founded in 1994, this journal initially published two issues per year and has gradually expanded so that eight will be issued in 2013, signaling the growth of interest in space and place within feminist and gender research.

Methodological diversity is being further extended by use of the technologies of geographic information systems (GIS), whose use in feminist projects has been contested as being objectifying and contrary to feminist concerns with subjectivity. Feminist geographers such as Mei-Po Kwan (2002) with substantial GIS expertise, have, however, demonstrated that GIS is useful in examining fine spatial scales and for integrating time and space, and can complement qualitative and quantitative information derived from interviews or time diaries. Also drawing increasing interest are participatory action research methods. Though not unique to feminist studies, they are seen as a way of bridging the divide between theorizing and contributing to social change (Kindon et al., 2007). Employing more traditional geographic methods to serve feminist goals, yet another initiative has been the creation of atlases to demonstrate spatial patterns of gender and women's lives. They address themes not commonly mapped such as the incidence of domestic violence or engagement in unpaid work. In so doing, these atlases offer information of value for local and national agencies. Examples include the multiple editions of *The State of Women in the World Atlas* (Seager, 2009) and gender atlases of India (Raju et al., 1999), Japan (Takeda et al., 2007), and Switzerland (Bühler, 2001).

Putting Feminist Geography into Professional Contexts

Important concerns of feminist geographers have been the representation of women within the profession, to whom and how feminist/gender geography is taught, and bringing feminist geographic research into community life. The first of these initially gained wide attention when Wilbur Zelinsky (1973) addressed *The Strange Case of the Missing Female Geographer*. Subsequent accounts have addressed changes over time and place and explored the climate for women geographers in a number of settings, exemplified in a theme section of the *Journal of Geography in Higher Education* (34 (1), 2004) which illustrated patterns in Hungary, the Netherlands, Singapore, and Spain, and the experiences of women of color geographers in Britain, the United States, and Canada. Addressing trajectories from the nineteenth into the twentieth centuries and including women geographers outside academia, Maddrell studied women in British geography (2009), and Monk (2004) examined women in US geography. They took into account national, political, and social contexts and academic structures and cultures. One theme is the place of women in teaching and its implications for inclusion of gender themes and perspectives in the curriculum and pedagogy, for example, in field courses in higher education and in school-level geography. Most recently a section of *International Research in Geographical and Environmental Education* (20 (3), 2011) was devoted to the implications of neoliberal and postsocialist political regimes for opportunities to include gender-focused courses in university curricula. Articles on feminist pedagogy and teaching gender recur in the *Journal of Geography in Higher Education* (Monk, 2000).

Given the roots of gender studies in feminist political movements, it is also important to recognize that direct and indirect connections with and commitments to action in communities remain a commitment within feminist work. The creation of atlases and the application of participatory research methods (discussed above) are such expressions. So too are projects that involve direct collaboration with community organizations as in Geraldine Pratt's work. Other recurring engagements are the collaborations of geographers in Switzerland with community planning agencies, such as those of geographers with the equal opportunity office of the Municipality of Zurich to deal with uses of the public spaces of city parks (Kaspar, 2012). These projects combined research with public exhibitions and coverage in the press and media (E-mail, Elisabeth Bühler to Janice Monk, 8 July 2011). In Berne, such collaborations have addressed employment for migrant women involving members of the city's Gender Equality Board in leadership, workshops designed not only to gather research data but to assist women in developing their social capital, publications that take the form of recommendations to policy maker and nongovernmental organizations, and establishment of a mentoring project by the city government (E-mail, Yvonne Riano to Janice Monk, 11 July 2011).

Status and Emerging Directions

In a little over three decades, gender and feminist studies have become well established in geography and spread internationally, if unevenly, from their main origins in Britain and the United States. Their critical perspectives have contributed to reshaping epistemologies across subfields of human geography. Attention to gender and feminist geography is now

commonly found in texts introducing human geography, in compendia on trends in social, political, cultural, and economic geography including one devoted to feminist geography (Nelson and Seager, 2005), and in the reviews of directions in the discipline in the journal *Progress in Human Geography*. Nevertheless, there are still areas of the discipline where feminist scholars consider that attention needs to be strengthened. Though a feminist political ecology emerged in the 1990s (Rocheleau et al., 1997), resource management studies have recently been criticized for inadequate attention to gender (Reed, Christie, 2009) and for inadequate recognition of feminist perspectives in studies of global concerns about climate change and water resources (Elmhirst, 2011). While feminist geographies continue to focus primarily on aspects of women's lives, the work on bodies and sexualities certainly prompted new perspectives and widened related interpretations of gender across the discipline to include socially and culturally shaped masculinities, geographies of children and youth, race, and religion. There is every indication that gender and feminist studies will remain a vital field within geography and that this geographical work will sustain cross-disciplinary attention.

See also: Children's Geographies; Cultural Geography; Knowledge, Geography of; Postcolonial Geography; Qualitative Methods in Geography; Sexuality and Space.

Bibliography

Awumbila, M., 2007. Challenging contexts: gender studies and geography in Anglophone African countries. Belgeo (3), 261–274.

Bühler, E., 2001. Frauen-und Gleichstellungsatlas Schweiz. Seismo Verlag, Zurich.

Datta, A., 2011. Introduction. Gender, Technology and Development 15 (3), 339–343.

Diaz-Cortés, F., Garcia-Ramon, M.D., Ortiz, A., 2007. Engendering Spanish geography. Belgeo (3), 323–333.

Droogleever Fortuijn, J.C., 2011. Teaching geography and gender: the case of the Netherlands. International Research in Geographical and Environmental Education 20, 175–178.

Elmhirst, R., 2011. Introducing new feminist political ecologies. Geoforum 42, 129–132.

Hanson, S., 1992. Geography and feminism: worlds in collision? Annals of the Association of American Geographers 82, 569–586.

Kaspar, H., 2012. Erlebnis Stadtpark Nutzung un Wahrnehmung urbaner Grüneaüme. Verlag VS, Wiesbaden.

Kindon, S., Pain, R., Kesby, M., 2007. Participatory Action Research: Origins, Approaches, and Methods. Routledge, London.

Kwan, M.-P., 2002. Feminist visualization: re-envisioning GIS as a method in feminist geographic research. Annals of the Association on American Geographers 92, 645–661.

Longhurst, R., 2001. Bodies: Exploring Fluid Boundaries. Routledge, London.

Maddrell, A., 2009. Complex Locations: Women's Geographical Work in the UK 1850–1970. Wiley-Blackwell, Oxford and Malden, MA.

McDowell, L., 1997. Capital Culture. Gender at Work in the City. Blackwell Publishers, Oxford.

Monk, J., 1994. Place matters: comparative international perspectives on feminist geography. The Professional Geographer 46, 277–288.

Monk, J., 2000. Looking out, looking in: the 'other' in the "Journal of Geography in Higher Education". Journal of Geography of Higher Education 24, 163–177.

Monk, J., 2004. Women, gender and the histories of American geography. Annals of the Association of American Geographers 94, 1–22.

Monk, J., Hanson, S., 1982. On not excluding half of the human in human geography. The Professional Geographer 34, 11–23.

Murata, Y., 2005. Gender equality and progress of gender studies in Japanese geography: a critical review. Progress in Human Geography 29, 260–275.

Nagar, R., 2013. Storytelling and co-authorship in feminist alliance work: reflections from a journey. Gender, Place and Culture 20, 1–18.

Nelson, L., Seager, J (Eds.), 2005. A Companion to Feminist Geography. Blackwell Publishing, Malden, MA; Oxford, UK and Carlton VIC, Australia.

Raju, S., Atkins, P.J., Kumar, N., Townsend, J., 1999. Atlas of Women and Men in India. Kali for Women, Delhi.

Raju, S., Lahiri-Dutt, K. (Eds.), 2011. Doing Gender, Doing Geography Emerging Research in India. Routledge, New Delhi and Abingdon.

Reed, M., Christie, S., 2009. Environmental geography: we're not quite home – reviewing the gender gap. Progress in Human Geography 33, 246–255.

Rocheleau, D., Thamas-Slayter, B., Wangari, E. (Eds.), 1997. Feminist Political Ecology: Global Perspectives and Local Experience. Routledge, London.

Samarasinghe, V., 2008. Female Sex Trafficking in Asia: The Resilience of Patriarchy in a Changing World. Routledge, London and New York.

Seager, J., 2009. The Penguin Atlas of Women in the World, fourth ed. Penguin, New York and London.

Takeda, Y., Inoshita, R., Takashi, N., Wakabayashi, Y., Kamiya, H., Yano, K., 2007. Chizu de miru Nihon no Josei (Gender Atlas of Japan). Akashi Shoten, Tokyo.

Thomas, M.E., 2009. The identity politics of school life: territoriality and social subjectivity of teen girls in Los Angeles. Children's Geographies 7, 7–19.

Veleda da Silva, S., Lan, D., 2007. Geog, and gender studies: the situation in Brazil and Argentina. Belgeo (3), 371–382.

Yui, Y., 2009. Japanese women's work overseas and the activities of recruitment agencies in Singapore. Journal of Geographical Science 57, 55–70.

Zelinsky, W., 1973. The strange case of the missing female geographer. The Professional Geographer 25, 101–106.

Gender and Health Care

Ute Habel and Birgit Derntl, RWTH Aachen University, Aachen, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.R. Anspach, volume 9, pp. 5948–5954, © 2001, Elsevier Ltd.

Abstract

In 2001, the Institute of Medicine (USA) called on biomedical researchers to step up their investigation of sex and gender as critical variables affecting health (Wizemann and Pardue, 2001). However, the authors also pointed out that the terms gender and sex have been used inconsistently and confusingly. While the definition of sex is relatively clear in that it is "the classification of living things, generally as male or female according to their reproductive organs and functions assigned by the chromosomal complement," the confusion seems to lie with the term gender that seems to have become almost synonymous with the word sex (Ristvedt, 2014). Gender refers to the socially constructed aspects of being female or male, i.e., the stereotypic psychological and behavioral characteristics presumably shaped by societal expectations (Haig, 2004). For future research it is mandatory that both aspects are investigated with a sound definition, which is acknowledged in the scientific and public discourse. Gender is of substantial relevance in every respect of health care and requires intensified research in basic and clinical areas, analyzing gender stereotypes and roles with respect to their influence in patient care, and checking for gender differences in research designs and results. Research evidence revealed that the field of health has to deal with numerous gender differences requiring not only an awareness in experts, clinicians, patients, and researchers for these differences but also a consciousness for more implicit gender influences exerted by each involved individual, being it researcher, clinician, or patient.

To be female or male is one of the most important biological determinants of life with critical consequences on many aspects ranging from cognitive and emotional behavior as far as human health (Wizemann and Pardue, 2001). Due to culture-specific gender norms and values, the distinct roles and behaviors of men and women give rise to gender differences. Gender norms and values, however, also implicate gender inequalities – that is, differences between men and women, which systematically empower one group to the detriment of the other. Both gender differences and gender inequalities can give rise to inequities between men and women in health status and access to health care (WHO, 2014).

As a consequence of Women's health movement and the feminist critique of science in the 1980s, the field of gender medicine evolved. Gender medicine is defined as "an unbiased comparison between women and men and the inclusion of gender as a socio-cultural process into medical hypotheses. It includes the recognition of biological sex differences among women and men, i.e., sex differences, but goes well beyond that" (Regitz-Zagrosek, 2012a,b). The subject of gender in health care has become even more prominent after 2000, with the increase in interest in biological differences between men and women and with the installation of specific centers for research in gender medicine, for example, in Vienna the Gender Medicine Unit at the Medical University, in Berlin the Institute of Gender in Medicine at the Charité, and in Stockholm the Center for Gender Medicine at the Karolinska Institute. Furthermore, gender interests are also represented within specific societies, such as the Organization for the Study of Sex Differences (OSSD) and the International Society of Gender in Medicine. General aim is to address women's and men's specific health issues in research and clinical practice to better understand and tackle the individual needs in health and disease.

Sociocultural and biological sex-specific influences exert their effects already in the fetus and are active lifelong. For example, lifestyle and fitness may lead to epigenetic modifications that not only affect the adult but also a potential developing fetus. Moreover, endocrinology and differential sex hormonal influences represent biological factors that affect structural brain development in the fetus but in addition to these organizing effects also have persisting major active effects that contribute to differences between men and women. And both these factors interact at every age and level, making it nearly impossible to exactly differentiate the influence of sex and gender, but rather taking into account the complex interaction in the field of health care and research (Haig, 2004; Rieder and Lohff, 2004; Ristvedt, 2014).

Gender and Medical Research

Up to the end of the last century, much of the human research has been performed in men, defining the male norm, and studies just neglected or failed to address specific gender differences. In the United States the Office of Women's Health Research at the National Institute of Health (NIH) achieved a change of the standard by which women as objects of research were excluded (Klinge, 2010), to be included in clinical research from 1994 onward. The same applies to studies on the pathophysiology of somatic or psychiatric diseases, which often show a marked difference in the prevalence for women and men. The divergent frequencies are also in most of the cases far from being clarified. In animal models, too, male animal models still predominate the field (Zucker and Beery, 2010). In 8 of 10 analyzed biological disciplines, authors found a significant sex bias in animal experiments, with a female bias only in studies on reproduction or immunology. Looking at the rate of female animals in studies of particular diseases, it emerged, that less than 45% of the studies on anxiety and depression used female animals, although these disorders have a more than two times higher prevalence in women. The same applies to drug studies on behavior, nearly exclusively performed in male animals, despite the fact that

absorption and effect rates are different in men and women. This is most often based on the assumption that female animals are less valid and reliable due to the estrous cycle, which has however been contradicted (Gold and Voskuhl, 2009).

The implications for human health care are obvious and perpetuate the gender bias. Zucker and Beery (2010) clearly voted for a correction of this bias by a refuse to tolerate it in publications and research applications. Other steps in that direction come from the formation of the OSSD in Washington DC (USA), and a new journal on the 'Biology of Sex Differences.' Certainly, adding female animals and humans to the research design not only makes it more complex and raises potential interaction effects, it also increases sample sizes and the number of influencing factors, that have to be taken into account, such as cycling phase. Based on the aim to advance the idea of personalized medicine we would like to stress that there is no way to neglect gender and its interactions in research, whatever the costs are, as it is one of our most formative factors in life. At the same time, this also ensures and increases the generalizability, quality, and validity of our research results and it guarantees equality in health care for men and women. As Ineke Klinge puts it: "Intriguingly, the nondiscriminatory proposition put forward by Gender Medicine is: to treat persons equally in health care often comes down to making a difference!" (Klinge, 2010).

Clinical studies face the fact that gender may affect the pathophysiology, the symptomatology, and the outcome of disorders, not overlooking the influence of gender stereotypes on behavior, as well as social gender roles and gender-specific environmental factors, such as prevention use, nutrition, stress, work, and lifestyle. Gender-specific behavior can initiate epigenetic changes, which are affected by estrogen receptors and finally may lead to differential results in male and female fetuses, which may determine gender-specific risk factors for certain diseases (Regitz-Zagrosek, 2012). Gender stereotypes and roles, for example, influence help-seeking behavior, with men seeking help less often and later (Möller-Leimkühler, 2013), and females showing a higher adherence to treatment and a different risk perception. Another very famous example is coronary heart disease (CHD), once highly associated to male gender and stress in managers, but now being acknowledged as one of the leading causes of death in women, with a differential symptom profile. This long neglected clinical field of relevant differences has produced a lot of research since its first discoveries and led to numerous relevant gender differences within CHD. In consequence, aspirin was established as a preventive measure for CHD, but this was based on trials with mainly male subjects. In the meantime, meta-analyses revealed that aspirin is effective in reducing the myocardial risk in men only, whereas in females it rather risks bleeding events (Berger et al., 2006).

Focusing on the specific needs and challenges in male health, it became obvious in a large EU project including 34 countries, that men have a higher mortality rate than women, that 50% of early deaths in men could be prevented, that CHD is still the main cause of death in men and men use preventive screenings less often than women. Moreover, depression and other mental disorders are underdiagnosed and not sufficiently treated in men (Bardehle, 2013). This shows that depending on where one sets the focus, on men or women, inequalities, and problems in health care arise in both. This should be addressed accordingly and where necessary in a gender-specific way.

Gender in neuroscience is another topic that merits attention (Cahill, 2006). Brain differences between men and women have been an issue neglected for a long time in neuroimaging research and sometimes even attacked since there has been a long-standing tradition, supported even by famous scientists, to use brain differences for the discrimination and to the disadvantage of women. In the last decades, the enormous potential of imaging techniques and their further development have led to a more specific characterization of the structure and the functional neuroanatomy of gender differences. Imaging studies shed light on neurobiological differences and potential influencing factors of the cerebral substrates of emotional and cognitive processes. An increasing number of recent publications on gender differences in neural correlates of different cognitive–emotional functions in healthy subjects demonstrate an augmenting interest in and acknowledgment of such differences. The number of studies on gender differences grew since then dramatically. Considering methodological differences and study constraints, oversimplifications and overinterpretations have to be avoided and interpretations have to be made with caution. The recognition of structural and functional cerebral differences may however have clinical implications with respect to individually optimized diagnoses and therapy of disorders.

In this field we are also embedded in a controversial discussion on biological determinism vs individual experience affecting the brain. We would like to stress the view, that behavior and brain are closely intertwined and as such, not only behavior, but the brain too is a product of biological influences as well as environmental experiences in their interaction. There are numerous findings to support this view plus we have to acknowledge the enormous plasticity of the brain over time. We always only assess a momentum of its structure and function, but the brain is subject to continuous dynamic changes again stemming from biological (i.e., aging) as well as environmental individual sources (i.e., training). The brain shows sexual dimorphism in structure, with a larger volume and a higher percentage of white matter and cerebrospinal fluid in men (overview in Habel and Derntl, 2013) whereas women reveal a higher rate of gray matter. However, the direct relationship between brain structure and function is still unknown and the same applies to brain–behavior relationships. The fact that there are no behavioral sex differences does not automatically imply that there are no brain functional differences and vice versa. Such functional gender differences are often a matter of debate due to contradictory results and have indeed a number of methodological challenges (see Ihnen et al., 2009). One example is, that is has often been stated, that women show better verbal abilities and this may be based on a differential functional lateralization. This stimulated a number of imaging studies with quite equivocal results. First, looking at this verbal superiority in greater detail, neuropsychological findings most often show a functional superiority of one or the other gender only restricted to certain subfunctions and not to more general abilities. The frequently held assumption, that language functions are less lateralized in females could finally not be supported by a large meta-analysis on 14 studies from 1995 to 2004 (Sommer et al., 2008). Besides cognitive abilities, the same

inconsistencies can also be found regarding emotional functions (Kohn et al., 2011; Derntl et al., 2010). In the meantime the growing interest in gender differences risks to provoke a publication bias with a higher probability for publication of studies on gender differences, which in turn strongly impacts on researchers to stress such differences.

Hence, these research examples show that much effort has been put in reducing the inequality, but still there are a number of challenges that have to be tackled. Gender and sex differences have to be taken into account in all domains of research, from basic animal research to neuroscience and public health, and also in data analyses and interpretation of results. This has been already stressed in the EU policy for research, acknowledging and pointing the attention to the fact that biological mechanisms, causes, clinical manifestation, consequences, and treatment of disease and disorders often differ between men and women being the topic of several calls starting with the Framework Programme 6 (2002–07). It is however also mandatory to promote the distribution of knowledge on these gender and sex influences to experts and the public.

Gender and Health Care Users

Whether the patient's gender leads to differences in treatment is still a matter of debate. Both, the structure of the medical interview and the treatment of similar medical problems may be affected by gender of the patient. Research comparing different aspects of physician communication observed that female patients have more contacts in general and more interactions characterized by in-depth talks compared to male patients (Bertakis et al., 2003). Studies analyzing pharmacological interventions found an undertreatment of women in the field of cardiology, especially concerning heart failure (Baumhakel et al., 2009) as well as CHD (Bischoff et al., 2006). Further, studies showed that female patients obtain less evidence-based medical therapy and fewer investigations such as coronary angiography in case of stable angina and coronary disease (Daly et al., 2006) as well as laboratory tests (Hamberg et al., 2002). In contrast, diagnostic support of physiotherapists and orthopedists were more often recruited for female patients (Hamberg et al., 2002). This tendency seems to be influenced by gender-specific expectations and perceptions creating a priori assumption about the needs of the patient.

Research in dental offices also has shown that the physician–patient interaction seems to be influenced by gender stereotypes and gender identities in both directions: Considering the percentage of total talk, the gender-specific patient–physician interaction ranged from the lowest percentage for male gender–concordant interactions to the highest percentage in female gender–concordant interactions. Interestingly this did not affect the patients' satisfaction with the physician (Peck and Denney, 2012).

Gender and Health Care Providers

According to several studies, female and male health care providers differ in practice styles with potential implications for patient health care use and mortality. For example, female physicians show a more patient-centered communication style by devoting more time to aspects such as psychosocial counseling, providing encouragement and reassurance as well as social exchange (Bertakis et al., 2009). In their review article, Roter and Hall (2004) conclude that data from the few previous studies on the effects of physician gender on the health care process indicate that female physicians spend more time during a typical visit talking with their patients than do male physicians. During this time, they engage in communication that more broadly relates to the larger life context of the patient's condition by addressing psychosocial issues through related questioning and counseling and through greater use of emotional talk, positive talk, and active enlistment of patient input.

In contrast to the higher levels of psychosocial and socio-emotional exchange, there is little evidence that physician gender is related to the more task-specific communication elements of care. Moreover, the authors point out that results from previous studies do not suggest that physician gender was related to the provision of biomedical information (including discussion of the diagnosis, prognosis, and medical treatment).

There is an ongoing debate whether female physicians are more likely than men to provide certain morbidity- and mortality-reducing preventive services to women. While some studies observed this effect particularly in cancer screening, for example, use of mammography (e.g., Franks and Bertakis, 2003), very recent data however indicate no significant association of gender of health care provider and health care utilization or mortality (Jerant et al., 2013). The authors speculate that the heterogeneity of findings might be based on the different methodological approaches as they relied on a huge sample of the US national data and used adjusted analyses of prospective data.

In sum, both patient and physician contribute to differences in health care by their gender-specific expectations. Subsequently, gender differences have significant consequences on medical practice and therapy effects. This does not mean that men and women are always treated differently or should be treated differently; studies just endorse that gender expectations are present.

Regarding the working conditions and their impact on health and well-being of health care providers, recent data indicate that high work load and work stress particularly affect female and older physicians (Fletcher et al., 2011). Especially mental health issues, ranging from burnout, depressive mood to suicidal thoughts, are a more frequent consequence of work stress and long working hours in female than male physicians (Hayasaka et al., 2007). This, however, seems to be influenced by discipline; for example, female surgeons report more work overload and are less satisfied with their work compared to female internal specialists (Bergman et al., 2003). Besides discipline, the work setting also seems to influence work satisfaction and stress response. Several studies indicate that the physicians and nurses working in intensive care units are specifically prone to developing depressive symptoms and burnout, and being female seems to increase the risk (Michalsen and Hillert, 2011).

Research suggests that empathic engagement in patient care can lead to positive clinical outcomes by improving quality of

life and reducing emotional stress (e.g., Hojat et al., 2013). More than 10 years of studies on empathy in medical students and physicians indicate that empathy tends to erode during medical school and in residency training (Neumann et al., 2011). Furthermore, higher empathy scores have been linked to better ratings of clinical competence in core medical school clerkships and interest in primary care specialties that often require first-encounter and long-term face-to-face interaction with patients (Hojat et al., 2002a). Of note, female medical students as well as physicians reported significantly higher empathy scores than males (Hojat et al., 2002b), which corroborate findings from empathy studies outside the health care system (Derntl et al., 2010). On the other hand, recent findings from Hojat et al. (2013) indicate that empathy can be enhanced and sustained in female *and* male medical students as measured via self-report questionnaires.

Conclusion/Outlook

Our overview illustrates that bringing gender in the last decades into the focus of health care and research has had beneficial effects with respect to an optimized personalized patient care. It also created an awareness of the topic and advanced the field and knowledge on the gender-specific pathophysiology, symptomatology, and illness course for different disorders. Although a number of achievements have been accomplished there is still a lot more to do, to minimize gender inequality in health for patients, clinicians, and researchers.

See also: Adolescent Health and Health Behaviors; Doctor–Patient Interaction in the West: Psychosocial Aspects; Health Psychology in Childhood; Human Development and Health; Life Course Epidemiology; Social Integration, Social Networks, and Health.

Bibliography

Baumhakel, M., Muller, U., Bohm, M., 2009. Influence of gender of physicians and patients on guideline-recommended treatment of chronic heart failure in a cross-sectional study. The European Journal of Heart Failure 11 (3), 299–303.

Berger, J.S., Roncaglioni, M.C., Avanzini, F., Pangrazzi, I., Tognoni, G., Brown, D.L., 2006. Aspirin for the primary prevention of cardiovascular events in women and men: a sex-specific meta-analysis of randomized controlled trials. The Journal of the American Medical Association 295 (3), 306–313.

Bergman, B., Ahmad, F., Stewart, D.E., 2003. Physician health, stress and gender at a university hospital. Journal of Psychosomatic Research 54 (2), 171–178.

Bertakis, K.D., Franks, P., Azari, R., 2003. Effects of physician gender on patient satisfaction. Journal of the American Medical Women's Association 58 (2), 69–75.

Bertakis, K.D., Franks, P., Epstein, R.M., 2009. Patient-centered communication in primary care: physician and patient gender and gender concordance. Journal of Womens Health (Larchmt) 18 (4), 539–545.

Bardehle, D., 2013. Psychische Gesundheit des Mannes – international und national. In: Weißbach, L., Stiehler, M. (Eds.), Männergesundheitsbericht 2013. Verlag Hans Huber, Bern.

Cahill, L., 2006. Why sex matters for neuroscience. Nature Reviews Neuroscience 7 (6), 477–484. http://dx.doi.org/10.1038/nrn1909.

Daly, C., Clemens, F., Lopez Sendon, J.L., Tavazzi, L., Boersma, E., Danchin, N., Fox, K.M., 2006. Gender differences in the management and clinical outcome of stable angina. Circulation 113 (4), 490–498.

Derntl, B., Finkelmeyer, A., Eickhoff, S., Kellermann, T., Falkenberg, D.I., Schneider, F., Habel, U., 2010. Multidimensional assessment of empathic abilities: neural correlates and gender differences. Psychoneuroendocrinology 35 (1), 67–82.

Fletcher, J.M., Sindelar, J.L., Yamaguchi, S., 2011. Cumulative effects of job characteristics on health. Health Economics 20 (5), 553–570.

Franks, P., Bertakis, K.D., 2003. Physician gender, patient gender, and primary care. Journal of Womens Health (Larchmt) 12 (1), 73–80.

Gold, S.M., Voskuhl, R.R., 2009. Estrogen and testosterone therapies in multiple sclerosis. Progress in Brain Research 175, 239–251.

Habel, U., Derntl, B., 2013. Geschlechtsabhängige Effekte. In: Schneider, F., Fink, G. (Eds.), Funktionelle MRT in Psychiatrie und Neurologie. Springer Berlin Heidelberg, pp. 203–214.

Haig, D., 2004. The inexorable rise of gender and the decline of sex: social change in academic titles, 1945–2001. Archives of Sexual Behavior 33 (2), 87–96.

Hamberg, K., Risberg, G., Johansson, E.E., Westman, G., 2002. Gender bias in physicians' management of neck pain: a study of the answers in a Swedish national examination. Journal of Women's Health & Gender-Based Medicine 11 (7), 653–666.

Hayasaka, Y., Nakamura, K., Yamamoto, M., Sasaki, S., 2007. Work environment and mental health status assessed by the general health questionnaire in female Japanese doctors. Industrial Health 45 (6), 781–786.

Heinke, W., Dunkel, P., Brahler, E., Nubling, M., Riedel-Heller, S., Kaisers, U.X., 2011. Burnout in anesthesiology and intensive care: is there a problem in Germany? Anaesthesist 60 (12), 1109–1118.

Hojat, M., Gonnella, J.S., Nasca, T.J., Mangione, S., Veloksi, J.J., Magee, M., 2002a. The Jefferson Scale of Physician Empathy: further psychometric data and differences by gender and specialty at item level. Academic Medicine 77 (Suppl. 10), S58–S60.

Hojat, M., Gonnella, J.S., Nasca, T.J., Mangione, S., Vergare, M., Magee, M., 2002b. Physician empathy: definition, components, measurement, and relationship to gender and specialty. The American Journal of Psychiatry 159 (9), 1563–1569.

Hojat, M., Louis, D.Z., Maio, V., Gonnella, J.S., 2013. Empathy and health care quality. American Journal of Medical Quality 28 (1), 6–7.

Ihnen, S.K., Church, J.A., Petersen, S.E., Schlaggar, B.L., 2009. Lack of generalizability of sex differences in the fMRI BOLD activity associated with language processing in adults. Neuroimage 45 (3), 1020–1032.

Jerant, A., Bertakis, K.D., Fenton, J.J., Franks, P., 2013. Gender of physician as the usual source of care and patient health care utilization and mortality. The Journal of the American Board of Family Medicine 26 (2), 138–148.

Klinge, I., 2010. Sex and gender in biomedicine: promises for women and men. How incorporation of sex and gender in research will lead to a better health care. In: Klinge, I., Wiesemann, C. (Eds.), Sex and Gender in Biomedicine. Theories, Methodologies, Results. Universitätsverlag Göttingen.

Kohn, N., Kellermann, T., Gur, R.C., Schneider, F., Habel, U., 2011. Gender differences in the neural correlates of humor processing: implications for different processing modes. Neuropsychologia 49 (5), 888–897.

Michalsen, A., Hillert, A., 2011. Burnout in anesthesia and intensive care medicine: part 2: epidemiology and importance for the quality of care. Anaesthesist 60, 31–38.

Möller-Leimkühler, A., 2013. Psychische Gesundheit von Männern: Bedeutung, Ziele, Handlungsbedarf. In: Weißbach, L., Stiehler, M. (Eds.), Männergesundheitsbericht 2013. Verlag Hans Huber, Bern, p. 20.

Neumann, M., Edelhauser, F., Tauschel, D., Fischer, M.R., Wirtz, M., Woopen, C., Scheffer, C., 2011. Empathy decline and its reasons: a systematic review of studies with medical students and residents. Academic Medicine 86 (8), 996–1009.

Peck, B.M., Denney, M., 2012. Disparities in the conduct of the medical encounter. SAGE Open 2 (3).

Regitz-Zagrosek, V., 2012a. Sex and gender differences in cardiovascular disease. In: Oertelt-Prigione, S., Regitz-Zagrosek, V. (Eds.), Sex and Gender Aspects in Clinical Medicine. Springer, London.

Regitz-Zagrosek, V., 2012b. Sex and gender differences in health. Science & Society series on sex and science. EMBO Reports 13 (7), 596–603.

Rieder, A., Lohff, B., 2004. Gender medizin. Dtsch Arztebl International 101 (48), A-3259.

Ristvedt, S.L., 2014. The evolution of gender. JAMA Psychiatry 71 (1), 13–14.

Roter, D.L., Hall, J.A., 2004. Physician gender and patient-centered communication: a critical review of empirical research. The Annual Review of Public Health 25, 497–519.

Sommer, I.E., Aleman, A., Somers, M., Boks, M.P., Kahn, R.S., 2008. Sex differences in handedness, asymmetry of the planum temporale and functional language lateralization. Brain Research 1206, 76–88.

Wizemann, T.M., Pardue, M.L., 2001. Exploring the Biological Contributions to Human Health: Does Sex Matter? The National Academies Press.

World Health Organization, 2014. Gender, women and health: why gender and health?. Retrieved 11.02.14., 2014 from: http://www.who.int/gender/genderandhealth/en/index.html.

Zucker, I., Beery, A.K., 2010. Males still dominate animal studies. Nature 465 (7299), 690.

Gender and Language: Cultural Concerns

Ruth Wodak, Lancaster University, Lancaster, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Research on gender and language started in the United States in the late 1960s and early 1970s and quickly took off across the Western World. At that time it was closely linked to the women's liberation movement and had many practical implications such as the creation of guidelines for nondiscriminatory language behavior for official purposes in bureaucracies, media, publishers, and so forth. There are many different theoretical and methodological approaches, which also imply a range of differing ideological positionings. Currently, in the twenty-first century, we encounter an important focus on more critically inspired work and on the integration of various dimensions such as social class, gender, ethnicity, and so forth in the actual empirical investigation, an approach termed intersectionality. Moreover, the performative aspect of social and thus also gendered language practices is increasingly salient.

Setting Out the Problems

Research on gender and sex in sociolinguistics and discourse analysis started systematically in the early 1970s, probably as a result of women's liberation movements. Investigators examined two domains of language behavior in particular: speech behavior of men and women on the phonological level, and interactions (conversational styles) between women and men in discourse (for overviews and the history of the field see Hall and Buchholtz, 1995; Crawford, 1995; Harrington et al., 2008; Holmes and Meyerhoff, 2003; Kotthoff and Wodak, 1997; Lazar, 2005; Mills, 2004; Sunderland, 2004; Talbot, 1998; Wilkinson and Kitzinger, 1995; Wodak, 1997; Wodak and Benke, 1996).

Studies of gender-specific language behavior depend – as does all research in the social sciences – on the authors' implicit assumptions about sex and gender, methodology, and samples used, etc. For example, some studies suggest that women use more standard language than men because they try to adapt to social norms (Trudgill, 1974). On the other hand, other large-scale surveys have shown that young women tend to speak dialect and nonstandard language as well (Wodak, 1984), due to specific situations like family conversations and peer group talk.

The socialization of infants is often regarded as primarily the duty and work of mothers; mothers use 'motherese,' a very specific conversational style adapted to the needs of infants; as a result of caretaking, children acquire the regional dialects and vernacular, which their mothers use at home. Other investigations illustrated that professional women are under more pressure than their male colleagues; therefore they adopt high standard sociolects or even hypercorrect features (see Kerswill, 2012). However, hypercorrectness is also clearly related to social class: Working-class and lower-middle-class men also tend to speak hypercorrectly in certain organizational contexts (Wodak, 1996). Very detailed case studies of code switching, moreover, have shown that the use of either standard or nonstandard styles as well as minority and majority languages depends highly on the context of the utterances as well as on social class (Duchene, 2008; Martin-Rojo and Esteban, 2003). Nonstandard lects can also be the manifestation of certain political opinions and attitudes as in the case of Ebonics (see Baugh, 1999), i.e., 'African-American vernacular.'

The same variation of correlations is true for studies on conversational styles: In some cultures, men clearly interrupt women more frequently and dominate topic choice, but in other cultures this need not be the case (see Sunderland, 2004). As a result, women's language has been said to reflect their conservatism, prestige consciousness, upward mobility, insecurity, deference, nurturance, emotional expressivity, connectedness, sensitivity to others, solidarity, and men's language is heard as evincing their toughness, lack of affect, competitiveness, independence, competence, hierarchy, and control (Eckert and McConnell-Ginet, 1992: p. 90). Of course, many differences are due to cultural variation as well which has to be accounted for. Many empirical studies have, however, neglected the context of language behavior and have often analyzed gender by merely looking at the speakers' biological sex (see the arguments in Wodak and Benke, 1996). By focusing only on the biological sex, the whole group of women is perceived as homogenous and contrasted to the whole group of men – also perceived as homogenous – without taking differences in these groups into account (see below).

Such a dichotomization clearly does not make sense when looking at many studies, which have shown significant differences that occur with women or with men due to culture, social class, ethnicity, etc. (Wodak and Schulz, 1986; Kotthoff and Wodak, 1997). Moreover, gender is related to and overlaps with the sociocultural and ethnic background of the interlocutors, with their age, their level of education, their socioeconomic status, their emotions, and the specific power dynamics of the discourse investigated. The focus nowadays lies on the discursive construction of identities in interaction where gender is but one part of this attempt (Kotthoff, 2010; Pavlidou, 2010; Wodak et al., 2009). Thus, research on gender should take the whole interaction and context into account; of course, the biological sex will always be part of a categorization but other categories might prove more valid in a specific research that makes it possible to operationalize gender roles (the inclusion of professional roles and organizational roles with gender is of utmost importance as well as culture and ethnicity; the socialization of women who come from Islamic cultures is

obviously different than the background of Western European women so that intercultural problems clearly arise in interactions, see Hellinger, 2010). All these results suggest that gender has an enormous impact in many relevant situations, like schools, hospitals, courtrooms, etc., but only in combination with other variables; this phenomenon is called *intersectionality*. However, it is usually not possible to try to decide which variable generally would have more impact. Only very detailed qualitative studies can provide insight into such complex theoretical questions (Wodak, 2008).

Important Theoretical Developments in the Field

Different models of *gender* have influenced theorizing and methodology over the years. The 'dominance' *and* 'difference' studies of the 1970s and 1980s saw gender as 'differential tendencies' between women and men, boys and girls. Whereas investigations following the 'deficit model' have interpreted the use of such indicators as manifestations of 'powerless language of women,' research in the paradigm of the 'difference model' views the linguistic behavior of men and women as originating in opposed modes of socialization and equal in their meaning and impact (Tannen, 1994). Hedging, tag questions, or indirectness can be viewed 'negatively' or 'positively,' either as signs of female insecurity or as supporting conversational work, depending on the context of the discourse and the theoretical approach adopted. This controversy led to many polemical debates, which finally illustrate the invalidity of such dichotomous distinctions. Thus, nowadays, other research questions have become more important, viewing gender as constructed due to *context*.

Social constructionism, however, sees gender as linguistic dealings *with* (individual, and groups of) women, men, boys, and girls, for example, how they are addressed, what is said *to* them and, more importantly what is said and written *about* (individual, and groups of) women, men, boys, and girls. The first understanding is arguably the 'popular' understanding of gender, the focus being on *who* (speaker sex) and *how*. The second two, however, de-emphasize gendered speakers (and writers) as agents, focusing rather on *what* is said (written) *to*, and especially *about*, women, men, boys, and girls (Pavlidou, 2010). Importantly, it was *poststructuralist thinking* more generally, associated with social constructionism that allowed gender to be seen as something *talked and written about* (as well as enacted in interaction), and thus as a *set of ideas* (rather than something embodied), hence an ideology, a set of values and norms, and simultaneously, frequently sexist stereotypes. Indeed, the gender differentiation *constructed* in talk or writing (i.e., the substance) may be more significant than any stylistic or interactional gender differences in language use (see Kotthoff, 2010). Most gender and language studies today encompass the last two meanings of *gender* together with the nuanced version of the first. It thus evidences poststructuralist characteristics, in that its concerns are (see Cameron, 2005): (1) gender diversity (e.g., class, ethnicity, and how these both interact with gender; multiple masculinities/femininities; differences among 'women' and 'men'); (2) gender being 'performed' (see below) in an ongoing way, allowing for *agency*; performance being achieved partly through language (which is therefore *constitutive*); similarly, power being 'done' rather than something speakers 'have'; (3) 'local' or 'contingent' explanations for gendered language patterns and the importance of specific contexts.

The notion of *performance* (Butler, 1990) originates from the field of philosophy. Judith Butler has been very influential in the gender and language field, and we now read about gender not only being constructed and performed through language, but also *accomplished*, *achieved*, *enacted*, and *effected* by language. 'Local' or in-context explanations for gendered language patterns have increasingly been framed in terms of *Communities of Practice* (Lave and Wenger, 1991). Communities of practice are defined as "(a)n aggregate of people who come together around mutual engagement in some common endeavor. Ways of doing things, ways of talking, beliefs, values, and power relations – in short, practice – emerge in the course of their joint activity around that endeavor" (Eckert and McConnell-Ginet, 1992: p. 95). Gender identity is produced, reproduced, and also contested through participation in such communities, in particular through engagement with gendered practices. As Litosseliti observes, the communities of practice notion allows us to "distinguish between speakers' assumed gendered behavior and the range of identities available in the gendered communities that speakers inhabit" (Litosseliti, 2006: p. 66; see Wodak, 2011 about communities of practice to which Members of the European Parliament (MEPs) belong). While researching gender, the emotions, attitudes, values, intergroup, and intragroup relationships have to be accounted for as well. Eckert and McConnell-Ginet studied peer groups of girls and boys and could illustrate why big differences on a phonological and conversational level occur between two groups due to all the above mentioned factors. Eckert and McConnell-Ginet argue very convincingly and provide many examples to support their view that survey studies are too general and their level of abstractness are too high because all the important sociological and psychological parameters cannot be integrated. Moreover, and this criticism applies to survey studies in general about other topics as well (e.g., research on anti-Semitism and racism, Wodak et al., 2009), questionnaires and opinion polls usually manifest self-assessments of the interviewees. A bias due to education must be considered as well: Better-educated interviewees usually know what kind of answers is expected from them (e.g., Wodak et al., 2009). They answer accordingly even though this might not be their opinion at all. This bias is very important in all studies related to stereotypes and prejudice research. Therefore, such quantitative studies must be complemented with qualitative case studies using authentic data tape-recorded in everyday interactions as well as ethnographic observations (see Titscher et al., 2000).

Sexism and the Language System

Analogous to the term 'racism,' the word 'sexism' was invented in the 1960s. It refers to discrimination within a social system on the basis of sexual membership. With the concept of 'sexism,' women defined themselves for the first time as a suppressed minority expressing basic political interests of being judged and valued in equal terms as men. Of course, the history

of women's fight for equal rights is much older (see Saurer, 2014) but the political movements in the 1960s, together with innovations like the invention of the contraceptive pill and birth control on a larger scale started a qualitative new time for women's emancipation. As such, they sought to reveal the mechanisms of suppression, making others aware of and fighting these devices. Social groups often define themselves by means of their common language, which plays an important role in identity creation and, for subcultures, serves as a means of differentiating themselves from the outside world. This specific identity manifests itself in certain conversational styles, manifestations of emotions, etc. (see Coates, 1997; Buchholtz et al., 1999).

Feminists conceive language as "a symbolical reflection of androcentric structures" (Günthner and Kotthoff, 1991: p. 7), as one of the means of patriarchal society to discriminate, disregard, and incapacitate women. In their view, language systems already reflect the patriarchal structure of Western societies. Such an assumption manifests a basic consideration of sociolinguistics, namely that language and social norms are connected and that language manifests and also constructs social practices (Wodak, 2008, 2011). If this claim is accepted – and such theories stem from Wittgenstein's 1967 concept of 'language games' – then the manifestation of women's domination must also be manifest in language.

The assumption that an individual language system has lexical elements and morphological and grammatical rules that are already sexist is based on the premise that "due to their long history as public decision makers, men not only determine the economic, political and social orientation of social life, but also influence the functioning and the semantic contents of each individual language" (Postl, 1991: p. 89). Once the language system has become the object of investigation for feminist linguists, it is interesting to note how the linguistic structure of an individual language is connected to the structure of society, how the structure of the language is conditioned by the structure of society, and vice versa. Unlike studies investigating divergent gender linguistic behavior, studies on the sexist use of language focus on the possibilities of reference to both genders or their practice that exist in an individual language. According to Pusch (1984: p. 7), the objects of 'feminist system linguistics' are 'patriarchalisms in diverse language systems.'

Feminist Linguistics

According to Hellinger (2010), feminist linguistics (FL) is distinguished from all other disciplines by the following three aspects, which similarly analyze the relationship between language and gender:

1. FL places female and male linguistic behavior and the linguistic phenomena connected with the designations of women and men at the center of its considerations.
2. FL interprets person-related asymmetries in the field of language systems and language use, as expressions of the linguistic discrimination of women (sexism) and links these directly to the plane of social discrimination. Traditional studies usually make do with descriptive results. Clearly, different languages necessarily differ in their linguistic realizations of such inequality: In German or Polish, for example, suffixes and certain prefixes manifest gender. Kargl et al. (1997) have submitted a whole register of possibilities of equal treatment of women and men in many everyday genres in communication, starting with names and labels up to whole reports in newspapers, as well as conversations. These attempts are valid in written as well as oral speech and have led to huge debates in the media (see Hellinger, 2010); in German, specifically, the use of capital 'I' in written texts was at the center of such political discussions (i.e., '*Lehrer-Innen*' vs '*Lehrer*' and '*Lehrerinnen*' or '*Lehrer/innen*').
3. FL does not accept phenomena as given, but seeks alternatives in keeping with the principle of the linguistic equal treatment of women and men. It pursues explicitly political goals by criticizing ruling linguistic norms and understanding the linguistic change it advocates as part of an overall change in society.

Sociolinguistic Studies of Sex/Gender

In the 1970s, 'sex' was established within sociolinguistic research as a social variable next to the already existing variables of social stratum, age, nationality, ethnic affiliation, religion, class, and region. In correlative-quantitative sociolinguistic investigations on the prestige and stigma variants of languages (see Kotthoff, 1992), the sex variable became a factor which significantly affected the use of language. The best known representatives of this quantitative-correlative approach in sociolinguistics, with its focus on urban groups of persons, are William Labov, Peter Trudgill, Lesley Milroy, and Paul Kerswill. The working methods within this sociolinguistic approach diverge considerably from one another in certain aspects (such as endorsing correlationist sociolinguistics or network analyses, and so forth) but in one respect they are identical: All of them operate quantitatively, that is, linguistic variables are defined that are realized in different forms in different varieties. Then, the different variants are correlated with sociological parameters (Kerswill, 2012).

Several other studies have taken different approaches and have used new and other methodologies. In particular, a lot of criticism related to the implicit chauvinistic ideologies in these first studies and also to the undifferentiated analysis of the two sexes, which totally neglected the range of variation in each of the genders under observation. In a study of the language of mothers and daughters in Vienna, for example (Wodak, 1984; Wodak and Schulz, 1986), it could already be proved that, apart from these sociological parameters, other variables, such as psychological ones, are also responsible for the variation concerning each gender. This study stressed in particular the variation that was found among the women and girls investigated, one that relativized the factor of 'sex.' It could be shown that mothers with a positive relationship to their daughters spoke differently than mothers with a conflicting relationship. The same is true for the daughters (30 mother–daughter pairs were interviewed and tested). These results illustrate that emotions and relationships as well as educational styles and the content/discontent of mothers are more important than social class in an isolated way. Moreover, this study proves

that the differences in one category of sex can be statistically significant due to all these parameters.

Such results and tendencies are of utmost importance for the – frequently controversial – debate on gender and sex as well as social class. First, it could be shown that significant differences exist in one category (either women or men). Second, many assumptions about the differences between working-class and middle-class socialization could be contradicted – educational styles do not vary so much by social class but more by the self-assessment of mothers and their attitudes to their own life and achievements. Third, it could be shown that no significant differences appear between mothers who work at home and mothers who have professional lives outside of their homes. Therefore, many claims that children need mothers who only stay at home are wrong; the satisfaction of women with their own lives is the most important factor that affects their relationship with others, in their families, or outside of their homes. Fourth, generalizations about the use of vernacular and dialect could also be contradicted: Even daughters from upper-class families used Viennese dialect to distinguish themselves from the phonological styles of their mothers and to be able to manifest their autonomy. This study illustrates the necessity of a very complex and interdisciplinary approach to gender research. Currently, the complex relationship between social class, gender, culture, communities of practice, and so forth are termed as *intersectionality* (Sunderland et al., 2013).

The Analysis of Conversation and Discourse

Unlike the above mentioned research field, FL – which examines the use of language – concerns itself with gender differences in conversation and discourse (oral as well as written). Differences are investigated in the following domains of language use: voice, pronunciation, intonation, choice of words, argumentation, lexicon, syntax, interactional and conversational behavior, as well as visual features and modes and nonverbal communication. The gender-induced differences in the use of language are not treated simply as divergent variants standing side by side. Due to a patriarchally organized society, the interpretation of the differences ascertained is of major social and political significance. Research on African communities and their gendering as well as the intersectionality have hugely broadened our perspective (Sunderland et al., 2013). Moreover, research on successful women (Wagner and Wodak, 2006) and female leadership (Baxter, 2006) are drawing our attention to women who 'seem to have made it'; that is, who have achieved high status positions in their professions. It is important to discuss which factors might be of relevance for such trajectories and what the context-specific costs and benefits might be.

Recent Theoretical Approaches in the Feminist Study of Discourse

The Construction of the 'Sexes'

In her arguments against the tenability of the categories of 'woman' and 'sex,' Judith Butler refers to postmodernist theories (Butler, 1990). In the manner of Foucault she ascribes normative power to the idea of gender identity and the attempt to describe it. By the mere act of defining a gender identity, many bodies, practices, and discourses are excluded or devalued, whereas the constructed character of this gender identity is simultaneously concealed.

Thus, in such feminist theories (Butler, 1990), the idea of two sexes is criticized as being a construction, a prediscursive factor taking the place of reality. Following this interpretation, 'sex' cannot be separated into a biologically inherent aspect and a socially acquired one. The category of 'sex' itself, in Butler's view, is a purely cultural product of discourse. Such an opinion relates well to Foucauldian theories of discourse. 'Sex,' therefore, cannot be viewed as a universal, suprahistorical and extrasocial entity, it is understood to be an "integral component of every form of life" (Butler, 1990: p. 23). Such a view relates well to the assumptions of context sensitivity discussed above. On the other hand, 'sex' possesses a material basis: In many relevant contexts in our societies, women and men are still evaluated, judged, promoted, or paid according mainly to their biological sex. These facts are not accounted for in Butler's approach. More recently, some authors have differentiated Butler's dogmatic position: It is a question of revealing "the reproduction mechanisms, networks, and institutional compulsions that ensure that constructions become persistent and resistant and appear timeless, immovable, and identical with themselves" (Hirschauer, 1992: p. 333). It is a matter of exposing the arbitrary construction of this binary opposition and hence also its mutability, and not of criticizing the binarily organized perception as unreal.

Doing Gender

Unlike a research approach that accepts sexual differences as an aggregation of qualities and deals with the qualitative behavioral tendencies of women and men, ethnomethodologically oriented studies produced a new focus of research: 'doing gender' (Kotthoff, 2010). Such an approach complements traditional theories, which viewed gender as an aggregation of static attributes and are concerned with investigating and displaying the peculiarities of women and interpreting them as 'gender-specific' or 'gender-typical' attributes so as to reveal the asymmetry of the difference between the sexes, to criticize it, and to make it politically visible. The use of a bundle of characteristics and attributes to define gender complicates or renders an impossible interactional approach since attributes are 'entities' and not processes. This concept entails that definitions of gender cannot change, transform, and are not influenced by, spontaneous interactions. A further problem raised by the concept of gender as a concept of attributes is the possibility of individualization; that is, the individual who has been seen to possess or not to possess certain attributes becomes the center of focus, and the level of the social system is neglected. Unlike this non-interactive approach, 'doing gender' regards membership of a sex not as a pool of attributes 'possessed' by a person, but as something a person 'does.' In this sense, membership of a gender constitutes a performative act and not a fact. Gender is continually realized in interactional form. Gender is created not only in the everyday activities that characterize 'doing gender,' but also in the asymmetry

of the relationship between the sexes, the dominance of the 'male' and its normativeness. Patriarchal inequality is produced and reproduced in every interaction. This concept of 'doing gender' stresses the creative potential and the embedding of gender-typical behavior in a social context.

Female MEPs, for example, have created very different ways of positioning themselves in interactions with male colleagues in the European Commission or in the European Parliament itself. Interviews with female MEPs (Wodak, 2003, 2004, 2011) illustrate that they are very conscious of their gender roles and that they tend to construct different ways of self-presentation in interactions: in particular, one MEP stated the following, in talking about her first experience as a rapporteur.

> when I – entered the parliament – *Orientation* (lines 1–3) on my first report it was about Leonardo I don't know if you know: ((laughing)) well – I said 'I'm going to speak to the commissioner' and – I – /I knew – he only speaks very bad French and my er my French was very bad as well. so I said 'I want to have interpretation' So – I went to the commissioner *Complicating Actions* (lines 4–14) with a very good int/int/interpreter and I/I/I/I talked more than an hour with him. because we talked the same about it and at the end he said – 'well: I have here the advice of my: civil servants but I – agree with you:
>
> and this and this and this all goes through – ' so you have to be: – er – I don't know/how do we call it in English in / in the Netherlands we say [brutal] so you have to: ((laughing)) be polite *Evaluation* (lines 16–20) but you have to – you:/you mustn't be/ you mustn't *sit* behind your -/your *desk* – because that doesn't help. ((laughing)) but then then you have the worse system that I tried several times *Coda* (22–31). …

In this example, which has been marked for basic narrative structure according to Labov and Waletzky's (1967) model, we see that this MEP's story is objectively about having a successful meeting with a Commissioner while acting as rapporteur on a report about Leonardo – an EU youth and education related program established in 1995 to provide financial support for professional development and job training. In lines 4–6, the complicating actions, she shows how she went to the commissioner with an interpreter, and because she and the commissioner had the same understanding of the issues involved, he was willing to support her, despite contrary advice by his 'civil servants' on the matters involved. The main point of the story, or evaluation, from the MEP's perspective, is to show that as an MEP, to get things done, you must be active and assertive, 'not sit behind your desk.' Thus, in this narrative, she positions herself as an MEP who is proactive and who will do what it takes, including arguing directly with Commissioners, to see that her voice is heard. She also orients to being a rapporteur (line 2), which carries some responsibility in a committee, and to being from the Netherlands (line 17), although this last identity is evoked only to characterize her style of work ('brutal' in Dutch, or 'assertive') (for further details of the analysis see Wodak, 2011). This narrative also manifests the complexity of the European Parliament, the interwovenness of languages, cultures, positions, political affiliations, and gender roles, as well as professional experiences.

Another of the interviewees is oriented to a particularly wide range of identities (left, woman, Swedish, mother, political outsider, etc.) during her interview. Most striking is the way in which she repeatedly positions herself as being an 'atypical MEP.' Here we see one such occasion.

> I figure here the most common – ah civil – job – for an MEP is er to be a lawyer. me myself *I'm far from that* the job I had doesn't even *exist* outside *Scandinavia*. so: it's a sort of a social teacher – so so I'm/I'm very in/an/a very special bird in this er: Interviewer mhm mhm so now you don't feel like you – fit into sort of a *typical* MEP er ME *no. no: no: I'm not. I'm left I'm a woman I'm Swedish* and I'm also everything – /everything's wrong (laughs).

In this example, the MEP contrasts herself with what she considers to be a typical profile for an MEP (lawyer by profession), emphasizing the degree to which she feels different (I'm far from that … I'm a very special bird … everything is 'wrong'). She also points out many of the identities that she associates with and that she perceives as marking her as different (social teacher, left, female, Swedish). This MEP is very much concerned with her self-presentation as being different, different than other female MEPs and also male MEPs. Moreover, she explains her success as a female MEP by being exotic and narrates different strategies which she applies in 'doing politics and gender.'

Even such short examples illustrate the diversity and complexity of gender roles and make clear why an interdisciplinary, qualitative, and context-sensitive approach is necessary to access the whole domain of 'gender and language.' To be able to make more general observations, however, much more cross-cultural comparable research is urgently needed.

See also: Discourse, Anthropology of; Gender Differences in Personality and Social Behavior; Gender: Anthropological Aspects; Language and Gender; Sociolinguistics.

Bibliography

Baugh, J., 1999. Out of the Mouth of Slaves. University of Texas Press, Austin, TX.
Baxter, J., 2006. Speaking Out: The Female Voice in Public Spaces. Palgrave, Basingstoke.
Buchholtz, M., Liang, A.C., Sutton, L.A. (Eds.), 1999. Reinventing Identities: The Gendered Self in Discourse. Oxford University Press, Oxford, UK.
Butler, J., 1990. Gender Trouble. Routledge, New York.
Cameron, D., 2005. Language, gender, and sexuality. Applied Linguistics 26/4, 482–502.
Coates, J., 1997. Women's friendships, women's talk. In: Wodak, R. (Ed.), Gender and Discourse. Sage, London, pp. 245–262.
Crawford, M., 1995. Talking Difference: On Gender and Language. Sage, London.
Duchene, A., 2008. Ideologies across Nations. De Gruyter, Berlin.
Eckert, P., McConnell-Ginet, S., 1992. Communications of practice: where language, gender, and power all live. In: Hall, K., Buchholtz, M., Moonwomon, B. (Eds.), Locating Power: Proceedings of the Second Berkeley Women and Language Conference. Berkeley Women and Language Group. University of California, Berkeley, CA, pp. 89–99.
Günthner, S., Kotthoff, H., 1991. Von fremden Stimmen. Zum Stand der Forschung. In: Kotthoff, H., Günthner, S. (Eds.), Von fremden Stimmen. Weibliches und männliches Sprechen im Kulturvergleich. Suhrkamp, Frankfurt, Germany.

Hall, K., Buchholtz, M. (Eds.), 1995. Gender Articulated. Language and the Socially Constructed Self. Routledge, London.
Harrington, K., Litosseliti, L., Sauntson, H., Sunderland, J. (Eds.), 2008. Gender and Language Research Methodologies. Palgrave, Basingstoke.
Hellinger, M., 2010. Guidelines for non-discriminatory language use. In: Wodak, R., Johnstone, B., Kerswill, P. (Eds.), The SAGE Handbook of Sociolinguistics. Sage, London, pp. 565–582.
Hellinger, M., 1990. Kontrastive Feministische Linguistik. Hueber, Ismaning, Germany.
Hirschauer, S. (Ed.), 1992. Die soziale Kkonstruktion der Transsexualitat. Suhrkamp, Frankfurt, Germany.
Holmes, J., Meyerhoff, M. (Eds.), 2003. Handbook of Language and Gender. OUP, Oxford.
Kargl, M., Wetschanow, K., Wodak, R., Perle, N., 1997. Kreatives Formulieren. Anleitungen Zum Geschlechtergerechten Sprachgebrauch, vol. 13. Schriftenreihe der Frauenministerin, Vienna.
Kerswill, P., 2012. Social factors behind language variation: class and ethnicity. In: Chambers, H., Trudgill, P., Shilling-Estes, N. (Eds.), Handbook of Variation and Change.
Kotthoff, H., 1992. Unruhe im Tabellenbild? Zur Interpretation weiblichen Sprechens in der Soziolinguistik. In: Günthner, S., Kotthoff, H. (Eds.), Die Geschlechter im Gespräch: Kommunikation in Institutionen. Metzler, Stuttgart, Germany, pp. 126–146.
Kotthoff, H., 2010. Sociolinguistic potentials of face-to-face interaction. In: Wodak, R., Johnstone, B., Kerswill, P. (Eds.), The SAGE Handbook of Sociolinguistics. Sage, London, pp. 315–329.
Kotthoff, H., Wodak, R. (Eds.), 1997. Communicating Gender in Context. Benjamins, Amsterdam.
Labov, W., 1966. The Social Stratification of English in New York City. Center for Applied Linguistics, New York.
Labov, W., Waletzky, J., 1967. Narrative analysis: oral versions of personal experience. In: Helm, J. (Ed.), Essays on the Verbal and Visual Arts. University of Washington Press, Seattle, pp. 12–45.
Lave, J., Wenger, E., 1991. Situated learning: legitimate peripheral participation. Cambridge University Press, Cambridge.
Lazar, M. (Ed.), 2005. Feminist Critical Discourse Analysis. Palgrave, Basingstoke.
Leodolter, R., 1975. Das Sprachverhalten von Angeklagten bei Gericht. Scriptor, Kronberg, Germany.
Litosseliti, L., 2006. Gender and Language: Theory and Practice. Arnold, London.
Martin-Rojo, L., Esteban, C., 2003. Discourse at work: when women take on the role of manager. In: Weiss, G., Wodak, R. (Eds.), CDA and Interdisciplinarity. Palgrave, Basingstoke, pp. 241–271.
Mills, S., 2004. Discourse: The New Critical Idiom. Routledge, London.
Milroy, J., Milroy, L., 1978. Belfast: change and variation in an urban vernacular. In: Trudgill, P. (Ed.), Sociolinguistic Patterns in British English. Arnold, London, pp. 19–36.
Pavlidou, T.-S., 2010. Gender and interaction. In: Wodak, R., Johnstone, B., Kerswill, P. (Eds.), The SAGE Handbook of Sociolinguistics. Sage, London, pp. 412–428.
Postl, G., 1991. Weibliches Sprechen. Suhrkamp, Frankfurt, Germany.
Pusch, L., 1984. Das Deutsche Als Männersprache. Suhrkamp, Frankfurt, Germany.
Saurer, E., 2014. Liebe und Arbeit: Geschlechterbeziehungen in Europa 19 und 20 Jahrhundert. Fischer–Taschenbuchverlag, Frankfurt am Main.
Sunderland, J., 2004. Gendered Discourses. Palgrave, Basingstoke.
Sunderland, J., Atanga, L, Ellece, S., Litosseliti, L. (Eds.), 2013. Gender and Language in Sub-Saharan Africa: Tradition, struggle and Change. John Benjamins, Amsterdam.
Talbot, M., 1998. Gender and Language: An Introduction. Polity Press, Cambridge.
Tannen, D., 1994. Gender and Discourse. OUP, Oxford.
Titscher, S., Meyer, M., Wodak, R., Vetter, E., 2000. Methods of Text and Discourse Analysis. Sage, London.
Trudgill, P., 1972. Sex, covert prestige, and linguistic change in the urban British English of Norwich. Language in Society 1, 179–196.
Trudgill, P., 1974. The Social Differentiation of English in Norwich. Cambridge University Press, Cambridge, UK.
Wagner, I., Wodak, R., 2006. Performing success: identifying strategies of self-presentation in women's biographical narratives. Discourse & Society 17 (3), 385–411.
Wilkinson, S., Kitzinger, C., 1995. Feminism and Discourse. Sage, London.
Wittgenstein, L., 1967. Philosophische Untesuchungen. Suhrkamp, Frankfurt, Germany.
Wodak, R., 1984. Hilflose Nähe? Mütter und Töchter erzählen. Bundesverlag, Vienna.
Wodak, R., 1996. Disorders of Discourse. Longman, London.
Wodak, R., 2003. Multiple identities: the role of female parliamentarians in the EU Parliament. In: Holmes, J., Meyerhoff, M. (Eds.), The Handbook of Language and Gender. Blackwell, Oxford, pp. 671–698.
Wodak, R., 2004. National and transnational identities: European and other identities oriented to in interviews with EU officials. In: Hermann, R., Risse, T., Brewer, M. (Eds.), Transnational Identities: Becoming European in the EU. Rowman & Littlefield, New York, NY, pp. 97–128.
Wodak, R., 2008. Controversial issues in feminist critical discourse analysis. In: Harrington, K., Litosseliti, L., Sauntson, H., Sunderland, J. (Eds.), Gender and Language Research Methodologies. Palgrave Macmillan, Basingstoke, pp. 193–210.
Wodak, R., 2011. The Discourse of Politics in Action: Politics as Usual. Palgrave, Basingstoke.
Wodak, R. (Ed.), 1997. Gender and Discourse. Sage, London.
Wodak, R., Benke, G., 1996. Gender as a sociolinguistic variable. In: Coulmas, F. (Ed.), The Handbook of Sociolinguistics. Blackwell, Oxford, UK.
Wodak, R., Johnstone, B., Kerswill, P. (Eds.), 2010. The SAGE Handbook of Sociolinguistics. Sage, London.
Wodak, R., de Cillia, R., Reisigl, M., Liebhart, K., 2009. The Discursive Construction of National Identity. Edinburgh University Press, Edinburgh, UK.
Wodak, R., Dressler, W., 1978. Phonological variation in colloquial Viennese. Michigan Working Papers in Germanic Studies 4 (1), 30–67.
Wodak, R., Pelikan, J., Nowak, P., Gruber, H., de Cillia, R., Mitten, R., 1990. 'Wir Sind Alle Unschuldige Taa Ter!' Diskurshistorische Studien Zum Nachkriegsantisemitismus. Suhrkamp, Frankfurt, Germany.
Wodak, R., Schultz, M., 1986. The Language of Love and Guilt. Benjamins, Amsterdam.

Gender and Physical Health

Gerdi Weidner, San Francisco State University, Tiburon, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G. Weidner, volume 9, pp. 5904–5907, © 2001, Elsevier Ltd.

Abstract

This article reviews factors that contribute to the gender gap in life expectancy, focusing on the leading cause of death, heart disease. A brief summary of gender differences in biological factors is followed by a presentation of behavioral and psychosocial contributions to women's greater longevity, and implications for prevention and treatment are discussed.

Gender Gap in Life Expectancy

Considerable gains in life expectancy have been observed for both sexes during the past decades. However, men still die about 5 years earlier than women (Murphy et al., 2013). This gender gap in life expectancy is largest when measured at birth, but shrinks with age. For example, if a man lives to 75 years of age in 2010, his life expectancy is only 1.9 years shorter than that of a woman's of the same age (Murphy et al., 2013). This shrinking gender gap with age is because men are more likely to die at younger ages than women. In the United States, the 15 leading causes of death in 2010 were (1) heart disease, (2) cancer, (3) chronic lower respiratory diseases, (4) stroke, (5) accidents (unintentional injuries), (6) Alzheimer's disease, (7) diabetes mellitus, (8) kidney disease, (9) influenza and pneumonia, (10) suicide, (11) septicemia, (12) chronic liver disease and cirrhosis, (13) hypertension, (14) Parkinson's disease, and (15) pneumonitis (inflammation of the lung). These 15 causes of mortality accounted for 80.4% of all deaths in 2010 (Murphy et al., 2013).

Almost all of these 15 leading causes of death show men to be at a greater risk than women. That is, the male-to-female mortality ratios (i.e., male age-adjusted death rate divided by the female age-adjusted death rate) ranged from 4.0 for suicide to 1.2 for septicemia. For heart disease, the major cause of death, the ratio was 1.5, and 1.4 for cancer, diabetes mellitus, kidney disease, and influenza/pneumonia. The only cause of death for which the ratio of male-to-female death rate was less than 1 was Alzheimer's disease (0.8), and women were as likely as men to die from stroke and hypertension (ratios = 1.0; Murphy et al., 2013). Several factors contribute to the gender gap in life expectancy. These can be grouped into three categories: biological, behavioral, and psychosocial (for more extensive coverage, see section on Gender in Whitfield et al., 2012).

Factors Contributing to the Gender Gap in Life Expectancy

Biological Factors

In her classic article, 'Why Do Women Live Longer Than Men?' Waldron concluded that "physiological differences have not been shown to make any substantial contribution to higher male death rates" (Waldron, 1976: p. 356). This conclusion has not changed much over the past decades (Gorman and Read, 2007; Waldron, 2005). Interestingly, international data on heart disease mortality from 46 communities in 24 countries have shown that male:female ratios vary widely, ranging from 10:1 in Iceland to 10:6 in Beijing, China (Jackson et al., 1997; also see Waldron, 2005). The fact that the differences between countries are larger than the difference between the sexes suggests that male anatomy is not destiny, at least in regard to heart disease. Additionally, the epidemic of cardiovascular diseases among young and middle-age Eastern European men has widened the gender gap in life expectancy over a very brief time span, which suggests that other than biological factors play a role in men's greater mortality risk (Weidner, 2010; Weidner et al., 2002; Weidner and Cain, 2003).

Behavioral Factors

Behavioral factors, such as smoking, excessive alcohol use, poor diet, and physical inactivity, are clearly involved in many of the major chronic diseases and causes of death worldwide (Lopez et al., 2006). For example, cigarette smoking has been linked to heart disease, lung cancer, chronic obstructive pulmonary disease, and pneumonia. Excessive alcohol consumption increases the risk for a number of diseases, foremost among them heart and liver disease. Excessive alcohol consumption, along with lack of seat belt use, also plays a major role in accidents and injuries in general. Other accidental deaths, such as homicide and suicide, often involve firearms. Overeating, unhealthy diets, and lack of exercise (associated with obesity) contribute to almost all chronic diseases, including diabetes, cancers, and strokes (Weidner, 2012).

An evaluation of lifestyle practices (i.e., smoking, physical inactivity, and unhealthy eating habits) in the INTERHEART study confirms their importance for myocardial infarction (MI), collectively accounting for more than 55% of the population attributable risk (PAR) in women and men (Anand et al., 2008), that is, the reduction in incidence that would be observed if the population were unexposed to these health-damaging lifestyle factors.

Generally, men are more likely to engage in health-damaging behaviors (e.g., smoking, unhealthy eating, excessive alcohol use) when compared to women. For example, women's diets have a lower ratio of saturated to polyunsaturated fat and higher vitamin C content than men's diets (Connor et al., 2002).

The only behavioral gender difference favoring men consistently appears to be physical activity. However, this finding may be due to the use of questionnaires designed for men, which focus on sports and neglect physical activities associated with housework (Barrett-Connor, 1997). Furthermore, stress may play a greater role for health-damaging behaviors among men than among women. For example, job strain appears to be associated with increases in health-damaging behaviors (e.g., cigarette smoking, excessive alcohol and coffee consumption, lack of exercise) among men, but not among women (Weidner et al., 1997; Weidner and Collins, 1993). Thus, considering the major behaviors involved in the leading causes of death, women appear to fare better than men.

Psychosocial Factors

Systematic assessment of psychosocial factors has not always been a part of large population studies of health outcomes. Most of what we know about psychosocial contributions to physical health conditions comes from studies of heart disease, the number one cause of death in the United States, accounting for 24.2% of total deaths in 2010 (Murphy et al., 2013). Psychosocial factors such as hostility/anger, depression/vital exhaustion, lack of social support, and work stress have prospectively been linked to premature mortality from all causes and heart disease mortality in both sexes (e.g., Holt-Lunstad et al., 2010; Kuper et al., 2002; Orth-Gomér et al., 2010; also see Depression, Pessimism, and Health in this section).

The importance of psychosocial factors for MI in both women and men has been most strongly illustrated by findings from the largest global study of MI, the INTERHEART study (Anand et al., 2008; Yusuf et al., 2004). This study compared over 15 152 cases with 14 820 sex- and age-matched controls in 52 countries on all continents. Most interestingly, in addition to eight commonly assessed risk factors of heart disease (e.g., hypertension, abnormal lipids, diabetes, smoking, exercise, moderate alcohol consumption), INTERHEART also included an assessment of eating a healthy diet and a psychosocial stress index. This index consisted of depression, stress at work or at home, financial stress, major life events, and low locus of control. Together, these risk factors were significantly associated with MI in both women and men and accounted for more than 90% of the PAR in both sexes. Interestingly, the lower burden of MI among women at younger ages (<60 years) was largely explained by their lower levels of risk factors: after statistical adjustment of these factors, the gender difference in the probability of MI before the age of 60 was reduced by 80% (Anand et al., 2008). The main contributors to the increased risk among the younger men were a high prevalence of lipid abnormalities and smoking, although men did appear to benefit from their greater (moderate) alcohol consumption. Examining gender differences in the PARs across all ages revealed several differences: higher PARs in women than in men were observed for hypertension, diabetes, abdominal obesity, physical inactivity, lack of moderate alcohol intake, and the psychosocial stress index. Men's PAR was significantly larger for smoking. In sum, findings from the INTERHEART study indicate that most premature MI is preventable and that interventions may benefit from targeting risk factors in a gender-specific fashion.

Gender-specific associations of personality attributes (hostility/mistrust), negative emotions (particularly depression), and social support/social isolation, as well as work stress, with health and well-being are discussed elsewhere (Barefoot et al., 1995; Leifheit-Limson et al., 2010; Orth-Gomér et al., 2010; Schwarzer and Rieckmann, 2002; Weidner and Spaderna, 2013). Generally, high levels of hostility/mistrust and depression and low levels of social support are related to harmful health outcomes in both sexes. Men consistently score higher on measures of hostility/mistrust and report less social support than women, which contributes to their increased health risk. Interestingly, both of these attributes are characteristics of the male ('macho') gender role, which has been linked to behavioral risk factors such as smoking, excessive alcohol consumption, and lack of seat belt use (Courtenay, 2009), as well as to decreased motivation to learn stress-management skills (Sieverding, 2002).

Additionally, women not only report more social support than men, but also have more sources of social support, thus decreasing their dependency on a single source. For example, studies of middle-aged people in Massachusetts found that men were more than twice as likely as women to name their spouse (or their partner) as their primary provider of social support (65.5 vs 26.4%). Furthermore, 24.2% of men (but only 6.1% of women) said this was their only source of support (New England Research Institutes, 1997). These data may, in part, explain why men's health is more seriously affected by partner loss through separation, divorce, or widowhood (Miller and Wortman, 2002).

In regard to negative emotions, gender differences appear to favor men. In most studies, women report more negative emotions such as depression than men (although this is not consistently found in populations where women and men have similar roles, such as college students; Nolen-Hoeksema and Girgus, 1994). Although women may report more depression, they may be coping more effectively than men. Generally, men are more likely to use avoidant-coping strategies, such as denial and distraction, whereas women are more likely to use vigilant-coping strategies, paying attention to the stressor and its psychological and somatic consequences (Weidner and Collins, 1993). Which style is more adaptive depends largely on the situation. Most stressful experiences consist of uncontrollable daily hassles, which are short-lived and typically of no great consequence. Here avoidant strategies would be more adaptive ("what I cannot control and what can't hurt me is best ignored"). Thus, men's strategies are likely to pay off for these types of events, contributing to their lesser experience (or report) of emotional discomfort or distress. But what if disaster hits? How do people cope with uncontrollable events requiring long-term adaptation, such as divorce, loss of a loved one, job loss, and sudden financial crisis and economic uncertainty? Here it may be women's greater vigilance that is more adaptive, such as preparing for the crisis, and seeking help and advice. Consistent with this reasoning are data from the Hungarian population that showed that women tended to accept their negative mood as a disorder to be treated, whereas

men were more likely to engage in self-destructive behavior, such as excessive alcohol consumption (Réthelyi et al., 2002).

Similarly, research on how people cope with disasters (e.g., hurricanes and tornadoes) has supported the notion of men's maladaptive coping: Increases in alcohol consumption and depression were related to personal disaster exposure among men, whereas no such direct relationship was evident among women (Solomon, 2002; Solomon et al., 1987). Furthermore, socioeconomic deprivation appears to be more closely related to depression in men than in women (Kopp et al., 2002). Thus, men's psychosocial risk factor profile appears to further enhance their risk for ill health.

Implications for Prevention and Treatment

Gender differences in behavioral and psychosocial lifestyle factors are likely contributors to the gender gap in several major causes of death. In regard to prevention, several large studies have now demonstrated significant health benefits to adopting a healthy lifestyle (e.g., a combination of a prudent diet, regular exercise, maintaining healthy body weight, and not smoking) for both women and men (Chiuve et al., 2006; King et al., 2007; Knoops et al., 2004). However, analyses of data from a community study of 15 708 middle-aged adults found that only 8.5% practiced healthy lifestyles (King et al., 2007). Although the percentage of women following healthy lifestyle practices was significantly larger than that of men, the low overall percentage is disconcerting. During a 4-year follow-up, 8.4% of participants newly adopted a healthy lifestyle (i.e., diets high in fruits and vegetables, regular exercise, maintaining a healthy body weight, and not smoking), and experienced a rather prompt benefit: mortality and cardiovascular disease risk were significantly reduced (40 and 35%, respectively) compared to participants with less healthy lifestyles (King et al., 2007). The authors also noted that men in general, African-Americans, and individuals with lower socioeconomic status were less likely to adopt a healthy lifestyle.

Although this study clearly demonstrates the benefits of adopting a healthy lifestyle, it also illustrates the problems associated with behavior change in primary prevention. Overall, efforts to reduce risk factors (via educational methods) in the general population have been disappointing. In their review of 39 trials, Ebrahim et al. (2006) conclude that "different approaches to behavior change are needed … the availability of healthier foods and better access to recreational and sporting facilities may have greater impact on dietary and exercise patterns … than health professional advice" (p. 2; also see Weidner and Kendel, 2010).

In contrast, attempts to change lifestyle behaviors in secondary and tertiary prevention trials (i.e., in people at high risk for disease and patients with the disease) have been more successful (Weidner and Kendel, 2010), but not without problems, especially in regard to overcoming barriers to participation. For example, the low utilization of cardiac rehabilitation by eligible patients has been attributed to several factors, including female gender (due to referral bias), medical insurance coverage, education, and ease of access to rehabilitation (Jackson et al., 2004).

In regard to psychosocial interventions, a meta-analysis of psychological treatment studies with heart disease patients concludes that psychological treatment is effective in reducing 2-year mortality by 27% (Linden et al., 2007). However, mortality was not reduced for women, for patients who perceived no reductions in stress, and for those whose treatment started right after the cardiac event (<2 months). Interestingly, the majority of women included in this meta-analysis were enrolled in the Enhancing Recovery in Coronary Heart Disease (ENRICHD) study. This study was a major multicenter, randomized clinical trial that tested the effects of a psychosocial intervention (aimed at decreasing depression and increasing social support) on reinfarction and mortality in 3000 post-MI patients at high psychosocial risk (i.e., depressed and/or socially isolated patients). Fifty percent of the patients were women. The study's results, with respect to the main outcome, were discouraging (Writing Committee for the ENRICHD Investigators, 2003). Considering that the intervention in ENRICHD started within 2 months after the cardiac event (a condition found to be ineffective in the later meta-analysis by Linden et al., 2007), the study's overall disappointing results may not be surprising. The unfortunate timing of the intervention may also explain the lack of benefit for women observed by Linden et al., as most of the women included in this meta-analysis were participants in ENRICHD.

Regardless of the reasons for the lack of benefit for women in Linden et al.'s meta-analysis, psychological treatments may have to be gender specific to yield beneficial outcomes for both sexes. The need for gender specific interventions may be most obvious for social support. For example, social support interventions seeking to elicit support from one's partner may be effective for men, who tend to see their spouse as their primary source of social support, but not for women, whose primary source of social support consists of friends and family members (New England Research Institutes, 1997). Consistent with this reasoning is the observation that single marital status and lack of spousal support predicted dropout from a comprehensive lifestyle change program among men with coronary artery disease (Koertge et al., 2003). Thus, soliciting social support from one's partner may be a promising intervention strategy for men but not for women, who could even experience exacerbated stress responses, as suggested by Kirschbaum et al.'s (1995) findings.

The need for gender-specific interventions also received support from a study conducted by Frasure-Smith and her colleagues. In their study aiming to reduce stress in male and female post-MI patients, significant gender differences were observed: direct advice about lifestyle changes seemed helpful for men, whereas listening to worries reduced stress in female patients (Cossette et al., 2001). Another investigation demonstrated a remarkable decrease in mortality among women with heart disease who were assigned to a stress reduction intervention specifically tailored to women (Orth-Gomér et al., 2009).

Finally, considering the fact that chronic diseases are influenced by a multitude of risk factors that are interrelated, implementing interventions aimed at the reduction of one factor alone may not be the most effective treatment strategy. As suggested by the findings from the INTERHEART study,

clinical, behavioral, and psychosocial risk factors contributed overwhelmingly to the risk of an initial acute MI. Most important, all identified risk factors are potentially modifiable. Thus, it would make sense to design interventions that target as many of these factors as possible. One intervention that targets multiple risk factors (diet, exercise, stress management, and social support) has been found to be beneficial for both women and men at various stages of heart disease and at different socioeconomic levels (for a review of this program of research, see Weidner and Kendel, 2010).

In sum, interventions targeting standard health risks (e.g., traditional medical risk factors such as hypertension, dyslipidemia), unhealthy behaviors (e.g., smoking, physical inactivity, unhealthy eating patterns), and psychosocial factors (e.g., social isolation, depression, maladaptive coping styles) are likely to benefit both women and men. However, given that there are many situational differences between men's and women's lives, a broader perspective on gender differences in health and well-being appears to be indicated (cf Courtenay, 2009). As a consequence, the design of gender-specific interventions may be required to yield effective outcomes.

See also: Behavioral Medicine; Cardiac Disease, Coping with; Coronary Heart Disease: Psychosocial Aspects; Depression, Pessimism, and Health; Gender and Health Care; Health Care Markets: Theory and Empirical Evidence; Health Insurance: Economic and Risk Aspects; Health Interventions, Community-based; Health-Related Support and Self-help Mutual Aid Groups; Men's Health; Public Health as Social Science; Social Epidemiology; Social Integration, Social Networks, and Health.

Bibliography

Anand, S.S., Islam, S., Rosengren, A., Franzosi, M.G., Steyn, K., Yusufali, A.H., Keltai, M., Diaz, R., Rangarajan, S., Yusuf, S., INTERHEART Investigators, 2008. Risk factors for myocardial infarction in women and men: insights from the INTERHEART study. European Heart Journal 29, 932–940.

Barefoot, J.C., Larsen, S., von der Lieth, L., Schroll, M., 1995. Hostility, incidence of acute myocardial infarction and mortality in a sample of older Danish men and women. American Journal of Epidemiology 142, 477–484.

Barrett-Connor, E., 1997. Sex differences in coronary heart disease: why are women so superior? The 1995 Ancel keys lecture. Circulation 95, 252–264.

Chiuve, S.E., McCullough, M.L., Sacks, F.M., Rimm, E.B., 2006. Healthy lifestyle factors in the primary prevention of coronary heart disease among men. Benefits among users and nonusers of lipid lowering and anti-hypertensive medications. Circulation 114, 160–167.

Connor, S.L., Ojeda, S., Sexton, G., Connor, W.E., Weidner, G., 2002. The role of pathogenic and protective dietary factors and non-traditional risk factors in the coronary epidemic of eastern and central Europe. In: Weidner, G., Kopp, M., Kristenson, M. (Eds.), Heart Disease: Environment, Stress and Gender, NATO Science Series, Series I: Life and Behavioural Sciences, vol. 327. IOS Press, Amsterdam, Netherlands, pp. 296–313.

Cossette, S., Frasure-Smith, N., Lespérance, F., 2001. Clinical implications of a reduction in psychological distress on cardiac prognosis in patients participating in patients participating in a psychosocial intervention program. Psychosomatic Medicine 63, 257–266.

Courtenay, W.H., 2009. Theorizing masculinity and men's health. In: Broom, A., Tovey, P. (Eds.), Men's Health: Body, Identity and Social Context. Wiley-Blackwell, London, UK, pp. 9–32.

Ebrahim, S., Beswick, A., Burke, M., Davey-Smith, G., 2006. Multiple risk factor interventions for primary prevention of coronary heart disease. Cochrane Database of Systematic Reviews 4. Art No. CD001561.

Gorman, B.K., Read, J.G., 2007. Why men die younger than women. Geriatrics and Aging 10 (3), 182–191.

Holt-Lunstad, J., Smith, T.B., Layton, J.B., 2010. Social relationships and mortality risk: a meta-analytic review. PLoS Medicine 7, e1000316.

Jackson, R., Chambless, L., Higgins, M., Kuulasmaa, K., Wijnberg, L., Williams, D., 1997. Gender differences in ischemic heart disease and risk factors in 46 communities: an ecologic analysis. Cardiovascular Risk Factors 7, 43–54.

Jackson, L., Leclerc, J., Erskine, Y., Linden, W., 2004. Getting the most out of cardiac rehabilitation: a review of referral and adherence predictors. Heart 91, 10–14.

King, D.E., Mainous, A.G., Geesey, M.E., 2007. Turning back the clock: adopting a healthy lifestyle in middle age. American Journal of Medicine 120, 598–603.

Kirschbaum, C., Klauer, T., Filipp, S.H., Hellhammer, D.H., 1995. Sex-specific effects of social support on cortisol and subjective responses to acute psychological stress. Psychosomatic Medicine 57, 23–31.

Knoops, K.T., de Groot, L.C., Kromhout, D., Perrin, A.E., Moreiras-Varela, O., Menotti, A., van Staveren, W.A., 2004. Mediterranean diet, lifestyle factors, and 10-year mortality in elderly European men and women: the HALE project. Journal of the American Medical Association 292, 1433–1439.

Koertge, J., Weidner, G., Elliott-Eller, M., Scherwitz, L., Merritt-Worden, T.A., Marlin, R., Lipsenthal, L., Guarneri, M., Finkel, R., Saunders Jr., D.,E., McCormac, P., Scheer, J.M., Collins, R.E., Ornish, D., 2003. Improvement in medical risk factors and quality of life in women and men with coronary artery disease in the multicenter lifestyle demonstration project. American Journal of Cardiology 91, 1316–1322.

Kopp, M., Skrabski, A., Székely, A., 2002. Risk factors and inequality in relation to morbidity and mortality in a changing society. In: Weidner, G., Kopp, M., Kristenson, M. (Eds.), Heart Disease: Environment, Stress and Gender, NATO Science Series, Series I: Life and Behavioural Sciences, vol. 327. IOS Press, Amsterdam, Netherlands, pp. 101–113.

Kuper, H., Marmot, M., Hemingway, H., 2002. Systematic review of prospective cohort studies of psychosocial factors in the etiology and prognosis of coronary heart disease. Seminars in Vascular Medicine 2, 267–314.

Leifheit-Limson, E.C., Reid, K.J., Kasl, S.V., Lin, H., Jones, P.G., Buchanan, D.M., Lichtman, J.H., 2010. The role of social support on health status and depressive symptoms after acute myocardial infarction: evidence for a stronger relationship among women. Circulation: Cardiovascular Quality and Outcomes 3, 143–150.

Linden, W., Phillips, M.J., Leclerc, J., 2007. Psychological treatment of cardiac patients: a meta-analysis. European Heart Journal 28, 2972–2984.

Lopez, A.D., Mathers, C.D., Ezzati, M., Jamison, D.T., Murray, C.J.L., 2006. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. Lancet 367, 1747–1757.

Miller, E., Wortman, C.B., 2002. Gender differences in morbidity and mortality following a major stressor: the case of conjugal bereavement. In: Weidner, G., Kopp, M., Kristenson, M. (Eds.), Heart Disease: Environment, Stress and Gender, NATO Science Series, Series I: Life and Behavioural Sciences, vol. 327. IOS Press, Amsterdam, Netherlands, pp. 251–266.

Murphy, S.L., Xu, J., Kochanek, K.D., 8 May 2013. Deaths: final data for 2010. National Vital Statistics Reports 61 (4), 1–117.

New England Research Institutes, 1997, Spring/Summer. Gender differences in social supports: data from the Massachusetts male aging study and the Massachusetts women's health study. Network, 12.

Nolen-Hoeksema, S., Girgus, J.S., 1994. The emergence of gender differences in depression during adolescence. Psychological Bulletin 115, 424–443.

Orth-Gomér, K., Schneiderman, N., Hui-Xin, W., Walldin, C., Blom, M., Jernberg, T., 2009. Stress reduction prolongs life in women with coronary disease: the Stockholm Women's Intervention Trial for Coronary Heart Disease (SWITCHD). Circulation: Cardiovascular Quality and Outcomes 2, 25–32.

Orth-Gomér, K., Weidner, G., Anderson, D.E., Chesney, M.A., 2010. Psychosocial influences on the heart. Epidemiology of cardiovascular diseases. In: Crawford, M.H., DiMarco, J.P., Paulus, W.J. (Eds.), Cardiology, third ed. Mosby, Edinburgh, UK, pp. 1819–1824.

Réthelyi, J.M., Purebl, G., Kopp, M., 2002. Sociodemographic and behavioral correlates of depression in Hungarian men and women. In: Weidner, G., Kopp, M., Kristenson, M. (Eds.), Heart Disease: Environment, Stress and Gender, NATO Science Series, Series I: Life and Behavioural Sciences, vol. 327. IOS Press, Amsterdam, Netherlands, pp. 114–120.

Schwarzer, R., Rieckmann, N., 2002. Social support, cardiovascular disease, and mortality. In: Weidner, G., Kopp, M., Kristenson, M. (Eds.), Heart Disease: Environment, Stress and Gender, NATO Science Series, Series I: Life and Behavioural Sciences, vol. 327. IOS Press, Amsterdam, Netherlands, pp. 185–194.

Sieverding, M., 2002. Gender and health-related attitudes: the role of a "Macho" self-concept. In: Weidner, G., Kopp, M., Kristenson, M. (Eds.), Heart Disease: Environment, Stress and Gender, NATO Science Series, Series I: Life and Behavioural Sciences, vol. 327. IOS Press, Amsterdam, Netherlands, pp. 237–250.

Solomon, S.D., 2002. Gender differences in response to disaster. In: Weidner, G., Kopp, M., Kristenson, M. (Eds.), Heart Disease: Environment, Stress and Gender, NATO Science Series, Series I: Life and Behavioural Sciences, vol. 327. IOS Press, Amsterdam, Netherlands, pp. 267–274.

Solomon, S.D., Smith, E.M., Robins, L.N., Fishbach, R.L., 1987. Social involvement as a mediator of disaster-induced stress. Journal of Applied Social Physiology 17, 1092–1112.

Waldron, I., 1976. Why do women live longer than men? Social Science & Medicine 10, 349–362.

Waldron, I., 2005. Gender differences in mortality – causes and variation in different societies. In: Conrad, P. (Ed.), The Sociology of Health and Illness – Critical Perspectives. Worth, New York, pp. 38–55.

Weidner, G., 2010. Why do men get more heart disease than women? An international perspective. Journal of the American College Health 48, 291–294.

Weidner, G., 2012. Sustainability in medicine – a case for the prevention of chronic non-communicable diseases. Environmentalist 32, 353–359.

Weidner, G., Boughal, T., Connor, S.L., Pieper, C., Mendell, N.R., 1997. The relationship of job strain to standard coronary risk factors in women and men of the family heart study. Health Psychology 16, 239–247.

Weidner, G., Cain, V., 2003. Gender gap in heart disease: lessons from Eastern Europe. American Journal of Public Health 93, 768–770.

Weidner, G., Collins, R.L., 1993. Gender, coping, and health. In: Krohne, H.W. (Ed.), Attention and Avoidance. Hogrefe & Huber, Seattle, WA, pp. 241–265.

Weidner, G., Kendel, F., 2010. Prevention of coronary heart disease. In: Suls, J.M., Davidson, K.W., Kaplan, R.M. (Eds.), Handbook of Health Psychology and Behavioral Medicine. Guilford Press, New York, pp. 354–369.

Weidner, G., Kopp, M., Kristenson, M., 2002. Heart Disease: Environment, Stress and Gender. In: NATO Science Series, Series I: Life and Behavioural Sciences, vol. 327. IOS Press, Amsterdam, Netherlands.

Weidner, G., Spaderna, H., 2013. Emotions and cardiovascular disease. In: Goldman, M.B., Troisi, R., Rexrode, K.M. (Eds.), Women and Health, second ed. Academic Press, San Diego, pp. 991–1002.

Whitfield, K.E., Weidner, G., Thorpe, R.J., Edwards, C.L., 2012. Cultural aspects of health psychology. In: Nezu, A.M., Nezu, C.M., Geller, P.A. (Eds.), Health Psychology, second ed. In: Weiner (Editor in Chief), Handbook of Psychology, vol. 9. I.B. Wiley & Sons, New York, pp. 538–563.

Writing Committee for the ENRICHD Investigators, 2003. Effects of treating depression and low perceived social support on clinical events after myocardial infarction: the Enhancing Recovery in Coronary Heart Disease Patients (ENRICHD) randomized trial. Journal of the American Medical Association 289, 3106–3116.

Yusuf, S., Hawken, S., Ounpuu, S., Dans, T., Dans, T., Avezum, A., Lanas, F., McQueen, M., Budaj, A., Pais, P., Varigos, J., Lisheng, L., INTERHEART study investigators, 2004. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet 364, 937–952.

Gender and Place

Daphne Spain, University of Virginia, Charlottesville, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The analysis of gender and place is interdisciplinary in that the field seeks to combine the spatial orientation of geographers, architects, and urban planners with the social focus of anthropologists, sociologists, and historians to understand the ways space and gender mutually reinforce each other. Places range from the geographic scale of the city or neighborhood to the building, and now to cyberspace. Gender identities include masculine, feminine, gay, lesbian, and an ambiguous 'third gender' of queer that recognizes gender identity as variable both historically and individually over the life course.

Introduction

'Gender and place' is a nuanced term for 'sex and space.' Both phrases refer to the ways in which gender or sex-role stereotypes are built into, and reinforced by, spatial arrangements. Take the adage that 'woman's place is in the home.' When ideology dictates that women should occupy only the domestic sphere, they seldom learn skills that would prepare them to enter offices or factories. Conversely, men who spend their days in corporate boardrooms or on the assembly line rarely are expected to excel at housework and childcare. Since single-use zoning laws typically separate homes from offices and factories, the association of gender with place is geographically inscribed on the city as well as attributed to individual buildings.

'Gender' represents socially constructed masculine and feminine (or some combination thereof) characteristics, while 'sex' refers to the biologically determined categories of male and female. One's sex is determined at conception, but one's gender identity develops over the life course and can fluctuate across a wide continuum of masculine and feminine characteristics. Girls who work on the family car in the garage and boys who cook meals in the kitchen are exhibiting gender characteristics incongruent with those expected of their sex; they are also occupying spaces inconsistent with gender expectations. A mismatch between space and gender is especially difficult for lesbian, gay, bisexual, and transgendered individuals to navigate (Doan, 2011).

Place is to space as gender is to sex: each pairing juxtaposes objective and subjective characteristics. Actual spaces can be designated with commonly understood boundaries of size or shape, but 'places' transcend simple attempts at measurement. Places are spaces imbued with significance. According to sociologist Gieryn (2000), space must meet three criteria to qualify as a place: A place is a unique geographical location, from a favorite chair to a region of the country; place has a physical, material form like streets, buildings, rocks, or trees; and place is invested with meaning and value. The space of a neigh borhood park may be a place to play for children, a place to sleep for the homeless, and a place to walk the dog for apartment dwellers. The park can be all of these places simultaneously or sequentially. The park may occupy a *space* the size of a city block, but it functions as a *place* for fun, shelter, or exercise depending on one's construction of its purposes. In short, space can be analyzed and measured; place is experienced. Cyberspace presents an enigma. It occupies an ambiguous position between the tangible characteristics of space and the subjective experience of place. Lacking physicality, chat rooms nevertheless create a sense of place, of being and belonging, for their participants.

Origins of the Field

The study of social structure and spatial structure emerged in the 1970s when geographer Harvey (1973) proposed that urban spaces and social relations are so intricately linked that a new language should be invented to capture the simultaneous occurrence of spatial form and social processes. During the 1970s, when feminist political activism was at its height, women were, quite literally, entering spaces previously open only to men. Colleges, factories, and offices became increasingly integrated by sex as women's place expanded beyond the home. Young women attending, and teaching at, universities during that era were among those who challenged the boundaries of traditional scholarship. The literal and symbolic collision of place and gender soon became a conscious subject of study in new interdisciplinary women's studies programs.

It is not surprising, then, that the first publications on gender and place were edited volumes generated from academic conferences about sex and space. Four of the most important volumes appeared almost simultaneously. *New Space for Women*, edited by sociologist Gerda Wekerle et al. (Westview Press, 1980) was a collection of papers presented at the 1976 United Nations Conference on Human Settlements that met in Vancouver. The editors organized a session addressing the role of women in the design and planning of urban settlements because they recognized that women's concerns about housing, transportation, health care, and the environment transcended categories of developed and developing nations. The next year, British anthropologist Shirley Ardener published *Women and Space: Ground Rules and Social Maps* (Croom Held London, 1981) as part of a series based on lectures sponsored by the Oxford University Women's Studies Committee. The edited volume included ethnographies of sex differences in the use of domestic and public spaces in the Andes, Greece, Iran, Russia, Nigeria, and South Africa.

In the United States, the Lincoln Institute of Land Policy sponsored a conference that resulted in the publication of *Building for Women* (Heath, 1981) edited by sociologist Suzanne Keller. It dealt with housing, community, and open space from a variety of academic disciplines; one of the chapters summarized the US Department of Housing and Urban Development's (HUD) Women and Mortgage Credit Project to facilitate women's ability to buy their own homes. HUD played a role in producing the fourth edited volume on women and space when it awarded a research grant to *SIGNS: Journal of Women in Culture* to commission essays for a special issue titled *Women and the American City*. The editor of *SIGNS*, Catherine Stimpson, subsequently published the volume (with Elsa Dixler, Martha Nelson, and Kathryn Yatrakis) through the University of Chicago Press in 1981. Contributors constituted the pantheon of researchers on women and space: Galen Cranz, Dolores Hayden, Jacqueline Leavitt, Susan Saegert, and Gerda Wekerle.

Donna Shalala, Secretary of Health and Human Services during President Clinton's administration, supported research on gender and place. Shalala was at HUD when the Mortgage Credit Project began and when *SIGNS* received funding for its special issue. As president of Hunter College, Shalala co-chaired the Ford Foundation–Hunter College Women and Housing Seminar with urban planning professor Eugenie Ladner Birch. Members of the seminar investigated the housing needs of single mothers, the elderly, and working women from the standpoint of design and affordability. The fruits of that pedagogical venture resulted in *The Unsheltered Woman: Women and Housing in the 1980s*, edited by Eugenie Ladner Birch (Center for Urban Policy Research, 1985), one of the first books to alert urban planners to the gendered nature of the housing market.

Urban planners Beth Moore Milroy and Caroline Andrews strengthened the tradition of edited volumes with *Life Spaces: Gender, Household, Employment* (UBC Press, 1988), a volume written entirely by Canadians about Canada. Milroy and Andrews introduced American audiences to geographer Suzanne Mackenzie, a founder of the field of feminist geography. Mackenzie, in turn, linked Canadian and American authors to British geographers Jo Foord, Sophie Bowlby, and Linda McDowell. *Life Spaces* departed from the model of previous edited volumes in one important respect: Wekerle's chapter on women's housing cooperatives included floor plans and photographs. Wekerle was one of the few social scientists to illustrate spatial concepts on the written page. During the first wave of scholarship, only Hayden and Birch (an architect and an urban planner, respectively) provided images of their topics. Today it seems obvious that books about space should include drawings and photographs, yet at the time authors rarely incorporated them.

Among pioneering scholars of gender and place, Dolores Hayden deserves special mention. Hayden's *Redesigning the American Dream: The Future of Housing, Work, and Family Life* (W.W. Norton, 1984) addressed the disparity between the massive single-family detached housing stock constructed after World War II and the decline of the nuclear family as the predominant household type. It opened debate among architects and planners about appropriate designs for new American families. Franck and Ahrentzen's edited volume on collective and single-parent housing, *New Households, New Housing* (Van Nostrand Reinhold, 1991) was a direct response to Hayden's challenge.

Differences in Scale: Cities, Neighborhoods, and Buildings

Cities, neighborhoods, and buildings reflect the gender assumptions of the era in which they are constructed. The post-World War II American metropolitan area took shape in the context of massive suburban expansion during a strong economy. Married-couple families were the norm, and middle-class white women could afford to stay home and raise the baby boom generation. The vast majority of architects, developers, and urban planners were men. Thus the suburbs were designed by men to appeal to mothers who seldom worked outside the home. Single-use zoning insured that residential neighborhoods were separated from commercial or industrial corridors, and public transportation was organized around men's commuting schedules. Only after divorce rates rose, fertility fell, and women entered the labor force did the gendered nature of cities reveal itself. Homes separated from employment and daycare opportunities, with little access to flexibly scheduled public transportation, became a liability for women.

Cities

Because almost all design and planning professionals were men at the turn of the twentieth century, the assumption is that women had little impact on the urban landscape. Yet there are historical precedents for both men and women building the city. In *Women and the City: Gender, Space, and Power in Boston, 1870–1940* (Oxford University Press, 2000), historian Sarah Deutsch argued that women reshaped the city's landscape to enhance their participation in public life. Working collectively through the Women's Educational and Industrial Union (WEIU), the Boston Women's Trade Union League, and the social settlement Denison House, women of diverse racial and social class backgrounds lobbied the municipal government to provide health clinics, vocational schools, and kindergartens. The WEIU located its headquarters at 264 Boylston Street, the city's legislative power corridor, to publicize women's accomplishments.

How Women Saved the City (Spain, University of Minnesota Press, 2001) explored similar themes. I examined four voluntary associations in which women played prominent roles at the turn of the twentieth century (The YWCA; The Salvation Army; the College Settlements Association; and the National Association of Colored Women) in Boston, Chicago, and New York to reveal the new spaces they created. Settlement houses, boarding homes, public baths, and institutions for unwed mothers were 'redemptive places' that saved the industrializing city from the onslaught of strangers streaming in from Europe and the rural South. While prestigious architects like Daniel Burnham were building monuments for the City Beautiful Movement, the less famous Vida Scudder was establishing settlement houses in New York and Boston to address problems in impoverished immigrant neighborhoods.

Cities and gender gained attention on the international front as well. In *A Women's Berlin: Building the Modern City* (University of Minnesota Press, 2008), architecture professor Despina Stratigakos examined the network of women's spaces in Wilhelmine Berlin from 1871 to 1918. The author took as her departure point a 1913 guide to the city written by and for women: 'What a Woman Must Know about Berlin.' The booklet introduced female residents and visitors to art schools that accepted women students; medical schools with gender-segregated anatomical theaters to avoid the risks of men and women viewing naked bodies together; a women's credit union; and women's clubhouses. The pamphlet exposed women to an entire range of places they could occupy outside the home.

In summer 2000, the International Women's University met at the University of Kassel (Germany) to discuss how the 'ideal' city would function and what it would look like. The conference attracted approximately 100 participants from different countries and cultures. They presented papers on the gendered aspects of transportation mobility, quality of life, public space, and urban design. This and other presentations were published in *City and Gender: International Discourse on Gender, Urbanism, and Architecture* (Terlinden, Leske and Budrich, 2003). A case study of a residential neighborhood in Haifa, Israel, for example, mapped women's interpretations of safe and unsafe places to walk. Their subjective assessments of the risks of crime affected how they navigated the city (Kallus, 2003; see also Fenster, 2005). Another book examining the way women and men experience urban space differently, *Cities and Gender* (Jarvis, Cloke, and Kantor, Routledge, 2009) illustrates the intersection of urban and gender studies in European, North American, and global south cities. Leslie Kern's *Sex and the Revitalized City* (UBC Press, 2010) applies a gendered lens to condominium development and urban citizenship in Toronto, Canada.

Neighborhoods

Studies at the neighborhood scale have produced a body of literature on gays' and lesbians' use of space. Manuel Castells' mapping of San Francisco's gay male culture was one of the first explorations into the relationship between homosexual subculture and space. Castells proposed that spatial concentration is fundamental to the development of gay identity. He argued that gay men are more likely to establish urban neighborhoods than lesbian women because men's tendency toward dominance extends to territorial aspirations, while lesbians' emphasis on relationships transcends space, making lesbians 'placeless' and more politically radical (Castells, 1983, p. 140; Castells and Murphy, 1982). Urban planners Sy Adler and Johanna Brenner (1992) challenged Castells' analysis, arguing that women's lower incomes, responsibility for children, and vulnerability to male violence had more to do with their location decisions than the bonds of lesbian relationships.

Gay men, and to some extent lesbian women, have played significant roles in the revitalization of American urban neighborhoods (Forsyth, 1997; Lauria and Knopp, 1985). As early as the 1920s, gay men led efforts to protect and restore New Orleans's French Quarter (Reed, 2012). Commercial areas catering to gays and lesbians, such as San Francisco's Market Street and Chicago's Halsted Street, display highly visible rainbow flags to attract clientele and, increasingly, tourists. Many cities now host Gay Pride parades. Even Walt Disney World, that bastion of family values, introduced Gay Days in 1991; by 2010 the event had expanded to all of Orlando and attracted more than 135 000 people. Gay and lesbian community centers, bookstores, and health clinics have become staples in many cities; gay and lesbian parades and festivals temporarily transform city streets. Being out in public often subjects gays and lesbians to harassment. Houston formed a community-based, police-affiliated Q-patrol to ensure that teens would not gather in key gay locations to cause trouble. Police departments in Chicago and San Diego recruited gay and lesbian officers. These efforts to make public spaces safer were initiated by queer activists (Forsyth, 2011). Making gays and lesbians welcome, in fact, has become a hallmark of Creative Cities (Florida, 2005).

Pressure is mounting to nominate prominent gay and lesbian sites as historic landmarks, like New York City's Stonewall Inn, where a police raid sparked the riot acknowledged as the catalyst for the gay rights movement (Dubrow, 1998). Such places have become easier to identify as gay and lesbian histories with a spatial focus have emerged. Chauncey (1994) has described how gay men appropriated public and private spaces to provide settings for the expression of gay culture in New York City between 1890 and 1940, while Kennedy and Davis (1993) have mapped the landscape of lesbian meeting places in Buffalo, New York, between the 1930s and 1960s. The National Trust for Historic Preservation lists many homes of prominent figures without including their sexual orientation in the narratives of place. Georgia O'Keeffe's property in Abiquiu, New Mexico; Philip Johnson's Glass House in New Canaan, Connecticut; and Willa Cather's childhood home in Red Cloud, Nebraska, are a few of the places gay and lesbian preservationists would like to see recognized (Dubrow, 2011).

Buildings

Homes, schools, offices, factories, and civic monuments are gendered places in at least three ways. One is in the principles underlying 'good' design; one is in the sex of the person who designs them; and another is in the connotations the building evokes.

Gendered principles of design originated with the Roman builder Vitruvius, who wrote *De Architectura* (*The 10 Books of Architecture)* around 25 BC. In Book III Vitruvius identified the male figure as the principal source of geometry for Classical architecture. Leonardo da Vinci's late fifteenth-century drawing, based on these ideal human proportions, became known as 'The Vitruvian Man.' It depicted a nude male figure in two superimposed positions, arms and legs outstretched, inscribed within a circle and a square. American adaptations of Vitruvian ideals are evident in Thomas Jefferson's neo-classical designs for Monticello and the University of Virginia, and in Daniel Burnhams City Beautiful monuments of the late nineteenth century. The highly influential mid-twentieth century French architect Le Corbusier worshipped geometric form to the extreme. American public housing residents have Corbusier to thank for the international style inflicted on them during

the 1950s. Would the history of architecture be different if there had been a 'Vitruvian Woman'? Impossible to know, of course, but many artists have proposed their female version of the famous diagram; a web search produces nearly 9000 images of 'Vitruvian Woman.'

Whether women and men design differently is a matter of debate. It was a moot point until recently, since for most of history architects have been men. In the United States, the 75 female architects registered by the American Institute of Architects and practicing in the first half of the twentieth century remain largely unrecognized. The exception may be Julia Morgan, who designed San Simeon castle for William Randolph Hearst during the Depression (Allaback, 2008). The question about gender differences in design has arisen as more women have joined the profession. Progress has been slow, however, in 2011 only 21% of practitioners were women (U.S. Bureau of Labor Statistics, 2012). Several of these, like Denise Scott Brown, are in husband–wife partnerships.

Does this small minority of women introduce feminine features into their buildings that their male colleagues omit? Do they conduct business differently from men? The stereotypes are that women tend to work in small firms, specialize in residential or human service projects, listen carefully to clients, and collaborate with their colleagues. In contrast, men tend to work in and for the corporate world, creating prize-winning signature buildings that express their rather sizable egos. Frank Gehry and Rem Koolhaas come to mind.

The most famous female architect today, Zaha Hadid, also has a distinctive style (and an ego to match any man's), as did Eileen Gray and Lucy Hillebrand, two other pioneering women architects. But their work has little in common, each exhibiting an 'exceptional peculiarity' that defies gender stereotyping (Dorhofer, 2003). The same could be said of Andrea Leers and Jane Weinzapfel, principals of Boston's Leers Weinzapfel, and Carol Wedge, president of Boston's Shepley Bullfinch (Landsmark et al.,, 2010). Chicago architect Jeanne Gang, however, has imbued her 82-storey apartment building, Aqua, with feminine touches in the very heart of the city that invented the skyscraper. Architectural critic Paul Goldberger (2010) credits Gang with figuring out how to give "brawny materials" "soft, silky lines, like draped fabric." She achieved the architectural and engineering feat by wrapping a rectangular glass slab with curving concrete balconies, turning the façade into an 'undulating landscape' that shades apartments from summer sun and dissipates high winds. Goldberger attributes Gang's success to her attempts to see the building from the perspective of its users rather than through a strict ideological lens of modernism.

Buildings also invite associations with masculine and feminine characteristics. Grand civic monuments are meant to convey the power of the state, represented by the men who traditionally occupied their corridors. Less imposing 'human-scale' buildings, on the other hand, are typically more comfortable places for women. Then there is the skyscraper. Noting the phallic connotation of skyscrapers is a cheap shot, since they are also the most cost-effective way to create commercial and residential space on expensive urban land. But business tycoon Donald Trump's insistence that his buildings are the tallest is a silly reminder of the analogy.

Some buildings designed specifically for women or men lack any exterior evidence of their identity. Architect Lupkin (2010) argues that the typical nineteenth-century Main Street YMCA, a brick building standing only three or four stories tall, was instrumental in redefining masculinity at the turn of the twentieth century. The YMCA's leisure spaces, its dorms, and its educational facilities were key elements in a spatially segmented system of class, race, and gender that helped define a 'culture of corporate capitalism' in the modern city (p. xvi). The YWCA, constructed specifically for single women, was equally as innocuous as the YMCA, and equally dedicated to defining proper gendered behavior in the industrializing city (Mjagkij and Spratt, 1997). Like women's and men's clubs, YWCAs and YMCAs were sites of gender formation for Americans entering the modern age.

A significant British book addressing differences in scale was Moore's *Space, Text, and Gender: An Anthropological Study of the Marakwet of Kenya* (Cambridge University Press, 1986). Moore drew on the theories of Foucault (1972) to analyze household space as a 'cultural text' representing gender relations in Marakwet society. The study placed the Marakwet in their geographic context with a map of Kenya and a diagram of their Sibou village, and in their building context with photographs of their dwellings. Moore's work progressed from a base of great geographical expanse to the smaller scale of the building. *Space, Text, and Gender* marked another important milestone. It was one of the first full-length ethnographies devoted to the topic of gender and place. Although edited volumes would still fill a niche, the field acquired stronger intellectual credentials with the publication of books posing theoretical arguments. My own work sought to extend that tradition. *Gendered Spaces* (Spain, University of North Carolina Press, 1992) was a cross-cultural and historical study proposing that women's status is lower in societies in which men and women are separated spatially in homes, schools, and workplaces because spatial segregation prevents women from gaining the 'masculine' knowledge most highly valued by society.

The Future of Gender and Place

The original scholarship on gender and place equated gender with women, as illustrated by titles like *New Space for Women* and *Women and the American City*. Two developments since the 1980s have enlarged the focus of research. First, contemporary work addresses men and masculinities in addition to women. Second, just as current work has broadened the concept of gender, future work promises to extend concept of place into cyberspace.

Sociologist Michael Kimmel was among the first scholars to introduce males into the conversation about gender. From *Manhood in America: A Cultural History* (Free Press, 1996) to *Guyland: The Perilous World Where Boys Become Men* (Harper, 2008), Kimmel has insisted that discussions of gender are incomplete as long as the focus is on women only. In *Guyland* he argues that it takes time for boys to learn how to become adults, and that the process is facilitated by "a place, or rather a bunch of places where guys gather to be guys with each other" safe from the scrutiny of parents or girlfriends (p. 4). Sports bars, TV or game rooms (aka 'man-caves'), dorms, and

fraternity houses are all part of 'Guyland,' at least for the middle class. Lower-income adolescents might learn how to be men in gangs that defend their turf. Like the nineteenth-century YMCA, these contemporary places are sites where youth can try out new identities in a safe space.

In 1998 Kimmel founded the journal *Men and Masculinities*. The first issue published articles titled "Masculinities and globalization" and "What is problematic about masculinities?" The peer-reviewed interdisciplinary journal was established to 'explore the evolving roles and perceptions of men across society' with empirical and theoretical scholarship from the perspectives of feminism, queer theory, and multiculturalism. A 2010 special issue was devoted to men in women's jobs, specifically men performing domestic service (Sarti and Scrinzi, 2010). It included articles on the masculine aspects of international migration and the globalization of care, topics typically covered from the women's standpoint (see Ehrenreich and Hochschild, 2002). The June 2013 volume included a review of the journalist Joel Stein's book *Man Made: A Stupid Quest for Masculinity* (Grand Central, 2012), about his search for ways to act manly when he became the father of a son. One might think of it as '*Guyland* lite' for adults. Then there's the *Journal of Men, Masculinities, and Spirituality*. This list of sources with masculinity in the title illustrates that after more than three decades of equating gender with women, the field of gender studies has become more inclusive. (Or, as some feminists might complain, men have found yet another way to become the center of attention.)

What does the future hold for the study of gender and place? It will definitely involve cyberspace, as the literature already suggests (Boyer, 1999; Light, 1995). For its female participants, the *World of Warcraft (WoW)*, a massive multiplayer online role-playing game, is a space connected to the real world; women take their off-line gender identities and social contexts with them when they go into the virtual world (Eklund, 2011). As the matrix of cyberspace supplements the grid of the city, women may experience the environment differently from men. Possibilities for gender ambiguities abound on the Internet in 'chat rooms' that allow more anonymity than face-to-face interactions. The digital landscape has the potential to change women's status, if it changes the way knowledge is disseminated. The gendered spaces that once limited women's access to knowledge, and thus power, may dissolve. For example, Muslim women living in the Western world use the Internet as a discursive space in 'fatwa sites,' where they raise questions about which rules govern gender behavior in a non-Muslim culture. Lacking a neighborhood mosque, women can consult an online imam to clarify Sharia law regarding dress code, rights to abortion, and mixing with the opposite sex. These fatwa sites have the potential to change Muslim family law (Ali, 2010).

Conclusion

The study of 'Gender and place' emerged during the 1980s to identify ways in which gender stereotypes and spatial arrangements reinforce each other. Over its brief history, the topic has become increasingly differentiated with respect to definitions of both place and gender. Cities, neighborhoods, buildings, and even cyberspace are spaces that become 'places' to the people who produce and use them. 'Gender' extends beyond the dichotomy of masculine and feminine to encompass lesbians, gays, bisexual, transgendered, and queer identities.

See also: Gender History; Gender and Feminist Studies in Geography; Gender and Feminist Studies in Geography; Gender, Economics of; Gender, Feminism, and Sexuality in Archaeological Studies; Lesbians: Social and Economic Situation; Property Rights and Urban Planning; Spatial Analysis in Geography; Spatial Thinking in the Social Sciences; Suburbanization and Suburbanism; Urban Geography; Urban Internal Spatial Structure.

Bibliography

Ali, A., 2010. Cyberspace as emerging muslim discursive space? Online fatawa on women and gender relations and its impact on muslim family law norms. International Journal of Law, Policy, and the Family 24, 338–360.

Allaback, S., 2008. The First American Women Architects. University of Illinois Press, Urbana.

Boyer, C., 1999. Crossing cybercities: urban regions and the cyberspace matrix. In: Beauregard, R., Body-Gendrot, S. (Eds.), The Urban Moment: Cosmopolitan Essays on the Late-20th-century City. Sage, Thousand Oaks, CA.

Castells, M., 1983. The City and the Grassroots. University of California Press, Berkeley, CA.

Castells, M., Murphy, K., 1982. Cultural identity and urban structure: the spatial organization of San Francisco's gay community. In: Fainstein, N., Fainstein, S. (Eds.), Urban Policy under Capitalism. Sage, Beverly Hills, CA.

Chauncey, G., 1994. Gay New York: Gender, Urban Culture, and the Making of the Gay Male World, 1890–1940. Basic Books, New York.

Doan, P. (Ed.), 2011. Queerying Planning: Challenging Heteronormative Assumptions and Reframing Planning Practice. Ashgate, Surrey, England.

Dorhofer, K., 2003. Symbols of gender in architecture and design. In: Terlinden (Ed.), City and Gender: International Discourse on Gender, Urbanism, and Architecture. Leske + Budrich, Opladen, Germany.

Dubrow, G., 1998. Blazing trails with pink triangles and rainbow flags: new directions in the preservation and interpretation of gay and lesbian heritage. Historic Preservation Forum 12, 31–44.

Dubrow, G., 2011. Lavender landmarks revisited: advancing an LGBT preservation agenda. In: Doan, P. (Ed.), Queerying Planning: Challenging Heteronormative Assumptions and Reframing Planning Practice. Ashgate, Surrey, England.

Ehrenreich, B., Hochschild, A., 2002. Global Woman: Nannies, Maids, and Sex Workers in the New Economy. Henry Holt, New York.

Eklund, L., 2011. Doing gender in cyberspace: the performance of gender by female world of warcraft players. Convergence: The International Journal of Research into New Media Technologies 17, 323–342.

Fenster, T., 2005. The right to the gendered city: different formations of belonging in everyday life. Journal of Gender Studies 14, 217–231.

Florida, R., 2005. Cities and the Creative Class. Routledge, New York.

Forsyth, A., 1997. 'Out' in the valley. International Journal of Urban and Regional Research 21, 36–60.

Forsyth, A., 2011. Queerying planning practice: understanding non-conformist populations. In: Doan, P. (Ed.), Queerying Planning: Challenging Heteronormative Assumptions and Reframing Planning Practice. Ashgate, Surrey, England.

Foucault, M., 1972. The Archeology of Knowledge. (A. Sheridan, Trans.). Tavistock, London.

Gieryn, T., 2000. A space for place in sociology. Annual Review of Sociology 26, 463–496.

Goldberger, P., 1 February 2010. Wave Effect: Jeanne Gang and Architecture's Anti-divas. The New Yorker.

Harvey, D., 1973. Social Justice and the City. Johns Hopkins University Press, Baltimore, MD.

Kallus, R., 2003. Gender reading of the urban space. In: Terlinden (Ed.), City and Gender: International Discourse on Gender, Urbanism, and Architecture. Leske + Budrich, Opladen, Germany.

Kennedy, E., Davis, M., 1993. Boots of Leather, Slippers of Gold: the History of a Lesbian Community. Penguin, New York.

Landsmark, T., Schwennsen, K., Van Lengen, K., 2010. Women in architecture. Chronicle of Higher Education 56, B10–B11.

Lauria, M., Knopp, L., 1985. Toward an analysis of the role of gay communities in the urban renaissance. Urban Geography 6, 152–169.

Light, J.S., 1995. The digital landscape: new space for women. Gender, Place and Culture 2, 133–146.

Lupkin, P., 2010. Manhood Factories: YMCA Architecture and the Making of the Modern Urban Culture. University of Minnesota Press, Minneapolis.

Mjagkij, N., Spratt, M. (Eds.), 1997. Men and Women Adrift: the YMCA and the YWCA in the City. New York University Press, New York.

Reed, J.S., 2012. Dixie Bohemia: a French Quarter Circle in the 1920s. Louisiana State University Press, Baton Rouge.

Sarti, R., Scrinzi, F., 2010. Introduction to the special issue: men in a woman's job, male domestic workers, international migration, and the globalization of care. Men and Masculinities 13, 4–15.

U.S. Bureau of Labor Statistics, 2012. Women in the labor force: a data book. www.bls.gov/cps/wlf-databook-2012.pdf (accessed 24.05.13.).

Gender and Social Research

Fatima A Castillo, University of the Philippines Manila, Quezon City, Philippines

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Globally, in the past two decades, there has been a remarkable expansion in social research on gender. But gaps and obstacles remain, including contentious issues with regard to ethics of research. This article provides a general description of the current situation of social research on gender – particularly the participation of women researchers, a discussion of methodological developments in gender research, moral and ethical issues on scientific validity, representation, and exclusion of women in clinical research.

Introduction

Indicative of the expansion in social research on gender since the 1990s is the proliferation of centers and programs for gender studies in academia, development agencies, and government and nongovernment organizations. For example, there are now more than 900 gender/women/feminist studies centers worldwide that have websites (Women's Studies Programs, Departments, & Research Centers). Many of these focus on women studies.

Public and private international funding institutions, government and semi-government bodies as well as the UN have set up gender programs aimed to encourage and support research on gender and women.

Gender studies tend to be interdisciplinary where researchers from diverse disciplines and fields bring their expertise together. Additionally, intersectoral and international collaboration among academic, civil society, and government researchers is becoming a trend.

Women Researchers

There have been improvements in the participation of women in research, but significant gaps remain. In most OECD member-countries (OECD), for example, the percentage of women among the total number of researchers in science and technology (S&T) has increased between 2004 and 2009, but no country registered a parity. The lowest participation is in Japan, at 13.6%, followed by Korea at 15.8%.

Globally, it is in S&T where the gap is alarming (WISAT). The situation is far better in Latin America and the Caribbean where women researchers are 46% of all the S&T researchers (ECLAC, 2013). However, senior posts are dominated by men. One reason is women scientists have to choose between career and motherhood. *The history of women scientists has been one of moving from exclusion to segregation* (ECLAC Report of October 2013 quoted in Elena Hurtado, 27 November 2013).

In psychology, the social sciences, and life sciences, there has been an upsurge in the number of women who earned doctoral degrees (Fox, 2004). The rapid expansion in the ranks of women in academia prompted Long to say that women have moved from *scarcity to visibility* (quoted in Ginther and Kahn, 2012: p. 2). But it is still an uphill climb for women. Only 30% of tenured faculty of US academics are women. The gender tenure gap is largest in economics and the social sciences. Worldwide, women constitute on average only 29% of all researchers (UNESCO, 2011, report cited in UN ECLAC report, October 2013). Again, having children is the most influential factor that causes the gender gap (Ginther and Kahn, 2012).

A researcher's productivity and quality of work in international academic culture are weighed in terms of publications, especially the citation counts for the author and the publication which have "historically been viewed as a relatively objective measure of the quality and impact of research" (Maliniak et al., 2013: 1). Using a gender lens, Maliniak et al. did the first study of its kind in examining patterns in citation and peer reviewed publications in international relations from 1980 to 2007. After controlling for a large number of variables, they found that women authors are *systematically less cited than men*. The authors think that this is because women tend to cite themselves less than men, and men tend to cite men authors more than women authors.

This is corroborated in a global review of published authors in international relations by Lariviere et al. (2013) where they found persistent gender gaps. In countries with the highest number of published researchers, women as the main authors were less cited than men in the same position. The gender disparity is being sustained by a system of using citations in the evaluation of researchers, and much more so for those in the social sciences, where there are more women authors than in the other disciplines.

The iniquitous system not only works against women but also against researchers in less developed economies. "Research evaluation that use quantitative measurements such as bibliometrics pose a severe limitation for countries where only a few researchers get published in international peer reviewed journals" (UNESCO, 2010: p. 5).

Shifts in Usage of 'Gender'

The term 'gender' is now commonly used including in the mass media. But there have been shifts in its usage such as exchanging sex and gender, exemplified in the US Food and Drug Administration's (FDA) use of gender (Miller, 2001). It is

now common to see gender instead of sex in survey questionnaires (i.e., for data on respondent's personal profile, similar to age). Used in the FDA and survey contexts, gender is used when the data being collected is on the biological characteristic, male or female. The socio-political meaning of gender is absent.

This is not the way feminists who brought the concept to public consciousness and political discourse mean in their use of gender. Gender is a social construct that attributes particular characteristics, abilities, roles, privileges, and entitlements to males and females. This attribution is manifested in every facet of social, economic, and political life.

As an analytic concept, it is the lens with which power, its location, and its distribution in various levels and dimensions of social life are examined including in relations in the research setting (Smith, 1987). But gender has become a "taken-for-granted buzzword" (Cornwall quoted in Sabea, 2008: p. 16) – the "workhorse concept" (Olesen, 2011: p. 135). The loose usage of gender for sex could erode the political import of the terminology. "The question remains how to refuse being incorporated, co-opted, or rendered superfluous and nonthreatening" (Sabea, 2008: p. 22).

Focus on Women

There have been criticisms on the focus on women and the marginalization of men in gender work (Proudlock, n.d.). An extreme example not in research but in medical practice is the controversial prohibition issued by the American Board of Obstetrics and Gynecology in September this year for members against treatment of men. Larry Gilstrap, a high official of the Board said: "if you want to voluntarily be known as a specialist in obstetrics and gynecology, it is restricted to taking care of women … There are plenty of physicians out there to take care of men." (Grady, Denise, 23 December 2013. Gynecology's gender question. The New York Times. From: http://www.nytimes.com/2013/12/24/health/gynecologys-gender-question.html?smid=tw-nytimes&_r=1& (accessed 23.12.13.)). A forceful argument was made by bell hooks (1984), considered a major thinker for radical feminism, that while men benefit from patriarchy more than women, it also oppresses them. "… I believe that if they (men) knew more about feminism they would no longer fear it, for they would find in feminist movement the hope of their own release from the bondage of patriarchy" (p. 1X).

An experience of a women's rights group in Asia shows that excluding men resulted in violence against the women who participated in the projects perpetrated by dominant males in the family (Estandarte and Alvarez Castillo, 1999).

Women rather than men are the central focus in gender research for three main reasons: (1) women bear the greater burden of oppression and inequality (Lengermann and Niebrugg-Brantley, 2003); (2) women's real life situation has been largely ignored in mainstream research (Lather, 1988); and (3) men and their life situation cannot represent women and their life situation (Mies, 1983).

Gender Bias in Research

Gender bias in research is documented in the literature (Ritchie, 2009). Common examples of gender bias in research are exclusion of women of reproductive age in clinical trials (discrimination – this will be discussed later in this article), using the household as unit of poverty measurement in economic research (Pogge and Rippin, 2013) (gender blindness), and women-only reproductive health studies (stereotyping).

Although not as pervasive and obvious as before, gender bias is still present, unrecognized, in various forms and degrees (Ritchie, 2009). Masculine characteristics are privileged over feminine characteristics. Contemporary social science research inherited this fundamental bias. This is why knowledge and knowing in the social sciences for a long time has been male oriented (Reinharz, 1992; cf Americas, Sociocultural Overviews: Caribbean).

There can be gender bias in every stage and process of social research, from the determination of the research purpose to the dissemination of research findings. It can be found in relationships among researchers in a project or in policies of funding institutions. Even in research ethics guidelines, in the ethics review system, there could be gender bias.

Research Gaps

Very important and wide-ranging findings about gender are now available, many of which have informed international and local policies and programs. There is so much that is now known and understood about the nature, dynamics, and consequences of, for example, gender bias, discrimination, and inequity, on women, families, communities, and societies in various contexts and historical periods. This being so, it is at this juncture that effects of the reforms on structures or systems, particularly on power shifts in socio-political life, should be examined. For example, there is no sufficient knowledge of the impact of policy reforms on the real life situation of women and girls. Susan Franceschet (2003) asked whether or not 'state feminism' has been good for women. There are questions whether invited participation (women invited by the state) is better than demanded participation (women's participation is fruit of their insistence) in forging ahead the struggle for justice for women (Murthy et al., 2005). On the other hand, there is scant attention to the impact on research on gender of the behavior and changing agenda of donors. Deep and comprehensive analysis beyond the statistical data on performance vis-à-vis the UN gender gap indicators and the Millineum Development Goals is needed.

With this background, among the issues that have not been given attention relative to the other issues are as follows:

- effects of global and state reforms on women's situations in various levels and contexts;
- structural or systemic effects of increase in the number of women and girls in education, in the domestic and migrant labor force;
- the effects of globalization of neoliberal agendas on women and girls;

- the interconnections of the women's rights movements and the larger human rights movements;
- the specific forms and character of oppression due to the intersection of social class, ethnicity or race, and gender;
- political power of, for, and by women; and
- power, dynamics, and effects of donors on research and the women's struggle.

Paradigm Shift, Methodological Revolution

There have been a paradigmatic shifting (Denzin and Lincoln, 2011) and a methodological revolution (Fonnow and Cook, 1991) in the social sciences since the 1970s. Perspectives long held to be universally correct about the nature of knowledge and the best ways of producing it were challenged while radically different approaches and innovative research methodologies are put forward (cf Feminism in Anthropology; Objectivity of Research: Ethical Aspects).

The contributions of feminist scholars are central in the paradigmatic turn and methodological revolution (Ramazanoglu, 1992) (cf Feminism in Anthropology). Feminist perspectives and reflections on their research experience constitute a distinct body of knowledge in the social sciences today.

The twentieth century women's rights movements provided impetus for contemporary feminist epistemology and methodology. Ann Oakley (1974), for example, whose sociological classic, Housewife said in an interview that the emerging feminist movement in the late 1960s was a major influence in her choice of the subject matter, housework.

Feminist Methodology

Feminist perspectives on the nature of knowledge and methods for knowing have influenced researchers worldwide. The methodology evolves as researchers try to find ways for knowing, describing, and making sense of the complexities and realities in women's lives usually hidden or unrecognized (Mama, 2013).

Feminist research is usually qualitative. However, there is diversity and contentiousness in feminist qualitative research (Olesen in Denzin and Lincoln, 2011). While Maynard (1994) thinks there is a distinct feminist methodology, Hammersley (1992) argues that feminists did not invent feminist methods. The commonly used data collection techniques are unstructured interviews, story telling, narratives, role play, observation, and conversation. While some of these are not invented by feminists and not all who use them are feminists, what distinguishes feminist methodology is their use with sensitivity to context, to the power relation between the researcher and researched, and with reflexivity.

Qualitative instead of quantitative methods are preferred because these are less alienating and researcher controlled or directed. Many critical researchers including feminists use both qualitative and quantitative methods because these are seen as not mutually exclusive (Jayaratne, 1983). What is important is attention to context; social conditions and hidden structures of social relations are surfaced and critically examined (Neuman, 1996). These promote rapport and empathy and provide women space and context to talk, and for both the researcher and the women to negotiate the meanings of emerging data. Because researchers must be sensitive to the context of the women, the research plan is tentative, not fully mapped out, and structured. "Feminist research in Africa … have nearly always been improvised and adapted for local usage, often in ways that are not fully conscious, and which often go unreported … a skilled feminist researcher gathering information in a community she is familiar with will be nuanced and adapted by her knowledge of the cultural context, its particular modes of orality and locally specific speech codes inflected by gender, ethnicity, class and age," (Mama, 2013).

There are those who are skeptical about the scientific rigor of feminist research (Hammersley, 1992; Hussain and Asad, 2012). This and other contentious points on representation/voice and the exclusion of women in clinical trials are discussed as ethical/moral contentions in the following section.

Scientific Soundness and Ethics

The scientific quality of feminist research has been questioned on three grounds: (1) lack of methodology or unstructured methodology; (2) highly subjective process of data collection and interpretation (cf Objectivity of Research: Ethical Aspects); and (3) blurring of the political and scientific boundaries.

If feminist research is unscientific, then it is also unethical based on three major guidelines for ethics in research being used by ethics review committees for health research (including social research in health) worldwide: the Declaration of Helsinki (DoH; World Medical Assembly, 2013), Council for International Organizations of Medical Sciences (CIOMS) Guidelines (2002 International Guidelines for Biomedical Research Involving Human Subjects), and the UNESCO Ethical Guidelines for International Comparative Social Science Research (in the Framework of MOST (Management of Social Transformations) Program of UNESCO, 1994–2003). It may be noted that there is no international ethical guideline for social science research that is comparable to the influence of DoH and CIOMS in biomedical research.

What is the 'generally accepted standard of scientific research' that is unbiased and whose findings are ' generalizable'? It is the positivist standard of scientific research. The major guidelines have an implicit bias for positivist, quantitative research methods. *Ipso facto*, these are biased against feminist, openly political, qualitative research that are not attuned to making generalizations but to theory building.

Feminist research has made the clearest and most comprehensive critique of positivist notions of scientific objectivity. For example, Bennett said, "Designing research methodologies capable of addressing the questions which compel us constitutes a politics in its own right, demanding a re-evaluation of received approaches and sophisticated reflection on the intersections of theory and practice" (2008: p. 1).

Feminist research rejects positivism's reliance on highly standardized and formalized techniques of quantitative measurements because these alienate the very vulnerable, powerless, and unheard; are insensitive to context; cannot

uncover hidden inequities; and gloss over specific realities and experiences (Harding, 1986; Maynard, 1994), yet they are used to make generalizations about social life (Scheyvens and Leslie, 2000; Finch, 1984).

It rejects positivism's norm of objectivity (Maguire, 1987) defined as value free and therefore bias free because of the following:

- It is impossible to remove values/bias in the researcher as an embodied, situated social being.
- Bias is masked/hidden by quantitative techniques that create a semblance of absence of the researcher's subject position in the study (Meis, 1983).

Conversely, qualitative methods "can be highly effective for creating a feeling for the whole, for grasping subtle shades of meaning, for pulling together divergent information, and for switching perspectives. It is not an excuse for doing poor quality research, and it has its own discipline and rigor" (Neuman, 1996: 319).

There is ideology in both positivist and feminist methods. In feminist research, the method is simultaneously several things at the same time: a way of collecting data; a language; and a message about the researcher's values regarding the place of the women in the study. The women are viewed as experts closely involved in the process of constructing knowledge about their own lives (Craige, 1992).

In positivism, method is a tool to get data. People are treated simply as data sources and the researcher due to his expertise is privileged to interpret and use the data.

In other words, there is bias in both positivist and feminist research – the questions are: what is the bias, for whom it is, and why this bias? Lather (1986) observed that "research which is openly valued based is neither more nor less ideological than is mainstream positivist research" (64) (cf Objectivity of Research: Ethical Aspects).

Feminist Politics and Scientific Rigor

This is the period of post-positivism (Guba and Lincoln, 1981). No researcher speaks anymore of doing research for knowledge' sake or that they are value free (Hesse, 1980). There is a disconnect between this observation and the major ethics guidelines identified earlier, which stipulate that research must be unbiased and should follow mainstream scientific methods. Ann Oakley (1992) describes the negative reception from funders and scientific review bodies because of her research methods. Ackerly and True (2010) observe that review committees push hard on qualitative researches but do not provide enough guidance on ethical problems for qualitative researchers. On the other hand, critical researchers openly define their political stand with regard to the research (Mies, 1983).

Commitment to the interest of people being studied is a matter of justice. "Siding with the people one researches inevitably means an emotional as well as intellectual commitment to promoting their interests. How else can one justify having taken from them the very private information which many have given so readily?" (Finch, 1984: p. 178).

The point then is not that researcher bias makes poor research. The point is to ensure that bias does not stand in the way of credible, trustworthy data. "The issue is not theoretical vigor versus methodological rigor. The vitality of postpositivist research programs necessitates the development of credibility checks that can be built into the design of openly ideological (and phenomenologically based) research. Both our theory and our empirical work will be the better for the increased attention to the trustworthiness of our data" (Lather, 1986: p. 77). Trustworthiness rather than validity is central to the feminist concept of scientific rigor. Lather (1986) discusses different validity tests in feminist research.

Validity in Feminist Research

Commitment to justice for women is not enough. "My concern is that efforts to produce social knowledge that is helpful in the struggle for a more equitable world pursue rigor as well as relevance … we must formulate self-corrective techniques that will check the credibility of our data and minimize the distorting effect of personal bias upon the logic of evidence" (Lather, 1986: pp. 67–68).

Self-corrective methods have been evolved over time. One is transparency with political standpoint and research background.

Another is detailed description in the research report of the entire research process, from inception to data collection, interpretation, validation, dissemination, and use of findings (Oakley, 1992).

A third method is reflexivity. Derived from Freirean philosophy, reflexivity in research is a complex and deeply personal, sometimes painful process of sensitizing oneself. The researcher reflects on her own consciousness, biases, where she is coming from, and how all these affect her relationship with the people she is studying as well as the data that she is collecting and analyzing. The practice of reflexivity and the perspectives underlying it are among the most distinctive contributions of feminist research in epistemology and methodology.

Fourth method is triangulation. There are three procedures: triangulation by method, by data source, and by researcher (Patton, 2001). More than one method is used in collecting similar sets of data from different data sources by more than one researcher. In cases where data collection is done by only one researcher, triangulation by researcher involves the reinterpretation of the data set by a different researcher. There are, however, variations in triangulation by researcher – randomly chosen datasets, instead of the entire data, may be interpreted by more than one researcher.

Fifth method is member check (Patton, 2001). In this method, interpretation or findings are brought back to research participants for confirmation. Some researchers, to protect confidentiality and participant's anonymity, provide feedback of findings to individual participants in separate sessions.

Integrity and Validity in Research

In the DoH and CIOMS guidelines, researcher integrity is considered in terms of absence of conflict of interest, technical competence, and compliance with intellectual property

regulations. In qualitative research, researcher integrity is a central issue (Neuman, 1996). "We set stringent statistical criteria not because logically they are crucial for establishing the truth of a theory but because our intellectual community is socially distrustful of the honesty of investigators" (Collins, 1984 quoted in Neuman: 321). Where feminist research uses self-correcting procedures, "to some degree a quantitative researcher substitutes explicit descriptions of standard techniques and statistics for trust" (Neuman, 1996: p. 321).

In feminist research, there is an overlap of politics, methodological rigor, and ethics. Ethics could be an inaccurate term to capture the depth and complexity of dilemmas experienced by feminist researchers, especially those who study highly vulnerable groups such as victims of violence, homeless, and indigenous women. In her reflexive writing about her research experience, Finch (1984) observes, "my commitments as a feminist raised moral questions for my work as a sociologist" (p. 176). Her dilemmas were moral and she finds "sanitized intellectual discussion of ethics irrelevant" (p. 178).

Politics of Representation

The second moral dilemma is the question of representation or voice in feminist research. Postmodern and other feminist scholars, especially women of color and indigenous women, raise the issue of universalizing the western, white middle class women's views about women and their life realities for the rest of the world (Collins, 2000). They identify fundamental differences in the issues of women contextualized in specific intersections of race, culture, ethnicity, class, history, and the prevailing political economy. It is not fair to impose categories born from specific contexts on other contexts (Fontes, 2004).

The criticisms affected many researchers not only in the west but also in other countries doing cross-cultural studies. They question themselves if indeed what they are doing will be useful to the women in these cultures (Scheyvens and Leslie, 2000). For example, one has to grapple with the dilemma: should it be simply a description without making an analytic judgment of the oppression? Yet to describe is to interpret, and interpretation involves the deployment of a conceptual frame. The moral dilemma is, whether it is fair to study women in other cultures and describe to the world their realities (cf Scheyven and Leslie, 2000).

The dilemma is sharp with researchers whose purpose is not only to know but to use what they have known to empower women. For example, "when women themselves do not see gender inequality as a material issue, or where women themselves do not see the need for their active participation in political activities, should researchers or external actors insist on taking up the issues?" (Alvarez Castillo and Cook Lucas, 2010: p. 34).

Others think that asking the marginalized, the unheard about their situation, listening to their stories, and sharing these with the rest of the world could contribute to their empowerment. These could affirm their self worth in knowing that they can contribute to knowledge. By thinking about hitherto unspoken subjects or unasked questions, consciousness change can take place (Acker et al., 1991).

There are two kinds of initial reaction to the criticisms: many western researchers inhibited themselves from doing cross-cultural studies; others took a relativist stance (Scheyven and Leslie, 2000; cf Cultural Relativism, Anthropology of). Relativism belongs to a theoretical school that rejects the notion of absoluteness of truth or values like goodness or justice. What is true or good is relative. At the same time, there are those who, despite their self doubt over their relevance, experience affirmation from women in other cultures of the usefulness of what they are doing. Reinhraz (1992) thinks that women's concerns in the west are universal concerns of women, but they should know and understand women and their specific contexts in other countries. It is crucial to recognize the presence of multiple perspectives and to be sensitive to situations where these perspectives come (Lugones, 1987).

The researcher must acknowledge that she made choices in constructing her representations of other women (Haraway, 1988; cf Research Ethics, Cross Cultural Dimensions of); that her social situation affects such representations (Harding, 2004).

Women's Exclusion in Clinical Studies

While this debate is in the field of medical research, it is taken up here because it involves issues of gender in research. The importance of the question on the exclusion/inclusion of women with childbearing potential (henceforth to be referred to as WCP) in clinical trials can be contextualized in two developments: the evidence-based medicine (EBM) movement and the women's rights movement. Ironically, EBM while not associated with the women's rights movement, the latter can draw arguments from EBM for the inclusion of women in clinical trials.

Evidence-Based Medicine (EBM)

For a larger ambit, evidence-based practice (EBP) is used to include other fields in health care such as nursing and public health. There are scholars, however, who point to the conflict between EBM and EBP (Fischer and Ferlie, 2013). The Cochrane Collaboration global network, leading international research centers, and even the pharmaceutical industry have embraced EBM as the gold standard (cf Research Funding: Ethical Aspects). The medical insurance industry in the west is now basing its payments on whether or not the intervention used EBM standards. EBM is defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" (BMJ Editorials, 1996).

In EBM, there is less reliance on individual practitioner's experience and judgment (believed to have high degrees of uncertainty) and more on formal evidence from quality research with high certitude as basis for best practice. Critical appraisals of evidence from the literature is a central method in EBM where evidence is assessed according to a hierarchy of validity through meta-analysis, statistical tests, or double-blind random trials. Outcomes from population trials are applied to individual patients.

In EBM, decision making on interventions should be guided by the following: effectiveness, safety, efficiency (cost and benefit ratio), and ethics (cf Therapeutic Effectiveness Assessment: Approaches in Contemporary Research).

Exclusion of Women in Clinical Trials

The exclusion of women in clinical trials has been historical (Prout and Fisch, 2001). Rogers and Ballantyne (2008) reviewed 400 clinical trials involving Australians and found that more women than men are included. That women are overrepresented in clinical trials is a controversial subject in the literature (see Prout and Fisch, 2001, on this controversy). However, according to Holdcroft (2007) EBM may be fundamentally erroneous due to the absence of research tools on sex differences. If EBM is to be consistently applied, no drugs that have not been tested by gold standard methods on WCP should be given to them (Holdcroft, 2007; Jones, 2009). But the fact is this is being done quite commonly.

Due to the exclusion of WCP in many drug trials, physicians do not have enough robust evidence to use medication for pregnant women. There is very little reliable information on the safety of drugs for pregnant women both for the women and physicians (Lipman, 2006).

Why Are WCP and Pregnant Women Excluded in Clinical Trials?

There are ideological, scientific, legal – and ironically – ethical explanations. Ideologically, it is the bias that what is good for men is good for women (Eckenweller, Feinholz et al. in commentary submitted to WMA in 2008). Scientifically, the hormonal changes in women could confound findings. Legally, there is fear of potential legal suits against researchers (and study sponsors) (Society for Women's Health Research, n.d.), (i.e., by women participants who suffered miscarriage during the trial). And ethically, it is to protect the fetus from potential harm. The fetus is now included in the list of vulnerable groups and in the case of pregnant women, it is the fetus who is considered the actual vulnerable population (Prout and Fisch, 2001). To protect the fetus, WCP and pregnant women must be excluded. The fact that many pregnant women take maintenance medication, for example, means that "they are essentially participating in an uncontrolled and unmonitored experiment for which the data will most likely never be assessed" (Berlin and Ellenberg, 2009).

The HIV/AIDS crisis challenges this norm when HIV+ pregnant women cannot participate in antiviral trials. To many of these women, the risks of harm are outweighed by the potential for treatment of the disease (Berlin and Ellenberg, 2009). While there is now more awareness about the question of women's exclusion and attempts to include women in clinical studies (Epstein), the general pattern remains exclusionary rather than inclusionary. The arguments for their inclusion are grounded on justice, science, and ethics. Those who advocate for the inclusion of WCP in clinical trials as a matter of policy argue that exclusion puts WCP at great risk when they take medication whose safety for them has not been established scientifically. The very reason for excluding them is in fact the very reason why they must be included: the pharmacologic effects of the medication including dosing must include their specific hormonal and physiologic characteristics during pregnancy when physicians make decisions (Prout and Fisch, 2001).

There must be respect for women's decision making, their ability to weigh the risks, and benefits where clear and adequate information is given to them. "The net effect of gender bias in medical research is that women are at risk for adverse drug reactions and may unnecessarily suffer and die. Such adverse reactions occur approximately twice as often in women as in men" (Schiebinger, 2003: p. 976).

It is injustice against pregnant women to be placed at risk of harm both for them and their babies when they take medication the safety and effectiveness of which have only been established for males. The ethical argument for exclusion flies in the face of the fact that no one is actually being protected because "exposure to a drug being tested is merely delayed until the drug is released and on the market" (Society for Women's Health Research, n.d.: accessed 30.12.13.).

It is in accordance with justice and ethics that WCP and pregnant women are not made to bear the burden of risks simply because science and funders have not given enough attention (and investment) in developing appropriate scientific methods and tools.

The Feminist Association for Bioethics (FAB) in a commentary solicited by the WMA calls for revisions in the DoH guidelines (Eckenweller et al., 2008) that will explicitly and proactively ensure the inclusion of WCP in research that will directly benefit them. Barriers to participation of marginalized groups should be addressed. Exclusion cannot be justified on the basis of potential harm to the fetus alone. Potential harm can be scientifically established with more studies on animals and humans, and how physiologic and hormonal specificities affect the pharmacologic processes of the drugs. FAB also suggests that where WCP are recruited, information about possible termination of pregnancy, or long-term medical followup, and the requisite use of contraception should be clearly explained by researchers when soliciting consent (cf 11029).

Reforms

There have been reforms to address the discrimination against WCP in clinical trials. For example, the DoH, in Principle 13, says, "Groups that are underrepresented in medical research should be provided appropriate access to participation in research" (cf Ethics for Biomedical Research Involving Humans: International Guidelines). The US NIH in 1993 issued the policy that all its grants must ensure that women and members of minorities and their subpopulations are included in all human subject research; for Phase III clinical trials, ensure that women and minorities and their subpopulations be included so that variations in the effects of the intervention are determined; and that cost cannot be used as a reason for exclusion. The NIH Office of Research in Women's Health maintains a database on the inclusion of women in NIH-funded trials.

The US FDA and the Centers for Disease Control and Prevention have also made similar policy reforms.

But reforms have not been enough. The US NIH funds only about 20% of clinical research with the rest of the funding coming from private groups such as the pharmaceutical industry. While there has been more participation of women in clinical studies, this is not enough. What is needed is sex-specific analysis that can determine specific pharmacologic effects such as response to the drug, adverse events, and drug interactions.

It is also recommended that women's inclusion is not limited to Phase III trials because the correct dosage and use of drugs are usually established in the earlier phases (SWHR, n.d.; Prout and Fisch, 2001).

Political Economy in Women's Exclusion

It is more costly, more lengthy, and more inconvenient for researchers to include WCP (Schiebinger, 2003). The physiologic differences between men and women could mean larger sample sizes (Prout and Fisch, 2001). Women who become pregnant may require long-term followup that could include counseling. Women's caregiving roles could increase dropout rates. Women tend to have lower insurance coverage than men because they are more likely to work part time (Prout and Fisch, 2001).

Bibliography

Acker, J., Barry, K., Esseveld, J., 1991. Objectivity and truth: problems in doing feminist research. In: Fonnow, Cook (Eds.), Beyond Methodology: Feminist Scholarship as Lived Research. Indiana University Press.

Ackerly, B., True, J., 21 May 2010. Framework for Research Ethics and Evaluation: Justification and Guidelines. http://www.palgrave.com/methodology/doingfeministresearch/resources/guides/framework_Research_Ethics_Evaluation.pdf.

Alvarez Castillo, F., Lucas, Julie Cook, 2010. Fairness and Gender in Benefit Sharing: Learning from the Kani, San, Nigerian, Kenyan and Icelandic Cases for Moving Forward. http://www.uclan.ac.uk/research/environment/projects/assets/cpe_genbenefit_gender.pdf.

Anderson, E., Fall 2012 edition. Feminist epistemology and philosophy of science. In: Zalta, Edward N. (Ed.), The Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/archives/fall2012/entries/feminism-epistemology/.

Barker, J., 2010. Evidence-Based Practice for Nurses. Sage, London.

bell hooks, 1984. Feminist Theory: From Margin to Center. South End Press, Cambridge, MA.

Belsey, J. What is …? Series. second ed. What Is Evidence – Based Medicine? http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/ebm.pdf.

Bennet, J., December 2008. Researching for life: paradigms and power. Feminist Africa (11).

Berlin, J., Ellenberg, S., 9 October 2009. Commentary: inclusion of women in clinical trials. BMC Medicine 7, 56. http://dx.doi.org/10.1186/1741-7015-7-56. Published online 9 October 2009. No pagination.

Canella, C., Lincoln, Y., 2011. Ethics, regulations and critical social science. In: Denzin, N., Lincoln, Y. (Eds.), The Sage Handbook of Qualitative Research. Sage, Thousand Oaks, New Delhi.

Christians, C., 2011. Ethics and politics in qualitative research. In: Denzin, N., Lincoln, Y. (Eds.), The Sage Handbook of Qualitative Research. Sage, Thousand Oaks.

Collins, P.H., 2000. Black Feminist Thought: Knowledge, Consciousness and the Politics of Empowerment. Routledge, New York.

Craige, B.J., 1992. Laying the Ladder Down: The Emergence of Cultural Holism. University of Massachusetts Press, Amherst.

Denzin, N., Lincoln, Y., 2011. The Sage Handbook of Qualitative Research. Sage Publications, Thousand Oaks.

DuBois, P., 1994. The platonic appropriation of reproduction. In: Tuana, N. (Ed.), Feminist Interpretations of Plato. The Pennsylvania State University Press, Pennsylvania, pp. 67–86.

DuBois, Barbara, 1983. Passionate scholarship: notes on values, knowing and method in feminist social sciences. In: Bowles, Gloria, Klein, Renate Duelli (Eds.), Theories of Women's Studies. Routledge and Kegan Paul, London, pp. 105–117.

Eckenweller, Feinholz, D., Ells, Carolyn, Schonfeld, Toby, 2008. The declaration of Helsinki through feminist eyes. In: Solicited Commentary Submitted to the World Medical Assembly.

Editorials, 13 January 1996. Evidence-based practice: what it is and what it isn't. BMJ 312, 71–72. http://www.bmj.com/cgi/content/full/312/7023/71.

Ehrenreich, B., 19 May 2004. What Abu Ghraib Taught Me. http://www.alternet.org/story/18740/what_abu_ghraib_taught_me.

Epstein S. Health research and the remaking of common sense. Excerpt from Inclusion: the Politics of Difference in Medical Research. At: http://press.uchicago.edu/Misc/Chicago/213095.html (accessed 30.12.13.).

Estandarte, N., Alvarez Castillo, F., 1999. Sungki-Sungking Daan Women Empowering Women. Philippine Health Social Science Association, Manila.

Finch, J., 1984. Its great to have someone to talk to: ethics and politics of interviewing women. In: Bell, C., Roberts, H. (Eds.), Social Researching: Policies, Problems and Practice. Routledge and Kegan Paul, London.

Fine, M., 1992. Disruptive Voices: The Possibilities of Feminist Research. University of Michigan Press, Ann Arbor.

Fischer, M.D., Ferlie, E., 1 January 2013. Resisting hybridisation between modes of clinical risk management: Contradiction, contest, and the production of intractable conflict. Accounting, Organizations and Society 38 (1), 30–49.

Fonnow, M.M., Cook, 1991. Beyond Methodology: Feminist Scholarship as Lived Research. Indiana University Press, Bloomington.

Fontes, L.A., 2004. Ethics in violence against women research the sensitive, the dangerous and the overlooked. Ethics and Behavior 14 (2), 141–172.

Fox, M.F., March 2004. Women in Scientific Fields: Doctoral Education and Academic Careers. Workshop on Women's Advancement. American Political Science Association. https://www.apsanet.org/imgtest/maryfrankfox.pdf.

Francheschet, S., 2003. "State Feminism" and women's movements: the impact of Chile's Servicio Nacional de la Mujer on women's activism. Latin American Research Review 38 (1), 9–40. http://dx.doi.org/10.1353/lar.2003.0006.

Ginther, D., Kahn, S., June 2006. Does Science Promote Women? Evidence from Academia 1973–2001. At: http://users.nber.org/~sewp/Ginther_Kahn_revised8-06.pdf (accessed 05.11.12.).

Grey, S., 2004. Decolonising feminism: aboriginal women and the global 'Sisterhood'. In: Enweyin. The Way We Speak, vol. 8 (2003–04), pp. 9–22. http://www.academia.edu/1330316/Decolonising_Feminism_Aboriginal_Women_and_the_Global_Sisterhood.

Guba, E.G., Lincoln, Y.S., 1981. Effective Evaluation Inquiry: The Usefulness of Evaluation Results through Responsive and Natural Approaches. Jossey-Bass, San Francisco, CA.

Hammersley, M., 1992. On feminist methodology. Sociology 26 (2), 187–206.

Haraway, D., 1988. Situated knowledges: the science question in feminism and the privilege of partial perspective. Feminist Studies 14 (3), 575–599.

Harding, Sandra, 1986. Feminism and Methodology. Indiana University Press, Bloomington.

Harding, S., 1986. The Science Question in Feminism. Cornell University Press, Ithaca.

Harding, S., 2004. Rethinking standpoint epistemology: what is 'Strong objectivity'? In: Harding, Sandra (Ed.), The Feminist Standpoint Theory Reader: Intellectual and Political Controversies. Routledge.

Hesse, M., 1980. Revolutions and Reconstructions in the Philosophy of Science. Harvester Press.

Holdcroft, A., January 2007. Gender bias in research: how does it affect evidence based medicine? Journal of the Royal Society of Medicine 100 (1), 2–3.

Hurtado, E., quoted ECLAC Report of October 2013, 27 November 2013. News Report titled Latin America's Scientists Outnumbered at the Top. At: http://www.scidev.net/global/gender/news/latin-america-female-scientists-outnumbered.html (accessed 12.12.13.).

Hussain, B., Asad, A.Z., 29 November 2012a. A critique on feminist methodology. Journal of Politics and Law 5 (4), 202–207. At: www.ccsenet.org/jpl (accessed 30.11.13.).

Hussain, B., Asad, A.Z., 2012b. A critique on feminist research methodology. Journal of Politics and Law 5 (4), 202–207.

Jaggar, A., 2013. Does poverty wear a woman's face? some moral dimensions of a transnational research project. Hypatia 28 (2), 240–246.

Jayaratne, Toby, 1983. The value of quantitative methodology for feminist research. In: Bowles, Gloria, Klein, Renate Duelli (Eds.), Theories of Women's Studies. Routledge and Kegan Paul, London, pp. 140–162.

Jones, N., 8 June 2009. Too Few Women in Clinical Trials? Nature News. http://dx.doi.org/10.1038/news.2009.549.

Katz, C., 1996. The expeditions of conjurers: ethnography, power and pretense. In: Wolf, D. (Ed.), Feminist Dilemmas in Fieldwork. Westview Press, pp. 171–184.

Kritz, F.L. Drug Risks for Pregnant Women. http://www.parenting.com/article/the-pregnancy-prescription.

Larivière, V., Ni, Chaoqun, Gingras, Yves, Cronin, Blaise, Sugimoto, Cassidy R., 11 December 2013. Bibliometrics: global gender disparities in science. Nature 504 (7479). http://www.nature.com/news/bibliometrics-global-gender-disparities-in-science-1.14321.

Lather, P., 1988. Feminist perspectives on empowering research methodologies. Women's Studies International Forum 11 (6), 569–581.

Lather, P., 1986. Issues of validity in openly ideological research: between a rock and a soft place. Interchange 17 (4), 63–84.

Lavery, J.V., Green, S.K., Bandewar, S.V.S., Bhan, A., Daar, A., et al., 2013. Addressing ethical, social, and cultural issues in global health research. PLoS Neglected Tropical Diseases 7 (8), e2227. http://dx.doi.org/10.1371/journal.pntd.0002227 (pub 8 August 2013).

Lengermann, P.M., Niebrugg-Brantley, J., 2003. "Contemporary feminist theory". In: Ritzer, G. (Ed.), Contemporary Sociological Theory and its Classical Roots: The Basics. McGraw Hill, New York, pp. 309–358.

Lipman, A., March 2006. The inclusion of women in clinical trials: are we asking the right questions? Women and Health Protection. http://www.whp-apsf.ca/pdf/clinicalTrialsEN.pdf.

Lugones, Maria, 1987. Playfulness, 'world'-traveling, and loving perception. Hypatia 2, 3–19.

Maguire, P., 1987. Doing Participatory Research: A Feminist Approach.

Maliniak, D., Powers, Ryan, Walter, Barbara F., 2013. The Gender Citation Gap in International Relations. https://www.princeton.edu/politics/about/file-repository/public/IO-Final-April-2-2013.pdf.

Mama, A., 2013. What does it mean to do feminist research in African contexts? Local/global interrogations panel. Feminist Review Journal. In: Feminist Theory & Activism in Global Perspective: FR Conference Proceedings, e4–e20. http://www.palgrave-journals.com/fr/conf-proceedings/n1s/full/fr201122a.html.

Martin, D., 2007. Bureaucratizing ethics: institutional review boards and participatory research. ACME Editorial Collective 6 (3), 321–328.

Maynard, Mary, 1994. Methods, practice and epistemology: the debate about feminism and research. In: Maynard, Mary, Purvis, Jane (Eds.), Researching Women's Lives from a Feminist Perspective. Taylor and Francis, London, pp. 10–27.

Mies, Maria, 1983. Towards a methodology for feminist research. In: Bowles, Gloria, Klein, Renate Duelli (Eds.), Theories of Women's Studies. Routledge and Kegan Paul, London, pp. 117–140.

Miller, M., 2001. Gender-based differences in the toxicity of pharmaceuticals – the food and drug administration's perspective. Office of Women's Health, Food and Drug Administration, Rockville, Maryland, USA International Journal of Toxicology 20, 149–152. http://www.fda.gov/downloads/ScienceResearch/SpecialTopics/WomensHealthResearch/UCM247552.pdf.

Mohanty, C., 2003. Under western eyes: feminist scholarship and feminist discourses. In: Feminism without Borders: Decolonizing Theory, Practicing Solidarity. Duke University Press.

Murthy, R., Klugman, B., Weller, S., 2005. Service accountability and community participation. In: Ravindran, Sundari, de Pinho, Helen (Eds.), The Right Reforms? Health Sector Reforms and Sexual and Reproductive Health. Women's Health Project, School of Public Health, University of Witwatersrand.

Neuman, L.H., 1996. Social Research Methods: Qualitative and Quantitative Approaches.

Nobilius, A.M., 23 June 2004. Gender Competence. http://www.med.monash.edu.au/gendermed/competence.html.

Oakley, A., 1974. The Sociology of Housework. Martin Robertson, London.

Oakley, A., 1992. Social Support and Motherhood: The Natural History of a Research Project. Blackwell.

Olesen, V., 2011. Feminist qualitative research in the millenium's first decade: developments, challenges, prospects. In: Denzin, N., Lincoln, Y. (Eds.), Sage Handbook of Qualitative Research. Sage Publications, Thousand Oaks, pp. 129–146.

Patton, M.Q., 2001. Qualitative Research and Evaluation Methods. Sage, Thousand Oaks.

Pogge, T., Rippin, Nicole, May 2013. Universal Agenda on Multiple Dimensions of Poverty. Background Paper, High Level Panel on the Post-2015 Development Agenda. At: http://www.genderpovertymeasure.org/publications/ (accessed 02.12.13.).

Proudlock, n.d. 'New' Masculinities, Old Shortcomings? The Implications of a New Male: Focus for Gender and Development. At: https://www.sussex.ac.uk/webteam/gateway/file.php?name=proudlock&site=10 (accessed 02.12.13.).

Prout, M., Fisch, S., 2001. Participation of women in clinical trials of drug therapies: a context for the controversies. Disclosures Medscape General Medicine 3 (4). http://www.medscape.com/viewarticle/408956 (accessed 15.11.13.).

Ramazanoglu, Caroline, 1992. On feminist methodology: male reason versus female empowerment. Sociology 26 (2), 207–212.

Reinhraz, S., 1992. Feminist Methods in Social Research. Oxford University Press.

Ritchie, T.D., 2009. Gender bias in research. In: O'Brien, J., Fields, J., Shapiro, E. (Eds.), Encyclopedia of Gender and Society, vol. 2. Sage Publications, Thousand Oaks, CA, pp. 713–715.

Rogers, W.A., Ballantyne, A.J., May 2008. Exclusion of women from clinical research: myth or reality? Australian Gender Equity in Health Research Group. Mayo Clinic Proceedings 83 (5), 536–542. http://dx.doi.org/10.4065/83.5.536. http://www.ncbi.nlm.nih.gov/pubmed/18452682.

Sabea, H., December 2008. Transnational what? Encounters and reflections on questions of methodology. In: Feminist Africa 11 Researching for Life: Paradigms and Power. Published by African Gender Institute, University of Capetown, pp. 13–28.

Scheyvens, R., Leslie, H., 2000. Gender, ethics and empowerment: dilemmas of development fieldwork. Women's Studies International Forum 23 (1), 119–130.

Schiebinger, L., 2003. Women's health and clinical trials. Journal of Clinical Investigation 112 (7), 973–977. http://dx.doi.org/10.1172/JCI19993.

Science and Technology: Key Tables from OECD. http://dx.doi.org/10.1787/2075843x. At: http://www.oecd-ilibrary.org/science-and-technology/women-researchers_2075843x-table3 (accessed 25.12.13.).

Smith, D., 1987. The Everyday World as Problematic: A Feminist Sociology. Northeastern University Press, Lebanon.

Society for Women's Health Research, n.d. Issue: Inclusion of Women and Minorities in Clinical Research. http://www.womenshealthresearch.org/site/PageServer?pagename=policy_issues_clin.

Starman, E., June 2009. Meeting Report. A South Asia Consultation on Working with Boys and Men for Gender Equality and Prevention of Gender-based Violence. Dhaka.

Tuhiwai Smith, L., 1999. Decolonizing Methodologies: Research and Indigenous People. Zed, London.

UNESCO. World Social Science Report 2010. www.unesco.org/shs/wssr (accessed 25.12.13.).

UNESCO, 2011. UN Commission for Latin America and the Caribbean (ECLAC) cited UNESCO 2011. Report in ECLAC report of October 2013.

Wallerstein, I., et al., 1996. Open the Social Sciences: Report of the Gulbenkian Commission on the Restructuring of the Social Sciences. Stanford University Press, Stanford.

Westmarland, Nicole, 2001. The Quantitative/Qualitative debate and feminist research: a subjective view of objectivity [28 paragraphs]. Forum Qualitative Sozialforschung/Forum: Qualitative Social Research 2 (1). Art. 13. http://nbn-resolving.de/urn:nbn:de:0114-fqs0101135.

Women in Science and Technology (WISAT). National Assessments on Gender and STI. At: http://www.wisat.org/gender-sti/ (accessed on 15.12.13.).

Women's Studies Programs, Departments, & Research Centers. At: http://userpages.umbc.edu/~korenman/wmst/programs.html (accessed 28.12.13.).

Gender and Technology: From Exclusion to Inclusion?

Vivian A Lagesen, Norwegian University of Science and Technology, Trondheim, Norway

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Wajcman, volume 9, pp. 5976–5979, © 2001, Elsevier Ltd.

Abstract

The field of gender and technology is a rather broad interdisciplinary scholarly field that has existed for more than 30 years now. It includes various aspects of technology and its social and cultural relationship with men, women and gender. The most prominent strand of research has been feminist technology studies (FTS), which has been more or less connected to science and technology studies (STS). This strand has engaged scientifically as well as politically with the *exclusion* of women from many technological areas and spheres. However, there have also been more optimistic strands of research in this field, for example so-called cyberfeminist studies, which have focused on technology as a means for the *inclusion* of women. Alongside such approaches, a social shaping approach has studied the mutual and reciprocal *social shaping* or *coconstruction* of technology and gender. These three narratives – exclusion, social shaping/ coconstruction, and inclusion – have been the most dominant and important perspectives or narratives in this field, and are the point of departure for this article. In the end, I will propose a fourth emerging narrative/theoretical perspective in the field, drawing on *material feminism*.

Introduction

The field of gender and technology is a rather broad interdisciplinary scholarly field that has existed for more than 30 years now. It includes various aspects of technology and its social and cultural relationship with men, women and gender. Even if it is a dispersed and multifaceted field, its scholarship has given particular attention and concern to particular technologies with regard to gender. These are Information and Communication Technologies (ICT), reproduction and biomedical technologies, and household technologies. The most prominent strand of research has been the field of feminist technology studies (FTS), which is related to science and technology studies (STS). This strand has engaged scientifically as well as politically with the *exclusion* of women from many technological areas and spheres. However, there have also been more optimistic strands of research in this field, which has focused on technology as a mean for *inclusion* of women, for example so-called cyberfeminist studies. Alongside these studies there has been a social shaping approach that has studied the mutual and reciprocal *social shaping* or *coconstruction* of technology and gender. These three narratives, exclusion, social shaping/coconstruction, and inclusion have been the most dominant and important perspectives or narratives in this field and will be the point of departure for this article. In the end I will propose a fourth emerging narrative/theoretical perspective in the field, drawing on *material feminism*.

The Exclusion Narrative

The field of gender and technology emerged from a feminist scientific and political concern to investigate the impact on technology on women's lives. Historically, the interest in the study of gender and technology developed from the second wave of the feminist movement at the end of the 1970s and the beginning of the 1980s, when there was a particular interest in the gendered power dimensions of technology. One important source of inspiration was the Marxist-inspired studies of industry, capitalism and labor (e.g., Baxandall et al., 1976; Braverman, 1974). Thus, the early historical accounts of men's domination of industrial technology were framed in terms of power (Cockburn, 1981) and leaned on Marxist accounts of labor process change, but also on object-relations theory from psychoanalysis. Many feminists wrestled with what Heidi Hartmann (1979) called 'The Unhappy Marriage of Marxist and Feminism.' The early feminist studies of technologies brought up a range of different challenges and potential opportunities for women's development of new technology. The most prominent ones were household technologies. These studies raised questions such as, did innovation in household technologies liberate women from household tasks? (Bose et al., 1984; Thrall, 1982; Vanek, 1974). Although contested, the answer was, and has continued to be, predominantly 'no' (Bittman et al., 2004). Ruth Schwartz Cowan's (1983) classic work of women's history shows that modern middle class women, despite the obvious technological changes in household appliances (most of them presumably labor-saving), worked as long if not as hard in their home as before. They just struggled to keep up with ever-higher standards of cleanliness.

An important concern in feminist technology studies has been the consequences of new reproduction technologies for women's lives. The quest for opportunities brought forward by biomedical technologies started with Shulamith Firestone's (1970) famous book, which argued that patriarchy was imposed upon women through their biology and the restrictions that pregnancy, childbirth and child-rearing entailed. She imagined human reproduction taking place in laboratories, thus emancipating women from their biological burdens by means of technology. Different biomedical technologies have been seen to represent new opportunities for women to control and also to design reproduction, to be able to choose when and how to have children, or whether or not to have them at all (Clarke, 1988; Casper and Berg, 1995; Thompson, 2005). On the other hand, both in a critical patriarchal perspective and in an ecofeminist perspective these technologies have

been interpreted as a way to control women and their bodies, as technology in itself has been considered a masculine project to control women and nature (Merchant, 1980; Mies and Shiva, 1993; Wajcman, 1991).

Another important issue has been the exclusion of women from technology and particularly the consequences and causes of women's lack of participation in the producer side of technologies (Faulkner and Arnold, 1985; Cockburn, 1983). Also, men's and women's (different) relationships to machines have been investigated (Cockburn, 1981; Form and McMillen, 1983), as well as the disqualification and degradation of women's work (Phillips and Taylor, 1980) and the polarization of men and women's job when new technologies are introduced (Feldberg and Glenn, 1983).

Two areas in which the exclusion of women has been seen as particularly severe have been in engineering (Frehill, 2004; McIlwee and Robinson, 1992) and computer science (Margolis and Fisher, 2002). Both areas/fields have been perceived as enacting 'masculine cultures.' The question 'why so few' has been, and still is, the most common one. There have been different answers to this question, but usually one or several deficit-models have been drawn upon. Women's 'deficits' have been one common explanatory model (in both feminist and nonfeminist literature). It has been argued that women lack interest, skills, courage, motivation, etc. and the solution is to inform/encourage more women to make their way into these areas. Deficits in the culture, meaning either a 'chilly culture' with discrimination and harassment of women and/or a 'masculine' culture that makes women feel alienated, have often been described and used in explanations (McIlwee and Robinson, 1992; Wajcman, 1991, 2004).

Thus, in many ways there has been a rather pessimistic view of technology as patriarchal, malignant and fostering the exclusion of women. Also, the narrative of exclusion has been – and still is – the most common one in the literature about gender and technology as well as in reviews of the field. What has been seen as the source of men's domination of (and with) technology has been a debated issue. Is it due to men's interests and motives, and patriarchal structures, or is there something inherently gendered about technology? Quite early on, feminist studies of technology became occupied with explaining the exclusion of women by showing how technology in itself was gendered. Berg and Lie (1995) polemically asked 'Do artifacts have gender?', a question which they answered affirmatively. But what does that mean? *How* do artifacts have gender? Does technology always have gender? And how might or should we study technology and gender?

The Social Shaping/Coconstruction of Gender and Technology

The most common answer to the last of these questions has been the idiom of *coconstruction* or *mutual shaping*. Feminist studies of technology have been deeply inspired by social constructivism, especially the social construction of technology approach (SCOT), and by ideas about the mutual shaping of society and technology, and of technology and society as a 'seamless web.' The main tenets of the coconstruction idiom are quite simple: a reciprocal interaction, where two dependent variables or a set of variables interact or influence each other simultaneously to create new outcomes. Feminist studies of technology have produced a substantial body of groundbreaking empirical research in a wide range of areas concerning women and technology. One such area of concern is with women's everyday uses of technology, in which women are usually at the receiving end but not of the design of the technologies they use (Faulkner and Arnold, 1985; Berg, 1996; Cockburn and Fürst-Dilic, 1994; Lie and Sørensen, 1996). Another has to do with technological 'genderscripts': technology designed to cater to men's or women's particular needs, or rather to the designers stereotypical configurations of these needs (Oudshoorn et al., 2002). Faulkner has termed the first issue gender *and* technology and the second gender *in* technology (Faulkner, 2001).

There have been two major ways of which gender and technology has been seen as coconstructed or mutually shaped: one focuses on the designers' role, and the other is concerned with the users of the technology. The main ideas of the coconstruction approach is illustrated in two exhibitions on gender and things (Oudshoorn et al., 2002). The first exhibition was from an idea that emerged during a workshop on gender and technology at the University of Twente, and was organized first on campus, and then later by the science museum in Amsterdam. Later, the same exhibition was set up at the campus at the university in Trondheim. The authors of the article accompanying the exhibitions proclaim that their goal was to show how things are inscribed with gender, "in order to make people aware that we live in a technological and gendered culture" (Oudshoorn et al., 2002, 472). Oudshoorn et al.'s article uses two approaches to do this. The first one is called a 'genderscript' approach, which draws on Akrich's (1992) concept of technological script. The gender script approach is seen to entail two processes of gendering; one is the gendering that takes place through a 'configuring' of a particular version of masculinity or femininity. The producers or designers anticipate the user's preferences, motives, tastes and skills, combined with general cultural norms, and materialize them into a technological object. The design, then, produces a mutual shaping of gender and technology. The next process is that these artifacts are seen as shaping and defining the agency of men and women:

> Due to the norms and values that are inscribed into a technical artifact, objects can attribute and delegate specific roles, actions, and responsibilities to their users.
>
> (Oudshoorn et al., 2002, 473).

The initial exhibition in the Netherlands focused on how designers and other actors were involved in the construction of gendered artifacts. The examples shown were shavers from Phillips, bicycles and microwave ovens from a famous study of Cockburn and Ormrod (1993) (see also below). In addition, the exhibition included a section on objects and artifacts that transcended the traditional separation between feminine and masculine. For example Dr. Marten shoes (once shoes for English policemen that were later used by girls disassociating themselves from the social imperative to be pretty). Further, these are all described as studies of the mutual shaping of gender and (technological) artifacts.

The second approach focuses on how users 'domesticate' technology (Lie and Sørensen 1996). Domestication is a concept that has been used to capture what happens when users integrates the technology into their daily lives. It consists of practices as well as symbolic work. In this perspective users are perceived as active agents in shaping technology, rather than passive adopters of the designer's scheme. Instead of adopting the script, users can choose to alter, modify, or even reject it, thus creating a whole new pattern of use. Domestication is also described as a reciprocal process that changes both people and the technological objects they use (ibid: 478), thus involving a coconstruction of gender and technology.

Faulkner (2001) takes a slightly different approach to the coproduction of gender and technology. While arguing that a mutual shaping of gender and technology occurs, she differentiates between what she calls gender in technology and gender of technology. The latter includes the gendered *associations* that are made in relation to a particular technology, and is based on a different kind of symbolism that connects to different kinds of technology. This symbolism for example is associated with a gendered division of labor (see also Sørensen and Berg, 1987 for a similar argument). The argument is that a technology is perceived as 'feminine' or 'masculine' depending upon who 'normally' uses it. This perspective opens up a substantial degree of interpretative flexibility. For instance, Faulkner shows how a computer keypad was seen as masculine by those who associated it with mathematics, assuming that it something that more men than women would use. Others associated the keypad with data entry, a stereotypical women's activity. On the other hand, Faulkner also stresses that in practice many sociotechnical arrangements are quite stable, irrespective of the theoretical potential for fluidity. Another example put forward by Faulkner is the labeling of technology with gendered language that refers to it as male or female. According to Faulkner (2001, 83), this also reinforces heteronormativity in the sense that it enacts heterosexual relations as 'normal.'

The gender *in* technology perspective focuses on designers and producers of technology. Some artifacts may be designed according to the designer's assumptions about gender. One example that is frequently used to illustrate this is Cockburn and Ormrod's (1993) study, which shows how the microwave oven was originally designed and marketed as a 'brown good' product like televisions, videos and hi-fi systems. It was first designed and marketed for men, assumed to be single and not particularly keen to cook. When this market failed, the producers decided to change the microwave and its configured user. The newly targeted group was women and families/couples shopping for their kitchens, and the microwave was redesigned for 'serious' cooking and sold as 'white goods' appliances such as stoves, refrigerators and freezers.

The gendering process in this case involved symbolic associations as well as a strict division of labor (women cook, men do not). 'Brown goods' were associated with men and men's enthusiasm for technology, and even the sales personnel were men. When the microwave was placed in the 'white goods' department, it was depicted as 'feminine.' Also, managers thought that customers would want to buy white goods from a saleswoman, rather than from a man. Thus, Cockburn and Ormrod (1993) showed how perceptions of gender roles were projected onto the microwave oven, thus making it into a gendered artifact.

Another example is Berg's (1996) study of the design of so-called smart houses or automated homes. This case also is referred to frequently as an important exemplar of coconstruction. Berg shows that the designers of the houses, men engineers, imagined themselves as users of the houses ('I-methodology'), and their design of these houses did not include technological solutions to ease household work (considered to be women's work). As the male designers left potential household tasks out of the development and design process, the smart house was characterized as a stereotypical 'masculine' construction. Thus, gender was assumed as an integral part of the construction. The underlying, counterfactual assumption was that women would have been less interested in the kind of technological solutions that the men designers favored, such as control and security systems, and would be more interested in automating housework. Again, at the core of Berg's observations was the gendered division of labor related to housework as well as design, and the case illustrates the problem of 'translating' gender to technology (see Sørensen, 1992 for a discussion). The gendering of the technology can only be made by the actors themselves and not by the interpretation done by the researchers.

These vanguard studies and many others show very convincingly how technologies are gendered; in particular, how they are shaped by and linked to the pervasive sexual division of labor. But, these studies have not investigated how gender is constructed and what role technology may play in the production of gender. In other words, the assumed dynamic relationship has not been so dynamic after all. Instead, gender often is invoked as a static, a priori category and also is constructed as a binary opposition. The lack of symmetrical attention in many allegedly coconstructional studies of gender and technology has been criticized; in fact, Ormrod (1994, 32) criticized early feminist work for presuming that "the content and behavior of gender relations is assumed to be common knowledge, and their meanings are stabilized and no longer need to be considered."

Arguably, such a structural lag produces an implicit conception of gender as fixed and nondynamic, opposite to the emphasis on 'doing' or 'acting' in many feminist studies of technology. As Faulkner (2001) points out, there are some technologies that are 'barely gendered at all.' In addition, there is the possibility of flexible interpretation, which means that technology may be interpreted in new and diverse ways.

Other scholars have studied the role of technology as a part of the production of 'masculinities' (Hacker, 1990; Mellström, 2003; see also Faulkner, 2001 for an elaboration of the technologies role for men's gender identities). A key issue in this literature has been how technology plays an important role in men's lives and is used to improve and build their identities, for example within such fields as engineering, mechanics and computer science. Still, many of these analyses treat the masculine gender as static, and they seem to take-for-granted what is considered to be 'masculine' and 'feminine.'

In fact, these deficiencies may be more serious than they appear to be at first glance. While the tendency to treat gender as a static category in relation to technology may be explained

by a feminist tendency to identify and critique patriarchy, it may in fact be even more damaging to feminist goals. Reproducing stereotypical notions of men, women and technology by labeling them 'masculine' and 'feminine' tends to ignore the complexity of the relations within and between them and in turn diminishes the analytical potential for developing new and complex gender–technology hybrids.

What does it mean to treat inclusion and exclusion as points of departure when investigating gender and technology? Perhaps we should turn the question on its head and ask: What has been performed through feminist studies of technology? Undoubtedly, such studies have provided much better understanding of how gender and technology are coproduced in a variety of circumstances, but in practice they still treat technology as a man's domain and as a symbol of masculinity. It may be paradoxical that gender and technology have been massively studied, and yet not much has happened to change the paucity of women involved with technology – the generally low and even decreasing share of women who design and study technology, although there remains substantial number of women who uses technologies. Why has it been so difficult to change the gendering of technology? The modernist assumption usually is to treat critique as a basis for enlightenment: to identify problems as a basis for rational intervention.

As we saw from the review above, this has been a central concern for feminist studies of technology. The postmodern belief is that critique may inform, but the interpretation and action are stable and shifting, thus making intervention more difficult. The question though is: how does this work perform on the public? Giddens' (1976) double hermeneutic illustrates the interaction between those who study and those who are being studied, showing that such a dynamic makes knowledge (as well as action) unstable. So, what is performed by feminist studies of technology? One claim would be that gender and technology are being continuously coproduced by involved actors, and that the accounts of how technology is gendered is potentially fed back to the coproduction circle and also becomes part of it. We may ask whether this contributes to change the gender–technology relationship, or actually reinforce it?

The Interpretation of Numbers

With regard to the process of the gendering of technology, researchers tend to interpret activities and artifacts as masculine or feminine, either based on their appearance (e.g., the exhibitions in the Netherlands and Norway) or on whether men or women enact that process. In short, they assume that what women do is feminine, and what men do is masculine. Accordingly, shifts in the relative numbers of women and men will change the gendering of the activity (Lagesen, 2007; Sørensen and Berg, 1987). An example is the mobile phone, which changed from being masculine to 'unisex.' There are other examples of previously men-dominated areas, such as in medicine, in which an influx of women has made the activity's gendered identity less masculine.

Such examples show some potential for change. But none of these changes was a result of feminist studies of technology. Thus, we could ask: Do artifacts have gender – or do their gendered associations derive from feminist technology studies? In other words, how can we avoid performing a conserving, double coproduction of gender and technology? Research-based accounts of gender and technology/ICT may reinforce the problem we want to solve, in part due to the widespread engagement in unwarranted interpretations of assumed symbolic properties. Also, arguments playing on the concepts of feminine and masculine tend to reproduce binaries that should be questioned. What might we do instead? Are there other possibilities?

Perspectives on Inclusion

Some narratives or theoretical perspectives on gender and technology are based on an inclusive, rather than exclusive, logic. One such perspective has been referred to as *cyberfeminism*. Cyberfeminists tend to take more optimistic approaches to the relationship between gender and ICTs, and claim to observe progress in the inclusion of women in new technological arenas such as the Internet. While cyberfeminists sometimes acknowledge that there are forces that tend to exclude women they also claim that many women have become highly motivated and skilled 'cybercitizens.' Plant (2000) is one of the most optimistic cyberfeminists, in the way she argues that women, computers, virtual reality, and cyberspace are linked together in dispersed, distributed connections – a matrix, which, because of its inherent feminine character, will emerge as a new society that will destroy patriarchy. However, Plant's visions have been criticized as too extreme, and also as essentialist and for lacking critical perspective and even hyping new technologies by overstating the women-friendliness of cyberspace. As Wajcman (2004) has pointed out, cyberfeminism does not bring much to the table in terms of feminist politics, and the approach can be criticized for assigning too much agency to new technology and not enough to feminist politics. Are there other options?

A Fourth Inclusion Narrative: Material Feminism?

Mainstream feminist theory traditionally has not been much occupied with technology, and that topic almost exclusively has been a concern for feminist technology studies. Moreover, feminist technology studies have not been very influential with feminist theory, unlike feminist studies of science that have had more resonance with feminist theory building. Donna Haraway's cyborg feminism (1985) has perhaps been the most successful concept for drawing attention to technology as an integral part of feminism, and her work has been used to include a concern for technology in feminist theory (see e.g., Alaimo and Hekman, 2008). *Material feminism* is an emerging area of interest within feminism, which argues for taking matter and materiality more into account when studying gender (Barad, 2003). This emphasis on materiality draws theoretical initiative in part from science and technology studies, which emphasize nonhuman agency (Latour, 1987) as well as the 'trickster' quality and unpredictability of technology, materiality and nature (Haraway, 2004).

Haraway's (1985) cyborg feminism represents a way of thinking about gender and technology that has more potential

for inclusion. The cyborg figure represents an implosion of human and machine. This perspective suggests a focus on disruptions and ambiguities in relation to gender and technology. It advocates the exploration of the women-computing relationship, paying particular attention to the complexities and contradictions of this relationship, and makes gender less binary and essential than in the coconstruction perspective, which tends to focus on exclusions in the gender–technology relationship. But, given that cyborg feminism is a philosophical perspective, how can we make feminist politics out of it? Some have advocated that a version of actor-network theory (ANT) is a way to study gender and technology (Lagesen, 2012; Ormrod, 1994; Singleton, 1995). A main tenet of ANT is that society is an achievement of actors (human and nonhuman) engaging in producing a variety of associations among human and nonhuman elements, and that we, as researchers, should trace these associations by following the actors. Thus, there is potential in Latour's theory for sociologically studying the role of technology in studying how associations are assembled when people construct themselves and their actions. This resonates with the widely used feminist theory that is concerned with how people 'do gender' (West and Zimmerman, 1987). In both a 'doing' perspective and in actor-network theory, gender as well as technology become an outcome of, and not a starting point for, the analysis. Gender–technology hybrids thus can be seen as outcomes of processes of reassembling associations among human and nonhuman elements. How is this perspective useful for thinking about (and acting upon) inclusion of women into technology? Sørensen et al. (2011) explicitly try out such an inclusive perspective in their analysis of several case studies about women and ICTs. They are inspired by actor-network theory and argue that thinking about inclusion would mean asking questions about why women should be interested in technology (as opposed to why they are not). Focusing on inclusion rather than exclusion may provide better insight into how women-technology assemblages are initiated and stabilized, but also how they are destabilized (Sørensen et al., 2011), thus complicating and disrupting the gender binary. Also, it is obviously easier to design inclusion strategies and politics based on knowledge about what contribute to include women to technology, rather than on what excludes women.

See also: Cultural Studies of Science; Engineering Studies; Gender: Gendering of Categories; Science and Technology, Social Study of: Computers and Information Technology; Situated Knowledge, Feminist and Science and Technology Studies Perspectives; Standpoint Theory, in Science; Technology, Social Construction of.

Bibliography

Akrich, M., 1992. The description of technical objects. In: Bijker, W., Law, J. (Eds.), Shaping Technology/building Society: Studies in Sociotechnical Change. MIT Press, Cambridge, MA, pp. 205–244.

Alaimo, S., Hekman, S., 2008. Material Feminism. Indiana University Press, Bloomington and Indianapolis.

Barad, K., 2003. Posthumanist performativity: toward and understanding of how matter comes to matter. Signs 28 (3), 801–831.

Baxandall, R., Ewen, E., Gordon, L., 1976. The working class has two sexes. Monthly Review 28 (3), 1–10.

Berg, A.-J., 1996. Digital Feminism, vol. 28. Centre for Technology and Society, Norwegian University of Science and Technology, Trondheim. STS Rapport.

Berg, A.-J., Lie, M., 1993. Feminism and constructivism: do artifacts have gender? Science, Technology & Human Values 20 (3), 332–351.

Berg, A.-J., Lie, M., 1995. Feminism and constructivism: Do artifacts have gender? Science, Technology & Human Values 20 (3), 332–351.

Bittman, M., Mahmud Rice, J., Wajcman, J., 2004. Appliances and their impact: the ownership of domestic technology and time spent on household work. The British Journal of Sociology 55 (3), 401–423.

Bose, C.E., Bereano, P.L., Malloy, M., 1984. Household technology and the social construction of housework. Technology and Culture 25 (1), 53–82.

Braverman, H., 1974. Labor and Monopoly Capital: the Degradation of Work in the Twentieth Century. Monthly Review Press, New York.

Casper, M.J., Berg, M., 1995. Constructivist perspectives on medical work: medical practices and science and technology studies. Science, Technology & Human Values 20 (4), 395–407.

Clarke, A., 1988. Disciplining Reproduction: Modernity, American Life Sciences and the "Problems of Sex". University of California Press, Berkeley, CA.

Cockburn, C., 1981. The material of male power. Feminist Review 9, 41–58.

Cockburn, C., 1983. Caught in the wheels. Marxism Today 27, 16–20.

Cockburn, C., Fürst-Dilic, R. (Eds.), 1994. Bringing Technology Home: Gender and Technology in a Changing Europe. Open University Press, Buckingham.

Cockburn, C., Ormrod, S., 1993. Gender and Technology in the Making. Sage, London.

Cowan, R.S., 1983. More Work for Mother. Basic Books, New York.

Faulkner, W., Arnold, E., 1985. Smothered by Invention. Pluto Press, London.

Faulkner, W., 2001. The technology question in feminism: a view from feminist technology studies. Women's Studies International Forum 24 (1), 79–95.

Feldberg, R.L., Glenn, E.N., 1983. Technology and work degradation: effects of office automation on women clerical workers. In: Rotschild, J. (Ed.), Machina Ex Dea. Pergamon Press, New York.

Firestone, S., 1970. The Dialectic of Sex: The Case for Feminist Revolution. William Morrow & Co, New York.

Form, W., McMillen, D.B., 1983. Women, men and machines. Work and Occupations 10 (2), 147–178.

Frehill, L.M., 2004. The gendered construction of the engineering profession in the United States, 1893–1920. Men and Masculinities 6, 383–405.

Giddens, A., 1976. New Rules of Sociological Method. A Positive Critique of Interpretative Sociologies. Basic Books, New York.

Hacker, S., 1990. Doing It the Hard Way. Investigations of Gender and Technology. Unwin Hyman, Boston.

Haraway, D., 1985. Manifest for cyborgs: science, technology and socialist feminist in the 1980s. Socialist Review 80, 65–108.

Haraway, D., 2004. The Haraway Reader. Routledge, New York and London.

Hartmann, H., 1979. The unhappy marriage of marxism and feminism: towards a more progressive Union. Capital and Class 3 (2), 1–33.

Lagesen, V.A., 2007. The strength of numbers: strategies for including women into computer science. Social Studies of Science 37 (1), 67–93.

Lagesen, V.A., 2012. Reassembling gender: actor-network theory (ANT) and the making of the technology in gender. Social Studies of Science 42 (3), 442–448.

Latour, B., 1987. Science in Action. Open University Press, Milton Keynes.

Lie, M., Sørensen, K.H., 1996. Making Technology Our Own? Domesticating Technology into Everyday Life. Scandinavian University Press, Oslo.

Margolis, J., Fisher, A., 2002. Unlocking the Clubhouse. Women in Computing. The MIT Press, Cambridge, MA.

McIlwee, J.S., Robinson, J.G., 1992. Women in Engineering: Gender, Power, and Workplace Culture. State University of New York Press, Albany, NY.

Mellström, U., 2003. Masculinity, Power and Technology. A Malaysian Ethnography. Ashgate, Hampshire.

Merchant, C., 1980. The Death of Nature: Women, Ecology and the Scientific Revolution. Harper and Row, New York.

Mies, M., Shiva, V., 1993. Ecofeminism. Zed Books, London.

Oudshoorn, N., Sætnan, A.R., Lie, M., 2002. On gender and things: reflections on an exhibition on gendered artifacts. Women's Studies International Forum 25 (4), 471–483.

Ormrod, S., 1994. "Let's nuke the dinner": discursive practices of gender in the creation of a new cooking process. In: Cockburn, C., Fürst-Dilic, R. (Eds.), Bringing Technology Home: Women, Gender and Technology in Europe. Open University Press, pp. 42–58.

Phillips, A., Taylor, B., 1980. Sex and skill. Feminist Review 6, 79–88.

Plant, S., 2000. On the matrix: cyberfeminist simulations. In: Kirkup, G., Janes, L., Woodward, K., Hovenden, F. (Eds.), The Gendered Cyborg: A Reader. Routledge, London.
Singleton, V., 1995. Networking constructions of gender and constructing gender networks. In: Grint, K., Gill, R. (Eds.), The Gender/technology Relation. Taylor & Francis, London, pp. 146–173.
Sørensen, K.H., 1992. Towards a feminized technology? Gendered values in the construction of technology. Social Studies of Science 22 (1), 5–31.
Sørensen, K.H., Berg, A.-J., 1987. Genderization of technology among Norwegian engineering students. Acta Sociologica 2, 151–171.
Sørensen, K.H., Faulkner, W., Rommes, E., 2011. Technologies of Inclusion. Gender in the Information Society. Tapir Academic Press, Trondheim.
Thompson, C., 2005. Making Parents. the Ontological Choreography of Reproductive Technologies. MIT Press, Cambridge.
Thrall, C.A., 1982. The conservative use of modern household technology. Technology and Culture 23 (2), 175–194.
Vanek, J., 1974. Time spent in housework. Scientific American 231, 116–120.
Wajcman, J., 1991. Feminism Confronts Technology. Polity Press, Oxford.
Wajcman, J., 2004. Technofeminism. Polity Press, Cambridge.
West, C., Zimmerman, D.H., 1987. Doing gender. Gender & Society 1, 124–151.

Gender and the Law

Martha Chamallas, The Ohio State University, Moritz College of Law, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by K. Plett, volume 9, pp. 5980–5984, © 2001, Elsevier Ltd.

Abstract

This article chronicles the various meanings of 'gender and the law' from nineteenth century 'separate spheres' ideology to present-day legal interventions related to women's rights, sexuality, and LGBT (lesbian, gay, bisexual, transgender) issues. Tracking the different forms of injustice that feminist legal reformers have confronted during the period, it defines and illustrates struggles over gender difference, subordination, devaluation, essentialism, victimization, and normalization. It contains an overview of the major strands of legal feminism, explaining the attributes of liberal, dominance, and cultural feminism as well as newer postmodern theories and theories that focus on intersecting forms of bias, including race, ethnicity, class, and sexual orientation.

Varied Meanings of 'Gender and the Law'

The concept of gender and its relationship to law has fluctuated over time and in different contexts. It may refer to a distinct set of legal rules and responsibilities governing men and women. Use of the term 'gender' (rather than 'sex') often signals that the topic extends beyond biological differences between men and women and includes the legal significance of gender roles, gender norms, and common understandings of masculinity and femininity. As a category, 'gender and the law' has been tied to sexuality and personal identity, also covering laws relating to sexual orientation and gender identity, as well as legal responses to gender 'performances' that turn on how individuals exhibit gender through their appearance, speech, dress, and other outward manifestations.

Separate Spheres and Exclusion

Until the latter part of the twentieth century, one's gender (i.e., one's status as a man or a woman) was a crucial determinant of formal legal rights. Prior to that time, the law largely embraced a 'separate spheres' ideology that relegated women to the private sphere of the home and family, while men dominated the public spheres of work, politics, and intellectual life. 'Separate spheres' reflected the dominant culture's ideal of woman's role, even though it did not match the reality of the lives of many women – particularly women of color and working-class women – who often worked outside the home to support themselves and their families (Kerber, 1988). In tune with the dominant ideology, the law was riddled with explicit sex-based distinctions and classifications, often grounded in stereotypes about the different roles, abilities, and interests of the sexes. Many of the stereotypes attached to women and femininity were disparaging and depicted women as unsuitable for numerous lines of work and most professions. Women were presumed to be weaker, fragile, more emotional, less rational, less competitive, and poor at math and sciences. Accordingly, many elite educational institutions refused to admit women, followed strict quotas, or had more stringent admission requirements for women (Murray, 1971). The separate legal treatment of women, however, was often couched in benign terms and justified by the need to 'protect' women from the harsh realities of public life. During this time, gender determined not only who received financial support after divorce (only women) or who would be conscripted into the military (only men), but also virtually every facet of life subject to legal regulation (Chamallas, 2012).

Marriage, in particular, was a highly gendered legal institution. The very definition of marriage was gender-specific, that is, a union between a man and a woman, excluding same-sex couples from entering into legally recognized marriages. Although by the mid-nineteenth century married women no longer lost their legal identity upon marriage (as they did under earlier 'coverture' or 'marital unity' doctrines), the rights and duties of husbands and wives differed sharply. Men were designated the 'head of the household,' with an obligation to provide financial support to their wife and children. By contrast, women were legally obligated to provide care and household services to their children and their husband. Married women's economic dependence, whether presumed or actual, often meant that they were required to obtain their husband's consent to secure credit or loans or to engage in other financial transactions. The law privileged the institution of marriage by punishing those who had sex outside of marriage, through criminal statutes prohibiting fornication, sodomy, and adultery. Additionally, broad legal restrictions on the sale and distribution of contraceptives and on abortions left women with little ability to control their fertility and the size of their families.

During this period, women were often excluded from or had only a token presence in legal institutions. Very few women attended law school, practiced law, or even served on juries. The proportion of female judges and legislators was also exceedingly small, meaning that men were largely responsible for writing and interpreting the law.

Equal Access and Formal Equality

Although women were afforded the right to vote in the United States (1920) and in most European nations in the early part of the twentieth century (Switzerland was a notable latecomer,

extending the franchise to women only in 1971 (with the last Swiss canton, Appenzell Innerrhoden, allowing women the right to vote at canton-level as late as 1990).), widespread changes in the law affecting women's legal status did not emerge until the latter part of the twentieth century, on the heels of the Second Wave feminist movement. Women's rights advocates in the 1970s emphasized women's similarity to men and argued that because women were the same as men in all relevant respects, they deserved access to public institutions, benefits, and opportunities on the same terms as men. Early reform efforts concentrated on dismantling the intricate system of 'protectionist' sex-based legal distinctions, with an eye toward integrating women into formerly male-dominated domains, including blue-collar jobs, the professions, the sciences, and elite universities. The prominent 'liberal feminist' discourse of the period often centered on individual rights, with the central message that gender should be irrelevant to the distribution of legal benefits and burdens and that persons should be treated as individuals, not as members of a class.

The campaign against the use of explicit gender classifications met with early success in the United States, prompting the US Supreme Court to invalidate a wide variety of gender lines in statutes, for example, striking down provisions that presumed that wives, but not husbands, were economically dependent on their spouses for support. These reforms are commonly referred to as having established a legal system of formal equality, requiring that the form of the law (the law on its face) be gender neutral (Franklin, 2010). Pushing back against separate spheres ideology, another theme of the time was that marriage and motherhood should not be regarded as compulsory for all women. Accordingly, courts began to recognize a woman's right to choose to terminate a pregnancy grounded on notions of individual privacy and autonomy. One enduring legacy of the era was to target difference or the differential treatment of the sexes as a touchstone of inequality, marking as suspicious any scheme that provided one set of rules for women and another set of rules for men.

Substantive Equality and Women's Subordination

By the latter part of the twentieth century, it became clear that formal legal equality would do little to change the lives of most women, as evidenced by such phenomena as the feminization of poverty, working mothers' second shift, and the 'glass ceiling.' To uncover the material or substantive causes of women's inequality, feminist theorists began to revise the concept of 'equality' to mean something other than identical treatment of men and women. They argued that if men and women did not start from the same position, identical treatment of each group might never produce meaningful equality. Feminists of this era targeted implicit male norms underlying facially neutral laws, frequently objecting to standards or rules designed around male needs, male social biographies or male life experiences. They shifted attention to areas in which women's situation differed most from men's, focusing on the legal treatment of pregnancy, rape, sexual harassment, domestic violence, and pornography. The rape law reform movement, for example, targeted male-centered definitions of 'consent' that presumed consent unless the rape victim offered physical resistance. Another target of attack was the common practice of refusing to prosecute date and acquaintance rapes in the absence of proof of overpowering physical force.

Despite their attention to gender difference, most feminist theorists of the time were social constructionists, locating the origin of gender differences in cultural attitudes, ideology, discourse, socialization, or organizational structures. In particular, the influential writings of Catharine MacKinnon sought to explain how the legal system was fundamentally opposed to women's interests and designed to perpetuate male dominance. MacKinnon emphasized the difference in power between men and women and contended that the sexual use and abuse of women was the principal mechanism by which women as a class were subordinated in society (MacKinnon, 1989). Her radical brand of feminism (sometimes called 'dominance feminism') was a driving force behind campaigns in the 1980s aimed at eliminating 'violence against women,' a construct used to tie together many forms of eroticized dominance of women. More recently, the antisubordination framework has gained traction globally, generating new understandings of international law, and new prohibitions against sex trafficking and mass rapes in war-torn and conflict-ridden societies (MacKinnon, 1994). The influence of MacKinnon can also be seen in changes in domestic laws, with the dramatic increase in the number of countries enacting laws against sexual harassment, such as in Israel, which passed comprehensive landmark legislation in 1998.

In contrast to earlier individual rights strategies, the dominance feminist emphasis on subordination is a group-based conception that measures injustice by whether a practice or policy deepens power disparities between men and women in society. Akin to the notion of caste in race, gender-based subordination frequently entails processes of sexual objectification that operate to silence women's perspectives and women's ability to discover and define their own desires and objectives. Somewhat paradoxically, given their deep-seated criticisms of law and legal institutions, antisubordination advocates have often favored legal interventions to curb sexual violence and have proven adept at institution building at both the national and international levels.

Devaluation, Caregiving, and the Work/Family Conflict

Another influential strand of legal feminism that first emerged in the 1980s proceeded from an assumption of gender difference but regarded difference in a positive light. The starting point for these cultural feminists (sometimes called 'relational feminists') is the distinctive ways women tend to approach problems, make moral choices, view the world, and construct their identity. Catalyzed by Carol Gilligan's popular book, *In a Different Voice*, feminists with this theoretical orientation celebrated the distinctive contributions women made to society, singling out women's capacity for nurturing, empathy, and preservation of human relationships (Gilligan, 1982). In contrast to liberal feminism's de-emphasis of the mothering role, cultural feminists supported the maternal role and other traditional activities associated with women.

The interdisciplinary literature in this vein drew a contrast between an ethic of care and connection, most often associated with women, and an ethic of justice and autonomy, most often associated with men. Rather than prioritize the rights of the various parties in a hierarchical fashion, the care 'voice' was more interested in finding solutions that would cause the least harm to all involved. Using the image of the 'web' to portray the cultural feminist approach, writers contrasted it to the image of the 'ladder,' which captured a more masculine mode of reasoning. Eschewing abstraction, proponents preferred to examine the particulars of interpersonal situations and to approach disputes contextually. The pioneers of the emerging field of alternative dispute resolution used cultural feminist theory to place greater emphasis on cooperation, preservation of relationships, and attention to process in an effort to reform the adversarial legal system and produce more positive outcomes for clients (Menkel-Meadow, 1989).

Beyond acknowledging differences between men and women, this vision of women's culture argues for a revaluation of the activities and attributes commonly associated with women and femininity. It is linked to the concept of devaluation, identified as the primary source of gender injustice under this orientation. Devaluation refers to the cultural assignment of a lower value to an entire category or class of activities cognitively associated with women, such as the devaluation of nursing, caregiving, or feminine behavior. In contrast to difference, devaluation does not typically involve comparing one individual to another 'different' individual, or comparing the treatment of different social groups. Although devaluation may operate to construct a hierarchy, in that a lower value is assigned to a given category or activity, marking it as inferior or deficient, it often operates to disadvantage individual members of privileged groups, such as a man who is a primary caretaker of young children (Chamallas, 2001).

The effects of gender devaluation have been detected in various sectors. With respect to employment, feminist reformers argued that pay scales in predominantly female jobs and occupations were unjustifiably low, reflecting gender bias in evaluating the relative worth of such jobs, given the skill, effort, and responsibility involved. The movement for 'comparable worth' – with the stated goal of upgrading compensation in female dominated jobs – stalled in the United States, but many other industrialized countries have instituted some sort of comparable worth policy, including the European Union, Canada, and Australia (Bowman et al., 2011). Similar efforts have been made to extend pay equity protection to part-time workers, a female-dominated group made up largely of mothers who are often denied valuable fringe benefits, even on a pro rata basis. While the European Court of Justice has ruled in favor of part-time workers' claims, US courts have rejected similar challenges, relying on the formal equality justification that male and female part-time workers are treated the same.

Women's groups have tried to ease the work/family conflict of working mothers by addressing the devaluation of mothering and caregiving activities. Public support for caregiving activities, including paid parental leave, is now the norm in France and other European nations, while the financial burden for caring for children in the United States remains privatized, resting largely on individual parents. In a new species of 'family responsibility discrimination' lawsuits, litigators in the United States have sought to convince courts that workplace discrimination against mothers and other caregivers is a form of forbidden stereotyping that unfairly labels such persons as less competent and less committed to their jobs (Williams and Bornstein, 2008). Because of the perceived inadequacy of laws addressing pregnancy discrimination, many scholars in the United States have pressed for enactment of laws requiring employers to offer accommodations, such as light-duty assignments, to allow pregnant women to continue working during their pregnancy. Such efforts in the United States have generally been stymied, with the legal fixation with formal equality getting in the way of change. Although the term 'devaluation' is now commonly used to signal bias, it has yet to acquire a precise legal meaning and devaluation remains a kind of injury the law has difficulty comprehending.

Complex Identities, Essentialism, and Intersectionality Theory

Since the 1990s, the theme of diversity among women and the complexity of women's identities has captivated gender and law scholars. Earlier scholarship was often preoccupied with making comparisons between men and women, displaying a tendency to overplay the differences between men and women and to underplay the differences among women. Contemporary feminist scholarship has attempted to correct for these overgeneralizations by highlighting the social and theoretical importance of differences among women, most prominently those of race, class, ethnicity, sexual orientation, disability, and immigration status. The salience accorded to differences other than gender has led to the development of new theories about the interaction of various kinds of oppression in society, referred to as postessentialist, anti-essentialist, critical race feminist, or intersectional approaches.

Intersectional legal scholars – most prominently, Kimberlé Crenshaw in the United States – objected to the use of exclusionary terms and techniques that focused attention on more privileged subgroups (e.g., white, middle-class women) while obscuring the plight of marginalized subgroups (e.g., African-American women) (Crenshaw, 1989). In law, the tendency to frame litigation as if 'race' and 'gender' were mutually exclusive categories has meant that women of color were often forced to 'split their identities' and required to choose between asserting a sex discrimination claim or a race discrimination claim, obscuring important features of the case that did not fit within the chosen framework. Intersectional theorists have called for recognition of distinctive forms of discrimination against subgroups of women that arise at the intersection of race and gender, observing that experiences of racism are often shaped by gender and experiences of sexism are often shaped by race. They assert that it is impossible to separate the racial and gender aspects of one's identity, preferring to describe the self in 'both/and' (rather than 'either/or') terms, in order to capture the full extent and meaning of individual and group experiences. For intersectional theorists, both personal identity and the multiple types of discrimination confronting different subgroups is not simply an additive phenomenon (e.g., race discrimination plus sex discrimination), but an interactive

process, that, for example, gives rise to distinctive brands of racialized sexism.

The intersectional approach is alert to the danger of regarding women as a monolithic group and assuming that all women share certain essential common experiences (Barlett and Rhode, 2006). Gender essentialism refers to the tendency to presume commonalities among all groups of women and to mistake the part for the whole by, for example, making sweeping statements or generalizations about women based on the situation of the most privileged subgroup of women. In many contexts, essentialism exaggerates the importance of gender bias as compared to other forms of discrimination and subordination. When unaddressed, essentialism can tarnish feminist-inspired reforms, rendering them inadequate to meet the needs of large numbers of women. In the United States, intersectional critics have charged that the standard liberal feminist analysis of separate spheres ideology failed to recognize that feminine traits of passivity, dependence, and fragility were culturally associated with white women only, missing the stereotyping of black women as too domineering, rather than too passive and weak. Similarly, they charged that dominance feminism failed to analyze black women's distinctive experience of rape and mistakenly assumed that black women's experience was merely a more intense version of white women's experience (Harris, 1990).

Intersectional scholars have also critiqued the form of essentialism that tends to set Western feminism as the standard for judging women's condition throughout the world. They maintain that the widespread belief that the 'West is the best' creates a danger of misjudging a society from the outside and is too quick to condemn unfamiliar practices associated with the Third World and immigrant populations. Western observers may thus be predisposed to attribute sex-subordinating practices in other countries as stemming from their culture, while regarding incidents of sexual violence in the West as aberrant, reflective of the behavior of deviants rather than constituting part of Western culture itself. The tendency to see 'culture' only in others has the dual negative effect of portraying Third World women as helpless victims of their culture and obscuring the structural and political causes of sex-subordinating practices in other cultures (Volpp, 1996).

Recent intersectional feminist scholarship has taken a comparative turn, investigating high-profile gender and law controversies around the globe. Scholars often aim to understand how women negotiate their multiple identities in different legal and political environments, for example, the gendered struggles over laws restricting the wearing of headscarves ('hijabs') in France and Turkey. In France, Muslim girls and young women have acted against a backdrop in which the predominant French view regards the hijab as a symbol of female subjugation, while the predominant view among Muslim religious authorities considers the hijab to be a symbol of female modesty and religious fidelity to Islam. Between these countervailing forces, Muslim girls and women have defied legal prohibitions against wearing hijabs, choosing to affirm a complex identity that rejects the opposing subjugation/modesty labels and exhibits pride in being both French and Muslim (Wing and Smith, 2006). In Turkey, women found themselves caught in a masculine power struggle between the secular government and male Islamist leaders, struggling to express an identity that opposes the state's compulsory deveiling of women, yet resists Islamist demands to cover and lead a traditional, private life. To disrupt the dominant narratives, they have refashioned the headscarf, pairing it with form-fitting, fashionable jackets and raincoats to signal a complex identity that is both modern and religious (Vojdik, 2010).

Activists have embraced intersectional theory as a tool of social mobilization and to influence public policy. Organizing under the banner of promoting 'social justice,' grassroots groups linked feminism to other movements for social justice in the United States and globally, targeting oppression of social groups and larger systems of inequality and concentrations of wealth (Kalsem and Williams, 2010). In the United States, a reproductive justice movement, which appeals to younger women, has reframed the debate from a narrow right to choose to a more material and affirmative conception of women's reproductive health and freedom. Reproductive justice advocates urge increased access to birth control and abortion, funding for prenatal care, support in childbirth and postpartum, support for breast-feeding mothers, availability of child care for infants and toddlers, income support for caregivers, and high-quality public school education (West, 2009).

While still in the early stages, intersectional theory and its activist offshoots have demonstrated a capacity to forge coalitions among progressives and to create some of the necessary conditions for broader social and legal change. One drawback is that intersectional scholarship is not always 'coded' as 'gender and the law' scholarship. Instead, it is often categorized as scholarship relating to race, ethnicity, or social class (but not gender), missing a key point of intersectionality theory and setting it apart from other developments related to women and sexuality.

Agency, Sex-Positive Feminism, and Victimization

In the twenty-first century, many feminist writers have recoiled against stereotyped portrayals in the popular culture and mass media of women as passive sexual objects, perpetually vulnerable to violence or exploitation. Placing the blame partially on the influence of Catharine MacKinnon's brand of dominance feminism, critics have sought to redirect the focus away from sexual violence and what they regard as a fixation on women's victimization. Paralleling the move to study intersectional and multiple forms of oppression, revised accounts of the complex nature of personal identity seek to present women as simultaneously victims of oppression and agents of their own destiny. There is a renewed emphasis on women's 'agency,' meaning the capacity to direct one's own life through individual action and choice. Related to the concept of autonomy in liberal theory, the concept of agency highlights women's ability to resist, make decisions, and contribute to the cultural meaning of gender in society.

One strand of the new autonomy-focused feminism, called 'sex-positive' feminism, has set out to reclaim the pleasures of sex and sexual agency for women and to distance itself from accounts of sex as an inevitable site of danger for women. It stresses the importance of female sexual desire and pleasure

and charges that dominance feminist accounts of sex as dangerous for women have played into the hands of social conservatives who continue to regard sex as something to be indulged in only for purposes of reproduction. One prominent writer remarked that feminists have gotten stuck in 'No' and have yet to theorize why women say 'Yes' to sex (Franke, 2001). Sex-positive feminism resonates with a younger, 'third-wave' generation of feminists who often embrace a 'fun, playful aesthetic,' that finds joy in 'make up, feminine styles of dress, and traditional "girlie" behavior' (Crawford, 2007). Third-wave writers are more apt to cast pornography in a positive light, arguing that pornography is not invariably bad for women. In this view, pornography allows some women to connect to their sexual feelings and acquires meaning only through the person who views it.

Other critics of victimization concentrate on complicating the account of women's agency beyond the sexual sphere. They have highlighted the contexts in which women are often required to make strategic choices – and exercise what is called 'partial agency' – against a backdrop of subordinating structures and ideologies (Abrams, 1995). Trying to break through the victimization/agency dichotomy, legal practitioners have developed concepts, such as 'separation assault,' to provide a fuller description of the situation of women who are battered by spouses or other intimates. They have explained how batterers often intensify their violence to prevent a woman from leaving or to retaliate against her for leaving. Such accounts challenge the stereotype of the 'battered woman' who suffers from 'learned helplessness' and is too terrified to act. They emphasize the strategic choices women make when they decide whether to stay in or to leave a relationship (Mahoney, 1991).

Critics of victimization are often wary of the benefits of legal intervention. They advocate pulling back from the law, for example, the deregulation of pornography or the elimination of mandatory arrest and no-drop laws that disregard the victim's wishes in domestic violence cases. The emergence of such new forms of legal feminism has deepened the split between legal feminists who are ready to enlist law to further a feminist agenda and those who mistrust the power of law and the state and are skeptical of legal entanglements.

Postmodernism, Identity Performance Theory, and Normalization

Lagging somewhat behind their counterparts in other disciplines, contemporary legal feminist scholars have generated a body of scholarship that bears the hallmarks of the larger intellectual trend known as postmodernism. Postmodern legal feminists explore gender not as an essence but as a social process, specifically, the process of assigning meaning and significance to gender differences. Postmodernists place paramount importance on language, discourse, and habitual social practices to explain how gender differences and gender deviancy are constructed in a given culture. In law, they have questioned the conventional legal understanding of gender identity as an ascribed or fixed status, focusing instead on the 'performative' dimensions of a person's identity. They examine, for example, how individuals present their 'difference' in a variety of settings through dress, language, personal style, and everyday behaviors.

This fluid, non-status-based account of identity multiplies the range of identities a person may project, making it possible to speak of an individual as being more 'black' or more 'gay' in certain contexts. In contrast to earlier brands of legal feminism, postmodern legal feminism tends to shift focus away from the existence and maintenance of gender hierarchies to analyze the sex-based categories themselves, particularly the dominant conception that the sexes are opposite and that there are only two sexes, that is, men and women. They contend that instead of being natural, binary conceptions of sex and gender are socially constructed, with individuals being conditioned through everyday social practices to perform their gender in binary ways. Scholars have explored the 'policing' of the boundaries between the genders, such as sex segregation and the stigmatization of individuals who do not conform to dominant gender norms. Postmodern feminists typically embrace the interrelated goals of destabilizing or disrupting traditional meanings of sex and gender, resisting the binary, and promoting and fostering gender nonconformity.

The postmodern emphasis on 'performing' one's identity has led to a deeper understanding of the dynamics of intra-group bias, a complicated process by which only certain members of a particular group or subgroup are singled out for adverse treatment, while others are able to rise to the top. Thus, an individual may be targeted for discrimination because she performs her identity in ways that call attention to her difference and make those in power feel uncomfortable, for example, the black female employee who wears braids or African-style clothing or socializes mainly with other black women (Carbado and Gulati, 2000). Legal scholars have criticized the demands placed on individuals to assimilate to the dominant norms of an institution or a community, characterizing such demands to 'cover' or downplay one's identity as a central mechanism of discrimination in the twenty-first century (Yoshino, 2002). Performance theorists have also documented ways that minorities and other 'outsiders' are called upon to enact 'racial comfort strategies,' gestures, and behaviors designed to make insiders feel at ease, imposing a special burden that results in extra work, anxiety, and resentment.

Performance theory has provided a valuable tool of analysis in 'masculinities' studies, a subgenre of scholarship of particular interest to legal feminists. Masculinities theorists investigate the multiple ways masculinity is socially constructed in different contexts, focusing especially on how men perform their identity (Cooper and McGinley, 2012). Although most such scholars accept that men as a group exercise power over women as a group, they are most interested in understanding how men achieve and retain power and how power differentials among men can harm some men as well as women. The operative term ('masculinities') is deliberately plural to signify that different groups of men perform their masculinity differently. Masculinities scholars see the diversity of male gender performances as partly a product of differential power, using the term 'hegemonic masculinity' to describe the ideal of masculinity traditionally associated with white, Christian, straight, upper class men. The fundamental tenets of masculinities theory reveal a close connection to

legal feminism, particularly the belief that dominant masculinity norms send the message that masculinity is to be privileged over femininity. Similar to postmodern feminists, they regard masculinity not as an essence but as an ideal that can never be finally possessed or achieved, causing men to be constantly anxious about their masculinity and leading them to engage in aggressive, sometimes destructive, masculinity contests. A recurring feature of men's identity performances is the need to denigrate 'contrast figures,' most notably women, gay men, and racial minorities. Built into the struggle among men is thus an imperative to disparage others, not necessarily to inflict harm on those others, but to win the intramale competition for power and esteem.

These postmodern versions of scholarship have identified normalization as the key problem to address. In this context, normalization refers to discourses and ideologies that simultaneously privilege socially 'acceptable' or 'normal' practices and ideas and marginalize and discredit the 'deviant' lives of nonconforming individuals. Postmodern legal writers typically express concern that even when law is invoked by reformers bent on change, the normalizing discourse of law carries the potential to backfire and reinforce, rather than subvert, dominant ideologies. For this reason, many postmodernists have decidedly deregulatory impulses, often presuming that the law will not forego its reliance on status-based identities and that even feminist-inspired legal reform will cause negative unintended consequences.

Lesbian, Gay, Bisexual, and Transgender Scholarship, Queer Legal Theory, and the Transgender Movement

For the past several decades, legal feminism has been closely tied to scholarship focused on lesbian, gay, bisexual, and transgender (LGBT) issues, a literature that has grown exponentially with the rise of the marriage equality movement. Particularly now that 16 nations and 13 US states permit same-sex couples to marry (writing in August 2013), the field has expanded beyond analysis of specific legal restrictions facing LGBT individuals to development of complex theories that map and explain the nature of cultural forces and discourses inhibiting change. Reflecting the fast-changing developments on the ground, LGBT scholars have struggled with and debated four key constructs that lie at the core of their inquiries: sex, gender, sexual orientation, and gender identity. Early LGBT theorists warned against conflating the first three categories, insisting that one should not presume that a person's sex (i.e., biological sex) determined his or her social style (gender) or choice of sexual partner (sexual orientation) (Valdes, 1995). Until the recent transgender movement, however, little attention was paid to 'gender identity,' defined by the medical profession as a deeply felt psychological sense of sex that departs from the sex assigned to a person at birth. Transgender (or 'trans') scholars have since argued for a more fluid concept of gender identity that encompasses not only transsexuals desirous of sex-change operations and medical interventions, but also a broader class of gender variant or gender nonconforming individuals whose gender identities may shift over time and who may even identify as neither male nor female (Vade, 2007).

Queer legal theory is a popular new strain of scholarship that rejects identity-based categories, such as heterosexual and homosexual, and envisions freeing people from the constraints of identity. Like other performance theorists, queer scholars often shift the focus from what individuals 'are' to the various ways individuals 'do' or 'perform' their sexuality or gender. In its reclaimed form, the term 'queer' signals a disdain for, and opposition to, dominant norms of sexuality, gender, intimacy, and kinship, often placing queer theorists against marriage and conventional family formations (Romero, 2007). Many queer theorists profess a positive stance on sex and are critical of laws or legal reforms that address sexual conduct as a moralistic restriction. This prosex commitment often leads them to affirm expressions of sexuality in all its forms and to value individual freedom over equality. Some queer critics have gone so far as to distrust legal parties' claims of sexual injury in sexual harassment or rape cases (Halley, 2002). Queer theorists' investments in sexual pleasure, including the 'dark' pleasures of sexual domination and submission, have often pitted them against dominance feminists who have used the law to eradicate exploitive or nonmutual sexual practices. This tension threatens to sever the historic alliance between feminist and LGBT constituencies (McCluskey, 2009).

Conclusion

Since the latter twentieth century, theorists and activists who operate under the banner of 'gender and the law' have identified a variety of different 'enemies' as the principal source of injustice facing women and other marginalized people. Early efforts called for eliminating gender differences between men and women, attacking women's subordination, and addressing the devaluation of women's culture and activities commonly associated with women. These foundational theories primarily compared men to women and were aimed at promoting women's equality, albeit defined in different ways.

In contrast, contemporary theories have embraced a more complex view of gender and its relationship to law. They recognize that women are a highly diverse group whose lives are affected by other aspects of their identity, that women are not only victims but exercise agency and choice, and that gender is a fluid concept that assigns meaning to myriad social practices. These newer theories have targeted new 'enemies,' cautioning against essentialist discourses that presume that all women share common experiences or that trade on simplistic portraits of women's victimization. Influenced by postmodernism, many recent writers have reconfigured the field of gender and law away from an exclusive focus on women to a new terrain that critiques normalizing discourses and ideologies that limit the freedom of nonconforming individuals whatever their gender. While the complex theories of the twenty-first century have enriched the field of study, they have also uncovered new tensions, particularly among those who advocate using law as a vehicle for change and those who distrust legal intervention and fear the unintended consequences of legal reforms.

See also: Critical Race Theory; Discrimination and the Law; Equality and Inequality: Legal Aspects; Family Law in the United States; Feminist Legal Theory; Queer Theory and Intersectionality; Race and the Law; Sexual Orientation and The law.

Bibliography

Abrams, K., 1995. Sex wars redux: agency and coercion in feminist legal theory. Columbia Law Review 95, 304.

Barlett, K., Rhode, D., 2006. Gender and Law: Theory, Doctrine, Commentary, fourth ed. Aspen Publishers, New York.

Bowman, C., Rosenbury, L., Tuerkheimer, D., Yuracko, K., 2011. Feminist Jurisprudence: Cases and Materials, fourth ed. Thompson Reuters.

Chamallas, M., 2001. Deepening the legal understanding of bias: on devaluation and biased prototypes. Southern California Law Review 74, 747.

Chamallas, M., 2012. Introduction to Feminist Legal Theory, third ed. Wolters Kluwer, New York.

Carbado, D., Gulati, M., 2000. Working identity. Cornell Law Review 85, 1259.

Cooper, F., McGinley, A., 2012. Masculinities and the Law: A Multidimensional Approach. New York University Press, New York.

Crawford, B., 2007. Toward a third-wave feminist legal theory: young women, pornography and the praxis of pleasure. Michigan Journal of Gender & Law 14, 99.

Crenshaw, K., 1989. Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago Legal Forum, 139.

Franke, K., 2001. Theorizing yes: an essay on feminism, law and desire. Columbia Law Review 101, 181.

Franklin, C., 2010. The anti-stereotyping principle in constitutional sex discrimination law. New York University Law Review 85, 83.

Gilligan, C., 1982. In a Different Voice: Psychological Theory and Women's Development. Harvard, Cambridge, MA.

Halley, J., 2002. Sexuality harassment. In: Brown, W., Halley, J. (Eds.), Left Legalism/Left Critique. Duke University Press.

Harris, A., 1990. Race and essentialism in feminist legal theory. Stanford Law Review 42, 581.

Kalsem, K., Williams, V., 2010. Social justice feminism. UCLA Women's Law Journal 18, 131.

Kerber, L., 1988. Separate spheres, female worlds, woman's place: the rhetoric of women's history. The Journal of American History 75, 9.

MacKinnon, C., 1989. Toward a Feminist Theory of the State. Harvard, Cambridge, Massachusetts.

MacKinnon, C., 1994. Rape, genocide, and women's human rights. Harvard Women's Law Journal 17, 5.

Mahoney, M., 1991. Legal images of battered women: redefining the issue of separation. Michigan Law Review 90, 1.

McCluskey, M., 2009. How queer theory makes neoliberalism sexy. In: Fineman, M., Jackson, J., Romero, A. (Eds.), Feminism and Queer Legal Theory: Intimate Encounters, Uncomfortable Conversations. Ashgate Publishing, Förlag.

Menkel-Meadow, C., 1989. Exploring a research agenda of the feminization of the legal profession: theories of gender and social change. Law and Social Inquiry 14, 289.

Murray, P., 1971. Economic and educational inequality based on sex: an overview. Valparaiso University Law Review 5, 237.

Romero, A., 2007. Methodological descriptions: "feminist" and "queer" legal theories. Yale Journal of Law and Feminism 19, 227.

Vade, D., 2007. Expanding gender and expanding the law: toward a social and legal Conceptualization of gender that is more Inclusive of transgender people. Michigan Journal of Gender & Law 11, 253.

Valdes, F., 1995. Queers, sissies, dykes, and tomboys: deconstructing the conflation of "sex," gender," and "sexual orientation" in Euro-American law and society. California Law Review 83, 1.

Vojdik, V., 2010. Politics of the headscarf in Turkey: masculinities, feminism, and the construction of collective identities. Harvard Journal of Law and Gender 33, 661.

Volpp, L., 1996. Talking "culture": gender, race, nation and the politics of multiculturalism. Columbia Law Review 96, 1573.

West, R., 2009. From choice to reproductive justice: de-constitutionalizing abortion rights. Yale Law Journal 118, 1394.

Williams, J., Bornstein, S., 2008. The evolution of "FReD": family responsibilities discrimination and developments in the law of stereotyping and implicit bias. Hastings Law Journal 59, 1311.

Wing, A., Smith, M., 2006. Critical race feminism lifts the veil? Muslim women, France and the headscarf ban. UC Davis Law Review 39, 743.

Yoshino, K., 2002. Covering. Yale Law Journal 111, 769.

Gender and Women's Studies, Applied Research on

Mitchell Brown and Regina Moorer, Auburn University, Auburn, AL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Gender and women's studies refers to scholarship about and the academic study of the meaning and implications of gender and gendered roles across society, including a focus on power, oppression, and the intersection of identities. As an interdisciplinary field, it draws upon, addresses, and adds to topics covered by all of the hard and social sciences and the arts. In this article we discuss the development of women's studies as an academic field; its intellectual backgrounds and offshoots; the current state of the discipline; current academic programs, methodological issues and problems; and emerging directions in theory and research.

Gender and women's studies refers to scholarship about and the academic study of the meaning and implications of gender and gendered roles across society, from arts and culture, to politics and the economy, to education, to philosophy and religion, and to the hard sciences. These studies include a focus on power, oppression, and the intersection of identities. Gender and women's studies have historically had a lens that privileged a white, middle-class, and US-based perspective, but over the past 25 years has increasingly focused on experiences outside of this narrow range through discussions of intersectionality as well as through a focus on 'Third World' women. Gender and women's studies originated as an academic field in the 1970s as a spin-off from the second wave of feminism.

Overview

Women's Studies, also referred to as Gender Studies or Feminist Studies, is the interdisciplinary study of how intersectionality of gender, race, age, class, nationality, ethnicity, ability, sexuality, and other differences impact almost every facet of the social, political, and cultural experience. The basis for the academic field of Women's Studies was laid in the student, civil rights, and women's movements of the 1960s and 1970s, when women in academia argued that academic knowledge production failed to recognize gender as a lens of analysis. Many women, concerned about the often inaccurate and disparaging patriarchal narratives about them, seized the opportunity presented to them in the 1960s and 1970s to tell their own stories. Women took advantage of their formal educated status and the cultural atmosphere of radical, social, and political uproar to insist on a more systematic and potentially transformative narrative about women's lives. The discipline examines how gender not only influences popular culture and private life, but also impacts laws and social policies. By examining a cross section of disciplines such as political science, sociology, literature, psychology, and other subjects, Women's Studies seeks to understand gender roles in past and contemporary societies. Women's and gender studies also examines how the lives of individual women and men are shaped by broader structural forces in both historical and contemporary contexts, for e.g., nation building, globalization, economic developments, and the legal system. Women's Studies continues to reflect in its curriculum and faculty research the constantly changing directions that multiple First and Third World feminisms are taking today.

Women's Studies is an interdisciplinary field that examines socially- and culturally-defined gender roles. Drawing upon academic areas such as history, psychology, media sociology, sociology, literary criticism, and anthropology, it crosses the boundaries of traditional disciplines raising important questions regarding the way we have organized ourselves, our chief social and political institutions, and knowledge itself. As an area of academic study, Women's Studies provides new frameworks that are sensitive not only to issues of gender, but also race, class, and ethnicity. By analyzing the powerful and problematic impact of sexual inequalities, Women's Studies revises the way we see ourselves and our world. Because of the growing importance of women's concerns globally, undergraduates with a concentration in Women's Studies are well prepared for professional programs in law, medicine, and business, as well as graduate programs in social work, education, and arts and sciences. Indeed, students who major in Women's Studies, whether or not they choose to go on to professional or graduate school, are especially well positioned to work in education, policy institutions, development, media, social work, and the private sector.

Even though Women's Studies is a relatively new phenomenon in higher education, it is today well established as an interdisciplinary field of study, which draws on knowledge from both the humanities, the social sciences, medicine, and natural science. Today the field's interrogation of identity, power, and privilege go far beyond the category 'woman.' The first Women's Studies course is purported to have been created by American historian Mary Ritter Beard, who in 1934 constructed a 54-page syllabus for a course titled *A Changing Political Economy as It Affects Women* (Tuttle, 1986). Although the course was never taught, it prepared the foundation for the development of the first Women's Studies program in the United States. The first Women's Studies department was founded at San Diego State College (now San Diego State University) in 1970. At San Diego State, the on-campus Women's Liberation group formed a committee with faculty and community members called the Ad Hoc Committee for Women's Studies. The group collected hundreds of petition signatures in support of a Women's Studies program. In 1970, five San Diego State faculty members from existing departments taught courses related to Women's Studies. In the fall of

1970, the first Women's Studies Department was officially approved. By 1974, San Diego State had begun a national faculty recruitment to strengthen the Women's Studies Department. By the 1974–75 school year, the once fledgling department had two full-time and four part-time faculty and an enrollment of nearly 400 students. The Women's Studies department continued to grow, and in 1995 began offering a master's degree in addition to the bachelor's degree. The program also hosts a Women's Resource Center on campus. The second Women's Studies department was created at Cornell University in 1970 as well. Cornell's program was renamed Feminist, Gender, and Sexuality Studies in 2002.

Women's Studies has made the conceptual claims and theoretical practices of intersectionality, the study of how modes of identity such as race, class, and gender, and the sociopolitical construction of inequality are mutually comprised and must continually be understood in relationship to one another, and transnationalism, which emphasizes cultures, structures, and relationships that are formed as a result of the movement of people and resources across geopolitical borders, foundations of the discipline. Eventually Women's Studies was institutionalized into thriving academic programs across the country. Currently there are over 600 academic Women's Studies programs that offer minors, majors, and even graduate degree options. As a discipline, Women's Studies is concerned with achieving a level of equality for women in society. This dedication to systemic social change makes Women's Studies scholarship feminist and links the discipline to social movements that seek to end practices of inequity such as sexism, racism, and heterosexism around the world.

The place of Women's Studies within the academy has, of course, always been highly contested. Within the university, the discipline was said to not be academic enough. At the same time, however, feminist activists outside of the academy argued that the discipline was being too academic. Moreover, Women's Studies was further problematized by the emergence of disciplines like gender studies, and queer, transgender, and postcolonial theories in the 1980s and 1990s, which challenged the very category of 'woman' and called into question the founding premises of its identity politics. This social activism appears to have given way to academic practice as the dominant mode of 'Western' feminism at the turn of the millennium. However, the goals of the women's movement have not been fully endorsed. Women still strive to address issues such as equal pay for equal work, free childcare, and autonomy of sexuality.

Intellectual Backgrounds

The intellectual backgrounds of gender and women's studies began in the West with discourse about the equality of women in the late-eighteenth century (Donavan, 1991). Two strains of conversations emerged, one in the former Colonies and one on the Continent, but both centered on growing disquiet about the unequal legal, political, and economic status of women. The interaction of changes in wealth and technology with Enlightenment-based revolutionary thought allowed upper-class women access to education and ideas that caused them to question their roles in the political, economic, and social structure of the day, while lower-class women began leaving the home for outside word as industrialization spread.

The most fundamental argument that the first feminists addressed concerned women's legal status. Men were considered heads of households, and women and children were considered his property, and thereby denied property, legal, and political rights. Thus when Elizabeth Cady Stanton drafted the *Declaration of Sentiments* in 1848 declaring that men and women were essentially equal with respect to natural rights and thereby should be granted legal, political, and social rights, including suffrage, she was proposing changes that were simultaneously radical and moderate, in that while women's inclusion in these arenas would fully disrupt the social, economic, and political structure of the day, the philosophy upon which these proposed changes are built were comparable with the accepted beliefs about men's rights.

The focus of intellectual development about women and their role in society during the last half of the nineteenth century and into the first quarter of the twentieth century focused on the potential and role for women's inclusion in public decision-making through the franchise. In conjunction with Stanton, Susan B. Anthony led the start of the women's suffrage movement, whose most famous quote was "Resistance to tyranny is obedience to God." The early women's suffrage movement quickly included women's organizations that also focused on prohibition, and the argument was that because of women's central role in the home and their charge to protect the family, women should be allowed to vote so that they can also help the prohibition movement. Female antisuffragists countered that this central role in the family was anathema to entering the public sphere through voting. Through Westward expansion, the notion that women were intellectually and physically less capable than their male counterparts was weakened, and this argument was further deteriorated through the roles that women took over during the war efforts during World War I. The American women's suffrage movement was radicalized through the experiences of some of its leaders taking part in the suffrage movement in Europe, and the work of women like Alice Paul forced open the political opportunity structure allowing the Congressional and state strategies led by Carrie Chapman Catt to lead the movement to victory.

The strides in the understanding of gender and gender roles made by these early feminists were significant but limited. Advances in understanding the nature of gender diverge at this point, and these divergences were significantly influenced by emerging intellectual traditions in Europe and the New World. These traditions include transcendentalism, Marxism, existentialism, psychoanalysis, and the works of Freud.

Cultural feminism grew out of the transcendentalist movement and focused on the tight bonds between women and the potential for the development of a matriarchal culture. No writings of the time capture the spirit of this tradition more than Charlotte Perkins Gilman's *Herland*. Gilman describes a community of women cut apart from the rest of society for over a 1000 years that survived by asexual reproduction. The women existed in harmony with each other and nature, escaping the trappings faced by women in Western patriarchal societies, and romanticized notions of woman as nature. Cultural feminism laid the groundwork for later feminist offshoots, including Essentialism and radical separatism.

Marxist feminism developed through an extension of the Marx–Engels analysis of the nature of oppression extant between the bourgeoisie and the proletariat. Marxist feminists contend that the nature of oppression does not emanate from the separation of workers from the means of production, but instead from a more fundamental form of oppression, that of women by men. These feminists argue that all other forms of oppression develop out of this fundamental one. Probably, the chief contribution of Marxist feminists to the second wave of feminism was a focus on the hidden economy of housework.

Feminist reactions to Freud have been largely negative, rejecting what is seen as biological determinism (as well as fundamentally misunderstanding female biology) and bias in favor of masculine superiority. A strand of feminism, nonetheless, developed around psychoanalysis that focuses on the developmental role of individuals becoming part of a familial system. A key distinction is made in this line of analysis that explains gender differences based on men's and women's different preparations for adult life, in the market (capitalist production) and in the home (reproduction), respectively. The role that language plays in this system is to undergird and reproduce gendered oppression.

The roots of feminist existentialism come from the works of Simone de Beauvoir who extends the Hegelian dialectic to gender, noting that in gendered systems male becomes the positive and norm, whereas female is defined in opposition to him as the Other. Extensions of feminist transcendentalism focus on women's liberation through the destruction of the self as other. Here, language becomes important in the active renaming of self and the creation of women-centered communities and societies.

A common thread in both the Marxist and existentialist feminist traditions is throwing off oppression through consciousness-raising and an intentional focus on marginalization and otherness. By locating perspective and analysis from the outsider position, the development of a fuller analysis becomes possible, thus increasing the potential for liberatory efforts. In neither of these traditions do the primary proponents limit themselves to otherness based in gender only, and extend their understanding of oppression to race, ethnicity, class, and social position (Alcoff and Potter, 1993).

The second wave of feminism developed out of women's participation in the Civil Rights movement and the reaction of many movement leaders to gender-based organizing. Initial calls for a separate women's movement focused on the destructive roles of marriage, love, rape (as both an interpersonal tool of subjugation as well as a more widespread political tool of subjugation), pornography, imperialism and colonialism, and work. A radical separatist arm of the movement developed that focused on alternative family structures. Simultaneously, an arm of the second wave of feminism developed by women of color and focused not only on their exclusion by leaders of the Civil Rights movement, but by the white leaders of the women's movement and a movement ideology focused primarily on the issues and problems faced by middle-class white women as well. But the winner of these battles were the liberal feminists who focused on making incremental changes through court challenges and legislation under the premise that with the exception of gestation and childbirth, women are the same as men and therefore laws should develop to ensure equality of opportunity.

By the mid-1980s the women's movement began experiencing backlash and with it significant critique. Among the critiques were that the movement was developed for and about the needs of middle-class, primarily heterosexual white women. Third-wave feminism developed from these critiques, with a rejection of the male–female dichotomy in favor of a more nuanced understanding of both gender and sexuality, and a greater emphasis on the perspectives of women of color and embracing Grrrl and Queer theory. The intellectual core of third-wave feminism is the idea of intersectionality, or the idea that the multiple forms of oppression in society are interactive and interconnected. Intersectionality is an extension of the focus on identity politics and standpoint theorists in the 1980s and is used for analysis, critique, and movement development.

Current State of the Discipline

As a discipline, Women's Studies programs faced challenges in the early years and Women's Studies professors have had to challenge statements about the purpose and necessity of their programs. In the early years there were no Women's Studies 'departments' or funding for those who wanted to teach Women's Studies courses. The first Women's Studies faculty had to create a new discipline: they had to create their own textbooks, compile reading lists, develop new curricula, and establish journals. These newly developed courses provided the first opportunity for women's lives and experiences to be studied critically in the academy from a gender-sensitive point of view. Furthermore, these faculty members drew upon a new pedagogy that was more participatory and personal and later became known as feminist pedagogy. For a number of the all-male colleges and universities that became newly coeducational during the 1970s, Women's Studies programs were created as a way of incorporating and integrating women into the campus community. These new courses also offered faculty and students the opportunity to conduct in-depth scholarly work on subjects that had previously not been a part of the college curriculum, or were at the margins of more traditional disciplines, such as domestic violence, women's roles in historical periods, and women's literature. Moreover, given its origins in the women's movement, Women's Studies was from its beginnings activist in its design. The goal of the discipline was not 'disinterested' academic inquiry, but was clearly focused upon ending of oppression against women and challenging traditional paradigms.

Academic Programs

Growing rapidly as a discipline, by 1975 there were about 150 Women's Studies programs in the United States. Academic programs in women's and gender studies include a range from nondegree course offerings, to undergraduate certificates, to academic minors, to bachelor's degrees, to graduate certificates, to master's degrees, and to doctorates. In addition, many academic institutions allow students enrolled in interdisciplinary studies programs to develop their own emphasis on gender issues. Women's studies courses frequently fulfill general education requirements for undergraduate degrees. Among these institutions, many also include women's centers.

These nonacademic units frequently offer student development and support programming and activities beyond the academic approach.

Using a variety of resources and web-based lists, we see a range of these programs across the United States. (These include http://www.artemisguide.com/; http://userpages.umbc.edu/~korenman/wmst/prog; rams.html; http://www.smith.edu/swg/graduate.php; http://creativefolk.com/directories.html; and www.nwsa.org.) While there are over 600 programs in the United States, Alaska is the only state without a formal degree program, and there are several states with only one or a handful. The state with the most institutions with degree programs in women's and gender studies include New York with 43 and California with 41. Among these programs there are roughly 350 with academic minors, 50 with undergraduate certificates, 200 that offer a bachelor's degree, almost 50 that offer a graduate certificate, and another 50 that offer a master's degree, and about 20 doctoral programs (though there are more doctoral programs that offer interdisciplinary programs of some sort through which a concentration in gender or women's studies is possible).

Both the academic programs and the resource centers are supported by a variety of groups, but chief among them is the National Women's Studies Association (NWSA). NWSA was established in 1977, includes both individual and institutional members, and receives additional support from a variety of private foundations. NWSA hosts both national and regional meetings, hosts constituency groups, promotes the advancement of women's studies in education through task forces, and facilitates caucuses for underrepresented groups within the discipline (see www.nwsa.org for more information).

Outside of the United States, degree-granting women's studies programs exist in every region, though not in every country. By far the United States has the most degree-granting programs, followed by Canada with 45 and the United Kingdom with 34. Programs exist in every country in Western Europe, as well as many others in Eastern Europe and some in the former Soviet states. There are a handful of programs in Asia, South America, Africa, and the Middle East.

Journals

Among mainstream academic journals, there currently exist over 100 that focus on gender and gendered issues specifically. Among them, they focus on a variety of social sciences, arts, business, and law. However, the most common type of these journals is interdisciplinary, which include a variety of topics. With the exception of one journal on geography and about 10 concerned with women's health and health care, there are no hard science journals focused just on gender and gendered issues. There are a little more than a handful of journals with a specific focus on people of particular ethnic backgrounds, and these focus on women and gender issues related to people of India, Asia, African-American, Middle East, and Nordic descent.

Research

Trends in academic literature tend to follow the availability of journals with some notable exceptions. The major issues that gender and women's studies focus on contemporarily include: reproduction and reproductive rights; criminal justice, particularly interpersonal violence and rape; economic development and justice; intersectionality; LGBTQ rights; gendered roles in the media; and equal representation in leadership and economic institutions. Much of the literature does not take on one topic at a time, but instead takes up many of these themes simultaneously.

Recent examples from the journal *Signs* demonstrate overlap among these areas (see Table 1). Most notably, an analytic framework focused on intersectionality as opposed to patriarchy dominates most analyses, and combines other theoretical traditions within women's studies to traditional academic fields, particularly within the liberal arts and social sciences. Typically, the literature does so by focusing on one traditional field, reflecting the reality that while there are doctoral programs in women's and gender studies, most feminist scholars are trained and hired into these more traditional fields and use a feminist analytic frame to guide their scholarship.

Criticisms and Methodological Issues

As a discipline, Women's Studies sought to empower women to take an active role in creating and critiquing the language about and reality of women's existence. Women's Studies devotes itself to the study of women, whose experiences have been neglected in conventional scholarship. However, the discipline faces strong criticism from several areas. For example, Pat Robertson was quoted in a fund-raising letter stating that "The feminist agenda is not about equal rights for women. It is about a socialist, anti-family political movement that encourages women to leave their husbands, kill their children, practice witchcraft, destroy capitalism, and become lesbians." Others critique the activist nature of the discipline, arguing that it is inappropriate to develop an academic discipline based on a liberatory agenda, which necessarily requires a biased analysis as opposed to a value-free or neutral analysis.

While these assessments of the discipline are far from accurate, Women's Studies has differed widely over time and across institutions. Always centered on the experience of women, the status of the feminine, or the construction of sexual difference, gender or women's studies theorists have addressed an array of patriarchal underpinnings that have tainted the way that we view various issues. Women's Studies demands that through interdisciplinary investigation we reconstruct the very way we think about thinking, rejecting the normative claims of positivism that it is possible to have a truly value-neutral investigation of anything.

As seen in Table 1 (below), the methodological approaches taken by most gender and women's studies primarily rely on either analysis of text for critique and theory development or through qualitative data collection and analysis. Since the second wave of feminism, a common analytic approach employed by women's studies and gender scholars revolves around feminist critical theory. Feminist critical theory or literary theory developed as an analytic tool to simultaneously understand gender in text and to develop liberatory projects. To a lesser extent, feminist scholars will engage in quantitative analysis primarily utilizing gender as a dummy (dichotomous)

Table 1 Examples of recent scholarship

Area	*Recent examples*	*Major question*	*Fields*	*Methodology*
Criminal justice	McKim (2014)	How do gender-sensitive treatment programs inside and outside of the criminal justice system compare?	• Criminal justice • Intersectionality • Psychology • Marginality	Ethnography
	Fili (2013)	How does the binary view of females in prison as either victims or resisters frame how we understand identity?	• Criminal justice • Political theory • Epistemology • Identity	Theoretical reflection; case study
Equality	Akchurin and Lee (2013)	How do forms of women's activism, professional, labor, and movement impact earnings cross-nationally?	• Development • Economic inequality • Social movement theory • Sociology • Political science	Surveys; quantitative modeling
Culture & norms	Sallee (2013)	Within the higher education setting, what role do gender norms play on men's involvement at home?	• Gender norms • Higher education • Business, management, and leadership	Interviews; case comparison
	Durham et al. (2013)	What is the state of the field of hip-hop feminism?	• Popular culture • Black feminism • Intersectionality	Review of literature; theoretical reflection
Intersectionality	Cho et al. (2013)	What is the potential for the development of a new field of intersectional studies? What are the dynamics of scholarly intersectional studies compared with applied change models?	• Intersectionality • Political science • Social movements • Feminist theory • Advocacy	Theoretical reflection
	Patil (2013)	To what extent has the shift from patriarchy to intersectionality in feminist theory been made?	• Feminist theory • Transnational scholarship	Theoretical reflection; case study process tracing

variable to understand the differences in behavior, responses, beliefs, and opinions between men and women. While a more common approach taken by scholars in traditional departments, particularly in fields in which quantitative analysis is given primacy over qualitative analysis, this approach flies in the face of the central tenants of third-wave feminism and their understanding of gender as a range of attributes as opposed to a binary opposition.

Current and Emerging Directions in Theory and Research

The focus of women's studies has changed significantly over time. Early scholarship focused on documenting the state of being female – what do women do; where are they (and where are they not); how are they treated. The focus quickly shifted to an analysis of why, usually accompanying advocacy for positive change. This focus on why also delved into the nature of gender and gender distinctions.

The continued growth in Women's Studies has been attributed to two major trends in higher education – the growth of a diverse student body increasingly older, female, racially and economically diverse, first-generation college students, and the differently abled; and a movement to transform the traditional academic curriculum into one that is more expansive and inclusive. Women and gender studies draws upon an interdisciplinary core that places women's diverse experiences and the issue of gender at the center of inquiry. One of the prominent contemporary trends in the field of Women's Studies is the model of incorporating intersectionality – race, class, gender, and other dimensions of human identity – into research.

Though topics such as sexuality, reproductive health, race, and equality continue to dominate the discipline, there are two significant and emerging themes in the discipline. The first focuses on the representations of children, children's experiences, and children's place in the world. The rights of women are often discussed around the question of their biological capacity to bear and raise children. Children's identification and development represents an emerging trend in women's and gender studies. The second is a reimagining of solidarity, or the act or expression of unity in the face of power. Solidarity is an open-ended concept, but at the most basic level it implies a community based on mutual responsibility of each member toward the other, and a moral obligation to fight for the welfare of other members and of the community as a whole. Contemporary feminist reifications of solidarity as an important quality of the movement harken back in some ways to the traditions of the early cultural feminists as well as the radical separatists of the 1970s but in this iteration place a greater emphasis on inclusivity.

See also: East Asian Studies: Gender; Ecofeminism; Feminism in Anthropology; Feminisms and Acculturation around the

Globe; Feminist Political Theory and Political Science; Gender History; Gender and Social Research; Gender, Feminism, and Sexuality in Archaeological Studies; Gender: Anthropological Aspects; Intersectionality and Crime; Intersectionality; Near Middle East/North Africa Studies: Gender; Queer Theory and Intersectionality.

Bibliography

Akchurin, Maria, Lee, Cheol-Sung, 2013. Pathways to empowerment: repertoires of women's activism and gender earnings equality. American Sociological Review 78, 679–701.

Alcoff, L., Potter, E. (Eds.), 1993. Feminist Epistemologies. Routledge, New York.

Cho, Sumi, Crenshaw, Kimberlé Williams, McCall, Leslie, 2013. Toward a field of intersectionality studies: theory, applications, and praxis. Signs 38, 785–810.

Crenshaw, K., 1989. Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago Legal Forum 139–167.

Donavan, J., 1991. Feminist Theory: The Intellectual Traditions of American Feminism. The Continuum Publishing Company, New York.

Durham, Aisha, Cooper, Brittney C., Morris, Susana M., 2013. The stage hip-hop feminism built: a new directions essay. Signs 38, 721–737.

Fili, Andriani, 2013. Women in prison: victims or resisters? Representations of agency in women's prisons in Greece. Signs 39, 1–26.

Howe, F. (Ed.), 2000. The Politics of Women's Studies: Testimony fro Thirty Founding Mothers. Feminist Press, New York.

Maher, F.A., Tretreault, M.T., 1994. The Feminist Classroom. Basic Books, New York.

McKim, A., 2014. Roxanne's dress: governing gender and marginality through addiction treatment. Signs 39, 433–458.

Olson, H.A., 2002. Information Sources in Women's Studies and Feminism. K.G. Saur, Munchen.

Patil, V., 2013. From patriarchy to intersectionality: a transnational feminist assessment of how far we've really come. Signs 38, 847–867.

Pilcher, J., 2004. Fifty Key Concepts in Gender Studies. Sage, Thousand Oaks, CA.

Reynolds, M., Shagle, S., Venkataraman, L., 2007. A National Census of Women's and Gender Studies Program in U.S. Institutions of Higher Education. The National Women's Studies Association, Baltimore, MD.

Sallee, M.W., 2013. Gender norms and institutional culture: the family-friendly versus the father-friendly university. The Journal of Higher Education 84, 363–396.

Scott, J.W. (Ed.), 2008. Women's Studies on the Edge. Duke University Press, Durham.

Tuttle, L., 1986. Encyclopedia of Feminism. Facts on File Publications, New York.

Gender Development: Cultural Differences

Deborah L Best and Kelsey L Luvender, Wake Forest University, Winston-Salem, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores culture-based differences in gender development. In spite of the noticeable differences seen in clothing and other somewhat superficial behaviors, there are indeed many similarities between men and women within various cultural groups, along with notable differences in gendered behaviors across cultural settings. This article examines the role of culture in the development of gendered behaviors, looking at both proximal and distal factors that affect gender development in various cultural settings.

Culture and Gender

The role of culture is often overlooked in discussions of gender roles and biological differences are usually considered the primary source of influence. Yet, visitors to other countries and cultures are often struck by the obvious gender differences they see in the clothing and behavior of women, men, girls, and boys. Western visitors in predominantly Muslim countries quickly notice that women are more 'modest' than men, covering more of their bodies with long dresses, headscarves, hijab, or burqa. Indeed, when South Asian visitors arrive in Western countries wearing their saris or dhotis, they encounter women with bare shoulders and midriffs, short skirts, and high heel shoes, and men in sport coats, trousers, and neckties. In spite of these differences in clothing, and that globalization has made blue jeans and tee-shirts standard fare around the world, one wonders whether differences in clothing reflect differences below the surface. Are women and men, girls and boys different in different cultural settings?

Before proceeding, it is useful to define several terms. *Sex* usually describes the anatomical and physiological differences between males and females, which are often considered to be the biological basis of observed gender differences. *Gender*, on the other hand, is a more subjective term and is used to describe the psychological and behavioral differences between males and females. The development of gender-related behaviors and beliefs are influenced by cultural practices and beliefs, such as what tasks children are taught, the clothing they wear, and the division of labor within the family (Whiting, 1963; Whiting and Edwards, 1973). In spite of the distinctions made between sex and gender, across the lifespan biological and social influences on the development of gender roles and identity are inseparable and it is not always easily distinguishable (Fausto-Sterling, 2000; Hoyenga and Hoyenga, 1993).

Viewing gender development through the lens of probabilistic epigenesis demonstrates the interplay between genes, the sociocultural environment, and evolutionary processes. The probabilistic epigenetic framework emphasizes the reciprocal influences within and between the levels of the organism's *developmental manifold* (genetic activity, neural activity, behavior, and the physical, social, and cultural influences of the external environment) in the realization of observable phenotypes (Gottlieb, 2002).

Proximal and Distal Influences on Gender Development

Although developmental psychologists often categorize factors that influence gender development as proximal and distal, these are not orthogonal categories and indeed are somewhat overlapping and interactive. Distal factors within the macrosystem (Bronfenbrenner, 1989) are culture-wide and objectively are shared by many if not all people within a culture. They are often represented by global comparison statistics (e.g., United Nations, World Health Organization, Organisation for Economic Co-operation and Development), which are assumed to be consistent across the culture or country being described. The political system and government regulations, the national economy and labor opportunities, dominant religious organizations and beliefs are among the distal macrosystem factors that can differentially influence women and men within and between cultures. Along with these more structural cultural characteristics there are also prevailing social norms and expectations for 'proper' gender role behaviors. Children grow up within this complex of distal influences that shapes their gender-related learning and experiences.

In contrast to these distal influences, proximal microsystem developmental factors are more individualized and may differ between families or specific groups within a larger culture (Bronfenbrenner, 1989). Although distal aspects of the culture may influence or be represented within the proximal factors, they are not subjectively experienced in similar fashion by all members of the society. Within a single cultural setting, for example, there is often wide variability in resources and opportunities between families and individuals; not everyone is rich, nor are they rich to the same degree. When a country experiences an economic downturn, the subjective experiences of people at the top and bottom of the economic scale differ. Low-wage earners may lose their jobs and homes while high-income families simply eat in restaurants less often. These distal and proximal factors make up the developmental niche (Super and Harkness, 1986), the ecocultural framework (Bronfenbrenner, 1989) within which gender development occurs.

Proximal Biological Influences on Gender Development

Setting the stage for cultural influences within the developmental manifold are the proximal biological factors that lay the

foundation for gender differences to develop. Sex differences between females and males are found in sex chromosomes, hormones, neural structures, physical size and strength, activity levels, and secondary sex characteristics. These biological differences are often considered to be immutable and the hallmarks of gender differences. Yet, there is considerable variation across cultural settings in how these differences unfold and are perceived. Behaviors and developmental milestones have long been considered biologically driven, such as infant sleep patterns or the onset of sitting and walking. However, these sorts of behaviors have been shown to vary across cultures and are influenced by parental practices and ethnotheories about child development (Super and Harkness, 1986). Along with inheriting genes, children also inherit parents, peers, and the places they inhabit. Thus, it should not be surprising to find that gender development is influenced by the sociocultural context.

Proximal Cultural Factors in Gender Development

Along with biological influences on the development of sex and gender, culture, a vital part of the developmental manifold, provides a context in which biology unfolds. Around the world, children grow up within the cultural context of the family – the developmental niche composed of the physical environment where the family lives, the parents beliefs, or ethnotheories about what children are like and how they should be cared for, and the actual care they receive from the mother, father, siblings, and others within the family circle (Super and Harkness, 1986). As should be obvious, girls and boys may be treated differently across these different cultural contexts.

Gender Socialization and the Role of Parents

Gender socialization is the process by which cultural information about gender is transmitted from one generation to the next within the ecocultural context (Best and Williams, 1993; Munroe and Munroe, 1975/1994; Rogoff et al., 1984). Mothers, fathers, siblings, relatives, teachers, peers, and others teach boys and girls the gendered behaviors, values, and roles that are expected or are necessary for living successfully within their cultural setting. At birth (or before given the frequent use of ultrasound), a newborn is labeled a girl or boy, and gender role socialization begins and is shaped by parents' and cultural expectations.

Parents' Expectations about Gender

Parents describe their newborn girls as 'finer' and 'quieter' than boys who are described as 'more coordinated' than girls (Sweeney and Bradbard, 1989). Girls are handled gently and are talked to more than boys who are handled more physically and are expected to be 'tough.' Such parental notions are even born out in 'Baby X' studies where unbeknownst to the parents, they are describing the same baby that has simply been labeled as a girl or boy (Seavey et al., 1975; Sidorowicz and Lunney, 1980). These expectations about girls and boys reflect the parents' memories of their own childhood and how the future will unfold for their child.

However, cultures change with time, and roles and behaviors reflect those changes. Little girls growing up in the United States in the twenty-first century will certainly have many different opportunities than did their mothers or grandmothers. They could become engineers, air force fighter plane pilots, or champion soccer players, roles unavailable to their grandmothers. Similarly, cultural changes have made it more acceptable for boys in the United States to grow up to be preschool teachers or stay-at-home dads. Likewise, Zinacanteco parents in Mexico have expectations for their children based on their own experiences and interpretation of their world (Greenfield et al., 1989). To prepare their children for their future roles, Zinacantec parents sometimes give gender-appropriate objects to newborns, such as green chilies to boys who are expected to know how to buy them as adults.

Culture shapes the way that parents interpret the child's biological sex characteristics (e.g., 'It's a girl.') and how they treat the child. The names that girls and boys are given, how they are dressed and cared for, what behaviors are considered appropriate, and what tasks and roles they are taught are among the many ways that culture influences gender. Children grow up within parents' culturally derived scripts (e.g., girls should learn to be obedient, boys should be brave and not cry) that assume continuity in the expectations and experiences that parents and children encounter. These scripts guide children's behavior long before they understand their own gender label or appropriate gender role behaviors.

Culturally based scripts and expectations shape the way that girls and boys are socialized. In their classic study of socialization practices in more than 100 societies, Barry et al. (1957) found that girls are usually reared to be nurturant, obedient, and responsible while boys are reared to be independent, self-reliant, and achieving. These differences are often referred to as the expressive and instrumental constellation of traits, which are assumed to be bipolar opposites, suggesting important differences in the socialization of boys and girls. However, a later reanalysis of these data suggested that these dimensions are independent, rather than opposite ends of one dimension and that both boys and girls are socialized to be independent and self-reliant (Hendrix and Johnson, 1985). These studies suggest that when looking within various cultural groups, differences in the way that girls and boys are socialized seem to be consistent across cultures. Yet, bear in mind that cultures certainly vary in how differently girls and boys are socialized. Girls may be reared to be more independent and self-reliant in one culture than are boys in another culture.

Mothers' and Fathers' Roles

Among the gender influences found in the family context, an important socialization mechanism is found in the differential roles and modeling provided by mothers and fathers. Weisner and Gallimore (1977) examined parenting roles in 186 societies and found that mothers, female adult relatives, and female children were the primary caregivers of infants, though siblings of both sexes were often caring for younger children. Although children generally see mothers as responsible for children, in some cultures 2- to 4-year-olds spend more than 70% of their time with child caregivers. In these cultures, mothers are actively contributing to the livelihood of the family and even young children are expected to assume responsible roles (Mintern and Lambert, 1964).

In comparison with mothers, fathers generally spend less time with their infants and young children and they participate

less in childcare (Katz and Konner, 1981; West and Konner, 1976). Both boys and girls see fathers' role as breadwinner and mothers' as responsible for the domestic realm, and these beliefs persist into adolescence (Goldman and Goldman, 1983). Best and colleagues (Best et al., 1994) confirmed these beliefs when observing parents and children on playgrounds in three European countries. Mothers in France, Italy, and Germany showed more caregiving behaviors than fathers, but French and Italian fathers engaged in more play than mothers. Indeed, Lytton and Romney (1991) found that fathers pay less attention to daughters than sons. Fathers tend to be more important as caregivers of sons (Rohner and Rohner, 1982), but mothers are equally important as caregivers of sons and daughters. When observed in public places, girls are seldom found with fathers or adult males, but boys are often found in all male groups and these differences increase with the age of the child (Mackey and Day, 1979).

Although fathers in most cultures have less direct contact with their infants and young children than do mothers, they do play an important role in gender socialization. Fathers are more likely than mothers to reward children for gender-appropriate behaviors and to discourage cross-sex behaviors, particularly for sons (Leve and Fagot, 1997). Indeed boys 'get the message' and they report that their fathers would think it was 'bad' if they play with cross-gender toys (Raag and Rackliff, 1998). Boys believe that their fathers have more restrictive rules than their mothers, but girls do not share this belief. When fathers are absent for long periods of time, boys display effeminate behaviors, dependence, excessive fantasy aggression, and overly masculine aggression (Gronseth, 1957; Lynn and Sawrey, 1959; Stevenson and Black, 1988; Stolz, 1954).

In contrast to these typical gender differential patterns of parenting is that found among the Aka forest foragers (Hewlett, 1992, 2000). Among the Aka, adults and children forage together caring for family needs. Men and women participate in collaborative net hunting and sex roles are egalitarian. Women are held in high esteem and food preparation and childcare are not considered feminine activities. Aka fathers are with their infant children 88% of time compared with !Kung and Logoli fathers who are with their infants 30 and 9% of the time, respectively. Aka fathers are affectionate caregivers for infants and children, not playmates, and they do not engage in rough-and-tumble play. Aka infants and fathers become attached and seek each other out. A different pattern of parenting is found among the Hadza, hunter-gatherers of East Africa, as well as among other African agricultural groups. Compared with the Aka, the Hadza are more gender stratified and fathers do not participate in childcare (Marlowe, 2005).

These different parenting strategies show the variability in how parents and children adapt to the demands of their physical and social ecological settings. Fathers can be nurturant caregivers and mother can be playmates and disciplinarians. As would be expected children who see their parents perform cross-gendered behaviors (e.g., fathers cooking, mothers washing the car) are less aware of gender stereotypes and show less preference for stereotypic adult occupations (Turner and Gervai, 1995).

Differential Treatment of Girls and Boys

When looking at actual parenting behaviors, Lytton and Romney (1991) found in their meta-analysis of 172 studies involving more than 27 000 children that there were few systematic differences in parents' rearing of girls and boys. In the North American studies, both parents encouraged gender-typed activities and characteristics, with fathers showing more gendered expectations than mothers. In 17 Western countries, boys received more physical punishment than girls, but disciplinary differences decrease with age. Fathers were more likely to use physical punishment, particularly for boys, and mothers were more likely to reason with children and explain rules and consequences (Conrade and Ho, 2001; Russell et al., 1998).

There appear to be only minor differences in the ways that parents treat girls and boys. Often studies comparing mothers and fathers behavior with their children observe them in the same, relatively representative settings. Yet, there are certainly different behavioral expectations and constraints for parents and children relative to the setting. Lindsey and Mize (2001) observed parent–child dyads in pretend and physical play settings. Mothers engaged in more pretend play with their daughters and fathers in more physical play with their sons, and these play patterns predicted similar play patterns observed with children and their peers. When parents were observed helping their preschool children on a balance beam, mothers were equally attentive with daughters and sons, but fathers were more protective of daughters (Hagan and Kuebli, 2007).

Weisner and colleagues (Gallimore et al., 1993; Weisner, 1984) developed the concept of 'activity settings,' which highlights the behavioral consequences of children's culturally regulated social settings. In Kenyan families that divided time between urban and rural settings, mothers were more nurturant toward their young sons when they were in their crowded Nairobi apartments than when they were at their farms. The Munroe and Munroe (1971) also demonstrated the effects of household density on Kenyan Logoli mothers' behavior when looking at infant caregiving. Infants in households with many potential caregivers were responded to more quickly when they cried, but they received less care from their mothers who were usually busy with gardening or other household tasks. The Munroes point out that in many cultural settings infant caregiving is a system, often with multiple caregivers, rather than the function of a single one-on-one caregiving relationship.

Thus, it appears that there may be subtle differences in how parents treat daughters and sons, and occasionally these differences are statistically significant. Yet, it is likely that the same parental treatment could affect girls and boys differently, particularly in transitional developmental periods, such as toddlerhood or adolescence, when new abilities are emerging (Beal, 1994).

Task Assignment

One of the important ways that children learn differential gender roles is through the tasks that they are taught and asked to perform. The role of the child's learning environment was demonstrated in the classic Six Culture Study that investigated socialization practices with 3- to 11-year-olds in India, Kenya, Mexico, Okinawa, the Philippines, and the United States (Edwards and Whiting, 1974; Minturn and Lambert, 1964; Whiting and Edwards, 1973). Mothers' work determined children's participation in chores, such as carrying wood and

water, preparing food, gardening, cleaning, and caring for younger children and animals. Training for girls began at an earlier age than for boys, girls were engaged in more chores, cared for younger children, and participated in more economic activities than boys (e.g., shelling maize for sale). Boys ran errands were involved in more undirected free play, tended animals, and were further from home than girls.

In the context of these differential learning environments, girls who spent more time caring for infants displayed more nurturance than boys. Boys showed more egoistic dominance (i.e., changing others' behaviors to suit one's own desires) than girls. There were fewer gender differences in the cultures where both boys and girls helped with sibling care and household chores (i.e., Kenya, Philippines, United States) and greater differences where girls did more of these (i.e., India, Mexico, Okinawa). In general, girls had more responsible productive work, and boys spent relatively more time playing. Parents communicate the importance of appropriate gender roles by their reactions to their children's behavior and by the organization of activities within the family.

Siblings and Peers

In addition to parents, across all cultural settings, siblings and peers play important roles in the socialization of gender roles, and their influence tends to increase with age. Boys with older brothers and girls with older sisters have more gender-typed activity preferences than same-sex singletons who, in turn, are more gender-typed than children with other-sex siblings (Rust et al., 2000). For both girls and boys, having an older brother is associated with more masculine and less feminine behavior. Boys with older sisters are more feminine but not less masculine, and girls with older sisters are less masculine but not more feminine. The greater influence of older brothers may be due to the higher cultural value of typical masculine role behaviors than feminine ones. In most societies it is more acceptable for girls to behave like boys than it is for boys to behave like girls (Golombok and Fivush, 1994). Indeed, among school-age children, girls imitate male role models to a greater extent than boys imitate female models (Bussey and Bandura, 1984, 1992). For girls, being a tomboy is relatively common and acceptable in most Western cultures, but for boys, being a sissy is frowned upon.

Peers also play an important role in gender socialization. When toddlers, 21–24 months of age, were observed in free play, both girls and boys received positive responses when they played with same-sex peers (Fagot, 1985). However, when boys participated in typically feminine play, the negative feedback from their male peers directly told them such behaviors were inappropriate and to play with others like you. Girls expressed disapproval of girls who played with boys, but they did not show concern about the types of activities involved. It is evident that male peer groups start defining what is *not male* at an early age, and those activities that are not male quickly drop out of the boy's repertoire.

Maccoby (1998) observed three different ways that gender influences children's social interactions with peers: gender segregation, dissimilarity of boys' and girls' interaction styles, and gender group asymmetry (e.g., boys groups are more cohesive, sexists, exclusionary, and separate from adult culture than girls' groups). Most often, boys are in packs and girls in pairs. These differences are influenced by both biological differences (e.g., boys higher activity levels) and by gender socialization that occurs within the cultural context. As early as 30–36 months of age, children favor same-sex playmates, and gender segregation occurs across cultures and increases with age (Bock, 2005; Leaper, 1994). By age $4^1/_2$, children in the USA spend three times more time with same-sex peers than with other-sex peers, and by age $6^1/_2$ this differential has grown to 11 times more time (Maccoby and Jacklin, 1987).

Similar patterns of peer interaction were found in the Six Culture Study with children as young age two preferring same-sex peers and this tendency increased throughout middle childhood. In these same-sex groups, boys strive for dominance and take more risks, and girls work to maintain positive social relationships and avoid conflict. Gender segregation found in the childhood years is sometimes reinforced by adolescent initiation customs and ceremonies found in many cultures.

Initiation Rituals

Initiation rites and rituals are a socially prescribed way to show others that the initiate has achieved adulthood. Separation from the family and other-sex peers, initiation into adult sexuality, instantiation of male dominance and aggression, and demonstration of peer group and political loyalties are seen in initiation rituals. Group initiation is found more often for boys than girls and in cultures that emphasize adult gender difference (Edwards, 1992).

The pubertal process does not hold a universal definition across cultures. In some cultures the crossover from childhood to adulthood is marked by a biological change, while for others it is more of a social occurrence. In some non-Western societies, 'coming of age' occurs at a particular age with traditional initiation ceremonies that may include male circumcision and female genital cutting. After initiation, the boy or girl is considered an adult and is expected to work, marry, and have children (Herdt, 2010).

The Sambia of Papua New Guinea initiate boys before their tenth birthday as a way of controlling gender roles in their culture. Initiation rites 'defeminize' the boy's body and separate him from his mother (Herdt, 2010). Initiation occurs at age 10 because of the 'emergence of sexual attraction to others' at that age, especially for boys (Herdt and McClintock, 2000). In societies that view sexual behavior as strictly reproductive, adolescents are less likely to feel comfortable engaging in such actions. Yet in societies, such as the Trobriand Islanders and the !Kung people of Southern Africa, sexuality is encouraged as recreational and pleasurable outside of marriage (Herdt and McClintock, 2000).

In Western cultures, there is a noted difference between gonadal puberty, with later onset for boys than girls, and sexual maturation (Herdt and McClintock, 2000). This difference is not typically distinguished in more traditional, non-Western cultures where biological changes are not considered the determinant of puberty. Western cultures recognize pubertal changes that begin around the age of eight as critical for cognitive and behavioral development that leads to mature awareness of sexuality.

In some cultures, unlike many Western cultures, open discussion about sexuality is rare and discouraged. Young Pakistani women report gaining information about their own

sexual development from sources within their immediate home while young men gain information from external sources (Hennink et al., 2005), though the information may be minimal. Sex education is neither taught in schools nor explained in the home in Pakistan. Few Pakistani adolescents discuss pubertal bodily changes with family members, and the majority of those who do are boys. Girls receive little information about pubertal changes in themselves or the opposite sex or how a female becomes pregnant. Pakistani women may only learn about sexuality and reproduction just prior to marriage, which often leads to high anxiety and fear on a woman's wedding night (Hennink et al., 2005).

Parents, siblings, peers, and other socialization agents help shape a child's gender-appropriate behaviors, toy choices, playmates, activities, and gender role behaviors. These proximal influences on gender development occur within the larger cultural context and interact with more distal cultural influences, such as the division of labor and the political and education systems. These distal influences help to shape the stereotypic beliefs that children develop about gender and their beliefs about appropriate gender roles.

Distal Cultural Factors in Gender Development

Division of Labor

With the wide variation in physical settings and environmental demands, it is quite surprising that from rural, subsistence cultures to urban, industrialized ones, there are few difference in the gender-related behaviors that are found. This consistency is usually attributed to the almost universal division of labor within societies. For example, across 244 different societies D'Andrade (1966) found that women were responsible for food preparation and domestic chores, activities consistent with childcare (Brown, 1970; Segal, 1983). In contrast, men were responsible for hunting and activities further from home, with little involvement in childcare (D'Andrade, 1966; Katz and Konner, 1981).

In spite of these consistencies in division of labor, environmental demands modify these distinctions. Van Leeuwen (1978) found that in hunting or gathering societies with low food accumulation, both males and females must contribute to the subsistence of the group, which minimizes the division between the sexes. In such settings women's contributions are valued. In sedentary societies with high food accumulation, the usual division of labor is found, and women's roles involve nurturance and compliance.

Cultural changes in women's fertility rates, opportunities for childcare, and technology have permitted women to increase their participation in the labor force outside the home. However, they still maintain most of their domestic responsibilities, doubling their workload (Dixon, 1978; Population Crisis Committee, 1988). Looking at workforce measures across 56 countries from 1960 to 1980, Jacobs and Lim (1992) found that women's labor force participation and fertility rates were inversely related. Surprisingly, modernization, measured by gross national product and women's education, were positively related to gender segregation and women were paid less than their male counterparts (Ottaway and Bhatnagar, 1988). Even more surprising, in a study in the USA, Great Britain, West Germany, and Austria more educated people and women with employed husbands were less favorable toward gender equality than less educated people or women with no male wage earner (Davis and Robinson, 1991). These findings suggest that the relation between the division of labor and beliefs about appropriate gender role behaviors is a complex one that is influenced by environmental demands, as well as long-standing cultural customs and practices which are embedded within the history, economics, and prevailing religious beliefs of a society (Mukhopadhyay and Higgins, 1988).

Status, Religion, and Political Influence

The multidimensional relations among status, power, and privilege in various cultures are influenced by the physical demands of the environment, and the economic and political organization of the society. As a result, the status afforded women is multifaceted (Mukhopadhyay and Higgins, 1988; Quinn, 1977) reflecting their economic contributions, political involvement, the autonomy, and power of their various positions, as well as the prevailing ideological and religious beliefs that prescribe appropriate gender roles (Best, 2009).

Along with economic activities, men generally play more powerful roles in religious practices and hold more influential positions in political/governmental organizations. Religious doctrines vary greatly in the status and freedoms they grant to women and men and their teachings about appropriate gender relations and roles. However, in the three most predominant religions throughout the world, Christianity, Islam, and Hinduism, religious practices are primarily controlled by men (Browning et al., 2006). When religion and government are largely separate, religious beliefs have less influence on the interactions of women and men. Yet in countries where religious doctrine and governmental authority are closely tied, religious beliefs exert a stronger influence on gender roles and child-rearing practices (Browning et al., 2006).

Across cultures, men's status is consistently superior to that of women (Hoyenga and Hoyenga, 1993; Population Crisis Committee, 1988; Whyte, 1978). Recently, the Global Gender Gap Index 2013, which is independent of the income and resource level of an economy, measured the equitable distribution of available income, resources, and opportunities for women and men. Review of 136 countries demonstrated that no country has yet achieved gender equality. However, all Nordic countries, except Denmark, have closed over 80% of the gender gap, providing benchmarks for international comparison (World Economic Forum, 2013). Similar shift in status and power are occurring in other cultures as well. In the last decade, Ethiopia passed the most progressive abortion laws in Africa to combat unsafe abortion rates, and Mali passed a law that women are no longer required to obey their husbands. Although the status of women's rights and freedoms continue to improve, men are more involved in economic and political activities and possess greater status and power than women.

Educational System and Practices

Certainly one of the most influential mechanism of cultural influence is found in the educational practices endorsed by

a society. Usually within these educational settings, there are cultural messages about appropriate gender behaviors and socialization practices that support them. Parents may value education, both formal and informal, differently for girls and boys. In some, less egalitarian societies, it is considered more important to educate boys than girls. Serpell (1993) found that in Zambia, it was considered more important to educate boys than girls, and even though mothers were responsible for children, schooling for boys was so important that fathers made the arrangements. Similarly, for some traditional Muslim ulama teachings, secular education for girls, especially higher education, is thought to lead them astray. Arab families have a tendency to pay for boys' education but not girls' because they may believe that money spent on boys will eventually be repaid. There is little point in paying for girls' education because girls are destined to become mothers and will not be in careers where they could make money (Faisal, 2003).

Across 48 countries, women who received some education were significantly less likely to be married as children (below age 18) than those who received no education (UNICEF, 2004). Marriage before the age of 18 is a reality for many young women, but education is often seen as key to preventing child marriage. In Zimbabwe, only 4% of girls who had attended primary school were married, compared with 40% of those who had received no education. In many parts of the world, parents encourage the marriage of daughters while they are still children hoping to benefit financially and socially, and also to relieve financial burdens on the family (UNICEF, 2005).

Parents' beliefs about the importance of school as well as their expectations about children's performance greatly influence children's achievements. Mothers in the USA, China, and Japan expect boys to be better at math and girls to be better at reading (Simpkins et al., 2012; Lummis and Stevens, 1990) even though for some aspects of both disciplines, they perform equally well. Boys show greater confidence in mathematics, relative to girls, than would be expected based on the relatively small differences in performance. Yet in most countries girls are more anxious about math abilities than boys (OECD, 2004b).

Teachers treat girls and boys differently in the classroom. Fifth grade teachers in the USA and Japan devote more than half their attention, especially negative attention, to boys than girls. However, this increased attention is not due to more off-task or bad behavior (Hamilton et al., 1991).

In many Western countries over the second half of the twentieth century, girls' educational performance has overshadowed boys' and women's level of education has surpassed men's. According to USA census data, since 2000 women in the USA have held 58% of the advanced degrees (master's, professional, doctoral), and in 2012 37% of women and 30% of men attained a bachelor's degree.

In 21 of the 27 OECD countries for which comparable data exist (OECD, 2004a), university-level graduation rates for women now equal or exceed those for men. In these primarily Western countries, significant progress has been made toward reducing the gender gap in formal education. It is evident that in these cultures the value and importance of education is clearly shifting in relation to the roles and expectations for girls and boys.

Development of Culturally Endorsed Gender Roles and Stereotypes

As a result of the various cultural forces on gender role learning that children experience, they develop culturally derived stereotypes about male–female differences and gender roles. As early as 2 years of age, children in the USA stereotype objects as feminine or masculine (Thompson, 1975; Weinraub et al., 1984). By 3–4 years of age they identify activities, toys, and occupations as appropriate for one sex or the other (Edelbrook and Sugawara, 1978; Guttentag and Longfellow, 1977). Similar gender-related learning takes place around the world (Best, 2013; Bloch and Adler, 1994; Edwards, 2000; Prosser, et al., 1986).

As children are learning about gender roles during the childhood years, they also begin to associate particular personality traits with women and men, such as women are emotional, gentle, and men are strong, dominant. European American children's knowledge of these gender-trait stereotypes increase from kindergarten through high school, with the largest increases during the elementary school years (Best et al., 1977; Williams and Best, 1982/1990). The same pattern was found with 5-, 8-, and 11-year-olds in 25 countries though the rates of learning varied (Williams and Best, 1982/1990). In 17 of the countries, children knew more of the male traits than female, with German children being the only country showing the opposite pattern. There was also evidence that female items were learned earlier in Latin/Catholic countries (Brazil, Chile, Portugal, Venezuela) where the adult female stereotype had more positive characteristics than the male. Gender stereotype trait knowledge was highest in Pakistan where 5-year-olds knew more of the stereotyped items, particularly the male items, than 8-year-olds in other countries. Studies have shown that urban, middle class children who have more educational opportunities learn the global, pancultural stereotypes at an earlier age than their rural, lower class counterparts (Lee and Sugawara, 1982; Montero, 1978; Tarrier and Gomes, 1981). However, as children grow, stereotypes incorporate more gender-incongruent information (Hanover, 2000) and by young adulthood stereotype scores show little relation to education or urban/rural status. Thus, it appears that although culture modifies the rate of learning as well as aspects of content, gender-trait stereotypes are remarkable universal (Berry et al., 2002; Best, 2010). Such similarity perhaps reflects the fact that the relations between females and males are remarkably similar in cultural groups around the world.

Conclusions Regarding Gender Development and Cultural Transmission

Perhaps growing from the biological differences between females and males, cultural practices and traditions evolve and become part of the belief systems that are seen in religion, oral and written history, and the myths and literature of cultures. From these grow expectations about appropriate gender roles and stereotypic notions about the differential characteristics associated with the sexes. These are woven into the framework of the culture and become socialization models for children.

Although the developmental mechanisms that lead to gender-related behaviors may differ across cultures, the more distal ones, such as the division of labor, religious beliefs, political and social status, and the educational system provide the backdrop for more proximal mechanisms, such as socialization by parents, siblings, and peers, differential treatment and tasks for girls and boys, and initiation rituals. Such practices result in the gender-related beliefs, attitudes, roles, and behaviors that are observed in cultures around the world and lead to individual variation as well as differences between cultural groups.

See also: Concept Formation in Childhood: Cross-Cultural Perspectives; Friendship During Childhood and Cultural Variations; Gender Differences in Personality and Social Behavior; Gender and Academic Motivation; Parenting Attitudes and Beliefs across Cultures; Pretend Play and Cognitive Development; Self-Development in Childhood and Adolescence; Sexism; Social and Emotional Development in the Context of the Family; Stereotypes in Social Psychology.

Bibliography

Barry III, H., Bacon, M.K., Child, I.L., 1957. A cross-cultural survey of some sex differences in socialization. Journal of Abnormal and Social Psychology 55, 327–332.

Beal, C.R., 1994. Boys and Girls: The Development of Gender Roles. McGraw-Hill, New York.

Berry, J.W., Poortinga, Y.H., Segall, M.H., Dasen, P.R., 2002. Cross-Cultural Psychology: Research and Applications, second ed. Cambridge University Press, New York.

Best, D.L., 2009. Another view of the gender-status relation. Sex Roles 61, 341–351.

Best, D.L., 2010. Gender. In: Bornstein, M.H. (Ed.), Handbook of Cultural Developmental Science. Taylor & Francis, New York, pp. 209–222.

Best, D.L., 2013. African perspectives on gender development. In: Tchombe, T.M.S., Nsamenang, A.B., Keller, H., Fülöp, M. (Eds.), Cross-Cultural Psychology: An Africa-Centric Perspective. Design House, Limbe, Cameroon. Bukhum Communications, Gainesville, FL, pp. 149–161.

Best, D.L., House, A.S., Barnard, A.E., Spicker, B.S., 1994. Parent-child interaction in France, Germany, and Italy: the effects of gender and culture. Journal of Cross-Cultural Psychology 25, 181–193.

Best, D.L., Williams, J.E., 1993. A cross-cultural viewpoint. In: Beall, A.E., Sternberg, R.J. (Eds.), The Psychology of Gender. Guilford Press, New York, pp. 215–248.

Best, D.L., Williams, J.E., Cloud, J.M., Davis, S.W., Robertson, L.S., Edwards, J.R., Giles, H., Fowles, J., 1977. Development of sex-trait stereotypes among young children in the United States, England, and Ireland. Child Development 48, 1375–1384.

Bloch, M.N., Adler, S.M., 1994. African children's play and the emergence of the sexual division of labor. In: Roopnarine, J.L., Johnson, J.E., Hooper, F.H. (Eds.), Children's play in Diverse Cultures. State University of New York Press, Albany, NY, pp. 148–178.

Bock, J., 2005. Farming, foraging, and children's play in the Okavango Delta, Botswana. In: Pelligrini, A.D., Smith, P.K. (Eds.), The Nature of Play: Great Apes and Humans. Guilford Press, New York, pp. 254–284.

Bronfenbrenner, U., 1989. Ecological systems theory. In: Vasta, R. (Ed.), Annals of Child Development, Theories of Child Development: Revised Formulations and Current Issues, vol. 6. JAI Press, Greenwich, CT, pp. 187–249.

Brown, J.K., 1970. A note on the division of labor by sex. American Anthropologist 72, 1073–1078.

Browning, D.S., Green, M.C., Witte Jr., J., 2006. Sex, Marriage, & Family in World Religions. Columbia University Press, New York.

Bussey, K., Bandura, A., 1984. Influence of gender constancy and social power on sex-linked modeling. Journal of Personality and Social Psychology 47, 1292–1302.

Bussey, K., Bandura, A., 1992. Self-regulatory mechanisms governing gender development. Child Development 63, 1236–1250.

Conrade, G., Ho, R., 2001. Differential parenting styles for fathers and mothers: differential treatment for sons and daughters. Australian Journal of Psychology 53, 29–35.

D'Andrade, R.G., 1966. Sex differences and cultural institutions. In: Maccoby, E.E. (Ed.), The Development of Sex Differences. Stanford University Press, Stanford, CA, pp. 174–204.

Davis, N.J., Robinson, R.V., 1991. Men's and women's consciousness of gender inequality: Austria, West Germany, Great Britain, and the United States. American Sociological Review 56, 72–84.

Dixon, R.B., 1978. Rural Women at Work: Strategies for Development in South Asia. Johns Hopkins University Press, Baltimore, MD.

Edelbrock, C., Sugawara, A.I., 1978. Acquisition of sex-typed preferences in preschool-aged children. Developmental Psychology 14, 614–623.

Edwards, C.P., 1992. Cross-cultural perspectives on family-peer relations. In: Parke, R.D., Ladd, G.W. (Eds.), Family-Peer Relationships: Modes of Linkages. Erlbaum, Mahwah, NJ, pp. 285–315.

Edwards, C.P., 2000. Children's play in cross-cultural perspective: a new look at the six cultures study. Cross-Cultural Research 34, 318–338.

Edwards, C.P., Whiting, B.B., 1974. Women and dependency. Politics and Society 4, 343–355.

Fagot, B.I., 1985. Beyond the reinforcement principle: another step toward understanding sex-role development. Developmental Psychology 21, 1097–1104.

Faisal, S., Wednesday September 24, 2003. Muslim Girls Struggle for Education. BBC Arabic Service. http://news.bbc.co.uk/2/hi/middle_east/3130234.stm.

Fausto-Sterling, A., 2000. Sexing the Body: Gender Politics and the Construction of Sexuality. Basic Books, New York.

Gallimore, R., Goldenberg, C.N., Weisner, T.S., 1993. The social construction and subjective reality of activity settings: implications for community psychology. American Journal of Community Psychology 21, 537–559.

Goldman, J.D., Goldman, R.J., 1983. Children's perceptions of parents and their roles: a cross-national study in Australia, England, North America, and Sweden. Sex Roles 9, 791–812.

Golombok, S., Fivush, R., 1994. Gender Development. Cambridge University Press, New York.

Gottlieb, G., 2002. On the epigenetic evolution of species-specific perception: the developmental manifold concept. Cognitive Development 17, 1287–1300.

Greenfield, P.M., Brazelton, T.B., Childs, C.P., 1989. From birth to maturity in Zinacantan: ontogenesis in cultural context. In: Bricker, V., Gosen, G. (Eds.), Ethnographic Encounters in Southern Mesoamerica: Celebratory Essays in Honor of Evon Z. Vogt. Institute of Mesoamerican Studies, State University of New York, Albany, NY, pp. 177–216.

Gronseth, E., 1957. The impact of father absence in sailor families upon the personality structure and social adjustment of adult sailor sons. Part I. In: Anderson, N. (Ed.), Studies of the Family, vol. 2. Vandenhoeck and Ruprecht, Gottingen, pp. 97–114.

Guttentag, M., Longfellow, C., 1977. Children's social attributions: development and change. In: Keasy, C.B. (Ed.), Nebraska Symposium on Motivation. University of Nebraska Press, Lincoln, NE, pp. 305–341.

Hagan, I.K., Kuebli, J., 2007. Mothers' and fathers' socialization of preschoolers' physical risk taking. Journal of Applied Developmental Psychology 28, 2–14.

Hamilton, V.L., Blumenfeld, P.C., Akoh, H., Miura, K., 1991. Group and gender in Japanese and American elementary classrooms. Journal of Cross-Cultural Psychology 22, 317–346.

Hanover, B., 2000. Development of self in gendered contexts. In: Eckes, T., Trautner, H.M. (Eds.), The Development of the Social Psychology of Gender. Erlbaum, Mahwah, NJ, pp. 177–206.

Hendrix, L., Johnson, G.D., 1985. Instrumental and expressive socialization: a false dichotomy. Sex Roles 13, 581–595.

Hennink, M., Rana, I., Iqbal, R., 2005. Knowledge of personal and sexual development amongst young people in Pakistan. Culture, Health & Sexuality 7, 319–332.

Herdt, G., 2010. Sex/gender, culture, and development: issues in the emergence of puberty and attraction. In: Worthman, C.M., Plotsky, P.M., Schechter, D.S., Cummings, C.A. (Eds.), Ormative Experiences: The Interaction of Caregiving, Culture, and Developmental Psychobiology. Cambridge University Press, New York, pp. 356–374.

Herdt, G., McClintock, M., 2000. The magical age of 10. Archives of Sexual Behavior 29, 587–606.

Hewlett, B.S., 1992. Husband-wife reciprocity and father-infant relationship among Aka Pygmies. In: Hewlett, B. (Ed.), Father-Child Relations: Cultural and Biosocial Contexts. Aldine de Gruyter, New York, pp. 153–176.

Hewlett, B.S., 2000. Culture, history and sex: anthropological perspectives on father involvement. Marriage and Family Review 29, 324–340.

Hoyenga, K.B., Hoyenga, K.T., 1993. Gender-Related Differences: Origins and Outcomes. Allyn & Bacon, Boston, MA.
Jacobs, J.A., Lim, S.T., 1992. Trends in occupational and industrial sex segregation in 56 countries, 1960–1980. Work and Occupations 19, 450–486.
Katz, M.M., Konner, M.J., 1981. The role of the father: an anthropological perspective. In: Lamb, M.E. (Ed.), The Role of the Father in Child Development, second ed. Wiley, New York, pp. 155–185.
Leaper, C., 1994. Childhood Gender Segregation: Causes and Consequences. In: New Directions for Child Development, vol. 65. Josey-Bass, San Francisco, CA.
Lee, J.Y., Sugawara, A.I., 1982. Awareness of sex-trait stereotypes among Korean children. Journal of Social Psychology 117, 161–170.
Leve, L.D., Fagot, B.I., 1997. Gender-role socialization and discipline processes in one- and two-parent families. Sex Roles 36, 1–21.
Lindsey, E.W., Mize, J., 2001. Parent-child physical and pretense play: links to children's social competence. Merrill-Palmer Quarterly 46, 565–591.
Lummis, M., Stevenson, H.W., 1990. Gender differences in beliefs and achievement: a cross-cultural study. Developmental Psychology 26, 254–263.
Lynn, D.B., Sawrey, W.L., 1959. The effects of father-absence on Norwegian boys and girls. Journal of Abnormal Social Psychology 59, 258–262.
Lytton, H., Romney, D.M., 1991. Parents' differential socialization of boys and girls: a meta-analysis. Psychological Bulletin 109, 267–296.
Maccoby, E.E., 1998. The Two Sexes: Growing up Apart, Coming Together. Belnap Press, Cambridge, MA.
Maccoby, E.E., Jacklin, C.N., 1987. Gender segregation in childhood. In: Reese, H.W. (Ed.), Advances in Child Development and Behavior, vol. 20. Academic Press, Orlando, FL, pp. 239–287.
Mackey, W.C., Day, R., 1979. Some indicators of fathering behaviors in the United States: a cross-cultural examination of adult male-child interaction. Journal of Marriage and the Family 41, 287–299.
Marlowe, F.W., 2005. Who tends Hadza children? In: Hewlett, B.S., Lamb, M.E. (Eds.), Hunter-Gatherer Childhood. Aldine/Transaction, New Brunswick, NJ, pp. 177–190.
Minturn, L., Lambert, W.W., 1964. Mothers of Six Cultures: Antecedents of Child Rearing. Wiley, New York.
Montero, M., 1978. Estereotipos sexuales en niños de areas rurales. In: Paper presented at the Congreso de Sexualidad Femenina. Sociedad Venezolana de Psiquiatria, Caracas.
Mukhopadhyay, C.C., Higgins, P.J., 1988. Anthropological studies of women's status revisited: 1977–1987. Annual Review of Anthropology 17, 461–495.
Munroe, R.H., Munroe, R.L., 1971. Household density and infant care in an East African society. Journal of Social Psychology 83, 3–13.
Munroe, R.L., Munroe, R.H., 1975/1994. Cross-Cultural Human Development. Waveland Press, Inc., Prospective Heights, IL.
Organisation for Economic Co-operation and Development (OECD), 2004a. Education at a Glance. OECD, Paris.
Organisation for Economic Co-operation and Development (OECD), 2004b. Learning for Tomorrow's World: First Results from Programme for International Student Assessment (PISA) 2003. OECD, Paris.
Ottaway, R.N., Bhatnagar, D., 1988. Personality and biographical differences between male and female managers in the United States and India. Applied Psychology: An International Review 37, 201–212.
Population Crisis Committee, June 1988. Country Rankings of the Status of Women: Poor, Powerless, and Pregnant (Issue Brief No. 20). Population Crisis Committee, Washington, DC.
Prossner, G.V., Hutt, C., Hutt, S.J., Mahindadasa, K.J., Goonetilleke, M.D.J., 1986. Children's in Sri Lanka: a cross-cultural study. British Journal of Developmental Psychology 4, 179–186.
Quinn, N., 1977. Anthropology studies of women's status. Annual Review of Anthropology 6, 181–225.
Raag, T., Rackliff, C.L., 1998. Preschoolers' awareness of social expectations of gender: relationships to toy choices. Sex Roles 38, 685–700.
Rogoff, B., Gauvain, M., Ellis, S., 1984. Development viewed in its cultural context. In: Bornstein, M.H., Lamb, M.E. (Eds.), Developmental Psychology: An Advanced Textbook. Erlbaum, Hillsdale, NJ, pp. 533–571.
Rohner, R.P., Rohner, E.C., 1982. Enculturative continuity and the importance of caretakers: cross-cultural codes. Behavior Science Research 17, 91–114.
Russell, A., Aloa, V., Feder, T., Glover, A., Miller, H., Palmer, G., 1998. Sex-based differences in parenting styles in a sample with preschool children. Australian Journal of Psychology 50, 89–99.
Rust, J., Golombok, S., Hines, M., Johnston, K., Golding, J., 2000. The role of brothers and sisters in the gender development of preschool children. Journal of Experimental Child Psychology 77, 292–303.
Seavey, C.A., Katz, P.A., Zalk, S.R., 1975. Baby X: the effect of gender labels on adult responses to infants. Sex Roles 1, 103–110.
Segal, E.S., 1983. The structure of division of labor: a tentative formulation. Behavior Science Research 18, 3–25.
Serpell, R., 1993. The Significance of Schooling: Life-Journeys in an African Society. Cambridge University Press, New York.
Sidorowicz, L.S., Lunney, G.S., 1980. Baby X revisited. Sex Roles 6, 67–73.
Simpkins, S.D., Fredricks, J.A., Eccles, J.S., 2012. Charting the Eccles' expectancy-value model from mothers' beliefs in childhood to youths' activities in adolescence. Developmental Psychology 48, 1019–1032.
Stevenson, M.R., Black, K.N., 1988. Paternal absence and sex role development: a meta-analysis. Child Development 59, 793–814.
Stolz, L.M., 1954. Father Relations of Warborn Children. Stanford University Press, Palo Alto.
Super, C.M., Harkness, S., 1986. The developmental niche: a conceptualization at the interface of child and culture. International Journal of Behavioral Development 9, 545–569.
Sweeney, J., Bradbard, M.R., 1989. Mothers' and fathers' changing perceptions of their male and female infants over the course of pregnancy. Journal of Genetic Psychology 149, 393–404.
Tarrier, N., Gomes, L., 1981. Knowledge of sex-trait stereotypes: effects of age, sex and social class on Brazilian children. Journal of Cross-Cultural Psychology 12, 81–93.
Thompson, S.K., 1975. Gender labels and early sex role development. Child Development 46, 339–347.
Turner, P.J., Gervai, J., 1995. A multidimensional study of gender typing in preschool children and their parents: personality, attitudes, preferences, behavior, and cultural differences. Developmental Psychology 31, 759–772.
United Nations Children's Fund (UNICEF), 2004. Girls' Education: Introduction. Website Update as of November 2004. www.unicef.org/girlseducation/index.html.
United Nations Children's Fund (UNICEF), 2005. Early Marriage: A Harmful Traditional Practice. A Statistical Exploration. http://www.unicef.org/publications/files/Early_Marriage_12.lo.pdf.
Van Leeuwen, M.S., 1978. A cross-cultural examination of psychological differentiation in males and females. International Journal of Psychology 13, 87–122.
Weinraub, M., Clemens, L.P., Sockloff, A., Etheridge, R., Gracely, E., Myers, B., 1984. The development of sex role stereotypes in the third year: relationships to gender labeling, gender identity, sex-typed toy preferences, and family characteristics. Child Development 55, 1493–1503.
Weisner, T.S., 1984. A cross-cultural perspective: ecocultural niches of middle childhood. In: Collins, A. (Ed.), The Elementary School Years: Understanding Development during Middle Childhood. National Academy Press, Washington, DC, pp. 335–369.
Weisner, T.S., Gallimore, R., 1977. My brother's keeper: child and sibling caretaking. Current Anthropology 18, 169–190.
West, M.M., Konner, M.J., 1976. The role of the father: an anthropological perspective. In: Lamb, M.E. (Ed.), The Role of the Father in Child Development. Plenum Press, New York, pp. 185–216.
Whiting, B.B., 1963. Six Cultures: Studies of Child Rearing. Wiley, Oxford, England.
Whiting, B.B., Edwards, C.P., 1973. A cross-cultural analysis of sex differences in the behavior of children aged 3 to 11. Journal of Social Psychology 91, 171–188.
Whyte, M.K., 1978. The Status of Women in Preindustrial Societies. Princeton University Press, Princeton, NJ.
Williams, J.E., Best, D.L., 1982/1990. Measuring Sex Stereotypes: A Multination Study (Rev. ed., 1990). Sage, Newbury Park, CA.
World Economic Forum, 2013. The Global Gender Gap Report 2013. World Economic Forum, Switzerland. http://www.weforum.org/issues/global-gender-gap.

Gender Differences in Personality and Social Behavior

Marco Del Giudice, University of New Mexico, Albuquerque, NM, USA

© 2015 Elsevier Ltd. All rights reserved.

This article replaces the previous edition article by J.S. Hyde, volume 9, pp. 5989–5994, © 2001, Elsevier Ltd.

Abstract

The scientific study of gender differences has yielded a wealth of robust generalizations about the way males and females differ across domains, cultures, and developmental stages. This article provides a descriptive overview of gender differences in personality, social cooperation and competition (including aggression and play), and verbal and nonverbal communication. Qualitative statements about the typical behavior of males and females are supplemented with quantitative data from *meta-analyses* and other large-scale studies. Key methodological issues in the measurement of gender differences are discussed to aid the reader in the interpretation of research findings.

The scientific study of gender differences has a long history, spanning more than a century of research (Ellis et al., 2008). Competing theories of gender vary in the role they assign to evolutionary history, cultural practices, endocrine and neurobiological mechanisms, and individual learning; they also vary in the extent to which they regard these levels of explanation as complementary or mutually exclusive. While the field is rife with theoretical debate and controversy, decades of empirical research have yielded a wealth of robust generalizations about the way males and females differ across domains, cultures, and developmental stages. This process has been accelerated by the diffusion of *meta-analytic* techniques over the last 30 years. Thanks to *meta-analysis*, the results of multiple studies can be aggregated and corrected for various sources of error; also, the effects of potential moderators can be analyzed in detail.

This article provides a descriptive overview of gender differences in personality; social cooperation and competition (including aggression and play); and verbal and nonverbal communication. Whenever possible, qualitative statements about the typical behavior of males and females are supplemented with quantitative data from *meta-analyses* and other large-scale studies. To aid the reader in the interpretation of these findings, the initial section presents and discusses key methodological issues in the measurement of gender differences.

Methodological Issues

The Interpretation of Effect Sizes

In psychological research, group differences (including differences between males and females) are usually expressed in terms of standardized effect sizes. The most common effect size for group and gender differences is Cohen's d, the difference between means divided by the pooled standard deviation of the two groups (e.g., $d = 0.50$ represents a distance of half a standard deviation between male and female means). Conventionally, positive values of d indicate that the male mean is higher than the female mean, whereas negative values indicate that the female mean is higher. Values of d can easily be translated into estimates of statistical overlap (**Figure 1**). For example, $d = 0.50$ corresponds to a 67% overlap between male and female distributions (assuming normality). The precision with which a variable is measured can dramatically affect the resulting effect size, since d becomes artificially smaller as measurement error increases. Many psychological variables are measured with a substantial margin of error; thus, empirical values of d should be interpreted as lower bound estimates of the actual differences. Effect sizes that have been corrected for measurement error are explicitly flagged as such in the remainder of this article.

Regrettably, many researchers routinely interpret effect sizes as 'small,' 'moderate,' or 'large' based on the conventional cutoffs ($d = 0.20$, 0.50, and 0.80) originally proposed by Cohen (1988). This practice has little basis in statistical theory and is discouraged by most methodologists (Breaugh, 2003; Hedges, 2008; Vacha-Haase and Thompson, 2004). Ironically, Cohen himself advised against using his tentative cutoffs except as a last resort method for determining sample size in exploratory studies (Cohen, 1988). The substantive

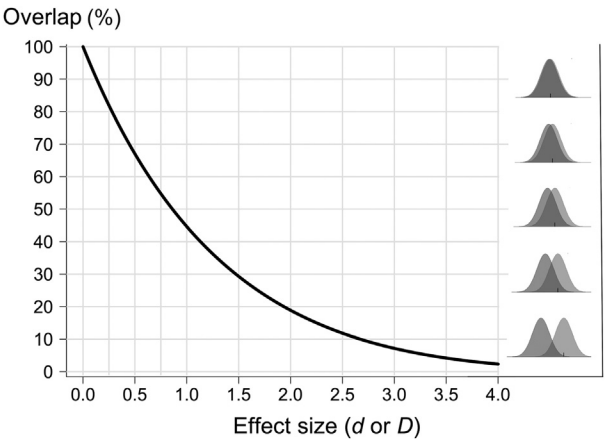

Figure 1 Relation between effect size (d or D) and percent overlap between male and female distributions. The illustrative plots on the right show two univariate distributions overlapping by 10%, 30%, 50%, 70%, and 90% of their joint area. Overlap is computed assuming multivariate normality. See Del Giudice, M., 2009. On the real magnitude of psychological sex differences. Evolutionary Psychology 7, 264–279.

interpretation of an effect size may vary dramatically depending on measurement issues, theoretical considerations, and the scientific question being asked. For example, differences between group means have a progressively larger impact as one moves toward the distribution tails. Even when male and female distributions are largely overlapping in the region surrounding the mean of a trait, individuals with *extreme* values of the same trait may still be predominantly male or female. In the light of these considerations, debating whether males and females are 'more similar' or 'more different' based on a rigid interpretation of effect sizes (e.g., Hyde, 2005, 2013) is unlikely to prove a fruitful strategy for understanding the meaning of gender differences and their societal implications.

Univariate versus Multivariate Differences

The standard approach to gender differences in psychology is to measure groups of males and females on multiple variables – for example physical, verbal, and relational aggression – and compute an effect size for each individual variable. This approach becomes problematic in dealing with intrinsically multidimensional domains such as personality and aggression. To begin with, small differences in multiple traits can add up to a much larger effect if the traits are considered simultaneously. For example, facial features – nose length, eye size, face width, and so forth – show considerable overlap between genders when they are considered one by one. However, their *combined* effect results in nearly complete separation between male and female faces, and indeed, human observers can classify faces by gender with more than 95% accuracy (see Del Giudice, 2013). Another limitation of univariate effect sizes (such as d) is that they cannot take into account the patterns of *correlation* between the variables that constitute a given psychological domain.

The limitations of univariate effect sizes can be overcome by supplementing them with multivariate effect sizes that simultaneously consider multiple variables and their correlational structure. The natural extension of Cohen's d is the Mahalanobis D, a standardized difference between two groups in multivariate space. When employed to measure gender differences, D can be interpreted as the distance between group means on an abstract dimension of *masculinity–femininity* in the relevant domain. The interpretation of D in terms of distribution overlap is exactly the same as that of d (**Figure 1**). Unlike d, D is always positive since the concept of a higher or lower mean is meaningless in multidimensional space (Del Giudice, 2009). The application of multivariate effect sizes to gender differences has been debated by some authors (e.g., Hyde, 2013); however, the arguments leveled against D turn out to be either incorrect or logically inconsistent (Del Giudice, in press). In this article, both univariate and multivariate effect sizes are reported whenever the latter are available.

Dimensional versus Taxonic Differences

Finding evidence that, on average, males and females differ on a certain set of variables does not tell the whole story about the nature of gender differences in that domain. Another important question is whether gender differences are differences in *degree* – with males and females varying along a single continuum – or differences in *kind*, implying a clear-cut discontinuity in the distribution of variables between groups. This question is conceptually independent from questions about the *size* of those differences.

Taxometrics is a set of statistical methods that attempt to determine whether a certain group is a *taxon*, i.e., a clearly demarcated category that differs in kind and not just in degree from another group. In a recent paper, Carothers and Reis (2013) employed taxometrics to determine whether binary gender categories can be regarded as dimensional or taxonic in a number of domains including personality, social behavior, and sexuality. The results overwhelmingly supported a dimensional model of gender, with the exception of *sociosexuality* (the preference for promiscuous, uncommitted sex versus monogamous, committed partnerships). However, these pioneering findings must be taken with caution, since taxometric methods are unlikely to detect an existing taxonic structure if the mean difference between groups is smaller than about $d = 1.20$ (Beauchaine, 2007). Since the size of gender differences in many psychological traits falls below this threshold (see below), taxometric results are difficult to interpret with any confidence.

Gender Differences in Trait Variability

Males and females may differ not just in the mean value of a trait or behavior, but also in their *variability* around the mean. When such differences are found, the general tendency is for males to display more variability than females (Archer and Mehdikhani, 2003; Hyde, 2013). The more variable gender will tend to be overrepresented at both tails of the distribution, even if mean differences are small or absent (see Voracek et al., 2013). The standard metric for gender differences in variability is the variance ratio (VR), the ratio of male variance to female variance. $VR = 1$ indicates equality of variances, $VR > 1$ indicates higher variability in males, and $VR < 1$ indicates higher variability in females.

Gender Differences in Personality

Descriptive models of personality have a hierarchical structure, from broad and inclusive traits to narrow and specific ones. Typically, 3–6 broad traits ('domains' or 'global traits') are subdivided into 10–20 narrower traits ('aspects' or 'specific factors'); at the lowest level of description, dozens of specific personality 'facets' (usually 30–45) can be identified. Patterns of gender differences are already apparent at the level of broad traits, but become stronger and more recognizable at a finer-grained level of analysis.

Broad Personality Traits

Big Five Domains

The most widely adopted model of personality is the Five Factor Model (FFM), whose five domains are known as the Big Five: *neuroticism* (negative emotionality and emotional instability), *agreeableness* (altruism and cooperation), *conscientiousness* (self-control, self-discipline, and organization), *extraversion* (sociability, assertiveness, and positive emotionality), and *openness* (imagination, intellectual curiosity, and aesthetic appreciation). Across countries, gender differences in the Big

Five are largest in agreeableness ($d = -0.40$) and neuroticism ($d = -0.34$); females consistently score higher than males on both traits. Smaller differences favoring females are found in conscientiousness ($d = -0.09$) and extraversion ($d = -0.11$); finally, there are no reliable differences in openness ($d = 0.01$). There is some cross-cultural variability around these averages; as a rule, gender differences tend to be *larger* in individualistic, gender-egalitarian societies (Feingold, 1994; Lippa, 2010; Schmitt et al., 2008). In a reanalysis of a large US data set, Del Giudice (2009) found a multivariate difference of $D = 0.98$ (corrected) for the Big Five domains, corresponding to a 45% overlap between male and female distributions.

A recent study of variation in Big Five domains (Borkenau et al., 2013) found equal variance between genders in self-reported traits ($VR \approx 1$), but higher male variance in informant-reported traits ($VR \approx 1.20$). Neuroticism was an exception: males showed lower variability in self-reports ($VR = 0.89$) and equal variability in informant-reports ($VR = 1.01$). Given the comparatively small size of the study, these findings should be regarded as preliminary.

The Interpersonal Circumplex

In the interpersonal circumplex model (IPC), individual differences in personality are mapped on two independent axes of variation, *dominance* (assured-dominant versus insecure-submissive) and *nurturance* (warm-agreeable versus cold-hearted). The IPC can be regarded as an alternative representation of the extraversion and agreeableness domains of the FFM (DeYoung et al., 2013). The interpersonal dimensions of the IPC bring into sharp relief a pattern of gender differences that is partially masked by the structure of Big Five domains: on the traits that best correspond to those dimensions, males consistently score higher in dominance ($d \approx 0.20$–0.50) and lower in nurturance ($d \approx -0.25$ to -1.00; Costa et al., 2001; Del Giudice et al., 2012; Feingold, 1994). Based on data for dominance and warmth reported in Del Giudice et al. (2012), the multivariate effect size for the IPC can be estimated at $D = 1.14$ (corrected).

Temperament

In infants and children, individual differences in behavior are usually framed as 'temperament' rather than 'personality'; standard dimensions of temperament overlap only partially with adult personality traits. In a *meta-analysis* of temperament ratings from 3 months to 13 years of age, Else-Quest et al. (2006) found that females scored higher in all dimensions of *effortful control* (self-control and delay of gratification; $d = -0.05$ to -1.01), whereas males tended to score higher in dimensions of *surgency* (activity and sociability; $d = -0.11$ to 0.55). No reliable gender differences were observed in *negative affectivity* ($d = -0.17$ to 0.13), a surprising finding when contrasted with the robust gender differences in adult neuroticism. The mean variance ratios were $VR = 1.08$ for dimensions of effortful control, $VR = 1.03$ for dimensions of surgency, and $VR = 0.98$ for dimensions of negative affectivity.

Narrow Personality Traits

While broad traits such as the Big Five provide a compact description of individual differences, they are not necessarily the best level of analysis to investigate gender differences in personality. In many cases, the structure of broad traits masks the existence of lager gender differences at the level of narrower traits. For example, comparatively small differences in extraversion actually result from the combination of larger, opposite-sign differences in assertiveness–sensation seeking (higher in males) and sociability–gregariousness (higher in females). Weisberg et al. (2011) examined the 10 aspects of the Big Five and detected contrasting patterns of gender differences within extraversion ($d = -0.23$ for enthusiasm, $d = 0.09$ for assertiveness), conscientiousness ($d = -0.18$ for orderliness, $d = 0.06$ for industriousness), and openness ($d = -0.27$ for aesthetic openness, $d = 0.22$ for intellect). At the level of 30 personality facets, the *meta-analysis* by Costa et al. (2001) found differences of $d = -0.09$ to -0.44 for neuroticism facets, $d = -0.33$ to 0.38 for extraversion facets, $d = -0.35$ to 0.32 for openness facets, $d = -0.17$ to -0.43 for agreeableness facets, and $d = -0.13$ to 0.20 for conscientiousness facets.

The 15 primary factors (PF) of the 16 PF model are narrow personality traits roughly comparable to Big Five aspects. In a large US sample, Del Giudice et al. (2012) found differences ranging from $d = -2.29$ to 0.54 (corrected). The largest differences were in *sensitivity* (aesthetic, intuitive, and tender-minded versus utilitarian, objective, and tough-minded; $d = -2.29$), *warmth* ($d = -0.89$), *apprehension* (worried, insecure, and self-doubting versus secure and self-assured; $d = -0.60$), *dominance* ($d = 0.54$), and *emotional stability* ($d = 0.53$). The multivariate effect size for the 15 factors was $D = 2.71$ (corrected), corresponding to a 10% overlap between male and female distributions (Del Giudice et al., 2012).

Impulsivity, Risk-Taking, and Sensation Seeking

Impulsivity, risk-taking, and sensation seeking are a group of correlated traits that deserve special consideration in virtue of their social and clinical implications. In a *meta-analysis* of risk-taking that employed a broad definition of the term and a wide range of indicators – including for example 'smoking' and 'guessing' – Byrnes et al. (1999) found that overall levels of risk-taking were higher in males ($d = 0.13$). Cross et al. (2011) employed a narrower definition of risk-taking and found even stronger gender differences ($d = 0.37$ across methods). In the same *meta-analysis*, males also showed higher levels of sensation seeking ($d = 0.22$). Gender differences in impulsivity were considerably smaller ($d = 0.08$). This finding can be contrasted with the robust gender differences in effortful control (the reverse of impulsivity) found in studies of infant and child temperament. Males displayed more variability in risk-taking ($VR = 1.10$) and sensation seeking ($VR = 1.08$), but not in impulsivity ($VR = 1.00$). Gender differences in risk-taking are not limited to stable dispositions, but extend to the role of contextual factors in determining moment-to-moment attitudes toward risk. For example, several studies have shown that acute stress tends to increase risk-taking in males, but decrease it in females (Mather and Lighthall, 2012).

Summary

The typical personality profiles of males and females differ in a number of ways. On average, males are more dominant,

assertive, risk-prone, tough-minded, cold-hearted, emotionally stable, utilitarian, and open to abstract ideas. Females are more nurturant, warm, altruistic, submissive, risk-averse, tender-minded, emotionally unstable, and open to feelings and aesthetic experiences. These differences tend to be larger in more gender-egalitarian countries, and together define a global dimension of personality masculinity–femininity. The global overlap between adult men and women in the distribution of personality traits can be estimated at about 10% – close to that between male and female facial features (Del Giudice, 2013). In addition, men are somewhat more variable in most (but not all) dimensions of personality. Gender differences in temperament are already apparent in infancy and childhood, but correspond only in part to those in adult personality. In particular, early differences in impulsivity seem to dissipate in adults, while differences in negative emotionality emerge during development.

Gender Differences in Cooperation and Competition

Cooperation and competition are the basic polarities of social life; their constant interplay provides the background for the rich tapestry of human relationships. From childhood to adulthood, cooperation and competition outside the family occur with particular frequency and intensity between same-gender individuals (Benenson, 2013; Geary, 2010; Geary et al., 2003). Both one's closest friends and allies and one's fiercest adversaries are likely to be other individuals of the same gender. The dynamics of same-gender relations are an essential aspect of human social behavior, and tend to carry over into mixed-gender interactions.

Cooperation and Competition in Males

While both genders participate in many kinds of social interactions, ranging from dyadic friendships to large groups, large-scale cooperation is a distinctive attribute of male sociality across the life span. Males are more likely to engage in cooperative group activities, form larger groups than females, and engage in frequent and intense between-group competition against other male groups. Male groups are characterized by stable internal hierarchies of status and dominance; within groups, competition tends to be individualistic or one-on-one (Benenson, 2013; Geary, 2010; Geary et al., 2003). A *meta-analysis* of cooperation in experimental games played between strangers (Balliet et al., 2011) showed that male–male interactions tend to be more cooperative than female–female interactions ($d = 0.16$). At the same time, males tend to show less altruistic and prosocial behavior across a range of domains, from sharing possessions to organ donations (Ellis et al., 2008). This is consistent with males' lower levels of agreeableness and nurturance (see above); the main exception to this pattern is 'heroic' helping behavior, especially when directed toward women (Eagly and Crowley, 1986). In group settings – including mixed gender groups – males display higher levels of task-oriented behavior and lower levels of social-emotional behavior ($d = 0.59$ and -0.59 in the *meta-analysis* by Carli, 1982).

Cooperation and Competition in Females

The dynamics of cooperation and competition in females are more complex and nuanced than their male counterparts. Female friendship tends to be more complex and less focused on shared activities than male friendships. Starting from middle childhood, females are more likely to form dyadic relationships characterized by high levels of intimacy, emotional support, exclusivity, and reciprocity. In turn, dyads are embedded in larger social networks in which hierarchies tend to be more fluid than those of male groups (Ellis et al., 2008; Geary, 2010; Geary et al., 2003). Peer relations in females are strongly influenced by norms of caring and equality. Superiority and overt competition are discouraged, and – unlike in male groups – high-status peers are often the target of dislike and denigration. Accordingly, female competition is often indirect and balanced by egalitarian concerns; displays of overt competitiveness are usually restricted to high-status females (Benenson, 2013). In group settings, females display higher levels of social emotional behavior focused on maintaining and managing personal relationships ($d = -0.59$ in Carli, 1982). This pattern is consistent with females' higher levels of agreeableness and nurturance and lower levels of dominance and assertiveness (see above). A prominent feature of female social relationships is social exclusion, an indirect aggression tactic in which a target individual is ostracized by the others. Both threats of social exclusion and actual episodes of ostracism are more common in female groups, especially in childhood and adolescence (Benenson, 2013).

Aggression

The existence of gender differences in aggression is one of oldest and most robust findings in this area. In psychology, the standard distinction is that between *physical*, *verbal*, and *indirect* (or *relational*) forms of aggression. Indirect aggression is used to damage another individual's social standing by ostracizing, gossiping, spreading malicious rumors, and so forth. In a comprehensive *meta-analysis*, Archer (2009) found that, across age groups, males engage in more direct aggression, both physical ($d = 0.58$) and verbal ($d = 0.29$), while females tend to engage in more indirect aggression ($d = -0.16$). A *meta-analysis* of studies of children and adolescents (Card et al., 2008) showed similar results for physical and verbal aggression ($d = 0.73$ and 0.38, respectively), and a smaller effect for indirect aggression ($d = -0.06$). Gender differences in aggression – especially physical aggression – peak between adolescence and young adulthood (Archer, 2009). Overall, the multivariate effect size for male versus female patterns of aggression can be estimated at about $D = 0.89$–1.01 (corrected; Del Giudice, 2009).

While males and females tend to display similar levels of indirect aggression (especially compared with the much larger difference in physical aggression), the finer-grained dynamics of aggressive behavior differ between male and female groups. In particular, attractive females are the target of *more* indirect aggression by other females, whereas attractive males tend to receive *less* indirect aggression from their peers

(Vaillancourt, 2013). Also, experimental studies show that males tend to engage in *unprovoked* aggression more frequently than females ($d = 0.33$; Bettencourt and Miller, 1996).

Social Play

Social play is a universal feature of human development, and – like many other social activities – involves a complex balance of cooperation and competition. Gender differences in social play can be summarized as follows. Across cultures, girls engage more frequently in cooperative, nonaggressive social play, whereas boys show a higher frequency of play fighting and 'rough-and-tumble' play (about 3–6 times as much as girls; Geary, 2010), as well as higher levels of between-group competition. Both genders engage in sociodramatic play, in which social episodes are enacted based on everyday or fantastic themes. However, boys' themes more frequently involve power, dominance, and aggression, whereas girls' themes tend to involve interpersonal and family relationships (including play parenting). Gender differences in play behavior peak in middle childhood, around 8–10 years of age (Ellis et al., 2008; Geary, 2010).

Summary

Males tend to form larger, activity-oriented, competitive groups in which hierarchies tend to be stable and individual relationships require comparatively little emotional investment. Competition often involves direct forms of aggression. Intimate, high-intensity dyadic relationships play a bigger role in female social networks. Female groups are more emotion-focused and are characterized by unstable hierarchies and strong egalitarian norms; competition is often indirect and less openly confrontational. Aggression shows robust patterns of gender differences. Direct aggression – especially physical aggression – is higher in males; indirect forms of aggression are higher in females, though the size of the effect is considerably smaller. Gender differences in play behavior peak in middle childhood and mirror the broader pattern of gendered social dynamics: males engage in more play fighting and between-group competition, while females engage in more cooperative play centered around relational themes.

Gender Differences in Communication

Social behavior would be all but impossible without communication. In humans, the interplay between language and nonverbal behavior forms the background for an exceedingly complex communication system. Research on gender differences in linguistic communication has focused mainly on verbal ability (e.g., vocabulary, reading ability), language use (e.g., assertive communication), and conversational style (e.g., interruptions). Gender differences in nonverbal communication have been extensively studied, and include both the production of nonverbal displays – such as gestures and facial expressions – and the ability to accurately decode other people's nonverbal behavior.

Verbal Ability

In a classic meta-analysis, Hyde and Linn (1988) found a female advantage in general verbal ability ($d = -0.20$). Specific abilities showed a more nuanced pattern of results: females performed better on measures of speech production ($d = -0.33$) and anagram solving ($d = -0.22$); smaller differences were found in writing and reading ($d = -0.09$ and -0.03, respectively), while males showed an advantage in verbal analogies ($d = 0.16$). No reliable gender differences were detected in vocabulary skills ($d = -0.02$). The size of gender differences was largely independent of age, and showed no clear developmental trend from childhood to adulthood.

Language Use and Conversational Style

Given the robust gender differences observed in dominance- and nurturance-related dimensions of personality (see above), it is reasonable to expect males and females to differ in their typical patterns of language use and conversational style. Indeed, *meta-analytic* data in both children and adults show that, averaging across contexts, males tend to use more assertive speech ($d \approx 0.10$) whereas females tend to use more affiliative speech ($d \approx -0.10$). In addition, girls are more talkative than boys ($d = -0.11$), but the difference is reversed in adults ($d = 0.14$; Leaper and Ayres, 2007; Leaper and Smith, 2004). A related finding (Leaper and Robnett, 2011) is that females use more tentative speech, including more tag questions, e.g., "isn't it?" ($d = -0.23$) and hedges, e.g., "I think" ($d = -0.15$). A conspicuous aspect of gender differences in language use is their strong dependence on contextual factors such as group size, age and gender of conversation partners, and familiarity. Depending on the specifics of the conversational setting, these differences can easily increase, decrease, or even change sign (Leaper and Ayres, 2007; Leaper and Smith, 2004). On average, males interrupt their conversation partners more than females do ($d = 0.15$); gender differences are especially large in the use of *intrusive* interruptions ($d = 0.33$), which can be unambiguously interpreted as a form of dominant behavior (Anderson and Leaper, 1998).

Nonverbal Communication

Nonverbal Behavior

A consistent finding in nonverbal behavior research is that females are generally more expressive than males. A *meta-analysis* by Hall (1984) found an overall effect size $d = -1.01$ for facial expressiveness and $d = -0.58$ for bodily expressiveness, even if males tend to adopt more expansive, open postures ($d = 1.04$). Females are also better at deliberately expressing emotions ($d = -0.52$). Gender differences in expressiveness show a clear developmental trend from childhood to adulthood (Chaplin and Aldao, 2013). In children, a smaller female advantage is found on positive emotions ($d \approx -0.08$) and *internalizing* negative emotions such as sadness and anxiety ($d \approx -0.10$); at the same time, boys tend to be more expressive on *externalizing* negative emotions such as anger ($d \approx 0.09$). By adolescence, however, females become more expressive across the board

and effect sizes increase ($d \approx -0.35$ for negative emotions and -0.28 for positive emotions).

Smiling and crying are two prototypical manifestations of expressive behavior. A *meta-analysis* of smiling from adolescence to adulthood showed that females tend to smile more frequently across contexts and cultures ($d = -0.41$). Average differences peak in adolescence ($d = -0.56$) and decrease with age ($d = -0.11$ in seniors); also, effect sizes are especially large in North American countries (LaFrance et al., 2003). As with language use, gender differences in smiling are highly contingent on contextual factors; for example, males and females differ more when they believe they are being observed than when they believe they are alone ($d = -0.46$ vs -0.19). Gender differences in crying are larger and more robust than those in smiling. Across cultures, females show a stronger tendency to cry ($d = -1.11$) and cry more often than males do ($d = -0.94$). Also, gender differences in crying tend to be *larger* in more individualistic, gender-egalitarian countries (van Hemert et al., 2011).

Nonverbal Decoding

In addition to being more expressive, females are generally better at processing and decoding other people's nonverbal behavior. In her *meta-analysis*, Hall (1984) found an overall effect size $d = -0.43$ for nonverbal decoding skills across communicative modalities. The female advantage in processing facial expressions tends to be smaller in childhood and adolescence ($d \approx -0.20$; McClure, 2000). There is some evidence that gender differences in expression processing may peak in infancy, decrease in early childhood, then peak again in middle childhood (McClure, 2000).

Summary

Females enjoy a general advantage in communication skills, both in the verbal domain (with the exception of verbal analogies) and in the production and decoding of nonverbal displays. The higher expressiveness of females is reflected in a higher frequency of both smiling and crying; gender differences in crying are especially robust, and tend to increase in more gender-egalitarian countries. Consistent with the data on personality and social relations, males tend to use more assertive speech and interrupt their conversation partners more often, while females tend to use more affiliative and tentative speech. However, gender differences in language use are highly dependent on contextual factors, testifying to the remarkable strategic flexibility of human communication.

See also: Crime and Gender; Evolutionary Social Psychology; Gender Development: Cultural Differences; Language and Gender; Personality Development and Temperament; Personality, Evolutionary Models of; Psychometrics; Systematic Reviewing and Meta-Analysis.

Bibliography

Anderson, K.J., Leaper, C., 1998. Meta-analyses of gender effects on conversational interruption: who' what' when' where' and how. Sex Roles 39, 225–252.
Archer, J., 2009. Does sexual selection explain human sex differences in aggression? Behavioral and Brain Sciences 32, 249–266.
Archer, J., Mehdikhani, M., 2003. Variability among males in sexually selected attributes. Review of General Psychology 7, 219–236.
Balliet, D., Li, N.P., Macfarlan, S.J., Van Vugt, M., 2011. Sex differences in cooperation: a meta-analytic review of social dilemmas. Psychological Bulletin 137, 881–909.
Beauchaine, T.P., 2007. A brief taxometrics primer. Journal of Clinical Child and Adolescent Psychology 36, 654–676.
Benenson, J.F., 2013. The development of human female competition: allies and adversaries. Philosophical Transactions of the Royal Society B 368, 20130079.
Bettencourt, B.A., Miller, N., 1996. Gender differences in aggression as a function of provocation: a meta-analysis. Psychological Bulletin 119, 422–447.
Borkenau, P., Hřebíčová, M., Kuppens, P., Realo, A., Allik, J., 2013. Sex differences in variability in personality: a study in four samples. Journal of Personality 81, 49–60.
Breaugh, J.A., 2003. Effect size estimation: factors to consider and mistakes to avoid. Journal of Management 29, 79–97.
Byrnes, J.P., Miller, D.C., Schafer, W.D., 1999. Gender differences in risk taking: a meta-analysis. Psychological Bulletin 125, 367–383.
Card, N.A., Stucky, B.D., Sawalani, G.M., Little, T.D., 2008. Direct and indirect aggression during childhood and adolescence: a meta-analytic review of gender differences, intercorrelations, and relations to maladjustment. Child Development 79, 1185–1229.
Carli, L.L., 1982. Are women more social and men more task oriented? A meta-analytic review of sex differences in group interaction, reward allocation, coalition formation, and cooperation in the Prisoner's Dilemma game. Unpublished manuscript, University of Massachusstes, Amherst.
Carothers, B.J., Reis, H.T., 2013. Men and women are from earth: examining the latent structure of gender. Journal of Personality and Social Psychology 104, 385–407.
Chaplin, T.M., Aldao, A., 2013. Gender differences in emotion expression in children: a meta-analytic review. Psychological Bulletin 139, 735–765.
Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. Erlbaum, Hillsdale, NJ.
Costa, P.T., Terracciano, A., McCrae, R.R., 2001. Gender differences in personality traits across cultures: robust and surprising findings. Journal of Personality and Social Psychology 81, 322–331.
Cross, C.P., Copping, L.T., Campbell, A., 2011. Sex differences in impulsivity: a meta-analysis. Psychological Bulletin 137, 97–130.
Del Giudice, M., 2009. On the real magnitude of psychological sex differences. Evolutionary Psychology 7, 264–279.
Del Giudice, M. 2013. Multivariate misgivings: is D a valid measure of group and sex differences? Evolutionary Psychology 11, 1067–1076.
Del Giudice, M., Booth, T., Irwing, P., 2012. The distance between Mars and Venus: measuring global sex differences in personality. PLoS One 7, e29265.
DeYoung, C.G., Weisberg, Y.J., Quilty, L.C., Peterson, J.B., 2013. Unifying the aspects of the Big Five, the interpersonal circumplex, and trait affiliation. Journal of Personality 81, 465–475.
Eagly, A.H., Crowley, M., 1986. Gender and helping behavior: a meta-analytic review of the social psychological literature. Psychological Bulletin 100, 283–308.
Ellis, L., Hershberger, S., Field, E., Wersinger, S., Pellis, S., Geary, D., Palmer, C., Hoyenga, K., Hetsroni, A., Karadi, K., 2008. Sex Differences: Summarizing More than a Century of Scientific Research. Psychology Press, New York.
Else-Quest, N.M., Hyde, J.S., Goldsmith, H.H., Van Hulle, C.A., 2006. Gender differences in temperament: a meta-analysis. Psychological Bulletin 132, 33–72.
Feingold, A., 1994. Gender differences in personality: a meta-analysis. Psychological Bulletin 116, 429–456.
Geary, D.C., 2010. Male, Female. The Evolution of Human Sex Differences, second ed. APA Press, Washington, DC.
Geary, D.C., Byrd-Craven, J., Hoard, M.K., Vigil, J., Numteee, C., 2003. Evolution and development of boys' social behavior. Developmental Review 23, 444–470.
Hall, J.A., 1984. Nonverbal Sex Differences. Johns Hopkins University Press, Baltimore, MD.
Hedges, L.V., 2008. What are effect sizes and why do we need them? Child Development Perspectives 2, 167–171.
van Hemert, D.A., van de Vijver, F.J.R., Vingerhoets, A.J.J.M., 2011. Culture and crying: prevalences and gender differences. Cross-Cultural Research 45, 399–431.
Hyde, J.S., 2005. The gender similarities hypothesis. American Psychologist 60, 581–592.
Hyde, J.S., 2013. Gender similarities and differences. Annual Review of Psychology 65, 373–398. http://dx.doi.org/10.1146/annurev-psych-010213-115057.
Hyde, J.S., Linn, M.C., 1988. Gender differences in verbal ability: a meta-analysis. Psychological Bulletin 104, 53–69.

LaFrance, M., Hecht, M.A., Levy Paluck, E., 2003. The contingent smile: a meta-analysis of sex differences in smiling. Psychological Bulletin 129, 305–334.

Leaper, C., Ayres, M.M., 2007. A meta-analytic review of gender variations in adults' language use: talkativeness, affiliative speech, and assertive speech. Personality and Social Psychology Review 11, 328–363.

Leaper, C., Robnett, R.D., 2011. Women are more likely than men to use tentative language, aren't they? A meta-analysis testing for gender differences and moderators. Psychology of Women Quarterly 35, 129–142.

Leaper, C., Smith, T.E., 2004. A meta-analytic review of gender variations in children's language use: talkativeness, affiliative speech, and assertive speech. Developmental Psychology 40, 993–1027.

Lippa, R.A., 2010. Gender differences in personality and interest: when, where, and why? Social and Personality Psychology Compass 3, 1–13.

Mather, M., Lighthall, N.R., 2012. Risk and reward are processed differently in decisions made under stress. Current Directions in Psychological Science 21, 36–41.

McClure, E.B., 2000. A meta-analytic review of sex differences in facial expression processing and their development in infants, children, and adolescents. Psychological Bulletin 126, 424–453.

Schmitt, D.P., Realo, A., Voracek, M., Allik, J., 2008. Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. Journal of Personality and Social Psychology 94, 168–182.

Vacha-Haase, T., Thompson, B., 2004. How to estimate and interpret various effect sizes. Journal of Counseling Psychology 51, 473–481.

Vaillancourt, T., 2013. Do human females use indirect aggression as an intrasexual competition strategy? Philosophical Transactions of the Royal Society B 368, 20130080.

Voracek, M., Mohr, E., Hagmann, M., 2013. On the importance of tail ratios for psychological science. Psychological Reports: Measures & Statistics 112, 872–886.

Weisberg, Y.J., DeYoung, C.G., Hirsh, J.B., 2011. Gender differences in personality across the ten aspects of the Big Five. Frontiers in Psychology 2, 78.

Gender, Economics of

Francine D Blau, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the contribution of economics to the analysis of gender differences in behavior and economic outcomes, including labor supply and the allocation of time; the gender division of labor in the family; analyses of gender differences in wages and occupations, including the human capital explanation and models of discrimination; as well as considerations of the gender pay gap in the context of the overall structure of wages.

Gender and the Economics of Gender

This article describes the relationship of gender to economics, focusing on economic analyses of gender differences in behavior and economic outcomes. Historically, gender has not been perceived to be a central concept in economic analysis. However, as the force of current events has thrust gender-related issues to the fore, economists have responded. The outcome has been not only a better understanding of the nature and causes of gender differences, but also an enrichment of the discipline itself.

One reason for the historical neglect of gender issues by economists is that economists traditionally focused on market transactions and, traditionally, most married women were primarily engaged in nonmarket activities. As women have come to comprise a larger share of the paid labor force and as market work has loomed larger in the typical woman's life, interest in the determinants of gender differences in behavior and labor market outcomes has grown. At the same time, the growing interest of economists in analyzing nonmarket work and demographic outcomes such as marriage, divorce, and fertility also played an important role in increasing economists' attention to gender-related issues.

Time Allocation in the Family Context

The traditional theory of labor supply is a simple extension of consumer theory. Individuals maximize their utility, which is derived from market goods and leisure, subject to budget and time constraints. The labor supply decision is determined by the opportunity wage (the wage rate the individual could earn in the labor market), nonlabor income (the amount of income available to them from other sources), and the individual's tastes or preferences for work vs. leisure. While this theory is sufficient for many purposes, it has substantial limitations for understanding the determinants of the gender division of labor in the family and the factors influencing women's labor force participation. The key to addressing these issues is a fuller understanding and analysis of the household production process. Prompted in part by their desire to understand the causes of the rising labor force participation of married women in the post–World War II period, economists extended the traditional theory of labor supply to consider household production.

Household Production and the Allocation of Time

The work of Becker and Mincer in the 1960s was particularly influential in bringing the analysis of household production into the mainstream. In analyzing married women's labor force participation in the US, Mincer (1962) pointed to the importance for women of the three-way decision among market work, nonmarket work, and leisure. He attributed the growth in married women's participation during the early postwar period to women's rising real wages which induced them to substitute market work for housework. This response was large enough to more than offset the negative income effect on their labor supply associated with the growth in their husbands' real wages.

Becker (1965) advanced this process further by proposing a general theory of the allocation of time. In this and other work (summarized in Becker, 1981), he laid the foundations for the 'new home economics', including economic analyses of time allocation, marriage, divorce, and fertility. Interestingly, while Mincer opened a window on household production by distinguishing nonmarket work from leisure, Becker was able to provide a further advance by again eliminating the distinction. However, while in the traditional labor supply model all nonmarket time is spent in leisure, in Becker's model all nonmarket time is spent in household production. Becker assumes that households derive utility from 'commodities' that are produced by inputs of market goods and nonmarket time.

Consider the relationship between children and women's labor force participation. Young children may be viewed as a 'time-intensive' commodity. Traditionally, the mother has been the primary care giver. While it is possible to substitute market goods and services (e.g., babysitters, day care centers) for home time in caring for children, these alternative production techniques tend to be costly and may be difficult to implement. Thus, at any point in time, the probability that a woman will participate in the labor force is expected to be inversely related to the number of children present. Over time, the increase in women's labor force participation rates has been associated with decreasing birthrates, as well as increases in the availability of alternative child care arrangements.

The Determinants of Tastes for Market Work

While economists have generally not examined the determinants of tastes for market work, this is an increasing focus of

research. Perhaps not surprisingly, the evidence suggests that family background influences such tastes. For example, mother's labor supply behavior has been found to be positively associated with their daughter's labor force participation decision and also to have a positive association with the labor force participation of their son's wife (Farré and Vella, 2013; Fernández et al., 2004).

Culture also plays a role. For example, it has been found that women migrating to the US from countries with relatively high female labor force participation rates have persistently higher labor supply in the US than women coming from countries with lower female participation rates (Blau et al., 2011). Similarly, the labor supply behavior of the US-born daughters of immigrants (the second generation) has been found to be positively associated with both female participation rates in their parents' country of origin (Fernández and Fogli, 2009) and the participation patterns in the US of immigrants from those origin countries in the parents' generation (Blau et al., 2013).

Economic Development and Trends in Labor Supply

Labor force participation is important for women because it is associated with female empowerment in the family and in the broader society. The relationship between women's labor force participation and economic development has been found to follow a U-shaped pattern (Goldin, 1990, 1995). Female labor force participation is high at earlier stages of economic development, when women tend to be heavily involved as family workers (often unpaid) on the farm or in family businesses, or otherwise working for pay or producing for the market within the household. Their labor force participation initially falls in the course of economic development, as the locus of production moves out of the household and family enterprises and into factories and offices. Goldin argues there is a stigma against wives doing wage work in manufacturing and agriculture that contributes to reducing women's participation at this stage. As women's education levels rise and their job opportunities become more attractive, the stigma against married women working is diminished and female labor force participation increases.

Goldin (1990) shows that, in the US, these shifts in participation patterns were accompanied by changes in the elasticity (responsiveness) of women to their wages and nonlabor incomes (generally husbands' income). Around the turn of the twentieth century (toward the bottom of the U), married women's wage elasticity (responsiveness to their own wages) was very small, but the elasticity associated with their husbands' income was negative and very large in absolute value. That is, women worked primarily if their husband's income was inadequate. As development proceeded (along the rising portion of the U), women's labor supply decisions became more responsive to their own wages and labor market opportunities and less responsive to (determined by) their husband's income. Goldin (1990) predicted that, as divorce rates rose and women's jobs increasingly became careers as opposed to merely a means to earn income, not only should wives' responsiveness to husbands' income continue to decline, but also should their responsive to their own wages. That is, as women became more firmly attached to the labor force, like men, they would be less likely to change their labor supply in response to changes in their wages or nonlabor (husbands') income. Consistent with this, Blau and Kahn (2007) found that, in the US, women's wage and income elasticities were each cut roughly in half between 1980 and 2000 (although they still remain higher than the corresponding male elasticities). Another indicator that women have become more firmly attached to the labor force is that most studies for the US find that the negative effect of children on women's labor force participation has diminished over time (e.g., Macunovich, 2010).

Although female labor force participation has risen in most countries, there are, nonetheless, substantial differences even among the economically advanced countries. These differences are due to a variety of factors ranging from religion and culture to economic organization to government policies. Some evidence suggesting a potentially important role for government policy is suggested by international comparisons of the impact of family-friendly policies; this research also shed light on recent developments in the US.

After a half century of strong growth after World War II, US female labor force participation rates began to stagnate in the mid-1990s. In 1990, the United States had one of the higher female participation rates among economically advanced countries – it ranked sixth out of a set of 22 OECD countries in that year. By 2010, however, as other nations caught up to and surpassed the US, the US ranked 17th (Blau and Kahn, 2013). One major difference between the US and other OECD countries is that most offer more generous parental leave programs (both in length of the leave and benefits received) and many require employers to provide greater part-time options. Blau and Kahn (2013) find that the greater expansion of these family-friendly policies in other OECD countries accounts for a substantial portion (nearly 30%) of the decrease in US women's labor force participation rates relative to these other countries between 1990 and 2010. These results suggest a positive employment effect for such policies and also that work–family issues are a factor in the US stagnation in participation rates.

While more expansive family-friendly policies may promote women's labor force participation, they come at the risk of channeling women into 'mommy track' jobs that tend to be associated with fewer opportunities for advancement and lower pay. Blau and Kahn (2013) report that US women were more likely than women in these other countries to have full-time jobs and to work as managers or professionals. These results are consistent with the negative wage effects of longer leave entitlements (over 26 weeks) reported by Ruhm (1998). Thus there may be a trade-off between some policies that make it easier for women to combine work and family and women's advancement at work.

The Gender Division of Labor

In the discussion of the allocation of time earlier, it was simply assumed that women tend to bear the primary responsibility for childcare and housework. However, the gender division of labor in the family is also an issue that the new home economics addresses. According to Becker (1981), the division of labor will be dictated by comparative advantage. To the

extent that women have a comparative advantage in household and men in market production, it will be efficient for women to specialize at least to some extent in the former while men specialize in the latter. In Becker's view, the increased output resulting from this arrangement maximizes family utility and constitutes one of the primary economic benefits to marriage. Thus, women's rising labor force participation is expected to reduce the gains from marriage, thereby contributing to higher divorce and lower marriage rates. However, offsetting this expectation, at least to some extent is that, as Stevenson and Wolfers (2007) and others have pointed out, with the growth of the two-earner family, marriages may be increasingly based on complementarity between the spouses (enjoyment of time spent together and joint consumption), rather than on substitution between them (a gender-based division of labor). Moreover, the benefits to complementarity may be increasing as employed spouses share more common interests.

Economists have also turned their attention to better understanding family decision making. Led by Manser and Brown (1980) and McElroy and Horney (1981), there has been a growing application of bargaining models to family decision making, These models assume that the preferences of husbands and wives may diverge and that outcomes for each spouse can be modeled using game theory. In the most popular application of these models to the family, the cooperative bargaining model of marriage, the utility received by each spouse in the Nash bargaining solution is positively related to the utility of each partner at the 'threat point', most commonly taken to be the utility that each would receive if the marriage irretrievably broke down. This suggests that a woman who works outside the home will have a higher utility at the threat point and hence a more favorable distribution within marriage. Studies reviewed by Lundberg and Pollak (1996) provide empirical support for this model in their finding that earned or unearned income received by the wife has a different effect on demand patterns than income received by the husband. Moreover, it has generally been found that when women have greater control over household resources, spending patterns are changed in ways that benefit children. In developing countries, where resources are especially scarce, such a reallocation can be quite important for child health and well-being.

Gender Differences in Labor Market Outcomes

Women earn less than men, on average, in all advanced economies, though the extent of the gender pay gap varies considerably across countries and has been declining in many of them in recent years (e.g., Blau et al., 2014). There is also considerable employment segregation of men and women – by occupation, industry, and firm. Segregation of men and women by occupation, for which summary data are most readily available, has been found to be prevalent in all economically advanced countries, though it too has been declining in recent decades in most OECD countries, including the US (e.g., Blau et al., 2014).

There are two broad approaches to analyzing gender differences in labor market outcomes. First is the human capital explanation which lays primary emphasis on men's and women's own choices in explaining occupation and pay differences. Second are a variety of models of labor market discrimination which place the onus for unequal outcomes on differential treatment of equally qualified men and women in the labor market. Some advances have also been made by considering the gender pay gap and other demographic differentials in the context of the overall structure of wages.

The Human Capital Explanation and Other Supply-Side Sources

The human capital explanation for gender differences in occupations and earnings, developed by Mincer and Polachek (1974) and others, provides an explanation for gender differences in market outcomes on the basis of the traditional division of labor by gender in the family. Anticipating shorter and more discontinuous work lives as a consequence of their family responsibilities, women have lower incentives to invest in on-the-job training than men. Their resulting smaller human capital investments will lower their relative earnings. Women are also expected to choose occupations for which human capital investments are less important and in which the skill depreciation that occurs during time spent out of the labor force is minimized. They will avoid especially jobs requiring large investments in firm-specific skills (i.e., skills which are unique to a particular enterprise), because the returns to such investments are reaped only as long as one remains with a particular employer. At the same time, employers are expected to be reluctant to hire women for such jobs because they bear some of the costs of firm-specific training (see the discussion of statistical discrimination below).

There also appears to be a motherhood wage penalty in many advanced countries, that is, lower earnings for women who are mothers, all else equal (Sigle-Rushton and Waldfogel, 2007). This likely reflects the loss of their tie to the firm and associated firm-specific skills when women withdraw from the labor force upon the birth of a child. Some evidence (e.g., Waldfogel, 1998) suggests that provision of parental leaves may reduce the size of the family penalty by maintaining the tie to the firm, although, as our previous discussion suggests, extended leaves may have a negative effect on women's wages, so the length of the leave is likely important.

The human capital model suggests not only that women should be less likely than men to invest in on-the-job training but also that they will be reluctant to make large investments in formal education, again because they do not expect to be in the labor force long enough to reap sufficient returns on their investment. However, although, traditionally, men were more likely to go to college than women, in the US and other economically advanced countries, women have not only caught up to men in college graduation rates but have now surpassed them.

Women's increasing labor force attachment is undoubtedly one factor contributing to their increased incentives to invest in higher education. However, that does not explain why they have surpassed men in most countries. One reason for this appears to be that girls and young women have lower non-pecuniary costs of investing in education than boys and young men – that is, investing in education appears to entail more disutility for males due to their slower social development and more serious behavioral problems (Goldin et al., 2006).

In recent years, labor economists have become increasingly interested in the effect of noncognitive or 'soft' skills – including psychological attributes, preferences, and personality – on labor market outcomes and behavior (Heckman and Kautz, 2012). With respect to gender, intriguing findings suggest a number of psychological attributes that differ between women and men. For example, women have been found to be less willing than men to negotiate and compete and to be more risk averse (for reviews, see Bertrand, 2010; Croson and Gneezy, 2009). Such differences in attitudes could help to explain the observed gender differences in pay and occupations. This line of work is relatively new in economics and, while of considerable interest, the sources of the observed gender differences (e.g., nature vs. nurture), as well as their consequences for labor market outcomes are not fully known, particularly since much of the evidence on gender differences in psychological attributes has been gleaned from laboratory experiments. Moreover, it is important to bear in mind that gender differences in all psychological attributes do not necessarily favor men. As we have seen, the greater behavioral problems of boys and young men appear to contribute to male's lower rate of college going. It will be interesting to learn more about the contribution of psychological factors to gender gaps as we accumulate more evidence, including from field experiments.

Models of Labor Market Discrimination

To the extent that gender differences in outcomes are not fully accounted for by productivity differences due to gender differences in human capital and other supply-side sources, models of labor market discrimination offer an explanation. Theoretical work in this area was initiated by Becker's (1957) model of racial discrimination. Becker conceptualized discrimination as a taste and analyzed three cases, those in which the discriminatory tastes were held by employers, co-workers, and customers or clients. Under certain circumstances, such discrimination will cause a wage differential between men and women. Discriminatory employers will only hire women at a sufficient wage discount that compensates them for the disutility of employing women. Discriminatory male workers will demand a wage premium to work with women thus raising men's relative wages, and the reluctance of discriminatory customers or clients to buy goods or services provided by women will make women less productive in terms of revenue brought in, thus depressing women's relative wages.

Models based on tastes for discrimination are also consistent with employment segregation, but do not necessarily predict it will occur. If wages are flexible, it is altogether possible that discrimination will result in lower pay for women, but produce little or no segregation. However, if discriminatory tastes against women in traditionally male pursuits are both strong and prevalent, women may tend to be excluded from these areas. If such segregation does occur, it may or may not be associated with gender pay differentials. In the presence of sufficient employment opportunities in the female sector, equally qualified women may earn no less than men. The relationship between occupational segregation and earnings differentials is clarified in Bergmann's (1974) overcrowding model. If potentially equally qualified men and women are segregated by occupation, the wages in male and female jobs will be determined separately by the supply and demand for labor in each sector. Workers in male jobs will enjoy a relative wage advantage if the supply of labor is more abundant relative to demand for female than for male occupations.

One question that has been raised about the Becker model is how employer discrimination can persist in the long run in the face of competitive forces. The least discriminatory firms, which hire more lower-priced female labor, would have lower costs of production. In the long run, they should expand and drive the more discriminatory firms out of business. One answer, suggested initially by Becker himself, is that discrimination will be located in sectors of the economy that are not competitive. While Becker emphasized monopolistic elements in the product market, a related approach targets monopsonistic power on the part of the employer in the labor market (e.g., Madden, 1973; Black, 1995).

In addition, models of statistical discrimination were later developed, in part to explain the persistence of discrimination in the long run in the face of competitive forces. Such models assume a world of uncertainty and imperfect information; thus differences between groups in the expected value of productivity or in the reliability with which productivity may be predicted are expected to result in differences in the treatment of members of each group. So, for example, employers may believe that women are more likely to quit their jobs than men with similar observed characteristics (education, experience, etc.) or they may believe that women's turnover can be less reliably predicted by their measured characteristics than men's. As a consequence, they may pay women less, exclude them from jobs requiring substantial firm-specific training, or deny them promotions.

It has been argued that such statistical discrimination (making decisions on the basis of the average characteristics of the group) is consistent with profit maximization and can thus persist in the face of competitive forces. However, Aigner and Cain (1977) contend that such models are no more convincing in explaining the persistence of discrimination than models based on tastes. To the extent that employers' views are correct, the lower expected productivity of women will reduce their wages but women as a group will be paid their expected productivity. This does not constitute labor market discrimination as economists define it, that is, pay differences that are not accounted for by productivity differences. Moreover, they note that when employer beliefs regarding average differences are erroneous, discrimination clearly exists but in their view discrimination based on such misperceptions is even less likely to persist in the long run than discrimination based on tastes. Finally, though they acknowledge that less reliable predictions of a group's productivity combined with risk aversion by employers could produce a discriminatory differential, a perfectly elastic supply of risk neutral entrepreneurs would be expected to erode discriminatory differentials based on this factor.

However, such models may have greater usefulness in explaining the persistence of gender differentials in pay than Aigner and Cain's analysis suggest. Consider first the situation

where employer perceptions are correct, is it appropriate to consider this a form of 'discrimination' in any sense? From a normative perspective, the answer may be yes, to the extent that basing employment decisions on a characteristic like sex – a characteristic that the individual cannot change – could be viewed as inequitable. Indeed, the practice of judging an individual on the basis of group characteristics rather than upon his or her own merits seems the very essence of stereotyping or discrimination. Such behavior is certainly not legal, for example, under antidiscrimination laws and regulations.

Now consider the situation where employer perceptions are incorrect. If statistical discrimination is accompanied by feedback effects, this may be a credible source of persistent discriminatory pay differences. For example, if employers incorrectly expect that women are more likely to quit their jobs, they may respond by giving women less firm-specific training or assigning them to dead-end jobs. Faced with few incentives to remain on the job, women may respond by exhibiting exactly the higher turnover that employers expect.

This criticism of the Becker model – its inability to explain the persistence of discrimination against women in the long run – has proved to be a double-edged sword in that it has led some economists to doubt that labor market discrimination is responsible, in whole or part, for observed gender differences in labor market outcomes. Yet it is important to recognize that discrimination is an intrinsically complex process. Thus, it is not surprising that no simple solution has been found to the question of why discrimination may persist in the face of competitive forces. Similarly, the various models of discrimination, each emphasizing different motivations and different sources of this behavior, need not be viewed as alternatives. Rather, each may serve to illuminate different aspects of this complex reality. Moreover, as gender discrimination has become less socially acceptable, it has likely become less overt and more subtle. Indeed, there is growing evidence from social psychologists that discriminatory attitudes and stereotyping may even be unconscious (e.g., Fiske, 1998; Bertrand et al., 2005). Finally, as our discussion earlier suggests, discrimination can adversely affect women's human capital investments and labor force attachment by lowering the market rewards to this behavior (e.g., Weiss and Gronau, 1981).

Evidence on Human Capital, Discrimination, and the Gender Pay Gap

The typical approach to analyzing the sources of the gender pay gap is to estimate wage regressions specifying the relationship between wages and productivity-related characteristics for men and women. The gender pay gap may then be decomposed statistically into two components: the first, due to gender differences in measured characteristics and the second, which is 'unexplained' and potentially due to discrimination.

Such studies provide considerable empirical support for both gender differences in human capital and other measured characteristics and labor market discrimination in explaining the gender pay gap (for a review, see Hersch, 2006). For example, Blau and Kahn (2006) report that, in 1998, for a sample of full-time workers, US women's wages were 80% of men's, on average. Adjusting for gender differences in human capital (i.e., education and experience) raises the ratio to 81%, while additionally adjusting for gender differences in industry, occupation, and unionism further increases it to 91%. Thus, while measured characteristics, particularly industry and occupation, are important, women still earn considerably less than similar men, even when all measured characteristics are taken into account. (Interestingly, these findings imply that human capital factors taken together do not explain much of the gender difference in pay. This is because although women had less labor market experience than men, on average, which does help to account for the gender difference, they had *higher* levels of education, which works in the opposite direction.)

One problem with studies of this type is that discrimination may be overestimated to the extent that unexplained pay differentials between men and women are due in whole or part to gender differences in unmeasured qualifications. On the other hand, to the extent that unmeasured factors favor women or that the regression models include controls for characteristics like industry or occupation that could themselves be due to discrimination, discrimination may be underestimated.

Studies that use an experimental design provide additional evidence of discrimination that addresses concerns regarding unmeasured characteristics. For example, Neumark et al. (1996) analyzed the results of a hiring 'audit' in which male and female pseudo-job seekers were given similar résumés and sent to apply for waiter or waitress jobs at the same set of Philadelphia restaurants. The results provided statistically significant evidence of discrimination against women in high-priced restaurants where earnings of workers are generally higher. A second study, Goldin and Rouse (2000), examined the impact of the adoption of 'blind' auditions by symphony orchestras in which a screen is used to conceal the identity of the candidate. Using data from actual auditions, they found that the screen substantially increased the probability that a woman would advance out of preliminary rounds and be the winner in the final round. Finally, there is evidence that suggests women, but not men, face discrimination based on their parental status. Based on both laboratory and field experiments, Correll et al. (2007) found that study participants had less favorable views regarding the résumés of equally qualified mothers relative to those of nonmothers, while fathers were not disadvantaged relative to nonfathers.

The Role of Wage Structure

Work by Juhn et al. (1991) on trends in the race pay gap over time in the US has highlighted an additional factor, overall wage structure, which is also relevant to analyzing the gender pay gap. Wage structure is the array of prices determined for labor market skills and the rewards to employment in particular sectors. Since men and women tend to have different levels of labor market qualifications (especially work experience) and to be employed in different occupations and industries, the returns to skills and the size of premiums for employment in predominantly male sectors may impact the gender pay gap. All else equal, the larger the returns to skills like experience and the larger the rewards to employment in predominantly male

occupations and industries, the larger will be the gender pay gap.

Blau and Kahn (1996) have applied this insight to understanding international differences in the gender pay gap in order to address a paradox. The relative qualifications of American women are high compared to women in other countries, and the US has had a longer and often stronger commitment to antidiscrimination laws than most economically advanced nations. Yet the US has traditionally been among the countries with the largest gender pay gaps. The resolution to this paradox lies in the decentralized system of wage setting in the US that has resulted in a highly dispersed wage distribution and high returns to labor market skills and employment in particular sectors. In contrast, the greater role of unions and government in the labor markets of other economically advanced nations has resulted in higher wage floors and more compressed wage distributions from which employed women, as a relatively low-wage group, disproportionately benefit.

This raises the question of how, given the rising wage inequality that has occurred in many of the advanced industrialized countries, women have nonetheless been able to narrow the wage gap with men in many cases. Results for the US from Blau and Kahn (2006) suggest that women have been able to compensate for the adverse effects of rising inequality and higher rewards to skills by improving their relative qualifications, particularly their average work experience and their occupational distribution. In addition, the unexplained portion of the pay gap declined, reflecting an upgrading of women's unmeasured labor market skills, a decline in discrimination against women, and/or broader economy-wide shifts in the demand for labor favoring the types of occupations and industries in which women tend to be employed (for more on demand, see also, e.g., Welch, 2000; Bacolod and Blum, 2010).

Of course, international differences in the gender wage gap and changes over the time in the gap are affected by a variety of other factors, in addition to wage structure. One that is interesting to consider is selection: which women choose to participate in the labor force. So, for example, in Southern European countries, where female participation rates are quite low, gender wage gaps also tend to be relatively low. Olivetti and Petrongolo (2008) suggest that the reason for these relatively low gender gaps is that in these countries where market work is uncommon for women, it is concentrated among women who can earn especially high wages when they enter the labor market. This lowers the measured gender wage gap. In countries with higher participation rates, on the other hand, more low-wage women are included in the labor force and the measured gender wage gap is larger. Such composition issues have also been examined in the context of explaining the trends in the gender wage gap in the US, with Blau and Kahn (2006) assigning such compositional shifts a relatively modest role and Mulligan and Rubinstein (2008) according them more importance.

Conclusion

This review suggests that women's economic roles in the family and their labor market outcomes are closely related. This has undoubtedly played a role in the persistence of gender differences in both spheres. However, increasing opportunities for women in the labor market may be expected to reduce gender differences in family roles, while such a diminution of adherence to the traditional division of labor in the family is likely to enhance women's position in the labor market further. Thus, just as a fuller understanding of the interrelationships between women's roles in the family and their status in the labor market helps us to understand the persistence of gender inequality in economic outcomes, it also enables us to appreciate how changes in either one of these spheres, or both, can induce a mutually reinforcing process of change. Reductions in the pay gap in many of the economically advanced countries may well signal the operation of such a process.

See also: Behavioral Economics; Cooperative Game Theory; Discrimination, Economics of; Econometric Software; Experimental Economics; Game Theory: Noncooperative Games; Gender and Feminist Studies in Geography; Income Distribution and Inequality: Measurement Issues; Labor Supply; Market Structure and Performance; Migration, Economics of; Wage Differential and Structure.

Bibliography

Aigner, D.J., Cain, G.G., 1977. Statistical theories of discrimination in labor markets. Industrial and Labor Relations Review 30 (2), 175–187.

Bacolod, M., Blum, B.S., 2010. Two sides of the same coin: U.S. 'residual inequality' and the gender gap. Journal of Human Resources 45 (1), 197–242.

Becker, G.S., 1957. The Economics of Discrimination, second ed. University of Chicago Press, Chicago.

Becker, G.S., 1965. A theory of the allocation of time. Economic Journal 75 (299), 493–517.

Becker, G.S., 1981. A Treatise on the Family. Harvard University Press, Cambridge, MA.

Bergmann, B.R., 1974. Occupational segregation, wages and profits when employers discriminate by race or sex. Eastern Economic Journal 1 (2), 103–110.

Bertrand, M., 2010. New perspectives on gender. In: Ashenfelter, O.C., Card, D. (Eds.), Handbook of Labor Economics, vol. 4b. Elsevier, Ltd., Oxford, pp. 1545–1592.

Bertrand, M., Chugh, D., Mullainathan, S., 2005. Implicit discrimination. American Economic Review 95 (2), 94–98.

Black, D.A., 1995. Discrimination in an equilibrium search model. Journal of Labor Economics 13 (2), 309–334.

Blau, F.D., Ferber, M.A., Winkler, A.E., 2014. The Economics of Women, Men, and Work, seventh ed. Pearson/Prentice-Hall, New Jersey.

Blau, F.D., Kahn, L.M., 1996. Wage structure and gender earnings differentials: an international comparison. Economica 63 (250), S29–S62.

Blau, F.D., Kahn, L.M., 2006. The US gender pay gap in the 1990s: slowing convergence. Industrial and Labor Relations Review 60 (1), 45–66.

Blau, F.D., Kahn, L.M., 2007. Changes in the labor supply behavior of married women: 1980–2000. Journal of Labor Economics 25 (3), 393–438.

Blau, F.D., Kahn, L.M., 2013. Female labor supply: why is the US falling behind? American Economic Review 103 (30), 251–256.

Blau, F.D., Kahn, L.M., Papps, K.L., 2011. Gender, source country characteristics and labor market assimilation among immigrants. Review of Economics and Statistics 93 (1), 43–58.

Blau, F.D., Kahn, L.M., Liu, A.Y., Papps, K.L., 2013. The transmission of women's fertility, human capital, and work orientation across immigrant generations. Journal of Population Economics 26 (2), 405–435.

Borghans, L., ter Weel, B., Weinberg, B., 2006. People People: Social Capital and the Labor-Market Outcomes of Underrepresented Groups. NBER Working Paper No. 11985.

Correll, S.J., Benard, S., Paik, I., 2007. Getting a job: is there a motherhood penalty? American Journal of Sociology 112 (5), 1297–1338.

Croson, R., Gneezy, U., 2009. Gender differences in preferences. Journal of Economic Literature 47 (2), 448–474.
Farré, L., Vella, F., 2013. The intergenerational transmission of gender role attitudes and its implications for female labor force participation. Economica 80 (318), 219–247.
Fernández, R., Fogli, A., 2009. Culture: an empirical investigation of beliefs, work, and fertility. American Economic Journal: Macroeconomics 1 (1), 146–177.
Fernández, R., Fogli, A., Olivetti, C., 2004. Mothers and sons: preference formation and female labor force dynamics. Quarterly Journal of Economics 119 (4), 1249–1299.
Fiske, S.T., 1998. Stereotyping, prejudice, and discrimination. In: Gilbert, D.T., Fiske, S.T., Lindzey, G. (Eds.), The Handbook of Social Psychology, fourth ed., vol. 2. Oxford University Press, New York, pp. 357–411.
Goldin, C., 1990. Understanding the Gender Gap. Oxford University Press, New York.
Goldin, C., 1995. The u-shaped female labor force function in economic development and economic history. In: Schultz, T.P. (Ed.), Investment in Women's Human Capital. University of Chicago Press, Chicago, pp. 61–90.
Goldin, C., Katz, L.F., Kuziemko, I., 2006. The homecoming of American college women: the reversal of the college gender gap. Journal of Economic Perspectives 20 (4), 133–156.
Goldin, C., Rouse, C., 2000. Orchestrating impartiality: the impact of 'blind' auditions on female musicians. American Economic Review 90 (4), 715–741.
Heckman, J.J., Kautz, T., 2012. Hard evidence on soft skills, Adam Smith Lecture. Labour Economics 19, 451–464.
Hersch, J., 2006. Sex discrimination in the labor market. Foundations and Trends in Microeconomics 2 (4), 281–361.
Juhn, C.H., Murphy, K.M., Pierce, B., 1991. Accounting for the slow-down in black-white wage convergence. In: Kosters, M. (Ed.), Workers and Their Wages. AEI Press, Washington, DC, pp. 107–143.
Lundberg, S., Pollak, R.A., 1996. Bargaining and distribution and marriage. Journal of Economic Perspectives 10 (4), 139–158.
Macunovich, D.J., 2010. Reversals in the patterns of women's labor supply in the United States, 1977–2009. Monthly Labor Review 133 (11), 16–36.
Madden, J.F., 1973. The Economics of Sex Discrimination. Heath and Co., Lexington, MA.
Manser, M., Brown, M., 1980. Marriage and household decision-making. International Economic Review 21 (1), 31–44.
McElroy, M.B., Horney, J.M., 1981. Nash-bargained household decisions: toward a generalization of the theory of demand. International Economic Review 22 (2), 333–349.
Mincer, J., 1962. Labor force participation of married women. In: Lewis, H.G. (Ed.), Aspects of Labor Economics. National Bureau of Economic Research. Princeton University Press, New Jersey.
Mincer, J., Polachek, S., 1974. Family investments in human capital: earnings of women. Journal of Political Economy 82 (2), S76–S108.
Mulligan, C.B., Rubinstein, Y., 2008. Selection, investment, and women's relative wages over time. Quarterly Review of Economics 123 (3), 1061–1110.
Neumark, D., Bank, R.J., Van Nort, K.O., 1996. Sex discrimination in restaurant hiring: an audit study. Quarterly Journal of Economics 111 (3), 915–941.
Olivetti, C., Petrongolo, B., 2008. Unequal pay or unequal employment? A cross-country analysis of gender gaps. Journal of Labor Economics 26 (4), 621–654.
Ruhm, C.J., 1998. The economic consequences of parental leave mandates: lessons from Europe. Quarterly Journal of Economics 113 (1), 285–317.
Sigle-Rushton, W., Waldfogel, J., 2007. Motherhood and women's earnings in Anglo-American, continental European, and Nordic countries. Feminist Economics 13 (2), 55–91.
Stevenson, B., Wolfers, J., 2007. Marriage and divorce: changes and their driving forces. Journal of Economic Perspectives 21 (2), 27–52.
Waldfogel, J., 1998. Understanding the 'family gap' in pay for women with children. Journal of Economic Perspectives 12 (1), 137–157.
Weiss, Y., Gronau, R., 1981. Expected interruptions in labor force participation and sex-related differences in earnings growth. Review of Economic Studies 48 (4), 607–619.
Welch, F., 2000. Growth in women's relative wages and in inequality among men: one phenomenon or two? American Economic Review 90 (2), 444–449.

Gender, Feminism, and Sexuality in Archaeological Studies

Ericka Engelstad, Institute of Archaeology and Social Anthropology, University of Tromsø, Tromsø, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Gender archaeology combines critical studies of the ways in which archaeological knowledge of gender, sex, and sexuality is produced, presented, and represented. Traditional archaeology was concerned with artifacts, cultures, societies, and people in both prehistory and history. An increasing interest in people in the past revealed that archaeologists were mostly concerned with men – man the hunter and warrior, and little concerned with women and gender. The 1980s saw the beginning of gender archaeology in Norway and the United States. Since then, there has been an increasing interest in gender archaeology and numerous studies of how gender varied over time as seen in studies of social relations in a wide range of ancient societies, variation in making and using tools, technological innovations, settlements, and ways of living.

Gender archaeology combines critical studies of the ways in which archaeological knowledge of gender, sex, and sexuality is produced, presented, and represented. Feminist/gender studies in archaeology started with critiques of the overdetermination of women's lives by nature or biology both in the past and present, thereby reducing multifaceted lives to an essentialist category 'woman.' Rather gender, sex, and sexuality are not constrained by essentialist and narrowly defined categories, but are historically constituted, varied, and changing over time and space. The aims of feminist archaeology are to investigate and challenge the assumed content of, and relationships between, gendered categories such as women and men or femininity and masculinity. Feminist archaeology also seeks to study gender divisions and sexuality over time and in different social and cultural contexts in the past and in contemporary archaeology. Thus feminist/gender archaeology challenges traditional constructions of archaeological knowledge.

In the Beginning

Although tied to the goals of Second Wave feminism, which critiqued gender relations in contemporary society, feminism was slow in coming to archaeology. In the mid-1970s the first feminist critiques of equity issues in archaeology started in Scandinavia, particularly Norway. This concern quickly expanded to critiques of traditional analyses and presentations of women in the past and resulted in the founding of the first, and as yet only, feminist archaeology journal, 'KAN: kvinner i arkeologi i Norge' ('KAN: Women in Archaeology in Norway') in 1985. That same year Margaret Conkey and Janet Spector (1984) published their groundbreaking article critiquing androcentrism in archaeology and discussing the possibilities of studying gender in the past in the United States. It was not until the early 1990s, however, that the archaeology of gender and feminist critiques exploded on the archaeological scene. Since the first critical articles, there have been numerous publications on the archaeology of gender in journals, as well as monographs from symposia and conferences. Few of these publications use the word 'feminist' in their titles, preferring various combinations of the words 'gender' and 'archaeology.' However, the majority are often feminist in content and intention, and in their interpretations of the past.

Gender/feminist archaeology began largely as an Anglo-American, Australian, and Scandinavian phenomenon, but quickly expanded to other European countries and Latin America. The slow development of gender and feminist archaeology in some countries can be partially explained by how archaeology is understood within different academic systems. In North America archaeology and anthropology are closely linked, and archaeology benefited from its close connection with feminist anthropology. In European countries, however, archaeology is a humanity and thus society and social relations were traditionally not in the forefront of many studies. Archaeology is one of the few humanities that works with material culture and was relatively isolated from the intense feminist discussions in the humanities. This basic disciplinary difference between North American and European archaeology of gender is probably a reason for the different perspectives these two archaeologies initially had. North American gender archaeology has been concerned more with social processes while European gender archaeology continued with traditional interests in cultural (now gender) identity and distinctions. Feminist and gender theories were part of broader theoretical shifts in archaeology and anthropology that drew on social theories like poststructuralism and postprocessualism.

The feminist philosopher of science Alison Wylie (1991) asked why it took so long for feminist critiques to come to archaeology and emphasized the essentialist assumptions in archaeological theorizing inherent in macroscale analyses. Such essentialist assumptions were based on biological determinism and it was vital to make a distinction between sex, a biological construct, and gender, a social construct. Here one must note the difference between the English-language separation of sex and gender, and the fact that many other languages have but one word for both. Recent work in feminist archaeology, both in English and in non-English speaking countries, resurrects sex and sexuality from the merely biological, but while viewing sex/sexuality as a social construction associated with gender, this work does not neglect the relevance of biological processes for understanding identities.

Back to Basics

Much feminist work in the 1980s and 1990s concentrated on gender rather than sex and sexuality. This was, perhaps, necessary in order to negate the importance of previous biological determinism and essentialism. Androcentric archaeology considered women essentially the lesser part of the male–female, masculine–feminine, man–woman dichotomies. Women were often reduced to their biological function of giving birth. This also placed women on the lesser side of a whole series of related hierarchical dichotomies, such as public–domestic, public–private, dominant–dominated, active–passive, etc., thus giving biological reasons for social differences and social formations. This pattern was established in the archaeological literature before feminist critiques turned this set of universal hierarchical dichotomies into a problem to be investigated rather than assumed.

Conkey and Spector (1984) noted that despite its professed objectivity and neutrality, archaeology was 'neither objective nor inclusive on the subject of gender.' Traditional gender paradigms implicitly treated women and men differently, consistently linked certain activities to each gender/sex, and made men central and active in all societies in the past, while women were passive and 'at home.' Feminist interpretations of the same past critiqued the subordinate role given to women in archaeological analyses, texts, exhibits, and illustrations. Feminist archaeologists worked from the premise that universal subordination was not the case in the past, that gender ideologies and women's roles varied and were given different statuses over time and space, again emphasizing that these concepts are something to be investigated rather than assumed. Both feminist theoretical and empirical work assumed that gender categories were not exclusive and that women were an active part of culture and society in the past. Studies were concerned with defining women's various roles, contributions, and activities. The concept of task differentiation was developed by Janet Spector as a framework and methodology to study the various dimensions of women's and men's activities in society.

Following feminist critiques of science, feminist archaeologists critically examined the basic assumptions behind previous understandings of the past. Studies addressed archaeological understandings of what counts as data, what counts as interpretation of that data, and what is considered acceptable interpretation and theory in the archaeological community. Thus, they examined the selective processes through which scholars determined relevant data for interpretation and how these processes overdetermined interpretations and results by ignoring gendered processes and structures. Why are 'men's' artifacts simply assumed to be of primary interest? Why did important interpretational frameworks consider only men to be the prime movers of social and cultural change. In addition, they examined androcentric theoretical concepts that were previously, and often currently, considered to be the only reasonably legitimate, 'natural,' and 'rational' theoretical frameworks. Engendering the past not only redefines and extends what is considered as data, interpretation, and theory, but also challenges many taken-for-granted concepts and models.

Approaches to Gender and Sexuality

The archaeology of gender is grounded in the analysis and understanding of women's and men's lives as these are reflected in archaeological remains of the past and especially in the material culture of past societies. Richly detailed empirical studies, which put women, men, and children back in to our understanding of the past, abound. These studies are tied to differing theoretical views of gender as a structuring principle and how material culture is used to express and construct gender roles, differences, ideologies, and gendered identities in past societies. Foregrounding gender effects changes in traditional archaeological thinking and leads to an interest in social relations between genders and between gendered individuals of differing age, class, ethnicity, and race.

Engendered studies of Stone Age archaeology, for instance, have gone further than a concern with ecology and technology by considering the wider perspective of social relations between individuals and groups. Doing so leads to a restructuring of traditional archaeological interpretations, offering new insights into social relations, ideologies, and sociocultural institutions. Margaret Conkey (1991) considers the wider contexts in which tools were made, used, and reused at a Magdalenian site where several families lived ('aggregation sites'). By viewing tools not as single objects, but as objects representing a complex set of technological, productive, and engendered social relations, she is able to show the changing contexts of action and contexts for power in which people and objects were involved. Gender is always in process, the sexual/gender division of labor is often changed and negotiated, and these changing contexts of action are associated with the renegotiation of power relations and structures. Such analyses challenge traditional views of stable and exclusive gender categories and lead to analyses of archaeological material that can reveal different models of gender in society that may change in relation to particular contexts.

European archaeologists have made considerable use of burial material in discussions of the construction of gender. In her study of the Viking period Oseberg burial, a rich grave with the burial of two women in a ship, Elisabeth Arwill-Nordbladh (1998) uses Conkey's 'context of action' concept in relation to textile production, farm economy, communication, and ritual performance. Artifacts in the grave were analyzed in order to see how gender was constructed through various contexts of action. Gender construction is not seen as following a set of strict gender categories, but was maintained, changed, negotiated, or resisted through practice or performative acts. In this manner gender is seen as difference between and within gender categories, and the complexity of gender dynamics comes to the forefront. Working with grave material from the North European Bronze Age, Marie Louise Stig Sørensen (1991) presents a series of studies on how objects are active in expressing gendered identity. She especially examines how gender is constructed through appearance by examining the composition and placement of grave goods in relation to the body and in relation to their wider sociocultural contexts of use and meaning. In this manner it is possible to analyze not only gender but also crosscutting social categories such as age, ethnicity, and status.

Initially, the importance of gender in archaeology led to the neglect of the complexities of sex and sexuality. Concentrating

on gender lead to an initial neglect of sex and sexuality, but as feminist theory developed, so did interest in sexuality and the body as socially constructed and historically constituted, thus becoming important themes in feminist archaeology. Much of this work is on representations of the body in art, particularly in classical Mediterranean cultures and in studies of prehistoric rock art. Scandinavian rock art has a variety of human figures representing both female and, in particular, male figures. Inspired by feminist theory and Judith Butler's work on gender and the body, archaeologists critiqued traditional two-sex or no sex understandings of rock art figures using diverse analogies tied to variation in representations of bodies, ethnicity, mythologies, and rituals (Engelstad, 2001).

In the 1990s it became clear that gender archaeology was too concerned with 'finding women' and that the importance of men and masculinities, queer theories, and alternative interpretations and variation in gender entered the scene. Men and masculinity have been taken for granted categories that were assumed to be stable over time and space. Such male stereotypes as 'Man the Hunter' and 'Man the Warrior' are presented in the archaeological literature and are unchanging over diverse archaeological periods and cultures. A. Bernard Knapp was one of the first archaeologists to take up the neglected 'masculinist' approaches to archaeology, which he saw as 'derivative of feminist scholarship' (Knapp, 1998). He also saw Skogstrand (2010) ask the question "Is androcentric archaeology really about men?" since it actually recreates the same hunter/warrior stereotypes over time and space. She contends that this is a problem of objectivity and lack of critical reflection that has only recreated the same male stereotypes in archaeological analyses, and uses feminist critiques of androcentrism to decenter masculine stereotypes in historic and prehistoric archaeology. It is interesting that critique of these male stereotypes is now emerging in the context of feminist and gender archaeology.

There are various threads in feminist and gender archaeology, perhaps the most important of which is the distinction between the 'archaeology of gender' and the 'gendered archaeology' (Roberts, 1993), the latter of which has been discussed here as feminist archaeology. The archaeology of gender, while often premised on a feminist perspective and concerned with putting women and children back into our understanding of the past, reiterates traditional archaeological concepts and principles of interpretation. A gendered archaeology is explicitly feminist and closely tied to feminist critiques of science in an effort to generate new engendered ways of thinking, doing, and presenting archaeology. A gendered archaeology/feminist archaeology does not simply add women's contribution to past societies, but is concerned with interrogating and reworking our unquestioned starting premises and taken for granted concepts.

Practicing Archaeology

Although often considered to be neutral and objective, archaeological practice is highly gendered though often implicitly so. Feminist interests include archaeological practice, particularly fieldwork and analysis, workplace issues, pedagogy, cultural heritage management including museums, new ways of writing and presenting archaeology, and concern for the local communities in which archaeologists work.

Joan Gero (1996) looked at how fieldwork, particularly excavations and artifact analysis, is gendered. Gero has shown that observation and documentation are not neutral, objective processes but always contain an element of interpretation. Her work on excavations in Argentina shows that there are gender differences among who does different kinds of archaeological work and how documentation can be tied to differences in gender, class, and age. In North America and Europe men often lead large fieldwork projects, while women analyze extant collections. In Australia men dominate the universities while women dominate in work with cultural heritage management. In other countries gendered specialties and research foci take other forms; and although there were some prominent women archaeologists in the past, they had little access to status and recognition in the profession. Thus, the status of women and equity issues in the archaeological workplace continue to be major feminist concerns.

While an important aspect of feminist/gender archaeology is a concern with institutional structure and workplace dynamics that hinder women from fully participating in the discipline, feminist archaeologists are concerned with the communities within which they work. Another concern is with archaeology's and archaeologists' relationship to the indigenous communities within which they work. In Australia, Laurajane Smith has written about the dominance of women in work with cultural heritage, or cultural resource management related to indigenous communities. Feminists have been particularly open to the importance and meaning of local indigenous communities' heritage and its place in their past and future.

From Margin to Center

In a review article, Conkey and Gero (1997) noted the lack of explicit feminist theorizing and critique of basic archaeological concepts in much of gender archaeology. Detailed empirical analyses were perhaps necessary in the beginnings of gender archaeology. With the beginning of the second millennium it is clear that diverse theoretical perspectives are becoming common, although there is still a tendency for some implicit essentialism. A lack of theory only marginalizes the archaeology of gender, not only in relation to mainstream archaeology but also in relation to other feminist disciplines. Future work should strengthen efforts to include gender/feminism/sexuality in our construction of knowledge about the past.

Work grounded in feminist critiques, principles, and perspectives not only have made women visible but also shown the diversity of their lives and relations to others. Gendered and feminist archaeology has developed and changed considerably during its short existence. Feminist archaeology is exceptionally dynamic, drawing on new theoretical insights from other feminist disciplines, and expanding the scope and direction of analyses of gendered relations and ideologies in the past. Both gender archaeology and feminist archaeology will need to grapple with the fundamentals of, and variation in, archaeological material across time and space. Archaeological

interpretation and theorizing is now increasingly incorporating feminist theoretical insights on gender and sexuality.

Despite the seemingly initial small impact of many aspects of gender, feminist, masculinist, and queer archaeology (Danielsson, 2012; Danielsson and Thedéen, 2012), there are numerous important articles and books that are concerned with these themes in the archaeological literature. Gender/feminist archaeology is an important, complex, dynamic, and constantly expanding area of interest and is well on its way to realizing its critical and reconstructive potential.

See also: Archaeology, Theory in; Europe, Archaeology of; Historical Archaeology; Household Archaeology; Hunter-Gatherer Societies, Archaeology of; Marxist Archaeology; Prehistoric Art; Ritual and Symbolism, Archaeology of; States and Civilizations, Archaeology of; Trade and Exchange, Archaeology of.

Bibliography

Arwill-Nordbladh, E., 1998. Genuskonstruksjoner i Nordisk vikingtid, förr og nu. GOTARC, Gothenburg Archaeological Theses, Series B No 9 (Gender Constructions in the Nordic Viking Period, Before and Now).

Conkey, M., 1991. Contexts of action, contexts for power: material culture and gender in the Magdalenian. In: Gero, J., Conkey, M. (Eds.), Engendering Archaeology: Women and Prehistory. Blackwell, Oxford, pp. 57–92.

Conkey, M., Gero, J., 1997. Programme to practice: gender and feminism in archaeology. Annual Review of Anthropology 411–437.

Conkey, M., Spector, J., 1984. Archaeology and the study of gender. In: Advances in Archaeological Theory and Method, vol. 7, pp. 1–38.

Danielsson, I.-M.B., Thedéen, S. (Eds.), 2012. To Tender Gender. The Pasts and Futures of Gender Research in Archaeology. Stockholm Studies in Archaeology, vol. 58.

Danielsson, I.-M.B., 2012. Much ado about nothing? Gender research in journals during the last 30 years within archaeology. In: To Tender Gender. The Pasts and Futures of Gender Research in Archaeology. Stockholm Studies in Archaeology, vol. 58, pp. 17–32.

Diaz-Andreu, M., Sørensen, M.L.S., 1998. Excavating Women: A History of Women in European Archaeology. Routledge, London.

Engelstad, E., 2001. Desire and body maps: all the women are pregnant, all the men are virile, but. In: Helskog, K. (Ed.), Theoretical Perspectives in Rock Art Research, The Institute for Comparative Research in Human Culture. Serie B, vol. CVI. Skrifter, Oslo.

Engelstad, E., 2007. Much more than gender. Journal of Archaeological Method and Theory 14 (3), 217–234.

Gero, J., 1996. The social world of prehistoric facts: gender and power in Paleoindian research. In: Wright, R. (Ed.), Gender and Archaeology. University of Pennsylvania Press, Philadelphia.

Gero, J., Conkey, M., 1991. Engendering Archaeology: Women and Prehistory. Blackwell, Oxford.

Gilchrist, R., 1999. Gender and Archaeology. Routledge, London.

Hollimon, S.E., 2011. Sex and gender in bioarchaeological research: theory, method, and interpretation. In: Agarwal, S.C., Glencross, B.A. (Eds.), Social Bioarchaeology, pp. 149–182, 1993.

Hager, L. (Ed.), 1997. Women in Human Evolution. Routledge, New York.

KAN: Kvinner i arkeologi i Norge. 1985–2002.

Knapp, B., 1998. Who's come a long way, baby? Masculinist approaches to a gendered archaeology. Archaeological Dialogues 2, 91–104.

Marshall, Y., 2008. Archaeological possibilities for feminist theories of transition and transformation. Feminist Theory 9 (1), 25–45.

Moore, J., Scott, E., 1997. Invisible People and Processes: Writing Gender and Childhood into European Archaeology. Leicester University press, London.

Roberts, C., 1993. A critical approach to gender as a category of analysis in archaeology. In: du Cros, H., Smith, L. (Eds.), Women in Archaeology: A Feminist Critique. Australian National University, Department of Prehistory, Occasional Papers in Prehistory No 23, Canberra, pp. 16–23.

Skogstrand, L., 2010. Is androcentric archaeology really about men? Archaelogies: Journal of the World Archaeological Congress, 56–68.

Spector, Janet, 1993. What This Awl Means: Feminist Archaeology at a Wahpeton Dakota Village. Minnesota Historical Society Press, St. Paul, Minnesota.

Sørensen, M.L.S., 1991. Gender archaeology and Scandinavian Bronze age studies. Norwegian Archaeological Review 25 (1), 31–49.

Wylie, Alison, 1991. Gender theory and the archaeological record: why is there no archaeology of gender? In: Gero, J.M., Conkey, M.W. (Eds.), Engendering Archaeology: Women and Prehistory. Blackwell, Oxford, pp. 31–54.

Relevant Website

http://www.archaeology-gender-europe.org/ – AGE Archaeology and Gender in Europe.

Gender History

Kathleen Canning, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Gender history emerged from women's history and represented a rejection of biological notions of sexual difference in favor of analysis of the cultural and social relations of the sexes. Its rise in Anglo-Saxon scholarship coincided with and was a vital part of the 'linguistic turn' during the late 1980s and early 1990s. Gender history has since widened its scope to encompass the study of men and masculinity. Gender history is also closely related to history of the body, while the history of sexuality/homosexuality is by now a distinct field of scholarly inquiry that is related to the history of gender but not dependent upon it.

Gender history examines the social and cultural, as opposed to natural or biological, relations of the sexes. The English term 'gender' is usually distinct from sex, which is understood as rooted in biological or bodily differences. The term gender does not exist in some languages, while in others its meaning diverges significantly from the English usage. The German term *Geschlecht*, for example, encompasses both 'sex' and 'gender' and thus blurs the core distinctions of the English term. In French the term 'sex' or *sexuel* is more commonly used than *genre*, which can refer to either grammatical or literary genre or serve as a classifying category in natural history (Riot-Sarcey, 1999). The historical category 'gender' emerged from feminist study of women's history during the mid 1980s. When the word 'gender' first gained currency among historians of women in the early 1980s, it was defined as signifying the social relations between the sexes. As the range and scope of scholarly research on gender expanded across the social, scientific, and humanistic disciplines, gender was increasingly understood as a symbolic system or signifier of relations of power in which men and women are positioned differently. By the end of the 1980s, the concept of gender came to encompass the languages, symbols, and social practices that defined sexual differences and sexual inequalities, signaling a more definitive departure of feminist social scientists from foundationalist notions of biological sex (Haraway, 1991).

The Analytic Origins of Gender History

By the turn to the 1990s, gender history – both in terms of its object of analysis and methodological approach – had distinguished itself from women's history, as well, although most practicioners continued to recognize women's history as a formative foundation for the study of gender in history. From the early 1960s on, feminist historians had set out to recover women's participation in and exclusion from processes of social transformations and political change. Empirical investigation of the private (familial, sexual) and public (political) inequality of women in past societies drove the first decade of women's history. The fact that women's history took shape during its formative years in interdisciplinary arenas, such as Women's Studies programs, explains why women's history became one of the most theoretically engaged fields of historical study during its early years, as feminists debated and critiqued the keywords of social analysis.

In this first phase of women's history, sex and class often figured as parallel forms of oppression: the female *sex* was viewed as a subordinate *class*, subjugated by a dominant *class* of men. In feminist scholarship the notion of *patriarchy* signaled a shifting understanding of sexual exploitation as the primary form of women's oppression (see, e.g., Rubin, 1975; Eisenstein, 1978). Taking patriarchy as a hierarchical sexual order, women's historians located its origins in 'the private family' and sought to examine the reproduction of patriarchy in social modes of production, divisions of labor, and property relations (Kelly, 1976). Analysis of the origins and practices of patriarchal structures and ideologies, along with the recovery of lost stories of female actors and women's agency, significantly expanded the scope of historical knowledge about women, family, and labor, while elevating 'sex' to a keyword of social analysis.

Yet in furnishing new insights into the place of women in past societies, the pioneering work in women's history soon revealed its own limitations. Unitary analytical categories, such as 'woman'/'women' and 'man'/'men,' based on fixed, rather than historically or socially variable, notions of biological sex, fostered narratives of women's experience or oppression that ignored differences of race, class, ethnicity, and sexual preference. Moreover, a fuller understanding of 'the significance of the *sexes*' in the historical past would require the study not only of 'the subjugated sex,' but of both women and men (Davis, 1976). At the end of the 1970s, feminist historian Joan Gadol called for a 'doubled vision' of society, one that emphasized the ways in which the identities of both men and women were shaped by sex and class (Kelly, 1979). In emphasizing the inextricable links between sex and class, women's history also indisputably widened the scope of the political to include family and household, bodies and sexualities – all once considered as belonging to the sphere of the 'private.' By the end of the 1970s 'sex' had emerged as a crucial category of feminist social analysis, one that was parsed in relation to race and class in scholarly debates.

Joan Kelly's formulation of this doubled vision of society marks a critical opening for the concept of gender, as feminist scholarship turned its attention to the ways in which ideologies, norms, and symbolic systems shaped sexual identities, relations, and sites of governance. The editors of the pioneering collection, *Sex and Class in Women's History* (Newton et al., 1983), for example, used the term 'gender' in order to understand the systematic ways in which 'sex differences' cut through

society and culture and conferred inequality upon women. Karin Hausen's path-breaking essay on the origins of 'sexual character,' situated the polarization of sexual stereotypes in Enlightenment philosophy, suggesting a temporal framework for the origins of modern gender distinctions (Hausen, 1981; Gleixner and Gray, 2006). Feminist historians delivered powerful critiques of the public/private dichotomy associated with the polarization of the sexes, widening the scope of the political to include family and household, bodies, and sexualities.

Even as gender began to forge its own analytical ground in the mid 1980s, it remained entangled for another decade or more in the analytics of women and sex. While feminist philosophers and theorists grappled with the sex/gender distinction (Nicholson, 1994), the terms 'women' and 'gender' remained loosely interchangeable in feminist history well into the 1990s and are still sometimes also described with the conjoined phrase 'history of women and gender' (Smith, 2000; Canning, 2006; Rose, 2010; Downs, 2010). The work of analytically distinguishing gender from women and from notions of biological sex was also propelled by transformations outside of academia. Through the broader public debates about 'identity politics' in the United States and Britain, once-cohesive ethnic, racial, national, or sexual identities were increasingly understood as multiple, mutable, and contradictory. Emphasizing the inextricability of gender and racial identities, feminist scholars of race issued powerful critiques of the unitary category of women and notions of patriarchal oppression (Higginbotham, 1992). Like gender, race also came to be viewed as a social construct rather than a genotypic or phenotypic form of difference.

Although the shift from women to gender was already well underway by the mid 1980s, the publication of Joan Scott's essay, Gender: A Useful Category of Historical Analysis (Scott, 1986), marked a definitive turn away from studies of patriarchy and of women's economic, sexual, or reproductive subordination grounded in notions of physical differences between the sexes. Scott summoned 'mainstream' historians to consider gender as a constitutive element of social relationships based on the 'perceived differences' between the sexes, as expressed in culturally available symbols, normative concepts, notions of politics and institutions, and subjective identities. Moreover, Scott argued, gender had become an essential category of historical analysis because it is a primary way of signifying relationships of power. Gender thus gained an analytic status of its own, first by severing the links between sex-gender and biology, and then by freeing it from the remnants of Marxism that hinged the production of gender to changing economic structures. Scott's article, followed by her essay collection, *Gender and the Politics of History* (Scott, 1988), unleashed a torrent of debate over the implications of gender for methodological approaches to historical causality and processes of change. Scott called upon historians to shift their attention from 'objective processes,' such as industrialization and urbanization, to the discursive, linguistic, or symbolic representations of sexual difference that, for example, institutionalized the sexual division of labor in late nineteenth century France or that produced women workers as 'objects of (state) investigation and subjects of history' (Scott, 1988, 1993).

Understood as a cultural, linguistic, or discursive construct, gender had a central place in the broader paradigm shift known as 'the linguistic turn' in the discipline of history. The controversies accompanying the rise of gender history thus involved the objects of historical investigation – women of the past versus historical discourses, languages, or symbols – and the methods of historical analysis, which in the wake of the linguistic turn came to include Foucauldian social theory, new historicism, as well as poststructuralist linguistic and literary theory (Poovey, 1988b; Canning, 1994; Surkis, 2012). At the same time feminist philosopher Denise Riley interrogated the category 'woman' in her *Am I That Name?* (1988), noting the inherent and historically grounded instability of the category termed 'woman' and identifying feminism as the precise site at which that instability was systematically contested, taken apart and reconstructed. These dual impulses – the critique and disassembly of the category of 'woman' and the advancement of a new analytic concept of 'gender' – effectively redefined and resituated the keywords of feminist history: woman, sex, and feminism. Soon other keywords of social history, such as 'experience,' 'identity,' and 'agency,' were also subject to critical debate and redefinition.

Yet the rise of gender as a category of historical analysis did not have the effect of asserting the primacy of gender relative to other forms of inequality, such as race, class, or ethnicity. Rather, the emphasis on the inextricable links between gender and other social identities and categories of difference sparked critical engagements with categories of race and class as well. In the United States and Britain, feminist scholars of race intervened at this juncture with powerful critiques of the categories and practices of women's history. Defining race as a 'metalanguage,' and the 'ultimate trope of difference,' African–American historians explicated the powerful effects of race on the construction and representation of gender, class, and sexuality. Understanding race as a 'double-voiced discourse' revealed the ways in which it shaped both the oppression and self-representation of minorities (Higginbotham, 1992).

Rewriting History after the Turn to Gender

The turn to gender, like all theoretical and methodological innovations, was an uneven and protracted process that has differed markedly across different temporal and geographical fields. In some fields scholars sought to redefine methodologies, concepts, and theories to meet the challenge posed by the concept of gender, while in other fields women's history was quietly renamed 'gender history' without a recognizable shift in historical methodology. In national historiographies, in which the work of recovering absent female subjects had a later start, historical research on women has continued to provide the foundation for a more explicitly theoretical engagement with gender well after the turn to the twenty-first century. In still other settings, where the turn to gender history coincided with the rediscovery of masculinity and men, gender history came to denote a more thoroughly relational engagement with the histories of both masculinity and femininity.

While women's history once sought to fit female subjects into existing categories of social analysis, such as class, nation, and citizenship, the turn to gender prompted more fundamental critiques and redefinitions of these keywords of social analysis. The concept of 'class,' for example, was at the heart of debates about both gender and the linguistic turn, leading some feminist scholars to call for an emancipation from class as a 'privileged signifier of social relations and their political representations' (Alexander, 1984), while others analyzed the formation of both the middle-class (Davidoff and Hall, 1987) and the working-class (Clark, 1995) as processes of differentiation in which gender was always centrally implicated. Eschewing models of class formation that sought to delineate economic, social, cultural, and political levels or stages of development, feminist historians of gender examined class as a political language, emphasizing the significance of gender in the process of assigning and contesting the boundaries and meanings of class (Scott, 1988; Sewell, 1990; Canning, 1992). The 'public/private divide' and its specific historical and gendered meanings for class, citizenship, public sphere, and nation-state formation was the topic of two issues of the *Journal of Women's History* (2003/1 and 2003/2) that offered theoretical perspectives and historical comparisons with case studies of Britain, Brazil, the United States and the Middle East. Other social-scientific keywords, like 'public sphere' (Fraser, 1992; Ryan, 1992), 'civil society' (Pateman, 1988; Hull, 1996), and 'citizenship' (Lister, 1997; Canning and Rose, 2002) have also been sites of fruitful rethinking by scholars of gender. Similar to the location of class in socioeconomic structures, citizenship is embedded in legal and constitutional frameworks and theories of state and nation formation. Feminist social scientists critiqued teleological models of citizenship, like T.H. Marshall's classic, *Citizenship and Social Class* (1950) that outlined the progressive acquisition of civil, political and social rights, pointing instead to the nonlinear trajectories by which women acquired rights, often first as social citizens and only much later, as political citizens. Feminist scholarship has probed the meanings and practices of citizenship beyond the law, exploring the rhetorics, claims, and subjectivities of citizenship, which are also taken up by those lacking and thus demanding citizenship rights.

Breaking up the binary oppositions of public and private/political and personal, also led to fruitful new research on the history of consumption from a gendered perspective. Although consumption has historically been identified with femininity, Davidoff and Hall's *Family Fortunes* (1987) argued convincingly that domesticity, the ideology of the British middle classes, was vested in complex and elaborate practices of consumption and display. With gender as a framing concept, the historical study of consumption has expanded in recent years to encompass analysis of commodification, spectatorship, material culture, popular culture, leisure, and fashion, which are vital dimensions of gendered social identities, not least as an arena in which these gendered identities were performed, critiqued, recast, and subverted (DeGrazia and Furlough, 1996; Peiss, 1998). Studies of consumption also explore its place in shaping gendered boundaries of class, citizenship, and cultural belonging (Cohen, 2003) and have also spurred new explorations of visual culture, fashion and beauty, marketing and advertising, both in transnational dimensions (Modern Girl, 2008) and as a vital aspect of the relationships between empires and metropoles (McClintock, 1995).

Masculinities

If the redefinition of keywords of social analysis was one outcome of gender history, its explicit attention to the relations between the sexes prompted the opening of gender history toward a more serious study of men and masculinity. The embrace of gender had been controversial in part because women's historians had feared that emphasis on the mutual constitution of masculinity and femininity would obscure both male oppression of women and female agency (Bock, 1991). The advent of men's studies in the early 1990s – marked, for example, by the founding of the American Men's Studies Association in 1991 and its journal *The Journal of Men's Studies* in 1992 – led feminists to observe that men, unlike women, had hardly been 'hidden from history.' While feminists worried that inquiry into the ideals of manliness and their imposition on men of the past might conceal the ways men had participated in and benefited from the oppression of women (Brod, 1994), historian John Tosh concluded in 1994 that the 'gendered study of men' and masculinities had become 'indispensable to any serious feminist historical project' (Tosh, 1994). As the category of 'masculinity' soon overtook that of 'men,' theoretical approaches to masculinity were refined in the disciplines of sociology and philosophy by Harry Brod (1994), R.W. Connell (1995), and Michael Kimmel (1996). A founding text in the field of history was George Mosse's *The Image of Man: The Creation of Modern Masculinity* (1996), a comparative study of ideal types of masculinity in European and American history since the early nineteenth century. Subsequent histories of masculinity examined ideal or hegemonic masculinities, masculinities as shaped by class (Hall, 1992; Clark, 1995; Frevert, 1995), citizenship (Dudink et al., 2007), race (Bederman, 1995), colonialism (Sinha, 1995), militarism (Kühne, 1996), and homosexuality (Chauncey, 1994; Puff, 1998; Rocke, 1998; McLaren, 1999b). One achievement of the history of masculinity has been to 'demonstrate that gender is inherent in all aspects of social life, whether women are present or not' (Tosh, 1994). While masculinity is generally regarded as an inherently relational category, femininity has scarcely had a parallel significance in the study of gender. Historian Stefan Dudink contends that the focus on masculinity reflects the 'asymmetry of gender relations' by which masculinity has remained unmarked amidst the universal categories of 'men,' 'mankind,' and 'humanity' (Dudink, 1998), while femininity, which was 'never allowed to pass as the universal,' was more visible and thus less in need of a similar kind of unmasking.

Entwined Terms: Genders/Bodies/Sexualities

Body/Bodies

Other key terms in the vocabulary of interdisciplinary gender studies, which were redefined in the course of the paradigm shift

from women to gender are 'body' and 'sexuality.' During the first wave of women's history, the body served as the foundation for the shared experiences and identities of women. As a term deeply embedded in biologism and essentialism, body was inextricably linked to sex before the 'linguistic turn' began to disentangle these terms. The 'discursivation' of the body, which took place in the course of the 'linguistic turn,' detached it from unchanging notions of physical difference, but it also appeared to sweep away any sense of bodily physicality or materiality. No longer the foundation of experience, nor the somatic location of identity, the ambiguous status of the body between discourse and experience prompted a wave of debate through the end of the 1990s (Haraway, 1991; Butler, 1993; Bynum, 1995; Canning, 1999). As these debates raged, the body moved to the center of methodological debates in feminist theory, which established conceptual frameworks for the 'full range of bodies' that had emerged from two decades of scholarship – body image; body performance; bodies in science, medicine, and technology; bodies in space; virtual bodies; lived or situated bodies; the body politic; and embodied ideals/ideologies (Bordo and Jaggar, 1989; Martin, 2001; Schiebinger, 2000).

At the same time 'body history' came to designate a pluralistic field of scholarly inquiry of bodies as significant dimensions of specific social, cultural, and political processes, such as the laboring body, the regulated/disciplined body; the medicalized or scientized body; the idealized body of nation, state, or king; the embodiment of theology and the place of the body in cultural or religious rituals; or body and memory in situated life histories (e.g., Outram, 1989; Roper, 1994). An important outcome of historians' engagement with the legacy of Michel Foucault for the study of bodies as sites of regulation and governmentality is the broad recognition of 'biopolitics' as a crucial form of modern governance. While Rublack (2012) suggests that early modern persons' sense of self – or subjectivity – was experienced as 'deeply embodied' and was both interlinked with the physical and emotional yet also 'interrelated with the external world,' the modern body is defined by the ways it is viewed, measured, mobilized, disciplined, and governed by the ideologies and technologies of states, social reformers, medical experts, nationalists, and industrial paternalists. Moreover, the modern body had individual and collective or social dimensions, reorganizing social and material life around hygienic principles while instilling in individuals the skills of self-care and self-discipline. The collectivized body or social body was cut through with social distinctions of class, ethnicity, and gender (Poovey, 1995).

If research on the hygienic and social body relied upon well-documented practices of states, as well as medical and social reform institutions, the gendered body as site of experience or agency was much more difficult to trace historically. Hence few historical monographs could be considered histories of the body in this sense, with the important exception of Barbara Duden's enduring classic *Woman in the Body: A Doctor's Patients in Eighteenth Century Germany* (1991), which analyzes the testimony patients left behind about their own bodily perceptions, sensations, and agency in an era when women still possessed 'interpretive authority' over their bodies. The transfer of this interpretive authority to male experts is one of the markers of the modern body. More recent studies of empire, colonialism, and global and transatlantic traffic in people and goods have placed bodies at the heart of the colonial encounter (Burton and Ballantyne, 2005; Brown, 2011) and analyzed bodily sensations, sensory experiences, and memories as part of the 'emotional economy of the everyday' in colonialism (Stoler and Strassler, 2000; Stoler, 2002; Hunt, 2008).

While 'body history' as a distinguishable field of historical inquiry has diminished in the last decade, two of its keywords remain salient for current historical inquiry – corporeality and embodiment. Corporeality can be understood as encompassing the social, cultural, and/or political significance of bodily being; encompassing bodily images/symbols, regimes, and strategies of bodily governance; and bodily perceptions, interpretations, and self-representations that are specific to a given temporal and spatial context (Mazzio, 1997). Why and how the corporeal dimensions of state policies or social movements, of class or ethnic conflict become visible, varies according to the specific historical setting. The notion of embodiment refers to the ways subjects and the social collectivity assign meaning to bodies, interpret and act upon them (Grosz, 1994; Gatens, 1996). No longer embedded in a dichotomy between discourse and agency/experience, embodiment does not presume resistance or subversion, encompassing instead the capacity of humans to assign meaning to and act upon their own bodies, often in unexpected ways.

Sexuality

Gender and body remain crucial analytics in the history of sexuality, an expanding and highly productive arena that has recently repositioned both gender and body as concepts and fields of scholarly inquiry. The focus of early histories of sexuality melded sexuality with gender, reproduction, and body, examining topics such as prostitution, contraception, abortion, sexual reform, adultery, masturbation, and venereal disease (e.g., Walkowitz, 1982, 1992; White, 1990; Grossmann, 1995). Studies of lesbianism and male homosexuality had the productive effect of destabilizing gender categories by breaking up the binary oppositions of male/female, masculine/feminine, and by unmasking 'the epistemic regime of presumptive heterosexuality' (D'Emilio, 1983/1998; Chauncey, 1994; Puff, 1998; Butler, 1990; Fausto-Sterling, 2000). With the founding of the *Journal of History of Sexuality* in 1990, the history of sexuality marked out its own terrain, distinct from the history of gender, coinciding with a decisive turn toward studying sexual politics as situated within and vital to broader historical processes of transformation, such as the formation of civil society, the inception of citizenship, or the pursuit of empire (Hall, 1992; Engelstein, 1992; Stoler, 2002; Surkis, 2006). Influenced by the work of Michel Foucault, the history of sexuality also focused on the specific domains within nation-states where sexualities were produced, such as systems of medical knowledge and state regulation that distinguished normal from 'deviant' or 'marginal' sexualities and that dispersed these definitions throughout the institutions of state and social reform (Mort, 1987; Nye, 1999; Eder et al., 1999).

In the last several years, as the influence of the Foucauldian paradigm has loosened, scholars have posed new questions about the periodicity, i.e., the implicit modernity of homosexuality and of sexuality more generally (Puff, 2012; Wiesner-Hanks, 2012). With histories of male and female sexuality as one of the 'prime movers' in the history of sexuality (Puff, 2012), its practitioners have examined the meanings of sex acts (Puff, 2003) and the emergence of homosexual subcultures and movements for emancipation (Chauncey, 1994; Lybeck, 2012). The history of sexuality has also been influenced by the transnational or global turn, exploring more recently the ways in which sexual cultures and identities transcend national boundaries and are shaped by the transnational flows of individuals, ideas, and movements (Aldrich, 2006; Rupp, 2009; Canaday, 2009; Herzog, 2011). Recent literature also reveals a marked shift away from their longstanding focus on such matters as population politics, hygiene, sexual science, and the normative construction of sexual identities toward examination of the formation of the sexual subject or a modern sexual self (Eder, 1999; Oosterhuis, 2000). Matysik (2012) proposes a view of sexuality that is situated in body and mind, while other scholars call for closer examination of intimacy, emotions, and memory in the construction and experiences of sexuality within life-worlds (Dickinson and Wetzell, 2005; Spector et al., 2012). While sexual histories of war, Nazism, and post-war reconstructions remained unexplored or were still highly controversial a decade ago, these topics have formed some of the most fruitful inquiries in the field during the past several years (Heineman, 2002; Herzog, 2004, 2007, 2011; Evans, 2011).

Outlook for the Future

After nearly three decades of scholarship, the history of gender is a well-established field, as evidenced by the number of chairs and teaching positions assigned to this scholarly field and the continued vitality of the field as represented in scholarly journals and academic book publishing. The debates of the past about the prospects of gender history entering the 'mainstream,' i.e., moving from the status of the margins to the center of historical scholarship have dissipated, not because gender history has achieved parity or equal representation therein, but because the very notions of mainstream and margins have themselves changed considerably during the last two decades. Rather, gender history has become a much more pluralistic undertaking, both in terms of historical methodologies and the scope of topics encompassed under the rubric of 'gender.' Gender history is most often a cultural historical inquiry, one that probes the rhetorics, languages, symbols, and discourses that shape femininity and masculinity as ideologies, practices, and performances. Yet historians have renewed interest in recent years in the social dimensions of gender, in the gendered relations and experiences of empire, colonialism, slavery, labor, poverty, immigration, and other dimensions of 'the social.' While the study of masculinity is a well-established inquiry within gender history by now, women – as subjects and objects of historical investigation – remain a vital presence in many studies of gender. In fact, 'gender and women' are still frequently paired in the titles of books, courses, or conference panels today. If gender history had a part in spawning the study of the body and of sexuality, today the history of sexualities is an established field that is in no sense beholden to gender history, not least because queer studies and the history of homosexuality have constituted its driving force during the last two decades (Puff, 2012). The analytic of 'gender' has thus been enriched and enlivened internally, as historical practices attended to the study of both femininity and masculinity and the relationships between them. Gender history has also been repositioned by the terms that surround it – body and sexuality – and by historical study that considers the connections and distinctions between and among them in specific historical settings.

See also: Gender and Education; Gender and Language: Cultural Concerns; Gender and Place; Gender and Women's Studies, Applied Research On; Gender, Feminism, and Sexuality in Archaeological Studies; Gender: Anthropological Aspects.

Bibliography

Aldrich, R., 2006. Gay Life and Culture: A World History. Universe Publishers, New York.

Alexander, S., 1984. Women, class and sexual differences in the 1830s and 1840s: some reflections on the writing of a feminist history. History Workshop 17, 125–214.

Bederman, G., 1995. Manliness and Civilization: A Cultural History of Gender and Race in the United States, 1880–1917. University of Chicago Press, Chicago.

Bland, L., 1999. Sexology in Culture: Labelling Bodies and Desires. University of Chicago Press, Chicago.

Bock, B., 1991. Challenging dichotomies: Perspectives on women's history. In: Offen, K., Pierson, R.R., Rendall, J. (Eds.), Writing Women's History. Indiana University Press, Bloomington, IN.

Bordo, S., Jaggar, A. (Eds.), 1989. Gender/Body/Knowledge: Feminist Reconstructions of Being and Knowing. Rutgers University Press, New Brunswick, NJ.

Brown, K., 2011. Foul Bodies. Cleanliness in Early America. Yale University Press, New Haven.

Butler, J., 1990. Gender Trouble. Feminism and the Subversion of Identity. Routledge, New York/London.

Butler, J., 1993. Bodies that Matter. On the Discursive Limits of 'Sex'. Routledge, New York/London.

Brod, H., 1994. The Making of Masculinities. The New Men's Studies. Allen & Unwin, Boston, MA.

Burton, A., Ballantyne, T., 2005. Bodies in Contact: Rethinking Colonial Encounters in World History. Duke University Press, Durham, NC.

Bynum, C., 1995. Why all the fuss about the body? A medievalist's perspective. Critical Inquiry 22, 1–33.

Canaday, M., 2009. Thinking sex in the transnational turn: an introduction. American Historical Review 114 (5), 1250–1257, 2009.

Canning, K., 2006. Gender History in Practice: Historical Perspectives on Bodies, Class and Citizenship. Cornell University Press, Ithaca.

Canning, K., 1992. Gender and the politics of class formation: rethinking German labor history. American Historical Review 97 (3), 736–768.

Canning, K., 1994. Feminist history after the linguistic turn: historicizing discourse and experience. Signs 19 (2), 368–404.

Canning, K., 1999. Body as method? Reflections on the place of the body in gender history. Gender & History 11 (3), 499–513.

Canning, K., Rose, S., 2002. Gender, Citizenships and Subjectivities. Blackwell, London.

Chauncey, G., 1994. Gay New York. Gender, Urban Culture and the Making of the Gay Male World. Basic Books, New York, 1890–1940.

Clark, A., 1995. The Struggle for the Breeches. Gender and the Making of the British Working Class. University of California Press, Berkeley, CA.
Cohen, L., 2003. A Consumers' Republic: The Politics of Mass Consumption in Postwar America. Vintage, New York.
Connell, R.W., 1995. Masculinities. University of California Press, Berkeley, CA.
Davidoff, L., Hall, C., 1987. Family Fortunes: Men and Women of the English Middle Class, 1780–1950. University of Chicago Press, Chicago.
Davis, N., 1976. Women's history in transition: the European case. Feminist Studies 3 (3/4), 83–103.
DeGrazia, V., Furlough, E. (Eds.), 1996. The Sex of Things: Gender and Consumption in Historical Perspective. University of California Press, Berkeley, CA.
D'Emilio, J., 1983/1998. Sexual Politics-Sexual Communities. The Making of a Homosexual Minority in the United States. University of Chicago Press, Chicago.
Dickinson, E.R., Wetzell, R., 2005. The historiography of sexuality in modern Germany. German History 23 (3), 291–305.
Downs, L.L., 2004/2010. Writing Gender History, second ed. Hodder, London. 2010.
Duden, B., 1991. Woman beneath the Skin: A Doctor's Patients in 18th Century Germany (Dunlap T, Trans.). Harvard University Press, Cambridge, MA.
Dudink, S., 1998. The trouble with men: problems in the history of 'masculinity'. European Journal of Cultural Studies 1, 419–431.
Dudink, S., Hagemann, K., Clark, A. (Eds.), 2007/2012. Representing Masculinity. Male Citizenship in Modern Western Culture. Palgrave Macmillan, New York.
Eder, F., Hall, L., Hekma, G. (Eds.), 1999. Sexual Cultures in Europe, National Histories and vol. II: Sexual Cultures of Europe, vol. I. Manchester University Press, Manchester.
Eder, F.X., Hall, L., Hekma, G. (Eds.), 1999. Sexual Cultures in Europe, vol 2. Themes in Sexuality. Manchester University Press, Manchester.
Eisenstein, Z., 1978. Capitalist Patriarchy and the Case for Socialist Feminism. Monthly Review Press, New York.
Engelstein, L., 1992. The Keys to Happiness: Sex and the Search for Modernity in Fin-de-Siecle Russia. Cornell University Press, Ithaca, NY.
Evans, J., 2011. Life Among the Ruins: Cityscape and Sexuality in Cold War Berlin. Palgrave Macmillan, New York.
Fausto-Sterling, A., 2000. Sexing the Body: Gender Politics and the Construction of Sexuality. Basic Books, New York.
Fenemore, M., 2009. The recent historiography of sexuality in 20th c. Germany. The Historical Journal 52 (3), 763–779.
Foucault, M., 1995. Discipline and Punish. The Birth of the Prison. Vintage Books, New York.
Foucault, M., 2010. The Birth of Biopolitics. Lectures at the Collège de France, 1978–1979. Picador, New York.
Frader, L., Rose, S. (Eds.), 1996. Gender and Class in Modern Europe. Cornell University Press, Ithaca, New York.
Fraser, N., 1992. Rethinking the public sphere: a contribution to the critique of actually existing democracy. In: Calhoun, C. (Ed.), Habermas and the Public Sphere. MIT Press, Cambridge, MA.
Frevert, U., 1995. Men of Honour: A Social and Cultural History of the Duel. Polity Press, London.
Gatens, M., 1996. Imaginary Bodies: Ethics, Power and Corporeality. Routledge, London.
Gleixner, U., Gray, M., 2006. Gender in Transition. Discourse and Practice in German-Speaking Europe 1750-1830. University of Michigan Press, Ann Arbor.
Grossmann, A., 1995. Reforming Sex. The German Movement for Abortion and Birth Control. Oxford University Press, Oxford/New York.
Grosz, E., 1994. Volatile Bodies: Toward a Corporeal Feminism. University of Indiana Press, Bloomington, IN.
Hall, C., 1992. White, Male and Middle Class. Explorations in Feminism and History. Polity Press, London.
Haraway, D., 1991. Simians, Cyborgs, and Women: The Reinvention of Nature. Routledge, New York/London.
Harris, V., 2010. Sex on the margins: new directions in the historiography of sexuality and gender. Historical Journal 53 (4), 1085–1104.
Hausen, K., 1981. Family and role-division. The polarisation of sexual stereotypes in the nineteenth Century – an aspect of the disocciation of work and family life (Cathleen Catt, Trans.). In: Evans, Richard J., Lee, W.R. (Eds.), The German Family: Essays on the Social History of the Family in Nineteenth and Twentieth Century Germany. Croom Helm, London, pp. 51–83.
Heineman, E., 2002. Sexuality and Nazism: the doubly unspeakable? Journal of the History of Sexuality 11 (1–2), 22–66.
Herzog, D. (Ed.), 2004. Sexuality and German Fascism. Berghahn, New York, 2004.
Herzog, D., 2007. Sex after Fascism. Memory and Morality in Twentieth-Century Germany. Princeton University Press, Princeton, NJ.
Herzog, D., 2009. Syncopated sex: transforming European sexual cultures. American Historical Review 114 (5), 1287–1308.
Herzog, D., 2011. Sexuality in Europe. A Twentieth-Century History. Cambridge University Press, Cambridge/New York.
Herzog, D. (Ed.), 2011. Brutality and Desire. War and Sexuality in Europe's Twentieth Century. Palgrave Macmillan, New York.
Higginbotham, E.B., 1992. African-American women's history and the metalanguage of race. Signs 17 (2), 251–274.
Hull, I., 1996. Sexuality, State, and Civil Society in Germany, 1700–1815. Cornell University Press, Ithaca, NY.
Hunt, L., 1998. The challenge of gender. Deconstruction of categories and reconstruction of narratives in gender history. In: Medick, H., Trepp, A.-C. (Eds.), Geschlechtergeschichte und Allgemeine Geschichte, Herausforderung und Perspektiven. Wallstein, Göttingen, Germany, pp. 57–97.
Hunt, N.R., 2008. An acoustic register, tenacious images, and Congolese scenes of rape and repetition. Cultural Anthropology 23 (2), 220–253.
Kelly, J., 1976. The social relation of the sexes: methodological implications of women's history. Signs 1 (4), 809–823.
Kelly, J., 1979. The doubled vision of feminist theory. Feminist Studies 5 (1), 216–227.
Kelly, J., 1984. Women, History, and Theory: The Essays of Joan Kelly. University of Chicago Press, Chicago.
Kimmel, M., 1996. Manhood in America: A Cultural History. Free Press, New York.
Kühne, T. (Ed.), 1996. Männergeschichte. Geschlechtergeschichte. Männlichkeit im Wandel der Moderne. Campus Verlag, Frankfurt a.M.
Lister, R., 1997. Citizenship: Feminist Perspectives. New York University Press, New York.
Lybeck, M., 2012. In: Spector, et al. (Eds.), Writing Love, Feeling Shame: Rethinking Respectability in the Weimar Homosexual Women's Movement, pp. 156–168.
Martin, E., 2001. The Woman in the Body. A Cultural Analysis of Reproduction. Beacon, Boston, MA.
Mazzio, C., 1997. The Body in Parts: Fantasies of Corporeality in Early Modern Europe. Routledge, London/New York.
Matysik, T., 2012. In: Spector, et al. (Eds.), Beyond Freedom: A Return to Subjectivity in the History of Sexuality, pp. 185–201.
McClintock, A., 1995. Imperial Leather. Race, Gender & Sexuality in the Colonial Contest. Routledge, New York.
McLaren, A., 1999a. The Trials of Masculinity: Policing Sexual Boundaries, 1870–1930. University of Chicago Press, Chicago.
McLaren, A., 1999b. Twentieth Century Sexuality: A History. Wiley Blackwell, London.
Modern Girl Around the World Research Group, (Eds.), 2008. The Modern Girl Around the World: Consumption, Modernity, and Globalization. Duke University Press, Durham, NC.
Mort, F., 1987. Dangerous Sexualities: Medico–moral Politics in England Since 1830. Routledge, New York.
Mosse, G., 1996. The Image of Man: The Creation of Modern Masculinity. Oxford University Press, Oxford.
Newton, J., Ryan, M., Walkowitz, J. (Eds.), 1983. Sex and Class in Women's History. Routledge, New York.
Nicholson, L., 1994. Interpreting gender. Signs 20 (1), 79–105.
Nye, R., 1999. Sexuality. Oxford Reader. Oxford University Press, Oxford.
Oosterhuis, H., 2000. Stepchildren of Nature: Krafft-ebing, Psychiatry, and the Making of Sexual Identity. University of Chicago Press, Chicago.
Outram, D., 1989. The Body and the French Revolution. Yale University Press, New Haven, CT.
Pateman, C., 1988. The Sexual Contract. Stanford University Press, Stanford, CA.
Peiss, K., 1998/2011. Hope in a Jar: The Making of America's Beauty Culture, second ed. University of Pennsylvania Press, Philadelphia. 2011.
Poovey, M., 1988a. Uneven Developments. The Ideological Work of Gender in Mid-victorian England. University of Chicago Press, Chicago.
Poovey, M., 1988b. Feminism and deconstruction. Feminist Studies 14 (1), 51–65.
Poovey, M., 1995. Making a Social Body: British Cultural Formation 1830–1864. University of Chicago Press, Chicago.
Puff, H., 1998. Männergeschichten/Frauengeschichten. Über den Nutzen einer Geschichte der Homosexualitäten. In: Medick, H., Trepp, A.-C. (Eds.), Geschlechtergeschichte und Allgemeine Geschichte, Herausforderung und Perspektiven. Wallstein, Göttingen, Germany.
Puff, H., 2003. Sodomy in Reformation Germany and Switzerland, 1400–1600. University of Chicago Press, Chicago.
Puff, H., 2012. After the History of (Male) Homosexuality. In: Spector, S. (Ed.), pp. 17–30.

Riley, D., 1988. Am I that Name? Feminism and the Category of 'Women' in History. University of Minnesota, Minneapolis, MN.
Riot-Sarcey, M., 1999. The difficulties of gender in France: reflections on a concept. Gender & History 11 (3), 489–498.
Rocke, M., 1998. Forbidden Friendships. Homosexuality and Male Culture in Renaissance Florence. Oxford University Press, New York.
Roper, L., 1994. Oedipus and the Devil. Witchcraft, Religion and Sexuality in Early Modern Europe. Routledge, London/New York.
Rose, S., 2010. What Is Gender History? Polity Press, London.
Rublack, U., 2010. Gender in Early Modern German History. Cambridge University Press, Cambridge.
Rublack, U., 2012. In: Spector, et al. (Eds.), Interior States and Sexuality in Early Modern Germany, pp. 43–62.
Rubin, G., 1975. The traffic in women: notes on the 'political economy' of sex. In: Reiter, R. (Ed.), Toward an Anthropology of Women. Monthly Review Press, Routledge, New York.
Rupp, L., 2009. Sapphistries: A Global History of Love between Women. New York University Press, New York.
Ryan, M., 1992. Gender and public access: women's politics in nineteenth-century America. In: Calhoun, C. (Ed.), Habermas and the Public Sphere. MIT Press, Cambridge, MA.
Scott, J., 1986. Gender: a useful category of historical analysis. American Historical Review 91 (5), 1053–1075.
Scott, J., 1988. Gender and the Politics of History. Columbia University Press, New York.
Scott, J., 1993. The woman worker. In: Fraisse, Genevieve, Perrot, Michelle (Eds.), A History of Women: Emerging Feminism from Revolution to World War. Harvard University Press, Cambridge, MA, pp. 399–426.
Scott, J. (Ed.), 1996. Feminism and History. Oxford Readings in Feminism. Oxford University Press, Oxford/New York.
Schiebinger, L., 2000. Feminism and the Body. Oxford Readings in Feminism. Oxford University Press, Oxford/New York.
Schiebinger, L., 2004. Nature's Body. Gender in the Making of Modern Science. Rutgers University Press, New Brunswick, NJ.
Sewell Jr., W.H., 1990. How classes are made: critical reflections on E P Thompson's theory of working-class formation. In: Kaye, H., McClelland, K. (Eds.), E. P. Thompson: Critical Perspectives. Temple University Press, Philadelphia, PA.
Sinha, M., 1995. Colonial Masculinity. The 'Manly Englishman' and the 'Effeminate Bengali' in the Late Nineteenth Century. Manchester University Press, Manchester.
Smith, Bonnie, 2000. The Gender of History. Men, Women and Historical Practice. Harvard University Press, Cambridge, MA.
Spector, S., Puff, H., Herzog, D. (Eds.), 2012. After the History of Sexuality. German Genealogies with and beyond Foucault. Berghahn Books, New York.
Stoler, A., 2002. Carnal Knowledge and Imperial Power: Race and the Intimate in Colonial Rule. University of California Press, Berkeley.
Stoler, A., Strassler, K., 2000. Castings for the colonial: memory work in 'new order' Java. Comparative Study of Society and History 42 (1), 4–48.
Surkis, J., 2006. Sexing the Citizen: Masculinity and Morality in France, 1870–1920. Cornell University Press, Ithaca.
Surkis, J., 2012. When was the linguistic turn? A Genealogy. American Historical Review 117 (3), 700–722.
Tosh, J., 1994. What should historians do with masculinity? Reflections on nineteenth century Britain. History Workshop 38, 179–202.
Tosh, J., 2007. A Man's Place: Masculinity and the Middle-class Home in Victorian England. Yale University Press, New Haven, CT.
Walkowitz, J., 1982. Prostitution and Victorian Society: Women, Class and the State. Cambridge University Press, Cambridge/New York.
Walkowitz, J., 1992. City of Dreadful Delight. Narratives of Sexual Danger in Late Victorian London. University of Chicago Press, Chicago.
White, L., 1990. The Comforts of Home: Prostitution in Colonial Nairobi. University of Chicago Press, Chicago.
Wiesner-Hanks, M., 2012. In: Spector, et al. (Eds.), Sexual Identity and Other Aspects of 'Modern' Sexuality: New Chronologies, Same Old Problem?, pp. 31–42.
Wiesner-Hanks, M., 2011. Gender in History. Global Perspectives. John Wiley, Oxford.

Gender, Time-Use, and Fertility Recovery in Industrialized Countries

Javier García-Manglano, Nuffield College, University of Oxford, Oxford, UK
Natalia Nollenberger and Almudena Sevilla, School of Business & Management, Queen Mary University of London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores gendered patterns of time use as an explanatory factor behind fertility trends in the developed world. We review the theoretical foundations for this link, and assess the existing evidence suggesting that a more equal division of labor within the home leads to more children, both at the household (micro) and country (macro) levels. After decades of unprecedented fertility decline in the industrialized world, only a handful of countries in the West exhibit replacement fertility rates, around two children per woman. Paradoxically, birth rates are substantially lower in countries in which family units (and women within families) remain primarily responsible for familial care obligations, and where the role of the state in the provision of care services is marginal. Very low-fertility poses challenges for economic growth and threatens the sustainability of pay-as-you-go welfare systems. It is thus important to understand why some developed countries managed to sustain near-replacement fertility rates or to recover from very low fertility, while others fell and still remain at lowest-low rates of about 1.3 children per woman. Looking at time-use and fertility trends for a few representative industrialized countries, we hint at the existence of a threshold ratio of gender equity in the distribution of domestic work that low-fertility countries need to cross before being able to enter a phase of fertility recovery.

Introduction

During the 1980s, total fertility rates (TFR) in most advanced societies fell below 2.1 children per woman (Van De Kaa, 1987), a level known as *replacement fertility* – two children to replace their parents, and the remaining fraction to compensate for those who will remain childless or die before the end of their reproductive lives. Within the next decade, an important proportion of Southern and Eastern European countries – e.g., Russia – reached rates below 1.3 children per woman (Billari and Kohler, 2004; Kohler et al., 2002), a level known as *lowest-low fertility*. Very low-fertility poses challenges for academics and policy makers alike, as shrinking and aging populations jeopardize economic growth and threaten the sustainability of pay-as-you-go welfare systems. In the absence of offsetting changes in employment rates and productivity, structural adaptations may be required, with important implications for public pensions, health expenditures, and welfare (Weil, 1997).

Recently, scholars have documented a process of fertility recovery in several countries, particularly in the US and in Northern Europe (Goldstein et al., 2009; Sánchez-Barricarte and Fernández-Carro, 2007). This uptick in fertility might currently be spreading to other countries (Myrskylä et al., 2013; Sobotka, 2011) and possibly to more advanced regions within lowest-low fertility countries such as Italy (Caltabiano et al., 2009). Even if these trends might partially be attributed to methodological differences in the measurement of fertility (Bongaarts and Sobotka, 2012), the fact remains that recent fertility declines were not homogeneous across the West, stabilizing near replacement in some countries and falling to substantially lower levels in others. Paradoxically, fertility sank deepest in countries with a strong *familialistic* social policy (e.g., Southern Europe), where the role of the State in the provision of care services is marginal, and family units – and women within families – remain primarily responsible for familial care obligations (Esping-Andersen, 2009).

This article contributes to the existing fertility research by taking a multidisciplinary approach from demography, sociology, and economics to identify an emerging area of research at the intersection between fertility and the gendered distribution of household labor. We use time diary data to present new cross-country evidence on the association between time use, gender, and fertility. This approach can provide new insights as we seek to understand lowest-low fertility and the forces behind fertility recovery.

The article is organized as follows. We begin by documenting the unprecedented fertility decline observed in industrialized countries after 1970, and highlight some of the sociological and economic explanations that have been advanced to explain these trends. Next, we review the theoretical link between fertility and gendered patterns of time use, and assess the existing evidence indicating that a more equal gender division of labor within the home leads to more children. We further investigate this hypothesis by comparing, for a sample of representative industrialized countries, trends in fertility and in the gender distribution of domestic work during the past 50 years. We close with the identification of avenues for future research.

Fertility Decline, Ideational Change, and Opportunity Cost

The theory of the second demographic transition (SDT) (Lesthaeghe and Surkyn, 1988; Lesthaeghe, 1995, 1983) attributes recent fertility declines to a process of ideational social change, by which fertility is no longer driven by concerns regarding basic human needs (food production in the home economy, old-age support and insurance, child survival, etc.), but geared instead toward the satisfaction of higher-order, post-materialistic needs (self-realization, expressive values, etc.) which do not require high fertility to be met. The SDT predicts structural below-replacement

fertility in advanced populations (Lesthaeghe, 2007). Some authors even suggested that most countries would be trapped into very low-fertility rates due to a combination of demographic (fewer women being born resulting in fewer future births) and sociological (declining fertility preferences derived from increased exposure to very low fertility) factors, among others (Testa et al., 2006).

The economic perspective relates fertility decline with increases in the cost of children, in particular the opportunity cost of time. Becker's original model of time allocation suggested that, as time became more valuable, the production and consumption of time-intensive 'commodities,' such as children, would be lower (Becker, 1965). Thus, increasing labor market opportunities for women would lead to higher female labor force participation and to fewer children (Becker, 1965; Willis, 1973). The increase in female wages over the past few decades is assumed to be unrelated to fertility declines; instead, female wage gains are attributed to factors such as technological change, which decreased the reliance on physical-intensive labor in favor of cognitive-intensive work, a transition which, arguably, suited better the labor market abilities of women (Galor and Weil, 1996). In addition, the diffusion of oral contraception contributed to marriage postponement and facilitated greater career investments from women (Goldin and Katz, 2002). Complementarily, the diffusion of home appliances played an important role in easing the demands of housework, increasing the responsiveness of female labor supply to changes in wages (Greenwood et al., 2005). Becker's prediction of a negative relationship between (female) wages and fertility was initially confirmed at the macro (Mincer, 1985) and micro levels (see Schultz, 2001, for a review of this literature). As we will see in the following sections, this relationship would be challenged later using more recent data.

The Heterogeneous Fertility Decline in Advanced Economies

Fertility declines were not homogeneous across countries, resulting in varying birth rates among industrialized nations. As a result, today some countries exhibit TFRs near two children per woman, having managed never to drop far from replacement rates or to recover fertility. In contrast, fertility in some other countries remains substantially below replacement. This is illustrated in **Figure 1**, which presents fertility trends in select OECD economies from 1970 to 2011. We distinguish between two groups of countries: those where fertility did not decline as much or recovered (United States, Denmark, Finland, and Norway) and those in which fertility decreased substantially over this period and remained at very low levels. The latter – Italy, Spain, Germany, and Japan – present TFRs approximately 1.3 children per woman since the early 1990s, a level known as lowest-low fertility.

The experience of fertility recovery in some countries challenged SDT's predictions of long-term low fertility and the hypothesis of a low-fertility trap, by showing that countries can approach near-replacement fertility from below, that is, after falling to lower levels. Indeed, SDT theory has been recently updated to account for rebounding fertility rates (Lesthaeghe, 2010). Similarly weakened is Becker's prediction of an inverse relationship between female wages and fertility, which no longer holds at the macro level. In the past, countries with lower female labor force participation rates had higher-fertility levels, and vice versa (Mincer, 1985), suggesting role incompatibility between the roles of mother and worker (Stycos and Weller, 1967). However, this relationship reversed in some nations after the mid-1980s (Brewster and Rindfuss, 2000; Murphy, 2001) and, today, fertility is lowest in countries with lower female labor force participation rates (Ahn and Mira, 2002; De Laat and Sevilla-Sanz, 2011).

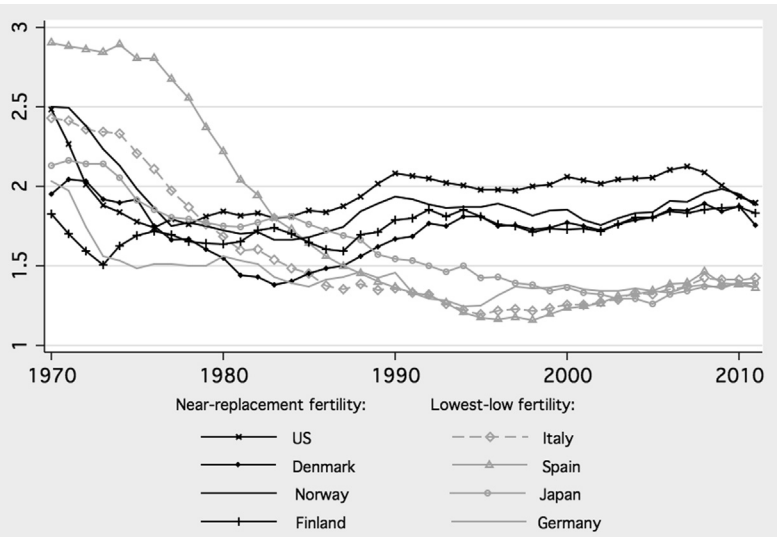

Figure 1 Total fertility rates in selected industrialized countries. Note: Total fertility rate (TFR) indicates the number of children that an average woman would have at the end of her reproductive life were she to follow the age-specific fertility patterns observed in a given population at a given time. Source: OECD, 2012. OECD Family Database, OECD, Paris. www.oecd.org/social/family/database.

Gendered Patterns of Time Use and Fertility Recovery

Two main developments are behind the fertility trends described above. First, male and female employment rates have been converging in OECD countries since the 1970s (Jaumotte, 2003). Second, men are increasingly involved in housework and childcare, especially in high-fertility countries (Aguiar and Hurst, 2007; Bianchi, 2000; Gimenez-Nadal and Sevilla, 2012; Kan et al., 2011). From these observations, a theory has emerged in parallel within sociology, demography, and economics, focusing on gender equity in the distribution of family responsibilities. This theory predicts that couples in which the man and the woman share equitably in home and market production will exhibit higher-fertility outcomes than couples with a more gendered time use pattern. Gender equity in domestic work allows women to combine the rearing of children with market work, a preference shared by most women in advanced economies (Hakim, 2000). Higher-fertility levels can thus be compatible with greater female labor force participation, as long as men take up a fair share of household responsibilities: see De Laat and Sevilla-Sanz (2011) for formalization and application of this theory, and Bettio and Villa (1998) for a similar argument.

This theory could also help explain the process of fertility recovery observed in some countries, and the lack of recovery observed in others. Fertility recovery would thus depend on changes in the household division of labor, driven by gender norms. In particular, when women's wages are low, as they were before 1970 in most developed countries, women specialize in childcare and housework, having children comes with a low opportunity cost, and couples have more of them. In this context, social norms and economic considerations align as observed in the neoclassical home economics approach (Becker, 1991). Esping-Andersen (2009) calls this the traditional 'male breadwinner, female homemaker' equilibrium. However, as women's wages rise, as they did after 1970, women's bargaining power increases and efficiency considerations (i.e., staying at home becoming relatively more costly) lead to less household gender specialization. A 'gender egalitarian' equilibrium is eventually reached, characterized by greater gender equity in the distribution of family duties, high female labor force participation, and fertility rates near replacement (De Laat and Sevilla-Sanz, 2006; Esping-Andersen et al., 2013; Feyrer et al., 2008; Goldscheider, 2012). However, there is an intermediate phase in which women's new labor market opportunities conflict with slowly evolving social norms, expectations that women are chiefly responsible for home production. In this middle stage, the high opportunity cost of time derived from increased female wages places additional stress on women, who face a 'double burden' of market work and childcare (Hochschild and Machung, 1989; Milkie et al., 2009).

Using this theoretical approach, current cross-country differences in fertility rates could be explained, *ceteris paribus*, by the different gendered time-use patterns. Why have some countries completed the transition to a more gender equitable use of time faster than others? Borrowing from McDonald (2000a,b), some authors argue that family policy is the main force behind these trends, since it promotes gender equity in settings, such as the tax code and paternity leave policy, that deal with people as members of families (Esping-Andersen, 2009). Other scholars stress the importance of social norms, particularly those dealing with men's role in the family, and argue that the gender revolution will complete only when norms are more gender neutral (Goldscheider et al., 2013). In any case, both approaches hinge on the growing body of evidence indicating that when men and women share more equitably in the worlds of work and home, couples reach higher levels of fertility (Esping-Andersen, 2009; Goldscheider, 2012).

As with aggregate fertility rates, similar patterns are observed at the household level: greater intra-household gender equity is associated with faster transitions to parenthood and to higher-order births, and with higher completed fertility. This general pattern is confirmed in a variety of countries, using several estimation techniques (Brodmann et al., 2007; Cooke, 2009, 2004; Goldscheider et al., 2013; Oláh, 2003). There is evidence for a U-shaped curve at the household level: couples with a traditional distribution of household responsibilities (male breadwinner, female homemaker), as well as those with more equitable gender roles, are more likely to transition to a second birth than couples with unequal but not specialized division of labor (Torr and Short, 2004).

Evidence from Time Diary Data

Studies of fertility and the distribution of household responsibilities usually approximate time use patterns using gender attitudes and other proxies (such as father's use of parental leave). In contrast, time diary data provide a more accurate and direct picture of the gender division of domestic labor. A diary is completed by respondents on selected days, which are divided into intervals. For each of these intervals, respondents report the main activity they are engaged in, among other things (such as whether they were performing that activity alone or with others). In the remainder of this article, we use time diary data from the past five decades to explore fertility trends in eight industrialized countries, and present some plausible hypotheses that could inform further research.

Table 1 shows average fertility rates and gender patterns of time use for the eight countries in **Figure 1**, for the early 2000s. Gendered patterns of time use are defined as the men-to-women ratio of time spent in a given activity. We distinguish between housework, childcare, and paid work. Countries are ranked by fertility rates, with those at the top (USA, Norway, Finland, and Denmark) exhibiting a more gender egalitarian distribution of time. Men in these four near-replacement fertility countries devote over 50% of woman's time to housework and only between 50% and 85% more time than women to paid work. In contrast, in lowest-low fertility countries (Germany, Japan, Italy, and Spain) men's share in housework time is less than half of women's, with over twice as much time devoted to paid employment. There are only slight differences across groups in time devoted to childcare (these are uncommonly large

Table 1 Fertility and gender patterns of time use, 2000–04

		Ratio of men/women time devoted to:		
	TFR	*Childcare*	*Housework*	*Paid work*
Near-replacement fertility				
USA	2.04	0.52	0.54	1.84
Norway	1.80	0.45	0.67	1.81
Denmark	1.76	0.51	0.73	1.54
Finland	1.75	0.42	0.59	1.52
Lowest-low fertility				
Germany	1.35	0.43	0.48	2.48
Japan	1.32	0.17	0.06	2.11
Italy	1.28	0.41	0.22	2.67
Spain	1.27	0.42	0.27	2.62

Note: Multinational Time Use Study (MTUS) is an ex-post harmonized cross-time, cross-national comparative time use database, constructed from national random-sampled time-diary studies. The sample is restricted to individuals aged 20–59 years, living with a partner in a household with at least one child under the age of 18 years.
Sources: OECD, 2012. OECD Family Database, OECD, Paris. www.oecd.org/social/family/database and MTUS, Jonathan Gershuny and Kimberly Fisher, with Evrim Altintas, Alyssa Borkosky, Anita Bortnik, Donna Dosman, Cara Fedick, Tyler Frederick, Anne H. Gauthier, Sally Jones, Jiweon Jun, Aaron Lai, Qianhan Lin, Tingting Lu, Fiona Lui, Leslie MacRae, Berenice Monna, José Ignacio Giménez Nadal, Monica Pauls, Cori Pawlak, Andrew Shipley, Cecilia Tinonin, Nuno Torres, Charlemaigne Victorino, and Oiching Yeung, 2012. Multinational Time Use Study, Versions World 5.5.3, 5.80, and 6.0 (released October 2012). Centre for Time Use Research, University of Oxford, United Kingdom. www.timeuse.org/mtus/. Because of lack of more recent harmonized time diary data, we report averages for the period 2000–04 period for all outcomes. Time use data for Japan come from the Survey on Time Use and Leisure Activities (2001).

in Japan) with women doing on average about twice as much as men; still, men's contribution to childcare is marginally higher in near-replacement countries.

Two other interesting facts emerge from these data. Firstly, there are greater cross-country gender differences in housework and paid work time than in childcare time, suggesting that parents treat childcare, but not housework, as a form of investment (Guryan et al., 2008); as such, childcare may be less subject to change as a result of evolving social norms. Secondly, perfect gender equality (a value near 1) does not seem to be a requisite for near-replacement fertility levels; in fact, even the more egalitarian societies are far from full equality in these three important daily activities: childcare, housework, and paid work.

To further explore the fertility recovery hypothesis, **Figure 2** presents the time series of fertility, female labor force participation, and domestic work (measured here as the sum of housework and childcare time) for selected countries within each fertility group. In countries with current fertility levels near replacement (USA and Norway), fertility declined until the 1980s, but then a threshold ratio of men-to-women domestic work time (about 0.5) was crossed, and they began to recover fertility. In contrast, countries with lowest-low fertility levels (Italy and Spain) have yet to cross such threshold, and show no sign of sustained fertility recovery. Even more, these countries seem to be at different stages of a similar process: the gender pattern of domestic work observed in low-fertility countries at the beginning of the 2000s is remarkably similar to that of near-replacement-fertility countries in the 1960 to 1970s. However, new research (Sullivan et al., 2014) shows that, among lowest-low fertility countries, early signs of fertility recovery are associated with greater participation in domestic work among educated young men; to the extent that gender equitable attitudes diffuse, this could lead to future fertility gains among lowest-low fertility countries.

Conclusion and Further Research

In the industrialized world, fertility declines varied in rates and timing across countries, resulting in two broadly defined fertility regimes. In some countries, such as the USA and Northern Europe, fertility never fell far below the replacement rate of 2.1 children per women, or bounced back and stabilized near that level. In Southern Europe, Germany, and Japan, fertility declined to substantially lower levels without recovery. In this article, we reviewed an expanding area of research located at the intersection between fertility and the gender distribution of household labor, and presented cross-country evidence from time diary data suggesting that increases in men's share of home production might partially explain fertility recovery in advanced economies.

Some questions remain which deserve further consideration. More could be said on the role of attitudes in mediating the relationship between gendered time use patterns and fertility. Goldscheider et al. (2013) report that, in Sweden, what most depresses couples' fertility is the inconsistency between gender attitudes and behaviors: when 'reality' in the distribution of housework (but not childcare) disagrees with women's 'ideals' of fairness, second births (but not other-parity births) are significantly delayed. Similarly, Brodmann et al. (2007) find that the effect on fertility of a higher male contribution in housework is stronger for career-oriented women.

This area of inquiry could also benefit from interdisciplinary dialogue. Complementary to arguments about gender equity and fairness – prevalent in sociology and demography – policy intervention could attend to reasons of economic efficiency. Prior economic research has identified peer effects and positive externalities in the spread of social norms and the diffusion of gender-equitable behaviors (De Laat and Sevilla-Sanz, 2011), which in turn may affect childbearing decisions (Fernández and Fogli, 2006). Under positive externalities, couples who do not engage in egalitarian behavior can still free-ride on the benefits from higher aggregate fertility resulting from couples engaging in more equal division of labor. As a result, the number of egalitarian couples would be less than socially optimal. This suggests that, even if targeted to a limited number of individuals, intervention to reach a more equal sharing of housework and to increase fertility, would have rippling effects by increasing the fertility of other couples. Similarly, inefficiencies in the distribution of household chores and in the achievement of desired fertility could arise from imperfect information among partners. Given the 'public good' nature of children in the family (Lundberg and Pollak, 1993), and the difficulty in legally enforcing marital contracts designed to guarantee cooperation in the allocation of domestic time (De Laat and Sevilla-Sanz, 2011; Rasul, 2008), partners could free-ride on each other's contributions.

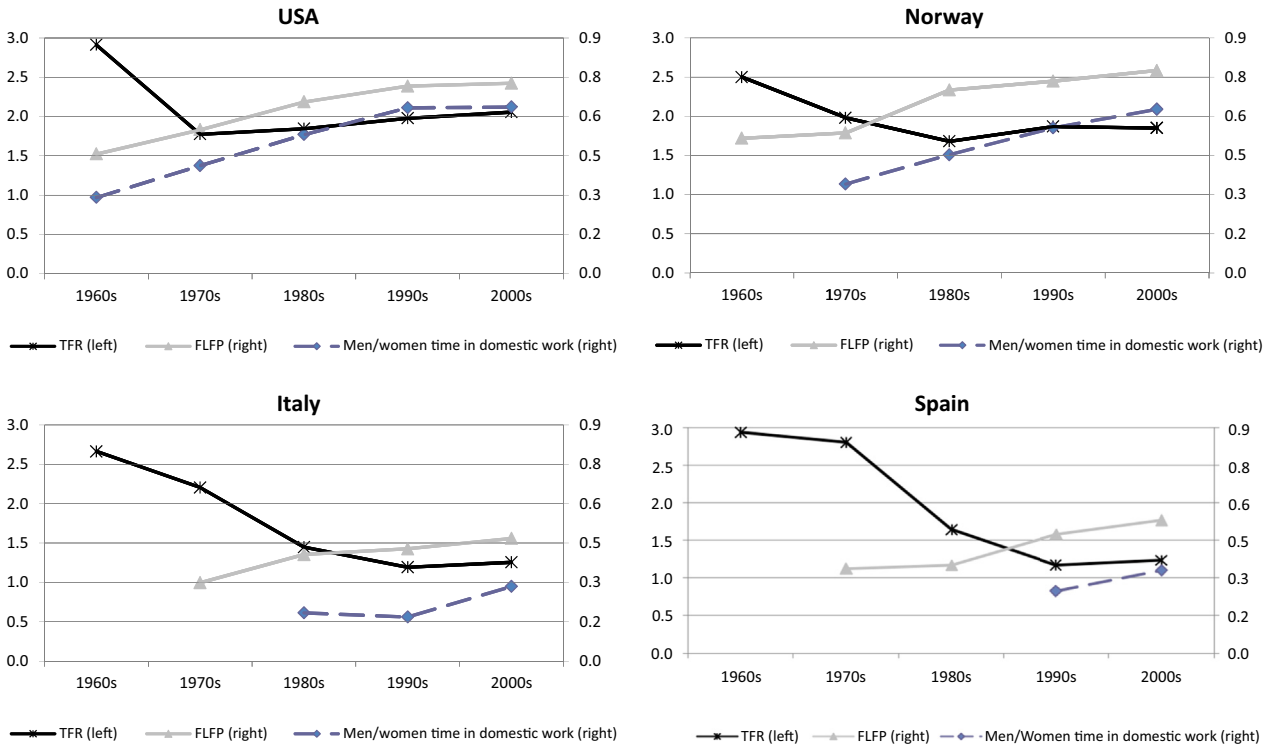

Figure 2 Trends in fertility, female labor force participation, and gender division of domestic work. Note: Figure displays total fertility rate (TFR), female labor force participation (FLFP), and the men–women ratio of time in domestic work (includes housework and childcare). Given that it is not possible to restrict the sample to households with children in the MTUS time series, we include all individuals aged 20–59 years, unlike the numbers reported in **Table 1**. Given that parenthood exacerbates gender specialization in domestic work, we can consider the time estimates above a bit conservative. Sources: OECD, 2012. OECD Family Database, OECD, Paris. www.oecd.org/social/family/database and MTUS, Jonathan Gershuny and Kimberly Fisher, with Evrim Altintas, Alyssa Borkosky, Anita Bortnik, Donna Dosman, Cara Fedick, Tyler Frederick, Anne H. Gauthier, Sally Jones, Jiweon Jun, Aaron Lai, Qianhan Lin, Tingting Lu, Fiona Lui, Leslie MacRae, Berenice Monna, José Ignacio Giménez Nadal, Monica Pauls, Cori Pawlak, Andrew Shipley, Cecilia Tinonin, Nuno Torres, Charlemaigne Victorino, and Oiching Yeung, 2012. Multinational Time Use Study, Versions World 5.5.3, 5.80, and 6.0 (released October 2012). Centre for Time Use Research, University of Oxford, United Kingdom. www.timeuse.org/mtus/.

Questions also remain on the relative importance of cultural and social norms vis-à-vis institutional context and policy intervention. Distinguishing between norms and institutions is empirically hard, given the simultaneity of both processes; and findings so far are mixed. Some studies have used immigrants to isolate the effect of social norms (from the country of origin) from institutional factors (at destination), finding that characteristics of the country of origin play an important role in determining women's employment behavior (Fernández and Fogli, 2006; Fogli and Fernández, 2009). In contrast, Klüsener et al. (2013) exploited a natural experiment in a culturally and linguistically German region in Belgium (two countries with diverse fertility outcomes), to find that institutional and policy circumstances trump cultural norms when predicting fertility.

The descriptive associations presented in this article also deserve further investigation. From our comparison across countries of recent trends in fertility and in the gender distribution of time in domestic activities, we advanced a tentative hypothesis: that there is a threshold level of domestic gender equity that needs to be crossed to trigger a process of fertility recovery in industrialized nations. Testing whether this threshold is similar across countries, or whether it is shaped by cultural or institutional factors, might constitute a fruitful avenue for future research. To that end, time diary data provide an invaluable resource. Important time diary resources are already available, such as the MTUS (2012) with more countries and waves being included and more data being harmonized every year, both from developed and developing countries. Such efforts should be continued and more widely used by researchers in the behavioral and social sciences.

See also: Family Theory: Economics of Childbearing; Family and Work; Fertility Change: Quantum and Tempo; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Fertility and Social Networks; Gender, Economics of; Second Demographic Transition; Time Use Research Methods; Time Use and Gender.

Bibliography

Aguiar, M., Hurst, E., 2007. Measuring trends in leisure: the allocation of time over five decades. The Quarterly Journal of Economics 122 (3), 969–1006.

Ahn, N., Mira, P., 2002. A note on the changing relationship between fertility and female employment rates in developed countries. Journal of Population Economics 15 (4), 667–682.

Becker, G.S., 1965. A theory of the allocation of time. The Economic Journal 75 (229), 493–517.
Becker, G.S., 1991. A Treatise on the Family. Harvard University Press.
Bettio, F., Villa, P., 1998. A mediterranean perspective on the breakdown of the relationship between participation and fertility. Cambridge Journal of Economics 22, 137–171.
Bianchi, S.M., 2000. Maternal employment and time with children: dramatic change or surprising continuity? Demography 37 (4), 401–414.
Billari, F., Kohler, H.-P., 2004. Patterns of low and lowest-low fertility in Europe. Population Studies 58 (2), 161–176.
Bongaarts, J., Sobotka, T., 2012. A demographic explanation for the recent rise in European fertility. Population and Development Review 38 (1), 83–120.
Brewster, K.L., Rindfuss, R.R., 2000. Fertility and women's employment in industrialized nations. Annual Review of Sociology 26, 271–296.
Brodmann, S., Esping-Andersen, G., Güell, M., 2007. When fertility is bargained: second births in Denmark and Spain. European Sociological Review 23 (5), 599–613.
Caltabiano, M., Castiglioni, M., Rosina, A., 2009. Lowest-low fertility: signs of a recovery in Italy? Demographic Research 21, 681–718.
Cooke, L.P., 2004. The gendered division of labor and family outcomes in Germany. Journal of Marriage and Family 66 (5), 1246–1259.
Cooke, L.P., 2009. Gender equity and fertility in Italy and Spain. Journal of Social Policy 38 (1), 123–140.
De Laat, J., Sevilla-Sanz, A., 2011. The fertility and women's labor force participation puzzle in OECD countries: the role of men's home production. Feminist Economics 17 (2), 87–119.
De Laat, J., Sevilla-Sanz, A., 2006. Working Women, Men's Home Time and Lowest-low Fertility. ISER Working Paper Series No. 2006–23.
Esping-Andersen, G., 2009. Incomplete Revolution: Adapting Welfare States to Women's New Roles. Polity, Cambridge.
Esping-Andersen, G., Boertien, D., Bonke, J., Gracia, P., 2013. Couple specialization in multiple equilibria. European Sociological Review 29 (6), 1280–1294.
Fernández, R., Fogli, A., 2006. Fertility: the role of culture and family experience. Journal of the European Economic Association 4 (2–3), 552–561.
Feyrer, J., Sacerdote, B., Stern, A.D., 2008. Will the stork return to europe and Japan? understanding fertility within developed nations. The Journal of Economic Perspectives 22 (3), 3–22.
Fogli, A., Fernandez, R., 2009. Culture: an empirical investigation of beliefs, work, and fertility. American Economic Journal: Macroeconomics 1 (1), 146–177.
Galor, O., Weil, D.N., 1996. The gender Gap, fertility, and growth. American Economic Review 86 (3), 374–387.
Gimenez-Nadal, J.I., Sevilla, A., 2012. Trends in time allocation: a cross-country analysis. European Economic Review 56 (6), 1338–1359.
Goldin, C., Katz, L., 2002. The power of the pill: oral contraceptives and women's career and marriage decisions. Journal of Political Economy 110 (4), 730–770.
Goldscheider, F., Bernhardt, E., Brandén, M., 2013. Domestic gender equality and childbearing in Sweden. Demographic Research 29, 1097–1126.
Goldscheider, F., 2012. The gender revolution and the second demographic transition: understanding recent family trends in industrialized societies. In: 2012 European Population Conference "Gender, Policies and Population". Stockholm.
Goldstein, J.R., Sobotka, T., Jasilioniene, A., 2009. The end of 'lowest-low' fertility? Population and Development Review 35 (4), 663–699.
Greenwood, J., Seshadri, A., Yorukoglu, M., 2005. Engines of liberation. Review of Economic Studies 72 (1), 109–133.
Guryan, J., Hurst, E., Kearney, M., 2008. Parental education and parental time with children. The Journal of Economic Perspectives 22 (3), 23–46.
Hakim, C., 2000. Work-lifestyle Choices in the 21st Century: Preference Theory. Oxford University Press, USA.
Hochschild, A.R., Machung, A., 1989. The Second Shift: Working Parents and the Revolution at Home. Viking, New York.
Jaumotte, F., 2003. "Female Labour Force Participation: Past Trends and Main Determinants" in OECD Countries. OECD Economics Department Working Paper No. 376. OECD Publishing.
Kan, M.Y., Sullivan, O., Gershuny, J., 2011. Gender Convergence in domestic work: discerning the effects of interactional and institutional barriers from large-scale data. Sociology 45 (2), 234–251.
Klüsener, S., Neels, K., Kreyenfeld, M., 2013. Family policies and the WESTERN European fertility divide: insights from a natural experiment in Belgium. Population and Development Review 39 (4), 587–610.
Kohler, H.-P., Billari, F.C., Ortega, J.A., 2002. The Emergence of lowest-low fertility in europe during the 1990s. Population and Development Review 28 (4), 641–680.
Lesthaeghe, R., 1983. A Century of demographic and cultural change in Western Europe: an exploration of underlying dimensions. Population and Development Review 9 (3), 411–435.
Lesthaeghe, R., 1995. The second demographic transition in Western countries: an interpretation. In: Mason, K., Jensen, A. (Eds.), Gender and Family Change in Industrialized Countries. Clarendon Press, Oxford, pp. 17–62.
Lesthaeghe, R., 2007. Second demographic transition. In: Ritzer, G. (Ed.), Encyclopedia of Sociology, Blackwell.
Lesthaeghe, R., 2010. The unfolding story of the second demographic transition. Population and Development Review 36 (2), 211–251.
Lesthaeghe, R., Surkyn, J., 1988. Cultural dynamics and economic theories of fertility change. Population and Development Review 14 (1), 1–45.
Lundberg, S., Pollak, R.A., 1993. Separate spheres bargaining and the marriage market. Journal of Political Economy 101 (6), 988–1010.
McDonald, P., 2000a. Gender equity in theories of fertility transition. Population and Development Review 26 (3), 427–439.
McDonald, P., 2000b. Gender equity, social institutions and the future of fertility. Journal of Population Research 17 (1), 1–16.
Milkie, M.A., Raley, S.B., Bianchi, S.M., 2009. Taking on the second shift: time allocations and time pressures of US parents with preschoolers. Social Forces 88, 487–518.
Mincer, J., 1985. Intercountry comparisons of labor force trends and of related developments: an overview. Journal of Labor Economics 3 (1), S1–S32.
MTUS, Gershuny, Jonathan, Fisher, Kimberly, Altintas, Evrim, Borkosky, Alyssa, Bortnik, Anita, Dosman, Donna, Fedick, Cara, Frederick, Tyler, Gauthier, Anne H., Jones, Sally, Jun, Jiweon, Lai, Aaron, Lin, Qianhan, Lu, Tingting, Lui, Fiona, MacRae, Leslie, Monna, Berenice, Giménez Nadal, José Ignacio, Pauls, Monica, Pawlak, Cori, Shipley, Andrew, Tinonin, Cecilia, Torres, Nuno, Victorino, Charlemaigne, Yeung, Oiching, 2012. Multinational Time Use Study, Versions World 5.5.3, 5.80 and 6.0 (released October 2012). Centre for Time Use Research, University of Oxford, United Kingdom. www.timeuse.org/mtus/.
Murphy, M., 2001. Family theory and the realities of childbearing behavior. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Pergamon, Oxford, pp. 5361–5367.
Myrskylä, M., Goldstein, J.R., Cheng, Y.A., 2013. New cohort fertility forecasts for the developed world: rises, falls, and reversals. Population and Development Review 39 (1), 31–56.
OECD, 2012. OECD Family Database. OECD, Paris. www.oecd.org/social/family/database.
Oláh, L.S., 2003. Gendering fertility: second births in Sweden and Hungary. Population Research and Policy Review 22 (2), 171–200.
Rasul, I., 2008. Household bargaining over fertility: theory and evidence from Malaysia. Journal of Development Economics 86, 215–241.
Sánchez-Barricarte, J.J., Fernández-Carro, R., 2007. Patterns in the delay and recovery of fertility in Europe. European Journal of Population 23 (2), 145–170.
Schultz, T.P., 2001. The Fertility Transition: Economic Explanations. SSRN Scholarly Paper No. ID 286291. Social Science Research Network, Rochester, NY.
Sobotka, T., 2011. Fertility in Central and Eastern Europe after 1989: collapse and gradual recovery. Historical Social Research 36 (2), 246–296.
Stycos, J.M., Weller, R.H., 1967. Female working roles and fertility. Demography 4, 210–217.
Sullivan, O., Billari, F.C., Altintas, E., 2014. Father's changing contributions to child care and domestic work in very low fertility countries: the effect of education. Journal of Family Issues 35 (8), 1048–1065.
Testa, M.R., Skirbekk, V., Lutz, W., 2006. The low fertility trap hypothesis. Forces that may Lead to further postponement and fewer births in europe. Vienna Yearbook of Population Research 4, 167–192.
Torr, B.M., Short, S.E., 2004. Second births and the second shift: a research note on gender equity and fertility. Population and Development Review 30 (1), 109–130.
Van De Kaa, D.J., 1987. Europe's second demographic transition. Population Bulletin 42 (1), 1–59.
Weil, D.N., 1997. Chapter 17. "The economics of population aging". In: Rosenzweig, Mark R., Stark, Oded (Eds.), Handbook of Population and Family Economics. Elsevier, pp. 967–1014.
Willis, R.J., 1973. A new approach to the economic theory of fertility behavior. Journal of Political Economy 81 (2), S14–S64.

Gender: Anthropological Aspects

Beverly Stoeltje, Indiana University, Bloomington, IN, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by B. Aretxaga, volume 9, pp. 5915–5920, © 2001, Elsevier Ltd.

Abstract

Having moved from the margins of anthropology to a position of centrality in the discipline, studies of gender now tend to focus on ethnographic and historical particularities and on interrelationships between gender and other aspects of society and culture, including prominently language and law, health and reproduction, religion, violence, and gender identity as a situated and performative accomplishment. The role of gender under conditions of regional and world historical change, such as postcolonial and postsocialist transformations, is a significant organizing principle in these lines of inquiry.

Conceptualizing Gender in the Postmodern Era

The term *gender*, to address and acknowledge the concepts of female and male as constructed by specific societies, entered into widespread anthropological use in the 1980s. Although a select few female anthropologists in the first half of the twentieth century conducted pioneering field research and published studies of women in a wide range of societies, it was not until the 1970s that feminist anthropologists established women as an independent topic. This newly established interest led to edited volumes in the 1980s linking gender with classic anthropological topics such as colonization, kinship, language, and regional studies. Studies explored women's reproductive lives, their role in shaping and maintaining kinship relations and their participation in the new transnational labor market; the relation of gender to colonialism and nationalism became a primary topic, and class and ethnic differences were identified.

A transitional decade for gender studies, the 1980s shifted the direction of research from recognizing gender differences and providing explanations to showing the effects of social and political systems on gender relations, mediated by discourse.

Crossing a Threshold to a New Configuration

From the marginal position it occupied in the early years, gender studies claimed a space at the dynamic center of anthropology in the 1990s. More than a continuation, this centrality represents the crossing of a threshold to a position that recognizes gender as foundational to every set of social and political circumstances. Attributed in part to the United Nations decade for women (1975–85) and the Beijing conference in 1995, epistemological changes were also occurring within anthropology. This repositioning can best be described as a *reconfiguration*.

In contrast to the view of gender as a static, stable category, studies of the 1990s and 2000s have recognized that 'gender' is a process, a situated accomplishment in the conduct of social life, and further, that multiple variations of the concept of gender may exist within one culture as well as across cultures. Like anthropologists generally, those who focus on gender have responded to the social and political developments of the late twentieth century with engagement and activism in the effort to link global and transnational movements to the local and new knowledge to emerging situations. Consequently, relevance to the conditions and events of contemporary social and political life has become an important measure, an indicator of valuable work in the anthropology of gender.

Theoretical approaches that shape the analytical perspectives of twenty-first century scholars are sometimes imported from the larger domain of social theory or from intellectual movements in the field of social and behavioral sciences. Rather than relying upon theories that strive for universality, or those of Western feminism that have been much criticized, research in the 2000s delves into complexity, recognizing the interaction of continuity with modernity in situations often characterized by ambiguity and contingency, leading to a focus on specific locations, people, and circumstances, in contrast to the more general term, *culture*.

Erasing Boundaries and Fusing Categories

Evidence of the centrality of gender to anthropology can be found in the sections on gender that appear in sociocultural textbooks, often coupled with 'sex' or 'sexuality.' Also important are the anthologies devoted solely to articles on gender as studied by anthropologists. The major anthropological journals publish the most recent gender research, and ethnographic studies featuring gender continue to occupy a prominent place in the catalogs of scholarly presses.

Conceptually characteristic of the new configuration is the erasure of boundaries separating disciplines and a fusing of categories and terms. For example, the long-separated terms, *sex* and *gender*, have now been reunited (McConnell-Ginet, 2011) while gender is combined with numerous other terms relevant to a topic such as identity, politics, Islam, and land (Hoffman, 2008). Some authors conflate the terms *feminism* and *gender* while others argue for a distinction, stating that the goal of *feminism* is the identification of inequality and the improvement of women's position whereas *gender* encompasses both females and males and their interaction as well as the production of knowledge related to these concepts. However, in many parts of Africa, the Middle East, and Asia, *feminism* continues to be a contested concept associated with the West,

whereas *gender* has largely been acknowledged as relevant and studies increased in the past decade.

Taking a broad overview of gender studies reveals a radical change in the division of labor among disciplines. Those that have conventionally studied domestic societies or restricted themselves to the past have now extended their research to non-Western societies and/or to the present. Working alongside anthropologists are sociologists, historians, political scientists, and scholars from the newer fields of Women's Studies or Gender Studies and area studies departments. The results of this common interest have brought gender scholars together and demonstrated the value of collaboration. The transdisciplinary approach has been a principled effort to transcend the biases of the various Western disciplines. The outcome has been conferences and exchanges and a rich body of edited volumes on gender: Joseph (2000), Lutkehaus and Roscoe (1995), Lamphere (1997), and Cohen et al. (1996).

Global Developments: Region, Revolution, and Religion

In regions of the world in transition studies of gender have flourished, some of which address large-scale social change, others that identify a regional focus, and some that have religion as the basis of the work. Researchers working in these areas select for ethnographic specificity and the identification of key variables rather than generalization, a scholarly strategy that tends to erase gender differences and plural perspectives on any subject.

A major development in addition to the twenty-first century turn to regions and religion is the prevalence of scholars who are native to the region or members of the religion, bringing insider and outsider views together, and, at times, into question. Equally significant are the numerous border intellectuals, those who occupy an interstitial cultural space, allowing them to speak with freedom from any national or cultural perspective.

Although anthropological studies in Eastern Europe are new since the fall of the Soviet Union, a rich body of research has been undertaken. Gender issues are a high priority for a group of scholars from several disciplines. Anthropologists have emphasized the importance of recognizing continuities with the past at the same time noting that the presence of the West in the region is indispensable today. These scholars have argued for a theoretical framework based on reciprocal effects: states produce ideologies and policies that circumscribe relations between men and women, but ideas about gender differences also shape the construction of the state.

Moreover, a gendered perspective highlights important aspects of the postsocialist transformations that have been taken for granted. Although the independent states in this region were once a part of the regime of state socialism, they have emerged with many different concepts about gender relations. Among the issues of greatest concern to gender scholars is the question of rights and policies relating to reproduction in the new independent or newly unified states: the discourse and policies surrounding reproductive rights (abortion, sterilization, sexuality), although issues of labor, class, and identity are also of concern (Gal and Kligman, 2000).

Muslim women were introduced to anthropologists in the West by Elizabeth Warnock Fernea with her pathbreaking book in 1965. Within the space of Fernea's lifetime, women in Muslim societies have captivated the attention of scholars, Muslim and non-Muslim alike. Gender in the Middle East and women in Muslim societies became relevant topics in the 1980s (Abu-Lughod, 1986), to extend in the 2000s from Africa to Central Asia, and taking account of the construction of gender in circumstances as different as neocolonial modernity and Islamic fundamentalism. Recognition of the scale of the subject is the six-volume *Encyclopedia of Women and Islamic Cultures* (Joseph, 2003–2008).

As Muslim women have so often been represented as the 'Other,' a practice revived since 9/11, studies focused on gender in Islam have sought to unravel the effects of global change and to unmask the discursive practices through which the image of the repressed woman has been portrayed. Research has centered on Muslim women's political involvement (through mothering practices, religious piety, and social activism), the gendering of citizenship, the legal privileging of males, religious nationalism, and other issues that result in inequities.

As gender scholars have argued in the past, women are often made into symbolic markers of the nation, and, indeed, research on Middle Eastern societies show that when women are made into symbolic markers, institutions and processes of the nation subsidize efforts to ensure control over reproduction, maternal rights, productivity, and wealth. The symbolic equation of women with the nation often leads to subordination of actual women through calls for the preservation of tradition, the family, codes of ethic, values, and other claims. Of particular interest, these studies have shown that patriarchy weaves together civil society, the state, marketing, and the family.

Two specific locations deserve comment due to events of consequence. In Turkey, a country where the dominant religion is Islam, gender has emerged as a fixture in national discourse due to major changes in the country. Since independence in 1923, Turkey has been defined as a secular country, but a change in government has allowed major shifts with regard to the public expression of religion. The resulting controversies frequently focus on gender and on the wearing of 'headscarves' by Muslim women. The headscarves function as an icon of the larger discussion on the relationship of women and religion, a topic taken up by numerous scholars. In North Africa, the Arab Spring has created upheaval and uncertainty. While most questions about the future remain unanswered, anthropologists have already begun documenting the experience, building on a significant body of scholarship on the women's movement in Egypt especially. Of renewed interest also is the life of Um Kulthum, the singer who became the national symbol of Egyptian independence in the colonial era.

Among Africanists, gender and the related concepts – feminism, women, masculinity – have generated a significant body of literature that characteristically addresses colonialism and postcolonialism in Africa or challenges the domination of Western paradigms on gender. Especially significant in the past two decades, the number of scholars from the African continent has increased notably. Some have degrees from African universities, others from England or France, as well as the United States, and they hold positions in agencies as well as universities. Research on gender is being published in Africa

and in Europe and also by the development agencies that publish in-house reports. Consequently, the literature on gender in Africa is widely dispersed and issues discussed among one set of scholars do not necessarily reach into another set of scholarly discussions.

Central to the controversies among gender scholars are models developed by scholars and activists in the West, considered not applicable in Africa due to the major differences between African societies and Western ones. Kinship systems are radically different as are religions, political organizations, and economies. Coupled with the powerful effects of colonialism and modernity on African societies that intensified gender hierarchy and female subordination, these differences have been highlighted and revised agendas formulated for research. The most frequently cited example of difference concerns the family and the position of the individual person. In the West, individual autonomy is the goal, and the nuclear family is the fundamental social unit, emphasizing the dichotomy – female and male. In contrast, the African family model has been described as a corporate group – one based on lineage and clan through which members maintain their cultural forms and ties, rather than loosen or sever them, making contributions and receiving benefits. Western models tend to create fixed categories whereas African ones are quite fluid, making adaptation, substitution, and transformation possible. Gender and sexuality are complex concepts, embedded in all dimensions of social life in most African societies. Abstract as well as material, they have rich metaphorical capacities, adding layers of significance to social life, which often remain opaque to outsiders.

Yet another difference concerns the goals of women. In the many different societies of Africa, women are struggling to achieve gender equity, but that does not mean that women want to be similar to men; alternatively, African women value their unique female traits and styles (particularly dress), their economic activities, and their role as mothers, all together.

The past two decades have witnessed a move among anthropologists well beyond the paradigms of the previous century, taking account of the complexities of African gender systems. At the forefront have been studies demonstrating that gender is at the core of African cultural models. Of importance has been the recognition that high-status females in positions of leadership (both political and religious), usually paralleling those of males in a dual gender system, characterize a great many African societies (Kaplan, 1997). Due to colonialism, their power was reduced considerably in many societies though Asante queen mothers in Ghana remain strong. Equally relevant is women's participation in and contribution to the economy, a factor that links the household to the larger entities of community, clan, and local political systems (Mikell, 1997). Developed in the 2000s, research has shown that African economic strategies operate within a plurality of different systems of value; many systems of exchange are regulated by popular convention rather than the formal systems of the state. Pointing to the effectiveness of popular economic activity, it has been argued that viewing formal and popular economic systems as a single configuration would advance our understanding of global economic realities, a consideration that would count women's contribution legitimate (Guyer, 2004).

Major developments in social and political theory have resulted in a rich body of gender studies, some of which feature women speaking for themselves (Stoeltje, 2009) while others have directed attention to the symbolic, religious, and economic value women place on cloth as it circulates, is displayed, or collected; yet another group of scholars has focused on fashion, dress, and clothing, while anthropologists have also focused on the female body as symbolic communication – the meaning of naked and nude, as vehicle for spirits, and the body as bathed, revealing one's moral and civic personhood (Masquelier, 2005). Important also are studies addressing the relation between women and the state, the impact of HIV/AIDS on women (whether wives or sex workers), health and reproduction, women's resistance to authorities (wicked women), and gender and the law (state law and legal pluralism).

Emergent Syntheses and Twenty-First Century Configurations

Identity, Sexuality, and Subjectivity

Not surprisingly, issues of identity have proven to be longstanding in anthropology because identity becomes a conscious feature when others are perceived as different. In recent years, sexuality, subjectivity, and gender are frequently conflated with identity and used to address issues of belonging, recognition, and legitimation. While the concept of subjectivity is broadly applied by scholars, recent work in anthropology on subjectivity argues that gender is both the embodiment of the symbolic and the expression of freedom (Moore, 2007).

As one gendered feature of identity, masculinity studies have explored how manhood is constituted, especially through initiation rites, and how it is performed and evaluated in societies in every continent. Noting the connection between masculinity and subjectivity, scholars have observed competing notions of masculinitiy within groups and different expectations of masculinity for different age groups.

Questions of sexual orientation are intimately associated with subjectivity and identity: gay and lesbian, transexual, two spirit, third social gender, and others. Anthropological research on sexual subjectivities has been refined theoretically in the past two decades and now takes non-Western gay and lesbian subjectivities as legitimate forms of selfhood and links up with research on sex work and trafficking (Boellstorff, 2007). Research on subjectivities in the United States previously overlooked includes the rural gay and lesbian youth in the Midwest of the United States (Gray, 2009) and gay fatherhood (Lewin, 2009). Rediscovered are the two-spirit people of Native American groups (earlier labeled *berdache*) who have the potential to recognize gender diversity (Jacobs, et al., 1997).

Turning to Southeast Asia, a region where alternative sexualities were tolerated until the twentieth century, when they were criminalized, a historical perspective shows a region known for its wide array of legitimate sexual roles and practices. Research employing the concept of pluralism demonstrates that the dynamics of gender and sexuality are deeply imbricated in the politics of race, ethnicity, religion, and class (Peletz, 2009). Anthropologists are also documenting issues of gender

identity in other Asian countries, including the emergence of gays and lesbians in India, and their legitimation by a high court in New Delhi, questions of being gay and Muslim in Indonesia, and the public discourse about sex and sexuality in China. In the neighboring Pacific island of Tonga, gender, identity, and subjectivity are publicly performed, negotiated, and displayed through beauty pageants: the Miss Heilala contest for young women and the Miss Galaxy "other pageant" that showcases transgender roles (Besnier, 2011).

Gender as an Analytic Category

Among the most significant but subtle theoretical developments in the past decade has been the recognition of gender as a resource for shaping phenomena beyond recognizable gender categories. Although lacking a specific name, this approach represents gender as an analytical category. Tracing it as it diffuses throughout dimensions of social life and domains of policy and practice at the cultural, national, or global level, it penetrates aspects of social relations otherwise opaque, including national and international events and movements. A widespread but seldom recognized practice involves the construction of knowledge in any institution or nation-state whereby one kind of knowledge obstructs another. Research has analyzed the means by which the authority of professional knowledge elides gendered knowledge in specific instances, raising the question, when does attention to gender difference emerge (Boyer, 2006).

As an analytic category, gender can shape the organization of specific cultures as well, affecting social relations, ritual, and identity. For the Kabre of Togo, the fluid concept of gender permeates both practical and symbolic life, organizing the material arrangement of houses, and courtyards, designing initiation, defining clans, naming the seasons, and ultimately determining the burial site of the individual. Gender is the construct that balances marriage, kinship, nurturance, and rites of passage, and organizes rituals that bring about the changing of seasons, including powerful gendered symbols. As the dynamic force diffused through all domains of Kabre life, gender provides the symbolic map for the construction and interpretation of the cosmological as well as the social (Piot, 1999).

Health and Healing, Disability, and Disease

Among the specific areas that have attracted gender scholars in the 2000s is the domain of health, a topic of increasing concern in the twenty-first century in every society. Women are centrally engaged in matters of health, first because of childbirth and second because of the care and nurture of children in most societies. Sexuality and sexual behavior of males and females, including sex workers and sexual relations within marriage, all have drawn the attention of anthropologists because of the crisis around HIV/AIDS. These particular issues and closely related ones have risen to prominence due to wars and conflict, poverty, the spread of disease, and government refusal to respond to these conditions (Briggs, 2004), yet in some instances women have organized to confront the failures of the government (Susser, 2009). At the same time, women healers continue to practice their art in some societies. Among the Tuareg, for example, the medicine women who have expertise with herbs and natural cures continue to treat the female population, also providing psychological support (Rasmussen, 2006).

Disability, too, creates gendered responses, one of which is the phenomenon of male attraction to females with disabilities (such males are known as *devotees*). Approaching the problem of disability on a national/global scale, scholars have found it to be a widespread condition in the former Soviet Union, and rampant in war zones.

Gender and the Dynamics of Violence

Theoretical perspectives have developed around topics that have made headlines in the twenty-first century. Violence dominates in that news. Consequently, scholars of the past two decades have been in conversation around that subject, developing theories to account for the dynamics of violence and its relationship to state power as well as domestic violence that can disrupt and unmake the world of those who inhabit the space. As most anthropologists have argued persuasively, violence links to larger social and political ideologies and understandings that connect the national to the domestic and empires to colonies, with the category of gender as essential for understanding that connection which is found to reside in foundational stories about the nation-state.

Within liberal philosophy, the emphasis on giving life to the nation and dying for the nation are powerful because they seem to normalize violence as part of gendered belonging to the nation-state. The implications of this position are that men should be ready to bear arms and to die for the nation and that women's reproduction rightly belongs to the state (Das, 2008; Aretxaga, 1997). More specific studies have addressed gender violence in context, revealing underlying ideologies that legitimate the many forms resulting in violence. The topic of one such study, the women's role in the Dirty Protest at the Armagh prison (Ireland), linked the shocking power of the women's primordial symbols with the domains of prison violence, colonial history, and gender discourse and argued that gender is an intrinsic component of violence (Aretxaga, 1995).

Yet another specific study is located in a South African neighborhood (where unemployment rises to 30%). Drawing on gender theory focused on males, this research found that young men who join gangs and engage in acts of violence are shaped by values anchored in the ideology of masculinity that held sway during the apartheid era (Salo, 2006).

Advances in Language, Law, and Performance

Performance

Performance theory, in combination with ethnography, directs attention to gender issues and ideologies as they are enacted for an audience. Whether the more individualized performance that focuses on the spoken word or song or a large-scale event with many participants (ritual, festival, markets, sports, and drama), these events capture the imagination with a shift to a time, place, and action set apart from the mundane and quotidian present. As developed within anthropology, a performance approach identifies what activities women and

men are engaged in before an audience as an expression of their preoccupations and commentaries on social life. Performance has the potential to enact resistance, memory, irony, satire, devotion, competition, transformation, or other forms. Considered from two perspectives – one in which the verbal is the focus and the other, a cultural performance – these are often integrated.

Studies concentrating on women as performers within gendered societies around the globe have identified rich repertoires of female genres and performance spaces. Moreover, these studies have proven to yield subtle understandings of how women are inscribing new values into their lives in situations of rapid social transformations. In markets in Morocco, for example, women's presence has increased and with that they have created new roles and carved out unique discursive domains, mixing categories and revoicing expressive forms (Kapchan, 1996). Equally significant, Chicanas using a performance approach have taken a critical stance toward the concept of tradition, exploring how women musicians and performers in *charreada* (Mexican rodeo) understand the power and politics inherent in performing (Najera-Ramirez and Cantu, 2002).

Performances involving males as well as females reveal relations between the two, often challenging, contesting, and negotiating those roles. During annual ritual/festival events, women and men in societies of the South Pacific engage in clowning performances, reversing gender roles, and in the process providing cultural critique on gender roles (Mitchell, 1992).

In South India, female and male performers on the Tamil stage (called Special Drama) have provided sophisticated insight into their gendered positions in society and how they respond with a range of verbal skills (Seizer, 2005).

Viewing gender as performed permits the female to emerge as subject, and both female and male to perform and/or explore or transform roles, creating an opportunity for reflection and critique. Above all, performers experience achievement, accompanied by varying degrees of power and a space filled with promise.

Language and Law

In the 1970s and 1980s, linguistic anthropologists tackled the issues of difference in the speech of females and males through ethnographic studies in societies around the world, including the United States. In almost every instance, significant inequities were found, leading to new questions concerning speaking and silence. Studies of language use in relation to gender continued to captivate scholars' interest in the 1980s, who observed that men performed roles and genres associated with political authority and public settings, while women controlled other genres and settings such as lullabies and stylized mourning. A major issue was whether differences in language could be considered a women's language and a men's language as some scholars proposed; further studies argued that language and social status could be correlated, and features of language use considered to be characteristic of women were also used by men of low status; though women used this language, known as 'powerless language,' men of low status also used it. Other studies showed the importance of context and how the organization of gender behavior (men's groups, women's talk) affected language; an additional variable showed the use of 'women's language' as a communicative strategy.

The next two decades elaborated on these questions, considering who has access to speaking in any situation. Studies showed that the silencing of women's voices generally rests on an underlying ideology that endows women's speech with a negative evaluation, expressed through conversational strategies that discourage women's participation. Theories of the social use of language encouraged the analysis of how social relations are created, negotiated, or denied through language. Noting that some linguistic strategies and genres are more highly valued and carry more authority than others, and these practices occur within institutions where speakers impose their definitions of events, peoples, and actions on others, they have been labeled with the term, *symbolic domination* (Gal, 2006). Consequently, it was concluded that language serves as an index of position and of ideology. Exploration of the links between gender, language, and power further revealed the greatest impact to be in power's ability to impose its visions of reality, inscribed in language and enacted through interaction.

Ethnographic attention to women's voices in female linguistic genres from different classes, ethnic groups and societies found a wide range of genres, some of which were used to express opposition to the dominant vision of reality. An example from North India documented Hindu rural women who perform songs at women's rituals and tell stories explicating them through which they express poetic resistance to structures of power (Raheja and Gold, 1994). Shifting to the Warao of Venezuela and comparing male and female speech genres, it has been shown that the only public genre accessible to women is ritual wailing, performed at funerals; through it, however, women unmask hierarchies that sustain the power of the postcolonial institutions and the nation-state; in contrast men's gossip serves as a contested space in which women's discourse is delegitimized and men's is legitimized (Briggs, 1992).

Little known until recently is an instance of women's language. In China, literacy and language were linked together as a communicative form for peasant women in Jiangyong County during the late imperial period. This gender-specific language was known as *nushu* (female writing). Associated also with a special ritual friendship, these sisterhood communities often failed to survive after marriage due to village exogamy (Liu, 2004).

Intertwined with the analysis of language is the anthropology of law; whether considered indigenous, customary, or modern, legal systems employ language, translating everyday categories into legal language to effect powerful changes. It will be apparent in any court or legal proceeding that a legal system utilizes discursive practices embodying concepts of power consistent with the gender ideologies that characterize specific cultures' assumptions about gender. Characteristic of all legal systems also is status, embodied in the roles occupied by individuals in a court or other legal setting. The speech in legal settings, of all individuals, will reflect status and, often, a position on gender.

Although anthropologists initially analyzed language in legal settings in small-scale societies, by the 1990s, research included a focus on legal systems in the United States and other

Western-style legal systems. These included the language of judges, lawyers, witnesses, and litigants, and a focus on gender. Moreover, anthropologists working in countries with relatively new state legal systems began looking at gender as defined and exercised by the state and the courts in those countries. In spite of the claims of modern legal systems, the courts in postcolonial settings often require lengthy wait times and high fees, proving not helpful to those with low incomes. Often knowledge of gender and/or women in relationship to the state has been located in the area of family law, the legal domain that has the most direct influence on women, men, and children. Family law has been criticized because the colonial powers who created it often placed religion in the domain of family law. Religion-based family law divides the citizenry into separate legal communities which then exacerbates differences as it has in Egypt, and those differences are played out with women as the symbol standing for one religion or the other all too often. Furthermore, family law that incorporates religion becomes oppressive for women if the religious perspective privileges males.

Viewing women's speech in courts often proves to be the means by which we identify how much access women have to the law and what constraints women encounter in legal experiences. Studies of women in indigenous/customary courts as well as state courts have emphasized the importance of the links between social and legal identities and revealed the advantages and disadvantages women face in any legal system. For example, Swahili women in Kenya take their problems to an Islamic court where a generally sympathetic male judge (the kadi) hears their stories of marriage (Hirsch, 1998). In the customary courts of the Asante in Ghana, women, as well as men, speak for themselves, and queen mothers and their courts hear cases involving women. In other societies, however, women frequently encounter difficulties when dealing with either indigenous courts or modern ones (Stoeltje, 2002).

The effort to define women's rights as human rights has attracted legal anthropologists who have brought their skills to the subject to argue for the legitimation of women's rights and to pursue specific issues such as prostitution, the labor force, land rights, gender violence, sex trafficking, and other issues that impact the lives of women. These topics are intertwined with global networks, international law, and the United Nations, leading anthropologists engaged in gender studies into these territories (Merry, 2006).

See also: Fieldwork in Social and Cultural Anthropology; Queer Anthropology; Socialist Societies: Anthropological Aspects.

Bibliography

Abu-Lughod, Lila, 1986. Veiled Sentiments: Honor and Poetry in a Bedouin Society. University of California Press, Berkeley, CA.

Aretxaga, B., 1995. Dirty protest: symbolic overdetermination and gender in Northern Ireland ethnic violence. Ethos 23 (2), 123–148.

Aretxaga, B., 1997. Shattering Silence: Women, Nationalism and Political Subjectivity in Northern Ireland. Princeton University Press, Princeton, NJ.

Besnier, Niko, 2011. On the Edge of the Global: Modern Anxieties in a Pacific Island Nation. Stanford University Press, Stanford, California.

Boyer, Dominic, 2006. Gender and the solvency of professionalism: Eastern German journalism before and after 1989. East European Politics and Societies 20 (7), 152–179.

Boellstorff, Tom, 2007. Queer studies in the house of anthropology. Annual Review of Anthropology 36, 17–35.

Briggs, Charles (Ed.), 1992. "Since i am a woman, i will chastise my relatives": Gender, Reported Speech, and the (Re) production of Social Relations in Warao Ritual Wailing. *American Ethnologist* 19, pp. 337–361.

Briggs, Charles (Ed.), 2004. Stories in the Time of Cholera: Racial Profiling during a Medical Nightmare. University of California Press, Berkeley, CA.

Cohen, Colleen, Wilk, Richard, Stoeltje, Beverly (Eds.), 1996. Beauty Queens on the Global Stage. Routledge, New York, NY.

Das, Veena, 2008. Violence, gender, and subjectivity. Annual Review of Anthropology 37, 283–299.

Fernea, Elizabeth Warnock, 1965. Guests of the Sheik: An Ethnography of an Iraqi Village. Doubleday, Garden City, NY.

Gal, Susan, Kligman, Gail (Eds.), 2000. Reproducing Gender: Politics, Publics, and Everyday Life after Socialism. Princeton University Press, Princeton, NJ.

Gal, Susan, 2006. Introduction: guest editor. East European Politics and Societies 20 (7), 152–179.

Gray, Mary, 2009. Out in the Country: Youth, Media, and Queer Visibility in Rural America. New York University Press, New York.

Guyer, Jane I., 2004. Marginal Gains: Monetary Transactions in Atlantic Africa. University of Chicago Press, Chicago, IL.

Hirsch, Susan F., 1998. Pronouncing and Persevering: Gender and the Discourses of Disputing in an African Islamic Court. The University of Chicago Press, Chicago, IL.

Hoffman, Katherine E., 2008. We Share Walls: Language, Land, and Gender in Berber Morocco. Blackwell Publishing, Malden, MA.

Jacobs, Sue-Ellen, Thomas, Wesley, Lang, Sabine (Eds.), 1997. Two-Spirit People: Native American Gender Identity, Sexuality, and Spirituality. University of Illinois Press, Urbana, IL.

Joseph, Suad, 2000. Gender and Citizenship in the Middle East. Syracuse University Press, Syracuse, NY.

Joseph, Suad (Ed.), 2003–2008. Encyclopedia of Women and Islamic Cultures, vols. 6. Brill, Leiden, The Netherlands.

Kapchan, Deborah, 1996. Gender on the Market: Moroccan Women and the Revoicing of Tradition. University of Pennsylvania Press, Philadelphia, PA.

Kaplan, Flora (Ed.), 1997. Queens, Queen Mothers, Priestesses, and Power: Case Studies in African Gender. New York Academy of Sciences, New York, NY.

Lamphere, Louise (Ed.), 1997. Situated Lives: Gender and Culture in Everyday Life. Routledge, New York.

Lewin, Ellen, 2009. Gay Fatherhood: Narratives of Family and Citizenship in America. University of Chicago Press, Chicago, IL.

Liu, Fei-Wen, 2004. Literacy, gender, and class: nushu and sisterhood communities in southern rural Hunan. Nan NU: Men, Women, and Gender in Early Imperial China 6 (2), 241–281.

Lutkehaus, Nancy C., Roscoe, Paul B. (Eds.), 1995. Gender Rituals: Female Initiation in Melanesia. Routledge, New York, NY.

Masquelier, Adeline, 2005. Dirt, Undress, and Difference. Indiana University Press, Bloomington, IN.

McConnell-Ginet, Sally, 2011. Gender, Sexuality and Meaning: Linguistic Practice and Politics. Oxford University Press, New York.

Merry, Sally Engle, 2006. Human Rights and Gender Violence. The University of Chicago Press, Chicago, IL.

Mikell, Gwendolyn, 1997. African Feminism: The Politics of Survival in Sub-Saharan Africa. University of Pennsylvania Press, Philadelphia, PA.

Mitchell, William E. (Ed.), 1992. Clowning as Critical Practice. University of Pittsburgh Press, Pittsburgh, PA.

Moore, Henrietta, 2007. The Subject of Anthropology: Gender, Symbolism, and Psychoanalysis Polity. Polity Press, Cambridge, UK.

Najera-Ramirez, Olga, Cantu, Norma E. (Eds.), 2002. Chicana Traditions: Continuity and Change. University of Illinois Press, Urbana, IL.

Peletz, Michael G., 2009. Gender Pluralism in Southeast Asia. Routledge, New York, NY.

Piot, Charles, 1999. Remotely Global: Village Modernity in West Africa. University of Chicago Press, Chicago, IL.

Raheja, Gloria Goodwin, Gold, Ann Grodzins, 1994. Listen to the Heron's Words: Reimagining Gender and Kinship in North India. University of California Press, Berkeley, CA.

Rasmussen, Susan, 2006. Those Who Touch: Tuareg Medicine Women in Anthropological Perspective. Northern Illinois University Press, De Kalb, IL.

Salo, Elaine, 2006. Mans is ma soe: ganging practices in Manenberg, South Africa, and the ideologies of masculinity, gender, and generational relations. In: Bay, Edna,

Donham, Donald (Eds.), States of Violence: Politics, Youth, and Memory in Contemporary Africa. University of Virginia Press, Charlottesville, VA.
Seizer, Susan, 2005. Stigmas of the Tamil Stage. Duke University Press, Durham, NC.
Stoeltje, Beverly J., 2002. Women and language, and law in Africa I and II. Africa Today 49 (1 and 2) (special issue).
Stoeltje, Beverly J., 2009. Asante traditions and female self-assertion: sister Abena's narrative. Research in African Literatures 40 (1), 27–41.
Susser, Ida, 2009. AIDS, Sex, and Culture: Global Politics and Survival in Southern Africa. Wiley-Blackwell, Malden, MA. This page intentionally left blank

Gender: Gendering of Categories

Zorica Siročić, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term gender first appears in psychological discourse of the 1950s as a concept designed to overcome biological determinism. The constructionist approach of conceptualizing sex as a biological distinction and gender as a cultural category became central during the 1970s for the feminist aim of transforming disciplinary paradigms and power relations. The critique of ahistorical and universalizing accounts of gender influenced its several reformulations, most notably a deconstructive approach defining sex as a social construct. This article illustrates the manner in which disciplines were gendered, the most salient changes implemented, and the ways in which epistemic contestations were translated into practical analysis.

Introduction

The history of the nongrammatical use of the term gender starts with an article by psychologist and sexologist John Money from 1955 in which the concept of 'gender role' was introduced (Haig, 2004). Influenced by the Parsons' concept of the 'sex role' and Freudian psychoanalysis, Money defined 'gender role' as a term "used to signify all those things that a person says or does to disclose himself or herself as having the status of boy or man, girl or woman, respectively. It includes, but it is not restricted to sexuality in the sense of eroticism" (Money, 1955: 254). In this psychological study of hermaphroditic patients manifesting somatic 'ambisexual anomalies,' Money observed how the variables defining one's sex (including assigned sex and sex of rearing, external genital morphology, internal accessory reproductive structures, secondary sexual characteristics, as well as hormonal , gonadal, and chromosomal sex) relate to one's gender role and orientation as male or female. The study suggested as the prognosticator of a person's gender role, not gonadal and hormonal sex or internal accessory organs as expected, but assigned sex and sex of rearing. By concluding that one's orientation as male or female does not have an "innate, instinctive basis" (Money, 1955: p. 308) and that the dichotomy male/female needs to be reconsidered, the article surpassed the dominant hereditary doctrine and pointed out the strong influence of the social environment on sexual behavior.

Based on a survey of academic articles from 1945 to 2001, Haig (2004) found that authors using gender in the 1960s seem to derive the term from Money. These authors were mostly publishing in psychological journals and were dealing with phenomena that did not conform with the heteronormative matrix (hermaphrodites, transvestites, homosexuals, etc.). The concept was further developed by psychoanalyst Robert J. Stoller (1964) by endorsing the term 'gender identity,' which he defined as the "sense of knowing to which sex one belongs, that is, the awareness 'I am a male' or 'I am a female'" (Stoller, 1964: p. 220). Similar to other early writings on gender, these articles were influenced by the psychoanalytical school of thought, which has played a crucial role in suggesting that 'natural' attributes of sex are in fact the effect of physic representation of gender (Elam, 2000: p. 169).

However, the widespread use of the term gender and its rise in the social sciences is connected with the feminist adoption of the term. While Simone de Beauvoir already provided the theoretical framework for the sex/gender distinction in her 1949 classic *The Second Sex* (de Beauvoir, 1949), a closer elaboration of this relation was delayed until the 1970s (Millet, 1970; Oakley, 1972; Rubin, 1975). An analysis of articles in the Web of Science database, which address the topic of gender, reveals a steady rise across social science disciplines such as psychology, history, political science, sociology, and economics from the beginning of the 1980s (see **Figure 1**).

By the mid-1980s gender became the central category of the feminist thought and has influenced the formation of new study areas including women's; gender; men's; lesbian, gay, bisexual, and transgender (LGBT); and queer studies (as well as studies of masculinity/femininity).

Gender as a Central Category

Feminist thought of the 1970s and part of the 1980s predominantly conceptualized sex as biologically given in contrast to the culturally and socially constructed gender. Judith Squires (2000) labeled this a 'constructionist approach' to gender theory. This differentiation became extremely theoretically influential and politically significant because it enabled feminists to contest the domination of concepts such as 'sex' or 'sexual difference' that were closely related to 'biological determinism.' The latter signifies the belief that anatomy dictates the disposition and social role of the individual and that differences between women and men (as binary and opposing sexes) are biologically rooted and therefore immutable and resistant to change (Hawkesworth, 2013; Nicholson, 1994). The rejection of biological determinism and the accompanying constructionist perspective on gender have their origins in psychology and historically overlap with, and were partly empowered by, the broader epistemological shift to poststructuralism and criticisms of empiricism (Scott, 1986). The viewpoint of understanding gender as the social institutionalization of sexual difference (Moller Okin, 1998: p. 116) was at the time a crucial part of the critical and emancipatory idea that women's scholarship and activism could transform disciplinary paradigms

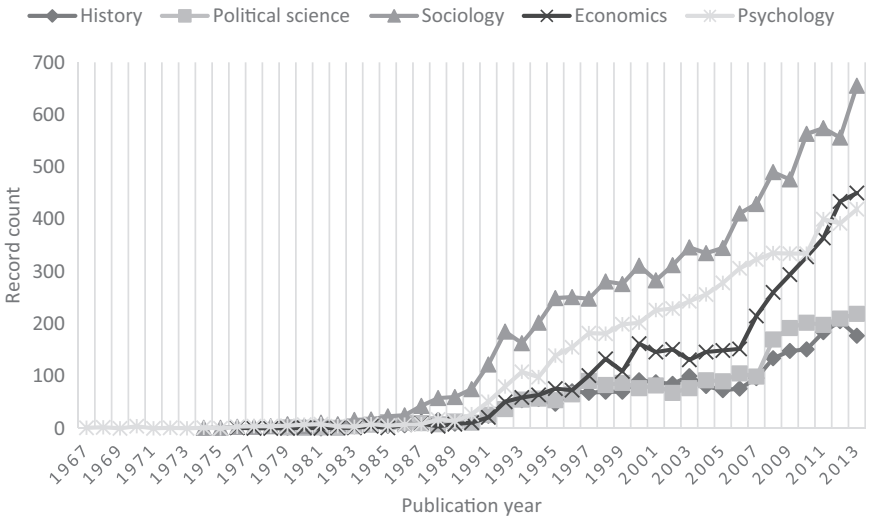

Figure 1 Trends in coverage of gender in social sciences.

(Scott, 1986) and power relations in society. In that sense, a thorough understanding of the beginnings and significance – as well as the discontents – of the constructionist approach to gender is inseparable from the study of feminist positions on scientific knowledge (*see* Feminist Research and Social Work; Situated Knowledge, Feminist and Science and Technology Studies Perspectives) and the history of women's movements and activism.

Structuralist studies in the feminist tradition from the 1970s reflected the need to analyze the mechanisms of social forces and power relations that construct gender identities. Presented by some of the landmark studies of second-wave feminist theory, each of them emphasized different aspects of social relations such as reproduction (Firestone, 1970; Rubin, 1975), production (Eisenstein, 1979), mothering and caring (Chodorow, 1978; Gilligan, 1982), or sexuality (MacKinnon, 1983; Millett, 1970) as the ones crucial for the reproduction and perpetuation of gender inequalities. Despite their influence, these readings were criticized, among other things, on the account of gender conceptualization. Firstly, the analyses relied on monocausal explanations of inequality (Moller Okin, 1998) while disregarding the effect of a broader socioeconomic and political context. Secondly, it produced ahistorical and universalizing accounts of gender (Scott, 1986) that were possible due to the widespread acceptance of 'biological foundationalism' – an understanding according to which gender is a matter of cultural specificity, while sex is treated as a cross-cultural and cross-historical commonality (Nicholson, 1994). In other words, the representatives of the constructionist approach took for granted the idea that sex is a stable, biological, i.e., materially given, category upon which gender as a matter of culture and character is constructed. Thirdly, issues of race, class, and ethnicity were neglected. This created a sense of false unity combined of otherwise conflicting and heterogeneous elements while masking the privileges of the white, middle-class, and heterosexual women (Hawkesworth, 2013).

This critique has influenced a twofold conceptual reformulation. Firstly, the scholars answered the objections of ahistorical and universalizing tendencies by providing complex and multifaceted accounts of gender and by incorporating intersecting impacts of race, ethnicity, class, sexuality, disability, etc. The latter was most prominently elaborated in the intersectionality theory (*see* Intersectionality), which through its multiperspectivity challenged the domination of the standpoint theory characteristic for the second-wave feminist scholarship (*see* Standpoint Theory, in Science). Yet, many of them seem to have preserved the assumption of the male/female gender binary. The literature refers in various terms to this conceptual move, as a shift from theorizing single accounts of social structures to multiple ones (Squires, 2000), a turn from difference feminism based on a belief in universal commonalities among women to diversity feminism (Dietz, 2003), or as a shift from identity to difference feminism.

The second part of the reformulation was analytically more challenging and the epistemic contestations it has caused persist until today. More precisely, under the influence of contributions that read genealogies of power and histories of sexuality differently (such as Foucault's (1976) and Laqueur's (1990)), combined with refined insights from literary theory, psychoanalysis, and postmodern theories, feminist theoreticians endorsed a new approach to sex and gender. This approach – whose most prominent advocate was Judith Butler (1990) – deconstructed the sex/gender binary, conceptualizing sex itself a social construct. Furthermore, the sex/gender relation based on a corresponding material/ideological framework is being reversed in the claim that gender performativity produces the sex (Butler, 1990). In a similar vein, Suzanne Kessler (1990), in her essay "Medical Construction of Gender," suggests the "backward" logic in sex/gender relation. When being confronted with the infants of ambiguous genital morphology, medical staff will be guided with the culturally conditioned image of gender in an attempt to attach the assumed 'natural' sex. Consequently, sex is being observed no longer as a biological category that precedes gender as a cultural category, but as a relational construction that requires a shift from a material to a linguistic, discursive (Squires, 2000), or symbolic understanding (Dietz, 2003).

The deconstructive turn, indicating the beginning of the so-called third-wave feminism, challenged the uniform existence of the subject of feminism, i.e., the 'Woman' suggesting instead multiplication and diversity of sex/gender combinations (Elam, 2000: p. 180), and called the (im)possibility of the political act into question. This has turned the direction of feminism and found its reflection in the split between feminist theory and feminist activism, within the theory itself (see Benhabib et al., 1995), and left the question of the appropriate methodology to operationalize these challenging theoretical insights open. Gradually, understandings of gender proliferated and theoretical approaches nowadays often reflect negotiations, as Mary Dietz (2003) labels it, between difference, diversity, and deconstruction feminism.

Gendering Categories

The majority of what was described up until now refers to fields where gender played a central role, such as feminist theory or women studies. At this point it is worth clarifying that all theory referring to gender is not necessarily feminist, while all feminist theory centers significantly or even exclusively on gender (Saltzman, 2006). The reason for the departure of gender studies from feminism can be found in the role that gender as analytical category played in gaining academic legitimacy of women's scholarship in the 1980s and the rise of popularity of gender studies within the past three decades. Instead of referring to 'women,' gender, similar to class or race, symbolized scientific neutrality and dissociation from politics of feminism (Scott, 1986). For that reason, it was accepted by those traditionally skeptical voices that perceived the activist and political agenda of feminism irreconcilable with scientific principles. Some of the consequences, as Scott (1986) and Lovenduski (1998) notice, are the often inconsistent and contradictory use of gender as a synonym for 'women' to ensure the political acceptability of the study and the fact that a great deal of gender studies no longer show explicit activist commitment to women's equality and social change, which causes predictable resentment by many feminist scholars.

Gendering categories, on the contrary, imply that the gender dimension was omitted in the 'classical' scientific disciplines resulting in hidden biases and distortions of standard conceptualizations. Accordingly, the process of gendering is required because research, methodology, and theory, which ignore gender as an analytical concept, are not good science (Goertz and Mazur, 2008: p. 6). This insight is particularly valuable because it overturns and contradicts the objection that feminism as political agenda does not have a place in value-free sciences. Gary Goertz and Amy Mazur (2008) define three distinct ways of gendering categories. The first way is to keep the name of the standard concept, but to change its substance by adding sex as an additional variable. Although many are skeptical of 'add-women-and-stir' approach, objecting that it has limited transformational potential, the addition of women in the regime classification of democracies or welfare states, for instance, shows the opposite. The second way of gendering categories is to 'adjectivize' the standard concept implying a new aspect to it, for instance, 'women's (social) movement.' The third way is to create a new concept as in the case of state feminism. As Goertz and Mazur (2008) rightfully warn, due to the changes in concepts, case selection can change and the results of previous analysis can be challenged. In other words, gendering has significant methodological and theoretical consequence that one needs to take into account.

Common paths and consequences of gendering categories across different disciplines encompass (at least) three levels. First, units of analysis were gendered, which in practice meant that women and women's experiences became a subject of inquiry. Second, the constructionist notion of sex/gender relation enabled a conceptual turn that facilitated the break with biological determinism and reductionism. Third, this has led to the proliferation of new areas of inquiry as well as to interdisciplinary approaches to methods and epistemologies. The paths through which gender as a category entered particular disciplines, and the ways in which theoretical contestations were translated into the analysis of some of the key concepts in disciplines of social and behavioral sciences, are elaborated below.

Psychology

The earliest contributions to the study of gender in psychology originated in the psychoanalytical tradition (Money, 1955; Stoller, 1964). The initial accounts of gender in psychology were critiqued on the similar basis to the studies of sexual difference, primarily for their ignorance and neglect of the social and political context of gender (Burr, 1998: p. 10). Although the impact of the social was recognized in some of the early studies (for instance, Money's previously described account of gender), the broader incorporation of the constructionist approach to gender in psychology that happened in the late 1960s and 1970s is most closely related with the second-wave feminist theory and activism (Chrisler and McCreary, 2010).

Starting with the Naomi Weisstein's (first published in 1963, reprinted in 1971) essay "Psychology Constructs the Female; or, The Fantasy Life of the Male Psychologist (With Some Attention to the Fantasies of His Friends, the Male Biologist and the Male Anthropologist)," feminist psychologists opened a range of questions and criticisms on the methodological and epistemological account of the discipline. Drawing on insights from feminist critique of science (Burr, 1998: p. 132; Chrisler and McCreary, 2010), they raised the awareness on how numerous studies used restricted samples in their research designs. The consequences were reported gender biases in theories of intelligence, learning, memory, thinking, identity, clinical diagnosis, and therapy (Crawford and Unger, 2000).

Additionally, the political role of psychology was questioned and its alleged neutrality challenged. By embracing the famous slogan 'the personal is political' feminist psychologists began to treat personal distress, not merely as a product of hormonal or other bodily disbalance, but as embedded in social, economic, and political inequalities (Worell, 2000). Consequently, the concept of gender enabled a shift from focusing on internal (individual) to external (structural) causes in psychological theory, research, and therapy.

The critique of early feminist works in psychology for providing monocausal, universalistic, and ahistorical accounts of gender (Scott, 1986; Nicholson, 1994) influenced the formation of the multicultural movement in psychology (Chrisler and McCreary, 2010) and integrative feminist psychology (Worell, 2000: p. 187). Nowadays, the predominant clashes are between essentialism and social constructionism (Worell, 2000), which correspond to the broader epistemological split related to diverging conceptualizations of gender.

Last, the concept of gender enabled the advent of new fields of psychological inquiry. The fields having gender as a central category include women's and feminist psychology, psychology of men and masculinity, and LGBT psychology. They originated in the 1970s and now have specialized organizations, conferences, and journals dedicated to these specific fields.

History

The first appearances of the term gender in academic historical discourse can be traced to early feminist writings that were primarily trying to establish narratives of women as legitimate voices within the discipline. Consequently, previously neglected areas of research such as private life, sexuality, and reproduction entered the area of historical inquiry. In the 1970s this was part of the idea that women's scholarship would transform the disciplinary paradigm and by including class, race, and gender would contribute to the creation of a 'new history' (including her-story) which would explain the organization of domination/subordination along these three axes. One of the classic texts for understating the role of gender in history is the essay *Gender: A Useful Category of Historical Analysis* by feminist historian Joan W. Scott (1986) in which she inaugurated a path-breaking debate in mainstream historical research (see Butler and Weed, 2011).

Based on an exhaustive overview of the available feminist scholarship, Scott argued that attempts to include gender, while remaining within traditional scientific frameworks, resulted in reductive or simplified generalizations that undermined the complexity of the analysis and stifled motivation for more substantive change. Furthermore, as gender was often used as a synonym for 'women' to assure scientific and political acceptability of the study, its analysis was restricted to areas traditionally related to women, leaving uncovered topics such as war, diplomacy, and high politics. Last, to the influential feminist studies of that time in the (radical) feminist, Marxist, and psychoanalytical traditions, Scott (1986) objected to the aforementioned ahistoricity of gender, to the reification of the sexual antagonism between men and women, as well as to the tendency to universalize the categories of male and female, which did not leave the space for historical specificity and variability. In order to develop gender as an analytical category, Scott, influenced by Foucault's genealogy of power and psychoanalysis, endorsed the refusal of the fixed and permanent quality of binary opposition and demanded a genuine historization and deconstruction of the terms of sexual difference in line with the criticism of Riley (1988) and Nicholson (1994). Almost 15 years later in her *Gender: Still a Useful Category of Analysis?*, Scott (2010) insists that gender is only useful when it goes beyond 'men' and 'women,' taken as fixed categories that only need to be described. Instead one needs to think critically about how these meanings are produced in relation to one another, how they are deployed and changed by focusing not on the roles assigned to women and men, but on the construction of sexual difference itself.

In conclusion, in the 1970s and the 1980s theoretical discussions on the category of gender went beyond the simple recognition of women's roles in history; they opened a range of questions about how, when, and why different roles have been assigned to sexes; how they varied across time and place; and under which circumstances were they enforced and constituted. Due to the criticism for ahistorical and universalizing accounts of gender, which implied a certain essence common to female identity regardless of time and place, the prevailing tendency in the contemporary mainstream gender and history is to incorporate deconstructive approaches while including multiple axes of specific sociohistorical circumstances that form gender identities.

Political Science

Drude Dahlerup (2010) reports the first use of gender within political science in the 1980s. Having in mind that the term dates back to the 1950s, the postponement reveals a lot about the discipline. Since the mid-1950s political science aspired to be established as an objective and value-neutral discipline adhering to behaviorist and positivistic principles. Within such a context it was first necessary (and possible) to introduce the variable of sex into the research. The concept of gender, on the other hand, was more difficult to incorporate into the research design. The reluctance to incorporate postmodern theories, as well as the preference of parsimonious models for hypothesis testing explains the discipline's animosity to the study of gender (Lovenduski, 1998: p. 352).

Nevertheless, gender entered the discipline due to the efforts of feminist scholars who recognized the practical and theoretical power inherent in the constructionist perspective to the gender/sex relation for dismantling some of the basic postulates of the discipline such as public/private divide.

According to Celis et al. (2013) the representative approaches of incorporating the conceptualization of gender into political scholarship can be systematized in the following ways. The first approach is based on gendering the classical units of analysis such as the citizen, the voter, parties, or states. In particular, sex is added as an additional variable or the analysis examines women as an afterthought (Goertz and Mazur, 2008), while asking traditional questions and using traditional methods. Characteristic for this direction is the strategy of inclusion defined by Squires (2000) as the one based on constructionist approach that aims to achieve equality in the political sphere by denaturalizing the presumption of sexual difference and eroding the link between sex and gender so that sex becomes irrelevant for the political engagement.

The second direction according to Celis et al. (2013) is the one focusing on arenas traditionally seen as being outside the domain of political science. Here the notion of 'the political' is being broadened and by studying women in their own terms, new areas of study are being introduced. According to Squires (2000), this approach predominantly uses the strategy of

reversal – one based on the constructionist approach to gender, but emphasizing how the form of engagement was influenced by gender while recognizing and revalorizing women's engagement.

The third approach is oriented toward the analysis of gender as a structure of social organization – study of gender regimes, gendered states, and institutions focusing on broader social patterns that reproduce gender inequalities (Celis et al., 2013). It questions not only the ways in which structures and institutions are gendered but also how the gendered subjects of politics are being produced (production of masculinities, femininities, and sexualities). Distinctive for this approach is the strategy of displacement (Squires, 2000) – based on the deconstructive approach to gender. As this directly touches the question of subjectivity as well as political and collective action, the epistemic contestations are translated and pertain to a number of theories and issues crucial for political science such as citizenship, representation, multiculturalism, democratic discourse, and political action (Dietz, 2003).

Sociology

The use of the term gender in sociology seems to have been pioneered by Ann Oakley in the early 1970s. By applying ideas about socialization, she sought to understand how gender is learned and how femininity and masculinity are socially constructed (Holmes, 2007: p. 42). In her later work on the sociology of housework, Oakley extended the theories on gender in order to explain the division of labor in the household and labor market segregation.

When providing systematization of gender in sociology, several conceptual levels are possible. This variety reflects, on the one hand, the discipline's multiparadigmatic character and, on the other hand, the theoretically challenging intersection of biological, social, psychological, and economical components inherent to the concept of gender. This explains why a significant number of the available overviews of gender in sociology provide either a general history of the concept or a history of feminist thought, which are connected, but still separate issues.

Amy Wharton's systematization (2005) of dominant frameworks within sociology to the study of gender as individualist, interactional, and institutional seems to reflectively circumvent this common slip, without reducing the complexity intrinsic to the matter. As one of the organizing principles of the social world, gender shapes the identities and behavioral dispositions of individuals, influences social interaction, and organizes social institutions (Wharton, 2005: p. 9).

Sociological scholarship on gender based on an individualistic approach would include, according to Wharton (2005), biosocial and evolutionary psychological perspectives, as well as theory and research on gender socialization and sexual differences that can be found both in psychology and in sociology. Although generally adopting a constructionist approach to gender, authors in this tradition conceptualized sex to be the basis for the development of gender that was treated as a stable set of personality characteristics.

A perspective challenging these assumptions offered an interaction-based perspective on gender instead. The seminal article representing this approach is "Doing Gender" by Candace West and Don H. Zimmerman from 1987. Largely inspired by the work of Erving Goffman, the authors approached the study of gender from an ethnomethodological tradition. West and Zimmerman (1987: p. 125) defined gender as a "routine accomplishment embedded in everyday interaction." In this way, the distinction between sex as a stable biologically given category (nature) and gender as culturally constructed (nurture) (*see* Nature/Nurture Controversy, History of) is openly contested, announcing the arrival of deconstructivist approach to gender in sociology. Besides the ethnomethodological tradition, interaction-based approach to gender is evident in status characteristics theory as well as in theory and research on homophily (Wharton, 2005).

The third approach presents the institutional perspective to gender. It encompasses a broad area of theory and research that treats gender as an organizing principle of social, economic, and cultural life. Accordingly, the practices and policies of organizations such as education systems, labor systems, health systems, work, sports, or family, to name only a few, are being studied.

Besides the well-developed women's studies, women's sociology, and feminist theory (see Smith, 1987), other developments within sociology related to gender include the development of men's studies and masculinity studies presented by Raewyn Connell's (1995) classic "Masculinities," as well as attempts to take into account variations within gender as a category and to accommodate complex intersections of gender with other categories.

Economy

A gendered perspective on economic processes reflected in the feminist critique of economy dates back to the nineteenth century (Bauhardt and Çağlar, 2010) and was present in feminist scholarship across various disciplines in the social sciences from 1970. However, feminist economics as a research direction within economics was institutionalized only in the early 1990s. Since then, the progress of the field and an overview of the major topics can be found in the classics of feminist economics *Beyond Economic Man: Feminist Theory and Economic* (1993) and *Feminist Economics Today: Beyond Economic Man* (2003) edited by Marianne A. Ferber and Julie A. Nelson. This direction is not a unified body of thought, as it was influenced by and stretches across all major streams of economic thought including neoclassical, Marxist, and institutionalist traditions.

The example of economics is especially illustrative of how far reaching the implications of gendering categories can be. In the 1970s under the influence of Marxism and socialism, second-wave feminists started questioning the role of the patriarchy within capitalist systems. By focusing on the category of work, they pointed out the significance of unpaid labor, done predominantly by women, for the maintenance of economic systems. Being perceived as marginal, private, and unproductive both by the neoclassical and by the Marxist tradition, unpaid housework was omitted from economic analysis. To illuminate its multidimensionality and above all gender-segregated nature, feminists adopted the notion of

social reproduction to cover biological reproduction and unpaid production at home (household maintenance, care work). The unjust distribution of the burden of social reproduction provoked the movement 'Wages for Housework' taking place in the mid-1970s in several countries of Europe and North America.

These insights disclosed the ways in which essentializing accounts of gender do not remain in the cultural and symbolic sphere, but are reflected in material consequences. More precisely, culturally defined norms determine the gender division of labor in the sphere of paid work (vertical and horizontal segregation) and between paid and unpaid work (women being disproportionally burdened by the housework). Where women's mass entrance on the labor market is concerned, gender-specific exploitation was reflected in disproportionally higher percentages of women in low-skill and low-paid occupations because of women's 'natural' capabilities for the job's performance implying no educational investment (Elson and Pearson, 1981). In contemporary labor relations, concepts such as the 'gender pay gap' and 'glass ceiling' indicate the globally documented difference in the earnings, i.e., positions acquired, and women continue to be more represented in outworking, contract and part-time labor, homework, and other forms of labor unprotected by labor regulations.

Conclusion

The first conceptualizations of gender as a socially and culturally constructed category have their roots in psychology from the mid-1950s. However, they became prominent mostly due to feminist scholars and activists from the 1970s. Additionally, the incorporation of gender as category into all major social and behavioral sciences was possible due to the broader poststructuralist turn, which influenced the characteristic feminist epistemological positioning and methodological framework. As they are historically varied, there are significant differences in the above-mentioned approaches to gender (biological determinism, constructionist (multi- or single structure), and deconstructive approach). Furthermore, favoring one or the other approach to gender differs across disciplines. For instance, while history or sociology seemed to be open toward deconstructive conceptualizations, others are more reserved and skeptical, such as political science, or appear to circumvent the issue, such as economics.

Irrespective of the degree of acceptance of the particular approach, based on this brief historical overview of the term's origins, it is possible to summarize its major contributions and ranges of significance. First, studies of gender, gender identities, and gender relations caused not only the proliferation of previously neglected areas of scientific inquiry within existing disciplines but also triggered the advent of new disciplines such as women's, gender, men's, LGBT, and queer studies. Second, the inclusion of gender as an analytical category and the gendering of scientific categories exposed the bias of scientific research that produced distorted results typically detrimental for women and thereby added to the broader debate on problems in the conduct of the social research. Consequently, women's experiences and voices were introduced into analysis, while gender contributed to the conceptual break with biological determinism and reductionism. At the same time, new areas of inquiry proliferated, which have challenged the epistemologies and methodologies of the scientific canon. Nevertheless, the reluctance on the part of academic disciplines to mainstream gender analysis, the irreconcilability of some parts of gender theory with feminist activism, as well as the question of the appropriate methodology to operationalize seemingly challenging theoretical insights are all issues that remain unsolved.

See also: Feminism in Anthropology; Feminist Criminology and Gender Studies; Feminist Legal Theory; Feminist Political Theory and Political Science; Feminist Psychotherapy; Gender History; Gender and Feminist Studies in Geography; Gender and Social Research; Gender and Women's Studies, Applied Research on; Gender and the Law; Gender, Economics of; Gender, Feminism, and Sexuality in Archaeological Studies; Gender: Anthropological Aspects; Gendered Care and Social Work Practice; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Bauhardt, C., Çağlar, H. (Eds.), 2010. Gender and Economics. Feministische Kritik der Politischen Ökonomie. VS Verlag für Sozialwissenschaften, Wiesbaden.
Benhabib, S., et al., 1995. Feminist Contentions: A Philosophical Exchange. Routledge, New York.
Burr, V., 1998. Gender and Social Psychology. Psychology Press, Routledge, London and New York.
Butler, J., 1990. Gender Trouble: Feminism and the Subversion of Identity. Routledge, New York.
Butler, J., Weed, E., 2011. The Question of Gender. Joan W. Scott's Critical Feminism. Indiana University Press, Bloomington.
de Beauvoir, S., 1972(1949). The Second Sex (H.M. Parshley, Trans.). Penguin, New York.
Celis, K., et al., 2013. Gender and politics: a gendered world, a gendered discipline. In: Waylen, G., et al. (Eds.), The Oxford Handbook of Gender and Politics. Oxford University Press, New York.
Chodorow, N., 1978. The Reproduction of Mothering. University of California Press, Berkeley.
Chrisler, J.C., McCreary, D.R. (Eds.), 2010. Handbook of Gender Research in Psychology, vol. 1 and 2. Springer, New York.
Connell, R., 1995. Masculinities. University of California Press, Berkeley.
Crawford, M., Unger, R., 2000. Women and Gender: A Feminist Psychology, third ed. McGraw-Hill, New York.
Dahlerup, D., 2010. The development of gender and politics as a new research field within the framework of the ECPR. European Political Science 9, 85.
Dietz, M., 2003. Current controversies in feminist theory. Annual Review of Political Science 6, 399–431.
Eisenstein, Z. (Ed.), 1979. Capitalist Patriarchy and the Case for Socialist Feminism. Monthly Review Press, New York.
Elam, D., 2000. Gender or sex. In: Tripp, A. (Ed.), Gender (Readers in Cultural Criticism). Palgrave Macmillan, pp. 168–182.
Elson, D., Pearson, R., 1981. 'Nimble fingers make cheap workers': an analysis of women's employment in third world export manufacturing. Feminist Review 7, 87–107.
Ferber, M.A., Nelson, Julie A. (Eds.), 1993. Beyond Economic Man. Feminist Theory and Economics. University of Chicago Press, Chicago.
Ferber, M.A., Nelson, Julie A. (Eds.), 2003. Feminist Economics Today. Beyond Economic Man. University of Chicago Press, Chicago.
Firestone, S., 1970. The Dialectic of Sex: The Case for Feminist Revolution. William Morrow and Company, New York.

Foucault, M., 1976. The History of Sexuality (R. Hurley, Trans.). In: An Introduction, vol. 1. Vintage, New York.
Gilligan, C., 1982. In a Different Voice: Psychological Theory and Women's Development. Harvard University Press, Cambridge, MA.
Goertz, G., Mazur, A., 2008. Politics, Gender, and Concepts. Theory and Methodology. Cambridge University Press, New York.
Haig, D., 2004. The inexorable rise of gender and the decline of sex: social change in academic titles, 1945–2001. Archives of Sexual Behavior 33 (2), 87–96.
Hawkesworth, M., 2013. Sex, gender, and sexuality: from naturalized presumption to analytical categories. In: Waylen, G., et al. (Eds.), The Oxford Handbook of Gender and Politics. Oxford University Press, New York.
Holmes, M., 2007. What Is Gender? Sociological Approaches. SAGE Publications Ltd, London.
Kessler, S.J., 1990. The medical construction of gender: case management of intersexed infants. Signs 16 (1), 3–26.
Laqueur, T., 1990. Making Sex: Body and Gender from the Greeks to Freud. Harvard University Press, Cambridge, MA; London.
Lovenduski, J., 1998. Gendering research in political science. Annual Review of Political Science 1, 333–356.
MacKinnon, C., 1983. Feminism, marxism, method and the state: an agenda for theory. In: Abel, E., Abel, E.K. (Eds.), Signs Reader: Women, Gender and Scholarship. Chicago Press, Chicago, pp. 227–256.
Millett, K., 1970. Sexual Politics. Doubleday, New York.
Money, J., 1955. Hermaphroditism, gender and precocity in hyperadrenocorticism: psychologic findings. Bulletin of the Johns Hopkins Hospital 96, 253–264.
Nicholson, L., 1994. Interpreting gender. Signs: Journal of Women in Culture and Society 20, 79–105.
Oakley, A., 1972. Sex, Gender, and Society. Harper Colophon, New York.
Okin, S.M., 1998. Gender, the public and the private. In: Phillips, A. (Ed.), Feminism and Politics. Oxford University Press, New York.
Riley, D., 1988. Am I that Name?: Feminism and the Category of "Women" in History. Macmillan, London.
Rubin, G., 1975. The traffic in women: notes on the "political economy" of sex. In: Reiter, R.R. (Ed.), Toward an Anthropology of Women. Monthly Review Press, New York.
Saltzman, C.J., 2006. Handbook of the Sociology of Gender. Springer Science and Business Media.
Scott, J.W., 1986. Gender: a useful category of historical analysis. The American Historical Review 91 (5), 1053–1075.
Scott, J.W., 2010. Gender: still a useful category of analysis? Diogenes 57, 7.
Smith, D.E., 1987. The Everyday World as Problematic: A Feminist Sociology. UPNE, Lebanon, NH.
Squires, J., 2000. Gender in Political Theory. Polity Press, Cambridge.
Stoller, R.J., 1964. A contribution to the study of gender identity. International Journal of Psychoanalysis 45, 220–226.
Weisstein, N., 1971. Psychology Constructs the Female, or, the Fantasy Life of the Male Psychologist: With Some Attention to the Fantasies of His Friends, the Male Biologist and the Male Anthropologist. New England Free Press, Boston.
West, C., Zimmerman, D., 1987. Doing gender. Gender & Society 1 (2), 125–151.
Wharton, A.S., 2005. The Sociology of Gender: An Introduction to Theory and Research, second ed. Wiley-Blackwell Publishers, Malden.
Worell, J., 2000. Feminism in psychology: revolution or evolution? Annals AAPSS 571.

Gender: Grammatical

Greville G Corbett, University of Surrey, Guildford, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 9, pp. 6335–6340, © 2001, Elsevier Ltd.

Abstract

The term 'gender' requires discussion, since different linguistic traditions describe gender phenomena using different terms. Moreover, while in many languages there is no dispute as to the number of genders, there are other languages where the question is far from straightforward. This requires consideration of the analytical problem of determining the number of genders in a given language. The central problem, however, is gender assignment, that is, the way in which the native speaker allots nouns to genders. Examination of languages from different families reveals that genders always have a semantic core, which may be biological sex, or animacy, with other features also having a role. In some languages nouns are assigned to genders solely on the basis of semantics, but in others this semantic information is supplemented by formal information, which may be morphological or phonological.

Gender is a remarkable category, pervasive in some languages, for instance in the Indo-European and Dravidian families, and lacking in others, as in the Uralic family. We start with a brief consideration of terms, then examine how the number of genders in a given language can be established. The number varies considerably: two and three genders are often found, four and five are not unusual, while Fula (a Niger-Kordofanian language) has around 20. Most attention will be given to the problem of how the speaker assigns nouns to genders. Gender systems may have sex as a component, as in languages with masculine and feminine genders; but equally sex may be irrelevant, as in the Algonquian languages where the distinction is between animate and inanimate. Finally we consider prospects for investigating the category further.

Terms

The defining characteristic of gender is agreement; a language has a gender system only if noun phrases headed by nouns of different types control different agreements. The evidence that nouns have gender in a given language thus lies outside the nouns themselves. There is no substantive difference between 'genders' and 'noun classes'; the different terms are merely the products of different linguistic traditions. For instance, we find systems with three genders, to which nouns are assigned by similar rules, in both Kannada (a Dravidian language) and in Godoberi (a Nakh-Daghestanian language). By tradition the first is said to have three genders, and the second three noun classes. We shall use the term 'gender' in both instances.

The Number of Genders

The approach to gender sketched in Sect. 1 is based on Zaliznjak's (1964) notion of 'agreement class.' Two nouns are in the same agreement class only if they take the same agreements under all conditions, that is if we control for other grammatical categories such as case and number. Conversely, if two nouns differ in their agreements when factors such as case and number are held constant, then they belong to two different agreement classes and it will normally be the case that they belong to two different genders. While in many languages there is no dispute as to the number of genders, there are other languages where the question is far from straightforward. In the case of Romanian there has been a lengthy debate. The analytical problem of determining the number of genders and the tests for deciding the gender of a given noun depends on separating out the classes into which nouns are divided (the controller genders) from the number of different genders marked on agreement targets (the target genders). Frequently the two match up, but in several languages they do not. A full treatment of the subject is not possible here; for a detailed account and extensive references see Corbett (1991).

Gender Assignment

We now ask how nouns are distributed over the genders of a given language, in other words, how does a native speaker 'know' what gender a noun belong to? Models of the mechanisms by which nouns are allotted to genders are called 'assignment systems.' Assignment may involve two sorts of information about the noun: its meaning and its form. We shall consider semantic systems first.

Strict Semantic Assignment Systems

This type is found in Dravidian languages like Kannada, spoken in the state of Karnataka, southern India (Sridhar, 1990, p. 198). Nouns denoting male humans are masculine, those denoting female humans are feminine. There are also deities, demons, and heavenly bodies in these genders. All remaining nouns, including those denoting infants and animals, are neuter. Thus *appa* 'father,' and *candra* 'moon' are masculine, *amma* 'mother' is feminine, and *na:yi* 'dog' is neuter. Certain other Dravidian languages like Kolami have only two genders: nouns denoting male humans are masculine and all others belong in the nonmasculine gender (Emeneau, 1955, p. 73). Elsewhere we also find the reverse. Thus in Diyari, a language of South Australia, there is a gender for nouns with female referents (such as women, girls, doe kangaroos),

and the other is for all remaining nouns (Austin, 1981). In Alamblak, a Sepik Hill language of Papua New Guinea, the masculine gender includes nouns denoting males and those denoting things like crocodiles, pythons, and arrows, which are long and thin, while the feminine is for nouns denoting females or short, squat items like turtles, frogs, and chairs (Bruce, 1984, pp. 96–98; Foley, 1986, pp. 80–81). An interesting system with some similar features is found in Taiap (Kulick and Stroud, 1992).

Predominantly Semantic Assignment Systems

Many languages have semantic assignment rules, which, however, lack the coverage of those seen in Kannada. An example of what we shall call predominantly semantic assignment is found in Tsakhur, which is a member of the Lezgic group of the Nakh-Daghestanian family. Its four genders are shown in **Table 1**.

Assignment to genders I and II is straightforward: I is for male humans (also gods, angels, and so on) while II is for female humans (and female mythical beings). Most remaining animates are assigned to gender III, though a few belong to gender IV, along with some mythical beings. Inanimates are found in genders III and IV (Melínikov and Kurbanov, 1964; Ibragimov, 1999, pp. 54–6; and references therein, Kibrik, 1999, pp. 48–9).

Languages of this general type are widespread. For some of them, researchers have proposed that abstract semantic criteria partly miss the point, and that if we can gain a better grasp of the world view of the speakers, we can then understand the assignment system. The most discussed case is that of Dyirbal, a language of North Queensland (Dixon, 1972), which has four genders, primarily for: male humans and nonhuman animates (I), female humans (II), non-flesh food (III) and the residue (IV). There are numerous apparent exceptions: the moon is in the first, masculine gender and the sun is the second, feminine gender. However, in Dyirbal mythology, as indeed in much of Australia, the moon is the husband of the sun; in Dyirbal the mythological significance of referents is a key factor in assignment. Dixon's data and analysis continue to provoke debate, as shown by the recent paper by Polinsky (1998). There is similar debate as the role of world view in the gender assignment system of Ojibwa (and in other Algonquian language of southern Canada and northern USA).

Table 1 Gender assignment in Tsakhur

Criterion	*Gender*	*Examples*	*Glosses*
male rational	I (masculine)	dak$_j$	father
		gade	boy
female rational	II (feminine)	jed$_j$	mother
		sos	bride
animate (also some inanimates)	III (animate)	balkan	horse
		dadal	cock
		(qow)	(nose)
other (including a few animates)	IV (neuter)	t'et'	flower
		niīse	cheese
		(kabaj)	(butterfly)

In all the languages discussed so far, the meaning of the noun determines gender. In the strict assignment systems, the rules are immediately obvious; in the predominantly semantic systems, there may be a minority of exceptions, but these exceptions have been claimed to be largely apparent in some languages, once the cultural setting of the language is taken into account. If we ask which are the semantic criteria on which semantic systems can be based, we see recurring patterns. Quite often we find animate ~ inanimate, human ~ nonhuman, and male ~ female. Sometimes there is a gender for diminutives, as in various Bantu languages. There are also less usual genders, such as that for nonflesh food (Dyirbal) and the gender for insects (found in the Rikvani dialect of the Nakh-Daghestanian language Andi). A criterion which is sufficient to define a gender in one language may be just one factor in the assignment to a gender in another. Thus the Bantu language Chichewa has a gender for diminutives, while in Dizi, diminutives together with nouns denoting females form a gender.

Formal Assignment Systems

In many languages, assignment by semantic rules would fail to assign many nouns to a gender. While in languages like Kannada the nouns not assigned by the semantic rules (the 'remainder' or 'semantic residue'), all belong to a single gender, in many languages they are spread over more than one gender. In such languages we find additional rules for assigning nouns to genders according to their form. Note the asymmetry: languages may use semantic rules, or semantic *and* formal rules, but not just formal assignment rules. In no language are nouns assigned to genders, as defined earlier, by purely formal rules. An example would be a language in which there were two agreement classes, and the nouns in the first all ended in a consonant cluster, and those in the second did not, and there was no semantic regularity for the distribution of nouns. I claim that this hypothetical type does not exist. Formal assignment rules may appeal to two types of information: phonological and morphological.

Phonological

A clear example of assignment depending on phonological information is provided by Qafar (Afar), an East Cushitic language spoken in north-eastern Ethiopia and in Djibouti (data from Parker and Hayward, 1985). Qafar has rather standard semantic assignment rules, namely that for sex-differentiable nouns, those denoting males are masculine and those denoting females are feminine. It is the nouns which fall outside these rules, the residue, which are of interest. For them the phonological rules apply: nouns whose citation form ends in an accented vowel are feminine (for example, *karmà* 'autumn') while all others are masculine (for example, *gilàl* 'winter,' which does not end in a vowel, and *tàmu* 'taste,' which does end in a vowel, but not an accented one). These rules operate with few exceptions. Moreover, nouns denoting males and females typically fit them too (for example *bàqla* 'husband' and *barrà* 'woman, wife'). It might seem that we could dispense with semantic rules for Qafar, and treat it as being of a quite different type. However, while the phonological rule works for the vast majority of cases, we find the crucial example *abbà* 'father,' which is masculine, even though it ends in an accented

vowel, which 'should' be an indicator of feminine gender. Conversely, *gabbixeèra* 'slender-waisted female' is feminine, though the accent is nonfinal. Qafar is remarkably for the simplicity of its phonological assignment rules, and for the fact that when the semantic and phonological rules are both applicable, they almost always give the same result. However, in cases of conflict, the semantic rules take precedence (as is the normal situation in gender assignment systems).

Other examples can occur in various parts of the world. Even French, often claimed to have no system to its gender assignment, has been demonstrated to have a phonological system, though not as simple as that of Qafar. For example, of 938 nouns ending in $\tilde{\varepsilon}$, 99% are masculine (like *le pain* [p $\tilde{\varepsilon}$] 'the bread'). And of 1,453 nouns in/3/, 94.2% are masculine (like *le ménage* [mena:3] 'the household' (for details see Tucker et al., 1977).

Morphological

In Russian, which is typical of many other Indo-European languages, nouns denoting male humans are masculine and those denoting female humans are feminine. But unlike the situation in languages like Kannada, it is not the case that the semantic residue is assigned to the neuter gender; in Russian the residue is shared between the three genders, with the neuter gender not even receiving the majority. Searching for additional semantic criteria is not at all promising, as the following data suggests (Table 2).

If we look instead at the morphology of the nouns, then we can make progress. There are four main inflectional classes in Russian, each with several thousands of nouns (for justification of this not uncontroversial view see Corbett, 1982, pp. 202–11). There are six cases and two numbers (though no paradigm has 12 distinct forms because of various syncretisms). We give just the singular forms in Table 3.

The native speaker needs to know how a noun inflects, in order to produce grammatical utterances. On the basis of that information, the assignment rules are straightforward. Nouns in class I are masculine, those in classes II and III are feminine, and those in IV are neuter. At first we might think that the semantic assignment rules are superfluous, since *mal'čik* 'boy' is in class I, while *devuška* 'girl' is in class II, and *mat'* 'mother' is in class III. In other words, many of the sex-differentiable nouns would be assigned to the appropriate gender by the morphological assignment rules. But there are also instances where this is not so, for instance, *djadja* 'uncle,' which denotes a male but is in class II, whose nouns are typically feminine. *Djadja* 'uncle' is masculine. Nouns like this show, once again, that we do not find languages where formal assignment rules are sufficient. Further rules are required in Russian for indeclinable nouns (like *taksi* 'taxi,' which is indeclinable and neuter). Morphological assignment systems are found in various other Indo-European languages. Looking further a field, in Arabic too, gender is assignable in the main according to morphology (Cowell, 1964, pp. 372–5).

The different types of assignment criteria may overlap in ways which make it difficult to establish their relative weight in a given language. Thus there may be small clusters of nouns which can be accounted for by semantic criteria (apart from those covered by the main semantic rules) even within systems where formal rules have a major role (as shown for German by Zubin and Köpcke, 1983, Köpcke and Zubin (1996). However, whenever gender languages are analyzed in sufficient detail, the gender of the vast majority of nouns turns out to be predictable.

Table 2 Russian nouns belonging to the semantic residue

Masculine	*Feminine*	*Neuter*
žurnal 'magazine'	gazeta 'newspaper'	pis'mo 'letter'
dom 'house'	izba 'hut'	zdanie 'building'
čaj 'tea'	voda 'water'	pivo 'beer'
dub 'oak'	sosna 'pine'	derevo 'tree'
ogon 'fire'	peč' 'stove'	plamja 'flame'
avtomobil' 'car'	maš ina 'car'	taksi 'taxi'
stul 'chair'	taburetka 'stool'	kreslo 'armchair'
den' 'day'	noč' 'night'	utro 'morning'
čas 'hour'	minuta 'minute'	vremja 'time'
nerv 'nerve'	kost' 'bone'	serdce 'heart'
nos 'nose'	ščeka 'check'	uxo 'ear'
lokot 'elbow'	lodyžka 'ankle'	koleno 'knee'
flag 'flag'	èmblema 'emblem'	znamja 'banner'

Table 3 Inflectional classes in Russian

	I	*II*	*III*	*IV*
NOMINATIVE	žurnal	gazeta	kost'	pis'mo
ACCUSATIVE	žurnal	gazetu	kost'	pis'mo
GENITIVE	žurnala	gazety	kosti	pis'ma
DATIVE	žurnalu	gazete	kosti	pis'mu
INSTRUMENTAL	žurnalom	gazetoj	kost'ju	pis'mom
LOCATIVE	žurnale	gazete	kosti	pis'me
gloss	'magazine'	'newspaper'	'bone'	'letter'

Prospects

Studies on grammatical gender continue to appear regularly, including valuable studies of previously unresearched languages. Here we highlight two main areas where we can expect substantial new developments: lexicology and modeling.

Lexicology

Gender is a key category for lexicology, since it is crucial for understanding the nature of lexical entries. In a gender language the gender of a noun must be available. On the other hand, as we have seen, it does not normally need to be specified, since the information can be derived, by assignment rules, from other information which must be stored. We still do not know the full range of information which may be relevant (for instance, the possible semantic features which may determine gender assignment).

There are further surprises: it is possible too for certain nouns to be of different genders according to the sex of the speaker; this is found in Garifuna, a member of the Arawak family spoken in Belize, Honduras, Nicaragua, and Guatemala (Taylor, 1977, p. 60; Munro, 1998). These two points raise interesting questions as to the shape of lexical entries,

a general question which is of concern to lexicologists and increasingly to psycholinguists (see, for instance, Berkum (1996).

Modeling

When discussing some types of assignment system, the evidence is so clear-cut that alternative analyses are hard to imagine. However, with morphological assignment systems, particularly when the number of genders and the number of inflectional classes is similar in a given language, there are other possible analyses, in particular some would specify gender in the lexical entry and derive inflectional information from it. Arguments in favour of the approach taken in Sect. 3.3.2 include the following; first, and primarily, it is claimed that in languages like Russian the internal evidence shows that the gender from inflectional class approach is preferable (it proves impossible to predict inflectional class from gender but it is possible to predict gender from inflectional class); and second, there is the typological claim that one or other of the possible types of assignment system (or some combination) will always be in play in a gender language such that gender is predictable for the vast majority of nouns and so if the Russian type system were treated differently this would create an anomaly.

However, there is a further step which we can take, of a rather different nature. Since the Russian type languages are crucial, it worth investing additional time in clarifying them. Hence in Fraser and Corbett (1995) we give an account of the Russian gender and inflectional class relations in the Network Morphology framework. The analysis is implemented in the lexical knowledge representation language DATR (Evans and Gazdar (1996, Gazdar forthcoming), which means that the analyses can be verified by computer. Given minimal lexical entries for nouns our account does indeed predict the right genders for a wide range of nouns. An account like this which is both explicit and testable is not available for the alternative approach.

A more complex type is that in which gender is again assigned according to morphological class, but where the morphological class is itself predictable. One such case is Arapesh; a language of the Torricelli family, spoken on the north coast of Papua New Guinea. Genders and morphological classes form a substantial part of Fortune's (1942) grammar, which led to work based on it by Aronoff (1992, 1994, pp. 89–114). Once again, there are semantic assignment rules based on sex. But then there are morphological assignments rules, which determine 13 genders on the basis of 22 morphological classes (which are in turn predictable largely from phonological information). Again we have demonstrated how this system can be analyzed within the Network Morphology framework and have provided an implementation to demonstrate that our claims are indeed valid (Fraser and Corbett (1997). There is a complex interaction of assignment to gender and to morphological class in Mayali, a non-Pama-Nyungan language of north Australia, which has four genders and five morphological classes for nouns. For the essential data see Evans (1997) and for details of a formal model of this system see Evans et al. (2001). Finally, there is interesting work on the development of assignment systems over time, including some modelling: see Comrie and Polinsky (1998), Polinsky and Everbroeck (1998), Polinsky and Jackson (1999).

Conclusion

Gender remains a puzzling category, apparently deeply embedded into the structure of some languages, and missing in other. This is found even in related languages: thus several Indo-Aryan languages have two or three genders, with substantial agreement, while another, Bengali, has no gender even in the personal pronouns. Where we find gender, the assignment rules can be simple, as in Kannada and Qafar, or complex as in French and German. It is an area in which gradual progress is being made and one in which there is still much to learn.

See also: Grammatical Agreement; Number, Grammatical.

Bibliography

Aronoff Mark, 1992. Noun classes in Arapesh. In: Booij, G., van Marie, J. (Eds.), Yearbook of Morphology 1991. Kluwer, Dordrecht, The Netherlands, pp. 21–32.

Aronoff, M., 1994. Morphology By Itself: Stems and Inflectional Classes (Linguistic Inquiry Monograph 22). MIT Press, Cambridge, MA.

Austin, P., 1981. A Grammar of Diyari, South Australia. Cambridge University Press, Cambridge, UK.

Berkum, J., van, J.A., 1996. The psycholinguistics of grammatical gender: Studies in language comprehension and production. Ph.D. dissertation. Nijmegen University Press, Nijmegen, The Netherlands.

Bruce, L., 1984. The Alamblak Language of Papua New Guinea (East Sepik). Department of Linguistics. Research School of Pacific Studies, Australian National University, Canberra, ACT.

Comrie, B., Polinsky, M., 1998. Gender in a historical perspective: Radial categories meet language change. In: Justus, C.F., Polome, E.C. (Eds.), Language Change and Typological Variation: In Honor of Winfred P. Lehmann on the Occasion of His 83rd Birthday: II Grammatical Universals and Typology. Institute for the Study of Man, Washington, DC, pp. 566–589.

Corbett, G.G., 1982. Gender in Russian: An account of gender specification and its relationship to declension. Russian Linguistics 6 (2), 197–232.

Corbett, G.G., 1991. Gender. Cambridge University Press, Cambridge, UK.

Cowell, M.W., 1964. A Reference Grammar of Syrian Arabic (based on the dialect of Damascus) (Arabic series 7). Georgetown University Press, Washington, DC.

Dixon, R.M.W., 1972. The Dyirbal Language of North Queensland. Cambridge University Press, Cambridge, UK.

Emeneau, M.B., 1955. Kolami: A Dravidian Language. University of California Press, Berkeley and Los Angeles.

Evans, N., 1997. Head classes and agreement classes in the Mayali dialect chain. In: Harvey, M., Reid, N. (Eds.), Nominal Classification in Aboriginal Australia. John Benjamins, Amsterdam, pp. 105–146.

Evans, N., Brown, D., Corbett, G.G., 2001. The semantics of gender in Mayali: Partially parallel systems and formal implementation. Language 77.

Evans, R., Gazdar, G., 1996. DATR: A language for lexical knowledge representation. Computational Linguistics 22, 167–216.

Foley, W.A., 1986. The Papuan Languages of New Guinea. Cambridge University Press, Cambridge, UK.

Fortune, R.F., 1942. Arapesh. J J Augustin, New York.

Fraser, N.M., Corbett, G.G., 1995. Gender, animacy and declensional class assignment: A unified account for Russian. In: Booij, G., van Marle, J. (Eds.), Yearbook of Morphology 1994. Kluwer, Dordrecht, The Netherlands, pp. 123–150.

Fraser, N.M., Corbett, G.G., 1997. Defaults in Arapesh. Lingua 103, 25–57.

Gazdar, G., forthcoming Ceteris paribus. In: Kamp J A W, Rohrer, C. (Eds.) Aspects of Computational Linguistics. Springer, Berlin

Ibragimov, G.X. (Ed.), 1999. Caxurskij jazyk. Nauka, Moscow.

Kibrik, A., 1999. E. In: Ĕlementy grammatiki caxurskogo jazyka v tipologičeskom osveščenii. Nasledie Press, Moscow.

Köpcke, K.M., Zubin, D., 1996. Prinzipien für Genuszuweisung im Deutschen. In: Lang, E., Zifonoun, G. (Eds.), Deutsch typologisch: Jahrbuch des Instituts für deutsche Sprache 1995. de Gruyter, Berlin, pp. 473–491.

Kulick, D., Stroud, C., 1992. The structure of the Taiap (Gapun) language. In: Dutton, T., Ross, M., Darrell, T. (Eds.), The Language Game: Papers in Memory of Donald C. Laycock. Department of Linguistics, Research School of Pacific Studies, Australian National University, Canberra, ACT, pp. 203–226.

Mel'nikov, G.P., Kurbanov, A.I., 1964. Logičeskie osnovanija imennoj klassifikacii v caxurskom jazyke. In: Ivanov, V.V. (Ed.), Voprosy struktury jazyka. Nauka, Moscow, pp. 157–170.

Munro, P., 1998. The Garifuna gender system. In: Hill, J.M., Mistry, P.J., Campbell, L. (Eds.), The Life of Language: Papers in Linguistics in Honor of William Bright. de Gruyter, Berlin, pp. 443–461.

Parker, E.M., Hayward, R.J., 1985. An Afar-English-French Dictionary (with Grammatical Notes in English). SOAS, University of London, London.

Polinsky, M., July 1998. Women are not dangerous things: Grammatical gender and categorization. Paper presented at the Workshop 'Processing of Grammatical Gender,' Max Planck Institute of Cognitive Neuroscience and Centre of Cognitive Sciences. University of Leipzig, Leipzig, 24–25.

Polinsky, M., Everbroeck, E. van, 2000. Noun classes: Language change and learning. Development of gender classifications: Modeling the historical change from Latin to French. Unpublished paper. UCSD.

Polinsky, M., Jackson, D., 1999. Noun classes: Language change and learning. In: Fox, B.A., Jurafsky, D., Michaelis, L.A. (Eds.), Cognition and Function in Language. CSLI, Stanford, pp. 29–50.

Sridhar, S.N., 1990. Kannada (Descriptive Grammars Series). Routledge, London.

Taylor, D., 1977. Languages of the West Indies. Johns Hopkins University Press, Baltimore, MD.

Tucker, G.R., Lambert, W.E., Rigault, A.A., 1977. The French Speaker's Skill with Grammatical Gender: An Example of Rule-governed Behavior. Mouton, The Hague.

Zaliznjak, A.A., 1964. K voprosu o grammatičeskix kategorijax roda i oduševlennosti v sovremennom russkom jazyke. Voprosy jazykoznanija 4, 25–40.

Zubin, D., Köpcke, K.M., October 1983. 1986 Gender and folk taxonomy: The indexal relation between grammatical and lexical categorization. In: Craig, C. (Ed.), Noun Classes and Categorization: Proceedings of a Symposium on Categorization and Noun Classification, Eugene, Oregon. John Benjamins, Amsterdam, pp. 139–180.

Gendered Care and Social Work Practice

Joan Orme, Glasgow School of Social Work, Glasgow, Scotland, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article plots the contribution of feminism to theory and practice in social work. Explorations of practice through a gender lens reveal that essentialist assumptions and stereotypes of male and female characteristics and behavior have underpinned welfare provision and social work practice within it, especially in the area of care work. The article documents how social work has responded to developments in feminist thought especially in relation to how the 'ethic of care' has informed debates about the gendered nature of care in social work.

Introduction

Social work in western society has its roots in the philanthropic movements of the late nineteenth and early twentieth centuries but has developed into a professional activity supported by high-quality education (see Social Work Education) and research. The organization of services may differ between countries, but the rationale for its existence and assumptions about what social work entails are universal. Globally, social work emerged initially as a profession of women dealing mainly with issues pertaining to women's situation. These issues, or domains of social work, involve primarily family and child welfare and the welfare of adults including older people, people with physical disabilities and learning difficulties, and those with mental health problems. The latter categories include both women and men as users of services, but the focus of social work has been on women as sources of support and informal care provision.

In some countries, social work has also had a role in criminal justice systems. The predominance of men in the criminal justice system and expectations about who should work with them has at times led to challenges and assumptions that social work was for women and by women. Also, simple gender analysis which attributed certain characteristics to men and women had implications for the treatment of women in the criminal justice system. A simple gendered analysis that men predominated in the criminal justice system because they were 'naturally' more delinquent and/or violent than women had unhelpful consequences for women (Gelsthorpe, 1989). As Gelsthorpe and others (Heidensohn, 1985) argued, women in the criminal justice system were seen to be doubly deviant: they offended against the law but they also offended against concepts of femininity. Violent women were triply deviant and women who were violent toward their own children were demonized. Analysis of this portrayal and treatment of women in the criminal justice system led to more complex discussions about understandings of gender and gendered social work practice (Orme, 1994).

Social Work and Feminism

Early theorists of social work practice gave little attention to the implications of the dominance of women in the workforce. The welfare system was at best gender unaware or at worst misogynist. For example, there were expectations of a caring profession influenced by feminine characteristics of passivity and maternal instincts. Women dominated in 'caring' professions such as nursing, education, the hospitality industry, and social work. The gendered nature of social work, that is, recognizing certain qualities as female, was first identified by Alice Salomon, a German feminist social work scholar writing at the beginning of the twentieth century who described social work as 'motherly love' (see Kuhlmann, 2008). The positive aspects of such a concept were associated with the responsibility of social workers to provide care and be caring on behalf of the state. But female-dominated professions commanded little respect or influence, despite the invaluable contribution they made to welfare provision (Williams, 1993).

In the United Kingdom, the first observations on the female nature of the workforce was in the Younghusband reports of 1947 and 1951 when it was seen as a deficit of the profession and led to conclusions that efforts should be made to recruit men into the workforce. This was done by professionalizing social work through increased qualifications and salaries. This led, from the 1960s onward, to an increase in men joining the profession. However, the front line workforce remained predominantly female while men who were recruited were quickly promoted to managerial positions (Nottage, 1991). This was in part a result of the very maternal instincts identified as necessary for the professions. Policies were based on expectations that women's careers would be interrupted by childbirth and theories of attachment in child care and development which underpinned social work policy and practice constrained women to be primary full-time carers. Hence, a spiral of gendered assumptions based on notions of caring led to a lack of recognition of women and opportunities for them in the paid workforce and confined them to domestic and caring roles.

It was this conundrum of women in the caring workforce and the challenges from mainstream feminism that precipitated a response from women in social work and focused their attention on issues of gender in the policies, practices, and theories relating to the profession (Hudson, 1985; Dominelli and McLeod, 1989; Dominelli, 1991).

The challenge from feminism focused on women's domination of and their role in the social work workforce. In the 1970s, radical traditions in social sciences of which

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28034-3

second-wave feminism was an important strand began to provide a critique of social work and women working within it (Wilson, 1977). The relationship between gender and welfare was analyzed by social policy feminists (see Orme, 2001; Daly, 2012 for an overview). It was argued that feminists working within welfare may be compromised because they were involved in shoring up state provision that contributed to the continuing oppression of women. The individualized approaches of social work focusing on counseling and casework were associated with a 'band aid approach' (Jagger, 1995) which denied or ignored the need for social change. Women social work practitioners were accused of preserving the status quo.

The engagement of social work theorists within feminism led to a response which initially emphasized women-centered practice. This approach both identified deficiencies in mainstream social work theory in the attention it gave to women and documented the positive contribution that women in general and feminism in particular made to practice (Brook and Davis, 1985; Collins, 1986; Marchant and Wearing, 1986; Hanmer and Statham, 1988; Dominelli and McLeod, 1989). This growing body of work concentrated on both the condition of being a woman and the conditions that women experienced (Orme, 2002a) and led to specific recommendations for intervention identified as feminist social work practice (see Hanmer and Statham, 1988 for a specific code of practice). These included among other things recognizing the conditions of women; always listening to women; raising consciousness of women who were oppressed by their conditions; providing women-only spaces; and identifying commonalities and diversities between women workers and those who used the services.

This codification of practice provided prescriptions for working with women, but it was argued by some (Payne, 2005) that this offered no extensive account of how feminist theory could be applied to generalist social work practice. However, analysis of the impact of feminist theories on social work over time demonstrates that it has had significant influence (Orme, 2012). This influence is not in the form of a particular way of 'doing' social work. The different political strands within feminism are variously described and include liberal, socialist, Marxist, radical, and black. This diversity means that feminism does not provide a single, coherent, and unified way of analyzing and conceptualizing the world. This is a consequence of its exploration of theory but it can be problematic for those practitioners looking for prescriptions for action. Over time, feminist social work theory has engaged with these different positions and now provides a multifaceted approach to social work. This includes particular approaches to practice such as recognizing commonalities and diversities between women workers and service users in one-to-one work (Hanmer and Statham, 1988); narrative approaches (Milner, 2001); community approaches (Dominelli and McLeod, 1989); and group work (Cohen and Mullender, 2003). Developments informed by feminist thinking can be summarized in three main approaches:

- synergies between feminism and social work based on shared values. This involved comparing the mantra 'the personal is political' the social work requirement to deal with the person within their environment (Collins, 1986);
- prescriptions for women-centered practice developed solely from the standpoint of women (Dominelli, 2002) and the implications of this for working with men (Cavanagh and Cree, 1996); and
- use of feminist perspectives to understand the gender dimensions of social work practice that challenge stereotypical assumptions of male and female qualities and behavior (Orme, 2001).

Underpinning these developments was the challenge presented by postmodernist thinking to feminism in general and feminist social work in particular. The prescriptions for action based on women-centered social work were thought to 'essentialize' both women and men according to stereotypical gender definitions. However, while some in social work embraced postmodernism arguing that it reflects feminism's own challenge to overarching male-dominated theories (Fawcett et al., 2000), others suggested that it undermined the political project of both feminism and social work (Dominelli, 2002) or that it was associated with 'academisation' (Gray and Boddy, 2010) which distanced it from practice. Although Gray and Boddy saw the need for acknowledging diversity and difference both between women and men and within the categories 'women' and 'men,' they argue that this was more effectively offered by a different political perspective, postcolonial feminism (2010).

The debates about the different political approaches within feminism and their potential contribution to social work can be illustrated by two significant welfare issues which became the specific focus of feminist social work and help explore understandings of gendered care. These were the response of feminist social work to men who were violent or abusive to women and children and the impact of responsibilities for caring on women's lives.

Gendered Care and Social Work Practice

As has been said, the original gender assumptions in welfare stemmed from beliefs about the essential nature of femininity. Women were expected to be 'the angel in the house' because of their capacity for nurturing which it was said, in a position which echoed Salomon's earlier statements, stemmed from their maternal instincts. These characteristics, assumptions of the 'feminine,' were thought to be necessary for maintaining care services but they were also associated with attributions of passivity which carried negative connotations (Hudson, 1989). It is this paradox that has led to the work on an ethic of care (or alternatively care ethics) becoming influential in understandings of gendered care in social work.

Care Ethics

Over the past two decades, debates involving the deconstruction of dichotomous thinking about men and women have been influential in both theorizing about care in general and the implications of the ethic of care for policy and practice. Originally based in social psychology on the work of

Gilligan (1982), care ethics has become a body of theory that has expanded to inform both theory and practice. Significantly, initially there was surprisingly little attention to the implications of an ethic of care from practitioners in professions that provided care such as nursing and social work. For some, the notions of social work and care were synonymous to the extent that the terms social work and social care were interchangeable, but attention to debates about care ethics within the wider social sciences was largely ignored in social work literature until the early twenty-first century. This may have been because they saw it as an uncontested concept and took for granted an understanding of care because, as is documented earlier, their professions were based on particular constructions of care. Recent work has not only focused on understandings of care ethics in general and their implications for social work practice in particular (Orme, 2001; Parton, 2003) but also debates about notions of justice (Held, 1995; Orme, 2001, 2002b) and identity (Fawcett, 2000; Orme, 2001). This work, as will be discussed later, has unraveled certain complexities.

Gilligan (1982) set out to challenge the work of Kohlberg (1981) whose studies in moral development had concluded that women had inferior moral reasoning to that of men: or more accurately women stayed at an inferior stage of moral development. Initially, Gilligan appeared to reinforce differences based on definitions of male and female in that from her empirical studies she argued that men operated from a logical rational ethic of justice while women operated from a relational ethic of care. It was this ethic of care that constituted women's 'different voice.' Significantly, Gilligan rooted her arguments in object relations theory. This is a branch of psychoanalysis which emphasizes the impact of early childhood experiences on an individual's subsequent interpersonal relationships (see Fairbairn, 1952 as one of many writers on this subject). The experiences focused on were seen as 'objects' but these objects included people, particularly mothers. Gilligan theorized that because mothers experienced their daughters as less separate, these daughters had less experience of separateness and boundaries and were therefore more connected in later relationships. This in turn gave them a greater capacity for empathy and sensitivity toward others. According to Gilligan, this enabled them to operate from an ethic of care. Boys, however, experienced separateness from their mothers at an early age and had to suppress feelings of intimacy and dependency which made them operate from the basis of an ethic of justice. There are, therefore, echoes here of the understandings of feminine caring based on 'maternal' qualities and motherhood that had permeated the profession of social work.

Although Gilligan's work made a major contribution to understandings of care by challenging orthodox thinking about moral development, there appeared to be orthodoxy in her thinking around gendered roles. Hence, there were criticisms of her position (see Featherstone and Morris, 2012). Her distinction between an ethic of care and an ethic of justice was seen by some as unhelpfully essentializing differences between men and women. In defense, Gilligan (1982) argued that the gendered distinctions were not based on biological differences but that the 'different voice' highlighted alternative ways of relating to and providing services for others. What was significant, Gilligan suggested, was that this voice did not represent an inferior form of moral reasoning but should be privileged above, or at least be equated to, the rational technical ethic of justice.

While embracing an ethic of care, Tronto (1993) and others (Sevenhuijsen, 1998, Koggel and Orme, 2013) focus less on the negative, passive elements of care and more on relational and dialogic approaches. They identify a number of reasons for moving beyond an emphasis on object relations and notions of mothering. In the first instance, many involved in care work (both women and men) are not mothers; and, second, those who are mothers do not always demonstrate the enabling nurturing that is associated with an ethic of care. Tronto, therefore, argues that care is not necessarily associated with either male or female behavior but is a species activity that includes everything we do. Care involves the capacity to take the other's needs as the starting point for what must be done. She identifies four phases: caring about; taking care of; caregiving; and care receiving. These phases are based on four ethical elements: attentiveness, responsibility, competence, and responsiveness. Such an analysis is pertinent to social work because social work practice operates at the public/private interface of welfare provision. Also the importance given to context, interdependence, relationships, and responsibilities to concrete others emphasized by such an analysis highlights that an ethic of care is relevant in the political as well as the personal realm (Tronto, 1993). Care ethics is relevant to both interpersonal caring relationships and the institutions responsible for providing them, including those in the political realm.

In terms of the association of feminist analysis and gendered care some argue that Gilligan's ethic of care was not necessarily feminist (Koehn, 1998). In her attempt to identify specifically feminist ethics, Koehn considers an ethic of broad empathy as a development of, or alternative to, an ethic of care. Empathy is a core concept in social work and is one that has presented paradoxes for feminist social workers. While recognizing the objective conditions of those with whom they work, social workers are expected to understand, accept, and respect individual understandings and expressed wishes. Early women-centered practice advocated emphasizing commonalities and differences as a way of understanding women's experience (Hanmer and Statham, 1988) but in doing so recognized that this challenged the expectations of objectivity by overidentifying and/or becoming overemotionally involved with women service users. Koehn's understanding is useful in that it argues that empathy means more than trying to understand the other's feelings but involves engaging with the other by utilizing analytic discernment and attentive reciprocity (achieved by dialogue). It also means going beyond what can usually be understood and accepted, to recognize "tropes and images that might strike us initially as quite radical and even repellent" (Koehn, 1998: p. 58). This is significant in applying care ethics to gendered social work because it requires the acceptance of men and those, women and men, who have gone beyond the experiences of practitioners. For example, gendered practice requires social workers to accept and work with those who

have committed the 'unthinkable' in child abuse (Featherstone and Lancaster, 1997).

The implications for working with some of the most marginalized within 'care' services – men who abuse women – present the greatest challenge to social work practice based on a standpoint or women-centered approach. Social work within the criminal justice system has had a particular remit to work with such men but this has presented certain challenges to the social work value base if they are seen to be beyond 'care.' Social work more frequently works with those who have been the victims of violence, sexual abuse, and other forms of oppression – and many of these are women and female children – and this has meant that gendered social work has been accused of ignoring the needs of the offender and the oppressor. However, the different feminist approaches in social work identified above meant that it is possible to go beyond a standpoint approach (i.e., one that argued the world should be perceived solely from the 'standpoint' of women). This meant that the needs of male perpetrators should also be considered and led to the reinterpretation of an ethic of justice (Orme, 2002b).

It was the wider understanding of an 'ethic of care' and the way it was set in opposition to an ethic of justice that informed thinking about work with male perpetrators of domestic violence and abuse. While early literature on domestic violence focused primarily on the conditions of women, developments in gendered practice required reconsideration of understandings of gender and the implications of this for social work with perpetrators (Orme et al., 2000). Koehn's understandings of empathy does not mean agreeing with abusive behavior, nor does it involve imputing wants and needs to individuals, nor constraining them by assumptions based on gender stereotypes, but, rather, being prepared to hear the conflicts and confusions. This leads to understandings of an ethic of care that challenge the process of dichotomizing care and justice. Care becomes the mercy that tempers justice (Jagger, 1995). Hence, an ethic of care that incorporates the principles identified by Tronto (1993) and based on a strong conception of rights leads to a caring justice. The obverse of this is that gendered practice also requires 'just care' (Orme, 2002b). Such a concept develops from the critique of 'traditional' caring that has come from women with disabilities and men who care.

As has been outlined earlier, early feminist critiques of care policy focused on the burden of care that fell on women in society (Dalley, 1988). This negative representation of those requiring care was challenged by women who required care. The categorizing of those requiring care into groups defined by their need or their condition (e.g., disabled, old, or mentally ill) denied everyone their personhood, and in particular denied women their womanhood (Orme, 2001). Feminists within the disability movement argued strongly against such categorization pointing out not only that they were women too, but also that they might be providers of care for others (Morris, 1992). Hence questions of identity were being challenged by those involved in the care system.

Universal assumptions of care as burden also challenged the nature of care offered. That care associated with maternal instinct was seen to be positive, denied the potential negative consequences of such care. Mothering can be overbearing and deny personhood often infantilizing those who require care. Rather than reflecting altruism, the provision of care can involve surveillance and the disciplining of those being cared for. Such negative conceptions of care also have expectations of passivity and gratitude on the part of the person receiving it. All of which denies them personhood. Hence, the stages of care outlined by Tronto (1993) have not only liberated those expected to provide care but also those who require care.

An ethic of care based on the processes of care rather than on who provides it has enabled gendered practice to respond to the challenge that came from men who provided care. While early statistics showed women as carers, further analysis led to suggestions that men might be the 'hidden' carers. When evidence emerged that men did provide care, a gendered perspective based on care as a feminine quality suggested that the nature of care provided differed between men and women. It was argued that men provided predominantly spouse care or care in close personal relationships; they did not care for members of the wider family or for strangers (Arber and Ginn, 1990). Men provided care for those they cared about (Rose and Bruce, 1995). This caring 'about' also informed the hypothesis that the nature of the care provided differed between women and men: women provided intimate personal care while men looked after those requiring care, they 'managed' the care rather than providing hands-on care. Also with the involvement of gay men in the buddy system that required intimate caregiving for those suffering from AIDS, the gendered assumptions based on stereotypical male and female characteristics and qualities were challenged (Orme, 2001). Again, Tronto's (1993) phases of care (caring about; taking care of; caregiving; and care receiving) based on the quality of care derived from the four ethical elements (attentiveness, responsibility, competence, and responsiveness) rather than who provides the care have led to changed understandings of gendered care.

A final postscript to these developments is provided by considering the implications of care ethics for the organization of care within the changing nature of welfare. Care ethics is relevant to both interpersonal caring relationships and the institutions based on these including those in the political realm. Developing theories about the application of an ethic of care to policy development and political organization offer a critique of neoliberalism and associated policies (Barnes, 2012; Featherstone and Morris, 2012). This is crucial for social work practice which, even when operating outside statutory organizations, is influenced by political changes. In the United Kingdom, policy development over the last 20 years has become increasingly based on individual responsibility and the marketization of welfare including principles of personalization (Dominelli, 2004).

As was said at the outset, the availability of women as carers was crucial in the development of welfare in the latter part of the twentieth century. Since then in western societies generally, and in the United Kingdom in particular, emphasis on care in the community has meant the closure of institutions and the need for resources for caring to be available in communities. These policies involved care by communities: they moved responsibility for care to individuals, families, communities, and private- and third-sector organizations rather than the state. Later developments involved the creation of 'markets' in welfare

which have involved the commodification of care. Care is no longer based on either altruism or empathy but is something to be purchased. In the first instance, this involved private companies contracted by statutory organizations to provide care previously provided directly by the state (Ungerson, 2003). Most recently, a more individualistic culture developed. Policies which were said to be related to user empowerment led to a personalization agenda where the state rather than provide services allots personal budgets to individuals requiring care, making them responsible for securing and managing their own care. They purchase the services of other individuals. As Barnes (2012) argues, such policies emerge from different but intertwining ideologies. The first is the primacy of the rational, autonomous, self-sufficient individual which can be defined by an ethic of justice and the second is the right of those who are in receipt of services to be empowered, to be involved in the service delivery required to meet their needs which is a manifestation of just caring.

Such developments have implications for understandings of gendered practice. The creation of a 'care market' associated with the commodification of care leads to a set of paradoxes (Orme, 2003). Payment for care, especially in times of economic depression, means that such employment may become attractive to men. While the finances mean that wages will be low, being able to purchase care by male carers might offer more choice to those who require it. Payment for care challenges assumptions that it is the expression of a maternal instinct or an act of altruism associated with 'feminine' behavior. Hence, the myth of feminine care is exploded and opportunities are created for men to behave differently, to operate with, as Gilligan (1982) argued, a different voice. However, such policies also raise questions about the nature of services that are paid for; if it is not motivated by altruism or empathy does providing service constitute care? Having said that when there are insufficient resources either to pay for the level of care required by the individual or to provide the appropriate level of staffing required to meet all the care needs of a personalization agenda, or when the person requiring care is not competent to manage budgets, then the provision of care will again fall disproportionately on women. This is because, despite developments in theory, practice, and policy, public attitudes still perceive care as gendered and that the gender of the carer is feminine.

Conclusion

This article has attempted to demonstrate that gender is integral to social work practice in a number of ways but that understandings of gendered practice have changed over time. Social workers initially argued that social work theory was gender-neutral, feminism challenged this. However, initial attempts to define the gendered nature of the profession led to paradoxes. When a wider and more flexible lens of gender was adopted, this led to a more nuanced analysis of social work practice. First, it highlighted that women-centered practice did not necessarily uphold the values of social work and feminism in its attitudes to working with men. Second, it complexified the role of men in the care workforce. Reworking social work concepts such as individualization by drawing on notions of identity led to an approach to social work that was more inclusive, just, and caring. This drew initially on Gilligan's (1982) theory of an ethic of care but incorporated later developments in moral philosophy which widened the implications beyond individualized social work interventions to the organization of the provision of services.

Gendered care, therefore, constantly challenges how the underlying philosophies and ideologies of policy influence the conditions in which social work can practice. At one level, policy which appears to democratize care, to move it from the power of the professionals and involve those who require it, might be seen to reflect a 'caring' approach. However, if such caring is not based on a true dialogic approach which informs all relationships and interventions from the highly political to the intensely personal, then it is not gendered care informed by an ethic of care because in the words of the feminist mantra the 'personal is political.'

See also: Community Care in England; Feminist Research and Social Work; Masculinity and Social Work Practice; Social Work Education.

Bibliography

Arber, S., Ginn, J., 1990. The meaning of informal care: gender and the contribution of elderly people. Ageing and Society 10, 429–454.
Barnes, M., 2012. Care in Every Day Life: An Ethic of Care in Practice. Policy Press, Bristol.
Brook, E., Davis, A. (Eds.), 1985. Women, the Family and Social Work. Tavistock, London.
Cavanagh, K., Cree, V. (Eds.), 1996. Working with Men: Feminism and Social Work. Routledge, London.
Collins, B.G., 1986. Defining feminist social work. Social Work 31 (3), 214–219.
Cohen, M.B., Mullender, A. (Eds.), 2003. Gender and Groupwork. Routledge, London.
Dalley, G., 1988. Ideologies of Caring. Macmillan, Basingstoke.
Daly, M., 2012. Gender and Welfare. In: Gray, M., Midgley, J., Webb, S.A. (Eds.), The Sage Handbook of Social Work. Sage, London, pp. 81–93.
Dominelli, L., 1991. 'Race' and Gender in Social Work. In: Davies, M. (Ed.), The Sociology of Social Work. Routledge, London, pp. 182–201.
Dominelli, L., 2002. Feminist Social Work Theory and Practice. Palgrave Macmillan, Basingstoke.
Dominelli, L., 2004. Social Work: Theory and Practice for a Changing Profession. Polity Press, Cambridge.
Dominelli, L., McLeod, E., 1989. Feminist Social Work. Macmillan, Basingstoke.
Fairbairn, W.R.D., 1952. An Object-Relations Theory of the Personality. Basic Books, New York.
Fawcett, B., 2000. Feminist Perspectives on Disability. Prentice-Hall, Harlow.
Fawcett, B., Featherstone, B., Fook, J., Rossiter, A. (Eds.), 2000. Practice and Research in Social Work: Postmodern Feminist Perspectives. Routledge, London.
Featherstone, B., Lancaster, E., 1997. Contemplating the unthinkable: men who sexually abuse children. Critical Social Policy 17 (4), 57–71.
Featherstone, B., Morris, K., 2012. Feminist ethics of care. In: Gray, M., Midgley, J., Webb, S.A. (Eds.), The Sage Handbook of Social Work. Sage, London, pp. 341–354.
Gelsthorpe, L., 1989. Sexism and the Female Offender: An Organizational Analysis. Gower, Aldershot.
Gilligan, C., 1982. In a Different Voice: Psychological Theory and Women's Development. Harvard University Press, Harvard, MA.
Gray, M., Boddy, J., 2010. Making sense of the waves: wipeout or still riding high? Affilia: Journal of Women and Social Work 25 (4), 368–389.
Hanmer, J., Statham, D. (Eds.), 1988. Women and Social Work: Towards a Women-Centred Practice. Macmillan, Basingstoke.

Heidensohn, F., 1985. Women and Crime. Macmillan, Basingstoke.
Held, V. (Ed.), 1995. Justice and Care: Essential Readings in Feminist Ethics. Westview Press, Boulder, CO and Oxford.
Hudson, A., 1985. Feminism and social work: resistance or dialogue? British Journal of Social Work 15, 635–655.
Hudson, A., 1989. Changing perspectives: feminism, gender and social work. In: Langan, M., Lee, P. (Eds.), Radical Social Work Today. Routledge, London, pp. 70–96.
Jagger, A., 1995. Caring as a feminist practice of moral reason. In: Held, V. (Ed.), Justice and Care: Essential Readings in Feminist Ethics. Westview Press, Boulder, Colorado and Oxford.
Koehn, D., 1998. Rethinking Feminist Ethics: Care Trust and Empathy. Routledge, London.
Koggel, C., Orme, J. (Eds.), 2013. Ethics and Social Welfare. Routledge, London.
Kohlberg, L., 1981. The Philosophy of Moral Development. Harper and Row, San Francisco.
Kuhlmann, C., 2008. Alice Salomon (Germany), President 1928/29–1946. Social Work & Society International Online Journal 6 (1), 128–141. http://www.socwork.net/sws/article/view/99/388 (accessed 20.06.13.).
Marchant, H., Wearing, B. (Eds.), 1986. Gender Reclaimed: Women in Social Work (Perspectives in Social Welfare). Hale & Iremonger, Sydney.
Milner, J., 2001. Women and Social Work: Narrative Approaches. Palgrave, Basingstoke.
Morris, J., 1992. Personal and political: a feminist perspective on researching disability. Disability, Handicap and Society 7 (2), 157–166.
Nottage, A., 1991. Women in Social Services: A Neglected Resource. SSI, London.
Orme, J., 1994. Violent women. In: Lupton, C., Gillespie, T. (Eds.), Working with Violence. Macmillan (now Palgrave), Basingstoke.
Orme, J., 2001. Gender and Community Care: Social Work and Social Care Perspectives. Palgrave, Basingstoke.
Orme, J., 2002a. Feminist social work. In: Adams, R., Dominelli, L., Payne, M. (Eds.), Social Work: Themes, Issues and Critical Debates. Palgrave Macmillan, Basingstoke.
Orme, J., 2002b. Social work: gender, care and justice. British Journal of Social Work 32 (6), 799–814.
Orme, J., 2003. Social work and the paradox of care (Chapter 1.14). In: Labonté-Roset, C., Marynowicz-Hetka, E., Szmagaiski, J. (Eds.), Social Work in Education and Practice in Today's Europe: Challenges and the Diversity of Responses. Slask, Katowice.
Orme, J., 2012. Feminist social work. In: Gray, M., Webb, S.A. (Eds.), Social Work Theories and Methods. Sage, London, pp. 87–98.
Orme, J., Dominelli, L., Mullender, A., 2000. Domestic violence: working with men from a feminist perspective. International Social Work 43 (1), 69–105.
Parton, N., 2003. Rethinking professional practice: the contribution of social constructionism and the feminist 'ethic of care'. British Journal of Social Work 33 (1), 1–16.
Payne, M., 2005. Modern Social Work Theory. Palgrave Macmillan, Basingstoke.
Rose, H., Bruce, E., 1995. Mutual care but differential esteem: caring between older couples. In: Arber, S., Ginn, J. (Eds.), Connecting Gender and Ageing: A Sociological Approach. Open University, Buckingham, pp. 114–128.
Sevenhuijsen, S., 1998. Citizenship and the Ethics of Care. Routledge, London.
Tronto, J., 1993. Moral Boundaries: A Political Argument for an Ethic of Care. Routledge, New York and London.
Ungerson, C., 2003. Commodified care work in European labour markets. European Societies 5 (4), 377–396.
Williams, C.L. (Ed.), 1993. Doing 'Women's Work': Men in Non-Traditional Occupations. Sage, California.
Wilson, E., 1977. Women and the Welfare State. Tavistock, London.

Genealogy in Anthropology

Mary Bouquet, Utrecht University, Utrecht, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

More than a century after W.H.R. Rivers formalized the genealogical method of anthropological inquiry, anthropologists are still debating the assumptions written into it through the notion of pedigree, considering the impact of genealogical thinking beyond kinship, and exploring new ethnographic cases. The historical and cultural specificity of Rivers' method percolated into colonial practice. New reproductive technologies used for humans and animals draw upon genealogical thinking in unexpected ways. Record linkage for commercial biotechnological purposes triggered controversy in Iceland and new forms of digital genealogy. New ways of examining genealogy use multidisciplinary approaches explored in three-dimensional exhibition space.

Introduction

More than a century after W.H.R. Rivers formalized genealogy as a method in anthropology, anthropologists are debating the premises of the genealogical model while exploring its historical and contemporary manifestations. (Others caution the emphasis on genealogy at the expense of addressing the unfinished business of kinship terminology (see, for example, Read, 2007).) Some have argued the case for developing alternative ways of theorizing relatedness (e.g., Carsten, 1997, 2000; Ingold, 2000; Franklin and McKinnon, 2001). Others defend the continuing relevance of genealogical thinking for understanding the human condition (Zerubavel, 2012); and the on-going pertinence of the distinction between biological and the social relatedness written into the genealogical framework for understanding kinship (e.g., Astuti, 2009) Historical and ethnographic studies covering a wide spectrum of subjects – from state and colonial uses of genealogy to the intertwining of human and animal pedigrees through breeding ideologies, practices, and stories, to everyday relations intercepted by new reproductive technologies, to popular and digital genealogy – extend, in often unexpected ways, the range and depth of this cultural phenomenon (see Henare, 2007; Bamford and Leach, 2009; Cannell, 2011). On the one hand, the peculiarity of genealogy as a history of relatedness among persons involving the transfer of bodily substance and cultural knowledge that is somehow abstracted from their relationships with the environment has been clearly analyzed (Ingold, 2000). On the other hand, the re-surfacing of genealogical thinking in unexpected places has given a new impetus to empirical research and pushed the study beyond kinship as originally understood (Bamford and Leach, 2009). Anthropologists investigate the way human and non-human pedigrees are interconnected (Cassidy, 2002, 2009). They study the remaking of genealogy whereby the lives of animals, such as sheep, are bound together with humans through old and new technologies of reproduction (Franklin, 2007); and they acknowledge genealogy as a practice shaping and informing everyday life where relations have been forged through the new reproductive technologies (Edwards, 2000, 2009). They have looked at the effects of extending the logic of pedigree and recording genealogies to work undertaken by amateur ethnographers in the early twentieth-century colonial service (Holmes, 2009). They have investigated how the mapping of the human genome has prompted new sets of linkages (in the case of Iceland) between medical records, DNA records, and genealogical records and at the link between pedigree and capital in digital genealogies (Pálsson, 2009; Cunningham, 2009). They delve into the significant contemporary role of genealogy, such as Maori *whakapapa*, whereby a person is the living face of their ancestors and the interweaving of Maori concepts with legal discourse (Henare, 2007). And they examine state suppression of genealogies elsewhere – as, for example, by the Mongolian authorities in the 1930s, which resulted in the destruction, concealment or mutation of Buryat genealogies into other forms – such as photomontages (Empson, 2007), or in the imaginative forms of *bilum* trees (Herle et al., 2009). In arguing against the overpowering effect of the genealogical framework, others have emphasized different geometries – horizontal relatedness and diagonal links – as well as the significance of sustenance, land, memory, and knowledge in constructing relatedness.

The following sections will consider the definition, method, and assumptions of genealogy, bringing in some of the new ethnographic cases of genealogical thinking and imagination and some of the ongoing debates that have taken place.

Genealogy defined

> Genealogy, *n* 1 an account of the descent of a person, family, or group from an ancestor or from older forms; 2 the descent of a person, family, or group from an ancestor or from older forms; pedigree, lineage; 3 the study of family pedigrees. (*Longman Dictionary of the English Language*, 1984: p. 611)

This definition of the term genealogy involves three components: first, it is a *representation* of the descent, which may be written (for example, the Book of Genesis), or graphic (for example, the Tree of Jesse). Second, there is descent as "…the *fact* or *process* of derivation from an ancestral stock, lineage; … transmission or devolution of an estate by inheritance, usu in the descending line … a transmission from a usu earlier source that has had a shaping or developing effect; a derivation-…"(*Longman*, 1984: p. 394). Third, the definition refers to the *study* of family pedigrees: the activities of amateur genealogists

and local historians avidly using censuses and parish records to create new accounts of descent.

The several component layers of genealogy – as an 'account,' as a 'fact or process,' and as a 'study' – hold complex implications for anthropology. The written or graphic account introduces a bureaucratic dimension whereby knowledge, in the form of record-keeping, opens up the potential for manipulation and the exercise of power. The exclusion of those deemed illegitimate is as compelling a dynamic of genealogy as is the cultivation of those seen as rightfully composing the ancestral line. The account is supposed to deal with descent: the links of between parent and child repeated over the generations – backward and forward – with social emphasis placed on the whole series of such links, which vary (see Parkin, 1997). Generational links imply the transmission of 'something' that is shared by descendents and this has been the focus of anthropological study. While some distinguish between genealogy and category (Parkin, 1997: pp. 50–51), others connect the two, arguing that the metaphor of transmission implies that both genetic information and cultural classification are being passed across the generations (Ingold, 2009: p. 196). This re-convergence of the biological and the social, which had been separated in twentieth-century kinship theory, subtends much of the renewal of theoretical interest in genealogy. The study of pedigree dynamics and uses, whether human or animal, or connecting the two, for example in biotechnology, has generated new questions that have given rise to a renewal of interest in genealogy as a field of empirical inquiry in anthropology.

The Genealogical Method

These developments require some historical contextualization. The genealogical method of anthropological inquiry was developed by W.H.R. Rivers during the 1898 Cambridge Torres Strait Expedition, and formalized in articles published in 1900, 1910, making its entry to *Notes and Queries in Anthropology*, in 1912, and subsequent editions. It could be argued that Rivers was building upon an insight by Charles Darwin in *The Origin of Species* (1859). In his natural system of genealogical arrangements, Darwin removed "the connotations of rank and status attached to the fact of knowing one's pedigree" replacing it with "the assumption that *a genealogy is a recorder of natural relations*" (Strathern, 1992: p. 91). Crucially, Rivers's reconfiguration of pedigree into the genealogical method re-focused on human relations, seeking to establish the limits of people's knowledge of kin, by birth and by marriage; and from there, recording their ways of referring to and addressing the persons they had identified. The nineteenth-century intellectual context of this method lay in a rich combination of biblical, philological, heraldic, and evolutionary ideas. Darwinian theories concerning the origin of species and descent of man had recourse to both narrative and graphic genealogical imagery (Beer, 1983; Strathern, 1992). The familiarity of biblical and historical family trees made genealogy a ready vehicle for translating new ideas about the relationships between all forms of life, and the slow transformation (or evolution) of these forms through the immense abysses of geological time that opened up in the nineteenth century. Graphic phylogenies, such as those drawn by Ernst Haeckel (1874), expressed these evolutionary connections in the form of a tree (Bouquet, 1995, 1996).

Genealogies are central to the creation story in Genesis, which was a source of inspiration for late eighteenth- and early nineteenth-century philologists and ethnologists who sought to explain the relationships between the languages (and hence peoples) of the world in genealogical terms. Still earlier, culturally diverse objects brought to Europe from the New World starting in the sixteenth century, were accommodated by Renaissance collections using medieval categories which included 'paganism.' "'Paganism' both allowed the admittance of cultural diversity and provided it with a biblical pedigree," as Anthony Shelton (1994: p. 201) puts it. "The date and circumstances of the divergence of the pagan races from the Christian world in the remote past greatly occupied the minds of the chroniclers who endlessly hypothesized on the origins of the Aztec and the Inca, who they usually related to an obscure offspring of Noah or a lost tribe of Israel … Small wonder, then, that racial genealogies and conjectural histories partly informed written compendiums of customs and arrangements of material culture" (Shelton, 1994: p. 203).

Philological, biblical, and Darwinian genealogical ideas influenced L. H. Morgan, author of *Systems of Consanguinity and Affinity of the Human Family* (1871), in his kinship and beaver research among the Objibwa around Marquette, which took place against a background of wilderness receding from American farms (Feeley-Harnik, 1999). This was rather different from Darwin's invocation of genealogical imagery in his theory of evolution, which was filtered through the ideas of British animal breeders for whom keeping records was an important means of exercising control over procreation. The fact that the British aristocracy could trace their genealogies, just as they could trace the pedigrees of their favorite animals, such as thoroughbred horses, helped to distinguish them in terms of descent and breeding from other ranks. Written pedigrees of thoroughbred racehorses, recorded in the *General Stud Book* (established in 1791), transform knowledge about people and horses in very specific ways (Cassidy, 2009, 2002, and below).

Rivers' genealogical method became a means of organizing fieldwork and written ethnography, despite being later overshadowed by, for example, participant observation (see Malinowski, 1922). The historical incorporation of pedigree into anthropological methodology is a remarkable maneuver in terms of class and cultural capital in the context of early twentieth-century Cambridge (where Rivers was based), which was still then dominated by a few great families. Borrowing pedigree was an assertion of being above this sort of hereditary ranking system, while harnessing its power as a form of intellectual capital for deployment in the comparative study of mankind (see Bouquet, 1993).

Recent research on Kenya shows that genealogical thinking percolated into British colonial administration in early twentieth-century in East Africa. Amateur ethnographers in the colonial service attempted to make sense of the 'turbulent collection of tribes' by privileging descent through men as the source of indigenous identity. "[F]luid and shifting social spaces of indigenous communities were materially reordered as colonial agents established, or invented, the rigid and internally coherent structure of 'the tribe' as the conceptual framework for colonial administration and control" (Holmes, 2009: p. 57).

Given the importance of pedigree in British society generally at that time, Teresa Holmes argues that the link between pedigree and the recording of genealogy can be extended to the activities of amateur ethnographers in the colonial service. Tribal representatives identified through this recorded descent information were seen as providing a natural connection to land 'passed on in the blood from original ancestral settlers' (Holmes, 2009: p. 63). These ideas were particularly problematic in the Nzoia region of Kavirondo, in western Kenya, where Luo had intermarried with Bantu and shared resources. The area was characterized by much movement between communities and disputes over land and trade. Bantu Wanga were identified by the British as primary occupants of the land by means of a map and a genealogy supporting Wanga connections with the land. By 1916, Kager Luo were contesting this interpretation by claiming landowner status for themselves and inviting 'tenants' to support their claims. Holmes' Kenyan research demonstrates how the genealogical model informed both the anthropological paradigm of segmentary lineage systems and the colonial model of the tribe.

Rivers' genealogical method would have been a welcome development not only for anthropologists but also for colonial administrators. It maps informants' knowledge of their relatives by proceeding from the concrete level of personal names to the abstract system of relationships. Rivers explains the steps by which he collected a pedigree (he uses the terms pedigree and genealogy interchangeably.) First, the descendents of the informants' 'real' (as opposed to classificatory) parents were listed, including previous or other marriages, and children listed in order of age, together with their marriages and offspring. The mother's and father's respective pedigrees were collected by following the same procedure. Rivers claimed that the most extensive pedigrees could be collected using only five terms of relationship: 'father, 'mother,' 'child,' 'husband,' and 'wife.' The fact that Rivers treated other categories – 'brother,' 'sister,' 'cousin,' 'uncle,' and 'aunt' – as problematic gave the entire method a lineal bias. Later scholars have argued that lineal relationships were seen as less ambiguous than either collateral or affinal ones (Barnard and Good, 1984). Information concerning social condition, residence, totemic, and other group affiliation, along with other relevant facts, were to be recorded for each person on the pedigree as far as possible. An extraordinarily wide range of topics could be studied in this way. At the heart of the method was the assumption that kinship was about the social recognition of certain genealogical relationships; and genealogy was assumed to be about biology.

Genealogical Thinking

Despite Bronislaw Malinowski's (1930) criticism that genealogy led to the overabstraction of kinship studies and the emphasis he placed on long-term participant observation, the genealogical method proved to have remarkable staying power in the fieldwork toolkit. It became one of the principal methods of ethnographic fieldwork during subsequent decades providing material for structural-functionalist analyses and theories of social organization. Although the study of kinship had faded from the forefront of discussion by mid-century, collecting genealogies remained part of the ethnographer's stock-in-trade, and genealogical diagrams became one of the minimum requirements for making fieldwork intelligible (Barnes, 1967). Genealogy began to be critically scrutinized from the late 1960s when David Schneider published his cultural account of American kinship (Schneider, 1968): there were calls for a social history of the genealogical tool (Bourdieu, 1977), and anthropological assumptions about the 'genealogical unity of mankind' underlying kinship studies were called into question (Schneider, 1984).

A number of critical studies and ethnographies of kinship, during the 1990s, helped to reinvigorate the study of genealogy as a matter of theoretical interest and ethnographic interest (e.g., Strathern, 1992; Bouquet, 1993; Carsten, 1997; Franklin, 1997). Since 2000, ethnographic research and theoretical debate concerning genealogy have taken several distinctive directions: there is a strong – though certainly not exclusive – focus on Britain. This trend, seen in Holmes' work discussed above, is also exemplified by Rebecca Cassidy's work on thoroughbred horse breeding (Cassidy, 2002, 2009), and Sarah Franklin's research on sheep breeding (Franklin, 2007). Cassidy found that since, among horses, breeding sells, pedigree must be regulated. Among racing enthusiasts, breeding suggests entitlement. Cassidy found the grammar of pedigree compelling "for people who act on the principle that 'like begets like'"(Cassidy, 2009: p. 26). In fact, she found that pedigrees run through all the stories the racing world tells about itself. The conflation of human and animal is used to express the logic of pedigree, including the idea of improvement as a means of increasing status; and the exclusion of nonracing family members and prevalence of patriarchy.

The intertwining of human and animal genealogies in the racing world recurs in the shared genealogies of humans and sheep (Franklin, 2007). Sarah Franklin draws upon 'the very conservative idiom of genealogy' to understand the sheep Dolly as a genetic mix – the merging of two cells through a technique of 'decapacitation' (Franklin, 2007: p. 41). The analysis stretches back into the past of sheep domestication, the role of sheep in the colonization of Australia, and forward through new breeds such as Dolly, to try to make sense of the life and death of this particular animal. Franklin delineates a "method of genealogical reckoning that not only looks back at how certain conditions came into existence but at how they are orientated toward the future, and how they can be seen as promissory repositories of the future" (Franklin, 2007: p. 127). The case of Dolly thus provides a window on the remaking of genealogy: the human–animal relations that emerge from new reproductive technologies, but which belong to a descent pattern already known from agricultural experimentation in pursuit of new commercial markets (Franklin, 2007: p. 155).

Genealogical thinking and breeding practices criss-cross the human–animal boundary, calling into question the notion of a genealogical unity of mankind.

Digital Genealogy

Information technologies play a major role in the new commercial, biomedical, and cultural uses of genealogy. The connection between biomedical research, genealogical records, health records, and state sanctioned commercial record linkage

in Iceland generated controversy and attracted the attention of theorists to online genealogy (Pálsson, 2009; Cunningham, 2009). Iceland's relatively homogeneous gene pool, combined with excellent medical and genealogical records (reflecting the Icelandic cultural passion for genealogy), made it eminently suitable for commercial research to identify the genetic codes responsible for common diseases and the development of drugs for their treatment. However, record linkage and the creation of commercially financed centralized databases – the Book of Icelanders and the Health Service Database – by the (American venture capital financed) company deCODE together with Frisk Software led to controversy about the role attributed to private companies by state legislation. The companies were challenged in the courts by publishing company Genealogia Icelandorum on the grounds of violation of property rights (Pálsson, 2009: p. 99). DeCODE's response was that 'genealogical records are the common heritage of all Icelanders' – a tradition they proposed to perpetuate by making genealogical records accessible to the public online (Cunningham, 2009: p. 131). Accessing the highly visible web version of the Book of Icelanders database became a popular pastime among Icelanders. However, it also had unforeseen consequences as unknown relations were exposed, or people were omitted; it created anxieties about citizenship, personhood, and property rights, as genealogies were removed from the sphere of social history and relocated in the sphere of biological commerce (Cunningham, 2009: p. 134). Although the capitalization of human genes involves reconfiguration of traditional alignments between heredity, family, and character, this reorientation always takes place in geopolitical, economic, social, and cultural contexts such as Iceland. One significant finding was that not only new, horizontal connections are important but that old ideas about the vertical movements of genes, orchestrated by consumer choice and directed genetic mobilities, also remain important in the era of genomic technologies (*idem*).

Diagonal Genealogy

The question of how the genealogical model might be mobilized to narrate connections forged through novel forms of conception is posed by Jeanette Edwards in reflecting on her research into the new reproductive technologies in Alltown (Edwards, 2009). Edwards questions whether all English folk draw on genealogy in the same ways and, if they do, whether genealogical links are as fixed and uncompromising as the model might suggest (Edwards, 2009: p. 138). Of course, the early twentieth-century middle-class English folk model of pedigree that Rivers formalized into the genealogical model was both historically and class-specific. As Holmes' work shows, its extension into administrative procedures by amateur ethnographers in the colonial service coincided historically with Rivers' formalization of the genealogical method. So the ethnographic question Edwards raises is also a historical one. It concerns the specificities of cultural capital formation and deployment among historically specific classes and class fractions as that exemplified in the genealogical method. If genealogy was further democratized over the course of the twentieth century, through digital genealogy, through the availability of new reproductive technologies, and through commercialization, then it is rightly called an ethnographic question.

Hilary Cunningham's discussion of Icelandic genealogy identified both horizontal and vertical movements arising from commercialization and popular understanding of gene mobility. The 'diagonal' genealogical links identified by an Alltown youth and community worker, when discussing sperm donation between fathers and sons, echoes these new geometries of kinship. Edwards' interest in what she terms the less pure side of the genealogical model, prompts her to examine the limits of the model, and the 'trickiness' with which actual and lived relationships cut across it (Edwards, 2009: p. 139).

'Young mums,' meeting weekly at a community center, commenting on the newspaper reports of widespread sperm donation by fathers to their sons, rejected as repugnant the scenario this conjured up for them – of sexual intimacy with their boyfriend's father, physical incorporation of his sperm, and the confusion of kinship roles (Edwards, 2009: p. 142). In a conversation with an older woman, the thinking through of complexity went a step further: what about boyfriend's mother's involvement in the negotiation? And what about the legal and emotional implications of being simultaneously father and grandfather? The confusion of generational roles and temptation to intervene in child rearing were seen as substantial obstacles to proper family relations. Problems were anticipated for the man who wanted to continue to help after the child's birth due to their double relationship, which would compound existing tensions between parents and grandparents over child raising.

While Edwards focuses on the way genealogy informs and is informed by relations forged through the new reproductive technologies, Fenella Cannell examines the uses of popular genealogy in remaking kinship relations with the dead (Cannell, 2011). Cannell emphasizes the social dimension of English amateur genealogists work, contrasting it both with the 'narcissistic' Parisian genealogists portrayed by Martine Segalen and Claude Michelat (2002) as being engaged in a compensatory activity and with American kinship, where the dead do not count as kin. Amateur genealogists in East Anglia, by contrast, invest considerable effort in researching their dead as part of a process of recuperation: knowledge allows the deceased to become actual rather than potential family members. While media dramatization of celebrity family histories – such as Jeremy Paxton's – gained mainstream viewing time, Cannell emphasizes the richness of off-screen genealogy: visiting houses, long rolls of paper trees, albums and boxes of photos, pamphlets about local history, and many stories – all of which create a space for restitution (Cannell, 2011: pp. 469–470). Although genealogical talk often seems to revolve around the self, Cannell does not see this as an 'attenuation accompanying modernity,' but rather a construction of the self through past others (Cannell, 2011: p. 473). She sees the access to databases and the Internet during the 1990s as expanding an existing interest. East Anglian genealogists can be seen as part of a democratizing trend, and a long tradition of asserting that history is not a possession of the privileged. Contrary to the idea of ancestral pedigree, popular genealogy, after World War I, can be seen as part of the shift toward democratic commemoration.

The underlying questions of morality to which both Edwards' diagonal genealogies and Cannell's popular ones refer,

demonstrate that genealogical thinking is more flexible and adaptable than was thought at the end of the twentieth century. One of the reasons why anthropologists have been so critical of the genealogical model as it has been used in anthropology is because its assumptions obscure the life worlds of the very people anthropologists claim to understand (see for example, Viveiros de Castro, 2009). Tim Ingold argues that the genealogical model profoundly misrepresents the way people we (anthropologists) class as indigenous "actually constitute their identity knowledgeability, and the environment in which they live" (Ingold, 2000: p. 133). In this, the genealogical model is deeply implicated in the discourse of the state; indeed, it is 'fundamentally a colonial model' (Ingold, 2000: p. 151). But what are the assumptions of the genealogical model to which Ingold refers?

Genealogical Assumptions

David Schneider was the first to draw attention to the assumptions underlying American kinship, which distinguishes between nature (blood) and law (marriage) (Schneider, 1968). His point was that this assumption colored anthropologists' understandings of kinship: taking for granted the facts of sexual reproduction and classifying relatives from there profoundly shaped the way non-western people's kinship and lives have been studied. An assumption about the genealogical unity of mankind was central to this process (Schneider, 1984). Early twentieth-century formalization of the genealogical method exemplifies how this perspective became embedded in anthropological method (Bouquet, 1993, 1995, 1996).

Ingold (2000: p. 136) points to the peculiarity of linear genealogy as a form that unfolds without regard to people's relationships due to the assumption that "original ancestry lies at the point where history rises from an ahistorical substrate of 'nature'; that the generation of persons involves the transmission of biogenetic substance prior to their life in the world; that ancestral experience can be passed on as the stuff of cultural memory, enshrined in language and tradition; and that the land is merely a surface to be occupied, serving to support its inhabitants rather than to bring them into being" (Ingold, 2000: p. 133).

Returning to the way Rivers' genealogical diagram 'upends' the potent image of the family tree (*cf* Bouquet, 1995, 1996), Ingold argues that the effect of this inversion is to replace the tree as a living, growing entity with "an abstract dendritic geometry of points and lines" (Ingold, 2000: p. 135). He asks, "What, exactly, is implied by this line? … [W]hat does it leave out?" (idem).

An assumption at the core of the genealogical model is that "individuals are specified in their essential genetic and cultural constitution – as genotype and culture type – in advance of their life in the world, through the bestowal of attributes from ancestors" (Ingold, 2009: p. 196). Since this transmission takes place at a remove from the surrounding environment, the genealogical model implies the transmission of 'vertically integrated classificatory knowledge.' Ingold proposes the reverse is also true: "the project of classification, combined with a principle of transmission by descent, generates the genealogical model" (*idem*).

Sandra Bamford and James Leach (2009: p. 2) argue that ideas about sequence, essence, and transmission behind the genealogical model "structure other modes of practice and knowledge making in domains that lie well beyond what we normally label 'kinship': … race, personhood, ethnicity, property relations, and the relationship between human beings and nonhuman species" (Bamford and Leach, 2009: p. 2). What does this mean for the genealogical method? Not all would agree with the harshly critical line taken on the comparative value of the genealogical method. Rita Astuti imagines W.H.R. Rivers enquiring into real versus classificatory kin among the Vezo of Madagascar (Astuti, 2009). She argues that much can be learned from the fact that people can make this distinction but are reluctant to do so. The Vezo ignore distinctions that are central to the method (such as parent–child bonds being distinct from all others), and their cultural practices obscure the pedigrees that the method is supposed to reveal – such as finding resemblances between parents and children (Astuti, 2009: p. 220). Astuti argues that there are two ways of interpreting these findings: either by asserting that Vezo ontology does not reveal a distinction between the 'facts of biology' and the 'facts of sociality,' and declaring Rivers' method inadequate; or by questioning whether Vezo cultural practice can be seen as reliable evidence of their 'ontological commitments.' The problem, as Astuti succinctly puts it, is that the method fails to capture the efforts through which these distinctions are made irrelevant (*idem*).

Vezo consciously play down the ties of biological kinship, to obscure pedigrees and to reduce the distinction between birth and nurture so as to build a community where children are 'generated, nurtured and moulded' through a wider circle of relations than the 'facts of biology' would admit (Astuti, 2009: p. 229). One of the unintended consequences of the genealogical method and its assumptions may thus be to reveal difference and complexity in a positive rather than a negative sense. These qualifications help to explain why anthropologists have continuing recourse to genealogy, despite – perhaps because of – the critical deconstruction of both method and model.

The Genealogical Imagination

What effect does the representational process of genealogy have upon the perceived rigidness or flexibility attributed to the model? As Cannell observed, "genealogy is a process by which the dead are brought back to some form of social life, the life of kinship, through the work and journeys (both literal and emotional) which the living descendants undertake on their behalf" (Cannell, 2011: p. 469). But what about people who keep, or kept, genealogies in other parts of the world? The close association between genealogy and the state, to which Ingold pointed, had implications for the Buryats, who migrated from Russian Buryatia to Mongolia in the early 1900s, and who were persecuted during the 1930s early Socialist period (Empson, 2007). During mass purges, almost all male Buryat were taken away, leading their survivors to conceal evidence of kin reckoning and personhood. Genealogical diagrams going back seven or eight generations were burned or concealed at the bottom of chests – along with Buddhist icons and shamanic implements (Empson, 2007: p. 119).

Rebecca Empson, in her analysis of the Buryat household chest, describes how photomontages came to replace

genealogies, which were displayed on each side or above the central mirror, as a means of representing kin (Empson, 2007: p. 120). She observes that unlike kin diagrams, Buryat genealogical representations start from a cluster of male descendants, expanding outward from a single founder in the middle or at the top of the page. The photographic montages mirror some of the compositional forms used to represent relations in traditional Buryat genealogies: a photograph of the patrilineal elder, with his wife, is placed at the center of the frame, surrounded by siblings and children, extending out to the frame. However there are significant differences: those linked in the photographic displays include both mother's and father's sides, school friends, people from summer pastures, colleagues from work, and people at other social events. Furthermore, the piling of photographs, with new ones constantly being added, means that the old ones at the back become obscured. 'Photographic montages thus replicate Buryat genealogical diagrams in their form, but extend their content to include other types of relations' (*idem*). These visible, infinite relations contrast with things hidden inside the chest (umbilical cords, locks of hair), which constitute a site for the subterranean relations between those of 'shared blood' – too close to live together but intimately connected by 'blood feelings.'

The exhibition 'Assembling Bodies: Art, Science and Imagination' at the Museum of Archaeology and Anthropology in Cambridge, 2009–10, staged a Mongolian Buryat chest based on Empson's research. Also on display was a *Bilum* Tree installation, designed by Marilyn Strathern, comprising netbags collected in Papua New Guinea between 1922 and 2005, strung plumply around a pole. Both installations powerfully materialized different ways in which genealogy is being re-thought, as well as the value of thinking through the things that go to make relations – and re-shape them (see Herle et al., 2009; Geismar, 2009).

By moving beyond the two-dimensional abstractions of written and diagrammatic genealogy into the complex three-dimensional ways genealogy is being materialized and transformed, in contemporary art as well as by science, this way of thinking and modeling the world is being put in its place. This place includes the three-dimensional space of exhibition and the assembly of materials there: as a space of modeling, experimentation, and reflection.

See also: Biomarkers in Demographic Research; Evolutionary Demography: A Darwinian Renaissance in Demography; Evolutionary Demography: A Synthesis of Two Population Sciences; Families and Households, Formal Demography of; Family and Kinship, History of; Family: The Anthropology of the Concept and Its History; Fertility and Culture: Anthropological Insights; Genetic Ancestry Testing; Kinship in Anthropology; Kinship, Evolution of.

Bibliography

Astuti, R., 2009. Revealing and obscuring Rivers's pedigrees: biological inheritance and kinship in Madagascar. In: Bamford, S., Leach, J. (Eds.), Kinship and Beyond. The Genealogical Method Reconsidered. Berghahn Books, New York & Oxford, pp. 214–236.

Bamford, S., Leach, J., 2009. Introduction: pedigrees of knowledge: anthropology and the genealogical method. In: Bamford, S., Leach, J. (Eds.), Kinship and Beyond. The Genealogical Method Reconsidered. Berghahn Books, New York & Oxford, pp. 1–23.

Barnard, A., Good, A., 1984. Research Practices in the Study of Kinship. Academic Press, London.

Barnes, J.A., 1967. Genealogies. In: Epstein, A.L. (Ed.), The Craft of Social Anthropology, Social Science Paperbacks in Association with Tavistock Publications, London, pp. 101–127.

Beer, G., 1983. Darwin's Plots, Evolutionary Narrative in Darwin, George Eliot and Nineteenth Century Fiction. Routledge & Kegan Paul, London.

Bouquet, M., 1993. Reclaiming English Kinship: Portuguese Refractions of British Kinship Theory. Manchester University Press, Manchester.

Bouquet, M., 1995. Exhibiting knowledge. The trees of Haeckel, Dubois, Jesse and Rivers at the *Pithecanthropus* centennial exhibition. In: Strathern, M. (Ed.), Shifting Contexts (Vol. 1 of the ASA Decennial Series, the Uses of Knowledge: Global and Local Relations). Routledge, London, pp. 31–55.

Bouquet, M., 1996. Family trees and their affinities. The visual imperative of the genealogical diagram. Journal of the Royal Anthropological Institute (N.S.) 2, 43–66.

Bourdieu, P., 1977. Outline of a Theory of Practice. Cambridge University Press, Cambridge.

Cannell, F., 2011. English ancestors: the moral possibilities of popular genealogy. Journal of Royal Anthropological Institute (N.S.) 17, 462–480.

Carsten, J., 1997. The Heat of the Hearth: The Process of Kinship in a Malay Fishing Community. Clarendon Press, Oxford.

Carsten, J. (Ed.), 2000. Cultures of Relatedness. New Approaches to the Study of Kinship. Cambridge University Press, Cambridge.

Cassidy, R., 2002. The Sport of Kings: Kinship, Class and Thoroughbred Breeding in Newmarket. Cambridge University Press, Cambridge.

Cassidy, R., 2009. Arborescent culture. Writing and not writing racehorse pedigrees. In: Bamford, S., Leach, J. (Eds.), Kinship and Beyond. The Genealogical Method Reconsidered. Berghahn Books, New York & Oxford, pp. 24–49.

Cunningham, H., 2009. Genes, mobilities and the enclosures of capital. Contesting ancestry and its application in Iceland. In: Bamford, S., Leach, J. (Eds.), Kinship and Beyond. The Genealogical Method Reconsidered. Berghahn Books, New York & Oxford, pp. 111–137.

Edwards, J., 2000. Born and Bred: Idioms of Kinship and New Reproductive Technologies in England. Oxford University Press, Oxford.

Edwards, J., 2009. Skipping a generation and assisting conception. In: Bamford, S., Leach, J. (Eds.), Kinship and Beyond. The Genealogical Method Reconsidered. Berghahn Books, New York & Oxford, pp. 138–158.

Empson, R., 2007. Separating and containing people and things in Mongolia. In: Henare, A., Holbraad, M., Wastell, S. (Eds.), Thinking through Things. Theorising Artefacts Ethnographically. Routledge, London & New York, pp. 113–140.

Feeley-Harnik, G., 1999. 'Communities of Blood': The natural history of kinship in nineteenth-century America. Comparative Studies in Society and History 41, 215–262.

Franklin, S., 1997. Embodied Progress. A Cultural Account of Assisted Conception. Routledge, London & New York.

Franklin, S., 2007. Dolly Mixtures: The Remaking of Genealogy. Duke University Press, Durham, NC.

Franklin, S., McKinnon, S. (Eds.), 2001. Relative Values. Reconfiguring Kinship Studies. Duke University Press, Durham, NC.

Geismar, H., 2009. Assembling Bodies: Art, Science and Imagination. Anita Herle and Mark Elliott, *Material World Blog*. http://www.materialworldblog.com/2009/06/assembling-bodies-art-science-and-imagination/ (accessed 09.07.13.).

Henare, A., 2007. *Taonga Māori*. Encompassing rights and property in New Zealand. In: Henare, A., Holbraad, M., Wastell, S. (Eds.), Thinking through Things. Theorising Artefacts Ethnographically, Routledge, London & New York, pp. 47–67.

Herle, A., Elliott, M., Empson, R., 2009. Assembling Bodies. Art, Science and Imagination, Exhibition Catalogue. Museum of Archaeology and Anthropology, Cambridge.

Holmes, J.T., 2009. When blood matters: making kinship in colonial Kenya. In: Bamford, S., Leach, J. (Eds.), Kinship and beyond. The Genealogical Method Reconsidered. Berghahn Books, New York, Oxford, pp. 50–83.

Ingold, T., 2000. The Perception of the Environment: Essays in Livelihood, Dwelling and Skill. Routledge, London.

Ingold, T., 2009. Stories against classification. Transport, wayfaring and the integration of knowledge. In: Bamford, S., Leach, J. (Eds.), Kinship and Beyond. The Genealogical Method Reconsidered. Berghahn Books, New York & Oxford, pp. 194–213.

Malinowski, B., 1922. Argonauts of the Western Pacific. Routledge & Kegan Paul, London.

Malinowski, B., 1930. Kinship. Man 30, 19–39.

Pálsson, G., 2009. The web of kin: an online genealogical machine. In: Bamford, S., Leach, J. (Eds.), Kinship and Beyond. The Genealogical Method Reconsidered. Berghahn Books, New York, Oxford, pp. 84–110.
Parkin, R., 1997. Kinship. An Introduction to the Basic Concepts. Blackwell, Oxford.
Read, D.W., 2007. Kinship theory: a paradigm shift. Ethnology 46, 329–364.
Rivers, W.H.R., 1900. A genealogical method of collecting social and vital statistics. Journal of the Royal Anthropological Institute 30, 74–82.
Rivers, W.H.R., 1910. The genealogical method of anthropological inquiry. The Sociological Review 3, 1–12.
Schneider, D.M., 1968. American Kinship: A Cultural Account. University of Chicago Press, Chicago.
Schneider, D.M., 1984. A Critique of the Study of Kinship. University of Michigan Press, Ann Arbor.
Shelton, A., 1994. Cabinets of transgression: renaissance collections and the incorporation of the new world. In: Elsner, J., Cardinal, R. (Eds.), The Cultures of Collecting. Reaktion Books, London, pp. 177–203.
Strathern, M., 1992. After Nature: English Kinship in the Late Twentieth Century. Cambridge University Press, Cambridge.
Viveiros de Castro, E., 2009. The gift and the given: three nano-essays on kinship and magic. In: Bamford, S., Leach, J. (Eds.), Kinship and Beyond. The Genealogical Method Reconsidered. Berghahn Books, New York & Oxford, pp. 237–268.
Zerubavel, E., 2012. Ancestors and Relatives. Genealogy, Identity, and Community. Oxford University Press, Oxford.

Gene–Culture Coevolution

Jeremy R Kendal, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by K. Aoki, volume 9, pp. 6084–6087, © 2001, Elsevier Ltd.

Abstract

Gene-culture coevolution theory explicitly accounts for the inheritance of genetic and cultural traits whose fitness and evolutionary trajectory are nonindependent. After describing how a gene-culture coevolutionary account is constructed, the framework is compared and contrasted against other evolutionary anthropology approaches to the study of human behavior. This is followed by highlighting a series of examples of gene-culture coevolution in human evolution, focusing largely on nutrition, disease, and the role of cumulative cultural evolution, which also shows how gene-culture coevolutionary mechanisms can affect both rates and evolutionary trajectories compared with a standard population genetic or cultural account. The approach is a powerful tool to understand both ancestral and contemporary human evolution.

Introduction

Coevolution refers to a form of evolution in which heritable traits affect each other's fitness. Within biology, famous cases include the antagonistic coevolution in either predator-prey or parasite-host systems, where an arms race can ensue as adaptations for predation or infection coevolve with countermeasures in the prey or host species. Most of these coevolutionary systems require evolution purely through genetic inheritance. By contrast, gene-culture coevolution concerns the evolutionary interaction of genetically and culturally inherited traits (Feldman and Laland, 1996). This requires that change in cultural variation across generations can be modeled as an evolutionary system, including mechanisms for the generation of variation, the inheritance of information, and selection (*see* Cultural Evolution: Overview).

Gene-culture coevolution occurs when there is a statistical association between the differential fitness of genetic and cultural traits. This means that the fitness of genetic and cultural variants is nonindependent. In such cases, the evolution of phenotypic variation is underspecified by either a purely genetic or a purely cultural model of evolution, and the appropriate unit of selection is the combination of genetic and cultural variant, called the *phenogenotype*.

Originally, analytically tractable gene-culture coevolutionary models were developed from population genetic theory. For instance, the multilocus population genetic model was modified to consider the coevolution of genetic and cultural traits in place of multiple genetic traits at separate loci (Feldman and Zhivotovsky, 1992). Adjustments have been made to the genetic models to account for biases in cultural transmission that are unusual or not commonly applicable in a purely genetic model (*see* Cultural Evolution: Theory and Models). For instance, cultural information that is vertically transmitted from parents to offspring may in some cases be subject to parental biases, as can be found in many sex-biased cultural preferences. Cultural information also can be transmitted obliquely between generations such as from teacher to pupils, or horizontally, within a generation. This information, however, often is short-lived, such as fashions and fads, and may not persist long enough to affect genetic selection.

A gene-culture coevolutionary model can be developed by explicitly accounting for (1) the transmission of genetic and cultural variants between generations, accounting for phenogenotype mating probabilities and specific pathways of cultural and genetic transmission; and (2) the differential fitness of phenogenotypes. From these two components, the phenogenotype frequencies can be iterated across generations. Analytical work has generated predictions that are qualitatively distinct from most population genetic models of evolution, on account of both characteristic rates and patterns of cultural transmission (Feldman and Laland, 1996). Typically, cultural variants can spread much faster than genetic variants, and this can have a particularly strong effect on genetic selection. Below, I consider several cases concerning nutrition and disease, where cultural traits have enhanced the intensity of genetic selection. In contrast, a fast cultural response to a changing environment, such as the invention of health interventions, also can buffer or protect otherwise-deleterious genetic variants from purifying selection.

Patterns of transmission also can drive coevolutionary dynamics, as exemplified by the case of sex-biased patterns of cultural transmission. These can become statistically associated with genetic variants that have sex-biased expression, including those on sex chromosomes. An example of this effect is found in the coevolution of a cultural preference for sons over daughters with sex-ratio distorter genes, which can affect sex at conception by altering the proportion or viability of sperm carrying an X or Y chromosome. The models suggest that different types of cultural bias can affect the gene-culture coevolutionary dynamics (Kumm et al., 1994). The first case concerns parents that are inclined to increase their proportion of sons by a fixed amount while compensating for their offspring lost to infanticide by having additional children. This might occur if restrictions on the number of permitted offspring make it unlikely for the cultural bias to affect the total number of offspring (Laland et al., 1995). In this situation, the excess of males favors the mating success of females over males. This provides a selective advantage to female-biased distorter genes, which then can evolve to give a female-biased primary sex ratio (i.e., the sex ratio at conception before sex-specific mortality) and an unbiased adult sex ratio. This result is consistent with standard sex-ratio

theory. The dynamics may be different if parents attempt to bring their offspring sex ratio closer to their ideal preference but without having more children in response to offspring loss. Here, the models suggest that the primary sex ratio may become biased until it matches the male-biased preference exhibited in the adult sex ratio. Thereafter, there would be no further sex ratio adjustment by parents as the primary sex ratio matches their ideal preference.

This analytical modeling approach is particularly useful for generating qualitative predictions and understanding the principles of evolutionary mechanisms. This type of model was developed before sequencing of the human genome, and development of a haplotype map (HapMap), which collates patterns of human genetic variation. Since then, the opportunity to analyze real genome sequences has identified a considerable number of human genes that appear to have undergone recent and rapid selection. In many cases, the function and estimated age of these genes appear consistent with culturally derived modification of their selective environment. These functions include metabolism, immune response, exposure to extreme climates, brain function, and skeletal development (Laland et al., 2010 Richerson et al., 2010;). Large simulation studies can be developed to reveal the most likely mechanisms underpinning the geographically distributed gene frequency data (e.g., Itan et al., 2009).

In the Context of Evolutionary Anthropology

Three broad approaches to the study of the evolution of human behavior can be distinguished by their assumptions (Laland and Brown, 2011). This section provides brief descriptions of each and shows how gene-culture coevolution fits in with these schemes. The descriptions will be stylized in their brevity, and it can be argued that new generations of researchers are drawing across these fields to address evolutionary questions appropriate for the context in which they work.

Perhaps, the most high-profile approach to study the evolution of human behavior has been evolutionary psychology, which puts an emphasis on the genetic evolution of psychological adaptations. Commonly, the assumption is made that behavior and expressed preferences, for instance, in mate choice, reflect psychology that was shaped by natural selection in ancestral environments. By contrast, human behavioral ecology enquiry is focused on the effects of contemporary ecological variation on adaptive behavior (*see* Human Behavioral Ecology). Here, two common assumptions are that human behavior is currently adaptive (maximizes inclusive fitness), and that according to the phenotypic gambit, the mechanism underpinning the behavior largely can be ignored. Thus, optimality is considered not to be constrained by mechanism but rather reflects an often-complex trade-off resulting from the interaction of marginal fitness benefits in a local ecology.

By distinction from these schools, gene-culture coevolution is principally consistent with the third approach, which is the field of cultural evolution (Cavalli-Sforza and Feldman, 1981; Boyd and Richerson, 1985; Mesoudi, 2011). This field focuses on the micro (within population) and macro (between population) patterns of spread of socially transmitted information, in the form of ideas, beliefs, behaviors, and material culture. The cultural evolution approach does not require the evolutionary psychologists' prior assumption of past adaptation or the behavioral ecologists' assumption of optimality. Instead, heritable traits are assumed to be subject to descent with modification, through some form of social transmission and innovation. The framework accounts for demic selection, considered equivalent to natural selection, by differential survival and reproduction, and cultural selection, which accounts for differential adoption and reversion from the use of a culturally acquired trait (Feldman, 2008). Other standard evolutionary features – such as drift, migration, and trait linkage – also can be considered as appropriate.

The modeling of cultural evolutionary or gene-culture coevolutionary dynamics is typically distinct from memetics in that there is no need to consider the socially transmitted trait as a replicator; however, there does need to be some effect of social learning, such that the probability that a learner acquires a trait is affected in some way by the behavior of another individual or its products (e.g., learning directly from material culture). One focus of cultural evolutionary theory is to consider the impact of individual-level learning mechanisms and cognitive biases on population-level patterns of cultural evolution.

In some regards, gene-culture coevolution can be seen to share common aspects with both evolutionary psychology and human behavioral ecology. Of interest to evolutionary psychology, gene-culture coevolution models have been used to examine the genetic evolution of predispositions for various biases in social learning cognition, including nonlinear frequency-dependent biases, such as conformity (to copy the majority) and payoff-related biases, such as copying successful individuals. Indeed, it is thought that the gene-culture coevolution of prestige bias can result in the cultural evolution of maladaptive behavior (Boyd and Richerson, 1985), a result consistent with evolutionary psychology and not human behavioral ecology. Consistent with human behavioral ecology, however, gene-culture coevolution commonly predicts the evolution of adaptive behavior in contemporary environments, where genetic variation continues to evolve in response to culturally derived selective environments (the cultural process of modifying the selective environment can be referred to as cultural niche construction; Kendal et al., 2011).

The gene-culture coevolutionary framework is also consistent with a relatively recent cross-disciplinary perspective on evolutionary theory interested in the role of so-called soft inheritance, or nongenetic inheritance, on the evolutionary processes. Nongenetic modes of inheritance have been carved up in a number of different ways. For instance, while Jablonka and Lamb (2005) have distinguished epigenetic, behavioral, and symbolic forms of inheritance, niche construction theorists have highlighted ecological inheritance of semantic information and resources, including forms of cultural inheritance (Odling-Smee et al., 2003).

Soft inheritance is considered important in genetic evolution as it accounts for genetic selective environments that are a function of the distribution of soft-inherited variants, which

is not, itself, predicted by contemporary or ancestral genetic variation (if it were, then, a purely population genetic model would be most parsimonious). Thus, although it may be perfectly reasonable to distinguish the orthogonal questions of why a trait evolved and how it develops within an individual, the gene-culture coevolutionary perspective deals with cases in which the most parsimonious model for why a genetic trait evolved requires explicit accounting for the soft inheritance of acquired characteristics during development.

Nutrition

Perhaps the most well-known case of gene-culture coevolution, that of dairy farming and lactase persistence, provides a good example of fast cultural change having a dramatic effect on genetic selection. In young mammals, milk is an extremely important source of nutrients, including the carbohydrate, lactose, which is broken down into its constituent sugars by the enzyme lactase. After weaning, the capacity to absorb lactose does not continue in most mammals. Approximately 35% of humans worldwide exhibit lactase persistence beyond weaning, reaching high frequencies across different regions including parts of India, Europe, and parts of Africa (Gerbault et al., 2011).

Lactase persistence in these separate regions appears to be associated with different genetic single nucleotide polymorphisms (SNPs) affecting promoter activity of the lactase gene itself. Thus, it would appear that lactase persistence has evolved more than once. Estimates for the age of these alleles coincide approximately with the Neolithic, and their rapid increase in frequency to contemporary levels indicates unusually high strengths of selection. For instance, simulation modeling suggests that the European mutation ($-13910*T$), now commonly more than 90% in Northern European populations, arose only 7500 years ago in the Central Balkans (Itan et al., 2009).

A principal advantage of drinking fresh milk is its high calorific value, which may be particularly valuable when other food sources are low. It is also possible that there are advantages in northern climes where ultraviolet radiation (UVR) levels are low and vitamin D in milk can facilitate calcium absorption, and in malarial regions, by disrupting the pathogen. Furthermore, it is possible that drinking milk may act as a sign of prestige where status was linked with reproductive success in relatively hierarchical societies.

A diverse range of evidence from archaeological, phylogenetic, and simulation studies lend support to the culture-historical hypothesis, that dairy farming preceded the spread lactase persistence, rather than the alternative reverse-cause hypothesis, that dairy farming evolved in response to high levels of lactase persistence (Gerbault, et al. 2011). For instance, milk residues in Neolithic archaeological pottery typically predates or coincides with a low frequency of lactase persistence. Gene-culture coevolution most likely followed the original culturally derived selection for lactase persistence, leading to a reciprocal effect for the spread of dairy farming among populations in which lactase persistence was high.

Unsurprisingly, gene-culture coevolution is common between farming practices and domesticates, and dairy farming is no exception. For instance, a high correlation has been shown for the genetic diversity of cattle milk proteins with both the frequency of human lactase persistence and the distribution of early European cattle pastoralists. The inference is that dairy farming by early pastoralists in different regions may have resulted in high allelic diversity that then was protected from drift by continued strong selection for milk yield and protein composition (Beja-Pereira et al., 2003).

In addition to domesticated animals, the cultural evolution of crop use also has affected the genetic evolution of human capacity to digest new dietary components, such as the case of high-starch diets. It appears that the number of copies of the gene for salivary amylase, *AMY1*, which catalyzes the breakdown of starch into sugars, correlates strongly with the proportion of starch in the diet. For instance, amylase copy number was found to be high in samples from Japanese and Hadza, whose diets are high in rice and tubers, respectively, but low in the pastoralist Turkic Yakuts inhabiting the Sakha Republic in the North East Russian Federation, which typically has a high fat and protein diet based on milk, meat, and blood (Perry et al., 2007).

The between-group differences in *AMY1* copy number are significantly greater than is typical at other loci that vary in copy number. The number of copies appears to correlate with gene expression, thus assisting the digestion of starch. It is likely that gene-culture coevolution would ensue in light of the potential adaptive benefits of a high-starch diet, including a source of nutrition and protection against intestinal disease.

Disease

The preceding diet-related examples involve culturally driven positive genetic selection, resulting in adaptive behavior. In other cases, however, dietary preferences can be extremely maladaptive, resulting in the selection of compensatory mutations to counteract the effects of the cultural behavior. In an extreme example, the prion disease, kuru, was transmitted between members of the Fore linguistic group in the Eastern highlands of Papua New Guinea through their ritual practice of consuming deceased relatives at mortuary feasts. Recent research suggests that the kuru epidemic resulted in the selection of a resistant G127V allele of the host prion protein (PRNP) gene (Mead et al., 2009). Strictly speaking, this example deviates slightly from the topic, as the protein-based, rather than nucleotide-based, replication system of the prion disease makes this a case of protein-culture coevolution.

Although ritual and social norms can be important contributors to the genetic evolution of pathogens and host resistance, arguably the greatest culturally derived impact has been through farming and domestication. A classic example concerns malarial resistance. Kwa-speaking West African populations clear areas of forest to plant yams (Durham, 1991). Holes in these clearings from the removed trees fill with water during rains and provide a suitable habitat for mosquitoes to lay eggs and for their larvae and pupae to grow in the sunlit pools. Thus, their farming practice encourages the growth of mosquito populations, which are vectors for *Plasmodium* protists. Once adult, the female mosquitoes inject the microorganisms into the human host

while feeding on blood, resulting in malarial infection in both the liver and red blood cells.

The sickle-cell allele *HbS* is not common because it causes anemia through sickling of the red blood cells. Sickled cells, however, confer some resistance against malaria by clearing the parasite along with the defective red blood cells to the spleen. In homozygotes, where both copies of the gene are *HbS*, this disease can be fatal, following symptoms, including restriction of blood flow round the body and an acute drop in hemoglobin levels. The heterozygote form, in which only one of the two copies is *HbS*, confers resistance to malaria with only mild sickling effects. The frequency of the *HbS* allele appears to correlate strongly with rainfall and surplus water among Kwa-speaking yam-cultivating populations but not among neighboring populations that do not cultivate yams (Durham, 1991). Thus, the culturally transmitted farming techniques would appear to have influenced genetic selection for resistance. Furthermore, the high incidence of malaria also may have enhanced cultural selection for the farming of yams as these crops appear to counteract the effects of sickling by helping the production of hemoglobin (O'Brien and Laland, 2012).

Although culturally transmitted activities, such as farming, have affected genetic disease resistance in the host, it is unsurprising to find culture coevolving with the pathogens themselves. These are special cases of host-parasite coevolutionary systems, where the host phenotype is culturally, rather than genetically, derived. For instance, heavy use of antimalarials, such as quinine and chloroquine, have resulted in selection for resistant strains of *Plasmodium*, and in reaction, the development of new efficacious drugs. In some cases, cultural evolution in this parasite-host arms race has exploited solutions in a parallel race between the parasite and plant hosts. Plants, such as fava beans (broad beans), which have evolved oxidizing action that kills the plasmodia, are cultivated and administered to protect against malaria. The exploited plant defense mirrors antimalarial mechanisms found in sickle-cell and similar disorders, such as glucose-6-phosphate dehydrogenase (G6PD) deficiency, which also exhibit elevated oxidation within infected erythrocytes (Etkin, 2003).

Antagonistic gene-culture coevolution is evident between normative administration of antibiotics and the genetic evolution of bacterial strain resistance. Typically, arms races can ensue between the invention of antibiotics, such as penicillin and then vancomycin, which provide selective environments for resistant strains, for instance, of Enterococcus. Although the resistance typically incurs a fitness cost that is absent in strains that are susceptible to antibiotics, this cost is insufficient to outweigh the fitness benefits of protection from these drugs.

The evolution of resistant strains can result not only in investment to develop new drugs, but also to alter normative prescribing practices, in an attempt to reduce the concentration of antibiotics in the environment. Models predict that when resistant-strain frequency is high, discouragement of high-prescribing practices can maintain bacterial strain polymorphism in cases in which it otherwise would not be expected. A strong conformist bias toward high antibiotic use, however, can lead to the fixation of the resistant strain. This work shows how cognitive biases affecting cultural transmission have population-level consequences for the trajectory of gene-culture coevolutionary dynamics (Boni and Feldman, 2005).

The gene-culture coevolutionary dynamics also can be affected by the trajectory of bacterial evolution. For instance, resistant strains can evolve compensatory mutations that retain their resistance but at a reduced fitness cost. Although antibiotic use encourages selection first for resistance and then for these new compensatory strains, a subsequent reduction in antibiotic use does not necessarily lead back to the evolution of the susceptible strain. This is because, in the absence of antibiotics, the fitness of both the compensatory and susceptible strains is greater than that of the original resistant strain. If the compensatory strain can only evolve back to the susceptible strain via an intermediate resistant but non-compensatory strain, the evolutionary trajectory would have to pass through what is effectively a valley in the fitness landscape (Bergstrom and Feldgarden, 2008). It has been shown that escape from the compensatory peak, through what is known as an 'evolutionary lobster-trap' (easy to get under high concentration of antibiotics but difficult to leave under low concentration of antibiotics) is possible under some conditions through a process called stochastic tunneling, which relies heavily on random sampling effects to succeed (Tanaka and Valckenborgh, 2011).

Cumulative Cultural Evolution

The examples discussed thus far typically are characterized by strong culturally derived selection over sufficient generations of the genetically evolving organism. This type of cultural dynamic is characteristic of a unique facet of evolution in the hominin lineage called cumulative cultural evolution. Here, learned information is passed across generations and developed so that cultural traits typically become more efficient or more complex and beyond the capacity for individual invention within a lifetime. A key feature of this process is that, in general, significant information is not lost between generations, but rather 'ratchets up' with its inheritance and modification across multiple generations (Tennie et al., 2009). This process can provide large changes in the selective environment for genetic evolution, which are sustained across sufficient generations for fast genetic evolution to occur.

In the absence of any ratcheting effect, a socially transmitted trait would risk dissolving away to extinction, particularly in variable environments. Thus, it is perhaps unsurprising not to find strong evidence for gene-culture coevolution in nonhuman primates, where cumulative culture appears to be absent, and thus culturally modified genetic selective environments may not be sustained across many generations in variable environments (*see* Social Learning and Culture in Nonhuman Organisms). In the hominin lineage, however, cumulative cultural evolution is likely to have ensured that the modified genetic selective environments are retained across generations, and further, that the magnitude of selection may be enhanced as the cumulative process proceeds.

Overall, a gene-culture coevolutionary dynamic that involves cumulative cultural change can undergo what is effectively a bootstrapping process, where the *rate* of cumulative cultural evolution is affected by coevolving genetic traits.

For instance, the cumulative cultural evolution of adaptive technology and societal norms in the hominin lineage may have affected, and been affected by, the capacity for social learning. The coevolution of social learning and adaptive information is integral to a number of theories exploring the evolution of human behavior, social systems, and technology, such as the cultural niche hypothesis (Boyd et al., 2011).

The capacity for the cumulative cultural evolution of technology has allowed humans to colonize diverse and extreme environments, resulting in genetic selection. For instance, populations living on the Tibetan and Andean plateaus carry several genes that are likely to have undergone selection in response to high-altitude hypoxia. Andeans have enhanced oxygen saturation compared with individuals from sea-level populations, but Tibetans appear to have elevated lung capacity, breathe more rapidly, and exhibit high levels of nitric oxide, which dilates blood vessels to enhance blood flow (Brown, 2012).

In response to migration and colonization of a diverse range of environments, gene-culture coevolution most likely contributed to the evolution of various external human phenotypes, such as hair color, skin color, and stature, through degrees of culturally affected sexual selection and natural selection. For instance, sexual dimorphism in skin pigmentation may have been enhanced in some populations by male cultural preference for lighter female skin (Jablonski and Chaplin, 2010). Arguably, however, a stronger predictor of skin pigmentation is latitude, or more specifically UVR, which surely was affected by culturally transmitted patterns of migration and the cultural evolution of tool kits allowing humans to have colonized novel environments. Dark and tanning skin most likely evolved to protect against UVR-mediated folate depletion, which can lead to fetal neural tube defects. By contrast, human migration to high latitudes may have resulted in selection for low pigmentation to enhance UVR-mediated vitamin D production in the skin, which has many important functions in the body, including bone metabolism and immune response (Jablonski and Chaplin, 2010).

The importance of cultural learning in hominin evolution has been situated within a broader context, referred to as the sociocognitive niche, which considers five principal components: 'cooperation, egalitarianism, mindreading, language, and cultural transmission' (Whiten and Erdal, 2012: p. 2119). The interactions between these facets often have been considered within a gene-culture coevolutionary framework. For instance, hypotheses have suggested a coevolutionary relationship between symbolic representation associated with material culture and technological development, and the genetic evolution of the parietal cortex, which can affect the capacity for language and motor-skills when using tools to solve tasks (Iriki and Sakura, 2008).

The capacity for cooperation is on many researchers' shortlist of key features affecting the hominin evolutionary trajectory (*see* Human Cooperation, Evolution of). For instance, ostensive sharing of learned information within and between societies that extends beyond close kin groups may catalyze the cumulative cultural evolution required for adaptive behavior in novel environments. One gene-culture coevolutionary proposal for the evolution of cooperation concerns the effect of between-group patterns of cultural evolution found in cultural group selection. The suggestion is that there can be cultural selection between groups, whose distinctions are retained by local social norms and cultural group markers, through differential normative practices affecting demographic and economic success, success in warfare, and prestige bias encouraging the adoption of norms of individuals from successful groups. These patterns of cultural evolution, which often may involve selection on adaptive cooperative norms, then may influence the genetic selection of predispositions that support such norms, such as prosociality, character virtues (e.g., honesty and trustworthiness), norm internalization, and conformity (Boyd et al., 2005; Henrich and Henrich, 2007).

Simulation models have shown that normative variation across groups can suppress genetic diversity at neutral loci and that this may explain relatively low-estimated Pleistocene hominin genetic diversity compared with that in extant great apes. Assuming both selective and neutral genetic and cultural loci, if there is a high threshold of cultural similarity required for migration between groups in a metapopulation, genetic diversity initially accumulates as cultural distinctions limit gene flow between groups. This is followed, however, by a sharp reduction in genetic diversity of neutral loci, as daughter groups of the most successful group sweep across the metapopulation (Premo and Hublin, 2009).

Conclusion

Gene-culture coevolution most likely played an important role in the evolution of modern humans and continues to do so today. In fact, it is possible that gene-culture coevolutionary dynamics that resulted from the advent of farming have yet to reach their zenith. Recent work suggests that although selective sweeps of beneficial mutations, spreading rapidly to fixation in a population, are not particularly frequent (Ryan et al., 2011), it is estimated that approximately three-quarters of all autosomal protein-coding single nucleotide variants, most of which were deleterious, evolved within the last 5000–10 000 years (Fu et al., 2013). Purifying selection (acting on deleterious alleles) has been unable to keep up in the face of culturally mediated exponential population expansion, from 5 million to 7 billion people, during this period. As a consequence, there is a high level of rare functional genetic variation – for instance, in relation to disease susceptibility – and the potential for adaptive evolution in our rapidly changing, culturally constructed, selective environments.

See also: Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Human Behavioral Ecology; Human Cooperation, Evolution of; Social Learning and Culture in Nonhuman Organisms.

Bibliography

Beja-Pereira, A., et al., 2003. Gene-culture coevolution between cattle protein genes and human lactase genes. Nature Genetics 35, 311–313.

Bergstrom, C.T., Feldgarden, M., 2008. The ecology and evolution of antibiotic resistant bacteria. In: Stearns, S., Koella, J. (Eds.), Evolution in Health and Disease, second ed. Oxford University Press, pp. 125–137.

Boni, M.F., Feldman, M.W., 2005. Evolution of antibiotic resistance by human and bacterial niche construction. Evolution 59, 477–491.

Boyd, R., Richerson, P.J., 1985. Culture and the Evolutionary Process. Chicago University press, Chicago.

Boyd, R., Gintis, H., Bowles, S., Richerson, P.J., 2005. The evolution of altruistic punishment. In: Gintis, H., Bowles, S., Boyd, R., Fehr, E. (Eds.), Moral Sentiments and Material Interests. The Foundations of Cooperation in Economic Life. MIT Press.

Boyd, R., Richerson, P.J., Henrich, J., 2011. The cultural niche: why social learning is essential for human adaptation. Proceedings of the National Academy of Sciences of the United States of America 108, 10918–10925.

Brown, E., 2012. High-Altitude-Hypoxia: Many Solutions to One Problem. Science in the News. Harvard University. sitn.hms.harvard.edu.

Cavalli-Sforza, L., Feldman, M.W., 1981. Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press, Princeton, New Jersey.

Durham, W., 1991. Coevolution: Genes, Culture and Human Diversity. Stanford University Press, Stanford.

Etkin, K.L., 2003. The co-evolution of plants, and parasites: biological and cultural adaptations to malaria. Proceedings of the Nutrition Society 62, 311–317.

Feldman, M.W., 2008. Dissent with modification: cultural evolution and social niche construction. In: Brown, M. (Ed.), Explaining Culture Scientifically. University of Washington Press.

Feldman, M.W., Zhivotovsky, L.A., 1992. Gene-culture coevolution: toward a general theory of vertical transmission. Proceedings of the National Academy of Sciences of the United States of America 89, 11935–11938.

Feldman, M.W., Laland, K.N., 1996. Gene-culture coevolutionary theory. Trends in Ecology and Evolution 11, 453–457.

Fu, W., et al., 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493, 216–220.

Gerbault, P., et al., 2011. Evolution of lactase persistence: an example of human niche construction. Philosophical Transactions of the Royal Society of London Series B Biological Sciences 366, 863–877.

Henrich, J., Henrich, N., 2007. Why Humans Cooperate. A Cultural and Evolutionary Explanation. Oxford University Press.

Iriki, A., Sakura, S., 2008. The neuroscience of primate intellectual evolution: natural selection and passive and intentional niche construction. Philosophical Transactions of the Royal Society of London Series B Biological Sciences 363, 2229–2241.

Itan, Y., et al., 2009. The origins of lactase persistence in Europe. PLoS Computational Biology 5 (8), e1000491.

Jablonka, E., Lamb, M.J., 2005. Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life. MIT Press.

Jablonski, N.G., Chaplin, G., 2010. Human skin pigmentation as an adaptation to UV radiation. Proceedings of the National Academy of Sciences of the United States of America 107, 8962–8968.

Kendal, J.R., Tehrani, J.J., Odling-Smee, J., 2011. Human niche construction in interdisciplinary focus. Philosophical Transactions of the Royal Society of London Series B Biological Sciences 366, 785–792.

Kumm, J., Laland, K.N., Feldman, M.W., 1994. Gene-culture coevolution and sex ratios: the effects of infanticide, sex-selective abortion, sex selection, and sex-biased parental investment on the evolution of sex ratios. Theoretical Population Biology 46, 249–278.

Laland, K.N., Kumm, J., Feldman, M.W., 1995. Gene-culture coevolutionary theory: a test case. Current Anthropology 36 (1), 131–156.

Laland, K.N., Brown, G.R., 2011. Sense and Nonsense, second ed. Oxford University Press, Oxford.

Laland, K.N., Odling-Smee, F.J., Myles, S., 2010. How nature has shaped the human genome: bringing genetics and the human sciences together. Nature Reviews Genetics 11, 137–148.

Mead, S., et al., 2009. A protective prion protein variant that colonizes with kuru exposure. The New England Journal of Medicine 361, 2056–2065.

Mesoudi, A., 2011. Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesise the Social Sciences. Chicago University Press, Chicago, IL.

O'Brien, M., Laland, K.N., 2012. Genes, culture and agriculture. An example of human niche construction. Current Anthropology 53, 434–470.

Odling-Smee, J.F., Laland, K.N., Feldman, M.W., 2003. Niche Construction: The Neglected Process in Evolution. Princeton University Press.

Perry, G.H. et al., 2007. Diet and the evolution of human amylase gene copy number variation. Nature Genetics 39 (10), 1256–1260.

Premo, L., Hublin, J.-J., 2009. Culture, population-structure, and low genetic diversity in Pleistocene hominins. Proceedings of the National Academy of Sciences of the United States of America 106, 33–37.

Richerson, P.J., Boyd, R., Henrich, J., 2010. Gene-culture coevolution in the age of genomics. Proceedings of the National Academy of Sciences of the United States of America 107, 8985–8992.

Ryan, D., et al., 2011. Classic selective sweeps were rare in recent human evolution. Science 331, 920–924.

Tanaka, M.M., Valckenborgh, F., 2011. Escaping an evolutionary lobster trap: drug resistance and compensatory mutation in a fluctuating environment. Evolution 65, 1376–1387.

Tennie, C., Call, J., Tomasello, M., 2009. Ratcheting up the ratchet: on the evolution of cumulative culture. Philosophical Transactions of the Royal Society of London Series B Biological Sciences 364, 2405–2415.

Whiten, A., Erdal, D., 2012. The human socio-cognitive niche and its evolutionary origins. Philosophical Transactions of the Royal Society of London Series B Biological Sciences 367, 2119–2129.

Gene–Environment Interactions in Health and Well-Being

Jason Schnittker, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Gene–environment interactions represent an important frontier in science and play an increasingly important role in forging interdisciplinary research. This article outlines what gene–environment interactions are, illustrates the different types of gene–environment interactions, distinguishes gene–environment interactions from gene–environment correlations, and reviews the findings of a number of exemplary studies. It also discusses some of the difficulties in studying gene–environment interactions and some of the reasons for lingering uncertainty in the empirical literature.

Enthusiasm for identifying the genetic determinants of health has quickly developed into the pursuit of discovering gene–environment interactions. The study of gene–environment interactions was hastened by the development of molecular genetics in the 1980s and 1990s and accelerated in recent years by the expansion of high-throughput genotyping technologies (Rutter et al., 2006). Although much of the early research sought to identify a single gene or small set of genes that caused a given condition, it quickly became apparent that genetic influences were much more complex and probabilistic (Kendler, 2005). Indeed, the additive effects of even genes that have survived replication are generally quite small, so that, for example, the odds of developing a psychiatric disorder increases by less than a factor of two (Kendler, 2005). Thus, efforts to identify 'a gene for' a given condition have largely disappeared.

Against this background, research has turned to the study of gene–environment interactions. The study of such interactions allows researchers to assess under what conditions genes do or do not matter, rather than how much they matter on average. In addition, the study of gene–environment interactions allows researchers to address a well-known feature of environmental risk: even for the strongest environmental risk factors, such as stress in the case of psychiatric disorder (Kessler, 1997) or smoking in the case of heart disease (Talmud et al., 2000), variation in response to these risks is enormous. Some individuals seem unusually robust to stress and see no increase in depression as a result, for example, whereas others are greatly affected by stress. Genes provide a potentially powerful explanation for this variation. At present, the gene–environment literature represents both an exciting and promising frontier and a site of ongoing controversy, much of it stemming from the uncertainties of common research designs.

Definition of a Gene–Environment Interaction

The study of gene–environment interactions is concerned with joint influences: with how variation in genes leads to different responses to environmental exposures or, conversely, how variation in environmental exposures leads to different effects of genes (Hunter, 2005; Ottman, 1996). In the study of interactions, the environment is defined broadly and can refer to proximate influences that are close to or within the individual, including chemical or biological exposures; more distal influences that are features of the social environment, including living in a high-poverty neighborhood or losing a job; or intermediary influences, such as stress, which link the environment and the person. The genetic element of the interaction is usually defined in terms of specific genes, but can also refer to generic and unspecified genetic influences, as in the case of research employing the classic twin design to examine variation in the degree of heritability as a function of measured environmental influences (Plomin et al., 2005). Because twin studies speak only to nonspecific genetic influences, researchers tend to regard molecular studies of gene–environment interaction as more compelling, although twin studies provide an important foundation to molecular research and, indeed, have set the stage for the modern study of interaction. Another method for detecting interactions is to explore interactions between a phenotype that is unambiguously the product of genes and an environment, as when research explores the effects of sunlight on skin cancer among fair- and dark-skinned persons, but interactions of this sort are necessarily focused because the number of phenotypes that are unequivocally genetic is small.

A generic definition of a gene–environment interaction does not describe the many varieties of interaction that are possible. In general, there are three types, and differentiating among them is an important if often-overlooked activity (Ottman, 1996; Ottman and Rao, 1990). The three types are represented in **Figure 1**, based on the framework of Ottman (1996). In Type A, genetic influences moderate the effects of an environmental factor, but genes have no influence among individuals not exposed to the relevant environmental factor. In Type B, the parameters are switched. An environmental factor moderates the influence of genes, but the environment has no direct effect on its own. In Type C, the presence of both the relevant genes and environments is necessary to realize an effect. Individuals without the relevant genes will not be affected even in environments conducive to the outcome, and individuals unexposed to the relevant environment will not be affected even if they have the relevant genes.

Within these three types of interaction, there are further refinements, as outlined by Shanahan and Hofer (2005). In these cases, the type of interaction reflects the magnitude and direction of the moderation. First, the environment may 'trigger' a genetic vulnerability, as when stress activates

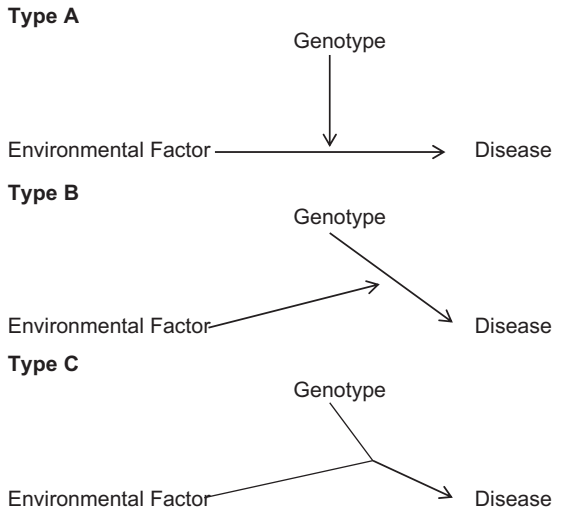

Figure 1 Three types of gene–environment interactions. Adapted from Ottman, R., 1996. Gene–environment interaction: definitions and study design. Preventive Medicine 25, 764–770.

a predisposition for depression as in the case of the diathesis–stress model (e.g., Rende and Plomin, 1992). Second, the environment may compensate for a genetic vulnerability, as when an enriched environment prevents a genetic predisposition for depression from materializing (e.g., Caspi et al., 2003). Third, the environment may reduce the impact of genetic factors by limiting opportunities, such as in the case of social control, as when some individuals are born with a genetic predisposition for alcoholism but laws, rules, or norms prevent alcohol consumption altogether or severely curtail its use (e.g., Koopmans et al., 1999). Fourth, the environment may accentuate positive genetic influences, as when genetic influences on educational attainment grow stronger as the educational system is opened up to a larger fraction of the population (e.g., Heath et al., 1985).

A gene–environment interaction is different from a gene–environment correlation, although both are features of gene–environment interplay. A gene–environment correlation pertains to when genetic influences are related to an individual's exposure to particular kinds of environments, as when an environmental exposure depends on genes. In this case, there is some genetic control over environmental exposure. For example, genes might increase exposure to stress by way of personality, which is at least partly heritable: a neurotic person may put themselves in more stressful situations. This implies an association between genes and environments: genes affect environments, but it does not imply that the effects of genes are contingent on the environment, as in a gene–environment interaction. Understanding correlations of this sort is important, but correlations reflect a different set of processes than interactions. A correlation can emerge in three ways: first, if individuals are put into environments that are related to their genes but not by virtue of their own behavior, as in the behavior of parents directed toward their biological offspring (a passive correlation); second, if individuals select themselves into an environment based on their genes (an active correlation); or, third, if individuals induce in others behavior that is related to their genes, as when a depressed person induces negative behavior from their spouse because of their depression (an evocative or reactive correlation) (Plomin et al., 1977). The range of environments affected by gene–environment correlations is quite large, reflected in the heritability of many measures of stress commonly regarded as 'environmental' (Kendler and Baker, 2007). Furthermore, there is substantial heritability even in more upstream measures of the environment, including socioeconomic status and social support (Schnittker, 2008a). Ideally, studies of gene–environment interactions should insure that the gene being studied is not correlated with the environment being studied or try to break any correlation using an appropriate research design (see Liu et al. (2004) on research designs to detect interaction). Nevertheless, the complexity of the processes behind gene–environment correlation makes this difficult, and the empirical evidence for gene–environment interactions is complicated as a result.

Empirical Evidence for Gene–Environment Interactions

There are abundant studies of gene–environment interactions now in the literature. Below is a summary of only a few such studies. These studies were selected because they either are emblematic of the literature or have been subject to widespread replication. These studies focus on psychiatric disorders, a topic for which the study of interactions has been particularly active (Hamer, 2002). Furthermore, these studies represent studies of humans. It is notable, however, that there is a parallel literature on gene–environment interactions among nonhuman animals, and these studies do not suffer from some of the research design problems discussed shortly. Studies of model organisms allow for strict manipulation of both environments and genotypes and can speak effectively to gene–environment interaction among humans (McClearn, 2004).

In a landmark investigation, Caspi et al. (2003) demonstrate an interaction between life stress and polymorphisms in the promoter region of the serotonin transporter gene (5-HTT). The study focuses on life stress and depression, but is expansive with respect to both concepts. With respect to stress, it explores interactions with both stressful life events and childhood maltreatment, and with respect to depression, it explores diagnosable depression, depressive symptoms, and suicidality. Caspi and colleagues find that stressful life events increase depression among those with a copy of a short allele of the 5-HTT promoter polymorphism, but not among those with two copies of the long allele (who made up 31% of the sample). On its own, the serotonin transporter gene has only a marginally significant main effect, although stress has a highly significant main effect, reflecting the genetic composition of the sample and the well-known effects of stress. The gene–environment interaction is specific to 5-HTT. Interactions with a related gene affecting monoamine oxidase A (MAOA) were insignificant.

The Caspi study received a good deal of attention and was widely embraced in scientific circles. *Science* declared it the second most important scientific discovery of 2003,

characterizing it as a 'huge catch' in a sea of potential interactions (Holden, 2009, p. 1628). The study received a good deal of media attention as well (Horwitz, 2005). There were several reasons for this enthusiasm. For one, the interaction was entirely plausible. The widespread use of serotonin-specific reuptake inhibitors in the treatment of depression and the well-established relationship between stress and depression gave the study a biologically credible mechanism wed to a clear environmental risk. In addition, the study helped resolve the uncertainty surrounding the main effect of 5-HTTLPR, which, to that time, had shown only inconsistent effects. The study fit quite well in the existing research, even as it subtly shifted the paradigm.

A host of related studies soon followed. Related studies looked at other disorders and different biological processes, but reached similar conclusions regarding the general importance of gene–environment interactions. Ducci et al. (2008), for example, explore an interaction between childhood sexual abuse and a functional MAOA locus, predicting alcoholism and antisocial personality disorder among adult women. They find a significant interaction, such that sexually abused women who were homozygous for the low-activity allele had higher rates of alcoholism and antisocial personality disorder than abused women who were homozygous for the high-activity allele. Among women who were not sexually abused, the MAOA genotype had no relationship with these disorders, revealing, in this way, an especially strong contingency.

Young et al. (2007) expand the potential range of interactions with the serotonin transporter gene 5-HTT. They do so by exploring its relationship with the limbic system and cognition, rather than a single specific psychiatric disorder. They find an elevated number of pulvinar (the medial part of the posterior end of the thalamus) neurons in subjects with two short copies of the 5-HTTLPR allele, which may heighten the impact of aversive emotional stimuli in the limbic system and, in turn, increase the risk for major depression, bipolar disorder, schizophrenia, and other psychiatric disorders. Indeed, the authors speculate that excessive levels of limbic activity increase fear, anxiety, and cognitive interference, setting the stage for a variety of interactions relevant to emotion and behavior.

The appeal of studying gene–environment interactions has extended to more traditional social sciences as well, including sociology, political science, and economics (e.g., Benjamin et al., 2012). Although traditional social sciences continue to use the twin design, molecular studies are appearing as well. In an article appearing in the *American Journal of Sociology*, for example, Pescosolido et al. (2008) explore interactions between the GABRA2 gene and assorted features of the environment in predicting alcoholism. Although generally interested in the effects of social position and resources, the authors test two specific interactions, both of which are significant. They find that family support attenuates the relationship between GABRA2 and alcoholism, while childhood deprivation accentuates it. Among those with low genetic risk, childhood material deprivation is unassociated with alcohol dependence, meaning the effect of childhood background depends entirely on a genetic predisposition. Family social support has a stronger relationship with alcohol dependence among those at high genetic risk, but the relationship is found even among those at low genetic risk. The main effect of GABRA2 is significant for the full sample, but further analyses show that this effect is apparent only among men and, in any case, is not terribly large. These assorted interactions are compelling as evidence for gene–environment interaction, but they are perhaps equally compelling for introducing some conceptual ambiguity. For instance, Pescosolido and colleagues find an interaction with family support, but not an interaction with friend support, implying something more than a generic effect of social support or stress in moderating genetic influences. The authors advance a network-based explanation, but future research will have to explore the mechanisms further.

Other studies do not explore interactions between specific genes and the environment, but rather explore the role of the environment in reducing or increasing the heritability of a phenotype, using the classic twin design. Such a design asks a question similar to that asked by molecular studies, but provides a different type of answer. In general, heritability is defined as the fraction of variance in a phenotype that can be explained by genetic variation in a population. Multivariate models allow researchers to ask whether this fraction is more or less dependent on other covariates. For example, research has explored the degree to which individual differences in cognitive ability can be explained by genetic influences more in low- or high-socioeconomic environments. In this vein, Tucker-Drob et al. (2011) find a significant gene–environment interaction that emerges as children develop. At age 10 months, genes account for a small amount of the variation in mental ability, and socioeconomic status is unrelated to ability. By the age of 2 years, however, there is a large gene–environment interaction. Among children in high-socioeconomic-status homes, genes account for nearly one-half of the variation in mental ability. Among children raised in low-socioeconomic-status homes, however, the environment is the dominant influence, with genes accounting for a very small fraction of the variance. Findings of this sort are consistent with the idea that genetic influences emerge in reciprocal interactions between an individual and her or his environment, but they call for a better understanding of the process. In the case of cognitive abilities, it is likely that low socioeconomic status impairs a child's ability to interact with environments that foster intellectual growth, resulting in suppressed genetic influence, whereas high socioeconomic status allows children to seek out environments consistent with their genetic endowments (see Scarr and McCartney, 1983).

Studies using the twin design are perhaps more common among social scientists than are molecular studies, which may explain their generally stronger engagement with the environment. Indeed, studies of this sort embrace an unusually wide assortment of environmental influences, including influences related to time period and culture. Boardman et al. (2010), for example, explore trends in the heritability of smoking and demonstrate how these trends can be explained in reference to the history of smoking-related knowledge and policies. In particular, they find strong genetic influences on smoking among those born in the 1920s, 1930s, and 1950s, but weak genetic influences among those born in the 1940s and 1960s. These birth cohorts map approximately onto certain historical events: the publication of the first Surgeon General's

Report in 1964 increased genetic influences, but subsequent smoking legislation decreased genetic influences. These countervailing effects reflect different processes. On one hand, the public acceptance of the Surgeon General's Report left the population of remaining smokers disproportionately reflective of those with a strong genetic proclivity to smoke, as those who smoke only in social settings will find it easier to quit in light of the recognition that smoking is harmful. On the other hand, legislation increased social control over smoking and, thereby, restricted variability in smoking among genetically diverse individuals (see Shanahan and Hofer, 2005). In general, genetic influences increase when a behavior becomes more widespread and social sanctions are reduced, which is what occurred in the case of smoking.

Related to this interpretation, Boardman et al. (2008) also show that genetic influences on smoking among adolescents are higher among those who attend schools where popular students smoke. A similar result was produced by Heath et al. (1989), albeit for a different outcome and a different environment. They find the heritability of alcohol consumption is lower in married than in unmarried women, reflecting differences in social control. Similar results have been found with respect to religious upbringing, which suppress genetic influences on alcohol consumption relative to nonreligious upbringings (Koopmans et al., 1999).

Findings of this sort can even be expressed with respect to the influence of culture. Dunne et al. (1997), for example, explore the heritability of sexual behavior and find differences between cohorts that can be accounted for by trends in sexual tolerance. Among those born in the early twentieth century, the heritability of age at first intercourse was relatively low, approaching zero in men. Among those born after 1950, however, the heritability of sexual behavior climbed enormously, reaching 72% for men. Similar results have been reported with respect to fertility, suggesting a process whereby a decline in social sanctions and improvements in birth control allow individuals to act on their genetically influenced preferences in a way that was not possible in earlier eras (Kohler et al., 2004).

Challenges in Identifying Gene–Environment Interactions

Although there is sufficient evidence to suggest that gene–environment interactions are ubiquitous and much of the evidence is compelling from the standpoint of social theory, there are numerous methodological challenges in studying interactions. The replication record of gene–environment studies is far from perfect. In a meta-analysis of the Caspi study, Risch et al. (2009) find little evidence for an interaction between 5-HTT and stressful life events, prompting some to encourage psychiatric genetics to go 'back to the drawing board' with respect to genes (Holden, 2009, p. 1628). In their meta-analysis, Risch and colleagues find a relationship between the number of stressful life events and depression, but no association between 5-HTTLPR and depression, either as a main effect (which Caspi and colleagues found as well) or as an interaction with stressful life events. Others have likewise concluded that evidence for an interaction is largely consistent with a chance finding (Munafò et al., 2009). These meta-analyses also suggest some limitations that were not anticipated in the initial study and are difficult to explain. For example, Eley et al. (2004) find a significant interaction, but only for girls. The conclusion of these meta-analyses is disputed. Some question the set of studies they use, which represent only a fraction of the total number of studies available. Analyses that include a wider set of studies find more evidence for an interaction (Karg et al., 2011). Regardless of whether the interaction is significant or not, it seems clear that the particular gene–environment interaction documented by Caspi and colleagues – and, indeed, virtually all other interactions documented in the literature – is sensitive to study design. The question is where exactly this sensitivity comes from and what it means. It is premature to disregard gene–environment interactions altogether, although it is important to reconsider what interactions reveal.

Part of the problem stems from our lack of understanding of genes themselves. The main effects of genes are generally not well understood, undermining the ability of researchers to evaluate the significance of interactions. In the case of psychiatric disorders, the effects of genes are often not replicated, even prior to testing interactions: genes with significant average effects in one study are not significant in another. This lack of replication reflects many things, but it at least partly reflects the presence of false negatives driven by underpowered studies (Lohmueller et al., 2003). The presence of weak effects in a small sample diminishes the capacity of researchers to evaluate interactions. Part of the problem also reflects how environments can be the product of genes. Exposure to stressful life events can be influenced by genes, and most studies of gene–environment interaction do not control for a sufficient number of genes to eliminate this possibility entirely. If the environment reflects the influence of genes, the apparent effects of stressful life events could reflect the effects of other unmeasured genes and, thus, the apparent gene–environment interaction could reflect a gene–gene interaction instead. In their study, Caspi and colleagues do, in fact, find a relationship between genes and stress: those with the short alleles that increase responsiveness to stressful life events also report more stressful life events. Caspi and colleagues make a logical case for a gene–environment interaction over a gene–gene interaction on the basis of time ordering and information regarding the onset of depression. They reason that if the gene–environment interaction they find reflects a gene–gene interaction instead, then they would find an interaction between 5-HTT and life events even for events that occurred after the onset of depression. However, they did not find this interaction in their data.

Some studies have attempted to subject gene–gene interactions to a more direct empirical test by using exogenous sources of environmental stress. In a sweeping study, Conley and Rauscher (2010) attempt to replicate several gene–environment interactions found in the literature using a combination of twin data, molecular information, and birth weight. The use of twins provides substantial leverage over unobserved genetic characteristics by providing a natural experiment: twin births effectively randomize a strong environmental influence, fetal position, which results in birth weight differences among genetically identical

twins who grow up in very similar environments. In this way, it is possible to use birth weight as an exogenous environmental insult within twin pairs, if one accepts that birth weight represents a conceptually meaningful product of the environment and not just a statistically exogenous one. Using this method, Conley and Rauscher find no evidence for gene–environment interactions or for the main effects of genes. In the case of their replication of the Caspi study, they find an interaction in the opposite direction. To be sure, birth weight is not stress per se – replication demands a strict mapping with the initial study – but it is an indicator of environmental insult that will, by design, be uncorrelated with genetic influences, an aspect of the study that is not possible in most observational studies.

Part of the problem with replication also stems from inconsistent measures of the environment. In reference to studies of stress and 5-HTT, Monroe and Reid (2008) characterize the measures of stress as 'ad hoc' (p. 947) and argue that the inconsistency this introduces is one reason for the apparent lack of reproducibility. Studies that measure stress in different ways do, indeed, reveal different interactions. For example, studies have produced stronger results when using objective measures of stressful life events rather than self-reports (Karg et al., 2011). Replication is also stronger for mildly threatening events than for severely threatening events (Kendler et al., 2005). Similarly, replication is stronger when exploring chronic stress than when exploring eventful stress (Karg et al., 2011). Even within the set of studies that use a simple count of the number of stressful events, there are inconsistencies that are difficult to explain. A review of studies attempting to replicate Caspi and colleagues' results found widely varying time frames for measuring stress, ranging from 1 month to lifetime measures (Monroe and Reid, 2008). The rapid decay of the effects of stress on depression introduces considerably heterogeneity between these measures. Interestingly, however, the intuitions that scholars have regarding gene–environment interactions may not align neatly with our understandings of how the effects of stress decay. In replicating the Caspi study, Wilhelm et al. (2006) use both a five- and one-year window for life events and find significant interactions only for the longer window, even though the main effects of stress operate over a shorter time frame. Without a well-characterized measure of stress wed to a robust theory of interaction, it is impossible to interpret these inconsistencies. Researchers may be tempted to conclude that gene–environment interactions are too complicated to be meaningful. This inconsistency also betrays the lack of engagement of the literature with the long-standing literature on stress. In fact, research knows a good deal about stress and how it can be measured (e.g., Brown and Harris, 1978). At present, this literature is not well integrated with the gene–environment literature.

The inconsistent measurement of stress reflects a more general problem. In general, the empirical characterization of the environment in gene–environment research is often abstract, which contrasts sharply with the very specific measurement of genes (Rutter et al., 2006, p. 251). Studies of the moderating influence of socioeconomic status, for instance, often use omnibus measures (i.e., measuring various factors together; Tucker-Drob et al., 2011) and, thereby, obscure variation in the effects of different components, such as income, education, and occupation. Related to this is a problem specific to studies documenting change in the heritability of a phenotype between cohorts. It is common to argue that an increase in the heritability of a phenotype such as fertility or sexual behavior reflects a decline in social control. This is a plausible interpretation, but studies rarely measure social control directly and, instead, infer it from historical events or social trends. Studies of this sort provide an invaluable set of observations and findings, but their interpretations are often incomplete in ways that prevent the development of a fully realized theory.

The Reception and Interpretation of Genetic Research

No review of gene–environment interactions would be complete without discussing the reception of gene–environment research in public opinion. The social and historical background to gene–environment research is, of course, controversial. Early eugenic interventions cast genetic influences in a deterministic fashion and linked genes to the inferiority of some populations (Duster, 2003). Although eugenic research has been discredited and the conceptual sophistication of the literature has advanced considerably, vestiges of the controversy remain in the distrust that some scientists maintain regarding genetic research (Rutter, 2006). Gene–environment research provides a potential antidote to this concern because it, in principle, puts genes and environment on the same footing. Yet there is also a lingering tendency to regard the influence of genes as somehow superordinate to those of the environment, especially in popular media. In particular, there is a tendency to regard gene–environment research as revealing more about genes than the environment, even if by definition an interaction means genes and environments cannot be interpreted apart from each other. In a review of media coverage of the Caspi study, for example, Horwitz (2005) finds a systematic emphasis on genetic influences rather than environmental influences, with much of the coverage referencing the possibility of finally discovering a gene for depression. Recognizing the interpretive difficulties in reporting gene–environment research, some scientists have encouraged standards for the presentation of results, disambiguating the influence of main and joint effects in tabular format (Botto and Khoury, 2001).

The public reception of genetic knowledge has its own complexities. For some advocates, the rise of research on genetic influences held the promise of finally reducing the stigma associated with certain health conditions, especially psychiatric disorders. The assumption was that identifying genetic influences would eliminate the idea that those who were afflicted brought the condition on themselves or were too weak to cope. Public opinion data reveal that the public is more receptive to genetic arguments than they were in the past, but that genes do little to obviate the public's insistence on personal responsibility. In particular, the public is increasingly likely to believe psychiatric disorders are caused by genes, in tandem with family background and stress, which they have long recognized as relevant (Schnittker, 2008b). This understanding is not inconsistent with the scientific understanding of psychiatric disorders. Nevertheless, the stigma surrounding psychiatric disorders has not changed

and for some psychiatric disorders, such as schizophrenia, the growing acceptance of genetic arguments may have accentuated the stigma by making those with schizophrenia appear even more dangerous. In the area of gene–environment interactions, there is much left to learn.

See also: Comparative Method in Evolutionary Studies; Cultural Evolution: Overview; Evolution and Language: Overview; Evolution and Language: Phylogenetic Analyses; Language Families, Archaeology and History of.

Bibliography

Benjamin, D.J., Cesarini, D., Chabris, C.F., Glaeser, E.L., Laibson, D.I., Guðnason, V., Harris, T.B., Launer, L.J., Purcell, S., Smith, A.V., Johannesson, M., Magnusson, P.K.E., Beauchamp, J.P., Christakis, N.A., Atwood, C.S., Hebert, B., Freese, J., Hauser, R.M., Hauser, T.S., Grankvist, A., Hultman, C.M., Lichtenstein, P., 2012. The promises and pitfalls of genoeconomics. Annual Review of Economics 4, 627–662.

Boardman, J.D., Blalock, C.L., Pampel, F.C., 2010. Trends in the genetic influences on smoking. Journal of Health and Social Behavior 51, 108–123.

Boardman, J.D., Saint Onge, J.M., Haberstick, B.C., Timberlake, D.S., Hewitt, J.K., 2008. Do schools moderate the genetic determinants of smoking? Behavioral Genetics 38, 234–246.

Botto, L.D., Khoury, M.J., 2001. Commentary: facing the challenge of gene-environment interaction: the two-by-four table and beyond. American Journal of Epidemiology 153, 1016–1020.

Brown, G.W., Harris, T.B., 1978. Social Origins of Depression: A Study of Psychiatric Disorder in Women. Tavistock Publications, London.

Caspi, A., Sugden, K., Moffitt, T.E., Taylor, A., Craig, I.W., Harrington, H., McClay, J., Mill, J., Martin, J., Braithwaite, A., Poulton, R., 2003. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science 301, 386–389.

Conley, D., Rauscher, E., 2010. Genetic Interactions with Prenatal Social Environment: Effects on Academic and Behavioral Outcomes. National Bureau of Economic Research, Cambridge, MA.

Ducci, F., Enoch, M.A., Hodgkinson, C., Xu, K., Catena, M., Robin, R.W., Goldman, D., 2008. Interaction between a functional MAOA locus and childhood sexual abuse predicts alcoholism and antisocial personality disorder in adult women. Molecular Psychiatry 13, 334–347.

Dunne, M.P., Martin, N.G., Statham, D.J., Slutske, W.S., Dinwiddie, S.H., Bucholz, K.K., Madden, P.A.F., Heath, A.C., 1997. Genetic and environmental contributions to variance in age at first sexual intercourse. Psychological Science 8, 211–216.

Duster, T., 2003. Backdoor to Eugenics. Routledge, New York.

Eley, T.C., Sugden, K., Corsico, A., Gregory, A.M., Sham, P., McGuffin, P., Plomin, R., Craig, I.W., 2004. Gene-environment interaction analysis of serotonin system markers with adolescent depression. Molecular Psychiatry 9, 908–915.

Hamer, D., 2002. Rethinking behavior genetics. Science 298, 71–72.

Heath, A.C., Berg, K., Eaves, L.J., Salaas, M.H., Corey, L.A., Sundet, J., et al., 1985. Education policy and the heritability of educational attainment. Nature Reviews Genetics 314, 734–736.

Heath, A.C., Jardine, R., Martin, N.G., 1989. Interactive effects of genotype and social environment on alcohol consumption in female twins. Journal of Studies in Alcohol 50, 38–48.

Holden, C., 2009. Back to the drawing board for psychiatric genetics. Science 324, 1628.

Horwitz, A.V., 2005. Media portrayals and health inequalities: a case study of characterizations of gene X environment interactions. Journal of Gerontology Series B: Psychological Sciences and Social Sciences 60, S48–S52.

Hunter, D.J., 2005. Gene-environment interactions in human disease. Nature Reviews Genetics 6, 287–298.

Karg, K., Burmeister, M., Shedden, K., Sen, S., 2011. The serotonin transporter promoter variant (5-HTTLPR), stress, and depression meta-analysis revisited: evidence of genetic moderation. Archives of General Psychiatry 68, 444–454.

Kendler, K.S., 2005. 'A gene for…': the nature of gene action in psychiatric disorders. American Journal of Psychiatry 162, 1243–1252.

Kendler, K.S., Baker, J.H., 2007. Genetic influences on measures of the environment: a systematic review. Psychological Medicine 37, 615–626.

Kendler, K.S., Kuhn, J.W., Vittum, J., Prescott, C.A., Riley, B., 2005. The interaction of stressful life events and a serotonin transporter polymorphism in the prediction of episodes of major depression: a replication. Archives of General Psychiatry 62, 529–535.

Kessler, R.C., 1997. The effects of stressful life events on depression. Annual Review of Psychology 48, 191–214.

Kohler, H.-P., Rodgers, J.L., Christensen, K., 2004. Is fertility behavior in our genes? Findings from a Danish twin study. Population and Development Review 25, 253–288.

Koopmans, J.R., Slutske, W.S., van Baal, G.C.M., Boomsma, D.I., 1999. The influence of religion on alcohol use initiation: evidence for genotype X environment interaction. Behavior Genetics 29, 445–453.

Liu, X., Fallin, M.D., Kao, W.H.L., 2004. Genetic dissection methods: designs used for tests of gene–environment interaction. Current Opinion in Genetics & Development 14, 241–245.

Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S., Hirschhorn, J.N., 2003. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nature Genetics 33, 177–182.

McClearn, G.E., 2004. Nature and nurture: interaction and coaction. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 124B, 124–130.

Monroe, S.M., Reid, M.W., 2008. Gene-environment interactions in depression research. Psychological Science 19, 947–956.

Munafò, M.R., Durrant, C., Lewis, G., Flint, J., 2009. Gene × environment interactions at the serotonin transporter locus. Biological Psychiatry 65, 211–219.

Ottman, R., 1996. Gene–environment interaction: definitions and study design. Preventive Medicine 25, 764–770.

Ottman, R., Rao, D.C., 1990. An epidemiologic approach to gene-environment interaction. Genetic Epidemiology 7, 177–185.

Pescosolido, B.A., Perry, B.J., Long, J.S., Martin, J.K., Nurnberger Jr., J.I., Hesselbrock, V., 2008. Under the influence of genetics: how transdisciplinarity leads us to rethink social pathways to illness. American Journal of Sociology 114, S171–S201.

Plomin, R., DeFries, J.C., Loehlin, J.C., 1977. Genotype-environment interaction and correlation in the analysis of human behavior. Psychological Bulletin 84, 309–322.

Plomin, R., DeFries, J.C., McClearn, G.E., McGuffin, P., 2005. Behavioral Genetics. Worth Publishers, New York.

Rende, R., Plomin, R., 1992. Diathesis-stress models of psychopathology: a quantitative genetic perspective. Applied and Preventive Psychology 1, 177–182.

Risch, N., Herrell, R., Lehner, T., Liang, K.-Y., Eaves, L., Hoh, J., Griem, A., Kovacs, M., Ott, J., Merikangas, K.R., 2009. Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression. The Journal of the American Medical Association 301, 2462–2471.

Rutter, M., 2006. Genes and Behavior: Nature-Nurture Interplay Explained. Blackwell Publishing, Malden, MA.

Rutter, M., Moffitt, T.E., Caspi, A., 2006. Gene-environment interplay and psychopathology: multiple varieties but real effects. Journal of Child Psychology and Psychiatry 47, 226–261.

Scarr, S., McCartney, K., 1983. How people make their own environments: a theory of genotype → environment effects. Child Development 54, 424–435.

Schnittker, J., 2008a. Happiness and success: genes, families, and the psychological effects of socioeconomic position and social support. American Journal of Sociology 114, S233–S259.

Schnittker, J., 2008b. An uncertain revolution: why the rise of a genetic model of mental illness has not increased tolerance. Social Science & Medicine 67, 1370–1381.

Shanahan, M.J., Hofer, S.M., 2005. Social context in gene–environment interactions: retrospect and prospect. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 60, 65–76.

Talmud, P.J., Bujac, S.R., Hall, S., Miller, G.J., Humphries, S.E., 2000. Substitution of asparagine for aspartic acid at residue 9 (D9n) of lipoprotein lipase markedly augments risk of ischaemic heart disease in male smokers. Atherosclerosis 149, 75–81.

Tucker-Drob, E.M., Rhemtulla, M., Harden, K.P., Turkheimer, E., Fask, D., 2011. Emergence of a gene × socioeconomic status interaction on infant mental ability between 10 months and 2 years. Psychological Science 22, 125–133.

Wilhelm, K., Mitchell, P.B., Niven, H., Finch, A., Wedgewood, L., Scimone, A., Blair, I.P., Parker, G., Schofield, P.R., 2006. Life events, first depression onset and the serotonin transporter gene. The British Journal of Psychiatry 188, 210–215.

Young, K.A., Holcomb, L.A., Bonkale, W.L., Hicks, P.B., Yazdani, U., German, D.C., 2007. 5HTTLPR polymorphism and enlargement of the Pulvinar: unlocking the backdoor to the limbic system. Biological Psychiatry 61, 813–818.

Gene–Environment Interplay

Jonathan Daw and Jason D Boardman, University of Colorado Boulder, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although the social and behavioral sciences were long divided over the relative sway of nature and nurture, an emerging field of inquiry, the study of gene-environment interplay, seeks to undermine these distinctions by looking at how these influences on social and behavioral outcomes are interdependent. This article provides a brief overview of three forms of gene-environment interplay: additive genetic and environmental effects, gene-environment correlations, and gene-environment interactions. A distinction is drawn between statistical genetic and molecular genetic models of gene-environment interactions. Finally, a variety of models of gene-environment interactions are described, and the limitations of current research on gene-environment interplay are briefly assessed.

Introduction

The days in which the social and behavioral sciences were polarized by debates concerning the sway of nature and nurture over human behavior are, and ought to be, long past. It is beyond serious debate that many most behaviors are influenced by genetic factors, and it is equally obvious that few behaviors and traits of interest to social scientists are solely biologically determined. It was not always thus. Sociologists have long concerned themselves with the intergenerational transmission of behavioral traits (e.g., similarity of smoking among grandfathers, fathers, and sons) without making serious efforts to link this common behavior to common genes within families. In parallel, biological scientists frequently argued with little persuasive evidence that biological influences on behavior were responsible for the majority of variation in human behaviors.

Rather than torturously resurrecting these tired debates, this article provides an overview of a nascent yet exciting field of inquiry: gene–environment interplay. The field of gene–environment interplay concerns itself with the myriad ways in which genetic and environmental processes function interdependently. That is, it is rarely the case that one can account for differences in behaviors because of nature or nurture. In fact, it is increasingly obvious that nearly all human traits are influenced by nature and nurture. This approach has grown exponentially in popularity in the past 20 years. Yet it is not a mere academic fad. Rather, the promise of gene–environment interplay research is that it provides a framework for thinking about social and behavioral science that is integrative, not oppositional, and based on evidence, not assumptions. This approach to behavioral and social science has the potential to bring together valuable lessons from sociology, demography, economics, anthropology, political science, genetics, development, and neuroscience to provide a framework for thinking about humans as both biological and social beings. Although it will likely not be the dominant framework in any other subfield of social science, it has the potential to inform the thinking of all fields of social scientific inquiry.

For example, consider the results of a paper by Boardman et al. (2010) that are summarized in **Figure 1**. In this study, the researchers used samples of adult twins to estimate the extent to which genes influenced the likelihood that an individual would become a regular smoker by the age of 30. As shown in this figure, genetic factors were very important for the earliest cohorts but became increasingly less relevant. For those born in 1942, genes had virtually nothing to do with whether or not they became regular smokers. This birth cohort was also at the prime age of smoking onset when the 1964 Surgeon General's report on smoking was released. Following the release of this report genes become more important over time because the composition of smokers began to change such that entry into smoking status was increasingly among those for whom smoking onset was being pushed by genetic factors. However, this changes noticeably when increasingly restrictive social policies were instituted the limited entry among many individuals such

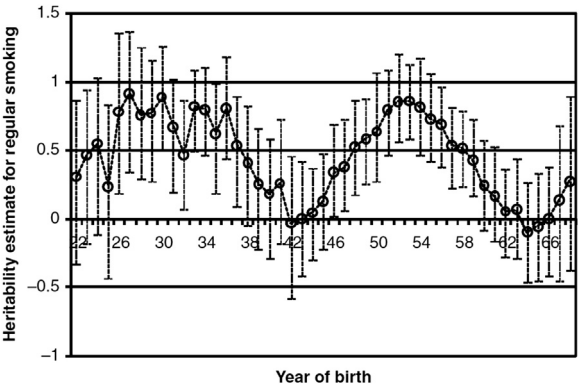

Figure 1 Birth cohort differences in the genetic effects on regular smoking (Boardman et al., 2010). *Note*: Data come from the 1995 National Survey of Midlife Development in the United States (MIDUS) (Brim et al., 1996). Estimates are obtained from a series of multivariate Cox regression models with shared frailty among twin pairs, and they describe the heritability of regular smoking for each year of birth. The bars around each point estimate describe the 95% confidence intervals for each year. These confidence intervals were bootstrapped from 200 runs for each zygosity using the coxph package for R 2.7.1 for all years between 1922 and 1968. Each birth year contains the 4 years before and after and thus describes a 9-year window for the estimated genetic influences on smoking at that time.

that genes, again, became less relevant for entry into the smoking status. This example illustrates that it is very difficult to ever say for certain if the reason that people smoke is because of nature or nurture. The relative influence of each factor is always changing so your answer depends on when, where, and for what behavioral trait one is interested in.

Genetic Research Terminology

First, we provide some very brief background in the terminology of genetics for those who are not familiar with this field. The human *genome* consists of an ordered series of paired *nucleotides* in two matching strings, arrayed in a double helix as deoxyribonucleic acid, or DNA, in 23 *chromosomes*, or discrete strands of DNA. As shown in Figure 2, cells in the human body contain each of these 23 chromosomes. There are four nucleotides in the human genome: adenine (A), thymine (T), guanine (G), and cytosine (C). These are paired together determinatively, such that A is always paired with T and G with C. A *polymorphism* is a section of the genome in which variations in this base pair sequence are found between humans; a *single nucleotide polymorphism* (SNP) occurs when this consists of a single base pair difference. The different proteins are illustrated by the different colors in Figure 2. This variation in the human genome produces *alleles*, or DNA sequences which vary among humans. The observed combination of alleles at a location in one's genome is known as one's *genotype* at that location. Genetic researchers frequently seek to link genotypic variation to variation in *phenotypes*, which is the biological term for an organism's characteristics or traits. Phenotype is a synonym for 'outcome variable' in social scientific research.

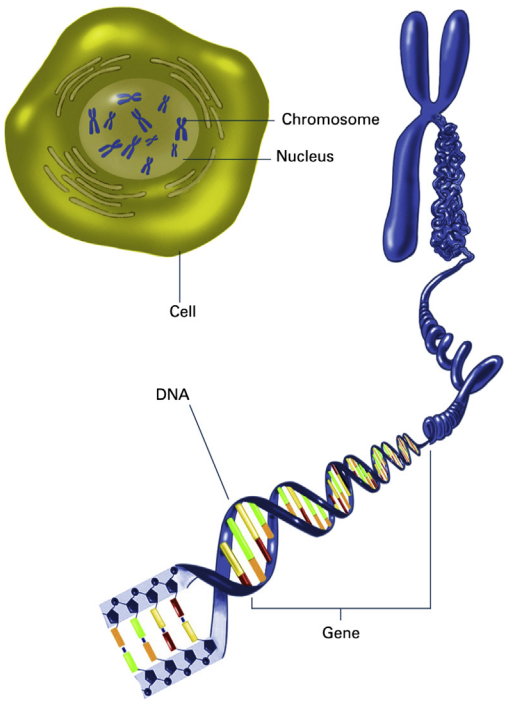

Figure 2 Structure of the human genome: the key components. *Note*: This figure was taken from the collection of figures organized by the National Institute of General Medical Sciences: http://images.nigms.nih.gov/index.cfm?event=doSearch&searchTerm=genetics&typeID=&sizeID.

Behavioral genetics is a field of behavioral science which attempts to summarize the separate contributions of genetic and environmental influences on a given phenotype. Frequently behavioral genetic researchers attempt to decompose this variance into three sources: (1) *heritability*, which describes the proportion of variance in an outcome that is thought to be explained by all genetic influences; (2) *shared environments*, which characterizes the environmental influences which make pairs of individuals more phenotypically similar; and (3) *unshared environments*, which characterizes the environmental influences which make pairs of individuals more phenotypically dissimilar. Behavioral genetics studies are frequently conducted using twin studies (which compares the phenotypic similarity of monozygotic twins, who are genetically identical, and dizygotic twins, who share 50% of their genes on average by descent) and adoption studies (which compares the phenotypic similarity of adopted sibling pairs to that of nonadopted members of their biological family).

When genetic information is available from subjects in a study, researchers can do two types of research. The first is called a *candidate gene* study in which individuals are classified into groups in which they are comparable to one another across segments of the human genome that are responsible for known physiological functions. These different groups are then compared with one another to see whether they are different with respect to some behavior like smoking, height, etc. There are many different statistical techniques for *candidate gene* studies, but using the correct method, researchers can conclude that the gene is actually responsible for differences in the trait and accounts for the similarity of family members who are much more likely to share the gene than two unrelated individuals. *Genome-wide association studies* (GWAS) incorporate data on a large number of SNPs (typically between 100 000 and 2 000 000 at this time) in order to identify loci associated with a phenotype of interest. Whereas the *candidate gene* method is a deductive exercise in which a specific hypothesis is tested, GWAS methods are entirely inductive and are much better characterized as a discovery science. In essence, GWAS models compare those with a specific illness (e.g., cancer) to those who do not have this illness (in which large efforts are made to provide the most meaningful comparisons) across each SNP in the human genome. Currently, genome-wide genotyping provide nearly 2.5 million SNPs. The strongest associations that emerge from this data mining exercise are then investigated to see whether they have any physiological pathway through which they might actually affect the trait of interest.

Additive Genetic and Environmental Effects

Sometimes both biological and environmental processes influence an outcome, but these influences are not interdependent. In other words, a given biological condition will have a certain influence on an outcome regardless of the environmental condition, and vice versa. Although this situation is not

traditionally grouped with gene–environment interplay, it is nonetheless an instance where considering biology and environmental influences jointly is potentially important. For instance, suppose that the odds of experiencing a myocardial infarction (heart attack) conforms to a threshold model, wherein a heart attack occurs when sufficient stress has been placed on the cardiovascular system from any source. In such an instance, one might suppose that any combination of genetic and environmental influences might combine non-interactively to produce the disease outcome in question.

To take an empirical example, consider the findings of Daw and Guo (2011) on the odds of an adolescent reporting contraceptive usage in their most recent sexual encounter. These researchers found that the odds of contraceptive nonuse were strongly linked to the dopamine transporter gene *DAT1*, the dopamine receptor gene *DRD2*, and the monoamine oxidase gene (*MAOA*). Furthermore, using multivariate logistic regression techniques, they found that these results held true when a wide range of social environmental and attitudinal variables that had previously been linked to contraceptive behavior were controlled. Finally, they found that these genetic and environmental influences on contraceptive behavior were noninteractive, although the genetic associations did vary somewhat by gender. In other words, environmental and attitudinal determinants of contraceptive behavior were independent from the genetic determinants. Recognizing that unprotected sex can occur as a function of both environmental *and* genetic influences was, however, an unprecedented contribution to social scientific understanding of this important outcome.

Gene–Environment Correlations

Gene–environment correlations (rGE) occur when genetic factors are associated with the causally important aspects of an environment for a given outcome. This can create the appearance of a genetic association with a phenotype where none exists – for instance, if gene G is correlated with environment E which is associated with phenotype P, a regression of P on G without a control for E will result in the appearance of an association of G with P. This is statistically identical to the more familiar problem of statistical confounding.

However, gene–environment correlations are also a substantively interesting social process. For instance, Price and Jaffee (2008) found that children whose parents have lower verbal abilities are reared in environments with more household disorganization, which is associated with children's intellectual development. In other words, children's verbal abilities resemble their parents' in part because of the association of genes linked with verbal ability with household disorganization as well as those children's inheritance of those same genes. This is known as a *passive gene–environment correlation*, in which environments which one did not select or evoke are correlated with one's genes and influence one's phenotypes.

Another form of gene–environment correlation occurs when one's genetically linked traits elicit different environmental influences on one's phenotype. This is known as an *evocative gene–environment correlation*. This type of rGE is rarely studied in humans, but one can conjure some hypothetical examples. Suppose, for instance, that one's physical attractiveness is partially genetically influenced, which appears to be the case. (For instance, in the National Longitudinal Study of Adolescent Health (Add Health), interviewers were asked to rate respondents' physical attractiveness on a 1–5 scale. The correlation of this measure among monozygotic twins is 0.620, and the correlation among dizygotic twins is 0.330. According to the standard twin model calculation, this indicates that interviewer-rated attractiveness is $2*(r_{MZ} - r_{DZ}) = 2(0.620 - 0.330) = 0.58$ heritable, a relatively high value.) Suppose further that one's physical attractiveness is positively associated with one's popularity, which also appears to be the case. (In Add Health, the correlation between one's interviewer-rated physical attractiveness and the number of others at one's school who nominate one as a friend is 0.158.)

Finally, suppose that one's self-esteem is positively associated with one's popularity, which also appears to be the case, albeit only moderately. (In Add Health, they ask how often, in the past week, the respondent felt that they 'were just as good as other people' on a 0–3 scale, which we take as a crude measure of self-esteem. This measure has a polychoric correlation of 0.059 with popularity.) In this case, one could argue that genetic influences on physical attractiveness have an evocative rGE relationship with self-esteem due to their correlation with popularity. Although this evidence does not prove that this is the case, it illustrates how this form of gene–environment interplay operates.

A final form of gene–environment correlation occurs when individuals select into different types of environments as a result of their genetic influences. This is known as an *active gene–environment correlation*. An example was identified by Cleveland et al. (2005), who found evidence for genetic associations with having friends who smoke and drink. Because this environment is associated with subsequent adoption of smoking and drinking behaviors, this is likely an example of an active rGE.

Gene–Environment Interactions

A *gene–environment interaction* describes a statistical relationship in which the effects of a gene G on a phenotype P depends on the value of an environment E. (It is equally correct to say that the environmental effect depends on the genotype.) For example, perhaps the most famous recent gene–environment interaction finding is that by Caspi et al. (2003), who found that the effects of stressful life events on depressive symptomology were conditional on an individual's genotype for the serotonin transporter gene *5-HTT*, such that those with more short alleles showed greater depressogenic effects. This is shown in **Figure 3** in which all individuals are susceptible to the effects of stress exposure on their mental health (e.g., all three lines have positive slopes) but the s/s group (e.g., those with two copies of the short allele) are the *most sensitive* to this same environmental influence. Similarly, the l/l group (e.g., those with two copies of the long allele) seems to be far more resilient after facing an increasing number of stressful life events. These findings help to explain

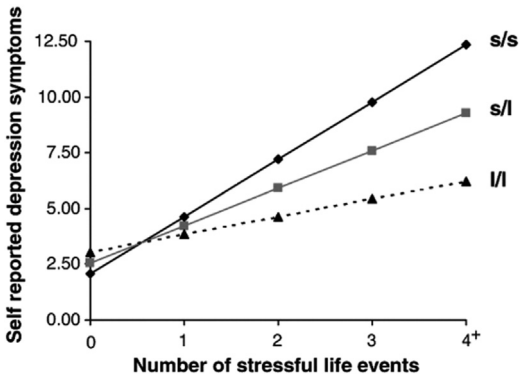

Figure 3 The effects of stressful life events on self-reported depression by genotype (Caspi et al., 2003). Results of multiple regression analyses estimating the association between number of stressful life events (between ages 21 and 26 years) and depression outcomes at age 26 as a function of 5-HTT genotype. Among the 146 s/s homozygotes, 43 (29%), 37 (25%), 28 (19%), 15 (10%), and 23 (16%) study members experienced zero, one, two, three, and four or more stressful events, respectively. Among the 435 s/l heterozygotes, 141 (32%), 101 (23%), 76 (17%), 49 (11%), and 68 (16%) experienced zero, one, two, three, and four or more stressful events. Among the 264 l/l homozygotes, 79 (29%), 73 (28%), 57 (21%), 26 (10%), and 29 (11%) experienced zero, one, two, three, and four or more stressful events. The main effect of 5-HTTLPR (i.e., an effect not conditional on other variables) was marginally significant ($b = -0.96$, SE $= 0.52$, $t = 1.86$, $P = .06$), the main effect of stressful life events was significant ($b = 1.75$, SE $= 0.23$, $t = 7.45$, $P < .001$), and the interaction between 5-HTTLPR and life events was in the predicted direction ($b = -0.89$, SE $= 0.37$, $t = 2.39$, $P = .02$). The interaction showed that the effect of life events on self-reports of depression symptoms was stronger among individuals carrying an s allele ($b = 2.52$, SE $= 0.66$, $t = 3.82$, $P < .001$ among s/s homozygotes, and $b = 1.71$, SE $= 0.34$, $t = 5.02$, $P < .001$ among s/l heterozygotes) than among l/l homozygotes ($b = 0.77$, SE $= 0.43$, $t = 1.79$, $P = .08$).

why some people may respond very differently to the same environmental signal.

Allele by Environment Interactions

However, this approach – in which the effect of a single environment is conditioned on one's genotype for a single gene – is but one of many that one can adopt when investigating gene–environment interactions. We call this approach the *allele by environment* strategy. To conduct such a study, one should identify a key environmental determinant of a dependent variable one wishes to investigate. Next, one should identify a gene that influences a hypothesized biological pathway influencing this outcome. Frequently genes involved in the reward and neurotransmitter pathways of the brain have been popular choices due to their obvious behavioral relevance and, recently, their common availability in social science datasets, but such investigations need not be restricted to these sorts of genes. Genes with documented main effects on the phenotype in question are often tempting targets for use in allele by environment studies, but theoretically these investigations need not be limited to these sorts of genes, either. For instance, variation in the *5HTT* gene has been linked to differential response to a range of environments for a variety of phenotypes (e.g., Caspi et al., 2003; Daw et al., 2013; Simons et al., 2011).

Heritability by Environment Interactions

An additional strategy for investigating gene–environment interactions adopts a behavioral genetic approach, modeling changes in the heritability of traits across different types of environments. We call this the *heritability by environment* approach. As briefly discussed above, heritability estimates are typically obtained by comparing the phenotypic similarity of pairs of individuals with different known genetic relationships. For instance, monozygotic (MZ) twins are genetically identical, whereas dizygotic (DZ) twins share about half of their genes on average by common descent, as do full siblings and parent–child pairs. (However, parent–child pairs show far less variance in the proportion of shared alleles than do full sibling pairs.) Similarly, half siblings, grandparent–grandchild, and aunt/uncle–niece/nephew pairs share about 25% of their genome by common descent, and cousins share about 12.5%. Using family-based data and this type of basic genetic information, one can calculate the heritability as the proportion of variance in a phenotype explained by genetic similarity between biological family pairs. Using extensions of this basic family-based model, one can also estimate the degree to which this estimate varies systematically with an environmental characteristic.

A recent example of this approach was published by Boardman (2009), who used a multilevel twin and sibling model to show that the heritability of smoking was lower in states with high cigarette taxes, and higher in states with lower cigarette taxes. This paper further found that heritability of cigarette smoking is higher in states with lower overall smoking rates. An additional example was documented by Turkheimer et al. (2003), who investigated whether the heritability of intelligent quotient (IQ) varies by the poverty rate of the respondents' neighborhood. This investigation showed that the heritability of IQ increased dramatically, and the shared environmental component decreased dramatically, with increasing neighborhood socioeconomic status. Findings such as these emphasize that heritability estimates are frequently contingent on the environmental context.

Genome-Wide Gene–Environment Interactions

Finally, an emergent approach to the study of gene–environment interactions has been facilitated by the declining costs of genotyping human DNA. Using GWAS datasets, one may now estimate *genome-wide gene–environment interactions* (GWGEI) using a phenotype and environment of interest without the need to select the gene of interest a priori (Cornelis et al., 2012; Mukherjee et al., 2012; Thomas et al., 2012). With available methodologies, this can be done using at least two approaches. First, as with standard GWAS, one can estimate the main effect and interaction of each SNP and a selected environment, and then perform Bonferroni adjustments on the resultant interactions to determine which interactions are statistically significant above the level expected by chance. One could theoretically perform such an analysis for every combination of genes and environments available in a dataset,

but this has not thus far been attempted in any published article to our knowledge. Second, a recently developed method/software called Genome-wide Complex Trait Analysis (GCTA) permits calculation of heritability estimates using GWAS data instead of pairs of individuals with known genetic relationships (Yang et al., 2010). One can also extend this method to estimate interactions between this molecular genetic heritability estimate and an environment of interest (Boardman et al., forthcoming). We will refer to this as the GCTAxE method. Although these are new approaches to the study of gene–environment interactions, these and related methods are very likely to be increasingly pursued given the increasing availability of genome-wide data in popular social science datasets such as the Health and Retirement Study and the National Longitudinal Study of Adolescent Health.

Theoretical Approaches to Heritability–Environment Interactions

When investigating gene–environment interactions, the functional form of the interaction is often of substantive theoretical interest. For instance, a heritability by environment interaction is more than simply a statistical novelty – it is a guide to identifying the type of environments in which biology can 'shine through' (Raine, 2002) (or cannot). When heritability is high, genetic influences hold relative sway; when heritability is low, social environments dominate. Frequently, the types of environments in which heritability are higher and lower is relatively predictable. Accordingly, a set of conceptual models have been characterized to describe the patterns observed. These are illustrated for the reader's reference in **Figure 4**.

The *social distinction* model hypothesizes that heritability estimates will be highest in the most advantageous environments and lowest in the most disadvantaged environments, with approximately linear change in between. In other words, the primary effect of the social environment is to preclude genetic influence. Many authors have argued that this is likely to be the case (Boardman et al., forthcoming; Nielsen, 2006) at least in certain situations. In contrast, the *social push* model (Raine, 2002) makes the key distinction that between middle-range and extreme environments, positing that heritability estimates will be highest in the former and lowest in the latter. The underlying theory argues that in extreme social environments, the environment is a leveling force, leading to lesser overall phenotypic variation. However, within normal social environments, genetics continue to exert significant influence. Boardman et al. (2012b) show that the heritability of body mass index (BMI) is the highest in schools with average BMI levels ($h^2 = 0.87$) and substantially lower in schools with a very low average BMI ($h^2 = 0.74$) and those with very high average BMI ($h^2 = 0.73$).

The most frequently identified empirical pattern of heritability by environment interactions is characterized by the *social control* model (Shanahan and Hofer, 2005), which predicts that heritability will be inversely related to environmental advantage, in the opposite direction of the predictions of the social distinction model. The theoretical conjecture underlying this empirical pattern is that more advantageous social environments also tend to have norms and laws which restrict individual actions for the phenotype in question. As an example of this model, consider the work of Boardman (2009) who shows that genetic influence on regular smoking are significantly reduced in states that have the highest rates of taxation on cigarettes. Finally, the *differential susceptibility* predicts the opposite patterns of those found in the social push model, with high heritability estimates in environmental extremes and comparatively low ones in middle-range environments. One likely possible explanation for such a pattern would be that there is greater overall phenotypic variation in middle-advantage environments compared with less- and more-advantaged environments. If this were the case and molecular genetic effects remained constant, heritability estimates would be inflated in the environmental extremes and deflated in the middle-range (since heritability is the ratio of variance explained by genetics and overall variance).

Theoretical Approaches to Allele–Environment Interactions

Allele–environment interactions also carry theoretical implications for social scientists, though perhaps less sweeping ones than those for the heritability–environment interactions. Perhaps the most important implication of these models is that individuals exhibit differential phenotypic responses to common environmental determinants. In other words, persons

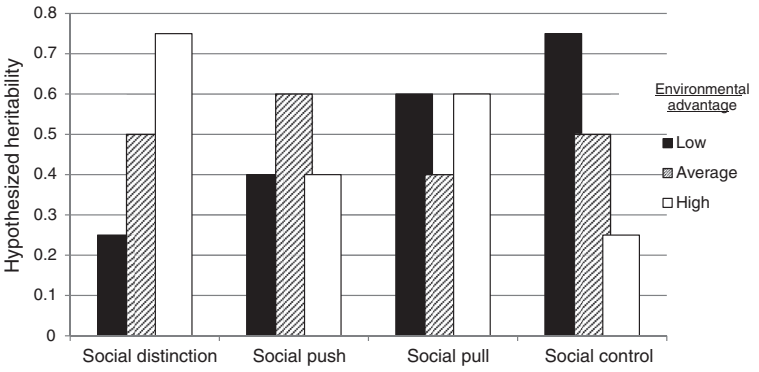

Figure 4 Heritability by environment conceptual models.

with one genotype may exhibit strong responses to environmental stimuli or conditions, whereas those with other genotypes may show a weaker response, no response, or even a response in the opposite direction. Such findings, which are relatively common now, remind us that frequently the environmental effects that social scientists typically document are only averages. (The same can be said for genetic main effects documented by biologists.)

These interactions come in a variety of functional forms, which are graphically depicted in **Figure 5**, and we discuss with reference to health outcomes (the most commonly studied phenotypes in allele–environment interaction research). The most common theoretical model in the allele–environment interaction literature is the *diathesis-stress* model, also known as the *social trigger* model (Ellis et al., 2011; Shanahan and Hofer, 2005). The intuition that underlies this model is that unhealthy environments increase the odds of poor health outcomes – but not for everyone. Instead, those with more 'risk' alleles at the locus in question bear the brunt of the insalubrious effects of these environments, while those with fewer risk alleles show little change in their expectation of this outcome. The Caspi et al. (2003) paper discussed earlier is the most widely cited example of this model.

However, the *differential susceptibility* model has become increasingly common in recent years of research. This model predicts a statistically significant 'crossover' in expected odds of a poor health outcome, such that persons with one type of genotype – more 'sensitivity alleles' – will be showing a far steeper relationship between the environment and the phenotype than those with fewer such alleles, resulting in higher odds of poor health outcomes in poor environments and lower odds thereof in favorable environments. This pattern sometimes characterizes the population as consisting of two groups – 'dandelions,' which can grow in any environment, and 'orchids,' which require very beneficial environments to grow at all but flourish beautifully in the right one (Boyce and Ellis, 2005). For instance, Simons et al. (2011) finds that persons with more sensitivity alleles at the *5HTT* and *DRD4* genes exhibit higher antisocial behavior in adverse environments, and lower antisocial behavior in beneficial ones, than their less genetically sensitive counterparts. Similarly, recent work by Daw et al. (2013) extends this model to cigarette smoking and alcohol use, finding that individuals with more *5HTT* sensitivity alleles smoke and drink more in high-smoking and high-drinking schools, respectively, but show the opposite pattern in low-smoking and low-drinking schools.

A less common theoretical model of allele–environment interaction (at least in social science research) is what we term the *alternative optima* model. This is the implicit model underlying the pursuit of personalized medicine, or the (thus far unrealized) notion that one can use patients' genetic codes to determine the optimal genome-specific therapy for their ailments. If one thinks of therapeutic interventions as the environment in question, then the hypothesis underlying this medical research program is that persons with one genotype

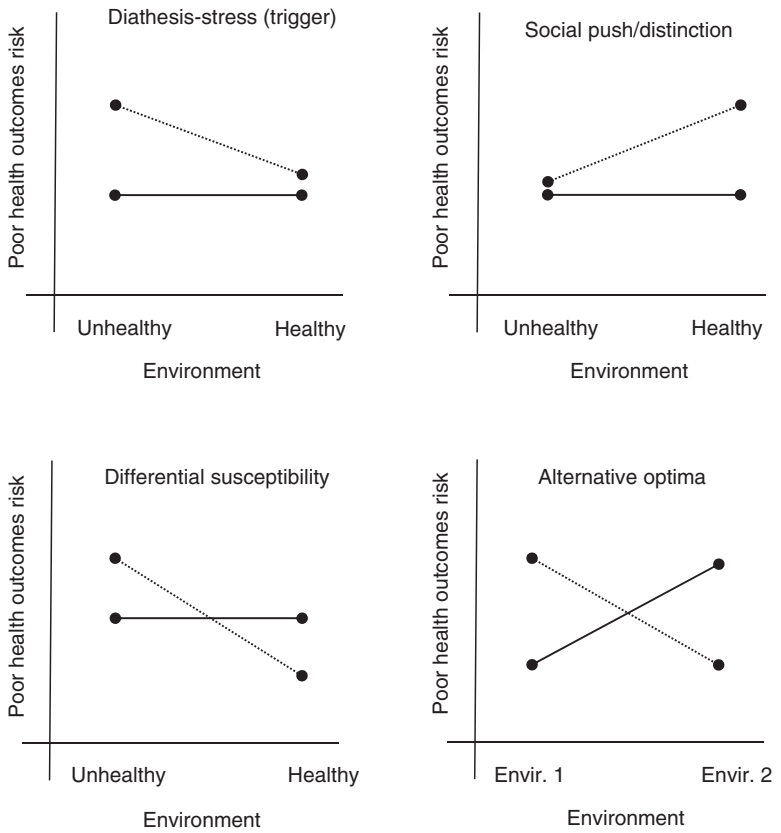

Figure 5 Conceptual models of allele by environment interactions.

will have optimal outcomes under one treatment whereas others with different genotypes will best thrive in a categorically different environment. Note that, unlike other allele–environment models, this does not imply a rank-ordering of environments. Rather, the environments are qualitatively different.

Finally, the *social push/distinction* model of allele–environment interaction argues that a particular genotype will be associated with a health outcome in the most advantageous environments, but not otherwise. Boardman et al. (2012a) demonstrate support for this perspective in which they examine the link between a gene responsible for producing a protein that, when combined with fats, created lipoproteins which move cholesterol around the bloodstream. An allele in this gene (e4) is linked to Alzheimer's disease and cognitive functioning among older adults. These researchers show that the e4 allele is only linked to cognitive functioning among adults from the *most stable, safe, and organized neighborhoods*. This is because there is too much environmental variation in the poorest and most disorganized neighborhoods in Chicago such that genetic differences cannot differentiate between people in these communities. This is shown graphically in **Figure 6** in which each dot corresponds to the estimated effect of the risky e4 allele in the ApoE gene for each of the neighborhoods in this study. While there is variation about this general theme, on average, the effect of the gene is significantly reduced among residents who live in the most disadvantaged neighborhoods. In these neighborhoods, the environment (the social environment) is 'pushing' the phenotype rather than genes.

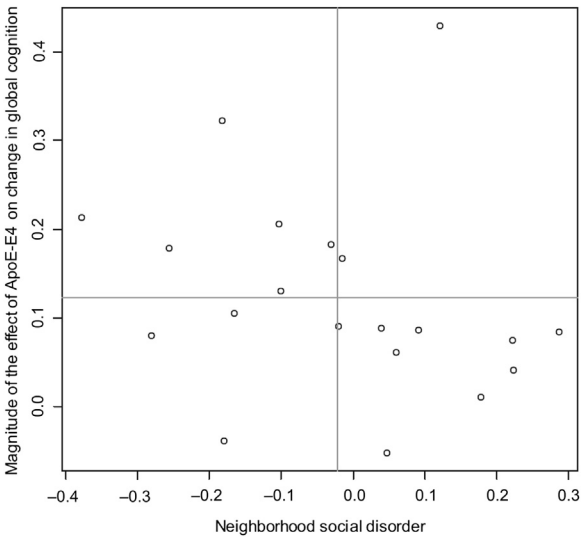

Figure 6 Neighborhood variation in the effect of ApoE genotype and cognitive functioning among older adults. *Note*: Values on the horizontal axis describe the level of social disorder that is present across the 20 neighborhoods in the Chicago Health and Aging Project. The values on the vertical axis denote the absolute value of the effect of the ApoE4 allele on changes in cognitive function from the baseline to the follow-up for each of the 20 neighborhoods. The lines correspond to the mean values for the vertical and horizontal axes.

Limitations of Gene–Environment Interplay Research

We hope that by now the promise of gene–environment interaction research is clear: at its best, it holds the potential to integrate our understandings of humans as both biological and social beings. In practice, this line of research is far from fulfilling this grand promise, and progress toward this goal has not proceeded monotonically.

A number of obstacles impede this goal, of which we emphasize a few. One primary limitation is that, amidst the spiraling numbers of findings on gene–environment interplay that have been reported in academic journals and the press, little in the way of a coherent theoretical framework has emerged to aid in understanding this huge agglomeration of findings. Instead, progress in this field has tended to come primarily in the form of methodological and data-collection advances. Twin and adoption studies have led to expanded family-based methodologies, *candidate gene* studies have led to GWAS (and, soon, full-genome-sequencing of more than 3 billion base pairs), and allele–environment interactions have led to GWGEI and GCTAxE. Yet as increasingly sophisticated techniques for characterizing genotypic variation have flooded the research community with data and findings, little progress has been made in aiding the interested party in predicting the sorts of environments in which one will see higher or lower genetic influence (or complementarily, higher or lower environmental influence). It is possible that there are no robust patterns across phenotypes, genotypes, and environments. At present the best one can hope to do is characterize one's isolated findings in terms of the conceptual models described above. And even with this humbler goal, obstacles remain. Crucially, it is often quite difficult to distinguish between these alternative models outside of conceptual charts.

Another key limitation for this line of research is that gene–environment interplay findings are frequently not replicated. The most highly cited allele–environment interaction, by Caspi et al. (2003), is an interesting case study. This is the article that launched a 1000 replications. Yet still no consensus has been reached on the generalization of this finding (Risch et al., 2009). Progress in understanding gene–environment interplay depends heavily on the ability to test the reproducibility of these findings, since enough researchers studying a given interaction using different datasets and modeling techniques will naturally produce a certain set of false positives (Duncan and Keller, 2011). Furthermore, gene–environment interaction research often has enormous power requirements (Duncan and Keller, 2011), and funds are limited to collect survey and genotypic data on samples of the requisite size. Accordingly, the ability of the scientific community to adequately test the validity of these types of findings is hamstrung by the fact that many researchers in this field work with the same set of large datasets incorporating genotypic data (the National Longitudinal Study of Adolescent Health, the Framingham Heart Study, the Wisconsin Longitudinal Study, the National Health and Nutrition Examination Study, and the Health and Retirement Study).

Finally, research on gene–environment interplay faces a particular challenge known as *population stratification* (e.g., Price et al., 2006). This is a form of confounding for the association between a genotype and a phenotype (or a gene–

environment correlation). Simply put, due to historical patterns of migration, genetic bottlenecks, and *de novo* mutations, some genotypes are more common in populations hailing from certain geographic regions than others. If being a descendant from a group where such an allele is more common is associated with a phenotype compared with descent from a group in which such an allele is less common, a spurious association may be observed if membership in these groups is not controlled for. Fortunately, numerous techniques exist to address this problem: principal components analysis using GWAS data (Price et al., 2006), controls for ancestrally-informative markers (Nassir et al., 2009), and using within-family estimators such as fixed effects regression (Daw et al., 2013), and most academic journals familiar with this research would now require that these concerns be addressed before publishing research reports using gene–environment interplay.

Of course, many research subfields discussed in this volume face challenges such as these. Population stratification is simply the geneticists' term for confounding, with which nearly all social and behavioral scientists should be familiar. All quantitative subfields in the social and behavioral sciences wish they had more and larger datasets with better measures with which to maximize the power of their analyses and likelihood of reproducing 'real' associations. And, while the theoretical stagnation of the gene–environment interplay field is concerning, the rapid rate of technological and methodological advances seen in this field is the envy of most. We hope that the theory will catch up soon.

Conclusion

Gene–environment interplay research addresses the ways in which genetic and environmental determinants of social and behavioral outcomes are interdependent. These primarily include additive genetic and environmental effects, gene–environment correlation, and gene–environment interactions. Different conceptual models of each of these forms of gene–environment interplay were described. The limitations and major obstacles to progress in this field were also noted. While gene–environment interplay research holds great promise and has seen rapid advancement, it also faces a number of substantial obstacles that impede its ability to achieve that promise. Future research should seek to address these limitations while expanding the theoretical, technological, and empirical boundaries of the field as it stands today.

See also: Behavioral Genetics: Psychological Perspectives; Familial Studies: Genetic Inferences; Genetic Studies of Behavior: Methodology; Genetics and Anthropology; Genetics and Economic Behavior; Genetics and Society; Genetics and Sociology; Social Epigenetics; Twin Studies and Epigenetics.

Bibliography

Boardman, J.D., 2009. State-level moderation of genetic tendencies to smoke. American Journal of Public Health 99, 480–486.

Boardman, J.D., Barnes, L.L., Wilson, R.S., Evans, D.A., de Leon, C.F.M., 2012a. Social disorder, APOE-E4 genotype, and change in cognitive function among older adults living in Chicago. Social Science & Medicine 74, 1584–1590.

Boardman, J.D., Roettger, M.E., Domingue, B.W., McQueen, M.B., Haberstick, B.C., Harris, K.M., 2012b. Gene-environment interactions related to body mass: school policies and social context as environmental moderators. Journal of Theoretical Politics 24, 370–388.

Boardman, J.D., Blalock, C.L., Pampel, F.C., 2010. Trends in the genetic influences on smoking. Journal of Health and Social Behavior 51, 108–123.

Boardman, J.D., Domingue, B.W., Blalock, C.L., Haberstick, B.C., Harris, Kathleen Mullan, McQueen, Matthew B. Is the gene-environment interaction paradigm relevant to genome-wide studies: the case of education and body mass index? Demography, forthcoming.

Boyce, William T., Ellis, Bruce J., 2005. Biological sensitivity to context: I. An evolutionary-developmental theory of the origins and functions of stress reactivity. Development and Psychopathology 17, 271–301.

Brim, O.G., Baltes, P.B., Bumpass, L.L., Cleary, P.D., Featherman, D.L., Hazzard, W.R., Kessler, R.C., Lachman, M.E., Markus, H.R., Marmot, M.G., Rossi, A.S., Ryff, C.D., Shweder, R.A., 1996. National Survey of Midlife Development in the United States (MIDUS), 1995–1996 [MRDF], 2nd ICPSR version. DataStat, Inc, Ann Arbor, MI; Harvard Medical School, Department of Health Care Policy [producers], Boston, MA; Inter-University Consortium for Political and Social Research [distributor], Ann Arbor, MI.

Caspi, A., Sugden, K., Moffitt, T.E., Taylor, A., Craig, I.W., Harrington, H., McClay, J., Mill, J., Martin, J., Braithwaite, A., Poulton, R., 2003. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science 301, 386–389.

Cleveland, H.H., Wiebe, R.P., Rowe, D.C., 2005. Sources of exposure to smoking and drinking friends among adolescents: a behavioral genetic evaluation. Journal of Genetic Psychology 166, 153–169.

Cornelis, M.C., Tchetgen, E.J.T., Liang, L.M., Qi, L., Chatterjee, N., Hu, F.B., Kraft, P., 2012. Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. American Journal of Epidemiology 175, 191–202.

Daw, J., Guo, G., 2011. The influence of three genes on whether adolescents use contraception, USA 1994–2002. Population Studies 65, 253–271.

Daw, J., Shanahan, M., Harris, K.M., Smolen, A., Haberstick, B., Boardman, J.D., 2013. Genetic sensitivity to peer behaviors: 5HTTLPR, smoking, and alcohol consumption. Journal of Health and Social Behavior 54 (1), 92–108.

Duncan, L.E., Keller, M.C., 2011. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. American Journal of Psychiatry 168, 1041–1049.

Ellis, B.J., Boyce, W.T., Belsky, J., Bakermans-Kranenburg, M.J., van Ijzendoorn, M.H., 2011. Differential susceptibility to the environment: an evolutionary-neurodevelopmental theory. Development and Psychopathology 23, 7–28.

Mukherjee, B., Ahn, J., Gruber, S.B., Chatterjee, N., 2012. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. American Journal of Epidemiology 175, 177–190.

Nassir, R., Kosoy, R., Tian, C., White, P.A., Butler, L.M., Silva, G., Kittles, R., Alarcon-Riquelme, M.E., Gregersen, P.K., Belmont, J.W., De La Vega, F.M., Seldin, M.F., 2009. An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. BMC Genetics 10, 39.

Nielsen, F., 2006. Achievement and ascription in educational attainment: genetic and environmental influences on adolescent schooling. Social Forces 85, 193–216.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38, 904–909.

Price, T.S., Jaffee, S.R., 2008. Effects of the family environment: gene-environment interaction and passive gene-environment correlation. Developmental Psychology 44, 305–315.

Raine, A., 2002. Biosocial studies of antisocial and violent behavior in children and adults: a review. Journal of Abnormal Child Psychology 30, 311–326.

Risch, N., Herrell, R., Lehner, T., Liang, K.-Y., Eaves, L., Hoh, J., Griem, A., Kovacs, M., Ott, J., Merkangas, K.R., 2009. Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: a meta-analysis. Journal of the American Medical Association 301, 2462–2471.

Shanahan, M.J., Hofer, S.M., 2005. Social context in gene-environment interactions: retrospect and prospect. Journals of Gerontology Series B-Psychological Sciences and Social Sciences 60, 65–76.

Simons, R.L., Lei, M.K., Beach, S.R.H., Brody, G.H., Philibert, R.A., Gibbons, F.X., 2011. Social environment, genes, and aggression: evidence supporting the differential susceptibility perspective. American Sociological Review 76, 883–912.

Thomas, D.C., Lewinger, J.P., Murcray, C.E., Gauderman, W.J., 2012. Invited commentary: GE-Whiz! Ratcheting gene-environment studies up to the whole genome and the whole exposome. American Journal of Epidemiology 175, 203–207.

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., Gottesman, I.I., 2003. Socioeconomic status modifies heritability of IQ in young children. Psychological Science 14, 623–628.

Yang, J.A., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Goddard, M.E., Visscher, P.M., 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42, 565–569.

Generalizability Theory

Georg E Matt, San Diego State University, San Diego, CA, USA
Marisa Sklar, San Diego State University/University of California San Diego Joint Doctoral Program in Clinical Psychology, VA San Diego Health Care System, La Jolla, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G.E. Matt, volume 9, pp. 6032–6037, © 2001, Elsevier Ltd.

Abstract

Generalizability theory (G theory) provides a framework for conceptualizing, investigating, and designing reliable observations. It was originally introduced by Cronbach and colleagues in response to limitations of the popular true-score model of classical reliability theory. G theory reinterprets classical reliability theory as a theory regarding the adequacy with which one can generalize from a sample of observations to a universe of observations from which it was randomly sampled. Because observations may conceivably be generalized to many different universes, observations may vary in how reliably they permit inferences about these universes and, therefore, be associated with different reliability coefficients. G theory also shows how the traditional distinction between reliability and validity can be overcome to design dependable observations. In G theory, a universe, its facets, and the conditions for admissible observations are defined through careful construct explication, the traditional domain of validity theory. The use of the terms 'dependability' and 'generalizability' instead of 'reliability' reflects the interest in unifying reliability and validity. How to go about designing and investigating dependable measures is the subject of G theory.

Overview

Generalizability theory (G theory) provides a framework for conceptualizing, investigating, and designing reliable observations. It was originally introduced by Cronbach et al. (1963, 1972) in response to limitations of the still popular true-score model of classical reliability theory (Spearman, 1904, 1910). Classical reliability theory centers on the notion that each observation or test score has a single true score, belongs to one family of parallel observations, and yields a single reliability coefficient (Nunnally and Bernstein, 1994). While this model may be reasonable for carefully equated parallel forms of tests, it is overly restrictive and often unrealistic in situations where, for instance, raters differ in the central tendency and variance, observations depend on the context in which they occur, and constructs are patently heterogeneous. Several writers argued in the 1950s and 1960s that the same observation could conceivably belong to more than one set of parallel tests and thus have more than one reliability coefficient (Cronbach et al., 1963; Guttman, 1953; Tryon, 1957). The fact that the internal consistency reliability estimate of a multidimensional measure tends to be low even when test–retest and alternate forms reliability estimates are high illustrates the contradictions and limitations of the classical reliability model.

It was Cronbach's vision to reinterpret classical reliability theory as a theory regarding the adequacy with which one can generalize from a sample of observations to a universe of observations from which it was randomly sampled. G theory acknowledges that the reliability of an observation depends on the universe about which the investigator wants to draw inferences. Because a particular measure may conceivably be generalized to many different universes, a measure may vary in how reliably it permits inferences about these universes and, therefore, be associated with different reliability coefficients. G theory explicitly requires investigators to specify a universe of conditions over which they wish to generalize. The extent to which different conditions are associated with different observations has implications for designing dependable observations.

G theory not only reinterprets classical reliability theory but also shows how the traditional distinction between reliability and validity can be overcome to design dependable observations. Traditionally, construct validity is concerned with drawing inferences about a latent construct based on observable measures, whereas reliability is concerned with drawing inferences about a true score from observations across parallel measures. In G theory, a universe, its facets, and the conditions for admissible observations are defined through careful construct explication, the traditional domain of validity theory. G theory defines observations as dependable if they permit accurate inferences about the latent construct (i.e., the universe of admissible observations) they are meant to represent. In doing so, G theory blurs the traditional distinction between reliability and validity. The use of the terms 'dependability' and 'generalizability' instead of 'reliability' reflects the interest in unifying reliability and validity. How to go about designing and investigating dependable measures is the subject of G theory.

Key Concepts

The Decision Maker

Measurements are designed with applications in mind, and their quality has to be examined in the context of their application. This context is symbolized by the 'decision maker' who is interested in measuring a particular construct in a particular population of persons, under a particular universe of conditions, and with particular types of decisions in mind.

Types of Decisions

The types of decisions for which a measurement instrument is to be used have to be known before an instrument that yields dependable observations can be designed. Two major types of decisions have to be distinguished. The first concerns decisions that rely on the relative ranking of individuals. Accepting the three top-scoring applicants for a position is an example of a decision relying on the relative interpretation of scores. Measurements that emphasize the interpretation of the relative differences between scores are the domain of classical test theory with its traditional focus on interindividual differences.

The second major type of decisions is based on the interpretation of the absolute level of scores. Mastery tests in education provide an example of a decision based on the absolute interpretation of a score. Similarly, preestablished admissions standards for college (e.g., minimum SAT score) or pass/fail standards on job selection tests rely on the interpretation of absolute score levels.

The Universe of Admissible Observations

A measurement consists of a sample from the universe of admissible observations. What defines a universe for a particular measurement is based on what the decision maker is willing to treat as interchangeable for the purposes of making a decision. Thus, a universe may consist of interchangeable three-letter syllables, symptoms of depression, or interviewers if the decision maker is willing to treat as interchangeable three-letter syllables, depressive symptoms, and interviewers. A universe may consist of interchangeable time points for urine collection, survey questions about voting behavior, or situations for observing altruistic behaviors if the decision maker considers those aspects of measurement as interchangeable.

Estimating a Person's Universe Score

Given a particular universe of admissible observations, a person's universe score (μ_p) can be defined as the average score based on all admissible observations (X_{pi}) of the universe of interest. The purpose of a measurement is to accurately estimate this universe score ($\hat{\mu}_p$) based on a sample of observations.

The degree to which a measurement is generalizable depends on how accurately the measurement permits us to estimate the universe score. The accuracy of an observation is captured in the variance associated with the different presumably interchangeable observations. The extent to which different items, different observers, or different measurement occasions yield different observations determines how dependable a single specific observation is. In other words, if different items, observers, etc., yield similar observations, a single measurement may allow accurate inferences about the universe of observations. If different items, observers, occasions, etc., yield dissimilar observations, inferences about the underlying universe based on a single observation may be questionable.

The Universe and Its Facets

Universes can be simple or complex, homogenous or heterogeneous, and small or large depending on the construct of interest and the decision maker's interest in investigating different aspects or facets of generalizability. The terms 'facets' and 'conditions' are analogous to 'factors' and 'levels' in the literature on experimental design. A universe of observations is said to consist of one facet if the generalizability of observations regarding one source of variation in the universe, say arithmetic questions of varying difficulty, is at stake. For instance, a particular arithmetic test includes a sample of items, covering different addition, subtraction, multiplication, and division problems of one- and two-digit numbers. The decision maker is interested in general arithmetic achievement and is indifferent to the particular questions on the test. In the one-facet design, one is interested in estimating the universe score of each person based on the sample of items included in the test.

A universe is said to have two (or more) facets if the generalizability of observations regarding two (or more) sources of variation in the universe, say items and occasions, is at stake. For instance, to the extent that arithmetic achievement depends on the difficulty of the items and the test occasions (e.g., beginning or end of school year), the generalizability of a particular measure taken early in the school year on a sample of difficult items may be compromised.

Generalizability Studies

Generalizability studies (G studies) are conducted to investigate the relationship between an observed score and a universe score. They are designed to provide information regarding the sources of variability (i.e., facets of the universe) that influence the generalizability of observations. Because the dependability of a measure rests on its purpose (and thus the universe of admissible observations it is meant to represent), different G studies may be needed depending on the proposed application of the measure. Alternatively, a single G study has to anticipate the multiple uses of a measurement to provide as much information as possible about potentially important sources of variation.

In the above example, arithmetic items varying in difficulty is an obvious choice of a facet to be examined in a G study. In addition, decision makers may be interested in whether measures taken at different occasions, based on different response formats, based on different presentation modes, etc., yield different observations. To examine these issues, a universe has to be defined more broadly to include items of different difficulties, response formats, presentation modes, and taken at different occasions. Having defined such a multifaceted universe, a G study has to be designed to investigate the contributions of these facets and their interactions to the overall variance in observations.

Decision Studies

G studies are designed to assess the dependability of a particular measurement technique. In contrast, decision studies (D studies) are designed to gather data based on which decisions about individuals are made. D studies rely on the evidence generated by G studies to design dependable measures for a particular decision and a particular set of facets about which a decision maker would like to generalize. The goal is to design

a measure that samples sufficient numbers of instances from different facets of a universe of observations to yield a sufficiently dependable estimate of the universe score that the measure is meant to represent.

Illustrative Generalizability Design

Generalizability Design

For illustration, the statistical models and analyses for a one-facet G study will be presented. This most simple generalizability design requires making observations (X_{pi}) on individual subjects (1, 2, …, p,…, n_p) under different conditions of a facet (1, 2, …, i, …, n_i), where subjects and conditions represent random samples from their respective universes. In a one-facet design, the universe of admissible observations consists of observations that differ with respect to one characteristic, say item difficulty. The generalizability question is, how dependable are observations made under different conditions to draw inferences about the universe consisting of all conditions?

The Statistical Model

A one-factorial analysis of variance (ANOVA) model can be used to describe how the observed score of person p under condition i can be partitioned into a grand mean (μ), an effect for each person (π_p), an effect for each item (α_i), and a residual (ε_{pi}).

$$X_{pi} = \mu + \pi_p + \alpha_i + \varepsilon_{pi} \quad [1]$$

In eqn [1], the grand mean (μ) is a constant value for all persons, positioning the score for all persons on the particular scale of measurement. The person effect (π_p) describes the difference between a person's universe score and the grand mean ($\mu_p - \mu$), indicating interindividual differences in universe scores. The item effect (α_i) indicates the difference between an item's difficulty and the average item difficulty ($\mu_i - \mu$), reflecting the variability associated with the facet of error being investigated. The residual combines the influence of the person-by-item interaction, systematic sources of error other than items, and random events.

According to this model, specific observations X_{pi} vary due to differences between persons (i.e., person effects), differences between items (i.e., item effect), and other unaccounted for sources (i.e., residual). Accordingly (see eqn [2]), the variance of a set of observed scores ($\sigma^2_{X_{pi}}$) can be partitioned into components due to persons (σ^2_p), items (σ^2_i), and the residual (σ^2_ε). The residual combines two sources of variance that cannot be distinguished from each other in this model. They are the person-by-item interaction (i.e., items vary in difficulty between persons; persons vary in their ranking on different items) and chance fluctuations in scores.

$$\sigma^2_{X_{pi}} = \sigma^2_p + \sigma^2_i + \sigma^2_{pi,\varepsilon} \quad [2]$$

Generalizability Study: Estimation of Variance Components

To estimate these variance components, a G study has to be conducted in which data are collected on a representative sample of persons and items. Variance estimates can be derived from the mean squares (MS) reported as part of the standard ANOVA printout of statistical analysis programs (see eqns [3]–[5]), in which persons are identified as between-subjects factor and items as within-subjects factor.

$$\hat{\sigma}^2_{pi,\varepsilon} = \mathrm{MS}_{\mathrm{residual}} \quad [3]$$

$$\hat{\sigma}^2_p = (\mathrm{MS}_p - \mathrm{MS}_{\mathrm{residual}})/n_i \quad [4]$$

$$\hat{\sigma}^2_i = (\mathrm{MS}_i - \mathrm{MS}_{\mathrm{residual}})/n_p \quad [5]$$

A Numerical Example

Table 1 shows findings from a G study in which 1275 persons rated the frequency with which they experienced 20 depressive symptoms. This data set may be downloaded at http://x.doi.org/10/1016/B978-0-08-097-86-8.44027-4. The table presents the numerical estimates and the percentage of total variance accounted for by persons, items, and the residual.

Interpretation of Variance Components

Variance components reveal information about how different sources of variability affect the response to a single item. For instance, the variance component associated with items in the above example is 0.0195, accounting for 5.1% of the total variance. This variance component indicates how much the mean of any one randomly selected item is expected to vary from the mean of all items in the universe. Clearly, the magnitude of a variance component depends on how the items were scored (e.g., 1, 2, 3, 4 or 10, 20, 30, 40). In this example, items were scored 1, 2, 3, 4, where 1 indicates the absence of depressive symptoms and 4 indicates the symptoms were present all the time. The average item mean (i.e., estimate of the mean of all items in the universe, $\hat{\mu}_p$) was 1.36, indicating that most respondents reported the absence of depressive symptoms. Taking the square root of the variance component $\sqrt{0.0195}$ yields a standard deviation of 0.1396. Because 95% of item means will fall within 1.96 standard deviations ($1.96 * \pm 0.1396 = \pm 0.2736$) of the overall mean, we would expect 95% of item means to fall between 1.09 and 1.63 (i.e., 95% confidence interval (CI) = [1.09; 1.63]). Given that the possible range of item scores is 1–4, the expected range indicates that the item means in the universe of admissible items will mostly vary between 1 and 2.

The variance component for persons is 0.0687 or approximately 17.9% of the total variance. This variance component estimates the universe score variance, that is, the between-person variability in the average item scores across all the

Table 1 ANOVA MS and estimates of variance components for a one-facet G study

Source of variation	*MS*	*Numerical estimate*	*Percentage of total variance*
Persons (p)	1.6701	$\hat{\sigma}^2_p = 0.0687$	17.9
Items (i)	25.1355	$\hat{\sigma}^2_i = 0.0195$	5.1
Residual (pi, ε)	0.2955	$\hat{\sigma}^2_{pi,\varepsilon} = 0.2955$	77
Total	0.3827	$\hat{\sigma}^2_{X_{pi}} = 0.3842$	100

admissible items of the universe. Based on the variance component, the standard deviation is 0.2621, indicating that 95% of universe scores are expected to fall between ±0.5137 of the grand mean (95% CI = [0.85; 1.87]).

Similar to the reliability coefficient in classical measurement theory, G theory yields a coefficient of generalizability (ρ^2) based on a particular generalizability design. It is defined as the ratio of universe-score and observed-score variances. In this example, the G coefficient for a single item is $\hat{\rho}^2 = \hat{\sigma}_p^2/\hat{\sigma}_{X_{pi}}^2 = 0.1788$. Similarly, intraclass correlation coefficients can be determined. For one item, it is $\alpha_{(1)} = (\mathrm{MS}_p - \mathrm{MS}_{\mathrm{res}})/(\mathrm{MS}_p + (n_i - 1)\mathrm{MS}_{\mathrm{res}}) = 0.1887$, and for the set of 20 items is $\alpha_{(20)} = (\mathrm{MS}_p - \mathrm{MS}_{\mathrm{res}})/\mathrm{MS}_p = 0.8231$. The latter is also known as the Cronbach α coefficient or KR20 formula.

By far the largest variance component is associated with the residual (i.e., 0.2955), accounting for approximately 77% of the total variance. This suggests that there are important sources of variance not accounted for by differences between persons or differences in item difficulties (i.e., item means). Because the residual consists of the item-by-person interaction, this variance component reflects the fact that subjects indicate somewhat different levels of depressive symptoms depending on which item they are responding to. Based on the variance component, the standard deviation is 0.5436, indicating that the response to a specific item is expected to fall within ±1.07 of a person's universe score in 95% of observations.

In summary, the results show that responses to a single item are mostly influenced by measurement error. Information about interindividual differences make up only 17.9% of the total variance, and information about differences in test items make up only 5.1% of the total variance in item responses. Clearly, a single test item is of little use in estimating a person's universe score; that is, a single item provides little foundation for generalizing to a person's universe score. Consequently, D studies must sample many items to achieve more dependable estimates. Exactly how many items is the subject of D studies.

Decision Study: Designing Generalizable Observations for Decisions

To conduct a D study, the relevant universe of observations and its facets must have been investigated in the G study. The type of decisions to be made have to be known, and the accuracy necessary to make the decisions has to specified. With this information at hand, the number of items from each facet of the universe can be determined such that dependable inferences can be drawn about the entire universe of observations at stake.

For instance, the decision maker in the previous example may want to design a measure that yields a 95% margin of error of ±0.20 for the estimate of a person's universe score. That is, the decision maker wants to be 95% confident that a person's universe score is within ±0.20 of the observed score. Based on this 95% margin of error (i.e., $0.20 = 1.96 * \sigma^2_{\mathrm{error}}$), the error variance of the universe score can be determined: $\sigma^2_{\mathrm{error}} = (0.20/1.96)^2 = 0.0104$.

For decisions based on the relative ranking of scores, the required number of items is

$$n_i' = \frac{\sigma^2_{pi,e}}{\sigma^2_{\mathrm{error|relative}}} \qquad [6]$$

For decision based on the absolute ranking of scores, the required number of items is

$$n_i' = \frac{\sigma_i^2 + \sigma^2_{pi,e}}{\sigma^2_{\mathrm{error|absolute}}} \qquad [7]$$

Thus, for relative decisions, we need to sample $n_i' = 0.2955/0.0104 = 28.41$ or 29 items. In addition, for absolute decisions, we need to sample $n_i' = (0.0195 + 0.2955)/0.0104 = 30.29$ or 31 items. In this example, very little variance is contributed by differences between items in their difficulties (i.e., σ_i^2), indicating that regardless of which items are sampled, similar absolute scores will be obtained. Consequently, only two additional items have to be sampled to yield equally dependable estimates of universe scores for relative and absolute decisions.

Extensions

The principles presented for a simple generalizability scenario can be extended to more complex designs consisting of two or more factors, which may be random or fixed, and which may be crossed or nested. The interested reader is referred to the works by Cronbach et al. (1972), Shavelson and Webb (1991), and Brennan (2001a). G theory can also be applied to the analysis of score profiles, composites, and difference scores, and extensions have been developed to estimate variance components via maximum likelihood, Bayesian, and covariance structure methods and for studying the dependability of any facet of observation. The interested reader is referred to the work by Brennan (2011), Gao and Harris (2012), Marcoulides (1996, 1999), Shavelson et al. (1989), Shavelson and Webb (2006), Wittmann (1988), and Webb et al. (1983).

Software

More than 40 years after its introduction, none of the major software packages for statistical analyses in the behavioral sciences have integrated dedicated G or D study analyses as part of their basic data analytic procedures, although most software packages provide estimates of α and intraclass coefficients as part of their reliability analyses. This is likely to have impeded the more widespread adoption of G and D studies in the behavioral sciences. The development of user-friendly programs to perform generalizability analysis should broaden the application of G and D studies. One such program developed principally for generalizability analysis is Generalized analysis of variance (GENOVA). Although originally developed in the early 1980s (Crick and Brennan, 1983), it has since been updated to a suite of computer programs consisting of GENOVA to evaluate univariate balanced designs without missing data, urGENOVA to evaluate unbalanced random effects designs, and mGENOVA to evaluate univariate and multivariate designs (Brennan, 2001b,c). EDUG (Cardinet et al., 2009) is another software package designed to perform generalizability analyses that promises ease of use. Readers are referred to Clauser's (2008) review for a comparison of the GENOVA and EDUG programs. Two kinds of user-written G theory programs were created for both SPSS and SAS. The SPSS and SAS G1 programs are reported to be easier to use but more

restricted in the kinds of permissible research designs than the G2 program. One such restriction is that missing values are not permitted with the G1 programs. Readers are referred to Mushquash and O'Connor (2006) for a comparison of the SPSS and SAS generalizability analysis programs.

See also: Classical (Psychometric) Test Theory; Content Validity; Measurement Theory: Conjoint; Measurement Theory: History and Philosophy; Measurement, Representational Theory of; Psychometrics: Preference Models with Latent Variables; Psychometrics; Reliability: Measurement; Replication.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/B978-0-08-097086-8.44027-4.

Bibliography

Brennan, R.L., 2001a. Generalizability Theory. Springer, New York.

Brennan, R.L., 2001b. Manual for mGENOVA. Iowa Testing Programs, University of Iowa, Iowa City, IA.

Brennan, R.L., 2001c. Manual for urGNOVA. Iowa Testing Programs, University of Iowa, Iowa City, IA.

Brennan, R.L., 2011. Generalizability theory and classical test theory. Applied Measurement in Education 24, 1–21.

Cardinet, J., Johnson, S., Pini, G., 2009. Applying Generalizability Theory Using EDUG. Psychology Press, New York.

Clauser, B.E., 2008. A review of the EDUG software for generalizability analysis. International Journal of Testing 8, 296–301.

Crick, J.E., Brennan, R.L., 1983. Manual for GENOVA: A Generalized Analysis of Variance System. ACT, Inc, Iowa City, IA.

Cronbach, L.J., Nageswari, R., Gleser, G.C., 1963. Theory of generalizability: a liberation of reliability theory. The British Journal of Statistical Psychology 16, 137–163.

Cronbach, L.J., Gleser, G.C., Nanda, H., Rajaratnam, N., 1972. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. John Wiley, New York.

Gao, X., Harris, D., 2012. Generalizability theory. In: Cooper, H., Camic, P.M., et al. (Eds.), APA Handbook of Research Methods in Psychology. Foundations, Planning, Measures, and Psychometrics, vol. 1. American Psychological Association, Washington, DC, pp. 661–681.

Guttman, L., 1953. A special review of Harold Gulliksen's theory of mental tests. Psychometrika 18, 123–130.

Marcoulides, G.A., 1996. Estimating variance components in generalizability theory: the covariance structure analysis approach. Structural Equation Modeling 3, 290–299.

Marcoulides, G.A., 1999. Generalizability theory: picking up where the Rasch IRT model leaves off? In: Embretson, S.E., Hershberger, S.L., et al. (Eds.), The New Rules of Measurement: What Every Psychologist and Educator Should Know. Lawrence Erlbaum, Mahwah, NJ, USA, pp. 129–152.

Mushquash, C., O'Connor, B.P., 2006. SPSS and SAS programs for generalizability theory analyses. Behavior Research Methods 3, 542–547.

Nunnally, J.C., Bernstein, I.H., 1994. Psychometric Theory, third ed. McGraw Hill, New York.

Shavelson, R.J., Webb, N.M., 1991. Generalizability Theory: A Primer. Sage, Newbury Park, CA.

Shavelson, R.J., Webb, N.M., 2006. Generalizability theory. In: Green, J.L., Camili, G., et al. (Eds.), Handbook of Complementary Methods in Education Research. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, pp. 309–322.

Shavelson, R.J., Rowley, G.L., Webb, N.M., 1989. Generalizability theory. American Psychology 44, 922–932.

Spearman, C., 1904. The proof and measurement of association between two things. American Journal of Psychology 15, 72–101.

Spearman, C., 1910. Correlation calculated from faulty data. British Journal of Psychology 3, 271–295.

Tyron, R.C., 1957. Reliability and behavior domain validity: reformulation and historical critique. Psychological Bulletin 54, 229–249.

Webb, N.M., Shavelson, R.J., Maddahian, E., 1983. Multivariate generalizability theory. New Directions in Testing and Measurement: Generalizability Theory 18, 67–82.

Wittmann, W.W., 1988. Multivariate reliability theory: principles of symmetry and successful validation strategies. In: Nesselroade, J.R., Cattell, R.B. (Eds.), Handbook of Multivariate Experimental Psychology, second ed. Plenum Press, New York, pp. 505–560.

Relevant Websites

Interested readers can obtain the generalizability analysis programs from the following websites:

EduG: http://www.irdp.ch/edumetrie/englishprogram.htm.

GENOVA suite: http://www.uiowa.edu/~casma/computer_programs.htm.

SPSS, SAS, and MatLab: http://people.ok.ubc.ca/brioconn/gtheory/gtheory.html.

Generalization: Conceptions in the Social Sciences

Thomas D Cook, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 9, pp. 6037–6043, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

A major goal in the behavioral and social sciences is to offer conclusions that are valid for observed particulars extensive to abstract entities. Four different approaches to generalization are considered in this article. The first describes random sampling with its advantages and limitations. The second considers purposive selection in contrast to randomization. The third covers the possibility of generalizing causal statements and its connection to meta-analysis. And the last refers to extrapolation from a sample to nonsampled entities. This article concludes by arguing that random sampling, though limited, is the best formal method for generalization.

This article examines how social scientists go about probing the generality of causal connections. It distinguishes between on the one hand methods for generalizing to populations of persons and settings and for generalizing to target cause and effect constructs as and, on the other methods for generalizing across different person, setting, cause, and effect constructs. Random sampling is discussed as the best warranted method for generalizing, but its reach for probing the generality of causal connections is unfortunately limited. Methods used for generalizing to construct validity of causes and effects are less well-warranted and more cumbersome, but apply to persons and settings as well as to causes and effects. The generalizing across entities depends on stratification methods that can be used in individual studies. But they are most productively used when synthesizing across multiple studies of the same apparent question, as with meta-analysis. However, the most esteemed method for generalizing causal relationships is by identifying the mediating processes that explain why a cause and effect are related so that these same processes can be instantiated in novel circumstances that have never been examined before. So, causal explanation also helps describe the boundary conditions determining variation in the direction and size of a given causal relationship.

All social scientists use observed particulars to draw conclusions about abstract entities. In measurement theory, the argument that 'IQ is what an IQ test measures' is discredited, and informed debates take place to interpret what the observed scores mean in general language. Some qualitative researchers are also reluctant to generalize beyond observables, but their language is replete with summarizing statements and they invite readers to construct theoretical meanings beyond the descriptions offered. In laboratory research, scientists often study local student samples, not to construct theories about them or about students in general, but to make inferences about humans. Some laboratory researchers use animals to substitute for humans, and their case for cross-species generalization has to depend on arguments – e.g., that cats and humans share the anatomical and physiological attributes central for understanding whatever aspect of vision is under study. This article describes and evaluates many of the procedures for generalization that social scientists actually use.

Cronbach (1982) has perhaps the most comprehensive framework for understanding generalization. We use his theory, though slightly modifying it. He claims that social scientists typically draw conclusions about four not completely independent entities – human populations, physical settings, causes, and observables (including effects and moderators of cause–effect relationships). Time, understood as historical period, might also be added. Cronbach further contends that scientists work with samples, instances, cases, exemplars, or particulars (terms he uses almost interchangeably), and that these represent study-specific populations, categories, entities, universes, classes, constructs, or categories (terms he also uses interchangeably). A major task of all scientists is to identify the class a given set of lower-order instances represents.

Cronbach also understands generalization in terms of inferences about classes that the sampling specifics manifestly do not represent. Such knowledge transfer is particularly salient, for instance, when animals are studied as proxies for humans or when laboratories substitute for other social settings. But transfer is also involved when research on American senior citizens, say, is used to extrapolate to Americans in midlife or to seniors in Western Europe; or when findings about factories are extrapolated to schools on grounds that both are complex organizations. Extrapolation also pertains to causes and effects, as when studies of health care subsidies are applied to housing subsidies, or when a finding about a subsidy's effect on hospital utilization is applied to effects on mental health. Cronbach believes that methods are needed both for justifying inferences from particulars to generals (as from samples to populations or from measures to constructs) and for justifying extrapolations beyond the categories studied.

His schema also includes subcategories – e.g., the racial groups within some human population or the components constituting a construct-like stress. Researchers routinely partition data both to test a relationship's stability across the subcategories examined and to identify the conditions under which it varies. The more heterogeneous the subpopulations are over which a relationship holds, the greater its generality is usually assumed to be. In addition, the more specifically identified are the boundary conditionals limiting a relationship, the better its generality is understood. Campbell's external validity (Campbell, 1957) also emphasizes generalization across entities, but without restricting it to subtypes. For him, a relationship found among Austrians should be examined not just by partitioning Austrians into regional or generational

subgroups, but also by examining the same relationship across other nations. Findings that dependably replicate across very different circumstances are the foundation of all general science.

Using the Cronbach framework, we structure this exposition into four parts: (1) methods based on formal sampling theory, as they are and are not used for generalizing to clearly designated universes of people, settings, times, causes, and effects; (2) purposive sampling methods, especially as they are used in validating constructs; (3) methods for assessing the range of circumstances over which a causal connection holds; and (4) methods for extrapolating knowledge to entities with different characteristics from those on which the knowledge was originally established.

This phrasing acknowledges that many social scientists use universe (and population) differently from construct, though the later Cronbach tends to use them almost interchangeably. In statistics, populations are ostensive; their elements are real and can be pointed to. But constructs are hypothetical and more obviously theory dependent. Moreover, the formal methods statisticians prefer when sampling elements from a population cannot be used with constructs, because the necessary enumeration and sampling of elements cannot be readily achieved with measures of abstract constructs. This is why Cook and Campbell (1979) used external validity to refer to people and settings and construct validity to refer to instances of more hypothetical causes and effects. However, the distinction is partly arbitrary. Constructs have constitutive elements theoretically specified as their components, and instances of any one construct vary in which components they incorporate. Moreover, human populations are not totally ostensive per se. Despite official definitions, there is still room to disagree about what being an Australian means: what about someone with an Australian passport who has always lived abroad, or the illegal immigrant who has always lived in Australia without a passport? There is room for legitimate debate about the definition of any human class and about whether certain individuals fit within that class. Also, whether human populations or abstract constructs are involved, the functional task is always the same: justifying inferences from particular cases to general categories. Hence, we see no fundamental reason for treating populations and constructs differently, though we do respect social science terminology and often use the one to refer to people and settings and the other to cause and effect variables as well as causal moderators.

Formal Sampling Theory

The best way to represent a population is to enumerate its members before selecting a random sample from that population. When properly implemented, this guarantees that the sample will formally represent the population within known limits of sampling error. The sample and population distributions will match on all measured and unmeasured attributes. This makes random sampling the centerpiece of survey research, often complemented by other sampling details. To reduce costs, cases can be clustered within a geographic area before being randomly selected. To ensure that subpopulations are reliably represented, stratification can also occur before random selection. Formal sampling theory provides an impeccable justification for generalization to well-designated populations. And methods for implementing are now so improved that its use in describing human populations and some settings is almost routine.

The problem with random selection is its own limited generality. It rarely applies to causes and other abstract constructs (or to time). One cannot readily imagine enumerating all instances of a specific causal agent before selecting a random sample of these instances for study. In practice, causal agents are purposively chosen because of a presumed correspondence between substantive theory and operational instance(s). Likewise, it is rare for all possible measures of a construct to be enumerated before some are randomly chosen. The selection of measures to represent a construct is again purposive. Also, it is hardly practical to sample historical periods at random for generalizing, if not to eternity, then at least to some meaningful interval. Detailed measures are not always available from the past; and anyway, many theories predict temporal contingency rather than constancy. This last justifies the purposive selection of time periods for the unique light each throws on some causal process.

Even when selection is at random, unambiguous inference still depends on large samples that are not always readily forthcoming. How often can we randomly sample 50 or 100 time points for assessment? Or 50 or 100 institutions, unless the aim is to describe a particular class of institutions? What about manipulating causal agents within institutions? What proportion of those asked to participate will agree to do so? What types are these volunteer organizations likely to be – the most typical or the most proud?

Random sampling rarely applies to assessing the conditions under which cause and effect are related. We know of only six experiments where individuals or households were randomly selected before being randomly assigned to a treatment (see Shadish et al., 2001). However, survey researchers often use random sampling to test causal hypotheses, using item responses to identify variation in a possible causal agent and to provide outcome data. But this is widely acknowledged to generate inferior causal conclusions compared to experimenter-controlled intervention. So how can causal relationships be better generalized via the random selection that surveys afford if these relationships are more uncertain because a survey was used? Moreover, surveys do not routinely use random selection for other generalization-relevant features of the research. For instance, data are typically collected at one or a few time points only, and then in one type of setting (usually the home).

Random selection was not developed for extrapolation, generalizing beyond the sampling particulars. So, it is not surprising that it is not relevant to this function. Indeed, none of the extrapolation methods we later detail depends on random sampling. Other premises are involved, as they are for generalizing causal relationships and for selecting measures to represent constructs. Thanks to its logical and empirical warrant, random selection is *the* method of choice for generalizing to populations of persons and sometimes settings. But it is rarely relevant to other tasks with which a comprehensive theory of generalization must struggle.

Purposive Selection and Construct Validity

Purposive samples are selected without randomization and require only that the selected cases have certain prespecified characteristics. If these characteristics coincide with those of a population, then a match is achieved between the details in the population description and the details achieved in the sampling particulars. Unfortunately, such proximal matching cannot guarantee that the cases and categories are matched on all unmeasured characteristics, let alone perfectly matched on the measured ones. Purposive selection is less well theoretically justified than random selection. By itself, it justifies claims about category membership but not category representativeness.

Yet purposive selection is used throughout the social sciences. Even passionate advocates of random sampling measure behavior by first describing the construct to be assessed and then ensuring that measures of it include all attributes in the theoretical description. Empirical pattern matching between constructs and measures is the basis for generalization in this case, and substantive theory provides the template for choosing how to measure (or manipulate). The correspondence between construct and operational representation is usually multidimensional, with the most important attributes being the prototypical ones that most clearly define the construct's uniqueness (Brunswik, 1956).

Such pattern matching (also called proximal similarity) is not restricted to measurement contexts. When random selection is infeasible it also applies to selecting persons, settings, and times. Thus, experimenters might partition their data into gender subgroups. The resulting subsamples clearly correspond to the categories 'male' and 'female,' as inferred from body shape or valid self-reports. Selecting this way ensures a proximal similarity between the class 'female' and the sampled instances of that class. But do they represent their gender if all attend the University of Berlin? In such a case, the gender specifics and the construct match on some prototypical attributes; but the sample is much more internally homogeneous than the population with respect to age, intelligence, geography, and culture, among other factors. Belonging in a class does not entail formally representing that class. Proximal similarity is necessary but not sufficient for inferring the representativeness of sampled instances. Sophisticated construct validation requires more.

One auxiliary principle concerns 'the heterogeneity of irrelevancies' (Campbell, 1988), ensuring that the sample is not homogeneous on attributes that are irrelevant to the inference being made – such as age, intelligence, and nation in the case of females students in Berlin. To make these conceptual irrelevancies heterogeneous requires adding samples of women of different ages, social classes, and countries. In measurement contexts, heterogeneity requires multiple measures of each construct or varying multiple instances of a treatment, provided that each instance has unique irrelevancies and that no irrelevancy is shared by all instances. It also requires testing the degree of triangulation achieved across the sources of presumed irrelevancy.

Another way to deal with homogeneous irrelevancies requires focused comparison. If female students in Berlin are compared with their male counterparts, this holds constant irrelevancies like country, language, and city of testing, while also varying gender. But such internal comparisons are not always feasible; and groups may differ in other unmeasured irrelevancies that affect responding – e.g., if female students in Berlin have higher intelligence than males on the average, or if data collection there occurs in chemistry classes that attract different kinds of females than males. Whatever the construct being targeted – gender, intelligence, or being Australian – purposive sampling requires critically examining whether the instances selected are more homogeneous than their presumed population.

Construct validation also requires 'discriminant validity' (Campbell and Fiske, 1959), demonstrating that a relationship obtained with constructs labeled one way is not due to another construct that is correlated but nonetheless conceptually unique constructs. Thus, the validity of an anxiety measure requires not only showing that its measured attributes behave as hypothesized but also that they, as a whole, behave differently from conceptually related concepts like depression. Likewise, testing a hypothesis about how new computer firms are unique requires sampling more than just this type of firm. Also needed are firms making new products in other sectors. Relationships that hold across sectors can be generalized to firms in general, including computer firms. However, only relationships limited to computer firms help define what is unique about computer firms.

All sophisticated construct validity is about testing unique predictions entailed by the substantive theory of a population. We always validate interpretations of categories rather than categories themselves, even with human populations (Cronbach, 1989). Lawyers might pronounce that an Australian is anyone eligible for an Australian passport. But this definition is not adamantine. It includes individuals with Australian parents who have never lived in the country; it excludes illegal immigrants who have lived there for decades; and what about persons eligible for a passport who deliberately choose not to claim one despite living in the country for decades? No specification of a construct or population is beyond reasonable criticism and so each preferred interpretation must be defended. Usually, the interpretation can be formulated as hypotheses about some combination of components, antecedents, consequences, or correlates. Construct validation requires specifying and testing the nearest to unique of these hypotheses. Only then can we be truly confident about generalizing from purposive samples and measures to their target constructs rather than to some other entity correlated with them. However, constructing the arguments needed is more laborious and uncertainty-riddled than random sampling. Still, most social scientists sample purposively most of the time, though they vary greatly in how much auxiliary information they collect after a modicum of proximal similarity has been achieved.

Generalizing Causal Connections

Much effort goes into testing causal hypotheses – statements that variation in one construct leads to variation in another and that no alternative cause can account for their covariation. Generalizing causal statements depends on methods other

than formal sampling theory, given that experiments with both random selection and random assignment are rare and surveys provide causal tests that are flawed when compared to experiments. So, how is generalization achieved when the object to be generalized is a causal claim?

One way is when a population has recently been described using measures that are also available on a purposive sample of persons or sites. Then, any causal relationships obtained at the sample level can be weighted to the population level, using the empirical relationship between sample and population for this purpose. But such weighting only approximates what the population effect size would have been if the sample had been randomly selected. Correct inference depends on the sample level effect sizes not being affected by unmeasured sample characteristics that are not fully captured by the shared, measured variables. Still, we can anticipate that weighting procedures will be used more often in the future as methods for social experiments and estimating missing data advance. After all, survey researchers now describe human populations from biased sample data using basically the same weighting technique.

Sometimes, substantive theory is specific enough to detail the conditions under which a causal relationship holds. Then, a program of research can test explicit hypotheses about specific sources of population, setting, or time variation in the strength or direction of the causal relationship. This involves deliberate attempts to falsify the hypothesis that a causal connection is general (Gadenne, 1976). Rejecting the hypothesis entails identifying specific causal moderators; failure to reject implies that the hypothesis of a general cause–effect relationship remains viable. A less formal variant of this describes common practice in the laboratory sciences. A causal claim is made and then independently replicated, often under conditions that approximate the original ones. Successful replication leads to concluding that the causal relationship is provisionally true – that is, true until future failures to replicate indicate the need to invoke causal moderator variables. In both approaches, special emphasis is on identifying moderators that reduce the causal relationship to zero or that change its sign, as opposed to those that merely vary its size without affecting its sign.

Purposive sampling is also used in other less formal generalization probes. In clinical trials, hospitals recruit patients with a given disease and then adhere to the study's intervention and data collection protocols. Hospitals are basically chosen for their willingness to comply and for their heterogeneity with respect to, say, region of the country, predominant ethnic group served, or type of hospital sponsorship. Statistical power is often only adequate for cross-site analysis, suggesting that the goal is not to identify site-level main and interaction effects, but to see if the causal estimate exceeds zero despite heterogeneity in the hospitals studied. If it does, the causal effect is held to be general. If it does not, several alternative claims are viable and need to be tested, especially the following: there is no effect; the causal sign varies by hospital type and obscures a main effect; or the methods used cannot detect an effect of the size expected. Only increasing the sample of sites allows direct tests of which hospital types are involved in site-level interactions. Even so, relying on a purposive sample of sites that are heterogeneous on many attributes except for volunteering is only logically defensible if researchers know from other sources that no confounds are associated with all the hospitals being volunteers for the clinical trial.

Exploratory analyses often involve measuring many constructs from among a large array of possible causal modifiers in order to probe which ones affect the strength of a causal relationship. Without a priori hypotheses about modifiers, the approach can be time-consuming and run afoul of error rate problems due to the many tests conducted and the small sample sizes for some values on the moderator. Nonetheless, subcategory breakdowns are often used within experiments to probe how robust a relationship is across persons, settings, times, and measures. This can prevent the overgeneralization that occurs when researchers conclude that a relationship found with the entire sample holds in each subgroup within the sample. The entire sample allows one to conclude that the effect is found despite the internal heterogeneity achieved. It does not imply there is independent replication within any of the individual subgroups. The story is essentially the same when partitioning data from a single survey. With adequate sample sizes per group, one can probe the robustness of a relationship across the subgroups sampled. But only those potential conditional variables directly sampled can be assessed.

The best tests of causal conditionals come from synthesizing multiple studies on a topic rather than from subgroup breakdowns within a single study (Cooper and Hedges, 1994). Experiments and surveys relevant to the same causal hypothesis accumulate and can be used in meta-analysis, the best-known form of synthesis. Meta-analysis combines data across multiple studies asking the same question, taking advantage of the between-study variation in measures, manipulations, populations of people, settings, and times. The dependent variable is each study's standardized effect size(s) and these are manipulated (1) to generate an average effect size across all the studies and all the irrelevant sources of heterogeneity they contain, and (2) to identify the specific contingency factors responsible for variation in the effect sizes.

What logic buttresses meta-analysis? It is not random sampling since individual studies are selected for their relevance to the hypothesis and the hope is to achieve a census of all published and unpublished studies. However, the more relevant target population is all studies that could be conducted on the topic, not all those done on it. In meta-analytic practice, generalization depends on the sample of achieved studies including a wide range of human populations, setting types, historical periods, and operationalizations of the cause and effect. Each study is first coded to label its sampling particulars in terms of theoretical constructs and to examine the sources of irrelevancy with which the particulars are and are not collectively imbued. Thus in a hospital experiment, hospitals might be described as being university-based, public, or private, and as being large or small. Patients might be described in terms of diagnosis, age and gender mix, etc. The treatment might be described in terms of its components, duration, and manner of delivery; and the outcome might be described in similar terms. Then researchers typically test whether the same causal relationship holds despite these sources of sampled irrelevancy.

Sample sizes permitting, researchers also test whether the relationship holds across individually identified sources of

potential irrelevancy so as to generalize to hospitals in general, people diagnosed with cancer, or whatever the target inferences might be for a particular study. Important here are the number and especially the heterogeneity of the types examined. Of course, some sources of variation might have a priori substantive meaning, as when university hospitals are a special target. Then one has to do a discriminant validity test and show that the effect size for such hospitals is different from that for other hospitals. Otherwise, there is nothing unique about university hospitals. So, meta-analysis depends on the very tests for proximal similarity, heterogeneous irrelevancy, and discriminant validity that are so central in construct validation. In addition, to achieve such tests multiple replication is needed across studies that may each be individually very narrow in the range of people, settings, times, and cause and effect operations it includes. Central is the heterogeneity across studies and the pattern of results achieved despite this sampled heterogeneity.

Extrapolating to Nonsampled Entities

Some social scientists probably think it silly to use samples for generalizing to populations that are not even proximally similar to the samples used. So, many animal researchers abjure all talk of generalizing to humans. But their language is often anthropomorphic, and their research is usually funded because others think it is relevant to humans. So, elaborate reasoning has evolved over time to justify generalizing from animals to humans (Steel, 2008).

In this regard, consider the use of cats in visual research (Blake, 1988). The rationale for their use is in two parts. The first is entirely pragmatic and stresses the conveniences of working with cats (or *Drosophila* in genetics research), usually because of cost and ethics. The second is theoretical and is based both on cats sharing some anatomical and physiological features with humans and on behavioral research showing that cats and humans behave similarly when confronted with similar visual stimuli. However, the empirical work necessary for showing such correspondences has also revealed that cats are much less like humans in some physical ways – especially features related to night vision – and that some behaviors are not replicated across species when the same stimulus is presented. So, a contingency theory has evolved of the conditions under which cats are better or worse substitutes for humans.

But when asked, honest vision researchers will avow that convenience is the paramount rationale for using cats. Would they rather work with chimpanzees? Undoubtedly, since the chimps' visual system is much closer to humans' than the cats.' But the expense of maintaining chimps precludes using them. Would the researchers rather use humans than chimps? Undoubtedly, if ethical research procedures are possible – which they often are not. However, now that new imaging technologies are available that allow observing the human eye reacting to planned stimuli in great microdetail and real time, many vision researchers are turning to work with humans rather than cats because the pragmatic obstacles to using humans now hardly apply. Thus within the community of vision researchers the preference for generalization to humans via within-species proximal similarity is clear. So, too, is the distaste for arguments that depend on the (only partial) proximal similarity between cat and human on some vision-related attributes given their obvious dissimilarity in others. A similar story could be told for conducting research in the laboratory rather than in the field (Shadish and Cook, 2009). Convenience is paramount, and the reasons given for preferring the lab often contain elements of rationalization. But even so, extrapolation across species and settings is surely superior when there is an empirically defensible contingency theory of the conditions under which the extrapolation is better and worse warranted.

How is extrapolation otherwise achieved? Causal explanation is often held to play a major role, explaining how or why two or more things are related (Mackie, 1974). Once an explanatory process has been identified, the argument is that it can be reproduced in other settings, with other kinds of people, at future times, and with unique ways of instantiating the process. Indeed, this is the major public rationale for basic science. Once one knows why fire happens, one can replicate it with new kinds of materials (say, some newly invented polymers), in new settings (an uninhabited desert island), and with people who have never seen fire before (so long as they can understand the explanatory recipe). The problem with this view is that full explanation is hard to conceptualize and achieve. It is difficult enough to use data for discriminating among explanatory claims even at one level of explanation, let alone for simultaneously investigating factors at the biological, individual, and social levels. Even so, causal explanation is the holy grail of science, not just for its aesthetic parsimony, but also for the utilitarian causal generalization it promotes in the form of extrapolation.

Extrapolation is also achieved through careful induction. If patient education reduces the length of hospital stays and does so with patients suffering from 25 different diagnoses, then it is likely to do so with the 26th. Of course, the logical warrant for such induction is limited, and seems more reasonable the more numerous and heterogeneous are the patient types already examined. If all 25 were different forms of cancer diagnosis, or involved patients aged between 25 and 40, or were restricted to men, this limited heterogeneity would reduce our confidence in any inductive-empirical extrapolation. Meta-analysis is particularly important in this regard, especially when it permits replication across multiple different sources of irrelevancy. A relationship that has proven to be robust in whatever form it was operationalized is more likely to reoccur when other ways of operationalizing its components are attempted – more likely, but not certain. We see here the value traditionally assigned in science to discovering both general dependencies and the contingencies that describe when a given relationship does and does not hold.

The final context for extrapolation relates to formal modeling. When subgroups can be arrayed along a single continuum, the functional form of a causal relationship can then be described. Is it stable? Is there a linear increase or decrease? If a pattern clearly emerges, one might want to argue that the strength of the relationship can be inferred for unobserved subpopulations with values that fall within the continuum range. Simple interpolation is involved here, while the strength of the relationship for values beyond the continuum is inferred by extrapolating from the observed function. But comfort with such extrapolation depends on how

distant the nonobserved subpopulation values are from the sampled ones.

Model-based extrapolation is always tricky, but especially so when the models are more complex. Consider much of macroeconomics. Models of a given economic system are developed from theory and past findings, and changes in parameter values are introduced to see how the system reacts to this perturbation. This is basically an extrapolation exercise to discover what would happen if some details of the economic system were not as they actually are. Can we extrapolate from what we think we know, in order to learn what would happen under novel circumstances? The validity of any extrapolation depends on the model being correctly specified with respect to the variables included, the relationships postulated between them, and the strengths of these relationships. Be wrong about any of these assumptions and any conclusions about the effects of the novel perturbation will be biased to some unknown extent. The assumption that the model is complete and accurate is difficult to test, though crucial to the credibility of the results (Glymour et al., 1987). Still, the creation of valid multivariate models of causal interdependency is another major goal of science, and it is not difficult to see why. It promises the extrapolation of results beyond whatever circumstances were studied in the past.

Conclusion

In the behavioral and social sciences at the beginning of the twenty-first century, theory and practice are most developed when individuals or households are sampled to describe a human population. Here, best practice requires randomly sampling units from the population. Theory and practice are also well developed for generalizing from a measure to an abstract construct or from an experimental treatment to a more general causal agent. Here, best practice requires an explicit theory of construct validity that necessarily invokes proximal similarity, but preferably also the heterogeneity of irrelevancies, discriminant validity, and causal explanation. However, the purposive selection and theory-linked methods needed for construct validation are not as formally well supported as random selection is. Theory and practice are even less developed for generalizing to settings and historical time periods and for identifying the conditions under which a causal relationship holds. However, various deliberate sampling and replication strategies are relevant here, including meta-analysis. At their best, such strategies borrow much of the logic underpinning sophisticated construct validation. Least well developed are theory and practice for extrapolating results to nonstudied circumstances. But even here, understanding causal explanatory processes can permit re-creation of the same causal forces in novel settings. Also, there are now decades of research and reflection about using animals to extrapolate to humans and about using the laboratory to extrapolate to other social settings. In addition, substantive modeling is sometimes used to examine what would happen to a system if one or more of its parameters were changed. Of all the methods examined, only random sampling provides an impeccable formal rationale for generalization. But its reach is, alas, limited. It is a great irony that the best formal method for generalization in the behavioral and social sciences can be used for only a small part of the generalization tasks that practicing social scientists routinely face. Given this, practicing social scientists and theorists of method reflecting on these practices have cobbled together more complex methods of generalization that depend more heavily on substantive theory, purposive sampling, heterogeneity of irrelevancies and demonstrated replication and the identification of causal contingencies.

See also: Causal Counterfactuals in Social Science Research; Experimental Design: Randomization and Social Experiments; External Validity; Generalizability Theory; Internal Validity.

Bibliography

Blake, R., 1988. Cat spatial vision. Trends in Neuroscience 11, 78–82.
Brunswik, E., 1956. Perception and the Representative Design of Psychological Experiments, second ed. University of California Press, Berkeley, CA.
Campbell, D.T., 1957. Factors relevant to the validity of experiments in social settings. Psychological Bulletin 54, 297–312.
Campbell, D.T., 1988. In: Overman, E.S. (Ed.), Methodology and Epistemology for Social Science: Selected Papers. University of Chicago Press, Chicago, pp. 290–314.
Campbell, D.T., Fiske, D.W., 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin 56, 81–105.
Cook, T.D., 1993. A quasi-sampling theory of the generalization of causal relationships. In: Sechrest, L., Scott, A.G. (Eds.), New Directions for Program Evaluation: Understanding Causes and Generalizing about Them, vol. 57. Jossey-Bass, San Francisco, pp. 39–82.
Cook, T.D., Campbell, D.T., 1979. Quasi-experimentation: Design and Analysis Issues for Field Settings. Rand-McNally, Chicago.
Cooper, H., Hedges, L.V. (Eds.), 1994. The Handbook of Research Synthesis. Russell Sage Foundation, New York.
Cronbach, L.J., 1982. Designing Evaluations of Educational and Social Programs. Jossey-Bass, San Francisco.
Cronbach, L.J., 1989. Construct validation after thirty years. In: Linn, R.L. (Ed.), Intelligence: Measurement, Theory, and Public Policy. University of Illinois Press, Urbana, IL, pp. 147–171.
Gadenne, V., 1976. Die Guttigkeit psychologischer Untersuchungen. Kohlhammer, Stuttgart.
Glymour, C., Scheines, R., Spirtes, P., Kelly, K., 1987. Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling. Academic Press, San Diego, CA.
Mackie, J.L., 1974. The Cement of the Universe: A Study of Causation. Oxford University Press, Oxford, UK.
Steel, D., 2008. Across the Boundaries: Extrapolation in Biology and Social Science. Oxford University Press, Oxford.
Shadish, W.R., Cook, T.D., Campbell, D.T., 2001. Experimental and Quasi-experimental Designs for Generalized Causal Inference. Houghton-Mifflin, Boston.
Shadish, W.R., Cook, T.D., 2009. The renaissance of field experimentation in evaluating interventions. Annual Review of Psychology 60, 607–629.

Generalized Linear Mixed Models

Charles E McCulloch and John M Neuhaus, University of California, San Francisco, San Francisco, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides an overview of generalized linear mixed models (GLMMs), how they are fit to data, and the inferences possible when using them. GLMMs are a class of statistical models that handle a wide variety of distributions for the outcome, accommodate nonlinear models, and model correlated data. As regression methods, they are not only capable of estimation and testing of covariate effects but also can be used to draw inferences about correlation structures in the data and are able to calculate predicted values that take into account not only covariates but also observed outcomes. We briefly describe software available for fitting GLMMs.

Introduction

Research questions often focus on analysis of change over time and use longitudinal and repeated measures designs. For example, in the Sacramento Area Latino Study of Aging (SALSA), a cohort of elderly Latinos was recruited to study, among other goals, the influence of diabetes on cognitive decline (Wu et al., 2003). In SALSA, cognitive functioning was assessed at baseline and in annual follow-ups by the Modified Mini-Mental State Examination (3MSE), which measures cognitive ability on a numeric scale that ranges from 0 (the worst) to 100. If the study was a cross-sectional study, then we might model 3MSE score as a function of diabetes status and other covariates (e.g., age, gender, hypertension, education) using linear regression. However, linear regression assumes that the outcomes are independent so that the method would not be able to accommodate the repeated measurements over time and the resulting correlated outcomes within a person.

Furthermore, the values of 3MSE are often dichotomized into cognitively impaired or not, with values under 79 indicating cognitive impairment. While dichotomizing a numeric outcome engenders a loss of information, it is sometimes an advantage to simplify clinical interpretations. With a cross-sectional study, we might use logistic regression to model cognitive impairment (Yes/No) as a function of the same predictors of diabetes status, age, gender, hypertension, and education. But, again, as with linear regression, this would not be appropriate with repeated, correlated outcome measures. Another feature of this design is the fact that some of the predictors (e.g., gender) do not vary with the repeated measurements over time and are measured at the person level, whereas other will vary, for example, hypertension.

In this article, we describe a flexible class of statistical models that are capable of accommodating correlated outcome data of a variety of different outcome types (e.g., numeric versus binary) and also predictors that are measured at different 'levels' in the study (e.g., at the person or within-person level). These are known as *generalized linear mixed models* (GLMMs) with the various components of the name being informative. The term 'generalized' indicates that a variety of distributional assumptions are possible (Normal or Gaussian, binomial, Poisson, etc.) to handle outcomes of different types. The term 'linear' indicates that the predictors enter the model together as a linear combination, just as in a linear regression where the effects of predictors enter the statistical model through an equation of the form,

$$\beta_0 + \beta_D DIAB + \beta_A AGE + \beta_E EDUC + \beta_G GEND \qquad [1]$$

Finally, the term 'mixed' indicates the method by which correlated outcomes are modeled as correlated, which we describe in more detail below.

We begin by describing how GLMMs accommodate correlated data and then formally define the class of models that comprise GLMMs. In the Section Types of Inferences Possible from GLMMs, we list the statistical inferences possible using GLMMs and then (in the Section Fitting and Inferential Methods) how they are fit to data. The Section Best Prediction discusses a unique feature of GLMMs, namely, prediction, and we close with brief sections on diagnostics and missing data.

Accommodating Correlation in GLMMs

The key idea in GLMMs is to incorporate correlation by modeling one or more of the predictors as random. Predictors that are not random are called fixed and because the model includes both fixed and random predictors it is described as a 'mixed' model. The impact of including a random predictor is easiest to describe in the context of a linear model, so we consider a simple one using the SALSA investigation.

Suppose we wish to create a model for a person's 3MSE score as a function of their diabetes status at the current visit, with M_{it} denoting the 3MSE for person i at time t, a simple model would be

$$M_{it} = \beta_0 + P_i + \beta_D DIAB_{it} + \varepsilon_{it} \qquad [2]$$

where P_i is person effect and $DIAB_{it}$ is 1 if the person is diabetic at time t and 0 otherwise. Larger values of P_i will tend to lead to larger measurements on a person, allowing the model to capture the feature that some participants' 3MSE scores are consistently higher and others are consistently lower. In a longitudinal context, $\beta_0 + P_i$ would be a person-specific intercept, allowing individuals to start at different values. The ε_{it} is an error term with mean 0 and with a variance we will denote by σ_ε^2.

To make statistical inferences from our experiment, we will need to assume that the people in the study are a random sample from some population of people, real or hypothetical.

Consequently, the person effects are necessarily a random sample of the person effects from that population, which has a distribution in the population with a variance, we will denote by σ_P^2. This simple idea, of regarding the person effects as random, introduces a covariance between outcomes (say outcome at time t and outcome at time s) within a person.

$$\begin{aligned}\text{cov}\,(M_{it}, M_{is}) &= \text{cov}\,(\beta_0 + P_i + \beta_D DIAB_{it} + \varepsilon_{it}, \\ &\qquad \beta_0 + P_i + \beta_D DIAB_{is} + \varepsilon_{is}) \\ &= \text{cov}\,(P_i + \varepsilon_{it}, P_i + \varepsilon_{is}) \qquad [3] \\ &= \text{cov}\,(P_i, P_i) \equiv \text{var}\,(P_i) \qquad [4] \\ &= \sigma_P^2\end{aligned}$$

In the above, equality [3] holds because the omitted terms are fixed constants and equality [4] holds under the assumption that the error terms are uncorrelated with each other (and uncorrelated with the P_i). In a similar fashion, it is straightforward to calculate the variance of M_{it} or M_{is}, both of which are equal to $\sigma_P^2 + \sigma_\varepsilon^2$, so that their standard deviation is $\sqrt{\sigma_P^2 + \sigma_\varepsilon^2}$. With these facts in hand the correlation between two measurements taken on the same person would be

$$\begin{aligned}corr(M_{it}, M_{is}) &= \frac{\sigma_P^2}{\sqrt{\sigma_P^2 + \sigma_\varepsilon^2}\sqrt{\sigma_P^2 + \sigma_\varepsilon^2}} \\ &= \frac{\sigma_P^2}{\sigma_P^2 + \sigma_\varepsilon^2}\end{aligned} \qquad [5]$$

the well-known intraclass correlation coefficient.

This simple model with only random intercepts generates equal correlation between two outcomes on the same person. This will not be sufficiently flexible for many situations, but can easily be added upon. For example, SALSA recruited participants across a number of census tracts, all of which had high percentages of Latinos. It could well be that observations on individuals within census tracts tend to be systematically higher or lower than other census tracts, perhaps because of unmeasured characteristics. Letting M_{cit} denote the 3MSE measurement at time t for participant i in census tract c, an elaboration of eqn [2] would be

$$M_{cit} = \beta_0 + T_c + P_{ci} + \beta_D DIAB_{cit} + \varepsilon_{cit} \qquad [6]$$

where T_c is a census tract effect. If we assume that the T_c are random effects then calculations similar to those in eqns [4] and [5] give the following correlations for, respectively, outcomes measured on the same person (and in the same census tract) at different times and different people in the same census tract (at any time):

$$\begin{aligned}corr(M_{cit}, M_{cis}) &= \frac{\sigma_T^2 + \sigma_P^2}{\sigma_T^2 + \sigma_P^2 + \sigma_\varepsilon^2} \\ corr(M_{cit}, M_{cjs}) &= \frac{\sigma_T^2}{\sigma_T^2 + \sigma_P^2 + \sigma_\varepsilon^2}\end{aligned} \qquad [7]$$

Model [2] also has the feature that correlations between measurements taken on the same person are the same irrespective of how far apart they are in time, which might not be realistic. There are also logical extensions of this model that are more realistic in longitudinal studies. For example, suppose we were interested in a model to address whether diabetes status at baseline was predictive of decline in cognitive functioning or not. Since a time effect measures overall decline in cognitive functioning, to capture the concept of decline depending on diabetes status we would incorporate a diabetes status by time interaction.

Because some participants will maintain cognitive function with age while others will decline, we will need to allow not only person-specific intercepts but also person-specific slopes. Figure 1 shows the values of 3MSE plotted versus age for

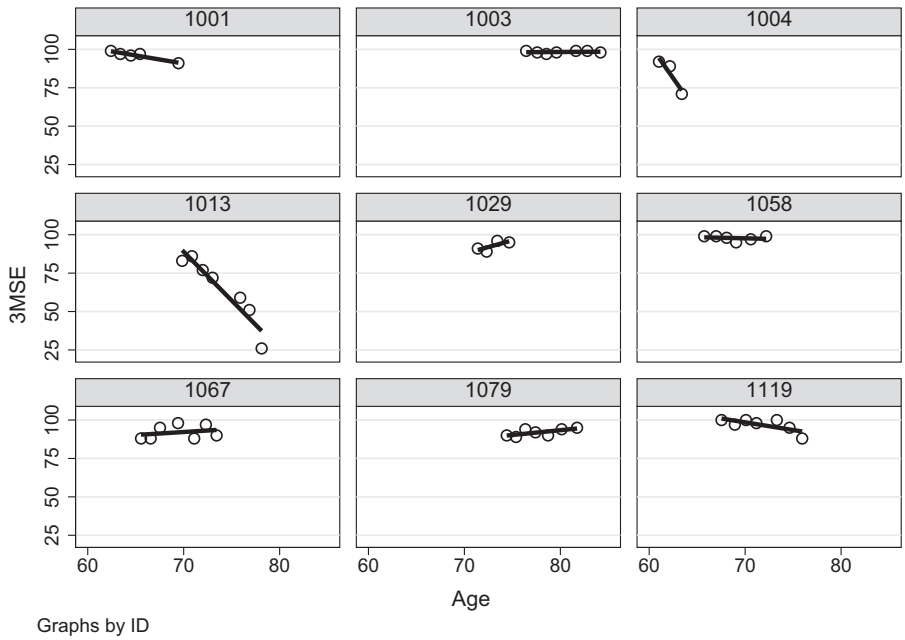

Figure 1 3MSE values by age for nine SALSA participants showing linear fits and actual values.

a selected subset of the SALSA participants, making it clear the need for both person-specific intercepts and slopes.

Here is a model that includes person-specific intercepts and slopes, as well as the diabetes and diabetes by time interaction:

$$M_{it} = (\beta_0 + P_i) + (\beta_T + PT_i)TIME_{it} + \beta_D DIAB_{i0} + \beta_{DT} DIAB_{i0} \times TIME_{it} + \varepsilon_{it} \quad [8]$$

In this model, $\beta_0 + P_i$ is the person-specific intercept and $\beta_T + PT_i$ is the person-specific decline over time. By the same random sampling argument above, the PT_i form a random sample from a population of person-specific slopes with some variance. That variance will be different from the variance of P_i (since they are measured on different scales), and it is quite possible that P_i and PT_i are correlated. For example, participants who started with low values (negative P_i) might decline more rapidly.

As illustrated in the examples above, the specification of a GLMM involves four basic steps:

1. Specify the distribution of the outcome Y. This is the distribution for fixed values of the predictors and conditional on the random effects. For example, in the SALSA examples we might use a normal distribution for 3MSE and a Bernoulli distribution for cognitive impairment (3MSE < 79).
2. Specify the predictors to be included in the model. As noted in eqn [1], these will enter the model in the form of a linear combination. This is typically called the *linear predictor* and is denoted by η. In more generic notation, it would be defined as $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, where the βs are regression parameters to be estimated from the data and x_k are the predictors.
3. Specify which function of the mean, $\mu \equiv \mathrm{E}[Y|x]$, is equal to the linear predictor. In a GLMM, this is of the form $g(\mu) = \eta$, where g is a known monotonic function, called the *link* function. For example, with a binary outcome, we might specify g as the logit function, i.e., $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$.
4. Declare one or more of the categorical predictors to be random factor(s) by assigning a distribution to the effects associated with the levels of that factor.

The Model

We are now in a position to more formally define a GLMM using the following notation. For the ith observation, let Y_i denote the outcome and let μ_i be the mean of Y_i, conditional on the random effects. Further, let $\mathbf{x}_i$ be the vector of fixed effect (nonrandom) predictors, $\mathbf{z}_i$ be the model matrix for the random effects, $g(\cdot)$ be the link function, β be the regression coefficients, and $\mathbf{u}$ be the vector of random effects. Our GLMM is then

$$\begin{aligned} Y_i|\mathbf{u} &\sim \text{indep. } f_Y(Y_i|\mathbf{u}) \\ g(\mu_i) &= \mathbf{x}_i'\beta + \mathbf{z}_i'\mathbf{u} \\ \mathbf{u} &\sim f_u(\mathbf{u}) \end{aligned} \quad [9]$$

where f_Y and f_u are, respectively, the density functions of Y and $\mathbf{u}$.

Using the cognitive impairment example above to illustrate the definition of a GLMM, define C_{it} to be 1 if the 3MSE score is less than 79 and 0 otherwise so that the outcome is binary. We therefore could use a logit link and might include random intercepts (u_{0i}) and random slopes over time (u_{1i}) and assume that the random effects follow a normal distribution:

$$C_{it}|\mathbf{u}_i \sim \text{indep. Bernoulli}(p_{it},), i = 1, \ldots, m; t = 1, \ldots, n_i$$

$$\log[p_{it}/(1-p_{it})] = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i})TIME_{it} \quad [10]$$

$$\text{or } p_{it} = 1/[1 + \exp\{-[(\beta_0 + u_{0i}) + (\beta_1 + u_{1i})TIME_{it}]\}]$$

$$\mathbf{u}_i \sim \text{indep. } \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{u}})$$

where the joint distribution of $\mathbf{u}_i = (u_{0i}\, u_{1i})'$ is bivariate normal.

Types of Inferences Possible from GLMMs

GLMMS are based on generalized linear models, which are nonlinear regression models. So GLMMs allow the usual inferential goals that regression models are used for. They can be used to test the association of an outcome with a predictor, to quantify the degree of association, or to estimate the mean value of the outcome for given values of the predictors. For example, in SALSA there was an interest for testing if diabetes was related to cognitive impairment and in quantifying the strength of the association. Another possible inferential goal would be to predict the chance of cognitive impairment for subgroups of the participants, e.g., diabetic women over the age of 60.

There are, however, two types of inferences that are possible from GLMMs that are not possible from generalized linear models, both related to the inclusion of random effects in the model [9]. First, we can analyze features of the random effects distribution. For example, in model [10], the variance of u_{2i} (the lower right element of $\mathbf{\Sigma}_{\mathrm{u}}$) measures the variability in the cognitive impairment declines across different participants, so testing the null hypothesis of $H_0 : \sigma^2_{u2} = 0$ tests the hypothesis that all participants decline at the same rate. This might be useful as a precursor to identifying if there were some participants with 'healthy aging,' i.e., not losing cognitive ability, whereas others were declining.

Second, GLMMs can be used to obtain predicted values of the random effects, thereby allowing person-specific predictions of, in the SALSA example, declines in cognitive ability over time. These predicted values are based not only on the measured characteristics (e.g., female, diabetic, over the age of 60) but also on the observed values of the outcome for that individual.

Fitting and Inferential Methods

The fitting of GLMMs to data is computationally challenging and a number of approaches have been proposed. By far the most common method and a method implemented in many software packages is maximum likelihood based on numeric evaluation of the likelihood.

Maximum Likelihood

Maximum likelihood or minor variations of maximum likelihood, such as restricted maximum likelihood, are standard methods of estimation for linear mixed models and generalized linear models (e.g., logistic regression), the building blocks of GLMMs. It is, therefore, unsurprising that maximum likelihood is regularly used for GLMMs. Maximum likelihood (ML) is attractive because of features associated with the method in general, namely, the following:

Under mild conditions:

- ML estimators are asymptotically unbiased and normally distributed.
- The standard errors can be approximated using second derivatives of the log likelihood.
- ML estimators are asymptotically efficient, i.e., have the smallest variance.
- Hypothesis testing can be conducted with likelihood ratio, score, or Wald tests.

With ML estimators being asymptotically unbiased and normally distributed, it is straightforward to derive Wald tests based on large sample normality or chi-square distributions.

An exception in GLMMs to the usual large-sample distributions for hypothesis testing is when testing the null hypothesis that one (or more) of the variances of the random effects is zero (as noted above in testing for no variation in cognitive decline) using a likelihood ratio test. In such a case, the likelihood ratio test does not have the usual large sample chi-square distribution under the null hypothesis. Its large sample null distribution is a mixture of chi-square distributions (Self and Liang, 1987) and is sometimes called a $\overline{\chi}^2$ distribution. For a model with a single random effect, the use of the $\overline{\chi}^2$ test is straightforward and is achieved by halving the p-value from a standard, 1 degree of freedom chi-square test or, equivalently, by referring the test statistic to a chi-square distribution with the cutoff based on 2α. This is essentially the same as converting the p-value from a two-sided Z- or t-test to a one-sided hypothesis test.

Computing Maximum Likelihood Estimates

While the use of maximum likelihood for GLMMs is straightforward in concept, it is computationally difficult. This is because the model is specified conditional on the random effects, and calculation of the likelihood requires integrating out the random effects to obtain the marginal likelihood. Consider a simple logistic mixed model with a single, time-invariant predictor (x_i) and with only random intercepts that follow a normal distribution. Further assume that, conditional on the random intercepts, the outcomes, Y_{it}, are independent. The contribution to the likelihood from the data associated with the ith person is

$$\begin{aligned} \mathrm{P}\{\mathbf{Y}_i = y_i|\mathbf{x}_i\} &= \int \mathrm{P}\{\mathbf{Y}_i = y_i|u_i = u, \mathbf{x}_i\} f_{ui}(u)du \\ &= \int \prod_{t=1}^{n_i} \mathrm{P}\{Y_{it} = y_{it}|u_i = u, x_i\} f_{ui}(u)du \end{aligned} \quad [11]$$

$$\begin{aligned} &= \int \prod_{t=1}^{n_i} p_{it}^{y_{it}} (1-p_{it})^{1-y_{it}} f_{ui}(u)du \\ &= \int \prod_{t=1}^{n_i} p_i^{y_{it}} (1-p_i)^{1-y_{it}} f_{ui}(u)du \qquad [12] \\ &= \int \left(\frac{p_i}{1-p_i}\right)^{y_{i\cdot}} (1-p_i)^{n_i} f_{ui}(u)du \\ &= \int_{-\infty}^{\infty} \frac{e^{y_{i\cdot}(\beta_0+\beta_1 x_i+u)}}{(1+e^{\beta_0+\beta_1 x_i+u})^{n_i}} \frac{e^{-u^2/2\sigma_u^2}}{\sqrt{2\pi\sigma_u^2}} du \qquad [13] \end{aligned}$$

with $y_{i\cdot} = \sum_t y_{it}$ and the simplification in eqn [12] and the line below it occurring because x is constant within a person and therefore $p_{it} = p_i$. Because eqn [13] cannot be evaluated in closed form neither can be the likelihood, much less maximized in closed form. So even for this very simple logistic regression model with a single predictor and random intercepts, evaluation of the likelihood requires numerical methods.

Numerical Computations for Maximum Likelihood

Evaluation of the value of the likelihood for GLMMs is typically achieved either by numerical integration or other forms of numerical approximation. A typical method is to use an approach called Gauss–Hermite quadrature, which is appropriate for normally distributed random effects and numerically evaluates the integrals, such as those in eqn [11], with a quickly computed summation. The accuracy of the approximation of Gauss–Hermite quadrature can be improved by choosing more 'quadrature points,' albeit at the cost of making the computations slower. To improve the approximations, the integrands are often centered and scaled to make sure that the calculations are used in the region where eqn [11] has the most nonzero contribution. To speed computation some algorithms adaptively choose the number of quadrature points to be as small as possible but still achieve acceptable accuracy. In our experience using Gauss–Hermite quadrature in applications and when using centered and scaled algorithms we have found that using 20 quadrature points usually gives a good degree of approximation. For numerically difficult problems we have encountered inaccurate fits when using values less than 10.

The steps outlined above are merely to calculate the likelihood. However, to obtain model fits, the likelihood still has to be maximized. To avoid calculation of derivatives (which also cannot be evaluated in closed form and require integral approximations), quasi–Newton algorithms (with derivatives approximated with differences) or derivative-free maximization methods are often used to maximize the numerically evaluated likelihoods. Standard errors or predicted values require even more involved numerical approximations.

Other Numerical Approximations

Because numerical integration of the likelihood can sometimes be very slow or even infeasible, especially when dealing with multiple random effects, other approaches have been considered. One is the so-called *Laplace approximation* (e.g., Wolfinger, 1993; Raudenbush et al., 2000), which is an alternative method

of approximating the integrals in eqn [11]. It is also mainly applicable when the random effects are normally distributed.

Another approach to fitting GLMMs is motivated by adapting the most common fitting method for generalized linear models, which is to fit models by maximizing the log of the *quasi-likelihood*. This quasi-likelihood approach is attractive because it does not explicitly make distributional assumptions, but instead is based only on the assumed relationship of variance to mean. But integrating out the random effects from the quasi-likelihood is no easier than integrating them out of the true likelihood, so numerical approximations are again needed.

This leads to 'penalized quasi-likelihood,' which consists of a quasi-likelihood contribution with an added penalty term of $\sum_i u_i^2/2\sigma_u^2$. The penalty term can be motivated either as arising from assuming that the random effects follow a normal distribution or simply as a device to keep the individual random effects u_i relatively 'small.'

Because these methods do not have direct theoretical justification like maximum likelihood, their performance needs to be quantified with separate investigations, even for large samples. In general, these methods are considerably faster than numerical maximum likelihood but can give estimators with higher bias. Pinheiro and Chao (2006), via simulation, and Joe (2008), via evaluation of the asymptotic limit of maximizers of the Laplace-approximated likelihood show that the Laplace approximation approach can work well in some cases, but suffers from some degree of bias, especially with binary outcomes and with small cluster sizes. To address this bias some authors have explored using higher order Laplace approximations (Raudenbush et al., 2000).

Penalized quasi-likelihood methods seem to have somewhat more bias than Laplace methods. While they generally give acceptable answers when the outcome distribution is close to normal they can fail when the distributions are far from normal, e.g., binary or Poisson with small mean counts (Breslow and Clayton, 1993; Breslow and Lin, 1995; Lin and Breslow, 1996; Neuhaus and Segal, 1997). If outcomes are approximately normal, use of linear mixed models and reliance on their robustness (see below) is an option. So the very situation in which GLMMs are arguably most needed is when quasi-likelihood and Laplace methods work less well.

Bayesian Methods

GLMMs can also be modeled using Bayesian methods instead of frequentist approaches based on likelihood inference. Bayesian inference would also begin with a GLMM such as eqn [9]. As usual, a Bayesian approach would need to make further assumptions about the distributions for the parameters themselves, such as β, rather than regarding them as fixed, unknown constants, as does the likelihood approach. So, in a sense, all parameters are random effects under the Bayesian paradigm.

Some have argued that random effects models are inherently Bayesian because they assume a distribution for the parameters, but we feel differently. In the mixed model approach, random effects are modeling tangible distributions, while assuming distributions (in the Bayesian approach) for parameters does not. In the SALSA example, the random effect distribution represents the variation in person-specific quantities from the mixed model across the population of people. That is quite different than assuming a distribution for the effect of diabetes, β_D, from, say, eqn [6].

Software

In recent years, a number of software packages have incorporated and improved the analysis methods available for GLMMs. For example, Stata, SAS, SPSS, and R all have routines that can fit GLMMs. In Stata, the main routine is meglm (for mixed-effects generalized linear models) with many other more specialized mixed effects command, e.g., xtologit for ordinal logit models with random effects. In SAS, the glimmix and nlmixed procedures fit GLMMs. In SPSS, GLMMs are fit with genlinmixed, available in their Advanced Statistics Module and accessible through the menuing system under Analyze > Mixed Models. R has a number of routines including those in lme4, glmmML, and glmmPQL. The defaults and options for the methods for fitting GLMMs (e.g., maximum likelihood versus Laplace versus penalized quasi-likelihood) vary widely by package. In light of the possible bias with different fitting methods (see Section Other Numerical Approximations), it is prudent for the user to be clear on which method is being used.

Best Prediction

As noted in Section Types of Inferences Possible from GLMMs, an advantage of using GLMMs is the ability to derive person-specific predictions based not only on the covariates but also on the outcomes observed for that person. For example, in SALSA we could derive predicted values for the person-specific intercepts and slopes and thereby derive predicted trajectories of cognitive change. Working from the basic GLMM model [9], we would need an estimated value for $\mathbf{x}_i'\beta + \mathbf{z}_i'\mathbf{u}$. Maximum likelihood or another fitting method will yield estimates of β and the parameters in the random effects distribution, $f_u(u)$, but not estimated values of the $\mathbf{u}_i$. This is because the marginal likelihood is integrated over the random effects and they do not appear in the likelihood. So a different approach is needed to find estimates of the $\mathbf{u}_i$.

Most approaches calculate estimates derived to have good performance in prediction. The most common approach is to use estimators that minimize the squared prediction error, namely, to choose a predictor, $\tilde{u}$, of the random effect u according to

$$\tilde{u} \text{ minimizes } \mathrm{E}\left[(\tilde{u}-u)^2\right]. \qquad [14]$$

Simple mathematics shows that the minimizer is given by $\tilde{u} = \mathrm{E}[u|\mathbf{Y}]$. Intuitively this makes sense because the predicted value is chosen as the average of the random effects consistent with the observed data.

In general, the quantity $\mathrm{E}[u|\mathbf{Y}]$ depends on the unknown parameters β, as well as the parameters in the random effects distribution. So, in practice, the estimated values (e.g., from a maximum likelihood fit) are inserted in place of the true values. Predicted values are then calculated by using $\mathbf{x}_i'\widehat{\beta} + \mathbf{z}_i'\tilde{\mathbf{u}}$ in the model eqn [9].

The predicted values, while not necessarily calculated using Bayesian methods, are nevertheless closely related to Bayesian

quantities, as $E[u|\mathbf{Y}]$ is the posterior mean of the random effects. The predicted values are also sometimes characterized as 'shrinkage' estimators since they are closer to zero than estimates that would be obtained by treating the random effect as a fixed effect and deriving separate estimates for each level of the effect.

Diagnostics

Specification of a GLMM involves the steps outlined in Section The Model. Choices involve specifying the distribution of the outcome as well as the distribution of the random effects and choosing a link function.

Distribution of *Y*

As with modeling outcomes that are not correlated, choosing the distribution for the outcome is often challenging. When the data are binary the choice is easy, since only the Bernoulli distribution can be correct. But for numeric or count data, there can be little information to guide the choice. For example, consider an outcome such as number of emergency room visits during a year. On one hand it is a count variable, so the temptation might be to model it as following a Poisson distribution. In our experience, it almost certainly will have a distribution with a variance that is much larger than that required by a Poisson distribution. Thus, conducting an analysis assuming that the Poisson distribution holds could easily lead to misleading inferences with standard errors being too small, confidence intervals too narrow, as well as *p*-values that are falsely small.

With linear regression models a common diagnostic tool is the residual plot, with analysts considering histograms, quantile plots and scatterplots of residuals versus predictors and versus predicted values. Unfortunately, residual plots for GLMMs are not as easy to interpret. Consider the relatively simple case of a linear mixed model, for which the marginal mean is $\mathbf{x}_i'\beta$ but the mean conditional on the random effects is $\mathbf{x}_i'\beta + \mathbf{z}_i'\mathbf{u}$. So we can define two types of residuals, marginal ($r_{i,m}$) or conditional ($r_{i,c}$):

$$r_{i,m} = y_i - \mathbf{x}_i'\widehat{\beta} \quad \text{or} \qquad [15]$$

$$r_{i,c} = y_i - \left(\mathbf{x}_i'\widehat{\beta} + \mathbf{z}_i'\tilde{\mathbf{u}}\right) \qquad [16]$$

where $\tilde{\mathbf{u}}$ is the predicted value of $\mathbf{u}$. These are useful for identifying some basic issues: extreme nonnormality of the errors, outliers and nonlinearities (when plotted versus a predictor). However, they fail to have some of the other common features of residuals for linear regression models, making interpretation less straightforward. For example, they are not homoscedastic and are correlated with the predicted values. For nonnormal distributions and for nonlinear link functions, residual-based diagnostics tend to be of even less use. Even so, some rough guidance can be gained by plotting residuals versus predicted values to get a crude indication of the relationship of the standard deviation to the mean, which may inform a distributional choice.

A different tactic is to take a model-based approach to diagnosing specification of the distribution by trying more general or alternate distributions. For example, the negative binomial distribution contains the Poisson as a special case. So by fitting a negative binomial model and comparing it to a Poisson model, one can judge whether the extra flexibility allowed by the negative binomial model is needed (Booth et al., 2003). When more general distributional specifications are not available, one can try alternate distributions and rough comparisons can be made on the basis of the log likelihood values for the different fits. GLMM routines available in software allow a wide variety of distributional choices and procedures such as SAS's nlmixed even allows user-specified distributions.

The linear mixed model does have some favorable robustness properties (Jacqmin-Gadda et al., 2007), so it may be attractive to use linear mixed models for outcomes that are approximately normally distributed or can be transformed to be approximately normally distributed. Even when the outcome is nonnormally distributed, maximum likelihood estimators in linear mixed models are consistent (Richardson and Welsh, 1994).

Distribution of Random Effects

We (McCulloch and Neuhaus, 2011a) wrote a review of the performance of fitting GLMMs when various assumptions about the random effects distribution are violated, focusing on models with random intercepts, and have some additional work on random intercept and slope models (Neuhaus et al., 2013). In many situations, using maximum likelihood fitting with an assumed normal distribution for the random effects (which is what most software assumes) was highly robust. There were, however, two main violations of assumptions that can cause bias: (1) an association of the mean of the random effects distribution with a predictor (Neuhaus and McCulloch, 2006) and (2) an association of the variance of the random effects distribution with a predictor (Heagerty and Kurland, 2001).

An extensive literature exists on the effect of misspecifying the *shape* of the random effects distribution, for example, modeling it as normal when it is, in fact, skewed. The general conclusion is that statistical inferences (with few exceptions) are highly robust to this type of misspecification and that fits assuming normally distributed random effects perform quite well. Robustness extends to estimation of covariate effects, estimation of the random effects variance, and the mean square error of prediction of the realized value of random effects. Particularly robust in the face of a misspecified random intercepts distribution was estimation of the effects of covariates that mainly vary within clusters. This is especially important because evaluation of such covariates is a primary reason for conducting longitudinal or clustered data studies.

Minor exceptions to robustness to shape misspecification are estimation of the parameters that have associated random effects that are badly misspecified (e.g., the intercept when the random intercept distribution is badly misspecified). Also, although the mean square error of prediction of random effects is robust to misspecification, the distribution of the predicted values takes on the shape of the *assumed* form of the random effects distribution and thus the actual predicted values can be affected.

Between and within Covariate Decompositions

It is useful to distinguish three covariate types in clustered and longitudinal data: (1) between-cluster covariates, such as baseline covariates in a longitudinal study, that take the same value for all units in the cluster; (2) purely within-cluster covariates, such as the time variable in a longitudinal study with no missing data, that take different values for every unit in a cluster; and (3) within-cluster covariates such as subject age in the SALSA study that vary both within- and between-clusters. For both substantive and statistical reasons, it is useful to decompose covariates of the third type into between- and purely within-cluster components. In particular, motivated by the form covariates take on in conditional likelihood approaches (Neuhaus and Kalbfleisch, 1998), it is useful to decompose a covariate such as age in the SALSA study, AGE_{it}, into a between-cluster component $\overline{AGE}_i = (1/n_i)\sum_{t=1}^{n_i} AGE_{it}$ and a within-cluster component $AGE_{it} - \overline{AGE}_i$.

Covariate decomposition methods are useful statistically because they can remove confounding at the cluster level (Neuhaus and McCulloch, 2006) and because they focus interest on within-cluster effects which are robust to many forms of model misspecification (McCulloch and Neuhaus, 2011b). Comparison of the between- and within-covariate effects can therefore serve as a diagnostic method. Covariate decomposition methods apply generally to noncanonical link functions and multiple random effects and extend conditional likelihood like methods to these more general settings.

Social science investigators find it useful to separate covariate effects of individuals from those related to a collection of individuals such as a neighborhood and to allow these effects to differ (Enders, 2013). In the social science literature, the differing covariate effects are known as contextual effects and one could estimate them using the covariate decomposition methods above.

Missing Data

Missing data are ubiquitous in practice. People drop out of studies and refuse to answer some questions, or data values are out of range and need to be dropped. Software typically deals with missing data by dropping the observation if it is missing the outcome or any of the covariates, called a *complete case analysis*. This can be inefficient (when there is a patchwork of missing data) and can lead to bias, because data are often missing because of reasons related to the outcome being studied. For example, people experiencing cognitive impairment may be less likely to come in for visits.

The approach to dealing with missing data depends crucially on whether the outcomes are missing (perhaps along with all the covariates, such as when a person drops out of a study) or some of the predictors are missing. When predictors are missing, the most common method is *multiple imputation*. In multiple imputation, predicted values of the missing data are filled in using both the outcomes and predictors that are present. This requires building a model for the missing data as a function of the present data. This prediction is done with noise (to reflect the uncertainty involved in filling in the data) and is done multiple times so as to be able to assess standard errors and calculate confidence intervals and *p*-values while properly accounting for the uncertainty in missing data.

When it is primarily the outcomes that are missing, multiple imputation can still be used. However, simply by using maximum likelihood fits to the data gives considerable protection against missing data, and will often give nearly unbiased results (Laird, 1988). This is easier than multiple imputation because no prediction equations for the missing data need to be derived.

The maximum likelihood and multiple imputation strategies for missing data described above work under the technical assumption that the data are 'missing at random,' detailed discussion of which is beyond the scope of this article. For an applied discussion on missing data methods, see Chapter 11 of Vittinghoff et al. (2012); for a more technical discussion, see Little and Rubin (2002).

Summary

Our goal in this article was to give a brief overview of GLMMs, how they are fit to data, and the inferences possible when using them. The hallmark features of GLMMs are the ability, within a single class of statistical models, to handle a wide variety of distributions for the outcome, accommodate nonlinear models, and model correlated data. As regression methods, they are capable of estimation and testing of covariate effects. But they go beyond nonlinear regression modeling by being able to accommodate correlated data, draw inferences about correlation structures in the data and by being able to calculate predicted values that take into account not only covariates, but also observed outcomes. Software for fitting GLMMs is widely available, though care should be taken with which fitting method is being used since some of the methods can give biased answers for high nonnormally distributed outcomes.

Bibliography

Booth, J.G., Casella, G., Friedl, H., Hobert, J.P., 2003. Negative binomial loglinear mixed models. Statistical Modelling 3, 179–191.

Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88, 9–25.

Breslow, N.E., Lin, X., 1995. Bias correction in generalized linear mixed models with a single component of dispersion. Biometrika 82, 81–91.

Enders, C.K., 2013. Centering predictors and contextual effects. In: Scott, M.A., Simonoff, J.S., Marx, B.D. (Eds.), The SAGE Handbook of Multilevel Modeling. Sage, pp. 89–107.

Heagerty, P.J., Kurland, B.F., 2001. Misspecified maximum likelihood estimates and generalised linear mixed models. Biometrika 88, 973–985.

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molinac, J.-M., Thibaut, R., 2007. Robustness of the linear mixed model to misspecified error distribution. Computational Statistics & Data Analysis 51, 5142–5154.

Joe, H., 2008. Accuracy of Laplace approximation for discrete response mixed models. Computational Statistics and Data Analysis 52, 5066–5074.

Laird, N., 1988. Missing data in longitudinal studies. Statistics in Medicine 7, 305–315.

Lin, X., Breslow, N.E., 1996. Bias correction in generalized linear mixed models with multiple components of dispersion. Journal of the American Statistical Association 91, 1007–1016.

Little, R., Rubin, D., 2002. Statistical Analysis with Missing Data. Wiley, New York.

McCulloch, C.E., Neuhaus, J.M., 2011a. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. Statistical Science 3, 388–402.

McCulloch, C.E., Neuhaus, J.M., 2011b. Prediction of random effects in linear and generalized linear models under model misspecification. Biometrics 67, 270–279.

Neuhaus, J.M., Kalbfleisch, J.D., 1998. Between- and within- cluster covariate effects in the analysis of clustered data. Biometrics 54, 638–645.

Neuhaus, J.M., McCulloch, C.E., 2006. Separating between and within-cluster covariate effects using conditional and partitioning methods. Journal of the Royal Statistical Society, Series B 68, 859–872.

Neuhaus, J.M., McCulloch, C.E., Boylan, R., 2013. Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes. Statistics in Medicine 32, 2419–2429.

Neuhaus, J.M., Segal, M.R., 1997. An assessment of approximate maximum likelihood estimators in generalized linear mixed models. In: Gregoire, T.G., Brillinger, D.R., Diggle, P.J., RussekCohen, E., Warren, W.G., Wolfinger, R.D. (Eds.), Modelling Longitudinal and Spatially Correlated Data. Springer, New York, pp. 11–22.

Pinheiro, J.C., Chao, E.C., 2006. Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. Journal of Computational and Graphical Statistics 15 (1), 58–81.

Raudenbush, S.W., Yang, M.-L., Yosef, M., 2000. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. Journal of Computational and Graphical Statistics 9 (1), 141–157.

Richardson, A.M., Welsh, A.H., 1994. Asymptotic properties of restricted maximum likelihood (REML) estimates for hierarchical mixed linear models. The Australian Journal of Statistics 36, 31–43.

Self, S., Liang, K.-Y., 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association 82, 605–610.

Vittinghoff, E., Glidden, D., Shiboski, S., McCulloch, C., 2012. Regression Methods in Biostatistics: Linear, Logistic, Survival and Repeated Measures Models. Springer, Berlin.

Wolfinger, R.W., 1993. Laplace's approximation for nonlinear mixed models. Biometrika 80, 791–795.

Wu, J., Haan, M., Liang, J., Ghosh, D., Gonzalez, H., Herman, W., 2003. Impact of diabetes on cognitive function among older latinos. A population-based cohort study. Journal of Clinical Epidemiology 56, 686–693.

Generation in Anthropology

Sarah Lamb, Brandeis University, Waltham, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Generation has been a productive lens through which to study many facets of social–cultural life in anthropology. The concept of generation has been used by anthropologists to explain social change over time, to examine the ways people organize and envision intergenerational ties within the family, to explore principles of social organization beyond the family, and to identify differences among members of a society.

The concept of generation has been used by anthropologists to explain social change over time, to identify differences among members of a society, and to examine the ways people across cultures organize and envision family relationships, the life course, and the social-moral order. Generation also figures into the concept of culture central to anthropology: Anthropologists define culture as 'transgenerational,' transmitted from one generation to the next, at the same time that it is open to creative revision and change.

The term generation has two primary meanings. In one of its senses, generation refers to a group of people who are living through a time period together and participate in some kind of shared identity, practices, and beliefs. Generation is also used by anthropologists in its more genealogical, kinship-related sense, to refer to the relationship between parents and their children, these children and their children, and so on, and/or to forms of prestige and identity tied to one's position in such a system of kinship descent. Using one or both of these senses of the term, anthropological interest in generation has focused on four primary topics, pertaining to understandings of social change, of family moral systems and elder care, of social organization beyond the family, and of forms of identity and inequality. This article will look at each approach in turn.

Generation and Social Change

Generation has long been viewed as a lens through which to understand and imagine social change, both in the kinds of popular discourses studied by anthropologists and in scholarly models within anthropology and the related social sciences. Nineteenth-century social theorist August Comte located the impetus for social change in the younger generation, suggesting that increases in life span would slow down the tempo of social progress, because the conservative, restrictive influence of the older generation would operate for a longer time, should they live longer, thus thwarting the 'innovating instinct that distinguishes the young' (1974[1855]: 518). Karl Mannheim in his 1920s classic essay, *The Problem of Generations*, went on to speculate that cultural transformation comes about precisely through the continuous emergence of new generations: As each new generation comes into 'fresh contact' with its social and cultural heritage, it remodels what is found (1952[1927]: 293). French anthropologist, sociologist, and philosopher Pierre Bourdieu likewise implicitly thinks about the social change in generational terms when developing his influential theory of the habitus. Bourdieu writes of how the habitus of people born of different generations will often be in conflict, because their habitus have been produced under different circumstances:

> This is why generation conflicts oppose not age-classes separated by natural properties, but habitus which have been produced by different modes of generation, that is, by conditions of existence which, in imposing different definitions of the impossible, the possible, and the probable, cause one group to experience as natural or reasonable practices or aspirations which another group finds unthinkable or scandalous, and vice versa. (1977: 78)

Anthropologist Meyer Fortes similarly muses that a generational model is an 'apt imagery by means of which to depict continuities and discontinuities in a community's social and cultural life over a stretch of time' (1984: 106). It is within the framework of the succession of generations, Fortes explains, that all people acquire their cultural heritage: "Mediated from earliest infancy by the natal family, the society readily takes on a generational form to the individual" (p. 106). The popular folk model of the 'generation gap' also envisions social change in generational terms, imagining the young generation moving forward, separated by a 'gap' from, the more static, 'old-fashioned' old.

Despite the enormous explanatory power given to generational change by both local people themselves and in classic social science texts, the concept of generation has never taken center stage in anthropological ethnography or theory. Yet over recent years, several anthropologists have claimed that generation is a key site through which to understand in particular the social processes associated with modernity and globalization (Cole and Durham, 2007; Cole, 2010; LiPuma, 2000). Edward LiPuma's ethnography examines the process of encompassment of the Maring of Highland New Guinea by colonialism, capitalism, and Christianity – "what is known as modernity" (2000: xi), and he argues that "there can be no theory of transformation without a theory and ethnography of generations, a concept often presupposed in anthropological discourse but rarely spoken about" (2000: xiii). LiPuma examines the period from roughly 1955 to 1980, when (1) the oldest generation (those generally over about age 60) could recall a life unfettered by foreign interventions, (2) the middle generation in power (aged 30s to mid 50s) had been brought up in a traditional world but lived now wholly in the

modern, and (3) the younger generation had never known a world untouched by missions and government and capitalism (p. 16). LiPuma contends that by examining the changing categories of knowledge and structures of desire across these generations, he is able to illuminate the broader dynamics of social-cultural transformation in Melanesia and beyond (p. 63), arguing that "an understanding of generations is essential to an account of… the creation of the modern" (p. 64).

In her ethnography of the coming of age of a generation of women in neoliberal Madagascar, Jennifer Cole (2010) examines "the ways new generations come to have a special force in social change" (p. 190). Her text foregrounds generation as a productive theoretical concept while cautioning that both popular and scholarly ways of interpreting generational change "offer a symbolic shorthand for understanding more complex dynamics" (p. 5). "This is never a mechanical, unreflexive matter of one cohort replacing another as some theories of generational change have implied," Cole reflects. "Rather, it is a highly imaginative undertaking in which people's desire to be good and valued, their culturally shaped dilemmas, and their visions for the future meet the broader forces of history and culture in unpredictable, uneven ways" (p. xii).

Because youth figures so prominently in how social change is imagined in both scholarly and colloquial discourses, it is perhaps not surprising that youth, more than any other age group, has been the focus of anthropological studies of generational change. A recent surge of work emphasizes the transformative impact of 'youth cultures' and 'youth agency' in the contemporary global era (Lamb, 2009: 15–16, 276, note 30). Yet one of the central arguments of Lamb's (2009) ethnography focusing on older people in India and abroad is that it is inadequate to think of the young as the only or perhaps even primary locus of social-cultural change. Lamb's data show that, even though they tend to be viewed by others and to view themselves as in many ways emblematic of 'tradition,' those of the senior generation are at the same time very often actively involved in fashioning new modes of life for themselves and their descendants.

The Cultural Construction of Intergenerational Ties

A second major body of anthropological work examines relations between generations of kin. Such work on the cultural shaping of intergenerational ties foregrounds generation in its genealogical sense as a descent relationship. Interest in intergenerational family relationships has existed since the beginnings of anthropology as a discipline, largely due to the analysis paid by anthropologists to kinship systems. Since about the 1970s, the majority of work on intergenerational ties has come out of the subfield of anthropology of aging, where one of the central research issues has surrounded the degree to which the elderly can expect to receive various forms of support from their younger relatives. Research on intergenerational ties, however, goes well beyond the (interesting and important) problem of how the elderly are cared for, to speak to broader conceptions about the nature of the family, the relationship among individuals, genders, families, the market and the state, and people's visions of the social-moral order.

Lamb (2000, 2009) examines such problems through a focus on West Bengal, India. She finds that Bengalis perceive intergenerational family relations to entail long-term bonds of reciprocal indebtedness extending throughout life and even after death. People widely believe that it is natural for adult children to provide care for their elderly parents, reconstruct relations with parents as ancestors after death, and ritually nourish these ancestors, as a means of repaying the tremendous 'debts' they owe their parents for producing and caring for them in infancy and childhood. What parents once gave their children to produce and raise them – such as a body in birth, food, material goods, money, a home, forms of love, the cleaning of urine and bowel movements – children are expected to provide in return for their parents years later in old age and after death. (Children, for instance, construct new spirit bodies for their parents as ancestors, parallel to the bodies parents create for their children in birth.) Other anthropologists have found that relationships of lifetime intergenerational reciprocity, conceptualized and practiced in varying ways, are not uncommon cross-culturally (Alber et al., 2008; Ikels, 2004; Sokolovsky, 2009).

The shaping of intergenerational ties speaks not only to aging and the family per se but also, frequently, to wider debates concerning the social, moral, and political conditions of a broader society and nation. Indian discourses, for instance, commonly present close, multigenerational families as signifying a 'good' and 'Indian' society; whereas disintegrating generational ties – the breakup of joint families, a rise in individualism, and the emergence of old-age homes – are held to be paradigmatic signs of an ambivalently valued modern, Westernizing, and globalizing era (Lamb, 2009). Similarly, people in Japan and China are reworking and debating the long-held principle of filial piety in the face of transitions such as population aging, the one-child policy in China, resistance on the part of daughters-in-law, and the younger generation's increasing desire for independence and generational equality (Ikels, 2004).

Anthropologists have thus found that the shaping of intergenerational ties is powerfully linked with surrounding cultural values. Andrei Simic (1983) suggests links between an American model of intergenerational relations and a broader cultural orientation toward self-fulfillment and individualism. Children in middle-class, Anglo-American households maintain their own rooms and possessions, and by about age 18 are expected to move out and live separately from their parents. "Though parents often extend financial aid to their children after they have left home," Simic observes, "it is almost always with the overt rationale of helping them get established as independent units, rather than creating a binding reciprocity of mutual dependence between the two generations" (p. 86). In later life, this same ideology prevails: among the aged economic independence is a fundamental value; dependency is 'demeaning' (p. 86).

In general, research has shown that in small-scale, developing societies, intergenerational family relations tend to be more mutually interdependent than in capitalist, postindustrial settings, although many important exceptions can be found (Albert and Cattell, 1994: 87–96; Sokolovsky, 2009: 110–111). A wealth of careful, detailed, cross-cultural data on generational ties has been produced by anthropologists, revealing a profound variety in

the ways intergenerational relationships are conceived, maintained, practiced, and given value.

Generation as a Principle for Structuring Society

Anthropologists have also examined the ways social and cultural groups employ some concept of generation to organize social relationships beyond the family. Many societies group people into generation-defined categories, taking some notion of genealogical succession within the family and extending it to think about broader social relations (Kertzer, 1979). Examining generation in this sense, as a principle of social organization, has allowed especially social anthropologists and those concerned with social structure to address questions about the very ordering and continuity of societies (Alber et al., 2008; Baxter and Almagor, 1978; Eisenstadt, 1956; Fortes, 1984).

In some societies, social generational ordering is highly formalized and elaborate. Rituals, variations in dress, forms of speech, permitted behaviors, specialized tasks, and places of residence may sharply distinguish persons of one generational group from another. The most highly institutionalized organizations of generations into social groupings, often referred to as *generation-set systems*, have been found among a number of East African groups (Baxter and Almagor, 1978). In such groups, the generation model is extended so widely as to provide the basis for regulating social relationships in an entire society. Every male is a member of a named generation set, and the generation set a man is in is determined by the generation set his father is in.

The crucial rule in such generation-set systems is that successive genealogical generations of fathers and sons must be socially and politically separated. Some scholars suggest that this is a way of minimizing generational conflict and competition, for instance over wives, land, and political leadership, for members of different generation sets generally cannot compete over the same resources and privileges. Generation-set systems have almost always been organized around males, and women may be tied to them peripherally as daughters, wives, and mothers. Versions of such systems have also been found in Melanesia and Polynesia, among Australian aborigines, and among some Native American groups.

Even in societies where people are not organized into formal generation sets, some notion of generation as a principle of social organization is almost universally employed (Kertzer, 1979). For instance, the social roles or identities a person assumes may depend on whether he or she represents the junior or senior generation in a family. In many societies, a person's transition to social adulthood is linked to the birth of his or her children; and in India, the transition to a 'senior' stage in life and society comes about not so much through reaching a particular chronological age but through the marriage of one's children (Lamb, 2000).

It is necessary to distinguish between generation and chronological age as principles of social organization and of cultural value. Most societies use some version of both in ordering social relations, but these categories are often confused or overlapping in social scientific analyses and among the people being studied (Kertzer, 1979). Age, however, is distinct from generation: it has to do with the chronological passage of time, and rests on the formally isolated individual as the unit of reference; whereas generation is family generated. There are many situations in which relationships by generation conflict with relationships by age. For instance, among the Tallensi, no man, even if very advanced in age, can achieve jural autonomy until his father dies (Fortes, 1984). In the United States, social criticism may ensue if a man marries a woman 'young enough to be his daughter' (the woman as wife becomes the man's generation, but has the chronological age of his children's generation).

Several anthropologists have been intrigued by the fact that what is essentially an intrafamily principle of genealogical generations has been used so widely by human groups to think about social relations outside the family. People of diverse cultures seem to find families 'good to think with.'

Generational Identity as a Facet of Social Difference

Finally, a fourth way anthropologists have approached the study of generation is as a form of identity. Since around the 1980s, anthropology has turned away from generalizations based on culture to consider the importance of difference within structures of power. The concept of generation as a form of differential identity can be used to heed the multiple, competing perspectives, experiences, and positions of inequality that make up any society or culture, as it intersects with other axes of difference such as class, race, gender, sexuality, ability, and age.

Sherry Ortner is one anthropologist to consider generation as a feature of differentiating prestige systems in her 1996 collection of essays, *Making Gender: The Politics and Erotics of Culture*. Ortner explores how generation can intersect with principles such as gender, class, and birth order in diverse ways in a range of societies, allowing her to move beyond what she describes as the 'by now rather sterile debates' over the universality of male dominance (p. 172). In a reanalysis of A.R. Radcliffe-Brown's 1920s work on the Andaman islanders, Ortner suggests a "basically gender-undifferentiated prestige hegemony" in which "age and generation were more significant axes of authority than was gender" (1996: 152–153).

Jean and John Comaroffs (1999) take a powerful look at generation as one of the dominant idioms through which youths in postcolonial South Africa frame their identities and place in the world. Widespread anxiety over the hardening materialities of life in an era of global capital translates frequently into bitter generational opposition. Many young black men blame their current incapacity to ensure a viable future for themselves on a perceived sinister, all-consuming aged elite, referred to as 'witches' or 'old ladies' (even when they are men). The Comaroffs cite one example of a group of youths pelting an old woman with sticks and stones. "'Why are you killing me, my grandchildren?' she wailed. Her assailants responded, 'Die, die you witch. We can't get work because of you!'" (p. 287). The major lines of opposition in contemporary South Africa, the Comaroffs find, are not race or class, but generation, mediated by gender. Several related essays are also published in Alber et al. (2008).

Other studies use a language of generational identity (as in post-Soviet generation, post-Civil Rights generation, one-child-policy generation) to examine historically situated processes

of identity construction that can explain significant differences in experiences, ideologies, and practices among diverse members of a society or nation.

Such work reveals the potential salience of generation as a category of social experience and analysis. A society at any given moment may be perceived, by the people within it, and/or by the anthropologists studying it, as being crosscut by generations representing powerfully divergent values, agendas, and positions of prestige.

The Importance of Generation in Future Work

Generation has been a productive lens through which to study many facets of social-cultural life in anthropology. As Alber, van der Geest, and Whyte argue, generation is a powerful analytic tool for studying society, as "it implies relations in time. Whether we think of intergenerational links within families or across historical periods, generation is about connections and contrasts – and often conflicts – in a temporal perspective" (2008: 1). The concept of generation has not emerged as a central analytic issue in the discipline, however. Many works (including several of those cited in this article) that do, in fact, examine generation in fruitful ways do not explicitly foreground their focus on generation as a central dimension of their projects. Interest in the concept of generation may, however, be increasing in the field. The category of generation can be fruitfully exploded to explore problems of current critical interest to the discipline – pertaining to the ways people envision and experience social change, how family moral systems and forms of elder care are fashioned under changing cultural and historical circumstances, and the ways diverse power structures, voices, and identities make up particular cultural worlds.

See also: Age: Anthropological Aspects; Family: The Anthropology of the Concept and Its History; Generations, Sociology of; Globalization, Anthropology of; Identity in Anthropology; Kinship in Anthropology; Tradition, Anthropology of; Youth Culture, Anthropology of.

Bibliography

Alber, E., van der Geest, S., Whyte, S.R. (Eds.), 2008. Generations in Africa: Connections and Conflicts. LIT Verlag, London.

Albert, S.M., Cattell, M.G., 1994. Old Age in Global Perspective: Cross-Cultural and Cross-National Views. Macmillan, New York.

Baxter, P.T.W., Almagor, U. (Eds.), 1978. Age, Generation and Time: Some Features of East African Age Organisations. C. Hurst, London.

Bourdieu, P., 1977. Outline of a Theory of Practice (R. Nice, Trans.). Cambridge University Press, Cambridge.

Cole, J., 2010. Sex and Salvation: Imagining the Future in Madagascar. University of Chicago Press, Chicago, IL.

Cole, J., Durham, D. (Eds.), 2007. Generations and Globalization: Youth, Age, and Family in the New World Economy. Indiana University Press, Bloomington, CA.

Comaroff, J., Comaroff, J.L., 1999. Occult economies and the violence of abstraction: notes from the South African postcolony. American Ethnologist 26, 279–303.

Comte, A., 1974[1855]. The Positive Philosophy of Auguste Comte (translated and condensed by H. Martineau, with an introduction by A.S. Blumberg). AMS Press, Blumberg, NY.

Eisenstadt, S.N., 1956. From Generation to Generation: Age Groups and Social Structure. Free Press, New York.

Fortes, M., 1984. Age, generation and social structure. In: Kertzer, D.I., Keith, J. (Eds.), Age and Anthropological Theory. Cornell University Press, Ithaca, NY, pp. 99–122.

Ikels, C. (Ed.), 2004. Filial Piety: Practice and Discourse in Contemporary East Asia. Stanford University Press, Palo Alto, CA.

Kertzer, D.I., 1979. Generation and age in cross-cultural perspective. In: Riley, M.W., Abeles, R.P., Teitelbaum, M.S. (Eds.), Aging from Birth to Death: Volume II: Sociotemporal Perspectives, AAAS Selected Symposium, vol. 79. Westview Press, Boulder, CO.

Lamb, S., 2000. White Saris and Sweet Mangoes: Aging, Gender and Body in India. University of California Press, Berkeley, CA.

Lamb, S., 2009. Aging and the Indian Diaspora: Cosmopolitan Families in India and Abroad. Indiana University Press, Bloomington, CA.

LiPuma, E., 2000. Encompassing Others: The Magic of Modernity in Melanesia. University of Michigan Press, Ann Arbor, MI.

Mannheim, K., 1952[1927]. The problem of generations. In: Keckskemeti, P. (Ed.), Essays on the Sociology of Knowledge. Routledge, London, pp. 276–320.

Ortner, S., 1996. Making Gender: The Politics and Erotics of Culture. Beacon Press, Boston, MA.

Simic, A., 1983. Aging in the United States and Yugoslavia: contrasting models of intergenerational relationships. In: Sokolovsky, J. (Ed.), Growing Old in Different Societies: Cross-Cultural Perspectives. Wadsworth, Belmont, CA.

Sokolovsky, J. (Ed.), 2009. The Cultural Context of Aging: Worldwide Perspectives, third ed. Bergin and Garvey, New York.

Relevant Website

http://faculty.usfsp.edu/jsokolov/ – Cultural Context of Aging home page.

Generations in History

Anthony Esler, College of William and Mary, Williamsburg, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article offers a summary of the theory of social generations as a force in human history. It points out that rebellious youth has often challenged authority and that rebellious younger generations have become more common in recent times. It traces the emergence of a theory of generational conflict over the last century and offers some critique of this view of change in history and its consequences.

Introduction

Generations are groups of human beings born about the same time and shaped by the same social structure and the same span of history. These subdivisions of society have played a more important part in the flow of human events than is commonly realized. During the nineteenth and twentieth centuries in particular, younger generations played colorful, sometimes revolutionary roles in modern history. As awareness of the importance of generational conflict has grown, an evolving *theory of social generations* has developed to explain this phenomenon. Thinkers such as Karl Mannheim, José Ortega y Gasset, and a host of more recent scholars have analyzed the impact of social and cultural changes and of traumatic events in the shaping of human generations. Their research has also suggested that generational activism itself has increasingly become a force for social change.

Age Groups in Human Society

Age groups have been recognized in most societies and have been elaborately institutionalized in some. Like such other subsets of society as race, class, and gender, age groups have been both active forces in social conflict and creative agents of social change.

Anthropologists have detected age groups (age grades or age sets) among many preurban peoples. In these cultures, progress from one age grade to the next may be celebrated by elaborate rites of passage, and membership in a particular age group often brings both obligations and privileges. Young men will be expected to fight wars, young men and women may look forward to marriage, and both genders will have traditional work duties. Older people of both sexes may serve as decision-making elders or perform important ceremonial functions in the community.

But age groups are detectable in more complex urban societies also. Apprentices in guilds, students in schools, and conscripts in armies are subgroups in society normally defined by age. Economic and political authority is commonly exercised by middle-aged and older people. Even the elderly may be consulted for their long experience or be assigned some modest function in the community.

In most preurban cultures, younger and older people tend to accept traditional roles, rights, and social limits defined in significant part by age. In more complex societies, however, *generational conflict* between age groups is detectable – and seems to have become considerably more common in recent centuries.

There is evidence of this sort of conflict between age groups in earlier urban societies than ours. In ancient Greece, for instance, Aristotle saw differences between youth and age at the heart of the social wars that frequently rent the fabric of Greek city-states. Medieval observers were dismayed by the riotous behavior of guild apprentices or the 'town and gown' violence of university students. Many of the leaders of the European Reformation were young preachers, and historians have noted that the 'founding fathers' of the United States were actually young men.

The clearest and most impressive examples of generational conflict, however, seem to have emerged in the nineteenth and twentieth centuries.

Youth in Modern History

Western youth in 1800 had inherited an unprecedented surge of social transformation. The 1700s had seen continuing expansion of Europe's overseas empires, the beginnings of the industrial revolution, and early steps toward democracy in Europe and America. Through the nineteenth and twentieth centuries, younger generations would continue to grow up in a world of challenge and change – and would add their own vigorous demands for more of the same. Surges of reform, revolution, or even utopian subcultures thus recur throughout the history of this period.

Generational revolt was common in the nineteenth century as young people swarmed into the streets of European cities demanding political, economic, and social changes in society. *Youth movements* like the German Student Unions, the youthful French barricade fighters of the 1830s and 1848, Mazzini's Young Italy, Young Ireland, the Young Czechs, and a series of increasingly violent Russian youth movements challenged authority across the continent.

The twentieth century saw the spread of rebellious younger generations around the globe. The first half of the century saw Lenin's Communists overthrow the tsars in Russia, the Young Turks seize power in Istanbul, Gandhi build youthful opposition to British rule in India, and a series of Chinese student movements paving the way for Mao's seizure of power after World War II.

During the second half of the century, young rebels liberated colonies from European rule across Africa and parts of Asia, and young women moved to transform the workforce of the western world. In the turbulent 1960s in particular, the United States was shaken by the youthful black liberation and the anti-Vietnam War movements; young demonstrators swarmed into the streets in West Germany, France, and other European countries; a young Fidel Castro cemented his power in Cuba; and an aged Chairman Mao turned his young Red Guards loose to revive the revolutionary spirit – and saw them wreak havoc from the public squares of Chinese cities to the monasteries of far Tibet.

These two centuries of change and challenge also saw rebellious youth exercise another option besides rebellion – *withdrawal from society* in increasing numbers to carve out new lifeways for themselves.

Thus the 1800s saw young English Romantic poets praising the beauties of the natural world and the simple lives of country folk. Later in the century, young French 'bohemians' – secessionist youth identifying with wandering gypsies – appeared in the slums of Left Bank Paris, rejecting materialism to live free and poor along the wrong side of the Seine. In the first half of the twentieth century, secession from bourgeois society spread to London's Soho, New York's Greenwich Village, and similar enclaves across the western world. Later generations in many places substituted social withdrawal for political action. Arguing over the latest wave of new music, and inhaling or injecting new drugs, they turned their backs on the world their parents had built.

The sheer scale of these multifarious challenges to the existing order clearly needed an explanation. The idea that seems to answer this need best is the notion of the social generation.

Emergence of the Generational Idea

Theoretical understanding of the generational dimension of human society had in fact evolved as conflict between generations grew more visible.

Earlier cultures had developed at least some awareness of this factor in society. In ancient times, Old Testament genealogies and Homeric poetry reveal clear concern with the succession of human generations. The traditional father of history, Herodotus, posited 33-year generations in his *History of the Persian War*. The Arab world historian Ibn Khaldun proposed a basic theory of generations in his *Muqaddimah: An Introduction to History*.

Only in the nineteenth century, however, did modern social thinkers of the caliber of Auguste Comte, John Stuart Mill, and Wilhelm Dilthey begin to develop a coherent and convincing theory of social generations. Julián Marías, a leading historian of the generational idea, emphasized the major contributions of these men. Comte, Marías pointed out, had clearly recognized the succession of generations as an important causal factor in social change. Mill was well aware that the generational concept must be expanded beyond the traditional familial or genealogical framework to encompass society as a whole. And Dilthey had attempted to analyze the basic structure and broader significance of the social generation in human society.

The first half of the twentieth century – particularly the period between the two world wars – saw classic formulations of the theory of social generations by the Spanish philosopher Jose Ortega y Gasset (1923, 1933) and the German sociologist Karl Mannheim (1928).

Ortega's account of the phenomenon focused on the internal development of the typical generation. He divided the human life span into four 15-year stages – childhood, youth, initiation to positions of power, and domination of society. This stages-of-life perspective, rooted in the psychological and social stages of human development, has influenced many generationists since.

Karl Mannheim's analysis emphasized the roots of generational differences in the changing social and cultural circumstances that shape each new generation. Emerging generations are thus influenced by their social *location* in a particular social class, by the rural or urban environment in which they grow up, and by new trends in the society in which they live. In changing times, generations may develop different points of view and patterns of behavior from those prevalent among their elders. Mannheim also analyzed the structure of the social generation. He suggested that a given birth cohort might be subdivided into a number of distinct *generation units*, social subgroups who 'work up' the common generational experience in very different, sometimes ideologically conflicting ways.

In the wake of the global pandemic of youth revolts in the 1960s, finally, scholars and other concerned commentators probed the patterns of generational development and conflict as never before. Such students of the generational phenomenon as T. Allen Lambert (1972), A. B. Spitzer (1973), Marvin Rintala (1979), Richard Braungart (1986), and Shmuel N. Eisenstadt (2005) developed a convincing model of the social generation as a force as powerful as the social class in human history. Other generationists, including D. Sinha (1972), Anthony Esler (1972), Wilhelm Pinder (1979), Mark Roseman (1995), and Joe Austin and Michael N. Willard (1998), focused on the role of generational conflict in the history of particular countries and cultures over the nineteenth and twentieth centuries.

These scholars probed the sociopsychological roots of generational consciousness in immense detail. They documented the persistence of a sense of generational identity even after a strong involvement with one's youthful peer group has faded. And they collected masses of contemporary evidence of the existence of generational conflict and of its influence on social change.

The Theory of Social Generations

In the twenty-first century, generational analysis is used by historians, political scientists, sociologists, anthropologists, and critics of literature and the arts. The concept has also proved useful to journalists, politicians, marketing experts, and many others concerned with understanding or influencing society. These diverse commentators explore *age groups* just as they do groups defined by social class, gender, or ethnicity. The brief exposition of a developed, essentially Mannheimian view of the social generation that follows will also draw on this broad range of insights.

A *social generation*, then, is a *birth cohort*, a group of individuals born in the same place and period and in similar circumstances. Shaped by common cultural, social, economic, and other influences, members of the same generation develop similar perspectives during their early years. Children of wealth, private schooling, and unlimited opportunity will grow into a very different kind of adult from the offspring of poverty, little opportunity, and no preparation to take advantage of any that should appear.

Within this framework, common family structures, child-rearing methods, and educational practices begin the process of *socialization*. A patriarchal philosophy of "spare the rod and spoil the child" and an education built around rote memorization, for instance, may tend to produce one sort of *generational conscious*. Permissive child-rearing methods and educational patterns dedicated to developing the individual propensities of students may begin the shaping of a very different sort of social generation.

As the cohort matures through adolescence and youth, the larger social framework within which this process takes place also exerts powerful shaping influences. Hierarchical or more egalitarian societies, traditionalist or rapidly evolving cultures, authoritarian or democratic political systems – all of these have an impact on young people as they emerge into the world beyond the family.

Within this larger milieu, the specific *social location* within which a particular subgroup of a given generation matures will also have its influence. Young people raised in working-class communities may grow up with a very different view of the world from that of youth raised in middle-class suburbs or among the even more affluent. At every level, the pages young people read, the songs they hear and sing – and, in more recent times, the films they watch and the electronic world they explore – shape the perspectives, beliefs, and values of a social generation.

To these influences of home, school, and society, however, historical study of the generational phenomenon would add a special emphasis on *traumatic historical events* or *disturbing social trends* in the shaping of a rebellious younger generation.

Accounts of specific generations have tended to stress this impact of disturbing events upon the young. Unsuccessful wars or brutal economic downturns may turn the younger generation against a society that seems to have failed. Such historic cataclysms may deeply undermine the view of the world – and their place in it – that these young people had been taught in childhood. The result may be a disillusioned and rebellious generation that turns against the world their elders made.

The mores of a younger generation may thus be drastically reshaped as they shrug off the prescriptions of their 'deauthoritized' elders (Feuer, 1969). Especially where antiestablishment ideas are already in the air, such historical traumas may turn advocates of change into leaders of generational rebellion.

It is sometimes difficult to trace the long-term impact of generational revolts. Such youthful upheavals as the wave of political revolutions against European monarchies that exploded in 1848 – or the surge of demonstrations by young black Americans against a segregated society in the 1960s – did not transform either European politics or American society overnight.

In time, however, the activism of rebellious youth often does leave its mark on history. The European 'forty-eighters' were vigorously suppressed by royal troops in the Revolutions of 1848. But in the decades that followed, the legislatures of at least some European nations grew more democratic and claimed increasing power. And in the wake of World War I, the hereditary monarchies and aristocracies faded away as a primary force in European politics.

In the 1960s in the United States, the young black demonstrators who followed Martin Luther King, Jr., into the streets got some liberalizing laws passed but did not abolish the United States' inner-city ghettos, and King himself fell to an assassin's bullet. But the demonstrators of the 1960s had opened doors for some black Americans, and white attitudes did begin to change. Fifty years on, King's statue stood on the Mall in Washington, and a black president sat in the White House.

Problems and Potential of Generational Analysis

The range of specific uses to which generational analysis has been put is broad and continues to expand.

There is a large literature focusing on younger generations mobilized in politically motivated student or youth movements. The generational approach has also been utilized by historians and critics to characterize the periods of transformation in literature and the arts. Other scholars have probed social innovations and social trends pioneered by younger generations. Studies of the liberation of women over the last century and of the more recent electronic transformation of work and play have found generational analysis useful. And attempts to estimate the damage done to a society by such group experiences as a war or a major depression often detect clear generational divides.

Like any attempt to explain the group lives of human beings, however, the theory of social generations has raised its share of challenging questions.

A generation is by definition a group of people born during the same span of time – but how long should that span be? Thirty-three years as Herodotus declared, 15 years with Ortega y Gasset, or as little as 3 years, as some modern demographic studies might suggest? In the end, the length of a generation that makes a difference in history seems to depend on the pace of social change in the period in question. The more rapidly the shaping social matrix changes, the more quickly enough changes accumulate to produce a new generational point of view.

Another important question for the skeptic, perhaps, would be "Are social generations in fact as *real* a force in society as race, class, or gender?" Some observers have noted that even when the streets surge with young demonstrators, most young people seem to be at home, sharing their parents' lives and views. In history, however, numerical majorities are not necessary to produce significant consequences. When socialization fails with a critical mass of the young, the resulting surge of generational activism can produce revolutionary changes in the world they live in.

But, yet another body of generational research suggests, are there not also identifiable generations whose shared qualities

of mind and feeling do not in fact exercise powerful influence on the course of history? Are there cohorts identifiable through lifestyles and mind-sets, customs and costumes, who often seem to be not shapers of the course of events in their time but victims of it? Highly developed generational subcultures may in fact emerge with neither a transforming impulse nor a belief that significant change is possible. But romantics and bohemians, and the 'Lost' and 'Beat' generations, may still leave artistic or cultural legacies that are no small gift to their successors.

And a final query, rooted in the present time – early in the twenty-first century – might be "How active a force for change *are* younger generations in our time?"

A decade into the 2000s – and half a century after the 1960s – a new surge of youthful protest seems in fact to be building around the world. The 'Arab Spring' of 2011 toppled old dictatorships in Egypt, Libya, and other Muslim lands in North Africa and the Middle East. Desperate economic times have recently filled European streets from Greece to Portugal with demonstrators and burning buildings. In China, an astonishing 200,000 local protests, often over land-devouring urbanization, have challenged the authoritarian government in Beijing over the last 5 years. And in the United States, jobless young people moved in to 'Occupy Wall Street' at the end of 2011 – and youthful dissenters were soon mobilizing to 'occupy' key protest sites in their own countries around the globe.

It might not be 'the 1960s' yet, but it does suggest a future for generational activism as a force in human history.

See also: Prosopography (Collective Biography).

Bibliography

Austin, J., Willard, M.N., 1998. Generations of Youth: Youth Cultures and History in Twentieth Century America. New York University Press, New York.

Bengtson, V.L., Laufer, R.S. (Eds.), 1974. Youth, generation, and social change. Journal of Social Issues 30 (2 and 3).

Braungart, R.G., Braungart, M.M., 1986. Life course and generational politics. Annual Review of Sociology 12, 205–231.

Eisenstadt, S.N., 2003. From Generation to Generation. Transaction Publishers, London.

Esler, A., 1972. Bombs Beards and Barricades: 150 Years of Youth in Revolt. Stein and Day, New York.

Feuer, L.S., 1969. The Conflict of Generations: The Character and Significance of Student Movements. Basic Books, New York.

Lambert, T.A., 1972. Generations and change: toward a theory of generations as a force in historical process. Youth and Society 4 (1), 21–45.

Mannheim, K., 1928. Das problem der generationen. Kölner Vierteljahrshefte für Soziologie 7, 157–185. 309–329 [1952 The problem of generations. In: Essays on the Sociology of Knowledge. Oxford University Press, New York 276–320].

Marías, J., 1967. El método histórico de las generaciones. Revista de Occidente, Madrid [1970 Generations: A Historical Method. University of Alabama Press, University, ALA].

Ortega y Gasset, J., 1923. El tema de nuestro tiempo. Revista de Occidente, Madrid [1933 The Modern Theme. W. W. Norton, New York] and 1933 En torno a Galileo. Revista de Occidente, Madrid [1958 Man and Crisis. Norton, New York].

Ortega y Gasset, J., 1933. En torno a Galileo. Revista do Occidente, Madrid [1958 Man and Crisis. W.W. Norton, New York].

Pinder, W., 1926. Das Problem der Generation in der Kunstgehichte Europas. Frankfurter Verlags-Anstalt, Berlin.

Rintala, M., 1979. The Constitution of Silence. Essays on Generational Themes Greenwood Press, Westport, Conn.

Roseman, M., 1995. Generations in Conflict: Youth Revolt and Generation Formation in Germany 1770–1908. Cambridge University Press, Cambridge.

Sinha, D., 1972. The Mughal Syndrome: A Psychological Study of Intergenerational Differences. Tata McGraw-Hill, Bombay.

Spitzer, A.B., 1973. The historical problem of generations. American Historical Review 78, 1353–1385.

Generations: Political

Paul R Abramson, Michigan State University, East Lansing, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P.R. Abramson, volume 9, pp. 6050–6053, © 2001, Elsevier Ltd., updated by Editor.

Abstract

Political generations are politically distinctive age groups delimited by their years of birth or by some major political event. Societies are transformed through the processes of birth, aging, and death, but this transformation may or may not lead to political change. This article examines the way replacement has contributed to the rise of the Democratic Party in the United States during the 1930s and 1940s, increased tolerance in the last two decades, and the decline of electoral participation since 1960. In Britain, replacement led to the dominance of the Labour Party after World War II, but it later led to the decline in party loyalties. In Israel, it led to the decline of Labour. Through his study of value change, Inglehart has made the most important contribution to the study of generational change, and the data he collected between 1970 and 2006 strongly suggest that replacement was a major factor contributing to the rise of Postmaterialism.

'Political generations' are politically distinctive age groups delimited by their years of birth or some major political event, such as their years of entry into the electorate. Rintala (1979: p. 8) defines a political generation as "a group of human beings who have undergone the same basic historical experiences during their formative years." According to Braungart and Braungart (1989: p. 283) "a political generation is an age group in history that mobilizes to work for political change," although in principle an age group could work to maintain the status quo.

Generation and Cohort

Birth, aging, and death gradually transform societies. Unless there is a world war (Snyder, 2010; Weinberg, 2005), genocide (Rummel, 1994), pestilence (Herlihy and Cohn, 1997), or massive human emigration or immigration, these natural processes are the main forces transforming societies. Birth, aging, and death may or may not transform societies. The young may resemble the old in their political behavior and, even when young adults do differ from their elders, they may come to resemble them when they themselves age.

But although each generation perpetuates society, it may also change it. If the young could be isolated from their elders, it would be easier to create a new social order, and Plato argued that philosophers could establish a new society more quickly if they expelled all citizens above the age of 10 from the Republic (*Republic* 54a). The *Book of Numbers* relates that the generation freed from Egypt, which again and again showed its faithlessness, was condemned to have their corpses consumed in the wilderness until a new generation, reared in freedom, was ready to conquer Canaan (14:29; 14:32). Both the *Hebrew Bible* and the *New Testament* reveal great interest in generations.

Generations can apply to personal records or to the ancestors of a group. As Wilson (1992) writes, genealogies are rare in Ancient Near Eastern literature. They sometimes appear in lists of Mesopotamian and in 2d-milllennium texts dealing with the political history of the Amorites. However, Wilson notes that the *Hebrew Bible* contains 25 genealogies of various complexity. The *New Testament* provides two genealogies of Jesus (*Matthew* 1:1–17 and *Luke* 3:23–38).

Other political systems have their founding legends, and in his epic Virgil explains how Aeneus, the son of Anchises and Venus, becomes the head of the surviving Trojans and leads them to found Rome. Some legends trace the British royal family back to the Trojans. Generational conflict is a basic theme of several of Shakespeare's tragedies, such as *Hamlet*, *King Lear*, *Macbeth*, and *Coriolanus*, and generational conflict plays a major role for twentieth-century writers, most notably in Faulkner's novels, especially his *Absalom, Absalom* (1936).

Although the concept of generations has biblical, Greek, and Roman roots, the social scientific study of political generations began with Mannheim's essay, 'The Problem of Generations' (1952 [1928]). According to Mannheim (1952) the concept of generations is both biological and sociological. Generations are produced through both "the biological rhythm of human existence," but "individuals who belong to the same generation, who share the same year of birth, are endowed, to that extent, with a common location in the historical dimension of the social process" (p. 209).

Mannheim (1952) maintains that historical experiences are most likely to be shared by persons in their late adolescence and early adulthood, postulating that 'personal experience with life' usually begins at about age 17. Whether young people are actually influenced by historical events depends upon social and historical conditions. Furthermore, young adults living in some historical conditions may experience quite different events, even within the same society. Nor do all historical conditions produce conditions that influence young adults in distinctive ways. Mannheim emphasizes that distinctive generations are more likely to form in periods of rapid social change. But his basic point is that there is no guarantee that socially or politically distinctive generations will emerge. "Whether a new *generation style* emerges rhythmically at all, depends entirely on the trigger action of the social and cultural process" (p. 310).

Distinctive generations emerge in times of rapid social change. Many contemporaries viewed the French Revolution as a time of unprecedented change. As Wordsworth (*The Prelude*

[1805, Book Tenth, lines 108–109]) wrote, "Bliss was it in that dawn to be alive, But to be young was very heaven!" Although Condorcet lacked both the ability to write flowery poetry or clear prose, he saw the French Revolution as an event that might bring female equality and the emancipation of slaves (1802). The Americans who served in World War II, were often in their thirties, but Brokaw (1998) famously labels them *The Greatest Generation*.

Following Mannheim's (1952 [1928]) cautions, it is unwise to label a group of people a political generation unless one can establish that they are actually distinctive in their attitudes and behaviors. Therefore, social scientists need a term describing people born during the same years, but making no other assumptions about their distinctiveness. The term 'cohort' is used to describe a distinctive birth group. The term derives from a military unit in a Roman legion, but the term now has a much broader meaning. Scholars usually define a cohort by its years of birth, often grouping respondents in 10-year-old categories in order to save one column on the IBM punch cards that once were used to be used to store data. Scholars who want to work with these early data needed to use these arbitrary age categories. With computerized storing, survey data are usually coded according to year of birth giving analysts greater flexibility. For example, in a major cohort analysis, Converse (1976b) divides his data into 3-year categories, and throughout their study of the electorate, Miller and Shanks (1996) generally divide their cohort results into 4-year categories. In the 2012 American National Election Study (ANES) survey, the policy of providing actual year of birth was abandoned and 5-year categories were used. In 2012 face-to-face interviews were supplemented by an Internet survey. Researchers feared that if this "Big *N*" approach included the respondents' year of birth anonymity might be compromised. But by not measuring the respondents' years of birth ANES is jeopardizing over-time comparability.

The Impact of Generational Replacement

The population is gradually transformed through generational replacement. (Some scholars use the term 'cohort succession.') Replacement can play a major role in transforming the political attitudes and behavior among mass publics. For example, during the 1930s and 1940s replacement helped make the Democrats the majority party in the United States (Campbell et al., 1960), and during the postwar years replacement contributed to eroding party loyalties (see Abramson 1976, 1979, 1983, 1989b; but see Converse, 1976b). Replacement benefited the British Labour Party during the 1960s (Butler and Stokes, 1969, 1974[1969]), but it contributed to the erosion of British party loyalties between 1964 and 1987 (Abramson, 1992). Between 1969 and 1988, replacement contributed to the decline of the Labour Party in Israel (Abramson, 1989a). Replacement may also help transform political values. For example, replacement helped to increase tolerance in the United States between the mid-1950s and the mid-1970s (Abramson, 1983). Women who came of age after the enfranchisement of women by the Nineteenth Amendment played a major role in increasing women's turnout (Firebaugh and Chen, 1995).

Miller (1992) argues that the decline of turnout in American elections largely results from the entry of a post-New Deal generation into the electorate (see also Miller and Shanks, 1996). That change, Miller contends, resulted not only from their youth, but also from generational differences that contributed to low levels of participation. During the 1960s and early 1970, a series of events – the Vietnam War, the Watergate scandal, and the perception that Gerald R. Ford and Jimmy Carter were failed presidents – created a generation that withdrew from political activity. In his *Bowling Alone*, Putnam (2000) explores the impact of aging and generational succession, and argues that political disengagement largely resulted from the baby-boom generation coming of age. He writes (p. 265), "The declines in church attendance, voting, political interest, campaign activities, associational memberships, and social trust are attributable almost entirely to generational succession." Along with Campbell (Putnam and Campbell, 2010), he further documents that replacement contributed to the decline of religiosity and for growing tolerance toward homosexuality (Putnam and Campbell, 2010). For more specific information on generational change in Canada, Québec, Great Britain, Italy, and France, see the volume edited by Crête and Favre (1989).

Whenever there is a dramatic shift in public opinion, along with substantial differences between age groups, the shift probably results from a combination of changes among the existing public and generational replacement. The growing support for same-sex marriage in the United States provides an example, although further research will be necessary to understand the dynamics at work. For example, three public opinion polls (accessed through http://pollingreport.com) show increased support for same-sex marriage. For example, a Gallup poll of about 1500 adults showed that in May 2006, 42% said that same-sex couples should have the right to have legally valid marriages; by May 2013, 53% did. A CBS/*New York Times* poll of about 2000 respondents found that in March 2004, only 24% said Americans should be allowed to marry legally; in February 2012, 40% did. And in an ABC/*Washington Post* poll of about 1000 respondents conducted in November 2003, only 37% said it should be legal for gay and lesbian couples to legally marry; by March 2013, that figure had risen to 58%. Young Americans are consistently more likely to support legal same-sex marriages than their elders. In July 2003, a large CBS/*New York Times* poll of 3000 Americans showed that among those between 18 and 29 years old, 61% said that homosexual couples had the right to marry, whereas among those 65 and older, only 18% did. And in a March 2013 CBS poll of 1000 respondents, 73% of Americans between 18 and 29 thought it should be legal for same-sex couples to marry; among those 65 and older, some 35% did.

Ronald Inglehart and the Study of Generational Change

By studying values for over four decades, Inglehart has made a major contribution to the study of political generations. In 1971 he reported that younger Europeans had substantially different values than did their elders. Whereas older Europeans

tended to value material security and domestic order, younger Europeans were more likely to value political liberties. Inglehart (1971) initially labeled these priorities as 'acquisitive' and 'post-bourgeois,' he subsequently (1977) used the terms 'Materialist' and 'Postmaterialist.' Although Inglehart initially had surveyed in Belgium, Britain, France, West Germany, Italy, and the Netherlands, he speculated that age-group differences probably resulted from the formative socialization of younger and older Europeans, although he acknowledged that they might result partly from higher levels of education among the young. Even if economic conditions did not remain as rosy after 1971, it seemed highly unlikely that Europeans born after World War II would suffer from the extreme deprivation of Europeans who grew up before or during World War II or from the deprivations of the post-War years. And there were many reasons to expect them to have higher levels of education than their elders.

Many scholars viewed Inglehart's prediction as a major insight. Almond (1990) argued that 'Inglehart's work is one of the few examples of successful prediction in political science.' The authors of the five-volume *Beliefs in Government* series find an overwhelming of a shift toward a redefinition of the Left–Right continuum. Scarbrough, who analyzes Materialist/Postmaterialist orientations, concludes (1995: p. 156), "indisputably, across much of Western Europe, value orientations are shifting." General editors, Kasse and Newton (1995: p. 61), sum up the findings: "We find substantial support for the model which traces social change to value changes, and value changes into political attitudes and behaviour, especially through the process of generational replacement." And in his discussion of value change Dalton (2014: 89) concludes, "The most influential research is Ronald Inglehart's theory of value change in advanced industrial societies."

Beginning in the mid-1980s, Inglehart and I began collaborative work to demonstrate that the trend toward Postmaterialism was largely the result of generational replacement. We employed an algebraic standard procedure that I used to study attitude change in the United States (Abramson, 1983). To demonstrate how this procedure works, I examined the percentage of high school graduates among the white electorate between 1956 and 1980, assuming that no new cohorts were allowed to vote. But this procedure would not account for the tendency of older Americans to die at higher rates than the younger. Thus, I algebraically 'immortalized' the older cohorts, assuming that their distribution remained unchanged. This procedure seemed justified since people do not become less educated with age, although, as Swift's (2002[1726]) *Struldbruggs* discovered, immortality can carry a heavy price. But when other political attitudes or behaviors are imputed to a cohort that has dropped out of the sampling frame through death or institutionalization, we do not know what their attitudes would have been if they had survived. For example, my first cohort analysis of the decline of party loyalties among American whites began with seven 8-year cohorts, the youngest of which was born between 1924 and 1931. In 1952, when party identification was first measured, that birth cohort was between the ages of 21 and 28. By the 2012 election, this cohort was between the ages of 81 and 88. Moreover, given changes in ANES sampling procedures (mentioned earlier) this cohort cannot be identified. But even if we could, the U.S. Bureau of the Census data show that only 5% of the noninstitutionalized white population in 2012 was old enough to have voted in the first Eisenhower–Stevenson contest. Whites who were 75 years old and older made up 6% of the sample, but a small percentage of them would have been too young to have voted in 1952.

As Inglehart first began to survey the European public in 1970, and since he sampled Europeans as young as 16, problems of cohort attrition are not as great as those confronting social scientists who rely upon the ANES surveys. However, the method I developed to analyze the effects of generational replacement had begun to affect the ongoing data that Inglehart has collected as President of the World Values Survey.

Before discussing these problems, however, it is fair to note that some scholars disagree with Inglehart's basic thesis and question his basic measure, which is based upon respondents choosing their top two national goals out of four choices. (A less frequently used measure asks them to choose their top six choices out of 12 goals.) Inglehart's data are readily available. We describe them in our book and the SPSS procedures for constructing them are available in Inglehart (1997: p. 389).

Inglehart has always archived his data and made them readily available to the social science community. In doing so, he conforms to the fundamental tenets of scientific research, but he also exposes himself to criticism, although in some cases critics have generated their own data. In a recent paper (Abramson, 2011), I reviewed 54 critiques of Inglehart's work. Many of these critiques were helpful, although many were based upon a misunderstanding of Inglehart's basic thesis. In this essay, I will focus only on the way these arguments shed light on the study of political generations.

By the early 1980s it was clear that Postmaterialism was not following a simple trend upward (Inglehart, 1981), but that the views of the European public also responded to short-term forces, or in the language of cohort analyses, the public responded to period effects.

Our book (Abramson and Inglehart, 1995) presents our most extensive treatment. We examine evidence from 40 societies, seven wealthy societies in the European Community (now the European Union), four less wealthy societies in the European Community, five West European societies not in the European Community, two North American countries, four Latin American countries, five East European societies, six former Soviet societies, one Middle Eastern society (Turkey), two Asian countries (China and India), two East Asian countries (Japan and South Korea), and two African countries (Nigeria and South Africa). Although we present the relationship between Postmaterialism and age and the mean rate of economic growth for the previous three decades, our cohort analyses focus mainly upon eight European democracies: Belgium, Britain, Denmark, France, Ireland, Italy, the Netherlands, and West Germany. At times we use a weighted national sample of Belgium, Britain, France, Italy, the Netherlands, and Germany (the only countries in the series sampled in 1970).

Our main conclusion is based upon our algebraic standardization procedures. The score on the Postmaterialism index (PDI) is computed by subtracting the percentage of Postmaterialists

from the percentage of Materialists. The measure ranges from a low of −100 to a high of 100, with 0 indicating an equal number of Materialists and Postmaterialists.

We present the PDI scores among cohorts for all eight countries and then discuss the possible effects of replacement, drawing upon Abramson (1983). We then estimate the effects of replacement for all eight societies as well as for a combined weighted sample of the six countries sampled from 1970 to 1992. We then estimate the effect of replacement for all eight societies as well as for the combined European sample. In 1970, only 11% of the West European publics were Postmaterialists, whereas 40% were Materialists, a PDI of −29. In 1992, 16% were Postmaterialists and only 26% are Materialists, a PDI of −10. During these two decades there had been a 19-point shift toward Postmaterialism. In the 1992 samples, 42% of the respondents were too young to have been sampled in 1970 and some had not yet been born.

According to our estimates, if there had been no generational replacement Postmaterialism would have increased only about two points between 1970 and 1992. In other words, only about one-tenth of the shift toward Postmaterialism would have occurred without replacement. These results also correspond to multivariate analyses (Abramson et al., 1997; Abramson and Inglehart, 1996; Inglehart and Abramson, 1994) showing that when inflation and unemployment are controlled, PDI scores increase at about one point per year, very close to the rate Inglehart's model predicts.

Since our book was published there have been several critiques, but we have already responded to most of them. The two most important were published in the *American Political Science Review* (Clarke et al., 1999; Davis and Davenport, 1999) and we have replied to both of them (Inglehart and Abramson, 1999).

More Recent Changes in Postmaterialism

Inglehart's recent study with Welzel (2005) briefly discussed the impact of generational replacement on the trend toward Postmaterialism, but they do not present the algebraic standardization procedures employed by Abramson and Inglehart. Perhaps this is just as well because ongoing generational replacement has further reduced the cohorts that can be tracked over time. However, Inglehart supplied me with a figure based upon a combined weighted sample of the six European countries that can be tracked between 1970 and 2006. This figure is available in Abramson (2011). It shows the results for 10 cohorts. The dark line shows the overall level of Postmaterialism for the entire sample.

The figure reveals five basic findings. First, it demonstrates that replacement occurs. Three cohorts in the first samples, conducted in 1970, cannot be tracked over the 36-year period. Granted, there are some centenarians in these societies, but not enough to be sampled. By 2006, even the cohort born between 1916 and 1925 was between the ages of 81 and 90. Many were institutionalized and because of differential death rates the cohort would be very heavily female. During these three and a half decades, three full cohorts born between 1986 and 1991 entered the sample. In fact, 45% of these younger Europeans had not been born when these surveys began. Although Firebaugh (1989) makes excellent suggestions for dealing with the problem of 'non-surviving' cohorts, all are based upon problematic inferences.

Among the respondents sampled in 2006, only 40% were old enough to have been sampled in 1970. Among the cohorts that did survive, the cohort composition was substantially different than it was in 1970. As the figure shows, the largest cohort differences are between Europeans born between 1936 and 1945 and those born between 1946 and 1955; that is, between cohorts that suffered through World War II and the postwar deprivation and those who came of age when Western Europe began to prosper. In 1970 among cohorts born before 1956 (and thus old enough to have been sampled in 1970), 86% were born before 1946. But by 2006, only 57% were.

Second, the figure demonstrates consistent cohort differences across time among the cohorts born before 1956. Among these cohorts the younger cohorts have higher levels of Postmaterialism than their elders and there are few breaks in monotonicity. The 1956–65 cohort is significantly more Postmaterialist than the 1956–65 cohort, a finding consistent with Inglehart's thesis since the 1956–65 cohort grew up under far more secure conditions.

Third, there are few differences among the four youngest cohorts. Despite moderate economic growth, rising levels of economic inequality have led to little or no increases in real income for most Europeans. Economic security has been reduced by cutbacks in the welfare state and high levels of unemployment, especially among the young.

Fourth, the figure demonstrates no support for a life-cycle explanation for age-group differences. If a life-cycle interpretation for the Postmaterialism of the younger cohorts were valid, we would expect, other things being equal, these cohorts to become more Materialist as they age. They do not. Moreover, the older cohorts gravitate toward Postmaterialism. One can always assume that these cohorts would have become more Materialist with age, but that period effects prevented such a move. But I know of no plausible explanation supporting such an interpretation. Of course, we know that the trend toward Postmaterialism does not result from replacement alone, for there are also short-term fluctuations caused by changing inflation rates (Abramson and Inglehart, 1994; 1995: pp. 25–39).

Finally, the figure strongly suggests that the overall shift toward Postmaterialism is at least partly driven by replacement. Because fully 60% of the 1970 sample can no longer be sampled in 2006, using the algebraic standardization procedure Inglehart and I applied from 1970 through 1992 (Abramson and Inglehart, 1995) is problematic. All the same, even an informal examination of the figure demonstrates the importance of replacement. Recall that in 1970 the PDI for this six-nation sample was −29. In 2006, 19% were Materialists, while 21% were Postmaterialists, a PDI of 2. Overall, there was a 31-point shift toward Postmaterialism. But while this shift is reflected in the solid trend line representing the overall level of Postmaterialism, none of the cohort gains is close to this magnitude. Comparing the overall trend line with the trend line for the cohorts makes this clear. The overall rise in the PDI scores occurs at least in part because over this third of a century the older, more Materialist

cohorts were gradually being replaced by younger, more Postmaterialist cohorts.

Some of Inglehart's critics would argue that his measure is meaningless (for a summary, see Abramson, 2011). If so, so is the trend. But there is overwhelming evidence that Postmaterialists differ in important ways from the Materialists they are replacing. Dalton (2014) provides an excellent summary. In the workplace, Postmaterialists are more likely to want a flexible work environment. In social relations, the rise of Postmaterialism leads to declining deference toward authority. Postmaterialism has led to changing moral standards. The rise in Posmaterialism has led to a shift of the issue agenda, with Postmaterialists being more concerned about the environment, gender equality, and multiculturalism. Finally, Postmaterialism affects political participation because it leads to participation in decisions affecting one's life. Although the growth of Postmaterialism may not lead to genuine revolutions, it may have the more benign effect of moderately transforming Western societies.

See also: Party Identification and Nonpartisanship; Political Tolerance; Politics and Participation; Politics and Socialization.

Bibliography

Abramson, P.R., June 1976. Generational change and the decline of party identification in America:1952–1974. American Political Science Review 70, 469–478.

Abramson, P.R., 1979. Developing party identification: a further examination of life-cycle, generational, and period effects. American Journal of Political Science 23 (23), 78–96.

Abramson, P.R., 1983. Political Attitudes in America: Formation and Change. Freeman, San Francisco.

Abramson, P.R., Inglehart, R., 1996. Formative security, education, and post-materialism: a response to Davis. Public Opinion Quarterly 60 (3), 450–455.

Abramson, P.R., August 1989a. Generational replacement, ethnic change, and partisan support in Israel. Journal of Politics 55, 545–574.

Abramson, P.R., 1989b. Generations and political change in the United States. In: Braungart, R.G., Braungart, M.M. (Eds.), Research in Political Sociology. JAI Press, Greenwich, CT, pp. 235–280.

Abramson, P.R., July 1992. Of time and partisan stability in Britain. British Journal of Political Science 22, 381–395.

Abramson, P.R., 2011. Critiques and Counter-critiques of the Postmaterialism Thesis: Thirty-four Years of Debate. Paper Available Online at the Center for the Study of Democracy. University of California at Irvine. http://www.escholarship.org/uc/item/3f72v9q4.

Abramson, P.R., Ellis, S., Inglehart, R., March 1997. Research in context: measuring value change. Political Behavior 19, 41–59.

Abramson, P., Inglehart, R., August 1994. Education, security, and postmaterialism: a comment on Duch and Taylor's, 'postmaterialism and the economic condition. American Journal of Political Science 38, 797–814.

Abramson, P.R., Inglehart, R., 1995. Value Change in Global Perspective. University of Michigan Press, Ann Arbor, MI.

Almond, G.A., 1990. Statement on the back cover of R Inglehart, Culture Shift in Advanced Industrial Societies. Princeton University Press, Princeton, NJ.

The Aeneid of Virgil (R. Fagles, Trans.), 2006. Penguin, New York.

Braungart, R.G., Braungart, M.M., 1989. Political generations. Research in Political Sociology 4, 281–319.

Brokaw, T., 1998. The Greatest Generation. Random House, New York.

Butler, D., Stokes, D., 1969. Political Change in Britain: Forces Shaping Electoral Choice. St. Martin's Press, New York.

Butler, D., Stokes, D., 1974[1969]. Political Change in Britain: The Evolution of Electoral Choice. St. Martin's Press, New York.

Campbell, A., Converse, P.E., Miller, W.E., Stokes, D.E., 1960. The American Voter. Wiley, New York.

Clarke, H.D., Kornberg, A., McIntyre, C., Bauer-Kaase, P., Kaase, M., September 1999. The effects of economic priorities on the measurement of value change: new experimental evidence. American Political Science Review 93, 637–647.

Condorcet, J.A.N. de, 1802[2012]. Outlines of an Historical View of the Progress of the Human Mind. Forgotten Books, Baltimore.

Converse, P.E., July 1976a. Of time and political stability. Comparative Political Studies 2, 139–171.

Converse, P.E., 1976b. The Dynamics of Party Support: Cohort-analyzing Party Identification. Sage, Beverly Hills, CA.

Crête, J., Favre, P. (Eds.), 1989. Générations et politique. Economica, Paris.

Dalton, R., 2014. Citizen Politics: Public Opinion and Political Parties in Advanced Industrial Democracies, sixth ed. Sage, Los Angeles.

Davis, D.W., Davenport, C., September 1999. Accessing the validity of the post-materialism index. American Political Science Review 93, 649–664.

Faulkner, W., 1936. Absalom, Absalom. Random House, New York.

Faulkner, W., 1990. Novels, 1936–1940. The Library of America, New York.

Firebaugh, G., 1989. Methods for estimating cohort replacement effects. In: Clogg, C.C. (Ed.), Sociological Methodology. American Sociological Association, Washington, DC, pp. 243–262.

Firebaugh, G., Chen, K., January 1995. Vote turnout of nineteenth amendment women: the enduring effect of disfranchisement. American Journal of Sociology 100, 972–996.

Herlihy, D., Cohn, S.K., 1997. The Black Death and the Transformation of the West. Harvard University Press, Cambridge, MA.

Inglehart, R., December 1971. The silent revolution in Europe: intergenerational change in post-industrial societies. American Political Science Review 65 (04), 991–1017.

Inglehart, R., 1977. The Silent Revolution: Changing Values and Publics Among Western Publics. Princeton University Press, Princeton, NJ.

Inglehart, R., December 1981. Post-materialism in an environment of insecurity. American Political Science Review 75, 880–900.

Inglehart, R., 1997. Modernization and Postmodernization: Cultural, Economic, and Political Change in 43 Societies. Princeton University Press, Princeton, NJ.

Inglehart, R., Abramson, P.R., June 1994. Economic security and value change. American Political Science Review 88, 336–344.

Inglehart, R., Abramson, P.R., September 1999. Measuring postmaterialism. American Political Science Review 93, 665–677.

Inglehart, R., Welzel, C., 2005. Modernization, Cultural Change, and Democracy: The Human Development Sequence. Cambridge University Press, New York.

Kasse, M., Newton, K., 1995. Beliefs in Government, Vol. 5: Beliefs in Government. Oxford University Press, New York.

Mannheim, K., 1952 [1928]. The problem of generations. In: Kecskemeti, P. (Ed.), Essays on the Sociology of Knowledge. Oxford University Press, New York, pp. 276–322.

Miller, W.E., 1992. The puzzle transformed: explaining declining turnout. Political Behavior 14 (1), 1–43.

Miller, W.E., Shanks, J.M., 1996. The New American Voter. Harvard University Press, Cambridge, MA.

Putnam, R.D., 2000. Bowling Alone: The Collapse and Revival of Community in America. Simon and Schuster, New York.

Putnam, R.D., Campbell, D.E., 2010. American Grace: How Religion Divides and Unites Us. Simon and Schuster, New York.

Rintala, M., 1979. The Constitution of Silence: Essays on Generational Themes. Greenwood, Westport, CT.

Rummel, R.J., 1994. Death by Government. Transaction Publishers, New Brunswick, NJ.

The Republic of Plato Translation with Notes, an Interpretive Essay, and a New Introduction by Alan Bloom, 1968. Basic Books, New York.

Scarbrough, E., 1995. Materialist-postmaterialist value orientations. In: Van Deth, J.W., Scarbrough, E. (Eds.), Beliefs in Government: The Impact of Values. Oxford University Press, New York.

Snyder, T., 2010. Bloodlands: Europe between Hitler and Stalin. Basic Books, New York.

Swift, J., 2002[1726]. In: Rivero, A.J. (Ed.), Gulliver's Travels. Norton Critical Edition, New York.

Weinberg, G., 2005. The World at War: A Global History of World War II, new ed. Cambridge University Press, Cambridge.

Wilson, R.R., 1992. Genealogy, genealogies. In: Freedman, D.N. (Ed.), The Yale Anchor Bible Dictionary, vol. 2. Doubleday, now published by Yale University Press, pp. 929–932.

Wordsworth, W., 1979 [1799][1805][1850]. In: Wordsworth, J., Abrams, M.H., Gill, S. (Eds.), The Prelude. Norton Critical Editions, New York.

Generations, Sociology of

Shmuel N Eisenstadt, Van Leer Jerusalem Institute, Jerusalem, Israel
Bryan S Turner, The City University of New York, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by S.N. Eisenstadt, volume 9, pp. 6055–6061, © 2001, Elsevier Ltd.

Abstract

Generations represent an important element in the social structure, alongside social class, race, and gender, but they have often been neglected in contemporary sociology. However, from a historical and comparative perspective, generations have played an important role in social change in terms of youth revolts against the established order. In the twentieth century, generational conflict over values and resources was important in the aftermath of the First World War, the Great Depression, and the youth movements of the 1960s. Indeed, the Baby Boomer Generation is often regarded as a crucial aspect of the transition of the West to a consumer society.

Generations, Cohorts, and Consciousness

Generations, alongside social class, ethnicity, and gender, are basic building blocks of the social structure, but they have often been neglected in mainstream sociology. The definition of generations and recognition of their sociological importance are associated with the work of Karl Mannheim (1952) and Maurice Halbwachs (1992) who established the basic components of the sociological perspective on generational change in terms of the importance of a shared culture and rituals in constituting the collective memory of a generation. Halbwach's perspective on collective memory was important in following the tradition of Emile Durkheim's notion of collective representations. Mannheim's approach was shaped by his own experiences of the political instability of Europe, the rise of fascism, and the alienation of youth. The sociology of generations is consequently a well-established tradition in modern sociology. However, there are a number of analytical problems with the concept that should be distinguished from the related ideas of cohort and age grades. Cohorts are simply aggregates of individuals born at the same time and passing through time, whereas generations, in order to have an impact on social change, typically possess a generational consciousness. In this respect, there is a parallel between class analysis with its emphasis on class consciousness and class conflict and generational consciousness and generational conflict (White, 2013). Generations and generational differences can play an important role in both political and social movements. The nationalist/revolutionary movement of Central and Eastern Europe in the late nineteenth to early twentieth century and the youth and student rebellion of the 1960s are prominent examples of the influence of generational consciousness, couched in the symbolism of youth. In contemporary sociology, interest has focused on consumption patterns and generational change, and on youth cultures and generational differences, for example, around 'The Sixties' and the 'Baby Boomers.' Finally, there is the notion that, while in the past generations were national phenomena of relevance to nation-states, with globalization one can consider the possibility of global generations with an emerging cosmopolitan awareness (Edmunds and Turner, 2005).

Generations constitute – just as the closely related categories of age, youth, adulthood, and old age – seemingly universal categories in all human societies signifying the different stages of human life; the same is true of tensions between generations, as well as of concepts of the age group or set – in particular the youth group – as a special social category with its own distinct self-consciousness. But it is only in very specific civilizational and historical circumstances that these features come together. It is only in special circumstances that historical change becomes articulated in terms of generational consciousness in general, and that of youth in particular.

It was Karl Mannheim – as well as J. Marias and J. Ortega y Gasset – who emphasized the apparent universality of generational consciousness or at least posed it as a major problem of research, but this tendency, at least as reported by them, was indeed very closely connected to rather specific civilizational traditions – the European ones – and to specific historical circumstances in the development of this civilization. Accordingly, a central problem in the analysis of generations is under what conditions does historical change become articulated in terms of generational consciousness in general and that of youth in particular?

The consciousness of a generation is often defined by traumatic historical events that shape not only the consciousness and identity of a generation but also its life chances and resources. We can examine several well-known examples. The First World War had an immediate and devastating impact on the mental and physical health of the young men who were involved, but there were also long-term implications for their personal development and opportunities. The trenches were also seen as defining the manhood of an entire generation of young men – the so-called Generation of 1914 – and the following generation that came to manhood in the interwar years had a sense of lack (Wohl, 1979). However, perhaps the important illustration of this argument is the effect of the Great Depression on the generation that came to maturity in the economic crisis that severely limited their opportunities for education and employment over the course of their mature years. This generation was famously described in Glen Elder's

Children of the Depression (1974). In more recent history, Mao Tse Tung's cultural revolution that involved sending thousands of young people to work in the countryside created a generation that was defined by its lost opportunities. The Baby Boomer generation by contrast was a postwar generation that benefitted considerably by the expanding economies of the postwar period with the welfare state, Keynesian economic politics, and by the consumer boom of the 1970s (Binstock, 2010). Their political consciousness was also defined by wars primarily by opposition to the Vietnam War and by other historical events – the Kennedy Assassination, the killing of Martin Luther King, the Beatles, and the landing of men on the moon (Eyerman and Turner, 1998).

We can conclude that where exactly people are born and the historical events that shape their life chances have importance for defining the resources, culture, and consciousness of different generations passing through time. Sociologists have proposed that the distinction between a class for itself and a class in itself could by analogy be applied to generations (Edmunds and Turner, 2002). Thus, active generations are large cohorts that have strategic advantages that are made possible by the specific circumstances that open up opportunities to exercise political and cultural influence. Passive generations occur in historical circumstances that offer limited opportunities and life chances. This distinction is implicit in the contrast that is often made between the Baby Boomers and Generation X, where the latter is seen to be apolitical and passive.

Youth as a Liminal, Transitory Phase; Kernels of Generational Consciousness

The problems of age definition and the linkage of personal time and transition with cosmic time become very often accentuated in that age span usually designated as youth – which constitute in most societies the most liminal of age definitions – the transition from childhood to full adult status, to full membership in the society. In this period, the individual is no longer a child (especially from the physical and sexual point of view) and is ready to undertake many attributes of an adult and to fulfill adult roles. But adolescent individual is not yet fully acknowledged as an adult, a full member of the society. Rather, he or she is being, as it were, 'prepared,' or is preparing for such adulthood.

This image of youth – the cultural definition of youth – contains all the crucial elements of any definition of age, usually in a specially articulated way, and often in confrontation with other age categories. Moreover, in this phase the problem of the linkage of the personal temporal transition with cosmic or societal time becomes extremely acute.

The fullest, the most articulate, and definitive expression of these archetypal elements of youth and of their liminal potentials is best exemplified in the ritual dramatization of the transition from adolescence to adulthood, such as the various *rites de passage* and ceremonies of initiation in primitive tribes and in ancient civilizations. In more diluted form, these elements may be found in the various spontaneous initiation ceremonies of the fraternities and youth groups in modern societies (Bettelheim, 1954; Erikson, 1963, pp. 54–93). Hence, the youth period and its symbolism, and the liminal situations in which it becomes expressed, may become the focus both of continuity of the social order as well as of protest against it, and of its change.

The very emphasis on the transitory nature of this stage and on its essentially preparatory character, however, often create a somewhat paradoxical situation – namely, an image of youth as the purest manifestation and repository of ultimate cultural and societal values of all the major human virtues and primordial qualities. Such an image may be reinforced by the identification of physical vigor in youth with a more general flowering of the cosmos or the society. It is this situation that contains the kernel of the development of strong generational consciousness especially around the image of youth, with strong potentialities of protest.

The social organization of any age and especially of the period of youth is constructed in all societies according to two major criteria. One is the extent to which age in general and youth in particular form criteria for the allocation of roles in a society, whether in politics or in economic or cultural activity – aside from the family, of course, where they always serve as defining principles. The second is the extent to which within society there develop specific age groups, specific corporate organizations, composed of members of the same 'age,' such as youth movements or old men's clubs.

The Social and Cultural Contexts of Different Patterns of Age and Generational Symbolism and Organization

The importance of age as a criterion for allocating roles in a society is closely related to several major aspects of social organization and cultural orientation. The first aspect is the relative complexity of the division of labor. In general, the less complex an organization, the more influential age will be as a criterion for allocating roles. Therefore, in tribal or preliterary traditional societies, age and seniority constitute basic criteria for allocating social, economic, and political roles.

The second aspect is that the major value orientations and symbols of a society, especially the extent to which they emphasize certain general qualities, or types of activity (such as physical vigor, the maintenance of cultural tradition, and the achievement and maintenance of supernatural prowess) become expressed and symbolized in specific age spans. For instance, in those societies in which military values and orientations prevail, young adulthood is emphasized as a very important – sometimes even as the most important – age, while those communities in which sedentary activities prevail emphasize older age. Similarly, within some traditional societies, where older people may serve as important a receptacle of collective memory, old age may be emphasized if it is seen as the most appropriate one for expressing major cultural values and symbols – for instance, the very upholding, symbolization, and perpetuation of a given cultural tradition.

The social and cultural conditions that determine the extent to which specific age groups and youth groups develop differ from the conditions that determine the extent to which age serves as a criterion for the allocation of roles. At the same time, the two kinds of conditions may be, in some situations or

societies, closely related. Age groups in general and youth groups in particular tend to arise in those societies in which the family or kinship unit cannot ensure (or occasionally even impede) the attainment of full social and political status on the part of its members. These conditions appear especially, although not uniquely, in societies in which family or kinship groups do not constitute the basic units of the social division of labor. Membership in the total society (citizenship) is not defined in terms of belonging to any such family, kinship group, or estate, nor is it mediated by such a group.

The emergence of distinct age and generational organization and consciousness, as Meyer Fortes has shown in one of his last papers, lies in the political realm, in the search and definition of citizenship beyond the family sphere and not through the family, while generational differences are mostly elaborated in the structure of the family and kinship unit.

Historical Change, Revolutionary Orientations, Generational Consciousness in Modern Youth Movements

The development in modern societies of such a dissociation between generations, and the concomitant possibility of the development of a special type of a distinct generational consciousness, is due not only to the complex division of labor that characterizes these societies. They are also closely connected to the relatively quick tempo of historical change, and above all to the consciousness of such change. Such consciousness is often connected with changes rooted in the revolutionary orientations that constituted important components of the construction of modern civilization, cultural values, and orientations. These revolutionary orientations were most fully articulated by the major social movements that have developed as part and parcel of modern life, of the constructions of modern social and political order, and of struggle around such construction. The possibility of the development of distinct generational consciousness has been enhanced in modern societies by the occurrence of rapid historical changes or ruptures, the consciousness of which is shared by members of the same age group or age cohort.

While historical change is particularly rapid and the consciousness of such change very widespread in modern societies, it does not always give rise to fully articulated generational consciousness – although to some degree such consciousness is probably rather common in modern societies and is not necessarily confined to the symbolism and ideology of youth. It is only in rather special historical circumstances and situations – like wars, changes of regime, and the like, especially of events that entail strong ruptures – that such generational consciousness has become articulated in modern societies in terms of youth symbolism.

Among such situations in modern societies, two are probably of special importance. One was developed at the end of the nineteenth and the beginning of the twentieth century, and found its fullest expression in the free German youth movements, although it was also present in the less spectacular youth movements of Central Europe and to some extent in the youth culture of various more flexible youth groups that developed in most modern societies. In all these movements, very strong attempts have been made to overcome the dislocation between personal transition and societal and cultural time. It was in them that dreams of a new life, a new society, freedom and spontaneity, a new humanity, and aspirations to social and cultural change found utterance. It was in these youth movements that the forging of youth's new social identity became closely connected with the development of new symbols of collective identity or new social–cultural symbols and meanings, which thus become foci of protest aimed at the reconstruction of society and culture.

While some of these orientations could be found in many of the youth groups that developed in most modern societies – be they in England or the United States – they developed most fully in situations where these groups became closely interconnected with major social movements oriented at the reconstruction of the centers and symbols of their respective societies. This was the case above all in the major nationalistic and revolutionary movements in Central and Eastern Europe, in Italy, Germany, Hungary, and Russia, or in the Zionist movement, where these orientations and student movements played a crucial role.

One of the major characteristics of many, probably most, of these movements was that in their ideology the symbolism of youth was very closely connected with that of the construction of new modern, especially national, collectivities that were being constructed at the period, and among the aims of these youth movements the reconstruction of such collectivities – national, social, or cultural – was central. These movements aimed at changing many aspects of the social and cultural life of their respective societies, to reconstruct the centers and premises of their respective societies in terms of some of the major revolutionary orientations of modernity. They depicted the present in a rather shabby light; they censured it for its materialism, exploitation, restrictive views, and lack of opportunity for self-fulfillment and creativity. At the same time they held out hope for the future – seemingly the not very far-off future – when both self-fulfillment and collective fulfillment could be achieved and the materialistic civilization of the adult world shaken off. They tried to appeal to youth to forge its own self-identity in terms of these new collective symbols, and this is why they have been so attractive to youth, for whom they provided a set of symbols, hopes, and aims to which to direct its activities.

The influence of such movements and orientations was very great, not only in political and educational but also in many cultural arenas. In a way, education itself has tended to become a social movement. Many schools and universities, and many teachers, have been among the most important bearers of collective values. The expansion of education itself is often seen as a means by which a new epoch might be ushered in.

Youth and Student Rebellion of the 1960s

The second historical situation in modern societies in which generational consciousness, couched in the symbolism of youth, became very prominent in the new social movements burgeoned in the late 1960s and early 1970s, and became most fully manifest in the various student rebellions and movements of the time.

The background to these new developments was a combination of the growing institutionalization of many of the collective values and symbols of the social, national, and youth movements discussed above, the rapid expansion of consumerism through easy credit and the continuous expansion of education and changes in occupational structure. These developments had several important repercussions on the organization of the life spans of youth.

The area of social life that the specific youth or student culture encompasses has tended to expand continuously. First, it has extended over a long period of life, reaching, through the impact of the extension of higher education, to what before was seen as early adulthood. Second, it tends more and more to include areas of work, leisure time activity, and many interpersonal relations. Third, the potential and actual autonomy of these groups, and the possibility their members have for direct access to work, marriage, and family life, to political rights, and to consumption, have greatly increased, while their dependence on adults has greatly decreased.

Paradoxically enough, the growing direct access of young people to various areas of life has given rise to a growing insecurity of status and self-identity and to a growing ambiguity of adult roles. This insecurity and ambiguity has been enhanced, first, by the prolongation of the time span between biological and social maturity and by the extension of the number of years spent in 'preparatory' (educational) institutions, and, second, by the growing dissociation between the values of these institutions and the future – especially in terms of occupational and parental roles – of those participating in them. A third factor is that for a long period many 'young' people may as yet have no clear occupational roles or responsibilities and may be dependent on their parents or public institutions for their economic needs, while at the same time they constitute an important economic force as consumers and certainly exercise political rights. In turn, this situation may become intensified or aggravated by the growing demographic preponderance of the 'young' in the whole population and by the increasing possibilities of ecological mobility.

All these developments gave rise to new types of generational and youth consciousness – a new combination between the symbolism of youth and collective symbolism. The older classical movements of protest and youth movements tended to assume that the framework and centers of the nation-state constituted the major cultural and social reference points of personal identity and of the charismatic orientations to some sociocultural orders, and that the major task before them was to facilitate the access of broader strata to these centers and full participation in them and possible changes in them. In contrast, the new movements of protest have been characterized by their skepticism toward these new modern centers and by lack of commitments to them – all of which denoted important aspects of what has been called by Weber the demystification of the world.

The foci of protest promulgated by these movements shifted from demands for greater participation in national political centers or from attempts to influence their socioeconomic politics to new directions. The most important of these directions were first, attempts to 'disrobe' these centers of their charismatic legitimacy and perhaps of any legitimacy at all; second, continuous searches for new loci of meaningful participation beyond these centers and the concomitant attempts to create new centers that would be independent of the existing center; third, attempts to formulate the patterns of participation in their centers not so much in sociopolitical or economic terms but more through symbols of primordial or of direct social participation.

These developments have often been connected with the breakdown of consciousness of historical continuity and of the conception of the future in terms of such continuity. The image of the societal future has become flattened and has been deprived of its allure. No sharp ideological discontinuity has been conceived between present and future.

Through these developments, youth has been, as it were, robbed of the full experience of the dramatic transition from adolescence to adulthood and of the dramatization of the difference between present and future. Their own changing personal future has become dissociated from any changes in their societies or in cultural activities and values. As a result of all these processes, the possibility of linking personal transition to societal and cosmic time – which was so strongly emphasized in the youth movements and observable also to some extent even in the earlier, looser youth culture – has become greatly weakened.

One of the most important results of these developments has been the evolution of a new configuration of youth symbolism and generational consciousness – the major characteristic of which was the creation of strong antinomian intellectual tendencies as a central component of such symbolism and consciousness.

It is not just the contents of these antinomian tendencies that are new, but the convergence of these contents with the spread and location of these movements. Of special importance in this context is the fact that this type of protest was not confined only to small, closed intellectual groups but involved widespread circles of novices and aspirants to intellectual status. Given the expansion of the modern educational system and the parallel effects of the spread of mass communication media, a very large part of the educated public have become aspirant to such status, and impinge on the centers of intellectual creativity and cultural transmission and become integral, even if transient, parts thereof.

This rather unique combination of intellectual antinomianism with youth rebellion gave rise to a new type of youth ideology that was connected to the reversal of the hitherto existing relation between the definition of different age spans and the possibilities of social and cultural creativity. Unlike in the past – even the not-so-distant past – youth became more and more seen not only as a preparation for independent and creative participation in social and cultural life, but as the very embodiment of permissive, often unstructured creativity – to be faced later on with the constants of a relatively highly organized, constrictive, meritocratic, and bureaucratic environment. Thus, youth in the late twentieth century became, even more than before, the transmitters and symbols of a continuous antinomian confrontation between the pristine values it embodies and a demystified social reality, and the impact of this new type of generational consciousness has been of crucial importance in the crystallization of what has been designated as the postmodern trend in contemporary societies. While the intensity and the concrete manifestation of this consciousness will

naturally change from one situation to another, the consciousness itself with the various antinomian tendencies analyzed above seems to have become more and more prevalent.

Conclusion: Generations, Politics, and Identity

In the contemporary world, generations continue to play a major role in shaping both domestic and international politics as seen with the so-called Arab Spring commencing in Tunisia in 2011 and then evolving into the 25 January Revolution in Egypt. Similar protest movements developed in Yemen, Libya, and Turkey. What these movements have in common is the presence of a large cohort of young people (typically young men) who have been relatively well educated but who cannot find adequate or rewarding employment in their own societies. Their frustrations and unrewarded expectations spilled over into collective protests against governments and their geriatric political leaders who were seen to be corrupt and unresponsive to their economic and social plight. The availability of online communication through Facebook and Twitter helped to forge global generational identities and consciousness (Alexander, 2011). There were some parallels between these youth rebellions in North Africa and the Middle East and the Occupy Wall Street protests of 2011 in New York and other major cities in the United States (Gitlin, 2012). While these political protests were contained, they indicate the ongoing relevance of generations to modern politics.

See also: Age Stratification; Age Structure; Age, Sociology of; Age: Anthropological Aspects; Durkheim, Emile (1858–1917); Generation in Anthropology; Generations in History; Generations: Political; Life Course in History; Life Course: Sociological Aspects; Mannheim, Karl (1893–1947); Social Structure; Youth Culture, Anthropology of; Youth Culture, Sociology of; Youth Movements.

Bibliography

Alexander, J.C., 2011. Performative Revolution in Egypt. Bloomsbury Academic, London.

Aron, R., 1969. The Elusive Revolution: Anatomy of a Student Revolt. Pall Mall Press, London.

Bauer, R.A., Inkeles, A., Kulckhohn, C., 1956. Generational differences. In: Bauer, R.A., et al. (Eds.), How the Soviet System Works: Cultural, Psychological and Social Themes. Harvard University Press, Cambridge, MA, pp. 190–198.

Bettelheim, B., 1954. Symbolic Wounds, Puberty Rites and the Envious Circle. Free Press, Glencoe, IL.

Bettelheim, B., 1963. The problem of generations. In: Erikson, E.H. (Ed.), Youth: Change and Challenge. Basic Books, New York, pp. 54–93.

Binstock, R., 2010. From compassionate ageism to intergenerational conflict. The Gerontologist 50 (5), 574–585.

Bourges, H., 1968. The French Student Revolt: The Leaders Speak. Hill and Wang, New York.

Edmunds, J., Turner, B.S., 2002. Generations, Culture and Society. Open University Press, Buckingham.

Edmunds, J., Turner, B.S., 2005. Global generations: social change in the twentieth century. The British Journal of Sociology 56 (4), 559–577.

Eisenstadt, S.N., 1956. From Generation to Generation: Age Groups and Social Structure. Free Press, Glencoe, IL.

Eisenstadt, S.N., 1963. Archetypal patterns of youth. In: Erikson, E.H. (Ed.), Youth: Change and Challenge. Basic Books, New York, pp. 24–43.

Elder Jr., G.H., 1974. Children of the Great Depression. Social Change in Life Experience. Chicago University Press, Chicago.

Erikson, E.H. (Ed.), 1963. Youth: Change and Challenge. Basic Books, New York.

Eyerman, R., Turner, B.S., 1998. Outline of a theory of generations. European Journal of Social Theory 1 (1), 91–106.

Fortes, M., 1984. Age, generation and social structure. In: Kertzer, D.I., Keith, J. (Eds.), Age and Anthropological Theory. Cornell University Press, Ithaca, NY, pp. 99–123.

Gitlin, T., 2012. Occupy Nation. The Roots, the Spirit, and the Promise of Occupy Wall Street. HarperCollins, New York.

Glazer, N., 1970. Remembering the Answers: Essays on the American Student Revolt. Basic Books, New York.

Gusfield, J., 1957. The problem of generations in an organizational structure. Social Forces 35, 323–330.

Halbwach, M., 1992. On Collective Memory. University of Chicago Press, Chicago.

Heberle, R., 1951. Social Movements: An Introduction to Political Sociology. Appleton Century-Crofts, New York (see especially 'The Problem of Political Generations', pp. 118–127).

Lipset, S.M., Altbach, P.G. (Eds.), 1969. Students in Revolt. Houghton Miffin, Boston.

Lipset, S.M., Wellin, S.S. (Eds.), 1965. The Berkeley Student Revolt. Facts and Interpretations. Doubleday, Garden City, NY.

Mannheim, K., 1952. The problem of generations. In: Kecskemeti, P. (Ed.), Essays on the Sociology of Knowledge. Oxford University Press, New York, pp. 276–320.

Marias, J., 1949. El Metodo Historico de las Generaciones. Revista de Occidente, Madrid, Spain.

Ortega y Gasset, J., 1933. The Modern Theme. Norton, New York.

Pinder, W., 1926. Kunstgeschichte nach Generationen. Pfeiffer, Leipzig, Germany.

Preston, T.R., 1927. My Generation of Politics, and Politicians. Rose, Toronto.

Rintala, M., 1963. A generation in politics: a definition. Review of Politics 25, 509–522.

White, J., 2013. Thinking generations. The British Journal of Sociology 64 (2), 216–247.

Wohl, R., 1979. The Generation of 1914. Harvard University Press, Cambridge, MA.

Zeitlin, M., 1966. Political generations in the Cuban working class. American Journal of Sociology 71, 493–508.

Generative Grammar

Manfred Bierwisch, Humboldt Universität Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Generative grammar is the most influential linguistic theory of the last half century. Starting with the publication of Noam Chomsky's *Syntactic Structures*, it went through a number of stages, increasing the systematic insight into the structure of the human language capacity as a species-specific component of the biologically based cognitive endowment. The first stage proposed a system of levels that systematically characterize the formal properties of linguistic expressions. This system took up assumptions of American structuralism, turning it, however, into a complex theory of mental organization, which strongly contrasted with the behaviorist orientation of previous approaches of Bloomfield, Harris, and others. A characteristic feature of this stage was the central place assigned to the concept *grammatical transformation*, which for some time led to the identification of generative grammar with the more specific notion of transformational grammar. The second stage of the 'Generative Enterprise' was the so-called Standard Theory, according to which grammatical transformations relate two levels of syntactic organization of an expression, viz. its deep and surface structure, determining the semantic and phonetic interpretation, respectively. Further generalizations led to the third stage, called the Principles-and-Parameters Theory, which showed most of the properties accounted for by idiosyncratic syntactic rules to be derivable from quite general principles. This turned out to be a crucial change in theoretical orientation. Individual languages now seem to differ only with respect to lexical information and a restricted set of parameters to which the otherwise universal principles are sensitive. In subsequent stages of the so-called Minimalist Program, the language-specific principles of this theory are further reduced to general conditions of cognitive organization, leaving only a restricted set of formal features that account for the language-specific properties of the computational system that correlates invariant patterns of articulation with conceptually organized representations of experience, roughly sound and meaning. Generative grammar was from the very beginning one of the driving forces within the field of cognitive sciences, stimulating research in language production and comprehension, language acquisition, language change and typology, computational linguistics, neurolinguistics, and the biology of language, including the evolution of the language capacity. It has strongly influenced discussions in formal semantics and pragmatics, and the philosophy of language in general. A number of offsprings, such as case grammar, generalized phrase structure grammar, or lexical functional grammar, pursued issues in different though related ways.

Generative grammar is the label of the most influential research program in linguistics and related fields during the last half century. Initiated by a short book, Noam Chomsky's *Syntactic Structures* (1957), it became one of the driving forces among the disciplines jointly called the cognitive sciences. The term *generative grammar* refers to an explicit, formal characterization of the (largely implicit) knowledge determining the formal aspect of all kinds of language behavior. The program had a strong mentalist orientation right from the beginning, documented, e.g., in a fundamental critique of Skinner's *Verbal Behavior* (1957) by Chomsky (1959), arguing that behaviorist stimulus–response theories could in no way account for the autonomous and complex character of ordinary language behavior. The 'Generative Enterprise,' as the program was called in Chomsky (1982), went through a surprising development with remarkable changes and modifications, pursuing however the completely coherent and constant research strategy, attempting to characterize and to explain the properties of linguistic knowledge. Each stage of this research program extended, deepened, and simplified previous results and insights and was accompanied by consequences within the narrower domain of linguistics as well as the wider range of related fields, such as ontogenetic development, psychology of language use, problems of biological evolution, or philosophy of mind.

Transformational Structure and Levels of Description

The foundation of all further developments is to be found in the overwhelming monograph *The Logical Structure of Linguistic Theory* (Chomsky, 1975, henceforth LSLT), which was published only 20 years after the manuscript was completed. Starting from the notion of structural analysis as developed by Bloomfield and made precise by Harris (1951), LSLT introduces a radically new perspective on the nature of linguistic structure. On the one hand, LSLT takes up and systematically extends Harris's notion 'level of linguistic structure,' in terms of which the form of utterances, i.e., their phonetic shape and syntactic organization, is characterized. On the other hand, however, LSLT develops a notion of linguistic structure that differs fundamentally from the empiricist orientation of American structuralism. More specifically, Chomsky proposes a notion of grammar that provides a full and explicit account of the tacit knowledge a speaker/hearer has of the form of utterances of his or her language. Under this perspective, a grammar G characterizes the internal states of knowledge that relate only indirectly to the overt, observable behavior of the speaker. In other words, G is viewed as an empirical hypothesis about the mental structure underlying characteristic aspects of ordinary speech behavior. To make this view precise, LSLT highlights the importance of intuitive judgments a speaker is able to make about properties of utterances, such as the grammatical well-formedness of the famous

nonsense-sentence (1) as opposed to the ungrammaticalness of (2), or the rather different properties displayed by the superficially similar sentences (3) and (4), where (4) is deviant, even though one might make sense of it:

1. Colorless green ideas sleep furiously
2. Furiously sleep ideas green colorless
3. She is easy to convince
4. She is easy to be convinced

Hence G of English has to account for the properties on which intuitions like these rely. The study of the general features of G can furthermore be expected to reveal principles of the internal organization of mental states in general and of knowledge of language in particular. This is the reason for the impact generative grammar had on the neighboring fields.

More technically, LSLT gives an explicit formal account of two core notions of linguistic theory, viz. linguistic level and rule of grammar.

A linguistic level **L** is construed as an algebraic structure based on a set L of primes, in terms of which a set μ of so-called **L**-markers is constructed. The elements of μ exhibit **L**-specific properties and relations, providing, roughly speaking, the representation of linguistic expressions at the level **L**. The levels assumed in LSLT (following in part structuralist traditions) include the phonetic level **Pn**, the phonemic level **Pm**, the morpheme level **M**, the word level **W**, the level of word classes **C**, the phrase structure level **P**, and the transformational level **T**. For **Pm**, the set μ consists of strings of phonemes, for **W**, the elements of μ are sequences of words, and for **P**, the set μ consists of phrase markers, roughly tree structures indicating the constituency of an utterance. The relation between the different levels is anything but trivial and the topic of much further research. The complete structural representation of an utterance comprises the **L**-markers assigned to it at the various levels in question.

The rules of a grammar G consist of the conditions and operations that generate the expressions of a given language, together with their structural representations. Technically speaking, the rules of G determine for each level **L** the set of **L**-markers and the way they are related to the other levels. After a short account of phonological rules, defining the relation between the phonemic and phonetic level, LSLT deals extensively with phrase structure rules, generating phrase markers, and with grammatical transformations. They play in fact a central role in LSLT, such that for years the term 'transformational grammar' was taken as equivalent to 'generative grammar.' Transformations needed to capture central structural properties, which P-markers cannot account for. A simple case is shown in (5) and (6), which do not differ by constituency, but by their different transformational derivation, as informally indicated in (5′) and (6′):

5. this picture was painted by a real artist
6. this picture was painted by a new technique

5′. a real artist painted this picture == $T_{Passive}$ ==> this picture was painted by a real artist

6′. someone painted this picture by a new technique == $T_{Passive}$ ==>
this picture was painted by a new technique by someone == $T_{Agent\text{-}deletion}$ ==>
this picture was painted by a new technique

While phrase structure rules are based on a restricted type of formal operations, transformations are complex mappings, specified by pairs $\langle Q, t\rangle$, where Q is a structural index and t is a sequence of operations. Q determines the constituent parts an expression must exhibit for the elementary operations of t to apply to them. A feature that transformations in LSLT inherit from the kind of analysis proposed by Harris (1957), where transformations do not only relate individual sentences to their more elementary basis, as in (5′) and (6′), but are also used to explain the composition of complex sentences from their elementary parts, called 'kernel sentences.' Thus, e.g., (7c) is composed by embedding (7a) into (7b), replacing the bracketed pronoun:

7. a. this picture was painted by a new technique
   b. the director suspected [it]
   c. == $T_{Embedding}$ ==> the director suspected [this picture to be painted by a new technique]

Thus a distinction between simple transformations like $T_{Passive}$ and so-called generalized transformations like $T_{Embedding}$ must be made. Another problem was related to the distinction between obligatory and optional rules and the conditions underlying this distinction.

An important point of the framework developed in LSLT is the assumption that any linguistic theory must provide criteria determining the choice among different grammars of the same descriptive adequacy accounting for the same range of linguistic properties. To this effect, a simplicity metric is proposed which prefers a grammar that is simpler than its competitors if – but only if – greater simplicity corresponds to greater descriptive adequacy. Thus a grammar G that systematically accounts for the different properties of (5) and (6) is selected over a grammar G′ that can only assign to them the same constituent structure, even though G′ were formally simpler, as it is not descriptively equivalent. For this reason, a grammar with a transformational component could turn out to be more adequate than one without according to the simplicity metric, although it has more complex formal specifications. What is at issue here is not merely the general methodological maxim by which the simpler description is to be preferred; it rather concerns a substantive claim about linguistic knowledge, according to which the simpler grammar matches the way linguistic knowledge is spontaneously organized. This turns directly into an empirical issue with regard to the problems of language acquisition, i.e., the fact that each normal child acquires the notoriously complex knowledge within highly restricted limits of time and exposure. Facing this tension, the speculative answer given within the generative enterprise was right from the beginning the claim that the language faculty must provide the relevant structure as a biological endowment. To redeem this claim was a continuous stimulus of further developments.

The productivity of the proposed framework is demonstrated by the analysis of a large fragment of English with respect to phrase structure and transformations in LSLT and in a stimulating monograph on nominalization by Lees (1960). A seminal study of phonology is given in Halle (1959) for Russian and in Chomsky's unpublished master's thesis on the *Morphophonemics of Modern Hebrew*.

A characteristic feature that the Generative Enterprise can be said to inherit from American structuralism is the assumption that all grammatical structure, from the most concrete level **Pn** of phonetics to the most abstract level **T** of transformational structure, is a matter of linguistic form, which must be embedded in a wider context of semiotics that accounts for sense, reference, and use of linguistic expressions. Although there have been several proposals to include a level of semantics into the framework of generative grammar, the main emphasis was on the knowledge of linguistic form throughout.

The Standard Theory of Generative Grammar

Immediately after the publication of *Syntactic Structures* (actually lecture notes excerpted from LSLT), the theory it proposed attracted interest from various sides, including notably mathematical linguistics and psychology. An overview of formal properties of grammatical rules, especially of phrase structure grammars, in Chomsky (1963) established what became known as the Chomsky hierarchy of formal grammars as a branch of subrecursive function theory. Miller and Chomsky (1963) gave a first outline of the role of generative grammar in the study of language comprehension and production. Furthermore, Katz and Fodor (1963) made a first proposal to extend the theory of grammar by a semantic component, using so-called projection rules to compositionally integrate the meaning of lexical items according to their syntactic relations.

The main development, however, was due to continuous efforts to systematically constrain and strengthen the theory, increasing its explanatory power by reducing the range of possible grammars compatible with a given set of data, as each systematic step to that effect decreases the task of the evaluation metric. Efforts to this effect led to what was later called the 'Standard Theory' of generative grammar. Based on preparatory ideas discussed a.o. in Katz and Postal (1964), this is formulated in Chomsky's *Aspects of the Theory of Syntax* (1965). At its core is a reassessment of the nature of transformational structure. Instead of a separate level **T** of transformations, grammatical transformations are now construed as operations that derive the relation between two levels of syntactic representation, both of which are characterized by P-markers. This modification is related to one of the most influential distinctions proposed in the Standard Theory, viz. that between *deep structure* and *surface structure*. On this view, each sentence exhibits a deep and a surface P-marker, with transformations deriving the latter from the former. Thus except for the different choice of personal or wh-pronouns, (8)(a)–(d) are just different surface realizations of the same deep structure, where the bracketed Agent phrase might furthermore optionally be deleted by $T_{\text{Agent-deletion}}$, known from (6′).

8. a. John may expect someone to convince her
   b. someone may be expected by John to convince her
   c. who may John expect someone to convince
   d. who may be expected by John to be convinced (by someone)

On this account, the semantic interpretation of linguistic expressions seems to be determined by their deep structure, which would be an interesting step toward an integration of the grammar into an account of meaning and reference. Another important aspect of the Standard Theory is a more systematic distinction between (1) the lexical system, i.e., the complex system of rules providing lexical items with their (idiosyncratic) phonetic, syntactic, and semantic properties, and (2) the grammatical rules that integrate the lexical information into the structure of complex linguistic expressions. Altogether, a more perspicuous design of the components of grammar and their relation to the levels of structural representation appears to emerge, compared to the system assumed in LSLT. The syntactic levels that seem to be crucial are deep- and surface-structure, while word- and morpheme-structure do not have an independent status. The components of the grammatical rule system are (1) the base component, including the lexical system and the phrase structure rules, (2) the transformational component, (3) the phonological component, and possibly (4) the semantic component. Rules and representations are interrelated according to the schema in (9), with levels of syntactic representation underlined:

9.

```
                 ⎧ Phrase Structure Rules
Base Component:  ⎨          ⇓
                 ⎩   Lexical Rules
                            ⇓
                   Deep Structure   ⇒   Semantic Component
                            ⇓
                   Transformations
                            ⇓
                   Surface Structure  ⇒   Phonological Component
```

The base component and the transformations together constitute the syntactic component.

The semantic component raises problems to be taken up below, while the organization of the phonological component, extensively discussed in Chomsky and Halle (1968), is assumed to have a fairly clear organization, parallel to that of the syntactic component in relevant respects. The underlying phonemic structure, made up from the phonological features of lexical items, is similar to deep structure

with phonological rules deriving the phonetic surface organization:

10. $\underbrace{\underline{\textit{Phonemic Structure}} \Rightarrow \text{Phonological Rules} \Rightarrow \underline{\textit{Phonetic Structure}}}_{\text{Phonological Component}}$

Remarkably simplified is especially the form and function of grammatical transformations. Thus generalized transformations are completely dispensed with, as the recursive deep structure of complex sentences is generated already in the base component. A particularly elegant feature that emerges from this reorganization is the principle of cyclic operation of the rules, which is added to the assumption already made that rules are ordered with regard to their application. The problem that immediately arises is the distinction between extrinsic (i.e., stipulated) and intrinsic ordering, the latter emerging from independent properties of rules and their organization. Cyclic operation apparently captures a central property of natural languages: the same sequence of operations applies to increasingly larger parts of an utterance, defining its syntactic properties, its phonetic patterns, and the compositional integration of its semantic structure.

An important general effect of the Standard Theory is the perspicuous account it suggests for the traditional notion of linguistic universals, or universal grammar (UG), considered as the innate, biologically fixed structure underlying the species-specific language capacity. The range of possible rules and the architecture of their interaction, as schematized in (9) and (10), could now be construed as formal universals, the potential primitive elements entering the rules and representations as substantive universals. Under this perspective, the set of possible grammars G accessible as instantiations of UG determines the diversity and typological differentiation of possible natural languages, including the range of variation delimiting processes of linguistic change. By the same token, UG provides the innate, formal basis and predisposition that makes language acquisition possible. These accompanying considerations of the Standard Theory led to a new, principled approach to problems of language acquisition, as well as linguistic change.

Extensions and Modifications of the Standard Theory

Shortly after the Standard Theory was formulated, empirical analyses as well as efforts to improve the structure of the theory led to proposals modifying the model. Controversial arguments concerned the role of semantics as indicated in (9). Lakoff (1971) and others proposed a model, according to which semantic representations are the underlying structure of linguistic utterances, from which surface structures are derived step by step by means of transformations and lexical rules. The insights and difficulties created by these proposals triggered a counterproposal, called the Revised Standard Theory. On the basis of observations discussed in Jackendoff (1972), it became clear that meaning is not unaffected by surface structure phenomena. For example, (11) and (12) rely on the same deep structure, with passivization applying to (12) but not to (11), but they do not in general have the same meaning:

11. Three students in the class wrote only one paper
12. Only one paper was written by three students in the class

Due to what is called the scope of quantification, in (11) the restriction of having written one paper is ascribed to three students, while in (12) the restriction of being written by three students is ascribed to a single paper. While (11) does not decide whether the students jointly or individually wrote one paper, the latter option is practically excluded in (12). As phenomena of this sort are by no means marginal, the Revised Standard Theory recognized a more complex relation among syntax and semantics than that indicated in (9). Certain aspects of meaning are determined by deep structure, while scope and certain other relations depend on surface structure, such that transformations – contrary to the Standard Theory – might still affect the semantic interpretation.

A technical way to handle these problems is the so-called Trace Theory proposed in Fiengo (1977). According to this proposal, a constituent that is moved, i.e., reordered, by a transformation leaves behind a trace that is phonetically empty, but may play a role in semantics, as indicated in (13), where e_i is the trace left in the initial position of the co-indexed pronoun *who*:

13. who_i do you want to talk to e_i

In many cases, the trace turns out to be similar to what logicians consider a variable bound by the operator contained in the moved constituent. In this sense, invisible elements of the surface structure account for its contribution to semantic interpretation. Other types of empty elements of the surface structure are 'invisible' pronouns, as in (14) (a) and (b), where the actor of the infinitive clause is controlled by (i.e., coreferential with) different constituents of the main clause:

14. a. John_i promised Mary [e_i to take the train]
    b. John recommended Mary_i [e_i to take the train]

Now, representations like those in (13) are based on transformations and cannot therefore belong to the pretransformational deep structures of (9); but they cannot belong to the surface structure either, as they contain abstract elements like traces and empty pronouns, which don't have phonetic realization. To reconcile these problems, a reorganization of linguistic knowledge has been considered, providing a more abstract syntactic representation called S-structure, on which the semantic as well as the phonetic interpretation can depend. Thus (9) would have to be replaced by the schema sketched in (15), where the Logical Form LF is the place of all the information by which G determines the meaning of an expression, just

as the Phonetic Form PF specifies the conditions G imposes on its pronunciation:

15.
```
      Syntax
        |
  S-structure
    /     \
  PF       LF
```

The Trace Theory furthermore plays a central role for more general conditions by which transformations and their effects are constrained. Thus in cases like (16)(a), where the subject of the embedded Verb Phrase *take the train* ends up as the pronoun *who*, this relation is mediated by a chain of positions as marked in (16)(b). This chain is created by the cyclic application of one operation to a single constituent along the recursive embeddings indicated in (16)(b), as the deficiency of (16)(c) strongly suggests.

16. a. Fred knows who John might persuade to be ready to take the train
    b. Fred knows [who$_i$ John might persuade e_i [e_i to be ready [e_i to take the train]]]
    c. Fred knows [who$_i$ John might persuade him$_k$ [e_k to be ready [e_i to take the train]]]

Observations like these entered the more general endeavor initiated in Ross (1967), to heavily reduce the complexities of the structural index Q and the possible operations *t* of transformations. In this way, transformations are eventually to be reduced to one general operation '*Move* α,' where α is any constituent, all further constraints deriving from principles imposed by UG. These general principles also include a systematic simplification of phrase structure rules in terms of so-called X-bar structure. This structure takes up a proposal made already in Harris (1951), according to which phrase structure combines syntactic constituents X and Y to form a complex constituent $\bar{X}$ with X as its dominating head. In somewhat simplified terms, the properties of $\bar{X}$ (e.g., *drink wine*) are determined by its head X (viz. *drink*), where Y is either a selected complement of X, as in *drink wine*, or a free adjunct, as in *drink slowly*. The selectional properties of X make up its lexically determined Argument-Structure, called Thematic Grid, an important syntactic aspect shaping the semantic interpretation such as <Agent, Theme> for the verb *drink*.

All these modifications of the Standard Theory were motivated and supported by descriptive problems, arising from a wide range of well-known as well as unrecognized facts; at the same time, however, they were driven by the central aim to restrict the wide range of possibilities UG provides, thereby strengthening the explanatory capacity of the theory.

The Principles-and-Parameters Theory

A major breakthrough in this direction was achieved by the so-called Principles-and-Parameters Theory, formulated in Chomsky's *Lectures on Government and Binding* (1981). The basic idea of this new framework is to dispense with the notion of UG as a rich system of different types of rules, individual grammars must be made up from, relying instead on a restricted system of universal principles that constrain the effect of a few basic operations including *Move* α and the schema of X-bar theory, generating and interrelating representations according to the basic architecture indicated in (15). The principles in question specify the conditions these general operations are subject to, determining, e.g., the relation between the positions of moved constituents as in cases like (13) or (16), specifying the features that candidates of possible movement must or must not exhibit or the conditions that trigger the operation in question, and a fair range of further intricate properties. These principles constitute separate subsystems of linguistic organization, such as the Theory of Government, of Binding, of delimiting Boundaries, of Control, of Case-features, and the effect of Thematic roles. Presupposing the levels indicated in (15), the components of a grammar G in the framework of Principles-and-Parameters can roughly be indicated as follows:

17. a. Lexical system, fixing the idiosyncratic phonetic, semantic, and syntactic properties of lexical items
    b. Syntactic operations:
       X-bar syntax,
       Move α
    c. Principles of Government, Binding, Bounding, Thematic structure, Control, and Case-Assignment
    d. PF-component (assigning phonetic interpretation)
    e. LF-component (adjusting to referential and conceptual conditions)

Extensive studies dealt with further consequences of this framework, such as Kayne's (1984) exploration of general properties of phrase structure or Rizzi's (1990) clarification of more specific conditions on movement operations, among a wide range of other work.

Two important points of this conception must be emphasized. First, the systems of principles in (17)(c) are essentially different from the grammatical rules assumed in the Standard Theory: while rules were supposed to be necessarily different for different languages, the principles of Government, Binding, Case, etc. are construed as providing the universal organization of the language capacity; in other words, they are not elements of particular grammars, but are part of UG. Hence they are not subject to language acquisition, but are part of the biological endowment. Second, the obvious structural differences between different languages are no longer conceived as consequences of intricate systems of different rules. They are rather explained by a restricted set of parameters contained in the general principles, to which particular languages can assign different values. For instance, the head X of a constituent $\bar{X}$ may be in the initial position of $\bar{X}$, as in *drink wine*, or in the final position, as in German *Wein trinken*, depending on the so-called head-parameter, or the interrogative pronoun can be required to be moved to the clause initial position, as in English, or stay *in situ*, as in Japanese, etc. Most (or all) of these parameters seem to be related to classes of particular lexical items, so that all language-particular properties are basically related to lexical information.

This leads to a rather different view of how UG determines an individual grammar G. Besides lexical items making up the language-particular dictionary, grammars of different languages might differ only by their choice of parameter values,

while the principles presupposed by these parameters are given by UG and hence identical for all languages. Under this perspective, language acquisition consists essentially in identifying the lexical system of a language, thereby fixing also the particular values of the parameters contained in UG. This must not be misconstrued as reducing the importance of the implicit knowledge of G: the very essence of the language faculty is its combinatorial capacity.

These observations relate to one of the core issues of generative grammar, called the logical problem of language acquisition. This problem concerns the need to explain the surprising fact that the organism normally develops an appropriate grammar on the basis of rather restricted, deficient, and partially corrupt evidence which the child is normally exposed to. This problem could not seriously be approached, if UG were as unconstrained as the Standard Theory would have to assume, providing all the options the unconstrained transformational component would allow for. As a matter of fact, much of the overwhelming capacity of the Standard Theory is to be dispensed with, because on closer inspection complex operations such as passive transformation or relative clause formation usually had to be reduced to the combination of more elementary operations such as *Move* α under general constraints. Thus the Principles-and-Parameter Theory is an important step toward an explanation of the human language capacity.

In view of the continuous modification and improvement of generative grammar, the shift to the Principles-and-Parameters conception is certainly the most important change. The preceding concepts and changes can be considered as preparing the present model, which the following changes and amendments unfold and develop. Based on the instruments and tools explored so far, it is now the type of problems and the kind of explanation to be developed which come into focus.

Among the numerous stimulating consequences of this model is a new perspective on historical and typological variation. Like language acquisition, linguistic change and typological differences are not only constrained by UG, but also closely related to the options provided by the parameter values on the basis of universal principles.

The Minimalist Program of Generative Grammar

Taking the Principle-and-Parameter Theory as a starting point, a new phase in the Generative Enterprise, called *The Minimalist Program*, was initiated in Chomsky (1995). Striving, as before, for an explanatory account of the Language Faculty, two different though interrelated issues are systematically explored within this program. One of these issues goes on to reduce the rules and processes of the computational system of language to more basic operations, with the reduction of grammatical transformations to the simple operation *Move* α serving as the first programmatic step. In the same spirit, the whole X-bar syntax, i.e., the residue of the system of phrase structure rules, is now reduced to one single operation *Merge*, which just combines two syntactic objects X and Y into one unit Z, automatically turning the head of X or Y into the head of Z, requiring the selectional features of the head to be saturated, determining at the same time the categorial properties of Z. In a way, *Merge* includes a highly restricted version of generalized transformations, insofar as both X and Y may be the result of previous merging. On this background, the next important simplification comes as a natural step: *Move* α can be construed as the operation that takes up α and copies, or rather remerges, it at a predetermined position, usually the edge of the domain on which *Move* operates. In other words, *Move* turns out to be just another instance of *Merge*, which does not add a new element, but reassigns an already merged element on a new position. This so-called 'internal Merge' recombines an element, in contrast to ordinary 'external Merge,' which adds new material. In any case, the generative system reduces to the recursive application of *Merge*, with the various systems of (parameterized) principles and conditions controlling its effects.

The computational system underlying the knowledge of language can thus be construed as a single recursive operation that recruits for each expression E a number of lexical items N (a so-called Numeration) to combine them – constrained by the parameterized principles of UG – into the complex linguistic structure of E, as schematized in (18), where |UG| abbreviates the principles of UG with the parameters fixed, and E provides the PF and the LF of the expression.

18. N === *Merge* ===> E
⇑
|UG|

The second issue pursued in the Minimalist Program concerns the possibility to derive the principles of UG as far as possible from more general (possibly language independent) conditions of cognitive organization plus the boundaries that are conceptually necessary to constitute the domain of linguistic knowledge. Minimal assumptions require linguistic expressions to compositionally relate sound and meaning, or more technically to determine a systematic correspondence between patterns originating from the perceptual and articulatory system A–P, i.e., the sound structure of E, and representations in the range of conceptually and intentionally organized experience C–I (the conceptual–intentional (system)), roughly the overall domain of thought. This requires the internal or I-language to provide a compositional correspondence of PF and LF as the interfaces, in terms of which a linguistic expression E relates to or interacts with A–P and C–I, respectively, as sketched in (19):

19. signal <===> A–P ⇔ PF ←--→ LF ⇔ C–I <===>
environment

Hence, a grammar G determines a pair $<\pi,\lambda>$ for each linguistic expression E, with π a representation in PF and λ a representation in SF and E is derived from a numeration N of lexical items according to (18). Now, the research strategy of the Minimalist Program proposed in Chomsky (1995) and elaborated in Chomsky (2000, 2002) explores the possibility to derive the properties of UG as far as possible from these minimal assumptions together with the hypothesis that UG is subject to conditions of structural economy. To this effect, lexical items selected by a numeration N must be supposed to contain all and only the idiosyncratic specifications by

means of which the elements of N participate in computation of the pair <π,λ>. Besides the phonetic and semantic features to be interpreted in the extralinguistic domains P–A and C–I, respectively, these specifications include certain grammatical or formal features controlling by means of agreement, Case-selection, and other grammatical requirements the purely language internal computation of <π,λ> from N.

Two types of very general conditions are supposed to influence the computation of linguistic expressions, called full interpretation (FI) and shortest derivation (SD). As to FI, the ultimately derived interface representations must be appropriate to the interpretive substance of their respective domains. This means that representations in PF and SF must not contain uninterpretable, vacuous elements or formal features and configurations that cannot be interpreted in P–A and C–I, respectively. This requires formal features to be eliminated on systematic grounds in the derivation of PF–SF pairs, otherwise the resulting expression is not well formed, i.e., ungrammatical. As to SD, the admission of internal *Merge*, that is the movement of elements already merged, does on the one hand extend the derivation of E from N beyond the quest for the smallest number operations, but it allows on the other hand grammatical operations to delete uninterpretable features by meeting, e.g., agreement conditions, thereby avoiding violations of FI.

The intricate interaction of conditions and features on which these operations depend seems to proceed stepwise, barring movement of constituents and matching of features, once the PF-representation is spelt out, i.e., has access to phonetic interpretation. Thus the schema in (18) would have to be modified as in (20), assuming that Merge proceeds in what has been called phases, namely combining elements from N, until a unit with certain properties (presumably some kind of propositional completeness) is reached and the next phase can start, until N is exhausted – a procedure that takes up the cyclic operation of the Standard Theory.

20. N ====> *Merge* ====> LF
⇓
Spell out
⇓
PF

The requirement of FI presupposes external and internal *Merge* to check and eliminate step by step the formal features, leaving only information that can appropriately be interpreted in terms of articulatory and conceptual conditions at the two interface levels. A remarkable range of facts in different languages has been shown to follow from appropriate lexical information together with minimal assumptions about *Merge*, *Move*, plus the two sorts of general conditions regulating the economy in representation and derivation.

A specific issue raised by the Minimalist Program is the question whether the design of the computational system of language is optimal, given the external conditions A–P and C–I noted in (19) and the plausible requirements of FI and SD. With respect to this question, it is a peculiar fact about natural languages that the derivation from N to <π,λ> cannot in general be restricted to the simplest possible option, requiring only the operation of simple, external *Merge*. The language capacity obviously exploits freely the effect of what was called *internal Merge* (or *Move* α), using 'displaced' constituents as borne out even by simple cases like *Who*$_i$ *did he talk to* e_i. It is an interesting but undecided question whether this obvious deviation from what would be the apparently optimal design is still in line with the most efficient way to establish the correspondence between PF and SF – as it might in fact be the case, if different relations within distinct aspects of C–I would have to be captured. Two things must be emphasized in this respect. First, considerations like these (and in fact the Minimalist Program as a whole) go beyond just methodological concerns, following only the usual concern for the simplest possible theory, but have to do with the subject matter as such, namely the organization of the language faculty. The Minimalist Program raises the question: How far does the computational system of the language faculty represent the optimal solution for the task of relating the two heterogeneous interfaces LF and PF? To clarify and motivate the specific details of the pertinent operations, different proposals are intensely explored, e.g., with respect to the features that categorize the different combinations built up by Merge. Second, these matters make sense only to the extent to which conditions based on C–I (that is on structures of thought or meaning) are taken as seriously as questions arising from A–P. As a consequence of the symmetry inherent in the design noted in (19), the Minimalist Program made the relation to standard views of semantics more explicit. Semantics is usually concerned with truth and reference, construed as relations of expressions to the environment, i.e., to external objects and states of affairs. In contrast to that, Chomsky (2002, 2012) argues strongly that these relations are barely accessible to scientific investigation. Even standard examples like '*water* refers to H_2O' are spurious in view of what is actually known about the structure of the relevant molecules and the use of items like *water*. Discussing a wide range of similar problems, Chomsky notes that the remarkable insights, which semantics nevertheless provides, do not really deal with the relation of expressions to the environment, but with the structure of LF and its relation to C–I as sketched in (19), with lexical items referring not to external entities, but related to concepts and their combinations. Under this perspective, LF concerns the formal structure of meaning, the 'content' of which is its conceptual interpretation in C–I, and Chomsky notes that syntax to a large extent is concerned with problems that are usually considered as semantics. This clarification does not only rationalize and slightly reduce the skepticism with respect to semantics noted in the beginning, it also widens and delimits the perspective of generative grammar.

It is to be noted that, from this point of view, the species-specific language capacity might essentially consist in the availability of discrete infinity and in fact be related, as Chomsky (1982) speculates, to the computational capacity underlying arithmetic. The language capacity clearly recruits and enhances general cognitive and communicative capacities, but its evolution, which Pinker (1994) persuasively argues to be due to adaptive selection, might well be an independent step accompanying evolutionary changes in the architecture of the human brain.

As a final example for the continuous extension and modification the Minimalist Program pursues, the interest in what Chomsky (2005) calls the 'third factor' will be mentioned. According to assumptions explored in the generative enterprise, two different factors are involved in the genesis of the internal

language: (1) the biological basis, characterized by UG and anchored in the genome, and (2) experience, i.e., the data the child encounters. To these conditions, those of the third factor have to be added. They comprise principles not specific to the language faculty, viz. principles of data analysis, of structural architecture, of organic form and action, of efficient computation, and they are based in part on overall conditions of physical organization, as should be expected, if language is a natural phenomenon.

General Perspectives

The intellectual force of generative grammar comes from the combination of two equally important factors. First, it provides a suggestive methodological and theoretical perspective for the study of language as part of the biological endowment of the human organism, supporting a central system of the overall cognitive capacity; second, it shows how this perspective can be made precise, turned into an effective research program, and pursued with respect to a large domain of empirical facts not recognized before. It is indeed the wide range of phenomena – from syllable structure and inflection to quantifier scope, focus determination, and syntactic embedding – as well as the diversity of languages – from English, German, and Chinese to Hebrew, Malaialam, and Walbiri – that gives the Generative Enterprise its unusual place in the field. The explicit formulation of the theory, its technical means, and its cognitive orientation are the reasons for its strong impact on a wide range of subdisciplines, from research in language comprehension and production, acquisition, and aphasia to typology, historical linguistics, computer science, poetics, theory of music, and philosophy of language.

During the stages sketched in (1)–(5), various offshoots have developed different, though related, ideas. These are a.o. categorial grammar, dependency grammar, case grammar, lexical functional grammar, generalized phrase structure grammar, head-driven phrase structure grammar, and relational grammar, to mention the more influential proposals. Another type of offspring is optimality theory, which – starting from a different background – develops certain ideas of the Minimalist Program in a different way, assuming that conditions on linguistic structure may be of different strength and can be violated.

A peculiar relationship must finally be noted between generative grammar and the various theories of semantics and pragmatics. Sticking to the initial notion that grammar can and must account for the structure, but not properly for the complexities of interpretation and use of language, Chomsky did not extend his theory of language to the domain of semantics and pragmatics. It might be noted, in this connection, that he essentially considered LF as a syntactic level, pertaining to the form, rather than the interpretation of language. Therefore, Formal Semantics as developed by Montague (1974), Discourse Representation Theory as proposed by Kamp in Kamp and Reyle (1993), Speech Act Theory as conceived by Searle (1969), and a fair number of related approaches to aspects of interpretation and use are compatible with and in part stimulated by generative grammar, but they have not been included within its proper domain. Even though this skeptical delimitation of the field was occasionally met with disappointment, it could not really diminish the intellectual attraction exerted by the Generative Enterprise – and has moderately been relativized by the needs of the Minimalist Program.

See also: First Language Acquisition, Linguistic Theory of; Grammatical Agreement; Linguistic Anthropology; Morphology in Linguistics; Movement Theory and Constraints in Syntax; Phonology; Semantics; Speech Perception; Speech Production, Psychology of; Valency and Argument Structure in Syntax; Vocabulary Acquisition; Word Order.

Bibliography

Chomsky, N., 1957. Syntactic Structures. Mouton, S'Gravenhague.
Chomsky, N., 1959. Review of Skinner's verbal behavior. Language 35, 26–58.
Chomsky, N., 1963. Formal properties of grammars. In: Luce, R.D., Bush, R.R., Galanter, E. (Eds.), Handbook of Mathematical Psychology, vol. II. Wiley and Son, New York, pp. 323–418.
Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.
Chomsky, N., 1975. The Logical Structure of Linguistic Theory. Plenum Press, New York.
Chomsky, N., 1981. Lectures on Government and Binding. Foris Publications, Dordrecht.
Chomsky, N., 1995. The Minimalist Program. MIT Press, Cambridge, MA.
Chomsky, N., 1982. The Generative Enterprise: A Discussion with Riny Huybrigts and Henk van Rimsdijk. Foris Publications, Dordrecht.
Chomsky, N., 1986. Knowledge of language: its nature, origin, and use. Praeger, New York.
Chomsky, N., 2000. Minimalist inquiries: the framework. In: Martin, R., Michaels, D., Uriagereka, J. (Eds.), Step by Step. MIT Press, Cambridge, MA, pp. 89–155.
Chomsky, N., 2002. In: Belletti, A., Rizzi, L. (Eds.), On Nature and Language. Cambridge University Press, Cambridge.
Chomsky, N., 2005. Three factors in language design. Linguistic Inquiry 36, 1–22.
Chomsky, N., 2012. The Science of Language – Interviews with James McGilvray. Cambridge University Press, Cambridge.
Chomsky, N., 2013. Problems of projection. Lingua 130, 33–49.
Chomsky, N., Halle, M., 1968. The Sound Pattern of English. Harperand Row, New York.
Fiengo, R., 1977. On trace theory. Linguistic Inquiry 8, 35–62.
Halle, M., 1959. The Sound Pattern of Russian. Mouton, 'S-Gravenhage.
Harris, Z.S., 1951. Methods in Structural Linguistics. University of Chicago Press, Chicago.
Harris, Z.S., 1957. Co-occurance and transformations in linguistic structure. Language 33, 293–340.
Jackendoff, R.S., 1972. Semantic Interpretation in Generative Grammar. MIT Press, Cambridge, MA.
Kayne, R.S., 1984. Connectedness and Binary Branching. Foris Publications, Dordrecht.
Kamp, H., Reyle, U., 1993. From Discourse to Logic. Kluwer, Dordrecht.
Katz, J.J., Fodor, J.A., 1963. The structure of a semantic theory. Language 39, 170–210.
Katz, J.J., Postal, P., 1964. An Integrated Theory of Linguistic Descriptions. MIT Press, Cambridge, MA.
Lakoff, G., 1971. On generative semantics. In: Steinberg, D., Jakobovits, L. (Eds.), Semantics. Cambridge University Press, Cambridge, pp. 232–296.
Lees, R.B., 1960. The Grammar of English Nominalzations. Indiana University, Bloomington.
Miller, G.A., Chomsky, N., 1963. Finitary models of language users. In: Luce, R.D., Bush, R.R., Galanter, E. (Eds.), Handbook of Mathematical Psychology, vol. II. Wiley and Sons, New York, pp. 419–491.
Montague, R., 1974. Formal Philosophy. Yale University Press, New Haven.
Pinker, S., 1994. The Language Instinct. Harper Collins, New York.
Rizzi, L., 1990. Relativized Minimality. MIT Press, Cambridge, MA.
Ross, J.R., 1967. Constraints on Variables in Syntax (Ph.D. dissertation). MIT (1986) Infinite Syntax! Ablex, Norwood, NJ.
Searle, J.R., 1969. Speech Acts. Cambridge University Press, Cambridge.
Skinner, B.F., 1957. Verbal Behavior. Appleton-Century-Crofts, New York.

Genes and Behavior: Lessons from the Fruit Fly

Ralph J Greenspan, University of California, San Diego, La Jolla, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Behavior draws on the activities of a wide range of genes acting at many different times in an individual's life. Model organisms have provided insight into the role of genes in behavior that have major implications for humans. Many of the mutations and genetic variants (alleles) that affect behavior in the fruit fly *Drosophila* have proved to be subtle changes in genes that, when more severely mutated, produce wider ranging effects. These studies based on model organisms have also provided information on the circuitry underlying behavior. Their study has made major contributions to our understanding of innate and learned behaviors, and points to the importance of network interactions among genes as well as among neurons in producing behavior.

Introduction

The idea that behavioral traits are passed from one generation to the next is an ancient one. Humans have acted on it for centuries in the breeding of domesticated animals and have assumed that it accounted for much of our own behavior as well. With the advent of the science of genetics, the question of how genes influence behavior has been recast as to which genes influence a given behavior and how they influence it. For humans, the question of the nature and extent of genetic influences on behavior is still hotly debated, and little is known about the latter question of mechanism. Still unresolved is the issue of how to unravel the interactions between genes and the environment in human affairs. There is no question that all behavior is affected both by genes and the environment in animals and in humans, and our ability to dissociate the two in humans is inevitably limited. However, because laboratory animals are raised under conditions where environmental differences between individuals can be minimized, the study of genetic influences on behavior can be focused on more sharply.

The principal lessons from such animal studies are that the relationship between genes and behavior is a complex one: multiple genes contribute to each aspect of behavior, and each gene participates in multiple aspects of behavior. Behavior is an emergent property of the organism's many interacting systems. Genes encode proteins, proteins build the constituents of cells, cells form the nervous system and other organ systems, and the nervous system ultimately gives rise to behavior. Thus, genes can affect behavior by participating in all the major kinds of biological processes – development, physiology, and the biochemistry of intracellular signaling. We will examine examples of such genes in the organism for which the most is known, the fruit fly *Drosophila melanogaster*. The value of a model system, however, must be more than its experimental versatility; it would be of no help if the phenomena and principles are fundamentally different from those of humans. Fortunately, they appear to be similar by an ever increasing number of criteria. One of the most striking outcomes of modern biology and the sequencing of genomes is the similarity of genes between distantly related creatures. The fruit fly *D. melanogaster* and humans share roughly two-thirds of the same genes in common. Similarly, they share considerable conservation of metabolic and signaling pathways at the cellular level. Most strikingly, and most relevant for this discussion, there are also specific instances of conservation at the level of behavior and its molecular mechanisms.

From Single Genes to Many

The earliest studies of the connection between genetics and behavior date back to the dawn of animal genetics, and were mainly directed at testing whether behavioral traits showed the same propensity to Mendelian inheritance as did physical traits.

Tropisms

The fruit fly, *D. melanogaster*, made its research debut in the laboratory of William E. Castle at Harvard University in 1901, the first laboratory to pursue the newly propounded principles of Mendelian genetics in animals, concentrating primarily on the coat color genetics of guinea pigs and rats. However, mammalian breeding studies take up a lot of laboratory space and at least four to five weeks per generation. So Castle took the suggestion of his entomologist colleague C.W. Woodworth and began working with the faster breeding fruit fly, testing how well it tolerated inbreeding and selection, and probing some of its simple behavioral responses to light (phototaxis), gravity (geotaxis), and odors. These responses are known as tropisms, the innate tendency to move toward or away from a particular stimulus, such as light, gravity, heat, or an odor.

None of these behavioral studies involved the use of genetic variants, and one may wonder where the idea came from to look at these behaviors. The most likely influence was the German zoologist Jacques Loeb who had moved to the United States and became a key figure in establishing the University of Chicago as one of the first research universities in America. Loeb's research on tropisms in flatworms was well known to all zoology students, and is referenced in these early *Drosophila* papers. A decade later in T.H. Morgan's fly laboratory, R.S. McEwen was the first to show that a mutant (i.e., a strain of flies in which an allele is drastically altered in comparison to those usually found in natural populations), the body color mutant

tan, was defective in phototactic behavior (we now know that this gene encodes an enzyme for pigment synthesis that also affects synthesis of the neurotransmitter histamine, which is released from the photoreceptor cells). In the 1950s, Jerry Hirsch had taken up this neglected field and shown that fly behavior could be dramatically altered after selective breeding for divergent phototactic or geotactic behavior (Hirsch, 1959). Hirsch's work was directed at showing that behavior has a basis in heredity, but his flies could not be further analyzed for which genes were responsible, let alone the mechanisms responsible.

A new angle on the role of genes in behavior emerged in the 1960s, growing out of the successful studies on the molecular genetics of bacteria and viruses. The microbial studies had yielded basic principles about the molecular nature of the gene, and had also established basic strategies for using genetics to analyze biological processes. One strategy is to detect individual genes essential in a process by inducing and isolating mutants that show specific defects in that process. The identities of the affected genes then give an outline of the actual 'steps' in the process. Another strategy uses altered genes of known function as tools to disrupt a process at specific points.

Armed with these strategies, one of the premier practitioners of microbial genetics, Seymour Benzer, inaugurated a new approach to the study of behavior using genetics in the fruit fly *D. melanogaster*. His aim was to 'dissect' behavior into its component genetic parts. At the time, all that was known with any certainty was that genes encode proteins, with little idea of what the genetic mechanisms of behavior might be. Ideas ranged from one gene per synapse to one gene per behavior. The primary focus was on the role and importance of single genes, but researchers also began early on to use genetic tools to identify portions of the nervous system relevant for particular behaviors. This latter work served as a precursor to today's emphasis on identifying the circuits underlying behavior.

Benzer began with phototaxis, as did Castle 60 years earlier, isolating many nonphototactic mutants. The outcome was not tremendously encouraging, as they were mostly blind mutants, which offered minimal insight into the mechanisms underlying a tropism. A few, however, did provide a window into brain mechanisms of visually driven behavior (see Section Anatomical Mutants, below). But Benzer and his students did not limit their efforts to phototaxis, and soon expanded into more interesting fly behaviors.

Complex Innate Behavior

Tropisms are not the only innate behaviors that flies do; there are also social behaviors that do not require learning or experience, of which courtship is probably the most elaborate. As the activity that accounts for most of their social life, it is rich in content, complex in elicitation, and robust in execution. One must practically hit the fly over the head, genetically speaking, to keep it from courting. Consistent with this picture, a male fly is fully capable of performing the entire courtship sequence in all its aspects even if raised in complete isolation from egg to adult and then presented with a virgin female as his first encounter with another creature. As an essential piece in the fly's evolutionary repertoire, it is not surprising that reproductive behavior should be reinforced and buffered against random events, whether developmental or experiential, that might subvert it.

Male flies must go through a ritual set of actions before a female will be receptive to mating. Female courtship behavior has no overt actions and consists primarily of allowing the male to mate. It has thus been studied far less. When a male encounters a female, he will orient, stand facing in her direction about 0.2 mm away (flies are 1.5 mm long), tap her on the abdomen with his foreleg, follow her if she walks away, and then extend one wing and vibrate it to produce a 'love-song'. Depending on the female's level of interest at this point, he may go back and repeat these actions or he may go on to the next stage, in which he extends his proboscis and licks the female's genitals (Figure 1).

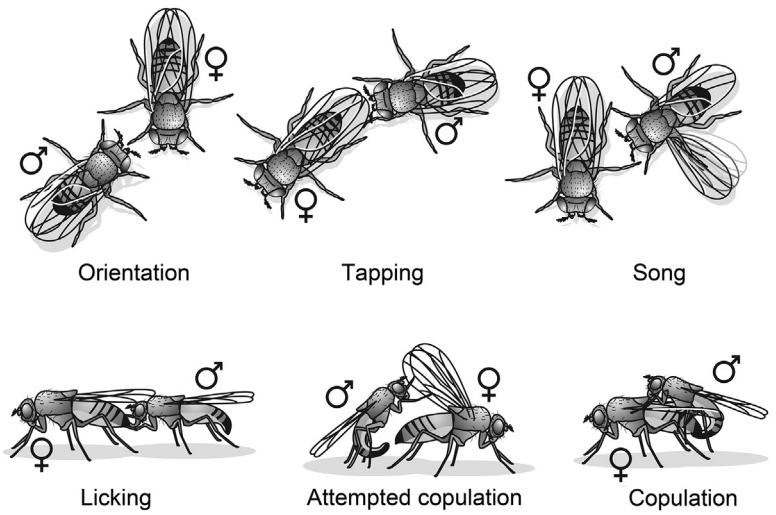

Figure 1 The sequence of behaviors in heterosexual courtship of *Drosophila melanogaster*. Reprinted from Greenspan, R.J., 2007. An Introduction to Nervous Systems, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY: pp. 111, Figure 7.5. Copyright 2007, with permission from Cold Spring Harbor Laboratory Press.

He then may go on to mount her and, if she is receptive, copulate. Female fruit flies will not mate unless the males have gone through this routine and the female has become receptive.

Sex-specific behavior in the fruit fly world is as much a part of the genetic difference between males and females as appearance and reproductive physiology. In flies, each cell determines its choice of sexual phenotype independently, based on its chromosomal makeup – one X for male and two Xs for female in normal diploid flies. This sets off a cascade of sex-determining genes that produce the various differences, or dimorphisms. In the cells that are precursors of the gonads, this genetic cascade produces the appropriate genitals for each sex. In the brain, it produces the circuitry for sex-specific behavior.

A gene called *fruitless* is essential for much of this male-specific development (Ryner et al., 1996). The gene encodes multiple forms of a transcription factor, one of which is male specific. In males that are missing the male-specific product, courtship does not occur. Conversely, when this male-specific product of the *fruitless* gene is expressed in a developing female brain, the adult female will perform malelike courtship behavior, which is never otherwise seen in females (Stockinger et al., 2005). The male-specific product of the *fruitless* gene is expressed in most of the brain regions shown to undergo sex-specific development for courtship, and this male-specific product appears to do little else apart from governing sex-specific development in the male nervous system (Figure 2). Other gene products from the same locus, which are not sex-specific, have more vital functions.

The *fruitless* gene's specificity would seem to belie the earlier claim about the relationship between genes and behavior being complex, but in 50 years of fly behavior mutant hunting, it is the only example of its kind. All other genes in which behavioral mutations have been found are notable for affecting many different phenotypes (technically referred to as 'pleiotropy'). Instead, the specificity of those mutations comes from the restricted nature of the mutation's effect on the gene's expression or activity (Greenspan, 2001).

Plasticity and Learning

Fruit flies can be trained to remember, to learn to tell the difference between smells or between things they see. If an odor (or image) is accompanied by an electric shock or a heat lamp, flies quickly learn to avoid it; the memory of this association will last for a day at most. What does it take to get them to remember it for longer? Regular conscientious practice (just like your piano teacher may have told you). In other words, as any student knows, cramming for an exam (i.e., all the training at once), can get you through the next day, but it does not stick. Studying at regular intervals, on the other hand, produces a much longer lasting memory. Fruit flies are no different. If multiple training trials are delivered all at once, memory still only lasts for a day. However, if the training trials are interrupted with 15-minute rest intervals, then the flies will remember it for up to a week. This is a substantial portion of a fruit fly's lifespan – midlife crisis has already set in by their second week of adulthood.

The first two genes identified from screens for learning and memory defects uncovered two components of a cyclic adenosine monophosphate (cAMP) cascade: *dunce* encodes a cAMP phosphodiesterase and *rutabaga* encodes a type I Ca^{2+}/calmodulin-stimulated adenylyl cyclase. These are two components of the same cascade that serotonin acts upon in the gill withdrawal reflex. Furthermore, cyclic AMP response element binding (CREB) turns out to be just as important in flies as in sea snails: flies with deficiencies of CREB also have poor memory (Yin et al. 1994). The cAMP pathway acts in mammals as well as in invertebrates; mutations in CREB, for example, alter memory in mice, and there is a human mental retardation syndrome due to mutations in the cyclic AMP response element-binding protein (CREB)-binding protein gene.

More recently, many of the genes required for the storage and retrieval of long-term memory have been identified. The classes of function found in these genes have shown that long-term memory storage makes use of cellular mechanisms involving the localization of mRNAs to particular regions of the neuron, as well as those that regulate translation of mRNAs.

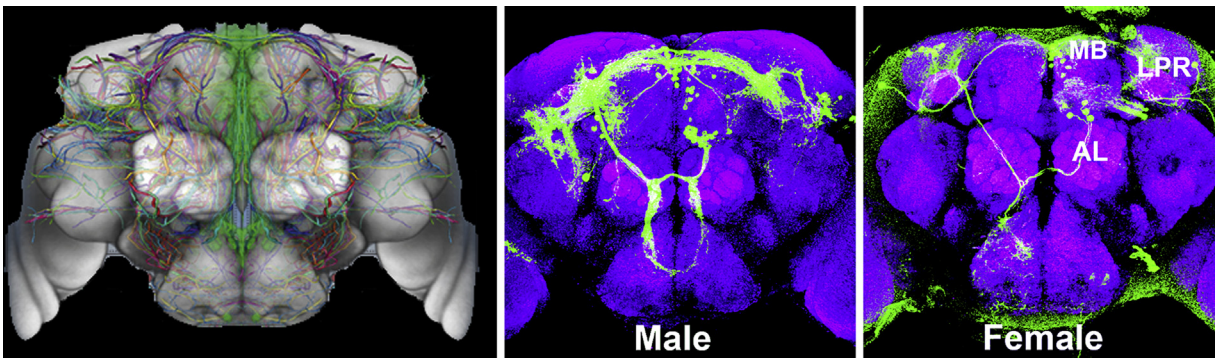

Figure 2 (Left) Male-specific neurons expressing the *fruitless* gene. (Center and right) Male vs. female difference in *fruitless* expressing neurons projecting from the olfactory lobes into the brain. Left panel modified from Yu, J.Y., Kanai, M.I., Demir, E., Jefferis, G.S.X.E. et al., Cellular organization of the neural circuit that drives *Drosophila* courtship behavior. Current Biology 20, 1602–1614. Copyright 2010, with permission from Elsevier. Center and right panels reprinted from Kimura, K., Ote, M., Tazawa, T., Yamamoto, D., 2005. Fruitless specifies sexually dimorphic neural circuitry in the *Drosophila* brain. Nature 438, 229–233. Copyright 2005, with permission from Nature, Inc. AL: antennal lobes; LPR: lateral protocerebrum; MB: mushroom bodies.

Unlike the case of courtship and *fruitless*, the genes implicated in learning and memory are involved in many different cellular functions. Thus, mutations in these genes only exert specific effects on learning when either temporal or spatial restrictions are imposed on the gene's action.

From Genes to Circuits

Genetic Mosaics

Before the discovery of *fruitless*, experiments involving flies that are a mixture of male and female cells (called 'mosaics') were shown to behave like males if only small portions of their brain are genetically male, even if the rest of their cells and tissues are genetically female. This made it possible to 'map' the regions of the fly's central nervous system that govern male-specific courtship behavior.

One way of producing mosaic flies is to generate fertilized eggs that start out having two X chromosomes but lose one during the first few rounds of chromosome replication and division (mitosis). The result is a fly that has some of its cells male (1 X) and the rest female (2 Xs), producing a sex mosaic, or gynandromorph. By incorporating visible marker genes into the unstable X chromosome, mosaic individuals can be easily recognized by their patchy expression of the markers. More importantly, male portions of the brain can be seen directly in sectioned tissue at the end of the experiment, after the gynandromorph's behavior has been tested. Since no two mosaics have exactly the same clusters of male and female cells, each one becomes a separate experiment in the subdivision of the brain into one type of sexual development or the other. The uniqueness of each mosaic means that the experimenter has no second chance, and that the consistency of a particular pattern cannot be tested directly.

After examining many gynandromorphs, Jeffrey Hall concluded that the different steps in male courtship behavior were controlled by sex-specific development in certain parts of the central nervous system (Hall, 1977). In particular, the initiation of courtship (orienting toward the female, tapping her abdomen, following, and wing extension) required that only one side or the other of a relatively small region near the top of the fly's brain be male. This finding points to a region called the posterior dorsal supraesophageal ganglion as the site of a trigger mechanism that is present in males but not in females, which can receive sensory information and send out motor commands through other parts of the brain that will function properly whether they are male or female. In fact, most of the brain's circuitry is the same in both sexes.

Later steps in courtship, particularly those requiring precise motor coordination, require male tissue in additional parts of the nervous system. To sing a proper courtship song, for example, a gynandromorph must have male cells in the dorsal brain and parts of the midbrain, as well as in parts of the thoracic ganglion, the fly's version of a spinal cord.

Anatomical Mutants

Some of a fly's repertoire of visual response during flight can be observed in the laboratory where a tethered fruit fly will fly in a stationary position and attempt to orient itself according to the visual patterns it perceives (Figure 3). The fly's efforts at orientation to movement in its visual field constitute its optomotor response. Rapid motor responses to moving patterns in the visual field are widespread among animals with sophisticated visual systems, including humans.

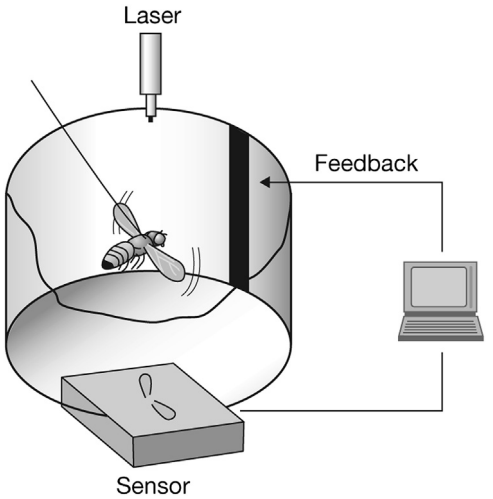

Figure 3 Fly virtual reality. A fruit fly is suspended (from a wire glued to its back) in the middle of a flight arena, consisting of a cylinder of light-emitting diodes, on which images or patterns are displayed. The fly's wing beats are detected by means of an infrared laser that casts the wings' shadows onto a pair of sensors. The differential amplitude of the left versus the right wingbeat is then fed back through a computer to the image display, allowing the fly to control the position of the image and thus exhibit virtual tracking behavior. Reprinted from Greenspan, R.J., 2007. An introduction to nervous systems. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, p. 103, Figure 6.16. Copyright 2007, with permission from Cold Spring Harbor Laboratory Press.

Many mutants in the fruit fly have been isolated that are defective in the optomotor response, and as with phototaxis, most of these were subsequently found to affect components of the phototransduction machinery in the retina. Several, however, affected the behavior more centrally in the brain. One, in particular, was normal for vision at the level of the retina and displayed normal attraction to light, but showed a selective deficit in motion detection. It responded normally to a single stripe in motion, but failed to respond as normal flies would to a grate or a random array of stripes.

The mutant, known as *optomotor-blind*, was subsequently found to affect selectively the development of the visual system in the fly's brain (Heisenberg et al., 1978). It fails to produce a specific subset of large neurons in the most central optic lobes called horizontal-motion-sensitive (HS) and vertical-motion-sensitive (VS) giant fibers (Figure 4). In other words, the behavioral defect it causes is actually a developmental defect in the formation of a very specific part of the visual system. These are neurons known from electrophysiological recordings made in larger flies to be responsive to motion over a wide portion of the field of vision, and they receive widespread inputs from neurons in the more peripheral optic lobes. The behavioral phenotype of the mutant demonstrates directly the critical role of these cells in a restricted aspect of the optomotor response, a role that could only be inferred from their physiological properties.

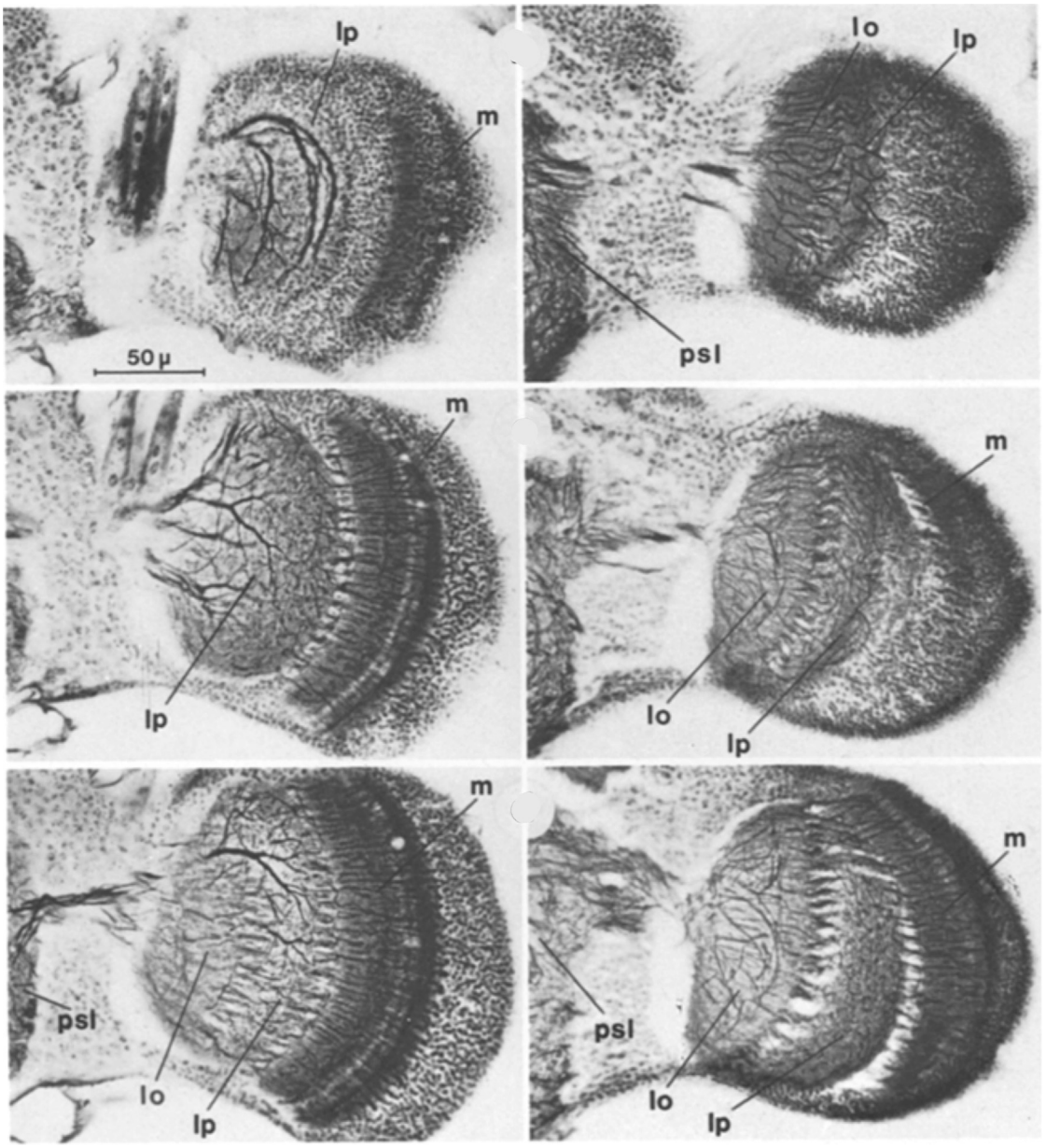

Figure 4 The optic lobes (lobula plate) of *Drosophila* showing horizontal giant neurons (HS cells) in normal (left) and *optomotor-blind* mutant flies (right). Reprinted from Heisenberg, M., Wonneberger, R., Wolf, R., 1978. Optomotor-blind^{H31}—a *Drosophila* mutant of the lobula plate giant neurons. Journal of Comparative Physiology A 124, 287–296. Copyright 1978, with permission from Elsevier. lo: lobula; lp: lobula plate; m: medulla; psl: posterior slope of protocerebrum.

The *optomotor-blind* gene has been cloned, and its DNA sequence determined (Pflugfelder et al., 1992). It turns out to encode a protein that has homology to transcription factors, proteins that regulate gene expression, consistent with a role in development. The gene's pattern of expression, however, indicates that it is more widespread than simply the HS and VS giant neurons, encompassing much of the proximal optic lobes (Poeck et al., 1993). In other words, it is not a gene exclusively for HS and VS neurons. How, then, can the *optomotor-blind* mutation produce such a specific defect?

Further analysis revealed that this is a gene playing multiple roles in development, as reflected by its numerous regulatory regions (control sites in the gene for activating it at specific times and places). Mutations in these various regions produce different developmental defects, some much more severe than the simple loss of HS and VS cells. The original *optomotor-blind* mutation was caused by a disruption of one of the several regions of its DNA sequence – that responsible for HS and VS neuron development. Disruptions of adjacent control regions produce progressively more drastic disruptions of optic lobe development up to a full knockout of the gene, which is lethal and majorly disrupts development of the optic lobes.

Thus, even though this gene has a major role in the development of the entire optic lobes, because of the organization of

its regulatory sequences, it was capable of mutating to a behaviorally specific phenotype. This, in turn, revealed its function in the organization of behavior: the specification of HS and VS giant motion-detecting neurons. Thus, *optomotor-blind* represents a paradigmatic example of a specific behavioral effect of a mutation in a pleiotropic gene.

Localization and Genetic Targeting

A major advance over the gynandromorph technique, described above, came with the introduction into behavioral studies of an alternative approach to generating intersexual mosaics, making use of gene targeting. This technique, known as the *GAL4* method (Ferveur et al., 1995), permits concerted control of expression of a transgene in specific cells of the brain. (*GAL4* is a transcription factor from yeast that was adapted for the purpose of expressing transgenes in restricted tissues in *Drosophila*.) The experiments on courtship relied on the ability of the female form of the sex determination gene transformer (*tra*) to act dominantly and autonomously in males to drive cells into the female pathway of development. In contrast to the gynandromorph method, where each individual is unique in its pattern of mosaicism, the *GAL4* method produces an essentially unlimited supply of individuals with ostensibly identical patterns of mosaicism. It also produces a finer grained mosaicism than gynandromorphs, and with it the potential for higher resolution mapping.

The *GAL4* method demonstrated that feminization of restricted sets of neurons in the male fly brain was sufficient to induce those flies to court other males and females (Ferveur et al., 1995), as well as to identify sites in many other parts of the nervous system whose sex-specific development was necessary for courtship (Ferveur and Greenspan, 1998). Ultimately, most of these sites, but not all, turned out to be sites of *fruitless* gene expression. The technique is widely used in studies of behavioral circuitry.

The 'Connectome' and Behavioral Circuitry

In recent years, the idea of producing a complete map of neuronal connections in the brain, known as the 'connectome' (Sporns et al., 2005), has generated broad interest. The first such case of a complete map was produced for the simple ~300-neuron nervous system of the little nematode *Caenorhabditis elegans* (White et al., 1986), and the goal of completing a human connectome map has been variously estimated to take anywhere from 100 to 7000 years. Major efforts in this direction are currently under way (The open connectome project <http://openconnecto.me/>).

Fortunately, the connectome of the 130 000-neuron *Drosophila* brain is well on its way to completion in a more livable timeframe, due to the efforts of Ann-Shyn Chiang in Taiwan (Chiang et al., 2011). The first draft, of roughly 10% of the brain, was published in 2011, and the remainder is expected to be completed within a few years. A summary diagram of the major connectivity pattern among brain modules (known as 'local processing units') is shown in Figure 5. The work represents a major milestone as the first connectome map of a complex brain.

The existence of this map and its accompanying database with a record for each neuron (FlyCircuit <http://www.flycircuit.tw/>) has revolutionized the study of the circuits underlying behavior. Previous efforts, using gynandromorphs or gene targeting (see above), all involved anatomical verification of the pattern of gene expression in the brain, but this required either serial tissue sections or whole mount staining, both of which introduced substantial distortion. Consequently, prior to the current connectome map, the only individually

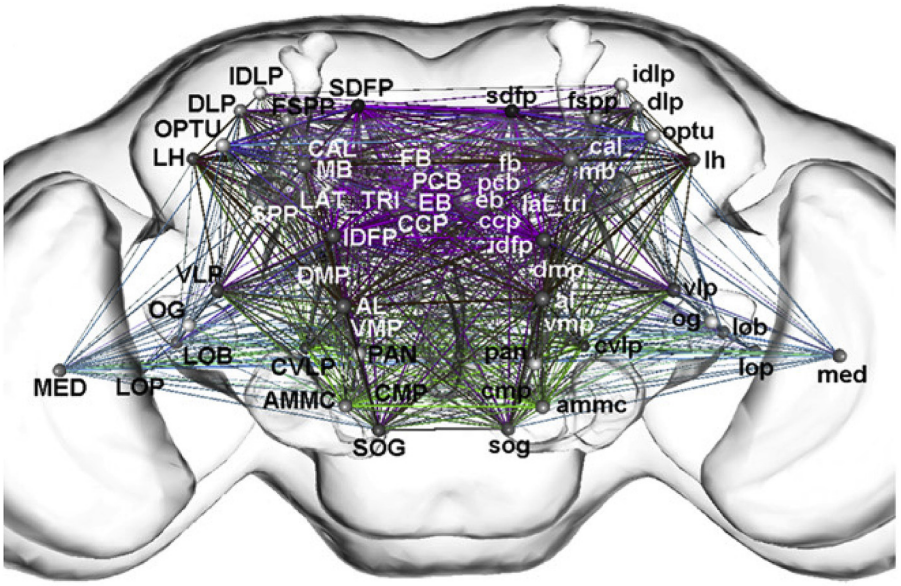

Figure 5 Major connectivity patterns of network connection in the *Drosophila* brain. Reprinted from Chiang, A.S, Lin, C.Y., Chuang, C.C., et al., Three-dimensional reconstruction of brainwide wiring networks in *Drosophila* at single cell resolution, Current Biology 21, 1–21. Copyright 2011, with permission from Elsevier. For abbreviations and explanation, see Chiang et al. (2011).

identifiable cells in the fly brain were the handful that were large enough to have an easily recognizable morphology, such as the HS and VS neurons described above.

These problems have all been obviated now, due to improved techniques of sample preparation, clearing, and staining that introduce little or no distortion. The result is the tagging of individual neurons in a coordinate system that gives each neuron a recognizable identity from fly to fly. Starting from the technology of targeted gene expression (see above), this new resource not only makes possible the reliable mapping of functional tests of individual cells onto the reference map but also allows for the rapid and accurate prediction of other connected neurons and the reconstruction of a hypothetical and testable complete circuit. Its potential is already being realized.

A logical corollary to the full connectome map is to have a comprehensive brain activity map (Alivisatos et al., 2012), a project that is now being mounted. It is currently most feasible in small-brained organisms such as *C. elegans* and *Drosophila*, but the development of new technologies to allow access to larger brains is also under way.

From Component Parts to Networks

Networks of Genes

The search for genes has been a major concern of geneticists since variants first began to be mapped – the more the better for filling in spaces along chromosomes, regardless of phenotype. The quest has become more directed in the last several decades of developmental and behavioral genetics. The guiding assumption has been that mutants reveal the key players and they will explain the underlying mechanisms. There is no doubt that this approach has been successful in expanding and shaping our view of when, where, and how genes act.

Central to this worldview is the genetic pathway, originally modeled in conjunction with the enzymatic transformations of intermediary metabolism. The extension of the analogy beyond metabolism made its debut in phage morphogenesis where it represented the sequential assembly of a phage particle. In more contemporary studies, it has been further adapted to represent several kinds of processes including pattern formation and cell fate determination. It has even been stretched to apply to morphogenetic events and, in behavioral studies, to phases in memory consolidation.

The aspiration of this approach, based again on the phage morphogenesis work, is saturation mutagenesis for mutants 'specifically' affecting the phenotype in question. Sadly, this may have been applicable only to phage morphogenesis. We now know that the mutant search for early embryonic patterning missed many critical genes despite the fact that it met the criterion for saturation. This was due not only to the significant role of maternal genes in the formation of the embryonic axes but also, more importantly for this discussion, to the fact that many of the relevant loci did not (and presumably could not) mutate to a phenotype restricted enough to be picked out. Eventually, mutants in this more refractory class of genes were found by more indirect strategies (e.g., looking mutations based on interaction with known mutants, or based on tissue-specific patterns of expression).

The ensuing proliferation of identified genes has created a problem of interpretation, e.g., which genes are most important, and how do they all interconnect? Rather than running in linear paths, the increasing complexity of relationships among genes is better described as a distributed network. Some produce more damage than others when mutated, but this depends heavily on the context of other variants that are present, and so it is difficult to arrange them in a simple order of importance. The interactions that have defined various pathways are not wrong, just not exhaustive. They are part of a much larger picture.

Pleiotropy, as discussed above, has been the hallmark of behavioral mutant hunts all along, and the majority of behavioral mutants in the fly are the result of rare phenotypically specific alleles of genes that act more widely. Behavioral mutants are also extremely sensitive to variations in genetic background – the natural, genetic heterogeneity in laboratory stocks. The dependence of mutant phenotype on strain background has been well documented for learning mutants in the mouse (Gerlai, 1996), and the sensitivity of behavior to genetic background is a special case of the more general observation that mild genetic differences can produce large phenotypic effects on behavior (Greenspan, 2001).

These conclusions are also consistent with the findings on genetic changes that occur during multigenerational selection for behavioral phenotypes in the laboratory. There, a wide range of genes has been found to be capable of affecting phenotypes, and the variants are demonstrably mild and acting in combinations (Greenspan, 2009). These conclusions were made possible in large measure by the advent of genomic technologies. Knowing all the organisms' genes and their sequences has spawned technologies for comprehensively measuring the level of expression of every gene across the entire genome. Thus, by comparing genome-wide expression levels between strains differentially selected for a behavior, it became possible to infer a gene's involvement, and then to confirm it by direct testing of a mutant in that gene.

Such extreme sensitivity to subtle genetic variation, and the ubiquitous presence of the requisite variation, argues strongly against any sort of linear pathway model for the action of genes on behavior. It is unlikely that the variation needed to cause 'background effects' would be found so often if only a narrow set of dedicated genes was capable of interacting. This network of genes, in turn, must operate through the networks of cellular machinery and anatomy in the nervous system to influence behavior. The filtering through successive networks provides greater potential for interactivity and synergy.

Networks of Neurons

As judged by engineering standards, nervous systems are noisy. The signals in neurons are often not reproducible. Synaptic release can vary tremendously depending on the strength of stimuli to a cell and also depending on the cell's immediate past history. These effects are magnified many times over by the fact that neurons are highly interconnected in a dense network. In the nematode *C. elegans*, it has been possible to estimate the degree of connectedness. Theoretically, a sensory stimulus to four of the animal's sensory neurons would reach 6% of the animal's ~300 neurons directly in one step (i.e., through their

first synaptic connections), and 50% of the nervous system in two steps. That is in theory, but what about reality?

Using a technology that allows hundreds of neurons to be monitored for their electrical activity at the same time, rather than the two or three usually monitored in conventional electrophysiology, one of the simplest of behavioral circuits showed signs of being far less simple. The behavior is the reflexive gill withdrawal response of the sea snail *Aplysia*, which, in the hands of Nobel Laureate Eric Kandel, became a model system for understanding the cellular and molecular mechanisms of synaptic plasticity. *Aplysia* will close up the gills when touched. Kandel and colleagues showed that this is a simple sensory-motor reflex, in which the touching of the gill activates sensory neurons that synapse directly onto motor neurons that drive the muscles to close the gill (Pinsker et al., 1970).

When this response was monitored under the microscope, with a voltage-sensitive dye to indicate the electrical activity of the entire abdominal ganglion, it became possible to measure the reproducibility of the population of neurons active in the gill withdrawal response during repeated stimuli. There was significant variation in the number of cells, the identity of cells, and the strength of response of cells from one trial to the next (Wu et al., 1994). The number of cells responding is in the hundreds.

These *Aplysia* findings reinforce the theoretical prediction obtained in the nematode *C. elegans* that it is almost impossible to insulate the activities of neurons in one region from those in another. This means that some other strategy, aside from insulation, must be exerting itself. If the intrinsic variability in the nervous system cannot be reined in, then perhaps it is exploited.

In reality, nervous systems are potentially chaotic places. If one estimates the actual extent of connectivity in a system such as *Aplysia*'s gill withdrawal response, the potentially relevant circuitry is huge and tangled, with many hundreds or perhaps thousands of potential pathways intervening between sensory neurons and motor neurons. The response is distributed widely in the nervous system. Granting that not all such pathways are necessarily productive or consequential in their output, if they all exist and have some probability of occurring, and if many of them produce at least some degree of consequential output, then we are faced with the question of how the animal sorts out the useful from the wasteful.

General Conclusions

Our understanding of the genetic influences on behavior has grown substantially in the 50 years since the question began to be directly addressed. During that time, the perspective has evolved from that of single genes to many genes acting in networks, and from rough localization of function in the nervous system to a fine-grained coordinate system, onto which behavioral circuits are being mapped. As so often happens in science, the advances have brought with them new challenges, which are becoming focused on the question of how the large-scale, dynamic interactions of genes acting through nervous systems can be addressed. Arriving at a solution to this complex problem may then actually bring us close to a deep understanding of how genes affect behavior.

See also: Behavioral Genetics: Psychological Perspectives; Intelligence, Genetics of: Cognitive Abilities; Sexual Orientation: Genetic Aspects.

Bibliography

Alivisatos, A.P., Chun, M., Church, G.M., et al., 2012. The brain activity map and the challenge of functional connectomics. Neuron, 74, 970–974.

Chiang, A.S., Lin, C.Y., Chuang, C.C., et al., 2011. Three-dimensional reconstruction of brainwide wiring networks in *Drosophila* at single cell resolution. Current Biology 21, 1–11.

Ferveur, J.F., Greenspan, R.J., 1998. Courtship behavior of brain mosaics in *Drosophila*. Journal of Neurogenetics 12, 205–226.

Ferveur, J.-F., Störtkuhl, K., Stocker, R.F., Greenspan, R.J., 1995. Genetic feminization of brain structures and changed sexual orientation in male *Drosophila melanogaster*. Science 267, 902–905.

Gerlai, R., 1996. Gene-targeting studies of mammalian behavior, is it the mutation or the background genotype? Trends in Neuroscience 19, 177–181.

Greenspan, R.J., 2001. The flexible genome. Nature Reviews Genetics 2, 383–387.

Greenspan, R.J., 2009. Selection, gene interaction, and flexible gene networks. Cold Spring Harbor Symposium on Quantitative Biology 74, 131–138.

Hall, J.C., 1977. Portions of the central nervous system controlling reproductive behavior in *Drosophila melanogaster*. Behavior Genetics 7, 291–312.

Heisenberg, M., Wonneberger, R., Wolf, R., 1978. Optomotor-blind^{H31}—a *Drosophila* mutant of the lobula plate giant neurons. Journal of Comparative Physiology A 124, 287–296.

Hirsch, J., 1959. Studies in experimental behavior genetics. Journal of Comparative Physiology 52, 304–308.

Pflugfelder, G.O., Roth, H., Poeck, B., et al., 1992. The lethal(1)optomotor-blind gene of *Drosophila melanogaster* is a major organizer of optic lobe development, isolation and characterization of the gene. Proceedings of the National Academy of Sciences of the United States of America 89, 1199–1203.

Pinsker, H., Kupfermann, I., Castellucci, V., Kandel, E., 1970. Habituation and dishabituation of the gill-withdrawal reflex in *Aplysia*. Science 167, 1740–1742.

Poeck, B., Hofbauer, A., Pflugfelder, G.O., 1993. Expression of the *Drosophila optomotor-blind* gene transcript in neuronal and glial cells of the developing nervous system. Development 117, 1017–1029.

Ryner, L.C., Goodwin, S.F., Castrillon, D.H., et al., 1996. Control of male sexual behavior and sexual orientation in *Drosophila* by the *fruitless* gene. Cell 87, 1079–1089.

Sporns, O., Tononi, G., Kötter, R., 2005. The human connectome: a structural description of the human brain. PLoS Computational Biology 1, e42.

Stockinger, P., Kvitsiani, D., Rotkopf, S., Tirián, L., Dickson, B.J., 2005. Neural circuitry that governs *Drosophila* male courtship behavior. Cell 121, 795–807.

White, J.G., Southgate, E., Thomson, J.N., Brenner, S., 1986. The structure of the nervous system of the nematode *Caenorhabditis elegans*. Philosophical Transactions of the Royal Society B, Biological Sciences 314, 1–340.

Wu, J.-Y., Tsau, Y., Hopp, H.-P., et al., 1994. Consistency in nervous systems: trial-to-trial and animal-to-animal variations in the responses to repeated applications of a sensory stimulus in *Aplysia*. Journal of Neuroscience 14, 1366–1384.

Yin, J.C., Wallach, J.S., Del Vecchio, M., et al., 1994. Induction of a dominant negative CREB transgene specifically blocks long-term memory in *Drosophila*. Cell 79, 49–58.

Relevant Websites

<http://openconnecto.me/> – The open connectome project.
<http://www.flycircuit.tw/> – FlyCircuit.

Genetic Admixture

Oscar Lao and Mannis van Oven, Erasmus MC – University Medical Center Rotterdam, Rotterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Genetic admixture plays a fundamental role in a species such as *Homo sapiens*, which has colonized the world on a relatively small evolutionary time scale, and has gone through complex migratory patterns in modern times. Several statistical methods exploiting different aspects of the genetic variation of admixed populations have been proposed. In this article, we describe the main consequences of genetic admixture for genetic variability, both in autosomal and paternally and maternally linked markers, introduce the state-of-the-art in methods for studying genetic admixture/genetic ancestry, and provide two key examples of genetic admixture in human populations: the human diaspora out of the African continent and the patterns of admixture in the American continent following Columbus' discovery.

Genetic admixture and genetic differentiation are two mutually dependent processes that have played a main role in shaping the neutral (not being under selective pressure) genetic variability of the human species. On the one hand, genetic admixture occurs when individuals from two or more different genetic backgrounds (or ancestries) migrate, mate, and produce descendants. On the other hand, a differentiated population genetic background is due to prolonged genetic isolation among populations, as a result of the absence of migration and genetic admixture between them. Since new populations appear by splitting from previous ones, genetic ancestry (and hence genetic admixture) has also a temporal dimension: one can define and infer genetic ancestry at different time scales, taking into account that all individuals from a species ultimately share a common genetic origin.

Detecting the genetic fingerprint of admixture and isolation processes in human populations is of main interest for human population geneticists, and the development of new methods for detecting such events has been constant in the literature (Verdu and Rosenberg, 2011), since 1931 (Bernstein, 1931). These methods, in addition to the availability of hundreds of thousands of single nucleotide polymorphisms (SNPs) genotyped in thousands of worldwide individuals (Novembre and Ramachandran, 2011), are providing a new and very detailed picture of the genetic landscape of human populations, as well as the underlying historical processes that shaped them (Gutenkunst et al., 2009).

The two basic models mostly used in human population genetics to describe genetic admixture events are depicted in Figure 1 (adapted from Long, 1991). Both models share the *sine qua non* condition of the existence of genetic differentiation between two parental populations (although by definition these parental populations ultimately share a common origin) that is acquired by means of long-term genetic isolation after the split from the ancestral population. Moreover, the number of breeding chromosomes (understanding that each diploid individual carries two chromosomes) in each generation and population (also known as effective population size or N_e, Hartl and Clark, 1997) is assumed to remain constant through time. In the *continuous-gene-flow* (*CGF*) model, *m* chromosomes from one population migrate to the other population since *t_admixture* generations ago, and replace the same proportion of chromosomes in the mating process. In the *hybrid-isolation* (*HI*) model, two parental populations create a new hybrid population in a single pulse event of admixture *t_admixture* generations ago, which then starts evolving independently from the parental populations. More complex models can be constructed from these basic models by considering different effective population sizes (Bertorelle and Excoffier, 1998), allowing successive migrations from the parental populations to the hybrid one (Verdu and Rosenberg, 2011), or considering more than two parental populations (Dupanloup and Bertorelle, 2001), among others.

A basic question of interest for population geneticists is the timing of admixture (*t_admixture*), the proportions of migrants (*m*) that were exchanged between populations and the number of parental ancestral populations that contributed to the current genetic variability. Furthermore, it can be important to determine the ancestry proportions in the genome of each individual or to detect fragments of genomic ancestry. Depending on the question of interest, the statistical methods focus on exploiting the genetic variability of the parental and admixed populations, the amount of individual ancestry or the genomic ancestry at a regional level. Depending on whether we are interested in studying sex-mating biases between the parental populations, or average migration patterns, current statistical methods use uniparentally inherited markers or autosomal markers, respectively, or a combination of both. We next describe some of these properties and the main statistical tools so far developed that take advantage of them.

Autosomal Genetic Admixture at the Population Level

Linkage Disequilibrium in the genome of the admixed population. Linkage disequilibrium (LD) refers to the nonrandom association between the allelic frequencies of two markers in a population (Slatkin, 2008 for more details). Recently admixed populations tend to show in their genome an overall excess in the levels of LD (Pfaff et al., 2001). The levels of LD decay exponentially since the time of admixture (Figure 2(a)) and therefore can be used to make inferences about this parameter (Patterson et al., 2012; Pfaff et al., 2001).

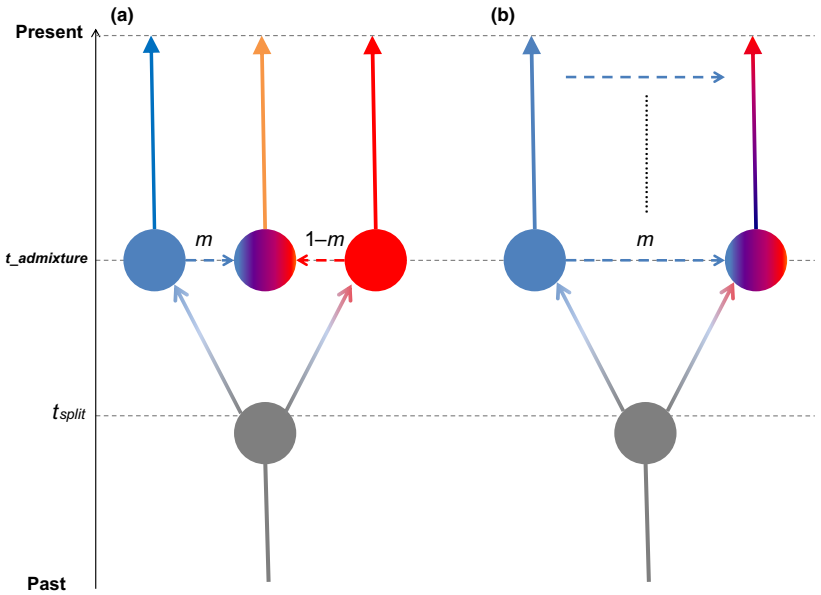

Figure 1 Basic admixture models commonly used in population genetics. Each circle represents a population. Both models consider one initial ancestral population (gray color) that splits into two new populations t_{split} generations ago. Each of the new populations evolves without exchanging migrants for a period of time, during which genetic differentiation between them can take place. (a) *HI* model. At *t_admixture* there is a single event of admixture, and a new hybrid population is created from *m* fraction of chromosome migrants from the blue population and $1-m$ fraction of migrants from the red population. After this event, each population continues to evolve independently; (b) *CGF* model. The blue population contributes *m* chromosome migrants to the red population, replacing the same number of chromosomes from this population. Adapted from Long, J.C., 1991. The genetic structure of admixed populations. Genetics 127 (2), 417–428.

Genetic similarity of the admixed population with the parental ones. Initially, admixed populations have allele frequencies that are linear combinations of the allele frequencies in their parental populations. With time after the admixture event, the allelic frequencies in the parental and admixed populations randomly fluctuate due to genetic drift (the change in allele frequency in a population due to random or chance events rather than natural selection), and genetic differentiation between them occurs. Nevertheless, the expectation of linearity still holds if a large number of unlinked loci are considered. Therefore, the degree of genetic similarity between the admixed and the parental populations can be used to make estimations of the migration rate in the admixed population (Long, 1991; Patterson et al., 2012; Pickrell and Pritchard, 2012) or to statistically compare particular population phylogenies (Patterson et al., 2012; Pickrell and Pritchard, 2012). A traditional statistic based on allelic frequencies summarizing the amount of genetic differentiation among populations is F_{st} (Holsinger and Weir, 2009). This statistic is proportional to the time of admixture and population split, although it is highly dependent on the demographic assumptions of the model (Verdu and Rosenberg, 2011), and does not consider other aspects of the genetic variants, such as the number of evolutionary steps separating different alleles, which can provide additional information about the admixture process. The *my* statistic (Bertorelle and Excoffier, 1998) and its extension to any number of parental populations (Dupanloup and Bertorelle, 2001), or maximum likelihood approaches such as the one implemented in the program LEADMIX (Wang, 2003) or MEADMIX (Wang, 2006) use the coalescence framework to estimate the different admixture parameters by taking into account this fact.

Autosomal Genetic Admixture at the Individual Level

Population methods for quantifying admixture require prior knowledge and a definition of what a population is, which in some cases can be particularly difficult (if not impossible) to obtain (Pritchard et al., 2000). Therefore, several methods for estimating individual ancestry proportions without taking into account any population sampling information have been proposed. Depending on the applied statistical strategy, these methods have been divided into *global ancestry-based methods* and *local ancestry-based methods* (Alexander et al., 2009).

Global ancestry-based methods are applied to estimate the number of the ancestral populations that shaped the observed current genetic diversity, infer their ancestral genomic composition, and quantify their contribution in the genome of each current individual (Figure 3). Depending on whether the statistical method takes into account the expected allelic frequencies of the ancestral populations or not, global ancestry-based algorithms have been subdivided into *model-based ancestry estimation* (Alexander et al., 2009) and *algorithmic ancestry estimation* (Alexander et al., 2009) (Figure 3). In the first case, the main aim is to infer the genetic variation of the ancestral populations, and their contribution to the genome of each current individual. Popular algorithms implementing such approach include STRUCTURE (Pritchard et al., 2000), FRAPPE (Tang et al., 2005), and ADMIXTURE (Alexander

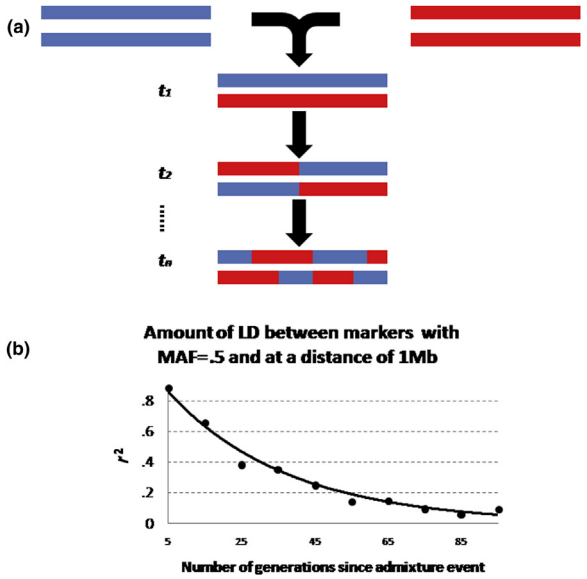

Figure 2 Effect of genetic ancestry at a local level. (a) Chromosomes of parental population 1 (with blue ancestry) and parental population 2 (red ancestry) admix into a new hybrid population. As time since admixture passes, chromosomes from the different ancestries collide in the same individual and recombine, creating a mosaic of ancestries (t_n). Adapted from Pugach, I., et al., 2011. Dating the age of admixture via wavelet transform analysis of genome-wide data. Genome Biology 12 (2), R19; (b) decay of the patterns of linkage disequilibrium (LD) in a simulated admixed population with the time since the admixture event (in number of generations). LD was estimated by means of r^2 between markers with a MAF of 0.5 and at 1 Mb distance. The effective population size (N_e) of each parental and admixed population was 2000 diploid individuals (i.e., 4000 chromosomes). Random monogamous mating was assumed. The admixed population was created for each simulation by admixing 1000 individuals from parental population 1 and the same number of individuals from parental population 2. Both parental populations merged into a single one 5000 generations ago. Eighty diploid autosomal genomes were sampled in each simulation; each genome comprised 22 pairs of autosomal chromosomes, each one proportional in length to the human chromosomes (Gravel, 2012). The mutation rate was set to 2.5×10^{-8} and the recombination rate to 1.8×10^{-8}. The demographic and genomic parameters used to generate the simulations were chosen to be illustrative of the different genomic processes, and are not intended to be representatives of human population history.

et al., 2009), among others. In the second case, the output is a set of coordinates for each individual in a new reduced Euclidean space (Jombart et al., 2009; Figure 3), in which each dimension can be interpreted as a locus with an infinite number of alleles (Menozzi et al., 1978), and the proximity between two individuals in this space as genetic similarity between them (Patterson et al., 2006). Principal component analysis (PCA; Patterson et al., 2006) and classical multidimensional scaling (classical MDS; Cox and Cox, 2001) are typical examples of algorithmic ancestry estimation, mainly applied to SNP data.

Besides inferring individual ancestry proportions and the number of ancestral populations, these methods can be used to get estimations of the demographic parameters associated to the admixture events. For example, ancestry proportions in older admixture events will tend to be homogeneously distributed among the individuals of the admixed population; in contrast, relatively recent admixtures are going to produce individuals carrying large proportions of the ancestries, thus increasing the ancestry variance among them (Figure 4). Therefore, the variance in the proportions of genomic ancestry between individuals of the admixed population can be used to make inferences about the magnitude, timing, and frequency of the admixture event (Verdu and Rosenberg, 2011). Similarly, it has been shown that the coordinates in the first dimension estimated by PCA tend to be proportional to the amount of genomic ancestries in the admixed individuals (McVean, 2009) (Figure 4). However, since global ancestry-based methods do not explicitly fit historical models, or unrealistically assume that all the ancestral populations have radiated from a single ancestral group (Figure 3), it has been suggested that the biological interpretation of their outcome can be cumbersome (Patterson et al., 2012). Nevertheless, despite these limitations, the methods for inferring genomic ancestry at individual levels are fundamental tools in interpreting the current genetic variability in human population genetics (see below).

Local ancestry methods. As already stated before, after an event of admixture between individuals of different genetic ancestry, recombination tends to break the chromosomes into blocks of genomic ancestry (Figure 2(a)); at a population level this is translated into patterns of LD; at an individual level, this process means that the genome of one individual can be considered as a mosaic of different ancestries (Figure 2(a)). *Local ancestry methods* (Alexander et al., 2009) try to estimate the ancestry of each locus in a given individual. If the genomic marker density is large enough, such as in the case of SNP microarrays (Novembre and Ramachandran, 2011), local ancestry methods will define blocks of genomic ancestry, also referred to as 'ancestry painting.' So far, several algorithms based on different statistical approaches and assumptions have been proposed for identifying such ancestry tracks (Pugach et al., 2011, for a list of algorithms). Most of these require previous knowledge of the genetic variation of the parental populations or, if such information is not available, of proxy parental populations (i.e., HapMix and StepPCO algorithms (Price et al., 2009; Pugach et al., 2011); Figure 3). Similarly to *global ancestry-based methods*, the output of local ancestry methods can provide insights into the values of the demographic processes at a population level because large ancestry fragments in the admixed population are indicative of recent and ongoing migration processes, whereas small ancestry fragments are suggestive of older migratory events (Gravel, 2012; Figure 3).

Sex-Biased Admixture

In contrast to the autosomes, there are two genetic loci that are haploid and do not undergo recombination: the maternally inherited mitochondrial (mt) genome (carried by males and females), and the paternally inherited Y chromosome (that is to say, the nonrecombining portion of it, referred to as NRY) (carried by males only). Despite differences in their size, genetic organization, mutation rate, and cellular localization (Jobling

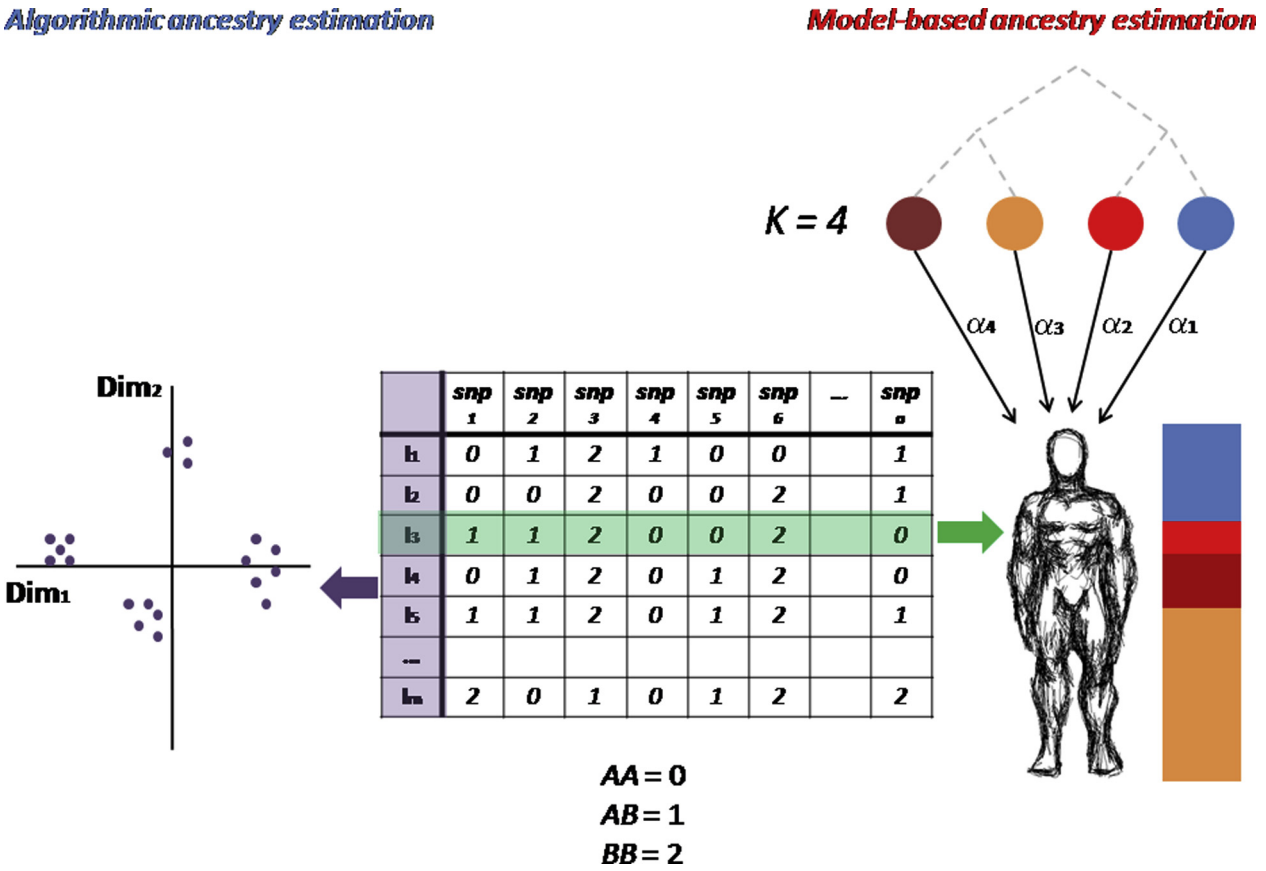

Figure 3 Schematic representation of the *Global ancestry-based methods* to detect individual genetic ancestry. Starting from a matrix of genotypes (coded 0, 1, and 2 for homozygotes for allele A, heterozygotes and homozygotes for allele B, respectively) of *m* individuals by *n* biallelic markers, the algorithms are classified into two main types. *Algorithmic ancestry estimation* methods apply algebraic transformations to obtain a new representation of the genomic coordinates of each individual (each purple dot) in a reduced space, so the genetic relationships between the individuals can be better identified. *Model-based ancestry estimation* methods aim for estimating the number of ancestral populations (in the current example: four) that shaped the genetic variation of the matrix and the ancestry proportions (α_i) for each individual, which can then be visualized by means of a bar. The underlying demographic history of the ancestral populations is shaded, since it is not modeled.

and Tyler-Smith, 2003; Pakendorf and Stoneking, 2005), mtDNA and NRY have in common their unique uniparental inheritance pattern. The process of hundreds of thousands of years of human evolution has yielded DNA variation in both NRY and mtDNA as a result of the accumulation of spontaneous mutations along radiating lineages. All haplotypes existing today are ultimately related genealogically and their relationships can be represented by detailed phylogenetic trees, which have been constructed for both NRY (Karafet et al., 2008; International Society of Genetic Genealogy, (http://www.isogg.org/tree/)) and mtDNA (Torroni et al., 2006; van Oven and Kayser, 2009). Shared (sets of) mutations between two haplotypes are indicative of a shared uniparental ancestor at some point of time in history. Such mutations can thus be used as evolutionary markers that define clusters of related haploid genetic lineages, or haplotypes (haplogroups). Each haplogroup has its own specific evolutionary history and associated geographic distribution, shaped by different demographic processes (Cavalli-Sforza and Feldman, 2003; Underhill and Kivisild, 2007). As a result of their mode of inheritance, NRY and mtDNA reflect very specific aspects of a person's ancestry, namely that of the father's father's father, etc. and that of the mother's mother's mother, etc., respectively (Figure 5; Shriver and Kittles, 2004). It is important to note that uniparental ancestry does not necessarily need to agree with an individual's overall ancestry (King et al., 2007). In fact, uniparental markers can provide additional insights into the evolutionary history of human populations, such as the mating process that gave rise to admixed populations, which cannot be inferred by other means.

The Complexity of Human Demographic History

The demographic history of *Homo sapiens* is far more complex than the models depicted in Figure 1. Although from an evolutionary point of view, anatomically modern humans can be considered relative newcomers, the species has managed to colonize the whole world after dispersing out from the African continent (Cavalli-Sforza et al., 1994; Liu et al., 2006). During this first main migration, anatomically modern humans

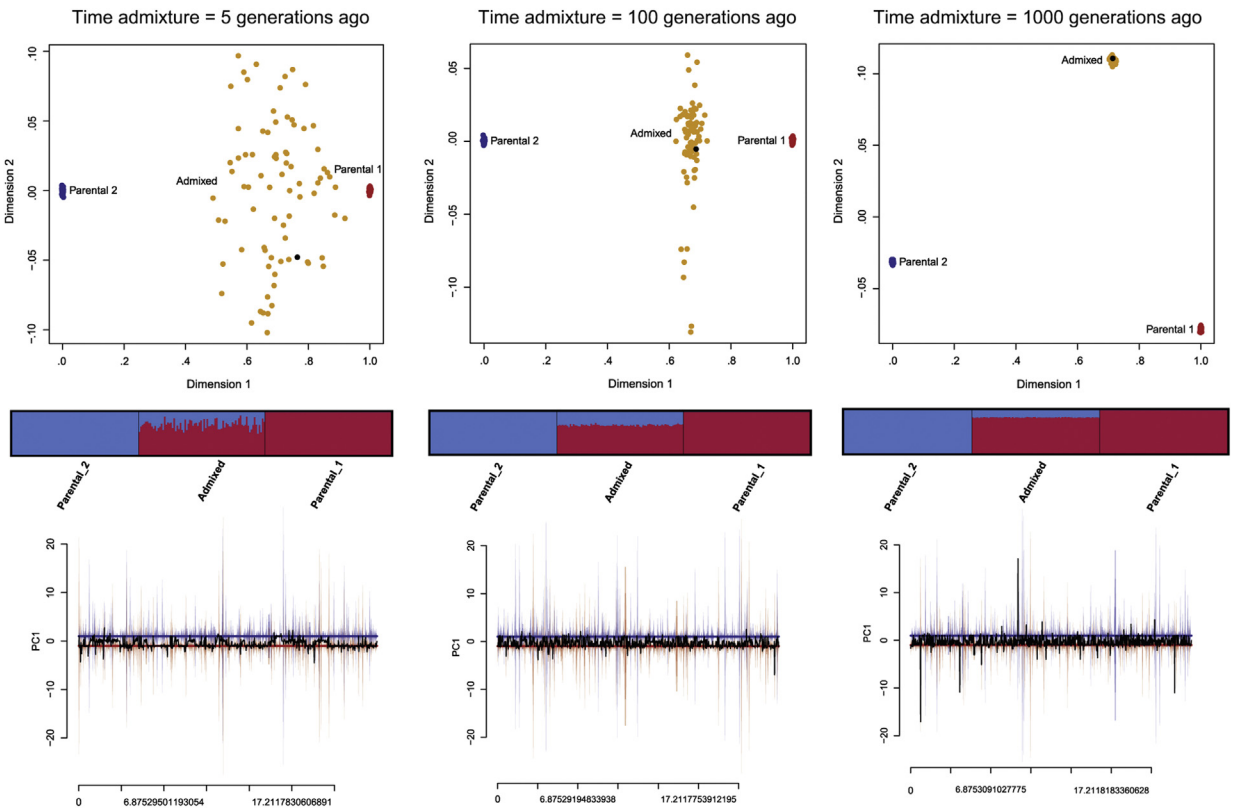

Figure 4 Three methods estimating individual ancestry on simulated data using the *HI* model at three different times of admixture (5 generations ago, 10 generations ago, and 1000 generations ago). The demographic parameters used to generate each simulated dataset is the same as in **Figure 2**, with the exception that in this case, 70% of migrants come from parental population 1 and the remaining 30% from parental population 2. Each dataset comprised 80 sampled individuals from each population. First row: classical MDS on identical-by-state (IBS) distance matrix between pairs of individuals. Each dot represents one individual. Red dots parental population 1, blue dots represent parental population 2, and orange dots represent the admixed population. One of the admixed individuals is marked as black since it is also used in StepPCO analyses. The coordinates of the first dimension have been rescaled so the mean value for parental population 2 is 0 and for parental population it is 1. As can be observed, the admixed individuals tend to be distributed around 0.7 in the first coordinate for times of admixture <100 generations, mimicking the initial proportion of migrants from the parental population 1. Second row: ADMIXTURE analyses on the same data assuming two ancestral populations. Each individual is represented by a bar (**Figure 2**), and individuals from the same population situated contiguous into the same block. The red component, indicating parental 1 ancestry, shows a mean of 0.7 in the admixed population for times <100 generations, and departs from this value as the amount of genetic drift increases (in the example, ∼0.8 at 1000 generations of split). Note also that the variation of the red ancestral component between admixed individuals decreases as the time since admixture increases. Third row: StepPCO analyses on one admixed individual (black dot in the MDS plots, the first admixed individual in the ADMIXTURE analyses) to detect tracks of ancestry. The plot shows the simulated genome of the individual in the *X*-axis and the ancestry values estimated by means of principal component analysis in the *Y*-axis for parental population 1 (red line), parental population 2 (blue line), and the individual of interest (black line). At a time of admixture of five generations ago, the black line tends to be on the red line (indicating long tracks of parental 1 ancestry). As the time since the admixture event increases, the black line starts alternating more frequently between the two parental populations, indicating short genomic fragments of ancestry.

encountered other *Homo* species (such as *Homo neanderthalensis* in Europe) and had the possibility to interbreed with them (Abi-Rached et al., 2011; Green et al., 2010). Following this first main migratory wave, other ancient migrations, mainly associated to technological developments such as farming, took place at a more regional level (Diamond and Bellwood, 2003). Whether these technological migrations were accompanied by demographic movements, admixture or replacement with the autochthonous populations is sometimes difficult to decipher (e.g., see the discussion in Pinhasi et al., 2012 in the case of the European continent). Moreover, the fact that subsequent migrations may have followed geographic patterns similar to the previous ones adds another layer of complexity for disentangling the effect of each historical process in the genetic variation of the genome (Jobling et al., 2004). Furthermore, humans tend to mate predominantly to individuals from the same (or close by) geographic area, which creates an effect of isolation by distance in the amount of genetic differentiation (Ramachandran et al., 2005). In addition, other factors, including language and cultural differences, can also influence the amount of genetic exchange among unrelated individuals (Oota et al., 2001).

One of the first main consequences of the demographic history of humans is a considerable lack of genetic differentiation within the species compared to that observed within other nonhuman primate species including our closest living relative the chimpanzee (Foley, 1998). A practical result of this

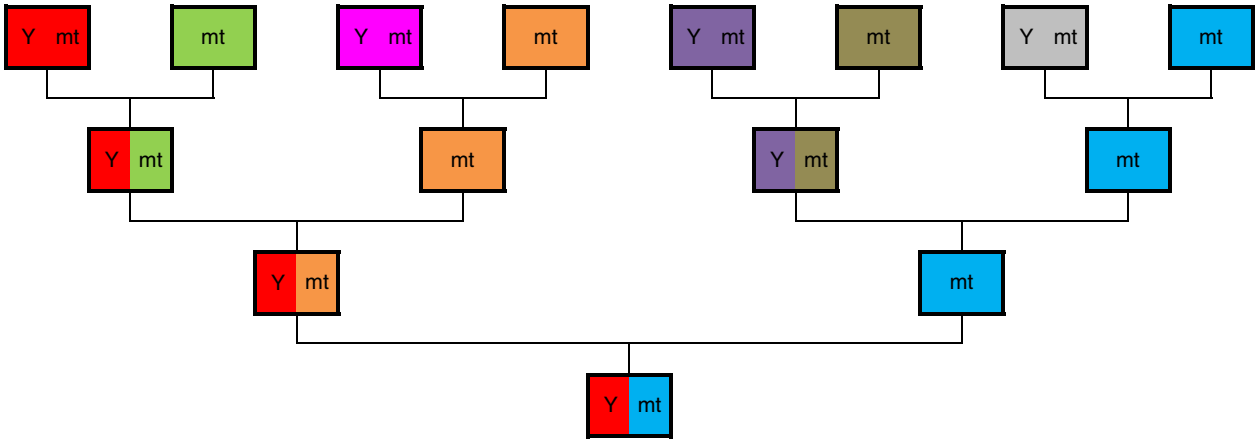

Figure 5 Inheritance patterns of NRY (Y) and mtDNA (mt). The eight different colors represent eight different genetic ancestries. Considering the four generations depicted, the individual at the bottom (a male) inherited his NRY type ('red') exclusively from his patrilineal great-grandfather (upper left) and his mtDNA type ('blue') exclusively from his matrilineal great-grandmother (upper right), whereas his autosomes (not shown) consist of fractions derived from all of his eight great-grandparents.

observation is that detecting population substructure and patterns of admixture in humans requires the use of powerful statistical tools such as these introduced in previous sections. A second consequence of the evolutionary history of *H. sapiens* is that its genetic variation follows clinal patterns in the space rather than sharp discontinuities (Handley et al., 2007), and genetic differences between individuals tend to be strongly correlated with sampling geographic distances (Wang et al., 2012; Yang et al., 2012). Admixture estimations tend to depend on which algorithm is applied (Barbujani and Belle, 2006), the type (i.e., short-tandem-repeat polymorphisms or SNPs) and number of markers (Wang et al., 2012), the considered populations (Xing et al., 2010) and the sample size of the populations (McVean, 2009). Taken altogether, it has been suggested that human genetic variation challenges the concept of 'population' in humans (Barbujani and Colonna, 2010).

1. Detecting ancient human migrations and admixture

Despite these caveats, the statistical methods previously introduced are substantially improving our knowledge about the past of the species. Figure 6 shows the application of three methods for studying the genetic ancestry and admixture in 53 worldwide populations from the Human Genome Diversity Project–Centre d'Étude du Polymorphisme Humain (HGDP-CEPH) panel commonly used in human population genetics (Cavalli-Sforza, 2005); see map in Figure 6(a), using the genotypic information of ~650 000 SNPs (Li et al., 2008).

The first analysis (Figure 6(a)) uses F_{st} distances between pairs of populations to infer a phylogenetic tree by means of the neighbor-joining algorithm (Saitou and Nei, 1987). The topology of the tree obtained is similar to the one depicted in Figure 1 of Li et al. (Li et al., 2008), even if in that case the authors used a different approach, and to the one proposed by (Pickrell and Pritchard, 2012), which also incorporates admixture events and used a subset of ascertained markers. The tree shows non-African populations as splitting from African sub-Saharan populations, thus supporting the out-of-Africa model (Excoffier, 2002). Outside the African continent, the tree suggests a series of successive population splits, which mostly matches continental subdivisions and, within each continent, with the geographic distances between populations (Ramachandran et al., 2005; Rosenberg et al., 2005). The second analysis (Figure 6(b)) is a classical MDS analysis based on IBS distances between pairs of individuals. The plot shows the genetic similarities between the HGDP-CEPH individuals in the first two dimensions. Supporting the previous analysis, African sub-Saharan individuals tend to be differentiated from the other individuals in both dimensions. Outside the African continent, individuals tend to cluster according to their continent of origin and are distributed in a West to East pattern. Such geographic dependence with the genomic landscape can be highlighted if both dimensions are counter clock rotated (Wang et al., 2010, 2012). The MDS also highlights the presence of some individuals showing different ancestries than the other individuals from the same population. This is, for example, the case with the Mozabite population (red dots), which on average are genetically similar to European and Middle Eastern populations. However, some Mozabite individuals show strong genetic similarities with African sub-Saharan populations, thus indicating recent genetic admixture (Henn et al., 2012). The third analysis, an ADMIXTURE analysis (Figure 6(c)), was performed to estimate the ancestry proportions in each individual assuming seven ancestral populations. This number was chosen to compare results with the ones obtained with FRAPPE algorithm, initially applied on the same data (Li et al., 2008). ADMIXTURE analysis showed that the ancestral components tend to be associated to the continent of origin of current human populations, yet showing some degree of admixture within each population and individual. Compared to FRAPPE results, the ADMIXTURE analysis detects additional population substructure within the African sub-Saharan populations,

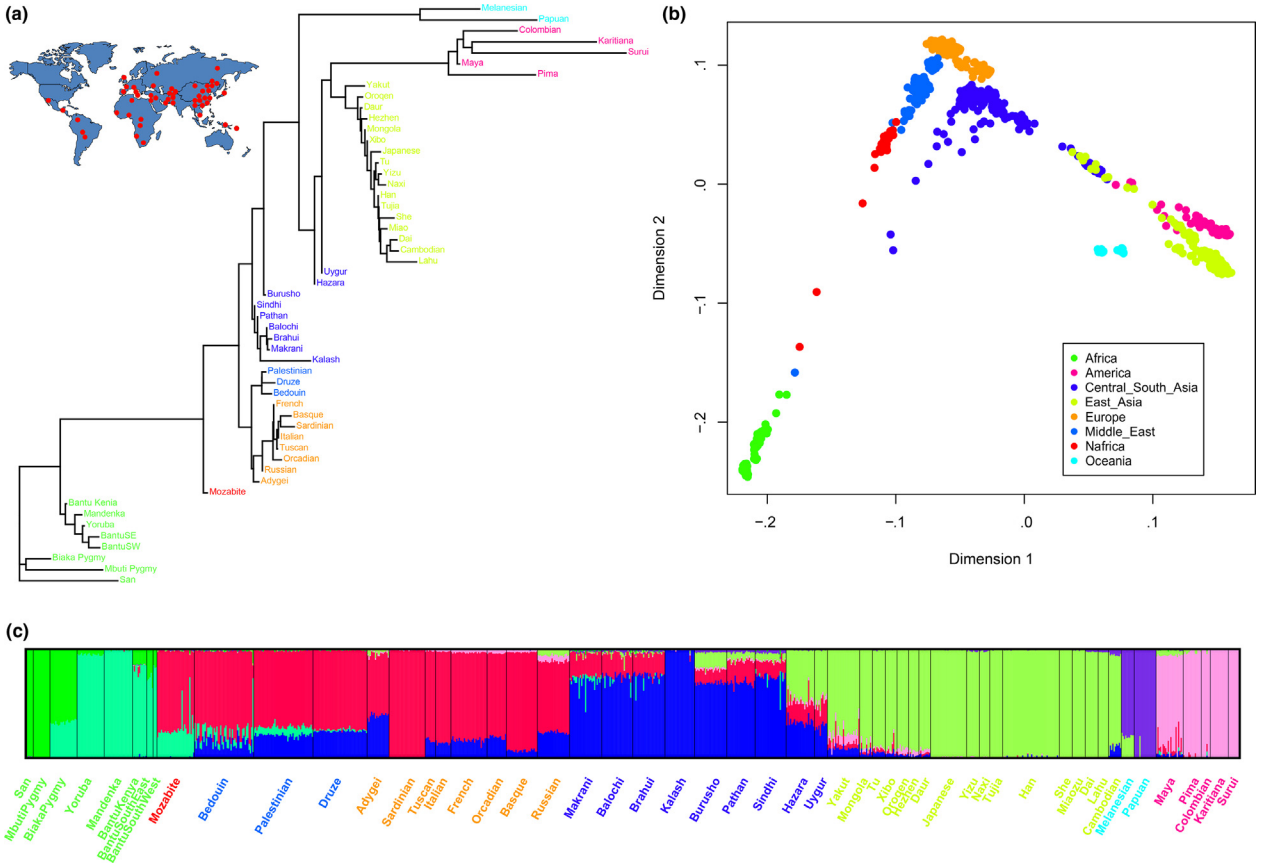

Figure 6 Application of three different methods to the study of the genetic variation present in ∼650 000 SNPs (Li et al., 2008) using the HGDP-CEPH panel, which comprises 53 worldwide human populations. (a) Neighbor-joining tree based on an F_{st} distance matrix between pairs of populations, using the San as out-group; (b) Classical MDS on an IBS distance matrix between pairs of individuals; (c) ADMIXTURE analysis using $K = 7$.

but fails to detect population substructure between Middle Eastern and European populations. Besides these discrepancies, both FRAPPE and ADMIXTURE results tend to correlate, in that they detect similar proportions of African sub-Saharan ancestry in Mozabite individuals (and thus supporting the previous observation based on MDS results), and in detecting East Asian ancestry in the Melanesian individuals, among others.

2. Detecting recent human admixture: the case of the post-Columbian admixture in the Americas

One prime example of large-scale recent admixture between humans originating from different continents can be observed in the Americas. The American continent was first colonized by anatomically modern humans from Siberia via the Bering route (O'Rourke and Raff, 2010). Native American populations tend to show genetic patterns in agreement with this route of migration, with a constant decay of genetic diversity from North to South of the American continent (Reich et al., 2012; Wang et al., 2007). However, following the European discovery of the Americas by Christopher Columbus in the year 1492, a complex admixture process, involving resident American natives, invading European colonizers, and forcibly moved African slaves, began. As a result, the genetic make-up of many contemporary Americans is a mosaic of those three ancestral populations, in some cases supplemented with other ancestry components (e.g., West Asian, South Asian, and East Asian) that reached the Americas during subsequent centuries through migration (Eltis, 2002). The relative contribution of each of the continental founder populations differs drastically depending on the subregion or the self-declared ethnic group within the Americas (Galanter et al., 2012; Figure 7), to the extent that in most of the cases these group assignations can be considered cultural constructs with little genetic support (Kittles and Weiss, 2003). Furthermore, additional sub-continental population substructure can also be detected within self-identified groups, such as within US European Americans (Campbell et al., 2005), US African Americans (Tishkoff et al., 2009; Zakharia et al., 2009), and US Hispanic Americans (Wang et al., 2008).

Interestingly, by analyzing the ancestry of NRY and mtDNA it has been shown that the admixture process was quite often sex-biased. For example, mestizo populations of Latin America usually carry predominantly Amerindian mtDNA types, but predominantly European NRY types (Bryc et al., 2010) indicating that the historical admixture involved mostly Amerindian women and mostly European men. A similar mating trend is also observed in the US African American

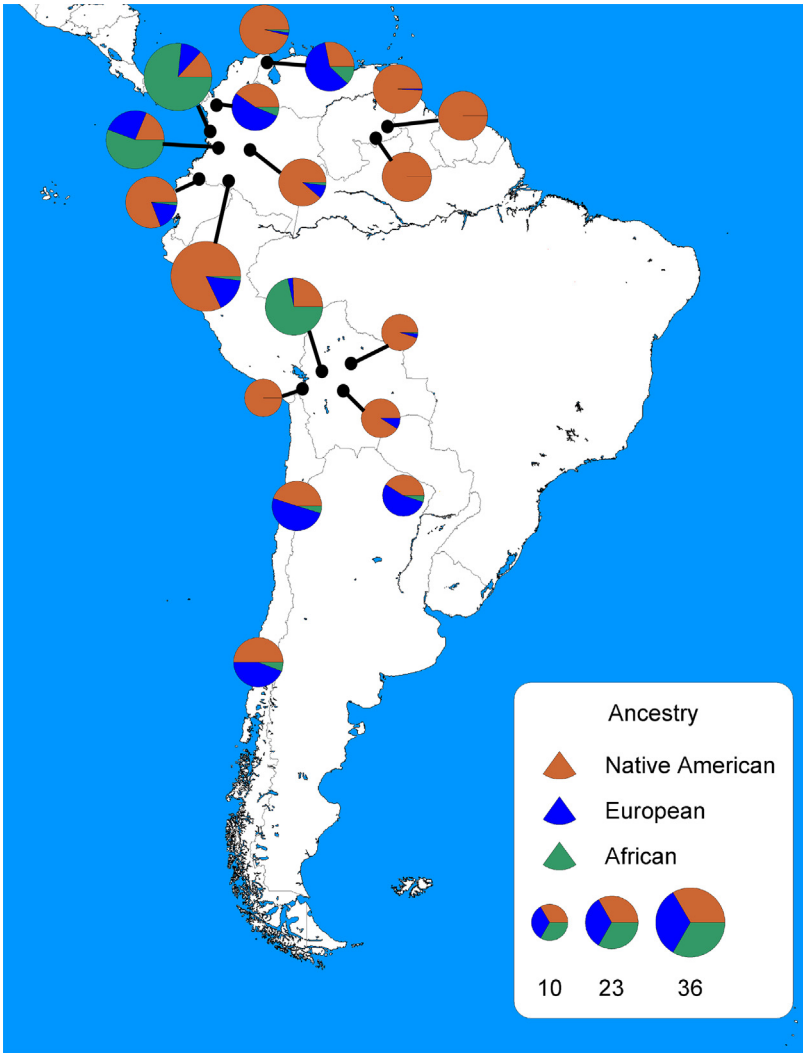

Figure 7 European, African, and native American components estimated in 18 Latin American populations, using ancestral informative markers (AIMs). Map adapted from Galanter, J.M., et al., 2012. Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. PLoS Genetics 8 (3), e1002554 using the ancestry frequency Table 4 of that article.

population (Figure 8; Lao et al., 2010; Lind et al., 2007; Stefflova et al., 2009), where the genetic ancestry of the mtDNA can be traced back almost entirely to the African sub-Saharan continent, whereas approximately 30% of the NRY ancestry is of European origin. In contrast, self-identified US European Americans show little evidence in their genomes of non-European ancestry (Figure 8; Lao et al., 2010; Lind et al., 2007; Stefflova et al., 2009).

This raises the question of why descendants from African and European ancestry couples tend to be perceived as US African Americans, rather than assigned at random to one of the two parental groups. One possible explanation that has been suggested (Lao et al., 2010) is that self-identification in admixed individuals tends to be biased toward visible traits with a biological basis such as skin pigmentation combined with cultural–sociological aspects (Bamshad and Guthery, 2007).

Concluding Remarks

Given the particular evolutionary history of humankind, mostly involving recent migrations, genetic admixture has played a major role on shaping the genetic variability of our genome. However, identifying such genomic fingerprints is complex and requires studying different aspects of the genomic variation within populations, individuals, and genomes. The development of new powerful statistical tools for detecting genetic admixture aids us to unravel details from our demographic past that could not be inferred otherwise. These statistical developments, together with the current technological advances in genome sequencing (Abecasis et al., 2012), hold the promise of an unprecedented level of resolution for detecting hidden genealogical relationships (Lawson et al., 2012) and genetic admixture processes in the near future.

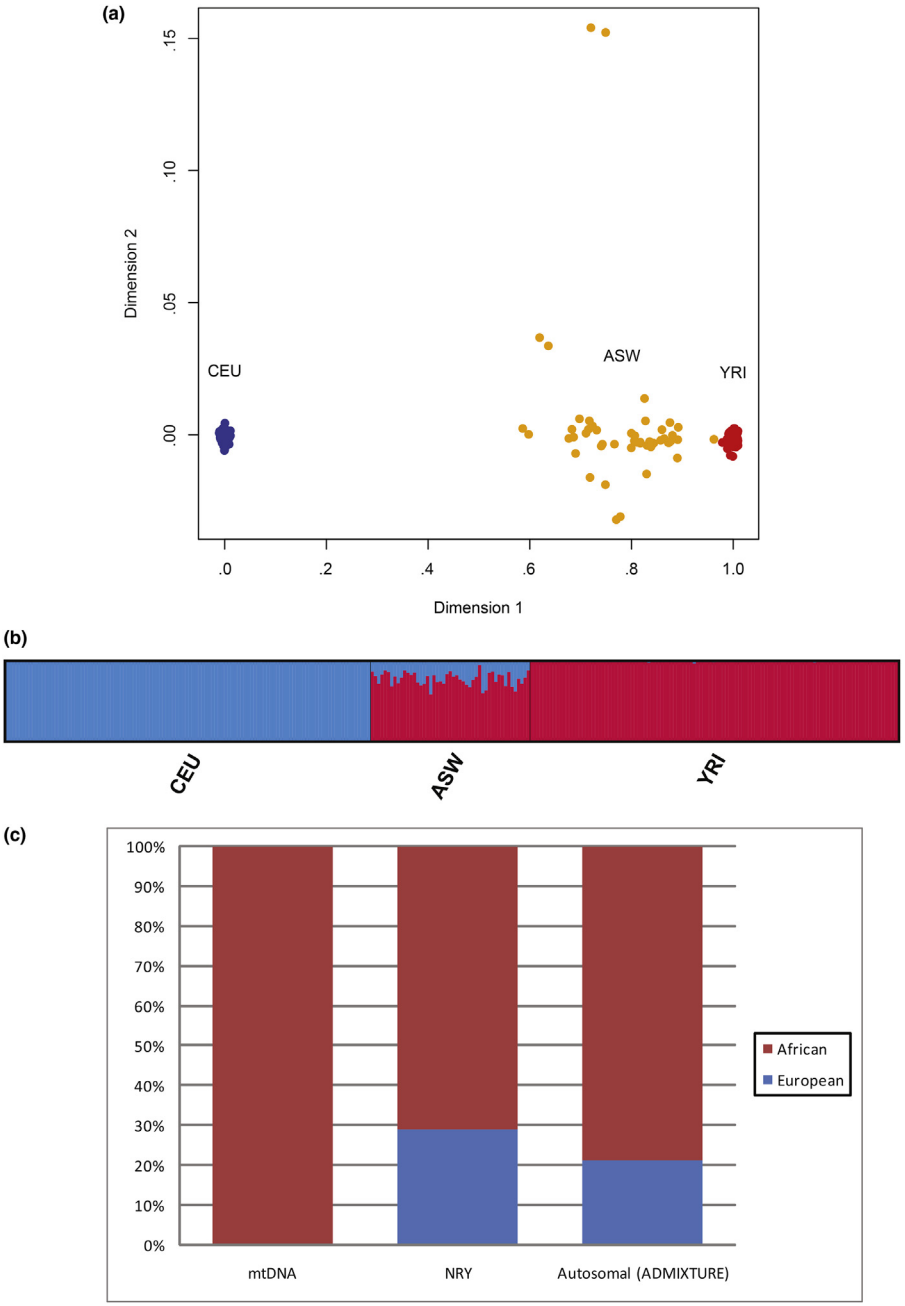

Figure 8 Estimation of genetic ancestry in unrelated African American individuals (ASW) from the HapMap III project (Altshuler et al., 2010) using Yoruban (YRI) and European American (CEU) individuals as proxy parental populations. (a) First two dimensions of a classical MDS on an IBS distance matrix computed with autosomal data (∼1.2 million SNPs). Each dot represents one individual. The first dimension has been rescaled so the mean value is 0 for CEU and 1 for YRI individuals. ASW individuals are distributed in the first dimension around 0.78 (ranging from 0.59 to 0.96); (b) ADMIXTURE analysis on the same data. Each bar represents one individual. In agreement with the MDS result, the African sub-Saharan ancestral component (red color) represents on average 78% (ranging from 59% to 96%) of the genome of ASW individuals; (c) Continental African sub-Saharan and European ancestry estimated on NRY variation, mtDNA, and autosomal data (ADMIXTURE result) in ASW individuals.

See also: Cultural Evolution: Theory and Models; Eugenics in the Americas; Evolutionary Demography: A Darwinian Renaissance in Demography; Genetic Ancestry Testing; Genetics and Anthropology; Human Evolutionary Genetics; Human Genome Diversity Project: History; Human Mate Choice, Evolution of; New Genetics and Race; Race: Genetic Aspects.

Bibliography

Abecasis, G.R., et al., 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491 (7422), 56–65.

Abi-Rached, L., et al., 2011. The shaping of modern human immune systems by multiregional admixture with archaic humans. Science 334 (6052), 89–94.

Alexander, D.H., Novembre, J., Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Research 19 (9), 1655–1664.

Altshuler, D.M., et al., 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467 (7311), 52–58.

Bamshad, M., Guthery, S.L., 2007. Race, genetics and medicine: does the color of a leopard's spots matter? Current Opinion in Pediatrics 19 (6), 613–618.

Barbujani, G., Belle, E.M., 2006. Genomic boundaries between human populations. Human Heredity 61 (1), 15–21.

Barbujani, G., Colonna, V., 2010. Human genome diversity: frequently asked questions. Trends in Genetics 26 (7), 285–295.

Bernstein, F., 1931. Die geographische Verteilung der Blutgruppen und ihre anthropologische Bedeutung. Comitato Italiano per o Studio dei Problemi della Populazione. Istituto poligrafico dello stato, Rome.

Bertorelle, G., Excoffier, L., 1998. Inferring admixture proportions from molecular data. Molecular Biology and Evolution 15 (10), 1298–1311.

Bryc, K., et al., 2010. Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proceedings of the National Academy of Sciences of the United States of America 107 (Suppl. 2), 8954–8961.

Campbell, C.D., et al., 2005. Demonstrating stratification in a European American population. Nature Genetics 37 (8), 868–872.

Cavalli-Sforza, L.L., 2005. The human genome diversity project: past, present and future. Nature Reviews Genetics 6 (4), 333–340.

Cavalli-Sforza, L.L., Feldman, M.W., 2003. The application of molecular genetic approaches to the study of human evolution. Nature Genetics 33 (Suppl), 266–275.

Cavalli-Sforza, L.L., Menozzi, P., Piazza, A., 1994. The History and Geography of Human Genes. Princeton University Press, Princeton, NJ.

Cox, T.F., Cox, M.A.A., 2001. Multidimensional scaling. In: Monographs on Statistics and Applied Probability, second ed., vol. 8. CHAPMAN & HALL/CRC, Florida.

Diamond, J., Bellwood, P., 2003. Farmers and their languages: the first expansions. Science 300 (5619), 597–603.

Dupanloup, I., Bertorelle, G., 2001. Inferring admixture proportions from molecular data: extension to any number of parental populations. Molecular Biology and Evolution 18 (4), 672–675.

Eltis, D., 2002. Coerced and Free Migration: Global Perspectives. Stanford University Press, Stanford, CA.

Excoffier, L., 2002. Human demographic history: refining the recent African origin model. Current Opinion in Genetics and Development 12 (6), 675–682.

Foley, R., 1998. The context of human genetic evolution. Genome Research 8 (4), 339–347.

Galanter, J.M., et al., 2012. Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. PLoS Genetics 8 (3), e1002554.

Gravel, S., 2012. Population genetics models of local ancestry. Genetics 191 (2), 607–619.

Green, R.E., et al., 2010. A draft sequence of the Neandertal genome. Science 328 (5979), 710–722.

Gutenkunst, R.N., et al., 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics 5 (10), e1000695.

Handley, L.J., et al., 2007. Going the distance: human population genetics in a clinal world. Trends in Genetics 23 (9), 432–439.

Hartl, D.L., Clark, A.G., 1997. Principles of Population Genetics, third ed. Sinauer Associates, Inc., Sunderland, MA, p. 542.

Henn, B.M., et al., 2012. Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genetics 8 (1), e1002397.

Holsinger, K.E., Weir, B.S., 2009. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nature Reviews Genetics 10 (9), 639–650.

Jobling, M.A., Hurles, M.E., Tyler-Smith, C., 2004. Human Evolutionary Genetics. Origins, Peoples and Disease. Garland Science, NY, p. 523.

Jobling, M.A., Tyler-Smith, C., 2003. The human Y chromosome: an evolutionary marker comes of age. Nature Reviews Genetics 4 (8), 598–612.

Jombart, T., Pontier, D., Dufour, A.B., 2009. Genetic markers in the playground of multivariate analysis. Heredity 102 (4), 330–341.

Karafet, T.M., et al., 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Research 18 (5), 830–838.

King, T.E., et al., 2007. Africans in Yorkshire? the deepest-rooting clade of the Y phylogeny within an English genealogy. European Journal of Human Genetics 15 (3), 288–293.

Kittles, R.A., Weiss, K.M., 2003. Race, ancestry, and genes: implications for defining disease risk. Annual Review of Genomics and Human Genetics 4, 33–67.

Lao, O., et al., 2010. Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA. Human Mutation 31 (12), E1875–E1893.

Lawson, D.J., et al., 2012. Inference of population structure using dense haplotype data. PLoS Genetics 8 (1), e1002453.

Li, J.Z., et al., 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319 (5866), 1100–1104.

Lind, J.M., et al., 2007. Elevated male European and female African contributions to the genomes of African American individuals. Human Genetics 120 (5), 713–722.

Liu, H., et al., 2006. A geographically explicit genetic model of worldwide human-settlement history. American Journal of Human Genetics 79 (2), 230–237.

Long, J.C., 1991. The genetic structure of admixed populations. Genetics 127 (2), 417–428.

McVean, G., 2009. A genealogical interpretation of principal components analysis. PLoS Genetics 5 (10), e1000686.

Menozzi, P., Piazza, A., Cavalli-Sforza, L., 1978. Synthetic maps of human gene frequencies in Europeans. Science 201 (4358), 786–792.

Novembre, J., Ramachandran, S., 2011. Perspectives on human population structure at the cusp of the sequencing era. Annual Review of Genomics and Human Genetics 12, 245–274.

O'Rourke, D.H., Raff, J.A., 2010. The human genetic history of the Americas: the final frontier. Current Biology 20 (4), R202–R207.

Oota, H., et al., 2001. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. Nature Genetics 29 (1), 20–21.

Pakendorf, B., Stoneking, M., 2005. Mitochondrial DNA and human evolution. Annual Review of Genomics and Human Genetics 6, 165–183.

Patterson, N., Price, A.L., Reich, D., 2006. Population structure and eigenanalysis. PLoS Genetics 2 (12), e190.

Patterson, N., et al., 2012. Ancient admixture in human history. Genetics 192 (3), 1065–1093.

Pfaff, C.L., et al., 2001. Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. American Journal of Human Genetics 68 (1), 198–207.

Pickrell, J.K., Pritchard, J.K., 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genetics 8 (11), e1002967.

Pinhasi, R., et al., 2012. The genetic history of Europeans. Trends in Genetics.

Price, A.L., et al., 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genetics 5 (6), e1000519.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155 (2), 945–959.

Pugach, I., et al., 2011. Dating the age of admixture via wavelet transform analysis of genome-wide data. Genome Biology 12 (2), R19.

Ramachandran, S., et al., 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences of the United States of America 102 (44), 15942–15947.

Reich, D., et al., 2012. Reconstructing native American population history. Nature 488 (7411), 370–374.

Rosenberg, N.A., et al., 2005. Clines, clusters, and the effect of study design on the inference of human population structure. PloS Genetics 1 (6), 1–12.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4 (4), 406–425.

Shriver, M.D., Kittles, R.A., 2004. Genetic ancestry and the search for personalized genetic histories. Nature Reviews Genetics 5 (8), 611–618.

Slatkin, M., 2008. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 9 (6), 477–485.

Stefflova, K., et al., 2009. Evaluation of group genetic ancestry of populations from Philadelphia and Dakar in the context of sex-biased admixture in the Americas. PLoS One 4 (11), e7842.

Tang, H., et al., 2005. Estimation of individual admixture: analytical and study design considerations. Genetic Epidemiology 28 (4), 289–301.

Tishkoff, S.A., et al., 2009. The genetic structure and history of Africans and African Americans. Science 324 (5930), 1035–1044.
Torroni, A., et al., 2006. Harvesting the fruit of the human mtDNA tree. Trends in Genetics 22 (6), 339–345.
Underhill, P.A., Kivisild, T., 2007. Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. Annual Review of Genetics 41, 539–564.
van Oven, M., Kayser, M., 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Human Mutation 30 (2), E386–E394.
Verdu, P., Rosenberg, N.A., 2011. A general mechanistic model for admixture histories of hybrid populations. Genetics 189 (4), 1413–1426.
Wang, C., Zollner, S., Rosenberg, N.A., 2012. A quantitative comparison of the similarity between genes and geography in worldwide human populations. PLoS Genetics 8 (8), e1002886.
Wang, C., et al., 2010. Comparing spatial maps of human population-genetic variation using Procrustes analysis. Statistical Applications in Genetics and Molecular Biology 9 (1), Article 13.
Wang, J., 2003. Maximum-likelihood estimation of admixture proportions from genetic data. Genetics 164 (2), 747–765.
Wang, J., 2006. A coalescent-based estimator of admixture from DNA sequences. Genetics 173 (3), 1679–1692.
Wang, S., et al., 2007. Genetic variation and population structure in native Americans. PLoS Genetics 3 (11), e185.
Wang, S., et al., 2008. Geographic patterns of genome admixture in Latin American Mestizos. PLoS Genetics 4 (3), e1000037.
Xing, J., et al., 2010. Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. Genomics 96 (4), 199–210.
Yang, W.Y., et al., 2012. A model-based approach for analysis of spatial structure in genetic data. Nature Genetics 44 (6), 725–731.
Zakharia, F., et al., 2009. Characterizing the admixed African ancestry of African Americans. Genome Biology 10 (12), R141.

Genetic Ancestry Testing

Jessica Bardill and Nanibaa' A Garrison, Stanford Center for Biomedical Ethics, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

From the large public interest in genealogy to academic research in human genetic diversity, genetic ancestry testing is a developing area that requires attention to interpretation of data and the ethical involvement of people and human populations. This article reviews the important definitions in genetic ancestry testing, ethical considerations in the uses of genetic ancestry testing, large-scale projects (including commercial ventures) that use the technology, and finally the limitations that have to be considered in uses of the tests.

Introduction

Ancestry can refer to different parts of an individual's lineage, such as close biological relatives or more distant ancestors. Primarily, it includes *where* someone's ancestors came from but also importantly *when* that person's ancestors lived in that place. Gathering information such as family stories and birth records provides information about both where and when one's ancestors are from and are important for reconstructing personal ancestral histories. To go further back into one's personal ancestral history, many people have turned to genetics to uncover their past. Due to genetic recombination, a process that shuffles what genes come from a person's mother and what genes come from the father, taking place at every generation, different gene segments are passed down from one generation to the next. Each gene segment, or small part of the parental genome, has its own ancestral history. Based on information about the connections between populations and particular gene segments, scientists can use that information to analyze an individual's genome to trace their lineage to different ancestral histories and populations. For example, one gene segment might be traced to Africa while another gene segment might be traced to Europe. When looking for individual lineage or family history in genealogical information searches, individuals may turn to genetic ancestry testing (Royal et al., 2010). Genetic ancestry tests use a person's DNA to gain information about where their ancestors were from many generations ago.

Genetic ancestry testing has grown in popularity among the general public as a tool to allow people to learn more about their own personal lineage, also referred to as their ancestral genetic history. Interest in genetic ancestry has also been stimulated by marketing from direct-to-consumer (DTC) testing companies. It remains a popular topic among researchers in areas as diverse as human migration and disease susceptibility. While popularity has driven technological and statistical innovations in these tests, genetic ancestry testing requires ongoing evaluation of its risks and limitations. This evaluation includes consultation or even partnerships between research and consumer communities, which involves making the tests' basic science, including their limitations, accessible to all involved in research and consumer testing. Scholars have noted the ethical implications of the concept of genetic ancestry and its testing (Lee, 2012) and remarked on the lack of standards and uniformity across DTC companies (Lee et al., 2009). However, other scholars are dedicated to identify more genetic markers to make these tests more precise and useful to individuals as well as for further research questions (Nassir et al., 2009). Recent research on consumer attitudes and knowledge about genetic ancestry testing (Wagner et al., 2012) has noted a gap between overall understanding that is related to different sets of information presented through commercial venues. An overview of genetic ancestry testing presented to the United States Congress also demonstrates the interrelations between these attitudes, lack of standards, and it argues for the place of policy and legislation or other protections for individuals and groups (Sarata, 2008). This policy briefing demonstrates how different motivations of consumers, variable knowledge of genetic science and statistics, as well as little to no regulation or standardization within the industry, create further confusion in the populace, as well as in between the companies.

This article reviews the important definitions in genetic ancestry testing, ethical considerations in the uses of the testing, large-scale projects (including DTC companies) that use the technology, and finally limitations of the tests themselves. This article attempts a comprehensive examination of the most widely used tests and their implications, but does not attempt a truly exhaustive examination, instead pointing to additional areas for exploration.

Concepts and Definitions

Ancestry

Ancestry as used here refers to a person's genetic lineage, which includes close familial relatives as well as more remote ancestors from hundreds or thousands of years previously. The difference in time, or *when* these relations occurred, combined with particular geographical locations, or *where* those ancestors came from, provides information about different parts of a person's ancestry.

Deep Ancestry versus Near Ancestry

When examining the uses of genetic ancestry testing for academics and the public, it is important to think about two types of ancestry: deep ancestry and near ancestry. Deep ancestry refers to many thousand years ago in a person's lineage, outside of the scope of most genealogical records. On

the other hand, near ancestry refers to close relatives, such as parents, grandparents, and other immediate relatives who are living or lived around the same time as the person who is interested in investigating his or her own ancestry. The scientific and statistical methods that are used to understand a person's ancestry vary for deep versus near ancestry. These methods produce an estimate of how much of one's genetic background matches up with database information to determine whether there is a match to one ancestry group or another. These results are referred to as ancestry estimates. Scientists use statistical analyses to make inferences about a person's ancestry, where that ancestor lived, and how long ago that ancestor lived.

Biogeographical Ancestry

Biogeographical ancestry refers to the connection drawn between a genetic sequence and a geographical location. These connections are also known as associations. This geographical location is further connected to a group of presumed ancestors by matching up information from the genetic variation of modern populations living in these geographical areas to a particular genetic sequence (Royal et al., 2010; Nassir et al., 2009). A researcher uses statistical estimates of multiple DNA markers to infer deep ancestry. DNA markers have been identified as being unique to certain modern populations living in particular geographical locations, also known as reference populations, and are used to estimate how similar or different a DNA sample is to that reference population. This genetic information is conveyed as 'ancestry proportions' that are correlated with human reference populations in specific geographical locations in order to infer how much of one's genome is similar to the reference population. A scientist might provide ancestry estimates about a person's genome that can be divided into multiple proportions of ancestry that add up to 100%. For example, a person's genome may comprise of three estimated proportions of ancestry: 50% European, 40% African, and 10% Native American. These proportions, or probable percentages of the genome distributed by geography, are determined through the use of multiple ancestry informative markers (AIMs) and a set of ancestry estimates. By itself, it may not provide sufficient genealogical information but can add to known genealogical information.

The related concept of 'continental ancestry' similarly focuses on geographical origins represented by five major parental populations (which are typically the continental areas of Africa, Europe, Asia, Native America, and Oceania) to which a person's sample is then correlated. This correlation refers to ancestry from hundreds or thousands of years previously.

Types of Genetic Ancestry Tests

Genetic ancestry testing relies on multiple kinds of genetic tests, including mitochondrial, Y-chromosome, and autosomal tests. Depending on the genetic test, an ancestry test can provide specific information about single branches of individual family trees, or they can additionally provide broad information about multiple branches of those family trees.

Much of the time, a kind of DNA marker known as a single nucleotide polymorphism (SNP) is used as a piece of information in a genetic ancestry test. When a given sample is genotyped, or processed to determine the genetic makeup of the sample, it can then be compared against reference sets of samples that have also been genotyped for similarities and differences. SNPs are allelic variations that occur on the level of one nucleotide change (A, T, G, or C) that differs from one person to another; for example, the difference between AATC[**C**]GT and AATC[**G**]GT. As a group, SNPs represent 90% of human genetic variation. SNPs have proven more abundant in certain populations than others, so a given sample is compared against a database of SNPs across populations. In genetic ancestry testing using SNPs, the number of SNPs between that sample and populations in the database can help determine a probability for origin. In order to be considered a useful marker, the SNP frequency has to be present in at least 1% of the populations under consideration, otherwise it would be too rare to be accurate. These results have not yet proven conclusive to the satisfaction of many in the field, including the presentation of results to the consumer without adequate background, interpretation, and analysis.

Mitochondrial DNA Testing

Mitochondria are organelles in eukaryotic cells that engage in energy production and cellular respiration. They also have their own genomes, known as mitochondrial DNA (mtDNA), that are distinct from the nuclear genome of a given cell. Mitochondrial genomes are found in all humans, but are inherited strictly maternally (from mother to offspring), and the mitochondrial genome does not undergo a process of recombination in the process of reproduction. Mitochondrial genetic variations are passed down through generations, from mother to child, and can therefore be tracked back through those ancestral generations along the maternal lineage. Mitochondrial lineages are defined by these genetic variations that are individually designated with letters and subgroup identifications (numbers and letters), also known as 'haplogroups.' The haplogroups are linked together in a large family tree that shows how different human groups are related to each other based on mtDNA information. The origins of different haplogroups are associated with given geographical regions. For example, the mtDNA haplogroup L is understood to have originated in Africa, and many African women are of this mitochondrial lineage.

Y-Chromosome Testing

Unlike the mitochondria that all humans inherit from their biological mother, only males have and share Y-chromosomes. The Y-chromosome is the sex-linked male chromosome that is passed from biological father to son, allowing a male individual to trace their paternal genetic lineage. The nonrecombining portions of Y-chromosomal DNA (also known as the NRY) are analyzed using genetic techniques, also known as genotyping, to identify SNPs and short tandem repeats (STRs), another type of allelic variation in the genetic sequence. While females do not have Y-chromosomes and therefore cannot use Y-chromosomal information to gain information about their genetic paternal lineages, they can learn more about that part of their genealogy if a close male relative in their father's line elects to have the testing and share information.

Ancestry Informative Markers

Ancestry informative markers (AIMs) are a collection of genetic markers that assist in providing estimations of genetic ancestry.

Markers were selected as AIMs if they had large allelic frequency differences (such as 30–50%) between populations. In most cases, SNPs are used as AIMs. There now exist multiple databases of AIMs that can be used in testing a sample or large set of samples to determine their probable origins. AIMs demonstrate shared genetic variation which typically correlates to shared ancestry, but which may also be the result of shared environmental changes and selective pressures without shared ancestry. Generally, around 50 AIMs could provide some information about the continental ancestry of a person, but more AIMs (such as several thousand) involved in the genotyping allow more certainty. A genetic ancestry test based on AIMs can provide consumers with an idea of their ancestry as best determined by comparing it against the company's database.

Uses and Ethical Considerations

Genetic ancestry tests have a particular utility for understanding genetic connections between persons. Their use to discern deep ancestry has been applied in myriad ways, each of which brings its own set of ethical considerations. One of the most important ethical considerations is the psychological impact for individuals or groups who receive genetic information from ancestry tests that may complicate or contradict family history or other sources of genealogical documentation, such as caused by issues of nonrelatedness, or claims to different national or tribal heritages.

Genealogical Information

Genealogical information is a major area in which genetic ancestry testing has been used to fill in gaps of traditional genealogical searches. While this use of genetic ancestry tests may be perceived as recreational, the results nevertheless can have serious ramifications. Some may view genetic ancestry testing as a definitive answer to their questions of heritage and identity. However, since testing results can contradict not only one another but also other genealogical information, their accuracy, assumptions, and limitations must be understood. For example, many African-Americans in search of their specific African ancestry have taken genetic ancestry tests that infer their ancestors came from particular places in West Africa, only to take another test with another company that contradicts that initial finding, presenting instead a separate ancestral connection, such as Southern Africa. The psychological impact on identity formation, as well as related group affiliations and removal from groups related to the sociocultural rules of ancestry (Royal et al., 2010), is importantly tied to this information.

Group Affiliation/Citizenship

Genetic ancestry testing may have important effects on group affiliation or citizenship in a tribe or nation. On the one hand, genetic ancestry testing might help to build connections with some groups while, on the other hand, it may sever social and cultural ties with other groups; for example, ancestry information may connect African-Americans to certain African peoples or it may lead to disenrollment of other individuals from an American Indian nation. A person's genetic test results might lead him or her to seek affiliation with a group or have additional genetic testing performed to find out specific ethnic ties to support a presumed affiliation. Others may go as far as changing their listed race or ethnicity due to results of the genetic ancestry test. On the other side, testing could also *disprove* a person's ascribed race or ethnicity and provide a challenge to their affiliation or citizenship within a group, tribe, or nation.

Human Migration Analysis

In the realm of scientific research, genetic ancestry testing primarily serves as a tool for human migration analyses. This research traces the histories of large migrations of populations of humans over time. In contrast, indigenous peoples have stories of their origins that are connected to particular places, which can relate to the *where* and *when* of their ancestral group formation. New research findings can undermine claims to territory or homelands as well as stories of group origins tied to particular geographies. These research studies have important ethical implications for individuals and groups whose samples are used for migration studies; for example, an individual from a group (such as a tribe) who participates in one type of research may inadvertently affect the whole group to whom the individual belongs because genetic information about that individual might also reveal genetic information about the whole group. On the other hand, genetic information that is revealed about a particular group can also affect individuals who identify with that group even if that individual decides not to participate in the research.

Pharmacogenomic Possibilities

Large-scale genetic ancestry testing and expansion of databases for genetic variants associated with certain populations bring with it the possibilities of applications in pharmacogenomics. Pharmacogenomics is an emerging field of study that incorporates genetic information with pharmaceutical drug information in efforts to better understand how drug interactions vary based on variation in genetic information. The promise of individualized medicine (also known as personalized medicine or precision medicine) would match up information from a patients' genetic variation with the best therapeutic options. However, to date only a few genes have been directly connected to specific disease states, such as Huntington's disease, and even fewer genes have been directly connected to better understanding effective drug interactions.

The introduction of race or other phenotypes in medical practice creates challenges in interpretation of genetic information for physicians and the public. By saying that a gene occurs more frequently in a continental population, which has correlations to racial categories, confusion with those racial categories can occur. Further, given the admixture of the overall population, physicians may have a harder time identifying members of continental population groups or even the stand-in of race, demonstrating a limited utility to the information. Even self-identification does not promise to assist in this issue, as the gene in question could occur in other modern populations or they could have no knowledge of that continental ancestry. For example, BiDil is a drug that was tested and then marketed to

African Americans, under the claim that it was more effective in that population because that population was genetically distinct, claims which are neither true nor founded in the clinical trials.

Truly personalized genomic medicine does have clinical utility, including the possibility for enhancing standard practice of care for individuals. The sequence and interpretation of one's personal genome could have promise in a healthcare system where that information can be used by physicians and clinical staff provided that the specific genes, and not substitutes for them such as ancestry, are used.

Large-Scale Projects Based on Genetic Ancestry Tests

Genetic ancestry testing has contributed to a large and profitable research and commercial sector. The most prominent projects, both public and commercial, will be reviewed here including their aims, progress, and critiques. These projects have built on the technology to discover more about variation in AIMs throughout the world, adding to the novel but limited information available after the completion of the Human Genome Project. These new projects continued in the private and public sectors, adding to the knowledge of larger connections between peoples through human migration analysis, or the study of how humans have moved throughout the globe and populated it, as well as the diversity of genetic variation.

Human Genome Diversity Project

Beginning shortly after the Human Genome Project, a separate effort called the Human Genome Diversity Project (HGDP) sought to catalog the genetic diversity of modern human populations around the world. Initially, this project received much attention and support from scientists who were interested in human genetic diversity. The HGDP organizers considered the indigenous peoples of the world as prime targets for testing because they primarily reproduced within their group and therefore were the least admixed in regards to their genetic ancestry (Cavalli-Sforza, 2005). The rationale for sampling isolated populations worldwide was to glean as much information about human genetic diversity as possible. However, HGDP organizers lost support from the very populations they wished to sample when they stressed the urgency to sample them before they *disappeared* through admixture with other populations.

In a historical account of the HGDP, Jenny Reardon's *Race to the Finish: Identity and Governance in an Age of Genomics* (Reardon, 2004) elucidates the flaws within the system including some of the researchers' beliefs in genetic purity versus genetic admixture, and the targeting of indigenous peoples. Reardon also documents the lack of community engagement. For example, HGDP researchers often quickly flew in and out of communities for genetic samples, a practice disparingly termed 'helicopter research.' Communities did not perceive benefits from participation in HGDP, and eventually many refused participation. Funding to the project from the federal level was eventually halted, though the project has continued in another form. Reardon's analysis presents why the HGDP became enmeshed in political troubles, but also concludes with the understanding that the idea of the project, to study human genetic variation, would continue and possibly much less publicly. This change prevents room for public debate and community consultation during the process.

At the end, 1064 participants from 51 distinct communities or tribes around the world were included in the HGDP database. These samples have subsequently been used by many researchers worldwide in studies on human diversity, human migration, and even in medical research. While these uses of the samples were not part of the original plans or consents, the community of scientists continues to use them for such varied research.

National Genographic Project

The National Genographic Project (NGP) aims to understand the deep ancestry of the modern human population by tracing human migrations within and out of Africa, and throughout the rest of the world. The NGP has sought to do research on a massive global scale, produce texts and films about the results, to elicit additional support and samples through a DTC effort to engage populations in developed countries to be a part of the project by sending in samples and paying for results, and puts the results into the public domain. It does so through mtDNA and NRY analyses performed as research worldwide, via population sampling and through interested individuals who submit their samples to contribute to the overall project and gain information about their ancestry. After the HGDP's controversies, the NGP took on a similar role in the world of research and genetic ancestry testing, though attempting to learn from the mistakes of HGDP, such as making sure to include indigenous peoples in their advisory board and creating tangible benefits for the indigenous populations sampled through a grants mechanism. The NGP seeks to analyze historical patterns in human genetic variation to better understand human ancestry. The NGP created the film *The Journey of Man* to showcase the project's initial findings based on Y-chromosomal DNA and their analyses of human migration based on those NRY variations. The project went on to look at the mitochondrial DNA relations in New York City, as a microcosm of the world, through *The Human Family Tree*, presenting not the current populations in their particular geographies as in the previous film but instead demonstrating the wide range of human diversity found within a specific modern geography.

Overall the NGP has come under much criticism for its practices, including lack of engagement with indigenous research subjects, conflation of modern peoples with ancestral populations, both in the rhetoric and the sampling, as well as results that reframe historical relations between groups. For example, throughout the films, mixed rhetoric demonstrates the ideological conflicts that emerge in the project, including problems to truly informed consent of participants, a master narrative of deep human relationships undermined by near racial and geographical relations, and the denial of modernity to current peoples, such as the living San Bushmen, who stand in for ancestral populations in the research (TallBear, 2007; Wald, 2006). The NGP continues despite these criticisms and has initiated a new phase known as Geno 2.0 that incorporates the use of 150 000 AIMs to provide information about regional specificity of ancestry and to allow more information than the

previous focus on mitochondrial and Y-chromosome testing had alone.

The International HapMap Project

The International HapMap Project is an effort to seek and catalog the genetic diversity and similarities of human beings. Multiple countries are contributing funding and research, including the United States, Canada, China, Japan, Nigeria, and the United Kingdom. Started in 2002, the HapMap aims to use information generated to support research in health fields by locating genes that affect disease and therapeutic drug responses as well as environmental factors. Importantly, the HapMap plans to and has been releasing its information into the public domain instead of private patent or other such limiting efforts.

Initially, the HapMap focused on four populations who were of European descent in Utah, USA, of Han Chinese descent in Beijing, China, of Japanese descent in Tokyo, Japan, and of Yoruban descent in Ibaden, Nigeria. These four populations represented European, Asian, and African ancestries and have been used as reference populations to understand the ancestry similarities and differences for other populations worldwide. Learning from the experiences from HGDP, the HapMap project organizers worked with the four communities of these populations in order to thoroughly assess informed consent before recruitment into the study. This work has now expanded to include seven additional populations, given the success of the applications of the initial data in biomedical research and promise for other groups.

DTC Testing Companies

Like the arm of the NGP's research that assists individuals who are interested in their ancestral history and in contributing to research, DTC genetic testing companies are emerging to respond to the demand of the public for ancestry information. These companies have also started to capitalize on their capacity, refining their own databases and contributing to novel research as well. Commercial DTC testing companies offer various kinds of genetic tests to map siblingship and distant kinship. These testing companies utilize their own proprietary databases of reference samples, which can lead to different tests from different companies producing different results (see section Contradictory Results). In a statement in 2008, the American Society of Human Genetics (ASHG) provided many recommendations regarding genetic ancestry testing, particularly regarding the DTC companies and other researchers who use it. The ASHG recommends that greater efforts to inform the public of the limitations of ancestry testing need to be made, more research needed to be done to prove the accuracy of the testing results, and guidelines should be developed to facilitate the return of and interpretations of results. They also call for greater accountability of the DTC companies (ASHG, 2008). In an update to this statement, a white paper came out in 2010 that further recommended that academic researchers should promote 'thoughtful and rigorous use of ancestry estimation' in their work, and that a national round table discussion should occur with DTC companies, researchers, consumers, and other key stakeholders to discuss the issues of concern and work to develop solutions (Royal et al., 2010).

Limitations

Many of the limitations of genetic ancestry testing have already been alluded to throughout this article, but the following section highlights the major areas of concern, some of which may change over time.

Databases

The current databases of samples, including mitochondrial DNA, Y-chromosomes, and SNPs, are limited in size, both in regards to the number of populations sampled and the individuals sampled within any given population to represent it and its variation. For a given geographical area, the populations and individual representatives thereof that are sampled do not reflect the entire genetic diversity of the area currently or that may have once existed in that geographical area. This limit is a particular concern with certain populations, like Native Americans, because lack of representation within the databases reflects less involvement in the research and translates to less accurate conclusions in test results. The limited size of these databases, as well as who owns which samples, can lead to misidentification via current sampling and differing conclusions when more SNPs are analyzed. For example, two different companies could have access to two vastly different databases of samples from global populations, and each could own their database and guard it as 'proprietary' information. However, a consumer who sent in a test sample to both companies could receive different ancestry estimates based on the different sets of AIMs or SNPs in their separate databases. The limited sampling of the vast amount of genetic diversity within the human population, both by number of populations and number of individuals within those populations sampled, further constrains the results available to only those gene variants present in sampled populations. From this limited sample, the probability of ancestry estimates from those populations is provided.

Ancestral Populations Derived from Modern Populations

In a related constraint, the use of modern peoples to stand in for ancestral populations itself has benefits and limits to usability. Given the recent developments in genetic testing technology and the paucity and degradation of sources of ancestral genetic data, there is benefit in using modern peoples to represent the development over time of human genetic variation. However, this analogy of modern genetic variation with ancestral, geographically specific genetic variation fails to take into account the genetic development that has occurred since the time period of the ancestral past. Again, the *when* of genetic ancestry test is important for the *where*, but that *where* is only dated in most cases to the present. It is important to recognize that the modern populations in a given geographical area have also undergone admixture and have acquired new genetic variations. They may closely resemble the ancestral populations to which they are correlated, but to what extent, there is uncertainty. While far fewer direct ancestral samples exist, such as preserved specimens or museum pieces from which DNA can be isolated and sequenced, even the use of these specimens produces additional ethical concerns about the uses of remains in ways which the original owner did not and could not have consented.

Conflation of Race and Biogeographical Ancestry

In connection to the conflation of modern peoples and ancestral populations is the conflation of the social construct of race and biogeographical ancestry. Race has been used as a proxy for biological difference, particularly in reference to phenotype differences as well as different areas of geographical origin. Biogeographical ancestry refers to the association between a geographical location and one's presumed ancestors, inferred to be related through genetic comparison with modern peoples in those particular geographical locations. Inherent to this conflation is the point that geography is socially constructed, meaning a genetic ancestry testing result that relates a proportion of estimated ancestry from Africa is understood by the recipient as related to the social construction of Africa, in modern and historical times, in regards to races, resources, and global positioning. Therefore, a result that ties someone ancestrally to any given place has to be understood in relation to how race has been used in and about that place. Given the additional layer of limitation, regarding the use of modern peoples for ancestral populations (Bolnick et al., 2007), this further conflation of race cannot be ignored. Race and its construction, as well as the influences of the genetic ancestry estimates, still permeate many lives and can lead those who receive testing results to reconsider their identity constructs despite the contrasting information, such as their communities, environments, or cultures.

Contradictory Results

Related to the limitations in the databases, when tested against different sets of reference AIMs or SNPs, some samples produce different results. In other words, in testing a given sample, one database will return ancestry estimates of a certain proportion (e.g., 50% Europe, 20% Africa, and 30% Asia) while another will return ancestry estimates of a contradictory proportion (e.g., 30% Europe, 30% Africa, and 40% Asia). While contradictory results are far less likely with mtDNA and Y-chromosome testing, they are an important concern for both research and commercial uses of genetic ancestry testing. For individuals seeking ancestry information, mtDNA and NRY only provide information about two members of a given individual's lineage. If the ancestral mix in a given person is different via different tests, what is their actual ancestral composition? How do they make sense of these contradictory results in regards to their identities? These questions demonstrate both the limits and ethical considerations that come along with using genetic ancestry testing. Beyond the individual research participant or consumer, additional questions emerge. How do researchers reconcile different results in the laboratory, particularly if these results undergird pharmacogenomics or migration study analyses? This multifaceted question emphasizes that different databases may yield different information and that using proxies for given set of variations in the genome, whether related to ancestry or not, is a poor substitute for the accurate variant.

Genetic Nonrelatedness

Another limitation, connected to certain ethical concerns, comes from the possibility of nonrelatedness revealed through genetic ancestry testing. Commonly, it is noted that nonpaternity occurs in 10% of the population, meaning that in a given group, 1 of 10 times the person who is believed to be the biological father is not such. Nonmaternity, while occurring in a much smaller portion of the population, is on the rise with innovations in reproductive technologies. This issue has been well documented in terms of its ethical constraints but also in how families given such information react, including continuing to claim kinship where they are told no genetic affiliation exists. For genetic ancestry testing, the issue with nonpaternity arises in that Y-chromosome testing and SNP analysis may reveal different deep ancestry and near ancestry than previously known, again providing information to the subject or consumer that can alter their understanding of self and have wider implications for their families and communities. Genetic tests could impact the way research participants and their families use the information. For example, a nonpaternity event in the data would provide both misleading and ethically challenging information. Further, given the commonality and possibility of genetic nonrelatedness between family members, genetic ancestry testing could be more complicated by a nonrelated event: if a sample is taken from a person who believes they have always been from a certain geographical area and a nonrelated event has occurred in their lineage (connecting them to a different geographical location), their sample may be correlated with a certain geographical location even though their mtDNA or NRY, and other genes, comes from a different geographical location.

Statistical Methods and Ancestry Estimates

A final and important limitation of genetic ancestry testing, and much genetic testing in general, is the fact that ancestry estimates are based on probability. This means that the results, or estimates, depend on the representation in the database, which also depends on the analysis of the samples, all of which can provide places for the estimates to be skewed in different directions. The statistical methods of ancestry testing carry a certain amount of uncertainty, and given the statistical nature of the estimates, definitive results beyond near ancestry are rare. Because of the inherent uncertainty in these statistical methods, researchers, test producers, and consumers all need to have a strong understanding of the probabilities. In other words, as in the contradictory multiple results, the correlation of genetic variants in a sample is made to other samples and though it might occur within multiple populations in the world, the likeliest population is chosen. The certainty of these numbers increases with the number of variants sampled as well as the robustness of the database, but they do not typically carry a 100% certainty, which would only come with an allelic variant private to a particular population present in the database. A gene variant may occur in multiple populations and therefore a test that indicates Middle Eastern descent also could have indicated Native American descent or East African descent in its place, but the probability of the variant being found in those populations is smaller.

Conclusion

This review of the terms, ethical considerations, uses and projects, as well as the limitations thereof has sought to show

the current operations of genetic ancestry testing, including why individuals, communities, and scientists have interest in its possibilities and development. Overall, the concerns regarding genetic ancestry testing must be taken seriously, and recommendations from places such as the ASHG (ASHG, 2008; Royal et al., 2010) should be followed to create better science and better understanding of that science by all involved, particularly those who receive results.

In order to avoid potential harms in the future, proper analysis, interpretation, and statements of limitations along with informed consent should occur for those participants involved in the research. Standards for DTC companies and researchers alike will help prevent contradictory information coming to the public, and the influence of that misunderstanding upon legislation, policy, or ethical conduct should also be a goal as the science moves forward. Consumers, communities, and participants must also educate themselves for the protection that such knowledge provides, through training or their own research outside of the transactions.

Future possibilities exist for the application of the testing to personal and group identity formations, kinship determinations, and group affiliation. These areas are controversial so it remain important to consider the way that the genetic ancestry testing operates within different cultures, where ideas such as race, family, and relations are socially defined. In turn, those ideas influence the uptake and understanding of information from the testing.

See also: Bioethics: Genetics and Genomics; Community Genetics; Direct-to-Consumer Personal Genetic Testing; Familial Studies: Genetic Inferences; Genetic Admixture; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genographic Project; Human Genome Diversity Project: History; Indigenous Knowledge and Intellectual Property Rights; Kinship, Evolution of; Race: Genetic Aspects.

Bibliography

Bolnick, Deborah., Fullwiley, Duana, Duster, Troy, Cooper, Richard S., Fujimura, Joan H., Kahn, Jonathan, Kaufman, Jay S., Marks, Jonathan, Morning, Ann, Nelson, Alondra, Ossorio, Pilar, Reardon, Jenny, Reverby, Susan M., TallBear, Kimberly, 2007. The science and business of genetic ancestry testing. Policy Forum. Science 318, 399–400.

Cavalli-Sforza, L. Luca, 2005. The human genome diversity project: past, present, and future. Opinion. Nature Reviews 6, 333–340.

Lee, Sandra Soo-Jin, 2012. Genetic ancestry. In: Chadwick, Ruth (Ed.), International Encyclopedia of Applied Ethics, second ed. Elsevier Inc, pp. 414–419.

Lee, Sandra Soo-Jin., Bolnick, Deborah A., Duster, Troy, Ossorio, Pilar, TallBear, Kimberly, 2009. The illusive gold standard in genetic ancestry testing. Policy Forum. Science 325, 38–39.

Nassir, Rami., Kosoy, Roman, Tian, Chao, White, Phoebe A., Butler, Lesley M., Silva, Gabriel, Kittles, Rick, Alarcon-Riquelme, Marta E., Gregersen, Peter K., Belmont, John W., De La Vega, Francisco M., Seldin, Michael F., July 2009. An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. BMC Genetics 24 (10), 39.

Reardon, Jenny, 2004. Race to the Finish: Identity and Governance in an Age of Genomics. Princeton UP, Princeton, NJ.

Royal, Charmaine D., Novembre, John, Fullerton, Stephanie M., Goldstein, David B., Long, Jeffrey C., Bamshad, Michael J., Clark, Andrew G., 2010. Inferring genetic ancestry: opportunities, challenges, and implications. The American Journal of Human Genetics 86, 661–673.

Sarata, Amanda, 2008. Genetic Ancestry Testing. Congressional Research Service Report for Congress. (Online) Available at: http://www.policyarchive.org/handle/10207/bitstreams/18866.pdf.

TallBear, Kim, 2007. Narratives of race and indigeneity in the Genographic Project. The Journal of Law, Medicine & Ethics 35 (3), 412–424.

The American Society of Human Genetics, 2008. Ancestry Testing Statement (Online, accessed 31.05.12.) Available at: http://www.ashg.org/pdf/ASHGAncestryTestingStatement_FINAL.pdf.

Turner, Trudy (Ed.), 2005. Biological Anthropology and Ethics; From Repatriation to Genetic Identity. State University of New York Press, Albany, New York.

Wagner, Jennifer K, Cooper, Jill D., Sterling, Rene, Royal, Charmaine D., 2012. Tilting at windmills no longer: a data-driven discussion of DTC DNA ancestry tests. Genetics in Medicine 14 (6), 586–593.

Wald, Priscilla, 2006. Blood and stories: how genomics is rewriting race, medicine and human history. Patterns of Prejudice 40 (4–5), 303–333.

Relevant Websites

http://ashg.org – American Society of Human Genetics.

http://indigenousgenomicsgovernance.org – Genomics, Governance, and Indigenous People Group.

http://genetics.ncai.org – National Congress of American Indians (NCAI), Policy Research Center, American Indian/Alaska Native Genetics Resource Guide.

https://genographic.nationalgeographic.com/ – National Genographic Project.

http://www.genome.gov/glossary/index.cfm – National Human Genome Research Institute (NHGRI) Talking Glossary of Genetic Terms.

http://hapmap.ncbi.nlm.nih.gov/ and http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html – The International HapMap Project.

Genetic Approaches to Memory

Alcino J Silva, Integrative Center for Learning and Memory Brain Research Institute, UCLA Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Genetic approaches have been extensively used in memory studies, from investigations of basic mechanisms of memory function to experiments designed to understand the mechanisms responsible for cognitive deficits. The approaches used vary and are under constant development, but can be divided into two major categories. Forward genetic studies are focused on the identification of genes responsible for memory processes of interest, including the discovery of genetic mutations that account for inherited disorders that affect cognitive function. These studies usually start with the identification of interesting memory phenotypes, and their aim is to identify and isolate the gene(s) responsible for those phenotypes. In contrast, reverse genetic studies, as the name implies, start with candidate genes that may be involved in memory, and the goal is to discover and understand the possible memory phenotypes (i.e., the function) associated with that gene. Reverse genetic studies involve the study of organisms with defined genetic changes. Although originally the goal of genetic approaches to memory was to understand the connections between genes and memory phenotypes, recently these approaches have been used to probe memory theories that go well beyond the function of the genes being manipulated.

The Role of Genetics in Biology

To understand the role of genetic studies in unraveling the mechanisms responsible for biological processes, including learning and memory, it is important to place these studies in the much broader perspective of biological investigations. The principal goal of biological studies is to understand complex phenomena such as learning and memory. Central to this process is the determination of causal connections that account for the properties of complex phenomena (e.g., learning) by the actions and functions of other relevant phenomena (e.g., genes, cells, circuits, brain regions). There are two major classes of strategies used in biological studies (including genetic studies) to test causal connections between phenomena of interest: convergence and reproducibility. All things being equal, hypotheses derived from reproducible data (i.e., reproducibility) are deemed more reliable than the alternative. Similarly, hypotheses supported by very different types of convergent evidence (e.g., elimination and overexpression of a gene) are usually considered to be in stronger footing than hypothesis derived from multiple pieces of similar evidence. Genetic approaches are capable of manipulating genes in a wide variety of ways, including studies that delete, alter, or increase the expression of a gene of interest, either in the whole organism or in a specific cell population or brain region (Silva et al., 1997). The only other biological strategy capable of manipulating molecular function is pharmacology. In this case, chemical agents are synthesized that target the function of a given biological agent (i.e., the function of a protein). Therefore, since genetics is one of only two general approaches to investigate biological function, it is easy to see why it has played a key role in every biological discipline. Beyond its key role in testing biological hypothesis (i.e., is a given kinase gene required for memory), the principal role that genetics has played in biology has been to suggest possible hypothesis or explanations of natural phenomena. Indeed, forward genetic screens have allowed biologists to make major discoveries, even in the absence of a well-delineated hypothesis.

Forward Genetics

Long before we had the ability to directly manipulate genes in animals, such as flies and mice, geneticists were busy using chemical mutagens to alter genetic information in living systems (forward genetics). The goal of classical or forward genetics, which continues to be used extensively to this day, is to identify the genes critical for biological processes of interest. The idea is that study of those genes is often a critical first hint for unraveling underlying biological processes. In forward genetic screens, animals are first exposed to a mutagen, for example, the DNA-altering compound ethyl-nitroso-urea, mated, and the progeny are screened for phenotypic changes of interest. The phenotype of a mutant is the sum total of observed biological changes caused by a genetic manipulation. Recent application of this approach in the study of mammalian circadian rhythms resulted in the identification of *clock*, a crucial link in the cascade of transcriptional events that marks molecular time in organisms as diverse as Drosophila and mice (Wilsbacher and Takahashi, 1998). Other molecular components of this pathway, such as *per*, were isolated in mutagenesis screens in Drosophila. By identifying novel and unexpected molecular components of biological processes of interest, forward genetics has often reshaped entire fields of research. At times, Science can go in circles, obsessively chasing its own tail of half-truths, incapable of escaping the gravitational pull of its worn-out paradigms. Forward genetics, in the hands of masters such as Edward Lewis (developmental mutants) and Seymor Benzer (learning mutants), has the ability to turn paradigms upside down, and initiate new lines of scientific inquiry. The key to the success of forward genetics is the design of biological screens with which the randomly mutagenized animals are tested. If the screens are too stringent, or if the fundamental biological insights underlying the screen's design are off mark, one runs the risk of ending up with empty hands, or even worse, with a number of misleading mutants. In contrast, nonstringent designs lead to overwhelming numbers of nonspecific mutants that are essentially useless.

The First Screens for Learning and Memory Mutants

Seymor Benzer and colleagues working with Drosophila at the California Institute of Technology designed the first successful screen for learning and memory mutants in the 1970s (Dudai, 1988). Benzer and colleagues developed a behavioral procedure with operant and Pavlovian components. During training, the flies were allowed to enter two chambers, each with a different odorant, but they only got shocked in one of the chambers. During testing, approximately two-thirds of the trained flies avoided the chamber with the odorant that previously had been paired with shock. With this procedure, Benzer and colleagues tested a number of Drosophila lines derived from flies treated with ethylmethane sulfonate. The first mutant line isolated from this screen was *dunce* (Dudai, 1988).

Remarkably, three out of the four learning and memory mutations, first discovered in genetic screens in *Drosophila*, code for members of the cAMP-signaling pathway. For example, *dunce* lacks a phosphodiesterase that degrades cAMP. Importantly, these findings have recently been extended into vertebrates, where electrophysiological and behavioral studies have confirmed the critical importance of cAMP signaling to learning and memory (Silva et al., 1998). Remarkably, in the early 1970s, Eric Kandel and his colleagues at the Colombia University also found evidence for the importance of cAMP signaling in learning and memory with a completely different approach. They used a reduced cellular preparation to study sensitization, a nonassociative form of learning, in the sea snail *Aplysia* (Byrne and Kandel, 1996). They also found that sensitization depends on cAMP signaling. This is a fascinating example of convergent evidence in science, but it also serves to illustrate that genetics, like any other tool in science, is most successful when used in parallel with other approaches. The persuasive power of convergent evidence cannot be overemphasized. Besides identifying new genes, genetics can also be used to test hypothesis about the function of cloned genes (reverse genetics).

Reverse Genetics

In classical genetics, an interesting phenotype is usually the driving force behind the molecular experiments required to identify the underlying mutant gene(s). In contrast, in reverse genetics, the interesting molecular properties of a gene usually drive the generation and study of the mutant animal (hence, the word reverse). It is now possible to delete and add genes to many species, ranging from bacteria to mice. For example, mice can be derived with the deletion (knock-outs) or overexpression (transgenics) of almost any cloned gene. These manipulations can involve whole genes or they can target specific domains or even single base pairs.

To generate knock-out mice, the desired mutation is engineered within the cloned gene, and this mutant DNA is introduced into embryonic stem (ES) cells. Since ES cells are pluripotent, they can be used to derive mice with the genetically engineered lesion. For that purpose, they are injected into blastocysts (early embryos), and the blastocysts are implanted in host mothers. The resulting chimeric (having mutant and normal cells) offspring are then mated to obtain mutant mice. In contrast, transgenic mice are derived by injecting the pronuclei of fertilized eggs with a DNA construct carrying a gene of interest under the regulation of an appropriate promoter. The injected eggs are transplanted into pregnant females and some of the resulting progeny will have the transgenic construct inserted randomly in one of its chromosomes.

With classical knock-out and transgenic techniques it is not possible to regulate the time and the regions affected by the mutation/transgene. However, recent techniques promise to circumvent these limitations with a variety of techniques. For example, the expression of the gene of interest can be regulated by gene promoters that can be controlled by exogenously provided substances, such as tetracycline derivatives (Mayford et al., 1997). Alternatively, it is also possible to regulate the function of a protein of interest by fusing it with another protein that can be regulated by synthetic ligands, such as tamoxifen (Picard, 1993). For example, our laboratory recently showed that a transcriptional repressor called CREB can be activated at will when fused with a ligand-binding domain (LBD^m) of a modified estrogen receptor. Addition of tamoxifen (the ligand of the modified receptor) activates the $CREB^r/LBD^m$ fusion protein. It is important to note that irrespective of the exact method used, the general idea of reverse genetic studies is that the function of a gene can be deduced from the phenotype of the mutant animal.

Knock-outs, Transgenics, and Learning

The first knock-out/transgenic studies of learning and memory analyzed mice with a targeted mutation of the α isoform of CaMKII (Grant and Silva, 1994). Pharmacological studies had previously shown that this family of calcium calmodulin-induced kinases present in synapses were required for LTP, a stable enhancement in synaptic efficacy thought to contribute to learning and memory. Remarkably, deleting αCaMKII resulted in profound deficits in hippocampal LTP and in hippocampal-dependent learning and memory. Additional studies showed that either the overexpression of a constitutively expressed form of the kinase, or a mutation that prevented its autophosphorylation, also disrupted LTP and learning. Importantly, studies of hippocampal circuits that fire in a place-specific manner (place fields) showed that these CaMKII genetic manipulations disrupted the stability of these place-representations (but not their induction) in the hippocampus. Altogether, these studies suggested the provocative hypothesis that this kinase is important for the induction of stable synaptic changes, that the stability of synaptic changes is crucial for the stability of hippocampal circuits coding for place, and that these circuits are essential for spatial learning (Elgersma and Silva, 1999). Even this very abbreviated summary demonstrates that these studies went well beyond gene function. Instead, they used mutations to test hypotheses that connected molecular, cellular, circuit, and behavioral phenomena. Although it is reasonable to claim that the αCaMKII has a direct role in the regulation of synaptic function (for example, by phosphorylating glutatmate receptors), it is more problematic to argue that the kinase is regulating spatial learning. There are a lot more

phenomenological steps between kinase function and the animal's ability to find a hidden-platform in a water maze than between the biochemical properties of this kinase and its role in synaptic plasticity. By comparison, it is easier to see how the stability of place fields in the hippocampus could be an important component of hippocampal-dependent spatial learning.

Common Concerns with the Interpretation of Transgenic/Knock-out Studies

Despite the power of genetics there are a number of concerns that must be kept in mind when using genetic approaches. One of the most commonly discussed is the possibility that developmental effects or any other change caused by the mutation preceding the study could confound its interpretation. Another pertains to the possible effects of genetic compensation. Since proteins do not work alone, but instead function in highly dynamic networks, it is often observed that specific genetic changes lead to alterations/compensations in the function of other related proteins. A related concern pertains to genetic background. Extensive studies have shown that the genetic background of a mutation has a profound effect on its phenotype. The concerns listed above are not limitations of genetics, but simply reflect the properties of the biological systems that genetics manipulates. At the heart of many of the concerns described above are two misconceptions concerning the nature and organization of biological systems.

First, genetics is essentially a lesion tool. As other lesion tools, it cannot be used in isolation. To establish causal connections between any two phenomena (A and B) in science, it is never enough to lesion A and document the alteration of B. As described above, it is also critical to fulfill three other criteria: first, A must be observed to precede B. Second, triggering A should result in B, finally, it is essential to have a clear hypothesis of how A triggers B. Fulfilling one or two of those four criteria is simply not enough to establish a causal connection between A and B. Therefore, although studying the effects of a deleted protein is an important component in determining its function, it is by no means sufficient.

Second, biological systems are dynamic and adaptive, and therefore, the lesion of any one component is always followed by changes in several other components. Although it is often a helpful simplification to think of biological components as independent functional units, it is important to remember that they are not. Thus, it is hardly surprising that the effect of a mutation is dependent on biological variables such as genetic background.

The Future of Genetic Manipulations

In the near future, it will be possible to delete or modify any gene, anywhere in most organisms of interest, and at any time of choice. Additionally, more powerful forward genetic strategies will allow the isolation of entire pathways of genes involved in any neurobiological phenomenon of interest, including learning, attention, emotion, and addition. In parallel with expected advances in genetics, there will also be advances in the methods used to analyze mutants. These advances are just as critical to genetic studies as advances in genetic methodology. For example, imaging the brain of mutant mice may yield insights into how molecular lesions affect the function of brain systems. Most genetic studies of learning and memory in mice have focused on the relationship between cellular phenomena (i.e., LTP) and behavior. Advances in small animal magnetic resonance imaging may enable the kind of functional system analysis in the brains of mutant mice that have so far only been possible in large primates. Similarly, multiple single-unit recording techniques are starting to yield systemwide snapshots of circuit activity in the brains of mutant mice. At a molecular level, small-size positron emission tomography devices will allow the imaging of molecular function in leaving animals, such as mice. For example, it may be possible to image the activation of a receptor, such as the dopamine receptor, during learning or memory. Microarray techniques and other molecular cloning approaches will allow the identification of gene profiles in mutant mice. These molecular profiles will be critical to delineating the molecular changes behind the expression of a mutant phenotype. It is important to note that genetics allows us to reprogram the biology or organisms. The finer and more sophisticated the phenotypic and genotypic tools that we have at our disposal, the deeper we may be able to probe the magical natural programs embedded in our genes.

See also: Epigenetics of Brain Plasticity and Behavior; Learning and Memory, Neural Basis of; Memory Development During Infancy and Early Childhood across Cultures; Neural Plasticity.

Bibliography

Byrne, J.H., Kandel, E.R., 1996. Presynaptic facilitation revisited: state and time dependence. Journal of Neuroscience 16 (2), 425–435.

Dudai, Y., 1988. Neurogenetic dissection of learning and short-term memory in Drosophila. Annual Review of Neuroscience 11, 537–563.

Elgersma, Y., Silva, A.J., 1999. Molecular mechanisms of synaptic plasticity and memory. Current Opinion In Neurobiology 9 (2), 209–213.

Grant, S.G., Silva, A.J., 1994. Targeting learning. Trends in Neurosciences 17 (2), 71–75.

Mayford, M., Mansuy, I.M., et al., 1997. Memory and behavior: a second generation of genetically modified mice. Current Biology 7 (9), R580–R589.

Picard, D., 1993. Steroid\binding domains for regulating functions of heterologous proteins in cis. Trends in Cell Biology 3, 278–280.

Silva, A.J., Kogan, J.H., et al., 1998. CREB and memory. Annual Review of Neuroscience 21, 127–148.

Silva, A.J., Smith, A.M., et al., 1997. Gene targeting and the biology of learning and memory. Annual Review of Genetics 31 (5352), 527–546.

Wilsbacher, L.D., Takahashi, J.S., 1998. Circadian rhythms: molecular basis of the clock. Current Opinion in Genetics and Development 8 (5), 595–602.

Genetic Counseling: Historical, Ethical, and Practical Aspects

Jörg Schmidtke, Hannover Medical School, Hannover, Germany
Heather Skirton, Plymouth University, Taunton, United Kingdom
Irmgard Nippert, Westfaelische Wilhelms-Universitaet Muenster, Muenster, Germany
Gerhard Wolff, Goethestraße, Freiburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Genetic counseling refers to the totality of activities that (1) establish the diagnosis of genetic disease, (2) assess the occurrence risk, (3) communicate to the patient and family the chance of recurrence, (4) provide full and unbiased information within a caring relationship regarding many problems raised by the disease and its natural history, including the potential medical, economic, psychological, and social burdens, and (5) provide information regarding the reproductive options. The understanding and goals of genetic counseling as they evolved over the last century exemplify the substantial shift in understanding, determining, diagnosing, and handling of hereditary/genetic risks. At present, nondirective concepts have found broad, though not universal, acceptance.

Genetic counseling is one of the most important direct applications of human genetic knowledge. Genetic counseling refers to the totality of activities that (1) establish the diagnosis of genetic disease, (2) assess the occurrence risk, (3) communicate to the patient and family the chance of recurrence, (4) provide full and unbiased information within a professional caring relationship regarding many problems raised by the disease and its natural history, including the potential ethical conflicts as well as the medical, economic, psychological, and social burdens, and (5) provide information regarding the reproductive options to be taken, including prenatal diagnosis, and referral of patients to the appropriate specialists. Genetic counseling should offer guidance but will allow those being counseled to arrive at their own choices, both before and after genetic procedures (Vogel and Motulsky, 2010; World Health Organization [WHO], 2012).

Paradigm Development and Concepts of Genetic Counseling

Prescientific knowledge of heritability in humans dates back to the early Greek physicians and philosophers, who presumably dispensed practical advice to individuals and couples as to how to use such knowledge (Vogel and Motulsky, 2010). The idea of an improvement of humankind by deliberate mate selection was a major aspect of ancient and Renaissance state and social utopias [Plato, *Politeia* (ca. 374 BC); Thomas More, *Utopia* (1516); Fra Tomaso Campanella, *The City of the Sun* (1623)]. In the medical literature of the eighteenth and nineteenth centuries, reports on the familial recurrence of particular disorders began to accumulate. An early and noteworthy example is the book by the English physician Joseph Adams (1756–1818), *A Treatise on the Supposed Hereditary Properties of Diseases* (1814), which attempts to lay out the fundamentals of genetic counseling for families affected by a genetic disorder.

Toward the end of the nineteenth century, two paradigms established human genetics as a science. Gregor Mendel (1822–84) originated the concept of the gene as the basic factor of inheritance (1866), while Francis Galton (1822–1911) introduced statistical methods to investigate and interpret the familial clustering of traits and thus founded the biometrical paradigm. Until the present, biometrics and Mendelian genetics have remained the basis of risk determination as the naturalistic component of genetic counseling. Galton coined the term 'eugenics' to mean the improvement of a population through measures to enhance propagation of individuals with supposedly high genetic qualities (positive eugenics) and discourage reproduction in those with undesired characteristics (negative eugenics). In Germany, around the turn of the twentieth century, Wilhelm Schallmeyer (1857–1919), Alfred Hegar (1830–1914), and Alfred Ploetz (1860–1940) developed the idea of negative eugenics–a concept of decreasing the number of offspring of individuals with supposedly below-average hereditary qualities, or of preventing congenitally diseased individuals from reproducing altogether, with the aim of improving the gene pool of future generations and especially that of the presumed superior Nordic race (see Ploetz, 1895).

In many countries during the first half of the twentieth century, positive and negative eugenics formed the ideological basis of state-enforced measures to exert reproductive control over individuals. Between 1911 and 1930, 24 U.S. states proclaimed sterilization laws, and eugenic thought also played a role in the enactment of restrictive immigration laws. In Europe, Denmark was the first country to proclaim a Law on Genetic Hygiene (1929), in which voluntary sterilization for genetic reasons was regulated. In Germany, the Law for the Prevention of Congenitally Diseased Progeny (1933), followed by a series of other laws and decrees, mandated and organized the sterilization of individuals with certain diseases, which were thought to be genetically determined. In Germany, eugenics, racial hygiene, and National Socialist ideology formed a gloomy alliance. The involvement of German geneticists in coercive measures and eventually the Holocaust to sections of the population that were declared to be inferior has been well documented by Müller-Hill (1997). It is beyond doubt that this historical burden strongly influences the current debate on ethics and genetics in Germany and elsewhere.

From about 1930 onward, the eugenic movement outside Germany began to lose power, but it was not until the 1950s that the population-directed eugenic orientation was slowly replaced by a medically oriented practice with genetic disease prevention as a major focus. Nevertheless, eugenic thinking, whether acknowledged or not, has continued to play a role until the present time, especially outside the Western world (Wertz, 1997). When, in 1947, Sheldon Reed coined the term 'genetic counseling,' it became clear that eugenics and what was then understood to be genetic counseling, were different entities. One historian (Stern, 2005) states that rather than clinical genetics being a complete departure from the eugenics movement, many of the concepts were retained, but the service was delivered in ways that enabled individual choice to be paramount. Clinical genetics differs in this sense from public health genetics, the focus of which is the improvement of the health of large populations in any public health setting. According to Reed, "genetic counseling is a type of social work entirely for the benefit of the whole family without direct concern for its effect upon the state or politics" (1974). This definition marks a change of paradigm in applied human genetics, away from the concern about the genetic makeup of the population and toward the attitude that decision making with regard to genetic risks should be left to the individual, even if this may create or increase social costs or be perceived as somehow worsening the gene pool. Thus, one of the fundamental ethical principles in medicine–respect for the autonomy and self-determination of persons–is also the cornerstone of genetic counseling.

Genetics in relation to human health has thus gone through three overlapping historical phases (Table 1). First, the eugenics phase, mainly rooted in England, the United States, Scandinavia, and Germany, was based on the social and biological utopia of general mental and physical health and the optimization of humankind. In order to attain these goals, depending on political circumstances, directive measures including direct and indirect coercion were applied. Eugenics was scientifically ill founded, and the definition of inferior or superior genetic qualities was a matter of ideology.

Second, the preventive-medical phase laid its emphasis on reducing individual suffering through prevention of genetic diseases and disabilities, in order to economize rare social welfare and medical resources. It was founded on the rational calculation that any means toward that end would find acceptance as being in the immediate interest of the individual. Strategies were those of benevolence, authority, and paternalism.

Third, in the biopsychosocial phase, the focus of genetic counseling is to help the individual or the family to manage a problematic situation caused by genetic disease. The basis of this help is a complex interaction between the counselor (a specialized physician or a nonmedically trained expert) and the client (patient) in a nondirective process of communication. Help is provided with the understanding that genetic diseases are often associated with special psychosocial problems and ethical conflicts that should be addressed individually. Thus, any direct or indirect coercive influence is to be avoided.

Nondirectiveness became a basic concept of genetic counseling (Wertz, 1997). However, genetic counselors have questioned the value of nondirective counseling from the patient's perspective for several decades, as it was interpreted by some as an inability to support patients in their decision-making process, rather than a lack of coercion (Wolff and Jung, 1995; Kessler, 1997). Patients frequently seek some reassurance from the counselor that the course of action they propose to take would be reasonable, and good counseling practice includes the exploration of the values and beliefs of the client and support for a decision that would be consistent with those values (Weil et al., 2006; Bartels et al., 1997). In addition, it should be clear that better health outcomes are likely to ensue from particular decisions, for example, having a genetic test that would clarify the need for invasive clinical surveillance (Schwartz et al., 2012). Under these circumstances, it is clear that provision of accurate information about potential benefits or disadvantages of testing should be explained as part of the genetic counseling process and that identifying health benefits should not be seen as coercive. However, coercion may exist where the counselor clearly tries to influence the patient to confirm to his or her own views.

The history of genetic counseling shows that its purpose has always been value-determined and has never been seen outside the context of social politics in which it was embedded. In countries that acknowledge human rights as defined in the democratic state theories of the eighteenth century, the only defensible position is that of the autonomy of the individual to make decisions regarding genetic counseling and genetic diagnosis, and to draw consequences for life and family planning.

The health care system, on the other hand, carries the ethical obligation to provide full information and the tools that may enable the individual, depending on the state of knowledge, the emotional situation, and moral positioning, to make use of that information.

Excursus: Ethics and Genetic Counseling: The Respect for Personal Autonomy

Today the term 'eugenics' is sometimes used to describe the cumulative result of individual choices after genetic counseling, especially in view of reproductive choices. The decreasing

Table 1 Paradigm shifts in applied human genetics

Paradigm	*Basis*	*Goal*	*Measure*
Eugenics	Emotionally influenced social utopia	Improvement of the gene pool	Directiveness, compulsion, coercion
Prevention	Rational calculation	Reduction of individual suffering, economic distribution of resources	Paternalism
Psycho(socio)-biology	Complex interaction between counselor (doctor) and client (patient)	Help in individual decision making	Empathy, nondirectiveness

prevalence of births of children with Down syndrome in western Europe, especially in the UK, may be regarded as a result of individual choices to undergo prenatal diagnosis and subsequently to terminate a pregnancy after a positive test result. Whether this type of individual eugenics is a benign development and socially desirable has been discussed with some controversy by Wertz (1998). Critics argue that free, autonomous choices (after genetic counseling) not only depend on surrounding social, cultural, and other factors, but also on the full and honest presentation of information in genetic counseling–that is, on access to valid information. Informed consent can only be said to exist under three conditions: The patient is given the appropriate information; the patient is competent to make the decision; and there is no coercion involved (Beauchamp and Childress, 2001). In the absence of correct and unbiased information, free choices are not possible and are compromised, and information and genetic counseling become a means of furthering the attitudes and goals of health professionals. A worldwide survey of 2091 genetics professionals in 36 nations conducted by Wertz (Nippert and Wolff in Germany) suggests that eugenic thought underlies the perception of the goals of genetics and genetic counseling in many countries. In developing nations of Asia and eastern Europe especially, directive counseling is practiced to influence individual decisions. Intentionally pessimistically biased information is routinely provided after positive results from prenatal diagnosis in order to avoid the birth of affected children (Wertz, 1998; Cohen et al., 1997). The data from this survey suggest that the concept of nondirective counseling found in English-speaking nations and western Europe is an aberration from the rest of the world.

Indications for Genetic Counseling

In medicine in general, a therapeutic activity is 'indicated,' when, after weighing the risks and benefits, it appears medically required. This definition is hardly applicable to genetic counseling because the recognition of risks and benefits of knowledge, as well as any actions that may result from such knowledge, will emerge only in the counseling process itself, and their appraisal is often more strongly determined by individual life circumstances than by medical judgment. On the other hand, the professionalization of genetic counseling as a specialization within medicine, and the need for correctness and completeness of any medical advice, have brought about a core service that every counselee may duly expect to receive: a scientifically founded statement regarding the etiology and the (recurrence) risk of a genetic condition, comprehensive information and counseling regarding diagnostic options, and referral to preventive, therapeutic, and/or socially alleviating measures.

Generally speaking, an indication for genetic counseling is given whenever an individual regards it as a means to help solve a personal problem related to an existing or suspected genetic risk. Table 2 summarizes the most frequently encountered situations. With the increasing awareness and availability of pre- and postnatal methods of genetic diagnosis, individuals may find themselves unexpectedly receiving risk information about conditions of which they had previously been unaware.

Table 2 Indications for genetic counseling

1. Birth of a child with congenital disease or disability
2. One or more relatives suffering from (suspectedly) genetic disease or disability
3. Affected counselee
4. Consanguineous couple
5. Recurrent miscarriage
6. Infertility
7. Mutagenic influences
   - 7.1 Advanced maternal age
   - 7.2 Pre- or postconceptional mutagen exposure
8. Teratogenic exposure
   - 8.1 Infections
   - 8.2 Drugs
   - 8.3 Alcohol
   - 8.4 Radiation

Concerns, including moral doubts, may have been raised, for example, by offers of prenatal screening (maternal serum marker and ultrasound screening), or carrier detection, within or outside the framework of a defined screening program because such offers may bring up the question of termination of the pregnancy.

The Counseling Process

According to an initial definition issued by an Ad hoc Committee on Genetic Counseling of the American Society of Human Genetics, genetic counseling is a communication process that:

> …deals with the human problems associated with the occurrence or risk of recurrence of a genetic disorder in a family. This process involves an attempt by one or more appropriately trained persons to help the individual or the family to 1. comprehend the medical facts, including the diagnosis, the probable course of the disorder, and the available management; 2. appreciate the way heredity contributes to the disorder and the risk of recurrence in specified relatives; 3. understand the options for dealing with the risk of recurrence; 4. choose the course of action which seems appropriate to them in view of their risk and their family goals and act in accordance with that decision; 5. make the best possible adjustment to the disorder in an affected family member and/or to the risk of recurrence of that disorder. (Fraser, 1974; Ad hoc Committee on Genetic Counselling, 1975)

However, developments in the profession resulted in a new definition of genetic counseling that involves an approach aimed at integrating the genetic information with other aspects of the client's life and encompasses more of the educational aspects of the activity. According to members of a task force of the National Society of Genetic Counselors (Resta et al., 2006, p. 77),

"Genetic counseling is the process of helping people understand and adapt to the medical, psychological and familial implications of genetic contributions to disease. This process integrates the following:

- Interpretation of family and medical histories to assess the chance of disease occurrence or recurrence.

- Education about inheritance, testing, management, prevention, resources and research.
- Counseling to promote informed choices and adaptation to the risk or condition."

The counseling process is usually initiated by clarifying the counselee's motivation for consultation. What is the background to the problem, and what are the expectations? The counselor should explain, in general terms, what genetic counseling can achieve and where its limits lie. This is important because expectations are frequently unrealistic. It needs to be explained, for example, that it is technically not possible to exclude any kind of genetic risk or to guarantee a healthy child. The next step is to record the personal and family medical history in the form of a genealogical tree, encompassing at least all first- and second-degree relatives. This is followed by providing the family with information regarding the most important medical facts concerning the disease in question (natural history of the disease, the likely degree of suffering it may confer, its treatability) as well as the likelihood of its occurrence or recurrence. It is important to give unbiased, objective information. During the later phases of the counseling process, the counselor will often return to this information to relate it to the personal views of the counselee.

The 'prior' risks for a genetic condition as deduced from its pattern of inheritance (monogenic dominant or recessive, autosomal or sex-linked, co-determined by several genes and/or the environment) may be modifiable by genetic diagnostic tests, other clinical test results, or other information (such as number of unaffected children in a family), so that a more precise risk figure emerges. Ever more often, with the near completion of sequencing and functionally annotating the human genome (Human Genome Project, 1997) and concomitant technical developments becoming routine practice, firm confirmation or exclusion of carrier status for any particular genetic disposition will be the rule rather than the exception. Nevertheless, the phenotypic outcome will often be variable, due to environmental factors, complex interactions between genes, and other 'non-Mendelian' genetic and epigenetic factors such as imprinting. In the counseling process, the resulting uncertainty needs to be explained and worked upon with great care. Without thoroughly understanding the chances and limits of genetic testing, an informed consent to any such procedure cannot be attained. Informed consent is generally held to be the prerequisite for any genetic testing, but as previously stated, the need for nondirectiveness in all situations is now questioned, even by potential users of the health service. A study by Martin et al. (2010) indicated that discussion among members of the public was more focused on individuals making choices rather than the principle of informed consent, although they did emphasize public education was a prerequisite to choice. They also expressed concern that coercion to be tested might stem from other members of family, rather than from health professionals.

During the counseling process it may be necessary to make a physical examination of the counselee and/or of family members in order to search, for example, for early signs of a late-onset disorder or microsymptoms of a disease. The impact of such an examination needs to be explained beforehand because counselees and their family members may wish to help clarify genetic risks for others, but not necessarily for themselves. It is sometimes difficult to decide to what extent information from the family history (such as risk for another condition that is apparent to the counselor but unsuspected by the counselee) needs to be discussed, if this information has not been central to the counselee's motivation for consultation. This is particularly delicate when no therapeutic or preventive options exist. Where these options do exist, however, the counselee has a right to comprehensive information and counseling.

The Determination of Genetic Risks

The determination of genetic risks in any individual case is the essential component of genetic counseling. It requires expertise in general and human genetics, including formal (mathematical) genetics, population genetics, and clinical genetics, as well as an awareness of novel developments, especially in the differential genetic diagnosis (attainable in this fast-moving field only through continuous further self-education). In the following discussion, we provide a summary of the most important groups of conditions in which a risk assessment is often sought and present a brief outline of risk provision. Often, the prior genetic risks need to be modified by using information from the pedigree or genetic tests (conditional probabilities). By employing Bayesian methodology, this information can be used to arrive at posterior (final) risks (Harper, 1998).

If a disease is (mainly) due to the effect of a single gene, the phenotype may be inherited in a recessive or a dominant fashion. The disease may manifest itself in a sex-dependent way, if the causative gene is not located on the autosomes (the chromosomes shared by both sexes in equal number) but rather on the sex chromosomes (X and Y). In an autosomal-recessive condition, the recurrence risk in a sibship is 1/4, but usually low (less than 1%) for second-degree relatives. If a parent has an autosomal-dominant condition, there is a 50% chance that each child will also carry the disease-causing mutation. If neither parent has a condition that is known to be autosomal-dominant, the recurrence risk in a sibship is usually negligible, because the presumption is that a spontaneous mutation has occurred in the germ line leading to the affected individual. In X-linked recessive disease, if the mother is a carrier, half of all her sons will develop the full clinical picture, and half of her daughters will be carriers. Mothers and daughters are usually unaffected or exhibit symptoms to a lesser degree (due to nonrandomness of X-inactivation during early development). In Y-linked disease (such as subfertility), the trait is passed on from fathers to sons.

In both recessive and, most notably, dominant conditions, the inheritance patterns may be obscured by variable penetrance (proportion of carriers manifesting the phenotype at all) and expressivity (degree of disease severity). Greatly reduced penetrance and expressivity–that is, where only a small fraction of carriers suffer from the condition–leads to the concept of genetic predisposition (susceptibility) for disease, which may be modulated by other genes or environmental factors. Polygenic and multifactorial inheritance are the etiology of most of the common diseases (such as coronary heart disease, obesity, and diabetes mellitus) and the majority of cases of birth defects. In these categories, empirical risk figures apply.

Numerical chromosomal aberrations are usually sporadic, with a low empirical recurrence risk in a sibship, but they may also run in families (e.g., structural aberrations, such as translocations, which may be inherited as balanced or unbalanced in offspring in a more or less apparent Mendelian fashion).

Genetic Counseling in the Social Context

How has genetic counseling and its supposed high standards fared in common practice? Throughout the world, the professional identity and training of providers of genetic counseling services varies. In Germany, for example, most counseling services are provided by physicians who are specialists in medical genetics (Nippert et al., 1997), while in many others (such as the United States, Canada, South Africa, Japan, France, the Netherlands, and the UK) a multidisciplinary team (including medical doctors, genetic counselors, and genetic nurses) is involved. While genetic counseling as an activity can be provided by professionals from a number of backgrounds, genetic counselors are specifically trained nonmedical professionals. A recent overview of the genetic counselor and genetic nurse in Europe (Cordier et al., 2011) indicated that differences in professional backgrounds and relative numbers of practitioners existed between countries. For example, in some European countries (e.g., France, the UK, Romania, Spain, and Portugal), genetic counselors are educated through a master's program in genetic counseling, while in others no such program exists and practitioners may have a wide range of backgrounds, including laboratory genetics, nursing, or social work. In Europe, only the UK has a national system for registration of genetic counselors. In the United States and Canada specialist genetic counselors must complete an approved master's degree in genetic counseling and be certified by the American Board of Genetic Counseling or the Canadian Association of Genetic Counselors, while a similar system exists in Australia and South Africa.

Despite differences in education and background, a survey of 216 European genetic nurses and counselors showed that clinical practice was remarkably similar across countries (Skirton et al, 2013). The majority of respondents stated that they alone or with a medical colleague took responsibility for making the first contact with the family (87.9%), drawing the pedigree (85.2%), explaining a genetic test to the patient (79.5%), and providing psychological support through the testing process. Over 81% managed some cases without the input of a medical doctor.

While published policies and guidelines on provision of genetic services have existed for some time (Fraser, 1974; German Society of Human Genetics, 1996; Andrews et al., 1994; WHO, 1998; Fryer and Lister Cheese, 1998; Holtzman and Watson, 1997), the type of information and the way information was communicated was found to vary within and between countries, and standard practices developed only slowly. Obviously, genetic counseling practices were shaped by the characteristics of the various health care systems, as well as by the personal characteristics of the counselor, such as sex, professional expertise, and type of employment. The process and outcomes of genetic counseling have been documented by a few empirical studies (Sorenson, 1992). Initially, outcomes were defined primarily on the basis of patient knowledge and patient satisfaction (Clarke et al., 1996), but this has altered in recent years. Researchers now approach the topic in a holistic manner to determine the outcomes of importance to patients and professionals (Skirton, 2001, Skirton et al., 2005; McAllister et al., 2008a; 2008b). These studies have shown that genetic counseling can help individuals adapt psychologically to their circumstances and enhance their sense of empowerment.

Concern has been voiced that in the future more counseling will be provided by professionals not trained in genetics. Although there is insufficient empirical evidence that genetic counselors successfully achieve autonomous, informed decision making, research from Europe and the United States suggests that nongeneticists do not set a high value on providing information that allows informed decision making before undergoing tests such as prenatal diagnosis (Dodds, 1997) or predictive testing (Burke, 2004).

In recent years, an initiative to harmonize the standards and practices of genetic counseling was started within the EU-funded Network of Excellence 'Eurogentest' (www.eurogentest.org). It included a survey of legislation regarding genetic counseling in European member states (http://www.eurogentest.org/professionals/info/public/unit3/regulations.xhtml; Soini (2012)), a survey of international bodies including political institutions, professional organizations, ethical boards, and patient associations (http://www.eurogentest.org/web/files/public/unit3/summaryofguidelinesMay06.pdf), and a survey among European National Societies of Human Genetics (http://www.eurogentest.org/web/files/public/unit3/Results_of_survey_1_WP_3-1_Dec06.pdf). What emerged from these activities were the Eurogentest "Recommendations for Genetic Counseling Related to Genetic Testing" (http://www.eurogentest.org/professionals/info/public/unit3/final_recommendations_genetic_counselling.xhtml). These activities respond to a widely felt need for standardization of genetic counseling practice, thereby improving the quality of genetic counseling especially in the context of old and new genetic testing options.

With the development of more new molecular tests, implementation of these tests in various fields of medicine is expected to drive the diversification of genetic counseling practice. While at present many general health care providers and nongenetics specialists do not have sufficient professional education and expertise in genetics, molecular testing, and genetic counseling to perform this task as shown by the European study on genetics education (Harris et al., 2006), there are certainly specialists in other areas who are providing effective and appropriate genetic counseling services and genetic testing in relation to their own fields. For example, in oncology, cardiology, and diabetes care, presymptomatic testing of those at risk has enabled resources, clinical surveillance, and treatment to be targeted at those who will benefit from them.

However, there is an issue around provision of education for providers via medical curricula and continuing education. Genetic specialists are arguably the most valid source for providing the needed expertise and information for their medical colleagues to facilitate their understanding of basic genetic principles and the application of molecular testing, including pre- and post-test counseling, in clinical practice.

However, the more genetics becomes part of general medicine and the more genetic testing is introduced into the various specialty areas, the more the task to provide information and to interpret test results will fall on primary care physicians and nongenetics specialists. Indeed, it could be claimed that primary care is the appropriate setting in which to utilize the power of genomic tests to create more truly personalized medicine (Gibbon, 2009). This requires greater genomics literacy among those health professionals in both secondary and primary care.

In its Federal Act on Genetic Testing put into force in 2010, Germany, for example, has not only implemented a legal obligation for all medical doctors to recommend to their patients genetic counseling in the context of genetic testing but also to undergo training in genetic counseling themselves (via continuous medical education measures), unless they refer to specialized genetic counselors. Given that genetic test utilization is rising steeply (Schmidtke et al., 2005) and genetic centers are sparse and understaffed, such referrals may become restricted to complicated cases, while the majority of genetic counseling may be given from nongenetic specialists.

The European Society of Human Genetics has recently published "Core Competences in Genetics for Health Professionals in Europe" (https://www.eshg.org/139.0.html), in which many of the skills currently expected from specialized genetic counselors are stated to be core competences in primary and secondary nongeneticist specialists' care. These competences are already being adopted for use in preparing specialists in other settings (e.g., nurses working in hemoglobinopathy or inherited bleeding disorder clinics in the UK) to offer appropriate care to patients. Importantly, the core competences stress the need for appropriate information, patient choice, and ethical practice, whatever the clinical context in which genetic health care is offered.

Most likely, the future of genetic testing and counseling will be shaped by economic interests; there is of course the need for financial returns on heavy investments made in the field of molecular testing, by reimbursement practices, and by the extent to which national health care policies will encourage genetic testing. It remains unclear whether the main ethical concepts attached to genetic counseling, such as nondirectiveness, individual autonomy, and greater equality in the doctor–patient relationship, will survive or will wither away during the integration of genetic tests into routine medical practice. Hopefully, genetic counseling will retain some of its best elements, namely, respect for patients' rights to valid, nonslanted information and for patient autonomy in decision, making.

See also: Ethical Dilemmas, Research, and Access to Treatment; Ethics and Psychiatry; Eugenics, History of; Genetic Screening for Disease; Genomics, Ethical Issues In; Risk Screening, Testing, and Diagnosis: Ethical Aspects.

Bibliography

Ad hoc Committee on Genetic Counselling of the American Society of Human Genetics, 1975. Genetic counselling. American Journal of Human Genetics 27, 240–242.

Assessing genetic risks. In: Andrews, L.B., Fullarton, J.E., Holtzman, N.A., Motulsky, A.G. (Eds.), Implications for Health and Social Policy. National Academy Press, Washington, DC.

Bartels, D.M., LeRoy, B.S., McCarthy, P., Caplan, A.L., 1997. Nondirectiveness in genetic counseling. Journal of Genetics Counseling 72 (2), 172–179.

Beauchamp, T.L., Childress, J.F., 2001. Principles of Biomedical Ethics, fifth ed. Oxford University Press, ISBN 0-19-514332-9.

Burke, W., 2004. Genetic testing in primary care. Annual Review of Genomics and Human Genetics 5, 1–14.

Clarke, A., Parsons, E., Williams, A., 1996. Outcomes and process in genetic counselling. Clinical Genetics 50, 462–469.

Cohen, P.E., Wertz, D.C., Nippert, I., Wolff, G., 1997. Genetic counselling practices in Germany: a comparison between East German and West German geneticists. Journal of Genetic Counselling 6, 61–80.

Cordier, C., Lambert, D., Voelckel, M.A., Hosterey-Ugander, U., Skirton, H., 2011. A profile of the genetic counsellor and genetic nurse profession in Europe. Journal of Community Genetics. http://dx.doi.org/10.1007/s12687-011-0073-x Published online 14 December 2011.

Dodds, R., 1997. The Stress of Tests in Pregnancy: Summary of a National Childbirth Trust Screening Survey. National Childbirth Trust, London.

Fraser, F.C., 1974. Genetic counselling. American Journal of Human Genetics 26, 636–659.

Fryer, A.E., Lister Cheese, I.A.F., 1998. Clinical Genetic Services: Activity, Outcome, Effectiveness and Quality. The Royal College of Physicians of London Report from the Clinical Genetics Committee of the Royal College of Physicians of London, London.

Gibbon, S., 2009. Genomics as public health? Community genetics and the challenge of personalised medicine in Cuba. Anthropology and Medicine 16 (2), 131–146.

Harper, P., 1998. Practical Genetic Counselling, fifth ed. Butter-worth-Heinmann, Oxford, UK.

Harris, R., Challen, K., Benjamin, C., Harris, H., 2006. Genetic education for non-geneticist health professionals. Community Genetics 9 (4), 224–226.

Holtzman, N.A., Watson, M.S., 1997. Promoting Safe and Effective Genetic Testing in the United States. Final Report of the Task Force on Genetic Testing, Johns Hopkins University Press, Baltimore.

Human Genome Project, http://www.ornl.gov/TechResources/Human_Genome/home.html

Kessler, S., 1997. Psychological aspects of genetic counselling XI: nondirectiveness revisited. American Journal of Medical Genetics 72, 164–171.

Martin, D.K., Greenwood, H.L., Nisker, J., 2010. Public perceptions of ethical issues regarding adult predictive genetic testing. Health Care Analysis 18, 103–112.

McAllister, M., Payne, K., Macleod, R., Nicholls, S., Dian, D., Davies, L., 2008a. Patient empowerment in clinical genetics services. Journal of Health Psychology 13 (7), 895–905.

McAllister, M., Payne, K., Macleod, R., Nicholls, S., Donnai, D., Davies, L., 2008b. What process attributes of clinical genetics services could maximise patient benefits? European Journal of Human Genetics 16 (12), 1467–1476.

Müller-Hill, B., 1997. Murderous Science: Elimination by Scientific Selection of Jews, Gypsies and Others in Germany 1933–1945. Oxford University Press, Oxford, UK.

Nippert, I., Horst, J., Schmidtke, J., 1997. Genetic services in Germany. European Journal of Human Genetics 5 (Suppl. 2), 81–88.

Ploetz, A., 1895. Racial hygiene. In: Weingart, P., Knoll, J., Bayertz, K. (Eds.), Rasse, Blut und Gene. Frankfurt/Main. Suhrkamp, Germany.

Resta, R., Biesecker, B.B., Bennett, R.L., Blum, S., Hahn, S.E., Strecker, M.N., Williams, J.L., 2006. A new definition of genetic counselling: national Society of genetic counselors' task force report. Journal of Genetic Counselling 15 (2), 77–83.

Schmidtke, J., Pabst, B., Nippert, I., 2005. DNA-based genetic testing is rising steeply in a national health care system with open access to services: a survey of genetic test use in Germany, 1996-2002. Genetic Test 9, 80–84.

Schwartz, M.D., Isaacs, C., Graves, K.D., Poggi, E., Peshkin, B.N., Gell, C., Finch, C., Kelly, S., Taylor, K.L., Perley, L., 2012. Long-term outcomes of BRCA1/BRCA2 testing: risk reduction and surveillance. Cancer 118 (2), 510–517.

Soini, S., 2012. Genetic testing legislation in Western Europe – a fluctuating regulatory target. Journal of Community Genetics 3, 143–153.

Skirton, H., 2001. The client's perspective of genetic counselling - a grounded theory study. Journal of Genetic Counselling 10 (4), 311–329.

Skirton, H., Cordier, C., Lambert, D., Lambert, D., Hosterey Ugander, U., Voelckel, M.A., O'Connor, A., 2013. A study of the practice of individual genetic counsellors and genetic nurses in Europe. Journal of Community Genetics 4 (1), 69–75.

Skirton, H., Parsons, E.P., Ewings, P., 2005. Development of an audit tool for genetic services. American Journal of Medical Genetics 136 (2), 122–127.

Sorenson, J.R., 1992. What we still don't know about genetic screening and counselling. In: Annas, G.J., Elias, S. (Eds.), Gene Mapping: Using Law and Ethics as Guides. Oxford University Press, New York, pp. 203–214.

Stern, A.M., 2005. Eugenic Nation: Faults and Frontiers of Better Breeding in Modern America. University of California Press, Berkeley, CA.

Vogel, Motulsky, 2010. In: Speicher, Michael, Antonarakis, Stylianos E., Motulsky, Arno G. (Eds.), Human Genetics. Springer, Berlin Heidelberg, New York.

Weil, J., Ormond, K., Peters, J., Peters, K., Biesecker, B.B., LeRoy, B., 2006. The relationship of nondirectiveness to genetic counselling: report of a workshop at the 2003 NSGC Annual Education Conference. Journal of Genetic Counselling 15 (2), 85–93.

Wertz, D.C., 1997. Society and the not-so-new genetics: what are we afraid of? Some future predictions from a social scientist. Journal of Contemporary Health Law Policy 13, 299–346.

Wertz, D.C., 1998. Eugenics is alive and well: a survey of genetic professionals around the world. Science in Context 11, 493–510.

Wolff, G., Jung, C., 1995. Nondirectiveness and genetic counselling. Journal of Genetic Counselling 4, 3–25.

Relevant Websites

https://www.eshg.org/139.0.html – European Society of Human Genetics.

www.eurogentest.org; http://www.eurogentest.org/professionals/info/public/unit3/regulations.xhtml; http://www.eurogentest.org/web/files/public/unit3/summaryofguidelinesMay06.pdf; http://www.eurogentest.org/web/files/public/unit3/Results_of_survey_1_WP_3-1_Dec06.pdf; http://www.eurogentest.org/professionals/info/public/unit3/final_recommendations_genetic_counselling.xhtml – Eurogentest.

http://www.gfhev.de – German Society of Human Genetics, 1996. Position paper.

http://www.who.int/genomics/publications/en/ethicalguidelines1998.pdf – World Health Organization, 1998. Proposed International Guidelines on Ethical Issues in Medical Genetics and Genetic Services. World Health Organization Human Genetics Programme, Geneva.

http://www.who.int/genomics/elsi/gentesting/en/index.html – World Health Organization, 2012.

Genetic Engineering

David J Harris, Boston Children's Hospital, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The techniques of genetic engineering have been used to determine elements of the genome of an organism, as well as copy, isolate, and identify single genes in many species including our own. Transferring human genes to other organisms allows investigators to understand the proteins and study the pathogenesis of disease. This approach is being used to either treat diseases by gene therapy or identify targets for drug development.

The term 'genetic engineering' stands for human alteration of the genetic code of an organism, so that its biosynthetic properties are changed. The major applications are for the industrial production of desired peptides or proteins or to alter the biological capabilities of the organism. These techniques have been used to develop crops with agronomically useful changes, such as pest resistance and ripening properties that allow for shipment. On the medical front, peptide hormones such as insulin have been produced as well as larger proteins to perform essential functions in individuals with disorders caused by mutations that act by either stopping the synthesis of a protein or producing an ineffective structure. There is a lot of interest in developing such products.

By silencing genes in experimental organisms, as well as by producing animal models of human disease, we have gained information about the pathogenesis of structural and biochemical disorders. Single human genes that follow classic Mendelian patterns are still being identified, however, more attention is being paid to complex traits. Most investigators agree that behavioral phenotypes fit this category.

Biological Background

The technology has developed in the context of the understanding of genetics. In order for a species to survive and evolve, variability must be produced and maintained in the population. One of the mechanisms that produces such variability is recombination. Information for the amino acid sequence of a number of peptides as well as the controlling mechanisms for regulating their synthesis are inherited together on a single chromosome. In diploid organisms, exchanges between homologous chromosomes are an important source of variation. This process occurs in the context of DNA replication, including the enzymatic unwinding of the double helix and synthesis of the new, complementary chain of nucleic acids.

Cellular properties may be altered by external mechanisms. One of the first systems that was extensively studied, and provided the tools that are currently used, is the phenomenon of infection of bacteria by bacteriophages, the bacterial viruses. These are the usual steps in protein synthesis. The message is coded in the DNA sequence. It is then transcribed into the sequence of the intermediate messenger RNA molecules, and then translated into protein. In the 'lytic cycle,' the viral DNA instructions are used to synthesize new viral DNA and protein that can assemble into new viral particles that destroy the cell and are then released into the environment to infect other host cells. In the 'lysogenic cycle,' the DNA of the virus integrates into the host DNA. In addition, there are a number of other particles called plasmids, which can transfer genetic information from one bacterium to another. They are small, circular DNA molecules that replicate independently of the DNA of the host bacterium. Their role is particularly prominent in the transmission of antibiotic resistance. The life cycles of most animal viruses are similar to the lytic cycle of bacteriophage, however, there are a number of viruses that can become latent. There are also agents like the retroviruses, whose genome gets transcribed into the host DNA, thereby changing cellular properties. Some of these agents will also incorporate host DNA into their sequences and transmit this sequence into a new host.

The tools that allow us to transfer genes from one organism to another are the restriction endonucleases. These enzymes are found naturally in bacteria and protect them from infection by breaking down foreign DNA. They are chemically very specific, only working on specific palindromic sequences. Consequently, the kind of infection discussed in the last paragraph can only occur if the bacteriophage sequence does not contain elements that will be broken down by the host bacterial system of endonucleases. A number of different enzymes are found in different bacterial species, and the enzymes referred to by an abbreviation of the bacterial species name and a number. Similarly, there are enzymes that are used to connect fragments, the ligases (Griffiths et al., 2010).

Cloning Genes

All manipulations of genetic engineering require multiple copies of the DNA sequence or gene of interest. The original methods of getting multiple copies relied on bacteriophage or plasmid vectors to introduce the foreign DNA into bacteria to produce these copies, as each modified cell produces multiple copies, and the bacterial culture itself increases. This is done by first physically isolating the vector, opening its DNA with a restriction enzyme, and binding in DNA from the organism being studied that has also been cleaved with a restriction endonuclease. A new population of bacteria is then infected with the altered vector. Given an appropriate way of selecting the population of bacteria so that it uniformly has the DNA of

interest multiplying within, one can isolate a large population of vector molecules with the desired sequence, which is then freed by enzymatic cleavage once again.

Fragments of DNA are identified by physically separating them by electrical charge and molecular weight through gels. The DNA of the vectors and bacteria are generally in the range of 1–10 000 bp, and there are a sufficiently small number so that the fragments can be identified with a simple staining technique, usually a compound that binds to DNA and fluoresces under ultraviolet light. The larger quantity of fragments that would be isolated from more complex organisms produces a smear with such dyes, so the base-pairing property of DNA, the obligate pairing of adenine with cytosine and guanine with cytosine that allows for both recognition and synthesis of the linear sequence, is used to identify the same sequence on the gel by labeling a known fragment with an isotope or fluorescent dye. The labeled molecules are called probes. This is also the basis for identifying genetic variation in organisms, either for basic studies or identification of mutations associated with disease.

Isolation of fragments produced by digestion with several enzymes, used both singly and in combination, allows for the construction of a physical, restriction fragment map. Smaller fragments may be replicated, followed by the chemical analysis of the base sequence within fragments, which are then assembled into the final base sequence of the gene. Once the sequence is known, production of useful amounts of a region of DNA may now be done enzymatically *in vitro* with the polymerase chain reaction. In this technique, the region between two primers, one from each strand of the final DNA molecule is copied in a logarithmic fashion by a heat-resistant DNA polymerase from a small amount of genomic DNA (it has been done with single cells), using multiple heating and cooling cycles. This technique is also used extensively in diagnostic work (Strachan and Read, 2010).

Modifying Cells to Produce Proteins

A number of strategies have been used to produce large amounts of peptide or protein. Multiple techniques have been necessary because of the different kinds of cells that have been used. The production of human insulin, one of the earlier products produced commercially, is a useful model, which is an example of 'expression cloning.' The sequences for the A and B chains were introduced into separate vectors in sequence with a bacterial promoter (a signal to turn a gene on) and the enzyme β-galactosidase that produces colored bacterial colonies. This arrangement is necessary because bacteria will not just produce the human protein if only the structural gene is introduced. The bacteria are tricked into producing a fusion protein: a bacterial β-galactosidase with a human insulin chain sequence at the end. After selecting the colonies by color, mass cultures are produced from which the fusion proteins can be purified. The normal insulin dimer is then produced chemically by removing the β-galactosidase, purifying the chains, combining the chains, and allowing them to refold and form the sulfide bridges between the two chains.

Expression cloning in bacteria works well if the peptide or protein is composed entirely of the amino acid sequence. Many functional proteins contain specific carbohydrate groups at precise locations on the chain that are attached after translation. These modifications are different in different classes of organisms, so techniques using yeast, plant, cultured insect, other mammalian, or human cells have been used to make the complete molecule. Another problem that is addressed in this approach is the elimination of bacterial toxins that cannot be removed from the desired protein product mixture. The engineered DNA can be introduced into these cells by a variety of methods. Physical techniques to get the DNA inside the cell include electroporation that employs a current, bombarding the cells with coated particles such as gold, or calcium salts, or heat. Lipid droplets are also being tried. Artificial yeast chromosomes have been produced to move gene constructs. For the cultured insect cells, modified baculoviruses are used (these are the viruses that have invertebrate hosts). In mammalian systems, several viruses have been used, including retroviruses (the class of agents involved in acquired immunodeficiency and tumor transformation in experimental organisms), adenoviruses (agents that typically produce respiratory disease, however, some strains are capable of producing more severe systemic disease), and herpes simplex.

The goal of expression cloning is to produce large amounts of protein that can be purified more easily and in much larger amounts than is possible using chemical techniques starting with cells from the normal tissue. On the research laboratory scale, the three-dimensional structure of the whole protein or the active region can be determined by a combination of nuclear magnetic resonance and X-ray crystallography. This information is useful for understanding function or designing small molecules that are expected to have pharmacological actions. This is then one element of the strategy in rational drug design. If large quantities of the protein is to be used as a product, the growth of the modified cells has to be done on an industrial scale that is an engineering problem similar to that used in the production of a number of antibiotics (Glick and Pasternak, 2003).

Modifying Organisms

An organism that has been engineered so that it has a gene from another species integrated into its genome is called a 'transgenic organism.' These techniques have been used to produce plants and animals expressing mutated genes of that species to study function, with mutated or genes introduced from another species with desired agronomic properties, as models of human diseases, and to produce therapeutic products. There is a large and continuously growing literature of studies in the mouse.

Two procedures have been widely performed in mice: the interruption of a mouse gene to eliminate its function, for which the term knockout has been used, and introduction of a transgene. The 'standard' procedure employs classical, cellular, and molecular genetic techniques. The gene to be studied is first cloned and then introduced into a targeting vector in such a way that the coding sequence of the gene is interrupted by the neomycin-resistance gene. The vector also has the thymidine kinase gene from herpes. Embryonic stem cells are obtained from a brown mouse and placed in culture. The cells are then exposed to the vector. There are three possible outcomes. The

vector does not insert at all, the whole construct inserts at random in a chromosome, or only the desired coding sequence and the neomycin-resistance gene insert into the normal location on the chromosome. Treating the cells with a neomycin analog and the antiviral agent ganciclovir will kill all the cells except the ones with the integrated sequence. The surviving stem cells are injected into the blastocyst stage embryo of a black mouse and placed in a surrogate mother. If a chimera is formed, it will have a coat of two colors. The chimeras are mated with black mice, and their brown offspring tested for the modified gene. Those that are heterozygous are mated, and one of four will be homozygous for the modification.

In addition to the introduction of genes into stem cells by vectors, injection of the desired sequence has been done by some of the physical methods discussed earlier such as electroporation, coated gold particles, and salt precipitation. Some transgenics have been derived from direct injection of the plasmid into the egg.

A newer experimental system, namely the use of knockdown of functions of genes by the use of morpholinos, i.e., synthetic oligonucleotides, has been used in zebra fish to study the role of mutations identified in human subjects. This system allows manipulation in a vertebrate that is more rapid and economical than a mouse.

The production of genetically identical organisms, cloning in the original sense, was first done in amphibians by nuclear transplantation into fertilized ova from which the original nucleus was removed. This was done widely in mammals by using embryonic cells as the nuclear donor. The production of the sheep named Dolly was noteworthy because the donor nuclei were obtained from adult mammary cells that were maintained in cell culture. This precipitated a round of discussion about the feasibility and the desirability of doing this in human beings. The claim was made that this would be an alternative for infertile couples and therefore acceptable. It is necessary to point out to people who would propagate themselves for narcissistic reasons that the derived individual would not be absolutely the same because of the developmental environment, both before and after birth (Hubbard and Wald, 1997).

Another possible use of this technology that has been proposed is the therapeutic cloning of replacement tissues or organs. Cells would be biopsied from the eventual recipient and placed in culture. Nuclei from these cells would be transplanted into eggs from which the nuclei have been removed; these eggs will be allowed to develop to the blastocyst embryonic stage for the isolation of embryonic stem cells that are treated to produce the desired cell types. This avoids the problem of using fertilized eggs but still involves the use of human eggs and early embryonic stages (Gurdon and Colman, 1999).

Gene Therapy

The goal of gene therapy is to provide a functional protein to an individual whose genotype leads to a disease because the protein is missing or modified so that it is not functional. There are two potential modes. In 'germ line therapy,' the modification would be heritable, and the individual's offspring potentially would not be affected. In 'somatic cell therapy,' only the treated individual would have the modification. There have been many arguments against germ line therapy including our inability to know whether potentially harmful changes in the sequence have been made inadvertently. All the trials to date have been in somatic cells for the treatment of inherited disease and cancer by control of cell death.

Two strategies are being attempted. For those disorders for which bone marrow transplantation has been effective, the gene is introduced into marrow cells, the population of which is then expanded and infused into the patient. The techniques for gene introduction are similar to the ones previously discussed: viral vectors and chemical or physical techniques. An immune deficiency disorder, adenosine deaminase deficiency, was the first attempt reported. Although the modified cells express the enzyme, repeated infusions of treated cells have been necessary, and the patients have continued to receive chemically modified enzyme, so it is not clear whether or not gene therapy alone is sufficient.

Other trials have used viral vectors to treat target organs, such as the lung in cystic fibrosis. The therapeutic trail of an adenovirus-mediated transfer of a gene to the liver has received general media attention because of the death of a participant. Evidently, other participants in gene therapy trials have died, although it is not clear how many of those deaths are a result of the procedure (Strachan and Read, 2010). Currently, there continue to be trials for immunodeficiency diseases, β-thalassemia, hemophilia, several forms of leukodystrophy, and retinal dystrophy.

Applications to Behavior

The application of these techniques in behavior is loaded with scientific, ethical, and legal complications. Part of the problem stems from the very human desire to have a simple answer to a complex problem. Genetic, environmental, and developmental factors interact to produce behavior, and it is difficult to distinguish the role of each component. One should ask if it is even desirable to do this because of the mixture of potential benefit and risk. It is clear that each behavioral trait is heterogeneous. Mutations of a single major gene resulting in deficiency of adenylosuccinase produces an autistic phenotype. For this class of disorder, a particular drug, expressed protein, or even gene therapy could work, but in the more usual instance of several genes and environmental factors, it is unlikely that such a unitary approach would be effective. The promise of a solution to disease problems by definition of molecular mechanisms has been difficult to achieve. We still do not have an answer in the case of sickle-cell anemia, the first disorder to be described in this manner. Common behavioral traits would be more difficult because of the nature of gene–-gene interactions as well as gene–environment interactions.

The example in the previous paragraph employed both standard pedigree analysis as well as biochemical techniques. The expansion of technology in the last few years has given new methods of genetic analysis and a new insight into human genetic variability, as well as possible explanations of mental disorders. The first technique is the chromosomal microarray that identifies changes from the standard, diploid copy number in regions of the genome. From cytogenetic analysis, such

changes had been associated with disorders that include mental retardation as part of the condition. Both additions, i.e., duplications, and subtractions, i.e., deletions, had been described. The array allows discovery of much smaller duplications and deletions. The smaller of these are often harmless, inherited traits. Several of these have been associated in multiple studies with autism spectrum disorder and schizophrenia. These genetic changes as yet only account for a minority of the affected individuals (Smith, 2011).

The new sequencing technology expands on the observations in the previous paragraph. This has allowed more identification of mutations of single, Mendelian genes as well as the existence of very complex chromosomal rearrangements (Talkowski et al., 2012). A pattern of new, detrimental mutations of genes involved with the development and maintenance of synapses is emerging in individuals with developmental issues and autism (Neale et al., 2012).

The technology therefore will create the possibility of diagnostic tests for the presence of variants at these genetic loci that might be involved in diagnostic confirmation, or even asymptomatic screening without necessarily having the ability to safely and ethically treat those in whom the genetic variant is identified. Furthermore, since it is well known that the genotype does not always predict the phenotype even for such things as emphysema associated with a variant α-1-antitrypsin, one should question this approach with respect to behavioral characteristics (Hubbard and Wald, 1997).

The second group to consider is the collection of positive and negative personality or behavioral traits that have been considered by educators and others interested in social welfare. This list of characteristics has included aggressiveness, criminality, laziness, or on the other side of the spectrum, altruism. This author is not aware of any studies that clearly show inheritance in a Mendelian pattern. The clustering in families may be biologic in part, but there are many social and environmental influences in these behaviors. A good explanation for part of the variability is cultural transmission, or learning (Creanza et al., 2012). Using genetic information to design pharmacologically active agents is being used for drugs interacting with receptors influencing neural function. A similar problem arises here. Will the intervention be beneficial to someone who is suffering or will it be used to control individuals who are designated by those in authority to be deviant? Modifying the environment to enhance development is effective and poses less of a biological risk than drugs or gene therapy (Duyme et al., 1999). The relative economic costs are unknown, but the human ramifications are immense.

See also: Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetics and the Media; Genomics, Ethical Issues In.

Bibliography

Creanza, N., Fogarty, L., Feldman, M.W., 2012. Models of cultural niche construction with selection and assortative mating. PLoS One 7 (8), e42744. http://dx.doi.org/10.1371/journal.pone.0042744.

Duyme, M., Dumaret, A.-C., Tomkiewicz, S., 1999. How can we boost IQs of 'dull children'?: a late adoption study. Proceedings of the National Academy of Science USA 96, 8790–8794.

Glick, B.R., Pasternak, J.J., 2003. Molecular Biotechnology: Principles and Applications of Recombinant DNA, third ed. ASM Press, Washington, DC.

Griffiths, A.J.F., Wessler, S., Carroll, S.B., Doebley, J., 2010. Introduction to Genetic Analysis, tenth ed. W. H. Freeman, New York.

Gurdon, J.B., Colman, A., 1999. The future of cloning. Nature 402, 743–746.

Hubbard, R., Wald, E., 1997. Exploding the Gene Myth: How Genetic Information is Produced and Manipulated by Scientists, Physicians, Employers, Insurance Companies, Educators and Law Enforcers. Beacon Press, Boston.

Neale, B.M., et al., 2012. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485, 242–245.

Smith, M., 2011. Phenotypic Variation: Exploration and Functional Genomics. Oxford University Press, Oxford.

Strachan, T., Read, A.P., 2010. Human Molecular Genetics, fourth ed. Wiley-Liss, New York.

Talkowski, M.E., et al., 2012. Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. Cell 149 (3), 525–537.

Genetic Factors in Organizational Psychology

Nikolaos Dimotakis and Jeff Schatten, J. Mack Robinson College of Business, Georgia State University, Atlanta, GA, USA
Richard Arvey, School of Business, National University of Singapore, Singapore

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Organizational psychology has increasingly integrated perspectives and findings from genetic approaches. In this article, we provide an overview of genetics-based investigations on variables of interest to organizations, such as attitudes, personality, and affectivity. We also discuss synthesizing perspectives and boundary conditions for genetic effects. Finally, we discuss ethical and legal issues surrounding genetic research in organizations, and suggest possible avenues for future research.

Introduction

Organizational research has been increasingly interested in the role of genetics in influencing the ways individuals think, feel, and behave. Over the past three decades, this program of research has become increasingly nuanced, detailed, and programmatic. Without doubt, these investigations have provided valuable insights on how genetic factors shape organizational outcomes of interest such as job satisfaction (Arvey et al., 1989), job choice (McCall et al., 1997), and leadership role occupancy (Arvey et al., 2007; Ilies et al., 2004). These developments are not limited to examinations of outcomes – other approaches have provided insights into antecedents and processes, ranging from personality dispositions (Ilies and Judge, 2003) to task persistence (McCartney and Berry, 2005) and information processing (Caldu and Dreher, 2007). This body of work has provided valuable insights, broadened our knowledge, and honed our predictive ability.

The success of these efforts is perhaps not surprising, as "individual differences in virtually all domains have a genetic component" (Ilies et al., 2006). Indeed, models that have integrated previously established personality and affectivity perspectives with genetic or heritable arguments and evidence (e.g., see Ilies and Judge, 2003) have demonstrated the importance of these factors for a variety of approaches that utilize individual differences as predictors or boundary conditions in the effects of interest. That is, because most traits, whose importance is established in the organizational literature have a genetic component, these components and mechanisms become a valuable tool of inquiry that helps us understand the ways people feel, think, and behave in (as well as feel, think, and behave in response to) their environments.

Apart from their important role in advancing the accumulation of knowledge and scientific consensus, these efforts further provide important theoretical and practical contributions. By enhancing the organizational literature with a theoretical background based on biological mechanisms, they set differences such as the personality/situation or nature/nurture debates in an arena where they can be more closely examined empirically, allowing us to do so with a better understanding of the variance to be explained and the processes behind such explanations (Ilies and Judge, 2003). Relevant to organizations, such an approach allows us to examine a variety of factors relevant to between-people as well as dynamic models of employee functioning, and potentially design better ways to intervene.

This article thus seeks to contribute to these efforts in general by providing an overview and synthesis of such research, adding to and supplementing prior synthesizing work in this area (see Ilies and Judge, 2003; Ilies and Dimotakis, in press; Song et al., in press). Furthermore, we aim to identify fruitful areas of future research, and provide suggestions for further integration. Finally, we discuss ethical and legal implications that integrating genetic factors into organizational research can have.

We begin with an overview of the literature, which we will generally summarize under a history of research on individual differences (including personality, cognition, and affectivity) as well as attitudes and behaviors. We then progress into a discussion of studies integrating individual differences with organizational outcomes, and discuss more recent and nuanced approaches to these areas of inquiry. Finally, we present an overview of recent issues of legal and ethical concern that are becoming increasingly relevant for all fields that border the genetic field. In each section, we will generally group the literature under organizationally important antecedents, processes, or outcomes, according to commonly utilized models in the field.

Overview of Genetic Research

Organizational research on genetic factors attempts to understand and explain where variation in individual variance in traits of interest arises; it seeks to establish the role of the environment and genetics in influencing a variety of individual factors. In the organizational research realm, such work aims to uncover the role of individual genotypes in a host of factors that matter. Such factors include individual differences (such as personality and cognitive ability), attitudes (such as job satisfaction), and behaviors (such as performance or leadership).

At its most basic form, this line of research aims to partition the variance in some construct into genetic and environmental components. For example, when examining a specific attitude, research has aimed to uncover the proportion of variance in this attitude that can be accounted for by genetic differences, with one such estimate being the parameter of heritability (Ilies et al., 2006; McGue and Bouchard, 1998). These influences can

be further decomposed. Genetic influences on attitudes, personality, or other traits can be separated into additive genetic effects (transmissible across generations) and *nonadditive genetic effects* (not transmissible). The effects of environmental variation on individual traits can also be further decomposed into those associated with *shared environments* (e.g., common experiences that individuals might have been exposed to by, for example, being raised in similar environments), and into these effects associated with *unshared environments* (i.e., environmental influences that are unique for each individual).

The proportion of variance in a trait accounted for by the influences described above can be estimated by examining similarities and differences among twins. In the simplest version, examining the correlation in a trait among monozygotic twin pairs that were raised apart will provide an initial estimate of genetic influences; as these twins are genetically identical, but were exposed to different circumstances growing up, any similarities between them must be largely attributable to genetic factors.

More sophisticated models utilize latent factors to compare the covariance in the traits of interest among four types of twin pairs; monozygotic twins raised together, monozygotic twins raised apart, dizygotic twins raised together, and dizygotic twins raised apart. This allows for the partitioning of variance in the variables of interest into additive genetic variance, shared environmental variance, and nonshared environmental variance (see Ilies and Judge, 2003). These models are then compared with alternatives to compare fit, and the best fitting approaches are used to estimate variance components. Recently, more sophisticated models are beginning to emerge that utilize multiple groups of related and nonrelated individuals to better examine the variance attributable to shared and nonshared factors – these models are showing great promise (see Hahn et al., 2013, for an example of this approach).

Furthermore, other efforts focus on the molecular level by examining the associations of distinct markers with personality, attitudinal or behavioral phenotypes. This approach involves directly examining the effects of the presence of a specific allele on the variable of interest instead of examining the effects of genetic factors overall (see Plomin et al., 2008, for an overview; Song et al., 2011, for an example of this approach). Finally, as with other analytic approaches, estimates from all of the above approaches can be further subjected to moderation and mediation analyses to either examine boundary conditions for genetic influences (or lack thereof), and to examine possible mediating variables in the causal chain (as we will discuss below).

Individual Differences in Personality, Cognition, and Affectivity

We begin our review of genetic factors in organizational research with an overview of individual differences, which include personality, cognitive ability, and affectivity. This is because these factors are commonly held to represent core antecedents of important organizational processes and outcomes, and as such, represent important starting points or foundations for subsequent research (see Ilies and Judge, 2003; Plomin, 1999). Even though their influence is by no means restricted to (or even mostly represented in) organizationally focused research, these factors remain central to models of workplace functioning.

In terms of personality, studies have been generally supportive of the importance of genetic influences in a variety of frameworks. For example, the personality dimensions included in the Big Five personality factors (see Goldberg, 1990) have been found to have mean narrow and broad heritability estimates of 0.34 and 0.45, respectively (Loehlin, 1992). Similar results, with an average heritability estimate of 0.41, were reported by Bouchard (1994) in an analysis of the Minnesota Study of Twins Reared Apart dataset. In terms of specific estimates, Jang et al. (1996) reported broad genetic influences of 41, 53, 61, 41, and 44% for Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness, respectively, utilizing a sample of 123 identical and 127 fraternal twins. These authors also reported significant estimates of personality facet heritability, although they did note that the latter showed significant influence of nonadditive traits. Similar genetic effects have been used when utilizing other personality frameworks, such as the Honesty and Humility (H) – Emotionality (E) – Extraversion (X) – Agreeableness (A) – Conscientiousness (C) – Openness to Experience (O) (HEXACO) model (see Veselka et al., 2010).

Cognitive ability has also been found to have a significant heritable component (Plomin, 1999). In a study utilizing the Minnesota Study of Twins Reared Apart, Bouchard et al. (1990) reported that 70% of the variance in cognitive ability could be attributed to genetic factors, while McClearn et al. (1997) report similar findings (62%) in an older sample of Swedish twins. However, such effects are not necessarily stable or universal. Haworth et al. (2009) report that the heritability of cognitive ability increases as people age, from 41% in childhood to 66% in young adulthood; such change is expected to persist or increase further with increasing age. Finkel et al. (1995) comparing twins from the Minnesota and Sweden samples found that estimates of heritability diverge in late adulthood between the samples. Turkheimer et al. (2003), on the other hand, report that socioeconomic status can affect heritability estimates in young children, with the vast majority of differences being accounted for not by environmental effects in impoverished families, but by genetic effects in affluent ones.

Finally, research has also examined the genetic component of affectivity, or the predisposition to experience positive and negative emotional states. The most commonly used framework in such examinations centers around Positive and Negative Affect (see Watson, 2000). Positive affect centers on feelings such as pride and happiness, while negative affect reflects anxiety and anger (Watson et al., 1999). These two factors have reflections in both personality (with extraversion corresponding to positive and neuroticism corresponding to negative affect; see Watson, 2000) as well as regulatory biobehavioral systems, with positive affect being reflected in the behavioral approach system and negative affect being reflected in the behavioral inhibition system (see Fowles, 1987; Gray, 1990; Watson et al., 1999). Research has demonstrated the importance of genetics for both the affectivity dimension as well as their associated factors. For example, Finkel and McGue

(1997) found that positive and negative emotionality have heritability estimates of 0.50 and 0.44, respectively, while Bouchard and McGue (2003) reported similar results (0.43 and 0.47). Furthermore, genetics can play a role in how affective experience varies across situations – that is, genetic factors are not only important for how individuals feel in general, but also for how they react affectively to different contexts. For example, Riemann et al. (1998) found that twin pairs exhibited similarities in affective experiences across different situations, in addition to the expected overall affective dispositions.

Attitudes and Values

Attitudes have long been represented in genetic and organizational research. In fact, attitudes such as job satisfaction represented the initial focus of research bridging these two areas. One of the first such studies was a report of significant intraclass correlations for general and intrinsic job satisfaction among monozygotic twins raised apart (Arvey et al., 1989). This work provided an initial indication of the importance of genetic factors in explaining workplace attitudes. Expanding on these findings, Arvey et al. (1994) reported that in two studies, utilizing samples from the Minnesota Twin Family Registry and the National Academy of Sciences and National Research Council twin sample, genetic factors accounted for significant variance in overall job satisfaction.

Job satisfaction studies have not always shown a genetic effect, however. Utilizing a sample of more than 500 monozygotic and dizygotic twins raised together and apart from the Swedish Adoption/Twin Study of Aging, Hershberger et al. (1994) reported no significant genetic influences on job satisfaction, although they did report significant genetic effects on other workplace-related attitudes and perceptions. Their item was, however composed of a one-item measure, meaning that unreliability could explain this lack of significance.

Naturally, research on attitudes has not been limited to job satisfaction. Tellegen et al. (1988) reported a heritability estimate of 48% for general subjective well-being, while Harris et al. (1992), utilizing a sample from Swedish Adoption/Twin Study of Aging, found that 32% of the variation in life satisfaction scores was attributable to genetic factors. In a later investigation, Stubbe et al. (2005) reported that nonadditive genetic factors explained 38% of the variance in life satisfaction. Additional evidence for the importance of genetic effects in explaining life satisfaction was provided by Johnson and Krueger (2006), using a sample of 719 twin pairs from the National Survey of Midlife Development, who found that genetic effects accounted for a substantial amount of variance in individual levels of life satisfaction. Furthermore, these authors found that these effects were moderated by level of financial position and perceived control, with genetic effects being stronger at higher levels of these factors. In a more recent effort, Hahn et al. (2013) found that a third of variance in life satisfaction was attributable to genetic effects, and further noted that shared environment effects were more pronounced for nontwins compared with twins.

In general, genetic effects can account for a wide variety of attitudes and beliefs. Keller et al. (1992) and Arvey at al. (1994) found that genetic effects accounted for a variety of workplace needs and values, although genetic effects varied in importance across values and needs. In a recent study, Weber et al. (2011) reported that genetic effects could also account for variance in group identification. Drawing from 691 twins from the National Survey of Midlife Development in the United States, these researchers found that genetic effects accounted for 41, 36, and 19% of racial, ethnic, and religious identification, respectively. Relationships have also been reported between genetics and other attitudes, such as conservatism (see Bergen et al., 2007).

Behaviors and Performance

An important body of work is also concerned with how genetics influence how people behave as well as how these factors ultimately affect performance in the workplace. Such work has focused on topics ranging from specific facets of workplace behavior, such as turnover or leadership role occupancy, to more general factors such as performance, discretionary behavior, and motivation.

As with attitudes, research has generally been supportive of the importance of genetic effects in this area. In terms of turnover across both occupations as well as specific jobs, McCall et al. (1997), utilizing a sample of 1236 monozygotic and 1165 dizygotic twin pairs, found that 36% of job change frequency and 26% in occupation change frequency could be attributed to genetic factors.

More specific classes of behavior have also been found to have a genetic component. In a study involving 1285 monozygotic and 849 pairs of dizygotic pairs of twins, Zhang et al. (2009) found significant heritability estimates for entrepreneurial behavior (although these effects were limited to the females in the sample). In a related study, Nicolaou et al. (2009) reported that a significant part of the variance in opportunity recognition (an important component of entrepreneurial behavior) could be attributed to genetic factors.

Similar results have been found in investigations of leadership role occupancy in general as well as in investigations of specific leadership behaviors. Li et al. (2012), utilizing a sample of 196 sets of twins, showed that genetic influences accounted for 48% of the variance in transformational leadership behaviors and for 30% of the variance in leadership role occupancy. More targeted investigations have also borne fruit; De Neve et al. (2013) found that a quarter of the variance in leadership role occupancy can be accounted for differences in a specific genotype; individuals with two A alleles of rs4950 were found to be more likely to have leadership positions compared with those with one or two G alleles in rs4950.

Genetic effects have also been found to relate to a variety of performance-related factors (although direct investigations of the role of genetics in job performance represent an underdeveloped area). In a recent investigation, Gottschling et al. (2012) found that, in a sample of 280 twin sets in Germany, genetic factors accounted for significant variance in a variety of scholastic achievement factors. Similarly, Deater-Deckard et al. (2005), in a sample of 92 monozygotic and 139 dizygotic pairs of twin children, found a significant heritability estimate for task persistence (0.59), which

furthermore tended to increase over time; these findings represent evidence for the role of genetics in the motivational predictors of performance.

More direct evidence can be perhaps found by looking at the undesirable components of contextual behaviors, such as aggression and altruism. In an initial study of 573 twin pairs from the University of London Institute of Psychiatry Volunteer Twin Register, Rushton et al. (1986) found significant genetic components for altruistic and aggressive tendencies (with genetic effects accounting for about 50% of the variance in this sample). The aggression findings were further supported by Bergen et al. (2007), who found a significant genetic effect for this class of behaviors. Specifically, externalizing behaviors were found to have a significant heritable component, and this was found to increase over time; this increase corresponds well with an established pattern of strengthening of genetic effects across the lifespan found in previous research. Rushton (2003), on the other hand, found additional evidence for a heritable component of prosocial tendencies (similar to altruism).

Mediating the Effects of Genetic Factors

The research efforts described above have provided important insights into the importance of genetic effects for organizational factors. However, examining organizational outcomes or their antecedents in isolation provides fairly limited contributions; approaches that synthesize genetic effects with established theoretical outcomes were needed, a fact that was actually identified fairly early in this literature. For example, the mediating role of personality in explaining the linkage between genetic effects and job satisfaction has been long hypothesized and expected (see Arvey et al., 1989). As a result, recent efforts have focused on providing more comprehensive models, providing linkages from genetics to attitudes and behaviors through individual differences in personality, cognitive ability, and affectivity. These approaches provide important contributions in demonstrating how genetic effects operate, and in linking this literature to broader organizational research, both informing and advancing the field.

In general, mediated models integrating genetic effects with organizational research either focus on how a dispositional factor can explain genetic influences in an outcome of interest or on how genetic effects can explain a relationship between two associated factors. While often mathematically similar, these two approaches represent different conceptualizations and investigate different parts of the causal chain. Both can provide important insights, and generate interesting questions for future research.

The former approach generally links individual differences in personality or affectivity to attitudes, values, and judgments, and ultimately to behavior in the workplace. Personality and affectivity are thus considered the mediators of genetic effects or, alternatively, the first part of a causal chain that considers how individual differences result in differences in perception or cognition, and ultimately behavior.

In terms of dispositional mediators, research efforts that examine how personality variables mediate the effects of genetic factors have provided important insights on the causal chain represented by these factors. This is not surprising, as personality factors have long been considered central to a number of models of workplace functioning (see Ones et al., 2007), as well as having recognized genetic components themselves (see Bouchard, 1994; Loehlin, 1992). In other words, researchers have expected and found that genetic factors are associated with workplace outcomes because of their association with individual differences.

In an initial synthetic investigation of the mediating role of personality and affective dispositions, utilizing extant meta-analyses and previously established heritability estimates, Ilies and Judge (2003) found that approximately 45% of the genetic influences on job satisfaction were mediated by positive and negative affectivity, while around 24% were mediated by the Big Five factors of personality. Similarly, in a study of 594 Swedish twins, Judge et al. (2012) reported that core self-evaluations accounted for almost half the effects of genetic factors on experienced work stress. In terms of other negative factors, Kendler and Myers (2010), in a sample of 542 same-sex twins, found that personality accounted for a significant amount of the relationship between genetic factors and depression (with the driving traits being neuroticism and, to a lesser extent, conscientiousness). Maas and Spinath (2012) found that personality factors explained part of the genetic effects on work engagement and resistance to stress (although the authors noted that the magnitude of these effects was lower than expected).

More general investigations of attitudes have provided results in line with the above. Olson, Vernon, Harris, and Lang, in a study involving 195 pairs of monozygotic and 141 pairs of dizygotic twins, found that personality factors (especially sociability) had an important mediating role in the relationship between genetic factors and a variety of general attitude factors (including such things such as gender differences, leadership importance, and intellectual pursuits).

Personality factors have been also found to mediate the effects of genetic factors on behavior. For example, Shane et al. (2010), utilizing a sample of 3412 monozygotic twins from The United Kingdom and 1300 monozygotic twins from the United States, found that openness to experience and extraversion were associated with entrepreneurial tendencies, a general class of work behaviors. Similarly, Zhang et al. (2009) found that personality (specifically, extraversion and neuroticism) accounted for a significant part of the relationship between genetic factors and entrepreneurial behaviors in the female component of their sample.

Genetic Factors as Mediators

It should be noted, however, that not all research has followed mediating approaches similar to the ones above. Instead of focusing on explaining what traits might mediate the effects of genetics on other individual factors, researchers have aimed to uncover whether an association between two traits are due to common genetic factors (as well as shared environments). For example, Bergeman et al. (1991) investigated the mediating effect of genetic influences on the relationship between perceived adequately of social support and a number of subjective well-being indicators in a sample of 424 older twin

pairs. Indeed, these authors reported that a substantial amount of the social support – well-being relationship was accounted for by genetic effects (56–65%).

Similarly, Nicolaou et al. (2009) reported that genetic factors accounted for a large percentage (53%) of the relationship between recognizing opportunities and the tendency to be an entrepreneur, while Li et al. (2012) found that genetic factors accounted for 78% of the correlation between transformational leadership and leadership role occupancy estimated in their sample. These results indicate that alternative conceptualizations of genetic influences in the causal structure linking attitudes and behaviors can also be supported by data, and highlight the importance of investigating these phenomena from a variety of viewpoints.

Other approaches have focused on the commonalities between traits or constructs. Veselka et al. (2010) found that the relationship between the HEXACO factors of personality and trait emotional intelligence can be fully attributed to common genetic and shared environmental factors. In an interesting application of this framework to a construct validation approach, Moore et al. (2010) found that genetic factors accounted fully for the demonstrated correlation between a verbal and a nonverbal measure of the Big Five. On the other hand, Kandler et al. (2012) utilized a multitrait – multimethod approach with a twin sample to show the areas of similarity and difference between temperament and personality frameworks.

Boundary Conditions

We have previously mentioned the results of Johnson and Krueger, who found that financial and perceptual factors moderated the importance of genetic effects for life satisfaction, as well as the findings of Turkheimer et al. (2003), related on the moderating effects of poverty on the relationship between genetics and cognitive ability. Other boundary conditions might reside with the factors under consideration. Conway et al. (2011) found that attitudes that are more complex demonstrated higher heritability estimates, even when controlling for a number of associated factors.

Genetic factors have also been hypothesized to influence the ways in which genetically derived traits can operate. For example, Tesser (1993) argued that attitudes with a stronger heritable component should be more readily accessible and harder to change. Indeed, Tesser et al. (1998) reported that attitudes that are more heritable were associated with greater availability of physiological responses, and were more resilient.

Directions for Future Research

A number of findings reported above hint at potentially fruitful areas of future research. Utilizing genetic factors as a biological cornerstone on which to base nomological networks of personality, affectivity, and temperament has provided important insights already, and can be expected to contribute to a variety of between-person models of workplace functioning and behaviors. The report of Kandler et al. (2011) of congruence and divergence between personality and temperament frameworks can represent an initial example of such efforts.

Other issues such as the similarities in jobs held by monozygotic twin pairs points at the importance of examining individual situation and context seeking, providing another avenue for how genetic effects can operate, generating future directions from literature ranging from turnover to job crafting research. More generally, investigating the reasons why genetic factors might predispose individuals to seek (or avoid) specific contexts, and the contextual factors that might stimulate or drive such decisions, could also be interesting.

Methodological advances both in terms of data analysis as well as available datasets will surely aid these and other efforts. Such advances can include increased capabilities to examine environment–gene interactions, and modeling trait transmission from parents to children (see Zyphur et al., 2013, for an example). Additionally, the increasing use of multilevel structural equation modeling can further refine estimations of genetic influence, and more accurately portray complex and comprehensive mediated models to test causal pathways between genetics, individual differences, and organizational outcomes of interest.

Legal and Ethical Issues

As scholarly understanding of genetic effects increased, policy makers were increasingly concerned with the potential implications of such advances for employers and employees. This interest has resulted in emergence of legislation both at the national (such as the Genetic Information Nondiscrimination Act (GINA) of 2008 in the US) and transnational levels (such as the European Union (EU) Charter of Fundamental Rights, and the 1997 United Nations Educational, Scientific, and Cultural Organization (UNESCO) Universal Declaration on the Human Genome and Human Rights).

Overall, we expect that advancement in research on genetic factors in the workplace, as well as advances in genetic testing in general, will be found to both have the potential for high levels of utility as well as abuse, on the part of employees as well as employers. Future legal discussion will be had on the rights of employees to safeguard privacy and secure against discrimination, and these discussions will likely become more pressing as screening technology and knowledge concerning genetics and individual workplace outcomes becomes more substantial.

Conclusion

The research coming out of the field of genetics has produced important, relevant findings for organizational studies. Researchers now understand the genetic components of core areas and central factors of the field, including personality, attitudes, and behavior. With increased growth in this area of research, in addition to an emphasis on multidisciplinary approaches to the social sciences, we expect there to be more opportunities in the future to integrate genetic factors into the organizational psychology literature and associated fields.

See also: Behavioral Genetics: Psychological Perspectives; Emotions and Work; Genetic Studies of Behavior: Methodology; Genetics and Anthropology; Genetics and Society; Genetics and Sociology; Genetics of Personality; Industrial–Organizational Psychology: Science and Practice; Job Satisfaction; Managerial and Organizational Cognition; Personality and Values at Work.

Bibliography

Arvey, R.D., Bouchard, T.J., Segal, N.L., Abraham, L.M., 1989. Job satisfaction: environmental and genetic components. Journal of Applied Psychology 74, 187–192.

Arvey, R.D., McCall, B.P., Bouchard, T.J., Taubman, P., Cavanaugh, M.A., 1994. Genetic influences on job satisfaction and work values. Personality and Individual Differences 17, 21–33.

Arvey, R.D., Zhang, Z., Avolio, B.J., Krueger, R.F., 2007. Developmental and genetic determinants of leadership role occupancy among women. Journal of Applied Psychology 92, 693–706.

Bergeman, C.S., Plomin, R., Pedersen, N.L., McClearn, G.E., 1991. Genetic mediation of the relationship between social support and psychological well-being. Psychology and Aging 6, 640–646.

Bergen, S.E., Gardner, C.O., Kendler, K.S., 2007. Age-related changes in heritability of behavioral phenotypes over adolescence and young adulthood: a meta-analysis. Twin Research and Human Genetics 10, 423–433.

Bouchard Jr., T.J., 1994. Genes, environment and personality. Science 264, 1700–1701.

Bouchard Jr., T.J., McGue, M., 2003. Genetic and environmental influences on human psychological differences. Journal of Neurobiology 54, 4–45.

Bouchard, T.J., Lykken, D.T., McGue, M., Segal, N.L., Tellegen, A., 1990. Sources of human psychological differences: the Minnesota study of twins reared apart. Science 250, 223–228.

Caldu, X., Dreher, J.C., 2007. Hormonal and genetic influences on processing reward and social information. Annals of the New York Academy of Sciences 1118 (1), 43–73.

Conway, L.G., Dodds, D.P., Towgood, K.H., McClure, S., Olson, J.M., 2011. The biological roots of complex thinking: are heritable attitudes more complex? Journal of Personality 79, 101–134.

De Neve, J., Mikhaylov, S., Dawes, C.T., Christakis, N.A., Fowler, J.H., 2013. Born to lead? A twin design and genetic association study of leadership role occupancy. The Leadership Quarterly 24, 45–60.

Deater-Deckard, K., Petrill, S.A., Thompson, L.A., DeThorne, L.S., 2005. A cross-sectional behavioral genetic analysis of task persistence in the transition to middle-childhood. Developmental Science 8, 21–26.

EU Charter of Fundamental Rights, 2000. Retrieved from: http://www.eucharter.org/home.php?page_id=28.

Finkel, D., McGue, M., 1997. Sex differences and nonadditivity in heritability of the multidimensional personality questionnaire scales. Journal of Personality and Social Psychology 72, 929–938.

Finkel, D., Pedersen, N.L., McGue, M., McClearn, G.E., 1995. Heritability of cognitive abilities in adult twins: comparison of Minnesota and Swedish data. Behavior Genetics 25 (5), 421–431.

Fowles, D.C., 1987. Application of a behavioral theory of motivation to the concepts of anxiety and impulsivity. Journal of Research in Personality 21, 417–435.

Goldberg, L.R., 1990. An alternative "description of personality": the big-five factor structure. Journal of Personality and Social Psychology 59, 1216–1229.

Gottschling, J., Spengler, M., Spinath, B., Spinath, F.M., 2012. The prediction of school achievement from a behavior genetic perspective: results from the German twin study on cognitive ability, self-reported motivation, and school achievement (CoSMoS). Personality and Individual Differences 53, 381–386.

Gray, J.A., 1990. Brain systems that mediate both emotion and cognition. Cognition and Emotion 4, 269–288.

Hahn, E., Johnson, W., Spinath, F.M., 2013. Beyond the heritability of life satisfaction: the roles of personality and twin-specific influences. Journal of Research in Personality 47 (6), 757–767.

Harris, J.R., Pedersen, N.L., McClearn, G.E., Nesselroade, J.R., Plomin, R., 1992. Age differences in genetic and environmental influences for health from the Swedish adoption/twin study of aging. Journal of Gerontology 47 (3), 213–220.

Haworth, C.M.A., Wright, M.J., Luciano, M., Martin, N.G., De Geus, E.J.C., Van Beijsterveldt, C.E., 2009. The heritability of general cognitive ability increases linearly from childhood to young adulthood. Molecular Psychiatry 15, 1112–1120.

Hershberger, S.L., Lichtenstein, P., Knox, S.S., 1994. Genetic and environmental influences on perceptions of organizational climate. Journal of Applied Psychology 79, 24–33.

H.R. 493: Genetic Information Nondiscrimination Act, 2000. Retrieved from: eeoc.gov.

Ilies, R., Dimotakis, N., in press. Genetic mechanisms explaining how evolution is linked to organizational behavior: The genetic basis of work attitudes and emotions. In: Stephen, M.C., Richard, D.A. (Eds.), Biological Foundations of Organizational Behavior. University of Chicago Press.

Ilies, R., Gerhardt, M.W., Le, H., 2004. Individual differences in leadership emergence: integrating meta-analytic findings and behavioral genetics estimates. International Journal of Selection and Assessment 12 (3), 207–219.

Ilies, R., Judge, T.A., 2003. On the heritability of job satisfaction: the mediating role of personality. Journal of Applied Psychology 88, 750–759.

Ilies, R., Arvey, R.D., Bouchard Jr., T.J., 2006. Darwinism, behavioral genetics and organizational behavior: a review and agenda for future research. Journal of Organizational Behavior 27, 121–141.

Jang, K.L., Livesley, W.J., Vemon, P.A., 1996. Heritability of the big five personality dimensions and their facets: a twin study. Journal of Personality 64, 577–592.

Johnson, W., Krueger, R.F., 2006. How money buys happiness: genetic and environmental processes linking finances and life satisfaction. Journal of Personality and Social Psychology 90, 680–691.

Judge, T.A., Ilies, R., Zhang, Z., 2012. Genetic influences on core self-evaluations, job satisfaction, and work stress: a behavioral genetics mediated model. Organizational Behavior and Human Decision Processes 117, 208–220.

Kandler, C., Bleidorn, W., Riemann, R., Angleitner, A., Spinath, F.M., 2011. The genetic links between the big five personality traits and general interest domains. Personality and Social Psychology Bulletin 37, 1633–1643.

Kandler, C., Held, L., Kroll, C., Bergeler, A., Riemann, R., Angleitner, A., 2012. Genetic links between temperamental traits of the regulative theory of temperament and the big five: a multitrait-multimethod twin study. Journal of Individual Differences 33, 197–204.

Keller, L.M., Bouchard Jr., T.J., Arvey, R.D., Segal, N.L., Dawis, R.V., 1992. Work values: genetic and environmental influences. Journal of Applied Psychology 77, 79–88.

Kendler, K.S., Myers, J.J., 2010. The genetic and environmental relationship between major depression and the five-factor model of personality. Psychological Medicine 40, 801–806.

Li, W., Arvey, R.D., Zhang, Z., Song, Z., 2012. Do leadership role occupancy and transformational leadership share the same genetic and environmental influences? The Leadership Quarterly 23, 233–243.

Loehlin, J.C., 1992. Genes and Environment in Personality Development. Sage Publications, Inc., Newbury Park, CA.

Maas, H., Spinath, F.M., 2012. Personality and coping with professional demands: a behavioral genetics analysis. Journal of Occupational Health Psychology 17, 376–385.

McCall, B.P., Cavanaugh, M.A., Arvey, R.D., Taubman, P., 1997. Genetic influences on job and occupational switching. Journal of Vocational Behavior 50, 60–77.

McCartney, K., Berry, D., 2005. Gene-environment processes in task persistence. Trends in Cognitive Sciences 9 (9), 407–408.

McClearn, G.E., Johansson, B., Berg, S., Pedersen, N.L., Ahern, F., Petrill, S.A., Plomin, R., 1997. Substantial genetic influence on cognitive abilities in twins 80 or more years old. Science 276, 1560–1563.

McGue, M., Bouchard, T.J., 1998. Genetic and environmental influences on human behavioral differences. Annual Review of Neuroscience 21, 1–24.

Moore, M., Schermer, J.A., Paunonen, S.V., Vernon, P.A., 2010. Genetic and environmental influences on verbal and nonverbal measures of the big five. Personality and Individual Differences 48, 884–888.

Nicolaou, N., Shane, S., Cherkas, L., Spector, T.D., 2009. Opportunity recognition and the tendency to be an entrepreneur: a bivariate genetics perspective. Organizational Behavior and Human Decision Processes 110, 108–117.

Ones, D.S., Dilchert, S., Viswesvaran, C., Judge, T.A., 2007. In support of personality assessment in organizational settings. Personnel Psychology 60, 995–1027.

Plomin, R., 1999. Genetics and general cognitive ability. Nature 402, C25–C29.

Plomin, R., DeFries, J.C., McClearn, G.E., McGuffin, P., 2008. Behavioral Genetics, fifth ed. Worth, New York.

Riemann, R., Angleitner, A., Borkenau, P., Eid, M., 1998. Genetic and environmental sources of consistency and variability in positive and negative mood. European Journal of Personality 12, 345–364.

Rushton, J., 2003. Review of 'The battlegrounds of bio-science: cross-examining the experts on evolutionary psychology, race, intelligence, & genetics, population, environment, & cloning'. Intelligence 31, 607–608.

Rushton, J.P., Fulker, D.W., Neale, M.C., Nias, D.K., Eysenck, H.J., 1986. Altruism and aggression: the heritability of individual differences. Journal of Personality and Social Psychology 50, 1192–1198.

Shane, S., 2009. Introduction to the focused issue on the biological basis of business. Organizational Behavior and Human Decision Processes 110, 67–69.

Shane, S., Nicolaou, N., Cherkas, L., Spector, T.D., 2010. Do openness to experience and recognizing opportunities have the same genetic source. Human Resource Management 49 (2), 291–303.

Song, Z., Li, W., Arvey, R.D., 2011. Associations between dopamine and serotonin genes and job satisfaction: preliminary evidence from the add health study. Journal of Applied Psychology 96, 1223–1233.

Song, Z., Wang, N., Fan, Q. Progress in molecular genetics and its potential implications to OB research. In: Stephen, M.C., Richard, D.A. (Eds.), Biological Foundations of Organizational Behavior. University of Chicago Press, in press.

Stubbe, J.H., Posthuma, D., Boomsma, D.I., De Geus, E.J.C., 2005. Heritability of life satisfaction in adults: a twin-family study. Psychological Medicine 35, 1581–1588.

Tellegen, A., Lykken, D.T., Bouchard, T.J., Wilcox, K.J., Segal, N.L., Rich, S., 1988. Personality similarity in twins reared apart and together. Journal of Personality and Social Psychology 54 (6), 1031–1039.

Tesser, A., 1993. The importance of heritability in psychological research: the case of attitudes. Psychological Review 100, 129–142.

Tesser, A., Whitaker, D., Martin, L., Ward, D., 1998. Attitude heritability, attitude change, and physiological responsivity. Personality and Individual Differences 24, 89–96.

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., Gottesman II, , 2003. Socioeconomic status modifies heritability of IQ in young children. Psychological Science 14 (6), 623–628.

UNESCO, 1997. Universal Declaration on the Human Genome and Human Rights, 29th General Conference. Retrieved from: http://www.unesco.org/new/en/social-and-human-sciences/themes/bioethics/human-genome-and-human-rights/.

Veselka, L., Petrides, K.V., Schermer, J., Cherkas, L.F., Spector, T.D., Vernon, P.A., 2010. Phenotypic and genetic relations between the HEXACO dimensions and trait emotional intelligence. Twin Research and Human Genetics 13, 66–71.

Watson, D., 2000. Mood and Temperament. Guilford Press, New York.

Watson, D., Wiese, D., Vaidya, J., Tellegen, A., 1999. The two general activation systems of affect: structural findings, evolutionary considerations, and psychobiological evidence. Journal of Personality and Social Psychology 76, 805–819.

Weber, C., Johnson, M., Arceneaux, K., 2011. Genetics, personality, and group identity. Social Science Quarterly 92, 1314–1337.

Zhang, Z., Zyphur, M.J., Narayanan, J., Arvey, R.D., Chaturvedi, S., Avolio, B.J., Lichtenstein, P., Larsson, G., 2009. The genetic basis of entrepreneurship: effects of gender and personality. Organizational Behavior and Human Decision Processes 110, 93–107.

Zyphur, M.J., Zhang, Z., Barsky, A.P., Li, W.D., 2013. An ACE in the hole: twin family models for applied behavioral genetics research. The Leadership Quarterly 24, 572–594.

Genetic Screening for Disease

Lidewij Henneman and Danielle Rosa Martina Timmermans, VU University Medical Center, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by L.P. ten Kate, D.R.M. Timmermans, volume 9, pp. 6104–6108, © 2001, Elsevier Ltd.

Abstract

Genetic screening aims to (1) ensure health gain by providing opportunities for disease prevention and treatment, or (2) increase reproductive choice. Genetic screening can be done at different stages in life such as preconception, prenatal, neonatal, and later in life. The advances in genetics research and technology have made screening for an increasing number of medical disorders possible. Genome-scale tests are expected to change the screening landscape even further. Before introducing these new screening initiatives, it is essential to weigh the benefits against the harms of screening, including addressing the ethical, economic, and societal issues.

Introduction

Definition

Genetic screening involves the medical examination of individuals for the early detection or exclusion of a hereditary disease or a genetic predisposition to a disease, or to determine whether a person carries a predisposition that may produce a hereditary disease in their offspring (Godard et al., 2003). A distinguishing feature of screening is that a test is offered without being requested. Typically, a country's health care system will take the initiative. Genetic screening is considered an activity of community genetics – the application of medical genetics for the benefit of large numbers of people. Screening can be offered systematically in the form of large-scale programs, aimed either at a whole population of asymptomatic individuals (universal) or at people in a predefined target group (selective), such as neonates or pregnant women. Early-detection screening means that testing is applied to those who have not (yet) sought medical aid as a result of physical signs, symptoms, or anxiety. Screening can also entail people responding to an advertisement (or a website offer) for a health check at a clinic or health organization (Health Council of the Netherlands, 2008) or direct-to-consumer (DTC) genetic testing (Borry et al., 2011). DTC genetic testing typically involves collecting a sample at home, mailing it back to the laboratory, and subsequently receiving results without having to consult a health care professional.

Characteristics of Genetic Screening

The main differences between diagnostic genetic testing and genetic screening are shown in Table 1. An important difference is that genetic diagnosis is often indicated or requested by those who have an a priori increased risk, for example patients who have symptoms of a disease, family members of patients already diagnosed, or prospective parents with a known increased risk of having an affected child. Genetic screening, in contrast, may be seen as being imposed on persons (Godard et al., 2003). Moreover, as will be explained in this article, in diagnostic testing results are usually definitive, whereas in screening results are sometimes preliminary.

Testing is not restricted to the analysis of DNA, but includes the use of any test (molecular, cytogenetic, biochemical, or imaging) that yields genetic data that are designed to detect and exclude hereditary disease, genetic predisposition to disease, or carrier state. One can thus test on a genotypic level (e.g., DNA, RNA, gene, and chromosome) or a phenotypic level (e.g., appearance and characteristics), or look at metabolites (e.g., proteins and enzymes). Screening for genetic risk can also be done by taking a family history (Burke et al., 2011). As with other types of medical screening, diagnostic confirmation of a positive screening test is an integral part of the screening procedure. For example, screening for Down syndrome during a woman's pregnancy can be done using a combined test consisting of an ultrasound examination to measure the nuchal translucency of the fetus (phenotypic level), and a blood test to check for levels of certain proteins and hormones of the mother. When combining these data with the mother's age, the chance of a child with Down syndrome can be determined. If the chance is higher than average, an invasive diagnostic test is offered by means of chorionic villus sampling or amniocentesis for definite diagnosis.

The difference between genetic screening and other types of screening is that with most forms of genetic screening, one test

Table 1 Differences between a (genetic) diagnostic test and screening test

	Diagnostic test	*Screening test*
Target	Individuals	(Part of) the population
Target characteristics	Persons often worried because of disease symptoms or relatives affected	Persons with no known signs or symptoms
Motive to test	On request or indication	Offer unasked for
Test results	Usually definitive	Sometimes preliminary
Expense and effort	May cost a lot of time, effort, and/or money	Should be cheap and easily applied
Example	Testing for hereditary breast cancer in the family	Blood spot screening in newborns

during a person's lifetime is enough – for example, when screening for a gene mutation that causes disease (e.g., familial hypercholesterolemia). The stability of the genetic characteristics also makes it possible to predict disease long before it would become manifest, thereby making early intervention possible. In contrast, in screening for infectious diseases or cancer, the majority of participants have to return for subsequent testing after a set time period. Genetic screening is also distinguished from other forms of screening because a positive test result may have consequences for family members. This does not apply to all types of genetic screening, however. For instance, finding a fetus with a nonhereditary trisomy 21 Down syndrome (the most common) does not increase the risk for sisters of the pregnant woman, since this is a spontaneous mutation. But when screening shows that someone carries, for example, two mutations of the gene involved in hemochromatosis, this puts his or her siblings immediately at a 25% risk of having the same mutations (Worwood, 2000). They may be grateful to be informed about this since manifestations of this iron overload disease can be prevented effectively by regular bleeding, but the point here is that an individual's decision to participate in a genetic-screening program is to some extent a decision that affects family members who might have preferred to make other choices.

A further characteristic of genetic screening that differentiates it from other types of medical screening is that it does not necessarily lead to the prevention or treatment of diseases.

Aim of Genetic Screening

The aim of genetic screening is twofold: (1) to minimize morbidity and mortality by providing opportunities for prevention and treatment through early detection and initiation of treatment, preferably even before symptoms are evident (e.g., neonatal screening) or by offering risk-reducing preventative measures (e.g., familial hypercholesterolemia); or (2) to increase reproductive choice and family planning (e.g., embryo selection and prenatal diagnosis).

The purpose of screening is thus not just the outcome of the test (e.g., early disease detection) but also ensuing health gain or other benefits to the person and population being tested (Health Council of the Netherlands, 2008). This is often referred to as 'clinical utility' (Grosse et al., 2010).

Types of Screening

Screening at Different Stages of Life

The timing of genetic screening can vary. The earliest possibility for genetic screening is preimplantation genetic screening (PGS) in the case of *in vitro* fertilization (IVF). Preimplantation genetic diagnosis (PGD) is performed when one is looking for a disorder for which this particular embryo has a particularly high risk, while in PGS the embryo in question does not have a much higher risk than other preimplantation or IVF embryos. PGS for chromosomal anomalies, for instance, is done to select embryos with the highest chance of survival.

Screening during pregnancy is a form of reproductive screening to allow for the possibility of prenatal diagnostics and selective termination of pregnancy, to prepare parents (psychologically, financially, etc.) for an affected child, or to timely facilitate medical care before or after birth. Prenatal screening includes prenatal aneuploidy screening (e.g., Down syndrome (trisomy 21), Edwards syndrome (trisomy 18), and Patau syndrome (trisomy 13)) and screening using an ultrasound for other structural fetal anomalies (e.g., neural tube defects or congenital heart disease).

Population-based reproductive carrier screening for autosomal recessive disorders such as cystic fibrosis, sickle cell anemia, thalassemia, and Tay–Sachs disease can be offered to individuals, but because being a carrier is significant only when the partner is also a carrier, it is logical to delay screening for carrier status until a partner has been found and reproductive issues arise. In this case, the couple rather than the individual would be the target. When both partners of a couple are carriers of the same disease, their children are at 1-in-4 risk for the disorder. Since carrier couple screening can be performed premaritally, preconceptionally, or prenatally, it is important to define the best moment for screening within the lifetime of the individual or couple. Premarital screening is seen in cultures where selection of marriage partners is done by third parties, without involving the partners at all. For example, to avoid marriage between carriers of the same disorder, Dor Yeshorim – an international anonymous carrier-screening program that primarily serves the Orthodox Jewish community – has elaborated a carrier-matching program for genetic diseases that are prevalent in the Ashkenazi Jewish population (Raz and Vizner, 2008). Premarital screening after the partners have matched (or been matched) has been normal practice in countries where consanguinity is high, such as Cyprus and Saudi Arabia, and may infringe personal autonomy if provided on a compulsory basis (Wertz et al., 2003).

Preconception carrier screening enables most reproductive options, including not only prenatal diagnosis but also the choices of using PGD, using donor sperm or oocytes, seeking adoption, or refraining from having children. Preconceptional screening is characterized by difficulties in gaining access to the target group in time. The majority of couple-screening programs therefore target pregnant women (and their partners), which has the disadvantage that limited reproductive options are available and decisions may have to be made under time pressure. Some people advocate carrier screening in secondary school, but the time interval between knowing the test results and using the information is long. Carrier-screening programs for X-linked disorders such as fragile X syndrome are offered to women either before or during pregnancy to detect their carrier status. Half of the male offspring of a carrier mother will typically be affected by the disorder, and a carrier daughter may or may not show symptoms (depending on the disease and/or mutation).

Screening of newborns typically aims to identify children with autosomal recessive disease to offer timely intervention to improve health outcome. Examples include a diet restricting phenylalanine to prevent severe cognitive impairment in phenylketonuria (PKU), or nutritional and respiratory interventions to prevent malnourishment and poor growth in cystic fibrosis. High-throughput tandem mass spectrometry makes it possible to screen newborns for many rare inborn errors of metabolism, and in most high-income countries the number of disorders included in neonatal screening is increasing (Javaher

et al., 2010). Proposed testing for untreatable disorders, such as fragile X syndrome or Duchenne muscular dystrophy, is the subject of ongoing debate. The aims of screening, then, are to enable reproductive choice for future children and save parents an often long, diagnostic quest, rather than the early detection to morbidity.

The final stage of screening during life is screening in adulthood (later in life), for example for familial hypercholesterolemia, an autosomal dominant disorder characterized by an increased risk of developing premature cardiovascular disease (CVD) caused by severe elevations in total cholesterol and low-density lipoprotein cholesterol. CVD risk can be reduced by cholesterol-lowering medication (statins) and the adoption of a healthy lifestyle (healthy diet, sufficient physical activity, and refraining from smoking) (Civeira, 2004).

Family-Based or Cascade Screening

Cascade screening is a special type of genetic screening targeted at relatives of identified carriers. Starting from an individual with the characteristic of interest (the proband), one actively offers screening to his or her first-degree relatives. Once the relatives with the same characteristic have been identified, their first-degree relatives are also offered screening, and so on. This type of screening is particularly suitable for dominant disorders, such as familial hypercholesterolemia, and for X-linked disorders with high penetrance (the frequency of cases in which a particular genotype comes to expression) and low mutation rate. It has also been proposed for the detection of couples who are carriers of an autosomal recessive disease such as cystic fibrosis. The disadvantage here is, however, that only a small proportion of at-risk couples can be identified in this way (Morris et al., 2004). In that case, population-based (mass) screening only, or in combination with cascade screening, is a better option. Considerably higher efficiency rates for cascade screening are, however, seen when screening is performed to detect homozygotes for frequent recessive mutations with reduced penetrance, such as hereditary hemochromatosis (Krawczak et al., 2001).

Genetic-Screening Criteria

There are many genetic, medically relevant characteristics for which, in theory, screening could be initiated. It is quite clear that it is not desirable to screen for all these characteristics, and careful selection is needed. In addition to potential benefits, potential harm should also be considered. Harm could be physical, psychological, social, or otherwise, and it is important to take not only the effects of screening on the participants into account, but also the effects on those who decide not to participate. Since the classic Wilson and Jungner screening criteria, developed in 1968 (Wilson and Jungner, 1968), screening criteria have been further refined and adapted to meet requirements for genetic screening (Godard et al., 2003; Nuffield Council on Bioethics, 1993). Some countries have also developed specific legislation.

To assist policy makers in their decisions to develop, implement, or continue a genetic-screening program, guidance is – with further refinements – built on the foundation of the Wilson and Jungner criteria. It is widely recognized that every individual screening program should be pilot-tested first to study the advantages and disadvantages arising from the screening, and to see if it meets the appropriate standards. An acceptable and reliable test should be available with known sensitivity, specificity, and predictive value. Practical courses of action should be available for the participants, such as preventive measures or reproductive choices. The psychological and social effects should be known for the participants and nonparticipants in general, and – in particular – for those among the participants who are faced with a positive test and those among the nonparticipants who are confronted with the disease in themselves or in a child. Participation in the screening program should be voluntary and based on good-quality information and informed consent, and an autonomous decision should be guaranteed, especially with regard to reproductive screening where couples may be faced with the choice of avoiding or terminating a pregnancy. In contrast, in some countries, and in most states in the United States of America, newborn screening is a legislatively mandatory program, and parental consent is rarely required. The argument here is that it is society's obligation to overrule the parental decision to refuse screening that has a minimal risk of harm while having potential life-saving consequences for the child.

In screening, people should be given enough time to reach a balanced, well-informed decision. This requirement may be difficult to realize in prenatal and neonatal screening, but every effort should be made to inform pregnant women about the decision to come forward as early in pregnancy as possible. People should have ample opportunities for pretest consultation and privacy protection, and permanent quality control of all aspects of the program should be ensured. Finally, the financial costs of the program should be taken into account, and alternative strategies for combating the health problem should be considered. All this information should make it possible to compare and weigh the benefits and drawbacks of the program in a given context. The final balance should be clearly biased toward the benefits. It is desirable to have an authority to supervise all (genetic) screening programs in a country or larger region.

The complexity of genetic screening makes it a multidisciplinary area. Genetic-screening programs can be envisioned as a conceptual framework with three integrated levels (i.e., laboratory, clinical, and management), involving many stakeholders such as laboratory specialists, medical doctors, psychologists, epidemiologists, sociologists, jurists, ethicists, economists, patient organizations, representatives of the community, and policy makers (Andermann et al., 2010).

Quality of the Screening Test Method

Before screening can be offered, a reliable and valid test should be available. A test result is termed positive when it points to the presence of the feature sought after. Similarly, a result is said to be negative when it indicates the absence of the feature. Both results, however, may be false, resulting in false positive and false negative test results, respectively. The probability that a certain screening test result is true or false depends on the prevalence of the feature and two test characteristics: sensitivity

and specificity. Sensitivity refers to the proportion of cases in which the presence of the feature is identified by the test. Specificity is the proportion of cases in which the absence of the feature is correctly identified by the test. The predictive value of a test result, either positive or negative, indicates how often that test result is correct. In many cases, it is not possible or feasible to have a test that is 100% sensitive or specific. In neonatal screening, the missed affected cases should be avoided as much as possible, which may result in a higher number of false positives. The lack of perfect specificity can be countered by subsequent diagnostic workup. When the screening involves a DNA test, specificity is usually very high. If, in the cascade approach, family members are tested for the same mutation as has been found in the proband, sensitivity is also very high. In the case of a test with imperfect sensitivity (e.g., 90%), the imperfection is multiplied in carrier couple screening (81%).

With an increasing number of disorders tested for, for example in neonatal screening, the challenges of interpreting results are increasing with ambiguous results, incidental findings, and overdiagnosis (i.e., a screening-based diagnosis in a person who is destined to remain asymptomatic or whose course of disease is not changed by early diagnosis) (Burke et al., 2011).

Psychosocial Impact of Genetic Screening

Because the potential impact of genetic screening is considerable, the advantages as well as the disadvantages of genetic screening should be considered carefully before deciding to offer a specific screening program as a matter of course. The availability of genetic screening may also have its drawbacks. Among the potential drawbacks of genetic screening are the difficulty of making an informed decision for a test that is offered but unasked for, and the psychosocial implications of a positive, false positive, or unfavorable test result.

Genetic risk information is complex and often replete with uncertainties that are associated not only with the hereditary nature of the disease but also with the degree of informativeness of the test results, the effectiveness of preventive measures, and the variability of expression of the disease (incomplete penetrance). The underlying assumption that accurate perception of risk could help people make informed and individual decisions seems to be based on the principle of the autonomous and rational decision maker. From this perspective, the most important goal of genetic education and pretest counseling is to provide people with all relevant information in an understandable way. It is therefore important to find the best format for communicating risks and to stimulate a rational trade-off of pros and cons of several options. Research has shown, however, that people find it difficult to understand risks and to make rational trade-offs (Van den Berg et al., 2005). In the case of screening for a hereditary disease, the decision to screen is complicated by the potential impact that test results could have on family members and may put familial relationships under pressure.

Multiple testing – for example, testing multiple carrier-screening panels for autosomal recessive conditions – increases the difficulty of giving sufficient information to reach an informed decision. Moreover, with the increasing possibilities and availability of genetic screening, more and more people are now being confronted with the choice of undergoing screening, for example during pregnancy for fetal anomalies. In many countries, young pregnant women with a low risk of having a child with Down syndrome are offered prenatal screening (Javaher et al., 2010). If they choose screening, they must cope with the possibility of a positive screening result and the implication that confirmation is needed using invasive diagnostics such as amniocentesis, which itself carries a risk of miscarriage. When they decide against screening, they must deal with the feelings that they could have made a different choice (in the case that things turn out badly and the child is affected). The decision for or against screening is not only emotionally complicated but also cognitively complicated.

Screening and waiting for the results may have an emotional impact on people. Many people at low risk are likely to undergo testing for reassurance and hope for a good result. Most of them will receive a negative – favorable – test result. For those who perceive themselves as being at risk, this is reassuring, but the news will perhaps be learned at the expense of knowing that they were at risk in the first place. A positive screening test is obviously upsetting even if it turns out to be false. In particular in the neonatal stage, an increasing number of parents receive a false positive result as the number of disorders tested for is increasing (Burke et al., 2011).

When test results become available, they need to be interpreted. Mostly, test results also involve uncertainties. A positive test result after prenatal screening for Down syndrome means that the risk of having a child with Down syndrome is higher than a selected cutoff point (e.g., 1 in 200). A decision must then be made of whether or not to confirm this chance estimate with invasive diagnostics, whereby the extra information must be traded off against the risk of a miscarriage. It has been shown that women who make an informed decision for prenatal screening showed less adverse emotional reactions when confronted with a positive screening outcome and seemed to feel more able to make a decision about prenatal diagnostic testing (Kleinveld et al., 2009). Most women having diagnostic tests as a result of a positive screening test for Down syndrome will turn out not to carry a child with Down syndrome, and their anxiety can be considered to be unnecessary.

Testing positive as a carrier of a recessive disease such as cystic fibrosis without actually having the disease can affect one's view of the future and one's optimism about it (Axworthy et al., 1996). Testing for carrier status of cystic fibrosis involves a very large number of mutations, but it is practicable to test only for the most common, resulting in test sensitivity of less than 100%. If tested negative for carrier status of cystic fibrosis, there is still a residual risk and thus no complete reassurance can be given that one is not a carrier, therefore demanding adequate pretest information. It has been shown that a significant proportion of participants who tested negative when undergoing carrier screening wrongly believed that they were definitively not carriers, while some carriers falsely believed that they were only likely to be carriers (Axworthy et al., 1996).

Another complex issue relates to the fact that as a consequence (by-product) of the screening method, in neonatal screening also carriers of autosomal recessive disorders are identified, for example for sickle cell disease and to a lesser

extent for cystic fibrosis. The implications of these results are not always adequately understood by those parents who are informed of these findings.

It could be argued that a poor understanding of genetics in general, the tests, and the meaning and consequences of test results lies at the basis of many of the problems or possible negative effects of genetic screening described in this article. It is therefore very important that the public as well as people with an at-risk status learn more about genetics, how best to make a decision, what it means to have a positive screening result, and so on. Professionals should learn the best way to provide genetic information in such a way that people can make a balanced decision for or against genetic screening and be able to cope with the test results so that the advantages of screening will be maximized and the disadvantages minimized.

New Developments in Genetic Screening

Traditionally, genetic screening was aimed at rare single-gene (monogenic) Mendelian disorders. As more has become known about the genetic background of disease, more attention is being given to screening susceptibility for common multifactorial disorders such as diabetes, heart disease, and psychiatric disorders (Andermann et al., 2010). However, testing based on these genetic variants, even in addition to traditional disease risk factors such as obesity and hypertension, still shows limited predictive value for disease and few health benefits to the individual user. Despite these limitations, testing is commercially offered directly to consumers, mostly on the Internet (*see* Direct-to-Consumer Personal Genetic Testing). In the future, as more genetic variants associated with cancer risk are identified, cancer-screening programs such as those aimed at breast cancer may be based on genetic risk profile, which has the potential to be more effective than the current mammography screening program based on age (Pharoah et al., 2008).

With the advances in genetics research and technology, genetic screening for an increasing number of medical disorders is possible. Questions are raised regarding whether screening for a particular disorder should be routinely performed or whether screening programs should be expanded, for example in neonatal screening. A successful newborn-screening program requires a suitable test and effective follow-up, such as rapid access to treatment, which may not always be the case.

New developments are also being achieved by high-throughput technologies such as genome-scale tests to measure multiple genetic differences in the genome simultaneously. These techniques are expected to have a great influence on the screening practice, especially since the costs of these tests will further decrease (Burke et al., 2011). One example is massively parallel DNA sequencing of cell-free fetal DNA in maternal blood to detect fetal chromosomal abnormalities. This means that for Down syndrome screening, a simple blood sample of the mother is enough to perform testing (noninvasive prenatal diagnosis (NIPD)). One of the main advantages of this technique for detecting Down syndrome, as opposed to the currently used combined test, is that the accuracy is very high and about 98% of the invasive diagnostic procedures could be avoided (Chiu et al., 2011). In addition to NIPD testing for aneuploidies, it has been shown that noninvasive whole-genome sequencing of a fetus is possible (Kitzman et al., 2012). Another example is the simultaneous analyses of carrier status of multiple recessive disorders using genome-scale screening techniques (Bell et al., 2011), which may be applied in preconception carrier screening. Genome-wide testing may also apply to the screening of IVF embryos. In that case, PGS is not just to select embryos with the highest chance of survival but also to select an embryo that will grow into a child that is as healthy as possible (Health Council of the Netherlands, 2010). These developments clearly have the potential to change the screening landscape, but obviously also raise new ethical, economic, and societal issues that need to be addressed.

See also: Behavioral Genetics: Psychological Perspectives; Community Genetics; Direct-to-Consumer Personal Genetic Testing; Eugenics as a Basis of Population Policy; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetics and Society; Genetics, Disease, and Reproduction; Genetics: Legal Aspects; Genetics: The New Genetics; Genomics, Ethical Issues in.

Bibliography

Andermann, A., Blancquaert, I., Déry, V., 2010. Genetic screening: a conceptual framework for programmes and policy-making. Journal of Health Services Research and Policy 15, 90–97.

Axworthy, D., Brock, D.J., Bobrow, M., Marteau, T.M., 1996. Psychological impact of population-based carrier testing for cystic fibrosis: 3-year follow-up. Lancet 347, 1443–1446.

Bell, C.J., Dinwiddie, D.L., Miller, N.A., et al., 2011. Carrier testing for severe childhood recessive diseases by next-generation sequencing. Science Translational Medicine 3, 65ra4.

Borry, P., Henneman, L., Lakeman, P., et al., 2011. Preconceptional genetic carrier testing and the commercial offer directly-to-consumers. Human Reproduction 26, 972–977.

Burke, W., Tarini, B., Press, N.A., Evans, J.P., 2011. Genetic screening. Epidemiology Review 33, 148–164.

Chiu, R.W., Akolekar, R., Zheng, Y.W., et al., 2011. Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study. British Medical Journal 342, c7401.

Civeira, F., 2004. Guidelines for the diagnosis and management of heterozygous familial hypercholesterolemia. Atherosclerosis 173, 55–68.

Godard, B., Ten Kate, L.P., Evers-Kiebooms, G., Aymé, S., 2003. Population genetic screening programmes: principles, techniques, practices, and policies. European Journal of Human Genetics 11 (Suppl. 2), S49–S87.

Grosse, S.D., Rogowski, W.H., Ross, L.F., et al., 2010. Population screening for genetic disorders in the 21st century: evidence, economics, and ethics. Public Health Genomics 13, 106–115.

Health Council of the Netherlands, 2008. Screening: Between Hope and Hype. Health Council of the Netherlands, The Hague (publication no. 2008/05E).

Health Council of the Netherlands, 2010. The 'Thousand-Dollar Genome': An Ethical Exploration Monitoring Report Ethics and Health. Centre for Ethics and Health, The Hague (publication no. 2010/2).

Javaher, P., Nyoungui, E., Kääriäinen, H., et al., 2010. Genetic screening in Europe. Public Health Genomics 13, 524–537.

Kitzman, J.O., Snyder, M.W., Ventura, M., et al., 2012. Noninvasive whole-genome sequencing of a human fetus. Science Translational Medicine 4, 137ra76.

Kleinveld, J.H., Ten Kate, L.P., Van den Berg, M., Van Vugt, J.M., Timmermans, D.R., 2009. Does informed decision making influence psychological outcomes after receiving a positive screening outcome? Prenatal Diagnosis 29, 271–273.

Krawczak, M., Cooper, D.N., Schmidtke, J., 2001. Estimating the efficacy and efficiency of cascade genetic screening. American Journal of Human Genetics 69, 361–370.

Morris, J.K., Law, M.R., Wald, N.J., 2004. Is cascade testing a sensible method of screening a population for autosomal recessive disorders? American Journal of Medical Genetics Part A 128A, 271–275.

Nuffield Council on Bioethics, 1993. Genetic Screening. Ethical Issues. Nuffield Council on Bioethics, London.

Pharoah, P.D.P., Antoniou, A.C., Easton, D.F., Ponder, B.A.J., 2008. Polygenes, risk prediction, and targeted prevention of breast cancer. New England Journal of Medicine 358, 2796–2803.

Raz, A.E., Vizner, Y., 2008. Carrier matching and collective socialization in community genetics: Dor Yeshorim and the reinforcement of stigma. Social Science and Medicine 67, 1361–1369.

Van den Berg, M., Timmermans, D.R., Ten Kate, L.P., Van Vugt, J.M., Van der Wal, G., 2005. Are pregnant women making informed choices about prenatal screening? Genetics in Medicine 7, 332–338.

Wertz, D.C., Fletcher, J.C., Berg, K., 2003. Review of Ethical Issues in Medical Genetics. Report of Consultants to WHO. World Health Organisation, Human Genetics Programme, Geneva.

Wilson, J.M.G., Jungner, G., 1968. The Principles and Practice of Screening for Disease. World Health Organisation, Geneva.

Worwood, M., 2000. Early detection of genetic hemochromatosis: should all young adults be offered the genetic test? Genetic Testing 4, 219–228.

Genetic Studies of Behavior: Methodology

Carol A Prescott, Thalida Em Arpawong, and John J McArdle, University of Southern California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The goal of genetic research on behavior is to understand how genetic variation influences behavioral variation. Multiple methods have been used, from genetic influences inferred from familial resemblance to measured genes using molecular genetic methods. In this article, first, we discuss some historical background in genetic studies of behavior. Second, we focus on contemporary issues in the study of gene structure and function, emphasizing the methodological unity of most approaches. Contemporary issues include (1) measuring variation, (2) experimental methods, and (3) statistical methods. Finally, we discuss methodological challenges likely to continue in future genetic studies of behavior.

The goal of genetic research on behavior is to understand how genetic variation influences behavioral variation. A wide variety of methods have been used, from genetic influences inferred from familial resemblance to measured genes using molecular genetic methods.

Humans have 26 pairs of chromosomes containing approximately 30 000 genes – identifiable units that contain the code for making specific proteins. Genes range in size, but altogether comprise $3.2\times10e^9$ *base pairs*, the fundamental letters in the genetic code. The genetic sequence of humans overlaps about 99% with that of chimpanzees, and all humans are identical at more than 99.9% of their base pairs. Genetic studies of individual differences focus on this relatively small amount of genetic variation, the *genotype*, and how it relates to variation in behavior, the *phenotype*.

Genetic methods and analyses have roots in the nineteenth century, specifically in Charles Darwin's well-known ideas on evolution and Gregor Mendel's work on the mechanisms of heredity in plants. Many techniques used in genetic analyses of behavior originated in the agricultural sciences in efforts to modify the size and health of plants and animals. These were further developed in the medical sciences to identify and control sources of disease, and are now used to study behavioral phenomena.

In this article, we discuss some of the historical background in genetic studies of behavior; then, focus on contemporary issues, emphasizing approaches used to study phenotypes that are common in the general population rather than rare conditions. This is an updated version of a prior article by McArdle and Allison (2001). We do not address important issues related to the applications of genetic information such as genetic testing and screening, ethical use of genetic information, cloning, and genetically modified organisms (see cross references for articles covering these topics).

Classical Methods in Genetic Studies of Behavior

Until recently, human genotypes were difficult to measure and often treated as unobserved or latent attributes. Analyses of family trees or *pedigrees* provided some of the first scientific evidence for establishing the contributions of genetic variation to clinical phenotypes. Tracing the pattern of inheritance through multiple generations of a family provides evidence about the mechanism of transmission – whether dominant or recessive, sex-linked (on the X or Y chromosome) or autosomal (carried on the other 22 pairs of chromosomes). For example, Sir Francis Galton studied the pedigree of the English Queen Victoria, Prince Albert, and their descendants to study the pattern of transmission of the bleeding disorder, hemophilia. The pedigree method has been highly successful for identifying Mendelian conditions – rare categorical outcomes determined by a single gene – but it is not as useful for studying behavioral variation in the normal range, which is mostly likely influenced by multiple genes as well as environmental factors.

Another source of evidence for genetic influences on behavior is selective breeding studies with nonhuman animals. Seminal selective-breeding studies of learning were conducted by Robert Tryon in the 1930s (Tryon, 1940). In these studies, a heterogeneous stock of rats was tested on a maze. The observed performance on the learning tasks was used to classify groups of 'dull' and 'bright' rats. Those within each group were allowed to mate and produce offspring, which were then tested on the maze and classified. After accounting for some deleterious effects from inbreeding, this successive selection led to distinct strains of 'maze-bright' and 'maze-dull' rats within relatively few generations, thereby demonstrating the importance of genetic influences on a complex behavioral phenotype.

Francis Galton also studied resemblance among family members for continuous phenotypes such as height and intellectual achievement. He reported that individuals born to higher social classes attained higher educational and social standing than those born to lower classes. These family studies were not controlled experiments, so the impacts of genotypes and environments could not be separated. However, the ideas from family studies were combined with experimental concepts derived from animal research to create a novel class of methods based on correlating the behaviors of individuals in atypical groups of families, including twins, adoptees, half-siblings, and other genetically informative family groupings. These methods were discussed and expanded in the early twentieth century by many researchers (e.g., Thorndike, Siemens, Merriman, Fisher, as described in Spector et al., 2000).

Statistical models of inheritance were defined by synthesizing the quantitative rules of Mendelian inheritance with the

correlation decomposition developed by statistician R.A. Fisher and subsequently by many others, including R.B. Cattell, J. Jinks, and D. Fulker (see Falconer and Mackay, 1996). Another important development was the recognition that most phenotypes are measured imperfectly, and this measurement error will reduce the observed correlation among relatives or between genotype and phenotype. Geneticist Sewell Wright (1920) developed path analysis to estimate genetic influences separately from environmental and 'chance' factors. Later, multivariate statistical methods were developed to allow analyses of *latent* phenotypes, derived from multiple measures or other statistical methods for reducing error (e.g., McArdle and Goldsmith, 1990; Neale and Cardon, 1992).

In the remaining sections of this article, we describe the methods for measuring genetic variation, contemporary methodologies used in genetic studies of behavior, and consider some issues for future research.

Measuring Variation in Gene Structure and Function

The final decade of the twentieth century brought a revolution in molecular genetic methodology, including direct measurement and experimental manipulation of variation in genotype and gene expression. Direct sequencing of genetic regions of interest is emerging as a method for studying behavioral phenotypes; but, more commonly applied methods utilize genetic *markers*, sequences with a known position on the genome at which there is measurable genetic variation between individuals. Studies based on positional information capitalize on the fact that Mendel's law of independent assortment does not hold for genes that are physically near each other on the same chromosome. This means that they are transmitted together from parents to offspring to a degree that is directly related to their physical proximity on a chromosome. Commonly used markers include smaller variants, or *single nucleotide polymorphisms* (*SNPs*) that differ at a single base pair and thus have at most four possible forms (A, G, C, or T). The larger variants have multiple forms based on size or structure (e.g., CNVs, microsatellite markers). The alternative forms of either the smaller or larger variants are referred to as *alleles*. Genes range in length from 800 to 2.4 million base pairs, providing thousands of locations within a gene at which individuals may vary in alleles and that could serve as markers. Some markers code for functional variation hypothesized to contribute directly to a phenotype, whereas others correlate with genetic function based on having a location proximal to the functional variant.

Cosegregation of markers and phenotypes is studied using several different designs, including linkage and association (Elston, 1998). Linkage studies are similar to the pedigree method in attempting to identify genetic regions that are more frequently transmitted with the phenotype than would be expected by chance based on the degree of genetic relatedness of family members. These family-based methods have been successful for identifying genetic variants responsible for rare conditions and conditions that have *etiological heterogeneity* – when there are different genetic mutations in different families. Their usefulness is less clear for studying traits that are continuously distributed in the population and for which there are many contributing genetic variants that are common in the population and have small effects. Another disadvantage of linkage studies is that they are typically designed to investigate a single phenotype (e.g., schizophrenia), requiring a great amount of effort to locate, assess, and genotype a large sample of affected individuals and their family members.

Association studies estimate the statistical association between a genetic marker and phenotype in a variety of designs, with the number of markers ranging from a single *candidate gene* (selected for its hypothesized functional relation to the phenotype) to exploratory methods such as *genome-wide association studies* (*GWAS*) that may include millions of markers. Sampling for association studies includes case control (selected for presence or absence of a categorical phenotype) and population samples, both of which allow estimates to be made for the association of markers with quantitative phenotypes. Association studies are useful for examining common genetic variants that contribute to complex traits and several genetic variants simultaneously that may have small effects on the traits. Disadvantages include the expense of collecting genetic data on large samples of individuals, accounting for variable gene frequencies within the study sample, and spurious positive associations (see Section Methodological Issues for Future Genetic Studies of Behavior).

Newer analysis methods contain features of both linkage and association, including family-based association tests and as well as methods for analyzing GWAS data obtained from relatives or multivariate phenotypes.

Experimental Methods to Study Genetic Variation

In addition to the correlational methods used for studying human phenotypes, studies of animals allow experimental methods for more precise understanding of gene activity. Most of this work is done with rodents as they have short-generation times, and mammals are most applicable for studying phenotypes that parallel human traits. The usefulness of animal models is most clear for physiological responses (e.g., drug sensitivity) but is less evident with more complex phenotypes.

Conceptually, the simplest method for locating trait-influencing genes is selective-breeding studies. These studies are based on measurements of the frequencies of genetic variants in an outbred (genetically heterogeneous) population. As subsequent generations are bred to produce higher and lower values of the trait, genetic variants that change in frequency across generations can be concluded to influence the trait or be linked to variants that do.

A more complex but efficient method uses recombinant inbred (RI) strains – lines of animals that have been inbred until they are genetically identical. If the lines differ in their average value for a phenotype (when all are reared under identical conditions), this implicates genetic variation between lines in producing the phenotype. Because members of a strain are identical, genotyping only needs to be conducted for one animal. Researchers can use databases cataloging the genetic variation among inbred strains to conduct genotype–phenotype studies. More complex methods involve intercrosses between inbred strains, analyses of multiple markers,

multivariate analyses with multiple phenotypes, selective sampling, and other enhancements.

In addition to variation in what a gene produces, there is variation in gene *expression*, how gene activity varies across development or in response to internal or external events. Genetic expression studies use microarrays, which are small silicon chips designed to simultaneously measure the expression of thousands of genes. Typically, expression is assessed across several different experimental conditions. For example, expression of genes active in neural response might be studied under several levels of ethanol exposure.

One limitation of approaches based on experimental organisms is that they study genetic variation as observed within a fixed range of environmental conditions. They do not identify genes that are fundamentally important to a phenotype but for which all individuals within a population have the same alleles. Similarly, some genes may not create phenotypic variation within the range of environments that are studied, but could produce important phenotypic variation under environmental manipulations produced experimentally or through ecological changes.

Other experimental techniques study gene function by altering particular genes and observing the behavioral effects. One of the most common types of these induced mutation studies is a 'knockout' animal, where the activity of a specifically targeted gene is disrupted. Another method is using transgenic designs that introduce genetic variation from one strain into another to study how gene effects change when expressed in a different genetic background.

Contemporary Statistical Methods Based on Measured Genotypes in Humans

Most behavioral phenotypes are assumed to be continuously distributed and multifactorial polygenic – arising from multiple genetic as well as environmental sources. Estimates of the variation attributable to these sources are based on classical biometric genetic theory, which specifies the relations among the underlying components contributing to a trait (e.g., Falconer and Mackay, 1996). Standard assumptions are that the effects of involved genes are additive, segregate independently in the population (i.e., the variants at one gene are uncorrelated with the variants at others), are not sex-linked, and do not influence spousal selection.

Estimating Sources of Family Resemblance

These biometric assumptions are important because they lead to statistical expectations about the observed correlations among pairs of relatives. Specifically, if these assumptions hold, the genotypes of pairs of first-degree relatives (full siblings, fraternal twins, and parent–offspring) correlate 0.50, genotypes of second-degree relatives (grandparents–grandchildren, aunt–niece, half-siblings) correlate 0.25, and those of third-degree relatives (cousins) correlate 0.125. Unrelated individuals living together, such as step-siblings or parents with their adopted children, are expected to have uncorrelated genotypes. If all resemblance between relatives in a family is due to genetic sources, the correlations between their observed phenotypic scores are expected to be proportional to the correlations between their genotypes.

Nongenetic familial influences are implicated in these models by the extent to which the pattern of familial resemblance differs from genetic resemblance. If phenotypes are influenced by shared environments (e.g., rearing or social background), the highest correlations would be found for pairs of twins and similar-aged siblings (including step or adoptive siblings), lower correlations for parents with children, and still lower from related individuals who have not lived together. Under a purely environmental etiology, the correlations of adoptees with their adopting parents and siblings should be as high as for any other children, and related individuals reared in different environments would be uncorrelated. A model that includes both genetic and environmental sources of familial resemblance would produce estimates intermediate between these two. All such models can include measurement error (unreliability) and the influence of experiences unique to particular individuals.

Structural equation models are used in behavior genetics to organize, estimate, and test aspects of these theories of underlying influences. In psychometric models, the observed phenotype or trait can be considered a combination of true score and specific factors, including error of measurement. Behavior genetic models partition true-score variation into unobserved variables representing biometric hypotheses: additive genetic scores (A), dominance and other nonadditive genetic scores (D), common family environment (C), and individual-specific environmental sources (E). The goal of the structural analysis is to use the data from relatives to estimate the total impact of each component. In studies of a single phenotype, the E component includes measurement error. These can be separated when the trait is a latent factor based on multiple measures.

Family, adoption, and twin designs rely on different assumptions and can provide different estimates of the theoretical components. In all cases, the goal is to obtain genetic and environmental estimates that will generalize to explain behavioral variation in the general population. To the extent that the samples studied are not representative, the resulting estimates may be biased.

Family and Adoption Designs

Figure 1 shows several path models representing the variance components that can be estimated using different configurations of relatives. In practice, nonadditive genetic sources (D) are less commonly found, so these examples focus on the other components. Figure 1(a) depicts a nuclear family design. In this model, resemblance between parents and offspring cannot be partitioned into separate additive genetic (A) and common family environment (C) sources.

In the adoption design shown in Figure 1(b), these A and C components are (theoretically) separable: the only source of parent–child resemblance is through environmental sources. By comparing differences in resemblance among family members in the adoptive versus biological family designs, an estimate of the additive genetic contribution to phenotypic variation can be obtained. A potential limitation of the adoption design is selective placement, in which adoptees are not

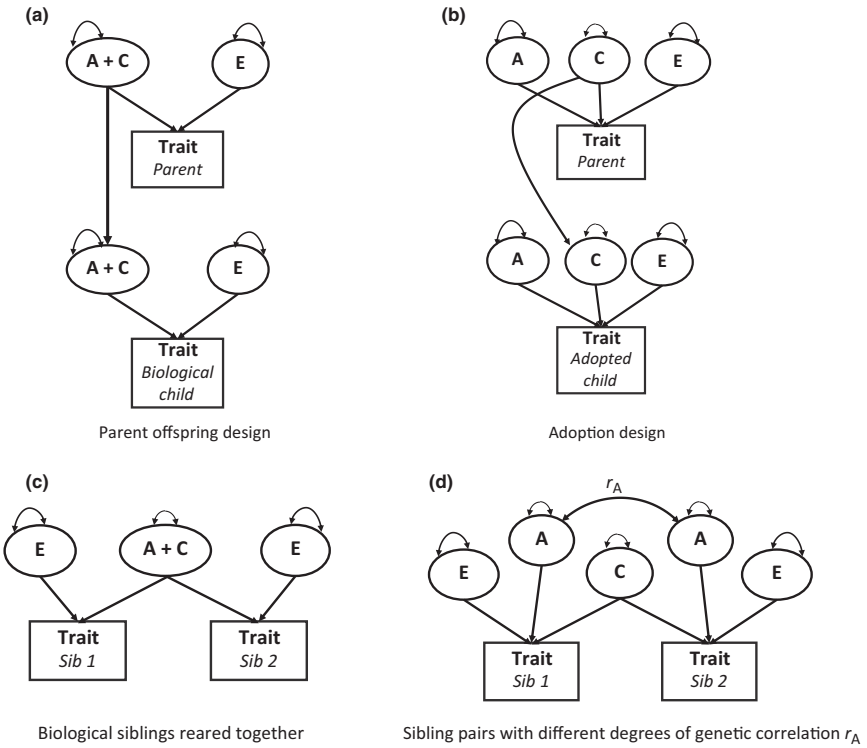

Figure 1 Latent variable path models for estimating sources of trait variation from different types of family designs. Boxes indicate measured variables (Trait score), circles represent unobserved theoretical variables. One-headed arrows are fixed regression coefficients. Two-headed arrows represent estimated variances or correlation coefficients. Latent variables are: A = additive genetic variation; C = common family environment; E = individual environment and measurement error. (a) Parent offspring design. Children resemble their biological parents due to transmitted genetic and environmental factors, which are not separable. (b) Adoption design. Children resemble their adoptive parents only due to environmental factors, allowing separation of genetic and family environmental effects. (c) Biological sibling design. Siblings reared together resemble each other due to transmitted genetic and environmental factors, which are not separable. (d) Sibling pairs with different degrees of genetic correlation can be used to separate genetic and family environmental effects.

randomly assigned to their adoptive families. To the extent there is positive selective placement that is not accounted for in the model, family environmental effects may be overestimated. Another limitation is that adoptees and adoptive parents may not be representative of the trait being studied – with adoptees more likely to have biological parents with disadvantaged backgrounds (including their prenatal environments) and adoptive parents coming from higher socioeconomic backgrounds screened for serious psychopathology.

Sibling and Twin Designs

Another common set of designs is based on the resemblance of siblings reared in the same family. These models do not address mechanisms for transmission between generations, as do the parent–offspring designs, but they are better for studying traits that are influenced by cohort- or age-specific factors. Figure 1(c) depicts sources of resemblance between pairs of full biological siblings. As in the parent–offspring model 1(a), genetic and shared environmental factors are confounded. However, if the design is expanded to include siblings with varying degrees of genetic correlation, represented by r_A in Figure 1(d), the A and C components can be distinguished.

The most influential sibling design is the twin design, which compares the resemblance for a trait among identical (monozygotic, MZ) twin pairs to that of fraternal (dizygotic, DZ) pairs. The genotypes of MZ pairs are correlated 1.0, whereas those of DZ pairs are correlated 0.50. MZ and DZ pairs are assumed to be equally similar for aspects of the common environment relevant to the trait being studied. If the pair correlation for a trait is the same for MZ and DZ pairs, common environmental sources (C) are implicated as the source of family resemblance and genetic sources (A) are not involved. To the extent MZ pairs are more similar than DZ pairs, genetic sources are implicated. Comparing the similarity of male–female DZ pairs to that of like-sex DZ pairs provides estimates of sex-specific genetic and environmental factors.

Key assumptions in the twin design are that mating of parents is random with respect to the trait and that the trait-relevant environments of MZ and DZ pairs are equally similar. To the extent that mating is trait-related, genetic influences will be underestimated. To the extent that the environments of MZ pairs are more similar than those of DZ pairs, genetic influences will be overestimated.

Comparisons of other types of siblings follow the same logic as in the twin design: if siblings with a higher genetic correlation are more phenotypically similar than pairs of

siblings with a lower genetic correlation, genetic sources are implicated as contributing to the trait. The assumed correlation of the genotypes of full siblings is 0.50, half-siblings is 0.25, and step-siblings is 0. In practice, there may be a small correlation between the genotypes of step-siblings because spouses in remarriages tend to be similar to those from prior marriages.

Twin studies are preferred to nontwin sibling designs because they are statistically more powerful (i.e., require smaller sample sizes) and the assumption of equal environment is considered more valid for twins than for siblings of different ages. Nontwin full siblings tend to be less similar than DZ twin pairs for most behavioral phenotypes. This suggests the importance of age or environmental influences on the phenotype. However, a genetic explanation is also possible; sibling groups usually include some half-siblings who are unaware they do not share a father, making the actual genetic correlation lower than the expected value of 0.50.

Two advantages of the sibling design compared to the twin design are ease of recruitment and sample representativeness. For most behavioral phenotypes, twins are similar to the general population, but they have higher rates of intellectual disabilities, possibly related to higher rates of complications in twin pregnancies and births. Recently, the frequency of DZ twin and other multiple births has increased markedly due to increases in maternal age and use of assisted reproductive technologies. Older age and hormonal treatments increase the chances that multiple eggs will be ovulated in the same cycle. With *in vitro* fertilization multiple zygotes are typically implanted together, resulting in more twin births. To the extent that these interventions are not random, but are differentially employed (e.g., by mothers with greater economic resources), DZ twins in younger samples may not be representative of the broader population. To some extent family, adoption, and twin studies complement each other and hybrid approaches help address the weaknesses in specific designs. Studies of the offspring of twins can distinguish different mechanisms for parent–offspring transmission. In typical twin studies, the 'common environment' includes all nongenetic sources of twin pair similarity. The composition of the environment can be clarified in studies of individuals with different degrees of genetic and environmental sharing including twins, siblings, and classmates, or twins reared apart.

Heritability and Interpreting Genetic Estimates

After biometric model coefficients are estimated, it is common to standardize the unobserved variables, and assuming that the latent scores (A, C, and E) are uncorrelated, calculate the unobserved genetic variance. To compare results for phenotypes with different measurement scales, it is common to calculate the *heritability ratio* (h^2): genetic variance divided by total phenotypic variance. This calculation may or may not account for other components such as shared environment and nonadditive genetic influence (dominance, epistasis). Comparing heritability estimates across samples is complicated by the fact that samples may differ in total variance, leading to much different heritability estimates for the same amount of genetic variation. For example, in a society that provides adequate nutrition for all its people, variation in height would be expected to be nearly all genetic (i.e., high heritability), whereas a society for which the nutritional environment was more variable would have greater phenotypic variation, of which a small proportion would be genetic (lower heritability).

So far, studies using measured gene designs account for only a fraction of the genetic variation that is implicated in studies using twin and family designs. A variety of factors may underlie this discrepancy, including nonadditivity of genetic effects, gene–environment interactions (GxE), and greater measurement error and heterogeneity in methodology in samples included in GWAS studies (see Section Methodological Issues for Future Genetic Studies of Behavior).

Gene–Environment Interplay

As genetic variants influencing behavior are identified, the next goal is to clarify how their expression varies over the life span and in response to environmental variation. Measured gene–environment interactions have been studied extensively in model organisms, but there are few replicated findings for complex behavioral phenotypes in humans. GxE studies are subject to a variety of methodological concerns including multiple testing and using measures of the 'environment' that also reflect genetic variation. Nevertheless, experimental evidence indicates that ideas about genetic effects need to be more dynamic and to consider the impact of variation due to environmental contexts, be they cellular, intrapersonal, or societal.

Methodological Issues for Future Genetic Studies of Behavior

The advent of measured gene technology has been accompanied by an expansion of study designs, analytical methods, and tools for data management. However, in order to apply these methods to the impressive amounts of accumulating genetic data, yet more novel approaches to the research are needed. This final section discusses some of the methodological issues involved and solutions being proposed in analyzing measured genotypes in future studies of behavior.

Sample Size and Statistical Power

GWAS uses a hypothesis-free design for marker selection that represents an unbiased method for identifying multiple SNP markers associated with a complex trait at the same time. Because complex traits are likely influenced by additive and individual effects of multiple genetic variants simultaneously, effect sizes for individual variants identified through GWAS are typically very small and very large sample sizes. Often samples including tens of thousands of individuals are required to detect statistically significant associations, depending on the effect size and design (i.e., population vs case-control). With this prohibitive size requirement, GWAS are often designed to test only alleles that are more commonly occurring in the population (i.e., SNP frequency greater than 5%).

To accommodate testing on such large samples, contemporary GWAS methods are designed for time and cost

efficiency. GWAS takes advantage of using fewer markers to capture allelic variation in a gene region, rather than testing each individual SNP. Of the millions of SNPs in the human genome, testing a fraction of SNPs in GWAS are sufficient to represent all common variation across the genome, depending upon the number of marker SNPs estimated to segregate together. Minimizing the number of SNPs optimizes GWAS for identifying associations between those occurring in higher frequency (versus rare SNPs) as well as traits that occur more commonly in the population (versus rare diseases).

Amassing the large sample sizes required for GWAS is often beyond the scope of a single research group. This has spurred the creation of large collaborative networks involving hundreds of investigators. Greater collaboration is a positive development for science, but the validity of combining existing data sets depends on comparability of measurement, sampling, and population ancestry. Phenotypes that can be combined across samples are ones that are easily measured (e.g., height) and may be of less interest for behavioral and social science.

As in other areas of science, the statistical power in measured gene studies can be increased by reducing measurement error. Using more precise measures or creating latent variables from multivariate and longitudinal data substitutes for greater sample sizes (Wong et al., 2003) and reduces problems associated with pooling data across studies.

Type I Error and Replication

The efficiency of being able to test millions of SNPs in single scans brings forth issues of multiple testing, which can yield a high rate of false positives (Type I error), and requires adjustment of statistical significance thresholds. A commonly cited threshold for genome-wide significance is $p < 5 \times 10^{e-8}$, based on an alpha set to 0.05 and testing 1 million SNPs (Risch and Merikangas, 1996). However, many researchers calculate their own probability-level adjustments or use less stringent criteria to avoid false negatives (Type II error). Current consensus among the field is that the reliability of association results should be based on replication in independent samples rather than significance levels. Many journals will not consider publication of association results unless there is evidence of replication in an independent sample. An alternative is using an internal replication design where GWAS is conducted in one portion of the sample, then the top SNPs are tested for association in the remaining sample.

Candidate gene studies reduce the multiple testing problem by hypothesizing an association between genetic loci and the phenotype, often from *a priori* evidence from GWAS, linkage, or gene expression studies. Early candidate gene studies typically employed a limited set of gene variants (or nearby markers), but customized chips for testing genes hypothesized to be involved in a particular biological pathway or condition (e.g., cardiovascular disease, addiction) include tens of thousands of SNPs, requiring modification of probability levels. It is common to apply Bayesian methods (e.g., Benjamini and Hochberg, 1995) to adjust for prior evidence of association in order to safeguard against spurious associations, while helping to maintain statistical power. Positive associations implicating a biological pathway require replication in independent samples or through experimental studies.

Population Stratification

A nontrivial concern for association studies is when the trait and genetic variants are correlated from population stratification rather than because the allele is causally involved. Population stratification, or population structure, occurs when ancestral differences between subpopulations in the study sample result in different allele frequencies between them (Tian et al., 2008). For example, spurious gene–trait associations can arise when ethnic groups differ in both allele frequencies and trait distributions. One way to detect stratification is genomic control, testing if genetic markers that are not hypothesized to be associated with the phenotype deviate from the expected distribution (Devlin and Roeder, 1999; Freedman et al., 2004). A common method for minimizing the effect of stratification is adjusting for ancestral similarity. This is done by conducting principal components analysis (PCA) to identify a subset of genetic variants, or ancestry informative markers (AIMS), that serve as quantitative markers for assigning ancestral similarity (Price et al., 2010). Then, top principal components, those that explain the greatest variance, are then used as covariates in association models.

Future Challenges

A great deal of progress has been made in developing new methods for studying the contribution of genetic variation to human behavior, with more innovative approaches rapidly being disseminated. It will be an ongoing challenge for data management and analytic tools to keep pace with the vast amounts of information generated by microarray chip technology, sequencing advancements, and future developments in genetic networks.

See also: Behavioral Genetics: Psychological Perspectives; Bioethics: Genetics and Genomics; Direct-to-Consumer Personal Genetic Testing; Familial Studies: Genetic Inferences; Gene–Environment Interactions in Health and Well-Being; Genetic Screening for Disease; Genetics and Forensics; Genetics of Personality; Human Mate Choice; Stem Cell Research and Cloning; Twin Studies and Epigenetics.

Bibliography

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 289–300.

Devlin, B., Roeder, K., 1999. Genomic control for association studies. Biometrics 55, 997–1004.

Elston, R.C., 1998. Linkage and association. Genetic Epidemiology 15 (6), 565–576.

Falconer, D.S., Mackay, T.F.C., 1996. Introduction to Quantitative Genetics, fourth ed. Longman, London.

Freedman, M.L., Reich, D., Penney, K.L., McDonald, G.J., Mignault, A.A., Patterson, N., Gabriel, S.B., Topol, E.J., Smoller, J.W., Pato, C.N., Pato, M.T., Petryshen, T.L., Kolonel, L.N., Lander, E.S., Sklar, P., Henderson, B., Hirschhorn, J.N., Altshuler, D., 2004. Assessing the impact of population stratification on genetic association studies. Nature Genetics 36 (4), 388–393.

McArdle, J.J., Allison, D.B., 2001. Genetic studies of behavior: methodology. International Encyclopedia of the Social and Behavioral Sciences 9, 6108–6116.
McArdle, J.J., Goldsmith, H.H., 1990. Some alternative structural equation models for multivariate biometric analyses. Behavior Genetics 20 (5), 569–608.
Neale, M.C., Cardon, L., 1992. Methodology for Genetics Studies of Twins and Families. Kluwer, Dordrecht, The Netherlands.
Price, A.L., Zaitlen, N.A., Reich, D., Patterson, N., 2010. New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics 11 (7), 459–463.
Risch, N., Merikangas, K., 1996. The future of genetic studies of complex human diseases. Science 273, 1516–1517.
Spector, T.D., Sneider, H., MacGregor, A.J. (Eds.), 2000. Advances in Twin and Sib-Pair Analysis. Oxford University Press, London.
Spencer, C.C., Su, Z., Donnelly, P., Marchini, J., 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genetics 5 (5), e1000477.
Tian, C., Gregersen, P.K., Seldin, M.F., 2008. Accounting for ancestry: population substructure and genome-wide association studies. Human molecular genetics 17 (R2), R143–R150.
Tryon, R.C., 1940. Genetic differences in maze learning ability in rats. Yearbook of the National Society for the Study of Education 39, 111–119.
Wong, M.Y., Day, N.E., Luan, J.A., Chan, K.P., Wareham, N.J., 2003. The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? International journal of Epidemiology 32, 51–57.
Wright, S., 1920. The relative importance of heredity and environment in determining the piebald pattern of guinea pigs. Proceedings of the National Academy of Sciences 6, 320–332.

Relevant Websites

http://www.dnalc.org – DNA Learning Center, Cold Spring Harbor Laboratory.
http://www.dnaftb.org – DNA from the Beginning, Cold Spring Harbor Laboratory.
http://geneed.nlm.nih.gov – GeneEd: Genetics, Education, Discovery; U.S. National Library of Medicine, National Institutes of Health.
http://ghr.nlm.nih.gov – Genetics Home Reference, U.S. National Library of Medicine, National Institutes of Health.
http://www.scientificamerican.com/topic/genetics/ – Genetics, Scientific American Magazine.
http://learn.genetics.utah.edu – Genetic Science Learning Center, University of Utah.
http://www.nature.com/ng/index.html – Nature Genetics, Nature Publishing Group.
http://www.genome.gov – National Human Genome Research Institute, National Institutes of Health.
http://www.plosgenetics.org – PLOS Genetics, Peer-Reviewed Open-Access Journal, Public Library of Science.
http://www.nature.com/scitable/topic/genetics-5 – Scitable, Nature Education, Nature Publishing Group.
http://www.yourgenome.org – Your Genome, Sanger Institute, Wellcome Trust.

Genetics and Anthropology

John H Moore, University of Florida, Gainesville, FL, USA
Jonathan Marks, University of North Carolina–Charlotte, Charlotte, NC, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.H. Moore, volume 15, pp. 6135–6141, © 2001, Elsevier Ltd.

Abstract

As soon as fundamental principles of genetic inheritance were clearly established in the early twentieth century, anthropologists began using these principles and new empirical data to illuminate long-standing problems of human variation and primate phylogeny. Initially focusing on human blood characteristics, geneticists quantified regional differences in ABO genotype, and tried to correlate these differences with more traditional osteometric assessments of variation. Anthropologists were especially interested in alleles like the sickling trait, which had adaptive significance. When blood samples became available from other primates, in midcentury, it became possible to construct phylogenetic trees based on genetic data, and to compare these results with the fossil record. After World War II, field expeditions were organized by anthropological geneticists to collect far-flung samples of human genetic material, along with new cultural and linguistic data, in the hope of finding correlations, thereby detecting evolutionary trends and reconstructing human migration. With the addition of ancient DNA techniques to the anthropological repertoire in the 1980s, all four fields of anthropology – ethnology, archaeology, linguistics, and human biology – became intimately involved in genetics research, both individually and in cross-disciplinary research. The immediate research agenda for anthropological genetics includes (1) human origins; (2) prehistoric migration; (3) coevolution of biology, language, and culture; and (4) variation and adaptation to modern environments.

The fields of anthropology and genetics maintained a tense coexistence throughout the twentieth century, sometimes complementary to one another (as in the decades immediately following World War II) and sometimes antithetical to one another (as in the eugenics movement of the 1920s and the genetic essentialism of the 1990s). The reason is that the self-interest of geneticists is sometimes perceived to lie in the assumption that genetics determines the most important aspects of a person's life or identity (Nelkin and Lindee, 1995), which would run counter to the anthropological focus on culture. Normative human genetics after World War II, led by Theodosius Dobzhansky, James V. Neel, and Luca Cavalli-Sforza, tended rather to emphasize the complementarity of the two fields. Here, human genetics would form a foundation for the study of human microevolution, and its primary data would be biochemical 'markers' of ancestry, or intergenerational continuity. Within anthropology, then, normative genetic research came to incorporate four kinds of questions: (1) patterns of human variation; (2) the phylogenetic relationships among humans and other primate species; (3) the migration and adaptation of human populations, or microevolution; and (4) the relationships among biological variation and other kinds of human variation, especially in language and culture. Additionally, there are areas of overlap between genetics and humanistic anthropology, including medical anthropology (Taussig, 2009), economic anthropology (Pálsson, 2007), bioethics (Brodwin, 2005), and political ecology (Stone, 2010).

Human Variation

When laboratory techniques for studying Mendelian traits first became available, international anthropology had just emerged from a controversy over racial typology during which Franz Boas, recognized by many as the founding father of American anthropology, had repudiated to the satisfaction of most anthropologists the racial theories to explain civilization that had been popular in the late nineteenth century. Originally trained as a physicist, Boas presented data and arguments between 1910 and 1913 showing that 'race,' 'language,' and 'culture' were three independent phenomena that, he asserted, should all be studied by anthropologists, but using different kinds of theories and techniques.

Shortly after World War I, the earliest population genetic studies of human blood groups seemed to undermine the ideas about race that physical anthropology had traditionally employed, although it would be several decades before the meaning of this discordance became clarified (Marks, 2012). After World War II, as human genetics became increasingly redirected toward real genetically based medical pathologies (such as sickle-cell anemia), rather than imaginary genetically based social pathologies (such as feeblemindedness), it became clear that the simple rules of Mendelian inheritance applied only to the inheritance of biochemical variants and pathological conditions resulting from the breakdown of specific genes (the repository for medical genetic conditions is Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/omim). The interesting physical attributes of human variation (such as height, complexion, facial form, hair form, and body build), however, are all the result of more complex hereditary and developmental physiologies, involving the coordinated activities of several genes and regulation of their products, in specific environmental, genetic, and cultural contexts.

For anthropologists, the major theoretical model that emerged to describe human genetic variation in this period was the cline, the gradual geographic change in the frequency of an allele, which also served to undermine further the idea of racial typology. Both physical form and genetic variation seemed to

vary across the human species in a continuous, quantitative fashion.

Some patterns that emerged from the analysis of clines conformed to more general biological principles pertaining to other animals. It was found, for example, that Gloger's, Bergmann's, and Allen's Rules for variation within a mammalian species (pertaining to pigmentation, mass, and limb length, respectively) were also crudely valid for the human species, reflecting adaptation to the different latitudes, climates, and environments inhabited by local human populations. Since these traits were so clearly adaptive, it was suspected that other patterned traits, with definite clines, likewise reflected genetic adaptation to local conditions. Alternatively, they might be the track of large-scale migrations of ancient peoples.

Adaptation

The most forceful new example of human genetic adaptation was presented by Livingstone (1958) with his study of sickle-cell disease, especially as it existed in Africa. In addition to presenting convincing evidence of several types to prove that sickled blood cells were a long-term genetic response to the presence of malaria, Livingstone's work illustrated several other lessons in anthropological genetics. First of all, it was the heterozygote (a person with two different alleles for a specific gene) or carrier of hemoglobin S (HbS) who had the adaptive advantage, not the HbS homozygote (possessing two identical alleles for a specific gene), who suffered from sickle-cell anemia. The second lesson was that an apparent 'disease,' sickle-cell anemia, was on closer inspection actually a successful genetic solution to a largely human-made environmental problem, namely, malaria – stimulated by the widespread presence of mosquitoes, which in turn were attracted by large areas of standing water that were promoted by early agricultural practices. The story does not end in the past, however, as medical anthropology and history are now illuminating complementary anthropological aspects of sickle-cell anemia (Wailoo and Pemberton, 2006; Fullwiley, 2011).

Another highly visible connection between genetic variability, culture, and environment concerns the ability of human adults to digest milk and milk products from dairy animals, an ability that seems to be under genetic control. When mapped geographically, the alleles responsible for adult lactose tolerance in human populations seem to be correlated, in the Old World, with the present and past locations of pastoral peoples (Durham, 1991). Lactose tolerance has arisen in parallel in several different areas, and the most common mutation in Europeans probably spread with the descendants of early dairy farmers (Ingram et al., 2009).

Many correlations have been found between pathogenic diseases and genetic variations. None, however, has been established as firmly as the relationship between malaria and sickle-cell anemia, and other related blood diseases. Different patterns of genetic variation among populations are explained as accidents of demographic history, or founder effects, rather than as adaptive responses of the gene pool. The most famous examples here are genetic diseases that are rare elsewhere but have attained an anomalously high frequency in a particular population, for example, porphyria variegata among the Boers of South Africa and Ellis–van Creveld syndrome among the Pennsylvania Amish – both of which have elevated frequencies because a single seventeenth-century immigrant was both a carrier of the genetic disease and an ancestor of the modern populations. In the absence of plausible and reliable physiology (as in the case of sickle-cell or thalassemia) or plausible and reliable historical records, most other diseases associated with genetic variants show more ambiguous patterns.

More generally, understanding the role that genetic variation may play in health care involves the analysis of risk factors and consequent probabilities (Terwilliger and Göring, 2000). Nevertheless, the presumption of a genetic basis for health disparities carries strong political implications and can hardly be taken at face value (Fullerton et al., 2012) Moreover, the saturation of medical genomics with modern capitalism may compromise the reliability of the scientific information it produces. BiDil, promoted as a cardiac medicine specifically for black patients, was actually supported by no valid evidence that it worked differently or better in blacks than in whites. Nevertheless, racializing the drug extended its patent protection. And far from being a public health service, BiDil's manufacturers charged their black patients far more than they needed to pay for it, since the drug itself was simply a mixture of two other generic drugs (Kahn, 2008).

Population Genetics

With increased funding for science in the United States after World War II, and increased access to field sites, anthropologists began to investigate isolated, small-scale human populations with renewed vigor, often including a biological anthropologist in the research team. The most visible of these joint enterprises was probably the comprehensive study of the Yanomamö, who live in the upper drainage of the Orinoco River in Venezuela and Brazil. In the 1960s, James Neel of the University of Michigan organized a comprehensive multidisciplinary series of field expeditions that investigated the environment, ecology, biology, culture, and social organization of these then little-known people. After a series of books and articles was published over the next two decades, the Yanomamö became not the least known, but one of the best-known tribal societies in the world.

For the genetic part of the research, Neel and Ward (1970) examined six blood systems relatively well known at the time (MNS, Rh, Kidd, Duffy, Diego, and haptoglobin) among 7 villages of Yanomamö, 7 villages of nearby Makiritare, and 12 other tribes of Central and South America. After comparing the different groups, the authors concluded that human evolution might have been proceeding at a pace '100 times more rapid' than previously thought, if these small-scale societies could be taken as analogs of earlier Pleistocene or Paleolithic human peoples.

In the same period, other biological anthropologists were proceeding with similar investigations elsewhere in the world. Working in the Solomon Islands from 1966 to 1973, Jonathan Friedlaender simultaneously undertook studies of demography, anthropometry, genetics, and social structure, seeking

to integrate his results into some general conclusions that cut across these traditional fields of research. Like Neel and Ward, he emphasized the genetics of blood systems, in this case ABO, MNS, Rh, Kell, and Duffy. From this kind of data, Friedlaender calculated phenotype frequency and then allele frequency by techniques described in his major ethnographic work, *Patterns of Human Variation* (Friedlaender, 1975). His most general conclusion, combining results from anthropometry, dermatoglyphics (the study of fingerprints), genetics, and demography, was simply that Bougainville communities tended to exhibit high biological variability within the group, but low variability among the groups, and that some traits, such as skin color, eye color, and hair form, were 'monotonous' on the island, while dermatoglyphic characteristics and some blood system alleles were discontinuous from village to village.

Those expecting that these early studies would provide fundamental insights into the mechanisms of human evolution were soon disappointed. Summarizing a widely shared consensus of opinion in 1974, Henry Harpending concluded that (1974: p. 229)

> *… studies of the genetic structure of small populations have made particular and incidental contributions to formal genetics, to regional history and prehistory, to epidemiology, and to several other fields to which they are peripheral, but … they have not advanced our understanding of human evolution in a global sense. The sample sizes available have been too small to allow reliable inferences about natural selection; the extensive occurrence of what is presumably local random genetic drift has little or no consequence for evolution over long time periods over large areas; and the presumed selective agents in the various environments of these peoples differ greatly so that few of the generalizations which have been put forward hold for many groups.*

Evolution of Primates

Early in the twentieth century, it was appreciated that the relations of primate blood appeared to replicate the relations of primate bodies; that is to say, animals who appeared more similar physically tended to have more similar biochemical properties discernible in their blood. Further, the blood of humans was especially similar to the blood of apes; indeed, by the 1920s, human blood was known to be even more similar to chimpanzee blood than horse blood was to donkey blood.

In the 1960s, Allan Wilson and Vincent Sarich (1967) showed that the biochemical variation distinguishable immunologically in the blood tended to track the time since the species being examined diverged from one another, rather than their degree of adaptive physical difference from one another. In the ensuing decades, this generalization has held up remarkably well over diverse kinds of genetic comparisons. Consequently, genetics tends to show how similar humans are to chimpanzees, not how different we are. The reason is that we do not know how to build a four-dimensional body (an organism) from a one-dimensional set of instructions for it (the DNA). Absenting that information, detectable DNA changes are largely uncorrelated with the adaptive physical variations that characterize the 'major features' of the history of life. Consequently, we know rather a lot about how similar genetically we are to chimpanzees, but almost nothing about the genetic basis of bipedality, language, evaporative heat loss, cooperative breeding, or any of the other significant ways in which we differ strikingly from the apes.

Sarich and Wilson (1967) used the primate molecular data in the 1960s to demonstrate convincingly that *Ramapithecus*, considered to be on a human line 14 million years ago, on the basis of teeth and jaw fragments, was about three times older than the human line itself was, and therefore could not be on the human line. Morris Goodman invoked the genetic similarity of humans and chimpanzees to argue for classifying them together as members of the genus *Homo* (Wildman et al., 2003). This would involve privileging the genetic relationships (in which chimpanzees and humans are extremely similar) over anatomical, behavioral, cognitive, and ecological relationships (in which they are considerably less similar).

Modern studies of primate genetics incorporate inferences about not only systematics and divergence times (Steiper and Young, 2006), but also demography, paternity, conservation, and microevolution.

Methods

The production of genomic data climbed so precipitously beginning in the 1980s that a new specialty, bio-informatics, had to be developed in order to detect patterns in genomic data and to make quantitative, statistical sense out of them.

The most fundamental tools used in genomic comparisons involve (1) the acquisition, isolation, manipulation, and amplification of genetic sequences to be studied; (2) the qualitative and quantitative assessment of patterns of similarity among the samples, for example, inferring the kinds and numbers of mutations needed to transform one DNA sequence into another; and (3) an explanation for the patterns in the history of the organisms.

The study of the human genome has revealed that traditionally defined genes (containing the DNA code for a specific protein) constitute only a small percentage of the total DNA in a human cell. About 90% of the human genome is intergenic, that is to say, lying between genes and thus not classically functional. Moreover, most of a gene itself does not actually code for a protein; it may be transcribed into RNA without being translated into protein. Two obvious conclusions are (1) our classical ideas about the functions of DNA were too narrow and (2) there are some interesting structural features of noncoding DNA. Some noncoding DNA is localized, and consists of simple sequences repeated side-by-side millions of times. Other repetitive DNA is interspersed among the genes, and consists of millions of sequences, each one hundreds or even thousands of bases long, and shot through the genome in apparently random places. This redundancy is characteristic of the genome. Genes themselves are commonly found in clusters, the ancient products of a 'rubber-stamp' process of gene duplication, followed by the mutation-mediated inactivation or alteration of the duplicate gene. Other sites in the genome differ from person to person in the number of segments of a short repetitive motif that may be present. These

polymorphisms are known as variable numbers of tandem repeats (VNTRs) and are particularly valuable in forensic contexts, where a match of several highly variable sites can provide a statistical basis for identifying the person that a particular DNA sample came from.

We can generally measure the intensity of selection on a particular genomic region by its rate of change. Noncoding sequences invariably are more different across species than coding sequences are. The simple reason is that coding sequences are expressed as phenotypes, and have evolved by natural selection to produce a functioning body. Any random change within them (mutation) is far more likely to make that body run worse than better. Consequently, mutations that occur in coding sequences are far more likely than mutations in unexpressed DNA sequences to be 'weeded out' by selection. Establishing the background extent of genetic difference between two species helps to identify regions that are evolving 'too slowly' and may thus be under selective constraints. More rarely, a genomic site may be of interest for evolving 'too fast' and thus possibly be related to the adaptive differences that arose between the species.

Stains

As soon as chromosomes could be observed directly through microscopes, attempts were made to stain them chemically so that bands would appear that might show differences in structure. Initially, the purpose was to discriminate among the different chromosomes of the human karyotype, or suite of chromosomes (22 pairs of autosomes, and a pair of sex chromosomes), so they could be counted and identified. During the course of this research, it also became apparent that patterns of stains on a particular chromosome might vary from one individual to another within a population. In addition, when it was noted that certain patterns of staining were characteristic of particular species, yet another method became available for evaluating the phylogenetic relationships among apes and humans. Although several correlations have been established between the observable banding pattern of chromosomes and the DNA sequences they contain, the staining difference presumably results from a higher-order property of DNA (i.e., its three-dimensional folding).

Biotechnological Breakthroughs

The age of biotechnology was made possible by the development of several laboratory methods. The first was the discovery that biomolecules migrate in an electric field, and can be separated from one another and isolated; this is known as electrophoresis. The phosphates in DNA give it a negative charge. Embedded in a semisolid matrix, then, DNA moves toward a positive pole, its shortest fragments moving fastest. Being able to separate DNA molecules from one another forms the basis of all subsequent DNA research.

The second technical breakthrough involved the isolation and purification of naturally occurring enzymes that modify DNA, and can be used to manipulate it in predictable ways. One such class of enzymes is endonucleases, which cut DNA at specific sites (such as GAATTC for the enzyme called EcoR1). Many such enzymes became available in the 1980s, permitting the construction of 'maps' of the DNA where different enzyme cut sites are located.

Another class of enzymes permitted the development of the third technical breakthrough, the polymerase chain reaction. Here, the enzyme is a DNA polymerase, which adds the appropriate complementary nucleotide to produce a stable, double helix from a single-stranded DNA template. By alternating cycles of heating (to dissociate the double helix into single strands of DNA) and replication (using the polymerase to fill in the opposite strand), small amounts of DNA can be amplified sufficiently to be analyzed (Hillis et al., 1996). This is applicable in modern forensic contexts as well as for organic remains that are thousands of years old.

Chromosome Mapping

A major objective of human genetics research is to locate all genes on their respective chromosomes as well as to identify all the other functional regions of DNA, including some formerly dismissed as 'junk.' One technique for gene mapping involves the use of genetic markers, like the restriction enzyme sites described in this article, or other DNA sequence variants. Because many of these are relevant to diagnosing human disease, high levels of funding were made available to the Human Genome Project to identify the location of these variable sequences or 'polymorphisms.' This support has led to the development of very fast computer and molecular technology and machinery, and much of the resultant data appear in easily searched form on the World Wide Web.

Mitochondrial DNA

Most human genes reside in the nucleus of the cell and occur in two copies, one each inherited from the mother and the father. However, the body or cytoplasm of each cell also contains structures known as mitochondria, which are involved in the production of energy for the cell and contain their own DNA. This DNA is special in many ways. At fertilization, the mitochondria of the ovum are transmitted to the offspring, but the mitochondria of the sperm are not. Thus, mitochondrial DNA (mtDNA) is inherited in a strictly maternal manner, such that (unlike chromosomal DNA, in which a child is equally related to father and mother), mitochondrially a child is a clone of its mother and unrelated to its father.

Since mtDNA is haploid (contributed by only one parent) rather than diploid (from both parents) in form, it does not recombine during replication, and over time, as sequence variations accumulate through mutation, it produces distinct, diverging lineages of mtDNA sequences. This variation can be used to reconstruct the mtDNA of the common ancestor (the coalescent sequence) from a sample of mtDNAs collected from individuals alive today. Because the mitochondria are so numerous in the cell, mtDNA can be recovered much more easily than DNA from the nucleus.

The mtDNA chromosome takes the form of a small ring of 16 500 base pairs (16.5 kB), which was entirely sequenced by 1981. In a famous 1987 study, the structure of the mtDNA tree from 147 people was announced, which 'coalesced' on a woman nicknamed 'mitochondrial Eve,'

deduced to have lived in Africa about 200 000 years ago, and whose 'modern human' progeny had emigrated from Africa and repopulated the world, which was at that time occupied entirely by earlier forms of human, such as Neanderthals (Cann et al., 1987).

mtDNA has become very popular in 'recreational ancestry studies,' in which a client's mtDNA is matched against a global reference sample, and clients receive information on their genetic ancestry. However, since (1) the matches are very nonspecific, (2) the world is poorly sampled, and (3) seven of your eight great-grandparents are invisible to this analysis (it only studies your mother's mother's mother), this kind of study has proven somewhat more valuable commercially than scientifically (Bolnick et al., 2007).

Y Chromosomes

Certain regions of the Y chromosome are very stable and do not recombine during meiosis, but are passed on clonally from father to son. Again, while it has some scientific value, by comparing the geographical distributions of mtDNA and Y-chromosome lineages in attempts to reconstruct the migratory practices of prehistoric peoples, much of its use is commercial. Here, clients can learn, for a price, whether they have a Y chromosome inferred (often dubiously) to have belonged to Genghis Khan, Thomas Jefferson, or Moses (Thomas et al., 1998; Jobling, 2012).

Repetitive DNA

Although the picture is constantly changing, researchers at this point recognize three general categories of tandemly repeated sequences (VNTRs), which are classified by length. They are satellites, the longest sequences; minisatellites, defined as sequences that are 500–40 000 base pairs in length, usually consisting of five or fewer repeated nucleotides; and microsatellites, which tend to be short and simple in motif, even consisting of repeats of a single base pair, or a doublet or triad of nucleotides (e.g., CACACA or CAGCAG).

These short, apparently neutral sequences are useful to anthropologists because their rate of mutational change is much higher than the single nucleotide mutation of DNA, so that more variation is detectable across humans and human populations. Thus, minisatellites and microsatellites, like mtDNA, may provide a finer control of historical relationships among individuals and the populations they represent than variation in the nuclear DNA.

Grand Synthesis

The history of gene pools is the history of the interaction of several biological and cultural forces. The first is mutation, or the production of new alleles; that is to say, the genetic variation on which the rest of microevolution rests. The second is natural selection, in which certain genetic variations come to predominate in descendant gene pools on account of the superior fit they consistently provide their bearers to some aspect of the environment. This is the cause of genetic adaptation. The third is genetic drift, in which certain genetic variations come to predominate in the descendant gene pool for nonadaptive reasons, simply blind chance. This would have the strongest effects in small populations, which are of course the kinds of populations that characterize most of human prehistory. For example, a mother who has 10 sons and no daughters will be unrepresented in the next generation's mitochondrial gene pool, while a mother who has only one son and one daughter will be represented; that is to say, the composition of the mtDNA pool is more strongly affected by sex ratio bias than by adaptive selection. The fourth is gene flow, the transformation of gene pools due to interbreeding (or admixture) among members of different communities (Cabana and Clark, 2011). The fifth is inbreeding, mating with genetic relatives, which increases the proportion of homozygotes in a population. If the population includes rare recessive alleles, inbreeding will increase the probability of a single individual inheriting two copies of them. Since we are all to some extent related, all populations are to some extent inbred. Small, circumscribed populations, however, would have higher amounts of genetic drift (randomly elevating the frequencies of rare alleles) and inbreeding (elevating the chances of those recessive alleles being expressed phenotypically).

While in principle the effects of these processes are mathematically distinctive, in practice it is often difficult to distinguish their different effects, except in extreme, rare cases. To establish selection requires not just genetic data but also physiology; to establish founder effect (genetic drift) requires demography. As noted in this article, sickle-cell anemia and thalassemia are solid cases of adaptation, where the selective regime is the endemic threat of malaria; Ellis–van Creveld syndrome in the Pennsylvania Amish and variegated porphyria among the South African Boers are well-established cases of founder effect. Most distinctive genetic variation in human populations lies somewhere in between, such as Tay–Sachs disease in Ashkenazi Jews.

Problems and Prospects

Reconstructions of human prehistory are commonly attempted with genetic data, and are useful for testing hypotheses generated with other kinds of data (Cavalli-Sforza et al., 1994; MacEachern, 2000; Terrell, 2000). The analysis of genetic data often contains hidden assumptions about the demography of ancient populations. Thus, standing alone, genetic conclusions about human history are often model dependent, and may be affected by the strength of selection operating on the DNA sequences being studied; the contractions, expansions, movements, and contacts of the prehistoric populations; the mutation rate; the relationship between the inferred ancestral gene pools and the descendant gene pools being studied; the inferred degree of homogeneity of ancient populations; the inferred modes of evolution (splitting and merging of lineages); and the complex relationship between modern ethnicities (the cultural labels that often guide the sampling process) and ancient gene pools.

The analysis of ancient DNA from human fossils is generating new information on human prehistory and evolution, but to retrieve usable DNA from scarce and damaged human

bones of great antiquity requires extensive precautions against contamination (Hermann and Hummel, 1994). Neanderthal DNA is especially similar to that of living Europeans and Asians, but also different from other archaic human samples (Hofreiter, 2011). We do not know whether the modern human gene pool provides a useful model for understanding the gene pool of our ancestors a few hundred thousand years ago.

Admixture mapping uses carefully chosen polymorphisms that differ significantly across indigenous populations to estimate the degree of relatedness of an unknown sample to each of them. These estimates can then be used to study the heterogeneous ancestry of recently admixed populations. Its applicability is limited, though, for the method uses living populations as proxies for ancestral gene pools, and makes assumptions that are probably unrealistic about the homogeneity (purity) and autochthony (static indigeneity) of ancestral populations.

See also: Anthropology, Genomics, and Human Variation: National Roots; Bioethics in the Post-genomic Era; Community Genetics; Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute; Eugenics as a Basis of Population Policy; Familial Studies: Genetic Inferences; Genetic Ancestry Testing; Genetics and Indigenous Communities: Ethical Issues; Genetics and Sociology; Genetics: The New Genetics; Human Evolutionary Genetics; Human Genome Project: History and Assessment; New Genetics and Race; Primates, Evolution of; Race: Genetic Aspects.

Bibliography

Bolnick, D.A., Fullwiley, D., Duster, T., Cooper, R.S., Fujimura, J., Kahn, J., Kaufman, J., Marks, J., Morning, A., Nelson, A., Ossorio, P., Reardon, J., Reverby, S., Tallbear, K., 2007. The science and business of genetic ancestry testing. Science 318, 399–400.

Brodwin, P., 2005. "Bioethics in action" and human population genetics research. Culture, Medicine and Psychiatry 29, 145–178.

Cabana, G.S., Clark, J.J. (Eds.), 2011. Rethinking Anthropological Perspectives on Migration. University Press of Florida, Gainesville, FL.

Cann, R.L., Stoneking, M., Wilson, A.C., 1987. Mitochondrial DNA and human evolution. Nature 325, 31–36.

Cavalli-Sforza, L.L., Menozzi, P., Piazza, A., 1994. The History and Geography of Human Genes. Princeton University Press, Princeton, NJ.

Durham, W.H., 1991. Coevolution. Stanford University Press, Stanford, CA.

Friedlaender, J.S., 1975. Patterns of Human Variation: The Demography, Genetics, and Phenetics of Bougainville Islanders. Harvard University Press, Cambridge, MA.

Fullerton, S.M., Knerr, S., Burke, W., 2012. Finding a place for genomics in health disparities research. Public Health Genomics 15, 156–163.

Fullwiley, D., 2011. The Enculturated Gene: Sickle Cell Health Politics and Biological Difference in West Africa. Princeton University Press, Princeton, NJ.

Harpending, H., 1974. Genetic structure of small populations. Annual Review of Anthropology 3, 229–243.

Hermann, B., Hummel, S., 1994. Ancient DNA. Springer, New York.

Hillis, D.M., Moritz, C., Mable, B.K., 1996. Molecular Systematics. Sinauer Associates, Sunderland, MA.

Hofreiter, M., 2011. Drafting human ancestry: what does the neanderthal genome tell us about hominid evolution? Commentary on Green et al. (2010). Human Biology 83, 1–11.

Ingram, C.J.E., Mulcare, C.A., Itan, Y., Thomas, M.G., Swallow, D.M., 2009. Lactose digestion and the evolutionary genetics of lactase persistence. Human Genetics 124, 579–591.

Jobling, M.A., 2012. The impact of recent events on human genetic diversity. Philosophical Transactions of the Royal Society B: Biological Sciences 367, 793–799.

Kahn, J., 2008. Exploiting race in drug development: BiDil's interim model of pharmacogenomics. Social Studies of Science 38, 737–758.

Livingstone, F.B., 1958. Anthropological implications of sickle cell gene distribution in West Africa. American Anthropologist 60, 533–562.

MacEachern, S., 2000. Genes, tribes, and African history. Current Anthropology 41 (3), 357–384.

Marks, J., 2012. The origins of anthropological genetics. Current Anthropology 53, S161–S172.

Neel, J.V., Ward, R.H., 1970. Village and tribal genetic distances among American Indians, and the possible implications for human evolution. Proceedings of the National Academy of Sciences of the United States of America 65 (2), 323–330.

Nelkin, D., Lindee, S., 1995. The DNA Mystique. The Gene as a Cultural Icon. Freeman, New York.

Pálsson, G., 2007. Anthropology and the New Genetics. Cambridge University Press, New York.

Sarich, V.M., Wilson, A.C., 1967. Immunological time scale for hominid evolution. Science 158, 1200–1203.

Steiper, M.E., Young, N.M., 2006. Primate molecular divergence dates. Molecular Phylogenetics and Evolution 41, 384–394.

Stone, G.D., 2010. The anthropology of genetically modified crops. Annual Review of Anthropology 39, 381–400.

Taussig, K.-S., 2009. Ordinary Genomes: Science, Citizenship, and Genetic Identities. Duke University Press, Durham, NC.

Terrell, J.E. (Ed.), 2000. Archaeology, Language, and History. Bergin and Garvey, Westport, CT.

Terwilliger, J.D., Göring, H.H., 2000. Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. Human Biology 72 (1), 63–132.

Thomas, M.G., Skorecki, K., Ben-Ami, H., Parfitt, T., Biadman, N., Goldstein, D., 1998. Origins of Old Testament priests. Nature 394, 138–139.

Wailoo, K., Pemberton, S., 2006. The Troubled Dream of Genetic Medicine: Disease and Ethnicity in Tay–Sachs, Cystic Fibrosis, and Sickle Cell Disease. Johns Hopkins University Press, Baltimore, MD.

Wildman, D.E., Uddin, M., Liu, G., Grossman, L.I., Goodman, M., 2003. Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus *Homo*. Proceedings of the National Academy of Sciences of the United States of America 100, 7181–7188.

Genetics and Disaster Victim Identification

Caroline Bennett, University of Kent, Canterbury, Kent, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The use of genetic methods (specifically DNA profiling) is increasingly popular in the identification of disaster victims. Where traditional modes of identification fall short, DNA profiling offers a statistically accurate means of providing positive identification of remains and of reassociating body parts. However, its use is not without problems, including logistical, practical, and social implications. Different cultures have different modes of identification, which may not be scientifically or biologically based, and DNA profiling may conflict with the strict norms and regulations related to treatment of the body after death.

The use of genetics in disaster victim identification (DVI) refers particularly to different methods of DNA profiling; it is one of a suite of methods, often used in conjunction with others, to provide positive, individual identification of human remains following a disaster. A disaster often produces conditions where human remains are degraded, fragmentary, and comingled; as such, the use of DNA is becoming increasingly relied upon to provide authoritative identification in preference over other, traditional, methods such as visual identification, fingerprint analysis, odontological comparison, and osteological analysis.

DNA profiling evolved from the development of other modes of identification, such as the assessment of blood types to determine relatedness, within molecular biology. First commercially available in 1987, the use of DNA profiling as a means of individual identification has developed at a rapid pace over the past few decades. Until recently, it assumed a backseat to other methods of identification, however, due to scientific advances in profiling techniques, combined with increased media coverage and widely publicized successful identification efforts across the globe, and its inclusion in popular media and television shows, it has attained the status among many experts and publics alike as being the ultimate provider of identity. As such it has become almost synonymous with the word 'identification' in the context of DVI; in post-disaster management its use has come to represent responsible, well-ordered, and capable authority; to use DNA analysis in victim identification has become normalized as both a universally appropriate and superior technological intervention. It is therefore often an expected intervention in disaster settings, where many people view identification efforts as central to the process of responsible postdisaster management.

What Is a Disaster?

A disaster is "a sudden, calamitous event that seriously disrupts the functioning of a community or society and causes human, material, and economic or environmental losses that exceed the community's or society's ability to cope using its own resources" (IFRC, 2013). These calamitous events may be natural (such as earthquakes, hurricanes, fires, floods, or other environmentally based events), or of human origin (air disasters or terrorist actions for example), but by definition, they cause events that require external assistance to manage. One of these events is often mass death. Management of the dead is one of the essential actions following a disaster, alongside care for survivors and reestablishment of basic services. It is a crucial aspect of reasserting control and order on a society and people thrown into chaos and disorder by the disaster event. In many (although by no means all) instances, this management includes individual identification of the deceased, and repatriation of the remains. Genetic profiling is becoming an increasingly popular mode of enabling this. Recent examples of disasters that have employed genetic means of identification include 9/11 (2792 deaths); the Asian tsunami of 2004 (over 220 000 deaths); Hurricane Katrina (1836 deaths); the 2009 Victorian bushfires in Australia (173 deaths); and the 2011 Japanese earthquake (over 16 000 deaths).

How Are Genetic Methods Used in DVI?

Whilst the forensic use of genetics is vast, ranging from paternity testing to perpetrator identification, in DVI it is primarily used in two contexts: to provide positive identification of recovered human remains, and to reassociate body parts with one another (to provide as complete a body for identification and repatriation as possible). In the latter case, particularly in disaster situations, the effectiveness of DNA profiling is without comparison as a mode of analysis. In addition, where appropriate, the use of DNA analysis may be used within a disaster to assess perpetrator identification.

The use of DNA profiling enables the matching of one person to another or of one body part to another, through the matching of particular sequences of information within the strands of DNA contained within biological remains. Although all human beings share 99.9% of their genetic information with other members of their species, the remaining 0.1% are unique for every individual (except between monozygotic twins and multiples). It is this variation that enables DNA profiling to be so accurate in its identification (*see* Genetics and Forensics). The specific loci of investigation within the DNA depends upon the method chosen for analysis, and this method depends upon many factors: the aims of the

investigation (identification or body part reassociation for example), the condition of remains, the type of biological material available (teeth, blood, or saliva for example), the environmental conditions of the disaster, and the time elapsed since the disaster.

Despite its reputation as the ultimate provider of identification, DNA profiling is neither a foolproof method of identification, nor is it always successful. Whilst the statistical accuracy of positive identification by DNA analysis is high, conditions must be favorable for a successful match to be made. There must be a reference sample for comparison, either from a close relative, or from belongings of the suspected victim. The remains recovered must be of good enough condition to enable a sample of DNA to be taken; this can depend on the type of remains available (blood, soft tissue, or bone for example), the level of degradation of remains, time lapsed since the disaster (DNA from different materials degrades at different rates – saliva degrades more quickly than blood, which degrades more quickly than bone and so on), and other environmental conditions. In addition, there must be the technical and logistical expertise and facilities, money, and political will to undertake the analysis. This political will depends not only on social norms and values of a given society, but also on the assessed social 'value' of the deceased. Even with the highest level of funding, facilities, expertise, and political will, identification by DNA profiling may not be possible: to date (September 2013) over 40% of the nearly 22 000 pieces of human remains recovered following 9/11 remain unidentified (representing over 1100 victims) (NYC Office of the Chief Medical Examiner, 2013).

Different Types of DNA Profiling in DVI

DNA profiling for genetic identification works by comparing sets of information within certain sections of the DNA to look for a match. Using polymerase chain reaction techniques, laboratories are able to extract and make copies of minute amounts of DNA information that show high variability between populations, or DNA 'markers,' that are useful molecular identifiers. The most common used for identification are short tandem repeat markers, although single nucleotide polymorphism markers may also be used. In DVI cases, the profiling is primarily of DNA markers found either in nuclear DNA or mitochondrial DNA (mtDNA). The relative advantages and disadvantages of either type are outlined in **Table 1**. For a more comprehensive outline of the science of DNA profiling for forensic use, please refer to Zietkiewicz et al. (2012).

There is constant scientific research into new and improved methods of DNA profiling and genetic identification methods (such as the analysis of RNA, next generation genome sequencing); however, these have yet to be useful effectively in a disaster setting.

The Complexities of DVI

Postdisaster management, including identification of the dead, is the responsibility of the country within which it occurs. However, a disaster, by definition, causes events that are either too big or too complex for individual countries to manage

Table 1 Advantages and disadvantages of the profiling of nuclear DNA (nDNA) vs mitochondrial DNA (mtDNA) following a disaster.

	nDNA	*mtDNA*
Cellular location	Nucleus of every human cell except red blood cells.	Cytoplasm of mitochondria – structures within cells that convert energy.
Mode of inheritance	Mendelian	Matrilineal
Advantages in DVI	Mendelian inheritance provides high level of statistical accuracy to identification. Flexibility of analysis methods: can be analyzed using STR markers or SNP markers. High mutation rate of STRs make them highly individualized and ideal for individual discrimination (the statistical probability of one STR frequency matching that of another individual is less than 1 in 1 billion). Low mutation rate of SNPs and their ability to be amplified from minute amounts of DNA makes them ideal in very degraded specimens. SNP analysis exhibits population variability and can therefore be useful in assessing ethnicity.	Matrilineal inheritance means any maternal relative can be used as a reference sample – useful therefore where insufficient information is available within the nucleus, such as very degraded specimens, or where only certain relatives are available for comparison. Less susceptible to degradation from environmental conditions or extraction techniques: Useful in very degraded remains or where only matrilineal identification is needed (e.g., in excluding unrelated individuals.
Disadvantages in DVI	Subject to degradation from environmental conditions or extraction techniques – therefore require well preserved tissue samples and optimal environmental conditions for successful analysis. SNPs can be difficult to assess in mixed samples as is common in disasters. Population variability must be known prior to analysis for accurate profiling.	Less statistically accurate identification due to matrilineal inheritance and cellular location only within the mitochondria.

STR, short tandem report; SNP, single nucleotide polymorphism.
Summarized from Zietkiewicz, E., Witt, M., Daca, P., Zebracka-Gala, J., Goniewicz, M., Jarzab, B., Witt, M., 2012. Current genetic methodologies in the identification of disaster victims and in forensic analysis. Journal of Applied Genetics 53 (1), 41–60.

alone. They are by nature outside the norm and often chaotic, and the management of the dead in such a context is an incredibly complex undertaking. Moreover, in our increasingly connected world, where media are available 24 h a day across the globe, disasters have both a local and global dimension. They become environments with an array of actors operating within them: police, military, government, forensic services, prosecutors, family, survivors, media, and NGOs for example. In addition, disasters often result in complex identification conditions, where bodies are severely damaged, fragmented and degraded by fire, water, or other environmental conditions. Remains of one person may be mixed with those of another, and it is not unknown in manmade disasters such as bombings, for remains of the perpetrators to be mixed with those of the victims. In such conditions, traditional methods of identification such as odontology or visual recognition are often unsuccessful and other methods are therefore required to enable positive identification of human remains. The use of genetic profiling is increasingly relied upon in these difficult circumstances. Where other methods are inconclusive or completely useless, DNA profiling of one form or another can provide some indication as to positive identification of remains.

Against this backdrop, the assumed primacy of genetic modes of identification can be problematic, because of its potential to clash with sociocultural norms and practices. Moreover, its use can cause conflicts within existing systems of power, knowledge, and governance. Often the most catastrophic disasters, particularly natural ones, occur in developing nations; their poor infrastructure and lack of logistical support, expertise, or money to undertake the necessary interventions makes the situation all the more devastating, and they may become reliant on intervening powers who offer support. These intervening authorities may provide consultation on the best way to manage the postdisaster situation, or, in some cases, they may take over its management. In many cases, several organizations will offer support, causing confusion and potential conflicts that invariably compromise effective action.

In these cases, deciding upon appropriate interventions becomes all the more complex. At present no accepted international standards exist on DVI, although many organizations have issued guidelines (e.g., the International Committee of the Red Cross, Interpol, the Pan American Health Organization, the US National Institute of Justice, the Association of Chief Police Officers in the United Kingdom, and many others besides). Although these guidelines mostly conform with one another, they vary in certain details, and the presence of so many different guidelines produces its own problems. Add to that the fact that not all nations necessarily adhere to Western scientific modes of understanding, where identity is biologically expressed through genes, and where individual identification and burial is the most pressing issue in dealing with the dead. The use of DNA profiling in DVI begs questions of place and authority. In particular, the concept of authority and the intervention of external agencies that determine the 'correct' approach to postdisaster management, along with the place of the 'expert' should be questioned, because those considered expert in one world may be very far from it in another place and social landscape.

Why Do We Need to Manage the Dead?

Following a disaster, appropriate management of the dead may become impossible in the struggle to regain control and reestablish stability, particularly where external agencies become involved with their own particular ideas of 'appropriate' postdisaster management. A disaster causes massive disruption in the social fabric of everyday life, throwing communities into chaos and disorder. People often live in a state of stress and panic, both individual and collective, and their lives need to be reordered and their worlds repaired for them to feel secure again. Order must be restored as quickly as possible. A failure to take swift action in managing the disaster invariably causes more problems and can cause further turmoil. The situation in New Orleans following Hurricane Katrina is an example of what happens when immediate action to manage the situation is not taken – moral and social chaos.

For a government or other authority to remain in control it must repair the ruptured cosmos whilst managing the increasingly globalized networks of people involved in the postdisaster management; gaining trust and confidence from its population to do this is imperative to prevent further disorder and chaos. Controlling the treatment of the dead, and effecting identification efforts as early as possible is often central to the process of order reconstruction. Identification of the dead brings them back into the control of the state, through a reassertion of control and order (Wagner, 2008). DNA profiling is an increasingly popular tool in this effort. Its use, or refusal, can support or negate the legitimacy of power and cultural or social ideologies of a society and its governing body; it can consequently be used to reinforce power mechanisms and morality ideals within society (Verdery, 1999). But who decides what kind of management is appropriate and how it should be conducted? And what effect does the choice of DVI by DNA analysis have on the wider community particularly when it is not a native technology?

Critiques of the Use of Genetic Profiling in DVI: Social and Scientific Definitions of Identity

Whilst studies in epigenetics have influenced scientific debates in recent years, encouraging them to explore the complexity of both behavior and identity as multifaceted and neither solely genetically nor solely socially located (see Carey, 2012), lay discourses have yet to embrace these complexities (Prainsack et al., 2014). Western media coverage and public discourse still favors a primarily essentialist notion of identity related to the genes (Dar-Nimrod and Heine, 2011); in this popular discourse, DNA analysis has attained the status of being the ultimate provider of 'truth' and identity for human remains. This is based on the assumption that identity is best assessed biologically. However, this assessment is based on a knowledge system based in Western postenlightenment scientific discourse, which, as with any knowledge system, is itself a product of the sociocultural environment and historical period within which it was born (Finkler, 2000).

Because of these lay beliefs, in the Western world, identification by DNA profiling is viewed as the *only* appropriate means of identification of human remains following a disaster.

Identification by DNA profiling, based solely on physical remains, suggests that identity is inextricably bound to the skeleton and located at the cellular level, and negates identifications produced more socially; through events, or by specific individuals within a community, such as spirit mediums, local healers, doctors, or others who may usually be involved in some places. The conflicts between different systems of knowledge related to identification comes to the fore in DVI when local forms of social identification may be undermined due to the many agencies from differing nations that are often involved. Identities are complexly intertwined presentations of biological, cultural, and social systems, which interact in different ways. How these systems interact and take preference over each other is little understood, and the impact of using one system over another – both on identity and on events beyond the act of identification – is rarely discussed.

The use of genetic profiling in DVI accompanies an assumption that individual identification and repatriation of remains is a central component of the postdisaster management process. Interpol, for example, asserts that identification of remains is 'a basic human right' (Interpol, 2009). Indeed, many people believe it is. However, there exists no directive to this end: neither the UN, nor any other international governing body has established declarations on this issue, and the legal status of remains depends upon the state from which they come. In many societies collective treatment is as legitimate, socially and legally, as individual identification, and even where individual identity is desired, DNA profiling or other genetic means may not be the most socially appropriate technology – other, less statistically accurate, scientifically based, or internationally respected methods may be preferred. In Timor-Leste, for example, identification has been conducted for many years by spirit mediums. The identifications they produce may not be genetically accurate, but they are socially accepted and adhered to. Meanwhile in Cambodia, identity relates not to the skeleton, but to the spirit, and therefore, the actual remains of a person are not necessary for proper funeral rites to be enacted. Over 1 million anonymous remains are scattered across the country following the Khmer Rouge genocide of 1975–79, and cause no social discomfort.

Interventions and assistance are often a necessary part of any postdisaster landscape. Many disasters occur in developing nations, which often lack the infrastructure, expertise, and money, to conduct extensive identification efforts. However, all interventions have social impacts. With DVI there are many questions to be asked of management of the dead, particularly related to identification efforts and specifically genetic identification efforts.

Who has the legal authority over the bodies? Whose dead have priority? Should remains stay within the country or can they be transported elsewhere? How should the dead be stored? How much of a body constitutes a person? Who can manage the dead? What result will improper management have on the dead and the living (e.g., dead intervening in lives of living, etc)? How is identity understood? Who can establish this identity? How does the identity given reflect power and control structures? How does identification by DNA profiling affect, or interfere with, identity by other means? What effect can its prescription have on people? And where identification by DNA undermines traditional modes of identification, and the authority of those who provide it (e.g., monks, spirit mediums, and doctors) how does this affect wider social stability of a locale? What is the actual ability of the managing organization to undertake identification efforts? What result will DNA profiling have in the long term (e.g., on expectations of the people involved)? When do identification efforts end? At what risk is forensic identification enforced at the cost of undermining local social systems and systems of rule? What is the potential effect of undermining these integral structures of stability and order in a society that has already undergone so much disruption and rupture?

Sociocultural Differences in the Treatment of Human Remains and Dead Bodies

The use of genetics in DVI must also be considered in relation to appropriate treatment of the dead. Most societies have strict norms regarding treatment and management of the dead, and breaking these norms can become one of the factors contributing to social unrest. In different cultural situations, belief systems on how the dead are to be treated, and the consequences of ignoring these, vary significantly from Western norms. Following the appropriate norms is imperative in ensuring the fate of the dead; to be able to progress to the next life in Hindu and Buddhism; to ascend to heaven in Christianity; and to lay at peace across the world. In Cambodia, lay ritual specialists work alongside Buddhist monks to free the spirit from the body and allow it to progress through its lives; the physical presence of the corpse is however, not necessary for this to be successful. Religious norms in Muslim and Jewish cultures usually provide that bodies be buried within 24 h, and the body should be buried in as much entirety as possible – the taking of samples for DNA analysis may be of some discomfort in both these cases if the sampling method is destructive (removal of soft flesh, teeth, or pieces of bones, for example). In the Comoros only men can have contact with the dead; in Tibet it is only members of the kin group within which they belong. In cases where (usually Western) experts take samples for DNA analysis, this rule is broken. In parts of Micronesia, to speak someone's name after death brings misfortune to the living (Dernbach, 2005); how then does forensic, individual identification of physical human remains, where the name must be spoken, affect the community in such a case? In most cultures, strict taboos dictate appropriate food, social interactions, and clothing to be used by those managing the dead and for the dead themselves, but it is unlikely DNA identification methods can take these into account in the chaotic and stressful postdisaster environment.

Even the process of identification can be distressing and may conflict with sociocultural norms and expectations. Whilst in some cases taking a sample for DNA profiling from human remains may be straight forward and easily collectible (blood or saliva, for example), due to the calamitous nature of many disasters, taking a sample for DVI may require removal of more substantial parts of the remains, such as pieces of soft flesh, teeth extracted from the head, or pieces of bone sawn from the rest of the skeleton. In such cases taking a sample is a destructive process, involving the removal, and destroying, of (albeit a small) part of the corpse. In some cases, to get at

the bone requires first the removal of the flesh; what then happens to the removed flesh? What implications do these treatments have for people who require as much integrity of the body as possible? When forensic identification efforts of the missing from Saddam Hussein's regime first began in Iraq in 2008, the Iraqi authorities initially attempted to demand that all skeletal samples taken for DNA analysis be returned so that the identified bodies could be buried intact. Of course this was not possible (the bone is ground up in the profiling process), and caused stagnation in the identification process for some years, with the authorities refusing to allow human tissue to leave the country, and yet not having the facilities or expertise to undertake the analysis themselves.

So Why Use DNA Profiling?

The use of DNA profiling in DVI is almost always presented in terms of humanitarian relief and human rights. Individual identification of the dead following a disaster is not, however, an established international human right or legal necessity, and no conventions have yet been created which deem it such. The legal status and rights of the dead vary between jurisdictions and depending on circumstances, and where international conventions do exist on the treatment of mass dead (the Geneva conventions, for example), these relate specifically to the dead resulting from war or conflict.

However, individual, positive identification of the dead may be necessary for many reasons, and not only to provide the assumed closure to family members and societies: establishment of death may be legally required for the production of death certificates, inheritance rights, access to bank accounts, pensions, and other legal actions. Individual identification of the deceased can be necessary to establish those who are still missing. It may be required in order to identify perpetrators. It may be used to give statistical figures for reporting of scale and severity of the disaster, or to provide emotive capital to political statements, which may be used to support international action, even potential military interventions.

In the case of a disaster, DNA profiling is often the most accurate mode of providing these necessary positive identifications. But its increasing use relates not only to its accuracy over other methods, but also to other social mechanisms in play across the globe: local, national and international pressures (real or imagined) that influence action, and the individual and collective benefits that result from undertaking such actions. That which is undertaken under the auspices of being 'for the common good,' or being the 'right' mode of approach, often functions in a much more pragmatic way for those involved (be it postdisaster or in longer term humanitarian aims); it is harnessed as a political tool and as a means of gaining social and cultural capital for all those involved, from individuals conducting the analysis to governments funding it. The use of DNA analysis in the identification of bodies sits within the array of political tools open to manipulation, enabling the harnessing of their symbolic power, the accumulation of social and political capital for those involved in DNA analysis, and, through the handling of these factors, a means of exercising control over a population and acting in the global community.

This is especially true after a disaster, when the social fabric of everyday life is ruptured; not only in the time and place where the event occurred, but in the case of manmade disasters such as terrorist attacks, across the globe as a whole. The very ambiguous nature of bodies – their silence – enables manipulation by one or many voices, and an imbuing of power that cannot be expressed but which is strongly emotive and affective. Their management is therefore paramount postdisaster in order to restore and maintain order following an event that has disrupted life – they possess a form of biopower that can be manipulated by those in authority and provide social and political capital to those involved in their management. By contrast, the disregard of the dead can cause the loss of control and power.

In the increasingly globalized world we live in, our social networks extend beyond our local neighborhoods to include the international community, and the groups with which we engage are ever increasing. Governments and large international organizations participate in these networks too, and are as subject to the social forces at play as are individuals. Disasters that are often entirely localized in their physical effects become globalized in their social impact, and the accumulation of capital, both social and political, affects actions across all these networks. The use of this highly advanced technology is prestigious; both for the individuals involved in taking samples and doing the analysis, and for the agency that undertakes it, or orders it undertaken, or manage its use postdisaster. In many cases it may not really matter whether the technology is used or not, nor its actual success rate. The simple declaration of intention may be enough to both provide social capital and control a chaotic and demanding population.

It is not just the technology that is used in this way. The bodies themselves are also valuable, and can be manipulated as tools of power and capital. Relations with the dead express ideological and moral norms within a society, but they can also expose aspirations (Verdery, 1999). They also have relative worth. The speed with which bodies are collected and identification efforts begin, particularly expensive, technologically advanced techniques such as DNA analysis, indicates the relative value of the dead to the agencies involved: that value often depends on the narrative within which the disaster occurs and how this can be manipulated to gain social capital and political power, both nationally and internationally.

The reality of assistance in international disaster situations is often that governments offer it primarily in order to support the victims of their own country and recover their own dead. This can lead to increasingly difficult situations and conflicts as experts from many nations attempt to recover and identify their dead. It also leads to unequal treatment of the dead from nations who do not have the means or expertise to enact identification efforts even should they wish to. Following the 2004 tsunami in Asia, over 30 international DVI teams worked on identification efforts: teams from Sweden, the United Kingdom, the United States, Spain, France, Germany, Austria, Singapore, and many other nations, worked on recovering and identifying their own dead (Beauthier et al., 2009; Lessig et al., 2006). The number of teams with different modes of working and different priorities led to many issues: bodies went missing, management conflicts caused delays in work, and unequal treatment of the dead was manifested in full sight.

Whilst identification efforts raged for Western victims, the dead from Asia were rapidly buried in mass graves. Identification efforts other than visual means only extended to these dead when the bodies started to decompose, and visual analysis could no longer distinguish identity or ethnicity. Even within the same nation, bodies have relative worth: a comparison of the identification efforts in the United States following 9/11 and Hurricane Katrina illustrate this: identification efforts for the 'national heroes' of 9/11 began within 1 month of the event, and unlimited funds have been pledged to identification, which will continue ad infinitum; in the case of Hurricane Katrina meanwhile, it took over 4 months for the use of DNA analysis to be agreed, and even then only $13 million of funding was provided (Wagner, 2008). The value of the living is also reflected in the treatment of the dead and related identification efforts: the dead of those who are valued are managed in culturally respectful ways; those who are not valued, are not.

Conclusion

The use of genetic modes of identification has become an expected technological intervention in the identification of disaster victims. Where traditional modes of identification fall short, DNA profiling offers a statistically accurate means of providing positive identification of remains and of reassociating body parts. However, its use is not without problems, including logistical, practical, and social implications, and these should be considered in its application.

See also: Bioethics: Genetics and Genomics; Bioethics: Philosophical Aspects; Genetics and Anthropology; Genetics and Forensics; Genetics and Indigenous Communities: Ethical Issues; Genetics and Sociology; Genetics: Legal Aspects; Genomics, Ethical Issues In; Indigenous Knowledge: Science and Technology Studies; International Law and Treaties; Norms in Science; Race: Genetic Aspects.

Bibliography

Association of Chief Police Officers, 2011. Guidance on Disaster Victim Identification. National Policing Improvement Agency, London.

Beauthier, J., De Valck, E., Lefevre, P., De Winne, J., 2009. Mass disaster victim identification: the tsunami experience. The Open Forensic Science Journal 2, 54–62.

Carey, N., 2012. The Epigenetics Revolution: How Modern Biology is Rewriting Our Understanding of Genetics, Disease and Inheritance. Icon Books Limited, London.

Dar-Nimrod, I., Heine, S.J., 2011. Genetic essentialism: on the deceptive determinism of DNA. Psychological Bulletin 137 (5), 800–818.

Dernbach, K.B., 2005. Spirits of the hereafter: death, funerary possession, and the afterlife in Chuuk, Micronesia. Ethnology 44 (2), 99–123.

Finkler, K., 2000. Experiencing the New Genetics: Family and Kinship on the Medical Frontier. University of Pennsylvania Press, Philadelphia.

Gonzales, A.R., Henke, T.A., Hart, S.V., 2005. Mass Fatality Incidents: A Guide for Human Forensic Identification (NIJ Special Report). US Department of Justice, Washington, DC.

Interpol, 2009. Disaster Victim Identification Guide (New). Interpol, Lyon, France.

Lessig, R., Grundmann, C., Dahlmann, F., Rötzscher, K., Edelmann, J., Schneider, P.M., 2006. Review article: tsunami 2004 – a review of one year of continuous forensic medical work for victim identification. EXCLI Journal 5, 128–139.

Lessig, R., Rothschild, M., 2011. International standards in cases of mass Disaster Victim Identification. Forensic Science Medicine and Pathology 8 (2), 197–199.

Morse, A., 1998. Searching for the Holy Grail: the human genome project and its implications. Journal of Law and Health 13, 219–256.

Pan American Health Organization, 2009. Management of Dead Bodies after Disasters – a Field Manual for First Responders. PAHO, Washington, DC.

Petrovich-Steger, M., 2009. Anatomising conflict – accommodating human remains. In: McDonald, M., Helen, L. (Eds.), Social Bodies. Berghahn, London.

Prainsack, B., Schicktanz, S., Werner-Felmayer, G., 2014. Geneticising life: a collective Endeavour and its challenges. In: Prainsack, B., Schicktanz, S., Werner-Felmayer, G. (Eds.), Genetics as Social Practice: Transdisciplinary Views on Science and Culture. Ashgate Publishing Limited, Farnham (Chapter 1).

Prinz, M., 2007. DNA Commission of the International Society for Forensic Genetics: recommendations regarding the role of forensic genetics for disaster victim identification. Forensic Science International: Genetics 1, 3–12.

Resnick, D.B., 1999. The human genome diversity project: ethical problems and solutions. Politics and Life Sciences 18, 15–23.

Tidball-Binz, M., 2007. Managing the dead in catastrophes: guiding principles and practical recommendations for first responders. International Review of the Red Cross 89, 421–442.

Verdery, K., 1999. The Political Lives of Dead Bodies: Reburial and Postsocialist Change. Columbia University Press, New York.

Wagner, S., 2008. To Know Where He Lies: DNA Technology and the Search for Srebrenica's Missing. University of California Press, Berkeley.

Zietkiewicz, E., Witt, M., Daca, P., Zebracka-Gala, J., Goniewicz, M., Jarzab, B., Witt, M., 2012. Current genetic methodologies in the identification of disaster victims and in forensic analysis. Journal of Applied Genetics 53 (1), 41–60.

Relevant Websites

http://www.ifrc.org/en/what-we-do/disaster-management/ – International Federation of Red Cross and Red Crescent Societies (IFRC), 2013. What We Do – Disaster Management.

http://www.interpol.int/INTERPOL-expertise/Forensics/DVI – Interpol, 2013. Interpol Expertise – Forensics.

http://www.nyc.gov/ – NYC Office of the Chief Medical Examiner, 2013. World Trade Center Operational Statistics.

http://www.swgdvi.org/ – The Scientific Working Group on Disaster Victim Identification (SWGDVI), 2013.

Genetics and Economic Behavior

Björn Wallace, Faculty of Economics, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides an overview of research on genetics and economic behavior. The field is still in its infancy, but the evidence to date suggests that observed variation in a number of economic behaviors is at least in part due to genetic factors. Despite substantial progress in recent years, more research is needed for a better understanding of how genes influence economic behavior.

Economics and biology have a long history of cross-fertilization. Perhaps most notably from the viewpoint of economics, Thomas Malthus' work on population inspired Charles Darwin's evolutionary theory (Brinley, 1991). More recently, evolutionary game theory has emerged in the intersection between economics and biology (Samuelson, 2002). But, despite this shared heritage, economics has at times lagged behind other disciplines when it comes to incorporating developments in biology. One such area that until recently has been left comparatively unexplored is the relationship between genes and economic behavior.

Like most human behavior, economic decisions – such as, for instance, how much to consume, study, and work or what financial risks to take – are functions of genes as well as environment. As a consequence, they are either directly or indirectly the result of evolutionary processes (Bowles, 1998). Nevertheless, evolution and genes are often overlooked in economic analysis, and while such omissions are often motivated, care has to be taken that important insights are not lost. For instance, ignoring the well-documented influence of genetic factors on education and socioeconomic status could easily lead to erroneous conclusions.

More generally, transmission of economic behavior and outcomes can be either genetic or cultural, and while it may not always matter for analysis and policy, these are by nature different (Cavalli-Sforza and Feldman, 1981; Bowles, 1998). Genetic transmission works from one generation to another, while cultural, that is nongenetic, transmission can occur both from one generation to another, as well as between individuals of the same generation. In addition, cultural transmission can involve more, or fewer, individuals than genetic transmission. Importantly, cultural transmission of behavior has presumably in part evolved as a more flexible complement to genetic transmission which allowed humans to adapt faster and better to complex and variable environments.

While some nineteenth- and early twentieth-century economists like Alfred Marshall drew parallels between economics and biology, it was only in the wake of the nature versus nurture debate in the 1970s that economists started to take a serious look at the influence of genes on economic behavior and outcomes (Witt, 1985; Brinley, 1991; Sacerdote, 2011). This early and important work followed two strands that remain influential to date. The first tried to provide a theoretical and evolutionary basis for the existence, and at times universal fixation, of the assumptions underlying economic models. The second instead studied the influence of genes on variation in economic outcomes. This development in economics mirrored both the more general scientific debate, but also in its focus on intergenerational transmission of economic outcomes and status, the political debate in society (cf Harwood, 1982).

In recent years, the literature on genes and economics has had somewhat of a renaissance. The reasons for this are many, but basically it follows from developments in behavioral economics, evolutionary game theory, and molecular sciences. The latter of these developments has arguably been the most dramatic as it has allowed economists to for the first time directly study the relationship between genes and economic behavior.

Genetics and Economics

Today, we typically think of genes as a segment of DNA associated with a specific function. However, the concept of the gene predates the advent of modern molecular biology and the discovery of DNA. Most work on genetics and economics remain faithful to the original concept and thus define genes as the basic units of biological inheritance. Genes come in different forms and the alternative versions thereof are referred to as alleles. An individual's combination of alleles is his or her genotype and the observable traits of an individual are his or her phenotype. In what follows the general discussion on genetics builds on Neale and Cardon (1992), Falconer and Mackay (1996), and Plomin et al. (2013).

Traditionally, economics has focused on models with no or little heterogeneity, and most key assumptions about the economic agents inhabiting the models are thought to apply universally (Kirman, 1992). Thus, it is hardly surprising that in the wake of the sociobiology debate on the biological and evolutionary influences on human behavior in the mid-1970s, a number of prominent economists suggested that these assumptions could have evolutionary origins (Becker, 1976; Hirshleifer, 1977). Despite an initial flurry of contributions focusing primarily on human altruism, and a series of later articles trying to explain, for instance, preferences for risk taking and timing of consumption, the relevant literature remains small (Witt, 1985; Robson and Samuelson, 2011).

Whereas theoretical work on genes and economics often focuses on universal behaviors, empirical work is by necessity

restricted to behaviors where there is observable variation. The reason for this is simple: in order to identify an association between genes and behavior, we need variation both in phenotype and genotype. We say that a behavior is 'heritable' if variation in phenotype, here economic behaviors can be attributed to variation in genotype. By 'heritability' we thus mean the fraction of variation in phenotype that is attributable to variation in genotype. In genetics it is common to distinguish between broad and narrow sense heritability. The first includes all genetic effects, linear as well as nonlinear, on variation in phenotype, while the second includes only linear, or additive, effects. If the variation in phenotype only depends on variation in one gene, then if the relationship between alleles and phenotype is additive, then so is the relationship between variation in genotype and phenotype. If instead the variation in phenotype is caused by several genes, then we say that variation is polygenic, and additive variation does not only require a linear relationship between alleles and phenotype, but also between genes and phenotype.

Since only additive genetic effects 'breed true' in the sense that the phenotypic effect of genes in offspring is a linear combination of parental genetic contributions, selection and its associated evolutionary processes work primarily on additive genetic effects. It is therefore common to focus on additive genetic effects and often heritability is used simply to denote narrow sense heritability.

That the empirical study of the relationship between genes and economics focuses on the contributions of genes to variation in observed economic behavior or outcomes is an important and often overlooked distinction. Thus, when we say that the heritability of income is 30%, we actually mean that 30% of the observed variation in income in a specific population is attributable to variation in genotype. In other words, an economic behavior that is universal will, by this definition, not be heritable.

Behavioral Genetics

Until recently, most empirical work on genetics and economics used standard behavioral genetic methods. The origins of these can be traced back to the late nineteenth century and the work of Sir Francis Galton. Behavioral genetics studies how genetic and environmental factors contribute to variation in behavioral phenotypes. The approach involves decomposing empirically observed variation into genetic and environmental influences. This is done by studying behaviors or outcomes for individuals of different, but known, degrees of genetic relatedness. For instance, the simplest behavioral genetic studies typically involve either twins or adoptees.

The appeal of twins and adoptees is that these types of family relationships allow us to make assumptions about genetic and environmental correlations. Adopted siblings are typically genetically unrelated, but grow up in the same home, so we can assume that their degree of genetic relatedness is zero, but that the environment is virtually perfectly correlated. In contrast, biological siblings have an expected degree of genetic relatedness of 50% on average, but if they grow up in different homes, then the correlation in environment is likely to be zero. The fact that we know this, allows us to pin down a structural model and decompose observed variation into genetic and nongenetic factors by studying and comparing correlations across the different types of family relationships for the phenotypes of interest. Likewise, twins reared together grow up in the same home, and twins can either be fraternal (dizygotic), in which case their expected degree of genetic relatedness is a half, or they can be identical (monozygotic), in which case their degree of genetic relatedness is one. If we, in addition, assume that there is no assortative mating, no gene–environment interactions, no nonadditive effects, and that environment does not influence identical twins differently compared to same-sex fraternal twins simply because they are identical, then we can calculate (narrow sense) heritability as two times the difference in correlation between identical and fraternal twins.

In psychology, behavioral genetics took off in the 1960s and 1970s, when a number of seminal twin studies established that important traits such as intelligence and personality were heritable. This work was of course supplemented by research on other traits, some of which (like education) were of substantial economic significance. Consequently, by the mid-1970s, the first papers in economics using twins and behavioral genetic techniques were published (Taubman, 1976; Sacerdote, 2011). These papers focused on economic outcomes rather than behaviors and in particular on education and income. In short, the findings were that education and income like other similar traits were moderately heritable, and that shared environment was of limited importance for explaining the observed variation. These early results spurred a fairly large literature, which also came to include adoptees and other types of family relationships, and that more or less confirmed the initial findings of moderate heritability. To begin with, behavioral genetic research in economics focused primarily on the US and used potentially biased samples, but research using data from more representative Northern European samples report similar findings.

Intriguingly, although a fairly substantial literature on the heritability of economic outcomes developed quite quickly along with a very large literature on the heritability of a number of economically relevant traits in psychology, little or no research was done on economic behavior or preferences until fairly recently. The first papers to study actual economic decision-making more directly did not appear until the early 2000s and they tried to estimate the heritability of observed behavior in standard economic experiments, such as the trust, ultimatum bargaining, and dictator games, as well as risk taking (Wallace et al., 2007; Cesarini et al., 2009). Heritability in these experiments was found to be low to moderate and the influences of shared environment small. While the estimates were in line with the preexisting literature on psychological traits, they are typically on the lower end of those observed elsewhere. One potential explanation for this is the comparatively high measurement error associated with eliciting economic preferences (Benjamin et al., 2012a).

The initial experimental papers were followed by a number of papers using survey as well administrative data to study risk taking and investment behavior in the field. Together, these papers demonstrate a low to moderate heritability for a wide range of economic behaviors, preferences as well as

susceptibility to behavioral biases. The most well-studied of these behaviors or preferences is risk taking, which has been studied in the laboratory using incentivized experiments and survey questions, as well as in large sample surveys and using actual real-world investment decisions (Cesarini et al., 2009, 2010; Barnea et al., 2010).

To summarize, the heritability of economically relevant traits is approximately as follows. Intelligence measured by standard psychometric tests is highly heritable, with a heritability of approximately 50–60% (Sacerdote, 2011; Plomin et al., 2013). Education, being in part a function of ability, and personality both have high to moderate heritability of approximately 40–60% (Plomin et al., 2013; Rietveld et al., 2013). Likewise, the heritability of income is about 40% for men, with estimates for women being somewhat lower and estimates using permanent income higher (Benjamin et al., 2012b). The heritability of risk taking in different situations including both experiments and real-life portfolio formation is also approximately 30%, with a range from 20% to 60% (Barnea et al., 2010). In contrast, the heritability of economic behavior observed in other experiments is somewhat lower and are typically around 25%, ranging from 15% to 40% (Cesarini et al., 2009). Finally, the documented heritability of susceptibility to behavioral biases is typically around 20%, but there is variation and zero estimates have been reported (Simonson and Sela, 2011; Cesarini et al., 2012).

Molecular Genetics

Recent developments in molecular biology have made it possible not only to study the contribution of overall genotype to variation in observed phenotypes, but also to study the more direct relationship between actual genes and economic behavior (Beauchamp et al., 2011; Benjamin et al., 2012b; Plomin et al., 2013). Broadly speaking, the research to date has followed two routes. The first is to use available evidence from other disciplines to identify candidate genes. The alternative is to conduct genome-wide association studies using a large number of genetic markers. While the latter type of studies is becoming increasingly popular, they were, until recently, prohibitively expensive and complicated. In addition, when identifying relationships between behavior and individual genes using genome-wide association studies there is the obvious risk of finding spurious relationships.

Candidate genes are typically found by using research in neuroscience as the starting point (Plomin et al., 2013). For instance, it is well known that the dopaminergic system plays an important role in the brain's reward and feedback system. There is also suggestive evidence that genes in the dopaminergic system are associated with novelty seeking and susceptibility to addiction. Consequently, it was suggested already early on that a number of genes related to the dopaminergic system may be associated with economic risk taking and early studies found that this was indeed the case (Dreber et al., 2009; Benjamin et al., 2012b). Similarly, associations were also found between a number of other genes and economic behavior such as that between oxytocin receptors genes and trust.

However, a number of later studies have failed to confirm these initial findings (Benjamin et al., 2012b). Just like the mass significance problem associated with genome-wide association studies, there are a number of problems with candidate gene studies. First of all, there is likely to be a substantial publication bias toward research demonstrating significant relationships between genes and economic outcomes. Secondly, most published studies demonstrating significant relationships, as well as those that failed to find any such relationships, use relatively small samples, often in the hundreds. This means that most studies are statistically underpowered. In fact, it has been argued that the sample sizes typically required to convincingly demonstrate a genetic association to economic behavior under reasonable effect sizes are probably in the tens of thousands rather than in the hundreds.

In this context, it should be noted that economic behavior and outcomes are typically to be considered complex traits involving a large number of different genetic and environmental influences, and that thus far the molecular genetic evidence for complex traits is in some respects rather disappointing (Chabris et al., 2013; Plomin et al., 2013). For instance, both height and intelligence are two highly heritable phenotypes for which there have been a number of genome-wide association studies. However, despite large sample sizes researchers have to date only been able to explain a fraction of heritability in these traits with specific genes. In fact, a recent and very large genome-wide association study on educational outcomes with reasonable power failed to find any substantial relationship between specific genes and education (Rietveld et al., 2013). Education is of course like so many other economic traits a complex trait, but it is also a composite of many underlying and more fundamental complex traits, such as, for instance, intelligence and conscientiousness, complicating matters further.

While the fact that there is no convincing evidence for any large effects of specific genes on economic behavior may at a first blush seem contrary to the behavioral genetic evidence, it is not necessarily so. Evolution has presumably, at least until recently, not selected strongly for economic phenotypes. Rather, these, for instance, active consumers or prudent savers of monetary assets, arise in a modern economic environment. Thus, the economic traits that we are typically interested in are like education composites of other underlying traits that may not even necessarily be correlated. This in turn suggests that rather than being mediated by large effects from a small number of genes, the heritability of economic behaviors may instead be mediated by small effects from a large number of genes (Chabris et al., 2013). In fact, this is what recent studies of complex traits suggest. Instead of studying the effects of specific genes, or using presumed genetic relatedness to infer genetic variation, these studies use genetic variation derived from actual genotypes. But, although such studies cover large parts of the genome, they rarely, if ever, use the complete genotypes of the sampled individuals. Therefore, the heritability estimates derived from these types of studies can arguably be considered lower bounds (Benjamin et al., 2012a; Rietveld et al., 2013). For economic traits the estimates for survey-based preferences range from 0% (fairness) to 34% (economic policy), while the average estimate for education is 24%. The fact that several of these estimates are significant strongly suggests that economic behaviors and outcomes are under genetic influence.

Broadening the Discussion

Some 40 years of research on genes and economics has demonstrated that economic outcomes such as education and income are in part heritable. The more recent literature on the relationship between genes and economic behavior is still in its infancy, but the evidence so far suggests that this is mostly likely also true, albeit to a lesser extent, for a number of key economic behaviors and preferences. While most research in the field supports the conclusion that economic traits are under genetic influence there are several reasons to be careful in interpreting the literature. One important reason for this is that a substantial share of the literature relies on data from twins, and some of the assumptions underlying the behavioral genetic models using twins are controversial. In particular, the assumption that same-sex fraternal and identical twins are not influenced differently by environmental factors due to their different genetic relationships with their co-twins has attracted a lot of criticism. Furthermore, the literature struggles with separating additive from nonadditive genetic effects as well as with accounting for interactions between genes and environments more generally.

However, based on the evidence from the extensive nature versus nurture debate on intelligence it is reasonable to assume that while the criticism against twin models is in part motivated, the qualitative finding that genes influence economic behaviors and that these are heritable is likely to be unchanged by future research. This conclusion is supported by the recent genome-wide association studies that were able to link variation in economic behavior and outcomes to actual genetic variation on the molecular level. Thus, as far as we can tell economic behavior is under genetic influence. As a consequence, the key problem arguably becomes not if genes influence economic behaviors, but rather to what extent and what the implications are. Empirically, more large scale traditional behavioral genetic and genome-wide association studies are needed for a better understanding of the heritability of various economic behaviors and the interaction between genes and environment. Perhaps more importantly, empirical research needs to be complemented by theoretical research and be better related to theory. On a fundamental level, common misunderstandings hamper research on genes and behavior. The perhaps most important of these is that if a trait is heritable, then it is immutable. That this is not so can easily be seen by considering the heritable traits myopia or religion. Likewise, not all economic behaviors are as likely as others to be under genetic control or to be heritable. Here, evolutionary and anthropological theory and evidence can help to guide research both on the planning and analysis stage. In addition, models that incorporate both heterogeneous agents and modes of transmission are presumably needed to better understand the implications of heritability of economic behavior.

See also: Behavioral Genetics: Psychological Perspectives; Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Epigenetic Inheritance; Evolutionary Developmental Psychology; Evolutionary Economics; Evolutionary Social Psychology; Familial Studies: Genetic Inferences; Gene–Culture Coevolution; Genetic Studies of Behavior: Methodology; Genetics and Society; Genetics and Sociology; Genetics of Personality; Genotype and Phenotype; Human Behavioral Ecology; Human Cooperation, Evolution of; Intelligence, Genetics of: Cognitive Abilities.

Bibliography

Barnea, A., Cronqvist, H., Siegel, S., 2010. Nature or nurture: what determines investor behavior? Journal of Financial Economics 98, 583–604.

Beauchamp, J.P., et al., 2011. Molecular genetics and economics. Journal of Economic Perspectives 25, 57–82.

Becker, G.S., 1976. Altruism, egoism and genetic fitness: economics and sociobiology. Journal of Economic Literature 14, 817–826.

Benjamin, D.J., et al., 2012a. The genetic architecture of economic and political preferences. Proceedings of the National Academy of Sciences of the United States of America 109, 8026–8031.

Benjamin, D.J., et al., 2012b. The promises and pitfalls of genoeconomics. Annual Review of Economics 4, 627–662.

Bowles, S., 1998. Endogenous preferences: the cultural consequences of markets and other economic institutions. Journal of Economic Literature 36, 75–111.

Brinley, T., 1991. Alfred Marshall on economic biology. Review of Political Economy 3, 1–14.

Cavalli-Sforza, L.L., Feldman, M.W., 1981. Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press, NJ, USA.

Cesarini, D., Dawes, C.T., Johannesson, M., Lichtenstein, P., Wallace, B., 2009. Genetic variation in preferences for giving and risk taking. Quarterly Journal of Economics 124, 809–842.

Cesarini, D., Johannesson, M., Lichtenstein, P., Sandewall, Ö., Wallace, B., 2010. Genetic variation in financial decision-making. Journal of Finance 65, 1725–1754.

Cesarini, D., Johannesson, M., Magnusson, P.K.E., Wallace, B., 2012. The behavioral genetics of behavioral anomalies. Management Science 58, 21–34.

Chabris, C.F., et al., 2013. Why it is hard to find genes associated with social science traits: theoretical and empirical considerations. American Journal of Public Health 103, S152–S166.

Dreber, A., et al., 2009. The 7R polymorphism in the dopamine receptor D4 gene (DRD4) is associated with financial risk taking in men. Evolution and Human Behavior 30, 85–92.

Falconer, D.S., Mackay, T.F.C., 1996. Introduction to Quantitative Genetics, fourth ed. Pearson Education, UK.

Harwood, J., 1982. American academic opinion and social change: recent developments in the nature-nurture controversy. Oxford Review of Education 8, 41–67.

Hirshleifer, J., 1977. Economics from a biological viewpoint. Journal of Law and Economics 20, 1–52.

Kirman, A.P., 1992. Whom or what does the representative individual represent? Journal of Economic Perspectives 6, 117–136.

Neale, M.C., Cardon, L.R., 1992. Methodology for Genetic Studies of Twins and Families. Kluwer Academic, NL.

Plomin, R., DeFries, J.C., Knopik, V.S., Neiderhiser, J.M., 2013. Behavioral Genetics, sixth ed. Worth Publishers, NY, USA.

Rietveld, C.A., et al., 2013. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science 340, 1467–1471.

Robson, A.J., Samuelson, L., 2011. The evolutionary foundations of preferences. In: Benhabib, J., Bisin, A., Jackson, M.O. (Eds.), Handbook in Social Economics, vol. 1A. North-Holland, CA, USA, pp. 221–310.

Sacerdote, B., 2011. Nature and nurture effects on children's outcomes: what have we learned from studies of twins and adoptees?. In: Benhabib, J., Bisin, A., Jackson, M.O. (Eds.), Handbook in Social Economics, vol. 1A. North-Holland, CA, USA, pp. 1–30.

Samuelson, L., 2002. Evolution and game theory. Journal of Economic Perspectives 16, 47–66.

Simonson, I., Sela, A., 2011. On the heritability of consumer decision making: an exploratory approach for studying genetic effects on judgment and choice. Journal of Consumer Research 37, 951–966.

Taubman, J., 1976. The determinants of earnings: genetics, family, and other environments: a study of white male twins. American Economic Review 66, 858–870.

Wallace, B., Cesarini, D., Lichtenstein, P., Johannesson, M., 2007. Heritability of ultimatum game responder behavior. Proceedings of the National Academy of Sciences of the United States of America 104, 15631–15634.

Witt, U., 1985. Economic behavior and biological evolution: some remarks on the sociobiology debate. Zeitschrift für die gesamte Staatswissenschaft/Journal of Institutional and Theoretical Economics 141, 365–389.

Genetics and Forensics

Simon A Cole, University of California, Irvine, CA, USA
Barbara Prainsack, King's College London, Strand, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides an overview of the use of genetic science and technologies for criminal investigation and forensics. After discussing the history of the use of biological markers for criminal identification and forensics, we outline the emergence of DNA analysis as the 'gold standard' in forensic identification. We then discuss a range of more recent practices and issues in the field of forensic genetics. We conclude with a brief discussion of novel technologies that have not yet found entrance into most courtrooms and routine crime scene work.

Background: Forensic Identification

In criminal investigation, it often is useful to try to link bodily traces to suspect bodies, which may have been present at sites of interest or had contact with persons or objects of interest. The earliest uses of such traces probably involved handwriting, but the paradigmatic forensic identifier is the fingerprint, which came into regular use for forensic purposes in the early twentieth century. Unlike handwriting, which is a volitional act, fingerprints are bodily attributes, which are in some sense outside the control of the individual, and which we call *markers*. Other forensic identification marker sources included hair, dentition, and blood. The foundation of bodily forensic identification lies in the claim that traces can, with some degree of accuracy and precision, be linked to the suspect body. A 'known' fingerprint trace is compelled from the suspect body (i.e., by 'rolling' the suspect's fingerprint), and the 'known' and 'unknown' traces are compared. For most of the twentieth century, fingerprinting often was called the 'gold standard' in forensic identification and, indeed, forensic science more generally. Fingerprints were particularly useful for forensic science for a number of reasons: People generally deposited fingerprints fairly prolifically, making the recovery rate high, despite rather well-known and easily adopted countermeasures like wearing gloves. Whereas handwriting identification may have been particularly useful in investigation of crimes like fraud and other financial crimes involving documents, fingerprints were deposited by the anatomical organs of touch, which were frequently employed in the commission of crimes, especially violence and theft. Fingerprints could be compelled from suspects easily and cheaply. These known prints could be indexed according to their 'pattern types' by trained classifiers and could be stored relatively cheaply. At the same time, however, the utility of fingerprinting was limited. Not all crime scenes yielded incriminating prints. If suspects were not identified by investigative means, the examiner would have no 'known' to which to compare the crime-scene print. The database could be consulted, but this often was prohibitively difficult because most databases were indexed according to the pattern types on all 10 fingers, whereas crimes scenes rarely yielded complete sets of 10 prints. This issue placed severe limits on the utility of fingerprint databases in suspectless crimes until the flourishing of computerized searching in the 1980s.

Development of Forensic DNA Profiling

Forensic serology, the analysis of blood, was not new. It had been possible for decades to analyze blood stains for blood type and also for the presence of various enzymes. These tests, however, were not sufficiently discriminating. Serology was mostly helpful for the exclusion of suspects (e.g., if a trace that must have come from the perpetrator had a different blood type than the suspect). Fingerprinting, at this time, was claiming to be able to achieve infinite discrimination, that is, to the level of a single individual (this claim later was found to be unfounded). With criminal investigation paying increasing attention to biological information, the utility of treating genetic information as a forensic marker was clear. It was in this milieu that the utility of Alec Jeffreys' discovery of probes for visualizing variable-number tandem repeats (VNTRs) (*see* Forensic Genetic Databases: Ethical and Social Dimensions) for forensic identification gradually became apparent. Given a sample of bodily fluid, the forensic investigator could posit the same sort of link between a bodily trace and a suspect body that had been posited for fingerprints. Indeed, Jeffreys called the technique 'DNA fingerprinting' in a deliberate evocation of the older technique (Lynch et al., 2008). Because DNA profiling – the term that later became standard – targeted a different type of bodily trace than fingerprint analysis did, it could be used in investigations in which no usable fingerprints were found, but in which body fluids from the likely perpetrator were present.

Briefly, Jeffreys' method utilized the fact that human beings have different variants of DNA in certain places on their chromosomes; in healthy individuals, DNA is the same in every cell in the body. The higher the number of places (loci) in the chromosome that are being analyzed, the smaller is the likelihood that two individuals have the same DNA variants in all the same places (unless they are identical twins or multiples; for more details, *see* Forensic Genetic Databases: Ethical and Social Dimensions). Jeffreys' method of DNA analysis went far beyond serology in the sense that it allowed for the testing of several genetic loci, which offered much higher discrimination in the analysis of bodily fluids. The technique was famously first put to use in 1986 in a serial rape-murder cases in Narborough in the United Kingdom near Jeffreys' laboratory at the University of Leicester. There was a prime suspect, who even confessed to the crimes; genetic analysis, however, excluded him as a possible

perpetrator. The police then conducted what later would be called a 'DNA dragnet,' demanding blood samples from every eligible male in the area. This pressure eventually flushed out a witness who reported that a suspicious colleague was seeking to evade giving a sample. Several themes that would become important for genetics and forensics, including the exoneration of the innocent, DNA dragnets, and the power of the perception of the technique's accuracy to flush out other kinds of evidence (e.g., Machado and Prainsack, 2012), were presaged in this case.

The technique's potential for forensics was apprehended relatively quickly, and some law enforcement agencies, including the (now defunct) UK Forensic Science Service (FSS) and the US Federal Bureau of Investigation (FBI) moved quickly to develop it. In addition, a number of private companies arose to market the DNA profiling services (Daemmrich, 1998). The technology developed rapidly. Systems based on the polymerase chain reaction (PCR) were developed during the 1990s. These systems enabled 'amplification' of forensic traces – small amounts of genetic material could be copied until sufficient material to test was available, and testing also became faster. But early PCR markers lacked the discrimination of Jeffreys' original technique. The technique of counting short tandem repeats (STRs) – repetitions of chains of nucleotides in given loci – combined PCR's amplification capability with high discrimination power. STR-based techniques remain the most common DNA profiling system used in criminal justice. According to the latest survey carried out by Interpol on this topic, 120 of the 172 Interpol member countries reported to have used DNA profiling (Interpol, 2008).

Effects on Criminal Investigations

The effect of genetics on criminal investigation may usefully be divided into two categories: cases in which genetic analysis inculpates a suspect and cases in which it exculpates a suspect or convict. There is, obviously, a quite significant third category of cases in which genetic analysis neither inculpates nor exculpates. Examples of such cases might include those in which the genetic samples are degraded, contaminated, or otherwise not fully analyzable a relatively robust profile is not linked to a person of interest in solving the crime, or the DNA profile of the suspect could have ended up at the crime scene for legitimate reasons (see following section).

Inculpation

Genetic analysis is of obvious utility in providing inculpatory evidence in the investigation of crimes. Evidence of a possible common source between a crime scene sample and a sample from a known individual can provide strong evidence of that individual's presence at the crime scene. Depending on the nature of the sample, that evidence may be more or less incriminating in the actual commission of the crime. For instance, consistency with a genetic sample derived from a semen sample from a rape-murder case may be incriminating. On the other hand, consistency with a genetic sample derived from a sweat stain at a public, highly trafficked location may be far less incriminating. Incriminating links need not necessarily involve a crime scene. For example, links might be drawn between a genetic sample recovered from a suspect's property and a genetic sample known to come from the crime victim.

Although it is commonplace to speak of genetic 'matches' in forensics, in fact, an association between two genetic samples is probabilistic in nature. Genetic forensic scientists assess whether the two samples appear to have the same alleles at the designated loci. This assessment sometimes can require interpretation, as when there are ambiguities as to whether a given result is genuine or is the result of noise, contamination, mixture with DNA from another source, and so on. Even if the findings of consistency are assumed to have been interpreted correctly, however, a further question concerns whether other individuals in the population might have the same array of alleles at the designated loci. A monozygotic (identical) twin is an obvious possibility. Even setting twins or multiples aside, however, genetically distinct individuals might by chance have the same array of alleles at the loci examined. It is, therefore, necessary to try to estimate the rarity of the array of alleles. A common such estimate used in forensics is a 'random match probability' (RMP). Genetic information has an unusual data structure not shared by most other forensic disciplines – the alleles are numbers of repeats and thus exist in nature as discrete whole-number values, rather than continuous measures – which makes the calculation of such estimates relatively straightforward in simple (e.g., nonmixture) cases.

It is necessary to take this estimate into account when considering the inculpatory value of a finding of genetic consistency between two samples. Many forensic scholars believe strongly that the proper framework for such considerations is what generally is known as a 'logical' or 'Bayesian' approach based on a likelihood ratio, but the penetration of this approach into routine practice and courtroom proceedings has been mixed. The approach has been embraced in some countries such as the Netherlands and Sweden, and it was strongly promoted by the FSS in the United Kingdom. In other countries, like the United States, however, the Bayesian approach remains rare. Instead, most forensic genetic expert witnesses testify about RMPs. In some cases, when RMPs are low, experts testify, controversially, to source attributions – that is, a single individual *is* the source of the forensic trace.

One paradox of advancing technology is that, as genetic tests have become increasingly sensitive, their incriminating value is in some cases beginning to erode. In the earliest days of genetics and forensics, blood or semen were necessary to contain sufficient genetic material to analyze. This left a lot of crime scenes impervious to genetic forensics, but if an association between a suspect and a blood or semen sample was established, that association often was quite difficult for the suspect to explain away without reference to increasingly fanciful scenarios involving other attackers or other less probable events. Contemporary technology, however, is capable of recovering analyzable samples from quite small traces of genetic material derived from sweat, hair, skin cells, body oils, and so on – so-called 'touch DNA.' This makes genetic analysis deployable for a much broader variety of crimes from auto thefts to financial crimes. At the same time, it increases greatly the number of samples that will be recovered from a particular crime of object, but it also increases the probability that the

sample came to be at the scene 'innocently' through a prior visit or transfer, thus increasing the range and plausibility of defense that might explain a finding of association. Planting scenarios are also increasingly plausible.

Exculpation

Many exculpations, presumably, occur at the investigative state. How many is not well understood, although during the early years of DNA profiling, the FBI reported that around one-quarter of prime suspects were exculpated by DNA analysis (Connors et al., 1996). The more spectacular exculpations are those that have occurred postconviction. Because biological evidence is sometimes preserved, the development of genetic forensics in the 1980s made it possible to retest evidence postconviction. This resulted in a significant number of spectacular exonerations of individuals who often had served long prison sentences, under sentence of death in some cases, for crimes they apparently did not commit. The primary locus of such exonerations has been the United States, possibly because of better preservation of evidence, where there have been more than 300 'post-conviction DNA exonerations' since 1989 (Garrett, 2011). Although there always have been other means of exonerating the convicted – and, indeed, during the period since 1989, there have been more non-DNA exonerations than DNA exonerations – DNA exonerations have an ability to achieve 'epistemological closure' (Aronson and Cole, 2009) that other exonerations usually lack. This means that 'innocence skeptics' are less disposed to challenge the factual basis of exonerations when they are vouched for by genetic evidence. In most cases, such challenges rely on changing the government's theory of the crime from single- to multiple-perpetrator rape. In recent years, initiatives of lawyers working *pro bono* to exonerate the wrongfully convicted have become more widespread in countries outside of the United States, including Canada, France, New Zealand, and Australia.

In the United States, the wave of DNA exonerations has been extremely influential in stimulating the movement for criminal justice system reform and also for supporting opposition to capital punishment. Seemingly paradoxically, a small number of the US DNA exonerations have involved individuals who also were convicted on the basis of genetic evidence, evidence that was either misinterpreted or presented in a misleading manner (Garrett, 2011). It has long been predicted that DNA exonerations eventually would wane, as DNA testing at the investigation stage became routine, but thus far exonerations continue.

Use as Evidence in Court

The 'DNA Wars'

During the earliest years in which DNA evidence was introduced, its admissibility was extensively litigated in the United States, in a series of contests sometimes labeled as the 'DNA wars' (Aronson, 2007; Kaye, 2010) and in the United Kingdom (Lynch et al., 2008). In the earliest US cases (in the late 1980s and very early 1990s), DNA evidence was either not challenged or not challenged expertly, and it was routinely admitted. In later cases, however, defense attorneys enlisted well-credentialed molecular biologists who were able to gradually expose sloppy practices, failure to adhere to protocols, and unprincipled (and biased) interpretations of data. These interventions and criticisms produced a few cases in which state courts excluded DNA results, perhaps most famously in *People v. Castro*. Not insignificantly, these exclusions quickly were followed by federal courts deeming similar evidence admissible. The courts also exercised caution in early cases in the United Kingdom and Australia (Edmond et al., 2013). In the United Kingdom, proposals to employ Bayesian statistical analysis to interpret DNA evidence were rejected by the courts (Lynch et al., 2008).

Successful challenges to the admissibility of DNA evidence in the United States drew on population genetics to challenge calculations of RMPs. Drawing on debates among geneticists about human mating patterns, defendants argued that the state's RMP calculations were not accepted in the scientific community. These cases helped to trigger the intervention of the US National Academy of Sciences – through its National Research Council (NRC) committees. The NRC issued two reports, in 1992 and 1996, each of which endorsed two different ways of estimating the RMP, the 'ceiling principle' and the 'product rule,' respectively. Following the first report, some courts excluded RMPs proffered by the government. The second report, however, practically eliminated admissibility challenges based on population genetics; to the extent that the government asserted that it was adhering to the NRC recommendation, admissibility challenges were unlikely to succeed, and subsequent cases upheld admissibility. Since the mid-1990s, DNA evidence in general has been universally admissible in the United States and many other countries (Edmond et al., 2013; Hindmarsh and Prainsack, 2010). Another consequence of these reports was that the issue of accounting for human mating patterns was resolved by allowing for the calculation of RMPs based on gross categories labeled 'race' (Kahn, 2008): for example, in the United States the categories were 'White,' 'Black,' and 'Hispanic'; in the United Kingdom, they were 'White,' 'Black,' and 'Asian.'

Perhaps the most famous motion to exclude DNA evidence was one that was never filed. Although the defendant's 'dream team' prepared an extensive motion to exclude DNA evidence in *People v. O.J. Simpson*, sometimes called 'the trial of the century,' they withdrew the motion early in 1995. Instead, the defense team famously – and generally, it would appear, successfully – opted to attack the weight of the evidence at trial by alleging sloppy procedures, inadvertent contamination, and possible planting of evidence (Aronson, 2007: p. 173; Kaye, 2010: p. 152).

Targeted admissibility challenges still are made. One avenue of challenge concerns how a 'cold' database search affects the calculation of the RMP. Because statisticians disagreed about how the fact that a DNA association was generated through a database search should be handled (although all agreed that it mattered), some defendants argued, unsuccessfully, that the government's method of calculating the RMP was not 'generally accepted.' There has been some litigation on statistical grounds about the conclusions that DNA analysts should be permitted to state in their testimony and that prosecutors

should be permitted to state in their summations. There has thus far been only a small amount of litigation over DNA profiling techniques more exotic than the STR testing that has become standard. The admissibility of Y-STR haplotyping (the analysis of DNA in the male sex chromosome, which can provide information on paternal lineage) and mitochondrial DNA profiling (i.e., the analysis of DNA that is not contained in the nucleus of the cell; mitochondrial DNA is passed on from mothers to children) both have been upheld in the United States (Kaye, 2010). In England, Wales, and Australia, there has been litigation concerning low-copy-number (LCN) DNA profiling (Gilder et al., 2009). LCN DNA profiling is a particularly sensitive method of analysis for which only tiny amounts of DNA – a few cells – are needed. Disagreement remains about the reliability of LCN DNA markers.

As the technology has improved and RMPs – that is, the chances that a DNA profile obtained from a crime scene matches the profile of a randomly drawn profile from the general population – have gotten smaller, the arguments for the importance of race and racial mating patterns have diminished. It originally was argued that race-specific calculations of RMP were important to generate more accurate RMPs, but some social scientists now argue that the purported accuracy gains derived from using racial categories 'fade into irrelevance' and are outweighted by the harms of reifying crude notions of genetic races (Kahn, 2008).

Effects on Forensic Technologies

Although the legal admissibility battles generally have resulted in admissibility for genetic forensics, they had unexpected and unintended consequences for other areas of forensic science. The sophisticated disputes involving high-powered scientists over population genetics and the statistical reporting of results provoked the question of how these issues were addressed in other areas of forensic identification, such as fingerprinting, firearms and toolmarks, bitemarks, handwriting, microscopic hair comparison, and so on. It turned out that all these fields had been indifferent to statistics, viewing it as unnecessary to seek to statistically characterize the size of the potential donor pool of forensic traces. Instead, these fields had relied on semantic workarounds that evaded these issues, and courts, historically, had permitted them to do so. The admission of genetic forensics into court thus had the effect of prompting defendants and scholars to demand that these older fields of forensic science account for themselves probabilistically, resulting in admissibility challenges to venerable techniques whose admissibility had seemed long settled. These developments have lent momentum to a movement toward probabilistic thinking in forensic science, which has held that all evidence can, and should, be understood probabilistically. It also generated a phenomenon that has been called an 'inversion of credibility' in which genetic forensics, which in the early 1990s routinely described itself as similar to, but weaker than, venerable techniques like fingerprinting, came to be widely seen as stronger. The period in which proponents of genetic forensics sought to bolster their credibility by invoking analogies with fingerprinting was followed by a period in which genetics was widely seen as having succeeded fingerprinting as the 'gold standard' in forensics (Lynch et al., 2008).

The 'CSI Effect'

Another potential unintended consequence of the development of DNA profiling is the so-called CSI effect, a term used somewhat inconsistently to mean a variety of things. Increasing awareness of the existence of forensic genetics may reasonably be expected to have a legitimate effect on the behavior of various criminal justice system actors, including attorneys, judges, jurors, police, and offenders. This effect has been dubbed the 'tech effect' to distinguish it from the 'CSI effect,' which generally is used to denote an exaggerated, and often false, belief in the power and capabilities of forensic science, including forensic genetics, fostered by popular fictional media, especially television programming. The most common claims have been advanced in regard to US juries, which, it is claimed, are acquitting in cases in which they would have convicted but for television programming. Evidence in support of such strong claims of a 'CSI effect' is thus far weak, but claims of changes in attorney, offender, or even individual juror behavior are better supported (Shelton et al., 2006; Cole and Dioso-Villa, 2007; Machado and Prainsack, 2012; Cole, in press).

Databases

In the early years, the results of DNA analysis from crime scene and suspect samples were compared in an ad hoc manner – that is, in the context of specific investigations and trials – and stored in a decentralized manner, if they were stored at all. In the mid-late 1990s, as an effect of both the increasing importance of STR-based DNA analysis in forensic investigation, and in court, centralized databases for police and forensic use were set up in many countries. The ways in which these databases were set up differ greatly in terms of when they were started, what the criteria for inclusion and expungement of profiles are, and in whose custody the database is held (*see* Forensic Genetic Databases: Ethical and Social Dimensions; see also Hindmarsh and Prainsack, 2010; Krimsky and Simoncelli, 2011). Not all countries had or have dedicated laws governing the establishment, maintenance, and use of DNA databases. What all DNA databases for police and forensic uses have in common, however, is that they include DNA profiles from crime scenes and from convicted offenders; some databases store profiles of suspects (e.g., England and Wales) or volunteers (e.g., Portugal) as well.

All databases allow so-called speculative searching, which means that profiles obtained from crime scene stains, or from subjects, can be compared against all other profiles in the database, without there being a particular suspect, or a hypothesis regarding how specific profiles could be linked. By making speculative searching possible, centralized DNA databases have given rise to many new 'cold case' investigations, that is, the reopening of investigations of cases that had ran out of leads (Innes and Clarke, 2009). At the same time, speculative searching has led to concerns about false matches within the database. This is particularly relevant in the context of very large databases, or networks of databases, such as the Prüm network: The Prüm network is based on EU law that obliges member states to render their DNA databases searchable to each other (at a hit–no hit basis; supposed profile matches then are followed up bilaterally by the countries between which the databases matches were

obtained; see Prainsack and Toom, 2013). In 2011, databases within the Prüm network included more than 9 million profiles (ENFSI, 2011); comparing all these profiles with each other would lead to several trillions of DNA comparisons. This means that, depending on how many loci are searched, even with very low RMPs (e.g., several billions), numerous false matches could result from every single search within the database. It is also for this reason that the European Union has endorsed the inclusion of five new markers in the European Standard Set of Loci (ESSOL; a voluntary agreement on a minimum set of genetic loci that DNA analysis in all countries include, irrespective of what chip and software they use for analysis; see Gill et al., 2006; European Union (EU) Council, 2009). These five markers will be added to the Combined DNA Index System (CODIS) used in the United States and elsewhere to ensure compatibility of DNA profiles across countries also outside of Europe, and within the DNA Gateway run by Interpol (*see* Forensic Genetic Databases: Ethical and Social Dimensions).

Function Creep

Over the course of the development of genetic forensics, the functions and uses of some DNA databases have expanded beyond those originally intended. This phenomenon has been referred to as function creep (Dahl and Sætnan, 2009). For example, the fact that biologically related people share more genetic markers in common than unrelated people could be used to search for biological relatives of a person in connection with a crime in the DNA database. Another example is the increasing use of genetic information to generate 'intelligence' about hitherto unknown crimes or linkages between crimes (Ribaux et al., 2010). Still another example is the linkage of DNA analysis results carried out by handheld devices with DNA databases. This may emerge as a result of the development of ultrarapid (i.e., a few hours) DNA analysis that does not need to be carried out in a laboratory but can be performed by crime scene officers, and analyses could be compared automatically against profiles held in the database (currently, this is not done in any country). Such a scenario would raise questions about accreditation, and about data leakage and misuse, and about the exacerbation of biases against certain population groups if handheld DNA analysis devices also would be used in police patrols. A third example is the storage of so-called single-nucleotide polymorphisms (SNP)-data or Y-chromosome STR data, alongside conventional STR-profiles in centralized DNA databases. SNPs are variations at the level of single nucleotides (G, A, T, C); forensic SNP-based tests scan places across the entire genome for such variants and highlight those that have been found to correlate with particular physical traits (e.g., eye color) or ethnicity in scientific studies (e.g., Kayser and Schneider, 2009). The question about storage of SNP-data thus emerges in connection with phenotypic profiling. Y-chromosome analysis, which can give clues about paternal lineage and – in societies where surnames are patrilineal – allow probabilistic inferences about a person's family name, has been used in criminal investigation for decades; however, the results of Y-chromosome analysis are not stored routinely in centralized DNA databases and thus are not regularly available for speculative searches. There is a Y-chromosome haplotype reference database that can aid case workers by providing Y-STR haplotype frequency estimates, but this database does not contain profiles of particular convicted individuals.

Access for Researchers

There is great variety between jurisdictions as to whether or not they allow researchers to access DNA samples or profiles stored in the database. Any research carried out with these materials or data normally will have gone through institutional research ethics reviews, and personal identifiers (names, addresses, etc.) will not be disclosed to researchers, although recent studies have shown that anonymization is no guarantee that reidentification will be impossible (Gymrek et al., 2013). The STR-based DNA profiles currently stored in centralized DNA databases are not useful for sensitive research into, for example, correlations between genetic markers and antisocial or criminal behavior, intelligence, disease propensity, character traits, sexuality, and ethnicity. This is because STR profiles – which are stored as strings of numbers in the database – represent random genetic variants of which we have millions, most with yet unexplored, or nonexisting, function. DNA *samples* – the physical material, such as blood or saliva stains, from which DNA profiles are derived – are stored in the laboratories carrying out the analysis in some countries, while other countries require that labs destroy samples once the DNA profile has been obtained. Retained DNA samples potentially could be useful for such future research, as would the analysis of existing SNP data (see next section on phenotypic profiling). In addition, some scholars have called for database custodians to allow increased access to (purportedly) anonymized data to better facilitate research on issues related to forensics, such as, for example, the statistical independence of the alleles used for profiling (Krane et al., 2009). Paradoxically for some, database custodians have emerged as defenders of the personal privacy of convicted offenders in these debates. The use of DNA samples and data collected and generated during criminal investigation for research raises a range of political and ethical concerns, especially in cases in which no transparent and publicly accessible information about the purposes and terms of such research exists.

Surreptitious Sampling

Concerns have been raised also about the practice of so-called surreptitious sampling. In the context of forensic genetics, the term 'surreptitious sampling' refers to analyses carried out on DNA that has not been obtained with the knowledge of the originator (the 'owner' of the DNA). Surreptitious sampling could use DNA left behind by suspects on coffee mugs, chewing gum, cigarette ends, or clothes, for example (Joh, 2006). At present, DNA profiles obtained by means of surreptitious sampling cannot be uploaded to the DNA database in most countries; they can be used only for comparisons in concrete cases, if at all. The issues around surreptitious sampling – or better, 'convenience' sampling in this context – would become pressing if portable DNA testing devices were deployed locally (e.g., at police stations or on police patrols), and if the people operating the device could compare the profile generated by the portable device with the profiles stored in the national database. This could lead to a situation in which individuals from whom it is not legal to demand a DNA sample could be submitted to

testing 'through the backdoor.' Moreover, as members of ethnic minorities are more likely to be affected by routine police controls in many countries ('racial profiling'), this practice also would exacerbate existing ethnic and racial biases.

Phenotypic Profiling

Another, more recent, practice that is sometimes discussed in connection with racial profiling so-called phenotypic profiling. Conventional, STR-based forensic DNA profiling looks only at genetic markers that do not allow for any meaningful inferences regarding a person's externally visible traits, such as her hair or eye color. If different markers on the genome are analyzed, however – namely, those that have been found to correlate with externally visible traits (Keating et al., 2012) – then inferences can be made regarding the person's likely characteristics. These inferences always will be probabilistic, because the predictive power of the genetic markers regarding expressed phenotypic traits is limited. In principle, this method can be highly useful in the context of some criminal investigations – for example, if there are no known suspects, phenotypic profiling could give an indication of what the perpetrator could look like and also what ethnic background she or he is likely to have. The latter is the case because some of the same markers that correlate with externally visible traits are typical for specific population groups. This example illustrates both the potential benefits as well as the potential risks inherent in phenotypic profiling (M'charek, 2008). Because the information obtained by this method is, as mentioned, not deterministic but probabilistic, its utility is limited, not only as evidence at court, but also as intelligence in a criminal investigation. For example, if in the context of a dragnet, the information that the perpetrator is 67% likely to have blue eyes, were used to target only people with blue eyes, the real perpetrator could be missed. (As dragnets are highly intrusive for those asked to volunteer a DNA sample, there is wide agreement among experts that they should be used sparingly, and only in a range of narrowly defined contexts; see Zadok et al., 2010.) DNA information used for phenotypic profiling purposes currently is not stored routinely in centralized DNA databases. Phenotypic profiling could be useful particularly in narrowing down the range of missing persons to which unidentified bodies could be matched (also in the context of disaster victim identification).

Familial Searching

'Familial searching' – which arguably would be more accurately called 'genetic proximity testing,' as not all genetically related people feel part of a family, and vice versa – can be performed if a DNA profile derived from a crime scene stain results in a partial match with a profile in the database. For example, two profiles might not match in all of the analyzed genetic loci but in most of them (e.g., in 11 of the 13 loci used in the US Federal CODIS database; for this reason, some authors call it 'low stringency,' 'near-miss matching,' or 'kinship matching'; e.g., Kaye, in press). The 'hit' in the database could have identified a genetic relative of the person who actually left the DNA at the crime scene. Police could then investigate whether the person identified has any biological relatives who could have committed the crime. In the United Kingdom and in the United States, for example, several cases have been solved by this method. In other countries, genetic proximity testing is explicitly or implicitly forbidden. Critics of genetic proximity testing argue that this method violates the privacy of the person whose profile has been identified on the database through the matching attempt, as well as the genetic privacy of the (genetic) relatives of the person whose DNA profile has been identified, and that it could reinforce the stereotype of 'crime running in families.' Moreover, some critics are concerned that a criminal investigation using genetic proximity testing could reveal a genetic link (or lack thereof) between members of a family that they currently were unaware of (Haimes, 2006, see also Greely et al., 2006). Finally, genetic proximity testing is seen to repeat existing demographic and ethnic biases within criminal justice systems. For example, in the forensic DNA database for England and Wales, black individuals are clearly overrepresented: According to the 2011 census in England and Wales, about 3.5% of those who filled in the questions about ethnicity self-identified as 'black' (in addition, just over 1% were of 'mixed' ethnicity, including 'black'; Office for National Statistics, 2011). The proportion of DNA profiles from 'black' individuals stored in the National DNA Database (NDNAD) is currently at more than 8%, that is, almost exactly double the proportion of black individuals in the general population (National Policing Improvement Agency, 2012). Although no data are available regarding socioeconomic background in this respect, a similar overrepresentation of certain groups is a possible, if not likely, scenario. This, in turn, means that individuals genetically related to people whose DNA profiles are stored in the database have a higher risk of being the subject of an investigation than genetic relatives of individuals whose profiles are not in the database, which raises questions about equity and fairness.

Future Prospects: Gene Expression and Microbiome Analysis

A limitation of the use of any kind of DNA analysis for forensic identification purposes is that it is impossible to differentiate monozygotic twins or multiples with this method. At least two strands of current research could provide solutions for this: The first is the comparison of gene expression profiles that do not look at the sequence of nucleotides but rather at what genes are expressed. Although monozygotic twins and multiples – who have developed out of one single egg and sperm – carry the same nucleotide sequences in their genes, this does not mean that all of these genes are *expressed* in the same manner. Lifestyle, environmental exposures, and other factors have the potential to switch certain genes 'on' or 'off' epigenetically without altering their structure. Epigenetic modifications – such as DNA methylation, that is, the addition of a methyl group to a DNA nucleotide – affect how genes are transcribed and expressed (see e.g., Bell and Spector, 2011; Li et al., in press). Another way of distinguishing people who carry identical DNA sequences is looking at the microbes they carry in their bodies, for example, in their throat. 'Throat-swabs' disclosing the composition of microbes could complement the traditional buccal swabs, in which cells are taken from the inside of the mouth to isolate DNA. It is unlikely that microbiome profiles will ever replace DNA profiles, however, because

they have the disadvantage that they are not stable – the composition of microbes changes over people's lifetime. For concrete investigations, however, micobiome profiles could distinguish people with identical DNA profiles in the future.

See also: Bioethics in the Post-genomic Era; Bioethics: Genetics and Genomics; Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute; Expert Testimony; Genetic Admixture; Genetic Ancestry Testing; Genetics and Disaster Victim Identification; Genetics and Social Justice; Genetics and Society; Genetics and Sociology; Genetics and the Media; Genetics: Legal Aspects; Genetics: The New Genetics; Genomics, Ethical Issues In; Human Genome Project: History and Assessment; Law and Society: Development of the Field; New Genetics and Race; Privacy: Theoretical and Legal Issues; Race: Genetic Aspects; Science and Law; Science and Technology Studies: Experts and Expertise; Scientific Controversies; Scientific Evidence: Legal Aspects; Truth and Credibility in Science.

Bibliography

Aronson, J.D., 2007. Genetic Witness: Science, Law, and Controversy in the Making of DNA Profiling. Rutgers University Press, New Brunswick, NJ.

Aronson, J.D., Cole, S.A., 2009. Science and the death penalty: DNA, innocence, and the debate over capital punishment in the United States. Law & Social Inquiry 34 (3), 603–633.

Bell, J.T., Spector, T.D., 2011. A twin approach to unraveling epigenetics. Trends in Genetics 27, 116–125.

Cole, S.A., Dioso-Villa, R., 2007. CSI and its effects: media, juries, and the burden of proof. New England Law Review 41, 435–469.

Cole, S.A., in press. A Surfeit of science: the 'CSI effect' and the media appropriation of the public understanding of science. Public Understanding of Science. http://dx.doi.org/10.1177/0963662513481294.

Connors, E., Lundregan, T., Miller, N., et al., 1996. Convicted by Juries, Exonerated by Science: Case Studies in the Use of DNA Evidence to Establish Innocence after Trial. National Institute of Justice, Washington.

Daemmrich, A., 1998. The evidence does not speak for itself: expert witnesses and the organization of DNA-typing companies. Social Studies of Science 28 (5–6), 741–772.

Dahl, J.Y., Sætnan, A.R., 2009. 'It all happened so slowly' – on controlling function creep in forensic DNA databases. International Journal of Law, Crime and Justice 37, 83–103.

Edmond, G., Cole, S.A., Cunliffe, E., Roberts, A., 2013. Admissibility Compared: The Reception of Incriminating Expert Opinion (i.e. Forensic Science) Evidence in Four Adversarial Jurisdictions. University of Denver Criminal Law Review 3, 31–109.

European Union (EU) Council, 2009. Council Resolution of 30 November 2009 on the Exchange of DNA Analysis Results (2009/C 296/01). http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2009:296:FULL:EN:PDF (accessed 25.02.13.).

European National Forensic Science Institutes (ENFSI), 2011. Survey on DNA Databases in Europe June 2011. http://www.enfsi.eu/page.php?uid=98 (accessed 11.05.12.).

Garrett, B.L., 2011. Convicting the Innocent: Where Criminal Prosecutions Go Wrong. Harvard University Press, Cambridge.

Gilder, J.R., Koppl, R., Kornfield, I.L., et al., 2009. Comments on the review of low copy number testing. International Journal of Legal Medicine 123 (6), 535–536.

Gill, P., Fereday, L., Morling, N., Schneider, P.M., 2006. The evolution of DNA databases – recommendations for new European STR loci. Forensic Science International 156 (2–3), 242–244.

Greely, H.T., Riordan, D.P., Garrison, N.A., Mountain, J.L., 2006. Family ties: the use of DNA offender databases to catch offenders' kin. Journal of Law, Medicine & Ethics 34 (2), 248–262.

Gymrek, M., McGuire, A., Golan, D., Halperin, E., Erlich, Y., 2013. Identifying personal genomes by surname inference. Science 338 (6117), 321–324.

Haimes, E., 2006. Social and ethical issues in the use of familial searching in forensic investigations: insights from family and kinship studies. Journal of Law, Medicine and Ethics 34 (2), 63–276.

Hindmarsh, R., Prainsack, B., 2010. Genetics Suspects: Global Governance of Forensic DNA Profiling and Databasing. Cambridge University Press, Cambridge.

Innes, M., Clarke, A., 2009. Policing the past: cold case studies, forensic evidence and retroactive social control. British Journal of Sociology 60 (3), 543–563.

Interpol, 2008. Interpol Global DNA Profiling Survey: Results and Analysis. Available at: http://www.dnaresource.com/documents/2008INTERPOLGLOBALDNASURVEYREPORTV2.pdf (accessed 14.05.13.).

Joh, E.E., 2006. Reclaiming 'abandoned' DNA: the Fourth Amendment and genetic privacy. Northwestern Law Review 100, 857–884.

Kahn, J., 2008. Race, genes, and justice: a call to reform the presentation of forensic DNA evidence in criminal trials. Brooklyn Law Review 74 (1), 325–375.

Kaye, D.H., 2010. The Double Helix and the Law of Evidence. Harvard University Press, Cambridge.

Kaye, D.H., in press. The genealogy detectives: a constitutional analysis of "familial searching". American Criminal Law Review 50 (1). Available at: http://ssrn.com/abstract=2043091.

Kayser, M., Schneider, P.M., 2009. DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations. Forensic Science International: Genetics 3, 154–161.

Keating, B., Bansal, A.T., Walsh, S., Millman, J., Newman, J., Kidd, K., et al., 2013. First all-in-one diagnostic tool for DNA intelligence: Genome-wide inference of biogeographic ancestry, appearance, relatedness, and sex with the identitas v1 forensic chip. International Journal of Legal Medicine 127 (3), 559–572.

Krane, D.E., Bahn, V., Balding, D.J., et al., 2009. Time for DNA disclosure. Science 326 (5960), 1631–1632.

Krimsky, S., Simoncelli, T., 2011. Genetic Justice: DNA Data Banks, Criminal Investigations, and Civil Liberties. Columbia University Press, New York.

Li, C., Zhao, S., Zhang, N., Zhang, S., Hou, Y., in press. Differences of DNA methylation profiles between monozygotic twins' blood samples. Molecular Biology Reports.

Lynch, M., Cole, S.A., McNally, R., Jordan, K., 2008. Truth Machine: The Contentious History of DNA Fingerprinting. University of Chicago Press, Chicago.

Machado, H., Prainsack, B., 2012. Tracing Technologies: Prisoners' Views in the Era of CSI. Ashgate, Farnham.

M'charek, A., 2008. Silent witness, articulate collective: DNA evidence and the inference of visible traits. Bioethics 22 (9), 519–528.

Prainsack, B., Toom, V., 2013. Performing the Union: the Prüm decision and the European dream. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 44 (1), 71–79.

Ribaux, O., Baylon, A., Roux, C., et al., 2010. Intelligence-led crime scene processing. Part I: forensic intelligence. Forensic Science International 195, 10–16.

Shelton, D.E., Kim, Y.S., Barak, G., 2006. A study of juror expectations and demands concerning scientific evidence: does the 'CSI effect' exist. Vanderbilt Journal of Entertainment and Technology Law 9, 331–368.

(UK) National Policing Improvement Agency, 2012. NDNAD by Ethnic Appearance. http://www.npia.police.uk/en/13852.htm (accessed 20.12.12.).

(UK) Office for National Statistics, London. Available at: http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-286262 (accessed 20.12.12.).

Zadok, E., Ben-Or, G., Fisman, G., 2010. Forensic utilization of voluntarily collected DNA samples: law enforcement versus human rights. In: Hindmarsh, R., Prainsack, B. (Eds.), Genetic Suspects: Global Governance of Forensic DNA Profiling and Databasing. Cambridge University Press, Cambridge, pp. 40–62.

Relevant Websites

http://www.councilforresponsiblegenetics.org/Projects/CurrentProject.aspx?projectId=10 – Council for Responsible Genetics Forensic DNA Project.

http://www.dnaresource.com/ – Gordon Thomas Honeywell Governmental Affairs.

http://www.interpol.int/content/download/8994/66950/version/2/file/GlobalDNASurvey.pdf – Interpol Global DNA Profiling Survey.

http://www.nij.gov/nij/topics/forensics/evidence/dna/welcome.htm – National Institute of Justice Forensic DNA.

http://www.npia.police.uk/en/13852.htm – (UK) National Policing Improvement Agency (NPIA), 2011. NDNAD by Ethnic Appearance.

http://www.npia.police.uk/en/13852.htm – (UK) National Policing Improvement Agency, 2012. NDNAD by Ethnic Appearance.

http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition$^{1}/_{4}$ tcm%3A77-286262 – (UK) Office for National Statistics, London.

http://www.yhrd.org/ – The Y Chromosome Haplotype Reference Database (YHRD). 0-08-097086-8.82050-4http://dx.doi.org/10.1016/B978-0-08-097086-8.82050-4doi:10.1016/B978-0-08-097086-8.82050-46.410.1016/B978-0-08-097086-8.82050-4noindexElsevierTrue

Genetics and Indigenous Communities: Ethical Issues

Emma E Kowal, University of Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Indigenous people stand to benefit from advances in genomic technology, but genetic research in Indigenous communities has been controversial. This article reviews the ethical issues that Indigenous people and others have raised with reference to genetic research projects and biobanks. The ethical issues that apply to Indigenous people should be seen as additional to 'conventional' ethical issues that apply to all people, rather than replacing them. The additional ethical concerns discussed include group harm; cultural beliefs relating to biospecimens and human origins; community engagement and collective consent; benefit; ownership; and whether biospecimens can and should be 'repatriated.'

Human genetic research is a rapidly developing field that addresses a range of scientific and health problems. Our understanding of past human population movements and microevolutionary processes has greatly improved in recent years with advances in genomics and bioinformatics. In the field of health, the future of genomics promises a 'personalized' approach to medicine where many health care decisions are based on one's individual genome, a future foreshadowed by the direct-to-consumer genetic testing industry (*see* Direct-to-Consumer Personal Genetic Testing).

Indigenous people have been involved in many genetic ancestry- and health-related projects. (The Australian convention to capitalize 'Indigenous' is followed in this article.) Their participation has been sought out on the basis that their genetic material is highly informative for studies of population history and disease, but also because they may benefit, particularly from genetic studies of rare heritable diseases that affect their communities. However, Indigenous involvement in genetic research has often resulted in acrimony. These experiences reinforce the need for a robust and broad approach to ethics in this specialized research context.

It has been proposed that genetic and genomic research could help address health inequalities (Ramos and Rotimi, 2009). This includes Indigenous health inequalities, with reduced life expectancy at birth relative to the general population in the order of 5–10 years (Australian Institute of Health and Welfare & Australian Bureau of Statistics, 2008; Cooke et al., 2007). At the same time, Indigenous peoples have been identified as particularly at risk of 'group harms' from genetic research (Foster and Sharp, 2000; McGregor, 2010; Scott et al., 2005). Indigenous peoples have expressed concern that they do not sufficiently benefit from genetic research, that genetics diverts attention and resources away from nongenetic causes of health disparities including racism, and that genetic research could actually reinforce racism by implying that Indigenous people have genetic susceptibilities to disease and are thus inherently unhealthy (Dodson and Williamson, 1999; Paradies et al., 2007; Pearce et al., 2004). Others have criticized the lack of involvement of Indigenous communities in the planning of projects and insensitivity to cultural beliefs in the practice of the research (Arbour and Cook, 2006; Harding et al., 2012; Wallace, 1998; Chang and Lowenthal, 2001).

This article reviews the ethical issues that Indigenous people and others have raised with reference to genetic research projects and biobanks. It focuses on Indigenous populations within nation-states founded through British colonialism, primarily Aboriginal and Torres Strait Islander populations in Australia; First Nations, Inuit and Metis populations in Canada; Native Americans, Native Hawaiians and Alaskan Natives in the United States; and Maori in Aoteoroa/New Zealand. While Indigenous communities differ dramatically both between and within these countries, their common experience of British settler-colonization and the resulting nation-states provides a useful comparative base from which to examine their responses to genetic research (Cunningham and Stanley, 2003). Before discussing the relevant ethical issues, a consideration of the definition and scope of 'indigeneity' is required.

Who Is Indigenous?

The question of who is Indigenous is complex and often contested. The United Nations, for example, resists defining indigeneity, and the United Nations *Declaration of the Rights of Indigenous Peoples* does not offer any definition (United Nations, 2008). However, a publication of the United Nations Permanent Forum on Indigenous Issues indicates that Indigenous people have the following characteristics:

> "Self-identification as Indigenous peoples at the individual level and accepted by the community as their member; historical continuity with precolonial and/or presettler societies; strong link to territories and surrounding natural resources; distinct social, economic or political systems; distinct language, culture and beliefs; form nondominant groups of society; resolve to maintain and reproduce their ancestral environments and systems as distinctive peoples and communities." (United Nations Permanent Forum on Indigenous Issues, undated)

Three major themes are clear in this account: descent from precolonial populations; self-identification and community acceptance; and maintenance of distinct beliefs and practices. Indigenous people from the United States, Canada, Australia, and New Zealand have been the driving forces behind the international Indigenous rights movement and Indigenous scholarship

 http://dx.doi.org/10.1016/B978-0-08-097086-8.82058-9

(Niezen, 2003). Given that the literature on Indigenous genetic research has overwhelmingly focused on these countries, the author will follow suit here. However, Indigenous minorities numbering 370 million people have been recognized across 90 countries, and the issues discussed here may differ in each national setting and language group. Some of these issues have been explored in the literature in other contexts, such as Africa, but often through the concept of 'ethnic minorities' rather than Indigenous people (de Vries et al., 2012).

The Diversity of Indigenous People

Indigenous people are highly diverse at multiple levels: within individual Indigenous communities, between different Indigenous communities, and between different nations. To take Australia as an example, 74% of Aboriginal and Torres Strait Islander people live in urban and regional areas, and 26% live in remote areas. Only 12% of Aboriginal and Torres Strait Islander people speak an Indigenous language at home, with the remainder speaking English at home (Australian Bureau of Statistics, 2010). Over half of Aboriginal and Torres Strait Islanders overall have non-Indigenous partners, a figure that rises above 80% in urban areas (Khoo et al., 2009). These patterns, which are similar in North America and New Zealand, contribute to a highly diverse Indigenous population in terms of appearance, knowledge of traditional cultural practices and socioeconomic position. This diversity should be kept in mind throughout this article as the author's use of the term 'Indigenous people' may erroneously imply a degree of homogeneity.

Ethical Issues and Indigenous People

The ethical issues that apply to Indigenous people should be seen as additional to 'conventional' ethical issues that apply to all people, rather than replacing them. The standard ethical principles of autonomy, beneficence, nonmaleficence and justice, and procedures such as informed consent and ethical review, must be provided to Indigenous participants in genetic research. The additional ethical concerns that are reported in the literature are discussed in sections on group harm; cultural beliefs relating to biospecimens and human origins; community engagement and collective consent; benefit; ownership; and whether biospecimens can and should be 'repatriated.' This range of concerns derives from three interrelated sources: the distinctive cultural beliefs of Indigenous peoples (e.g., concerning blood samples); the collective nature of Indigenous societies (e.g., the need for collective consent); and the dispossession and marginalization of Indigenous people by the state.

In particular, experiences of dispossession, marginalization, and oppression have led many Indigenous people to distrust science and the dominant society in general. This lack of trust underlies many of the issues discussed below. The history of racial classification and stratification of Indigenous people by science and the state has enhanced the extent of suspicion of genetics as compared to other fields of science. In Australia, for example, the assimilation era (1930s–60s) is remembered as a time when the state sought to 'breed out the color' and make Aboriginal and Torres Strait Islander people 'white' through the removal of children of mixed heritage from their families. Many Indigenous people also associate contemporary genetic research with the long history of the procurement, circulation, and research of Indigenous human remains without consent (Bowekaty and Davis, 2003).

More recently, genetic research in Indigenous communities has spurred a number of controversies. The Human Genome Diversity Project (HGDP) is paradigmatic. This effort to collect DNA samples from Indigenous populations worldwide was dubbed the 'Vampire Project' by some Indigenous people and prompted the formation of the US-based international organization the Indigenous People's Council on Biocolonialism (*Indigenous Peoples Council on Biocolonialism*). The Genographic project, launched in 2005, is seen by some as repeat of the HGDP (Tallbear, 2007), although it has avoided the level of public criticism directed against the earlier project. Other recent controversies involving genetic research concern the Havasupai and the Yanomami (see **Box 1 and 2**). Some Indigenous groups have sought to avoid or ban genetic

Box 1 Havasupai

The Havasupai are a Native American tribe who live in the base of the Grand Canyon in Arizona. In the early 1990s they initiated genetic research into diabetes, as they were concerned about the rate of amputations caused by diabetes in their communities. Researchers obtained broad written informed consent for 'medical/behavioral research' when blood samples were collected. Samples were later used for a number of other studies without specific consent, including research into ancient population migrations into the Americas from Asia. The Havasupai discovered their samples had been used in this way when tribe member Carletta Tilousi attended a seminar at Arizona State University (ASU) in 2003. Tribal members felt that this research contradicted their cosmological beliefs and threatened their sovereignty. The samples were used for other studies into schizophrenia that had the potential to stigmatize the Havasupai. The tribe initiated legal action against ASU that was resolved in a settlement in 2010. As part of the settlement, 151 remaining samples were returned to the Havasupai in a ceremony conducted by tribal elders in the ASU laboratory. (Harmon, 2010)

Box 2 Yanomami

Anthropologist Napoleon Chagnon and geneticist James Neel collected blood samples from the Yanomami in Brazil and Venezuela in the 1960s and 70s. The samples have been held since that time at various laboratories in the United States.

In the last decade, Chagnon and Neel have been accused of unethical research practices in their dealings with the Yanomami, although these accusations have not been proven. In the midst of the controversy, the Yanomami began a campaign for the return of their samples on cultural grounds. When a Yanomami Indian dies, their body is cremated and no physical remains or possessions of the dead are kept. In this way, the Yanomami ensure that the deceased departs the world and separate the world of the living from the world of the dead. A blood sample held in a freezer may prevent this process from occurring. The Yanomami requested that samples be returned to them to be ritually destroyed by shamans and elders. Unlike the Havasupai case, the researchers and institutions involved with the Yanomami samples have never fought the tribe's requests. However, the task of returning the samples presents significant challenges. At the time of writing, the quest to return the samples had not yet overcome diplomatic and bureaucratic hurdles. (Couzin-Frankel, 2010)

research in their communities (Kowal et al., 2012; Navajo Nation Council, 2002).

It should be noted that Indigenous critiques of genetic research are one aspect of a broader critique of western research (Smith, 2012). Since the 1980s, Indigenous scholars have demanded that Indigenous people are active participants in research concerning their communities, or ideally, conduct their own research. These critiques have led to specific national guidelines for conducting research in Indigenous communities in Australia, Canada, and New Zealand (CIHR, 2007; HRCNZ, 2008; NHMRC, 2003). Recent reviews of guidelines have argued that the Canadian national guidelines provide the most comprehensive guidance for the conduct of genetic research (Jacobs et al., 2010; Taniguchi et al., 2012).

Group Harms

Many of the ethical issues specific to Indigenous people can be understood through the concept of group harm. Group harms are risks to a group that can arise from the process or outcomes of research. They can impact on the group as a whole or on individuals by virtue of their group membership (Hausman, 2008). The process of conducting a genetic research project can cause disruption in a community. For example, a study that found lower than expected Native American ancestry among participants had the potential to create conflict between different segments of the tribe who questioned whether study participants had the minimum required blood quantum for tribal membership (Foster and Sharp, 2000).

A wider range of group harms is associated with the outcomes of research. Many of these result from the way that Indigenous people are portrayed and treated in mainstream society. A research finding that a group is susceptible to a particular disease has the potential to increase discrimination against Indigenous people in a range of settings including interpersonal racism, racism in health care settings, or discrimination by insurers.

The concept of outcome-related group harm is illustrated by a research project on genes related to the enzyme monoamine oxidase-A (MAO-A) in a Maori population. The study aimed to understand genetic determinants of tobacco use and response to cessation treatment, processes which involve the MAO-A enzyme. However, the particular variant of the gene in question has also been associated with risk-taking and aggression, and is consequently known as the 'warrior gene.' The study was widely reported in the international media as proof that Maori were genetically predetermined to be violent. A Maori academic argued that linking the gene variant with high levels of violence among Maori is scientifically unsound, contributes to racial stereotyping, and may lead to genetic and racial discrimination by insurance companies (Hook, 2009: p. 6).

This example illustrates that while the risk of group harm is greater when the condition studied is potentially stigmatizing (e.g., a study of the genetic associations of alcoholism), stigmatization may be associated with the gene studied rather than the condition itself. All people who identify as Maori were at risk of potential harm from the outcomes of this project, whether or not they personally participated in the study. As Rothstein has noted, measures to protect individuals in research, such as the deidentification of data, do not address outcome-related group harms (Rothstein, 2010). Outcome-related group harms are by nature difficult to control. This highlights the importance of robust consent processes as discussed in the 'community engagement and collective consent' section below.

Cultural Beliefs

Some Indigenous groups hold cultural beliefs that have implications for the handling of biospecimens in genetic research. Indigenous scholars in Australia, the United States, and New Zealand have expressed that 'DNA is sacred.' This generally refers to the extension of holistic approaches to the body, the self, culture and nature to encompass DNA. Any part of the body, including blood or DNA samples, may be considered a piece of collective cultural heritage, indivisible from the self, other members of the group and the environment. This view is illustrated by the words of Native American geneticist the late Frank Dukepoo: "To us, any part of ourselves is sacred. Scientists say it is just DNA. For an Indian, it is not just DNA, it is part of a person, it is sacred, with deep religious significance. It is part of the essence of a person" (Dukepoo, 1998).

Indigenous people across the Pacific have argued that the use of blood and tissue samples can negatively impact on traditional spiritual beliefs. Blood samples are imbued with the 'life force' of the individual, known as *mana* in Hawaii (Fong et al., 2006), *ngeia* in Tonga (Dickenson, 2005) and *tapu* in Maori culture (Mead, 1996). The use of samples for purposes that are not explicitly consented, particularly in the case where the donor has died, can undermine the dignity of that person. For Maori, this can also have implications for donor's descendants. As Mead explains, "A physical gene is imbued with a life spirit handed down from the ancestors, contributed to by each successive generation, and passed on to future generations" (Mead, 1996). The misuse of blood samples is a violation of *tapu*, diminishing the amount of *tapu* available to descendants and affronting one's ancestors.

Indigenous people have drawn upon their cultural beliefs in devising culturally grounded approaches to genetic research and biobanks. Some Maori genetic research projects have involved the formation of a 'kaitiaki' group composed of Maori people who oversee the use of biospecimens and ensure that research protects Indigenous culture and benefits the community (Durie, 2004; Hudson et al., 2007). Durie describes a project where two elders were employed as 'spiritual guardians' for blood and urine samples, accompanying them to the research laboratory, inspecting protocols for biospecimen handling and disposal, conducting a ceremony to make the environment culturally 'safe' for biospecimen handling, and even traveling with the samples to a laboratory overseas (Durie, 2004). The Christchurch Tissue Bank in New Zealand is another example of the recognition of Indigenous cultural beliefs. It offers participants the option of having their sample disposed of with a Maori blessing or *karakia* when it is no longer required for research (Morrin et al., 2005). Such measures can be understood as attempts to reduce cultural harm (Tsosie, 2007).

Indigenous cultural beliefs about human origins are another potential source of contestation between scientists and Indigenous people. The cosmological beliefs of many Indigenous

cultures involve the idea that the group was created *in situ*, and did not migrate to their traditional lands from elsewhere. The dominant scientific account is that *Homo sapiens* evolved in Africa and migration from there to the rest of the world beginning around 100 000 years ago. Conflict between these two accounts has been most pronounced in North America, where Native Americans have taken offense at scientific explanations of Native American migration through the Bering Strait 15 000–45 000 years ago (Foster and Sharp, 2000). This issue is often at the center of contestation over scientific research on Native American ancestral remains (Tallbear, 2003), and the HGDP was also criticized on these grounds. Some have argued that scientific accounts of population history threaten Indigenous people's land claims and sovereignty rights (Harmon, 2010), although to date scientific knowledge about ancient population movements has not been used by any government to erode these rights where they exist.

Community Engagement and Collective Consent

Collective consent with Indigenous organizations is an important strategy to manage the potential risks of group harm and cultural harm associated with genetic research. As the risks cannot be entirely alleviated or prevented, it is important that (1) members of the collective are engaged at all stages of the project in order to influence the research questions, data and biospecimen collection, analysis and communication of results to minimize potential risks (and maximize potential benefits, see below), and (2) collective consent to conduct the research is provided in addition to individual consent from participants. An organization with a representative function in the Indigenous community of interest will consider the potential risks and benefits of the project *to the collective* and make a decision about participation. In many cases, permission or consent from an Indigenous organization is required to obtain a research permit or ethical approval.

Organizations such as tribal organizations (North America), iwi (tribal) health boards or organizations (Aotearoa/New Zealand) and Aboriginal community-controlled health organizations or land councils (Australia) regularly receive and process requests from researchers to conduct research in the communities within their purview. Information provided to collective organizations to inform the decision-making process would include the scientific rationale, potential risks and benefits of the study, confidentiality and privacy, return of results and genetic counseling (if applicable), ownership of samples and data, storage and disposal of specimens and data, third party use of specimens and data, future contact for additional research, withdrawal from research, intellectual property arrangements, commercialization, benefit sharing, and community review of draft publications. In some cases, researchers and Indigenous representatives will develop a formal Memorandum of Understanding covering some or all of these issues.

Community engagement can also inform the individual consent process, as indicated by a recent study of the attitudes of Alaskan Natives toward biospecimen handling and genetic research (Hiratsuka et al., 2012). Many participants believed that consent processes should disclose more information than is usually provided. They wanted to know, for example, "what type of container specimens are stored in, how exactly specimens are preserved over time (e.g., how they are protected from freezer burn), and the security of the storage facility … individuals mentioned a "destroy by" date or the option for choosing to have the sample destroyed after a participant's death" (Hiratsuka et al., 2012: p. 3). Community engagement and collective consent can reveal local preferences for biospecimen handling and what should be discussed in the consent process.

Genetic research often involves storing samples for unknown future uses, either within an established biobank or an individual project collection. In such cases, researchers often obtain 'broad consent' for future uses. However, within Indigenous research there is a general movement against broad consent, which may be considered insufficient to protect against group harms. For example, the Indian Health Service guidelines for handling biospecimens note that broad or 'blanket' consent is by definition uninformed, as it allows for potentially stigmatizing research to occur, such as research into the genetics of alcoholism (Freeman, 2001). Instead, consent should be specific to a project or a limited range of future research.

While these approaches to limiting group harm address some important problems, they raise others. Both collective consent and specific consent invoke the possibility that the individual autonomy of an Indigenous person to consent to research, or to provide broad consent, could be overridden by the decision of a collective body. In some instances, such as biobanks, an injunction against broad consent could exclude Indigenous people from participating if other models of consent are not available (e.g., because there is insufficient funding to request individual consent for each future use). These problems are partially addressed through incorporate Indigenous governance into a biobank, as discussed in the conclusion.

Benefit

Research participation is thought to significantly depend on altruism, with research participants contributing toward scientific discovery in order to benefit society as a whole. Genetic research is likely to involve long lead times before any benefits accrue, such as improvements in health care. Indigenous people may not perceive that they benefit from general medical and scientific advances because of their marginalization from mainstream society and poor access to health care (Du Plessis et al., 2004). Bad experiences of genetic research, such as those examples discussed above, can reinforce the perception that Indigenous people will not benefit.

For these reasons, ethical guidelines indicate that research should offer direct benefits to Indigenous communities (CIHR, 2007; HRCNZ, 2008; NHMRC, 2003). For example, providing employment, training, and professional development opportunities for Indigenous people ensures that Indigenous communities directly benefit from research projects and build the capacity of the Indigenous research workforce.

Ownership

The cases of John Moore and Henrietta Lacks attest that donor property rights over samples and data has long been a vexed

issue in human biospecimen research. Current ethical approaches tend to focus on autonomy and sidestep issues of ownership (Charo, 2006). In an Indigenous context, cultural beliefs about biospecimens outlined above can encompass collective ownership of samples. Indigenous communities often view their natural and cultural resources as collectively owned. In cases this has been explicitly extended to DNA, whereby the DNA of an Indigenous person is the property of the collective as well as the individual (Mead, 1996; Dodson and Williamson, 1999). This is more likely to apply when researchers collect biospecimens in order to 'represent' the genetic variation of the collective (Indigenous Peoples Council on Biocolonialism, 2000).

The principle of collective ownership forms the basis of Arbour and Cook's concept of 'DNA on loan.' They argue that, "in keeping with cultural perspectives and the principles of participatory action research," biosamples should be considered the property of individual donors and the community. Samples are 'on loan' to researchers who act as 'stewards' (Arbour and Cook, 2006). Arbour and Cook do not discuss the legal implications of recognizing the body as property. In practice, the legal limitations of the concept may be of little consequence: the principle of 'DNA on loan' works to ensure that the individual and collective autonomy of Indigenous donors is respected. It addresses a core concern that Indigenous people from various countries have expressed about genetic research: the use of biospecimens (and data) for research that they do not agree with on cultural or moral grounds (Cunningham, 1998; Du Plessis et al., 2004; Foster and Sharp, 2000; Indigenous Peoples Council on Biocolonialism, 2000). It is not surprising, then, that a key consequence of retained donor ownership of samples is that specific consent (from individuals or community organizations) is required for any use of samples or data beyond the initial project, even if samples or data are used without identifiers (Arbour and Cook, 2006). This relates to the move away from broad consent discussed above.

'Repatriation' of Blood Samples

The concept of a loan implies an eventual return. Arbour and Cook do not discuss this issue, although the Canadian Institutes of Health Research guidelines (which draw on the concept of samples 'on loan') state that researchers should outline whether samples will be returned "if possible, and desired by the research participant" (CIHR, 2007: p. 32). This raises the question: is the return of blood samples to Indigenous people an instance of 'repatriation'?

The repatriation of ancestral human remains held by museums and universities to the Indigenous community of origin began in the 1980s in response to international campaigns by Indigenous activists. It has become a widespread practice in Australia, Canada, the United States, New Zealand, and more recently Africa. It is guided by national legislation such as the Native American Graves Protection and Repatriation Act 1990 (NAGPRA) in the United States as well as various institutional policies.

Under most definitions, human remains are bodies or parts of bodies originally intended for burial. Blood or saliva samples collected for scientific research are therefore not recognized as human remains for the purposes of repatriation. However, the term 'repatriation' is used by commentators discussing the return of blood samples to the Havasupai and Yanomami, and the 2004 case of the Nuu-chah-nulth in Canada (Reardon and TallBear, 2012; Wiwchar, 2004). This remains an emerging ethical issue in Indigenous genetic research that could affect many thousands of old biospecimens currently held in institutions around the world.

An important implication of the 'repatriation' paradigm is that it acts across national borders. While the various ethical issues the author has outlined in this article are enshrined in national guidelines and are built into ethical review, the use of previously collected Indigenous biospecimens in countries other than where they are collected occurs outside such guidelines. This issue was highlighted in 2011 when the 'first Aboriginal genome' was sequenced by a team of researchers based in Denmark using a hair sample collected in 1932 from a train station in remote Western Australia and held at the University of Cambridge. Researchers sought approval for the study from an Aboriginal land council in the region where the sample was collected, even though the Danish ethics committee that approved the study thought this unnecessary (Calloway, 2011). It is likely that the distinction between human remains destined for repatriation and biospecimens available for scientific study will continue to be a source of ethical uncertainty.

Conclusion

Scholars have expressed optimism about 'bridging the divide' between genomics and Indigenous peoples in order that Indigenous people can benefit from the potential of genomics to improve their health (Jacobs et al., 2010). Clearly, there is much work to be done to address the ethical concerns that Indigenous people have raised, particularly unresolved and emerging issues such as collective ownership of samples, repatriation, and the use of Indigenous biospecimens that exist beyond national borders.

The creation of biospecimen collections under robust Indigenous governance is one way to address many ethical concerns. A prominent example is the Alaskan Area Specimen Bank, a resource of over 500 000 biospecimens of which 76% were collected from Native Alaskans. The collection is governed by the Alaska Area Specimen Bank Working Group, a body that includes representatives from Alaskan Native Tribal Health Organizations and the Indian Health Service. All projects that seek to access the Bank need approval from a Tribal Health Organization, and also require individual consent if identifying information is to be used (Parkinson et al., 2013). Another newly established collection of Indigenous biospecimens under Indigenous governance is the National Centre for Indigenous Genomics at the Australian National University.

Tools to educate and empower Indigenous people to take an active role in genetic research projects are another mechanism of 'bridging the divide.' Again, the United States offers the most advanced models. The *American Indian and Alaskan Native Genetics Resource Center* is a Web site developed by the National

Congress of American Indians to provide resources for tribes who are considering participating in genetic research or developing their own research projects (*American Indian and Alaska Native Genetics Resource Center*). Another key initiative is the Summer Institute for Native Americans in Genomics, hosted annually at the University of Illinois since 2011 and open to Indigenous people from across the Americas (The Institute for Genomic Biology). It aims to introduce participants to the principles and practices of genomics and informatics; facilitate discussion of bioethics in Indigenous genetic research; and promote science careers for Indigenous people. Similar initiatives in other countries would promote ethical practices in Indigenous genetic research.

Understanding the ethical issues related to Indigenous genetic research is important to uphold the rights of Indigenous people, and also has wider relevance. Bioethicists are increasingly looking to Indigenous research ethics to inform wider practices (Mello and Wolf, 2010; Rothstein, 2010). As Taniguchi et al. have recently phrased it, "Indigenous groups are most vulnerable to research exploitation and harm; therefore, identifying principles that work for Indigenous people will lead to best practices for all populations" (Taniguchi et al., 2012). The ethical innovations and uncertainties of Indigenous genetic research are important considerations for everyone concerned with the genomic future.

See also: Biobanking: Ethical Issues; Bioethics in the Post-genomic Era; Bioethics: Genetics and Genomics; Bioethics: Philosophical Aspects; Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute; Genetic Ancestry Testing; Genetics and Anthropology; Genetics and Society; Genetics and Sociology; Genographic Project; Genomics, Ethical Issues In; Human Genome Diversity Project: History; Indigenous Knowledge: Science and Technology Studies; New Genetics and Race; Postcolonial Science Studies.

Bibliography

American Indian and Alaska Native Genetics Resource Center, 2013. http://genetics.ncai.org.

Arbour, L., Cook, D., 2006. DNA on loan: issues to consider when carrying out genetic research with aboriginal families and communities. Community Genetics 9, 153–160.

Australian Bureau of Statistics, 2010. Population Characteristics, Aboriginal and Torres Strait Islander Australians, 2006. Australian Bureau of Statistics, Canberra.

Australian Institute of Health and Welfare & Australian Bureau of Statistics, 2008. The Health and Welfare of Australia's Aboriginal and Torres Strait Islander Peoples, 2008. Australian Bureau of Statistics, Canberra.

Bowekaty, M., Davis, D., 2003. Cultural issues in genetic research with American Indian and Alaskan Native people. IRB: Ethics and Human Research 25 (4), 12–15.

Calloway, E., 2011. Aboriginal genome analysis comes to grips with ethics. Nature 477, 522–523.

Chang, R.M., Lowenthal, P.H., 2001. Genetic Research and the Vulnerability of Native Hawaiians. Public Health Dialogue 8, 364–367.

Charo, R., 2006. Body of research – ownership and use of human tissue. NEJM 355 (15), 1517–1519.

CIHR, 2007. CIHR Guidelines for Health Research Involving Aboriginal People. Canadian Institutes of Health Research, pp. 1–40.

Cooke, M., Mitrou, F., Lawrence, D., Guimond, E., Beavon, D., 2007. Indigenous well-being in four countries: an application of the UNDP'S human development index to indigenous peoples in Australia, Canada, New Zealand, and the United States. BMC International Health and Human Rights 7 (9).

Couzin-Frankel, J., 2010. Researchers to return blood samples to the Yanomamo. Science 328 (5983), 1218.

Cunningham, H., 1998. Colonial encounters in postcolonial contexts: patenting indigenous DNA and the human genome diversity project. Critique of Anthropology 18 (2), 205–233.

Cunningham, C., Stanley, F., 2003. Indigenous by Definition, Experience, or World View. BMJ 327, 403–404.

de Vries, J., Jallow, M., Williams, T.N., Kwiatkowski, D., Parker, M., Fitzpatrick, R., 2012. Investigating the potential for ethnic group harm in collaborative genomics research in Africa: is ethnic stigmatisation likely? Social Science & Medicine 75 (8), 1400–1407.

Dickenson, D., 2005. Human tissue and global ethics. Genomics, Society and Policy 1 (1), 41–53.

Dodson, M., Williamson, R., 1999. Indigenous peoples and the morality of the Human Genome Diversity Project. Journal of Medical Ethics 25 (2), 204–208.

Du Plessis, R., Scott, A., Phillips, H., Cram, F., Tipene-Matua, B., Winstanley, A., 2004. The Social, Cultural, Ethical and Spiritual Implications of Genetic Testing, 3. Social Science Research Centre, University of Canterbury, New Zealand.

Dukepoo, F.C., 1998. Genetic services in the new era: native american perspectives. Community Genetics 1, 130–133.

Durie, M., 2004. Understanding health and illness: research at the interface between science and indigenous knowledge. International Journal of Epidemiology 33, 1138–1143.

Fong, M., Braun, K.L., Chang, R.M., 2006. Native Hawaiian preferences for informed consent and disclosure of results from genetic research. Journal Cancer Education 21, S47–S52.

Foster, M., Sharp, R., 2000. Genetic research and culturally specific risks: one size does not fit all. Trends in Genetics 16 (2), 93–95.

Freeman, W.L., 2001. IHS Guidelines for Implementing and Complying with IHS Policy on Specimens. http://www.aatchb.org/epi/docs/ResearchEthics/T3-Specimen.pdf (viewed 25.05.13.).

Harding, A., Harper, B., Stone, D., O'neill, C., Berger, P., Harris, S., Donatuto, J., 2012. Conducting Research with Tribal Communities: Sovereignty, Ethics and Data-Sharing Issues. Environmental Health Perspectives 120, 6–10.

Harmon, A., 2010. Indian tribe wins fight to limit research of its DNA. New York Times p. A1.

Hausman, D., 2008. Protecting groups from genetic research. Bioethics 22 (3), 157–165.

Hiratsuka, V.Y., Brown, J.K., Hoeft, T.J., Dillard, D.A., 2012. Alaska Native people's perceptions, understandings, and expectations for research involving biological specimens. International Journal of Circumpolar Health 71 (18642), 1–6.

Hook, R., 2009. "Warrior Genes" the Disease of Being Maori. MAI Review, 1–11.

HRCNZ, 2008. Guideliners for Researchers on Health Research Involving Maori. Health Research Council of New Zealand, Auckland.

Hudson, M., Ahuriri-Driscoll, A., Lea, M., Lea, R., 2007. Whakapapa – a foundation for genetic research? Journal of Bioethical Inquiry 4 (1), 43–49.

Indigenous Peoples Council on Biocolonialism. http://www.ipcb.org (viewed 05.06.13.).

Indigenous Peoples Council on Biocolonialism, 2000. Indigenous Peoples, Genes and Genetics: What Indigenous Peoples Should Know about Biocolonialism. http://www.ipcb.org/publications/primers/index.html (viewed 15.08.10.).

Jacobs, B., Roffenbender, J., Collmann, J., Cherry, K., Lee Bitsói, L., Bassett, K., Evans, C.H., 2010. Bridging the divide between genomic science and indigenous peoples. The Journal of Law, Medicine & Ethics 38 (3), 684–696.

Khoo, S.-E., Birrell, B., Heard, G., 2009. Intermarriage by birthplace and ancestry in Australia. People & Place 17 (1), 15–28.

Kowal, E., Pearson, G., Peacock, C.S., Jamieson, S., Blackwell, J., 2012. Genetic research and aboriginal and Torres Strait Islander Australians. Journal of Bioethical Inquiry 9 (4), 419–432.

McGregor, J., 2010. Racial, ethnic, and tribal classifications in biomedical research with biological and group harm. The American Journal of Bioethics 10 (9), 23–24.

Mead, A., 1996. Genealogy, sacredness, and the commodities market. Cultural Survival Quarterly 20 (2), 46–52.

Mello, M.M., Wolf, L.E., 2010. The Havasupai Indian tribe Case—Lessons for research involving stored biologic samples. New England Journal of Medicine 363 (3), 204–207.

Morrin, H., Gunningham, S., Currie, M., Dachs, G., Fox, F., Robinson, B., 2005. The Christchurch Tissue Bank to support cancer research. Journal of the New Zealand Medical Association 118 (1225), 1–12.

National Centre for Indigenous Genomics, 2014. http://ncig.anu.edu.au.
Navajo Nation Council, 2002. Approving a Moratorium on Genetic Research Studies Conducted within the Jurisdiction of the Navajo Nation until Such Time that a Navajo Nation Human Research Code Has Been Amended by the Navajo Nation Council. Health and Social Services Committee of the Navajo Nation Council, Window Rock, AZ.
NHMRC, 2003. Values and Ethics: Guidelines for Conduct of Aboriginal and Torres Strait Islander Health Research. National Health and Medical Research Council, Canberra. http://www.health.gov.au/nhmrc/publications/pdf/e52.pdf.
Niezen, R., 2003. The Origins of Indigenism. University of California Press, Berkeley.
Paradies, Y., Montoya, M.J., Fullerton, S.M., 2007. Racialized genetics and the study of complex diseases: the thrifty genotype revisited. Perspectives in Biology and Medicine 50 (2), 203–227.
Parkinson, A.J., Hennessy, T., Bulkow, L., Smith, H.S., Alaska Area Specimen Bank Working Group, 2013. The Alaska Area Specimen Bank: a tribal-federal partnership to maintain and manage a resource for health research. International Journal of Circumpolar Health 72 (20607), 1–5.
Pearce, N., Foliaki, S., Sporle, A., Cunningham, C., 2004. Genetics, race, ethnicity, and health. BMJ 328, 1070–1072.
Ramos, E., Rotimi, C., 2009. The A's, G's, C's, and T's of health disparities. BMC Medical Genomics 2 (1), 29.
Reardon, J., TallBear, K., 2012. Your DNA is our history. Current Anthropology 53 (5), s233–s245.
Rothstein, M., 2010. Is deidentification sufficient to protect health privacy in research? American Journal of Bioethics 10 (9), 3–11.
Scott, A., Phillips, H., Moore, A., Du Plessis, R., 2005. Ethics in practice: conversations about biobanks. Critical Public Health 15 (4), 359–368.
Smith, L.T., 2012. Decolonizing Methodologies, Research and Indigenous Peoples, second ed. Zed Books Ltd, London and New York.
Tallbear, K., 2003. DNA, blood, and racializing the tribe. Wicazo SA Review Spring, 81–107.
Tallbear, K., 2007. Narratives of race and indigeneity in the Genographic Project. Journal of Law, Medicine and Ethics 35 (3), 412–424.
Taniguchi, N., Taualii, M., Maddock, J., 2012. A comparative analysis of indigenous research guidelines to inform genomic research in indigenous communities. International Indigenous Policy Journal 3 (1). Article 6.
The Institute for Genomic Biology, 2013. Summer Internship for Native Americans in Genomics. http://conferences.igb.illinois.edu/sing/.
Tsosie, R., 2007. Cultural challenges to biotechnology: native American genetic resources and the concept of cultural harm. Journal of Law, Medicine, and Ethics 35 (3), 396–411.
United Nations, 2008. United Nations Declaration on the Rights of Indigenous Peoples. http://www.un.org/esa/socdev/unpfii/documents/DRIPS_en.pdf.
United Nations Permanent Forum on Indigenous Issues, undated. Indigenous People, Indigenous Voices: Factsheet. United Nations, New York.
Wallace, R., 1998. The Human Genome Diversity Project: Medical Benefits Versus Ethical Concerns. Molecular Medicine Today 4, 364–367.
Wiwchar, D., 2004. Nuu-chah-nulth blood returns to West Coast. Ha-shilth-sa 31 (25), 1–4.

Genetics and Social Justice

Lindsay A Smith, University of New Mexico, Albuquerque, NM, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines four genetic technologies as they have been articulated as tools for social justice: kinship analysis in postconflict settings, forensic DNA and criminal exoneration, genetic testing and women's health, and golden rice and food security. The cases highlight the potentials and pitfalls of science in the service of justice and the importance of recognizing technologies as hybrid political and scientific tools best implemented within a politics of humility and governed by social actors rather than scientists.

From the earliest days of genetics, concerns for social justice have been embedded it its practice. Since the days of Galton's coining of the term eugenics in the years after reading his cousin Darwin's monograph *On the Origin of Species*, there has been an enduring preoccupation with the potential for genetic knowledge and technology to both explain and shift the inequalities of human society (Gillham, 2001). In the twentieth century, new technologies like karyotyping and recombinant DNA were directly applied to social concerns through medical and agricultural technologies like prenatal genetic testing and genetically modified crops. As the controversies surrounding these examples show, however, scientific solutions to complex social problems like human reproduction, disability, and the global population boom, have also generated intense ethical debate and, in many cases, contributed to real and lasting social harm. Given this complexity, can genetics be used as a tool for social justice?

Genetic technologies and social justice movements have coalesced around four major areas: (1) the family and kinship; (2) criminal justice and forensics, (3) human health and disease; and (4) environment and sustainability. Highlighting one technology in each of these categories, this article offers a review of the challenges of genetic technology-driven social justice movements. Rather than focus only on successful sites of intervention, these cases emphasize the need for deliberation and suggest three common challenges of genetics as a tool for social justice: the use of race as a proxy for genetic difference, biological reductionism, and commercialization. The review suggests that genetics best coalesces with the goals for social justice in cases where technology is governed by a politics of humility, owned and managed by social rather than scientific actors, and conceived of as one tool within broader social and political movements rather than a magic bullet capable of quickly cutting through messy social issues.

Genetics and Social Justice: Between the Individual and Society

The sequencing of the human genome in 2000 heralded the unlocking of the secrets of humankind. As a scientific benchmark, the human genome project promised revolutions in the management of disease, health, morphology, development, and evolution. The sequencing of the genome also offered an important platform for social justice by establishing humanity's 99.9% genetic similarity. Revealed were humans' deep interconnections, literally through ties of marriage and reproduction and metaphysically through millions of years of evolutionary development. At the same time, the genetic revolution also inaugurated a focus on the individual specificity of every human being and much of the technological innovation of genetics has focused on individuation.

Justice, like genetics, also straddles this uneasy divide between the collective and the individual. In contemporary usage, social justice can be synonymous with human rights, refering to reform or change at a collective level to ameliorate the deep and lasting disparities between particular segments of society. Social justice, as a theoretical concept, emerged as a critique of traditional theories of justice, which had focused on individual good and interpersonal harm through the rubric of property. With the rise of labor and social reforms in the face of the expanding inequalities of the Industrial Revolution, thinkers and activists sought to articulate new formulations of justice to account for the harm done by structural inequalities and the reproduction of social class. Where traditional forms of justice have focused on individual violations codified under penal and civil codes, social justice has worked at the level of the harms caused by the status quo of social and state structures (Capeheart and Milovanovic, 2007).

Justice and genetics meet in this uneasy space between the individual and the collective. For genetics, this has been articulated in debates about whether the potential of the field is best realized through greater personalization and individualization or through an emphasis on similarities and shared evolutionary history. Social analysts of genetic science have pointed to the importance of a balanced focus on the dynamic interplay of these population and individual forces rather than a single causal explanatory framework (Lewontin, 2001). These questions are subjects of important debate in society and highlight the paradoxical nature of both genomic science and social justice at the intersection of the individual and society. When genetic technologies are developed as a potential intervention into inequality, they necessarily invoke a third social sphere of institutional policy and power. In cases where science works toward social justice goals, the hybrid reality of all science and technology becomes apparent as innovations within laboratory walls are invested with social, legal, or state power.

Historical Background: Eugenics and the Progressive Spirit

Eugenics remains a necessary starting point for considering the tight linking of societal and genetic technologies. Although eugenics is now roundly rejected as ideologically driven pseudoscience, in its heyday this scientifically driven social reform was championed as a powerful way of enacting social justice. Many liberal reformers of the day saw eugenics as a humane and effective way to stamp out systemic social ills like poverty, alcoholism, and mental illness. Encouraging "better breeding" through selective marriages between individuals with socially valued, culturally determined traits from ethnic phenotypes to perceived intelligence and economic productivity and discouraging the reproduction of "feeble" stocks. Advocates of eugenics as social improvement argued that this science would, within a few generations, transform human populations, much like agricultural breeders had created new productivity and utility out of the germplasm of domesticated animals and plants. Oliver Wendell Holmes, the Chief Justice of the Supreme Court and renowned American progressive, even wrote the majority opinion upholding the right of the state of Virginia to sterilize Carrie Buck an institutionalized young woman whose mother had also been found social unfit. He summed up the eugenic ethos of selective breeding as public good in his 1927 majority opinion. Comparing tubal ligation and vaccination as legally equivalent public health measures, he declared that this preventative act would do much more for the social good than subsequent more punitive, responses to the social results of bad breeding: "It is better for all the world if, instead of waiting to execute degenerate offspring for crime or to let them starve for their imbecility, society can prevent those who are manifestly unfit from continuing their kind" (*Buck v. Bell*, 1927).

Genetic technologies in the form of early eugenics were complicit with some of the worst human atrocities of the twentieth century. This period of genetic research is notable for its employment of genealogical studies of traits and the early development of twin studies, which remain standard techniques of genetic research. However, eugenics ended in brutal practices of forced sterilization in Europe and the United States and ultimately genocide during the Second World War (WWII). It is important to consider eugenics within a broader analysis of genetics and social justice, because it highlights the ways in which justice and social good are not absolute qualities, but rather have been and continue to be the subject of fierce ideological and political battles. Scientific technologies are part of ideological systems of social transformation, coming into being not as value-neutral tools for truth but because of politics, moral beliefs, and visions for particular kinds of futures.

Given this violent history, human genetics in the 1950s reformulated itself through the Modern Evolutionary Synthesis toward a focus on development, fitness, and discrete measurable traits. As human and medical genetics reached a new level of utility with the ability to visualize chromosomes through karyotyping (e.g., tests for autosomal anomalies, chromosomal based sex determination) and later manipulate and sequence the genome, it also increasingly had direct applications in human life and biology. This emerged in legal contexts as paternity testing and later DNA analysis and kinship analysis, in medical contexts as organ donor matching, amniocentesis, genetic registries for recessive diseases, and most recently preimplantation genetic diagnosis, and in agriculture through increased experimentation and control over plant and animal breeding. With this rapid expansion of genetic technologies, new debates arose over their potential for both social good and the hardening of biological notions of human difference.

Family: Kinship Analysis and Transitional Justice

Nelida Navajas received a phone call right before bed on the 13 July 1976. The woman on the line told Nelida to come quickly to her daughter Cristina's apartment because the children were alone and crying. When she arrived at the apartment, she could hear the children screaming from the sidewalk as she parked her car. She ran upstairs to find the apartment in disarray and three distraught children, her grandsons and their cousin. Nelida's pregnant daughter and the two other women, who had been caring for the children, had been forcibly detained by masked men with guns. Nelida never saw her daughter again. Cristina became one of the 30 000 *desaparecidos*, disappeared men and women, who between 1976 and 1984 were forcibly detained by the Argentine military dictatorship as part of a violent military crackdown on social and political life under the guise of a 'dirty war' against communist subversives (CONADEP, 2003). For years, Nelida explained, she could not hear the sound of a child screaming without getting sick (Nelida Navajas. Interview by author. 2005.).

While her daughter and the child Cristina was pregnant with continuing to remain disappeared, Nelida spent her life searching for them. She was a founding member of the *Abuelas* (Grandmothers) of the Plaza de Mayo, an organization of women dedicated to finding their missing children and grandchildren. Through their activism the Abuelas discovered that the military dictatorship was trafficking in the children of young women like Cristina, rather than killing the babies when they were born. Nelida's grandchild and an estimated 500 other missing children were likely alive, appropriated by military families or others with close ties to the ruling class. The Abuelas had found evidence of the survival of their missing grandchildren, but did not know how to prove their relatedness to these children so that they could be legally returned to the families in which these babies were born. In 1976, when Nelida's grandchild would have been born and illegally trafficked by the government, genetic testing for relatedness was in its technological infancy. Developments in organ transplant technology and histocompatibility studies among families, led the scientists the Grandmothers consulted to believe it might be possible that, based on blood alone, genetics could establish the relatedness between a grandchild and grandparent in the absence of the parental generation. A transnational team of scientists under the aegis of the American Association for the Advancement of Science (AAAS) developed the first genetic tool for human rights, the Index of Grandpaternity, which in 1984 was used to return Paula Logares, who was born in captivity, to her grandmother (Arditti, 1999). This use of genetics to connect families separated by genocide, though routine in the present, was technologically impressive and politically unprecedented in 1984. Scientists, like Mary Claire King, who headed the AAAS

team as well as other members, developed a new agenda for politically engaged scientists to use their expertise as tools for social justice. From the tsunami in Thailand to the World Trade Center in the United States, family groups, scientists, and activists have worked together to create DNA databases for the purposes of family reunification. This work has focused both on identifying the dead as in Bosnia after the genocide at Srebrenicia, and the living, as in the case of children displaced or kidnapped during the chaos of war and disaster.

Criminal Justice: Forensic DNA and Wrongful Convictions

After serving 11 years of a 27-year criminal sentence, Mr Anthony Robinson was released on parole. Convicted in 1987 of raping a white college-aged girl in the University of Houston campus, Mr Robinson had maintained his innocence throughout his trial and prison term. When he was released on probation, he worked until he had made enough money to hire a lawyer. Even his lawyer did not believe his story and suggested he save the funds to have the biological samples from his case undergo forensic DNA analysis. They would wait and see what the results showed before pursuing the case aggressively. At the time of his trial, forensic DNA typing was not widely in use and blood typing was the primary forensic evidence presented. Although he petitioned up to the Supreme Court to have his case reexamined, his conviction stood. Robinson, now a lawyer, understood his case as emblematic of a racist judicial system. As a black man in the South, accused of raping a white woman, old prejudices ran deep, coloring the interpretations of everyone involved from the jury and judge to his own public defender. They could not see beyond their prejudices and believe his claims of innocence. Like many of the other 316 individuals who have been exonerated through the work of the Innocence Project, a group devoted to using forensic DNA to address wrongful convictions, Mr. Robinson found himself caught up in the structural inequalities of American justice. Although his words were disregarded, the forensic DNA test corroborated his claims of innocence, showing that the body fluids at the scene of the crime did not match him. In 2000, after the test was repeated by the Texas authorities, Anthony Robinson was officially pardoned (Anthony Robison. Interview by author. March 2002.).

The Innocence Project was founded in 1992 through collaboration between Barry Scheck and Peter Neufield at the Cordozo Law School. It is dedicated to helping prisoners gain accesss to resources to contest wrongful convictions. Although the organization now works on cases where DNA identification is not applicable, its initial mission was to bring the power of genetics to victims of an unjust system. DNA, in the eyes of the founders, both veterans of the OJ Simpson defense team, had the potential to offer a new kind of objectivity in some of the most problematic crimes: murder and sexual assault. As defense lawyers, they understood the system as inherently biased against the poor and people of color, leaving an overwhelmingly disproportionate number of black men behind bars. Forensic DNA analysis, although controversial scientifically when it was first introduced as a tool of criminal prosecution (Aronson, 2007), nonetheless had exceptional strengths for exoneration in these cases. Including someone as a potential perpetrator, of a crime especially in the earliest days of genetic technologies, when the statistical power of matches was relatively low, was legally problematic; but excluding people was straightforward. If the biological sample from the crime scene and the sample from the person convicted for the crime did not match at the loci tested in the genetic analysis, the DNA test provided powerful proof of their potential innocence. The logic behind this appeal to genetics was that the objectivity of science could serve as a corrective to the structural inequalities of the justice system. Genetic technologies would be the agents of social justice where other legal systems of review, like the appellate system, had failed. As of 2014, the Innocence Project had helped to exonerate 316 individuals, including Anthony Robinson. Although genetic analysis is not applicable in most criminal cases, the organization sees genetic science as a means of providing visible proof of the failures of the justice system and thereby galvanizing change.

Health: Genetic Testing for Breast Cancer and Health Disparities

In 2013, the actress Angelina Jolie, published an op-ed in the *New York Times* describing her decision to undergo a double mastectomy and breast reconstruction because of her family history of cancer and her positive finding of a mutation in her BRCA1 gene. She describes the difficulty of losing her mother to breast cancer and explaining this loss to her children especially knowing herself to be a carrier of a "'faulty' gene, BRCA1" which greatly increases her risk of the breast and ovarian cancer (Jolie, 2013). In radical detail, Jolie lays out her experience of the procedure of double mastectomy: the preparation to save sensation in her nipples, the marathon surgery to remove her breast tissue, and the pain and mess of recovery. Jolie offers an intimate portrayal of a radical choice. As an actress, who has routinely been sexualized and idealized as one of *People Magazine's* most beautiful woman, her frank discussion of risk, surgery, and recovery had important effects on popular conceptualizations of breast cancer genetics, risk, and the violent choices of prophylaxis. Her op-ed concludes by alluding to underlying social justice potential of genetic testing for a disease that claims the lives of so many women. She encourages women to get tested. Mentioning the cost of testing as a barrier, she alludes to, although does not discuss, the contemporaneous issue of BRCA1 and BRCA2 patents and their effect on research, the cost of testing, and by extension the lives and mortality of women at risk for hereditary breast cancer. Using her influence as a Hollywood star, Angelina Jolie offers a portrait of genetic testing for breast cancer as giving women choices and control over a terrifying illness.

Angelina Jolie's very public discussion of this topic and her own choices in the face of new genetic knowledge, come at the end of more than a decade of research on the genetics of breast cancer. Mary Claire King, who also played a central role in the development of human rights genetics in Latin America, was a pioneer in this field, honing in on the potential loci for these tumor suppressor genes which were known to play an important role in the etiology of breast and ovarian cancer. In the film, *In the Family*, a documentary chronicling Joanna Rudnick's

experience with BRCA testing, King is interviewed and she describes the importance of this focus on the genetics of women's health as concrete form of empowering women with knowledge and choices about their body (Rudnik, 2008). Gendered health disparities are addressed in the same language as the feminist body politics surrounding pregnancy and abortion. Genetic knowledge is coded as empowering choice through basic genetic research and increasing women's access to testing. Although advances in BRCA testing are heralded as an important first step to addressing women's health, advocates acknowledge that much work remains to be done to give women better options than preventive mastectomy in the face of a positive test result through more robust education about risk, prevention, and treatment (Parthasarathy, 2012).

Biodiversity: Golden Rice and Food Security

Upon his retirement, Ingo Potrykus, a European geneticist, whose career focused on agricultural biotechnology, published a summary of one of his major life achievements, the development of 'golden rice' (Potrykus, 2001). His early interest in the potential of transgenic crops to alleviate food insecurity, found form in the successes his laboratory had in the genetic modification of rice. Although he began his career working in more esoteric crops like tobacco and petunias, his realization of the social justice potential of genetics in food security converged on rice because of its centrality as a sustenance crop in the developing world. In the article, Potrykus describes the scourge of Iron and Vitamin A deficiency and the drawbacks of rice as a staple crop given its deficiency in these nutrients. With the support of the Rockefeller Foundation, what he later understands as a fortuitous reliance on public rather than private support, Potrykus' laboratory inserted two genes into rice which formed a molecular pathway for the production of Betacarotene and gave the rice its golden hue. Potrykus offers golden rice as a potential corrective for the general mistrust and negativity surrounding genetically modified crops because, from its funding model to its distribution model, golden rice innovates to address illness and food security in a single improved crop.

Despite the seeming promise of the technology, golden rice has not yet been implemented on a wide scale in any developing context because it has met with entrenched local and global resistance. It has, however, been successful in offering alternative public narrative for genetically modified organisms (GMO's) beyond the critique of the multinational control of food production represented in genetically modified plant and pesticide systems, like 'Roundup Ready' Soy and BT cotton (Stone, 2010). Proponents highlight the crop as a model because of its development through publicly funded research, innovative patent licensing that could make the rice available to the poor, and the potential of this fix to address preventable childhood blindness due to vitamin A deficiency. India has been the primary target of this intervention, and advocates point to the potential important cultural resonances that a yellow rice might have given the widespread use of saffron to achieve this same effect for ritual purposes. Rather than an economically driven attempt to control agricultural markets, golden rice is offered as a way to enhance local farming, health, and cultural life.

Opposition to Social Justice Genetic Technologies

Postconflict DNA identification, DNA-led exonerations, breast cancer genetic testing, and golden rice can be broadly understood as social justice genetic technologies, that is technologies whose development and application have been inextricably intertwined with claims of justice, equity, and societal amelioration. However, this framing tends to overemphasize the intentions of the scientists and advocates developing the technologies and ignore the inevitable friction between intentions and social realities. Each has also met with broad and sustained opposition. The controversies that have emerged when genetic technologies are taken up as pathways to social justice, point to the complexity of the very concept of justice. In addition, these technologies are embedded within a larger ideology of science which presents technology and knowledge as fundamentally apolitical and teleological with knowledge production marching unceasingly toward truth. Science and technology scholars, however, have shown that knowledge is political. From the macrostructures of funding and research agenda-setting to the micropolitics of laboratory life, science like all human endeavors is embedded in systems of power. Initiatives in genetics to address human inequality and social suffering do not escape these same power structures simply because of the good intentions of their creators. As the historical case of eugenics illuminates, broader social ideologies can be reinforced through technological development and practice.

Scientific social justice initiatives often have a familiar narrative structure to their development: socially committed scientists recognize that through their specialized skills they might be able to decrease inequality and address specific social harms. We see this in all the cases discussed, from eugenics to DNA-led exonerations and biofortified GMO crops. Scientists like Mary Claire King, who has focused her career on science in the public interest, offer an alternative to the increasing corporatization of the Academy, where technologies are funneled through technology transfer offices and patent regimes to biotechnology start-ups and industry. Socially committed science, rather than justifying its existence through an appeal to innovation or capital generation, instead offers the public good as its ultimate aim. Increasingly, this justification of public good has become normative as researchers are required to enumerate in funding applications the ways in which the knowledge they generate will reach and benefit the public.

The opposition to genetic technologies, even ones framed with explicit social justice goals, has fallen into three major arenas: the role of genetics in reinforcing biological notions of race and practices of racism, biological reductionism, and the corporate control of life. The sophisticated and insightful critiques of postconflict DNA identification, DNA-led exonerations, breast cancer genetic testing, and golden rice all point to the ways in which these technologies resonate with broader ideologies about genetics and human identity, meanings of family, life and ownership, and race and racism.

Race and genetics have a long and troubled history. The use of race as a biological category to explain patterns of human difference was a standard practice of nineteenth century science, within a broader political context of human exploitation and colonialism. With the sequencing of the human genome, ideas about race have been publicly eschewed in favor

of the concept of population. Despite the scientific consensus around the biological nonexistence of race, technologies have continued to rely on race as a proxy for ideas about group differences, which are further reinforced by the pervasive racial ideologies in American society, where much of the technology is developed. In criminal justice and medical settings, in particular, race becomes the default social category for genetic technologies. Statistical probabilities of genetic matches and the relative risk of disease are broken down by racial grouping. Developers and users of these technologies see race as a helpful proxy, while critics have clearly shown that this continued reinscription of race as a meaningful biological category in everyday interactions with genetics, reinforces ideas about inherent biological differences between social groupings (Roberts, 2012). An unintended consequence of genetic technologies aimed at social justice that embed race in their form and function can be the hardening of racist ideologies about groups already stigmatized and oppressed. This has been an important critique of initiatives around human identification like kinship analysis in postconflict settings and DNA-led exonerations. BRCA testing has also unintentionally reinforced racialized health disparities, with current tests remaining much more effective and available for white women than for women of color despite deep differentials in cancer-related deaths (Hall and Olopade, 2006).

Closely tied to the biological hardening of racial ideologies has been the critique of genetic technologies as reinforcing biological reductionism, the use of biology as the primary and fundamental explanatory framework for human behavior and social relations. This move toward biological explanation leaves out the role of social life and by extension power and inequality in human outcomes. Vandana Shiva's trenchant critique of golden rice illuminates how Vitamin A deficiency is not driven by rice or cultural practices of rice consumption but by poverty which forces entire segments of society to subsist on rice alone. Rather than modify rice, she argues, how might we modify society to foster farming and land tenure practices that restore traditional cultivars thereby addressing inequality-driven hunger. By reducing vitamin A deficiency to a biological problem, the answer becomes technological: improve rice. But when the broader social picture is restored, it becomes clear that the problems underlying nutrient deficiency are both biological and social and solutions must take both into account (Shiva, 2001). This same critique is important in cases of postconflict DNA identification when the widespread use of kinship analysis, which relies on biological notions of family and relatedness, interacts with diverse social practices of family formation. The technology, unintentionally through its form, reinforces narrow ideas of family based in heterosexual reproduction and biological relatedness. This leads to misunderstandings between scientists and victims in postconflict settings when individuals, which scientists understand as nonrelated, repeatedly attempt to donate DNA samples for identificatory purposes; it also points to the vibrant and varied human practices of love, obligation, and belonging that must be understood to successfully address the legacies of violent conflict.

One of the most visible and successful critiques of genetic technologies has been around the corporate control of life, specifically through patenting regimes. The environmental and antiglobalization movement have joined together in their resistance to intellectual property regimes. This critique understands the problem of biotechnology to be one of the ownership of knowledge and by extension the corporate control of the biodiversity of life. Golden rice has been a key player in these debates. Corporate sponsors highlight its potential as a public good emerging from a successful and benevolent intellectual property regime. Critics argue that a public good like rice should not fall into the realm of benevolent charity but instead remain part of a fundamental right to access the global commons of biodiversity. Genetic testing for BRCA genes has played a historic role in this fight, becoming in and of itself a rallying cry for social justice. The American Civil Liberties Union (ACLU) along with many partner organizations challenged the right to patent genes. The case focused on Myriad Genetics, which held the patent on the BRCA genes and was the sole purveyor of testing. In 2013, the United States Supreme Court ruled that human genes, in the form of a naturally occurring sequences, could not be patented, reversing years of biotechnology practices of patenting genes as soon as they were sequenced (*Association for Molecular Pathology v. Myriad*, 2013). This ruling was heralded as an important social justice accomplishment in the realm of genetics, one which enshrined the human genome as a public good.

The Governance of Genetic Technologies for Social Justice

The governance of technology focuses directly on the issue of power and politics within knowledge production. Who decides what technologies should be developed and how they should be implemented and regulated? Formally this often falls on the shoulders of governmental regulatory bodies but as in the case of the BRCA patents, courts of law also play central role as a space for the public and other interested parties to contest scientific technologies and practices and their impact on social life. In cases where urgent social justice ends are at stake, like the identification of missing children in Argentina, technologies are often developed and implemented under a logic of crisis, which can trump ethical deliberation. The controversies social justice genetic technologies have generated suggest that to be successful in their aims, geneticists must move beyond crisis thinking to deliberate with the public about how to best address social harms and avoid the pitfalls of racialization, biological reductionism, and privatization of genetic resources. Looking at the case studies together, an important theme in social justice genetics has been the public ownership and management of technologies. For example, the Abuelas of the Plaza de Mayo worked in collaboration with scientists to develop the Index of Grandpaternity to specifically address the brutal nature of disappearance in Argentina. Women directly affected by genocide decided the needs and parameters of genetic intervention and subsequently directed the use of the technology: who would be tested, in what ways, and to what ends. In contrast, the genetic technology of golden rice, although designed to address perceived problems of poverty, was not integrated with local communities' needs and approaches to hunger alleviation. Genetics, if it is to be a tool for social justice, must be practiced with a politics of humility (Jasanoff, 2007) that understands genetic technologies as

inherently political, as partial and socially embedded, and only effectively implemented and governed through a robust and equitable collaboration between science and society.

See also: Eugenics, History of; Forensic Genetic Databases: Ethical and Social Dimensions; Genetic Screening for Disease; Genetics and Society; Human Rights, Anthropology of; Kinship in Anthropology; Social Justice and Social Inequalities; War Crimes Tribunals, Crimes against Humanity and Transitional Justice; Wrongful Convictions.

Bibliography

Arditti, R., 1999. Searching for Life: The Grandmothers of the Plaza de Mayo and the Disappeared Children of Argentina. University of California Press, Berkeley.

Aronson, J.D., 2007. Genetic Witness: Science, Law, and Controversy in the Making of DNA Profiling. Rutgers University Press.

Association for Molecular Pathology v. Myriad, 2013. S. Ct.

Buck v. Bell, 1927. S. Ct.

Capeheart, P.L., Milovanovic, P.D., 2007. Social Justice: Theories, Issues, and Movements. Rutgers University Press, New Brunswick, NJ.

CONADEP, 2003. Nunca mas: informe de la Comision Nacional sobre la Desaparicion de Personas, sixth ed. EUDEBA, Buenos Aires.

Gillham, N.W., 2001. A Life of Sir Francis Galton: From African Exploration to the Birth of Eugenics, first ed. Oxford University Press, New York.

Hall, M.J., Olopade, O.I., 2006. Disparities in genetic testing: thinking outside the BRCA box. Journal of Clinical Oncology 24, 2197–2203.

Jasanoff, S., 2007. Technologies of humility. Nature 450, 33.

Jolie, A., 2013. My medical choice. New York Times, A25.

Lewontin, R.C., 2001. The Triple Helix: Gene, Organism, and Environment. Harvard University Press, Cambridge.

Parthasarathy, S., 2012. Building Genetic Medicine: Breast Cancer, Technology, and the Comparative Politics of Health Care. MIT Press Books, Cambridge.

Potrykus, I., 2001. The "golden rice" tale. Vitro Cellular & Devolpment Biology – Plant 37, 93–100.

Roberts, D., 2012. Fatal Invention: How Science, Politics, and Big Business Re-create Race in the Twenty-First Century. The New Press, New York.

Rudnik, J., 2008. In the Family. Kartemquin Films.

Shiva, V., 2001. Special report: golden rice and neem: biopatents and the appropriation of women's environmental knowledge. Women's Studies Quarterly 29, 12–23.

Stone, G.D., 2010. The anthropology of genetically modified crops. Annual Review of Anthropology 39, 381–400.

Relevant Websites

https://www.aclu.org/free-speech/brca-faqs – ACLU website for the supreme court challenge of BRCA patents.

http://www.aaas.org/page/srhrl-human-rights-activities – American Association for the Advancement of Science activities on human rights and science.

http://www.goldenrice.org/ – A golden rice advocacy website.

http://www.councilforresponsiblegenetics.org/ – Website of the council for responsible genetics, which advocates for public deliberation and governance of new genetic technologies.

http://inthefamily.kartemquin.com/ – Website of the film "In the Family" chronicling Joanna Rudnik's experience with BRCA testing.

http://www.abuelas.org.ar – Website of the Grandmothers of the Plaza de Mayo.

http://www.innocenceproject.org – Website of the innocence project dedicated to using genetics to contest wrongful convictions.

Genetics and Society

Peter Glasner, Cardiff University, Cardiff, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Following the breakthrough of the development of recombinant DNA in the 1970s, scientific research in genetics has grown exponentially. Social science, and in particular sociology, has taken a little longer to catch up following closely on the heels of ethics and philosophy. Strands from medical anthropology and sociology, law, environmental issues, science policy studies, and the public understanding of science have together formed a bricolage of approaches that underpin a discussion of the relationship between genetics and society. It follows with a discussion of a selection of both empirical and conceptual advances in genomics and stem cell technology, before concluding by indicating some future directions in which social science is heading. Finally there is a short appendix with online links to several useful Web sites.

Introduction and Intellectual Antecedents

The relationship between genetics and society has proved to be a complex one since the link between Herbert Spencer's Social Darwinism and the rediscovery of 'classical' (transmission) genetics in the work of Gregor Mendel was made at the end of the nineteenth century. The early decades of the twentieth century saw the applications of genetic research as relatively uncontentious. Much of the focus was on raising living standards by improving livestock strains and developing food production through more efficient plant breeding (Bud, 1993). Societal concerns focused on the growing eugenics movement with its aim, made explicitly by Francis Galton who coined the term in 1883, to improve the social condition by using similar methods. Both can be seen as responses to the rapid and unimagined consequences of the process of industrialization and urbanization overtaking many western societies at that time. As Koch (2009, p. 438) points out, linking science and social policy to move from natural to artificial selection was seen as a progressive way to enhance the quality of the citizenry. She also makes it clear, however, that eugenics is by no means a unitary concept, and its meanings are historically and culturally grounded. Much of the same can now be said for the concept of the gene (Keller, 2000, p. 67). The rise of Nazi Germany, the War, and the Holocaust all contributed to muting many of the advocates of eugenics in many of its forms (Kevles, 1992). The immediate postwar period also saw the molecularization of the concept of the gene culminating in the discovery of the double helical structure of DNA in 1953 (Watson and Crick, 1953).

In the 1960s and 1970s, several groups of social commentators and concerned scientists began to publicize the potential dangers of a wide range of scientific and technological advances, including most notably nuclear and biological weapons (Maddox, 1972). The ideological debate about the status and role of science and technology in contemporary society gathered momentum with the advent of antinuclear groups such as the Campaign for Nuclear Disarmament, environmentalist movements led by Rachel Carson's *The Silent Spring* (1962), and anti-Vietnam war protestors (Krimsky, 1982; Rose and Rose, 1976). Together these social movements formed the backdrop to the discovery of a key moment in the relationship between genetics and society: genetic manipulation. Popular culture has a long history of describing the intended and often unintended consequences of directing nature according to the whims of individuals or social groups with noteworthy contributions by among others, Mary Shelley, HG Wells, and Aldous Huxley (Turney, 1998). Now for the first time it seemed fiction might be turned into fact with the first publication using the new technique of recombining DNA (in *Escherichia coli*) appearing in 1972 (Jackson et al., 1972). However, the wider public expressed great concern about the dangers of the new technique, and even the scientists were not immune to the reservations being expressed within society at large. In 1973 a group published a letter in *Science*, the journal of the American Association for the Advancement of Science, asking for the dangers surrounding the creation of new hybrid viruses with unpredictable biological properties to be formally investigated. Following almost a year of investigation a second letter calling for a moratorium on certain kinds of rDNA experiments was published marking one of the very few occasions in the history of science when scientists themselves have called for their research to be stopped until appropriate restraints have been put in place (Krimsky, 1982; Bennett et al., 1986). The practical outcome of these events was the establishment of rigorous guidelines around the world that enabled researchers to continue with their experiments. This highly public controversy also contributed to the continuing erosion of the legitimacy of the institution of science within society. It also strengthened interest among social scientists in governance and regulatory issues (Wright, 1994). But by the late 1970s, regulatory oversight had become reduced to the management of risk in the laboratory (Gottweis, 1998).

In the early 1980s, leading scientists in the field of genetics began to discuss the possibility of mapping and sequencing all of the human genome, conceived through such metaphors as the Holy Grail, the 'moon shot' of genetics, and most significantly, the book of life. Much of the impetus came from the need to develop better diagnostic and therapeutic tools in the war on cancer, initiated by President Richard Nixon in 1971, with a view to eradicating it by 2015. In 1977, Walter Gilbert in USA and Frederick Sanger in the UK independently developed the gene sequencing methodology which, along with the technique for copying gene sequences (polymerase chain reaction (PCR)), underpinned the Human Genome Project (HGP). PCR was developed by Kary Mullis and the Cetus

Corporation in 1983 and speeded up the sequencing process (Rabinow, 1996a). By 1990 the HGP was officially launched, though some critics argued that straightforward curiosity-driven research was more important, leading Walter Gilbert to assert that possession of the genetic map and DNA sequence of a human being would transform medicine (Gilbert, 1992). The map of the human genome, it was argued, may also provide answers to such questions as the existence of a gene for alcoholism, aggression, or homosexuality, and perhaps even explain the biological basis for intelligence that eluded Francis Galton and the eugenicists. In this way, the biomedical and political systems came together to largely define health and disease, normality and abnormality, presenting genes as causal factors within a reductionist framework (Lippman, 1992). In February 2001, the international, publicly funded Human Genome Sequencing Consortium (in *Nature*) and Craig Venter of the private company Celera (in *Science*) published the first drafts of the full sequence of the human genome.

For social scientists, the HGP provided rich opportunities for investigation and analysis using resources developed mainly in the social studies of science, including social worlds analysis, social constructivism, and actor–network theory. In order for the appropriate funding for the HGP to be accomplished by science, government, and commerce, both in USA and elsewhere, the need to align their different social worlds had to be recognized. This was an example of heterogeneous engineering (Law, 1987) bringing together researchers, techniques, organizations, laboratories, databases, biological materials, funding sources, and political support to the task at hand (Hilgartner, 1995). The HGP became a boundary object – something that serves as an interface between communities of practice (Star and Griesemer, 1989) – allowing the techno-scientific work of mapping and sequencing to proceed by enrolling those from different social worlds, with their differing views about its scientific merit, to agree on a common goal. Much of the work of enrolling allies and translating their interests (Callon, 1986; Latour, 1987) was in the hands of leading spokespersons like Walter Gilbert in USA and Walter Bodmer in the UK, who achieved substantial funding at a time of financial stringency by constructing a doable problem (Fujimura, 1987; Balmer, 1996). The publication of the first maps of the human genome constituted the preliminary attempts to stabilize this technological ensemble of social and techno-scientific elements (Bijker and Law, 1992). It led to genetics moving out of the laboratory and becoming a big science enterprise (Glasner, 2002; Martin, 2001), while still leaving its development and applications in a state of flux.

Science was beginning to lose its authority, and expertise was becoming contested as risks, however real, became recognized as socially located. Experimentation appeared to take place in a global laboratory rather than on someone's laboratory bench (Beck, 1992; Giddens, 1990). A wide range of actors became increasingly involved in co-constructing (Jasanoff, 2004) the trajectory of the emerging technoscience of genomics. Several controversies surrounding the genetic basis of human behavior were revealed to be motivated more by ideology than by science. Lippman (1991) coined the term 'geneticization' to describe how differences between individuals including disorders, behaviors, and psychological variations are reduced to their DNA codes. One example surrounded the publication in 1994 of *The Bell Curve* (Herrnstein and Murray, 1994), which linked intelligence to the class structure in USA. It also implied that African Americans were not as intelligent as their neighbors, adding fuel to the fire of the controversy over race in USA. While presented as such, this was not a work of genetics per se but a secondary analysis of a number of intelligence test studies and social surveys. It revealed a bell curve of IQ, and linked normal variation to such factors as poverty, employment, illegitimacy, and welfare recipiency, mistaking correlation for causality. Social scientists had become involved in the controversy that the book generated, and in related controversies including the genetic basis of race and crime (Carson and Rothstein, 1999).

Within the social studies of science, the study of controversies (Collins, 1981; Nelkin, 1995) developed alongside the analysis of what became known as the crisis of legitimation and the rise of the risk society in late modernity (Habermas, 1976; Beck, 1992). These were often the focus of media attention, with landmarks in the development of genetics taking center stage, from the birth of the first test-tube baby in 1978 to the cloning of Dolly the sheep in 1997. The discovery of the structure of DNA in 1953 was not simply a scientific event, but one that opened new cultural understandings in ways that had previously belonged to the world of fiction (Haran et al., 2008). The image of the double helix became ubiquitous and appeared in many aspects of modern life including advertisements for motorcars and perfumes. It acquired the status of a cultural icon (Nelkin and Lindee, 1995).

In this context, the public understanding of science had become a matter of concern to senior scientists involved in shaping government policies (Royal Society, 1985). In the mid-1980s, as the scientific establishment sought to regain its authority, social science analysis drew on a wide range of disciplines including medical sociology, women's studies, and environmental studies to better understand the processes involved (Wynne, 1995). Underpinning the Royal Society's approach was the widely held view that the main reason for the increasing public concerns about the direction that scientific advance was taking, particularly in the public understanding of genetics, lay in public ignorance about the science itself. This 'deficit model' of public understanding attracted a variety of analyses by social scientists. Wynne and others pointed to the need to better understand the contextualized nature of how the various publics related to scientific knowledge (Irwin and Wynne, 1996) and the importance of recognizing the extent of lay expertise. Rather like the deficit model of public understanding of science, Lippman's geneticization thesis had depended more on the agency of the health practitioners than on the patients and their families (Gibbon, 2002). In contrast, studies were beginning to show that lay people were often knowledgeable about key issues that affected them directly. For example, the ways in which genetic services are used were found to be mediated by differing perceptions of the nature of hereditary conditions within families, and previous experiences of medical interventions (Parsons and Atkinson 1992, Lambert and Rose, 1996). The uptake of prenatal diagnosis provides another example of how, through the narratives of pregnant women, consumers of medical technology become lay experts (Rapp, 2000). Thus, a better understanding of the how lay people are expert in, rather than ignorant of, the ways in which

genetics shapes their lives contributed to discrediting both the deficit model and the geneticization thesis as the completion of the mapping of the human genome approached (Michael, 1992; Kerr et al., 1998).

Current State of Knowledge and Theory

Two key moments inform this section: the completion of the mapping of the human genome to reveal the book of life; and the subsequent discovery that, while extremely valuable, the book itself ignores the importance of epigenetic and other factors in complicating its interpretation. The current state of social science research on genomics is briefly explored along with attempts to provide new conceptual approaches to the study of genomics. There is also a discussion concerning these developments in stem cell research.

The main difference between genetics and genomics is that genetics scrutinizes the functioning and composition of the single gene, whereas genomics addresses all genes and their interrelationships in order to identify their combined influence on the growth and development of the organism. The publication of a final version of the map of the human genome in 2003 and 2004 revealed that, rather than 100 000 genes, it consists of some 20–25 000 genes, significantly fewer than the puffer fish. It also became clear that the concept of the gene was becoming less useful in scientific research with questions raised about the demarcation between one gene and another (Nature, 2006). A greater emphasis was placed on the role of RNA and how cells function over time through the study of epigenesis, or how the interaction between behavioral and genetic factors causes an organism's genes to be expressed in different ways without altering the underlying gene sequence (Moss, 2004).

The promise of the HGP was to improve human health by moving from diagnosis, particularly through identifying single-gene disorders such as cystic fibrosis or Huntington disease, to intervention through gene therapy and pharmacogenomics. The development of genetically based therapies where a disease-causing gene is replaced by a healthy one has made little progress since the first clinical trial in 1990, and the death of Jesse Gelsinger who died of multiple organ failure due to a severe immune response during a trial for ornithine transcarbamylase deficiency (OTCD) in 1999. However, pharmacogenomics appears to have been more successful, though perhaps more through what is promised than through any impact on clinical practice (Hedgecoe, 2004). One of the most common biomedical applications of genetics has been in the area of human reproduction particularly through preimplantation genetic diagnosis where parents with a history of genetic disease can, within clinical settings, select healthy embryos. Such procedures are presented as clear choices for parents to make, but appear in reality to be much more ambiguous (Roberts and Franklin, 2006), and hide the extent to which the social practices of genetics can be said to be eugenic (Shakespeare, 1999). Research suggests that the best types of decisions are made in settings where ambiguity and uncertainty are openly acknowledged by all parties (Kerr, 2009).

The completion of the HGP has generated a vast sea of data raising issues of privacy of information in a variety of contexts including health, insurance and employment, DNA profiling, biobanking, and data sharing. In particular, there are issues relating to whether relatives of carriers of single-gene disorders should have access to this information, or should this information be passed on to third parties such as insurance companies, leading to genetic exceptionalism and discrimination (van Hoyweghen, 2010; Rothstein and Joly, 2009). In the criminal justice system, there has tended to be a reliance on the unassailability and apparent infallibility of DNA data to determine innocence or guilt through DNA profiling (Lynch et al., 2008). The collection of forensic DNA evidence in police-held databases has also been a cause for concern, with data from 0.5% of the population held in USA in CODIS, the Combined DNA Index System, and with 1% in the UK in the NDNAD, the National DNA Database (Nuffield Council, 2007). There are also issues of data sharing between database curators, stakeholders, and users. Often these challenge the competitive nature of research either through possible future commercial application, or, more conventionally, in the publish or perish world of academia. Some of these are addressed through bioinformatics and the ability to retrieve information via online databases. However, there continue to be problems in standardization and cooperation between very locally contingent practices in data generation, management, and regulation (Stein, 2008).

Another focus of concern has been the development of genetically modified organisms (GMOs) and the risks they might pose to the environment, including herbicide resistance, transfer to wild relatives, insect resistance, threat to biodiversity, new viruses and traits, and changes to land use, as well as to the food we eat. The introduction of bioengineered ingredients in processed foods in USA had happened largely without public comment or indeed knowledge (Beckwith et al., 2003). In Europe, and in particular in the UK with a history of food-related health scares, the public has appeared much more resistant (Grove-White et al., 1997). Jasanoff (1995) has argued that the permissiveness of the US response to the introduction of genetically modified (GM) crops is a result of having in place the successful model of an existing precautionary regime for controlling chemicals and other toxic substances. This regulatory model was narrowly prescribed through close agreement between government and industry. The European response, on the other hand, while based on the precautionary principle (Levidow, 2001), was less unitary as countries in the European Union, with very different cultures and histories, responded differentially to the regulatory framework (Levidow, 2009). The introduction of innovatory technologies is thus not a simple linear process dependent on the knowledge of scientific experts, but, according to Jasanoff (2005) a reflection of a society's civic epistemology: how scientific debates are politically framed; how what counts as science is established; and who are the key actors. In India, where GM was introduced in the 1990s, cotton farmers have banded together to forcibly reject transgenic cotton brought in by Monsanto, a US multinational company, burning the crops and exhibiting a civic epistemology rooted in mass local protest within a loosely configured regulatory framework.

Social scientists have developed a number of different ways to theorize the new developments in genomics including concepts such as biosociality and the somatic individual,

expectation and promise, and biovalue and the bioeconomy. Early studies had identified a link between individual genetically based pathologies and social problems (Lippman, 1992). The concept of biosociality (Rabinow, 1996b) was developed to counteract the reductionism of sociobiology by suggesting that new types of group identities and citizenships around shared traits were being created (Rose and Novas, 2004). By appearing at the intersections of between clinical medicine and biotechnology companies, these groups empower their members to influence research agendas rather than be passive recipients. They become ethical pioneers who, as somatic individuals, actively cocreate the interpretation and application of new genetic technologies (Rose, 2007). They create new economies of hope, producing new research agendas and commercial opportunities, thereby breaking with the negative social effects of the earlier eugenic model underpinning geneticization. Recent studies of genomic research using these concepts have been into embryonic stem cell therapy, breast cancer, autism, Parkinson's, and IVF treatment, as well as on the expectations and education surrounding genomic research. The results have covered themes such as novel modes of identity and identification, including genetic citizenship, the role of institutions, ranging from disease advocacy organizations and voluntary organizations to the state, the production of biological knowledge, novel life-forms, and technologies, and the generation of wealth and commercial interests in biology (Gibbon and Novas, 2007).

Genetic advocacy groups need to be involved in shaping future research agendas to fit their particular interests and do so by making clear their hopes and expectations for the directions in which these will progress. However, what constitutes a vision of the future is a contested terrain as many different social and material interests are involved. It can be conceptualized as an object in its own right (Brown et al., 2000). (Many commentators on the new biosciences use the term 'objects' as a shorthand for a wide range of novel biological, social, technological, and regulatory entities.) Futures are created anew as different rhetorical, organizational, and material resources are mobilized including fears of risk and failure. Visions, however, provide a necessary framework, co-constructed between actors and resources, within which decisions about the shape and application of a developing technology can be formulated. In this way, promissory visions have a significant performative value (Brown and Michael, 2003). What appear to be speculative claims about future therapeutic interventions made by pharmaceutical companies and genetic interest groups can be considered fundamental to the process of creating viable sociotechnical networks and make progress. These promissory visions facilitate the enrollment and alignment of key players such as industry (seeking new profits), clinicians (new techniques for treatment), patients (new hope for care), and regulators (new governance frameworks) in establishing a new genetically based medical technology (Hedgecoe and Martin, 2003). As a result, expectations become embedded in novel practices, institutions, representations and inscriptions, and ethics and discourse. In addition to studying genetic advocacy groups (Novas, 2007), studies using this approach have focused on cord blood stem cell banking. Parents invest in storing tissue for possible future use should it be required by their offspring who in the parents' minds are going to be at future risk. The expectation envisaged by these parents is one in which techno-scientific innovation, resulting in therapeutic intervention, and embodied in deposited cord blood samples, becomes technologically feasible (Brown and Kraft, 2006; Martin et al., 2008). In turn interest groups facilitate and encourage their development (Franklin and Lock, 2003).

The attempt through promissory regenerative medicine to add value to life itself suggests that markets exist within which such added value attracts an economic price (Glasner, 2005). Such markets are to be found in global economies related to those of tissues and genetic material. New biological objects with their associated technologies produce genetic capital or biovalue, based on the surplus value that is produced from their extraction and manipulation. Vitality is manufactured in the laboratory (Waldby, 2002). There is a global knowledge economy relating to the flow of ideas and projects, with the publicly funded HGP providing an impressive illustration. Nation-states and global companies outsource their research activities, for example in India through the growth of clinical trials or therapies derived from stem cells, where poverty and ignorance provide willing subjects and regulatory governance is still in a fluid state (Rajan, 2006). Such trials are mainly run by US-based pharmaceutical companies with their Indian partners and are closely colocated with the twentieth century neoliberal capitalist modes of production (Cooper, 2008).

The fueling of bioeconomies through permissive regulation and governance frameworks in areas such as oocyte and embryo donation, and clinical and laboratory applications of stem cell research has a less clearly delineated trajectory than in the more tightly regulated European and US frameworks (Gottweis et al., 2009). The motives of nation-states in encouraging such research may vary widely, sometimes forcing leading scientists to fraudulently announce advances in therapeutic applications. The Korean scientist Hwang Woo-Suk, for example, achieved cult status associated with euphoric expectations of mass healing before his fraud was discovered (Kim, 2008). There are also questions about who benefits and how different public needs are represented, so in India the beneficiaries of stem cell research (as with large-scale clinical trials) are normally to be found elsewhere. However, in some cases, maverick scientists like Geeta Shroff exploit regulatory inconsistencies only for their own patients, and refuse to share their work with the wider scientific community (Bharadwaj and Glasner, 2009). Studies have shown that understanding the practice of how research regulations operate on the ground requires an intimate knowledge of existing modes of social control, the role of the media, and the functioning of local, provincial, and national authorities (see Sleebohm-Faulkner, 2011).

Stem cell banking has been seen as one way in which to ensure that appropriate governance oversight is in place and divert attention away from the more emotive debates that surround stem cell technologies. While a number of private banks are now in operation worldwide, they are often specialist in nature, curating cell lines from adult or cord blood tissue. The first publicly funded bank, the UK Stem Cell Bank, officially opened in 2004 for the deposit of regulated human embryonic stem cell (hESC) lines from approved laboratories, and made the first of these available for wider distribution in 2006. It accepts stem cells, tests them, grows

them, and then distributes them internationally. Through the progressive refinement and purification of persons and materials, the Bank can be seen as a moral refinery providing a visible location for the international legitimacy of hESC science (Stephens et al., 2008).

There is a history of public debates about the ethics and morals of stem cell research using in particular hESCs that has paralleled the discourse of hype and hope concerning future therapeutic applications (Prainsack et al., 2008). This debate, mainly in USA and Europe, has diverted attention from, on the one hand, existing uses of adult stem cells in (for example) bone marrow transplants successful for over 20 years, and on the other, more commercial applications, in (for example) the cosmetic industry. There are two kinds of DNA-based cosmetics. The first uses DNA research to come up with new kinds of ingredients to add to cosmetics, while the second promises to use your genetic profile to create a makeup just for you. Less high profile is the regulatory, manufacturing, and distribution infrastructure that is rapidly being put in place to enable the delivery of near-market human biological products. Here the controversies take a different form, asking questions about when, rather than if, they should come to market, and focusing on how, and by whom, products are considered to be safe (Franklin and Kaftantzi, 2008; Faulkner, 2009).

Some Future Directions

The science of genetics may already, as many commentators suggest, have left the unitary concept of the gene and its linear development behind. The heyday of injecting huge sums of public funding into genomics research is substantially over, being replaced by an emphasis on knowledge translation through privately or commercially funded initiatives. In doing so it has opened a variety of societal avenues of concern, hype and hope including personalized medicine, bioprospecting, biosustainability, human/species enhancement, and bio-weapons development. It has also put into question the accepted boundaries between human/animal, organic/nonorganic, and living/nonliving through the creation of new bio-objects. These include stem cells, trans-species animals, GM crops, or tissues that are cut from human bodies and stored in tissue repositories or biobanks. They are fluid and mobile, allowing discoveries in medicine, for example, to be applied in nutrigenomics and food production. They also challenge current modes of governance through multiple (or even contrasting) meanings as they circulate through different sectors of society. In this way, postgenomics – i.e., the study of the function of DNA and other molecules in the human organism and their interaction with environmental factors – contributes to and simultaneously generates novel social relations (Parry and Dupre, 2011; Vermeulen et al., 2011).

Synthetic biology provides one of the clearest attempts to generate new bio-objects as it attempts the design and construction of new biological parts, devices, and systems, and the redesign of existing natural biological systems for more useful purposes. The greatest success to date comes from the J Craig Venter Institute where, in 2012, a bacterial chromosome was created and was successfully transferred into a bacterium, replacing its existing DNA. The microbial cell then began replicating and producing a new set of proteins (Gibson et al., 2012). However, this was an expensive and time-consuming experiment costing US$40 million and 200 person years, leaving a long gap between its discovery and any future practical or scaled up applications such as artificial biofuels or novel pharmaceuticals. It also rules out immediate uses in warfare or terrorism, although the ethical and governance issues still need to be addressed. Should synthetic biology live up to its hype, it has the potential to shift the boundaries between the natural and the artificial in significant ways.

Another way in which genomics is contributing to making conventional social boundaries more fluid relates to the accumulation, sharing, mining, and integration of data. Following the completion of the HGP, and the development of faster and cheaper ways to sequence and analyze genetic material, researchers are drowning in sea of data. A significant concern has been the archival digital storage of this information with its attendant problems of damage and atrophy. A team from Harvard University has married next-generation sequencing technology with a novel strategy to encode 1000 times the largest amount of data previously stored in DNA (Church et al., 2012). They estimate that it would take only 4 g of DNA to store all human knowledge collected in 1 year, and do so in such a way that, even if lying dormant for millions of years, it would be as accessible then as now. While the cost of DNA-handling tools makes this storage option uncompetitive as yet, and the process is still a slow one, its wider impact has yet to be investigated while the principle has now clearly been established.

The issues of privacy and ownership have become more pressing as the data generated through genetics accumulate exponentially. Historically, the famous case of *John Moore v. Regents of the University of California* in 1990 highlighted the tensions surrounding property rights and profits. Genetic material derived from Moore's spleen (removed as treatment for leukemia) became, without his knowledge or consent, the source of a commercial cell line. Many biomedical research projects are dependent on tissues and data collected from participants and their families. Linking these large, detailed, and heterogeneous data sets enables researchers to generate statistically significant results for use by population geneticists. However, individuals agreeing to participate in biomedical research are normally required to provide informed and voluntary consent for the use of their contribution in these ways. In biobanking, which is now a worldwide feature of postgenomic science, it has proved to be a potentially contentious issue even though its ostensive focus is on the prevention and treatment of disease (Kaye and Stranger, 2009).

As an example of a bio-object, however, the data stored can also be used as a means of reducing costs in health care provision, speeding up drug development and generating commercial opportunities for pharmaceutical companies, and producing personalized therapies for certain disease groups. One by-product of this accessible data is found in Genome Wide Association Studies (GWAS) that compare the genomes of healthy people with the hope of locating genetic variation responsible for common diseases. Such studies are not without their critics as it is clear that GWAS provide data for only a part

of the genetic basis of the human trait. That is, as the US National Institutes of Health (NIH) define them, GWAS only focus on the common genetic factors in populations that influence health and disease. However, human traits consist of more than just genes. Some elements are acquired and we now know could affect how genes themselves are expressed. As a result GWAS are useful only for population-based health studies, and not applicable to individual diagnosis or treatment.

However, GWAS have since 2006 given rise to successful direct-to-consumer genetic testing companies such as Navigenics and 23andMe that are comparatively unregulated and easily accessible via the World Wide Web, raising serious concerns about how informed the users really are (Curnette and Testa, 2012). Biobanks are funded by an equally diverse range of actors ranging from publicly funded hospitals and universities, through pharmaceutical biotechnology and genomics companies, to single-disease advocacy groups. Together these make informed consent in the postgenomic age difficult to manage (Corrigan and Tutton, 2009). However, the generative aspects of bio-objects point to one possible solution through the use of social media to develop participant-centered initiatives in informed consent as the basis for long-term interactive partnerships between patients and researchers or clinicians (Kaye et al., 2012).

The post-HGP landscape of genomics following the demise of the central dogma – DNA makes RNA makes protein – has resulted in a variety of new ways of viewing life, including proteomics, metabolomics, epigenomics, and stem cell biology. The cell rather than the gene is becoming the central actor in biomedical and biotechnical settings once more. In tissue engineering and regenerative medicine, a multidisciplinary field involving biology, medicine, and engineering, new bio-objects abound. The foundation of tissue engineering/regenerative medicine is the ability to exploit living cells in a variety of ways. There are potential therapeutic applications, where the tissue is either grown in a patient or outside the patient and transplanted, restoring, maintaining, or enhancing tissue and organ function. Alternatively, tissue engineering can have diagnostic uses where the tissue is made *in vitro* and used for testing drug metabolism and uptake, toxicity, and pathogenicity. Much of this work is carried out with stem cells data garnered from animal models and applied in areas such as bone tissue, skin, and heart muscle, providing alternatives to organ transplantation (Atala, 2009).

However, studies of how research results are translated from bench to bedside are challenging the accepted rhetoric of a straightforward exchange of new knowledge between the two. The new states of social relations that co-construct the bio-objects of tissue engineering include not only a number of intermediary stages, but also significant differences between those working within them. Between groups, bench scientists differ from clinical practitioners, and proponents of animal models differ from cell biologists, while within groups, there are differing approaches between mouse modelers and primate modelers, or creators of dish cell lines versus those who insist on genetically engineered cell lines. These differences can be analyzed using such concepts as rhetorics of the future, cultural capital and boundary work, to understand the struggle for power between science and medicine in this emerging field (Wainwright et al., 2009).

Conclusion

An international consortium led by the National Human Genome Research Institute (NHGRI) in Bethesda, USA announced the successful completion of the first phase of the Encyclopaedia of DNA Elements (ENCODE) in September 2012. Made up of 32 groups and 440 scientists, it was established in 2003, with a pilot phase completed in 2007. It was the first systematic attempt to work out what the DNA outside protein-coding genes does. At the cost of around US$180 million, researchers have found that 80% of the elements of a human genome sequence have an identifiable function. They have mapped 4 million switches in what was once thought to be junk DNA, and expect that many will help them better understand a range of common human diseases, from diabetes to heart disease, that depend on the complex interaction of hundreds of genes and their associated regulatory elements.

ENCODE research should help guide the hunt for problem areas that lie in the DNA sequence rather than, more conventionally, looking for errors within the genes themselves (ENCODE Consortium, 2012). However, some critics have suggested that the project has been flawed from the start as it uses 50-year-old gene sequences that are known to have significantly altered in makeup. Others have argued that the money would have been better spent on curiosity-oriented and hypothesis-driven research into specific problems, rather than on such an expensive and ultimately open-ended project. ENCODE researchers have responded by making a case for both approaches to complement each other (Maher, 2012). After 40 years, even with a myriad of new discoveries stemming from it, the 1980s debates about whether or not to pursue the HGP have, in the second decade of the twenty-first century, come full circle.

Links

A number of research centers around the globe provide the basis for a growing network of researchers in genetics and society. These together with key genomics Web sites, journals, and book series provide the reader with introductory links to this still-developing subfield.

Centers and Networks

Centre for Biomedicine and Society, Brunel, UK [www.brunel.ac.uk/sss/research/research-centres].

Centre for Society and Genomics, Nijmegen, The Netherlands [www.society-lifesciences.nl/en/home.html].

Centre for Genetics and Society, Berkeley, CA, USA [www.geneticsandsociety.org]

ESRC Genomics Network, UK [www.genomicsnetwork.ac.uk] (this includes three centers and a forum across the UK).

Genome Canada, Ottawa, Canada [www. genomecanada.ca].

Genethon, Evry, France [www.genethon.fr].

National Human Genome Research Institute, Bethesda, USA [www.genome.gov].

The Wellcome Trust, London, UK [www.wellcome.ac.uk].

Journals

BioSocieties: An interdisciplinary journal for the social studies of the life sciences [www.jounals.cambridge.org/bio].

New Genetics and Society. An international platform for informed discussion of key issues raised by the exploitation of the new genetics [www.ingentaconnect.com/content/routledg/cngs].

Science as Culture. Places science within the wider debate on the values that constitute culture; it is not the journal for a particular academic discipline [www.ingentaconnect.com/content/routledg/csac].

Science, Technology & Human Values. An interdisciplinary journal containing research, analyses, and commentary on the development and dynamics of science and technology, including their relationship with politics, society, and culture [sth, sagepub.com].

Social Studies of Science. An international peer-reviewed journal that encourages submissions of original research on science, technology, and medicine [sss.sagepub.com].

Sociology of Health and Illness. An international journal that publishes sociological articles on all aspects of health, illness, medicine, and health care [www.blackwellpublishing.com/shil_enhanced].

Book Series

Cambridge Studies in Genetics and the Life Sciences (Cambridge University Press)

Promotes interdisciplinary works in the social sciences that focus on the social shaping, social meaning, and social implications of recent developments in the life sciences. Interdisciplinary projects will be encouraged with a strong empirical and ethnographic base that explores topics within the context of an internationalized world of science as well as recognizing national variations in ethical, moral, and legal codes and mores [www.cambridge.org/gb/knowledge/series/series_display/item3937957/?site_locale=en_GB].

Genetics and Society Book Series (Routledge)

This series provides an outlet for outstanding scholarship in the multiple fields of genetics and genomics social sciences research. Published with Routledge since 2006, the research monographs, handbooks, textbooks, and edited collections offer authoritative, cutting edge perspectives on issues covering the ethical, legal, social, economic, or political aspects of:

- tissue engineering, enhancement, and cloning
- genetic modification of foodstuffs and other organisms
- neuroscience and neuroethics
- genetic screening and testing
- stem cell research and reproductive technologies
- psychosocial aspects of medical genetics and gene therapy
- the social and ethical issues surrounding biomedical innovation
- public engagement and political discourse
- representations of genetics across the media and cultural spheres
- regulatory policy and governance of biomedical research and its human applications
- the sociology and anthropology of bioscience and biotechnology
- bioethics
- the economics of new biomedical technologies and their place in the 'knowledge economy'

[www.routledge.com/books/series/GANDS].

Health, Technology, and Society (Palgrave Macmillan)

Medicine, health care, and the wider social meaning and management of health are undergoing major changes. In part this reflects developments in science and technology, which enable new forms of diagnosis, treatment, and delivery of health care. It also reflects changes in the locus of care and the social management of health. Locating technical developments in wider socioeconomic and political processes, each book in the series discusses and critiques recent developments in health technologies in specific areas, drawing on a range of analyses provided by the social sciences. Some have a more theoretical, some a more applied focus but all draw on recent research by the authors. The series also looks toward the medium term in anticipating the likely configurations of health in advanced industrial society and does so comparatively, through exploring the globalization and internationalization of health [www.palgrave.com/products/series.aspx?s=HTE].

See also: Anthropology, Genomics, and Human Variation: National Roots; Biobanking: Ethical Issues; Biocitizenship; Bioethics in the Post-genomic Era; Bioethics: Genetics and Genomics; Direct-to-Consumer Personal Genetic Testing; Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute; Genetics and the Media; Genetics: Legal Aspects; Genetics: The New Genetics; Genomics, Ethical Issues In; Human Genome Project: History and Assessment; Human Genomics in Asia; Stem Cell Research and Cloning.

Bibliography

Atala, A., 2009. Engineering organs. Current Opinion in Biotechnology 20 (5), 575–592.

Beck, U., 1992. Risk Society: Toward a New Modernity. Sage, London.

Beckwith, J.A., Hadlock, T., Suffron, H., 2003. Public perceptions of plant biotechnology – a focus group study. New Genetics and Society 22 (2), 125–141.

Bennett, D., Glasner, P., Travis, D., 1986. The Politics of Uncertainty. Regulating Recombinant DNA Research in Britain. Routledge and Kegan Paul, London.

Bharadwaj, A., Glasner, P., 2009. Local Cells, Global Science. The Rise of Embryonic Stem Cell Research in India. Routledge, London.

Bijker, W.E., Law, J. (Eds.), 1992. Shaping Technology/Building Societies: Studies in Sociotechnical Change. MIT Press, Cambridge, MA.

Brown, N., Kraft, A., 2006. Blood Ties: banking the stem cell promise. Technology Analysis and Strategic Management 18 (3/4), 313–327.

Brown, N., Michael, M., 2003. A sociology of expectations: Retrospecting Prospects and prospecting Retrospects. Technology Analysis and Strategic Management 15 (1), 3–18.

Brown, N., Rappert, B., Webster, A. (Eds.), 2000. Contested Futures: A Sociology of Contested Technoscience. Ashgate, Aldershot.

Bud, R., 1993. The Uses of Life: A History of Biotechnology. Cambridge University Press, Cambridge.

Callon, M., 1986. Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St. Brieux Bay. In: Law, J. (Ed.), Power, Action and

Belief: A New Sociology of Knowledge? Routledge and Kegan Paul, London, pp. 196–229.
Carson, R.A., Rothstein, M.A., 1999. Behavioural Genetics: The Clash of Culture and Biology. Johns Hopkins University Press, Baltimore.
Church, G.M., Gao, Y., Kosun, S., 2012. Next-generation digital storage Information in DNA. Science. http://dx.doi.org/10.1126/science.1226355.
Collins, H.M. (Ed.), 1981. Knowledge and controversy: studies in modern natural science. Special Issue of Social Studies of Science, vol. 11, p. 1.
Cooper, M., 2008. Life as Surplus. University of Washington Press, Seattle.
Corrigan, O., Tutton, R., 2009. Biobanks and the challenges of governance, legitimacy and benefit. In: Atkinson, P., Glasner, P., Lock, M. (Eds.), Handbook of Genetics and Society. Routledge, London, pp. 302–318.
Curnette, M., Testa, G., 2012. Consuming genomes: scientific and social innovation in direct-to-consumer genetic testing. New Genetics and Society 31 (2), 159–181.
Durant, J., Hansen, A., Bauer, M., 1995. Public understanding and the new genetics. In: Marceau, T., Richards, M. (Eds.), The Troubled Helix: Social and Psychological Implications of the New Genetics. Cambridge University Press, Cambridge, pp. 235–248.
ENCODE Consortium, 2012. An integrated encyclopedia of DNA elements within the human genome. Nature 489, 57–74 (6 September).
Faulkner, A., 2009. Medical Technology into Healthcare and Society: A Sociology of Devices, Innovation and Governance. Palgrave Macmillan, Basingstoke.
Franklin, S., Kaftantzi, L., 2008. Industry in the Middle: Interview with Intercytex founder and CEO, Dr Paul Kemp. Science as Culture 17 (4), 449–462.
Franklin, S., Lock, M., 2003. Animation and cessation: remaking life and death. In: Franklin, S., M.Lock (Eds.), Remaking Life and Death. Toward an Anthropology of the Biosciences. School of American Research Press, Sante Fe, NM, pp. 3–22.
Fujimura, J.H., 1987. Constructing "do-able" problems in cancer research: Articulating alignment. Social Studies of Science 17, 257–293.
Gibbon, S., 2002. Re-examining geneticization: family trees in breast cancer genetics. Science as Culture 11 (4), 429–457.
Gibbon, S., Novas, C. (Eds.), 2007. Genetics and the Social Sciences: Making Biosociality. Taylor and Francis, London.
Gibson, D.G., et al., 2012. Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329 (5987), 52–56.
Giddens, A., 1990. The Consequences of Modernity. Polity Press, Cambridge.
Gilbert, W., 1992. A vision of the grail. In: Kevles, D.J., and Hood, L. (Eds.) The Code of Codes. Scientific and Social Issues in the Human Genome Project. Harvard University Press, Cambridge, MA, pp. 83–97.
Glasner, P., 2002. Beyond the Genome: reconstituting the new genetics. New Genetics and Society 21 (3), 167–177.
Glasner, P., 2005. Banking on Immortality? Exploring the stem cell Supply chain from embryo to therapeutic application. Current Sociology 53 (2), 355–366.
Gottweis, H., 1998. Governing Molecules. The Discursive Politics of Genetic Engineering in Europe and the United States. MIT Press, Cambridge, MA.
Gottweis, H., Salter, B., Waldby, C., 2009. The Global Politics of Human Stem Cell Science: Regenerative Medicine in Transition. Palgrave, London.
Grove-White, R., et al., 1997. Uncertain World: Genetically Modified Organisms. Food and Public Attitudes in Britain. CSEC, University of Lancaster, Lancaster.
Habermas, J., 1976. Legitimation Crisis. Heinemann Educational Books, London.
Haran, J., Kitzinger, J., McNeil, M., O'Riordan, K., 2008. Human Cloning in the Media: From Science Fiction to Science Practice. Routledge, London.
Hedgecoe, A., 2004. The Politics of Personalised Medicine – Pharmacogenetics in the Clinic. Cambridge University Press, Cambridge.
Hedgecoe, A., Martin, P., 2003. The drugs don't work. Expectations and the shaping of pharmacogenetics. Social Studies of Science 33 (3), 327–364.
Herrnstein, R.J., Murray, C., 1994. The Bell Curve: Intelligence and Class Structure in American Life. The Free Press, New York.
Hilgartner, S., 1995. The human genome project. In: Jasanoff, S., et al. (Eds.), Handbook of Science and Technology Studies. Sage, London, pp. 302–315.
van Hoyweghen, I., 2010. Taming the wild life of genes by law? Genes reconfiguring solidarity in private insurance. New Genetics and Society 29 (4), 431–455.
Irwin, A., Wynne, B. (Eds.), 1996. Misunderstanding Science? The Public Reconstruction of Science and Technology. Cambridge University Press, Cambridge.
Jackson, D., Symons, R., Berg, P., 1972. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: Circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. Proceedings of the National Academy of Sciences of the United States of America 69 (10), 2904–2909.
Jasanoff, S., 1995. Product, process or programme: three cultures and the regulation of biotechnology. In: Bauer, M. (Ed.), Resistance to New Technology: Nuclear Power, Information Technology and Biotechnology. Cambridge University Press, Cambridge, pp. 311–334.
Jasanoff, S. (Ed.), 2004. States of knowledge. The Co-production of Science and the Social Order. Routledge, London.
Jasanoff, S., 2005. Designs on Nature: Science and Democracy in Europe and the United States. Princeton University Press, Princeton, NJ.
Kaye, J., Stranger, M. (Eds.), 2009. Principles and Practice of Biobank Governance. Ashgate, Farnham.
Kaye, J., et al., May 2012. From patients to partners: participant-centric initiatives in biomedical research. Nature Reviews/Genetics 13, 371–376.
Keller, E. Fox, 2000. The Century of the Gene. Harvard University Press, Cambridge, MA.
Kerr, A., 2009. Reproductive genetics. From choice to ambivalence and back again. In: Atkinson, P., Glasner, P., Lock, M. (Eds.), Handbook of Genetics and Society. Routledge, London, pp. 59–75.
Kerr, A., Cunningham-Burley, S., Amos, A., 1998. The new genetics and health: mobilizing lay expertise. Public Understanding of Science 7, 41–60.
Kevles, D.J., 1992. Out of eugenics. In: Kevles, D.J., Hood, L. (Eds.), The Code of Codes. Scientific and Social Issues in the Human Genome Project. Harvard University Press, Cambridge MA, pp. 3–36.
Kim, B., 2008. Explaining the Hwang scandal: national scientific culture and its global relevance. Science As Culture 17 (4), 397–415.
Koch, L., 2009. Eugenics. In: Atkinson, P., Glasner, P., Lock, M. (Eds.), Handbook of Genetics and Society. Routledge, London, pp. 437–447.
Krimsky, S., 1982. Genetic Alchemy. The Social History of the Recombinant DNA Controversy. MIT Press, Cambridge, MA.
Lambert, H., Rose, H., 1996. Disembodied knowledge? Making sense of medical science. In: Irwin, A., Wynne, B. (Eds.), Misunderstanding Science? The Public Reconstruction of Science and Technology. Cambridge University Press, Cambridge, pp. 65–83.
Latour, B., 1987. Science in Action: How to Follow Scientists and Engineers through Society. Harvard University Press, Cambridge, MA.
Levidow, L., 2001. Precautionary uncertainty: regulating GM crops in Europe. Social Studies of Science 31 (6), 845–878.
Levidow, L., 2009. Making Europe unsafe for agbiotech. In: Atkinson, P., Glasner, P., Lock, M. (Eds.), Handbook of Genetics and Society. Routledge, London, pp. 110–126.
Lippman, A., 1991. Prenatal testing and screening: constructing needs and reinforcing inequities. American Journal of Law and Medicine 1, 15–50.
Lippman, A., 1992. Led (Astray) by Genetic Maps. The cartography of the human genome and health care. Social Science and Medicine 35 (12), 1469–1476.
Lynch, M., Cole, S., McNally, R., Jordan, K., 2008. Truth Machine: The Contentious History of DNA Fingerprinting. University of Chicago Press, Chicago, IL.
Maddox, J., 1972. The Doomsday Syndrome. Macmillan, London.
Maher, B., 2012. ENCODE: the human encyclopaedia. Nature 489, 46–48 (6 September).
Martin, P., 2001. Genetic governance: risks, oversight and regulation of genetic databases in the UK. New Genetics and Society 20 (2), 157–183.
Martin, P., Brown, N., Turner, A., 2008. Capitalizing hope: the commercial development of umbilical cord blood stem cell banking. New Genetics and Society 27 (2), 127–143.
Michael, M., 1992. Lay discourses of science: science-in-general, science-in-particular and self. Science, Technology and Human Values 17 (3), 313–333.
Moss, L., 2004. What Genes Can't Do. MIT Press, Cambridge, MA.
Nature, 2006. Coping with Complexity. Nature 441 (25 May), 383–384.
Nelkin, D., 1995. Science controversies: the dynamics of public Disputes in the Unites states. In: Jasanoff, S., et al. (Eds.), Handbook of Science and Technology Studies. Sage, London, pp. 444–456.
Nelkin, D., Lindee, S., 1995. The DNA Mystique: The Gene as a Cultural Icon. W.H. Freeman, New York.
Novas, C., 2007. Genetic advocacy groups, science and biovalue: creating political economies of hope. In: Atkinson, P., Glasner, P., Greenslade, H (Eds.), New Genetics, New Social Identities. Routledge, London, pp. 11–27.
Nuffield Council, 2007. The Forensic of Bioinformation: Ethical Issues. Nuffield Council on Bioethics, London.
Parry, S., Dupre, J. (Eds.), 2011. Nature after the Genome. Blackwell, Oxford.
Parsons, E., Atkinson, P., 1992. Lay constructions of genetic risk. Sociology of Health and Illness 14, 439–455.
Prainsack, B., Geesink, I., Franklin, S. (Eds.), 2008. Stem cell technologies 1998-2008: controversies and Silences. Special Issue of Science as Culture, vol. 17, pp. 351–484 (4).
Rabinow, P., 1996a. Making PCR. A Story of Biotechnology. Chicago University Press, Chicago, IL.
Rabinow, P., 1996b. Artificiality and enlightenment: from socio-biology to biosociality. In: Rabinow, P. (Ed.), Essays on the Anthropology of Reason. Princeton University Press, Princeton, NJ, pp. 91–111.

Rajan, K.S., 2006. Biocapital. The Constitution of Post-genomic Life. Duke University Press, Durham, NC.
Rapp, R., 2000. Testing Women, Testing the Fetus: The Social Impact of Amniocentesis in America. Routledge, London.
Roberts, C., Franklin, S., 2006. Born and Made: An Enthography of Pre-Implantation Genetic Diagnosis. Princeton University Press, Princeton, NJ.
Rose, N., 2007. The Politics of Life Itself. Princeton University Press, Princeton, NJ.
Rose, N., Novas, C., 2004. Biological citizenship. In: Ong, A., Collier, S.J. (Eds.), Global Assemblages: Technology, Politics and Ethics as Anthropological Problems. Blackwell, New York, pp. 439–463.
Rose, H., Rose, S., 1976. The Radicalisation of science. In: Rose, H., Rose, S. (Eds.), The Radicalisation of Science. Macmillan, London.
Rothstein, M.A., Joly, Y., 2009. Genetic information and insurance underwriting. In: Atkinson, P., Glasner, P., Lock, M. (Eds.), Handbook of Genetics and Society. Routledge, London, pp. 127–144.
Royal Society, 1985. The Public Understanding of Science. The Royal Society, London.
Sleebohm-Faulkner, M. (Ed.), 2011. Stem Cell Research in Asia: Looking beyond Regulatory Exteriors. New Genetics and Society, (2), 137–207. 30 (3), 213–288 (special issues).
Star, S.L., Griesemer, J., 1989. Institutional ecology, "translations" and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–1939. Social Studies of Science 19, 387–420.
Stein, L.D., 2008. Towards a cyberinfrastructure for biological sciences: progress, visions and challenges. Nature Genetics 9, 678–688.
Stephens, N., Atkinson, P., Glasner, P., 2008. The UK stem cell bank: Securing the Past, Validating the Present, Protecting the future. Science As Culture 17 (1), 43–65.
Vermeulen, N., Tamminen, S., Webster, A. (Eds.), 2011. Bio-Objects: Life in the 21st Century. Ashgate, Aldershot.
Wainwright, S.P., Williams, C., Michael, M., Cribb, A., 2009. Stem cells, translational research and the sociology of science. In: Atkinson, P., Glasner, P., Lock, M. (Eds.), Handbook of Genetics and Society. Routledge, London, pp. 41–58.
Waldby, C., 2002. Stem cells, tissue cultures and the production of biovalue. Health: An Interdisciplinary Journal 6 (3), 305–323.
Watson, J.D., Crick, F.H.C., 1953. A structure for Deoxyribose Nucleic Acid. Nature, 737–738. 4356, April 25.
Wright, S., 1994. Molecular Politics. Developing American and British Regulatory Practices for Genetic Engineering. Chicago University Press, Chicago, IL.
Wynne, B., 1995. Public understanding of science. In: Jasanoff, S., et al. (Eds.), Handbook of Science and Technology Studies. Sage, London, pp. 361–388.

Genetics and Sociology[a]

Dalton Conley, New York University, New York, NY, USA
Jason Fletcher, La Follette School of Public Affairs, University of Wisconsin–Madison, Madison, WI, USA
Christopher Dawes, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

For close to a century, the social and biological sciences have treated the role of biology and genetics in the explanation of human behavior in different ways; while the social sciences have tended to privilege 'nurture' as the cause for variations, biological sciences have leaned toward foregrounding 'nature.' Increasingly, scholars in both fields are viewing the roles of genetics and environment as complementary, a perspective that is also contributing to the breakdown of disciplinary barriers. Contemporary sociologists, for example, often take into account the potential role of genes in analyses of social behavior. This article describes the history of interactions of the disciplines of sociology and genetics and reviews the development of intellectual discourse of the relative roles of genes and environment, highlighting recent contributions by contemporary sociologists.

The controversy surrounding the relationship among evolution, genes, and society has ebbed and flowed over the 150 years since Charles Darwin penned *The Descent of Man* in 1871. Philosopher Herbert Spencer applied natural selection as a metaphor to human society in the mid-1800s, and Charles Darwin himself became embroiled in a debate about whether blacks and whites constituted separate species. (Ethnocentrists of the nineteenth century had sought to use Darwin's concepts to argue for fundamental biological differences between the groups we call races. When Darwin countered these claims, he curiously found himself allied with his one-time foes: the religious conservatives of the day who also argued for the unity of humankind.) Most recently, the revelation that the Beijing Genomics Institute is sequencing the genomes of 2000 individuals with IQs over 150 in search of intelligence genes has led to renewed concerns in the West of an emerging program of eugenics through embryo testing and selective abortion (Normile, 2002).

Part of the reason for such trepidation when it comes to the examination of genetics as it relates to human behavior is that it is commonly assumed that the answers obtained by examining genetics are deterministic ones: to the extent that genes explain any social outcome, this 'naturalizes' any inequality in that outcome. For example, in *The Bell Curve*, Richard Herrnstein and Charles Murray argued that thanks to an ongoing system of meritocracy, today's class stratification is based on innate (i.e., genetic) endowment (Herrnstein and Murray, 1994). Meanwhile, by selectively breeding with others of similar genetic stock, parents reinforce their offspring's advantages or disadvantages. In their view, social policy to promote equal opportunity is counterproductive since each individual has reached the level of social status best suited to his or her native abilities. This is the nightmare conclusion of progressive social scientists and the reason why most avoid genetic data like it was the plague. After all, if all our human outcomes and traits are now increasingly explained by discoveries in genetics and some of our basic social constructs becoming complicated by incorporating findings from genetics and the biological sciences, will the social sciences lose their relevance to understanding human social behavior?

History of Social Science and Genetics

The rise of professional social sciences was partly a response to 'Darwinistic' interpretations of human behavior; that is, human behavior as a product of natural selection. Efforts by scholars like Francis Galton and others to establish a science of eugenics engendered social scientific responses attempting to show the importance of the environment in shaping outcomes ranging from delinquency to class attainment to family relations.

However, the science of genetics never gave up on trying to explore the genetic bases of many human tendencies. During the 1970s a spate of studies emerged suggesting that between a third and half of socioeconomic outcomes could be explained by 'genetic' differences. Even though it was some of their own who authored many of these studies, social scientists wasted no time in questioning both the core assumptions in the models as well as the utility of such estimates themselves. The fire was further fanned by the 1994 publication of *The Bell Curve* which argued that meritocracy had led to the perverse outcome that inequalities were largely based in innate ability – and thus resistant to being remedied by public policy. *The Bell Curve* was, in short order, attacked for its poor methodology and ideological spin. However, the truth is that all of these studies that utilized – twins, adoptees, separated family members, and so on – were really just trying to estimate the size of a black box called genotype without being able to peer inside it. Meanwhile, until recently, no social scientists even bothered to defend the concept of genetic inheritance as a useful datum in the designing of policy.

Recently, however, a new group of sociologists, political scientists, and economists have joined forces with statistical geneticists to make serious arguments about the utility of genetic information to understanding social dynamics

[a] This article is adapted from The Emergence of Socio-Genomics. Contemporary Sociology: A Journal of Reviews 43, 458–467 (July 2014). http://dx.doi.org/10.1177/0094306114539640.

(e.g., Benjamin et al., 2012; Conley, 2009; Fowler and Dawes, 2008). Fast-forward to today and that black box has been cracked by lower-cost DNA genotyping platforms that allow a researcher to actually measure around a half-million or more of the base-pair differences between individuals. The genomics revolution that began in the 1990s ran into some stumbling blocks early out of the gate, however. These included inadequate sample sizes for studies that would search for statistically meaningful effects with hundreds of thousands of discrete 'hypotheses' (i.e., for each marker) and the fact that the markers that were typically measured by the big chip companies (e.g., Affymetrix, Illumina, etc.) were common variants and ignored rarer (and potentially more powerful) sites of difference among humans.

The result of these and other challenges was that these measured markers did not seem to explain the level of variation in outcomes – ranging from schizophrenia to height to IQ – that was meant to be due to 'genetics' according to the prior generation genetic investigations based on twin studies. Slowly but surely, statistical geneticists have made significant progress in solving this 'missing heritability' mystery using a range of newly developed tools. For example, perhaps it was the earlier twin studies that overestimated the role of genes to begin with? That is, the key assumption of 40 years of twin studies is the fact that the estimation of genetic influences rests on the notion that identical twins demonstrate more similar outcomes than do their same-sex fraternal twin counterparts because they share more DNA in common, a notion expressed by the term 'heritability.'

However, plaguing this research is the fact that these twins are probably experiencing more similar environments as well, both because of how others treat them (confusing them, for example) and because they may be closer to each other as well and thus mutually influence each other's choices and experiences than what the fraternal twins do. This 'equal environments assumption' has been the weak link in genetic studies – till now. Now that we have actual DNA markers of twin sets, a natural experiment emerged: Some twins turned out to be wrong about whether they were monozygotic or dizygotic, often because their mothers had been wrongly told by their physicians that the presence of two placentas meant that twins were dizygotic. These 'misclassified' twins allowed us to see if those twins who lived their entire lives as identical but were genetically fraternal (and vice versa) led us to the same results as those earlier studies. That is, if there was indeed any conflating of environmental effects with genetic ones, these misclassified twin sets should tease that out (Conley et al., 2013a). As it turned out, the equal environment assumption of the 1970s held, and the problem of 'missing heritability' – or the extent to which 'genes' explain behavior or diseases – remained unsolved.

Meanwhile, statistical geneticists have developed a number of innovative techniques to address this missing heritability problem. For example, while on average siblings share 50% of their genome, there is significant variability in that figure due to the randomness of recombination and segregation of the grandparental alleles when parents form sperm and eggs. So one pair of siblings may share 45% of their genome while another (even in the same family) may share 60%. By correlating these differences in relatedness among siblings to differences in their outcomes, we are able to generate an unbiased estimate of the genetic contribution to a given trait – that is, heritability (Visscher et al., 2006). Interestingly, those 1970s estimates based on early twin studies now hold up. Using this sibling and other approaches, earlier estimates of heritability were confirmed. And while any individual DNA base (or entire gene for that matter) may not explain much in terms of how we turn out, the sum total of all nucleotide differences did provide a measure of the extent to which genetic variation was predictive of important social outcomes, directly measurable and, to a certain extent, randomly assigned at birth (Rietveld et al., 2013).

At the same time, more and more national surveys were asking respondents to spit into a cup, adding genotype data to the rich tapestry of social variables that economists, sociologists and political scientists had worked with for decades. These additional data may help scientists better understand patterns of human behavior, enhance individuals' self-understanding, and design optimal public policy, especially when the answers we get from peering into the black box are not always – or even often – the kind that reify existing inequalities, assumptions, and policies. As it turns out, adding genetic data to social science upends the apple cart on many of our assumptions. For example, were Herrnstein and Murray right in *The Bell Curve* when they argued that meritocracy has perversely resulted in more intransigent inequalities today because we are now sorted by genetic ability? As it turns out, the data show that thanks to the magic of sexual reproduction – where the deck of genetic cards is reshuffled each generation – genetics does as much (if not more) to upset existing inequalities (i.e., create social mobility) than it does to reinforce social reproduction.

Additionally, a growing body of literature addresses the fluid and complex nature of the creation of individual and collective identities based on genetics and addresses aspects of the commercialization, politics, and regulation of genetic research (see, e.g., Almeling, 2011; Richards, 2013; Nelkin and Lindee, 1995; Conrad and Gabe, 1999; Petersen and Bunton, 2002; Duster, 2013; Hauskeller et al., 2013; Prainsack et al., 2014).

The Role of Environment

So far, we have focused our attention on what new discoveries in genetics can tell us about very thorny issues and ideas in modern society. In general we have neglected a focus on the environment – specifically how environmental factors upend received wisdom about genetic determination and the naturalization of human behaviors and societal structures. However, we now bring the environment back into the conversation and discuss the many complications that emerge. Indeed, genetics and environmental factors seem to interplay with one another in a complex and dynamic feedback loop that begins to explain further several aspects of the behavior of humans and societies – leading to terms and concepts like gene–environment interactions and gene–environment correlation.

One aspect of human behavior that is puzzling is the range of resilience and vulnerability to environmental conditions that we see. Two people (e.g., siblings) who experience the

same negative life event (famine, war, neglect, abuse) can have wildly different outcomes. Additionally, two people with the same genotype (identical twins) often have wildly different life outcomes. These simple observations lead to the (potentially obvious) suggestions that genetics and environments may interact to determine what we see in human behavior – it is not 'nature versus nurture' as much as 'nature and nurture.'

This idea of nature and nurture dynamically interacting and interplaying with one another has also led to various ideas that feed back into our understanding and theorizing in genetics and evolution. An early suggestion was the stress-diathesis hypothesis, where the idea was that some individuals were born with 'risky' genetic variants and others were born with 'safe' genetic variants – but both individuals would have about the same outcomes if placed in a neutral environment. However, if individuals with 'risky' genetic variants were placed in 'risky' environments, we would see disproportionately different outcomes – a gene–environment interaction. Indeed, there is quite a bit of (still controversial) evidence that children who are abused and have 'risky' gene variants related to the serotonin or dopamine systems in our brains experience incredible reductions in life outcomes, while abused children with alternate gene variants are worse off, but not multiplicatively so (Caspi et al., 2003; Conley et al., 2013b; Conley and Rauscher, 2013; Karg and Burmeister, 2011; Risch et al., 2009).

The next question for the stress-diathesis hypothesis, though, comes from an evolutionary perspective – why would humans have 'risky' genetic variants at all, if there were no benefits? Why would evolutionary pressures fail to wipe out these variants? A couple of theories have been proposed. The first set of theories is that these 'risky' variants were, in the not-too-distant past, beneficial. Perhaps the genetic variants that we seem to see producing 'over-reactions' to current stressful environments would have been excellent variants to have in the dangerous Serengeti, and humans are too recently displaced from these environments for the variants to have disappeared through evolution. Or perhaps these 'risky' variants are risky for some outcomes but actually protective for other, often unmeasured, outcomes.

Or … a second theory considers the gene at the level of the species rather than individual and supposes that there are variants that are 'orchids' and other variants that are 'dandelions.' The dandelions do fine in most environmental circumstances, within a reasonable range. However, the orchids thrive in some environments but wilt in others. Again, this is gene–environment interaction, and, at the level of a species, humans would like to have both types of genes in our pool in case of larger changes in the environment (e.g., climate change) that might wipe out all the dandelions and most of the orchids, but allow some orchids to thrive (Conley et al., 2013b).

These theories and the evidence for gene–environmental interactions also potentially have more pressing implications, while we are waiting for evolution to change genotype distributions, we would also like to figure out how the purposeful shaping of our environment that is under human control may affect us. This leads us into thinking about the built environment, schooling, health policies, energy use, and the whole array of government action.

Specifically, we might ask what allows some policies to be effective and others to fail? Many policies are guided by theory and data. We tax products we do not want people to use because economic theory posits a Law of Demand – when prices go up; people consume less of the product. And this theory is often verified by data – when states increase taxes on cigarettes or soda (or whatever), people shift their choices away from these products. We give people housing vouchers to move away from impoverished neighborhoods following sociological and economic theories about the presence of poverty traps and the importance of neighborhoods. However, our policies have a very mixed success rate. Some are successful for some people or during some periods of time but not others. New evidence merging genetics and public policy has started to uncover *why* we see such different impacts of the same policy for different people and how future policies might be targeted – taking the concept of 'personalized medicine' to allow 'personalized policy.'

A case in point is the example of tobacco taxes in the United States. The Institute of Medicine and Centers for Disease Control and Prevention, among others, have ranked this policy as one of the top 10 most impactful on increasing public health in the past century. This is because the United States has witnessed enormous changes in smoking since the first taxes were introduced following the Surgeon General's report in the mid-1960s – which have been cut in half. And raising taxes is cheap for the government, so the benefit–cost ratio is incredibly favorable. But the last decade has seen a reversal. In the face of the largest tax increases in our history, tobacco use has remained virtually unchanged. Has the Law of Demand been repealed? New evidence that combines genetics and social policy evaluation has pointed to an explanation: the pool of smokers in the 1960 was genetically different than the pool of smokers now. People carry around different variants of a group of nicotinic receptor genes with them – and as the name suggests, these genes decide how much dopamine (a pleasure chemical) is released when you smoke; how much you 'like' smoking. These genetic variants also interact with our environments. It looks like many of the smokers in the 1960s did not need a big push (tax increase or social pressure) to quit smoking (or never start). But over time, these small pushes have mostly pushed people to stop smoking if their genes did not put up a big fight. And what we have left in the smoking pool are disproportionately folks whose genes are fighting back, partly because nicotine is so pleasurable. This shows up in evaluations of tobacco tax policy, where only adults with low genetic risk of smoking will still respond to taxation and adults at higher genetic risk seem to be unshaken (Fletcher, 2012).

What does this mean for policy? Should we keep increasing taxes – forcing the remaining smokers to pay a greater and greater share of their incomes in part because they were unlucky in the parental (genetic) lottery of life? Or should we even subsidize their smoking because they love it so much (due in part to their genetics)? Smoking is not the only area where 'personalized' policy may come into play. Evidence has also shown that some educational interventions have greater or lesser effects depending on targeted students' genotypes. Ditto for crime prevention policy. And even the Earned Income Tax Credit spurs low-wage workers to labor more or less depending on their genes. While personalized medicine may be just over the horizon, the potential for custom-tailored social policy is

here today. Whether we want to go down that path is another question altogether.

Genetic analysis not only can inform microlevel analysis of human behavior, it can also add to macrohistorical analysis of social phenomena. For example, a fundamental question in macroeconomics is why some countries have thrived and others have stagnated over the past several hundred years. There has been a revival of interest in this topic in the past few decades and many novel and stark hypotheses have been proposed. Jared Diamond, in a series of books and articles, has suggested that environmental factors that contribute to differences in disease burden, soil efficiency, and similar 'endowments' across countries are a primary source of difference in what we currently see between rich and poor countries. For example, tropical diseases reduce life expectancy in places like Zambia so that trained workers can expect to have only 10 years of economic productivity – the same figure for the United States is over 35 years (Diamond, 1997). Agricultural grain yields in US soil are about 10 times greater per acre than for their African counterpart. And African River Blindness means that Africa is the only continent where populations concentrated *away* from the rivers and coasts, where the most fertile soil is and where growth-stimulating trade is facilitated.

More recently, economists Daron Acemoglu and Simon Johnson and others have dismissed that the environment is a key factor explaining economic success and instead consider the development and expansion of institutions (legal systems, democracy, property rights, corruption, etc.) as the key set of factors explaining why some countries are rich and others are poor (Acemoglu et al., 2002; Easterly and Levine, 2003; Easterly et al., 1993). Some countries arrive at having 'extractive' institutions in which small groups of individuals exploit the rest of the population (think diamond mines in Sierra Leone or Congo) while other countries arrive at 'inclusive' institutions where many people are included in governance (Acemoglu and Robinson, 2012). Key to this argument are examples where geography is held constant but institutions differ. For example, North Korea and South Korea share a peninsula and a microenvironment. The land is of similar quality, the topology of the environment not too different, even the populations are nearly identical. However, a history of inclusive governance in South Korea, and one of noninclusive (extractive) governance in North Korea, Acemoglu and Johnson argue, shines a bright light on the importance of institutions and not geography as the key to development. Think also about the economic performance of East versus West Germany or economic growth trajectories in towns on the US versus Mexico side of the Rio Grande – similar environments, similar people, similar disease burden, but different institutions and different growth and development.

As these new ideas are being sorted out, economists interested in the intersection of economic development, institutions, and geography have begun to explore another aspect of populations that might fit into these grand theories: Population Genetics. In particular, a new breed of macroeconomists has posited that genetic diversity within countries is a key to development. Quamrul Ashraf and Oded Galor published an article providing evidence that an optimum level of genetic diversity within countries might lead to higher incomes and better growth trajectories (Ashraf and Galor, 2013). The authors discuss the observation that there are many countries with relatively low diversity, such as in the Americas, as well as countries with high diversity (e.g., many sub-Saharan African countries) have experienced low economic growth, while many countries with intermediate diversity (such as in Europe and Asia) have been conducive for development in the precolonial as well as modern eras. Basically areas/countries face a trade-off with genetic diversity between cooperation and innovation. The idea is that areas with genetic diversity that is 'too high' have higher likelihoods of disarray and mistrust, which would reduce cooperation and disrupt socioeconomic order and lead to low productivity. However, areas with genetic diversity that is 'too low' might have less innovation, fewer new ideas, and therefore reduced technological capacity. The authors argue that higher diversity "enhances society's capability to integrate advanced and more efficient production methods, expanding the economy's production possibility frontier and conferring the benefits of improved productivity" – an idea motivated by Darwin's theory of evolution by natural selection. These conflicting costs and benefits of genetic diversity lead to the proposition that a 'middle level' of diversity will lead to the highest growth and development patterns. The authors find that, for countries that have low genetic diversity, small increases in diversity (1% point) could increase their population density (a measure of economic development) by 58%; likewise countries with current high genetic diversity who reduced this diversity by 1% point might see an increase in population density of 23%. It should be noted that Ashraf and Galor's premise has been critiqued on a number of fronts, including the fact that a correlation – such as between levels of genetic diversity and economic indicators – does not mean causation (d'Alpoim Guedes et al., 2013).

In addition to considering the 'optimal' level of genetic diversity in populations to maximize economic development, other economists have considered the role of population genetics as it interacts with environmental resources to affect growth patterns across countries. Justin Cook has shown that populations with the (genetic) ability to digest milk after weaning that appeared early in human history conferred large advantages in population density around CE 1500 (Cook, 2013b). As other studies have shown that economic development differences in history have been remarkably persistent, the implication is that (relatively) small changes in the genome, at the right time and in the right place (during the Neolithic Revolution in areas able to raise cattle) can lead to large, persistent, and accumulating differences in economic development across countries.

Macroeconomists have also begun exploring how population genetics might affect economic development through the health of populations that then affect productivity and incomes. For example, Cook has also shown that populations with immune systems that are 'genetically diverse' have had a health advantage in the premodern period (Cook, 2013a). The idea is that pathogens evolve to target specific immune function weaknesses, and populations with limited genetic diversity (and hence limited diversity in immune response) are at particular risk of infectious pathogens spreading and reducing the health of wide swaths of the population. However, at the population level, genetic diversity can be beneficial in inoculating against such widespread health insults

by constraining epidemic spread of illness. Because many pathogens specialize in their attack strategy, populations with multiple defenses (through genetic diversity) can potentially reduce the long run effects of infectious diseases on population health. Indeed, Cook finds that increases in immune genetic diversity (through the human leukocyte antigen system) leads to increases in population (country-level) life expectancies. He then further documents this causal argument by showing that the invention and widespread use of modern vaccinations and other medical technologies has led to a decline in the genetic advantage. That is, modern science and medicine is substituting for 'natural' (genetic) defenses against illnesses – at the population level – and in doing so is promoting convergence in life expectancy, and eventually growth and income, across rich and poor countries. This is another example of gene–environment interactions, at the population level. In earlier times (in diseased environments with lack of medications), genetic variation acted as a buffer against disease – leading to country level differences in life expectancy based in part on genetic differences. But now that the environment has changed, with new medications/vaccinations, the previous genetic advantages have been largely eliminated, so that genes interact with the larger environment in producing outcomes. Likewise, in our earlier example of the 'milk genes,' these genes only confer advantages when in environments which have the ability to foster agriculture. With no cows, goats, or other domesticatable mammals, the gene confers no population advantage.

Conclusion

Until recently, the study of human genetic variation has consisted mainly of behavior genetics studies, where twin and adoption designs were used to identify heritable, or genetic, variation in various traits (see, e.g., Björklund et al., 2006; Plomin et al., 1994; Plomin, 2009; Plug, 2004; Sacerdote, 2007). These studies are controversial and the assumptions underlying them have been questioned (e.g., Goldberger, 1979; for a defense see Conley et al., 2013a; Scarr and Carter-Saltzman, 1979). Whether or not one believes the estimates of genetic influence on phenotypes that emerge from such studies, the fact remains that they do not directly measure genotypes and thus have limited utility for social scientists.

But now, the cost of human genotyping is dropping faster than the rate at which Moore's law is bringing down the price of computer chips (Benjamin et al., 2012), and Americans are genotyping themselves in record numbers using consumer services such as 23andme, AncestryDNA, or FamilyTreeDNA. (In addition to citizen/consumer genotyping databanks, major funding bodies, including those in the social and behavioral sciences, have now begun to incorporate genetic and biological markers into major social surveys. For example, in addition to the pioneering study – the National Longitudinal Survey of Adolescent Health (Add Health) – more recently the Wisconsin Longitudinal Study and the Health and Retirement Survey have released datasets with comprehensively genotyped subjects. Similar efforts are also underway in Europe, for example with the UK Biobank Project in the United Kingdom (Ollier et al., 2005; Platt et al., 2010) and large-scale genotyping of subjects at several European twin registries (Rønningen et al., 2006). These datasets provide new opportunities for asking scientific questions that could not be explored until very recently.) And they, in turn, are acting on the information they receive – as recently demonstrated by Angelina Jolie's preventative, bilateral mastectomy. Meanwhile, genetic data is pouring into what were once purely social scientific studies, raising old debates about genes and IQ, racial differences, risk profiling, criminal justice, political polarization, and privacy. If sociologists seek to remain relevant to the study of human behavior, we must embrace, not flee, this deluge of data. Whether social scientists are interested in modeling gene-by-environment interaction effects to better understand heterogeneity in response to social influences, or, rather, if they merely wish to parse out the genetic effects on behavior in order to gain a less biased estimate of purely social effects, they would do well to integrate genotypic data into their analyses – and ignore it as their peril.

See also: Bioethics in the Post-genomic Era; Darwinism; Eugenics as a Basis of Population Policy; Eugenics as an International Movement; Evolutionary Economics; Gene–Environment Interactions in Health and Well-Being; Genes and Behavior: Lessons from the Fruit Fly; Genetic Studies of Behavior: Methodology; Genetics and Anthropology; Genetics and Society; Genomics, Ethical Issues in; Twin Studies and Epigenetics.

Bibliography

Acemoglu, D., Johnson, S., Robinson, J.A., 2002. Reversal of fortune: geography and institutions in the making of the modern world income distribution. The Quarterly Journal of Economics 117 (4), 1231–1294.

Acemoglu, D., Robinson, J., 2012. Why Nations Fail: The Origins of Power, Prosperity, and Poverty. Crown Business, New York.

Almeling, Rene, 2011. Sex Cells: The Medical Market for Eggs and Sperm. University of California Press, Berkeley, CA.

Ashraf, Q., Galor, O., 2013. The "out of Africa" hypothesis, human genetic diversity, and comparative economic development. American Economic Review 103 (1), 1–46.

Benjamin, D., Cesarini, D., Chabris, C., Glaeser, E.L., Laibson, D.I., Guðnason, V., Harris, T.B., 2012. The promises and pitfalls of genoeconomics. Annual Review of Economics 4, 627.

Björklund, A., Lindahl, M., Plug, E., 2006. The origins of intergenerational associations: lessons from Swedish adoption data. The Quarterly Journal of Economics 121 (3), 999–1028.

Caspi, A., Sugden, K., Moffitt, T., Taylor, A., Craig, I.W., Harrington, H., Poulton, R., 2003. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science 301 (5631), 386–389.

Conley, D., 2009. The promise and challenges of incorporating genetic data into longitudinal social science surveys and research. Biodemography and Social Biology 55, 238–251.

Conley, D., Rauscher, E., 2013. Genetic interactions with prenatal social environment effects on academic and behavioral outcomes. Journal of Health and Social Behavior 54 (1), 109–127.

Conley, D., Rauscher, E., Dawes, C., Magnusson, P.K., Siegal, M.L., 2013a. Heritability and the equal environments assumption: evidence from multiple samples of misclassified twins. Behavior Genetics 43 (5), 415–426.

Conley, D., Rauscher, E., Siegal, M., 2013b. Beyond orchids and dandelions: testing the 5-HTT "risky" allele for evidence of phenotypic capacitance and frequency-dependent selection. Biodemography and Social Biology 59 (1), 37–56.

Conrad, P., Gabe, J. (Eds.), 1999. Sociological Perspectives on the New Genetics. Blackwell Publishing.

Cook, J., 2013a. The Natural Selection of Infectious Disease Resistance and Its Effect on Contemporary Health. New Haven, CT.

Cook, J., 2013b. The Role of Lactase Persistence in Precolonial Development. New Haven, CT.
Diamond, J., 1997. Guns, Germs, and Steel. W.W. Norton, New York.
Duster, T., 2013. Emergence vs. reductionism in the debate over the role of biology in politics. Perspectives on Politics 11 (02), 495–499.
Easterly, W., Bank, W., Kremer, M., Pritchett, L., Summers, L.H., Bruno, M., Schiantarelli, F., 1993. Good policy or good luck? Journal of Monetary Economics 32 (3), 459–483.
Easterly, W., Levine, R., 2003. Tropics, germs, and crops: how endowments influence economic development. Journal of Monetary Economics 50 (1), 3–39.
Fletcher, J., 2012. Why have tobacco control policies stalled? Using genetic moderation to examine policy impacts. PloS One 7 (12), e50576.
Fowler, J.H., Christopher, T.D., 2008. Two genes predict voter turnout. The Journal of Politics 70 (3), 579–594.
Goldberger, A., 1979. Heritability. Economica 46 (184), 327–347.
Guedes, Jade d'Alpoim, et al., 2013. Is poverty in our genes? Current Anthropology 54 (1), 71–79.
Hauskeller, C., Sturdy, S., Tutton, R., 2013. Genetics and the sociology of identity. Sociology 47 (5), 875–886.
Herrnstein, R., Murray, C., 1994. The Bell Curve: Intelligence and Class Structure in American Life. Free Press, New York.
Karg, K., Burmeister, M., 2011. The serotonin transporter promoter variant (5-HTTLPR), stress, and depression meta-analysis revisited: evidence of genetic moderation. Archives of General Psychiatry 68 (5), 444–454.
Nelkin, D., Lindee, S., 1995. The DNA Mystique: The Gene as a Cultural Icon. W.H. Freeman.
Normile, D., 2002. Beijing genomics institute: from standing start to sequencing superpower. Science 296 (5565), 36–39.
Ollier, W., Sprosen, T., Peakman, T., 2005. UK biobank: from concept to reality. Pharmacogenomics 6 (6), 639–646.
Petersen, A., Bunton, R., 2002. The New Genetics and the Public's Health. Routledge, London.
Platt, A., Pullum, S., Lewis, D., Hall, A., Ollier, W., 2010. The UK DNA banking network: a "fair access" biobank. Cell and Tissue Banking 11 (3), 241–251.
Plomin, R., 2009. The nature of nurture. In: Experience and Development: A Festschrift in Honor of Sandra Wood Scarr, p. 61.
Plomin, R., Owen, M., McGuffin, P., 1994. The genetic basis of complex human behaviors. Science 264 (5166), 1733–1739.
Plug, E., 2004. Estimating the effect of mother's schooling on children's schooling using a sample of adoptees. The American Economic Review 94 (1), 358–368.
Prainsack, B., Schicktanz, S., Werner-Felmayer, G. (Eds.), 2014. Genetics as Social Practice. Ashgate, Farnham.
Richards, M., 2013. The new genetics: some issues for social scientist. Sociology of Health and Illness 15, 567–586.
Rietveld, C., Medland, S., Derringer, J., Yang, J., 2013. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science 340 (6139), 1467–1471.
Risch, N., Herrell, R., Lehner, T., Liang, K.-Y., Eaves, L., Hoh, J., Merikangas, K.R., 2009. Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression. The Journal of the American Medical Association 301 (23), 2462–2471.
Rønningen, K., Paltiel, L., Meltzer, H.M., Nordhagen, R., Lie, K.K., Hovengen, R., Hoppin, J.A., 2006. The biobank of the Norwegian mother and child cohort study: a resource for the next 100 years. European Journal of Epidemiology 21 (8), 619–625.
Sacerdote, B., 2007. How large are the effects from changes in family environment? A study of Korean American adoptees. The Quarterly Journal of Economics 122 (1), 119–157.
Scarr, S., Carter-Saltzman, L., 1979. Twin method: defense of a critical assumption. Behavior Genetics 9 (6), 527–542.
Visscher, P., Medland, S., Ferreira, M., 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genetics 2 (3), e41.

Genetics and the Media

Kate O'Riordan, University of Sussex, Falmer, Brighton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Genetics and genomics have been made in the media in significant ways. From the new genetics of the mid-twentieth century, to the Human Genome Project of the 1980s and 1990s to the personal and public genome projects of the twenty-first century, the media has been central. The new genetics ushered in the icon of the double helix and the figure of the heroic genetic scientist. The late twentieth century saw the ascendance of the gene as a powerful cultural trope and the emergence of media genres such as gene-hunting stories and documentary and reality television genres became partners in genetic research. The media became the site of scrutiny and creative intervention in turn of the century anxieties about public engagement with science, and science institutions in turn became more involved in media making. The science–media relationship has seen another shift, from representation to interaction with genomics, and current social media structures are used to enroll publics and research subjects alike.

Since the middle of the twentieth century, genetics has broken with negative associations (eugenics, Nazi science) through the 'new genetics' of the double helix and beyond. These shifts in scientific provenance and reputation have gone hand in hand with media representations of genetics. To a large extent the ascendance of genetics has been coupled with, and facilitated by, increasingly prolific media coverage over the last 50 years. More recently the media cultures of genetics and genomics have shifted from representing and publicizing the icons of genetics to a means by which identities and publics are incorporated into genomics. Examples of this incorporation include reality television genres in which research subjects and patients provide the content for production, and direct-to-consumer genome testing that uses social media structures to enroll and manage media consumers as research subjects.

Representation and Publicity: The Making of an Icon

The new genetics was marked by the image of the double helix (Watson and Crick, 1953) and narratives about its construction. In the late 1980s the 'century of the gene' (Keller, 2002) was being reworked as an era of scientific discovery in health benefits with commercial appeal. Between the 1950s and 1980s the image of the double helix had traveled across so many media forms that it could be said to be iconic. Anthropologist Professor Sarah Franklin argued that it had become a fetish, "Etched on the Western consciousness" (1988).

In the 1990s, Dorothy Nelkin and Susan Lindee (1995), Jose Van Dijck (1998), and Celeste Condit (1999) highlighted the ubiquitous appearance of the gene and its 'mystique' (Van Dijck, 1998) in multiple popular sites, including cartoons, advertising, and lifestyle magazines. One of the aims of their work was to examine the meanings of popular genetics, meanings that are made beyond the control of scientists, research projects, and institutions. Their work signals a strand of media and cultural studies scholarship that intersects with the history and philosophy of science and science communication. An interest in these areas also came to matter for studies in the public understanding of science, public engagement, ELSI (ethical, legal, and social implications) projects and in understanding genetic knowledge. Importantly these scholars also examined the interplay between popular and scientific meanings, concluding that the imaginaries of the gene were powerful, but limited, and both scientific and popular sites needed some kind of 'retooling' of the imagination (Van Dijck, 1998).

Watson and Crick's autobiographical versions of the new genetics (Watson, 1968; Crick, 1988) were media-friendly and news worthy and provided content for books, news, and TV drama in the 1970s and 1980s. The popularity of the story and of Watson's (1968) autobiography was sustained in the making and reception of the British television drama *Lifestory* in 1987. This was a Horizon special science documentary that dramatized James Watson and Francis Crick's construction of the double helix in the 1950s as a heroic discovery narrative. Greg Myers and Sarah Franklin wrote about *Lifestory* in 1988 and 1990. *Lifestory*'s fictionalized version of the work of science, and the vogue for celebrity media scientists ever since, helped to recast genetic scientists as heroic (and sometimes maverick) individuals. In the following decade Jon Turney (1998), Pat Spallone (1993), Jose Van Dijck (1998), Dorothy Nelkin and Susan Lindee (1995), and Donna Haraway (1997) all examined the iconic and fetishistic symbolism of the gene playing out in media cultures.

Despite the iconography of the double helix, media cultures of genetics in the late twentieth century retained some attachment to the darker associations of eugenics and Nazi science. Even as the heroes of modern genetics were being created as media figures, the influential film *The Boys from Brazil* (1978) reminded cinema audiences of the field's associations with evil through a plot linking genetics to cloning and Nazi science. The film constructed a fictional postwar scenario about the cloning of Hitler in secret facilities in Paraguay by Nazi scientists in hiding. The plot supported historical memory of the Second World War, representations of recent scientific cloning advances (Bromhall, 1975), and Joseph Mengele's sanctuary in Paraguay, to intimate that the horror of the film scenario lay in its proximity to the bounds of possibility. The association of genetics with cloning, and thus largely with horror, is one that has continued through media cultures despite the very recent, albeit ambivalent, reworking of the horror of cloning into stem

cell hope. (Haran et al., 2008; O'Riordan, 2012) (This article does not cover all media forms and art and film are notable absences. See Anker and Nelkin (2004) for a review of art in the genetic age and O'Riordan (2012: pp. 71–97) for a discussion of art and genomics. A film and cultural studies analysis of genetics in the cinema can be found in Jackie Stacey's *The Cinematic Life of the Gene* (2010).)

Genes, Genomics, and Media Genre: Celebrity Scientists, Concerns, and Publics

These darker associations were largely reworked as the new genetics of the double helix was consolidated by the advent of the Human Genome Project (HGP) in the 1980s. This project drew intense media attention, as it was represented as heralding a new era of global and public science. A new genre of 'gene-hunting' news stories was instigated which grew in the 1980s and 1990s. These were often linked to humorous examples of pop sociobiology. For example, the opening line of a UK newspaper read:

> A month does not go by, it seems, without some new advance being made in gene discovery. Last week, a team of U.S. scientists published their research in pinpointing a gene for dyslexia. As well as causing Alzheimer's disease, cystic fibrosis and haemophilia, genes have, more controversially, been blamed for predisposing individuals towards shyness, alcoholism and even aggression.
>
> Slater, 1994

The proliferation of these stories, and the ongoing attention to the double helix, the gene, and the institutionalization of genomics led to concerns with this attention and popularity. By the 1990s in the UK there were political concerns about the public understanding of science as a whole, the uptake of genomics specifically, and the beginning of significant research trajectories that would look at the role of the media in relation to these concerns.

The pattern of "the modern, secular, scientific hero" established by Watson (McNeil, 2007: p. 47) as a way of understanding the practices of genetic science has persisted through the HGP and beyond. John Sulston and Craig Venter were cast as the leading figures of this project. This way of figuring celebrity science (McNeil, 2007; Haran et al., 2008; O'Riordan, 2011) continues in the representations of Craig Venter, George Church, Steven Pinker – and most recently – Sergey Brin. Some problems with this particular narrative structure include the erasure of women from representations of science, and an image of science as driven by individual genius, rather than as a complex socially situated set of institutions and working conditions. The scientific hero has become a feature of representations of science, and the celebrity scientist a legitimate figure when crafted appropriately. However, too close an association with celebrity is also used to manage the boundaries of legitimate science and define its limits (Haran, 2007; Haran and Kitzinger, 2009; Gieryn, 1999).

By the late 1990s, multiple sources of scholarship identified the ongoing prominence of positive gene stories in the news (Durant et al., 1996; Nelkin and Lindee, 1995; Nerlich et al., 2002). In fact these gene-hunting stories did not give way to the cures stories that they invoked and concerns about hype also became part of these stories. The public discourse also featured concerns that these stories would mean that biology was understood as genetic determinism, and that eugenics would be ushered in by other means. These concerns were taken up in much of the social research in this area, as well as structuring debates and public engagement initiatives. Debates about whether life could be understood as information, blueprint, or code and whether genetics-enabled deterministic thinking structured much of the turn of the century news discourse (Nerlich et al., 2002; Condit, 1999) although these features were interpreted in different ways. Concerns about race and genetic selection also patterned some of the news stories together with a positive gloss on medical application. In a reworking of concerns raised in Van Dijick's (1995) work, humanities scholar Judith Roof argued that the "hyperbolized notion of DNA" of the new genetics enabled racist and homophobic ideologies, often thought of as historical, to be imported through descriptions of DNA's importance and the image of genomics as a progressive modern science (Roof, 2007). Something of a counter to this argument is Jon Turney's work on genetics in popular culture, *Frankenstein's Footsteps*, which tracks the ways in which nineteenth-century anxieties and negative associations of genetics rather more explicitly haunt the contemporary biosciences (Turney, 1998).

In the policy research around public understandings of science, media institutions have often been attributed blame for generating the hyperbole and for getting the complexity of the science wrong, and this attribution has driven some of the policy agenda around science education and engagement. However, Jenny Kitzinger's work on the reporting of the gay gene offers a challenge to this. "The mass media, in the UK at least, were no more deterministic or 'homophobic' than science journals and were arguably better at locating the research in its social context and discussing its limitations, assumptions and implications" (Kitzinger, 2006: p. 101). Changes in the structure of science journalism, the rise in the importance of publicity for science funding, and the reliance on press releases generated by scientific institutes and journals such as *Nature* also help to generate these features (Bubela et al., 2009). Haran and Kitzinger's work on stem cell science also points to the importance of understanding the contemporary life sciences as mediated science:

> Science and a generically defined "media" are often framed in opposing camps, not least by scientists who blame the media for sensationalism, inaccuracy and distortion …). Rather than accepting the notion of separate spheres (science versus media) it is thus more meaningful to explore the coverage of contemporary scientific advances (and setbacks) as forms of "mediated science" and unpack the role of the media–in our case, primarily the news media–in the "making" of science.
>
> Haran and Kitzinger, 2009: p. 634

The press release has become a ubiquitous feature in science media and often sets the agenda for science news. The completion of the Human Genome Project offers a particularly dramatic example. President Clinton and Prime Minister Blair's joint address in 2000 called publics powerfully into being. The platform of world leaders addressing a global media public

about an innovation in the life sciences, in the name of God and humanity, operated to sediment the "gene as cultural icon" (Nelkin and Lindee, 1995) and to consolidate the metaphor of the book of life, the genome as map and the promise of dramatic medical cures (Nerlich and Hellsten, 2004). In doing so, they also addressed a global genomic public through mass media circulation. Claims of new era marked the address, both state leaders emphasized an ethical, disease-free future in which biotechnology would partner government for the common good of humanity. Since that announcement genomics has shifted again in terms of its address. No longer a single human genome, there is a plurality of national and personal genomics projects, and at the same time genomic publics and audiences have proliferated and become multiple.

Although many genetic variants were identified during the HGP mapping project, the question of what the genome means for its publics has become a more complex question since its completion. In the absence of large-scale medical benefits, cures or discoveries beyond single gene conditions, familiar tropes about ancestry, and forensics came to the fore. Media dramas and factual programming alike saw a proliferation of gene dramas based on these themes. The biggest single media project was National Geographic's *Genographic Project* (2005). National programming about ancestry, paternity, race, and identity such as the *Face of Britain* (2007, UK), which popularized Walter Bodmer's population genetics, have continued. The *Genographic Project* championed the work of geneticist Spencer Wells, and its reach and impact in popularizing genetics as an epic ancestry project has been significant. It epitomizes a popular genre in the media cultures of genetics where reality or documentary forms enroll research subjects and genetic research is used to frame and narrate identity.

The promise of public health benefits coupled with commercial application persists as the dominant framing for the significance of genomics. However, the figure of the celebrity scientist has become supplemented by the vision of multiple public participants in genome sequencing. *Genomics England*, a national genome-sequencing project in the UK funded from the National Health Service, was launched in 2013 the basis of this vision. However, as well as legitimizing genomics, publics are also positioned as a potential threat, or an antipublic, represented as either too interested or irrational or religious to be good genomic publics. These constructions of dangerous publics threaten to exclude the legitimacy of some people by assigning them as irrational, having a special interest, or as a 'lobby' (Haran et al., 2008). These exclusions generate a form of media address where those who are not on the side of the progress narratives of genomics threaten the (public) benefits of genomics, where activist mobilization is not encouraged, while purchasing genome tests and joining genome projects is endorsed. The appeal to consumer publics and the rejection of agonistic publics constructs genomics as a form of consumption.

Twenty-First Century Media: Interactivity and Mediatization

Genomics is an informatic science that generates more data than can be searched or used on a human scale. Algorithmic structures and computer delegated storage, search, and hypothesis are requirements in contemporary genomic science. These interface directly with Internet media structures and the same conditions – market-led, consumer public uptake, user-generated content, crowd sourcing, and data visualization – come together in both media cultures and genomics. Companies like *23andMe* and initiatives such as the *Personal Genome Project*(s) instantiate this coming together: where to be a user, audience, or subject of genomics is to be a user, audience, or subject of digital media forms.

Since the HGP then there has been a shift in the kind of address made by the media. The completion of the HGP came together with the mass consumption of social media, and mediatization (Krotz, 2008), where media become ubiquitous and aspects of everyday life are mediated. If media audiences were asked to marvel at the iconography of the gene in the 1980s and 1990s and were regaled with gene-hunting stories, twenty-first-century media cultures change the mode of address. Digital media cultures of genomics invite audiences to interact directly with genomics, incorporating it into embodied practices. Rather than watching others become research subjects in the reality TV and documentary mode, digital culture demands participation. This shift from iconography to interactivity does not necessarily escape earlier questions about the quality of public genomic knowledge but it changes how knowledge is made.

The current mode of address in genomic media has also shifted from foregrounding life to foregrounding death. Genomics on a mythic scale is attached to the meaning of both life and death (Keller, 1992). However, the attachments to life, life stories, and books of life (Kay, 2000) that characterized the ascent of the genome have also seen an attention to death. Genome sequencing and genotyping was variously articulated in the UK as 'the doomsday test,' The Killer in Me, 'how will I die?' 'can I take the worst?' in a range of newspaper and television appearances around 2007.

Direct-to-consumer testing, personal whole genome sequencing–projects, and national sequencing initiatives have emerged as media forms. Audiences are addressed as though they can personalize the genome and interact with it. Genomics may be the book of life but the mode of address is to 'know thyself' (Knome Inc.). Audiences are now asked to identify with and narrate a version of 'my genomic life' (Pinker, 2009) and in this transition, in this context, the facts of life now also appear as the realities of death.

Media Futures

The incorporation of research subjects and publics through social media and other digital media forms seems likely to continue. Internet media are showing no signs of going away. These communication technologies as a feature of all kinds of knowledge production, including science, continue in ascendance. Examples of the current media cultures of genomics that have emerged from this genomics/media nexus, and which might point to future developments, include life logging, data tracking, big data, and the Quantified Self.

See also: Bioethics in the Post-genomic Era; Direct-to-Consumer Personal Genetic Testing; Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute; Genetics and Society; Genetics and Sociology; Genetics: The New Genetics; Genographic Project; Genomics, Ethical Issues in; Human Genome Project: History and Assessment.

Bibliography

Anker, S., Nelkin, D., 2004. The Molecular Gaze: Art in the Genetic Age. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Bromhall, D., 1975. Nuclear transplantation in the rabbit egg. Nature 258, 719–722.

Bubela, T., Nisbet, M.C., Borchelt, R., Brunger, F., Critchley, C., Einsiedel, E., Geller, G., Gupta, A., Hampel, J., Hyde-Lay, R., Jandciu, E.W., Jones, S.A., Kolopack, P., Lane, S., Lougheed, T., Nerlich, B., Ogbogu, U., O'Riordan, K., Ouellette, C., Spear, M., Strauss, S., Thavaratnam, T., Willemse, L., Caulfield, T., 2009. Science communication reconsidered. Nature Biotechnology 27, 514–518.

Condit, C., 1999. How the public understands genetics: non-deterministic and non-discriminatory interpretations of the 'blueprint' metaphor. Public Understanding of Science 8 (3), 169–180.

Crick, F., 1988. What Mad Pursuit: A Personal View of Scientific Discovery. Basic Books, New York reprint edition, 1990.

Durant, J., Hansen, A., Bauer, M., 1996. Public understanding of the new genetics. In: Marteau, T., Richards, M. (Eds.), The Troubled Helix: Social and Psychological Implications of the New Human Genetics. Cambridge University Press, Cambridge.

Franklin, S., 1988. Lifestory: the new gene as fetish object. Science as Culture 3, 92–101.

Gieryn, T.F., 1999. Cultural Boundaries of Science: Credibility on the Line. University of Chicago Press, Chicago.

Haran, J., 2007. Managing the boundaries between Maverick Cloners and mainstream scientists. New Genetics and Society 26 (2), 203–219.

Haran, J., Kitzinger, J., 2009. Modest witnessing and managing the boundaries between science and the media: a case study of breakthrough and scandal. Public Understanding of Science 18 (6), 634–652.

Haran, J., Kitzinger, J., McNeil, M., O'Riordan, K., 2008. Human Cloning and the Media: From Science Fiction to Science Practice. Routledge, London.

Haraway, D., 1997. Modest Witness@Second Millennium: Female Man Meets OncoMouse. Routledge, London and New York.

Kay, L., 2000. Who Wrote the Book of Life? A History of the Genetic Code. Stanford University Press, Stanford.

Keller, E.F., 1992. Secrets of Life, Secrets of Death: Essays on Language, Gender and Science. Routledge, London and New York.

Keller, E.F., 2002. The Century of the Gene. Harvard University Press, Cambridge.

Kitzinger, J., 2006. Constructing and deconstructing the 'gay gene': media reporting of genetics, sexual diversity and 'deviance'. In: Ellison, G., Goodman, A. (Eds.), The Nature of Difference: Science, Society and Human Biology. CRC Press, New york, pp. 99–118.

Krotz, F., 2008. Media connectivity: concepts, conditions, and consequences. In: Hepp, A., Krotz, F., Moores, S. (Eds.), Network, Connectivity and Flow: Key Concepts for Media and Cultural Studies. Hampton Press, New York, pp. 13–31.

McNeil, M., 2007. Feminist Cultural Studies of Science and Technology. Routledge, London.

Myers, G., 1990. The double helix as icon. Science as Culture 1 (9), 49–72, 12.

Nelkin, D., Lindee, M.S., 1995. The DNA Mystique: The Gene as a Cultural Icon. W.H. Freeman and Company, New York.

Nerlich, B., Dingwall, R., Clarke, D.D., 2002. The book of life: how completion of the human genome project was revealed to the public. Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine 6 (4), 445–469.

Nerlich, B., Hellsten, I., December 2004. Genomics: shifts in metaphorical landscape between 2000 and 2003. New Genetics and Society 23 (3), 255–268.

O'Riordan, K., 2011. Writing biodigital life: personal genomes and digital media. Biography 34 (1), 119–131.

O'Riordan, K., 2012. The Genome Incorporated: Constructing Biodigital Identity. Ashgate, London.

Pinker, S., 2009. My Genome, My Self. New York Times. January 11, 2009, p. MM24.

Roof, J., 2007. The Poetics of DNA. University of Minnesota Press, Minnesota.

Slater, L., 1994. Have You the Courage to Face Your Genes? Daily Mail. October 18, 1994.

Spallone, P., 1993. The Emperor's new genes. Science as Culture 18 (1993), 110–111, 33.

Stacey, Jackie, 2010. The Cinematic Life of the Gene. Duke University Press, Durham.

Turney, J., 1998. Frankenstein's Footsteps: Science, Genetics and Popular Culture. Yale University Press, London.

Van Dijck, Jose, 1998. Imagenation: Popular Images of Genetics. New York University Press, New York.

Watson, J.D., Crick, F.H.C., 1953. A structure for deoxyribose nucleic acid. Nature 171, 737–738.

Watson, J., 1968. The Double Helix: A Personal Account of the Discovery of the Structure of DNA. Atheneum, New York.

Genetics, Disease, and Reproduction

Carrie Friese, London School of Economics and Political Science, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This entry overviews the literature on genetics, disease, and reproduction, with a focus on what genetic diseases mean for people within their reproductive lives. The focus will be on the different technologies used to both assess and ameliorate these disorders at different stages of reproduction; the ways a genetic disease is understood and acted upon are thus shaped by biological time. For example, preimplantation genetic diagnosis is generally used to avoid giving birth to a child with a heritable disorder that will result in death within the first couple years of the child's life. Meanwhile, newborn screening is used to assess for the presence of a heritable disorder that can be ameliorated if treated early. As such, different social, ethical, and legal issues are raised within the varying reproductive contexts within which heritable diseases are both detected and treated. But in addition, the meaning of these techniques varies across different national and regional contexts.

The Human Genome Project (HGP) was hailed as a means to find the genetic causes of a full range of diseases, an assumption that tightly coupled genetics, disease, and reproduction through heredity. It was assumed that this would enable new treatments for genetic disorders, particular gene therapy. However, the completion of the HGP has somewhat ironically shown this to be a simplistic understanding of disease; an era of postgenomic thinking has been ushered in, wherein genes are not viewed as deterministic but are instead understood as parts within systems and processes (Chilibeck et al., 2011; Lock, 2005; Keller, 2000, 2010). In addition, the HGP has probably been more significant to date in helping to create tests for the *diagnosis* of mono- and polygenetic disorders rather than in the *treatment* of these diseases (e.g., via gene therapy). As a consequence, the HGP has had significant impacts upon reproductive medicine, where more genetic tests have been introduced than in any other area of health care delivery in order to screen for, rather than treat, heritable disease (Pavone and Arias, 2012: p. 236). Assessing and acting upon the risk of genetic disorders have thus become an increasing focus in reproductive medicine.

This entry reviews the social science literature on genetics, disease, and reproduction with a focus on what genetic disease means for people within their reproductive lives. I consider the social science literature that explores the different technologies used to both assess the risk for and ameliorate genetic diseases at different stages of reproduction. This includes prenatal genetic diagnosis, preimplantation genetic diagnosis (PGD), newborn screening, and genetic testing later in life. Taken together, this literature shows that the ways a genetic disease is understood and acted upon are informed by the temporal, social, and biological ordering of reproduction. For example, PGD, in which embryos produced through *in vitro* fertilization (IVF) are assessed for heritable diseases before implantation, was initially developed to avoid giving birth to a child with a severe illness that would result in a painful and premature death (Franklin and Roberts, 2006). Meanwhile, newborn screening was initially developed to screen for a heritable disorder that could be treated if detected early in life (Paul, 1998a). As such, different social, ethical, and legal issues are raised within the varying reproductive contexts in which genetic diseases are both detected and addressed.

A key contribution of social science research on genetics, disease, and reproduction has been to explore the social, economic, political, historical, and cultural contexts that shape reproductive decision making in the context of genetic disorders. Here, scholars have asked how clinicians and parents alike confront the dilemmas raised by these techniques. Because reproductive technologies are here used to select against genetic disorders, these techniques necessarily raise questions and concerns regarding the delineation of the 'normal' body (Vailly, 2008) and corresponding eugenic consequences for those deemed 'abnormal' (Duster, 1990). As a consequence, these techniques have, in some parts of the world, been heavily criticized by disability activists (Parens and Asch, 2003; Shakespeare, 1998, 1999). In addition, reproductive and genetic technologies have been resisted on religious grounds, particularly as a result of abortion opposition (Evans, 2010). A common theme across this literature is that parents and clinicians have had to become what Rapp (2000) has called 'moral pioneers' in adjudicating different responsibilities while assessing, using, or rejecting innovative reproductive technologies to avoid or treat genetic disorders (see also Kalfoglou, 2005; Franklin and Roberts, 2006; Williams et al., 2005).

Scholars have characterized moral pioneering as emblematic of the more widespread individualization and privatization of risk assessment in social life (Kerr, 2004; Hallowell, 1999 Helen, 2010; Williams et al., 2005), as part of the 'responsibilization' and 'ethicalization' of subjects in neoliberalism (Helen, 2010; Rose, 2007). However, as social science research has expanded to consider genetic testing and reproductive medicine beyond urban centers in North America and Western Europe, social scientists are increasingly finding that these techniques are differently constitutive of social life when used in varied political, economic, historical, cultural, religious, and national contexts (Prainsack and Siegal, 2006; Hashiloni-Dolev and Shkedi, 2007; Kelly, 2009; Ivry, 2006; Ettore, 2002).

Testing for Genetic Diseases in Utero: Prenatal Genetic Diagnosis

Much of the research in social studies of reproduction, genetics, and disease focuses on prenatal genetic diagnosis. This includes noninvasive tests such as blood tests for genetic mutations as well as assessing nuchal thickening through ultrasound. These

screens provide risk ratios, which indicate if more invasive, diagnostic tests should be conducted including amniocentesis and chorionic villus sampling. If the test indicates the fetus has a genetic disorder, the pregnant woman and, if applicable, her partner must decide if they will abort. Amniocentesis was originally used to detect Down syndrome amongst pregnant women 35 and older, but prenatal screening has expanded significantly since. Many countries now offer all pregnant women screening for a range of genetic disorders (Williams et al., 2005; Vassy, 2006).

In her now canonical multisited ethnography of amniocentesis, Rapp (2000) interrogated the ways in which different groups of patients and professionals use and understand amniocentesis in New York City. She situates amniocentesis within the historical processes of eugenics, feminist social movements rooted in reproductive choice, disability activism, unequal access to health services in the United States, and scientific developments in the area of genetics. Drawing on Colen's (1995) concept of stratified reproduction, Rapp illustrates the diversity of women's experiences with this technique and the corresponding question of whether or not to abort. She shows that the decision-making process is structurally delimited according to race, ethnicity, class, migration status, and religion. Rapp has since gone on to show how prenatal genetic testing is a form of 'flexible eugenics,' one that is rooted in neoliberal choice as opposed to state and professional mandates (Taussig et al., 2003). But she emphasizes that these choices are made within the socially structured constraints of women's everyday lives.

Social scientists have since explored the ways in which women outside the United States understand and experience prenatal genetic testing, allowing for comparisons of varying practices across different national contexts. In the United Kingdom, Ireland, and Iceland, scholars have explored women's experiences with the expansion of screening beyond high-risk groups to include the general population with tests that occur in the first rather than second trimester, also noting the ambivalences that women have regarding this information and the kinds of choices it enables (Williams et al., 2005; Gottfreosdottie, 2009; Lalor et al., 2009). Meanwhile, scholars have also expanded their analyses to consider men's experiences with genetic testing (Inhorn et al., 2009; Ivry and Teman, 2008; Reed, 2009). For example, where Rapp finds husbands were at times controlling, Reed's (2009) later research in the United Kingdom finds that women experienced their husband's involvement as supporting. Research in Israel similarly shows that women use or refuse prenatal screening in ways that negotiate diverse social and cultural terrain including religious beliefs, perceptions of support for disabled children, a sense of what counts as good mothering, and trust in physicians (Remennick, 2006). But there is overall a great deal of support for prenatal genetic testing in Israel that has been linked to the dominance of the medical profession (Remennick, 2006), bodily ideals related to Zionism (Weiss, 2002), and biocultural notions of life (Hashiloni-Dolev, 2006). Researchers have also found support for prenatal screening amongst disability activists in Israel, which contrasts with the United States and Western Europe (Raz, 2004). This is because disability activists see testing and abortion as a distinct issue when compared to ensuring social support for disabled people after birth. In contrast, social science research in Japan has found that prenatal genetic diagnosis, and the corresponding geneticization of prenatal care, is being actively resisted by health care professionals (Ivry, 2006).

Prenatal genetic diagnosis is organized so as to position women and their partners at the center of reproductive decision making through nondirective counseling. This shifts responsibility from the medical profession and state to women and their partners, which in turn distances prenatal genetic diagnosis from early twentieth-century eugenic programs (Kerr et al., 1998; Williams et al., 2002c; Paul, 1998b). Social science research in the United Kingdom has shown that the majority of health professionals support the idea that women should be allowed to make these decisions (Williams et al., 2002a,b; Farsides et al., 2004). However, this research also finds that some health professionals question if there should be a line drawn. That said, these health professionals were also aware that how to draw such a line is difficult, raising questions about what counts as a normal versus pathological, or mild versus severe, condition that frequently lack consensus (Williams et al., 2002b). Social scientists working in the tradition of disability studies have emphasized that people living with these conditions should be involved in such deliberations (Shakespeare, 1999), and that social science research on reproduction, genetics, and disease should include within its purview parenting children with disabilities (Landsmen, 2003). In turn, while early research found that clinicians failed to appreciate the structural conditions in which decisions about having a disabled child are made (Kerr et al., 1998), Williams et al. (2002d: p. 747) found that many of the health care workers they spoke with in the United Kingdom were troubled by the ways in which stigma surrounding disability along with structural constraints shape decision making.

Nondirective counseling is the primary technology through which reproductive decisions are made in the context of prenatal genetic diagnosis (Pilnick, 2004; Williams et al., 2002c; Paul, 1998b). There are, however, questions about the extent to which counseling can be nondirective, as research indicates that offering prenatal genetic diagnosis is often interpreted by women as an endorsement of the technique and the perception of it is routinization (Pilnick, 2004: p. 456; Williams et al., 2002c,d). As a consequence, some women are not aware that they are actively choosing to use these tests (Press and Browner, 1997). And miscommunication in the context of cultural and linguistic differences can create further barriers to nondirective counseling (Browner et al., 2003).

Testing for Genetic Diseases *In Vivo*: PGD

PGD builds upon and extends the use of IVF to treat infertility in order to alternatively avoid conceiving a fetus with a heritable disorder (Franklin and Roberts, 2006; Ehrich and Williams, 2010). The embryos created through IVF are tested for chromosomal indications of a disorder, and only those embryos not at risk are transferred back to the woman for gestation. As such, PGD allows women and couples to test for a heritable disease before pregnancy so as to avoid termination.

Due to the relationship between genetic abnormalities and advanced maternal age, PGD is also used to test for chromosomal alterations and aneuploidy so as to improve the chance of pregnancy using IVF amongst women over 35. The technique has expanded to include the creation of 'savior siblings.' Here, embryos are selected both to avoid a genetic disorder and for tissue matching, so that the resulting child could serve as a cord blood or bone marrow donor to another child suffering from a heritable disease (Hashiloni-Dolev and Shkedi, 2007: p. 2081; Devolder, 2005). And there are ongoing debates over the use of PGD for social sex selection.

Different countries have implemented different policies surrounding the use of PGD. Social science research has, on the one hand, examined the reasons for this variation. A significant amount of research has focused on Germany's resistance to PGD in comparison to regulated support in other national contexts within Europe (Krones and Richter, 2004; Meister et al., 2005; Hashiloni-Dolev and Shkedi, 2007). Surveys indicate that this is a reflection of the political culture of Germany as opposed to the general public's attitudes and opinions (Krones and Richter, 2004; Meister et al., 2005). Similarly, in considering the rapid expansion of PGD in Spain when compared to other permitting countries like the United Kingdom, Pavone and Avias (2012) have emphasized the importance of political economy in delineating varied policies. They explain this diffusion through the expansion of private IVF centers in Spain that cater to medical tourists, combined with the slow implementation of regulatory oversight. Meanwhile, Hashiloni-Dolev and Shkedi (2007) have compared German opposition to PGD for sibling donation with support in Israel, a country that is known to be very supportive of reproductive and genetic technologies. They argue that the difference between these two national policies is rooted in different family ethics, particularly different belief systems regarding familial obligations. They contend that German beliefs that parents should unconditionally love their children as individuals inform concern that PGD could be exploitative, whereas Israeli understandings of the family as a mutually dependent unit inform the support for PGD as a means for siblings and parents to help one another as part of a unified body. In doing so, Hashiloni-Dolev and Shkedi emphasize the cultural as opposed to the political economic in explaining national differences with regard to PGD regulation and use.

Within much of Europe, the popular discussion of PGD has been framed by the idea that parents are trying to 'design' and 'perfect' their children by using this technique. Some areas of social science scholarship have expressed similar concerns. Habermas (2003) has, for example, argued against PGD on the basis that parents, in making such choices about their child's life, restrict their child's freedom to become a unique individual. Meanwhile, much of the social science literature has alternatively sought to understand the experiences of parents who decide to use PGD (Franklin and Roberts, 2006; Kalfoglou, 2005). Franklin and Roberts (2006) conducted one of the first ethnographic studies of PGD in the United Kingdom, which included both observations and interviews within clinics as well as qualitative interviews with would-be parents in their homes. Countering the frames of 'design' and 'perfection,' Franklin and Roberts instead show parents who have either suffered multiple miscarriages or lost a very young child to a very painful death as a consequence of a heritable disease. Some women had undergone one or more terminations after tests showed that the fetus carried genetic markers for a disorder that they had already seen a child or loved one suffer from. In this context, parents had to continually make a number of complicated decisions within the uncertain trajectory of both IVF and PGD, where success rates are low.

Building upon this research, social scientists have also explored PGD and IVF staff perceptions and experiences with this form of genetic testing. Contra the polarized debates in the United Kingdom, research has found that staff have rather divergent, conflicted, and ambivalent feelings about embryos and PGD (Ehrich et al., 2008, 2007; Williams et al., 2007). This is because decisions are rarely clear-cut. For example, should a couple select against embryos that are unaffected or should they also select against embryos that are 'carriers' (e.g., embryos that will result in a healthy child who could nonetheless pass the genetic mutation to their children)? The latter could be seen as a stronger form of eugenics, as it goes beyond ridding the disease from the next generation to ridding the disease into the future (Shakespeare, 1998). Ehrich et al. (2007) found that staff members within UK PGD and IVF clinics held very different opinions when addressing this question in weighing the individual versus societal ramifications of the technique.

PGD has begun to raise new questions and dilemmas as its use has expanded from severe genetic disorders with high penetrance that result in the death of young children to include chromosomal abnormalities with lower penetrance, with an uncertain prognosis, and that affect adults (Williams et al., 2007; Kalfoglou, 2005). Research has found that providers in both the United States and the United Kingdom are less supportive of these developments, and express a range of opinions (Kalfoglou, 2005; Williams et al., 2007). The expansion of PGD for sex selection also raises new kinds of questions. While currently illegal in the United Kingdom, Ehrich et al. (2007) report that staff generally looked upon such developments unfavorably but nonetheless struggled with the fact that they had to withhold information regarding sex from a woman or couple who would later terminate a pregnancy as a means of sex selection. Scully et al. (2006) find the public in the United Kingdom to also be ambivalent about choice in the context of social sex selection. Meanwhile, in the United States, many PGD providers will not provide social sex selection even though it is permitted. In this regulatory context, however, some providers are more willing to provide information about sex in the context of using PGD for medical purposes (Kalfoglou, 2005).

By raising the dilemmas involved in actually using PGD, the social science literature has reassessed key ethical categories like 'autonomous choice' (Zeiler, 2004; Ehrich et al., 2007), individual versus social consequences (Ehrich et al., 2007, 2006), 'serious and significant risk' (Williams et al., 2007), as well as 'quality life' (Ehrich and Williams, 2010; Wahlberg, 2008). This research has found that such distinctions are rarely clear-cut in clinical practice, but must instead be actively managed within social, cultural, and institutional contexts. What makes these negotiations different from those that occur in prenatal

diagnosis is that women or couples come to PGD after having experienced firsthand the heritable disorder in question (Williams et al., 2007; Franklin and Roberts, 2006). In addition, the moral status of an embryo is often viewed as different from the moral status of a fetus (Williams et al., 2007), although in Germany this has had the somewhat ironic consequence of permitting termination while outlawing PGD (Krones and Richter, 2004).

Testing for Genetic Diseases upon Birth: Newborn Screening

Many parts of the world either offer or mandate newborn screening to detect heritable disorders that can be ameliorated if treated early in life, before symptoms appear (Raho, 2008). Newborn screening was initiated in the United States during the 1960s for the disease phenylketonuria (PKU). Here, a defect in a liver enzyme results in an inability to metabolize phenylalanine, which generally results in retardation as a consequence of accumulating toxic levels of the amino acids through normal dietary intake (Paul, 2008: p. 8). With the creation of the Guthrie test, wherein a newborn's heel is pricked so that a few drops of blood can be placed on a filter paper for an assay, newborn screening for PKU became universal in the United States between 1967 and the mid-1970s. Newborn screening continues to be mandatory today, which means that it is largely practiced without informed consent. This creates a significant amount of shock on the part of new parents who receive a positive diagnosis shortly after their child is born (Grob, 2008; Timmermans and Buchbinder, 2010). However, proponents of newborn screening for PKU have successfully argued that this problem is offset by the ability to treat afflicted newborns with diets that lack phenylalanine, thereby averting mental retardation. Mandated newborn screening continues to be justified according to this argument.

The number of disorders assessed by newborn screening has expanded significantly, a process that accelerated with the introduction of tandem mass spectrometry (MS/MS) in the 1990s. In 2005, the American College of Medical Genetics issued a report advocating for this expansion; by 2009, 44 states tested for all 29 recommended conditions while six screened for 21 of the 29 (Timmermans and Buchbinder, 2010: p. 408). While MS/MS has led to an expansion in the number of disorders being universally screened for in the United States, it has also led to an increasing number of countries legally mandating newborn screening worldwide (Raho, 2008: p. 1).

To date, much of the social science research on newborn screening has focused on its development and use within the United States. Diane Paul (2008, 1998a) historicizes the development of newborn screening in the United States, showing that its uptake for PKU was contentious amongst metabolic researchers and clinicians for a number of reasons, including lower disease incidence rates than those publicized; a lifelong (as opposed to 5-year) treatment requirement comprising expensive, difficult-to-prepare, and bland food; as well as high numbers of false positives that resulted in children being unnecessarily deprived of nutrients (Paul, 2008: p. 9, 1998a). Paul shows that patient advocacy groups as well as profit-oriented medical companies have increasingly come to play a significant role in the development of this health care policy (Paul, 2008, 1998a). In the process, Paul highlights that the admittedly worthy goal of treating newborns before they become symptomatic is not as straightforward in practice as is frequently represented by advocates.

Social science research has since empirically explored the dilemmas that arise when newborn screening is implemented. Timmermans and Buchbinder (2010) have, for example, shown that diagnosing and treating a newborn based on genetic information garnered from MS/MS are not always straightforward. Based on their ethnographic study in a California clinic, the authors explore 24 (out of a total of 55) cases in which the results of the screen were ambiguous. Timmermans and Buchbinder refer to these newborns and their parents as 'patients-in-waiting.' When test results are ambiguous, the newborn is treated as if he or she has a serious condition. This has consequences for parents. Specifically, Timmermans and Buchbinder find that, in instances when geneticists come to believe the condition is not serious due to subsequent tests and events, almost all parents nonetheless experienced sustained levels of extremely high concern for their child's health. Timmermans and Buchbinder's discussion of these dilemmas has reignited debates with proponents of newborn screening, who have seen this 'critical' perspective as a disservice to public health as opposed to an opportunity to work through the uncertainties that arise when implementing innovative medical technologies (e.g., Timmermans and Buchbinder, 2011; Watson et al., 2010).

Timmermans and Buchbinder (2010) emphasize that patients-in-waiting occupy a liminal, or ambiguous, space between normalcy and pathology in clinical medicine, which results in a lengthy period of medical surveillance. Rachel Grob (2008) follows the ways in which parents experience this liminal space outside of the clinic, in their family life. Grob conducted qualitative interviews with parents of children diagnosed with cystic fibrosis prenatally ($n = 4$), via newborn screening ($n = 16$), or after the development of symptoms ($n = 11$). Like the patients in Timmermans and Buchbinder's study, some parents in Grob's study also had to confront the ambiguity of genetic mutations that may not result in the child ever becoming symptomatic. Also like Timmermans and Buchbinder, Grob found that receiving the diagnosis via newborn screening affected the relationship between parent and child. Where Timmermans and Buchbinder focus on the concerns of illness when these worries are no longer viewed as medically founded, Grob alternatively finds that parents come to be actively involved in the medicalization of newborns. She points out that parents learn about their child in tandem with learning about cystic fibrosis (Grob, 2008: p. 1059). Receiving this information when one is in the early stages of learning how to parent led many to view physicians as the experts of their children's well-being. That said, Grob (2008: p. 1061) did find that parents over time became more confident and assertive advocates for their children.

Social science literature on newborn screening has thus emphasized the ways in which medical and scientific uncertainties surrounding the genotype–phenotype relationship become acutely felt in the practices of newborn screening, particularly by parents but also by clinicians who are uncertain of how to best treat potential patients.

Testing for Genetic Diseases Later in Life: Reproductive Decision Making

One of the rationales for using newborn screening to detect disorders that lack a cure has been to provide parents with information that will assist them with their reproductive decision making (Paul, 2008: p. 11), which will presumably have cost-saving implications for the state (Grob, 2006: p. 162). The ways in which genetic testing is used to inform reproductive decision making have similarly expanded as genetic tests are used to assess susceptibility for genetic disorders that present later in life, such as Huntington's disease (Taylor, 2004: p. 141; Konrad, 2005), breast cancer (Hallowell, 1999), and Alzheimer's disease (Chilibeck et al., 2011; Lock et al., 2006). This intersects with the development of community genetic programs that have been used to determine if couples are carriers of genetic disorders in communities that experience a high prevalence of a disease. Here decisions about whether or not to reproduce are made through the prism of susceptibility for having children with a heritable disease.

Due to a high rate of endogamy, a possible founder effect, and genetic drift, the incidence of heritable diseases is higher within Ashkenazi-Jewish communities (Prainsack and Siegal, 2006). After having four children die from Tay–Sachs, New York Rabbi Joseph Ekstein developed Dor Yeshorim, a community genetics program, in the 1960s. This is a system in which couples are assessed before marriage to find out if they are 'genetically compatible' (Prainsack and Siegal, 2006; Raz and Vizner, 2008; Wailoo and Pemberton, 2006). Here individuals are tested when they are teenagers. However, because of concerns that the entire family of a known carrier would be stigmatized, individual genetic tests are not given at this time. Rather, once a marriage between two people is being considered, Dor Yeshorim will report whether or not the couple is at risk of having children with a heritable disease. If the couple is not genetically compatible, the marriage is generally called off. Prainsack and Siegal (2006) emphasize that this testing system works because arranged marriages are commonplace, and so compatibility is assessed early in the marriage process and quite possibly before the couple has even met. They contrast this with the community testing program for thalassemia in Cyprus, where individuals learn of their own results and marriages continue despite 'genetic incompatibility' with an indication that prenatal genetic screening should be used.

In the general population, hereditary disorders are also likely to have consequences for the wider familial unit once detected (Lock et al., 2006; Finkler et al., 2003; Chilibeck et al., 2011; Konrad, 2005). This includes not only couples but also their siblings. For example, social science research has explored how couples make reproductive decisions after having a child diagnosed with a heritable disorder (Kelly, 2009; Raspberry and Skinner, 2011a,b). Both Kelly (2009) and Raspberry and Skinner (2011a,b) find that parents make a range of different decisions prenatal in this context, including not having more children, using prenatal diagnosis or PGD, as well as having additional planned and unplanned children in an unmediated manner. However, both studies found that the majority of participants decided not to have more children. Social science research has also explored if and how people disclose the presence of a heritable disorder to their wider family (Konrad, 2005). Questions regarding responsibility and autonomy are particularly acute here, as some siblings may or may not want to consider their genetic risk as part of their family-planning practices (Liao, 2009; Arribas-Ayllon et al., 2008a,b; Hallowell et al., 2006). For example, while some women would certainly want to know this information (Liao, 2009), many of the rural women that Kelly (2009) interviewed in the United States actively choose not to make the kinds of choices that such knowledge necessitates. Disclosure would have made it impossible for these women to 'choose not to choose.'

Conclusion

The ways in which a genetic disorder is understood and acted upon is informed by the social, temporal, and biological ordering of reproduction. Most of the techniques used to assess for genetic disorders are screens rather than treatments. Nonetheless, different screens are informed by and enable different kinds of reproductive trajectories. Prenatal screening and diagnosis offer women and their partners the choice to abort a fetus with a chromosomal disorder. PGD offers women and their partners the ability to select out embryos that have a heritable disease. Meanwhile, genetic testing later in life enables women and their partners to ask if they will have children at all, and if those pregnancies will be medically mediated or not. Newborn screening is the only genetic test that is associated with therapeutics. However, it nonetheless raises many uncertainties and ambivalences that social scientists have consistently found in personal and professional accounts of genetic screening.

In this context, women and their partners have had to become 'moral pioneers' (Rapp, 2000), adjudicating different responsibilities in the context of political, economic, social, cultural, historical, and religious constraints. This process has been linked with the increasing individualization, 'responsibilization,' and 'ethicalization' of subjects in neoliberalism (Helen, 2010; Rose, 2007). However, as social science research has expanded to consider genetic testing and reproductive medicine beyond urban centers in North America and Western Europe, social scientists are increasingly finding that these techniques are varyingly constituted in different places (Prainsack and Siegal, 2006; Hashiloni-Dolev and Shkedi, 2007; Kelly, 2009; Ivry, 2006; Ettore, 2002). Rural women in the United States reject the notion that one needs to know about the genetics of their fetus in order to be a responsible mother (Kelly, 2009). Japanese Ob–Gyns resist the use of prenatal genetic tests, emphasizing gestational responsibility (Ivry, 2006). And the priority of the family over the individual leads to different notions of what responsible parenting entails in Israel (Hashiloni-Dolev and Shkedi, 2007). As such, genetics, reproduction, and disease at times enact a kind of 'somatic responsibility' (Rose, 2007) rooted in the individual. But this nexus also generates what Prainsack and Siegal (2006) call 'genetic couplehood,' wherein genetic identity supersedes the individual and instead focuses on the couple in relationship to their community.

See also: Community Genetics; Eugenics as a Basis of Population Policy; Eugenics as an International Movement;

Eugenics in Asia; Eugenics in Europe; Eugenics in the Americas; Eugenics, History of; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Screening for Disease; Genetics: The New Genetics; Genomics, Ethical Issues in; Human Genome Project: History and Assessment.

Bibliography

Arribas-Ayllon, Michael, Sarangi, Srikant, Clarke, Angus, 2008a. Managing self-responsibility through other-oriented blame: family accounts of genetic testing. Social Science and Medicine 66, 1521–1532.

Arribas-Ayllon, Michael, Sarangi, Srikant, Clarke, Angus, 2008b. The micropolitics of responsibility vis-a-vis autonomy: parental accounts of childhood genetic testing and (non)disclosure. Sociology of Health and Illness 30 (2), 255–271.

Browner, C.H., Preloran, Mabel H., Casado, Maria Christina, Bass, Harold N., Walker, Ann P., 2003. Genetic counseling gone awry: miscommunication between prenatal genetic service providers and Mexican-origin clients. Social Science and Medicine 56, 1933–1946.

Chilibeck, Gillian, Lock, Margaret, Sehdev, Megha, 2011. Postgenomics, uncertain futures, and the familiarization of susceptibility genes. Social Science and Medicine 72, 1768–1775.

Colen, Shellee, 1995. Like a mother to them: stratified reproduction and West Indian childcare workers and employers in New York. In: Ginsburg, F.D., Rapp., R. (Eds.), Conceiving the New World Order: The Global Politics of Reproduction. University of California Press, Berkeley.

Devolder, Katrien, 2005. Preimplantation HLA typing: having children to save our loved ones. Journal of Medical Ethics 31, 582–586.

Duster, Troy, 1990. Backdoor to Eugenics. Routledge, New York and London.

Ehrich, Kathryn, Williams, Clare, 2010. A healthy baby: the double imperative of preimplantation genetic diagnosis. Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine 14 (1), 41–56.

Ehrich, Kathryn, Williams, Clare, Farsides, Bobbie, 2008. The embryo as moral work object: PGD/IVF staff views and experiences. Sociology of Health and Illness 30 (5), 772–787.

Ehrich, Kathryn, Williams, Clare, Farsides, Bobbie, Sandall, Jane, Scott, Rosamund, 2007. Choosing embryos: ethical complexity and relational autonomy in staff accounts of PGD. Sociology of Health and Illness 29 (7), 1091–1106.

Ehrich, Kathryn, Williams, Clare, Scott, Rosamund, Sandall, Jane, Farsides, Bobbie, 2006. Social welfare, genetic welfare? Boundary-work in the IVF/PGD clinic. Social Science and Medicine 63 (5), 1213–1224.

Ettore, E., 2002. Experts' views on prenatal screening and diagnosis in Greece, the Netherlands, England and Finland. In: Ettorre, E. (Ed.), Before Birth: Understanding Prenatal Screening. Ashgate, Aldershot, UK.

Evans, John H, 2010. Contested Reproduction: Genetic Technologies, Religion, and Public Debate. University of Chicago Press, Chicago, IL.

Farsides, Bobbie, Williams, Clare, Alderson, Priscilla, 2004. Aiming towards moral equilibrium: health care professionals' views on working within the morally contested field of antenatal screening. Journal of Medical Ethics 30, 505–509.

Finkler, Kaja, Skrzynia, Cecile, Evans, James P., 2003. The new genetics and its consequences for family, kinship, medicine and medical genetics. Social Science and Medicine 57, 403–412.

Franklin, Sarah, Roberts, Celia, 2006. Born and Made: An Ethnography of Preimplantation Genetic Diagnosis. Princeton University Press, Princeton, NJ.

Gottfreosdottie, Helga, 2009. Have you had the test? A discourse analysis of media presentation of prenatal screening in Iceland. Scandinavian Journal of Caring Sciences 24, 414–421.

Grob, Rachel, 2006. Parenting in the genomic age: the cursed blessing of newborn screening. New Genetics and Society 25 (2), 159–170.

Grob, Rachel, 2008. Is my sick child health? Is my healthy child sick?: changing parental experiences of cystic fibrosis in the age of expanded newborn screening. Social Science and Medicine 67, 1056–1064.

Habermas, Jurgen, 2003. The Future of Human Nature. Polity, Cambridge.

Hallowell, Nina, 1999. Doing the right thing: genetic risk and responsibility. In: Conrad, P., Gabe, J. (Eds.), Sociological Perspectives on the New Genetics. Blackwell, Oxford.

Hallowell, Nina, Ardenn-Jones, A., Eeles, R., Foster, C., Lucassen, A., Moynihan, C., Watson, M., 2006. Balancing autonomy and responsibility: the ethics of generating and disclosing genetic information. Journal of Medical Ethics 29, 74–83.

Hashiloni-Dolev, Yael, 2006. Between mothers, fetuses and society: reproductive genetics in the Israeli-Jewish context. NASHIM: A Journal of Jewish Women's Studies and Gender Issues 12, 129–150.

Hashiloni-Dolev, Yael, Shkedi, Shiri, 2007. On new reproductive technologies and family ethics: pre-implantation genetic diagnosis for sibling donor in Israel and Germany. Social Science and Medicine 65, 2081–2092.

Helen, Ilpo, 2010. Technics over life: risk, ethics and the existential condition in high-tech antenatal care. Economy and Society 33 (1), 28–51.

Inhorn, Marcia C., Tjornhoj-Thomsen, Tine, Goldberg, Helene, la Cour Mosegaard, Maruska (Eds.), 2009. Reconceiving the Second Sex: Men, Masculinity, and Reproduction. Berghahn, New York.

Ivry, Tsipy, 2006. At the back stage of prenatal care: Japanese Ob–Gyns negotiating prenatal diagnosis. Medical Anthropology Quarterly 20 (4), 441–468.

Ivry, Tsipy, Teman, Elly, 2008. Expectant Israeli fathers and the medicalized pregnancy: power, pragmatism and resistance. Culture, Medicine and Psychiatry 32 (3), 358–385.

Kalfoglou, Andrea L, 2005. PGD patients' and providers' attitudes to the use and regulation of preimplantation genetic diagnosis. Reproductive BioMedicine Online 11 (4), 486–496.

Keller, Evelyn Fox, 2000. The Century of the Gene. Harvard University Press, Cambridge, MA.

Keller, Evelyn Fox, 2010. The Mirage of a Space between Nature and Nurture. Duke University Press, Durham, NC.

Kelly, Suan E, 2009. Choosing not to choose: reproductive responses of parents of children with genetic conditions or impairments. Sociology of Health and Illness 31 (1), 81–97.

Kerr, A., Cunningham-Burley, S., Amos, A., 1998. Eugenics and the new genetics in Britain: examining contemporary professionals' accounts. Science, Technology, and Human Values 23 (2), 175–198.

Kerr, Anne, 2004. Genetics and Society: A Sociology of Disease. Routledge, London.

Konrad, Monica, 2005. Narrating the New Predictive Genetics: Ethics, Ethnography and Science. Cambridge University Press, Cambridge.

Krones, Tanja, Richter, Gerd, 2004. Preimplantation genetic diagnosis: European perspectives and the German situation. Journal of Medicine and Philosophy 29 (5), 623–640.

Lalor, Joan, Begley, Cecily M., Galavan, Eoin, 2009. Recasting hope: a process of adaptation following fetal anomaly diagnosis. Social Science and Medicine 68, 462–472.

Landsmen, Gail, 2003. Emplotting children's lives: developmental delay vs. disability. Social Science and Medicine 56, 1947–1960.

Liao, S.M., 2009. Is there a duty to share genetic information? Journal of Medical Ethics 35, 306–309.

Lock, Margaret, 2005. Eclipse of the gene and the return of divination. Current Anthropology 46 (S5), S47–S70.

Lock, Margaret, Freeman, Julia, Sharples, Rosemary, Lloyd, Stephanie, 2006. When it runs in the family: putting susceptibility genes in perspective. Public Understanding of Science 15 (3), 277–300.

Meister, U., Finck, C., Stobel-Richter, Y., Brahler, E., 2005. Knowledge and attitudes towards preimplantation genetic diagnosis in Germany. Human Reproduction 20 (1), 231–238.

Parens, E., Asch, A., 2003. Prenatal Testing and Disability Rights. Georgetown University Press, Washington, DC.

Paul, Diane B, 1998a. The history of newborn phenylketonuria screening in the U.S. In: Holtzman, N.A., Watson, M.S. (Eds.), Promoting Safe and Effective Genetic Testing in the United States. Final Report of the Task Force on Genetic Testing. Johns Hopkins University Press, Baltimore, MD.

Paul, Diane B, 1998b. The Politics of Heredity: Essays on Eugenics, Biomedicine, and the Nature–Nurture Debate. State University of New York Press, Albany, NY.

Paul, Diane B, 2008. Patient advocacy in newborn screening: continuities and discontinuities. American Journal of Medical Genetics Part C (Seminars in Medical Genetics) 148C, 8–14.

Pavone, Vicenzo, Arias, Flor, 2012. Beyond the geneticization thesis: the political economy of PGD/PGS in Spain. Science, Technology, and Human Values 37 (3), 235–261.

Pilnick, Alison, 2004. It's just one of the best tests that we've got at the moment: the presentation of nuchal translucency screening for fetal abnormality in pregnancy. Discourse and Society 15 (4), 451–465.

Prainsack, Barbara, Siegal, Gil, 2006. The rise of genetic couplehood? A comparative view of premarital genetic testing. BioSocieties 1, 17–36.

Press, Nancy, Browner, Carol H., 1997. Why women say yes to prenatal diagnosis. Social Science and Medicine 45, 979–989.

Raho, Joseph A, 2008. Appendix: newborn screening: an international survey. In: The Changing Moral Focus of Newborn Screening: An Ethical Analysis by the President's Council on Bioethics. The President's Council on Bioethics, Washington, D.C.

Rapp, Rayna, 2000. Testing Women, Testing the Fetus: The Social Impact of Amniocentesis in America. Routledge, New York.

Raspberry, Kelly Amanda, Skinner, Debra, 2011a. Enacting genetic responsibility: experiences of mothers who carry the fragile X gene. Sociology of Health and Illness 33 (3), 420–433.

Raspberry, Kelly Amanda, Skinner, Debra, 2011b. Negotiating desires and options: how mothers who carry the fragile X gene experience reproductive decisions. Social Science and Medicine 72, 992–998.

Raz, Aviad, 2004. Important to test, important to support: attitudes toward disability rights and prenatal diagnosis among leaders of support groups for genetic disorders in Israel. Social Science and Medicine 59, 1857–1866.

Raz, Aviad E., Vizner, Yafa, 2008. Carrier matching and collective socialization in community genetics: Dor Yeshorim and the reinforcement of stigma. Social Science and Medicine 67, 1361–1369.

Reed, Kate, 2009. It's them faulty genes again: women, men and the gendered nature of genetic responsibility in prenatal blood screening. Sociology of Health and Illness 31 (3), 343–359.

Remennick, Larissa, 2006. The quest for the perfect baby: why do Israeli women seek prenatal genetic testing? Sociology of Health and Illness 28 (1), 21–53.

Rose, Nikolas, 2007. The Politics of Life Itself: Biomedicine, Power, and Subjectivity in the Twenty-First Century. Princeton University Press, Princeton, NJ.

Scully, Jackie Leach, Banks, Sarah, Shakespeare, Tom W., 2006. Chance, choice and control: lay debate on prenatal social sex selection. Social Science and Medicine 63, 21–31.

Shakespeare, Tom, 1998. Choice and rights: eugenics, genetics and disability equality. Disability and Society 13 (5), 655–681.

Shakespeare, Tom, 1999. Losing the plot? Medical and activist discourses of contemporary genetics and disability. Sociology of Health and Illness 21 (5), 669–688.

Taussig, Karen-Sue, Rapp, Rayna, Heath, Deborah, 2003. Flexible eugenics: technologies of self in the age of genetics. In: Goodman, A.H., Heath, D., Lindee, M.S. (Eds.), Genetic Nature/Culture: Anthropology and Science beyond the Two-Culture Divide. University of California Press, Berkeley, CA.

Taylor, Sandra D, 2004. Predictive genetic testing decisions for Huntington's disease: context, appraisal and new moral imperatives. Social Science and Medicine 58, 137–149.

Timmermans, Stefan, Buchbinder, Mara, 2010. Patients-in-waiting: living between sickness and health in the genomic era. Journal of Health and Social Behavior 51 (4), 408–423.

Timmermans, Stefan, Buchbinder, Mara, 2011. Improving expanded newborn screening: a reply to Watson, Howell, and Rinaldo. Journal of Health and Social Behavior 52 (2), 279–281.

Vailly, Joelle, 2008. The expansion of abnormality and the biomedical norm: neonatal screening, prenatal diagnosis and cystic fibrosis in France. Social Science and Medicine 66, 2532–2543.

Vassy, Carine, 2006. From a genetic innovation to mass health programmes: the diffusion of Down's syndrome prenatal screening and diagnostic techniques in France. Social Science and Medicine 63, 2041–2051.

Wahlberg, Ayo, 2008. Reproductive medicine and the concept of quality. Clinical Ethics 3 (4), 189–193.

Wailoo, Keith, Pemberton, Stephen, 2006. The Troubled Dream of Genetic Medicine: Ethnicity and Innovation in Tay-Sachs, Cystic Fibrosis, and Sickle Cell Disease. The Johns Hopkins University Press, Baltimore, MD.

Watson, M.S., Howell, R. Rodney, Rinaldo, Piero, 2010. A disservice to advances in newborn genetic screening: comment on Timmermans and Buchbinder. Journal of Health and Social Behavior 52 (2), 277–278.

Weiss, Meira, 2002. The Chosen Body: The Politics of the Body in Israeli Society. Stanford University Press, Stanford, CA.

Williams, Clare, Alderson, Priscilla, Farsides, Bobbie, 2002a. Dilemmas encountered by health practitioners offering nuchal translucency screening: a qualitative case study. Prenatal Diagnosis 22, 216–220.

Williams, Clare, Alderson, Priscilla, Farsides, Bobbie, 2002b. Drawing the line in prenatal screening and testing: health practitioners' discussions. Health, Risk and Society 4 (1), 61–75.

Williams, Clare, Alderson, Priscilla, Farsides, Bobbie, 2002c. Is nondirectiveness possible within the context of antenatal screening and testing? Social Science and Medicine 54, 339–347.

Williams, Clare, Alderson, Priscilla, Farsides, Bobbie, 2002d. Too many choice? Hospital and community staff reflect on the future of prenatal screening. Social Science and Medicine 55, 743–753.

Williams, Clare, Ehrich, Kathryn, Farsides, Bobbie, Scott, Rosamund, 2007. Facilitating choice, framing choice: staff views on widening the scope of preimplantation genetic diagnosis in the UK. Social Science and Medicine 65, 1094–1105.

Williams, Clare, Sandall, Jane, Lewando-Hundt, G., Heyman, B., Spencer, K., Grellier, R., 2005. Women as moral pioneers? Experiences of first trimester antenatal screening. Social Science and Medicine 61 (9), 1983–1992.

Zeiler, Kristin, 2004. Reproductive autonomous choice – a cherished illusion? Reproductive autonomy examined in the context of preimplantation genetic diagnosis. Medicine, Health Care and Philosophy 7, 175–183.

Genetics: Legal Aspects

Shawn HE Harmon, University of Edinburgh, Edinburgh, United Kingdom

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The legal challenges of emerging genetic practices are numerous and fundamental. Not only has genetics destabilized concepts of personhood, community, health, and disease, but also it has complicated the regulation of many fields, throwing into question settled concepts and mechanisms. This article highlights a range of live legal issues and then focuses on issues found in four settings: newly emerging research models, commercial practices associated with genetics, food security, and the practice of law. It concludes by suggesting that we are little closer to determining the appropriate role and form of law in relation to genetics than we were a decade ago.

Introduction

Genetics has prompted dramatic shifts in diverse fields such as human reproduction, medical diagnostics, epidemiology, toxicology, agriculture, environmental rehabilitation, and more. In doing so, it has thrown into question traditional definitions of personhood, individuality, community, health, and disease, and traditional assessments of risks (to individual or public health or the environment) (Grodsky, 2005). With respect to disease, for example, genetics, which relies on the idea of 'biomarkers' (i.e., it approaches the study of the genome by looking at the placement and order of gene sequences), has suggested the existence of continuums from healthy to diseased, and from exposure to disease, which have become important to reproductive, developmental, neurological, immunological, and other health impairments, particularly those relating to cancer and environmental exposures (Perera, 1996; DeCaprio, 1997; Marchant, 2003).

The rapid evolution of genetics has also, importantly, given rise to ethico-legal concerns, some of them new, many of them well-rehearsed. From a legal perspective, some questions that have been much debated, though hardly settled, include the following:

- What is the most appropriate legal consequence of characterizing the genome as the common heritage of humanity? How do we reconcile the fact that an individual's genetic information is both intimately private and shared across family and (sometimes) community? For more on these questions, see Knoppers (2005) and Ossorio (2007).
- What are the appropriate limits of science? Where do we set the boundaries of acceptable research given our ambitions, our rudimentary understanding of biological mechanisms, and our relatively powerful technical capabilities (to clone animals and human tissue or to alter genetic germlines to both plants and animals)? For more on these questions, see Jasanoff (1987), Clayton (2003), and Wainwright et al. (2006).
- How do we balance the need (and desire) to facilitate science and technological innovation with the need to protect individual rights, cherished social values, and traditional cultural norms? What is the appropriate role of the scientific regulatory regime? How does it optimally interact with other regimes? How do we avoid regulatory overburden and complexity? For more on these questions, see House of Lords (2009), Academy of Medical Sciences (2011).
- How do we integrate new genetic understandings and the more precise data we are generating into existing science and risk-based assessments, whether environmental, public health, or other? How do we cope with the detail and complexity that are being generated? What does our greater understanding of the susceptibility of population subgroups mean for standard setting? For more on these questions, see Grodsky (2005) and Caulfield et al. (2012). In the US context, see American Lung Association v EPA (1998).
- How and to what extent should new genetic tools that assess predisposition (and the treatment options to which they give rise, which are often not yet genetic, e.g., abortion) be integrated into public health systems? How effective are existing genetic counseling services? How can the often expensive technologies to which genetics research gives rise be made more accessible around the world as part of our moral and human rights–based responsibilities to share in scientific knowledge and preserve the health of all humans? How practical and realistic are ambitions for personalized medicine, or even stratified medicine? In short, how do we cope with genetic knowledge and technologies at the clinical level? For more on these questions, see Acheson et al. (2005), Drahos (2007), and Mittra and Tait (2012).

As should be clear from this nonexhaustive list, the range of concerns with which the law is interested is broad, but all of them implicate to varying degrees notions of

- risk (to individual health, community identity, and environmental well-being);
- choice (of patients, health care providers, and consumers);
- values (dignity, autonomy, sanctity, solidarity, equality, and democracy); and
- rights (to tissue, data, patents, privacy, and decisional freedom).

Generally, in each category, the law would prefer certainty, which is a fundamental principle of the law (Alexy, 2004). However, while genetics does offer a previously unimagined precision when it comes to understanding the structure of biology (if not the diagnosis and treatment of conditions and

complex diseases), it has proven rather more destabilizing when it comes to the law. The result has been that how we permit genetic knowledge and innovations to be deployed remains open and contested, with most commentators encouraging caution, engagement, attention to context, and preservation of existing rights.

Obviously, the full range of legal issues could not possibly be addressed in a single article. This article, therefore, focuses on four issues that directly and deeply implicate the law, and that might be considered consequent of the relative maturation of genetics – a maturation that has seen genetic science taken up in applied fields despite, it must be admitted, the persistent and acknowledged uncertainty surrounding much of our knowledge. This article examines sociolegal concerns arising from the use or interaction of genetics with (1) new scientific methods, (2) the commercialization of science, (3) food production and security, and (4) the operation of law.

New Scientific Methods and Approaches

To begin with the science itself, genetics has given rise to new research models. Population biobanking and bioinformatics-reliant genome-wide association studies are just two examples of how the practice of science has transformed as a result of advances in genetics (House of Lords, 2009). Biobanks are collections of tissue and data for future use by diverse researchers, some of whom will be extraterritorial; they are complex research undertakings embedded in complex research networks (Kaye et al., 2012), and they raise critical and novel questions as to the following (Harmon, 2008):

- Appropriate recruitment and consent procedures
- Appropriate stewardship responsibilities and standards of conduct
- Appropriate long-term maintenance procedures for the resource
- Appropriate access criteria to the resource
- Appropriate monitoring of use and governance of the resource
- Appropriate benefits to participants, including feedback of incidental findings

Exhibiting structural diversity, purposive uncertainty, and temporal longevity, they are, by their nature, exercises in the unknown, and that makes them particularly challenging for law (Laurie, 2011). As such, key questions include the following:

- How do we regulate these new scientific models so that they are sustainable?
- How and where does law fit into that regulatory regime?

While some jurisdictions have opted for a legislative approach (e.g., Estonia, Latvia, Norway, Sweden, and Spain), many have not. The UK Biobank, which is subject to the Human Tissue Act 2004, has adopted an "ethics+" approach (Laurie, 2011). This means that legal standards, found in a range of instruments, are supplemented by an evolving ethical and oversight framework that goes beyond the law and contains a public participation component.

Of course, biobanking is not the only novel scientific method (or socioscientific undertaking) that has arisen. Assisted by innovations in high-speed gene-profiling technologies (e.g., miniaturized semiconductor chips, laser and robotics engineering, and DNA microarrays) that allow for comprehensive analysis of gene expression at unprecedented speeds and scales, a variety of new approaches to the study of genetics are being explored. These so-called postgenome approaches, such as systems biology, gene ecology, and epigenetics, are more organismic and relational in nature (Wynn, 2005; Powell and Dupré, 2009). They shift the analysis from intrinsic elements like functional outcomes of genetic sequences and their mutations to the interactions between systems; they have thus been described as the biology of the in between (Van de Vujver, 2009). Such new approaches trigger new legal questions and controversies, and reignite old ones.

For example, epigenetics adopts a more complex and contextual approach to gene expression than traditional genetics (Waddington, 1942). Appreciating that the body is embedded in different environmental circumstances, epigenetics suggests that these circumstances can be inscribed on gene expression (i.e., not on the structure of the sequence but on how the gene is expressed physiologically) (Kuzawa and Sweet, 2009; Barrès et al., 2012). It holds that heritable characteristics can arise from marks on the genome, or the scars of experience that influence gene expression. The temporalities considered to be relevant are

1. evolutionary (across generations);
2. developmental (early origins of adult disease);
3. life course (environmental exposure to conditions); and
4. real time (day-to-day).

This epigenetic model blurs the boundary between the biological and the social in which it exists. In doing so, it reemphasizes and intensifies the complexity and uncertainty that traditional genetics has already introduced (Niewöhner, 2011). A major scientific concern is that this complexity could unravel much of what we 'know' from past scientific methods or frames about biology and nature. This, in turn, could undermine the legal rules and presumptions that rely on scientific understandings. For example, as awareness of the role of epigenetics in normal and abnormal cell development grows, so has the realization of the potential importance that epigenetic disruptions (such as toxicity and environmental regulatory priorities) can have for human well-being, now and into the future (Dolinoy et al., 2007; Jirtle and Skinner, 2007). Shifts in accepted knowledge will affect a range of legislated standards, including those around air and water purity, toxins exposure, food quality, and more; swathes of legislative regimes may have to be revisited with respect to the thresholds of exposure and standards of conduct they permit (In the United States, implicated legislation includes the Clean Air Act, 42 USC §§ 7401, et seq.; the Clean Water Act, 33 USC §§ 1251, et seq.; the Toxic Substances Control Act, 15 USC §§ 2601, et seq.; the Comprehensive Environmental Response, Compensation and Liability Act, 42 USC §§ 9601, et seq., the Resource Conservation and Recovery Act, 42 USC §§ 6901, et seq.; the Safe Drinking Water Act, 42 USC §§300f, et seq.; the Federal Insecticide, Fungicide and Rodenticide Act, 77 USC §§ 136(a), et seq.; and the Occupational Health and Safety Act, 29 USC §§

651–78 (2000). In addition, FDA authorizations of food products and medicines under the Federal Food, Drug and Cosmetic Act, 21 USC § 301 et seq.; are implicated, particularly requirements for efficacy of drugs.).

Additionally, epigenetics will continue the trend of generating masses of information about an individual's likelihood to develop health problems in the future and transmit them to children. This means that more sensitive health information will be generated and will, by necessity, be shared across projects and health systems; it means that information disclosures will become more common and more diffusive, particularly as individuals claim rights of access to others' health information that they see as relevant to them (Laurie, 2002). Because of the social stigma that attaches to some conditions (and to ill health more generally), concerns about privacy, confidentiality, and the sufficiency of existing protections abound (Rothstein, 1997; Andorno, 2004). This will be a particular challenge for data curators charged with information protection and access duties under data protection legislation (In Europe, data management is governed by EU Directive on Data Protection 1995/46/EC, which has been transposed into Member State law. In the United Kingdom, the relevant statute is the Data Protection Act 1998.).

Finally, epigenetics raises the specter that we might (consciously or unconsciously) naturalize individual or community social situations through the creation of a biological frame for those situations (Busby and Martin, 2006; Niewöhner, 2011). In examining the reframing of food as a form of metabolic environmental exposure, Landecker argues that nutritional epigenetics, in addition to destabilizing our understanding of food, contributes to 'a molecular politics of eating' that implicates notions of personal and social responsibility for the future health of individuals and populations and the governmental stewardship of food (Landecker, 2011: 168). In other words, if we construct and institutionalize science to reflect or bolster social conditions, then we risk perpetuating those conditions and neglecting the (difficult) political, social, ethical, and legal solutions that we should be pursuing as part of our moral or legal responsibilities (such as making healthy food accessible, providing quality education, etc.).

These concerns do not exhaust the legal implications of epigenetics, but rather demonstrate that concerns over genetic determinism, reductionism, exceptionalism (which asserts that the occurrence of specific conditions or diseases is directly caused by specific genes without more, and that we are little more than the sum of our genes, which, from birth, describe our fate), and the medicalization of human conditions (which describes how we have turned all manner of social conditions into medical conditions to be remedied) remain very much alive despite our growing understanding of, and aversion to, them (Mittra, 2007). Therefore, we have to remain vigilant to issues such as characterization, issues that were highlighted by genetics and become more pronounced with further developments.

Commercialization of Genetic and Cellular Inventions

Genetic science is closely allied with commercial interests; indeed, the Human Genome Project was in a race against a commercial competitor to map the human genome. Given the relative infancy of the science and the relatively modest range of practical health care products to which it has so far given rise, commercial activity has concentrated on compiling intellectual property portfolios relating to research tools and techniques and to diagnostics. That is, research entities have focused on obtaining patents on genes, gene sequences, processes for culturing cells, and so on. This, of course, raises questions about the propriety of patenting genes and the research or diagnostic products derived from living human material (not to mention questions about how genetic and cellular products fit into existing regulatory frameworks applicable to the market authorization of medical devices and medicinal therapies).

According to the dominant liberal dogma, patents encourage and promote innovation, competition, and economic growth, so the expansion of patenting is good not only for the economy but also for science (European Patent Office [EPO], 2012). While patenting is governed by domestic laws, the field is very much harmonized internationally. European signatories must conform to the European Patent Convention (1973), and members of the World Trade Organization (WTO; 157 members by October 2012) must conform to the TRIPS Agreement (1994). Generally, patents afford patentees a nationally bounded, exclusive, 20-year monopoly to exploit an invention in that jurisdiction in exchange for public disclosure of the invention. Many inventions are patented in multiple countries. To be patentable, an invention must meet certain criteria, namely

- novelty;
- nonobviousness (contains an inventive step that takes it beyond the state of the art); and
- utility (has some industrial application).

In addition, the disclosure of the invention in the patent application must be such that an individual conversant in the field could replicate it. Importantly, a mere discovery of nature, or of a naturally occurring substance or phenomenon, is not patentable, nor, at least in Europe, is anything that contravenes morality or *ordre public*.

Diamond v Chakrabarty (1980) is attributed with extending patentability to living products and, therefore, to genes. In that case, the US Supreme Court held that "anything under the sun made by man" is patentable. The TRIPS Agreement (1994) dispelled any lingering doubts as to whether patents involving living material were patentable when it stipulated, in Article 27, that patents shall be available for any inventions in all fields of technology. A similar approach was adopted in the EU's Biotechnology Patenting Directive 1998/44/EC. Since these clarifications, genetic science has been the subject of frantic commercial activity, with thousands of gene patents having been issued around the world. By 2000, there were some 4200 gene patents worldwide (Jensen and Murray, 2005), and, by 2005, the United States had issued some 40 000 DNA-related patents (Torrence, 2010).

Given the significant wealth at stake (e.g., the global biotech industry is estimated as having an annual revenue of US$228.6 billion; IBISWorld, 2012), it is not surprising that numerous patents have been challenged as stakeholders try to determine how the patenting criteria will operate in the genetics field. In

the United Kingdom, Genentech Inc.'s Patent (1989) and Kirin-Amgen Inc. v Hoechst Marion Roussel Ltd. (2004) confirmed that isolated gene sequences are patentable inventions so long as the inventor has demonstrated how it can be usefully employed. In Biogen v Madeva (1997), the House of Lords held that a product produced by different processes from an existing product is not novel, and that overbroad claims can be fatal. In Harvard College v Canada (2002), the Canadian Supreme Court offered guidance on the sufficiency of disclosure in patent claims. In the United States, Amgen Inc. v Chagai (1991) and Fiers v Revel (1993) confirmed gene patentability and held that claims must not be too broad or generic; the gene must be chemically isolated and specific claims as to application must be made.

Recently, however, the general position of patentability has been thrown into some doubt by Association for Molecular Pathology v USPTO (2011), wherein the US Federal Circuit Court held that isolation of a gene does not change it from being a mere discovery of a natural phenomenon. The Court of Appeal reversed this decision, but in the much-criticized case of Mayo Collaborative Services. v. Prometheus Laboratories (2012), the US Supreme Court offered new guidance on the natural phenomenon doctrine that could limit gene patentability, and it remanded the case back to the Federal Court. These cases could, arguably, be seen as a tentative retreat from cases such as In re Deuel (1995) and Eli Lilly v Human Genome Sciences (2011), US and UK cases respectively, which were characterized as victories for the biotechnology industry over patent law requirements (Ducour, 1996).

Aside from these technical conflicts between competitors, gene patents have been challenged on a more fundamental basis. Arguments against so-called life patents include the following:

- They are contrary to the idea of the genome as the common heritage of humanity (i.e., as part of the commons that should be open to all to use; it is a global public good).
- Genes are natural products, so these patents are mere discoveries despite the fact that they have been purified and isolated by human ingenuity (a process that is no longer novel).
- Gene patents confer overly broad protections for largely unproven applications over things (DNA sequences) that have little utility on their own.
- Gene patents unduly instrumentalize humans, particularly vulnerable early-stage humans, and are therefore contrary to morality.

Opposition grounded on these arguments, which are value based, has led to legal challenges, particularly in Europe.

First, EU Member States challenged the EU Biotech Directive on both technical and substantive grounds; neither Netherlands v European Parliament and Council (2001) nor Commission of the European Communities v Italy (2005) was successful, but both elucidated important elements of European law (Plomer, 2012). Next, using a patenting procedure unique to Europe, stakeholders challenged the patenting of inventions the base material for which originated in human embryos and their destruction. Two EPO cases, EDINBURGH (2002) and WARF (2008), offered, for the first time, a more expansive view of the exclusions to patentability, giving patenting antagonists hope. More recently, in Brüstle v Greenpeace (2011), the European Court of Justice held that inventions relying on embryonic stem cells derived from methods destructive to the embryo are not patentable because they are contrary to public morality. In doing so, and contrary to common scientific views, it extended the definition of embryo to that collection of cells that exists from conception, thereby limiting the possibility of obtaining patents in the embryonic stem cell research setting. These value-based cases could be seen as victories for dignity- and rights-based advocates who view any interference with the building blocks of life with fear or skepticism. The bottom line is that the field is still very much contested and evolving.

Genetically Modified Food

Nutritional and environmental awareness are growing social phenomena, and food security is a growing legal concern globally as arable land diminishes and populations grow. The development of genetically modified organisms (GMOs) and genetically modified foods (GMFs) thus is a focus for a range of sociolegal issues.

GMOs generally trigger concerns over environmental risks, and legal disputes relating thereto have been fought. Very basically, some stakeholders are concerned about GMO contamination of non-GMO crops, fields, and habitat. Though few cases have been prosecuted, many incidents of contamination have been reported, and it is generally agreed that fines for infringements are derisory. As in the pharmaceutical setting, the conduct of large corporations is often viewed suspiciously, a disposition encouraged by their introduction of characteristics such as sterility into traditionally sustainable products that are needed by farmers often operating on very fine margins.

For example, Monsanto developed so-called terminator seeds, seeds with genes that render second-generation seeds sterile, thereby tying farmers to the corporate supplier. A seminal case is Monsanto Canada Inc. v Schmeiser (2004), in which the fields of a canola farmer were infected with his neighbor's genetically modified canola produced and patented by Monsanto. Upon discovery, he was sued by Monsanto for patent infringement. The Supreme Court of Canada found that Schmeiser was indeed using seeds that he had neither purchased nor licensed. Nonetheless, it held that he had made no excessive profit and so need not pay damages to Monsanto or the company's (hefty) legal costs. *Monsanto* highlights not only the weakened position of the farmer in this new agricultural order, a matter over which the law ought to be deeply concerned, but also the potential consequences of the superior survivability of GMOs, which gives them resilient, weed-like characteristics that could endanger the survivability of other plants.

These factors, together with the more fundamental concerns that stakeholders have with organisms that have been human altered at the genetic level, have led to calls for strong regulation of GMOs and GMFs. The US approach, which relies on the Department of Agriculture (USDA), the Food and Drug Administration (FDA), and the Environmental Protection Agency (EPA), can be described as minimalist, and is generally

supported by international institutions such as the WTO and the Food and Agricultural Organization (FAO). By contrast, and more in keeping with many developing countries, the European approach is more precautionary and interventionist. It has adopted instruments addressing the risk assessment of GMOs, the release of GMOs, and the import and sale of GMOs and GMFs (and GM animals) (See EU Directive on Genetically Modified Micro-Organisms 90/219/EC, EU Directive on Deliberate Release of GMOs 2001/18/EC, Regulation (EC) No. 1829/2003 on GMFs and Feed, and Regulation (EC) 1946/2003 on Transboundary Movements of GMOs.). This regulation is informed by the idea of advanced informed consent as articulated in the Cartagena Protocol on Biosafety (2000) and the precautionary principle (which asserts that, where there are threats of serious damage beyond a *de minimus* range from a given activity, lack of full scientific knowledge about, or proof of, those threats must not be used as a reason to avoid or postpone cost-effective measures to minimize or avoid them; Rio Declaration, 1992; Clarke, 2005). As one might imagine, as with other areas of genetics, regulatory burden and regulation-driven cost are issues for operators in the field.

Shifting from cash crops to dinner tables, strong public resistance to GMFs has arisen. This resistance is fueled by all or some of the following factors (Wagner et al., 1998):

- Revulsion toward GMFs (leading to unhelpful caricatures such as Frankenfoods)
- Inaccurate understandings about GMOs
- Legitimate concerns around issues such as novel allergenic proteins and zoonosis (the transfer of animal viruses and diseases to humans)

British and European responses to GMFs have been particularly vociferous (Gaskell et al., 1999), even implicating the criminal law. In the United Kingdom, for example, in DPP v Bayer (2003), an individual who believed that a crop of GM maize was dangerous to people and to nearby fields was convicted with aggravated trespass for entering onto land and disrupting the crop.

Because many people wish to avoid GMFs, the content and quality of food labels have become important legal issues. Unfortunately, the Codex Alimentarius, which ought to offer leadership in this regard, has failed to do so. Again, the United States takes a hands-off approach, and a 2004 report fell short of calling for GM labeling but did recommend greater scrutiny of GMFs, including improved postmarket surveillance (National Research Council, 2004). In Europe, GMFs and food labeling are governed by EU Regulation (EC) No. 258/97 on Novel Foods, and Regulation (EC) No. 1830/2003 on Traceability and Labelling of GMOs. Basically, they require labels to declare GMF status if a foodstuff contains more than a stipulated percentage of GM material.

The above concerns have given rise to the idea of "food forensics"; the practice of applying science to determine the extent and nature of genetic modification in any given food substance (Heller, 2006). This will become more important as labeling requirements tighten and GMFs become more common. Indeed, as foods begin to exhibit medicinal qualities (like the GM rice that was claimed to inhibit hay fever), their classification as food is called into question altogether. Such products will require food and pharmaceutical regulatory institutions to jointly determine which regulatory pathway is more appropriate. This will necessitate a good deal more joined-upness than currently exists.

All told, GMFs represent a particularly controversial example of how genetics has transformed a series of sectors that are fundamental to human existence. It has changed how food, which we all need, is manipulated, discussed, and protested. Food is now considered and contested on a molecular level; its molecular structure or genetic constitution is highlighted by innovations in science because it is partly human-made, and consumers wonder about the effect of those (artificially introduced) genes once digested (Landecker, 2011). Thus, in addition to giving rise to legal issues of environmental safety, biodiversity, and the ownership of agricultural production, GMFs give rise to the more social phenomenon of thinking of, visualizing, and controlling food as molecules that interact with our molecules (Landecker, 2011).

Genetics and the Operation of Law

The previous sections have focused on how genetics has given rise to new practices (research, commercial, and agricultural) to which the law must respond. This section focuses on how genetics can and is (1) eroding fundamental rights, (2) challenging long-held legal doctrines, and (3) being used by the law (or in legal proceedings and peoples' interactions with the state). Essentially, as will be shown, genetics is forcing us to revisit how we ensure that rights are enjoyed and how we conceptualize important legal concepts, and it is contributing to the reshaping of the practice of law and how legal conflicts are decided.

With respect to the first issue, the enjoyment of fundamental rights, concerns abound that genetic science might dilute, wither, or destroy existing rights and interests. Genetics has unquestionably given rise to new forms of categorizing people (according to genetic susceptibility), and this has, in turn, unquestionably given rise to new forms of discrimination: so-called genetic discrimination, which is differential treatment of asymptomatic individuals or their relatives on the basis of their presumed susceptibility to known diseases or conditions as extrapolated from genetic differences assumed or demonstrated by predictive genetic testing. While our tendency toward genetic determinism has been much lamented (Greely, 2001; Curley and Caperna, 2003), there is ample evidence that genetic discrimination has been widely experienced, particularly in insurance and employment (Mittra, 2007; Taylor et al., 2008).

Drawing on the values of dignity and autonomy, and on equality, privacy, and other rights, stakeholders have challenged the idea that genetic susceptibility is an appropriate ground on which to limit access to benefits (Rothstein and Anderlik, 2001; Mayor, 2003) and lawmakers have responded. With respect to insurance, they have either adopted specific statutes or agreed moratoria on the use of genetic information by insurers (Morrison, 2005; Lemmons, 2010). With respect to work, legislation has also been adopted protecting employment rights; in the United States, note the Genetic Information Non-Discrimination Act of 2008. At the same time, questions have been raised as to whether genetic susceptibility might

constitute a disability entitling individuals to make certain claims to benefits. In the United States, this issue was highlighted by Save Our Summers v Washington Department of Ecology (1999) without being settled.

Given the international nature of science, it is unsurprising (and indeed quite helpful) that leading international organizations have offered some guidance on how genetics might interact with rights. Indeed, many general and specific international legal instruments aimed at articulating norms and setting standards for the new genomic age have been fashioned, many of them intended to serve as models for national legislators drafting their own laws. Examples include UNESCO's Universal Declaration on Human Rights and the Human Genome (1997), International Declaration on Human Genetic Data (2003), and Universal Declaration on Human Rights and Bioethics (2005); the Council of Europe's Convention on Biomedicine and Human Rights (1997) and its related protocols; and instruments offered by entities such as the World Health Organization and the Organization for Economic Cooperation and Development.

With respect to the second issue, alteration of long-standing legal concepts, the quintessential example is property. It has long been the position in law that property cannot be held in the undivided human body, whether living or dead. Universal antislavery laws negate the former, and a range of instruments have resulted in the position that corpses can be possessed only for purposes of disposal. Precedents to this effect stretch back centuries: Haynes Case (1614), R v Lynn (1788), and Williams v Williams (1882). More recently, in R v Bentham (2005), the House of Lords held that a person does not 'possess' his body or any part of it, though excised body parts can be owned by the originator or a third party if skill has been applied (such as dissection or preservation) that gives the body part new properties. In this regard, Doodeward v Spence (1908), Moore v Regents of the University of California (1990), and R v Kelly (1998), Australian, American, and English cases respectively, are pertinent. With the rise of genetics and the fantastic economic value that potentially rests in cells and DNA, individuals are now claiming ownership in their bodies and bodily products. Lawyers, however, are unsure whether to characterize DNA as tangible property, intellectual property, or something else for the purpose of assigning control and entitlements to benefit, a fact that has been described as worrying (Foster, 2003). And as with many of the above issues, positions are shifting, a fact demonstrated by Yearworth and Others v North Bristol NHS Trust (2009), a UK Court of Appeal case that extended property rights in semen to its originators. For more on this, see Harmon (2010).

With respect to the third issue, reliance on new genetic knowledge and tools is expanding dramatically, making the jobs of lawyers, judges, and civil servants (e.g., border agents and police officers) more complicated and difficult insofar as they must be aware of how genetics can be used across a field and within any given case. Once that threshold is crossed, there are issues around actually obtaining samples for examinations; in many jurisdictions, individuals are entitled to rights of physical inviolability and to not be a witness against themselves. Once access is addressed, the reception of this evidence into legal proceedings, whether civil or criminal, must be managed, and judicial (and many administrative) proceedings have very specific methods for receiving technical evidence and accepting something as fact. Curiously, while much genetic science deals only with probabilities, it has been largely readily accepted by courts and the public (Cadlish, 2010). Indeed, scientific evidence presented by experts is sometimes viewed as more compelling than other, more traditional forms of evidence (such as witness observations), and popular portrayals of the use of DNA and science-derived evidence bolster this view.

Table 1 Applications of genetics in different legal fields

Areas of law	*Application*
Criminal	Matching tissue or forensic samples
Humanitarian	Determining time of death
	Determining or confirming identity
Immigration	Determining paternity and fidelity
Human rights	Determining (familial) relationship
Family	
Wills and estates	
Tort (privacy)	
Tort (negligence; medical)	Assessing causation for claimed injuries
Trade	Tracing products and materials
Environmental	Testing risks and risk management strategies
Sports	Testing for performance enhancements

Table 1 lists just some of the legal fields, both public and private, that are increasingly relying on genetic science to perform their functions effectively.

The difficulties are exemplified by a very brief look at just one field: Actors in the environmental field are increasingly confronted with new science-based challenges to regulatory decisions and decision-making processes, and new understandings and complexities will have to be taken into account when making or justifying policy and judicial decisions. Genetics is thus forcing a revisitation (and sometimes an alteration) of (1) decisional criteria, (2) permitted standards of conduct, and (3) entitlements to rights and/or benefits because it is constantly redefining and refining what we consider to be risks, and much of our environmental (as well as our medical and public health) law and policy relies on the idea of risk. Note the American Thoracic Society's recognition that evaluations of adverse health effects might shift its evidentiary reliance from clinical symptoms to biomarkers (American Thoracic Society, 2000), and note Chlorine Chemistry Council v EPA (2000), wherein the US District Court required the EPA to abandon its traditional assumptions and defaults for assessing carcinogenic properties (of chloroform) in favor of "the best available peer-reviewed science." Grodsky (2005: 240) has argued,

> The genomic revolution is likely to invigorate ongoing debates about the relative merits of "science-based"and "science-blind" regulatory approaches. Importantly, this fundamental debate may help reveal the structural elements of new approaches that could be considered in the future.

The author goes on to argue, quite rightly, that, because we do not fully understand the nature of future challenges or the solutions that science may provide, regulatory approaches that

encourage the generation, evaluation, and incorporation of new information are most useful. (So, bad news if you are a technophobe.)

Conclusions

This article has demonstrated that advances in genetic knowledge, and new approaches to studying genes, diseases, and the human condition, have given rise to difficult legal issues with respect to

- regulating the conduct of science and its commercialization;
- making decisions and setting standards in fields that rely on science (and scientifically grounded notions of risk);
- protecting and revising established rights (such as privacy and property rights); and
- prosecuting and defending claims, both civil and criminal, through the law.

Fundamentally, we remain unsure what role we want the law to play (or think it can play) in the development of genetic technologies, a matter identified as critical over a decade ago (Black, 1998). More practically, despite the now decades-long use of genetics in some legal or policy fields, we are still struggling with issues such as technical transparency, standardization of lab practices (particularly in the criminal law context), and challenges to professional and scientific competence.

See also: Biobanking: Ethical Issues; Bioethics in the Post-genomic Era; Bioethics: Genetics and Genomics; Direct-to-Consumer Personal Genetic Testing; Gene–Environment Interactions in Health and Well-being; Genetics and Forensics; Genetics and Society; Genomics, Ethical Issues in; Globalization and Religion; Human Behavioral Ecology; Human Genome Project: History and Assessment; Law and Risk; New Genetics and Race; Patenting in the Biological Sciences; Race: Genetic Aspects; Risk, Sociology and Politics of.

Bibliography

Academy of Medical Sciences, 2011. A New Pathway for the Regulation and Governance of Health Research. AMS, London.
Acheson, Louis, Strange, Kurt, Zyzanski, Stephen, 2005. Clinical genetics issue encountered by family physicians. Genetics in Medicine 7, 501–508.
Alexy, Robert, 2004. The nature of legal philosophy. Ratio Juris 17, 156–167.
American Thoracic Society, 2000. What constitutes an adverse health effect of air pollution? American Journal of Respiratory and Critical Care Medicine 161, 665–673.
Andorno, Roberto, 2004. The right not to know: an autonomy based approached. Journal of Medical Ethics 30, 435–439.
Barrès, Romain, Yan, Jie, Egan, Brendan, et al., 2012. Acute exercise remodels promoter methylation in human skeletal muscle. Cell Metabolism 15, 405–411.
Black, Julia, 1998. Regulation as facilitation: negotiating the genetic revolution. In: Brownsword, Roger, Cornish, William, Llewelyn, Margaret (Eds.), Law and Human Genetics: Regulating a Revolution. Hart, Oxford, pp. 29–68.
Busby, Helen, Martin, Paul, 2006. Biobanks, national identity and imagined communities: the case of UK Biobank. Science as Culture 15, 237–251.
Caulfield, Tim, Harmon, Shawn, Joly, Yann, 2012. Open science versus commercialisation: a modern research conflict? Genome Medicine 4, 17.
Clarke, Steve, 2005. Future technologies, dystopian futures and the precautionary principle. Ethics & IT 7, 121–126.
Clayton, Ellen, 2003. Ethical, legal, and social implications of genomic medicine. New England Journal of Medicine 349, 562–569.
Curley Jr., Robert, Caperna, Lisa, 2003. The brave new world is here: privacy issues and the Human Genome Project. Defense Counsel Journal 70, 22–35.
DeCaprio, Anthony, 1997. Biomarkers: coming of age for environmental health and risk assessment. Environmental Science and Technology 31, 1837–1848.
Dolinoy, Dana, Weidman, Jennifer, Jirtle, Randy, 2007. Epigenetic gene regulation: linking early developmental environment to adult disease. Reproductive Toxicology 23, 297–307.
Drahos, Peter, 2007. Four lessons for developing countries from the trade negotiations over access to medicines. Liverpool Law Review 28, 11–39.
Ducour, Phillippe, 1996. In re Deuel: biotechnology industry v patent law. European Intellectual Property Review 18, 35–46.
European Patent Office, 2001. Mission Statement. Available at: www.european-patent-office.org/epo/pubs/brochure/general/e/mission_e.htm.
Foster, Charles, 2003. Current issues in the law of genetics. New Law Journal 9, 29.
Gaskell, George, Bauer, Martin, Durant, John, Allum, Nicholas, 1999. Worlds apart? The reception of genetically modified foods in Europe and the U.S. Science 285, 384–387.
Greely, Henry, 2001. Genotype discrimination: the complex case for some legislative protection. University of Pennsylvania Law Review 149, 1483–1505.
Grodsky, Jamie, 2005. Genetics and environmental law: redefining public health. California Law Review 93, 171–270.
Harmon, Shawn, 2006. From engagement to re-engagement: the expression of moral values in patenting proceedings, present and future. European Law Review 31, 642–666.
Harmon, Shawn, 2008. Semantic, pedantic or paradigm shift? Recruitment, retention and property in modern population biobanking. European Journal of Health Law 16, 27–43.
Harmon, Shawn, 2010. *Yearworth v North Bristol NHS Trust*: a property/medical case of uncertain significance. Medicine, Health Care and Philosophy 13, 343–350.
Harmon, Shawn, Laurie, Graeme, Courtney, Aidan, 2012. Dignity, plurality and patentability: the unfinished story of Brüstle v Greenpeace eV. European Law Review.
Heller, Knut, 2006. Genetically Engineered Food: Methods and Detection. Wiley-VCH, Weinheim.
House of Lords Science and Technology Committee, 2009. 2nd Report of Session 2008–09: Genomic Medicine. London: Stationery Office.
IBISWorld, 2012. Global Biotechnology Industry Market Research Report. Available from: http://www.prweb.com/pdfdownload/10037492.pdf.
Jasanoff, Sheila, 1987. Contested boundaries in policy-relevant science. Social Studies of Science 17, 195–230.
Jensen, Kyle, Murray, Fiona, 2005. Intellectual property landscape of the human genome. Science 310, 239–240.
Jirtle, Randy, Skinner, Michael, 2007. Environmental epigenomics and disease susceptibility. Nature Reviews Genetics 8, 253–262.
Kaye, Jane, Gibbons, Susan, Heeney, Catherine, et al. (Eds.), 2012. Governing Biobanks. Hart Publishing, Oxford.
Knoppers, Bartha, 2005. Of genomics and public health: building public "goods"? CMAJ 173, 1185–1186.
Kuzawa, Chris, Sweet, Elizabeth, 2009. Epigenetics and the embodiment of race: developmental origins of US racial disparities in cardiovascular health. American Journal of Human Biology 21, 2–15.
Landecker, Hannah, 2011. Food as exposure: nutritional epigenetics and the new metabolism. BioSocieties 6, 169–194.
Laurie, Graeme, 2002. Genetic Privacy: A Challenge to Medical-legal Norms. CUP, Cambridge.
Laurie, Graeme, 2011. Reflexive governance in biobanking: on the value of policy led approaches and the need to recognise the limits of law. Human Genetics 130, 347–356.
Lemmons, Trudo, 2010. Selective justice, genetic discrimination, and insurance: should we single out genes in our laws? McGill Law Journal 45, 347–412.
Marchant, Gary, 2003. Genomics and toxic substances: part II genetic susceptibility to environmental agents. Environmental Law Reporter 33, 10641–10667.
Mayor, Susan, 2003. Report warns of danger of genetic discrimination in the workplace. BMJ 327, 702.
Mittra, James, 2007. Predictive genetic information and access to life assurance: the poverty of "genetic exceptionalism". BioSocieties 2, 349–373.
Mittra, James, Tait, Joyce, 2012. Analysing stratified medicine business models and value systems: innovation-regulation interactions. New Biotechnology 29, 709–719.
Morrison, Patrick, 2005. Insurance, unfair discrimination, and genetic testing. Lancet 36, 877–880.

National Research Council, 2004. Safety of Genetically Modified Foods: Approaches to Assessing Unintended Health Effects. National Academies Press, Washington.
Niewöhner, Jörg, 2011. Epigenetics: embedded bodies and the molecularisation of biography and milieu. BioSocieties 6, 279–298.
Ossorio, Pilar, 2007. The human genome as common heritage: common sense or legal nonsense? The Journal of Law, Medicine and Ethics 35, 425–439.
Perera, Frederica, 1996. Uncovering new clues to cancer risk. Scientific American 274, 54–55, and 58–62.
Plomer, Aurora, 2012. After Brüstle: EU accession to the ECHR and the future of European patent law. Queen Mary Journal of Intellectual Property 2, 110–135.
Powell, Alexander, Dupré, John, 2009. From molecules to systems: the importance of looking both ways. Studies in History and Philosophy of Biological and Biomedical Sciences 40, 54–64.
Rothstein, Mark (Ed.), 1997. Genetic Secrets: Protecting Privacy and Confidentiality in the Genetic Era. Yale University Press, Connecticut.
Rothstein, Mark, Anderlik, Mary, 2001. What is genetic discrimination, and when and how can it be prevented? Genetics in Medicine 3, 354–358.
Taylor, Sandra, Treloar, Susan, Barlow-Stewart, Kristine, et al., 2008. Investigating genetic discrimination in Australia: a large-scale survey of clinical genetics clients. Clinical Genetics 74, 30–40.
Torrence, Andrew, 2010. Gene concepts, gene talk and gene patents. Minnesota Journal of Law, Science and Technology 11, 177–178.
Van de Vujver, G., 2009. No genetics without epigenetics? No biology without systems biology? Annals NY Academy of Sciences 1178, 305–317.
Waddington, Conrad, 1942. The epigenotype. Endeavour 1, 18–20.
Wainwright, Steven, Williams, Clare, Michael, Mike, et al., 2006. Ethical boundary-work in the embryonic stem cell laboratory. Sociology of Health and Illness 28, 732–748.
Wynn, Brian, 2005. Reflexing complexity: post-genomic knowledge and reductionist returns in public science. Theory, Culture and Society 22, 67–94.

Cases

American Lung Association v EPA, 1998. 134 F.3d 388 (DC Cir).
Amgen Inc. v Chagai, 1991. 927 F.2d 1200 (Fed Cir).
Association for Molecular Pathology et al. v. United States Patent and Trademark Office, et al.
Biogen v Madeva, 1997. RPC 1 (HL).
Brüstle v Greenpeace e.V., Case C-34/10, 18 October 2011. Grand Chamber ECJ.
Chlorine Chemistry Council v EPA, 2000. 206 F.3d 1286 (DC Cir).
Commission of the European Communities v Italy, Case C-456/03, 2005. ECR I-0533.
DPP v Bayer, 2003. EWHC 257 (Admin.).
Diamond v Chakrabarty, 1980. 447 SC 303.
Doodeward v Spence, 1908. 6 C.L.R. 406 (Aust HC).
EDINBURGH/Animal Transgenic Stem Cells, Patent Application No. 94913174.2, 21 July 2002 (Opp Div).
Eli Lilly v Human Genome Sciences, 2011. UKSC 51.
Fiers v Revel, 1993. 884 F.2d 1164 (Fed Cir).
Genentech Inc.'s Patent, 1989. RPC 147 (CA).
Harvard College v Canada (Commissioner of Patents), 2002. SCC 76.
Haynes Case, 1614. 77 ER 1389.
In re Deuel, 1995. 51 F.3d 1552 (Fed Cir).
Kirin-Amgen Inc. v Hoechst Marion Roussel Ltd. 2004. UKHL 46.
Mayo Collaborative Services v Prometheus Laboratories, 2012. West Law 912952 (USSC).
Monsanto Canada Inc. v Schmeiser, 2004. 1 SCR 902.
Moore v Regents of the University of California, 1990. 793 P.2d 479 (Cal SC).
Netherlands v European Parliament & Council, Case C-377/98, 9 October 2001. ECR I-07079.
R v Bentham, 2005. 1 WLR 1057 (HL).
R v Kelly, 1998. 3 All ER 741 (CA).
R v Lynn, 1788. 2 T R 394.
Save Our Summers v Washington Department of Ecology, 1999. 132 F. Supp. 2d 896 (E.D. Wash.).
WARF/Primate Embryonic Stem Cells, Patent Application No. 96903521.1, 25 November 2008 (Enlarged Bd App).
Williams v Williams [1881–85] All ER 840.
Yearworth and Others v North Bristol NHS Trust, 2009. 2 All ER 986 (CA).

Relevant Websites

http://www.conventions.coe.int/ – Council of Europe Conventions.
http://eur-lex.europa.eu/en/legis/latest/chap16.htm – European Union Legislation on Science, Information, Education, and Culture.
http://www.ukbiobank.ac.uk/ – UK Biobank.
http://www.legislation.gov.uk/ – UK Legislation.
http://portal.unesco.org/en/ev.php-URL_ID=13652&URL_DO=DO_TOPIC&URL_SECTION=-471.html – UNESCO Legal Instruments on Social and Human Science.

Genetics of Personality

Sharon Niv and Laura A Baker, University of Southern California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by N.L. Pedersen, volume 9, pp. 6116–6118, © 2001, Elsevier Ltd.

Abstract

In this article, the findings from behavioral genetic studies of personality are summarized, including twin and adoption studies estimating heritability as well as molecular genetic studies of candidate genes and genome wide association studies. Studies of biological markers and other personality correlates are also reviewed, with suggestions for future directions in genetic studies of personality. An overview of personality theories and assessment instruments of personality is also provided.

With roots reaching back as far as ancient Greece and the four Hippocratic temperamental humors, the study of personality and its hereditary basis has a long history. 'Personality' refers to a set of tendencies, attitudes, emotions, and behavioral patterns that characterize an individual. Conceptualized as a series of traits rather than any single defining characteristic, personality encompasses one's temperament character, attitudes, and worldviews. Personality research suggests that certain traits and tendencies are enduring patterns, and not merely state dependent or fleeting. Consistent with this finding, we intuitively recognize that people's behaviors are largely stable and not arbitrary. Most widely used personality instruments in research adopt this 'trait approach,' although others using a typological approach are also used in clinical and other applied settings.

One consistent finding in personality research is that almost all traits show moderate genetic influence, or *heritability*. Genetic analysis has been applied in abundance to dimensional traits. The findings from behavioral genetic studies of personality are summarized in this article, starting with twin and adoption studies estimating heritability as well as molecular genetic studies of candidate genes and genome wide association studies (GWAS). Studies of biological markers and other personality correlates are also reviewed, followed by future directions in genetic studies of personality. An overview of personality theories and assessment instruments of personality is first provided.

What Is Personality?

Before considering its genetic and environmental basis, it is worth reviewing the major theories and definitions of personality. Although 'personality' is a common lay term, there has historically been a lack of consensus within the scientific community as to its definition and constitution. Modern personality theories consider such different areas as (1) understanding individual traits and the relationships amongst them; (2) investigating personality dimensions related to psychopathology; and (3) the use of typological approaches to understanding personality.

Personality Theories: Traits vs Typologies

No single theory of personality is accepted by all psychologists. Several theories posit different traits or dimensions of personality, although several major traits are often found across these models. The most commonly studied of these is the Five Factor Model (FFM), commonly referred to as the 'Big Five' (McCrae and Costa, 1987), which is comprised of extraversion (engagement with the external world), neuroticism (proneness to negative feelings and experiences), openness to experience (novelty and curiosity), conscientiousness (organization and efficiency), and agreeableness (friendliness and compassion). The FFM has also been found to reflect two higher order factors (Digman, 1997), the first of which (α) is defined by agreeableness, conscientiousness and neuroticism, which is thought to be similar to Adler's 'social interest.' The second factor (β) is comprised of extraversion and openness to experience and may reflect such concepts as Maslow's 'self-actualization.' The FFM has much in common with an earlier theory of personality that specified three traits, including two found in the Big Five (extraversion, and neuroticism), and a third (psychoticism – reflecting aggression and proneness to unusual thinking) (Eysenck and Eysenck, 1975). Another major personality theory is a seven-factor model proposed by Cloninger et al. (1994), which consists of four temperaments (harm avoidance, novelty seeking, reward dependence, and persistence), and three character traits (self-directedness, self-transcendence, and cooperativeness). The first two temperaments are similar to neuroticism and extraversion, respectively. Of these theories, the FFM has by far received the most attention from behavioral genetic research, in part due to its reliability and reproducibility cross-culturally. Among these instruments, it is also the most brief (as few as 40 items) and easy to administer (self-report questionnaire).

In addition to dimensional personality models (i.e., trait approaches that yield quantitative variation), there exist several notable models that use a qualitative or typological approach. The most well recognized of these is the Myers-Briggs Type Indicator (Myers, 1962), which is based on Jungian theory and focuses on four dichotomies (extraversion–introversion; intuitive–sensing; thinking–feeling; judging–perceiving) and yields 16 possible personality types reflecting all possible combinations of the four dichotomies. Another typological approach with considerable attention is the Enneagram, which examines nine interconnected personality types based on core motivations and fears. One challenge that emerges for these typological personality instruments is that traditional quantitative genetic methods (e.g., correlations among relatives) cannot be easily applied to their resulting categorical variables.

Personality across Development

In children, researchers typically refer to temperament characteristics rather than personality. One theory suggested three temperamental *types*: easy, difficult, and slow to warm up (Chess and Thomas, 1987, 1990), as well as nine temperament *traits*: activity level, distractibility, intensity, regularity, sensory threshold, approach/withdrawal, adaptability, persistence, and mood (Thomas and Chess, 1977). These have been found to predict personality traits at the age of 18, as well as interpersonal relationships, social support, employment status, and psychiatric disorders at the age of 21 (Caspi, 2000).

Although personality is often considered a stable set of traits, research has found that changes in personality occur throughout development, even in adulthood and into old age, with the most changes during adulthood occurring between the ages of 20 and 40 (Roberts and Mroczek, 2008). Personality tends to change across age toward increase in self-confidence, warmth, self-control, social dominance, and emotional stability. Consistent with the idea that personality can change across development are findings from a recent meta-analysis of 92 studies across the age span of 10–101 years, showing mean level changes over the course of life, with significant changes in 75% of traits in middle age and beyond (ages 40–60 and 60+; Roberts and Mroczek, 2008).

Personality Assessment

Personality is typically measured through self-report instruments. Individual differences in response patterns to questions about personal behavior and attitudes remain stable across even decades – that is, the rank ordering of individuals on personality traits remains similar across time, in spite of the fluctuations as described above (Costa and Mccrae, 1992). The instruments most often studied in genetic research of personality are those that assess the FFM, such as the 240-item NEO-PI Revised (Costa and Mccrae, 1992). The 100-item Eysenck Personality Questionnaire (EPQ; Eysenck and Eysenck, 1975) measures the traits of psychoticism, extraversion and neuroticism. The Multidimensional Personality Questionnaire (MPQ), consisting of 276 questions, measures 11 primary traits (well-being, social potency, achievement, social closeness, stress reaction, alienation, aggression, control, harm avoidance, traditionalism, and absorption) into three higher order factors that are potentially comparable to openness, neuroticism, and conscientiousness, namely positive emotionality, negative emotionality, and constraint, respectively (Tellegen, 1982). The 240-item Temparament and Character Inventory (TCI) measures personality through Cloninger's seven-factor model (Cloninger, 1987). Brief versions of these instruments are also available and widely used, for example, a 40-item FFM (Thompson, 2008) and a 48-item EPQ (Eysenck et al., 1985).

In addition to those inventories that examine facets of personality, there exist several prominent measures often used in clinical psychology research and that are concerned with pathological aspects of personality. The most common of these is the Minnesota multiphasic personality inventory (MMPI-2; Butcher, 1989), a 567-item self-report measure aimed at discerning tendencies toward deviance and psychopathology in addition to some normative personality features. The MMPI-2 has several validity scales built into it, which are designed to identify whether individuals are attempting to portray themselves more negatively (e.g., faking mental illness) or more positively (e.g., feigning social desirability). Similar mental health and personality assessments exist, such as the personality assessment inventory (Morey and Lowmaster, 1999) or the California personality inventory (CPI; Gough, 1957).

Genetics of Personality: Twin and Adoption Studies

Behavioral genetic studies often utilize twin or adoption designs and are aimed at answering questions of 'nature and nurture' by comparing covariation of variables between individuals who are known to share specific proportions of genes. For example, comparing correlations for a given trait between 'identical' monozygotic (MZ) twins (who share 100% of their genes) and 'fraternal' or non-identical dizygotic twins (who share 50% of their genes on average) can help discern how much genetic influence exists for that trait. These studies often constitute the first step in establishing heritability (h^2) of an observable trait (i.e., the proportion of observable variance in a trait that can be explained by genetic differences among people), and may serve as rationale for further molecular genetic investigation in order to identify specific genes involved in personality. The earliest and most consistent finding from behavior genetics studies is that all personality traits have some heritable component (Loehlin and Nichols, 1976). Another notable finding from most studies discussed below is that environmental influence on personality stems from nonshared environment, meaning environment that is unique to each individual and not shared by twins. That is, individual experiences appear to be most important to shaping personality, rather than the shared environment among twins and other family members.

A large portion of behavior genetics research conducted on personality has utilized the FFM. Typically, all five factors emerge as primarily influenced by genetics and nonshared environment, with little if any effect of shared environment. Heritability estimates are 40–50% for facets of the FFM (Loehlin, 1992), with remaining variance being explained by nonshared environment. This is consistent with broad estimates of heritability between 30 and 50% on all personality facets, including those from other instruments such as the EPQ, CPI, and Junior Temperament and Character Inventory (JTCI) (Plomin et al., 2008; Jang et al., 1998). Some studies have examined the cross-cultural structure of influences on the FFM, and have found that the structure, both genetic and environmental influences and correlations amongst them, remains stable across cultures from North America, Asia, and Europe (Yamagata et al., 2006). Personality stability appears to be influenced largely by stable genetic influences, whereas change over time (across the age span of adolescence to late middle age) is due largely to changes in environment (Kandler et al., 2010). At least one twin study analyzed the α (neuroticism, conscientiousness, agreeableness) and β (openness, extraversion) traits and found that genetic influences were best represented as exerting their effects through latent common variables of α and β (Jang et al., 2006).

Individual differences in other personality traits can be explained by similar proportions of genetic and nonshared

environment influences. Sensation-seeking, harm avoidance, reward dependence, and persistence are each influenced by genetics (h^2 = 35–60%) (Heiman et al., 2012; Zuckerman, 1994), as are Cloninger's seven personality factors (h^2 = 30–50%) (Gillespie et al., 2003). When Cloninger's and Eysenck's models of personality were jointly examined, six separate genetic and environmental sources of influences were found, suggesting the two sets of traits may not be entirely redundant with one another, but rather that each represents an incomplete description of personality with different underlying genetic etiologies (Heath et al., 1994). Akin to the g-factor of general cognitive ability or intelligence, a general factor of personality (GFP) has been proposed from factor analyses of the FFM that provide an alternative to α and β findings (Musek, 2007; Rushton et al., 2008). This GFP is comprised of desirable personality traits such as low neuroticism, extraversion, and conscientiousness, and has been found to be influenced by genetics (h^2 = 46–53%) (Veselka et al., 2009).

For childhood temperament specifically, heritable influences in five dimensions that somewhat overlap but are not entirely redundant with adult FFM traits have been examined. These temperament traits were activity level, anger–frustration, behavioral inhibition–fear, effortful control, and positive effect, each of which showed similar proportions of genetic influence to the adult FFM traits, with heritability estimates ranging between 38 and 58% (Gagne et al., 2009).

In spite of their popular interest, there has been a notable lack of behavioral genetic research on categorical measures of personality (i.e., typological approaches). There exist no reliable estimates of the heritability of typological personality assessments such as the Myers–Briggs or the Enneagram, regardless of their growing popularity and use outside of research settings. However, when the Myers–Briggs categories were examined as continuous variables, thinking–feeling and introversion–extraversion were found to be about 60% heritable and sensing–intuition and judgment–perception were found to be about 40% heritable (Bouchard and Hur, 1998).

Genetic Influences on Correlates of Personality

In addition to studying genetic influences on personality directly, it is also possible to investigate heritable influences on *behavioral correlates of personality*, including traits involved in both mate selection and the success (or failure) of romantic relationships. While no genetic influence has emerged for individual differences in mate selection itself (Lykken and Tellegen, 1993) or romantic style in love (Waller and Shaver, 1994), propensity to divorce was found to be 52% heritable in both men and women (McGue and Lykken, 1992). Moreover, 30–42% of the genetic influence on risk of divorce overlapped with genetic influences on the personality factors of positive emotionality, negative emotionality, and constraint of the MPQ (Jockin et al., 1996).

What about the abnormal or pathological aspects of personality? Some psychologists conceptualize personality disorders (PDs) to be pathological manifestations of normative traits. The 10 personality disorders outlined in the *Diagnostic and Statistical Manual* (DSM-IV) are theorized to cluster into three categories: A (odd/eccentric), B (dramatic), and C (anxious). Multivariate genetic analysis of traits corresponding to these 10 PDs found influence of three genetic and three environmental factors. The first genetic factor loaded onto all 10 PDs, while the second loaded only onto borderline and antisocial traits (both within the 'dramatic' B cluster), and the third loaded onto schizoid and avoidant traits (both within the 'odd/eccentric' A cluster). The environmental factors loaded onto the three defined DSM-IV clusters almost perfectly with the exception of obsessive-compulsive personality disorder, which was influenced only by unique environmental influences. Several PDs are influenced by unique genetic or environmental influences not shared with other PDs. In all, heritability estimates ranged from 20 (schizotypal) to 40% (antisocial), falling in line with personality heritability estimates for normative traits (Kendler and Aggen, 2008). Further evidence for similar ranges of heritability in pathological personality dimensions come from genetic studies of the MMPI or MMPI-2, which find h^2 = 25–57% in the clinical scales (Viken and Rose, 2007; Gizer et al., 2010). Moderate genetic overlap emerged amongst the scales, with positive genetic correlations among them, except for mania with depression and social introversion, and hypochondriasis with social introversion, which showed negative genetic correlations (Viken and Rose, 2007).

Another question is whether major psychological disorders might be extreme manifestations of normal varying personality traits. Genetically informative studies have attempted to address this by investigating the genetic overlap between major personality traits of extraversion and neuroticism with categories of externalizing and internalizing disorders, respectively. Indeed, internalizing disorders including anxiety, depression, and phobias appear to have some genetic overlap with the personality trait neuroticism (Hettema et al., 2008). Extraversion, on the other hand, does not appear to overlap genetically with externalizing disorders such as antisocial personality disorder and substance abuse (Khan et al., 2005). Instead, the externalizing disorders appear to be more related to disinhibition and (lack of) constraint (Krueger et al., 1996).

Biological Substrates of Personality

Certain endocrine and neurobiological measures have been associated with personality features. The study of the heritability of these measures has provided additional clues about the underlying physiological mechanisms of personality. For example, asymmetry of activation in the prefrontal cortex has been associated with positive emotionality and an approach tendency, which are related to low neuroticism and openness (Davidson and Fox, 1982; Coan and Allen, 2004). The heritability of frontal asymmetry is somewhat low in children (h^2 = 0–21%; Gao et al., 2009), and moderate in adults (h^2 = 27–37%; Anokhin et al., 2006; Smit et al., 2007), although there is little or no research on the genetic overlap of this asymmetry with personality.

Another biological substrate of personality is cortisol, a stress hormone related to anxiety, chronic stress, and irritability. Naturally occurring cortisol levels in blood, saliva, and urine were found to be heritable (h^2 = 60%; Bartels et al., 2003). In addition, cortisol reactivity to stress is also a relevant measure of one's natural response patterns, and this appears to be 56–97% heritable (Federenko et al., 2004).

Molecular Genetics of Personality

Although there have been many examinations of specific genes and their involvement in personality outcomes, the results have been marked by inconsistency. One common form of molecular genetic research is association studies, which involve candidate genes, or specific alleles with different variants that are differently related to measurable traits. Several candidate genes are frequently studied in personality and psychopathology research, typically using genes that are involved in neurotransmitter systems. Cloninger's theory of personality posits that dopamine is the dominant neurotransmitter system involved in novelty seeking, serotonin in harm avoidance, and norepinephrine in reward dependence (Cloninger, 1993). These neurotransmitters are related to the candidate genes most often explored.

The first of these neurotransmitter systems is dopamine, involved in mood, reward response, learning, attention, arousal, and other important functions. Dopamine receptor D4 (DRD4) was first found to have a long variant associated novelty seeking (Cloninger, 1993), but several subsequent studies failed to replicate this finding or found the contribution was negligible (Ebstein et al., 1997; Kluger et al., 2002). The dopamine receptor D2 (DRD2) gene was also found to be associated with novelty seeking (Berman et al., 2002; Light et al., 2007).

Another frequently examined candidate gene is the serotonin transporter gene (5HTTLPR), the short form of which has been associated with neuroticism (Lesch et al., 1996). However, this too has failed to replicate (Flory et al., 1998) A recent meta-analysis found the following associations significant: (1) approach motivation and the DRD4 c > t polymorphism, and DRD3 A1/A2 polymorphism; (2) avoidance traits and the 5HTTLPR polymorphism, DRD4 c > t polymorphism, DRD2 A1/A2 polymorphism and DRD3 A1/A2 polymorphism; (3) aggression traits and 5HTTLPR polymorphism, and DRD3 A1/A2 polymorphism (Munafò et al., 2003). However, due to some failures to replicate even within the studies in this meta-analysis, there remains skepticism in the field about any one of these genes' direct contributions to personality.

Some genes found to associate with personality have involvement with several neurotransmitter systems. For example, transcription factor AP-2β has binding sites that are involved with regulating several genes related to dopamine (DβH, DAT1, DRD1a), as well as serotonin (5HTTLPR, 5HT2a) and acetyl choline (ChAT), and is active largely during development. The short allele variant of this gene has been associated with higher anxiety and muscular tension in both men and women (Damberg et al., 2000). Another transcription factor, brain-derived neurotropic factor has a single nucleotide polymorphism (SNP), producing either a valine or a methionine amino acid in the outcome protein. The methionine variant of this SNP has been associated with lower trait neuroticism (Sen et al., 2003).

Overall, although candidate gene studies represent an optimistic avenue for genetic research into personality, study results so far have been contradictory and effect sizes small. One proposed problem with these studies is that they tend to be underpowered because the involvement of any particular single-locus genotype on a complex trait is very minimal (Martin et al., 1997). Several newer genetic research methods have similarly produced little insight into genetic origins of personality. For example, GWAS use large sample sizes (as high as $N > 10\ 000$) and approximately 500 000 to 2 million loci to search for SNP association to trait variation or disease vulnerability. These studies have not found significant associations of any one gene with any personality trait (van den Oord et al., 2008; Terracciano et al., 2010). Notably, these GWAS studies did not find significant associations with some of the candidate genes most often examined, such as DRD4 or 5HTTLPR.

The majority of linkage studies – which use within-family designs to test for chromosomal regions contributing to variance in personality traits – have concentrated on the trait neuroticism, and one consistent replication suggests involvement of chromosome 12 (Fullerton et al., 2003; Kuo et al., 2007; Gillespie et al., 2008). Additionally, individual differences in emotional liability (which is akin to neuroticism) have been associated with a broad region on chromosome 8p (Cloninger et al., 1998; Fullerton et al., 2003; Dina et al., 2005). Overall, the inconsistency of results across GWAS, linkage and association studies suggests that personality is immensely complex and likely shaped by intricate interactions among a large number of genes, as well as by environmental influences on those genes.

Conclusions and Future Directions

Personality traits have consistently emerged as moderately heritable, with genetic factors accounting for about 30–50% of the observed variance in each trait, regardless of the measuring instrument or theoretical approach. The environmental influences on personality have consistently emerged as being of the nonshared variety, such that the unique experiences of each individual rather than shared experiences among family members appears to shape the development of behavioral and emotional response patterns.

Among the many molecular genetic studies of personality, the vast majority has focused on a small number of well-studied candidate genes, which account for only small percentages of personality variance at best. There are likely thousands of genes that comprise the complex organization structure of personality and that have yet to be researched. As new sequencing, synthetic biology, and bioinformatics methods emerge for identifying genetic regions involved in behavior, our ability to extract more reliable information should increase considerably. In addition, it is important to consider the potential involvement of gene by environment interactions (GxE) in the development of personality. GxE refers to the interplay between genes and the environments that may affect them, and some evidence already exists to suggest that specific environment does activate genetic disposition. For example, harsh home environments have been found to interact with certain alleles of DRD2, 5HTTLPR, or the monoamine oxidase A gene in producing more antisocial behavior (Caspi et al., 2002; Retz et al., 2008). A GxE association between genetics and hostile environment has also been found in novelty seeking for DRD4 (Keltikangas-Järvinen et al., 2004) and DRD2

(Keltikangas-Järvinen et al., 2009). Arguments have been made to suggest that effects of any one genetic variant may be too minimal to be detected without the consideration of GxE (Bronfenbrenner and Ceci, 1994; Munafò and Flint, 2011).

Another field of genetic research that is gaining increasing attention is *epigenetics*, which studies molecular changes to DNA and gene expression due to environmental conditions. This represents direct research into GxE mechanisms, and could potentially explain how behavioral differences in MZ twins come about. Epigenetic changes to the DNA do not affect sequence but affect the shape of the molecule and its folding patterns. One commonly studied epigenetic mechanism is methylation – a process by which methyl groups become attached to DNA nucleotides. A strand's 'methylome' can affect not only the way in which it folds, but also its attractiveness for transcription, both of which can have profound effects on genetic expression. Evidence for epigenetic effects in personality have already emerged in MZ twins with different levels of risk-taking behavior (Kaminsky et al., 2009). As bioinformatics methods advance and become capable of more efficient calculation and simulation, more information regarding molecular genetics will become available.

Lastly, but perhaps most importantly, the challenge that faces geneticists seeking answers about personality is a lack of consensus about what personality actually is. Different theories of personality may overlap, but are not completely redundant. There are questions about whether personality should be measured within a person's own trait organization or in comparison to others (Cervone, 2005). It has also been suggested that in addition to self-report or behavior, endophenotypes (e.g., biological markers) and behavioral assessments need to be included in definitions of personality (Jang and Yamagata, 2009). Others have suggested altogether that personality is not nearly as stable as is assumed, but is rather situational, with patterns to the situational variance (Mischel and Shoda, 2008). Until personality concepts and their assessment are refined to a point of agreement among personality theorists, questions will remain about whether genetic studies are accessing the correct constructs.

See also: Behavioral Genetics: Psychological Perspectives; Familial Studies: Genetic Inferences; Genetic Studies of Behavior: Methodology; Genetics and Society; Genetics: The New Genetics; Genotype and Phenotype; Twin Studies and Epigenetics.

Bibliography

Anokhin, A.P., Heath, A.C., et al., 2006. Genetic and environmental influences on frontal EEG asymmetry: a twin study. Biological Psychology 71, 289–295.

Bartels, M., de Geus, E.J.C., et al., 2003. Heritability of daytime cortisol levels in children. Behavior Genetics 33, 421–433.

Berman, S., Ozkaragoz, T., et al., 2002. D2 dopamine receptor gene polymorphism discriminates two kinds of novelty seeking. Personality and Individual Differences 33, 867–882.

Bouchard, T.J., Hur, Y.M., 1998. Genetic and environmental influences on the continuous scales of the Myers-Briggs Type Indicator: an analysis based on twins reared apart. Journal of Personality 66, 135–149.

Bronfenbrenner, U., Ceci, S.J., 1994. Nature-nurture reconceptualized in developmental perspective: a bioecological model. Psychological Review 101, 568–586.

Butcher, J.N., 1989. Minnesota multiphasic personality inventory. Corsini Encyclopedia of Psychology.

Caspi, A., 2000. The child is father of the man: personality continuities from childhood to adulthood. Journal of Personality and Social Psychology 78, 158–172.

Caspi, A., McClay, J., et al., 2002. Role of genotype in the cycle of violence. Science 297, 851–854.

Cervone, D., 2005. Personality architecture: within-person structures and processes. Annual Review of Psychology 56, 423–452.

Chess, S., Thomas, A., 1987. Know Your Child: An Authoritative Guide for Today's Parents. Basic Books, New York.

Chess, S., Thomas, A., 1990. Continuities and discontinuities in temperament. In: Robins, L.N., Rutter, M. (Eds.), Straight and Devious Pathways from Childhood to Adulthood. Cambridge University Press., New York, NY, pp. 205–220.

Cloninger, C.R., 1987. A systematic method for clinical description and classification of personality variants: a proposal. Archives of General Psychiatry 44, 573–588.

Cloninger, C.R., 1993. A psychobiological model of temperament and character. Archives of General Psychiatry 50, 975–990.

Cloninger, C.R., Przybeck, T.R., et al., 1994. The Temperament and Character Inventory (TCI): A Guide to Its Development and Use. Center for Psychobiology of Personality, Washington University, St. Louis, MO.

Cloninger, C.R., Van Eerdewegh, P., et al., 1998. Anxiety proneness linked to epistatic loci in genome scan of human personality traits. American Journal of Medical Genetics 81, 313–317.

Coan, J.A., Allen, J.J.B., 2004. Frontal EEG asymmetry as a moderator and mediator of emotion. Biological Psychology 67, 7–49.

Costa, P.T.J., Mccrae, R.R., 1992. Four ways five factors are basic. Personality and Individual Differences 13 (6), 653–665.

Damberg, M., Garpenstrand, H., et al., 2000. A polymorphic region in the human transcription factor AP-2β gene is associated with specific personality traits. Molecular Psychiatry 5, 220–224.

Davidson, R., Fox, N., 1982. Asymmetrical brain activity discriminates between positive and negative affective stimuli in human infants. Science 218, 1235–1237.

Digman, J.M., 1997. Higher-order factors of the Big Five. Journal of Personality and Social Psychology 73, 1246–1256.

Dina, C., Nemanov, L., et al., 2005. Fine mapping of a region on chromosome 8p gives evidence for a QTL contributing to individual differences in an anxiety-related personality trait: TPQ harm avoidance. American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics 132B, 104–108.

Ebstein, R.P., Gritsenko, I., et al., 1997. No association between the serotonin transporter gene regulatory region polymorphism and the Tridimensional Personality Questionnaire (TPQ) temperament of harm avoidance. Molecular Psychiatry 2, 224–226.

Eysenck, H.J., Eysenck, S.B.G., 1975. EPQ (Eysenck Personality Questionnaire). Educational and Industrial Testing Service, San Diego, CA.

Eysenck, S.B.G., Eysenck, H.J., et al., 1985. A revised version of the Psychoticism scale. Personality and Individual Differences 6, 21–29.

Federenko, I.S., Nagamine, M., et al., 2004. The heritability of hypothalamus pituitary adrenal axis responses to psychosocial stress is context dependent. The Journal of Clinical Endocrinology and Metabolism 89, 6244–6250.

Flory, J.D., Mann, J.J., et al., 1998. Recovery from major depression is not associated with normalization of serotonergic function. Biological Psychiatry 43, 320–326.

Fullerton, J., Cubin, M., et al., 2003. Linkage analysis of extremely discordant and concordant sibling pairs identifies quantitative-trait loci that influence variation in the human personality trait neuroticism. American Journal of Human Genetics 72, 879–890.

Gagne, J., Vendlinski, M., et al., 2009. The genetics of childhood temperament. In: Kim, Y.-K. (Ed.), Handbook of Behavior Genetics. Springer, New York, pp. 251–267.

Gao, Y., Tuvblad, C., et al., 2009. Genetic and environmental influences on frontal EEG asymmetry and alpha power in 9–10-year-old twins. Psychophysiology 46, 787–796.

Gillespie, N.A., Cloninger, C.R., et al., 2003. The genetic and environmental relationship between Cloninger's dimensions of temperament and character. Personality and Individual Differences 35, 1931–1946.

Gillespie, N.A., Zhu, G., et al., 2008. A genome-wide scan for Eysenckian personality dimensions in adolescent twin sibships: psychoticism, extraversion, neuroticism, and lie. Journal of Personality 76, 1415–1446.

Gizer, I.R., Do, K.L. S.-s., et al., 2010. Heritability of MMPI-2 scales in the UCSF family alcoholism study. Journal of Addictive Diseases, 37–41.

Gough, H.G., 1957. The California psychological inventory. Major Psychological Assessment Instruments 2, 67–98.

Heath, A.C., Cloninger, C.R., et al., 1994. Testing a model for the genetic structure of personality: a comparison of the personality systems of Cloninger and Eysenck. Journal of Personality and Social Psychology 66, 762–775.

Heiman, N., Stallings, M.C., et al., 2012. Investigating the genetic and environmental structure of Cloninger's personality dimensions in adolescence. Twin Research 7, 462–470.

Hettema, J.M., An, S.S., et al., 2008. Association study between the serotonin 1A receptor (HTR1A) gene and neuroticism, major depression, and anxiety disorders. American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics 147B (5), 661–666.

Jang, K.L., Livesley, W.J., et al., 2006. Behavioral genetics of the higher-order factors of the big five. Personality and Individual Differences 41, 261–272.

Jang, K.L., McCrae, R.R., et al., 1998. Heritability of facet-level traits in a cross-cultural twin sample: support for a hierarchical model of personality. Journal of Personality and Social Psychology 74, 1556–1565.

Jang, K.L., Yamagata, S., 2009. Personality. In: Kim, Y.-K. (Ed.), Handbook of Behavior Genetics, Springer Science+Business Media, LLC, New York, pp. 223–237.

Jockin, V., McGue, M., et al., 1996. Personality and divorce: a genetic analysis. Journal of Personality and Social Psychology 71, 288–299.

Kaminsky, Z.A., Tang, T., et al., 2009. DNA methylation profiles in monozygotic and dizygotic twins. Nature Genetics 41, 240–245.

Kandler, C., Bleidorn, W., et al., 2010. Sources of cumulative continuity in personality: a longitudinal multiple-rater twin study. Journal of Personality and Social Psychology 98, 995–1008.

Keltikangas-Järvinen, L., Pulkki-Råback, L., et al., 2009. DRD2 C32806T modifies the effect of child-rearing environment on adulthood novelty seeking. American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics 150B, 389–394.

Keltikangas-Järvinen, L., Räikkönen, K., et al., 2004. Nature and nurture in novelty seeking. Molecular Psychiatry 9, 308–311.

Kendler, K., Aggen, S., 2008. The structure of genetic and environmental risk factors for DSM-IV personality disorders: a multivariate twin study. Archives of General Psychiatry 65, 1438–1446.

Khan, A.A., Jacobson, K.C., et al., 2005. Personality and comorbidity of common psychiatric disorders. British Journal of Psychiatry 186, 190–196.

Kluger, A.N., Siegfried, Z., et al., 2002. A meta-analysis of the association between DRD4 polymorphism and novelty seeking. Molecular Psychiatry 7, 712–717.

Krueger, R.F., Caspi, A., et al., 1996. Personality traits are differentially linked to mental disorders: a multitrait-multidiagnosis study of an adolescent birth cohort. Journal of Abnormal Psychology 105, 299–312.

Kuo, P.-H., Neale, M.C., et al., 2007. A genome-wide linkage analysis for the personality trait neuroticism in the Irish affected sib-pair study of alcohol dependence. American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics 144B, 463–468.

Lesch, K.-P., Bengel, D., et al., 1996. Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. Science 274, 1527–1531.

Light, K.J., Joyce, P.R., et al., 2007. An association study of DRD2 and COMT polymorphisms with novelty seeking and harm avoidance scores, in two independent samples of depressed patients. Behavioral and Brain Functions: BBF 3, 3.

Loehlin, J.C., 1992. Genes and Environment in Personality Development. Sage Publications, Inc, Newbury Park.

Loehlin, J.C., Nichols, R.C., 1976. Heredity, Environment, & Personality: A Study of 850 Sets of Twins. University of Texas Press, Austin, TX.

Lykken, D.T., Tellegen, A., 1993. Is human mating adventitious or the result of lawful choice? A twin study of mate selection. Journal of Personality and Social Psychology 65, 56–68.

Martin, E.R., Kaplan, N.L., et al., 1997. Tests for linkage and association in nuclear families. American Journal of Human Genetics 61, 439–448.

McCrae, R.R., Costa, P.T., 1987. Validation of the five-factor model of personality across instruments and observers. Journal of Personality and Social Psychology 52, 81–90.

McGue, M., Lykken, D., 1992. Genetic influence on risk of divorce. Psychological Science 3, 368–373.

Mischel, W., Shoda, Y., 2008. Toward a unified theory of personality. In: John, O.P., Robins, R.W., Pervin, L. (Eds.), Handbook of Personality: Theory and Research. Guilford Press, New York, pp. 208–241.

Morey, L., Lowmaster, S., 1999. Personality assessment inventory. In: Weiner, I.B., Craighead, W.E. (Eds.), Corsini Encyclopedia of Psychology. John Wiley and Sons.

Munafò, M.R., Clark, T.G., et al., 2003. Genetic polymorphisms and personality in healthy adults: a systematic review and meta-analysis. Molecular Psychiatry 8, 471–484.

Munafò, M.R., Flint, J., 2011. Dissecting the genetic architecture of human personality. Trends in Cognitive Sciences 15, 395–400.

Musek, J., 2007. A general factor of personality: evidence for the Big One in the five-factor model. Journal of Research in Personality 41, 1213–1233.

Myers, I.B., 1962. The Myers-briggs Type Indicator. Consulting Psychologists Press, Palo Alto, CA.

Plomin, R., DeFries, J.C., et al., 2008. Behavioral Genetics. Worth Publishers, New York.

Retz, W., Freitag, C.M., et al., 2008. A functional serotonin transporter promoter gene polymorphism increases ADHD symptoms in delinquents: interaction with adverse childhood environment. Psychiatry Research 158, 123–131.

Roberts, B.W., Mroczek, D., 2008. Personality trait change in adulthood. Current Directions in Psychological Science 17, 31–35.

Rushton, J.P., Bons, T.A., et al., 2008. The genetics and evolution of a general factor of personality. Journal of Research in Personality 42, 1173–1185.

Sen, S., Nesse, R.M., et al., 2003. A BDNF coding variant is associated with the NEO personality inventory domain neuroticism, a risk factor for depression. Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology 28, 397–401.

Smit, D.J.A., Posthuma, D., et al., 2007. The relation between frontal EEG asymmetry and the risk for anxiety and depression. Biological Psychology 74, 26–33.

Tellegen, A., 1982. Brief Manual for the Multidimensional Personality Questionnaire. University of Minnesota, Minneapolis, MN.

Terracciano, a., Sanna, S., et al., 2010. Genome-wide association scan for five major dimensions of personality. Molecular Psychiatry 15, 647–656.

Thomas, A., Chess, S., 1977. Temperament and Development. Brunner/Mazel, New York, NY.

Thompson, E.R., 2008. Development and validation of an international English Big-Five mini-markers. Personality and Individual Differences 45, 542–548.

van den Oord, E.J.C.G., Kuo, P.-H., et al., 2008. Genomewide association analysis followed by a replication study implicates a novel candidate gene for neuroticism. Archives of General Psychiatry 65, 1062–1071.

Veselka, L., Schermer, J.A., et al., 2009. Evidence for a heritable general factor of personality in two studies. Twin Research and Human Genetics: The Official Journal of the International Society for Twin Studies 12, 254–260.

Viken, R.J., Rose, R.J., 2007. Genetic variation and covariation in the original and restructured clinical scales of the MMPI. Journal of Abnormal Psychology 116, 842–847.

Waller, N.G., Shaver, P.R., 1994. The importance of nongenetic influences on romantic love styles: a twin-family study. Psychological Science 5, 268–274.

Yamagata, S., Suzuki, A., et al., 2006. Is the genetic structure of human personality universal? A cross-cultural twin study from North America, Europe, and Asia. Journal of Personality and Social Psychology 90, 987–998.

Zuckerman, M., 1994. Behavioral Expressions and Biosocial Bases of Sensation Seeking. Cambridge University Press, Cambridge.

Genetics: The New Genetics

Karen-Sue Taussig, University of Minnesota, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores what is new about contemporary genetic research. The article examines the everyday work of genetic research and clinical practices in and beyond the laboratory; understandings of human biological variation, including race and ancestry; and the intersections of science, citizenship, and subjectivity. In examining these aspects of contemporary scientific knowledge production, the article demonstrates the deeply social nature of every aspect of contemporary genomics.

The concept of a new genetics emerged in the mid-twentieth century as a term to refer to research newly informed by James Watson and Francis Crick's (1953) discovery and description of the helical structure of the deoxyribonucleic acid (DNA) found in every living cell (Ekberg, 2007). The new genetics can perhaps best be understood not simply as a subject of biological inquiry but also as a cultural object, simultaneously producing and produced by cultural meaning and practice. Its many dimensions operate in diverse social domains and geographical locales. With its etymological roots in the idea of generation, genetics has virtually always been a future-oriented phenomenon that simultaneously invokes the past. Here we see ideas about expectations (e.g., Brown and Michael, 2003), promise (e.g., Fortun, 2008; Tutton, 2011), hope (e.g., Brown, 2003), hype (Rajan, 2007), and potential (e.g., Taussig et al., 2013) at work, as genome researchers articulate genomics as the royal road to improving the human condition. The shifts in thinking and practice associated with this new knowledge occur simultaneously with institutional and cultural changes through which the new genetics is taught, learned, and experienced.

The idea of *new* raises the question of precisely what is new and what is old in this area of contemporary research. After discussing the idea of when genetics was new, this article highlights three overlapping themes associated with the new genetics broadly conceived, exploring both the emergent and historical antecedents. The first theme involves the everyday work of scientific research and clinical practices in and beyond laboratory and medical settings. The second theme focuses on issues related to human biological variation, including those having to do with race and ancestry. The third theme deals with issues at the intersections of science, citizenship, and subjectivity. Intersecting all of these themes are issues about the body, commerce, governance, and the nature of life. Further, examination of these themes illuminates not simply the 'social impact' of new knowledge and technologies. Rather, it illustrates the deeply social nature of every aspect of the new genetics, demonstrating that the social is not after the fact of scientific or technological innovation, but that the knowledge and practices associated with the new genetics are never outside of the grip of culture.

When Genetics Was New

We might say that genetics was first new in 1865 when Gregor Mendel presented his paper, *Versuche Über Pflanzenhybriden* (*Experiments on Plant Hybridization*), elaborating the heritability of traits in peas at the meetings of the Natural History Society of Brünn in Moravia, or in 1866 when he published the paper in the Society's journal. The paper was well received but essentially ignored, its significance only apparent when it was rediscovered virtually simultaneously by Hugo de Vries (1900) and Carl Correns (1900) more than three decades later. The history of genetics from this founding through its transformation into eugenics – a term coined by Francis Galton in 1883 to refer to the idea of improving the human stock through encouraging and discouraging specific kinds of human matings – and the explosion of eugenics into a worldwide movement promoted by cutting edge scientific theories of the time in the early decades of the twentieth century, is rather well rehearsed (e.g., Kevles, 1985; Lombardo, 2008; Paul, 1995; Stepan, 1991; Stoller, 1997).

The systematic adoption of eugenic theories in Nazi racial and health policies ultimately laid bare their logical ends (Ekberg, 2007; Kevles, 1985; Paul, 1995; Proctor, 1990), thereby producing the desire, indeed the need, to create genetics anew. Recognizing the potential horrors embedded in the eugenic theories that had come to dominate their discipline, in the post-World War II era geneticists sought to distance themselves from these theories. They did so by working to distinguish the science itself from what were perceived as the social uses to which that science might be put and by separating human and nonhuman genetics (Keller in Kevles and Hood, 1992). In particular, when attending to human genetics, they focused on basic processes of physiology, acknowledging the complexities of human behavior and leaving it outside the purview of this new genetics. This genetics focused on understanding processes of heredity in order to improve diagnosis of and develop treatments for heritable diseases (Keller in Kevles and Hood, 1992; Kevles in Kevles and Hood, 1992).

Watson and Crick's (1953) publication of their discovery of the structure of DNA offered a particularly robust opportunity for shifting genetics onto new terrain that focused on understanding inheritance through genetic structure. This watershed moment, along with Joe-Hin Tjio and Albert Levan's 1956 publication convincingly articulating the number of chromosomes in a human cell into which DNA is spooled, both validated the value of focusing on this basic science and energized such efforts (Kevles in Kevles and Hood, 1992). Watson and Crick's findings supported the flourishing of this new genetics as molecular biology, including the development of

recombinant DNA (rDNA). At least by 1969, scientists began imagining this genetics as providing tools for transforming human health, articulating possibilities of animating bodies, and extending life in new ways (Sinsheimer, 1969) through intervention in life at the molecular level.

The interplay between molecular theories and practices, on the one hand, and this imaginary of healthier and longer human lives, on the other, provided the impetus and foundation for the establishment of the Human Genome Project (HGP) in the late 1980s with an initial commitment of $3 billion in funding from the U.S. government. The goal of the project was to map a complete human genome (the full complement of DNA in a single organism) that would provide a foundation for improving human health and well-being. A first draft of the human genome was announced in June 2000 and the project was declared completed in 2003.

With the establishment of the HGP, the idea of *new* became increasingly tied to a genetics associated with the genome project and the knowledge and technologies the project was producing. *The New Genetics* has since come to refer not just to the HGP, but also to all of the allied genetics (e.g., epigenetics), genomics (e.g., genomics itself, pharmacogenomics), -omics (e.g., proteomics, microbiomics, metabolomics), and novel techniques (transgenics, cloning, stem cell science) that at least in part have been built, in both the academy and the private sector, on the foundation of knowledge and techniques developed through the HGP. These endeavors have also supported the development of new medical technologies (e.g., assisted reproduction; pre- and postnatal genetic screening and testing), of new or expanding medical disciplines (e.g., genetic counseling, medical genetics, bioethics), and of direct-to-consumer commercial ventures selling genetic information (e.g., 23andMe). While framed by the idea that knowledge of the human genome could lead to dramatic medical breakthroughs, the technologies developed through the project have been adopted in other, highly diverse, domains as well – in, for example, genetic ancestry research, proving or disproving paternity or other forms of kinship, practices associated with identifying human remains after natural disasters or from mass graves, agricultural practices associated with crops and animal husbandry, and policing/criminology where DNA evidence and identification have been added to the forensic tool kit. Taken together, these efforts involve enormous financial, institutional, and cultural investments including the establishment of new academic disciplines and departments in new buildings, the development of biotechnology as a multibillion dollar sector of the economy, and the emergence of new ways of understanding humanness and human relations. As such, they are part of global commercial enterprises that have been identified under the rubric of biocapital (Helmreich, 2008). These phenomena have received significant attention from scholars interested in understanding this complex social field.

Scientific Research and Clinical Practices

In its contemporary form, we see research and clinical practices associated with the new genetics organized around an active imaginary that is simultaneously about the nature of life broadly conceived and the potential of human capacities and longevity. In contemporary scientific research and clinical practices associated with the new genetics, we see the reimagining of life, efforts to transform human capacity including the capacity for health, and emergent forms of scientific governance.

Reimagining Life

The new genetics has emerged in the context of knowledge and practices that have allowed biological materials to exist and reproduce in diverse ways outside of the bodies in which they previously resided (e.g., the reproduction of cells through tissue culture – the growth of cells derived from living tissue in an artificial medium – or the production of persons through *in vitro* fertilization). Often moving genetic material or cells across species, such practices challenge understandings of life constituted in specific (species) form and, thereby, reshape how we are able to think about life. Here I focus on three key insights into how these endeavors challenge assumptions about life. First, examining the development of tissue culture, Hannah Landecker argues that this phenomenon "changed the placticity and temporality of living things [which] thus may be harnessed to human intention" and, in so doing, "changes what it is to be biological" (Landecker, 2007: 1 and 235). The ability to grow cells into tissue and, thus, sustain life outside of the body in which those cells originated provides the technological grounds for many of the assisted reproductive technologies and related applications associated with the new genetics.

This shift in what it is to be biological was perhaps made most publicly visible in the 1997 announcement of the production of Dolly, a sheep, described as a clone, produced through the laboratory technique known as somatic cell nuclear transfer (Wilmut et al., 1997). In her analysis of Dolly, Sarah Franklin offers a second insight into life in the context of the new genetics. What was revolutionary about the cloning of Dolly, Franklin tells us, was not the cloning itself. Prior to Dolly, scientists virtually unanimously believed that the cloning of a large mammal was impossible. The cloning of Dolly did not simply prove that it was possible to clone a sheep but, rather, it did away with the notion of the biologically impossible itself (Franklin, 2001; see also Rheinberger, 1995). In other words, the advent of practices associated with the new genetics such as tissue culture and cloning have opened up new horizons for creative thinking about life in terms of what cells and the genetic materials they contain can do and can be made to do.

These horizons involve the third insight into life in the context of the new genetics – that is, the ability to redistribute life. In this potential for redistribution, as Stefan Helmreich elaborates, we find both 'life forms' and 'forms of life' in flux. For Helmreich, life forms refer to organisms, species, types of creatures, or ecological relations among creatures, while forms of life point to the "social, symbolic, and pragmatic ways of thinking and acting that organize human communities" and through which "action becomes meaningful" (Helmreich, 2009: 6). Here Helmreich is in conversation with the work of biologist Lynn Margolis and the science writer Dorian Sagan who theorize life not in terms of concrete organisms, but in terms of an open thermodynamic system "always emerging into an assortment of bodied manifestations" (Helmreich,

2009: 279). It is in relation to the dynamics of life illuminated here that we can best understand the ongoing emergence of the new genetics since the advent of the HGP.

Transforming Human Capacities

One of the most persistent characteristics of the HGP has been the promises made by those who have supported and worked on it that the new knowledge the project would develop would lead to a molecular medical tool kit that would facilitate dramatic improvements in human health. In thinking about the effects of the HGP, it is useful to distinguish between the project itself – the mapping of 'the human genome,' which can be understood as a project of structural genomics – that is an effort to understand the basic structure of a human genome – and the ongoing activities that have increasingly moved to center stage since the June 2000 announcement of a first rough draft of the genome, which can be understood as a project of functional genomics – that is, the aim of developing greater understanding of gene function through understanding the complex pathways from genetic sequence to the expression of particular traits, especially those associated with complex human capacities and diseases.

The first domain in which this new knowledge has had significant effects involves assisted reproductive technologies. Here we see efforts focused on the prevention of genetic conditions (e.g., cystic fibrosis, Tay–Sachs disease, Huntington's chorea, sickle cell disease) in the explosion of techniques through which genetic material is separated from the bodies in which it originates for amniocentesis and prenatal genetic testing; preimplantation diagnosis involving *in vitro* fertilization; and postnatal genetic testing. With the new genetics, researchers increasingly identified chromosome differences and genetic mutations associated with particular genetic anomalies. With the technical ability to isolate fetal cells from amniotic fluid, culture them, and evaluate them for such anomalies and, in the context of legal abortion, amniocentesis as a means of evaluating the genetic qualities of a fetus has become a routine aspect of reproduction in those places where such technologies are socially available and culturally meaningful (Rapp, 1999). With these practices and techniques in place, as new mutations are identified, it is a relatively simple matter to develop a prenatal diagnostic kit through which any newly recognized mutation might be identified in a particular fetus (see, for example, Taussig et al., 2003).

Although researchers certainly realized there would be a 'therapeutic lag' between the mapping of the genome and the development of useful clinical interventions, the promises made by HGP researchers were not simply about prevention, but about treatment. The goal was and remains a revolution in medicine through which a molecular medical tool kit will be mobilized to intervene in genetic conditions. Importantly, genome project proponents always emphasized that they were not just talking about rare single-gene conditions such as cystic fibrosis for which they imagined the establishment of gene therapy as a useful therapeutic intervention, but about widespread common conditions such as cancer, diabetes, heart disease because, as they argued, every embodied condition has a genetic component related to risk for and susceptibility to disease. For these more common conditions, they imagined not just gene therapy, but the tailoring of medical and pharmacological treatment to the specifics of an individual patient's genome (Kevles and Hood, 1992).

While efforts are still underway to develop gene therapy, to date it has repeatedly failed and, in fact, appears to have killed more people than it has helped. In this light, researchers have turned to other avenues, including genome-wide association studies, stem cell research, epigenetics, and synthetic biology, each heralded in turn as the big thing that would dramatically improve human health. All of this work has led to remarkable insights into life and to a handful of significant clinical applications, but none of it has yet led to the hoped for revolution in medicine. This is a space that is suffused with ideas about realizing particular human futures and efforts to shape the social relations perceived as necessary for realizing those futures.

Governance

The new genetics associated with the HGP emerged in tandem with the expression of concerns to address issues of governance in scientific research and in related clinical and forensic practices (Bunton and Peterson, 2005; Gottweis, 2002, 2005; Hindmarsh and Prainsack, 2010). A precursor to these efforts can be seen in the 1975 Asilomar Conference on Recombinant DNA. The conference brought together key researchers to discuss the risks to lab and public safety posed by new research involving rDNA technologies (which involve combining the DNA from distinct organisms) that had been made possible through techniques developed by Stanley Cohen and Herbert Boyer (Cohen et al., 1973). The conference established voluntary guidelines for using rDNA in research, including a moratorium on certain kinds of experiments, in order to ensure the safety of those working in such labs and of the broader public. Asilomar offers an example of scientific self-governance in which issues of safety in and beyond the lab were of central concern to scientists themselves (Hindmarsh and Gottweis, 2005; Weiner, 2000).

In contrast, in the face of concerns about the broad social implications of the new genetics and worry that such concerns might inhibit research, the HGP established the ethical, legal, and social implications (ELSI) of the HGP program. In so doing, the project devoted a percentage of HGP funds to the ELSI program for the study of, among other things, governance issues associated with human subject protections (e.g., privacy, informed consent, individual autonomy, the regulation of data) and intellectual property. Rather than following the voluntary self-governance model of Asilomar, these efforts were explicitly intended to identify issues of concern and shape formal government regulations associated with the research, knowledge, and clinical practices involved in the new genetics (see e.g., Meslin et al., 1997).

The establishment of ELSI led to a major infusion of money – $125 million in the first 13 years of the program – into scholarship in this domain, supporting, in particular, a flourishing of bioethics. Indeed, in the early days of the HGP, bioethicist Arthur Caplan described the project as "a full employment act for bioethicists," while others refer to the development of an 'ELSI industry' that has emerged internationally in relation to the new genetics (Hoeyer and

Tutton, 2005). From its inception, the HGP's ELSI program was enmeshed in controversy over its engagement – or lack thereof – with broader issues of power and distributive justice. The very framing of 'ethical, legal, and social,' these critiques suggest, elides the political, the economic and the moral. In exploring the political implications of contemporary governance of the new genetics, Klaus Hoeyer and Richard Tutton suggest that ethics is used here to as a "kind of assurance that potential or actual ethical problems have been identified, attended to and resolved." They argue that the result of this use of ethics is that "no matter how ambitious or controversial, biomedical research is then able to progress" (Hoeyer and Tutton, 2005: 386). These critiques suggest that tensions between concepts of regulation and enablement are inherent in the governance of the new genetics and its emergent biotechnological practices, raising questions about science subjectivity, and citizenship to which I will return after attending to issues of human biological variation also engendered by new genetic knowledge.

Race, Ancestry, and Human Biological Variation

Questions about the nature of human variation have been central to the new genetics. Here we see interest in using the tools produced by the new genetics to elucidate an evolutionary story about the history of the human family, including the relatedness of all species (e.g., Helmreich, 2009; Marks, 1995, 2002). At the same time, the idea that humans are profoundly similar at the molecular level is always in tension with the recognition that it is the differences that lead to disease and embodied affliction. The June 2000 announcement of the completion of a first draft of the human genome celebrated the unity of the species, with President Clinton stating that at the level of the genome, "the only race is the human race." Yet, those working on the project almost immediately intensified work on understanding the differences that might aid in illuminating ancestry, on the one hand, and disease processes, on the other. An early effort in this direction involved the Human Genome Diversity Project of the 1990s (Reardon, 2004 and cross reference). More recently, such efforts are visible in the 2002 establishment of the International HapMap Project that aimed to describe patterns of human variation in order to facilitate research on human disease and pharmaceutical responses. This project explicitly sought to understand human variation by gathering genetic samples initially from four populations with continental origins in Africa, Asia, and Europe. In developing the project, researchers were careful to characterize the samples they collected in locally specific terms. For example, samples from 'the Yoruba people of Ibadan, Nigeria.' Nevertheless, as the cell lines were taken up in subsequent research, they often were described simply as 'African.' Here we see how genetic material can become racialized in practices associated with the new genetics, a process Jonathan Kahn (2012: 34–36) describes as "racial recursion."

In such approaches to understanding human variation, we find ideas and assumptions related to folk theories about the relations between and within species, including those associated with race and racial difference (e.g., Abu El-Haj, 2007; Bolnick et al., 2007; Fujimura et al., 2008; Hartigan, 2008; Koenig et al., 2008; Wade et al., 2014; Wailoo et al., 2012). Here we see the tools of the new genetics utilized to engage a racialized human history and its contemporary manifestations in research and clinical practice. Given that much contemporary research associated with the new genetics focuses on the implications of human variation at the molecular level, of particular concern in this domain is the persistent ways that race gets used in medical practice as a proxy for genetic difference (Duster, 2007; Fullwiley, 2007; Kahn, 2012; Root, 2003) and that, at the same time, in forensic DNA profiling, genetic material is used to signify race and race is used to organize genetic data (Duster, 2006; Hindmarsh and Prainsack, 2010; Kahn, 2009; Krimsky and Sloan, 2011). Thus, while the completion of a first draft of the human genome provoked statements about the unity of the species, the new genetics also has been marked by a peculiar reemergence of race in genetic research and in clinical and forensic practices.

Science, Citizenship, and Subjectivity

In their essay on biological citizenship, Nikolas Rose and Carlos Novas make the point that many citizenship projects have been "framed in biological terms; in terms of race, blood lines, stock intelligence, and so forth" (2005: 440). Pointing to the significance of biology and citizenship in his book about the politics of inclusion in medical research, Steven Epstein (2007: 318 n. 13) points to a growing list of citizenship concepts related to science and understandings of biology – e.g. scientific citizens, biological citizenship (noting two distinct approaches to this concept), genetic citizenship, and therapeutic citizenship. The contemporary emergence of these concepts highlights the complex ways that new scientific knowledge implicates understandings of both citizenship and subjectivity.

The biological dimensions of citizenship associated with the new genetics can be found in at least two arenas of work in this area. The first involves the social conditions of enablement that facilitate access to the material means of contemporary knowledge production. The second can be seen in what, with Deborah Heath and Rayna Rapp, I have elsewhere described as genetic citizenship.

Social Conditions of Enablement

It has long been understood that knowledge production in the life sciences depends upon the "right tools for the job" (Clarke and Fujimura, 1992). While genetic research has now long relied on access to genetic material, what is new in contemporary research in this area is that the right tools for the production of new genetic knowledge requires access to vast quantities of data. The demand for this data is part of the move, described above, from the structural genomics of the HGP to the functional genomics of contemporary genetic knowledge production. In the structural project aimed at developing a map of a human genome, access to virtually any human genetic material would suffice. Functional genomics, which seeks to understand the complex pathways from the presence of a particular genetic sequence to the specifics of how it is

expressed in any one individual, requires access to large quantities of genetic material attached to medical records and family histories in order to help illuminate how any particular sequence is expressed in a unique individual in a particular family at a specific time. These material demands of contemporary knowledge production associated with the new genetics have engendered a range of efforts to establish the social relations perceived as necessary for researchers to gain access to such materials. Emblematic of these efforts are calls to make participation in medical research an obligation of citizenship (Schaefer et al., 2009) and attempts to establish genetic biobanks in diverse locales, including both national banks – e.g. Iceland, the United Kingdom, Japan, Italy, Sweden, and Taiwan – and banks associated with specific medical centers – e.g. the Mayo Clinic, Vanderbilt University Medical Center, and the recently established Million Veterans Project.

In some locales, efforts to encourage participation in such biobanking have led to a range of projects that aim to invite diverse citizens to engage in genetic thinking and practice. That is, they attempt to educate people about the new genetics in part to encourage them to participate in genetic research or to contribute to a genetic biobank. In the United States, for example, the National Institutes of Health (NIH) recently contracted with the Johns Hopkins University's Genetic & Public Policy Center to conduct focus groups and a set of national town hall meetings to engage a wide array of U.S. citizens in a conversation about whether and under what conditions they would be willing to participate in a national NIH-based genetic biobank. In the context of the Hapmap project discussed above, significant attention was paid to engaging members of target communities in conversations about the effort prior to collecting the genetic samples that would be made into cell lines for the project's biobank.

At the same time, lack of knowledge about genetics among physicians is widely seen as inhibiting the translation of new genetic knowledge into clinical practice. Repeated studies have shown that, as a group, physicians have little sense of when to order genetic tests or how to interpret them (e.g., Gaff et al., 2007; Hayflick et al., 1998; Hunter et al., 1998). Significant efforts are now underway to incorporate the new genetics more systematically into medical education. Such education is also viewed as a crucial component of establishing the social relations necessary for the flourishing of the new genetics, because physicians are often the ones who will refer patients to participate in genomic research or draw their blood for inclusion in a biobank.

Genetic Citizenship

The concept of genetic citizenship (Heath et al., 2004; see also Taussig et al., 2003) points to classic issues of citizenship having to do with the politics of inclusion and exclusion. Deborah Heath, Rayna Rapp, and I explore the reconfiguration of activism in the context of citizenship claims made by lay advocacy organizations in relation to the new genetics.

Knowledge and technologies associated with the new genetics persistently raise questions about belonging, as people are faced with reproductive choices about genetic testing, and about what makes an acceptable member of their family and of the human species that such testing provokes. National policies that shape the public (national health insurance) and private (corporate insurance) logics behind the coverage of such medical procedures also come in to play here, since the calculus of covering prenatal genetic testing involves balancing that cost with the cost of covering the care of a fragile child and it always occurs in the context of the politics of access to abortion.

In the United States and elsewhere, this push to perfectibility has emerged virtually simultaneously with the disability rights movement and the claims for dignity and inclusion of which lay advocacy groups supporting those with genetic conditions are actively involved. It is, for example, in this context that the lay advocacy organization for people of short stature produced a t-shirt emblazoned with the words 'Dwarf: Endangered Species' upon the publication of the discovery of the FGFR3 gene responsible for achondroplasia, the most common form or dwarfism in humans. While many people who have this dwarfing condition welcome this finding because it allows them to undergo prenatal testing to discover whether they are carrying a fetus with the nearly always fatal (at birth) condition of double dominance (two copies of the already dominant gene for achondroplasia), the t-shirt points to their recognition that this knowledge may be used for eugenic purposes by those in the wider population (Taussig et al., 2003).

In developing the concept of genetic citizenship, Heath, Rapp, and I point to how the experiences of those at the vanguard of genetic decision-making – whether researchers, clinicians, individuals or families bearing distinctive genetic conditions, or members of the lay public – connect "discussions of rights, recognitions, and responsibilities to intimate, fundamental concerns about heritable identities, differential embodiment, and an *ethics of care*" (Heath et al., 2004: 157 italics in the original). Genetic citizenship points both to "the political economy of state funding and regulation and the unruly aspirations of scientists, health advocates, health-care service providers, and families, sometimes struggling against the market, sometimes seeking, whether ardently or ambivalently, to join it in an effort to imagine other possible futures" (p. 165). Those involved in lay advocacy organizations associated with genetic conditions undertake crucial work in care, research, and policy. They often fund significant research on the condition with which they are concerned; facilitate essential access to the genetic material, family histories, and medical records so crucial to understanding the epidemiology of a particular condition and to discovery of the genetic sequence to which it is connected; intervene with legislators to ensure policy supporting that research; and even work to shape commercial aspects of that research through claiming a share in resulting intellectual property. In the activism of those involved in lay advocacy organizations, we see the complex ways the new genetics implicates citizenship and subjectivity as it opens up new or newly configured domains for political engagement and demands for rights and inclusion.

See also: Anthropology, Genomics, and Human Variation: National Roots; Biobanking: Ethical Issues; Bioethics in the Post-genomic Era; Bioethics: Genetics and Genomics;

Bioethics: Philosophical Aspects; Direct-to-Consumer Personal Genetic Testing; Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute; Eugenics as an International Movement; Eugenics in Europe; Eugenics in the Americas; Eugenics, History of; Genetic Ancestry Testing; Genetics and Anthropology; Genetics and Disaster Victim Identification; Genetics and Social Justice; Genetics and Society; Genetics, Disease, and Reproduction; Genomics, Ethical Issues in; Human Genome Diversity Project: History; Human Genome Project: History and Assessment; New Genetics and Race; Race: Genetic Aspects.

Bibliography

Abu El-Haj, N., 2007. The genetic reinscription of race. Annual Review of Anthropology 36, 283–300.

Bolnick, D.A., Fullwiley, D., Duster, T., Cooper, R.S., Fujimura, J., et al., 2007. The science and business of genetic ancestry testing. Science 318, 399–400.

Brown, N., 2003. Hope against hype – accountability in biopasts, presents and futures. Science Studies 16, 3–21.

Brown, N., Michael, M., 2003. A sociology of expectations: retrospecting prospects and prospecting retrospects. Technology Analysis & Strategic Management 15 (1), 3–18.

Bunton, R., Peterson, A. (Eds.), 2005. Genetic Governance: Health, Risk and Ethics in the Biotech Era. Psychology Press, New York.

Clarke, A., Fujimura, J., 1992. The Right Tools for the Job: At Work in 20th-Century Life Sciences. Princeton University Press, Princeton, NJ.

Cohen, S., Chang, A., Boyer, H., Helling, R., 1973. Construction of biologically functional bacterial plasmids *in vitro*. Proceedings of the National Academy of Sciences of the United States of America 70 (11), 3240–3244.

Correns, C., 1900. G. Mendel's regel über das verhalten der nachkommenschaft der rassenbastarde. Berichte der deutschen botanischen Gesellschaft 18, 158–168.

Duster, T., 2006. The molecular reinscription of race: unanticipated issues of biotechnology and forensic science. Patterns of Prejudice 40 (4–5), 427–441.

Duster, T., 2007. Medicalisation of race. Lancet 369, 702–704.

Ekberg, M., 2007. Second opinions: the old eugenics and the new genetics compared. Social History of Medicine 20 (3), 581–593.

Epstein, S., 2007. Inclusion: The Politics of Difference in Medical Research. University of Chicago Press, Chicago.

Fortun, M., 2008. Promising Genomics: Iceland and DeCODE Genetics in a World of Speculation. University of California Press, Berkeley.

Franklin, S., 2001. Culturing biology: cell lines for the second millennium. Health 5 (3), 335–354.

Fujimura, J.H., Duster, T., Rajagopalan, R., 2008. Introduction: race, genetics, and disease. Social Studies of Science 38, 643–656.

Fullwiley, D., 2007. The molecularization of race: institutionalizing human difference in pharmacogenetics practice. Science as Culture 16, 1–30.

Gaff, C., Aitken, M.A., Flouris, A., Metcalfe, S., 2007. A model for the development of genetics education programs for health professionals. Genetics in Medicine 9, 451–457.

Gottweis, H., 2002. The governance of genomics. Critical Public Health 12 (3), 2007–2220.

Gottweis, H., 2005. Governing genomics in the 21st century: between risk and uncertainty. New Genomics and Society 24 (2), 175–194.

Hartigan, J., 2008. Is race still socially constructed? The recent controversy over race and medical genetics. Science as Culture 17, 163–193.

Hayflick, S., Eiff, P., Carpenter, L., Steinberger, J., 1998. Primary care physicians' utilization and perceptions of genetic services. Genetic Medicine 1, 13–21.

Heath, D., Rapp, R., Taussig, K.S., 2004. Genetic citizenship. In: Nugent, D., Vincent, J. (Eds.), A Companion to the Anthropology of Politics. Blackwell, Oxford, pp. 152–167.

Helmreich, S., 2008. Species of biocapital. Science as Culture 17 (4), 463–478.

Helmreich, S., 2009. Alien Ocean: Anthropological Voyages in Microbial Seas. University of California Press, Berkeley.

Hindmarsh, R., Gottweis, H., 2005. Recombinant regulation: the Asilomar legacy 30 years on. Science as Culture 14 (4), 299–307.

Hindmarsh, R., Prainsack, B. (Eds.), 2010. Genetic Suspects: Global Governance of Forensic DNA Profiling and Databasing. Cambridge University Press, Cambridge.

Hoeyer, K., Tutton, R., 2005. "Ethics was here": studying the language-games of ethics in the case of UK Biobank. Critical Public Health 15 (4), 385–397.

Hunter, A., Wright, P., Cappelli, M., Kasaboski, A., Surh, L., 1998. Physician knowledge and attitudes towards molecular genetic (DNA) testing of their patients. Clinical Genetics 53, 447–455.

Kahn, J., 2009. Race, genes, and justice: a call to reform the presentation of forensic DNA evidence in criminal trials. Brooklyn Law Review 74 (2), 325–375.

Kahn, J., 2012. Race in a Bottle: The Story of BiDil and Racialized Medicine in a Post-Genomic Age. Columbia University Press, New York.

Kevles, D.J., 1985. In the Name of Eugenics. University of California Press, Berkeley.

Kevles, D.J., Hood, L. (Eds.), 1992. The Code of Codes: Scientific and Social Issues in the Human Genome Project. Harvard University Press, Cambridge, MA.

Koenig, B.A., Lee, S.S.-J., Richardson, S. (Eds.), 2008. Revisiting Race in a Genomic Age. Rutgers University Press, Piscataway, NJ.

Krimsky, S., Sloan, K. (Eds.), 2011. Race and the Genetic, Myth, and Culture. Columbia University Press, New York.

Landecker, H., 2007. Culturing Life: How Cells Became Technologies. Harvard University Press, Cambridge, MA.

Lombardo, P., 2008. Three Generations, No Imbeciles: Eugenics, the Supreme Court and Buck V. Bell. Johns Hopkins University Press, Baltimore, MD.

Marks, J., 1995. Human Biodiversity: Genes, Race, and History. Aldine de Gruyter, New York.

Marks, J., 2002. What It Means to Be 98% Chimpanzee. University of California Press, Berkeley.

Mendel, G., 1866. Versuche uber pflanzenhybriden. Verhandlungen des Naturforschenden Vereines in Brunn 4, 3–47.

Meslin, E.M., Thomson, E.J., Boyer, J.T., 1997. Bioethics inside the beltway: the ethical, legal, and social implications research program at the National Human Genome Research Institute. Kennedy Institute of Ethics Journal 7 (3), 291–298.

Paul, D., 1995. Controlling Human Heredity: 1865 to the Present. Humanity Books, Amherst, NY.

Proctor, R., 1990. Racial Hygiene: Medicine under the Nazis. Harvard University Press, Cambridge, MA.

Rajan, K.S., 2007. Biocapital: The Constitution of Post-Genomic Life. Duke University Press, Durham, NC.

Rapp, R., 1999. Testing Women, Testing the Fetus: The Social Impact of Amniocentesis in America. Routledge, New York.

Reardon, J., 2004. Race to the Finish: Identity and Governance in an Age of Genomics (In-formation). Princeton University Press, Princeton, NJ.

Rheinberger, H.-J., 1995. Beyond nature and culture: a note on medicine in the age of molecular biology. Science in Context 8 (1), 249–263.

Roberts, D., 2011. Fatal Invention: How Science, Politics, and Big Business Re-create Race in the Twenty-First Century. New Press, New York.

Root, M., 2003. The use of race in medicine as a proxy for genetic differences. Philosophy of Science 70, 1173–1183.

Rose, N., Novas, C., 2005. Biological citizenship. In: Ong, A., Collier, S.J. (Eds.), Global Assemblages: Technology, Politics, and Ethics as Anthropological Problems. Blackwell, Oxford, pp. 439–463.

Schaefer, G.O., Emanuel, E., Wertheimer, A., 2009. The obligation to participate in biomedical research. Journal of the American Medical Association 302 (1), 67–72.

Sinsheimer, C., April 1969. The prospect of designed genetic change. Engineering and Science 32, 8.

Stepan, N.L., 1991. The Hour of Eugenics: Race, Gender and Nation in Latin America. Cornell University Press, Ithaca, NY.

Stoller, A., 1997. Tensions of Empire: Colonial Cultures in a Bourgeois World. University of California Press, Berkeley.

Taussig, K.S., Rapp, R., Heath, D., 2003. Flexible eugenics: technologies of the self in the age of genetics. In: Goodman, A., Heath, D., Lindee, M.S. (Eds.), Genetic Nature/Culture: Anthropology and Science beyond the Two Culture Divide. University of California Press, Berkeley, pp. 58–76.

Taussig, K.S., Hoeyer, K., Helmreich, S., 2013. The anthropology of potentiality in biomedicine: an introduction to supplement 7. Current Anthropology 54 (S7), S3–S14.

Tutton, R., 2011. Promising pessimism: reading the futures to be avoided in biotech. Social Studies of Science 41 (3), 411–429.

de Vries, H., 1900. Sur la loi de disjunction des hybrids. Comptes Rendus de l'Academie des Sciences 130, 845–847.

Wade, P., Beltrán, C.L., Restrepo, E., Santos, R.V. (Eds.), 2014. Mestizo Genomics: Race Mixture, Nation, and Science in Latin America. Duke University Press, Durham, NC.

Watson, J., Crick, F., 1953. A structure for deoxyribose nucleic acid. Nature 171, 737–738.

Wailoo, K., Nelson, A., Lee, C. (Eds.), 2012. Genetics and the Unsettled Past: The Collision of DNA, Race, and History. Rutgers University Press, Piscataway, NJ.

Weiner, C., 2000. Recombinant DNA, policy, Asilomar. In: Murray, T., Mehlman, M. (Eds.), Encyclopedia of Ethical, Legal, and Policy Issues in Biotechnology. John Wiley & Sons, New York, pp. 909–914.

Wilmut, I., Schnieke, A.E., McWhir, J., Kind, A.J., Campbell, K.H., 27 February 1997. Viable offspring derived from fetal and adult mammalian cells. Nature 385, 810–813.

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 10

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser—Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 10

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 10 *xliii*

VOLUME 10

Genocide and War 1
Martin Shaw

Genocide, Historical Aspects of 6
Gunnar Heinsohn

Genocide: Social and Economic Aspects 15
Ross Kleinstuber

Genographic Project 22
Theodore G Schurr

Genomics, Ethical Issues in 32
Henry T Greely

Genotype and Phenotype 42
Richard C Lewontin

Gentrification 46
Loretta Lees

Geoarchaeology 53
Marshall I Weisler and Serena Love

Geographic Information Systems 58
Michael F Goodchild

Geographic Information Systems and Remote Sensing 64
Steven M Manson, Dudley B Bonsal, Melinda Kernik, and Eric F Lambin

Geographic Thought, History of 69
Paul Claval

Geographies of Food: Politics, Place, Power Reprised 76
Nick Lewis and Richard Le Heron

Geography: Overview 83
Ron Johnston

Geopolitics 91
Klaus Dodds and Chih Yuan Woon

Geospatial Applications in Archaeology 97
Devin A White and Katie A Corcoran

Geriatric Psychiatry 104
Florian Riese

Gesture in Linguistics 109
David McNeill

Ghetto 121
Loic Wacquant

Gibson's Theory of Perceptual Learning 127
Karen E Adolph and Kari S Kretch

Giddens, Anthony (1938–) 135
Steven Loyal

Gift: History of the Concept 139
Christian Papilloud

Global Agenda for Social Work and Social Development 144
Rory Truell and David N Jones

Global Development: Markets and Wellbeing 150
Alberto Arce

Global Governance 155
Kennette Benedict

Global History: Universal and World 162
PK O'Brien

Global Migration 169
Jonathan Crush

Global Positioning System 174
Mauro Salvemini

Global Public Policy 178
Inge Kaul

Global South 186
Don Kalb and Luisa Steur

Global Violence Against Women 192
Elizabeth A Swart

Globalization and Health 198
Ronald Labonté

Globalization and Human Trafficking 206
Mónica Méndez

Globalization and Religion 213
Peter Beyer

Globalization and Transnational Corporations 218
Alessandro Bonanno

Globalization and World Culture 225
John Boli and Frank J Lechner

Globalization, Anthropology of 233
Arjun Appadurai

Globalization: Geographical Aspects 239
Andrew Jones

Globalization in Education 247
Yong Zhao

Globalization: Legal Aspects 254
Alfred C Aman, Jr

Globalization: Welfare and Social Work 262
Karen Lyons

Goal-Oriented Behaviour Among Human and Non-Human Primates 268
Heidi L Marsh and Maria Legerstee

Goffman, Erving (1922–1982) 275
Herbert Willems

Goldthorpe, John H. (1935–) 279
Christopher T Whelan

Gorz, André (1923–2007) 286
Claus Leggewie

Governance 292
Gunnar F Schuppert

Government Statistics 301
M Straf

Governmentality 308
Marc Abélès

Governments 313
Jean Blondel

Grammatical Agreement 316
Dieter Wunderlich

Grammatical Relations 324
Wei-wen Roger Liao

Grammaticalization 330
Paolo Ramat

Grandparent–Grandchild Relationships 336
Teun Geurts and Theo van Tilburg

Graphical Markov Models: Overview 341
Nanny Wermuth and David R Cox

Graphical Methods, Analytic 351
Antony Unwin

Graphical Methods: Presentation 358
Leland Wilkinson

Green Jobs 372
Olivier Deschenes

Green Parties 379
Neil Carter

Green Social Work 385
Lena Dominelli

Greening of Business 392
John A Mathews

Grit 397
Angela L Duckworth and Lauren Eskreis-Winkler

Grounded Theory: Methodology and Theory Construction 402
Kathy Charmaz

Group Processes in Organizations 408
Susanne Braun, Dieter Frey, Felix Claus Brodbeck, and Tanja Hentschel

Group Processes in the Classroom 416
Angela M O'Donnell

Group Processes, Social Psychology of 422
Monique H Matelski and Michael A Hogg

Group Psychotherapy, Clinical Psychology of 428
Donelson R Forsyth

Groups, Sociology of 434
Siegwart Lindenberg

Growth Curve Analysis 441
John J McArdle

Growth Management and Urban Planning in the United States 447
Arthur C Nelson

Guilt 453
Amy Farrell

Guns and Criminals 457
James D Wright and Rachel L Rayburn

Gynecological Health: Psychosocial Aspects 462
Pearl La Marca-Ghaemmaghami and Ulrike Ehlert

Habermas, Jürgen (1929–) 469
William Outhwaite

Habit: History of the Concept 475
Charles Camic

Habituation 480
RF Thompson

Habitus: History of a Concept 484
Gisèle Sapiro

Halévy, Élie (1870–1937) 490
Michael Hill

Hall, Granville Stanley (1844–1924) 495
Lester F Goodchild

Hallucinations, Neural Basis of 499
Rachael Goldsworthy and Harry Whitaker

Handedness and Language, Relationship between 504
Brigitte Stemmer

Happiness and Work 515
Annabelle Krause

Happiness: History of the Concept 521
Ruut Veenhoven

Harlow, Harry Frederick (1905–81) 526
Stephen J Suomi

Harmony and Social Work Practice 529
Terry Tse Fong Leung

Hart, Herbert Lionel Adolphus (1907–92) 536
Nicola Lacey

Hate Crimes 540
Jack Levin and Jack McDevitt

Hazards and Disaster Research in Contemporary Anthropology 546
Anthony Oliver-Smith

Hazing in Fraternities and Sororities 554
Hank Nuwer

Healing Practices: Alternative and Complementary 562
Adrian Furnham

Health and Illness: Mental Representations in Different Cultures 567
Ross G White

Health: Anthropological Aspects 571
Adriana Petryna

Health Behavior, Psychosocial Theories of 577
Stephen Sutton

Health Behaviors 582
Mark Conner

Health Behaviors, Assessment of 588
Britta Renner, Verena Klusmann, and Gudrun Sproesser

Health Care: Legal Issues (United States) 594
Nicole Huberfeld

Health Care Markets: Theory and Empirical Evidence 598
Stefan K Lhachimi and Johannes Siegrist

Health Care, Rationing of 603
David Mechanic

Health: Definitions 607
Machteld Huber

Health Economics 614
Åke G Blomqvist

Health Education and Health Promotion 620
Gerjo Kok and Nanne K de Vries

Health, Geography of 628
Jennifer O'Brien

Health in Low-Income Countries 633
Solomon R Benatar

Health Insurance: Economic and Risk Aspects 640
Conchita D'Ambrosio and Simone Ghislandi

Health Interventions, Community-based 646
John W Farquhar

Health Literacy 653
Kristine Sørensen

Health Outcomes, Assessment of 657
Robert M Kaplan

Health Politics and Policy 663
Claus Wendt

Health Promotion in Schools 669
Inge Seiffge-Krenke and Annika Krick

Health Psychology 674
Denise de Ridder

Health Psychology in Childhood 679
Petra Warschburger

Health-Related Support and Self-help Mutual Aid Groups 686
Benjamin H Gottlieb and Karla D Stroud

Health Research, Qualitative 692
Thorsten Meyer

Health Risk Appraisal and Optimistic Bias 698
Neil D Weinstein and William MP Klein

Health Risk Perception 702
Britta Renner, Martina Gamp, Ralf Schmälzle, and Harald T Schupp

Health Self-Regulation, Motivational and Volitional Aspects of 710
Ralf Schwarzer

Health Social Work 716
Cecilia Lai Wan Chan and Bo Yan

Health Surveys 725
Jennifer H Madans and Kassi M Webster

Hebb, Donald Olding (1904–1985) 731
Peter M Milner

Hegel, Georg Wilhelm Friedrich (1770–1831) 736
Miriam Wildenauer

Hegemony and Cultural Resistance 742
Sujeong Kim

Hegemony: Anthropological Aspects 747
Donald V Kurtz

Heider, Fritz (1896–1988) 750
Bernard Weiner

Helmholtz, Hermann Ludwig Ferdinand von (1821–94) 755
David Cahan

Helping Intergroup Relations 759
Arie Nadler and Samer Halabi

Hemispatial Neglect, Neural Basis of 766
Marlene Behrmann and Sarah Shomstein

Hempel, Carl Gustav (1905–97) 773
Gereon Wolters

Henry, Louis (1911–91) 776
Etienne van de Walle

Hermeneutics 778
Oliver R Scholz

Hermeneutics, History of 785
William Outhwaite

Heterosexism and Homophobia 790
Jodi O'Brien

Heterosexual Expression at Older Ages 796
Amelia Karraker

Heterosexuality 801
James Iveniuk and Edward O Laumann

Heterosexuality in Africa 812
Signe Arnfred

Heterosexuality in China 819
Emil Man-Lun Ng

Heuristics in Social Cognition 825
Sascha Topolinski and Fritz Strack

Heuristics, History of 829
Ralph Hertwig and Thorsten Pachur

Hierarchical Models: Random and Fixed Effects 836
Dennis V Lindley

Hierarchies and Markets 841
Oliver E Williamson

Higher Education Management 845
Juergen Enders

Higher Education Market 850
Felix Maringe

Higher Education Research 862
Ulrich Teichler

Hilberg, Raul (1926–2007) 870
Detlev Claussen

Hintze, Otto (1861–1940) 873
Witold Neugebauer

Hippocampal Amnesia 876
Andrew Mayes

Hippocampus and Related Structures 886
Menno P Witter

Hirschman, Albert O. (1915–2012) 892
Jeremy Adelman

CONTRIBUTORS TO VOLUME 10

Marc Abélès
Ecole des Hautes Etudes en Sciences Sociales, Paris, France

Jeremy Adelman
Princeton University, Princeton, NJ, USA

Karen E. Adolph
New York University, New York, NY, USA

Alfred C. Aman
Indiana University Maurer School of Law, Bloomington, IN, USA

Arjun Appadurai
NYU Steinhardt, Media, Culture, and Communication, New York, NY, USA

Alberto Arce
Wageningen University and Research Centre, Wageningen, The Netherlands

Signe Arnfred
Centre for Gender, Power and Diversity, Roskilde University, Roskilde, Denmark

Marlene Behrmann
Carnegie Mellon University, Pittsburgh, PA, USA

Solomon R. Benatar
University of Cape Town, Observatory, South Africa; and University of Toronto, Toronto, ON, Canada

Kennette Benedict
Bulletin of the Atomic Scientists, Chicago, IL, USA

Peter Beyer
University of Ottawa, Ottawa, ON, Canada

Åke G. Blomqvist
University of Western Ontario, ON, Canada; and Carleton University, Canada; Ottawa, ON, Canada

Jean Blondel
European University Institute, Florence, Italy

John Boli
Emory University, Atlanta, GA, USA

Alessandro Bonanno
Sam Houston State University, Huntsville, TX, USA

Dudley B. Bonsal
University of Minnesota, Minneapolis, MN, USA

Susanne Braun
Ludwig-Maximilians-Universität München, Munich, Germany; and Technische Universität München, Munich, Germany

Felix Claus Brodbeck
Ludwig-Maximilians-Universität München, Munich, Germany

David Cahan
University of Nebraska-Lincoln, Lincoln, NE, USA

Charles Camic
Northwestern University, Evanston, IL, USA

Neil Carter
University of York, Heslington, York, UK

Kathy Charmaz
Sonoma State University, Rohnert Park, CA, USA

Detlev Claussen
Institute for Sociology, Leibniz University Hannover, Hannover, Germany

Paul Claval
Université Paris-Sorbonne, Paris, France

Mark Conner
Institute of Psychological Sciences, University of Leeds, Leeds, UK

Katie A. Corcoran
University of Tennessee, Knoxville, TN, USA

David R. Cox
Nuffield College, Oxford University, Oxford, UK

Jonathan Crush
Balsillie School of International Affairs, Waterloo, ON, Canada; and University of Cape Town, Cape Town, South Africa

Conchita D'Ambrosio
Université du Luxembourg, Walferdange, Luxembourg

Denise de Ridder
Utrecht University, Utrecht, The Netherlands

Olivier Deschenes
University of California Santa Barbara, Santa Barbara, CA, USA; IZA, Bonn, Germany; and National Bureau of Economic Research (NBER), Cambridge, MA, USA

Nanne K. de Vries
Maastricht University, Maastricht, Netherlands

Klaus Dodds
Royal Holloway University of London, Egham, Surrey, UK

Lena Dominelli
Durham University, Durham, UK

Angela L. Duckworth
The University of Pennsylvania, Philadelphia, PA, USA

Ulrike Ehlert
Institute of Psychology, University of Zurich, Zurich, Switzerland

Juergen Enders
School of Management, University of Bath, Bath, UK

Lauren Eskreis-Winkler
The University of Pennsylvania, Philadelphia, PA, USA

John W. Farquhar
Stanford University, Stanford, CA, USA

Amy Farrell
School of Criminology and Criminal Justice, Northeastern University, Boston, MA, USA

Donelson R. Forsyth
Jepson School of Leadership Studies, University of Richmond, Richmond, VA, USA

Dieter Frey
Ludwig-Maximilians-Universität München, Munich, Germany

Adrian Furnham
University College London, London, UK

Martina Gamp
University of Konstanz, Konstanz, Germany

Teun Geurts
Ministry of Security and Justice, Research and Documentation Centre, The Hague, The Netherlands

Simone Ghislandi
Università Bocconi, Milan, Italy

Rachael Goldsworthy
Central Michigan University, Mt. Pleasant, MI, USA

Lester F. Goodchild
University of Denver, CO, USA

Michael F. Goodchild
University of California, Santa Barbara, CA, USA

Benjamin H. Gottlieb
University of Guelph, Guelph, ON, Canada

Henry T. Greely
Center for Law and the Biosciences, Stanford University, CA, USA

Samer Halabi
School of Behavioral Sciences, Tel Aviv — Jaffa Academic College, Jaffa, Israel

Gunnar Heinsohn
Raphael Lemkin Institute for Comparative Genocide Research, University of Bremen, Bremen, Germany

Tanja Hentschel
Technische Universität München, Munich, Germany

Ralph Hertwig
Max Planck Institute for Human Development, Center for Adaptive Rationality, Berlin, Germany

Michael Hill
Victoria University, Wellington, New Zealand

Michael A. Hogg
Claremont Graduate University, Claremont, CA, USA

Machteld Huber
Louis Bolk Institute, Driebergen, The Netherlands

Nicole Huberfeld
University of Kentucky, Lexington, KY, USA

James Iveniuk
NORC PRC, Chicago, IL, USA

Ron Johnston
University of Bristol, Bristol, UK

Andrew Jones
School of Arts and Social Sciences, City University London, Northampton Square, London, UK

David N. Jones
People Need People Consulting, Northampton, UK

Don Kalb
Central European University, Budapest, Hungary; and Utrecht University, Utrecht, The Netherlands

Robert M. Kaplan
University of California, San Diego, CA, USA

Amelia Karraker
University of Michigan, Ann Arbor, MI, USA

Inge Kaul
Hertie School of Governance, Berlin, Germany

Melinda Kernik
University of Minnesota, Minneapolis, MN, USA

Sujeong Kim
University of California, San Diego, CA, USA

William M.P. Klein
National Cancer Institute, NIH, Bethesda, MD, USA

Ross Kleinstuber
University of Pittsburgh at Johnstown, Johnstown, PA, USA

Verena Klusmann
University of Konstanz, Konstanz, Germany

Gerjo Kok
Maastricht University, Maastricht, Netherlands

Annabelle Krause
Institute for the Study of Labor (IZA), Bonn, Germany

Kari S. Kretch
New York University, New York, NY, USA

Annika Krick
Johannes Gutenberg-Universität, Mainz, Germany

Donald V. Kurtz
University of Wisconsin, Milwaukee, WI, USA; University of Texas, San Antonio, TX, USA; and Southwest Texas State University, San Marcos, TX, USA

Ronald Labonté
University of Ottawa, Ottawa, ON, Canada

Nicola Lacey
All Souls College, Oxford, UK

Pearl La Marca-Ghaemmaghami
Institute of Psychology, University of Zurich, Zurich, Switzerland

Eric F. Lambin
Georges Lemaître Centre for Earth and Climate Research, Earth and Life Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium; and School of Earth Sciences and Woods Institute for the Environment, Stanford University, Stanford, CA, USA

Edward O. Laumann
University of Chicago, Chicago, IL, USA

Frank J. Lechner
Emory University, Atlanta, GA, USA

Loretta Lees
King's College London, Strand, London, UK

Maria Legerstee
York University, Toronto, ON, Canada

Claus Leggewie
Kulturwissenschaftliches Institut Essen (KWI), Essen, Germany

Richard Le Heron
School of Environment, University of Auckland, Auckland, New Zealand

Terry Tse Fong Leung
The Chinese University of Hong Kong, Hong Kong, People's Republic of China

Jack Levin
Northeastern University, Boston, MA, USA

Nick Lewis
School of Environment, University of Auckland, Auckland, New Zealand

Richard C. Lewontin
Harvard University, Cambridge, MA, USA

Stefan K. Lhachimi
University of Bremen, Bremen, Germany; and Leibniz Institute for Preventive Research and Epidemiology, Bremen, Germany

Wei-wen Roger Liao
Institute of Linguistics, Academia Sinica, Taipei, Taiwan

Siegwart Lindenberg
University of Groningen, Groningen, The Netherlands; and Tilburg University, Tilburg, The Netherlands

Dennis V. Lindley[†]
Minehead, Somerset, UK

Serena Love
School of Social Science, University of Queensland, QLD, Australia

Steven Loyal
School of Sociology, University College Dublin, Dublin, Ireland

Karen Lyons
London Metropolitan University, London, UK

Jennifer H. Madans
National Center for Health Statistics, Hyattsville, MD, USA

[†] Deceased.

Steven M. Manson
University of Minnesota, Minneapolis, MN, USA

Felix Maringe
Division of Education Leadership and Policy Studies, Faculty of Humanities, School of Education, University of the Witwatersrand, South Africa

Heidi L. Marsh
University of Western Ontario, London, ON, Canada

Monique H. Matelski
Claremont Graduate University, Claremont, CA, USA

John A. Mathews
MGSM Macquarie University, Sydney, NSW, Australia

Andrew Mayes
School of Psychological Sciences, University of Manchester, Manchester, UK

John J. McArdle
University of Virginia, Charlottesville, VA, USA

Jack McDevitt
Northeastern University, Boston, MA, USA

David McNeill
University of Chicago, Chicago, IL, USA

David Mechanic
Rutgers, The State University of New Jersey, Institute for Health, Health Care Policy and Aging Research, New Brunswick, NJ, USA

Mónica Méndez
House of Mentoring & Empowerment, Inc., Orlando, FL, USA

Thorsten Meyer
Institute for Epidemiology, Social Medicine, and Health Systems Research, Hannover Medical School, Hannover, Germany

Peter M. Milner

Arie Nadler
Tel Aviv University, Ramat-Aviv, Israel

Arthur C. Nelson
University of Arizona, Tucson, AZ, USA

Witold Neugebauer
University of Wurzburg, Wurzburg, Germany

Emil Man-Lun Ng
Academy of Family Therapy, Hong Kong SAR, China

Hank Nuwer
Franklin College, Franklin, IN, USA

Jennifer O'Brien
School of Environment, Education and Development, University of Manchester, Manchester, UK

Jodi O'Brien
Seattle University, Seattle, WA, USA

P.K. O'Brien
London School of Economics, University of London, London, UK

Angela M. O'Donnell
Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

Anthony Oliver-Smith
University of Florida, Gainesville, FL, USA

William Outhwaite
Newcastle University, Newcastle upon Tyne and wear, UK

Thorsten Pachur
Max Planck Institute for Human Development, Center for Adaptive Rationality, Berlin, Germany

Christian Papilloud
Institute for Sociology, Martin-Luther University in Halle-Wittenberg, Germany

Adriana Petryna
University of Pennsylvania, Philadelphia, PA, USA

Paolo Ramat
Center for Postgraduate Education and Research LETiSS, Istituto Universitario di Studi Superiori (IUSS) (Retired), Pavia, Italy

Rachel L. Rayburn
Indiana University Purdue University — Fort Wayne, Fort Wayne, IN, USA

Britta Renner
University of Konstanz, Konstanz, Germany

Florian Riese
Psychiatric University Hospital Zurich, Zurich, Switzerland

Mauro Salvemini
Sapienza Università di Roma, Rome, Italy

Gisèle Sapiro
École des hautes études en sciences sociales and Research, Paris, France; and CNRS (Centre européen de sociologie et de science politique), Paris, France

Ralf Schmälzle
University of Konstanz, Konstanz, Germany

Oliver R. Scholz
Westfälische Wilhelms-Universität Münster, Münster, Germany

Harald T. Schupp
University of Konstanz, Konstanz, Germany

Gunnar F. Schuppert
Social Science Research Center Berlin, Berlin, Germany

Theodore G. Schurr
University of Pennsylvania, Philadelphia, PA, USA

Ralf Schwarzer
Freie Universität Berlin, Berlin, Germany

Inge Seiffge-Krenke
Johannes Gutenberg-Universität, Mainz, Germany

Martin Shaw
Institut Barcelona d'Estudis Internacionals, International Relations and Human Rights, University of Roehampton, London, UK; and International Relations, University of Sussex, Brighton, UK

Sarah Shomstein
George Washington University, Washington, DC, USA

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany

Kristine Sørensen
CAPHRI, Maastricht University, Maastricht, Netherlands

Gudrun Sproesser
University of Konstanz, Konstanz, Germany

Brigitte Stemmer
Université de Montréal, Montreal, QC, Canada; Brock University, St. Catherines, ON, Canada; and CRIUGM, Montreal, QC, Canada

Luisa Steur
University of Copenhagen, Copenhagen, Denmark

Fritz Strack
University of Wuerzburg, Wuerzburg, Germany

M. Straf
National Research Council, Washington, DC, USA

Karla D. Stroud
University of Guelph, Guelph, ON, Canada

Stephen J. Suomi
National Institute of Health, Bethesda, MD, USA

Stephen Sutton
Institute of Public Health, University of Cambridge, Cambridge, UK

Elizabeth A. Swart
University of Southern California, Los Angeles, CA, USA

Ulrich Teichler
International Centre of Higher Education Research Kassel (INCHER-Kassel), University of Kassel, Kassel, Germany

R.F. Thompson
University of Southern California, Los Angeles, CA, USA

Sascha Topolinski
University of Cologne, Cologne, Germany

Rory Truell
International Federation of Social Workers, Bern, Switzerland

Antony Unwin
University of Augsburg, Augsburg, Germany

Etienne van de Walle[†]
University of Pennsylvania, Philadelphia, PA, USA

Theo van Tilburg
VU University Amsterdam, Amsterdam, The Netherlands

Ruut Veenhoven
Erasmus University Rotterdam, Rotterdam, The Netherlands; and North-West University, South Africa

Loic Wacquant
University of California, Berkeley, CA, USA

Cecilia Lai Wan Chan
The University of Hong Kong, Hong Kong

Petra Warschburger
University of Potsdam, Potsdam, Germany

Kassi M. Webster
National Center for Health Statistics, Hyattsville, MD, USA

Bernard Weiner
University of California Los Angeles, Los Angeles, CA, USA

Neil D. Weinstein
Rutgers University, Tucson, AZ, USA

Marshall I. Weisler
School of Social Science, University of Queensland, QLD, Australia

[†] Deceased.

Claus Wendt
University of Siegen, Siegen, Germany

Nanny Wermuth
Chalmers University of Technology, Gothenburg, Sweden and Medical Psychology and Medical Sociology, Gutenberg-University, Mainz, Germany

Christopher T. Whelan
School of Sociology, Social Policy & Social Work, Queen's University Belfast, Ireland; and School of Sociology & Geary Institute, University College Dublin, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA

Devin A. White
Oak Ridge National Laboratory, Oak Ridge, TN, USA

Ross G. White
Institute of Health and Well-being, University of Glasgow, Glasgow, UK

Miriam Wildenauer
University of Heidelberg, Heidelberg, Germany

Leland Wilkinson
SPSS Inc., Chicago, IL, USA

Herbert Willems
University of Giessen, Germany

Oliver E. Williamson
University of California, Berkeley, CA, USA

Menno P. Witter
Norwegian University of Science and Technology NTNU, Trondheim, Norway

Gereon Wolters
Universität Konstanz, Konstanz, Germany

Chih Yuan Woon
National University of Singapore, Kent Ridge, Singapore

James D. Wright
University of Central Florida, Orlando, FL, USA

Dieter Wunderlich
University of Düsseldorf, Germany; and Centre for General Linguistics (ZAS) at Berlin, Germany

Bo Yan
The University of Hong Kong, Hong Kong

Yong Zhao
University of Oregon, OR, USA; and Victoria University, VIC, Australia

Genocide and War

Martin Shaw, Institut Barcelona d'Estudis Internacionals, International Relations and Human Rights, University of Roehampton, London, UK; and International Relations, University of Sussex, Brighton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article considers genocide as a general type of political violence and violent conflict, together with its relationships to the larger fields of political and armed conflict, especially war. The article deals first with the conceptual delimitation of these terms and then with the analytical issues involved in their relationships in different historical periods up to and including the present.

Introduction

Genocide is increasingly considered as a major phenomenon of modern world history, and has become the subject of the growing interdisciplinary field of genocide studies. This field generally studies 'genocides,' understood as discrete episodes of group destruction, generally in a comparative vein, so that the name of the field is often qualified as comparative genocide studies. However, from the first discussion of the topic in the closing stages of the World War II, genocide has been seen as having particularly close connections with war. Most major genocides have either taken place during wars (for example, the Armenian Genocide and the Holocaust during the World Wars I and II respectively, and the Rwandan Genocide during the civil war in that country), or as a consequence of wars (for example, the Cambodian Genocide at the conclusion of the wars in Indo-China). Almost all genocides involve the use of military power against civilians, and many are committed by militarized regimes or armed organizations of one kind or another. Indeed, genocide can even be seen as a particular form of war, distinguished by the fact that armed force is targeted not at armed opponents, but at civilian social groups. This apparently close connection raises problems, however, for the idea that genocide can be a distinct field of study, and equally for the idea that we can study war without problematizing genocide. This article explores the roots of this connection and the range of issues for the study of organized violence that arise from it.

The Origins of the Genocide Idea in War

The concept of genocide was first developed in the early 1940s by the lawyer, historian, and campaigner Raphael Lemkin as a general sociological and legal category to describe targeted destruction of social groups. The word first appeared in print in Lemkin's critique of *Axis Rule in Occupied Europe* (1944), in which he analyzed a 'Nazi genocide' against the many national groups of German-occupied Europe. Lemkin stressed the multimethod attack on populations (he did not confine genocide to physical violence but included economic, legal, cultural, and many other forms of coercion too), and the general character of national destruction across the occupied countries (he did not restrict genocide to the destruction of the Jews, even though he was well aware of the extreme murderousness directed at them in particular). Lemkin recognized that many of the acts of genocide were already covered by existing laws, but he argued for the need to define a general crime of group destruction.

The original publication of the idea in a book about Nazi occupation policies underlined the strong connections between Lemkin's conception of genocide and the context of war. The earliest uses of the idea – for example, when it was cited in the Nuremberg trials of Nazi war criminals in 1946 – emphasized this context. Nevertheless, Lemkin aimed to establish genocide as an international crime, whether it was committed in war or in peace, and this was reflected in the Convention on the Prevention and Punishment of the Crime of Genocide, adopted by the United Nations in 1948. Subsequently, as the Cold War developed, the idea was rarely invoked in legal settings (the UN did not establish the International Criminal Court, envisaged when the Convention was adopted, until the twenty-first century). However, it gained wide political and cultural currency, and was increasingly invoked whenever mass violations of human rights were committed.

Owing to the origins of the idea in the World Wars of the twentieth century, the type of war–genocide connection that was originally remarked was the linkage with total war, in which states targeted each others' populations as a means of defeating the enemy state. This type of war – degenerated because of the centrality of direct civilian targeting – culminated in the atomic bombing of Hiroshima and Nagasaki, and has sometimes been seen as genocidal in itself. However, it can be argued that the USA, like Britain in its bombing of Germany, did not see the opposing population as an enemy to be destroyed in themselves, and therefore ceased attacking them when their states surrendered. Nazi Germany, on the other hand, clearly regarded Slav as well as Jewish populations as enemies, the destruction of whose societies was a goal of the war, alongside the destruction of enemy states.

Thus in the Nazi case, it is difficult to distinguish the conventional wars that the state was fighting against the USSR and the Western Allies from their antipopulation policies. At the macro level of state policy, Germany's war, at least in Eastern Europe, was generally a racial war to enslave inferior peoples, hence one might say a genocidal war. At the meso level of strategic implementation, while the regime created specialized agencies like the paramilitary Schutzstaffel (SS) that were particularly charged with pursuing genocidal goals, the

Reichswehr (regular army) was also involved in the practice of genocide. During the invasions of Poland, the Soviet Union, Yugoslavia, and other countries, this went hand in hand with the regular campaign on the front, as well as being pursued after conquest in the occupied territories. At the micro level of implementation, German soldiers and civilian personnel together with local auxiliaries, including sometimes local civilians, had both 'war' and 'genocide' ideas in their heads as they dealt with local populations.

The understanding of these linkages has been blocked because the Nazi genocide has subsequently been misunderstood solely or mainly as a campaign against the Jews, and the latter has been misunderstood solely as a campaign of industrial extermination in death camps. However, until 1942, the Germans had been responsible for killing more Slav than Jewish civilians; and the campaigns against the Jews, which began with discrimination and pogroms in Germany itself before the war, had proceeded after the invasion of Poland through phases of expulsion and ghettoization, combined with violence against Slavs, before culminating in industrial gassing after the invasion of the USSR. Anti-Jewish policies were a distinct component of Nazi practice, but their implementation was combined with antipopulation policies against Slavs, Roma, and others, all of which conditioned each other in, and were conditioned by, the context of war.

It can be seen from this classic example that the connections of war and genocide can be complex: the same German actors were fighting a relatively conventional – albeit total and degenerate – war against the Western Allies and a genocidal war against the Soviet Union, while pursuing an array of specific genocidal policies against Jews, Slavs, and others. Where did war end and genocide begin? How far were the connections internal, because the war aims were themselves genocidal, or external and causal, in that war conditioned genocide and vice versa?

Genocide and Genocidal Violence

These issues only become more difficult when we consider a wider range of cases of both phenomena. The Nazi genocide, as Lemkin called it, was manifestly a large complex of policies and practices emanating from a large imperial state, centrally directed by a totalitarian regime. Although historians now emphasize the roles that local Nazi leaders, army commanders, allied states, auxiliaries, and even non-German civilians played in producing genocide, there is little doubt that the main drivers came from the top, indeed from a single leader, Adolf Hitler. This was mostly true of other European genocide in the era of total war, from the Ottoman killing of the Armenians to the Soviet 'terror-famine' and deportations of 'treacherous' nationalities and, after the World War II, the great mass deaths are produced by Chinese and Cambodian Communism.

However, not all genocide can be considered top-down in this way. Colonial genocide, during and after the European conquests of parts of the Americas, Australasia, Asia, and Africa, was often carried out by settlers, even against the opposition of the distant imperial state and its local representatives. Sometimes settlers' local authorities, and later the independent governments that colonists were able to be established, were the centers of genocidal policies against indigenous peoples. Often, however, it was local militias or simply ad hoc groups of settlers who expelled, killed, and raped them. These perpetrators were not fighting large-scale wars, but often short-lived local campaigns, or isolated skirmishes. They did not necessarily develop complex, formalized policies but improvised genocidal goals and practices as they saw fit in changing local circumstances. Often they did not carry out 'genocides' in the sense of complex episodes of large-scale dispossession and murder, but localized genocidal violence.

The literature uses the term 'genocidal massacre' (introduced by Kuper, 1981: 32) to describe more limited killing episodes, and by analogy we can talk of limited genocidal expulsions, genocidal rapes, etc. I use the general term 'genocidal violence' to refer to localized and limited acts of genocide, by whichever violent method. It is necessary to comment on the inclusion of expulsions and other acts of forced population removal – deportations, terrorized flight, etc. – within the framework of genocidal violence. These are sometimes considered 'ethnic cleansing' rather than genocide, the term being adapted from language long used in the Balkans and which entered international and academic discourse during the 1990s Yugoslav wars. 'Cleansing' terminology is manifestly euphemistic, and since forced removal of populations has long been a prime means of destroying particular societies and cultures, the phenomenon belongs to genocide. However, the terminology is highly politicized, with specific mention of forced removal omitted from the UN Convention because UN states carried it out, and 'ethnic cleansing' often preferred by statesmen because it does not imply the same legal obligations as 'genocide.' Here forced removal is considered a method of genocide and one form of genocidal violence.

Distinguishing and Relating Genocide to Other Types of Violence in War

This terminology is useful in analyzing not only colonial but also more recent genocide. Historically, highly centralized, statist genocide is the exception, and since the 1940s, genocide has appeared in a wide range of contexts but often with more limited scope and decentralized, less state-centric direction. During the Cold War period, genocide was a dimension of armed conflicts between Communists and anti-Communist forces, wars of decolonization, secessionist wars, and post-colonial civil wars. Since the end of the Cold War, it has continued to be a feature of secessionist and national power conflicts, and increasingly of electoral struggles, particularly in democratizing countries, often influenced by the UN and Western interventions and global media coverage. In short, genocide and genocidal violence are phenomena of a wide range of political and armed conflicts in the twenty-first century.

However, the complex tasks of distinguishing and analyzing the relationships of genocide and war have been affected by changes both in conflict itself and in the discourses surrounding it. The very use of the term 'conflict' reflects these changes. Whereas the twentieth century was dominated by large, long, often formally declared interstate wars, in the twenty-first (so far), wars are often relatively small, short, only

rarely interstate in a straightforward sense, and hardly ever declared. Although 'small wars' have always been more numerous than major interstate wars, there has been a marked trend of decline in the latter. There is a widespread consensus that war has changed its appearances: the majority of contemporary wars are civil wars, albeit usually internationalized. Moreover, while any use of organized armed (military) violence against a similarly armed enemy can be considered an act of war, and any conflict between organized armed opponents belongs to the realm of war, analysts have often been hesitant to apply the label 'war' to relatively small, short, localized conflicts, especially where the protagonists are mainly not recognized state actors. Hence there is a resort to 'armed conflict' as a description of the field of organized violence previously described as 'war,' and a frequent use of 'political violence' to recognize the overlap of war with other types of violent action and conflict. Thus problems in identifying contemporary trends reflect mutations in the phenomenology and perceptions of violence. According to Mary Kaldor (1999), we are in an era of 'new wars' combining civil and interstate dimensions, paramilitarism, identity politics, and ethnic targeting with parasitic, globalized war economies. Her account emphasizes 'ethnic cleansing' as a novel feature, although this begs questions about historical patterns of 'cleansing' and genocide in what she calls 'old' war. Yet, while the depth of this novelty is often disputed, the idea that the appearance of conflict has changed is widely shared.

The recognition of genocide has naturally been affected by these changing perceptions of war. Even the best recognized 'genocides' took place in contexts of general war and were often seen at the time as excesses of war. Their analytical recognition depended on simultaneously distinguishing genocide from and understanding its linkages to war, as well as overcoming denialist ideas representing genocidal violence as 'only' war. In today's wars, genocide's distinction from the wars and political conflicts of which it is a part may be less obvious. In considering these conflicts as a whole, moreover, it may be appropriate to use 'war,' 'armed conflict,' or 'political conflict' rather than 'genocide' as the dominant frame, and to conceptualize genocidal violence as a component of such conflicts.

The problem of recognizing genocidal elements in contemporary conflicts is exacerbated by the fact that genocide is often perceived as linked to the 'totalitarian/radical party' type of regime, the prevalence of which has declined together with twentieth-century types of extreme ideology. Actors' commitments to clear ideological frames, through which relatively coherent antigroup policies are formulated, have become less common; antigroup drives seem to reflect more contingent political interests. At the same time, the 'sacred evil' quality of the Holocaust has increasingly been extended to genocide in general. Together with the beginnings of international enforcement of anti-genocide law and policy, this has increased the stakes in using the term. The term 'genocide' has become an even more crucial ideological prize, claimed by representatives of victim groups to elevate their victim experience to the highest level, and used in the new activism. For the opposite reason, states, interstate organizations, and NGOs mostly avoid the terminology because of its political and legal baggage. Academic analysts of antipopulation violence, armed conflict, and war have often followed suit, inventing new terminologies to describe violence that could be considered within the genocide frame. I have already commented on the adoption of the euphemism 'ethnic cleansing'; another widely used euphemism is 'humanitarian crisis,' a term as applicable to natural disasters as to the results of targeted violence against civilians, which serves to create vagueness about the latter.

There have also been more focused attempts to conceptualize violence against civilians in twenty-first century conflict. The Uppsala Conflict Data Program (2009) proposes the concept of 'one-sided violence,' defined as "[t]he use of armed force by the government of a state or by a formally organized group against civilians which results in at least 25 deaths in a year." The low quantitative criterion (matching the Uppsala criterion for 'armed conflict') usefully recognizes that such force can be used on a much smaller scale than that recognized by the widely used older 1000-battle-death criterion for a 'war.' Ekaterina Stepanova (2009: 39–40) explains the idea: "[o]ne-sided violence is not armed conflict as such, as it directly and intentionally targets civilians who cannot defend themselves with arms. It is also distinct from battle-related violence that incidentally harms civilians, for example when civilians are caught in crossfire between combatants." In "contrast to indiscriminate violence, one-sided violence is a matter of conscious political and strategic choice and not a technologically inseparable part of modern warfare."

By implication, 'one-sided violence' is selective and political; but since it includes 'massacres, arbitrary killings, and terrorist attacks,' it covers violence with a range of objectives. Yet, this Uppsala idea echoes influential definitions of genocide: the widely cited sociological authorities, Frank Chalk and Kurt Jonahsson (1991: 23) actually use the term 'one-sided mass killing' in their definition. Like genocide, 'one-sided violence' captures the crucial distinction of combatant–noncombatant from combatant–combatant violence. Indeed, Roberta Cohen and Francis Deng (2009: 20) confirm that the Uppsala terminology was adopted as much for reasons of political context as of intellectual coherence, describing 'one-sided violence' as "a UCDP term that moves away from the highly emotive words 'terrorism' and 'genocide'." However, the genocide theorists who emphasize one-sidedness are also those who narrow the modalities of genocide to mass murder, and similarly Uppsala *defines* violence in terms of killings, even if it recognizes mass population displacement, for example, as increasingly "a central and direct goal of armed violence, rather than a side effect" (Stepanova, 2009: 42) and a 'devastating transformation' of peoples' lives (Walter Kälin, quoted by Cohen and Deng, 2009: 16).

Extermination and Control

More theoretically conceived work in social science also tends to write genocide out of armed conflict, often by simply not referring to it. Even where genocide is recognized conceptually, it tends to be marginalized analytically. For example, in Stathis Kalyvas' influential *The Logic of Violence in Civil War*, whose "focus is on violence committed intentionally against noncombatants" (Kalyvas, 2006: 5), genocide is mentioned only briefly. Kalyvas recognizes, as I do, that actual civil wars

comprise different types of violence, and in distinguishing between different types, he argues that violence "can be used to *exterminate* a group or to *control* it" (2006: 29, emphasis added), and focuses on the latter type, 'also known as coercive violence.' In particular, Kalyvas concentrates on a particular type of violence, 'civil war violence,' which he explains as a strategically rational form based on the need of parties to civil wars to control populations. Although he too restricts his attention to homicidal violence, he claims that his theory also does well for "violence that stops short of homicide (e.g., arrest, torture, displacement)" (Kalyvas, 2006: 7), and recognizes that "[a]nalytically akin to physical destruction is mass deportation, sometimes referred to as 'ethnic cleansing'" (Kalyvas, 2006: 26). This analytical kinship lies, presumably, in the fact that forced population removal (of which direct deportation is only one method) is also a means of 'exterminating' a group, in the sense of destroying its physical, social, and cultural presence within its historic territory.

More formally, Kalyvas (2006: 30, Table 1.1) also presents a typology of mass political violence, differentiated by whether or not a political actor 'intends to govern the population targeted.' He sees 'state terror' and 'civil war violence' as forms of violence reflecting the intention to govern, and 'genocide and mass deportation' and 'reciprocal extermination' as forms that reflect the lack of this intention. Genocide is also defined as 'unilateral': this captures the defining one-sidedness of the relationship between perpetrator center and target group, but seems an insufficient characterization since genocide may be practiced by both sides in a conflict. (Presumably, this is what Kalyvas indicates by 'reciprocal extermination,' but this indicates, in my view, a different type of conflict rather than of violence.)

The two ways of distinguishing violence that Kalyvas uses, 'control/extermination' and 'intention to govern/not' seem similar (he apparently uses them interchangeably) but they should be separated. The intention to govern or not is a more precise formulation of the criterion for genocide: the point when a collective actor enters into the realm of removing, exterminating, or destroying a given population under its control, it abandons any intention of governing it. Control, on the other hand, is a more overriding concern that may be a rationale for either governance or extermination, and is a feature of both governing and exterminating processes. Thus 'intention to govern' helps *define* the difference between nondestructive coercion and genocide; the logic of 'control' helps *explain* the choice between the two.

Kalyvas regards 'civil war violence' as a specific type of coercive rather than exterminatory violence, explicable by the logic of control. What he fails to explain is why violence goes beyond coercion toward extermination: the escalation of violence from one type to another. As Laia Balcells (2010) suggests, discussing Franco's killing of Spanish Republicans after the end of the civil war, exterminatory violence may be explained by the same logic of control that is seen in war violence. Yet Kalyvas, while distinguishing 'civil war violence' as a particular type from civil war as such, tends to assimilate most actual violence in civil wars to this type, so occluding the question of exterminatory or genocidal violence. Kalyvas and Nicholas Sambanis' analysis of Bosnia (2007) examines Serbian 'ethnic cleansing' without reference to its exterminatory meaning for Bosniak, Croat, and plural communities, apparently because they explain it as a function of population control in conquered territory. But this is to confuse the question of 'government/not' with the question of dynamics of control.

The overriding reason why we cannot satisfactorily distinguish 'control' from 'extermination' is that *control is an essential aspect of extermination*. Questions of control are apparent at every stage in the movement to and practice of genocidal violence. Movement toward deportation may be, as Kalyvas and Sambanis suggest it was in Bosnia, a function of the failure of more limited coercion to control a population. But deportation does not necessarily end dilemmas of control: a population needs to be controlled during deportation, and a deported population usually still needs to be controlled where it is deported to. If the deport-to location is within the direct control of the deporters, deportation simply gives them new dilemmas of direct control, as for example, for the Nazis with Jews in Polish ghettos and the Stalinist regime with the remnants of deported minorities in their regions of exile. If the deport-to location is outside the deporters' control, different issues of control may still arise: Israel managed to remove most Palestinian Arabs from its territory in 1948, but has since spent over six decades attempting to control the responses of the resulting refugee populations. Moreover, since deportation is rarely total, removal of the majority of a population will usually be combined with issues of the control of the remaining minority, as Israel has also found.

Fully murderous policies certainly involve additional issues, but still include questions of control. To starve or kill on any scale requires prior control over the target population and control during destruction processes. Moreover, it is important to *represent* murderous policies as policies of control. For example, deportation may be presented as relocation for work, even where it is taking people to death, since target populations and third parties acquiesce more willingly in policies that they perceive as for control rather than for destruction. Mass murder, moreover, is usually combined with less destructive measures of control, because the killing of some is often designed as a means of controlling larger populations. For example, Feierstein (2010) argues that the Argentinian dictatorship's 'disappearance' of leftists in the 1980s was a means of reordering the whole national society.

Moreover, as Kalyvas (2006: 26) notes, "the methods used to achieve compliance and physical destruction may be similar," even if the objectives differ. To operationalize the distinction (govern or destroy) we therefore need to infer perpetrators' objectives from the context in which violence occurs. In the case of victims who have surrendered – whether originally combatants or not – it should be relatively simple to determine genocide. The same should be true of civilians murdered or expelled *after* armed conflict has concluded, and of course *outside* the context of armed conflict. In the case of those killed or expelled *during* civil war, however, the distinction will be less straightforward, requiring detailed knowledge of the relationships between the targets of violence and the armed conflict. Like the distinction between victims of 'civil war violence' and of 'collateral damage' from intercombatant violence, it is easier to outline in principle than to apply conclusively in practice. Actors may *combine* these types of

objectives: by having partially destructive and partially governing objectives toward the same population; by using destructive methods against one section of a population in order to reinforce government of a larger population; and of course by having governing objectives at one point in time and destructive objectives at another. Mann (2005: 6–7) notes that escalation toward more destructive ends is a normal feature of genocide: but *de-escalation* is also normal, in the sense that governance may be reasserted once a certain level of destruction has been accomplished. And of course, among the 'members' of a collective actor which has a governing policy, elements or individuals may pursue destructive policies – or vice versa.

All these elements of indeterminacy, arising from common issues of control in both governance and destruction, may complicate the ascription of *either* destructive *or* governing objectives to a given actor in a given situation. When control is to be achieved through selective violence, any perception of a distinct section of the population as an 'enemy' in a fundamental sense (whether 'ethnic,' 'political,' or other) will offer a rationale to extend civil war violence to genocidal violence. These questions arise, moreover, on the interfaces between all three levels of civil war that Kalyvas (2006) specifies: macro, meso, and micro. Introducing the genocide question means reproblematizing the macro level, because genocidal drives often crystallize mainly at that level. However, it does not necessarily mean a macro-down approach. Research on the Holocaust has emphasized the interconnections between levels in both directions, for example, how Nazi officials managing conquered populations in different jurisdictions 'worked to' Hitler and the central leadership, evolving 'solutions' to racial 'problems' which influenced the regime's overall policies. Gerlach's (2010) work on 'extremely violent societies' shows how central militaries and local movements have interacted to produce mass killing, for example, in Indonesia in 1965 and in Bangladesh in 1971. Toal and Dahlmann (2011) show how local political elites, controlling municipalities, were central to mass deportation and the consolidation of its results in Bosnia. Historical work, for example, on 'genocidal moments' in settler colonization in Australia and on the local genocide of Jews by Polish civilians in Jedwabne, also suggests that genocide may arise at the micro level.

Conclusions

Although the connections between genocide and war are as old as the idea of genocide, the issues raised have been shown to be particularly relevant in the twenty-first century context, which is widely recognized to involve simultaneous shifts in patterns of armed conflict and of antipopulation violence. The relationship is a core analytical challenge for social scientists in making sense of these phenomena, closely linked to the shared core insight that actual 'conflicts' or 'wars' *combine* what should be understood as different *types* of violence. Theoretically, distinct types of violence are not to be equated with different *methods*. Killing, for example, even of civilians, can be a function of war or of genocide, and indeed of different subtypes of violence within 'war,' as suggested by the distinction of 'civil war violence' from 'collateral damage,' as well as the difference between genocidal and military killing. The typology of violence reflects, rather, understandings of the different purposes of violence, which is why the typology has a potential explanatory significance.

It is important to emphasize, too, that analytical clarity on these issues enables social science to contribute to clearer public debate on anti-civilian violence in armed conflict. A clear distinction of genocide from war, combined with an understanding of its relationships to war, can help to cut through the self-serving obfuscations that are practiced on all sides in violent conflicts. While would-be spokespeople for victims will invariably claim all violence against particular groups of civilians as 'genocide,' spokespeople for perpetrators and often for third-party states and international organizations will typically represent genocidal violence as 'only' war, thus using the idea of war to deny genocide. From both sides of such arguments, the idea is presented that 'war' and 'genocide' are exclusive terms. Social science can show their complex interdependence and hybridity and thus contribute to more careful discussion of the related public issues.

See also: Civil Wars; Ethnic Cleansing, History of; Ethnic Conflicts; Holocaust, The; Nation-State and War; Violence, History of; War and Nationalism; Warfare in History.

Bibliography

Balcells, L., 2010. Rivalry and revenge: violence against civilians in conventional civil wars. International Studies Quarterly 54 (2), 291–313.
Cohen, R., Deng, F., 2009. Mass Displacement Caused by Conflicts and One-Sided Violence: National and International Responses. Oxford University Press, Oxford. SIPRI Yearbook 2009, pp. 15–38.
Feierstein, D., 2010. Political violence in Argentina and its genocidal characteristics. In: Esparza, M. (Ed.), The Cold War Years State Violence and Genocide in Latin America. Routledge, London, pp. 43–63.
Chalk, F., Jonahsson, K., 1991. The History and Sociology of Genocide: Analyses and Case Studies. Yale University Press, New Haven.
Gerlach, C., 2010. Extremely Violent Societies: Mass Violence in the Twentieth Century World. Cambridge University Press, Cambridge.
Kaldor, M., 1999. New and Old Wars. Polity, Cambridge.
Kalyvas, S., 2006. The Logic of Civil War. Cambridge University Press, Cambridge.
Kalyvas, S., Sambanis, N., 2007. Bosnia's civil war: origins and violence dynamics. In: Collier, P., Sambanis, N. (Eds.), Understanding Civil War: Evidence and Analysis, vol. 2. The World Bank, Washington, DC, pp. 191–229.
Kuper, L., 1981. Genocide. Penguin, Harmondsworth.
Lemkin, R., 1944. Axis Rule in Occupied Europe. Carnegie, New York.
Mann, M., 2005. The Dark Side of Democracy. Cambridge University Press, Cambridge.
Shaw, M., 2003. War and Genocide: Organized Killing in Modern Society. Polity, Cambridge.
Shaw, M., 2007a. What Is Genocide? Polity, Cambridge.
Shaw, M., 2007b. The general hybridity of war and genocide. Journal of Genocide Research 9 (3), 461–473.
Stepanova, E., 2009. Trends in Armed Conflicts: One-Sided Violence against Civilians. Oxford University Press, Oxford. SIPRI Yearbook 2009, pp. 39–68.
Toal, G., Dahlmann, C., 2011. Bosnia Remade: Ethnic Cleansing and Its Reversal. Oxford University Press, Oxford.
Uppsala Conflict Data Program (UCDP), 2009. UCDP Definitions. http://www.ucdp.uu.se/ (accessed 15.05.11.).

Genocide, Historical Aspects of

Gunnar Heinsohn, Raphael Lemkin Institute for Comparative Genocide Research, University of Bremen, Bremen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mass atrocities – that were not defined as genocide before 1948 or as crimes of war before 1899 – are followed through the ages from tribal communities to modern nation-states with a special focus on ecological, demographical, and ideological factors behind major annihilations.

Initiatives to criminalize genocide are exemplified from first steps in ancient Hebrew law via similar approaches after the atrocities in sixteenth-century European colonies to the international conventions passed after the Holocaust, and the most recent efforts like the *Responsibility to Protect* (*RtoP*) initiative of 2005. The probability for contemporary and future perpetrations is illuminated by the youth bulge concept that helps to develop an index for mega-killing potential.

The Term Genocide

On 9 December 1948, the United Nations *Convention on the Prevention and Punishment of the Crime of Genocide* was adopted unanimously (by a vote of 55:0:0). By 2012, 140 nations had ratified or acceded to the treaty. The text of the Convention was mostly written by a Polish lawyer and Holocaust refugee, Raphael Lemkin (1900–59; Szawłowski, 2005) who, in 1933, had already proposed a measure against the 'crime of barbarity' to the *Legal Council of the League of Nations* in Madrid based on the Ottoman extermination of Armenians after 1915, and the Simele massacre of 1933 in which the Army of Iraq had slaughtered Christian Assyrians. In his 1944 study *Axis Rule in Occupied Europe*, Lemkin coined the term 'genocide' (from Greek *genos* = tribe, and Latin *caedere* = kill). In the Convention's article II

> genocide means any of the following acts committed with intent to destroy, in whole or in part, a national, ethnical, racial or religious group, as such: a. Killing members of the group; b. Causing serious bodily or mental harm to members of the group; c. deliberately inflicting on the group conditions of life calculated to bring about its physical destruction in whole or in part; d. Imposing measures intended to prevent births within the group; e. Forcibly transferring children of the group to another group.

The crimes mentioned in paragraphs a–e are generalizations of practices employed by Hitler's Germany after 1 September 1939 (beginning of the World War II) mostly against Jews and Slavs. Because 'politicide' – the murder of any person or people because of their politics or for political purposes – as well as 'ideological genocide' – the murder of any person or people because of their ideology, theory, or knowledge – were excluded from the UN Genocide Convention, scholars later broadened the concept. Rudolph Rummel's 'democide' may be considered as the most comprehensive term. It is the murder of any person or people by a political movement or a government, including genocide, politicide, and mass murder (Rummel, 1994: p. 39). Genocide and democide are used side by side with the term 'crimes against humanity' coined on 24 May 1915 by France, Great Britain, Russia, and the United states on behalf of the Ottoman Armenians as "new crimes of Turkey against humanity and civilization." The term 'crimes against mankind,' created on 12 October 1942, in London ('Declaration of St. James') to denounce the Holocaust, is also still in use. Less known is the French term "populicide" coined by Gracchus Babeuf (1760–97) as a shorthand for his 'système de dépopulation,' which was carried out – during the French Revolution – in the Vendée (1793/94) by killing its entire royalist peasantry of some 120 000 (Babeuf, 1795). A more recent formula used by the *International Criminal Court* defines genocide as "the deliberate and systematic destruction, in whole or in part, of an ethnic, racial, religious, or national group" (Funk, 2010: p. 9).

The Study of Genocide

History of the Discipline

The discipline of comparative genocide research was established by Lemkin, who by 1959 had drafted three volumes on the *History of Genocide*. Several publishing houses turned him down. The manuscripts are kept (under "*Series III: History of Genocide*") by the *American Jewish Historical Society: Center for Jewish History* in New York (Elder, 2012).

The 1967–70 massacres of Ibos, who were trying to secede from Nigeria, with an estimated death toll of 500 000, triggered a worldwide interest in genocide (Gould, 2012). In 1970, Yvan van Garsse published *A Bibliography of Genocide Crimes against Humanity and War Crimes*. The genocide, in 1971, of up to 1.5 million citizens of secessionist East Pakistan (today's Bangladesh) by West Pakistan's army and militias, also drew attention to the subject (Mascarenhas, 1986; Linton, 2010). Both mega-killings had worldwide television coverage. However, it was not understood that both Nigeria (20–60 million inhabitants from 1940 to 1970) and Pakistan (60–135 million) were going through massive 'youth bulges.' Youth bulges are defined by at least 30% of the male population in the age bracket 15–29 years, i.e., there are three or more sons per father desperate for positions in society.

The Twentieth Century Book of the Dead by Gil Elliot (1972) became the first major monograph devoted to comparative genocide research. In 1976, sociology entered the field with Vahakn Dadrian's "A Typology of Genocide." Irving Louis Horowitz's *Genocide: State Power and Mass Murder* followed a year later. In 1981, Leo Kuper found a major academic publisher, Yale University Press, with *Genocide: Its Political Use in the Twentieth Century*. More systematically and historically minded works followed in 1990 with Helen Fein's *Genocide: A Sociological Perspective* as well as Frank Chalk's and Kurt Jonassohn's *The History and Sociology of Genocide*.

The first – one person – institute devoted to this subject – *Institute on the Holocaust and Genocide* – was founded in 1979 by Israel W. Charny in Jerusalem. Helen Fein followed, 1982 in New York, with the *Institute for the Study of Genocide*. Next were Frank Chalk and Kurt Jonassohn with the Canadian *Institute for Genocide and Human Rights Studies* (Montreal, Concordia University). Late starters were Australia with the *Centre for Comparative Genocide Studies* (Macquarie University) initiated in 1993 by Colin Tatz as well as Europe with the *Raphael-Lemkin-Institut für Xenophobie- und Genozidforschung* (University of Bremen/Germany) founded in the same year by Gunnar Heinsohn, Klaus von Muenchhausen, and Léon Poliakov. An *International Association of Genocide Scholars* (http://genocidescholars.wsg.net) was established in 1994. In 2005 followed the *International Network of Genocide Scholars* with more than 300 members in 2012 (http://www.inogs.com).

The first encyclopedia of genocide was published in Germany (Heinsohn, 1998) after Rummel's seminal *Statistics of Democide* (1997). A similar endeavor in English followed 2 years later (Charny, 2000). Meanwhile, there are quite a few general works available (Valentino, 2004; Levene, 2005; Kiernan, 2007; Moses, 2008; Goldhagen, 2009; Jones, 2010; Pinker, 2011: Chapters 1–7; Staub, 2011). *Necrometrics* by Matthew White keeps death tolls up to date (http://necrometrics.com/index.htm). Leading journals are *Holocaust and Genocide Studies* (since 1986) and the *Journal of Genocide Research* (since 1999).

Major Subjects of Genocide Research

Historical, sociological, and sociopsychological studies of genocide aim at the assessment of genocidal potential as well as the formulation of means to prevent the crime. Interventions against ongoing genocides are rather dealt within studies of warfare than in comparative genocide research. Eight predictive economic, demographic, ideological, ethnic, and ethical variables indicating genocidal potential have been identified: (1) a history of minorities outside the moral obligation of the majority (e.g., Jews, Armenians, Chinese); (2) a relevant percentage of young men with no hope for positions to satisfy their ambitions (youth bulge); (3) a preponderance of Marxist–Leninist, clericalist, nationalist/tribalist, racist, and/or other antihumanitarian ideologies; (4) an accepted political role for the military; (5) at least one totalitarian political party in power or well positioned to conquer it; (6) the totalitarian party has territorial claims against neighbors; (7) a civil war is waged against the internal resistance and/or a war against neighbors to realize these aims; (8) the legitimate secrecy in times of war is used as a cover for the killings with those minorities most vulnerable which have no retaliatory power of their own, no foreign allies or are hardly known to the rest of the world (for a pioneering work see Porter, 1982; cf also Stanton, 1998).

The prevention of these crimes requires an early warning system. Of the many failed attempts to establish 'Genocide Watch' (term created by this author) *The International Alliance to End Genocide* by Gregory Stanton has been the most enduring (http://www.genocidewatch.org/home.html). Israel Charny's *Genocide Prevention Now* (GPN; http://www.genocidepreventionnow.org) is online since 2010. On 23 April 2012, US President Obama declared, at the *United States Holocaust Memorial Museum*, that preventing mass atrocities was a 'core national security interest.' The declaration is an offshoot of the United Nations' RtoP initiative of 2005. That norm (not a law) focuses on preventing and halting four *Mass Atrocity Crimes* – genocide, war crimes, crimes against humanity, and ethnic cleansing (Evans, 2008).

Historically, only a minority of high-risk nations have transformed their perpetrator potential into actual genocide (the rate is higher for democide). Since a genocide needs preparation, all nations with perpetrator potential have to be scanned for red alert signals. At least six – difficult to conceal – indicators for the imminence of a genocide have been identified: (1) registration and/or marking of prospective victims; (2) staging a propaganda campaign which paints the prospective victims (most frequently the Jews of Israel) as a mortal danger to the prospective perpetrators; (3) ousting of police or military leaders who refuse to join in genocide operations; (4) recruiting and training of special killing units (Cete (Young Turks), Tsheka (USSR), SS (Schutzstaffel; Hitler Germany), Chetniks (Serbia), Interhamwe (Rwanda), etc.); (5) employing new euphemisms like 'ethnographic restoration' or 'final solution' (Hitler on Poles and Jews, respectively), 'ethnic cleansing' (Serbs on fellow Yugoslavians), 'civil war' (Hutus on Tutsis), etc.; (6) sudden appearance of – nonfamine or disaster-induced – refugees belonging to endangered minorities.

To predict genocidal potential within movements not yet capable of reaching national positions of power, a special grid of 14 or more designation points was developed to scan them (Littell, 1969: Chapters 2 and 3). A red alert pattern is reached when 8–10 actions are detected: (1) The cadre prints, distributes, and uses prejudicial material for attracting sympathizers and recruiting members. (2) It makes prejudicial appeals through the media or in mass meetings against any ethnic, religious, or cultural community. (3) Members cultivate the politics of verbal assault and physical violence-publishing slanderous charges, bombing meeting places and homes and media, beating and assassinating opponents. (4) The cadre pursues the politics of polarization, scorning the middle ground of compromise or consensus, rejecting the politics of moderation and orderly change. If it becomes numerous enough, instead of functioning as a parliamentary opposition it builds the structures of 'a state within a state.' (5) It uses the *Language of Assault* toward political opponents, rather than using the *Language of Dialogue* and participating in good faith in the political forum. (6) It deliberately drives a wedge between the generations, creating youth groups hostile to their parents' generation and to their heritage. (7) The cadre maintains camps for paramilitary training, including the use of antipersonnel weapons, outside the control of public officials (either police or

military). (8) It trains and maintains private armies, demonstrating in public in uniform, parading and marching to intimidate loyal citizens. (9) Leaders of the movement elaborate a quasi-religious structure of authority and sanctions, with political hymns, shrines, martyrs, and liturgies. (10) Archaic tribal, clannish, or religious symbols are worn by members as recognition signs, etc. (11) Induction and termination of membership are observed as quasi-religious rites. Straying members are treated as 'heretics.' (12) The movement's basic unit is the closed cell, with three to seven the standard number. This is the classic unit of a revolutionary party or intelligence operation, but it lies outside the style appropriate for loyal citizens to use in exercising influence upon a legitimate government. (13) The movement practices deception and confuses public opinion by launching one-issue 'fronts' – without clear identification of sponsorship, financing, and control. A great deal of money is raised this way, ostensibly for some good public purposes – but actually to serve the party's internal interests in the pursuit of power. These devices are known as 'the large net' within which 'innocents' are caught. (14) The cadre's studied tactics include infiltration and subversion of public institutions and voluntary associations, turning their direction from public service to bolstering the movement's drive for power.

Less developed concepts are scenarios for postgenocidal reconciliation. Even the systematization of Germany's pertinent efforts – so far second to none – is still in its infancy (Heinsohn, 1997). Low-key approaches like Truth Commissions are much more advanced. They were first tried, in 1974, in Uganda. By 2012, at least 23 nations had experience with such commissions (Hayner, 2001; Martin, 2009).

History of Genocide

Tribal (Primitive) Society

Genocide and acts of war were hardly distinguishable in tribal societies – covering the longest stretch of human history – because there was no separation of noncombatant civilians and warriors. Warfare was "'ecological' in motivation" and led to the redistribution of hunting and fishing grounds or arable land 'from the weak to the strong' (Keegan, 1993: p. 101). These wars were characterized by 'battles of annihilation' or, if space allowed, 'displacement of the weaker' (Keegan, 1993: pp. 29, 387). However, due to a lack of reliable data up to second millennium CE or later we simply cannot know to what degree tribal societies engaged in such atrocities. Much of what Europeans have reported as inhuman cruelty may just have been the desperate reaction against merciless invaders. Still, we know that genocidal tribal raids continued well into the second half of the twentieth century in territories of the Amazon, New Guinea, the Philippines, and elsewhere: "Typically the goal in such raids is to destroy an entire village or settlement, the common intention being to kill everyone, perhaps excepting a few of the younger women and children who are carried off for incorporation into the home group" (Glick, 1994: p. 45).

In resisting conquering European powers entire tribes – not only their adult males – went to war and perished together. A typical example is provided by the Hereros. By employing the most efficient technique of genocide – artificial famine – imperial German troops under General von Trotha, in 1904, drove them into the Omaheke desert (South West Africa), blocked the escape routes and let them die of thirst (Zimmerer et al., 2008). The annihilation, e.g., of the Tehuelche and Puelche by Argentina (1878–85; Hasbrouck, 1935; Maybury-Lewis, 2002: p. 45) or the Ndebele (Zimbabwe) by Great Britain in 1896/97 (Dodds, 1998) was executed in a similar manner.

Antiquity/Feudalism (Bronze Age)

The first stage of civilization (Early and Middle Bronze Age with priest-kingship feudalism) saw tribal genocidal practices on a larger scale. Tiglath-Pileser I of Assyria (conventionally dated from 1116 BC to 1077 BC) may exemplify this practice for Mesopotamia: "I burned their cities with fire, I devastated, I destroyed. / I captured, I killed multitudes of them. […] Their orchards I cut down" (Luckenbill, 1989: pp. 75/99). "I caught the survivors and impaled them on stakes in front of their towns" (Pritchard, 1969: p. 276). After unconditional surrender, pledges of loyalty and payments of tribute the victor may have granted mercy but there was no international law binding him to it as the same ruler made clear: "I destroyed them. The rest of their host embraced my feet, and I had mercy upon them. Tribute and tax I laid upon them, and counted them among the subjects of Assur" (Luckenbill, 1989: p. 80).

Further west, Egypt behaved in the same way: "I slew them without sparing them, they sprawled before my horses, and lay slain in heaps in their blood," Ramses II (conventionally dated from 1279 BC to 1212 BC) boasted after a campaign against the Hittites whose leader begged him for mercy: "Look, you spent yesterday killing a hundred thousand and today you came back and left no heirs. Be not hard in your dealings, victorious king" (Lichtheim, 1976: pp. 70f.).

In the West, the Greek expedition against Troy (conventionally dated to the twelfth-century BC) matched the campaigns of Asian and African rulers: "My good Menelaos' (king of Sparta), said he (Agamemnon, king of Mycene), 'this is no time for giving quarter. Has, then, your house fared so well at the hands of the Trojans? Let us not spare a single one of them – not even the child unborn in its mother's womb; let not a man of them be left alive, but let all in Ilius (Troy) perish, unheeded and forgotten'. Thus he did speak and his brother was persuaded by him, for his words were just" (Homer, 1952, *Iliad* VI: pp. 55–62).

The prototype of Western man, Ulysses, did not fail to live up to the same standard of justice: "When I had set sail thence the wind took me first to Ismarus [Thracia], which is the city of the Cicons. There I sacked the town and put the people to the sword. We took their wives and also much booty which we divided equitable amongst us so that none might have reason to complain" (Homer, 1952, *Odyssey* IX: pp. 39–42). Three millennia later, in Srebrenica – not that far from Ismarus – Bosnian Serbs dealt, on 11 July 1995, with Bosnian Muslims very much in Ulysses' manner. They separated 7000 males from their families and shot them. Before and after they raped young women left alive for that very purpose (Rohde, 1997; Nuhanovic, 2007).

The still pagan Early Israelites were hardly different from their neighbors: "Now go (Saul, 1st king of Israel; from 1022

BC to 1002 BC) and smite Amalek, and utterly destroy all that they have, and spare them not; but slay both man and woman, infant and suckling, ox and sheep, camel and ass" (*1 Samuel* 15: pp. 2–3). However, with monotheistic Judaism a set of moral rules was set down that has inspired antigenocidal legislation from its very beginning in the sixteenth century (see below).

Antiquity/City States (Iron Age)

The genocidal practice of time immemorial was cast into a principle by Athenians. When, in the Peloponnesian War (431–404 BC), Athens decided to wipe out Sparta's ally Melos, the admirals in charge tried to formulate an eternal law underlying their action: "It is a general and necessary law of nature to rule wherever one can. / We are merely acting in accordance with it, and we know that you or anybody else with the same power as ours would be acting in precisely the same way." After a few attempts at defending themselves "the Melians [in 415 BC] surrendered unconditionally to the Athenians, who put to death all the men of military age whom they took, and sold the women and children as slaves. Melos itself they took over for themselves, sending out later a colony of 500 men" (Thucydides, 1954, *The Peleponnesian War*, V: pp. 85ff.).

As a first step away from mutually slaughtering their males, the Greeks developed the practice of exchanging prisoners of war. Within the Roman Empire the practice of destroying enemies who had caused problems in the past was continued. The annihilation of Corinth (metropolis of mainland Greeks) and Carthage (metropolis of Phoenicians) in 146 BC, of Numantia (capital of Spanish Celtiberians) in 134 BC, or the elimination of the Etruscans (92 BC) and the Samnites (82 BC) as major opponents in Italy proper provide typical examples with 50 000–150 000 dead in each case. Of course, the opponents acted in the same manner as, e.g., in the slaughter of the inhabitants of the Greek cities Selinus and Himera (Sicily) in 409 BC by the Phoenicians of Carthage. However, within the Roman Empire a 'ius gentium' was developed for strangers who did not hold citizenship with – sometimes – more protection for these minorities than for full citizens.

The concept of 'politicide' – the elimination of local rivals for positions of power – was also already practiced in the Greek City States. Periander, despotic ruler of Corinth, had asked Thrasybulus, in the same position at Miletus, "about the safest political establishment for administering the city best. Thrasybulus led out Periander's messenger, outside the city, and with him entered a sown field […] And ever and again as he saw one of the ears of corn growing above the rest he would strike it down, and what he struck down he threw away, until by this means he had destroyed all the fairest and strongest of the corn. […] Periander understood the act of Thrasybulus and grasped in his mind that what he was telling him was that he should murder the most eminent of the citizens" (Herodotus, 1987, *The History*, V: 92).

The best known example of politicide in modern times was the annihilation, after eight inconclusive civil wars, of the Protestant leadership of France (*c.*100 000 victims), starting on Saint Bartholomew's Day (24 August 1572) when the Huguenots – easily recognizable by their black clothing – were guests at the royal marriage of Henri of Navarra and Marguerite of Valois (Garrisson, 2000). The politicide, in February 1570, against Novgorod, Russia's richest city, by Ivan IV (the Terrible) was peculiar insofar as the tsar had put together special killing units – the 'Oprichniki' – because the army was not trusted to act ruthlessly enough (Skrynnikov and Graham, 1981: pp. 213ff.). The suppression, on June 30 of 1934, of the so-called Roehm revolt in which Hitler's Gestapo assassinated more than 90 left wing Nazi leaders (Gritschneder, 1993). Above are well-researched politicides of the twentieth century.

Politicide does not necessarily include the killing of the families of the victims. However, in a so-called 'Vesper' – a politicide that is turned against a foreign elite – can easily grow into a full-blown genocide. It took its name from the Easter (=Vesper) festival in Palermo on 29 March 1282, when the French in control of the Kingdom of Sicily were killed with their families ('Sicilian Vespers'; Runciman, 1958). It had been preceded, in 1268, by the butchering of some 5000 members of the German Hohenstaufen elite of Sicily by the hands of the French conquerors. The earliest well recorded Vesper took place in 88 BC when all the Romans and Italians – including women and children – living in Greek Asia Minor were put to the sword by orders of Mithradates VI (Mayor, 2010). The Romans and Italians living in Britannia met a similar fate in 61 BC. In the 'Ulster Massacre' at Portadown (23 October 1641), the last Vesper of Western Europe, Irish rebels killed some 100 protestant British settlers and administrators (MacCuarta, 1993). The Creoles of Haiti delivered their French masters a Vesper in 1791/1805 (Dubois, 2005) as did, in 1857/58, albeit with no lasting success, the Sepoy to British masters of India (Anderson, 2007).

Parallel to genocidal practices of Greek City States, monotheistic Judaism – evolving after the Babylonian exile (conventionally ending in 539 BC) – began to formulate a code of ethics that – very slowly – came to influence international ethics. To overcome expectations of salvation tied to the ritual slaughter of humans and animals, the prophets found a new moral vision: "I desired mercy, not sacrifice; and the knowledge of God more than burnt offerings" (*Hosea* 6:6). The sanctity of life became the highest principle: "Thou shalt not kill" (*Exodus* 20:13; *Deuteronomy* 5:17) including – in its universal character – the strict outlawing of infanticide. The Torah emphasizes the sanctity of life as the core of the law and as a value identical with goodness ('the good law'):

> See, I have set before thee this day life and good, and death and evil. / Therefore choose life, that both thou and thy seed may live. / The stranger who dwelleth with you shall be unto you as one born among you, and thou shalt love him as thyself. / Keep thee far from a false matter, and the innocent and righteous slay thou not; for I will not justify the wicked. / If thou forbear to deliver them that are drawn unto death, and those that are ready to be slain. If thou sayest, "Behold, we knew it not" - doth not He that pondereth the heart consider it? And He that keepeth thy soul, doth He not know it?
>
> *Deuteronomy* 30: 15/19; / Leviticus 19: 34; / *Exodus* 23: 7; / *Proverbs* 24: 10–12.

Against monotheistic Judaism, the first recorded order (of Haman) for an ideological – though abortive – genocide (also called religiocide) was issued in the Ackhaemenid Empire of the fourth-century BC: "There is a certain people scattered abroad / and their laws are diverse of all people. […] And the

letters were sent by posts into all the king's provinces, to destroy, to kill, and to cause to perish all Jews, both young and old, little children and women, in 1 day" (*Esther* 3: 8/13).

Christians who had adopted the Jewish rejection of blood sacrifice to planetary deities actually suffered the first documented religiocide. The edict of Emperor Decius of CE 250 forced them to sacrifice to the pagan gods and eat sacrificial meat or suffer death (Bleckmann, 2006).

Middle Ages

The Middle Ages saw attempts at a 'Pax Romana' monitored by the Papal See. Its concept of a 'Pax Dei' (peace of God) – developed since CE 975 and extended over all Christian lands – outlawed war from Wednesday night to Monday morning as well as on Christian holidays. Excommunications and interdicts were the only means of enforcement. It was the beginning of international law based on the Judeo-Christian code of ethics. Yet, frequent transgressions were severe and – in modern terminology – genocidal.

The crusades – starting in CE 1096 – not only brought massacres of thousands of Jews in Germany and ten thousands of Jews and Muslims in Jerusalem (CE 1099), but also of fellow Christians. Against Pope Innocent III's interdict, Dalmatia's port Zadar – arch rival of Venice – was bathed in the blood of all its inhabitants in November 1202 during the 4th Crusade by French and Italian troops transported on Venetian ships. The same men – for a starter – butchered 2000 inhabitants of Greek Constantinople on 13 April 1205. Before the Latin army set mankind's finest city on fire it went on a three-day rampage of indiscriminate slaughter, not sparing women and children (Philipps, 2004).

Since 1167, the Cathars (from Greek katharoi = pure), called Albigensians after the French city of Albi, suffered a series of religiocides by Catholic forces because they rejected any spilling of blood and did not believe in the resurrection of Christ or the Holy Communion as the residue of blood sacrifice within Christianity (Pegg, 2008). By 1234 the Cathars were wiped out. In 1208, Innocent III who had tried to protect Zara and Constantinople, ordered the massacre of Béziers, a center of the Cathars, where most of the 15 000 inhabitants were slaughtered or burnt in their churches (Berlioz, 1996).

A few years later, in CE 1211, Gengis Khan's Mongols – with no authority to arouse their conscience by threats of excommunication or invoking the image of all men as protected children of the one God – began their terrifying campaigns from Japan to the Baltic Sea. To turn China's fertile areas into horse pastures the Mongols have, up to CE 1290, eliminated some 25 million Chinese. Another 5 million people perished in Persian, Arab, and Christian territories where the Steppe warriors revived the memory of Attila's fifth-century Huns. Behind their lines, the Mongols used to kill everyone perceived as a future threat, i.e., all males not suitable to join their ranks. Defeated armies were allowed a brief lease on life by serving as shields or battering rams for advancing Mongol troops. Tamerlane, a pious Muslim, let, in a single day of 1398, some 100 000 Hindus be strangled in sight of the walls of Delhi, which had rejected surrender. In 1402, he dealt a mortal blow to Arab civilization by killing 90 000 inhabitants of Baghdad (Morgan, 1986; May, 2007).

Early Modern Times

A sovereign's right to war ('ius ad bellum') dominated the period of sixteenth to nineteenth century in which Europe came to dominate the world. What happened within a war was more or less up to the sovereign. In Europe as well as overseas, the results were clearly genocidal. Tenochtitlan, the Aztec capital in the valley of Mexico with its 100 000 inhabitants, was annihilated by Cortes' Spaniards in 1521 (Sánchez, 2004). Around 20 000 inhabitants of German Magdeburg, the richest city of Protestantism, were slaughtered on 10 May 1631 by Catholic troops under General Tilly in the 30 Years' War (Lorenzen, 2006).

Yet, in response to the extermination of indigenous Americans at the hand of Spanish Europeans, scholars began to set down most of the principles that have guided international criminal law ever since. Thus, the Dominican friar of Jewish converso descent, Francisco de Vitoria (1483–1546), stated in *De Jure belli Hispanorum in barbaros* of 1532:

> Diversity of religion is not a cause of just war. Extension of an Empire is not a just cause of war. / Wrong done is the sole and only just cause for making war. // The deliberate slaughter of the innocent is never lawful in itself. This is proved, firstly, by Exodus, ch. 23: 'The innocent and righteous slay thou not.' / It is not lawful within a State to punish the innocent for the wrongdoing of the guilty. / Hence it follows that even in war with the Turks it is not allowable to kill children. / And the same holds with regard to the women of unbelievers. / On this principle it follows that it is not lawful to slay either foreigners or guests who are sojourning among the enemy, for they are presumed innocent. / The same principle applies to clerics and members of a religious order / unless / they engage in actual fighting. / [It is] the obligation to see that greater evils do not arise out of the war than the war would avert. / It would not be right, for the purpose of assailing a few guilty, to slay the many innocent by use of fire or engines of war.
>
> Vitoria, 1532: Part III, Summary 10–11/13; // Chapters 35–37.

Parallel to the upbringing of Vitoria, a special knowledge, the expertise of birth control in the hands of midwives, was wiped out in Europe making his principles inoperative for centuries to come. To repopulate the Old World that had fallen from some 80–50 million inhabitants after the population catastrophe of the Black Death (1348–52) some 100 000 persons, mostly women, were put to death in the great witch hunt between 1350 and 1700 (Heinsohn and Steiger, 1999). As a consequence, the *European Population Explosion*, four centuries of youth bulges, pushed this small continent from 12 or 13 to 27% of world population between 1485 and 1914. This not only led to world domination but also to an increase of Europe's own population from 60 to 500 million between 1500 and 1916. In 1914, out of 1000 young men worldwide aged 15–29, 350 were either born in Europe or European in origin. They had no one to fear but themselves until their total fertility (children in a woman's lifetime) had fallen from five to six in 1890 to 1.5 in 1990.

During the *European Population Explosion*, some 60 million young Europeans, with the superior military technology of their ownership-based economies forced to innovate to meet the rate of interest, but too few positions in an overcrowded

continent, conquered their 'Lebensaum' (settlement space) in the Americas, Australia, Africa, and Northern Asia. In the wake of Europe's 'emigration,' a population of 55–85 million native Americans (low estimate of 8.4 to high of 112.5 million) for 1492 (Spanish conquest) was reduced by 85–95% around 1650. Even a reduction of 100–1.8 million of the indigenous population of the Western Hemisphere between 1490 and 1890 is not only considered due to emigration (Stannard, 1992) but also seriously contested by identifying epidemic diseases as the main killer (Cook, 1998).

Between 250 000 and 500 000 native Australians estimated for 1788 (British conquest) were reduced to 60 000 by 1921. All means of killing were employed. Yet, there is a controversy (Dawson, 2012) between those regarding outright genocide as the major cause of deaths (e.g., Reynolds, 1999) and others that do not deny massacres but see infections as the main culprit (e.g., Windschuttle, 2002).

Not much is known on internal African genocides before the twentieth century. However, between 1500 and 1870, the continent lost up to 20 million inhabitants into American slavery with 2 million dying on transports. The slave trade within Africa may have cost, for the same period, the life of 4 million. On the slave routes to – mostly Islamic – Asia, another 11 million may have perished in these gigantic democides (Rummel, 1994: p. 48). The depopulation of the so-called Congo Free State, between 1885 and 1908, by ruthless war, starvation, and diseases under Belgium's King Leopold II (1835–1909) is considered the most atrocious example of European colonialism in Africa with a death toll of 5 million (Forbath, 1977) to 10 million (Hochschild, 1998).

Twentieth Century to the End of World War II

The entire century is on record with close to 180 million victims of democide according to Rummel's (1997) statistics.

The Hague Conventions of 1899 and 1907 with "The Regulations Respecting the Law and Customs of War on Land" prohibited, in articles 22–28, the murder, ill-treatment, or deportation of enemy civilians and prisoners of war. The use of poison and special ammunitions was banned. Yet, a nation's harm to its own civilians was not taken into consideration. This had to wait for the 'Istanbul Trials' (Dadrian and Akçam, 2011) installed in January 1919 to inflict punishment on those responsible for the Armenian massacres. On 10 April 1919, Mehmet Kemal Bey, procurator of the province of Bogazliyan, became the first person executed for genocide, then called 'crime against humanity.'

In the Nuremberg Trials of Hitler-Germany's leadership (1945/46; Weinke, 2006) as well as in the Tokyo Trials of Imperial Japan's leaders (1946–48; Totani, 2009), *International Military Tribunals* sharpened the legal instruments for punishing perpetrators of crimes against humanity. After these first steps, however, punishment of this crime had to wait 45 years after passage of the UN Genocide Convention in 1948 for ad hoc tribunals to prosecute the perpetrators of 'ethnic cleansing' in Yugoslavia (1993) and – a year later – in Rwanda (genocide of Tutsis in spring 1994). On 1 July 2002, the *International Criminal Court* came into being as a permanent tribunal located in Rome (for its legal basis see Schabas, 2000).

The extermination of the Ottoman Armenians (1894–96; 1909; 1915/16; 1919/23) combined strategic, racial, and ideological intentions (Dadrian, 1997). A shrinking empire was to be consolidated by a homogenous Pan-Turkic population within an undivided settlement space under Islam as the only religion. That is why not only more than 1 million Armenians suffered genocide but also other Christian minorities like Maronites (1860), Assyro-Chaldaeans (1915), and Greeks (1915–23) with a combined death toll of some 440 000.

Since 1917, the Marxist–Leninist Regimes have killed around 110 million people (Rummel, 1994; Courtois, 1997) (**Table 1**). "One-party communist states are 4.5 times more likely to have used genocide than are authoritarian states" (Fein, 1993: p. 79). The killings came as a combination of politicides (elimination of like minded rivals for power), genocides (elimination of minority nations, mostly in the Soviet Union and Communist China), and ideological genocides (foremost the elimination of proprietors or ownership-'infected' people). The latter category became the largest victim group in history with some 50 million dead (55% in China, 40% in the Soviet Union, 5% in the other Marxist–Leninist regimes). The Marxist ideology that ownership caused the fettering of productive forces, crises, and mass poverty guided these mega-killings. Since, however, earning the rate of interest, which is tied to the creditory encumbrance of property, is the real engine of accumulation and technological progress, the Marxist expropriations stalled economic development. Before this nexus was sensed, the economic failures were usually blamed on sabotage and/or party comrades ideologically tied to the past. This analysis led to massive party purges like the 'Great Terror' in the Soviet Union (1934/36–38) with more than 7 million (Conquest, 1990) or the 'Cultural Revolution' in China (1964–75) with at least 3 million dead (Clark, 2008) while, from 1949 to 1976, some 27 million perished in the Laogai forced labor camps (Chang and Halliday, 2005; they give a total of 70 million victims of Maoism).

Annihilation through forced labor in camps was initiated by Trotzki on 26 June 1918. By 1922, 23 Soviet camps were at work forming the nucleus of Stalin's Gulag emerging in 1928 and eventually running some 8000 camps. Working people to

Table 1 Rough estimates of numbers of victims by genocides, politicides, and ideological genocides of Marxist–Leninist regimes

Soviet Union, 1917–87 (Stalin alone 20 million)	30–40 000 000
Red China plus Maoists before 1949	40–70 000 000
Khmer Rouge in Cambodia, 1975–79	2 000 000
North Korea, 1945 onward	1 700 000
Yugoslavia of Tito, 1941–87	1 100 000
North-Vietnam, 1945–87	1 700 000
Poland (deportations of Germans), 1945ff.	1 400 000
Chekhovslovakia (deportations of Germans), 1945ff.	240 000
Ethiopia, 1974–91	4 000 000
Albania, 1944–90	100 000
Communist Germany, 1945–52	96 000
Yugoslavia–serbia, since 1990	150 000

Rummel, R.J., 1994. Death by Government. Transaction Publishers, New Brunswick, NJ; Courtois, S. (Ed.), 1997. Le Livre Noir du Communisme: Crimes, terreur, répression. Robert Laffont, Paris; Chang, J., Halliday, J., 2005. Mao: The Unknown Story. Anchor Books, New York.

death became, in the long run, the most devastating killing tool of communism. It allowed the elimination of opponents by simultaneously using them for the creation of big projects – like canals, railroads, subways, etc. – to prove by 'miracles' the promised productive superiority of communism that otherwise was wanting.

Communists of the Soviet Union and Ethiopia turned the ancient siege tool of artificial famines into the most efficient killing techniques used so far. To crush national ambitions of the Ukraine, e.g., a Soviet made famine killed at least 3 million in the single winter (4 months) of 1932/33 (Snyder, 2010). Ethiopian communists managed to kill between 1 and 2 million peasants between 1983 and 1985 by the same means (Kissi, 2006).

Of all twentieth-century authoritarian regimes, Hitler's Germany was the most genocidal. Only its Fascist partner Croatia – albeit on a much smaller scale – came close. Since 1 September 1939, deformed newborn and mentally handicapped Germans as well as German soldiers severely wounded in the military campaign against Poland were killed by injections or in gas chambers. Close to 130 000 eventually ended in this genocidal euthanasia operation (Benzenhöfer, 2010). Later exterminations struck so-called socially inferior Gypsies ('drop-outs') as well as homosexuals, etc. A bigger target, however (outlined in the 'Generalplan Ost' (Masterplan East)), was 100 million Slavs (Czechs, Poles, White-Russians, Ukrainians, and Russians) living between Germany's Eastern border and the Ural Mountains, a territory marked to become German 'Lebensraum.' Seventy five thousand members of the Polish elite – most of them holders of academic degrees – were killed right in autumn 1939. The second Eastern campaign, against the Soviet Union (22 June 1941), was aimed at the direct killing of 30 million Slavs. Another 30 million were destined to be worked to death to lay the infrastructure for German settlements. The remaining 40 million were to be driven into Siberia with a calculated death toll of another 10 million. With 13 million Slavic civilians and prisoners of war killed by 1945, Hitler-Germany had fallen short of its 70 million goal but still had committed the largest genuine genocide of the twentieth century. Hitler-Germany's Eastern campaigns, thus, were not common wars (Hillgruber, 1993) but militarily forced genocides.

Parallel with Hitler-Germany's genocidal wars, the destruction of European Jewry – beginning 1939 in Poland and in full force since 25 June 1941 – was taking the life of 5–6 million Jews up to May 1945 (best overview Longerich, 2010). After more than 50 different theories research on the Holocaust is tangled up in perplexity. Joachim Fest (1992), the most experienced of all Hitler researchers, summarized more than 40 years of work: "I do not understand it [the murder of the Jews], and nobody who has ever been dealing with it has even come close […] to a convincing interpretation" (Fest, 2004: p. 3). Ian Kershaw, the highly respected English Hitler expert (1998a, 2000) goes along with his German opponent: "Auschwitz comes to the limits of what can be explained: historians can describe how it happened, but why it happened is a completely different question. // [Hitler was] / obsessed with a / hardly explainable desire for destruction" (Kershaw, 1998b: pp. 4f. // 2005: p. 36).

A minority position put forward by this author (Heinsohn, 1995, 2000; adopted by Léon Poliakov, 1996; Jonathan Friedman, 2011: p. 509) claims that Hitler wanted to reintroduce the archaic rights to (1) infanticide – to strengthen Germany internally by killing handicapped including newborn – and to (2) genocide – to give Germany strategic superiority externally – by annihilating the monotheistic people of the 10 Commandments whom he regarded as the first abolishers of these archaic practices. To Hitler, thus, the sanctity of life was the ultimate Jewish *Weltanschauung* (ideology), which could be eliminated like any other ideology – e.g., the communism of Red Army commissars, etc. – by exterminating its 'carriers.' To turn the entire German military into a gigantic SS, to change every German male into a killer was no small task. This would not work if the soldiers were time and again exposed to the Jewish code of ethics with the sanctity of life at its core. In August 1930, Hitler had diagnosed the influence of this Jewish code: "In every human being the Jew destroys the natural instinct of self-preservation."

As Marxists had found out by historical research that property was nothing natural and, therefore, could be abolished, Hitler had found out that sanctifying life was nothing eternal but a human Jewish invention that could be removed from Germany's consciousness by exterminating its standard bearers.

The World since 1945

Whereas during the war some 900 000 Japanese and some 600 000 German civilian victims of allied carpet and/or nuclear bombing may be categorized as crimes of war or democide, genuine genocide with a death toll of 3.3 million accompanied the treatment of German prisoners of war as well as the expulsion of 16 million ethnic Germans from Russia (since 1944), Eastern and Southern Europe, and the territories Germany had to cede. In the postwar world only the genocidal crimes of Communist China surpassed the perpetrations of Germany's neighbors.

Worldwide (exempting the Soviet Union and Communist China) there were 68 conflicts with at least 10 000 casualties from 1950 to 2010 with a combined death toll of 36 million (Heinsohn, 2010). More than 100 crimes fulfilling the definition of the UN Genocide Convention have been committed since 1945 (Harff and Gurr, 1996). In the three decades from 1980 to 2010 alone, some 80 distinct ethnic and religious groups – mostly in Africa, Asia, and Latin America – have been victimized with at least 20 of them going through full blown genocide: Aceh and Timorese (Indonesia), Albanians (Serbian Kossovo), Bosnian Muslims, Cambodians, Isaaq (Somalia), Kurds (Iraq), Mayans (Guatemala), Tibetans, Tutsi (Rwanda) as well as 12 first nations in Brazil.

Many post-1945 conflicts in Afghanistan, Africa, Indonesia, Latin America, and Pakistan run on ethnic divisions and usually include acts of genocide when strife erupts. The cause for conflict, however, cannot be found in ethnicity as such because many regions were ethnically diverse before, during and after periods of killing. Therefore, another factor must be at work. Youth bulges – driving Europe's violence from 1500 to 1945 – again are

the main candidate with these very regions multiplying, between 1945 and 2010, from 450 million to nearly 3 000 000 000 inhabitants.

The Future

To evaluate future chances of mass killing, a more precise approach can be tried by computing what the author has called the 'fratricidal war index' (Heinsohn, 2011). It expresses the ratio between 15- and 19-year-old young men entering the strife of life and 55- to 59-year-old men approaching retirement. The range stretches from 0.7 (Japan, i.e., 10 retiring men are succeeded by only seven young men) to more than 6 (e.g., Afghanistan, Palestine, Uganda, Yemen), where 10 retirees are followed by 60 angry young men ready to embrace any justification – religious, racial, economical, etc. – for fighting their way up. An index of 3 (e.g., Egypt, Libya, Saudi Arabia) or higher points to violence with genocide and cross border wars the least but gang warfare, coups, and civil wars the most likely outcome. Sub-Sahara Africa with most nations indexed 5 or 6 in 2010 may still face decades of bloodshed. However, the index also proves helpful for forecasting which nations will not return to major killings because their index has already peaked and fallen back to 2 (e.g., Lebanon, Tunisia, Turkey).

Since most powers honoring human rights have indexes below 1 with America marking the maximum of 1.2, humanitarian intervention has the best prospects in nations indexed 2.5 or lower. After all, how many times can aging nations send a family's only son or even only child into mortal danger to hinder 10 or 20 superfluous young men in youth bulge nations from killing each other?

See also: Ancient Greece and Rome, History of; Antisemitism; Conflict: Organizational; Crowds in History; Dehumanization; Discrimination: Racial; Ethnic Cleansing, History of; Ethnic Conflicts; Eugenics as a Basis of Population Policy; Holocaust, The; Judaism; Mass Killings and Genocide, Psychology of; Middle Ages, The; Militaries and Militarization, Anthropology of; Minorities; National Socialism and Fascism; Nationalism, Sociology of; Racism, History of; Racism, Sociology of; Revolutions, History of; Terrorism; Tribe; Violence, History of; War Crimes Tribunals, Crimes against Humanity and Transitional Justice; Warfare in History; Youth Gangs.

Bibliography

Anderson, C., 2007. Indian Uprising of 1857–8: Prisons, Prisoners and Rebellion. Anthem Press, New York.
Arnaud, M., 2009. La mémoire et le pardon: Les commissions de la verité et de la réconciliation en Amérique latine. L'Harmattan, Paris.
Babeuf, F.N., 1795. Du système de la dépopulation…avec la guerre de la Vendée. Franklin, Paris.
Benzenhöfer, U., 2010. Euthanasia in Germany before and during the Third Reich. Klemm & Oelschläger, Münster, Ulm.
Bleckmann, B., 2006. Zu den Motiven der Christenverfolgung des Decius. In: Johne, K.P., et al. (Eds.), Deleto paene imperio Romano: Transformationsprozesse des Römischen Reiches im 3. Jahrhundert und ihre Rezeption in der Neuzeit. Franz Steiner, Stuttgart, pp. 57–72.
Berlioz, J., 1996. Tuez-les tous, Dieu reconnaîtra les siens: Le massacre de Béziers (22 juillet 1209) et la croisade contre les Albigeois vus par Césaire de Heisterbach. Loubatieres, Portet-sur-Garonne.
Chalk, F., Jonassohn, K., 1990. The History and Sociology of Genocide. Analysis and Case Studies. Yale University Press, New Haven and London.
Chang, J., Halliday, J., 2005. Mao: The Unknown Story. Anchor Books, New York.
Charny, I.W., 2000. Encyclopedia of Genocide. ABC-CLIO, Santa Barbara, CA et al.
Clark, P., 2008. The Chinese Cultural Revolution: A History. Cambridge University Press, Cambridge and New York.
Conquest, R., 1990. The Great Terror. Oxford University Press, New York and Oxford.
Cook, N.D., 1998. Born to Die: Disease and New World Conquest, 1492–1650. Cambridge University Press, Cambridge, England.
Courtois, S. (Ed.), 1997. Le Livre Noir du Communisme: Crimes, terreur, répression. Robert Laffont, Paris.
Dadrian, V.N., 1976. A typology of genocide. International Review of Modern Sociology 5, 201–212.
Dadrian, V.N., 1997. The History of the Armenian Genocide: Ethnic Conflict from the Balkans to Anatolia to the Caucasus (1995). Berghahn Books, Providence, RI and Oxford, UK.
Dadrian, N., Akçam, T., 2011. Judgement at Istanbul: The Armenian Genocide Trials. Berghahn, Books, Brooklyn, NY, and Oxford, UK.
Dawson, J., 2012. Washout: On the Academic Response to the Fabrication of Aboriginal History. New Expanded Edition. Macleay Press, Sydney, NSW.
Dodds, G.L., 1998. The Zulus and Matabele, Warrior Nations. Weidenfeld Military, London, UK.
Dubois, L., 2005. Avengers of the New World: The Story of the Haitian Revolution. Belknap Press of Harvard University, Cambridge, MA.
Elder, T., 2012. Guide to the Raphael Lemkin (1900–1959) Collection. American Jewish Historical Society: Center for Jewish History. http://findingaids.cjh.org/?pID=109202.
Elliot, G., 1972. Twentieth Century Book of the Dead. Allan Lane/The Penguin Press, London.
Evans, G., 2008. The Responsibility to Protect: Ending Mass Atrocity Crimes once and for All. Brookings Institution Press, Washington, DC.
Fein, H., 1990. Genocide: A Sociological Perspective. Sage Publications, London.
Fein, H., 1993. Accounting for genocide after 1945: theories and some findings. International Journal of Group Rights I, 79–106.
Fest, J., 1992. Hitler. Harcourt, Brace, Jovanovich, New York (1973, 1st).
Fest, J., 10 September 2004. Mitleidlosigkeit bis zum allerletzten Punkt. Die Welt, p. 3.
Forbath, P., 1977. The River Congo: The Discovery, Exploration, and Exploitation of the World's Most Dramatic River. Harper & Row, New York.
Friedman, J.C. (Ed.), 2011. The Routledge History of the Holocaust. Routledge, London and New York.
Funk, M.T., 2010. Victims' Rights and Advocacy at the International Criminal Court. Oxford University Press, New York.
Garrisson, J., 2000. 1572: la Saint-Barthélemy. Editions Complexe, Bruxelles, Belgium.
Garsse, Y.v., 1970. A Bibliography of Genocide, Crimes against Humanity and War Crimes. Studiecentrum voor Kriminilogie en Gerechtelijke Geneeskunde, Sint Niklaas Waas.
Glick, L.B., 1994. Religion and genocide. In: Charny, I.W. (Ed.), The Widening Circle of Genocide. Genocide: A Critical Bibliographic Review, vol. 3. Transaction Publishers, New Brunswick and London, pp. 43–74.
Goldhagen, D., 2009. Worse than War: Genocide, Eliminationism, and the Ongoing Assault on Humanity. Public Affairs, New York.
Gould, M., 2012. The Struggle for Modern Nigeria: The Biafran War 1966–1970. I. B. Tauris, London, UK.
Gritschneder, O., 1993. Der Führer hat Sie zum Tode verurteilt…": Hitlers" Röhm-Putsch"-Morde vor Gericht". C.H. Beck, München.
Harff, B., Gurr, T., 1996. Victims of the state: genocides, politicides and group repression from 1945 to 1995. In: Jongman, A.J. (Ed.), Contemporary Genocides: Causes, Cases, Consequences. PIOOM, Leiden, pp. 33–58.
Hasbrouck, A., May 1935. The conquest of the desert. The Hispanic American Historical Review 15 (2), 195–228.
Hayner, B., 2001. Unspeakable Truths: Facing Challenge of Truth Commissions. Routledge, New York.
Heinsohn, G., 1995. Warum Auschwitz? Hitlers Plan und die Ratlosigkeit der Nachwelt. Rowohlt, Reinbek.
Heinsohn, G., 1997. Post-Genocidal Reconciliation in Rwanda: Are There Lessons from Germany? Raphael-Lemkin-Institut für Xenophobie- und Genozidforschung, Bremen.
Heinsohn, G., 1998. 1999 Lexikon der Völkermorde, first ed. Rowohlt, Reinbek.
Heinsohn, G., 2000. What makes the holocaust a uniquely unique genocide? Journal of Genocide Research 2 (3), 411–430.

Heinsohn, G., 2003. Söhne und Weltmacht: Terror im Aufstieg und Fall der Nationen. Orell & Fuessli, Zurich.
Heinsohn, G., 2010. Is international terror caused by the creation of Israel? Genocide Prevention Now (GBN) (3), Summer. www.genocidepreventionnow.org/Home/GPNISSUES/Issue3Summer2010/.
Heinsohn, G., 2011. Gesellschaftliche Kriegstheorien: Demografische Faktoren. In: Jäher, T., Beckmann, R. (Eds.), Handbuch Kriegstheorien. VS Verlag, Wiesbaden, pp. 72–95.
Heinsohn, G., Steiger, O., 1999. Birth control: the political-economic Rationale behind Jean Bodin's 'Démonomanie'. History of Political Economy 31 (2), 423–448.
Herodotus, 1987. The History (D. Greene, Trans.). The University of Chicago Press, Chicago and London.
Hillgruber, A., 1993. Hitlers Strategie: Politik und Kriegfuehrung 1940–1941 [1963[1]]. Bernard & Graefe, Bonn.
Hochschild, A., 1998. King Leopold's Ghost: A Story of Greed, Terror, and Heroism in Colonial Africa. Mariner Books, New York.
Homer, 1952. The Iliad and the Odyssey (S. Butler, Trans.). Encyclopedia Britannica, Chicago et al.
Horowitz, I.L., 1976. Genocide and State Power. Transaction Books, New Brunswick, NJ. The fourth edition of 1997 (same publisher) was titled Taking Lives: Genocide and State Power.
Jones, A., 2010. Genocide: A Comprehensive Introduction. Routledge, London.
Jonassohn, K., Björnson, K.S., 1998. Genocide and Gross Human Rights Violations in Comparative Perspective. Transaction Publishers, New Brunswick, NJ.
Keegan, J., 1993. A History of Warfare. Alfred A. Knopf, New York.
Kershaw, I., 1998a. Hitler: 1889–1936 Hubris. Allan Lane, London.
Kershaw, I., October 1, 1998b. In gewisser Weise war er der Mann ohne Eigenschaften: Die Geschichte Hitlers ist auch die Geschichte seiner Unterschätzung. Frankfurter Allgemeine Zeitung, p. 44f.
Kershaw, I., 2000. Hitler 1936–1945: Nemesis. Allan Lane, London.
Kershaw, I., March 19, 2005. Was Wäre Gewesen, Wenn? Frankfurter Allgemeine Zeitung, p. 36.
Kiernan, B., 2007. Blood and Soil: A World History of Genocide and Extermination from Sparta to Darfur. Yale University Press, New Haven, CT.
Kissi, E., 2006. Revolution and Genocide in Ethiopia and Cambodia. Lexington Books, Lanham, MD.
Kuper, L., 1981. Genocide: Its Political Use in the Twentieth Century. Yale University Press, New Haven and London.
Lemkin, R., 1944. Axis Rule in Occupied Europe. Carnegie Endowment for International Peace, Washington, DC.
Levene, M., 2005. Genocide in the Age of the Nation State. Palgrave Macmillan, New York.
Littell, F.H., 1969. Wild Tongues: A Handbook of Social Pathology. Macmillan, New York.
Lichtheim, M., 1976. Ancient Egyptian Literature. Volume II: The New Kingdom. University of California Press, Berkeley et al.
Linton, S., 2010. Completing the circle: accountability for the crimes of the 1971 Bangladesh war of liberation. Criminal Law Forum 21, 191–311.
Longerich, P., 2010. Holocaust: The Nazi Persecution and Murder of the Jews. Oxford University Press, Oxford, England.
Lorenzen, J.N., 2006. 1631-Die Zerstörung Magdeburgs. In: Idem (Ed.), Die Großen Schlachten. Mythen, Menschen, Schicksale. Campus, Frankfurt am Main, pp. 55–100.
Luckenbill, D.D., 1989. Ancient Records of Assyria and Babylonia (1926), vol. 1. Histories & Mysteries of Man Ltd, London.
MacCuarta, B., 1993. Ulster 1641, Institute of Irish Studies. Queen's University of Belfast, Belfast, NI.
Martin, A. (Ed.), 2009. La mémoire et le pardon: Les commissions de la vérité et de la réconciliation en Amérique latine. L'Harmattan, Paris.
Mascarenhas, A., 1986. Bangladesh: A Legacy of Blood. Hodder and Stoughton, London, UK.
May, T., 2007. The Mongol Art of War. Westholme Publishing, Yardley, PA.
Maybury-Lewis, D., 2002. Genocide against Indigenous People. In: Hinton, A.L. (Ed.), Annihilating Difference: The Anthropology of Genocide. University of California Press, Berkeley, CA, et al., pp. 43–53.
Mayor, A., 2010. The Poison King: The Life and Legend of Mithridates. Princeton University Press, Princeton, NJ.
Morgan, D., 1986. The Mongols. Blackwell, Oxford.
Moses, A.D. (Ed.), 2008. Empire, Colony, Genocide: Conquest, Occupation, and Subaltern Resistance in World History. Berghahn Books, Brooklyn, NY.
Nuhanovic, H., 2007. Under the UN Flag: The International Community and the Srebrenica Genocide. DES, Sarajevo.
Pegg, M.G., 2008. A Most Holy War: The Albigensian Crusade and the Battle for Christendom. Oxford University Press, New York.
Philipps, J.P., 2004. The Fourth Crusade and the Sack of Constantinople. Viking, New York.
Pinker, S., 2011. The Better Angels of Our Nature: The Decline of Violence in History and its Causes. Viking Adult, New York.
Poliakov, L., 1996. Les vraies raisons des crimes hitlériens. L'Infini: Littéarture, Philosophie, Art, Science, Politique, 76–79 (46), winter.
Porter, J.N. (Ed.), 1982. Genocide and Human Rights: A Global Anthology. University Press of America, Washington, DC.
Power, S., 2009. A Problem from Hell: America and the Age of Genocide. Harper Perennial, New York.
Pritchard, J.B. (Ed.), 1969. Ancient Near Eastern Texts Relating to the Old Testament. Princeton University Press, Princeton, NJ.
Reynolds, H., 1999. Why Weren't We Told? Viking, New York.
Rohde, D.W., 1997. Die letzten Tage von Srebrenica: Was geschah und wie es möglich wurde. Rowohlt, Reinbek.
Rummel, R.J., 1994. Death by Government. Transaction Publishers, New Brunswick, NJ.
Rummel, R.J., 1997. Statistics of Democide: Genocide and Mass Murder since 1900. Center for National Security Law, School of Law, University of Virginia, Charlottesville, VA. http://www.hawaii.edu/powerkills/NOTE5.HTM.
Runciman, S., 1958. The Sicilian Vespers: A History of the Mediterranean World in the Later Thirteenth Century. Cambridge University Press, Cambridge, England.
Sánchez, J.L., 2004. Tenochtitlan. Die letzte Schlacht der Azteken. Unionsverlag Taschenbuch, Zurich.
Schabas, W., 2000. Genocide in International Law: The Crime of Crimes. Cambridge University Press, Cambridge, England.
Skrynnikov, R.G., Graham, H.F., 1981. Ivan the Terrible. Academic International, Gulf Breeze, FL.
Snyder, T., 2010. Bloodlands: Europe between Hitler and Stalin. Basic Books, New York.
Stannard, D.E., 1992. American Holocaust: The Conquest of the New World. Oxford University Press, New York.
Stanton, G.H., 1998. The 8 Stages of Genocide. http://www.genocidewatch.org/aboutgenocide/8stagesofgenocide.html.
Staub, E., 2011. Overcoming Evil: Genocide, Violent Conflict and Terrorism. Oxford University Press, New York.
Szawłowski, R., 2005. Diplomatic file: Raphael Lemkin (1900–1959) - the polish lawyer who created the concept of 'Genocide'. The Polish Quarterly of International Affairs 1 (2), 98–133.
Thucydides, 1954. The Peleponnesian War (R. Warner, Trans.). Penguin Books, Harmondsworth.
Totani, Y., 2009. The Tokyo War Crimes Trial: The Pursuit of Justice in the Wake of World War II. Harvard University Press, Cambridge, MA.
Valentino, B.A., 2004. Final Solutions: Mass Killing and Genocide in the 20th Century. Cornell University Press, Ithaca, NY.
Vitoria, F de, 1532. De Jure Belli Hispanorum in Barbaros. http://en.wikisource.orgMartin/wiki/De_Indis_De_Jure_Belli/Part_3.
Weinke, A., 2006. Die Nürnberger Prozesse. C.H. Beck, München.
Windschuttle, K., 2002. The Fabrication of Aboriginal History, Volume One: Van Diemen's Land 1803–1847. Macleay Press, Sydney, NSW.
Zimmerer, J., Zeller, J., Neather, E., 2008. Genocide in German South-West Africa: The Colonial War (1904–1908) in Namibia and Its Aftermath. Merlin Press, London, UK.

Genocide: Social and Economic Aspects

Ross Kleinstuber, University of Pittsburgh at Johnstown, Johnstown, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews some of the major recent theoretical developments in the criminological study of genocide, discusses the possibilities for interdisciplinary theory integration, and discusses themes common among existing theories. This article of the existing genocide literature suggests that genocide is likely when a dominant group perceives a vulnerable and dehumanized group as threatening its position of dominance and the leaders of the dominant group are able to direct such insecurities into targeted violence in the absence of social control mechanisms. The article concludes by calling for empirical tests of these theoretical claims.

Introduction

According to Goldhagen, "[T]he problem of genocidal killing is worse than war. Until now, the world's peoples and governments have done little to prevent or stop mass murdering. Today, the world is not markedly better prepared to end this greatest scourge of humanity" (2009: p. xi). In the twentieth century, genocide claimed four times more lives than war (Alvarez, 2001: p. 10); yet, Goldhagen's indictment could be directed easily against criminologists too. As recently as 2005, Hagan et al. called out the discipline for its "denial and neglect of these crimes" and argued that criminology is capable of explaining genocide and other international law violations; rather, its failure to do so "itself needs explanation" (2005: p. 556). Although other disciplines – most notably political science – have studied and theorized about genocide, the unique perspective of criminologists is essential for a complete understanding of genocidal violence.

Fortunately, recent years have seen an increase in criminological interest in the study of genocide, and numerous theories have been offered. Although these theories are in their infancy and still require further empirical testing, they point to some important social, economic, and political factors that are likely at work during genocidal incidents: most notably dehumanization, threat, a vulnerable target group, fanatical leadership, and absent control mechanisms. These theories also point to the importance of developing a multidisciplinary approach that can expand understandings of genocide and allow each discipline to learn from the insights and perspectives of others. As such, this article reviews some of the major recent theoretical developments in genocide studies and discusses some of the commonalities among the theories that seem like promising avenues for theoretical integration, empirical testing, and policy making.

Political Science Theories

The academic study of genocide is the most mature in the political science literature, where the act is viewed as a form of political (rather than criminal) violence. Contemporary political theories of genocide can be divided into two main paradigms: strategic theories that look at the importance of war and ideological ones that look at genocidal leaders' motives (Straus, 2012). Strategic theorists have shown that government mass killings are more likely to occur during civil wars or armed uprisings (Ulfelder and Valentino, 2008) and have suggested that the territorial loss that occurs during war creates insecurity and threat that drive genocide (Midlarsky, 2005). Straus (2006) has claimed that war legitimizes violence, increases feelings of insecurity, and empowers hardliners. The ideological paradigm focuses on the motives and goals of genocidal leaders. According to Straus, "The most consistent focus in the literature is on ideological visions of and for the state …" (2012: p. 548). For example, Weitz (2003) focused on genocidal leaders' visions of racial or national utopia, and Harff (2003) found that 'exclusionary ideologies' are central in the beliefs of genocidal leaders. Integrating these two paradigms suggests that war contributes to defining some groups as enemies (Semelin, 2007) and that when leaders with visions of ethnic, racial, or national purity feel their positions of power threatened, they are likely to resort to violence against a threatening subordinate group in an attempt to maintain their grip on power and to establish a more homogenous polity (Straus, 2012).

The macro-level, top-down approach of political theories fails to consider the actions of those who carry out the violence. It is one thing to understand why the political elites seek to maintain power or create an ethnically pure state; it is quite another to explain how they are able to get the masses to overcome their moral prohibitions and start killing their neighbors, friends, coworkers, and classmates. It also fails to consider how the target group is selected. Humans are simultaneously members of multiple different groups. How does one group affiliation take primacy and why? Although the political theories are good for understanding the politics of genocide and the motives of leaders, from a criminological perspective, they leave many unanswered questions as to what social forces are at work, why genocide is likely to and able to take place, and what the relationship between the political elites and the ordinary citizens is like. As will be shown, however, the two ideas of threat and ideology are important contributions that dovetail well with the emerging

criminological theories and provide an excellent nexus for cross-disciplinary theorizing and testing.

Early Criminological Theorizing: Out-Group Hatred

Browning (1992) and Goldhagen (1996) both challenged the idea popular in the social psychological literature that Nazi Germany was overtaken by psychopaths who used the power of obedience to authority to convince the populous to follow their command. They both pointed out the times when Germans refused to comply with commands and noted Germany's long history of anti-Semitism, which caused most Germans to view the Jews as subhuman before Hitler's ascendency. Drawing from these historical accounts, Brannigan (1998) offered a criminological theory of genocide that suggests that genocidal propaganda often serves the purpose of forging in-group solidarity, which eliminates the normal inhibitions against killing out-group strangers. Brannigan (1998), however, suggested that propaganda campaigns like the ones waged by the Nazis are not necessary for the creation of such hostilities toward out-group members. Brannigan (1998) contended that intergroup conflict and prejudice may be one of the by-products of cooperative human society. He argued, "The coalitionist or 'tribal' instinct is at the core of human nature. We cooperate to compete. But the dark side of this sociability is group prejudice and an intolerance of out-groups. In fact, the maintenance of group boundaries is premised on instinctual hatred of the Other" (Brannigan, 1998: p. 271, citations omitted). He even claimed that as societies become more cooperative, they also become more violent, confrontational, and xenophobic.

On the basis of his reading of the literature, Brannigan (1998) claimed that the root of genocidal behavior is xenophobic out-group hatred, but he also realized that this was not enough to explain genocide. After all, if out-group hatred is a universal part of the human condition, it becomes meaningless in attempting to explain or predict genocide since it does not vary. According to Brannigan, "Evolution may have created numerous appetites and instincts, but cultures are mechanisms for both switching them on and exciting other 'instincts' or mechanisms that can curb them or switch them off" (1998: p. 273). The question is, what are those other factors that trigger xenophobic instincts to annihilate the 'Other'? In other words, what factors must be present and interact with dehumanizing beliefs to trigger genocide, and which group affiliations will take precedence?

Techniques of Neutralization

Drawing from the insights of Milgram's (1963) obedience study and using a speech from Heinrich Himmler to support his proposition, Neubacher (2006) proposed a theory of genocide based on Sykes and Matza's (1957) techniques of neutralization. Neubacher (2006) argued that the Milgram study teaches us that when orders are given by someone with a strong position of authority, the actor is able to see him- or herself as a tool and therefore not responsible for his or her behavior. In addition, he relied on Kelman's (1973) expansion of Milgram's work, which suggested that the routinization of an act and the dehumanization of the victim makes it easier for perpetrators to engage in massacres. Neubacher (2006) connected these ideas to three techniques of neutralization: denial of responsibility, denial of victim, and appeal to higher loyalties.

The challenge, of course, is to apply this theory to mass violence perpetrated by the state. It would seem that the neutralization needed to engage in such atrocities is far greater than that needed for involvement in traditional acts of delinquency. Neubacher argued, "States violating human rights use the same neutralization techniques as individuals. State neutralization, however, is spread via the state's propaganda machine, thus having a far greater impact than individual neutralization" (2006: p. 795). He then analyzed a 1943 speech by Himmler and pointed out the neutralization techniques employed by the Schutzstaffel (SS) commander; Neubacher (2006) concluded that neutralization theory offers an explanation for both state involvement and why ordinary citizens are able to perpetrate the violence needed. Something seems to be missing, however. Sykes and Matza's (1957) original theory applied to juvenile delinquents and was an attempt to expand Sutherland's (1947) differential association theory by explaining what exactly juveniles learned from each other. Sykes and Matza (1957) believed that juveniles learn the techniques of neutralization through peer interaction and peer pressure. Although this may apply to the propaganda campaign waged by genocidal states on their own citizens, it fails to explain why the state itself desires the destruction of the targeted population and how they learned their neutralization techniques. Whereas street crime is a relatively constant phenomenon with a fairly obvious motive and a steady supply of peers from whom another juvenile can learn the techniques of neutralization, it is not clear from whom states and leaders learn these techniques or why they seek to destroy the victims. In other words, what causes the state-level motivation and neutralization that leads to the propaganda campaign and the eventual genocide, and how can this theory be used to predict or prevent genocide? For example, Goldhagen (1996) pointed out that Germans had a long history of anti-Semitism (i.e., denial of victim), and did not need to learn to neutralize anti-Semitic acts. So why did the Holocaust begin when it did? What other conditions caused the Germans to turn their hatred of the Jews into overt actions of annihilation?

Recent Criminological Approaches

Recently, criminologists have started to acknowledge that genocide involves complex processes that occur contemporaneously and at multiple levels of analysis. So although these earlier theories are lacking, their main points are useful starting places for building complex, multilevel models, and criminologists have started to propose such models. Four of these theories – Winton and Unlu's (2008) integration of violentization and circumplex theories, Hagan and Rymond-Richmond's (2009) collective action

theory, Maier-Katkin et al.'s (2009) theory of social conformity, and Rothe and Mullins' (2009) integrated theory of international criminal law violations – are described in the following sections.

Violentization and Circumplex Theory

In an attempt to bridge the micro- and macro-level interactions present in genocide, Winton and Unlu (2008) proposed integrating violentization theory from the homicide literature and the circumplex model from family therapy research. According to violentization theory, violent criminal behavior is the result of a microlevel process of development that occurs in four stages: brutalization, defiance, violent dominant engagements, and virulency (Athens, 2003). In the brutalization stage, one learns about violence through experience, demonstration, or witnessing violence. In the defiance stage, one learns a belief system that helps to legitimate and justify violence, such as needing to use violence as a means of stopping violence. In the third stage, violent perpetrators test out what they have learned in the first two stages by actively engaging in violence, and in the final stage, the perpetrator has come to identify him- or herself as violent (Athens, 2003). Winton and Unlu (2008) applied these ideas to genocide and suggested that those individuals who engage in genocidal violence are prepared to act violently toward each other through assaults, threats, observation, and coaching; are presented with a belief system and violent repetition that justifies violence; are encouraged to start engaging in violent acts of 'collective self-defense'; and eventually come to define themselves as violent. Winton and Unlu (2008) added a fifth stage to genocidal violentization – extreme virulency – to address the fact that genocidal violence often involves brutalization beyond simple killing. At the macro-level, Winton and Unlu (2008) drew from the circumplex model, which suggests that families are better when they are balanced in regards to cohesion and flexibility (Olson and DeFrain, 1997). In other words, families that are too enmeshed or too disengaged and families that are too flexible or too rigid are likely to experience dysfunctions (Olson and DeFrain, 1997). In their theory, Winton and Unlu (2008) applied these concepts to societies.

In a study of 13 genocides from the twentieth century, Winton (2005) concluded that genocidal societies are overly rigid and enmeshed, and in their examination of the Bosnian genocide, Winton and Unlu (2008) determined that when there is a crisis or conflict in a society that is becoming more rigid with increasing group cohesion, they are less able to cope with the crisis and are more likely to engage in the violentization process. Furthermore, in his examination of court documents from both the International Criminal Tribunal for the former Yugoslavia and International Criminal Tribunal for Rwanda, Winton (2011) demonstrated how both societies went through the violentization process (see **Figure 1**). This theory attempts to explain both the macro- and the micro-level factors involved in genocidal processes: the level of social cohesion and flexibility plus the microlevel process of violence escalation. It also jives well with political theories. As Straus explained, "Genocide is the product of a process or spiral of escalation in which alternative strategies of violence might have been previously tried or exercised" (2012: p. 554). Yet, this theory fails to explain why societies become less flexible or more enmeshed and why the violence becomes genocidal or is even directed toward the target group. Even Winton and Unlu (2008) admitted that their theory is not useful for predicting and preventing genocide – just explaining the process post hoc.

Macro–Micro–Macro Theory of Collective Action

Using data from the Darfur genocide collected by the US State Department's Atrocities Documentation Survey (ADS), Hagan and Rymond-Richmond (2009) created and tested a collective action theory of genocide that also drew upon differential association and collective efficacy theory in an attempt to link macro- to micro- and back to macro-level processes. Using ADS data, Hagan and Rymond-Richmond (2009) found that genocidal intent was expressed at both the individual and the collective levels, so they attempt to create a theoretical model that is able to account for this dynamic. The thrust of their

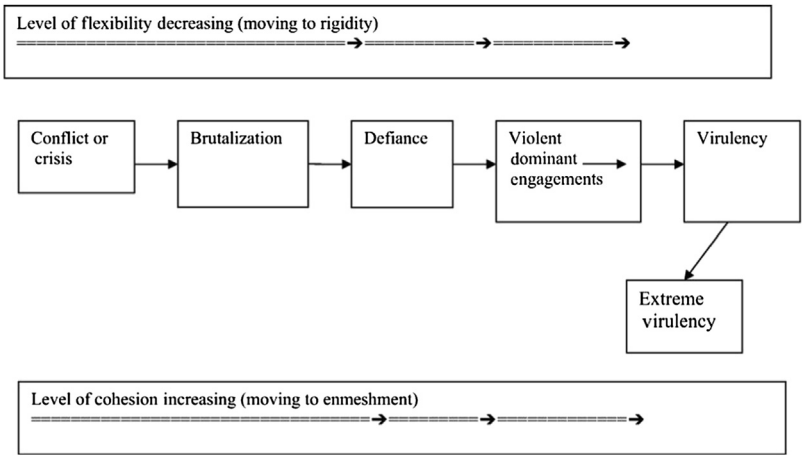

Figure 1 Winton and Unlu's (2008) violentization and circumplex model. Reprinted with permission from Elsevier.

argument is that at the macro level, there was increasing competition for land and resources between Arabs and Africans and a state-led ideology of Arab Muslim supremacy that constructed the two competing groups. At the individual level, members of the Arab group internalized the dehumanizing images and then engaged in violent actions. These individual acts of violence combined into collective action, causing macro-level violence.

As shown in **Figure 2**, Hagan and Rymond-Richmond (2009) have proposed that an Arab supremacist ideology in Darfur was used to justify an Arabization policy in the face of growing resource competition. Because of strained resources, genocide provided many Arabs with the opportunity to acquire goods and property from their black African neighbors. Hagan and Rymond-Richmond (2009) proposed that these macro-level features led to a socially constructed division between Arabs and Africans in Sudan. These socially constructed identities, in turn, incited the use of racial epithets by individuals, which helped to dehumanize the targeted group. When these individual expressions of racial hatred were concentrated within a geographic and social location, they became aggregated into collective racial intent. This collective intent, however, needed to be transformed into a violent fury that connected the racial intent to targeted violence. This is the final link in their model, in which a genocidal state directs the anger and hatred of a group into lethal destruction. Once the fury of one group has been unleashed onto another group, violence becomes genocidal. A key point for Hagan and Rymond-Richmond (2009) is that some kind of spark must ignite the angers and fears of the dominant group; they hypothesized that this spark is racial and collective. They suggested that their macro–micro–macro model explains both the state's involvement in genocide and "how and why ordinary people enlist in genocidal killing" (Hagan and Rymond-Richmond, 2009: p. 168). Hagan and Rymond-Richmond (2009) provided compelling evidence from the Darfur genocide to support their theory, including evidence that racial epithets were linked to increased severity of attacks, that the state was instrumental in leading the genocide and in identifying black Africans as the target, and that competition for resources and land intensified.

Theory of Conformity

Using the 1941 massacres of Jews in Jedwabne as a case study, Maier-Katkin et al. (2009) developed a theory of crimes against humanity (including, but not limited to, genocide) based on conformity. Drawing off of strain, social learning, and differential association theories, they presented a theory composed of six propositions: social strain and angry aggression, primary group affiliations, normal processes of socialization, group structure, ideology and self-interest, and displacement.

According to their first proposition, a necessary (but not sufficient) precondition to any crime against humanity is "intense and widely prevalent societal-level strain—caused by insecurity, danger, scarcity, deprivation, political and economic oppression, and/or military occupation," and that the more intense the strain is, the more likely a crime against humanity is (Maier-Katkin et al., 2009: p. 239). They argued in their second proposition that this "societal-level strain produces widespread individual-level fear, arousal and angry aggression" (Maier-Katkin et al., 2009: p. 241). This insecurity inhibits intergroup solidarity and drives people to seek safety in their most immediate primary groups. The third proposition is that once intergroup solidarity breaks down, primary group socialization takes on increased importance and can cause individuals to participate in horrific acts of violence. In other words, rather than being deviant, crimes against humanity are the result of conformity to social roles and expectations of the immediate social group (i.e., appeal to higher loyalties). Drawing off Milgram's obedience studies and Zimbardo's prison experiment, the fourth proposition is that during times of societal strain, individuals are more submissive and more accepting of social roles, thus giving socialization more influence. In this stage, leadership is crucial. If those in positions of influence happen to be

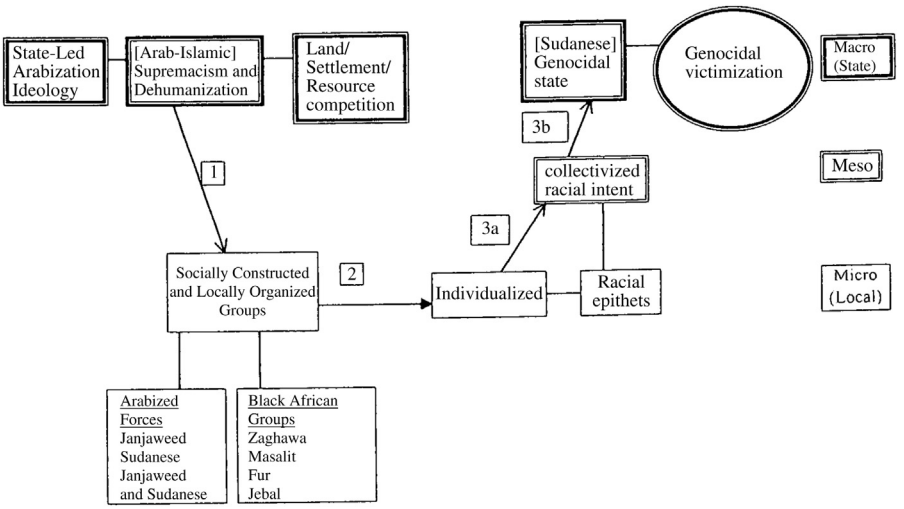

Figure 2 Hagan and Rymond-Richmond's (2009) macro–micro–macro transformation model of genocide. Reprinted with permission from Cambridge University Press.

reasonable or moderate, there is a strong likelihood of avoiding atrocities; if, however, radical elements that appeal to emotion take control, the stage is set for mass slaughter. Fifth, there also must be some ideology that motivates violence. It is not enough to have strain, fear, and intensified in-group solidarity. For crimes against humanity to occur, Maier-Katkin et al. stated there need to be "belief systems that dehumanize others, justify or recommend violence and associate 'higher' values with immediate self-interest" (2009: p. 246). Finally, when the first five propositions are present, for crimes against humanity to occur, the perpetrators need to have a nonthreatening scapegoat to target. If the aggression is directed at the actual source of strain, it represents self-defense or resistance rather than a crime; however, when the actual source is not vulnerable, the anger and aggression of the stressed group may be displaced onto a vulnerable group, if such a group exists. These six propositions provide a guide for understanding genocidal processes and also suggest that crimes against humanity cannot be studied as pathological or deviant. If we are to understand the actual criminal acts, we need to situate them within their context and understand that they are often a result of conformity to existing (but changing) social norms. Furthermore, numerous interacting factors are needed in sufficient quantity. None of the variables discussed by Maier-Katkin et al. (2009) are sufficient by themselves. They must coexist in just the right volatile mixture to produce crimes against humanity.

An Integrated Theory

Finally, Rothe and Mullins (2009) proposed an integrated theory of international criminal law violations. Like the other theories reviewed in this section, Rothe and Mullins (2009) believed that a single level of analysis is insufficient to explain such complex crimes. Drawing on theories of socialization, social learning, symbolic interaction, anomie, control, phenomenology, routine activities, and deterrence, Rothe and Mullins (2009) suggested that four forces operate at each the international, macro, meso, and micro levels of analysis: motivation, opportunity, constraints, and controls. Routine activities theory presumes motivation simply exists, but Rothe and Mullins (2009) believed that it is essential to explain motivation when trying to understand international crimes. In other words, some motivation is needed for the state, the organization, and the individual actors to engage in such crimes. These motivations may include political power, economic or resource gain, revenge, or some other variable. Beyond motivation, there must be an opportunity to commit a crime. Motivated offenders cannot commit a crime unless they encounter an opportunity to do so. For an act as complex as genocide or some other international crime to occur, however, the illegitimate opportunity must extend beyond the individual to other levels. This would include things like a vulnerable population, international relations, global economics, military supremacy, diffusion of responsibility, and propaganda. Rothe and Mullins (2009) then added a crucial concept that is not considered by other theorists: the role of social control forces (i.e., capable guardians). According to them, once a motivated offender is presented with an opportunity, an international crime becomes likely in the absence of constraints and controls. According to Rothe and Mullins (2009), a constraint is an informal inhibitor – something that makes a criminal act more difficult. A control, on the other hand, either completely prevents an act or ensures that the act will be punished formally. Because they presume that punishments are more likely to deter potential international law violators than street criminals, the threat of punishments should be effective at preventing such crimes.

As shown in **Figure 3**, Rothe and Mullins used their ideas to develop a causal model in which "a variety of independent factors [come] together to influence, as a combined force, the manifestation of" an international crime (2009: p. 110). According to their model, although motivational and opportunity forces exist at all four levels of analysis, they all function through motivations and opportunities they create within individuals. Therefore, they theorize that opportunity at the international, meso, and macro levels combine to create micro-level opportunity. This micro-level opportunity then directly can influence the decision to offend and can produce micro-level motivation. In addition to the influence of micro-level opportunity, micro-level motivation is influenced by motivation at the international, macro, and meso levels. As with opportunity, micro-level motivation directly influences the decision to offend but not the actual commission of the act. Constraints and controls, however, exist outside of individuals. Therefore, controls and constraints at all four levels of analysis combine to create a constraint node and a control node, each of which directly affects the decision to offend. The decision to offend, the constraint node, and the control node then directly affect the actual commission of the crime. This theory has the advantage of suggesting clear causal connections among the various variables in the model; however, Rothe and Mullins (2009) provided no data to support their theory.

Conclusion: Connecting the Dots

Although these theories each provide slightly different conceptions about what causes genocide, there is substantial overlap regarding dehumanization, threat, vulnerability, and leadership. The clearest parallel is the presence of a dehumanization campaign. As Chalk and Jonassohn observed, "We have no evidence that a genocide was ever performed on a group of equals. The victims must not only not be equals, but also clearly defined as something less than fully human" (1990, pp. 27–28). Several other scholars have pointed out this reality, and each of the theories reviewed here identifies something similar. Brannigan (1998) identified xenophobia, Neubacher (2006) focused on techniques of neutralization, Winton and Unlu (2008) discussed belief systems and enmeshed cohesion, Hagan and Rymond-Richmond (2009) described the importance of supremacist ideology and racial epithets, Maier-Katkin et al. (2009) detailed the process of primary group socialization and the role of ideology, and Rothe and Mullins (2009) explained the importance of absent internal constraints. Each one of these concepts describes the important role of belief systems that promote strong in-group identification and dehumanize other social groups. Even the political science literature suggests that exclusionary or supremacy ideologies are central to comprehending genocide (Straus, 2012).

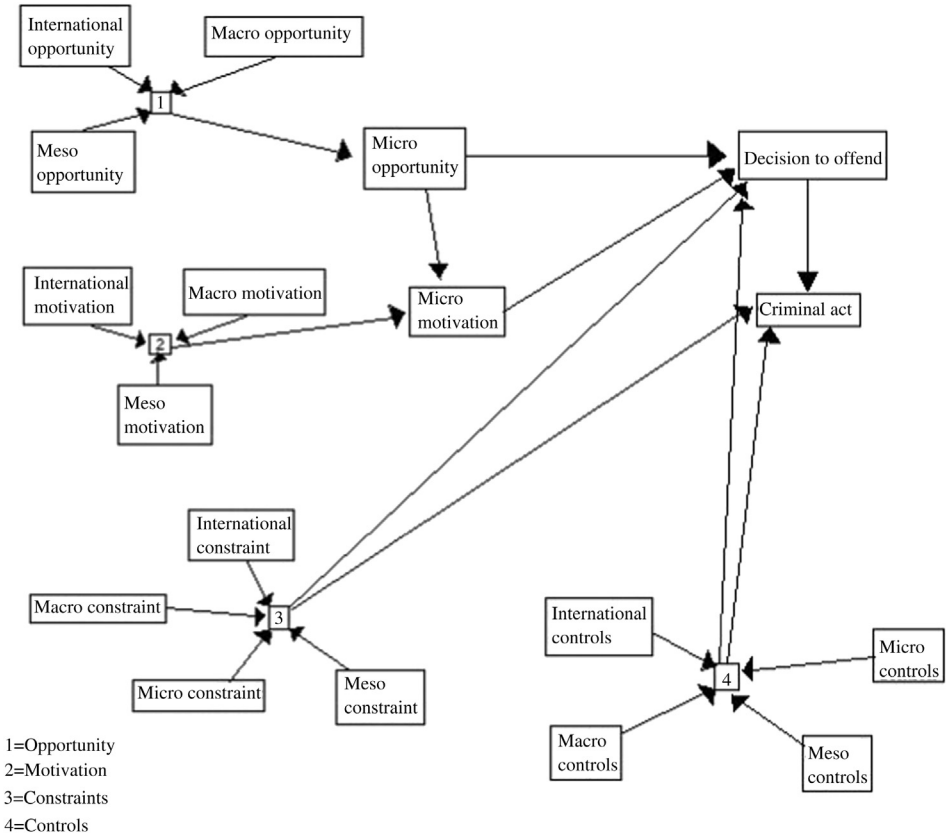

Figure 3 Rothe and Mullins' (2009) causal model of genocide. Recreated by author. Used with permission from Taylor & Francis.

In addition to having a dehumanized victim group, there must be some motivational force. A common thread among each of these theories is the motivational power of perceived threats to a dominant group position. The political science literature discusses the importance of the threat and insecurity that occurs during war (Midlarsky, 2005; Straus, 2012). According to Winton and Unlu (2008), some crisis is needed to initiate the violentization process, and in the defiance stage, individual perpetrators learn that the victim group is threatening and that they must kill them or be killed themselves. Hagan and Rymond-Richmond (2009) described the importance of resource competition, and Maier-Katkin et al. (2009) mentioned that societal-level strain and insecurity are the precursors to crimes against humanity. Lastly, Rothe and Mullins (2009) suggested that, among other factors, genocidal motivation can include political, economic, and ideological goals; anomie; and rewards. Despite their differences, each of these theories is describing a perceived threat (economic, social, or political) to the dominant group's position or survival that can be perceived at several levels of analysis. At the state or macro levels, political scientists suggest that leaders are likely to target groups perceived as threatening if they hold ideological visions of racial purity. At the meso and micro levels, this same process also may be at play. At the group level, Blalock (1967) hypothesized that intergroup violence against a minority group is likely to increase when the dominant group experiences either a political or an economic threat, and at the individual level, Blumer (1958) suggested that racial prejudice emerges out of a sense of relative group position and is more likely to appear when members of the dominant group feel their position of superiority threatened by a subordinate group.

Furthermore, vulnerability and leadership are important ingredients. Opportunity is a key component to any crime, and if there is no vulnerable group against whom aggression can be directed or that can be effectively scapegoated, a crime against humanity cannot occur. Yet, because genocide is an extreme form of group-directed and group-perpetrated violence, leadership matters. As Hagan and Rymond-Richmond (2009) described, a genocidal state needs to focus and concentrate individual and collective racial intent into genocidal violence and usually is needed to propagate the dehumanizing or supremacist ideology. Similarly, Rothe and Mullins (2009) mentioned the importance of authoritarian pressures at the meso level and of propaganda and the structure of government at the macro level. The fourth proposition of Maier-Katkin et al.'s (2009) theory talks about authority figures being important for directing the activities of primary groups, and Winton and Unlu (2008) implied that someone needs to brutalize potential genocidal actors and teach them the supporting belief system. As such, when theorizing about genocide, in addition to considering the social and economic factors of a society, one needs to consider the relative power and vulnerability of different groups as well as the structure and motives of the authority figures within a society. This is

why an interdisciplinary approach that incorporates the theories of political science is important. Political scientists already have focused attention on the role of those in power; by adding the criminological focus on social and economic conditions, individual actors, and differences across different sociodemographic groups, a more complete understanding of genocide can be achieved.

One final point to consider is Rothe and Mullins' (2009) important insight about the role of constraints and controls, which have the capacity to prevent genocide even if all the other necessary forces are present. In other words, the existing literature suggests that when a dominant group perceives that a vulnerable and dehumanized group is threatening its position of dominance, it is likely to engage in genocidal violence when their insecurities can be concentrated and directed against the vulnerable group by those in power and when there is an absence of formal and informal control mechanisms. As promising as these theories appear to be for understanding genocide and its causes, they never have been fully tested. Therefore, it is imperative that the variables suggested in these theories – dehumanization, threat, vulnerability, leadership, constraints, and controls – be operationalized and subjected to empirical testing to determine which social and economic forces actually influence the genocidal process, what their relative influence is, and what factors can be addressed to prevent genocide.

See also: Collective Efficacy and Crime; Control Theories and Crime; Deterrence Theory: Crime; Deterrence: Sanction Perceptions; Hate Crimes; Homicide; International and Comparative Criminology; Race: Prespectives and Crime; Routine Activities, Criminal Opportunities, Crime and Crime Prevention; Social Learning Theory and Crime; Strain Theories and Crime; Transnational Organized Crime.

Bibliography

Alvarez, A., 2001. Governments, Citizens, and Genocide: A Comparative and Interdisciplinary Approach. Indiana University Press, Bloomington, IN.

Athens, L., 2003. Violentization in larger social context. In: Athens, L., Ulmer, J.T. (Eds.), Violent Acts and Violentization: Assessing, Applying, and Developing Lonnie Athens' Theories. Elsevier Science, Boston, pp. 1–41.

Blalock Jr., H.M., 1967. Toward a Theory of Minority-Group Relations. John Wiley & Sons, New York.

Blumer, H., 1958. Race prejudice as a sense of group position. The Pacific Sociological Review 1, 3–7.

Brannigan, A., 1998. Criminology and the Holocaust: xenophobia, evolution, and genocide. Crime and Delinquency 44, 257–276.

Browning, C., 1992. Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland. Harper Collins, New York.

Chalk, F., Jonassohn, K., 1990. The History and Sociology of Genocide: Analyses and Case Studies. Yale University Press, New Haven, CT.

Goldhagen, D.J., 1996. Hitler's Willing Executioners: Ordinary Germans and the Holocaust. Alfred A. Knopf, New York.

Goldhagen, D.J., 2009. Worse than War: Genocide, Eliminationism, and the Ongoing Assault on Humanity. Public Affairs, New York.

Hagan, J., Rymond-Richmond, W., 2009. Darfur and the Crime of Genocide. Cambridge University Press, New York.

Hagan, J., Rymond-Richmond, W., Parker, P., 2005. The criminology of genocide: the death and rape of Darfur. Criminology 43, 525–562.

Harff, B., 2003. No lessons learned from the Holocaust? Assessing risks of genocide and political mass murder since 1955. American Political Science Review 97, 57–73.

Kelman, H.C., 1973. Violence without moral restraint: reflections on the dehumanization of victims and victimizers. Journal of Social Issues 29, 25–61.

Maier-Katkin, D., Mears, D.P., Bernard, T.J., 2009. Towards a criminology of crimes against humanity. Theoretical Criminology 13, 227–255.

Midlarsky, M., 2005. The Killing Trap: Genocide in the Twentieth Century. Cambridge University Press, New York.

Milgram, S., 1963. Behavioral study of obedience. Journal of Abnormal and Social Psychology 67, 371–378.

Neubacher, F., 2006. How can it happen that horrendous state crimes are perpetrated? An overview of criminological theories. Journal of International Criminal Justice 4, 787–799.

Olson, D.H., DeFrain, J., 1997. Marriage and the Family: Diversity and Strengths, second ed. Mayfield, Mountain View, CA.

Rothe, D.L., Mullins, C.W., 2009. Toward a criminology of international criminal law: an integrated theory of international criminal violations. International Journal of Comparative and Applied Criminal Justice 33, 98–118.

Semelin, J., 2007. Purify and Destroy: The Political Uses of Massacre and Genocide (C. Shoch, Trans.). Columbia University Press, New York.

Straus, S., 2006. The Order of Genocide: Race, Power, and War in Rwanda. Cornell University Press, Ithaca, NY.

Straus, S., 2012. "Destroy them to save us": theories of genocide and the logics of political violence. Terrorism and Political Violence 24, 544–560.

Sutherland, E., 1947. Principles of Criminology. Lippincott, New York.

Sykes, G.M., Matza, D., 1957. Techniques of neutralization: a theory of delinquency. American Sociological Review 22, 664–670.

Ulfelder, J., Valentino, B., 2008. Assessing Risks of State-Sponsored Mass Killing. Political Instability Task Force, Washington, DC.

Weitz, E., 2003. A Century of Genocide: Utopias of Race and Nation. Princeton University Press, Princeton, NJ.

Winton, M.A., 2005. A circumplex model of genocide. In: Salfati, C.G. (Ed.), Homicide Research: Past, Present, and Future: Proceedings of the 2005 Meeting of the Homicide Research Working Group. HRWG, Chicago, pp. 47–56.

Winton, M.A., 2011. Violentization theory and genocide. Homicide Studies 15, 363–381.

Winton, M.A., Unlu, A., 2008. Micro–macro dimensions of the Bosnian genocides: the circumplex model and violentization theory. Aggression and Violent Behavior 13, 45–59.

Relevant Websites

http://www.amnesty.org/ – Amnesty International.
http://www.chgs.umn.edu/ – Center for Holocaust & Genocide Studies.
http://www.yale.edu/gsp/index.html – Genocide Studies Program.
http://www.genocidewatch.org/ – Genocide Watch.
http://www.hrw.org/ – Human Rights Watch.
http://www.instituteforthestudyofgenocide.org/index.html – Institute for the Study of Genocide.
http://migs.concordia.ca/ – Montreal Institute for Genocide and Human Rights Studies.
http://savedarfur.org/ – Save Darfur.
http://endgenocide.org/ – United to End Genocide.

Genographic Project

Theodore G Schurr, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Genographic Project is a major international effort to assemble genetic data from populations located around the world and use them to reconstruct the migratory history of our species. A consortium of scientists based at 10 regional centers worked with indigenous human populations from various parts of the world to investigate their genetic history, while members of the general public learned about their genetic ancestry through having their own DNAs analyzed for sequence variation. These studies have expanded our understanding of human migration patterns around the globe, the dynamics of human settlement through time, the influence of religion, culture, and language on population genetic diversity, and the evolution of human mitochondrial and Y-chromosome phylogenies. Counting both indigenous and public participants, some 675 000 people have participated in the project, making it the largest anthropological genetics study ever conducted. In addition, the Legacy Fund has supported numerous educational and linguistic initiatives of indigenous communities, thereby helping them sustain their cultural and linguistic heritage.

Inception of the Project

The Genographic Project was conceived as an initiative for addressing fundamental questions concerning our origin as a species and dispersal around the world. While certain aspects of human evolution and migration had been defined by the time that this project began, such as the fact that modern humans speciated in Africa (e.g., Cann et al., 1987; Quintana-Murci et al., 1999; Ingman et al., 2000), many other details about the migrations that led to the settlement of other continents of the world remained unclear (e.g., Forster, 2004; Metspalu et al., 2004; Schurr, 2004). For anatomically modern humans, the major dispersals across the globe took place in the latter stages of the Paleolithic period, between 75 000–15 000 years ago.) In addition, relatively little was known from a genetic perspective about more recent events in human evolutionary history, such as those occurring over the past 10 000 years (e.g., Wells et al., 2001; Salas et al., 2002; Richards, 2003) or even more recently (within the last 1000 years) (Gresham et al., 2001; Zerjal et al., 2003; Yao et al., 2004). Thus, while anthropological research (linguistic, archeological, ethnographic, biogenetic) had helped to clarify some aspects of these questions (e.g., Polynesia: Bellwood, 1996; Oppenheimer and Richards, 2001), in other respects, geneticists had really only just begun to elucidate them.

The Genographic Project was officially launched in April 2005 at the National Geographic Society headquarters in Washington, D.C. Conceived and led by Dr. Spencer Wells, an Explorer-in-Residence at the National Geographic Society, the project was designed to substantially expand our knowledge of human ancestry, migration, and history and expected to run over the course of 5 years. To achieve this goal, a set of anthropological and human geneticists was assembled and tasked with undertaking project-related research in different areas of the world. These researchers were to conduct field research and collect DNA samples for genetic analysis that would allow them to address outstanding questions important to the regions in which they were working. By employing state-of-the-art technology for genetic analyses, project researchers would also increase the resolution of the genetic data being used to explore human migratory history. Moreover, by generating a large data set based on genetic markers analyzed in tens of thousands of individuals representing global population diversity, these data could be used to delineate large scale as well as more localized patterns of genetic diversity in different regions of the world.

Once the project was initiated, Genographic researchers began working with numerous indigenous populations living in traditional contexts to explore their genetic histories. Here, the term 'indigenous populations' refers to groups living in their natal territories or homelands and following traditional cultural practices, as described above. Hence, indigenous populations can be found in the Americas (Native Americans) and Australia (Aboriginal Australians), as well as Sardinia, Laos, Lebanon, and South Africa, among other places. These populations were targeted because they represented communities that have lived in their respective locations or territories for many generations, preserved many aspects of their traditional life ways and cultures, and remained relatively unadmixed with populations from other regions of the world. Due to increasing globalization and contact between peoples of different backgrounds, many indigenous communities are undergoing significant changes, with some of their members moving to urban areas and becoming integrated into the broader nonindigenous cultures. Because of these rapid cultural and demographic processes, the geographic and cultural context for the genetic diversity that these individuals bear could potentially be lost or obscured. Thus, the project was attempting to capture a genetic 'snapshot' of the history of our species before processes of modernization erased or blurred some of the patterns of genetic variation that reflected the ancient migratory history of our species.

At the end of the 5-year period, the Genographic Project planned to have created a public-domain anthropological genetic database that could be used by other scientists and the participants in the study to learn more about human genetic diversity. This database would also serve as a resource for educating the general public about the history of our species. An explicit goal of the project was sharing the benefit of

knowledge acquired during the course of the research efforts with participating indigenous groups and individuals.

While ambitious in its scope, the Genographic Project wanted to avoid some of the pitfalls of the Human Genome Diversity Project (HGDP) (NRC, 1997), an earlier attempt to survey global human genetic diversity (Weiss et al., 2002). Accordingly, the Genographic Project created a research framework that was informed by concerns raised over the HGDP. These included putting measures in place to protect the privacy of participants and maintain the anonymity of samples obtained for genetic analysis. Participants were informed that they owned their DNA samples, had full control over the use of them for this research project, and could withdraw from the project at any time during their involvement in it. The project also recognized that informed consent could involve both communal and/or individual consent, depending on the populations being consulted. To obviate concerns about biopiracy and exploitation of biological specimens, the project also explicitly stated that its researchers would not use DNA samples for medical testing or to generate cell lines for use in future human genetic studies. All DNA samples would only be used for the scientific objectives of the project, which were to map the migratory history of humankind over the past 200 000 years.

Due to the complexities of undertaking a global research study, the first phase of the Genographic Project went beyond the originally projected 5 years and continued through 2012. In the second phase of the project (Geno 2.0), which began in that year, researchers began analyzing additional DNA markers to gain a broader view of genetic variation in different human populations. This effort was stimulated by the creation of the GenoChip, an Illumina microarray designed to interrogate human genomes for autosomal, mitochondrial, and Y-chromosome variation (see below). The mitochondrial DNA (mtDNA) is passed down each generation from mothers to daughters, hence, represents maternal lineages in human populations. Similarly, the Y-chromosome is passed down each generation from fathers to sons, hence, represents paternal lineages in human groups. By contrast, autosomal DNA is passed down to children equally from each parent, hence, the markers or genes in the autosomes (chromosomes) are biparental in their inheritance. They reflect the contributions of numerous ancestors from both the mother and father's sides of the family. An expanded analysis of autosomal variation in Genographic populations is now in progress.

Overview of Project Objectives and Components

Objectives

The primary objective of the Genographic Project is to generate a global perspective on human genetic history by establishing and analyzing the world's potentially largest public database of genetic markers devoted to the study of human origins and migration. In collaboration with indigenous communities, the project intends to raise awareness in the global community about the threat to indigenous cultures and populations through the dissemination of unique narratives describing their histories. These efforts have also helped to build scientific capacity in developing countries through collaborations with local laboratories, scientific organizations, and researchers. Furthermore, the project has functioned as a unique educational platform that has enlarged public knowledge about genetics and human history, and also showed individuals in the global community how they are related to other people from around the world through their participation in the project.

Project Components

To achieve these goals, the Genographic Project initiated three interrelated components of the project, namely, Research, Education (Public Participation), and the Legacy Fund. These components are described in some detail below. Because the Genographic Project has been conducted in two phases, these components have been pursued to somewhat different extents over the 9 years that the project has been underway. For example, in Phase I, Genographic researchers collected the bulk of the DNA samples to be used for genetic analysis through field expeditions in various parts of the world. In Phase II, more limited fieldwork and sample collection has taken place, with most project work being focused on publishing research findings generated from and analyzing autosomal variation in DNA samples collected during Phase I. Each regional center initiated its part of Genographic Project research based on the dates at which its Principal Investigators (PIs) completed financial contractual negotiations with National Geographic Society, obtained research permissions in the countries in which the regional team wanted to work, and assembled the research team needed to undertake project research. Thus, where appropriate, further distinctions in the timing of work on different project components will be made in the text.

Research

The Research Component has been the primary focus of this project. Its main objective was to collect oral and family history data together with biological samples from people from different geographical regions of the world to help reconstruct the history of human migrations and interactions over the past 200 000 years. During Phase I, laboratory analysis was directed toward analyzing genetic markers in the mtDNA and Y-chromosome (NRY) that are useful for inferring genealogical relationships and discerning these historical migration patterns. Here, 'NRY' means the 'nonrecombining portion of the Y-chromosome,' the region in which most of the anthropologically informative genetic markers are found. No markers of known medical relevance were investigated.

A total of 10 research centers were established to conduct this research, each with its own PIs to direct this research. They included centers in North America (Theodore Schurr), South America (Fabricio Santos), West/Central Europe (Lluis Quintana-Murci, David Comas, Jaume Bertranpetit), Middle East/North Africa (Pierre Zalloua), Sub-Saharan Africa (Himla Soodyall), North Eurasia (Oleg Balanovsky, Elena Balanovskaya), Australasia/Pacific (John Mitchell, Elizabeth Matisoo-Smith), East/Southeast Asia (Li Jin), South Asia (Ramasamy Pitchappan), and Ancient DNA (Alan Cooper). The PIs assembled research teams of technicians, postdoctoral researchers, graduate students, and undergraduate students to conduct field and laboratory research in these regions. The

project also involved two Associate Researchers who served as technical and research advisors, these being Chris Tyler-Smith (The Wellcome Trust Sanger Institute) and Doron Behar (Family Tree DNA). Each of the 10 regional centers followed its own localized sampling and research agenda, with the overarching goal of working with as wide a range of populations as possible. During the course of Phase I, each center endeavored to enroll approximately 10 000 individuals in the project, for a total of 100 000 participants.

Education (Public Participation)

The Education, or Public Participation, component of the project invited individuals from around the world to participate in the real-time anthropological study that is, the Genographic Project. Public Participation kits were sold through the Genographic Project Web site. Participants used the buccal (cheek) swabs in the kits to obtain DNA samples for genetic analysis, and then sent the swabs to Family Tree DNA, which, in collaboration with the Arizona Research Labs at the University of Arizona, extracted and analyzed the DNA samples. All samples were analyzed for the basic set of genetic markers used for the Research Component of the project. Based on the resulting data, members of the general public were able to trace their own genetic ancestry and also donate their DNA test results to the project database for use in research studies.

Another important part of this component was the interactive Web site for mapping and presenting the results of Genographic Project research to the general public (https://genographic.nationalgeographic.com/). On the Web site, viewers could find details about the project, such as maps tracing human migration routes, descriptions of the genetic methods used in the project, updates on project related studies, and profiles of project researchers. Through this Web site, public participants could also access their own DNA test results, which include both their basic genetic data and a corresponding description of their ancestors' migratory and cultural journeys as inferred from these results, thereby promoting their understanding of the objectives and results of Genographic Project research.

The Legacy Fund

The Genographic Legacy Fund (GLF), funded by a portion of the net proceeds from the sales of the Genographic Public Participation kits, links the scientific side of the project and public participants to the indigenous communities. The GLF was also intended to reflect the values and missions of the Genographic Project funding partners, as well as raise awareness of the challenges facing indigenous groups while building community capacity. National Geographic Society's mission is to increase and diffuse geographic knowledge while promoting conservation of the world's cultural and natural resources. IBM is committed to providing diverse and sustained programs that support education, workforce development, arts, and culture, and communities in need on a global basis. The Waitt Family Foundation is dedicated to supporting projects that seek a deeper understanding of human history, expand the potential of the human mind and improve humankind's knowledge through historical and scientific exploration. The Genographic Project's Advisory Board, comprised of leaders from various scientific and other fields – including indigenous advocacy – provided oversight and guidance to foster meaningful community involvement and representation in the decision-making process surrounding the operation of the GLF. Overall, the GLF was designed to support indigenous communities and build capacity by funding relevant programs and projects initiated by these groups that would be sustainable beyond the established time frame of the Genographic Project.

Phase I

Research Component

Community Outreach and Engagement

To undertake this research, Regional Center PIs established relationships with individuals of a given community (e.g., tribal elders, village council members, members of a particular indigenous group, elected leaders, etc.). These relationships were developed through several means, including trips to the communities by project members to discuss the details of the project with them, and contacting anthropologists, archaeologists, and colleagues from medical fields who had previously worked, or are currently working, with these communities to initiate a dialogue with their leaders about the group's possible involvement in the project. Each center also developed collaborations with scholars working in different countries within the region for which the PI was responsible for working, with these individuals participating in sample collection and analysis.

Sample and Data Collection

Members of the Genographic scientific team chose populations to approach for participation in the project on the basis of different scientific or historical criteria, i.e., because their inclusion would help to answer specific questions about human migration history that were relevant to specific regions of the world. To work with indigenous populations, the PIs and their team, along with collaborators, traveled to the locations where these communities reside. As noted above, individuals from indigenous groups were chosen because they were more likely to have not experienced extensive recent admixture with other population groups, hence, to preserve a crucial record of their own and their communities' biological history and genetic ancestry.

Whenever possible, unrelated adult men and women were enrolled in the study. Researchers confirmed that individuals were genetically unrelated to each other through at least two generations through interviews, family history information, written genealogies, tribal and village information, and written documents in local and archival records. Such data were also invaluable for interpreting patterns of genetic variation within a population, as they provided insights into marriage patterns, places of residence, and family relocations.

Biological samples used for genetic analysis included buccal swabs, mouthwash, and blood samples, depending on the community involved in the research. The collection method was agreed with the participating indigenous groups and individuals during the consultation process. Upon their arrival at the regional center laboratory, researchers extracted DNA from the samples for genetic analysis. For ethical reasons, no transformed cell lines were established from any of the blood

samples collected in the course of the Genographic Project. In addition, no medically relevant information was collected from either the DNA samples or the participants themselves during interviews.

Human Subjects

The research teams followed local laws concerning genetic sampling and anthropological fieldwork, and obtained the requisite research permissions to work in each location. Project researchers obtained written or oral informed consent from participants, as specified in the permission requirements for each region, and asked them questions about their ancestry and duration of residence in the region. These included details about their family histories and genealogies, the language(s) that they spoke, and the oral histories of their communities. Following the interview, researchers then collected biological samples for DNA analysis.

Genographic Project IDs

All persons who provided samples were assigned a randomly generated alphanumeric code called a Genographic Project ID (GPID). Participants were able to use their GPIDs to retrieve their DNA test results and accompanying explanations from the secure public Genographic Web site. If individuals did not have access to the Internet and requested information about their results, researchers mailed the DNA test results to them or otherwise communicated these findings to the participants through an appropriate means.

Project Database

Raw genetic data were interpreted at the regional centers and prepared for database submission. The interpreted data were submitted to the central Genographic Database and Repository (DAR), which is being hosted at National Geographic Society headquarters in Washington, D.C. No identifying information obtained during the sample collection process was entered into the DAR. No one outside of the individual regional centers had access to the GPIDs such that the names and genotype data from participants could be connected. In addition, all family history information was securely maintained by the regional PIs and their team, and used primarily to interpret the genetic data. Such information was also very helpful in reporting the results of the genetic analyses to participants in the project, which, in turn, facilitated dialogue between the participants and the PIs about the nature and interpretation of their DNA data.

Genetic Data Reports for Participants

In addition to providing GPIDs to participants for accessing their DNA test results online, some regional centers produced a set of explanatory documents to send to all of their participants. These documents provided additional details to help participants understand the DNA test results. As such, they served as tools for educating participants about genetics and helping them to better understand aspects of their biological ancestry, complementing information accessible through the Genographic Web site. Genographic researchers also communicated project results to indigenous participants through return visits to their communities to present the genetic findings, preparing summary documents for tribal authorities, and having conversations with them via telephone and e-mail.

Laboratory Analyses

As noted above, laboratory analysis of the samples focused on analyzing genetic markers useful for inferring genealogical relationships, namely those found on the paternally inherited Y-chromosome and maternally inherited mtDNA. For the Y-chromosome, the markers included an array of short tandem repeats (STR) and single nucleotide polymorphism (SNPs) drawn from a more extensive list of Y-chromosome STR and SNP markers described in the literature (e.g., Kayser et al., 2001; Underhill et al., 2001; Butler et al., 2002; Jobling and Tyler-Smith, 2003; Y-Chromosome Consortium, 2002; Karafet et al., 2008). In addition, mtDNA sequences from the non-coding hypervariable segments I and II and lineage-defining SNPs located in the coding regions of the mitochondrial genome were analyzed (Torroni et al., 2006). In some cases, researchers sequenced whole mitochondrial genomes to obtain the highest resolution maternal haplotypes possible (e.g., Kong et al., 2003; Palanichamy et al., 2004; van Oven and Kayser, 2009). Standardized laboratory protocols for analyzing Y-chromosomal and mtDNA were used with ABI 3130XL Gene Analyzers and ABI 7900HT Real-Time PCR Machines based in regional center labs to characterize indigenous participant DNA.

Public Participation

For the Public Participation component of the project, subject recruitment occurred through information presented on the Internet, radio, television, and print media outlets. All members of the public were encouraged to participate, and the purchase of the kits was made through the project Web site, as explained above. Participants purchased swab kits and had their DNA samples analyzed, with the resulting data being sent from Family Tree DNA/Arizona Research Labs to the DAR. The data residing in this repository were accessible to participants at the project Web site through the use of their unique GPID identifiers. Upon viewing their DNA test results, public participants were asked if they wanted to include their data in the Genographic database. Those who decided to include their data in the DAR were also asked to complete a questionnaire about their family history, one very similar to that used with indigenous participants to assure the equivalency of demographic data collected from all participants in the project.

Research Support

The Genographic Project was financially supported by several partnering organizations, including the National Geographic Society, IBM, and the Waitt Family Foundation. During Phase I, the National Geographic Society provided support for laboratory analysis in the regional centers, IBM provided computational and software support, and the Waitt Family Foundation provided funds for field research in various parts of the world. Each PI also obtained additional funding through intramural and extramural sources to help carry out various aspects of project-related research.

Genographic Project Results: Phase I

Research Component

After 7 years of focused research, Genographic Project scientists have made a number of important contributions to our understanding of human migratory history and population dynamics. To attain these achievements, they worked with hundreds of indigenous communities and nearly 75 000 participants from around the world. Because many of these populations were not represented in existing sample panels (e.g., Altshuler et al., 2010; Durban et al., 2010; 1000 Genomes Project Consortium, 2010), the data obtained from their DNAs were novel and highly informative about the demographic and migratory history of the regions of the world in which they lived, as well as the specific histories of the participating communities themselves. Although page limits place constraints on a full description of the results of Genographic Project research to date, the following discussion summarizes many of the highlights of this work.

To begin with, Genographic research has elucidated the deeper branches of the human mtDNA phylogeny in Africa, particularly those found largely in contemporary Khoisan-speaking populations (Behar et al., 2008b). This study also suggested that the early settlement of humans in Africa was matrilineally structured and involved small, separately evolving isolated populations (Behar et al., 2008b). More recent work has revealed differentiation in the genetic affinities of Middle Eastern populations with African and European populations, with Y-chromosome data suggesting greater affinities with Africa and mtDNA data indicating stronger genetic affinities with Europe (Badro et al., 2013).

Related research with populations from the Near East has further defined the influences of history, geography, and religion on population genetic structure in Lebanon. Y-chromosome analysis indicated that Islamic expansion from the Arabian Peninsula in the CE seventh century introduced lineages typical of this area into populations that subsequently became Lebanese Muslims, whereas Crusader activity in the eleventh–thirteenth centuries CE introduced Western European lineages into Lebanese Christians (Zalloua et al., 2008a). Similar work also defined the genetic influences of Phoenicians around the circum-Mediterranean region during the flourishing of their trade empire (Zalloua et al., 2008b).

Y-chromosome studies have also illuminated the history of populations from other regions of the world. For example, in Afghanistan, an analysis of paternal lineages suggested that interpopulational differences emerged during the Bronze Age after the formation of the first civilizations in the region, with later migrations and invasion reshaping them, giving Afghans a unique pattern of genetic diversity in Central Asia (Haber et al., 2012). Related work showed the parallel evolution of genes and languages in the North Caucasus resulting from high levels of isolation, differentiation, and genetic drift *in situ*, due in part to the mountainous terrain of the region (Balanovsky et al., 2011). Likewise, Y-chromosome analysis demonstrated that population differentiation of male lineages in Southern India (Tamil Nadu) correlated with agricultural expansions that predated the formation of the caste system (Arunkumar et al., 2012).

Genographic research has provided fine-grained details about the population history of specific regions of the world. In Europe, ancient DNA (aDNA) analyses confirmed prehistoric gene flow from Siberia as part of the complex human population history of Northeast Europe (Der Sarkassian et al., 2013, 2014), and revealed that the formation of Central European mtDNA diversity occurred in several different stages (Brandt et al., 2013). A related aDNA study of the Linear Pottery Culture in Germany further demonstrated that this archeological population had affinities with groups from modern-day Near East and Anatolia, supporting a major genetic input from this area during the advent of farming in Europe (Haak et al., 2010). From a modern DNA perspective, the analysis of uniparental genomes revealed evidence of pre-Roman tribal genetic structure in Basques, who are otherwise linguistically distinct from their Indo-European speaking neighbors (Spanish and French) (Martínez-Cruz et al., 2012a), as well as evidence for partial genetic continuity of contemporary Basques with preceding Paleolithic/Mesolithic settlers of their homeland (Behar et al., 2012). In addition, Y-chromosome analysis of individuals bearing the Basarab name was used to examine the genealogy of the first dynasty of Wallachian kings in Romania, which included Vlad III the Impaler, also known as Dracula (Martínez-Cruz et al., 2012b).

Genetic data from Genographic studies have also contributed to our understanding of East Asian prehistory. The analysis of Y-chromosome lineage O3 diversity in Sino-Tibetan populations revealed that human groups entered the eastern Himalayas through two different routes (Kang et al., 2012). An analysis of mtDNA variation in Tibetans also indicated both pre-last glacial maximum (LGM) and post-LGM expansions into the region, with greater diversity in the southern part of the Tibetan Plateau suggested that it was a population refugium during the LGM (Qin et al., 2010). A Y-chromosome analysis of Mon-Khmer and Hmong-Mien-speaking populations also showed that human groups expanded from Southeast Asia into East Asia through a series of bottlenecks during the LGM, based on the distributions of lineages O3a3b-M7 and O3a3c1-M117 in the region (Cai et al., 2011).

Genographic research has made important advances in our understanding of the people of the Americas. A recent Y-chromosome analysis revealed genetic divergence and new founding native lineages in Athapaskan and Eskimoan-speaking populations of northern North America, with these data providing new insights into the human settlement of the Arctic (Dulik et al., 2012). Similarly, an analysis of mtDNA variation showed that clan, language, and migration history have shaped genetic diversity in Haida and Tlingit populations from Southeast Alaska (Schurr et al., 2012). In South America, researchers explored the genetic history of Uros populations of the Peruvian and Bolivian Altiplano, and found that they likely derived from ancient Andean lineages that were partially replaced during more recent farming expansion events and the establishment of complex civilizations in the Andes (Sandoval et al., 2013).

From a methodological standpoint, Genographic research has produced a number of new approaches for analyzing

genetic data. These included tracing patterns of chromosomal recombination to estimate the effective population size and the history of Old World human populations (Melé et al., 2012), using these same recombinational events to refine methods of producing gene trees (Parida et al., 2008, 2009), and taking knowledge about rates of mutation in DNA sequences to better identify the lineages to which they belong (Rosset et al., 2008). The effective population size (N_e) of a human group is the number of reproductively active individuals in it. Although N_e can be influenced by genetic drift and endogamy, it is usually less than the census size of the group. Project researchers have also created a new multiplex TaqMan genotyping assay for distinguishing most Y-chromosome lineages and sublineages in Europe (Martínez-Cruz et al., 2011), found new Y-chromosome SNPs in lineage Q through resequencing work (Jota et al., 2011), and developed an 'A to Z' protocol for obtaining complete human mitochondrial (mtDNA) genomes (Clarke et al., 2014).

An extended list of Genographic Project publications is provided in the Literature Cited, with a number of other manuscripts currently being in preparation or under evaluation at peer-reviewed journals. Collectively, they represent the enduring scientific legacy of this project.

Public Participation

Based on the enthusiastic response to this initiative, nearly 600 000 individuals from around the world have become involved in the Public Participation component of the Genographic Project. Their contributions to the DAR represent the largest database ever assembled for anthropological genetics research. In fact, an analysis of mtDNAs from approximately 80 000 individuals enrolled in the first 18 months of the study produced the largest standardized human mtDNA database ever collected at that time (Behar et al., 2007). This database was quite informative with respect to providing new details about mtDNA phylogeny and mutational dynamics, and also facilitated the development of a new nearest neighbor lineage prediction tool. Future work will utilize the abundance of mtDNA and Y-chromosome data obtained from public participants in similar fashion.

Legacy Fund

To date, the GLF has supported a number of initiatives by indigenous communities from around the world. These include the documentation of traditional Otomí stories in the Sierra of Xichú, a community located in Guanajuato State, Mexico; the recording of oral literature in Dangaura Tharu (an Indo-Aryan language) and in Limbu (a Tibeto-Burman language) in Nepal; the creation of a digital archive of the photographs, stories, and historical documents of the Cherbourg-Aboriginal settlement of Southeastern Queensland, Australia; and the revitalization of indigenous language and ecological knowledge on the Marquesas Islands in French Polynesia, among others. Additional details about the GLF can be found at the Genographic Project Web site: https://genographic.nationalgeographic.com/legacy-fund/.

Phase II: Geno 2.0

The second phase of the project, called Geno 2.0, began in late 2012. Its focus is the characterization of autosomal SNP variation in different human populations, as these data provide a more complete genetic history for individuals in encompassing contributions from many different ancestors. Based on recent studies, it is clear that genome-wide methods are becoming a crucial component of anthropological genetic studies (e.g., Li et al., 2008). This work is being partially supported by National Geographic through research grants awarded to Genographic Project researchers and other scientists interested in using the new GenoChip to conduct anthropological genetics studies, and supplemented by funds from other public and private organizations obtained by PIs from different regional centers.

More specifically, Genographic scientists and other investigators are analyzing autosomal SNP variation in human DNA samples (many collected during Phase I, some previously obtained in unrelated studies) with the novel GenoChip microarray, which is specifically designed for anthropological genetics research (Elhaik et al., 2013). This SNP array is an Illumina HD iSelect genotyping bead array with over 130 000 highly informative autosomal and X-chromosomal SNPs ascertained from over 450 worldwide populations analyzed in the 1000 Genomes Project (2010) and the HapMap Project (Altshuler et al., 2010; Durban et al., 2010) and other published studies (Rasmussen et al., 2010, 2011). To avoid any inadvertent medical testing, all SNPs known or suspected to have health or functional associations were filtered out from the GenoChip microarray, and SNPs occurring in coding regions were not selected.

The chip was validated by genotyping over 1000 samples from the 1000 Genomes Project, Family Tree DNA, and Genographic Projects (Elhaik et al., 2013), and served as a worldwide reference database for the project. Initial results indicate a 99.5% concordance between the GenoChip and the 1000 Genomes Project data sets, as well as SNP density of approximately (1/100 000) bases over 92% of the human genome. This array also yields extremely highly resolved maternal (mtDNA) and paternal (Y-chromosome) haplotypes. In addition, when compared to the results from Affymetrix and Illumina microarrays, the GenoChip showed a greater capacity for detecting population substructure (Elhaik et al., 2013). Furthermore, its SNPs overlap with 40–76% of those present in commercial microarrays, making the GenoChip data compatible with those generated with other human genotyping tools.

The GenoChip also contains autosomal SNPs found to be variable in archaic humans such as Neandertals (Green et al., 2006, 2010) and Denisovans (Reich et al., 2012) but which are also present in modern humans (Meyer et al., 2010). The presence of these markers in modern humans suggests their ancestors interbred to some extent with archaic humans as they expanded outside of Africa some 60 000 years ago. Thus, it is possible to estimate the relative percentage of Neandertal and Denisovan markers in participants' genomes by analyzing autosomal SNPs with the GenoChip, hence, determine how 'archaic' they really are.

In addition to this genotyping tool, project researchers have developed a new Geographic Population Structure

Figure 1 North American Regional Center researchers and members of the South Naknek Alutiiq community from Southeast Alaska enjoying a salmon cookout together. Photo by T.G. Schurr.

admixture-based algorithm to explore the ancestry of different human populations. The accuracy and power of this method to infer the biogeography of worldwide individuals down to their country or, in some cases, village, of origin, with autosomal SNP data has recently been demonstrated (Elhaik et al., 2014). These results affirmed the utility of admixture-based methods for reconstructing human biogeography, and raised interesting questions about how population size, genetic diversity, and environmental influences have shaped human population structure.

Public participants are also able to have their DNA analyzed with the GenoChip. Similar to Phase I genotyping, individuals can purchase a cheek swab kit to have their DNA analyzed on the GenoChip at Family Tree DNA (now ancestry by DNA). When the genotyping is completed, participants can access their results at the project Web site through the use of their unique GPID identifiers, as well as opt to include their data in the Genographic database. As a result of their involvement, public participants in this phase of the project will gain a broader understanding of their ancestry in learning what different parts of their genome can tell them about genetic genealogy and inheritance. The inclusion of their data in the project database will also enhance the accuracy of the algorithm used to make proportional ancestry estimates for both indigenous and public participant samples.

Summary

As evident from the preceding discussion, Genographic studies have added substantially to our knowledge of the origin and dispersal of our species. They have expanded our understanding of modern human migration patterns around the globe, the dynamics of human settlement through time, the influence of religion, culture, and language on population genetic diversity, and the evolution of human mitochondrial and Y-chromosome phylogenies. The success of the project has resulted in large part from the involvement of indigenous participants, whose keen interest in learning about their genetic ancestry and collaborative efforts have made this study a mutually enriching experience for researchers and participants alike (**Figure 1**). In addition, the Legacy Fund has supported numerous educational and linguistic initiatives of indigenous communities, thereby helping them sustain their cultural and linguistic heritage. Finally, the knowledge about our deep ancestry that has been contributed by this project demonstrates that we are part of one large, extended family with genetic ties that bind us together, despite our outward biological, cultural, and linguistic differences.

See also: Anthropology, Genomics, and Human Variation: National Roots; Biocitizenship; Bioethics in the Post-genomic Era; Bioethics: Genetics and Genomics; Direct-to-Consumer Personal Genetic Testing; Genetic Admixture; Genetic Ancestry Testing; Genetics and Anthropology; Genetics and Society; Genetics: The New Genetics; Genomics, Ethical Issues in; Human Genome Diversity Project: History; Human Genome Project: History and Assessment; Indigenous Knowledge: Science and Technology Studies; Microevolution; New Genetics and Race.

Bibliography

Adler, C.J., Haak, W., Donlon, D., Cooper, A., The Genographic Consortium, 2011. Survival and recovery of DNA from ancient teeth and bones. Journal of Archaeological Science 38 (5), 956–964.

Arunkumar, G., Soria-Hernanz, D.F., Kavitha, V.J., Arun, V.S., Syama, A., Ashokan, K.S., Gandhirajan, K.T., Vijayakumar, K., Narayanan, M., Jayalakshmi, M., Ziegle, J.S., Royuru, A.K., Parida, L., Wells, R.S., Renfrew, C., Schurr, T.G., Smith, C.T., Platt, D.S., Pitchappan, R., The Genographic Consortium, 2012. Population differentiation of southern Indian male lineages correlates with agricultural expansions predating the caste system. PLoS ONE 7 (11), e50269.

Altshuler, D.M., Gibbs, R.A., Peltonen, L., et al., 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467, 52–58.

Badro, D.A., Douaihy, B., Haber, M., Youhanna, S.C., Salloum, A., Ghassibe-Sabbagh, M., Johnsrud, B., Khazen, G., Matisoo-Smith, E., Soria-Hernanz, D.F., Wells, R.S., Tyler-Smith, C., Platt, D.E., Zalloua, P.A., The Genographic Consortium, 2013. Y-chromosome and mtDNA genetics reveal significant contrasts in affinities of modern Middle Eastern populations with European and African populations. PLoS ONE 8 (1), e54616.

Balanovsky, O., Dibirova, K., Dybo, A., Mudrak, O., Frolova, S., Pocheshkhova, E., Haber, M., Platt, D., Schurr, T.G., Haak, W., Kuznetsova, M., Radzhabov, M., Balaganskaya, O., Druzhinina, E., Zakharova, T., Koshel, S., Renfrew, C., Wells, R.S., Tyler-Smith, C., Balanovska, E., The Genographic Consortium, 2011. Parallel evolution of genes and languages in the caucasus. Molecular Biology and Evolution 28 (10), 2905–2920.

Behar, D.M., Blue-Smith, J., Soria Hernanz, D.F., Tzur, S., Hadid, Y., Bormans, C., Moen, A., Tyler-Smith, C., Quintana-Murci, L., Wells, R.S., The Genographic Consortium, 2008a. A novel 154-bp deletion in the human mitochondrial DNA control region in healthy individuals. Human Mutation 29 (12), 1387–1391.

Behar, D.M., Harmant, C., Manry, J., van Oven, M., Haak, W., Martinez-Cruz, B., Salaberria, J., Ovharcabal, B., Bauduer, F., Comas, D., Quintana-Murci, L., The Genographic Consortium, 2012. The Basque paradigm: genetic evidence of a maternal continuity in the Franco-Cantabrian region since pre-neolithic times. The American Journal of Human Genetics 90 (3), 486–493.

Behar, D.M., Rosset, S., Blue-Smith, J., Balanovsky, O., Tzur, S., Comas, D., Mitchell, R.J., Quintana-Murci, L., Tyler-Smith, C., Wells, R.S., The Genographic Consortium, 2007. The genographic project public participation mitochondrial DNA database. PLoS Genetics 3 (6), e104.

Behar, D.M., Villems, R., Soodyall, H., Blue-Smith, J., Perriera, L., Metspalu, E., Scozzari, R., Makkan, H., Comas, D., Bertranpetit, J., Quintana-Murci, L., Tyler-Smith, C., Wells, R.S., Rosset, S., The Genographic Consortium, 2008b. The dawn of matrilineal diversity. The American Journal of Human Genetics 82, 1130–1140.

Benn, Torres, J., Vilar, M.G., Torres, G., Gaieski, J.B., Bharath, Hernandez, R., Browne, Z.E., Stevenson, M., Waters, W., Schurr, T.G., The Genographic Consortium. Genetic diversity in Carib and Garifuna populations of the Lesser Antilles reveals significant indigenous ancestry and insights into Caribbean settlement history. PLoS ONE, submitted.

Boattini, A., Martinez-Cruz, B., Sarno, S., Harmant, C., Useli, A., Sanz, P., Yang-Yao, D., Manry, J., Ciani, G., Luiselli, D., Quintana-Murci, L., Comas, D., Pettener, D., The Genographic Consortium, 2013. Uniparental markers in Italy reveal a sex-biased genetic structure and different historical strata. PLoS ONE 8 (5), e65441.

Brandt, G., Haak, W., Adler, C.J., Roth, C., Szécsényi-Nagy, A., Karimnia, S., Möller-Rieker, S., Meller, H., Ganslmeier, R., Friederich, S., Dresely, V., Nicklisch, N., Pickrell, J.K., Sirocko, F., Reich, D., Cooper, A.C., Alt, K.W., The Genographic Consortium, 2013. Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity. Science 342 (6155), 257–261.

Brotherton, P., Haak, W., Templeton, J., Brandt, G., Soubrier, J., Adler, C.J., Richards, S.M., Der Sarkissian, C., Ganslmeier, R., Friederich, S., Dresely, V., van Oven, M., Kenyon, R., Van der Hoek, M., Korlach, J., Luong, K., Ho, S.Y.W., Quintana-Murci, L., Behar, D.M., Meller, H., Alt, K.W., Cooper, A., The Genographic Consortium, 2013. Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. Nature Communications 4, 1764.

Bellwood, P., 1996. The origins and spread of agriculture in the Indo-Pacific region: gradualism and diffusion or revolution and colonization? In: Harris, D.R. (Ed.), The Origin and Spread of Agriculture and Pastoralism in Eurasia. Smithsonian Institution Press, Washington, D.C, pp. 465–498.

Butler, J.M., Schoske, R., Vallone, P.M., et al., 2002. A novel multiplex for simultaneous amplification of 20 Y-chromosome STR markers. Forensic Science International 129, 10–24.

Cai, X., Qin, Z., Wen, B., Xu, S., Wang, Y., Lu, Y., Wei, L., Wang, C., Li, S., Huang, X., Jin, L., Li, H., The Genographic Consortium, 2011. Human migration through bottlenecks from Southeast Asia into East Asia during last glacial maximum revealed by Y-chromosomes. PLoS ONE 6, e24282.

Cann, R.L., Stoneking, M., Wilson, A.C., 1987. Mitochondrial DNA and human evolution. Nature 325, 31–36.

Clarke, A.C., Prost, S., Stanton, J.L., White, W.T., Kaplan, M., Matisoo-Smith, E.A., The Genographic Consortium, 2014. From cheek swabs to consensus sequences: an A to Z protocol for high-throughput DNA sequencing of complete human mitochondrial genomes. BMC Genomics 15, 68. http://dx.doi.org/10.1186/1471-2164-15-68.

Committee on Human Genome Diversity, National Research Council, 1997. Evaluating Human Genetic Diversity. US National Academy of Sciences, Washington, D.C.

Deng, Q.-Y., Wang, C.-C., Wang, X.-Q., Wang, L.-X., Wang, Z.-Y., Wu, W.-J., Li, H., The Genographic Consortium, 2013. Genetic affinity between the Kam-Sui speaking Chadong and Mulam people. Journal of Systematics and Evolution 51 (3), 263–270.

Dulik, M.C., Owings, A.C., Gaieski, J.B., Vilar, M.G., Andre, A., Lennie, C., Mackenzie, M.A., Kritsch, I., Snowshoe, S., Wright, R., Martin, J., Gibson, N., Andrews, T.D., Schurr, T.G., The Genographic Consortium, 2012. Y-chromosome analysis reveals genetic divergence and new founding native lineages in Athapaskan and Eskimoan-speaking populations. Proceedings of the National Academy of Sciences of the United States of America 109 (22), 8471–8476.

Durbin, R.M., Abecasis, G.R., Altshuler, D.L., et al., 2010. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073.

Der Sarkissian, C., Balanovsky, O., Brandt, G., Khartanovich, V., Buzhilova, A., Koshel, S., Zaporozhchenko, V., Moiseyev, V., Gronenborn, D., Kolpakov, E., Shumkin, V., Alt, K.W., Balanovska, E., Cooper, A., Haak, W., The Genographic Consortium, 2013. Ancient DNA reveals prehistoric gene flow from Siberia in the complex human population history of northeast Europe. PLoS Genetics 9 (2), e1003296.

Der Sarkissian, C., Brotherton, P., Balanovsky, O., Templeton, J.E.L., Llamas, B., Soubrier, J., Moiseyev, V., Khartanovich, V., Cooper, A., Haak, W., The Genographic Consortium, 2014. Mitochondrial genome sequencing in Mesolithic North East Europe unearths a new sub-clade within the broadly distributed human haplogroup C1. PLoS ONE 9 (2), e87612. http://dx.doi.org/10.1371/journal.pone.0087612.

El-Sibai, M., Platt, D.E., Haber, M., Xue, Y., Youhanna, S.C., Wells, R.S., Izaabel, H., Sanyoura, M.F., Harmanani, H., Bonab, M.A., Behbehani, J., Hashwa, F., Tyler-Smith, C., Zalloua, P.A., The Genographic Consortium, 2009. Geographical structure of the Y-chromosomal genetic landscape of the Levant: a coastal-inland contrast. Annals of Human Genetics 73 (Pt 6), 568–581.

Elhaik, E., Greenspan, E., Staats, S., Krahn, T., Tyler-Smith, C., Xue, Y., Tofanelli, S., Francalacci, P., Cucca, F., Pagani, P., Jin, L., Li, H., Schurr, T.G., Greenspan, B., Wells, R.S., Genographic Consortium, 2013. The GenoChip: a new tool for genetic anthropology. Genome Biology and Evolution 5, 1021–1031.

Elhaik, E., Tatarinova, T., Chebotarev, D., Piras, I.S., Calò, C.M., De Montis, A., Atzori, M., Marini, M., Tofanelli, S., Francalacci, P., Pagani, L., Tyler-Smith, C., Xue, Y., Cucca, F., Schurr, T.G., Gaieski, J.B., Melendez, C., Vilar, M.G., Owings, A.C., Gomez, R., Fujita, R., Santos, F.R., Comas, D., Balanovsky, O., Balanovska, E., Zalloua, P., Soodyall, H., Pitchappan, R., GaneshPrasad, A., Hammer, M., Matisoo-Smith, L., Wells, R.S., The Genographic Consortium, 2014. Geographic population structure (GPS) of worldwide human populations infers biogeographical origin down to home village. Nature Communications 5, 3513. http://dx.doi.org/10.1038/ncomms4513.

Forster, P., 2004. Ice ages and the mitochondrial DNA chronology of human dispersals: a review. Philosophical Transactions of the Royal Society of London B 359, 255–264.

Gaieski, J.B., Owings, A.C., Vilar, M.G., Dulik, M.C., Gaieski, D., Gittelman, R.M., Lindo, J., Gau, L., Schurr, T.G., The Genographic Consortium, 2011. Genetic ancestry and indigenous heritage in a Native American descendant community in Bermuda. American Journal of Physical Anthropology 46 (3), 392–405.

Gan, R.-J., Pan, S.L., Mustavich, L.F., Qin, Z.-D., Cai, X.-Y., Qian, J., Liu, C.-W., Peng, J.-H., Li, S.-L., Xu, J.-S., Jin, L., Li, H., The Genographic Consortium, 2008. Pinghua population as an exception of Han Chinese's coherent genetic structure. Journal of Human Genetics 53, 303–313.

Green, R.E., Krause, J., Briggs, A.W., et al., 2010. A draft sequence of the Neandertal genome. Science 328, 710–722.

Green, R.E., Krause, J., Ptak, S.E., et al., 2006. Analysis of one million base pairs of Neanderthal DNA. Nature 444, 330–336.

Gresham, D., Morar, B., Underhill, P.A., et al., 2001. Origins and divergence of the Roma (gypsies). The American Journal of Human Genetics 69, 1314–1331.

The 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073.

Haak, W., Balanovsky, O., Sanchez, J.J., Koshel, S., Zaporozhchenko, V., Adler, C.J., Der Sarkissian, C.S.I., Brandt, G., Schwarz, C., Nicklisch, N., Dresely, V., Fritsch, B., Balanovska, E., Villems, R., Meller, H., Alt, K.W., Cooper, A., The Genographic Consortium, 2010. Ancient DNA from European early neolithic farmers reveals their near eastern affinities. PLoS Biology 8 (11), e1000536.

Haber, M., Platt, D.E., Badro, D.A., Xue, Y., El-Sibai, M., Ashrafian Bonab, M., Youhanna, S.C., Saade, S., Soria-Hernanz, D.F., Royyuru, A., Wells, R.S., Tyler-Smith, C., Zalloua, P.A., The Genographic Consortium, 2011. Influences of history, geography, and religion on genetic structure: the Maronites in Lebanon. European Journal of Human Genetics 19, 334–340.

Haber, M., Platt, D.E., Bonab, M.A., Youhanna, S.C., Soria-Hernanz, D.F., Martínez-Cruz, B., Douayhi, B., Ghassibe-Sabbagh, M., Rafatpanah, H., Ghanbari, M., Balanovsky, O., Wells, R.S., Comas, D., Tyler-Smith, C., Zalloua, P.A., The Genographic Consortium, 2012. Afghanistan's ethnic groups share a Y-chromosomal heritage structured by historical events. PLoS ONE 7 (3), e34288.

Hammer, M.F., Zegura, S.L., 2002. The human Y-chromosome haplogroup tree: nomenclature and phylogeography of its major divisions. Annual Review of Anthropology 31, 303–321.

Ingman, M., Kaessmann, H., Pääbo, S., et al., 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408, 708–713.

Javed, A., Melé, M., Pybus, M., Zalloua, P., Haber, M., Comas, D., Netea, M.G., Balanovsky, O., Balanovska, E., Jin, L., Yang, Y., Arunkumar, G., Pitchappan, R.M., Bertranpetit, J., Calafell, F., Parida, L., The Genographic Consortium, 2013. Recombination networks as genetic markers in a human variation study of the old world. Human Genetics 131 (4), 601–613.

Jota, M.S., Lacerda, D.R., Sandoval, J.R., Vieira, P.P., Santos-Lopes, S.S., Bisso-Machado, R., Rodrigues Paixão-Cortes, V., Revollo, S., Paz-y-Miño, C., Fujita, R., Salzano, F.M., Bonatto, S.L., Bortolini, M.C., Santos, F.R., The Genographic Consortium, 2011. A new subhaplogroup of Native American Y-chromosomes from the Andes. The American Journal of Physical Anthropology 146 (4), 553–559.

Jobling, M.A., Tyler-Smith, C., 2003. The human Y chromosome: an evolutionary marker comes of age. Nature Reviews Genetics 4, 598–612.

Kang, L., Lu, Y., Wang, C., Hu, K., Chen, F., Li, S., Jin, L., Li, H., The Genographic Consortium, 2012. Y-chromosome haplogroup O3 diversity in Sino-Tibetan populations reveals two migration routes into the eastern Himalayas. Annals of Human Genetics 76 (1), 92–99.

Karafet, T.M., Mendez, F.L., Meilerman, M.B., et al., 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Research 18, 830–838.

Kayser, M., Krawczak, M., Excoffier, L., et al., 2001. An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. The American Journal of Human Genetics 68, 990–1018.

Kivisild, T., Bamshad, M.J., Kaldma, K., et al., 1999. Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. Current Biology 9 (22), 1331–1334.

Kong, Q.-P., Yao, Y.-G., Liu, M., et al., 2003. Mitochondrial DNA sequence polymorphisms of five ethnic populations from northern China. Human Genetics 113, 391–405.

Li, D., Wang, C.-C., Yang, K., Qin, Z.-D., Lu, Y., Lin, X.-J., Li, H., The Genographic Consortium, 2013. Hainan indigenous substitute for genetic lineages of Champa exile Utsat people. Journal of Systematics and Evolution 51 (3), 287–294.

Lu, Y., Pan, S.-L., Qin, S.-M., Qin, Z.-D., Wang, C.-C., Gan, R.-J., Li, H., The Genographic Consortium, 2013. Genetic evidence for the multiple origins of Pinghua Chinese. Journal of Systematics and Evolution 51 (3), 271–279.

Lu, Y., Qin, Z., Wen, B., Jin, L., Li, H., The Genographic Consortium, 2012. Mitochondrial origin of the matriarchal Mosuo people in China. Mitochondrial DNA 23 (1), 13–19.

Li, J.Z., Absher, D.M., Tang, H., et al., 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100–1104.

Martínez-Cruz, B., Harmant, C., Platt, D.E., Manry, J., Ramos-Luis, E., Hernanz, D.F., Bauduer, F., Salaberria, J., Oyharçabal, B., Quintana-Murci, L., Comas, D., The Genographic Consortium, 2012a. Evidence of pre-Roman tribal genetic structure in Basques from uniparental genomes. Molecular Biology and Evolution 29 (9), 2211–2222.

Martínez-Cruz, B., Ioana, M., Calafell, F., Rubio, L., Sanz, P., Ioenscu, R., Boengui, S., Plantinga, T., van der Meer, J.W.M., Comas, D., Netea, M.G., The Genographic Consortium, 2012b. Y-chromosome analysis in individuals bearing the Basarab name of the first dynasty of Wallachian kings. PLoS ONE 7 (7), e41803.

Martínez-Cruz, B., Ziegle, J., Sanz, P., Sotelo, G., Anglada, R., Plaza, S., Comas, D., The Genographic Consortium, 2011. Multiplex screening of the human Y-chromosome using TaqMan probes. Investigative Genetics 2, 13.

McAllister, P., Nagle, N., Mitchell, R.J., The Genographic Consortium, 2013. The Australian Barrineans: southern negritos or rainforest-adapted aborigines? Human Biology 85 (1–3), 485–494.

Melé, M., Javed, A., Pybus, M., Calafell, F., Parida, L., The Genographic Consortium, 2010. A new method to reconstruct recombination events at a genomic scale. PLoS Computational Biology 6 (11), e1001010.

Melé, M., Javed, A., Pybus, M., Zalloua, P., Haber, M., Comas, D., Netea, M.G., Balanovsky, O., Balanovska, E., Jin, L., Yang, Y., Pitchappan, R.M., Arunkumar, G., Parida, L., Calafell, F., Bertranpetit, J., The Genographic Consortium, 2012. Recombination gives a new insight in the effective population size and the history of the old world human populations. Molecular Biology and Evolution 29 (1), 25–30.

Metspalu, M., Kivisild, T., Metspalu, E., et al., 2004. Most of the extant mtDNA boundaries in South and Southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. BMC Genetics 5 (26). http://dx.doi.org/10.1186/1471-2156-5-26.

Meyer, M., Kircher, M., Gansauge, M.-T., et al., 2010. A high-coverage genome sequence from an archaic Denisovan individual. Science 338, 222–226.

Oppenheimer, S., Richards, M., 2001. Fast trains, slow boats, and the ancestry of the Polynesian islanders. Science Progress 84 (3), 157–181.

van Oven, M., Kayser, M., 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Human Mutation 29, E386–E394.

Parida, L., Javed, A., Melé, M., Calafell, F., Bertranpetit, J., The Genographic Consortium, 2008. Optimizing recombination events in coalescing networks. Journal of Computational Biology 15 (9), 1133–1153.

Parida, L., Javed, A., Melé, M., Calafell, F., Bertranpetit, J., The Genographic Consortium, 2009. Minimizing recombinations in consensus networks for phylogeographic studies. BMC Bioinformat. 10. Article S72.

Palanichamy, M.G., Sun, C., Agrawal, S., et al., 2004. Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. The American Journal of Human Genetics 75, 966–997.

Qin, Z., Yang, Y., Kang, L., Cai, X., Zhu, D., Fei, D., Li, S., Jin, L., Li, H., The Genographic Consortium, 2010. A mitochondrial revelation of early human migrations to the Tibetan plateau before and after the last glacial maximum. American Journal of Physical Anthropology 143 (4), 555–569.

Quintana-Murci, L., Semino, O., Bandelt, H.-J., et al., 1999. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nature 23, 437–441.

Rębała, K., Martínez-Cruz, B., Tönjes, A., Kovacs, P., Stumvoll, M., Lindner, I., Büttner, A., Wichmann, H.-E., Siváková, D., Soták, M., Quintana-Murci, L., Szczerkowska, Z., Comas, D., The Genographic Consortium, 2012. Contemporary paternal genetic landscape of Polish and German populations: from early medieval Slavic expansion to post-World War II resettlements. European Journal of Human Genetics 21, 415–422.

Rosset, S., Wells, R.S., Soria, D., Tyler-Smith, C., Royyuru, A., Behar, D.M., The Genographic Consortium, 2008. Maximum likelihood estimation of site-specific mutation probabilities from partial phylogenetic classifications. Genetics 180, 1511–1524.

Rasmussen, M., Guo, X., Wang, Y., et al., 2011. An aboriginal Australian genome reveals separate human dispersals into Asia. Science 334 (6052), 94–98.

Rasmussen, M., Li, Y., Lindgreen, S., et al., 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463, 757–762.

Reich, D., Green, R.E., Kircher, M., et al., 2012. Genetic history of an archaic hominin group from Denisova cave in Siberia. Nature 468, 1053–1060.

Richards, M.B., 2003. The Neolithic invasion of Europe. Annual Review of Anthropology 32, 135–162.

Sandoval, J.R., Lacerda, D.R., Jota, M.S.A., Salazar-Granara, A., Vieira, P.P.R., Acosta, O., Cuellar, C., Revollo, S., Fujita, T., Santos, F., The Genographic Consortium, 2013. The genetic history of indigenous populations of the Peruvian and Bolivian Altiplano: the legacy of the Uros. PLoS ONE 8 (9), e73006.

Schurr, T.G., Dulik, M.C., Owings, A.C., Zhadanov, S.I., Gaieski, J.B., Vilar, M.G., Ramos, J., Moss, M.B., Natkong, F., The Genographic Consortium, 2012. Clan, language and migration history have shaped genetic diversity in Haida and Tlingit populations from southeast Alaska. American Journal of Physical Anthropology 148, 422–435.

Salas, A., Richards, M., De la Fe, T., et al., 2002. The making of the African mtDNA landscape. The American Journal of Human Genetics 71, 1082–1111.

Schurr, T.G., 2004. The peopling of the new world: perspectives from molecular anthropology. Annual Review of Anthropology 33, 551–583.

Torroni, A., Achilli, A., Macaulay, V., Richards, M., Bandelt, H.-J., 2006. Harvesting the fruit of the human mtDNA tree. Trends in Genetics 22 (6), 339–345.

Underhill, P.A., Passarino, G., Lin, A.A., et al., 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Annals of Human Genetics 65, 43–62.

Vilar, M.G., Dulik, M.C., Owings, A.C., Gaieski, J.B., Lennie, C., Zillges, H., Gittelman, R., Lindo, J., Keating, K., Schurr, T.G., The Genographic Consortium, 2014a. Genetic variation in Inuvialuit populations from the Northwest territories: Implications for the human colonization of the Arctic. The American Journal of Physical Anthropology (in review).

Vilar, M.G., Melendez, C., Sanders, A., Walia, A., Gaieski, J.B., Owings, A.C., Schurr, T.G., The Genographic Consortium, 2014b. Genetic diversity in Puerto Rico and its implications for the peopling of the West Indies. Article first published online: July 16, 2014b. The American Journal of Physical Anthropology 155, 352–368. http://dx.doi.org/10.1002/ajpa.22569.

Wang, C.-C., Yan, S., Qin, Z.-D., Lu, Y., Ding, Q.-L., Wei, L.-H., Li, S.-L., Yang, Y.-J., Jin, L., Li, H., The Genographic Consortium, 2013. Late Neolithic expansion of ancient Chinese revealed by Y-chromosome haplogroup O3a1c-002611. Journal of Systematics and Evolution 51 (3), 280–286.

Weiss, K.M., Kidd, K.K., Kidd, J.R., 2002. Human genome diversity project. Evolutionary Anthropology 1 (3), 80–82.

Wells, R.S., Yuldsheva, N., Ruzibakiev, R., et al., 2001. The Eurasian heartland: a continental perspective on Y-chromosome diversity. Proceedings of the National Academy of Sciences of the United States of America 98 (18), 10244–10249.

Yan, S., Wang, C.C., Li, H., Li, S.L., Jin, L., The Genographic Consortium, 2011. An updated tree of Y chromosome haplogroup O and revised phylogenetic position of mutations P164 and PK4. European Journal of Human Genetics 19, 1013–1015.

Yang, K., Zheng, H., Qin, Z., Lu, Y., Farina, S.E., Li, S., Jin, L., Li, D., Li, H., The Genographic Consortium, 2011. Positive selection on mitochondrial M7 lineages among the Gelong people in Hainan. Journal of Human Genetics 56, 253–256.

Y-Chromosome Consortium, 2002. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Research 12, 339–348.

Yao, Y.-G., Kong, Q.-P., Wang, C.-Y., et al., 2004. Different matrilineal contributions to genetic structure of ethnic groups in the silk road region in China. Molecular Biology and Evolution 21 (12), 2265–2280.

Zalloua, P.A., Platt, D.E., El Sibai, M., Khalife, J., Makhoul, N., Haber, M., Xue, Y., Izaabel, H., Bosch, E., Adams, S.M., Arroyo, E., Aler, M., Picornell, A., Ramon, M., Jobling, M.A., Comas, D., Bertranpetit, J., Wells, R.S., Tyler-Smith, C., The Genographic Consortium, 2008a. Identifying genetic traces of historical expansions: Phoenician footprints in the mediterranean. The American Journal of Human Genetics 83 (5), 633–642.

Zalloua, P.A., Xue, Y., Khalife, J., Makhoul, N., Debiane, L., Herrera, R.J., Soria, D., Blue-Smith, J., Wells, R.S., Comas, D., Bertranpetit, J., Tyler-Smith, C., The Genographic Consortium, 2008b. Y-chromosomal diversity in Lebanon is structured by recent historical events. The American Journal of Human Genetics 82, 873–882.

Zhadanov, S.I., Dulik, M.C., Markley, M., Jennings, G., Gaieski, J.B., Elias, G., Schurr, T.G., The Genographic Consortium, 2010. Genetic heritage and native identity of the Seaconke Wampanoag tribe of Massachusetts. The American Journal of Physical Anthropology 142 (4), 579–589.

Zerjal, T., Xue, Y., Bertorelle, G., et al., 2003. The genetic legacy of the Mongols. The American Journal of Human Genetics 72, 717–721.

Relevant Websites

https://genographic.nationalgeographic.com/ – Genographic Project.
http://www.1000genomes.org/ – 1000 Genomes Project.
http://www.isogg.org/ – International Society of Genetic Genealogists.
http://www.phylotree.org/ – Phylotree.
http://www.yhrd.org/ – Y-chromosome Haplotype Reference Database.

Genomics, Ethical Issues in

Henry T Greely, Center for Law and the Biosciences, Stanford University, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Improvements in our ability to analyze and understand DNA will have ethical implications for society. The most important issue stem from DNA's ability to predict an individual's future, during his or her life, at birth, as a fetus, as an *in vitro* embryo, or even before conception. There will also be ramifications for health care, in establishing identity, in revealing the past, through genetic manipulation, on ownership and control of genetic material and information, and in cultural understandings of humanness, race, and the roles of nature and nurture.

Genetics, both as a field and as a word, was born in the early twentieth century. The 'new' genetics referred to the vast increase in the power of genetics that followed the increased understanding of the biochemical basis for genetics in deoxyribonucleic acid (DNA). That understanding has made it possible to 'read' the sequence of DNA in organisms, including humans; to discover, in some cases, the physiological significance of that sequence; and to contemplate changing the sequence in order to attain desired ends. These heightened powers have both created new ethical and social concerns and exacerbated older ones.

The first edition of this article was written in 2001. The major development since then has been the growth of 'genomics.' Impressive advances in testing, sequencing, and manipulating DNA mean we can do much more with and to DNA – faster, cheaper, and more accurately – than imagined 13 years ago. Various methods of testing first hundreds, and ultimately millions, of 'single nucleotide markers' in a human genome using 'SNP chips' now cost less than $100, leading to their widespread use in the last 15 years in scientific 'genome-wide association studies' (GWAS) as well as in consumer genomics companies. GWAS still test under a thousandth of the whole human genome, but the first reasonably complete human genome sequence was finished in 2003. It cost about $500 million (depending on the accounting conventions used). In 2014, high-quality whole human genome sequences are routinely available for about $3000 with several firms recently announcing their '$1000 genomes.' These new large-scale capabilities have shifted the focus of scientific interest from individual genes to whole or partial genomes; the scientific language has similarly been moving from 'genetics' to 'genomics,' as will, largely, this article.

At the same time, the major 'nondevelopment' has been the creeping pace of our ability to understand the effects of different DNA sequences. A major question of the last decade has been the mystery of the 'missing heritability' – researchers know, for example, that about 90% of a person's adult height comes from his combination of his parents' genes, but thus far only 10–20 percentage points of that heritability has been attributed to specific DNA sequences. As a result, the consequences of the unexpectedly enhanced sequencing capabilities of genomics have been smaller than we would have imagined.

Concern about implications of the new genetics for human societies grew with scientists' ability to decipher and manipulate DNA. The late 1960s saw discussions of human cloning, followed in the 1970s and early 1980s by safety concerns about recombinant DNA and social concerns about genetic discrimination. When the Human Genome Project was proposed for substantial funding in the United States in the late 1980s, the Project's first director, Dr James Watson, suggested that 5% of the funding be set aside to study the 'ethical, social, and legal implications' (ELSI) of human genetics (Cook-Deegan, 1994). The subsequent financial commitment led to the rapid expansion of 'ELSI' studies across many disciplines and the publication of a vast number of articles and books on these topics. The technical advances brought by genomics have accelerated the timing of some ethical, social, and legal issues, but, because of the slow pace of understanding, neither broadly nor dramatically.

Three general caveats are useful in surveying the resulting discussions of either the 'new' genetics or the 'new' genomics. First, although these new technologies often provide great power, for prediction, or for intervention, they seem unlikely to provide the extent of power often assumed in the writings about their implications. Oftentimes, their predictions will be weak; the percentage of people with a particular set of genetic variations (or 'genotype') who develop a particular trait of condition (or 'phenotype'), a percentage known as 'penetrance,' will often be low. Second, although the technologies are quite novel, their implications almost always have parallels in the social effects of other modern technologies. And third, although much of the discussion of the consequences of genomic technologies has focused on individuals or families, these technologies usually also have implications for broader human groupings. With those caveats in mind, this article will discuss the ethical, social, and legal effects of the new genomics in seven areas: health care, establishing identity, predicting the future, revealing the past, genetic manipulation, ownership and control of genetic material and information, and cultural understandings.

Issues Arising from Medical Successes

Writing on social issues of genomics focuses on the dangers of human genomics misused – lost privacy, genetic discrimination, state-enforced eugenics. It rarely looks at the potential implications of the hoped for uses of genomics to prevent and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.82011-5

treat diseases. The vast sums, both private and public, being spent on research in genomics are not being committed from an altruistic search for knowledge, but in the expectation that the research will bring immense medical benefits. The diseases in question may not be limited to 'genetic diseases'; genomic tools are allowing unprecedented understanding, at the molecular level, of the proper and diseased functioning of human cells and of human pathogens. The most powerful lesson of the first 40 years of the biotechnology industry has been that the human body and its functions are more complicated than expected, but those complications are steadily being tackled with genomic tools. A stream of new treatments, derived from increased knowledge of genomes and their associated proteins, is flowing through research, development, and regulatory approval, not as quickly as one would like, but persistently.

What will happen if these drugs succeed? It is possible that the results might include, among other things, a significant extension of average human life span; an increase in pharmaceutical costs; and changes, up and down, in the demand for various medical services. The social implications could be substantial, on everything from pension plans and political voting blocks to health care systems. Some more focused medical consequences of genomics research, such as individualizing patient therapies as a result of genomic tests (known as 'pharmacogenomics'), or some kinds of proposed genome therapy or tissue replacement will require that treatments be created just for one patient, thus straining the existing mechanisms for drug development, manufacturing, approval, and financing. And any new treatment will raise questions of availability, both within any one country and between countries. Genomics has created a uniquely high expectation of medical progress, which it has used to acquire funding, both public and private. If it comes close to the successes it has promised, these consequences may be raised in dramatic form.

Establishing Identity

Genomics raises four main issues centering around identity: forensic identification, personal identity, family identity, and ethnic identity.

Forensic Identification

By the mid-1990s, the new genetics was leading to major changes in forensic identification. Any tissue from a person that contains DNA can be tested for a pattern of identifying markers, sections of the genome that vary substantially from one person to another. Those markers can be compared with the markers analyzed from human tissue found in connection with a crime (National Research Council, 1992). New and powerful techniques for analyzing DNA mean that samples with only tiny amounts of genetic material can be useful, whether derived from flesh, blood, semen, or even the cells found in saliva. A negative test is very powerful. If the markers are significantly different, the tissues cannot be from the same person (unless the sample was swapped or contaminated or in a very rare case where a single person because of embryonic merger, has some tissues with one genome and some with the genome of a never-born twin). Initially, there was concern that a positive match might not be very compelling because patterns of markers might be relatively common among particular ethnic populations. That could make the odds that a match was a coincidence vanishingly small among the general population but relatively high among people from that particular ethnic group. These fears were largely assuaged through both more information on the distribution of the markers used in forensics and different techniques (National Research Council, 1996).

This forensic use of DNA is not essentially different from the use of fingerprints, blood groups, or dental records. Like those techniques, it has had to prove its effectiveness in a host of judicial cases. And, like the earlier techniques, the DNA itself can never be conclusive. Good defense counsel will always investigate the possibilities of an innocent explanation for the presence of the DNA or for mistakes, contamination, or fraud in the DNA collection and analysis process. Juries and judges might pay more attention to DNA evidence because of the high status and reputation of the science, but that effect should also fade with greater familiarity (and good defense counsel).

The real issues about the forensic use of DNA now concern not the individual case, but the collection and retention of DNA or DNA records for identification purposes. One issue concerns surreptitious collection by the police of DNA from suspects, either through an interaction with them, such as offering them something to drink and getting their DNA from the discarded container, or merely by collecting DNA from items they have had contact with, from cigarette butts to eating utensils. More broadly, many countries or states have created DNA repositories with DNA samples taken from convicted, or even merely accused, criminals. These samples can be used to seek matches in criminal cases or for the identification of human remains. If markers from the samples are analyzed in advance, the results can be put into a database and used to seek 'cold matches' to samples or tissue of unknown origin. These repositories raise two different issues. The first involves their creation. Should samples be accepted only from those who consent or should they be mandatory? If mandatory, what classes of people should be required to provide samples – convicted sex offenders, convicted felons, all those arrested, or the whole population? Litigation on this issue has become increasingly common in the United States, where all states have some form of mandatory DNA repository (*Maryland v. King*). A subtler, but equally important, issue concerns *what* the repository should contain: information on what, and how many, genetic markers? And should the DNA samples from which the profiles were derived be retained or destroyed? The markers currently used for forensic identification, as far as scientists can tell, are not strongly associated with any traits (other than sex) (Kaye and Greely, 2013). If the markers are kept, no information about a person other than identity (and possibly ancestry) can be deduced – nothing about genetically influenced disease susceptibility or other traits. (Using new markers as is currently under serious consideration, could change that.) If samples of whole DNA are kept, however, they can be reanalyzed to confirm that the profile is accurate, which may be important in some contexts, such as some international sharing arrangements. But the samples could also be analyzed for DNA associated with a wide variety of traits. This raises

much broader implications for privacy than a regime where the samples are destroyed after analysis.

Personal Identity

The furor about the possible use of somatic cell nuclear transfer (SCNT) technology to clone not just a sheep, but a human, involves a host of issues (National Bioethics Advisory Commission, 1997). Most of those issues have already been raised as a result of *in vitro* fertilization and other forms of advanced reproductive technologies. The issue unique to cloning is, in essence, a concern about human identity. Would a person with the same genome as another living individual be, in some meaningful sense, a copy of that person?

This and other speculative questions about clones were widely discussed in the aftermath of the announcement in 1997 of the birth of the cloned sheep, Dolly. Concern died down as scientists were consistently unable to produce viable human embryos using SCNT, in spite of fraudulent claims in 2005 by Dr Hwang Woo-suk. In 2013 and 2014, however, new methods did successfully produce apparently viable human embryos and embryonic stem cells from those cloned human embryos. Concern about the birth of human clones has, to some extent, revived.

Monozygotic twins, who account for about 1 in 240 live births around the world, have always shared the same genome. Although their physical similarity is strong, the individual identities, and separate personalities, of monozygotic twins have long been recognized. Cloning by SCNT, if possible in humans, could not produce individuals *more* alike than monozygotic twins. Indeed, the clone and the source of the clone's DNA would most likely be much less alike. They would have spent their first 9 months of development in the wombs of different women, subject to different environmental influences *in utero*. Once born, they would be subjected to different environments, often in very different times. The exact degree to which identical human genotypes produce similar phenotypes would be an interesting research question, in the unlikely event that it could ethically be studied. It seems quite likely that the popular perception of identity would prove grossly exaggerated.

Family Identity

Because genetic variations are inherited, parents and children share half of their genomic variations, as do siblings. Grandparents and grandchildren, as well as first cousins, share, on average, one-fourth of their genomic variations. This fact, along with cheap genomic analysis, makes it possible to use DNA to establish, or refute, close family relationships. This kind of family analysis, using inexpensive SNP chips as well as analysis of variations on the Y chromosome (found only in men and inherited directly from father to son) or the mitochondrial DNA (found in both men and women but inherited only from the mother), has led to the growth of a consumer genetic genealogy industry. At least three major firms each had hundreds of thousands of customers pay to get information about their genetic variations and to look for previously unknown genetic family members. At a more immediate level, genomic analysis can be used to try to find unknown parents, either when a mother (or a government) is trying to identify a genetic father for child support or when a child who was adopted or the product of an egg or sperm donor goes looking for genetic parents. With large enough genomic databases, close family relationships can be picked out easily. Similarly, the police can use genomic analysis to try to determine whether the source of crime scene DNA came from a close genetic relative of someone in a forensic database (Greely et al., 2006). This will become easier if the forensic databases adopt more markers, particularly on the paternally inherited Y chromosome. The ethical implications of using DNA to establish, or to refute, family relationships will vary, depending often on the consent of the family members being found or 'lost.'

Ethnic Identity

Culturally defined human groups – ethnicities or nations – often share more than common cultures, but also share a substantial degree of common ancestry. This genealogical connection implies a genetic connection; particular variations in DNA sequences, or patterns of those variations, will be more common in some populations than in others. This may apply to both common and rare variations. For example, the four ABO blood groups are genetically determined. Each one is found in all (or almost all) of the world's populations, but in different proportions. Other variations in DNA sequence may be found at elevated levels among particular groups and extremely rarely among others (Cavalli-Sforza et al., 1994).

This kind of variation can be used for anthropological research; it could also be used in an effort to define a person's ethnic identity. The latter use would have limited scientific validity; no known DNA variations are found in all members of an ethnic group and in no nonmembers of that group and given the human history of enthusiastic sexual intermixing (and of adoption or conversion), no such variations are likely to be found. Nonetheless, the rise of inexpensive genome-wide testing has led 'biogeographical' tests of an individual's ancestry becoming an important feature of popular genetic genealogy products sold to consumers, who learn that they are, for example, 87% European in origin, 11% Middle Eastern, and 2% sub-Saharan African. There have been reports of governments or ethnic groups using this kind of information for some legal purposes. The ethical implications of the naïve use of such inherently inaccurate methods for defining ethnicity are likely to depend on the circumstances, including the uses made of such definitions and whether they were adopted by the group itself or imposed on it. Any such use, however, would likely have the negative effect of reinforcing the false concept that race and ethnicity have some strong 'genetic' meaning.

Predicting the Future

The most discussed issues arising from the new genetics come from its perceived ability to make predictions about individuals' futures as a result of associations between genetic variations and physical and behavioral traits. These predictions in themselves can have implications for living people; the attempt to make such predictions prenatally can have implications for what people are born. The strength of the predictions will vary

dramatically. Some conditions, such as Huntington's disease, appear to follow inevitably from possession of a particular genetic variation. For other conditions or traits, a person with a particular genetic variation may have a somewhat increased but still small likelihood of having that condition or trait. The ethical significance of such predictions will vary largely based on who uses which predictions for what ends.

Prenatal Selection to Avoid Disease

The new genetics brought the ability to test directly the DNA of the parents and of the embryos and fetuses they conceive; genomics has made such testing increasingly practical. Where the analyzed genotypes are strongly or completely associated with particular phenotypes, whether the children might (when the parents are tested) or would (when the embryos or fetuses were tested) show those traits can often be confidently predicted. This kind of prenatal testing has been used for several decades now in efforts to avoid serious genetic diseases. Early efforts included testing for as sickle cell anemia and for Tay-Sachs disease. In some cases, prospective parents have been tested, in order to alert to them to whether such a disease is a possibility for their children. The prospective parents could use that information to avoid marriage, to avoid having children, or to seek prenatal testing during any pregnancy. This carrier testing, at least when consented to by informed patients, has been relatively uncontroversial; it is now becoming increasingly common, and broad, as commercial firms offer carrier testing for over a hundred genetic disease traits for only a few hundred dollars.

Testing of fetuses is possible only with DNA from the fetus. In the past that DNA was retrieved from cells taken from the amniotic fluid about 16 weeks into the pregnancy through amniocentesis or from the chorionic villi, part of the placenta, at about 10–12 weeks of pregnancy. Both procedures are expensive and increase the risk of a miscarriage. Since late 2011, another method has been available, the so-called 'noninvasive prenatal testing' (NIPT). This makes use of small, broken bits of the fetus's DNA that can be found in the pregnant woman's blood (along with more small bits of her own DNA). It requires only a small blood draw from the pregnant woman and can be done at least as early as the 10th week of pregnancy. By 2014, at least four U.S. companies offered this test for various chromosomal abnormalities, notably trisomy 21, which causes Down syndrome. This test also easily reveals the fetus's sex. It seems likely that, eventually, this relatively inexpensive procedure will be useful to test the fetus for any genomic variation of interest; it has already been used, in a proof of principle experiment, to recreate the fetus's entire genome sequence. Currently only a few percent of pregnancies receive genetic testing; NIPT seems likely to increase that percentage enormously (Greely, 2011).

Whatever the method, after a direct genetic test of the fetus, parents could choose to abort or could prepare themselves better for the birth of a child with the particular genetic condition. Testing of fetuses for genetic conditions has been controversial. Some have opposed it because it often leads to abortion, which they condemn as murder. Some advocates for people with disabilities have been concerned about this kind of testing, especially when followed by abortion, because they believe it implies that the lives of people with those genetic conditions are not worth living. In the United States abortion, at least until the fetus is viable outside the womb, is legal for any reason, so there has been little discussion of outlawing abortion for late onset, low penetrance, or mild genetic conditions; many countries do regulate the reasons for abortion and so face this issue. Even in the United States, a question remains as to whether governments, health care systems, and health care professionals should encourage, discourage, or be neutral on prenatal genetic testing.

A different method may change some of the discussions about prenatal genetic testing. 'Preimplantation genetic diagnosis' makes it possible to test a very young embryo before it is implanted. This procedure can only be done as part of *in vitro* fertilization because the clinicians need to have access to the early stage embryo. One cell is detached from an embryo and the DNA in that cell tested. Based on the results, the parents can decide whether to implant that embryo, or, more realistically, which of several tested embryos to implant. No fetus would be aborted; instead, some embryos would just not be implanted. In the past this procedure has been uncommon; less than 0.1% of annual births in the U.S. follow from it. That percentage may increase. New methods of genomic analysis now mean that rather than testing the embryos for one or two genetic conditions, their whole genomes could be sequenced. Furthermore, within the next few decades the need for egg harvest, the most expensive, unpleasant, and risky aspect of *in vitro* fertilization, may be avoided by using skin or other cells from a woman to make new eggs.

Prenatal Selection for Genetic 'Enhancement'

By eliminating the need to abort in order to choose genetic traits, preimplantation genetic diagnosis also increases the possibility of parents selecting children based on genetic traits that would, the parents believe, 'enhance' their children. The idea of using genetic testing to select 'enhanced' genetic traits for offspring has raised many concerns (Parens, 1998). Some fear that allowing parents to choose, or to believe they can choose, the traits of their children would deny those children their right to 'an open future' not of their parents' choosing. Others worry that the rich will be able to buy genetic advantages for their offspring, denied to those who conceive in the traditional manner. Still others worry that parental selection will lead to a homogenization of the human gene pool, as all parents opt for children with similar traits. On the other hand, some fear that some parents will want to 'enhance' their children by selecting genetic variations that most people would view as *disabilities*. For example, deaf parents might want a deaf child. Finally, some fear that these parental choices could lead to a self-perpetuating genetic caste system. Most of these results require both that genetic variations be identified that strongly predict 'success' and that many parents must be willing to undergo preimplantation genetic diagnosis. Neither is certain.

One might also question whether parents choosing to guide the choices or to enhance the chances of their children through genetics is fundamentally different from parental efforts to do the same through the environment. Rich parents send their children to 'better' schools. Parents with particular interests may push their children from an early age to excel in those

activities – Wolfgang Mozart and Tiger Woods are both successful examples. Arguably, a major goal of parenting is to 'deny' one's children certain unhappy or illegal futures. The argument that enhancement through genotype selection is inherently different or more threatening from other kinds of enhancement is not an obvious one.

Sex selection is a special case of 'genetic enhancement,' though one that can be done through sonography as well as genetic testing. In some countries, prenatal sex selection by abortion is common with female fetuses being aborted much more frequently than males. This seems likely to produce striking imbalances in the sex ratio in those countries, which may have broad social consequences. It also, some argue, reinforces the subordinate position of women in those cultures. Easier genetic testing, through NIPT or greater use of preimplantation genetic diagnosis could increase the use of sex selection (Hvistendahl).

State-Sponsored Prenatal Genetic Selection

The discussion thus far has focused on parental choices about their children's genetic makeup, but parents need not make those decisions. Governments have intervened to prevent people with 'inferior' genes from having children; such eugenics laws could come again in more precise forms. 'Eugenics' is discussed in detail in other articles and will not be reviewed here except to note two things. First, state-sponsored eugenics measures do not have to be compulsory; state actions to encourage prenatal genetic selection can also be, in a sense, eugenics. Second, the ethics of government-sponsored eugenics depends on striking a balance between government public health powers and individual procreative liberty that remains in debate.

Postnatal Genetic Testing

The possibilities of genetic testing do not end at birth. Genetic tests can help diagnose diseases or predict future disease risks. Like all medical interventions, their value depends on them being used appropriately (President's Commission, 1983; Institute of Medicine, 1993). Postnatal genetic tests provide information that can lead to a medically useful intervention, that can help the patient make life plans, and that might just satisfy a patient's desire to know her future. Genetic testing also has costs. Test results may pose psychological problems for those tested or affect the patient's relationships with family members. They may have implications for the patient's dealings with the broader society, through employment or insurance discrimination. Where tests, genetic, or otherwise, do not have powerful implications for medical intervention, the strong influence of the individual's personality and circumstances on the value of testing argues a for particularly good process of informed consent, especially when the patient's understanding of the links between genetic variations and disease may be weak. Testing children is particularly ethically suspect unless there are useful medical interventions that need to be done during childhood. Otherwise, the benefits of testing for children could be put off until they reached adulthood and were able to make their own informed decision on testing. On the other hand, when there are good early interventions, testing for children becomes more important. Most jurisdictions require, or strongly encourage, neonatal genetic testing for diseases, such as phenylketonuria, where early diagnosis can lead to interventions with enormous benefits.

Improvements in DNA analysis offer the realistic possibility of whole genome sequencing as an affordable method of genetic testing, for diagnosis or for screening. Instead of looking for just one, or a handful, of genetic conditions, whole genome sequencing offers the possibility of seeing all genomic variations. Many issues remain concerning both the analytic validity of evolving whole genome sequencing techniques – how accurately they actually analyze the DNA sequence – and, even more, the clinical validity of interpretation of whole genome sequencing – how accurately medical conclusions can be drawn from them. Other problems concern what findings to return to patients, particularly, if those problems were not the reason for the test. And determining how to provide patients with meaningful information on hundreds or thousands of genomic risks (or advantages) will prove challenging (Ormond et al., 2010). All of these issues will be particularly troubling with respect to tests of children, especially if the currently mandatory neonatal genetic testing is replaced with whole genome sequencing.

Of course, many medical tests predict higher or lower disease risks. The diagnosis of a disease, whether through genetic or nongenetic methods, can trigger many of the same costs as prediction. Genetic diseases do implicate family members to an unusual extent, but shared environments or diets can also put family members at shared risk.

Genetic Discrimination

Great concern has been expressed about the possible uses of genetic tests to discriminate against people based on genetically predicted susceptibilities. Employment, life insurance, and, in the United States, health insurance are the fields where genetic discrimination seems to be most feared. The likelihood of genetic discrimination depends on two different sets of factors: the strength of the prediction possible from genetic information and the social structure governing the relevant fields.

For someone with the allele that leads inevitably to the fatal condition, Huntington's disease, genetic prediction is quite powerful. Fortunately, such powerful genetic predictions are rare; it is not clear that employers or insurers would find significant an increased risk of diabetes from 10 to 15%. Similarly, discrimination in health insurance has been mainly a concern in the United States; all other rich countries guarantee health coverage to all, or almost all, of their residents, a direction in which the United States is moving. Even in the United States, the great majority of Americans with health insurance obtain it through methods that are not directly susceptible to genetic discrimination. Medicare, Medicaid, and employer-provided health insurance are not medically underwritten – they must accept everyone eligible without using any medical exclusions. They cover about 90% of Americans with health coverage. There is little good evidence that genetic discrimination in employment or insurance exists in the United States, although there is clear evidence of public fear of such discrimination.

Until 2008, no federal law banned genetic discrimination in insurance or in employment, although many state laws banned one or both. In 2008, the United States adopted the Genetic Information Non-Discrimination Act, known as GINA, which largely banned the use of genetic information in employment and in health insurance. GINA does not cover discrimination in life insurance, disability insurance, or long-term care insurance. The United Kingdom, where the National Health Service makes private health insurance of limited importance, has set up a regulatory authority to pass on whether or not life insurers may require tests for particular genetic conditions; the body recently approved the use in life insurance of predictive genetic testing for Huntington's disease. How well statutes limiting genetic discrimination will work remains unclear, although GINA has not yet been expensive or difficult to enforce. If some of the statutes banning genetic discrimination are able to function with relatively low costs, one might justify them on the pragmatic ground that they increase the research on and clinical use of human genetics by reducing public fears, whether or not those fears are justified (Greely, 2001).

Prediction of Behavioral Traits

The successes of the new genetics have largely been in associating genetic variations with physical traits or conditions. Much of the public interest in genetics and genomics, though, seems to lie in its possible to use to predict human behavioral traits, such as intelligence, sexual orientation, violence, or mental illness. Some human behavioral traits have been linked to genetic variations, including some forms of intellectual disability. For the most part, though, efforts to link genetic variations with human behavioral variations have, thus far, had little success and even the successes have been controversial. At this point, it is not clear to what extent genetic variations will prove able to lead to strong behavioral predictions. If such predictions were possible, the ethical implications could be substantial (Carson and Rothstein, 1999).

Strong links between genetic variations and behavior might undermine notions of personal responsibility for the involved behaviors. At the broadest level, they could affect society's view of the extent of individual free will. More narrowly, they could affect the society's view of, for example, 'genetically determined' criminal behavior or sexual orientation, although whether the result would be more or less tolerance for the behavior is not clear. These kinds of genetic predictions could also lead to interventions in individual's lives – people predicted to be likely to act violently might be put into custody before they committed any offense. The successful association of behaviors perceived as important with genetic variations would boost the interest of both parents and governments in prenatal genetic selection.

Societies make predictions about future individual behavior in many ways, from aptitude tests to past records. Except for its prenatal uses, it is not clear that doing a genetic test to determine, for example, musical ability, would offer any advantages over doing a direct test of such ability. Genetic tests would have an advantage only when the genetic prediction is strong and the behavior is not easily observed directly. At this point, it is not clear that many strong associations between human behavior and genomic variations will be discovered; the problem of 'missing heritability' has been particularly great with behavioral traits. The real ethical issues with behavioral genetics may turn out to be preventing misuse by cultures that believe in its reality even in the face of uncertain science.

Uncovering the Past

DNA can reveal aspects of the past as well as of the present. In some cases, DNA, by providing another line of evidence, can raise in new contexts issues of the privacy of historical figures in matters such as Thomas Jefferson's relationship with his slave, Sally Hemmings, or whether Abraham Lincoln had Marfan's syndrome. Other historical uses of DNA raise broader questions.

Individual Ancestry

Genetic variations can be used in a straightforward way to establish close biological relationships. Blood groups or isoforms of particular proteins have long been able to provide such information; direct DNA testing can increase the accuracy of such determinations. The ethical implications of such paternity or maternity testing vary according to its use. If done with the full consent of all relevant parties, it seems unproblematic. If done without consent of one of the parties, it might be justified, as perhaps in establishing child support obligations, but would require some consideration. One problem arises where genetic testing done for some other reason, such as attempting a diagnosis of a disorder of possibly genetic origin in a child, reveals unsought information about parentage, such as 'false paternity,' that might be unwelcome and even dangerous. Another problem could be its use by children to find sperm or egg donors who were promised anonymity.

Group Histories

Analysis of patterns of genetic variation among different human populations can reveal how closely related two populations are. This information then becomes some evidence about human history and migrations. It is not conclusive evidence, but would join with linguistic, archaeological, anthropological, historical, and other sources to improve understanding of the human past (Cavalli-Sforza et al., 1994). The information may not always be welcome. It may contradict a population's own beliefs about its history. In some situations, that information alone might destabilize or disrupt the entire culture. It could also have modern political implications where ancestral origins, and the length of time in possession of particular territories, could affect land disputes. Whether a historical researcher has ethical obligations to avoid doing research that could provoke such consequences is unclear. Similarly unclear is whether there is an ethical obligation to seek the informed consent of the entire group to such genetic research that might have consequences for the group (Greely, 1997). Such consent has not generally been thought necessary for other historical investigations affecting contemporary peoples; some might argue that 'genetic history' is different, either because of the nature of the materials being used or the different degree of certainty perceived for such 'scientific'

evidence. Somewhat similar issues arise about the necessity of obtaining permission for genetic analysis of human remains.

Manipulating Genes

The new genetics made it possible not just to read an organism's genes but to change them. The newer genomics is greatly enhancing that ability, creating additional concerns.

Somatic Cell Gene Therapy

Somatic cell gene therapy involves the placement of a human gene into a living person's somatic cells – cells that do not produce the eggs and sperm that in turn produce the next generation. Somatic cell gene therapy would aim to cure a disease only in the patient, not in the patient's descendants. It was initially conceived as introducing a properly functioning copy of a gene into a person who had a genetic disease as a result of inheriting only improperly functioning copies. Different types of somatic cell gene therapy have since been investigated for the treatment of diseases that are not primarily caused by inherited genes, such as AIDS and most kinds of cancer. Over one thousand clinical trials of somatic cell gene therapy have taken place; very few have, thus far, shown any success, although by 2014, good results were beginning to accumulate.

The genetic aspects of somatic cell gene therapy have been largely uncontroversial. In essence, gene therapy is merely another drug delivery system, a different way to get a normal human protein to the right place in the body. Somatic cell gene therapy therefore stands in the same position as most experimental therapies. Like such therapies, it has prompted concerns that desperate patients are not truly giving informed consent and that the possible benefits of the treatment are exaggerated. Gene therapy may face the latter problem to a greater extent than most experimental treatments because of the exaggerated public view of the power of anything genetic.

Germ Line Gene Therapy

Germ line gene therapy has been much more controversial (Nelson, 2000). In its usual form, it would introduce 'normal' human genes into the eggs or sperm of parents, or into the fertilized egg or early embryo of the offspring. The goal would be to change the eventual child's genetic inheritance. This could be done in order to avoid a genetic disease or in order to introduce an 'enhancing' genetic variation. No one has tried human germ line gene therapy in its classic sense; indeed, there has been an informal moratorium in the scientific community on trying such experiments in humans. Both its feasibility and its value are unclear.

New genes have been successfully introduced into the germ lines of other mammals, but with low efficiency. At the same time preimplantation genetic diagnosis allows parents to choose embryos based on their genetic variations, as long as the parents themselves produced the desired variations. If not, donated eggs or sperm would likely be a much safer and easier way to introduce the desired genes than germ line cell gene therapy; ultimately, somatic cell gene therapy might provide another way for parents to change their offspring's genetic variations.

One kind of germ line gene therapy is plausible, has been tried, and is currently being debated: mitochondrial transfer. The mitochondria, organelles found in all eukaryotic cells, including human cells, have their own very small genome. In humans, mitochondria, and hence the mitochondrial genome, are inherited only from a child's mother. If the mother's mitochondrial genome has variations that cause disease, all of her children will carry the same pathogenic DNA. This can be avoided by combining the nucleus of an egg from the woman who wants children with mitochondria from a woman with a healthy mitochondrial genome. At least one *in vitro* fertilization clinic in the United States tried this procedure in the late 1990s and early 2000s, but abandoned it when the Food and Drug Administration asserted jurisdiction over the procedure. In 2014, researchers have sought approval from authorities in both the United Kingdom and the United States to resume experiments with this mitochondrial transfer procedure, which would allow afflicted women to have children without their disease, but would change the genomes of all cells of the children, including their egg and sperm – though only the female children could pass on their mitochondrial DNA. These efforts have been quite controversial, raising concerns about children with 'three parents.'

Chimeras

The manipulation of genes permits the creation of 'chimeras,' creatures that are genetically a mix of two species, including organisms that could not possibly mate. In agriculture, this mixing of genes from very different organisms has been perceived by many as unethical and potentially dangerous. Unless one takes a strong view of the reality and sanctity of sharp lines between species, it is hard to develop a strong ethical argument against such mixes. On the other hand, one might argue that a gene in the new setting of a different species might develop different and possibly harmful functions than in its home species.

Chimeras involving human genes could provoke greater concern. Intentionally moving nonhuman genes into humans could be seen as lessening the humanity of the recipient; moving human genes into nonhumans could be seen as investing nonhumans with some aspects of humanity. The former has not been attempted; the latter, however, is routine. Most of the products of the biotechnology industry, including insulin, erythropoietin, and human growth hormone, are made by creatures that are genetically partially human and partially nonhuman. Human genes are transferred into useful host cells such as yeast, which then produce large amounts of the human protein for which the gene codes. Of course, an organism with 5300 yeast genes and one human gene does not seem very human; to be concerned about this kind of chimera would require a very strong belief in the essentially human, or sacred, nature of any human DNA sequence. Moving more human genes or moving human genes into more closely related organisms, such as chimpanzees, could provoke more serious concern about blurring the definition of humanity.

Artificial Genes or Genomes

It is possible for scientists to create 'new' genes, not found in any existing species, that would create new or modified proteins. One branch of the biotechnology industry specializes in such efforts, named 'directed evolution.' The technology takes related genes from different species, recombines them or makes mutations in them, and then examines how well the resulting protein functions. Similarly, scientists could construct, piece by piece, a novel genome. Placed inside a cell, it might be able to generate a new living organism. One research team has already synthesized the genome of one bacterium, inserted it into a related bacterium, and the resulting organism lived, reproduced, and seemed ultimately indistinguishable from the first bacterium. In 2014, another group synthesized, with a few changes, chromosome 3 of the 16 in a common yeast, used it to replace a natural copy of chromosome 3, and produced a viable yeast cell. New methods of synthesizing DNA, coupled with new methods for targeted replacement of specific existing DNA with synthesized DNA (particularly a method called CRISPR) greatly increase the possibilities of genomic editing. These techniques are unlikely to be used soon in humans, where the safety consequences are potentially enormous, but may well be used to transform bacteria, plants, and nonhuman animals. Recreating recently extinct species may be possible; creating quite new species – from algae for biofuels to unicorns for entertainment – is also plausible.

The safety of making new genes, and hence new proteins, or new genomes, and hence new organisms, is an obvious concern. The deeper concern would be whether humanity should 'play God' by intentionally creating such new entities. Of course, since the invention of agriculture, humans have created new kinds of organisms by cross-breeding between species and by selective breeding within species. New genes are created whenever a gene is mutated, which is sometimes a conscious step in experiments with nonhuman organisms. The speculative possibilities outlined here are more significant, and less accidental, interventions. It is not clear whether that makes them ethically different.

Ownership and Control

Research

Rules governing the conduct of human genetics research control the research subject's genetic materials and information. These rules generally require the informed consent of the subject to the research. Research subjects, however, rarely have any control over the subsequent uses of materials or information derived from them (Greely, 2007). The California Supreme Court decision in *Moore v. Regents of the University of California* (1990) ruled that the plaintiff could not assert a property interest in a cell line derived from cells taken from his body, although it allowed him to try to prove that the physician–scientist had not gotten his informed consent. There have been few court decisions on this point, but current practice generally follows *Moore* and gives research subjects no control over who can use their previously collected genetic materials and information, for what purposes, or for how long. Current practices also discourage any sharing of financial benefits of the research with the subjects on the theory that the hope of such benefits could be an 'undue inducement' to them to take part in the research.

Since the late 1990s, the improved ability to analyze DNA has led a new approach to human genomics research. This associational research seeks to find weak connections between genetic variations and disease by searching for correlations between the genotypic and phenotypic data of large groups of people. Originally, it was undertaken using SNPs, increasingly it can be done with whole exome or whole genome sequence data. This kind of research requires the creation of a resource made up of health records and DNA samples (which are then analyzed for their genotypes) of hundreds of thousands of people. These resources would be too expensive to be created for the study of one kind of illness. Instead, the resources would contain complete medical information and be available to test a wide range of hypotheses. Iceland provided the first large example of this kind of 'genotype/phenotype resource' or biobank, when the government gave a private firm, deCODE Genetics, the license to create a health records database with clinical, medical, and genetic information on all 275 000 residents of the country.

For the genealogical information and health records of Icelanders, the Icelandic legislation provided that consent was presumed; unless an Icelander returned a government form 'opting out' of the database, their records would be included. Explicit consent was required for the inclusion of a person's DNA information. The Icelandic plan raised numerous ethical concerns, including the use of presumed consent, the absence of discussion of the risks and benefits of specific research done on the resource, the degree that privacy would be protected, the access (if any) of other researchers to the company's data, the propriety of having a for profit company control this information, and the financial fairness of the agreements between Iceland and deCODE (Greely, 2000). The Iceland/deCODE plan ultimately foundered, largely for financial reasons, but other countries have created similar biobanks, notably the United Kingdom with its UK Biobank made up of samples and data from 500,000 volunteers with current plans to sequence the whole genomes of at least 100,000 of the participants. These biobanks raise similar issues. Soon, however, even bigger 'virtual' biobanks are likely to exist. Once a substantial number of patients have their whole genomes sequenced for medical purposes, all of their health information and their entire genome sequence will be stored in their electronic medical records. This will be an irresistible and potentially population-wide source of data for researchers; what kinds of consent or control will be required from the patients is unclear.

Patents

No single case unequivocally announced that genes, or human genes, could be patented, but when the U.S. Supreme Court, in *Diamond v. Chakrabarty*, approved a patent on a genetically altered bacterium, it started a continuing debate over the patenting and genetics (*Diamond v. Chakrabarty*, 1980). To many patent lawyers, DNA is just another organic molecule. Patenting of forms or parts of it seems them to fall within long-established principles concerning the composition of matter patents on molecules not found in pure or refined form in

nature (Eisenberg, 1990). Others objected to some or all gene patents on a variety of grounds. Those objections can be put into two categories: fundamental and technical.

Some people have fundamental objections to patents on genes, on human genes, or on genetically modified life forms. One set of objections, focusing on genes, contends that they are 'discoveries,' not 'inventions,' and so should not be patented. Others point out that genes were made by God, nature, or evolution – in any event, something other than those filing the patent claims – and should not be allowable as intellectual property. Another objection is that human genes are the common heritage of mankind and should not be 'owned' by anyone but should be held in common. Still others oppose gene patents as part of broader opposition to biotechnology; forbid genetic patents and the technology would, at least according to the industry, be slowed. Proponents of gene patents put forward answers to each of these objections; throughout the industrialized world, the proponents won. Patents on genes, human genes, and to a somewhat lesser extent genetically modified life forms were commonly allowed in the United States, Europe, and Japan. It is worth noting, though, that most patents expire 20 years from the date of application. The first wave of gene patents has already expired; presumably all patents on human genes will have expired by 2021, 20 years from the publication of a draft of 'the human genome.'

This broad picture was upset when the United States Supreme Court ruled in *Association for Molecular Pathology v. Myriad Genetics* (*Association for Molecular Pathology v. Myriad Genetics*, 2013) that naturally occurring genomic sequences could not be patented as 'compositions of matter' because they were 'products of nature.' The Court did say that humanly modified genomic sequences, such as 'complementary DNA,' could be patented. The *Myriad* case did not deal with patents on uses of DNA, although a slightly earlier Supreme Court decision, *Mayo Collaborative Services v. Prometheus Laboratories, Inc.* (*Mayo Collaborative Services v. Prometheus Laboratories, Inc.*, 2012) has limited the power of patents on genetic tests. The Court held that a test that compared the level of a molecule in a patient's blood to a known set of ranges and made dosing decisions as a result could not be patented because it was a 'law of nature.' That doctrine has already been used to invalidate patents on the process of comparing the sequence of a patient's BRCA 1 and BRCA 2 genes with known 'safe' and 'dangerous' sequences for the purpose of assessing cancer risk. The ultimate results of these Supreme Court decisions on the patentability of genes, genomes, and genomic technologies remain unclear.

Privacy

Protecting the privacy of genetic information is one way to control its use (Annas et al., 1995; Rothstein, 1997). But such 'genetic privacy' legislation raises its own set of problems.

The first problem is definitional – what is the 'genetic information' that should be protected? Information about a person's genetic variations can come from DNA tests, from other biochemical tests, from a naked eye examination, or from family history. A narrow definition, focusing on the results of 'genetic tests,' would miss such powerful information as the 50% chance that the child or sibling of a Huntington's disease patient carries the Huntington's allele. A broad definition, covering everything from which an inference can be drawn about a person's genotype, would cover almost all medical information. A very high cholesterol level is strong evidence that a person carries two alleles for familial hypercholesteremia. An average or low level is strong evidence that he or she does not. GINA deals with this problem, but mostly in the context of forbidding the use of genetic information to discriminate against people in employment or health insurance; not as a general privacy protection of such information.

If almost all medical information leads to inferences about genotypes, it may not make sense to try to protect 'genetic information' separately from medical information. And it may not be possible to 'protect' a person's medical information from the many institutions that have legitimate uses for it, including physicians; hospitals; and those employers, insurers, and governments that pay health care bills. At the same time, the passage of legislation expressly to protect 'genetic' information sends the public the inaccurate message that genetic information is always much more powerful and important than other medical information. Legislation to protect genetic privacy has been introduced in the United States Congress but has not been passed. Several states have passed such legislation, but it is too early to tell how effective that legislation will prove.

Cultural Consequences

The most far-reaching, but hard to predict, ethical implications of genomics may lie in its effects on society's beliefs. Three areas stand out.

First, genomics demonstrates, in a very powerful way, that all life on earth is related. Over one-third of the genes found in the single-celled brewer's yeast have recognizable relatives in the human genome. Some stretches of DNA appear in close to identical form between humans, mice, and fruit flies. Indeed, there appear to be few specifically 'human genes' but many human variants of genes shared by all mammals, vertebrates, multicellular creatures, or life forms generally. Darwin's thesis that all earthly life is related by descent from a common ancestor can be seen in these similarities in DNA. It is unclear what cultural significance, if any, this will have. It could not promote vegetarianism, because genomics shows that carrots and corn, like sheep and cattle, are our relatives. It might, however, promote greater respect for nonhuman life.

Second, the new genetics shows that all humans are closely related. Humans are all cousins. Our DNA differs, on average, at one spot in 1000. In the regions of the genome that code for protein, the differences are one base pair in 10 000. Humans from opposite ends of the earth are far more similar to each other genetically than are chimpanzees from the same band. Genetic theories were used to provide support for a 'scientific' racism in the first part of the twentieth century. Genomics should provide evidence against such racism.

Third, genetics may shift the balance in the cultural debate between nature and nurture as the source of human characteristics (Degler, 1991). Dr James Watson, codiscoverer of the structure of DNA, was famously quoted as saying "We used to think that our fate was in the stars. Now we know that our fate

lies in our genes." In a debate that has raged for centuries, DNA may appear to be proof that individuals are powerfully shaped by forces beyond their control. In fact, genomics paints a more complicated picture. Genes play a role in the development of many traits or diseases, but the environment or luck may also be essential. The general population seems to hold a much stronger belief in the power of genes. For that reason, genomics could end up promoting a more closed and fatalistic view of human life and abilities than either current society holds or science would support (Nelkin and Lindée, 1995). That reaction may prove to be the most significant ethical challenge of the new genetics.

See also: Biobanking: Ethical Issues; Bioethics in the Post-genomic Era; Bioethics: Genetics and Genomics; Direct-to-Consumer Personal Genetic Testing; Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute; Eugenics as an International Movement; Eugenics, History of; Forensic Genetic Databases: Ethical and Social Dimensions; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Engineering; Genetic Screening for Disease; Genetics and Forensics; Genetics and Social Justice; Genetics and Society; Genetics, Disease, and Reproduction; Genetics: Legal Aspects; Genetics: The New Genetics; New Genetics and Race; Race: Genetic Aspects.

Bibliography

Annas, G.J., Glantz, L.H., Roche, P.A., 1995. The Genetic Privacy Act and Commentary. Boston University School of Health, Boston, MA.

Association for Molecular Pharmacology v. Myriad Genetics, Inc., 133 S. Ct. 2107, 2013.

Bobinski, J.A., 2000. Genetic information, legal, ERISA preemption, and HIPAA protection. In: Murray, T.H., Mehlman, M.J. (Eds.), Encyclopedia of Ethical, Legal, and Policy Issues in Biotechnology. John Wiley & Sons, New York.

Carson, R.A., Rothstein, M.A. (Eds.), 1999. Behavioral Genetics: The Clash of Culture and Biology. Johns Hopkins Univ. Press, Baltimore, MD.

Cavalli-Sforza, L.L., Menozzi, P., Piazza, A., 1994. The History and Geography of Human Genes. Princeton Univ. Press, Princeton, NJ.

Cook-Deegan, R.M., 1994. The Gene Wars: Science, Politics, and the Human Genome. Norton, New York.

Degler, C.N., 1991. In Search of Human Nature: The Decline and Revival of Darwinism in American Social Thought. Oxford Univ. Press, New York.

Diamond v. Chakrabarty, 444 U.S. 397, 1980.

Eisenberg, R.S., 1990. Patenting the human genome. Emory Law Journal 39, 721–745.

Greely, H.T., 1997. The control of genetic research: involving the "Groups Between". Houston Law Review 33, 1397–1430.

Greely, H.T., 2000. Iceland's plan for genomics research: facts and implications. Jurimetrics Journal 40, 153–191.

Greely, H.T., 2001. Genotype discrimination. Pennsylvania Law Review 149, 1483–1505.

Greely, H.T., Riordan, D.P., Garrison, N.A., Mountain, J.L., 2006. Family ties: the use of DNA offender databases to catch offenders' kin. Journal of Law, Medicine & Ethics 34, 248–262.

Greely, H.T., 2007. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. Annual Review of Genomics and Human Genetics 8, 343–364.

Greely, H.T., 2011. Get ready for the flood of fetal gene screening. Nature 469, 289–291.

Hvistendahl, M., 2011. Unnatural Selection: Choosing Boys over Girls, and the Consequences of a World Full of Men. Public Affairs, New York.

Institute of Medicine, U.S. National Academy of Sciences, 1993. Assessing Genetic Risks: Implications for Health and Social Policy. National Academy Press, Washington, DC.

Kaye, D.H., Greely, H.T., 2013. Brief of genetics, genomics and forensic science researchers as Amici Curiae in support of neither party. Jurimetrics 54, 43–64.

Maryland v. King, 133 S.Ct. 1958, 2013.

Mayo Collaborative Services V. Prometheus Laboratories, Inc., 132 S. Ct. 1289, 2012.

Moore v. Regents of the University of California, 51 Cal.3d 120, 1990.

National Bioethics Advisory Commission, 1997. Cloning Human Beings. NBAC, Rockville, MD.

National Research Council, 1992. DNA Technology in Forensic Science. National Academy Press, Washington, DC.

National Research Council, 1996. DNA Technology in Forensic Science: An Update. National Academy Press, Washington, DC.

Nelkin, D., Lindée, M.S., 1995. The DNA Mystique: The Gene as a Cultural Icon. W.H. Freeman, New York.

Nelson, R., 2000. Gene therapy, ethics, germ cell gene transfer. In: Murray, T.H., Mehlman, M.J. (Eds.), Encyclopedia of Ethical, Legal, and Policy Issues in Biotechnology. John Wiley & Sons, New York.

Ormond, K.E., Wheeler, M.T., Hudgins, L., Klein, T.E., Butte, A.J., Altman, R.B., Ashley, E.A., Greely, H.T., 2010. Challenges in the clinical application of whole-genome sequencing. The Lancet 375, 1749–1751.

Parens, E., 1998. Enhancing Human Traits: Ethical and Social Implications. Georgetown Univ. Press, Washington, DC.

President's Commission for the Study of Ethical Problems in Medicine and Biomedical Research, 1983. Screening and Counseling for Genetic Conditions: A Report on the Ethical, Social, and Legal Implications of Genetic Screening, Counseling and Education Programs. The Commission, Washington, DC.

Rothstein, M.A. (Ed.), 1997. Genetic Secrets: Protecting Privacy and Confidentiality in the Genetic Era. Yale Univ. Press, New Haven, CT.

Genotype and Phenotype

Richard C Lewontin, Harvard University, Cambridge, MA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 9, pp. 6159–6162, © 2001, Elsevier Ltd.

Abstract

The distinction between genotype and phenotype is the distinction between the state of genes of an organism (genotype) and the developmental manifestation of those genes in the physiology, morphology and behavior of the organism (phenotype). The phenotype is a consequence of a unique interaction between the genes of the organism and the sequence of environments in which the organism's development takes place. In addition, random events at the level of cell metabolism and cell division cause differences in phenotype between organisms that are identical genetically and have undergone the same developmental environments.

The fundamental distinction between genotype and phenotype was first introduced into biology by Wilhelm Johannsen in 1911 to make explicit a distinction that was implicit in Mendel's explanation of heredity, and which marked it off from previous theories of inheritance. Mendel found that when he crossed two pure-breeding varieties of garden pea, the first generation of hybrid offspring were indistinguishable from one of the two parental strains, but if these hybrids were crossed with each other, both of the original types appeared in the offspring of these hybrids. So, for example, a pure bred red-flowered crossed with a pure bred white-flowered form produced all red-flowered plants, but these, when interbred with each other produced both red and white-flowered plants.

In order to explain this extraordinary result, Mendel postulated the existence of internal, unseen 'factors' contributed by each parent to the offspring, factors that were causal determinants of the outward appearance of the plants, but whose nature was unaffected by their passage from one generation to the next. So, pure bred white-flowered plants had two white 'factors' one contributed by the pollen and one by the egg of their male and female parents, while pure bred red-flowered plants had received two red 'factors' from their two parents. The hybrids of red-flowered and white flowered forms would then contain one red and one white 'factor,' yet these plants were indistinguishable in appearance from a parent that had two red 'factors.' There was then a fundamental distinction between the state of the internal causal 'factors' and the outward state of the plant that carried them. Moreover, to explain the reappearance in the next generation of some white-flowered plants, Mendel had to suppose that the temporary coexistence of red and white 'factors' in the hybrids had no effect on their nature, but that they could be passed on uncontaminated to the offspring of the hybrids. An offspring of the hybrids that happened to contain two white 'factors' was then white like its white grandparent.

This scheme of explanation makes a radical separation between the internal factors that behave according to one set of physical relations in their passage between generations, the laws of *heredity*, and the external appearance of the organism that results from a different set of physiological relations which determine the consequences to the organism of carrying those factors, the laws of *development*. This distinction between the phenomenon of heredity and the phenomenon of development which we usually associate with August Weissman's separation of the germplasm from the somatoplasm had already been made by Mendel 20 years earlier. Earlier theories of heredity had not made such a separation of inner factors from outward manifestation and had been posed as the problem of the inheritance of the organism's characters themselves. It was Mendel's realization that characters are not inherited, they develop. In Johanssen's terminology the description of an organism's internal factors that were passed on to it in the gametes from its parents is a specification of the organism's *genotype*, while the description of the total physical and behavioral state of the organism is a specification of its *phenotype*.

Description of Genotypes and Phenotypes

As a result of the discoveries in molecular and cellular biology since Johanssen's time genotypic and phenotypic descriptions have become more concrete, more detailed and greatly amplified in their scope. Originally genotypes had no independent physical descriptors but were distinguished from each other entirely on the basis of the external phenotypic manifestations and the patterns of similarity in phenotypes of related individuals. 'Genes' were simply abstract placeholders of unknown physical composition, inferred entities that could be characterized only by their phenotypic effects.

The modern identification of DNA as the physical material that behaves in heredity in the manner inferred for Mendel's mysterious 'factors' has allowed a precise physical description of the genotype of an organism. It is the complete spatially-ordered sequence of the four alternative nucleotide bases, adenine, cytosine, guanine and thymine contained in the DNA of the organism's cells. In its most complete form this genotypic description includes not only the DNA in the nucleus of the cell but in certain cellular organelles such as mitochondria or chloroplasts that carry the DNA that is related to their physiological function and that is passed across generational lines. Phenotypic description has also been greatly amplified beyond the external characteristics of individuals. It now includes the gross structure of the various cells of the organism, the physico-chemical description of the various

molecules within cells, including their shape and intra-cellular localization, and the network of biochemical reactions in which these molecules participate.

The concepts of phenotype and genotype require a distinction between *types* and *tokens*. Phenotype and genotype are, as the words indicate, types, sets of which individual organisms are members, sets delineated by the complete descriptors of those organisms. Any individual member of these type sets is a token of the type. In addition it is necessary to distinguish the descriptors of organisms from the physical object or behavior that is being described. The concepts phenotype and genotype refer to the descriptors. The physical objects that are being described are the *phenome* and *genome* of an organism, the actual physical material or behavior of some particular individual and its actual DNA molecules. If we take the definitions literally, then no two organisms belong to the same phenotype because there is always some physical difference in morphology, physiology and behavior between any two organisms, even between identical twins, cloned individuals, or cells that result from asexual reproduction. Moreover, except for the results of twinning, cloning or vegetative reproduction, no two individuals belong to the same genotype, as the process of mutation guarantees some variation in their DNA between any two individuals.

The rate of mutation per nucleotide per cell division is of the order of one in a billion, so for mammals, with over a billion nucleotides of DNA making up their genes, even identical twins are likely to have some small difference in total genotype. There is, in addition, an ambiguity in the assignment of an individual to a genotype because mutations occur during the millions of cell divisions that occur in development, so that all cells in the body do not have identical DNA.

In practice phenotypic and genotypic descriptions are not total but partial, referring to some aspect of the genome and phenome that are of particular interest, while ignoring the rest. The partial phenotypic and genotypic descriptions are partial in two senses. First, only some aspect of the morphology or function of the organism is characterized; say the rate of enzymatic splitting of alcohol, a phenotype, and the DNA sequences that code for the enzyme proteins involved in the reaction, a genotype. Second, a decision must be made about the collection of actual partial phenotypes and genotypes that will be regarded as indistinguishable. At the phenotypic level, given the actual variation in enzymatic rates from individual to individual and from time to time, what range of rates will be regarded as belonging to the same phenotypic class? For any phenotypic variable that is continuous, some boundaries of distinguishable phenotypic classes must be laid down in order to make a mapping between phenotype and genotype because genotypes, by their nature, are discrete classes.

The problem of aggregation exists at the genotypic level as well. The relation of DNA differences to differences in the amino acid sequence of a protein is not one-to-one. Because the DNA triplet code is redundant, substitutions of nucleotides in the third position of the triplet often do not change the amino acid that is specified, so for some purposes many different alternative DNA sequences of a gene count as the same genotype. On the other hand, nucleotide substitutions that do not affect the amino acid sequence of a protein may affect the rate of synthesis of that protein and so may, for some questions, be counted as belonging to different genotypes.

The decision about which aspects of total phenotype and genotype are relevant in creating genotypic and phenotypic classes is a deep question of biology that is largely ignored because everything cannot be considered at once. Sometimes biologists act as if the rest of the organism really were constant, sometimes they recognize the variation but claim that it is causally irrelevant to the phenomenon of interest and sometimes they admit its causal relevance but claim that the variation can be treated as background experimental noise that can be averaged out by repeated observations on a large class of individuals. These decisions are usually made either for the sake of observational convenience without a coherent rationale or from a commitment to a reductionist view of the organism that asserts that an understanding of the whole system can be obtained by breaking it down into 'obvious' small components to be studied in isolation.

Relationship Between Genotype and Phenotype

It is clearly true that organisms do indeed consist of a collection of quasi-independent genotypic and phenotypic subsystems, within which there are strong interactions and between which there are very weak interactions. But the boundaries of these genetic and phenotypic subsystems can only be found by actual experiment and differ from function to function. The hand and eye are, as far as is known, developmentally independent systems for which different gene combinations are relevant during embryogenesis, but handeye coordination is essential for reaching out and picking up observed objects and so cannot be regarded as phenotypically or genotypically independent in considering this elementary behavior.

The problem for genetics and development is how the genotypic and phenotypic spaces map into each other. It is now certain that Weissman was correct in claiming that the genotype-phenotype mapping is causally asymmetrical. Genotypic differences may cause phenotypic differences during the process of development, but there is no material pathway by which changes in the phenome can result in related changes in the genome. There is no biological inheritance of acquired characters. This causal asymmetry sets the program for developmental, physiological and behavioral genetics which is to discover the mechanisms by which the reading of DNA by the cell machinery is causally efficacious in influencing the phenotypic state of the organism. This is the problem of making the forward mapping of genotype on to phenotype.

There remains, then, the problem of the backward mapping, how the causal knowledge of epigenetic processes can be used to infer the genotype, given the phenotype. Even if a knowledge of the genotype gave perfect information about the phenotype, the inverse would not be true, as discovered by Mendel. The phenomenon of dominance alone makes it impossible to distinguish between heterozygotes and homozygotes from phenotypic information at an arbitrary level. In addition, the same phenotype may appear as a result of changes at different locations in the DNA. There are, for example, more than 200 different nucleotide substitutions known in the genes coding for blood-clotting proteins, any one of which leads to hemophilia. The major difficulty of mapping between genotypes and

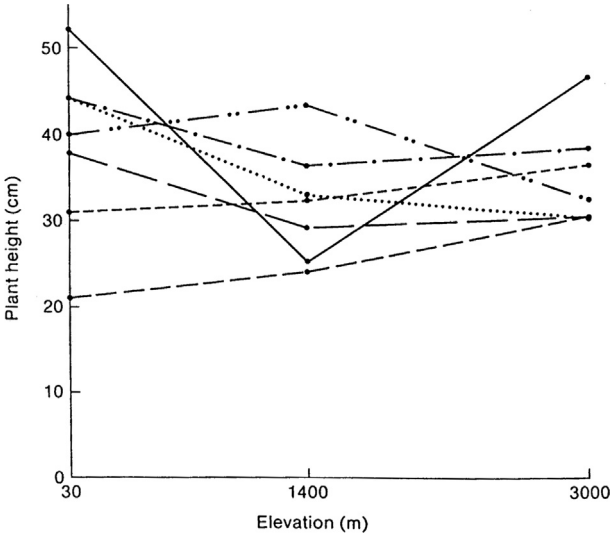

Figure 1 The growth of cloned plants of *Achillea* at three different elevations. The ordinate shows final the plant height and the abscissa the elevation. Each connected line is the pattern of growth for a different cloned genotype. Data of Clausen, J.D., Keck, D., Hiesey, W.W., 1958. Experimental Studies on the Nature of Species, vol. 3. Environmental Responses of Climatic Races of Achillea. Carnegie Institution of Washington Publication No. 581.

phenotypes is, however, in the forward direction. The complete DNA sequence of an organism does not contain the information necessary to specify that organism.

Genotype, Phenotype, and Environment

First, the outcome of development and the ongoing functioning of an individual is the consequence of a unique interaction between the molecules contained in the fertilized egg, including the DNA, and the sequence of environments in which development occurs. Figure 1 shows the result of a cloning experiment in a plant, *Achillea*, grown at three different elevations in the Sierra Nevada. Immature plants were sampled from nature, each plant was cut into three pieces, and one piece was replanted at each of the three elevations. The figure shows the height to which the pieces grew in the different environments, each line connecting the three genotypically identical pieces in the three environments. A line, giving the phenotype of a given genotype as a function of different environments is a *norm of reaction*. The norms of reaction in this case show that there is no genotype with unconditionally better growth and, in fact, there is no way to predict the order of growth of the different genotypes in one environment from their order of growth in another environment. Thus, the phenotype is not only contingent on the environment but in an unpredictable and irregular way.

Such irregular and particularistic relations between genotype and phenotype in different environments is characteristic of norms of reaction whenever they have been measured for morphological and behavioral characteristics in experimentally accessible organisms. There are, of course, extreme exceptions; genotypes that result in the same phenotype in any environment and environments that are in no way contingent on genotypes for their effect. Eye color variation among humans is entirely a consequence of genotypic differences, uninfluenced by environment, while regional and social class differences in the phonemic structure of speech have no influence of genetic differences. Moreover, major differences between species are a consequence of genetic differences that are manifest in all environments. There is no environment in which a gorilla could learn to write this article. Within species, however, the vast majority of phenotypic variations in morphology, physiology, and behavior are a consequence of particular contingent relations between genotype and phenotype.

Developmental Noise

Second, even a complete specification of the genotype and the environment is insufficient to specify the organism. All cells contain a very small number of copies of biologically active molecules, spatially segregated within the cell, so that the number of various reactions per unit time vary greatly from cell to cell and time to time. Moreover, when cells divide they do not distribute the molecules exactly evenly, except for DNA which has a special machinery for guaranteeing its equal distribution. The consequence of these small molecular numbers is that there is random variation from cell to cell in time to the next cell division, in cell size, in cell shape and in the movements of cells during development. These random events, random at the level of thermal and quantum chemical molecular states, can have considerable effect on the formation and behavior of the whole organism. (Goss and Pecoud, 1998)

So-called 'bilateral symmetries' are not, in fact, symmetrical but vary between sides in an irregular way from individual to individual. The fingerprints on the left hand are not identical, although similar to, those on the right. Such *random developmental noise* is a nontrivial source of variation in phenotype. During the development of the central nervous system of a multicellular organism, superimposed on the major regular patterns of neuronal connection are random variations in the growth and connections of specific neuronal pathways, even in organisms like the roundworm *Caenorhabditis* with a very simple central nervous system. It has yet to be determined how much of the behavioral variation among individuals is a consequence of such random variation in the development of the central nervous system, variation that is inborn, but not genetic.

The environmental and chance contingency of the mapping between genotype and phenotype arises from the nature of organisms as physical objects. They are intermediate in size between particles and planets and they are internally functionally heterogeneous. As a result, unlike simple physical systems like the movement of planets around the sun, their functions are not the consequence of a few strong causal forces whose parameters govern the outcome. Organisms are the nexus of a very large number of individually weakly determining forces subject to noise at all levels. The consequence for the understanding of the physical properties of organisms, including their individual and social behavior, is that we cannot expect any simple and universal generalizations like Newton's Laws. That is certainly true for the relation between genotype and phenotype.

See also: Body, Evolution of; Brain, Evolution of, Developmental Behavioral Genetics and Education; Human Evolutionary Genetics; Human Genome Project: History and Assessment; Intelligence, Genetics of: Cognitive Abilities.

Bibliography

Clausen, J.D., Keck, D., Hiesey, W.W., 1958. Experimental Studies on the Nature of Species, vol. 3. Environmental Responses of Climatic Races of Achillea. Carnegie Institution of Washington Publication No. 581.

Goss, P., Pecoud, J., 1998. Quantitative modeling of stochastic systems in molecular biology using stochastic petri nets. Proceedings of the National Academy of Science USA 95, 6750–6755.This page intentionally left blank

Gentrification

Loretta Lees, King's College London, Strand, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Gentrification has grown from being a small-scale inner city process pioneered by a left-liberal new middle class to being a mass-produced process around the world – 'gentrification generalized.' There are debates over definitions of the process, the kinds of theory we might use to explain and indeed predict the process, and over who is responsible for a process that most academic work is clear, despite its positive rhetoric, has a negative impact on low-income populations.

Origins of the Term

The term *gentrification* first appeared in the writings of the British Marxist sociologist Ruth Glass in 1964, although Glass is said to have used the term earlier than this in an unpublished study of housing in North Kensington in 1959, and an early reference to the word 'gentrification' can be found in the 1888 *Memoirs and Proceedings of the Manchester Literary & Philosophical Society* (p. 39). In coining the term 'gentrification,' Glass was describing a new and somewhat peculiar process that she was witnessing in areas of inner London such as Kensington, Chelsea, and Islington:

> One by one, many of the working class quarters of London have been invaded by the middle classes – upper and lower. Shabby, modest mews and cottages – two rooms up and two down – have been taken over, when their leases have expired, and have become elegant, expensive residences. Larger Victorian houses, downgraded in an earlier or recent period – which were used as lodging houses or were otherwise in multiple occupation – have been upgraded once again. Nowadays, many of these houses are being subdivided into costly flats or 'houselets' (in terms of the new real estate snob jargon). The current social status and value of such dwellings are frequently in inverse relation to their status, and in any case enormously inflated by comparison with previous levels in their neighbourhoods. Once this process of 'gentrification' starts in a district it goes on rapidly until all or most of the original working class occupiers are displaced and the social character of the district is changed. (Glass, 1964: xviii–xix)

Although Ruth Glass coined the term *gentrification* as we know it in 1964, Eric Clark (2005: 260) has argued that "it is careless to turn this into an assumption that we have here the origin of the phenomenon." In fact, Neil Smith (1996: 34–40) had already argued that earlier processes, such as the Haussmannization of Paris, when Baron Haussmann demolished the poor residential areas of central Paris to build new, wealthy tree-lined boulevards to show case the city's famous monuments were examples of gentrification. Gale (1984) also pointed out that gentrification was happening in parts of New York City, New Orleans, Charleston, and the Georgetown area of Washington DC, in the late 1930s. Lees et al. (2008: 5) argued, however, that "the emergence of gentrification proper … began in postwar advanced capitalist cities. It's earliest systematic occurrences were in the 1950s in large metropolitan cities like Boston, Washington DC, London and New York City." The process also had different names in different places: in New York City it was called brownstoning; in Baltimore, homesteading; in Toronto, whitepainting or whitewalling; and in San Francisco, red-brick chic. As Williams (1986: 65) argued: "Many American analysts have been uncomfortable with the term 'gentrification' (with its obvious class connotations), preferring instead labels such as 'back-to-the-city movement', 'neighborhood revitalization', and 'brownstoning', all of which were indicative of underlying divergences in what was believed to be central to this process."

The Stages/Waves of Gentrification

Stage models of gentrification, based on the ecological models of the time, began to develop in the 1970s and 1980s both to explain the process and to predict the future course of gentrification. These models were based on ecological ideas about invasion and succession, ideas that have long been reflected in definitions of gentrification: "The process of gentrification is often similar to that of *invasion and succession*. A few gentrifiers obtain properties in a relatively run-down condition within a small area and improve them, thereby increasing the attractiveness of the area to others who would prefer such a location, so that eventually the entire area (often only a few streets) changes its socio-economic status, and property values are substantially enhanced. Real estate agents and property developers may participate in the process, as they seek to enhance the exchange value of an area and to reap substantial profits from promoting *uneven development* at the intra-urban scale" (The Dictionary of Human Geography, 1994: 216–217, emphases in original).

Clay (1979) was one of the first to produce a schema of the different stages of gentrification. He based his 'predictive' model on observations and data from a number of US cities, including Boston's South End, Philadelphia's Society Hill, San Francisco's Western Addition, and Washington's Capitol Hill. Clay's model moved from Stage 1 (where risk-oblivious 'pioneer' gentrifiers moved into historic neighborhoods, which were 'red-lined,' and renovated property through 'sweat equity' causing little displacement at that time), through Stage 2 (when similar people moved in and small-scale speculation began, more displacement ensued, and neighborhood names were often changed), through Stage 3 (when media and official interest began, developers became interested, displacement continued,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.74013-X

new incomers were much less tolerant of the preexisting community, and banks began to 'green-line' the area), and finally, to Stage 4 (when the process matured and the neighborhood was stabilized, the business and managerial classes moved in rather than the professional classes, rent and sales prices of property were now very high). Writing at the same time, Gale (1979) also designed a stage model of gentrification based on research in Washington DC, he was especially interested in the class and status distinctions between old and new residents in a gentrifying neighbourhood and in displacement of former working-class residents. The differences between Clay's and Gale's stage models of gentrification indicate the different emphases that different researchers place on analyses of gentrification. However, as the process of gentrification developed, it became apparent that models such as these needed to be updated and were losing their predictive value. Hackworth and Smith (2001) produced a more updated model that divided gentrification into three different waves running from 1968 to 1999, divided by two transition phases. The model demonstrated 'the changing state' of gentrification. First-wave gentrification occurred between 1968 and 1972 (this is akin to what Clay and Gale called pioneer gentrification). The global oil crisis of 1973 triggered a transition phase that lasted until 1978 during which gentrification continued and set the stage for a second-wave gentrification from 1978 to 1989, when the process anchored itself in neighborhoods and triggered both displacement and political struggles. Then another transition phase was triggered by recession until 'post-recession' or third-wave gentrification kicked off in 1993/1994. By now, the process had attracted large-scale capital, large-scale developers, and the state. Persuasive as this model was, Lees (2003) showed that mutations of gentrification, such as supergentrification, were already beginning to contradict such models. Supergentrifiers set out to regentrify already gentrified neighbourhoods, such as Brooklyn Heights in New York City, demonstrating no mature or end stage to the process of gentrification.

Definitions, Types, and Shifting Terms of the Debate

Gentrification is a slippery term; definitions and typologies have changed as the process has evolved. Early definitions of gentrification were closely aligned to Glass (1964) description, for example: "By gentrification I mean the process by which working class residential neighborhoods are rehabilitated by middle class homebuyers, landlords and professional developers. I make the theoretical distinction between gentrification and redevelopment. Redevelopment involves not rehabilitation of old structures but the construction of new buildings on previously developed land" (Smith, 1982: 139). But as developers became more involved in gentrification it was apparent that the residential rehabilitation that Ruth Glass had described was only one facet of the gentrification process. As Sassen (1991: 255) explained: "Gentrification was initially understood as the rehabilitation of decaying and low-income housing by middle-class outsiders in central cities. In the late 1970s a broader conceptualization of the process began to emerge, and by the early 1980s new scholarship had developed a far broader meaning of gentrification, linking it with processes of spatial, economic and social restructuring. Gentrification emerged as a visible spatial component of this transformation. It was evident in the redevelopment of waterfronts, the rise of hotel and convention complexes in central cities, large-scale luxury office and residential developments, and fashionable, high priced shopping districts." The types of gentrification that Sassen mentions go well beyond the classic rehabilitation of residential historic properties, to include rehabilitation of nonresidential properties and new-build development. As such, new types of gentrification became parlance–retail gentrification, commercial gentrification or boutiqueification (Bridge and Dowling, 2001; Zukin, 2010), new-build gentrification (Davidson and Lees, 2005, 2010), tourist gentrification (Gotham, 2005), and so on.

Some authors, though, questioned the definition of gentrification earlier than this. The term *rural gentrification* can be traced back to Parsons (1980). It refers to the gentrification of rural areas, describing the link between new middle-class settlement, socioeconomic and cultural transformations of the rural landscape, and the subsequent displacement and marginalization of low-income groups (Phillips, 1993). Phillips (2004) documents the parallel between rural and urban gentrification; both reflect distanciation from the suburbs. The process in the United Kingdom has also been called *greentrification* (Smith and Phillips, 2001) and in the United States *wilderness gentrification* (Darling, 2005): "It is a different story in the wilderness, largely because of the kind of rent that is being capitalized. What gets produced in the process of urban gentrification is residential space. What gets produced in the process of wilderness gentrification is recreational nature" (p. 1022). Phillips (2004) writes persuasively about the need to widen the spatial lens of gentrification studies, and Hines (2010) does this in his discussion of the colonization of rural communities by the ex-urban middle class. Mills (1988) also questioned previous definitions of gentrification in her study of the newly built postmodern landscape of Fairview Slopes, a small neighborhood in inner city Vancouver: "Yet Fairview Slopes does not fit the usual image of a gentrified neighborhood. It is a landscape of redevelopment, and renting is probably still as common as owner occupancy" (p. 186).

Others have produced more focused descriptions of certain gentrifiers associated with certain types of gentrification, for example, the students who are 'proto-gentrifiers' in Darren Smith's (2005) 'studentification' thesis. Studentification refers to the process of social, environmental, and economic changes effected by large number of students (caused by the massive expansion of higher education in the United Kingdom) invading particular areas of the cities and towns in which popular universities are located. Though originating in Britain, studentification has also been adopted in American English (see also Davison, 2009 on Australia) to refer to similar situations in many US college towns. Increasingly though, attention has turned to the role of universities themselves as gentrifiers (see http://www.blackballot.com/harlem-gentrification/columbia-university-and-the-gentrification-of-harlem) – an interesting parallel to the evolution in the process from individual gentrifiers to institutional gentrifiers.

Lees (2003) discusses the supergentrifiers who have led the process of supergentrification or "financification" in certain neighbourhoods that have become the focus of intense

investment and conspicuous consumption by a new generation of super-rich "financifiers" fed by the fortunes from the global finance and corporate service industries. Butler and Lees's (2006) position revolves around Sassen's (1991) argument about the creation of a new class of financial engineers who have successfully commodified the financial services industries, creating new products and great wealth for themselves. The spread of this industry has been such that it has also included those in supporting industry sectors such as marketing, IT, and crucially legal services. It is these very wealthy, 'masters of the universe' who enact supergentrification in neighborhoods, like Barnsbury in London, which have already entered a stage of mature gentrification. Supergentrification is an interesting phenomenon because it poses important questions about the historical continuity of current manifestations of gentrification with previous rounds of neighbourhood change. It also stands against neo-Marxist rent gap–type explanations (see below), which focus on the dialectic of disinvestment and reinvestment, for here the neighbourhood is not disinvested.

Gentrification then is "a highly dynamic process, it is not amenable to overly restrictive definitions; rather than risk constraining our understanding of this developing process by imposing definitional order, we should strive to consider the broad range of processes that contribute to this restructuring, and to understand the links between seemingly separate processes" (Smith, 1986: 3).

Key Theories and Debates

In the early days, gentrification studies were dominated by neoclassical economic theory, and this approach continues today (e.g., several economists have refined bid-rent models to chart gentrification and other shifts in the contours of urban spatial structure–e.g., Brueckner et al., 1999). Social scientists, however, for the most part were not enamored of such an approach. As such, by the 1980s and 1990s, the pages of urban studies journals were dominated by two different theoretical framings of gentrification: a Marxist theoretical camp led by geographer Neil Smith and a humanist (or cultural postmodernist) theoretical camp led by geographer David Ley. The debates were also framed as production versus consumption, economic versus cultural, and structure versus agency (for summaries of these debates and key readings associated with them, see Lees et al., 2008, 2010 and Brown-Saracino, 2010). Neil Smith, like Ruth Glass, was a Marxist, but he was a geographer and not a sociologist, as such he was more interested in space and spatiality with respect to the gentrification process. Smith drew on Marxist notions of uneven development to theorize gentrification; he was interested in the 'locational seesaw' of capital as it was reflected in the successive development, underdevelopment, and redevelopment of spaces in the city. Smith showed how capital in the twentieth century jumped about the city looking for places in which to create profit, first in the suburbs and then in the inner city. Smith (1979) summarized the microdynamics of uneven development in relation to gentrification through his theory of the rent gap. His rent gap theory described the difference between capitalized ground rent – the rent captured under present land use–and potential ground rent – the maximum rent that could be attained if the land were put to a better or higher use. Disinvestment was simply urban creative destruction. In urban areas, new developments (like suburbanization) undermined older investments (in the central city), and land/property owners saw a poor return on sinking more capital into aging property in inner cities. There was a clear spatial imprint – the old inner city and the new suburbs. But as profit margins began to diminish in the suburbs, capital moved back to the city in a search for value. As such, certain neighbourhoods became ripe for gentrification: "The movement of capital into suburban development led to a systematic devalorization of inner and central city capital, and this, in turn, with the development of the rent gap, led to the creation of new investment opportunities in the inner city precisely because an effective barrier to new investment had previously operated there" (Smith, 1982: 149). In inner London as landlords saw the money that could be made from evicting tenants and selling their disinvested properties on to developers – some resorted to winkling (also known as Rachmanism) to get tenants out (see Lees, 1994a).

The rent gap thesis, however, involves concepts that are very hard to measure: nothing like capitalized ground rent appears in any public database; as such the researcher has to construct specialized indicators. For example, Smith et al. (1989) used tax arrears data on the Lower East Side in New York City. Hammel (1999) talks through some of the difficulties with the thesis and how the integration of the rent gap with theories of scale resolves a number of crucial difficulties. However, debate over the rent gap has been complicated by the publication of similar hypotheses – for example, Hamnett and Randolph's (1986) theory of a value gap driving the conversion of London rental flats into owner-occupation (and gentrification), and Sykora's (1993) notion of a functional gap describing the mismatch between urban core land uses under state-socialist conditions in Prague, as market conditions created a land-market gradient in the early 1990s. It became obvious how important local context was to theorizations of uneven development.

David Ley, a humanist geographer, developed new ideas about the "postmodern condition" in his "post-industrial thesis" on gentrification. He was especially taken with American sociologist Daniel Bell's (1973) *The Coming of the Post-Industrial Society*. Ley (1996) argued that a "cultural new class," an expanding middle class of professionals, had instigated gentrification, not capital per se. This new middle class could be identified by their left-liberal politics, lifestyles, consumption habits, aesthetics, and tastes. This new class desired an alternative to the hegemony of postwar suburbanization – urbanism. They were pro-urban, and Ley argued that gentrification represented an "imagineering of an alternative urbanism to suburbanization" (p. 15) and that this could not be captured by explanations of the process that privileged the structural forces of production and housing market dynamics. Ley's pioneer gentrifiers were politically progressive hippies who eventually morphed into more corporate type gentrifiers – or yuppies. The crucial point, as Rose (1984: 56) stated: "[was] that 'gentrifiers' are not the mere bearers of a process determined independently of them. Their constitution, as certain types of workers and as people, is as crucial an element in the production of gentrification as is the production of the dwellings they occupy."

Canadian sociologist Jon Caulfield (1994) interviewed this cultural new class in Toronto and found that gentrification was a form of resistance not just to suburbia but also to the large-scale inner city urban renewal of the 1960s. British sociologist, turned geographer, Tim Butler's (1997) work on London detailed well how this new middle class was different from the traditional middle class and indeed the upper class – in terms of their pro-urban, pro-social diversity, left-liberal politics, and their education and professions. In addition, Butler (1997) showed how gentrification in Hackney was different from gentrification in other areas of London, for example, Islington. Butler with Robson (2003) continued this theme when they looked at six gentrified neighbourhoods in London and found that each of them formed a different 'metropolitan habitus.' They showed how gentrifiers ensured their hegemony over the localities in which they settled and argued that gentrification should be seen as a middle class "coping strategy" (p. 27) in unstable economic times when the middle classes were facing various structural constraints and where the inner city did not provide everything they desired – for example, quality (middle-class) schools.

The proliferation of work on gender and gentrification (on social reproduction, on dual career households, etc.), and on sexuality and gentrification (gay gentrification) was similar to Ley's thesis (see Lees et al., 2008: Chapter 3). As Ann Markusen (1981) summarized: "[G]entrification is in large part a result of the breakdown of the patriarchal household. Households of gay people, singles, and professional couples with central business district jobs increasingly find central locations attractive … Gentrification … corresponds to the two-income (or more) professional household that requires both a relatively central urban location to minimize journey-to-work costs of several wage earners and a location that enhances efficiency in household production (stores are nearer) and in the substitution of market-produced commodities (laundries, restaurants, child care) for household production" (p. 32).

By the mid-1990s, however, researchers had become frustrated with these (arguably) polarized theoretical views (what became known as the gentrification battleground) and began to adopt a more complementary approach (see Clark, 1992; Lees, 1994b). In the late 1990s, attention moved away from theorizing the process toward investigating the influence of gentrification on urban policies and looking at what Martin Phillips (2004) has called gentrification's others. Today, British and North American researchers are focused more on the effects, rather than the causes, of gentrification and on finding alternatives. We can say without any hesitation, however, that gentrification is an economic, cultural, political, social, and institutional phenomenon and that various actors have been and are involved, from individuals to private institutions to the state.

The State Gets Involved

Over time, the state in Europe and North America has become increasingly involved in gentrification. In the early days of gentrification, as is seen in the stage models discussed above, the state acted as a facilitator of gentrification. For example, in London in the 1970s, the state provided home improvement grants (Lees, 1994a; Lees et al., 2008: Chapter 1) to aid gentrification, and in New York City in the 1980s and 1990s the state undertook more revanchist activities, such as zero tolerance initiatives and eviction of the homeless and drug dealers from the Lower East Side (Smith, 1996). By the late 1990s, however, the state was beginning to emerge as a major player in gentrification, and state-led gentrification began to affect neighborhoods in cities across North America and Europe (see Lees et al., 2008: Chapter 5).

In the United States, the Federal Department of Housing and Urban Development (HUD) set about the state-led gentrification of public housing through their HOPE VI program. Drawing on poverty deconcentration arguments, HUD was able to sell gentrification through the concept of mixed communities/social mixing (Lees et al., 2008). They began to replace existing high-rise, high-density social housing projects with new, lower density, mixed-income communities. An infamous example is Cabrini Green in Chicago. In 1994, despite being located next to some of the most expensive real estate in Chicago, Cabrini Green qualified, under HUD guidelines, as the worst case of public housing in the United States and received $50 million to redevelop a portion of the site (Lees et al., 2008). By 'manufacturing' a socially mixed community in areas of poverty concentration, such as Cabrini Green, US policy makers thought that gentrification would

> ameliorate the social isolation of the poor. New more affluent residents will rub shoulders with poorer existing residents on the streets, in shops, and within local institutions, such as public schools. Such newcomers may exhibit possibilities of social mobility and a determination to secure adequate public services that provide existing residents with the kind of role models and contacts the absence of which Wilson [W.J. Wilson, 1987] finds debilitating in the ghetto. (Byrne, 2003: 422)

New evidence collated on mixed communities in Bridge et al. (2011) shows that such a positive scenario is rarely the case. Hyra (2008) has tellingly labeled HOPE VI "the new urban renewal," referring to the displacement of the urban poor as formerly notorious ghettos became dotted with expensive restaurants, health spas, and chic boutiques.

State-led gentrification in the form of mixed communities policy was to emerge not just in the United States, but also in the United Kingdom, France, the Netherlands, Ireland, Canada, Australia, among other countries.

In the Netherlands, a policy of housing redifferentiation began in 1996. Housing redifferentiation was a policy of adding more expensive dwellings to low-income areas by removing inexpensive dwellings through demolition, together with the sale and upgrading of existing dwellings. The idea was to create a more socially mixed population in neighbourhoods via gentrification (see Hulsbergen and Stouten, 2001; Uitermark, 2003). Uitermark et al. (2007) are clear that an influx of middle-class residents into a disadvantaged neighbourhood does not increase social cohesion; rather the contacts between low-income and higher income households tend to be superficial at best and downright hostile at worst.

In the United Kingdom, the New Labor government initiated an Urban Task Force that was especially excited by the Dutch

policy, as it was by HOPE VI. The Urban Task Force Report stated: "In responding to social problems we must avoid repeating the mistakes of the past. Developing large amounts of social housing in one location does not work. Many existing social housing estates have a strong sense of community – often more so than wealthier neighbourhoods – but there is not the economic capacity to make these neighbourhoods work over the long term. As a result, jobs and investment go elsewhere, exacerbating the physical isolation of many of these estates. In future, we must develop on the basis of a mix of tenures and income groups" (DETR, 1999: 45). The Urban White Paper (DETR, 2000) that followed implemented state-led gentrification in the United Kingdom in the guise of urban sustainability. The British government's intention in the late 1990s was to bring the middle classes back to the central city, read gentrification. This was marketed as an attempt to reduce socio-spatial segregation and strengthen the social tissue of deprived neighborhoods; they sold gentrification as something positive (see Cameron, 2003 on positive gentrification or gentrification as a positive public policy tool).

In the 2000s state-led gentrification in Europe and North America (and subsequently further afield) was given a boost by Richard Florida's (2003) creative city thesis: "… since the image of the 'livable city' has become a key aspect of a city's ability to compete in a globalized, knowledge-based economy (Florida, 2003), post-industrial cities have a growing interest in marketing themselves as being built on a foundation of 'inclusive' neighbourhoods capable of harmoniously supporting a blend of incomes, cultures, age groups and lifestyles …." (Rose, 2004: 281). Florida's work on the "creative class" allowed policy makers to promote gentrification while describing their strategies in a new, popular language summarized in the three T's of technology, talent, and tolerance. The creative class that Florida discusses is none other than Ley's (1996) cultural new class described earlier, even if they seem more interested in technology now! State-led gentrification did not just travel in terms of mixed communities/social mix policy but also increasingly so in terms of creative city policy. As Jamie Peck (2005: 740–741) has noted, "In the field of urban policy, which has hardly been cluttered with new and innovative ideas lately, creativity strategies have quickly become the policies of choice, since they license both a discursively distinctive and an ostensibly deliverable development agenda. No less significantly, though, they also work quietly with the grain of extant 'neoliberal' development agendas, framed around interurban competition, gentrification, middle-class consumption and place-marketing." Creative city strategies "presume and work with gentrification, conceived as a positive urban process, while making a virtue of selective and variable outcomes, unique neighborhood by unique neighborhood. And with almost breathtaking circularity, it is now being proposed that these gentrification-friendly strategies should be evaluated, not according to hackneyed metrics like job creation or poverty alleviation, but according to more relevant measures like… increased house prices! … As if this were not circular enough, it is increasingly common for cities to evaluate the effectiveness of their creativity strategies according to their shifting position in Florida's league tables" (Peck, 2005: 764).

Policies of state-led gentrification are intricately tied up with the globalization of gentrification. But as yet it is not proven whether policy transfers and mobilities brought gentrification to cities as far away as Shanghai, Tapei, Cape Town, or Santiago, or whether gentrification in these other places was home grown (Lees, 2012). As Clark (2005) says: "confident proclamations ring out: Gentrification is now global! The problem with this is not if gentrification can be observed in places around the world, but it is again an issue of time: it is now global … The extent of occurrence of the phenomenon from a global historical perspective remains however largely uncharted" (p. 260).

The Globalization of Gentrification

Gentrification traveled beyond the inner city to rural areas and the suburbs; it developed architecturally and aesthetically beyond the historic to the new; it emerged in metropolitan cities, provincial cities, and cities further down the urban hierarchy; but importantly gentrification also went global – it spread beyond North American and western European cities to locations all around the world. Neil Smith (2002) was perhaps the first gentrification researcher to highlight the relationship between globalization, neoliberalism, and the changing role of the state in gentrification. Smith (2002) argued that gentrification is now a global urban strategy linked to a new globalism and a related new urbanism. He made two central arguments: first, that the neoliberal state was now the main agent of gentrification worldwide, and second, that gentrification had gone global and had become what he called gentrification generalized, a strategy connected into the circuits of global capital and cultural circulation. Atkinson and Bridge (2005) followed suit arguing that gentrification was pressurizing cities and neighbourhoods around the globe, and that processes of class colonization were dislocating the urban poor. They claimed: "Gentrification is now global. It is no longer confined to Western cities" (p. 1), and they critiqued this as a 'new form of urban colonialism.' They asked some important questions, for example: "does the fact that gentrification has moved into very different urban contexts – rapidly urbanizing, postcolonial, postcommunist, or commune-capitalist all overlaid by a diversity of cultural and religious forms – inflect, or indeed generate, a very different process?" (p. 1). Unfortunately, their collection of essays did not really answer this question. More recently, Lees (2012) has asked such questions in reference to the new literature on comparative urbanism. She argues that if we take on board ideas from comparative urbanism, we must not assume that gentrification emerged later in non-Western cities, nor that the types of gentrification there are the same as in the West/Global North. Earlier it was stated that the term *gentrification*, with its references to the British gentry and social class system, was not seen as appropriate in the US context. Indeed, as the process has gone global, the process itself has different terms – for example, gentrification began in Finland around 2000 and is called *syrjäytyneiden akanvirta* meaning 'migration of the disadvantaged.' The question Lees (2012) asks is, how important are language and meaning with respect to gentrification? She states that we should avoid the assumption that gentrification is simply 'expanding' toward the global periphery, as if it is an imported new phenomenon that builds upon emerging real estate markets (Peck, 2010). Indeed, it

may be possible to hypothesize that gentrification has taken place in various guises in different parts of the Global South since the 1960s. It may be possible that only the systematic and social scientific observation of these processes is something new, but not the actual phenomena themselves. In this regards, the experiences in the Global South deserve careful attention and revised conceptualization beyond the usually accepted definitions of gentrification outlined here that are confined to the Global North.

Conclusion

Over the past decade, we have seen the rapid and visceral emergence of state-led gentrification in the Global South: processes of gentrification are now changing the centers of cities in China, India, Pakistan, South America, and South Africa (among others). Gentrification began to take off in the Global South (or at least it began to attract the attention of the media and certain academics) at the turn of the twenty-first century, but even then the geographies of a global gentrification presented (e.g., Atkinson, 2003; Atkinson and Bridge, 2005) all but omitted the Global South in any meaningful way. Despite a good discussion in Atkinson and Bridge (2005) about gentrification as a form of neocolonialism – the White Anglo appropriation of the central city – there was no discussion about appropriate theory to analyze either this or how it might play out differently in the predominantly nonwhite cities of the Global South.

What is now required of the gentrification literature is a comparative imagination that can respond to the postcolonial challenge of 'decentering the reference points for international scholarship' (Robinson, 2006: 169), and this will have implications for how gentrification is conceived (questioning the usefulness and applicability of the term *gentrification* in the Global South and indeed elsewhere) and how research is to be conducted (this will push researchers to learn new kinds of urbanism and involve multiple translations throughout the world). Importantly, it entails unlearning the existing dominant literature from the Global North/the West that continues to structure how we think about gentrification, its practices and ideologies.

Bibliography

Atkinson, R. (Ed.), 2003. The gentry in the city: upward neighbourhood trajectories and gentrification. Urban Studies 40 (12), 2343–2584 (special issue).

Atkinson, R., Bridge, G. (Eds.), 2005. Gentrification in a Global Context: The New Urban Colonialism. Routledge, London.

Bell, D., 1973. The Coming of the Postindustrial Society. Basic Books, New York.

Bridge, G., Butler, T., Lees, L. (Eds.), 2011. Mixed Communities: Gentrification by Stealth? Policy Press, Bristol.

Bridge, G., Dowling, R., 2001. Microgeographies of retailing and gentrification. Australian Geographer 32 (1), 93–107.

Brown-Saracino, J. (Ed.), 2010. The Gentrification Debates. Routledge, New York. Brown-Saracino, J., 2010. In: The Gentrification Debates. Routledge, New York.

Brueckner, J.K., Thisse, J.F., Zenou, Y., 1999. Why is central Paris rich and downtown Detroit poor? an amenity-based theory. European Economic Review 43, 91–107.

Butler, T., 1997. Gentrification and the Middle Classes. Ashgate, Aldershot.

Butler, T., Lees, L., 2006. Super-gentrification in Barnsbury, London: globalisation and gentrifying global elites at the neighbourhood level. Transactions of the Institute of British Geographers 31, 467–487.

Butler, T., Robson, G., 2003. London Calling: The Middle Classes and the Remaking of Inner London. Berg, London.

Byrne, J.P., 2003. Two cheers for gentrification. Howard Law Journal 46 (3), 405–432.

Cameron, S., 2003. Gentrification, housing redifferentiation and urban regeneration: 'Going for Growth' in Newcastle upon Tyne. Urban Studies 40 (12), 2367–2382.

Caulfield, J., 1994. City Form and Everyday Life: Toronto's Gentrification and Critical Social Practice. University of Toronto Press, Toronto, Canada.

Clark, E., 2005. The order and simplicity of gentrification – a political challenge. In: Atkinson, R., Bridge, G. (Eds.), Gentrification in a Global Context: The New Urban Colonialism. Routledge, London, pp. 256–264.

Clay, P., 1979. Neighborhood Renewal: Middle-Class Resettlement and Incumbent Upgrading in American Neighborhoods. Heath and Company, Lexington, MA.

Darling, E., 2005. The city in the country: wilderness gentrification and the rent-gap. Environment and Planning A 37 (6), 1015–1032.

Davidson, M., Lees, L., 2005. New build "gentrification" and London's riverside renaissance. Environment and Planning A 37 (7), 1165–1190.

Davidson, M., Lees, L., 2010. New-build gentrification: its histories, trajectories, and critical geographies. Population, Space and Place 16, 395–411.

Davison, G., 2009. Carlton and the campus: The university and the gentrification of inner Melbourne 1958–75. Urban Policy and Research 27, 3.

DETR, 1999. Towards an Urban Renaissance, London.

DETR, 2000. Our Towns and Cities – the Future. Delivering an Urban Renaissance, London.

Dictionary of Human Geography, 1994.

Florida, R., 2003. The Rise of the Creative Class. Basic Books, New York.

Gale, D.E., 1979. Middle class resettlement in older urban neighborhoods: the evidence and the implications. Journal of the American Planning Association 45, 293–304.

Gale, D.E., 1984. Neighborhood Revitalization and the Postindustrial City: A Multinational Perspective. Heath and Company, Lexington, MA.

Glass, R., 1964. 'Introduction: aspects of change' in Centre for Urban Studies. In: London: Aspects of Change. MacKibbon and Kee, London.

Gotham, K.F., 2005. Tourism gentrification: the case of New Orleans' Vieux Carre (French Quarter). Urban Studies 42 (7), 1099–1121.

Hackworth, J., Smith, N., 2001. The changing state of gentrification. Tijdschrift voor Economische en Sociale Geografie 22, 464–477.

Hammel, D.J., 1999. Re-establishing the rent gap: an alternative view of capitalized land rent. Urban Studies 36 (8), 1283–1293.

Hamnett, C., Randolph, B., 1986. Tenurial transformation and the flat break-up market in London: the British condo experience. In: Smith, N., Williams, P. (Eds.), Gentrification of the City. Allen and Unwin, London, pp. 121–152.

Hines, D., 2010. In pursuit of experience: the postindustrial gentrification of the rural American West. Ethnography 11 (2), 285–308.

Hulsberg, E., Stouten, P., 2001. Urban renewal and regeneration in the Netherlands: integration lost or subordinate? City 5 (3), 325–337.

Hyra, D., 2008. The New Urban Renewal: The Economic Transformation of Harlem and Bronzeville. The University of Chicago Press, Chicago, IL.

Lees, L., 1994a. Gentrification in London and New York: an Atlantic gap? Housing Studies 9 (2), 199–217.

Lees, L., 1994b. Rethinking gentrification: beyond the positions of economics and culture. Progress in Human Geography 18 (2), 137–150.

Lees, L., 2003. Super-gentrification: the case of Brooklyn Heights, New York city. Urban Studies 40 (12), 2487–2509.

Lees, L., 2008. Gentrification and social mixing: towards an urban renaissance? Urban Studies 45 (12), 2449–2470.

Lees, L., 2012. The geography of gentrification: thinking through comparative urbanism. Progress in Human Geography 36 (2), 155–171.

Lees, L., Slater, T., Wyly, E., 2008. Gentrification. Routledge, New York.

Lees, L., Slater, T., Wyly, E. (Eds.), 2010. The Gentrification Reader. Routledge, London.

Ley, D., 1996. The New Middle Class and the Remaking of the Central City. Oxford University Press, Oxford.

Manchester Literary and Philosophical Society, 1888. Memoirs and Proceedings of the Manchester Literary & Philosophical Society.

Markusen, A., 1981. City spatial structure, women's household work, and national urban policy. In: Stimpson, C., Dixler, E., Nelson, M.J., Yatrakis, K.B. (Eds.), Women and the American City. University of Chicago Press, Chicago, IL, pp. 20–41.

Mills, C., 1988. Life on the upslope: the postmodern landscape of gentrification. Environment and Planning D: Society and Space 6, 169–189.

Parsons, D., 1980. Rural Gentrification: The Influence of Rural Settlement Planning Policies. Department of Geography Research Paper 3. University of Sussex, Brighton, UK.

Peck, J., 2005. Struggling with the creative class. International Journal of Urban and Regional Research 29 (4), 740–770.

Peck, J., 2010. Constructions of Neoliberal Reason. Oxford University Press, New York.

Phillips, M., 2004. Other geographies of gentrification. Progress in Human Geography 28 (1), 5–30.

Phillips, M., 1993. Rural gentrification and the processes of class colonization. Journal of Rural Studies 9, 123–140.

Robinson, J., 2006. Ordinary Cities: Between Modernity and Development. Routledge, London.

Rose, D., 1984. Rethinking gentrification: beyond the uneven development of Marxist urban theory. Environment and Planning D: Society and Space 1, 47–74.

Rose, D., 2004. Discourses and experiences of social mix in gentrifying neighbourhoods: a Montreal case study. Canadian Journal of Urban Research 13 (2), 278–316.

Sassen, S., 1991. The Global City: New York, London and Tokyo. Princeton University Press, Princeton, NJ.

Smith, D., 2005. Studentification: the gentrification factory. In: Atkinson, R., Bridge, G. (Eds.), Gentrification in a Global Context: The New Urban Colonialism. Routledge, London, pp. 72–89.

Smith, D., Phillips, D., 2001. Socio-cultural representations of greentrified Pennine rurality. Journal of Rural Studies 17, 457–469.

Smith, N., 1979. Toward a theory of gentrification: a back to the city movement by capital, not people. Journal of the American Planning Association 45 (4), 538–548.

Smith, N., 1982. Gentrification and uneven development. Economic Geography 58 (2), 139–155.

Smith, N., 1986. Gentrification, the frontier, and the restructuring of urban space. In: Smith, N., Williams, P. (Eds.), Gentrification of the City. Allen & Unwin, London, pp. 15–34.

Smith, N., 1996. The New Urban Frontier: Gentrification and the Revanchist City. Routledge, London and New York.

Smith, N., 2002. New globalism, new urbanism: gentrification as global urban strategy. Antipode 34 (3), 427–450.

Smith, N., Duncan, B., Reid, L., 1989. From disinvestment to reinvestment: tax arrears and turning points in the East Village. Housing Studies 4 (4), 238–252.

Sykora, L., 1993. City in transition: the role of the rent gap in Prague's revitalization. Tijdschrift voor Economisce en Sociale Geografie 84 (4), 281–293.

Uitermark, J., 2003. 'Social mixing' and the management of disadvantaged neighbourhoods: the Dutch policy of urban restructuring revisited. Urban Studies 40 (3), 531–549.

Uitermark, J., Duyvendak, J., Kleinhans, R., 2007. Gentrification as a governmental strategy: social control and social cohesion in Hoogvliet, Rotterdam. Environment and Planning A 39 (1), 125–141.

Williams, P., 1986. Class constitution through spatial reconstruction? A re-evaluation of gentrification in Australia, Britain and the United States. In: Smith, N., Williams, P. (Eds.), Gentrification of the City. Unwin Hyman, London, pp. 56–77.

Zukin, S., 2010. Naked City: The Death and Life of Authentic Urban Places. Oxford University Press, Oxford.

Geoarchaeology

Marshall I Weisler and Serena Love, School of Social Science, University of Queensland, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Geoarchaeology uses the techniques, methods, and concepts of the physical sciences to address archaeological questions. Examined here are the major techniques of geoarchaeology that are geared toward: discovering archaeological sites and documenting their internal structure; surveying site formation and disturbance processes; the analysis of soils and sediments; paleoenvironmental reconstruction and the impact of humans on the landscape; the physical analysis of archaeological materials; and the integration of geoarchaeology with social archaeology. Geoarchaeology must be integrated into research programs at the design, excavation, and analytical stages to be most advantageous.

Research that uses the techniques, methods, and concepts of the earth sciences to address archaeological problems is called geoarchaeology. In the important early exposition, *Archaeology as Human Ecology: Method and Theory for a Contextual Approach* (1982), Karl Butzer described geoarchaeology as a collection of techniques for contributing substantive data critical to understanding prehistoric and historic human ecosystems (1982: p. xi). Although geoarchaeology is primarily a methodology, its ultimate objective is to interpret the human past, through an understanding of natural, physical processes that are involved in the formation of archaeological sites.

There are at least eight major concerns of geoarchaeology that must be integrated into research programs at the design, excavation, and analytical stages and include (1) site survey techniques that use geochemical, electromagnetic, and other remote sensing procedures to locate sites or features within a known site; (2) documentation of site-formation processes and the regional context of a site; (3) using a wide array of approaches to characterize activity areas and differentiate cultural from natural (nonhuman) sediments and features, including disturbance processes by biological, soil formation-related pedological, and geological processes; (4) the development and correlation of intra- and extrasite temporal contexts by relative and/or absolute dating; (5) integration of ancient landforms and biological information for the reconstruction of paleolandscapes (Dincauze, 2000); (6) reconstructing paleoenvironments by the identification of pollen, wood charcoal, phytoliths, and seeds; (7) the geochemical analysis of artifacts for procurement, provenance and technology and to document interaction between social groups; and (8) microanalysis of stratigraphy, sediments, and artifacts through micromorphology. Examined here are a few techniques of geoarchaeology that are geared toward: discovering archaeological sites and documenting their internal structure; surveying site formation and disturbance processes; the analysis of soils and sediments; paleoenvironmental reconstruction and the impact of humans on the landscape; the physical analysis of archaeological materials; and the integration of geoarchaeology with social archaeology.

Site Discovery and Definition

Traditionally, locating archaeological sites was accomplished by asking local residents in conjunction with a pedestrian survey; that is, simply walking over the landscape and looking for artifacts and food remains or larger features such as stone house foundations, terraces, or earthworks. In the 1960s with the advent of public archaeology programs – also known as 'contract' archaeology or heritage management studies formulated in response to development projects – archaeological surveys became more systematic, utilizing sophisticated sampling procedures, predictive modeling, and techniques of remote sensing, or archaeological prospecting as it is called in Europe.

Remote Sensing or Archaeological Prospecting

Archaeological prospecting "is considered as an exercise in the deliberate planned measurement and evaluation of physical properties which lead to the detection and mapping of archaeological sites" (Scollar et al., 1990: p. 27). Since all surface indications of many archaeological sites have been obliterated by modern urban encroachment or been buried under soils, excavation has been the technique of choice for locating subsurface cultural deposits. Remote sensing techniques are nondestructive and include aerial photography as well as resistivity, magnetic, electromagnetic, thermal prospecting, and LiDAR (see Conyers, 2012; Hesse, 2010; Lillesand et al., 2004 for a detailed explanation of these techniques).

Aerial Photography and Satellite Imagery

The oldest form of remote sensing, aerial photography, is used in archaeology primarily for discovering field monuments, whether earthen or stone constructions. A huge corpus of aerial photographs exist, some of which were taken during various wars for reconnaissance purposes, others for urban planning, to record changing shorelines, and to document vegetation patterns. However, aerial photos made exclusively for archaeological site discovery are the most useful since they are taken at the appropriate scale, angle, and time of day for enhancing surface features. Prehistoric agricultural field systems in areas such as Mexico and the Pacific Islands consist of low terraces and earthen mounds and embankments which are readily mapped from aerial photographs, saving countless hours of field survey and mapping (Ladefoged et al., 1996).

More recently, technological advances have allowed archaeologists to use satellite images and Google Earth

providing low cost survey methods that are easily accessible (Parcak, 2009). These methods have successfully located previously unknown archaeological sites, can monitor changes through time, such as erosion and the encroachment of development, and can identify and quantify illicit activity (Contreras and Brodie, 2010). Using only an internet connection from any location, vast expanses of land can be quickly and cheaply surveyed without having to physically conduct a ground survey. A further benefit is the ability to survey and monitor potentially dangerous, off-limit or war-torn areas, such as Afghanistan, Iraq, and Syria. However, these methods are not a substitute for fieldwork and can be as easily used by looters or archaeologists.

Other Techniques of Remote Sensing

Once archaeological sites are located, identifying internal structures (intrasite variability) can be accomplished with other techniques of remote sensing including resistivity prospecting, where subsurface anomalies (such as burial pits, compacted ancient floors, or buried house foundations) are mapped by identifying the variation in an electric charge as it passes through the ground (Lillesand et al., 2004). Magnetic prospecting can locate the shape, size, and depth of subsurface anomalies that can then be mapped. Thermal prospecting uses variations in surface temperatures as an indication of subsurface (often cultural) structures (Clark, 2004).

Coring Archaeological Sites

Identifying subsurface archaeological site boundaries, internal features, and differential distributions of artifacts and faunal material can be accomplished by coring (Stein, 1986), especially in conjunction with the analysis of sediment chemistry. Small cores or bucket augers (about 10 cm in diameter) are used to collect sediment samples for the analysis of cultural content and to differentiate cultural from noncultural layers. Since organic matter, carbonates (e.g., marine shells), and phosphates (deposited in the form of urine, excrement, ash, and organic refuse) can increase relative to the intensity of human occupation, these residues are good indicators of past site use and anthropogenic modification, such as agriculture and animal husbandry (Holliday and Gartner, 2007). Distribution of phosphate is used to identify subsurface cultural deposits with a minimum of site disturbance.

Site Formation Processes

The archaeological record is not static, but changes in response to a number of processes – both cultural and environmental (or noncultural). Once deposited in, or on, the surface of an archaeological site, cultural material (artifacts and food remains) and noncultural matter, such as pollen grains and terrestrial sediments, can shift their position or, in some cases, be removed completely. Since archaeological interpretations gain their strength from identifying and analyzing the vertical and horizontal distributions of material within and between sites, it is vital to understand how the positions of these materials may have changed since they were originally deposited – this is the challenge of site formation research defined as the processes, both natural and cultural, that build, disturb, redeposit, and destroy archaeological sites (Schiffer, 1987).

Nonhuman Disturbance Processes

Since the soils and sediments that make up an archaeological site are part of a dynamic open system, archaeological sites are best viewed in a regional context. It is at this level that soil development is best understood because it is influenced by such large-scale processes as precipitation and temperature and erosional agents – including wind and surface water (e.g., rivers, marine shore waves, and sheet wash). Rapp and Hill (1998); (see also Wood and Johnson, 1978) have outlined a number of disturbance processes that affect archaeological deposits. Some of the more important processes include faunalturbation, or the mixing of deposits by animals, especially the burrowing kind (rodents, crabs, insects, earthworms, and birds). Digging burrows not only brings buried material to the surface, but open holes are traps for surface material to accumulate beneath the ground. The mixing of deposits by plants that occurs during root growth and decay is known as floralturbation. Uprooted trees can also bring material to the surface, often from a meter or more below the ground.

Discrete layers within archaeological sites can be homogenized by alternating freezing and thawing episodes which sort sediments by size. This is a problem with sites in the higher latitudes that consist of clay soils that swell and contract in response to precipitation and temperature. Gravity can cause mixing and movement of soil and rock debris downslope, as well as settling or subsidence of relatively large areas. Aeroturbation occurs when wind winnows fine sediments from deposits leaving a 'pavement' of artifacts. This is a significant problem in areas with active sand dune systems. Despite the range of disturbance processes that operate on a regional and site-specific basis, most cultural deposits maintain some order of integrity and, after the site formation processes are understood, valid interpretations are possible.

Human Disturbance Processes

In addition to natural or environmental site formation processes, people intentionally or unintentionally contribute to the formation of archaeological deposits by bringing raw materials to a site such as finegrained rock to fashion into tools; finished artifacts of bone, shell, wood, and stone; cooking fuels, building materials, and food. Many of these items are discarded on site as waste. Even sand grains adhering to feet, after countless thousands of trips to a site, can add significantly to the cultural matrix. Digging pits for earth ovens, trash disposal, burials, small holes excavated for wall or fence posts, or trenches dug for building foundations can displace subsurface cultural deposits. Even stratigraphy that looks intact can contain cultural materials that have been displaced through multiple layers as demonstrated by Khaweerat et al. (2010) for

a coastal midden. Accurate interpretations of the archaeological record can only be made after understanding both noncultural and cultural site formation processes.

Definition of Soils and Sediments

Soils and sediments figure prominently in discussions of site formation processes, so it is essential to know how these terms differ. Soils develop *in situ* at or near the ground surface through the various processes of weathering in conjunction with plant growth and decay, other biological elements, time, climate (temperature and moisture), and local topography. Sediments are particulate matter that has been transported from one location to another by one or more processes such as wind, water, or people.

Principles of Sedimentation

All particles, including artifacts and food remains, found in archaeological deposits can be viewed as sediments (Stein and Farrand, 2001: p. 339) and, consequently, the four principles of sedimentation (Garrison, 2003) apply equally to natural and cultural sediments. First, the source of sediments found at an archaeological site can be unconsolidated particles of bedrock as well as humanly transported stone tools, shells, bones or building stones, to name but a few examples. Second, the transport history of sediments is usually determined by textural and compositional analysis in that biological agents (people, plants, and animals), wind, and water move sediments according to certain physical constraints. For example, wind normally transports evenly sorted sand-sized sediments, while water – in the form of rivers, sea waves, and glaciers – can move anything from clay particles suspended in its volume to huge boulders that are pushed under enormous power. Third, the environment of deposition, in this example, is the archaeological site, but offsite examples can include sedimentary basins, aggrading shorelines, sand dunes, and till and eradict boulders deposited from melting glaciers. And last, postdepositional alterations, discussed above, are caused by people, other animals, plants, and geologic erosional processes.

Sampling Soils and Sediments

Noncultural soils and sediment samples are usually taken from a profile in an erosional exposure or from the sidewall of an excavation unit or trench. Beginning from the bottom up, samples are taken from separate stratigraphic layers or in a continuous column. In the same manner, cultural sediments can be collected from architectural features (walls and floors), middens, streets, burials, and other features such as pits, post-holes, and benches. Collecting micromorphology blocks from these same contexts is also advantageous for understanding mircrostratigraphic events not visible through traditional excavation methods (see Courty et al., 1989). The sample volume varies according to the kinds of analyses undertaken and the size of sediments. For example, small samples of about 20 g are needed to analyze the texture of clays and silts, while several kilograms are necessary for analyzing cultural content or if the sediments are gravel to cobble size. The key is to collect a representative sample of the sediment.

Grains and Particle Size Analysis

Among other things, particle size analysis permits the detection of the agent of deposition (e.g., wind, river, sea) and the environment of deposition such as beach, flood plain, and dune (Waters, 1992). As mentioned above, well-sorted sand is often wind-deposited and certain agents of deposition have particular sediment signatures. Mineral identification of individual grains can suggest a likely source of the sediments, especially if the regional geology is well known. Grain shape (roundness and flatness) can indicate the transport agent, depositional environment, and history. Grains can be, for example, platy or flat; angular, or granular. Roundness refers to the general grain surface curvature and is more a function of mineral composition (certain minerals are harder or softer than others and respond to friction differently), depositional history, and final depositional environment. Grain surface textures, examined under high-power magnification, can be: unworn and angular, suggesting a fresh or minimally transported history; rounded and glossy, the result of transport by running water; or matt-surfaced grains, the result of wind transport. A good review of the analytical methods of sediment analysis can be found in Goldberg and Macphail (2006).

Paleoenvironmental Reconstruction and Human Impact

The relationship of people to the landscape (and here we use 'landscape' to mean the physical and biotic environment) is of worldwide interest, not only to archaeologists, but to biologists and politicians alike, to name just a few. How humans managed their resources through time – whether successfully or with tragic consequences – is of utmost importance to contemporary society, especially as the world becomes increasingly one 'global village.' Archaeology is better positioned than any other discipline to understand the historical details of human–landscape relationships over tens of thousands of years and across an enormously varied set of environmental conditions that posed opportunities and problems for our ancestors.

To document and understand the variation in human–landscape interactions, the analysis of soils and sediments is especially useful when used in conjunction with the differing distribution of floral remains (see Pearsall, 2000, for a review of analytical methods). The analysis of faunal material is also of importance here and is addressed in *Zooarchaeology*. Whether of carbonized wood fragments, pollen, opal phytoliths (minute silica bodies formed in the cells of living plants that can be diagnostic of certain species), or leaves, plant remains are often indicative of vegetation on or near an archaeological site. Wind-dispersed pollen can reveal regional floral patterns. Recovered from dated stratigraphic contexts, vegetation history can be charted over time – before, during, and after human settlement. In essence, archaeological sites hold an historical record of local and regional

paleoenvironmental change, often in response to human modifications to the landscape. The journal *Vegetation History and Archaeobotany* is an excellent source for articles on these topics.

Physical Analysis of Raw Materials and Archaeological Artifacts

A growing area of interest in geoarchaeology is the mineralogical and elemental analysis of raw materials and artifacts such as ceramics, stone tools, amber, glass, metal, and building materials. Analytical results connect artifacts to the geological source and occasionally to the site of production, but also help to identify the human agent in technology and the interrelated web of social practices associated with the procurement and production of objects. The emphasis is on establishing lines of interaction or communication between or within social groups. The diachronic and synchronic distribution of exotic materials is thought to reflect changing aspects of social organization, trade, exchange, early commerce, and extended spheres of influence – topics of great interest to archaeologists.

Instrumental Analytical Techniques

There are many useful techniques for determining the elemental composition of archaeological materials and artifacts (see Garrison, 2003 for a description of various techniques). These include, X-ray fluorescence (XRF), X-ray diffraction (XRD), infrared spectroscopy (IR), neutron activation analysis (NAA), inductively coupled plasma (ICP) and isotope analysis (Collerson and Weisler, 2007), to thin-section petrography which has been used in geologic investigations for over a century.

Ceramics and earthen architecture are manufactured from clay, water, and nonplastic inclusions (temper). Pots were often decorated with an exterior clay slip, incised or stamped with various intricate patterns. Fired bricks occasionally have stamps or makers marks, where sun-dried or rammed earth construction can provide information on production, organization of labor and resource exploitation. Of interest to interaction or provenance studies is the elemental composition of the clay and temper. If the latter consists of crushed rock or terrigenous (land-based) sands, then chemical analysis of the temper may isolate its source. Stone tools, manufactured from, for example, obsidian, basalt, chert, or other finegrained rocks, also can be geochemically analyzed to determine the artifact's geological source. Economically important rock sources were generally well known in prehistory and had a limited distribution, thus facilitating sourcing studies.

X-ray Fluorescence Analysis, Inductively Coupled Plasma-Mass Spectrometry and Inductively Coupled Plasma-Optical Emission Spectroscopy

There are several geochemical techniques that are routinely used to determine the elemental composition of source material and artifacts. Simply stated, XRF analysis consists of bombarding a specimen with primary X-rays. The sample emits fluorescent X-rays that are indicative of the kind and frequency of its elemental composition (oxides, such as iron or silica, are measured in weight percentage, while elements are counted in parts-per-million or ppm). Dedicated computers are used for data analysis, while statistical clustering programs can aid in grouping artifacts with possible sources. Once sources are adequately characterized in terms of their geochemical variability, then unknowns or artifacts can be more confidently assigned to a specific source (see Weisler, 1997 for provenance studies using basalt). Although limited in the range and precision of analyzed oxides and elements, new portable devices are making XRF more accessible to areas where samples cannot be exported or destructively sampled (Shackley, 2011). The most accurate and precise results – and the widest array of elements and isotopes that can be analyzed – must be obtained from more sophisticated geochemical techniques such as inductively coupled plasma mass spectrometry and inductively coupled plasma optical emission spectroscopy. Although the operation of these techniques is beyond the scope of this discussion (see Pollard, 2007), of great interest to archaeologists is the ability to analyze the greatest range of trace elements and isotopes using very small samples (often less than 10 mg) with the highest accuracy and precision thus making artifact source assignments more robust than with other kinds of elemental analyses (Collerson and Weisler, 2007).

Thin-Section Petrography

The study of rocks and minerals in thin section (archaeomineralogy) provides information about how ancient people used materials over time and space as tools, ornaments, pigments, construction materials and other manufactured artifacts, such as ceramics, metals, and glass (Rapp, 2009). Thin-section petrography is used extensively today in sourcing studies when mineral identification is desired. Samples measuring up to 25 mm by 40 mm are cut from ceramic sherds and ground thin enough to pass light through them. Placed under magnification, minerals (such as olivine and feldspar) are identified and texture measured. Within a geologically well-known region, temper composition can be assigned to a particular river drainage (in the example of placer sands), a marine beach, or a larger region such as a group of hills or a mountain range.

Social Geoarchaeology

Geoarchaeology describes and characterizes materials and artifacts but a social geoarchaeology contextualizes and interprets culturally modified materials, contributing to issues of temporality, household archaeology, identity, social memory, and object biographies. Too often, archaeological science is unnecessarily distinct from archaeological theory, even though scholars in both areas share a similar objective: to understand past societies through the medium of material culture. Social archaeology is defined as the "process of knowledge production mediated by material culture and experience" (JSA editorial, 2001), and recent scholarship has demonstrated how all aspects of our physical experience are social, such as landscape,

environment, weather, and climate (Evans, 2003). Social geoarchaeology looks at nature as something nondistinct from culture, where artifacts and deposits are all results of cultural behaviors, actions, gestures, and movements. The geographical distribution and physical and chemical properties of sources and materials allows geoarchaeologists to interpret the material impact on human culture, materiality, and life histories of objects (see Boivin and Owoc, 2004 for case studies). For example, analysis of floor sediments informs on activity areas, storage patterns, and the use of space. Micromorphology of wall plasters and floor deposits suggests the frequency and seasonality of plastering events and the life-history of buildings (Matthews, 2005; Boivin, 2000). Examination of mudbrick compositions revealed individual recipes and household autonomy (Love, 2012). Sediment cores from Crete are used to interpret the process of sedimentation and landscape change from the Neolithic-Bronze Age. Using Pierre Bourdieu's concept of *Habitus*, Jusseret (2010) argues that the sediments and the environment structure and frame the lives of past people, where the environment can impact human actions. Geoarchaeology has long been undertheorized but by providing empirical data on artifacts, sediments, and landscapes, geoarchaeology can significantly contribute to social archaeology.

Geoarchaeology uses the techniques, methods, and concepts of the physical sciences to address archaeological questions. In this short article, we have only touched on a few of the techniques and methods applied to research domains that are of worldwide interest. This subdiscipline of archaeology is continually expanding the range of techniques and their applications to innovative questions. The successful melding of interdisciplinary research, typical of geoarchaeological pursuits, continues to have a bright future.

See also: Bioarchaeology; Ceramics in Archaeology; Geographic Information Systems; Zooarchaeology.

Bibliography

Boivin, Nicole, Owoc, Mary A., 2004. Soils, Stones and Symbols: Cultural Perceptions of the Mineral World. UCL Press, London.

Boivin, Nicole, 2000. Life rhythms and floor sequences: excavating time in rural Rajasthan and Neolithic Çatalhöyük. World Archaeology 31, 367–388.

Butzer, Karl, 1982. Archaeology as Human Ecology: Method and Theory for a Contextual Approach. Cambridge University Press, Cambridge, UK.

Clark, Anthony, 2004. Seeing beneath the Soil: Prospecting Methods in Archaeology. B.T. Batsford, London.

Collerson, Kenneth, Weisler, Marshall, 2007. Stone adze compositions and the extent of ancient Polynesian voyaging and trade. Science 317, 1907–1911.

Contreras, Daniel, Brodie, Neil, 2010. Quantifying destruction: an evaluation of the utility of publicly available satellite imagery for investigating looting of archaeological sites in Jordan. Journal of Field Archaeology 35, 98–111.

Conyers, Lawrence, 2012. Interpreting Ground-Penetrating Radar for Archaeology. Left Coast Press, Walnut Creek.

Courty, Marie A., Goldberg, Paul, Macphail, Richard, 1989. Soils and Micromorphology in Archaeology. Cambridge University Press, Cambridge.

Dincauze, Dena, 2000. Environmental Archaeology: Principles and Practice. Cambridge University Press, Cambridge.

Evans, John, 2003. Environmental Archaeology and Social Order. Routledge, London.

Garrison, Ervan, 2003. Techniques in Archaeological Geology. Springer, Berlin.

Goldberg, Paul, Macphail, Richard, 2006. Practical and Theoretical Geoarchaeology. Blackwell, Oxford.

Hesse, Ralf, 2010. LiDAR-derived local relief models – a new tool for archaeological prospection. Archaeological Prospection 17 (2), 67–72.

Holliday, Vance, Gartner, William, 2007. Methods of soil P analysis in archaeology. Journal of Archaeological Science 34, 301–333.

Jusseret, Simon, 2010. Socializing geoarchaeology: insights from Bourdieu's theory of practice applied to Neolithic and Bronze age Crete. Geoarchaeology 25, 675–708.

Khaweerat, Sasiphan, Weisler, Marshall, Zhao, Jian-xin, Feng, Yuexing, Yu, Kefu, 2010. Human caused stratigraphic mixing of a coastal Hawaiian midden during prehistory: implications for interpreting cultural deposits. Geoarchaeology 25, 527–540.

Ladefoged, Thegn, Graves, Michael W., Jennings, Richard P., 1996. Dryland agricultural expansion and intensification in Kohala, Hawai'i Island. Antiquity 70, 861–880.

Lillesand, Thomas, Kiefer, Ralph, Chipman, Jonathan W., 2004. Remote Sensing and Image Interpretation. Wiley, London.

Love, Serena, 2012. The geoarchaeology of mudbricks in architecture: a methodological study from Çatalhöyük. Geoarchaeology 27, 140–146.

Matthews, Wendy, 2005. Life-cycle and life-course of buildings. In: Hodder, Ian (Ed.), Çatalhöyük Perspectives: Reports from the 1995–99 Seasons. MacDonald Institute for Archaeological Research, Cambridge, pp. 125–150.

Pollard, Mark, 2007. Analytical Chemistry in Archaeology. Cambridge University Press.

Parcak, Sarah, 2009. Satellite Remote Sensing for Archaeology. Routledge, London.

Pearsall, Deborah, 2000. Paleoethnobotany: A Handbook of Procedures, second ed. Academic Press, New York.

Rapp, George, 2009. Archaeomineralogy. Springer-Verlag, Berlin.

Rapp, George, Hill, Christopher, 1998. Geoarchaeology: The Earth-Science Approach to Archaeological Interpretation. Yale University Press, New Haven.

Schiffer, Michael, 1987. Formation Processes of the Archaeological Record. University of New Mexico Press, Albuquerque, NM.

Scollar, I., Tabbagh, A., Hesse, A., Herzog, I., 1990. Archaeological Prospecting and Remote Sensing. Cambridge University Press, Cambridge, UK.

Shackley, M. Steven, 2011. X-ray Fluorescence Spectrometry (XRF) in Geoarchaeology. Springer, New York.

Stein, Julie, 1986. Coring archaeological sites. American Antiquity 51, 505–527.

Stein, Julie, Ferrand, William (Eds.), 2001. Sediments in Archaeological Context. University of Utah Press, Salt Lake City.

Waters, Michael, 1992. Principles of Geoarchaeology. University of Arizona Press, Tucson.

Weisler, Marshall (Ed.), 1997. Prehistoric Long-Distance Interaction in Oceania: An Interdisciplinary Approach. New Zealand Archaeological Association Monograph 21, Auckland, New Zealand.

Wood, W. Raymond, Johnson, Donald, 1978. A survey of disturbance processes in archaeological site formation. Advances in Archaeological Method and Theory 1, 315–381.

Geographic Information Systems

Michael F Goodchild, University of California, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Geographic information systems (GISs) are defined as software systems. In this article, the relationships between GIS and other activities having to do with geographic information are reviewed. The use of GIS in social and behavioral sciences is discussed as an increasingly essential component of the research infrastructure and as a tool for acquiring and communicating geographic knowledge. Examples are used to discuss the importance of GIS across the social and behavioral sciences. Sources of data are reviewed, and GISs are discussed from the perspectives of client–server architectures, the Internet, Internet-based services, data archives, and digital libraries. GIS use is intimately related to the role of space in scientific explanation. The article ends with a discussion on the future of GIS.

Introduction

A geographic information system (GIS) is best defined as a software system designed to provide services related to geographic data: such services include storage, analysis, transformation, maintenance, editing, visualization, and modeling. That somewhat dry definition masks a record of rapid growth and widespread applications, beginning some 40 years ago. GIS (this article follows the common practice of using the term GIS both in a specific and in a generic sense) is now the focus of a substantial software industry and a user community that probably numbers in the hundreds of thousands or even millions worldwide, and many simpler GIS functions are now familiar to the users of online maps and GPS-based navigation systems. While many of those applications lie in areas of social and behavioral sciences, they also extend across the full range of sciences, industries, and government agencies that deal in one way or another with information about the surface of the Earth, the subsurface, and the atmosphere.

GISs are one of several geographic information technologies that have appeared in recent years. Geographic information is also handled in image processing systems, in the form of aerial photographs or images from space satellites, and many GISs include image processing capabilities. The Global Positioning System (Kennedy, 2010) and its Russian and European equivalents use satellites to allow users to measure their positions on the Earth's surface, with accuracies that range from millimeters to tens of meters depending on the complexity of the measuring device. Earth-observing satellites (Jensen, 1996) capture images of the surface, and are now widely used to map, detect change, discover and invent resources, forecast weather, and measure variables of importance to Earth science. The Internet and the Web are widely used to distribute geographic data from archives and digital libraries and to provide GIS services to remote users.

Today, the potential user of GIS might employ any of a number of hardware and software configurations. A simple Web browser, running on a laptop, an office machine, or a mobile device, can be used to access Web sites that offer a range of GIS services. Such services include generation of standard map products, lookup of the latitude and longitude of a street address (geocoding), and simple analysis and mapping of data. In this configuration, the browser software might be free, though the accessed online services might require a subscription. Increasingly, online services are accessed through apps available for the iPhone, iPad, or Android-based system. GIS might also be employed using a desktop system, such as a PC or Macintosh, running a powerful GIS such as ArcGIS (Environmental Systems Research Institute, *www.esri.com*), MapInfo (MapInfo Corporation, *www.mapinfo.com*), or a free open-source equivalent. Data might be resident on the system's hard drive, might have been created locally, or might have been downloaded from a data supplier.

A social scientist 'doing GIS' might use any one of these configurations, depending on personal skills and needs. The system would provide services that include preparation of simple maps; manipulation of data to overcome issues of map projections; running of statistical tests; and fitting of complex models to data.

Courses in the use of GIS and in methods of spatial analysis and modeling are now offered in almost all universities and colleges and in many high schools. A wide range of educational materials are available in the form of textbooks (see, for example, DeMers, 2008; Clarke, 2011; Longley et al., 2011), Web sites (see, for example, *www.esri.com*), and Web-based courses (see, for example, *www.unigis.net*). GIS software vendors also maintain Web sites with much useful information.

The history of GIS (Foresman, 1998) originated in the 1960s, with the development of the Canada Geographic Information System. Analysis of paper maps is notoriously difficult, because of the awkwardness and unreliability of techniques for measuring simple properties such as length or area, and because of the difficulties associated with overlaying different maps of the same area. Computer technology was applied in the late 1960s and early 1970s to the editing of maps by producing agencies, such as the US Geological Survey, and to the automation of various functions associated with the taking of the Census. But early efforts were crude and expensive, and GIS became widely available to government agencies and corporations only in the early 1980s. Today, the general public is increasingly aware of GIS through Web sites such as *maps.google.com*, which generates several million maps a day for users who need driving directions, and through powerful, easy-to-use packages such as Google Earth (*earth.google.com*).

GIS Essentials

General computer packages to support analysis of data began to appear in the 1960s, led by BMD, SAS, and SPSS. The table is a powerful model for organizing data, by arraying cases in the rows and devoting each column to some property of each case, and early packages took advantage of this commonality by organizing housekeeping and analytic functions around it. Thus means and variances could be computed by column; regression could be readily applied to two columns; and data could be aggregated by computing the mean value of a column for every unique value of some other categorical column. In essence, these packages exploited an economy of scale in software production: once the housekeeping functions for a simple data model like a table had been built, it became very cost-effective to add new functions based on the same data model.

GIS similarly exploits economies of scale, but in this case by supporting not one but several data models. The real geographic world is enormously complex, and there are many different ways of capturing its contents in digital form, all of them involving some form of approximation. Remotely sensed images use a *raster* model, capturing the Earth's reflectance in each of an array of square cells. Because of its simplicity, the raster model has been used as the basis for many GISs, each of them representing data as a series of *layers* of co-registered raster cells (Figure 1), with each layer capturing some distinct variable, such as land elevation, county name, average income, population, or class of land use.

In the *vector* model, the world is viewed as occupied by collections of points, lines, and areas. Points are located by *coordinates*, often latitude and longitude or some universal coordinate system such as the universal transverse mercator (UTM). Lines are represented as *polylines* by connecting an ordered sequence of points with straight lines. Areas are similarly represented as *polygons*, that is, figures with point vertices connected by straight edges. Each point, line, or area object in the vector model is associated with one or more *attributes* which capture the object's significant characteristics. A set of objects of the same type is associated with an *attribute table* by assigning one row of the table to each object and one column to each characteristic.

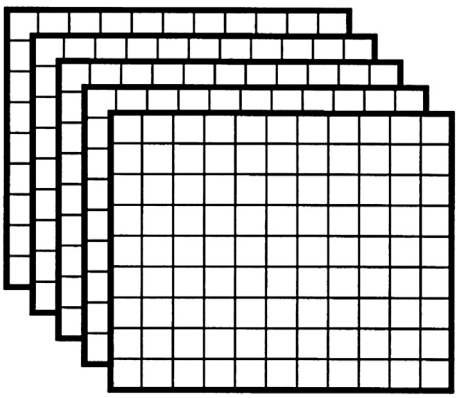

Figure 1 The concept of layers. Geographic reality is represented by a set (five in this case) of co-registered rasters, each representing the spatial pattern of one variable. Example layers might include soil type, vegetation cover type, ownership, mean annual temperature, or elevation.

Figure 2 shows a simple example of raster and vector data. The same set of census tracts is shown in Figure 2(a) in the form of a raster and in Figure 2(b) in vector form. Today, most GISs include support for both models and for transformations between them. They also support various kinds of map projections, allowing the user to switch readily between latitude and longitude, UTM, and other coordinate systems and to display any part of the Earth's surface in the form of a flat paper map.

One of the most compelling arguments for GIS relies on the ability to integrate information around a common geographic location. A project might involve many different kinds of information, about different themes or variables for a geographic area. Every point, line, area, or raster cell in a GIS database has an associated location, so it is possible to determine the distance between two locations or to place locations correctly on a map. But more significantly, it is also possible to identify all of the information available about a location or to link together two data sets through common location. GISs include functions to determine how much overlap exists between two areas, whether a given point lies inside a given area, or to select all of the points lying within a given distance of a line. To return to a point made earlier, the economics of software production provide great advantages to GIS designers once a basic structure has been laid down, because new functions that work on the same structure can be added very easily.

Certain key GIS functions are particularly relevant to social and behavioral science applications:

1. *Geocoding*, or the ability to convert street addresses to geographic coordinates, and vice versa. This function might be used to associate coordinates with surveyed households, allowing locations to be mapped, integrated with other data, or subjected to various kinds of analysis.
2. The *point-in-polygon* function, or the ability to compare a set of points to a set of polygons, to identify the polygon containing each point, and to combine the characteristics of the containing polygon with those of the point. For example, this function might be used to associate census-tract properties with those of the individuals who live there, or to count the number of points lying in each polygon, or to average their attributes.
3. The *polygon overlay* function, or the ability to combine two sets of polygons, creating a new polygon for every unique area of overlap, and associating it with the attributes of the two polygons that overlap to create the area. By overlaying a map of county boundaries on a map of school districts, it would be possible to estimate the number of children living in each school district based on knowledge of the population density of each county (subject to the assumption that population densities are spatially uniform within each county).
4. The *buffer* function, which identifies the area lying within a specified distance of a given point, line, or area, and gives it the attributes of this object. By combining this with the geocoding and point-in-polygon functions, it would be possible, for example, to determine the addresses of households within a certain distance of a shopping center.

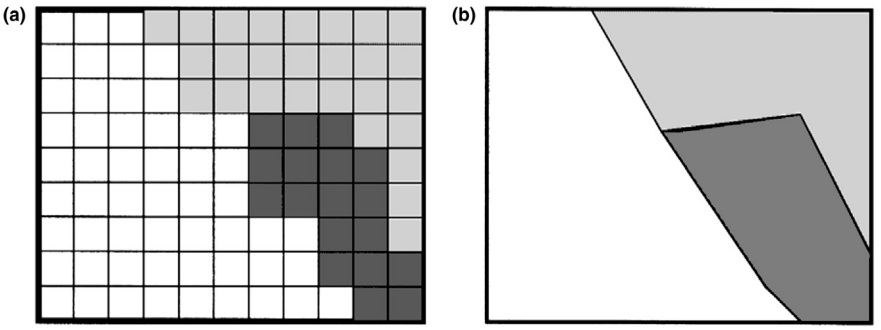

Figure 2 Representation of three adjacent census tracts using raster cells (a) and vector polygons (b). In the raster case, cell size determines the accuracy of boundary positions. In the vector case, each tract is located by specifying the coordinates of its vertices in order.

GIS as Research Infrastructure

GIS software first became widely affordable in the late 1980s, when personal computers powerful enough to manipulate geographic data became available for a few 1000 dollars. Social and behavioral scientists began to invest in its capabilities, particularly in those disciplines that had traditionally adopted a spatial perspective. Archaeologists and anthropologists, for example, began to use GIS software to map and store artifacts at the scale of the dig, to build digital maps of reconstructed landscapes, and to make use of spatial analysis to understand how early societies operated in space (Aldenderfer and Maschner, 1996). Criminologists use GIS to map crime, correlate crime rates with social variables, and examine detailed spatiotemporal patterns of criminal behavior (Wang, 2005). Political scientists use GIS to evaluate districting and to explore the spatial diffusion of political movements (Eagles, 1995). Market researchers and location analysts use GIS to map market potential, model the spatial behavior and choices of consumers, and search for optimal locational strategies (Martin, 1996).

Of all the social and behavioral disciplines, geographers have long claimed a special interest in space, and much of the early work on GIS design and application was strongly associated with the discipline. Geography is often the home department of GIS, though many other arrangements exist, especially at institutions lacking a geography department. The largest number of GIS courses are offered by geography departments, and geography is often the sponsor of campus GIS infrastructure, in the form of labs, software licenses, and data repositories.

In recent years, efforts have been made to develop wider linkages. In the United States, the University Consortium for Geographic Information Science (*www.ucgis.org*) now links over 80 member institutions in an organization dedicated to advancing GIS and its applications through research and education. In Europe, the Association of Geographic Information Laboratories in Europe has a similar mission. National centers have been funded in several countries to pursue GIS research: in the United States, the National Center for Geographic Information and Analysis (*www.ncgia.ucsb.edu*) was founded by the National Science Foundation in 1988, while in Canada, the GEOIDE network was founded by the three national research councils in 1998.

Several journals are devoted to advances in GIS, including the *International Journal of Geographical Information Science*, *Cartography and Geographic Information Science*, *Geoinformatica*, the *Journal of Geographical Systems*, and *Transactions in GIS*. Descriptions of research using GIS can be found in major journals in many application disciplines.

Comparisons have often been drawn between GIS and statistics, as generic tools and concepts used widely across the sciences (e.g., Goodchild, 1987). It is argued that GIS provides the implementation of methods of spatial data analysis in easy-to-use computer tools, just as the statistical packages implement methods of statistical analysis. In this context, spatial data analysis can be defined as a set of analytic methods whose results are not invariant under changes in the locations of the objects of analysis – in other words, location matters. This serves to distinguish these techniques from more familiar numerical and statistical analyses, such as the calculation of a mean, or the preparation of a scatterplot, that do not normally include location or respond to it. Spatial analysis predates GIS, and major texts include those by Berry and Marble (1968), Haining (1990), and Bailey and Gatrell (1995). But the advent of GIS has greatly improved access to these methods. Moreover, it has come at a time when there is increasing interest in visual, intuitive, and exploratory techniques, all characteristics that are readily compatible with GIS. For discussions of the impacts of GIS on spatial analysis, see, for example, Fischer and Getis (2009).

Sources of Data

Maps are highly compressed representations of large amounts of information and extremely tedious to convert to digital form. Two methods are available, both with advantages and disadvantages:

1. *Digitizing*, a process in which a human operator traces information from a map, capturing the locations of features and their attributes using a pointing device and a keyboard. The process is very expensive, since it is difficult to capture complex geographic features rapidly and accurately. With more and more information available in digital form, digitizing is now no longer as important to GIS applications.
2. *Scanning*, in which an automated device captures the information from a map. Extensive editing and error correction are necessary, especially if the map is of low quality, such that scanning is typically as expensive as digitizing.

In the social and behavioral sciences, data are often highly specific to applications and must be rendered into digital form using one of these methods. Thus, an archaeologist may have to digitize the locations of sites or artifacts from manuscript maps or drawings. But, in other cases, the process of database creation will take advantage of the widespread availability of *framework* data or geographic data sets of generic use that already exist in digital form. These include

1. *Topographic maps*, in both raster (in the United States, produced by the US Geological Survey and known as digital raster graphics) and vector (digital line graphs) forms. Various mapping scales are commonly available (in the United States, the most detailed scale is 1:24 000).
2. *Elevation data* in raster form (in the United States, produced by the US Geological Survey and known as digital elevation models). These data sets vary according to the spacing of the grid: for the entire Earth, the best spacing available is several kilometer, but 30 m spacing is available for much of the United States. The Shuttle Radar Topography Mission is a useful source of data at 90 m spacing for all but high latitudes.
3. *Street centerline data* in vector form. These data sets record the locations, names, address ranges, and other information on streets. The TIGER database created and maintained by the US Bureau of the Census covers all of the United States and includes boundaries of statistical reporting zones. These data sets represent streets by single lines and have positional accuracies of tens of meters. Open Street Map is a global project in which volunteers have created a remarkably rich and detailed data set that is freely available.
4. *Imagery from aerial photography*. The digital orthophoto quad is typical of these data sets, with a 1 m ground resolution and a 6 m positional accuracy and available from the US Geological Survey for much of the US.
5. *Imagery from satellites*. These sources include images collected by a number of Earth-orbiting satellites, with spatial resolutions as low as 50 cm.
6. *Municipal databases* created by local government agencies, often in collaboration with utility companies. These are in vector form and often include property boundaries, locations of buildings, and streets represented by double lines. In some countries, notably the United Kingdom, they are produced by government for the entire land area. In the United States, coverage is patchy. Positional accuracies are typically in the meter range.
7. *Gazetteers*, the digital versions of the indexes found in atlases, linking place names to geographic coordinates. These indexes make it possible to translate information recorded by place name to map form for analysis and display. Recently, gazetteer data have been extended to commercial establishments and local points of interest in support of online map applications.

Many of these data sets are created and maintained by longstanding programs of national mapping agencies. Recently, many countries have attempted to place them within broader policy frameworks that emphasize sharing of data, avoidance of duplication, and the setting of standards. In the United States, the National Spatial Data Infrastructure (NRC, 1993) is a leading example.

These data sets provide the framework on which social and behavioral scientists can often hang their own features and attributes. For example, a criminologist might use a municipal database, adding records of crime locations; a sociologist might use reporting zone boundaries from the TIGER database to create maps and analyze data at the census-tract level; and a planner might use imagery to monitor urban growth. Using data that are already in digital form greatly reduces the cost and tedium of database creation.

Unfortunately, the conditions under which such framework data sets are accessible vary widely. In the US, policy requires that geographic data produced by the federal government be distributed to anyone at a cost no greater than that of dissemination. It is often argued that this policy has stimulated growth in the use of GIS and in the geographic information industry generally. But while US courts have held that geographic data are facts and therefore not subject to copyright, in other countries national governments have adopted cost-recovery policies, protected by copyright. Academic scientists in such countries have found various ways of accessing geographic data at lower cost through bulk purchase agreements. Lower levels of government in the US are also not subject to the cost-of-dissemination policy and have made efforts at cost recovery under trademark protection.

While the public sector has traditionally supplied much of the geographic data used by the scientific community, the private sector is today a major supplier, and companies such as Google now manage massive volumes of georeferenced data. Much of this volume originates from users, in the form of user-generated content. The popular Web site Flickr, for example, contains hundreds of millions of geographically referenced photographs, and many tweets are also geographically referenced. Very large geographic databases are being created for internal purposes by insurance, retail, banking, and other corporate sectors. Point-of-sale systems now routinely capture data on grocery purchases, and direct-mail and delivery companies make use of detailed geographic data for scheduling and targeting. These activities have raised privacy concerns, and it is clear that the levels of detail now available to corporations, together with the ability to integrate and merge data based on geographic location, have very significant privacy implications. Much of this activity remains largely outside any regulatory framework, at least in the United States.

Networks

The Internet and Web have had major impacts on GIS. They have made it much easier to share data and to access data remotely. Data can be easily published and disseminated, as can the more sophisticated results of analysis. Indeed, the entire process of scholarly communication and publication is being radically restructured by networking technologies.

It is useful to distinguish two forms of processing of geographic information. The term 'geographic information system' suggests a combination of hardware and software that may or may not be connected to a larger network. It is likely controlled directly by a single user and provides capabilities and functions directly related to the user's needs. On the other hand, the term 'geographic information services' is used to

describe Web sites that offer specific GIS functions in response to user requests. Instead of sending the data to a GISystem, a GIService retains its data, sending only the results of analysis. Esri's ArcGIS Online, for example (*www.arcgis.com*), provides a range of services that include the generation of maps at the request of the user.

GIServices are growing rapidly and are expected to grow much further in the coming years. In addition to avoiding lengthy transmission of data, they allow the custodian of a sophisticated analysis tool to retain control and to upgrade the tool, rather than undertake a complex process of distribution to user sites. Pricing based on services is very different, and perhaps simpler, than pricing based on data (Guenther and Mueller, 1999). GIServices also assume a level of adherence to standards and encourage the development of interoperable tools using common formats and protocols (the Open Geospatial Consortium, *www.opengeospatial.org*, is a major influence toward greater interoperability in GIS).

Several major initiatives are making geographic data more easy to find and access via the Web. In the United States, the federal government sponsors the Geospatial One-Stop (*geo.data.gov*), a catalog and portal to sites offering geospatial data. In Europe, the INSPIRE geoportal (*inspire-geoportal.ec.europa.eu*) provides unified access to the geographic data resources of European Commission's member states. The terms *geolibrary* and *geoportal* have been coined to describe Web sites that allow search for data based on geographic location, including information that is associated with a geographic area but is not a map or an image, such as a planning report.

Space and Scientific Explanation

Ultimately, the value of GIS and related technologies in the social and behavioral sciences will be intimately tied to the role of space in scientific explanation and understanding. Before expanding on that point, however, it is useful to distinguish between normative and explanatory applications. Many applications of GIS may be scientific in their objectivity and rigor, but nevertheless aimed at supporting decisions, developing policy, or solving day-to-day problems. Such *normative* efforts often apply the principles and knowledge of social and behavioral sciences. Moreover, many do so in a spatial context, by combining principles about behavior and society with the specific conditions existing in a geographic area. Thus there are abundant normative applications of GIS, and in recent years the term geodesign has come to encompass this paradigm.

In this section, however, the focus is on those applications that advance scientific knowledge about the world, through observation, experiment, hypothesis testing, and deduction, because such explanatory efforts lie at the heart of the social and behavioral sciences. To what extent can GIS be used as a tool to advance scientific knowledge?

Ideally, scientific knowledge advances through controlled experiment, by varying the conditions under which society operates and observing the response. But although such experiments are common in areas such as agriculture, and GIS is often used as a tool to manage experimental plots, controlled experiments are unlikely to be possible in the case of social and behavioral processes operating at geographic scales. Instead, science must advance through inference, either inductively by inferring the nature of a process from its geographic or temporal expression or deductively by comparing the predictions of theory with reality.

GIS is most strongly associated with *cross-sectional* approaches to observation and inference. A phenomenon is compared across a geographic area, and its spatial variation is compared with the variation in relevant conditions. Inferences are drawn about process, but since the process is not observed directly it is impossible for inference to be definitive. Alternatively, patterns may be used to provide confirmation of hypothesized processes, though again confirmation is never definite. The tradition of using maps as a basis for inference has a very long history, one of the most compelling historic examples being the map made by Dr John Snow of a cholera outbreak in the Soho district of London in the 1850s, which provided evidence of the role of drinking water as the disease vector.

It is often argued that *longitudinal* observation and inference based on change through time is more definitive than cross-sectional inference. But cross-sectional data are widely available from sources such as the Census, and powerful technologies like GIS increasingly make it possible to examine data from both cross-sectional and longitudinal perspectives. Moreover, in recent years, there has been an upsurge in data that are specific in both space and time, raising the possibility of spatiotemporal GIS and real-time analysis. Tweets and other social media, together with large numbers of digitally captured tracks from mobile devices, offer vast new potential for studies of human behavior and social processes.

Geographers draw a distinction between the properties of a *site* and the site's *situation* in spatial context. Many social processes respond both to the conditions at a place and the conditions in the place's surrounding area. For example, choice of residential location is affected by the properties of the available sites and by their surrounding communities and neighborhoods. GIS makes it possible to observe and analyze a site's situation, in the form of maps and images, and also to link one property of a site to other properties through common geographic location. Further discussions of the spatial dimension of explanation, and of the role of GIS, are available in many sources (e.g., Goodchild and Janelle, 2004).

GIS in the Future

There is every indication that the use of computers to handle geographic information will continue to increase dramatically in the future and that the user base will continue to broaden. But trends in the software industry and in society more generally suggest that the nature of GIS will continue to evolve.

In recent years, there has been a discernible trend toward the solution of problems at the local level. It is reflected, for example, in the popularity of *place-based decision making* in the US Environmental Protection Agency, a policy that encourages the solution of problems locally, by local stakeholders. It is also reflected in the efforts to devolve power to lower levels in the administrative hierarchy. In the 1960s, the so-called *quantitative revolution* in geography emphasized the *nomothetic* search for general laws affecting spatial phenomena and saw traditional

geography's approach as *idiographic*, with its stress on cataloging the unique properties of places. Today, there is much interest in quantitative techniques that accommodate spatial variation in parameters and emphasize the role of spatial analysis in exploring local variation. Fotheringham, for example, has identified such variation as one of the most important themes in contemporary spatial analysis and with colleagues has developed the novel technique of *geographically weighted regression* (Fotheringham et al., 2002). GIS provides an ideal platform for such methods, since it makes it easy to analyze spatial data and to display spatial variation.

At the same time, there is increasing interest in several disciplines in the incorporation of spatial effects directly into theory, on the grounds that a system that is spatially distributed, and forced to communicate over distance, is fundamentally different from the perfectly connected systems that have often been assumed in the past. Spatial economics and spatial ecology are now significant areas within the theoretical wings of those disciplines, while geography has a long tradition of explicit theorizing about space. Again, GIS provides an ideal platform for subjecting such spatially explicit theory to empirical test. Finally, the spatially explicit nature of GIS provides an ideal environment for linking normative and explanatory science. The implications of theories for real geographic areas are readily evaluated, and models that incorporate social and behavioral processes can be combined with geographic boundary conditions to provide readily visualized predictions. Thus GIS offers an ideal environment for bridging the gap between theory and practice and a significant clue in the eternal search for relevance.

See also: Global Positioning System; Spatial Analysis in Geography; Spatial Data Infrastructures; Spatial Data.

Bibliography

Aldenderfer, M., Maschner, H.D.G., 1996. Anthropology, Space, and Geographic Information Systems. Oxford University Press, New York.

Bailey, T.C., Gatrell, A.C., 1995. Interactive Spatial Data Analysis. Longman Scientific and Technical, Harlow, UK.

Berry, B.J.L., Marble, D.F., 1968. Spatial Analysis: A Reader in Statistical Geography. Prentice-Hall, Englewood Cliffs, New Jersey.

Clarke, K.C., 2011. Getting Started with Geographic Information Systems, fifth ed. Prentice Hall, Upper Saddle River, New Jersey.

DeMers, M.N., 2008. Fundamentals of Geographic Information Systems, fourth ed. John Wiley & Sons, New York.

Eagles, M. (Ed.), 1995. Spatial and Contextual Models in Political Research. Taylor & Francis, London.

Fischer, M.M., Getis, A. (Eds.), 2009. Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications. Springer, Berlin.

Foresman, T.W. (Ed.), 1998. The History of Geographic Information Systems: Perspectives from the Pioneers. Prentice Hall PTR, Upper Saddle River, New Jersey.

Fotheringham, A.S., Brunsdon, C., Charlton, M., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, Hoboken, New Jersey.

Goodchild, M.F., 1987. A spatial analytic perspective on geographic information systems. International Journal of Geographical Information Systems 1, 327–334.

Goodchild, M.F., Janelle, D.G. (Eds.), 2004. Spatially Integrated Social Science. Oxford University Press, New York.

Guenther, O., Mueller, R., 1999. From GISystems to GIServices: spatial computing in the Internet marketplace. In: Goodchild, M.F., Egenhofer, M.J., Fegeas, R., Kottman, C.A. (Eds.), Interoperating Geographic Information Systems. Kluwer Academic Publishers, Norwell, Massachusetts, pp. 427–442.

Haining, R.P., 1990. Spatial Data Analysis in the Social and Environmental Sciences. Cambridge University Press, Cambridge.

Jensen, J.R., 1996. Introductory Digital Image Processing: A Remote Sensing Perspective, second ed. Prentice Hall, Upper Saddle River, New Jersey.

Kennedy, M., 2010. The Global Positioning System and ArcGIS, third ed. CRC Press, Boca Raton, Florida.

Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W., 2011. Geographic Information Systems and Science, third ed. John Wiley & Sons, New York.

Martin, D.J., 1996. Geographic Information Systems: Socioeconomic Applications, second ed. Routledge, London.

National Research Council, 1993. Toward a Coordinated Spatial Data Infrastructure for the Nation. National Academy Press, Washington, DC.

Wang, F., 2005. Geographic Information Systems and Crime Analysis. Idea Group, Hershey, Pennsylvania, Philadelphia.

Geographic Information Systems and Remote Sensing

Steven M Manson, Dudley B Bonsal, and Melinda Kernik, University of Minnesota, Minneapolis, MN, USA
Eric F Lambin, Georges Lemaître Centre for Earth and Climate Research, Earth and Life Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium; and School of Earth Sciences and Woods Institute for the Environment, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E.F. Lambin, volume 13, pp. 13150–13155, © 2001, Elsevier Ltd.

Abstract

Remote sensing and geographic information systems (GIS) analysis involves the use of technology to gather, manipulate, and analyze spatial data to understand a range of phenomena. Remote sensing entails obtaining information about the Earth's surface by examining data acquired by a device, which is at a distance from the surface, most often satellites orbiting the earth and airplanes. GIS are computer-based systems that are used to capture, store, analyze, and display geographic information. These two approaches are used widely, often together, to assess natural resources and monitor environmental changes. Social scientists can gain insights into fine spatial and temporal dynamics of a range of social phenomena in environmental contexts by analyzing time series of remote sensing data, by linking remote sensing to socioeconomic data using GIS, and developing with these data a range of digital models and analyses. This article examines remote sensing and GIS in general, with an emphasis on the former, and then explores how these approaches may be used together to address a range of issues. It also emphasizes the role of remote sensing and GIS for use by social scientists engaged in the environmental and ecological scholarship.

Analysis with Remote Sensing and Geographic Information Systems

Remote sensing and geographic information systems (GIS) analyses are complementary approaches that social scientists use to gather, manipulate, and analyze spatial data. Remote sensing is the science and art of obtaining information about the earth from a vehicle, which is at a distance from the surface, most often satellites orbiting the earth and airplanes. These craft carry sensors that capture images of the earth surface across a variety of spectra including visible light, heat, and radar. GIS are computer-based systems that are used to capture, store, analyze, and display spatial information (see the entry on GIS for details). These two technologies are used widely across the social and natural sciences to assess natural resources, monitor environmental changes, and understand a range of social phenomena as they pertain to environmental and ecological issues.

In this article, we examine remote sensing, GIS, and their application to social science scholarship in environmental contexts. The modern field of remote sensing began with aerial photography in the twentieth century and matured with the launch of satellite sensors that have acquired multispectral data over the entire globe since the 1970s. A range of natural and human features can be identified within an image via a variety of methods. The analysis and interpretation of remote sensing data typically is performed through image classification and through change detection after the raw data have been preprocessed. GIS facilitates the interpretation of remote sensing data by linking biophysical information measured by remote sensing to maps of natural and cultural landscape attributes, to field measurements, and to socioeconomic information. GIS also offers many different forms of spatial analysis and modeling that draw on these data. Remote sensing and GIS are important approaches for studies aimed at understanding land-use dynamics, its driving forces, and its impact on society. Remote sensing and GIS support the analysis of changing landscapes in a variety of ways: land-cover mapping, monitoring of land-cover change, identification of land-use attributes, spatial stratification and extrapolation of regions based on land-use patterns, development of spatial models, and application to an ever-increasing number of social science questions.

Remote Sensing and the Environment

Humans 'see' the world by sensing electromagnetic radiation in the form of visible light with their eyes and interpreting the resultant signals with their brains. In many respects, remote sensing functions in a similar fashion; a sensor housed in a satellite or aerial vehicle captures images of different wavelength ranges reflected and emitted by the earth's surface. These images are then interpreted by computer, often with human assistance, and processed into spatial data that are then stored, manipulated, mapped, and analyzed. The character of the returned wavelengths depends on surface attributes, such as the composition of the material, surface moisture, and state of vegetation. Different features can be identified within an image based on the specific 'spectral signature' (or specific combinations of wavelengths) of different surfaces, such as the green of a forest or gray of concrete, and their spatial patterning, such as the readily identifiable edges of agricultural fields or the branching of road networks. For environmental and ecological research, a good understanding of the biophysical processes controlling the radiation of a surface is necessary to interpret the remotely sensed signal in terms of surface attributes. Forest cover, for example, can exhibit varying spectral signatures depending on age of the trees, their dryness, or degree of thinning.

Spectral, Spatial, and Temporal Domains

Remote sensing platforms vary in their spectral, spatial, and temporal characteristics or domains. Spectral range refers to the part of the electromagnetic spectrum in which the sensor operates. The term 'multispectral' denotes a sensor that can capture many different slices or 'bands' of the spectrum (e.g., several shades or red, green, and blue in the visible wavelengths) while panchromatic refers to a sensor that collects one electromagnetic band (e.g., resulting in black and white or greyscale images). Some sensors are termed hyperspectral because they can capture hundreds of discrete parts of the spectrum (Jensen, 2005; Campbell, 2007). Consider the range of different approaches and sensor characteristics:

- In the visible and near-infrared (heat) range of the electromagnetic spectrum, the land surface generally is characterized using vegetation indices, which are arithmetic combinations of spectral bands. The most widely used of these indices is the normalized difference vegetation index. Vegetation indices derived from remote sensing data are related to several variables quantifying vegetation activity and state, such as the fraction of photosynthetically active radiation absorbed by the vegetation canopy, canopy attributes (e.g., green biomass or green leaf area index), or state of the vegetation (i.e., vegetation vigor or stress). Seasonal variations of vegetation indices are related to vegetation phenology and biome seasonality.
- In the thermal range of the spectrum, the heat of the land surface can be derived and is related to surface energy balance, moisture availability, and evapotranspiration. Empirical studies have demonstrated the usefulness of thermal data to locate dense forest boundaries and monitor the growth cycle of crops, including for irrigated systems.
- Active sensors are those that emit radiation and measure how much is returned. Light Detection and Ranging (LIDAR) involves pulses of laser light that can create high-resolution representations of 3D surfaces. Microwaves are also used, where the intensity of microwaves returning from a given surface depends on the geometric characteristics of the surface (i.e., shape and orientation of objects, as well as their surface roughness) and its electrical characteristics (i.e., reflectivity and conductivity of materials). Both active and passive microwave sensing are possible, where active systems (commonly known as radar) supply their own source of energy or illumination while passive systems sense energy that is naturally emitted or reflected from the Earth's surface, albeit at extremely low levels. The capacity of microwave remote sensing to collect data in all weather conditions combined with the fine spatial resolution of the data opens promising avenues for land-cover assessment and monitoring. This technology has, however, not yet reached the same operational level as optical remote sensing.

The spatial domain refers to the resolution of the sensor (the smallest discrete area it can identify on the ground), its range or extent (the amount of area it can capture at one time), and the spatial patterns that may be discerned from an image. A major attribute of a landscape is its spatial pattern, namely the arrangement in space of its different elements. The concept of landscape spatial pattern encompasses, for example, the patch size distribution of forests, the location of agricultural plots, the shapes of fields, or the number and configuration of landscape elements. Landscape spatial pattern is seldom static due to natural changes in vegetation and to human intervention. Remote sensing offers the possibility to analyze changes in spatial pattern at the scale of landscapes. Textural information of images can be measured using a variety of indices such as variance or mathematical morphology.

The temporal domain of a sensor can refer to the amount of time any given snapshot covers and how frequently a given area is sampled by the sensor. Snapshots are usually instantaneous (akin to the way that a camera takes a photo at a single instant of time) but some sensors will average electromagnetic radiation collected over a short period of time. Aerial imagery is collected at an interval chosen by the producers but is usually flown on annual or less-frequent basis given the expense (although the advent of drones, or unmanned aerial vehicles, may make aerial remote sensing much more available). Satellites, in comparison, are expensive to place in orbit, but once aloft they capture imagery on a very frequent basis, ranging from a couple of weeks down to a few days. This frequency, also termed the return interval, can affect the utility of imagery for any given use. For instance, fast-moving phenomena like flooding may require frequent aerial imagery, while seasonal monitoring of deforestation trends can be accomplished by platforms with a lower frequency.

Remote Sensing Platforms

There is an ever-increasing number and variety of sensing platforms suitable for environmental and ecological investigations. Technological considerations combined with cost and desired use have resulted in several families of sensors with comparable characteristics.

- Moderate spatial-resolution platforms are characterized by a number of satellite sensors that have been acquiring multispectral data over the entire globe since the 1970s, at a spatial resolution of dozens of meters. The most widely used among these sensors are Landsat Multispectral Scanner (spatial resolution of 80×80 m), Landsat Thematic Mapper (30×30 m), and SPOT High Resolution Visible Imaging Instrument (20×20 m for the multispectral product and 10×10 m for the panchromatic product).
- Coarse resolution sensors include the Advanced Very High Resolution Radiometer (AVHRR) on the NOAA series of orbiting platforms, which have been used widely for monitoring changes in the attributes of the Earth's surface. These data cover the period from June 1981 to date. The daily NOAA AVHRR Local Area Coverage data have a spatial resolution of about 1.1 km, which is well suited for monitoring seasonal vegetation dynamics at the scale of the landscape. Other useful sensors with a 1-km resolution are the ERS Along Track Scanning Radiometer, SPOT VEGETATION, and the Moderate Resolution Imaging Spectroradiometer, which also has some sensors with a finer-scaled 250-m resolution that is used for a range of applications.
- Fine spatial resolution sensors are capable of generating imagery with spatial resolutions as fine as 1 m or lower in

panchromatic mode and 4 m in multispectral mode. These systems are also characterized by short temporal resolution, with repeat intervals on the order of days. Such imagery also provides greater spatial details on the land surface and opens new applications especially relevant for social sciences, such as urban studies. Much of the data from fine resolution sensors comes from commercially operated satellites such as GeoEye and WorldView-2.

- Several active microwave sensors are also available: the Synthetic Aperture Radar (SAR) on the ERS-1 and JERS-1 satellites, and the Advanced SAR sensor as part of the ENVISAT mission. The cloud-penetrating capability of microwaves makes them an ideal tool for work in areas with a permanent cloud cover. LIDAR is increasingly common through national and state initiatives. While coverage is patchy, a growing number of governments are conducting surveys of entire jurisdictions, and the library of private data is expanding very quickly.

Developing Remotely Sensed Data

Just as humans must mentally interpret signals garnered with their eyes, remotely sensed images must be developed from the raw electronic signals collected by sensors into data suitable for human interpretation and computer processing. This process is comprised of two stages: preprocessing and analysis. The analysis stage in turn consists of image classification to generate land-cover maps and, if desirable, change detection to monitor land-cover changes (Jenson, 2005; Warner et al., 2009).

Preprocessing turns the raw images from the sensor into a format suitable for analysis. This stage corrects geometric distortions and georectifies the data, which entails positioning the image with respect to geographic coordinates. It may also modify and enhance the images, including transforming multispectral data into simpler formats amenable for subsequent analysis. Overall, the purpose of these operations is to transform the image data into a form for which earth surface attributes can be readily interpreted by computers and humans.

Images are then classified, in that labels are attached to spatial sampling units (the pixels) in view of their response in different spectral ranges. Vegetation and water, for example, have characteristic responses in the visible and near-infrared ranges that make their associated pixels straightforward to identify. It may also include their spatial attributes, such as texture or pattern, or temporal changes, such as predictable seasonal changes in vegetation cover. The most commonly used classification techniques are based on statistical methods, although more advanced techniques are continuously being developed. For example, contextual classifiers label a pixel in the context of its neighbors in space and thereby exploit spatial information of an image, such as more readily discerning the boundaries of objects on the surface. Another approach is the use of neural networks, computer programs that mimic the operation of the human brain to bring together a collection of simple classifier algorithms to solve a complex classification problem.

Another analytical step, usually pursuant to classification, is creation of time series of remote sensing data that can be used to monitor changes in landscapes. Classic land-cover change detection techniques are based on the comparison of sequential land-cover classifications. For every sampling unit of these classified images, the land-cover categories at the two dates are compared. The comparison of successive maps can result in failure to detect subtle changes when land-cover classes are too broad. Thus, rather than detecting changes on the basis of land-cover categories, change detection is often better performed on the basis of the continuous variables defining these categories, whether these are reflectance values measured by a satellite sensor or biophysical attributes derived from these sensors (e.g., percent forest cover). Change detection is an area of ongoing research and development (Coppin et al., 2004; Kennedy et al., 2009).

Geographic Information Systems

A GIS offers a powerful approach for handling spatial data. It is designed for the collection, storage, and analysis of objects and phenomena where geographic location is an important characteristic (Longley et al., 2010). GIS data are of two primary kinds: locational and attribute. Locational data specify the spatial position of entities or phenomena, while attribute data describe their characteristics. The location of a building, for example, can be specified in a spatial coordinate system such as longitude and latitude, while attributes describe characteristics such as stories or roofing style. Both components – location and attributes – are usually joined into a single data model, or an internally consistent system of storing data on the computer. The two most common models are termed raster and vector. The raster model has its origin in remote sensing images and their square pixels or cells, where a layer of these cells represents a discrete variable such as elevation or land-cover classes. The vector model instead encodes the world as points, lines, and areas and stores a series of variables associated with these geometric entities.

Once data are stored in an appropriate data model, the GIS offers a range of operations to operate on these data. Data management capabilities include functions needed to store and retrieve data from an underlying database or across the internet and transform the data model, projection, or coordinate system. GIS automates many analysis and data manipulation functions, including variable classification and reclassification, overlay operations, neighborhood operations, spatial interpolation, statistical analysis, and analysis of connectivity. The output or reporting functions of a GIS can be used to produce maps, tables, or statistical results (see the entry on GIS for details). GIS also offer a range of modeling and analytical methods for sophisticated analyses such as representing hydrological flows, transportation modeling, or view-shed derivation.

GIS facilitates the interpretation of remote sensing data by virtue of matching data from images with other forms of spatial data via their location (Merchant and Narumalani, 2009). Examples include linking biophysical information measured by remote sensing to maps of natural and cultural landscape attributes (e.g., soil type, topography, accessibility as measured by various distances), field measurements (e.g., crop yield or water quality), and socioeconomic information derived through household surveys and other sources (e.g., indicators of well-being, rate of population growth, agricultural input

use). Scholars across the social and behavioral sciences use GIS to understand myriad issues and by focusing on quantitative spatial analysis and using basic spatial concepts, such as the neighborhood or population density, in their research (UNESCO, 2010).

Application of Remote Sensing and GIS to Social Sciences of Environment

Social and behavioral scientists use remote sensing and GIS to study social phenomena within the environmental and ecological realms. These studies often rely on 'socializing the pixel,' or discerning information embedded within spatial imagery that is directly relevant to the core themes of the social sciences, and using it to inform the concepts and theories pertinent to those themes (Geoghegan et al., 1998). Creating a direct link between spatially explicit land-cover information, as derived by remote sensing, and information on land-use change processes requires the ongoing development of new methods and models, which merge landscape data with data on human behavior. The majority of this work uses land use and land cover as the linchpin for analysis.

Land Use and Land Cover

One of the preconditions for using GIS and remote sensing to understand social processes within an environmental context is an understanding of land-use dynamics, including the driving forces and societal impacts (Liverman et al., 1998; Verburg et al., 2011). Remote sensing is used widely to map land cover and land use. Classification algorithms generally succeed in identifying from about a dozen land-cover classes, including several succession stages of vegetation. Achieving this level of detail requires, however, a set of georeferenced field observations from a sample of locations. Many current sensors do not identify settlement types or different cropping systems very well, objectives that are increasingly desirable and possible with fine-scaled platforms. Remote sensing also facilitates monitoring changes in the spatial patterns of landscapes caused by land use. Change detection can be used to identify dynamics such as an expansion of the area under permanent or shifting agriculture, urban growth, or increase in eroded areas. Quantitative estimates of rates of deforestation are derived from time series of remote sensing data, for example, and remote sensing has been used in semiarid regions to assess dryland degradation and to identify transformations of land use and cultivation. Remote sensing can also identify land-cover change 'hot spots,' defined as areas where high rates of land-cover change have been observed based on indicators such as land-cover fragmentation, occurrence of fires, and opening of new roads. Finally, remote sensing in combination with GIS allows identification and measurement of key socioeconomic and ecological characteristics of land-use systems, such as different forms of cultivation and land-use trends (Buenemann et al., 2011).

Spatial Stratification and Spatial Extrapolation

Remote sensing and GIS allow for the extrapolation of local observations to larger areas. A common problem of in-depth investigations of land-use dynamics is that they require a large data-collection effort in the field, at the household and plot levels. Remote sensing, by measuring indicators of land use, helps to define regions characterized by a similar land use and to identify spatial stratification of a study area in terms of land-use patterns. As a result, findings obtained over specific sites within a sampling unit that is homogenous in terms of land use can be extrapolated more readily to other areas with the same characteristics. Spatial stratification that is performed before field surveys can be used to fine-tune the sampling design and identify villages or households that will be used for in-depth data collection and investigation. This strategy leads to an optimized data-collection effort. Relatedly, the spatial pattern of a changing landscape can provide insight into the processes of land-cover change. Certain types of land-use change tend to fragment the landscape (e.g., expansion of smallholder farming, small-scale logging, overgrazing around deep wells), while other land-use changes increase landscape homogeneity (e.g., mechanized cultivation or ranching over large areas). For example, in a classic study of land-use intensification (Guyer and Lambin, 1993), remote sensing analysis provided a means of matching the patterns from a haphazard sample of households with that of a larger population, based on a few variables that were estimated both by household survey (on the sample) and by remote sensing (on the entire population). The results tended to give greater confidence to inferences from the sample about ecological dynamics.

Linking Household-Level Data with Remote Sensing Data

There is an ever-expanding array of methods for merging socioeconomic household survey data with remote sensing-based information (Liverman et al., 1998). The major challenges of this work are twofold. First, defining appropriate spatial observation units, especially the appropriate level of aggregation of information derived from the domains of social phenomena and natural environment. Second, developing appropriate linkages between household-level and remote sensing data sets. While conceptually straightforward, these links can be difficult to implement operationally. In remote sensing, the unit of observation is the pixel, a division of space that is usually not directly associated to any social science unit of observation, such as individuals, households, or villages. Establishing the correspondence between biophysical and socioeconomic variables requires taking into account the unit of decision-making, units of landscape transformation, and the spatial scale of ecological processes, such as a watershed or landscape form. Linking remote sensing observations to socioeconomic data at the scale of the administrative units, which were used for the collection of socioeconomic data leads to a loss of information as it obscures the variability within the units. Researchers have attempted to integrate remote sensing observations and field surveys at finer levels of aggregation, such as at the scale of individuals, households, or villages (Fox, 2003).

Spatial Modeling

Modeling-based scholarship joins remote sensing and GIS. With the application of remote sensing technology to the

monitoring of land-use changes, for example, it becomes possible to integrate more explicitly spatial relationships into models of land-use change. It is also possible to apply and test these models at a finer spatial resolution and with samples comprising smaller geographic units, such as individual landscape units. Spatial models are born from the combination of remote sensing, GIS, and modeling approaches ranging from multivariate mathematical models to hybrids of agent-based modeling and other simulation methods (Parker et al., 2003). Their emphasis is on the spatial distribution of landscape elements and on changes in landscape patterns, since spatial patterns detectable at a variety of spatial scales via remote sensing provide the most obvious link between spatially explicit land-use change models and remote sensing analysis. Remote sensing also offers a powerful means to calibrate and validate these models by comparing actual and predicted land-use and land-cover patterns, and by comparing the level of fragmentation of the landscape. The goal of these models is the projection of future landscape patterns, often under varying regimes. This joint remote sensing and spatial modeling approach has been applied to a range of land change domains and their associated ecological and socioeconomic dimensions, including tropical deforestation or dryland degradation, desertification, and coastal erosion (Lambin, 1997; Parker et al., 2003).

Environment and Society

While land-use and land-cover research serves as the linchpin for many social science applications of remote sensing and GIS, there are other research domains that use these approaches to address social issues within the realm of environment and ecology. A fast growing area of research concerns the application of remote sensing and GIS techniques to analyze the relationship between public health and the environment, as researchers use these techniques to map disease clusters, model exposure to environmental hazards, and evaluate access to health services (Cromley and McLafferty, 2012). Research is not limited to present-day issues, as archaeologists, for example, have used GIS software to map the locations of artifacts at the scale of the dig, to build digital maps of reconstructed landscapes, and to make use of spatial analysis to understand how early societies operated in space (Wheatley and Gillings, 2003). Remote sensing and GIS are increasingly coupled to examine hazards, including short-term events such as coastal flooding and fires, to longer terms threats posed by sea-level rise (Allen and Wang, 2010). The combination of GIS and remote sensing is also increasingly used to understand features of the urban environment, such as water quality and imperviousness or the changing composition of greenspace (Liang and Weng, 2011; Patino and Duque, 2013). These examples are just a few of the growing number of the application of remote sensing and GIS to social and behavioral science topics within an environmental context.

See also: Geographic Information Systems; Geography: Overview; Globalization: Geographical Aspects; Spatial Analysis in Geography.

Bibliography

Allen, T.R., Wang, Y., 2010. Selected scientific analyses and practical applications of remote sensing: examples from the coast. In: Bossler, John D., McMaster, R.B., Rizos, C., Campbell, J.B. (Eds.). Taylor and Francis, London, pp. 467–485.

Buenemann, M., Martius, C., Jones, J.W., Herrmann, S.M., Klein, D., Mulligan, M., Reed, M.S., Winslow, M., Washington-Allen, R.A., Lal, R., Ojima, D., 2011. Integrative geospatial approaches for the comprehensive monitoring and assessment of land management sustainability: rationale, potentials, and characteristics. Land Degradation and Development 22 (2), 226–239.

Campbell, J.B., 2007. Introduction to Remote Sensing. Guilford Press, New York.

Cromley, E.K., McLafferty, S., 2012. GIS and Public Health. Guilford Press, Chicago.

Coppin, P., Jonckheere, I., Nackaerts, K., Muys, B., Lambin, E., 2004. Digital change detection methods in ecosystem monitoring: a review. International Journal of Remote Sensing 25 (9), 1565–1596.

Fox, J. (Ed.), 2003. People and the Environment: Approaches for Linking Household and Community Surveys to Remote Sensing and GIS. Springer, Berlin.

Geoghegan, J., Pritchard, L., Ogneva-Himmelberger, Y., Chowdhury, R.R., Sanderson, S., Turner II, B.L., 1998. 'Socializing the pixel' and 'pixelizing the social' in land-use and land-cover change. In: Liverman, D., Moran, E.F., Rindfuss, R.R., Stern, P.C. (Eds.), People and Pixels: Linking Remote Sensing and Social Science. National Academy Press, Washington, DC.

Guyer, J., Lambin, E.F., 1993. Land use in an urban hinterland: ethnography and remote sensing in the study of African intensification. American Anthropologist 95, 839–859.

Jensen, J.R., 2005. Introductory Digital Image Processing: A Remote Sensing Perspective. Prentice Hall, Upper Saddle River, NJ.

Kennedy, R.E., Townsend, P.A., Gross, J.E., Cohen, W.B., Bolstad, P., Wang, Y.Q., Adams, P., 2009. Remote sensing change detection tools for natural resource managers: understanding concepts and tradeoffs in the design of landscape monitoring projects. Remote Sensing of the Environment 113 (7), 1382–1396.

Lambin, E.F., 1997. Modelling and monitoring land-cover change processes in tropical regions. Progress in Physical Geography 21, 375–393.

Liang, B., Weng, Q., 2011. Assessing urban environmental quality change of Indianapolis, United States, by the remote sensing and GIS integration. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 4 (1), 43–55.

Liverman, D., Moran, E.F., Rindfuss, R.R., Stern, P.C. (Eds.), 1998. People and Pixels: Linking Remote Sensing and Social Science. National Academy Press, Washington.

Longley, P.A., Goodchild, M., Maguire, D.J., Rhind, D.W., 2010. Geographic Information Systems and Science. Longman, Harlow, UK.

Merchant, J., Narumalani, S., 2009. Integrating remote sensing and geographic information systems. In: Warner, T.A., Nellis, M.D., Foody, G.M. (Eds.), Handbook of Remote Sensing. Sage, London, pp. 257–268.

Parker, D.C., Manson, S.M., Janssen, M., Hoffmann, M.J., Deadman, P.J., 2003. Multi-agent systems for the simulation of land use and land cover change: a review. Annals of the Association of American Geographers 93 (2), 316–340.

Patino, J.E., Duque, J.C., 2013. A review of regional science applications of satellite remote sensing in urban settings. Computers, Environment and Urban Systems 37, 1–17.

UNESCO, 2010. World Social Science Report. United Nations Educational, Scientific and Cultural Organization, Paris.

Verburg, P.H., Neumann, K., Nol, L., 2011. Challenges in using land use and land cover data for global change studies. Global Change Biology 17 (2), 974–989.

Warner, T.A., Nellis, M.D., Foody, G.M. (Eds.), 2009. Handbook of Remote Sensing. Sage, London.

Wheatley, D., Gillings, M., 2003. Spatial Technology and Archaeology: The Archaeological Applications of GIS. CRC Press, Boca Raton, FL.

Geographic Thought, History of

Paul Claval, Université Paris-Sorbonne, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Out of the set of practices, know-hows, and knowledge, which allowed human beings to transform the Earth into their habitat, geography grew up under the guise of vernacular, corporate, and then scientific narratives. It then developed as a science of terrestrial locations and cartography, a description of regional diversity, a natural science of environments, an analysis of the impact of distance on human life, an interpretation of people's experience of nature, place, and society. This history is more that of a coexistence of perspectives than that of displacement of paradigms.

Research on the history of geography developed early, but it remained for long centered on the discovery of the Earth by Western explorers. The emphasis on the evolution of geographic thought developed mainly in the twentieth century. It took momentum in the 1960s and 1970s, as exemplified, for instance, by the series *Geographers: Biobibliographical Studies*, published by Mansell since 1977 and then as an e-book by Continuum. Even with this new interest in the history of ideas and their social context and significance, this research remained focused on Western tradition. Contributions relative to precontact geographies and their influence on modern forms of geographic thought in non-Western countries remained exceptional.

The history of scientific geographic thought was influenced from its beginning, in the fifth century BC, by vernacular practices, corporate skills, and the dynamics of narratives. It was built along five programs: (1) the geometric representation of the Earth's surface and its cartographic expressions, (2) the literary description of the Earth's surface, (3) the search for a naturalist form of explanation, (4) the adoption of a social science type of explanation, and (5) the understanding of the experience of the Earth developed by human beings. These traditions are not exclusive paradigms, but complementary approaches which coexisted for long periods.

General Conditions

Vernacular, Corporate, and Scientific Geographies

As a scientific discipline, geography has a long history. Its practice goes back to the time of Herodotus (484–425 BC) in the fifth century BC. The term itself was used in *De Mundo* attributed to Aristotle in the fourth century BC. Geography deals with orientation, the location of places, the surface of the Earth, the features which characterize it and the way people conceive and experience places, the Earth, and the Cosmos. It means that geography existed as a practice of daily life before the development of scientific thought. People had to find places, evaluate the value of soils and the best way to farm them, detect and exploit mineral ores, etc. Through the names they gave to places, they developed the first – and purely oral forms – of geographic information systems. Many of these vernacular practices and knowledge coexisted simultaneously with the formal development of the discipline.

Since geography gave information regarding distant places, the ways to reach them, their economy, population, and culture, it was important for rulers. Governments relied either on empirical knowledge accumulated by local persons or on objective evidence gathered by their cartographic, statistical, and ethnographic services for the management of a territory. In the first case, the local agents of the sovereign knew the country they were in charge of so well that they could develop policies of their own. Hence, there was a tendency for all governments to develop centralized forms of control, for which they needed special forms of geographic knowledge – and more sophisticated forms of geographic information systems. Modern enterprises also had to develop geographic skills. They had to locate resources, exploit them, and distribute goods. A part of modern geography was linked with their activity.

The history of scientific geographic thought had always been situated in this context, with an eye to the type of practices used by ordinary people as well as to the skills developed by administrative services in gathering and interpreting spatial data. The discipline was institutionalized for several reasons:

1. the will to provide the state (or other power institutions) with valuable knowledge for domestic control or overseas expansion;
2. a curiosity for the low culture forms of geographic lore and their preservation; and
3. the will to provide the citizens of modern nation-states with an image of their country and the rest of the world in order to build or strengthen their national identity and explain how the contemporary world is working.

Geographic Narratives and the History of Geographic Thought

Geographic thought was expressed through standardized forms of narratives. The history of the discipline is partly that of the transformation of geographic discursive fields. What were the publics aimed at, the rhetoric used to attract or convince them, and the type of knowledge judged convenient for them? In Greek and Roman times, for instance, according to one of the established geographic genres, geography was often presented as the account of a journey made by Icarus.

From the Renaissance onward, geography often developed as narratives on the newly discovered lands and the strangeness of their inhabitants. It sometimes took the form of universal

geographies, i.e., pictures of all the places and countries of the Earth. In other cases, it was made of quantitative assessments of the population, activities, wealth, and military assets of the provinces of kingdoms in order to facilitate their control. During the nineteenth and twentieth centuries, geographers developed other types of narratives: the description of exotic places for tourists and the analysis of national territories for the civic instruction of young people.

Geography is a complex discipline. Its changes are partly explained by the ways it was institutionalized and partly by the evolution of the discursive fields that geographers developed in order to gain a large readership. However, its transformation mainly reflects the dynamics of its scientific purposes. It is convenient to distinguish five different scientific programs that were transformed into traditions later.

The Five Approaches in Geography

Geography as a Science of Terrestrial Locations: Cartography

The first program was linked with the problems of orientation. In order to build a reliable system of place locations, Greek geographers discovered that the best solution was to rely on an abstract grid of reference (Jacob, 1991). The idea appeared with the first forms of mapmaking. For the Ionian mapmakers in the late sixth and early fifth centuries BC, all territories had geometric shapes. The idea of spherical coordinates, which appeared in the third century BC, was conducive to the first scientific cartography. At the same time, it provided a coherent grid of reference for locating places – a scientific basis for geographic information systems. As expressed by Ptolemy (AD 98–168), geography became "the sublime science which reads in the sky the image of the Earth."

The Earth is a sphere and terrestrial positions, i.e., longitudes and latitudes, can be measured through the observation of the Sun and stars. Erastothenes (273–192 BC) was the first to develop this line of research. Hipparchus (second century BC) proposed the idea of geographic coordinates. In the second century AD, Ptolemy synthesized the accumulated data in this field. His contribution was preserved and improved by Arabic geographers during the Middle Ages. The rediscovery of his texts in the fifteenth century marked a new start in the history of geographic thought. In the seventeenth century, Bernhardus Varenius (1622–50) modernized the cosmographic approach through the introduction of the Copernican view of a Sun-centered universe.

The major problem for this conception of geography came from the impossibility of keeping time, so that astronomical observations were unable to give a measure of longitude – a weakness which persisted from antiquity to the discovery of the marine chronometer by John Harrison (1693–1776) in 1735. Until the mid-eighteenth century, one of the main tasks of geographers was to compile data relative to distances in travel accounts. From the mid-eighteenth century, mapmaking became the task of specialized engineers and scientists. They provided geographers with the topographical maps they needed.

Geographers ceased to be responsible for the basic cartography, but the problems of location, mapmaking, and the building of geographic information systems remained important for them. Their role changed; they learnt in the nineteenth century to use statistical sources to produce thematic maps. During the twentieth century, remote sensing facilitated the gathering of data on land use, vegetation, temperatures, atmosphere, etc. Thanks to computing facilities, the building of geographical information systems has recently become easier; one of the most prized skills of contemporary geographers is to provide administrations or firms with this new cartographic aid to decision-making.

Thanks to cartography, geographers were the first who were able to describe the Earth as seen from the sky – the Icarus' viewpoint, according to Greek geographers. It was essential for getting an objective description of the face of the Earth and its regional divisions.

Geography as Description: The Regional Approach

The second agenda developed by geographers was the description of the Earth's surface and its regional differentiation. During antiquity, it was illustrated by Herodotus, Strabo (64 BC–AD 20) and Marinus of Tyre (first or second century AD). Arab geographers improved much of the knowledge of the Middle East, India, the Far East, and Africa south of the Sahara. With the great discoveries, the program became really worldwide.

The main problem was what to describe and how to describe it? Thanks to the map and the vertical perspective it offered, geographic description was different from the travelers'; it was not structured along itineraries, but on the apprehension of regional units as seen from above. Such a perspective allowed for the contraction of much information, but it destroyed the link between direct observation and description, hence the dryness of many descriptions until the eighteenth century.

Because there was no standardized vocabulary to speak about the details of topography, rocks, plants, animals, and human activities, geographers were unable to grasp the real world in their narratives. Sometimes, they preferred to provide their readers with information on local mythology (in the Greek and Roman times), the history of Christianity (during the Middle Ages), and the local coats of arms and heraldry (in feudal Europe). In a way, the travel accounts of the Middle Ages and the Renaissance are generally more interesting for today's geographers than the geographical publications of the time.

Everything changed in the eighteenth century, with the development of natural history and a new curiosity for human activity. The work of Linné in botany and zoology, a few geologists for rocks, and the first agronomists for rural life, created the technical terms geographers needed in order to write precise descriptions of the world. Thanks to Jean-Jacques Rousseau (1712–78) and Johann-Heinrich Pestalozzi (1746–1827), geographers began to offer descriptions on two scales: one was based on both fieldwork and direct observation, and the other on the synthetic vision of the map-reader. Carl Ritter (1779–1859), one of the pioneers of modern geography, has been trained in a Pestalozzian school, just as Elisée Reclus (1830–1905). Alexander von Humboldt (1779–1859) discovered the value of first-hand contact with landscapes through his training as a natural scientist and his contacts with

Georg Forster (1754–94), who was on Cook's second voyage around the world.

The quality of geographic description kept increasing with the development of new scientific fields: pedology, meteorology, and climatology on the natural side of the discipline, and agronomy, demography, and social and cultural anthropology on the human one. Many of these research fields were initially explored by geographers and then turned independent.

Because of the improvement of geographical descriptions, geographers perceived their discipline in a different way: they began to consider that one of its major advantages was to deal with the real world. The idea that geography was a science of concreteness appeared. Geographers did not consider themselves only as scientists; description was an art, so that they had to display a literary bent.

During all of the nineteenth century and a good part of the twentieth, when writing descriptions, geographers felt the need to provide impartial views of the world since their work was based on objective observations. They considered that the descriptions and modes of regionalization developed by laymen were subjective and as such, not reliable. They preferred to ignore them.

Geography as Explanation: The Perspective of Natural Philosophies and Natural Sciences

As soon as description progressed in the eighteenth century, a third task appeared for geographers: they had to explain the distributions they observed. A precise description was necessarily conducive to questions about the causes responsible for the patterns read on the maps or the features discovered through fieldwork. Geographers had to give an explanatory dimension to their analyses.

During the first half of the nineteenth century, the main interrogation was about the interaction of natural features, their mutual conditioning, and their influence on the destiny of human groups. In the seventeenth century, Varenius had opened the way through his analysis of position. In order to explain terrestrial distributions, he stressed the role of latitude and of the major topographic accidents. He opposed inland and coastal locations. In the first half of the nineteenth century, this approach was transformed into a coherent paradigm: the analysis of position, well exemplified by Alexander von Humboldt and systematized by Carl Ritter. The major factors in the shaping of the Earth and the explanation of regional distributions were latitude, altitude, continental versus coastal locations, and the patterns of mountain ranges and coastlines. This kind of analysis had to be developed at different scales.

Geographers used these tools to explain the distribution of particular natural features: the altitudinal patterns of vegetation in tropical mountains by Alexander von Humboldt. Ritter wished to explain how the destiny of peoples was shaped by the countries they inhabited. Behind these projects, the influence of early nineteenth-century philosophies was obvious. Humboldt stressed the harmony between the different levels of the geographic reality according to the natural philosophies of his time. Ritter followed Herder (1744–1803) when exploring the influence of environment on cultures.

This first type of explanation, linked with the German philosophies of late eighteenth and early nineteenth centuries, was displaced by a natural science perspective when Charles Darwin (1809–82) stressed the role of the environment in the transformation of life. Geographers of the late nineteenth century, Friedrich Ratzel (1844–1904) in Germany and Paul Vidal de la Blache (1845–1918) in France, built human geography on this evolutionist, either Darwinist or neo-Lamarckian, basis. Their ambition was to explain human distributions through the analysis of the mainly 'vertical' (or local) relations people had with their milieu. In fact, they combined the analysis of position inherited from Ritter with an ecological approach. They stressed both the significance of local environments (the ecological approach) and of circulation (the analysis of position). The risk with the ecological approach was to accept a deterministic interpretation of human–nature relationships. Circulation offered a way to escape this danger.

Vidal de la Blache used the idea of *genre de vie* as the main concept of the new approach. In order to exploit an environment, people developed sets of techniques and practices that were soon transformed into a way of life because of the force of habits. The main interest of the *genre de vie* notion was to give a social dimension (the force of habits) to human geography and to stress the role of techniques (either in the fields of production or exchange) in the mediation between human groups and the environment.

Even if Ratzel and Vidal de la Blache were conscious of the significance of circulation, their human geography tried mainly to explain the influence of physical geography on human distributions and the way people took advantage of natural endowments. The role of social relations and of distance as an obstacle to interaction was practically ignored by many of their followers. Human geography was at its best when it used the *genre de vie* paradigm to explain regional structures through the interaction between physical forces and social groups. Unfortunately, this approach could not work in urbanized and industrialized societies, where *genres de vie* became too complex to be a useful means of investigation.

Because human geography was unable to explain satisfactorily the distribution of social groups and artifacts on the surface of the Earth, some geographers, especially in Germany, preferred to define geography as the science of landscape, which avoided splitting the discipline between natural and social parts.

The analysis of position played a central role in the first efforts to investigate political geography. Alfred T. Mahan (1840–1914) and Halford Mackinder (1861–1947) stressed the opposition between inland locations and coastal ones. This theme became central to geopolitical studies in the first half of the twentieth century.

Thanks to the natural sciences, an explanatory dimension had been introduced in geography. Geographers succeeded in organizing efficiently physical systematic disciplines: geomorphology, climatology, biogeography, pedology, hydrology, oceanography, etc. On the human side of the discipline, the results were less impressive. Geographers capitalized more on the analysis of position than on the ecological paradigm. Even cultural geographers like Carl Sauer (1890–1975) did not try to explore systematically the role of techniques in the distribution

of human groups. By the mid-twentieth century, a feeling of dissatisfaction had developed in human geography.

Geography as Explanation: The Social Science Approach

Many geographers tried to give a new start to their discipline. The most successful was Edward Ullman (1912–76). His diagnosis was clear. Human geography failed because it was conceived as a natural science. Overemphasis had been given to the study of human beings–milieu relationships. Circulation, the other face of human geography as defined at the end of the nineteenth century, had been neglected. In order to build a modern geography, emphasis had to be given to social interaction (Ullman, 1980). At the same time Jean Gottmann (1917–95) developed parallel conceptions.

Geographers discovered that economists had developed an interest for spatial economics since the beginning of the nineteenth century. They had explored the influence of distance upon the location of farming, industrial activities, and services. Geographers decided to take advantage of the results of spatial economics to give a more systematic dimension to human geography. It was the basis of the new geography of the late 1950s and 1960s.

In a way, this new geography translated Ritter's position into social terms. Instead of focusing on the opposition between inland and coastal locations, for instance, it focused on the dialectics between center and periphery. The center was defined in terms of social physics as the point or the area that enjoyed the highest accessibility to the totality of a social group. Procedures were invented to measure centrality and to map it. The idea of population or income potential changed the traditional vision. Instead of the center as a point, it presented the center as an area. In the United States, a central core, the manufacturing or industrial belt, appeared as the main feature of the territorial structure.

Geographers relied on spatial economics to develop their new understanding of human geography, but they went far ahead of economists in their analysis of spatial organization. The regional geography of the first half of the twentieth century fell short of a systematic formulation. Thanks to the works of Edward Ullman, Torsten Hägerstrand (1916–2004) and a few other geographers, the way space was organized – through specialization, the development of networks, the dynamics of scale and external economies, and the opposition between center and periphery – became clear.

During the 1970s and 1980s, a few geographers, for instance, Paul Claval (born 1932), tried to extend the analysis of spatial interaction to noneconomic aspects of social life, either social relations or political structures. Instead of focusing on transport costs, they analyzed communication, the structuring influence of commutation costs, the role of medias, and the significance of symbolism. In this way, they provided a new understanding of the dynamics of globalization, so evident from the 1960s.

The emphasis on social interaction had positive effects on geography in many respects, but it reduced the interest in the ecological approach at a time when ecology came really of age and pollution became a major problem at the local, regional, and global levels.

Geography and Postmodernity: The Cultural Approach and the Experience of Nature, Place, and Society

From the early 1970s, a strong critical movement developed in geography. Young colleagues struggled against the conception of their discipline as a natural science and the interpretation that was given to its social dimension. The positivist approach considered social relations as natural facts. It ignored their historical dimensions and their creation through human agency. As a result, neopositivist geographers did not criticize the hierarchical structures and spatial organizations of the groups they observed and explained, even if they were conducive to exploitation and oppression.

From the eighteenth century, a different conception of geography had developed. In his *Géographie politique*, Turgot (1727–81) wished to promote new geographies much more than to explain the existing one (Turgot, 1808–11). It was one of the main features of the geographic projects of the Enlightenment. The critical tradition was represented during the nineteenth century by anarchists like Elisée Reclus and Peter Kropotkin (1842–1921), who made it illustrious, but the prevailing option for a natural science approach restricted their influence.

At the beginning of the 1970s, David Harvey (born 1935) played a decisive role in the development of a new interest for social justice (Harvey, 1973). However, the break with the neopositivist positions of the 1950s and 1960s had other correlates. If social realities are the product of human agency, it is urgent to study the ways people experience places and conceive nature and society, the dreams they have about their future, and the significance they give to their lives. Geographers can no more ignore the ways people organize their existence. The rigid division between vernacular or corporate geographies on the one side, and scientific ones on the other, fades out. Scientific geographers have to explore the geographies of lay people in order to understand the landscape they shape and the institutions they create. Ethnogeography becomes a new field of investigation.

There was an old but limited interest in people's conceptions of place thanks to the analysis of choronyms. The development of new attitudes among geographers had not been possible without the influence of Heidegger (1889–1979). Eric Dardel (1902–68), in France, was well aware of the significance of his philosophy, as Yi-Fu Tuan (born 1930) or Edward Relph (born 1944) a generation later.

When societies are conceived as a product of human activity, i.e., in an historical perspective, it becomes important to understand the practices, attitudes, knowledge, and beliefs shared by social groups, the techniques they master in the fields of material activity and social life, and the significance they give to their lives. The cultural approach, which had been mainly developed by Carl Sauer in the first half of the twentieth century and had experienced a crisis in the 1960s and early 1970s, benefited from a new popularity thanks to Denis Cosgrove (1948–2008), James S. Duncan (born 1945), Joël Bonnemaison (1940–97), and Augustin Berque (born 1942). It was one of the main frontiers in human geography from 1980 onward. Culture ceased to be considered as a collective heritage. It was analyzed as a continuous process of communication, adaptation, evaluation, and refurbishing.

The criticism of the neopositivist conceptions of social sciences was by no means restricted to geography. The transformations of the discipline were parallel with new orientations in sociology, anthropology, and history. Geographers discovered the interest of the Clifford Geertz's type of thick description as a tool to explore ethnogeographies and enter the complexities of cultures.

Under the aegis of French deconstructionist philosophers, Jacques Derrida (1930–2004) and Michel Foucault (1926–84), for instance, social knowledge was increasingly analyzed as a tool to legitimize the Establishment and gain support for governmental policies, whatever their content. Geographers began to scrutinize geographical narratives and look for their hidden significance. They became increasingly aware of the role of geography as an instrument to build efficient forms of social control. Thanks to Derek Gregory (born 1951), the study of geographical imaginations and of their role in social life and power situations became popular (Gregory, 1994).

The new epistemologies are increasingly critical of the great narratives of the past. Geographers are reluctant to rely on macrotheories or philosophies—historical materialism, for instance. They develop a growing interest in the local and privileged micro- or mesotheories. They explore realms that had been previously neglected: gender geography, the geography of young and aged people, the varied forms of segregation, and exclusion or sexual differences that most societies build.

Men and women do not live in a standardized, sterilized, and acclimatized space. They inhabit specific places, develop territorial identities, and use landscape markers as symbols of what unite them and provide them with a sense of shared community. Consequently, geographers are increasingly concerned with the religions, philosophies, or ideologies people create in order to give significance to their lives. They explore the moves the human mind performs in order to discover perspectives of society, nature, and collective destiny, and build values. Transcendence and immanency are wrongly considered as purely philosophical concepts. They have a spatial dimension; geographers have to explore them.

Men and women are part of nature; they live in nature, with nature (Harvey, 1996). The ecological perspective has been completely recovered. Geographers are increasingly involved in the study of the energy and material relations which develop between human beings and their environment, and in the analysis of the attitudes toward nature, the significance given to natural landscape, and the management of planet Earth.

Contemporary geography goes much deeper into the analysis of social and psychological processes than was the case in the 1970s. At the same time, it has become a more critical and modest discipline. For many of the young radicals, scholars have to refrain from building mega-narratives, as in the past. It is better to concentrate on limited problems.

The Coexistence of Perspectives Instead of the Displacement of Paradigms

Kuhn's interpretation of scientific revolutions was very popular among geographers in the 1970s. They considered the new geography of the 1960s as a new paradigm. In fact, the history of geographic thought does not fit well into the Kuhnian model. We have distinguished five traditions, but when a new one appeared, it did not displace the older ones. The interest in location and mapmaking is less central today than in antiquity or modern times, but it is still fundamental to the discipline. Description as an intellectual venture worth pursuing for its own sake has long ceased to be fashionable, but progress in the analytical instruments offered by other natural or social sciences allows for new readings of landscapes and maps, and the discovery of new patterns. It makes geographers aware of new problems.

Geography as a natural science and geography as a positivist social sciences are still alive. They provided geographers with many of the tools they use, the analysis of position both relative to physical factors along Ritter's views, or social distributions according to Ullman's conceptions, and the ecological approach to human beings–milieu relationships. Many colleagues try to develop more performing models of economic or social distributions. The emphasis on communication is a new development of an old curiosity.

When building a periodization of the history of geographic thought, the development, fading out, and renewal of geographical traditions constitute only one component of the whole picture.

Greek geography was born out a Kuhnian revolution, the invention of the grid of meridians and parallels, but its successes came partly from its role as a repository of Greek culture. With the development of empires, first in Alexandria and then in Rome, it became a useful tool for the new bureaucracies because of the maps it produced and the descriptions it provided.

The decline of the classical tradition in the Middle Ages reflected the disappearance of imperial polities in Western Europe and their displacement by feudal structures, which did not involve explicit corporate geographies. Maps were used as symbols of the monarch's power, not as a tool for his or her action. Because of the development of trade from the eleventh century onward, new cartographic and geographic instruments were needed – hence the invention of Portolano charts. In the Arab world where imperial polities played an important role, the fate of medieval geography was far better.

The new phase of development of geography which started from the Renaissance was linked with the rediscovery of Ptolemy and the cartographic perspective, the voyages of discovery, the rise of the modern state with its proto-bureaucracies, and new forms of corporate economic activities. For two centuries, geography remained based on the two perspectives of cartography and regional description.

The eighteenth century was a major turning point. The cartographic limitations disappeared thanks to an easier measure of longitudes. Because of the progress of natural sciences and the birth of social sciences, geographic description improved much and ceased to be problematical. This kind of description necessarily led to an interest in explanation. The transition from premodern to modern forms of geography was a difficult one and lasted for about a century, since it meant a change in the institutionalization of the discipline: from state cartographic services to universities.

The naturalist orientation that was born at the beginning of the nineteenth century prevailed until the mid-twentieth. From

that time, geography became first a neopositivist social science, and then a discipline, partly scientific, partly humanistic, or philosophical, based on human experience and the significance of the human presence on Earth. Institutionalization began once again to change. Geography is increasingly dependent on research institutions and private funding.

Geography offers the intellectual tools that are needed to understand the contemporary acceleration of globalization. It explains why technical uniformity is conducive to new forms of regional organization and does not entail the destruction of all spatial differentiation.

What will be the future of the discipline? As in the past, a part of the specialty fields it nurtured will develop as fully fledged sciences of their own. The technicalities of contemporary physical geographies will deepen the gap between them and human geography. The future of geography as a unitary discipline is perhaps more on the reflective side than on the analytical side of human knowledge.

Critical Approaches and Postmodern and Postcolonial Interpretations of the History of Geographical Thought

During the twentieth century, the history of geographical thought was mainly developed as an Academic enterprise: its aim was to provide geographers with a better interpretation of their discipline than that offered by the oral traditions developed by any group of scholars wishing to give a meaning to their enterprise. It was, in this way, a critical enterprise, but which remained encapsulated in the Western vision of science.

From the 1990s, the history of geographic thought adopted a more critical orientation, which meant a new interest in non-Western traditions, the collusion between Western geographies and imperialism, and the role of gender in the development of the discipline.

During the last generation, this critical dimension was extended in two ways.

1. By the end of the twentieth century, the interest for non-Western forms of geographic thought and non-Academic forms of geographic knowledge began to develop. Keiichi Takeuchi explored, in this way, the introduction in Japan of Western conceptions of geography, but also pictured the Japanese traditional interpretations of the field and the way they influenced – up to a point – the new orientations (Takeuchi, 2000). This kind of study is today popular in many countries, the Muslim World or Latin America: How *altepelt*, the Aztec vision of the city, was incorporated into the corporate geography of Spanish colonial authorities in Mexico, for instance (Fernandez-Christlieb and Garcia Zambrano, 2006)?
2. A new form of critical interpretations of the history of geographic thought appeared at the end of the twentieth century. It tried to renew itself with the critical perspective of the Enlightenment as analyzed by the Frankfort school in the interwar period and after World War II. It took advantage of the ideas developed by Jacques Derrida (1967) on deconstruction, Michel Foucault (1975) on the role of the observing gaze in the building of societies of surveillance, Yves Lacoste (1976) on the military uses of geographic knowledge, and Edward Said on *Orientalism* (1978). These orientations began to influence the history of geographic thought by the early 1990s, as shown by Gregory (1994). Initially, this critical turn mainly stressed the links between Western imperialism, the development of cartography, and Academic geography (Godlewska and Smith, 1994). It focused on the role of the observing gaze – and of cartography, as its main application – in the development of imperial policies (ÓTuathail, 1996; Driver, 2000).

The critical studies are now more diverse. Denis Cosgrove (2001) gave more attention to the building of Western imagination. The significance of exoticism or the gender dimension of geographies is explored. Jean-François Stazsak (2003) showed how, by the end of the nineteenth century, Western countercultures began to reevaluate non-Western civilizations and geographies. Their views influenced the geographical perception of the non-European world developed in already industrialized countries.

In *Civic Geography: Geography in America, 1860–1890*, Karen Morin (2011) offers a good illustration of the new orientations in the study of geographic thought. In a way her book is a classical one: a biography of Charles Daly, who presided over the *American Geographical Society* for 35 years. Its main interest lies in the way she explores the interplay of Daly's conception of geography, the liberal vision of the world he shared with most Americans of the time and the role of the institution he was responsible for.

> For Morin, geography is not just civic but also disciplined, in the Foucauldian sense. For Morin this means telling the story of the dialectic between thought and action: geographical thought and whether it disciplines how people think about the world and act on it. Without the right thought the action cannot take place in the way that it does. In this sense [Morin] is tackling the history of those who 'thought out space' in the words of Foucault (1984: 244).
>
> Crampton, 2012: 539–540

Through the evocation of Daly's role at the head of the *American Geographical Society*, Morin shows how Daly "explicitly promoted a type of commercial geography in Africa" (Morin, 2011: 21). Was gender important in shaping the scientist Daly's conceptions of geography and the action of the *American Geographical Society*? Morin explores this problem.

Such a study is a good example of the new ways to explore the history of geographic thought. It is strongly contextual and shed new light on the relations between vernacular, corporate, and scientific geographies.

See also: Environmental Determinism; Geography: Overview; Nature and Society in Geography; Regional Geography; Spatial Thinking in the Social Sciences.

Bibliography

Berdoulay, V., 1981. La Formation de l'Ecole française de géographie. Bibliothèque nationale, Paris.

Capel, H., 1981. Filosofia y ciencia en la Geografia contemporanea. Una introduccion a la Geografia. Barcanova, Barcelona, Spain.

Claval, P., 1998. Histoire de la géographie française de 1870 à nos jours. Nathan, Paris.
Cosgrove, D., 2001. Apollo's Eye: A Cartographic Genealogy of the Earth in the Western Imagination. Johns Hopkins University Press, Baltimore.
Crampton, J., 2012. Disciplining the A.G.S., Geographical Review 102 (4), 539–541.
Derrida, J., 1967. L'Ecriture et la différence. Editions du Seuil, Paris.
Driver, F., 2000. Geography Militant: Cultures of Exploration and Empire. Blackwell, Oxford.
Dunbar, G., 1991. Modern Geography. An Encyclopedic Survey. Garland, New York.
Fernandez Christlieb, F., Garcia Zambrano, A.J. (Eds.), 2006. Territorialidad y paisage en el Altepetl del siglo XVI. Fondo de Cultura Economica, Mexico.
Foucault, M., 1975. Surveiller et punir. Gallimard, Paris.
Foucault, M., 1984. Des espaces autres, n° 360, p. 752–762, Dits et écrits. Gallimard, Paris.
Godlewska, A., Smith, N. (Eds.), 1994. Geography and Empire. Blackwell, Oxford.
Godlewska, A., 1999. Geography Unbound. French Geographic Science from Cassini to Humboldt. The University of Chicago Press, Chicago.
Gould, P., 1985. The Geographer at Work. Routledge & Kegan Paul, London.
Gregory, D., 1978. Ideology, Science and Human Geography. Hutchinson, London.
Gregory, D., 1994. Geographical Imaginations. Blackwell, Oxford, UK.
Hartshorne, R., 1939. The Nature of Geography. The Association of American Geographers, Lancaster, PA.
Harvey, D., 1973. Social Justice and the City. Arnold, London.
Harvey, D., 1996. Justice, Nature and the Geography of Difference. Blackwell, Oxford, UK.
Holt-Jensen, A., 1988. Geography. History and Concepts. Paul Chapman, London.
Jacob, C., 1991. Géographie et ethnographie en Grèce ancienne. A. Colin, Paris.
James, P., 1972. All Possible Worlds. A History of Geographic Ideas. Bobbs-Merrill, Indianapolis, IN.
Jones III, J.P., Natter, W., 1999. Space 'and' representation. In: Buttimer, A., Brunn, S.D., Wardenga, U. (Eds.), Text and Image. Social Construction of Regional Knowledges. Institut für Länderkunde, Leipzig, pp. 239–247.
Lacoste, Y., 1976. La Géographie, ça sert, surtout, à faire la guerre. Maspéro, Paris.
Livingstone, D.N., 1992. The Geographical Traditions. Blackwell, Oxford, UK.
Morin, K.M., 2011. Civic Discipline: Geography in America, 1860–1890. Ashgate, Burlington, VT.
Phillips, R., 2006. Sex, Politics, and Empire: A Postcolonial Geography. Manchester University Press, Manchester.
Stazsak, J.-F., 2003. Géographies de Gauguin. Vincent Bréal, Paris.
Said, E.W., 1978. Orientalism. Routledge and Kegan Paul, London.
Schultz, H.D., 1980. Die deutschsprachige Geographie von 1800 bis 1970. Selbstverlag des Geographischen Instituts der Freien Universität Berlin, Berlin.
Stoddart, D.R., 1986. On Geography. Blackwell, Oxford, UK.
Ó Tuathail, G., 1996. Critical Geopolitics. Routledge, London.
Takeuchi, K., 2000. Modern Japanese Geography. Kokon Shoin, Tokyo.
Turgot, A.R., 1808–1811. Sur la géographie politique. In: Dupont de Nemoun, P.S. (Ed.), Plans et projets d'ouvrages, Oeuvres completes Paris.
Ullman, E.L., 1980. Geography as Spatial Interaction. University of Washington Press, Seattle, WA.

Geographies of Food: Politics, Place, Power Reprised

Nick Lewis and Richard Le Heron, School of Environment, University of Auckland, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article we work from our positionings in geography, Australasia, and our teaching/researching work to lay out a contemporary landscape of agri-food knowledge and knowledge making. It traces our evolving reading of contemporary agri-food knowledge through participation in the globally prominent Australasian Agri-food Research Network, our own research program, and our graduate agri-food class. The article forms a pedagogical and research development resource based on the situatedness of food thinking and politics that issues a call for more enactive forms of food research as the world enters a period of shifting geopolitics, society–nature relations, and food demands in the Anthropocene.

Introduction

Social science explorations of food are set in trajectories of thought and the material and particular contexts that frame the questions asked by researchers. The choices available around food research are always worked out in situated contexts and situated/grounded ways, while the knowledge we build up around them is forged in an interplay between a politics of food and a politics of knowledge about food. The food literature has a history of geographical silences, while also featuring narratives that bring places together. To borrow a phrase from Campbell (2009), food knowledge is always 'from somewhere' despite any efforts that might be made to globalize it. Any exploration of food and its socioecological relations requires a point of entry into the worlds of knowledge and of material events, a site and moment of framing that is itself a political act, and one that unleashes a politics of knowing. This means knowing and knowledge are framed by shifting trajectories and assemblages in knowledge production that are of both their place and their intellectual field. Critical social science contends that such entry points and political projects of knowledge making have transformative aspirations, and demands that its scholars reflect on the genealogy of their knowledge making but also its politics and that they make these matters explicit. The food literature encompasses many political projects of knowing and a diversity of starting points in geographically situated material concerns, theoretical understanding, and cultures of reading and learning knowledge.

Our starting points are three generative experiences of agri-food knowledge production: our University of Auckland graduate course on agri-food relations and practices (see course description (see http://www.env.auckland.ac.nz/en/about/our-postgraduate-courses/pg-geography-courses.html#a25db098d0b5d458d3ed76723fcdae8b)); 20 years of Australasian Agri-Food Research Network (AFRN) conferences from which we have drawn much of our intellectual inspiration (http://afrn.org.au/); and the Biological Economies (BE) research program, which has dominated our recent agri-food research (Lewis et al., 2013; Campbell et al., 2009). These interconnected sources of diverse knowledge allow us to trace trajectories in the literature that define the field from where we practice our research and teaching and from where we interpret literature shifts. Importantly, ours is a scholarship (and a chapter) grounded in geography. When we say 'geography' we refer less to a well-charted disciplinary field or the location and movements of material things and more to a critical disposition to reveal where things are known and done, disturb extant understandings and presumed universalities in knowing, and highlight both territorialized and otherwise networked configurations of knowing and doing. We see territorialized assemblages as technologies of knowledge production and political mobilization. At the same time, we see critical agri-food knowledge making and practice as a commitment to reframing knowledge and practice of things food in particular constituencies and/or to reterritorializing food politics. Our own knowing and doing of agri-food relations is of our place.

At the heart of the article lies our course, 'Future Food and Resilient Biological Economies' (Geog738), run at The University of Auckland since 2008. We engage our students with and trace local and global issues of agri-food from an Auckland setting. Geog738 privileges a particular reading of a set of debates in the social science of food. Each debate has its own disciplinary or transdisciplinary starting points ranging from rural sociology (family farm to financialization), anthropology (food identities, following as methodology), geography (postproductivism and multifunctionality), economic sociology (calculative practices), to rural sociology science technology studies (performativity of food science), political studies (food regimes (FR)), and/or development (food security), health (food and the body), and environmental studies (nature–society relations). Reflecting our 'Northern Antipodean' setting we take particular interest in the agri-food relations of the affluent North, especially supermarket controlled supply relations, the interplay of food, health, and well-being, and alternative food networks (AFNs); the challenges to the productivism of the practices and policying of New Zealand agriculture. Our location, the School of Environment astride the Arts and Science Faculties, means we draw students from Development Studies, Sociology, and Environmental Management and Environmental Science, a rich mix. Students have included permaculture activists, environmental activists, community volunteers, economists, development NGO workers, local and national politicians, alternative agriculture practitioners, health policy

analysts, chemists, and others interested in all of these fields, as well as cultural theory and life science. This architecture of engagement resonates with what many social scientists are now striving to do.

The annual conferences of the AFRN and their themes have given impetus to a distinctive trajectory Australasian agri-food research, from critiques of Australasian leadership of the GATT-supporting Cairns Group to concerns with the demise of the family farm to engagements with organics, AFNs, community economies, urban food provisioning, and food activism in the Pacific and beyond. The involvement of global protagonists from across the disciplines has in part enabled Australasian researchers to be at the forefront of a number of fields. The significance of AFRN's global agency is demonstrated in the 2009 special issue of *Agriculture and Human Values* on FR. Its coeditors, from Australia and New Zealand, and its contributors had been conspicuous participants or key note speakers at AFRN meetings, where much of the critique and refashioning of FR ideas had been thrashed out. An Australasian synthesis articulated the emergence of deepening divides in food production and consumption with the phrasing 'food from somewhere' and 'food from nowhere.'

The intellectual milieu of the AFRN has analyzed new lines of agri-food research in both New Zealand and Australia and fostered a deep critique of the agricultural research establishment and its grip on policy. The BE project has sought to reunderstand New Zealand in its own terms, shifting away from the family farm as a platform for knowing and politics. The aim was to identify novel theoretical and methodological approaches that might allow us to draw on the literature's new lines of flight and to put the political question of 'why' into the shifting configurations of 'who,' 'what,' and 'how' being grasped at in different literature (Lewis and Rosin, 2013). This required a stretch in imagination and geographic reach. Working with food movements, relations, and translations from paddock to 'someone else's' plate, 'someone else's' environment, and 'someone else's' consumption values demands deep consciousness of priorities of others, an appreciation of epistemic barriers to reimagining, and a recognition that doing food differently is transformative of self and collectiveness. Mobilizing the values of others (predominantly urban consumers in other places) to build prosperity in our settings by reshaping our productivist values and approaches is not straightforward.

In the article we turn first to the major shift in the grand currents in agri-food theorizing that has taken place in the last decade or so involving the opening up of the master narrative of FR thinking. We then highlight six distinctive projects in knowledge production that have emerged from the questioning of FR premises, reorientations from material and discursive conditions, and cross-fertilizations of wider social knowledge into the agri-food literature. The final part of the article uses the work of US rural sociologist Michael Carolan to work through one assemblage of these shifts in knowing foodscapes, an assemblage that directs attention to an engaged politics of knowledge production in food, and one that offers a basis for knowing and enacting food futures differently in what is becoming known as the Anthropocene.

Conceptual Reframings of Food Relations

The late twentieth-century food literature has been forced to respond to significant changes in both global and local food relations and food landscapes, and in approaches to knowing the social world in the wider literature. Here we identify the key shifts in food thinking as researchers have brought a diversity of theoretical and methodological approaches to changing material challenges and priorities. In short, the late 1980s dominance of political economy in the guise of FR thinking and development on the one hand, and a politics of the struggling small farmer and food access (the calorie) on the other, has given way diverse conceptual framings used by different researchers in different contexts to address different concerns in different political projects in different parts of the world.

The FR theorists periodized agriculture and food through a political economy critique of capitalism (Campbell and Dixon, 2009; McMichael, 2009). Emphasizing national subjects and spaces, they linked colonial and neocolonial geopolitics to food production and trade, regulation, and the institutions and institutional sites where regulation takes form (Friedmann, 1993, 2000; McMichael, 1992). They placed food and agriculture at the center of tectonic shifts in global political economy, identifying three distinctive and relatively stable periods of global food economy or historical regimes: the colonial-diasporic (pre-World War II), mercantilist industrial (post–World War I), and 'corporate' or 'corporate environmental' (contemporary) (Friedmann, 2005; McMichael, 2005, 2009). FR theorists focused attention on the rules and governance relations at the heart of the coordination of circuits of production, distribution, and capital (Friedmann and McMichael, 1989) and the dialectical tensions between the regime of coordination and shifting strategies of capitalist accumulation (Pritchard, 1998).

The contemporary, 'third' FR is highlighted by the increasing globalization of production and consumption markets and the inter- and intracorporate relations, and state and private regulation that enable corporations to be "geared to the global or regional markets that are replacing national markets" (McMichael, 1992). It identifies the links between broader investment trajectories in and across nations combined with the supportive rhetoric and institutional development of GATT/WTO with farm-level change, the rise of globalizing supply chains, and the concentration of capital in agribusiness. The actual geographies of agriculture and food began to exceed the FR framings and the territorializations they implied, spawning new lines of theoretical and empirical work to match a moving world of food. The dominance of this political economy reading was undermined further by scholarly reactions against structuralism and Marxian *meta*-narratives and the turns to culture, consumption, and microanalyses of governance focused on practices and technologies (Campbell, 2009).

A more diverse literature now deploys different concepts to address a far wider set of themes. Emerging from post-structuralist, humanist, environmentalist, and posthumanist thought, as well as a fracturing of the political economy challenge to the 'globalization' and the emerging corporatist FR, they offer multiple challenges to contemporary food

economy. (The references that follow in this paragraph are drawn from the Geog738 reading list, and therefore a particular disciplinary reading of broader shifts in the literature.) These can be understood as more or less distinctive trajectories that include concerns with consumers, consumption values, food culture, and identity politics in food (Miele, 2006; Jackson, 2010); society–nature relations and the interplay between human and nonhuman actors (Whatmore, 2006; Carolan, 2008a); environmental change, sustainability, and the challenge to productivism from AFNs (Maye et al., 2007; Goodman et al., 2011); science and technology in food (Enticott, 2003; Miele, 2011, 2013); new ethical and moral economies of food (Jackson et al., 2008); food, nutrition, and well-being (Dixon, 2002, 2009); embodied food experiences and food justice; and ethics (Lockie, 2009; Guthman and DuPuis, 2006; Guthman, 2009; DuPuis et al., 2011; Goodman et al., 2010) the changing countryside of amenity migrants and environmental values (Woods, 2007). Political economy interests have also been recomposed and refocused on retailer control, new private orders of regulation, and questions of food security and food sovereignty, financialization (Campbell, 2005; Campbell and Rosin, 2011; Burch and Lawrence, 2013; Edelman, 2014; Clapp, 2014). New efforts to bring political economy to the materialities of social relations in food supply chains (Pritchard, 2013; Pritchard and Connell, 2011), as well as to food questions at the level of the body (Guthman, 2009), have also been made. As we are compiling this list early in 2014, a new political conceptual framing of food issues is emerging under labels like 'urban agriculture' and 'urban food provisioning' (see Morgan, 2013; Blay-Palmer and Donald, 2006), which promises to refocus questions of food on cities and use this rescaling to bring into dialogue critiques of foodscapes in Northern and Southern worlds, and explicitly to draw lessons for the Global North from the Global South.

Four features of this list are notable. First, the foci are distinctive, but are related. They intersect in the nature of the questions addressed, the politics of knowledge production and common methodologies such as critical discourse analysis and following. Taken together, the different approaches reflect a deepening of knowledge production capabilities around the social and reveal the enrichment of food scholarship over the last decade as authors have drawn more widely upon this capability. This in turn suggests renewed potential for a more varied institutional engagement. Second, the different foci have emerged in relation to each other and the changing priorities of global and local food politics. This is a challenging aspect, not least because it is hard to keep up with what is going on elsewhere. Third, as a result of both of these factors, the list must be considered a particular reading, where the categories are conceived and the borders (albeit porous and changing) are drawn for particular tasks. This cannot be a definitive categorization, though studies of food by social scientists have only recently begun to consider what the impossibility of universality might mean. And fourth, FR thinking has in turn been enlivened by the interest in food sovereignty, consumption, and calculative practices, and by other materialities in food supply chains. Here we point to the growing influence of actor network ideas and the determination of food scholars to put these to particular uses in critical and constructive analyses of food materialities (see, for example, Mather, 2014; Miele, 2013). The screening and narrowing work of the once hegemonic FR perspective is being recognized and credibility, legitimacy, and potential of initiatives missed out in FR narratives are being assiduously investigated.

Key Lines of Engagement with Prominent Trajectories of Agri-Food Knowledge Production

The previous section establishes a conceptual appreciation of social science engagements with the foodscape as a set of interrelated explorations of interests in food relations and associated material questions. We now drill down into these explorations using differently scaled and positioned research trajectories focusing on the work of particular authors with whom we have engaged. We use the metaphor of 'a geography' to characterize these programs, so as to emphasize that there is a defining geography to the engagement, the conceptual armory deployed, and the foodscape of interest. Read from the material world to the research approach, there are clearly immediacies in different places to do with food challenges and food politics, as well as political projects of knowledge production, be they responses to ideological commitments, funding regimes, or a gritty material politics of food security or sovereignty (Edelman, 2014; Bernstein, 2014; Agarwal, 2014). The section is intended as a nonhierarchical intervention of illustrative explorations. By this we mean it does not matter where we start but the consequential links arising from 'an ordering' are likely to yield differing reflections.

Hidden Labor Behind All Seasons Food

Joerg Gertel and colleagues work on temporary food labor migration regimes in the Mediterranean (Gertel and Sippel, 2014). This is a hidden face of food affluence – it is a much more gritty and grounded account of the conditions that are gestured to in the later 'following' accounts of Cook and others, conditions that are built up into mesoscale configurations of processes, and actors and their dynamics. The research tracks multiple investment relationalities of nations, regions, corporate, and local food actors which translate into large shifts and movements of people; continued existence of the state and multiple forms of the state; the conditionality of rights in different places, the relations among them and how embodied and differentially righted labor underpins intensive agriculture of the Mediterranean; the aesthetics of food choice and food health in affluent Europe; and the linkages between the material and discursive. The research is closely tied to Gertel's own following of the emergence of these local–regional–global issues – in mesoscale and longitudinal configurations (Gertel, 2005). This non-Anglo work with its German–French–Arabic focus and German-EU funding, grapples explicitly with the difficult questions that transcend national interests.

La via Campesina – Food Sovereignty and a Radicalizing and Expansionary Initiative

This literature comes out of the food movements and non-Anglo Global South, and draws on activist and Latin American understandings of the possibilities and practice of localized politics. It is land tenure, food entitlement, and livelihood based but animated by the activist's sense of urgency, rights, and possibility in contexts where private land owners and investors have commonly escaped state oversight to more fully appropriate the means of survival. Its focus on peasant production has located it stubbornly in smallholder production. This situatedness has given it an autonomous trajectory, but it has also been aligned to a new generation of FR thinking and political economy by the work of McMichael (2008, 2014) and others, many of whom are exploring the value of food sovereignty as a concept that might transcend the politics of La via Campesina (LVC) (see Bernstein, 2014; for example). It is now also being written into Northern urban agricultural activism through the work of RUAF and others in Europe – bringing a politics of food and political practices from the South to the North and a new geography to the activist politics at its base. (Edelman (2014) points out LVC is now headquartered in Zimbabwe, after periods in Honduras and Indonesia.)

Food and Development

Placing food in the extreme complexities and embodied experiences of development has always proven challenging, often too challenging for agri-food researchers in Western settings. Among AFRN researchers and others with whom we have contact the recent work of Kirsten Lyons (2014), Bill Pritchard (Pritchard and Connell, 2011), and Charles Mather (2014) exemplify a diversity of approaches. Mather's work over the years has targeted major food groups in Southern and South Africa, especially dairying, fruit and vegetable, wine and fish, with a view to showing the impacts of market liberalization on farm workers. Like Mather, Pritchard (Pritchard's 2013 "The realities of food aid" TEDxSydney broadcast, is a valuable teaching resource) uses the global commodity chain lens to look at the fairness and ethicality of supermarket procurement practices in farming communities in India. Lyons's concerns are more enactive, emphasizing how to mobilize networks for environmentally responsible food production and teasing out the possibilities of democratic engagement and impacts for smallholder farmers.

Institutional Innovation for Food Security

In contradistinction to the local scale food activist literature, others are bringing the intellectual to activism via a different route – an enactive commitment to building concrete and cooperative institutions collaboratively with community groups and local government. In our work with students we use Morgan et al. (2006) reflection that the distanced oppositional stance of political economy was disabling thoughts and performance of alternative thinking and economies to introduce the current enactive research of both Morgan and Marsden. That is, while a defining contribution to agri-food scholarship at a time when the literature desperately needed efforts to bring post-structuralist ideas into dialogue with political economy, the book was not in itself politically engaged. Its concluding provocations, however, informed Marsden's later Chatham House research initiatives on UK food futures (Marsden, 2013) and are embodied in Kevin Morgan's recent institution-building work on school lunches and food policy councils (Morgan, 2007, 2010, 2013). This work demonstrates the value of 'safe' spaces of relative autonomy in mobilizing and stabilizing hybrid collectivities of institutional and individual actors.

Enacting Difference from Knowing Difference

A distinctive program of agri-food research in the 2000s coalesced around detailed empirical exploration of alternative food projects, networks, and practices. In the most obvious expression Maye et al. (2007) showed multiple projects: vegetarianism, local food, slow food, and so on. Located firmly in northern settings, the research focused attention on the multiple politics of alterity. As with the later, similarly themed book (Goodman et al., 2011), Maye and colleagues dance between celebrating a politics of alterity in food production and provisioning and a critique of its partialities (class, geographical, and other), and between focusing on the overlaps between different projects and their contradictions – animated by alterity. The two volumes move well beyond farmers markets and the consumption preferences of young middle class professionals in the United States and European cities to enable a connection between discourses of environmental sustainability and more socioeconomically inspired political projects around food. The multiple chapters in the earlier volume expose a rich variety of place-based initiatives and systematic research, thereby not only celebrating diversity but empowering its enactment. The work established a potential bridge to that of the Australasian inspired Community Economies project (Gibson-Graham, 2008; Cameron et al., 2013) and its particular grounded politics of alterity.

Following as a Methodological Intervention

The landmark paper of Ian Cook (2004), "Follow the thing: Papaya," unhinged many traditions in agri-food research. Distinguished by its lack of formalism, it became the first in a series of papers that pressed open inquiry into consumption-end ethics and moral economy, and implicitly traversed the 'who-what-how-where-why-when' landscape that we see as so vital in future work (Lewis and Rosin, 2013). Most powerfully it signaled overt commitment to exploring practice and micromaterialities, as well as discourses; borrowing from the anthropological gaze and thereby ethnography and practice-centered notions of identity and culture; innovation around performing, journeying, and motion in both foods and text; and thus in relating practices of knowing and materiality. The trilogy carried by *Progress in*

Human Geography (Cook et al., 2011) is startlingly different in their experimental exploration of methodologies of knowledge making. The trajectory brings consumption, its ethics, aesthetics, and its politics onto the stage, together with anxieties over the other and the hands that grew their food. It does so in a socialized form and while it tends to be about individual engagement it provides a language and a sensitivity to/of identity and subject formation around food which is not present in any other of the literature that we study – albeit focused on the affluent consumer. Others working back toward the consumer from production places through distribution and regulatory networks have also deployed a cultural political economy that deploys techniques of following (see, for example, Freidberg, 2004; Jackson, 2010).

Healthy Gendered Foodscapes

Domestic foodscapes and food and health meet in the research of Jane Dixon. The basic questions of equity and human dignity presuppose an enactive disposition (Olsen et al., 2009). This has been important for our students. Dixon takes her work to the heart of the political economy of health (Dixon, 2009), but from very situated positions (Dixon and Banwell, 2004; Dixon and Isaacs, 2013). Her work points to the lively materialities of intrabody engagements with food and the way the materialities are so tightly interwoven with the major questions of political economy. This takes us beyond the fetishization of calories because of the framings she employs. Her work considers how corporate strategy, government policy, and civil society in countries, such as China, Thailand, and Australia influence cultural transitions in food. When Jane visited the University of Auckland in 2009, she encountered both the warmth and earnestness of a postgraduate class dedicated to questioning food and the critical ambivalence of more mainstream health researchers elsewhere in the university to her work.

Reassembling How We Know Foodscapes: An Ontological Politics

These particular projects of knowledge production and their food politics are necessarily partial, situated, and incomplete with respect to incorporating the shifts discussed earlier in the article. In 2006, as this richness of conceptual reframing was coming into view, Morgan et al. (2006) challenged agri-food social scientists to articulate their interests and approaches to what they called 'worlds of food' around a clear politics and a rethinking of power, place, and provenance. They had begun to reassemble insights drawn from different intellectual traditions (actor network approaches, FR thinking, local food activism, and institutional engagement with food actors). The response to the underlying challenge that we emphasize here is that crafted by rural sociologist Michael Carolan, whose own project of reassemblage of thinking about food politics coincided with the 'Worlds of Food' intervention (Carolan, 2006). Carolan has a privileged place in our entry points, having been a highly valued guest of the BE project (Carolan, 2013b) and having spoken as a guest lecturer to one iteration of the Geog738 class.

Carolan's work has brought new insights from environmental politics and post-structural thinking to the food literature (Carolan, 2004, 2006, 2008b, 2009), but from within the grounded traditions of North American rural sociology. He recognizes the crucial place of representing the real, but as a preliminary to informing and moving into ways to perform food worlds differently. In doing so it has taken seriously the invocation of 'who, what, how, where, why, when' in making food knowledge and of enactive research (Carolan, 2013b). This has led to books that break the convenient canons that detune people from the realities of the contemporary world. Grouped, his writings are an intellectual expose on questions such as the real cost of cheap food and the embodied nature of food politics. His work has asked why is 'who' so important a question when it comes to the production of food knowledge (both who is representing foodscapes and who is living them), and offers a humanist moral project around the quality of food. These framings have destabilized taken for granted ontological positions and politics alike, but more in the traditions of an engaged sociology of food than the deconstructionist tradition of cultural theory.

Carolan's *Reclaiming Food Security* (2013c), for example, fundamentally reworks the terrain of food security discussion by introducing the Food and Human Security Index (FHSI). An FHSI score is calculated for 126 countries based on indicators of objective and subjective well-being, nutrition, ecological sustainability, food dependency, and food system market concentration. The final FHSI ranking produces many counterintuitive results. As he notes, why, for example, does Costa Rica top the ranking, while the United States comes in at number 55? The question is a highly political question that connects with the work of food researcher–activists who highlight food inequalities such as food desserts.

Carolan's book is an invitation to do several things: work with demonstrable objects of comparison, including where appropriate with numbers; engage in food knowledge in experimental ways so fracturing existing modes of analysis and framing (Le Heron, 2013); and politicize food by revealing and challenging assumptions and myths. His latest book (Carolan, 2014a) *Cheaponomics: The High Cost of Low Prices* takes aim at the false economy of cheap food and makes available a politics for alternative food movements to bridge the gap to wider audiences and dispel myths that they are out of touch and range of food security questions. In other recent work (Carolan, 2014b), he attempts to shift the discourse of individual level responsibility and commitment from food consumer to food citizen, a project that parallels some of the recent shifts in food sovereignty thinking. For us his work is that of a knowledge activist, translating critical and alternative ideas into coherent and cohesive knowledge projects beyond the academic and research communities, rather than being a political activist of food like Tim Lang (2013) or Raj Patel (2007). In each case, the knowledge they produce about food draws on their situatedness. Lang is deep in the UK and global policy circles, Patel has both worked at the World Bank and protested against it, while Carolan reaches to his theoretical work.

This politics of performative knowledge making adds new purchase to the moral economies of and suggestive politics of

Embodied Food Politics and begins to turn the technologies of control in the new materialities of FR back on power. It is informed, by a politics of social and environmental justice that owes much to shifts in understandings of nature–society relations, the food imperatives of environmental change, the embodied nature of food relations, health, and nutrition, and epistemological demands of post-structuralist ontologies. And in very interesting ways it is mirrored by the recent engagements of Terry Marsden with institutional actors and those of Kevin Morgan with local food activists.

Conclusion

This article emphasizes the inescapable situatedness to the social science of food, and is written to highlight that this must be acknowledged even in an encyclopedia article. Depending upon where you start you will see different things. This is a reality of the politics of place and its representation, the making of relevance, funding materialities, meaningful teaching resources, and the more mundane performativities of institutional setting, publication, and career building. This is how a body of literature is (and must be) built up, such that knowledge making, knowing, and knowledge as text or practice is always situated. In Geog738, our own reading of the different literature and efforts to draw meanings and linkages across them, the reaction of our students to them, and the different ways in which the students have assembled the trajectories in the literature, their own projects, and their association of the literature to material worlds, renders social knowledge of agri-food inherently situated both in its making and its refashioning through learning. The AFRN has nurtured particular place-centric intellectual development, while BE has taken us into enactive research, assemblage of provenance, and experimentation, all of which we are now writing into the agri-food literature. There are myriad agri-food projects, many with radical new lines of inquiry, multiple food courses all with their particular privilegings and relevances, and there are likely other food research networks existing at some kind of level in other places. Our reading and this encyclopedia cannot be the only, the most authoritative, or the final word.

In the politics of food, the challenges of 'what next' in the Anthropocene appear almost overwhelming. There are new imperatives for, and questions of, food safety, security, and sovereignty. Food, energy, and water are positioned at an interconnected core, which we are poorly equipped to address. Place-based food initiatives are contesting the 'who, what, why, and how' of putting food into mouths in place. Community economies and other urban agriculture initiatives offer some hope, as do the renewal of AFNs despite 'cheaponomics' and intensifying pressures on household food budgets, and efforts to link AFN concerns with the practices and possibilities of alterity to via campesina-like concerns with smallholder and collectivized models of agricultural production (Martínez-Torres and Rosset, 2014). This is not to deny political economy, the dominance of corporate FR realities, and the emergence of new forms of 'not food' (be these nonfood products or monster foods). Rather it is to trouble and disrupt them and as Carolan (2013c) observes, to search for new ways of putting the 'alter' into alternative foods.

Taking food as a lens seriously is what we ask our students to do. The article contends we can only do this in situated conditions and our gaze is situated. Partly following from this, every social science discipline offers significant insight. Food research agendas can only be situated, despite the appearance of universal research agendas. What we have shown throughout the article is that there are many initiatives but few connections. We have not talked about pedagogy per se but we have centered the article on situated learning. When interrogated we find situatedness in place is the 'where-when' of living food, a crucial but often forgotten domain of the nature of knowing and of knowledge making and fundamentally related to 'why' we might act to know, what it exists to be known, and how we might come to know. In worlds of food place becomes the site of the actualities of politics and power as agency configures 'who, why, what, how' relations affecting food. This is the insight that Michael Carolan (2004, 2006) harnesses when he speaks of epistemological distances and their disabling influences. Cartesian mappings and distance are always an unacknowledged cartography of situated knowledge production of food. The challenge in the twenty-first century is to connect our way out of the entrapments of a Cartesian mind frame and instead map hopefulness into the realities of fair food for all in places.

See also: Actor-Network Theory; Agricultural Sciences and Technology; Development Geographies; Eco-cities and Sustainable Urbanism; Epistemic Communities and Collaborative Research; Food Deserts in US Cities; Gender and Feminist Studies in Geography; Health, Geography of; Knowledge, Geography of; Labor, Geography of; Nature and Society in Geography; Place in Geography; Postcolonial Geography; Property Rights and Urban Planning; Scale in Geography; Slow Food as a Social Movement; Territoriality: Geographical.

Bibliography

Agarwal, B., 2014. Food sovereignty, food security and democratic choice: critical contradictions, difficult conciliations. The Journal of Peasant Studies. http://dx.doi.org/10.1080/03066150.2013.876996).

Bernstein, H., 2014. Food sovereignty via the 'peasant way': a sceptical view. The Journal of Peasant Studies. http://dx.doi.org/10.1080/03066150.2013.852082.

Blay-Palmer, A., Donald, B., 2006. A tale of three tomatoes: the new food economy in Toronto, Canada. Economic Geography 82, 383–399.

Burch, D., Lawrence, G., 2013. Financialization in agri-food supply chains: private equity and the transformation of the retail sector. Agriculture and Human Values 30, 247–258.

Cameron, J., Healy, S., Gibson-Graham, J.K., 2013. Take Back the Economy: An Ethical Guide for Transforming Our Communities. University of Minnesota Press, Minneapolis.

Campbell, H., 2005. The rise and rise of EurepGAP: the European (re) invention of colonial food relations? International Journal of Sociology Agriculture and Food 13, 6–19.

Campbell, H., 2009. Breaking new ground in food regime theory: corporate environmentalism, ecological feedbacks and the 'food from somewhere' regime? Agriculture and Human Values 26, 309–319.

Campbell, H., Burton, R., Cooper, M., Henry, M., Le Heron, E., Le Heron, R., Lewis, N., Pawson, E., Roche, M., Rosin, C., White, T., 2009. From agricultural science to "biological economies"? New Zealand Journal of Agricultural Research 52, 91–97.

Campbell, H., Dixon, J., 2009. Introduction to the special symposium: reflecting on twenty years of the food regimes approach in agri-food studies. Agriculture and Human Values 26, 261–265.

Campbell, H., Rosin, C., 2011. After the 'organic industrial complex': an ontological expedition through commercial organic agriculture in New Zealand. Journal of Rural Studies 27, 350–361.

Carolan, M., 2004. Ontological politics: mapping a complex environmental problem. Environmental Values 13, 497–522.
Carolan, M., 2006. Do you see what I see? Examining the epistemic barriers to sustainable agriculture. Rural Sociology 71, 232–260.
Carolan, M., 2008a. From patent law to regulation the ontological gerrymandering of biotechnology. Environmental Politics 17, 749–765.
Carolan, M., 2008b. More-than-representational knowledge/s of the countryside: how we think as bodies. Sociologia Ruralis 48, 408–422.
Carolan, M., 2013a. The wild side of Agro-food studies: on Co-experimentation, politics, change, and hope. Sociologia Ruralis. http://dx.doi.org/10.1111/soru.12020.
Carolan, M., 2013b. Doing and enacting economies of value: thinking through the assemblage. New Zealand Geographer 69, 176–179.
Carolan, M., 2013c. Reclaiming Food Security. Earthscan/Routledge, London and New York.
Carolan, M., 2013d. Final word: putting the "alter" in alternative food futures. New Zealand Sociology 28 (4), 145.
Carolan, M., 2014a. Cheaponomics: The High Cost of Low Price. Earthscan/Routledge, London and New York.
Carolan, M., 2014b. Future food "Needs": from consumer to citizen choice. Sociologia Ruralis 54, 98–100.
Clapp, J., 2014. Financialization, distance and global food politics. The Journal of Peasant Studies. http://dx.doi.org/10.1080/03066150.2013.875536.
Cook, I., 2004. Follow the thing: papaya. Antipode 36, 642–664.
Cook, I., Hobson, K., Hallett, L., Guthman, J., Murphy, A., Hulme, A., et al., 2011. Geographies of food: afters. Progress in Human Geography 35, 104–120.
Dixon, J., 2002. The Changing Chicken: Chooks, Cooks and Culinary Culture. University of New South Wales Press, Sydney.
Dixon, J., 2009. From the imperial to the empty calorie: how nutrition relations underpin food regime transitions. Agriculture and Human Values 26, 321–333.
Dixon, J., Banwell, C., 2004. Heading the table: parenting and the junior consumer. British Food Journal 106, 182–193.
Dixon, J., Isaacs, B., 2013. There's certainly a lot of hurting out there: navigating the trolley of progress down the supermarket aisle. Agriculture and Human Values 30, 283–297.
DuPuis, M., Harrison, J., Goodman, D., 2011. Just food? In: Alkon, A.H., Agyeman, J. (Eds.), Cultivating Food Justice: Race, Class, and Sustainability. The MIT Press, Cambridge, MA, pp. 283–307.
Edelman, M., 2014. Food sovereignty: forgotten genealogies and future regulatory challenges. The Journal of Peasant Studies. http://dx.doi.org/10.1080/03066150.2013.876998.
Enticott, G., 2003. Risking the rural: nature, morality and the consumption of unpasteurised milk. Journal of Rural Studies 19, 411–424.
Freidberg, S., 2004. French Beans and Food Scares: Culture and Commerce in an Anxious Age. Oxford University Press, Oxford.
Friedmann, H., 1993. The political economy of food: a global crisis. New Left Review 197, 29–57.
Friedmann, H., 2000. What on Earth is the modern world-system? Food-getting and territory in the modern era and beyond. Journal of World-System Research VI (2), 480–515.
Friedmann, H., 2005. From colonialism to green capitalism: social movements and the emergence of food regimes. In: Buttel, F., McMichael, P. (Eds.), New Directions in the Sociology of Global Development. Elsevier, Amsterdam, pp. 227–264.
Friedmann, H., McMichael, P., 1989. Agriculture and the state system. The rise and decline of national agriculture. Sociologia Ruralis 19, 93–117.
Gertel, J., 2005. Inscribed bodies within commodity chains: global wheat and local insecurity. In: Fold, N., Pritchard, B. (Eds.), Cross-continental Agro-food Chains. Structures, Actors and Dynamics of the Global Food System. Routledge, Oxford, pp. 109–124.
Gertel, J., Sippel, S. (Eds.), 2014. Seasonal Workers in Mediterranean Agriculture: The Social Costs of Eating Fresh. Routledge, Earthscan.
Gibson-Graham, J.K., 2008. Diverse economies: performative practices for other worlds. Progress in Human Geography 32, 613–632.
Goodman, D., Dupuis, M., Goodman, M., 2011. Alternative Food Networks: Knowledge, Practice and Politics. Routledge, London.
Goodman, M., Maye, D., Holloway, L., 2010. Ethical foodscapes?: premises, promises and possibilities. Environment and Planning A 42, 1782–1796.
Guthman, J., 2009. Teaching the politics of obesity: Insights into neoliberal embodiment and contemporary biopolitics. Antipode 41 (5), 1110–1133.
Guthman, J., DuPuis, M., 2006. Embodying neoliberalism: economy, culture, and the politics of fat. Environment and Planning D 24, 427.
Jackson, P., 2010. Food stories: consumption in an age of anxiety. Cultural geographies 17, 147–165.
Jackson, P., Ward, N., Russell, P., 2008. Moral economies of food and geographies of responsibility. Transactions of the Institute of British Geographers 34, 12–24.
Lang, T., 2013. Food Matters: An Integrative Approach to Food Policy. France, Paris (OECD).
Le Heron, R., 2013. Emerging neo-liberalising processes in New Zealand's land-based sector: a post structural political economy framing using emergence diagrams. Applied Geography 45, 392–401.
Lewis, N., 2009. Thinking/learning about policy to enact different food futures. Regional Science Policy & Practice 1 (2), 187–192.
Lewis, N., Le Heron, R., Campbell, H., Henry, M., Le Heron, E., Pawson Perkins, H., Roche, M., Rosin, C., 2013. Assembling biological economies: region-shaping initiatives in making and retaining value. New Zealand Geographer 69, 180–196.
Lewis, N., Rosin, C., 2013. Epilogue: emergent (re-) assemblings of biological economies. New Zealand Geographer 69, 249–256.
Lockie, S., 2009. Responsibility and agency within alternative food networks: assembling the "citizen consumer". Agriculture and Human Values 26, 193–201.
Lyons K., 2014. Critical engagement, activist/academic subjectivities and organic agri-food research in Uganda. Local Environment, 19(1), 103–117.
Marsden, T., 2013. From post-productionism to reflexive governance: contested transitions in securing more sustainable food futures. Journal of Rural Studies 29, 123–134.
Martínez-Torres, M., Rosset, P., 2014. Diálogo de saberes in La Vía Campesina: food sovereignty and agroecology. The Journal of Peasant Studies. http://dx.doi.org/10.1080/03066150.2013.872632.
Mather, C., 2014. Avian influenza multiple: enacting realities and dealing with policies in South Africa's farmed ostrich sector. Journal of Rural Studies 33, 99–106.
Maye, D., Holloway, L., Kneafsey, M. (Eds.), 2007. Alternative Food Geographies. Representation and Practice. Elsevier Science, Amsterdam.
McMichael, P., 1992. Tensions between national and international control of the world food order: contours of a new food regime. Sociological Perspectives 35, 343–365.
McMichael, P., 2005. Corporate development and the corporate food regime. In: Buttel, F., McMichael, P. (Eds.), New Directions in the Sociology of Global Development. Research in Rural Sociology and Development. Elsevier, Amsterdam, pp. 265–299.
McMichael, P., 2008. The peasant as 'canary': not too early warnings of global catastrophe. Development 51, 504–511.
McMichael, P., 2009. A food regime genealogy. Journal of Peasant Studies 36 (1), 139–169.
McMichael, P., 2014. Historicizing food sovereignty. Journal of Peasant Studies 41 (6), 933–957.
Miele, M., 2006. Consumption culture: the case of food. Handbook of rural studies 344–354.
Miele, M., 2011. The taste of happiness: free-range chicken. Environment and Planning A 43, 2076–2090.
Miele, M., 2013. Religious slaughter: promoting a dialogue about the welfare of animals at time of killing. Society and Animals 21, 421–424.
Morgan, K., 2007. The polycentric state: new spaces of empowerment and engagement? Regional Studies 41, 1237–1251.
Morgan, K., 2010. Local and green, global and fair: the ethical foodscape and the politics of care. Environment and Planning A 42, 1852–1867.
Morgan, K., 2013. The rise of urban food planning. International Planning Studies 18, 1–4.
Morgan, K., Marsden, T., Murdoch, J., 2006. Worlds of Food: Place, Power, and Provenance in the Food Chain. Oxford University Press, Oxford.
Olsen, A., Dixon, J., Banwell, C., Baker, P., 2009. Weighing it up: the missing social inequalities dimension in Australian obesity policy discourse. Health Promotion Journal of Australia 20, 167–171.
Patel, R., 2007. Stuffed and Starved. Markets, Power and the Hidden Battle over the World's Food System. Portobello Books, London.
Pritchard, W., 1998. The emerging contours of the third food regime: evidence from Australian dairy and wheat sectors. Economic Geography 74, 64–74.
Pritchard, W., 2013. Food chains. In: Murcott, A., Belasco, W., Jackson, P. (Eds.), The Handbook of Food Research. Bloomsbury Academic, USA, pp. 167–176.
Pritchard, B., Connell, J., 2011. Contract farming and the remaking of agrarian landscapes: insights from South India's chilli belt. Singapore Journal of Tropical Geography 32, 236–252.
Whatmore, S., 2006. Materialist returns: practising cultural geography in and for a more-than-human world. Cultural Geographies 13, 600–609.
Woods, M., 2007. Engaging the global countryside: globalization, hybridity and the reconstitution of rural place. Progress in Human Geography 31, 485–507.

Geography: Overview

Ron Johnston, University of Bristol, Bristol, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The practice of geography – literally 'earth writing,' from the Greek *geo* and *graphein* – extends back at least 2000 years, but it has a much shorter history as an academic discipline, with little formal presence in either universities or schools before the twentieth century. Its key concepts are environment, place, scale, and space and it seeks to understand not only what is where, and why, but also how people experience, and behave in, different places, and why. Although most definitions (and university departments) of geography embrace both human and physical worlds, studies of human and physical geography are now substantially separate components of the discipline. In any period, the nature of geography and geographical practices – what it studies, how and why – is conditioned by the contexts of time and place. Geography emerged as an academic discipline just over a century ago out of a subject that had long been practiced without formal institutionalization in universities. Its evolution since then has passed through various stages as its practitioners respond to society's needs – in all levels of education, and in fundamental and applied research, for example. Geographers seek not only to understand the earth and its human occupance but also to improve it.

The practice of geography – literally 'earth writing,' from the Greek *geo* and *graphein* – extends back at least 2000 years, but it has a much shorter history as an academic discipline, with little formal presence in either universities or schools before the twentieth century. There have been many attempts to define it: one widely used in the mid-twentieth century saw it "concerned to provide accurate, orderly, and rational description and interpretation of the variable character of the Earth surface" (Hartshorne, 1939: p. 21) but its focus is now much wider. Its key concepts are environment, place, scale, and space and it seeks to understand not only what is where, and why, but also how people experience, and behave in, different places, and why.

Although most definitions (and university departments) of geography embrace both human and physical worlds, studies of human and physical geography are now substantially separate components of the discipline. This article focuses on human geography, that part normally affiliated with the social sciences and the humanities, while recognizing a recent partial coming together of the two sections, reflecting not only contemporary concerns with environmental change, its causes and impacts, but also the realization that separating humans and nature creates a false dichotomy – each is a part of the other (Whatmore, 2002).

In any period, the nature of geography and geographical practices – what it studies, how and why – is, like that of many other disciplines, conditioned by the contexts of time and place. Thus, as outlined here, geography emerged as an academic discipline just over a century ago out of a subject that had long been practiced without formal institutionalization in universities. Its evolution since then has passed through various stages as its practitioners respond to society's needs – in all levels of education, and in fundamental and applied research, for example. Geographers seek not only to understand the earth and its human occupance but also to improve it.

Intellectual Origins

As a source of information about the earth, geography has been practiced in all societies – if not always by that name. In most of them, creating and storing locational information involved some form of cartography, as the multivolume history of cartography project is revealing (Harley and Woodward, 1987). Mapmaking was part of the geography taught at Cambridge and Oxford universities in the late sixteenth century, for example (Cormack, 1997), and, along with map interpretation, was central to most degree curricula until well into the twentieth century. A second major strand of early geographical education was chorology, description of different physical environments and their inhabitants. Geographers portrayed the rich mosaic of world regions – large and small – to meet the needs of traders, diplomats, and the military as well as whet the curiosity of the 'educated classes.' Much of this development of the subject was closely linked to imperial projects – both formal, through colonization, and informal (Driver, 1998; Kearns, 2009; Morin, 2011) – and many university geography departments were established in British Commonwealth countries during the twentieth century.

In the early- and mid-nineteenth century geographical societies were founded in European and North American cities to promote geography, in particular exploration (Livingstone, 1992; Bell et al., 1995), and its commercial applications (Chisholm, 1885). Some societies later actively promoted geographical education: in the United Kingdom, for example, the Royal Geographical Society realized that geography could only be successfully introduced as a school subject if it was first established as a separate academic discipline in the country's ancient universities, where it funded some of the first appointments of geographers (Stoddart, 1986).

Preceding the British developments, the first university chairs and departments in the new discipline were established by royal decree in Germany in the 1870s: the departments were to prepare school teachers, as part of the creation of an educated citizenry, and sustain colonial and military enterprises. (The first established chair in the Netherlands was explicitly created in 'colonial geography.') In the United Kingdom, the first separate honors degree in geography was only established in 1917 – at Liverpool. The British departments then were small, with few teaching staff (for whom research was not a priority) and students – most of whom

became secondary (mainly grammar) school teachers after graduation. Virtually every UK university had a separate department of geography by 1950, however, and they formed the foundations for major expansion thereafter (Johnston and Williams, 2003). There were parallel developments in the United States, except that geography never gained a substantial presence in high school curricula there (Schulten, 2001).

Creation of a Discipline

The subject matter taught in these new departments focused on people–environment interrelations, as conceived by several pioneering German and French geographers. Among them, the early works of, for example, Varenius (Mayhew, 2010) and Kant (see May 1970; Harvey, 2009; Elden and Mendieta, 2011), followed in the nineteenth century by Hettner (see Hartshorne, 1939: Harvey and Wardenga, 2006), Humboldt (1845–1972; see Bowen, 1981), Ritter (1817–1859; Linke, 1981), Ratzel (1882–1891), and Vidal de la Blache (1903; Buttimer, 1971) became disciplinary canons. Some of these pioneers' followers promoted their ideas oversimplistically (such as the environmental determinism advanced by one of Ratzel's American disciples, Ellen Churchill Semple, 1911 – see Keighren, 2010) but most portrayed geography's subject matter as the myriad ways in which the nature–society interactions created different landscapes (as Mackinder – the first appointed geographer at the University of Oxford – defined the discipline in 1887; his definition was paralleled by US contemporaries – e.g., Davis, 1909).

According to Freeman (1961), up to the 1950s geography was characterized by six separate but linked concerns: encyclopedic, colonial, educational, political, generalizing, and specialization. Subsequently, the first two continued their long-term decline (exploration of terrae incognitae by European societies had reached its near-conclusion and decolonization was rapid) and the last two came to dominate: the educational and political components changed to reflect these shifts. The emerging practices were strongly influenced by their national and local milieux (Livingstone, 1992; Livingstone and Withers, 2011), but for a few mid-twentieth-century decades the disciplinary core was widely identified as the study of regions, relatively homogeneous segments of the earth's surface defined on criteria reflecting the natural environment and its modification by humans. In the United States, Hartshorne (1939: p. 462) presented geography as

> … a science that interprets the realities of areal differentiation of the world as they are found, not only in terms of the differences in certain things from place to place, but also in terms of the total combination of phenomena in each place, different from those at every other place.

Regions are unique areas, and so, according to a British geographer (Wooldridge, 1956: p. 53), regional geography – promoted 25 years later as 'the highest form of the geographer's art' (Hart, 1982) – aimed

> … to gather up the disparate strands of the systematic studies, the geographical aspects of other disciplines, into a coherent and focused unity, to see nature and nurture, physique and personality as closely related and inter-dependent elements in specific regions.

This identifies two long-recognized strands in geographical practice (stemming from the writings of Humboldt and Ritter respectively): the systematic – focusing on particular subject matter – and the regional, synthesizing material on specific areas (at varying scales). The European roots weakened after 1945, however (Johnston and Claval, 1984), and a specific Anglo-American discipline emerged (albeit not without national differences: Johnston and Sidaway, 2004): 'Anglo' dominance is now the subject of considerable contemporary criticism in other language realms – although in some of them (the Scandinavian countries and the Netherlands being the main examples) English is now the dominant language in geographical research and, increasingly, teaching too (Johnston and Sidaway, 2015).

A 'New' Geography

Systematic specialization increasingly dominated the expanding academic discipline from 1950 on, involving two intradisciplinary divides: a split between physical and human geography, reflecting their growing links to the natural and social sciences respectively; and the creation of separate specialisms within each. Regional geography declined rapidly as a subfield, although some scholars continued to maintain the importance of each geographer becoming an expert on a specific (usually 'foreign') part of the world (Mead, 1963); by the century's end it was only a minuscule segment of the discipline. Individual human geographers increasingly defined themselves by their systematic specialisms: historical geography, led by Darby (2002), was (and remains) especially strong in the United Kingdom, as was cultural geography – led by Sauer (1952), with strong links to anthropology – in the United States. Economic, political, and social geography were small enterprises until the century's last quarter.

The 1950s/1960s were characterized by the discipline's so-called 'quantitative and theoretical revolutions' (Burton, 1962). These involved both human and physical geography, leading to a brief period of intradisciplinary collaboration focused on methods. Among human geographers, the focus was the identification of spatial order in human occupance of the earth's surface, rather than a concentration on the unique characteristics of individual regions. Schaefer (1952) was among the first to promote geography as a spatial science, a nomothetic (as against an ideographic) orientation that placed the discipline within the burgeoning – quantitatively oriented – social sciences, whose findings were having an increasing impact on public policies. The 'revolutions' involved embracing a particular form of scientific method (positivism) that was theory-led and involved the application of quantitative methods to test hypotheses and draw general (lawlike?) conclusions (Harvey, 1969; Amedeo and Golledge, 1975). Geography's niche within the social sciences was to be location theory, and its identifying rationale a focus on space as an influence on (determinant of) individual and collective behavior; geography was to be a 'discipline in distance' with geometry the key mathematical stimulus (Bunge, 1966; see also

Sack, 1981). Space was usually interpreted as the cost involved in the movement of goods, people and information, and classic theories – most initiated outwith geography – of land use patterns (von Thünen: Chisholm, 1962; Hall, 1966), central places (Christaller, 1966), industrial location (Smith, 1981), and gravity models (Olsson, 1965), for example – were appropriated as the foundations for this 'new geography'; statistical modeling was adopted as the means for hypothesis-testing to identify general patterns in observed geographies.

The 'revolutions' were initially North American in provenance, rapidly spreading from a few innovation centers (Johnston and Sidaway, 2015), but soon infiltrated British geography, notably through two seminal anthologies (Chorley and Haggett, 1965, 1967). Further spatial theories were added – particularly the spatial diffusion of innovations (Hägerstrand, 1968) – and the entire corpus synthesized in influential textbooks (notably Haggett, 1965; Abler et al., 1971) which covered the relevant theories and their empirical explication plus the methods to be deployed in spatial analysis. Within those general frameworks, systematic specialisms rapidly emerged, initially in economic and urban geography, within each of which separate, overlapping, sub-subdisciplines were formed: economic geography spawned industrial, agricultural, transport, and development geographies, for example: urban geography begat a substantial urban social geography and was paralleled by increased interest in rural geography; out of social geography emerged both population and medical (later health) geographies; and a revived political geography incorporated electoral geography. By the late 1970s these various disciplinary components achieved institutional status through the creation of specialist groups operating within the overall umbrella of the major learned societies, such as the Association of American Geographers and the Institute of British Geographers; the groups held their own conferences and sessions within larger meetings, and published their own journals alongside the longer-established generalist serials.

This burgeoning activity was possible because the discipline was expanding rapidly within universities in response to increased student demand. The revolutions' promoters achieved major intradisciplinary change within a relatively short time by portraying spatial science as filling contemporary gaps within the social sciences, while at the same time criticizing, even denigrating, both regional geography and those parts of the discipline which failed to be converted to the new ways (Johnston, 2006).

'-isms,' 'Schisms,' and 'Turns'

Although some promoted this 'new geography' as surpassing its predecessor, there was no complete revolution to a new hegemonic paradigm in Kuhnian terms. Some geographers continued to teach and research in their established ways, sustaining the traditions of historical, cultural, and (to a lesser extent) regional geography. More importantly, spatial science was soon the subject of two main criticisms. The first attacked the narrow economism of location theories which privileged profit-maximizing (expressed in geographical terms as transport-cost-minimizing) as a determinant of decision-making, thereby (at least implicitly) denigrating the wide range of other cultural influences on behavior. Spatial science, it was argued, assumed that individuals respond to stimuli in predetermined ways, and the elevation of space as a predominant influence led to a derogation of place – the context of so much human learning – and the importance of human free will. Some critics explored alternative philosophies to positivism – such as phenomenology, idealism, and existentialism – in work characterized as humanistic geography which emphasized the variety of human decision-making criteria and behaviors. Although those alternatives attracted few permanent adherents, they presaged an exploration of a much wider range of approaches to social theory (Jackson and Smith, 1984). This became known as geography's 'cultural turn'; it was not a return to Sauerian cultural geography and its emphasis on landscapes, however, but a new focus on the contingent nature of much individual and group behavior – with much of that contingency being place-based. It involved greater interaction with not only other social sciences but also the humanities, within which a greater appreciation of the role of places as crucial contextual locales led, in turn, to a 'geographical or spatial turn' within some of those disciplines.

The second set of criticisms concerned spatial science's approach to explanation. Location theories, it was argued, could not account for the major contours of the geographies of development and underdevelopment at various scales, of inequality and injustice, wealth, power, disadvantage, and discrimination: the theories were apolitical and could not be used to achieve fundamental change in an inherently unequal socioeconomic mode of production. Explaining those geographies required an appreciation of how economic systems – especially capitalism – operate. A major stimulus to such arguments was Marxism, notably through Harvey's (1973) provocative essays, which presaged a series of seminal volumes building a historical materialist geography (e.g., Harvey, 1982, 1996). This body of work became known as radical geography, which spawned its own journal – *Antipode*. Not all who accepted the critique's key arguments were prepared to adopt the Marxist political program, but the shift to studying capitalism as a system stimulated work in (spatial) political economy more generally, which became a major focus of work in economic geography.

Both critiques of spatial science attracted adherents, but others expressed dissatisfaction with the alternatives. As Gregory (1978) argued, whereas some spatial science strayed very close to naive economic determinism, some humanistic geography approached another extreme – of voluntarism, tearing the individual out of context; meanwhile, some radical geography offered little more than an alternative economic determinism. But those three broad approaches – spatial science, radical-structuralism, and humanistic geography – became the cores of much debate over subsequent decades as the discipline divided – if not fractured – into increasingly separate, sometimes antagonistic, components.

The Contemporary Discipline

Sheppard (1995) divided human geographers into two main groups – spatial scientists and social theorists. Although each – especially the second – embraces a wide range of subject matter and approaches to the discipline, nevertheless that binary

divide encapsulates the two major strands of contemporary scholarship, characterized, among other ways, by their separate methodological foundations – quantitative and qualitative, respectively. Such is the divide between them that few introductory human geography textbooks now fully cover both and some either explicitly dismiss spatial science as part of a 'discredited past' (Peet, 1998) or simply ignore it (e.g., Cloke et al., 1999): against that, US volumes promoting the discipline to the national scientific research community focus almost entirely on spatial science (NRC, 1997, 2010). Some encyclopedia (e.g., Kitchin and Thrift, 2009) and dictionaries (e.g., Gregory et al., 2009) are more all-embracing but attempts at comprehensive overviews lack integration (e.g., Gaile and Willmott, 1989, 2001) and the editors of a volume on the "pervasive themes [and] common elements within a discipline whose practitioners are in danger of forgetting their shared heritage and ideals" concluded presciently that not only had they failed to bridge the human–physical divide but they also feared a similar split opening up within human geography (Abler et al., 1992). One popular text – which appeared in several editions – did successfully combine perspectives from human and physical geography, though most of the material from the former was of the spatial science variety (Haggett, 1972, 1990).

Spatial Science

Spatial science remains a substantial minority component of contemporary human geography, but its orientation has changed considerably over the last half-century. It is strongly, and increasingly sophisticatedly, quantitative but the formal (geometrical) location theories based on a single causal variable (distance) have largely been abandoned: the search for spatial order neither anticipates the discovery of regular patterns in the landscape nor seeks universal laws of human spatial behavior. Space has been replaced by place as the key underpinning geographical concept – increasingly recognizing its multiple scales. Sayer (1984) drew an important distinction between extensive and intensive research: the former seeks empirical regularities whereas the latter explores the causal chains responsible for particular outcomes. Much spatial science is extensive, a necessary precursor to later intensive investigations: by not eschewing empirical generalizations it identifies significant features and trends in the mass of numerical data that characterize modern societies. An underpinning assumption of much of the research in this mold is that as nonexperimental disciplines the social sciences cannot establish causality in the same way as the natural and some of the environmental sciences: what matters is the quality of the theory underpinning the empirical investigations: "Causal inference resides at a theoretical level and is not something that comes directly from the data" (Sampson, 2011: p. 380).

Geographers in the 1950s to 1970s assumed that standard statistical procedures could be applied unproblematically to the analysis of point, line, flow, and areal data. This was challenged by work on two particular technical areas:

- spatial autocorrelation: particular features of spatial data encapsulated in Tobler's first law of geography – "Everything is related to everything else, but near things are more related than distant things": Sui et al. (2004) require bespoke analytical procedures to produce valid results (Cliff and Ord, 1973); and
- the modifiable areal unit problem which showed that scale and aggregation effects both substantially impact on the results of analyses of areal data (such as population characteristics) of small areas: Openshaw (1984) demonstrated that different aggregations of small areas into larger blocks can produce divergent statistical results – a finding applied in Gudgin and Taylor's (1979) pioneering modeling of the impact of different districting schemes on election outcomes.

Building on these foundations, geographers developed a range of 'local statistical methods' to identify significant spatial patterns (Anselin, 1988), and adapted other methods – such as multilevel modeling (Jones, 1991), which integrates local variations and autocorrelation at a variety of spatial scales.

These increasingly sophisticated changes in spatial science were facilitated by – indeed would have been impossible without – developments in computing power, stimulating the conception of geocomputation to encompass such work in both human and physical geography (Longley et al., 1998); there is an annual international geocomputation conference and a National Centre for Geocomputation has been established at the National University of Ireland, Maynooth. More powerful and faster computers allowed analysis of increasingly large and complex data sets – including those generated through remote sensing. The most significant developments were in geographical information systems (GIS), combined hardware and software for the organization, integration, analysis, and display of spatially referenced data, with the display media underpinning the growth of visualization strategies and the rapid production of maps and diagrams: the satellite and keyboard have replaced the field and drawing board in cartography. Alongside their widespread research use, GIS have underpinned a range of commercial applications, and their development has led to the identification of a geographical information science (Goodchild, 1992). Their application has stimulated many teaching programs providing graduates with skills relevant to a range of potential careers (NRC, 1997, 2010): many textbooks (e.g., Longley et al., 2005) serve those audiences and geography/ers are at the core of a substantial service industry.

Social Theory

Although spatial science remains a substantial and vibrant component of contemporary human geography, to many the approaches grouped together under the 'social theory' heading now form the disciplinary mainstream (e.g., Massey et al., 1999). Humanistic and radical geographies have metamorphosed into new forms – although some maintain allegiance to the original formulations, as in Harvey's (2005a) continued reworking of Marxist theories in contemporary contexts.

As in spatial science, place (rather than space: Taylor, 1999; discusses the tensions between the two) and scale are the key underpinning concepts: places are the milieux within which

differences are produced and reproduced, within which people learn, act, and influence others. But places are not fixed containers: they too are produced and continually reproduced yet changed through human action. In studying people–place interactions, geographers were influenced by Giddens' (1977) concept of structuration: each is continually transforming the other – people make their own histories but within spatial contexts that are both constraining and enabling. (Giddens',1984 – work was an early example of the 'spatial turn' in the other social sciences, influenced by Hägerstrand's, 1982 – time geography, which drew attention to the geographical constraints on how individuals pursue their personal and collective projects.)

A very important stimulus to appreciating the role of milieu/locale in the structuration of difference came from feminist geography. Initially launched in the 1980s to highlight discrimination against women within a male-dominated and male-oriented discipline, and the consequent silences in what geographers studied, this extended into a much wider critique of 'single-factor' approaches, such as Marxists' concentration on class and the economic infrastructure that ignored other important social cleavages (Women and Geography Study group, 1984, 1997; Rose, 1993). People socialized into, and viewing society from, different positions create separate identities of self and other, which are potential foundations for identity politics. The class struggle is not the only cleavage within capitalist society and others (based on not only sex but also race, religion, and sexual orientation, for example) cannot be folded within it. Indeed, these various cleavages interact; a realization emerged within feminist geography, for example, that women's identity politics in more 'developed' countries differed from their contemporaries' in excolonies.

Other stimuli came from postmodernism, which also emphasized difference and rejected the search for grand theories. Soja (1989) and Dear (1998) introduced it to geographers as a way of addressing the argued transition from a Fordist regime of accumulation characterized by production of uniform commodities for mass consumption to post-Fordism and its small niche markets with consumption for status as well as use – sometimes termed flexible accumulation – in which tastes changed frequently, reflecting local concerns and influences.

More important have been impulses from post-structuralist arguments that human subjectivity is not pregiven but rather produced in contexts: geographers have argued that those contexts are necessarily spatially and temporally contingent, so that concepts of truth and knowledge are too. Representation of these contingencies is crucial to social conduct, and post-structuralists stress the importance of language not only as a communication medium but also as a social construction itself (as in the relationships between signified and signifier and their interpersonal transmission through hermeneutics). Language is only one means of representation and communication, however, and alongside the interpretation of written texts geographers have focused on other images – works of art, maps, and landscapes – implicated in the production of situated knowledges (Barnes and Duncan, 1992; Cosgrove, 1998). But not all events and interactions are recorded let alone 'archived' for later analysis: much transitory but meaningful spatial behavior goes unrecorded, and analysis of such practices has led geographers into nonrepresentational theory and studies of performativity (Thrift, 2007).

Whereas much of the work stimulated within human geography's 'cultural turn' focuses on the individual condition and small scale events, including the body as a site of identity-formation (the most intimate of 'personal spaces' – or 'the geography closest in': Pile, 1996), there have also been changes in larger-scale concerns. Work in geopolitics, for example, has been much influenced by Said's (1978) Orientalism – the West's creation of a 'mythical East' – as a stimulus to investigations of the role of images of places and their peoples and, as illustrated by Gregory's (2004) book on Afghanistan, Iraq, and Israel–Palestine, their influence on political actions. Territory and territoriality are key in the spatial construction of these geographies (Sack, 1986; Elden, 2009, 2013), which embrace postcolonial movements that challenge not only the economic and political components of colonialism and imperialism (and geographers' participation therein – Smith, 2003; Morin, 2011) but also their roles in creating imaginative geographies at variance with those of people who experienced that external power: the latter are now creating their own alternative imaginative geographies, knowledges, and identities which articulate their desires, fantasies, and fears. As Harvey (2005b) argued, there is a myriad geographical imaginations, created worlds which people not only occupy but also portray to others: 'professional' geographers are not the only creators and purveyors of geographies – the editors of popular geography magazines present the world in different ways, as do travel agents and film/TV producers.

Geography in Action

Many geographers have long put their knowledge and expertise to use beyond the research laboratory and library in what has been generally termed 'applied geography.' After the 'age of exploration' had passed, geographers' skills in mapping and regional synthesis were promoted as valuable foundations for spatial planning, and this continued with the claimed utility of spatial science to, for example, identify optimal locations for shopping centers and transport links. From the late twentieth century on major changes in the political environment within which universities operated involved increased pressure to undertake research which had applied value to both the state apparatus and, particularly, private sector 'wealth creators' (Breheny, 1989).

One area of geography which this growing emphasis on applied work stimulated was GIS, presented as a key set of tools that geographers could apply and transferable skills that they could inculcate in their students (NRC, 1997). Some of the applications continue to advance the 'geometrical' approach to spatial science and the optimal location of facilities (Birkin et al., 2010); others illustrate the turn to place-based analyses – in geodemographics, for example, the classification of small areas according to their population characteristics to facilitate targeting of potential consumers for particular goods and services (Harris et al., 2005). Early critiques of GIS argued that they were readily implicated in the exercise of both commercial and geopolitical power (Pickles, 1995), a trend since partially countered by the development of 'participatory

GIS,' intended to equalize access to the relevant resources (Elwood, 2006), and 'neogeography,' which provides online tools to enable mapmaking by nonprofessionals (Goodchild, 2009). Many GIS applications involve human–environment interactions, such as changes in land cover, the management of physical resources, environmental changes, and risk analyses. These involve more contact between human and physical geographers than characterized previous decades (Pappenberger et al., 2012; Landstrom et al., 2011) but are largely separate from the growing interest in wider issues regarding nature (Castree, 2005; Whatmore, 2002). Much of the work undertaken for private sector sponsors is not published in academic journals for commercial reasons, but papers on environmental management and optimal locational decision-making dominate the journal *Applied Geography*.

Outwith spatial science, there are fewer potential applications of geographical expertise as generally understood, but this does not preclude engagement between professional geographers and the worlds they study. There has been growing interest, for example, in participatory geographies, action-orientated research involving engagement with, rather than just study of, individuals, groups, and communities (Pain, 2003); the research itself is collaborative and involves not just enhancing appreciation of situations but assisting campaigns for change. These have their roots in the anarchist writings of Kropotkin and Reclus (Springer, 2012), later developed by Bunge (1971) as 'geographical expeditions.' Whereas such work is emancipatory in its aims, its context is generally local, although Harvey links grassroots movements for change with the larger goal of replacing capitalism (Harvey, 2012).

More generally, there has been a substantial interest in 'critical human geographies,' the subject of a series of major international conferences focused on critical and emancipatory discussion aimed at radical explanation and change. These place as much emphasis on the 'ought' as the 'is,' on not only unmasking power in all its forms and the inequality it breeds but also emancipating people so that they are not just aware of how spaces and places are being manipulated but can also develop geographies of resistance (Pile and Keith, 1997; Blunt and Wills, 2000). Linked to this is work on the geographies of justice and ethics (Smith, 2000). Geography in such contexts is being applied in the interest of transformation, reflecting Harvey's earlier arguments that applications of spatial science may alleviate some superficial problems but not contribute toward the creation of a different, fairer world (Harvey, 2000).

Conclusion

Human geography is a large, diverse, expanding, and vibrant discipline characterized, like most others, by fragmentation and tensions – it is divided philosophically, methodologically, and politically. It has an established niche within the academic division of labor, with many productive interdisciplinary interactions: its core concerns – environment, place, space, scale – crystallized over a century ago remain its foundations, but its practices – what it studies, how and why – have changed markedly (as illustrated by the five editions of *The Dictionary of Human Geography*; Gregory et al., 2009). Buttimer (1993) has suggested that throughout their discipline's history, geographers have deployed four foundational metaphors (mosaic, machine, organism, and arena), with each more relied upon in some periods than others – though without the academic revolutions associated with Kuhn's (1970) paradigm model of the history of the natural sciences. Contemporary spatial science employs the mosaic metaphor (having largely abandoned the machine metaphor used in earlier decades to account for the mapped mosaics); contemporary social theory concentrates on organism and area – the spatial contexts (places, locales, milieux) implicated in the production of humanness while themselves being produced and reproduced by human action.

See also: Location Theory; Marxist Geography; Place in Geography; Regional Geography; Regional Science; Space and Social Theory in Geography.

Bibliography

Abler, R.F., Adams, J., Gould, P.R., 1971. Spatial Organization: The Geographer's View of the World. Prentice-Hall, Englewood Cliffs, NJ.
Abler, R.F., Marcus, M.G., Olson, J.M. (Eds.), 1992. Geography's Inner Worlds: Pervasive Themes in Contemporary American Geography. Rutgers University Press, New Brunswick, NJ.
Amedeo, D., Golledge, R.G., 1975. An Introduction to Scientific Reasoning in Geography. John Wiley, New York.
Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer, Dordrecht.
Barnes, T.J., Duncan, J.S. (Eds.), 1992. Writing Worlds: Discourse, Text and Metaphor in the Representation of Landscape. Routledge, London.
Bell, M., Butlin, R., Heffernan, M. (Eds.), 1995. Geography and Imperialism: 1820–1940. Manchester University Press, Manchester.
Birkin, M., Clarke, G., Clarke, M., 2010. Refining and operationalizing entropy-maximizing models for business applications. Geographical Analysis 42, 422–445.
Blunt, A., Wills, J., 2000. Dissident Geographies: An Introduction to Radical Ideas and Practice. Prentice-Hall, Harlow.
Bowen, M., 1981. Empiricism and Geographical Thought from Francis Bacon to Alexander von Humboldt. Cambridge University Press, Cambridge.
Breheny, M., 1989. Chalkface to coalface: a review of the academic-practice interface. Environment and Planning B: Planning and Design 16, 451–468.
Bunge, W., 1966. Theoretical Geography. C.W.K. Gleerup, Lind.
Bunge, W., 1971. Fitzgerald: Geography of a Revolution. Schlenkman, Cambridge, MA.
Burton, I., 1962. The quantitative revolution and theoretical geography. The Canadian Geographer 7, 151–162.
Buttimer, A., 1971. Society and Milieu in the French Geographical Tradition. Rand McNally, Chicago.
Buttimer, A., 1993. Geography and the Human Spirit. Johns Hopkins University Press, Baltimore, MD.
Castree, N., 2005. Nature. Routledge, London.
Chisholm, G.G., 1885. Handbook of Commercial Geography. Longman, London (This book went through twenty editions, with the last one published in 1980).
Chisholm, M., 1962. Rural Settlement and Land Use. Hutchinson, London.
Chorley, R.J., Haggett, P. (Eds.), 1965. Frontiers in Geographical Teaching. Methuen, London.
Chorley, R.J., Haggett, P. (Eds.), 1967. Models in Geography. Methuen, London.
Christaller, W., 1966. Central Places in Southern Germany (C.W. Baskin, Trans.). Prentice-Hall, Englewood Cliffs, NJ.
Cliff, A.D., Ord, J.K., 1973. Spatial Autocorrelation. Pion, London.
Cloke, P.J., Crang, P., Goodwin, M. (Eds.), 1999. Introducing Human Geographies, second ed., 2005. Edward Arnold, London.
Cormack, L., 1997. Charting an Empire: Geography at English Universities, 1580–1620. University of Chicago Press, Chicago.
Cosgrove, D.E., 1998. Social Formation and Symbolic Landscape. University of Wisconsin Press, Madison, WI.
Darby, H.C., 2002. In: Williams, M., Clout, H., Coppock, J.T., Prince, H.C. (Eds.), The Relations of History and Geography: Studies in England, France and the United States. University of Exeter Press, Exeter.

Davis, W.M., 1909. In: Johnson, D.W. (Ed.), Geographical Essays. Dover Publications, New York.
Dear, M.J., 1998. The postmodern challenge: reconstructing human geography. Transactions of the Institute of British Geographers NS13, 262–274.
Driver, F., 1998. Geography Militant: Cultures of Exploration in an Age of Empire. Blackwell, Oxford.
Elden, S., 2009. Terror and Territory: The Spatial Extent of Sovereignty. University of Minnesota Press, Minneapolis, MN.
Elden, S., 2013. The Birth of Territory. University of Chicago Press, Chicago.
Elden, S., Mendieta, S. (Eds.), 2011. Reading Kant's Geography. SUNY Press, Albany, NY.
Elwood, S., 2006. Critical issues in participatory GIS: deconstructions, reconstructions and new research directions. Transactions in GIS 10, 693–708.
Freeman, T.W., 1961. One Hundred Years of Geography. G. Duckworth, London.
Gaile, G.L., Willmott, C.J. (Eds.), 1989. Geography in America. Merrill Publishing, Columbus, OH.
Gaile, G.L., Willmott, C.J. (Eds.), 2001. American Geography at the Millennium. Oxford University Press, New York.
Giddens, A., 1977. Studies in Social and Political Theory. Hutchinson, London.
Giddens, A., 1984. The Constitution of Society. Polity Press, Cambridge.
Goodchild, M.F., 1992. Geographical information science. International Journal of Geographical Information Systems 6, 31–45.
Goodchild, M.F., 2009. NeoGeography and the nature of geographic expertise. Journal of Location Based Services 3, 82–96.
Gregory, D., 1978. Ideology, Science and Human Geography. Hutchinson, London.
Gregory, D., 2004. The Colonial Present: Afghanistan, Palestine, Iraq. Blackwell, Oxford.
Gregory, D., Johnston, R.J., Pratt, G., Watts, M.J., Whatmore, S.J. (Eds.), 2009. The Dictionary of Human Geography, fifth ed. Blackwell, Oxford.
Gudgin, G., Taylor, P.J., 1979. Seats, Votes and the Spatial Organization of Elections. Pion, London. (Reprinted 2012, Colchester: ECPR Press).
Hägerstrand, T., 1968. Innovation Diffusion as a Spatial Process (A. Pred, Trans.). University of Chicago Press, Chicago.
Hägerstrand, T., 1982. Diorama, path and project. Tijdschrift voor Economische en Sociale Geografie 73, 323–339.
Haggett, P., 1965. Locational Analysis in Human Geography. Edward Arnold, London. (second edition – with A.D. Cliff and A.E. Frey – 1977).
Haggett, P., 1972. Geography: A Modern Synthesis. Harper and Row, New York. (latest edition, Geography: a Global Synthesis, 2001. Prentice-Hall, Harlow).
Haggett, P., 1990. The Geographers' Art. Blackwell, Oxford.
Hall, P., 1966. Isolated State: An English Edition of Die Isolierte Staat (Carla M. Wartenberg, Trans.). Pergamon Press, Oxford.
Harley, J.B., Woodward, D. (Eds.), 1987. The History of Cartography, vol. 1. University of Chicago Press, Chicago.
Harris, R.J., Sleight, P., Webber, R., 2005. Geodemographics, GIS and Neighbourhood Targeting. John Wiley, London.
Hart, J.F., 1982. The highest form of the geographer's art. Annals of the Association of American Geographers 72, 1–29.
Hartshorne, R., 1939. The Nature of Geography. A Critical Survey of Current Thought in the Light of the Past. Association of American Geographers, Lancaster, PA.
Harvey, D., 1969. Explanation in Geography. Edward Arnold, London.
Harvey, D., 1973. Social Justice and the City. Edward Arnold, London.
Harvey, D., 1982. The Limits to Capital. Blackwell, Oxford.
Harvey, D., 1996. Justice, Nature and the Geography of Difference. Blackwell, Oxford.
Harvey, D., 2000. Spaces of Hope. University of California Press, Berkeley and Los Angeles, CA.
Harvey, D., 2005a. The New Imperialism. Oxford University Press, New York.
Harvey, D., 2005b. The sociological and geographical imaginations. International Journal of Politics, Culture and Society 18, 211–255.
Harvey, D., 2009. Cosmopolitanism and the Geographies of Freedom. Columbia University Press, New York.
Harvey, D., 2012. Rebel Cities: From the Right to the City to the Urban Revolution. Verso, London.
Harvey, F., Wardenga, U., 2006. Richard Hartshorne's adaptation of Alfred Hettner's system of geography. Journal of Historical Geography 32, 422–440.
Humboldt, A. von, 1845–1972. Kosmos. J.C. Gotta Verlag, Stuttgart.
Jackson, P.A., Smith, S.J., 1984. Exploring Social Geography. George Allen & Unwin, London.
Johnston, R.J., 2006. The politics of changing human geography's agenda: textbooks and the representation of increasing diversity. Transactions of the Institute of British Geographers NS31, 286–303.
Johnston, R.J., Claval, P. (Eds.), 1984. Geography Since the Second World War: An International Survey. Croom Helm, London.
Johnston, R.J., Sidaway, J.D., 2004. The trans-Atlantic connection: 'Anglo-American' geography reconsidered. GeoJournal 59, 15–22.
Johnston, R.J., Sidaway, J.D., 2015. Geography and Geographers: Anglo-American Human Geography Since 1945, seventh ed. Taylor and Francis, London.
Johnston, R.J., Williams, M. (Eds.), 2003. A Century of British Geography. Oxford University Press for the British Academy, Oxford.
Jones, K., 1991. Specifying and estimating multi-level models for geographical research. Transactions of the Institute of British Geographers NS16, 148–160.
Kearns, G., 2009. Geography and Empire: The Legacy of Halford Mackinder. Oxford University Press, Oxford and New York.
Keighren, I.M., 2010. Bringing Geography to Book: Ellen Semple and the Reception of Geographical Knowledge. I. B. Tauris, London.
Kitchin, R., Thrift, N.J. (Eds.), 2009. International Encyclopaedia of Human Geography. Elsevier, Oxford.
Kuhn, T.S., 1970. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Landstrom, C., Whatmore, S.J., Lane, S.N., Odoni, N.A., Ward, N., Bradley, S., 2011. Coproduced flood risk knowledge: redistributing expertise in critical 'participatory modelling'. Environment and Planning A 43, 1617–1633.
Linke, M., 1981. Carl Ritter 1779–1859. Geographers: Biobibliographical Studies 6, 99–108.
Livingstone, D.N., 1992. The Geographical Tradition: Episodes in the History of a Contested Enterprise. Blackwell, Oxford.
Livingstone, D.N., Withers, C.W.J., 2011. Geographies of Nineteenth Century Science. University of Chicago Press, Chicago.
Longley, P.A., Brooks, S.M., McDonnell, R.A., McMillan, B., 1998. Geocomputation: A Primer. John Wiley, New York.
Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W., 2005. Geographic Information Systems and Science, second ed. John Wiley, New York.
Mackinder, H.J., 1887. On the scope and method of geography. The Geographical Journal 9, 141–174.
Massey, D., Allen, J., Sarre, P. (Eds.), 1999. Human Geography Today. Polity Press, Cambridge.
May, J.A., 1970. Kant's Concept of Geography and its Relation to Recent Geographical Thought. University of Toronto, Department of Geography, Research Publication 4, Toronto.
Mayhew, R.J., 2010. Printing posterity: editing Varenius and the construction of geography's history. In: Ogborn, M., Withers, C.W.J. (Eds.), Geographies of the Book. Ashgate, Farnham, pp. 157–187.
Mead, W.R., 1963. The adoption of other lands. Geography 48, 241–254.
Morin, K.M., 2011. Civic Discipline: Geography in America 1860–1890. Ashgate, Farnham.
NRC, 1997. Rethinking Geography: New Relevance for Science and Society. National Research Council, Washington, DC.
NRC, 2010. Understanding the Changing Planet: Strategic Directions for the Geographical Sciences. National Research Council, Washington, DC.
Olsson, G., 1965. Distance and Interaction: A Review and Bibliography. Regional Science Research Institute, Philadelphia, PA.
Openshaw, S., 1984. The Modifiable Areal Unit Problem. CATMOG 38. GeoBooks, Norwich.
Pain, R., 2003. Social geography: on action-orientated research. Progress in Human Geography 27, 677–685.
Pappenberger, F., Stephens, E., Thielen, J., Salamon, P., Demeritt, D., van Andel, S.J., Wetterhall, F., Alfieri, L., 2012. Visualizing probabilistic flood forecast information: expert preferences and perceptions of best practice in uncertain communication. Hydrological Processes 27, 132–146.
Peet, R., 1998. Modern Geographic Thought. Blackwell, Oxford.
Pickles, J. (Ed.), 1995. Ground Truth: The Social Implications of Geographical Information Systems. The Guilford Press, New York.
Pile, S., 1996. The Body and the City. Routledge, London.
Pile, S., Keith, M. (Eds.), 1997. Geographies of Resistance. Routledge, London.
Ratzel, F., 1882–1891. Anthropogeographie. J. Engelhorn, Stuttgart.
Ritter, K., 1817–1859. Die Erdkunde. G. Reiner, Berlin.
Rose, G., 1993. Feminism and Geography. Polity Press, Cambridge.
Sack, R.D., 1981. Conceptions of Space in Social Thought. Macmillan, London.
Sack, R.D., 1986. Human Territoriality: Its Theory and History. Cambridge University Press, Cambridge.
Said, E.W., 1978. Orientalism. Pantheon Books, New York.
Sampson, R.J., 2011. Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago Press, Chicago.
Sauer, C.O., 1952. Agricultural Origins and Dispersals. American Geographical Society, New York.

Sayer, A., 1984. Method in Social Science: A Realist Approach. Hutchinson, London. (second edition, 1992, Routledge, London).
Schaefer, F.K., 1952. Exceptionalism in geography: a methodological examination. Annals of the Association of American Geographers 42, 226–249.
Schulten, S., 2001. The Geographical Imagination in America, 1880–1950. University of Chicago Press, Chicago.
Semple, E.C., 1911. Influence of Geographic Environment. Henry Holt, New York.
Sheppard, E.S., 1995. Dissenting from spatial analysis. Urban Geography 16, 283–303.
Smith, D.M., 1981. Industrial Location: An Economic Geographical Analysis, second ed. John Wiley, New York.
Smith, D.M., 2000. Moral Geographies: Ethics in a World of Difference. University of Edinburgh Press, Edinburgh.
Smith, N., 2003. American Empire: Roosevelt's Geographer and the Prelude to Globalization. University of California Press, Berkeley, CA.
Soja, E.W., 1989. Postmodern Geographies. Verso, London.
Springer, S., 2012. Anarchism: what geography still ought to be. Antipode 44, 1605–1624.
Stoddart, D.R., 1986. On Geography, and Its History. Blackwell, Oxford.
Sui, D.Z., et al., 2004. Forum: on Tobler's first law of geography. Annals of the Association of American Geographers 94, 269–310.
Taylor, P.J., 1999. Places, spaces and Macy's: place-space tensions in the political geographies of modernities. Progress in Human Geography 23, 7–26.
Thrift, N.J., 2007. Non-representational Theory: Space, Politics, Affect. Routledge, London.
Vidal de la Blache, P., 1903. Tableau de la Géographie de la France. Hachette, Paris.
Whatmore, S.J., 2002. Hybrid Geographies: Nature Cultures Spaces. SAGE Books, London.
Women and Geography Study Group, 1984. Geography and Gender: An Introduction to Feminist Geography. Hutchinson, London.
Women and Geography Study Group, 1997. Feminists Geographies: Explorations in Diversity and Difference. Longman, Harlow.
Wooldridge, S.W., 1956. The Geographer as Scientist. Thomas Nelson, London.

Geopolitics

Klaus Dodds, Royal Holloway University of London, Egham, Surrey, UK
Chih Yuan Woon, National University of Singapore, Kent Ridge, Singapore

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by S.B Cohen, volume 9, pp. 6199–6205, © 2001, Elsevier Ltd.

Abstract

Geopolitics is a term encompassing a diversity of thought and practice that extends far beyond the academy. From its initial preoccupation with imperialist struggles closely linked to the survival of states and empires, geopolitics continues to attract attention in the policy realm with the unfolding of great power rivalries and global hegemonic agendas. The infusion of 'critical' perspectives into the study of geopolitics has challenged the existence of objective scriptings of political space, highlighting geopolitics as a power-laden pursuit instead. As critical geopolitics diverge into 'newer' interests such as investigations of affect/emotions and assemblage thinking, there remains the need to be vigilant of the geostrategic knowledges that are invoked in the service of warfare and their multifarious impacts on society and its people.

Introduction

Geopolitics is a term that continues to attract a range of reactions including approval and opprobrium, although not always in equal measure. From its formal inception in the 1890s, the term 'geopolitics' has mushroomed to embrace a wide variety of thought and practice, and as a consequence a diverse range of stakeholders that extends far beyond the academy. While recognizing diversity is important, we can identify nonetheless a series of themes and motifs that continue to prevail. One area worth highlighting right at the start is a concern for how the physical environments of the earth shape the operationalization of power. In other words, geopolitical writers continue to be fascinated by the power of the earth, in the form of mountains, rivers, seas, deserts, and the like to shape political geographical relations. One contemporary example would be the Arctic region and the arguments that are raging about what continued sea ice thinning will do to the regional dynamics of this part of the world. The consensus of opinion, among the geopolitical community, seems to be that the Arctic is going to become more accessible and more exploitable. Where the community divides is whether this will encourage more consensual or conflict-based behavior.

The second area of interest is in state territory and state power (*see* State, Geography of). Ever since the earliest European commentators in the late nineteenth century embarked on explicitly geopolitical exchanges, interest loomed large in the role of populations, resources, and territorial advantage. Such an interest provoked further reflection either on how states/governments might better secure their national territories and/or resist other states who were operating in an international political world imagined to be uncooperative and even anarchical. Armed with ideas associated with social Darwinism, the state was frequently conceived as a living organism needing protection from predatory others eager to take advantage of additional territory and resources. Such a dystopian view of the world order played its part in ensuring that geopolitical ideas and practices were later to be condemned as intellectually poisonous, as critics claimed that the German Nazis and other fascist governments in Italy and Japan drew intellectual and practical inspiration from geopolitical doctrines.

The third theme worth stressing is alternative geopolitical ideas and practices. Geopolitics remains a contested enterprise. While it has long appealed to the powerful, there have always been counterprevailing tendencies (resistance, subversion, dissent and alternative imaginations, and practices) within the academy and beyond. Such struggles to resist dominant geopolitical imaginations find expression, therefore, in a diverse range of sites and spaces involving multiple agents such as anti-imperialists, governments in the global South, opposition movements, journalists, activists, and dissenters throughout the world. One example of someone who actively articulates a counter-geopolitical imagination is the long-serving president of Zimbabwe, Robert Mugabe. Since his election to political office in 1980, Mugabe has argued that Zimbabwe has to resist the dominant geopolitical machinations directed against his country by the former imperial power, the UK and other Western allies such as the United States and European states. As he claimed in 2002 at the Earth Summit in South Africa following criticism of his political rule and associated human rights abuses, "We have fought for our land, we have fought for our sovereignty, small as we are we have won our independence and we are prepared to shed our blood …. So, Blair keep your England, and let me keep my Zimbabwe." Mugabe might well think that he is practicing a form of critical geopolitics, one that highlights the discursive and material inequalities in contemporary global geopolitics alongside inherited imperial legacies.

Bearing the above in mind, this short article has two substantial sections. The first addresses some of the central tenets associated with classical geopolitics and the second section touches upon the literature associated with critical geopolitics. We conclude briefly with an assessment of the outstanding areas of research.

Classical Geopolitics

The earliest classical geopolitical writings were informed by imperial preoccupations and social Darwinist anxieties about the survival of states and empires. Writing in the latter part of the nineteenth century, the German writer Friedrich Ratzel

(1844–1904) expressed some of the dominant trends in classical geopolitics including racial and environmental determinism. Distinguishing between settled and nomadic communities, Ratzel argued that the settled Aryan races in particular were vulnerable to the marauding and thus hypermobile Ural–Altaic races. As a consequence of this perceived vulnerability, the Aryan races were the earliest to develop statelike structures designed to organize defense against mobile and, in all probability, threatening others. Ratzel contended that contemporary Germany should strive to secure additional land and resources so that it was better able to secure the survival of the nation-state in the face of Eastern races and their traditions of mobile existence. The term *Lebensraum* (living space) is particularly associated with the writings of Ratzel as a consequence of his interest in the interrelationship between environment, state, and culture.

While Ratzel's writings were infused with racial and environmental indices, the British writer Halford Mackinder (1861–1947) highlighted another aspect of classical geopolitics, namely a concern for grand strategy. Mackinder, a reader in geography at Oxford University and director of the London School of Economics, was primarily worried about Britain and the British Empire (Blouet, 1987; Kearns, 2009). Mindful of gathering imperial competition, Mackinder warned that traditional sea powers such as Britain were under threat from new land-based powers that might, with the help of new transport technologies such as the railway, be able to mobilize their populations and resources in a decisive manner. Intrigued by the historic significance of migrant empires such as the Mongols, Mackinder divined a future possibility based on new great powers (such as the latter day Soviet Union) using what he termed the 'heartland' to project power over the European continent. Vast quantities of coal, oil, gas, and other minerals, transported by railways, would quite literally empower those who controlled the heartland. In his famous epithet, Mackinder warned his readers that "who rules East Europe commands the Heartland. Who rules the Heartland commands the World-Island. Who rules the World-Island commands the World" (Mackinder, 1919).

These brief sketches of two classical geopolitical writers highlight the dominant strands of classical geopolitics – on the one hand, social Darwinism and environmental determinism and, on the other hand, imperial rivalries and great-power projection. In the case of the former, the earliest geopolitical writers were overwhelmingly informed by racial and environmental determinism. Most contributors were eager to warn their political masters and the wider public about the challenges facing their societies from competitor races and states, both past and present. Geopolitics was, and for many authors still is, the study of statecraft and the divination of patterns of global politics. If geopolitics has an intellectual value, it lies in a capacity to uncover the challenges facing the state and empire, and a willingness to use force if necessary to protect vital interests.

The controversial reputation inherited by geopolitics owes much to a particular moment in the interwar period when the subject became linked with more widespread fascist and authoritarian strands of thought in Germany, Italy, and Japan. With its focus on territorial and resource-related factors, geopolitical thought attracted the interest of writers such as the German military officer and professor of geography at the University of Munich, Karl Haushofer. In 1924, he founded the *Zeitschrift fur Geopolitik*, a journal that was to explore how Germany might recover from the humiliating loss of the World War I and the sorts of geographical strategies need to stimulate revival. Geopolitics, according to Haushofer, should be an accomplice of the state and encourage an expansion of Germany, possibly leading to leadership of the Euro-Asian landmass in alliance with Russia and Japan. Combining the global strategizing of Mackinder with social Darwinist insights including the state as a living organism, a geopolitical perspective armed with maps and statistics was believed to be an indispensible part of the state and its intellectual armoury. As with Fascist Italy, this academic interest was supplemented by popularizing trends with a new generation of school children being instructed on the spatial extent of a Greater Germany and Italy (Atkinson, 1995). Under this worldview, the state needed to expand into new territory and acquire new resources in order to restore itself to health.

For American-based geographers, including refugees from Nazi Germany, geopolitics was at best a 'pseudoscience' and at worst a 'Nazi science.' Either way it stood accused of acting as a willing accomplice to the worst excesses of territorial domination and spatial expansionism. With the onset of the World War II, this jaundiced view of geopolitics hardened. American and European geographers such as Isaiah Bowman, Derwent Whittlesey, George Kiss, and Jean Gottmann warned that geographers should steer clear of the subject matter and concentrate their energies on developing a scientific political geography (as noted in Dodds, 2010). While this alleged association between Hitler and geopolitical scheming was perhaps not as straightforward as implied by the critics, it did prevent the Roosevelt administration from commissioning a special study of German geopolitics in order to consider its resonance and influence among Hitler's cabal. Senior geographers such as Karl Haushofer, while hardly bystanders, did not share Hitler's preoccupation with race and hatred of Jewry. They were, however, eager to put forward proposals for Germany to restore its proper place in the world and supportive of the idea that the country needed to expand into new territories. Popular magazines such as *Reader's Digest* played their part in warning international readers that the Nazis created a shadowy Institute of Geopolitics in Munich and that there were a '1000 scientists behind Hitler' using charts, maps, statistics, and plans to facilitate further expansion and conquest.

Accused of being an accomplice to Nazism, geopolitics became an academic field best avoided for many geographers in the USA and Europe. The end result was twofold. First, there was a general reluctance to use the term 'geopolitics' explicitly and, second, attention turned to seeking a new social scientific language based on modeling and testing that would replace any notion that the state was a living organism. For much of the post-1945 period, explicit references to geopolitics were limited in number and scope even if well-known figures such as George Kennan were to talk and write about continental powers, the Euro-Asian landmass and resource potential of regions. As Richard Hartshorne reminded his readers: "We may have produced no atom bombs in political geography, but the field

is nonetheless strewn with dynamite – it is no place for sophomores to play with matches" (cited in Kearns, 2011: 614).

Henry Kissinger is famously credited with making geopolitics respectable again in US policy-making and academic circles (1979). In the early 1970s, geopolitics became a short hand for highlighting great power rivalries and associated regional dimensions, especially in the Middle East and Southeast Asia. As National Security Advisor and Secretary of State in the Nixon and Ford administrations, Kissinger was well placed to survey the global political scene and ruminate on the geopolitical consequences and implications, more often than not involving military force and assertion. As Leslie Hepple reminds us, Kissinger's geopolitics was often fuzzy and vague, even if it had as its basis an interest in superpower rivalries and the global stage (1986). Other contemporaries such as President Carter's National Security Advisor, Zbigniew Brzezinski, also used geopolitical language to promote his view that control of the 'heartland' (a point of view articulated by Mackinder some 70 years earlier), was critical to the future patterns of global politics. As Brzezinski noted, "whoever controls Eurasia dominates the globe. If the Soviet Union captures the peripheries of this landmass … it would not only win control of vast human, economic and military resources but also gain access to the geostrategic approaches to the Western Hemisphere – the Atlantic and Pacific" (1986: 22). For the USA to retain its global prominence, it would need, by force if necessary, to ensure that vital resource supplies and territorial access were secure. Moreover, as US covert support for anti-Soviet forces in Afghanistan in the 1980s suggests, it would require investment in opportunities that might disrupt any attempt by the Soviet Union to project further its domination of the Eurasian landmass.

The net result of this resurrection of interest in geopolitics in the 1970s was to remind us why geopolitics continues to inspire a great deal of interest and, as noted earlier, distrust. If there was a revival – acknowledging the long-established interest in the field in the former Soviet Union and Latin America, where it has been closely tied to forms of military geography (*see* Military Geography), then it was based on a predilection for global perspectives; skepticism about international diplomacy and law; an understanding of power relations as a zero-sum game; a belief in the importance of resources such as oil, gas, coal, and minerals; a stark contrast between land- and sea-based powers; and finally a willingness to urge intervention (or sometimes nonintervention) where national security interests were at stake (for a critique, see Kearns, 2011).

Critical Geopolitics

The renewed interest in geopolitics in the international political/policy arena garnered the attention of many scholars, not least Les Hepple who published an influential article in 1986 titled, "The Revival of Geopolitics." In his paper, Les Hepple was not only concerned with the "'social life' of geopolitical ideas: how they 'travel' between places, how they were adapted, contextualised and repoliticised through specific tactical deployments or denouncements" (Megoran, 2008: 403). More importantly, he argued that this exercise of tracing the adaptations and mutations of geopolitical discourses will form the necessary premise for challenging dangerously misleading geopolitical doctrines and policies in discrete contexts. Les Hepple's provocative call for deconstructive geopolitical engagements through responsible scholarship has characterized the field of critical geopolitics, conceived of by O'Tuathail (1996: 256) as one of the "cultures of resistance to Geography as imperial truth, state-capitalized knowledge and military weapon." Indeed as Atkinson (2001: 426) has noted, Hepple's insights foreshadowed concerns that would later occupy critical geopoliticians and presaged the inquiries of critical geopolitics, conveying "a sense of the arbitrary and diverse use of the label, and an impression of the plural contexts wherein geopolitics surfaced." As a broad set of writings that first emerged in the early 1990s, critical geopolitics came under the radar of the geographical discipline through the seminal works of O'Tuathail and Agnew (1992), O'Tuathail (1996) and Dalby (1990, 1991). These pioneering research quickly proliferated into an extensive array of publications that helped to shape the intellectual boundaries and trajectories of the now established field of critical geopolitics. Influenced by post-structuralist philosophies particularly emanating from Michel Foucault and Jacques Derrida's works, the initial impetus of critical geopolitics was to generate new ways of thinking about how global political space is being constructed, represented, and practiced. In Dalby's (2010: 281) words, there were "concerns about disrupting the hegemonic practices of statecraft, challenging the taken for granted specifications of the world in various mappings by elites and academics." What this signifies is that there is a more explicit focus on the dual issues of agency and power: who is sculpting/representing the terrains of geopolitics and for what purposes and agendas. In doing so, there can be more nuanced expositions of the deployment of geopolitical discourses in the service of exclusionary and violent state power.

Indeed, there have been ongoing efforts to recover "textuality within practices which are represented as objective or practical, as 'beyond the text'" (O'Tuathail in Jones and Sage, 2010: 316). In such a conception, geopolitics is not something 'out there' that can be objectively observed and analyzed; rather it is actively played out through the complex web of language and (con)text. As such, geopolitical discourse in global politics is understood to be the outcome of perpetual 'geo-graphing': the production and dissemination of strategic maps and texts (Dodds and Sidaway, 1994). Aligning to this position, critical geopolitical writers have demonstrated that the ability of political elites to depict and represent places in certain ways should be an important component of investigating international politics. This line of inquiry emerged primarily as a challenge to American military power during the Cold War era: critical geopolitics were keen to expose how geographical knowledges were invoked to bequeath certain characteristics to specific places so as to make them amenable to US's intercession and mode of governance (e.g., the Antarctica) (Dodds, 1997). With the end of the Cold War, critical geopolitics began investigating the shifting of global security from the dangers posed by Communism/the Soviet Union to matters of development and insecurity at the fringes of global political economy. New

insecurities came in the form of weak/failed states and wild zones, as portrayed by many policy-making bureaucracies (O'Tuathail and Luke, 1994). These caricatures continue to find resonance in a world dominated by conspiracies of terror in the aftermath of the 11 September 2001 attacks on New York and Washington. Mapping dangerous places and formulating effective interventions once again captured the imaginations of military cartographers and geostrategists (Barnett, 2004; Kaplan, 2009). Critical geopolitics has proved particularly instructive in reminding us that the planet's transformation into a battlefield as a consequence of the 'global war on terror' cannot be divorced from the ways in which threats are imagined and spatialized so as to legitimate (America's) political power and hegemony (Gregory, 2004).

While critical geopolitics involves fundamentally unpacking the tropes and epistemologies of dominant geographs and scriptings of political space, its proponents are quick to note that there are multiple pathways to attain such a goal. Beyond state-centric readings of world politics, critical geopoliticans are keen to illuminate the messy and mundane practices that constitute the modern interstate system. In other words, geopolitics is inescapably plural and this has opened up the range of sites/texts and practices where 'geopolitics' is seen to take place (O'Tuathail in Jones and Sage, 2010). Examples include works on anti-geopolitics that study nonstate actors such as social movements and transnational organizations (Routledge, 1996, 2002); and on popular geopolitical representations contained within magazines (Sharp, 2000), cartoons (Dodds, 2007), films (Dodds, 2006; Dalby, 2008), Internet platforms (Mamadouh, 2003; Woon, 2011a), and digital video games (Power, 2007) along with a burgeoning focus on geopolitics, visuality, and visual culture (Hughes, 2007; McDonald et al., 2010). Such a diversification of geopolitics has also led to heightened interest in different traditions and historical contexts in which geopolitics has been written and interpreted. Indeed, following Hepple's initial exposition of the Latin American and Rumanian strands of geopolitical thought, critical geopolitics have since delved into understanding a growing range of formal geopolitical traditions (Dodds and Atkinson, 2000). From Italy (Atkinson, 1995) to France (Fall, 2007) to Spain and Portugal (Sidaway, 2000; Sidaway and Power, 2005), reflections abound on the historicity of geopolitics, drawing upon the debates over writing more contextual accounts of the histories of geographical knowledge (Livingstone, 1992).

Despite its vibrant development over time, critical geopolitics as a field of study has been criticized for being overly normative and not living up to its political potential. Dalby (1996) notes that the 'critical' in critical geopolitics usually refers to the problematization of discourse rather than the presence of a worked out, alternative political project (see also Dowler and Sharp, 2001). Hyndman (in Jones and Sage, 2010: 317) likewise argues that "in its quest to destabilize the normative, critical geopolitics rarely engages in transformative and embodied ways of seeing and knowing." Kelly (2006: 43) on the other hand, lambastes critical geopolitics for offering "neither a clear characterization of a better society nor a specific road map for attaining such an improvement." In lieu of such critiques, there has been the rise of numerous intellectual initiatives that seek to steer the subdiscipline into new directions for critical engagements. For instance, Jennifer Hyndman (2004, 2010) has, for some time now, made the case for a feminist geopolitics and an embodied politics of shared fate so as to realize the possibility of nonviolence in facilitating interactions in the global arena. Similarly, Mercille (2008) proposes a 'radical geopolitics' to reinvigorate the political spirit of geopolitics, suggesting that political economy might be the way forward in terms of identifying some of the structural underlying practical geopolitical decision-making and global inequalities, helping us to explore the 'why' as well as 'how' of political/economic policy decisions. Additionally, Koopman (2011) evokes the term 'alter-geopolitics' to foreground the connections between critique and political activism. Specifically, she casts attention on the grassroots movements that build international relations of solidarity for active peacemaking, arguing that academics might simultaneously take on the activist persona to translate their geopolitical theories into positive, material outcomes.

Although diverging in their exact focus, these aforementioned geopolitical agendas are united in their insistence to ground geopolitics in particular places. This links to Thrift's oft-cited claim that critical geopolitics' apparent mesmerized attention to texts and images has led it to neglect the 'little things.' Thrift calls for more sensitive inquiry into the workings of geopower, by attending to objects and issues related to embodiment. This simultaneously entails critical geopolitics to expand its methodological repertoire (beyond discourse analysis) to experiment with 'traditional' modes of doing geography: fieldwork, ethnography, and interviews (Megoran, 2006; O'Tuathail, 1998; Woon, 2013). These various dimensions and challenges have been undertaken by a plethora of work in recent years. While it is impossible for us to cover such research in their entirety, we will zoom in on what we think are two most exciting developments in critical geopolitics.

First, matters of percept, affect, and emotion, as well as the most ordinary ('precognitive') forms of sociality have entered into the research domain of critical geopolitics (Power and Campbell, 2010). A few examples will suffice to illustrate this point: O'Tuathail's (2003) work on how affect is implicated in the US's invasion of Iraq, Carter and McCormack's study on film and the affective logic of intervention and Pain's (2009, 2010) investigations of the emotional geographies of fear that underpin the 'war on terror.' Most recently, these textual themes have been complemented by looking at their localized politics of reception and interpretation. The current discussion of audience reception, fandom and how readers and viewers consume movies and other cultural products extends analysis of critical geopolitics in another useful direction (Dodds, 2006; Dittmer and Dodds, 2008). And so do research focusing on the ways in which the emotive infrastructures of warfare are responded to by affected communities on the ground (some culminating in hopeful outcomes whereby emotions are stimulating actions and shaping the practices and progress of radical praxis at different geographical scales) (Pain and Smith, 2008; Woon, 2011b).

Second, the infusion of assemblage thinking into critical geopolitics has also offered stimulating prospects for the field. According to Anderson et al. (2012: 173), the concept of assemblage provides something significant and unique

insofar as it allows analyses to "attend to how often disparate activities become entangled with one another, but nonetheless have potential agency beyond those interactions which may later become parts in other assemblages." Adopting such a relational approach, Anderson (2010), for example, illuminates how a form of anticipatory actions such as preemption in US counterterrorism policy is constituted through the redeployment of elements from other ways of acting over futures as well as the intensification of dormant tendencies in linked areas of public policy. Preemption emerges as a coherent and distinct way of reordering space, even as it is always involved in a series of partial connections with other ways of acting over threats and hazards. Alternatively, Muller (2012) casts scrutiny on organizations as one underresearched actor in critical geopolitics. Conceptualizing organizations as assemblages that are constituted through complex social material networks, Muller exhorts for organizations to be given more attention in the subdiscipline due to their complicity in the continuous processes of geopolitical ordering. Hence, through these writings, we can see that assemblage offers a sustained account of the multiple ways in which geopolitical orders endure across differences and amid transformations, in addition to a sensitivity to how such orders are changed and reworked.

Concluding Remarks

We are not trying to posit a strict dichotomy between classical and critical geopolitics nor are we suggesting a linear progression from the former to the latter. As O'Tuathail (in Jones and Sage, 2010: 316) reminds us, critical geopolitics is not simply a more advanced or superior version of its classical variant. Rather "in exposing geopolitics as a convenient fiction, critical geopolitics reveals itself as a similarly convenient fiction of opposition. It is merely the starting point for a different form of geopolitics …." After all, it must be remembered that the fundamental impetus of critical geopolitics is precisely to question the power/knowledge imbued in classical geopolitical discourses. In other words, the development of critical geopolitics cannot be adequately analyzed or fully appreciated without reference to its classical counterpart.

Indeed, the continued need for critical geopolitics to deconstruct and expose how geostrategic knowledges are used to legitimate warfare has been demonstrated by a number of scholars. By underscoring the geographical tropes of warfare embedded within Robert Kaplan's (2009) work, Dalby (2010), for instance, concurs with Fraser MacDonald's (in Jones and Sage) assessment that the persistence of classical geopolitical reasoning makes sustained critique necessary. Both MacDonald and Dalby are wary that critical geopolitics' forays into a host of other research avenues (the 'little things' such as affect/emotions, lived experiences of people, etc.) will dilute its original purpose concerning the writing of global political space (i.e., the 'big things'). But if we accept O'Tuathail's (in Jones and Sage, 2010: 316) claim that critical geopolitics is "no more than a general gathering place for various critiques of the multiple geopolitical discourses and practices that characterize modernity," the eclecticism of the field can only be seen as a stimulating and necessary intellectual and political project to engage with some of the most pressing issues of our times (terror, diseases, migration, human rights, and environmental risk) as part of a wider set of critical approaches to political geography (*see* Political Geography). What perhaps needs to be further pursued then, are critical studies on the ways in which the 'big' and 'little' things of geopolitics are intimately connected. This necessitates a multiscalar approach to understanding geopolitics and its multifarious impacts and relevance on society and its people.

See also: Geography: Overview; Military Geography; Political Geography; State, Geography of; Territoriality: Geographical; War and Terrorism, Geography of.

Bibliography

Anderson, B., 2010. Preemption, precaution, preparedness: anticipation action and future geographies. Progress in Human Geography 34, 777–798.

Anderson, B., Kearnes, M., McFarlane, C., Swanton, D., 2012. On assemblages and geography. Dialogues in Human Geography 2, 171–189.

Atkinson, D., 1995. Geopolitics, cartography and geographical knowledge: envisioning Africa from Fascist Italy. In: Bell, M., Butler, R.A., Heffernan, M. (Eds.), Geographical Knowledge and Imperial Power 1820–1940. Manchester University Press, Manchester, pp. 290–332.

Atkinson, D., 2001. Classics in human geography revisited: Hepple L.W. 1986: the revival of geopolitics. Progress in Human Geography 25, 423–430.

Barnett, T., 2004. The Pentagon's New Map: War and Peace in the Twenty-First Century. G.P. Putnam's Sons, New York.

Blouet, B.W., 1987. Halford Mackinder, first ed. Texas A & M Press, College Station, TX.

Brzezinski, Z., 1986. Game Plan, first ed. Atlantic Monthly Press, Boston.

Brzezinski, Z., 1997. The Grand Chessboard, first ed. Basic Books, New York.

Bullitt, W., 1947. A report to the American people on China. Life 23, 35–37.

Bush, G., 1991. The Hard Work of Freedom, Public Papers of the Presidents of the United States. US Government Printing Office, Washington, DC.

Carter, S., McCormack, D., 2006. Film, geopolitics and the affective logics of intervention. Political Geography 25, 228–245.

Chrone, G.R., 1969. Background to Political Geography. Pittman Publishers, London.

Cohen, S.B., 1973. Geography and Politics in a World Divided, second ed. Oxford University Press, New York.

Cohen, S.B., 1982. A new map of global geopolitical equilibrium: a developmental approach. Political Geography Quarterly 1, 223–242.

Dalby, S., 1990. American security discourse: the persistence of geopolitics. Political Geography Quarterly 9, 171–188.

Dalby, S., 1991. Critical geopolitics: discourse, difference and dissent. Environment and Planning D: Society and Space 9, 261–283.

Dalby, S., 1996. Writing critical geopolitics: Campbell, O'Tuathail, Reynolds and dissident skepticism. Political Geography 15, 655–660.

Dalby, S., 2008. Warrior geopolitics: black hawk down and the kingdom of heaven. Political Geography 27, 439–455.

Dalby, S., 2010. Recontextualising violence, power and nature: the next twenty years of critical geopolitics. Political Geography 29, 280–288.

Dittmer, J., Dodds, K., 2008. Popular geopolitics past and future: fandom, identities and audiences. Geopolitics 13, 437–457.

Dodds, K., 1997. Geopolitics in Antarctica: Views from Southern Oceanic Rim. John Wiley, Chichester.

Dodds, K., 2006. Audience dispositions and popular geopolitics: James Bond and the Internet Movie Database IMDb. Transactions of the Institute of British Geographers 31, 116–130.

Dodds, K., 2007. Steve Bell's eye: cartoons, geopolitics and the visualization of the 'war on terror.' Security Dialogue 38, 157–177.

Dodds, K., 2010. Classical geopolitics revisited. In: Denemark, R. (Ed.), The International Studies of Encyclopedia. Wiley Blackwell, Oxford, pp. 302–322.

Dodds, K., Atkinson, D. (Eds.), 2000. Geopolitical Traditions: Critical Histories of a Century of Political Thought. Routledge, London.
Dodds, K., Sidaway, J., 1994. Locating critical geopolitics. Environment and Planning D: Society and Space 12, 515–524.
Dowler, L., Sharp, J., 2001. A feminist geopolitics? Space and Polity 5, 165–176.
Fairgrieve, J., 1915. Geography of World Power. University of London Press, London.
Fall, J., 2007. Lost geographers: power games and the circulation of ideas within Francophone political geographies. Progress in Human Geography 31, 195–216.
Gregory, D., 2004. The Colonial Present: Afghanistan, Palestine, Iraq. Blackwell, Malden, MA.
Gyorgy, A., 1944. Geopolitics – The New German Science. University of California Press, Berkeley and Los Angeles, CA.
Haushofer, K., Obst, E., Lautensash, H., Maull, O., 1928. Foundations of Geopolitik. Kurt Vowinckel Verlag, Berlin.
Hepple, L., 1986. The revival of geopolitics. Political Geography Quarterly 5, S21–S36.
Hughes, R., 2007. Through the looking blast: geopolitics and visual culture. Geography Compass 1, 976–994.
Huntington, S.P., 1993. The clash of civilizations? Foreign Affairs 72, 22–49.
Hyndman, J., 2004. Mind the gap: bridging feminist and political geography through geopolitics. Political Geography 23, 307–322.
Hyndman, J., 2010. The question of 'the political' in critical geopolitics: querying the 'child soldier' in the 'war on terror.' Political Geography 29, 247–255.
Ingram, A., Dodds, K., 2011. Counterterror culture. Environment and Planning D: Society and Space 29, 89–97.
Jones, L., Sage, D., 2010. New directions in critical geopolitics: an introduction. GeoJournal 75 (4), 315–325.
Kaplan, R., 1994. Coming anarchy. Atlantic Monthly 273, 44–46.
Kaplan, R., May/June 2009. The revenge of geography. Foreign Policy, 96–105.
Kearns, G., 2009. Geopolitics and Empire. Oxford University Press, Oxford.
Kearns, G., 2011. Geopolitics. In: Agnew, J., Livingstone, D. (Eds.), The Sage Handbook of Geographical Knowledge. Sage, London, pp. 610–622.
Kelly, P., 2006. A critique of critical geopolitics. Geopolitics 11, 24–53.
Kissinger, H., 1979. The White House Years, first ed. Little Brown, Boston, MA.
Kjellén, R., 1916. The State as a Form of Life. Hugo Gebers Forlag, Stockholm.
Koopman, S., 2011. Alter-geopolitics: other securities are happening. Geoforum 42, 274–284.
Livingstone, D., 1992. The Geographical Tradition. Blackwell, Oxford.
Mackinder, H., 1904. The geographical pivot of history. Geographical Journal 23, 421–437.
Mackinder, H., 1919. Democratic Ideals and Reality. Constable, London.
Mackinder, H., 1924. The Nations of the Modern World. George Philip, London.
Mackinder, H., 1943. The round world and the winning of the peace. Foreign Affairs 21, 595–605.
Mahan, A.T., 1890. The Influence of Sea Power Upon History. Sampson Low Marston, London.
Mahan, A.T., 1900. The Problem of Asia and Its Effect Upon International Policies. Little Brown, Boston, MA.
Mamadouh, V., 2003. 11 September and popular geopolitics: a study of websites run for and by Dutch Morocans. Geopolitics 8, 191–216.
McDonald, F., Hughes, R., Dodds, K. (Eds.), 2010. Observant States: Geopolitics and Visual Culture. I B Tauris, London.
Megoran, N., 2006. For ethnography in political geography: experiencing and re-imagining Ferghana Valley boundary closures. Political Geography 25, 622–640.
Megoran, N., 2008. The task and responsibility of geopolitical analysis. Geopolitics 13, 403–407.
Mercille, J., 2008. The radical geopolitics of US foreign policy: geopolitical and geoeconomic logics of power. Political Geography 27, 570–586.
Modelski, G., 1978. The long cycle of global politics and the nation state. Comparative Studies of Society and History 20, 214–235.
Muller, M., 2012. Opening the black box of the organization: social-material practices of geopolitical ordering. Political Geography 31, 379–388.
O'Loughlin, J. (Ed.), 1994. Dictionary of Geopolitics. Greenwood Press, Westport, CT.
O'Sullivan, P., 1982. Antidomino. Political Geography Quarterly 1, 57–64.
O'Tuathail, G., 1996. Critical Geopolitics: The Politics of Writing Global Space. University of Minnesota Press, Minneapolis, MN.
O'Tuathail, G., 1998. Thinking critically about geopolitics. In: O'Tuathail, G., Dalby, S., Routledge, P. (Eds.), The Geopolitics Reader. Routledge, London, pp. 1–12.
O'Tuathail, G., 2003. Just out looking for a fight: American affect and the invasion of Iraq. Antipode 35, 856–870.
O'Tuathail, G., Agnew, J., 1992. Geopolitics of discourse. Political Geography Quarterly 11, 190–204.
O'Tuathail, G., Luke, T., 1994. Present at the (dis)integration: deterritorialization and reterritorialization in the new wor(l)d order. Annals of the Association of American Geographers 84, 381–398.
Pain, R., 2009. Globalized fear?: towards an emotional geopolitics. Progress in Human Geography 33, 466–486.
Pain, R., 2010. The new geopolitics of fear. Geography Compass 4, 226–240.
Pain, R., Smith, S., 2008. Fear: Critical Geopolitics and Everyday Life. Ashgate, Aldershot.
Painter, J., 1995. Politics, Geography and 'Political Geography.' Arnold, London.
Parker, G., 1998. Geopolitics – Past, Present and Future. Pinter, London.
Power, M., 2007. Digitized virtuosity: video war games and post-9/11 cyber-deterrence. Security Dialogue 38, 271–288.
Power, M., Campbell, D., 2010. The state of critical geopolitics. Political Geography 29, 243–246.
Ratzel, F., 1882. Anthropogeographie. Engelhorn, Stuttgart, Germany (pt.I)/ 1892 (pt.2).
Ratzel, F., 1897. Politische Geographie. Oldenbourg Verlag, Munich, Germany.
Routledge, P., 1996. Critical geopolitics and terrains of resistance. Political Geography 12, 509–552.
Routledge, P., 2002. Travelling east as Walter Kurtz: identity, performance and collaboration in Goa, India. Environment and Planning D: Society and Space 20, 477–498.
Sharp, J., 2000. Condensing the Cold War: Reader's Digest and American Identity. University of Minnesota Press, Minneapolis, MN.
Sidaway, J., 2000. Iberian geopolitics. In: Dodds, K., Atkinson, D. (Eds.), Geopolitical Traditions: Critical Histories of a Century of Political Thought. Routledge, London, pp. 118–149.
Sidaway, J., Power, M., 2005. The tears of Portugal: empire, identity, 'race' and destiny in Portuguese geopolitical narratives. Environment and Planning D: Society and Space 23, 527–554.
Taylor, P.J., 1989. Political Geography. John Wiley, New York.
Thrift, N., 2000. It's the little things. In: Atkinson, D., Dodds, K. (Eds.), Geopolitical Traditions: A Century of Geopolitical Thought. Routledge, London, pp. 380–387.
Wallerstein, I., 1983. Historical Capitalism. Verso, London.
Walsh, E., 1944. Geopolitics and international morals. In: Weigert, H., Stefansson, V. (Eds.), Compass of the World. George Harrap, London, pp. 12–39.
Whittlesey, D., 1939. The Earth and the State. Holt, New York.
Woon, C.Y., 2011a. 'Protest is just a click away': responses to the 2003 Iraq War on a bulletin board system in China. Environment and Planning D: Society and Space 29, 131–149.
Woon, C.Y., 2011b. Undoing violence, unbounding precarity: beyond the frames of terror in the Philippines. Geoforum 42, 285–296.
Woon, C.Y., 2013. For 'emotional fieldwork' in critical geopolitical research on violence and terrorism. Political Geography 33, 31–41.

Geospatial Applications in Archaeology

Devin A White, Oak Ridge National Laboratory, Oak Ridge, TN, USA
Katie A Corcoran, University of Tennessee, Knoxville, TN, USA

Abstract

Archaeology is an inherently spatial discipline and archaeologists have long created and used maps to help them make sense of the past. Among field scientists, archaeologists were also some of the earliest adopters of geospatial technology like geographic information systems, spectral imaging, and laser scanning, all of which have enabled significant discoveries in recent years. As those technologies have become more widespread and sophisticated, so has their use within the field. This article provides an overview of core concepts, technologies, and analytical methods.

Introduction

Archaeology is an inherently spatial discipline. Beginning with excavation, space and time are controlled to the maximum extent possible. Artifacts and features are gradually exposed and mapped through careful horizontal excavation. Site formation processes are gradually understood and mapped through careful examination of stratigraphic profiles. Dates are assigned to everything that can be dated. This attention to spatiotemporal detail allows archaeologists to reconstruct past events and to explain what happened in a particular place at a particular time. The same approach holds at coarser scales, where the locations of households, villages, cultural areas, language groups, exchange systems, migration routes, chiefdoms, and states are mapped in space and time in an attempt to understand broader patterns of human behavior. In this context, the term 'geospatial applications' can be interpreted to mean putting the well-established cartographic and analytic methods used by archaeologists in a more exact locational context, where maps are produced with ever-increasing accuracy and coordinates are tied to real-world projection systems like Universal Transverse Mercator to facilitate interoperability and collaborative research. Doing so, however, misses the point. While archaeologists have taken full advantage of improvements in technology to make better maps, they have also started to leverage more sophisticated forms of spatiotemporal data collection and analysis to generate more meaningful insights. This article will discuss both aspects of geospatial applications in the discipline, beginning with advances in digital cartography, moving to prominent data collection platforms, and concluding with common analytical activities.

Geographic Information Systems

The geographic information system (GIS) concept has transformed and democratized the cartographic aspects of archaeology. At its most basic level, a GIS is a digital framework that allows a researcher to assemble layers of spatially referenced information into a single package that can then be modified as needed to easily produce professional-quality maps. The two most common types of layers are raster (a regularly spaced grid of values, like an image) and vector (points, lines, and polygons). The layers can come from almost anywhere, including satellite images, digital elevation models (DEMs), virtual reconstructions, data collected via GPS in the field, and scanned/digitized maps pulled from older reports or historical sources, but the unifying core concept is that they can all exist within a common real-world coordinate space. This framework gives archaeologists unprecedented control over what kind of stories their maps tell and the flexibility to play with multiple display and layer ordering concepts in a relatively risk-free and low-cost way. Until fairly recently, the only robust GIS software package available to archaeologists was ArcGIS, developed by ESRI. While it is still considered to be the gold standard, significant advances in the open source software domain – in direct response to the relatively high cost of an ArcGIS license – have led to the emergence of two free competitors: GRASS and QGIS. For basic cartographic tasks, all three applications enjoy widespread use within the discipline.

GIS is much more than digital cartography, though. Once information layers are brought into a common space, they can be visually altered, manipulated through algorithms, and/or mathematically combined to produce new information and, ultimately, new insights. This is one of the biggest growth areas for archaeological analysis, given the ubiquity of geospatial data in today's world and the wide variety of other disciplines working on similar or at least somewhat related problems.

GIS has permeated archaeology, but it is largely up to the individual researcher to learn how to use it, leading to the development of bad habits, fundamental knowledge gaps, and a lack of standards and best practices. Some academic archaeology programs, especially outside of the United States, offer training courses in the basics of GIS; a very small number have GIS degree emphasis or certificate options. Also, there is currently no standard reference volume for how GIS can and should be used within archaeology. This situation must improve if the discipline is to fully embrace the technology and all that it has to offer.

Popular Online Geospatial Data Visualization Tools

The world is currently experiencing a geospatial data revolution, fed in large part by the activities of large technology companies like Google, which develops and freely distributes

two widely used data visualization platforms: Google Maps and Google Earth. Both are in essence simplified GIS software packages, which, at their most basic levels, allow you to view street maps and satellite imagery for almost any point on the earth's surface from the comfort of your desk in two or three dimensions, respectively. For the public in general and archaeologists in particular, this is usually their first exposure to remotely sensed data like satellite images and other geospatial products like hydrology maps and DEMs. The currency and quality of the maps and imagery hosted by Google varies widely, but for the most part the company endeavors to provide a good experience in most populated places as well as remote areas that are well-known by the general public (e.g., Grand Canyon). All other remote areas receive updates far less often. Due to the fact that both platforms can be used for free, many professional and avocational archaeologists have spent large amounts of time exploring hosted imagery, looking for previously unknown sites or assessing change at known ones. These explorations have led to amazing and dubious discoveries, but both outcomes reinforce the fact that fieldwork must be done to verify anything found in imagery collected from a significant distance.

Because of how Google Maps and Google Earth are developed and distributed, it is difficult, but not impossible, to do more than view hosted maps and imagery. Users can create their own maps for viewing in Google Maps from geospatial content hosted elsewhere on the Internet, but geospatial analysis is not currently supported. Google Earth does allow the importing of custom raster and vector datasets created using third party software and relying upon a web standard known as Keyhole Markup Language. It also allows users to import three-dimensional models built using free software like Google SketchUp, so reconstructions of archaeological sites can be placed in their approximate locations on the Earth's surface to aid in contextual analysis. With that said, importing large datasets of any type tends to drastically slow down the application. When this happens, it is time to switch to a more robust solution.

While the names and capabilities of online geospatial tools will likely change in the coming years, the trend of working with large amounts of data remotely, through a lightweight application like a web browser, is here to stay. Most desktop computing tasks will be transferred into this domain, which will allow researchers to ask larger and more complex questions without having to invest in specialized hardware and software.

Prominent Geospatial Data Collection Platforms

This section will highlight the two most prominent types of geospatial data collection platforms in use today: spectral imaging and laser scanning. These platforms are complementary, with the former recording how electromagnetic energy either reflects off of or is emitted by objects and the latter using reflected electromagnetic energy to record the precise shape of the objects and their locations. Either one can be used separately, but leveraging both can produce much better results. For the purpose of brevity, the following discussion is primarily focused on how the platforms are deployed via aircraft (spectral and laser) and satellite (spectral), but the same general principles apply to data collected on the ground at much finer spatial resolutions.

Spectral imaging is a dynamic technology applied to archaeological site detection, site characterization, and predictive modeling. The imagery available through applications like Google Maps and Google Earth is produced by spectral systems, but the user is only seeing a small fraction of the data collected (just what our eyes can see) and is not able to perform quantitative analysis. While often used to supplement other remotely sensed data or a GIS, spectral imagery is usually delivered as a stack of images, with each one representing a different portion of the electromagnetic spectrum. These 'slices' can be analyzed individually or as a whole, supporting manipulation to derive secondary products for additional analysis, classify objects in an image, build a reference library for use with similar imagery, or simply emphasize certain visual features of a landscape. Like other remote sensing technologies, spectrometry is attractive to archaeologists due to the large amount of land area that can be quickly imaged. Further, spectral imagery can reveal qualities of archaeological sites and features, or their proxy effects on the surrounding environment, that are not visible to the naked eye. For example, buried archaeological materials might affect how plants grow on the surface, which can be detected by sensors as differences in how certain wavelengths of light are absorbed or reflected by those plants when compared to their neighbors (Parcak, 2009). Near-infrared light, which cannot be seen by humans, is among the most widely examined. It is important to note that spectral imaging is considered to be a passive remote sensing technology, in that the sensor opportunistically collects the energy that reaches it. In contrast, active systems send out their own energy pulses and wait for them to return.

Introduced to archaeology a bit earlier than other remote sensing technologies, spectral imaging became popular in the mid- to late-1970s, becoming what is now one of the most common modern remote sensing approaches in the discipline. Most of the examples of spectral imagery use in archaeology are of work done in Western Europe (Campana, 2002; Ciminale et al., 2009), the Americas (Custer et al., 1986; Masini et al., 2009), and Western Asia. By comparison, surprisingly little work using spectral imagery has been performed in Africa (Clark et al., 1998; Parcak, 2009).

The availability of no- or low-cost satellite spectral imagery has resulted in the increase of its use within many disciplines. Archaeologists are using the imagery in creative ways, often optimizing it by correlating spectral responses with other data sources like Synthetic Aperture Radar (Vining and Wiseman, 2006), satellite photography (Scardozzi, 2008), GIS (Sarris et al., 1996; Vaughn and Crawford, 2009), and even other spectral sensors with differing spatial and spectral resolutions (Leucci et al., 2002; Sarris et al., 2013). Spectral data are often leveraged to create proxy indications of archaeological site presence and distribution, to varying degrees of success. The most well-known proxies involve the cross-examination of different spectral bands for detailed analyses of vegetation presence, plant species identification, soil and plant moisture content, and overall vegetation health. Of the well-known proxies, the most used is

Normalized Difference Vegetation Index (cf Masini et al., 2009).

Beyond data fusion or predictive modeling, spectral imagery can be used to manually or automatically classify objects in an image. Buildings, roads, vegetation, and water are commonly classified early on so a researcher can focus on spectral signatures within the image that are anomalous or patterned, and are perhaps of significant interest to an archaeological investigation. Similarly, the spectral signatures of known archaeological sites in an image can be defined by a user and leveraged to locate similar signatures in the same image, or a different image – areas which may also contain archaeological sites. These methods are common practice in spectral imagery analysis. If done carefully, spectral reference libraries can be established using these observations, ones that can be used in related contexts as a way to save time and maintain consistency between analyses and researchers.

Laser scanning, generally referred to as light detection and ranging (LIDAR), is an active three-dimensional data collection method that has been used by archaeologists for identifying or confirming the spatial extents of sites, for visualizing known and unknown features on the ground surface, and in combination with other data to develop predictive models of archaeological site location. In brief, a single laser beam or collection of beams is directed toward the ground and the time it takes for the beam(s) to return to the sensor, in combination with data about the sensor's current position and orientation, is used to calculate precise coordinates for each return with very small error estimates. Those returns are stored as points, collectively referred to as a point cloud, and are then subjected to one or more processing techniques in order to highlight features of archaeological interest. In many cases, the spatial complexity of the point cloud is reduced to that of a DEM (height measurements on a regularly spaced grid) for software compatibility and ease of use.

The mid-2000s brought about an increased popularity of laser scanning technologies in archaeological investigation. The use of LIDAR in archaeology has grown rapidly and case studies have been published for work done throughout the world, with a clear emphasis on prehistoric sites in the UK and North America (Chase et al., 2011; Dorshow, 2012). Historically, archaeologists have spent their limited time and project funds investigating features discovered through ground-based surveys, native accounts, and/or aerial photography – all which rely on the ability to identify and record areas of interest without much visual assistance. Since the introduction of LIDAR to archaeology, it is becoming increasingly common to scan landscapes in search of subtle artificial features that are not otherwise obvious, then follow up with fieldwork to confirm their significance. In several recent studies, the use of LIDAR has greatly expanded our understanding of the size and complexity of even extremely well-known sites (Chase et al., 2011; Evans et al., 2013; Fisher et al., 2011).

Two scales of LIDAR are common in archaeology: airborne and terrestrial. Airborne platforms are flown overhead, while terrestrial scans are generally collected from a stationary tripod-mounted platform. Airborne LIDAR has traditionally been preferred, but terrestrial LIDAR is becoming increasingly popular due to its relatively low acquisition cost, ease of use, and ability to capture a significant amount of detail over very small areas (e.g., individual sites and features). This historical trend is demonstrated by an overwhelming number of case studies using airborne LIDAR data in archaeology as compared to the handful of terrestrial LIDAR surveys, but the balance is expected to shift in coming years. It is also important to note that the two scales are not mutually exclusive. Some archaeological projects call for analysis at the landscape and site levels.

LIDAR scans generate point clouds that must be filtered to remove obstructive objects like vegetation, before subtle archaeological features become apparent (White, 2013). Many different strategies exist for point cloud filtering, and there are techniques designed for airborne and terrestrial point clouds – with a separation of techniques between data types firmly maintained in the literature by researchers (i.e., airborne filtering techniques are always applied to airborne LIDAR point clouds, and the same rule is followed for terrestrial LIDAR) (Zhang and Whitman, 2005). This is due to distinct geometric differences between the two types of LIDAR data that stem from different sensor orientations relative to the ground or object. There is a relative lack of research on terrestrial LIDAR point cloud processing and analysis within the discipline, likely because it is a newer approach to archaeological investigation. As more processing techniques are developed and become popular, there may be a proportional increase in the use of terrestrial LIDAR by archaeologists.

Since many archaeologists prefer to visually browse their LIDAR data in search of potential archaeological features, effective visualization methods are critical. For example, after filtering out all nonground points, 'bare earth' DEMs can be generated and displayed in different ways to emphasize features on the surface. It should be noted that there is no one correct way to produce a bare earth DEM and none of the currently available methods are optimized for archaeological research. Each one requires experimentation with available software settings and may not be appropriate in all environmental contexts. Most archaeologists prefer the shaded relief visualization method (Horn, 1981), which simulates the projection of light from a single point onto the Earth's surface, simulating sunlit terrain. While shading in this way is by far the most popular visualization technique in published case studies, it is not without problems. Two issues in particular are (1) the 'shaded' areas often contain features that are easily overlooked because they are not directly illuminated and (2) linear features that are oriented parallel to the direction of sunlight are mostly illuminated, resulting in a partial or complete loss of visual detail (Zaksek et al., 2011). This is problematic for the archaeologist whose goal is to emphasize subtle features that may require further investigation. There are recent developments in DEM visualization that simulate diffuse rather than direct light, such as Sky-View Factor (Zaksek et al., 2011), that have improved the way LIDAR data are viewed. Diffuse lighting techniques are expected to gain traction in the archaeological community as an intermediate step to more advanced analyses that involve direct use of the filtered or unfiltered point cloud.

Common Analytical Activities

As mentioned above, archaeologists have begun to move far beyond mapmaking and simple visual analysis of geospatial

data. This last section highlights some of the most common analytical activities undertaken by researchers, albeit at a very high level due to space restrictions.

Spatial Analysis

The topic of spatial analysis in archaeology covers a lot of territory, but at its most basic level, geospatial data are supplied to one or more algorithms that (1) take sparse observations and attempt to fill in the gaps, (2) highlight spatial patterns that are not readily visible to the naked eye through the use of statistics and applied mathematics, or (3) reduce complexity so that the data become more digestible.

With respect to filling in the gaps, generally referred to as interpolation, there are several standard approaches, all of which operate under the assumption that you are starting with point observations and desire to create a continuous surface from them that speaks of trends over a landscape (**Figure 1**). The first, nearest neighbor, moves outward from each supplied point and fills in gaps based on the value of the nearest original point to the current location. Sometimes a decay model is applied during the search process to lessen the influence of known points as distance increases. The second, inverse distance weighting, starts with the use of an exponential decay model, but then either limits the number of points to consider or their distance from the current location during interpolation. The user's choice of how to limit the search will impact the final result. The third, natural neighbor, is the most complex. It begins by connecting all of the known points to one another through a series of nonoverlapping triangles, referred to as Delaunay Triangulation, then uses those triangles to create a set of nonoverlapping complex polygons that represent the neighborhood of influence for each known point. Gaps are filled in by weighing the influence of each neighborhood around a location and its distance from their centers. It is important to note that no one approach is better than another, but the researcher should be very clear in indicating which one they use, and why, when communicating analytical results and their significance.

Trying to discover hidden spatial patterns is another popular research activity. The most common approach is spatial autocorrelation, which is based on the principle that things closer together in space tend to be more like each other than things that are far apart. At its heart is a spatially-aware regression technique that attempts to highlight data patterns that are spatially dependent, which is of great use to archaeologists looking for similarities and differences between entities, and how those might change over time, at multiple scales of analysis. This kind of analysis can be done with both vector and raster data. Another area of emerging interest is spatial network analysis. Social network analysis has been around in the social sciences for decades (Borgatti et al., 2009), but the spatial form of this analytic technique is fairly new and of particular interest to archaeologists who want to examine synchronic and/or diachronic interconnectedness between communities. This can be done by building temporal graphs that connect communities through the one- or two-way exchange of particular resources, languages, and/or iconography, for example. Subjecting these graphs to algorithms that factor in not only the locations of those communities and time but also the strengths of connections and how the change can make the nature of the relationships between those communities much more apparent (Mills et al., 2013).

Finally, in terms of data complexity reduction, it is worth briefly mentioning a technique known as principal component analysis (PCA). Instead of indicating the degree of correlation between data that are spatially near one another, PCA removes the correlation in an attempt to highlight true differences and suggest which aspects of the complex data are the biggest drivers of those differences. This technique, in an aspatial form, is commonly used in artifact sourcing studies (e.g., X-ray fluorescence (XRF) and neuron activation analysis (NAA)) to

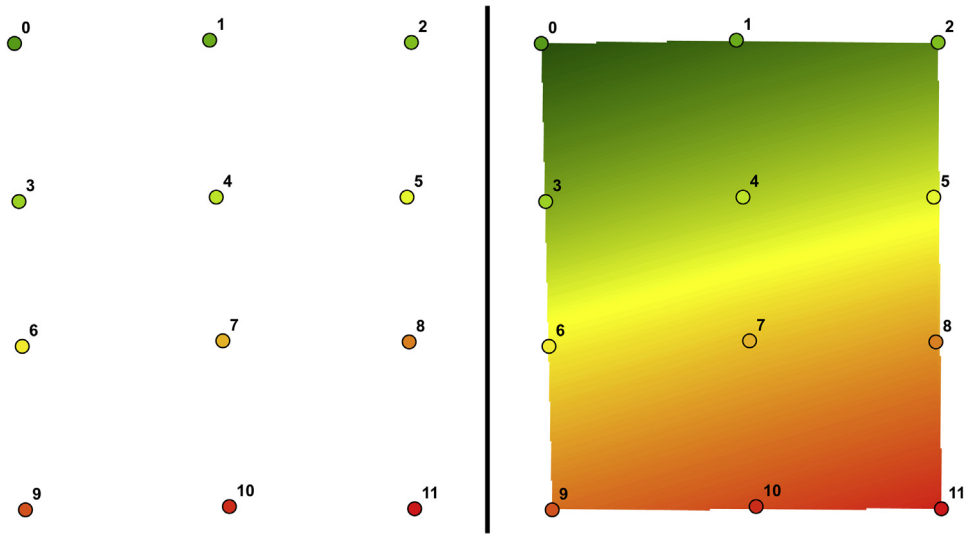

Figure 1 An example of interpolation, where the empty space surrounding individual data points (left) is filled in using the closest known values (right) to form a continuous surface. Colors are coded by data values, which appear next to each point.

determine which elements do the best job of discriminating source groups.

Terrain Analysis

Like spatial analysis, terrain analysis is an expansive subject that cannot be covered in detail here. There are four common analytical approaches that are used to aid in archaeological research, with each one focused on a different aspect of how physical terrain influences human behavior and interaction. For the sake of simplicity, all approaches discussed below relate to the analysis of DEMs. They can be extended to point clouds, but currently available GIS software does not handle that scenario well. Instead, point clouds are reduced in complexity to DEMs and fed into existing algorithms.

The first, morphological analysis, tends to be used to create inputs for the other three, and to aid in predictive modeling, but can also stand on its own. This type of analysis includes slope and aspect derivation, surface roughness derivation, surface type derivation (e.g., valley, ridge, plain), and the generation of both shaded relief maps and related products based on diffuse light.

The second, viewshed analysis, is concerned with how much of the landscape can be seen from a user-selected location on that landscape. It uses line of sight to paint regions of a DEM that are visible and those that are not. This technique is particularly useful when trying to determine which, if any, archaeological sites are visible to one another in a region, if a particular site is in a defensible location, or if any sacred landscape features are visible from a given site (or vice versa). In recent years, it has even been used to help delineate archaeological heritage protection zones where the viewshed is an integral part of the experience and should not be cluttered with signs of modern life (Davis et al., 2013).

The third, least cost analysis, deals with how people can move efficiently across the physical landscape. There are several variations to this type of analysis, depending on the interests of the researcher. The simplest one is proximity analysis, which generates a raster based on the input of a vector feature like a river network. Each cell in the raster indicates how far it is from the nearest part of the feature. Proximity rasters are often used as inputs to predictive models. A more complex approach is to pick a single spot and calculate the minimum cost, usually based on an easily quantifiable metric like distance traveled or time expended, from that location to every other location on the landscape, as well as the routes required to reach them. The most efficient route to each location may not necessarily be a straight line and those routes can be reconstructed by picking a location and tracing the decisions that led to reaching that spot in reverse, using an intermediate product known as the backlink, which most users of GIS software never see. What they do see is the route itself, delivered as either a vector line or a raster with only a small number of cells highlighted that indicates the route. This kind of analysis can be done with many locations, not just one, in order to get a more complete understanding of how people can move across a landscape. Knowing the most efficient ways to move between locations, and how they may or may not line up with known travel routes, is extremely useful information to archaeologists interested in patterns of movement related to exchange, migration, and warfare, just to name a few topics.

The fourth, hydrological analysis, can in some ways be considered a variant of least cost analysis. Morphological inputs derived from DEMs are used to determine the most likely routes water would take from the highest points on a landscape to the lowest. This kind of analysis can be used to determine watershed boundaries and extract potential stream and river networks, with associated flow rates. Knowing where water could be, especially on landscapes with only intermittent water flow through time, is integral to understanding decisions on matters like site placement and the construction of irrigation systems.

Chained Analysis

It is important to note that the analytical methods mentioned above rarely exist in a vacuum. Results from one are often fed into another, and then those results are passed into a third, and so on, until a final product is generated. This is generally referred to as chained analysis, or model building, and all of the currently available GIS software packages allow you to do it, usually through a drag-and-drop visual interface. The outputs of one process become the inputs to another, with the end result often being something like a predictive map of archaeological site locations. One of the main advantages of chained analysis, beyond automating or semiautomating an analytical process, is the repeatability of the results and the ability to share a complex workflow with other researchers, who may not be as technologically savvy, in a way that is easy to use.

Predictive Modeling

Predictive modeling is a useful tool in archaeological investigation that is typically based on multivariate input leading to estimates of settlement distribution, human behavioral trends, and/or archaeological site locations. Predictive modeling has been helpful to archaeologists because of the time and financial limitations that exist for nearly every project. Archaeological landscapes can be vast, making ground-based inspections costly. As geospatial datasets have become more accessible and sophisticated, researchers have experimented with modeling site locations to minimize cost.

Emerging from the needs of cultural resource management firms, the central objective of predictive modeling has been to locate high-probability areas that may contain archaeological sites, which could affect modern planning projects. As a trend, high-probability areas identified by predictive models are designated high priorities for field investigation. Low-priority areas are not necessarily eliminated as prospective archaeological sites, but might require further investigation; these areas would simply not be identified as having an urgent need for investigation without considering additional information. With reliable site prediction, time and funds can be allocated to areas that are likely to contain archaeological sites – a potentially significant way to save resources, if done well. While not universally popular among archaeologists, predictive modeling has been continually explored for the past 50 years, with increasing sophistication of various approaches and degrees of reliability.

Predictive modeling has been used throughout the world, but with a particular emphasis on two areas: the British Isles (Siart et al., 2008) and North America (Chen et al., 2013; Custer et al., 1986; Saturno et al., 2007). In many cases, satellite spectral imagery, airborne laser scanning, ground-penetrating radar, and/or synthetic aperture radar are used, sometimes in combination with each other or with a GIS to inform a predictive model (Clark et al., 1998; Custer et al., 1986; Saturno et al., 2007; Siart et al., 2008). Models have been used to explore regional cultural distribution (Clark et al., 1998; Siart et al., 2008), settlement location (Custer et al., 1986; Graves, 2011; Saturno et al., 2007; Vaughn and Crawford, 2009), and ancient human movement (White and Barber, 2012; Whitley and Hicks, 2003), just to name a few applications.

A variety of multivariate statistical methods have been used to derive predictive models including logistic regression (Custer et al., 1986; Graves, 2011; Vaughn and Crawford, 2009), multiple linear regression (Kohler and Parker, 1986), and linear discriminant analysis (Chen et al., 2013). Some models do not rely on statistical methods, but instead combine datasets to narrow site potential based on the presence of visible patterning between them, or through variable scoring systems based on predefined criteria of site suitability. Sometimes statistical tests are performed on these scored data (Dorshow, 2012; Graves, 2011). Scoring methods depend on the visual interpretation of data by the researcher; the accurate assessment of which rely, presumably, on experience. Currently, archaeologists seem to be in favor of using weighted scoring systems over statistical methods, but example case studies of both can be found in the literature.

The importance of verifying data quality before analysis and following up with ground truth fieldwork cannot be overstated. Predictive models are only as good as the data analyzed, the methods used to interpret those data, and calibration to reality, but in general, model building is an adaptable process that offers the freedom to constantly explore new variables that might improve a model. As statistical and geospatial analytic methods become more integrated within archaeology, data collections become more widely accessible to researchers, and geospatial analysis programs become more sophisticated through time, predictive models will continue to become more coherent and useful to archaeologists.

Agent-Based Modeling

Last, but not least, is spatially explicit agent-based modeling, which is a technique designed to help archaeologists gain insights into the workings of complex adaptive systems (i.e., societies). Originally built for exploring interactions between particles, insects, and even herds of animals and flocks of birds, it is starting to enjoy widespread use in archaeology. Each 'agent' is a person, household, village, etc., that makes decisions based on a programmed set of rules. As simulated agents interact with one another in space and time, their decisions form a complex, nonlinear path toward a set of outcomes that can then be compared to what is known through the archaeological record. By running lots of models with variations in the rules, not only can existing hypotheses be tested, but new ones can be generated. At present, there are two main software suites for agent-based modeling that are used by archaeologists: NetLogo and Repast.

Conclusion

Recent advances in geospatial technology, and the sophisticated analytical methods that come with it, have revolutionized how archaeologists conduct fieldwork and make sense of the data they collect. With that said, archaeologists have always thought about their research problems both spatially and visually, so many of the topics discussed above can be viewed as extensions of established practice. As costs come down and barriers to entry are lowered, a broader swath of the discipline will embrace geographic information science – hopefully as full contributors, not just consumers. Archaeologists have a unique perspective on the complex relationship between people, space, and time that could benefit many other fields.

See also: Geographic Information Systems; Global Positioning System; Spatial Analysis in Geography; Spatial Data; Spatial Sampling; Spatial Thinking in the Social Sciences; Survey and Excavation (Field Methods) in Archaeology.

Bibliography

Borgatti, S.P., Mehra, A., Brass, D.J., Labianca, G., 2009. Network analysis in the social sciences. Science 323, 892–895.

Campana, S., 2002. High Resolution Satellite Imagery: A New Source of Information to the Archaeological Study of Italian Landscapes? Case Study of Tuscany. Conference on Space Applications for Heritage Conservation. European Space Agency, Strasbourg.

Chase, A., Chase, D.Z., Weishampel, J.F., Drake, J.B., Shrestha, R.L., Slatton, K.C., Awe, J.J., Carter, W.E., 2011. Airborne LIDAR, archaeology, and the ancient Maya landscape at Caracol, Belize. Journal of Archaeological Science 38, 387–398.

Chen, L., Comer, D.C., Priebe, C.E., Sussman, D., Tilton, J.C., 2013. Refinement of a method for identifying probable archaeological sites from remotely sensed data. In: Comer, D.C., Harrower, M.J. (Eds.), Mapping Archaeological Landscapes from Space. Springer, New York, pp. 251–258.

Ciminale, M., Gallo, D., Lasaponara, R., Masini, N., 2009. A multiscale approach for reconstructing archaeological landscapes: applications in Northern Apulia (Italy). Archaeological Prospection 16, 143–153.

Clark, C., Garrod, S.M., Pearson, M.P., 1998. Landscape archaeology and remote sensing in Southern Madagascar. International Journal of Remote Sensing 19, 1461–1477.

Custer, J., Eveleigh, T., Kelmas, V., Wells, I., 1986. Application of landsat data and synoptic remote sensing to predictive models for prehistoric archaeological sites: an example from the Delaware coastal plain. American Antiquity 51, 572–588.

Davis, S., Brady, C., Megarry, W., Barton, K., 2013. Lidar survey in the Brú na Bóinne world heritage site. In: Cowley, D., Opitz, R. (Eds.), Interpreting Archaeological Topography: Airborne Laser Scanning, 3D Data and Ground Observation. Oxbow Books, Oxford, pp. 223–237.

Dorshow, W., 2012. Modeling agricultural potential in Chaco Canyon during the Bonito phase: a predictive geospatial approach. Journal of Archaeological Science 39, 2098–2115.

Evans, D., Fletcher, R.J., Pottier, C., Chevance, J.B., Soutif, D., Tan, B.C., Im, S., Ea, D., Tin, T., Kim, S., Cromarty, C., De Greef, S., Hanus, K., Baty, P., Kuszinger, R., Shimoda, I., Boornazian, G., 2013. Uncovering archaeological landscapes at Angkor using LIDAR. PNAS 110, 12595–12600.

Fisher, C., Leisz, S.J., Outlaw, G., 2011. LIDAR – a valuable tool uncovers an ancient city in Mexico. Photogrammetric Engineering and Remote Sensing 77, 962–967.

Graves, D., 2011. The use of predictive modelling to target neolithic settlement and occupation activity in Mainland Scotland. Journal of Archaeological Science 38, 633–656.

Harmon, J., Leone, M.P., Prince, D.S., Snyder, M., 2006. LIDAR for archaeological landscape analysis: a case study of two eighteenth-century Maryland plantation sites. American Antiquity 71, 649–670.

Horn, B., 1981. Hill shading and the reflectance map. Proceedings of the IEEE 69, 13–47.

Kohler, T., Parker, S., 1986. Predictive models for archaeological resource location. Advances in Archaeological Method and Theory 9, 397–452.

Leucci, G., Negri, S., Ricchetti, E., 2002. Integration of High Resolution Optical Satellite Imagery and Geophysical Survey for Archaeological Prospection in Hierapolis (Turkey). International Geoscience and Remote Sensing Symposium.

Masini, N., Lasaponara, R., Orefici, G., 2009. Addressing the challenge of detecting archaeological adobe structures in Southern Peru using QuickBird imagery. Journal of Cultural Heritage 10, e3–e9.

Mills, B.J., Clark, J.J., Peeples, M.A., Haas Jr., W.R., Roberts Jr., J.M., Hill, J.B., Huntley, D.L., Borck, L., Breiger, R.L., Clauset, A., Shackley, M.S., 2013. Transformation of social networks in the late pre-hispanic US Southwest. PNAS 15, 5785–5790.

Parcak, S., 2009. Satellite Remote Sensing for Archaeology, first ed. Routledge, New York.

Sarris, A., Papadopoulos, N., Agapiou, A., Salvi, M.C., Hadjimitsis, D.G., Parkinson, W.A., Yerkes, R.W., Gyucha, A., Duffy, P.R., 2013. Integration of geophysical surveys, ground hyperspectral measurements, aerial and satellite imagery for archaeological prospection of prehistoric sites: the case study of Veszto-Magor tell, Hungary. Journal of Archaeological Science 40, 1454–1470.

Saturno, W., Sever, T.L., Irwin, D.E., Howell, B.F., Garrison, T.G., 2007. Putting us on the map: remote sensing investigation of the ancient Maya landscape. In: Wiseman, J., El-Baz, J. (Eds.), Remote Sensing in Archaeology. Springer, New York.

Sarris, A., Weymouth, J., Cullen, B., Stein, C., Wiseman, J., 1996. The Nikopolis Project – Integration of Geophysical Prospection, Satellite Remote Sensing, and GIS Techniques in the Study of Epirus, Greece. International Symposium on Archaeometry, Urbana, IL.

Scardozzi, G., 2008. Old high resolution satellite images for landscape archaeology: case studies from Turkey and Iraq. Proceedings of SPIE 7110, 1–14.

Siart, C., Eitel, B., Panagiotopoulos, D., 2008. Investigation of past archaeological landscapes using remote sensing and GIS: a multi-method case study from Mount Ida, Crete. Journal of Archaeological Science 35, 2918–2926.

Vaughn, S., Crawford, T., 2009. A predictive model of archaeological potential: an example from Northwestern Belize. Applied Geography 29, 542–555.

Vining, B., Wiseman, J., 2006. Multispectral and synthetic aperture radar remote-sensing-based models for Holocene coastline development in the Ambracian Gult, Epirus, Greece. Archaeological Prospection 13, 258–268.

White, D.A., 2013. LIDAR, point clouds, and their archaeological applications. In: Comer, D., Harrower, M.J. (Eds.), Mapping Archaeological Landscapes from Space. Springer, New York, pp. 175–186.

White, D.A., Barber, S.B., 2012. Geospatial modeling of pedestrian transportation networks: a case study from precolumbian Oaxaca, Mexico. Journal of Archaeological Science 39, 2684–2696.

Whitley, T., Hicks, L., 2003. A geographic information systems approach to understanding potential prehistoric and historic travel corridors. Southeastern Archaeology 22, 77–91.

Zaksek, K., Ostir, K., Kokalj, Z., 2011. Sky-view factor as a relief visualization technique. Remote Sensing 3, 398–415.

Zhang, K., Whitman, D., 2005. Comparison of three algorithms for filtering airborne LIDAR data. Photogrammetric Engineering and Remote Sensing 71, 313–324.

Geriatric Psychiatry

Florian Riese, Psychiatric University Hospital Zurich, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Geriatric psychiatry is the subspecialty of psychiatry that deals with mental health in older persons. Its importance is growing due to the aging of populations worldwide. The practitioners of geriatric psychiatry are physicians working in close collaboration with other medical and nonmedical professionals and carers. The principal mental health problems covered by geriatric psychiatry are cognitive disorders, such as the dementias and delirium, and affective disorders, such as depression. While pharmacological and nonpharmacological treatments exist for the majority of disorders encountered in geriatric psychiatry, their efficacy is often limited. Further research is needed to develop more effective strategies for treatment and prevention and decrease the burden associated with mental illness in the elderly.

Scope

Geriatric psychiatry (also referred to as old age psychiatry, psychogeriatrics, gerontopsychiatry, geropsychiatry) is the subspecialty of psychiatry dealing with mental health in the elderly. Geriatric psychiatry covers the prevention, diagnosis, acute and long-term treatment, rehabilitation, and palliative care of mental illness in this population. To qualify for geriatric psychiatry services, the most frequently applied age limit is 65 years. In principle, the users of geriatric psychiatry services fall into two categories: First, persons who suffer from mental illness throughout their lives and continue to need psychiatric assistance in old age (e.g., for treatment of schizophrenia, depression, anxiety disorders). Second, persons who become mentally ill during their old age (e.g., Alzheimer's disease, late-onset depression). Certain diseases like early-onset forms of dementia may be treated by geriatric psychiatry also in patients younger than 65.

Conceptual Background

As part of medicine, geriatric psychiatry is organized as other medical specialties. Its practitioners are physicians who have first studied medicine, then specialized in general psychiatry and then subspecialized in geriatric psychiatry. However, geriatric psychiatry does not exist as a separate specialty in many countries and training in geriatric psychiatry varies considerably between countries (Camus et al., 2003; Toot et al., 2012). As subspecialty of psychiatry, geriatric psychiatry follows the biopsychosocial model (Engel, 1977), which emphasizes the contribution of biological, psychological, and social factors to mental well-being and illness. Consequently, geriatric psychiatry uses approaches from all three of these areas to improve mental well-being. Recently, the biopsychosocial model has been complemented by a focus on recovery and person-centered care (Hill et al., 2010; Daley et al., 2013), which emphasizes patient empowerment and care provision according to individual health needs and priorities. Like in other fields of psychiatry, neuroscientific research continues to have a major impact in geriatric psychiatry by improving the understanding of disease mechanisms and introducing novel diagnostic and therapeutic tools. In 2009, the Section of Old Age Psychiatry of the World Psychiatric Association (WPA) identified the respect for autonomy, beneficence/nonmaleficence, justice, and veracity as ethical principles of care for mental illness in the elderly (Katona et al., 2009). The statement endorses the human rights of older people with mental illness and takes a stand against ageism, discrimination, and stigmatization.

Development of the Field

While care for mental health problems of the elderly has existed for much longer, geriatric psychiatry as distinct subspecialty of medicine has only developed recently. Around 1900, important foundations for the field were laid by the neuropathologic findings of Alzheimer, Pick, Binswanger, and others. However, in the UK, elderly care developed into a focus of psychiatric research and clinical practice only in the 1930–1950s. Facilitating factors were requisitioning of mental hospital beds for military purposes, employment of the traditional nonprofessional caregivers – women – in the military industry, and increasing life expectancy of the older population strata (Hilton, 2005). Neuropsychiatric assessment scales were applied to measure function and Alzheimer's disease was recognized as disease of the elderly – and not as a disease of middle age as in the first case described by Alzheimer. Most importantly, mental illness in the elderly was begun to be considered not a mere consequence of normal aging (Hilton, 2005). Consequently, the first inpatient geriatric psychiatry wards were set up approximately at the same time as geriatric medicine established itself as an independent specialty. Following the rise in research in the field, international peer-reviewed journals specifically dedicated to geriatric psychiatry were established, including the International Journal of Geriatric Psychiatry (1985), International Psychogeriatrics (1989), and the American Journal of Geriatric Psychiatry (1993) (Meyers and Jeste, 2010). Formal criteria for geriatric psychiatry training were introduced in the UK in 1989, the US in 1993, Canada in 1994, Ireland in 1995, and in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.27031-1

Australia and New Zealand in 1999. In Continental Europe, geriatric psychiatry was only introduced as independent specialty after 2000 (UEMS Section of Psychiatry, 2014). The principles of old age psychiatry were finally laid out by the World Health Organization (WHO) in 1996: A consensus statement on 'Psychiatry of the Elderly' was issued together with the Section of Geriatric Psychiatry of the WPA and other stakeholder organizations that covers the definition of geriatric psychiatry, its treatments, service organization, and training requirements (World Health Organization, 2001).

Professional Organizations

In many countries there are now professional organizations in which the practitioners of geriatric psychiatry are organized, e.g., the American Association for Geriatric Psychiatry. International organizations include the European Association of Geriatric Psychiatry, Section of Geriatric Psychiatry of the European Psychiatric Association, International Psychogeriatric Association, and the Section of Old Age Psychiatry of the WPA. Beyond these professional associations that primarily have physicians in the field as their members, there is a wide variety of other stakeholder organizations that usually center on specific diseases and are organized following a charity model (e.g., the Alzheimer's Association in the US).

Importance

Due to improved social and medical conditions, more people worldwide live into old age than ever. According to the WHO (World Health Organization, 2012b), the number of people aged 60 doubled between 1980 and 2012. The number of people aged 80 will almost quadruple to 395 million between 2012 and 2050. For the first time in history, the number of adults aged 65 and over will soon outnumber children under the age of 5. The majority of older people already now live in low- or middle-income countries. By 2050, this number will have increased to 80%. Consequently, an increasing number of persons falls into the age group treated by geriatric psychiatry. Furthermore, there is a particularly steep increase in the number of nonegarians (90+ years old) and centenarians (100+ years old), i.e., the very old. Special patient populations such as chronically medically ill patients, users of illicit substances, or prison inmates now also reach old age and may require psychiatric attention. At the same time, many of today's elderly people are more healthy and lead more active lives than ever before and try to conserve their abilities and lifestyle as long as possible. As a consequence, there is a high interest in preventing mental illness and improving psychiatric care in the elderly, as psychiatric disorders such as depression and dementia are associated with high levels of disability and burdens of disease (Murray et al., 2012).

While the number of elderly people is rising, the number of people in younger age groups is declining in many populations worldwide. Furthermore, with the current increase in national and international mobility, many younger people do not live close to their parents or grandparents anymore. Both trends exert pressure on the organization of care: Professional caregivers may have to take over a larger share of elderly care, but fewer people have to shoulder the cost of care and fewer professional caregivers are available for demographic reasons. Current strategies to address this problem are task shifting to nonmedical professionals (Bartels and Naslund, 2013), the improvement of working conditions, remuneration and academic standing of nursing staff, international recruitment of health care professionals, increased use of lay volunteers, and introduction of technology including robotics into the care process. However, none of these approaches is likely to fully compensate for the increasing demand for elderly care, so that there is urgent need for research and innovation that may delay the onset of need for care and facilitate independent living until higher age.

Demographic Change and Risk for Dementia

The impact of population aging on mental health is especially visible in the case of dementia, since old age is an important risk factor for the development of dementia. The prevalence for those aged 60 years and older is 5–7% in most world regions, with a higher prevalence in Latin America (8.5%) and a lower prevalence in sub-Saharan Africa (2–4%). The prevalence approximately doubles with every 5 years in age and reaches 40% or even more of all persons >90 years of age. In 2010, an estimated 35.6 million people lived with dementia worldwide. Almost a doubling of the worldwide number of dementia cases is expected every 20 years and while in 2010 58% of all people with dementia lived in low- and middle-income countries, this proportion may rise to 71% in 2050 (Prince et al., 2013). The total worldwide costs of dementia – including for informal caregiving – were estimated to be US$604 billion in 2010, which is equivalent to 1% of the world's gross domestic product (Wimo and Prince, 2012). For the US, the costs associated with dementia were calculated to be similar to the cost associated with cancer or cardiovascular disease (Hurd et al., 2013). Due to the strain for health care systems that is associated with dementia, the WHO recognized it as a public health priority in 2012 (World Health Organization, 2012a).

Some authors suggest that up to 50% of dementia cases are attributable to preventable risk factors (Barnes and Yaffe, 2011). Recent evidence indicates that the prevalence of dementia per age group is indeed declining in high-income countries, supposedly due to the protective effect of education and primary and secondary prevention of cardiovascular disease (Matthews et al., 2013; Christensen et al., 2013). Low- and middle-income countries seem to undergo the opposite trend (Prince et al., 2013; Chan et al., 2013), possibly because risk factors for dementia such as obesity and diabetes are on the rise in these countries. Harnessing the benefits of prevention may thus be crucial to decrease the health care burden associated with dementia, but will also be important in other mental illnesses in old age such as depression, delirium, or substance abuse.

Delivery of Care

The majority of mental health care for the elderly is performed by primary care physicians, geriatricians, and general

psychiatrists, while specialist care by geriatric psychiatry is usually only started after referral from primary or secondary care. Similar to other medical fields, geriatric psychiatry services are delivered in hospitals, in outpatient clinics, through consultation and liaison services, and in community services. In a consensus statement, the WHO and WPA defined good quality care for older people with mental health problems (World Health Organization, 1998). According to this consensus statement, quality geriatric psychiatry services are characterized by comprehensive, accessible, responsive, individualized, transdisciplinary, accountable, and systemic care ('CARITAS'): "A comprehensive service should take into account all aspects of the patient's physical, psychological and social needs and wishes and be patient-centered. An accessible service is user-friendly and readily available, minimising the geographical, cultural, financial, political and linguistic obstacles to obtaining care. A responsive service is one that listens to and understands the problems brought to its attention and acts promptly and appropriately. An individualized service focuses on each person with a mental health problem in her/his family and community context. The planning of care must be tailored for and acceptable to the individual and family, and should aim wherever possible to maintain and support the person within her/his home environment. A transdisciplinary approach goes beyond traditional professional boundaries to optimize the contributions of people with a range of personal and professional skills. Such an approach also facilitates collaboration with voluntary and other agencies to provide a comprehensive range of community-orientated services. An accountable service is one that accepts responsibility for assuring the quality of the service it delivers and monitors this in partnership with patients and their families. Such a service must be ethically and culturally sensitive. A systemic approach flexibly integrates all available services to ensure continuity of care and coordinates all levels of service providers including local, provincial and national governments and community organizations (World Health Organization, 1998)."

Following the CARITAS principle and the general trend for deinstitutionalization of psychiatry, current geriatric psychiatry emphasizes person-centered care in the community. Still, many elderly persons with or without mental health problems live in long-term care facilities, such as residential care or nursing homes. Delivery of care in institutions – albeit not hospitals – therefore remains an important part of geriatric psychiatry.

For delivery of care, specialists in geriatric psychiatry work in close collaboration with other medical and nonmedical professionals. In the medical field, the closest links exist with general medicine, geriatrics, neurology, and palliative care medicine. Other professions that are major constituents of the geriatric psychiatry care process are nursing, (neuro-) psychology, psychotherapy, social work, physical therapy, as well as other nonmedical therapeutic professions. The involvement of family members in the care process is often essential. Due to their medical background and understanding of mental health and old age, geriatric psychiatrists have a central role in planning of such multidisciplinary care and in assuring its continuity.

Mental Disorders in Old Age

In principle, all mental disorders encountered in younger patients can also be found in the elderly. Frequent mental disorders in old age include dementia, depression (Unützer and Park, 2012), anxiety disorders (Blay and Marinho, 2012), and substance abuse (especially the abuse of prescription medications and alcohol (Wang and Andrade, 2013)). Old age is also an important risk factor for suicide (J. Chan et al., 2007), so that assessment and treatment of suicidality is a priority. Other mental disorders such as schizophrenia or bipolar disorder rarely have their onset in old age, but may continue to be active throughout life. A special toll on mental health in the elderly is taken by neurodegenerative diseases such as Alzheimer's disease, Parkinson's disease, or one of several types of frontotemporal dementia. The prevalence of certain mental disorders like autism spectrum disorders, which are more readily associated with younger age is unknown for old age. The frequent comorbidity with medical disease in the elderly (e.g., diabetes, cardiovascular disease) may aggravate mental disorders but also cause mental illness itself, e.g., in the case of vascular dementia (Iadecola, 2013) or delirium (Inouye et al., 2014). Medical comorbidity can also lead to polypharmacy, which increases the risk for medication interactions and adverse events.

As part of the diagnostic routine in geriatric psychiatry, standardized assessments are used that cover the most important psychiatric, general medical, and social health needs in the elderly (Elsawy and Higgins, 2011). Areas that are frequently assessed in this process are cognition, mood, functional capacity, physical health, fall risk, nutritional status, dental status, continence, sexual function, sensory function, medications, living situation, financial situation, social support, spiritual needs, advanced care preferences, and goals of care. Treatment outcomes of patients who undergo such comprehensive geriatric assessment were consistently shown to be better (Ellis et al., 2011). In most countries worldwide, clinical diagnoses in geriatric psychiatry are currently made according to the International Statistical Classification of Diseases and Related Health Problems version 10 (ICD-10) issued by the WHO or according to the Diagnostic and Statistical Manual of Mental Disorders version 5 (DSM-5). Certain disorders may present differently in old age than they do in younger people, e.g., depression or personality disorders. Since ICD-10 and DSM-5 do not encompass specific old age diagnostic criteria for some of those, the diagnostic fit may not be optimal at all times and ultimately relies on clinical expertise. For research purposes, diagnoses in geriatric psychiatry are sometimes made according to criteria other than ICD-10 or DSM-5. Especially for diseases leading to dementia (e.g., Alzheimer's disease, Lewy Body diseases), the diagnostic procedure in research increasingly involves advanced neuroimaging techniques or other biomarkers that help to detect disease phenotypes before the onset of the clinical dementia syndrome.

Based on the assessment, the diagnosis, and the individual goals of care, a treatment plan is made through shared decision making between patient and physician. Frequently in geriatric psychiatry – especially in persons with advanced dementia – decision-making capacity is impaired so that treatment decisions are taken together with health care proxies, usually either

family members of professional proxies. Due to the complexity of health needs in the elderly, treatment plans often involve interventions in multiple areas and by multiple professions. Available psychiatric treatments include pharmacological (primarily medications with antidepressant, antipsychotic, mood-stabilizing, or sedative action) and nonpharmacological interventions (among others, psychoeducation and various forms of psychotherapy, but also biological therapies like electroconvulsive therapy or bright light therapy). To avoid immediate harm to self or others, involuntary treatment is sometimes necessary in severe mental disorders. Involuntary treatments include application of sedative medications and use of physical restraints and are tightly monitored, short-term emergency measures. Many mental disorders in old age are underdiagnosed ('diagnostic gap') and affected persons may not receive adequate treatment ('treatment gap'). Insufficient diagnosis or treatment may be due to lack of resources (including lack of specialized services), stigma of mental illness, but also due to the pervasive misconception that certain mental disorders are inherent part of aging (e.g., dementia).

Research

Research continues to contribute substantially to our understanding of disease mechanisms, diagnosis, and treatment of mental illness in the elderly. An especially active field is the research on neurodegenerative diseases such as Alzheimer's disease. Several lines of evidence suggest that the disease process in Alzheimer's disease starts decades before the diagnosis of dementia (Jack et al., 2013). Consequently, current clinical Alzheimer's research focuses on the development of biomarkers such as radioactive tracer molecules for neuroimaging and biochemical blood or cerebrospinal fluid assays for the detection of proteins that are presumed drivers of the disease process and may thus be suitable markers for early detection. With the increasing possibility for early detection, both pharmacological and nonpharmacological studies are now under way to test the possibility for prevention of the full-blown clinical dementia syndrome.

Still, important gaps remain in our understanding of mental illness in the elderly, and the overall evidence-base in geriatric psychiatry is lagging behind other medical fields. An important contributing factor to this deficit is the exclusion of participants from many clinical trials based on their higher age or medical comorbidity. Furthermore, study inclusion may be more difficult to perform for persons who lack capacity to give informed consent. However, according to a consensus statement by the WPA Section of Old Age Psychiatry, "older people with mental illness should have the right (and be given the opportunity) to participate in research, even if they lack capacity (Katona et al., 2009)." Open questions in the field of geriatric psychiatry include: Are common mental disorders like depression or schizophrenia in old age mechanistically different from young age? What are the shared disease mechanisms that convey risk for medical diseases like diabetes and mental disorders like depression and dementia? Can accurate and reliable diagnostic tests based on the biological disease process be established? Can strategies for prevention of mental illness in the elderly be identified and introduced into clinical practice? How can impaired cognition be restored? What are the active ingredients in nonpharmacological interventions for mental illness in old age? What defines recovery in elderly people with mental illness? How can stigma be reduced and social participation of elderly people with mental illness be improved? How can the growing need for care of elderly people with mental illness be met – particularly in the light of demographic change in low- and middle-income countries?

See also: Aging Mind: Facets and Levels of Analysis; Aging and Health in Old Age; Aging, Theories of; Alzheimer's Disease, Neural Basis of; Alzheimer's Disease: Behavioral and Social Aspects; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Caregiver Burden; Co-morbidity of Mental and Physical Conditions; Cognitive Aging; Dementia: Psychiatric Aspects; Depression; Emotions and Aging; Language Processes in Delirium and Dementia; Mental Health; Population Aging: Economic and Social Consequences.

Bibliography

Barnes, D.E., Yaffe, K., 2011. The projected effect of risk factor reduction on Alzheimer's disease prevalence. The Lancet Neurology 10 (9), 819–828.

Bartels, S.J., Naslund, J.A., 2013. The underside of the silver tsunami–older adults and mental health care. The New England Journal of Medicine 368 (6), 493–496.

Blay, S.L., Marinho, V., 2012. Anxiety disorders in old age. Current Opinion in Psychiatry 25 (6), 462–467.

Camus, V., et al., 2003. Teaching and training in old age psychiatry: a general survey of the World Psychiatric Association member societies. International Journal of Geriatric Psychiatry 18 (8), 694–699.

Chan, J., Draper, B., Banerjee, S., 2007. Deliberate self-harm in older adults: a review of the literature from 1995 to 2004. International Journal of Geriatric Psychiatry 22 (8), 720–732.

Chan, K.Y., et al., 2013. Epidemiology of Alzheimer's disease and other forms of dementia in China, 1990–2010: a systematic review and analysis. Lancet 381 (9882), 2016–2023.

Christensen, K., et al., 2013. Physical and cognitive functioning of people older than 90 years: a comparison of two Danish cohorts born 10 years apart. Lancet 382 (9903), 1507–1513.

Daley, S., et al., 2013. Development of a framework for recovery in older people with mental disorder. International Journal of Geriatric Psychiatry 28 (5), 522–529.

Ellis, G., et al., 2011. Comprehensive geriatric assessment for older adults admitted to hospital. In: Ellis, G. (Ed.), Cochrane Database of Systematic Reviews (Online), (7), p. CD006211.

Elsawy, B., Higgins, K.E., 2011. The geriatric assessment. American Family Physician 83 (1), 48–56.

Engel, G.L., 1977. The need for a new medical model: a challenge for biomedicine. Science (New York, NY) 196 (4286), 129–136.

Hill, L., et al., 2010. Recovery and person-centred care in dementia: common purpose, common practice? Advances in Psychiatric Treatment 16 (4), 288–298.

Hilton, C., 2005. The origins of old age psychiatry in Britain in the 1940s. History of Psychiatry 16 (3), 267–289.

Hurd, M.D., et al., 2013. Monetary costs of dementia in the United States. The New England Journal of Medicine 368 (14), 1326–1334.

Iadecola, C., 2013. The pathobiology of vascular dementia. Neuron 80 (4), 844–866.

Inouye, S.K., Westendorp, R.G.J., Saczynski, J.S., 2014. Delirium in elderly people. Lancet 383 (9920), 911–922.

Jack, C.R., et al., 2013. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurology 12 (2), 207–216.

Katona, C., et al., 2009. World psychiatric association section of old age psychiatry consensus statement on ethics and capacity in older people with mental disorders. International Journal of Geriatric Psychiatry 24 (12), 1319–1324.

Matthews, F.E., et al., 2013. A two-decade comparison of prevalence of dementia in individuals aged 65 years and older from three geographical areas of England: results of the cognitive function and ageing study I and II. Lancet 382 (9902), 1405–1412.

Meyers, B., Jeste, D., 2010. Geriatric psychopharmacology: evolution of a discipline. The Journal of Clinical Psychiatry 71 (11), 1416–1424.

Murray, C.J.L., et al., 2012. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010. The Lancet 380 (9859), 2197–2223.

Prince, M., et al., 2013. The global prevalence of dementia: a systematic review and metaanalysis. Alzheimer's & Dementia 9 (1), 63–75 e2.

Toot, S., et al., 2012. A survey of geriatric psychiatry training across Europe. International Psychogeriatrics/IPA 24 (5), 803–808.

UEMS Section of Psychiatry, 2014. Old Age Psychiatry in Europe, pp. 1–16. Available at: uemspsychiatry.org (accessed 16.07.14.).

Unützer, J., Park, M., 2012. Older adults with severe, treatment-resistant depression. JAMA: The Journal of the American Medical Association 308 (9), 909–918.

Wang, Y.-P., Andrade, L.H., 2013. Epidemiology of alcohol and drug use in the elderly. Current Opinion in Psychiatry 26 (4), 343–348.

Wimo, A., Prince, M., 2012. Alzheimer's Disease International World Alzheimer Report 2010 – The Global Economic Impact of Dementia. Available at: www.alz.org (accessed 15.07.14.).

World Health Organization, 2012a. Dementia: A Public Health Priority. Available at: www.who.int (accessed 15.07.14.).

World Health Organization, 2012b. World Health Day 2012: Ageing and Health: Toolkit for Event Organizers. Available at: www.who.int (accessed 15.07.14.).

World Health Organization, 1998. Organization of care in psychiatry of the elderly-a technical consensus statement. Aging & Mental Health 2 (3), 246–252.

World Health Organization, 2001. Psychiatry of the Elderly – a Consensus Statement, pp. 1–6. Available at: www.who.int (accessed 17.08.12.).

Gesture in Linguistics

David McNeill, University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by S. Kita, volume 9, pp. 6215–6218, © 2001, Elsevier Ltd.

Abstract

Gestures touch classical linguistic themes, especially what can be called the dynamic dimension of language. The article begins by describing a continuum of gesture kinds, and then zeroes in on co-expressive, speech-synchronized 'gesticulations.' The rest of the article describes how gesticulations exhibit and 'fuel' language dynamics in growth points, psychological predicates, catchments and prosody, material carriers and communicative dynamism, the social context of speech, and 'new' actions.

As a field of study, gesture has become energized in recent years. There is now an organization (International Society for Gesture Studies) and a journal (*GESTURE*, Adam Kendon and Cornelia Müller, founding editors). (A web page describing the Society and *GESTURE* is at http://www.utexas.edu/coc/cms/International_House_of_Gestures/.) In preparing this article, I have aimed to balance the empirical basics of gestures with a theoretical perspective from which to regard gestures and see what insights they bring to an understanding of the nature of language.

The word 'gesture' covers a multiplicity of communicative movements, primarily but not always of the hands and arms. Often, gestures are assumed to comprise a channel distinct from speech, but careful investigation challenges this traditional view. Gestures and language are best thought of as a single system, larger than either language or gesture as traditionally assumed. The aspects of gesture I shall emphasize for their importance to linguistics and the dynamic dimension of language is called 'gesticulation.' The hallmark of gesticulation (as opposed to emblems or pantomimes) is that it and speech are synchronous and co-expressive of the same idea units. 'Co-expressive' is not redundant; gesture can express its own aspect of the shared idea unit, but it and speech are expressing the same underling idea. I will give examples below.

The Dynamic Dimension

A prominent aspect of gesture, important for the study of language, but likely to be overlooked, is its relation to the dynamic side of language. Gestures not only provide a window into language dynamics but, in some interpretations, are the engine of it. Without gestures, in this view, language is a crystal palace, but uninhabited.

Speaking is more than uttering speech sounds with meaning; more deeply it is 'inhabiting' language (a term from Merleau-Ponty, 1962). Gesture is a part of how this multilayered inhabitance takes place. The dimensions have classically been called 'linguistic' and 'psychological' but better (and less proprietary) terms are static and dynamic. Some phenomena are accessible or prominent on one dimension, others on the other, but the dynamic and static dimensions cannot be isolated: they intersect and interact. The dimensions are structured on different principles and draw on different methods of description and analysis (the field of linguistics specializing on the static dimension). Yet, both are dimensions of language, and to explain the origin of language we must understand how they work together.

- The static dimension is classically accessed through the synchronic method. Language in this method is viewed as a totality at a single theoretical instant. Saussure (1959) argued that only in this way, with the whole of language laid out panoramically, could the contrasts be discerned that define *langue* (the 'systematic' aspects of a given language) and the linguistic values of words and larger constituents that contrasts generate (and two kinds of value – the syntagmatic on the axis of combination (Jakobson, 1960) and the paradigmatic on the axis of selection). "In language," Saussure said, "there are only differences" (he offered words to illustrate the concept but it applies to phrases and full sentences as well). To use Saussure's example, in English sheep on the hoof versus sheep cooked contrasts lexically, 'sheep' versus 'mutton.' French has no lexical contrast at this point – both states are 'mouton.' The words in the two languages therefore translate each other in one sense, but have different values, necessarily so because value is difference, and 'sheep' and 'mouton,' despite translating each other, differ in their internal-to-*langue* contrasts (speakers of course can distinguish the two kinds syntactically, as in 'viande de mouton'). (Thanks to Gonia Jarema for this example.) Difference and linguistic value are the foundation of the static dimension ('static' does not mean 'stasis' – we are not speaking of moments of repose between bursts of activity: every linguistic event can be regarded statically or dynamically or, as here, both).
- The dynamic dimension could be termed the 'activity' of language but I am calling it (again invoking Merleau-Ponty) 'inhabiting' language with one's being, thought, and action. This terminology has the advantage of alluding to both *langue*, the static system synchronically revealed, and to whatever animates it, the dynamic dimension. A historical figure associated with language so regarded dynamically is Vygotsky (1987). On the dynamic dimension, units come and go, emerge and disperse in real time. It crosses at 90° the abstractions from time and the unmoving totality of synchronic *langue*.

The concept of competence (as in Chomsky, 1965) aligns with Saussurian *langue*, the system of language, but with the difference that 'competence' attributes the system of language to the individual as a personal property whereas *langue* is what Saussure called a "social fact," a socially constituted set of norms. These definitions are not contradictory but take different perspectives on the same linguistic forms.

Parole and 'performance,' the other terms from Saussure and Chomsky, respectively, are defined in relation to their own static counterparts. The formula is *langage* minus *langue* = *parole* (where *langage* is the semiotic totality of language). So *parole* is the 'non-systematic' part (i.e., what remains after subtracting the systematic part or *langue*) of *langage*. 'Performance' taps a different metaphor. It is the 'use' of competence (the static totality of language in a personal knowledge base) subjected to the shortcomings of memory, attention, etc. that are irrelevant to competence. 'Use' indeed belongs on the dynamic dimension but is not the same as 'inhabiting' language with one's imagery and thought. Performance (the technical concept) is the attempt to deploy competence. Inhabitance (our concept) is not 'deploying' – an inappropriate manipulative metaphor – but 'living in' and 'through' language. An imagery–language dialectic is neither 'subtraction' nor 'use.' It is a positive process of its own that requires *langue*/competence (as one pole of the dialectic) but does not derive from it. Linguistic form (regarded either as *langue* or competence) and imagery are both necessary and are equally part of the dialectic.

The Gesture Continuum

To begin, we need to distinguish different kinds of gesture. Not all gestures are included. As mentioned, we focus on one kind, but it is important to understand the others. Adam Kendon once distinguished gestures of different kinds along a continuum that I nicknamed "The Kendon's Continuum" (McNeill, 1992), but now, at his request, have renamed the 'Gesture Continuum.' The gestures we are concerned with are the 'gesticulations.'

'Gesticulation' is motion that embodies a meaning relatable to the accompanying speech. Gesticulation is by far the most frequent type of gesture in daily use and it covers many variants and usages. It is made chiefly with the arms and hands but is not restricted to these body parts – the head can take over as a kind of third hand if the anatomical hands are immobilized or otherwise engaged, and the legs and feet too can move in a gesture mode. In a large sample of gestures, Shuichi Nobe found that the stroke phase of the gesticulation is synchronous with the co-expressive speech about 90% of the time (gesture phases are defined below). When strokes are asynchronous, they slightly precede the speech to which they link semantically, usually because of brief hesitations, and the time gap is small. Gesticulations rarely if ever follow their co-expressive speech (Kendon, 1972). There is no basis for the assertion that strokes occur during hesitations. Such view has attained urban legend status, but it is based on a misrepresentation of the original study by Butterworth and Beattie (1978). They reported that the rate of gesture occurrence was higher during speech pauses than phonations. However, far more gestures occur during phonation than pauses, so the 90% figure results (Nobe also did not replicate their higher gesture rate during pauses, possibly because of different communicative situations: Nobe was looking at narrations, while Butterworth and Beattie had analyzed college tutorials, where gestures during pauses are likely to have had 'turn-suppression' functions not prominent in narrations). The expression 'co-expressive speech and gesture' is explained below. Other controversies have revolved around the issue of whether gesticulations are communicative – 'made for the listener' – or beneficial primarily for speech production – 'made for the speaker' (cf Krauss et al., 2000; Alibali et al., 2000). Gesticulations combine both 'universal' and language-specific features. Speakers of every language studied thus far (and this is a sizable list: in our lab alone, besides English, Japanese, Mandarin, Korean, Spanish, French, German, Italian, Turkish, Georgian, Russian, ASL, Taiwanese Sign Language, and a few African languages) produce them, and the gesticulations for the same events in a cartoon stimulus show clear similarities across these languages. Yet, there are also striking differences which are traceable to characteristics of the languages the gestures are co-occurring with, in particular whether the language is, in Leonard Talmy's typology (Talmy, 2000), S-type or V-type (see McNeill and Duncan, 2000). Gesture space is oriented in terms of absolute compass direction by speakers of Guugu Yimithirr (an aboriginal language with obligatory absolute orientation in its verb morphology) and also speakers of Tzotzil (a Mayan language that lacks the lexical precision of directional reference as seen in Guugu Yimithirr, but whose mode of living promotes exact spatial orientation, which is then embodied in gestures; see John Haviland, 2000).

'Speech-framed gestures' are part of the sentence itself. The term is from Karl-Erik McCullough. Such gestures occupy a slot in a sentence, e.g., "Sylvester went [gesture of an object flying out laterally]," where the gesture completes the sentence structure. These gestures time differently from gesticulations – they occupy a gap that fills a grammatical slot, rather than synchronizing with speech that is co-expressive.

'Emblems' are conventionalized signs, such as thumbs-up or the ring (first finger and thumb tips touching, other fingers extended) for 'OK,' and others less polite. Kendon prefers the term 'quotable gesture,' referring to a potential to be quoted as a gesture, like a word – 'OK' translating into terms of approbation for example. Emblems or quotable gestures are culturally specific, have standard forms and significances, and vary from place to place. Kendon (1995, 2004) has for some years studied the gesture culture of Naples, a locale with an exceptionally rich repertoire of quotable gestures (cf de Jorio, an early nineteenth-century figure, in Kendon, 2000). These gestures are meaningful without speech, although they also occur with speech. They function like illocutionary force markers, rather than propositions, the mode of gesticulation, and when they occur they time with speech quite differently. A single Neapolitan emblem for 'insistent query' (the 'purse' or *mano a borsa*: prototypically, the hand palm up, the fingers and thumb loosely bunched together at the top, and rocking up and down) was observed in one case stretching over several utterances and then continuing into the next speaker's turn, still demanding clarification. This gesture is not employed in North America (a similar looking gesture is used to mean that

something is 'precisely so'), which illustrates the cross-cultural variation of the emblem. Emblems can blend both sequentially and simultaneously with gestures of other kinds. Many emblems have deep historical roots, far outlasting the spoken languages with which they occur. Some go back to Roman times (Morris et al., 1979), including the infamous 'finger,' beloved of the American road – it would have been understood by Julius Caesar. Most emblems have iconic or metaphoric components. The contact of the thumb and forefinger in the 'OK' sign captures the idea of precision. But the emblem is also specified by a convention pairing the form of the gesture to the approbation meaning. The fixity of the emblem is the evidence of this. Putting the *second* finger in contact with the thumb is still precision but no longer is the 'OK' sign of approbation.

'Pantomime' is dumb show, a gesture or sequence of gestures conveying a narrative line, with a story to tell, produced without speech.

And at the other extreme of the Continuum

'signs' are lexical words in a sign language such as ASL. Sign languages have their own linguistic structures, including grammatical patterns, stores of words, and morphological patterns. The linguistic code of ASL is quite unlike that of English. Sign languages have evolved without the requirement of being coordinated with speech. In fact, hearing signers find producing speech and signs simultaneously to be disruptive to both. For an authoritative description, see Liddell (2003).

As one moves along The Gesture Continuum, two kinds of reciprocal changes occur. First, the degree to which speech is an obligatory accompaniment of gesture *decreases* from gesticulation to signs. Second, the degree to which a gesture shows the properties of a language *increases*. Gesticulations are obligatorily accompanied by speech but have properties unlike language. Speech-framed gestures are also obligatorily performed with speech, but relate to speech in a different manner – sequentially rather than concurrently, and in a specific linguistic role. Signs are obligatorily *not* accompanied by speech and have the essential properties of a language. Clearly, therefore, gesticulations (but not the other points along The Gesture Continuum) combine properties – gesture with language – that are unalike, and this combination occupies the same communicative instant. A combination of unalikes, at the same time, is a key psycholinguistic fact and a framework for an imagery–language dialectic. The remainder of this article focuses on gesticulations. If no ambiguity results, from here on I shall use the simpler term 'gesture.'

Traditions of gesture study not summarized in this article because of length are gestures of the theater (Fischer-Lichte, 1992), histories of gesture studies (Bremmer and Rooddenburg, 1991), 'neurogestures' (McNeill, 2005), gestures in human–computer interface design (Wachsmuth et al., 2008), methods of gesture transcription and measurement (Quek et al., 1999), and the gestures of children (Bates and Frederic, 2002), including home signs (Goldin-Meadow, 2003).

The Iconic, Metaphoric, Deictic, Beat Quartet

These categories, or as I will say, dimensions are inspired by the semiotic categories of C.S. Peirce. Elena Levy and I (1982) proposed a classification scheme with four categories: *iconic*, *metaphoric*, *deictic*, and *beat*. All are gesticulations or speech-framed gestures on The Gesture Continuum.

Iconic

Such gestures present images of concrete entities and/or actions. For example, appearing to grasp and bend back something while saying "and he bends it way back." The gesture, as a referential symbol, functions via its formal and structural resemblance to event or objects.

Metaphoric

Gestures are not limited to depictions of concrete events. They can also picture abstract content, in effect, imagining the unimageable. In a metaphoric gesture, an abstract meaning is presented as if it had form and/or occupied space. For example, a speaker appears to be holding an object, as if presenting it, yet the meaning is not presenting an object but an 'idea' or 'memory' or some other abstract 'object' (for examples, see McNeill, 1992; Cienki and Müller, 2008; Müller, 2008). This is a gestural version of the 'conduit' metaphor that appears in expressions like 'he packed a lot into that lecture,' where the lecture is presented as a container and the message as its contents (Lakoff and Johnson, 1980). Recent work on metaphoric gestures has greatly expanded the subject. Cornelia Müller (2008) has developed a new theory of metaphor as a dynamic process (whereby 'sleeping' metaphors are 'awakened' in context) in which metaphoric gestures play an essential part. Metaphoric gestures often indicate that the accompanying speech is meta-rather than object-level – for example, saying "the next scene of the cartoon" and making a conduit cup of meaning gesture (iconic gestures, in contrast, favor the object level).

Deictic

The prototypical deictic gesture is an extended 'index' finger, but almost any extensible body part or held object can be used. Indeed, some cultures prescribe deixis with the lips (Enfield, 2001). Deixis entails locating entities and actions in space vis-à-vis a reference point, which Bühler called the origo (Bühler, 1982; Haviland, 2000). Much of the pointing we see in adult conversation and storytelling is not pointing at physically present objects or locations but is abstract pointing, which Bühler referred to as *deixis at phantasma*. The emergence of abstract pointing is a milestone in children's development. In striking contrast to concrete pointing, which appears before the first birthday and is one of the initiating events of language acquisition, abstract pointing is not much in evidence before the age of 12 and is one of the concluding events (McNeill, 1992).

Beats

It is so called because the hand appears to be beating time. Other allusions to the musical analogy use the term 'baton' (Efron, 1941). As forms, beats are mere flicks of the hand(s) up and down or back and forth, zeroing in rhythmically on the prosodic peaks of speech. This rhythmicity has made beats seem purely speech-related. However, they also have discourse functionality, signaling the temporal locus of something the speaker feels to be important with respect to the larger context. One can think of a beat as a gestural yellow highlighter.

With these four categories, Levy and I were able to classify nearly all gestures in the narrative materials we collected. Other researchers have proposed more finely subdivided categories.

But Not Kinds

I wish to claim, however, that none of these 'categories' is truly categorical. We should speak instead of *dimensions* and say: *iconicity*, *metaphoricity*, *deixis*, '*temporal highlighting*' (for beats), *social interactivity*, or some other equally unmellifluous (but accurate) terms conveying dimensionality.

The essential clue that these are dimensions and not categories is that we often find iconicity, metaphoricity, deixis, and other features mixing in the same gesture. Beats often combine with pointing, and many iconic gestures are also deictic. We cannot put them into a hierarchy without saying which categories are dominant, and in general this is impossible. A practical result of dimensionalizing is improvement in gesture coding, because it is no longer necessary to make forced decisions to fit each gesture occurrence into a single box.

Gestures 'Fuel' the Dynamic Dimension of Language

Tight Binding

The temporal binding of speech and gesture is almost impervious to forces trying to interrupt it. The very heterogeneity of the following observations shows the inviolability of the speech–gesture unit (see McNeill, 2005 for details). Tight binding is an essential feature of how gesture unleashes dynamism, its intrinsic fluidity infiltrating the static dimension and bringing it to life.

Gesture Synchrony and Delayed Auditory Feedback

Delayed auditory feedback or DAF has a dramatic effect on the flow of speech, which slows down, becomes hesitant and is subject to drawling and metatheses (Spoonerisms). Nonetheless, despite the interruptions, speech and gesture remain in synchrony.

Gesture Inoculates against Stuttering

Mayberry and Jaques (2000) made two noteworthy observations. First, the onset of a gesture stroke inoculates against the onset of stuttering. Second, if stuttering begins *during* a stroke, the speaker's hand freezes in midair and may fall to rest. In both observations, we see an incompatibility between the interruption of speech in stuttering and the occurrence of the meaningful gestures.

Gestures of the Blind

Congenitally blind speakers, who have never observed gestures, nonetheless do gesture and do so as frequently as sighted subjects, and gesture even when they know that they are talking to another blind person (Iverson and Goldin-Meadow, 1998). This is dramatic evidence of a speech–gesture bond.

Information Exchange

Information received in a gesture may be recalled later as speech (not gesture) (Cassell et al., 1999). Symmetrically, Kelly, et al. (1999) observed subjects recalling information presented in speech as having been gestural.

Gestures and Fluency

Speech and gesture become complex or simple in tandem, even to the point of jointly disappearing (the gesture disappears along with speech, rather than replacing it).

To Sum Up Binding

Speech and synchronous co-expressive gestures form a tightly bound unit, capable of resisting outside forces attempting to separate them, such as DAF, stuttering, lack of visual experience of gesture, and loss of fluency. Speech and gesture also spontaneously exchange information in memory, so that when something is recalled the speaker cannot tell the original format. Tight binding clearly fosters an imagery–language dialectic by creating unbreakable psycholinguistic units within which it can take place.

Context and Psychological Predicates

Gesticulations (but not other points along the Gesture Continuum) have the property that *strokes synchronize with co-expressive speech*. This section explains this concept. Gesture–speech unity is argued to have been the starting point of language evolution in McNeill (2012). An example is illustrated in **Figure 1** but many other examples can be found. Mostly, the gestures we studied, of which **Figure 1** is an example, were collected in a standardized manner: a participant is shown an approximately 8-min long animated Tweety and Sylvester cartoon. For readers unfamiliar with the genre, Tweety is a pear-shaped, pugilistic canary belonging to an elderly, far from retiring character named Granny. Sylvester is an enterprising cat, endlessly pursuing Tweety with culinary goals, whose lot is frustration and disaster. The genre was familiar to our narrators but not the specific cartoon ('Canary Row,' Warner Brothers, 1950). The cartoon consists of eight episodes, all with the same pursuit-catastrophe theme in amusing variations. It was originally chosen by McNeill and Levy (1982) to be shown to children but most adults also

Figure 1 'Rising hollowness' gesture with: "he goe[s *up through the* pipe] this time." Computer art by Fey Parrill, now on the faculty of Case Western University. Used with permission of University of Chicago Press.

find it engaging. One participant (the 'speaker'), chosen at random, was shown the cartoon in its entirety. Immediately after viewing it, the speaker recounted the story to a second participant "as accurately and completely as possible, as your listener will have to retell the story based on your narration" (or words to this effect). The second participant was a genuine listener, not one of the experimenters (usually a friend or spouse). The performance was audio-video recorded with the speaker in full camera view and at least the front half of the listener as well. The instructions emphasized that the experiment was about storytelling and gesture was not mentioned.) The speaker was describing a cartoon episode in which one character tries to reach another character by climbing up a drainpipe on the inside. The speaker is saying, "and he tries going *up thróugh* it this time," with the gesture occurring during the boldfaced portion (the illustration captures the moment at which the speaker is saying the vowel of 'through'). Co-expressively with 'up' her hand rose upward, and co-expressively with 'through' her fingers spread outward to create an interior space. The upward movement and opening of the hand took place simultaneously, and both were synchronized with 'up through,' the linguistic package that carries the same meanings. (What Kendon (1980) called the gesture phrase – one complete "manifestly expressive action" – is enclosed within "[" and "]." In any gesture phrase up to five distinguishable phases can occur. Not all phases need be present, but one or more *strokes*, the image-bearing phase, marked in boldface, is obligatory; without it, a gesture is not said to have occurred. **Figure 1** had one stroke during 'up through.' The *preparation* phase is the hand getting into position to make the stroke and is indicated by the span from the left bracket to the start of the boldface stroke. The preparation shows that a gesture is coming into being. *Holds* are cessations of movement, either a prestroke, the hand frozen awaiting the stroke, or a poststroke, the hand frozen in the stroke's position and hand shape even though movement has ceased. Holds of either kind are indicated with underlining. Speech notations: a *slash* '/' mans a silent pause; *repeated letters* a filled pause; and an *asterisk* an interruption.)

The effect is a uniquely gestural way of packaging meaning – something like 'rising hollowness,' which does not exist as a semantic package of English at all. The gesture and the linguistic construction synchronize as a whole, not component by component. Thus, speech and gesture synchronize at the point where they are co-expressive and this, not the components, is the unit that aligns them.

Growth Points and Context

Synchronous combinations of such unlike modes of cognition – visuospatial-actional gesture synchronized with analytic-combinatoric speech – may be operative psycholinguistic units, termed 'growth points,' or GPs (McNeill and Duncan, 2000). One way to think of a GP is as imagery that is categorized linguistically, an image with a foot in the door of language. A GP is inferred from the totality of communicative events, with special focus on speech–gesture synchrony and co-expressivity. It is called a GP because it is meant to be the initial form of a thinking-for-speaking unit out of which a dynamic process of utterance and thought organization emerges.

The Psychological Predicate

Regarding the GP as a psychological predicate (a term from Vygotsky, 1987 – not always a grammatical predicate) suggests a mechanism of GP formation in which *differentiation of a focus from a background* plays an essential part.

The concept of a psychological predicate illuminates the theoretical link between the GP and the context of speaking. Defining a psychological predicate requires reference to the context; this is because the psychological predicate and its context are mutually defining. The psychological predicate:

1. Marks a significant departure in the immediate context; and
2. Implies this context as a background.

We have in this relationship the seeds for a model of real-time utterance generation and coherent text formation.

First, when gestures and speech synchronize they jointly form the contrast underlying a psychological predicate. Second, the form of the gesture embodies the content that makes this differentiation meaningful. These correspondences can be demonstrated by exploiting a quirk in the cartoon stimulus that we have employed, in which Sylvester attempts to reach Tweety by climbing a drainpipe conveniently running up the side of the building from street level to the floor where Tweety is perched; he does this twice – first on the outside of the pipe, then on the inside. If a speaker recalls both attempts and in the correct outside–inside order, the gesture–speech combination relating to the second attempt includes a focus on interiority; this is the differentiating element. If a speaker misses the outside attempt but does recall the inside attempt, or recalls them in reverse order, interiority does not now contrast with exteriority, and the gestures with such recall do not include it as a particular feature. These results have been reported by Susan Duncan at workshops; I summarize them, with illustrations, in McNeill, 2005.

Catchments and Prosody

The field of meaningful oppositions a psychological predicate differentiates can be discovered directly in the gestures themselves. We then uncover yet another form of gestural discourse cohesion. Catchments are when space, trajectory, hand shape, etc. recur in two or more (not necessarily consecutive) gestures. Catchments show the effective contextual background and provide an empirical route to the discovery of the discourse context.

- A catchment is recognized from recurrences of gesture form features over a stretch of discourse.
- It is a kind of thread of consistent visuospatial action imagery running through the discourse and provides a gesture-based window into discourse cohesion.
- The logic is that discourse themes produce gestures with recurring features; these recurrences give rise to the catchment.
- Thus, reasoning in reverse, a catchment offers clues to the cohesive linkages in the text with which it co-occurs.

Adam Kendon, in 1972, published a detailed analysis of a filmed conversation, and identified and correlated three hierarchies – kinesic, prosodic, and discursive. We can follow up on Kendon's analysis making use of the concepts of a psychological predicate and catchment. In our study (McNeill et al., 2001),

subjects were asked to describe their living quarters to an interlocutor. In one case, describing her house, the following occurred:

1. so you're in the kitchen
   n' there's a sss-
   the back stairc*
   oh I forgot to say
   when you come though the*
   when you enter the house from the front
   and you open the door with the*
   the glass in them
   there's a* the front staircase runs right up there on your left
   so you can go straight up stairs to the second floor from there if you want
   but if you come around through the kitchen into the back
   there's a back staircase that winds around like this
   and puts you up on the second floor

The following four catchments can be identified covering this passage:

C1 consists of right-hand gestures elevated above the right knee; all are associated with *the kitchen* at the back of the house. See **Figure 2(a)**

C2 consists of two-similar hand gestures; all are associated with the theme of *the front doors* of the house. See **Figure 2(b)**.

C3 consists of left-hand gestures made with the arm extended and lifted up; all are associated with *the front staircase and second floor*. See **Figure 2(c)**.

C4 consists of right-hand gestures where the hand rises and turns in a spiral motion, while the left hand remains in an elevated hold; all are associated with *the back staircase and its relationship to the second floor*. See Figures 2(d) and 2(e).

Each catchment is distinctive in form, location, and/or movement and has nonconsecutive occurrences. The *C2* 'front door' catchment links back to a much earlier description of the front doors of the house. The centerpiece of this discourse is the back staircase (*C4*) and its location at the back of the house where it connects the kitchen to the second floor. The first mention of the back staircase is immediately aborted ("oh I forgot to say") and is replaced by *C2* 'front door' and then *C3* 'front staircase.' *C3* is held as the kitchen *C1* and back staircase *C4* catchments resume. *C3* is hierarchically dominated by *C1*/*C4*, as we infer because it is motionless. One interpretation of the discourse is that the speaker aborted the first mention of the back staircase when she recalled that she had yet to mention another way to reach the second floor of the house that was linked to the front entrance. A repair was undertaken by introducing *C2* (the front doors, going back to the beginning of the house tour) and from there presenting the front staircase and second floor (*C3*). At the end of this repair, an overlap of catchments took place when the elevated left hand from *C3* was held, now representing the second floor, and the back staircase *C4* resumed and connected to it. The two-different hands gesture thus created an overlap of the *C3* and *C4* catchments.

The arrangement in **Table 1** shows a hierarchy of *discourse purposes* as revealed by the Nakatani et al. (1995) query procedure. Position in the hierarchy is indicated by indentation; gesture location is shown with boldface. The text is broken up so that each line is a single prosodic phrase. The numbers refer to the hierarchical level of the purpose, as determined by the procedure, which consists of asking and answering with a purpose why each line was uttered (the procedure applies only to statements that fulfill a purpose, not always the case, but here it was). The hierarchy and the discourse's catchment structure correspond closely – in fact, 100%!

Each catchment had its own purpose level or levels, not shared by the other catchments. At this degree of delicacy, there is a perfect mapping of the discourse structure onto gesture and the speaker apparently created discourse segments on the basis of consistent WHY purposes. Gestures are thus here accurately accounted for as presenting information that is relevant to the WHY? questions of the purpose hierarchy. For example, the answer to purpose 1.1 is 'in the kitchen,' and a gesture was performed that conveyed this content (the hand held in a space identified as the kitchen). Predicting gestures from purposes suggests that the WHY? hierarchy was guiding the speaker by defining what was newsworthy at each point, hence her psychological predicates, and that gestures expressed this content. The catchment is the base from which the communicative weight of the gesture is formed. Each gesture is simultaneously shaped by its semantic content and its relationship to a catchment.

Thus several dimensions converged: The catchment determined the form of the gesture – it was formed to differentiate the next step in the catchment; the utterance purpose defines the communicative weight; the gesture provides the content.

The third leg of Kendon's triad is prosody. To access this dimension, we made use of the ToBI (Tone-Break-Index) analytic system. Following Beckman and Hirschberg (1994),

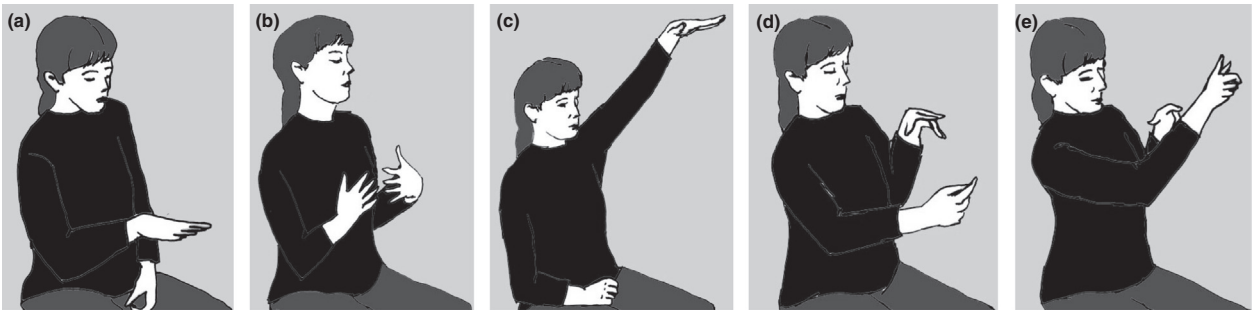

Figure 2 (a). *C1* KITCHEN gesture with "so you're in the kitchen." (b) *C2* FRONT DOORS gesture with "when you enter the house." (c) *C3* front staircase/second floor gesture with "to the second floor." (d) *C4* back staircase gesture with "there's a back staircase that winds around like this." (e) *C3* + *C4* BACK STAIRCASE + SECOND FLOOR gesture with "and puts you up on the second floor." Used with permission of John Benjamins Publishing Company, Amsterdam/Philadelphia. www.benjamins.com.

Table 1 Hierarchy of purposes

WHY? to locate the back staircase (1.1) *C1*
[so *you're in the* kitchen]
WHY? ways of getting to the second floor (1) *C4*
['n' then there's an *s<sss>**]
[The *back stairc**]
WHY? to note the existence of the first staircase (1.1.1)
[*I forgot to say*]
Why? to restart the tour (1.1.1.1) *C2*
[When you come *through* the*]
[When *you enter* the house from the front]
[an*nd you*<ou> *open*m the doors with t][he*]
[<uum*m*> %*smack*/]
[*I the glas*][s *inn them#*]
WHY? to explain first staircase (1.1.1) *C3*
[there's a*
The *front staircas*e] [runs
*Right up the*re
o][n* on yo*ur lef*t]
[So you can go *straight up*][*stair*]
s to th*e se*][ec*ond floo*][r fro*m the*re]
[If *you wan*nt]
WHY? to locate the back staircase (1.1) *C1*
[But if *you come around throu*gh the ki]
[tch*en into t*he bac][k
WHY? ways of getting to the second floor (1) *C4*
there's a back s sta]
[Ircase th*at winds around like th*is]
WHY? to connect to the second floor (1.2) *C4+C3*
[And *putss you up on the second floor*]

Used with permission of John Benjamins Publishing Company, Amsterdam/Philadelphia.www.benjamins.com.

a ToBI transcription occupies four tiers: (1) orthographic transcription (as above), (2) a tone tier in which phrasal tones and pitch accents are marked, (3) a break-index tier in which juncture degree is rated between each pair of words and after the final word, and (4) a miscellaneous tier (comments by the coder). We focused on tiers (2) and (3). (ToBI coding by Karl-Erik McCullough.)

The more deeply embedded a segment in the discourse, the higher the final boundary tone, conveying a 'more is to come' meaning. The more dominant the segment, correspondingly, the lower the final tone, reflecting a declarative pattern. This contrast evidently reflects a general characteristic of intonation contours with independent and dependent content.

Prosodically, each catchment had its own distinctive boundary tone. The *C1* (kitchen) and *C4* (the back staircase) catchments, which come together at the end of the discourse in terms of the spatial layout of the house, were Low. *C2* and *C3* were preponderantly High.

What factors influence the boundary tone of a given catchment? Two seem important. One is the discourse embeddedness of the content, noted above, in which embedded content tended to have high final tones and main-line content low final tones. Iconicity is a second factor. The high and low tones seem to show semantic motivation (cf Bolinger, 1986) in that phrases having to do with the base of the stairs tended to end Low and those with the second floor High. Thus aspects of intonation behave like gestures themselves, and are predictable from knowledge of both semantic content and position in the discourse structure.

The Break Index code also shows the degree of phonological distinctiveness at each 'break' point in the speech stream, ranging from 0 for the highest degree of phonetic reduction to 4 for a full intonation phrase boundary. The *C4* catchment (the back staircase) had a high level of phonetic reduction (Level 0), reflecting more internal continuity of speech. In a seeming paradox, *C4* also had the highest proportion of full intonation phrase boundaries (Level four). The paradox is only apparent, however, and makes sense if we consider that the back staircase was the dominant catchment of the full discourse. This central position yielded tightly configured gesture-prosody packages – internal boundaries glossed over while external boundaries were maximized.

Our general conclusion is that the organization of discourse is inseparable from gesture and prosody: the three components are different sides of a single mental-communicative process. A purely text-based approach, as in the narratology tradition, is blind to two-thirds of this discourse structure. Indeed, the principal themes of the living space discourse were *gestural* – the four catchments embodied the speaker's intentions and were the foundations of the discourse purposes of successive utterances. The utterance hierarchy grew out of these images but only partially encoded them (as we see in the greater delicacy of discourse information in gesture). It was to present catchment themes that the discourse, at each moment and step by step, was organized. Prosody, the other component of the triad, is gesture in spoken form, as in Bolinger (1986). The psychological predicate is the basis for integrating all these components. It provides the co-equal generation of gesture and speech from the same semantic intent. The catchment is the locus around which this integration proceeds.

Material Carriers and Inhabitance

An abiding feature of the dynamic dimension is the creation of what Vygotsky (1987) termed material carriers – a phrase referring to the embodiment of meaning in enactments or material experiences. (Pointed out to me originally by Elena Levy. The quote (rediscovered by Tae Kunisawa) is: "That which is specific to this particular form of sound has remained unexplored. As a consequence, this research has not been able to explain why sound possessing certain physical and mental characteristics is present in human speech or how it functions as a component of speech. In a similar manner, the study of meaning has been defined as the study of the concept, of the concept existing and developing in complete isolation from its material carrier. To a large extent, the failure of classic semantics and phonetics has been a direct result of this tendency to divorce meaning from sound, of this decomposition of the word into its separate elements" (Rieber and Carton, 1987, p. 46).) The material carrier deepens insight into gesture and also into other forms of motor execution, writing something down for example, the very act of it, whatever it is enhancing the symbolization's existential power. Materialization implies that the gesture, this natural material image, the actual motion of the gesture itself,

is a dimension of meaning. Enhancement of language is possible if the gesture is the image in an imagery–language dialectic, not an 'expression' or 'representation' of it, but is it. From this viewpoint, a gesture, the global-synthetic whole, is an image in its most developed – that is, in its most materially, naturally embodied – form. The absence of a gesture is the converse, an image in its least material form. The material carrier concept thus explains how an imagery–language dialectic still is possible in the absence of gestural movement. When there is no gesture there is still imagery and a dialectic with linguistic categorization, still the simultaneous rendering of meaning in opposite semiotic modes – the dialectic in its essentials – but it is bleached, experienced at the lowest level of materialization. This leads us to expect that gestures are more elaborate, more materialized, and more frequent – more present – when the psychological predicate has greater newsworthiness.

In Figure 3 we see a range of materializations, from zero to full, on two continua – speech and gesture. "Communicative dynamism" (CD) is the extent to which a given spoken or gestured form "pushes the communication forward" (Firbas, 1971). The continua correlate positively with CD in terms of the elaboration of the quantity of 'substance' involved. The most elaborate NPs are accompanied by the most developed gestures, the least with the least. So the more discontinuous an utterance is with the previous context, the greater the CD, the more probable a gesture, the more internally complex it will be, and the more complex the synchronous speech as well.

The concept of a material carrier is raised to a whole new level when we turn to Merleau-Ponty (1962) for insight into the unity of gesture and language and what we can expect of gesture in a dual semiotic process. Gesture, the instantaneous, global, nonconventional component, is 'not an external accompaniment' of speech, which is the sequential, analytic, combinatoric component; it is not a 'representation' of meaning, but instead meaning 'inhabits' it:

> The link between the word and its living meaning is not an external accompaniment to intellectual processes, the meaning inhabits the word, and language 'is not an external accompaniment to intellectual processes.' (Merleau-Ponty's quotation is from Gelb and Goldstein (1925: p. 158).) We are therefore led to recognize a gestural or existential significance to speech.... Language certainly has inner content, but this is not self-subsistent and self-conscious thought. What then does language express, if it does not express thoughts? It presents or rather it *is* the subject's taking up of a position in the world of his meanings
>
> p. 193; emphasis in the original. (I am indebted to Jan Arnold (pers. comm.) for this quotation.)

The GP is a mechanism geared to this 'existential content' of speech – this 'taking up a position in the world.' Gesture, as part of the GP, is inhabited by the same 'living meaning' that inhabits the word (and beyond, the discourse). The correlation of speech and gesture complexity in Figure 1 thus involves greater 'inhabitance' (if this concept can be gradual, half in, half out); it involves more of the body, in movements more coordinated by the significance being differentiated from context. It is a thought–language–hand link simultaneously affecting both motor domains.

A striking illustration of materialization is Cornelia Müller's (2008) "waking of sleeping metaphors," the breath of new life given by a gesture to an inactive or 'sleeping' metaphor. It shows the effect of the gesture as a material carrier. Müller gives the example of a German metaphor ('gefunkt,' 'sparked,' the equivalent to English 'clicked,' for suddenly falling in love). The metaphor is usually inactive, experienced as an arbitrary term. However, a speaker, describing her first love, said, "between us somehow it sparked ['clicked']" (Müller's translation). As she said 'between us' her hand rose upward touching her face at eye level, in a ring shape but with an unusual orientation – the fingers pointing at her own face; then, as she uttered the metaphor itself, 'gefunkt,' her hand abruptly turned outward – her gesture materializing the 'dead' (actually,

Communicative dynamism (CD)

Most continuous/predictable				Least continuous/predictable
Less materialization →				More materialization
Linguistic form continuum				
∅	Unstressed pronoun	Noun phrase	Modified noun phrase	Clause or verb phrase
Gesture form continuum				
Referring term included in ongoing iconic	Referring term excluded from adjacent iconics	Iconics that cover clause or verb phrase	O-VPT iconic with noun phrase	Deictics with clause or VP O-VPT iconics (one hand) O-VPT iconics (two different hands) C-VPT iconics

Figure 3 Communicative dynamism. From McNeill, David, 2005. Gesture and Thought. University of Chicago Press, Chicago, p. 55. Linguistic continuum based on Givón 1985. Used with permission of University of Chicago Press.

sleeping) metaphor as a sudden event, an electrical spark out to her loved one.

A deeper answer to the query, therefore – when we gesture, what are we doing? – is that we are updating our current cognitive being, our very mental existence, while it is occurring. By performing the gesture, a core idea is brought into concrete existence and becomes part of the speaker's existence at that moment. The Heideggerian echo in this statement is not accidental. Following Heidegger's emphasis on being, a gesture is not a representation, or is not only such: It is a form of being. From a first-person perspective, the gesture is part of the immediate existence of the speaker. Gestures (and words, etc., as well) are themselves thinking in one of its many forms – not only expressions of thought, but thought, i.e., cognitive being, itself (see Dreyfus, 1994). To the speaker, gesture and speech are not only 'messages' or communications, but are a way of cognitively existing, of cognitively being, at the moment of speaking. (The 'H-model' avoids the homunculus problem encountered by the third person perspective inherent in the concept of a 'representation' and with it the 'theater of the mind' problem highlighted by Dennett (1991). The theater of the mind is the presumed central thinking area in which representations are 'presented' to a receiving intelligence. The possibilities for homunculi – each with its own theater and receiving intelligence – spiraling down inside other similarly endowed homunculi in this theory are well known. In the H-model, there is no theater and no extra being; the gesture is, rather, part of the speaker's momentary mode of being itself, and is not 'watched.')

The speaker who creates a gesture of Sylvester rising up fused with the pipe's hollowness is, according to this interpretation, embodying thought in gesture, and this action – thought in gesture–action over the thought–language–hand link – was part of the person's being cognitively at that moment.

To make a gesture, from this perspective, is to bring thought into existence on a concrete plane, just as writing out a word can have a similar effect. There is not a causal sequence, thought → speech/gesture; speech and gesture are the thought coming into being at that instant. The greater the felt departure of the thought from the immediate context, the more likely its materialization in a gesture. Thus, gestures are more or less elaborated depending on the importance of material realization to the existence of the thought (Figure 4 again).

Social Context

Vygotsky (1987), in his debate with Piaget, famously asserted that everything appears in a child's development twice, first on the social plane, later on the psychological. The concept of a social-to-psychological transition can be applied to gestures as well. Gestures imply a social other. If one denies access to a listener, the frequency of gesture declines (Alibali et al., 2000). Increasing the number of listeners changes the shape of the gesture space and this in turn changes the direction of gestures depicting movements of characters in the story (Özyürek, 2000).

Mimicry and Gestures in Two Bodies

Gestural mimicry occurs when one conversational partner, within a short time, reproduces the gesture of the other partner; the effect is often to cement a kind of social solidarity and is accordingly seen commonly between close friends. The example in Figure 4 from research by Irene Kimbara (2006) shows the woman on the left reproducing the gesture just performed by her friend on the right.

A different kind of shared gesture is also possible between strangers. The following illustrations are from Nobuhiro Furuyama (2000) who studied the gestures of learners as they were instructed in a new motor task (*origami* paper folding). In each illustration, the learner is seated at the left (two different training sessions).

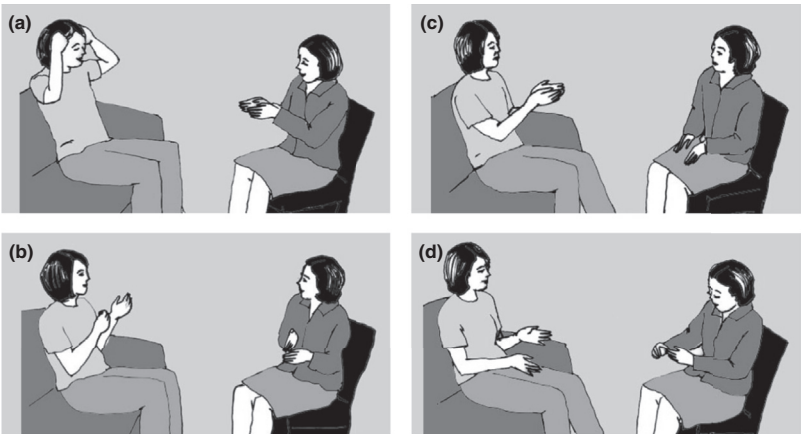

Figure 4 Interpersonal mimicry. (a and b) Speaker on the right: describing the line as 'irregular'; her gesture depicts lines of waiting passengers; the separation of her hands may depict the density of the crowding. Speaker on the left: in (b), hands entering the gesture space and preparing to perform gesture in panels c and d. (c and d) Continuous with panels a and b. Speaker on the left mimics right speaker's two-line gesture as she emphatically agrees ('yes, yes, yes, yes'), including absolute direction (in both figures, the hands are moving toward camera). Meanwhile, in (d), right speaker is preparing her next gesture. Computer art by Fey Parrill. From Kimbara (2006), a, b, c, d. Used with permission of the University of Chicago Press.

In the Figure 5(a) example, the learner mimics the gesture of her tutor. The mimicry occurred without speech by the learner, but her *gesture* synchronized with the *tutor's speech*. As the tutor said, "[pull down] the corner," the learner performed the gesture during the bracketed portion.

The Figure 5(b) example, in contrast, shows the learner appropriating the tutor's *gesture* by combining it with her *speech*. The learner is saying "[you bend this down?]," and during the bracketed speech moved her hand to the tutor's hand, and then moved her hand downward and away (the illustration shows the start of the gesture). As Furuyama observes, the tutor had turned in his chair so that the same left–right gesture space was before him and the learner, a maneuver that might have invited her to enter his gesture space. It is striking that any taboo against strangers coming into physical contact was overridden while the hands were in a symbolic mode.

Examples of these kinds address what can be called collectively the 'social resonance' of gesture. Janet Bavelas has long been interested in this phenomenon. Bavelas et al. (2000) have demonstrated resonance not from the viewpoint of shared feelings, the more traditional approach, but as joint cognition and storytelling; gestural repetition seems to act as grounding and in general conveys mutual understanding.

Gestures – Do They Stem from Practical Actions?

Speech is a 'new action' and so also is gesture, in the sense that any plausible primate antecedents have been altered by the evolution of language itself. They and speech have been unified. The origin of language brought gesture and speech together in an imagery–language dialectic, and this spawned new forms of each. Does anything remain of 'old actions' of the hands and body in this new action gesture world? Many of my fellow researchers of gesture and action consider practical actions to be the source of gestures in evolution as in daily life. Kendon regards gestures as ritualized practical actions. Streeck (2010), LeBaron and Streeck (2000), and Müller (1998) have also proposed that gestures stem from practical, world-directed actions. Kendon in his plenary talk at the Frankfurt/Oder International Society for Gesture Studies conference (2010) likens a gesture of something being flat to "*making* something flat" (p. 14, emphasis in original) and more generally, alluding to Streeck, that "many of the actions [i.e., gestures] understood as meaningful can be understood as being derived from actions of the hands acting in relation to objects in the environment" (p. 14, emphasis in original).

The idea that gestures derive from actions is plausible at first glance but there is more (or perhaps less) than meets the eye. A gesture may look like a pragmatic action but the action has changed at its core. To describe a gesture as 'outlining' or 'shaping' is useful as a description but to also say that such a practical action is still within the gesture is to disregard what makes the gesture a human sign. Our initial example, the 'rising hollowness' gesture (Figure 1), looks like the action of lifting something in the hand, but the gesture is not lifting at all. It is an image of the character rising, of the interior of the pipe through which he rose, and of the direction of his motion upward – all compacted into one symbolic form to differentiate a field of meaningful oppositions having to do with HOW TO CLIMB A PIPE: the gesture shows ON THE INSIDE. This complex idea, as a unity, orchestrated the hand shape and movement; it is the same motor response but it is not the action of lifting up an object.

While a gesture may engage some of the same movements and tap in part the same motor schemas as an 'action–action' it has its own thought–language–hand link in the brain. In keeping with the unity of speech and gesture, the manual movements of gestures as well as the actions of speech are co-opted and orchestrated in new ways by significances other than those of the original vocal tract and manual actions. We observe the separation of 'action–actions' and 'gesture–actions' directly in the IW case where, without vision, action–actions are impossible but gesture–actions are normal (see McNeill et al., 2008).

To show the difference that 'new actions' make for explaining gesture and speech, I will take, for my purposes, an especially challenging example from a talk by Cornelia Müller (at the University of Chicago, 7 May 2011). The example is challenging because, at first sight, it compellingly shows a gesture coming from an 'action–action' source. But we can also see that if there was an action–action element, it was secondary, recruited by a *sui generis* gesture that had nothing to do with the action–action but was a contrast in a field of meaningful oppositions that the speaker had created as part of speaking itself. Müller described a conversation she had filmed in which a speaker (of Spanish) first outlined an oval shape with a two-handed open-palm gesture for the shape of the wall plaque, then outlined it again with extended forefingers. Was he tracing an outline of the plaque with his fingers? In speech,

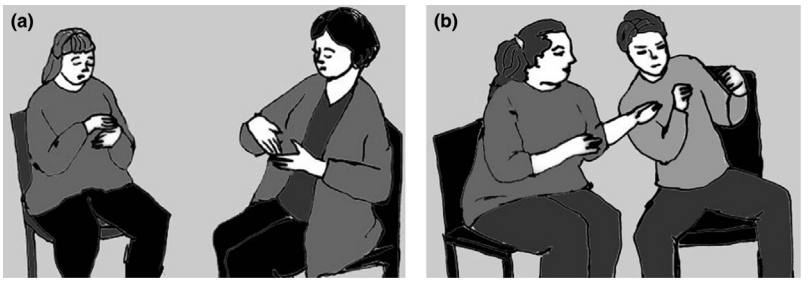

Figure 5 Embodiment in two bodies. (a) Mimicry by learner (on left). (b) Appropriation by learner (left). Computer art by Fey Parrill From Furuyama, Nobuhiro, 2000. Gestural interaction between the instructor and the learner in origami instruction. In: McNeill, D. (Ed.), Language and Gesture. Cambridge University Press, Cambridge, pp. 99–117. Used with permission of the University of Chicago Press.

he said "lo tenía *con un a marco redondo*" [lit. it he-was holding *with a frame round*]. (Thanks to Gale Stam for translation.) From a GP viewpoint, the speaker highlighted exactly the roundness of the plaque; this was the psychological predicate and GP in his immediate context of speaking. His forefingers, circling around the plaque shape, embodied the differentiation specifically of 'ROUND' in this field of meaningful oppositions (rather than the size, as in the first two-handed open-palm gesture) – something like THE PLAQUE'S RIM: ROUND. So, while the gesture looks like 'tracing,' the stroke was integrated with the synchronous speech at the point where they were co-expressive, differentiating a field of alternative possible rim-shapes. In this event, tracing fingers may not be 'tracing' (with this significance) at all but a contrast in imagery of rim-to-plane that was an integral part of speaking.

Some gestures do recreate actions; they take action stubs and ritualize them, much as Kendon described. These gestures are more than actions – they are a kind of performance, a replication of the action, and may also include posture, spatial location, as well as the manual action. These gestures are of two kinds. Some are pantomimes at their own locus on the Gesture Continuum. Others are 'character viewpoint' (C-VPT) gestures, gesticulations with the viewpoint of the character that is being recounted. And here again the difference of pantomimes from gesticulations applies; pantomime relates to speech loosely, gesticulation combines with it at the point of co-expressiveness. The C-VPTs are co-expressive with speech and their viewpoint is part of the semiotic with which they oppose it in a dialectic – a C-VPT is, among other things, *not* an 'observer viewpoint,' or O-VPT, a significance that it has but that is not part of a pantomime. Pantomime is simulated action and has no such contrastive value. In the 'it down' catchment example the speaker's hands were Tweety's hands, the space his space, her body his, and so forth – the speaker stepped into Tweety's shoes and added this to the utterance itself. It was part of a caused-motion unpacking ("… and $\varnothing_{tw}$ drops *it* (= a bowling ball) *do*wn the pipe"), making Tweety the agent. The gesture's C-VPT perspective was shaped by this caused-motion agent significance. The most telling difference from a replicated action is that, agent-wise, the hands thrust the ball down, in contrast to the deputation of this motion to gravity in the stimulus. All of this could happen because the gesture was a 'new' action, formed over the thought–language–hand link.

What looks to be an 'action–action' source of gesture is often really a metaphor in gesture form. The metaphor uses the action–action as its vehicle. The gesture embodies this vehicle but is not itself the action. It is an iconic image and in is fact two steps removed from the action. The metaphor itself is one step. Müller's (2008) awakening process adds another. (Müller views metaphor dynamically, as a *process* by which the speaker and her listener generate metaphoricity in the context of the speech event; clearly a conception in alignment with the arguments of this paper.) If there is an action source of the *metaphor* then that action can also be brought to the surface. The 'old' action circuit is activated along with the 'new' thought–language–hand circuit as a kind of meta-gesture comment. The question is whether this old action is necessarily active whenever the metaphoric gesture arises, and the answer is no. To see why, notice what happens to the metaphoric gesture when you do awaken the action. In the case of a rejection metaphor (the hand slicing side to side, palm forward, in typical form) it becomes a simulated removal of the imagined offending object (as if flicking a fly off with a finger, etc.). This action–action pantomime the metaphor co-opts into a new version of the metaphor. As a result, the action changes character; the opposite of the metaphor coming out the action. This is not to say that awakenings do not occur. They do but, to remain part of speaking rather than detaching from it, the meaning also changes; the pantomime becomes a meta-gestural layer commenting on the metaphor vehicle, as if saying that the metaphor is *decorated* with pantomime ("I am totally against your argument" plus flicking: the meta-comment of the awakened action is something like saying (if you dared), "your argument is puny, repellent and can be flicked away like a fly"). None of this shows that metaphoric gestures derive from action-actions. Quite the contrary, the gesture *recruits* the action as an ornament.

Liebal et al. (2010), following Kendon 2004, describe what they identify as 'gesture families,' which in their study are families of chimp gestures based on shared action sources – slapping, pushing, offering, etc. If these gestures are what they seem, miniaturized and ritualized editions of the source actions and not chimp metaphors, they do what human gestures do not, preserve in the gesture the original significance of the action, and would count therefore as yet another morsel of what, on comparison, arose from language evolution – new kinds of actions with their own significances that do not occur with the apes. Indeed, the key difference between ape emblems (if this is what they can be called) and human emblems is that those of apes are action stubs, while those of humans are metaphors – a precision metaphor within 'OK,' the image of a removal metaphor within 'negation,' and so forth. It all points to the mode of human cognitive action that evolution engendered.

Part of the scale, creativity and distinctiveness of human thought lies precisely in its freedom from pragmatic action constraints. When the hands make a gesture, it is thought that controls them and not a hidden action with purposes in relation to the physical world.

See also: Christianity in Latin America; Communicative Competence: Linguistic Aspects; Functional Approaches to Grammar; Language and Thought: The Neo-Whorfian Hypothesis; Mental Imagery: Visual Cognition; Piaget's Theory of Cognitive Development; Pragmatics, Linguistic; Psycholinguistics: Overview; Religious Organizations; Speech Production, Psychology of.

Bibliography

Alibali, Martha. W., Kita, Sotaro, Young, Amanda, 2000. Gesture and the process of speech production: we think, therefore we gesture. Language and Cognitive Processes 15, 593–613.

Bates, Elizabeth, Dick, Frederic, 2002. Language, gesture and the developing brain. In: Casey, B.J., Munakata, Y. (Eds.), Special Issue: Converging Method Approach to the Study of Developmental Science, vol. 40. Developmental Psychobiology, pp. 293–310.

Bavelas, Janet B., Coates, Linda, Johnson, Trudy, 2000. Listeners as co-narrators. Journal of Personality and Social Psychology 79, 941–952.

Beckman, Mary. E., Hirschberg, Julia, 1994. The ToBI Annotation Conventions. http://www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html (accessed 05.11.12.).

Bolinger, Dwight, 1986. Intonation and Its Parts: Melody in Spoken English. Stanford University Press, Stanford, CA.

Bremmer, Jan, Roodenburg, Herman, 1991. A Cultural History of Gesture. Cornell University Press, Ithaca, NY.

Bühler, Karl, 1982. The deictic field of language and deictic words. In: Jarvella, R.J., Klein, W. (Eds.), Speech, Place, and Action. John Wiley & Sons Ltd, Chichester, England, pp. 9–30.

Butterworth, Brian, Beattie, Geoffrey, 1978. Gesture and silence as indicators of planning in speech. In: Campbell, R.N., Smith, P. (Eds.), Recent Advances in the Psychology of Language: Formal and Experimental Approaches. Plenum Press, New York, pp. 347–360.

Casssell, Justine, McNeill, David, McCullough, Karl-Erik, 1999. Speech-gesture mismatches: evidence for one underlying representation of linguistic and nonlinguistic information. Pragmatics & Cognition 7, 1–34.

Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.

Cienki, Alan, Müller, Cornelia, 2008. Metaphor and Gesture. John Benjamins, Amsterdam.

Dennett, Daniel T., 1991. Consciousness Explained. Little, Brown & Company, Boston.

Dreyfus, H., 1994. Being-in-the-World: A Commentary on Heidegger's Being and Time, Division I. MIT Press, Cambridge, MA.

Efron, David, 1941. Gesture and Environment. King's Crown Press, Morningside Heights, NY.

Enfield, Nick J., 2001. 'Lip-pointing': a discussion of form and function with reference to data from Laos. Gesture 1, 185–221.

Firbas, Jan, 1971. On the concept of communicative dynamism in the theory of functional sentence perspective. Philologica Pragensia 8, 135–144.

Fischer-Lichte, Erika, 1992. The Semiotics of Theater (J. Gaines, D.L. Jones, Trans.). Indiana University Press, Bloomington, IN.

Furuyama, Nobuhiro, 2000. Gestural interaction between the instructor and the learner in origami instruction. In: McNeill, D. (Ed.), Language and Gesture. Cambridge University Press, Cambridge, pp. 99–117.

Gelb, A., Goldstein, K., 1925. Über Farbennamenamnesie. Psychologische Forschung 6, 127–186.

Goldin-Meadow, Susan, 2003. The Resilience of Language: What Gesture Creation in Deaf Children Can Tell Us about How All Children Learn Language. Psychology Press, New York.

Haviland, John, 2000. Pointing, gesture spaces, and mental maps. In: McNeill, D. (Ed.), Language and Gesture. Cambridge University Press, Cambridge, pp. 13–46.

Iverson, Jana M., Goldin-Meadow, Susan, 1998. Why people gesture when they speak. Nature 396, 228.

Jakobson, R., 1960. Concluding statement: Linguistics and poetics. In: Sebeok, T. (Ed.), Style in Language. MIT Press, pp. 350–377.

Kelly, Spencer D., Barr, Dale J., Church, R. Breckinridge, Lynch, Katheryn, 1999. Offering a hand to pragmatic understanding: the role of speech and gesture in comprehension and memory. Journal of Memory and Language 40, 577–592.

Kendon, Adam, 1972. Some relationships between body motion and speech. In: Siegman, A., Pope, B. (Eds.), Studies in Dyadic Communication. Pergamon Press, New York, pp. 177–210.

Kendon, Adam, 1980. Gesticulation and speech: two aspects of the process of utterance. In: Key, M.R. (Ed.), The Relationship of Verbal and Nonverbal Communication. Mouton and Co., The Hague, pp. 207–227.

Kendon, Adam, 1995. Gestures as illocutionary and discourse structure markers in southern Italian conversation. Journal of Pragmatics 23, 247–279.

Kendon, Adam, 2000. Gesture in Naples and Gesture in Classical Antiquity: A Translation of Andrea de Jorio's La mimica degli antichi investigata nel gestire napoletano (A. Kendon, Trans.). Indiana University Press, Bloomington, IN.

Kendon, Adam, 2004. Gesture: Visible Action as Utterance. Cambridge University Press, Cambridge.

Kendon, Adam, 25 July 2010. Accounting for Forelimb Actions as a Component of Utterance: An Evolutionary Approach. Plenary Lecture. International Society for Gesture Studies, Frankfurt/Oder.

Kimbara, I., 2006. On gestural mimicry. Gesture 6, 39–61.

Krauss, Robert M., Chen, Yihsiu, Gottesman, Rebecca F., 2000. Lexical gestures and lexical access: a process model. In: McNeill, D. (Ed.), Language and Gesture. Cambridge University Press, Cambridge, pp. 261–283.

Lakoff, George, Johnson, Mark, 1980. Metaphors We Live by. University of Chicago Press, Chicago.

LeBaron, Curtis, Streeck, Jürgen, 2000. Gestures, knowledge, and the world. In: McNeill, D. (Ed.), Language and Gesture. Cambridge University Press, Cambridge, pp. 118–138.

Liddell, Scott K., 2003. Grammar, Gesture, and Meaning in American Sign Language. Cambridge University Press, Cambridge.

Liebal, K., Bressem, J., Müller, C., 2010. Recurrent Forms and Contexts: Families of Gestures in Non-Human Primates. Conference of the International Society of Gesture Studies, Panel 13: Towards a Grammar of Gesture: Evolution, Brain and Linguistic Structures, Frankfurt/Oder, Germany.

Mayberry, Rachel, Jaques, Joselynne, 2000. Gesture production during stuttered speech: insights into the nature of gesture-speech integration. In: McNeill, D. (Ed.), Language and Gesture. Cambridge University Press, Cambridge, pp. 199–214.

McNeill, David, 1992. Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press, Chicago.

McNeill, David, 2005. Gesture and Thought. University of Chicago Press, Chicago.

McNeill, David, 2012. How Language Began: Gesture and Speech in Human Evolution. Cambridge University Press.

McNeill, David, Levy, Elena, 1982. Conceptual representations in language activity and gesture. In: Jarvella, R.J., Klein, W. (Eds.), Speech, Place, and Action. John Wiley & Sons Ltd, Chichester, UK, pp. 271–296.

McNeill, David, Duncan, Susan D., 2000. Growth points in thinking for speaking. In: McNeill, D. (Ed.), Gesture and Language. Cambridge University Press, Cambridge, pp. 141–161.

McNeill, D., Quek, F., McCullough, K.-E., Duncan, S., Furuyama, N., Bryll, R., Ma, X.-F., Ansari, R., 2001. Catchments, prosody and discourse. GESTURE 1, 9–33.

McNeill, David, Duncan, Susan, D., Cole, Jonathan, Gallagher, Shaun, Bertenthal, Bennett, 2008. Growth points from the very beginning. Interaction Studies 9 (1), 117–132 (Special Issue on Proto-Language, Bickerton, D., Arbib, M. (Eds.)).

Merleau-Ponty, Maurice, 1962. Phenomenology of Perception (C. Smith, Trans.). Routledge, London.

Morris, Desmond, Collett, Peter, Marsh, Peter, O'Shaughnessy, Marie, 1979. Gestures: Their Origins and Distribution. Stein & Day, New York.

Müller, Cornelia, 1998. Iconicity and gesture. In: Santi, S., Guaïtella, I., Cavé, C., Konopezynski, G. (Eds.), Oralité et Gestualité: Communication Multimodale, Interaction. L'Harmattan, Paris, pp. 321–328.

Müller, Cornelia, 2008. Metaphors, Dead and Alive, Sleeping and Waking: A Cognitive Approach to Metaphors in Language Use. University of Chicago Press, Chicago.

Nakatani, Christine H., Grosz, Barbara J., Ahn, David D., Hirschberg, Julia, 1995. Instructions for Annotating Discourses. Technical Report TR-21–95. Center for Research in Computer Technology, Harvard University, MA.

Özyürek, Asli, 2000. The influence of addressee location on spatial language and representational gestures of direction. In: McNeill, D. (Ed.), Language and Gesture. Cambridge University Press, Cambridge, pp. 64–83.

Quek, Francis, Ma, Xin-Feng, Bryll, Robert, 1999. A parallel algorithm for dynamic gesture tracking. In: ICCV'99 International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (RATFG-rts'99). Corfu, Greece, September 26–27, pp. 64–69. Available at: http://vislab.cs.wright.edu.

Rieber, Robert W., Carton, Aaron S. (Eds.), 1987. The Collected Works of L.S. Vygotsky. Problems of General Psychology. Including the Volume "Thinking and Speech". (Norris Minick, Intro. and Trans.), vol. 1. Plenum, New York.

Saussure, Ferdinand de, 1959. Course in General Linguistics (Charles Bally, Albert Sechehaye (Eds.), Wade Baskin, Trans.). The Philosophical Library, New York.

Streeck, Jürgen, 2010. Gesturecraft: The Manufacture of Meaning. John Benjamins, Amsterdam/Philadelphia.

Talmy, Leonard, 2000. Toward a Cognitive Semantics. The MIT Press, Cambridge, MA.

Vygotsky, Lev S., 1987. Thought and Language (Hanfmann, E., Vakar, G., Ed. and Trans. (A. Kozulin, Rev. and Ed.)). MIT Press, Cambridge.

Wachsmuth, Ipke, Lenzen, Manuela, Knoblich, Guenther (Eds.), 2008. Embodied Communication in Humans and Machines. Oxford University Press, Oxford.

Ghetto

Loic Wacquant, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the online update, pp. 1–7, © 2004, Elsevier Ltd.

Abstract

Although the social sciences have made extensive use of the term 'ghetto' as a descriptive term, they have failed to forge a robust analytical concept of the same, relying instead on the folk notions taken for granted at each epoch in the society under examination. This article constructs a relational concept of the ghetto as a Janus-faced instrument of ethnoracial closure and control by drawing on the historiography of the Jewish diaspora in Renaissance Europe, the sociology of the black American experience in the Fordist metropolis, and the anthropology of ethnic outcasts in East Asia. This reveals that a ghetto is a social-organizational device composed of four elements (stigma, constraint, spatial confinement, and institutional encasement) that employs space to reconcile the two antinomic purposes of economic exploitation and social ostr! acization. The ghetto is not a "natural area" coterminous with 'the history of migration' (as Louis Wirth argued), but a special form of collective violence concretized in urban space. Articulating the concept of ghetto makes it possible to disentangle the relationship between ghettoization, urban poverty, and segregation, and to clarify the structural and functional differences between ghettos and ethnic clusters. It also enables us to spotlight the role of the ghetto as symbolic incubator and matrix for the production of a spoiled identity, and suggests that it should be studied by analogy with other institutions for the forced confinement of dispossessed and dishonored groups such as the reservation, the refugee camp, and the prison.

It is a paradox that, while the social sciences have made extensive use of the 'ghetto' as a *descriptive term*, they have failed to forge a robust *analytical concept* of the same. In the historiography of the Jewish diaspora in early modern Europe and under Nazism, the sociology of the black American experience in the twentieth-century metropolis, and the anthropology of ethnic outcasts in Africa and East Asia, its three traditional domains of application, the term 'ghetto' variously denotes a bounded urban ward, a web of group-specific institutions, and a cultural and cognitive constellation (values, mindset, or mentality) entailing the sociomoral isolation of a stigmatized category as well as the systematic truncation of the life space and life chances of its members. But none of these strands of research has taken the trouble to specify what makes a ghetto *qua* social form, which of its features are constitutive and which derivative, as they have, at each epoch, taken for granted and adopted the *folk concept* extant in the society under examination – which explains that the notion, appearing self-evident, does not figure in most dictionaries of social science, including previous publications of this encyclopedia.

A Wooly and Shifting Notion

Thus the semantic range of the 'ghetto' in American society and social science, which has dominated inquiry into the topic both quantitatively and thematically, has successively expanded and contracted in keeping with how political and intellectual elites have viewed the vexed nexus of ethnicity and poverty in the city (Ward, 1989). At first, in the latter half of the nineteenth century, the term designated residential concentrations of European Jews in the Atlantic seaports and was clearly distinguished from the 'slum' as an area of housing blight and social pathology. It dilated during the Progressive era to encompass all inner-city districts wherein exotic newcomers gathered, namely lower-class immigrants from the Southeastern regions of Europe and African Americans fleeing the Jim Crow regime of caste subjugation in the US South. Expressing ruling-class worries over whether these groups could or should assimilate into the predominant Anglo-Saxon pattern of the country, the term then referred to the intersection between the ethnic neighborhood and the slum, where segregation combined with physical disrepair and overcrowding to exacerbate urban ills such as criminality, family breakdown, and pauperism, and thwart participation in national life. This conception was given scientific authority by the ecological paradigm of the Chicago school of sociology. In his classic book *The Ghetto*, Louis Wirth (1928: p. 6) assimilates to the Jewish ghetto of medieval Europe the 'Little Sicilies, Little Polands, Chinatowns, and Black Belts in our large cities,' along with the 'vice areas' hosting deviant types such as hobos, bohemians, and prostitutes – all of which are said to be 'natural areas' born of the universal desire of different groups to 'preserve their peculiar cultural forms' and each fulfilling a specialized 'function' in the broader urban organism.

The notion contracted after World War II under the press of the Civil Rights movement to signify mainly the compact and congested enclaves to which African Americans were forcibly relegated as they migrated into the industrial centers of the North. The growth of a 'Black Metropolis in the womb of the white,' wherein Negroes evolved distinct and parallel institutions to compensate for and shield themselves from unflinching exclusion by whites (Drake and Cayton, 1945), contrasted sharply with the smooth residential dispersal of European Americans of foreign stock. Writing at the acme of the black uprisings of the 1960s, Kenneth Clark (1965: p. 11) made this relationship of ethnoracial subordination epicentral to his dissection of the *Dark Ghetto* and its woes: 'America has contributed to the concept of the ghetto the restriction of persons to a special area and the limiting of their freedom of choice on the basis of skin color. The dark

ghetto's invisible walls have been erected by the white society, by those who have power.' This diagnosis was confirmed by the Kerner Commission (1968: p. 2), a bipartisan taskforce appointed by President Johnson whose official report on the 'civil disorders' that rocked the American metropolis famously warned that, due to white racial intransigence, America was 'moving toward two societies, one black, one white – separate and unequal.' But over the ensuing two decades, the dark ghetto collapsed and devolved into a barren territory of dread and dissolution due to deindustrialization and state policies of welfare reduction and urban retrenchment. And, as racial domination grew more diffuse and diffracted through a class prism, the category was displaced by the duet formed by the geographic euphemism of 'inner city' and the neologism of 'underclass,' defined as the substratum of ghetto residents plagued by antisocial behaviors, acute joblessness, and social isolation (Wilson, 1987). By the 1990s, the neutralization of the 'ghetto' in policy-oriented research culminated in the expurgation of all mention of race and power to redefine it as any tract of extreme poverty, irrespective of population and institutional makeup, in effect dissolving the ghetto back into the slum.

The extension of the term to the study of the distinctive sociocultural patterns elaborated by homosexuals in the cities of advanced societies 'in response to both stigma and gay liberation' after the Stonewall riots (Levine, 1979: p. 31) and its recent resurgence in Western Europe in heated scientific and political debates over the links between postcolonial immigration, postindustrial economic restructuring, and urban dualization (Mingione, 1996) would seem only to further muddle its meaning. Yet one can extract out of these varied literatures common threads and recurrent properties to *construct a relational concept* of the ghetto as an *instrument of closure and control* that clears up most of the confusion surrounding it and makes it a powerful tool for the social analysis of ethnoracial domination and urban inequality. For this it suffices to return to the historical inception of the word and of the phenomenon it depicted in Renaissance Venice.

A Janus-Faced Institution of Ethnic Closure and Control

Coined by derivation from the Italian *giudecca*, *borghetto* or *gietto* (or from the German *gitter* or the Talmudic Hebrew *get:* the etymology is disputed), the word 'ghetto' initially refers to the forced consignment of Jews to special districts by the city's political and religious authorities. In medieval Europe, Jews were commonly allotted quarters wherein they resided, administered their own affairs, and followed their customs. Such quarters were granted or sold as a privilege to attract them into the towns and principalities for which they fulfilled key roles as money-lenders, tax collectors, and longdistance tradesmen. But, between the thirteenth and the sixteenth century, in the wake of the upheavals caused by the Crusades, favor gradually turned into compulsion (Stow, 1992). In 1516, the Senate of Venice ordered all Jews rounded up into the *ghetto nuovo*, an abandoned foundry on an isolated island enclosed by two high walls whose outer windows and doors were sealed while watchmen stood guard on its two bridges and patrolled the adjacent canals by boat. Jews were henceforth allowed out to pursue their occupations by day, but they had to wear a distinctive garb and return inside the gates before sunset on pain of serious punishment. These measures were designed as an alternative to expulsion to enable the city-state to reap the economic benefits brought by the presence of Jews (including rents, special taxes, and forced levies) while protecting their Christian residents from contaminating contact with bodies perceived as unclean and dangerously sensual, carriers of syphillis and vectors of heresy, in addition to bearing the taint of money-making through usury, which the Catholic Church equated with prostitution (Sennett, 1994: p. 224).

As this Venetian model spread in cities throughout Europe and around the Mediterranean rim (Johnson, 1987: pp. 235–245), territorial fixation and seclusion led, on the one hand, to overcrowding, housing deterioration, and impoverishment as well as excess morbidity and mortality, and, on the other, to institutional flowering and cultural consolidation as urban Jews responded to multiplying civic and occupational restrictions by knitting a dense web of group-specific organizations that served as so many instruments of collective succor and solidarity, from markets and business associations, to charity and mutual aid societies, to places of religious worship and scholarship. The *Judenstadt* of Prague, Europe's largest ghetto in the eighteenth century, even had its own city hall, the *Rathaus*, an emblem of the relative autonomy and communal strength of its residents, and its synagogues were entrusted not only with the spiritual stewardship but also with the administrative and judicial oversight of its population. Social life in the Jewish ghetto was turned inward and verged 'on overorganization' (Wirth, 1928: p. 62), so that it reinforced both integration within and isolation from without.

One can detect in this inaugural moment the four constituent elements of the ghetto, viz., *stigma, constraint, spatial confinement, and institutional encasement*. The ghetto is a social-organizational device that employs space to reconcile two antinomic purposes: to maximize the material profits extracted out of a group deemed defiled and defiling and to minimize intimate contact with its members so as to avert the threat of symbolic corrosion and contagion they carry. This same dual rationale of *economic exploitation cum social ostracization* governed the genesis, structure, and functioning of the African-American ghetto in the Fordist metropolis during most of the twentieth century. Blacks were recruited into northern US cities after World War I because their unskilled labor was indispensable to the industries that formed the backbone of the expanding factory economy. But there was no question of them mixing and consorting with whites, who regarded them as inherently vile, congenitally inferior, and shorn of ethnic honor owing to the stain of slavery. As blacks moved in from the South in millions, white hostility increased and patterns of discrimination and segregation that had hitherto been informal and inconsistent hardened in housing, schooling, and public accommodations and were extended to the economy and polity (Spear, 1968; Osofsky, 1971). African Americans had no choice but to seek refuge inside the bounded perimeter of the Black Belt and to endeavor to develop in it a network of separate institutions to procure the basic needs of the

castaway community. Thus arose a parallel city anchored by black churches and newspapers, black block clubs and lodges, black schools and businesses, and black political and civic associations, nested at the core of the white metropolis yet sealed from it by an impassable fence built of custom, legal suasion, economic discrimination (by realtors, banks, and the state), and by violence manifested in the beatings, fire-bombings, and riots that checked those who dared stray across the color line.

This forced institutional parallelism predicated on envelopping and inflexible spatial seclusion – not extreme poverty, housing blight, cultural difference, or mere residential separation – is what has distinguished African Americans from every other group in US history, as noted by leading students of the black urban experience from W.E.B. Du Bois and E. Franklin Frazier to Drake and Cayton to Kenneth Clark and Oliver Cox (Wacquant, 1998). It also characterizes the trajectory of the Burakumin in the Japanese city after the close of the Tokugawa era (Hane, 1982). As the descendants of the *eta*, the lowest of the four castes forming the estate order of feudal Japan, the Burakumin were untouchables in the eyes of the Buddhist and Shinto religions, and they were legally confined from sundown to sunup in out-of-the-way hamlets *(buraku)*, obliged to wear a yellow collar and to walk barefoot, expected to drop on their hands and knees when addressing commoners, and restricted to marrying solely among themselves. Though they were officially emancipated in 1871, as they moved into cities they were funneled against their will into notorious neighborhoods near garbage dumps, crematoria, jails, and slaughterhouses, widely viewed as nests of criminality and immorality. There, they were barred from industrial employment and locked in low-paying and dirty jobs, sent to separate schools, and compelled to remain endogamous by the indelible taint of their blood as traced through 'family registration records' (DeVos and Wagatsuma, 1966). In the late 1970s, according to the Burakumin Defense League, they were estimated to number 3 million, trapped in 6,000 ghettos in some thousand cities across the main island.

Spread over three continents and five centuries, the Jewish, African-American, and Burakumin cases demonstrate that the ghetto is not, *pace* Wirth (1928: pp. 284–285), a 'natural area' arising via environmental adaptation governed by a biotic logic 'akin to the competitive cooperation that underlies the plant community.' The error of the early Chicago school here consisted in falsely 'converting history into natural history' and passing ghettoization as 'a manifestation of human nature' virtually coterminous with 'the history of migration' (Wirth, 1928: p. 285) when it is a highly peculiar form of urbanization warped by asymmetric relations of power between ethnoracial groupings: a special form of *collective violence concretized in urban space*. That ghettoization is *not* an 'uncontrolled and undesigned' process, as Robert E. Park asserted in his preface to *The Ghetto* (Wirth, 1928: p. viii), was especially visible after World War II when the black American ghetto was reconstructed from the top down through state policies of public housing, urban renewal, and suburban economic development intended to bolster the rigid separation of blacks from whites (Hirsch, 1983). It is even more glaring in the instance of the 'caste cities' built by colonial powers to inscribe in space the hierarchical ethnic organization of their overseas possessions, such as Rabat under French rule over Morocco and Cape Town after the passage of the Group Areas Acts under the apartheid regime of South Africa (Abu-Lughod, 1980, Western, 1981).

Recognizing that it is a product and instrument of group power makes it possible to appreciate that, in its full-fledged form, the ghetto is a *Janus-faced institution* as it serves opposite functions for the two collectives that it binds in a relation of asymmetric dependency. For the dominant category, its rationale is to *confine and control*, which translates into what Max Weber calls the 'exclusionary closure' of the dominated category. For the latter, however, it is an *integrative and protective device* insofar as it relieves its members from constant contact with the dominant and fosters consociation and community building within the constricted sphere of intercourse that it creates. Enforced isolation from the outside leads to the intensification of social exchange and cultural sharing inside. Ghettos are the product of a mobile and tensionful dialectic of external hostility and internal affinity that expresses itself as ambivalence at the level of collective consciousness. Thus, although European Jews consistently protested relegation within their outcast districts, they were also deeply attached to them and appreciative of the relative security they afforded and the special forms of collective life they supported: Frankfurt's ghetto in the eighteenth century was 'not just the scene of confinement and persecution but a place where Jews were entirely, supremely, at home' (Gay, 1992: p. 67). Similarly, black Americans took pride in having 'erected a community in their own image,' even as they resented the fact that they had done so under duress, as a result of unyielding white exclusion aimed at warding off the specter of 'social equality,' that is, sexual mixing (Drake and Cayton, 1945: p. 115).

Disentangling Poverty, Segregation, and Ethnic Clustering

Articulating the concept of ghetto makes it possible to disentangle the relationship between ghettoization, urban poverty, and segregation, and thence to clarify the structural and functional differences between ghettos and ethnic neighborhoods. It also enables us to spotlight on the role of the ghetto as a symbolic incubator and matrix for the production of a spoiled identity.

1. *Poverty is a frequent but derivative and variable characteristic of ghettos:* the fact that most ghettos have historically been places of endemic and often acute misery owing to the paucity of space, the density of settlement, and the economic exploitation and generalized maltreatment of their residents does not imply that a ghetto is necessarily a place of destitution, nor that it is uniformly deprived. The *Judengasse* of Frankfurt, instituted in 1490 and abolished in 1811, went through periods of prosperity no less than penury and it contained patches of extraordinary opulence as court Jews helped the city become a vibrant center of trade and finance – part of its glamor to this day comes from it being the ancestral home of the Rothschild dynasty (Wirth, 1928: Chap. 4). James Weldon Johnson (1930: p. 4)

insisted that the Harlem of the 1930s was 'not a slum or a fringe' but the 'cultural capital' of black America, where 'the Negro's advantages and opportunities are greater than in any other place in the country.' Likewise, Chicago's 'Bronzeville' was far more prosperous at mid-century than southern black communities and harbored the largest and most affluent African-American bourgeoisie of its era (Drake and Cayton, 1945). Whether a ghetto is poor or not depends on extraneous factors such as demography, ecology, state policies, and the shape of the surrounding economy.

Conversely, not all dispossessed and dilapidated urban districts are ghettos. Declining white neighborhoods in the deindustrializing cities of the US Midwest and British Midlands, depressed rural towns of the former East Germany and southern Italy, and the disreputable *villas miserias* of the greater Buenos Aires at the close of the twentieth century are territories of working-class demotion and decomposition, not ethnic containers dedicated to maintaining an outcast group in a relationship of seclusive subordination. They are not ghettos other than in a metaphorical sense, no matter how impoverished – if extreme rates of poverty sufficed to make a ghetto, then large chunks of the former Soviet Union and most third-world cities would be gargantuan ghettos. The *favelas* of the Brazilian metropolis often portrayed as segregated dens of dereliction and disorganization turn out to be working-class wards with finely stratified webs of ties to industry and to the wealthy districts for which they supply household service labor. As in the *ranchos* of Venezuela and the *poblaciones* of Chile, the families that dwell in these squatter settlements span the color continuum and have extensive genealogical bonds to higher-income households; they are 'not socially and culturally marginal, but stigmatized and excluded from a closed class system' (Perlman, 1976: p. 195; also Quijano, 1968). Given that not all ghettos are poor and not all poor areas are ghettos, one cannot collapse and confound the analysis of ghettoization with the study of slums and lower-class districts in the city.

2. Similarly, *all ghettos are segregated but not all segregated areas are ghettos.* The select boroughs of the West of Paris, the exclusive upper-class suburbs of Boston or Berlin, and the 'gated communities' that have mushroomed in global cities such as São Paulo, Toronto, and Miami are monotonous in terms of wealth, income, occupation, and often ethnicity, but they are not for all that ghettos. Segregation in them is entirely voluntary and elective, and for that reason it is neither all-inclusive nor perpetual. Fortified enclaves of luxury package 'security, seclusion, social homogeneity, amenities, and services' to enable bourgeois families to escape what they perceive as 'the chaos, dirt, and danger of the city' (Caldeira, 2000: pp. 264–265). These islands of privilege serve to enhance, not curtail, the life chances and protect the life-styles of their residents, and they radiate a positive aura of distinction, not a sense of infamy and dread.

   This suggests that residential segregation is a necessary but not a sufficient condition for ghettoization. For a ghetto to emerge spatial confinement must, first, be imposed and all-encompassing and, second, it must be overlayed with a distinct and duplicative set of institutions enabling the group thus cloistered to reproduce itself within its assigned perimeter. If blacks are the only ethnic group to be 'hyper-segregated' in American society (Massey and Denton, 1992), it is because they are the only community to have combined involuntary segregation with organizational parallelism entrapping them in a separate and inferior social cosmos of their own, which in turn bolstered their residential isolation. That even involuntary segregation at the bottom of the urban order does not *eo ipso* produce ghettos is demonstrated by the fate of the declining French *banlieues* after the 1980s. Although they have been widely disparaged as 'ghettos' in public discourse and their inhabitants share a vivid feeling of being cast out in a 'penalized space' suffused with boredom, anguish, and despair (Pétonnet, 1982), relegation in these depressed concentrations of public housing at the urban periphery is based on class, not ethnicity; as a result they are culturally heterogeneous, typically harboring native French families along with immigrants from two dozen nationalities; and their inhabitants suffer not from institutional duplication but, on the contrary, from the lack of an ingrown organizational structure capable of sustaining them in the absence of gainful employment and adequate public services. Like the British or Dutch inner cities and the immigrant clusters of urban Germany or Italy, the French *banlieues* are, sociologically speaking, *anti-ghettos* (Wacquant to be published).

3. *Ghettos and ethnic neighborhoods have divergent structures and opposite functions:* moving beyond a gradational perspective to scrutinize the peculiar patterning of social relations within the ghetto as well as between it and the surrounding city throws into sharp relief the differences between the ghetto and the ethnic clusters or immigrant neighborhoods such as newcomers to the metropolis have formed in countless countries. The foreign 'colonies' of interwar Chicago that Robert Park, Ernest Burgess, and Louis Wirth – and after them the liberal tradition of assimilationist sociology and historiography – mistook for so many white 'ghettos' were scattered and mobile constellations born of cultural affinity and occupational concentration. Segregation in them was partial and porous, a product of immigrant solidarity and ethnic attraction instead of being imposed by outgroup hostility. Consequently, residential separation was neither uniformly nor rigidly visited upon these groups: in 1930, when the all-black Bronzeville harbored 92% of the city's Afro-American population, Chicago's Little Ireland was 'an ethnic hodge-podge' of 25 nationalities composed of only one-third Irish persons and containing a paltry 3% of the city's denizens of Irish ancestry (Philpott, 1978: pp. 141–145).

   What is more, the distinctive institutions of European immigrant enclaves were turned outward: they operated to facilitate adjustment to the novel environment of the US metropolis. They neither replicated the organizations of the country of origin nor perpetuated social isolation and cultural distinctiveness, and so they typically waned within two generations as their users gained access to their American counterparts and climbed up the class order and the corresponding ladder of places (Nelli, 1970; a similar process of spatial diffusion via class incorporation of Belgian, Italian,

Polish, and Iberian immigrants in the French industrial city is reported by Noiriel, 1988), all of which is in sharp contrast with the immutable racial exclusivity and enduring institutional alterity of the Black Belt. This Chicago illustration dramatizes the fact that immigrant neighborhood and ghetto serve diametrically opposed functions: one is a springboard for *assimilation* via cultural learning and social-cum-spatial mobility, and the other a material and symbolic isolation ward geared toward *dissimilation*. The former is best figured by a bridge, and the latter by a wall.

A Tainted Identity Machine

The ghetto is not only the concrete means and materialization of ethnoracial domination through the spatial segmentation of the city but also a potent *collective identity machine* in its own right, for it helps incrustate and elaborate the very division of which it is the expression in two complementary and mutually reinforcing ways. First, the ghetto sharpens the boundary between the outcast category and the surrounding population by deepening the sociocultural chasm between them: it renders its residents objectively and subjectively more dissimilar from other urban dwellers by submitting them to unique conditionings, so that the patterns of cognition and conduct they fashion have every chance of being perceived by outsiders as singular, exotic, even aberrant (Sennett, 1994: p. 244; Wilson, 1987: pp. 7–8), which feeds prejudicial beliefs about them. Second, the ghetto is a cultural combustion engine that melts divisions among the confined group and fuels its collective pride even as it entrenches the stigma that hovers over it. Spatial and institutional entrapment deflect class differences and corrode cultural distinctions within the relegated ethnoracial category. Thus Christian ostracism welded Ashkenazic and Sephardic Jews under an overarching Jewish identity such that they evolved a common 'social type' and 'state of mind' across the ghettos of Europe (Wirth, 1928: pp. 71–88, 1964). America's dark ghetto similarly accelerated the sociosymbolic amalgamation of mulattos and Negroes into a single 'race' and turned racial consciousness into a mass phenomenon fueling community mobilization against continued caste exclusion (Drake and Cayton, 1945: p. 390).

Yet this unified identity cannot but be stamped with ambivalence as it remains tainted by the very fact that ghettoization proclaims what Weber calls the 'negative evaluation of honor' assigned to the group confined. It is therefore wont to foster among its members sentiments of self-doubt and self-hatred, dissimulation of one's origin through 'passing,' the pernicious derogation of one's kind, and even fantasical identification with the dominant (Clark, 1965: pp. 63–67). And, because ghettoization is typically bound up tightly with ethnicity, segregation, and poverty, it is difficult to discern empirically which of the properties exhibited by ghetto dwellers are 'ghetto-specific cultural traits' as opposed to properties expressive of class, community, or masculinity (Hannerz, 1969: p. 79). Also, cultural forms forged in the ghetto seep across its boundaries and circulate through the surrounding society where they often become outward signs of cultural rebelliousness and social eccentricity – as indicated by the fascination of bourgeois teenagers around the world for black American 'gangster rap.' This makes it difficult to distinguish between cultural forms effectively in currency among ghetto residents and the public imagery of them diffused in the broader society (including via scholarly writings).

It is fruitful to think of ghetto and ethnic cluster as *two ideal typical configurations situated at opposite ends* of a continuum along which different groups can be located or travel over time depending on the intensity with which the forces of stigma, constraint, spatial confinement, and institutional duplication and completeness coalesce with each other and impinge upon them. Ghettoization can then be turned into a *multilevel variable* for comparative analysis and empirical specification. It can become attenuated to the point where, through gradual erosion of its spatial, social, and mental boundaries, the ghetto devolves into an elective ethnic concentration operating as a springboard for structural integration and/or cultural assimilation into the broader social formation. This describes well the trajectory of the Chinatowns of the United States from the early to the late twentieth century (Zhou, 1992) and the status of the Cuban immigrant enclave of Miami that fostered integration through biculturalism after the Mariel exodus of 1980 (Portes and Stepick, 1993). It also characterizes the 'Kimchee Towns' in which Koreans have converged in the metropolitan areas of Japan, which sport a blend of features making them a hybrid between ghetto and ethnic cluster (DeVos and Chung, 1981): they are places of infamy that first arose through enmity and constraint, but over the years their population has become ethnically mixed and they have enabled Koreans to socialize and intermarry with Japanese neighbors as well as obtain Japanese citizenship through naturalization. This schema also fits the so-called gay ghetto, which is more aptly characterized as a 'quasi-ethnic community' since 'most gay persons can "pass" and need not be confined to interacting with their "own kind" and none are forced to reside in the areas of visible concentration of gay institutions (Murray, 1979: p. 169).

The double-sidedness of the ghetto as a weapon and shield implies that, to the degree that its institutional completeness and autonomy are abridged, its protective role for the subordinate group is diminished and risks being swamped by its exclusionary modality. In cases where its residents cease to be of economic value to the dominant group, ethnoracial encapsulation can escalate to the point where the ghetto serves as an apparatus merely to warehouse the spoiled group or to prepare it for the ultimate form of ostracization, i.e., physical annihilation. The first scenario fits the evolution of the black American 'hyperghetto' in the post-Civil Rights era: having lost its function of reservoir of unskilled labor power, it has become symbolically linked to the hypertrophied carceral system of the United States by a triple relationship of structural homology, functional surrogacy, and cultural fusion (Wacquant, 2004). The second scenario is that implemented by Nazi Germany, which revived the *Judenghetto* between 1939 and 1944, first, to impoverish and concentrate Jews with a view toward relocation and later, when mass deportation turned out to be impractical, to funnel them toward extermination camps (Friedman, 1980, Browning 1986).

The unchecked intensification of its exclusionary thrust suggests that the ghetto might be most profitably studied not by analogy with urban slums, lower-class neighborhoods, and immigrant enclaves but alongside the reservation, the refugee camp, and the prison, as belonging to a broader class of institutions for the forced confinement of dispossessed and dishonored groups. It is not by happenstance that the Bridewell of London (1555), the *Zuchthaus* of Amsterdam (1654), and the *Hospital general* of Paris (1656), designed to instill the discipline of wage work to able-bodied vagrants, beggars, and criminals via incarceration, were invented around the same time as the Jewish ghetto. And that today's sprawling refugee camps in Rwanda, Sri Lanka, and in the occupied territories of Palestine look ever more like a cross between the ghettos of late medieval Europe and gigantic gulags.

Bibliography

Abu-Lughod, J.L., 1980. Rabat: Urban Apartheid in Morocco. Princeton University Press, Princeton, NJ.

Browning, C.R., 1986. Nazi ghettoization policy in Poland, 1939–1941. Central European History 19 (4), 343–368.

Caldeira, T., 2000. City of Walls: Crime, Segregation and Citizenship in Sâo Paulo. University of California Press, Berkeley, CA.

Clark, K.B., 1965. Dark Ghetto: Dilemmas of Social Power. Harper, New York.

DeVos, G., Chung, D., 1981. Community life in a Korean ghetto. In: Lee, C., De Vos, G. (Eds.), Koreans in Japan: Ethnic Conflict and Accomodation, University of California Press, Berkeley, CA, Chapter 10, pp. 225–251.

DeVos, G., Wagatsuma, H. (Eds.), 1966. Japan's Invisible Race: Caste in Culture and Personality. University of California Press, Berkeley, CA.

Drake St, C., Cayton, H.R., 1993. Black Metropolis: A Study of Negro Life in a Northern City [1945]. University of Chicago Press, Chicago.

Friedman, P., 1980. The Jewish ghettos of the Nazi era. In: Roads to Extinction: Essays on the Holocaust. The Jewish Publication Society of America, New York, Chapter 3, pp. 59–87.

Gay, R., 1992. The Jews of Germany: A Historical Portrait. Yale University Press, New Haven, CT.

Hane, M., 1982. Peasants, Rebels, and Outcastes: The Underside of Modern Japan. Pantheon, New York.

Hannerz, U., 1969. Soulside: Inquiries into Ghetto Culture and Community. Columbia University Press, New York.

Hirsch, A., 1983. Making the Second Ghetto: Race and Housing in Chicago 1940–1970. Cambridge University Press, Cambridge. repub. University of Chicago Press.

Johnson, J.W., 1981[1930]. Black Manhattan. Da Capo, New York.

Johnson, P., 1987. Ghetto. In: A History of the Jews. HarperPer-ennial, New York, pp. 230–310.

Kerner Commission 1989[1968]. The Kerner Report. The 1968 Report of the National Advisory Commission on Civil Disorders, Pantheon, New York.

Levine, M.P., 1979. Gay Ghetto. Journal of Homosexuality 4 (4) (Summer). Reprinted in expanded form as 'YMCA': The Social Organization of Gay Male Life. In: Gay Macho: The Life and Death of the Homosexual Clone. New York University Press, New York, pp. 30–54.

Massey, D., Denton, N., 1992. American Apartheid: Segregation and the Making of the Under! class. Harvard University Press, Cambridge.

Mingione, E. (Ed.), 1996. Urban Poverty and the 'Underclass'. Basil Blackwell, Oxford.

Murray, S.O., 1979. The institutional elaboration of a quasi-ethnic community. International Review of Modern Sociology 9 (July), 165–177.

Nelli, H.S., 1970. Italians in Chicago: A Study in Ethnic Mobility. Oxford University Press, New York.

Noiriel, G., 1988. Le Creuset francais. Editions du Seuil, Paris.

Osofsky, G., 1971. Harlem: The Making of a Ghetto—Negro New York, 1890–1930, second ed. Harper and Row, New York.

Perlman, J., 1976. The Myth of Marginality: Urban Poverty and Politics in Rio De Janeiro. University of California Press, Berkeley, CA.

Pétonnet, C., 1982. Espaces habités. Ethnologie des banlieues. Galilée, Paris.

Philpott, T.L., 1978. The Slum and the Ghetto: Neighborhood Deterioration and Middle-Class Reform, Chicago 1880–1930. Oxford University Press, New York.

Portes, A., Stepick, A., 1993. City on the Edge: The Transformation of Miami. University of California Press, Berkeley, CA.

Quijano, A., 1968. Notas sobre el concepto da marginalidad social. Commission for Latin American Report, Santiago (Chile).

Sennett, R., 1994. Fear of touching. In: Flesh and Stone: The Body and the City in Western Civilization. W.W. Norton, New York, Chapter 7, pp. 212–51.

Spear, A.H., 1968. Black Chicago: The Making of a Negro Ghetto, 1890–1920. University of Chicago Press, Chicago, IL.

Stow, K.R., 1992. Alienated Minority: The Jews of Medieval Europe. Harvard University Press, Cambridge, MA.

Wacquant, L., 1998. A black city within the white: revisiting America's dark ghetto. Black Renaissance 2 (1), 141–151.

Wacquant, L., 2004. Deadly Symbiosis: Race and The Rise of Neoliberal Penality. Polity Press, Cambridge.

Wacquant, L. Urban Outcasts: Toward a Sociology of Advanced Marginality. Polity Press, Cambridge (in press).

Ward, D., 1989. Poverty, Ethnicity, and the American City 1840–1925. Cambridge University Press, Cambridge.

Western, J., 1981. Outcast Cape Town. University of Minnesota Press, Minneapolis, MN.

Wilson, J., 1987. The Truly Disadvantaged: The Inner City, The Underclass and Public Policy. University of Chicago Press, Chicago.

Wirth, L., 1928. The Ghetto. University of Chicago Press, Chicago.

Wirth, L., 1964. The ghetto. In: Reiss , Jr., A.J. (Ed.), On Cities and Social Life. The University of Chicago Press, Chicago [1956], Chapter 5, pp. 84–98.

Zhou, M., 1992. Chinatown: The Socioeonomic Potential of an Urban Enclave. Temple University Press, Philadelphia, PA.

Gibson's Theory of Perceptual Learning

Karen E Adolph and Kari S Kretch, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes the key ideas of the influential psychologist Eleanor J. Gibson, developed over 70 years of research with infants, children, adults, and a wide range of nonhuman species. Gibson's ecological approach to perceptual learning and development describes how perception – extracting meaningful information from the environment to guide actions adaptively – improves with experience, the acquisition of new means of exploration, and the development of new perception–action systems.

Learning to Perceive or Perceiving to Learn?

Eleanor Jack Gibson built her theory of perceptual learning over a 70-year research career. She published her first paper on perceptual learning in 1932 (Gibson et al., 1932) and her last book in 2002 (Gibson, 2002). There is a clear thread from beginning to end, but she was not dogmatic in her ideas; her theories were always informed by data, and data collection was often inspired by real life and serendipity. Her theory is consistent with but not identical to James Gibson's (1979) ecological approach to perception. Although the Gibsons were married and shared many arguments and ideas about perceptual learning and development, they wrote only five articles together (Gibson, 2002). Thus, in this article, 'Gibson' refers to Eleanor Gibson unless otherwise noted. Gibson's 1969 book, *Principles of Perceptual Learning and Development*, described her theory in detail and jump-started a new field of inquiry. However, as new methods and findings became available, notably new ways of studying perception in young infants, Gibson questioned things that she had once taken for granted and broadened her domain of inquiry. In later writings, Gibson critiqued her 1969 account for failing to capture how infants learn to detect the perceptual information for guiding action adaptively as their perceptual-motor systems are developing. In fact, she considered infancy to be the perfect place to study perceptual learning (Gibson, 1992) and she was instrumental in building the field of infant perception (Pick, 1992).

On Gibson's (1969) account, perceptual learning entails an increased ability to extract relevant information from a stimulus array as the result of experience. The traditional view of perceptual learning, dating back to Bishop Berkeley in the 1700s, is that animals must learn to perceive; the information at sensory receptors is impoverished and meaningless and thus a complete percept requires learning. In Gibson's view, the information at receptors is sufficient to support complete percepts from the start, and thus animals need not *learn to perceive*; rather, they *perceive to learn* (Gibson, July 1989). Perceptual learning is the key to knowledge and where it all begins.

Gibson (1992) maintained that a theory of perceptual learning must answer basic questions: What is learned and what is the function? What instigates learning and what terminates the process? As for the question of mechanism, Gibson held that meaningful explanations of psychological processes must be at the level of behavior. She always maintained that the job of a perception psychologist is to describe and explain perception at the level of individual behavior, not in terms of the underlying neurophysiology (Gibson and Pick, 2000). Perceptual processes should be consistent with what we know about neural mechanisms and physiology, but these facts should not drive research into perceptual processes. Rather, the appropriate research program should begin with a consideration of what there is to be perceived in an ecological context – what animals need to perceive so as to act adaptively in the natural environments in which they evolved and in the changing environments in which they develop.

Starting Assumptions

Gibson's theory grows out of her commitment to several related starting assumptions that guided her thoughts about perceptual learning and shaped the trajectory of her research career. She believed that J.J. Gibson's (1966, 1979) *ecological approach* to perception is the proper starting place for a theory of perceptual learning and development (Gibson and Pick, 2000). Animals must be considered doing the things that they naturally do in their particular ecological niche. Their activities are guided by perceptual information and such information is available in the ambient arrays of energy to which their perceptual systems are sensitive. Because perception involves the whole animal, nested in its environment, her approach was necessarily a *systems approach*. The total system includes animals in the environment that surrounds them in the here and now and which has shaped them in the course of evolution and development, and which they in turn influence (Gibson, 1992).

Because of her focus on animals in their species-typical environments, hers was a *comparative approach*. In contrast to the traditional practice of focusing nearly exclusively on humans and the chambered eye, she maintained that perception psychologists should consider the perceptual systems and actions of varied species as they build theories of perceptual learning and development in humans (Gibson and Walk, 1960; Walk and Gibson, 1961). All animals have solved the problem of using perceptual information to guide activity: houseflies with compound eyes, fish using

electrolocation, bats using echolocation, and so on. Studying such diversity of form and function will lead to general principles of perception and its development. Evolutionary and developmental pressures are integral to a comparative approach, and perceptual learning – like all behavior – can only be understood in the context of development (Gibson, 1984b). Thus, her theory also begins with a *developmental approach*.

Above all, hers was a *functional approach* to perception and perceptual development (Gibson, 1982). A functional view considers the purposes of perception in everyday activity and over the evolutionary history of the species. Perception does not exist to discriminate Gabor patches or to support other reified laboratory phenomena. The ultimate question for perception researchers should be "how [animals] perceive what is going on around them so as to make good use of what the world offers" (Gibson, 1997: p. 42).

Perceptual Information

The ecological approach differs from traditional theories of perception in the conceptualization of perceptual information (Gibson and Pick, 2000). In the Gibsons' view, the ambient arrays of energy surrounding the observer – light, sound waves, patterns of pressure on tactile receptors, and so on – are structured by the objects and surfaces in the environment in ways that specify those objects and surfaces; thus, information arrives at sensory receptors already richly imbued with structure. This structure is not carried in a static image; it is only apparent in relations that emerge over transformations in space and time (movement of objects or the observer, edges, gradients, flow, etc.). Structured arrays of energy contain the information through which perception of the self and environment occurs (Gibson, 1970). If one only considers rays of light or sound vibrations, specification is not possible.

The task for the perceiver is not to add structure to impoverished sensory stimulation, but to detect the structure that already exists. We do not perceive ambient energy arrays and we do not perceive information. Rather, by detecting information in ambient arrays, we perceive the things in the world that structured the information and that are specified by the information (Gibson and Spelke, 1983).

Starting with this view of perceptual information, it might seem as if the *ecological approach* is at odds with a theory of learning. If information is already structured to specify things in the world, what is there to learn? But Gibson did not see a tension between the two. The organism must learn through experience how to generate and detect the appropriate perceptual information.

Animal–Environment Reciprocity

Animals and their surrounding environment constitute an interactive system, with each constituent reciprocal to the other (Gibson, 1997). Animals generate information about the environment and must tailor their actions to the environment. Reciprocally, the environment provides the animal with "opportunities and resources for action" (Gibson and Pick, 2000: p. 14) and with information that specifies those opportunities and resources. An animal fits into its ecological and developmental niche. The animal is scaled to the environment, as the environment is scaled to it. Muscle fibers and neurons exist, as do light waves and sound waves. But descriptions in such terms do not provide a meaningful link between animals and their environments (Gibson, 1984b). Physicists describe the environment in terms of molecules and galaxies, events in terms of nanoseconds and light years. But the psychologist's description should be on the scale of the animal concerned and the niche it occupies. The appropriate description is an ecological one: The animal's body, capabilities, and propensities must be described relative to the environment; reciprocally, the properties and features of the environment must be described relative to the animal. For example, "a cliff is a drop-off that is large relative to the size of the animal, and a step is a drop-off that is small relative to its size. A falling-off edge is dangerous, but a stepping-down edge is not" (Gibson, 1979: p. 157). Perception takes animal–environment reciprocity into account because to perceive the world is to coperceive the self (Gibson, 1991).

Perception–Action Reciprocity

A corollary of animal–environment reciprocity is perception–action reciprocity. Traditionally, researchers study perception by projecting displays onto observers' eyes or into their ears. But on Gibson's (1988a) account, we do not just see, we look. We do not just hear, we listen. Perceiving is an active process. The visual system, for example, is a motor system as well as a sensory one. It involves more than the retina or even the moving eye. Looking involves the eyes in the head on a mobile body. Perceivers seek information and optimize it rather than passively receiving it. The animal must do something to obtain information – scan the scene, turn the head, palpate the surface, kick the tires. Learning what to do – how to gather perceptual information – is a major part of perceptual learning. Because perceiving goes on over time, it can correct itself in real time and improve over development.

What is all this perceptual activity for? Perception functions to gather information about objects, events, and places and what the perceiver can do about these things (Gibson, 1988b). Perception guides action in a changeable environment. Thus, perception and action operate as a continuous cycle, whereby "perception obtains information for action, and action has consequences that inform perception about both the organism itself and the events that it perpetrates" (Gibson, 1997: p. 25). All actions – walking, cooking, typing, driving a car – must be guided by perceptual information. Reciprocally, action turns up perceptual information about the environment, what the animal is doing, and what it might do next. Gibson (1997: p. 25) differentiated between exploratory action ("foraging for information") and performatory action ("controlling environmental consequences"). But both types of action generate perceptual information, and both are guided by perceptual information.

Learning in Development

Gibson (1994) wrote, "There is just one way to understand behavior, and that is to take a developmental approach. There

is no typical or standard moment of maturity; besides, we gain our understanding from change and becoming" (p. 71). Development, for Gibson (1970), included changes over evolution and in the individual's lifetime. Although fashionable long before she arrived on the scene, and still considered central to current work, Gibson (1994) eschewed the nature–nurture dichotomy. She referred to it as a "hob-goblin that has haunted us" in developmental science (p. 71).

Development involves changes in animals' bodies, perceptual sensitivity, action capabilities, and environments. The developmental course of these changes creates changes in the perception–action loop, and thereby constrains or facilitates perceptual learning and development (Gibson and Pick, 2000). In particular, new action capabilities bring about new ways of generating information and also alter the relevance of the information generated (Gibson, 1988a, 1991, 1992). Recent work on relations between action development and perceptual learning back this up. Objects, for example, take on special relevance once infants acquire manipulatory skills. Accordingly, infants exhibit sensitivity to the three-dimensional form of objects after they can produce coordinated visual–manual exploration, which in turn, depends on the ability to sit independently (Soska et al., 2010).

What Is Learned

Gibson (1970, 1984b, 1992) assumed that a theory of perceptual learning must start with a functional analysis of what is out there to be perceived and learned – the general tasks of perceptual systems across development and species. Only then can researchers tackle subsequent questions regarding processes of perceptual learning. Gibson's focus regarding the content of perceptual learning changed over the course of her career. Her earlier work, influenced by her studies of reading, focused on learning distinctive features through higher order invariant relations (Gibson, 1969). She later asserted that she had been "premature" in concluding that "the principal thing learned in perceptual learning was distinctive features of objects" (Gibson, 1991: p. 394); "the old mistake was to start with static displays in formulating a theory of perceptual learning" (p. 615). Her later work, influenced by her studies of infant action, emphasized learning to perceive affordances and events (Gibson and Pick, 2000). This learning is more dynamic and fundamental.

Affordances

Above all, what animals learn to perceive are affordances for action (Gibson, 1980, 1982, 1992, 2000). *Affordances* are possibilities for action, what the environment offers the animal (Gibson, 1979). Possibilities for action depend on the fit between the animal's bodily capabilities and the physical properties of the environment. For example, a vertical wall affords walking for a housefly but not for a human; a 200-lb dumbbell affords lifting for a bodybuilder but not for a kindergartener. The animal–environment relations exist objectively in the interface between self and world. Affordances thus are real, regardless of whether they are perceived or used. Affordances reflect both animal–environment reciprocity and perception–action reciprocity because affordances must be perceived, perception must guide action, and actions are implicit in affordances. In Gibson's (1982) view, "we do not perceive stimuli or retinal images or sensations or even just things; what we perceive are things that we can eat, or write with, or sit down on, or talk to" (p. 60). We do not only perceive the size or distance of objects. We perceive whether something is within arms' reach and whether it will fit into our grasp. What we perceive are the functional relations between self and world (Gibson, 1980).

Perceptual development is a process of learning about affordances, becoming better able to detect appropriate supports and resources, and discovering new affordances as action capabilities change (Gibson, 1992). Affordances themselves develop. The acquisition of new motor skills – looking, reaching, walking, weight lifting, swimming, driving, sewing, hand writing – produces new affordances to be learned throughout life.

Perceiving affordances is essential for survival; it is what perceptual systems evolved to do. Certainly the best-known empirical example of animals perceiving affordances is Gibson and Walk's (1960) classic work using a 'visual cliff.' The visual cliff studies were Gibson's first foray into work with human infants and the first experiment to show the close tie between perception and action in the human infant. This was an experiment on perceived affordances for action, although she did not conceive it as such at the time (Gibson, 1991). All animals tested, including rats, cats, chickens, turtles, monkeys, and human infants, readily cross a visible surface that affords locomotion (the 'shallow side') and refuse to cross an apparent drop-off that does not (the 'deep side'). Later work inspired by these classic studies revealed that, for some species including cats and humans, a period of self-produced locomotor experience is necessary to learn to perceive affordances and thereby produce avoidance of a drop-off (Adolph, 2000; Bertenthal et al., 1984; Held and Hein, 1963; Kretch and Adolph, 2013).

Events, Invariants, and Multimodal Information

According to Gibson (1969), perception is a process of information pickup that happens over time. *Events* therefore are the primary source of perceptual information and a critical component of what is learned during perceptual learning and development. "Happenings over time" (Gibson, 1969: p. 16) include events perpetrated by the perceiver (throwing a ball), events that will soon involve the perceiver (an approaching ball to be caught or dodged), and events that are merely observed and external to the perceiver (watching others throw and catch). Infants are sensitive to all these types of events (Gibson and Pick, 2000).

Many critical sources of information only emerge in the context of an event (Gibson, 1979). Optic flow and motion parallax, for example, emerge as an animal moves through the world. Accretion and deletion of visual texture elements occur when an object or part of the layout becomes progressively uncovered or occluded. Time-to-contact information – the rate of change in visual flow, air pressure, sound, and so on – emerges as an object approaches the observer or another target (Lee, 2009). In the course of an event, some things change and some things do not. The relations that remain

invariant under transformation can specify what is permanent and what is changing and how the change is occurring (Gibson, 1979). As an object moves through different locations, it causes particular patterns of transformation in the optic array. Through these patterns of transformation, we perceive it as the same, unitary object and we simultaneously perceive its trajectory and manner of motion through the layout. Invariant information for unity and constancy is abstract and relational; it must be invariant over change, and it has a higher order structure (Gibson, 1984a). A looming object on an approach trajectory is an obstacle if the expansion pattern causes increasing, symmetrical deletion of background texture. The object is an aperture – a window to look through – if it progressively reveals more background texture in the middle of the display. By 3 months of age, infants blink and retract their heads at the approach of a looming obstacle on a collision path to their face, but lean forward to peer through an approaching aperture (Gibson, 1982; Gibson and Schmuckler, 1989; Schmuckler et al., 2007). Crabs, frogs, kittens, and other animals do likewise (Gibson, 1970).

Information can be detected in different modalities – through looking, feeling, hearing, and so on. In the natural environment, information for affordances and events is typically multimodal (Gibson, 1984a, 1992). Direction of self-locomotion, for example, is specified redundantly by patterns of visual flow, vestibular information, proprioceptive information from muscles and joints, and occasionally auditory input. A communication event between infant and mother is specified redundantly by mother's facial movements, the synchronous sounds emanating from her mouth, and the contingent, causal relations as the infant responds in turn. Infants do not need to learn to paste percepts from different modalities together. Rather, information from different modalities belongs together when it is unified by the same invariant relations, typically occurring in the same event. Infants demonstrate multimodal perception of rigid and deformable motions, for example, through viewing and feeling hard and spongy objects, either held in their hands or in their mouths (Gibson and Walker, 1984). They do likewise for rigid and deformable ground surfaces by viewing an event on the surface and by creating events through actively palpating the surface (Gibson et al., 1987; Joh and Adolph, 2006). In cases involving temporal synchrony, rhythm, and intensity, the invariants are amodal, meaning the information is not specific to a particular modality; it is the same in every modality (Gibson, 1984a, 1984b).

Distinctive Features

Based on her research on learning to read, in her earlier formulations of what is learned, Gibson (1969, 1977) emphasized learning to detect the *distinctive features* of things in the world. Distinctive features are the relational contrasts or minimal set of attributes that distinguish one kind of thing from another. For example, the distinctive features of written letters include the presence or absence of straight lines, curves, and symmetry. When learning to read, children must learn to detect these distinctive features that, in different combinations, specify particular letters (Gibson and Levin, 1975). Distinctive features are relational properties that are invariant under certain transformations; growing a letter in size or distorting its slant does not change the critical information, whereas reversing the letter or changing the topography does. Importantly, not every characteristic of a stimulus is a distinctive feature. In the case of letters, color and line thickness are not distinctive features that distinguish one letter from another. Discrimination occurs when, "out of a mass of stimulus properties emanating from a set of objects, the perceiving organism learns to choose only those necessary for distinguishing between the objects" (Gibson, 1969: p. 140).

Although Gibson's early work on distinctive features was inspired by her research on reading, the processes are presumably more general. Reading relies on perceptual processes that evolved for other purposes and are shared by other animals. Controlled rearing studies in animals provide evidence of learning to discriminate among sets of objects by discovering the particular features that distinguish them (Gibson and Walk, 1956). Nonetheless, Gibson (July 1989) asked, "Do infants progressively learn to select out distinctive features of the things around them, as I once thought? Maybe. But that is not where perceptual learning begins. I think it begins with learning affordances: what properties of the environment have utility for what actions" (p. 14).

Process of Perceptual Learning

Perceptual learning is an increased ability to detect information specifying affordances, events, and distinctive features. In Gibson's (1992) view, "perceptual learning is first and foremost a process of selection" (p. 217); perceptual information becomes increasingly differentiated and specific to the self, the world, and the relations between them. Exploratory activity is the primary behavioral mechanism for generating perceptual information and for differentiating information from the flux. Through spontaneous exploratory activity, animals seek information for guiding actions adaptively. This search for information is as integral to animals' nature as breathing (Gibson, 1970, 1977). What is extracted at any moment depends on the animal's evolutionary history, its level of development and learning, and the activities it is engaging in – what it is currently doing and what it wants to do next (Gibson, 1994). During a bout of exploration, the search for information terminates with a significant reduction of uncertainty (Gibson, 1969). However, animals are always searching for increasing specificity of information for affordances; over learning and development, there is a continual increase in predictability and efficiency of perceiving what is doable (Gibson, 1991, 1992).

Differentiation and Increasing Specificity

According to Gibson (1977), infants perceive objects, events, and the layout of the environment from the start. Perceptual learning is the process whereby perceptual information becomes increasingly differentiated and specific to the things in the world and to what one can do with those things (Gibson, 1992).

Gibson's focus on *differentiation* and *increasing specificity* sets her theory apart from traditional theories of perceptual

learning. On the traditional account, perception requires more than the available sensory stimulation; thus, perceptual learning is a process of learning to construct, supplement, infer, hypothesize, interpret, organize, associate, or otherwise enrich impoverished input. Gibson and Gibson (1955) pointed out that the traditional enrichment view requires perceptual learning to produce percepts that *decrease* in correspondence with the available information. The Gibsons proposed to the contrary that perceptual learning should result in percepts that *increase* in correspondence with the available information. Rather than an add-on process, on their account, perceptual learning is a process of differentiation, selection, and extraction of information; the information was always present, just not previously detected. They theorized that learning is a process of "differentiating previously vague impressions," not "enriching previously meager sensations" (Gibson and Gibson, 1955: p. 34). Perception does not improve through constructing new descriptions of the world, but through discovering new information about it (Gibson, 1978; Gibson and Spelke, 1983).

A good, intuitive example of increased specificity through differentiation is wine tasting; a novice wine taster can only distinguish reds from whites, whereas a connoisseur perceives differences that correspond to specific grapes, regions, and sometimes even harvest years. Over years of practice, small differences – differences that truly exist in the chemical signatures of different wines – become more easily distinguished; an expert taster can identify a Bordeaux from France and distinguish it from a similar mix of Cabernet Sauvignon, Merlot, and Cabernet Franc grapes from California. Other everyday examples abound: Experienced drivers can easily perceive whether the spot is sufficient for parallel parking or the gap in moving traffic will allow merging or switching lanes; experienced listeners can distinguish the cellos from the violins and the clarinets from the oboes; expert radiologists can distinguish a cancerous spot from normal breast tissue on a mammogram or a fracture from a normal vascular groove on an X-ray. In the laboratory, perceptual learning follows the same course, proceeding from murky undifferentiated percepts to increasing specificity. For example, with practice, adults can learn to use gradients of texture density on a large lawn of mown grass to estimate the distance between targets (Gibson and Bergman, 1954; Gibson et al., 1955; Purdy and Gibson, 1955). Infants learn to distinguish between two types of structure in optic flow, radial versus lamellar flow, using the former for steering and the latter to control upright balance (Stoffregen et al., 1987). In each case, the invariant information was there all along to be differentiated. Targets of perception that were once confusable and undifferentiated become increasingly discriminable with practice (Gibson, 1963).

Not all the available information is relevant for the task at hand. We never detect the total stimulation reaching our receptors. Instead, we sample from this vast pool so that only part of the potential information becomes effective. The key to perceptual learning is the *education of attention* – learning which variables to attend to and which to ignore. Through practice and experience, attention becomes fine-tuned toward the relevant information (Gibson and Spelke, 1983). Attention relates perception to action and to a person's needs and motives (Gibson and Rader, 1979). Search is part of attending. The more perceivers know what is wanted and where and how to look for it, the less they bother with irrelevant and unhelpful information in performing a task. Perceptual learning seeks the minimal information that differentiates one thing from another. Reducing the information to just what is needed is typical of skilled performance. Detecting order, invariance, and redundancy are ways of achieving specificity and economy.

Exploratory Activity

For Gibson (1994), exploratory activity is integral to the process of perceptual learning. Animals forage for information about the relations between self and the surrounding environment (Gibson, 1997). Evolution provides animals with perception–action systems that equip them to discover what the world is all about; the development of these systems provides the motivation to explore the accessible surround, act on it, and further extend explorations as capacities grow. Vision is an especially powerful motivator because it provides information about the world at large (Gibson, 1978). Human infants begin by looking around. Anything in a baby's field of view can provide an incentive for exploration. Before they acquire independent mobility, infants reach for the things that they see. After they can locomote, babies crawl toward and finally walk and run for the attractive goal at a distance.

Exploratory activity includes behaviors such as scanning, fixating, feeling, mouthing, licking, listening, whisking, sniffing, and manipulating – seeking information using whatever actions and body parts are most currently suited for efficient search (Gibson, 1970). Multimodal exploration is the norm. Although all movements produce perceptual information, Gibson (1988a) believed that deliberate, focused exploratory behavior has special importance for perceptual learning: "A sequence of acts termed exploratory will have some outcome and will not be random. It will have a perceptual aspect, a motor aspect, and a knowledge-gathering aspect" (p. 5). Observing the consequences of exploratory actions leads to learning.

Exploration develops and the development of exploratory activity facilitates and constrains perceptual learning (Gibson, 1988a). Long before birth, animals have movements in their repertoires and are sensitive to ambient stimulation. So these movements and sensitivities can be harnessed immediately. As new perception–action systems become available, they are tried out for use in the service of searching (Gibson, 1997). In human infants, the development of postural control (head, trunk, and finally the whole body) opens up new opportunities for exploration and perceptual learning (Gibson, 1988a). In the first few weeks of life, infants gain control over their eyes and head. This enables active scanning of the visual world. In this period of development, infants learn about objects, people, and events by actively directing their gaze toward points of interest. Several months later, infants acquire independent sitting. This frees up their hands and leads to improvements in manual skills. They gain the ability to explore objects with their hands, to coordinate visual inspection with manual actions, and to pass objects back and forth from mouth to eyes. In this period

of development, they learn about the physical properties and affordances of objects. Toward the end of the first year, when infants begin to crawl and to walk, the wider world becomes available to them. In this period, they use locomotor exploration to learn about the properties and affordances of surfaces, places, and the spatial layout. In this way, perceptual-motor development is the driver of early perceptual learning. Emergence of a new action system, such as walking, instigates exploratory activity that generates information about the body's new and changing capacities and about properties of the environment that are implicated in the new activity (Gibson, 1988b). Practice with particular exploratory activities may result in optimizing the search or elaborating it with more complex forms of coordination like using the hands for different roles in manipulation. At any point in development, exploratory activity can be extended with tools such as a rattle or hammer in hand, or a handrail lining a staircase, or by putting instrumental means at infants' disposal by yoking their leg kicks to an overhead mobile, or connecting a pressure-sensitive pacifier to a recording or slide show (Gibson, 1992).

Hallmarks of Human Behavior

Toward the end of her career, Gibson (1994, 1997) expanded her theory of perceptual learning to encompass the hallmarks of human behavior – the fundamental aspects of human behavior that psychologists need to understand and explain. These hallmarks are inherent in behavior and are specific to behavior; they do not exist in lower level physiological descriptions. Each hallmark develops, and Gibson believed that describing their development is at the heart of developmental psychology. Three major hallmarks are agency (the self in control), prospectivity (the forward looking direction of activity), and behavioral flexibility (transfer of means and strategies to new situations). All three require and promote perceptual learning.

Agency

One of infants' great discoveries is that it is possible to effect change on the world by controlling an observable event through one's own action. This realization comes about through motor activity and perceptual learning. It is the hallmark of *agency* and it is discovered very early in development (Gibson, 1995). Gibson and Pick (2000) wrote, "infants learn at a remarkably early age that their actions have an effect on the environment, that some events going on around them are amenable to their control, and that they can in fact regulate their own actions" (p. 160).

One of Gibson's favorite examples of achieving control is that infants increase their rate of kicking when their foot is tied to an overhead mobile, such that the jiggling mobile elements are contingent on their kicks (Rovee-Collier and Geloski, 1979). Another favorite finding is that infants increase their rate of sucking on a nonnutritive pacifier when the sucks are linked to the sound of mother's voice or bring a slide show into focus (DeCasper and Fifer, 1980; Kalnins and Bruner, 1973). Infants in control groups do not kick or suck differently from baseline when the contingencies do not exist. Although the mobile and nonnutritive sucking paradigms are examples of instrumental conditioning, in Gibson's (1992) view, these are not simply examples of associative pairing. The reinforcer is not the jiggling mobile elements or the voice. It is control over the environment that is rewarding. Indeed, infants' emotional expressions are linked with the sense of control. They display happy expressions when the contingency is working and angry expressions when the contingency is broken during extinction (Lewis et al., 1990, 1985). Yoked controls who see the same displays exhibit only interest and neutral facial expressions. Infants detect the relation between their own exertion and the environmental change just as they perceive the sequence of cause and effect.

What is learned in the development of agency? Infants acquire a sense of a self that is separate from the world, and that can act intentionally on its own powers on the world (Gibson, 1997). Perceiving oneself as a source of control that causes perceptible changes in the world is the epitome of perceiving oneself (Gibson, 1995). In answer to William James' (1879) famous question, "Are we automata?" Gibson's (1995) answer was no.

Prospectivity

Behavior extends over time. At any moment, perceptual consequences of the preceding actions inform the planning of subsequent actions (Gibson and Pick, 2000). *Prospectivity* refers to the forward-looking quality of behavior. To select and perform actions adaptively, perception must be planful and prospective. Many of the ideas that were most important to Gibson were linked to prospectivity. For example, perception of affordances implies perceiving possibilities for potential future actions, and guides planning of future actions. Exploratory behavior is used prospectively to generate information for planning actions. Agency and control also imply prospectivity because agents have an eye toward the future consequences of their actions.

Infants show evidence of prospectivity very early in development, although planning improves with learning and experience. For example, before infants raise an arm to reach, their abdominal muscles activate to stabilize their posture, anticipating the disruption to balance caused by raising the arm (von Hofsten, 1993). Infants bring their hand to the location where a moving object will arrive so as to intercept it (von Hofsten, 1980) and they begin to close their hand in anticipation of grasping the object rather than as a reaction to encountering the object (von Hofsten and Ronnqvist, 1988). As infants learn more about what leads to what, and pick up information about the causal relations in events, the prospectivity of behavior increases.

Flexibility

One of the most impressive hallmarks of human behavior is its flexibility. *Flexibility* refers to the generation of new means and strategies and their transfer to novel and variable environmental and bodily conditions and to the task at hand (Gibson, 1994). Affordances constantly change, so "stereotypy of reaction is perilous" (Gibson and Pick, 2000: p. 169).

For example, goats produce a variety of reactions when they receive an electric shock signaled by a buzz or light – walking/running backward or forward, walking in circles, side-stepping, independent leg movements, foreleg flexion, foreleg extension, humping the back, rearing, and so on – until they discover the

behavior that works to avoid the shock (Gibson, 1952). Humans also show a variety of strategies, and they can transfer means that work in one context to another. In her first published paper (Gibson et al., 1932), Gibson found that adults who were conditioned to withdraw the index finger of one hand to avoid a shock transferred the means of avoidance to the index finger on the other hand when told to place that finger on the electrode.

Infants also generate a variety of means and strategies. When tested at the brink of shallow and steep slopes, infants correctly perceived affordances for walking (Adolph et al., 1993). Moreover, on the steepest slopes where walking was impossible, infants generated new means of descent: sliding down backward feet first, in a sitting position, and belly-down headfirst. Are these alternative means of descent transferable? Yes, when discovered over weeks of crawling, infants can use the same sliding strategies to descend slopes after they begin walking (Adolph, 1997).

Flexibility improves with learning and development. Expansion of the repertoire of available actions provides more options for confronting novel challenges. Moreover, experiencing new environmental conditions provides opportunities to generalize old skills to new settings and to develop new solutions on the fly (Gibson, 1997).

Conclusions Beyond Perception

Gibson's theory of learning extends beyond perception and the hallmarks of human behavior. Many psychologists think of cognition exclusively in terms of problem solving, reasoning, conceptualizing, remembering, and so on. However, Gibson (1991) points out that these processes, like the hallmarks of behavior, "begin with and depend on knowledge that is obtained through perception" (p. 494). Cognition is built on a foundation of perceptual knowledge. "Perhaps knowledge eventually becomes a system of representations and beliefs about the world … but … representations and beliefs must be grounded by detection of the surfaces, events, and objects of the layout—the 'stuff' of knowledge must somehow be obtained from the world" (Gibson, 1988a: p. 34).

Acknowledgments

The writing of this article was supported by National Institute of Health and Human Development Grant R37-HD33486 to Karen Adolph and a predoctoral National Science Foundation fellowship DGE-0813964 to Kari Kretch. We are grateful to Mark Blumberg, Anne Pick, and Harry Heft for their helpful comments and to Eleanor "Jackie" Gibson for sharing her unpublished manuscripts and talks.

See also: Cognitive Development During Infancy and Early Childhood across Cultures; Cognitive Development: Mathematics Learning and Instruction; Comparative Evolutionary Psychology: Current Status and a Proposal for a More Integrated Future; Goal-Oriented Behaviour Among Human and Non-Human Primates; Human Behavioral Ecology; Infancy and Human Development; Learning Theories and Educational Paradigmstnq; Learning to Learn; Motor Skills: Development in Infancy and Early Childhood; Piaget's Theory of Cognitive Development; Play During Infancy and Early Childhood: Cultural Similarities and Variations; Problem Solving During Infancy and Early Childhood, Development of; Reading Skills, Acquisition of: Cultural, Environmental, and Developmental Impediments; Self in Culture: Early Development.

Bibliography

Adolph, K.E., 1997. Learning in the development of infant locomotion. Monographs of the Society for Research in Child Development 62 (3, Serial No. 251).

Adolph, K.E., 2000. Specificity of learning: why infants fall over a veritable cliff. Psychological Science 11, 290–295.

Adolph, K.E., Eppler, M.A., Gibson, E.J., 1993. Crawling versus walking infants' perception of affordances for locomotion over sloping surfaces. Child Development 64, 1158–1174.

Bertenthal, B.I., Campos, J.J., Barrett, K.C., 1984. Self-produced locomotion: an organizer of emotional, cognitive, and social development in infancy. In: Emde, R.N., Harmon, R.J. (Eds.), Continuities and Discontinuities in Development. Plenum Press, New York, pp. 175–210.

DeCasper, A.J., Fifer, W.P., 1980. Of human bonding: newborns prefer their mother's voices. Science 208, 174–176.

Gibson, E.J., 1952. The role of shock in reinforcement. Journal of Comparative and Physiological Psychology 45, 18–30.

Gibson, E.J., 1963. Perceptual learning. Annual Review of Psychology 14, 29–56.

Gibson, E.J., 1969. Principles of Perceptual Learning and Development. Appleton-Century Crofts, New York.

Gibson, E.J., 1970. The development of perception as an adaptive process. American Scientist 58, 98–107.

Gibson, E.J., 1977. How perception really develops: a view from outside the network. In: Basic Processes in Reading: Perception and Comprehension. Erlbaum, Hillsdale, NJ, pp. 155–173.

Gibson, E.J., 1978. C'est moi. Contemporary Psychology: A Journal of Reviews 23, 609–614.

Gibson, E.J., 1980. Eleanor J. Gibson. In: Lindzey, G. (Ed.), A History of Psychology in Autobiography, vol. 7. W H Freeman & Co, San Francisco, pp. 239–271.

Gibson, E.J., 1982. The concept of affordances in development: the renascence of functionalism. In: Collins, W.A. (Ed.), The Concept of Development: The Minnesota Symposia on Child Psychology, vol. 15. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 55–81.

Gibson, E.J., 1984a. Development of knowledge about intermodal unity: two views. In: Liben, L.S. (Ed.), Piaget and the Foundations of Knowledge. Erlbaum, Hillsdale, NJ, pp. 19–41.

Gibson, E.J., 1984b. Perceptual development from the ecological approach. In: Lamb, M.E., Brown, A.L., Rogoff, B. (Eds.), Advances in Developmental Psychology, vol. 3, pp. 243–286.

Gibson, E.J., 1988a. Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge. Annual Review of Psychology 39, 1–41.

Gibson, E.J., 1988b. Levels of description and constraints on perceptual development. In: Yonas, A. (Ed.), Perceptual Development in Infancy: Minnesota Symposium on Child Psychology, vol. 20, pp. 283–299.

Gibson, E.J., July 1989. Learning to Perceive or Perceiving to Learn? Paper Presented to the International Society for Ecological Psychology, Oxford, OH.

Gibson, E.J., 1991. An Odyssey in Learning and Perception. MIT Press, Cambridge, MA.

Gibson, E.J., 1992. How to think about perceptual learning: twenty-five years later. In: Pick, H.L., van den Broek, P., Knill, D.C. (Eds.), Cognition: Conceptual and Methodological Issues. American Psychological Association, Washington, DC, pp. 215–237.

Gibson, E.J., 1994. Has psychology a future? Psychological Science 5, 69–76.

Gibson, E.J., 1995. Are we automata? In: Rochat, P. (Ed.), The Self in Infancy: Theory and Research. Elsevier Science, Amsterdam, pp. 3–15.

Gibson, E.J., 1997. An ecological psychologist's prolegomena for perceptual development: a functional approach. In: Dent-Read, C., Zukow-Goldring, P. (Eds.), Evolving Explanations of Development: Ecological Approaches to Organism-Environment Systems. American Psychological Association, Washington, DC, pp. 23–45.

Gibson, E.J., 2002. Perceiving the Affordances: A Portrait of Two Psychologists. Lawrence Erlbaum Associates, Mahwah, NJ.

Gibson, E.J., Bergman, R., 1954. The effect of training on absolute estimation of distance over ground. Journal of Experimental Psychology 48, 473–482.

Gibson, E.J., Bergman, R., Purdy, J., 1955. The effect of prior training with a scale of distance on absolute and relative judgments of distance over ground. Journal of Experimental Psychology 50, 97–105.
Gibson, E.J., Levin, H., 1975. The Psychology of Reading. MIT Press, Cambridge, MA.
Gibson, E.J., Pick, A.D., 2000. An Ecological Approach to Perceptual Learning and Development. Oxford University Press, New York.
Gibson, E.J., Rader, N., 1979. Attention: the perceiver as performer. In: Hale, G.A., Lewis, M. (Eds.), Attention and Cognitive Development. Plenum Publishing Corporation, New York.
Gibson, E.J., Riccio, G., Schmuckler, M.A., Stoffregen, T.A., Rosenberg, D., Taormina, J., 1987. Detection of the traversability of surfaces by crawling and walking infants. Journal of Experimental Psychology: Human Perception and Performance 13, 533–544.
Gibson, E.J., Schmuckler, M.A., 1989. Going somewhere: an ecological and experimental approach to development of mobility. Ecological Psychology 1, 3–25.
Gibson, E.J., Spelke, E.S., 1983. The development of perception. In: Flavell, J.H., Markman, E. (Eds.), Cognitive Development. In: Mussen, P. (Ed.), Handbook of child psychology, vol. 3. Wiley, New York.
Gibson, E.J., Walk, R.D., 1956. The effect of prolonged exposure to visually presented patterns on learning to discriminate them. Journal of Comparative and Physiological Psychology 49, 239–242.
Gibson, E.J., Walk, R.D., 1960. The "visual cliff". Scientific American 202, 64–71.
Gibson, E.J., Walker, A.S., 1984. Development of knowledge of visual-tactual affordances of substance. Child Development 55, 453–460.
Gibson, J.J., 1966. The Senses Considered as Perceptual Systems. Houghton-Mifflin, Boston.
Gibson, J.J., 1979. The Ecological Approach to Visual Perception. Houghton Mifflin Company, Boston.
Gibson, J.J., Gibson, E.J., 1955. Perceptual learning: differentiation or enrichment? Psychological Review 62, 32–41.
Gibson, J.J., Jack, E.G., Raffel, G., 1932. Bilateral transfer of the conditioned response in the human subject. Journal of Experimental Psychology 15, 416–421.
Held, R., Hein, A., 1963. Movement-produced stimulation in the development of visually guided behavior. Journal of Comparative and Physiological Psychology 56, 872–876.
von Hofsten, C., 1980. Predictive reaching for moving objects by human infants. Journal of Experimental Child Psychology 30, 369–382.
von Hofsten, C., 1993. Prospective control: a basic aspect of action development. Human Development 36, 253–270.
von Hofsten, C., Ronnqvist, L., 1988. Preparation for grasping an object: a developmental study. Journal of Experimental Psychology: Human Perception and Performance 14, 610–621.
Joh, A.S., Adolph, K.E., 2006. Learning from falling. Child Development 77, 89–102.
Kalnins, I.V., Bruner, J.S., 1973. The coordination of visual observation and instrumental behavior in early infancy. Perception 2, 307–314.
Kretch, K.S., Adolph, K.E., 2013. Cliff or step? Posture-specific learning at the edge of a drop-off. Child Development 84, 226–240.
Lee, D.N., 2009. General tau theory: evolution to date. Perception 38, 837–858.
Lewis, M., Alessandri, S.M., Sullivan, M.W., 1990. Violation of expectancy, loss of control, and anger expressions in young infants. Developmental Psychology 26, 745–751.
Lewis, M., Sullivan, M.W., Brooks-Gunn, J., 1985. Emotional behavior during the learning of a contingency in early infancy. British Journal of Developmental Psychology 3, 307–316.
Pick, H.L., 1992. Gibson, E.J.,: learning to perceive and perceiving to learn. Developmental Psychology 28, 787–794.
Purdy, J., Gibson, E.J., 1955. Distance judgment by the method of fractionation. Journal of Experimental Psychology 50, 374–380.
Rovee-Collier, C.K., Geloski, M., 1979. The economics of infancy: a review of conjugate reinforcement. In: Reese, H.W., Lipsitt, L.P. (Eds.), Advances in child development and behavior, vol. 13. Academic, New York.
Schmuckler, M.A., Collimore, L.M., Dannemiller, J.L., 2007. Infants' reactions to object collision on hit and miss trajectories. Infancy 12, 105–118.
Soska, K.C., Adolph, K.E., Johnson, S.P., 2010. Systems in development: motor skill acquisition facilitates three-dimensional object completion. Developmental Psychology 46, 129–138.
Stoffregen, T.A., Schmuckler, M.A., Gibson, E.J., 1987. Use of central and peripheral optical flow in stance and locomotion in young walkers. Perception 16, 113–119.
Walk, R.D., Gibson, E.J., 1961. A comparative and analytical study of visual depth perception. Psychological Monographs 75 (15, Whole No. 519).

Giddens, Anthony (1938–)

Steven Loyal, School of Sociology, University College Dublin, Dublin, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Anthony Giddens is perhaps Britain's leading sociologist, renowned especially for his theories of structuration and late modernity, and his championing of 'third-way' politics. This article outlines Giddens work in terms of four overlapping phases focusing on his early writings, his theory of structuration, his theory of modernity, and finally his writings in politics and social policy. It concludes by arguing that although Giddens provides some insights in his sociology, his work is underpinned by a number of theoretical and substantive failings which ultimately derive from his liberal worldview and include an overemphasis on agency and a downplaying of conflict, and social and material interests as fundamental characteristics of the social world.

Anthony Giddens is perhaps Britain's leading sociologist, renowned especially for his theories of structuration and late modernity, and his championing of 'third-way' politics. Born in London in 1938, he attended the University of Hull and graduated in 1959 in Sociology and Psychology. He studied for a master's degree in Sociology at the London School of Economics, and in 1961 became a lecturer in the Department of Sociology at Leicester University. He held positions at Simon Fraser University and the University of California at Los Angeles before moving to King's College Cambridge. He formed Polity Press in 1985 and became director of the LSE in 1997. His growing status as the most visible intellectual proponent of 'third-way' politics drew him into the inner circle of British Prime Minister, Tony Blair, and in 2004 he was given a life peerage as Lord Giddens of Southgate.

Giddens's intellectual career is best understood in terms of four overlapping periods, each marked by a distinctive set of theoretical concerns. His early work, 1970–75, focused on the exposition of the classical tradition of European sociology, and Giddens was influential in establishing the canonical trilogy of Marx, Weber, and Durkheim as the basis of social theory. Subsequently, until 1989, he focused on the possibility of transcending a series of perceived dualisms within social theory, most significantly that between agency and structure. The resulting theory of structuration led to an attempted rewriting and reperiodizing of human history. The third phase of his career, 1990–93, developed these theoretical and temporal insights into a more substantive analysis of the contours of modernity and the contemporary stage of what he referred to as 'late modernity.' This prepared the way for the most recent phase in his work, in which he moved from sociology to more directly political–theoretical concerns. Giddens presentation of the 'third way' can be seen as an application of his earlier theoretical work on 'dualisms': an attempt to transcend the dichotomy between left- and right-wing political ideologies.

His first major book, *Capitalism and Modern Social Theory* (1971), not only set out a comprehensive exposition of the sociological ideas of Marx, Weber, and Durkheim, but also reexamined significant areas of convergence and divergence. Rather than converging around an implicit, if abstract, Hobbesian 'problem of order,' these three were instead concerned with the substantive historical–sociological problem of the profound rupture separating capitalism from prior feudal or traditional social formations. Their politics sought to synthesize liberalism and revolutionary or radical forms of thought in different ways. Giddens developed some of these ideas in, *The Class Structure of Advanced Societies* (1973), synthesizing Weberian and Marxist views of class.

His most significant contribution to social theory, the theory of structuration, was an attempt to overcome the division between sociological approaches that emphasize agency and those that emphasize structural constraints. The former focus on self-conscious agents – their intentionality, subjectivity, knowledgeability, and ability to construct, create, or make the social world they find themselves in. Advocates have included phenomenology, ethnomethodology, symbolic interactionism, and rational choice theory. Structural theories, on the other hand, place more emphasis on objective factors, the social determination of the self, and the invisible forces and emergent dynamics that shape the actions, perceptions, and 'second nature' of individuals. These have included functionalism, structuralism, and the many varieties of Marxism. The agency/structure dualism that informs these two traditions runs through a series of related perspectival and methodological tensions that plague the discipline: individual *versus* society, micro *versus* macro, and subjective *versus* objective.

In his theory of structuration, Giddens argues that the agency/structure dualism can be overcome only by synthesizing insights from a variety of otherwise flawed perspectives. This involves a reformulation of the lexicon of sociological concepts. For Giddens, prior theoretical applications of the term 'structure' – most significantly those found in functionalism and Marxism – tended to define structure as causally efficacious patterned social relationships that are not only external to human agency but also constraining upon it. He discerns a very different understanding of 'structure' in Levi-Strauss's structuralism. Structure here refers to abstract models in the form of binary oppositions and dual relations, existing in and through human beings and that do not exist in time and space but as relations of presence and absence. Utilizing this latter approach, Giddens offers his own novel definition of structure. Structures, like languages, are 'virtual' since they exist 'outside of time and space,' are 'subjectless,' and are, for the most part, unintentionally reproduced in practices. By identifying structure with language, he hoped to effect a dynamic

juxtaposition between the speech or action implemented by an agent and the structure that forms the condition of possibility for generating this speech and action. This had two effects. First, there is what Giddens refers to as a 'duality of structure': structure is no longer understood as simply constraining but also as enabling. Structure not only limits action through rules of 'syntax,' it also makes possible the generation of action. Secondly, structure is both the medium and the outcome of action. Hence, the 'instantiation' of structure in individual action recursively draws upon and reproduces structure in a manner akin to a speech–act drawing upon and reproducing the totality of language. Every act of social production is therefore simultaneously an act of reproduction.

For Giddens, structure and agency form two sides of the same coin and are connected through social practices. They are inseparable dimensions of the flow of activities in which individuals participate during the course of their day-to-day lives. Specifically, he sees structures composed of rules and resources. Rules may be explicit or tacit, intensive or shallow, formal or informal, strongly or weakly sanctioned, but should generally be understood, in Wittgenstein's sense, as practical forms of knowledge that 'allow us to go on' in novel circumstances. Resources are 'authoritative' capabilities that generate command over persons or 'allocative' capabilities that generate command over objects or other material phenomena.

In addition to rethinking the concept of structure, Giddens also reconceptualizes the concept of agency. In his stratification model of action, he draws upon and modifies the Freudian schema of agency to argue that an actor's consciousness has three aspects: discursive consciousness, practical consciousness, and unconsciousness. Loosely corresponding to this threefold division of the consciousness, he refers to reflexive monitoring, rationalization, and the deep motivations for action. Practical consciousness and its rationalization as the tacit or 'mutual' knowledge that provides agents with the ability to 'go on' in relation to rule-bound social life is most significant for understanding social life.

Although free to choose, agents follow routines to avoid ontological insecurity and any disruption of the basic security system internalized during childhood. Drawing on Merton's classic essay on the unintended effects of purposive action, however, Giddens argues that an individual's intentional acts often produce unintended consequences. These unintended consequences then become the future unacknowledged conditions that structure the subsequent action of the agent.

A second aspect of Giddens's structuration theory relates agency to power, arguing that an agent ceases to be such if s/he loses the ability to 'act otherwise.' However, given that there exists a 'dialectic of control' built into the very nature of agency and involving the interplay of autonomy and dependence, a total loss of agency is rare.

In parallel with the development of the theory of structuration, Giddens produced his own historical sociology as a 'positive critique' of Marx's historical materialism, which he considered as economically reductionist and methodologically suspect. In *A Contemporary Critique of Historical Materialism* (1981), he repeats a number of criticisms he has made against Marxism. These include the recognition that (1) there is a need for a distinction between capitalism and industrialism; (2) political factors have played a far more significant role in the latter-day development of advanced societies than Marx had envisaged; (3) the preeminence of 'traditional' elements in the capitalist societies is tied to the rise of nationalism; (4) the study of the nation state within its international context is crucial; (5) there is a need for a move away from evolutionary and endogenous models of social change; and (6) Marxism fails to recognize the influence of military factors in social change. He then elaborates a nonfunctionalist, nonevolutionary, nonteleological, multidimensional, and historically contingent view of social change. His historical schema incorporates a tripartite societal typology that distinguishes tribal class-divided societies from class societies. These societies are defined according to their level of social and system integration and time–space distanciation. Social integration, refers to 'systemness' in circumstances of 'copresence' or 'face-to-face' interaction, not only emphasizes the spatial significance of presence but also attempts to convey the importance of the 'positioning' of the body within social space. By contrast, system integration refers to the reciprocity which exists between actors or collectivities across extended time–space or in relation to absent others outside of the conditions of copresence. This involves a process of 'time–space distanciation': the 'stretching' of social systems across time and space. In *Nation-State and Violence* (1985), he develops and modified this framework, exploring the complex symbiotic codevelopment of capitalism, industrialism, and the nation-state. For Giddens these three spheres provide the basis for the four irreducible, though connected, 'institutional clusterings' that characterize modern society: capitalistic enterprise, industrial production, heightened surveillance, and centralized control of the means of violence. This characterization of modernity is developed further in the third phase of his work. In *The Consequences of Modernity* (1990), Giddens argues that the critical feature of the dynamic social formation that began to develop in Europe from about the seventeenth century was its sharp, qualitative discontinuity from the previous, traditional social order. This break involved a profound transformation, that is, both global and personal.

Underlying modernity are three sources of dynamism: distanciation, disembedding, and reflexivity. The 'separation and recombination of time and space' facilitates and promotes an increased 'zoning' of social life. Everyday encounters and interactions become less tied to fixed locales and less dependent on the copresence of the individuals involved. This, in turn, facilitates the development of modern rationalized organizations, and permits the emergence of a radical historicity in which the past can be appropriated with the aim of shaping the future. More importantly, it facilitates the 'disembedding' of social systems. Disembedding refers to the 'lifting out' of social relations from their local contexts of interaction, which permits their restructuring across larger spans of time–space. Modernity is inherently globalizing, so that time–space distanciation links the local with the global through disembedding, although there may also be processes of 'reembedding' in which disembedded social relations are again pinned down. The two major types of disembedding mechanism are 'symbolic tokens' and 'expert systems.' Symbolic tokens are media such as money that can be exchanged, regardless of who uses them, while expert systems are systems of technical accomplishment and professional

expertise such as those of doctors, lawyers, architects, and scientists.

Fundamental to both mechanisms, as well as to the reflexivity of modernity more generally, is the concept of trust. A sense of trust in processes, people, and things is a crucial factor in maintaining a sense of ontological security in the modern world, since its absence results in existential angst or dread. By contrast with premodern societies, where trust and risk were anchored in the local circumstances of place, rooted in nature and characterized by hazards from the physical world or violence in social life, modernity offers a new 'risk profile' characterized by 'manufactured risk.' Here the pervasiveness of socially organized knowledge in the form of abstract systems means that risk becomes the defining parameter of modern culture and life, even replacing the preoccupation with wealth. As Giddens notes "The possibility of nuclear war, ecological calamity, uncontainable population explosion, the collapse of global economic exchange, and other potential global catastrophes provide an unnerving horizon of dangers for everyone" (Giddens, 1990: 125). Giddens's analysis of modernity leads to the conclusion that in place of the two classic sociological accounts of the experience of modernity – Marx's theory of alienation and Weber's 'iron cage' of bureaucracy – the discipline would be better served by his own image of modernity as a careering juggernaut.

Although in his structuration theory, reflexivity is a fundamental feature of social action, it takes on a special meaning in Giddens's theory of modernity. Under emerging conditions of 'wholesale reflexivity' everything (including both individuals and institutions) becomes open to reflection and self-monitoring, including reflexivity itself. Social practices are continually examined and reexamined in the light of incoming information and processes of self-evaluation.

This leads Giddens to consider the transformation of the intimate and personal features of day-to-day existence in modernity especially in his books, *Modernity and Self Identity* (1991) and *The Transformation of Intimacy* (1992). Increasingly, the pressures of work and domestic life push individuals toward the continual reconstruction of self-identities as part of a reflexive project. In this increasingly unavoidable autobiographical project, individual choices are made in the context of an array of trajectories and options engendered by abstract systems. Modernity involves both a transformation of lifestyle and a 'transformation of intimacy' in which personal and erotic ties are formed as 'pure relationships.' According to Giddens, 'pure relationships' involve 'commitment' and demands for intimacy, such that trust develops through mutual disclosure alone, rather than (as in more traditional societies) through criteria – such as kinship ties, social duty, or traditional obligations – that exist outside of the relationship itself.

Building on this theory of modernity, Giddens set out a general political program. In *Beyond Left and Right* (1994) and *The Third Way* (1998), he considered how radical politics might be rethought, both theoretically and in practice, in the context of a changing modern world. What he refers to as 'third-way' politics involves a reevaluation of socialism, social democracy, and conservatism in the light of the altered social conditions of modernity that ultimately renders them inoperable. He argues that the collapse of communism has made the distinction between the political 'Left' and 'Right' superfluous. Consequently, the idea of a fixed left/right binary for Giddens needs to be supplanted by the notion of a radical center or 'active middle' embodying a 'utopian realist' position.

The fostering of an active civil society became a central focus for third-way politics. This involves the state and civil society acting in partnership with one another to provide material and social support for local groups to engage in 'generative politics' orientated toward empowerment. Echoing Beck's discussion of the importance of 'subpolitics' in an era of risk society, Giddens also argues that increasing individualization and reflexivity has led to new forms of democratization or 'dialogic democracy.' In some ways following Habermas's theory of communicative action, this 'democratization of democracy' presupposes uncoerced discussion as a basis for agreement. Rather than referring to the extension of social and civil rights, dialogic democracy points toward a 'deliberative democracy' where forms of social interchange, social solidarity, and cultural cosmopolitanism can be established. Widening democracy also requires a decentralization of the state, or what he calls 'double democratization,' and a move away from the paternalism of the welfare state.

A further consequence of 'reflexive individualization' has been the reevaluation of traditional left-wing understandings of emancipation as pertaining to 'life chances' and freedom. Such emancipatory politics must now be supplemented with a 'life politics' that breaks out of the restrictive cast of class politics. Concerning rich and poor groups alike, this life politics addresses universal issues, less driven by the polarities of social class: that is, issues pertaining to lifestyle, leisure, consumption, and identity. Giddens also argues for a 'new mixed economy' that would correspond to the realities of 'post-scarcity society' and in which a production orientation is replaced by ideals of self-actualization.

These public policy arguments are further developed in *Europe in a Global Age* (2006) in which he argues that European societies must rethink their welfare state model in the light of globalization, and that they must transform them in terms of addressing changing people's changing lifestyles. Two further books develop and extend these arguments. In *Over to You, Mr Brown* (2007), while reviewing the successes and failures of the New Labor Government between 1997 and 2007, he argues that overall the legacy of the party has been positive though new ideas and a new policy outlook are needed in the future. These include a move away from the New Democrats in the United States, which initially shaped the third-way agenda. They also include a more market and business friendly outlook and a commitment for Labor not to put up taxes. Although societies need to be more egalitarian this has to be done in tandem with maintaining economic dynamism and creating jobs. Here he advocates institutionalizing a policy of 'flexicurity' (flexibility plus security) in the labor market. Devolution in terms of power to local and regional authorities needs to be combined with decentralization in healthcare and education. While lifestyle change has to take center stage in Labor's new political concordant with the public and be regulated by the four major forces in society working together: state, business, voluntary or third sector groups, and citizens. Giddens also argues that multiculturalism in which different cultures and communities come together and respect one another through dialogue is to be championed but within a context where immigration is

regulated to prevent the entry of unskilled workers into the country. In *The Politics of Climate Change* (2007), written in the same year, Giddens discusses the political implications and policy parameters pertaining to climate change and energy security, introducing what he rather immodestly calls 'Giddens's Paradox' into the climate debate. This states that since the dangers posed by global warming are not tangible or visible in day-to-day life people will generally remain passive in relation to climate change. However, waiting until they become visible and acute will be too late. The book is a developed analysis of the political innovations required to limit climate change and to this end a plethora of further policy concepts are unraveled. These include an 'ensuring state' – a state that facilitates emissions capping, promotes technological innovations, and supports the innovative role of markets. Other proposals and concepts include the fostering of 'political convergence' – which refers to how far climate change policy overlaps with other values and political goals, as well as 'economic convergence' which concerns how far economic and technological innovations that are developed to combat global warming also generate competitive advantage to those who deploy them. Giddens also talks of a 'development imperative' in which poorer nations, which have only added marginally to global warming, are allowed to develop and are aided by technology transfer from richer nations. At the same time as rejecting green and conservationist arguments based on the precautionary principle of not interfering with nature Giddens posits a positive view of risk, seen as opportunity rather than danger, as the basis for a model of a low-carbon future. This in turn would be based on a politics of convergence in which the left and right on the political spectrum, as well as rich and poor countries, would come together and cooperate on forging long-term climate policies and sharing the fruits of technological advances. Such long-term planning policy though requiring state regulation and intervention would not be dependent upon a top–down state but rather the aforementioned ensuring state.

Giddens's has often provided commendable commentaries on many sociological debates as well as offering unique insights into the nature of the social world. And it is probably here that his strength as a sociologist lies. He has been criticized, however, for his eclecticism, inconsistency, and the lack of empirical utility of his theoretical input. The theory of structuration has been of little use in substantive sociological case studies or field work, and it does not even appear in his own later policy-oriented studies and arguments. More specifically, he has been criticized for placing an excessive emphasis on individual agency, especially in his theory of structuration. It can be argued that Norbert Elias and Pierre Bourdieu provide more fruitful approaches for transcending various sociological dualisms that he was concerned with. It has also been noted that his politics evidences a strong liberal idealism, which underpinned his theory of structuration and historical sociology as a worldview. His more recent political and policy discussions, rather than renewing social democracy as claimed, reflect the influence of a post-Thatcherite neoliberal consensus, often entailing anodyne commentaries on how common goals can be pursued by divergently positioned social actors. These claims are most evident in the sweeping prescriptions of the putative 'Third Way,' which because again based ultimately on a liberal doctrine lead him to ignore the entrenched and unequal distribution of power, and the embedded nature of conflict and social interests – including ethnonational conflicts, struggles over money, private property, and wealth – that exist at both the national and global level in the contemporary world.

See also: Class Consciousness; Class: Social; Marx, Karl (1818–83); Materiality and Culture; Modernity; Postmodernism in Sociology; Social Identity in Sociology; Social Structure; Sociological Theory; Weber, Max (1864–1920).

Bibliography

Beyond Left and Right, 1994. Polity Press, Cambridge.
Bryant, Chris, Jary, David, 1991. Giddens' Theory of Structuration. Routledge, London.
Capitalism and Modern Social Theory, 1971. Cambridge University Press, Cambridge.
A Contemporary Critique of Historical Materialism, 1981. Macmillan, London.
The Class Structure of Advanced Societies, 1973. Hutchinson, London.
The Consequences of Modernity, 1990. Polity Press, Cambridge.
The Constitution of Society, 1984. Polity Press, Cambridge.
Europe in a Global Age, 2006. Polity Press, Cambridge.
Held, David, Thompson, John (Eds.), 1989. Social Theory of Modern Societies: Anthony Giddens and His Critics. Cambridge University Press, Cambridge.
Kasperson, L.B., 2000. Anthony Giddens: An Introduction to a Social Theorist. Blackwell, Oxford.
Loyal, Steven, 2003. The Sociology of Anthony Giddens. Pluto, London.
Modernity and Self-Identity, 1991. Polity Press, Cambridge.
The Nation State and Violence, 1985. Polity Press, Cambridge.
The New Rules of Sociological Method, 1976. Macmillan, London.
Over to You Mr Brown, 2007. Cambridge Polity Press.
Profiles and Critiques in Social Theory, (Essays 1975–82), 1982. Macmillan, London.
The Politics of Climate Change, 2007. Polity Press, Cambridge.
Social Theory and Modern Sociology (Essays 1982–86), 1987. Polity Press, Cambridge.
Studies in Social and Political Theory, (Essays 1966–76), 1977. London, Hutchinson.
The Third Way: The Renewal of Social Democracy, 1998. Polity Press, Cambridge.
The Transformation of Intimacy: Sexuality, Love and Eroticism, 1992. Polity Press, Cambridge.

Gift: History of the Concept

Christian Papilloud, Institute for Sociology, Martin-Luther University in Halle-Wittenberg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

It is impossible to speak about the gift without evoking the figure of the ethnologist and sociologist Marcel Mauss, nephew of Emile Durkheim, who anchored the concept of the gift in the tradition of social sciences with his essay *The gift* (1923/24). Mauss shows that the gift is polysemous. Its meaning changes according to the objects or the events to which it has been attributed, even if Mauss thinks of being able to propose a general scheme of the gift as the triple obligation to give, to take, and to return the given things. This scheme was the object of many debates in social sciences which contributed to highlight the forms of the gift in the various cultures of the world and in their history. This article reconstructs the main arguments put forward about the gift starting with Marcel Mauss contribution on the gift as a total social fact.

The Gift as a Total Social Fact

Marcel Mauss (1872–1950) and his *Gift* build the convergence point of the broad majority of the works carried out on the gift at an international level in several disciplines such as ethnography and ethnology, the sciences of religions, sociology, economy, and the linguistic. In *The Gift*, Mauss succeeds in synthesizing these research results regarding the operations related to the gift, putting them at the same time in an historical and intercultural perspective. The gift is an exchange formed by three obligations: the obligation to give, the one to take, and the one to return the given things. Two words are critical for Mauss. First, the 'given things' are not only presents, but every possible things, as well as human beings, services, everyday and magic/religious rituals, or offerings. Second, the gift always means a 'mutual obligation' shared by all actors in society. It is not a mechanical give-and-take. It involves all actors in societies in broad exchanges taking place in the long run. Finally, the gift retains the ambivalent property suggested by his name, meaning at the same time 'present' and 'poison.' To give the gift, and to take it put the actors in a cycle of obligations to return the gift in a certain manner and in a given time. If these manners are not respected, the actors are likely to lose face, or even to die. Thus, the gift is not a generous sacrifice only, or the expression of gratuity. It challenges the actors who should show that they can give, take, and return gifts in an appropriate way.

Mauss works out his conception of the gift considering the sumptuary forms of consumption in the Kwakiutl and Eskimo societies described by Franz Boas. Mauss pays attention to one of the most spectacular of them, the *potlatch*, the ceremonial destruction of the most invaluable goods of a tribe at regular intervals of time in order to balance property in society. Inspired by James Frazer (1894), John Swanton (1910), and Charles Seligman and Brenda Seligman (1911), Mauss formulates assumptions regarding the relation of Melanesian Papuan to the *potlatch* (1913), which will influence the views of anthropologists on exchanges in the Melanesian societies. Finally, the works of Marcel Granet (1920), Georges Davy (1920), and the materials delivered by Bronislaw Malinowski about the *kula* on the Trobriand Islands (1922) convince Mauss that something like a 'rock' (Mauss, 1999[1923/24]: 264) of all social practices exists in all of the known cultures. This 'rock' is the gift as universal principle of exchange leading to the peaceful association of the human beings through the mutual obligations to give, take, and return presents. What motivates the actors to do so?

There are two forces that Mauss describes using the Melanesian terms *mana* and *hau*. The *mana* is the force at the core of the magic and the religion which denotes the sacral character of the objects embedded in gifts (Firth, 1940: 483–510). The *hau* is the spirit of the giver belonging to the thing given which forces the return of the thing to its owner – an interpretation which will raise many controversies in ethnology (e.g., Lévi-Strauss, 1999: XXXVIII; Lefort, 1979[1950]: 1402; Weiner, 1985: 211–215; Firth *in* James and Allen, 1998: 23). Critics blame Mauss for not having interpreted the term *hau* correctly, and for postulating with the *mana* a mystical origin of human exchanges. According to Mauss, though, the *mana* and the *hau* are mythical expressions used by the Melanesian societies in order to describe the power of practical actions embedded in the gift. These actions are powerful because they regulate the entire society, i.e., they enable society to return to a balanced state where the power relationships as well as the properties are best distributed among the actors.

Mauss' gift is also at the origin of a more epistemological consideration regarding the nature of the social facts. Like the gifts reflecting all the facets of a society – domestic economy, religious ceremonies, festivities, the exchanges with foreign countries, diplomatic and power relationships, etc. – the social facts also crystallize this totality because they rest on gifts. The idea of totality is not an abstract one. On the contrary, totality means the material complexity of societies which enter the gift exchanges. These exchanges are vast systems of reciprocity which bind the human beings together within groups and societies. These lead Mauss to see in the gift a political ideal described in the conclusions of his essay. Mauss sees the gift as a mean to change the modern society making human beings more aware of their solidarity. While giving, one does not only give a present, but also gives oneself to the others. According to Mauss' conception of the gift, it is not a phenomenon specific to foreign societies and cultures. It is at the core of our societies which remains veiled by the rationalization and the individualism in modernity.

Mauss' gift stimulates many works in ethnology, cultural anthropology as well as in history, economy, philosophy, and sociology which gradually open the meanings of the gift, and tend to put in question the three obligations described by Mauss. With few exception regarding the critical considerations on Mauss' gift by Raymond Firth, Claude Lévi-Strauss, or Georges Bataille, most of these debates start in the early 1970s.

Reciprocity Instead of Obligations

In ethnology and cultural anthropology, the major works of Lévi-Strauss on the kinship structures and of Marshall Sahlins on exchanges take there starting point in Mauss' conception of the gift.

Lévi-Strauss and Sahlins criticize Mauss' global conception of the gift based on the sciences of religion. If the gift is important in ethnology and cultural anthropology, it is not because of the three obligations which Mauss uses in order to describe the kind of 'magical' power of the gift in regulating society. The gift shows us something more, i.e., the fundamental social link between us – reciprocity. According to Lévi-Strauss, reciprocity is the universal mental structure of human beings leading to exchanges and kinship (Lévi-Strauss, 1967: 100). These exchanges build the structure of society, and they rest on two main forms. There are restricted exchanges embedding two partners or a multiple of two, or generalized exchanges with far more partners (Lévi-Strauss, 1967: 159).

Sahlins operates in a similar way taking into account what Mauss describes as the system of total reciprocities made of the interrelations between the generalized, the balanced, and the negative reciprocities (Sahlins, 1972: 192–195). For Sahlins, these three reciprocities do not build a system. On the contrary, they form a continuum going from the extreme sociability (the generalized reciprocity) to the extreme unsociability (the negative reciprocity). Thus, the challenge of the gift is not to convince modern societies that they should rediscover these archaic structures of social life in order to empower societies with the obligations of the gift – in order to make it more solidary. It rather supposes, Sahlins suggests, that we examine how the reciprocity can be reproduced in our societies, a view which the contemporary literature on Mauss' gift will largely adopt (e.g., Case, 1976: 323–347; Price, 1978: 339–351; Weiner, 1980: 71–86).

The Gift as Expenditure and Scapegoats

Besides the structuralist cultural anthropology of Lévi-Strauss and Sahlins, the authors brought together around Roger Caillois and his *Collège de Sociologie* – a group of writers, ethnologists, and sociologists belonging to the surrealism built in the 1930s in France – also will return to Mauss' gift. Rather than to see in the gift the warranty of the social order, or the symbol of the reciprocal nature of human exchanges, they propose to consider it as revealing the irrational roots of societies and cultures. The works of Georges Bataille (1991) and René Girard (1986) particularly contribute to situate the gift within the framework of an economy of expenditure and sacral violence granted to the profit of society.

For Bataille, the gift is the symbol of a global economy primarily based on debts – because we are in debt toward the others and the natural environment in which we live, we have to give the gift. This does not erase our debts, rather on the contrary. It puts us in a cycle not of obligations, but of growing debts which lead to growing expenditure. According to Bataille, the gift symbolizes the structure underlying every economy which is in its essence an economy working on the basement of growing expenditures and debts which determine the development of our modern economy paradoxically pretending to be driven by the accumulation of profits.

Girard depicts a similar situation based on the often violent sacrifice of innocent actors described as 'scapegoats' which he finds out in every possible human society. Each society has to designate scapegoats who will be sacrificed, giving their life to the community in order for society to survive. According to Bataille and Girard, the existence of gifts shows that our economic and social structures are not resting on rationality. If gift has nothing to do with rationality, does it make sense to speak about the gift in a modern society dominated by the economic rationality and the financial transactions (Komter, 2005)?

The Gift as a Symbolic Market

Since the 1970s, the majority of authors consider against Bataille, Girard, and the surrealists of the *Collège de Sociologie* that the gift does not represent any irrational root of societies and cultures. On the contrary, it is a kind of exchange – in the broad sense of a social exchange, but also in a narrower sense of an economic exchange – whose rationality relies on the acquisition and the delivery of symbolic goods such as prestige values or altruistic qualities.

For this reason, Cheal (1988) and Camerer (1988) consider the gifts as a kind of economy which is not a market economy in a strict sense, but whose principles are similar to those of the market economy (Swedberg, 1990: 33). We offer and demand gifts, some gifts are carried out for altruistic reasons, others for egoistic ones (Stark, 1995). In this particular gift economy, the actors do not seek a material remuneration of their investment only. They also want to obtain a remuneration in a symbolic form – they want to be individually and socially upvalued by their gifts (Fehr and Gächter, 2000; Thomas and Worrall, 2002).

This symbolic rationality of the gift is at the origin of Pierre Bourdieu's ethnographical reflections on power relationships in North Africa. Bourdieu takes Mauss' gift into account in his description of gift exchanges in Kabylie (Bourdieu, 1972). After the work of René Maunier on the Kabyle *taoussa* (Maunier, 1927: 11–97) – a ceremonial exchange of money returned with interest – Bourdieu indicates that not only the *taoussa*, but every gift in Kabylie veils the power relationships between actors who do not have same economic and social capitals. According to Bourdieu, in the Kabyle society as in much of other ones, these power relationships are seldom apprehended as such. They remain dissimulated by the honor and the disinterest which the actors show when they exchange gifts. If there is any obligation to give in return, this obligation is never clarified, neither in person (who must give back?), nor in place

(where must the gift be returned?), nor in time (when must the gift be returned?). This is the reason why gifts are intrinsically ambiguous. They veil the interests of actors while exerting at the same time their honor and prestige. Everything looks like as if the Kabyle society refuses to face its own economic reality and social hierarchy in order to rather show them as controlled by laws depending on noble virtues. Hence, every gift rests on a spirit of calculation which is not different from the one expressed in the economic rationality of our societies. And as in our societies, the Kabyle actors spend a part of their efforts to systematically dissimulate this logic of interest and power which governs every exchange. In this sense, the gift is a mark of distinction between social groups bound to the affirmation of their nobility and social power.

The Gift as a Distinction Mark in Power Relationships

Parallel to economists and sociologists, several historians have also supported that gifts have always been distinction marks used in order to stabilize relationships of power within societies. Aristotle and Plato already depict gifts as associated as well with hospitality (Finley, 1988) as with the means by which the Greek tyrants establish diplomatic relations with their neighborhood (Van Der Mije, 1987: 241–267; Morris, 1986: 1–17). These exchanges of gifts also show up in societies with an important population and, for some of them, with a state. This appears more distinctively when we consider the history of gifts in the transition from classical antiquity to the Christian era. Paul Veyne, for instance, (Veyne, 1976) who writes on Roman cities in the time from 300 BC to AD 300 describes how they evolve from a use of the gifts restricted to the family members and the friends, to a use of gifts in broader circles, coming from the persons of rank in the city, and going to the population. Veyne calls these gifts *evergetism*, which is a neologism derived from the Greek terms *evergesia* and *evergetai*. The former means the desire to contribute to the well-being of the city. The latter depicts the civic attitude of benevolence of the persons of rank – the courtesy which Seneca describes as *beneficiis*. The *evergetism* secures the authority of the persons of rank in the city.

Peter Brown (1992) describes a similar phenomenon in the time from AD 300 to AD 450 regarding the relationship between the governors of the cities, their representatives, and the higher social classes. The honesty of the higher social classes toward the governors of the city and their representatives rests on the higher classes' expectation to have an impact on the political decisions using the patrician art of persuasion – the *paideia*. It is a benevolent way to influence them in order to make the political decisions following a noble art to rule the city. Brown indicates to which point the Christians govern this process using donations as the frame of the relationships between dominating and dominated social groups – for example, between rich and poor, or men and women – relationships which have been conceived similar to those toward God.

In the Dark Ages, the Renaissance and up to the French Revolution, gifts build the core of the relations between a population and its leaders. To give presents is associated with human qualities such as honesty or heroism, or with the grace (*gratia*) – which Raphaël represents in his painting 'The Three Graces' at the beginning of the sixteenth century. The grace remains an important societal topic at the end of the fifteenth century when the Spanish Kingdom stages Christopher Columbus' discovery of the American continent. The Amerindians and the objects brought back from America symbolize Columbus' power relationships to the Spanish monarchs (Vilches, 2004: 201–225). With the growing activities on economic markets, the donations are also often translated into amounts of money. If that does not make them loose their symbolic value – they symbolize the power, the honor, the reliability of relationships – these quantified donations nevertheless lead to a corresponding socioeconomic characterization of social groups (Howell, 2010). For example, if the European courts in the sixteenth and seventeenth centuries are willing to engage in more commercial activities, the nobility remains distant to economic exchanges not taking part in commercial activities in order to attest its social prestige. To the nobleman – the man of donation – goes the honor, to the tradesman and the salesman the money. This distinction is not a hermetic one because the tradesmen and the salesmen often solicit the nobility to get fortunate customers in order to make money (Kettering, 1993: 69–87). Nevertheless, the nobility claims for itself the gifts which are associated with the merit, the generosity, the honesty toward the monarchs who in return give them their grace to reward their loyalty. This relation of reciprocity and honor is for the nobility a confirmation of their social status, as well as a distinctive mark which makes them socially different from the upper- or middle-class rich *bourgeois gentillhomme*. The gift is, in this sense, against the modern economic market – a statement which will be brought to modern society in the 1980s by the interdisciplinary M.A.U.S.S. group around Alain Caillé (*Mouvement Anti-Utilitariste en Sciences Sociales – Anti-Utilitarian Movement in Social Sciences*).

The Gift against the Market Economy

Alain Caillé and the authors affiliated to his program – the M.A.U.S.S. – support the same consideration about the gift which is not irrational, but linked to honor, prestige, and nobility describing another kind of rationality. However, this rationality is not supported by any kind of calculation, and it is not the veil hiding the logic of interests transforming all of our social exchanges in power relationships. The gift does not describe another market economy. It supposes radical anti-utilitarian practices which are not compatible with the economy in the modern sense of this term (Caillé, 2000; Testart, 2001: 723; Anspach, 2002: 76). According to Caillé, antiutilitarianism is not only the radical criticism of the hegemony of economic rationality framing our daily life as well as the theoretical narratives of social sciences. It is not about to spend efforts and time in the solely critics of the economic semantics. It is rather a matter of delivering a perspective able to rank all social practices within the gift. The gift as a paradigm, so Caillé, is the reintegration of the market and the state in a social order made of gift exchanges which make sense for all social actors (Caillé, 1994). Pushed to the extreme, it is a cosmological principle and the axiom of the social theory, the origin of the unique paradigm which can be conceived and legitimated in the social sciences. In the framework of the

M.A.U.S.S., the gift consequently links to a wide range of contemporary problems, such as the donation of organs, sponsoring, technology transfers, build of civil associations, forms of national and international – governmental or nongovernmental – supports. As Mauss said, the gift is not an exotic character of the exchanges in foreign countries and cultures. It is at the heart of our modernity, and it represents the original structure of our social practices.

The Gift as an Impossible Exchange

In the 1980s, Jacques Derrida shows another position on the gift, assuming that, contrary to the thesis of the M.A.U.S.S. group, there cannot be any gift as a pure gift, and therefore, there is no possible gift. When we give something, we cannot give everything, we cannot give what belongs to us only, so that for Derrida, there is no gift in the strong acception of the term. The examples of time and death, his own ones and the ones of the others, lead Derrida to question the gift in the most radical way. If I can give the life or if I can kill others, I can neither live, nor die in their place, and I can neither give them eternity, nor immortality. It is the same regarding the gift. I can live the given, received, and returned gifts as events, but I cannot possess them. Even if I preserve a trace of them in me, this trace does not refer to these gifts, but only to me (Derrida, 1992: 47). Thus while giving, one can never be sure that the gift will be recognized by others as such. Therefore, it is unsure that the others receive it, return it, or give it further. The obligation, the duty which makes the circulation of the gifts happened only refers to the 'I' of the giver. He is responsible for the gift, and he cannot delegate his responsibility to others without being at the same time irresponsible. He cannot give without doing at the same time an irresponsible act leading to the negation of himself as 'I,' as the legitimated identity of the giver, without denying the gift. When he does it, then giving the gift turns into calculation which destroys the value of what is given and the act itself (Derrida, 1992: 104). Giving cancels the gift and makes it impossible.

Jean-Luc Nancy avoids this extreme conclusion while following a similar path to the one of Derrida. Derrida shows that the absolute gift does not exist. But it does not mean that the gift is not possible. One can even benefit from the criticism of Derrida in order to investigate how, if the gift cannot exist as an absolute gift, the phenomena occur in our daily life, that we name 'gift.' Taking this into account, we can then put an end to the myth of a pure gift (Nancy, 1988: 33), and we can also manage to deconstruct – as Marcel Hénaff proposes it – the relationships between gifts and economic exchanges which, as Hénaff suggests, maybe never have been bound together (Hénaff, 2002).

The Gift as an Ethnocentric and (Post)colonial Scheme

Derrida's deconstruction of the gift has also ramifications to very different theories supported by antiglobalization authors like Serge Latouche. Since 1990, Latouche casts doubts on the viability of a social theory using the gift (Latouche, 1998: 311–322). In all noncapitalist societies in the world, there is no true gift because there is nothing truly free, since nothing has to be paid. At the opposite, where one must pay for goods and services and where it is normal to do so, where the capitalist economic system pertains to almost all cultures and to all facets of the social life, there are gifts of various forms. Thus, the gift is not a strong paradigm – it is a Western phenomenon with the help of which one tries to ennoble and excuse the violence of the capitalist economy. Who defends the gift defends a (post)colonial kind of ideology linked to the market economy of industrialized societies. All in all, there is no viable alternative to capitalism apart from a firmly antiglobalization attitude. It can lead to give up helping countries in need of economic and technological support like Third World countries. But this attitude must favor their autonomous development. If another world is possible, it comes from the world of the others and not from our societies. The only thing which our societies must make is to express their solidarity while decreasing their life standards at the same level as the ones of these countries (Latouche, 2003: 18–19). The gift is dead, long live the gift!

Conclusion

Between economy and society, freedom and constraint, gratuity and interest, us and the others, the gift is a multiform and deeply ambivalent concept as early noted by Emile Benveniste regarding the ancient Anatolian verb *da* (Benveniste, 1966 [1951]: 315–327). The conclusions drawn from the first analyses on the gift show it – in proximity with Karl Polanyi (1944) and Alvin Gouldner (1960: 161–178) – as the point where morals and economy converge, and as the rock of all human exchanges. The contemporary analyses of the gift have deconstructed these certainties. Without denying the interest of the researches on the gift, they tend to relativize its obligatory and reciprocal characters, and they criticize the link between the gift and the idea of a morally viable social economy in industrialized societies. Is the gift carrying the fundamental standards of our exchanges? Or does the gift only show our blind idealization of these standards, unveiling our ethnocentrism when it comes to analyze the various features of social and cultural exchanges? These questions are likely to find neither an immediate, nor a final answer. Nevertheless, they contribute to underline a vast field of interdisciplinary problems which the history of the gift echoes while prolonging the reflections on the gift, namely that of the transformations of the forms of our exchanges in the history and through the cultures.

See also: Altruism and Prosocial Behavior, Sociology of; Mauss, Marcel (1872–1950); Potlatch in Anthropology; Prosocial Behavior, Cultural Differences in.

Bibliography

Anspach, M.R., 2002. A charge de revanche. Seuil, Paris.
Bataille, G., 1991. The Accursed Share 1: Consumption. Zone Books, New York.
Benveniste, E., 1966[1951]. Problèmes de linguistique générale, vol. I. Gallimard, Paris.

Bourdieu, P., 1972. Esquisse d'une théorie de la pratique, précédée de Trois Etudes d'ethnologie kabyle. Droz, Genève.
Brown, P., 1992. Power and Persuasion in Late Antiquity. Towards a Christian Empire. University of Wisconsin Press, Madison.
Caillé, A., 1994. Don, intérêt et désintéressement. La Découverte, Paris.
Caillé, A., 2000. Anthropologie du don: le tiers paradigme. Desclée de Brouwer, Paris.
Camerer, C.F., 1988. Gifts as economic signals and social symbols. American Journal of Sociology 94, 180–214.
Case, S., 1976. Social distance and exchange: the coast Salish case. Ethnology 15 (4), 323–347.
Cheal, D., 1988. The Gift Economy. Routledge, London.
Davy, G., 1920. La foi jurée. Alcan, Paris.
Derrida, J., 1992. Donner la mort. In: Rabaté, Jean-Michel, Wetzel, Michael (Eds.), L'éthique du don. Jacques Derrida et la pensée du don. Galilée, Paris, pp. 11–108.
Fehr, E., Gächter, S., 2000. Fairness and retaliation: the economics of reciprocity. Journal of Economic Perspectives 14, 159–181.
Finley, M.I., 1988. The World of Odysseus. Chatto and Windus, London.
Firth, R., 1940. The analysis of mana: an empirical approach. The Journal of the Polynesian Society 49 (196), 483–510.
Frazer, J.G., 1894. The Golden Bough. A Study in Comparative Religion. I&II. MacMillan, New York.
Girard, R., 1986. The Scapegoat. Johns Hopkins University Press, Baltimore and London.
Gouldner, A., 1960. The norm of reciprocity: a preliminary statement. American Sociological Review 25, 161–178.
Granet, M., 1920. La Polygynie Sororale et le Sororat dans la Chine Féodale. Étude sur les formes anciennes de la polygamie chinoise. Ernest Leroux, Paris.
Hénaff, M., 2002. Le Prix de la Vérité: Le don, l'argent, la philosophie. Editions du Seuil, Paris.
Howell, M.C., 2010. Commerce before Capitalism in Europe, 1300–1600. Cambridge University Press, Cambridge.
James, W., Allen, N.J. (Eds.), 1998. Marcel Mauss. A Centenary Tribute. Berghahn Books, New York.
Janssen, J.J., 1981. Die Struktur der pharaonischen Wirtschaft. Göttinger Miszellen 48, 59–77.
Kettering, S., 1993. Brokerage at the court of Louis XIV. The Historical Journal 36 (1), 69–87.
Komter, A., 2005. Social Solidarity and the Gift. Cambridge University Press, New York.
Latouche, S., 1998. Le don est-il l'autre paradigme? Revue du MAUSS 12, 311–322.
Latouche, S., November 2003. Pour une société de la décroissance. Le Monde Diplomatique, 18–19.
Lefort, C., 1979[1950]. L'échange et la lutte des hommes. Les Temps Modernes 6, 63–68, 1400–1417.
Lévi-Strauss, C., 1967. The Elementary Structures of Kinship. Bacon, Boston.
Lévi-Strauss, C., 1999. Introduction. In: Marcel Mauss, Sociologie et anthropologie. PUF, IX-LII, Paris.
Maunier, R., 1927. Recherches sur les échanges rituels en Afrique du Nord. Année Sociologique, Nouvelle Série 2, 11–97.
Mauss, M., 1999[1923/24]. Sociologie et anthropologie. PUF, Paris.
Morris, I., 1986. Gift and commodity in archaic Greece. Man 21 (1), 1–17.
Nancy, J.-L., 1988. Die undarstellbare Gemeinschaft. Patricia Schwartz, Stuttgart.
Polanyi, K., 1944. The Great Transformation. Farrar & Rinehart, New York.
Price, S., 1978. Reciprocity and social distance: a reconsideration. Ethnology 17 (3), 339–351.
Sahlins, M., 1972. Stone Age Economics. Aldin de Gruyter, New York.
Seligman, C., Seligman, B., 1911. The Veddas. Cambridge University Press, Cambridge.
Stark, O., 1995. Altruism and beyond: An Economic Analysis of Transfers and Exchange within Families and Groups. Cambridge University Press, Cambridge.
Swanton, J.R., 1910. Ethnological Problems of the Lower Mississippi Valley. Proceedings of the Mississippi Valley Historical Association 112–127.
Swedberg, R., 1990. Vers une nouvelle sociologie économique. L'évolution récente des rapports entre la science économique et la sociologie. Revue du MAUSS 9, 33–71.
Testart, A., 2001. Echange marchand, échange non marchand. Revue française de sociologie 42 (4), 719–748.
Thomas, J., Worrall, T., 2002. Gift-giving, quasi-credit and reciprocity. Rationality and Society 14 (3), 308–352.
Veyne, P., 1976. Le Pain et le cirque. Sociologie historique d'un pluralisme politique. Seuil, Paris.
Vilches, E., 2004. Columbus's gift: representations of grace and wealth and the enterprise of the Indies. Modern Language Notes 119 (2), 201–225.
Van Der Mije, S.R., 1987. Achiles' god-given strength. Mnemosyne XL (3–4), 241–267.
Weiner, A.B., 1980. Reproduction: a replacement for reciprocity. American Ethnologist 7 (1), 71–86.
Weiner, A.B., 1985. Inalienable wealth. American Ethnologist 12 (2), 210–228.

Global Agenda for Social Work and Social Development

Rory Truell, International Federation of Social Workers, Bern, Switzerland
David N Jones, People Need People Consulting, Northampton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Global Agenda for Social Work and Social Development (*The Agenda*) is being developed by three global, professional organizations – International Association of Schools of Social Work, International Council on Social Welfare, and International Federation of Social Workers. Launched in 2010 to strengthen the international profile of social work and social development, it enables social workers to make a stronger contribution to policy development. *The Agenda* identifies four fields for action by global institutions, local communities, and the three organizations: promoting social and economic equalities, promoting dignity and worth of peoples, working toward environmental sustainability, and strengthening human relationships. This article links *The Agenda* process with parallel developments within the worldwide social work professional community.

The Global Agenda for Social Work and Social Development (*The Agenda*) is being developed and promoted jointly by the International Association of Schools of Social Work (IASSW), the International Council on Social Welfare (ICSW), and the International Federation of Social Workers (IFSW) (Jones and Truell, 2012). It is designed explicitly to strengthen the profile and visibility of social work and social development, to help these organizations to acquire new partnerships, to boost the confidence of social workers, and to enable social workers to make a stronger contribution to policy development. *The Agenda*, as a document and a process, is responding to increased social complexity. It is designed to reposition the global social work professional community, together with social development professionals, as a core source of information about social problems and of creative responses to these. The aim is to ensure that the experience and skills of social professionals are recognized and applied to achieve sustainable, collaborative outcomes drawing on the acknowledged skills of social workers in creating multifaceted, pragmatic solutions to highly complex problems, both individual and social (e.g., see Sucharipa, unpublished; Sweifach et al., 2010; Dominelli, 2012).

The mission of the social work profession and the development of social policy are rooted in a set of core values (International Federation of Social Workers and International Association of Schools of Social Work, 2004). These core values, embraced by social workers, social work educators and social development practitioners, are the foundation of social work's unique purpose and perspective (Hugman, 1995; Chan, 2000; Banks and Nøhr, 2011). That perspective has its foundations in the frontline work of social workers, whose task is to respond to some of the most complex individual, family, and community challenges. There are many causes behind human problems, but they always involve an interweaving of personal and social factors. The development of the Global Agenda gives the opportunity to build upon past experiences and to articulate a new and crucial vision for engagement that can mobilize social workers, their representative organizations and their partners in the new global context.

Global Context

The worldwide recession, globalization of economic systems, heightened inequality within and between nations, extensive migratory movements, increased pandemics, climate change, natural catastrophes and new forms of conflict, compel social work practitioners, educators, and social development professionals to find new ways of responding to new realities (International Federation of Social Workers, 2004; Jones et al., 2004).

The impact of global crises has been generally acknowledged, for example, in the United Nations report 'The Global Social Crisis' (United Nations Department for Economic and Social Affairs, 2011) reflecting concerns identified in the report of the World Commission on the Social Dimension of Globalization (2004). This context was addressed in the foreword to the 2011 report by Sha Zukang (former Under Secretary General, Department of Economic and Social Affairs, United Nations and keynote speaker at the Hong Kong world social work conference 2010). He stated that it is essential for governments to "take into account the likely social implications of their economic policies. It has been shown time and again that economic policies considered in isolation from their social outcomes can have dire consequences for poverty, employment, nutrition, health and education, which, in turn, adversely affect long-term sustainable development" (United Nations Department for Economic and Social Affairs, 2011: p. iv). The Report also found that many governments did not pay enough attention to the social implications of the recent global financial crisis and urged that social investments be given priority in recovery programs.

Given this global context, the challenge for social work and social development is how best to build effective linkages between the global realities and the local community response (Pettifor, 2004; Lyons et al., 2006; Payne and Askeland, 2008; Healy and Link, 2011).

Geopolitical, social, and economic shifts also have their impact on social work and social policy. The predominance of western theories and practices is being challenged by articulate

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28035-5

practitioners and educators in the global south, creating lively philosophical and practical exchanges (see, e.g., Yip, 2004; Cox and Pawar, 2006; Netto, 2006; Brazilian Centre for Cooperation and Exchange of Social Services (CBCISS), 2008; Laird, 2008; Law and Gu, 2008; Hatta, 2009; Mwansa, 2010; Nikku, 2010; Ofreneoa, 2010; Rankopo and Osei-Hwedie, 2011; Truell, 2011; Abdullah, 2013).

Organizational Context

This Agenda is the product of a collaborative initiative undertaken by three international organizations representing social work practice, social work education, and social development. These international bodies grew out of the world social service conference in Paris in 1928 and have formal consultative status with the UN Economic and Social Council and other UN and related agencies.

The IASSW is an international community of schools and educators in social work, promoting quality education, training and research in the theory and practice of social work, administration of social services, and formulation of social policies. IASSW speaks on behalf of 2000 schools of social work including 500 000 students.

The ICSW is a global, nongovernmental organization which represents tens of thousands of organizations around the world that are actively involved in programs to promote social welfare, social development, and social justice.

The IFSW is the global federation of national social work organizations representing 100 countries with over one million social workers.

The three organizations have cooperated more or less closely over the 80 or more years of their existence. They jointly own the journal *International Social Work* and have organized several joint conferences regionally and globally. IASSW and IFSW have jointly published the Global Definition of Social Work (currently under review) (International Federation of Social Workers and International Association of Schools of Social Work, 2001), the statement of ethical principles (International Federation of Social Workers and International Association of Schools of Social Work, 2004), and the Global Standards for the Education and Training of the Social Work Profession (International Association of Schools of Social Work and International Federation of Social Workers, 2004). In the current context, they recognize the professional and organizational value of sharing strategy and common purpose. They recognize the need to link social work education and practice and sustainable social development into a single collaboration. They are also committed to promoting wider partnerships, such as the links with the Social Work Health Inequalities Network (Bywaters et al., 2009; Fish and Karban, 2013) and the Commonwealth Organization for Social Work, as well as with service user organizations.

History of the Agenda

The Agenda process started in 2004, arising from a series of parallel developments in the three global bodies. The IFSW Executive agreed that the Federation needed to develop strategies to provide clearer professional leadership in response to the changing social context and evidence of world-wide low morale and loss of confidence felt by social worker practitioners (Stevens and Higgins, 2002; Jones, 2005; Depanfilis and Zlotnik, 2008; Vyas and Luk, 2011). At the same time, IASSW was exploring strategies for increasing global influence and ICSW was changing the relationship between global conferences and its advocacy strategy. To pursue each of these goals, the three organizations had decided to hold joint world conferences and started planning a major conference in Hong Kong in 2010. Building global partnerships and planning global conferences take time, involving internal democratic processes and relationship building.

Having made the agreement, the three organizations sustained the commitment and gradually developed plans and a shared analysis. The 2010 conference program was shaped to support the development of a world social agenda and the three organizations began examining how the process could be supported and sustained.

Background papers were commissioned and published on the website (see IFSW website) and existing partnerships with UN and other bodies were used to explore the evolving ideas. Members and organizations were invited to examine the papers and to come prepared to participate in the strategic developments. The organizations consulted with social workers, social work educators, and social development practitioners.

When *The Agenda for Social Work and Social Development* process started in 2004, it had not been envisaged that the initiative would be launched in a time of major financial and social crisis. When around 3000 social work practitioners, educators, social development professionals, and policy makers came together in the Hong Kong 2010 world conference (see conference website) the new global economic and social reality was becoming clear and the need for shared action seemed even more necessary and urgent. Rapporteurs collated ideas emerging from the conference workshops and presented the outcomes in the closing session. The three Global Rapporteurs drew the elements together and, working with the three Global Presidents, presented themes and strategies to the closing session.

The following four pillars of *The Agenda* emerged and were endorsed:

- promoting social and economic equalities,
- ensuring the dignity and worth of the person,
- promoting sustainable communities and environmentally sensitive development, and
- promoting well-being through sustainable human relationships.

This formulation was a minor refinement of the four pillars which emerged in Hong Kong and was agreed at the Accra meeting in 2011. A commitment to ensuring an appropriate environment for practice and education was added later. The *Global Agenda* received overwhelming support from delegates, who supported the launch of a global movement led by IASSW, ICSW, and IFSW.

The preamble to *The Agenda* which emerged from this process stated its aims as "developing multiple platforms to foster the emergence of a Global Agenda for the profession

through the shared, collective voice of its members." The shared intention was to engage the conference participants and those not able to attend in a new form of dialog which would identify agreed priorities and shape the action plans of each organization. This was recognized as an ambitious and daring process, with significant risks, but the conference evaluation showed that it had succeeded.

The three founding organizations represent only the beginning of the collaboration. It is intended that further alliances will be built with likeminded movements to maximize the impact of creating positive social change. It is recognized that such change will be most likely to occur when common strategies link *The Global Agenda* with regional and national priorities.

Following the conference, the text of *The Agenda* and the supporting documents and action plans were posted on the website and widely distributed throughout the organizations. The core themes identified by the 3000 participants in Hong Kong were subject to extensive, global debate and feedback, in particular widespread debates in schools of social work and social work agencies on World Social Work Day 2011. The responses broadly endorsed the four priority areas.

The final round of postconference consultation culminated in a tripartite meeting in Ghana in November 2011, which focused on actions needed by each organization and the partnership to ensure a strategic focus on the four priority areas, resulting in a renewed determination to promote social work and social justice. The representatives agreed that:

- the full range of human rights is available to only a minority of the world's population;
- unjust and poorly regulated economic systems, driven by unaccountable market forces, together with noncompliance with international standards for labor conditions and a lack of corporate social responsibility, have damaged the health and well-being of peoples and communities, causing poverty and growing inequality;
- cultural diversity and the right to self-expression facilitate a more satisfactory intellectual, emotional, moral, and spiritual existence, but these rights are in danger due to aspects of globalization which standardize and marginalize peoples, with especially damaging consequences for indigenous and first nation peoples;
- people live in communities and thrive in the context of supportive relationships, which are being eroded by dominant economic, political and social forces;
- people's health and well-being suffer as a result of inequalities and unsustainable environments related to climate change, pollutants, war, natural disasters, and violence to which there are inadequate international responses.

Recognizing these realities, the representatives formulated key objectives related to each of the four key themes of *The Agenda* which had been agreed in Hong Kong (see above). Each theme targets commitments to three areas. The first set of commitments focuses our joint activities on presenting a social work and social development perspective in our work at the United Nations and other international agencies. The second set of commitments recognizes the importance of strong and resilient communities to achieve stable well-being and the importance of the role of social work and social development practitioners in facilitating healthy and strong communities. The third set of commitments relates to the internal activities of our own organizations, directed toward ensuring that policies and standards are consistent with addressing the root causes of poverty and oppression and promoting sustainable social environments which make a reality of respect for human rights and dignity. Finally, the platform recognized the significance of education and training and of the working environment for effective and ethical social work practice, and includes commitments to coordinate research and activity to improve these elements.

The Global Agenda for Social Work and Social Development: Commitment to Action was formally released in the week of World Social Work Day and UN Social Work Day in March 2012 (International Federation of Social Workers et al., 2012) at simultaneous events in Geneva and New York.

The Agenda was received on behalf of the United Nations by Helen Clark (Head of the United Nations Development Program). She commented:

> It is an honour for me to accept the Global Agenda for Social Work and Social Development on behalf of the UN, and to congratulate the International Federation of Social Workers, the International Association of Schools of Social Work, and the International Council on Social Welfare on developing The Agenda.
>
> I understand that the Global Agenda was adopted with the overwhelming support of delegates to the joint conference of your three organisations in Hong Kong. That suggests that there is a widely shared and clear vision across the international social worker community, which gives the Global Agenda great credibility…
>
> We too are pledged to address the root causes of poverty, oppression, and inequality. We support the call in the Global Agenda "to create a more socially-just and fair world which we will be proud to leave to future generations"…
>
> For more than a century, the social work profession has been at the forefront of promoting human rights and supporting people to realize their full potential. Various arms of the United Nations have worked with your organisations in global fora, and alongside social workers in-country to advance progress on the Millennium Development Goals, engage citizens in development, and strengthen social protection systems…
>
> …the formal presentation of your Global Agenda also reminds us of the shared vision and values we have, and of the importance of working together for a better world. (Clark, 2012)

The Agenda was also received in the Geneva office of the United Nations by Assane Diop, International Labour Organisation (ILO) Executive Director. He referred specifically to the Social Protection Floor (SPF) Initiative and said that large numbers of social workers and social development workers were needed to help:

> Create the political will to build the Social Protection Floor
>
> Create public awareness about the need for the social protection floor, and
>
> Design, implement and monitor the Social Protection Floor. (Diop, 2012)

Mr Diop said there was "formidable support for the mandate of the Global Agenda" in the United Nations. He

further said that social protection in all aspects of social and economic life was a necessity for all people to live in dignity and that the SPF should be visible in all the countries of the world.

The Agenda has also been received by the United Nations in Nairobi, Santiago de Chile and Bangkok and by a number of international bodies including the African Union, Commonwealth Foundation, Council of Europe, and Social Platform of NGOs in Europe.

The Stockholm Joint World conference in July 2012, commissioned by IFSW, IASSW, and ICSW, enabled participants to focus on implementation of *The Agenda* and strategies for action (see conference website). The three Presidents, in separate speeches on behalf of the three organizations, restated the commitment to joint work on *The Agenda* and to monitor its implementation.

The next stage of the implementation strategy includes a commitment to launch and promote a Global Observatory, drawing from a global network of regional centers, to support implementation of *The Agenda*, research its impact and study the work environments which promote positive outcomes in social work and social development. The plans for the Observatory were agreed at a tripartite meeting in Paris in November 2012 and launched in January 2013. The Observatory will present its first report to the 2014 world conference in Melbourne, drawing on regional and national reports, focusing on the first Agenda theme – promoting social and economic equality. This theme is also the focus of World Social Work Days in March 2013 and 2014. The report will be published, to coincide with the conference, as a supplement to a special issue of International Social Work which will include papers on social and economic development.

The Impact of the Agenda on the Social Work Profession

The objectives set out in *The Agenda* are having and will continue to have far reaching implications for the social work profession. Each of the five objective areas of *The Agenda* specifies that each organization will adjust its policies, and other official documents, to ensure they are consistent with *The Agenda*.

IFSW, for example, is undergoing strategic change as *The Agenda* evolves the professional focus and culture of international social work. This is in part due to the outward looking nature of *The Agenda*, with a strategic commitment to use new media, to consult and involve social workers, to plan and coordinate global activity more effectively, and to engage systematically with UN agencies and other partners. Whilst the Federation has had many past successes working with others to influence global policy statements, such as the UN Declaration on Indigenous Peoples and work on extreme poverty and human rights, there has not been an overarching policy framework and strategy. In practice, the international profession has tended to be more focused on developing the identity and professional status of social work at global level. This was demonstrated in the official Aims of the Federation, which stressed the establishment of national organizations of social workers and bringing the organizations together for the exchange of ideas and the formulation of polices.

The Agenda objectives clearly indicate a new set of political intentions for the three global organizations. The strategy more explicitly focuses on changing the dynamics that affect the people who use social work services, as demonstrated in *The Agenda* commitments document:

> We commit ourselves to supporting, influencing and enabling structures and systems that allow people to have power over their own lives.
>
> We commit ourselves to supporting, influencing and enabling structures and systems that positively address the root causes of oppression and inequality.
>
> We commit ourselves wholeheartedly and urgently to work together, with people who use services and with others who share our objectives and aspirations, to create a more socially-just and fair world that we will be proud to leave to future generations.

The Influence of the Agenda on Global Policies

The Agenda compels social work professional organizations and social workers to advocate a social work perspective in the political systems that affect the well-being of their clients. For example, writing in *International Social Work*, Truell (2012) called for changes to the trade system to regulate economic activity, in the context of reflections on the Occupy demonstrations:

> Through The Agenda Commitments (2011), The International Federation of Social Workers (IFSW), The International Association of Schools of Social Work (IASSW) and The International Council on Social Welfare (ICSW) supports the International Labour Organization's call for international standards to regulate the global workforce conditions and trade arrangements to create a socially just and sustainable global economy. …Socially-just global economic frameworks would be based on: human rights, fair pricing, international standards of labor conditions, corporate social responsibility, capacity building, agreed forms of dialogue underpinning supply and demand agreements – enabling all parties to participate, and enhancing communal rights and benefit sharing. (Truell, 2012: p. 593)

Another example of the impact of *The Agenda* is seen in the IFSW announcement of a conference in South Africa that will 'give voice to peoples who are targeted by social development policies and for their voices to influence the policy development frameworks' (see conference website). The conference scheduled for 2013 was planned to influence the policies and structures that will replace the UN Millennium Development Goals when they expire in 2015.

The Agenda partners have also been actively influencing the United Nations by issuing joint statements through their formal consultative status. One such example is a statement on *social and human rights questions: social development* (International Association of Schools of Social

Work and International Federation of Social Workers, 2012):

> As Social Workers, Educators and Social Development practitioners we witness the daily realities of people living in poverty and the multiple systems that alleviate hunger, ill health and exploitation. We therefore make the following statement to UN member-states that are responding to their people living in poverty and to member-states that benefit from the exploitation of impoverished countries. 1, People cannot be developed and managed by others: Our frontline experience has taught us that to escape from poverty people need to be actively involved in their own futures. They need to be able to contribute to best practice ideas, and as leaders, they need to be able to form collaborative engagements with funding bodies, managers of resources and other systems for support. 2, Basic minimum levels of social protection: As global bodies representing international social work and social development we know that poverty is worst where there are no social protection systems covering the lifespan. We support the UN's call for a minimum social protection floor to be built in every country. 3, Poverty eradication through systemic responsibility: As social workers and social development practitioners we understand that in many instances poverty is sustained through exploitive systems. We advocate for trade systems to be based on human rights, fair pricing, international standards of labor conditions, corporate social responsibility, capacity building, agreed forms of dialogue underpinning supply and demand agreements – enabling all parties to participate, and benefit sharing.
>
> We, as frontline workers, social work educators and members of affected communities commit ourselves to assist in bridging the links between people living in poverty, national governments, regional authorities and international frameworks. We welcome working with UN member-states and the UN agencies to fulfill the vision of eradicating poverty and fulfilling the MDGs.

Another regional example of translating the objectives of *The Agenda* into political advocacy was the Statement on Austerity submitted by IFSW to the European Union in 2013 jointly with the IFSW member organizations living through the negative effects of austerity (International Federation of Social Workers, 2013). This statement highlighted the need to ensure people's well-being was not subject to the whims of the economic market:

> This is not a crisis of the Euro; it is a crisis for people. The EU's market driven strategy of pursuing 'austerity' and 'competition' is not producing the anticipated 'growth'. Instead it is producing entrenched poverty, placing the peoples of the affected countries at risk of civil violence, and critically undermining the possibilities for rebuilding the affected societies.
>
> The EU directions are driving massive job losses, price increases, economic/social and personal depression, loss of the skilled workforce, business failure and increased inequality. Further, as people have lost access to adequate housing, health and education services critical problems are multiplying and will have lasting implications for the future generations. Such environments, as expected, have seen the significant growth of extreme right wing political parties/organizations that attack the vulnerable, disabled and immigrants, and hold the potential to destabilize democratic processes.
>
> As international professionals who see the results of both strong and weak communities and societal structures, we witness that resourced and confident communities are the cornerstone of a stable and thriving economy. Businesses want to operate in environments where the workforce is skilled, secure and reliable and there is sustainable demand from consumers. We therefore call upon the EU to abandon austerity and free market approaches and urgently to introduce new processes that bring the affected parties together to establish long-term and planned solutions, which emphasize: People living in coherent, stable and equitable environments, and sustainable economies.

Concluding Remarks: Building Capacity and Impact

The 2014 Global Observatory report will be a significant landmark for the three organizations and global social work and social development. This first global report will consolidate *The Agenda* strategy and provide new avenues for interacting with the United Nations and other global policy bodies. It is envisaged that the biannual reports will increase in significance, not only within the social work community but also in wider social policy arenas. To achieve this ambitious objective, a more robust infrastructure needs to be developed; discussions are beginning on funding for a global observatory structure which will be owned by the three organizations, with the potential to involve new partners.

The process of developing *The Global Agenda* has re-established the connection between the practice of social work, wherever that might be in the world, the education of social workers and the development of social policy at national, regional, and global levels. Making statements on social policy is not new for international social work organizations, but *The Agenda* platform, which was developed by social workers throughout the world, gives new energy and commitment for taking such political positions, with and alongside the people who use social work services. *The Agenda* represents the beginning of a long-term commitment to advocate for new social systems that address the root causes of poverty and oppression and to challenge the social forces which erode personal well-being. It is recognized that many more policies and action platforms, underpinned by the values of social work, will be needed to ensure that *The Agenda* has impact in daily practice and on the education and training of tomorrow's social workers.

Acknowledgments

The authors acknowledge the contributions of the Joint Coordinators of The Global Agenda process – Abye Tassé with Lena Dominelli (IASSW) and Charles Abbey with Sergei Zelenev (ICSW) and, the Presidents and other colleagues from the three global organizations.

See also: International Council on Social Welfare; International Association of Schools of Social Work (IASSW); Social Work Theories.

Bibliography

Abdullah, S., 2013. Multicultural social work and national trauma: lessons from South Africa. International Social Work (Online) http://isw.sagepub.com/content/early/2013/01/24/0020872812461019.abstract?papetoc (accessed 01.03.13.).

Banks, S., Nøhr, K. (Eds.), 2011. Practising Social Work Ethics around the World. Routledge, London.

Brazilian Centre for Cooperation and Exchange of Social Services (CBCISS), 2008. The dynamics of welfare in globalization: lessons from the past, challenges for today and tomorrow. 33rd ICSW Global Conference: Brazilian Preparatory Conference Meeting. CBCISS, Rio de Janeiro.

Bywaters, P., Mcleod, E., Napier, L. (Eds.), 2009. Social Work and Global Health Inequalities: Practice and Policy Developments. The Policy Press, Bristol.

Chan, C., 2000. Chinese culture and values in social work intervention. In: Tan, N.-T., Envall, E. (Eds.), Social Work around the World. IFSW, Bern.

Clark, H., 2012. Helen Clark: Opening Remarks on the Occasion of World Social Work Day. United Nations Development Programme, New York. http://www.undp.org/

content/undp/en/home/presscenter/speeches/2012/03/26/helen-clark-opening-remarks-on-the-occasion-of-world-social-work-day (accessed 22.08.13.).
Cox, D., Pawar, M., 2006. International Social Work: Issues, Strategies and Programmes. Sage Publications, New Delhi.
Depanfilis, D., Zlotnik, J.L., 2008. Retention of front-line staff in child welfare: a systematic review of research. Children and Youth Services Review 30, 995–1008.
Diop, A., 2012. Opening Remarks on the Occasion of Social Work Day at the United Nations, Geneva. http://www.youtube.com/watch?v=0BgCCkRnJgA&;feature=youtu.be (accessed 22.08.13.).
Dominelli, L., 2012. Green Social Work: From Environmental Crises to Environmental Justice. Polity Press, Cambridge.
Fish, J., Karban, K., 2013. Health inequalities at the heart of the social work curriculum. Social Work Education: The International Journal (Online) http://www.tandfonline.com/doi/abs/10.1080/02615479.2012.742502 (accessed 10.01.13.).
Hadorn, M., 2012. Wir müssen beweisen, dass Soziale Arbeit effektiv ist: UN-Welttag der sozialen Arbeit in Genf - David N Jones präsentiert die "Global Agenda". Sozial Aktuell 2012 (5), 40–41.
Hall, N. (Ed.), 2006. Social Work: Making a World of Difference. Social Work around the World. IFSW/FAFO, Bern, Oslo.
Hatta, Z.A., 2009. The Sufic paradigm for stress management and its contribution to a model of social work practice. Asian Social Work and Policy Review 3, 113–121.
Healy, L.M., Link, R.J. (Eds.), 2011. Handbook of International Social Work: Human Rights, Development, and the Global Profession. Oxford University Press, Oxford.
Hugman, R., 1995. Ethical Issues in Social Work. Routledge, Abingdon.
Hugman, R., 2010. Understanding International Social Work. A Critical Analysis. Palgrave Macmillan, Basingstoke.
International Association of Schools of Social Work, International Federation of Social Workers, 2005. Global Standards for the Education and Training of the Social Work Profession. IASSW/IFSW, Bern. http://ifsw.org/policies (accessed 08.04.13.).
International Federation of Social Workers, 2004. International Policy Statement on Globalisation and the Environment. International Federation of Social Workers, Bern. http://www.ifsw.org/en/p38000222.html (accessed 21.04.06.).
International Federation of Social Workers, 2012. Social and Human Rights Questions: Social Development. Statement submitted by International Federation of Social Workers, a non-governmental organization in consultative status with the Economic and Social Council. United Nations, New York. http://daccess-dds-ny.un.org/doc/UNDOC/GEN/N12/312/05/PDF/N1231205.pdf?OpenElement (accessed 08.04.13.).
International Federation of Social Workers, 2013. Statement on Austerity Presented to Commissioner László Andor, European Commissioner for Employment, Social Affairs and Inclusion. International Federation of Social Workers, Bern (accessed 08.04.13.).
International Federation of Social Workers, International Association of Schools of Social Work, 2001. Statement on the Definition of Social Work. International Federation of Social Workers, International Association of Schools of Social Work, Bern. http://ifsw.org/policies (accessed 08.04.13.).
International Federation of Social Workers, International Association of Schools of Social Work, 2004. Statement of Ethical Principles. International Federation of Social Workers, Bern. http://ifsw.org/policies (accessed 08.04.13.).
International Federation of Social Workers, International Association of Schools of Social Work, International Council on Social Welfare, 2012. The Global Agenda for Social Work and Social Development: Commitment to Action. http://www.globalsocialagenda.org (accessed 20.03.12.).
International Labour Organisation, 2013. Jobs, Growth and Social Justice. International Labour Organisation, Geneva.
Jones, D.N., 2005. Social work on the global stage: time to assert our vision and self confidence. Professional Social Work.
Jones, D.N., 2011. Global Agenda For Social Work and Social Development. Crossing Borders: Kyosei and Solidarity. T. Akimoto, Tokyo, Japan, pp. 61–65, 73–83. (accessed 24.08.12.).
Jones, D.N., Johannesen, T., Dodds, I., 2004. International social work, globalization and the environment. In: Tan, N.-T., Rowlands, A. (Eds.), Social Work around the World III. International Federation of Social Workers, Bern.
Jones, D.N., Truell, R., 2012. The global agenda for social work and social development: a place to link together and be effective in a globalized world. International Social Work 55, 454–472.
Laird, S.E., 2008. Social work practice to support survival strategies in sub-Saharan Africa. British Journal of Social Work 38, 135–151.
Law, A.K.-C., Gu, J.X., 2008. Social work education in mainland China: development and issues. Asian Social Work and Policy Review 2, 1–12.
Lyons, K., Manion, K., Carlsen, M., 2006. International Perspectives on Social Work. Palgrave Macmillan, Basingstoke.
Lyons, K., et al. (Eds.), 2013. The SAGE Handbook of International Social Work. Sage Publishing, London.
Mwansa, L.-K.J., 2010. Challenges facing social work education in Africa. International Social Work 53, 129–136.
Netto, J.P., 2006. Ditadura e serviço social. Cortez, São Paulo.
Nikku, B.R., 2010. Social work education in Nepal: major opportunities and abundant challenges. Social Work Education 29, 818–830.
Ofreneoa, R.E., 2010. Will Asia and the world waste the crisis? Policy coherence and social protection for all (Special Issue: Social Security and Development in East Asia). China Journal of Social Work 3, 5–17.
Payne, M., Askeland, G.A., 2008. Globalization and International Social Work: Postmodern Change and Challenge. Ashgate, London.
Pettifor, J.L., 2004. Professional ethics across national boundaries. European Psychologist 9, 264–274.
Rankopo, M.J., Osei-Hwedie, K., 2011. Globalization and culturally relevant social work: African perspectives on indigenization. International Social Work 54, 137–147.
Stevens, M., Higgins, D.J., 2002. The influence of risk and protective factors on burnout experienced by those who work with maltreated children. Child Abuse Review 11, 313–331.
Sucharipa, E., 2001. Managing Conflicts on the International Level: Can We Draw Parallels with Social Work? Danube Conference: Managing Conflicts in Social Work, 30 August 2001, Vienna/Bratislava, unpublished.
Sweifach, J., Laporte, H.H., Linzer, N., 2010. Social work responses to terrorism: balancing ethics and responsibility. International Social Work 53, 822–835.
Truell, R., 2011. 2nd Latin American and Caribbean Meeting 2011 Held in Mendoza, Argentina. IFSW, Bern.
Truell, R., 2012. News and views from IFSW international social work. International Social Work 55, 593–594.
United Nations Department for Economic and Social Affairs, 2011. The Global Social Crisis: Report on the World Social Situation 2011. United Nations, New York. http://www.un.org/en/development/desa/publications/report-on-the-world-social-situation-2011-the-global-social-crisis.html (accessed on 22.08.13.).
Vyas, L., Luk, S., 2011. Frazzled care for social workers in Hong Kong: job stress circumstances and consequences. International Social Work 54, 832–851.
World Commission on the Social Dimension of Globalization, 2004. A Fair Globalization: Creating Opportunities for All. International Labour Office, Geneva.
Yip, K.-S., 2004. A Chinese cultural critique of the global qualifying standards for social work education. Social Work Education 23, 597–612.

Relevant Websites

http://www.globalsocialagenda.org – Global Social Agenda.
http://www.swsd2010.org – Hong Kong 2010 Conference.
http://www.ifsw.org/get-involved/agenda-for-social-work – Hong Kong 2010 Papers.
http://www.iassw-aiets.org – IASSW.
http://www.icsw.org – ICSW.
http://www.ifsw.org – IFSW.
http://www.ilo.org – International Labour Org.
http://www.un.org/millenniumgoals – Millennium Development Goals.
http://www.ifsw.org/get-involved/voices-for-development-conference – South Africa 2013 Conference.
http://www.swsd-stockholm-2012.org – Stockholm 2012 Conference.
http://www.undp.org – UN Development Program.
http://www.who.int – World Health Organization.

Global Development: Markets and Wellbeing

Alberto Arce, Wageningen University and Research Centre, Wageningen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The relationship among global development, markets, and well-being is an important theoretical and methodological association to address and analyze theoretical ways of 'value framing and talking' among experts, institutions, and people. These relations may take the forms of interfaces, dislocations, and disjunctions, which are expressed under the notion of 'new materialities', encompassing the encounter of social and biological sciences. Main social concerns still are civic manifestations, knowledge encounters, and policy dilemmas present on contemporary global modernities. This interrelatedness of concepts, ethnography studies, and imagined spaces represents a cosmopolitan configuration that focus on the reflexivity of development studies and direct us to critically reconsider dualistic ways of thinking about the world. Today actors involve themselves arduous laboring of becoming engaged in development processes through social action to organize and materialize their meaning, values, and new subjectivities, as an expectation of becoming 'new' social actors. This orientation reaches the core of the anthropology of development endeavor.

Introduction

Conceptually, global development is an important representational site in which the meaning and intent of governance and development policies can be critically grasped and analyzed. The language of 'good government' and 'the market' (Stirrat, 1992) is one instance of this representation. During the 1990s, official discourses equated markets with democracy and heralded this relationship as part of a contemporary cosmopolitan configuration of empowerment; one based on market forces and consumer choice as a source of social inspiration to construct a new subjectivity – the enterprise spirit – that could shape the identity and economic relevance of social groups.

Well-being studies have encompassed the importance of material and nonmaterial resources in processes of development (Gough and McGregor, 2007). This orientation reaches to the core of the anthropology of development endeavor: Equal attribution to the importance of material and nonmaterial aspects of social life invigorates understanding of how people's experiences reflect a fast changing world. This permits explanation of how actors' expectations and struggles to attain development and overcome poverty are achieved. This orientation focuses on questions of reflexivity in development studies, including the researcher's own encounters and experiences with the lifeworlds of actors in the development arena. Recognizing that there are some actors wishing to reshape and others to contest policy measures that threaten the livelihood or human rights of human populations around the world provides the possibility of constructing ethnographies that are concerned with the depiction of particular actions and values (Graeber, 2001) for territorializing and reterritorializing elements of multiple 'modernities and traditions'.

This article takes a broad view of what constitutes the theoretical and methodological *bearings* of the association of global development, markets, and well-being. This relation is seen as centrally connected to theoretical ways of talking about the manifestations and dilemmas of contemporary global modernities (Arce and Long, 2000). The frequent emphasis placed on the interaction between discourse and social action in texts that explore the makeup of globalization processes and local responses can be traced to the contribution of Giddens' theory of structuration (1979, 1984) and to his conceptualization of structure and agency and to how this shapes, and is itself shaped by, 'modern' social and individual life (1990).

This theoretical orientation facilitated the counterrepresentation of various forms of 'modern economic development' as freedom (Sen, 1999) and to appreciate the capability for actors to perform the lives that they themselves value (Long, 2001). This has brought to the fore new ways of framing how actors are enacting the mainstream economic political development and institutional challenges. This directed anthropology to interrogate modes of intervention (Mosse, 2005) and civic participation (Foweraker, 2001; Fox, 2007; Arce, 2010), this clearly ensure the critical depiction of the advance of democratic values (Paley, 2002) and rights (Merry and Goodale, 2007). As 'world audiences' interests lead to an increasing interrelatedness of global development, markets, and well-being, since the end of the 1990s, actors started to perform actions to meet a wide-range of basic needs. These actions became socially relevant, as civic participation came to equal the importance of global developments and markets (Carrier, 1997). These contemporary associations challenged lineal cause-effect methodologies to explain processes of development and extend the significance and 'language' of markets to rethink types of development and poverty.

Our concern here is with the conditions of possibilities for the democratization of research about global development in the context of an ethnographic methodology, critically engaged with academic and policy dominant orientations, especially those organized by the social sciences in the West (Appadurai, 2003). Thus, there is a need for researchers to rethink critically the nature of anthropological studies in a global era, and the kind of contributions that anthropology of development can make to actors and their well-being.

Methodological Encounters and Tribulations

A glimpse at the contemporary literature on development and social change indicates that development subjects are not a mechanical outcome of apparently objective circumstances. That is the global associations between events apart in space and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10152-7

adjacent in time are uneven, fragile, and heterogeneous in contemporary social change. People living in different zones have come to realize that the world and their zones are not unified systems. These zones cannot be controlled and administered by stationary technologies, the conservation and protection of classical styles of organization, the domestic shelter of economic activities, the regulation of transformations in the international markets, and the use of new large-scale mediation policies among international, national, and regional contexts.

In their attempts at interpreting these global developments, social scientists have recognized these new range of conditions, uncertainties, and sociopolitical responses. And in this manner, states, transnationals, markets, technologies, and global entities themselves enter (in deterritorialized and territorialized forms) actors' zones, where the reordering of social processes is taking place. This generates conflict and negotiations. However, this reality is quite different from reports of global conflict on Internet media and television. Hence, these parallel sets of global representations of people, conflicts, and 'realities' introduce a high degree of tension in individual actors 'global consciousness' and their sense of multiple world realities (Beck and Grande, 2010).

Underpinning the description and analysis of these encounters are methodological advancements over disciplinary concerns (Rabinow et al., 2008). These have come to be recognized as the implicit sociopolitical characteristics of science and technology is continually generating new discourses that are demanding new accounts of agency and practices implicated in a complex field of relationships that may take us to think afresh notions of development, markets, and well-being. The recent surge in interests of matter and material experience (Coole and Frost, 2010) contributes to reshape the development field, as an activity rather than just a ground for representations and it has yielded a varied and colorful mosaic of ethnographic fieldwork experiences, directing us to reconsider dualistic ways of thinking about the world. Hence an anthropology of development rests on the malleable boundaries between social anthropology and sociology in an attempt to work together with natural science fields, such as agriculture, climate, ecology, and nutrition.

The inspiration of the new materiality issues has resonance with a renewed emphasis on the value of ethnography and a contribution to a relevant agenda for an anthropology of development. Our focus on development is characterized by both 'outside' intervention and 'inside' creativity. This perspective orients us to consider development interactions, as an assemblage (Ong and Collier, 2005) of socioeconomic, cultural, environmental, and biological conditions and an engagement with the new cosmopolitanism (Werbner, 2008), rather than a priority with the physicality of emerging processes, enabling the presence of global and localized discourses to exteriorize social forms of contemporary 'material realities'. The challenge is then how to approach analytically these entities and 'new' realities (De Landa, 2009).

Entities and Actors' Subjectivities

The existence of entities and realities (global development, markets, and well-being) are repositioning and reshaping actors' subjectivities. Some of these are presented as new political alliances between science and the social (Coole and Frost, 2010; De Landa, 2009). Others, introduced as innovative forms of capitalist coordination, are carving indenture pockets of control across physical distances, while dispossessing people by the privatization of public assets (Harvey, 2003). These associations are displayed as new and flexible development strategies of 'mature' capitalism, but also these are seen as a manifestation of forms of contemporary managerial failure. The inability to control agency in the functioning of a fluid world of various flows and ebbs of discourses, people, objects, and images is turning around values and the material world, repositioning, commodities, people's identities and the material organization of the world, precipitating various kinds of actions, interactions, performances, and enactments.

One of these entities is quality. This sort of entity which occurs around social relations is generating contests among public, private, and markets goals. Quality is and emergent entity around the creation of new spaces that affect actors' feelings and emotions. The intensification of commoditization (Scheper-Hughes, 2005) attached emotions of lost security, justice, and fairness to 'pure reason' searching for economic profitability and we are witness to new an unexpected assemblage of everyday life.

The recent concern in social research with consumer awareness of food risk has added a social relevance impetus to the notion of quality. The apparent reduction of quality in modern food has generated a significant body of critical research on the growth of alternative food markets. As consumer reaction to industrialized food circuits has increased, fears over contamination have risen, as have concerns over social justice and environmental sustainability (Kloppenburg et al., 2000; Allen et al., 2003; Trauger, 2007). Modern convenient products may imply industrial and chemical food processing accompanied by impacts on the rise of non-communicable diseases (NCDs), such as those associated with food and nutrition, contamination, and the environment. Despite a growing consensus on the global rise of obesity and concerns over future health consequences, such as cardiovascular diseases and diabetes, there is far less international agreement on the appropriate public policy response.

This is an example that highlights consumer concerns, however, these would be deficient if an appreciation of entities such as quality are missing from an understanding how these markets' 'fears' have resulted in a 'real' and 'fragile' world for consumers, where commodities are enabling insecurity and may reduce the well-being of individuals' health. In the absence of suitable market conditions for quality food consumption, the formulation of quality standards is seen as the only way forwards; a governance device to promote diverse and at times contradictory sets of scientific, social, ethical, and policy issues. These are bureaucratic and hierarchical attempts to manage and control assemblages of resources, commodities, and people's demands for food. The recognition that a consumer stunted by malnutrition or able to eat in excess, particularly 'modern' foods to an undesirable state of obesity, both are part of emerging materialities clustered as the field of NCDs. These diseases are starting to be recognized, as reducing the potential of people for development and centrally important in decreasing actors' well-being – quality of life – in a physical

sense. These materialities are molded from a world surrounded by constant ambiguity and in continuous institutional conflicts, which blur existing private and public boundaries for policy intervention and individual/collective creativity. This proposes to reconceptualize previous notions of agency across social, cultural, and natural actual characteristics of development to explain variations in external factors, such as income, work, education, community, governance, values, religion, and political factors in personal choices, such as the access to food quality consumption.

However, the food quality is not a puzzle to be solved exclusively through private-voluntary standards and public policy, but a heavily contested, often chaotic field characterized by high rates of uncertainty and vulnerable to competing beliefs and ideals over well-being desirabilities. It also is an expression of people's agency in mobilizing public policy, ideas about needs, potential desires, which are reasonable for family-level purposes within which an individual is born and lives. This agency interacts with the increasingly 'bads' of modernization and development, conditioning many aspects of both the birth of civic counteractivity, in the form of growing and diversifying food movements, and the individual improvement of expectations about his or her quality of life. These ranges of civic movements and individuals' expectations along with a spectrum from a country or region with indigenous populations recovering their food heritage to an industrialized country with a rural population producing organically and locally, it has over time shifted well-being concerns from the agronomy of food production to processes of food circulation and consumption, thereby forging new, more nuanced understandings and practices in ecology, economy, and health.

Countries and their people are trying to adjust to changing needs within the increasing mobile global market places. Thus, the market is a key site through which producers and consumers are expressing and pursuing their experiences and understanding of well-being, as the multidimensionality of life that includes physical health and safety, community and family social relations. Actors are striving to actively define and improve the notion of well-being. They involve themselves in the arduous laboring of becoming engaged in development processes through social action to organize and materialize their meaning, values, and new subjectivities, as an expectation of becoming 'new' social actors.

New Subjectivities: The Consumer–Citizen in Ecuador and Beyond?

Bekkering (2011) presented a group of food basket housewives that she described as consumer–citizens, and the 'ultimate expression of care', both in term of cooking and food provision, as well as meaningful participation in family community social relations. This case from Ecuador shows these experiences have begun to address concerns over the changing of subjectivities, which are triggered by feeling of social justice and environmental concerns associated with food production, the mainstream market, and its food provision chains. One of the achievements of the food movement is the establishment of a shopping routine that is subsequently carried into 'good' eating practices. Members of the food basket obtain fresh products for discount prices and the meeting themselves are opportunities to socialize consumer experiences as well as to network around emerging ideas and opportunities, such as short-term employment or the cleaning up of a local playground for their youth. The development of a more 'responsible' shopping routine has done much to continue the consumption of whole, home-cooked meals that are generally quite healthy. Organization around 'good food' to counteract the surge of obesity represents a tangible, civic political outcome, and highlights the potential of new subjectivities and social relationships. The food basket experience represents a source of citizenship empowerment for low to middle-class families in improving the household budget and procuring environmental safer, healthier food for a segment of Ecuadorian society (Sherwood et al., 2013).

In order to comprehend the magnitude of these social transformations under way, a brief characterization of these subjectivities is necessary. The power of environmental and social justice values that have evolved to include the health concerns of members has meant a break from simple material affluence, trust in command-based economies, and their inefficiencies and inflexibility to create values. New social relations and the changing subjectivities are becoming less distinct from notions of organic entities and more self-governing from economic and cultural mechanic determinisms. These relations are more responsive to actual national, regional, and international markets, because actors' affect are part of the intersect movement of food between producers, consumers, and new spaces by which commodity circulation circuits (markets) are formed and valorized. In key areas of market development, greater attention is given to variant processes taken place in the social and private sector, market forces, agriculture, and agriculture's alternative forms of production and consumption, which are in tune with people's corporeal capacities to resituate actions, such as eating, cooking, and enjoyment within the material sphere of nature, new nationalisms, and socioeconomic divisions of the broader economy (Fisher, 2012). Somewhat paradoxically, but not unexpectedly, these material and non-material dimensions of global development, markets and how we live the impact of commodities on our stress and well-being have emerged as critically important in an era of neoliberalism institutional discourse, where increasingly it has become socially visible that the advancement of markets and the agency of producers and consumers depend significantly upon certain values and affects of global development as freedom, well-being, and multidimensional commitments to material and nonmaterial developments of actors' quality of live.

Thus, we are concerned with actors' agency, mobility, and their social actions, as part of the process by which commodity and other social values enter the relationship on actors and their notions of well-being, shaping their practices. This can lead to an extension of the partial connections approach (Strathern, 1991) to focus on the blending and juxtaposition of elements of modernity and tradition in the creation of various modernities, whereby people, ideas, and representations of space are interconnected, but never totally symmetrical or fully

integrated. The field of global development, markets, and well-being offers an especially interesting area of inquiry because it inherently throws up a complex mixture of material and nonmaterial notions of quality of life and explanations of causes of health issues and their implications for an appropriate understanding of the consequences of specific modern global pathways and actor's civic participation.

There is in this characterization a certain mode of fragmented dynamism of experts' global development interventions (i.e., the millennium program of development) and a better appreciation of the ways in which global properties, market relations, and well-being values are generated, and challenged, through the ambivalence and ambiguity for differences practices, interests, and knowledge that are appropriated by local people, delimited, and reorganized by active self-organization, provisional building arrangements, and knowledge interactions that are not local or global, internal or external. The appropriation and incorporation of these exteriorities to the lifeworld of actors allow for the resolution of the practicalities of life (Parkin, 1995). From the actor's standpoint, people's self-organizing practices can only effectively be grasped if we give close attention to actors' identity-recursiveness processes and their representations, images, and normative schemata of what constitutes their quality of life. In this vein a concern for people's well-being is part of a generation of imagined worlds and new situated development forms and practices and not simply the reproduction of existing global arrangement regimes.

To be a member of an imagined – well-being – world is of course to belong to a world populated by material and nonmaterial resources, in the sense that there is no actor that have an immutable and static identity not a single and overarching network that exactly match the qualities or profiles of those who are conceived as the guardians of the existing differences that made visions of the accepted order. If this representation is just partially context defining, then this implies that the examination of global development, markets, and well-being invoke the notion of multiple representations and arrangements. The social space itself is capable of being iterated by violent fluctuations in produce market prices. This has exacerbated a trend among farmers to switch crops following commercial and technological advice, inhibiting local experimentation and innovation (Sillitoe, 2007). This has also contributed to create the occurrence of social unrest among producers and consumers. This social unrest is a distinctive collective process, operating in ways and leading to the generation of 'other' mutual coordination practices, between consumers and producers. In this vein, the notion of well-being is a concept that recognizes the existence of poverty, but it does not recognize it determinacy on peoples' agency and social inactivity.

The Positioning of International Agencies and Knowledge in the New Global Environment

The connection of terms: global development, markets, and well-being have been heavily influenced by the successive rethinking of policy options from within international development agencies (DFID and World Bank 2003, UNDIP report on human resources 1990, World Bank 1989). Hence, its meaning has shifted in accordance with the different policy orientations of the global experts and bureaucrats involved. Dissolving the boundaries between 'externalities' and the social forms that represent responses to policies and programs. This process has tended to promote market models and the Western neoliberal optic of consumerism. This blueprint becomes significant because it entails a technique of representing, organizing and describe the global development affording concepts and policies generated by decision makers themselves. This, once again, challenges researchers to develop further theoretical understanding of agency, material and nonmaterial dimensions of life. Ultimately is important to shift the orientation that demonstrates the effectiveness of state institutions, science, and technology, but also that any scientific knowledge requires an endorsement from peoples' grounded worlds in order to legitimize researchers' observations and new information in explaining health conceptions and practices as part of the notion of peoples' well-being.

Finally, global development, markets, and well-being are ambiguous, political, and pragmatic categories that facilitate rethinking social and policy changes and serves to make us aware of the potential difficulties and different interpretations that 'good governance' and the policy of the development intervention need to address. In the last three decades development studies have shown that aid processes are far more complicated and disrupted in terms of actions and emergent properties evolving between actors and policy makers in charge of distributing aid objectives (Lewis and Mosse, 2000). Chambers (2005) has stressed that perhaps one of the central issues in these policy disruptions is the incompatibility of values between interventions expanding 'global modernities' into multiple localities and the subsequent resistance, struggle, and many unintended consequences emerging, when local people encounter these development policies and representation of Westernized lifestyles.

Indeed, global development, markets, and well-being are part of a research-oriented policy vocabulary attempting to shift the debate from the discourse of neoliberal policy to actors' interpretations of its contents, consequences, and actual social practice. Such shifts underline the need to explore how global development, market values are reproduced, transformed, or subsumed by other values. We need to analyze how these categories intersect reordering peoples' everyday life. This is to forge a new path of policy research, which accords due importance to the appropriation and resituation of multiple forms of value and materialization. In other words, we need to explore the ways in which global development properties are embedded into 'real markets' and quality of life values and social action. This challenge beliefs in the universality and uniformity of the global development that unfolds into the world unproblematically, whatever their market and social value context, offering instead an analysis of the associations and linkages of situated well-being values and markets, against the dominant thinking that opposes citizenry and consumption, neglecting or denying the responsible practices that value-oriented consumers' (i.e., fair-trade) already play in the enactment of learning to be satisfied as global 'reflexive' shoppers.

Conclusion

This article has underlined that the main focus of the relation among global developments, markets, and well-being lies at the very core of grappling with an understanding of contemporary change. Global properties, market beliefs, and well-being values constitute the conceptual field upon which counterworks, countertendencies, and counterpoints to neoliberal ideology and discourse are founded and remain one of the main issues of anthropology of development.

This type of research orientation should be considered a borderland across theoretical representations and ethnographic particularities of self-organizing processes to explain interests and values for the analysis of multidimensional social processes resulting from the incorporation of groups and individuals to develop a far-reaching understanding on what drives well-being of people and nations, and what needs to be done to achieve greater democracy, rights, and development. This orientation may contribute to the democratization of world research. This idea that anthropology of development can generate a global and critical collaborative rethinking may not resolve the differences of power that characterized the world of global development experts, but it might contribute to make less linear and more complex the problems of globalization from below. This means recognizing how people bridge with their practices and actions existing taken for granted differences of the material, immaterial, and hierarchical world of knowledge.

See also: Alternative Food Movements; Citizenship: Political; Cosmopolitanism; Development: Social-Anthropological Aspects; Ethical Turn in International Development; Geographies of Food: Politics, Place, Power Reprised; Modernity: Anthropological Aspects; Reflexivity in Anthropology; Slow Food as a Social Movement; Subjective Wellbeing and Culture.

Bibliography

Allen, P., FitzSimmons, M., Goodman, M., Warner, K., 2003. Shifting plates in the agrifood landscape: the tectonics of alternative agrifood initiatives in California. Journal of Rural Studies 19, 61–75.

Appadurai, A., 2003. Grassroots globalization and the research imagination. In: Appadurai, A. (Ed.), Globalization. Duke University Press, Durham, pp. 1–21.

Arce, A., 2010. State policy intervention in an era of civic participation. In: Long, N., Jingzhong, Y., Yihuan, W. (Eds.), Rural Transformations and Development – China in Context. Edward Elgar, Cheltenham, UK; Northampton, MA, pp. 279–311.

Arce, A., Long, N., 2000. Reconfiguring modernity and development from an anthropological perspective. In: Arce, A., Long, N. (Eds.), Anthropology, Development and Modernities: Exploring Discourses, Counter-tendencies and Violence. Routledge, London and New York, pp. 1–31.

Beck, U., Grande, E., 2010. Varieties of second modernity: the cosmopolitan turn in social and political theory and research. British Journal of Sociology 61 (3), 409–443.

Bekkering, E., 2011. The multiple realities of alternative food networks: an ethnography of the Canastas Comunitarias in Ecuador (MSc Thesis). Wageningen University, The Netherlands, 84 pp.

Carrier, J.G. (Ed.), 1997. Meanings of the Market: The Free Market in Western Culture. Berg, Oxford and New York.

Chambers, R., 2005. Ideas for Development. Earthscan, London.

Coole, D., Frost, S. (Eds.), 2010. New Materialisms: Ontology, Agency and Politics. Duke University Press, Durham, 352 pp.

De Landa, M., 2009. A New Philosophy of Society: Assemblage Theory and Social Complexity. Continuum, London and New York.

Fisher, E., 2012. The "Fair trade nation": market-oriented development in devolved European regions. Human Organization 71 (3), 255–267.

Foweraker, J., 2001. Grassroots movements and political activism in Latin America: a critical comparison of Chile and Brazil. Journal of Latin American Studies 33, 839–865.

Fox, J., 2007. Accountability Politics: Power and Voice in Rural Mexico. Oxford University Press, Oxford, New York.

Giddens, A., 1979. Central problems in Social Theory. Macmillan/Berkeley: University of California Press, London.

Giddens, A., 1984. The Constitution of Society: Outline of the Theory of Structuration. Polity Press, Cambridge.

Giddens, A., 1990. The Consequences of Modernity. First published in the United Kingdom 1990 by Polity Press & Blackwell Publishers Ltd.

Gough, I., McGregor, A. (Eds.), 2007. Wellbeing in Developing Countries: From Theory to Research. Cambridge University Press, Cambridge.

Graeber, D., 2001. Toward an Anthropological Theory of Value: The False Coin of Our Own Dreams. Palgrave, New York.

Harvey, D., 2003. The New Imperialism. Oxford University Press, Oxford.

Kloppenburg Jr., J., Lezberg, S., De Master, K., Stevenson, G.W., Hendrickson, J., 2000. Tasting food, tasting sustainability: defining the attributes of an alternative food system with competent, ordinary people. Human Organization 59 (2), 177–185.

Lewis, D., Mosse, D., 2000. Encountering order and disjuncture. contemporary anthropological perspectives on the organisation of development. Oxford Development Studies 34 (1), 1–13.

Long, N., 2001. Development Sociology: Actor Perspectives. Routledge, London and New York.

Merry, S.E., Goodale, M., 2007. The Practice of Human Rights: Tracking Law between the Global and the Local. Cambridge University Press, Cambridge.

Mosse, D., 2005. Cultivating Development: An Ethnography of Aid Policy and Practice. Pluto Press, London and Ann Arbor.

Ong, A., Collier, S. (Eds.), 2005. Global Assemblages: Technology, Politics and Ethics as Anthropological Problems. Blackwell Publishing, Oxford and New York.

Paley, J., 2002. Toward an anthropology of democracy. Annual Review of Anthropology 31, 469–496.

Parkin, D., 1995. Latticed knowledge: eradication and dispelling the unpalatable in Islam, medical and anthropological theory. In: Fardon, R. (Ed.), Counterworks: Managing the Diversity of Knowledge. Routledge, London and New York, pp. 143–163.

Rabinow, P., Marcus, G., Faubion, J., Rees, T., 2008. Designs for an Anthropology of the Contemporary. Duke University, Durham.

Scheper-Hughes, N., 2005. The last commodity: post-human ethics and the global traffic in "fresh" organs. In: Ong, A., Collier, S. (Eds.), Global Assemblages: Technology, Politics and Ethics as Anthropological Problems. Blackwell Publishing, Oxford and New York, pp. 145–168.

Sen, A., 1999. Development as Freedom. Belknap Press, Cambridge.

Sherwood, S., Arce, A., Berti, P., Borja, R., Oyarzun, P., Bekkering, 2013. Tackling the new materialities: modern food and counter-movements in Ecuador. Food Policy 41, 1–10.

Sillitoe, P. (Ed.), 2007. Local Science Vs. Global Science: Approaches to Indigenous Knowledge in International Development. Berghahn Book, New York and Oxford.

Stirrat, L.R., 1992. Good governance and the market. In: Dilley, R. (Ed.), Contesting Markets: Analyses of Ideology, Discourse and Practice. Edinburgh University Press, Edinburgh.

Strathern, M., 1991. Partial Connections. Rowman & Littlefield Publishers, Inc., Balt.

Trauger, A., 2007. In: Maye, D., Holloway, L., Kneafsey, M. (Eds.), Connecting Social Justice to Sustainability: Discourse and Practice in Sustainable Agriculture in Pennsylvania in Alternative Food Geographies: Representation and Practice. Elsevier, Amsterdam, pp. 39–54.

Werbner, P., 2008. Introduction: towards a new cosmopolitan anthropology. In: Werbner, P. (Ed.), Anthropology and the New Cosmopolitanism. ASA Monographs, vol. 45. Oxford International Publishers, Berg-New York, pp. 1–29.

Global Governance

Kennette Benedict, Bulletin of the Atomic Scientists, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Global governance is the purposeful order that emerges from institutions, processes, norms, formal agreements, and informal mechanisms that regulate action for the common good. Global governance encompasses activities that transcend national boundaries at the international, transnational, and regional levels and is based on rights and rules that are enforced through a combination of economic and moral incentives. Mechanisms of global governance are composed of elements and methods from both the public and private sectors. These elements include agreed upon standards, evolving norms based on shared values, and directives issued by private authorities and ultimately enforced by states. Methods of global governance include harmonization of laws among states, international regimes, global policy issue networks, and hybrid institutions that combine functions of state agencies and private sector organizations.

The exercise of global governance raises three sets of issues. The first has to with the legitimate exercise of authority, the second with democratic values, and the third with effectiveness. In contrast to governance at local and national levels, a social contract between citizens and institutions of global governance has not developed to constitute a sufficient basis for legitimate action in all domains. In its current formulations, global governance implies democratic governance by emphasizing openness, accountability, and fairness. However, the reliance on scientific and professional bodies to set standards, on bureaucratic agencies of the state to implement policies, and on voluntary organizations to monitor compliance, none of which are based on democratic principles of representation or equal participation, raises questions about the compatibility of democratic values with the practice of global governance. The effectiveness of global governance is challenged by the worldwide scope of problems, such as climate change, and by the large number of public institutions, private firms, and civil society groups that require coordination to address global problems.

Definition

Global governance is the combination of intentional and patterned human interactions that regulate action worldwide for the common good. It is a purposeful order that emerges from institutions, processes, norms, formal agreements, and informal mechanisms that provide a field of action for human activities. Governance is based on shared expectations, as well as on intentionally designed institutions and mechanisms (Rosenau and Czempiel, 1992; Biersteker, 2009).

In nation-states, the authority of public institutions lies ultimately in the government's capacity to provide societal order and socialize individuals to a common identity, as well as its monopoly of the legitimate use of force (March and Olsen, 1998). In the global context, no world governmental body is responsible for order and stability, nationalism stirs rivalries, and no one institution enjoys a monopoly of the legitimate use of force; no single global authority exists. Yet commerce among countries and firms expands, information is communicated across national boundaries in an orderly fashion, and expectations about rule implementation are shared through transnational networks of private and public agents. Cooperative action is based on rights and rules that are enforced through a combination of financial, material, and moral incentives as well as through the use or threat of force. How this happens is the domain of global governance.

Global governance is distinguished from international relations, which refers to relations among states and to the formal institutions in which states participate, such as the United Nations, the North Atlantic Treaty Organization, or the Association of South East Asian Nations. Global governance includes these formal organizations, but also encompasses activity at the international, transnational, and regional levels that is both within and outside the purview of states. For some analysts, the term includes recognition of the effects of local activity on the global, such as emissions of carbon dioxide on the atmosphere, and global effects on the local, such as transnational capital flows on the health of subnational economies. For most analysts, global refers to those activities in the public and private sectors that transcend nation-state boundaries.

Governance in the Context of Rapid Globalization

Heightened attention to global governance at the turn of the twenty-first century is the result of an increasing understanding of the interconnectedness of human activity on the planet and growing recognition that problems arising from global trade and finance, and from changes in the climate due to human activity, for example, need more cooperation than is possible through states acting together in existing international organizations, such as the United Nations or the International Monetary Fund. The growing interdependence of societies brought about by increased trade and capital flowing across national boundaries, a recognition of the effects of human activity on the biosphere, extensive population migration, and increasing communication between people in far-flung locations contribute to the globalization of human society. Regulating these transborder activities to prevent harm to economic livelihoods, to human health and dignity, and to the Earth's

atmosphere prompt increased attention to the need for improved global governance.

Informal Networks, Formal Institutions, and Incentives

Ordered and patterned interactions on a global scale then are the result of international agreements, of harmonization of laws among countries, of common interpretations of international law, of transnational cooperation among functional agencies, of self-enforcing contracts between private firms, and of standard operating procedures of governmental and nongovernmental organizations. Taken altogether, this array of formal legal arrangements, informal codes of conduct, incentives, and habitual practices based on shared expectations about legitimate action constitutes global governance.

So, for example, the United Nations is a formal international arrangement based on legal agreements among states and is considered to be part of global governance. It is one of many formal institutions, however. Others include the World Trade Organization, the North Atlantic Treaty Organization, the Organization of American States, as well as other regionally based treaty organizations, such as the Association of Southeast Asian Nations, the Organization of African States, and Asia–Pacific Economic Cooperation.

In addition, transnational cooperative actions among national law enforcement agencies, for example, also are part of global governance. Interpol combats illegal criminal activity, like drug and human trafficking, through police agencies that volunteer to cooperate and share information. In civil law, courts set precedents and decide cases based on case law from other national judicial systems. And in commercial law, private mediation is used by business firms, rather than government authorities, to settle disputes (Slaughter, 1997). Finally, over the past two decades, newly created international courts with compulsory jurisdiction are independently applying the rule of law as newly empowered nonstate litigants seek remedies, and courts call upon legal and political networks to pressure states to uphold court interpretations (Alter, 2014).

These informal arrangements are supported by networks of individuals, nongovernmental organizations (NGOs), and state agencies. Cross-government professional relationships provide opportunities to learn from the experience of judges, central bank governors (the Basle Committee), securities regulators (International Organization of Securities Commissioners), and regulators of insurance industries (International Association of Insurance Supervisors) from other countries. These organizations use Memoranda of Understanding to improve the harmonization of regulation on the international level as financial transactions have become increasingly global. Providing for orderly global financial transactions through transnational networks of domestic officials and industry representatives can be quicker and more flexible than negotiating through formal international institutions; new rules can be incorporated directly into domestic law, rather than requiring treaty action.

The development of networks has accelerated with the introduction of high-speed electronic communications technology. Even before these advances, however, travel and telephone communication contributed to a worldwide human rights movement, a global environmental movement, an international women's movement, transnational corporations, and international professional societies. Such networks have established common standards in a number of fields. In the case of human rights, for example, the diffusion of norms about the rights of prisoners and citizens came about through the use of international media to publicize cases of abuse. Similar tactics have exposed poor labor conditions of major transnational corporations such as Nike and Apple. Such public 'shaming' of governments and businesses is based on shared norms about the need to protect individual rights. Some of these values have been codified in formal national and international laws, while others are observed because they enhance the public reputations of businesses and legitimacy of governments.

At its core, the concept of global governance builds on behavioral understandings of law and society, where law is a dynamic and changing reflection of shared standards and values. This suggests that norms of appropriate behavior emerge from daily interactions to solve transborder problems and from a developing consensus about what is fair. These norms may subsequently be codified in law. Compliance is undertaken voluntarily rather than through the use of physical force, although the threat of sanctions, whether criminal or civil, is recognized as an underlying motivator of compliant action. The evolution of GATT into the World Trade Organization exemplifies the formal codification of rules, about international trade in this case, emerging from a process of informal negotiation and standard setting (Abbott and Snidal, 2000).

Both the threat of sanctions and an interest in avoiding harm to themselves are at work. In the private adjudication of transnational business disputes, for example, firms wish to preserve flexibility and avoid undue government interference, which they may view as harmful, so they comply with the judgment of a private mediator rather than taking the issue to a public institution for settlement. Transnational corporations, such as Nike and Apple, will implement higher labor standards if they believe their product sales will suffer if poor worker conditions are brought to light.

Elements of Global Governance

Several fundamental elements constitute the building blocks of contemporary global governance. The most prominent of these are standards, norms, states, and international organizations.

Standards

Agreeing upon and setting standards is one means of coordinating private actions. For example, the standardization of time through the use of Greenwich Mean Time permits coordination of communications, airline schedules, television broadcasting, business deals, and a myriad of daily transactions that we now take for granted. Rules of international navigation and airline procedures ensure the safety of air travel across national borders. A standard railroad gauge eases transborder crossings. Product standardization rationalizes the multinational

manufacturing of automobiles, televisions, refrigerators, and other appliances through ISO 9000. The Society for Worldwide Interbank Financial Telecommunications (SWIFT) ensures that financial wire transactions between banks in different countries are protected from corrupt or criminal practices.

In general, standards are accepted because they are voluntary and are arrived at through the exercise of professional expertise, in which peers or equal actors reach decisions through rational deliberations that are open to interested parties. In practice, the process of standard setting by bodies such as ISO 9000, for example, is completed by those with expertise in fields like accounting and engineering or the Internet, representing users and pledged to serve the public interest. The legitimacy of standards rests upon efficiency and on the desirable consequences of the rule, which allows no exceptions or variations. To the extent that efficiency is a value in itself, and to the extent that professions embody and transmit a set of values, standard setting is not entirely value-neutral. In comparison with hierarchy, markets, or norm-based communities, standardization is a minimal form of governance.

Norms

Norms are internalized rules based on commonly held values. In the international security area, nonuse of nuclear weapons is cited as such a norm. The norm is based in part on a shared revulsion at the devastation to human life and society that is inflicted indiscriminately in nuclear explosions. In the human rights domain, human rights abuses violate widely shared understandings of the inherent value of human life and the appropriate treatment of individuals by their governments.

Shared norms are developed and transmitted through informal networks of professionals and representatives from international institutions, national governments, nongovernmental voluntary associations, and private business (Finnemore and Sikkink, 1998). For example, the International Red Cross advocated for the humane treatment of wartime combatants that resulted in the Geneva Convention by convincing military leaders that protecting the wounded was compatible with their war aims. These commonly held values in many domains, in turn, contribute to the regulation of international conflict, the management of nuclear weapons, environmental pollution, human rights abuses, and humanitarian relief, among others. Some analysts view these networks as an essential component of global governance, providing a way of governing without government (Reinicke, 1998).

States

The organization of nation-states in the twentieth century has depended upon the development of bureaucracy, or hierarchy, as a means of governing. The growth of the modern state has included regulation of markets, of educational institutions, of environmental conditions, of criminal activity – in fact of nearly every aspect of daily life – as a means of coordinating action for the common good. While some of this regulation has come through standard setting and norms development, the characteristic mode of hierarchical government is through directives. These directives are enforced through the use and threat of punitive sanctions, including imprisonment and fines.

The enforcement power of the state is based on its monopoly over the legitimate use of force. This monopoly, in turn, is based on a social contract, whereby citizens voluntarily cede some of their sovereign rights to the government, in return for protection from hostile forces and for a social order maintained by the state. The central problem for theorists and practitioners of global governance is that there is no global social contract and no world government, and yet, transborder activities generate needs for coordination, as well as problems of collective action to deal with externalities, such as climate-changing carbon dioxide emissions.

Some observers locate global governance in a global public domain beyond the sphere of states (Ruggie, 2004), where supranational and private actors dominate in emerging structures that are undemocratic and unequal (Deudney, 2007). Others emphasize the continuing role of states in global governance and document how national laws anchor and promote transnational private financial transactions for economic gain at the same time that national laws are shaped by the actions and needs of private transnational firms (Sassen, 2002). Whether the state is part of or is marginal to global governance continues to be debated. Whatever the outcome of the debate, however, it is clear that events like the global financial crisis of 2007, violent attacks by global terrorist movements, and the lack of concerted action to reduce climate-changing gases challenge state sovereignty and the capacity to protect citizens from harmful transborder flows and activities.

International Organizations

International organizations, ranging from the United Nations, the World Trade Organization, the European Union, the North Atlantic Treaty Organization, to the International Court of Justice, have played more active roles since the end of the Cold War in the early 1990s. Their influence is sometimes difficult to evaluate, and their failings are often the subject of negative comment, as when the UN Security Council failed to intervene in genocides in Rwanda and Darfur. Yet the positive contributions to more efficient collective state action, as in state pooling of financial resources for economic development in poor countries through the World Bank and the International Monetary Fund, are often taken for granted (Abbott and Snidal, 1998).

In a particularly telling contrast, UN conferences that have sought to commit member states to reductions in their carbon dioxide emissions that disrupt the Earth's climate have failed over the past 40 years. Yet the UN-supported Intergovernmental Panel on Climate Change, which collects data from scientists worldwide and reports the findings in consensus documents, has been very successful. Indeed, it has presented the public and policy leaders with remarkably clear analyses of large bodies of data and provided a clear scientific basis for government action.

However, industrial countries belong to a plethora of international, regional, and issue-specific organizations that overlap in ways to actually cause conflict and complicate national decision making. The decade-long international dispute over bananas between the European Union and the

United States illustrates the complexities and costs of overlapping jurisdictions, in this case, the nesting of the banana regime within the EU, the Lome Convention, and the World Trade Organization (Alter and Meunier, 2006).

Newly established international courts are emerging as autonomous interpreters of the rule of law as well as arbiters of disputes between states. The rise of compulsory jurisdiction that allows international commissions, prosecutors, and private parties to initiate litigation enables international courts to interpret the meaning of international law independently from state interests. In effect, changes over the past 20 years have removed the monopoly power of national governments and domestic judges to define what international law requires even within their own jurisdictions. Much as states have tasked international financial institutions with aiding poor countries and regulating trade disputes, states have tasked international courts with helping to enforce international law (Alter, 2014). In cases ranging from prosecuting war criminals, to enforcing international conventions against human trafficking, to establishing labor rights, courts have interpreted and enforced new international law. And, whereas states have traditionally enforced laws through coercion and punishment, international courts enforce laws with the aid of transnational organizations, international public media, and legal actors to pressure governments to comply. As domestic courts often effect changes in legal interpretation by expressing a public consensus about the expansion of rights even at the expense of dominant interests, international courts are acting independently from powerful state interests to make findings from the perspective of rule of law.

Methods of Global Governance

Elements of governance – standards, norms, state capacities, and international organizations – are used in combination to coordinate action, solve transnational problems, and manage conflict, and together constitute global governance in practice. Among the most effective methods are harmonization of laws, international regimes, global issue networks, and hybrid public–private institutions.

Harmonization of Laws

The implementation of formal international agreements is accomplished through the adoption of national laws to bring local and national practices into compliance with international rules. For example, carrying out the provisions of the Montreal Protocol to protect the atmospheric ozone layer is accomplished through legislative action in each country to bring manufacturing practices into conformity with the Protocol's agreed directives to reduce production and release of chlorofluorocarbons.

Legislative action is one path to harmonization of laws. Another is through the courts and judicial systems of countries. A body of transnational law is developing informally in which judges in one country are basing their decisions on precedents and case law from other countries (Busch and Jorgens, 2005; Slaughter, 1997). This is taking place in areas of civil and commercial law, through observation and borrowing, where standard rules help to smooth business and legal transactions in what is conceived as global public administrative law (Kingsbury et al., 2005). In the area of human rights law, courts are citing cases in other national jurisdictions, in addition to the United Nation's Universal Declaration of Human Rights, as precedent and to lend credibility to decisions about the protection of civil and political rights in their own countries.

International Regimes

International regimes are defined as a "set of implicit or explicit principles, norms, rules, and decision-making procedures around which actors' expectations converge" (Krasner, 1983). The effectiveness of a regime rests on the operations of institutions, organizations, governments, and international bodies that share a set of principles, rules, and norms in a particular area of international action. Although regimes include formal treaties and national law, they also rely on informal norms and networks to develop and enforce standard behavior in an area of global policy.

For example, regimes to control the development and spread of nuclear weapons had been relatively successful in the context of the Cold War. They were critical to controlling the US–Soviet arms race, to preventing nuclear war, and to limiting the spread of nuclear weapons technology and weapons. These regimes had their basis in a number of formal treaties and international agreements, such as the Anti-ballistic Missile Treaty of 1972 and the Nuclear Non-proliferation Treaty of 1968, as well as selected enforcement powers through the United Nation's International Atomic Energy Agency. They were shaped as well by the informal meetings, discussions, networks, and personal relationships of scientists and other experts working across national boundaries to formulate these agreements (Evangelista, 2002). To ensure compliance, regimes relied on formal and informal surveillance techniques and the trust that emerged from shared norms about the dangers of nuclear weapons, their development, production, and proliferation (de Nevers, 1999).

The World Trade Organization (WTO), along with the OECD's system for accepting data on product testing, multilateral agreements on investment, and other regulations, constitute a regime that provides order in trading between and among countries. The regime provides for standards, regulates tariffs, and adjudicates disputes among countries to prevent trade wars. Although still regarded as an institution that favors developed countries, WTO inclusion of environmental standards and labor provisions has enhanced its legitimacy as a global organization that can effectively resolve disputes, especially between emerging economies and established industrial countries.

Global Policy Issue Networks

Development of global policy and implementation of common laws also depends on informal networks of organizations and individuals. These networks are composed of individuals who share an interest in providing a public good that may not be provided by any single nation-state. Through communication networks, sometimes built around professional associations and scientific bodies, these transnational

action networks (Keck and Sikkink, 1998) or epistemic communities (Haas et al., 1993) provide information, documentation, and alternative policy solutions to global policy problems.

Some of the most visible networks have formed since the mid-1970s on human rights, environmental protection, communications technology, and women's issues. In each of these areas, lawyers, scientists, and political and cultural leaders have found counterparts in other countries to form worldwide movements that have publicized the harm brought about by current national policies and practices. In the case of women's issues, the actions of nongovernmental organizations at United Nations conferences on women and on population were essential to changing norms and practice. At the 1995 population conference in Cairo, the recognition of women's empowerment as a key path to fertility reduction, rather than an emphasis on birth control, demonstrated the influence and effectiveness of women's issue networks in shaping global policy. The adoption by the United Nations of the concept of the 'responsibility to protect' emerged from policy networks, high-level commissions funded by private foundations and governments, and in response to the shameful genocides in Rwanda and Bosnia in the 1990s that violated shared norms about the protection of civilians in conflicts.

Hybrid Institutions

As transnational activity has increased, as nongovernmental actors have become more prominent, and as national governments have devolved authority to private sector bodies, new institutional arrangements have developed to manage global relations. These arrangements are referred to as hybrid institutions or public–private partnerships because they mix private sector organizations and a variety of incentive systems with governmental agencies to achieve public goods. They combine standard rules, private market arrangements, nongovernmental norm-based organizations, and government agencies – both national and international.

Some of the most inventive have emerged in the area of environmental protection. The certification of sustainably harvested coffee or forest products is an example of hybrid global governance institutions. Presented as an alternative to governmental regulation of the timber industry, environmental nongovernmental organizations have convinced major private firms in the forest products industry that consumers will pay premium prices for furniture and other lumber products that are sustainably harvested. To ensure compliance, private firms participate in a council of producers, scientists, and other independent interests that inspects production operations and certifies that manufactured products are derived from sustainably harvested lumber. The manufacturer uses that certification to appeal to consumers who are willing to pay higher prices to protect natural resources. The process takes place without government intervention, includes market incentives, and relies on professional expertise and standardization to enforce rules. Similar certification institutes have been established for textile manufacturing and diamond mining.

The prosecution of transnational criminal activity involves another kind of hybrid institution. Organized crime has become relatively border free, especially as liberalization of economic relations has taken place, yet law enforcement jurisdiction remains within fixed territorial legal frameworks. To deal with this challenge, law enforcement agencies have developed cooperative relationships with their counterparts in other countries, as well as with banking institutions, to share information and to apprehend criminals. Interpol, a voluntary association of police departments from around the world, has made 31 000 arrests in the past decade to bring malefactors in the illegal transnational drug trade, the arms trade, human trafficking, and other criminal activity, to justice (Barnett and Duvall, 2005).

Issues in Global Governance

While new modes of coordinating transborder activity are emerging that hold the promise of effective global governance, theoretical and normative challenges have been raised. Three are addressed here: effectiveness, the legitimacy of global governance bodies, and the democratic nature of these global governance processes.

Effectiveness

The problems that global governance is intended to address are beyond the scope of any one national government to solve. Yet, without the usual tools of government, including legitimate use of force to induce compliance, governance at the global level often fails. Enforcement of the Nuclear Nonproliferation Treaty is compromised by lack of automatic sanctions for countries such as North Korea, Libya, and Iran, when they withdraw from the treaty or engage in bomb-making activity, or for European suppliers that provide equipment to the illicit nuclear weapon network of Pakistan's nuclear scientist A.Q. Khan.

When the interests of a coalition of states can be met through international organizations, as in the case of removing chemical weapons from Syria in 2014 by the Organization for the Prohibition of Chemical Weapons, global agencies can provide expertise and coordination of state parties to international conventions. But more often, global governance is frustrated by lack of agreement about how goals should be met, as in the failure to agree on goals and quotas for reducing carbon emissions that are warming the Earth's atmosphere, or in devising common rules for regulating securities and financial transactions worldwide with the result that the global economy suffers in 2014 from volatility that dampens recovery from the world recession of 2008. In another example, prompt coordination to prevent the spread of deadly pandemics is hampered by lack of state cooperation in reporting outbreaks. Officials may fear a halt in international tourism and loss of trade if they do report emerging epidemics, but they also may suffer from a lack of faith in the effectiveness of global institutions like the World Health Organization. This lack of information sharing, in turn, further reduces the effectiveness of global governance.

One way to increase the effectiveness of global governance is to instill heightened belief in the legitimate action and authority of global institutions. This is the next challenge to be addressed.

Legitimacy and Authority

Successful governance requires the consent and compliance of those whose activity is being governed. In other words, to be effective, practices and institutions that constrain actions must be accepted as legitimate. This aspect of contemporary global governance remains problematic. In theories of the state, social contracts between citizen and government have depended on citizens relinquishing some rights in exchange for physical protection. But no such social contract exists at the global level (Barnett and Duvall, 2005). At the most, legitimacy claims have been based on what works well, what is efficient, or what is reasonable in the eyes of those seeking solutions to transborder coordination and public goods problems (Deudney, 2007).

Issues of legitimacy become even more vexed with the increased participation of nongovernmental organizations in issue networks and hybrid institutions. While the authority of governments and the legitimacy of private firms are based, in some measure, on implicit contracts with citizens or consumers, respectively, the authority and legitimacy of nongovernmental organizations is based on their claims of acting in the public good – as they define it. Whether protecting individuals by calling attention to the human rights abuses of their own governments, or conserving biodiversity in the face of extractive private industry, or saving women from customary rituals that cause physical harm, nongovernmental actors believe they are upholding standards that are in the interests of all humanity.

These organizations, however, are not assembled through an election process or any other democratic procedure. They are self-appointed and claim to be acting on the basis of values that serve a global public interest. Without an authoritative institution that provides a process for establishing common values and agreement on what constitutes the public good, issues of authority and legitimacy in global governance will most likely remain unresolved.

Democratic Values and Fairness

While some definitions of global governance emphasize the neutral management of human affairs, there is a strong normative component to the term that leaves underlying assumptions about who governs and with whose standards largely unexamined (Barnett and Finnemore, 2004). The strongest challenges to current practices of global governance come from developing country analysts who observe that emerging global standards and norms are based on industrial countries' interests and values (Hurrell, 2007). They suggest that the dominance of power politics in international institutions, the disparity of wealth between North and South, the intolerance for differences in values, the resort to force to resolve international conflicts, and the neglect of environmental degradation are major deficiencies of global governance.

To reform current practices, critics encourage equal participation by developing countries in international institutions and suggest treating dominant powers like the United States as normal countries that must abide by international law. Chinese analysts, in particular, would strive for a harmonious world of justice and common prosperity. They observe that globalization has disproportionately benefited developed countries and contributed to inequality everywhere which is undemocratic and which undermines peace and stability. And global governance mechanisms to manage globalization have grown out of and are imposed by Western ideology and values on the rest of the world. To rectify these disparities, global governance should encourage technology transfer, provide more aid to poorer countries, and in general, promote shared development and common prosperity. Furthermore, recognizing the diversity of civilizations and tolerance for alternative methods of problem solving would provide a richer set of ideas for managing transnational issues and would enhance the legitimacy of global governing mechanisms (Wang and Rosenau, 2009).

While the challenges to current conceptions of global governance may be growing, practices continue to evolve along with calls for transparency, accountability, and protection of individual rights (Ruggie, 2004). The informality and pluralism of global governance both recognizes and requires the contributions of private firms, nongovernmental actors, and informal networks in policymaking. Yet the reliance on unrepresentative scientific and professional bodies to set standards, on bureaucratic national and international agencies to implement policies, and on unelected voluntary organizations to monitor compliance raises questions about the compatibility of democratic values and the practice of global governance.

The increasingly acute challenges to human welfare from rising inequality within and between countries, and from the stresses on societies from climate change will propel transborder activity, on the one hand, and, on the other hand, may induce fear and a retreat to state-centered isolation. The web of networks, institutions, rules, and norms that constitute global governance has thickened dramatically over the past 30 years to meet some of the challenges of economic globalization, promote the spread of universal human rights, and conserve biodiversity of species in the world, among others. Whether the effectiveness, legitimacy, and the authority required to resolve the global crises of the environmental commons, of growing inequality, and of increasing instability and violence will emerge quickly enough to prevent cataclysmic suffering remains to be seen.

See also: Globalization: Legal Aspects; Law and Society; Nongovernmental Organizations: Anthropological and Historical Aspects.

Bibliography

Abbott, K., Snidal, D., 1998. Why states act through formal organizations. Journal of Conflict Resolution 42, 3–32.

Abbott, K., Snidal, D., 2000. Hard and soft law in international governance. International Organization 54 (3), 421–456.

Alter, K.J., 2014. The New Terrain of International Law. Princeton University Press, Princeton, NJ.

Alter, K.J., Meunier, S., 2006. Nested and overlapping regimes in the transatlantic banana dispute. Journal of European Public Policy 13 (3), 362–382.

Avant, D.D., Finnemore, M., Sell, S.K. (Eds.), 2010. Who Governs the Globe?. Cambridge University Press, Cambridge, UK.

Barnett, M., Finnemore, M., 2004. Rules for the World: International Organizations in Global Politics. Cornell University Press, Ithaca.

Barnett, M., Duvall, R. (Eds.), 2005. Power in Global Governance. Cambridge University Press, Cambridge, UK.

Biersteker, T.J., 2009. Global governance. In: Cavelty, M.D., Mauer, V. (Eds.), Routledge Companion to Security. Routledge Publishers, New York and London.
Brunsson, N., Jacobsson, B. (Eds.), 1995. Standardisering. Nerenius and Santerus, Stockholm.
Busch, P.O., Jorgens, H., 2005. The international sources of policy convergence: explaining the spread of environmental policy innovations. Journal of European Public Policy 12 (5), 860–884.
Buzan, B., 2011. A world order without superpowers: decentred globalism. International Relations 25 (1), 3–25.
Commission on Global Governance, 1995. Our Global Neighborhood. Oxford University Press, Oxford, UK.
Deudney, D., 2007. Bounding Power. Princeton University Press.
Evangelista, M., 2002. Unarmed Forces: The Transnational Movement to End the Cold War. Cornell University Press, Ithaca, NY.
Finnemore, M., Sikkink, K., 1998. International norm dynamics and political change. International Organzation 52 (4), 887–917.
Haas, P., 2004. Addressing the global governance deficit. Global Environmental Politics 4 (4), 1–15.
Haas, P.M., Keohane, R.O., Levy, M.A. (Eds.), 1993. Institutions for the Earth. The MIT Press, Cambridge, USA.
Held, D., 1995. Democracy and the Global Order: From the Modern State to Cosmopolitan Governance. Stanford University Press, Stanford, CA.
Hurrell, A., 2007. On Global Order: Power, Values, and the Constitution of International Society. Oxford University Press, Oxford.
Karns, M., Mingst, K., 2009. The united nations: centerpriece of global governance. In: International Organizations: The Politics and Processes of Global Governance. Lynn Rienner, Boulder.
Kaul, I., Grunberg, I., Stern, M. (Eds.), 1999. Global Public Works: International Cooperation in the 21st Century. Oxford University Press, New York.
Keck, M., Sikkink, K., 1998. Activists Beyond Borders: Advocacy Networks in International Politics. Cornell University Press, Ithaca, NY.
Kingsbury, B., Krisch, N., Stewart, R., Summer/Autumn 2005. The emergence of global administrative law. Law and Contemporary Problems 68 (3 & 4).
Krasner, S., 1983. International Regimes. Cornell University Press, Ithaca, NY.
Lake, D., 2010. Rightful rules: authority, order, and the foundations of global governance. International Studies Quarterly 54, 587–613.
March, J.G., Olsen, J.P., 1998. The institutional dynamics of international political orders. International Organization 54 (3), 421–456.
de Nevers, R., 1999. Regimes as Mechanisms for Global Governance. Rockefeller Brothers Fund, New York.
Reinicke, W., 1998. Global Public Policy: Governing without Government? Brookings Institution Press, Washington, DC.
Rosenau, J., Czempiel, E. (Eds.), 1992. Governance without Government: Order and Change in World Politics. Cambridge University Press, Cambridge, UK.
Ruggie, J.G., 2004. Reconstituting the global public domain—issues, actors, and practices. European Journal of International Relations 10 (4), 499–531.
Sassen, S., 2002. The state and globalization. In: Hall, R.B., Biersteker, T.J. (Eds.), The Emergence of Private Authority in Global Governance. Cambridge University Press, New York.
Simmons, B., 2001. The international politics of harmonization: the case of capital market regulation. Interantional Organization 55, 589–620.
Slaughter, A.M., 2005. A New World Order. Princeton University Press, Princeton, NJ.
Slaughter, A.M., 1997. The real new world order. Foreign Affairs 76 (5), 183–197.
Wang, H., Rosenau, J.N., 2009. China and global governance. Asian Perspective 33 (3), 5–39.
Xue, H., 2007. Chinese observations on international law. Chinese Journal of International Law 6 (1), 83–93.

Global History: Universal and World

PK O'Brien, London School of Economics, University of London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 9, pp. 6237–6244, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Global (universal, world) history includes attempts to construct histories concerned with the evolution of humankind that are intellectually inclusive (no discipline or approach to the subject is ruled out); unbounded by geographies, let-alone national frontiers and resolutely unfettered by established chronologies for the construction of European, American, African, Asian, or other national histories. This article elaborates upon the long historiographical tradition of the writing of global history, explains why a renaissance of that tradition is now clearly underway, deals with mainstream and postmodern objectives to the field, and elaborates upon two distinct approaches that exemplify and categorize the increasing flow of modern publications labeled as global history.

Global (universal, world) history includes attempts to construct histories concerned with the evolution of humankind that are intellectually inclusive (no discipline or approach to the subject is ruled out); unbounded by geographies, let-alone national frontiers and resolutely unfettered by established chronologies for the construction of European, American, African, Asian, or other national histories. This article elaborates upon the long historiographical tradition of the writing of global history, explains why a renaissance of that tradition is now clearly underway, deals with mainstream and postmodern objectives to the field, and elaborates upon two distinct approaches that exemplify and categorize the increasing flow of modern publications labeled as global history.

Historiographical Traditions

Global histories have been written since Herodotus (495–425 BCE), whose cosmopolitan concerns were commended by Cicero but deplored by Thucydides – who opined that Greeks had no business investigating alien mythologies, religions, customs, and traditions. History writing, Thucydides suggested, should be modeled upon his own study of the Peloponnesian Wars – sharper in focus, shorter in time horizons, based upon verified facts, and prescriptive in purpose. Fortunately, Herodotus ignored his 'Eurocentric' advice and ranged widely beyond the Hellenic world to include Egypt, India, Babylonia, Arabia, and Persia in his histories in order 'to preserve the memory of the past by placing on record the astonishing achievements both of our own and of the Asiatic peoples.' He used oral testimony, archeological remains as well as written sources and imposed some chronology and order on streams of events that had occurred on several continents and over long spans of time. His (still popular) book ends in 'European triumphalism.' Its story turns into celebration of the victories of the Greek *polis* (especially Athens) over the Persian Empire, which Herodotus represented as a conflict between occident and orient, despotism and freedom.

Global historians applaud Herodotus for his reflexive interest in barbarian virtues and regret the long hiatus that occurred in the writing of secular world histories until Voltaire and his generation carried the project forward again during those brief decades of Enlightenment, before the outbreak of the French Revolution. They are puzzled, however, to observe that with the exception of a few Stoics (Diodorus, Polybius, and Dionysius) and a handful of geographers and ethnographers (Strabo, Ptolemy, and Pliny), histories of the world did not attract more scholars writing during Graeco-Roman times. Although they lived in empires that maintained contact with an array of cultures (African, Arabian, Persian, Indian, and Chinese), historians of the classical era remained overwhelmingly concerned with Europe. Their histories (e.g., Tacitus) deal mainly with the politics and scandals of Rome, or (e.g., Thucydides) with wars among Greeks. Their equally provincial oriental contemporaries, residing in that celestial empire on the other side of the world, asserted with quintessentially Chinese superiority that they remained unimpressed with anything the Romans had to impart, or to trade for, China's silks, ceramics, medicines, perfumes, and spices.

In Europe, steps in an ecumenical direction occurred when Christian and Jewish historians started their narratives with the Creation and set out to ensure that Greeks and Romans should not be left (in the words of Josephus) 'as the arrogant possessors of antiquity.' That was why Orosius (a student of St. Augustine) proclaimed that "The Roman Empire arose in the west but was nourished in the east." Narratives in Christian historiography are dominated by chronicles about the evolution of mankind over stages of history, no longer bounded by the frontiers or cultures of the Greco-Roman world. In 1158, Bishop Otto of Freising recognized that "all human power and learning had its origins in the East." Medieval historians included Oriental and African civilizations right down to the Reformation when Catholic and Protestant intellectuals began to look for a usable past to help princes to integrate populations and secure the frontiers required for the formation of nation states. Even then both parties to Europe's conflict over religion and national identities remained aware of places, peoples, and cultures not merely outside the borders of their enclosed and ideologically encompassing polities, but beyond Christendom itself, and that is hardly surprising.

Residing on an underdeveloped promontory on the western edge of the great Eurasian landmass, with ice to the north,

uncharted oceans and territory to the west, and the hostile power of Islamic communities to the south and east, Europeans had not secured their frontiers, religions, cultures, and technological superiority over other civilizations, long before the eighteenth century. It took centuries to clear the Iberian peninsular of Arab power. Moslem fleets threatened European shipping in the Mediterranean for another two centuries after the surrender of Grenada. Turkish armies crossed into the Balkans in 1350, took Constantinople in 1453, and menaced Vienna as late as 1683. For nearly 1000 years after the fall of Rome, Christendom braced its defenses, imported useful knowledge, shaped its collective identity, and composed ideologically charged histories in the context of conflicts (and encounters) with the powerful civilization of Islam on all its frontiers. Memoirs, travelogs, diplomatic correspondence, and investigations into Arabic medicine and astronomy, together with commercial intelligence on societies under Ottoman dominion, flowed from east to west. Much of this knowledge came through Genoa, Florence, and Venice, along with spices, herbs, sugar, botanical drugs, jewels, chinaware, silks, cottons, and metal goods, imported through the Middle East, from India and China. The expense of procuring and transporting luxuries moving slowly, over long distances and subject to extortionate charges for protection along the way, prompted Europeans to seek ways to surmount the barriers to commerce with South Asia and the Far East by their Islamic enemies.

Shortly after the voyages of discovery, Europeans broke out of their political, commercial, and mental encirclement by Islam. Over the sixteenth century, flows of information into Europe widened to include the Americas, and increased exponentially as a result of the establishment of regular and direct trade and contacts by sea with India, China, South-East Asia, and Africa. Yet nearly two and a half centuries elapsed before Europeans produced a secular and enlightened school of world historians. Its godfather was Voltaire, who composed a history concerned with "only that which is worth knowing; the spirit, the customs, the practices of the principal nations based on facts which one cannot ignore." Voltaire and his followers in France, the Netherlands, Scotland, Naples, Germany, and other parts of Europe deliberately constructed histories of their expanding world in ways which they conceived as a departure from the providential narratives of their clerical predecessors and the 'nationalism' of historians writing for princes and their obedient subjects. That aspiration was propitious. As a result of regular contacts and commerce, the trickle of information about peoples, geographies, technologies, politics, warfare, customs, manners, and the import of all sorts of commodities and artifacts turned into a river of knowledge about places, populations, societies, and economies beyond Europe. For radical intellectuals, China became an alternative model, favorably contrasted with Europe's political systems, social arrangements, and religious beliefs. European histories became not only wider but also deeper as scholarship took over from propaganda, sycophancy, fantasy, religion, and national ideologies.

For something like half a century before the French Revolution and the takeover of India and South-East Asia, European historians resituated Europe's and Christendom's place in world history by including more ancient as well as classical Mediterranean civilizations and by widening their gaze to take China, India, and Japan as well as the newly discovered Americas into account. Europe's new school of world historians dealt with the natural histories, religions, and ethnographies of most parts of an increasingly interrelated universe. Favorable interpretations of oriental polities, societies, cultures, and religions by philosophers and physiocrats did not go unchallenged. Bernier and Leibniz, Montesquieu, Defoe, Winckelman, and Hume remained resolutely unimpressed with the Chinese empire. Nevertheless historians of that era engaged in debates about cultures, geographies, and spans of time that their modern counterparts (confronted with libraries of knowledge and professional interpretations) can only envy.

Alas, the ironic reflexion that their narratives of Enlightenment brought to the task of repositioning Europe in a secular history of the world was overshadowed by a tradition of Western triumphalism which emerged in the aftermath of the Revolutionary and Napoleonic Wars, 1789–1815 when Europeans acquired massive additions to the populations, territories, and natural resources of other continents under direct or informal control. During an 'imperial meridian,' European naval and military superiority over the states and societies of other continents emerged as irresistible. Furthermore, by 1825, European settlers throughout the Americas and the South Pacific had escaped from all metropolitan and aristocratic restraints upon their economic activities and their relations with the indigenous peoples of colonized worlds. White settlers became free to exploit 'their' heritage under rules of their own making, tempered only by religion. As its technological and economic lead over other continents widened, Europe moved clearly onto a path which led, as time went on, to clear divergences in levels of productivity and living standards between East and West and North and South.

Against this background, the writing of history by Europeans departed from its moment of cosmopolitan enlightenment, moved in a Hegelian direction, and returned to the introspection and cultural arrogance associated with imperial Rome. Here and there, aspirations to write universal histories survived. Isolated scholars eschewed the condescensions embodied in the nineteenth century science, technology, liberalism, and colonialism, and a scatter of historians continued to work outside the encompassing national frameworks, and dominant teleologies associated with the rise, rationality, and hegemony of Europe. Nevertheless, it took the destruction and cultural shock of two world wars in the twentieth century to encourage a return to the rewriting of histories of mankind. That program, conducted by a small group of erudite scholars, began in the climate of pessimism and breakdown in the liberal consensus that followed the Great War, when Spengler, Toynbee, Sorokin, Wells, Mumford, and Dawson wrote universal histories of humanity to help Europeans comprehend the decline and barbarity of the West. In history, they attempted to find some substitute for Neitzche's proclaimed demise of Christianity and rationality. Their scholarship is inspiring, but their explicit search for spiritual meanings in the histories of the rather amorphous entities called 'civilizations' did become a compelling framework for scholars and students to work within. The attempt by UNESCO, after the World War II, to fund the writing of a multivolume collaborative history of the world (designed to promote peace across nation states) failed to provide

a paradigm for the reconstruction of secular of histories on an ecumenical scale. Long after the war, historical scholarship continued to be located where its founders had situated it from the time of the Renaissance and Reformation onwards – namely, within the confines, chronologies, and parameters established for comprehending the origins and evolution of European (including North American and Australasian) societies and nation states.

Globalization and the Revival of Global History

Fortunately pockets of global history survived – notably at the universities of Chicago, Northwestern, and Leipzig. Complementary branches of scholarship continued to flourish in the forms of survey courses in Western civilization; area studies, histories of trade and imperialism, investigations into the impact Western technology and culture upon other peoples and above all, as flows of encyclopedias, chronicles, and textbooks for schools and popular education concerned with the world as a whole. Thus, 'tribes' of professional historians with interests in fields that transcended nation states and Europe possessed the credentials and remained ready to refashion their intellectual identities, when demands for global history reappeared from the late 1970s onwards. That revival could be retraced to canonical publications by M. Hodgson, W. McNeill, and L. Stavrianos, but research and teaching has only matured into university programs for the sustained study of world history over the past quarter of a century of 'globalization.' Although there is nothing particularly new about the current phase of economic integration and interdependence, which are the core features covered by that label. Long-distance trade goes back millennia to prehistory, and the exchange of commodities across frontiers and overseas had always involved arrangements for counterpart flows of funds, conversions of currencies, and the operation of transnational, commercial, and financial organizations of a more or less sophisticated kind within and across Europe, Africa, and Asia.

Nevertheless, real discontinuities had occurred. For example, by the late nineteenth century, supported by steam-powered ships and trains, the volume of commodities and services traded across national borders, continental boundaries, and oceans had increased dramatically compared to 1800. Commerce was promoted and sustained by movements of capital, migrations of labor, and transfers of technology and information around the world on an unprecedented scale and at ever-increasing speeds. Political impediments to international flows of exports, imports, money, credit, capital, labor, technology, and information diminished sharply during the liberal international order that prevailed between 1846 and 1914.

State controls revived again during nearly four decades of neomercantilism that succeeded World War I; when governments everywhere sought to insulate their domestic economies from inflows and outflows of commodities, from the emigration of their nation's capital and skilled labor, and from fluctuations associated with unregulated rates of exchange. In the interests of narrowly defined national competitive advantages and political stability, sovereign states proceeded to delink economies from participation in an evolving global economy. Between 1914 and 1948, overall levels of world trade fluctuated in volume, but ceased to grow. Warfare and political regulations depressed movements of capital and labor across frontiers. Controls on the convertibility of currencies hampered international exchanges of every kind.

Normalcy returned after World War II when a discernible shift toward liberalism in the political and legal framework for commerce among nations was also promoted by even more radical improvements in new technologies for the transportation of goods and people across space and time; and by vastly more efficient communications systems for the widespread diffusion of commercial intelligence, messages, and cultural information between distant locations and separated populations. Revolutions in transport, television, and communications widened markets and created opportunities for the organization of transcontinental and multinational business across an expanding range of production, distribution, finance, and commerce.

Unsurprisingly, the implications of 'globalization' for hitherto separated communities, bounded spaces, distinctive cultures, partially connected local economies, and nominally sovereign states have been analyzed by social scientists. Intensified demands have appeared for historians to supply perspective on trends toward integration – that often appear to economists, political scientists, sociologists, and philosophers (as well as natural scientists) to represent profound discontinuities in the ways that ubiquitous technological, economic, and cultural forces are clearly shaping our interconnected world. This demand is supported by Western governments, uneasily aware that their powers to control societies, nominally under their rule, are transforming their claims to effective sovereignty over the populations, territories, and assets within their borders into something more tenuous than used to be the case during the long era of competing nation states. Their anxieties (eloquently fostered by 'official' intellectuals) have been compounded by recent military, technological, and economic achievements of several Asian societies, especially Japan and, the 'Tigers' but latterly (and more significantly) by the reawakening of China. The rest are perceived to be catching up and modern mercantilists (who continue to conceive the world in terms of winners and losers or clashes of civilizations) worry about the relative decline of Western hegemony. Politicians and publics turn to historians for explanations of the past and more reassuring predictions of the future.

National frameworks for political action and academic enquiry are now widely recognized as unsatisfactory. That comprehension is patently functional for natural sciences investigating environmental issues, for biological sciences concerned with the health of human populations, and for social sciences analyzing international relations, crime, migration, communications, and whole ranges of economic activity, which can no longer be regulated within the boundaries of nation states. It is also clear that national and local cultures are being subverted by advertising, by fashion, and by the popular arts, especially music. The means and media of modern transportation, travel, and communications (cheap and available to the masses) are opening up cosmopolitan discourses (usually in English) that contribute to the reshaping of personal identities around the world and transform frameworks for behavior, especially among younger generations.

Traditional, Postmodern, and Global Histories

Pressures to make, write, and teach histories of the world that might become more contemporary and relevant has encouraged the discipline to become less intellectually, politically, spatially, and chronologically confined. They are intensifying, seem irresistible, and are being welcomed as a return to the ecumenical programs of Herodotus, Vico, Voltaire, Leibniz, and von Ranke who opined "there is no history but universal history as it really was."

Intellectual arguments for resituating national and European histories into frameworks that might better address the problems of the new century have been persuasively rehearsed by proselytizers for global history in recent years – as indeed have their antitheses – recommendations to concentrate attention upon microhistories of difference, diversity, and a stance of humility toward the complex histories of 'unknowable others.' That postmodern exhortation is understandably suffused with a concern that the revival of global history will implicitly celebrate 'triumphs of the West' in science, technology, military power, and economic development. Unfortunately that has already occurred in several recent publications by journalists and even by historians whose pretensions to offer overviews of world history seem truncated in time, confined in space, and selective in areas of social endeavor emphasized for recognition as universal achievements; and whose understanding of environmental, biological, and other natural constraints on human action look unscientific.

As humanistic scholarship, global history aims to mature into a renaissance and not a return to the arrogance of Rome or to Victorian triumphalism. Modern history has matured into a capaciously catholic subject, willing to take many kinds of hitherto myopic exclusions into its narratives, to reposition knowledge of the past into more relevant spaces, models, and chronologies, and to plunder the natural and social sciences (as well as archeological remains) for insights into the evolving human condition. In finding voices and viewpoints to construct histories on a global scale, historians (alive to the second cosmopolitan enlightenment that emerged during the closing years of an otherwise dark twentieth century) have not found it difficult to detach their narratives from another politically correct but romanticized appeasement of the underprivileged on one hand, or from the implicit denigration of non-Western people and cultures that flows from Eurocentric convergence and modernization theory on the other.

In an ongoing search for strategies to teach and write compelling histories on a global scale, they will respect the preferences of their colleagues, whose postmodern anxieties about their capacities to comprehend 'others' or to offer plausible conjectures and metanarratives about major trends and significant events, leads them toward dense description, microhistories, and biography. Such flowers should certainly bloom, but the intellectual effort required to landscape the garden also demands attention and imagination from historians.

Global historians confront more challenging intellectual problems in trying to persuade the majority of their academic colleagues (writing bounded histories and bunkered in the security of national archives) that their attempts to write on a global scale could meet the standards of a discipline deeply concerned with detail and documentation that discourages generalization beyond strictly confined geographies and periods. Having rejected Toynbee's quest for spiritual guidance as a tenable mission, global historians tend to make a general case behind arresting metaphors extolling the need for histories constructed as 'maps on the largest scale,' as 'photographs taken through the lenses of wide-angled cameras,' as 'architectural frameworks' and 'as views; contemplated from space.' Many contend that the boundaries and chronologies of parish, let-alone national or continental histories seem equally problematical to define. Nevertheless, coming as most do, as recognized specialists in but one or two fields of history, global historians remain conscious of their credentials and are entirely sensitive to their epistemological vulnerability. It continues to be much easier to conduct research, claim territory, and acquire status as national or local historians.

Current Paradigms: Connections and Comparisons

Nevertheless, historiographical and methodological discussion as well as the volume of publications (off- and online) has increased exponentially. Most of the bibliography is concerned with connections and comparisons across national frontiers and cultural boundaries. Most of it displays a scientifically informed appreciation of evolutionary environmental and biological constraints on all forms of human activity, and represents a welcome reunification of natural and human history. Celebrated and influential books display a talent for the construction of synoptic histories about matters that have been important for humanity as a whole. As more historians risk writing on a grand scale, the field looks set to produce plausible, but competing metanarratives into which the overwhelming flow of parish, regional, and national histories could be reconnected.

Meanwhile two distinct approaches have already been developed. First (and central) is the model, persuasively exemplified and elaborated in several books by McNeill. He has inspired a program of historical investigation into connections across continents, countries, and geographies as well as through very long spans of time. McNeill's assumption is that such connections, interactions, encounters, and contacts with outsiders can be represented as the origins and engine of economic, social, political, military, cultural, religious, technological, and all other conceivable types of change studied by historians; and that connection can be pursued with scholarly detachment and a proper appreciation of the long chronologies required to construct histories that aim to select and analyze major forces at work in the evolution of humanity, McNeill's recommendations aim to avoid the condescension of cultures, the restrictions of time, and the arrogance of nations built into currently dominant styles of history. His (and Crosby's) books show that 'interactions' were often malign: the spread of plagues, diseases, and parasites; those waves of destructive invasions by nomads; wars of conquest, plunder, and imperial expansion; forcible conversions to alien systems of religious belief; and the subversion of indigenous traditions, cultures, societies, and communities by strangers have been part of the universal experience of suffering.

Well-specified types of connection can be investigated over significant periods of contact and transformation. Its taxonomical divisions include trade, investment, warfare, religion, migrations, transfers of useful knowledge, technological diffusion, botanical exchanges, and the spread of diseases. Proper attention is also accorded to the means of transportation and modes of communication that secured, facilitated, and cheapened contacts by sea, land, and latterly by air. Once a relevant range of major connections are traced, understood, and their significance evaluated global history will broaden our education and provide the spatial and chronological perspectives required for the appreciation of the histories of nations, peoples, localities, and groups embodied within particular histories that the majority of their colleagues study in greater depth and sophistication, but too often in isolation.

Put simply, the second major approach extends geographical catchment areas for comparative histories of topics that could be the subject of illuminating study across national boundaries, continents, and separable cultures. Although the linguistic qualifications for such research are extremely difficult to acquire, and the scholarship needed to make sense of contrasts discovered is formidable, the methodological problems are similar to those encountered for exercises conducted within European, national, or indeed regional frameworks.

Comparative history helps scholars to surmount the complexity and tyranny of local detail and seeks to offer coherent but sensitive answers to a great variety of questions that historians select for investigation. For global historians, the illumination derived from the method tends to be realized when they can concentrate upon well-defined artifacts, institutions, organizations, social practices, attitudes, and beliefs that are found in dispersed cultures, and which have already been studied in scholarly depth, in the context of particular locations and which exhibit comparable, but, more importantly, dissimilar geographical economic, political, social, and other features. Then, as Marc Bloch anticipated, the "comparative method can elicit from the chaotic multiplicity of circumstances those contrasts which were generally effective."

In order to communicate salient differences, global historians will be tempted to aggregate and average contrasts across wider spaces and larger populations (continents, oceans, and civilizations) than their colleagues who have the sources to engage in exercises in comparative history within circumscribed boundaries. For purposes of research, designed as building blocks for the production of metanarratives that are truly universal, the comparative method looks set to dominate the field for years to come. It has, for example, already generated a bibliography of global histories concerned with the family, youth, marriage, diet, housing, health, military organization, government, human rights, parliaments, nationalism, religions, fundamentalism, and revolutions, and there are serious debates on two seminal topics: gender and long-run economic growth.

Comparative Histories of Gender and Material Progress

Gender is perhaps the quintessential 'analytical category' for purposes of universal social history and it interacts with other categories including class, ethnicity, race, nation citizenship, and colonialism for purposes of constructing world histories that 'keep humanity in view.' Yet no surprise will be occasioned by the appearance in recent decades of histories seeking to encompass the world economy – going way back in time and written to expound and to explain the disparate levels of material progress achieved by countries, societies, and communities located on all continents. Most people, in most places for most of history, have been preoccupied with obtaining food, shelter, clothing, and other manufactured artifacts that they required to sustain a basic, a comfortable, and, only latterly, an agreeable standard of living.

A modern generation of historians is resituating the scientific, technological, and economic achievements of Western societies in global context. They call upon libraries of books and articles dealing with African and Asian agricultures, industries, towns, commercial networks, communications, trade, science, technologies, cultures, business organization, taxation, state systems, Government policies, etc., covering the last millennium and written in large part by historians from universities not long emancipated from imperial rule. This enormous, but still far from comprehensive, volume of historical research has, moreover, been communicated to the West by specialists in area studies from North American, European, Australian, and Japanese universities. Not long after World War II and during an era of decolonization, historians were offered the opportunity – provided by the accumulation of a large body of archival knowledge – long available in Europe and North America, but emerging in Asia, the Middle East and Africa and Latin America – to reposition their hitherto disconnected histories of wealth and poverty one against another in order to construct global histories of material progress that might have satisfied the aspirations of Montesquieu, Voltaire, Smith, and their 'enlightened' followers, and pleased Max Weber.

Recent attempts to survey this monographic literature have been dominated by scholars from Western universities. Syntheses from China, Japan, India, Africa, and the Near East would be a salutary correction to what is a plural but nevertheless European perspective. Meanwhile, scholars have carried forward a Weberian program for teaching and writing a truly global history of material progress. They have found that the quantitative evidence for such purposes would certainly not be regarded as sufficient to underpin a serious program of historical research designed to account for long-term economic success and retardation among the national economies of Western Europe, North America, and Australasia. In the absence of hard data or even of an acceptable range of contemporary observations (which could be included in a narrative designed to track chronologies of divergence in productivities and living standards between Europe and Asia), global histories of material progress continued to be marred – by three powerful Marxist and Weberian assumptions. First, that Europe's technological and economic superiority over Asia (which became first visible in the eighteenth century and then unmistakable during the nineteenth and twentieth centuries) had evolved gradually, but cumulatively, over several centuries before 1815. Over that time, Europe's political, legal, cultural, and institutional heritage for the conduct of economic activity continued to be exceptionally favorable for the promotion,

first, of Smithian growth, and, following on from that, for technological progress. Third (and in contrast to Europe), other populations lived and worked in more hostile environments and this big ecological fact restrained and often arrested long-term economic development in Asia and Africa – although our knowledge of precolonial African economies remains entirely thin.

Those who follow Weber, Marx, and a tradition of writing that can be tracked back to Montesquieu insist that Europe's competing economies had been on a potentially more promising trajectory for long-run growth for centuries. They emphasize persistent contrasts between European and Asian institutional arrangements, political systems, legal frameworks, cultural conditioning, and religious beliefs, within which economic activities continued to be embedded for at least four centuries before 1815. Cruder Weberians juxtapose the preconditions for long-run Smithian growth that they found operating efficiently in Europe (or rather in some societies in north western Europe) in sharp contrast to Asia, where political, social, familial, cultural, and other negative institutions and beliefs restrained the emergence of competitive markets for commodities, capital, land, and labor.

Their critics, who seem more deeply and widely read in the now massive bibliography of economic and social history dealing with South, East, and South-East Asia, and the Ottoman Empire, have produced samples of data (related largely to population growth, levels of urbanization, balances of commodity trade, yields per hectare and real wages) for selected regions of Asia and the Middle East that at the very least qualify any impression that labor productivity levels and standards of living afforded by European economies to their populations were discernibly (let alone significantly) superior long before the outbreak of the French Revolution. Of course, at this level of 'continental' aggregation, the selection of samples of data for comparative history remains problematical because the Asian economy includes much larger shares of the world's cultivable land, resources, income, and population than early modern Europe.

Nevertheless, economic historians are a long way from offering statistically secure comparisons for average levels of productivity and living standards. Meanwhile, the accumulating volume of research now available on Asian agricultures, industries, trades, credit systems, mercantile enterprise, transportation networks, markets for commodities, and factors of production degrades simplistic Weberian perceptions that Europe alone had evolved the political, institutional, legal, cultural, and religious frameworks required for an effective process of Smithian growth long before other continents into dubious propositions. As Marshal Hodgson observed decades ago, "All attempts to invoke premodern seminal traits in the occident can be shown to fail under close historical analysis." Braudel, Chaudhuri, Frank, Pomeranz, Goldstone, McNeil, Sugihara, Washbrook, and other historians of long-run material change concur. Jones would probably (as he did in Growth Recurring, 1988) continue to revise and reformulate some of the positions he adopted in the first edition of the European Miracle in 1981. From his comparisons of levels and types of development achieved by Europe and Asia for the early modern period, Braudel inferred that "the populated regions of the world faced with demands of numbers seem to us to be quite close to each other." But there is he continued "a historiographical inequality between Europe and the rest of the world. Europe invented historians and made good use of them. Her own history is well lit and can be called as evidence or used as claim. The history of non-Europe is still being written. And until the balance of knowledge and interpretation has been restored the historian will be reluctant to cut the Gordian knot of world history."

That acute observation applies across the entire field because histories of gender and capitalism are but two current areas of controversy that display the unrealized potential of universal history. In truth, it seems difficult to conceive of any grand theme that attracts attention from scholars engaged in writing national and local histories that could not be repositioned within wider geographies and longer temporal periods. For purposes of research, teaching, and debate, comparisons and connections are now the two dominant styles of global history. Pursued with sensitivity, they should deepen our understanding of difference and diversity. In time, they will meet demands for perspective from those trying to comprehend accelerated trends toward interdependence and integration on a world scale and allow for a less ethnocentric appreciation of the manifold achievements of more peoples, communities, and cultures over long spans of human history.

See also: Civilization, Concept and History of; Comparative History; Gender History; Globalization and World Culture; Historical Thought and Historiography: Classical Period (Especially Greece and Rome); Historical Thought and Historiography: Current Trends; Historical Thought and Historiography: East Asia; Historical Thought and Historiography: Indigenous Cultures in the Americas; Historical Thought and Historiography: Jewish Tradition; Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; Historical Thought and Historiography: South Asia; Historical Thought and Historiography: Southeast Asia; Historical Thought and Historiography: Sub-Saharan Africa; History: Overview; Modernization and Modernity in History; Science, History of; Time, Chronology, and Periodization in History.

Bibliography

Albrow, M., 1996. The Global Age: State and Society beyond Maturity. Polity Press, Cambridge.

Anderson, B.S., Zinsser, J.P., 1988. A History of Their Own: Women in Europe from Prehistory to the Present. Harper & Row, New York.

Armitage, D., 2007. The Declaration of Independence: A Global History. Harvard University Press, Cambridge, Mass.

Barnes, H.E., 1962. A History of Historical Writing. Dover Publications, New York.

Bentley, J.H., 1993. Old World Encounters. Oxford University Press, Oxford, UK.

Brook, T., Blue, G. (Eds.), 1999. China and Modern Capitalism. Genealogies of Sinological Knowledge. Cambridge University Press, Cambridge, UK.

Chaudhuri, K.N., 1990. Asia before Europe, Economy and Civilisation of the Indian Ocean from the Rise of Islam to 1750. Cambridge University Press, Cambridge, UK.

Costello, P., 1993. World Historians and Their Goals. Twentieth Century Answers to Modernism. Northern Illinois University Press, De Kalb, IL.

Crosby, A.W., 1986. Ecological Imperialism: The Biographical Expansion of Europe, 900–1900. Cambridge University Press, Cambridge, UK.

Curtin, P.D., 1984. Cross Cultural Trade in World History. Cambridge University Press, Cambridge, UK.

Crossley, P.K., 2008. What Is Global History? Polity, Cambridge.
Ching, F., Jarzombek, M., Prakash, V., 2007. A Global History of Architecture. J. Wiley & Sons, Hoboken, NJ.
Diamond, J.M., 1998. Guns, Germs and Steel: The Fates of Human Societies. Vintage Random House, London.
Darwin, J., 2008. After Tamerlane: The Global History of Empire since 1405. Bloomsbury Press, New York.
De, S.P., 2008. The Ancient World at War: A Global History. Thames & Hudson, London.
Edgerton, D., 2007. The Shock of the Old: Technology and Global History since 1900. Oxford University Press, Oxford.
Frank, A.G., 1998. ReOrient: Global Economy in the Asian Age. University of California Press, Berkeley, CA.
Goody, J., 1996. The East in the West. Cambridge University Press, Cambridge, UK.
Gills, B.K., Thompson, W.R., 2006. Globalization and Global History. Routledge, London.
Heiser Jr., C.B., 1990. Seed to Civilization: The Story of Food. Cambridge University Press, Cambridge, UK.
Hirst, P., Thompson, P., 1996. Globalization in Question: The International Economy and the Possibilities of Governance. Routledge, London.
Hodgson, M.G.S., 1993. Rethinking World History. Cambridge University Press, Cambridge, UK.
Jones, E.L., 1998. Growth Recurring: Economic Change in World History. Oxford University Press, Oxford, UK.
Kishlansky, M., Geary, P., O'Brien, P., Wong, R.B., 1995. Societies and Cultures in World History. Harper Collins, New York.
Kotkin, J., 2005. The City: A Global History. Modern Library, New York.
Landes, D.S., 1998. The Wealth and Poverty of Nations: Why Some Are so Rich and Some so Poor. Little Brown, London.
Lewis, B., 1995. The Middle East. 2000 Years of History from the Rise of Christianity to the Present Day. Weidenfeld and Nicholson, London.
Livi Bacci, M., 1992. A Concise History of World Population. Blackwell, Oxford, UK.
McNeill, W., 1991. The Rise of the West: A History of the Human Community. University of Chicago Press, Chicago, IL.
Mazlish, B., Buultjens, R. (Eds.), 1993. Conceptualizing Global History. Westview Press, Boulder, CO.
Mokyr, J., 1990. The Lever of Riches: Technological Creativity and Economic Progress. Oxford University Press, Oxford, UK.
Mazlish, B., 2006. The New Global History. Routledge, New York.
Neiberg, M.S., 2005. Fighting the Great War: A Global History. Harvard University Press, Cambridge, MA.
O'Brien, P.K. (Ed.), 1999. Philips Atlas of World History. George Philip Limited, London.
Offen, K., Roach, R.R., Rendall, J. (Eds.), 1991. Writing Women's History: International Perspectives. Macmillan, London.
Pomeranz, K., 2000. The Great Divergence: Europe, China and the Making of the Modern World Economy. Princeton University Press, Princeton, NJ.
Sanderson, S.K., 1995. Social Transformations: A General Theory of Historical Development. Blackwell, Oxford, UK.
Stavrianos, L.S., 1989. Lifelines from Our Past: A New World History. Pantheon Books, New York.
Smil, V., 1994. Energy in World History. Westview Press, Boulder, CO.
Weber, M., 1961. General Economic History. Collier, New York.
Wong, R.B., 1997. China Transformed: Historical Change and the Limits of European Experience. Cornell University Press, Ithaca, NY.
Worster, D. (Ed.), 1998. Perspectives on Modern Environmental History. Harvard University Press, Cambridge, MA.
Woolf, D.R., 2011. A Global History of History. Cambridge University Press, Cambridge, UK.

Global Migration

Jonathan Crush, Balsillie School of International Affairs, Waterloo, ON, Canada; and University of Cape Town, Cape Town, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In his work on the nature of migration in the late nineteenth century, E.G. Ravenstein highlighted a basic tension between human mobility and the desire of states to control and regulate movement in their own interests. In this article, I apply Ravenstein's insight to the contemporary global migration regime, identifying seven basic tensions that characterize the governance of mobility in the twenty-first century. These include the tension between globalization and territorial sovereignty, policies governing temporary vs permanent movement, states and citizens, citizens and migrants, exploitation and rights, origin and destination countries, and migration and development. Each of these tensions is considered in turn, highlighting the nature and impacts of each on migration and migrants.

The first attempt to systematically describe and explain spatial patterns of migration is to be found in the work of the Anglo-German geographer, Ernst Ravenstein, in the late nineteenth century (Ravenstein, 1885). Ever since, Ravenstein's ideas about migration have served as a jumping-off point in courses, textbooks, and articles on migration and as a source of hypotheses in empirical studies (Grigg, 1977; Skeldon, 1997; Alexander and Steidl, 2012). His basic argument was that observable regularities in human migration could be distilled into seven immutable laws of migration. A critic of Ravenstein noted at the time that migration is "rather distinguished for its lawlessness than for having any definite law" (quoted in Lee, 1966: 47). The idea that there are laws governing migration has long since been abandoned, but this abandonment does not mean that migration is as chaotic and disorderly as Ravenstein's detractors suggested. A vast body of geographical and interdisciplinary research affirms that it is possible to make generalizations and build theories of migrant behavior and explanations for observable patterns, processes, and trajectories of migration (King, 2012).

Ravenstein did concede, in a response to his critics, that the 'laws' of migration were subject to the vagaries of 'human agency' by which he meant government interference. He was pointing to the inherent tension between the propensity of humans to migrate and the regulatory frameworks that seek to control, dam, or divert such movements. The tension between mobility and control was becoming increasingly apparent in the late nineteenth century and became far more acute in the twentieth as states began to erect physical barriers to entry and introduced documentation regimes to regulate entry and exit. Political regimes in many parts of the world (ranging from European colonialism to apartheid South Africa to the Soviet Union to Maoist China) also imposed draconian controls on the internal movement of populations. Ravenstein's insight that there is an inherent tension between mobility and governance, is therefore, worth revisiting in the light of international cross-border migration and globalization in the early twenty-first century. Ravenstein articulated seven 'laws' of late nineteenth-century migration; in the same vein, this article focuses on seven recurrent governance 'tensions' in the early twenty-first century global migration regime.

The first tension is between greatly increased human mobility accompanying globalization and the drive of sovereign states to enforce territorial sovereignty by putting up and enforcing barriers to entry (Collier, 2013). The global migrant population (defined as people living outside their country of birth who have officially or unofficially circumvented these barriers) increased from 92 million in 1950 to 232 million in 2012. Ranked against national populations, the global 'migrant nation' would come in behind China, India, and the United States and just ahead of Indonesia and Brazil. While the number of international migrants may seem relatively small (at just over 3% of the total world population), the desire to migrate across borders (or latent migration potential) is high. The 2007–09 Gallup World Survey found that about 16% of the world's adults would move to another country permanently if they had the chance; this would mean 700 million people worldwide. What holds them back are the opportunity costs of migration and the vast institutional apparatus of gatekeeping – walls, fences, border guards, electronic surveillance, documentation regimes, and virtual 'smart borders' – which constitute formidable barriers to free movement. Many regional economic groupings of contiguous states give lip service to the idea of free movement between member states. For example, the African Union has enshrined free movement as an essential principle of continental integration. Regional subgroupings such as the Economic Community of West African States and the Southern African Development Community have taken a similar stance. However, with the notable exception of the European Union (EU), the implementation of the free movement principle invariably meets with strong opposition from member states who water down the provisions or block them altogether. At the same time, barriers to entry to the EU as a whole (Fortress Europe) are continuously maintained and reinforced (Geddes, 1999).

The second major tension in the contemporary global migration regime is between temporary or short-term and permanent or long-term migration. For much of the twentieth century, international migration was characterized by its permanent or semipermanent character. People emigrated from one country and immigrated to another with little intention of return. Destination countries built immigration

policies around the idea that people would come permanently and that entry could be regulated by moving the bar up or down. One way in which this is done is through points-based systems where potential immigrants score points for various educational, skills, language, and personal criteria. Once the threshold is crossed, immigration is possible for that individual. In this, and like, ways states reasoned that they could pick and choose who they wanted to come and match the new immigrants with domestic labor market needs and priorities. In the first half of the twentieth century, racial or ethnic criteria were fundamental to immigrant selection. However, these were progressively abandoned as the century progressed although they remained in place in countries like South Africa until the collapse of apartheid (Peberdy, 2009). Skills, education, and experience increasingly became the primary criteria for immigrant selection. There were exceptions to the preference of migrants and states to permanent immigration: for example, the Bracero program in the United States for agricultural workers, the Gastarbeiter (guest worker) program in West Germany and the mine migrant labor system in Southern Africa were all designed to ensure that migrants stayed only for as long as their labor was needed. The greatest fear of states importing temporary, often lower-skilled, workers was that impermanence would lead to permanence or, as Castles (1986: 761) put in an 'obituary' on European guest worker programs, they 'inevitably lead to permanent migration in the long run.' States that once feared this 'inevitability' have reembraced temporary migration over the last two decades. The 'resurrection' of temporary migration programs fills short-term needs at the lower end of the labor market and, when tightly regulated, ensures that migrants will go home (Castles, 2006). In Canada, for example, long seen as the epitome of a permanent immigration country, the number of temporary migrant workers now exceeds the number entering permanently as immigrants. In the Global South, circular migration has long been, and remains, the major form of intraregional mobility, Hugo (2009), for example, calculates that there are now at least 15 million temporary migrants within Asia and the Middle East working outside their countries of origin. He identifies nine different types of temporary migration in Asia and the Pacific with varying degrees of legal status, potential to move to permanent residence, ability to be accompanied by dependents, and human and labor rights (Table 1). Some temporary movements take place under formalized recruitment and contract systems (such as migration to the Middle East from countries such as India and Bangladesh and migration to the South African mines from countries such as Mozambique and Lesotho). However, much temporary, circular migration is informal (and oftentimes undocumented) in nature because of the absence of formal policies and channels.

The third tension concerns the contrasting views of states and citizens on the desirability of migration. Almost without exception, the majority of citizens favor exclusionary policies involving restrictions on migration and entry by noncitizens. The World Values Survey asked a representative sample of citizens, what government should do about migrants wishing to come to their country to work (Table 2). At one extreme was the South African citizenry where 78% would like to see strict limits or a total prohibition on entry and only 6% favor a policy that would allow entry to anyone wishing to come and work in the country. The latter figure is consistent with many other citizenries polled: less than 10% of citizens in countries as diverse as Ghana, Brazil, Indonesia, Italy, the United States, and Canada favor such a policy. Lowest of all are Australia (2%) and Germany (5%). Less than half of the citizens in all countries surveyed (except Indonesia, Malaysia, and China) favored an employment-based immigration policy. Over 40% of citizens in major immigration countries such as Italy (57%), Australia (57%), India (46%), the US (44%), and Canada (41%) favor strict limits or a total ban on entry for work. That is a greater proportion of the popular vote than is needed to form a government in many of these countries. In general, states that favor controlled immigration policies are out of step with their own citizens. If citizens could vote on the issue, there would certainly be far more restrictions on movement and much less international mobility. One of the consequences of the tension between states and citizens over the terms and conditions of the entry of noncitizens is that states tend to exaggerate the nature and threat of forms of migration that they do not authorize. Public rituals of denunciation of undocumented migrants and 'bogus' refugees are often accompanied by

Table 1 Typology of temporary migration in Asia and the Pacific

Type	*Status in destination*	*Potential to move to permanent residence*	*Ability to bring family*	*Rights*
1 Less-skilled contract labor temporary migration	Documented and undocumented	Very low	Not possible	Very limited
2 High-skilled temporary labor migrants	Mostly documented	High	Allowed	Substantial
3 Low-skilled seasonal labor migration	Documented and undocumented	Low	Not possible	Very limited
4 Working holiday makers	Mostly documented	Moderate	Possible	Substantial
5 Student migration	Documented	High	Allowed	Substantial
6 Trainee migration	Initially documented but many become undocumented	Low	Not possible	Limited
7 Border commuters/circulatory	Documented and undocumented	Low	Not necessary in many cases	Limited
8 Project tied labor migration	Documented	Low	Possible	Limited

Hugo, G., 2009. Best practice in temporary labour migration for development: a perspective from Asia and the Pacific. International Migration 47 (5), 23–74.

Table 2 Citizen attitudes to migrant entry for work, 2006–07

Country	*Impose complete prohibition on entry (% yes)*	*Enforce strict limits on entry (% yes)*	*Allow entry as long as jobs available (% yes)*	*Let in anyone who wants to enter to work (% yes)*
South Africa	30	48	16	6
India	23	23	25	30
Ghana	18	39	36	7
Zambia	11	30	44	15
Brazil	11	33	47	9
China	8	21	51	20
Italy	8	49	37	6
The USA	7	37	49	8
Germany	7	43	45	5
Indonesia	5	15	72	8
Thailand	5	16	65	14
Australia	3	54	41	2
Canada	2	39	51	8
Malaysia	2	8	72	18

draconian enforcement mechanisms, including raiding homes and workplaces, and arrest, detention, and deportation with a minimum of due process in order to 'purge' national territory of undesirable migrants (Aas and Bosworth, 2013). In the United States, for example, the number of forcible deportations more than doubled from 185 000 in 2000 to 410 000 in 2012. Over this time period, a total of 3.8 million people were deported from the country (Ternes, 2013).

The fourth tension is connected to the third, namely the nature of the relationship between citizens and migrants once the latter are on national territory. One of the fundamental characteristics of national identity in a world of nation-states is that it defines and constantly reinforces the literal and figurative boundaries of belonging, distinguishing between those who belong and those who do not, between insiders and outsiders, between us and them (Amin, 2012). Not only are insiders generally unhappy about policies that do not impose very high administrative and physical barriers to entry by outsiders; they are profoundly ambivalent about the presence of the latter once they do arrive. In Europe, for example, longitudinal data from Eurobarometer surveys between 1997 and 2003 show that 60% of Europeans believed that the limits to a multicultural society have been reached, effectively opposing further acceptance of immigrants into their countries (Coenders et al., 2003). Nearly 40% of participants opposed civil rights for legal migrants and 22% endorsed the deportation of legal migrants, above all, the unemployed. A 2012 Special Eurobarometer Report found that 73% of Europeans feel that non-EU citizens have difficulty integrating because they do not wish to do so. Interestingly, 60% also acknowledge that non-EU citizens faced discrimination in the EU. When intolerance and hostility is based purely on the 'foreignness' of the migrant, xenophobia exists. Xenophobia consists of highly negative perceptions and practices that discriminate against noncitizen groups on the basis of their foreign origin or nationality (Crush and Ramachandran, 2010). It affects all categories of migrants, immigrants, refugees, and asylum-seekers. Xenophobia is perpetuated through a dynamic public rhetoric that actively stigmatizes and vilifies migrants by playing up the 'threat' posed by their presence and making them scapegoats for social problems. Xenophobia takes on various forms but is usually accompanied by a powerful stigmatizing and stereotypical set of myths about migrants (Lucassen, 2005). Xenophobia is not simply an attitudinal orientation, however. Hostile and skewed perceptions of migrant groups generally go hand in hand with discriminatory practices and poor treatment of such groups. Acts of violence, aggression, and brutality toward migrant groups represent extreme and escalated forms of xenophobia. Xenophobia is traditionally associated with right-wing political parties in Europe, but its pernicious influence extends well beyond Europe. The present-day surge in intolerance and xenophobia in the South can be seen in migrant-receiving countries as diverse as South Africa, India, Malaysia, Thailand, and Libya (Crush and Ramachandran, 2010). South Africa, which has one of the most xenophobic citizenries in the world, was rocked in May 2008 by nationwide mob violence against migrants in which 60 people died, 600 were seriously injured, and over 100 000 were driven out of their communities.

The fifth tension relates to the extension and denial of rights to migrants. States and citizenries display a consistent pattern of resistance to according the same rights to migrants that are enjoyed by citizens. As a result, migrants are generally vulnerable to a range of abuses by governments and employers. The poor treatment of migrant workers, especially those in temporary work situations, has been well documented worldwide. The latest example concerns the situation of migrant construction workers in Qatar hired from various countries outside the region to build venues for the 2022 FIFA World Cup. Amnesty International's (2013) expose, *The Dark Side of Migration*, systematically documents widespread discrimination against a general backdrop of discriminatory attitudes toward migrant workers. These abuses include not being paid for long periods of time, inhumane working conditions, and squalid living arrangements. According to Amnesty International many migrants are in "severe psychological distress due to the treatment they had received and their sense of powerlessness to resolve their own situations." At a global level, the struggle for migrant rights pits migrants, migrant organizations, and human rights advocacy groups against governments and employers. The general attitude of states toward rights for migrants is evident in the tortuous history of the 1990 UN International Convention on the Protection of the Rights of All

Migrant Workers and Members of Their Families (ICRMW). Cholewinski et al. (2009) characterize the ICRMW as "a vital instrument to ensure protection of migrants' human rights" and yet it has been treated with 'marked indifference' by states. In 20 years, the Convention has been ratified by only 41 states, none in the North and no major migrant-receiving countries. By contrast, the 1989 UN Convention on the Rights of the Child was ratified by 193 states within 5 years. A global survey of policy responses to the ICRMW revealed a wide range of reasons why states refuse to ratify the Convention. There are concerns about the costs of implementation and, in some cases, specious arguments (made by countries like Canada) that the Convention is unnecessary because domestic legislation provides all the same protections, making ratification redundant. The overall message is clear, however. States do not want to accede to any kind of international normative framework that would potentially divest them of complete discretion over how they choose to treat migrants and what rights they seek to provide and deny.

The sixth tension is between countries of migrant origin in the South and countries of migrant destination in the North. This tension takes various forms but has been most intense in relation to the 'brain drain' of skilled migrants from South to North (Kapur and McHale, 2005). Tanner (2005) argues that attitudes toward brain drain in countries such as Canada are "at best unconscious, at worst indifferent." The actual impacts of the brain drain on countries of origin are hotly debated. As Goldin and Reinert (2006: 183) then of the World Bank suggest: "While high skilled migration in sectors such as IT seems to have played an integral role in helping spur economic development in a few source countries, high-skilled migration in other sectors – health and medicine, in particular – have done considerable damage to source countries." Most absence effects discussed in the general literature are negative. These include larger skill premiums that increase inequality; fiscal losses through lost tax revenue; diminished scale economies; loss of role models and spillover knowledge from most-skilled to lesser-skilled individuals; loss of entrepreneurs; and changed comparative advantage (Kapur and McHale, 2005; Tanner, 2005). Most attention has been paid to the negative impact of the migration of medical professionals on countries of origin. These countries, with underresourced health systems, the highest rates of infant mortality, and ravaged by infectious disease, cannot afford the loss of health care personnel. In Ghana, for example, half of the physician graduates between 1985 and 1994 had left the country within 4.5 years, and three-quarters had left within 9.5 years. The whole 2002 graduating class of doctors had left or planned to leave. Awases et al. (2003) provide information on reasons for and patterns of migration (with reference to doctors, nurses, pharmacists, and midwives), effects on health care, and policies undertaken in six countries for the period 1991–2000. They conclude that the skilled migration of health professionals from Africa has adversely affected the quality of care to the poor, given that there are fewer personnel available, heavier workloads, fewer drugs, and performance of duties by underqualified personnel. Overall they conclude that "the importance of the impact of the loss of human resources in terms of skills and numbers in relation to the total size of the health workforce cannot be overemphasized." Among the demands of countries in the South are a cessation of recruiting, closing of the doors to skilled migrants from developing countries, and financial compensation for the loss. The countries of the North have consistently rejected these demands.

The seventh, and final, tension concerns the relationship between migration and development (Skeldon, 1997). Global policy debates about this relationship have intensified dramatically since 2000. Key moments include the appointment and report of the Global Commission on International Migration, two UN High Level Dialogues on International Migration and Development (in 2006 and 2013), the establishment of the annual Global Forum on Migration and Development in 2007, and the formation of the interagency Global Migration Group. The relationship between international migration and development is often conceptualized as a 'nexus' to denote the reciprocal nature of the linkage: migration has impacts on the development process; similarly development can influence flows of migration. Most commentators agree that the links are exceedingly complex and hard to unravel. Three general perspectives on the nexus can be identified in contemporary policy debates: (1) migration has a positive impact on the development of countries of origin in the Global South; (2) migration has negative implications for the development of those countries; and (3) the relationship between migration and development is a 'balance sheet' of pros and cons. de Haas (2010) argues that since the 1960s, the pendulum has swung backward and forward between optimistic and pessimistic views of the relationship. The current, post-2000, phase of optimism focuses on two main features of the contemporary global migration regime. First, there are the massive global flows of remittances from migrants to their countries of origin. In 2013, these flows are estimated at USD410 billion or three times the amount of Official Development Assistance. Second, many migrants maintain close social and economic ties with their countries of origin that can have development-related implications and impacts. Migrants are increasingly seen as diasporas, and countries of the South are actively attempting to engage 'their' diasporas in development activities (Mohamoud, 2010). Another, contentious, aspect of the relationship is whether development leads to a decline in migration. One line of argument is that poverty causes migration and that development will therefore dampen migration flows. In the case of the EU, the particular policy outcome is that the 'root causes' of migration should be identified and development assistance targeted to reduce 'migratory pressure' on Europe. There is little evidence that this approach works or is worth pursuing. Indeed, many analysts argue precisely the opposite that development may actually increase rather than decrease migration (de Haas, 2007; Bakewell, 2008).

Researchers are active in all seven of the tension areas identified in this article. However, three are particularly relevant to understanding the geographies of international migration and are consistent with Ravenstein's concern with how state actions 'interfere' with migration flows. The first concerns the massive effort – measured in resources, personnel, and surveillance technologies – expended by states in the physical and symbolic construction and reinforcement of barriers to movement, the strategies of circumvention by migrants, and the coercive efforts by states to 'cleanse' their territories of

'outsiders' (Fassin, 2011; Mountz et al., 2013). The second concerns the flow of skilled migrants from the Global South to the Global North, which is facilitated by state interventions in the form of proactive immigration policies and recruiting. These policies and their impacts are well studied. More controversial is the growing debate about impacts between 'brain drain' and 'skills flow' researchers (Mills et al., 2011; Clemens, 2013). The latter group takes issue with virtually every aspect of arguments critical of the brain drain – including the very term itself – and contend that the downstream benefits of migration to countries of origin easily outweigh the loss of skills. This debate has rapidly become an ideological contest with divergent policy prescriptions but little hard evidence either way. And third, with regard to the migration and development relationship, there is a need to break with the pervasive view that only migration from the South to the North has development impacts in the South. South–South migration is an increasingly important global phenomenon, and its impacts on development in both countries of origin and destination is starting to emerge as a major research issue (Anich et al., 2014).

See also: Globalization: Geographical Aspects; Immigration Policy; Migration, Economics of; Population Geography; Refugees: Geographical Aspects.

Bibliography

Aas, K., Bosworth, M. (Eds.), 2013. The Borders of Punishment: Migration, Citizenship, and Social Exclusion. Oxford University Press, Oxford.

Alexander, J., Steidl, A., 2012. Gender and the 'Laws of Migration': a reconsideration of nineteenth-century patterns. Social Science History 36 (2), 223–241.

Amin, A., 2012. Land of Strangers. Polity, Cambridge.

Amnesty International, 2013. The Dark Side of Migration: Spotlight on Qatar's Construction Sector Ahead of the World Cup. Amnesty International, London.

Anich, R., Crush, J., Melde, S., Oucho, J. (Eds.), 2014. A New Perspective on Migration in the South. Springer, New York.

Awases, M., Nyoni, J., Gbary, A., Chatora, R., 2003. Migration of Health Professionals in Six Countries. World Health Organization, Brazzaville.

Bakewell, O., 2008. 'Keeping them in their place': the ambivalent relationship between development and migration in Africa. Third World Quarterly 29, 1341–1358.

Castles, S., 1986. The guest-worker in Western Europe - an obituary. International Migration Review 20 (4), 761–778.

Castles, S., 2006. Guest workers in Europe: a resurrection? International Migration Review 40 (4), 741–766.

Cholewinski, R., de Guchteneire, P., Pécoud, A. (Eds.), 2009. Migration and Human Rights: The United Nations Convention on Migrant Workers' Rights. UNESCO, Paris.

Clemens, M., 2013. What Do We Know about Skilled Migration and Development? Policy Brief No. 3 Migration Policy Institute, Washington, DC.

Coenders, M., Lubbers, M., Scheepers, P., 2003. Majorities' Attitudes towards Minorities in European Member States: Results from the Standard Eurobarometers 1997, 2000, 2003 (Report 2). EUMC, Vienna.

Collier, P., 2013. Exodus: Immigration and Multiculturalism in the 21st Century. Allen Lane, London.

Crush, J., Ramachandran, S., 2010. Xenophobia, international migration and development. Journal of Human Development and Capabilities 11 (2), 209–228.

Fassin, D., 2011. Policing borders, producing boundaries: the governmentality of immigration in dark times. Annual Review of Anthropology 40, 213–226.

Geddes, A., 1999. Immigration and European Integration: Towards Fortress Europe? Manchester University Press, Manchester.

Goldin, I., Reinert, K., 2006. Globalization for Development. World Bank and Palgrave Macmillan, Washington, DC.

Grigg, D., 1977. E.G. Ravenstein and the 'Laws of Migration'. Journal of Historical Geography 3 (1), 41–54.

de Haas, H., 2007. Turning the Tide? Why development will not stop migration. Development and Change 38 (5), 819–840.

de Haas, H., 2010. Migration and development: a theoretical perspective. International Migration Review 44 (1), 1–38.

Hugo, G., 2009. Best practice in temporary labour migration for development: a perspective from Asia and the Pacific. International Migration 47 (5), 23–74.

Kapur, D., McHale, J., 2005. Give Us Your Best and Brightest: The Global Hunt for Talent and Its Impact on the Developing World. Center for Global Development, Washington, DC.

King, R., 2012. Geography and migration studies: retrospect and prospect. Population, Space and Place 18 (2), 134–153.

Lee, E., 1966. A theory of migration. Demography 3 (1), 47–57.

Lucassen, L., 2005. The Immigrant Threat: The Integration of Old and New Migrants in Western Europe since 1850. University of Illinois Press, Chicago.

Mills, E., Kanters, S., Hagopian, A., Bansback, N., Nachega, N., Alberton, M., Au-Yeung, C., Mtambo, A., Bourgeault, I., Luboga, S., Hogg, R., Ford, S., 2011. The financial cost of doctors emigrating from sub-Saharan africa: human capital analysis. British Medical Journal 343, d7031.

Mohamoud, A. (Ed.), 2010. Building Institutional Cooperation between the Diaspora and Homeland Governments in Africa. African Diaspora Policy Centre, The Hague.

Mountz, A., Coddington, K., Catania, R., Loyd, J., 2013. Conceptualizing detention: mobility, containment, bordering, and exclusion. Progress in Human Geography 37, 522–541.

Peberdy, S., 2009. Selecting Immigrants: National Identity and South Africa's Immigration Policies, 1910–2008. Wits University Press, Johannesburg.

Ravenstein, E., 1885. The laws of migration. Journal of the Statistical Society of London 48 (2), 167–235.

Skeldon, R., 1997. Migration & Development: A Global Perspective. Longman, Harlow.

Tanner, A., 2005. Emigration, Brain Drain and Development. Migration Policy Institute, Washington, DC.

Ternes, B., 2013. Immigration nation: raids, detentions and deportations in Post-9/11 America. Ethnic and Racial Studies 36 (6), 1119–1120.

Global Positioning System

Mauro Salvemini, Sapienza Università di Roma, Rome, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The contemporary society benefits from using global positioning system (GPS) for locating almost everything in the world. Data originated by GPS technology are easily usable and can be integrated in a large variety of systems for purposes ranging from safety and territorial intelligence to land use change detection, from navigation to mapping, from social networking to rescue intervention. Many personal devices, such as mobile phones, contain GPS microchips allowing the common user to play the role of surveyor and produce an enormous amount of geographic data and georeferenced data. GPS is the electronic device, together with the mobile phone, which underpin the paradigm of a liquid society.

A global positioning system (GPS) is a radio wave receiver used to provide coordinates that give the exact position of an element in a certain space. The radio waves are emitted by a constellation of orbiting satellites, which function as a reference system for GPS. In reality GPS is based on technology, which is capable of reaching very high levels of precision, even to less than a millimeter, if certain aspects of the technology are exploited and if the period of observation can be maintained for a sufficient amount of time. GPS technology is deeply influencing contemporary society. Only recently have mobile phones and tablet computers been equipped with GPS chips; as well, many new car models and other public and private transportation systems now have built-in navigation devices. The stand-alone GPS with a built-in system for managing downloadable maps is now so affordable that it has become the common companion of travel, leisure, and sport time.

GPS gives human beings the ability to access the geographical coordinates of any object on the earth's surface and to increase our knowledge of the object's characteristics with geographical data, collected in real time and directly from the field at very low cost.

Along with extensive and easy use of location-based data and linked information and images, problems arise: the limitation of privacy and a reduction in people's orientation abilities are prime examples. On the one hand, personal wearable devices such as glasses and watches may give directions, show and record images, and interact with the owner because of knowing, through GPS technology, the spatial coordinates of the user. On the other hand, sharing a person's specific location on the web through social networks or through the street view, his house and parked car, or even himself entering garage is a potential threat to privacy.

GPS helps solve three fundamental questions that from the beginning of time humans have sought to answer: where am I now, where am I going next, and what is the best route from here to there? The observation of the sun and stars, and then later the use of the compass, the sextant, navigational maps, and charts of the heavens have all been essential aids and instruments in this never-ending quest. GPS is the most recent and easy to use contribution to this search and changes the expertise needed to find in the real world and to map the geographic coordinates of everything; in short, everyone can produce geographic information; anyone can generate the path, street, point, and area or location of any object. As a result, volunteered geography (VG) is steadily developing a large database of volunteered geographic information published on the WWW by user communities that subscribe to the open-data geographic information approach. While changing the praxis of having geographic information available for use by citizens, VG is affecting, differently in various nations, the authoritative process of public administrations providing maps for institutional, legal, and financial purposes. The provided data and map precision are consistent parts of government-provided official cartography; the use of GPS to create maps does not automatically solve issues of data accuracy or data transformations. GPS helps people primarily to move about in space by providing their location relative to other objects, which most often have fixed locations but can also be mobile. The semantic difference between location and position has to be taken into account when analyzing physical and social – cultural phenomena.

Origins and Instances

Over the years the need to answer to location questions has been the driving force behind the development of many civil and military applications. During World War II, one of the nations in combat realized that, even if it managed to build warships relatively rapidly, it was much more difficult to find suitable commanders for the ships. Such positions required candidates with a high level of experience, especially for mapping the ship's course. From this dilemma long-range radio navigation (LORAN) was born, a system essentially for use at sea, based on the reception and phase difference of radio waves transmitted from two antennae. These two stations transmitted a continual flow of coordinates to navigators equipped with the correct passive antenna, which they then used to pinpoint their exact location on the earth's surface. This facilitated the commander's decision making enormously, as he knew the ship's position at any time and, by using additional electronic aids or by simply mapping the points already crossed on a nautical chart, also knew exactly where the ship was headed. The GPS was the natural evolution of LORAN.

At any given moment and point in the area covered by the system, GPS is capable of providing the three-dimensional

 http://dx.doi.org/10.1016/B978-0-08-097086-8.72022-8

coordinates (x, y, z, or latitude, longitude, and height). Rapid progress in electronic component design, has miniaturized GPS equipment to the extent that any personal wearable device can contain GPS and can provide the wearer with her exact position, the route she followed, and where she has to go.

In flights, ships, trains, buses, and cars, a screen displays a schematized map of the territory and, using GPS, a pointer indicates current position, path, and, among other information, how much time remains before reaching the destination. Users draw comfort from knowing where they are and when they will arrive. They may be able to use the features mapped on the screen to recognize physical characteristics of the surrounding territory and look to specific points of interest. Increased safety can be an added benefit of GPS use.

Aside from the passive and temporary use of GPS to determine position, route, and destination, the system is also important for activities with long-lasting and consequential effects. The most important example is the use of GPS in cartographic, topographic, and geodetic practice and research, and, by extension, to many forms of inquiry in social science. Through the use of this device, mapping becomes easy even in difficult mapping situations such as *favelas*, migrant communities, and natural constituents (forest, swamps, polluted areas, etc.), which had long posed challenges for cartography and thematic mapping. With GPS, the earth's surface becomes an enormous digitizer, and any point can be identified instantly by its three-dimensional coordinates and archived in a suitable information system, potentially increasing the intensity of sampling events and objects in space and time by several degrees of magnitude. Such measurements affect research and practice in archaeology, economics, geography, planning, and sociology, as well as in the natural sciences. By determining precise coordinates of points on land or on the surface of water, GPS makes it possible to update maps by locating physical and social features from which GPS gathers data instantaneously and/or detecting changes that have occurred in reality but have yet to be recorded (e.g., new construction, physical changes in the environment, transient events, or those in constant transformation such as people, vehicles, coastlines, woods, glaciers, and so on). GPS plays a unique role in disaster and relief management giving exact location of people, settlement components, infrastructures, and disaster origins. For example, volunteer citizens used GPS to map the Haiti earthquake zone in 2010 as they have in many other community-originated initiatives during and after disasters such as flooding, fires, and hurricanes.

The General Principles of GPS Functioning

GPS receivers use radio waves to determine positional coordinates. The radio waves are emitted by a constellation of orbiting satellites, which function as a reference system for GPS. The constellation consists of 24 satellites, which are located in six orbits, inclined at an angle of approximately 50°, at a height of around 20 000 km. Known as NAVSTAR, the constellation is the property of the US government and is controlled and managed by the Department of Defense. The orbits and the distances between satellites are organized in such a way that, ideally, every point on the earth's surface can be seen by at least five satellites at the same time, although this is not always the case in practice. The GPS carries out the triangulation of three satellites visible at the same time above the horizon, measuring the distance that separates it from each of these by calculating the time between each satellite (transmitter) to the GPS itself (receiver) using precise time signals emitted by the satellites. The triangulation allows the GPS to determine the latitude and longitude, though a fourth satellite is also necessary to determine a third coordinate giving the height at which the GPS receiver is located. To measure the time required by the signal to travel the distance from each satellite to the receiver, it is essential to use extremely precise equipment to transmit the signal, to calculate the time lapse, and also to pinpoint with extreme accuracy the position of each satellite in space. To this end each satellite is equipped with an atomic clock, and its position is constantly monitored and regulated by a series of earth stations.

In place at the end of 1990, GPS achieved initial operational capability by a full constellation (24 satellites) by December 1993, having been massively used for the first time during the Gulf War, and was able to provide the standard positioning service. Since 1993, the GPS system has been the object of several US governmental initiatives.

It is essential to consider the problem of errors incurred by external agents when dealing with instruments as sophisticated as the GPS. To determine the exact time the signal takes to reach the GPS receiver, it is also necessary to understand all the modifications and distortions the signal may undergo while passing through the ionosphere and the atmosphere. The causes of error and/or the malfunctioning of the system are varied, and can range from atmospheric conditions to the position of the satellite constellation in the sky and their visibility from the receiver. The location of the GPS receiver itself may cause problems. For example the receiver may be unable to detect a satellite if it is located under a canopy of thick vegetation or in a city canyon, where its horizon may be modified by the height of surrounding buildings.

Those managing the system itself may introduce certain deliberate errors. GPS, for example, has been subject to selective availability (SA), a randomized error deliberately incurred in the time signal from the satellites to degrade the positional accuracy of the receiver to the order of 10–100 m. Acting on a declaration made in 1996, on 1 May 2000, US President Bill Clinton announced the end to the use of SA in GPS, with the immediate effect of increasing and encouraging the use of GPS for civil and commercial purposes throughout the world. In fact, the demand for GPS precision is high among civilian users. Every context has its own GPS accuracy measurement needs. In an urban situation for example, the user may wish to know what side of the road he is on or which traffic lane to be in, and in the development of technical charts, cadastres, and engineering plans, the precise coordinates of the point on the ground are needed in order to represent it on a topographic map; accuracy needs increase even more in geodetic applications. In reality, GPS is capable of reaching precision of less than a millimeter, if certain components of the technology are exploited and if the period of observation can be maintained for a sufficient length of time.

The precision of GPS measurements is also influenced by factors such as whether the station is mobile or fixed (at least

for the time being) and whether the receiving station functions alone or in connection with other stations. The ability for a station to operate in conjunction with at least one other fixed station provides optimal precision. This technique is called differential GPS, and it is based on the registration of signals transmitted by the satellites and simultaneously picked up by a fixed or base station (master), whose precise geographical coordinates are known, and by a mobile station (rover), located anywhere within a range of around 150 km from the master station. Processing the data received by the master station eliminates the errors in reception by the rover. Local public administrations have been aware of the social demand for and usefulness of GPS data and have created local master stations as a public service to support user communities.

GPS models suitable for specific applications are available in professional and consumer markets; some are equipped to receive radio signals and different location signals produced by specific systems such as the Russian-developed GLONASS. The issue of automatic location detection has reached such relevance during these years that China, India, and Japan have developed their own systems covering various areas. The Galileo global system developed by the European Union and other partner countries is planned to be fully deployed by 2019. Societal needs are important in Galileo deployment, which is designed to support industrial development and social progress.

GPS and other mentioned positioning systems do not work indoors because the signals encounter difficultly being received in buildings, built areas, and closed spaces. This limitation is presently solved by indoor positioning systems, which by using different or combined technologies enable one to survey the location of objects indoors. This technology is expected to develop rapidly along with the evolution of mobile devices (e.g., mobile telephones) and the wide use of wireless networks and sensors while the social implications discussed below will increase.

The Relationship between GPS and Information Systems

The most relevant aspect of GPS to the information society in 2013 is its ability to provide the geographical coordinates of any object on the earth's surface and to increase our knowledge of the object's characteristics with ancillary data, collected in real time and directly from the field, at very low cost. Without doubt, the geographical or location component is an essential part of current databases and is generally deduced from a classification of objects by name (city, road, house number, etc.) or is represented by the cartographic coordinates of the single object gathered from maps (e.g., the centroid of a manhole cover, telephone or electricity pole, the axis of a water or gas pipe). In each case the geographical component, if obtained by traditional means, is approximate, expensive, or unreliable in relation to both time and space: Time, because things can change from one moment to the next (an electricity pole can be moved relatively rapidly in reality but not so rapidly on the charts representing it); Space, because an element located in a certain territory undergoes continual transformation; the shores of a lake or a coastline for example can be modified by environmental factors.

Conferring a geographical address on each item of data allows for the application of the analytical, inquiry, and management functions of a geographic information system (GIS). Real-time availability of the geographic address means that the spatial analysis can be carried out instantaneously in the field. Typical spatial analysis functions include: investigating relationships of proximity to, or distance from the object, superimposition of information layers, the search for the shortest or any route that satisfies specific conditions, identification of spatial environments that group objects with specified characteristics, monitoring transformations of the object, and so on. Spatial analysis is based on the amount and the accuracy of geographic data and their attributes. GPS is dramatically increasing the possibility to easily input data into databases and on the other hand it is enabling the descriptive attributes of real world to be geographically referred. This makes spatial analysis substantially more efficient. Typical functions of inquiry analysis are those that identify the information linked to an object by selecting that object from a spatial database rather than selecting it from the alphanumeric database and obtaining thematic maps ad hoc. Typical functions of information system management are those that enable continuous updating of the constituent elements by refreshing the data and integrating the geographic database whenever necessary.

Constant information flux poses a major problem in an information-dependent society, and updating coordinates in three-dimensional space is extremely challenging. Connecting GPS to other techniques helps solve this problem. For example, we can consider digital cartography, necessary for a decision-making process in a territorial or urban context. Decision makers rely on territorial information systems and the digital cartography that are archived in their databases. But what happens when digital cartography no longer reflects reality? A city or region can change very rapidly, and if the digital map is not updated, the information system using those spatial data will no longer be very useful. Invalid information may be merely inconvenient, or it can lead to serious social or legal complications. GPS helps greatly, but in addition to the maps that need continual updating, the geographic addresses of data points are particularly useful in social science. Among these are census tracts, political districts, administrative units, pollution zones, and epidemiological information, or research concerning traffic and movement. More controversially we may monitor individuals.

Examples of the Use of GPS and Its Social Implications

The development of GPS and its diffusion is a step toward having geographic address (x, y, z coordinates in space, or longitude, latitude, and height) and exact time as a component of any type of information. GPS equipment is widely used by utility companies to identify the components of their networks (poles, pylons, wells, technical components, etc.) and has long been used to identify the position of submarine oil wells from offshore oil platforms. Additionally anyone, at any moment in

time, can visualize the location or destination of any vehicle transmitting its coordinates to be collected in real time by GPS. For example when mounted on a public bus, a GPS can supply immediate information on the vehicle's location, the journey time, and anything else that is necessary to satisfy the needs of the driver, passenger, or manager alike. For some time insurance companies have been using this system to reduce the incidence of car thefts, and traffic management systems, publically available on WWW, now collect data from cars, which constantly communicate their own location. Car manufacturers now equip cars with navigators: electronic navigation systems containing a digital map of the area in which the car is traveling and on which the positioning and movement of the vehicle itself is traced, thanks to the continual tracking of its location through GPS. Integrated with other systems, GPS permits a vehicle to proceed automatically, without the aid of the driver or pilot. Systems already exist in marine and aeronautical applications that are capable of functioning as autonomous autopilots. In agriculture so-called precision-farming is transmitting accurate GPS-produced data to automatic farm machinery. Exact details of military applications are not available to public knowledge, but it is easy to guess that the same basic technology is used with more sophisticated developments. An internal GPS in a cellular phone provides the user with his or her exact coordinates, and can be used to visualize the location of similarly equipped phones in social network systems. Such devices are in regular use at large conventions to facilitate (or avoid) meetings. The same device alerts emergency services to incidents on lonely stretches of unfamiliar road or, worn as clothing, GPS may guide the sight-impaired or guard children against becoming lost. All fields that use geographic coordinates may derive benefits from the use of GPS.

As mentioned above, two principal problems are linked to the use of GPS: a reduction in the capability of orientation and limitations on privacy.

The problem of the reduction in the capacity to orientate oneself is due in part to the use of digital coordinates, rather than analog information such as: go to the square, turn right at the newsstand, go straight on to the white building, etc. With GPS a person can supply a visitor with the coordinates of his or her position by simply pressing a button; the visitor receives the coordinates and transfers them to his autopilot or navigation system, and the vehicle navigator directs the visitor to the desired location. Nothing will remain in the visitor's memory, unless he has paid attention to the features of his route and not simply paid attention to the screen. It is a question of personal choice whether to learn to navigate in one's environment or to rely wholly on digital assistants. The so-called systems of computerized street view together with GPS use or the use of GPS-originated coordinates may help to understand, analyze, and better use space, especially in a city environment; nevertheless navigation through these systems does not substitute for the reality experienced by walking.

The privacy problem is a crucial aspect of the use of GPS in society. The power of knowing at any given moment the location of someone equipped with a GPS can be a source of enormous security in certain situations such as the occurrence of an accident. But knowing someone's movements is potentially a great intrusion on personal privacy. For example, with GPS and GIS we could discover the shop, office, or facility someone is currently visiting, or recreate their activities during the past 24 h or longer. The surveillant society is the flip side of the information society. Separating valuable monitoring from unwarranted intrusion into personal privacy, and deciding who has the power to distinguish between them, poses a major challenge to society.

Summary

GPS systems and GPS receivers (integrated in mobile phones, tablets, cameras, watches, and even glasses) have become practically ubiquitous. Citizens are pushed to contribute through the volunteered and open-source paradigm to survey GPS data, to share geographic data, and to deliver data to the WWW making them available to web community.

Some of the benefits deriving from the enormous quantity of publicly originated GPS data have already been realized and some others are expected to emerge as the WWW develops together with communication systems.

Cartographers, engineers, planners, archaeologists, and geographers have drawn enormous advantages from the active use of GPS and will continue to enjoy the benefits of a new technology that allows for rapid and much less costly collection and updating of geographic information. The integration of engineering designs realized with CAD, digital mapping, and GPS, introduces the possibility of devices capable of elaborating designs for civil engineering and proceeding through interactive surveys, to the construction of roads, buildings, etc. GPS also integrates well with the analysis of data that have been collected via satellite for the observation of the earth, in that it allows for the exact identification of the coordinates of specific features on the satellite image. Social applications range across all walks of life. Inevitably, as with all technological innovations, its use should be evaluated critically, and the limits of its appropriate use determined.

See also: Applied Geography: A US Perspective; Cartography; Geographic Information Systems; Location, Absolute and Relative; Navigation; Spatial Analysis in Geography; Spatial Data; Time-Space in Geography; Transportation Geography; Transportation Planning; Urban Planning: Methods and Technologies.

Relevant Websites

http://www.colorado.edu/geography/gcraft/notes/gps/gps_f.html – The University of Colorado at Boulder Global Positioning System Overview.
http://www.trimble.com/ – Trimble.
https://www.myconnectedsite.com/site/trimbletraining/TrainingVideos/ – Trimble Training Videos.
http://www.navtechgps.com/ – NavtechGPS.
http://www.navtechgps.com/departments/books/ – NavtechGPS Bookstore.

Global Public Policy

Inge Kaul, Hertie School of Governance, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although we live in a globalizing world there exists so far no systematic conceptualization of global public policy. Accordingly, this article explores the question of whether such a policy has emerged in practice. It finds that qualifying initiatives do exist, and in rising numbers, but that they have not yet coalesced to form a new, identifiable policy strand. The conjecture that the article offers for further research and debate is that defining a sovereignty-compatible rationale for global public policy is the critical missing link. This will first require well-framed proof that in global-issue areas marked by interdependence, fair and effective international cooperation is in states' enlightened self-interest and, second, consensus on a notion of 'smart' sovereignty, that is, an exercise of sovereignty that seeks to combine the pursuit of national interests with respect for the sovereignty of other states.

Awareness is growing among policy analysts that there exists no well-established concept of global public policy (Reinicke, 1998; Severino and Ray, 2010; Stone, 2008). This despite the reality that more and more policy challenges are of a global, border-transgressing nature – just think of the excessive volatility in international financial markets witnessed since 2008 or of global climate change. Indeed, examples of challenges that pay no heed to national borders are rampant; among them are also illicit international trade, an accompanying rise in crime, cross-border migration, Internet governance, and the risks of cyber-war, not to mention the persisting challenges of nuclear proliferation, communicable diseases, and global inequity.

So, how are these policy challenges actually being addressed? Could it be that a new type of policy, a global public policy, has evolved? Perhaps its development has progressed incrementally and in an ad hoc fashion, pushed by a series of crises that have called for exceptional, innovative measures? If so, what are the distinctive features of this global public policy? And what might it take for theory to catch up with policy practice?

These are the questions at the center of this article.

Section How to Recognize Global Public Policy outlines the approach. It suggests that to recognize global public policy when we see it, we need an analytical grid that helps us to spot relevant phenomena. Accordingly, and considering that many, if not most global challenges possess the characteristics of a global public good, it proposes to define the purpose of global public policy as the provisioning of global public goods. Furthermore, taking into account the particular governance challenges that these goods pose, as well as current trends in global power relations, Section How to Recognize Global Public Policy identifies four functions that a global public policy would need to fulfill, namely (1) averting risks of dual – market and state – failure in the presence of global public goods; (2) promoting global fairness; (3) facilitating the actual 'getting to the good'; and (4) supporting global systemic sustainability.

Drawing on select empirical evidence, Section Global Public Policy: A Reality? then reviews recent policy responses to global challenges to determine whether those appear to have contributed to any of the four functions. The review reveals that qualifying initiatives exist, and in rising numbers. Perhaps it is this fact – the take-off of global public policy at the operational level – that gives rise to sovereignty concerns and inspires, at the same time, a reluctance to cooperate especially among state actors.

As Section The Missing Link: A Sovereignty-Compatible Rationale for Global Public Policy argues, this leads to a policy trap: the sovereignty paradox. The more state actors shy away from international cooperation to shield their sovereignty, the more often they fail to meet global challenges, losing, piece-by-piece in these failures, the very thing they seek to protect – and allowing problems to linger and assume crisis proportions.

Thus, what it might take for policymakers and their constituencies to break out of this paradoxical situation might be the fashioning of a sovereignty-compatible rationale for global public policy.

In more detail, the argument follows.

How to Recognize Global Public Policy

The term 'policy' is being used here in a broad sense. It denotes all the actions taken to achieve desired goals. Accordingly, we should first specify the overall purpose and functions of a global public policy, taking into account the nature of the challenges to be addressed and what might be needed to address them under the present policy-making realities.

If we define the overall purpose of global public policy as helping to resolve global challenges, we ought to bear in mind that many, if not most of those challenges possess the properties of global public goods or might, as a corrective step, require inputs with these attributes. Moreover, these challenges need to be addressed amidst shifting power relations, specifically the intensifying trend toward multipolarity and a balance between state and nonstate actors that has, in recent decades, tipped in favor of the latter.

Taking these factors into account, it appears that a global public policy would, among others, need to perform the following four functions.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.75019-7

Function 1: Averting Risks of Dual Actor – Market and State – Failure

As Box 1 notes, the defining property of public goods, including global public goods, is their publicness in consumption, that is, they affect everyone, for better or worse. This characteristic poses an important risk for the provisioning of these goods, as it may tempt individual actors into free riding. They may conceal their true preferences for these goods, waiting for others to step forward and provide them. Then, when the goods emerge in the public domain, the free-riders may enjoy them, too – free of charge.

Because some individual actors succumb to such free riding temptations markets are likely to fail in ensuring an adequate provisioning of public goods. Therefore, as economists say, the existence of public goods constitutes a condition of market failure. As the risk of free riding tends to increase with the number of actors involved, global public goods are especially prone to underprovision due to market failure.

At the national level, the state can intervene and potentially correct public good-related market failure due to its unique ability to create incentives. At the international level, however, there is no full equivalent of state institutions. Moreover, when appearing internationally, state parties, for the most part, pursue national interests which may or may not overlap with global concerns. Hence, internationally, they, too, are actors with particularistic – quasi-private – interests, and, as studies on international cooperation have repeatedly shown, they, too, succumb to free riding temptations.

This implies that global public good-type challenges run the risk of dual actor failure: market failure and state failure. (The type of state failure identified here is not to be confounded with the types of government failure that public choice scholars tend to focus on, namely the self-interested behavior of politicians and bureaucrats. The state failure referred to here has systemic roots that lie in the fact that the present world order is composed of individual sovereign nation states.)

Thus, one function of global public policy would need to be averting or at least minimizing the risk of such dual actor failure in the presence of global public goods, most notably the risk of state failure. Preventing the latter is of special importance, as effective multilateral cooperation is often needed for states to 'tame' globalizing markets, as well as themselves.

Function 2: Promoting Global Fairness

Alone, function 1 might not suffice to bring about the level of willingness to participate required to make cooperation work effectively under today's polycentric conditions, that is, without pressure from a hegemon or club of major powers. Under these conditions, an important incentive for actors to engage in a public policy initiative is that the engagement clearly 'pays' – economically, politically, or otherwise.

Considering the wide differences and disparities that exist in the world, the most feasible way for parties to perceive international cooperation as fair and legitimate is perhaps process fairness: to afford all concerned an opportunity to have an effective say in matters that affect them, while striking a balance between representativeness and decision-making efficiency.

Risks of policy stalemates or even breakdowns due to irreconcilable differences in the negotiating parties' preferences could perhaps be reduced by coupling process fairness with positive-sum thinking. The latter approach is characterized by willingness to seek mutually beneficial bargains and broadly share the gains from cooperation among the stakeholders, internationally and nationally. This could be in the parties' enlightened self-interest because, under conditions of openness, unresolved global challenges continue to roam the global public domain, potentially affecting all or any one country or actor group haphazardly.

Box 1 Introducing public goods and global public goods.

Standard economic theory distinguishes between two main categories of goods: private goods and public goods.

Private goods are goods that can be parceled out and made excludable so that clear property rights can be attached to them.

Public goods, by contrast, are goods that are nonexcludable, meaning that the goods' effects (benefits or costs) are there for all.

If a good is nonexcludable and nonrival in consumption so that one person's use of the good or one person's being affected by it does not diminish its availability for others, the good is said to be pure public. Examples are the light of a candle and peace and security. If a good possesses only one of these properties it is impure public. The atmosphere, for example, is nonexcludable but rival in consumption, because unrestricted pollution can change its gas composition and contribute to global warming. Patented pharmaceutical knowledge illustrates a nonrival good, whose use has, at least for a limited period of time, been made excludable. So, it, too, falls into the category of an impure public good.

The public effects of a good can be of different geographic – local, national, regional, or worldwide – reach, and they can span across one generation or several generations.

Global public goods are goods whose benefits or costs are of nearly universal reach or potentially affecting anyone anywhere. Together with regional public goods they constitute the category of transnational public goods.

Important is to emphasize that, in the present context, the term 'good' has no value connotation. It is used as a short form for the goods or products, as well as services and conditions that exist in the public domain.

Also, in most cases, publicness and privateness are not innate properties of a good but the result of a social or political choice. Therefore, it is important to distinguish between the potential and *de facto* publicness of a good. For example, land can be freely accessible to all; or it can be fenced in, be made excludable.

Globalness is a special form of publicness, and in most cases, it, too, results from a policy choice, for example, a decision to promote free trade or financial liberalization. Thus, while some global public goods are naturally global and public like the sunlight, many others are human made, including the international communication and transportation systems or the goods of communicable disease control, financial stability, or peace and security.

Source: This box draws on Kaul and Mendoza (2003). The interested reader may also wish to consult Barrett (2007) and Sandler (2004).

Hence, a desirable further function of global public policy would be the promotion of process fairness coupled with positive-sum thinking.

Function 3: Facilitating the 'Getting to the Good'

While publicness in consumption is the defining property of a public good, many goods are also public in provision. They require state or nonstate actors, who wish to alter the provision status or form of a public good, to pursue the cooperation of others. This holds especially for global public goods. In many cases, their provision path is multiactor, multisector, and, importantly, multilevel.

As Figure 1 shows, goods of this type emerge from a summation process. National policy measures and outcomes – provided in an internationally concerted manner – tend to be the main building blocks, complemented, as and if needed, by international building blocks such as an agreement defining the shared goal to pursue or an organization through which parties undertake pooled initiatives.

Consequently, effective issue management and result orientation, that is, actually producing the desired good, can be seen as a third key function of global public policy. This entails, in particular, recognizing the provision path of global public goods as a new, added policy space that may partially overlay the established policy spaces defined by geographic criteria, economic sector, and actor group. In many instances, meeting a global challenge calls for inputs from all of these other spaces – in a multiactor, multisector, multilevel approach, and, hence, for an issue manager or anchor institution to facilitate the pulling together of these diverse input threads.

Function 4: Supporting Global Sustainability

Alongside the global issue space there is another new public policy realm which a global public policy would need to take into account: the global public domain.

Various global public goods may have entered the global public domain at different times, owing to newly arising challenges or swings in economic thought and evolving global power balances. Moreover, they may have been negotiated in different venues and by different actor groups. Thus, conflicts between them may exist (Petersmann, 2011), with the result that overall global efficiency and equity suffer and opportunities for progress are constrained.

Also, as recent studies on future global trends show (e.g., WEF, 2011–2013), more and more global challenges are lingering in a state of serious underprovision. These are beginning to interact and form risk clusters of potentially catastrophic impact.

Hence, the fourth function of global public policy would be global risk and opportunity management.

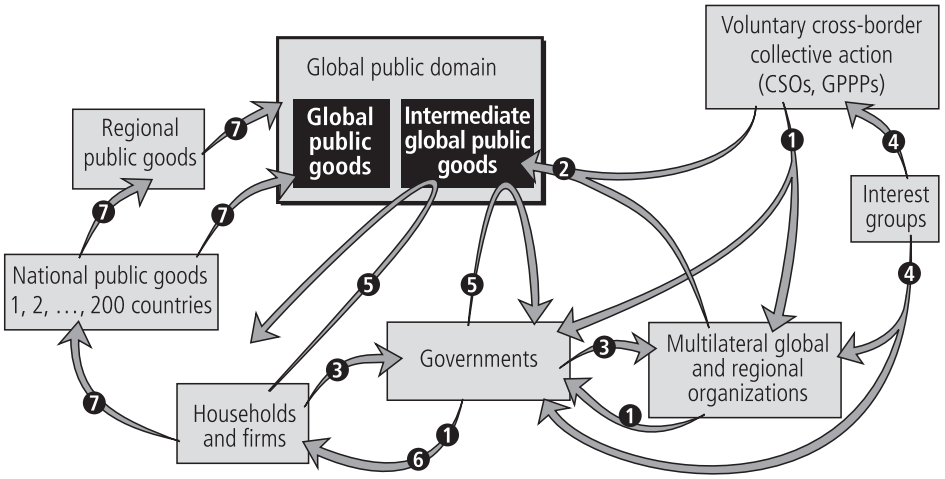

❶ **Incentives**
Encouraging actors to deliver direct and indirect inputs or to change behavior to account for social concerns

❷ **Opportunity**
Offering households and firms the possibility of consuming goods and services that generate externalities that enhance the provision of the public good

❸ **Demand for international cooperation**
Reflecting national preferences for cooperation beyond borders

❹ **Political Pressure**
Lobbying governments to fund or deliver goods

❺ **Consumption**
Consuming goods and services made available to enhance the provision of the public good

❻ **Coercion**
Compelling individuals and firms to change their behavior to account for social concerns

❼ **Externality**
Emerging as a result of individual action

Based on the assumption that the good follows a "summation" aggregation technology

Figure 1 Provision of global public goods. Source: Kaul, I., Conceição, P. (Eds.), 2006. The New Public Finance: Responding to Global Challenges. Oxford University Press, New York, p. 14.

Do we see empirical evidence that these predicted functions are actually being performed? If so, what operational forms do they take and who is performing them?

These are the questions explored in Section Global Public Policy: A Reality?

Global Public Policy: A Reality?

Even a cursory look at recent policy practices in global issue areas reveals that many qualifying initiatives exist. So far, however, as the following facts and figures suggest (and discussed in more detail in Kaul, 2013), they occur in a relatively spontaneous, ineffectively coordinated, even muddled way. They are being undertaken by myriad state and nonstate actors, acting alone or in association with others. But, before describing the existing response pattern in more detail, let us first look at some of the evidence, function by function.

Evidence of Function 1: Averting Risks of Dual – Market and State – Failure

As Figure 2 shows, during the post-1990 era, there has been a marked increase in activities aimed at monitoring and assessing the performance of countries, notably that of their governments, in an ever-larger number of issue areas.

For one, governments are keeping a closer check on each other through mutual monitoring and assessment of national-level compliance with international agreements, ranging from oversight of country-level progress toward the Millennium Development Goals (MDGs) and respect for human rights to control of international terrorism, curbing of greenhouse gas emissions, and tackling of excessive financial volatility. In addition, states are also subject to increased monitoring, review, and ranking by nonstate actors, including international traders and investors, rating agencies and civil society organizations.

In fact, as Keohane (2012) shows, moralism and legalism, as well as legalization, have been on the rise in the post-1990. Among the indications are the increased use of sanctions by the United Nations Security Council, the new norm of responsibility to protect, and the creation of the International Criminal Court and the Appellate Body of the World Trade Organization.

While pressure on state actors to cooperate is intensifying, nonstate actors, for their part, appear more forthcoming in revealing their preferences and demonstrating willingness to contribute to global public goods. Indicators of this willingness include the growing number of individuals and associations that engage in global advocacy, voluntary norm and standard-setting, corporate social responsibility initiatives, the creation of private foundations supporting global public policy purposes, and the volunteering of time, knowledge, and expertise. The demand for, as well as the production of, 'green,' 'safe,' or 'fair price' private goods also signals a desire to contribute (Edwards, 2011; Granville and Dine, 2013; Hale and Held, 2011).

Of course, many private-sector actors still fail to internalize the externalities of their actions such as excessive risk taking. Yet, as the follow-up to the 2008 financial crisis has once again shown, it is often hesitant regulatory behavior on the part of governments that allows this behavior to continue. State failure frequently precedes market failure, not least due to market pressure on states.

Thus, we see that the problem of free riding of states in the presence of global public goods is clearly being recognized and responded to – by state and nonstate actors. Both, state and nonstate actors fail in the presence of global public goods, and both intervene to nudge others into action. Yet, as policy tools like monitoring and review are merely tools, we also see that their increased use does not necessarily tell us anything about the direction in which policy is moving, for example, to which side the market/state balance is tipping. The policy content depends more on the openness of policymaking, the aspect reviewed next.

Evidence of Function 2: Promoting Global Fairness

Turning first to the process dimension of global fairness, the days of relatively closed, 'statist' international cooperation are clearly gone, as suggested by the foregoing discussion. This

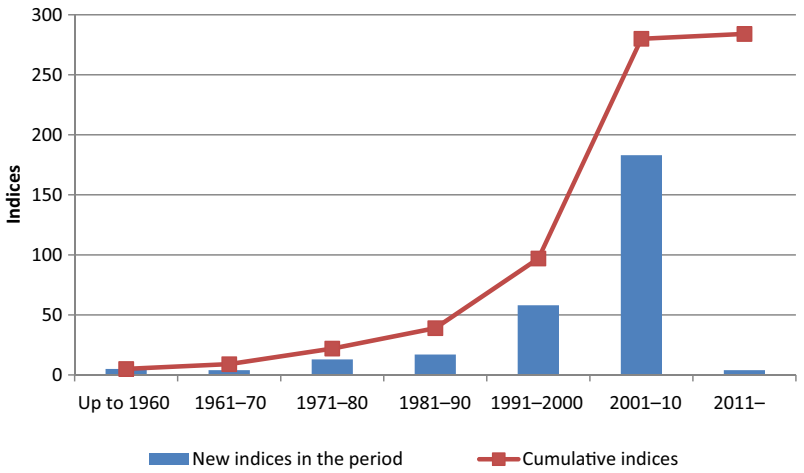

Figure 2 Increase in indices measuring and ranking the performance of states in select issue areas. Source: Bandura, R., 2011. Inventory of Indices. Available at https://sites.google.com/site/inventoryofindices/ (accessed 06.08.14.).

change did not – or not only – happen because some international agency or global leader took a deliberate decision to encourage broader participation of nonstate actors. Rather, it has been a 'bottom-up' process (Kaldor et al., 2012). More nonstate actors became internationally engaged and connected, sought avenues of consultation with multilateral organizations, and set up their own platforms like the World Economic Forum and the World Social Forum. By now, although the democratic credentials of civil society organizations themselves still pose a challenge, most multilateral organizations have institutionalized channels of consultation with civil society (Scholte, 2010).

Like the nonstate actors, the developing countries did not sit by waiting to be granted a more effective voice in multilateral organizations by the conventional powers. They, too, created their own associations, including regional and issue-specific cooperation venues, as well as their own leadership forums like the summit meetings of the BRICS (BRICS is a forum of leading emerging economies, including Brazil, Russia, India, China, and South Africa.) – and they succeeded in influencing international agendas and negotiation outcomes (see, e.g., UNDP, 2013; De Lombaerde et al., 2012).

Yet, unlike in the case of the nonstate actors, developing countries have been granted only modest increases in their decision-making power, for example, through quota reform in the International Monetary Fund and through their participation in the Leaders' G-20. Moreover, opposite trends can also be observed. For example, venue shifting on the part of the major industrial countries has given preference to smaller, informal negotiating groups, as well as bilateralism (Skidmore, 2012; WTO, 2011).

Not surprisingly, the major multilateral negotiations today are increasingly stuttering. Just think of the Doha Round of Multilateral Trade negotiations, or the global climate change conferences in Copenhagen, Cancun, and Durban. However, within both developing and industrial countries, international cooperation in the form of cross-border policy harmonization appears to be on the rise. This is notably the case when such steps generate 'immediate' national cobenefits, as they often do in the areas of climate change mitigation and energy security (see, e.g., REN21, 2013), and when a market-based, private-good approach to meeting global concerns can be taken.

Thus, what we seem to be witnessing is a multiplication of venues and events, and with it, a growing deconcentration, as well as informalization, of decision-making on global issues. This includes the wielding of more responsibility and initiative by nonstate actors and, frequently, bypassing of the conventional organizations.

In due course, and to the extent that they will not only be happening ad hoc but form part of a deliberate strategy of vertical, as well as horizontal, subsidiarity, these trends could have important positive effects in affording different actor groups enhanced policy space. For the time being, however, they seem to be of an ad hoc, experimental and unsettled nature, and they are placing constraints on the second dimension of global fairness: the pursuit of mutually beneficial cooperation bargains.

Signs of the pursuit of positive-sum thinking and win/win strategies are barely evident. The conventional powers, notably the United States, are no longer as willing as they previously were to 'foot the bill' for international cooperation, and the new, emerging powers are not yet prepared to step into the resulting void (Ikenberry, 2008; Skidmore, 2012). Moreover, this retreat from multilateralism means more challenges remain unresolved and, in turn, fewer arguments remain for policymakers to convince electorates about the merits of international cooperation. This task is especially difficult given that in many wealthy countries the notion of 'embedded liberalism' is on the wane and inequity is rising (Galbraith, 2012; Ruggie, 1997). In fact, some analysts (e.g., Bremmer, 2012, Rachman, 2010) warn that we are heading not toward a positive-sum world but away from it.

Evidence of Function 3: Facilitating the 'Getting to the Good'

The trend toward actor pluralism is perhaps even more pronounced on the operational side of international cooperation than on its decision-making side. Most striking is the proliferation of single-issue delivery and financing mechanisms since the 1990s (Conceição, 2006; Reisen, 2010). An important driving force behind this trend has been the encouragement and facilitation of a greater role for private-sector actors in international cooperation, including through new and innovative ways of tapping private finance and the creation of incentives for private business to engage in pro-poor production efforts. Another propelling force has been the realization that some summation-type provision processes – for example, polio eradication or terrorism control – include 'weak-links' that require targeted support in order for the good to emerge (Kaul and Conceição, 2006). And a third impetus for the rise in single-issue mechanisms has, obviously, come from civil society and their strengthened international engagement.

Although these varieties of global issue management are important and perhaps contribute to a strengthened result orientation, they concern only select subcomponents of global public goods. To illustrate this point, a new drought-resistant crop may be a critical input in enhanced food security, but much more than this lone input will be required to achieve the overall goal.

Awareness of the importance of overseeing the provisioning of global public goods in a more holistic and consolidated way appears to be growing slowly (as also shown by Dubash and Florini, 2011; Elson, 2012; Kickbusch et al., 2013). The UN Secretary General has, for example, appointed a growing number of eminent personalities as his Special Representatives to issues of global concern. Yet their role is mainly one of advocacy.

Similarly, states are now appointing 'issue ambassadors,' that is, diplomats not assigned to countries but to topics like HIV/AIDS, climate change, energy, multilateral trade, or human rights, and issue-specific transnational networks of government officials have become a reality (Slaughter, 2004; Hale and Held, 2011; Reinicke, 1998). Examples are the Financial Stability Board and the International Association of Insurance Supervisors. However, national institutional arrangements that allow global issues to be addressed in a comprehensive manner are still rare, both on the legislative and the executive side of policy-making.

Thus, the provisioning of global public goods today is still a fractured process, and therefore, it should not surprise that despite countless practical–political initiatives at all levels many of these good continue to linger in a state of under-provision. They lack an anchoring agency or agent to nudge all concerned actors to go the extra mile: to jump over their narrowly conceived private or national interest hurdles and do what is required to actually get to the good. And a reason for this might be that the transnationalness of the provision path of global public goods has so far found only faint recognition.

In fact, in some cases not only is more focused global issue management absent but also attention to the global public domain as a whole, as discussed next.

Evidence of Function 4: Supporting Global Sustainability

When the 2008 financial crisis erupted, the G-20 was established as the world's premier forum for international economic cooperation. However, member states' resolve to cooperate in tackling the crisis weakened over time. Yet, when assessing the G-20's performance, one must bear in mind that what "was done in 2008–2009 was by nature exceptional and the following steps were bound to be of lower intensity" (Pisani-Ferry, 2012: p. 33).

Also, Kharas and Lombardi (2012) remind us that the G-20 is not an operational body itself. Rather it tries to help build global consensus and to provide the required political momentum to nudge others into action. In the authors' view, "the G-20 appears to be the 'best available option' for global economic governance" (p. 16).

While this might be so, the fact is that many critical global issues have not yet received the Group's full attention, despite multiplying warning calls from reports such as the one on why we must avoid a 4° global temperature rise (PIK and Climate Analytics, 2012). Similarly, the broader issue of a new, more sustainable, and inclusive growth and development paradigm, while finding considerable popular and business interest, is still lacking decisive support from world leaders, as shown by the 2012 Rio +20 conference (Khor, 2012). And even more limited issues remain unresolved, such as the continuing inconsistencies between the global intellectual property rights regime and global health (Kickbusch et al., 2013).

So again, while many initiatives are materializing, a systematic approach to managing the overall structure and functioning of the global public domain is not yet in sight. Global leadership in support of the provisioning of global public goods occurs haphazardly, mainly in response to major crises. This corroborates what we already saw in the previous point and what also earlier studies (e.g., Shah, 2012; Stone, 2008) found: the lack of a clear notion of global space.

In Sum: Relevant Initiatives Exist, but Not yet a Policy

Measures fitting the functional profile of global public policy are clearly discernible. However, they do not yet seem to be coalescing to form a recognizable, new policy strand.

While many state and nonstate actors are involved, in large numbers and at all policy-making levels, it seems that rights and responsibilities are unclear and roles blurred. No one appears fully in charge. In fact, governments can often be counted among the more reluctant actors, hesitant to restrain their own free riding attempts, as well as those of nonstate actors in the presence of global public good-type challenges.

Certainly, we live in a multiactor world, and important resources now reside in the private sector. Public–private partnering is, in many instances, essential to getting things done. Nevertheless, governments seem to be the best positioned actor group to foster the delivery of what is perhaps the most essential global public good: an institutional architecture that allows the currently messy process to take on a more organized, focused, and, thereby, effective shape. But, if we want governments to do that, we need to understand what currently limits their willingness to cooperate and how they could be persuaded to become more forthcoming facilitators of global public policy.

Section The Missing Link: A Sovereignty-Compatible Rationale for Global Public Policy will point to some of the existing impediments and propose possible next steps that might be required in order for global public policy to advance.

The Missing Link: A Sovereignty-Compatible Rationale for Global Public Policy

Reluctance to engage more actively in international cooperation exists on the part of the weaker states, as well as on the part of the major powers, and each side recriminates (Subramanian and Mattoo, 2012). One driver of this reluctance to cooperate is the noticeable lack of global fairness. For example, the perception of unfair burden sharing has been found to weaken public support for international cooperation in the more powerful, richer countries (Skidmore, 2012).

The way in which many policymakers, as well as scholars, journalists, and hence, the public at large, think and speak about international cooperation provides a clue to another – and perhaps, even more fundamental – possible impediment. International cooperation is often perceived as running counter to the core principle on which the present world order rests: noninterference by external forces into states' internal affairs, or, put differently, into their national policy-making sovereignty.

At first sight, global public policy appears to imply just that. It requires the recognition of common interests and a bridging of the foreign/domestic divide. This suggests a blending of national and external policy concerns. To achieve that, the state would need to act – at least in fields of policy interdependence among countries – as an intermediary between national and global concerns. In many instances, states are already playing this role (see, Kaul, 2006). Yet, they are often playing it reluctantly, because many perceive this role as entailing a loss or transfer of national policy-making sovereignty to international organizations and other globally engaged actors. As a result, policymakers frequently shy away from international cooperation even where it would certainly constitute the better approach to meeting national interests. Because, how sovereign are policymakers and countries when, due to a lack of international cooperation in key areas, they come under attack from forces such as financial crises, catastrophic storms or floods, terrorist actions, or new and resurgent communicable diseases that hold no respect for national borders?

In these cases, reluctance to engage in multilateral cooperation leads to a 'sovereignty paradox': The harder governments hold on to a conventional, absolute notion of sovereignty, the more they lose their policy-making capacity.

Thus, it can be conjectured that to persuade policymakers and their constituencies to break out of this paradoxical situation, it is important to define a sovereignty-compatible rationale for global public policy. This is might entail taking two main steps.

First, it must be established that, especially in policy areas involving interdependence challenges, it is rational and feasible for states to engage in effective – and to that end, fair – international cooperation. This could, for example, be done by complementing studies on the economics and politics of particular global issues, such as the Stern Review (Stern, 2007), with country-specific analyses.

Second, based on such evidence, it would be useful to forge global consensus on a notion of mutually respectful sovereignty. This notion of sovereignty would combine the pursuit of national interests with respect for the sovereignty of other states and, therefore, with respect for planetary boundaries and other global systemic balances (De Brabandere, 2013; Ţuţuianu, 2013). Such consensus would reassure countries that other nations will come with similar – positive sum-oriented – strategies to the negotiating table.

If framed in this way, sovereignty would 'allow' policymakers to engage in global public policy without being concerned about betraying – or being perceived as betraying – the national interest. It would then be clear that sovereignty as such is not in question. To the contrary, the strategies of exercising sovereignty would be adjusted to current policymaking realities, notably the challenge of interdependence, to safeguard national sovereignty.

In other words, a sovereignty-compatible rationale for global public policy would state that mutually beneficial international cooperation is the preferred strategy for responding to interdependence challenges because (1) it fosters effectiveness and maintains national policymaking capacity and (2) it shows respect for the national concerns of others, reinforcing, whenever it is applied, the norm of mutually respectful sovereignty. Importantly, this norm, if widely adopted, would protect countries against negative spill-ins from abroad.

It is perhaps not accidental that policymakers and scholars have begun noticing the absence of a systematic conceptualization of global public policy. The global challenges that figure prominently today on both the international and national agendas seem to be of a different character than the global challenges we confronted in the pre-2000 phase of globalization. The focus then was on fostering economic openness and market integration. Many governments followed the calls for greater openness, in part, because of the promise of access to new, expanded economic opportunities, and, in part, because they were nudged along by the then superpowers. Now, the focus is on constraining openness, specifically on states' reigning in of cross-border externalities in issue areas of high global risk. Moreover, under today's multipolar power relations, the adjustment burden falls on all states.

Therefore, to avoid heading toward a zero-sum world, we need to establish a sovereignty-compatible rationale for global public policy and to develop the conceptualization and operationalization of this policy strand, offering a persuasive narrative on why and how engaging in effective multilateralism can, under certain conditions, be potentially desirable and feasible.

See also: Alternative Schools of Economic Thought; Climate Change; Global Governance; Governance; Health Politics and Policy; Organizational Fields; Policy Analysis: Evidence Based Policy-Making; Policy Networks; Policy Process: Nonprofit Institutions in the United States; Public Choice Economics; Public Policy: Ethics; Public Policy; Quality of Government; Spatial Thinking in the Social Sciences.

Bibliography

Bandura, R., 2011. Inventory of Indices. Available at https://sites.google.com/site/inventoryofindices/ (accessed 06.08.14.).

Barrett, S., 2007. Why Cooperate? The Incentive to Supply Global Public Goods. Oxford University Press, New York.

Bremmer, I., 2012. Every Nation for Itself; Winners and Losers in a G-Zero World. Penguin, London.

Conceição, P., 2006. Accommodating new actors and new purposes in international cooperation: the growing diversification of financing mechanisms. In: Kaul, I., Conceição, P. (Eds.), The New Public Finance: Responding to Global Challenges, pp. 269–280.

De Brabandere, E., 2013. The impact of 'Supranationalism' on State Sovereignty from the Perspective of the Legitimacy of International Organizations. File name: SSRNid2175478 (accessed 06.08.14.). Available at http://papers.ssrn.com/.

De Lombaerde, P., Baert, F., Felício, T. (Eds.), 2012. The United Nations and the Regions; Third World Report on Regional Integration. Springer, New York.

Dubash, N.K., Florini, A., 2011. Mapping global energy governance. Global Policy 2, 6–18.

Edwards, M., 2011. The Oxford Handbook of Civil Society. Oxford University Press, New York.

Elson, A., 2012. International financial architecture; a report card. World Economics 13 (2), 155–170.

Galbraith, J.K., 2012. Inequality and Instability: A Case Study of the World Economy Just before the Great Crisis. Oxford University Press, New York.

Granville, B., Dine, J. (Eds.), 2013. The Processes and Practices of Fair Trade: Trust, Ethics and Governance. Routledge, New York.

Hale, T., Held, D. (Eds.), 2011. The Handbook of Transnational Governance: Institutions and Innovations. Polity Press, Cambridge.

Howse, R., Teitel, R., 2010. Beyond compliance: rethinking why international law really matters. Global Policy 1 (2), 127–136.

Ikenberry, G.J., 2008. America and the reform of global institutions. In: Alexandroff, A.S. (Ed.), Can the World Be Governed? Possibilities for Effective Multilateralism. Wilfrid Laurier University Press, Waterloo, pp. 110–138.

Kaldor, M., Morre, H.L., Selchow, S. (Eds.), 2012. Global Civil Society 2012: Ten Years of Critical Reflection. Palgrave Macmillan, Basingstoke.

Kaul, I., 2013. Meeting global challenges; assessing governance readiness. In: Hertie School of Governance, the Governance Report 2013. Oxford University Press, Oxford, pp. 33–58.

Kaul, I., 2006. Blending external and domestic policy demands; the rise of the intermediary state. In: Kaul, I., Conceição, P. (Eds.), The New Public Finance: Responding to Global Challenges, pp. 73–108.

Kaul, I., Conceição, P., Le Goulven, K., Mendoza, R.U. (Eds.), 2003. Providing Global Public Goods; Managing Globalization. Oxford University Press, New York.

Kaul, I., Conceição, P. (Eds.), 2006. The New Public Finance: Responding to Global Challenges. Oxford University Press, New York.

Kaul, I., Mendoza, R.U., 2003. Advancing the concept of global public goods. In: Kaul, I., Conceição, P., Le Goulven, K., Mendoza, R.U. (Eds.), Providing Global Public Goods; Managing Globalization. Oxford University Press, New York.

Keohane, R.O., 2012. Twenty years of institutional liberalism. International Relations 26 (2), 125–138.

Kharas, H., Lombardi, D., 2012. The Group of Twenty: Origins, Prospects and Challenges for Global Governance. Brookings, Washington, DC.
Khor, M., 9 July 2012. RIO+20 Was Not All In Vain. In: IDN-InDepth News Analysis.
Kickbusch, I., Lister, G., Told, M., Drager, N. (Eds.), 2013. Global Health Diplomacy: Concepts, Issues, Actors, Instruments, Fora and Cases. Springer, New York.
Petersmann, E.-U., 2011. International economic law, 'public reason', and multilevel governance of interdependent public goods. Journal of International Economic Law 14 (1), 23–76.
Pisani-Ferry, J., 2012. Macro-economic coordination: what has the G-20 achieved?. In: Think Tank 20: New Challenges for the Global Economy, New Uncertainties for the G-20, pp. 28–34.
Potsdam Institute of Climate Impact Research, Climate Analytics, 2012. Turn Down the Heat: Why a 4°C Warmer World Must Be Avoided. World Bank, Washington, DC.
Rachman, G., 2010. Zero-Sum World; Politics, Power and Prosperity after the Crash. Atlantic Books, London.
Reinicke, W., 1998. Global Public Policy; Governing without Government? Brookings Institution Press, Washington, DC.
Reisen, H., 2010. The multilateral donor non-system: towards accountability and efficient role assignment. Economics 4, 2010–2015.
REN21, 2013. Renewables 2011 Global Status Report. REN21 Secretariat, Paris.
Ruggie, J.G., 1997. Globalization and the Embedded Liberalism Compromise: The End of an Era? Max-Planck-Institute for the Study of Societies. Working Paper 97/1.
Sandler, T., 2004. Global Collective Action. Cambridge University Press, Cambridge.
Scholte, J.A., 2010. Civil society in multi-level governance. In: Enderlein, H., Wälti, S., Zürn, M. (Eds.), Handbook on Multi-level Governance. Edward Elagr, Cheltenham, UK, pp. 383–396.
Severino, J.-M., Ray, O., 2010. The End of ODA (II): The Birth of Hypercollective Action. Center for Global Development, Washington, DC.
Shah, N., 2012. The territorial trap of the territorial: global transformation and the problem of the state's two territories. International Political Sociology 6, 57–76.
Skidmore, D., 2012. The Obama presidency and US foreign policy: where's the multilateralism? International Studies Perspectives 13, 43–64.
Slaughter, A.M., 2004. A New World Order. Princeton University Press, Princeton.
Stern, N., 2007. The Economics of Climate Change: The Stern Review. Cambridge University Press, Cambridge.
Stone, D., 2008. Global public policy, transnational policy communities, and their networks. The Policy Studies Journal 36 (1), 19–38.
Subramanian, A., Mattoo, A., 2012. Brookings Institution Press, Washington, DC.
Ţuţuianu, S., 2013. Towards Global Justice: Sovereignty in an Interdependent World. Asser, The Hague.
UNDP (United Nations Development Programme), 2013. Human Development Report 2013 – the Rise of the South – Human Progress in a Diverse World. UNDP, New York.
WEF (World Economic Forum), 2011–2013. Global Risks. World Economic Forum, Geneva.
WTO (World Trade Organization), 2011. World Trade Report 2011: The WTO and Preferential Trade Agreements: From Co-existence to Coherence. WTO, Geneva.

Relevant Websites

www.cfr.org/thinktank/iigg/ – International Institutions and Global Governance Program.
www.g20.utoronto.ca/ – University of Toronto G20 Information Centre.
www.globalpolicy.org/ – Global Policy Forum.
www.globalpolicyjournal.com/ – Global Policy Journal.

Global South

Don Kalb, Central European University, Budapest, Hungary and Utrecht University, Utrecht, The Netherlands
Luisa Steur, University of Copenhagen, Copenhagen, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The global south is a complex and dynamic concept that straddles multiple social science and humanist disciplines. Emerging around 2000, it reflects the agenda of two ascending forces in those years: the antiglobalist (alterglobalist/global justice) movement and the World Social Forum, on the one hand, and an alliance of Southern states within the World Trade Organization on the other. Generally seen as an inheritor of the emancipatory thought behind the notion of the 'third world,' in the social sciences the idea of the 'global south' is also entangled with more classical academic themes such as modernity, development, and (post)colonialism.

The concept of the global south is a very recent theoretical assemblage. One of the first formal appearances of the word was in the title of a new advocacy journal, *Focus on the Global South* (http://focusweb.org/) established in 1995 by the sociologist Walden Bello in Bangkok. A quick search through academic journals shows that the term was rarely used before the year 2000, however. From then on, it was exceptionally swiftly established in politics and academia. It was adopted as a key programmatic term by emergent global justice movements and nongovernmental organizations (NGOs) that crystallized into the World Social Forum (WSF) in Porto Alegre (Brazil) in 2001, a process still ongoing through regular meetings and publications. It was also featured as a banner for a new alliance among key 'southern' states such as Brazil, Argentina, and India during trade negotiations at the 2003 World Trade Organization (WTO) meetings in Cancun, Mexico, and was the topic of a major United Nations (UN) conference in the same year. Around that time, it began to be featured in leading academic publications as diverse as *Third World Quarterly* and *Anthropological Theory*. And in 2007, its fast rise on the intellectual-political agenda culminated in it becoming the actual title of a new academic journal, *The Global South Journal* (Indiana University Press, see http://www.jstor.org/action/showPublication?journalCode=globalsouth). A Google search shows the first appearance of the term in book titles in 1999, with a continuous upward slope afterward.

Notwithstanding its obvious geographical meaning (see further below), the term has in fact a very complex and contested referent. This complexity partly explains, and is explained by, its use in very different fields. It functions within and between those fields as an intellectual magnet bringing various separate long-standing themes together in new ways. It is used among others in international relations, political science, development studies, postcolonial studies, cultural studies, anthropology, and economics. On the one hand, it serves as a synonym for the older term 'the Third World.' This is how it is mostly used in international relations, development studies, and economics, and indeed how it was largely used at the WTO in Cancun and in the UN. Global south and global north are then the present day equivalents of the First World and the Third World, the core and the (semi)periphery, the developed and the underdeveloped world. But this older state-driven and rather 'upscale' referent is crisscrossed and energized by a very different and more informal play of 'downscale' populist democratic signifiers that evoke themes of diversity, alterity, indigeneity, injustice, dispossessions, and neocoloniality. These themes are articulated by antineoliberal popular mobilizations and subaltern resistance movements and they are aimed against the contemporary mechanisms of global capitalist disempowerment. In between these populist and state-driven fields, there is an interlocking layer of educated NGO actors and practices that brings the two levels in conversation. These upscale, midscale, and downscale meanings all have their own academic representatives. Upscale and downscale fields also seem to get blurred in academic discussions, making them slightly unpredictable and exciting.

Interwoven through these fields, there is also a certain homology between the idea of 'the global south' and the older scholarly theme of postcoloniality. Here, the idea of the global south has been used to bring the histories and complexions of various ex-European postcolonies together, and to associate them with the injustices and inequalities signaled by the new cycle of subaltern political energies in the postcolonies. In the process, a new political life was given to a field that had become almost equated over the years with high-brow postmodernism, in particular in its Indian variants. But this amalgamation of the idea of the 'global south' and postcoloniality seems only to have arisen after the first two forces, the movement and the state level, had done their combined intellectual work around the notion of 'the global south.' Postcoloniality/global south has swiftly become an influential mode of writing in academia and is perhaps most obviously articulated in the work of the Portuguese public intellectual and sociologist, Boaventura de Sousa Santos. In anthropology it features, for example, in the work of Arturo Escobar and the recent writings of Jean and John Comaroff. Other important authors are the Chinese American historian Arif Dirlik, the Indian publicists Arundhati Roy and Vandana Shiva, among others.

This multiple field of meanings betrays the origin of the concept of the global south in the partial coming together of a new cycle of left wing popular mobilizations in both the south and the north under the umbrella of a broadly defined 'alterglobalist' movement ('antiglobalization movement' or 'global justice movement') as it emerged from the Seattle

anti-WTO demonstrations in 1999, on the one hand, and the politics of trade, development, and globalization at the top echelons of global governance in the early 2000s on the other. It is therefore fair to say that the notion of the 'global south' was, and still is, a 'movement' concept. It is a movement concept stemming from and feeding into movements of both high and low politics, state driven as well as populist, and is generally geared to a renewal of the Left, with intriguing linkages upward and downward, and with different sections and agendas of academia, public intellectuals, and the NGO world involved in both tiers.

To understand its rise, use, and possible meanings, one must place it in its proper historical context, and indeed in relation to the historical process to which it responds and which it is meant to challenge, that is Northern led neoliberal globalization.

The 'Third World' and Its Neoliberal Undoing

Unsurprisingly, various authors have argued that the idea of the 'global south' is the successor to the older notion of 'the third world.' More precisely, it is argued, it is the inheritor of the emancipatory impulse behind the Bandung conference and the 'Nonaligned Movement' (NAM) operating as a bloc at the United Nations. The NAM came into being at the Bandung conference in Indonesia in 1955 and was led by Sukarno (Indonesia), Zoe En Lai (China), Nehru (India), and Nasser (Egypt), shortly to be joined by Tito (Yugoslavia). It started with around 90 member states, all just recent or soon to become independent former Western, Ottoman, or Japanese colonies. The NAM was designed to keep the domestic choices of the newly independent states free from interference or intervention by the First or the Second World, which continued to monopolize the Security Council. And it tried to keep the United Nations to its promise to help 'develop' the new nations rather than being a forum for great power confrontations between East and West in the Cold War. The NAM thus was a major engine behind 'development' as a world goal.

The activism of the NAM had been preceded in Latin America by the formation of CEPAL (1948, Chile; the Economic Commission of Latin America) and was followed by the creation of the United Nations Conference on Trade and Development, the UN organization for industrial development led by the left wing Argentine economist Raul Prebisch, former director of CEPAL. The NAM was immediately confronted with the less than gentle realities of 'development' in a system that remained under control by Western capitalist states, which were demonstrably ready to use violence if their interests were threatened, for example, by nationalization: the coup against Mossadeq in Iran (1952), the Suez crisis (1956), the coup against Sukarno, the very organizer of Bandung, plus the subsequent Indonesian mass massacres (1964/5). Against this, the NAM succeeded to push through at the UN General Assembly its radical proposal for a New International Economic Order (NIEO) (1968). The NIEO entailed nothing less than a demand for far-going control by Third World states over international capital and trade flows, and a serious redistribution of technical expertise to the new nations. It was an outright claim by the Third World to a good chunk of actual power over the global system and a radicalization of the development framework. This was then one of the most memorable political markers of the rising 'New Left' in the world of the 1960s and 1970s. Intellectually it was supported by 'new' structuralist theories in development economics, 'dependency theory,' world systems theory, various new Marxisms, and a popular 'third worldism' actively endorsed by the New Left in the West too. These intellectual streams were united in their emphasis on the ways in which global linkages between cores and peripheries had historically been working against the interests of the peripheries and they led to a claim for reversing the skewed balances. After years of foot-dragging by the West, the compromise at the UN of all this was the infamous 'Brandt Report' (1981), supervised by the former socialist mayor of West Berlin and dynamic left wing Chancellor of West Germany in the early 1970s. The Brandt report endorsed the moral and political claim for a global redistribution of power. The report, however, came 10 years too late to effectively influence the way the world was evolving, and thus remained a dead letter. The 'global south' now in fact grasps back to it as the historical embodiment of broken promises.

The transformation of the capitalist world system through diplomacy between states at the UN General Assembly, then, did not happen. Instead of a 'new order,' the new nations experienced crisis and indebtedness, with the International Monetary Fund (IMF) leading 'structural adjustment.' This implied privatization of publicly owned assets, and the spread of a 'new international division of labor' through the proliferation of 'export processing zones.' The IMF had been transformed around 1980 from a Keynesian tool to manage trade imbalances into a debt collection agency for Northern capital in the indebted South and East. It began rolling over sovereign debt to private banks with public loans, in exchange for which the recipient sovereigns had to accept a substantial loss of sovereignty. It imposed debt interest payments equaling 10% of the gross domestic product (GDP) or more on its clients, as well as 'Structural Adjustment Programs,' which forced states to privatize public assets, liberalize trade, reduce subsidies and entitlements, and stabilize the currency. It led, predictably, to a massive export of capital from the periphery to the center. Africa and Latin America would thus experience lost decades in the 1980s and 1990s, with poverty multiplying, inequality going completely out of hand, a rural exodus translating into mega-urbanization without industrialization, and therefore into ghettoization, increased precariousness of life, and finally the veritable fracturing of states, in particular, in Africa. This, in the midst of an IMF-enforced national policy theater that was tightly focused on the myths, rituals, and protocols of 'globalization' and 'economic growth' (GDP) facilitated by global investment and export-driven earnings in the private economy. This did create new business classes, farmers, and entrepreneurs in the south, but earnings would consistently fail to trickle down, let alone structurally improve common life in the postcolony (until the 'rise of China' after 2000). Instead, then, of a NIEO on behalf of universal social emancipation, the 'Third World' received the 'Washington Consensus' and a neoliberal imperial hold over national politics. The goal of state politics transformed from emancipation for all to the administration of private accumulation. In the process, the NAM lost its relevance and coherence, and the 'Third World,'

once united by a desire for emancipation, liberation, and autonomy, broke down into an incoherent collection of mere 'emerging markets.'

While the Third World was collapsing, the 'Second World' of socialist people's republics cracked under very similar stresses of stagnation, indebtedness, and the swift evaporation of its rulers' popular legitimacy. It subsequently underwent not fundamentally different processes of privatization, unequalization, and immiseration, advocated under the rubric of 'democratic transition.' Democracy was also the promise with which the West would sell its undoing of 'corrupt' statist 'Third World' elites to their constituents. Following this came in the 1990s 'transparency,' 'civil society,' 'human rights,' and 'good governance.' By the late 1990s, it was realized that while these were all perfectly noble 'goods' in themselves, in this wider historical context of worldwide social polarization and declining popular sovereignty they did not by themselves bring much improvement to a quite generalized dismal situation. The emergent alterglobalist movement in the late 1990s realized that democracy and human rights, in short 'the liberal goods' produced by what Huntington had called the 'third wave of democratization,' had served more as smoke screens than as tools for mass empowerment. One can point to the Seattle demonstrations of 1999 against the WTO meetings as the moment at which globalization as the global resurrection of the royal right of property over everything else was unmasked. What was demanded was a different globalization, a globalization 'from below' rather than one for capital and Northern/Western ruling classes. This, then, was the moment at which those twin notions were publicly discovered that, together, would help shape a new global left and a new sense of universal emancipation versus mere administration: 'globalization unmasked as global neoliberalism,' and 'the global south' as its prime victim and its foremost challenger.

Antinomies of the 'Global South'

The anti-WTO protests in Seattle 1999 had been supported by an unprecedentedly wide array of organizations from various parts of the world. The groups had been crystallizing during the neoliberal 1990s, often building on older initiatives. They included Brazilian and Indian peasant organizations, urban black and left wing groups from South Africa, South African and Korean labor organizations, *afficionados* of the Zapatistas and the indigenous groups from Chiapas, whose rebellion in 1994 against the North American Free Trade Agreement had prefigured the emerging alterglobalist movement. But, happening in Seattle, they featured above all American, Canadian, and European labor union delegations protesting the offshoring of jobs, the decline of the Northern welfare state, and the dwindling of workers' influence. Newly emerging Northern left wing youth groups, some of them inspired by the Zapatistas, played a dynamic role too, as did Northern peasant and farmer groups. If the Zapatista uprising in Chiapas had been a forerunner for the Southern mobilization, the French labor and student mobilizations of 1994/95 had laid some of the complementary organizational and intellectual groundwork in the north. Northern and Southern protest had not grown independently of each other. The 'left hand' of the state in the north, including the associated Northern development organizations, had been key sponsors of peasant organizations in the South, while Southern Left authors had contributed regularly to *Le Monde Diplomatic*, the quintessential Left intellectual forum in 1990s Europe, in which Pierre Bourdieu and other Left intellectuals had become active.

In Seattle, all this came together for all to see in the form of a '*new* new Left internationalism.' This new internationalism then came fully into its own at the first WSF in Porto Alegre, 2001. This new 'Left internationalism' did not deny its historical association with the first two 'Northern' Internationals – the First International of 'the working men's association' under Marx, and the Second International of late nineteenth-century European social democracy, although they played them down somewhat – but it rejected the heritage of the Third International, controlled by Moscow. The rejection of the latter, however, remained somehow more significant than the endorsement of the former, which was under suspicion of having deep connections with the latter. The WSF rejected equality through centrally administered state-bureaucratic homogeneity, and declared diversity and the participatory democracy of human communities as its ultimate objectives. In a sense, therefore, one could say that it prioritized the heritage of the Mexican Revolution of the 1910s over the French and Russian revolutions, this time peaceful and worldwide, a revolution of the subalterns not aimed at taking power, but at rejecting all concentrated power. This was a revolution on behalf of indigeneity, ecology, communities, sodalities, villages and cities, and women, against the homogenizing militarized Leviathan of Northern capitalism embodied in WTO, IMF, World Bank, transnational capital and corporations, and the neoliberal policies of the United States and Europe – in short, against the Washington-Wall Street complex and its satellites.

Seattle and the WSF, thus, were the outcome of a longer run process of mobilization, ideological articulation, and alliance formation 'from below.' However, they did not come about without crucial inputs by NGOs, academics, political parties, and the left hand of Northern and Southern government organizations. Tellingly, many of the ideas of the WSF had already been voiced in the worldwide protest movements of the 1970s and had been institutionalized in their consequent NGO-ization in the 1980s and 1990s. Here clearly was a worldwide protest cycle coming to full fruition in a moment of global coordination and radicalization. The notion of the global south, in this context, expressed the will to create a set of intra-south alliances among movements and new knowledge positions. But it also served as a universal metaphor for the excluded and oppressed peoples of global neoliberalism in general, including those in the north, which could become aligned like in Seattle and form a global force, not necessarily only in the Southern hemisphere. Here was a tension in its references that would not anymore go away.

At exactly the same time, a not unrelated shift of governmental power to a more radical Left was happening in Latin America, with Brazil and to a lesser extent Argentina and Venezuela as the most important players. India also alternated back to the Left. As a result, the notion of the global south was now taken into the UN and in particular into the calamitous WTO meetings in Cancun 2002. There, it served as a tool for unifying the Southern positions on trade versus the north. 'The global

south,' generally pro-free trade in the 1990s, was now demanding protection against free trade, against the imposition of intellectual copyrights and the wielding of corporate and financial power. And it did so explicitly on behalf of global social justice and the popular sovereignty of its indigenous peoples, peasants, and workers, indeed of its whole cultural patrimony and ecology. The claims of the movements and the claims of states had become closely aligned. In the process, unsurprisingly, the notion of the global south became harder, more realist, more law oriented, less metaphorical and fluid, now often denoting a hard geopolitical and geocultural divide between identifiable world regions. The potential northern antineoliberal globalist connection had faded.

Here are some of the antinomies inscribed into the notion of the global south that have resulted from this shift and tension. Most social science authors who use it do so in a low key and are keenly aware that the geographical referent is problematic. For one, it is blind to the consequences of the collapse of the Soviet Union: Central Asian nations such as Tajikistan and Kyrgyzstan are among the poorest peasant countries on earth. Postsocialist countries in the Balkans such as Bulgaria and Romania, now even members of the European Union, have on some counts lower average incomes than the averages for Latin America. The Baltic countries as well as large parts of Ukraine could well be located in the Southern Cone for their income and dependency situation. Also, Pakistan, India, Vietnam, and Haiti are in the northern hemisphere, Australia and New Zealand in the southern. The whole Arabic and Islamic world remains unidentified. Most importantly, China does not figure at all in this mapping. Sao Paulo, Mexico City, and Mumbai have among their citizens some of the very wealthiest and most powerful people on earth. The inequality indices of the Latin American global south are by now roughly matched by those in Russia and China. And while it should be admitted that their social problems and social structures are not simply comparable, the same can be said about the comparison between Latin America and Africa, or Argentina and Bolivia for that matter. The 'global south' suggests a factual unity where that unity is only a claim, and it excludes those who could conceivably be included.

Geopolitically, the concept remains beholden to the conjuncture of the late 1990s, early 2000s. It entirely failed to anticipate some of the most crucial macrosocial developments of the 2000s, such as the eye-popping rise of China – which in 1999 was still confidently registered somewhere close to 'the global south' but now seems to have left the club; the crisis of capitalism in the United States and Europe; the next wave of mass contentious politics around the 'Arab Spring,' and the citizens' mobilizations against inequality and austerity in the global north. It has not anticipated the decline of Northern hegemony and the rise of East Asia. Its close focus on peasants, indigeneity, and complex natural and subaltern ecologies was indeed a new internationalist vision but it turned a blind eye toward ongoing processes of worldwide industrialization and urbanization, in fact the globalization and multiplication of the proletariat, flanked by Marxian 'reserve armies' and 'surplus populations.' These were the sorts of world processes for which the intellectual heritage of the Mexican revolution as compared to the French and the Russian revolutions could not deliver much of a compass. But they were the transformations responsible for the unprecedented rise of East Asia, the crisis of the Northern Atlantic countries since 2007, and the new waves of mass political contention. Significantly, while Brazil was the key state player in the emergence of 'the global south' around 2000, the elite of Lula's and Roussef's 'Workers' Party' has now joined China and Russia in the altogether different league of 'the BRICS' (Brazil, Russia, India, China, and South Africa), as has India, perfectly reflecting the nature of global social change in the 2000s. It also reflects the interests of Brazil's own mighty engineering sector plus agricultural export machine. But the BRICS alliance is in radical defiance of the 'global south' agenda, with the latter's concern for democracy, equality in difference, and the protection of smaller communities and knowledges overruled by straightforward capitalist geopolitics combined with brazen disregard of even Western liberal forms of democracy among its most formidable members. Also significant is that the key social policy act of the Lula years in Brazil has been the 'Bolsa Familia,' a poverty eradication program of universal social rights that, again, derives in an intellectual sense probably more from the heritage of the Northern socialist internationalisms than from the political innovations of the WSF.

Similar antinomies around the harder and 'realist' 'global south' notion have emerged in the key intellectual works of the alterglobalist movement. In David Harvey's influential work on 'the new imperialism' (2005) and 'neoliberalism' (2007), there is a continuous reminder of the mutual dependencies, intellectually and politically, of Northern and Southern anticapitalist mobilization and critique. Significantly, he emphasizes the links between the earlier 'Marxian' internationalisms, with their key symbol of the worker, and the new 'subaltern internationalism' of the WSF with its iconic embrace of indigenous people and peasants. Instead of ontological oppositions, he sees learning, social change, new times, and new alliances. He is, therefore, much better at anticipating China and the crisis of 'the West' than, for example, Boaventura de Sousa Santos. Intellectuals such as Hardt and Negri are decidedly less sympathetic to 'the worker' and the history of the European Left than Harvey is, but do explicitly include Northern subalterns and modern migrants in their alternative for 'the proletariat,' the 'multitude,' which is of course a pre-Marxian European notion. These works all indicate the tensions between the earlier movement notion of the global south and the later statist one, including the differential alliances, intellectual heritages, and social icons that are involved.

In the more intellectual circles affiliated with the movements of the global south, many embraced a 'critique of modernity' and a rejection of 'development' while endorsing ideas of 'postdevelopment,' 'multiple modernities,' 'altermodernities,' or 'postmodernities.' Here, postcolonial theory and 'global south' theorizing overlapped. In their radical versions, most of them Latin American and Indian, these ideas went together with a call for 'different epistemologies' and a recognition of 'different ontologies.' De Sousa Santos and Arturo Escobar are good examples. These ideas often hinge, again, on a radical separation of the north and the south, or in the Indian case the west and the east. Significantly, they also tend to picture the intellectual patrimony of the north as relatively homogeneous with few internal contradictions or contestations. Santos is explicit about this. Northern epistemologies and ontologies boil down to the rational actor and

are best embodied in the dominant science of neoliberalism: neoclassical economics. An Indian version of a similar argument in a celebrated book by Chakrabarty makes identical homogenizing claims in relation to European epistemologies and ontologies, here aiming against Marx rather than neoclassical economics. Paradoxically, Chakrabarty then embraces Heidegger contra Marx, as if the two German philosophers are extractions from a different soil. Santos' framework, too, rests in important respects on another European, Antonio Gramsci, but the Italian Marxist is never mentioned. The 'north,' the 'west,' or 'Europe,' in these arguments, stand for a narrow and arbitrary cut of politically dominant Anglo-Saxon outlooks presented as absolute cultural others.

Both Escobar and Santos emphasize the need to intervene on behalf of local and indigenous knowledges, ecologies, and indeed whole 'ontologies,' terms that are closely equated within this vision. Such local knowledges, ecologies, and ontologies are assumed to be driven by the desire to sustain and conserve nature and humanity, rather than to dominate and exploit. Escobar's ethnography of the Colombian South Pacific coast is a good example. Such authors argue that this could also inspire 'postdevelopment': a development process attuned to local needs and away from development as large-scale capital accumulation. As with their idea of 'the north,' such visions have a habit of homogenizing southern local cultures, skipping over internal social differentiations, erasing cultural and political contestation, and seeing internal change and discord as coming basically from the outside. The outside, similarly, becomes just that, an absolute and singular outside. The Northern NGOs and development monies that helped to install Escobar's local NGOs are apparently not part of the north. The Colombian constitution of 1991, which emphasized decentralization and the rights of indigenous populations, is not readily seen as part of a much larger wave of Northern promoted liberal constitution-making in the early 1990s but rather as an original product of the Colombian soil.

While Latin American postcolonial notions of the global south such as Escobar's often seem to invoke or advocate the end of development and modernity *tout court*, African research, in contrast, tends to emphasize the continuing popular appeal of 'the modern,' but then a modernity of local African forms, at best in conversation with but not simply derived from European examples. Jean and John Comaroff's rich work has partly been spun around this argument. James Ferguson agrees with the ongoing popularity of modernity in Subsaharan Africa, but seems more suspicious than the Comaroffs of a mere superficial copying of European global modern forms. He describes a widespread popular political critique of the absence of once promised development European style. Other work also points to a copying of modern political forms in Africa without such forms attaining the real substance; democracy, bureaucracy, or human rights are cases in point. In India, some anthropologists like Arjun Appadurai have emphasized, like the Comaroffs, 'non-Western modernities.' But this is far from a dominant position in the South Asian literature: many postcolonial Indian authors have begun rejecting modernity altogether and identifying it as another erasure of local subjectivities by colonialism, not unlike Escobar and Santos. These are, once more, antinomies within and of 'the global south.'

The newest branch on the politicointellectual tree of 'the global south' idea is the emergence of 'southern theory' or 'theory from the south.' While Santos, Escobar, and others could certainly claim that they had been doing something worthy of the name, it only emerged as an explicit epithet in the later 2000s. The notion has first been coined in the work of the Australian sociologist Raewyn Connell. The global south was here firmly reconnected with its antineoliberal and anticapitalist social movement origin, including, although still hesitantly, the origins in the first and second internationals, and was separated once more from the 'georealist' circle of meanings that it had acquired by high state politics around 2000. The global south was taken as a vital location to theorize about the injustices, dispossessions, and political dialectics of Northern dominated world capitalism as such. The work of Connell, however, is largely about 'southern' intellectuals, and is clearly a first approach with significant omissions and some mistakes of judgment. Jean and John Comaroff gave the idea a more robust and anthropological twist – including a subtle shift of meaning expressed in the change from 'southern theory' to 'theory from the south.' They thus completed the genealogical circle by suggesting that the experiences of the violently dispossessed populations of the global south form a rich and necessary vantage point for understanding the ongoing dispossessions and disenfranchisements everywhere, including in the Northern metropolises and the deindustrialized landscapes of postsocialism. For them, the north and the south, the west and the east are caught up in the same 'all embracing process' of stepped-up capitalism. One can read their text, offered as an introduction to a set of earlier South African ethnographic episodes, as a radicalization of their earlier perspectives on modernity and the postcolony, or as an ethnographic deepening of what one used to call class struggle. But now, class struggle is in the context of populations that have been made expendable rather than directly exploitable, the lumpen undersides of capitalist transformations driven by Northern as well as local capitalists, power wielders, rentiers, and profiteers. Theory from the south, in their approach, is offered, too, as a counterpoint to northern neoliberal modes of knowing that, they claim, are increasingly forced into a bleak empiricism as sociality becomes reduced to law, contracts, and administration. In the global south, they argue, one does not have the luxury to know without theory.

Acknowledgment

We thank Dr Marc Edelman, CUNY, and Stefan Guga, Central European University, for advice.

See also: Capitalism: Global; Capitalism: The History of the Concept; Capitalism; Colonization and Colonialism, History of; Globalization and Religion; Globalization and World Culture; Globalization, Anthropology of; Globalization: Geographical Aspects; Modernity: Anthropological Aspects; Modernity; Modernization and Modernity in History; Neoliberalism; Social Action and Social Justice; Social Justice and Social Inequalities; Third World; Urban Neoliberalism.

Bibliography

Appadurai, Arjun, 1996. Modernity at Large: Cultural Dimensions in Globalization. University of Minnesota Press, Minneapolis.

Borras, Saturnino, Edelman, Marc, Kay, Christobal (Eds.), 2008. Transnational Agrarian Movements Confronting Globalization. Wiley, London.

Broad, Robin, Heckscher, Zahara, 2003. Before Seattle: the historical roots of the current movement against corporate-led globalization. Third World Quarterly 24 (4), 713–728.

Chakrabarty, Dipesh, 2007. Provincializing Europe: Postcolonial Thought and Historical Difference. Princeton University Press, Princeton.

Comaroff, Jean, Comaroff, John, 2012. Theory from the South: Or, How Euro-America Is Evolving toward Africa. Paradigm Publishers, Boulder.

Connell, Raewyn, 2007. Southern Theory: The Global Dynamics of Knowledge in Social Science. Polity Press, London.

Davis, Mike, 2007. Planet of Slums. Verso, London.

Dirlik, Arif, 2004. Spectres of the Third World: global modernity and the end of the three worlds. Third World Quarterly 25 (1), 131–148.

Dirlik, Arif, 2007a. Global Modernity: Modernity in the Age of Global Capitalism. Paradigm Publisher, Boulder.

Dirlik, Arif, 2007b. Global south: predicament and promise. The Global South 1 (1 and 2), 1–11.

Edelman, Marc, 2000. Peasants against Globalization: Rural Social Movements in Costa Rica. Standford University Press, Standford.

Escobar, Arturo, 2004. Beyond the Third World: imperial globality, global coloniality and anti-globalisation social movements. Third World Quarterly 25 (1), 207–230.

Escobar, Arturo, 2008. Territories of Difference: Place, Movements, Life, Redes. Duke University Press, Durham and London.

Ferguson, James, 2006. Global Shadows: Africa in the Neoliberal World Order. Duke University Press, Durham and London.

Hardt, Michael, Negri, Antonio, 2001. Empire. Harvard University Press, Boston.

Hardt, Michael, Negri, Antonio, 2006. Multitude: War and Democracy in the Age of Empire. Penguin, London.

Harvey, David, 2005. The New Imperialism. Oxford University Press, Oxford.

Harvey, David, 2007. A Brief History of Neoliberalism. Oxford University Press, Oxford.

Juris, Jeffrey, 2008. Networking Futures: The Movements against Corporate Globalization. Duke University Press, Durham and London.

Kalb, Don, 2005. From flows to violence: politics and knowledge in the debates on globalization and empire. Anthropological Theory 5 (2), 176–204.

Klein, Naomi, 2008. The Shock Doctrine: The Rise of Disaster Capitalism. Penguin, London.

Koonings, Kees, Kruijt, Dirk (Eds.), 2009. Mega-Cities: The Politics of Urban Exclusion and Violence in the Global South. Zed Books, London.

Prashad, Vijay, 2013. The Poorer Nations: A Possible History of the Global South. Verso, London.

Santos, Boaventura de Sousa, 2006. The Rise of the Global Left: The World Social Forum and Beyond. Zed Books, London.

Santos, Boaventura de Sousa (Ed.), 2007. Another Knowledge Is Possible: Beyond Northern Epistemologies. Verso Books, London.

Global Violence Against Women

Elizabeth A Swart, University of Southern California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Gender-based violence exists in virtually every culture around the globe, although local manifestations are highly dependent on structural inequalities; cultural definitions of gender; and the levels of violence in the larger society. Responses to gender-based violence in the West began as criminal justice interventions but expanded in the 1990s to include a global perspective of women's human rights, as well as greater attention to the collection of internationally comparable data sets. The current discourse (2014) incorporates dialogue on the intersectionality of oppressions, rethinking the concept of agency and focusing attention on indigenous women's perspectives.

Introduction

In the 1970s, with the growth of the women's movement in the Western world, advocacy about violence against women emphasized reforming the criminal justice model so that it would be responsive to cases of intimate partner and familial violence. In the United States, that campaign culminated in the passage of the Violence Against Women Act (VAWA) in 1994, which for the first time provided federal funding for emergency women's shelters, hotlines, and other social services. The VAWA, as reauthorized in 2013, includes protections for immigrant women, LGBT populations, and Native American survivors of gender-based abuse.

In 1995, the UN Fourth World Conference on Women issued a precedent-setting articulation of women's human rights. Known as the Beijing Platform, this document is considered by many to be the clearest and most compelling argument to date, for seeing women's rights as 'human' rights. With the globalization of the women's movement have come data-collection initiatives that aim for international comparability (WHO, 2006; Johnson et al., 2008; DHS Macro, 2014), as well as ongoing refinement of the definition of global violence to women, which now includes 'cultural' forms of violence, such as female genital mutilation (FGM). The discourse on 'women's agency' in relation to global violence is increasingly understood as inseparable from community.

International Institutional Response

The effort to address global violence against women is inextricably linked to the globalization of the women's movement. In 1947, the establishment of the United Nations' Commission on the Status of Women 'was among the most significant early developments' in the globalization of women's struggle for rights and equality. The Commission's mission was to 'raise the status of women, irrespective of nationality, race, language or religion' and to create "equality with men in all fields of human enterprise, and to eliminate all discrimination against women in the provision of statutory law, in legal maxims or rules, or in interpretations of customary law" (Penn and Nardos, 2003). These goals were, however, not specifically equated with the "need to eliminate sexual and physical violence against women and girls" (Penn and Nardos, 2003: p. 3). Nevertheless, the Commission was responsible for shaping the language of the 1948 Universal Declaration of Human Rights so that gender-insensitive language was removed and the document was tailored to affirm the inherent equality of women and men. Progress continued in 1952 when the UN General Assembly adopted the Convention of the Political Rights of Women (1954), which affirmed the inherent entitlement of women around the globe to "vote in any election, run for any political office, exercise any public function, and hold any public position that a man might hold" (Penn and Nardos, 2003). The Commission also began a partnership with UNESCO to develop equal access educational programs. And, although inequality in education is an ongoing problem in the twenty-first century, the beginning of consciousness about the disparity between girls' and boys' educations began with the Commission's early work.

From 1955 to 1962, the United Nations adopted two measures related to marriage, which were intended to reduce violence to women that was a consequence of marriage. The first was the Convention on the Nationality of Married Women (UN, 1958; Penn and Nardos, 2003), which was intended to secure nationality rights for women that is independent of the nationality of their spouses. The second measure related to the consent to marriage and minimum age for marriage (UN, 1964; Penn and Nardos, 2003), the purpose of which is to reduce the forced marriage of very young girls. The Convention, however, was weak, requiring premarital consent by both parties but leaving the minimum age of marriage to national governments.

The UN declared that 1975 would be known as International Women's Year and that the following 10 years would be the Decade of Women. During this fertile period for the confrontation of global violence against women, 'cultural' practices, such as FGM, were redefined as types of gender-based violence that constituted violations of women's human rights. The definition of violence continued to expand, incorporating structural and cultural concerns as well as health and survival-related issues. The Commission on the Status of Women, as well as General Assembly and the UN Economic and Social Council, began to focus on issues of poverty and structural inequality (Penn and Nardos, 2003). The first global conference on women's issues in Mexico City in June 1975 articulated

the World Plan of Action for the Implementation of the Objectives of International Women's Year, which consisted of three main objectives: promoting gender equality; facilitating the integration of women into all development efforts; and strengthening the contribution of women to the global movement for world peace (Penn and Nardos, 2003: p. 6). In 1979, the landmark Convention of the Elimination of all Forms of Discrimination against Women (CEDAW) was adopted by the General Assembly. Since its adoption, more than 150 nations have agreed to be bound by its provisions, although many countries have ratified the Convention without passing national implementing legislation. The Convention establishes an international bill of rights for women and mandates the establishment of a Committee on the Elimination of all Forms of Discrimination against Women, whose job it is to see that the provisions of the Convention are observed.

The International Year of the Woman was followed by global conferences on women in Copenhagen in 1980 and in Nairobi in 1985. Both gatherings brought nongovernmental organizations into prominence and influence. But it was the Nairobi conference, in particular, which cut across traditional lines of race, class, and culture to form alliances and coalitions among women from the Global North and South. A document entitled the 'Nairobi Forward-Looking Strategies for the Advancement of Women' expressed an increased urgency in the need to address cultural issues, such as FGM, as human rights issues. Because of the strength of this platform, CEDAW rectified the omission of gender-based violence from its own Convention in 1992 and categorized gender-based violence for the first time as a kind of discrimination under CEDAW. In 1993, building on the wave of grassroots human rights activism that had culminated in the Vienna Conference on Human Rights in June of that year, the UN General Assembly passed the Declaration on the Elimination of Violence Against Women, the "first international human rights instrument designed exclusively to deal with violence against women" (Penn and Nardos, 2003).

In September 1995, the United Nations held the Fourth World Conference on Women in Beijing, China. It was the largest UN conference to date, significant not only for the number of nations represented but also for the largest-ever representation of nongovernmental organizations. The Platform for Action which emerged from that conference is "among the most comprehensive human rights documents ever articulated on behalf of the world's women" (Penn and Nardos, 2003: p. 11). It established a list of 12 'priority action' areas including the persistent and increasing burden of poverty on women; inequalities in and unequal access to education and training; inequalities in and unequal access to health care; violence against women; effects of armed conflict on women; inequality in economic structures and policies and resources; inequality between women and men in sharing power and decision-making at all levels; insufficient methods to promote the advancement of women; lack of respect for and protection of women's human rights; stereotyping of women in the media and unequal access to media and communication systems; gender inequalities in the management of natural resources and in safeguarding the environment; persistent discrimination against and violation of the rights of the girl child (UN Women, 1995; Penn and Nardos, 2003: p. 12). The Beijing Platform is appropriately lauded as a landmark for women's human rights and much progress has been made in achieving the platform's goals. Although an increasing number of advocates and organizations recognized that legal and political measures were insufficient to confront the magnitude of global violence against women, new definitions were elusive.

Beyond Legal Interventions

By the turn of the twenty-first century, the discourse had turned from political definitions to ones rooted in situated understandings of social inequalities. In many parts of the world, women and girls have no access to legal mechanisms that can aid them, regardless of the lofty ideals that are codified by national and international assemblies. Violence between intimate partners began to be viewed as inseparable from "the violence of poverty, hunger, social exclusion, and humiliation" (Merry, 2009) and as

> … an umbrella term for a wide range of violations from rape during wartime, to sexual abuse in prisons, to insults and name-calling within marriages. Although the early movement against gender violence in the USA and internationally centered on rape and battering in intimate relationship international, activists continue to expand the scope of violence against women, to include cultural practices such as genital-cutting, illegal acts such as dowry deaths, the trafficking of women as sex workers, the effects of internal wars such as displaced people and the vulnerability to violence experienced by migrants in the context of contemporary globalization.
>
> Merry, 2009: p. 3

Indeed, understanding the intersectionality of gender with other forms of oppression locates violence not only in the gendered identities of the parties, but also in relation to cultural definitions and gender norms. Issues such as education and labor have become crucial in analyses of both definition and intervention. The World Bank reports that "two thirds of the 960 million illiterate people in the world are women and, of the 130 million children who received no primary education in 1990, 81 million were girls" (Penn and Nardos, 2003: p. 25). Such inequalities increase the likelihood that women and girls will experience gender-based violence through early pregnancy, forced marriage, sexual harassment, sexual violence, and early death in child-bearing (Penn and Nardos, 2003).

Similarly, women and girls work more daily hours than men but own very little of the world's wealth – performing nearly two-thirds of the work hours, receiving one-tenth of the world's income, and owning less than one-hundredth of the world's property (Frankenhaeuser et al., 1991). Women's unpaid workload and the increasing number of households headed by poor single women have resulted in more women and girls seeking employment outside the home, as well as in women migrating from rural to urban areas for jobs; with both trends exposing women to risk of gender-based violence. According to the UN's Special Rapporteur on Violence against Women, economically disadvantaged women are more vulnerable to sexual exploitation, harassment, and slavery than other women (Penn and Nardos, 2003: p. 25). Globalization, migration, and urbanization have created a market for women

in global sweatshops, which have created black holes of lawlessness and impunity where gender-based violence can flourish. The maquiladora factories along the US–Mexican border, particularly in Ciudad de Juarez, are the sites of hundreds of unsolved 'femicides' since 1993 (Segura and Zavella, 2007; Eisenstein, 2009).

Such black holes have left women and girls vulnerable to trafficking for sexual exploitation, the fastest growing criminal industry in the world. At least 20.9 million adults are bought and sold worldwide into commercial sex slavery and forced labor annually, as well as about 2 million children (ILO, 2012; UN, 2012). Women and girls make up about 98% of victims exploited for sexual purposes. Human trafficking of US citizens within the UN is also a growing business (UN, 2012). Experts estimate that at least 244 000 American children and youth – most of them very young girls – are trafficked within the United States each year for sexual exploitation. The average age of recruitment of girls into forced prostitution is 12–14 years (Estes and Weiner, 2001).

At the time of this writing (2014), the socio-cultural definition of gender-based violence has become instrumental in both grassroots and national campaigns (Bahun-Radunovic and Rajan, 2008; Merry, 2009), although efforts to reform legal and criminal justice systems continue (Bond and Phillips, 2008; Bahun-Radunovic and Rajan, 2008).

Types of Global Violence against Women

Domestic and Intimate Partner Violence

Violence against women by their male partners sustains patriarchal family structures, creates a hierarchy of male domination, and maintains women in a position of subordination (Seager, 2009: p. 28). Domestic or intimate partner violence is the most ubiquitous form of violence against women around the world. There "is virtually no place where it is not a significant problem … and women of no race, class or age are exempt from its reach" (Seager, 2009: p. 28). Although prevalence rates are difficult to compare because of definitional and measurement issues, a UN global review reports that 35% of women worldwide have experienced either physical and/or sexual intimate partner violence or non-partner sexual violence. Some national violence studies show that up to 70% of women have experienced physical and/or sexual violence in their lifetime from an intimate partner (UN Women, 2012). In Australia, Canada, Israel, South Africa, and the United States, intimate partner violence accounts for between 40 and 70% of female murder victims (Krug, 2002). Although advocates continue the struggle to expose and eradicate intimate partner violence, in many countries it is still condoned by the state as a 'private matter' (Seager, 2009: p 28).

Sexual Violence

Women around the world experience rape, often by men they know. In many countries, advocates have succeeded in changing the definition of rape from nonconsensual sex to a "violent act intended to assert male domination and control" (Seager, 2009: p. 58). Nevertheless, in most if not all countries, rape carries a social stigma that reduces reporting. Experts estimate that the actual incidence of rape may be 50 times greater than the number actually reported. In the US, up to 700 000 rapes are estimated to occur each year (Seager, 2009: p. 58). In South Africa, a woman is raped every 83 seconds (Seager, 2009: p. 58.) Around the world, the problem of institutional rape is on the rise, as is the rape of women in prisons, jails, border-crossings, and refugee camps (Segura and Zavella, 2007; Leatherman, 2011).

The subject of rape as a weapon of war drew international attention through the testimonies of female survivors of genocide in Bosnia–Herzegovina, as well as through the subsequent establishment of the International Crimes Tribunal for Yugoslavia in 1993. The Tribunal "created new categories of crime that were themselves gendered" (Giles and Hyndman, 2004) and for the first time in history, rape was prosecuted as a weapon of war and a crime against humanity (Giles and Hyndman, 2004). Women told stories of rape as a genocidal tool; of being imprisoned in rape camps; of being tattooed and marked with enemy insignia; and of being forced to bear their captors' children. Radhika Coomaraswamy commented on these atrocities by noting that "to rape women with impunity and to mark their bodies with the symbols of the other side is to assert domination and to symbolically assault ethnic identity in its most protected space" (Coomaraswamy, 1999: p. 10). Mass rape is also an efficient means of displacing surviving postconflict populations, whose families and communities are torn apart by shame, horror, and trauma (Coomaraswamy, 1999; Giles and Hyndman, 2004).

According to Leatherman (2011), sexual violence as a weapon of war has expanded to epidemic proportions. In 2006, the UN reported 27 000 sexual assaults during conflicts in Congo's South Kivu province alone (Leatherman, 2011: p. 120). Conflict-related sexual violence crosses thresholds related to type of sexual violence; target of sexual violence; extreme coercion of the sexual subject; and the loss of neutrality or safe space for the victims of sexual predation. The type of sexual violence is often public gang rape, perpetrated in schools, marketplaces, churches or other locations not normally conflict-related. Violated bodies are often left on display. Rape may also be performed with material objects – bottles, guns, spears – or victims may be knowingly raped by individuals carrying HIV/AIDS, with the intention of infecting the victim. The range of targets of sexual violence continues to expand – women in their 80s and children as young as 2 years old are currently (2014) raped with impunity in the Congo (Leatherman, 2011). Extreme coercion – forcing an unwilling perpetrator to commit acts of sexual violence, such as forced incest – imbues a sense of powerlessness and loss of agency among victim populations. Conflict-related sexual violence acknowledges no neutral zones or white flags and is perpetrated in churches, sanctuaries, refugee camps, homes, and other supposedly safe places. Such egregious sexual violence destroys the fabric of families; creates lasting wounds, such as PTSD and mental health disorders; neutralizes and disempowers communities that might once have considered fighting back; and normalizes sexual violence in postconflict society.

In 2000, the UN Security Council unanimously adopted Resolution 1325 (S/Res/1325), which aims to protect women and girls from violence, including rape as a weapon of war. It also advocates for the inclusion of women in conflict

resolution programs. The Inter-Agency Task Force on Women, Peace and Security facilitates collaboration on this Resolution among more than 20 UN agencies, as well as *between* UN entities and nongovernment organizations working on the ground in specific conflict areas (Olonisalom et al., 2011).

Traditional and Cultural Practices

FGM is a "cultural practice involving the removal of various parts of the genitalia of girls and young women" (Penn and Nardos, 2003: p. 89). It is performed in at least 27 African nations and in several countries in Asia and the Middle East (Yoder et al., 2013), as well as by immigrant and refugee communities in Western nations. Experts cite Sudan, Ethiopia, Mali, Egypt, Sierra Leone, and India as countries with the highest rates of FGM (Penn and Nardos, 2003: p. 88). Although the practice has come under fierce criticism by medical and mental health professionals around the world, the World Health Organization (WHO) estimates that in 1994, at least 114 million girls and women endured some type of FGM (Penn and Nardos, 2003: p. 89). FGM procedures, usually performed without anesthesia and in unsanitary conditions, can result in complications such as shock, hemorrhaging, urinary retention and infection, fever, painful intercourse, complications during childbirth, and death. There are three types of FGM. The first, 'clitorectomy' (sunna) involves partial or complete removal of the clitoris. 'Excision' involves the removal of the clitoris and the labia minora. 'Infibulation,' the most dangerous and most common form of FGM, involves the removal of the clitoris, the labia minora, as well as the inner surface of the labia majora, followed by the sewing together of the vulva so that a small opening for urine and menstrual blood remains. A process known as 'defibulation' – a surgical reopening – must sometimes be performed before sexual intercourse is possible. Scar tissue created by infibulation often inhibits childbirth or makes it extremely painful.

In countries where FGM is pervasive, women sometimes report submitting to it (for themselves or for their daughters) as a means of community acceptance – a necessary prerequisite for marriage and employment. Uncut girls are often considered ugly, unclean, promiscuous, and unmarriageable. For this reason, women's empowerment within communities and education about health issues must be addressed if FGM is to be eradicated. In many countries, indigenous women are speaking out against the procedure. In Kenya, for example, advocates have created alternative rites of passage for girls, which include a community-based initiation to womanhood that does not include genital cutting (Chege et al., 2001).

Dowry murder (bride-burning or dowry death) is "the murder of a woman who is set on fire by her husband or in-laws for the purpose of keeping the dowry that the woman brought into the marriage. By burning his wife, the husband can remarry and obtain another dowry" (Penn and Nardos, 2003: p. 97). Dowry murder, particularly common in India but also occurring in other parts of Asia and the Middle-East, is most commonly done by pouring kerosene on a woman and setting her on fire. Because many homes have kerosene-fueled stoves, those who commit the murder may attribute the death to a cooking accident (Penn and Nardos, 2003: p. 97). Studies on rate and prevalence of dowry murder in India show that it persists in all levels of society, despite the passage of the Dowry Prohibition Act in 1961. Families in lower socio-economic levels may use it as a way to escape subsistence poverty. Educated classes may use it as a method of retaining their status (Penn and Nardos, 2003: p. 99). Ironically, because the Dowry Prohibition Act criminalizes both dowry-giving and receiving, it discourages reporting of the dowry-related crime. Experts estimate that 5000 women per year in India may be victims of dowry-related murder. However, it is important to understand that violence in India, as in many developing countries, is often viewed by the West "in its most extreme incarnation" (Narayan, 1997: p. 101). It is helpful to remember that "the level of reported homicide by a partner in proportion to the country's population is, in fact, similar in India as in the United States …" where spousal murders are related not to dowry but to gun violence (Narayan, 1997: p. 99).

Honor killing, an ancient practice in which men kill female relatives in the name of family honor, are also on the rise in the twenty-first century. These killings occur in societies where women are not only considered the property of male relatives but also the standard bearers of male familial honor. Although exact numbers are difficult to obtain, experts estimate that in 1997 in one province of Pakistan alone, 300 women were killed to protect or avenge family honor. In Yemen, approximately 400 honor killings took place in 1997, and 52 honor killings in Egypt during the same year (Penn and Nardos, 2003; Bahun-Radunovic and Rajan, 2008). Women are killed for refusing arranged marriages; for being victims of rape; for wearing clothes male relatives deem inappropriate; or for adhering to Western standards of behavior (Penn and Nardos, 2003; Patel and Gadit, 2008). Unfortunately, honor killings are rapidly spreading to the West, with the increase in refugee and immigrant populations from countries where the tradition is entrenched.

Measuring Gender-Based Violence

The past three decades have shown increasing refinement of methods for collecting data on gender-based violence. The first population study of gender-based crimes among the general population was carried out in the late 1960s in the United States. These victimization surveys interviewed large samples of adults, asking about their perceptions and experience of the crime in the 12 months prior to the interview (Johnson et al., 2008: p. 11). This methodology became increasingly popular in the 1970s as a tool for assessing the level of crime in society (Johnson et al., 2008) and was further refined in other Western developed nations. Researchers considered these data important complements to official police data because many crimes – particularly rape and domestic violence – are never reported to the police (Johnson et al., 2008). But although these data provided information on crime from the victims' perspective, it became increasingly clear that they were not adequately measuring gender-based crimes. Factors included lack of sensitivity in wording of questions; lack of special training for interviewers; lack of safety supports for women interviewees who might be in danger for disclosure; the traumatization caused by the reporting; and the undercounting of women's experiences of lifetime or long-term violence. Such critiques led

to the redesign of the US-based National Crime Victimization Survey (NCVS) in 1992 to improve and expand screening questions, which increased prevalence rates.

Statistics Canada was the first agency to take on a major revamping of the methodology of measuring women's victimization. In 1993, "a survey dedicated to exploring a wide range of women's experiences of violent victimization was developed and implemented, using the basic crime victimization survey methodology" but with female interviewers who were specially trained and the referral of respondents to community agencies and supports for interviewers (Johnson, 1996, 1998). Since then, similar surveys in the US, Australia, New Zealand, Finland, and Sweden have been successfully carried out.

Population-based surveys of representative samples of women are now generally recognized as the "most reliable method of acquiring statistically reliable estimates of the nature and extent of violence against women. At least 70 countries have conducted population surveys on violence against women but, although there are commonalities among the surveys, there are enough important differences to prevent reliable country-to-country comparison of results" (Johnson et al., 2008: p. 14). Those problems include the definition of violence; the specific wording of questions; reference periods (previous year or adult lifetime); and sampling methods.

Despite differences in definition and cultural manifestation of gender-based violence, there exist several exemplary efforts to measure gender-based violence internationally. In 2005, the World Health Organization (WHO) issued a report on violence to women by intimate partners. It was based on 24 000 interviews with women in 10 countries (Bangladesh, Brazil, Peru, Tanzania, Ethiopia, Thailand, Japan, Samoa, Namibia, and Serbia/Montenegro) (WHO, 2006). The study found important variations among nations, as well as among rural and urban populations within nations. Finding showed that between 15 and 71% of women who had ever been in an intimate relationship had experienced physical or sexual assault during their lifetime (WHO, 2006; Merry, 2009).

The Demographic and Health Surveys (DHS) are household surveys sponsored by Macro International and the US Agency on International Development (DHS Macro, 2014), which focus on health-related practices and behaviors in 90 developing countries. DHS country surveys contain a Domestic Violence Module which collects representative data on lifetime experience of intimate partner violence. DHS surveys use large nationally representative samples, as well as a standardized questionnaire and a methodology to ensure comparisons over time and among countries. The module includes nine questions related to acts of physical and sexual violence, as well as a series of questions describing psychologically controlling or abusive behaviors. Comparability problems remain, however, because some countries opt out of the domestic violence module or reduce it to a single prevalence-gathering question (Johnson et al., 2008).

The International Violence Against Women Survey is an important data collection method which "incorporates a standardized methodology and has as a primary focus the strengthening of legal measures and improving criminal justice responses to violence" (Johnson et al., 2008: p. 15). Analysis of survey results from nine countries provide information not only on women's experiences but also on how the criminal justice system responds and how civil society may effectively intervene (Johnson et al., 2008: p. 167).

Rethinking Agency

When the Beijing Platform was still hot off the press, indigenous women pointed out that the platform was flawed due to its overemphasis on gender discrimination and on an individual rights-based philosophy (UN Women, 1995; FIMI, 2006). Before the conference adjourned, more than 100 indigenous women from 26 countries had written a declaration outlining their critique of the Platform's theoretical viewpoint (UN Women, 1995). For indigenous women, human rights violations "are based not only on gender but on the interplay of gender" with other aspects of identity (Cunningham, 2006; FIMI, 2006: p. 10). Issues omitted in the Beijing Platform include the dispossession of territories, pollution of water, forced relocations, and other activities which, although not gender-based in their intent, have a "disproportionate effect on women" (FIMI, 2006: p. 13). Indigenous women call for an integrated analysis of violence against women that "recognizes both the near-universality of violence against women and the specificity of violence perpetrated on the basis of distinct but overlapping identities" (FIMI, 2006: p. 10). They point out that "for indigenous women, the systematic violation of their collective rights as indigenous people is the single greatest risk factor for gender-based violence – including violence perpetrated within their communities" (FIMI, 2006: p. 7).

The current discourse on global violence against women challenges scholars in the Global North *and* South to rethink agency as it relates to gender-based violence and to women's alternative ways of "being and seeing in the face of violence" (Campbell and Mannell, forthcoming). Scholars hope to 'push the boundaries of how we conceptualize agency' and to break down the neoliberal assumption that encourages women to think that leaving situations of violence is the only appropriate course of action. Current discourse pays particular attention to situations of extreme poverty, armed conflict, and migration as they challenge an individualistic-behavioral approach to gender-based violence (Campbell and Mannell, forthcoming). It also acknowledges the need to actively engage men on all levels of violence prevention and cessation (Kimball et al., 2013). Seeking global cooperative strategies capable of adapting to localized contexts (Bahun-Radunovic and Rajan, 2008) advances the struggle to end global violence against women.

See also: Crime and Gender; Feminist Criminology and Gender Studies; Gender and Education; Gender and Feminist Studies in Geography; Gender and Place; Gender and Women's Studies, Applied Research On; Gender and the Law; Gender, Economics of; Gender: Gendering of Categories; Intimate Partner Abuse, Applied Research On; Teens, Gender, and Self-Presentation in Social Media; Trafficking of Women and Children in Latin America; Women and Criminal Justice; Women and Psychiatry; Women's Status and National and International Security.

Bibliography

Bahun-Radunovic, Rajan, J.V.G. (Eds.), 2008. Violence and Gender in the Globalized World: The Intimate and the Extimate. Ashgate, Burlington, VT.

Bond, J., Phillips, R., 2008. Violence against women as a human rights violation: international institutional responses. In: Renzetti, Edleson, Bergen (Eds.), Sourcebook on Violence against Women. Sage, Thousand Oaks, CA.

Campbell, C., Mannell, J. Intimate partner violence in marginalized settings: re-theorizing agency for global public health. International Journal of Global Public Health, forthcoming.

Chege, J.N., Askew, I., Liku, J., 2001. An Assessment of the Alternative Rites of Passage for Encouraging Abandonment of Female Genital Mutilation (FGM) in Kenya. USAID, Washington, DC.

Coomaraswamy, R., 1994. Preliminary Report Submitted by the Special Rapporteur on Violence against Women, Its Causes and Consequences. Doc. E/CN.4/1995/42. UN, New York.

Coomaraswamy, R., 1999. Question of Honor: Women, Ethnicity, and Armed Conflict. Third Minority Rights Lecture, May 25, Geneva, Switzerland. Retrieved: http://www.sacw.net/Wmov/RCoomaraswamyOnHonour.html.

Cunningham, M., 2006. Indigenous women's visions of an inclusive feminism. Development 49 (1), 55–59.

Demographic and Health Survey, 2014. http://www.measuredhs.com.

Eisenstein, H., 2009. Feminism Seduced: How Global Elites Use Women's Labor and Ideas to Exploit the World. Paradigm Publishers, Boulder.

Estes, R.J., Weiner, N.A., 2001. The Commercial Exploitation of Children in US. University of Pennsylvania: School of Social Work, Canada and Mexico.

Foro Internacional de Mujeres Indigenas, 2006. Mairin Iwanka Raya – Indigenous Women Stand against Violence: A Companion Report to the United Nations Secretary-general's Study on Violence against Women. International Indigenous Women's Forum, New York.

Frankenhaeuser, M., Lundberg, U., Chesney, M., 1991. Women, Work and Health: Stress and Opportunities. Plenum, New York.

Giles, W., Hyndman, J., 2004. Sites of Violence: Gender and Conflict Zones. University of California Press, Berkeley.

International Labor Organization (ILO), 2012. ILO Global Estimates of Forced Labor: Risk and History. ILO, New York.

Johnson, H., 1996. Dangerous Domains: Violence against Women in Canada. Nelson Publications, Toronto.

Johnson, H., 1998. Rethinking survey research on violence against women. In: Dobash, R.P., Dobash, R.E. (Eds.), Rethinking Violence against Women. Sage, Thousand Oaks, CA, pp. 23–52.

Johnson, H., Ollus, N., Nevala, S., 2008. Violence against Women: An International Perspective. Springer, New York.

Kimball, E., Edleson, J.L., Tolman, R.M., Neugut, T.B., Carlson, J., 2013. Global efforts to engage men in preventing violence against women: an international survey. Violence Against Women 19 (7), 924–939.

Kishor, S., Johnson, K., 2004. Profiling Domestic Violence: A Multi-country Study. ORC Macro, Calverton, MD.

Krug, E.G., 2002. World Report on Violence and Health. WHO, Geneva.

Leatherman, J.L., 2011. Sexual Violence and Armed Conflict. Polity Press, Malden, MA.

Merry, S.E., 2009. Gender Violence: A Cultural Perspective. John Wiley, Malden, MA.

Narayan, U., 1997. Cross-cultural connections, border-crossings, and "death by culture." In: Narayan, U. (Ed.), Dislocating Cultures/Identities, Traditions, and Third World Feminism. Routledge Press, New York, pp. 83–117.

Olonisalom, F., Barnes, K., Ikpe, E., 2011. Women, Peace and Security: Translating Policy into Practice. Routledge, New York.

Patel, S., Gadit, A.M., 2008. Karo-kari: a form of honour-killing in Pakistan. Transcultural Psychiatry 45 (4), 683–694.

Penn, M.L., Nardos, R., 2003. Overcoming Violence against Women and Girls: The International Campaign to Eradicate a Worldwide Problem. Rowman & Littlefield Publishers, Inc., Lanham, MD.

Rudd, J., 2001. Dowry-murder: an example of violence against women. Women's Studies International Forum 24 (5), 513–522.

Seager, J., 2009. The Penguin Atlas of Women in the World. Penguin, New York.

Segura, D.A., Zavella, P., 2007. Women and Migration in the US–Mexico Borderlands. Duke University Press, Durham, NC.

Spindel, C., Levy, E., Connor, M., 2000. With an End in Sight: Strategies from the UNIFEM Trust Fund to Eliminate Violence against Women. UNIFEM, New York.

UN Office of Drugs and Crime, 2012. Global Report on Trafficking in Persons. UN, New York.

UN Women: UN Entity for Gender Equality and the Empowerment of Women,1995. Beijing Platform for Action. Retrieved at: www.un.org/womenwatch/daw/beijingplatform/plat1.htm.

UN Women: UN Entity for gender Equality and the empowerment of Women, 2012. Violence against Women Prevalence Data: Reports by Country. UN, New York.

United Nations, 1954. Convention on the Political Rights of Women. UN, New York.

United Nations, 1958. Convention of the Nationality of Married Women. UN, New York.

United Nations, 1964. Convention and Recommendation on Consent to Marriage, Minimum Age for Marriage and Registration of Marriages. UN, New York.

United Nations, 1979. Convention on the Elimination of All Forms of Discrimination against Women. UN, New York.

World Health Organization, 2006. Multi-country Study on Women's Health and Domestic Violence. WHO, New York.

Yoder, P.S., Wang, S., Johansen, E., 2013. Estimates of female genital mutilation/cutting in 27 African countries and Yemen. Studies in Family Planning 44 (2), 189–204.

Globalization and Health

Ronald Labonté, University of Ottawa, Ottawa, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by B. Chalmers, volume 9, pp. 6258–6261, © 2001, Elsevier Ltd.

Abstract

Globalization describes how nations, peoples, and economies are becoming increasingly interconnected and interdependent. Globalization has contributed to health improvements through diffusion of new health knowledge, low-cost health technologies, and human rights. Economic globalization based upon a neoliberal model of liberalized trade and investment and minimal government regulation, however, has brought many health risks. This article outlines the health risks of neoliberal globalization from structural adjustment of the 1980s, through economic financialization of the 1990s and the 2000s, to the financial crisis and global recession of 2008 and subsequent 'austerity agenda' of state retrenchment. It concludes with a discussion of high-level public policies that would ensure that globalization works for 'health for all'.

Introduction

Globalization is a relatively new term, first appearing in the 1970s and quickly supplanting earlier concepts sharing roughly similar terrain (internationalization, postcolonialism, new international economic order). Broadly defined, globalization describes how nations, businesses, and people worldwide are becoming more connected and interdependent through increased economic integration and communication exchange, cultural diffusion, and travel. Lee (2002) further considers globalization to be effecting change in spatial, temporal, and cognitive dimensions: social activities increasingly transcend borders and occur more rapidly, as do peoples' understanding of global events and their place in the world. By such an account, globalization is not a new phenomenon; rather, apart from small and geographically isolated settlements, this push, prod, and sharing against 'borders' has characterized much of the history of human societies. At base, it is driven by economic and political expansion, whether by trade or conquest.

Some historians date the 'modern' era of (Western) globalization with European colonial expansion and the Westphalia formation of nation states in the late seventeenth century. Mercantilism dominated globalization's policies and practices over the next two centuries, as governments supported their local merchant class in competition for an expanding global trade facilitated by technological advances in maritime navigation. A series of late nineteenth-century economic recessions, rising social inequalities, staggering inflation, and nationalist protectionism became the fertile ground from which the Great War emerged, which devastation and lessons failed to find a sufficiently firm political foothold to avoid a recurrence in the Second World War. This foothold became more secure in the 30 years following the Second World War, which saw a new pattern in globalization emerge with the creation of stronger multilateral governance bodies. The United Nations (UN) replaced the League of Nations. The International Monetary Fund (IMF) was created to maintain macroeconomic stability and avoid the recessionary crises underpinning both world wars. The World Bank (WB) was founded to finance reconstruction of war-torn Europe (and later, developing country economies and social infrastructure). The General Agreement on Tariffs and Trade provided a negotiating platform intended to interconnect national economies (at that time, of the high-income developed countries only) in mutually beneficial ways, with the expectation that this would also provide economic disincentives to future war.

Contemporary Globalization

A new 'contemporary' globalization began to emerge in the 1970s. More economic recessions, two oil price shocks, profligate international bank lending, an end to fixed currency exchange rates, and a 'stagflation' leading to sharp interest rate spikes in the world's (then) major economies combined to create a developing world debt crisis. This crisis provided an opportunity for neoliberal economic theories first advanced by Friedrich Hayek in the 1920s to become policy practice, in the form of structural adjustment programs imposed largely on the heavily foreign indebted countries of Africa and Latin America, which needed financial assistance from the IMF and WB to avoid government default and bankruptcy. The key elements of these programs, later codified as the Washington Consensus, are well known and included such requirements as:

- privatization of state assets to generate revenue,
- deregulation of economic markets to promote private sector growth,
- lower corporate and individual taxes to attract foreign direct investment,
- reduced government spending and increased cost recovery (user pay for public services), and
- increased trade and financial liberalization.

These policy prescriptions were intended to increase economic growth and the ability of governments both to service their foreign debt obligations and to reduce their debt burden. Evidence of their effectiveness in doing so is mixed, with other regions of the world not adopting these policies (notably Southeast (SE) Asian countries) performing much better over this period (Chang, 2005). The policies also resulted in a number of health-negative impacts, first documented with respect to children's health in the landmark United Nations International Children Emergency Fund publication, *Adjustment with a Human Face* (Cornia et al., 1987). Subsequent studies similarly found singularly negative health impacts of

structural adjustment programs, attributed to user fees that created barriers to health and education services, labor market adjustments from trade liberalization that led to unemployment or insecure informal work, and tax reductions and the loss of tariffs revenues that reduced government health and social protection spending (Breman and Shelton, 2001). But even as evidence and argument over the economic and health impacts of structural adjustment continued, the neoliberal policy context shaping contemporary globalization became firmly embedded.

Globalization's Positive Health

Even as this neoliberal model of global market integration strengthened its grip on countries' economic policies, civil society movements, the foreign aid arms of governments, and several UN agencies were emphasizing the social dimensions of global health development. Key innovations at this time included the 'child health revolution' of the 1980s, with its emphasis on low-cost health interventions: GOBI (Growth monitoring, Oral rehydration, Breastfeeding, Immunization), later followed by the triple FFF (Family planning, Female education, and Food supplementation). Some have argued that much of the improvement in global health during the last half of the twentieth century arose not from economic growth in developing countries, but from health technologies and knowledge transfers between wealthier (developed) and poorer (developing) countries (Deaton, 2003): one indication of globalization's health beneficence.

Ratification by most of the world's countries of international human rights conventions is seen as another health-positive facet of contemporary globalization. Several of these conventions make specific reference to the right to health, and most others deal with rights associated with access to social determinants of health. Though sometimes critiqued for a perceived Western bias (emphasizing individual rather than collective rights) and for their lack of enforcement (apart from 'naming and shaming' states that fail to fulfill their obligations under these treaties), international human rights have become a set of global tools used in the struggle for better health throughout much of the world (Chapman, 2002). A global diffusion of concerns for gender equity has also been attributed to globalization, from increased funding to improve conditions for maternal/child health, to new employment opportunities in export-oriented industries allowing women to earn income outside patriarchal social structures. That the poor working conditions in these outsourced factories have become a focus of international and civil society activism attests to another distinguishing quality of contemporary globalization: the breadth and reach of digital communication technologies. These technologies enable not only the more exploitative elements of open markets but also the platforms for oppositional civil society mobilizations.

The past 15 years have also seen global attention to health rise considerably. Health concerns directly informed three of the UN Millennium Development Goals, have been central in bilateral and multilateral development aid policies, and have led to the creation of over 100 global public–private partnerships for health. International financing for health followed suit, increasing substantially over this period, partly on the accumulating evidence that health should be considered an investment in economic growth, rather than a cost (World Bank, 1993; CMH, 2001).

Globalization's Negative Health

The emphasis on health as an investment relates to one of the more contentious positive health claims made for globalization: that of the contribution of liberalized markets to economic growth. This argument holds that trade and investment liberalization improves growth, which generates wealth that reduces poverty. Poverty reduction, in turn, improves health (poverty being the single greatest risk condition for disease), creating more productive and skilled workers, which spur the economy to even greater growth and more trickle-down health. It is by interrogating the links in this 'virtuous circle,' however, that globalization's health risks begin to emerge.

Most econometric studies do find that trade liberalization is associated with better growth, but this positive relationship is neither automatic nor always observed (Thorbecke and Nissanke, 2006). There also remains doubt about the direction of causality in these studies: does openness lead to growth, or does growth give rise to openness? Many development economists are critical of the assumption that liberalization is inevitably a 'global public good' for the economic growth it is presumed to engender. Some argue that trade liberalization rules are 'kicking away the ladder' of development policies successfully used by high-income countries in the past, and by middle-income SE Asian economies in the two decades prior to the birth of the World Trade Organization (WTO) in 1995 (Chang, 2005). Others point to evidence that trade and investment liberalization disproportionately benefits larger, wealthier nations with greater factor endowments and the political strength to set the terms of liberalization rules in their favor (Birdsall, 2006). One study of four different scenarios of a completed 'Doha Development Round' of WTO negotiations (so named as it was intended to give more attention to the economic interests of developing countries) showed net income gains for wealthy countries but net income losses for the world's poorest countries (Polaski, 2006). While the number of people living in extreme poverty (US \$1.25 day^{-1}) has decreased since liberalized trade began its acceleration in the 1980s, the 'rising tide' of global economic growth has not lifted people very far with poverty at the US \$2.50 day^{-1} level increasing by almost the same number (Chen and Ravallion, 2008). Expanding external trade has not been a powerful force for poverty reduction in developing countries.

There are three other ways in which trade liberalization and global market integration poses specific health risks: spread of disease, loss of government policy space and capacity, and increased labor market insecurity.

Spread of Disease

Infectious diseases such as the plague, smallpox, and cholera have long followed trade and military routes; it was a freighter's dumping of infected bilge waters that caused the Latin American cholera pandemic of the 1990s, and the deployment of UN Nepalese peace keepers that introduced cholera into Haiti in 2010. Trade has also been responsible for the movements of

insect or animal species destructive to natural resources essential for health (food, water, and biomass). Modern transportation allows such pathogens and pests to spread faster and farther, with the severe acute respiratory syndrome outbreak in the early 2000s shaking public health systems out of their recent complacency. New International Health Regulations monitored by the World Health Organization (WHO) followed quickly, as did ongoing 'pandemic preparedness' within and between countries and the 'securitization' of health within global governance structures (the WHO and the UN), wherein pandemic disease is viewed as a risk to national and international security (OUNHCHR, 2000). The global trade in fake medicines alongside the misuse of antibiotics in many countries, especially those that lack well-regulated public health systems, is heightening concern over the spread of multiple or extremely drug-resistant infections.

While increased research and policy attention is being paid to the globalization of infectious diseases, chronic non-communicable diseases (NCDs) such as cancer, cardiovascular disease, and diabetes have been steadily growing in prevalence and now account for the largest proportion of the global burden of disease, outpacing infectious diseases in all developing countries except for those in sub-Saharan Africa. The rise in NCDs, particularly problematic in many low- and middle-income countries still coping with infectious epidemics, is partly attributable to international trade and investment policies that are globalizing 'Western' lifestyles, including increased consumption of unhealthy products. The nutrition transition in many of these countries toward a low-nutrient, high-energy diet is occurring much faster and at earlier stages of development than it did in high-income countries (Popkin, 2009). Increased trade in tobacco and alcohol products, in turn, is associated with higher levels of consumption and health-related problems (Labonté et al., 2011), at the same time that trade and investment treaties are being used by transnational companies (or governments on their behalf) to challenge public health restrictions on advertising, points of sale, taxation, and other measures now widely accepted as essential tools in health promotion.

Loss of Government Policy Space and Capacity

Another aspect of trade and investment liberalization with indirect health outcome is the loss of governments' 'policy space' and 'policy capacity.' 'Policy space' describes "the freedom, scope, and mechanisms that governments have to choose, design and implement public policies to fulfill their aims" (Koivusalo et al., 2009: p. 105). Policy capacity refers to the fiscal ability of states to enact those policies or regulations, which depends upon their ability to capture sufficient revenue through taxation for this purpose. The increasing 'behind-the-border' reach of an ever larger number of multilateral, regional, and bilateral trade and investment treaties is shrinking policy space by prohibiting a range of domestic regulatory options that could be used to promote more equitable population health outcomes. Although governments retain substantial policy flexibilities within existing trade treaties, these flexibilities continue to be eroded through ongoing negotiations, notably the shift to bilateral and regional treaty negotiations, which are often 'WTO+,' limiting policy space to a much greater extent than current WTO rules. Of particular concern is the proliferation of investor state dispute settlement provisions in bilateral and regional treaties, which allow corporations to sue governments for regulations that result in expropriation of their investments, with expropriation often defined in very broad terms. Anticipating the cost of such disputes can lead governments to choose not to exercise the health flexibilities that might still exist within these treaties: a 'regulatory chill.'

Policy capacity, in turn, can be affected by liberalization's requirements for progressive reductions in tariffs. Many low- and middle-income countries rely much more upon tariffs for their tax revenue than do high-income nations. These tariff rates have come under intense trade negotiation pressures from high-income countries to be 'locked in' and rapidly reduced. In theory, governments should be able to shift their tax base from tariffs to sales or income taxes, assuming their economies grow with increased liberalization. In reality, most low- and middle-income countries subject to tariff reductions under structural adjustment loans from the IMF and WB were unable to do so (Baunsgaard and Keen, 2010), partly the result of inadequate institutions to implement alternate tax regimes. For a majority of these countries, there was a net decline in overall public revenues – a loss in policy capacity – with implications for spending in health, education, or public regulations that can affect primary and secondary prevention of disease.

Turmoil in Labor Markets

At the same time that many countries saw their policy space and capacity erode under structural adjustment, a global reorganization of production was unfolding through the creation of 'export processing zones' (EPZs) in low- and middle-income countries. Industrial manufacturing previously centered in high-income countries began to outsource production to EPZs, which are tax-free and provide a low-paid and often minimally regulated workforce (Martinez, 2004). Most of the world's trade no longer occurs between nations, but between multiple subsidiaries of transnational companies. Much of the value added (and the profit) is captured by the firm at the top of the chain, often in offshore financial centers ('tax havens'), and often by way of relentless pressure on competing suppliers to bear most of the risk associated with investments in plant and equipment and to contain labor and other production costs. Corporations' ability to relocate production to lower cost jurisdictions limits the ability of workers to negotiate better wages and working conditions, and to expand social provision through the political process that was at the core of the postwar compromise between workers and capital in the high-income world. The result: 'most workers are being squeezed' (Woodall, 2006). This is occurring in all countries of the world. Even those in low- and middle-income countries who have seen their incomes rise as a result of the growth in global production chains live in conditions variously described as 'vulnerable' or 'precarious,' where their employment is often informal, temporary, part-time, or generally lacking in adequate remuneration, security, or benefits (Standing, 2011). Social protection programs can compensate, at least in part, for job and labor income losses that result from trade liberalization. However, given the extent of global market liberalization in capital, goods, and services, most of the world's governments find themselves converging on a policy model in which they strive to make their own geographic location as competitive as

possible in international markets. Both labor market 'flexibilization,' with its stress-related health risks, and limits on the taxes that could finance improved health and social protection policies constitute important elements of competition state policies (Schrecker and Labonté, 2011).

Globalization's New Health Challenges: Financial Crises and Austerity

In parallel with the globalization of the 'real economy' of production and consumption since the 1980s, a new form of capital accumulation arose in the rapid growth of a 'financialized' economy. Aided by new digital technologies, investors and financial institutions discovered that it was easier and faster to make money from money than lending it to the 'real economy' of production and consumption. The scale of this financialization is almost hard to imagine, with the value of outstanding derivatives in 2011 exceeding $700 trillion, or more than 10 times the total value of the world's gross domestic product. In late 2011, and despite the 2008 Global Financial Crisis (GFC), this daily arbitrage amounted to almost $5 trillion (Bech, 2012).

The Great Financial Crisis

The proximate cause of the GFC was the collapse of the US housing market, leading to publicly financed trillion-dollar bailouts and countercyclical spending. Public deficits were created to cover the risks taken by private financiers, with one estimate of the total amount of public financing that went into the bank rescues among the G20 countries placing it at $11.7 trillion (Ortiz and Cummins, 2012), several hundred billion of which was direct subsidies. Some of this amount has since been repaid or recovered through sale of government shares in quasinationalized failed banks, although not without large public losses. Less evident but more systemic is the interest rate spreads on what governments and their central banks provide to banks and what they then borrow back to cover this lending, leading to a massive transfer of public wealth to the very institutions and individuals responsible for the GFC (Altvater and Mankopf, 2012).

The depth of global production chains meant that the GFC's credit crunch and subsequent Great Recession rippled rapidly across supply chains in low- and middle-income countries. While there is some worry that this will lead to a rise in extreme poverty (the greatest risk condition for disease) and increased child mortality, the most striking concern is a sharp rise in global unemployment, much of it concentrated among young adults, and creating a surplus (unemployed) labor pool of over 200 million (ILO, 2013). The GFC also accelerated a longer standing trend in rising wealth and income inequalities within and between most nations of the world, notwithstanding some notable exceptions in South America. Although still subject to empirical contestation, such inequalities have been argued to pose a number of health and social risks (Wilkinson and Pickett, 2009). While the GFC has negatively affected the pensions and savings of many of the world's middle and working classes, the 24 million whom investment banks refer to as 'high- and ultrahigh-net-worth individuals' saw their balance sheets decline for a year or two, before increasing by over 20% (Baxter, 2011), a remarkable feat the pre-1929 oligarchs never accomplished. Billionaire wealth rose 20% alone in 2012 over 2011, in what Forbes described as 'a very good year for billionaires' (Forbes, 2013).

The Austerity Agenda

After a brief period of stimulus spending by many of the world's high-income countries and those low- and middle-income countries that could afford it, most governments quickly signed on to 'the austerity agenda': a protracted period of state retrenchment argued as necessary to reduce government debt (much of it incurred to rescue failed banks) and to stimulate economic growth. The key tenets of austerity differ little from earlier structural adjustment (Ortiz and Cummins, 2013). Unlike structural adjustment, the austerity agenda began in those high-income countries that were the epicenter of the GFC before becoming a more globalized set of policy prescriptions with fiscal contractions most severe in the developing world. Although there was little consensus among economists on the wisdom of pursuing austerity, the majority of IMF post-GFC loans pushed for elimination of food and fuel subsidies; wage bill cuts including salaries of education, health, and other public sector workers; rationalizing (reducing) safety net expenditures; and pension reform to delay eligibility. Several countries have been advised to reform their public health systems and increase labor market flexibilization, while many governments are attempting to generate revenue by broadening their consumption taxes to items disproportionately consumed by poor households (Ortiz and Cummins, 2013).

The austerity agenda is expected to have mixed health impacts on populations in high-income countries, on the one hand potentially reducing discretionary expenditures on tobacco or excess alcohol consumption, but on the other hand increasing poverty rates for those in less secure employment settings; homelessness, particularly in the United States; and increased reliance on low-cost, highly processed obesogenic foods. Unemployment- and poverty/insecurity-related stress levels are also predicted to rise; and suicide rates since the crisis have increased by 12–15% in the worst affected European countries, at the same time that health and social protection financing has been dramatically cut back and user charges imposed on many services (Stuckler and Basu, 2013). The health implications of austerity in low- and middle-income countries will likely follow that which accompanied earlier, regional financial crises: increased social stratification and inequalities; reduced food intake, health care utilization, and education expenditure; and increased morbidity and mortality, experienced first and worst by women, rural populations, and the poor.

A Deepening Neoliberalism

It is important to understand these recent globalization health risks as an extension of a 40-year uncontrolled experiment in neoliberal economic theory. From neoliberalism 1.0 (structural adjustment) to 2.0 (financialized economy) to 3.0 (austerity agenda), the rationale first proposed by Hayek – that the economy is too complex for government regulation

and best left to an open, unfettered market with minimal state interference – has deepened and proved resilient even in the face of its failures. The present fiscal contraction of austerity is due, in no small measure, to decades of advice to developing countries by the international financial institutions (the IMF and WB) to keep taxes low to attract foreign investment, and tax cuts, notably among the high-income Anglo-American liberal economies and defended in the name of debt reduction and international competitiveness, that saw substantial declines the 1980s in marginal tax rates (paid by highest income earners) and corporate taxes, and increases in regressive consumption taxes. There would be no fiscal crisis and no need for austerity in most of the world's countries if progressive taxation had been retained. Even in those countries facing structural deficits, such as Greece, or reliant on aid transfers, such as many sub-Saharan African countries, liberalization of capital markets and global banking integration (including branches in all the world's tax havens) continue to allow companies and wealthy elites to avoid taxation almost altogether. The African continent has lost more wealth over the past 40 years in illicit capital flight, partly due to criminality and corruption but most of it involving commercial tax evasion, than it has received in foreign aid (African Development Bank, 2013).

The two international institutions most implicated in the diffusion of neoliberal economics, the WB and the IMF, have begun to accept the empirical evidence of its shortcomings, calling for greater government caution in implementing austerity measures, partly in recognition of the 'fiscal multiplier' effect of government spending. In times of recession, public spending yields far more economic return to the 'real economy' of production and consumption (through purchasing and employment) than do tax cuts intended to spur private sector investments. Conservative estimates of this fiscal multiplier find that every dollar in public spending generates between 1.6 and 1.7 dollars in economic growth (Stanford, 2013), with recent estimates of European public spending by sector showing much greater multiplier impacts for public investments in health, education, and environmental protection (Reeves et al., 2013). The WB and IMF are also now cautioning against 'extreme austerity' and calling for higher taxation rates, especially for 'ultrahigh-net-worth individuals.' In 2012, the world's 1426 billionaires between them had as much wealth (USD 5.4 trillion) (Forbes, 2013) as the entire continent of Africa (USD 2.3 trillion) and India (USD 3.2 trillion) (Shorrocks et al., 2012). Given Africa's and India's 2011 combined population of 2.2 billion, this represents an inequality ratio of roughly 1 500 000:1.

Although this surge in private wealth has led to new philanthropies, some of them focused on global health issues, these are driven by individuals and their own ideas about how their wealth should be distributed, rather than through tax-funded, publicly representative, and accountable processes. A 2013 US review of econometric studies found that raising marginal tax rates from their present historic low of 35% to their previous high of 68% would have no statistically significant impact on factors driving economic growth, but would reduce poverty, inequality, and stimulate growth through public spending (Fieldhouse, 2013). Similar findings exist for imposition of global taxation measures, such as a financial transaction tax that, by one study's estimate and if applied at a very low rate (.05%, or 5 cents on every 100 dollars) on every currency transaction, would generate over US $8 trillion in annual revenues (McCulloch and Pacillo, 2011). Globalized capital calls for global systems of taxation, toward which some of the world's governments are now slowly turning their attention.

Simply stated, we do not have a problem of excess public debt. We have a problem of inadequate private taxation. We do not have a crisis of scarcity. We have a crisis of extreme inequality. These conditions and the health risks they pose are not 'natural,' but the result of policy choices for which there are alternatives.

Global Governance for Health

There is nothing inherently unhealthy in globalization as first defined: the growing interdependence of peoples across the world, through increased economic integration, cultural diffusion, and travel. It is the particular economic form of this interdependence – the 'neo'liberalization of capital from the checks and balances that once existed when capital was confined within national borders – that is problematic, not only for public health. It has been argued for some time that the crisis facing contemporary globalization is that national governments and powerful elites have created binding and enforceable rules for a 'free' global economy without a global governance system capable of regulating that economy for a social purpose. That may be slowly changing.

Putting More Health in Trade

There is some evidence that WTO panels are increasingly accepting the importance and legitimacy of public health arguments in trade disputes, albeit only if there is no obvious trade discrimination involved. A 2010 case involving a tobacco dispute between Indonesia and the United States is illustrative. To comply with domestic legislation to prevent adolescent smoking, the United States banned imports of flavored cigarettes. Indonesia argued that the domestic legislation, by exempting 'menthol' from the list of flavors, discriminated against its clove cigarettes in favor of US-manufactured menthol cigarettes, which are the favored choice of American teens. This constituted a violation of WTO rules on national treatment (there should be no discrimination against 'like' imported goods). Indonesia further argued that there were 'less-trade-restrictive' ways (another WTO requirement) to meet the public health goal of reducing adolescent smoking than a ban on clove cigarette imports. The WTO dispute panel ruled with Indonesia on the first argument (nondiscrimination) but with the United States on the second (agreeing that a ban was a necessary public health policy). Provided there is no obvious discrimination, public health regulations appear to have growing latitude within (at least some) WTO (McGrady, 2012). There is also greater public health attention being given to proposed trade and investment treaties, with efforts to ensure that before such treaties are finalized, public health concerns are given due consideration, and that 'policy space' for future public health regulation is assured, e.g., regarding trade or investment in food, alcohol, or other health-damaging commodities. The potential for a healthier governance of

trade and investment treaties, however, remains constrained by the secrecy in which the treaties are negotiated, the privileging of business interests over public interests in informing government positions during negotiations, and the still-dominant assumption that liberalization is an inherently positive development for all countries involved. As argued, this is not a universally held or observed assumption.

Building Up the Social Protection Floor

In a different vein, the International Labor Organization is actively promoting a 'Social Protection Floor Initiative' to encourage countries to systematically build up their social protection systems. The argument for such an initiative is simple and compelling – social protection equitably redistributes wealth, opportunity, and health – and it is rooted in globally shared principles of social justice and human rights. It is supported by a large number of UN agencies, development banks, and government development agencies. There are glimmers of its implementation in different parts of the world: extended health coverage (now embraced by the WHO and the WB in the call for 'universal health coverage'), social cash transfers, and public sector and rural employment guarantee schemes. The Initiative highlights the important role that a strong and strongly tax-supported social protection system can play in times of economic crises. Its implementation, however, rests on countries' willingness to increase rather than decrease progressive taxation (a race to the top, rather than to the bottom), and, in the case of many sub-Saharan African countries, an increase in the very low level of royalties or 'rents' charged to transnational extractive industries, and imposed during neoliberalism 1.0 as structural adjustment conditionalities (OSISA, 2009). Finally, while the Social Protection Initiative offers countries good technical advice, it does not and cannot address the political will needed to implement such schemes.

Moving Health Higher Up the Foreign Policy Agenda

Efforts to create that political will can be found on two fronts: governmental initiatives to raise health higher in the foreign policy positions and negotiations of states, and the international campaigning of civil society organizations. Regarding the first, there have been several recent initiatives to improve foreign policy coherence for health. The Foreign Policy and Global Health initiative launched in 2006 by the foreign ministers of Brazil, France, Indonesia, Norway, Senegal, South Africa, and Thailand and renewed in 2011, as one example, encourages nations to broaden the scope of national foreign policies to integrate health concerns in a sustainable way and not as a crisis driven 'one-off' (Ministers of Foreign Affairs, 2007). Resolution 63/33 'Global health and foreign policy' adopted by the UN General Assembly in November 2008 further recognizes the "close relationship between foreign policy and global health and their interdependence" and urges "member States to consider health in the formulation of foreign policy" (United Nations General Assembly, 2009). Several nations have pursued this aim through the development of coherent strategies (Labonté and Gagnon, 2010) or efforts to create platforms for a 'health-in-all-policies' approach to their foreign affairs. Despite its recent global prominence, however, health is still not typically prioritized in the foreign policy agenda of most countries, and when it is, it is driven by traditional 'high politics' of security or economic risks due to pandemics or the spread of drug-resistant diseases. The continuing funding crisis of the WHO, and its reliance on extrabudgetary (nontithed) funding from governments and private sector donors for much of its operational budget, weakens its ability to provide leadership on global health governance initiatives that may go against the short-term interests of powerful nations or corporate actors (Vaughan et al., 1996).

Supporting Civil Society Activism

Civil society activism, often aligned with organized labor movements, has been in the forefront of challenging the damaging social, environmental, and health impacts of neoliberal globalization. There have been some successes. The persistence of health activists in government and civil society moved tobacco control from indisputable evidence of smoking's harms to smoking restrictions, pricing measures, and other policies aimed at shifting cultural norms – always in conflict with tobacco transnational companies. South Africa's Treatment Action Campaign helped to break the impasse between new regimes of intellectual property rights and access to antiretroviral medicines, aligning with international groups mobilizing for a 'just' globalization and invoking international human rights treaties in their advocacy. A powerful network of 'peasant' organizations arose in Latin America in the 1990s, and has inspired similar mobilizations globally, linking together activism on agroecology, right livelihoods, food security, social protection, gender empowerment, indigenous rights, and environmental sustainability. In the same decade, a network of international civil society organizations defeated the Multilateral Agreement on Investments in 1998, an Organization for Economic Cooperation and Development (OECD) effort to further and more deeply liberalize cross-border capital flows than had been accomplished in WTO negotiations, although liberalized investment rules are increasingly being inserted into regional and bilateral trade treaties. Civil society activism demands persistence.

Conclusion: Accepting the Environmental Limits to Growth

While providing global health activists with a cautious optimism of the spirit, recent government and continued civil society activism in global health has yet to offer a sustained challenge to neoliberal globalization. Even if governments and activists succeed in curbing the predatory excesses of the financialized economy (neoliberalism 2.0), and exposing the inequitable and economically ineffective austerity agenda (neoliberalism 3.0), it will have to contend with ecological limits of returning to a pre-GFC 'real economy' of production and consumption. As the UK Commission on Sustainable Development noted in its 2009 study, "there is as yet no credible, socially just, ecologically sustainable scenario of continually growing incomes for a world of nine billion people" (Jackson, 2009). Indeed, the two sectors that are most driving climate change are energy production/consumption and transportation – the two activities central to our globalized economy.

Borrowing from an earlier study on the social dimensions of globalization, the Globalization Knowledge Network of the 2008 WHO Commission on Social Determinants of Health concluded that an equitable global health would only be secured under the 'three R's': systematic *resource* redistribution between countries and within regions and countries to enable poorer countries to meet human needs, effective supranational *regulation* to ensure that there is a social purpose in the global economy, and enforceable social *rights* that enable citizens and residents to seek legal redress (Labonté et al., 2007).

See also: Comparative Health Systems; Ecology and Health; Environmental Stress and Health; Migration and Health; Social Epidemiology; Social Exclusion, Social Deprivation and Health; Socioeconomic Status and Health; Urban Life and Health.

Bibliography

African Development Bank and Global Financial Integrity, 2013. Illicit Financial Flows and the Problem of Net Resource Transfers from Africa: 1980–2009 [pdf]. http://www.gfintegrity.org/storage/gfip/documents/reports/AfricaNetResources/gfi_afdb_iffs_and_the_problem_of_net_resource_transfers_from_africa_1980-2009-web.pdf.

Altvater, E., Mankopf, B., 2012. European integration and the crossroads: deepening or disintegration? In: Pons-Vignon, N., Ncube, P. (Eds.), Confronting Finance. International Labour Organization, Geneva.

Baunsgaard, T., Keen, M., 2010. Tax revenue and (or?) trade liberalization. Journal of Public Economics 94 (9–10), 563–577.

Baxter, J., 30 March 2011. The glossary of greed. Pambazuka News [Online]. http://www.pambazuka.org/en/category/features/72112.

Bech, M., 2012. FX volume during the financial crisis and now. BIS Quarterly Review [pdf]. http://www.bis.org/publ/qtrpdf/r_qt1203f.pdf.

Birdsall, N., 2006. The World Is Not Flat: Inequality and Injustice in Our Global Economy. World Institute for Development Economics, Helsinki.

Breman, A., Shelton, C., 2001. Structural Adjustment and Health: A Literature Review of the Debate, its Role-players and Presented Empirical Evidence [Working Paper WG6:6]. WHO, Commission on Macroeconomics and Health, Geneva.

Chang, H.J., 2005. Why Developing Countries Need Tariffs: How WTO NAMA Negotiations Could Deny Developing Countries' Right to a Future. South Centre, Geneva.

Chapman, A., 2002. Core obligations related to the right to health. In: Chapman, A., Russell, S. (Eds.), Core Obligations: Building a Framework for Economic, Social and Cultural Rights. Intersentia, Antwerp, pp. 185–215.

Chen, S., Ravallion, M., 2008. The Developing World Is Poorer than We Thought, but No Less Successful in the Fight against Poverty [Policy Research Working Paper 4703]. World Bank. http://ideas.repec.org/p/wbk/wbrwps/4703.html#download.

CMH (Commission on Macroeconomics and Health), 2001. Macroeconomics and Health: Investing in Health for Economic Development. World Health Organization, Geneva.

Cornia, G.A., 2001. Globalization and health: results and options. Bulletin of the World Health Organization 79 (9), 834–841.

Cornia, G.A., Jolly, R., Stewart, F. (Eds.), 1987. Adjustment with a Human Face: Protecting the Vulnerable and Promoting Growth, vol. 1. Oxford University Press, Oxford.

Deaton, A., 2003. Health, inequality, and economic development. Journal of Economic Literature 41, 113–158.

Fieldhouse, A., 2013. A Review of the Economic Research on the Effects of Raising Ordinary Income Tax Rates [Online]. Economic Policy Institute, Washington, DC. http://www.epi.org/publication/raising-income-taxes/.

Forbes, 2013. Inside the billionaires list: Facts and figures. Forbes [Online]. http://www.forbes.com/sites/luisakroll/2013/03/04/inside-the-2013-billionaires-list-facts-and-figures/.

ILO (International Labour Organization), 2013. Global Employment Trends 2013: Recovering from a Second Jobs Dip [pdf]. http://www.ilo.org/wcmsp5/groups/public/–-dgreports/–-dcomm/–-publ/documents/publication/wcms_202215.pdf.

Jackson, T., 2009. Prosperity without Growth? the Transition to a Sustainable Economy. Sustainable Development Commission, London.

Koivusalo, M., Schrecker, T., Labonté, R., 2009. Globalization and policy space for health and social determinants of health. In: Labonté, R., Schrecker, T., Packer, C., Runnels, V. (Eds.), Globalization and Health: Pathways, Evidence and Policy. Routledge, New York (chapter 5).

Labonté, R., Blouin, C., Chopra, M., Lee, K., Packer, C., Rowson, R., Schrecker, T., Woodward, D., 2007. Towards Health-Equitable Globalization: Rights, Regulation and Redistribution [pdf] Globalization Knowledge Network Final Report. http://www.who.int/social_determinants/resources/gkn_report_06_2007.pdf.

Labonté, R., Gagnon, M.L., 2010. Framing health and foreign policy: lessons for global health diplomacy. Globalization and Health 6 (14), 1–19.

Labonté, R., Mohindra, K., Lencucha, R., 2011. Framing international trade and chronic disease. Globalization and Health 21 (3), 273–287.

Lee, K., 2002. Health Impacts of Globalization: Towards Global Governance. Palgrave Macmillan, London.

Martinez, M.E., 2004. Women in the maquiladora industry: toward understanding gender and regional dynamics in Mexico. In: Kopinak, K. (Ed.), The Social Costs of Industrial Growth in Northern Mexico. Centre for US-Mexican Studies, La Jolla, CA.

McCulloch, N., Pacillo, G., 2011. The Tobin Tax – a Review of the Evidence. Institute of Development Studies. University of Sussex, Brighton.

McGrady, B., 2012. Tobacco Product Regulation and the WTO: Appellate Body Report, US – Clove Cigarettes [pdf]. O'Neill Institute for National and Global Health Law, Washington, DC. http://www.oneillinstitutetradeblog.org/wp-content/uploads/2012/04/briefing-paper.pdf.

Ministers of Foreign Affairs of Brazil, France, Indonesia, Norway, Senegal, South Africa and Thailand, 2007. Oslo ministerial declaration – global health: a pressing foreign policy issue of our time. The Lancet 369 (9570), 1373–1378.

Ortiz, I., Cummins, M. (Eds.), 2012. A Recovery for All: Rethinking Socioeconomic Policies for Children and Poor Households. UNICEF, New York.

Ortiz, I., Cummins, M., 2013. The Age of Austerity: A Review of Public Expenditures and Adjustment Measures in 181 Countries. UNICEF – Initiative for Policy Dialogue, New York.

OSISA (Open Society Institute of Southern Africa), Third World Network Africa, Tax Justice Network Africa, Action Aid International, Christian Aid, 2009. Breaking the Curse: How Transparent Taxation and Fair Taxes Can Turn Africa's Mineral Wealth into Development.

OUNHCHR (Office of the United Nations High Commissioner for Human Rights), 2000. Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children, Supplementing the United Nations Convention against Transnational Organized Crime. United Nations, Geneva.

Polaski, S., 2006. Winners and Losers: Impact of the Doha Round on Developing Countries. Carnegie Endowment, Washington, DC.

Popkin, B.M., 2009. Global changes in diet and activity patterns as drivers of the nutrition transition. In: Kalhan, S.C., Prentice, A.M., Yajnik, C.S. (Eds.), Emerging Societies – Coexistence of Childhood Malnutrition and Obesity. Karger, Switzerland.

Reeves, A., Basu, S., Mckee, M., Meissner, C., Stuckler, D., 2013. Does investment in the health sector promote or inhibit economic growth? Globalization and Health 9 (43).

Schrecker, T., Labonté, R., 2011. Global development and population health. In: Nriagu, J. (Ed.), Encyclopedia of Environmental Health. Elsevier, Amsterdam, pp. 976–986.

Shorrocks, J., Davies, J., Lluberas, R., 2012. Global Wealth Report 2012. Credit Suisse AG Research Institute, Zurich.

Standing, G., 2011. The Precariat: The New Dangerous Class. Bloomsbury, Cornwall.

Stanford, J., 2013. The failure of corporate tax cuts to stimulate business investment. In: Swift, R. (Ed.), The Great Revenue Robbery. Canadians for Tax Fairness, Ottawa.

Stuckler, D., Basu, S., 2013. The Body Economic: Why Austerity Kills. Basic Books, New York.

Thorbecke, E., Nissanke, M., 2006. Introduction: the impact of globalization on the world's poor. World Development 34 (8), 1333–1337.

United Nations General Assembly, 2009. Resolution Adopted by the General Assembly: 63/33 Global Health and Foreign Policy [pdf]. http://www.who.int/trade/events/UNGA_RESOLUTION_GHFP_63_33.pdf.

Vaughan, J.P., Mogedal, S., Walt, S., Kruse, S.E., Lee, K., Wilde, K.D., 1996. WHO and the effects of extrabudgetary funds: is the organization donor driven? Health Policy and Planning 11 (3), 253–264.

Wilkinson, R., Pickett, K., 2009. The Spirit Level: Why More Equal Societies Almost Always Do Better. Allen Lane, London.

Woodall, P., 16 September 2006. The new titans: a survey of the world economy. The Economist, vol. I.

Relevant Websites

http://www.globalhealthequity.ca/ – Globalization and Health Equity.
http://www.globalizationandhealth.com/ – Globalization and Health Journal.
http://www.ghwatch.org/ – Global Health Watch.
http://www.phmovement.org/ – Peoples Health Movement.
http://www.worldbank.org/en/topic/health – World Bank Health.
http://apps.who.int/globalatlas/ – World Health Organization Global Health Atlas.
http://www.who.int/gho/en/ – World Health Organization Global Health Repository.

Globalization and Human Trafficking

Mónica Méndez, House of Mentoring & Empowerment, Inc., Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since the 1990s, globalization, driven by economic interests and improved communications, technology, and transportation, has resulted in a more interconnected environment that allows human trafficking to thrive, while helping to combat it at the same time. The unequal social and economic conditions around the world perpetuate the idea that some humans are commodities that can be sold and traded in world markets. The lucrative nature of human trafficking has resulted in the growing demand for humans from countries of origin where the supply of people is both high and the transactions quick and low risk as a result of technology, advanced communication, and faster transportation. The demand for people comes from within the supply countries and from both developing and developed countries as the need for labor and the sexualization of primarily women and children continue to breed inhumane uses of humans.

This article defines globalization and human trafficking and then shows how the former has perpetuated and assisted in the evolution of the trafficking of people. At the same time, globalization has supported governments, nongovernmental organizations (NGOs), and individuals in combating human trafficking.

Globalization

As world economies, transportation, and communications expand, the world becomes increasingly integrated and interdependent. Commonly referred to as 'globalization,' this interconnectedness includes everything from the integration of governments, economies, transportation, and business to the melding of cultures, laws, and social movements. Beginning in the late nineteenth century, globalization intensified as a result of industrial, communications, technological, economic, and transportation advances (Wolf, 2001). It has played a key role in the growth of national economies and international businesses, both legitimate and illegitimate, primarily through the continuously emerging science, technologies, and communication mediums that facilitate the free flow of people, goods, services, and ideas (Shangquan, 2000; Stohl, 2008). The concept of globalization is complex, but it may be broken down into economic, cultural, and political components to better understand it.

Economic globalization refers to the connectedness of economic activity, such as cross-border trade and the spread of technologies, that creates global interdependence in markets, capital, and commodities (Shangquan, 2000). Olteanu and Radoi (2010) describe it as the internationalization of economies where liberalizing markets have created intense competition for control and location of the labor markets. Economic globalization can result in people, who have been displaced by agricultural modernization, needing to seek other lines of work in their own and other countries, and it can create corruption and violence, while destroying social norms, through rapid economic changes (Bales, 2004). Cultural globalization refers to the diffusion of ideas, values, and ways of life, as well as the creation of a global conscience, through technology, communication mediums, and transportation (Moghadam, 2005). It implies that the almost instantaneous exchange of ideas creates more connectivity, interaction, and interdependence among people of different cultures and can refer to the heavy influence of Western cultures on other societies. Finally, political globalization focuses on multilateralism and the increased influence of public policy across borders, affecting governing doctrines that deal with human rights, energy, the environment, and political governmental models, such as democracy (Moghadam, 2005; Zaleski, 2006). The connectivity that exists as a result of the tools of globalization (i.e., science, technology, and communication mediums) breaks through the time and space limitations that, before globalization, created barriers to the exchange of capital and commodities. Economic markets, businesses, intra- and intercontinental communications, and transportation are now more efficient and more accessible to people and regions who previously lacked such access.

Globalization's effects are vast and are both positive and negative, especially considering that different social groups are affected in diverse manners (Lee, 2011). From the positive perspective, global communication allows people to connect to one another more efficiently, making business, politics, and social exchanges more immediate and profitable by saving both time and money. The power to connect people and its potential to spread wealth are also seen as a means of providing opportunities to individuals who, because of location, language barriers, and/or economic situations, have little to no access to education and employment (Mohr, 2012). In addition, due to globalization more countries, individuals, and businesses have access to financial lending, allowing for improved social and civil infrastructures, a higher standard of living and access to markets that were not open before. Finally, people are exposed to other cultures and ideas, and they are more aware of world social problems and solutions.

Other perspectives argue that globalization increases social inequalities, specifically the gap in income distribution, and it allows for the continued exploitation of people, both in and out of the labor market, who are not as connected or protected as others (International Labour Organization, 2008; Mohr, 2012; World Commission on Social Dimension of Globalization, 2004). The interests of high-income countries drive the interests of the rest of the world, leaving low-income countries to bear the brunt of the negative economic, labor,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64131-4

environmental, and social consequences of globalization. The latter include the perpetuation of poverty, social inequalities, the oppression of women and other minorities, and the lack of afforded protections to these groups. Marginalized people – those who are impoverished, discriminated against, homeless, incarcerated, and otherwise socially and economically excluded – are particularly negatively affected and victimized by globalization (Hepburn and Simon, 2013; Polakoff, 2007).

Although globalization has improved many aspects of contemporary life around the world, it has also had a far-reaching and negative effect on one of the oldest and most egregious human rights offenses: human trafficking. The negative effects of globalization are in many ways embodied by the sexual exploitation of youth and adults. Interconnectedness has made it easy for sexual exploiters to connect to and sell to buyers of infants, children, youth, and adults in a sex industry that highlights the intersectionality of race, gender, and economics. The victims and survivors of these crimes – especially women and children – often do not have the means to fight back due to poverty and the lack of other protections, including a lack of power, opportunity, and a growing inequality gap (Farr, 2005; United Nations Office of Drugs and Crime, 2012).

Human Trafficking

Human trafficking has been defined by numerous governmental entities in an attempt to clarify what it looks like, who is involved and how it is happening. Although there are variations of the definition and debate about what should and should not be included (Lee, 2011), a widely accepted definition found in the Protocol to Prevent, Suppress and Punish Trafficking in Persons especially Women and Children (hereafter the Trafficking Protocol) (2000) defines human trafficking as the

> recruitment, transportation, transfer, harbouring or receipt of persons, by means of the threat or use of force or other forms of coercion, of abduction, of fraud, of deception, of the abuse of power or of a position of vulnerability or of the giving or receiving of payments or benefits to achieve the consent of a person having control over another person, for the purpose of exploitation. Exploitation shall include, at a minimum, the exploitation of the prostitution of others or other forms of sexual exploitation, forced labour or services, slavery or practices similar to slavery, servitude or the removal of organs…(p. 42)

There is a clear distinction between smuggling, which is a crime against state borders, and human trafficking, which is a crime against a person. The definitions of human trafficking that are widely used have made efforts to emphasize that the important aspect of trafficking is the crime against the person (Nagle, 2008).

Human beings have been seen as commodities since before globalization emerged, and the tools of globalization have made it easier for human trafficking to flourish in markets across the world. The increased ability to exchange money, communicate and sell through the Internet, and send commodities and people across borders faster than ever has made the trafficking of humans the third most widespread business for organized crime in the world, right behind drugs and arms dealing (Nagle, 2008). Since the beginning of the twenty-first century, more international attention has been given to human trafficking. Yet it is important to remember that attempts to combat it through antislavery movements and laws have been continuous, from the Universal Declaration of Human Rights in 1948 to more recent resolutions and legislation encompassing the rights of women and children and others. (see the 1956 Supplementary Convention on the Abolition of Slavery, the Slave Trade, and Institutions and Practices Similar to Slavery; the 1948 American Declaration on the Rights and Duties of Man; the 1957 Abolition of Forced Labour Convention; the International Covenant on Civil and Political Rights; the Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment; United Nations General Assembly Resolutions on Traffic in Women and Girls; the 1996 Final Report of the World Congress against Sexual Exploitation of Children; the 1995 Fourth World Conference on Women in Beijing; and the 1991 Moscow Document of the Organization for Security and Cooperation in Europe).

As a result of the transnational component of human trafficking, increased awareness of the issue and the increasing interconnectedness of countries, the fight against human trafficking has been progressively seen as a global responsibility and a gross violation of human rights, which is a contrast to what it was seen as in the 1980s as a 'women's issue' that was given little attention and funding (Lee, 2011). As such, international entities have dedicated conventions and protocols to the issue of trafficking. In 1998, at the UN Commission for Crime Prevention and Criminal Justice, the United States presented a resolution on the trafficking of women and children that called for the creation of protocols under the United Nations Convention against Transnational Organized Crime (Hyland, 2001). As a result, the United States and Argentina drafted the Trafficking Protocol in 1999; it became the first and main comprehensive international instrument to combat human trafficking. The 2000 final version of the Trafficking Protocol, first put into effect in 2003, was the first enforceable instrument to provide a universal definition of human trafficking, making it easier for countries to work internationally, and was the first to provide tough laws and protections for survivors of trafficking.

As defined by the International Labour Organization (2012), the U.S. State Department (2012), and the United Nations General Assembly (2000), human trafficking takes on many forms in both labor and sex trafficking. For instance, labor trafficking encompasses forcing, deceiving, coercing, manipulating, and/or threatening individuals into working on fishing boats, in manufacturing factories, private homes, agriculture, construction, massage parlors, and/or other businesses such as nail salons, restaurants, call centers, and hotels. Individuals may cross no borders, or travel through national or continental boundaries under the guise of working in legitimate businesses but then find themselves trapped. Traffickers hold victims and survivors without justification, under threat of harm to themselves or their families, or under the guise that the survivor owes the trafficker money for transportation, housing, and any food used during the working period. Traffickers may

also operate as legitimate businesses that take control of indigenous land or local farmers' lands and force the people to work on the land for little to no pay in order to buy back the land, knowing full well that the interest included in the price of the land accumulates faster than workers can accumulate pay.

Another means of labor trafficking, and one of the most prolific in the world, occurs through debt bondage, which is when an individual is born into the debt of a family, acquires a loan that is unpayable, or is deceived into traveling for work and then owes the money for the travel (Siddharth, 2009; United Nations, 1956). As time passes, the victims and survivors may be unable to pay back the money, and debt accumulates everyday that passes. Debt bondage has existed throughout history and continues to exist today primarily in developing countries, including South Asian countries, where caste systems enslave families generation after generation, and in South America and Africa, where banking systems, legitimate loan and credit sources are often scarce. However, debt bondage can exist in developed countries in the form of travel debt, drug debt, and domestic servitude. For example, sexual exploitation can occur when an individual is forced to sell her or his body for sex to pay for drugs she or he has used. Debt bondage is one of the most common forms of forced labor and has been outlawed internationally, but, like many other forms of trafficking, it persists as a result of inadequate social, political, and economic safety nets available to victims and survivors.

Forcing children and handicapped people to beg on streets is also a common form of labor trafficking (Cherneva, 2011). In some instances, children who might be healthy are purposely maimed to seem more sympathetic; for example, some homeless or poor children in South Asia are purposefully blinded for this reason. Similar to begging, a traveling 'sales crew' is a relatively new human trafficking method, being tracked in some countries, including in the United States. In these situations, teenagers and young adults are manipulated into working for little to no money by traffickers who transport them to varying locations to have them sell magazines, candy, or other trinkets under the guise of working for a nonprofit or a scholarship. Victims work long hours and are charged for any water or food they use, leaving them close to nothing in pay. There are few repercussions for the traffickers and few resources for the survivors as a result of social inequalities and insufficient laws or understanding surrounding of the issues.

Sexual exploitation is another type of human trafficking that is primarily, but not exclusively, been seen as affecting women and children (United Nations Office on Drugs and Crime, 2006, 2009, 2012; International Labour Organization, 2012). According to the ILO, sexual exploitation is a form of labor trafficking that can include infants, children, teenagers, and adults being sold into the pornography industry, brothels, hotels, massage parlors, strip clubs, the street, and homes around the world. Sexual exploitation occurs differently across the world, but the lack of opportunities for families and individuals, as well as gender and social inequalities, feeds the need to find some form of livelihood. In societies where women and girls are seen as economic weights on their families, girls and teens are sometimes sold to brothels as a means of disposing of them or to make them support their families. In the United States, domestic trafficking is aided by the hypersexualization of girls, the sexual abuse of girls, a lack of opportunity to escape abuse and poverty, and a system that is not yet ready to assist these survivors. Social factors, such as community rejections of gay, lesbian, bisexual, or transgender individuals, also make it easier to exploit individuals in the sex industry. Combined with poverty, these factors make individuals in all countries susceptible to the cons and traps of traffickers.

Adoptions, forced marriages, military service, and organ trafficking are other forms of human trafficking that are not necessarily included in some definitions of trafficking, including the International Labour Organization's (2012) definition. Adoptions can turn into human trafficking when children are abducted from their families, families are deceived or coerced into giving up their children for adoption, or families sell the children to adoption agencies in order to survive. As a result of war, civil conflicts, natural disasters (e.g., Haiti's 2010 earthquake), and other events, the practice of selling children, especially via international adoptions, is not easily measurable or preventable.

Forced marriages of women and girls are another form of human trafficking that defines a situation where one of the spouses, usually a woman or girl, does not have the ability to refuse marriage (United Nations, 1956). Women and girls as young as 8 years of age can be forced to marry men in exchange for dowries or because the girls and women are seen as less valuable than sons. This practice is common in the Middle East, South Asia, and Africa (Bokhari, 2008). Mail-order brides, common in areas of Russia, Eastern Europe, and South Asian countries, are also considered forced marriage because the women are deceived or forced into making themselves available to the highest bidder as a result of organized crime and/or poverty.

Forced military service of boys, men, women, and girls is another type of trafficking that may occur out of threat, force, coercion, deception, and manipulation. It has been occurring since World War II, but the social disruption caused by globalization, prolonged armed conflict, and the spread of diseases has been seen as causes for the increase in the use of child soldiers (Tiefenbrun, 2007). Children as young as 8 years of age have been identified as child soldiers, and the girls and boys who are forced into military services also often serve as sex slaves for the troops (United Nations Office for the Coordination of Humanitarian Affairs, 2007). Armed conflict, illness, and poverty may leave orphaned children who are then forced to serve in military organizations supported by governments, paramilitary organizations, or rebel groups.

Finally, organ trafficking must be included in this sometimes-excluded group of trafficked individuals. The trafficking of organs may occur with or without the consent of the victim, and, when one considers that consent may mean a poor individual from a developing country that sees no way of surviving than through the sale of her or his organs, consent is nothing more than exploitation by those involved in the globalization of organ transplants (Scheper-Hughes, 2000; Budiani-Saberi and Delmonico, 2008). Deception in organ trafficking could occur when a person believes she or he will be paid for the organ and is not paid or is paid less than the expected price, or when the person does not realize that she or he agreed to give up an organ. Individuals from poor developing countries, migrant workers,

homeless individuals, and illiterate individuals are some of the target victims of this type of trafficking (Budiani-Saberi and Delmonico, 2008). Everyone from the recruiter of the victim to the medical professionals who remove the organs to the individuals who buy for the organ banks is involved in the illegal trafficking of organs; however, due to a lack of systemic data and laws, the scope of the problem is not well known or understood.

Because people are seen as durable commodities that can be sold, exchanged, and/or traded in world markets for very high prices, human trafficking is extremely lucrative. For example, a drug or weapon sale provides a one-time profit. However, humans can be used and sold as many times as the individuals can withstand, with continuous gains going to the traffickers. According to the ILO, some estimates of human trafficking profits exceed US$31 billion a year (2005).

Estimates of Human Trafficking

Human trafficking takes on different forms around the world and estimates vary as to how many people are affected, what countries they originate in, and to what countries they go. In fact, there is much criticism about the methodologies used to gather estimates and the poor data propelling the movement to end trafficking (Sanghera, 2005; Lee, 2011). Estimates on the condition of human trafficking vary as a result of its different definitions that come from varying interests and stakeholders, different methodologies for gathering and analyzing data, a lack of systematic data gathering, varying access to institutional, country, and organizational data, and the use of different measures. In addition, some laws around the world may only apply to and define certain types of trafficking, types of victims and survivors (e.g., women and children but not men), and enforcement tools, making it difficult to obtain any data or any reliable data (United Nations Office of Drugs and Crime, 2006, 2009). Finally, as a result of a lack of protection for the victims and survivors of trafficking, as well as internal conflicts and discrimination, survivors of trafficking and their allies may not be able to report the trafficking out of fear for themselves and their families. In countries where governments do not have a history of victim services and comprehensive laws, the barriers to reporting are very real and can result in low levels of reporting. However, it is still important to outline what different entities around the world estimate when it comes to the trafficking of humans.

In 2006, the United Nations Office on Drugs and Crime provided its first Trafficking in Persons Global Report detailing trends of human trafficking in 127 countries of origin to 137 destination countries from 1996 to 2006 (United Nations Office on Drugs and Crime, 2006). The report outlined three basic distinctions among countries that are not mutually exclusive when it comes to trafficking in humans: countries of origin (from where the victims and survivors come), transit countries (countries through which transportation of victims and survivors is common), and destination countries. According to the findings, Albania, Belarus, Bulgaria, China, Lithuania, Nigeria, Moldova, Romania, Russia, Thailand, and Ukraine have the highest incidences of human trafficking when it comes to origins of victims and survivors. Albania, Bulgaria, Hungary, Italy, Poland, and Thailand are the most used transit countries, and victims and survivors of trafficking are most likely headed to Belgium, Germany, Greece, Israel, Italy, Japan, Netherlands, Thailand, Turkey, and the United States of America.

The ILO's most recent publication on human trafficking (2012) is based on a study conducted from 2002 to 2011 on the estimate of people forced into labor in the world. It estimated, conservatively, that there were about 20.9 million people around the world who were victims and survivors of forced labor, including sexual exploitation. It is a conservative estimate because human trafficking is a hidden crime not easily measurable and because it does not include trafficking for the removal of organs or numbers on forced marriage or adoption unless the practices lead to forced labor situations. The report further divides the 20.9 million into what sectors exploit them: 18.7 million by the private sector, encompassing 4.5 million in the sex industry and 14.2 million in agriculture, construction, domestic work, and manufacturing. Governments, through state-imposed forms of forced labor, exploit the remaining 2.2 million. This study is an example of how definitions of exploitation include and exclude different types of human trafficking; it is meant to include labor practices only, ignoring some important forms of human trafficking while including what other researchers may not consider forced labor by other definitions (i.e., state-imposed labor through prisons).

In contrast, the Walking Free Foundation conducted its own research on the prevalence of what they term modern day slavery that includes slavery, slavery-like practices (e.g., forced marriage, debt bondage, children forced into armed conflict), human trafficking, and forced labor (2013). The study relies on three definitions of modern day slavery provided by the 1956 Slavery Convention, the 1930 Forced Labour Convention, and the 2000 United Nations Trafficking Protocol. According to the Global Slavery Index report of 2013, there are an estimated 29.8 million people exploited through modern day slavery. The index provides estimates for 162 countries and showed that the highest prevalence of slavery (in order) was in Mauritania, Haiti, Pakistan, India, Nepal, Moldova, Benin, Cote D'Ivoire, The Gambia, and Gabon. When considering numbers in absolute terms, however, the countries with the highest number of slaves (in order) are India, China, Pakistan, Nigeria, Ethiopia, Russia, Thailand, the Democratic Republic of Congo, Myanmar, and Bangladesh. The latter group of countries, those with the highest number of slaves, accounts for up to 76% of the total estimated number of 29.8 million people enslaved worldwide.

These two reports are examples of often cited statistics on the current status of human trafficking, and, as is evident, they show how widely different the perspectives on who is a victim and survivor of human trafficking and what types of trafficking are most prevalent and where.

The Victims and Survivors

The victims and survivors of human trafficking across the world come from varying backgrounds and are sought for different types of exploitation. As previously mentioned, several factors make trafficking statistics and survivor identification difficult. What reports have found across the board is that human

trafficking appears to be intraregional, meaning that the traffickers are usually trafficking people of their same country or regional area (United Nations Office on Drugs and Crime, 2006, 2009; International Labour Organization, 2012).

When it comes to transnational human trafficking, the data on the type of exploitation and its estimate vary (United Nations Office on Drugs and Crime, 2009; International Labour Organization, 2012). As a result of most of the laws around the world focusing on child victims and victims and survivors of sexual exploitation, the majority of victims and survivors of trafficking who were identified in 2006 in 61 countries were 66% women and 13% girls, with men and boys only being identified in 12 and 9% of the cases, respectively (United Nations Office on Drugs and Crime, 2009). There were only slight changes in the 2012 United Nations report that indicated that 59, 17, 14, and 10% of trafficked survivors were women, girls, men, and boys, respectively. When the type of exploitation was identified, statistics showed that 79% of the victims and survivors were sexually exploited, a contrast to the 2012 ILO statistic showing that about 24% of its cases concerned sexual exploitation. However, the exploitation of men and boys may not be as visible or measurable as a result of their lack of inclusion in laws. The United Nations, ILO, Amnesty International, and others have recorded quantitative and qualitative data on the trafficking of men and boys in the shipping, fishing, agriculture, and manufacturing businesses, as well as soldiers in wars, domestic servitude, and sex industries.

The effects of labor and sex trafficking on individuals trafficked tend to be devastatingly negative (Zimmerman et al., 2011; U.S. Department of State, 2012). The physical, psychological, verbal, and sexual violence and torture experienced by trafficked individuals cause irreparable, long-lasting physical and psychological damage to individuals (Hossain et al., 2010). Trauma, sexually transmitted infections, anxiety, depression, post-traumatic stress disorder, alcohol and substance abuse, unwanted pregnancies, malnutrition, and social ostracism may all occur to one individual. For children, there is the added component of stunted physical, educational, and psychological growth, creating – if they survive – adults who have a hard time thriving in a healthy society. The victims, the survivors who die in the process of trafficking, suffer the ultimate effect of trafficking. While the international community has a hard time estimating the number of individuals being trafficked, it has a harder, even impossible, time estimating the number of individuals who die while being trafficked.

The Exploiters: Organized Crime, Pimps, Businesses, and Buyers

Researchers gathering data from criminal justice organizations around the world have shown that the international trafficker networks in the countries of origin are those controlling and driving the trafficking (United Nations Office on Drugs and Crime, 2009). Traffickers in the countries of origin are best able to gain the trust of the victims and manipulate the local crime networks, which sometimes include local law enforcement and politicians, to keep victims under control. However, in some parts of the world, like Eastern Europe, there are organized crime entities that consist of national and transnational groups that make the trafficking possible (Surtees, 2008).

Research on the continued existence of trafficking shows that traffickers must depend on a wide network of legal and illegal socioecological participants, including politicians, government workers, law enforcement, nonprofit organizations, and communities that survive on the trafficking and exploitation of humans (Surtees, 2008). For example, in a study of 91 Chicago pimps done by Raphael and Myers-Powell (2010), pimps mention that the only way that they can operate is through the bribing of police officers, cab drivers, hotel managers, and some politicians. As a result of increased globalization, these networks have been made stronger and faster than before the existence of our current communications and transportation tools.

For traffickers, keeping victims and survivors of trafficking under control is also a vital part of the process (Surtees, 2008). Mind games, physical harm, threats on family, threats of death, isolation, and debt are tools that traffickers use to keep survivors trapped (Bruce and Failla, 2004; Surtees, 2008). Trauma bond, or Stockholm syndrome, may also play a part in a survivor staying in the trafficking situation (National Colloquium, 2012). For example, girls who run away from home and are picked up and used by traffickers for sex exploitation may see the trafficker as a savior who has sheltered, fed, and provided them clothing and, most of all, provided them the love and attention that their family and community failed to provide. In these situations, the survivor is psychologically chained to the trafficker.

Globalization and Human Trafficking: Connectedly Thriving

The trafficking of persons is facilitated by social and economic conditions around the world that treat humans as lucrative and durable commodities that can be traded in markets. Globalization has enhanced the need for a cheap labor market for different sectors of society and, as governments, industries and people compete for footing in newly emerging industries and markets, groups of people find themselves exploited in a process that appears to leave human rights behind (Nagle, 2008). Poverty, underdevelopment, and a lack of opportunity or movement in social classes are but a few factors that add to the ability of traffickers to coerce, force, manipulate, and/or threaten people into trafficking (International Labour Organization, 2008; Mohr, 2012; United Nations General Assembly, 2000; World Commission on Social Dimension of Globalization, 2004). In fact, the displacement caused by wars, natural disasters, technological advances that replace humans, and the transient economies created by rapidly changing economies, force people who have no opportunities to take risks that result in labor or sex exploitation. In addition, trafficking thrives in countries that devalue and discriminate against women and children, providing few legal and social protections for their exploitation in both the labor and sexual commercial realms (United Nations General Assembly, 2000; Walk Free Foundation, 2013). When racial discrimination is recognized as an important variable, also affecting who is trafficked, the effect of globalization on trafficking is made much clearer.

The negative consequences that arise from the perpetuation of human trafficking through globalization includes the faster and more extensive spread of sexually transmitted diseases and infections, the tearing apart of families and communities through better and faster transportation, the breakdown of small, local economies, the further dehumanization and sexualization of women and girls and continued intergenerational violence and abuse experienced at a faster pace by those who are survivors of trafficking. These are a few of the negative results of human trafficking in general, but the increase and growth of globalization has made these results come in a faster and more interconnected manner. The labor force that emerges from inequalities around the world provides an almost free supply for the high demand of all kinds of labor, from sex work to agricultural work. The same tools that have been used to open markets and economies and spread ideas are used to exploit and manipulate individuals and communities who may not have access to these tools and are often oppressed within their regions.

Combating Trafficking in the Midst of Globalization

Although globalization has made it possible to interconnect countries, laws, economics, and societies, it has also facilitated trafficking itself (United Nations General Assembly, 2000). The efforts to decrease and eliminate the trafficking of humans, up to this point, have been fragmented by noncollaboration among countries and regions and diminished by faster and tech-savvy traffickers who do not have to stay within the boundaries of jurisdiction and laws to run their businesses (United Nations Office of Drugs and Crime, 2006). In addition, the trafficking efforts that have taken place thus far are driven by moral laws and agendas that have, at times, turned the tables on the victims and survivors trafficking, viewing them and their 'choices' as the problem instead of centering on the social problems that forced them into sex work, illegal migration, and exploitation (Lee, 2011; Weitzer, 2007).

According to the United Nations Global Report on Trafficking in Persons (2009), this millennium has seen an increase in the number of countries that have taken a stance against trafficking by implementing the Trafficking Protocol; 63% of the 155 countries on which it reported had, by 2008, passed laws against human trafficking that covered the major forms of trafficking defined in the protocol, while 16% had passed some laws that covered at least part of the definition components. With more than half of the 155 countries having established special trafficking police units and national action plans, the strategy for combating trafficking is evolving, especially when compared to a 2003 statistic, which tallied only 35% of countries as having specific laws on the trafficking of persons. Not all of the countries that have tough, comprehensive laws and successful prosecutions of traffickers are, however, high-income countries (United Nations Office on Drugs and Crime, 2009). In other words, national will, not wealth per se, is a driving force for combating trafficking.

In recent years, the fight against human trafficking has shifted into a crime control model that centers on the geopolitical, economic, security, and moral concerns of some of the countries controlling globalization. According to Lee (2011), the crime control antitrafficking movement has centered on five aspects: antiprostitution, securitization, criminalization, militarization, and privatization, all of which are directly connected to those who hold power in the current world of globalization. Antiprostitution and securitization have both directly and indirectly put the weight of trafficking on the victims and survivors; they are blamed for sexual exploitation and immigration issues that, to some groups, are at the root of trafficking. Criminalization takes into account the transnational crossing that globalization has broadened; however, criminalization, although referring to increased attention to organized crime, has also increased the immigration tensions from developing countries to developed countries (i.e., Latin Americans immigration to the United States). The latter three aspects of crime control, as defined by Lee (2011), continue to redefine victims and survivors of trafficking as 'worthy' and 'unworthy' and true victims vs criminals (Bosworth, 2007).

Globalization has also made the militarization and privatization of crime control in human trafficking more efficient. In an attempt to control the aspects of crime that are deemed to be affecting trafficking, countries, like the United States and the United Kingdom, have begun to militarize their borders so as to stop the intrusion of borders (Lee, 2011). Technology that includes advanced surveillance and communication methods has expanded countries' abilities to fight what they see as the underlying problems in human trafficking. However, since it can be difficult for governments themselves to provide the militarization and technological advances needed for combating trafficking, a new industry for private enforcement has emerged in the global markets (Bosworth, 2007). These management, technological, and policing organizations work transnationally in order to ensure that the criminalization, securitization, militarization are all taking place according to the definition of human trafficking accepted by the underlying power (Lee, 2011).

Although the methods of combating human trafficking have increasingly put the onus of the crime on the survivors and victims, there have also been increases in assistance to these victims and survivors. Victim services for survivors of human trafficking have been more widely available since the beginning of the twenty-first century, as the recognition of trafficking as a global, regional, and national problem has increased. Because there is no uniform source of victim services for survivors, assistance for them has been set up primarily independently at the international, national, and local levels (Méndez, 2010). International organizations and conventions have established funds that are being used globally to assist survivors of trafficking, covering basic shelter, food, and clothing needs as well as victim compensation, relocation, legal assistance, and immigration documentation. Méndez (2010) mentions that the improved technology, communications, and transportation currently in the hands of the traffickers makes it imperative to provide victim services that are transnational and interconnected so as to provide the best possible services to survivors in this era of increased globalization.

While globalization, through the use of advanced technology, communications mediums, and transportation, has allowed traffickers to more effectively exploit people around the world, it has also made combating trafficking easier for

governments and NGOs. The same tools that have abetted traffickers are now being used by governments and NGOs; however, in order to increase the effectiveness of these tools, governments and NGOs must use one more tactic that has made human trafficking more successful: international collaboration (Surtees, 2008). The formal, successful human trafficker organizations that control large segments of international trafficking have long skirted the lines of jurisdictions in order to better serve their bottom lines (United Nations Office of Drugs and Crime, 2006). Combining globalization tools and transnational collaboration will make it possible for governments and NGOs to prevent trafficking, protect survivors, and prosecute traffickers.

See also: Child Labor; Child Soldiers; Children, Rights of: Cultural Concerns; Crime and Gender; Gender and Feminist Studies in Geography; Global Migration; Human Rights and Social Work; Missing Girls: A Globalizing Issue; Rape and Sexual Coercion; Slaves and Slavery, History of; Trafficking of Women and Children in Latin America.

Bibliography

Bales, K., 2004. Disposable People: New Slavery in the Global Economy. University of California Press, Berkeley.

Bokhari, F., 2008. Child Trafficking for Forced Marriage. ECPAT UK Discussion Paper. http://www.ecpat.org.uk/content/discussion-papers.

Bosworth, M., 2007. Immigration detention in Britain. In: Lee, M. (Ed.), Human Trafficking. Willan, Cullompton, pp. 159–177.

Bruce, R., Failla, S., 2004. Human Trafficking/Prostitution Investigation Unit – Yearly Report/Assessment (Pristina, Kosovo).

Budiani-Saberi, D.A., Delmonico, F.L., 2008. Organ trafficking and transplant tourism: a commentary on the global realities. American Journal of Transplantation 5 (8), 925–929.

Cherneva, I., 2011. Trafficking for Begging: Old Game, New Name ((Kindle Version). Retrieved from Amazon.com).

Farr, K., 2005. Sex Trafficking: The Global Market in Women and Children. Worth Publishers, New York.

Hepburn, S., Simon, R.J., 2013. Human Trafficking around the World: Hidden in Plain Sight. Columbia University Press, New York.

Hossain, M., Zimmerman, C., Abas, M., Light, M., Watts, C., 2010. The relationships of trauma to mental disorders among trafficked and sexually exploited girls and women. American Journal of Public Health 100 (12), 2442–2449.

Hyland, K.E., 2001. The impact of the protocol to prevent, suppress and punish trafficking in persons, especially women and children. Human Rights Brief 8 (2), 30–31, 38.

International Labour Organization, 2005. A global alliance against forced labour. International Labour Conference. ILO Publications, Geneva.

International Labour Organization, 2008. ILO Declaration on Social Justice for a Fair Globalization. ILO Publications, Geneva.

International Labour Organization, 2012. ILO Global Estimate of Forced Labour: Results and Methodology. ILO Publications, Geneva.

Lee, M., 2011. Trafficking and Global Crime Control. Sage Publications Ltd, London.

Mason, M., 2013. Kids Pimping Kids Out for Sex in Indonesia. AP. Available at: http://m.apnews.com/ap/db_289563/contentdetail.htm?contentguid=jUqSkN0x (accessed 30.10.13.).

Méndez, M.M., 2010. Victim assistance programs, international. In: Fisher, B., Laub, S. (Eds.), The Encyclopedia of Victimology and Crime Prevention, vol. 2, pp. 966–967.

Moghadam, V.M., 2005. Globalizing Women: Transnational Feminist Networks. The Johns Hopkins University Press, Baltimore, MD.

Mohr, A., 2012. The Effects of Economic Globalization on Developing Countries. Demand Media. Available at: http://smallbusiness.chron.com/effects-economic-globalization-developing-countries-3906.html (accessed 06.11.13.).

Nagle, L.E., 2008. Selling souls: the effect of globalization on human trafficking and forced servitude. Wisconsin International Law Journal 26 (1), 131.

National Colloquium, 2012. National Colloquium 2012 Final Report: An Inventory and Evaluation of the Current Shelter and Services Response to Domestic Minor Sex Trafficking (Washington, D.C.).

Norwegian Ministry of Justice and the Police, 2011. United against Human Trafficking: The Government's Plan of Action against Human Trafficking (2011–2014). Norwegian Ministry of Justice and the Police.

Olteanu, A., Radoi, M.A., 2010. Positive and negative effects of globalization. Scientia International Journal 22 (2), 209–212.

Polakoff, E.G., 2007. Globalization and child labor: review of the issues. Journal of Developing Societies 23 (1–2), 259–283.

Raphael, J., Myers-Powell, B., 2010. From Victims to Victimizers: Interviews with 25 ex-pimps in Chicago. DePaul College of Law.

Sanghera, J., 2005. Unpacking the trafficking discourse. In: Kepadoo, K. (Ed.), Trafficking and Prostitution Reconsidered: New Perspectives on Migration, Sex Work and Human Rights. Paradigm Publishers, Boulder, pp. 3–24.

Scheper-Hughes, N., 2000. The global traffic in human organs. Current Anthropology 41 (2), 191–224.

Shangquan, G., 2000. Economic Globalization: Trends, Risks, and Risk Prevention. United Nations, New York.

Siddharth, K., 2009. Sex Trafficking inside the Business of Modern Slavery. Columbia University Press, New York.

Stohl, C., 2008. Globalization of organizations. In: Donsbach, Wolfgang (Ed.), The International Encyclopedia of Communication. Wiley Blackwell & International Communication Association.

Surtees, R., 2008. Traffickers and trafficking in Southern and Eastern Europe: considering the other side of human trafficking. European Journal of Criminology 5 (1), 39–68.

Tiefenbrun, S., 2007. Child soldiers, slavery and the trafficking of children. Fordham International Law Journal 31 (2), 417.

United Nations Office on Drugs and Crime (UNODC), 2006. Trafficking in Persons: Global Patterns. United Nations, Vienna.

United Nations, 1956. Supplementary Convention on the Abolition of Slavery, the Slave Trade and Institutions and Practices Similar to Slavery.

United Nations General Assembly, 2000. Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children, Supplementing the United Nations Convention against Transnational Organized Crime. United Nations, Palermo, Italy.

United Nations Office for the Coordination of Humanitarian Affairs, 2007. AFRICA: Fighting for Children. Available at: http://irinnews.org/report.aspx?reportid=70016.

United Nations Office on Drugs and Crime, 2009. Trafficking in Persons: Global Patterns. United Nations, Vienna.

United Nations Office on Drugs and Crime, 2012. Trafficking in Persons: Global Patterns. United Nations, Vienna.

U.S. Department of State, 2012. 2012 Trafficking in Persons Report – United States of America. U.S. Department of State.

Walk Free Foundation, 2013. The Global Slavery Index 2013. Walk Free Foundation.

Weitzer, R., 2007. The social construction of sex trafficking: ideology and institutionalization of a moral crusade. Politics & Society 35, 447–475.

Wolf, M., 2001. Will the nation-state survive globalization? Foreign Affairs 80 (1), 178.

World Commission on Social Dimension of Globalization, 2004. A Fair Globalization: Creating Opportunities for All. Available at: www.ilo.org/wcsdg (accessed 24.10.13.).

Zaleski, P., 2006. Global non-governmental administrative system: geosociology of the third sector. In: Gawin, Dariusz, Glinski, Piotr (Eds.), Civil Society in the Making. IFiS Publishers, Warszawa.

Zimmerman, C., Hossain, M., Watts, C., 2011. Human trafficking and health: a conceptual model to inform policy, intervention and research. Social Science & Medicine 73, 327–335.

Globalization and Religion

Peter Beyer, University of Ottawa, Ottawa, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the different ways that religion has been treated in the literature on globalization. Emphases in the literature tend to be on a global 'religious' function; on religious pluralism; on the religious dimension of local, ethnic, national, and civilizational identities; and on the way that transnational religious movements, religious organizations, and religious networks are able to go beyond and even ignore state and other local boundaries. Religion is a prime way of expressing both universality and particularity in global society. In contrast to the global capitalist economy, religions are more divergent than convergent; and yet, like culture and nation, religion is a social form to which all humans have access. Religions provide a way of multiplying ultimate narratives, and of doing difference in a situation where we all share the same place and time. Moreover, through the creation of diaspora communities, the amplification of readily accessible communication technologies, and the development of transnational organization, religions attain a more globalized presence. Globalization provides conditions for the multicentered reformation and reformulation of religious traditions, and for the creation of new religions whose existence is simultaneously global and local, tied to place and yet not tied to place.

Globalization, the historical process by which all the world's people come to live in a single social unit, brings with it religious changes as well. Prime among these is that globalization encourages religious pluralism such that religions identify themselves in relation to one another. Religions also become the objects of multiple centers of authority because of diasporas and transnational organization. In postmodern fashion, global religious pluralism creates a situation in which absolute visions end up living side by side and are engaging one another in highly diverse ways.

Globalization and Religion

Since the 1980s, the term globalization has taken on a certain currency in social scientific as well as in popular discourses. Its meaning in both these areas varies; but what unites economic, political, cultural, and even theological conceptions of globalization is the sense that, for better or worse, all people around the world live today in a single social unit much more than they ever have before. There is now and has been developing for some time a world society that serves as the encompassing social context of all more local social units. In this regard, most of the debates about globalization, whether scholarly or popular, have centered on the relation and even polarity between the global and the local. Roland Robertson (1992), a prime contributor to globalization debates since the 1980s, talks about the dialogical relation between the universal and the particular. The global context not only leads to a certain homogenization of cultures and social structures around the world, but also provides new conditions for constructing social and cultural difference. Global society, while single, is inherently pluralistic, not only permitting but even favoring a multitude of local and particular identities, both individual and collective. In that context, although this global society features great inequalities in power between what Immanuel Wallerstein (1979) calls core and peripheral regions, globalization is not just a matter of diffusion – say, of Western culture to the rest of the world; it is primarily the local appropriation of global models for such things as state, economy, or mass media in such a way that no region serves as the standard to which all others must conform. This is what Robertson and Zymunt Bauman (1998) call 'glocalization,' the global realized through the local.

The status of religion in debates about globalization has varied, to a significant extent according to academic discipline. Perspectives that see the development largely in economic terms, for instance as the increasing dominance of global capitalist institutions, tend to ignore religion almost entirely. However, an increasing number of approaches originating in the fields of international relations, political geography, and communication studies have, since the beginning of the twenty-first century, begun to pay attention to religion (Haynes, 2007; Calhoun et al., 2011). Yet it is still primarily the work of anthropologists, sociologists, and scholars of religion that accords religion a more visible place in the debate. Even here, however, analyses vary a great deal, for the most part depending on how the term 'religion' is understood. Two overlapping lines of variation appear to be central: the distinction between functional and substantive notions of what counts as religion, and whether one sees religion as an aspect of culture or as a differentiated and independent social domain.

A functional conception of religion regards as religion any social process that addresses certain problems, for instance the question of the ultimate meaning of life, or social integration and control. Here religion, or at least the religious, almost automatically becomes a necessary dimension in the formation of world society because problems of meaning or of self-description (integration) have to be addressed in any society, especially one as diverse as the global one. As examples under this heading, some authors have sought to analyze the emergence of a global culture that is the equivalent of a new global 'sacred canopy' (Meyer et al., 1997). One might even include a growing discussion about the development of a global 'civil society' under this heading (Rudolph and Piscatori, 1997; Juergensmeyer, 2005). In addition, a functional approach allows questions of national, regional, ethnic, or

civilizational identity in world society to take on religious character (Huntington, 1996; Robertson, 1992), and thus one will tend to see religion largely as an aspect of (global) culture. Substantive ideas of religion, by contrast, insist on the presence of identifiable and institutional sorts of activity that refer to a posited supernatural, transcendent reality. From this perspective, the relation of various specifically religious groups to globalization becomes the focus of attention, groups ranging from the Roman Catholic Church and Islamic organizations to Western Buddhist groups and Pentecostal movements (Martin, 2001; Learman, 2005; Rudolph and Piscatori, 1997). Such approaches also see religion as an independent domain of social activity and as more than an expression of something else, whether economics, politics, or culture (Beyer, 2006).

This variety in ideas about what counts as religion in the global context by itself already points to important aspects of the relation between globalization and religion. The fact that most of the literature that deals explicitly with globalization ignores religion or functionalizes it suggests that substantive or institutional forms are for these observers somewhere between irrelevant and antiglobal. One manifestation of this orientation is the great amount of attention that so-called fundamentalisms receive, and this, ironically, is very often from scholars who are quite positively oriented toward institutional or substantive religion. Thus, for example, the volumes issuing from Scott Appleby and Martin Marty's *Fundamentalism Project* (1991–95), although not directly about globalization, nevertheless assume globalizing processes; and they judge a wide variety of important religious movements all around the contemporary world as particularistic, and usually local reactions against them (cf. Juergensmeyer, 1993; Beyer, 1994). Among some of these authors, however, designating 'fundamentalisms' as, in effect, antiglobal movements parallels a positive attitude to the idea of a single world society, especially as manifest in the search for a specifically global ethos. In these cases, the search is on an explicitly religious basis, albeit not on the basis of a particular religion. What is somewhat ironic is that the resulting global ecumenical ethic (see esp. Küng, 1991) corresponds reasonably well to the intention of those who functionalize religion at the global level: that is, it avoids reference to a particular religion while addressing what it assumes to be the urgent and globally relevant religious task of providing ethical ideals and values that people all around the world can accept. In short, this sort of discussion seems to suggest that the distinction between functional and substantive conceptions of religion overlaps in a certain way those between the universal or global and the particular or local. Since these latter are principal organizing polarities in the entire globalization debate, it is best not to decide for one conception of religion or the other, but rather to consider the potential insights that flow from both.

Theoretical Perspectives

An excellent example of how a functionalized conception of religion can enter the globalization debate at the theoretical level is to be found in the work of John W. Meyer and his collaborators. In a 1997 article that clearly presents their basic outlook, they carry to the level of global society an analysis that the classical sociology of Max Weber and Emile Durkheim made for modern societies. For these scholars, the *world polity*, as they call it, is an extension of modern Western values and their corresponding rationalized structures to the rest of the world. Western religion had a critical role in bringing about this modernization and rationalization; but today the old religious elites have lost their power and have in any case also become believers in salvation through the pursuit of modern global values, especially progress (development) and equality. In their place, there is now a new 'religious' elite who are 'the professionals, researchers, scientists, and intellectuals who write secularized and unconditionally universalistic versions of the salvation story, along with the managers, legislators, and policymakers who believe the story fervently and pursue it relentlessly. This belief is worldwide and structures the organization of social life almost everywhere' (Meyer et al., 1997: 174). Meyer and his colleagues support the existence of this world polity by showing how the concrete policies of most countries around the world conform to its goals and values. In arguing in this fashion, however, they present a rather unitary and homogenizing picture that leaves room neither for the assertion of any meaningful particular or local variation nor, not coincidentally, for any role for institutional religion other than as futile reaction or powerless supporter.

A significantly more nuanced vision, as regards both the question of homogeneity vs heterogeneity and the possible role of institutional religion, is contained in the work of Roland Robertson. This author also emphasizes what he calls the 'globalization of societal modernity' (Robertson, 1991), which includes the sort of global religious function that Meyer et al. see. In addition, however, Robertson points much more positively to the proliferation of religious movements in the contemporary world, movements ranging from Latin American liberation theology and Küngian 'ecumenism' to Islamic 'fundamentalism,' Japanese Soka Gakkai, and the Sathya Sai Baba movement (Srinivas, 2010). His theoretical model makes room for these more institutional and substantive forms of religion, and not just as local defensive reactions against incorporation into the global system. Instead, Robertson allows that they can also be attempts to define the 'global-human' situation as a whole, and, *at the same time*, they can be ways of identifying a particular culture, region, or (in his terms) society in relation to the global whole and other subunits of it. The task of constructing such universalistic religious visions – whether functional or substantive – is always done on the basis of particular traditions, cultures, or movements. In consequence, the forms of institutional religion themselves change in the context of globalization, leading for instance to revisions of old traditions, syncretic or hybrid movements, and the rise of entirely new religions. For Robertson, then, the World Parliament of Religions, the Vishwa Hindu Parishad (World Hindu Organization), the Church of Scientology, a peripatetic pope or dalai lama, and Chinese Qi Gong groups are just as much aspects of globality and contributors to the shape of global society as are the aims and pronouncements of social scientists and technocrats. They all reflect and share a common social context, and one is not more modern or global than the other.

Religious Pluralism

From these theoretical starting points, one of the important questions that emerges concerns religious pluralism. Only with a radically functionalized concept of religion, such as that of the world polity school, can it appear possible for global society to share a single religious vision, a single worldview. All other conceptions point to the near impossibility of a hegemonic, global religion. Substantively speaking, contemporary global society shows no sign of moving in this direction. If anything, the opposite is the case: not only is there an increasing number of identifiable religions and religious traditions in the world, but also the variety of expressions within those religions is growing. Some sociologists of religion, such as Thomas Luckmann (1967), have long suggested that the increasingly dominant form of religious expression in today's global society is a highly individualized one that carries pluralism to the point where Luckmann can call it 'invisible religion.' Large institutionalized religions have been losing their authority and their social influence, but at the level of the individual, religious expression continues in myriad ways. These, however, are not publicly visible; they are private and in that sense invisible. Others do not go so far, but emphasize the degree to which even long-established institutional religions are manifesting themselves around the world in increasingly pluralistic fashion (Meister, 2011). Put somewhat differently, the question of religious pluralism is not simply one of how different religions, each one claiming an absolute vision, can live side by side in world society; it is also one regarding to what extent this society still favors authoritative institutions worthy of the name 'religion.' A partial answer is possible through a comparison of the idea of religion with those of culture and nation.

The model of global culture argues that, in the contemporary social world, the ideas of culture and nation serve as names for types of legitimate and fundamentally plural collective identities: cultures and nations, like individuals, have the right to equal existence. Within the context of homogenizing forces like global capitalism and mass media, one can assert one's difference and defend that difference in the name of one's culture or nation. Cultures and nations, like individuals, are legitimated agents or actors in global society: it is part of the world polity (of which Meyer et al. speak) that their right to 'self-determination' should be limited only by the right of other such identities to assert themselves in like fashion (Lechner, 1989). Religion is both a similar and different concept. On the one hand, religion in modern times has often formed part of cultural and national identities: it has very often helped to define them (Beyer, 2011). On the other hand, global culture also allows that religion is something independent of these other two notions. Religions are something to which people assumedly adhere voluntarily as individuals; and this may or may not flow from their cultural or national identities. Religions also enjoy the same 'protection' as cultures and nations; at least, such freedom of religion is an ideal expressed in most of the world's state constitutions (Richardson, 2004). Correspondingly, global cultural ideals regard religions, like cultures and nations, as inherently plural. Two seeming paradoxes emerge from this view. First, religious freedom is grounded not only in the individual's right to choose, but also in the right of the religious collectivity to exist and determine itself. Religious collectivities can therefore claim authority over their members, but only with the consent of these latter! Second, religions are particular and plural, and yet it is understood that they each offer a universal or cosmic vision, a total way of life. This holistic emphasis therefore includes the possibility of exclusivist religions, ones that deny the validity of all others. Thus, in true postmodern fashion, religions provide ways of multiplying ultimate narratives about how the world is, and they do this in such a way that each is regarded as of equal value, the choosing of one or the other being deemed a matter of preference, or upbringing, or culture. Ultimately, each is grounded only in the conviction of its adherents, and this in the knowledge that most others in global society have different convictions: even by the most generous estimates, the largest religion, Christianity, can claim far less than half of the world's population.

Religion in Postmodernity

The postmodern character of religion in the context of world society is manifest in other features as well. Embedded in the idea that religion is fundamentally plural are the questions of how religions distinguish themselves from one another, and what the difference is between religion and nonreligion. Neither question is susceptible to a simple and clear answer. The idea that there is a plurality of distinct religions is one that dates historically only from seventeenth-century Europe in the context of post-Reformation religious conflicts, and as European expansion brought its people into protracted contact with a much wider variety of other cultures, civilizations, and hence religious traditions. Which traditions emerged from this historical process as 'one of the religions' depended both on what Westerners observed as religions and on what non-Westerners were willing to style according to the implied model of a religion. Different and even diametrically opposite outcomes resulted. Thus, the elite carriers of 'Hinduism' took up the suggestion of Western observers and cooperated in reimagining South Asian religious traditions as somehow a single religion; whereas corresponding Chinese intelligentsia refused, in spite of Western observation, to reconstruct their Confucian thought and ritual traditions as a religion (Beyer, 2006). The contemporary list of (world) religions therefore has an inductive and even arbitrary character that cannot be reduced to abstract characteristics. Like nations and cultures, what comes to be seen as one of the religions depends in large part on how well the supposed members succeed in having others accept it as one.

The boundary between religion and nonreligion exhibits comparable ambiguities, the difficulty of deciding between functional and substantive ideas of religion being one example. At the core of this problem lies the notion that religion refers to an overarching view of the world and thus to a complete way of life; and yet religions are, in principle and in fact, plural. A prime symptom of this ambiguous status of religion in world society is that, over the past one and a half centuries, key representatives of almost all the so-called world religions have claimed that 'ours is not a religion.' Most often, protests of this

sort favor a broad functional conception of religion for their tradition, and they reject the sort of substantive one that limits religion to a special category of human action – say, one dealing only with interaction with gods or spirits. Similarly, opponents of various new religions have sought to exclude these from the category, in these cases so as to eliminate them from the range of legitimate religious choices. Such conflicts over categorization are, on occasion, ways of circumventing the global cultural value of 'freedom of religion,' justifying in these instances the obligatory adherence of everyone to the tenets and practices of a particular 'religion,' or preventing such adherence to another (Richardson, 2004). Thus, the category of religion in global culture allows a postmodern juxtaposition of absolute narratives without offering an overarching narrative to reconcile them.

Religious Authority, Transnational Organization, and Diaspora

Accompanying these various ambiguities regarding the category of religion and religions, the carrying out of institutional religion under conditions of globalization brings a number of other challenges and changes, especially with respect to questions of religious authenticity and authority. In any society, there are a limited number of ways through which conformity to authoritative religious traditions can come about. Thus, for example, in small-scale societies a relative lack of alternatives, and social structures centered on face-to-face interaction, could assure reasonable religious conformity. After the advent of writing and large-scale political empires, dedicated priesthoods or monastic orders, political state support, along with distinctions between elite and popular religiosity (e.g., religion vs superstition) could handle the problem according to hierarchical principles. Under conditions of globalization, neither relative social isolation nor hierarchical solutions are possible. Globalized communication media, from telephones and air travel to television and the internet, assure that religious social networks need not be limited to local interaction or particular places (Lyon, 2000). In fact, the existence of these possibilities positively encourages global or translocal religious networks and religious organizations. Moreover, the global cultural valuing of equality for both individuals and collectivities militates against the notion that the orthodoxy practiced or advocated by a certain group of religious professionals or adherents in a certain place should serve as the unquestioned standard for all others. Even in religions that have traditional centers of such authority (e.g., Rome for Roman Catholics, al-Azhar for Sunni Muslims, or Salt Lake City for Mormons), these centers have limited resources for enforcing their authority; and those that choose to ignore them can use the same techniques of individual religious freedom, (alternative) tradition, separate organization, and transnational networking to circumvent it. The only real exceptions are situations in which state power helps enforce religious authority; but in those cases, such as Iran, the state boundaries themselves set effective limits to such compulsion.

A prime illustration of the way that globalization encourages the multiplication of local centers of religious authenticity is the effect on religion of global migration and transnational religious networking and organization. Although the migration of people from one part of the world to others is a constant of world history, what is new under conditions of globalization is both the extent of the phenomenon and the fact that migrants leave one place for another far less definitively. The possibilities for maintaining communicative and even physical ties with the places of origin are far more developed, meaning that migrants are less dependent on their new contexts and can still influence the places that they have left. From a different angle, the difference between migrants and nonmigrants is not as great as it used to be, just as the difference between travel and tourism vs migration is not as clear. And the chief source of this new situation is the much greater and more rapid communicative links that now exist throughout the world.

Among the consequences for religion of migration and enhanced transnational or translocal linkages, two are of the greatest potential significance. The first is that, as migrants from a particular religion settle in different parts of the world outside of the traditional 'heartlands,' these diaspora communities will develop religious variations that respond to their new locations, and may furthermore re-export these to their places of origin as authentic and authoritative versions. Thus, for example, the temples that Hindu migrants typically set up in diaspora communities of Europe and North America often allow the worship of a range of gods and goddesses in the Hindu pantheon, unlike in India where temples are usually dedicated to one deity only. The diasporic 'one-stop supermarket' temple is primarily a response to the fact that Hindus from different parts of India migrate to the same places and that these communities do not have the resources to build temples for each subgroup. In spite of these strategic origins of the variation, however, the model is being transferred back to India, providing another authentic way of 'doing' this religion (cf. Waghorne, 2004). Similar analyses could be made for Muslim, Christian, Buddhist, or other migrants.

The second major consequence for religion is related but also in a sense opposite. It consists in the possibility of tying together disparate versions of a religion across the globe, expressly in spite of the difference in local variations. The Roman Catholic Church is perhaps the prime example of a religious organization that has taken great advantage of this possibility. In many ways, it is the 'same' no matter where one goes in the world, and yet local variations are not thereby suppressed or excluded. The links among the world's Islamic populations have also strengthened significantly in this context, such that one can speak of a 'global Islam' in spite of the significant variations in the practice of this religion all around the globe (see Levitt, 1998; Roy, 2004). Another very illustrative example is global Christian Pentecostalism. Here, the concrete ways of doing Pentecostal Christianity vary enormously when one compares places like South America, North America, Africa, Korea, and elsewhere in Asia. Yet transnational linkages in such forms as common publications, regular world conferences, connections among professionals, and migration encourage a consciousness in most Pentecostals of being part of the same worldwide religious movement, in spite of the differences among them (see Poewe, 1994).

Conclusion

What this article demonstrates is that globalization creates and enhances conditions in which religion can become more pluralistic and yet at the same time more unified. The local and the global in religion, the universal and the particular, are not opposite poles between which religious forms oscillate. Rather, local religion can now increasingly only be done with reference to the global context, especially in the awareness of religious pluralism. And global religious forms expressly include and are the expression of local manifestations. Global society does not present the choice between antiglobal, perhaps 'fundamentalist' resistance and proglobal, liberal accommodation. Instead, it recreates the conditions under which religions form and operate in relation to one another (Corten and Marshall-Fratani, 2001).

See also: Civil Religion; Diaspora; Global Governance; Globalization and World Culture; Globalization, Anthropology of; Globalization: Legal Aspects; Immigration Policy; Legal Pluralism; Migration: Cultural Aspects; New Religious Movements; Pluralism and Tolerance; Postmodernism in Sociology; Religion and Economic Life; Religion and New Media; Religion, Sociology of; Religion: Nationalism and Identity; Religious Fundamentalism: Cultural Concerns; Religious Nationalism and the Secular State; Secularization; Transnational Citizenship; Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns.

Bibliography

Bauman, Z., 1998. Globalization: The Human Consequences. Polity, London.
Beyer, P., 1994. Religion and Globalization. Sage, London.
Beyer, P., 2006. Religions in Global Society. Routledge, London.
Beyer, P., 2011. Religious pluralization and intimations of a post-Westphalian condition in a global society. In: Michel, P., Pace, E. (Eds.), Annual Review of the Sociology of Religion. Religion and Politics, vol. 2. Brill, Leiden, pp. 3–29.
Calhoun, C.J., Juergensmeyer, M., Van Antwerpen, J. (Eds.), 2011. Rethinking Secularism. Oxford University Press, Oxford.
Corten, A., Marshall-Fratani, R. (Eds.), 2001. Between Babel and Pentecost: Transnational Pentecostalism in Africa and Latin America. Indiana University Press, Bloomington, IN.
Haynes, J., 2007. An Introduction to International Relations and Religion. Pearson Education, Harlow, UK.
Huntington, S.P., 1996. The Clash of Civilizations and the Remaking of World Order. Viking Penguin, New Delhi.
Juergensmeyer, M., 1993. The New Cold War? Religious Nationalism Confronts the Secular State. University of California Press, Berkeley, CA.
Juergensmeyer, M., 2005. Religion in Global Civil Society. Oxford University Press, Oxford.
Küng, H., 1991. Global Responsibility: In Search of a New World Ethic. Crossroad, New York.
Learman, L. (Ed.), 2005. Buddhist Missionaries in the Era of Globalization. University of Hawai'i Press, Honolulu, HI.
Lechner, F., 1989. Cultural aspects of the world system. In: Swatos, Jr, W.H. (Ed.), Religious Politics in Global and Comparative Perspective. Greenwood, Westport, CT.
Levitt, P., 1998. Local-level global religion: the case of United States–Dominican migration. Journal for the Scientific Study of Religion 37, 74–89.
Luckmann, T., 1967. The Invisible Religion: The Problem of Religion in Modern Society. Macmillan, New York.
Lyon, D., 2000. Jesus in Disneyland: Religion in Postmodern Times. Polity, Cambridge, UK.
Martin, D., 2001. Pentecostalism: The World Their Parish. Blackwell, Oxford.
Marty, M.E., Appleby, R.S. (Eds.), 1991–1995. The Fundamentalism Project, vol. 5. University of Chicago Press, Chicago, IL.
Meister, C. (Ed.), 2011. The Oxford Handbook of Religious Diversity. Oxford University Press, Oxford.
Meyer, J.W., Boli, J., Thomas, G.M., Ramirez, F.O., 1997. World society and the nation-state. American Journal of Sociology 103 (1), 144–181.
Poewe, K., 1994. Charismatic Christianity as a Global Culture. University of South Carolina Press, Columbia, SC.
Richardson, J.T. (Ed.), 2004. Regulating Religion: Case Studies from around the Globe. Kluwer Academic/Plenum, New York.
Robertson, R., 1991. The globalization paradigm: thinking globally. In: Bromley, D. (Ed.), Religion and the Social Order. New Developments in Theory and Research, vol. 1. JAI Press, Greenwich, CT.
Robertson, R., 1992. Globalization: Social Theory and Global Culture. Sage, London.
Roy, O., 2004. Globalized Islam: The Search for a New Ummah. Columbia University Press, New York.
Rudolph, S.H., Piscatori, J. (Eds.), 1997. Transnational Religion and Fading States. Westview, Boulder, CO.
Srinivas, T., 2010. Winged Faith: Rethinking Globalization and Religious Pluralism through the Sathya Sai Movement. Columbia University Press, New York.
Waghorne, J.P., 2004. Diaspora of the Gods: Modern Hindu Temples in an Urban Middle-Class World. Oxford University Press, New York.
Wallerstein, I., 1979. The Capitalist World-Economy. Cambridge University Press, Cambridge.

Globalization and Transnational Corporations

Alessandro Bonanno, Sam Houston State University, Huntsville, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article first reviews debates that stress the contrasting interpretations of globalization existing at the social, political, and economic levels. Globalization is defined in terms of the combined processes of acceleration of time and compression of social space. The development of transnational corporations (TNCs) since the 1960s reveals that these organizations, that operate globally and do not necessarily identify with a particular country, play an increasingly important role in the global economy, politics, and society. This article reviews the primary interpretations of the position and role of TNCs under globalization.

Globalization

Definitions and the Debates

The concept of globalization has been employed to define processes at the social, political, and economic levels. Socially, it has been interpreted as a complex phenomenon that molds distant groups into unified networks, allows enhanced communication, promotes the exchange of ideas and experiences, fosters understanding and bringing people closer together (Croucher, 2004; Hopper, 2007; Jagdish, 2004). It has also been viewed as an entity that increases interdependence, mutual exchange awareness, and respect and fortifies society's cohesiveness (Friedman, 2005; Guillen, 2001). Globalization's greater circulation of capital, labor, and products has been heralded as one of the primary conditions for social growth (Wolf, 2004). These positive views have been countered by interpretations that analyze globalization as a phenomenon that engenders the exploitation and oppression of local groups and cultures, eliminates differences, represses identity, and instigates radical resistance and opposition (Barber, 1995; Giddens, 2000; Hosseini, 2009; Ritzer, 2008). It has been further regarded as a process that enhances the centrifugal power of capitalism, dismembers communities, and undermines the stability of society (Amoore, 2005; Harvey, 2006; Giddens, 2000).

Politically, globalization has been analyzed as a deterrent to unilateralism and a tool to enhance cooperation among people and nations. While the United States remains the 'benign hegemon' of the world, the working of the global system is based on, and nurtures, a greater participation of all people in governance (Friedman, 2000). This view is opposed by concerns about the reduced power of the nation-state and the limited ability of people to be represented in political debates and decision making. The crisis of political representation (the declining connection between the rulers and the ruled), and the growing inability of democratic political institutions to control the actions of global corporations and actors have often been mentioned among the most distinctive outcomes of globalization (Boggs, 2000; Wettstein, 2009).

Economically, globalization has been described as a condition for growth and a tool for the elimination of socioeconomic inequality. Because of globalization, income is now distributed more justly and the idea that one or few nations can dominate other nations is considered obsolete. The emergence of a global ruling class, it is argued, promotes just development (Becker and Sklar, 1999; Firebaugh and Goesling, 2004). Opponents of globalization have indicated that globalization is a class project that increases the exploitation of the lower classes and the global South (Harvey, 2005, 2010; Robinson, 2004). It has been also defined as the outcome of the growth of monopoly capital that engenders crises and stagnation (Bellamy Foster and McChesney, 2012). Others contend that the growth of transnational corporations (TNCs) and the declining power of political and economic institutions (i.e., the nation-state, unions, and social movements) resulted in the concentration of wealth in the hands to the top 1% of the world population and the worsening of the socioeconomic conditions of the rest of society (Milanovic, 2011; Stiglitz, 2003; Volscho and Kelly, 2012).

While contrasting interpretations characterize the debate on globalization, there is convergence on the definition of this phenomenon. It refers to the acceleration of the time and the compression of space within which social relations take place (Harvey, 2010; Bonanno and Cavalcanti, 2011). In other words, globalization means that the social world is 'smaller' and people 'move about it in a much faster way than before.' Viewed as the acceleration of social time and the compression of the social space, globalization does not appear to be a novel process. In effect, these have been constant characteristics of the development of capitalism. As indicated by classical social theorists – such as Karl Marx and Max Weber – the advent of capitalism mandates a constant acceleration of time and compression of space. The novel dimension of current globalization is that time acceleration and the compression of space occur in a quantitatively and qualitatively enhanced manner than in the past (Robinson, 2004).

At the spatial level, the establishment of capitalism meant that the feudal regional space was reorganized in favor of the creation of a greater (national) capitalist space and its political representation of the nation-state. Marx (Marx and Engels 1998 [1848]: pp. 39–41), for instance, underscores the historical importance of the creation of a 'national space' in which interdependent markets, business organizations, and labor networks are established, organized, and controlled and where commodity production and consumption are organized and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64132-6

executed more efficiently and rapidly. At the level of time, the increased velocity of the turnover time of capital and overall efficiency in the capitalist structure of accumulation were paramount for economic growth. Nation-states employed their powers to establish, organize, and control national markets and to assist their bourgeoisies in expanding the scope and size of the economy. Simultaneously, bourgeoisies expanded their space of action, and organized and controlled the forces of production contained within it, and, most importantly, the use of labor. Marx saw homogenization and standardization of production processes and their organization under national laws, interests, policies, and practices, via the state to be primary objectives of the capitalist class. Weber (1963: pp. 166–68) viewed the standardization of time and space as a new moral dimension that propelled the growth of capitalism.

Yet, classical social theorists recognized that capitalism was never exclusively domestic and that, in fact, it is global in nature. In the case of Marx, he stressed emphatically that capitalism's dynamic nature generates business activities far beyond the nation-state's organized and increasingly homogenous spatial dimension. He argued that capital mobility, under modern industry, transcends the nation and with that the regulatory and controlling dimensions through which the nation-state embeds the economy culturally and institutionally. Marx held that capitalism's tendency to expand would lead to a global capitalist system. It must create and recreate new spaces to continue the search for profit. The bourgeoisie's need to reshape space to fit the requirements of capital accumulation, organize its factors of production and, above all, to reproduce and control labor was, for Marx, a fundamental component of capitalism. Capital accumulation requires economic and political concentrations that are manifested through the organization and regulation of the use of the forces of production. Simultaneously, the regulation of the forces of production requires the existence of forms of social interdependence – such as that between capitalists and workers – that are codified in the spatial and temporal conditions associated with the nation. As these conditions are eroded and capital mobility transcends the nation-state, established forms of interdependence are destroyed displaying the contradictory and unsustainable dimension of capitalism.

The Origins of the Term Globalization and Debates in the Late Twentieth Century

Despite these classical insights into the development of a global economy and society, the term globalization was rarely used for most of the twentieth century. In effect, its origins and first use are not very clear. There are unconfirmed claims that references to global socioeconomic processes were made in the first portion of the twentieth century. However, it was only in 1951 that the Merriam-Webster dictionary of the English language officially acknowledged the word globalization. In academic circles, the term gained some recognition in the 1970s through some publications that discussed the growth of multinational corporations (MNCs) (Barnet and Muller, 1976; Hymer, 1971). In 1972, the political scientist Fuad Ajami employed the term globalization to stress the characteristics of the expansion of corporations at the global level. In the early 1980s, Theodore Levitt published the article 'The Globalization of Markets' in the Harvard Business Review (Levitt, 1983). Disregarding classical debates, Levitt argued that the world entered a new era in which national differences in consumer preferences had become largely irrelevant. Changes at the technological and social levels allowed corporations to market their products worldwide in ways that were impossible before. The essence of Levitt's argument was that a set of new and qualitatively different factors enhanced the ability of companies to conduct business globally. After this and similar works published in the 1980s, the term globalization gained worldwide popularity.

As the debate evolved in the 1980s and 1990s, it was characterized by three camps. The first referred to the once dominant Modernization School. Its key argument of the world development is characterized by the independent and autonomous growth of nations (Parsons, 1971; Rostow, 1960). Such views were criticized for blaming poor countries for their conditions, justifying the unequal redistribution of resources and power, disregarding the relationship between development and underdevelopment, and, ultimately, ignoring the historical conditions that shape development. In its ahistorical formulation of growth, modernization theory proposed a view in which each country would develop by adopting the culture and institutions of the United States. This process of cultural and economic modernization or Americanization (Parsons, 1971) would be reached through stages that would be achieved according to each country's own ability to mobilize the necessary resources to grow (Rostow, 1960).

Objections raised by Marxian structuralists were particularly strong. This camp consisted of two major subgroups. The first included the world system school led by Immanuel Wallerstein (1974; 1980; 2005) while the second referred to the regulationist school (Aglietta, 1979; Lipietz, 1987, 1992). The world system school borrowed from the Longue Duree School (The Longue Duree group refers to that group of scholars that were originally associated with the French journal *Annales d'histoirie économique et sociale*. First popularized in the 1930s and 1940 by the work of Marc Bloch and Lucien Febvre, and later in the post–World War II decades by that of Fernand Braudel, this group's historiography was centered on the importance of the evolution of long-term historical conditions (from this the name Longue Duree or long term) and the view that contemporary socioeconomic phenomena find their roots and explanations in established historical trends. Wallesterin, along with other important representatives of the world system school such as Giovanni Arrighi, was associated with the Braudel Center at the State University of New York, Binghamton.) and emphasized the long-term evolution of social relations. For Wallerstein, development could be understood only by analyzing the emergence of capitalism in Europe and its evolution. This evolution, he contended, culminated in a 'world system' in which all nations were connected together in patterns of growth but also underdevelopment. Rejecting the assumption that the growth of each nation evolves following indigenously based stages and strategies of development, world system theorists demonstrated that the development of advanced societies (core countries) was the result of processes of underdevelopment of countries in the periphery (underdeveloped countries) and semiperiphery (less developed countries). Under world capitalism, the socioeconomic growth of

core nations was the direct outcome of concomitant processes of underdevelopment of countries of the world periphery and semiperiphery.

At this stage of the debate, the contribution of world system theory was relevant for at least two reasons. First, it proposed a view in which the global system was understood as a process that initiated with the growth of capitalism. Accordingly, and following Marx's classical analysis, world system theorists rejected the thesis that attributed the genesis of globalization to recent events and insisted on the long-term dimension of the phenomenon. Second, world system theory reaffirmed the centrality of the nation-state in contemporary capitalism. While they stressed the historical nature of the nation-state – the nation-state is only one of the historical forms of the state – they also viewed the world system as a system among nations, an international system. In this respect, world system theorists remained skeptical of statements that associated globalization with the crisis of the nation-state.

World system theorists' skepticism about the novelty represented by globalization remained. And statements about this concept's inadequacy were re-proposed periodically. Wallerstein (2000), for instance, has frequently and overtly argued that globalization is not a new phenomenon. Christopher Chase-Dunn (1998) contended that "Most discussions of globalization assume that, however defined, it is a fairly recent phenomenon … [I]f we take a long-term view of the structural constants, cyclical processes, and secular trends that have operated in the Europe-centered system for several centuries we can understand that there have been no recent major transformations in the developmental logic of the world-system" (Chase-Dunn, 1998: ix).

Exemplifying the reaction to this position, Robinson (2004) argues that there are quantitative and qualitative differences between globalization and previous periods of capitalist development. Accordingly, it is misleading to argue that globalization is simply the continuation of trends established in the past. From a quantitative point of view, Robinson maintains that trade is much larger under globalization than in other historical periods and this is the case also for the scale of capital flows, and the level of information and communication technology. He also stresses the differences associated with the global presence of TNCs and a transnational labor force. Qualitatively, Robinson underscores that globalization "has entailed the fragmentation and decentralization of complex production chains and the worldwide dispersal and functional integration of the different segments of these chains … In this way globalization is unifying the world into a single mode of production and a single global system and bringing about the integration of different countries and regions into a new global economy" (2004: p. 15). He concludes that because of these changes the nation-state has lost the central role that it played for most of the twentieth century.

Members of the regulationist school maintained that each era of capitalism was defined by a 'regime of accumulation' and a 'mode of regulation.' The concept of regime of accumulation refers to the system of wage relations that defines a particular era of capitalism. Mode of regulation indicates the social norms that govern each regime of accumulation. In the 1980s, regulationists concentrated their efforts on the analysis of the end of Fordism and the emergence of a new and post-Fordism regime of accumulation. At the time, post-Fordism was used widely to refer to conditions that latter would be associated with globalization.

According to the regulationists, Fordism indicated the 'regulated capitalism' that characterized most of the twentieth century. Introduced by the Italian philosopher Antonio Gramsci (1971) to define the advanced Taylorist capitalism of the 1930s, it was re-proposed by the regulationists to refer to the combination of Keynesian political economy and state intervention at the social level that characterized the development of advanced Western societies in the post–World War II decades. The crisis of Fordism, they contended, was the result of the so-called rigidities of this regime of accumulation. Also theorized from a number of different perspectives (see O'Connor, 1974; Habermas, 1975), the concept of rigidities referred primarily to the fiscal and ideological dimensions of Fordism. At the fiscal level, the cost of maintaining state intervention became too high. The combination of an increased demand for state-sponsored services and the world economic crisis of the early 1970s prevented many nation-states from continuing their support of corporate activities at previous levels. State attempts to increase tax and introduce more regulations motivated corporations to export profit and bypass state requests. The result was a fiscal crisis of the nation-state. Simultaneously and at the ideological level, the inability of nation-states to address many of the existing social issues discredited the political elites and their ideologies. This legitimation crisis engendered widespread calls for changes in the administration of society and the economy. The regulationists viewed these changes as indications of the emergence of a new global regime that they called post-Fordism.

Postmodernity was the third major camp that characterized the debate on development and globalization in the 1980s and 1990s. Dwelling on poststructuralist contributions (Foucault, 1969, 1975; Derrida, 1978), postmodern theorists (e.g., Baudrillard, 1983, Laclau and Mouffe, 1985, 1987) argued that the issue of global development could not be adequately addressed employing 'grand narratives.' By grand narratives, they referred to holistic explanations that could be applied across time and space. In their critique, they maintained that the complexity of human existence could not be explained by modern theories and their assumption of universalism. Similarly, development could not be understood by using these modern theories. Denouncing the oppressive character of grand narratives, postmodernists maintained that the elimination of class inequality would not necessarily address other forms of oppression such as gender, race, and ethnic dominations. Furthermore, liberation strategies that spoke for the disenfranchised often silenced the very voices they claimed to support. In contemporary society, theories of development and liberation were transformed in theories of oppression. Their critique of left leaning and socialist theories was often more strident than those coming from the far right. Their argument centered on the historical failure of socialist projects and the 'exclusion' and 'oppression' of many groups that these theories and practices engendered. The global society, they contended, requires the acknowledgment of the existence of a plurality of voices and patterns of action as repression emerges from all sides of the political spectrum.

Globalization and TNCs

In the first two decades of the twenty-first century, the debate on globalization expanded significantly and included contributions on the growth of TNCs. In these contributions, TNCs are defined as enterprises that operate globally and have virtually severed connections with their home nation-states. They act disregarding the interests and well-being of their home countries and governments. They are different from MNCs that are enterprises that operate in more than one country through branches and subsidiaries. Their national origins are known and recognized as they are identified with, and claim membership in, a particular nation-state. They also differ from domestic corporations – enterprises that operate predominantly within one single nation. In relevant literature, the difference between MNCs and transnational or global corporations is often downplayed. In effect, the two terms are often used interchangeably. Simultaneously, it has been underscored that many MNCs have evolved into TNCs in recent decades.

The transformation of MNCs into TNCs has been associated with the growth and crisis of Fordism. During the Fordist twentieth century, MNCs strengthened the *mother–daughter* industrial system (Dunning, 1993). In this system, a *mother* company with headquarters and operations in the metropolitan homeland opened *daughter* subsidiaries overseas. These subsidiaries were directly controlled by, and managed from, the home headquarters and, therefore, operated with a highly restricted autonomy. As the twentieth century unfolded, MNCs expanded counting on the active military, political, and economic assistance of their home nation-states. Additionally, MNC's ownership and management, while including an international presence, remained largely in the hands of citizens of the company's home country, a situation that further legitimized the intervention of home governments in host countries. In return, host countries benefited from the repatriation of profit and the associated tax revenue, domestic investment and the creation of local jobs that contributed to social stability and economic growth.

The beneficial association between MNCs and the United States and other major nation-states reached a point of crisis in the late 1960s and 1970s. As the United States was increasingly asked to assist and protect MNCs, the search for more lucrative markets – including commodity and financial markets – and new cost-saving operations pushed MNCs to repatriate profits at an increasingly lower rate. While foreign direct investment (FDI) increased, this move signified less revenue for the home state and the concomitant escalation of fiscal pressure. In the United States, the declining rate of repatriation of profit exacerbated the fiscal crisis of the American state. The success of the postwar Fordist system mandated an expanded welfare state and social services that were supported through state expenditures and taxation. Corporations contributed to Fordism by maintaining high wages and salaries that sustained mass consumption and provided popular legitimacy to it. Internationally, the American government's actions to maintain the Pax Americana also became increasingly costly monetarily as well as politically. The Vietnam War crisis symbolized the high monetary and political costs associated with US control of the world economy and politics.

Responding to this new domestic and, and above all, international economic situation, first the United States, and later European, corporations expanded their foreign affiliates and FDI. Some developing countries counteracted through the expropriation of corporate assets and the implementation of import substitution strategies that favored local companies. Fueled by the negative economic consequences of the oil crisis, the loss of legitimacy of the modernization project and a move to the political left by many governments, more host countries overtly antagonized MNCs' presence on their territories. In order to overcome what they perceived as mounting adverse conditions, MNCs adopted novel means to reignite profitability. They began to decouple their presence overseas with the interests of the home countries. They began global searches for convenient production sites featuring favorable political contexts, support for deregulation, privatization of once publically controlled enterprises and resources, and open markets. They searched for locations in which national and regional governments adopting neoliberal policies allowed the creation of production networks that transcended the classical mother–daughter system. In essence, the corporate response to the crisis of Fordism consisted in the 'transnationalization' of their operations. Decentralization of production, the creation of production and distribution networks, and the global mobility of profit and investment defined the new TNCs (Robinson, 2004; Bonanno and Constance, 2008).

TNCs' moves were entwined with changes in the posture of many nation-states – and primarily the United States – that embraced neoliberal ideology and policies. Framed as an effective manner to attract fleeing corporate investment, nation-states adopted globally oriented macro-organizational strategies that shifted emphasis away from socially oriented goals (i.e., social stability, equitable socioeconomic growth, and welfare spending) to profit enhancing objectives. In this context, the allocation of FDI became increasingly linked to the 'regulatory and policy instruments' employed by nation-states and the existence of supportive economic policies. This trend was a common feature of national economic policies worldwide during the final years of the twentieth century. The neoliberal deregulation of markers eroded the long-standing association between corporations and home countries as TNCs began to look for the maximization of profit at the global level, paid increasingly less attention to the home country's domestic problems, and lack of corporate accountability emerged as a new 'social problem.' TNCs increase in FDI, for instance, while creating fresh opportunities for corporate profit, was one of the factors behind the *deindustrialization* process and the loss of many well-paid and stable jobs in many home countries.

In pertinent debates, the relationship of TNCs and the nation-state is discussed in contrasting terms. Extreme readings of this process (see Ohmae, 1995) portray TNCs as entities that have cut their 'home nation' attachments and operate as almost exclusively stateless global actors. Because of TNCs' power and ability to freely move about the globe, Ohmae and likeminded scholars argue that the nation-state has lost most, if not all, of its ability to control and regulate economic and social matters as it did in previous eras of capitalism. Framed as the ability of TNCs to 'bypass' state requirements, the crisis of the state has been identified as one of the most distinctive features of globalization (e.g., Ohmae, 1995; Constance and Heffernan,

1991). Reacting against this position, some authors, such as Arrighi (2005, 1998), Chase-Dunn (1998), and Hirst and Thompson (1996), contend that the nation-state continues to remain the centerpiece of capitalist development and an entity needed by corporations operating domestically and abroad. Corporations, therefore, maintain their links with nation-states even though conflict and divergence of interests may arise. According to this view, capitalism continues to be characterized by international social relations or relations among nations. It remains centered on the evolution of the nation-state and its capacity to forge a more or less unified national identity and culture while promoting and controlling the flows of resources and processes necessary for the development of the economy domestically and internationally. For Paul Hirst and Graham Thompson (1996), for instance, true TNCs do not exist so the purported socioeconomic changes that resulted in the creation of TNCs are unsubstantiated. They maintain that most corporations are still national as they are still connected to their countries of origin in significant ways regardless of their international activities. Additionally, data do not indicate the existence of trends toward the emergence of truly global corporations. In effect the examination of FDI reveals that it is generated by companies located in North America, Europe, or Japan. Hirst and Thompson further maintain that while the postwar system that regulated capitalism has exhausted its role, a new system of regulation has been set in place. This new mechanism allows the three superpowers (the United States, the European Union, and Japan) to govern the world and ensure adequate levels of socioeconomic stability.

According to other scholars (Dicken, 1998; Harvey, 2010, 2005; Robinson, 2004; Sklair, 2001), the recent evolution of the economy and society has brought about a qualitatively different set of conditions for the accumulation of capital of which the growth of TNCs is among the most relevant. The distancing between TNCs and their home nation-states alters the historically built connection between these two entities. Corporations increasingly bypass home state requirements and requests while they strategically forge associations with other nation-states and regional or continental groups of nation-states (Bonanno and Constance, 2008). Because of these changes, the modus operandi of TNCs is different from the historical role played by MNCs throughout most of the twentieth century. These scholars also stress that TNCs play an active role in the underdevelopment of many less advanced regions and nations of the world. The soaring corporate power, they argue, has negative consequences for the economy and society as economic inequality increases and the unchecked actions of corporations are destabilizing (Krugman, 2007). They identify TNCs as active contributors to the mechanisms that allow the transfer of wealth from periphery to core countries. Furthermore, they question these corporations' social and political practices. TNCs have been often identified as 'undesirable citizens' that contaminate the democratic institutions and political climates of host countries and engender unwanted social consequences such as environmental pollution, labor exploitation, political violence, physical abuse, and death. They contend that the power of TNCs reduces the level of democracy and prevents national and regional constituencies from participating in decision making processes (i.e., Boggs, 2000).

Scholars who support neoliberal positions express much more sympathetic views of the TNCs' actions. They maintain that those who oppose corporate actions and presence in less developed regions do not recognize the benefits that expanded business opportunities generate for these societies and the fact that they would suffer many more negative economic consequences in the absence of corporate investment and business activities. Because of their visibility, TNCs obey the law in much greater extent than smaller companies. However, also because of their visibility, they are often attacked by a variety of opponents, including unsympathetic nation-states. Advocates of this position hold that TNCs, to defend themselves, are sometimes forced to take actions that cause negative social and political consequences in the host country. However, they also argue that the claim that TNCs coerce nation-states to accommodate their interests through the use of illegal means is 'greatly exaggerated.' They contend that TNCs generate far more positive socioeconomic consequences to the communities where they operate than the negative outcomes often reported in scientific and popular venues (Rubner, 1990; Firebaugh and Goesling, 2004).

According to these accounts, the economic decline of the United States and the concomitant growth of Europe and Japan, as new world powers experienced since the 1960s, affected the regulation and coordination of world affairs based on the nation-state system. In this context, corporations could no longer depend on the United States to maintain global order and on nation-states' actions to support corporate interests domestically and internationally. The global development of anticorporate sentiments in advanced countries produced a public mood of indifference to the adverse treatments and conditions that corporations experience abroad. This situation affected advanced countries' governments that grew increasingly unwilling to intervene. Additionally, the growth of anticorporate movements – such as the environmental and labor movements – placed public opinion in direct opposition to corporate interests. Some of the consequences of this situation are that TNCs' objectives are, now more than ever, diverging from those of nation-states. Because of this emerging hostility, TNCs are often forced to operate to counter nation-states' behaviors that are damaging to their interests. This posture does not only pertain to the company itself, but also to the behavior of managers.

The view that TNCs are better world citizens than otherwise believed and that they act to defend their legitimate interests in hostile political environments is also shared by scholars who identify with the political left. Gavin Kitching (2001), for instance, articulates a 'qualified defense' of TNCs (2001: p. 42) whereby he contends that not only the number of TNCs (or 'supranationals' as he calls them) engaged in behavior that causes negative social and economic consequences in host countries is greatly exaggerated, but that this behavior is the result of structural conditions that ultimately dictate TNCs' actions. Supranationals, he argues, are far better employers than many 'local' companies as they provide better wages, working conditions, and obey local laws. One of the reasons for this behavior is that TNCs can afford to be more 'generous' than local companies. In effect, the cost of labor in developing regions is very much cheaper than in developed countries that it is still a bargain for TNCs to pay high 'local' wages. Also,

because TNCs lack local political legitimacy, they need to be on their best behavior and appear as model employers to win over the trust of host governments (Kitching, 2001: p. 43). Yet, it is the absence of a global regulatory entity (i.e., the absence of a global state that would compensate for the declining hegemony and capacity to control world industrial and political relations of the United States) that constitutes the most important structural reason behind the often publicized 'negative' corporate behavior. Due to the lack of this global regulatory entity and the consequent political and economic instability and uncertainty, TNCs tend to give preference to short-term goals (profit) over a more balanced and constructive local presence. It is the structural gap between a world economy and the local polity that ultimately affects TNCs' behavior. Corporations, he concludes, cannot change this situation as it can be altered only at the political level through the creation of a world political system that could regulate the economy and provide the basic conditions for social stability. For Kitching, the spread of TNCs presence worldwide creates new opportunities for developing regions and workers. While jobs may be 'lost' in developed regions with the consequent protest by local unions and left-leaning political groups, new jobs are created in less developed regions. He contends that if one job created is equal to one job lost, the fact that jobs reappear in the less developed South is a positive sign in the struggle to reduce socioeconomic inequality worldwide. He dismisses as unsubstantiated the objection that the creation of new jobs in less developed regions often does not translate into the actual lifting of workers out of the subsistence level. He contends that currency instability along with government's attempts to maintain export competitiveness depress wages and do not eliminate the threat of market withdrawal through plant relocations.

The growing power of TNCs has characterized globalization since its outset and has been a persistent theme in pertinent literature. In current debates, there is consensus that TNCs will continue to be powerful actors in the economy and society. Disagreement, however, exists in regard to (1) the extent to which TNCs have the same level of power that research attributed to them a decade ago and (2) the consequences generated by the application of this power. According to one camp (i.e., Busse, 2004; Macleod and Lewis, 2004), TNCs will continue to be powerful actors and their ability to control nation-states, social groups, and economic policies will continue unabated. A second camp consists of authors who argue that the power of TNCs is not as absolute as some analyses would contend. While TNCs remain powerful actors in the global context, their power is resisted at various levels. Resistance emerges from a number of actors including nation-states, local groups, and social movements. A number of studies have underscored the role that consumers play in shaping TNCs' actions (Humphery, 1998; Marsden, 2003; Wright and Midderndorf, 2007). It is argued that consumers' behavior is increasingly affected by social values – such as the protection of the environment; health concerns; the protection of workers and communities in less developed areas; fair processes of trade – rather than price. Accordingly, as consumers' social consciousness affects demand, it also affects the behavior of corporations and their power. It is further argued that the power of TNCs is also limited by internal organizational constraints and contradictions (Bonanno and Constance, 2008). Because they compete with other corporations and have relatively limited organizational strength given the scope of their actions, TNCs encounter difficulties in exercising power at the global level.

See also: Capitalism: Global; Capitalism: The History of the Concept; Capitalism; Culture and Economic Development; Development Geographies; Development Studies; Development Theory in Geography; Local Economic Development; Migration and Development; Multinational Corporations; Nations and Nation-States in History; Neoliberalism.

Bibliography

Aglietta, M., 1979. A Theory of Capitalist Regulation. New Left Books, London.
Ajami, F., 1972. Corporate giants: some global social costs. International Studies Quarterly 26 (4), 511–529.
Amoore, L. (Ed.), 2005. The Global Resistance Reader. Routledge, New York.
Arrighi, G., 1998. Globalization and the rise of East Asia: lessons from the past, prospects for the future. International Sociology 13 (1), 59–77.
Arrighi, G., 2005. Globalization and the world system perspective. In: Appelbaum, R., Robinson, W. (Eds.), Critical Global Studies. Routledge, New York, pp. 33–44.
Barber, B., 1995. Jihad vs. McWorld: Terrorism's Challenge to Democracy. Ballantine Books, New York.
Barnet, R.J., Muller, R.E., 1976. Global Reach: The Power of the Multinational Corporations. Touchstone, New York.
Baudrillard, J., 1983. Simulations. Semiotext(e), New York.
Becker, D., Sklar, R. (Eds.), 1999. Postimperialism and World Politics. Preager, Westport, CT.
Bellamy, F.J., McChesney, R.W., 2012. The Endless Crisis: How Monopoly-Finance Capital Produces Stagnation and Upheaval from the USA to China. Monthly Review Press, New York.
Bonanno, A., Cavalcanti, J.S.B., 2011. Globalization and Time-Space Reorganization. Emerald Group Publishing, Bingley, UK.
Bonanno, A., Constance, D., 2008. Stories of Globalization. Transnational Corporation, Resistance and the State. The Pennsylvania State University Press, University Park, PA.
Boggs, C., 2000. The End of Politics. Corporate Power and the Decline of the Public Sphere. The Guilford Press, New York.
Busse, M., 2004. Transnational corporations and repression of political rights and civil liberties: an empirical analysis. Kyklos 57, 45–65.
Chase-Dunn, C., 1998. Global Formation: Structure of the World-Economy. Rowman & Little field, Lanham, MD.
Constance, D.H., Heffernan, W.D., 1991. The global poultry agro-food complex. International Journal of Sociology of Agriculture and Food 1 (1), 126–142.
Croucher, S., 2004. Globalization and Belonging: The Politics of Identity in a Changing World. Rowman & Littlefield, New York.
Derrida, J., 1978. Writings and Differences. University of Chicago Press, Chicago.
Dicken, P., 1998. Global Shift. The Guilford Press, New York.
Dunning, J., 1993. The Globalization of Business. Routledge, New York.
Firebaugh, G., Goesling, B., 2004. Accounting for the recent decline in global income inequality. American Journal of Sociology 110 (2), 283–312.
Friedman, T.L., 2000. The Lexus and the Olive Tree. Anchor Books, New York.
Friedman, T.L., 2005. The World is Flat. Picador, New York.
Foucault, M., 1969. The Archaeology of Knowledge and the Discourse of Language. Harper Colophon, New York.
Foucault, M., 1975. The Birth of the Clinic: An Archaeology of Medical Perception. Vintage Books, New York.
Giddens, A., 2000. Runaway World: How Globalization is Reshaping Our Lives. Rutledge, New York.
Gramsci, A., 1971. In: Hoare, Quintin, Smith, Geoffrey Nowell (Eds.), Selections from the Prison Notebooks. International Publishers, New York.
Guillen, M.F., 2001. Is globalization civilizing, destructive or feeble? A critique of five key debates in the social science literature. Annual Review of Sociology 27, 235–260.
Habermas, J., 1975. Legitimation Crisis. Beacon Press, Boston.
Harvey, D., 2005. A Brief History of Neoliberalism. Oxford University Press, New York.

Harvey, D., 2006. Spaces of Global Capitalism. Verso, London.
Harvey, D., 2010. The Enigma of Capital. Oxford University Press, New York.
Hymer, S.H., 1971. The multinational corporation and the international division of labor. In: Cohen, R.B. (Ed.), The Multinational Corporation a Radical Approach. Cambridge University Press, Cambridge, UK, pp. 47–65.
Hirst, P., Thompson, G., 1996. Globalization in Question. Polity Press, Cambridge, UK.
Hopper, P., 2007. Understanding Cultural Globalization. Polity Press, Cambridge, UK.
Hosseini, H., 2009. Global complexities and the rise of global justice movement: a new notion of justice. The Global Studies Journal 2 (3), 15–36.
Humphery, K., 1998. Shelf Life. Supermarkets and the Changing Cultures of Consumption. Cambridge University Press, Cambridge, UK.
Jagdish, B., 2004. In Defense of Globalization. Oxford University Press, New York.
Kitching, G., 2001. Seeking Social Justice through Globalization: Escaping a Nationalist Perspective. University Park, PA.
Krugman, P., 2007. The Conscience of a Liberal. W.N. Norton, New York.
Laclau, E., Mouffe, C., 1987. Post-marxism without apologies. New Left Review 166, 79–106.
Laclau, E., Mouffe, C., 1985. Hegemony and Socialist Strategy: Towards a Radical Democratic Politics. Verso, London.
Levitt, T., 1983. The globalization of markets. Harvard Business Review 61 (3), 92–102.
Lipietz, A., 1987. Mirages and Miracles. Verso, London.
Lipietz, A., 1992. Towards a New Economic Order: Post-fordism, Ecology, and Democracy. Oxford University Press, New York.
Macleod, S., Lewis, D., 2004. Transnational corporations – power, influence and responsibility. Global Social Policy 4 (1), 77–98.
Marsden, T., 2003. The Condition of Rural Sustainability. Royal Van Gorcum, Assen, The Netherlands.
Marx, K., Engels, F., 1998[1848]. The Communist Manifesto. Verso, London and New York.
Milanovic, B., 2011. The Haves and the Have-Nots: A Short and Idiosyncratic History of Global Inequality. Basic Books, New York.
O'Connor, J., 1974. The Fiscal Crisis of the State. St. Martin's Press, New York.
Ohmae, K., 1995. The End of the Nation State: The Rise of Regional Economies. The Free Press, New York.
Parsons, T., 1971. The System of Modern Societies. Prentice-Hall, Englewood Cliffs, NJ.
Ritzer, G., 2008. The McDonaldization of Society. Pine Forge Press, Los Angeles, CA.
Robinson, W., 2004. A Theory of Global Capitalism. Production, Class and State in a Transnational World. The Johns Hopkins University Press, Baltimore and London.
Rostow, W.W., 1960. The Stages of Economic Growth: A Noncommunist Manifesto. Cambridge University Press, Cambridge, UK.
Rubner, A., 1990. The Might of the Multinationals: The Rise and Fall of the Corporate Legged. Praeger, New York.
Sklair, L., 2001. The Transnational Capitalist Class. Blackwell, Oxford, UK.
Sklair, L., 1993. Assembling for Development: The Maquila Industry in Mexico and the United States. UCSD Center of U.S.-Mexican Studies, San Diego, CA.
Stiglitz, J., 2003. The Roaring Nineties. W.W. Norton & Company, New York.
Volscho, T.M., Kelly, N.J., 2012. The rise of the super-rich: power resources, taxes, financial markets, and the dynamics of the top 1 percent, 1949 to 2008. American Review of Sociology 77 (5), 679–699.
Wallerstein, I., 1974. The Modern World-System: Capitalist Agriculture and the Origins of the European World Economy in the 16th Century. Academy Press, New York.
Wallerstein, I., 1980. The Modern World System II: Mercantilism and the Consolidation of the European World-Economy, 1660–1750. Academy Press, New York.
Wallerstein, I., 2000. Globalization or the age of transition. International Sociology 15 (2), 251–267.
Wallerstein, I., 2005. After development and globalization, what? Social Forces 83 (3), 1263–1278.
Wettstein, F., 2009. Multinational Corporations and Global Justice: Human Rights Obligations of a Quasi-Governmental Institution. Stanford University Press, Stanford, CA.
Weber, M., 1963. The Sociology of Religion. Beacon Press, Boston.
Wolf, M., 2004. Why Globalization Works. Yale University Press, New Haven, CT.
Wright, W., Midderndorf, G. (Eds.), 2007. The Fight over Food. Penn State University Press, University Park, PA.

Globalization and World Culture

John Boli and Frank J Lechner, Emory University, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores the origins, development, structure, and impact of world culture in relation to globalization over the past two centuries. It discusses the complex and largely rationalized content of world culture and the rise and proliferation of organizations that generate, debate, and propagate world culture, particularly international nongovernmental organizations. It describes world-cultural processes that promote homogenization and heterogeneity at the global, national, and local level, and the consequences for national identity of world-spanning cultural development. The article concludes with a discussion of the role and significance of religions and religiosity in a globalized world.

Globalization and the Construction of World Culture

Globalization entails the spread of conceptions of the world as a single entity – for example, a single and highly complex society, market, polity, technical system, or culture (Robertson, 1992). At a more concrete level, globalization entails the many organizations, associations, and social movements of global orientation that structure worldwide flows of goods and services, people, knowledge, religions, ideologies, and much more. World culture refers to the complex array of foundational assumptions, forms of knowledge, and prescriptions for action that underlie globalized flows, organizations, and movements (Lechner and Boli, 2005). Its webs of significance span the globe, providing knowledge, rules, policy prescriptions, conceptions of world society, purposes of human action, and models and methods of structuring social life that are assumed to have worldwide significance or applicability.

Some examples of the content of world culture will help convey its meaning. The rules of chess are elements of world culture; the game is defined by a standard set of procedures that are deemed valid everywhere, and departures from those rules are considered nonsensical. The principles of physics are world-cultural, assumed to apply at every point on the globe and even beyond. Human rights, as codified in the Universal Declaration of Human Rights and its associated conventions, are couched in world-spanning terms, applying to all humans in all countries. The technical standards (ISO/IEC 4909:2006) for the dimensions, thickness, tensile strength, and other properties of financial transaction cards produced by ISO, the International Organization for Standardization, are presumed to be useful and technically rational for any maker or issuer of such cards. Other examples of world culture include the 'model subcontract' produced by the International Chamber of Commerce, a template for international business relations; the pathologies identified in the World Health Organization's *International Statistical Classification of Diseases and Related Health Problems*; and the Six Sigma management methodology that is promoted worldwide as a means of improving business performance. What these examples have in common, despite their highly disparate social fields, is the presumed universality that is associated with them. They are elements of the culture of the globalized world, their proponents seek to spread them globally, and they are presumed to have meaning or value wherever they may be implemented.

World-cultural elements are not necessarily embraced or accepted universally, however. Like the cultures and subcultures of national societies, world culture is replete with competing cultural elements and contradictory principles. For example, democratic governance principles are strongly embedded in world culture, but many varieties of democracy are available and some segments of world society favor nondemocratic regime forms. The major world religions are elements of world culture, but they disagree sharply on such fundamental issues as the existence and nature of divinity, the purposes of life, and the fate of the soul. Some elements of world culture are hotly debated, such as the International Panel on Climate Change's claim that human action contributes to global warming and the neo-liberal economic claim that free markets are the best means of raising living standards. Hence, while many elements of world culture are largely taken for granted, many other elements are contentious, contradictory, or controversial.

World culture is the product of many forms of social activity and many types of social actors, reaching far back in human history. In its contemporary form, world culture began to coalesce in the second half of the nineteenth century. We can trace its emergence through the rise of international nongovernmental organizations (INGOs), which are the key structural form through which world culture is expressed (Boli and Thomas, 1999). The earliest of these voluntaristic, purposive, specialized associations emerged in the first half of the nineteenth century, but it was not until the 1870s that their numbers began to grow rapidly. It was also at this time that the range of human activities they address broadened widely, eventually covering almost every imaginable social field. INGOs are especially important with respect to world culture because they operate explicitly at the transnational level and much of their work deals directly with world-cultural principles and prescriptions. Science INGOs develop and propagate scientific principles; professional INGOs propagate knowledge and ethical principles relevant to their respective professions; sports INGOs codify and promote 'rules of the game' and principles of sportsmanship; and education INGOs promote teaching methods, curricular development plans, and policies to protect teacher autonomy in the classroom. Across the

board, these principles and norms are couched in universalistic terms, and universalism was strikingly evident in the outlook and work of the vast majority of INGOs that formed before the mid-twentieth century. These bodies welcomed members from all parts of the world and presumed that their knowledge, ethics, and prescriptions were applicable all over the world.

Since World War II, world-cultural structuration has become more complex; a layered transnationalism has appeared as regional organizations have proliferated. By one count, roughly half of the transnational organizations founded since 1950 have been regional in scope, activating a variety of types of regions: geographic (European, Latin American, Asian), linguistic (francophone, hispanophone) religious (Islamic, Christian, Jewish), climatic and vegetational (tropical areas, rain forest zones), and so on. This regional intensification has occurred within the larger global cultural frame, and it characterizes the internal structure of global organizations as well. For example, the World Health Organization (founded in 1948) eventually developed regional bodies for each major part of the world (the Americas, Africa Southeast Asia, etc.), while ISO helped with the establishment of regional standardization bodies (e.g., the European Committee for Standardization, the Pacific Area Standards Congress) that coordinate their activities with those of the larger global organization. The same process is common in global corporations. For example, CNNI (Cable News Network International) began as a unitary organization bringing a standardized English-language product to televisions around the world, but it eventually established numerous regional centers producing local-language broadcasting adapted to specific target audiences but tied closely to CNNI's overarching global operation.

An important mechanism for the spread of world culture in the post-war period was the rapid decolonization that yielded more than a hundred new states covering a large portion of the world's land mass. As colonies, these societies had been only weakly integrated into world culture, however deeply they may have been integrated into the world economy. Most new states have eagerly joined large numbers of intergovernmental organizations (IGOs), interacting with older states as they learn about and debate world-cultural principles and prescriptions. Citizens of the new countries became avid joiners of INGOs, expanding the range of their memberships much faster than that of citizens of older countries. Thus, political independence opened the door to much greater participation in world society by the formerly colonized populations, exposing them to world culture to a much greater degree than previously. This process works both ways; these newly empowered world citizens began contributing new or refashioned elements to world culture, propagating them through the same channels by which world culture reaches them but with an increasing flow of cultural elements in the reverse direction.

Dimensions of World Culture

Most discussions of world culture focus on its expressive or normative dimensions. Expressive (popular) culture includes media products, consumable goods that often attain iconic status as symbols of modernity or avant-gardism, and foods and clothing styles originating in particular cultures that become worldwide fads. Normative culture comprises the values and desiderata promoted by globally active organizations, be they corporations, foundations, IGOs, or INGOs: fetishistic consumerism and neoliberal economic ideologies, democracy and human rights, religious doctrines and world views, gender and racial equality, environmental stewardship, and so on. The expressive dimension has evoked intense debate regarding the purported homogenizing effects of world culture (discussed below). The normative dimension is most apparent in political struggles regarding the purposes and policies of states, major global governance IGOs such as the World Trade Organization (WTO) and the International Monetary Fund (IMF), and global corporations. Critics of these powerful global actors invoke world-cultural norms and principles in decrying, for example, the refusal of the WTO to consider working conditions in trade agreements, the harm done to the poor by IMF structural adjustment programs, and the damage done to the environment by mining companies. These normative clashes are often widely publicized – challenges to power centers tend to be noticed – so normative claims rooted in the global moral order are the most widely recognized elements of world culture, and they have drawn much scholarly attention as well.

Much less attention has been paid to the ontological and cognitive dimensions of world culture. At the ontological level, world culture is built on assumptions about the fundamental reality of the individual, the disenchantment of nature (the natural world is not inhabited by spirits but operates according to scientifically knowable laws), the usefulness and efficiency of rationalized organization and management, the indispensability of centralized authority structures (states), and so on (Meyer et al., 1987). More specific ontological assumptions stipulate, for example, that individuals are the primary units of action in society, they are educable, they are capable of great accomplishments, and they are subject to many types of physical and psychological maladies that are increasingly well specified. In short, world culture offers definitions of 'the nature of things' – the basic elements of physical and social reality, the technologies available to manipulate these elements, and theories about causal processes operative among them. These theories assure states, corporations, and other organizations that, for example, such major modern institutions as formal education, experimental science, national accounting systems, and advanced-technology engineering are effective means of realizing human goals. Much of this ontological and cognitive material is highly institutionalized and thus largely taken for granted. Science, mathematics, bookkeeping, management techniques, electronic technology, medical knowledge, land surveying methods, principles of roadway design – all these, in world culture – are deemed unassailably useful and effective tools. Their universal applicability is not easily challenged, and they are strongly supported by states, professional societies, consulting companies, and other organizations that deal with rationalized aspects of human life.

Despite their foundational importance for human understanding of the nature of (world) society and the possibilities and prospects of human endeavors, the ontological and cognitive dimensions of world culture are largely ignored by both the scholarly and general publics. These cultural elements are uninteresting, even invisible, for two reasons: one, they operate far below the surface of everyday life; two, they deal

with mundane, routine, matter-of-fact aspects of reality. Ontological assumptions are, by definition, hidden from plain sight; they tell us what kind of world we live in and what is possible in that world, providing the givens of the reality that underlies daily activity. Cognitive elements are, in most cases, produced and elaborated by specialized experts and professionals whose vocabularies and conceptual apparatus are not accessible to nonspecialists. Thus, these highly formalized and rationalized arenas seldom become the focus of ideological debate or controversy, and they therefore are largely absent from the global public realm. Notable exceptions are easily adduced – nuclear power, genetically modified organisms, and global warming have garnered much public attention – but most highly institutionalized forms of cultural globalization stay under the radar.

A third major dimension of world culture is the implicit moral order that underlies such global ideologies as human rights, progress, citizenship, and nationalism. While this dimension cannot be studied directly, it makes good sense to infer that the moral order of world culture embraces the individual as a sacred entity that is to be protected and empowered. The nation is also sacred, in considerable degree, as are national and indigenous cultures, but most collectivities – organizations, corporations, associations, classes, and so on – do not share this exalted status. Virtue is attributed to action, technologies, and structures that enhance or enable sacred entities, while vice is espied whenever these entities suffer injury, degradation, betrayal, or destruction. Variable conceptions of the specific elements of virtue and vice surely flourish in world culture, but a large number of claims about definitions of the good, the just, and the righteous are surprisingly free of controversy or debate throughout world society.

Origins and Expansion of World Culture

Many types of global flows have been rising cyclically for ages, but it was only in recent centuries that a transnational cultural complex took a sufficiently organized form to constitute a recognizable world culture. This crystallizing complex, derived in substantial part from the highly structured transnational culture of the Catholic Church, was primarily European, and from the sixteenth century the imperial Western European countries extended it to most corners of the globe. Europeans promoted political and philosophical principles, societal and individual goals, modes of organizing, and ways of conceiving and manipulating reality that they deemed universally applicable. Epitomized above all by technical and scientific principles and practical knowledge presumed to be invariant across time and space, emerging world culture also included more historically bound constructs and ideologies, such as citizenship, individualism, principles of international law, aesthetic sensibilities, and literary forms. This early version of world culture, perhaps more properly called 'transeuropean' culture, was carried far and wide by missionaries, traders, military expeditions, colonialists, intellectuals, and travelers.

The Western Church played a crucial role in world culture's formative period. The Church was the preeminent transnational organization of the medieval and early modern period. It was the only highly organized structure spanning a large geographical area and a great many local cultures. This structure was the principal means of long-distance communication, spreading all sorts of cultural elements throughout Christian lands. The Church embedded a decidedly transnational culture through the medium of Latin, which was the only widely spoken language, and its personnel circulated among the cities and towns of Europe with surprising facility. Alongside the Church proper, the religious orders that sprang up with such profligacy after the revival of the Church in the twelfth century can be seen as the first wave of INGO formation, in that the orders typically relied on voluntary membership, personal commitment, a transnational outlook, and specialized purposive action to achieve their goals.

In a sense, then, the Church was the original locus of a proto-world culture that, thanks to unparalleled European imperialism, became an important part of truly global culture. The Church provided motivations for some globalizers (missionaries and religious orders) and legitimations for others (traders, conquerors, colonizers) to incorporate the entire world in the nascent world society. It also contributed significantly to the cultural underpinnings of contemporary world culture. The conception of societal progress that prevails almost everywhere stems in part from the Christian legitimation of rational world mastery. The view of the individual as both sacred being and active citizen is rooted in Christian views of the individual, particularly after the Reformation when even Catholics began to stress the importance of individual belief and literacy for salvation. The global script for organizing society in a rational-legal manner derives from the Church's highly successful bureaucracy. More concretely, missionaries not only spread their faith, they also provided education and rudimentary health care in far-flung places, diffusing secular commitments that have since become globally entrenched. Today, distinctly religious views are less obviously influential in world culture yet they still shape its evolution (discussed further below). A vigorous Pentecostal movement, rapidly expanding in Latin America, Asia, and Africa, carries with it the gospel of material success and individual well-being, thus supporting global corporate culture. In counterpoint, the World Council of Churches contributed centrally to global environmental concerns with its initiative on 'Justice, Peace, and the Integrity of Creation,' launched in 1983.

The striking crystallization of world culture in the nineteenth century was evident above all in the rapid proliferation of INGOs, which became the 'structural backbone' of world culture. Increasingly interconnected with each other and with states and business organizations, these carriers of world culture formed an emerging 'global civil society' of voluntaristic, self-governing associations whose members took it upon themselves to organize the world in specific social domains. Typical early examples include the International Charity Association (founded in 1855), the International Sugar Union (1864), the Scandinavian Dental Association (1866), the Permanent International Committee of Architects (1867), and the International Meteorological Organization (1873). By the 1890s, INGOs were appearing at the rate of more than 10 per year across a wide range of social sectors, drawing participants mainly from Europe and the Anglo-American former colonies but also from Latin America and some Asian countries, particularly India. These bodies defined their arena of action as

the world entire, sponsoring conferences, publishing journals and newsletters, and entering into public debate about universalistic issues, problems, methods, and proposed solutions. They became prominent participants in a formalizing global public realm in which world culture was defined, documented, elaborated, and propagated to what the fast-growing number of INGO participants had come to think of as a single world society.

The cataclysmic world wars of the twentieth century severely interrupted world-cultural development but only temporarily. INGOs continued to proliferate between and after the wars, and states became increasingly engaged in transnational cooperative relationships through IGOs. The IGO population eventually burgeoned to comprise hundreds of organizations, while the number of INGOs soared into the many thousands (by the broadest definition, INGOs now number in the tens of thousands). For a good many social sectors, this expanding complex of global organizations came to center on the United Nations, whose agencies and programs emerged as core elements of global governance regimes in such major institutional areas as education, health, environmental protection, and development. INGOs flocked around the UN and many other IGOs to try to influence their programs, policies, and principles, and the confluence of these many forms of organization produced a high degree of formalization regarding many aspects of world-cultural development. This formalization proceeds apace, engaging all states and increasingly large proportions of the world's population.

Homogenization, Heterogeneity, and Conflict in World Culture

Until the late 1990s, globalization was generally seen as a strongly homogenizing cultural force. Globalization was 'Americanization' or 'Westernization', the implantation of the products and cultures of the dominant power centers on the rest of the world. The global popularity of American films, television shows, sports figures, and consumption patterns was undeniable. Western music, soap operas, blue jeans, Coca-Cola, fast foods, and morally lax lifestyles were romping unhindered around the world, and many analysts assumed that the market success of such products necessarily entailed the refashioning of local cultures in line with American or Western tastes and values. Many a worried eyebrow was raised about the spread of an ideology of materialistic consumerism that was obliterating local cultures.

This viewpoint proved too facile (Berger and Huntington, 2002; Appadurai, 1996). Homogenization is certainly evident in some respects: more people in more places are watching television, seeing films, buying brand-name products, and obsessing about one or another pop star. But when anthropologists and sociologists take a close look at the reception of Western cultural products, they find not passive, easily brainwashed consumers but active, engaged receptors who interpret cultural imports in highly varied ways, often with skeptical or dismissive attitudes rooted in local cultures and traditions. Local cultures are not easily abandoned; much creative interpretation and adaptation of cultural imports takes place (Cowen, 2002). Local cultures integrate new cultural elements while retaining their basic identities, in a process known as hybridization (Nederveen Pieterse, 2009). Cultures become more complex and internally contradictory, and hybridization evokes much culture work to make sense of and pacify the contradictions and tensions that result.

One striking implication of this viewpoint, particularly when conjoined with the massive human migration of recent decades, both voluntary and involuntary (refugees fleeing war, famine, and political unrest), is the sharp upswing in cultural and linguistic diversity in the world's cities. In Seattle and Rio de Janeiro, in Amsterdam and Budapest, in Dubai and Mumbai and Bangkok and Sydney, ethnic groups from all continents have become more numerous, more organized, and more active in local commercial and social life. Their restaurants blossom, their community newspapers multiply, their theater troupes become increasingly visible, their languages stand out on store fronts. The net result is a paradoxical form of homogenization: the world's cities are becoming ever more diverse but they are ever more familiar to visitors because they are diversifying in similar ways. This is a telling reflection of world culture in general: it is increasingly complex, diverse, and incoherent, as more cultural elements from a wider range of cultures become globally known, imported, and reinterpreted as they are embedded in local cultures.

The reinterpretation process, often labeled glocalization (Robertson, 1992), implies that even highly standardized elements of global popular culture are subject to varied uses and meanings. For instance, consumers in Hong Kong appreciate McDonald's restaurants as much for their clean facilities as for their food. They expect fast, not smiling, service, and they reserve the right to eat at their leisure. In a city pressed for space, the restaurant has become a gathering spot for youth and a favored locale for birthday parties. In Beijing, McDonald's American origins are attractive to upwardly mobile Chinese as a symbol of participation in global society, but to many Koreans in Seoul those same origins evoke criticism. Popular culture increasingly creates global tastes, but highly variable consumer bases keep this culture from becoming entirely uniform (Watson, 2006).

An anthropological assessment of this situation leads to the conclusion that the world's cultures (national and local, as well as global) are creolizing (Hannerz, 1996) – every locale is an ever greater mélange of cultural elements from many sources. A case in point is West African popular music, an amalgamation of variegated genres (Senegalese folk music, French *chanson*, American blues, British rock, and more) that constitute new and distinctive styles. Retaining indigenous instruments and rhythms, African musicians are not simply on the receiving end; their styles flow back into world culture as well, influencing some of the very traditions from which they borrow. Given the global valorization of difference and authenticity (below), commercial interests intensify this process by promoting diversity widely (UNESCO, 2000). Thus, as new musical and other fads make their way across the globe, they produce new forms of localized diversity and unforeseen local reactions and interpretations. In the expressive realm, the varied forms of local–global interaction are hardly resulting in a standardized, stifling, hegemonic world culture.

Despite this rising diversity, in many other ways world culture is decidedly homogenizing. The complex array of global

organizations that embody, debate, and propagate world culture have produced a wide range of standardized principles, models, and methods for the organization of social life. For example, a basic model of the modern state has crystallized (Meyer et al., 1997). The model is reflected directly in national constitutions, which reveal a remarkable degree of similarity despite some striking variability (regarding, for example, the relationship between religion and the state). It is institutionalized globally through international agreements, such as International Labor Organization conventions regarding labor–management relations, global health programs of the World Health Organization, women's rights declarations, and national accounting principles promoted by the International Accounting Standards Board. The basic model stipulates that the state is responsible not only for internal order and external defense but also for building a modern society and promoting citizen welfare and civic engagement. The state is to organize, regulate, or provide national educational systems, health care programs, retirement and unemployment insurance schemes, and cultural promotion programs, among many other duties. In many sectors, globally legitimated models have emerged to guide states in meeting these responsibilities, such as the standardized educational models developed under UNESCO's aegis in collaboration with educational INGOs and the standardized models for the regulation of stock exchanges promulgated by the International Organization of Securities Commissions.

Although states vary greatly in terms of their historical experiences, resources, and organizational capacity, they have implemented these models in remarkably uniform ways, in a process known as structural isomorphism (Meyer et al., 1997). Isomorphism has been identified in such diverse areas as citizen rights and state authority, school systems and curricula, environmental protection, science policy, children's rights, women's empowerment, same-sex sexual relations, and national planning. Few states successfully implement the models across all these domains, and many states adopt some or many features of the models structurally but make little attempt to actualize them in practice. Nonetheless, the isomorphism process has made states much more similar to one another than one would predict on the basis of the extremely varied capabilities, cultures, and historical experiences of their national societies.

Thanks in large part to the efforts of IGOs and INGOs, standardized global models of organizations have also coalesced in world culture. They can be found across an enormous set of domains: in business, science, medicine, health, technology, recreation and leisure (sports, tourism, entertainment), and much more. Pushed by the carriers of these models (which include business consultants, technical advisers, and academic consultants), corporations, hospitals, research institutes, social movement organizations, and many other types of social units implement the models in largely uniform ways. They also engage themselves in a learning process, constantly monitoring new developments in world culture so they can keep up with the leading edges of world-cultural development. The net result is both a tremendous force for the homogenization of the world and a persistent tendency for organizations in particular social domains to ride the waves of organizational fashion. Businesses are wont to adopt, quite uncritically, 'hot' management techniques, such as Total Quality Management or the Program Evaluation and Review Technique. When such methods go out of fashion, businesses replace them with new 'state-of-the-art' techniques whose usefulness or applicability is more presumed than demonstrated.

In sum, world culture generates both greater diversity and greater homogeneity, in world society as a whole and within national societies. The technical and rationalized forms of homogenization are accepted without much resistance, but expressive and normative homogenization meets with considerable suspicion and, not infrequently, outright hostility. Expressive/normative homogenization is seen as a threat to national and local cultures, and defenders of these cultures are found not only within the threatened societies but also in the rich, powerful (Western) societies that are generally assumed to be the source of the threat. Resistance is often grounded in the principle of cultural relativism (itself of Western origin) and its associated ideology of cultural authenticity. Relativism proclaims the fundamentally equal value of all human cultures. Ethnocentrism is seen as both an intellectual trap and a cause of injustice, while tolerance and, even more, the championing of difference occupy the moral high ground. Outsiders are in no position to judge foreign cultures because they do not understand the internal logics of these cultures or the interconnections among particular cultural elements that make the parts indispensable to the larger wholes. A natural implication of this viewpoint is the claim that, above all, cultures should retain their primordial authenticity. Authentic cultures are to be respected, even revered.

The emphasis on cultural relativism and authenticity undergirds a sense of heightened moral virtue associated with the poor, the excluded, the oppressed, that is, marginal peoples whose right to their own cultures is jeopardized by the onslaught of globalizing world culture. This universalistic form of particularism impels peoples to emphasize or invent traditions and distinctiveness that contrast with the universalistic world-cultural principles that are supposed to operate uniformly in all places. In Western societies, such particularism is stimulated by cosmopolitan connoisseurs seeking authentically exotic cuisine or ethnically distinct 'world' music. In other places, politically astute groups understand the power of reified cultural authenticity as a fulcrum for leveraging rewards from global systems. Hence, distinction and difference have become strategic resources for collective mobilization. In consequence, ethnonationalist movements of many sorts (Basques, Biafrans, Bretons, Croats, Quebecois, Scots, Tamils, etc.) and multiculturalist restructuring by states, churches, schools, and other organizations have become the order of the day.

As this discussion indicates, world culture engenders conflict in world society through a process that has, as it were, turned the West against itself (Leclerc, 2000). By the 1970s, when scores of new states had formed in Africa and Asia, vociferous opposition to continued economic and cultural domination by Western countries (labeled neo-colonialism and cultural imperialism, ideas that also originated in the West) began to penetrate global organizations, especially UN bodies. Various associations of less developed countries called for a restructuring of world society, the fashioning of a New World Economic Order and a New World Information Order that would put restraints on the operations of transnational

corporations and shift resources to the poorer countries. Former colonies began assertively invoking world-cultural principles of equality and development as basic human rights. Further tensions arose as the global human rights sector (itself increasingly involving individuals and groups from outside the West) turned its scrutiny on the new states, decrying their selective championing of certain global universals while ignoring or severely violating others, especially those relating to the integrity and political participation of citizens. Thus, both the particularism of world culture (the moral legitimacy attributed to national or ethnic units) and its universalism (the insistence that all units abide by basic principles and values) yield forms of disagreement and conflict that would not emerge in a less globalized world (Robertson, 1992; Lechner and Boli, 2005).

As non-Western cultures and regions have become more prominent in the world polity, it makes sense to speak of world cultures (in the plural) rather than a singular world culture. Dominant Western models have penetrated deeply in most places, but they have also evoked resistance as well as efforts to revivify and globalize alternative models (Leclerc, 2000). Most notable in this regard is the assertiveness of Islamic cultural carriers, particularly since the 1970s (Roy, 2004). Many Muslim leaders and organizations promote a societal model that infuses the state with religious precepts and recasts the relationship between state and citizen (codified in 1981 as the Universal Islamic Declaration of Human Rights). African and Asian models of social organization and development have also emerged, and some observers argue that conflict in the twenty-first century will revolve primarily around grand civilizational axes rather than the nation-state clashes that have dominated in recent centuries. By the same token, more fine-grained analyses of world culture(s) identify multiple models of central world-cultural constructs. For example, derived from the Western tradition are liberal, socialist, corporate, and welfare models of the state; from Asian sources, quasi-familial and state-led development models. Multiple models of the individual, the business enterprise, and the national polity are further examples. World-cultural complexity has increased rapidly in recent decades, perhaps most sharply since the collapse of Communism and the end of the Cold War, as more cultural centers generate more alternative cultural constructions. Poorer and more peripheral societies are less able to bring their cultural models to the world-cultural table, but many participants in the global arena from richer societies have become strong advocates of the poor and marginalized, helping to ensure that world culture becomes still more complex and incoherent but also more significant for national and local structuration and change.

Globalization, World Culture, and National Identity

Globalization creates ties across space and is therefore often equated with 'deterritorialization'. This feature would seem to spell trouble for nation-states: if globalization transcends national boundaries, states' role in world society should diminish. Focusing on the cultural side of the matter, however, it is clear that the scenario of decline is overstated. As indicated earlier, while globalization challenges nation-states in many ways, the idea of the nation-state is deeply embedded in world culture. World culture provides models of nation-state functioning and norms that legitimate states as constitutive units of world society.

This is no accident: the nineteenth-century wave of globalization occurred during an era of mostly Western nation-building, binding global processes to far more integrated 'national' societies, each with its own claim to cultural distinction. Over time, practice turned into principle, codified by organizations like the United Nations. Newly independent countries mostly adopted the established framework, reinforcing the legitimacy of a single global model for societal organization. The model allowed for considerable variation, for example in the extent to which nation-states comprised actual nations of a single people with a single culture. The striking world-cultural fact is that even highly diverse places witnessed a drive toward greater unity and the development of a distinct national identity, as in the case of Indonesia adopting a new national language.

Homogenization at the national level boosts the institutionalization of global diversity, since each nation's claimed identity differs in principle from those of other nations. A number of global practices endorse and enshrine the value of diversity, from a tourist industry seeking distinctive destinations to UNESCO designating world heritage sites. In this way, globalization 'universalizes particularism'. This is more than a legacy. State institutions continue to reinforce national identities, for example, through formal educational systems that follow a common global script. Important global events, such as the Olympic Games, serve as occasions for the display of distinct identities. Since practices, institutions, and events also press toward commonality, requiring more people in more places to play by the same rules, the reproduction of national identity is yet another instance of the globalizing interplay between homogenization and heterogenization.

Because national identities are so salient and highly legitimated, the glocalization of cultural forms and experiences often involves creative mixing of transnational and domestic elements, as in the case of popular music genres that match national symbolism to global (frequently American-initiated) forms, such as German or Japanese hip-hop. While this may appear to dilute 'authentic' national culture, in some ways globalization also heightens the significance of national identities. Changes associated with globalization – above all, immigration of peoples with very different cultures – are often interpreted as a threat to national identity (Lechner, 2008). Similarly, specifically cultural 'intrusions' associated with globalizing flows may also trigger defensive national responses, for example in the form of quota policies for cultural production or efforts to cleanse a language of English contamination.

Though strongly embedded in world culture, the legitimacy of nations and national identities is contested. New governance structures, such as the EU, have begun to gnaw away at the primacy of national distinctiveness. The sheer scale of global flows makes it ever more difficult to maintain a clear-cut national identity. As more and more individuals sample global cross-currents, they fashion ever more complex hybrid identities. Capitalizing on the growth of the human rights regime, many groups strive to cultivate a form of cosmopolitanism, focused on the well-being of humankind. To the extent

that globalization fosters such trends, it also exposes a continuing tension within world culture.

Religion in Globalized Cultural Context

While much of world culture is decidedly secular, religion plays a special role in it. In fact, all the features of world culture discussed thus far – construction, dimensions, origins, homogenization, and conflict – have roots or parallels in religion. Globalization affords new opportunities for religion and in turn is shaped by it.

'Axial' religions such as Buddhism and Christianity can reasonably be seen as the first world cultures by virtue of their translocal organization, universal message, and highly elaborated concept of 'the world'. As noted earlier, several aspects of contemporary world culture derive from the early history of the Catholic Church and its involvement in the first wave of globalization during European expansion. Religion also figured prominently in the crystallization of world culture in the late-nineteenth century period of intense globalization. Missionary societies were among the most active INGOs, spreading the Christian faith to new places. Other religiously inspired organizations, such as the YMCA, were similarly involved, embracing versions of 'muscular Christianity' that diffused both moral precepts and globalizing sports. Propagating new faiths also became easier, as evidenced by Pentecostalism taking hold in South America (Coleman, 2000) soon after it emerged in Los Angeles at the Azusa Street revival of 1906. Along with such efforts at Christianization, more subtle changes took place as well. As the world drew together in new ways, a new attitude toward religious pluralism emerged. The 1893 World Parliament of Religions in Chicago, for example, brought together representatives of many religions on a relatively equal footing, in an attempt to express the common spiritual concerns of mankind. The World Parliament also signaled that the various faith traditions increasingly were coming to be defined as proper 'religions,' a construct that is itself of Western origin (Beyer, 2006). Speaking of religion as a feature of world culture thus captures a global process in which highly variable 'local' religious experiences gained new significance as part of a more rationalized, standardized world religious domain. To a considerable extent, 'doing' religion is now subject to global norms and standards.

As in world culture more broadly, the Rest challenged the West in several ways after World War II and the decolonization that ensued. Not only countries gained independence, so too did many Protestant churches. In some instances, this occasioned a religious West-against-itself reversal. For example, African Anglicans resisted a more liberal approach to homosexuality in their communion, reflecting wider opposition on the continent to the expansion of human rights in the sexual sphere. As the Catholic Church expanded in the global South, in what some have called 'reglobalization', its membership, leadership, and actual practices further diversified, illustrated partly by the transnational Charismatic Renewal (Csordas, 2009). Overall, the center of gravity in global Christendom has shifted to the South, where worship tends to take a more revivalist form and socioreligious beliefs are generally more conservative (Jenkins, 2011).

Social conservatism also characterizes much of the Islamic world, where resistance to some aspects of modernity has gained prominence in the context of political suppression (Roy, 2004). Globalization, both as actual process and as perceived threat, has had much to do with that renewed religious fervor. Though by no means wholly characteristic of Islam, the resistance takes many forms: Taliban-style attempts at purification, symbolized by the destruction of Buddhist statues in Afghanistan; reform from within, as with Egypt's Muslim Brotherhood after the Arab Spring of 2011; and more violent 'jihad' against mostly Western targets, as perpetrated by globalized militants in loose transnational networks. With few exceptions, however, even such ostensibly oppositional activity still owes much to the world culture it confronts.

As all such examples indicate, religion is of more than historical interest for world culture. If the core of world culture consists of sets of ideas that enable people to act and think as part of one world society, whether in abstract terms or through specific practices, then the current world religions surely count as key components. What 'the world' is about, what human beings are supposed to do in the world, how humans must treat each other as members of a single species, and what lies beyond this world – for millions of people, meaningful answers to such questions derive from religious traditions. Some of those answers fit comfortably with the evolving content of world culture more broadly; witness the Catholic Church's support for human rights or Buddhist endorsement of environmentalism. Other answers fit less comfortably, such as some transnational African religions' conception of the self as inhabited by multiple beings or supporters. Even more challengingly, some world religions question the very categories world culture has inherited from its Western roots, notably the religious/political and transcendence/immanence distinctions, as in forms of Islamic and Asian spirituality, respectively.

World religion stands out as a counterpoint to the notion that globalization entails cultural homogenization. The historic diversity of faiths, reproduced in the modern era, provides the backdrop, but more than reproduction is at issue: globalization facilitates the diffusion of religious ideas and the expansion of religious communities across continents (Wuthnow, 2009), as in the case of Muslim diasporas in Europe and Indian/Hindu diasporas in North America, thus diversifying both those faith communities and the locales where they grow new roots. Nor does diffusion entail simple emulation; a key feature of many transnational religious ventures is the way in which they 'indigenize' or 'glocalize', adapting old models to new places, as in the case of Pentecostal churches absorbing Brazilian, Ghanaian, or Korean symbolism into their worship and teaching. Adding to the diversity of religious expression is the fact that many world religions are very much in flux, strikingly illustrated by the 'spiritual recycling' practiced by Nigerian and Korean Christian leaders reexporting their faith to the West (Micklethwait and Wooldridge, 2009).

Caveats are in order. However great the range of options in the overall religious market created by globalization, local supplies tend to be limited and at times monopolistic. In their daily lives, many religious adherents participate in the same routines as others around the world, guided by common norms. The countries where they reside are subject to the 'isomorphic' tendencies mentioned previously. However

strong their faith, adherents live in a world that operates largely in secular-rational terms set by a world culture that is not evidently grounded in any transcendent faith. Among those who enact and propagate world culture, such as the mostly secular INGOs, specifically religious actors are now in the minority. Yet that same world culture provides space for radically different world views. Deep cultural difference has by no means disappeared; indeed, under new interpretations of human rights, difference has acquired new significance and legitimacy. The differences matter to millions, if not billions. How to manage these differences peaceably, as a source of human enrichment rather than conflict, is a major contemporary challenge.

Overall, religion is a central aspect of cultural globalization: as traditions spread, transnational networks expand, national cultures become more mixed, and new ways of experiencing the world emerge. In many countries, religion mediates the pluralizing effect of world culture. It plays an important role in the intense contests concerning global values and world order. Yet the construction of world culture has become a mainly secular process; it has no transcendent content in the conventional sense. Cultural heterogeneity and conflict themselves take many forms, only some of them religious. While world religion is intimately connected with globalization and involved in the latter's dynamics, it is by no means a dominant force. Whether it can, or should, take on a greater role in defining the desirable world order is likely to be a central issue in future global cultural contestation.

See also: Capitalism: Global; Catholicism; Civil Liberties and Human Rights; Creolization: Sociocultural Aspects; Cultural Relativism, Anthropology of; Global Governance; Global History: Universal and World; Globalization and Religion; Globalization, Anthropology of; Hybridity; Nation-State as Symbolic Construct; Religion: Mobilization and Power; Social Ontology; Transnational Citizenship; Video Cultures: Television; Voluntary Associations, Sociology of.

Bibliography

Appadurai, A., 1996. Modernity at Large: Cultural Dimensions of Globalization. University of Minnesota Press, Minneapolis.

Berger, P.L., Huntington, S.P. (Eds.), 2002. Many Globalizations: Cultural Diversity in the Contemporary World. Oxford University Press, New York.

Beyer, P., 2006. Religions in Global Society. Routledge, London.

Boli, J., Thomas, G.M. (Eds.), 1999. Constructing World Culture: International Nongovernmental Organizations since 1875. Stanford University Press, Stanford.

Coleman, S., 2000. The Globalisation of Charismatic Christianity: Spreading the Gospel of Prosperity. Cambridge University Press, Cambridge.

Cowen, T., 2002. Creative Destruction: How Globalization is Changing the World's Cultures. Princeton University Press, Princeton.

Csordas, T.J. (Ed.), 2009. Transnational Transcendence: Essays on Religion and Globalization. University of California Press, Berkeley.

Hannerz, U., 1996. Transnational Connections: Culture, People, Places. Routledge, London.

Jenkins, P., 2011. The Next Christendom: The Coming of Global Christianity. Oxford University Press, Oxford.

Leclerc, G., 2000. La mondialisation culturelle: les civilisations à l'épreuve. Presses Universitaires de France, Paris.

Lechner, F.J., Boli, J., 2005. World Culture: Origins and Consequences. Blackwell, Malden.

Lechner, F.J., 2008. The Netherlands: Globalization and National Identity. Routledge, New York.

Meyer, J.W., Boli, J., Thomas, G.M., Ramirez, F.O., 1997. World society and the nation-state. American Journal of Sociology 103, 144–181.

Meyer, J.W., Boli, J., Thomas, G.M., 1987. Ontology and rationalization in the western cultural account. In: Thomas, G.M., Meyer, J.W., Ramirez, F.O., Boli, J. (Eds.), Institutional Structure: Constituting State, Society, and the Individual. Sage Publications, Newbury Park, pp. 12–37.

Micklethwait, J., Wooldridge, A., 2009. God is Back: How the Global Revival of Faith is Changing the World. The Penguin Press, New York.

Nederveen Pieterse, J., 2009. Globalization and Culture: Global Mélange, second rev. ed. Rowman & Littlefield, Lanham.

Robertson, R., 1992. Globalization: Social Theory and Global Culture. Sage Publications, Newbury Park.

Roy, O., 2004. Globalized Islam: The Search for a New Ummah. Columbia University Press, New York.

UNESCO, 2000. World Culture Report: Cultural Diversity, Conflict and Pluralism. UNESCO, Paris.

Watson, J.L. (Ed.), 2006. Golden Arches East: McDonald's in East Asia, second ed. Stanford University Press, Stanford.

Wuthnow, R., 2009. Boundless Faith: The Global Outreach of American Churches. University of California Press, Berkeley.

Globalization, Anthropology of

Arjun Appadurai, NYU Steinhardt, Media, Culture, and Communication, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the anthropological interest in globalization from two perspectives. First, in conversation with other social sciences and second, to reexamine classical questions of diffusion, identification, comparison from within the discipline itself. Central to this analysis is the idea of 'flow' that places objects, ideologies, technologies, and images within the dynamic sphere of exchange and circulation. This article also engages with contemporary concerns regarding forceful human dislocation and ensuing violence that are integral to the emerging structures of sovereignty.

The Disciplinary Context

Anthropological interest in globalization is difficult to trace precisely but can be recognized as early as Appadurai's work on the global cultural economy (1990, 1996), Hannerz's analysis of cultural complexity and creolization (1987, 1992), and Friedman's work linking global capitalism and cultural processes (1988, 1994), all of which built on earlier anthropological interest in cultural exchange and world systems. From the start, such work has been part of two sets of scholarly exchanges and debates. The first links anthropology with a wider debate about globalization in other social science fields, notably geography, political science, and sociology. The other is internal to anthropology and links it to longstanding concerns with diffusion, comparison, ethnography, and the study of large-scale historical change. A brief look at each of these contexts might help to frame the anthropological perspective on globalization.

As regards the social sciences in general, anthropologists in the late 1980s found themselves, like many other scholars, trying to understand the processes that led to the breakdown of the Soviet World and to guess at the shape of things to come. They soon found themselves having to engage powerful perspectives on the emergent world order, which came out of the Marxist tradition in geography (exemplified by Harvey, 1989), by a renewed interest in global political culture, exemplified by the polemical and much discussed work of Samuel Huntington (1996), and some prescient studies of the new forms being taken by global capitalism in the last decades of the twentieth century (Lash and Urry, 1987). In addition, anthropologists found themselves in a rich, sometimes competitive dialogue with scholars in literary and cultural studies, notably those influenced by British cultural Marxism, and most powerfully represented by Stuart Hall (1986). In addition, the publication of Benedict Anderson's landmark study of nationalism (1983) provoked a strikingly wide debate about the links between politics, the imagination, and national identification. These stimuli helped to shape anthropological research on globalization, which in many ways is marked by the ongoing effort to link broad structures and processes in the world economy to the subtleties of communication, interpretation, and translation that govern everyday life in all societies.

Within anthropology, the study of globalization is built on several well-established currents of interest. In the United States, the study of globalization fits well with Boasian traditions in the study of diffusion, cultural change, and culture contact and with a longstanding interest in urban settings and complex societies. In Europe, anthropologists were slower to take an interest in globalization; hence, too long-standing interests in problems of scale, social cohesion, and structural change made globalization a subject of growing interest after 1990. In other parts of the world, such as Latin America, Africa, and India, the anthropological interest in globalization was more closely linked to problems of ethics, development, and inequality.

In spite of these various currents of research and theorization within which the anthropological interest in globalization has developed, from the beginning there were serious doubts about whether globalization by its nature was a topic suitable to the special strengths of anthropology in the study of intimate social relations and of societies governed by nonmarket social principles. Many of these anxieties have been translated into methodological concerns, reflected in a significant body of methodological work about how anthropology ought to address the emergent world of globalization (e.g., Fox, 1991). There is still a considerable body of opinion among professional anthropologists that globalization may well be a mere trend, an artifact of academic fashion, and that, in any case, it is not an ideal subject for anthropological research because of its conceptual and social scale. This rearguard anxiety, which is sometimes not easy to distinguish from fear of disciplinary failure, has not prevented a rich set of theoretical debates and empirical studies from emerging out of various anthropological traditions, both in the Atlantic world and beyond. These studies, largely a product of anthropological work since the 1990s, are discussed in the following sections.

Global Cultural Flows

The image of 'flow' has been a central trope of anthropological work on globalization. In the work of Appadurai (1990, 1996), the idea of 'flow' was used to capture a complex dynamic in which objects, ideas, ideologies, technologies, and images were placed in a single economy of circulation, with an eye to distinguish different emergent mosaics of cultural form and social design. In Hannerz's use of this concept (1992, 1996), it was fruitfully linked to Kroeber's idea about large-scale

civilizational and interactional spheres – ecumenes – as well as to ideas about creolization, cosmopolitanization, and the professional networks associated with rapid flows of cultural commodities.

In these early usages of the idea of flow, and in subsequent elaborations of them by others, the idea of flow was sharpened to redress the assumption that cultural flows in the era of globalization were unrestricted in scope and range. A salutary emphasis on boundaries, limitations, selectivities, and exclusions reappeared, thus placing the classic debate between Marxist and cultural approaches once again at the center of attention. This broad interpretive struggle continues (Rockefeller, 2011), and the choices it poses are reflected in many studies associated with the anthropology of globalization, on such subjects as migration (Lomnitz, 2001; Eriksen, 2010; Gingrich and Banks, 2006), colonialism and post-colonialism (Comaroff and Comaroff, 2000a; Mbembe, 2001), media (Feld, 1999; Ginsburg et al., 2002), diaspora (van der Veer, 1995), markets (Hart, 1999; Bestor, 2004), and religion (Csordas, 2009; Rudnyckyj, 2010; van der Veer, 2013).

The positive contribution of the idea of flow was to open up new questions about the relation between legal and illicit flows of commodities and persons, the changing nature of borders and boundaries, the implications of new regimes of financial circulation and cybercommerce, and the evolution of new forms of diasporic identification and mobile tradition building (Nordstrom, 2007; Nordstrom and Carlson, 2014). The image of flows also opens a potential conversation, yet to be fully exploited, between economists and anthropologists on the question of the changes in the world economy suggested by the idea of globalization, and of the best ways of measuring their scale and significance. In a general way, the engagement with cultural flows, their associated economies and interpretations, has marked a decisive shift, in which anthropologists are no longer confined to the study of simple market societies but are addressing such complex phenomena as cybercommerce, intellectual property (Strathern, 2006), and new forms of genetic engineering (Lock and Nguyen, 2010) and marketing. Likewise, the anthropological focus on cross-border flows, as we will see shortly, has opened up a useful conversation between anthropologists, political theorists, and sociologists on new forms of citizenship and sovereignty (Geschiere, 2009; Held and Moore, 2008).

Above all, the focus on flows, regardless of context, commodity, or sphere, has brought anthropology back to one of its earlier insights, namely that most human societies have always been in interactive relations with others, forming spheres of exchange and circulation. Thus, the images of social stability, impermeable boundaries, and natural divisions between ethnic groups, polities, and communities, sometimes implicit in anthropological practice, have been put under sharp scrutiny. In turn, the widespread recognition that societies in the era of globalization are inevitably parts of wider circulatory systems and networks, has posed new challenges for how anthropologists will need to theorize classical models of cohesion, consensus, and order (Appadurai, 2013). At the end of the twentieth century, there was some uncertainty about how to reconcile the undoubted significance of large-scale cultural differences with the reality of the velocity of cross-border flows and the uneven thickness of the membranes between societies, whether organized as nation-states or not. It seems probable that the question of how social order and cultural coherence in lived communities are maintained in the face of complex and uneven cross-border flows will be a major theoretical challenge for the twenty-first century anthropology.

Many of these concerns have produced a special interest in the anthropology of nationalism and a series of contested views about the salience, form, and future of the nation-state as a form of global political organization. The pendulum appears to swing between those (both inside and outside anthropology) who predict the early demise of the nation-state and those who see it as growing stronger than ever. Sometimes, these debates are reflections of actual differences between existing states and sometimes, they are the result of blurring the lines between nations and nationalism on the one hand and states and their powers on the other. The emerging consensus seems to be that while nation-states are unlikely to disappear in the near future, they are evolving their own powers in a transformed ecology of sovereignty. This ecology is affected by changes in the velocity of global capital flows, scale and speed of cross-border shadow economies, and geography of dual, mobile, and diasporic political identifications among citizens. These processes are discussed further below (see Section Sovereignty, Citizenship, and Violence).

The interest in various forms of global cultural flow, and in the nonisomorphic relationship between them, has provoked a strong debate about whether, and the extent to which, globalization is a new social regime. Many voices within anthropology argue vigorously that globalization is no more than a semiotic conspiracy of the media and of global capital and its allies in the world of development. In this view, globalization is only a clever way to repackage colonialism, imperialism, modernization, or various combinations of these ideologies. It is old wine in new bottles. This view runs into the difficulty of having to accommodate too many new animals into an existing theoretical Ark and suffers from the further disability that it does not provide any criteria for recognizing genuine systemic change. It ends history in the name of theory. The more widespread band of opinion among anthropologists is that something surely has changed. The challenge is, as always, to establish what has changed and how it affects the things that have not changed. Thus, there is a renewed interest in how identities are negotiated across multiple locations, how existing structures like witchcraft (Geschiere, 1997) become vehicles and sites for new anxieties and energies, and of how new legal and economic arrangements (those sometimes lumped together under the rubric of neoliberalism) call forth new perceptions of danger, wealth, and risk (Comaroff and Comaroff, 2000b).

Finally, ideas and debates surrounding the idea of global cultural flows have cast fresh light on a classic concern of anthropology, namely the process of identification and the production of identities. These processes have to be reexamined in a world characterized by massive information flow, heightened media images of life possibilities for ordinary people, and new fantasies of wealth and mobility. The rapid growth in mass-market media penetration across national boundaries, the associated growth in the battle between global and national media barons, the projected increase in

various kinds of bandwidth, and the emergent possibility of new, hybrid instruments for cheap, high-speed communication, all combine to change the grounds on which stable cultural identities are produced and secured. Likewise, these factors make it impossible to presume the image of the local as an unchanging ground against which the tableau of global change plays out. Indeed, many anthropologists now have come to take as great an interest in 'the production of locality' (Appadurai, 1996) as in the dynamics of globalization.

With its special interest in language, everyday life, and small-scale social interaction, anthropology is evolving a special purchase on the vast terrain represented by globalization, by paying particular attention to the vagaries of translation and localization and to the special role of language and semiotic mediation in producing new forms of lived experience. This has resulted in a sustained effort to examine the cultural dimension of objects and technologies that have their primary home in the spheres of law, science, and the market. Therefore, anthropologists at the beginning of the twenty-first century studied the organ trade, refugee incarceration, NGOs, musical copyright law, and derivatives, to name just a few examples of social and technical forms whose cultural design is sometimes overlooked. In this proliferation of topics (some of them addressed by other authors in the Encyclopedia), there are some major crosscutting themes, two of which are briefly discussed below.

Heterogeneity or Homogeneity?

Most anthropologists share the widespread presumption that globalization is a new phase in the integration of the world economy, principally powered by new scales and speeds in the circulation of financial capital and the associated explosion in electronic communication technologies. When we recall that the major models of global economic processes have come out of the Marxist tradition, it helps to explain why there is a tendency among some anthropologists to argue that globalization is no more than a new phase in the history of capitalist commodification. From this, follows the possibility that, from a cultural point of view, the forces of commodification will produce increasing levels of cultural homogenization and social standardization, particularly in the realm of media icons, social styles, and consumption values. The further extrapolation, rarely stated explicitly, is that these processes will in turn erase cultural differences both within and across societies. This is the nightmare of globalization as coca-colonization (Hollywood, McDonald's, and Disney are the favored icons of this line of speculation). Yet the evidence is mounting that societies appropriate various forms of commodification in their own terms, media messages are downloaded idiosyncratically, weapons of standardization are frequently subverted to produce further brand differentiation, and the best-laid plans of the standardizers are frequently thwarted by local entrepreneurs and customers.

A splendid example of this process of differentiation is provided in a collection of anthropological essays on the role and style of McDonald's, as it functions in six distinct East Asian settings (Watson, 1997). The ethnographic essays in this collection illustrate that, as a global profit-making organization, McDonald's is certainly eager to create uniformity in its products and in its outlets wherever possible. But the essays also show that in each of these settings, McDonald's encounters a quite different set of needs, anxieties, and possibilities and is drawn into a specific local mosaic of social patterns and cultural orientations, involving age, leisure, work, and freedom. Though McDonald's is a heavily capitalized multinational corporation, with an unerring eye for its local and global profits, its sensitivity to local markets inevitably contributes to the emergence of different desires and practices in its outlets. This does not make it any less a global corporation, but its strategies are regularly bent to fit local wishes and projects. However, these essays also show that this is not a simple matter of Goliath being repackaged to suit David, culturally speaking. In fact, what McDonald's produces in these settings are new and unintended styles and contexts for consumption, sociality, and public interaction, each culturally distinct but not in any obvious or predictable way. Many other studies, especially in the sphere of consumption (involving clothing, music, food, leisure, and housing), also confirm this process.

Heterogenization is not a mechanical product of the sphere of consumption. It is itself produced by local makers of ideas, images, and commodities who give to the culture industries of different societies and nations their own distinctive stamp. Thus, anthropologists have been able to show that the simple distinction that places global producers and uniformity on the one side and local consumers and difference on the other is overdrawn. We have a growing series of studies of cultural production worldwide, especially in the areas of music, film, and advertising, which let us look into the sites and institutions through which global commodities are locally interpreted by producers as well as consumers. The study of 'world music' by ethnomusicologists is perhaps the best developed of these subfields. In general, these studies have produced a broad consensus that cultural differentiation tends to outpace homogenization, even in this most interactive of economic epochs. Naturally, this conclusion fits with the methodological and ethical proclivities of anthropology, with its professional investment in cultural difference and its ethical commitment to cultural diversity. Nevertheless, there appears to be a substantial empirical basis for this consensus. But anthropological interest has by no means focused solely on the good news of cultural differentiation and has not been uncritically celebratory in its treatment of globalization, as the following section will show.

Sovereignty, Citizenship, and Violence

Anthropologists have not failed to notice that the world of global cultural flows is hardly free of suffering, injustice, and dislocation. They have certainly been aware that the incidence and brutality of forced population movements have, in some respects, increased, producing many pockets of human dislocation and frequency of ethnic violence. They have paid attention to the dilemmas of proletarian labor migrants, who struggle with their own translocal loyalties – frequently to religious formations which are themselves increasingly

transnational in scope – against the pressures of the varied ideologies of cultural and legal inclusion that characterize their new homes.

Considered transnationally, poorer migrants face bewildering new combinations of democratic ideology and racist reality. In some cases, such as the oil kingdoms of the Middle East, nonimmigrants monopolize most of the benefits of citizenship. In others, the benefits of citizenship are more widely shared, but everyday racism is a constant threat. In some cases, refugees are drawn into multinational ethnic wars (as in Central Africa in recent years), while in other cases, the revitalization of longstanding ethnonational boundaries flies in the face of intermarriage and ethnic mixing as social realities. In Central Europe, after the death of Tito, the new national constitutions frequently encouraged political campaigns that exacerbated genocidal identities parading as 'primordial.' Many nation-states now practice various forms of cultural fundamentalism and majoritarianism, as they struggle for new forms of legitimation in a global economy. In the process, they increasingly both produce and demonize 'minorities' (Appadurai, 2006). In many of these cases, language, blood, and race are overdetermined icons of peoplehood that force anthropologists to reconcile their own knowledge of the fragility and historicity of these constructions against the popular, lived sense that these realities are timeless, natural, and nonnegotiable.

As anthropologists examine these bloody sites of ethnic violence and violent social purification (Appadurai, 2006; Das, 2006; Feldman, 1991; Malkki, 1995), they find themselves having to study a variety of antecedent and emergent social realities. These include new forms of 'flexible' citizenship (Ong, 1999), new ideologies of citizenship and law (Comaroff and Comaroff, 2000b; Sassen, 1996), new battles over indigenes and outsiders (Meyer and Geschiere, 1999), and new patterns of state-sponsored violence and legality. In engaging with these topics, anthropologists discover that they need to link longstanding professional interests in conflict, bodily ritual, mythologies of origin, and ethnocultural ideology, with new regimes of law, incarceration, borders, and migration.

In particular, anthropologists are now tackling the ways in which the forms of violence and dislocation produced by globalization may be related to new structures of sovereignty, new forms of nationally sponsored globalization, and new niches for the location of the global economy. They thus find themselves in a productive dialogue with scholars in adjacent fields who are making important contributions to these subjects (Castells, 1997; Held and McGrew, 2000; Sassen, 1998). This engagement promises a more sophisticated answer to the question of how new forms of global linkage are also producing new forms of cultural fundamentalism, a question to which somewhat Manichaean answers have been provided in the popular media and its favored public intellectuals in many national contexts. In the United States, and to some extent in Europe, there have been strong alarms raised by economists and development specialists to the effect that globalization is producing both greater wealth and greater inequality. Yet, in popular opinion, especially among the wealthier countries and classes, there seems to be an emergent sense that globalization is a world of the victors in the Cold War and their global allies who have cracked the codes of the global market, the information economy, and the new international order. The losers, in this view, are cultural holdouts at best, cultural terrorists at worst, people who have not understood the new alphabet of the globalized world.

Anthropological work on globalization, in this context, has a challenge and an obligation. The challenge is to take its classical strengths in the study of intimate relations, everyday life, and cultural difference and use them to cast light on circulatory processes that yield local and global results simultaneously. The obligation is to provide a way of thinking about why this era of global integration is yielding at least as much genocidal violence as its democratic voice. That may not seem like a particularly anthropological obligation, but given that anthropology has long tried to create new ways of showing how broad human processes are found in translation, globalization may need anthropology at least as much as vice versa.

Looking Ahead

It is not easy to predict what topics will find special favor among anthropologists interested in globalization in the coming decades. It is likely that some of the broad themes discussed in this article will continue to elicit empirical research and theoretical debate, but it also seems likely that the institutional and geographical range of the anthropological debate will widen. Research institutes, conferences, workshops, publications, and translations concerned with anthropological contributions to the study of globalization are appearing in Brazil, Turkey, Hong Kong, India, and in many other sites outside of the North Atlantic region. In these many sites, anthropologists are building their own critical resources as they engage with specific kinds of national debate, cultural conflict, and research tradition (Appadurai, 1999). Whatever the individual outcomes of this process, two results seem likely. Anthropology is unlikely to remain a bystander in any important national debate about globalization. Moreover, these debates are increasingly unlikely to be determined by scholars, projects, and institutions confined to Europe and the United States. What more appropriate topic could there be for international research than globalization, and what opportunity could be more appropriate for anthropology to contribute to a global scholarly debate about the world we are about to build?

As far as the United States and Western Europe are concerned, anthropologists working from these locations (who are frequently nationals of other countries in Asia, Africa, and the Middle East) are breaking new ground in numerous topical areas connected to globalization. A partial list of some of these important areas would include militarization and counterinsurgency (Kelly et al., 2010), new financial instruments and markets (Appadurai, 2011, 2012; Hart, 2011; Maurer, 2005, 2006; Schüll, 2012; Zaloom, 2006), religion and media (Engelke, 2013; Larkin, 2008; Houtman and Meyer, 2012; Pels, 2012; Spyer, 2009), human–animal relationships (Helmreich, 2009; Ingold, 1994; Raffles,2002, 2010), infrastructure (Anand, 2011; Chalfin, 2010; de Boeck, 2011; Larkin, 2013), ethics and everyday life (Faubion, 2011; Jackson, 2013; Lambek, 2010), the Internet and cyber sociality (Boellstorff, 2008; Coleman, 2012; Ito, 2009; Miller and Slater, 2000; Turkle, 2011). These

emerging interests deepen the conversation between anthropologists of globalization and scholars in other disciplines. They reflect the classical interest of anthropology in the question of what it means to be human, and ask also what this question means in an era in which the human, the global, and the planetary have become overlapping forces in our worlds.

See also: Center–Periphery Relationships; Colonialism, Anthropology of; Cosmopolitanism; Cultural Critique, Anthropological; Diaspora; Diffusion: Anthropological Aspects; Global South; Human Rights and Social Work; Hybridity; Migration: Anthropological Perspectives; Modernity; Multisited Field Studies; Postindustrial Societies; State: Anthropological Aspects; Xenophobia and Nativism.

Bibliography

Anand, N., 2011. Pressure: the polytechnics of water supply in Mumbai. Cultural Anthropology 26 (4), 542–564.

Anderson, B., 1983. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London.

Appadurai, A., 1990. Disjuncture and difference in the global cultural economy. Public Culture 2 (2), 1–24.

Appadurai, A., 1996. Modernity at Large: Cultural Dimensions of Globalization. University of Minnesota Press, Minneapolis, MN.

Appadurai, A., 1999. Grassroots globalization and the research imagination. Public Culture 12 (1), 1–19.

Appadurai, A., 2006. Fear of Small Numbers: An Essay on the Geography of Anger. Duke University Press, Durham, NC.

Appadurai, A., 2011. The ghost in the financial machine. Public Culture 23 (3), 517–539.

Appadurai, A., 2012. The spirit of calculation. Cambridge Anthropology 20 (1), 3–17.

Appadurai, A., 2013. The Future as Cultural Fact: Essays on the Global Condition. Verso, New York.

Bestor, T., 2004. Tsukiji: The Fish Market at the Center of the World. University of California Press, Berkeley, CA.

Boellstorff, T., 2008. Coming of Age in Second Life: An Anthropologist Explores the Virtually Human. Princeton University Press, Princeton, NJ.

de Boeck, F., 2011. Inhabiting ocular ground: Kinhasa's future in the light of Congo's spectral urban politics. Cultural Anthropology 26 (2), 263–286.

Castells, M., 1997. The Power of Identity, vol. II. Oxford, UK: Blackwell.

Chalfin, B., 2010. Neoliberal Frontiers: An Ethnography of Sovereignty in West Africa. University of Chicago Press, Chicago.

Coleman, E.G., 2012. Coding Freedom: The Ethics and Aesthetics of Hacking. Princeton University Press, Princeton, NJ.

Comaroff, J., Comaroff, J.L., 2000a. Millennial capitalism and the culture of neoliberalism. Public Culture 12 (2). Special issue.

Comaroff, J., Comaroff, J.L., 2000b. Millennial capitalism: first thoughts on a second coming. Public Culture 12 (2), 291–343.

Csordas, T.J. (Ed.), 2009. Transnational Transcendence: Essays on Religion and Globalization. University of California Press, Berkeley, CA.

Das, V., 2006. Life and Words: Violence and the Descent into the Ordinary. University of California Press, Berkeley, CA.

Engelke, M., 2013. God's Agents: Biblical Publicity in Contemporary England. University of California Press, Berkeley, CA.

Eriksen, T.H., 2010. Ethnicity and Nationalism: Anthropological Perspectives, third ed. Pluto Press, New York.

Faubion, J.D., 2011. An Anthropology of Ethics. Cambridge University Press, New York.

Feld, S., 1999. A sweet lullaby for world music. Public Culture 12 (1), 145–171.

Feldman, A., 1991. Formations of Violence: The Narrative of Body and Political Terror in Northern Ireland. University of Chicago Press, Chicago.

Fox, R.G. (Ed.), 1991. Recapturing Anthropology: Working in the Present. School of American Research Press, Santa Fe, NM.

Friedman, J., 1988. Cultural logics of the global system. Theory, Culture, and Society 5 (2–3), 477–490.

Friedman, J., 1994. Cultural Identity and Global Process. Sage, Thousand Oaks, CA.

Geschiere, P., 1997. The Modernity of Witchcraft: Politics and the Occult in Postcolonial Africa. University of Virginia Press, Charlottesville, VA.

Geschiere, P., 2009. The Perils of Belonging: Autochthony, Citizenship, and Exclusion in Africa and Europe. University of Chicago Press, Chicago.

Gingrich, A., Banks, M. (Eds.), 2006. Neo-Nationalism in Europe and beyond: Perspectives from Social Anthropology. Bergahn Books, New York.

Ginsburg, F.D., Abu-Lughod, L., Larkin, B. (Eds.), 2002. Media Worlds: Anthropology on New Terrain. University of California Press, Berkeley, CA.

Hall, S., 1986. Cultural studies: two paradigms. In: Collins, R. (Ed.), Media, Culture and Society: A Critical Reader. Sage, Beverly Hills, CA.

Hannerz, U., 1987. The world in creolisation. Africa 57, 546–559.

Hannerz, U., 1992. Cultural Complexity: Studies in the Social Organization of Meaning. Columbia University Press, New York.

Hannerz, U., 1996. Transnational Connections: Culture, People, Place. Routledge, New York.

Hart, K., 1999. The Memory Bank: Money in an Unequal World. Profile Books, London.

Hart, K., 2011. The financial crisis and the end of all-purpose money. Economic Sociology: The European Electronic Newsletter 12 (2), 4–10.

Harvey, D., 1989. The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change. Blackwell, Cambridge, MA.

Held, D., McGrew, A., 2000. The Global Transformations Reader: An Introduction to the Globalization Debate. Polity Press, Malden, MA.

Held, D., Moore, H.L. (Eds.), 2008. Cultural Politics in a Global Age: Uncertainty, Solidarity, and Innovation. Oneworld Publications, Oxford, UK.

Helmreich, S., 2009. Alien Ocean: Anthropological Voyages in Microbial Seas. University of California Press, Berkeley, CA.

Houtman, D., Meyer, B. (Eds.), 2012. Things: Religion and the Question of Materiality. Fordham University Press, New York.

Huntington, S.P., 1996. The Clash of Civilizations and the Remaking of World Order. Simon & Schuster, New York.

Ingold, T. (Ed.), 1994. What is an Animal?. Routledge, New York.

Ito, M., 2009. Engineering Play: A Cultural History of Children's Software. MIT Press, Cambridge, MA.

Jackson, M., 2013. Lifeworlds: Essays in Existential Anthropology. University of Chicago Press, Chicago.

Kelly, J.D., Jauregui, B., Mitchell, S.T., Walton, J. (Eds.), 2010. Anthropology and Global Counterinsurgency. University of Chicago Press, Chicago.

Lambek, M. (Ed.), 2010. Ordinary Ethics: Anthropology, Language, and Action. Fordham University Press, New York.

Larkin, B., 2008. Signal and Noise: Media, Infrastructure, and Urban Culture in Nigeria. Duke University Press, Durham, NC.

Larkin, B., 2013. The politics and poetics of infrastructure. Annual Review of Anthropology 42 (1), 327–343.

Lash, S., Urry, J., 1987. The End of Organized Capitalism. University of Wisconsin Press, Madison, WI.

Lock, M., Nguyen, V.K., 2010. An Anthropology of Biomedicine. Wiley-Blackwell, Malden, MA.

Lomnitz, C., 2001. Deep Mexico, Silent Mexico: An Anthropology of Nationalism. University of Minnesota Press, Minneapolis, MN.

Malkki, L.H., 1995. Purity and Exile: Violence, Memory and National Cosmology among Hutu Refugees in Tanzania. University of Chicago Press, Chicago.

Maurer, B., 2005. Mutual Life, Limited: Islamic Banking, Alternative Currencies, Lateral Reason. Princeton University Press, Princeton, NJ.

Maurer, B., 2006. The anthropology of money. Annual Review of Anthropology 35 (1), 15–36.

Mbembe, A., 2001. On the Postcolony. University of California Press, Berkeley, CA.

Meyer, B., Geschiere, P., 1999. Globalization and Identity: Dialectics of Flow and Closure. Blackwell Publishers, Malden, MA.

Miller, D., Slater, D., 2000. The Internet: An Ethnographic Approach. Berg, New York.

Nordstrom, C., 2007. Global Outlaws: Crime, Money, and Power in the Contemporary World. University of California Press, Berkeley, CA.

Nordstrom, C., Carlson, L., 2014. Cyber Shadows: Beyond Big Data, Botnets, and Privacy. Corby Press & EBooks.

Ong, A., 1999. Flexible Citizenship: The Cultural Logics of Transnationality. Duke University Press, Durham, NC.

Pels, P., 2012. The modern fear of matter: reflections on the protestantism of Victorian science. In: Houtman, D., Meyer, B. (Eds.), Things: Religion and the Question of Materiality. Fordham University Press, New York.

Raffles, H., 2002. In Amazonia: A Natural History. Princeton University Press, Princeton, NJ.

Raffles, H., 2010. Insectopedia. Random House, New York.
Rockefeller, S.A., 2011. Flow. Current Anthropology 52 (4), 557–578.
Rudnyckyj, D., 2010. Spiritual Economies: Islam, Globalization, and the Afterlife of Development. Cornell University Press, Ithaca, NY.
Sassen, S., 1996. Losing Control? Sovereignty in an Age of Globalization. Columbia University Press, New York.
Sassen, S., 1998. Globalization and Its Discontents: Essays on the New Mobility of People and Money. The New Press, New York.
Schüll, N.D., 2012. Addiction by Design: Machine Gambling in Las Vegas. Princeton University Press, Princeton, NJ.
Spyer, P., 2009. In and out of the picture: photography, ritual, and modernity in Aru, Indonesia. In: Morris, R.C. (Ed.), Photographies East: The Camera and its Histories in East and Southeast Asia. Duke University Press, Durham, NC.
Strathern, M., 2006. Intellectual property and rights: an anthropological perspective. In: Tilley, C., Keane, W., Kuechler-Fogden, S., Rowlands, M., Spyer, P. (Eds.), Handbook of Material Culture. Sage, Thousand Oaks, CA.
Turkle, S., 2011. Alone Together: Why We Expect More from Technology and Less from Each Other. Basic Books, New York.
van der Veer, P., 1995. Nation and Migration: The Politics of Space in the South Asian Diaspora. University of Pennsylvania Press, Philadelphia, PA.
van der Veer, P., 2013. The Modern Spirit of Asia: The Spiritual and the Secular in China and India. Princeton University Press, Princeton, NJ.
Watson, J.L., 1997. Golden Arches East: McDonald's in East Asia. Stanford University Press, Stanford, CA.
Zaloom, C., 2006. Out of the Pits: Traders and Technology from Chicago to London. University of Chicago Press, Chicago.

Globalization: Geographical Aspects

Andrew Jones, School of Arts and Social Sciences, City University London, Northampton Square, London, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A. Amin, volume 9, pp. 6271–6277, © 2001, Elsevier Ltd.

Abstract

This article examines the nature of geographical thinking about globalization and contrasts the nature of geographer's theoretical contributions with other social science disciplines. It argues that geographical thought on globalization provides a distinctive set of insights into the spatialities of globalization which can be of great utility to the central concerns of a range of social science disciplines, and which address some of the limitations identified in the wider globalization debate. It aims to illustrate these geographical ideas through consideration of the geographer's work on global economic developments and city regions.

Introduction

Globalization continues to be one of the most widely used concepts by social scientists to describe and explain how contemporary world society is changing in the twenty-first century. Yet much of the globalization debate of the last 30 years has only intermittently and often partially engaged with the complex spatialities of globalization processes and global societal integration. During the 1990s, debates about globalization were often framed around simplistic understandings of globalization as creating homogeneity and sameness (Bisley, 2007) and the long-standing simplistic idea that this constitutes an 'end of geography' (cf O'Brien, 1991; Friedman, 2007). While the last decade has seen a more diverse and sophisticated approach to understanding what globalization 'is' (as a complex set of tendencies rather than a singular process or end state (Dicken, 2011)), much discussion of globalization within a range of social science disciplines still does not foreground the intrinsic spatiotemporal nature of the transformations it is associated with (Jones, 2010). In this article, I want to argue that this is a significant ongoing limitation in the effectiveness of globalization theories, and that a wider integration of geographical theorization of globalization provides scope to address this. In that respect, it seeks to demonstrate how geographical thought – not always utilized by globalization theorists more widely – provides a range of epistemological insights and theoretical concepts that can both inform and enhance social scientific theories of globalization in a way that permits a more sophisticated understanding of the complex set of phenomena captured by the term. It is, therefore, an (albeit) modest manifesto for a much wider integration of geographical approaches to understanding globalization than has hitherto emerged across the social sciences.

Geographical Thought and Globalization

Geographical thinking has been criticized – certainly within the discipline – for standing on the margins of the so-called mainstream globalization debate that emerged during the 1990s (Dicken, 2011). With a few notable exceptions such as Doreen Massey, David Harvey, or Peter Dicken, geographical thinkers did not figure in the wide-ranging meta-theoretical accounts of globalization that emerged from a range of social science disciplines (and in popular commentaries) (cf Harvey, 1989, 2003, 2006; Massey, 1994, 1999; Dicken, 2011). It was political scientists, sociologists, and urban theorists who put forward all-encompassing theoretical accounts of the development of globalization as a phenomenon – obvious examples being the contribution by David Held and his co-authors, Jan Art Sholte, Manuel Castells, George Ritzer, or Saskia Sassen (cf Held et al., 1999; Sassen, 2001, 2007; Castells, 2009; Ritzer, 2009). From one perspective, geographical thought has missed an enormous opportunity (Dicken, 2011) in making both a distinctive and sophisticated contribution to understanding this complex and contested phenomenon.

Yet another reason for the lack of prevalence of geographical thinking in the so-called mainstream globalization debate is that geographical thinking quickly adopted a critical and skeptical perspective on social scientific meta-theories of globalization (Jones, 2010). This was not the skeptical 'school of thought' identified by Held et al. (1999) as doubting the pervasiveness of globalization both as state of society (defined as global societal integration as the planetary scale) and of globalization processes (e.g., Hirst and Thompson, 1999), but rather a skepticism at the utility of the concept and the coherence of a meta-theoretical approach that placed globalization at the center of explanations for a multitude of societal phenomenon (Jones, 2010). For geographical thinkers, the globalization debate that continued into the new century after a wave of debate in the 1990s often appeared simplistic and naive, deploying an unsophisticated spatial epistemology that failed to capture the nuances, diversity, and sheer complexity of contemporary global societal integration (Cox, 1998; Taylor, 1999; Sheppard, 2002; Amin, 2002). Geographers have long made use of a plethora of spatial concepts – space, place, proximity, territoriality, boundaries – that did not fit easily into the dynamics of a globalization debate that raged around polarized caricatures of integration (Amin, 2002). The proposition that globalization corresponded to 'the end of geography' (O'Brien, 1991), creating a 'borderless world' (Ohmae, 1995; Fung et al., 2007) or producing truly global corporations (cf Jones, 2005) – counterposed against historical analyses that nation-states remained the key building block of the international system and that trade was not more radically 'globalized' than the interwar period – did

not provide much theoretical traction in the context of geographical debates about uneven development (Cox, 1998; Brenner, 1999; Smith, 2008), the construction of global production networks (GPNs) (Yeung, 1998; Henderson et al., 2002; Coe et al., 2004), or the fractured nature of urban political movements in different world cities (Amin and Thrift, 2002; Brenner and Keil, 2006). From these vantage points, critical interventions within these debates were at times of limited value since it amounted metaphorically to shooting a series of conceptual and epistemological ghosts. None of the positions characterized by Held et al. (1999) – globalist, skeptic, or transformationalist – served to fully capture the way in which geographical thinking was engaging with the concept of globalization (Jones, 2010).

In this respect, geographers have developed what can be seen as a more multilayered and complex view of the spatiality of the phenomenon associated with globalization which has not been willing to treat globalization in the singular, and which accentuates the variable and uneven spatial configurations of social relations in the period of contemporary globalization. Such an approach contrasts with other disciplines insofar as it can be argued that geographical thinking about globalization has taken a broader and more holistic approach that has been less disposed to focus on a particular dimension or epistemological 'cut' on the concept. For example, in economics and business studies, for example, earlier studies of globalization focused on the implications of the emergence of a supposedly borderless economy (Ohmae, 1995) (driven by international market integration, global flows of money and information, and transnational production and regulation) on national and subnational economic sovereignty, competitiveness, and growth. By the 2000s, these disciplinary approaches were widely characterized by a proglobalization or 'globalist' (Friedman, 2000; Wolf, 2004) perspective that focused on the economic characteristics of globalization, and was grounded in a view of the benefits of liberal international markets and an increasingly cosmopolitan and democratic global society (Friedman, 2000; Wolf, 2004). The optimists celebrate the offer of new market opportunities, and the capacity of economic and social interconnectedness to deliver greater material well-being for a growing proportion of the world's population including those in less developed countries (Bhagwati, 2004).

In contrast, within political economy and international politics, the discipline has been characterized often by a more critical and skeptical understanding of globalization based on the disciplinary concern for the nation-state, national borders, and the nature of global governance and political structures (Berger, 2000; Swank, 2002). During the 1990s, historical-based analyses doubted that contemporary globalization was as unprecedented or novel as the globalists claimed (Hirst and Thompson, 1999), while more recent political economic contributions have focused on the discursive dangers of neoliberalism, the problems created by a global economy dominated by a small number of powerful transnational corporations (TNCs), and the growth of wealth inequality across the planet (Strange, 1997). In politics and international relations, thinkers have also been concerned with a wide-ranging debate centered on the erosion/reformulation of the nation-state as a unit of authority (e.g., Holton, 2001; McGrew and Lewis, 2013), following the rise of non-state institutions of regulation and governance (e.g., Garrett, 2000), and on the challenge to citizenship and democracy posed by the rise of transnational political organizations (e.g., global NGOs) as well as plural, non-national sources of political rights and identification (e.g., EU citizenship) (e.g., Jacob and Neumann, 2006). More recent work has grounded such interventions around increasingly vexed challenges in confronting the governance of global climate changes, conflict management, or the postcrisis global financial system (Palan et al., 2013).

Of course this argument should not be taken too far. Clearly there has been and there remains considerable overlap in the way geographers have engaged with globalization processes with other disciplines, and questions of the complex spatiality of globalization have also been substantially engaged within those fields. This is certainly the case in sociology where extensive consideration of the spatial constitution of globalization underpins much thinking (e.g., Giddens, 1990). However, much of the debates have still been dominated by contributions that seek to develop broad conceptual frameworks – or overarching meta-theoretical accounts – of globalization that span a diverse range of societal transformations. Within sociology and cultural anthropology or cultural studies, there has been wide-ranging engagement with globalization through the lens of the global informational society (Castells, 2009), the nature of cultural globalization (Appadurai, 1996; Tomlinson, 1999), or attempt to develop an overarching sociology of globalization (Sassen, 2007). Much attention has also fallen on the implications of information and media technologies (Castells, 2009), global consumerism (Hopper, 2006), and international mobility (Shamir, 2005) which is generating a mixing of identities and lifestyles in different local settings. Through these contributions, there is certainly overlap with the spatial concerns of geographers but the concern for spatial difference is secondary to the role of a global-scale meta-process or phenomenon. A good example would be George Ritzer's theories about the erosion of local difference under the weight of the same world products and consumption norms and the same global corporations (i.e., the so-called MacDonaldization of society) (cf Ritzer, 2011). A sensitivity to the persistence of local difference is engaged within this work, but there is less focus than a geographical perspective would adopt on the nuances of local change resulting from heightened global exposure and virtual and real linkage with different parts of the world.

In foregrounding the diverse nature of globalization's spatial implications, geographers have therefore taken a distinctive epistemological approach to globalization when contrasted to other social science disciplines. But this has also occurred in a context where over the last decade, the wider social scientific debate has become ever more fractured and diverse in its analysis of globalization (Jones, 2010). Bisley (2007) argues that debate can be characterized by at least five cross-disciplinary phases and in that sense it is becoming more difficult to characterize a particular disciplinary 'cut' at globalization. Political thinkers adopt both globalist and skeptical views of the phenomenon, while cultural thinking is becoming ever more influential on those concerned with the global economy or the nature of transnational firms or markets (e.g., Edwards and Rees, 2006; Gertsen et al., 2012). At the same

time, many of the subdebates within social science disciplines have become more nuanced in both their engagement with the specificities of globalization processes and also in their conceptualization of the complex spatialities manifest in contemporary global transformations. Management theorists have engaged with the complex spatial form of firms as transnational organizations (Dunning and Lundan, 2008), and political economists increasingly viewed debates on the interactions between state and suprastates as requiring a spatial epistemology that moves beyond states as territorial containers of institutions and political systems (e.g., Boyer and Drache, 1996). A consequence of this is that interest in some social science disciplines in the existing work of geographical thinkers on globalization has increased, and it has also opened up the interdisciplinary scope for a more sophisticated theoretical framing of the array of phenomena captured in the concept of globalization. In order to further argue for the distinctive potential of further engagement with geographical thinking in this respect, the next task is to provide further insight into how geographer's conceptualization of globalization has evolved in the last couple of decades.

A Distinctively Geographical Approach to Globalization?

During the 1990s, geographers developed a theoretically diverse and skeptical view of many of the globalization theories – and especially the meta-theoretical frameworks – in other social science disciplines. This stemmed from both a diversity of approach within the discipline, and also a spatial onto-epistemological perspective that intrinsically questioned the foundations of all-encompassing theories of globalization. Thus, while there are parallels to the sociological theorization of Giddens which that argued that globalization was at root a radical reconfiguration of social relations in space and time in the contemporary period (and developed by Held et al., 1999 in their (revised) transformationalist approach (cf Jones, 2010)), geographical thinking tended to deploy a more nuanced and complex understanding of spatiality that made use of a variety of concepts not widely developed in the rest of the debate. Such concepts as scale, place, distance, proximity, territoriality, and different forms of 'space' occupy the geographical literature. Geographical thinking was thus instinctively dissatisfied with the meta-theoretical deployment of concepts such as 'deterritorialization' (Elden, 2005), as manifesting one generalizable phenomenon applicable to every globalized social relationship – an underlying basis for some of the broad theories of globalization developed in other disciplines.

This appreciation of the complexity of the social spatial in the era of contemporary globalization thus in part accounts for the lack of an ambitious geographical meta-theory of globalization, but also important is the internal debate and diversity of theoretical perspective within the discipline when it comes to conceptualizing globalization. A key debate in this respect was that around the concept of spatial scale, with the broader globalization debate's sometimes binary usage (and conflation) of the global versus local scales serving to emphasize the need for a better theorization of different scales. A key contribution was that of Marston (2000) who reviewed the then state of the debate in political geographical thinking to conclude that within geographical theories, spatial scales (local, regional, national, global, supraregional) have no pre-given or fixed ontological status, but are produced and reproduced and continually transformed socially by the imperatives of social relations in world society and economy (cf Swyngedouw, 2004). In that sense, territories and scales were understood as being continually (re)produced (cf Agnew, 1999) and thus historically constituted (see also Agnew, 1994; Cox, 1998; Taylor, 2005; Smith, 2008). Neil Brenner, for example, argued that "the post-1970s wave of globalization has significantly decentred the role of the national scale as a self-enclosed container of socio-economic relations," and we are "witnessing a 're-scaling of territoriality'" which includes the increased "importance of both sub- and supranational forms of territorial organization" (Brenner, 1999: 435). In this geographical understanding of globalization, the phenomenon can be seen as the multiplication of geographical scales of social organization, linked to the changing spatial requirements of the latest phase of (capitalist) societal development. These are not seen as parallel or mutually exclusive scalar geographies, but intersecting and overlapping ones, with places/territories implicated in the full variety of scalar reorientations. Thus, this provides significant theoretical tractions to the contention that we are not witnessing an end to geography in a globalizing totality but rather a rescaling of boundaries (upward, outward, and downward) and a reconfiguration of territories as they engage in multiscalar processes and politics (Amin, 2002).

Importantly, all these geographical thinkers argued during the 1990s that contemporary globalization was associated with the rise of new territorial scales of organization, and new territorial boundaries, rather than the replacement of old, national scales and boundaries by a unitary global field of meaning and action. Contra to Manuel Castells, they did not see the replacement of an old world as 'space of places' with a new 'space of flows' – at least in any simplistic sense. Nor did geographical thinking sit comfortably with the idea that globalization could be easily reduced to one process or phenomenon (in line with Anthony Giddens, Jan Arte Scholte, or David Held and his co-authors), at least beyond the very broadest level of generalization. Rather, geographical approaches argued that to fully understand globalization, there was a need to recognize and engage with the persistence of geographical unevenness and diversity.

However, in the last 15 years, this debate around globalization in geography has not remained static, and the concept of scale in this context has become heavily contested and debated (cf Berg, 2004; Marston et al., 2005). This was because a range of contributions argued that scalar concepts did not permit a sufficient understanding of how places, spaces, and territories were themselves being transformed through global societal integration (Thrift, 1999; Murdoch, 2006). Amin (2002) argues that the scalar approach provided "little concession" to the "possibility that traditional demarcations between spatial and territorial forms of organizations might be blurring" – a consequence of globalization processes. Amin's well-cited argument is that theories of globalization need to account for the ontological presence of the proximate and the

remote at the same time in a globalized world. Geographical theories have thus moved to recognize that places such as cities and nations continue to exist as territorial units, but now with different external orientations (e.g., as sites in GPNs or places dependent upon international investment or competitiveness) (Coe et al., 2008; Massey, 2007; Dicken, 2011) and different scalar involvement (e.g., national welfare policies, continental trade agreements, global environmental regulations, local tax regimes) (Bulkeley, 2012; Bulkeley and Schroeder, 2012; Bridge and Le Billon, 2012). The boundary between place and space is not seen to be bending or disappearing – a line in shifting sand – but as a remaining line of distinction between 'in-here' relations and institutions, and those 'out-there' (cf Amin, 2002; Massey, 2005). Global connectivity is seen in that sense as leading to the 'nesting' of territories into multiple scales, rather than territorial places becoming uncoupled from a geography of scales within which they 'nest' (Lagendijk, 2002).

In the next section, we move to consider how this more recent work has opened up a new set of theoretical approaches to globalization which further problematized the dualist distinction between the 'local' and the 'global' in much globalization theory, and has moved to a 'postscalar' approach which increasingly regards the reconfiguration of the spatiality of social relations as a central aspect of contemporary globalization.

The Need for a Relational Approach to Space and Place

In the early 1990s, Doreen Massey (1994) coined the oxymoronic phrase 'a global sense of place' to deliberately seek to conceptualize localities in 'nonscalar' or 'relational' terms, arguing that places in today's globalized world increasingly should be understood as sites of juxtaposed social realities, shaped by the layered historical connections and geographical links that are to be found within and beyond them (Massey, 1994). Their character as sites of or physical contexts for agglomerated activity has thus far less to do with the properties of a territorial system (e.g., localized linkage, self-regeneration, local governance) than with the effects of spatial and temporal exposure and connectivity (e.g., continual and open-ended change, juxtaposition of differences, overlapping of networks of different global connections). Two of Massey's now famous examples of places as 'open, porous, hybrid' (Massey, 1999: 21) are Mexico City, with its varied architectures revealing layers of global Spanish and American-Indian historical connections, and Kilburn High Road in London, with its juxtaposition of English, Irish, and Indian diaspora lifestyles and connections. In her more recent book on London, Massey (2007) develops this argument about the way in which world cities are relational places, whose political, economic, and sociocultural nature can only be understood through an analysis of the intersection of local and distant relationships. Her account seeks to unravel how social inequality, industrial restructuring, and political identity are all caught up in a variety of global-scale interconnections of differing strengths and influences at different points in the last 50 years.

It is this approach that marks Massey's distinctly geographical approach to the spatiality of what are often loosely referred to as 'globalization processes,' and how they are manifest in place. For Massey, the spatial in general is "a sphere of the juxtaposition, or co-existence, of distinct narratives, as the product of power-filled social relations" (Massey, 1999: 21). In this respect, although exemplified well in a leading world city, a global sense of place is needed not only for cosmopolitan places such as London, but for *all places*. Every place may be imagined as 'particular articulations of these social relations' which includes 'local relations 'within' the place' and also 'those many connections which stretch beyond it' (Massey, 1999: 22). The important consequence of this insight is that places are open, porous, and hybrid – their uniqueness comes not from some mythological essence of rootedness or an isolated history, but from the absolute distinctiveness of all the mixtures of near and far influences that are captured by and through them over time (cf Massey, 2005).

Such a conceptualization of the spatiality of places stands in stark contrast to the rather simplistic homogenization view of globalization making all places the same – a perspective that had gained some traction in its most simplistic form across in popular debates about globalization. On the contrary, what globalization becomes in this framework is the contemporary and ongoing restructuring of space/time through places. For Massey, globalization is thus a "remaking of those, inherited but always temporary and provisional, spaces, places, cultures which are themselves the hybrid products of previous restructurings" (Massey, 2005). But of course places are themselves composites of multiple social relations – individual, collective, organizational, and institutional – that are differentially constituted through proximity and distance (Jones, 2009). This is where geographical thinking reaches the limits of the utility of understanding anything as existing at one scale; the nub of this issue is that it is arguable that any given set of social relations notionally 'within' a locality can in fact be understood to exist discreetly at one scale or the another. Local encounters are no more 'essential' than nonlocal practices and customs (Swyngedouw, 1997; Amin, 2002) (even if a purely local or nonlocal practice could be identified which is itself problematic (cf Jones, 2010)). Recent geographical thinking about economic practices, for example, has begun to seek the translocal extension of certain practices which are constituted simultaneously in their enactment by proximate co-present social relationships and those mediated at a distance (often by information or communications technologies) (Beaverstock et al., 2010; Hall et al., 2009; Faulconbridge and Muzio, 2009). The development of transnational firms and their associated global economic productive spaces is one fertile example of how these ideas have been developed, and recent work on the nature of cities and regions in today's global world is another. We will now consider each of these in turn to understand these arguments in more depth.

Globalized Corporate Spaces

Geographical thought on the nature of the global economy has made increasing use of a relational approach (Yeung, 2005; Bathelt, 2006; Jones, 2014) that has focused on the nature of the relationships between actors in the global economy. This

relational approach has been both at the level of firms and between key individuals within firms and other economic organizations in order to understand the complex spatiality of global economic transformations. In theorizing economic globalization as producing a range of global economic spaces (territorial, social, organizational, technological) (cf Jones, 2009), economic geographers have developed a range of conceptual cuts at demarcating sets of social relations that provide insight into the complex spatialities of global transformation. I want to briefly identify two of the most significant fields of work that have been developing this postscalar relational approach.

The first is the body of work that surrounds the concept of the GPN. Originating in the empirical of the so-called Manchester school on transnational firms and the geography of global production activities (Dicken et al., 2001), the development of these ideas predates the emergence of the mainstream globalization debate in the 1990s. Peter Dicken's work in the 1970s and 1980s began to outline the theoretical implications of an empirically identifiable 'global shift' in the location of manufacturing industries to the (then) newly industrializing economies of Asia (Dicken, 1986). By the 1990s Dicken argued that this shift constituted a radical geographical organization in production in the global economy, and developed with colleagues the concept of the GPN to conceptualize how transnational firms were locating different aspects of production process across a planetary scale and through complex relationship with subsidiaries and supplier firms (cf Coe et al., 2004). Important in these analyses has always been a reconceptualization of the spatial existence of firms as economic actors in an era of global economic integration. The GPN literature that developed during the 1990s posed an important counterbalance to the narrower epistemological framings of firm-level globalization that were being utilized in international business or management studies – notably, accounts of the emergence of multinational enterprises and their development into more transnational forms while remaining strongly wedded to their home economy territories (e.g., Rugman, 1996, 2001). GPN work within economic geography argued on the basis of detailed empirical studies of firms and sectors that TNCs were increasingly organizing production through highly complex and uneven planet-level networks through branch plants, subsidiaries, outsourced supplier networks, and just-in-time production techniques (Coe, 2012). Importantly, this work also demonstrated how corporate success in the global economy for these TNCs was increasingly tied into complex webs of relationships both embedded in but also transcending local and global-level contexts (Hess, 2004). In this respect, this important field of economic geography has reconfigured understanding of the spatial existence of firms and the complex relationship between territorial production sites and distant actors.

A second strand of economic geography has, however, also developed a relational approach using the role of key individual firm managers and economic practices as a contrasting means to better understand the organization and development of global economic activity. In particular, a body of work has used such concepts to theorize the development of a range of 'postindustrial' sectors that are becoming increasingly important in both developed and developing economies (for example, finance, advanced business services, cultural and creative industries). At both the theoretical and empirical levels, geographical work has sought to understand the key role of globally connected and mobile senior managers in shaping the way in which firms globalize production, enter new markets, and engage in innovation. Geographers have thus engaged with the nature of 'corporate globality' (Jones, 2003; Beaverstock, 2004; Strom and Mattsson, 2006; Hermelin and Rusten, 2007) with respect to understanding how transnational firms generate coherence in strategic, operational, and coherence terms. Geographical work has also shed new insight into how specific industries have globalized in different ways, shaped by the specific nature of the work processes (Gertler, 2003) and mobilities (Beaverstock et al., 2010), geographies of their markets (Zook, 2001; Gluckler, 2006), access to financial networks (Wray, 2012), or the need for certain types of labor (Cumbers et al., 2008). Examples include the uneven nature of internationalization in legal service firms, the transnational financing of large infrastructure projects (O'Neill, 2009), the very specific way in which the global architectural industry has emerged (McNeill, 2008), or 'global managers' who can manage complex global organizations in multiple intercultural contexts (Schlunze et al., 2012).

Cities and Regions in a Global World

Another strand of urban and regional geographical work has made significant contributions around how we might theorize contemporary economic agglomeration and the spatiality of global economic actors within regions, also increasingly drawing on nonscalar and relational approaches. Both economists and economic geographers have developed a body of work that seeks to better understand the clustering of economic activity in cities and regions as the outcome of powers of proximity, arguing that in an era of greater globalization locally embedded aspects of economic competitiveness have become ever more important. Much focus has thus been on the way that physical or territorial proximity leads to economic success by encouraging trust, reducing transaction and communication costs, and creating knowledge spill-over and/or creative buzz (cf Storper and Venables, 2004; Glaeser, 2011) and accounts which continue to see an economic place for regions or cities contrast the global as a scale of distanciated links, fast flows, and world-scale processes, with cities' regions as the sites of face-to-face-contact, or centers of agglomeration among global networks (e.g., clusters of transnational head offices) (Pain, 2008).

Yet a relational conception of these places provides a more sophisticated understanding of the nature of these territorial agglomerations of economic activity and reveals the complex nature of proximity and distance captured through these physical territorial spaces. Geographical work contrasts with the spatial epistemology of economists by seeing cities and regional economic clusters (for example, regions such as Silicon Valley) as contexts in which diverse geographies of linkage coalesce rather than just being concentrations of activity in territorial spaces (cf Ibert, 2010; Storper, 2013). Contemporary information and communications technologies, combined with global transportation infrastructure, mean that economic activity in a cluster of firms or an urban business

district is a continual mélange of near and distant relationships, played out across an uneven multiscalar set of dynamic relationships. For example, a software firm working with Facebook or Google in Palo Alto (in Silicon Valley) is reliant on a range offace-to-face interactions in its daily activities between co-present employees mixed with ICT-mediated interactions with others in firms elsewhere in the region, the United States, and planet-wide. The same applies to a media firm in London, small ICT firms in west Africa, or the research and development activities of engineering firms in northern Scandinavia (Grabher, 2006; Murphy, 2013; Ahlqvist, 2013). What matters is the interplay of a web of multiple relationships that produce proximity in a range of way that exceeds the physical around technological, social, cultural, and organizational dimensions (Jones, 2009). Global business spaces exist in this manner, always articulating in part in and through places, but not reducible to the attributes or confines of those places.

In this way, a relational understanding of global economic space recognizes places as clusters of overlapping network sites, without seeing in them only local economies of association restricted in their scope to a territorially localized economic system. It views with skepticism territorial metaphors such as 'local fixity vs global flow,' 'embedding the global,' or 'the local in the global' which often unhelpfully 'defend' the space of local economy against global 'encroachment.' It also dispels the binary division of these spatial categories. Instead, it highlights that local economic activity which is part of, and inseparable from, proximate and distanciated transactions, and that whatever counts as the local has to be seen as a networked space that exists dynamically in and through territorial spaces, not as a purely territorial phenomenon itself. Such work therefore seeks to develop a theoretical approach to understanding how place-specific and 'translocal' relationships are simultaneously and continuously caught up in the way veconomic activity becomes embedded in certain regions (Hess, 2004; Jones, 2008).

Conclusion

This article has aimed to provide an overview of how geographical thinking has offered a range of epistemiological and theoretical insights into the nature of contemporary globalization, and how social scientists in other disciplines might make use of its approach in developing more sophisticated approaches to the spatiality of globalized phenomenon. Geographical thinking has in many respects opened up a series of challenges for social scientific thinking to consider critically how many of the actors and entities that are the subject of social scientific enquiry – whether they are nations, cities, territories, corporations, organizations, or even individuals – exist as spatial phenomenon in today's increasingly integrated global society and economy. Relational and nonscalar thinking within geography reconfigures the terms of the social scientific debate about the impact globalization on places, as well as how we understand the nature of the spaces in which social phenomena exist and in which social practices are undertaken. This presents a potentially powerful toolkit for a range of objects of study that concern social scientists whether they are economic, political, or cultural phenomenon in today's globalized world.

The latter part of this article illustrated this approach through discussion of two strands of geographical thinking around global economic actors and the nature of city and regional economies, but within current human geographical work on globalization there are many more possible examples of how nonscalar and relational approaches (loosely defined) to space and place are being deployed. Other important areas of geographical work are concerned with the nature of urban politics and cultural/political practices are undertaken through different spatial configurations and are fluid in space and place (Amin and Thrift, 2002; Amin, 2012; Brenner, 2014). Equally political geographical work has sought to understand the multilayered spatialities of governance that perforate boundaries of territorial scalar units, as well as the role of nongovernmental actors in the politics of place in the contemporary global society (Taylor, 2005; Sparke, 2005). In that sense, the aim of this article has been to provide a partial insight into how geographers have addressed the concept of globalization, rather than a comprehensive overview of the breadth of this work. There is a great deal of human geography that could have been drawn upon to illustrate further examples, since with the discipline's central concern for understanding the nature of spatiality globalization is often at the center of contemporary geographical concern. Central to this is the capacity of geographical thinking to pose a challenge to a substantial proportion of ongoing work on globalization that still tends to construct globalization processes as unifying and homogenizing transformations (albeit in a more nuanced way than in the 1990s), rather than putting the diverse and uneven nature of global interconnectedness at the center of social scientific analysis. In that respect, geographers continue to have a significant and important offering to those who would better understand the nature of globalization across a wide range of social science disciplines.

See also: Economic Geography; Finance, Geography of; Global Migration; Industrial Geography; International Business: Geographic Aspects; Networks, Geography of; Region: Geographical; Service Economy, Geography of; Space and Social Theory in Geography; Territoriality: Geographical; Urban System in Geography.

Bibliography

Ahlqvist, T., 2013. Engineering spatial fixes: relational spatial practices, strategic topological approach and the construction of a Finnish science technology district. Geografiska Annaler: Series B, Human Geography 95 (1), 89–109.

Agnew, J., 1994. The territorial trap: the geographical assumptions of International Relations Theory. Review of International Political Economy 1 (1), 53–80.

Agnew, J., 1999. Mapping political power beyond state boundaries: territory, identity and movement in world politics. Millennium – Journal of International Statistics 28 (3), 499–521.

Amin, A., 2002. Spatialities of globalization. Environment & Planning A 34, 385–399.

Amin, A., 2012. Land of Strangers. Polity, Cambridge.

Amin, A., Thrift, N., 2002. Cities: Reimagining the Urban. Polity Press, Cambridge.

Appadurai, A., 1996. Modernity at Large: Cultural Dimensions of Globalization. University of Minnesota Press, Minneapolis.

Bathelt, H., 2006. Geographies of production: growth regimes in spatial perspective 3 – toward a relational view of economic action and policy. Progress in Human Geography 30 (2), 223–226.

Beaverstock, J.V., 2004. 'Managing across borders': knowledge management and expatriation in professional legal service firms. Journal of Economic Geography 4 (2), 157–179.
Beaverstock, J., Derudder, B., Faulconbridge, J., Whitlox, F. (Eds.), 2010. International Business Travel in the Global Economy. Farnham, Ashgate.
Berg, L., 2004. Scaling knowledge: towards a critical geography of critical geographies. Geoforum 35 (5), 553–558.
Berger, S., 2000. Globalization and politics. Annual Review of Political Science, vol. 3.
Bhagwati, J., 2004. In Defense of Globalization. OUP, Oxford.
Bisley, N., 2007. Rethinking Globalization. Palgrave Macmillan, Basingstoke.
Boyer, R., Drache, D., 1996. States against Markets: The Limits of Globalization. Routledge, London.
Brenner, N., 2014. Implosions/Explosions: Towards a Study of Planetary Globalization. Verlag, New York.
Brenner, N., 1999. Globalization as reterritorialization: the re-scaling of urban governance in the European Union. Urban Studies 36 (3), 341–351.
Brenner, N., Keil, R. (Eds.), 2006. The Global Cities Reader. Psychology Press, New York.
Bulkeley, H., 2012. Governance and the geography of authority: modalities of authorisation and the transnational governing of climate change. Environment and Planning A 44 (10), 2428–2444.
Bulkeley, H., Schroeder, H., 2012. Beyond state/non-state divides: global cities and the governing of climate change. European Journal of International Relations 18 (4), 743–766.
Bridge, G., Le Billon, P., 2012. Oil. Polity, Cambridge.
Castells, M., 2009. The Rise of the Network Society. In: The Information Age: Economy, Society and Culture, second ed., vol. I. Blackwell, Oxford.
Coe, N.M., Hess, M., Yeung, H., Dicken, P., Henderson, J., 2004. 'Globalizing' regional development: a global production networks perspective. Transactions of the Institute of British Geographers 29, 468–484.
Coe, N., Dicken, P., Hess, M., 2008. Global production networks: realizing the potential. Journal of Economic Geography 8 (3), 271–295.
Coe, N., 2012. Geographies of production II: a global production networks A–Z. Progress in Human Geography 36 (3), 389–402.
Cox, K., 1998. Spaces of dependence, spaces of engagement and the polities of scale or looking for local politics. Political Geography 17, 1–23.
Cumbers, A., Nativel, C., Routledge, P., 2008. Labour agency and union positionalities in global production networks. Journal of Economic Geography 8, 369–387.
Dicken, P., Kelly, P.F., Olds, K., Yeung, H., 2001. Chains and networks, territories and scales: towards a relational framework for analysing the global economy. Global Networks 1 (2), 89–112.
Dicken, P., 1986. Global Shift, first ed. Sage, London.
Dicken, P., 2011. Global Shift: Mapping the Changing Contours of the World Economy, sixth ed. Sage, London.
Dunning, J., Lundan, S., 2008. Multinational Enterprises and the Global Economy. Edward Elgar, Cheltenham.
Edwards, T., Rees, C., 2006. International Human Resource Management: Globalization, National Systems and Multinational Companies. Pearson, Harlow.
Elden, S., 2005. Missing the point: globalization, deterritorialization and the space of the world. Transactions of the Institute of British Geographers 30 (1), 8–19.
Faulconbridge, J., Muzio, D., 2009. The financialization of large law firms: situated discourses and practices of reorganization. Journal of Economic Geography 9 (5), 641–661.
Friedman, T., 2000. The Lexus and the Olive Tree. Harper Collins, London.
Friedman, T., 2007. The World Is Flat: The Globalized World in the Twenty-First Century, second ed. Penguin, New York.
Fung, V., Fung, W., Wind, Y., 2007. Competing in a Flat World: Building Enterprises for a Borderless World. Pearson, Harlow.
Garrett, G., 2000. The causes of globalization. Comparative Political Studies 33 (6–7), 941–991.
Gertler, M., 2003. Tacit knowledge and the economic geography of context, or the indefinable tacitness of being (there). Journal of Economic Geography 3, 75–99.
Gertsen, M., Søderberg, A., Zølner, M. (Eds.), 2012. Global Collaboration: Intercultural Experiences and Learning. Palgrave Macmillan, Basingstoke.
Giddens, A., 1990. The Consequences of Modernity. Polity, Cambridge.
Glaeser, E., 2011. Triumph of the City: How Our Greatest Invention Makes US Richer, Smarter, Greener, Healthier and Happier. Pan Macmillan, New York.
Gluckler, J., 2006. A relational assessment of international market entry in management consulting. Journal of Economic Geography 6, 369–393.
Grabher, G., 2006. Trading routes, bypasses, and risky intersections: mapping the travels of 'networks' between economic sociology and economic geography. Progress in Human Geography 30 (2), 163–189.
Hall, S., Beaverstock, J., Faulconbridge, J., Hewitson, A., 2009. Exploring cultural economies of internationalization: the role of 'iconic individuals' and 'brand leaders' in the globalization of headhunting. Global Networks 9 (3), 399–419.
Harvey, D., 1989. The Condition of Postmodernity. Blackwell, Oxford.
Harvey, D., 2003. A Brief History of Neoliberalism. OUP, Oxford.
Harvey, D., 2006. Spaces of Global Capitalism: Towards a Theory of Uneven Geographical Development. Verso, London.
Held, D., McGrew, A., Goldblatt, D., Perraton, J., 1999. Global Transformations: Politics, Economics and Culture. Polity, Cambridge.
Henderson, J., Dicken, P., Hess, M., Coe, N., 2002. Global production networks and the analysis of economic development. Review of International Political Economy 9 (2), 436–464.
Hermelin, B., Rusten, G., 2007. The organizational and territorial changes of services in a globalized world. Geografiska Annaler: Series B, Human Geography 89 (Suppl.1), 5–11.
Hess, M., 2004. 'Spatial' relationships? Towards a reconceptualisation of embeddedness. Progress in Human Geography 28 (2), 165–186.
Hirst, P., Thompson, P., 1999. Globalization in Question, second ed. Polity Press, Cambridge.
Holton, R., 2001. Globalization and the Nation-State, second ed. Palgrave Macmillan, Basingtoke.
Hopper, P., 2006. Understanding Cultural Globalization. Polity, Cambridge.
Ibert, O., 2010. Relational distance: socio-cultural and time-spatial tensions in innovation practices. Environment and Planning A 42, S187–S204.
Jacob, O., Neumann, I., September 2006. Governance to governmentality: analyzing NGOs, states, and power. International Studies Quarterly 50 (3), 651–672.
Jones, A., 2003. Management Consultancy and Banking in an Era of Globalization. Palgrave Macmillan, Basingstoke.
Jones, A., 2005. Truly global corporations? Theorizing organizational globalisation in advanced business-services. Journal of Economic Geography 5, 177–200.
Jones, A., 2008. Beyond embeddedness: economic practices and the invisible dimensions to transnational business activity. Progress in Human Geography 32 (1), 71–81.
Jones, A., 2009. Theorizing global business spaces. Geografiska Annaler: Series B, Human Geography 91 (3), 203–218.
Jones, A., 2010. Globalization: Key Thinkers. Polity, Cambridge.
Jones, A., 2014. Geographies of production 1: relationality revisited and the 'practice shift' in economic geography. Progress in Human Geography 38 (4), 605–615.
Lagendijk, A., 2002. Beyond the regional lifeworld against the global systemworld: towards a relational–scalar perspective on spatial–economic development. Geografiska Annaler: Series B, Human Geography 84 (2), 77–92.
Mann, M., 1997. Has globalization ended the rise and rise of the nation-state? Review of International Political Economy 4 (3), 472–496.
Massey, D.B., 1994. Space, Place and Gender. Polity, Cambridge.
Massey, D.B., 1999. Power-Geometries and the Politics of Space Time. Hettner, Berlin.
Massey, D., 2005. For Space. Sage, London.
Massey, D., 2007. World City. Polity, Cambridge.
Marston, S.A., 2000. The social construction of scale. Progress in Human Geography 24 (2), 219–242.
Marston, S., Jones, J.P., Woodward, K., 2005. Human geography without scale. Transactions of the Institute of British Geographers 30 (4), 416–432.
McGrew, A., Lewis, P. (Eds.), 2013. Global Politics: Globalization and the Nation-State. John Wiley & Sons, Oxford.
McNeill, D., 2008. The Global Architect: Firms, Fame and Urban Form. Routledge, New York.
Murdoch, J., 2006. Poststructuralist Geography. Sage, London.
Murphy, J.T., 2013. Transforming small, medium, and micro-scale enterprises? Information-communication technologies (ICTs) and industrial change in Tanzania. Environment and Planning A 45 (7), 1753–1772.
O'Brien, K., 1991. Global Financial Integration: The End of Geography. Council on Foreign Relations Press, New York.
Ohmae, K., 1995. The End of the Nation State. Harper Collins, London.
O'Neill, P.M., 2009. Infrastructure investment and the management of risk. In: Clark, G.L., Dixon, A.D., Monk, A.H.B. (Eds.), Managing Financial Risks: From Global to Local. Oxford University Press, Oxford, pp. 163–188.
Pain, K., 2008. Spaces of practice in advanced business services: rethinking London–Frankfurt relations. Environment and Planning D: Society and Space 26 (2), 264–279.

Palan, R., Murphy, R., Chavagneux, C., 2013. Tax Havens: How Globalization Really Works. Cornell University Press.
Ritzer, G., 2009. Globalization: A Basic Text. John Wiley & Sons, Oxford.
Ritzer, G., 2011. The McDonaldization of Society 6. Pine Forge Press, Thousand Oak, CA.
Rugman, A., 1996. The Theory of Multinational Enterprises: The Selected Scientific Papers of Alan M. Rugman, vol. 1. Edward Elgar, Cheltenham.
Rugman, A., 2001. The End of Globalization. Random House, London.
Sassen, S., 2001. The Global City, second ed. Princeton University Press, Princeton.
Sassen, S., 2007. A Sociology of Globalization. W W Norton, New York.
Schlunze, R., Agola, N., Baber, N., 2012. Spaces of International Management: Launching New Perspectives on Management and Geography. Palgrave Macmillan, Basingstoke.
Shamir, R., 2005. Without borders? Notes on globalization as a mobility regime. Sociological Theory 23 (2), 197–217.
Sheppard, E., 2002. The spaces and times of globalization: place, scale, networks, and positionality. Economic Geography 78 (3), 307–330.
Smith, N., 2008. Uneven Development: Nature, Capital, and the Production of Space. University of Georgia Press, Athens, GA.
Sparke, M., 2005. In the Space of Theory: Postfoundational Geographies of the Nation State. Minnesota University Press, Minneapolis.
Storper, M., 2013. Keys to the City: How Economics, Institutions, Social Interaction and Politics Shape Development. Princeton University Press, Princeton.
Storper, M., Venables, A.J., 2004. Buzz: face-to-face contact and the urban economy. Journal of Economic Geography 4, 351–370.
Strange, S., 1997. Casino Capitalism. Manchester University Press, Manchester.
Strom, P., Mattsson, J., 2006. Internationalization of Japanese professional business service firms. The Service Industries Journal 26 (3), 249–265.
Swank, D., 2002. Global Capital, Political Institutions, and Policy Change in Developed Welfare States. Cambridge University Press, Cambridge.
Swyngedouw, E., 1997. Neither global nor local: 'Glocalization' and the polities of scale. In: Cox, K.R. (Ed.), Spaces of Globalization. Guilford, New York.
Swyngedouw, E., 2004. Globalisation or 'glocalisation'? Networks, territories and rescaling. Cambridge Review of International Affairs 17 (1), 25–48.
Taylor, P.J., 1999. Places, spaces and Macy's: place-space tensions in the political geographies of modernities. Progress in Human Geography 23 (1), 7–26.
Taylor, P., 2005. Leading world cities: empirical evaluations of urban nodes in multiple networks. Urban Studies 42 (9), 1593–1608.
Thrift, N., 1999. Steps to an ecology of place. In: Massey, D., Allen, J, Sarre, P (Eds.), Human Geography Today. Polity Press, Cambridge, pp. 295–322.
Tomlinson, J., 1999. The Globalization of Culture. Polity, Cambridge.
Wolf, M., 2004. Why Globalization Works. Yale University Press, Yale.
Wray, F., 2012. Rethinking the venture capital industry: relational geographies and impacts of venture capitalists in two UK regions. Journal of Economic Geography 12 (1), 297–319.
Yeung, H., 1998. Capital, state and space: contesting the borderless world. Transactions of the Institute of British Geographers 23 (3), 291–309.
Yeung, H., 2005. Rethinking relational economic geography. Transactions of the Institute of British Geographers 30 (1), 37–51.
Zook, M., 2001. Old hierarchies or new networks of centrality: the global geography of the internet content market. American Behavioural Scientist 44, 1679–1696.

Globalization in Education

Yong Zhao, University of Oregon, OR, USA; and Victoria University, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Globalization has brought about tremendous changes to every aspect of human life in the twenty-first century. Education, traditionally a local practice confined by geographical and political boundaries, has too been affected. This article reviews the emerging trends of globalization in education. Specifically, the article discusses the trends of global standardization of education, global enterprising of educational systems and institutions, global migration of students, and the need for global citizenship. The article concludes with discussions of the most challenging issues facing education in the age of globalization.

Humans have entered the age of globalization. Thanks to advances in communication and transportation technologies, the physical distance that used to separate human societies has been disappearing in recent decades. For some practical purposes, physical distance is dead, in that it no longer matters in terms of time and costs. As a result, the primary living space, that is, the space a human individual or society can act in and impact, has become global, no longer constrained by the physical location of their presence at the time (Zhao, 2009a). Information, people, and goods move as freely, fast, and frequently from one place to another on earth, as they used to in much smaller spaces such as a village or geographical region (Legrain, 2002; Giddens, 1991; Friedman, 2007; Stiglitz, 2006).

While there are different definitions of globalization, which as a concept has only begun to be written about explicitly over the past three decades, the gist is the process and result of worldwide economic, social, cultural, and political networks that could only exist in much smaller geographical spaces before. Globalization may not be new, according to some (Friedman, 2007). It has happened before but the scale of participation, the pace, the intensity, and frequency were much less in the 1600s than what we witness today. The nineteenth century steamboats undoubtedly shrunk the distance between Asia and Europe, but the efforts, time, and costs it took to get to Asia then were still prohibitive to most individuals, in comparison to today's jet airplanes. The cultural and political ignorance then practically prohibited most people from equal participation in global commerce.

Today, despite the different views about the consequences of globalization (Friedman, 2007; Stiglitz, 2006), it is theoretically possible and economically real that the majority of the people on earth are participants of the globalized world. We wear shirts that are made by people in many different countries. Hundreds of millions watch the same sports event simultaneously from all parts of the world thousands of miles away. McDonald's, the American fast food giant born in Chicago, has stores in over 120 countries. Suhas Gopinath, an Indian boy from a poor family, developed a multimillion-dollar company creating Web sites for American businesses when he was 14 years old (Salkowitz, 2010). In 2013, over 230 million people lived in a country different from where they were born, compared to about 150 million merely 20 years before (Connor et al., 2013).

The potential for anyone to engage in activities on a global scale without considering geographical distance has enabled multinational corporations to spread their production all over the world, market products to people in the remotest areas of Africa, and employ people from any part of the globe. It has also made it possible for any individual to take advantage of the same mechanism to join the global production chain, to market their talents to people in faraway places, and to consume goods produced in remote places. The implications of this potential are not fully understood and may never be as globalization continues to evolve and will always be a dynamic process. However, it is obvious that from now on, all human activities must be considered in the context of globalization as they will always be affected by forces of globalization.

Education is a universal human activity. Although traditionally conducted mostly in local contexts, it has been affected by globalization in recent years. But because of its historical nature of being an extremely local activity, governed by local economy, culture, and politics, education has been slow in response to globalization in comparison to economic activities. Judging from recent activities, this is likely to change. Globalization in education is speeding up and will probably become one of the most impactful changes in the twenty-first century.

This article reviews developments in globalization in education in the recent past and discusses the challenges and prospects of globalization in education in the future.

Global Standardization

The first major development in globalization and education is increasing global standardization of educational policies, student experiences, and teaching practices. For a variety of reasons to be discussed in this section, education providers, which have typically been local, have engaged in activities that aim to comply with and compete against a global standard, just like their business counterparts, which have worked to develop international standards and compete with each other to demonstrate their ability to produce better quality products with better practices that are more efficient. As a result, global organizations have been established to develop and market standards, organize competitions, and identify good practices.

The Emergence of Global Standards: The Case of PISA

The growth of the Programme for International Student Assessment (PISA) exemplifies the increasing trend of global standardization of education. PISA is given every 3 years in three subjects: reading, math, and science. Each round focuses on one subject. First launched in the year 2000 by the Organisation for Economic Co-operation and Development (OECD), the PISA has become the world's most influential and successful educational assessment program within just over a decade. Since its first round of testing in 2000, the number of participating countries and economies has increased from 32 (28 OECD member countries and four nonmember countries) to 65 in the most recent round administered in 2012. With over 510 000 students taking the test in 2012, the OECD claims it represented "about 28 million 15-year-olds globally" (OECD, 2010).

Technically PISA is a large-scale international test in only three subject areas, similar to the cross-national testing programs that have been in operation by the International Association for the Evaluation of Educational Achievement (IEA) since 1959. By the time PISA was launched, the IEA had already conducted multiple rounds of large-scale studies in mathematics, literacy, science, civic and citizenship, information literacy, including the well-known Third International Mathematics and Science Study (TIMSS) in the 1990s. TIMSS gained so much traction that the acronym was kept but changed to represent Trends in International Mathematics and Science Study. Despite the success of TIMSS in promoting global standards, it is PISA that has established itself as the *de facto* gold standard for both educational quality and global competitiveness.

Global Standard for Economic Competitiveness

PISA scores have been marketed as indicators of the quality of a nation's human capital, hence its economic prosperity in the future. PISA claims to "assess to what extent students at the end of compulsory education, can apply their knowledge to real-life situations and be equipped for full participation in society" (OECD, 2013b). In a research paper published by PISA/OECD, the authors make a direct connection between PISA performance and national economic growth. According to the authors, an increase of 25 points in PISA scores over the next 25 years would result in an aggregate gain of USD 115 trillion in GDP in OECD countries over the lifetime of the generation born in 2010. The gain would be USD 260 if all OECD countries could raise their average PISA performance to that of the best OECD performer, Finland (Hanushek and Woessmann, 2010).

This marketing strategy has been extremely successful in the age of globalization when all nations are concerned about their competitiveness against others. As a result, political leaders and national governments in many countries have accepted PISA as a measure of their future economic prosperity and thus frequently refer to PISA performances in national policy contexts. For example, the former Australian Prime Minister Julia Gillard set the goal for Australia to be placed top five on the PISA ranking by 2025 (Ferrari, 2012). U.S. President Barack Obama called Shanghai's students' top PISA performance in 2009 another Sputnik moment, evoking the launch of the first man-made satellite by the former Soviet Union, as if Chinese students PISA scores meant the same as the science and technological prowess behind Sputnik.

Global Standard for Educational Quality

The idea that PISA is the global gold standard for educational quality has been firmed established. As the standard for educational quality, PISA now commands the attention of world political leaders, media pundits, and the general public, as well as educational researchers. Standings on the PISA ranking have been equated with nations' educational quality. As such countries that have produce high PISA scores have been automatically accepted as having better education quality and those with lower scores lower quality. Governments, education leaders, and the education sector in countries with low performance or a decline have been criticized, while high-performing education systems have been idolized. For example, after students in Shanghai became the number one on the PISA scale, media stories immediately call them the smartest, the best educated in the world. Germany's poor ranking in the first round caused a national uproar. Doubts about the PISA as a gold standard for educational quality are virtually nonexistent as the few reports that do present doubts are drowned in the overwhelming questioning about why some countries have so much better education (Breakspear, 2012).

Such questions result in global idolization of high-performing countries. Since its first round, PISA has created at least two major education miracles: Finland and Shanghai. Prior to its top performance in the first few rounds of PISA, Finland was little known to the rest of the world. But thanks to PISA, Finland has become the Mecca for education reformers, attracting hordes of pilgrims from all over the world eager to learn how Finland has such a high-quality education system. Without the PISA, few would be interested in reading the now popular book *Finnish Lessons: What Can the World Learn from Educational Change in Finland?* (Sahlberg, 2011). Likewise, Shanghai became the new Mecca as soon as it was announced that students from this Chinese city scored number one in the 2009 round. Prominent education researchers in the United States quickly came up with books like *Surpassing Shanghai: An Agenda for American Education Built on the World's Leading Systems* (Tucker, 2011) to encourage learning from Shanghai.

Global Standardization of Educational Policy and Practices

One of the original purposes of cross-country studies was to understand the effects of policies and practices within and across educational systems. Cross-country studies also provide the opportunity for educational policy makers and practitioners to learn from each other. But PISA has pushed such learning to the level of global standardization. Given the belief that PISA marks the potential for economic competitiveness and educational quality, it suggests that whatever factors contributes to PISA scores should become universally accepted benchmarks. As a result there have been global attempts by individual countries, research groups, international consulting firms, and the OECD itself to identify and summarize policies and practices that can be universally accepted as best practices.

The OECD, in connection with Pearson Foundations, for example, produced *Strong Performers and Successful Reformers in Education: Lessons from PISA for the United States*, a book and a collection of video cases (OECD, 2011). McKinsey & Co, the New York-based international consulting firm, summarized the best practices in its report *How the World's Best-Performing School Systems Come out on Top* (McKinsey & Company, 2007). The same firm, using PISA as the primary indicator of education quality, also came up with an influential on teacher quality based on studies of teacher recruitment, training, retention, and accountability in high-performing countries on PISA to offer lessons about how to improve educational quality via improving teachers (Citation). The Grattan Institute, an independent research group in Australia, conducted a study on the high-performing countries in Asia and published an influential report *Catching Up: Learning from the Best School Systems in East Asia* (Jensen, 2012). The OECD has not been hesitant to directly offer strategies either (Bieber and Martens, 2011). Among others, it suggests a focus on social and gender equity; decentralized school management; a cooperative school structure; strong centralize assessment system and accountability measures with high and demanding educational standards (Schleicher, 2006).

The results have been remarkable. PISA has prompted major policy changes in many countries (Breakspear, 2012; Knodel et al., 2013). Many of the changes are strikingly similar: centralizing curriculum and standards making, global benchmarking in curriculum and standards, strengthening testing and accountability measures, and encouraging local autonomy and market competition so as to reach the same outcomes (Breakspear, 2012). The Global Education Reform Movement or GERM described by Pasi Sahlberg (Sahlberg, 2012) is an excellent summary of the global standardization of education phenomenon. In parallel are similar efforts to standardize teacher recruitment, preparation, and incentivization. Students' experiences are also increasingly homogenized globally as countries emulate or attempt to emulate successes of high-performing nations: more time on academic studies and focus on the PISA subjects – reading, math, and science.

Global Spread of Local Standards and Programs

While the PISA is stimulating the evolution of global standardization of education through collaboration among local governments, there is another growing trend of global standardization, that is, the global spread of standards and programs developed for a specific local context. In the age of globalization, as schools, parents, educators, and governments look for high-quality programs and standards, they have started directly 'buying' standards and programs developed elsewhere.

International Baccalaureate

The rapid growth of the International Baccalaureate (IB) in recent years is a telling example. Founded in Geneva, Switzerland, IB was originally developed as "a single program for internationally mobile students preparing for university" by "a group of talented, forward-thinking teachers at the International School of Geneva, with assistance from several other international schools" (IBO, 2013a). At its early stage, the majority of schools that use the IB program were international schools, with a few government schools.

The rapid expansion of the IB did not start after the 1990s. There were only 269 IB programs in 1989. The last decade of the twentieth century saw some acceleration in growth. The number of IB programs increased to nearly 889. The growth skyrocketed in the twenty-first century. In 2012, there were 4231 IB programs in over 140 countries around the world, an astonishing fivefold increase in 12 years (IBO, 2013b). More importantly the IB is no longer a high school program for internationally mobile students. It has become a complete education program for students aged 3–19 years with the original IB Diploma Programme (DP), IB Middle Years Programme, and IB Primary Years Programme that were added later, and the newest addition the IB Career-related Certificate Program. Over half of the IB schools today are not international schools serving mobile students.

In its essence, IB is a curriculum similar to other curricula offered within a local school but it has become a global franchise. As a complete curriculum, carries its own philosophy and standards of education. A school that decides to adopt IB cannot pick and choose what IB courses to offer, but adopt the entire program and of course its standards and philosophy. Students go through the program and complete the requirements following the IB prescription. In other words, students in IB programs all over the world have the same experience, in very much the same way as customers of McDonald's restaurants. In other words, IB is grown from a local school curriculum to a global standard.

The quick spread of the IB in the past decade is largely driven by the desire for global high standards of education, the same that fueled the PISA movement. To many, the IB (and the Advanced Placement or AP) "represent the curricular gold standard for secondary education" (Byrd et al., 2007: p. 7). Governments, philanthropy organizations, schools, and local communities around the world have been eager to promote the franchise of IB and the AP in the hope to upgrade their education to a global standard (Byrd et al., 2007). Tony Blaire, former British Prime Minister, once announced his support to develop an IB DP school in every local authority in England (Abrioux and Rutherford, 2013).

Localization of International Schools

Another burgeoning phenomenon of global standardization is the localization of international schools. Like the IB, international schools were initially established to offer an education to a highly mobile population – children of diplomats, staff of international businesses and organizations, and military personnel stationed away from their country of origin. International schools typically offer a curriculum that bears the standards of their country of origin, or different from the local curriculum where they reside. In other words, international schools are outposts of national education systems in other countries.

International schools have gone through a rapid period of growth and transformation in the age of globalization. "In the past, international schools reflected the process of international migration, with expatriates and diplomats wishing their

children to be educated in systems compatible with those of their home country" (Hayden and Thompson, 2008: p. 9). Hence until very recently the majority of international school students were expatriates. But "Today, locally born residents fill about 80% of spaces in international schools and drive about two thirds of the market expansion" (ICEF Monitor, 2013). The expansion has been astronomical: 153% increase in the number of international schools in just 12 years. At the beginning of the 2012–13 academic year over 6300 international schools were recorded. The number is expected to reach 11 000 by 2020 and will serve over 6 million students (ICEF Monitor, 2013).

The growth of international schools in the last decade has been driven by the same desire to have high-quality education or "a certain dissatisfaction with the quality of the national education offer; the desire to sponsors and families to foster a global outlook that is not tied to a particular culture or education system; and the expressed need to prepare students to live and work in a globalized world" (Hayden and Thompson, 2008: p. 9). An international curriculum is deemed as high standard. "Not only are new schools continually opening, but their ranks also grow whenever existing schools convert to an international curriculum, begin instruction using a foreign language (usually English), or open a satellite campus in another country" (ICEF Monitor, 2013).

The Irony of Global Standardization: East versus West

There are two distinct trends in the globalization standardization movements in terms of the origin of the 'gold standard.' In the mostly government-driven movement inspired by PISA and other international tests, the East, that is, East Asian countries seem to be the source of high-quality education because they have consistently produced high scores. Ironically, the fastest growth of international schools with curriculum standards originated from the West, mostly the United Kingdom and the United States, has been seen in East Asia. Today, over 50% of international schools are in Asia. China, for example, with only 22 international schools just a decade ago, had over 340 in 2012. In Singapore, Hong Kong, South Korea, and China, all with high PISA scores that the West admires, "government funded efforts have resulted in the development of entire international education hubs" (ICEF Monitor, 2013). It seems that the gold standard of education for the East Asian countries is entirely different from what has been identified by the PISA.

Global Enterprising

Globalization has also created the possibility for existing educational institutions and educators to imagine and evolve from local providers into global enterprises. The desire of schools to pursue high-quality education, especially internationally recognized qualifications, presents an attractive market for entrepreneurially minded education entities to export their products globally. The result is the emerging phenomenon of global enterprising on multiple levels: systems, schools, and teachers.

Enterprising Systems

Related to the global spread of local standards and curriculum discussed in the previous section is the entrepreneurial efforts to export educational qualification courses and exams to students and schools globally. Unlike the IB, students and schools do not have to adopt the programs in their entirety. Local schools do not have to change their entire school curriculum, thus making themselves as a retail store of different programs. This flexibility creates a vast market for different programs to coexist and motivates a large of group of players to enter the global market.

National qualification systems that were historically developed to serve their citizens have been exported to other countries. One of the systems that entered the global market earlier is from the United Kingdom. The various versions of its qualifications such as the O-Level, the A-Level, GCSE, and the IGCSE designed for students in England, Wales, and Northern Ireland (Tattersall, 2007) have been used in many commonwealth countries or former British colonies for a while. They have recently entered more countries, even government schools in sovereign countries such as China. What is interesting is that the offerings of the qualifications and supports have become commercial products that schools and students in other countries purchase. As a result, providers such as Cambridge International Examination and Edexcel of the global education business giant Pearson have become global enterprises. IGCSE, for example, was taken by students in 120 countries in 2010.

The Victoria Certificate of Education (VCE), a qualification system owned by the government of the state of Victoria, Australia has been actively marketing the qualifications and licensing it to schools outside Victoria. The VCE is now taught in eight Chinese schools under license through schools in Victoria (VCAA, 2013). Canada is another example. While there is no national curriculum or qualification in Canada to export, education ministries of its seven provinces all accredit offshore schools and offer them the right to teach Canadian curricula and award Canadian diplomas, for a fee. British Columbia even let public "school districts operate as profit-making intermediaries" (Godsall, 2011).

The AP offered by the College Board was originally designed for students in the United States in 1950s, but in recent years it has been expanding its global reach as well. It is not a government-backed program but generally accepted by higher education institutions in the United States. In the last decade the number of non-U.S. students taking the AP increased by four times from around 20 000 in 2002 to over 80 000 in 2012.

Enterprising Schools

Some schools and universities have also taken an entrepreneurial approach to education in the age of globalization. Although still at its early stage, we are seeing a growing trend of educational institutions, both postsecondary and K-12, evolving into global enterprises by exporting their products to other countries. These enterprising efforts take a number of forms, but primarily in the following ways.

Joint Programs: Global Supply Chain

The first form is very much like a global supply chain, that is, the final product, in this case students, is created by different institutions located in different parts of world. Higher education institutions from developed countries have been offshoring parts of the product-making process to other institutions. For example, a university in the United States can have 2 years of the 4 years its degree program requires taught by an institution in India or China. Such joint programs have been extremely popular because it saves the students and institutions costs and resources. In China, the number of such joint programs has grown from 0 in 1990 to nearly 2000 in 2012, not counting those not officially approved by the Chinese authorities. Over 1 million Chinese students have graduated from these programs by 2012.

Independent Campuses: Global Outposts

Another common form is building or renting a campus. Traditional basic education schools such as the Harrow School and the Dulwich College from England have begun their global enterprising journey. In recent years, they have built campus in multiple locations in Asia. The University of Nottingham opened its Ningbo campus in China in 2004. Most recently, New York University opened two campuses, one in Abu Dhabi and one in Shanghai. While the nature of their operations may vary slightly, these campuses are no question examples of these institutions' global enterprising ambition.

Distance and Online: Virtual Campus

Another form of global enterprising is through technology. While distance education programs have been pioneered by dedicated institutions such as the Open University in the United Kingdom, recent technological development has enabled more traditional schools and universities to attract students globally. One of the most amazing developments is Massively Online Open Courses (MOOCs). Starting in 2011, a number of well-financed consortia of prestigious universities in the United States have begun to offer free online courses to students around the globe. While it is too early to know how MOOCs may or may not affect higher education or K-12 education around the globe, it is clear though that traditional educational institutions can have access to a global audience without being physically present.

Global Migration of Students

Students pursuing education in a country where they were born is not a new phenomenon but globalization has certainly enabled more students to do so in recent years (She and Wotherspoon, 2013). As a result, international students are one of the fastest growing parts of the global education system. "The number of students enrolled in tertiary education outside their country of citizenship increased more than threefold, from 1.3 million in 1990 to nearly 4.3 million in 2011, representing an average annual growth rate of almost 6%. This is a greater increase than the overall rise in tertiary enrolments globally," says to an OECD report in 2013 (OECD, 2013a: p. 1).

Globalization is unquestionably one of the primary drivers behind the rapid growth. First, the increasingly globalized and competitive world places new demand on skills and qualifications, placing more value on tertiary education obtained in countries which have traditionally been viewed to have better universities and colleges. It is thus no surprise to see the Western developed nations that have the reputation of offering excellent higher education as the top receiving countries. In 2011, 77% of international students were studying in OECD countries, with over half of them in the top five hosting countries: the United States (17%), the United Kingdom (13%), Australia (6%), Germany (6%), and France (6%) (OECD, 2013a: p. 1).

The second reason that globalization fuels the growth of international students is lowered costs and barriers to study abroad as well as rising income. China, for example, did not begin to send any students to the Western countries until 1978. But today it has become the largest source of international students. In 2011, China had over 700 000 students studying abroad, 85% of whom in OECD countries (OECD, 2013a). This is largely because of China's political change after the Cultural Revolution and the rising middle class as a result of its participation in the global economy.

These two reasons, if they continue to be true, will greatly change the landscape of global student migration in the global education market. First, if and when high education system develops in developing countries, they will become competitors to the existing providers. For example, the number of foreign students in Korea has increased over 17-fold between 2000 and 2011, while the share of international students in the United States dropped from 23 to 17% during this period (OECD, 2013a). It is likely that China will become the next major player in attracting international students. China has already seen exponential growth in the number of overseas students over the last decade. In 1999, about 52 000 students from 166 countries studied in China; the number increased by six times to over 320 000 in 2012 (China Association for International Education, 2013). Likewise, as their economies develop, we can expect more international students from countries that are not major sources of international students. The next likely group may be from Southeast Asia, Latin America, and Africa.

Global student migration is not only limited to higher education. Although the number remains small, students at the basic education level have begun to study in other education systems.

Global Citizenship: Globalization's Demand on Education

Globalization places new demand on education and one of the primary challenges is to change its orientation from preparing only local and national citizens to global citizens. Global citizens must be globally competent. Global competencies have no universally accepted definition but it generally include "the skills, knowledge, and attitude to work effectively in our increasingly interdependent world," according to the

University of Wisconsin Task Force on Global Competence (University of Wisconsin-Global Competence Task Force, 2008, p. 3). The task force further states: "Foremost amongst these 'global competencies' are the abilities to communicate effectively across linguistic and cultural boundaries, to see and understand the world from a perspective other than one's own, and to understand and appreciate the diversity of societies and cultures. Students need to appreciate the interdependence of nations in a global economy and to know how to adapt their work to a variety of cultures" (University of Wisconsin-Global Competence Task Force, 2008, p.3).

Fernando Reimers, a Harvard University education professor, recently offered a similar definition that includes three dimensions: the affective dimension, the action dimension, and the academic dimension. The affective dimension is also called the ethical dimension, which refers to "a positive disposition toward cultural differences and a framework of global values to engage difference." The action dimension is also the skill dimension, which primarily means the "ability to speak, understand, and think" in a foreign language. The academic dimension has to do with academic knowledge of the world, hence also referred to as the disciplinary and interdisciplinary dimension, and that includes "deep knowledge and understanding of world history, geography, the global dimensions of topics such health, climate, and economics, and the process of globalization itself" (Reimers, 2009).

Consistent across most definitions of global competence is the issue of others. The hallmark of globalization is the death of distance, which has brought 'others' previously separated by distance into our lives and how to view them, interact with, and live with them has become a significant issue facing all of us. Thus the foundation of our ability to live in the global world is the knowledge, perspectives, and skills required to live and work with people who are different from us culturally, linguistically, economically, and religiously (Zhao, 2009b).

Global competence and global citizenship have been identified by many international organizations and governments as crucial for our children to succeed in the age of globalization. In Australia, for example, Asia literacy, which is part of global competency in that it reorients Australian's focus to Asia, a more important part of the globe than their traditional partners – Europe and the United States, has been recognized as essential for young Australians to succeed in the future. In the United States, the Asia Society and a host of other organizations have been working promoting global competencies in schools (Asia Society, 2008). The increasing number of international students and international education programs represent the strong interest of students in developing countries to gain global competency. However sadly, global competency has not factored in as major component of education provided in most schools around the world.

Conclusions: Major Themes and Issues

Globalization in education is still relatively new and constant evolving phenomenon. The article provides preliminary observations of some recent developments, which by far do not capture the dynamic and complex possibilities globalization may bring to education. However, even this preliminary review offers some intriguing and challenging themes and issues brought about by globalization.

Homogenization and Diversification

It seems clear that globalization has brought many attempts by national authorities and international organizations to standardize education. Thus it warrants concern about possible global homogenization of education. Given the vast diversity and inequality that exists around the globe, a globally homogenized education may not serve everyone well. Moreover, although we have entered the age of globalization, when it may work better for nations to develop a diversity of talents instead of competing with each other in the same domains, nations in general seem to be stuck in the Cold War age by engaging in an education arms race through international tests. However, globalization has enabled more grassroots educational exchanges, which makes it possible for a more diverse education scene in the world. Furthermore, what is paradoxical is that although nations are all working hard to out-educate others, there is tremendous interest in collaboration and exchanges of educational ideas, strategies, and programs among different countries. No country is hiding their education secretes from each other.

Educational Sovereignty

Historically education, at least that offered by government schools, is considered a national matter. However, globalization results in "the erosion in the autonomy of the nation-state in all matters, including educational policy matters" (Burbules and Torres, 2000: p. 15). As shown in this article, there are significant growth of importing and exporting of education standards, content, and qualifications among countries. Such exporting and importing poses serious challenges to countries sensitive to the issue of educational sovereignty. For example, China still would like to maintain control of education for its citizens, thus it still does not allow its citizens to attend international schools in China. It also maintains control of international joint education programs by requiring prior approval for overseas educational programs to operate in China. At the same time, in the interest to bring high-quality education and stimulate educational innovations, it seems necessary to allow a variety of successful programs, regardless of their place of origin, to enter national markets. Countries will have to work on this issue in the coming years, if not decades.

Brain Drain, Brain Gain, and the Flight of Talents

Developing countries have traditionally been concerned about losing their talents when their students go for advanced studies in other, mostly developed, countries, hence the issue of brain drain. In the age of globalization, this issue has changed. Increased mobility as well as technology enables high-level talents to move around different locations with ease and work remotely. Additionally multinational corporations recruit and deploy talents globally. Thus developing countries, should

they desire and are able to provide the necessary conditions to attract talents, they cannot only have talents originally from their own country but also talents from around the world to work for them, hence changing brain drain to brain gain because these talents have received advanced education that their countries may not be able to provide. The issue is who can attract the best talents, not who can educate the best talents.

A Final Word

Education will surely change in the age of globalization. The changes will pose tremendous changes to nations at both the policy and practice levels. But today, fundamental changes have not been instituted yet. What has happened so far is not transformational in that education is still conceived the same way as it was a century ago: a group of students working with one adult (the teacher) belonging to an institution that educates and certifies the education. It is still driven by institutions. A true revolution will be a new paradigm of education in which students are the driving force and organizer of education. Globalization certainly presents such possibilities.

See also: Affordable Private Schools in Developing Countries; Anthropological Approaches to the Study of Education: The United States and Beyond; Educational Planning Worldwide; International Organisations and Education.

Bibliography

Abrioux, M., Rutherford, J., 2013. Introducing the IB Diploma Programme. Cambridge University Press, Cambridge.

Asia Society, 2008. Going Global: Preparing Our Students for an Interconnected World. Asia Society, New York.

Bieber, T., Martens, K., 2011. The OECD PISA study as a soft power in education? Lessons from Switzerland and the US. European Journal of Education 46 (1), 101–116. http://dx.doi.org/10.1111/j.1465-3435.2010.01462.x.

Breakspear, S., 2012. The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance. OECD, Paris.

Burbules, N., Torres, C.A., 2000. Globalization and education: an introduction. In: Burbules, N., Torres, C.A. (Eds.), Globalization and Education: Critical Perspectives. Routledge, New York.

Byrd, S., Ellington, L., Gross, P., Jago, C., Stern, S., 2007. Advanced Placement and International Baccalaureate: Do They Deserve Gold Star Status? Thomas B. Fordham Institute, Washington, DC.

China Association for International Education, 2013. Tongji Shuju (Statistics) [Online]. Beijing. Available: http://www.cafsa.org.cn/index.php?mid=6 (accessed 01.01.14.).

Connor, P., Cohn, D.V., Gonzalez-Barrera, A., 2013. Changing Patterns of Global Migration and Remittances: More Migrants in U.S. And Other Wealthy Countries; More Money to Middle-income Countries. Pew Research Center, Washington, DC.

Ferrari, J., 3 September 2012. Labor's 'top five' goal for schools. The Australian. Retrieved from. http://www.theaustralian.com.au/national-affairs/policy/labors-top-five-goal-for-schools/story-fn59nlz9-1226463502869.

Friedman, T.L., 2007. The World Is Flat: A Brief History of the Twenty-First Century. Farrar, Straus and Giroux, New York.

Giddens, A., 1991. The Consequences of Modernity. Polity Press, Cambridge.

Godsall, D., 2011. Selling Canadian Teaching in China [Online]. BC Business, Vancouver. Available: http://www.bcbusiness.ca/careers/selling-canadian-teaching-in-china (accessed 01.01.14.).

Hanushek, E.A., Woessmann, L., 2010. The High Cost of Low Educational Performance: The Long-run Economic Impact of Improving PISA Outcomes. OECD, Paris.

Hayden, M., Thompson, J., 2008. International Schools: Growth and Influence. UNESCO, Paris.

IBO, 2013a. History of the International Baccalaureate® [Online]. Available: http://www.ibo.org/history/ (accessed 30.12.13.).

IBO, 2013b. Programme Growth Statistics [Online]. Available: http://www.ibo.org/facts/schoolstats/growth.cfm (accessed 30.12.13.).

ICEF Monitor, 2013. International School Enrolment on Pace to Reach $60 Billion by 2022 [Online]. ICEF. Available: http://monitor.icef.com/2013/03/international-school-enrolment-on-pace-to-reach-60-billion-by-2022/ (accessed 31.12.13.).

Jensen, B., 2012. Catching up: Learning from the Best School Systems in East Asia. Grattan Institute, Melbourne.

Knodel, P., Martens, K., Niemann, D., 2013. PISA as an ideational roadmap for policy change: exploring Germany and England in a comparative perspective. Globalisation, Societies and Education 11, 421–441.

Legrain, P., 2002. Open World: The Truth about Globalization. Abacus, London.

Mckinsey & Company, Inc., 2007. How the World's Best-Performing School Systems Come Out on Top. McKinsey & Company, New York.

OECD, 2010. OECD Programme for International Student Assessment [Online]. OECD, Paris. Available: http://www.pisa.oecd.org/pages/0, 2987,en_32252351_32235731_1_1_1_1_1,00.html (accessed 10.01.10).

OECD, 2011. Strong Performers and Successful Reformers in Education: Lessons from PISA for the United States. OECD, Paris.

OECD, 2013a. How Is International Student Mobility Shaping Up? OECD, Paris.

OECD, 2013b. Programme for International Student Assessment (PISA) [Online]. OECD, Paris. Available: http://www.oecd.org/pisa/aboutpisa/ (accessed 28.12.13.).

Reimers, F., 2009. Educating for global competency. In: Cohen, J.E., Malin, M.B. (Eds.), International Perspective on the Goals of Universal Basic and Secondary Education. Routledge Press, New York.

Sahlberg, P., 2011. Finnish Lessons: What Can the World Learn from Educational Change in Finland? Teachers College Press, New York.

Sahlberg, P., 29 June 2012. How GERM Is Infecting Schools Around the World. The Answer Sheet [Online]. Available: http://www.washingtonpost.com/blogs/answer-sheet/post/how-germ-is-infecting-schools-around-the-world/2012/06/29/gJQAVELZAW_blog.html#comments (accessed 20.11.12.).

Salkowitz, R., 2010. Young World Rising: How Youth, Technology, and Entrepreneurship Are Changing the World from the Bottom Up. Wiley, Hoboken, NJ.

Schleicher, A., 2006. Fundamentos y cuestiones políticas subyacentes al desarrollo de PISA. Revista de Educación, extraordinario. Madrid, pp. 21–43.

She, Q., Wotherspoon, T., 2013. International student mobility and highly skilled migration: a comparative study of Canada, the United States, and the United Kingdom. Springer Plus 2.

Stiglitz, J., 8 September 2006. Make globalisation work for everyone, review. The Straits Times, p. 25.

Tattersall, K., 2007. A brief history of policies, practices and issues relating to comparability. In: Newton, P., Baird, J.-A., Goldstein, H., Patrick, H., Tymms, P. (Eds.), Techniques for Monitoring the Comparability of Examination Standards. Office of the Qualifications and Examinations Regulator, London, pp. 324–328.

Tucker, M. (Ed.), 2011. Surpassing Shanghai: An Agenda for American Education Built on the World's Leading Systems. Harvard Education Press, Boston.

VCAA, 2013. International Projects for Schools and Educators [Online]. Melbourne. Available: http://www.vcaa.vic.edu.au/pages/schooladmin/international/index.aspx#H3N1009F (accessed 01.01.13.).

Zhao, Y., 2009a. Catching up or Leading the Way: American Education in the Age of Globalization. ASCD, Alexandria, VA.

Zhao, Y., 2009b. Needed: global villagers. Educational Leadership 67, 60–65.

Globalization: Legal Aspects

Alfred C Aman, Jr., Indiana University Maurer School of Law, Bloomington, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by B. De Sousa Santos, volume 9, pp. 6277–6284, © 2001, Elsevier Ltd.

Abstract

Law's role in globalization is often misunderstood, mainly with respect to whether 'the global' is its own distinct and unified field. To a large extent, this misunderstanding reflects the influence of neoliberalism, some versions of which treat globalization as a function of capitalism (thereby relegating it to the preserve of economics). Accordingly, one might not think that law has much to do with this phenomenon, beyond stepping out of the way, except where law contributes to the creation of markets and to ensuring that they function efficiently. But the reality of globalization challenges such formulations, and this article argues that law has an important role to play. The contexts in which law is relevant reflect the great diversity of relationships, circumstances, and legal arrangements under which globalization now develops. This article will examine three such contexts. The first deals with multilateral treaties, the second with corporate social responsibility and voluntary regulation, and the third with domestic law.

Introduction

Law's role in globalization is often misunderstood, mainly with respect to whether 'the global' is its own distinct and unified field. To a large extent, this misunderstanding reflects the influence of neoliberalism, some versions of which treat globalization as a function of capitalism (thereby relegating it to the preserve of economics) (Harvey, 2005). Such a formulation leaves little room for law, except in relation to international law and such 'global' institutions as the International Monetary Fund (IMF), the World Trade Organization (WTO), and other legal arrangements primarily associated with the liberalization of trade. From this standpoint, one might imagine that globalization is all about competition – a competition for markets and investments that is global in scale and increasingly intense as markets expand. Accordingly, one might not think that law has much to do with this phenomenon beyond stepping out of the way, except where law contributes to the creation of markets and to ensuring that they function efficiently.

But the reality of globalization challenges such formulations. Law has an important role to play, particularly in contexts in which economic activity generates human rights questions, such as child labor (de Feyter and Gomez Isa, 2005; Hallo de Wolf, 2011; International Labour Organization, 1999; Roe v. Bridgestone, 2007). This article discusses law's prominence in the creation, implementation, and contestation of globalization. The contexts in which law is relevant reflect the great diversity of relationships, circumstances, and legal arrangements under which globalization develops (Calliess and Zumbansen, 2010; Kingsbury et al., 2005). This article will examine three such contexts: multilateral treaties, corporate social responsibility (CSR) and voluntary regulation, and domestic law.

Globalization is embedded in our institutions – domestic and international, public and private, by virtue of legal arrangements (legislation, agency regulations, contracts, etc.) that draw global 'forces' into everyday life, and vice versa. It is not a unilinear process or geography 'out there,' but a dynamic relation across multiple regimes of public and private ordering (Aman, 2001a, 2004, 2006; Delbrück, 1993). Globalization is subject to a wide array of influences and control, and yields pervasive social effects – some of them broadly homogenizing, some of them diversifying in highly specific ways (Aman, 2009; de Sousa Santos, 2006; Harvey, 2006; Sassen, 2003). Understanding the relationship of globalization to law requires analysis of the interactions of markets, rights and bodies of law at all levels of government, domestic and international, as well as diverse processes of governance that involve norm creation, enforcement, and adjudication by state and nonstate actors alike (Calliess and Zumbansen, 2010; Halliday and Carruthers, 2009).

Moreover, globalization is neither external and superior to national states, nor primarily state centered. The essence of globalization is its denationalizing effect (Sassen, 2006; Delbrück, 1993). Thus it has been said that, "Globalization … may be defined as the process of denationalization of clusters of political, economic, and social activities" (Delbrück, 1993). In a globalized world, various nonstate actors, including multinational corporations (MNCs), play key roles precisely because they are not bound to any one jurisdiction (Backer, 2011). Globalization includes global and regional institutions such as the WTO and the European Union, but globalization is also a domestic phenomenon, driven by private actors and local law.

While this article is concerned with some phenomena that are global in the geographic sense, this is only one subset of our concerns with transnational law (TL) more generally. For example, there are some issues such as ozone depletion (see *infra*) that are truly global (i.e., affecting the whole world), but many pressing transnational issues do not involve universal or planetary stakes. Indeed, there are many domestic contexts in which law may properly be understood as 'transnational' – that is, in its functioning across bodies of law, public and private, domestic and international, rather than in a particular geography or international field. TL potentially encompasses legal relationships involving domestic and international law as well as private ordering

involving a host of nonstate actors who operate transnationally. The complex sociopolitical interrelationships resulting from these interactions affect institutions and individuals in a variety of ways at home and abroad (wherever these may be). Law features in all of these arrangements, highlighting the pluralism and polycentrity of law's 'many rooms' (Galanter, 1981). States play important roles in global governance issues, but so does the private sector – the latter becoming increasingly important even rivaling or substituting for state powers in some contexts (Backer, 2011).

The emergence of a global private sector has been accelerated by technological advances in high-speed worldwide communication (including transfers of funds). The mobility of capital has been especially important to the development of modern 'global' firms and their prominence in global capitalism (Harvey, 2010; Strange, 1997). Private decisions involving, for example, production, finance, and investment occur increasingly free of direct, individual state involvement. Moreover, MNCs may shop around the globe for the locations most favorable to their interests. Transnational corporations thus decide where it is most cost effective to locate various activities in the sometimes long value chains connected with the production and marketing of their goods and services. They may locate Research & Development in one country, component assembly in another, final assembly in another, and distribution networks in yet another. They also decide how much to customize the globally conceived product for local markets (Dunning and Lundan, 2008). The General Agreement of Tariffs and Trade and other multilateral agreements facilitated trade in goods; a substantial portion of world trade today takes place between and among divisions of the same company doing business in various locations around the globe (Dunning and Lundan, 2008). Trade intensifies other transnational phenomena. Migration often follows the itineraries of trading relationships as well as military engagements (Sassen, 1998; see also Mosk, 2005).

These factors do not mean that states are no longer important as influences on these businesses or that they are powerless to affect environmental problems on a global scale. Still, the regulatory role nation-states play is substantially different than it was in the past, because deregulation, privatization, and market-oriented regulatory approaches are more prevalent than they were a generation ago (Akech, 2009; Aman, 1998, 2001b; Cerny, 1990; Prakash and Hart, 2000). For other reasons, too, the efficacy of states is becoming more limited; for example, problems such as the environment, security, or some areas of public health cannot be resolved effectively by the way of individual nation-state action alone. The political networks involved in these and other issues (such as climate change, ozone depletion, the destruction of the rainforest and other biohabitats, and human rights generally) transcend nation-states and state institutions. That said, global problems do not always demand exclusively global solutions; local engagement remains relevant (Myint, 2012; Ostrom, 2010; Valverde, 2009). For example, as we shall see, infra, the Montreal Protocol dealing with ozone depletion benefited considerably from the domestic laws and politics in the United States that preceded its drafting. New bodies of TL are developing to deal with problems that are neither wholly domestic nor comprehensively global – sometimes working through domestic law, sometimes creating new authorities and powers (Cole, 2011; Paavola, 2011; Ostrom, 2010; Koh, 1994).

Transnational Law

The term 'transnational law' was coined by Phillip Jessup in his seminal lecture in 1956 entitled *Transnational Law*. In that lecture, he claimed that TL includes "all law which regulates actions or events that transcend national frontiers. Both public and private international law is included, as are other rules which do not wholly fit into such standard categories" (Jessup, 1956). TL develops in the gaps that exist between states but also emerges in other contexts. For example, it involves supply contracts between and among MNCs, *lex mercatoria*, and other bodies of law developed by nonstate actors wholly outside or sometimes in the shadow of the state (Shaffer, 2013; Mattli and Woods, 2009; Zumbansen, 2006; Teubner, 1996, 1987). Laws created by worldwide sports federations like the International Olympic Committee or the Fédération Internationale de Football Association – world soccer's governing body – the best practices of various global industries, international standards (Mattli and Buthe, 2003), and voluntary rules including corporate codes of conduct and other modes of self-regulation also fall within the purview of TL (Blackett, 2001; Cutler, 1999; Compa and Hinchliffe-Darricarrère, 1995). In short, the term 'transnational' refers to pluralistic, multicentered, and dynamic processes involving interrelationships among states and nonstate entities, whether across national boundaries or within them.

Depending upon the context involved, TL may involve domestic law, international law, some form of voluntary regulation, or various combinations of all the three. Increasingly, TL involves hybrid forms of power sharing among public and private actors. These hybrid arrangements are one prominent result of various contemporary forms of delegation of state power. Sometimes state power is delegated 'up' to an international entity such as the WTO; other times there may be a delegation or devolution of state power 'downward' to a substate jurisdiction, such as in a federal system with its member states. States also delegate power horizontally to private entities – for example, when a governmental agency outsources its regulatory functions to a private provider of these services.

Given the horizontal aspects of state delegations of power, and the fact that states are limited in their regulatory reach by their own jurisdictional limitations, it is not surprising that, increasingly, voluntary regulation becomes one of the main ways in which law develops outside of the state (Backer, 2011; Blackett, 2001). Moreover, by partnering with transnational corporations, states are able to achieve more in certain arenas than if they operated solely on their own. Corporations' powers as transnational legal actors may be significant, especially given the increasing competition between and among states with respect to attracting and retaining private investment. The influence of transnational corporations in relation to regulatory and tax issues encourages the flow of power from states to market actors and back again. But such shifts are not automatic.

They occur through legislation and regulatory and deregulatory procedures, and reflect the impact of domestic, state-centered law and politics on international and domestic fronts (Aman, 2001b, 2004, 2008; Freeman and Minow, 2009).

For all of the reasons discussed in this section, understanding how denationalizing processes work is an important part of understanding law's role in relation to globalization.

Denationalization

The denationalizing processes associated with TL are integral to the decentering effects of globalization, understood more broadly. Pollution, drug trafficking, terrorism, and global investment opportunities involve diverse challenges, but the processes involved in creating, implementing, and adjudicating the rules are alike in not necessarily being bound to or by any one particular jurisdiction. One way for states to regulate such activities successfully is to cooperate on a long-term basis, and this usually means pooling their sovereignties in some way (Delbrück, 2002). Such pooling may take the form of a multilateral treaty or a delegation of power to an international organization such as the WTO. The pooling of sovereignty and the delegation of state power to international bodies inevitably create denationalizing effects, as new actors bring their policy perspectives and national interests into the negotiating arena.

Similarly, markets, capital, technology, ideas, and culture – like people – are mobile, not necessarily tied to jurisdictional boundaries (Appadurai, 1990). Economic globalization developed around a commitment to the free flow of investment capital across borders. It previously had been common for states to try to regulate international capital flows, but unregulated capital flow has been synonymous with globalization for decades. Currently, however, new regulatory initiatives are emerging in some states, and their potential value as a means of managing volatility has been recognized by the IMF (Klein, 2012; International Monetary Fund, 2012). These regulatory efforts are inevitably restrained by both the inherent difficulty of unilaterally regulating cross-border activities and the potential risk of repercussions.

The very nature of certain kinds of global issues may obviate state action or bypass states altogether. Globalization thus draws attention to processes that are essentially deterritorialized and often independent of states. This is especially true of economic, social, and cultural processes, but it also includes various law and norm-creating processes that may develop among private actors (Shaffer, 2013; Abbott and Snidal, 2009). Private entities themselves may make the law they believe is necessary to govern their activities in complex transnational situations. TL can be and often is made by states, but it can also be made outside the state or, to the extent private lawmaking is influenced though not necessarily governed by state law, it can be made in the shadow of the state (Abbott and Snidal, 2009; Calliess and Zumbansen, 2010; Kahler and Lake, 2003; Shaffer, 2013). Such *de facto* delegations of state power to the private sector are key features of globalization.

Recognizing globalization as intersecting and multilayered fields of transnational actors, both inside and outside the state sphere, yields a richer and more dynamic account of the multiple legal relationships involved in the development and application of TL.

To illustrate these various aspects of TL, this article now turns to the three contexts in which TL is made, through a series of case studies. The cases illustrate treaty processes, voluntary regulation, and domestic law. All three cases involve the United States in some way – not because American experience defines globalization, but because the variety of 'angles' on the 'Americanness' in these scenarios is illustrative of diverse modalities of TL that are more broadly relevant.

Transnational Legislation: Multilateral Treaties and the Global Environment

Quite apart from issues facing multiple jurisdictions, some global issues are the result of activities, interests, or problems connected to specific transnational places or spaces over which no one state has control – such as the oceans (Law of the Sea) or outer space or the Arctic Circle. One example of this phenomenon involves the ozone layer in the stratosphere of the Earth. To respond to this problem, many nations developed the Montreal Protocol to reduce chlorofluorocarbon (CFC) emissions from around the globe, emissions that were causing the ozone layer to rapidly deteriorate. Ozone depletion presented global actors with a new and daunting set of challenges, which required innovation as well as adherence to traditional treaty processes.

Ozone depletion represented a global commons problem with a clearly defined source, but a less-defined solution. The Montreal Protocol's negotiators confronted a problem affecting a resource that no one state could effectively regulate – the Earth's atmosphere – yet the destruction of which would negatively affect everyone. But international standards without domestic support would accomplish very little.

In the case of the Montreal Protocol, domestic law and politics played an important role in creating the domestic legal and political environment necessary for the United States to pursue such a treaty in the first place. Moreover, nation-states were not the only players. Various nonstate actors played crucial roles in devising the regulatory approaches ultimately adopted. Moreover, the politics necessary to accomplish these goals was fueled not only by technological innovation and the prospect of new markets for new products, but by the ability of science to dramatize the human stakes involved for everyone, regardless of where they lived, and a politics willing to accept this science (Benedick, 1998).

Domestic law laid the groundwork for the treaty process. In the United States, the Congress had passed legislation to curb domestic use of CFCs (Toxic Substances Control Act, 1976). The participants of the Montreal Protocol came to the table from markedly divergent viewpoints – several represented constituencies that were heavily invested in the production and use of CFCs. Those involved in the process agreed that science offered the possibility of an analysis of environmental problems that could be shared by all countries. This was essential. Reliable scientific evidence helped to unify the international community around a political solution (Aman, 1992; Benedick, 1998). In 1987, the Montreal Protocol – a binding multinational treaty

regulating CFC consumption – was signed (Montreal Protocol on Substances That Deplete the Ozone Layer, 1987).

As Benedick details in his analysis of the treaty process and its aftermath, it was not just science – industry leaders in the United States and abroad (Europe especially) carefully matched each other's actions in gradually 'disarming' their industries of CFC production, and substituting ozone-friendlier substitutes. For example, in September 1988, the US Alliance for Responsible CFC Policy announced that it was in favor of the CFC phaseout, and the UKs Imperial Chemicals Industries followed suit soon thereafter. Industrialists in the Federal Republic of Germany, Japan, and other major industrial nations, seeing the path set forth by the British and Americans, began to draw down their CFC use as well. This industrial cooperation strengthened the signatories' momentum going into the Protocol negotiations, and contributed to the relatively seamless implementation process that followed the signing of the Protocol.

In short, a primary reason for the Protocol's success was that the problem and solution that the Protocol sought to address was clear: (1) this precise phenomena; (2) caused by these specific chemicals; (3) with these demonstrated risks; (4) affecting all people; (5) could be remedied by these specific actions, with available substitutes; (6) over a specific timeline (Sunstein, 2006; Benedick, 1998; Aman, 1992). The negotiators had a comprehensible, focused, and universally applicable problem before them. The ozone depletion threat was truly global in terms of the scope of its potential consequences, as well as the contributing factors to the problem. The Montreal Protocol was a success story when it comes to the creation and implementation of this type of TL, but the almost 'perfect storm' nature of the elements of its very success are illustrative of why such cooperation among multiple states at the global level is, and is likely to remain, so rare. Global warming highlights the immense difficulty of crafting a multilateral treaty when nations are unable to reconcile divergent interests, let alone agree on the underlying science involved. Nevertheless, the Montreal Protocol is a model of one kind of intersection of globalization and law, albeit one not easily duplicated.

TL in the Shadow of the State

The polycentric nature of global environmental problems and the different interests that states bring to their resolution are not confined to environment and global commons issues. Discussion in this section moves to another context in which they arise – in human rights questions involving, for example, labor relations and conditions of employment. Global uniformity in such matters is highly unlikely for a number of reasons, including the variation that exists among states in their economic, political, legal, and cultural orientations toward providing for social welfare, as well as their capacity to regulate employers and supply contractors who seek a competitive advantage in a low-wage workforce (Harvey, 2010).

Beginning as early as the 1960s, responding to tax and regulatory incentives, wage differentials, and other factors, Western companies began to close their domestic factories and outsource manufacturing abroad in certain industries (Goldin and Katz, 2007). Companies outsourcing labor no longer had to deal directly with a workforce. Transnational supply chains created by contract began to substitute for local labor forces governed by national and state labor laws. Many of the jurisdictions to which labor was outsourced either had no effective labor laws in place, or no effective means of enforcing the laws in place (Sum and Ngai, 2005).

In the 1970s, a number of external codes of conduct were promulgated by major international agencies, such as the UN, the Organisation for Economic Cooperation and Development, and the International Labour Organization (Compa, 2008; Organisation for Economic Co-operation and Development, 2008; International Labour Office, 2006). They set forth various declarations and guidelines to protect workers' rights. They were external codes because they were created outside the corporations involved; their value was largely symbolic but they were significant nonetheless (Compa, 2008; Perez-Lopez, 1993; Sullivan, 1983).

CSR embodied in corporate codes of conduct developed internally by the corporations themselves began in earnest in the 1980s and early 1990s when firms such as Nike, Reebok, Levi Strauss, and others produced codes of conduct internally that were designed to regulate the supply contractors in their respective global production chains (Compa, 2008; Compa and Hinchliffe-Darricarrère, 1995; Perez-Lopez, 1993; Levi Strauss & Co., 1992; Swoboda, 1992). Moreover, Western media outlets began to run stories calling attention to, for example, the extensive use of child labor or other inhumane labor conditions found in foreign factories manufacturing sneakers and apparel for these and other companies.

There has been progress in the promulgation of meaningful supply chain management and regulation on two fronts in particular. The development of these CSR codes is becoming more democratic in the sense that companies, unions, human rights groups, and community organizations are increasingly all involved in the development phase (Fransen and Kolk, 2007). This new generation of codes is called 'multi-stakeholder' codes (Compa, 2008). At the same time, these new codes have monitoring mechanisms that rely on outside parties, beyond the corporation itself (Compa, 2008; Keller, 2008; Fransen and Kolk, 2007). Monitoring groups include groups such as the Fair Labor Association (FLA), which is a coalition of universities and nongovernmental organizations, Social Accountability International, and other third-party monitors that add independence and legitimacy to the auditing process (Compa, 2008; Keller, 2008).

Self-Regulatory Processes

In these contexts, we observe a very different lawmaking dynamic than the one that drove the Montreal Protocol. The labor conditions that came to the attention of the general public were driven not by science, but by worker fatalities.

The legitimacy of a firm or an industry's attempts to make and enforce their own law may turn on the degree to which those laws and the processes resemble those of the states in which they operate (Ochoa, 2011; Scott, 2010); however, neoliberal reforms do not automatically translate into traditional forms of government regulation (Shamir, 2005). Indeed, as Shamir writes, "[s]uch a theory must come to grips with, and identify the pitfalls of, working within the non-governmental organization institutional paradigm, because the latter is heavily biased

toward the corporate hegemonic model of organization and implementation" (Shamir, 2005). As a consequence,

> [We] must also worry about the way the perceived grievances of oppressed, marginal, and exploited populations…are transformed into a meaningful political and legal voice by relatively affluent and secured…experts who often speak the language of and deploy the instruments of hegemonic rational organizational and managerial systems characteristic of contemporary capitalism…. [T]here seems to have emerged a hegemonic cultural-institutional model concerning the 'right way of doing things'. This right way of doing things purports to be divorced from actual substance as it deals with models, blueprints, protocols, procedures, key words, and key concepts alone. (Shamir, 2005)

CSR is not a substitute for domestic law and law enforcement, nor is it a substitute for collective bargaining. TL made primarily by private actors and traditional state law are as likely to diverge as converge.

The Domestic Face of Globalization and the Globalizing State

Domestic legal systems are not just affected by global processes; they are embedded within them, and integral to them. This fact merits far more attention than it is generally accorded by globalization scholarship; however, it is crucial to understanding the relationship of globalization and law. By looking at globalization's domestic side, we can appreciate both the effects of globalization on domestic institutions as well as the potential effects of domestic institutions, especially legal institutions, on globalization. The domestic side of globalization may be counterintuitive for those who imagine globalization primarily in terms of a free-flowing private sector. However, the interplay between the globalization of capital, TL, and domestic legal processes is central to our topic.

My term for this complex and dynamic relationship between globalization and domestic law is the 'globalizing state' (Aman, 1998). The globalizing state is characterized by a fusion of public and private values, rhetorics, and regulatory practices (including privatization and deregulation). The fusion is itself integral to the interdependence of global and local economies. Privatization is both the engine and the outcome of that interdependence, increasing the exposure of the state to economic and political pressures that in turn tend to accelerate globalization, since private actors fully exposed to the global economy are now charged with carrying out delegated state tasks.

The global political economy places great pressures on all entities, public and private, to be cost effective. In the United States, this has further encouraged delegations of power to the private sector. Given the role that the public/private distinction plays in US administrative law, the globalizing state is one in which privatization tends to correspond to a democracy deficit (Aman, 2004). In lieu of public debate within the institutions of government, privatization favors other arrangements less likely to be transparent and accountable to the public, and less exposed to competing noneconomic value regimes. At the same time, the pragmatics of globalization make privatization the critical terrain in which a new public domestic law might respond to the democracy deficit.

Privatization in the public sector should be understood as a principal effect of globalization as it has been experienced in the United States. Privatization is not merely one means among others for making government more efficient or for expanding the private sector. Nor is it just an expression of current political trends as the regulatory pendulum swings from liberal to conservative. Rather, the increasing reliance on what is commonly referred to as 'the new governance' (Salamon, 2002) is indicative of recent transformations in the relationship between the market and the state itself – transformations that are inseparable from global economic competition and other forms of interdependence as these have developed in recent decades (see Stewart, 1993).

Marketizing Social Services

Outside of the legal spheres referenced above (treaties and corporate codes), the primary area in which law is involved in globalization is in the administrative transfer of state functions to markets. 'Privatization,' 'downsizing,' 'outsourcing,' 'contracting out,' and other forms of marketization are properly considered among the dynamics of globalization, once these are understood in context. The Reagan-era deregulatory reforms that set in motion the pervasive marketization of social services and other government functions in the United States were undertaken to improve US economic competitiveness at home and abroad. The pervasive neoliberalization of certain state functions in the United States and elsewhere in recent decades is inseparable from the development of globalization. National states – like corporations – 'function' in relation to maximizing their competitive advantage globally. This makes for what might seem like 'strange bedfellows' between the domestic and the foreign. In this section, I turn to the marketization of social services to emphasize these little-recognized but important connections.

Marketizing social services became politically popular as electorates became newly aware of global economic competition in the 1980s and 1990s, resulting in major changes in both the idea of 'the public' and the way the public sector functions in American life. The main regulatory reforms in ascendancy today consist of various forms of deregulation, privatization, outsourcing, and public/private partnerships – with the private sector often performing tasks once reserved primarily for government (Kamerman and Kahn, 1989). Indeed, the use of the private sector to deliver what once were governmentally provided services is a common form of privatization. Private contractors are pervasively involved in government functions, from military support to social services and myriad other roles. The primary governance tool in these cases is the contract. In the United States, for example, the management of prisons has been increasingly outsourced to the private sector at both the federal and state levels. Private providers also now commonly handle city services such as garbage and snow removal, and private entities in some states carry out various aspects of welfare administration, such as eligibility determinations (Aman, 2013, 2001b; Michaels, 2004; Bezdek, 2001).

Some forms of privatization result from administrative action; others derive from legislation. The legislature may, for example, sell off a governmentally owned entity to private parties, as was common in Europe in the 1980s. Like the deregulation that results from the wholesale statutory repeal of

a regulatory regime, the market in these instances is intended to replace the government completely when government-owned assets are sold to private buyers. Whether privatization takes place administratively or via statute, it is the context of these reforms that makes such domestic legal activity integral to globalization – just as in different but related ways corporate legal activity may make firms 'global.'

Understanding the role of domestic law in the current global environment enlarges the scope for comparison at the same time that it requires us to take into account the specifics of localized differences, as these are also meaningful. Globalization is deeply embedded within domestic institutions – both public and private – as ideas, relations, interests, and effects. This situation poses both opportunities and constraints on the relevance of law and the potential for law reform, especially as affecting vulnerable populations (Aman, 2009, 2013). Globalization is not just a top-down affair.

In the United States, recent developments in the social service sector illustrate the integral relationship of domestic law and globalization. These include welfare reform (disencumbering the federal and state treasuries), prisons (privatizing in the interest of efficiency), and immigration (setting checks on access to labor markets and social services). These rationales are, of course, highly politicized, but this does not negate their relevance as evidence of a relationship between domestic law (in these examples, Congressional legislation) and globalization. These and other areas of market-driven reforms reflect a particular conception of globalization as global economic competition (Aman, 2004). Along with other aspects of security more obviously linked to the global (e.g., domestic counterterrorism measures), privatization of social services is a primary zone where we can expect to find active channels between global and domestic legal activity and their associated purposes.

For example, welfare eligibility in many states increasingly depends upon judgments rendered by multinational companies whose mandate is to trim the rolls by depersonalizing eligibility and claims processes. Prisoners held in privately managed facilities are subject to companies' business interests and their need to satisfy their shareholders (Aman, 2005; Dolovich, 2009). Some states are eager to compete in the global economy by attracting the new investment and jobs associated with private prisons to their jurisdictions (Aman, 2013). A domestic perspective on globalization draws attention to the fact that the firms to which this work is outsourced in the welfare and prison contexts are MNCs that operate around the world (Bezdek, 2001; Fulcher, 2012; Raher, 2010; White, 2001). The expertise that these firms develop internationally in such public–private management arrangements can accelerate pressures for privatization at domestic and international levels. Immigration – conventionally debated as a border issue – is also a domestic 'front' of globalization. Immigration is in some ways a subset of the broader debate over welfare, as some of the problems facing immigrants are also the problems of the working poor and other populations at the social margins who are particularly reliant on social services (Aman, 2008, 2013; Aman and Rehrig, 2011; Gustafson, 2011; Michaels, 2004). But it is not just the role of multinational firms that make these sectors 'global'; rather, the pervasive formulation of the relationship between service providers and consumers in terms of economic efficiency renders state–individual relations generally as fundamentally economic in character.

The new relationships between private and public actors created by the marketization of welfare, borders, and prisons teach us important lessons about the domestic fronts of globalization and – in that light – we may see the potential for new kinds of law reform that can correct the limitations of a binary conception of law and markets referenced at the outset. Looking ahead, a role for law might be in the development of infrastructures for reforms – for example, contributing to the development of new forums for deliberation and stakeholder participation in relation to decisions regarding whether and how to outsource.

Conclusion

The changes in the world's political economy generally known under the rubric of globalization are fundamentally and simultaneously both structural and political in nature (Aman, 2004). Understanding law's role highlights the diversity of actors and interests involved, as well as the prominence of innovation, improvisation, and a host of ad hoc arrangements in TL – within states and beyond them. TL also highlights the broadening of economic and political networks across institutional and community lines, the technology-driven capacity for rapid change and response to change, and the growing powers of nonstate actors who operate simultaneously in multiple jurisdictions within and across national boundaries, and between public and private sectors. By associating competitiveness with national security, the rhetoric of global competition may sustain the impression that national states can be understood apart from globalization; however, attention to law's nature, sources, powers, and limits in relation to globalization compels a reassessment of such claims. The benefits of such reassessment include a more nuanced understanding of the heterogeneity within globalization, as well as a clearer appreciation of both the specificity of 'global' effects and the global significance of localized movements of protest or reform.

The three contexts considered in this article show some of the many complex ways that TL is both a source and effect of globalization. They represent three perspectives, seen from starting points in state lawmaking, private enterprise, and social security. However, these standpoints are not fixed; each is necessary to understanding the others. For example, the treaty process – though state driven and culminating in a multilateral treaty with global coverage – could not have developed but for the politics developed at the domestic level and domestic laws that made global negotiations feasible. Moreover, private players also played a crucial role in the process, incentivized by the prospect of new markets for new environmentally friendly substitute products. Public and private actors were able to utilize international and domestic law along with markets to reach a successful conclusion.

The importance of domestic law as a crucial component of globalization cannot be overestimated, especially when states privatize administration and services for reasons of economic competitiveness. It may seem obvious that they should do so;

however, that this should be a common sense notion is itself indication of the extent to which global competition has been normalized as the presumed context of public administration. Privatization of prisons and the marginalization of immigrants are integral to the global context in which they occur.

Law's adversarial process and its essential qualities of collaborative contestation remind us of law's potential relevance in the future development of globalization, as well as in relation to current models of globalization beyond the neoliberal one that predominates today. Law has an important role to play in supporting novel stakeholder partnerships and protecting the floor under social security – perhaps something of the reembedding of the market in society envisioned by Polanyi (1944). Critical reactions to the current politics of austerity around the world draw attention to the legal forums already in existence as arenas for experiments with pragmatic and flexible approaches to balancing efficiency and fairness (de Sousa Santos, 2007, 2008; de Sousa Santos and Rodriguez-Garavito, 2005; Evans, 2008; Rodriguez-Garavito, 2011). Law provides mechanisms and resources for the negotiation of public and private interests in the political arena – a political arena no longer contained within national states or political institutions in the narrow sense (Aman, 2004). Maximizing those opportunities will require yet new ways of thinking about law and its relationship to globalization.

See also: Capitalism: Global; Global Governance; Globalization and World Culture; Globalization, Anthropology of; Globalization: Geographical Aspects; International Business; International Law and Treaties; Internationalization: Patterns in Business Firms; Multinational Corporations.

Bibliography

Abbott, K., Snidal, D., 2009. The governance triangle: regulatory standards institutions and the shadow of the state. In: Mattli, W., Woods, N. (Eds.), The Politics of Global Regulation. Princeton University Press, Princeton, pp. 44–88.

Akech, M., 2009. Privatization & Democracy in East Africa: The Promise of Administrative Law. East African Educational Publishers, Nairobi.

Aman, A., 1992. Administrative Law in a Global Era. Cornell University Press, Ithaca, NY, 135–137.

Aman, A., 1998. The globalizing state: a future oriented perspective on the public/private distinction, federalism, and democracy. Vanderbilt Journal of Transnational Law 31, 769.

Aman, A., 2001a. Privatization and the democracy problem in globalization: making markets more accountable through administrative law. Fordham Urban Law Journal 28, 1477, 1481, 1495–1498, 1505–1506.

Aman, A., 2001b. The limits of globalization and the future of administrative law: from government to governance. Indiana Journal of Global Legal Studies 8, 379.

Aman, A., 2004. The Democracy Deficit: Taming Globalization through Law Reform. New York University Press, New York, pp. 7, 90–91, 87–127.

Aman, A., 2005. Privatization, prisons, democracy, and human rights: the need to extend the province of administrative law. Indiana Journal of Global Legal Studies 12, 511, 518–519, 529–532, 541, 544, 548–550.

Aman, A., 2006. Law, markets, and democracy: a role for law in the neo-liberal state. New York Law School Law Review 51, 802, 807–808.

Aman, A., 2008. Globalization from the ground up: a domestic perspective. In: Bertho, M. (Ed.), The Impact of Globalization on the United States: Volume 2, Law and Governance. Praeger Publishers, Westport, CT.

Aman, A., 2009. Privatization and democracy: resources in administrative law. In: Freeman, J., Minow, M. (Eds.), Government by Contract: Outsourcing and American Democracy. Harvard University Press, Cambridge, pp. 263–264.

Aman, A., 2013. Globalization and the privatization of welfare administration in Indiana. Indiana Journal of Global Legal Studies 20, 377, 378–390.

Aman, A., Rehrig, G., 2011. The Domestic Face of Globalization: Law's Role in the Integration of Immigrants in the United States. Legal Studies Research Paper #196, Indiana University Maurer School of Law. Indiana University, Bloomington.

Appadurai, A., 1990. Disjuncture and difference in the global cultural economy. Theory, Culture & Society 7, 295.

Backer, L.C., 2011. Private actors and public governance beyond the state: the multinational corporation, the Financial Stability Board, and the global governance order. Indiana Journal of Global Legal Studies 18, 751, 756–759.

Benedick, R., 1998. Ozone Diplomacy: New Directions in Safeguarding the Planet, second ed. Harvard University Press, Cambridge, MA, pp. 9–11, 24, 26, 28–34, 44–45, 47, 115, 118, 138, 204, 214.

Bezdek, B., 2001. Contractual welfare: non-accountability and diminished democracy in local government contracts for welfare-to-work services. Fordham Urban Law Journal 28, 1559, 1564.

Blackett, A., 2001. Global governance, legal pluralism and the decentered state: a labor law critique of codes of corporate conduct. Indiana Journal of Global Legal Studies 8, 401, 413–415.

Calliess, G., Zumbansen, P., 2010. Rough Consensus and Running Code: A Theory of Transnational Private Law. Hart Publishing, Oxford.

Cerny, P., 1990. The Changing Architecture of Politics: Structure, Agency, and the Future of the State. Sage, Newbury Park, CA.

Cole, D., 2011. From global to polycentric climate governance. Climate Law 2, 395–397.

Compa, L., 2008. Corporate social responsibility and workers' rights. Comparative Labor Law & Policy Journal 30 (1), 2–3, 5.

Compa, L., Hinchliffe-Darricarrère, T., 1995. Enforcing international labor rights through corporate codes of conduct. Columbia Journal of Transnational Law 33, 663.

Cutler, A., 1999. Locating 'authority' in the global political economy. International Studies Quarterly 43, 59.

de Feyter, K., Gomez Isa, F. (Eds.), 2005. Privatisation and Human Rights in the Age of Globalization. Oxford and Antwerp, Intersentia.

de Sousa Santos, B., 2006. Globalizations. Theory Culture Society 23, 393. http://dx.doi.org/10.1177/0263276406023002 68.

de Sousa Santos, B., 2007. Beyond abyssal thinking: from global lines to ecologies of knowledge. Eurozine. Electronic document: http://www.eurozine.com/articles/2007-06-29-santos-en.html (accessed 30.10.13.).

de Sousa Santos, B., 2008. The world social forum and the global left. Politics and Society 36, 247–270.

de Sousa Santos, B., Rodriguez-Garavito, C. (Eds.), 2005. Law and Globalization from below: Towards a Cosmopolitan Legality. Cambridge University Press, New York.

Delbrück, J., 1993. Globalization of law, politics, and markets—implications for domestic law—a European perspective. Indiana Journal of Global Legal Studies 1, 9–10.

Delbrück, J., 2002. Prospects for a "world (internal) law"?: legal development in a changing international system. Indiana Journal of Global Legal Studies 9, 401, 408–410.

Dolovich, S., 2009. How privatization thinks. In: Freeman, J., Minow, M. (Eds.), Government by Contract: Outsourcing and American Democracy. Harvard University Press, Cambridge, pp. 128–147.

Dunning, J., Lundan, S., 2008. Multinational Enterprises and the Global Economy, second ed. Edward Elgar, Cheltenham, pp. 213–232.

Evans, P., 2008. Is an alternative globalization possible? Politics and Society 36, 271–305.

Fransen, L., Kolk, A., 2007. Global rule-setting for business: a critical analysis of multi-stakeholder standards. Organization 14, 667, 672–678.

Freeman, J., Minow, M. (Eds.), 2009. Government by Contract: Outsourcing and American Democracy. Harvard University Press, Cambridge.

Fulcher, P., 2012. Hustle and flow: prison privatization fueling the prison industrial complex. Washburn Law Journal 51, 589, 601–603.

Galanter, M., 1981. Justice in many rooms: courts, private ordering, and indigenous law. Journal of Legal Pluralism 19 (1), 1–47.

Goldin, C., Katz, L., 2007. Long-run changes in the wage structure: narrowing, widening, polarizing. Brookings Papers on Economic Activity 38, 135.

Gustafson, K., 2011. Cheating Welfare: Public Assistance and the Criminalization of Poverty. New York University Press, New York.

Halliday, T., Carruthers, B., 2009. Bankrupt: Global Lawmaking and Systemic Financial Crisis. Stanford University Press, Stanford, CA, pp. 1–37.

Hallo de Wolf, A., 2011. Reconciling Privatization with Human Rights. Intersentia, Antwerpen, Belgium, pp. 1–17.

Harvey, D., 2005. A Brief History of Neoliberalism. Oxford University, New York, pp. 152–206.

Harvey, D., 2006. Spaces of Global Capitalism: Towards a Theory of Uneven Geographical Development. Verso, London.
Harvey, D., 2010. The Enigma of Capital and the Crises of Capitalism. Oxford University Press, New York, pp. 58–118.
International Labour Office, 2006. Tripartite Declaration of Principles Concerning Multinational Enterprises and Social Policy, fourth ed. International Labour Organization, Geneva.
International Labour Organization, June 17, 1999. Worst Forms of Child Labour Convention (No. 182), 38 I.L.M. 1207.
International Monetary Fund, November 14, 2012. Report: The Liberalization and Management of Capital Flows: An Institutional View. Electronic document: http://www.imf.org/external/np/pp/eng/2012/111412.pdf (accessed 26.10.13.).
Jessup, P., 1956. Transnational Law. Yale University Press, New Haven.
Kahler, M., Lake, D., 2003. Governance in a Global Economy: Political Authority in Transition. Princeton University Press, Princeton.
Kamerman, S., Kahn, A. (Eds.), 1989. Privatization and the Welfare State. Princeton University Press, Princeton, NJ, pp. 3–11, 15–45.
Keller, H., 2008. Corporate Codes of Conduct and Their Implementation: The Question of Legitimacy, pp. 6, 14–15. http://www.yale.edu/macmillan/Heken_Keller_Paper.pdf (accessed 16.08.13.).
Kingsbury, B., Krisch, N., Stewart, R., 2005. The emergence of global administrative law. Law and Contemporary Problems 68, 15.
Klein, M.W., 2012. Capital Controls: Gates versus Walls. National Bureau of Economic Research Working Paper 18526. Electronic document: http://www.nber.org/papers/w18526 (accessed 26.10.13.).
Koh, H., 1994. Transnational legal process. Nebraska Law Review 75, 181.
Levi Strauss & Co., 1992. Business Partner Terms of Engagement and Guidelines for Country Selection.
Mattli, W., Buthe, T., 2003. Setting international standards: technological rationality or primacy of power? World Politics 56, 1.
Mattli, W., Woods, N., 2009. The politics of global regulation. In: Mattli, W., Woods, N. (Eds.), The Politics of Global Regulation. Princeton University Press, Princeton.
Michaels, J., 2004. Deforming welfare: how the dominant narratives of devolution and privatization subverted federal welfare reform. Seton Hall Law Review 34, 573, 625–626, 666.
Montreal Protocol on Substances That Deplete the Ozone Layer, 1987. 1522 UNTS 3, 26 ILM 1550.
Mosk, C., 2005. Trade and Migration in the Modern World. Routledge, New York, pp. 34–36.
Myint, T., 2012. Governing International Rivers: Polycentric Politics in the Mekong and the Rhine. Edward Elgar Publishing, Northampton, MA.
Ochoa, C., 2011. Corporate social responsibility and firm compliance: lessons from the international law-international relations discourse. Santa Clara Journal of International Law 9 (169).
Organisation for Economic Co-operation and Development, 2008. OECD Guidelines for Multinational Enterprises. OECD, Paris.
Ostrom, E., 2010. Polycentric systems for coping with collective action and global environmental change. Global Environmental Change 20, 550–551.
Paavola, J., 2011. Climate change: the ultimate 'tragedy of the commons'? In: Cole, D., Ostrom, E. (Eds.), Property in Land and Other Resources. Lincoln Institute of Land Policy, Cambridge, MA, pp. 417–418.
Perez-Lopez, J., 1993. Promoting international respect for worker rights through business codes of conduct. Fordham International Law Journal 17 (1), 5–9, 23–26, 27–29.
Polanyi, K., 1944. The Great Transformation. Rinehart, New York.
Prakash, A., Hart, J. (Eds.), 2000. Coping with Globalization. Routledge, New York.
Raher, S., 2010. The business of punishing: impediments to accountability in the private corrections industry. Richmond Journal of Law and Public Interest 13, 209, 214.
Rodriguez-Garavito, C., 2011. Ethnicity.gov.: global governance, indigenous peoples, and the right to prior consultation in social minefields. Indiana Journal of Global Legal Studies 18, 263.
Roe v. Bridgestone, 492 F. Supp. 2d 988 (S.D. Ind. 2007).
Salamon, L. (Ed.), 2002. The Tools of Government: A Guide to the New Governance. Oxford, New York.
Sassen, S., 1998. Globalization and Its Discontents: Essays on the New Mobility of People and Money. The New Press, New York.
Sassen, S., 2003. Globalization or denationalization? Review of International Policial Economy 10 (1), 1–22.
Sassen, S., 2006. Territory, Authority, Rights. Princeton University Press, Princeton and Oxford, pp. 222–271.
Scott, C., 2010. Regulatory Governance and the Challenge of Constitutionalism. Private Regulation Series-02. European University Institute, Robert Schuman Centre for Advanced Studies, Florence.
Shaffer, G., 2013. Transnational Legal Ordering and State Change. Cambridge University Press, New York, pp. 26–28, 34.
Shamir, R., 2005. Corporate Social Responsibility: A Case of Hegemony and Counter-hegemony. Law and Globalization Below towards Cosmopolitan Legality. Cambridge University Press, New York, p. 113.
Stewart, R., 1993. Environmental regulation and international competitiveness. Yale Law Journal 102, 2039.
Strange, S., 1997. Casino Capitalism, second ed. Manchester University Press, New York.
Sullivan, L., 1983. Agents for change: the mobilization of multinational companies in South Africa. Law and Policy in International Business 15, 427–438.
Sum, N., Ngai, P., 2005. Globalization and paradoxes of ethical transnational production: code of conduct in a Chinese workplace. Competition and Change 9, 181.
Sunstein, C., 2006. Montreal vs. Kyoto: a tale of two protocols. In: John, M. Olin (Ed.), Law & Economics Working Paper No. 302. University of Chicago, Chicago.
Swoboda, F., 1992. Levi Strauss to drop suppliers violating its worker rights. The Washington Post, D1–D2.
Teubner, G., 1987. Juridification: concepts, aspects, limits, solutions. In: Teubner, G. (Ed.), Juridification of Social Spheres. Walter de Gruyter & Co, Berlin, pp. 3–48.
Teubner, G., 1996. Global Bukowina: legal pluralism in the world-society. In: Teubner, G. (Ed.), Global Law without a State. Dartsmouth, Hanover, NH, pp. 3–28.
Toxic Substances Control Act, 1976. 15 U.S.C. 2601–2671.
Valverde, M., 2009. Jurisdiction and scale: legal 'Technicalities' as resources for theory. Social and Legal Studies 18, 139, 144.
White, A., 2001. Rule of law and the limits of sovereignty: the private prison in jurisprudential perspective. American Criminal Law Review 38, 111, 134.
Zumbansen, P., 2006. Transnational law. In: Smits, J. (Ed.), Encyclopedia of Comparative Law. Edward Elgar Publishing, Northampton, MA, pp. 740–743.

Globalization: Welfare and Social Work

Karen Lyons, London Metropolitan University, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Globalization is not a new phenomenon, but it has received relatively little attention in relation to welfare systems and social work. This article describes the understandings of this and related terms, with particular reference to economic globalization. It identifies some contemporary characteristics and trends and discusses how these affect welfare systems; and considers why and how globalization is relevant to social work and social development.

Introduction

The notion of 'globalization' entered public discourse in the 1990s and has since been the subject of much debate across a range of disciplines, as well as in the media. The characteristics – and risks and benefits – of this process have been elaborated and contested; and related terms have been articulated. Advocates have stressed the importance of economic globalization for human and societal development while social movements have developed to counteract the costs to particular communities or the whole planet.

Central features of globalization are international trade and the spread of people and knowledge across national and continental borders. The current phase of globalization is not unique, and antecedents can be found over two millennia or more but it was extended and deepened in the twentieth century by increasingly rapid and cheap transport and then advances in information and communications technology (ICT). Transnational corporations (TNCs) are the agents of globalization and, in addition to those already existing in Europe and the United States, new ones have been established in the BRIC countries (Brazil, Russia, India, and China) and other emerging market economies (EMEs).

Another essential feature of current globalization is the extent to which national governments have adopted neoliberal policies favorable to free market enterprise and global competition. Policies which favor reduction in the role of the state across a range of activities and which promote privatization or outsourcing of goods and services impact on domestic circumstances, including in relation to the extent and form of welfare systems. In the postfinancial crises (originated in the United States in 2008), many European governments have adopted austerity measures affecting large proportions of their populations, both those who have traditionally been the focus of social work concern and also those who were previously self-supporting.

Conversely, the relocation of jobs to low labor cost countries has brought employment to large numbers of people, e.g., in India and the EMEs, but has so far failed to benefit many economies in the global south, particularly those countries deemed to be fragile states. The espoused benefits of economic globalization have not been equally shared and there is evidence of an increasing gap between the 'haves' and 'have-nots,' with implications for welfare policies and provisions; for social development programs; and for social work in its varied forms.

This article explores current definitions of globalization and related terms. The major actors in (economic) globalization are identified and other aspects of globalization are considered. Some aspects of the relationship between globalization and 'welfare' are presented before examining some of the ways in which globalization has relevance for social work and social development. Alongside the international processes of globalization, various forms of regionalism have also been established in or between different continents: these are mainly economic arrangements (with a wider remit and influence in the case of the European Union) but only passing reference to regionalization can be made here.

Characteristics and Meanings of Globalization and Related Terms

Globalization has been examined from a variety of disciplinary perspectives, with different definitions being advanced for this and related terms. However, economic integration is usually taken to be the dominant form of globalization affecting other processes, including social and cultural. Das (2011: p. 18) states that while "… globality, globalism and globalization … are often narrowly defined in their economic meaning, they are broad in their meaning and implications." Economic globalization has a major impact on other processes and systems, although not all countries are integrated into the world system of interrelated economies and finance, and the form that globalization takes varies over time (Wallerstein, 1974).

The origins of 'global systems' can be traced to the post Second World War period when various institutions were established to promote world peace, economic reconstruction, and social development (e.g., the United Nations (UN) and its specialist agencies; the World Bank; and the International Monetary Fund (IMF)) (see Wilson, 2012; Baker and Hinds, 2012). However, in 1974 Wallerstein used the specific term, globalization, to describe "the expansion of capitalism to the point where a world economy was operational, thus creating a world system" (Lyons et al., 2006: p. 22). The term became widely used in public discourse and social science literature in the 1990s and writers in different disciplines recognized the extent to which various processes are interrelated and interact on each other. For example, Axford (1995: p. 5) linked economic and political processes evident in globalization to culture change. However, Khan and Dominelli (2000)

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28036-7

described economic activity as the 'lynchpin' of globalization (p. 96) and identified the following key features:

- mobility of production processes from advanced capitalist countries to developing countries;
- the rising power and spread of TNCs worldwide;
- the liberalization of international finance systems; and
- reduction in transport and communication costs.

Das (2011: p. 18) describes "globalism" as "a network of connections spanning multi-continental distances, drawing them close together economically, socially, culturally and informationally." Globalization, in turn, is the process "promoting and intensifying multi-continental interconnectedness, and thereby increasing the degree of globalism" (Das, 2011: p. 18). As others have also observed, globalization overrides national boundaries and integrates societies and economies, leading to discussions about the role and power of the nation states (Held et al., 1999). As national power has apparently declined, due to both global and also regional processes (e.g., in Europe), the influence of TNCs has increased. They have played a major role in increasing the "international division of labor on the one hand and growing integration of national economies through trade in goods and services, cross-border corporate investment, and capital flows on the other" (Das, 2011: p. 18). While the power of individual states over national economies and the activities of TNCs and financial institutions has decreased, the role of the state has changed with an increase in powers related to regulation and surveillance. Such powers are evident in the field of migration and the limits placed on gaining citizenship in another country, as well as in relation to collective activities (e.g., trade unions) and the public sector. Such changes impact most on people who are least well off in specific societies and internationally, compounding the growing gap between corporate elites and others in national and global populations (Wichterich, 2000).

Alongside continuing questions about the power of nation states, the 1990s also saw expressions of concern about the possible decrease or even loss of national identity and culture, seeing globalization as responsible for an increasing homogenization of culture. For example, criticisms were voiced about the extent to which the rationalization of processes and the spread of uniform products of (what were then termed) 'multinational corporations' (MNCs, often US based) permeated countries in (economic) transition, struggling to (re)establish national identity and policies. Ritzer (1993: p. 99) suggested that the processes of globalization were leading to a situation where "people know what to expect at all times and in all places … (in) … a world that offers no surprises."

In the context of concerns in the 1990s about the social and cultural impact of globalization. The term, 'glocal,' was used to describe the relationship between the local and the global, with the concept of glocalization being developed by Bauman (1998). The terms have utility in a number of areas, including in indicating the transcultural relationships, which increasingly exist, not least in the personal/domestic arena through migration. While the terms themselves may not be in evidence, the balance of power between the global and the local has been further considered in this century. For instance, Gray and Fook (2004) "identified four sets of related tensions … globalization–localization; Westernization–indigenization; multi-culturalism–universalization; and universal–local standards" (cited in Healy, 2012: p. 13) and indigenization and globalization have been discussed together by Dominelli initially in 1997 as well as more recently (e.g., 2012).

As mentioned, the contemporary phase of globalization has been significantly enhanced by technological innovations, with ICT facilitating not only the operation of TNCs but also enabling an increase in information sharing and communication in a wider range of activities. These include not only an increase in international criminal activities (e.g., trade in arms, drugs, and people), but also the ability of migrants to retain connections with family and communities 'back home' and to forge transnational lifestyles (e.g., Negi and Furman, 2010). It has also underpinned the growth of critical responses to some of the consequences of globalization and the growth of global social movements (e.g., Dominelli, 2012) and the beginnings of a global civil society.

However, technological advances alone are not a sufficient condition for continuing globalization, which also requires the lowering and elimination of trade barriers; the deregulation of capital markets and domestic economies; the maintenance of balanced (national) budgets; and the opening up of banking and telecommunications sectors to private ownership and competition (Das, 2011). Many of these conditions were being met and advanced in the first decade of the twenty-first century although, in the aftermath of the economic crises in 2008, some are questioning the extent to which deregulation, for example, of financial institutions and capital markets, should be maintained. But it has two further conditions, privatization and the shrinking of government bureaucracies, which have had the most impact in countries where governments were heavily involved in welfare provision; and the changing economic conditions and political shifts have also had a bearing on social development initiatives in less-advanced economies (Midgley, 2012).

Globalization, Welfare, and Development

Measures to promote well-being and to support the development of various welfare measures require economic resources, but they also require particular policy directions and frameworks. A US social work academic described welfare as "the state of collective well-being of a society or community" (Van Wormer, 1997: p. 4). The term is variously used in conjunction with other words giving it more specific connotations. So, for instance, the term 'welfare state' suggests a range of provisions, including universal free education and health care as well as various income maintenance ('social security') measures and personal social/human services. Such systems were established in the second half of the twentieth century in the context of social democratic political systems then prevailing, e.g., in North West European and other countries (Gilbert, 2002). In other parts of the world, 'welfare' may have a different meaning, related to poverty relief (through state, voluntary, and private means) but income maintenance schemes are at very basic levels and/or limited in their reach (Midgley, 1997).

In addition, the lack of accountability of TNCs to national governments or international bodies allows free movement of capital and production to sites that benefit the corporations'

shareholders, with little regard for the social consequences in particular locations, notably, high levels of unemployment in areas/countries which firms have 'vacated' and, possibly, poor working conditions in developing countries (Akimoto and Sungkawan, 2012). The worst excesses of labor exploitation may be curtailed by international bodies (e.g., the International Labor Organization, ILO) or by public outcry, and some TNCs pride themselves on being good employers (e.g., providing 'decent work' and contributing to community resources), but problems related to unemployment and development of comprehensive welfare systems or even basic safety nets are not generally seen as the responsibility of commercial enterprises. However, transnational companies are increasingly involved in the provision of health and social care facilities, as well as the insurance schemes, which sometimes underpin access to these resources and pensions.

Teeple (2000) has argued that globalization marks the ascendency of economics over politics, of corporate demands over public policy, and of private interests over public ones; and that the rise of neoliberalism has undermined and reversed welfare state initiatives. New Zealand led the way in the 'dismantling of the welfare state' in the early 1990s although there had been anxieties about 'the crisis in welfare' in Europe and the United States from the 1980s (e.g., Pierson, 1996). By the twenty-first century, radical changes had taken place in many countries in the organization and provision of pension systems, health care, education, personal social services/human services, housing policies, and criminal justice systems. These systems all face further changes – downsizing of the public sector, outsourcing and increased emphasis on private sector provisions – in the light of 'austerity measures' being pursued by national governments in the wake of the 2008 economic crises (e.g., Hessle, 2012).

In some countries, moves to reduce the role of the state in welfare have been justified on the grounds that generous welfare provisions promote 'welfare dependency' and restrict choice and efficiency (Gilbert, 2002). Alongside efforts to reduce the welfare bill, countries have put considerable emphasis on employment as the main means of income and well-being for individuals and families, e.g., 'welfare to work' programs in the United States and social inclusion policies in the European Union (Lyons et al., 2006). However, the economic downturn and the shift to knowledge economies among postindustrial societies limit the extent to which jobs are available, particularly for less-skilled workers. Even the growth of the service sector (relative to manufacturing) offers diminishing opportunities for employment, for example, as high street retail stores go into liquidation in the face of competition from online purchasing.

These trends can be seen as an inevitable rebalancing of economies and employment opportunities at national and global levels, attributable to technological change and other global processes, but they have significant implications for the demands made on welfare and social development budgets, particularly as they occur in the context of demographic changes. Advances in health care and improved material benefits in the developed world have underpinned increased longevity and significant expansion in the proportion of older people (including frail elders) in many populations, including in Europe and North America. Conversely, improved health and education have reduced maternal and child mortality rates in the BRICs and emerging economies and also in some developing countries resulting in increased proportions of children and young people in their populations. The 'greying of the planet' places significant pressures on welfare provisions in the West, while a youthful population provides a ready supply of labor in many Asian and South American countries, but the lack of employment opportunities in many African countries has implications for development programs and affects migration patterns (e.g., Hokenstad and Roberts, 2012; Segal and Heck, 2012).

Cultural factors also affect the provision and take-up of such welfare provisions as exist or policies to increase welfare when up to 50% of national populations born female may be discriminated against, for example, in extreme cases being debarred from education or not having access to appropriate health services such as maternity care. The establishment of the UN's Millennium Goals has had some positive effects in increasing the take-up of education by girls in some countries and encouraging national governments to introduce policies and services which reduce maternal and child mortality rates, but it seems that in many places there has been little change in patriarchal societies where boy children are favored and girls are prevented from reaching their potential (Correll, 2008).

Finally, environmental factors, including climate change and extreme weather events, currently and potentially impact on the well-being of the worldwide population in one way or another (Alston and Besthorn, 2012). The concerted efforts of international bodies and national governments are needed to address these concerns, which also have implications for policy makers and practitioners in the welfare and development sectors. Such events have contributed to food insecurity (in the face of the expanding global population) exacerbated in some countries and regions by competition for water and fuel sources. In some cases, such competition contributes directly to interethnic and international conflicts. However, the latter also reflect significant economic inequalities and cultural differences, not least those related to religion, as exemplified in the Middle East (e.g., Al-Makhamreh and Libal, 2012).

Globalization, Social Work, and Social Development

References to globalization began to appear in the social work and social development literature from the 1990s. For example, in Europe, Trevillion (1997: pp. 7–8) suggested that social work was a product of the first wave of globalization (in this context, the late nineteenth/early twentieth century) and was now going through a transition to 'the new globalism of networks and markets,' while Midgley (1997) was writing about social development in the context of globalization. Globalization has been a theme addressed primarily in literature relating to international social work and Lyons (1999: p. 12) suggested that "internationalisation has been given an added impetus by the increased recognition of the relevance of globalisation to the welfare sector, and to the roles and tasks of social workers, whatever their arena."

'Social work' takes a range of forms nationally in the West, even in particular regions, e.g., Europe (Zavirsek and Lawrence, 2012), but is often restricted in its practices, relative to broader

understandings and practices in the social development field. Some writers have identified the need for a closer association between 'social work' and 'social development' as promoted, for instance by the UN and INGOs, in 'developing countries.' This would entail a more aware and integrated approach to the economic as well as the social aspects of human well-being and an appreciation that the flow of ideas and practices, generally from more-developed to less-developed countries, should be reversed so that 'western' social workers import and utilize some of the theories and practices common in the social development field. Midgley has long advocated this view and has suggested the term 'developmental social work' to describe the role of social investment in professional practice (Midgley and Conley, 2010). During the twenty-first century, the term 'international social work' has continued to be variously understood and the subject of a range of texts. Most recently, these include two handbooks (Healy and Link, 2012; Lyons et al., 2012), where globalization and its relationship to social work/development have been the focus of specific chapters as well as referred to elsewhere. Harris and Chou (2001: p. 161), acknowledging that the global economy would act as a major constraint on welfare development, utilized the term, 'glocalization,' for the processes whereby local policy and practices in relation to community care are impacted by global processes but mediated by national contexts and 'country specific institutional arrangements.' Dustin (2007), building on research in the community care field, illustrated the relationship between local and global trends when she drew on Ritzer's thesis to describe the commodification of care and effects of centralized managerial domination, evident in the growth of care management in the United Kingdom. Concerns have been expressed that the risks of educational colonization and professional imperialism are inherent in some forms of international social work and must be challenged (Payne and Askeland, 2008) while Dominelli has questioned whether there is such a thing, if social work is taken to be an inclusive entity in which each country has the right to specify its own modes of practice and traditions, even if borrowing from others (Dominelli, 2000).

A recognition of the influence and implications of globalization has been evident in the work of three major international professional organizations representing social work in its various forms, the International Association of Schools of Social Work (IASSW); the International Council on Social Welfare (ICSW); and the International Federation of Social Workers (IFSW). In 2000, general meetings of the IASSW and IFSW were held following a joint Congress, leading to an agreed international definition of social work (note, not international social work). Subsequently the two organizations collaborated in producing 'Global Standards for Social Work Education and Training' and 'Ethics in Social Work: statement of principles' accepted in 2004 (see IASSW or IFSW websites). These have raised questions about the universality of the assumptions, values, and practices on which these statements are based although it was not intended that they would be prescriptive or static (see, for example, Sewpaul and Jones, 2004).

Globalization has received increasing attention at international level in the social work/development fields, over the past two decades. For example, IFSW issued a policy statement on globalization in 1996 in response to a conference resolution expressing concern about the impact of neoliberal policies and structural adjustment programs. This document was revised and extended to include 'the Environment,' accepted by the membership in 2004. Most recently, the three international organizations – IASSW, ICSW, and IFSW, have collaborated to bring social work and social development into closer association at international level. In 2010, the ICSW issued three Briefing Papers (on The Impact of the Global Financial Crisis; Social Protection Floor Initiative; and Human Trafficking) and IFSW issued four Consultation Papers (including on Poverty Eradication and Aging and Older Adults) as part of a policy process aimed at establishing a Global Agenda, the main concerns of which were collectively agreed at a joint Conference in Hong Kong (2010) (see ICSW and IFSW websites). The revised version of the 'Global Agenda for Social Work and Social Development Commitment to Action' was presented to the joint social work and social development global conference in Sweden in 2012 (Global Agenda, 2012). The ICSW also issues a monthly 'Global Co-operation Newsletter,' updating readers on a variety of relevant developments (ICSW website).

Key and emerging themes in these documents and other literature have been increasing poverty levels and exclusion; migration; and environmental concerns. The first of these is discussed below as an example of issues with global dimensions, which are relevant to social and development workers.

Poverty and Social Exclusion

Concerns about poverty have been a central focus of social work since its inception (late nineteenth century) to address the negative impacts of industrialization and urbanization in Europe and North America. Similar processes have affected many other countries as they modernized during the twentieth century, while social development programs were established to address the needs of societies and communities lacking the infrastructure of industrial development. Poverty itself takes many forms and has been characterized as 'relative' in advanced industrial societies and absolute in the global south, where it can take extreme forms, when basic needs for survival cannot be met by individuals themselves (Desai and Solas, 2012).

Consequences of economic globalization have included the increase in the disparity in incomes between rich and poor people (both between different sectors of society within countries and between rich and poor nations) and the extent to which economic problems in one part of the world impact on populations elsewhere. It is ironic that, as the international community declared a Second United Nations Decade for the Eradication of Poverty (2008–2017), so the financial crises struck in Europe and the US, with ramifications for economies around the world. In both continents, previously self-supporting people lost their jobs and sometimes their homes and thus, new populations joined those groups traditionally identified as living in poverty and experiencing social exclusion. The latter term, deriving from French policy concerns (Lenoir, 1974), has been in evidence in both social work and social development literature from the 1990s. For example, it was introduced to European Union terminology from about 1990 and policies to promote 'social inclusion' are a core

feature of European regional policy as well as having spread (e.g., to Australia) (Lyons and Huegler, 2012).

Poverty and social exclusion reflect inequalities and power imbalances, evident in most societies: they are indicative of social injustices and violate the human rights of sections of the population. Local experiences and international perspectives have identified some people as being particularly vulnerable to chronic poverty or dramatic changes in national economic conditions and most likely to seek social work assistance. These include families where there is a lone parent or neither partner is in work; people with disabilities, or physical or mental health problems; older people; migrants, and children. Child poverty is a worldwide phenomenon, including in wealthy countries: it has been subject to various attempts to measure and address it over many decades but continues to be an enduring reflection of social inequalities (UNICEF, 2012).

Elder care is of particular concern to many western countries, as demand for residential and domiciliary services for older people increase, although physical infirmity may be less significant than various forms of dementia. The private sector has long played a major role in providing residential care homes for older people and increasingly, care home providers are transnational companies (Di Rosa et al., 2012). A growing trend is the recruitment of 'overseas workers' to staff such homes, either because of shortage of local labor or to reduce labor costs. For example, in Europe, staff are recruited from neighboring countries or further afield, with implications for both the residents and the workers themselves (e.g., Di Rosa et al., 2012). In addition, in the case of Europe, over the last two or three decades, a proportion of (healthy) older people from North European countries had sought cheaper living conditions in warmer climes, e.g., Spain, but recently anecdotal evidence suggests that they have had difficulties selling properties and face problems paying for the accommodation or care they need if returning to their country of origin. Most extremely, it has been reported that, while some older German people are living in care homes in Hungary, Greece, and even Thailand by choice, others are apparently being 'sent' to rehabilitation centers or long-term retirement homes outside Germany, e.g., in Slovakia. The reasons given are primarily economic (the rising costs of care homes relative to the level of pensions), but it is also suggested that older people's increased needs are met by a better standard of care abroad (Connolly, 2012: p. 21).

Concluding Comments

Globalization has been evident in various forms over many centuries, increasing in its reach and the extent to which markets and other aspects of human life are integrated to the current state of globalism. The economic aspects of globalization predominate over other aspects (including social and cultural) although these have been increasingly 'facilitated' by political processes and international agencies. Major changes are currently taking place in the balance of power between 'western' and 'eastern' economies. However, even before the twenty-first century financial crises in the United States and Europe, governments in many countries were keen to curb public expenditure and the role of the state in welfare in line with political ideology. Many governments have since adopted austerity measures affecting both those sectors of the population who have traditionally relied on welfare benefits and social work interventions and also new groups affected by the shifts in employment opportunities (from west to east) and the impact of ICT.

In addition to the increase in rapid and relatively cheap transport, the revolution in digital technology has played a significant role in the expansion of TNCs but has also been utilized in other sectors, ranging from the spread of international criminal organizations to the maintenance of transnational family networks. ICT has also facilitated the establishment of social movements (including some in opposition to capitalism) and plays a significant role in the spread of information and international conventions (promoted by bodies associated with the UN) and in the exchange of ideas and practices in welfare and professional fields.

An awareness of the impact and implications of globalization on welfare, social, and development work has been in evidence since the 1990s and further attention has been given to the topic in the context of increased literature about international social work (broadly conceived). Global demographic trends and migration have particular implications for varied forms of social work organization and practices, worldwide, while social development initiatives in less economically developed countries have also been impacted. Three international bodies representing social service organizations and professionals are increasingly engaging with globalization (not least as its processes affect the distribution of resources within and between nations) and with international bodies, which seek to address some of the trends and issues associated with globalism.

See also: Concept Learning and Representation: Models; Global Agenda for Social Work and Social Development; Globalization and Health; Globalization and World Culture; Globalization, Anthropology of; Globalization: Geographical Aspects; Green Social Work; Neoliberalism; Poverty in History; Poverty, Geography of; Poverty, Sociology of; Social Exclusion, Social Deprivation and Health; Social Work Theory; United Nations and its Agencies.

Bibliography

Akimoto, D., Sungkawan, D., 2012. Social work, economic conditions and livelihood. In: Lyons, K., Hokenstad, T., Pawar, M., Huegler, N., Hall, N. (Eds.), The SAGE Handbook of International Social Work. Sage, London.

Al-Makhamreh, S., Libal, K., 2012. The Middle East: expanding social work to address 21st century concerns. In: Lyons, K., Hokenstad, T., Pawar, M., Huegler, N., Hall, N. (Eds.), The SAGE Handbook of International Social Work. Sage, London.

Alston, M., Besthorn, F., 2012. Environment and sustainability. In: Lyons, K., Hokenstad, T., Pawar, M., Huegler, N., Hall, N. (Eds.), The SAGE Handbook of International Social Work. Sage, London.

Axford, B., 1995. The Global System: Economics, Politics and Culture. Policy Press, Cambridge.

Baker, P.-A., Hinds, K., 2012. Regulators of the global economy: the IMF, the World Bank and the WTO. In: Healy, L., Link, R. (Eds.), Handbook of International Social Work: Human Rights, Development and the Global Profession. Oxford University Press, New York.

Bauman, Z., 1998. Globalization: The Human Consequences. Polity Press, Cambridge.

Connolly, K., December 27, 2012. No place like home: Germany 'exporting' thousands of elderly to cheap foreign care. The Guardian, 21.
Correll, D., 2008. The politics of poverty and social development. International Social Work 51 (4), 453–466.
Das, D.K., 2011. Conceptual Globalism and Gloablization: An Initiation. Solbridge International School of Business, Woosong University, Korea. http://wrap.warwick.ac.uk/49071/1/WRAP_Das_27511.pdf (accessed 09.01.13.).
Desai, M., Solas, J., 2012. Poverty, development and social justice. In: Lyons, K., Hokenstad, T., Pawar, M., Huegler, N., Hall, N. (Eds.), The SAGE Handbook of International Social Work. Sage, London.
Di Rosa, M., Melchiorre, M., Lucchetti, M., Lamura, G., 2012. The impact of migrant work in the elder care sector: recent trends and empirical evidence in Italy. European Journal of Social Work 15 (1), 9–28.
Dominelli, L., 1997. Sociology for Social Work. Macmillan, London.
Dominelli, L., 2000. International comparisons in social work. In: Pearce, R., Weinstein, J. (Eds.), Innovative Education and Training for Care Professionals: A Providers' Guide. Jessica Kingsley, London, pp. 25–42.
Dominelli, L., 2012. Globalisation and indigenisation: reconciling the irreconcilable in social work? In: Lyons, K., Hokenstad, T., Pawar, M., Huegler, N., Hall, N. (Eds.), The SAGE Handbook of International Social Work. Sage, London.
Dustin, D., 2007. The McDonaldization of Social Work. Ashgate, Aldershot.
Gilbert, N., 2002. Transformation of the Welfare State: the silent surrender of Public Responsibility. Oxford University Press, Oxford.
Global Agenda, 2012. The Global Agenda for Social Work and Social Development Commitment to Action. http://cdn.ifsw.org/assets/globalagenda2012.pdf (accessed 21.01.13.).
Gray, M., Fook, J., 2004. The quest for a universal social work: some issues and implications. Social Work Education 23 (2), 625–644.
Harris, J., Chou, Y.C., 2001. Globalization or glocalization? community care in Taiwan and Britain. European Journal of Social Work 4 (2), 161–172.
Healy, L., 2012. Defining international social work. In: Healy, L., Link, R. (Eds.), Handbook of International Social Work, Human Rights, Development, and the Global Profession. Oxford University Press, Oxford and New York.
Healy, L., Link, R. (Eds.), 2012. Handbook of International Social Work, Human Rights, Development, and the Global Profession. Oxford University Press, Oxford and New York.
Held, D., McGrew, A., Goldblatt, D., Perraton, J., 1999. Politics, Economic and Culture. Polity Press, Cambridge.
Hessle, S., 2012. Variations and issues of survival of European welfare state models in the twenty first century. In: Healy, L., Link, R. (Eds.), Handbook of International Social Work: Human Rights, Development, and the Global Profession. Oxford University Press, Oxford and New York.
Hokenstad, M.C., Roberts, A.R., 2012. Older persons and social work: a global perspective. In: Lyons, K., Hokenstad, T., Pawar, M., Huegler, N., Hall, N. (Eds.), The SAGE Handbook of International Social Work. Sage, London.
Khan, P., Dominelli, L., 2000. The impact of globalization on social work practice in the UK. European Journal of Social Work 3 (2), 95–108.
Lenoir, R., 1974. Les exclus: Un francais sur dix. Seuil, Paris.
Lyons, K., 1999. International Social Work: Themes and Perspectives. Ashgate, Aldershot.
Lyons, K., Huegler, N., 2012. Social exclusion and inclusion. In: Healy, L., Link, R. (Eds.), Handbook of International Social Work, Human Rights, Development, and the Global Profession. Oxford University Press, Oxford and New York.
Lyons, K., Manion, K., Carlsen, M., 2006. International Perspectives on Social Work: Global Conditions and Local Practice. Palgrave Macmillan, Basingstoke.
Lyons, K., Hokenstad, T., Pawar, M., Huegler, N., Hall, N. (Eds.), 2012. The SAGE Handbook of International Social Work. Sage, London.
Midgley, J., 1997. Social Welfare in Global Context. Sage, Thousand Oaks, CA.
Midgley, J., 2012. Development. In: Healy, L., Link, R. (Eds.), Handbook of International Social Work, Human Rights, Development, and the Global Profession. Oxford University Press, Oxford and New York.
Midgley, J., Conley, A. (Eds.), 2010. Developmental Social Work. Oxford University Press, New York.
Negi, N.J., Furman, R. (Eds.), 2010. Transnational Social Work Practice. Columbia University Press, New York.
Payne, M., Askeland, G.A., 2008. Globaliztion and International Social Work: Postmodern Change and Challenge. Ashgate, Aldershot.
Pierson, P., 1996. The new politics of the welfare state. World Politics 48 (2), 143–179.
Ritzer, G., 1993. The McDonaldization of Society, first/third ed. Thousand Oaks, CA, Pine Forge.
Segal, U., Heck, G., 2012. Migration, minorities and citizenship. In: Lyons, K., Hokenstad, T., Pawar, M., Huegler, N., Hall, N. (Eds.), The SAGE Handbook of International Social Work. Sage, London.
Sewpaul, V., Jones, D., 2004. Global standards for social work education and training. Social Work Education 23 (5), 493–513.
Stiglitz, J.E., 2003. Globalization and Its Discontents. Norton, New York.
Teeple, G., 2000. Globalization and the Decline of Social Reform: Into the 21st Century, second ed. Garamond Press, Aurora, Ont.
Trevillion, S., 1997. 'The globalisation of European social work. Social Work in Europe 4 (1), 1–9.
UNICEF, 2012. Measuring Child Poverty: New League Tables of Child Poverty in the World's Rich Countries. Innocenti Research Centre Report Card 10. Available at: www.unicef-irc.org/publications/pdf/rc10_eng.pdf (accessed 28.11.12.).
Van Wormer, K., 1997. Social Welfare: A World View. Nelson Hall Publications, Chicago.
Wallerstein, I., 1974. The Modern World System. Academic Press, New York.
Wichterich, C., 2000. The Globalized Woman: Reports from a Future of Inequality. Zed Books, London.
Wilson, S., 2012. The United nations agencies. In: Healy, L., Link, R. (Eds.), Handbook of International Social Work, Human Rights, Development, and the Global Profession. Oxford University Press, Oxford and New York.
Zavirsek, D., Lawrence, S., 2012. Social work in Europe. In: Lyons, K., Hokenstad, T., Pawar, M., Huegler, N., Hall, N. (Eds.), The SAGE Handbook of International Social Work. Sage, London.

Relevant Websites

www.iassw-aiets.org – IASSW International Association of Schools of Social Work.
www.icsw.org – ICSW International Council on Social Welfare.
www.ifsw.org – IFSW International Federation of Social Workers.
www.ilo.org – ILO International Labour Organisation.
www.imf.org – IMF International Monetary Fund.
http://www.iss-ssi/2009/index.php?id=1 – ISS International Social Services.
www.oecd.org – OECD Organisation for Economic Co-operation and Development.
http://www.un.org./esa/population/ – UN POP UN Population Division.
http://www.undp.org/content/undp/en/home/ourwork/overview.html – UNDP United Nations Development Programme.
http://www.unhcr.org/cgi-bin/texis/vtx/home – UNHCR The UN Refugee Agency.
www.unicef.org – UNICEF United Nations Children's Fund.
www.who.int/ – WHO World Health Organisation.

Goal-Oriented Behaviour Among Human and Non-Human Primates

Heidi L Marsh, University of Western Ontario, London, ON, Canada
Maria Legerstee, York University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

We review the literature surrounding the phylogenetic and developmental emergence of goal-oriented behavior, among human and nonhuman primates. We define goal-oriented awareness as the ability to perceive goals and perceptions in others. We examine empirical literature involving gaze-following, shared attention, distinguishing between actions and intentions, and the ability to generate and understand communicative cues. We conclude that at least a rudimentary awareness of goal-oriented behavior is present from birth in humans, and in adult great apes. The evidence in other primate species shows phylogenetic differences as well as gaps in the empirical literature.

Social cognition includes any process that enables conspecifics to interact together effectively, including verbal and nonverbal signals. Although sociality in human adults is largely comprised of the former, the latter are informative because they exist in both preverbal children and nonhuman primates (hereafter, primates), and thus, represent the development of social cognition in its earliest forms. In this article, we examine one aspect of nonverbal social cognition – goal-oriented behavior – as it develops both ontogenetically and phylogenetically. We discuss the theoretical and methodological challenges inherent in the study of this behavior. Next, we trace its development in infants and primates, as seen in the ability to follow gazes and share attention, infer goals and distinguish intentions, and understand and use communicative cues. We conclude that both populations exhibit goal-oriented behavior, and further, that the evidence of such abilities so early in life (in the case of humans), and phylogenetically (in the case of primates) suggests a critical predisposition for more complex social abilities.

Defining Goal-Directed Behavior and Intentions

Predicting a social partner's goals is a critical, yet complex aspect of social cognition. Humans and other animals are similar to nonliving moving objects, in that they have physical characteristics (e.g., form, color, and size), but they are different in that living beings have inner lives that drive their behavior: They communicate, and have feelings and intentions, in ways that are often outwardly subtle (Gelman and Spelke, 1981). Moreover, unlike inanimate objects, social animals react to the physical and social environments in which they are embedded. This makes the behavior of social partners much less predictable than that of nonliving objects.

Because primates (human and nonhuman) have complex social lives, maintaining relationships demands flexible socio-cognitive abilities. Human adults are experts at predicting others' behavior because they react to overt behavioral cues, and they know that people act according to mental processes that allow them to plan and choose between different actions to attain goals. Without an understanding that behavior is organized by a governing sense of purpose, it would be nearly impossible to form accurate predictions or interpretations of others' behavior. Understanding goal-oriented behavior, therefore, requires the ability to both perceive what another is doing, and why (i.e., what is the desired outcome?) This would allow an individual to predict a social partner's behavior, even if the desired end state is not achieved. Moreover, as noted by Call and Tomasello (2008: p. 189), in order to perceive a social partner as goal-directed, an individual "must understand not only his goals but also his perceptions because what he sees and knows helps to determine what he does." Thus, an individual who has an understanding of goal-directed action should follow gazes and understand some connection between seeing and knowing; distinguish between action sequences that have the same external appearance, but which differ by the underlying intention; and read and use communicative gestures effectively.

Note that this constellation of abilities might not result in a comprehensive appreciation of the psychological states of others (i.e., a representational theory of mind, or ToM), as we know exists in human adults. For instance, the definition above does not imply an awareness of another's beliefs (which may differ from the true state of the world). The current literature suggests that neither infants nor primates have a representational ToM, in which they are aware not only of others' minds, but also that their thoughts and beliefs might differ from their own (Sabbagh et al., 2013; Call and Tomasello, 2008). Therefore, an understanding of goal-oriented behavior in others might be viewed as a rudimentary, but essential and precursory component of a representational ToM.

Theoretical Orientations

Although it is generally agreed that goal-oriented behavior plays a significant role in social cognition, several aspects of this ability remain under debate. Among developmentalists, there are differing views on the time line of its emergence. Most agree that by 4 years of age, children are aware of the minds and beliefs of others and further, those thoughts and beliefs may differ from their own (Wellman, 1990). However, preverbal children's understanding of others' minds remains a controversial subject. One viewpoint is that infants are born as

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23075-4

a 'blank slate,' and can only learn to socialize through experience (James, 1890; Piaget, 1954). Accordingly, infants must learn to distinguish between physical and underlying psychological states through a domain-general process. Thus, for the first 2 years of life, infants perceive social partners only in terms of their external physical behavior, rather than psychologically, and do not become aware of the thoughts underlying actions until as late as the end of the second year of life (Piaget, 1954).

In contrast, others assert that infants are born with a 'social brain,' containing domain-specific abilities that make newborns preadapted to social interaction (e.g., Grossmann and Johnson, 2007; Trevarthen, 1979). This allows them to process information about the social environment differently from the physical world (e.g., Grossmann and Johnson, 2007). According to this view, the mind consists of specialized domains that each process represents certain kinds of information (e.g., Legerstee, 2013). The social cognition domain directs infants to perceive social input appropriately. Within this framework, it is possible to examine rudimentary nonverbal social abilities during the first 2 years of life to trace the development of social cognition. It is from this latter perspective that we examine goal-oriented behavior in infants and toddlers. Because normal sociocognitive development is dependent not only on biological predispositions, but also on interactions with the social environment, it should be noted that there are cultural variations in the development of goal-oriented behavior, particularly with respect to the time line of the emergence of the abilities described below.

In comparative psychology, the problem is somewhat different; while we know that human children will eventually develop a representational ToM because this ability exists in adults, the same cannot be said with certainty about primates. Among comparative researchers, theoretical controversies focus on *whether* primates have goal-directed awareness, rather than *when* it emerges. In their seminal paper, Premack and Woodruff (1978) assessed whether a chimpanzee understood the goals of a human who was confronted with a series of physical problems, and tentatively claimed that chimpanzees showed evidence of a ToM. However, 20 years later, evidence seemed to suggest the opposite (see Tomasello and Call, 1997; for a review). Many argued that Premack and Woodruff's (1978) results could be explained in terms of 'lower-level' mechanisms such as learned associations between stimuli (e.g., Savage-Rumbaugh and Rumbaugh, 1979). Subsequent studies also failed to find evidence in support of complex social cognition, such as understanding others' goals, perceptions, or false beliefs, in apes (e.g., Call and Tomasello, 1999; Povinelli et al., 1994). This led many to conclude that primates did not perceive others as psychological beings, after all (e.g., Heyes, 1998; Tomasello and Call, 1997).

More recent evidence suggests the contrary. Although most agree that a representational ToM has not been established in primates (Call and Tomasello, 2008; Penn and Povinelli, 2007), much of the accumulating evidence suggests that primates can read below surface behavior to gain insights about the goals, perceptions, and knowledge of others (cf Call and Jensen, 2007; Rosati et al., 2010; Tomasello et al., 2003). These theorists argue that even if a 'lower-level' interpretation could explain a single empirical result, several such explanations would be required to account for the vast convergence of evidence favoring the mentalist explanation (Call and Tomasello, 2008). At present, it is more parsimonious to hypothesize a single unifying explanation: apes (and perhaps some monkeys) have knowledge about the goals and perceptions of others. It is from this perspective that we review the comparative literature examining primates' goal-oriented behavior.

Methodological Considerations

Infants, toddlers, and primates all lack the ability to verbalize their knowledge. As a result, developmental and comparative researchers are forced to rely on behavioral and/or biological measures to evaluate subjective understanding. This is particularly difficult when the question of interest is whether subjects have an awareness (which is not directly observable) about the internal states (goals, perceptions, and knowledge, which are also not directly observable) of others. Researchers have devised increasingly clever methods to gauge this understanding, but several issues remain.

One method that is often used in developmental psychology is visual habituation. This method assesses whether infants look differentially at actors that behave intentionally versus those that do not, suggesting that they appreciate the goal directedness of their actions (e.g., Trevarthen, 1979). However, it is difficult to prove conclusively that these attention biases reflect a comprehension of the underlying intentionality, rather than a superficial discrimination between external patterns of behavior. This issue has been addressed by using control conditions, in which external patterns of behavior are matched between intentional versus accidental situations, so that discriminative looking must be based on the underlying motives (e.g., Woodward, 1998). However, even these conditions are problematic, as infants might focus solely on the outcome, and thus perceive people's actions as no more than bodily movements. One solution is to adopt paradigms that involve unsuccessful or invisible outcomes, because they separate the visible end point from the underlying goal (e.g., Brandone and Wellman, 2009; Hamlin et al., 2008).

Other studies test whether infants can imitate incomplete actions, because this avoids some limitations of visual attention paradigms, such as reactions to novelty effects. If infants can complete an unseen goal, this suggests that they inferred the underlying motive, even though it was never physically demonstrated. However, requiring an infant to imitate an action sequence is physically demanding, particularly with younger infants. Alternatively, researchers can compare infants' communicative behaviors (e.g., reaching and banging arms) between situations in which a social partner's intentions differ (e.g., Behne et al., 2005). In these studies, the social partner's body movements are matched between conditions, as is the outcome of the interaction (e.g., a toy is not shared), so differential responses between conditions appears to result from an understanding of the underlying intention.

A final issue in developmental psychology is that studies occasionally use similar methods but find different patterns of results, with infants of comparable ages (cf Behne et al., 2005; Marsh et al., 2010). This highlights the need to select age-appropriate behavioral measures. For instance, it is possible

that infants at different ages use different responses, such as conventional (e.g., affective responses and instrumental reaches) at younger ages versus more stylized communicative responses at older ages (Crais et al., 2004; Mosier and Rogoff, 1994).

Like research with infants, comparative studies on social cognition in primates demonstrate the importance of methodology. Researchers testing the same or closely related species have found conflicting results depending on their methodology (Barth et al., 2005). Sometimes, the methodological detail responsible for the resultant difference can be ascertained; for instance, consistent differences have been found in social cognitive tasks that are competitive versus cooperative in nature (Hare and Tomasello, 2004). Unfortunately, it is not always obvious what caused the difference in results, because discrepancies between the reported methods are not apparent. Additionally, many experiments require primates to infer the goals of a human, rather than a conspecific. It is possible that their performance on these tasks might be facilitated if primates were given more naturalistic social contexts, in which to infer goals (cf Hare and Tomasello, 2004; Itakura et al., 1999). This underscores the importance of choosing methodological details with care, bearing in mind the natural ecology of the species under investigation.

Finally, an issue in many primate studies is the limited availability of captive primates. Indeed, much of what is known is based on only a few species – capuchins, rhesus monkeys, and chimpanzees – due to their prevalence in research facilities and zoos. Although this has resulted in a rather comprehensive body of knowledge on these species, it has come at the expense of learning about phylogenetic trends. Additionally, limited sample sizes mean that only a narrow range of ages are represented. Many studies include only adult subjects, and those including juveniles often have only a few, leading to a poor understanding of the developmental trajectories of socio-cognitive abilities in primates (chimpanzees are an exception – see Matsuzawa et al., 2006; Tomasello and Carpenter, 2005; Tomonaga et al., 2004). Often, trends emerge that suggest there may be developmental or sex differences, but firm conclusions cannot be drawn due to the limited sample size.

Development of Goal-Directed Behavior

Goal-Oriented Visual Behavior

From birth, human infants are attracted to agents that act intentionally (Trevarthen, 1979). They attend to faces, eyes, voices, and body movements (e.g., Markova and Legerstee, 2006). By 2–3 months of age, if a social partner suddenly halts a conversation with an infant, the infant shows a characteristic response to the adult's sudden 'still face' (Tronick, 1978): They reduce smiling and gazing, while trying to reengage the partner. Some theorists posit that these reactions reveal infants' understanding that when people look at them, they intend to communicate (e.g., Tronick, 1978).

Studies using visual habituation have also demonstrated that infants appreciate the goal directedness of successful actions during their first year (e.g., Woodward, 1998). Six-month-old infants respond to human actions, but not nonhuman ones, as though they are goal directed (e.g., Woodward, 1998). In an 'unsuccessful action' study (in which an internal goal is attempted, but not achieved), using visual habituation, 10- and 12-month-olds inferred the goal directedness of both successful and unsuccessful reaching actions, but 8-month-olds only recognized the goal directedness of successful actions (Brandone and Wellman, 2009). Other research using visual habituation suggests that the ability to infer unfulfilled goals might emerge even earlier, in 6-month-olds (e.g., Legerstee et al., 2000).

Between 5 and 7 months, infants also look to faces as a source of information during ambiguous situations. In a study, 7-month-old infants looked at an adult's face (as if to say, "Why are you doing that?") significantly more often when the adult teased the infant, by preventing a ball from being shared, than when the ball was given in the expected manner (Legerstee, 2005: ch. 5). It has been argued that this type of gaze-monitoring indicates mental state awareness, because typically developing infants, but not infants with autism, monitor the gazes of people more often in ambiguous situations than in nonambiguous situations (Phillips et al., 1992).

Visual behavior during triadic social interactions (a social dyad plus a third individual or object) is another indicator of infants' goal-oriented awareness. Gaze-following implies that infants are aware that their partner sees/perceives the same object that they do (coordinated attention), their speech concerns that object (labeling), and/or the emotion on the person's face is a result of looking at the object. These behaviors develop in increasing complexity during the first year of life (Legerstee and Barillas, 2003).

Infant gaze-following has also been tested using 'object-choice' tasks, in which a prize is hidden in one of several objects, and an adult indicates its location by gazing at it. The rationale is that if an infant chooses the correct object, based on the direction of the gaze, then the individual understands the communicative intent behind the behavior. Several studies have shown that infants as young as 18 months can do this successfully and appear to do so spontaneously (e.g., Itakura and Tanaka, 1998; Tomasello et al., 2007).

The visual behavior of some primates shows a striking resemblance to humans. For instance, mutual gaze as a means of communicating social intent is also seen between mother and infant chimpanzees. Matsuzawa (2006) has argued that this behavior – unique to hominoid species (humans and apes) – may be a precursor for the human mother–infant relationship. Mutual gaze is generally not present among monkey species, for which direct gaze is often perceived as a threatening signal. This behavior increases at 2 months of age, which coincides with the onset of social smiling in this species (Bard et al., 2005). Further, like humans, infant chimpanzees show a visual preference for stimuli that mimic goal-directed behavior versus those that do not (Uller, 2004).

Object-choice tasks have also been used to test primates' abilities to follow gazes. However, primate eyes differ morphologically from human eyes: The sclera in most primate species is not white, as it is in humans, and less of the sclera is visible, making it much more difficult to detect the direction of gaze. Consequently eye gaze is not as salient as a social cue in primates (Kobayashi and Kohshima, 2001), and great apes have been shown to rely more on the orientation of a head than gaze direction to locate a hidden object (Tomasello et al.,

2007). Thus, many studies examining this ability include not only gaze-orientation conditions, but also head orientation (e.g., Itakura, 1996; Marsh, 2012; Tomasello et al., 2007). The balance of research suggests that apes can use gaze or head orientation as a cue, and appear to do so spontaneously, or with minimal training (e.g., Call et al., 1998; Marsh, 2012; but see Barth et al., 2005; Itakura et al., 1999). The ability to follow gazes emerges in infant chimpanzees around 9 months of age (Okamoto-Barth and Tomonaga, 2006), although other research suggests it may occur later in ontogeny (Tomasello et al., 2001). This behavior has been cited as evidence that apes, like infants, understand the communicative intentions underlying visual gestures.

Research with monkeys and prosimians (which are more distantly related to humans than apes) provides less evidence to support spontaneous use of gaze. Although many species of monkeys follow gaze and/or coorient (e.g., Ruiz et al., 2009; Tomasello et al., 1998), they seldom use this social cue as a source of information in object-choice tasks, at least without extensive training (e.g., Anderson et al., 1995, 1996; Itakura, 1996). Thus, there may be a phylogenetic divide between hominoids and monkeys in terms of understanding the communicative significance of gaze and head orientation as an intentional social cue.

Another similarity between apes and infants is that apes also have a tendency to check back to a social partner's face (social referencing) when faced with uncertainty. For instance, if they follow a social partner's gaze, but fail to see any obvious target, they will check back to the social partner to ensure that they have followed the gaze correctly (Bräuer et al., 2005; Call et al., 1998). This suggests that they have expectations about the reasons underlying a social partner's gaze, and expect to see something interesting at the same location. In this case, however, it appears that the function of the apes' behavior may be to confirm the direction of the social partner's gaze (i.e., what are you looking at?), rather than checking intentions (i.e., why are you looking at that?). This behavior has not been observed in monkeys.

Finally, several experiments have investigated whether primates understand the connection between seeing and knowing. Although this remains a controversial topic, some research suggests that apes understand something about the perceptions of others, whereas monkeys do not. Specifically, apes adjust their rates and locations of gesturing to experimenters and/or conspecifics, depending on the attentional state and direction of the experimenter (e.g., Kaminski et al., 2004; Povinelli et al., 2003). They can also adopt the perspective of conspecifics, and adjust their behavior accordingly (Hare et al., 2000). In contrast, although some monkey species can follow gazes, it has been argued that this behavior is reflexive in nature and does not demonstrate an understanding about what others see (Rosati et al., 2010). This may again represent emerging complexity in social cognition, across phylogenetic time.

Goal-Oriented Behavioral Responses

From 9 months of age, infants imitate novel and combined action sequences (Meltzoff, 1988). When shown a human agent, a nonhuman agent (stuffed animal), and a surrogate object with mechanical pincers as hands, 10-month-old infants reenacted the successful actions of the human and nonhuman actors, but completed the goals of the human agents only (Legerstee and Markova, 2008). These findings were replicated with 7-month-old infants in a simplified imitation task (Hamlin et al., 2008). Further, 14- and 18-month-old infants imitate intentional action sequences, but not behavior that appears to be accidental (Carpenter et al., 1998).

Rather than relying on imitation, Behne et al. (2005) measured infants' natural behavioral responses (e.g., reaching and banging their arms) to a series of situations. They found that 9-, 12-, and 18-month-olds, but not 6-month-olds, reacted differently when an adult was unwilling rather than unable to pass a toy, despite perceptually matched, similar outcomes (the toy was not shared). Using a similar method but different behavioral measures, such as smiles and vocalizations, Marsh et al. (2010) demonstrated this ability in 6-month-olds.

This paradigm has also been used with primates. Both chimpanzees (Call et al., 2004) and capuchin monkeys (Phillips et al., 2009) stayed longer in a testing area and showed less frustration when an experimenter was trying, but unable, to give food than when she had no intentions of sharing. Moreover, capuchins showed this behavior with a human social partner, but not when analogous but inanimate objects were used (Phillips et al., 2009). In similar studies, in which the superficial movements of an experimenter were matched between conditions, but the context of the behavior suggested different intentions, apes once again demonstrated the ability to infer underlying goals (Warneken and Tomasello, 2006; Buttelmann et al., 2012). Other research shows that both monkey and ape species can discriminate between accidental and goal-directed behaviors in others (Call and Tomasello, 1998; Lyons and Santos, unpublished data). For instance, in object-choice tasks, apes choose objects that have been intentionally marked by an experimenter, rather than those that were inadvertently marked (Call and Tomasello, 1998).

Comparative researchers have also investigated primates' imitation of goal-directed actions. Like young children, apes are more likely to imitate goal-directed actions (e.g., Call et al., 2005), even when they are unsuccessfully demonstrated (Tomasello and Carpenter, 2005). Moreover, they can infer and complete goals using alternate means, rather than imitating goal-irrelevant behaviors (i.e., goal emulation; Buttelmann et al., 2007; Horner and Whiten, 2005). Whether this ability is shared more widely among primates is, as yet, unknown, as there has been little research conducted beyond hominoid species.

Use of Goal-Oriented Communicative Cues

Before infants use speech to communicate, they use gestures such as pointing, giving, and requesting to direct people's attention to aspects of their environment (Bates et al., 1975). Imperative pointing is a goal-oriented instrumental behavior to direct another's actions (e.g., "I want that thing"). In contrast, declarative pointing, when an indicated object or event is used as a means of making a nonverbal comment (e.g., "Wow, look at that!"), reflects recognition that one can share experiences and information with others and has been linked to the emergence of an understanding of intentions (Camaioni et al., 2004; Legerstee and Barillas, 2003). Current evidence suggests that imperative pointing emerges first, ontogenetically, but

declarative pointing is evident by the end of the first year of life (Camaioni et al., 2004; Liszkowski et al., 2006).

Infants' understanding of communicative gestures has been examined using object-choice tasks. Children as young as 1.5–2.5 years of age can use a social partner's pointed finger to locate a hidden prize, and do so spontaneously (e.g., Itakura and Tanaka, 1998; Povinelli et al., 1997). This suggests that they understand the communicative intention underlying the gesture, especially given that their performance with pointing cues is superior to that with inanimate markers (Tomasello et al., 1997). Sodian and Thoermer (2004) used a visual habituation paradigm to demonstrate that even 1-year-old infants understood the goal underlying a social partner's reaching gesture. They noted, however, that although infants followed a social partner's point, they did not yet understand the referential intent or goal underlying it.

In apes, there is also evidence from object-choice tasks that individuals can spontaneously use an experimenter's pointed finger as a cue (e.g., Itakura and Tanaka, 1998; Marsh, 2012), although contradictory results have been found (e.g., Povinelli et al., 1997; Tomasello et al., 1997). Moreover, one ape species (orangutans) learned to use socially given cues significantly faster than inanimate markers that were matched for spatial location (Marsh, 2012). This suggests that the social component (and perhaps the communicative intentions underlying the gestures) facilitated their understanding.

Monkeys have more difficulty in using communicative cues, such as pointing or reaching (e.g., Anderson et al., 1996), at least without extensive training (e.g., Anderson et al., 1995), suggesting that they do not understand these types of cues spontaneously (but see Hauser et al., 2007). Therefore, this is another aspect of goal-oriented understanding that may have emerged more recently, phylogenetically. Moreover, evidence for the *production* of spontaneous gestural communicative cues is also much stronger in ape species, than in monkeys (e.g., Blaschke and Ettlinger, 1987; Tanner and Byrne, 1999). However, whether or not apes grasp the intentional meanings behind their communicative gestures remains under debate (e.g., Call, 2010; Cartmill and Byrne, 2010).

Conclusion

Based on the evidence reviewed, it is clear that at least a rudimentary awareness of goal-oriented behavior is present from very early on developmentally, in humans, and phylogenetically, in primates. In particular, there are striking resemblances between human infants and great apes (chimpanzees, bonobos, gorillas, and orangutans), whereas the evidence in non-hominoid species (such as monkeys and/or prosimians) shows both demonstrated differences as well as empirical gaps in the literature.

With respect to goal-oriented visual behavior, newborn infants and infant chimpanzees show visual preferences for agents that display goal-directed behavior. By 2–3 months, human and chimpanzee infants engage in mutual gaze and increased social smiling, as a means of communicating social intent in mother–infant dyads. Older infants (5–7 months) and great apes also check back at faces of social partners, as a form of social referencing, although the reasons they do so may differ between species. Monkeys, to date, have not been recorded engaging in this behavior.

From 5 to 15 months, infants show increasing complexity in gaze-following and joint attention. They also follow others' gazes as a cue during object-choice tasks, and do so spontaneously, from as young as 18 months. Apes also show the ability to follow an experimenter's gaze in an analogous task and can use gaze (usually in conjunction with head orientation) to solve object-choice tasks spontaneously, although contrary evidence exists. To date, prosimians and monkeys appear to have much more difficulty in using gaze as an informative social cue, particularly without extensive training.

There is also strong evidence for the ability to distinguish intentions based on unsuccessful actions in both infants and primates. Infants have demonstrated this ability between 6 and 12 months, both through their visual behavior, as well as in their imitative and communicative behaviors. Great apes and capuchin monkeys can also infer underlying goals and/or distinguish between an experimenter's intentions, despite similar external patterns of behavior. Primates can also discriminate between accidental and goal-directed behaviors. Thus, this is one area in which monkey species show convergent patterns of behavior with apes and human infants. However, only apes and infants have, thus far, shown the ability to imitate goal-directed actions and engage in goal emulation.

Finally, human infants, from approximately 12 months of age, and at least some great apes, engage in communicative gesturing. Although infants appear to do so intentionally and referentially, it remains unclear whether this is the case for apes, and whether gesturing occurs spontaneously at all in monkeys. Human infants (from approximately 1.5–2 years of age) and apes also use a social partner's reach or point as a cue to locate prizes in object-choice tasks, and do so more effectively than with inanimate cues that are matched for spatial position. Monkeys have much more difficulty in using these types of cues, particularly without extensive training and experience.

Taken together, humans, across the first 2 years of life, and at least the great ape species, show abilities consistent with an understanding of goal-oriented behavior, across several behavioral categories. Given that the extant great ape species began to diverge from the human lineage some 10 million years ago, it is possible that this increasing sociocognitive complexity was, and continues to be, a necessary and precursory component of the multifaceted mental awareness that we see in modern day human adults.

See also: Cognitive Development During Infancy and Early Childhood across Cultures; Imitation During Infancy and Early Childhood; Intentionality During Infancy and Early Childhood, Development of; Joint Attention During Infancy and Early Childhood Across Cultures; Social Referencing During Infancy and Early Childhood across Cultures.

Bibliography

Anderson, J.R., Montant, M., Schmitt, D., 1996. Rhesus monkeys fail to use gaze direction as an experimenter-given cue in an object-choice task. Behavioural Processes 37, 47–55.

Anderson, J.R., Sallaberry, P., Barbier, H., 1995. Use of experimenter-given cues during object-choice tasks by capuchin monkeys. Animal Behaviour 49, 201–208.
Bard, K.A., Myowa-Yamakoshi, M., Tomonaga, M., Tanaka, M., Costall, A., Matsuzawa, T., 2005. Group differences in the mutual gaze of chimpanzees (*Pan troglodytes*). Developmental Psychology 41, 616–624.
Barth, J., Reaux, J.E., Povinelli, D.J., 2005. Chimpanzees' (*Pan troglodytes*) use of gaze cues in object-choice tasks: different methods yield different results. Animal Cognition 8, 84–92.
Bates, E., Camaioni, L., Volterra, V., 1975. The acquisition of performatives prior to speech. Merrill-Palmer Quarterly of Behavior and Development 21, 205–226.
Behne, T., Carpenter, M., Call, J., Tomasello, M., 2005. Unwilling versus unable: infants' understanding of intentional action. Developmental Psychology 41, 492–499.
Blaschke, M., Ettlinger, G., 1987. Pointing as an act of social communication by monkeys. Animal Behaviour 35, 1520–1523.
Brandone, A., Wellman, H.M., 2009. You can't always get what you want: infants understand failed goal-directed actions. Psychological Science 20, 85–91.
Bräuer, J., Call, J., Tomasello, M., 2005. All great ape species follow gaze to distant locations and around barriers. Journal of Comparative Psychology 119, 145–154.
Buttelmann, D., Carpenter, J., Tomasello, M., 2007. Enculturated chimpanzees imitate rationally. Developmental Science 10, F31–F38.
Buttelmann, D., Schütte, S., Carpenter, M., Call, J., Tomasello, M., 2012. Great apes infer others' goals based on context. Animal Cognition 15, 1037–1053.
Call, J., Carpenter, M., Tomasello, M., 2005. Copy results and copying actions in the process of social learning: chimpanzee (*Pan troglodytes*) and human children (*Homo sapiens*). Animal Cognition 8, 151–163.
Call, J., Hare, B., Tomasello, M., 1998. Chimpanzee gaze following in an object-choice task. Animal Cognition 1, 89–99.
Call, J., Hare, B., Carpenter, M., Tomasello, M., 2004. 'Unwilling' versus 'unable': chimpanzees' understanding of human intentional action. Developmental Science 7, 488–498.
Call, J., Jensen, K., 2007. Chimpanzees may recognize motives and goals, but may not reckon on them. Empathy and Fairness: Novartis Foundation Symposium 278, 56–70.
Call, J., Tomasello, M., 1998. Distinguishing intentional from accidental actions in orangutans (*Pongo pygmaeus*), chimpanzees (*Pan troglodytes*), and human children (*Homo sapiens*). Journal of Comparative Psychology 112, 192–206.
Call, J., Tomasello, M., 1999. A nonverbal false belief task: the performance of children and great apes. Child Development 70, 381–395.
Call, J., Tomasello, M., 2008. Does the chimpanzee have a theory of mind? 30 years later. Trends in Cognitive Sciences 12, 187–192.
Call, J., 2010. March. Action understanding and gesture acquisition in great apes. In: Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction. IEEE Press, p. 283.
Camaioni, L., Perucchini, P., Bellagamba, F., Colonnesi, C., 2004. The role of declarative pointing in developing a theory of mind. Infancy 5, 291–308.
Carpenter, M., Akhtar, N., Tomasello, M., 1998. Fourteen- through 18-month old infants differentially imitate intention and accidental actions. Infant Behavior and Development 21, 315–330.
Cartmill, E.A., Byrne, R.W., 2010. Semantics of primate gestures: intentional meanings of orangutan gestures. Animal Cognition 13, 793–804.
Crais, E., Douglas, D.D., Campbell, C.C., 2004. The intersection of the development of gestures and intentionality. Journal of Speech Language and Hearing Research 47, 678–694.
Gelman, R., Spelke, E.S., 1981. The development of thoughts about animate and inanimate objects: implications for research on social cognition. Social Cognitive Development: Frontiers and Possible Futures, 43–66.
Grossmann, T., Johnson, M.H., 2007. The development of the social brain in human infancy. European Journal of Neuroscience 25, 909–919.
Hamlin, J.K., Hallinan, E.V., Woodward, A.L., 2008. Do as I do: 7-month-old infants selectively reproduce others' goals. Developmental Science 11, 487–494.
Hare, B., Call, J., Agnetta, B., Tomasello, M., 2000. Chimpanzees know what conspecifics do and do not see. Animal Behaviour 59, 771–785.
Hare, B., Tomasello, M., 2004. Chimpanzees are more skillful in competitive than cooperative cognitive tasks. Animal Behaviour 68, 571–581.
Hauser, M.D., Glynn, D., Wood, J., 2007. Rhesus monkeys correctly read the goal relevant gesture of human agent. Proceedings of the Royal Society B 274, 1913–1918.
Heyes, C.M., 1998. Theory of mind in nonhuman primates. Behavioral and Brain Sciences 21, 101–114.
Horner, V., Whiten, A., 2005. Causal knowledge and imitation/emulation switching in chimpanzees (*Pan troglodytes*) and children (*Homo sapiens*). Animal Cognition 8, 164–181.
Itakura, S., Agnetta, B., Hare, B., Tomasello, M., 1999. Chimpanzee use of human and conspecific social cues to locate hidden food. Developmental Science 2, 448–456.
Itakura, S., Tanaka, M., 1998. Use of experimenter given cues during object-choice tasks by chimpanzees (*Pan troglodytes*), an orangutan (*Pongo pygmaeus*), and human infants (*Homo sapiens*). Journal of Comparative Psychology 112, 119–126.
Itakura, S., 1996. An exploratory study of gaze-monitoring in nonhuman primates. Japanese Psychological Research 38, 174–180.
James, W., 1890. The Principles of Psychology. MacMillan, New York.
Kaminski, J., Call, J., Tomasello, M., 2004. Body orientation and face orientation: two factors controlling apes' begging behavior from humans. Animal Cognition 7, 216–223.
Kobayashi, H., Kohshima, S., 2001. Unique morphology of the human eye and its adaptive meaning: comparative studies on external morphology of the primate eye. Journal of Human Evolution 40, 419–435.
Legerstee, M., 2005. Infants' Sense of People: Precursors to a Theory of Mind. Cambridge University Press, New York.
Legerstee, M., 2013. The development of the social brain, social connections and social bonds, social loss, and jealousy in infants. In: Legerstee, M., Haley, D.W., Bornstein, M.H. (Eds.), The Infant Mind: Origins of the Social Brain. Guilford Press, New York, pp. 223–248.
Legerstee, M., Barillas, Y., 2003. Sharing and pointing to objects at 12 months: is the intentional stance implied? Cognitive Development 18, 91–110.
Legerstee, M., Barna, J., Di Adamo, C., 2000. Precursors to the development of intention: understanding people and their actions at 6-months. Developmental Psychology 36, 627–634.
Legerstee, M., Markova, G., 2008. Variations in imitation: ten-month-old infant awareness of intentional action. Infant Behavior and Development 31, 81–91.
Liszkowski, U., Carpenter, M., Striano, T., Tomasello, M., 2006. 12-and 18-month-olds point to provide information for others. Journal of Cognition and Development 7, 173–187.
Lyons, D., Santos, L.R., 2007. Capuchin Monkeys (*Cebus paella*) Discriminate between Intentional and Unintentional Human Actions, unpublished data.
Markova, G., Legerstee, M., 2006. Contingency, imitation or affect sharing? Foundations of infants' social awareness. Developmental Psychology 42, 132–141.
Marsh, H.L., Stavropolous, J., Nienhuis, T., Legerstee, M., 2010. Six- and nine-month-old infants discriminate between goals despite similar action patterns. Infancy 15, 94–106.
Marsh, H.L., 2012. Orangutans' use of contiguous versus distal social and non-social cues in an object choice task. International Journal of Comparative Psychology 25, 299–308.
Matsuzawa, T., 2006. Evolutionary origins of the human mother-infant relationship. In: Matsuzawa, T., Tomonaga, M., Tanaka, M. (Eds.), Cognitive Development in Chimpanzees. Springer, Japan, pp. 127–141.
Matsuzawa, T., Tomonaga, M., Tanaka, M. (Eds.), 2006. Cognitive Development in Chimpanzees. Springer, Japan.
Meltzoff, A.N., 1988. Infants' imitation after a 1-week delay: longterm memory for novel acts and multiple stimuli. Developmental Psychology 24, 470–476.
Mosier, C.E., Rogoff, B., 1994. Infants' instrumental use of their mothers to achieve their goals. Child Development 65, 70–79.
Okamoto-Barth, S., Tomonaga, M., 2006. Development of joint attention in infant chimpanzees. In: Matsuzawa, T., Tomonaga, M., Tanaka, M. (Eds.), Cognitive Development in Chimpanzees. Springer, Japan, pp. 155–171.
Penn, D.C., Povinelli, D.J., 2007. On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. Philosophical Transactions of the Royal Society B: Biological Sciences 362, 731–744.
Phillips, W., Baron-Cohen, S., Rutter, M., 1992. The role of eye contact in goal detection: evidence from normal infants and children with autism or mental handicap. Development and Psychopathology 4, 375–383.
Phillips, W., Barnes, J.L., Mahajan, N., Yamaguchi, M., Santos, L.R., 2009. 'Unwilling versus unable': capuchins' (*Cebus paella*) understanding of human intentional action. Developmental Science 12, 938–945.
Piaget, J., 1954. The Construction of Reality in the Child. Basic Books, Oxford, UK.
Povinelli, D.J., Rulf, A.B., Bierschwale, D.T., 1994. Absence of knowledge attribution and self-recognition in young chimpanzees (*Pan troglodytes*). Journal of Comparative Psychology 108, 74.
Povinelli, D.J., Reaux, J.E., Bierschwale, D.T., Allain, A.D., Simon, B.B., 1997. Exploitation of pointing as a referential gesture in young children, but not adolescent chimpanzees. Cognitive Development 12, 423–461.
Povinelli, D.J., Theall, L.A., Reaux, J.E., Dunphy-Leli, S., 2003. Chimpanzees spontaneously alter the location of their gestures to match the attentional orientation of others. Animal Behaviour 66, 71–79.

Premack, D., Woodruff, G., 1978. Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences 1, 515–526.

Rosati, A., Hare, B.A., Santos, L.R., 2010. Primate social cognition: thirty years after Premack and Woodruff. Primate Neuroethology 1, 117–144.

Ruiz, A., Gomez, J.C., Roeder, J.J., Byrne, R.W., 2009. Gaze following and gaze priming in lemurs. Animal Cognition 12, 427–434.

Sabbagh, M.A., Benson, J.E., Kuhlmeier, V.A., 2013. False belief understanding in infants and preschoolers. In: Legerstee, M., Haley, D.W., Bornstein, H.M. (Eds.), The Infant Mind: Origins of the Social Brain. Guilford Press, Ney York, pp. 301–323.

Savage-Rumbaugh, E.W., Rumbaugh, D., 1979. Chimpanzee problem comprehension: insufficient evidence. Science 206, 1201–1202.

Sodian, B., Thoermer, C., 2004. Infants' understanding of looking, pointing, and reaching as cues to goal-directed action. Journal of Cognition and Development 5, 289–316.

Tanner, J.E., Byrne, R.W., 1999. The Development of Spontaneous Gestural Communication in a Group of Zoo-living Lowland Gorillas. The Mentalities of Gorillas and Orangutans. Comparative perspectives Cambridge University Press, Cambridge, 211–239.

Tomasello, M., Call, J., Gluckman, A., 1997. Comprehension of novel communicative signs by apes and human children. Child Development 68, 1067–1080.

Tomasello, M., Call, J., 1997. Primate Cognition. Oxford University Press, New York.

Tomasello, M., Call, J., Hare, B., 2003. Chimpanzees understand psychological states–the question is which ones and to what extent. Trends in Cognitive Sciences 7, 153–156.

Tomasello, M., Carpenter, M., 2005. The emergence of social cognition in three young chimpanzees. Monographs of the Society for Research in Child Development 70, 1–154.

Tomasello, M., Call, C., Hare, B., 1998. Five primate species follow the visual gaze of conspecifics. Animal Behaviour 55, 1063–1069.

Tomasello, M., Hare, B., Lehmann, H., Call, J., 2007. Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. Journal of Human Evolution 52, 314–320.

Tomasello, M., Hare, B., Fogleman, T., 2001. The ontogeny of gaze following in chimpanzee, *Pan troglodytes*, and rhesus macaques, *Macaca mulatta*. Animal Behaviour 61, 335–343.

Tomonaga, M., Tanaka, M., Matsuzawa, T., Myowa-Yamakoshi, M., Kosugi, D., 2004. Development of social cognition in infant chimpanzees (*Pan troglodytes*): face recognition, smiling, gaze, and the lack of triadic interactions. Japanese Psychological Research 46, 227–235.

Trevarthen, C., 1979. Communication and cooperation in early infancy: a description of primary intersubjectivity. In: Bullowa, M. (Ed.), Before Speech: The Beginning of Interpersonal Communication. Cambridge University Press, New York, pp. 321–347.

Tronick, E., 1978. The structure of face-to-face interaction and its developmental functions. Sign Language Studies 8, 1–16.

Uller, C., 2004. Disposition to recognize goals in infant chimpanzees. Animal Cognition 7, 154–161.

Warneken, F., Tomasello, M., 2006. Altruistic helping in human infants and young chimpanzees. Science 311, 1301–1303.

Wellman, H.M., 1990. The Child's Theory of Mind. MIT Press, Cambridge, MA.

Woodward, A., 1998. Infants selectively encode the goal object of an actor's reach. Cognition 69, 1–34.

Goffman, Erving (1922–1982)

Herbert Willems, University of Giessen, Germany

Published by Elsevier Ltd.
This article is reproduced from the previous edition, volume 9, pp. 6297–6301, © 2001, Elsevier Ltd.

Abstract

Erving Goffman (1922–1982) is considered to be one of the most important sociologists of the latter half of the twentieth century. He began his academic career at the University of Toronto and completed his sociology studies at the University of Chicago, where he primarily studied under Everett C. Hughes and W. Lloyd Warner. From 1962 Goffman was a Full Professor at Berkeley. In 1968 he accepted the Benjamin Franklin Professorship at the University of Pennsylvania in Philadelphia, where he wrote his *chef d'oeuvre 'Frame Analysis'* (1974). Goffman's main theme was the forms of immediate social interaction, 'face-to-face' encounters. He saw in them a specific type of social system and a sociological field in its own right. Goffman focused on tiny moments of people's behavior, which they themselves find more or less natural and are unaware of: fleeting glances, physical contact, apologies, greetings, compliments, silence, etc. All of Goffman's works, 11 books in total, are consistently oriented towards this empirical reality, but from the start Goffman also coined or redefined concepts and models (dramaturgy, role, identity, stigma, ritual, total institution, strategic interaction, frame, etc.). Goffman left behind a series of studies that are still relevant and influential today, and a network of concepts that is both highly complex and highly integrated.

The work of the sociologist Erving Goffman is regarded as a classic by most of his professional colleagues. Randall Collins (1988, p. 41) even considers him 'the greatest sociologist of the latter half of the twentieth century.' Goffman's influence, particularly on many younger researchers, is still powerful at the beginning of the twenty-first century and is by no means restricted to sociology. Goffman's works have been influential in numerous parallel disciplines, above all in social psychology, linguistics, philosophy, literature studies, pedagogics, psychiatry, and media studies. It cannot be denied that Goffman made important empirical, theoretical, and methodological contributions. The following essay will discuss them against the background of a short biographical sketch and relate them to relevant sociological contexts of discussions.

Biographical Sketch

Erving Goffman was born in the small Canadian town of Manville on 11 July, 1922, the son of Russian-Jewish immigrants. He began his academic career at the University of Toronto, where he took his BA in 1945. After completing his sociology studies at the University of Chicago, USA, where he primarily studied under Everett C. Hughes and W. Lloyd Warner, he took up his first post in Scotland, UK. From 1949 to 1951 he worked at the Department of Social Anthropology at Edinburgh University. After that he returned to Chicago, where he gained his doctorate with his thesis *Communication Conduct in an Island Community*, which he had prepared in Scotland. The empirical basis of his dissertation, which was supervised by W. Lloyd Warner, Donald Horton and Anselm L. Strauss, was a field study on the Shetland Islands. Following a spell as research assistant at the Department of Sociology at Chicago University, where his research included social stratification, Goffman was employed as a Visiting Scientist in the Laboratory of Socio-Environmental Studies at the National Institute of Mental Health (NIMH) in Bethesda (Maryland) from mid-1954 to the end of 1957. During this time Goffman carried out a one-year study in a mental hospital. This study on the basis of 'participant observation' formed the essential foundation of his book *Asylums* (1961a). In 1958 Goffman transferred to Berkeley as Assistant Professor of Sociology and member of the Center for the Integration of Social Science Theory at the University of California. From 1962 he was a Full Professor in Berkeley. In 1968 he accepted the Benjamin Franklin Professorship at the University of Pennsylvania in Philadelphia, where he substantially advanced his research program and wrote his *chef d'oeuvre 'Frame Analysis'* (1974).

His works, which reached a wide academic and non-academic audience and have been translated into many languages, earned Goffman numerous honors. He received a number of prestigious awards and was Visiting Professor at several universities at home and abroad. In 1981 Goffman became President of the American Sociological Association (ASA). At the peak of his powers after an academic career of almost 30 years, Goffman died on 20 November 1982.

Goffman's Work

Research Program and Intellectual Context

Goffman's main theme was the forms of immediate social interaction, 'face-to-face' encounters. He saw in them a specific type of social system and a sociological field in its own right. In his Presidential Address delivered at the end of his term as President of the American Sociological Association shortly before his death, he declared: 'My concern over the years has been to promote acceptance of this face-to-face domain as an analytically viable one – a domain which might be titled for want of any happy name, the interaction order – a domain whose preferred method of study is microanalysis' (Goffman, 1983, p. 2). This programmatic description of his position does not mean that Goffman regarded the interaction level as completely autonomous or socially fundamental. Rather he

stressed the manifold 'relativity' of the autonomy of the interaction order. In many places in his works Goffman pointed out its historical variability as well as links ('a loose coupling') between the micro-level of the interaction and social macro-structures (e.g., social stratification or bureaucratic organization). However, Goffman did not systematically consider the historical and social conditions and dependencies of his subject. Rather he was primarily concerned with observing and discovering seemingly insignificant aspects of daily behavior. He focused on tiny moments of people's behavior, which they themselves find more or less natural and are unaware of: fleeting glances, physical contact and distance, apologies, greetings, compliments, silence, etc.

All Goffman's works, 11 books in total, are consistently oriented towards this empirical reality, but from the start Goffman also adopted means of interpretation from various sociological schools and traditions, including above all the 'classics' William James (*see* James, William (1842–1910)), Emile Durkheim (*see* Durkheim, Émile (1858–1917)) and Georg Simmel (*see* Simmel, Georg (1858–1918)), structural functionalism (Alfred Radcliffe-Brown, Talcott Parsons (*see* Mead, George Herbert (1863–1931))), representatives of the Chicago School (George H. Mead (*see* Mead, George Herbert (1863–1931)), Charles H. Cooley, Louis Wirth et al.), existentialism (Jean-Paul Sartre), ethology, ethnomethodology (Harold Garfinkel, Harvey Sacks), and sociolinguistics (Dell Hymes). But Goffman never regarded his theoretical references, models and metaphors as ends in themselves, they were rather means to the end of discovering and describing social order.

Approaches, Positions, and Results

In the following attempt to briefly summarize and appraise Goffman's work, a number of key sociological terms either coined or redefined by Goffman will be used as a guideline.

Dramaturgy

Goffman became known for his analysis of social life based on the model of the theater (see 1959), and his model has remained influential. It is too simple to speak of Goffman's work, as for example Jonathan Turner does, as the 'dramaturgical school of interactionism,' but it can be demonstrated that the dramaturgical perspective consistently played an important role for Goffman. It can similarly be proved that Goffman's dramaturgical approach partly anticipated and formed a basis for many of his later constructs. This particularly applies to his works on strategic interaction, ritual, territoriality, identity, and stigma.

Within the context of the theater model Goffman reconstructed above all the strategies of social 'face work': forms of concealment, lying, self-stylizing, mystification, the manipulation of 'facades,' etc. His aim here was not just to show how individuals, selfish subjects in the fight for social recognition, attempt to present themselves in the best light; rather he constantly also saw dramaturgical everyday life as producing and reproducing social order, a restrictedly open and widely endangered process that is to be permanently established.

Goffman's dramaturgical approach and his later reflections on social territories (1971) and frames (1974) also help us understand the spatial organization (social ecology) of modern society. For example, it becomes clear that normality is a fictional impression to be produced by differentiating distinct spheres of perception, audiences, and representations; it is the foundation for the actors' basic social trust. Goffman's differentiation between 'frontstage' and 'backstage' (see 1959) became famous in this context.

Roles and identities

Goffman's early works in particular are to be seen in the context of role theory, influential at the time, which he criticized and developed further. Goffman treated role analysis primarily as interaction analysis, investigating the functioning and organization of the 'actual' practice of performing a role against the background of its normative frame. Thus Goffman developed concepts of patterns and styles of behavior, including the term 'role distance' (see 1961b), which has meanwhile acquired the status of a basic sociological term and which refers to a way of behavior which comments on the role and primarily serves either the interaction system or the selves relevant in the situation. For example, Goffman described a five-year-old boy riding a roundabout horse who by little irreverences in his behavior demonstrates that his current role does not correspond to his 'true,' 'more adult' self.

In *Stigma* (1963), Goffman differentiated the term 'self' in the context of an identity theory which distinguishes three types of identity in a basic sense that is still valid at the beginning of the twenty-first century: (1) social identity as a person's role set, (2) personal identity as a person's synchronic and diachronic individuality ascribed to him by observers, and (3) ego identity as a person's inner self-reference. Goffman used these concepts of identity together with the instruments of his dramaturgical approach above all in analyzing stigmatized deviations and deviants. He focused on the techniques of information control and emotion management of the stigmatized people as well as on the consequences of the stigmatization for their identities and socialization.

Total institutions

Goffman's book *Asylums* (1961a) represents an exception among his works in that it systematically goes beyond the interaction level. It describes a class of social organizations that has to do with confining, controlling and 'altering' people, such as mental hospitals, prisons, convents, homes, barracks, concentration camps, etc. However, Goffman's (1961a) investigated such 'total institutions', where the most elementary matters of course and norms of everyday interaction are broken, with the aim of exposing the interaction order and the self connected with it. Beyond that, Goffman's analysis demonstrated that total institutions form an ambivalent milieu for socialization. On the one hand they act as machinery for disciplining and normalizing behavior and interaction. On the other hand these institutions damage the moral self, e.g., self-respect and modesty, through various forms of humiliation.

Ritualizations

In his work Goffman repeatedly dealt with the moral sides of the interaction order. For Goffman the self is a moral fact which

implies rights and duties when showing respect and which can be observed in ritual practices and codes of everyday interaction. Every person appears as both subject and object of a ritual 'language,' a 'language' of politeness, decency, and tact, but also on occasion of contempt, wounding, humiliation, etc. (see Goffman, 1967).

At the end of his career Goffman elaborated his early ritual model to include ethological and frame theoretical considerations, on the basis of which he analyzed gender interactions and the interpretative patterns they are based on (see 1979). Goffman depicted gender as a combination of ritual styles of presentation representing beliefs about gender-specific essential attributes. These beliefs amount to a schematic asymmetry similar to that of the parent-child relationship. In daily interaction life and its media depictions (e.g., in advertising) men appear as superior beings to women by ritualizations of dress, physical contact, smiling, activity, etc.

Strategic interactions

For Goffman, ritualizations represented the most important social condition of profit-oriented strategic action on the level of interaction. Totally ignoring their contents, he analyzed this action within the framework of a model of strategic interaction oriented towards game theory (see 1969). The main point of departure and aim of Goffman's respective considerations were the social situation in which the subjects strive for control of the 'definitions' of the other players or rivals. Following on from his dramaturgical approach, Goffman above all investigated the conditions of competence and limits of 'impression management,' the strategically relevant interpretations of expression (e.g., of the body), the logic of reciprocal strategic observation and assessment, as well as the complexity of meaning and virtuosity of strategic moves. The empirical basis of the investigation consisted in particular in self-descriptions by a wide variety of strategic subjects (spies, confidence tricksters, adulterers, politicians, etc.). Goffman 'translated' their practical knowledge of action (knowledge of cleverness) into the categories of his analytic description, e.g., a typology of strategic moves.

Frames and framing subjects

Frame Analysis (1974) can be regarded as Goffman's main theoretical work, because it integrates, or could integrate, the many various analytic perspectives and terminological distinctions of his earlier works. *Frame Analysis* itself therefore forms something like a frame. It is characterized by two basic orientations. On the one hand, the interest is directed at 'structures' and procedures for establishing meaning which are effective in generating normality in social situations and which are valid, independent of these situations. On this level, which can be compared with the grammar of a language, Goffman revealed basic types of meaning and complex possibilities of transforming meaning (play, experiment, demonstration, etc.). On the other hand, on the level of the concrete situative events he aimed both at normal uncertainties, dangers, and the pressure to conform, as well as at extreme forms of abnormality, such as the 'negative experiences' of inmates of mental hospitals, whose habitual framings, which have proved their worth in everyday life, fail in the institution. The individual, whom Goffman revealed as the subject of the framing, is characterized by enormous powers of judgment and virtuosity, particularly in functionally complementing rules and also in manipulating them.

Forms of talk

In his last period, Goffman built on earlier studies and particularly concerned himself with conversations. The main result of this work is his book *Forms of Talk* (1981), which Goffman specifically placed in the context of *Frame Analysis*, one chapter of which was already devoted to conversations.

Goffman's late focusing on conversations can be seen in connection with the emergence of conversation analysis as a branch of ethnomethodology, which he approved of, but also criticized (see Bergmann, 1991). In contrast to conversation analysis, which tends to conceptualize the process of communication as an autonomously organized system, Goffman stressed the importance of ritual order, which is always culture-specific and which permits selves and consequently psychological states to develop. Goffman basically rejected the assumption by conversation analysts that sections of a conversation can be analyzed without reference to their situational and extra-situational contexts. Moreover, Goffman was mainly interested in a sociological approach to the contextuality of social processes. *Frame Analysis* represents this approach.

Goffman's Methods

Goffman's work is essentially the result of the author's talents, in particular his personal sensitivity, the subtlety of his powers of observation, and his ability to classify his material in his search for order. He displayed these talents in the course of practical research, which struck most observers as unconventional, even puzzling. Among the most irritating points are Goffman's data, whose heterogeneity can hardly be exceeded. In his works the reader comes across personal everyday experiences, novels, autobiographies, the daily press, essays by film critics, photographs, etc. In many other respects Goffman also deviated from the normal forms of sociological discourse. He abbreviated the traditional terminological vocabulary of his field, avoided lengthy argumentation, frequently changed perspective, focused on seemingly unimportant details, commented in an autoreflective manner on his own work, made ironic remarks and repeatedly told or referred to 'stories.' What is most important is that Goffman did not conform to any methodological canon, rather his methods entirely depended on the subject. This meant first of all that it was necessary to forgo certain traditional research methods. For example, the fact that those patterns of behavior that interested Goffman were generally unconscious ones meant that interviews were largely or totally impossible.

Goffman himself described his procedure as primarily 'naturalistic.' The scope of what he meant by that ranged from 'participant observation' (even in his own everyday life) to work with audio-visual recordings. His aim was always to get as close as possible to 'social nature' and to capture it as accurately as possible. On the other hand Goffman tried to achieve analytic distance and obtain information by various forms of 'alienation.' Alongside systematic work with models such as that of the theater, his most important technique was the strategy of alienation, which Paul Drew and Anthony

Wootton (1988, p. 7) called 'the investigation of the normal through the abnormal.' Goffman employed totally different forms of abnormality, such as stigmatizations, espionage, accidents, or celebrities as approaches to the logic of everyday normality.

In order to convert into texts the results that were generated by his methodological dialectic of approaching and distancing Goffman combined two strategies. On the one hand he developed a formal analytical language which contains the basic features of the interaction order and the self connected with it. On the other hand he employed a strategy of 'thick description' (C. Geertz). By using a literary style, a metaphorical language suited to the specific subject, and extensive quotations from his material, cleverly combined with his analytical observations, he succeeded in making his subject appear alive and viewing it in a new light.

Goffman's Importance for the Social Sciences Today

It is not easy to judge how relevant Goffman's work is for the social sciences at the beginning of the twenty-first century; its main significance probably lies in his contribution to discovering the interaction order. In this way Goffman opened up a specific research area and at the same time provided central categories and strategies for exploring it.

But quite apart from interaction sociology, Goffman also had a strong influence on a number of other research areas and he remains influential today. This is especially true of many aspects of research into identity and socialization. Of prime importance among them are Goffman's reflections on institutional control and discipline procedures, and on subjectivity. Goffman is especially appreciated today in research on normality and deviation as well as in gender studies, where some of his works are still standards now. In addition, he has a significant influence not only on the level of the everyday interaction order, but also on the sociology of certain social fields of action (politics, advertising, science, etc.) and in media research, where many of Goffman's analytical instruments have proved to be valid. Furthermore it can be seen that recently Goffman's works have increasingly been read and discussed from the point of view of general theories of civilization and society. For example, parallels have been drawn and connections made with the works of Norbert Elias (*see* Elias, Norbert (1897–1990)), Michel Foucault (*see* Foucault, Michel (1926–84)), and Pierre Bourdieu (see Willems, 1997). In general it is a remarkable fact that widely differing and contrasting schools have incorporated parts of Goffman's works. The spectrum of positions Goffman is 'relevant' for, ranges from modern systems theory through critical theory, right down to the rational-choice approach.

Just as the reactions to Goffman's work have been wide-ranging and many-sided, so evaluations of its scope and its identity have differed greatly. While most people see the importance of Goffman's work essentially on the level of empirical micro-observation, others (e.g., Anthony Giddens) believe that in it there is to be found a systematic theory that goes beyond the level of micro-sociology and is highly significant for the development of the social sciences. Similarly, various labels have been attached to Goffman's thinking (symbolic interactionism, phenomenology, role theory, structuralism, ethnomethodology, etc.).

No matter how the situation of his reception is to be judged, it cannot be denied that Goffman left behind a series of studies that are still relevant and influential today and a network of concepts that is both highly complex and highly integrated. Perhaps an equally important part of his legacy is the paradigm he provided by the style and ethos of his research.

See also: Durkheim, Emile (1858–1917); Identity and Identification, Social Psychology of; James, William (1842–1910); Simmel, Georg (1858–1918); Social Identity in Social Psychology; Social Psychological Theory, History of; Social Psychology.

Bibliography

Bergmann, J.R., 1991. Goffmans Soziologie des Gesprächs und seine ambivalente Beziehung zur Konversationsanalyse. In: Hettlage, R., Lenz, K. (Eds.), Erving Goffman—ein soziologischer Klassiker der zweiten Generation. UTB, Bern, Stuttgart, Germany, pp. 301–326.

Collins, R., 1988. Theoretical continuities in Goffman's work. In: Drew, P., Wootton, A. (Eds.), Erving Goffman. Exploring the Interaction Order. Polity Press, Cambridge, UK, pp. 41–63.

Drew, P., Wootton, A., 1988. In: Goffman, Erving (Ed.), Exploring the Interaction Order. Polity Press, Cambridge, UK, pp. 1–13.

Goffman, E., 1959. The Presentation of Self in Everyday Life. Doubleday Anchor, New York.

Goffman, E., 1961a. Asylums. Essays on the Social Situation of Mental Patients and Other Inmates. Doubleday Anchor, New York.

Goffman, E., 1961b. Encounters: Two Studies in the Sociology of Interaction. Bobbs-Merrill, Indianapolis, IN.

Goffman, E., 1963. Stigma. Notes on the Management of Spoiled Identity. Prentice Hall, Englewood Cliffs, NJ.

Goffman, E., 1967. Interaction Ritual: Essays on the face-to-face Behavior. Doubleday Anchor, New York.

Goffman, E., 1969. Strategic Interaction. University of Pennsylvania Press, Philadelphia, PA.

Goffman, E., 1971. Relations in Public. Microstudies of the Public Order. Harper and Row, New York.

Goffman, E., 1974. Frame Analysis. An Essay on the Organization of Experience. Harper and Row, New York.

Goffman, E., 1979. Gender Advertisements. Macmillan, London.

Goffman, E., 1981. Forms of Talk. Basil Blackwell, Oxford, UK.

Goffman, E., 1983. The interaction order. American Sociological Review 48, 1–17.

Manning, Ph, 1992. Erving Goffman and Modern Sociology. Stanford University Press, Stanford, CA.

Willems, H., 1997. Rahmen und Habitus. Zum theoretischen und methodischen Ansatz Erving Goffmans: Vergleiche, Anschlsse und Anwendungen. Suhrkamp, Frankfurt a. M., Germany.

Goldthorpe, John H. (1935–)

Christopher T Whelan, School of Sociology, Social Policy & Social Work, Queen's University Belfast, Ireland; and School of Sociology & Geary Institute, University College Dublin, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Goldthorpe's theoretical orientation involves a rejection of liberal and Marxist forms of historicism. It focuses on social patterns that can be established on a probabilistic basis and the manner in which they are created, sustained, and modified through the action and interaction of individuals. His macro- and microsociological concerns are combined in a wide-ranging exploration of multidimensional relational structures of inequality and the extent to which they result in people living in different social worlds.

Biography

John Goldthorpe was born in Yorkshire in 1935. He completed an undergraduate degree in history at University College London in 1956. In the same year he became a research student in the Department of Sociology at the London School of Economics (LSE). After 3 years as an assistant lecturer in the Department of Sociology at the University of Leicester he was appointed to a Fellowship at King's College Cambridge in 1960. In 1962, Goldthorpe and David Lockwood initiated what became known as the *Affluent Worker Study*. In 1969 Goldthorpe was appointed as Official Fellow of Nuffield College, Oxford. In the same year he was appointed the editor of *Sociology*, the newly established official journal of the British Sociological Association and became President of the Sociology Section of the British Association for the Advancement of Science.

In 1972, members of the Nuffield Social Mobility Group undertook a major enquiry into patterns of mobility of men in Britain. The findings from Goldthorpe's work were consolidated in *Social Mobility and Class Structure in Modern Britain* (1980/1987) published by Oxford University Press (OUP). His research on mobility was extended in the Comparative Analysis of Social Mobility in Industrial Nations project. The outputs from Goldthorpe's collaboration with Robert Erikson were consolidated in the landmark volume *The Constant Flux* published in 1990 by Clarendon Press. This work had a critical influence on the work of the Research Committee on Social Stratification of the International Sociological Association of which Goldthorpe acted as Vice President from 1974 to 1982 and as President from 1982 to 1985 and the European Consortium for Sociological Research (ECSR). Goldthorpe was a founding editor in 1985 of the *European Sociological Review*. In 2000, OUP published the first edition of *On Sociology* a collection of papers dealing with key theoretical and methodological issues particularly as they relate to the integration of theory and research. An extended two-volume edition was published by Stanford University Press in 2006.

Goldthorpe is an Emeritus Fellow of Nuffield College, Oxford, and a Distinguished Research Fellow at the Department of Social Policy and Intervention, University of Oxford, a Fellow of the British Academy, a Member of Academia Europea, and a Foreign Member of the Royal Swedish Academy of Sciences. In 1992, he was made a Commander of the British Empire (CBE) for his contribution to the advancement of the social sciences.

In what follows we will consider Goldthorpe's theoretical and methodological contributions before proceeding to a consideration of his substantive work in relation to social stratification and an overview of his contribution to sociology.

Theoretical Orientation

Macrosociological Perspective

As Gallie (1990: 29) observes, much of Goldthorpe's work can be seen as a prolonged engagement with the liberal theory of industrialism. Goldthorpe's (1971) critique involves a rejection of 'historicism' and a focus on its failure to specify the linkages between economic change and social action. Goldthorpe (1992) focuses on its neglect of the constraints deriving from relations with other nations, multinational business enterprises, and supranational agencies and those associated with structures and processes expressing established relations of power and advantage.

Social Action

The social action perspective employed in the *Affluent Worker Study* involved the worker's definition of the work situation becoming a key explanatory variable (Goldthorpe et al., 1968a: 26). Central to this approach was the assumption that workers inject certain meanings into the work situation, the sources of which pertain to some significant extent to the individual's wider life situation.

More recently Goldthorpe (2006a: 139–152) spelled out the key elements of his approach to a theory of social action. These include a commitment to methodological individualism; the belief that analytic primacy lies with the intended and unintended consequences of individual action. Goldthorpe (2007a: 140) argues that the appropriate phenomena of concern for sociologists are social patterns established on a probabilistic basis. The explanatory task of sociology he argues, following Boudon (1990) and Max Weber, is to show how they are created, sustained, modified, or disrupted through the action and interaction of individuals. Goldthorpe (2006a: 142) notes that Rational Action Theory (RAT) may

be understood as of an 'outcome-oriented' kind in which certain requirements are met regarding the nature of actors' goals, their beliefs relevant to their pursuit, and the courses of action which they then follow. Goldthorpe (2006a: 158–159) argues that, in relaxing the assumptions of the version of RAT employed in mainstream economics, the key idea that sociologists must exploit is that of *subjective* rationality. Actors may hold beliefs and pursue courses of action for which they have 'good reasons' in the circumstances in which they find themselves, even though they may fall short of the standard of rationality that utility theory would presuppose. The social structuring of knowledge should be treated as a crucial aspect of the situation. (The long-term consistency of Goldthorpe's position in this regard can be seen by considering his 1973 critique of ethnomethodology.)

Methodological Issues

For Goldthorpe (2004) the primary goal of sociology as social science (SSS) is the explanation of opaque social phenomena. He sees a natural relationship between SSS and descriptive sociology because of the role of the latter in establishing social regularities, particularly given the mutability of social phenomena. In a related fashion, Goldthorpe (1996) directs attention to the potential offered by the application of RAT to the quantitative analysis of large scale data sets. Statistical analysis plays an important role in defining the 'objects of study' of sociology, in conjunction with the concepts that sociologists employ.

Goldthorpe's (2001) preferred the style of sociology that is linked to his understanding of the relevance of different conceptions of causation. He distinguishes between causation as robust dependence, consequential manipulation and, his own preference, as a generative process. Further crucial distinctions relate to establishing the phenomenon, hypothesizing the generative process, and testing hypotheses. Goldthorpe (2012) stresses that while statistically analysis plays an important role in establishing sociological explananda, it cannot in itself lead to causal explanations of these regularities. Furthermore, he argues that if such explanation is the primary objective, then the counterfactual approach to causation will be of much reduced relevance.

From this perspective, Goldthorpe (2006a: 1–40) seeks to identify common methodological problems in qualitative styles of sociology. These included what he labels 'grand' historical sociology. However, he stresses that his objective is not to launch an all-out attack on such qualitative research. Rather he seeks to establish three specific points.

The first is that qualitative styles of sociology face methodological problems that their proponents largely neglect. These relate to the reliability and validity of the kinds of data on which chief reliance is placed and the manner in which they are used in the processes of theory construction and evaluation. Goldthorpe argues that claims to avoid complicity with 'positivistic' approaches lead their proponents into positions of an ur-positivistic kind. For 'grand' historical sociology this relates to a failure to acknowledge that historical fact involves inference from finite and incomplete artifacts. For case-oriented macrosociology his concerns relate to issues involving determinism vs probabilism and the failure to acknowledge the distinction between description, explanatory theory, and testing. These concerns are linked to the failure to acknowledge that the small *N* problem is one not of method but of insufficient information. He stresses the need to distinguish between recurrent social situations and processes that might be the subject of sociological theory and contingencies, distinctive conjunctures of events, or other singularities that such theory cannot comprehend.

Finally, Goldthorpe focuses on the problem that comparative macrosociology shares with ethnography arising from the difficulties associated with generalizing from insufficient samples by resorting to entirely deterministic or logical methods of analysis by invoking theory that presupposes knowledge of law like relations. Ultimately, he argues, it is necessary to deal with principles of selection and representativeness if one is to be in a position to answer the question relating to cases of *what*? (For a robust debate on Goldthorpe's views on comparative macrosociology see the collection of papers included in *The British Journal of Sociology*, 45. 1 March 1996.)

Social Stratification

Goldthorpe's macro- and microsociological concerns have been combined in his focus on a wide range of theoretical and substantive issues relating to the multidimensional nature of social stratification. His insistence on the contingent nature of the relationships between forms of stratification and their consequences forms a crucial part of his debates with theories of convergence, whether liberal or Marxist. Goldthorpe has consistently rejected reliance on 'umbrella' concepts of social stratification and synthetic indices. As Goldthorpe (2008: 352–353) makes explicit, his starting point does not invoke some counterfactual society but rather the existence a capitalist free market society. His program of social stratification research has involved in developing an understanding of the relational structures of inequality and their implications for different domains of social life.

Class and Status Analysis

For Goldthorpe, class analysis has as its central concern the relationships between class structures, mobility inequalities, and action. The focus is on the consequences of differentiation in employment relations for life chances and choices or, as Goldthorpe (2008: 352) puts it, the extent to which they result in people living in different social worlds. The objective is to investigate the interconnections between historically formed macrosocial structures and everyday experience. No assumption of the preeminence of class is required. No historicist assumptions are entailed nor do they require a theory of exploitation (Goldthorpe and Marshall, 1992: 382–383).

The class schema central to the pursuit of this agenda emerged from his work on social mobility. Its importance is reflected in its adoption as the basis of the British National Statistics Socio-Economic Classification and the European

Socio-economic Classification. Formalization of this schema, following Erikson and Goldthorpe (1992: 35–47) and Goldthorpe (2006b: 101–124), is based on two main principles of differentiation. The first is that of employment status. The second relates to the regulation of employment as a viable response to the weaker or stronger presence of monitoring and asset specificity problems in different work situations.

Goldthorpe addresses the theoretical question of how differences in employment relationships arise by drawing on RAT, particularly as employed in organizational and personnel economics. The general problems of the employment contract he argues relate to work monitoring and human asset specificity. The particular form in which they arise in terms of the principal–agent relationship is with regard to work monitoring of professional and managerial employees whom the employer engages the employee to act in the principal's interest in circumstances in which the agent's actions cannot be observed. The solution is sought in a diffuse exchange of services in return for compensation in which prospective rewards play a crucial role. This provides the basis for secure investment aimed at increasing employees' organization specific skills. The key distinction in relation to the latter is between regulation occurring via a 'labor contract,' involving a relatively short-term and specific exchange of effort for money, or a 'service' relationship, as typified in the kind of contract typical for professional staff, which involve important prospective rewards (Erikson and Goldthorpe, 1992: 41).

Goldthorpe (2006b: 120–124) notes that such differentiation is consistent with the assumption that employers will endeavor to modify contracts to shift the balance of risk toward employees. However, he is not persuaded by the arguments that globalization has led to service-type relationships being discontinued to a degree that undermines the significance of class (Goldthorpe, 2006a,b: 100–107, 122).

His approach involves an emphasis on social relations rather than individual attributes. The distinction between class and status is crucial. Goldthorpe argues that, while scales of social standing can be interpreted as indicative of general desirability of occupations, they cannot be seen as capturing social relationships of deference, acceptance or derogation (Goldthorpe and Hope, 1974). Status he understands as being expressed in differential association, especially in more intimate forms of sociability and lifestyles of differing distinction (Chan and Goldthorpe, 2007a: 1096–1097).

Goldthorpe (2008: 350) stresses the need to keep the measurement of social class independent of outcomes which it is intended to explain and of other dimensions of stratification to which we may wish to compare its explanatory power. Goldthorpe (2010) elaborates on the consequences of the failure to make such distinctions in his critique of the argument by Wilkinson and Pickett (2009) that income inequality has a negative impact on a range of social outcomes for all members of unequal societies. Goldthorpe notes that their understanding of the mechanism through which negative social outcomes are generated is crucially dependent on the existence of a close link between income inequality and social status but their failure to understand the complexities of this relationship leads them to inferences of a questionable kind.

Social Mobility

Goldthorpe (1980/1987) exploited developments in log-linear analysis that allowed for the reformulation of the exchange/structure distinction in terms of absolute mobility rates and relative chances of mobility. He proceeded to demonstrate that continuous economic growth, large-scale educational reforms, and the transformation of the division of labor were associated with the expansion of the salariat and substantial levels of absolute mobility. However, no significant reduction was observed in relative class advantages. Given Goldthorpe's explicit concern with the implications of extensive 'recruitment from below,' it is puzzling that one of the major criticisms of his work has been its supposed neglect of absolute mobility and structural change. This is so despite the fact that a focus on changing patterns of absolute mobility is central to Goldthorpe's critique of recent attempts in the UK to politicize social mobility issues (Goldthorpe and Jackson, 2007; Goldthorpe, 2012).

To understand this response it is necessary to return to Goldthorpe's understanding of what constitutes an appropriate explananda for sociology. Goldthorpe (1990: 417) argues that occupational change must be treated as 'exogenous,' permitting only description rather than explanation. It is not that Goldthorpe considers the relative mobility as more important than the absolute mobility but rather that he sees the former as constituting the kind of opaque regularities for which sociologists should seek to provide theoretical explanation.

Similarly, while Goldthorpe acknowledges that gender stratification in the labor market operates to the advantage of men, he sees the task of explaining such segregation and its impact on the class structure as lying outside the realm of class analysis. A choice must be made whether to allocate class position on the basis of individual attributes or on information available for the family unit or household as a whole. Adopting an individual approach has the consequence that women are observed to experience substantially greater downward mobility (Erikson and Goldthorpe, 1992: 239–245). The issue which arises is whether such differences should be taken at face value, in the context of the segregation of women into lower-level occupations together with discontinuous work histories. Adopting the individual approach requires accepting that women's class positions are less closely associated with their material circumstances, that their propensity for class-based partisanship is significantly lower and their class identification far more uncertain. The alternative interpretation is that it is rational and intelligible that women, whose partners occupy superior class positions, should regard their class interest and affiliations as being, in important respects, determined by their partner's employment (Erikson and Goldthorpe, 1992: 252). (For a detailed discussion of the debate in British sociology on gender and class mobility see Mc Rae, 1990.) Where class is determined by the occupational position of the member regarded as dominant in terms of work time and position, 'complete' mobility tables provide results that do not diverge significantly from those for men only tables (Erikson and Goldthorpe, 1992: 371).

The distinctions between absolute and relative mobility and between sociological and historical forms of explanation are

central to the extension of Goldthorpe's social mobility concerns on a comparative basis. This work forms part of his sustained critique of liberal industrialism which represents high and rising levels of mobility and openness as functional imperatives of industrial development. Erikson and Goldthorpe (1992) showed that industrial societies exhibited substantial variation in absolute mobility rates but that there was no support for convergent rates of such mobility or steadily increasing rates of upward movement. The observed variation in absolute mobility rates was not found to be amenable to explanation in sociological terms. Neither do the findings support the expectation that in relative rates tend to become more equal (Erikson and Goldthorpe, 1992: 253–277).

Extending such analysis, Goldthorpe focuses on the barriers to meritocracy involving countervailing class forces integral to any capitalist market society. In particular, he points to the crucial role of employers in deciding what importance should be attached to educational attainment. Goldthorpe and Jackson (2008) provide evidence of utilization of a wide variety of criteria intended to act as signals for productivity, many of which have nothing to do with educational attainment. A particularly relevant case study involves the Hungarian transition from state socialism. Bukodi and Goldthorpe (2010) examine this transition from two competing perspectives which they label as 'meritocracy as functional imperative' and 'market vs meritocracy.' The former sees the development of meritocracy as essential to the technological and economic dynamism of modern societies while the latter views it as incompatible, in important respects, with the principles of a free market economy and liberal democracy. None of the trends required to support the former argument were observed.

For Goldthorpe social mobility occurs within a class structure that conditions rates and patterns of intergenerational mobility through both the structure of opportunities and the typical attributes of class positions. What should be treated as common to individuals of all social classes is the objective of avoiding downward mobility. We will elaborate on this theoretical perspective below when considering class differentials in educational achievement.

More recently Goldthorpe's approach to analyzing class mobility has been criticized for operating at too macro a level and needing to be ratcheted down to the occupation level where 'real' social groupings emerge (Jonsson et al., 2009). Goldthorpe (2002) challenges the assumption that the study of the processes through which economic categories become real social groupings is the *sine qua non* of class analysis. Nor does he share the assumption that collective action has to proceed on the basis of transformation of class in itself to class for itself. His concern with class organization has focused on the levels at which interests are aggregated and the issues raised by the potential role of trade unions in corporate and social partnership (Goldthorpe, 1984). (The complexities of the relationships between occupational communities, images of society, and voting behavior were at the heart of *The Affluent Worker* study. Goldthorpe (1974) drew attention to the possibility that the consequences of inequality might be reflected not in class conflict but in a situation of Durkheimian anomie.) Erikson et al. (2012) conclude that the microapproach accounts for immobility rather than mobility and is substantially less successful in doing so for women. To be persuasive it needs to specify a mechanism that accounts for mobility and immobility. They accept that class analysis may not capture the highly specific processes involved in occupational mobility but they argue neither does the microclass approach, because they appear to be structured far more by father's class rather than father's occupation within class.

Class Analysis and Educational Inequalities

Goldthorpe (2006b: 21–44) notes the theoretical problems constituted by persisting class differentials in educational attainment are thrown into sharp relief by the reduction in gender differentials. Goldthorpe's explanatory account involves the development of an RAT which involves only features of class that derive directly from employment relations and avoids reference to class values, norms, and forms of consciousness. Two related arguments are involved in the initial explanatory propositions. Class differentials have been maintained because so too have the corresponding perceived differentials in costs and benefits. The consequence is that children in disadvantaged families require a greater assurance of success. Aspirations are understood in relative rather than absolute terms. Differences in choices are seen to reflect corresponding 'social distances' that must be traversed from differing class origins. This is consistent with the evidence that class differentials persist even when ability is held constant (Breen and Goldthorpe, 2001). The declining influence of costs is part of the explanation of growth in the number of children remaining in education beyond the compulsory period. However, for those from less advantaged class origins to choose to stay on at a more rapidly increasing rate than others would require a shift in *relative* risks and costs. The fact that gender disparities have been largely eliminated suggests a more positive cost–benefit evaluations of education for women in the light of changing gender relations and labor market conditions. The role played by gender differentiated patterns of inheritance also helps to account for the different choices made by the self-employed for sons and daughters.

Following Boudon (1974), Goldthorpe distinguishes between 'primary' and 'secondary' effects. The former captures the relationship between children's class origins and their average levels of demonstrated academic ability. Secondary effects operate through the decisions made by children and their parents regarding particular educational options.

Goldthorpe (2007b: 29) accepts that the degree of congruence between the subcultures into which children are socialized and those that prevail in schools can contribute to accounting for educational differentials. However, to explain persisting differentials it would be necessary to provide an account of how such cultural influences have maintained their differentiating power in the context of social transformations involving major educational expansion and reform. Goldthorpe (2006b) notes that Bourdieu's (1973) theory of 'cultural reproduction' offers such an explanation. 'Dominant classes' impose a pedagogy which requires of children an initial socialization into the 'dominant' culture. Educational inequalities are maintained by

unequal amounts of, largely arbitrarily defined, 'cultural capital' transmitted over a longer period than economic capital and with less possibility of reversibility.

Goldthorpe argues that this account founders on the predominance of 'first-generation' students entering more selective and academic forms of education during the post-war decades. With expansion and reform, the constraints on choice that primary effects impose tend to weaken and greater scope is created for the impact of secondary effects, as reflected in educational choices. It is assumed that, in their central tendencies, such choices can be understood as reflecting evaluations of the cost and benefits of possible alternatives. Three mechanisms contribute to the emergence of class differences. These comprise relative risk aversion, differences in ability, and expectations and resources. The expectations derived from this model differ from those associated with standard human capital theory in which class differences are viewed simply as the result of corresponding differences in average ability and resources. However, they contrast even more sharply with sociological explanations in which cultural capital is the key explanatory factor.

Goldthorpe (2008: 31) distinguishes between cultural resources and cultural values in order to accommodate class differences in the former that are necessary to the realization of values that are largely shared. From this perspective, the tendency of children from working class families to pursue less ambitious careers is not necessarily a consequence of poverty of aspirations but arises from the differences in 'social distances' that must be traversed. This approach allows a role for choice in relation to secondary effects that is conditioned by a structure of opportunities and constraints (these include not only social class of origin but also parental class and parental education and status (Bukodi and Goldthorpe, 2013)) including class variation in evaluations of risks and rewards. However, such choice involves more than the expression of unchanging and compelling subcultural values and norms (Goldthorpe, 2006b: 80–81). (For a robust set of exchanges on Goldthorpe's critical comments on Bourdieu's interpretation of class differentials in educational attainment see the contributions to *Sociologica* 2007, 2 and 2008, 1.)

Cultural Consumption

Chan and Goldthorpe (2007b) identify three rival lines of argument in the current sociological literature dealing with the relationship between social stratification and cultural taste and consumption. These comprise the homology, individualization, and omnivore–univore arguments.

The homology argument claims that social and cultural stratification map closely on to each other. Individuals in higher and lower social strata, respectively, prefer and predominantly consume 'high' or 'elite' culture, and 'popular' or 'mass' culture. For Bourdieu (1984) class and status are not to be understood as different forms of social stratification, rather status is the symbolic aspect of the class structure which is not in itself reducible to economic relations. The individualization thesis holds that differences in cultural taste and consumption are losing their grounding in social stratification and are becoming more a matter of individual 'self-realization.' In contrast with Bourdieu's position, the emphasis shifts from habitus to freedom. The omnivore–univore argument advances the hypothesis that the cultural consumption of individuals in higher and lower strata differs chiefly in comprising not only more 'high-brow' culture but also more 'middle-brow' and 'low-brow' culture. The crucial contrast is not that of 'snob versus slob' but that of cultural omnivore versus cultural univore. As Chan and Goldthorpe (2007b) note, this interpretation can developed in two rather different ways which come close, respectively, to the homology and individualization arguments.

Goldthorpe and Chan's (2007b) analysis of musical tastes in the UK leads them to conclude that the homology argument is undermined by the fact that they could identify a potential 'mass' of univore musical consumers but not a musical 'elite.' Univores make up a majority of even the professional and managerial classes. At the same time, the degree of social differentiation in consumption patterns undermines the individualization thesis. Social status and education are a good deal more important than class in differentiating between univores and omnivores, refuting the claim that status is simply a symbolic reflection of class. Chan and Goldthorpe (2004) conclude that, while a status order can be discerned in present day Britain, it is less sharply demarcated than previously with status differences being less openly asserted from above and less deferentially acknowledged from below.

The Contribution to Sociology of John H. Goldthorpe

In the introduction to the first edition of *On Sociology* John Goldthorpe argued that sociology was in a state of intellectual disarray. Goldthorpe's central concerns related to a manifest lack of integration of theory and research, a collective failure on the part of sociologists to decide just what kind discipline sociology should or ought to be and an appeal to a form of pluralism that encourages cognitive relativism. These concerns were not new, as reflected in a letter to the *Times Higher Education Supplement* in July 1982 in which he bemoaned the preoccupation with types of 'theory' that are often no more than mere word spinning and the scandalous disregard for maintaining the standards of research expertise. Goldthorpe is not an uncontroversial scientist, as reflected in the title of a volume devoted to his research: *John H. Goldthorpe: Consensus and Controversy*. Here we have referred to the vigorous debates in which he has been involved relating to macrosociology, RAT, gender, mobility, and cultural capital. These exchanges reflect the fact that even those who most vehemently disagree with his conclusions understand that his intellectual standing requires that they engage with him. Goldthorpe, for his part, has taken his critics seriously but has shown himself to be consistently unswayed by intellectual fashions and has held firmly to the view that arguments have to be judged on their merits rather than by reference to the celebrity of their authors (Goldthorpe, 2007b: 1).

Goldthorpe's fundamental concern has been with the cumulative development of the discipline. His programmatic essays in *On Sociology* are focused on the development of

a kind of sociology in which theoretical and methodological discussion can reflect a genuine pluralism rather than an 'anything goes' intellectual climate. Goldthorpe has pursued this objective not only through his own research but through his roles as a colleague, mentor, collaborator, and office holder at Nuffield College, the European Consortium for Sociological Research and the Research Committee on Stratification of the International Sociological Association (ISA).

Goldthorpe's approach is not intended to promote a school of sociology, Thus, Goldthorpe (2006b: 81–82) has responded to the critiques of his development of RAT for sociology as 'antisociological' with a clear statement that theories have to be judged not by their provenance but by their consequences in terms of whether or how far they lead to adequate explanations and stand up to empirical test. Goldthorpe's approach to such micro–macro linkage has had at its center issues relating to the social structuring of knowledge and the need to understand the logic of the situation from the actor's point of view. From a macro perspective, in rejecting historicist perspectives in both their functionalist and Marxist forms, he has consistently emphasized the importance that should be attached to historical and political contexts and choice. There is no need Goldthorpe (1992: 431) stresses to accept that constraints of whatever kind operate in a manner that issues relating to purposive social action and what the future *could*, as opposed to *should*, look like become irrelevant.

Goldthorpe has consistently argued that sociology will benefit from confronting problems that arise exogenously and that require a focus on causes of effects rather than 'cherry-picking' problems chosen to illustrate the power of some generative model by a concentration on 'effects of causes' (Goldthorpe, 2001: 102–103). The danger of the latter approach he argues is that it can give rise to a greater interest in mechanisms per se than in the actual extent of their explanatory potential.

Following Merton (1959), he has stressed the ultimate complementarity of pure and applied science. Being able to show that problems have a 'double rationale' in terms of internal scientific development of sociology and applied importance provides protection against what he refers to as 'sociological dandyism': a preoccupation with models on account primarily of their intrinsic elegance, refinement, and subtlety (Goldthorpe, 2001: 100–101). Goldthorpe's own research achievements provide a perfect example of the manner in which rigorous theoretical development and methodological sophistication can be combined to produce a body of sociological work that is of major substantive relevance from the standpoint of both pure and applied interests.

Acknowledgment

I would like to thank Robert Erikson, Tony Fahey, Duncan Gallie, John H. Goldthorpe, Walter Müller, Philip O' Connell, Toni Schizzerotto, and Sally Shortall for helpful comments on an earlier version of this article.

See also: Bourdieu, Pierre (1930–2002); Capitalism; Causation: Physical, Mental, and Social; Causes and Laws: Philosophical Aspects; Functional Explanation: Philosophical Aspects; Intentionality and Rationality; Marx, Karl (1818–83); Merton, Robert K (1910–2003); Methodological Individualism: Philosophical Aspects; Rational Choice Explanation: Philosophical Aspects; Social Inequality in History (Stratification and Classes); Social Mobility, History of; White-Collar Workers, History of; Working Classes, History of.

Bibliography

Boudon, R., 1974. Education, Opportunity and Social Inequality. Wiley, New York.
Boudon, R., 1990. Individualism and holism in the social sciences. In: Birnbaum, P., Leca, J. (Eds.), Individualism. Oxford University Press, Oxford.
Bourdieu, P., 1973. Cultural reproduction and social reproduction. In: Brown, R.K. (Ed.), Knowledge, Education and Cultural Change. Tavistock, London.
Bourdieu, P., 1984. Distinction: A Social Critique of the Judgement of Taste. Routledge & Kegan Paul, London.
Breen, R., Goldthorpe, J.H., 2001. Class, mobility and merit. European Sociological Review 17 (2), 81–101.
Bukodi, E., Goldthorpe, J., 2011. Class origins, education and occupational attainment in Britain. European Societies 13 (3), 347–375.
Bukodi, E., Goldthorpe, J., 2013. Decomposing social origins': the effects of parents' Class status and education an the educational attainment of their children. European Sociological Review 29 (5), 1024–1039. http://dx.doi.org/10.1093/esr/jcs079.
Chan, T.-W., Goldthorpe, J.H., 2004. Is there a status order in contemporary British society? evidence from the occupational structure of friendship. European Sociological Review 20, 383–401.
Chan, T.W., Goldthorpe, J.H., January 2007a. Social status and newspaper readership. American Journal of Sociology 1124, 1095–1134.
Chan, T.W., Goldthorpe, J.H., 2007b. Social stratification and cultural consumption: music in England. European Sociological Review 23 (1), 19.
Erikson, R., Goldthorpe, J.H., 1992. The Constant Flux. Clarendon Press, Oxford.
Erikson, R., Goldthorpe, J.H., Hällsten, M., 2012. No way back up from ratcheting down: a critique of the 'microclass' approach to the analysis of social mobility. Acta Sociologica 55 (3), 211–229.
Gallie, D., 1990. John Goldthorpe's critique of liberal theories of industrialism. In: Clark, J., Modgil, C., Modgil, S. (Eds.), John H. Goldthorpe: Consensus and Controversy. The Falmer Press, London.
Goldthorpe, J.H., 1971. Theories of industrial society: reflections: on the recrudescence of historicism and the future of futurology. Archives Européenes de Sociologie 12, 263–288.
Goldthorpe, J.H., 1973. A revolution in sociology? Sociology 7 (3), 449–462.
Goldthorpe, J.H., 1974. Social inequality and social integration in modern britain. In: Wedderburn, D. (Ed.), Poverty, Inequality and Class Structure. Cambridge University Press, Cambridge.
Goldthorpe, J.H., 1980/1987. Social Mobility and Class Structure in Modern Britain. Clarendon Press, Oxford.
Goldthorpe, J.H., 1984. Order and Conflict in Contemporary Capitalism: Studies in the Political Economy of Western European Nations. Clarendon Press, Oxford.
Goldthorpe, J.H., 1990. A response. In: Clark, J., Modgil, C., Modgil, S. (Eds.), John H. Goldthorpe: Consensus and Controversy. The Falmer Press, London.
Goldthorpe, J.H., 1992. The theory of industrialism and the Irish case. In: Goldthorpe, J.H., Whelan, C.T. (Eds.), The Development of Industrial Society in Ireland. Oxford University Press, Oxford.
Goldthorpe, J.H., 1996. The quantitative analysis of large scale data-sets and rational action theory: for a sociological Alliance. European Sociological Review 12 (2), 109.
Goldthorpe, J.H., 2001. Causation, statistics and sociology. European Sociological Review 17 (1), 1–20.
Goldthorpe, J.,H., 2002. Occupational sociology, yes: class analysis, no: comment on Grusky and Weeden's research agenda. Acta Sociologica 45 (3), 211–217.
Goldthorpe, J.H., 2004. Sociology as social science and cameral sociology: some further thoughts. European Sociological Review 20 (2), 97–105.
Goldthorpe, J.H., 2006a. On Sociology Volume 1: Critique and Program. Stanford University Press, Stanford.
Goldthorpe, J.H., 2006b. On Sociology Volume 2: Illustration. Stanford University Press, Stanford.
Goldthorpe, J.H., 2007a. Cultural capital some critical observations. Sociologica 2, 1–22.

Goldthorpe, J.H., 2007b. Cultural capital: a response to the comments. Sociologica 2, 1–7.
Goldthorpe, J.H., 2008. Two oppositions in studies of class: a reflection. In: Lareau, A., Conley, D. (Eds.), Social Class How Does it Work. Russell Sage Foundation, New York.
Goldthorpe, J.H., 2010. Analysing social inequality: a critique of two recent contributions from economics and epidemiology. European Sociological Review 26 (6), 731–744.
Goldthorpe, J.H., 2012. Understanding – and Misunderstanding – Social Mobility: The Entry of Economists, the Confusion of Politicians and the Limits of Educational Policy. Department of Social Policy and Social Intervention, University of Oxford. Barnett Papers in Social Research, 2012/1.
Goldthorpe, J.H., Jackson, M., 2007. Intergenerational mobility in contemporary Britain; political concerns and empirical findings. British Journal of Sociology 58.4, 525–546.
Goldthorpe, J.H., Jackson, M., 2008. Education based meritocracy: the barrier to its realization. In: Lareau, A., Conley, D. (Eds.), Social Class How Does it Work. Russell Sage Foundation, New York.
Goldthorpe, J.H., Lockwood, D., Bechhofer, F., Platt, J., 1968a. The Affluent Worker: Industrial Attitudes and Behaviour. Cambridge University Press, Cambridge.
Goldthorpe, J.H., Lockwood, D., Bechhofer, F., Platt, J., 1968b. The Affluent Worker in the Class Structure. Cambridge University Press, Cambridge.
Goldthorpe, J.H., Marshall, G., 1992. The promising future of class analysis: a response to recent critiques. Sociolog 26, 381–400.
Goldthorpe, J.H., Hope, K., 1974. The Social Grading of Occupations. Oxford Clarendon Press.
Jonsson, J.O., Grusky, D.B., Di Carlo, M., Pollak, R., Brinton, M.C., 2009. Microclass mobility: social reproduction in four countries. American Journal of Sociology 114 (4), 977–1036.
McRae, S., 1990. Women and class analysis. In: Clark, J., Modgil, C., Modgil, S. (Eds.), John H. Goldthorpe: Consensus and Controversy. The Falmer Press, London.
Merton, R.K., 1959. Notes on problem-finding in sociology. In: Merton, R.K., Broom, L., Cotterll, L.S. (Eds.), Sociology Today: Problems and Prospects. Harper and Row, New York.
Wilkinson, R., Pickett, K., 2009. The Sprit Level: Why More Equal Societies Almost Always Do Better. Allen Lane, London.

Gorz, André (1923–2007)

Claus Leggewie, Kulturwissenschaftliches Institut Essen (KWI), Essen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

André Gorz, who was born in Vienna but adopted France, was a French public intellectual and philosopher. Heavily influenced by the alienation brought on by National Socialism experienced during his younger years, his perspective is described as that of an existential Marxism influenced by Sartre. An advocate of a political ecology that combined critique of capitalism with environmental concerns and the importance of liberty, he was a leading philosopher in the French *Nouvelle Gauche* (New Left). Known as a philosopher under the pen name Gérard Horst, he published as a journalist under the pen name Michel Bosquet.

Life

André Gorz (pen names Gérard Horst and Michel Bosquet) was born in Vienna on 9 February 1923 and named *Gerhart Hirsch*, was the son of a Jewish timber trader Jakob Hirsch and a Catholic secretary from Dresden, Maria Hirsch. Owing to the pronounced anti-Semitism in Austria, Gorz's father converted to the Catholic faith in 1930 and the family assumed the 'Aryan' sounding surname (Robert) Horst. As a school pupil, Gorz, who had not experienced any Jewish religious or cultural instruction, was at first attracted to the 'vital values' that ironically National Socialism seemed to him, from a distance, to embody. The central motif of Gorz's life and work, namely the issues of doubt and change surrounding identity, had its roots in this early experience: "I began to lose my identity when one of my dear school buddies showed me a poster on which there was an illustration of a Jew with locks of hair hanging from the temples, a crooked nose, dribbly lips and hook-like fingers and said: that is your father! From then on I knew I was not a normal person and that, in the eyes of others, I was neither an equal nor of equal worth" (Leggewie, 1989: 17). The experience of having a pariah identity as a second-class Jewish person forced upon him by the majority society was one that Gorz shared with many of Europe's assimilated Jews.

Encouraged by his mother, who wished her son to be spared the ordeal of military service, Gorz attended 'Montana' boarding school in the Swiss canton of Graubünden one year after the annexation of Austria into the German Reich in 1938. Here, he assumed the name Gérard Horst. Upon graduation, he commenced the study of chemistry at the engineering college in Lausanne and began to write philosophical essays, stories, and political articles for the newspaper of a Swiss cooperative. Thus, he began his career as a freelance author, and the onset of the feeling of existential emptiness and political homelessness that caused Gorz to seek refuge in *république des lettres* (the *Republic of Letters*, i.e., a long-distance intellectual community).

In politically neutral Switzerland, Gorz experienced the scarcely concealed complicity with the fascistic *weltanschauung* of Hitler's Germany. As France was overrun by the Wehrmacht's *Blitzkrieg* ('*drôle de guerre*') and occupied in the northwestern part, he made this 'emptiness' the anchor of his personal and political identity – an existentialist design for an 'other,' imagined identity: "[…] when I learnt in 1940 that France had collapsed under the triumphant progress of the *Wehrmacht* and the 'master race,' I had the feeling that, yes, these are the people to whom I am closest. France is sinking into nothingness and it is in this nothingness that I shall declare my faith […]" (Leggewie, 1989: 17). With the political obliteration of the French Republic, which had collaborated with the Germans in the form of the fascistic Vichy regime, Gorz chose literary and philosophical France as his spiritual home for the period until 1946, during which he remained physically in the Swiss confederation: that is, neither the United States, which lent itself as a site of exile for many others who experienced a similar destiny, nor Stalin's Soviet Union (both of which would anyway have been, on a practical level, out of reach for the young Gorz).

Gorz's pattern of socialization during the interwar period was (not so) extraordinary: the Jew and secret admirer of 'vital' National Socialism in anti-Semitic Austria was brought up by a governess in the French language, became a half-Jew in the extended German *Reich*, then a Jew with a German passport in a Switzerland that had complied to the *Reich* and, finally, converted to a geographically displaced but biographically and linguistically rooted Frenchman. "France has never been a nation in the German sense of the national, that is, the reason that one turns to French culture as a foreigner has not to do with thoughts concerning the national but with the universality of French culture" (Leggewie, 1989: 20). This virtual naturalization was fleshed out through Jean-Paul Sartre, who during the occupation had begun his ascent toward becoming the country's leading intellectual. Gorz devoured Sartre's novels and works for theater, and also personally met with him during Sartre's lecture series in Switzerland.

It was also in Switzerland that Gorz met his wife, Doreen, who was a Scot by birth and with whom he would spend the rest of his life. The couple departed for Paris in 1949, where Gorz commenced his profession as a freelance writer in the proverbial attic digs. He earned his living as a secretary to the 'Citizens of the World,' a cosmopolitan society initiated by the American Gary Davis, a peace activist who created the *World Passport*, the first travel document originally based on world citizenship, and as a press agent at the Indian embassy, before becoming active as a journalist. It was upon joining the *Paris-Presse* newspaper that both his pseudonyms came into being under which the author and philosopher Gérard Horst would become renowned worldwide. While the journalistic alias Michel Bosquet was one possible translation of 'Horst' (small

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61207-2

wood) into French, the philosophical *alias* author of the first great work, *Le Traître* (1957), stemmed from a childhood memory of his father's military telescope, manufactured in the then Austrian town of Görz. This central European border town, located on the border between Austria, Italy, and Yugoslavia, was of particular interest as concerns the three names André, Michel, and Gérard, since, after 1919, the town assumed three linguistic identities: Italian Gorizia, Slowenian Gorica, and Friulian Gurize. Bosquet was now the name of the journalist, Gorz that of the philosopher, and both answered to the name in the passport of Gérard Horst.

After this elaborate self-construction, due to the complicated course of his time, the remainder of André Gorz's life story is cursorily outlined in the following paragraphs.

In 1954, Gorz, stateless until then, became a French citizen – apparently thanks to the support of the French politician Pierre Mendès (Contat, 2009). The following year, 1955, Bosquet took up the editorship of the left-liberal weekly magazine L'Express, following which, in 1964, he joined the editorial team of France Observateur, which was still more left-leaning, and later became Le Nouvel Observateur – of which Gorz was acting editor-in-chief until 1983. Following the publication of his philosophical novel *Le Traître*, with a forward by Jean-Paul Sartre, the philosopher Gorz in 1961 joined the editorial committee of the monthly magazine Les Temps Modernes, where he remained until 1974.

The 1960s saw the generally introvert Gorz become a 'public intellectual' and a leading intellectual force of the New Left. Apart from numerous articles in the aforementioned journals (and many marginal leftist reviews), his status was above all established by the books, of which there were many reprints as well as translations into numerous languages: S*tratégie ouvrière et néocapitalisme* (Editions du Seuil, 1964), *Le Socialisme difficile* (Seuil, 1967), and *Réforme et Révolution* (Seuil, 1969) (see Part 2). Subsequent to this phase, during which an alternative to the Communist Party presented itself (Gorz sympathized with the ex-Catholic, increasingly left-leaning, moderately anarcho-syndicalist CFDT (*Confédération française démocratique du travail*) and the left breakaway party PSU (Parti Socialiste Unifié)), there came works that were just as influential in a field to which Gorz contributed much: political ecology. These works are *Critique du capitalisme quotidien* (Editions Galilée, 1973), *Critique de la division du travail* (Seuil, 1973), *Écologie et Politique* (Galilée, 1975), and *Écologie et Liberté* (Galilée, 1977). Neither should the role of Bosquet in the founding of the ecology magazine *Le Sauvage* (1973ff.) be forgotten. As regards German-language publications, Bosquet also provided the idea for the series *Technologie und Politik* under the general editorship of the Hamburg publicist Freimut Duve and published by Rowohlt Verlag.

In the 1980s, Gorz became a reluctant supporter of the Union de la Gauche, which, under Francois Mitterrand, had found its way into power. At the same time, unlike a series of so-called *nouveaux philosophes*, he did not turn away from Marxism. In the publications that then appeared, beginning with the bestseller *Adieux au prolétariat* (Galilée, 1980; Le Seuil, 1983), followed by *Les Chemins du Paradis* (Galilée, 1983), the work of founding theory *Métamorphoses du travail* (Galilée, 1988), and the scattered pieces collected in *Capitalisme Socialisme Écologie* (Galilée, 1991), Gorz presented an original synthesis that drew on an empirically grounded critique of capitalism, on a Marxism enriched by humanist elements and existentialist philosophy, and on political ecology. During the last two decades of his life, Gorz left the *Nouvel Observateur*'s editorial team and moved to the countryside, where he concerned himself intensively with bioethical questions of life and death, and with the changes that the Internet had prompted in society.

The book by André Gorz that sold most and captured the most attention was the slim but striking volume *Lettre à D. Histoire d'un amour* (Galilée, 2006; Gallimard, 2008), an ode to his wife Dorine concerned with the themes of his earlier works. It begins with the sentence: "Tu vas avoir quatre-vingt-deux ans. Tu as rapetissé de six centimètres, tu ne pèses que quarante-cinq kilos et tu es toujours belle, gracieuse et désirable. Cela fait cinquante-huit ans que nous vivons ensemble et je t'aime plus que jamais" ("You'll be 82 years old. You have shrunk by 6 cm, you weigh but 45 kilos and you are always beautiful, graceful and desirable. It has been 58 years that we've lived together and I love you more than ever."), and ends with the prophetic sentence: "Récemment, je suis retombé amoureux de toi une nouvelle fois et je porte de nouveau en moi un vide débordant que ne comble que ton corps serré contre le mien […] Nous aimerions chacun ne pas survivre à la mort de l'autre. Nous nous sommes souvent dit que si, par impossible, nous avions une seconde vie, nous voudrions la passer ensemble." ("Recently, I fell in love with you again and I carry inside me an overflowing void that only your body held tightly against mine fills […] We each would like not to survive the death of the other. We often said that if, in the unlikely event, we had a second life, we would like to spend it together.")

The couple ended their 'first lives' by committing suicide together on 22 November 2007. There is no gravesite devoted to them; their earthly remains were scattered to the wind.

André Gorz's papers can be found in the Institut Mémoires de l'Édition Contemporaine (IMEC) in Caen (http://www.imec-archives.com/fonds/gorz-andre/).

Works

André Gorz left behind an extensive, thematically diverse and highly influential body of work that is of astonishing relevance to many significant questions in the current debate on social theory and sociopolitical issues. It comprises voluminous philosophical works as well as short, sharp political commentaries on contemporary questions. Although both of the author's names are no longer in circulation among the majority of today's young, his influence extends today far beyond the milieu of the 'New Left' within which his writings first began to be discussed.

It is hardly possible to do justice to Gorz's *oeuvre* within such a limited space. One can however point to some of the stages in the development of his thought and identify some of the themes that, in various nuanced ways, reoccur throughout his work and to which, the preoccupation with the autonomy and freedom of the individual must be seen as central, and which Gorz considered massively threatened by the structures surrounding capitalist economy, technological development, and technocratic systems. In the 1960s, Gorz concerned himself

with the strategic options available to the workers' movement and thus distanced himself from the dogmatic positions of those hailing from the political parties of the Left, whose programs and politics he called into question – particularly where the French Union of the Left was concerned. In the 1970s, Gorz became a leading intellectual force in terms of political ecology. Thereafter, he concerned himself with the transformation of the capitalist economy and society wrought in the face of digitalization, biopolitics, and the knowledge economy, and critiqued the inhumane dimensions of this transformation, while also drawing attention to its potential outcome in a freer society.

Although most themes and points of emphasis extend throughout the body of his work, one can divide them roughly into four phases: (1) existentialist Marxism (*Traître*, *Fondements*, *Lettre à D*); (2) New Left and self-organization in neo-capitalism (from *Stratégie* to *Adieux au prolétariat*); (3) political ecology (*Écologie et politique*/*Écologie et liberté*); and (4) immaterial economy (*L'Immateriel*). Marxist sources and influences always were and remained significant; there was no departure from Marxism as described by many castigators. Continuously, Gorz read Marx's texts without ever having become 'Marxist' in any dogmatic sense (the sense that Marx himself had mocked in those who succumbed to it):

1. The influence of Jean-Paul Sartre too remains recognizable in all phases of Gorz's life and work. At the same time, Gorz's philosophical work (*Le Traître*, *Fondements*, *Critique*) did not achieve the consistency of the role models with whom he had been so impressed and whom he studied so meticulously to begin with; as such, Gorz was neither a 'Sartrean' nor an 'existentialist.' Nonetheless, motifs from Sartre's existentialist philosophy remain significant and, above all, that foundational existential motif, the treatment of nothingness, the nothingness that can appear in the devastating gaze of the other, and the resulting, lifelong sense of homelessness; then the denial of a human nature that is reified in tradition and morality, collectives and conventions, and the resulting emancipation of 'objective compulsions,' the possibility of choosing and freedom to make decisions, the emphasis on freedom; lastly, the precarity of the subject and his tendency to flee in the face of his own freedom (see Häger and Meretz, 2012).
2. Gorz was never convinced of the capitalist economy's capacity of inclusion, nor of its productivity. For him, the invisible hand of the market ruled at the expense of the human workforce, freedom, and creativity; markets were no replacement for the organizing hand that held society together. In order to be able to analyze the development, logic, contradictions, and crisis-prone nature of capitalist production and reproduction, Gorz reached for Marx's works, the three volumes of *Capital* among them. But it was the early Marx that most interested him and who dealt with the alienation implicit in capitalism. Gorz did not so much seek anti- asnoncapitalistic ways out – 'ways to the paradise' that lie beyond the dictates and the primacy of the economy (which Marxist–Leninism and actually existing socialism have never overcome).

In this sense – how liberation can succeed and socialism can be practiced (Gorz later preferred the concept of 'conviviality,' which he took from Ivan Illich since 1968) – he emerged as the speaker of the New Left in the May 1968 revolt, which was on the one hand closer to the concrete *forces vives* of dissent and innovation than the old leftist parties of socialists and communists and, on the other, overcame the collective ideology of the monolithic proletariat and recognized in the diversity of social movements the motor of a 'revolutionary reformism.'

A large portion of work dealing with Gorz's reception is concerned with the question, which in the last analysis is irrelevant, as to whether Gorz interpreted Marx 'correctly,' whether or not he was an orthodox or dissenting Marxist, or whether in a given work a revisionist turn had taken place or not. Gorz had never located himself in this Marxologic context; he did not think in terms of dogma but of problems and challenges that the times in which he lived presented. Added to which, the question, already elaborated in Marx but ignored in actually existing socialism just as it was in capitalism, of the relations between humans and nature in capitalism also featured.

3. Building on parallel patterns of thought that were critical of civilization and had accompanied industrialization from the beginning, the 'ecological question' posed itself in the 1970s with a renewed sharpness. It was by no means down to chance that answers, which may retain their validity for a comprehensive analysis conducted in the context of the cultural sciences, came from an interpreter of modern times closely associated with the existentialist philosophy of Jean-Paul Sartre. In contrast to Sartre, who was at the time in the process of becoming a literary Maoist, André Gorz, for the most part under the *nom de plume* Michel Bosquet and as acting editor-in-chief of the left weekly magazine *Le Nouvel Observateur*, turned his attention in a series of striking essays to a critical revision of Marxism (Gorz, 1980a,b,c) and the relation between ecology and politics. Gorz went beyond the environmental management that had also emerged in the 1970s and become institutionalized in policy and administration, and provided a postindustrial and post-bureaucratic basis for political ecology. Here, there are affinities with his contemporary Alain Touraine, the theorist, and participant observer of new social movements.

Collected in the paperbacks *Écologie et Politique* (1975) and *Écologie et Liberté* (1977a), both widely available in France, Germany, and other European countries, are the author's contributions to *Nouvel Observateur* and the environment magazines *Sauvage* and *Lumière et vie* during the years 1973–77. Here, the 1973 'oil crisis' as well as the critique of growth following the report of Club of Rome provided the background, along with the efforts of French and European socialists and 'eurocommunists' to create a party-political Union of the Left. The little great book was at the time read above all as a strategy paper of the environmental movement that emerged at the end of the 1970s and as a 'leftist' basis for the alternative green parties, which were at the time very heterogeneous in terms of the support, ideology, and milieus upon which they drew and which in France and Germany experienced their first success at the polls (Leggewie and Miller, 1978).

Gorz disputed the basis and the potential of every kind of politics of environmental protection that was not at the same time a guarantor of freedom for such a politics only served to perpetuate the "development of the industrial technosphere […] without compromising the self-regenerating capacities of the ecosphere. This approach does not involve a fundamental break with industrialism and its hegemony of instrumental reason"; bans, administrative regulations, taxation, subventions, and fines only lead to the regulation of society from the outside. This approach makes eco-compatibility seem obtainable "without any need for change in the mentalities, values, motivations or economic interests of the social actors." (The quotes are referred to: November–December 1993. "Political Ecology: Expertocracy versus Self-Limitation" (John Howe, Trans.). New Left Review, pp. 55–67. http://newleftreview.org/I/202/andre-gorz-political-ecology-expertocracy-versus-self-limitation.)

If these should change, and Gorz believes that they should, they would not do so as a result of some kind of authoritarian measures taken by the state, as might occur under 'green' expert rule. Rather, the protection of the lifeworld would be defined through the realization that "the result of activities corresponds to the intentions that gave rise to them; in other words, that social individuals see, understand and accept the result as the outcome of *their* actions." (emphasis CL). Just like megatechnologies (of which the nuclear state developed in France in the 1960s is exemplary), a perfected environmental management would also further the technification and colonization of the lifeworld to which the ecological movement had specifically objected. 'The system invades and marginalizes the lifeworld", in Gorz's formulation (see above), which draws upon the work of Ivan Illich and the terminology of Jürgen Habermas, as also on older traditions of *Lebensphilosophie* (philosophy of life). Thus, he describes "the world accessible to intuitive understanding, to practical and sensory assimilation. [The system] deprives individuals of the chance to have a world, to share it with one another. It was against the different forms of this expropriation that resistance gradually came into being."

Gorz thought the ecological movement was capable of overcoming the same old divisions and conflicts within industrial society that had made the 'old Left' (in France, in the form of the 'Union of the Left' on its way to taking power) the advocate of the dominant industrialism and a follower of the 'religion of growth.' "So ecologism could become a political movement after all. Defence of the lifeworld was not just a sectional, local aspiration lacking general validity, but something in tune with the interests of humanity in general and the living world as a whole" (see preceding paragraphs). For Gorz, as a natural scientist by training with a liking for engineers and statisticians, it was this that constituted the decisive advantage, in democratic and political terms, of political ecology, namely "the problem of retroactive coupling of necessity with normativeness or, if you prefer, the transformation of objective necessities into normative behaviour corresponding to needs that have been experienced, and that influence the form in which the objective necessities are perceived." This social contract for sustainable development led him, with reference to the Marxist labor theory of value and criticism of capitalism, to formulate the 'norm of sufficiency' – which he now emphasized above all as an *ecosocialist* – to which the economic rationality of capitalism stood in opposition. In arguing his case, Gorz used the example of 'the societal ideology of the motorcar.' At the core of an ecosocial argumentation: "to restore politically the correlation between less work and less consumption on the one hand, and more autonomy and more existential security on the other, for everyone."

Gorz and other social theorists further developed this approach through conceptual and practical consideration of a guaranteed basic income, in accordance with which, the connection between labor (time) and wage income would be terminated and the resulting disposable time become 'the true measure of wealth' (Marx, *Grundrisse*). "This boils down to saying that economic activity has meaning only if it serves something other than itself."

4. Gorz's understanding of the ecosocial question always went hand in hand with a radical critique of the hostility of capitalism to 'life,' which he then, however, applied 'constructively' as well in *Paths to Paradise* and the collection of essays *Und jetzt wohin?* ('And now where?'), namely with regard to the opportunities that a digitally supported and networked knowledge society offered (*L'Immatériel*); these opportunities were only achievable through a radical reduction in formal labor time and the provision of a guaranteed basic income, that is, through terminating the connection between labor and income in order to enable the people's creative potential to develop. To which he perceived modern biotechnology and gene technology as being completely opposed. At the end of his life Gorz came back to a radical critique of capitalism without ignoring the liberating elements of knowledge production in 'digital capitalism.'

Influence and Impact

André Gorz did not belong to any school of thought, nor did he actively encourage followers. He was an exemplary maverick among the intellectuals of the New Left. And he maintained an intellectual distance from the activists of May 1968 and the global student revolt too. He had never been an insider in the circle surrounding Sartre and de Beauvoir, and the vanity of the Paris star-intelligentsia made him feel acutely uneasy; he felt like a stranger among the staff of *Le Nouvel Observateur*, the magazine he himself had helped found. Gorz appeared only with considerable reluctance at public events, avoided electronic media, and fled the nervosity and blasé attitude of the metropolitan city. He was more inclined toward face-to-face conversation and the sharing of ideas within a small circle, and had a preference for publishing in lesser-known journals.

Despite his personal disposition as a maverick and a literary cultivated manner of self-presentation as an outsider, his books of the 1960s had a considerable influence on the theoretical formulations of the *Nouvelle Gauche* or New Left globally and especially on those 'green' social movements emerging from both the New Left and other alternative currents. In this respect, it was the trained natural scientist's

empirical knowledge of natural phenomena that made him stand out and eschew the romanticization of the 'environment' and its discussion in absolute terms. As in the case of another of the period's great solitary figures, Ivan Illich, central to Gorz's political ecology was the safeguarding of individual freedom and collective autonomy against the politico-economic and techno-industrial accumulation of power that was, for Gorz, embodied in the idealized 'nuclear state' (Robert Jungk).

It was this essentially antiauthoritarian, anti-institutional, and libertarian mode of thought that gave rise during the 1980s to the distance he maintained from and his criticism of the bureaucratic industrial project of the socialists and communists in the Union of the Left, and to his rejection of a dogmatic, ossified Marxism. Politically, Gorz favored the independent candidates for president – including in opinion pieces in the *Nouvel Observateur* – the agronomist René Dumont (1974) and the head of the environmental association *Amis de la Terre* Brice Lalonde (1981). Gorz did not envisage a state led by the Left that administered redistribution from above in the industrial labor society, but a qualitative change that came from within the labor society itself, through a change in the way of life that would render a postindustrial society, which would not overexploit its natural resources and exceed its physical limits.

French was Gorz's chosen language, and it was thus self-evident that he addressed, in the first instance, French readers. However, it was Michel Bosquet in particular who was listened to and read, the sharp-tongued and aggressive interpreter of his time, rather than the social and moral philosopher André Gorz, who attempted to link existential philosophy with a critique of political economy and to broaden his outlook to take into consideration the latest developments in the postindustrial society, including the knowledge commons that digital media in their capacity as sources of information and a means of communication seemed to make possible. As a political philosopher he fell between two chairs as far as the French parties of the Left were concerned, which misunderstood him by dismissing him either as a Marxist dogmatist or a post-Marxist renegade. He remained closest to the left-socialist CFDT trade union confederation and the refined PSU, which offered a home to dissenters from the workers' parties, as well as to anticolonialists and independent intellectuals of the Left; they were, to begin with, led by Gorz's special supporter Pierre Mendès-France, and from 1968 by Michel Rocard, and propagated a self-organizing socialism that came very close to Gorz's own ideas. (It must be from this affinity that the common categorization of Gorz as a 'Trotskyist' stems.) The French Greens, who could recognize in him a genuine mentor, misunderstood him to a similar extent and could not use his ideas in their programs; within France's political system they emerged as something of a lost cause, despite having been founded earlier and originally being more powerful than green movements in Germany and other European countries.

See also: Alienation, Sociology of; Field Theory; Materiality and Culture; Personal Identity – Philosophical Aspects.

Bibliography

Bowring, F., 2000. Andre Gorz and the Sartrean Legacy. Palgrave Macmillan, London.

Contat, M., 2009. André Gorz. Vers la société libérée. Les éditions Textuel [=La voix au chapitre], Paris.

Gorz, A., 1957. Le Traître (with a preface by Jean-Paul Sartre). Le Seuil, Paris.

Gorz, A., 2008a, new ed.. Der Verräter. With the essay "Über das Altern". Rotpunktverlag, Zürich.

Gorz, A., 2010a. Ecologica. Translated from the French by Chris Turner. Seagull Books, London.

Gorz, A., 2005, rev. ed.. Le Traître (with a preface by Jean-Paul Sartre) [=Folio Essais].

Gorz, A., 1958/1997. Der Verräter. With the essay "Über das Altern" (first published 1958; with the essay 1961/62, translated from the French by Eva Moldenhauer.). Suhrkamp, Frankfurt am Main.

Gorz, A., 1959/1989. The Traitor. Translated from the French by Richard Howard. Simon and Schuster, New York (second ed. revised by Gorz. Verso, London).

Gorz, A., 1959. La Morale de l'histoire. Le Seuil, Paris.

Gorz, A., 1964. Stratégie ouvrière et néocapitalisme. Le Seuil, Paris.

Gorz, A., 1967a. Zur Strategie der Arbeiterbewegung im Neokapitalismus. Europäische Verlagsanstalt, Frankfurt am Main.

Gorz, A., 1967b. Strategy for Labor: A Radical Proposal. Translated from the French by Martin A. Nicolaus and Victoria Ortiz. Beacon Press, Boston.

Gorz, A., 1967c. Le Socialisme difficile. Le Seuil, Paris.

Gorz, A., 1975a. Socialism and Revolution (Norman Denny, Trans.). Allen Lane, London.

Gorz, A., 1969. Réforme et Révolution. Le Seuil, Paris.

Gorz, A., 1970. Die Aktualität der Revolution. Nachtrag zur "Strategie der Arbeiterbewegung im Neokapitalismus". Europäische Verlagsanstalt, Frankfurt am Main.

Gorz, A., 1973a. Critique du capitalisme quotidien. Galilée, Paris.

Gorz, A., 1973b. Critique de la division du travail. Le Seuil, Paris (Ouvrage collectif).

Gorz, A., 1982a. Kritik der Arbeitsteilung. Fischer Taschenbuch, Frankfurt am Main.

Gorz, A., 1975b. Écologie et Politique. Galilée, Paris.

Gorz, A., 1978, rev. ed.. Écologie et Politique. with the additional text Écologie et Liberté from 1977. Le Seuil, Paris [Éditions Points].

Gorz, A., 1980a. Ecology as Politics. Translated from the French by Jonathan Cloud and Patsy Vigderman. Pluto Press, London.

Gorz, A., 1977/1980. Ökologie und Politik. Beiträge zur Wachstumskrise. Rowohlt, Reinbek.

Gorz, A., 1977a. Écologie et Liberté. Galilée, Paris.

Gorz, A., 1977b. Fondements pour une morale. Galilée, Paris.

Gorz, A., 1980b. Adieux au proletariat. Galilée, Paris (éd. augmentée Le Seuil, 1983).

Gorz, A., 1980c. Abschied vom Proletariat: jenseits des Sozialismus (Translated from the French by Heinz Abosch). Europäische Verlagsanstalt, Frankfurt am Main.

Gorz, A., 1982b. Farewell to the Working Class: An Essay on Post-industrial Socialism (Michael Sonenscher, Trans.). Pluto Press, London.

Gorz, A., 1983a. Les Chemins du Paradis. Galilée, Paris.

Gorz, A., 1983b. Wege ins Paradies: Thesen zur Krise, Automation und Zukunft der Arbeit. Translated from the French by Eva Moldenhauer. Rotbuch Verlag, Berlin.

Gorz, A., 1985. Paths to Paradise: On the Liberation from Work (Malcolm Imrie, Trans.). Pluto Press, London.

Gorz, A., 1988/2004. Métamorphoses du travail. Galilée/Paris: Folio Essais, Paris.

Gorz, A., 1989/2009. Kritik der ökonomischen Vernunft: Sinnfragen am Ende der Arbeitsgesellschaft. Translated from the French by Otto Kallscheuer. Rotbuch, Berlin. new ed. Rotpunktverlag, Zürich.

Gorz, A., 1989. Critique of Economic Reason (Gillian Handyside, Chris Turner, Trans.). Verso, London.

Gorz, A., 1991. Capitalisme Socialisme Écologie. Galilée, Paris.

Gorz, A., 1994. Capitalism, Socialism, Ecology (Chris Turner, Trans.). Verso, London.

Gorz, A., 1997. Misères du présent, richesse du possible. Galilée, Paris.

Gorz, A., 2003. L'Immatériel. Galilée, Paris.

Gorz, A., 2006/2008. Lettre à D: Histoire d'un amour. Galilée, Paris reed. Gallimard, Paris [=Folio].

Gorz, A., 2008b, seventh ed.. Brief an D. Geschichte einer Liebe. Translated from the French by Eva Moldenhauer. Rotpunktverlag, Zürich.

Gorz, A., 2008c. Ecologica. Galilée, Paris.

Gorz, A., 2009. Auswege aus dem Kapitalismus. Translated from the French by Eva Moldenhauer. Rotpunktverlag, Zürich.
Gorz, A., 1999. Arbeit zwischen Misere und Utopie. Translated from the French by Jadja Wolf. Suhrkamp, Frankfurt am Main.
Gorz, A., 2010b, fourth edition. Wissen, Wert und Kapital: Zur Kritik der Wissensökonomie. Translated from the French by Jadja Wolf. Rotpunktverlag, Zürich.
Häger A., Meretz, S., 2012. Fahnder nach "Auswegen aus dem Kapitalismus" – eine Erinnerung an André Gorzin: Berliner Debatte. http://www.streifzuege.org/2012/fahnder-nach-auswegen-aus-dem-kapitalismus#more-12721
Krämer, H.L., Leggewie, C., 1989. Wege ins Reich der Freiheit. Rotbuch, Berlin.
Lesourt, E., 2011. André Gorz. Portrait du philosophe en contrebandier, ou l'écologie politique comme reconquête du sujet. L'Harmattan [= La Librairie des Humanités], Paris.
Leggewie, C., de Miller, R., 1978. Der Wa(h)lfisch. Okologiebewegungen in Frankreich, Berlin.
Leggewie, C., 1989. "Die ich wählte, wiesen mich ab, die ich abwies, wählten mich..." Das Leben des André Gorz zwischen Österreich, Frankreich und Deutschland, in: Krämer/Leggewie, S.12–35.
Little, A., 1996. The Political Thought of Andre Gorz. Routledge Chapman & Hall, London/New York.
Münster, A., 2011. André Gorz oder: Der Schwierige Sozialismus. Rotpunktverlag, Zürich.
Tatman, J., 1997. Andre Gorz: A Critical Introduction. Pluto Press, London.

Governance

Gunnar F Schuppert, Social Science Research Center Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

What accounts for the pervasive use of the term 'governance' and to what does it refer? This article aims to answer these questions by first examining the scientific roots of the much vaunted, but often contested concept. To clarify its actual meaning, the cornerstones of an emerging governance theory are then discussed, followed by an inquiry into how governance can be viewed in the context of changing statehood. The article ends with a review of how this concept relates to broader contemporary changes in statehood, transnational decision-making, and the architecture of modern governance.

Introduction

"The word 'governance' is ubiquitous. The World Bank and the International Monetary Fund make loans conditional on 'good governance.' Climate change and avian flu appear as issues of 'global governance.' The European Union issues a White Paper on 'governance' […]. What accounts for the pervasive use of the term 'governance' and to what does it refer? Current scholarship offers a bewildering set of answers." These remarks sound like and are indeed the classic opening of an introductory chapter in the 'Handbook of Governance' (Bevir, 2011: p. 1). Considering this opening, the aim of this article is quite obvious: it seeks to clarify the concept of governance and to offer some guidance and orientations for all those who have and want to move on the slippery ground of the governance concept. To achieve this, the article proceeds in three steps: We will start with some theoretical considerations, which will help us explain what is 'new' about governance and why the governance concept is so tremendously successful. At this point, it should be hopefully clear what governance means and we can therefore continue to look at governance in the policy world, presenting three examples for 'governance at work.' Finally, we conclude this overview by discussing the strengths and weaknesses of the governance concept.

Before we start it should be made clear that we talk about public governance, not about corporate governance. Corporate governance is a field of policy and research by its own, going back on the influential publications of Ronald Coase (1937) and Oliver Williamson (1985). In these studies, they pointed out that a decisive factor for economic success and effective transactions is to be found in the organizational structure of a corporation. Corporate governance (for an overview see Jürgens, 2006) encompasses a great variety of topics: from the basic aims of private firms as market actors including their role in society (Corporate Social Responsibility), to rules of transparency in the field of financial reporting, incentive structures for managers and the structures of supervisory boards. The importance of this branch of governance is mirrored in the numerous 'Corporate Governance Codices' as guiding rules for 'good corporate behavior.' As the focus of this article is governance in the public sector, we will not go deeper into this field of corporate governance.

Some Theoretical Considerations

At the beginning of the theoretical part, the term governance should be defined. Zürn et al. (2010: p. 2) suggest a generic definition of governance that denotes "the sum of rules and regulations […] processes as well as structures, justified with reference to a public problem." In greater detail, Michael Zürn defines governance as "[…] all the collective rules which aim at a certain set of problems or a certain societal constellation and which are justified by referring to the collective interest of the affected group" (2008: p. 554).

But these definitions do not tell us, what is new about governance and what can be regarded as the added value of the governance concept. The best way to answer these questions is to contrast the governance concept with other more 'classic' concepts; we start with contrasting (**Table 1**).

The lesson we can draw from this table is, that governance is a nonstate-centered concept. One of the critical reasons for the tremendous career of the governance concept is that it overcomes the state-centered approach, which dominates many publications in governance-relevant disciplines like political science, law, and history. Furthermore, the stateless term governance seems to offer the opportunity to jettison the unloved triangle of state/hierarchy/command and control by a 'semantic coup': from state/government to governance.

And indeed we can notice the broad consensus that the governance concept must be understood as a nonstate-centered concept. Scott Burris et al. (2008: p. 2) have put this consensus as follows:

> There is fairly wide agreement that it no longer makes sense to conceive of the state as the monopolist of governance. Once it was dogma that our collective world was divided into two fundamentally different spheres: the public sphere – which was the realm of governance; and, the private sphere – the realm of the governed. This crucial distinction has eroded. States do not enjoy a monopoly on governance, and themselves are often governed by non-state actors.
>
> It now makes more sense to describe our world as polycentric, with multiple agencies and sites of governance who govern through a variety of forms of power and largely in their own interests with far reaching collective impacts. This is as true at the inter-national level (where the proliferation of non-state governing actors has been characterized as a, post-Westphalian regime) as it is at the local (where, for example, more public order services are provided by private security guards than by public police).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.75020-3

Table 1 Government and governance

	Government *State versus market/society*	*Governance* *State, market, and networks as complementary methods of steering*
Polity	● Focus on the state ● Majority rule and hierarchy as most important institutions	● Institutional structures combining elements of hierarchy, bargaining systems, and mechanisms of competitions ● Networks
Politics	● Competition for power between political parties and for influence by lobby groups ● Conflict solving by decisions of state agencies and implementation of state made decision	● Conflict between governing/guiding and governed/addressed actors ● Guidance and control in the context of institutionalized regulatory regimes ● Negotiations between state and nonstate actors ● Adjustment of institutional regulatory regimes
Policy	● Lawmaking (commands and bans) ● Distribution of public goods	● Independence (in networks and associations), compromises, deals ● Coproduction of public goods ● Institution policy (management of institutional change)

Source: Benz, A., 2004. Einleitung: Governance - Modebegriff oder nützliches sozialwissenschaftliches Konzept? In: Benz, Arthur (Ed. and Trans.), Governance – Regieren in komplexen Regelsystemen. VS Verlag, Wiesbaden, pp. 11–28.

In the same direction argues Christoph Zürcher looking from the outside at the problems of governance in areas of limited statehood with the following really thought-provoking argument: "Consequently, I depart from the notion that statehood is provided solely by the state. Instead, I suggest we think of statehood as a product which is produced by the state in association with other actors. There are examples abundant when state outsource – intentionally or not – the provision of basic functions to external actors. It is sufficient to think of who provides security in Afghanistan or Tajikistan, domestic authority in Kosovo or Bosnia, or public services in Mozambique or Burundi" (Zürcher, 2007: p. 11).

The telling slogan of Burris et al. (2008: p. 15) "The king is dead – long live the extended Royal Family" seems to neatly sum up these findings of an eroding state monopoly as a governance actor and the simultaneously growing importance of nonstate actors.

A second contrast we want to introduce is the contrast between 'traditional governing' and 'modern governance' as Werner Jann (2011) has put it (**Table 2**):

Table 2 'Traditional governing' and 'modern governance'

	'Traditional governing'	*'Modern governance'*
Citizen	● Voter ● Consumer ● Client	● Active citizen and coproducer of public goods ● Partner of social cooperation and political agreements (empowerment)
Society	● Addressees of governmental steering ● Pluralistic representation of interests ● Corporatistic bargaining systems	● Societal self-regulation ● Active civil society

The lesson to learn from this table is that the relationship between government and society is in both cases significantly different and that the role of citizens changes from a passive addressee of policies to an active partner in governance. Stressing the important role of societal self-regulation means to take notice of the regulatory power of corporate societal actors such as professions, business associations, trade unions and so forth; and referring to the significant part of an active civil society as a basic element of modern governance means to recognize the contribution of different types of NGOs to what is called modern governance. So government and governance can not only be distinguished in terms of state-centeredness, of modes of guidance and control, but also in terms of different sets of actors and different roles of citizens.

The third and here last lesson to learn is that governance is about coordination. One of the main characteristics of the governance concept is its basic assumption that the reality of modern statehood is marked by the plurality of actors: a variety of state actors – the state is not and has never been a monolithic body – nonstate actors – they may be genuine private actors or institutions of the third sector/civil society – and – last but not least – hybrid actors with roots in the private and public sector. To coordinate the actions of these various actors and their specific contributions to the public interest is therefore one of the most important tasks of the modern pluralist state.

The modes of coordination may vary considerably and reach from hierarchical regulation to regulated self-regulation or even to the adoption of private regulations by state laws (for more details see Schuppert, 2011b). To become aware of this wide variety of choices, it is useful to imagine them on a scale with hierarchy and self-regulation as its cornerstones (**Figure 1**):

In nearly every case, these modes of coordination are to a certain degree institutionalized and therefore supported by and embedded in a regulatory structure, which simultaneously enables and restricts cooperative action. All forms of a closer cooperation and partnerships like public–private partnerships (PPPs), privatizations, delegation by contract and so forth, have to follow rules, not necessarily state laws like company law. However, these rules must be a result of negotiation or agreed rules of procedure monitored by the parties involved. Governance as coordination is insofar unavoidably governance as rule making.

Governance in the World of Politics and Policies

Good Governance and Global Governance by Governance Indicators

On the international stage most successful branch of governance is 'good governance' (Nuscheler, 2009). There is what might be

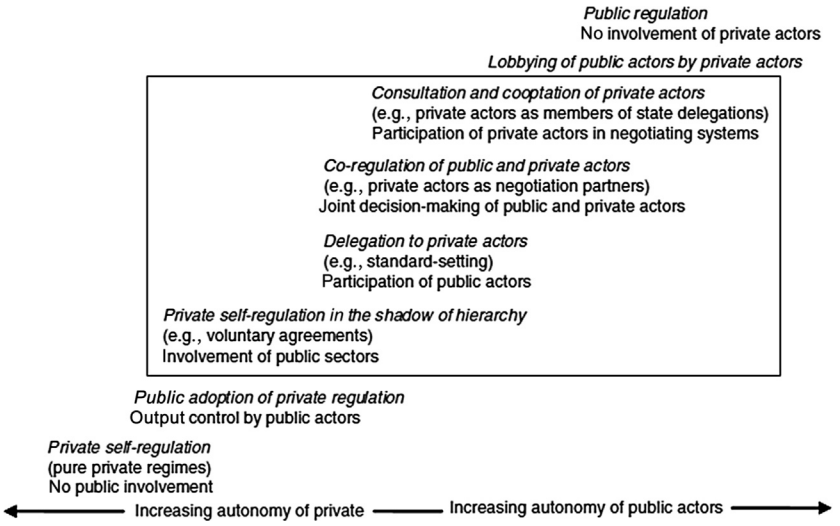

Figure 1 Source: Börzel, T.A., Risse, T., 2005. Public–private partnership. Effective and legitimate tools of transnational governance? In: Grande, E., Pauli, L.W. (Eds.), Complex Sovereignty: Reconstitution Political Authority in the Twenty-First Century. University of Toronto Press, Toronto, pp. 195–216.

called a 'good governance promotion industry,' headed by the World Bank and influential think tanks like Freedom House and others, which function as transfer agents throughout the world. Their powerful instrument is the use of governance indicators; the best known are the World Wide Governance Indicators developed for the World Bank by Kaufmann et al. (2010). These indicators are rooted in the values of the Western world and oriented toward the relationship between good governance and economic performance (see also the recently published book by Acemoglu and Robinson, 2012). Considering the impressive career of the good governance concept and of good governance indicators, it seems quite appropriate when Davis, Kingsbury, and Merry call this mode of governance 'Global Governance by Indicators' (2012).

In contrast to the fact that political discussions are in general dominated by the good governance concept and that politicians of all countries and political colors like to think in rankings and benchmarks of good governance practices it is important to make clear that the primary goal of the governance approach is an analytical not a normative one. It is widely agreed upon that theory building in the field of governance has to start with an empirical analysis of the actual governance structures in a policy field or a given society and that the normative judgment in terms of the rule of law and theories of democracy has to follow as a second step. So the governance approach can be characterized as a two-step approach: analysis first, normative judgment on the basis of the empirical findings afterward. Otherwise, it would not be possible to look at governance in areas of limited statehood or to investigate the governance structures of the Hamas, Hezbollah, or of institutions and states at the borderline of organized crime (for governance as an analytical concept see especially Benz et al., 2007; Zürn, 2008; Schuppert, 2011a).

Governance by and as Communication

To govern means the ability to get things done. To achieve this, governments can use 'command and control' policies and negative and positive economic incentives. But that is not enough: more and more governments are dependent on persuading and motivating their citizens, a mode of governance, which Claus Offe (unpublished) has recently called 'ruling through signals': "The third family of policies is designed to affect courses of action through the sending of signals, defined as the dissemination of information which, if accepted as valid and incorporated into the beliefs of target actors, will modify their course of action. 'Signals are suggestions from the government. It encourages or recommends certain choices (…) and discourages others. With signals, in contrast to regulations, there is no enforcement and no compulsory compliance'" (Ringen: p. 8).

Stein Ringen, the by Claus Offe quoted scholar, discusses the phenomenon of 'ruling through signals' in a broader context; in a discussion paper with the nice title 'The Powerlessness of Powerful Government' he starts with the already mentioned two types of decisions, compulsory regulation and noncompulsory signals: "The orders contained in government decisions may be strong or weak. Decisions containing strong orders are called regulations: compliance is compulsory. Decisions containing weak orders I call signals; they are suggestions about behaviours or mindsets without compulsory compliance" (Ringen, unpublished paper: p. 6).

But then he comes up with a second important distinction, the distinction between 'power and authority.'

> Power explains a government's ability to get decisions made. But power is not what regulates its ability to get co-operation from its bureaucracies and compliance from its citizenry. For this, it does not help governments to be able to give orders, at least not very much; they need to be able to persuade. Persuasiveness is not contained in the weight of power; it depends on the strength of authority. Power is a necessary condition for a government to be able to rule but not a sufficient one. Once it is in power, its ability to rule depends on

> authority. Weak governments struggle for power. Powerful governments are beyond that but not be-yond struggling. They need to find the authority that can enable them to translate power into rule (p. 5).

But with this mode of governance, governments will be only successful, if they communicate their message, showing and practicing what we can call 'communicative leadership' (p. 11); this way of ruling seems to us the most promising mode of governance to meet the governance challenges ahead.

Three Examples for 'Governance at Work'

- Governance as crisis management: the case of financial governance

  The takeover of Merrill Lynch in March 2008 and the collapse of Lehmann Brothers 6 months later touched off a financial crisis that spread to many parts of the world. This caused severe governance problems, which forced the nation-states and the international community to look for innovative solutions. The main difficulty was that the problems were transnational whereas the financial and economic policies remained national: "Problems extend across borders more than before. At the same time, the jurisdictions for most economic policy remain national. Decisions national policymakers take create externalities for others. This suggests that coordination of policies across borders may help all countries. The EU attempted to coordinate policy in 2008–09 during the 'financial' part of the crisis, and it has put in place some new institutions meant to reduce the likelihood of future crises and to coordinate responses to crises when they occur" (Clark et al., 2013: p. 60).

  The national and European responses to the challenges of the global financial and economic crisis were quite remarkable. On the national level, the German legislature, for instance, created a new type of crisis reaction laws using a new and characteristic language: the German Bundestag passed three so-called 'Finanzmarktstabilisierungsgesetze' (Financial Market Stabilization Act) in 2008 and 2009 and in April 2009 a law for rescuing ailing banks, the so-called 'Rettungsübernahmegesetz' (Rescue Takeover Act). On the European level, the innovation was an organizational one; in September 2012 the European Stability Mechanism was established, an international financial institution situated in Luxembourg. This mechanism succeeded the European Financial Stability Facility and is at the very heart of the stabilization package. So the described governance innovations seem to support the well-known thesis of Joseph Schumpeter that every crisis owns a specific creative capacity.
- Transnational governance: the case of transnational rule making

  The governance area or governance realm that is referred to as transnational governance (Djelic and Sahlin-Andersson, 2006) is a space located somewhere beyond the reach of the nation-state and below the legal regime of international law and the authority of international organizations. In this space without clear-cut frontiers a need for transnational rules and transnational governance structures arises, a finding that Dieter Grimm (2003) puts as follows. "But beyond the nation state and the reach of international organizations modes of law making come into being, which cannot be influenced by states and international organizations. Global markets create regulations by themselves – independent of politics."

  Taking these findings as starting point, we can now try to describe the phenomenon of transnational governance in more detail.

  First, thinking in the category of new actors or new actor constellations, not only do we find a large number of them in the transnational realm but this space seems to be a perfect environment for the mushrooming of the type of actor that is characteristic of governance beyond the nation-state. We meet transnational standard setters such as 'standard-setting committees' in the field of accounting, transnational norm-building networks such as stewardships in different policy fields, transgovernmental networks as institutional backbones of a new world order (Slaughter, 2004), and – last but not least – actors such as rating agencies that are of the utmost importance for the governance of financial markets.

  Second, if we look for new modes of regulation, we have to explore a different world of rules compared to the legal system of nation-states. This world is characterized by the growing importance of standards created, for instance, by the International Standards Organization, codes of conduct of international corporations such as Shell or BP, and transnational contracts that are part of a new branch of governance labeled 'contract governance' (Riesenhuber and Möslein, 2009).

  Third, the most striking and – in terms of legitimacy and the rule of law – most challenging finding is the observable blurring of traditional borderlines, especially between public and private, formal and informal, and national and international. Examples of this erosion of borderlines in the transnational realm are so numerous that it might indeed be helpful to develop a 'theory of hybridity' (Trubek et al., 2005; Gillespie, 2009). Only one example shall be mentioned here: The erosion of the classic borderlines between binding and nonbinding norms through the process referred to as 'the hardening of soft law' (for more information see Schuppert, 2010).

  Fourth, it is more than obvious that new actor constellations, new modes of rule making, and the blurring of traditional borderlines raise a lot of questions in terms of legitimacy and accountability. The accountability problem is especially difficult to solve now that traditional chains of accountability become less omnipresent: how do we hold a network responsible, and how shall we deal with hybrid organizations? These questions are often discussed but rarely answered, and if progress is to be made in this difficult area, rethinking the concepts of accountability and legitimacy is needed.

  If we look back to the examples for transnational rule making, it is quite obvious that all these modes of rule making are so different in origin and character – from security standards to hybrid rule making by state-approved corporate governance codices – that they do not fit into

a passe-partout frame. Despite all these differences, some common features can, however, be found; following the observations of Stefan Oeter (2008) and Christian Tietje (2001, 2002), we can identify four peculiarities of transnational rule making:

- The first point is the deterritorialization of lawmaking. Although, the concept of law in the modern state is an etatistic and territorial one (Sassen, 2006), lawmaking under the conditions of globalization is a complex arrangement of territorial and deterritorialized law structures.
- The second characteristic feature is the blurring of the traditional borderline between binding and nonbinding norms. Lawmaking in a globalized world is a complex mix of nonbinding, semibinding, and legally binding norms, a mix of soft and hard law, and – to make it even more complicated – processes of hardening of soft law.
- The third aspect is an observable change from top-down lawmaking to a bottom-up system. While traditional lawmaking is a state-centered top-down process, modern transnational rule making is a process in which different actors and mechanisms are involved in network-style 'law production.'
- Above all, we witness a common trend to informality (Oeter, 2008). Informal rule making is not only a way to regulate when a legislative power is missing (for the role of soft law see Senden, 2004), it also seems to have many advantages in terms of flexibility and effectiveness, which means it has become increasingly popular.

- Governance as governance by partnerships

Policymaking is about choices – choosing the instruments that best suit a certain public policy (instrumental choice); choosing the 'right' organizational form or type of agency for implementing the policy in question (institutional choice); and – last but not least – choosing the appropriate regulatory regime, especially for the governance of partnerships (regulatory choice). To make the 'good choices' essential for effective public administration a certain amount of theoretical reflection is needed – as Donahue and Zeckhauser (2006) have pointed out with regard to 'collaborative governance': "Good governance requires choosing the right implementation model as well as the right ends. The richer the repertoire of alternative models, the more important is analytic work to guide the assignment of tasks. As government increasingly shares the collective-action stage with private actors, both for-profit and not-for-profit, addressing this assignment problem – who should do what? – becomes both more complex and more consequential."

To make these consequential choices in a sensible way, it is indispensable to be aware of the rich spectrum of choices as illustrated by Christoph Reichard (2006) in the following figure (**Figure 2**):

At the very heart of this choice, models are three sector-crossing institutional arrangements: contracting out, PPPs, and networks. As we come to networks later, it seems useful to say a few words about the differences and similarities between the first two modes of sector-crossing governance. At first, contracting out (this term is also used here as a synonym for 'functional privatization') and PPPs seem to be very similar: in both cases, the partners come from different sectors and their collaboration is governed by contracts. At a closer look, however, the differences are considerable, as the following table shows (**Table 3**).

As this table shows, to operate a PPP is a complex endeavor. The transaction costs are high. Choosing the appropriate partner causes information costs; facilitating control for the branch of public administration involved leads to high costs for comprehensive information and monitoring; and especially, the problem of developing a fair risk sharing requires high quality of governance.

Therefore, one of the best experts on PPPs in Germany – Detelf Budäus – has proposed to make a transparency report a necessary part of every complex PPP with an expected

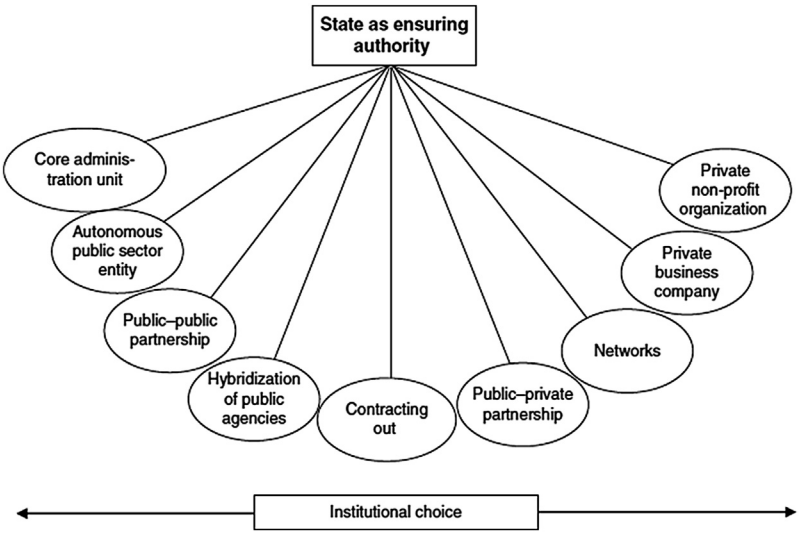

Figure 2 Variety of organizational possibilities. Source: Reichard, C. (2006). New institutional arrangements of public service delivery. In: Reichard, C., Kupke, S. (Eds.), The governance of services of general interest between state, market and society. Wissenschaftsverlag, Berlin, pp 35–48.

Table 3 A comparison between contracting out arrangements and partnerships

Contacting out	*Partnerships*
Government and company (or consortium) are involved in a principal–agent relationship	Government and company (consortium) are involved in joint decision-making and production
Government defines the problem, specifies the solution, and selects a private company that can produce results in a cost-efficient way	Both parties are involved in joint processes early on in order to develop joint products that contribute to both their interests
Benefits of contracting out arrangements especially relate to efficiency (quicker and cheaper)	Benefits of partnership arrangements especially relate to increasing effectiveness (synergy and enrichment of output)
Keys to success are unambiguous definitions of goals, projects, rules of tendering, rules of selection, and rules of delivery	Keys to success are an interweaving of goals, establishing rules for ongoing interaction, developing rules, and tailor-made assignments concerning joint effort and production commitments
Based on the principles of project management, because there is a clear principle, clear goals, and well-defined project specifications	Based on the principles of process management because the joint goals, the art of financing, realization, and utilization remain subject to joint decision-making
Contractual transparency regarding rules of tendering, selection, delivery, and rules of inspection are crucial for a good relationship	Mutual trust is crucial for lasting relationships between partners who maintain their own interest, ways of working, accountability, and financing principles

Source: Klijn, E., Teismann, G., 2000. Governing public–private partnerships. Analyzing and managing the processes and institutional characteristics of public–private partnerships. In: Osborne, S. (Ed.), Public–Private Partnerships. Theory and Practice in International Perspective. Routledge, London, pp. 84–105.

peril of additional costs, consisting of the following elements (Budäus and Grüb, 2008: p. 48; **Figure 3**):

Only such a precaution could minimize the specific problems, which are normally raised by the fact that in every PPP project there is a 'point of no return,' as the following phase model illustrates (Budäus and Grüb, 2008: p. 47; **Figure 4**): PPPs can be found in nearly every field of policy – as the following four examples show (for details see Sack, 2006):

- Very popular are PPPs in the field of public transport. Not only in Germany but also in other European countries, highways, bridges and tunnels are built by private companies on their own expenses. In the PPP agreement, they are allowed to levy charges from the users. A famous example for a really complex PPP is the development of toll systems for trucks – called Toll Collect – in cooperation between the department of transport, Daimler Chrysler, German Telekom, and the French company Cofiroute.
- PPPs are used worldwide for building constructions, including schools, hospitals, jails or airports. In this field cost-explosions are a regularly made experience, a finding which supports the above mentioned instrument of a transparency report.
- A third example of the usefulness of PPP arrangements are the challenges of 'E-Government.' In this field, administrative agencies are often dependent on the specific know-how of private companies, an expertise which they can buy (contracting out) or integrate in an organizational PPP.

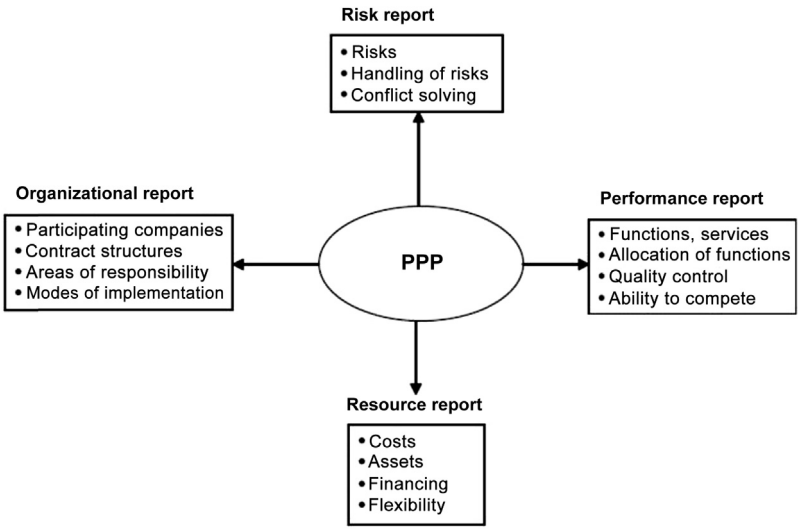

Figure 3 Elements of the transparency report.

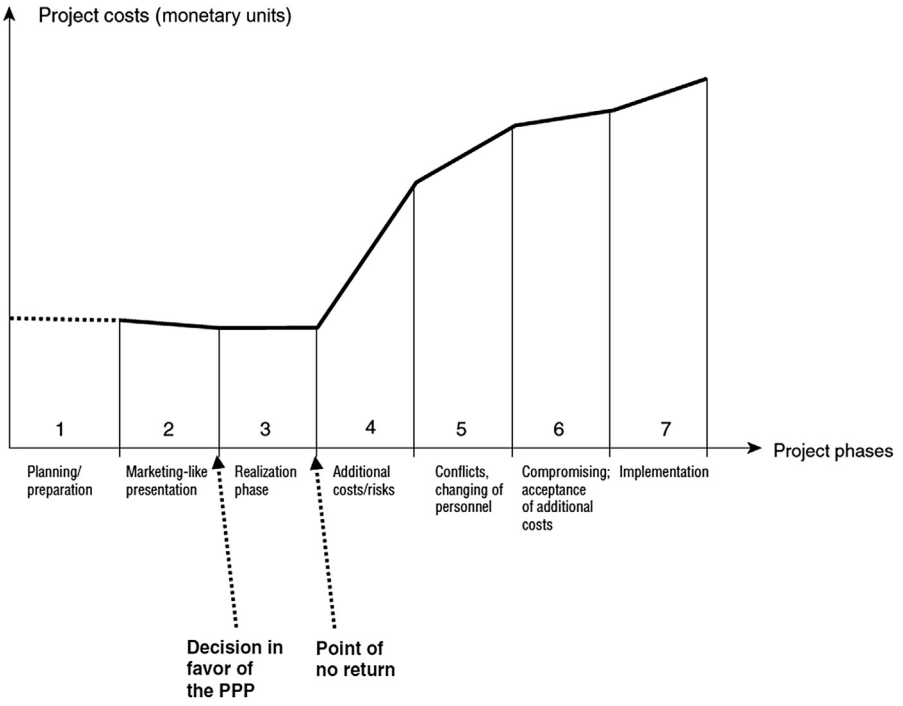

Figure 4 Phase model of PPPs at work.

Strength and Weaknesses of the Governance Concept

Beginning with the strength of the concepts I would like to make three points.

Governance as an Interdisciplinary Bridging Concept

That governance is a bridging concept seems to be widely accepted (van Kersbergen and van Waarden, 2004; Schuppert, 2006; Benz et al., 2007; Börzel and Risse, 2010). During the last 20 years, the governance paradigm has developed into a conceptual bridge enabling fruitful scientific exchange between political science, sociology, law, and history. At the same time, it enables scholars to talk about policies and politics beyond a narrow focus on the state.

However, following this metaphor of governance as a bridging concept implies that a coherent and comprehensive governance theory cannot exist on its own. If the pillars of the bridge are rooted in different disciplines – the recently published SAGE Handbook of Governance (2011) mentions 11 different approaches as 'theories of governance'– a 'governance theory' is by its very nature a multidisciplinary endeavor with all the weaknesses and advantages of a multifaceted scientific approach that seeks to explain the dynamics of politics and policies.

Governance as a Mirror of Changes in Statehood

There is – as it seems – a broad consensus in the governance community that governance and changes of statehood belong together and that changes in governance reflect changes of statehood. As a witness of this consensus, we want to quote some sentences, which can be found in the concluding remarks of the introductory article of Mark Bevier in the SAGE Handbook of Governance:

> The term 'governance' has risen to prominence in the last 30 years as a way of describing and explaining changes in our world. […] Typically, the new governance refers to changes in the nature and role of the state since the last quarter of the twentieth century. The state has become increasingly dependent on organizations in civil society and more constrained by international linkages. On the one hand, the public sector in many states has shifted away from bureaucratic hierarchy and toward markets and networks; governance thus captures the ways in which patterns of rule operate in and through groups within the voluntary and private sector. On the other hand, states have become increasingly embroiled with transnational and international settings as a result of the internationalization of industrial and financial transactions, the rise of regional blocks, and concerns over problems such as terrorism and the environment; governance thus captures the formal and informal ways in which states have attempted to respond to the changing global order.
>
> Bevir, 2011: p. 15.

Governance as an Analytical Tool for Studying the Institutional Architecture of Modern Statehood

A very good example of the analytical capacity of the governance concept is the contribution to explain the institutional architecture of modern statehood from the network governance research (Börzel, 1998; Kenis et al., 2008). In the famous book 'A New World Order,' Anne-Marie Slaughter (2004) has given 'governance by government networks' a very prominent place. She illustrates how this mode of

governance works by looking at the 'networking of finance ministers and top bankers' (p. 38):

> Networking among some regulators, such as central bankers, securities commissioners, and insurance supervisors, has become so established that they now have their own international organizations – the Basel Committee, IOSCO Commissioners and IAIS, among others. These organizations are transgovernmental networks that have become sufficiently formalized to warrant the title of association or organization and that have a staff and regular meetings. But they are not 'inter-state' organizations; they are not formed by treaty or even executive agreement; they have no place on the landscape of the international legal system.

The main tasks of such networks are – as Anne-Marie Slaughter points out – the following:

- exchanging information,
- coordinating policy,
- cooperating on enforcement issues,
- collecting and distilling best practices,
- exporting particular regulatory forms,
- bolstering their members in domestic bureaucratic politics, and
- transmitting information about their members' reputation.

If we look at this very illuminating list, it is obvious that networks cannot function without a very intensive communication; they practice what can be called 'governance by and as communication.' The best proof for that is the short answer of A.M. Slaughter to the question what networks are doing: 'Their members talk a lot' (p. 51).

But of course there is also some criticism of the governance concept:

The most popular criticism is the reproach that the governance concept neglects the dimensions of power and politics. Claus Offe, for instance, argues that governance has – in contrast to steering – no subject, only structures: "Something happens but no one has done it and no one would be responsible." (Offe, 2008: p. 61) Similar are the critical remarks of Edgar Grande: What does steering by governance structure mean, who is the actor behind governance institutions and governance arrangements and, does it really make sense to speak of institutional steering by governance structures (Regelungsstrukturen)? The answer is yes, it makes sense to talk about institutional steering (Schuppert, 2012). Politics is not only as mentioned above about institutional choices, but also about designing and creating new institutions or reforming traditional ones. All the political controversies of the last 20 years have been arguments about 'getting the institutions right'; that is true for the reform of the school system, the health care structure, the reform of financial federalism, the rescue of the Euro, labor market policies, and so forth: designing and creating institutions is making politics and exercising power (Olsen, 2010).

To sum it up, we would argue that the advantages of the governance concept outweigh its weaknesses in a very convincing manner.

See also: Global Governance; Global Public Policy.

Bibliography

Acemoglu, D., Robinson, J.A., 2012. Why Nations Fail. The Origins of Power, Prosperity, and Poverty. Crown Publishing, New York.

Benz, A., 2004. Einleitung: Governance - Modebegriff oder nützliches sozialwissenschaftliches Konzept? In: Benz, Arthur (Ed.), Governance - Regieren in komplexen Regelsystemen. VS Verlag, Wiesbaden, pp. 11–28; translation from German into English from author.

Bevir, M., 2011. Governance as theory, practice, and dilemma. In: Bevir, M. (Ed.), The SAGE Handbook of Governance. Sage Publications, London, pp. 1–16.

Börzel, T.A., 1998. Organizing Babylon – on the different conceptions of policy networks. Public Administration 76, 253–273.

Börzel, T.A., Risse, T., 2005. Public–private partnership. Effective and legitimate tools of transnational governance? In: Grande, E., Pauli, L.W. (Eds.), Complex Sovereignty: Reconstitution Political Authority in the Twenty-first Century. University of Toronto Press, Toronto, pp. 195–216.

Börzel, T.A., Risse, T., 2010. Governance without a state: can it work? Regulation and Governance 4, 113–134.

Budäus, D., Grüb, B., 2008. Public Private Partnership (PPP): Zum aktuellen Entwicklungs- und Diskussionsstand. In: Bauer, H., Büchner, C., Brosius-Gersdorf, F. (Eds.), Verwaltungskooperation. Public private partnerships und public public partnerships. Universitätsverlag Potsdam, Potsdam, pp. 33–50.

Burris, S., Kempa, M., Shearing, C., 2008. Changes in governance. A cross-disciplinary review of current scholarship. Akron Law Review 41, 1–66.

Clark, W.R., Copelovitch, M., Hallerberg, M., Quaglia, L., Walter, S., 2013. Challenge in focus. Financial and fiscal governance. In: Hertie School of Governance (Ed.), The Governance Report 2013. Oxford University Press, Oxford, pp. 59–81.

Coase, R.H., 1937. The nature of the firm. Economica 4, 386–405.

Davis, K., Kingsbury, B., Merry, S., 2012. Introduction: global governance by indicators. In: Fisher, A., Davis, K., Kingsbury, B., Merry, S.E. (Eds.), Governance by Indicators. Global Power through Quantification and Rankings. Oxford University Press, Oxford, pp. 3–28.

Djelic, M.-L., Sahlin-Andersson, K. (Eds.), 2006. Transnational Governance: Institutional Dynamics of Regulation. Cambridge University Press, Cambridge.

Donahue, J.D., Zeckhauser, R.J., 2006. Public–private collaboration. In: Muran, M., Rein, M., Goodin, R.E. (Eds.), The Oxford Handbook of Public Policy. Oxford University Press, Oxford, pp. 496–526.

Gillespie, J., 2009. The role of state, non-state and hybrid actors in localizing global scripts in East Asia. In: Gillespie, J., Peerenboom, R. (Eds.), Regulation in Asia: Pushing Back on Globalization. Routledge, New York, pp. 24–50.

Grimm, D., 2003. Gemeinsame Werte: Globales Recht? In: Däubler-Gmelin, H., Mohr, I. (Eds.), Recht schafft Zukunft: Perspektiven der Rechtspolitik in einer globalisierten Welt. Dietz Verlag, Bonn, pp. 14–29.

Jann, W., 2011. Executive Seminar "Managing Organizational Transformations". Challenges to Public Governance, Potsdam.

Jürgens, U., 2006. Corporate Governance. Anwendungsfelder und Entwicklungen. In: Schuppert, G.F. (Ed.), Governance-Forschung. Vergewisserung über Stand und Entwicklungslinien, second ed. Nomos, Baden-Baden, pp. 47–71.

Kaufmann, D., Kraay, A., Mastruzzi, M., 2010. The Worldwide Governance Indicators. Methodology and Analytical. The World Bank Policy Research Working Paper, No. 5430.

Kersbergen, K.V., Waarden, F.V., 2004. Governance as a bridge between disciplines. Cross-disciplinary inspiration regarding shifts in governance and problems of governability, accountability, and legitimacy. European Journal of Political Research 43, 143–171.

Kenis, P., Nölke, A., Raab, J., 2008. Transnationale Politiknetzwerke. Insitutionenkultur jenseits des Nationalstaates. In: Gosewinkel, D., Schuppert, G.F. (Eds.), Politische Kultur im Wandel von Staatlichkeit. WZB-Jahrbuch 2007. Edition Sigma, Berlin, pp. 163–180.

Klijn, E.H., Teisman, G.R., 2000. Governing public–private partnerships. Analyzing and managing the processes and institutional characteristics of public–private partnerships. In: Osborne, S.P. (Ed.), Public–private Partnership. Theory and Practice in International Perspective. Routledge Chapman and Hall, London, pp. 84–104.

Nuscheler, F., 2009. Good Governance. Ein universelles Leitbild von Staatlichkeit und Entwicklung? Institut für Entwicklung und Frieden. INEF-Report, 96.

Oeter, S., 2008. The Openness of International Organisations for Transnational Public Rule-Making. Paper presented during the workshop "The law beyond the state: Public governance, networks and organization in the transnational sphere", 19–21 June 2008.

Olsen, J.P., 2010. Governing through institution building Institutional Theory and Recent European Experiments in Democratic Organization. Oxford University Press, Oxford.

Offe, C., 2008. Governance: "Empty signifier" oder sozialwissenschaftliches Forschungsprogramm? In: Schuppert, G.F., Zürn, M. (Eds.), Governance in einer sich wandelnden Welt. VS Verlag für Sozialwissenschaften, Wiesbaden, pp. 61–76.

Offe, C., Signals and information as a resource of public policy, unpublished.

Ringen, S., The powerlessness of powerful government, unpublished paper, 25 p.

Riesenhuber, K., Möslein, F., 2009. Contract governance – a draft research agenda. European Review of Contract Law 5, 248–289.

Reichard, C., 2006. New institutional arrangements of public service delivery. In: Reichard, C., Mussari, R., Kupka, S. (Eds.), The Governance of Services of General Interest between State, Market and Society. Wissenschaftsverlag, Berlin, pp. 35–48.

Sack, D., 2006. Eine Bestandsaufnahme der Verbreitung, Regelungen und Kooperationspfade vertraglicher PPP in Deutschland – Effizienz, Kooperation und relationaler Vertrag. In: Budäus, D. (Ed.), Kooperationsformen zwischen Staat und Markt. Theoretische Grundlagen und praktische Ausprägungen von Public Private Partnership. Nomos, Baden-Baden, pp. 30–51.

Sassen, S., 2006. Territory, Authority, Rights: From Medieval to Global Assemblages. Princeton University Press, New York.

Schuppert, G.F., 2006. Governance im Spiegel der Wissenschaftsdisziplinen. In: Schuppert, G.F. (Ed.), Governance-Forschung. Vergewisserung über Stand und Entwicklungslinien, second ed. Nomos, Baden-Baden, pp. 371–469.

Schuppert, G.F., 2010. Governance und Rechtsetzung. Grundfragen einer modernen Regelungswissenschaft. Nomos, Baden-Baden.

Schuppert, G.F., 2011a. Partnerships. In: Bevier, M. (Ed.), The SAGE Handbook of Governance. Sage Publishing, London, pp. 286–299.

Schuppert, G.F., 2011b. Law without the state? In: Risse, T., Lehmkuhl, U. (Eds.), Governance without the State? Policies and Politics in Areas of Limited Statehood. Columbia University Press, New York, pp. 55–88.

Schuppert, G.F., 2012. Verwaltungsorganisation und Verwaltungsorganisationsrecht als Steuerungsfaktoren. In: Hoffmann-Riem, W., Schmidt-Assmann, E., Voßkuhle, A. (Eds.), Grundlagen des Verwaltungsrechts. C. H. Beck, München, pp. 1067–1159.

Senden, L., 2004. Soft Law in European Community Law. Hart Publishing, Oxford.

Slaughter, A.-M., 2004. A New World Order. Princeton University Press, New York.

Tietje, C., 2001. Internationalisiertes Verwaltungshandeln. Dncker & Humblot, Berlin.

Tietje, C., 2002. Transnationales Wirtschaftsrecht aus öffentlich-rechtlicher Perspektive. Zeitschrift für Vergleichende Rechtswissenschaft 101, 404–420.

Trubek, D., Cotrell, P., Nance, M., 2005. 'Soft law', 'hard law' and European integration: toward a theory of hybridity. In: In de Búrca, G., Scott, J. (Eds.), New Governance and Constitutionalism. Hart Publishing, Oxford, pp. 1–41.

Williamson, O.E., 1985. The Economic Institutions of Capitalism. Free Press, New York.

Zürcher, C., 2007. When governance meets troubled states. In: Beisheim, M., Schuppert, G.F. (Eds.), Staatszerfall und Governance. Nomos, Baden-Baden, pp. 11–27.

Zürn, M., 2008. Governance in einer sich wandelnden Welt. Eine Zwischenbilanz. In: Schuppert, G.F., Zürn, M. (Eds.), Governance in einer sich wandelnden Welt, PVS-Sonderheft, vol. 41, pp. 553–580.

Zürn, M., Enderlein, H., Wälti, S., 2010. Handbook on Multi-level Governance. Edward Elgar, Cheltenham.

Government Statistics

M Straf, National Research Council, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Government statistics are essential for a democratic government and they serve to advance businesses and the well-being of its citizens. Statistical agencies provide accurate, timely, and credible information, which informs current and future policies. Respect and trust of data providers are fostered by maintaining confidentiality of their information. Government statistics are collected through censuses and surveys and from administrative records and other sources. Defining practices of an effective statistical agency begin with a clear, well-accepted mission and a strong standing of independence. Today, statistical agencies are challenged to keep up with technological change and new sources of data in a rapidly changing world with a global economy.

Introduction

Government statistics are vital to advance the economic well-being and quality of life of a nation's citizens. Without such statistics, policy making would be less effective, economic markets would be less informed, and the countless decisions of private businesses and individuals would allocate economic and other resources less efficiently. But more, government statistics are vital to a democracy, as the information they convey provides the means by which citizens can hold their government accountable.

This article provides a general review of government statistics largely in the context of the USA, but it also addresses broader issues relevant to other countries. The focus is on the national government and, therefore, mostly on national statistics, but much is also relevant to governments at subnational levels, which also produce statistics.

The article begins with the purposes of government statistics and the rationale for why governments produce them. Next it takes up the principles by which the statistics should be produced to serve these purposes effectively; this is followed by descriptions of various types of data collection and of how the responsibilities for collection, analysis, and reporting of statistics are organized among statistical agencies. Effective practices of these agencies are identified. The article concludes with a discussion of some challenges faced by statistical agencies.

Statistics from government sources are referred to in many articles in this Encyclopedia (*see* Censuses: History and Methods; Censuses, Population: Comparative International Aspects; Census: Current Approaches and Methods; Census Microdata; Panel Surveys: Uses and Applications; Sample Surveys: Cognitive Aspects of Survey Design; Large-Scale Social Surveys, Analysis of; Data Bases and Statistical Systems: Applied Social Research; Data Bases and Statistical Systems: Demography; Data Bases and Statistical Systems: Economics (General)).

What Are Government Statistics?

'Government statistics' means factual information that is reported upon for statistical purposes by a unit of government. Some countries also use the term 'official statistics.' In this article, 'statistics' and 'data' are used interchangeably to mean factual information, and the terms are not restricted to quantitative information. Usually, the unit or agency of government that issues the statistics has collected or compiled the data to produce them. In most cases, the agency is a distinct entity of government entrusted with reporting the statistics. In some cases, the agency may analyze its data to provide more information about the statistics in its reports, especially information that makes the statistics more relevant to public policy concerns. For the agency to be effective, the focus of any analysis must be on ways to improve the relevance and accuracy of data for their use in public policy analysis and decision making and to explain what the data mean or imply, but not to advocate any particular policy.

The *sine qua non* of the definition of government statistics is that they are collected and reported upon only for *statistical purposes*:

> 'Statistical purposes include description, evaluation, analysis, inference, and research. Statistical purposes relate to descriptions of groups and exclude any interest in or identification of an individual person or economic unit. For these purposes, a statistical agency may collect data directly from providers, or it may obtain data from other sources, but it does not use these data for administrative, regulatory, or law enforcement purposes. The data are used solely to describe and analyze statistical patterns, trends, and relationships involving groups of individuals or other units.
>
> NRC, 2013b, p. 10

Why Do Governments Produce National Statistics?

A government provides statistics to inform its citizens, satisfying an essential requirement of a democracy (see, e.g., United Nations Statistical Commission, 1994). The government, however, is most often a primary user of the statistics it produces. The statistics are used in the formulation of, and decision making on, public policies, as well as in their implementation and evaluation. Information from statistics, for example, is used to inform economic markets and monetary policy; to determine the effects of proposed policy changes; to adjust taxes and benefits, as well as private contracts; to allocate government funds among state or regional

governmental entities; to inform decision making at these levels; and to inform private sector investment decisions and individual decision making.

In most cases, a government collects and reports the statistics itself, either directly or through the services of a private organization. In a few cases, the private sector satisfies the demand for national statistics. Examples are the Conference Board's compilation of US leading economic indicators, abortion statistics for many countries compiled by the Alan Guttmacher Institute, and many private market-research surveys and reports of economic forecasting firms. But even when the private sector can do so, there are reasons for a government to collect the statistics. Alternative sources may not exist. Timeliness may be a critical factor. And statistics collected by government agencies are often deemed to be of greater quality.

A more compelling reason may be trust and credibility. Potential subjects of government data collection may be more willing to volunteer their time and effort and divulge confidential personal or proprietary information for a government survey rather than for a private one, in part because of legal protections of confidentiality. And the public and other important users may find data collected by the government to be more credible.

A fundamental reason, however, for a government to collect data has to do with the very nature of data. If a commodity can be consumed by many and consumption by one does not reduce the consumption of others, then the commodity is a 'public good.' Data are manifestly public goods: use of data by one party in no way diminishes their usefulness to other parties. Individuals benefit from a public good but, unless each individual is charged for the consumption of the good, they are unlikely to pay for them voluntarily. Consequently, the goods may not be produced, unless the government intervenes. In the case of data, the government may intervene directly by collecting the data, or it may intervene indirectly, for example, by defining rights for a license or copyright to provide the means for the private sector to charge for the value added in producing secondary data (data used or prepared for uses other than the purpose for which they were collected). For these reasons, a government typically provides primary data (data collected for a specific purpose) and the information infrastructure. The private sector typically provides secondary data and may build on the infrastructure.

Principles of Government Statistics

A report of the US National Research Council of the US National Academy of Sciences, National Academy of Engineering, and Institute of Medicine (NRC, 2013b) defines a 'statistical agency' as a unit of government whose principal function is the compilation and analysis of data, and the dissemination of information for statistical purposes. The unit is recognized generally as a distinct entity, which may be located in a cabinet department or independent agency. The mission of a distinct entity or agency of government that issues government statistics may or may not be solely statistical.

Statistical agencies are established to be a credible source of accurate, timely, and relevant statistics in one or more subject areas. Accuracy requires proper concern for consistency in spatial and temporal dimensions, as well as statistical measures of error. Timeliness requires concern for the frequency of issuing data, as well as disseminating data to users when they need them. Relevance requires concern for anticipating future data needs, as well as meeting current ones. And credibility requires concern for the reality and appearance of impartiality, as well as the reality and appearance of independence from control for partisan political purposes. It is essential to the mission of statistical agencies to work to ensure the goals of accuracy, timeliness, relevance, and credibility of statistical information.

The US National Research Council report identifies four principles for a statistical agency. The first is that the agency is in a position "to provide objective, accurate, and timely information that is relevant to issues of public policy" (NRC, 2013b, p. 12). This principle requires the agency to be knowledgeable about public policy issues and be able to provide objective information that is relevant to the needs of formulating, implementing, and evaluating public policies. The second is that the agency has credibility with those who use its data and information. To develop that relationship, an agency must strive to maintain credibility for itself and for its data. In particular, the agency must be perceived to be free of political interference and policy advocacy. Other aspects that contribute to credibility are a commitment to quality and professional standards and openness about all aspects of the data, especially the quality of the data.

The third principle is that the agency has a relationship of mutual respect and trust with respondents who provide data, and with all data subjects whose information it obtains. Statistical agencies rely in large part on information supplied by individuals and organizations outside the federal government, such as state and local governments, businesses, and other organizations. Some of this information is required by law, such as employers' wage reports, and some of it is compiled from records of government programs. But much of it is obtained through the voluntary cooperation of respondents to surveys. Even when the response to a survey is required by law, cooperation by respondents promotes accuracy, which, in turn, promotes the credibility of the agency's data. In encouraging cooperation, it is important to statistical agencies that respondents perceive that the information they are being asked for is important; that it is being collected in an impartial, professional manner; that their privacy and autonomy are respected; and that their personal or proprietary information will remain confidential.

When statistical agencies pledge to keep respondents' information confidential, it means that the information will be used only for statistical purposes and that, in particular, it will not be used for any administrative, regulatory, or enforcement purpose regarding the individual (*see* Confidentiality and Statistical Disclosure Limitations).

Statistical agencies go to great lengths to maintain the confidentiality of individual responses (see NRC, 1993, 2000). In the USA, some agencies, such as the US Census Bureau, long had legislation to protect confidentiality, and some, such as the US Bureau of Labor Statistics, have had a history of successful reliance on executive order, custom, and case law. A federal law, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA), established a uniform set of safeguards to protect the confidentiality of individually

identifiable information acquired from the public for statistical purposes by any agency of the US government. With this protection, CIPSEA also authorizes the sharing of business data among the Bureau of Economic Analysis (BEA), the Bureau of Labor Statistics (BLS), and the Census Bureau (Census), to achieve efficiencies.

In Europe, the 1995 European Community Directive on Data Protection governs the transfer of individually identified or identifiable data. The European Commission has proposed a comprehensive reform to modernize and extend this Directive (European Commission, 2012), which was supported by the European Parliament (European Commission, 2014).

The respect and trust of data providers are also essential for statistical agencies to achieve credibility for their data. Protecting the confidentiality of individual responses engenders trust.

Respect for data providers also means valuing their time and effort to respond, in particular, designing data collection so that it is not unduly burdensome and valuing their responses so that they are put to important uses. It is important too that respondents perceive that the agency collecting their information is trustworthy and credible, in particular, that the information will not be used for a particular partisan interest. A statistical agency must be perceived, by its data providers as well as its data users, to be working in the national interest.

The fourth principle is that a statistical agency be independent from political and other undue external influence in developing, producing, and disseminating statistics. A statistical agency should encourage external input on its programs, of course. By an 'undue external influence' the National Research Council report means that which seeks to undermine the agency's impartiality and professional judgment. Independence from such influences – in appearance as well as in reality – is essential for credibility with data users and the public.

How Are Government Statistics Collected and Reported?

Statistical agencies collect data through censuses, surveys, and compilations from administrative records and other sources. Some agencies use nontraditional sources, such as private vendors for expenditure data, and some are exploring new sources, such as Internet websites, for example, price data for online purchases, and transaction data, for example, purchase data scanned in retail stores. The vast amount of electronic data available presents challenges to extract useful information and raises questions of quality and representativeness.

Many countries conduct a census of their population. In the UK and USA, one is conducted every 10 years; in Canada, every 5 years. To provide more timely data than which were previously collected only every decade in the census, the US Census Bureau conducts the American Community Survey of approximately 250 000 addresses monthly.

Other examples of US censuses are the quinquennial census of agriculture, which covers agricultural production on farms that are defined as places from which $1000 or more of agricultural products were produced and sold in the census year; the quinquennial economic census, which covers the economic activity of nonfarm establishments; and the quinquennial census of governments, which covers employment and finances of state and local governments.

Examples of surveys in the USA include the Current Population Survey, from which monthly employment and unemployment data are derived; the Consumer Expenditure Survey, from which data on expenditure patterns are used periodically to adjust the weights of goods and services in producing the Consumer Price Index; and the National Health Interview Survey, which provides annual estimates of health conditions and use of healthcare services. Data in government surveys may be collected by field staff of an agency (e.g., the Census Bureau conducts the Survey of Income and Program Participation with its field staff); by state or local governments under a cooperative program (the Current Employment Statistics program of the Bureau of Labor Statistics is a monthly survey, based on payroll records of business establishments, conducted by state employment security agencies); or by another federal agency (the Census Bureau conducts the Current Population Survey and the Consumer Expenditure Survey for the Bureau of Labor Statistics and the National Health Interview Survey for the National Center for Health Statistics). In addition, private organizations may collect data for a government agency under a contract or grant (the University of Michigan conducts the Health and Retirement Study under a grant from the National Institute on Aging).

Many of these surveys have counterparts in other countries: for example, the Labour Force Survey of the Office of National Statistics in the UK and the French Labor Force Survey of the National Institute for Statistics and Economic Studies in France.

The units of observation in surveys may be individuals, households, families, schools, establishments, or other economic units. Surveys may be cross sectional to provide an estimate at a point or interval of time, or they may be longitudinal in design in that they follow the same units of observation over time. Longitudinal surveys can provide information on changes in a unit over time: for example, transitions of families in an out-of-poverty status from the Survey of Income and Program Participation (*see* Longitudinal Data: Event–History Analysis in Discrete Time; Panel Surveys: Uses and Applications). Some surveys, such as the American Housing Survey, however, may also employ a longitudinal component in their design to reduce error or increase efficiency.

Statistical agencies also compile information from administrative records. For example, vital statistics are compiled by the National Center for Health Statistics from state agencies, and administrative records are the source of basic statistics for very small firms in the quinquennial economic census, thus reducing or eliminating the need to send these firms census questionnaires.

Statistical agencies may be dependent on the censuses and surveys of other agencies as well as their own data collection in order to produce national indicators. The Bureau of Economic Analysis relies heavily on economic statistics from other sources in compiling the national income and product accounts. And most agencies rely on the Census Bureau to produce the population denominators for reported rates.

Since the US statistical system is so decentralized, important information on a population of interest, such as children or the elderly, is often collected by many agencies and in different ways that may make relating the data difficult, if not

impossible. Sometimes, agencies collaborate to fill important information gaps. Two examples are key indicators of well-being for older Americans (Federal Interagency Forum on Aging-Related Statistics, 2012) and for children (Federal Interagency Forum on Child and Family Statistics, 2013).

Whatever the source of data, one of the most important activities of a statistical agency is measuring, controlling, and reporting upon error. Statistical agencies have been preeminent in the development of survey designs to reduce sampling error. And many require, as a standard for some forms of publication, that statistical errors be reported. The agencies too are concerned about measurement and other forms of nonsampling errors, and employ various means to study them, including reinterview studies, record checks, multiple indicators, and cognitive studies, which focus on how errors arise in the process of answering questions. Three statistical agencies, the Census Bureau, the Bureau of Labor Statistics, and the National Center for Health Statistics, have cognitive laboratories to evaluate questionnaires and conduct other research to better understand and reduce error (*see* Sample Surveys: Cognitive Aspects of Survey Design; Questionnaires: Cognitive Approaches).

How Are Government Statistics Organized?

Government statistical activities may be centralized into a single agency or not. The organization in the USA, which is described first, is highly decentralized with activities dispersed among many agencies. Some of the trade-offs between centralized and decentralized systems are described at the end of this section.

The US Statistical System

The USA has a highly decentralized statistical system, with over 100 agencies having direct funding of over $500 000 per year for statistical activities. Only some of these agencies are statistical agencies (US Office of Management and Budget, 2014b).

The highly decentralized statistical system in the USA arose historically because, as new subject matter areas emerged as a national priority, new statistical agencies were created to meet the needs for new data. Little attention was paid to the relevant data-collection activities of other agencies. Efforts at collaboration have met with some successes, such as the interagency forums on aging-related statistics and on child and family statistics, but achieving systemic collaboration has proved more difficult. Attempts to achieve collaboration through centralization, or at least a major consolidation, have run into stakeholders with a special interest in retaining strong, relatively independent agencies within a particular department of government (see Norwood, 1995).

Responsibility for coordinating the myriad statistical activities of the US government rests with a small unit in the Office of Management and Budget, the Statistical and Science Policy Office in the Office of Information and Regulatory Affairs. Authorization of that responsibility derives from legislation originally concerned with paperwork reduction. Despite its small size, however, the Statistical and Science Policy Office has brought statistical agencies together to address many matters of common concern.

The 1995 reauthorization of the Paperwork Reduction Act codified an Interagency Council on Statistical Policy, formalizing an arrangement whereby the heads of the principal statistical agencies meet regularly together with the US Chief Statistician, who heads the OMB Statistical and Science Policy Office. The Council comprises 13 principal statistical agencies: the Bureau of Economic Analysis and the Bureau of the Census in the Department of Commerce; the Bureau of Justice Statistics in the Department of Justice; the Bureau of Labor Statistics in the Department of Labor; the Bureau of Transportation Statistics in the Department of Transportation; the Economic Research Service and the National Agricultural Statistics Service in the Department of Agriculture; the Energy Information Administration in the Department of Energy; the National Center for Education Statistics in the Department of Education; the National Center for Health Statistics in the Department of Health and Human Services; the Statistics of Income Division of the Internal Revenue Service in the Department of the Treasury; the Office of Research, Evaluation and Statistics in the Social Security Administration; and the National Center for Science and Engineering Statistics in the National Science Foundation.

The Interagency Council, the Federal Committee on Statistical Methodology and other interagency committees it convenes, and the Committee on National Statistics of the US National Research Council (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine) have addressed many methodological and other issues that cut across the interests of most statistical agencies. These include issues of poverty measurement, definitions of race and ethnicity and of metropolitan areas, and confidentiality of government statistics. The Statistical and Science Policy Office is also concerned with international comparability of statistics, and it represents the USA at the United Nations Statistical Commission and other international organizations. Moreover, the Statistical and Science Policy Office has striven to extend the protection of confidential data in ways that would allow statistical agencies to share needed information under that protection.

The Statistical and Science Policy Office is the only unit of government charged to be concerned with the government statistical system as a whole. It conducts an annual review of all major statistical programs and publishes a special analysis of funds proposed for statistical activities in the President's budget (US Office of Management and Budget, 2014a).

Centralized versus Decentralized Systems

Most large countries of the world have centralized most or all of their statistical activities into a single agency with broad authority for collecting and compiling government statistics. Coordination of the different activities occurs within the agency itself. Examples are Statistics Canada, the statistical agency of Canada, and the Australian Bureau of Statistics.

A central statistical agency is usually seen to be more efficient in resource allocation; in reaping the benefits of scale; in sharing information across different subject matter areas in order to enhance data and to avoid duplicating activities; in coordinating data activities, such as developing a single longitudinal survey of youth that serves the purposes of both health and education policy makers; and in arranging for the effective

use of administrative data. Advantages of a decentralized system are that priorities for budget, relevance, and timeliness, and the urgency to meet needs for new data are set within the department with the most concern for the information and with policies that are most affected by it, and that different perspectives and ways of conducting statistical activities can bring strength to the system as a whole.

Practices of Effective Statistical Agencies

The operation of an effective statistical agency begins, first, with a clearly defined and well-accepted mission and, second, with necessary authority to protect independence. With these prerequisites, a number of practices contribute to the agency's effectiveness. According to the National Research Council (2013b) report on principles and practices for a statistical agency, these include continual improvement of and development of more useful data; openness about its data; wide dissemination of data; cooperation with data users; fair treatment of data providers, in particular respecting their privacy and autonomy and protecting the confidentiality of personal information they provide; commitment to quality and professional standards, including through a highly professional staff; an active research program; a strong program of evaluation; and coordination and collaboration with other statistical agencies. Effective statistical agencies also engage in the analysis and interpretation of their data and in advancing their effective use.

Mission

An agency's mission should include responsibility for assessing and meeting needs for information in its purview, determining the sources of the data and the measurement and collection methods, and ensuring that its data are available to the public. An agency mission statement often goes beyond its charge in legislation. In particular, it should articulate the principles that will govern all of its work. The mission should be well accepted externally by the agency's department and others who oversee the agency, and especially by data users. It should also be well accepted internally, so that all agency employees understand how their work furthers the mission.

Independence

A strong and widely acknowledged standing of independence from undue external influences is necessary for a statistical agency to have the credibility to carry out its mission effectively. In particular, the agency must be independent of other parts of its department engaged in policy making, enforcement, regulatory, or administrative activities. It must be impartial and avoid even the appearance that its data might be manipulated for political purposes, or that information on an individual might be used for enforcement, regulatory, or administrative purposes.

Forms of independence differ among statistical agencies. Some of the other characteristics of independence include authority to decide the scope, content, frequency, and manner of data collected, compiled, or reported; to select and promote professional staff; to release statistical information without prior clearance and with a clear separation from policy interpretations by others; to represent the agency's statistical program before legislatures and the public; and to establish and follow predetermined schedules for the release of important economic and other indicators.

Continually Developing More Useful Data

In order to make their data more useful and to meet changing public policy needs, statistical agencies must continually look to develop data systems that can serve broader purposes. Integrating data from surveys conducted by different agencies or integrating data from surveys with data from administrative records, such as tax or social welfare program records, are ways to achieve this goal. When separate data sets are collected and analyzed in such a way that they may be used together, the value of each may be greatly enhanced. Data sharing and statistical uses of administrative records, with appropriate safeguards for confidentiality, also make a statistical agency more efficient as well as more effective.

Openness

An effective statistical agency operates in a completely open manner. Openness requires providing full documentation of the data with acknowledgment of any uncertainty, a description of the methods used and assumptions made, and measures or other descriptions of error, for both sampling and nonsampling error. When a mistake occurs, the agency should acknowledge it promptly and publicly. For a correction to an important indicator, the agency should strive to inform as many current and potential users as possible.

Dissemination of Data

An effective statistical agency seeks the widest possible dissemination of its data through a variety of means, including public-use data files and the agency's website, and has policies to guide what data to archive and how.

Cooperation with Data Users

An effective statistical agency consults with a broad range of users of its data in order to make its products more useful. It seeks advice on data concepts, methods, and products in a variety of formal and informal ways, from data users as well as subject-matter and other experts. It endeavors to meet the needs for access to data while safeguarding the confidentiality of individual responses. And it strives to provide easy and equal access to all potential users.

Fair Treatment of Data Providers

A statistical agency must maintain a relationship of respect and trust with its data providers. It does so through fair information practices, which include protecting the confidentiality of data identified with an individual, whether the data are collected directly or obtained from administrative record sources; informing respondents of the purposes of data collection and the expected uses of the information; informing those asked

to report whether their participation is voluntary or mandatory, and using appropriate informed consent procedures for voluntary data collections; minimizing the contribution of time and effort asked of respondents, which may be achieved in seeking the advice of respondents in the planning and design of data collection; and respecting the willingness of respondents to contribute to society by representing accurately the information they provide, and by making it available for important uses.

Research

An active research program is integral to the activities of an effective statistical agency. Research may be conducted on the substantive issues for which the data were compiled, especially concepts and definitions; methods of measurement, collection, analysis, and dissemination; and topics to improve the utility of the agency's data in policy analysis and decision making.

Quality and Professional Standards

An effective statistical agency is committed to the highest standards of quality work and professional practice and ethics, in statistics and other relevant professions. A statistical agency should recruit, develop, and retain professional staff committed to high standards of quality, in particular through strong and continuous relationships with appropriate professional organizations in statistics and relevant subject-matter areas.

Evaluation

Statistical agencies must continually evaluate their statistical programs for accuracy, relevance, timeliness, efficiency, and cost-effectiveness. Regular formal reviews should consider ways to reduce duplication through effective uses of administrative records and other data sources and to adjust priorities to meet current policy needs. Evaluations should include both internal reviews by staff and external reviews by independent groups.

Coordination and Collaboration

In the US context, an effective statistical agency continually seeks opportunities to cooperate with other statistical agencies to enhance the value of its own and other agencies' information. Statistical agencies cooperate, not only with each other, but also with state and local statistical agencies in the provision of data for subnational areas. They cooperate, too, with foreign and international statistical agencies to exchange information, on both data and methods, and to develop common classifications and procedures to promote international comparability of information.

Integrating data collected or compiled by different statistical agencies invariably requires much coordination and cooperation among the agencies, each of which have different missions. But the rewards are data more relevant to policy concerns, and a stronger statistical system as a whole. For these reasons, statistical agencies must cooperate as partners with one another, not only on the development of data, but also on the entire panoply of statistical activities, including definitions, concepts, measurement methods, analytical tools, professional standards, dissemination modes, ways to advance the effective use of their data, and means to protect the confidentiality of responses.

Analysis and Interpretation

Agencies can make their data more relevant to public policy concerns by analyzing and interpreting the data. Care must be exercised, however, that analyses are not perceived as policy advocacy. Analyses of data also reveal where improvements in data are needed.

Advancing Effective Use of Statistics

Statistical agencies should also advance the uses of their data through making them easily accessible to users in convenient forms, involving data users in planning data collection and dissemination, and providing full documentation and other forms of technical support to users.

Challenges Faced by Statistical Agencies

We live in a rapidly changing, increasingly interdependent world with a global economy. Many countries face profound demographic changes. Primarily fueled by technological advances, widespread changes are occurring in society, the environment, and international relationships. Governments struggle to keep up with the changes. They are concerned with the need for new property rights, laws governing new markets, new forms of businesses, and investments in new technology for education. They are concerned too about mitigating human error in transportation, communications, healthcare, and other complex, critical systems and protecting the security of these systems; avoiding harmful effects on the environment and sources of food; and guarding against the loss of individual privacy. And governments are also concerned with disparities in economic growth and the distribution of income and wealth, to which disparities in access to technology may contribute, since these disparities could increase tensions among social groups and nations.

An efficient, flexible, and modern information system is essential for a government to address these and other concerns. The greatest challenge facing any statistical agency is for it to keep up with the times. "Public policy analysis often cannot wait until new surveys are undertaken, but must rely on knowledge drawn from the existing statistical database" (Cook, 2000). Those in charge of government statistics must anticipate and prepare for future needs.

Other challenges that statistical agencies face are the decline in survey response rates in the USA and other developed countries (NRC, 2013a), making greater use of administrative records for statistical purposes (OMB, 2014a), and understanding when nontraditional methods, such as nonprobability sampling, may have utility (Baker et al., 2013).

Technology, which has induced so many rapid changes, itself can give statistical agencies the means to respond to these challenges. The agencies must not only incorporate new

technology into collecting, processing, analyzing, and disseminating government statistics, but they must also prepare for the plethora of new data in digital form.

Advances in information technology are also making it more and more difficult to protect confidentiality, and the widespread use of data analytics and experiments on 'big data' that are easily captured have changed the meaning of privacy in our digital age. If a statistical agency is to respect its data providers by putting their information to use in the best way to benefit society, stronger ways to protect confidentiality must be sought. Statistical agencies should lead in confidentiality research, including research on data encryption and in fostering legal and administrative means to protect confidentiality (*see* Confidentiality and Statistical Disclosure Limitations).

Conclusion

Government statistics are essential for a government to function effectively and to advance businesses and the well-being of its citizens. The nature of data as a public good provides an economic argument for a government to collect statistics, but there are many other reasons for it to do so. One is to assure accurate, timely, and credible information that its citizens may use to hold their government accountable on any issue. A statistical agency must be able to provide information relevant to current and future policy issues, and to establish credibility for itself and its data, in particular with a strong standing of independence. Finally, the agency must be respected and trusted by respondents who provide it with data and with all data subjects whose information it obtains. That respect and trust can be fostered by protecting the confidentiality of individually identifiable information and by putting the information it obtains to use in the best ways to benefit society, including research.

More information on the principles and practices of statistical agencies can be found in the National Research Council report (NRC, 2013b) and on the organization of US statistical agencies in Norwood (1995). Statistics for over 200 countries and areas of the world are compiled by the United Nations Statistics Division (2011). Links to the websites of statistical agencies in the USA are available at www.fedstats.gov.

See also: Census Microdata; Censuses, Population: Comparative International Aspects; Censuses: Current Approaches and Methods; Censuses: History and Methods; Confidentiality and Statistical Disclosure Limitations; Data Bases and Statistical Systems: Demography; Data Bases and Statistical Systems: Economics (General); Large-Scale Social Surveys, Analysis of; Longitudinal Data: Event–History Analysis in Discrete Time; Panel Surveys: Uses and Applications; Questionnaires: Cognitive Approaches; Sample Surveys: Cognitive Aspects of Survey Design.

Bibliography

Baker, Reg, Brick, J. Michael, Bates, Nancy A., Battaglia, Mike, Couper, Mick P., Dever, Jill A., Gile, Krista J., Tourangeau, Roger, 2013. Report of the AAPOR Task Force on Non-probability Sampling. American Association for Public Opinion Research, Deerfield, IL. Available: http://www.aapor.org/AM/Template.cfm?Section=Reports1&Template=/CM/ContentDisplay.cfm&ContentID=5963.

Cook, L., 2000. Statement by the First National Statistician on the Launch of National Statistics on 7 June. UK Office of National Statistics.

European Commission, January 25, 2012. Commission Proposes a Comprehensive Reform of the Data Protection Rules. European Commission Data Protection Newsroom. http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm.

European Commission, March 12, 2014. Progress on EU Data Protection Reform Now Irreversible Following European Parliament Vote. European Commission Memo/14/186. http://europa.eu/rapid/press-release_MEMO-14-186_en.htm.

Federal Interagency Forum on Aging-Related Statistics, 2012. Older Americans 2012: Key Indicators of Well-being. U.S. Government Printing Office, Washington, DC. Available: http://www.agingstats.gov/agingstatsdotnet/main_site/default.aspx.

Federal Interagency Forum on Child and Family Statistics, 2013. America's Children: Key National Indicators of Well-being, 2013. U.S. Government Printing Office, Washington, DC. Available: http://www.childstats.gov/americaschildren/index.asp.

National Research Council, 1993. Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics. Committee on National Statistics, George T. Duncan, Thomas B. Jabine, and Virginia A. De Wolf, Editors. National Academy Press, Washington, DC.

National Research Council, 2000. Improving Access to and Confidentiality of Research Data: Report of a Workshop. Committee on National Statistics, Christopher Mackie and Norman Bradburn, Editors. National Academy Press, Washington, DC.

National Research Council, 2013a. Nonresponse in Social Science Surveys: A Research Agenda. Panel on a Research Agenda for the Future of Social Science Data Collection. Committee on National Statistics; Roger Tourangeau and Thomas J. Plewes, Editors. Division on Behavioral and Social Sciences and Education. National Academies Press, Washington, DC.

National Research Council, 2013b. Principles and Practices for a Federal Statistical Agency. Committee on National Statistics, Constance F. Citro and Miron L. Straf, Editors. Division of Behavioral and Social Sciences and Education, fifth ed. National Academy Press, Washington, DC.

Norwood, J.L., 1995. Organizing to Count: Change in the Federal Statistical System. The Urban Institute Press, Washington, DC.

US Office of Management and Budget, 2014a. Memorandum for the Heads of Executive Departments and Agencies: Guidance for Providing and Using Administrative Data for Statistical Purposes. Sylvia M. Burwell, Director, OMB Memorandum M-14–06, 14 February 2014. Available: http://www.whitehouse.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf.

US Office of Management and Budget, 2014b. Statistical Programs of the United States Government, Fiscal Year 2014. US Government Printing Office, Washington, DC.

U.N. Statistical Commission, 1994. Fundamental Principles of Official Statistics. Official Records of the Economic and Social Council, 1994, Supplement No. 9. U.N. Statistical Commission, New York. Available: http://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx.

United Nations Statistics Division, 2011. Statistical Yearbook Fifty-Sixth Issue. United Nations, New York. Available: http://unstats.un.org/unsd/pubs/gesgrid.asp?id=489.

Governmentality

Marc Abélès, Ecole des Hautes Etudes en Sciences Sociales, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The introduction of the concept of governmentality at the end of the twentieth century has given a new impulse to the study of contemporary political power and institutions in a global context. It can be seen as the result of two converging intellectual movements. One emerged from the critique of the ethnocentric category of the state by the political anthropologists; the other comes from the philosophy, more precisely from the genealogy of neoliberalism proposed by Michel Foucault who imposed this concept in the field of social sciences.

The introduction of the concept of governmentality at the end of the twentieth century has given a new impulse to the study of political power and its institutions. It can be seen as the result of two converging intellectual movements. One is to be found in the anthropological literature, specifically in the critique of the ethnocentric category of the state; the other comes from the philosophy, more precisely from the genealogy, of neoliberalism proposed by Michel Foucault in the field of social sciences.

Stress has often been laid on the contrast between traditional societies in which the political sphere is embedded and the modern world in which the autonomy of politics is manifested in the organization of states and their many institutions. For a long time, anthropology was confined to exotic societies, and the counterpart to this preference for far off places was the creation of boundaries between universes perceived as having different ontological properties. Two contrastive approaches were thus adopted: one appropriate to an understanding in which it is difficult to separate the political from other aspects of reality, and the other applicable to contemporary societies, where politics as an institution is clearly circumscribed. This implicitly set limits on the work of anthropologists and gave sociologists and political scientists a monopoly on research on modernity. The division of labor has certainly had some positive effects, since it has enabled the different disciplines to delve very deeply into the areas assigned to them.

At the same time, the boundary that has been created could not resist for long the curiosity that anthropologists felt for their own societies, which eventually led them to expand their areas of investigation. It is true that the modern state appears to be very remote from the archaic structures and faltering institutions that once attracted anthropologists. With its complex civil service, dense bureaucratic fabric and deployment of hierarchies, the functioning of the state in western societies bears the little resemblance to the much more diffuse operation of politics in the exotic world. There is a real disparity of scale between contemporary state formations and the systems that the anthropologists have previously described such as segmentary societies or chiefdoms. But it has not prevented the development of an anthropology of the state focused on the everyday practices of the bureaucracy with its official procedures and routines (Sharma and Gupta, 2006). A number of ethnographic studies have offered relevant analyses of the functioning of administrative and repressive state apparatuses and the way in which the ideological state apparatuses realize "the interpellation of 'individuals' as subjects" (Althusser, 2006: p. 108).

One of the main contributions of anthropologists to the study of politics has been to demonstrate how the state's rise as a political system was contingent on a specific historical context. Anthropologists have thus suggested that the state is susceptible to eventual disappearance and replacement by other political constructions (Abélès, 2005). From a comparative perspective, they have clearly shown that the state is only one form of government among many others. In the era of globalization, with the emergence of new transnational political forms, the Hegelian preeminence of the state has been challenged, and one aim of political anthropology is to offer a better understanding of this great transformation. As a consequence of globalization, in a world where economic processes subvert frontiers, state action is more and more dependent on economic constraints (Appadurai, 1996; Strange, 1996; Trouillot, 2003) and the rules of financial markets. We simultaneously find the limits of the concept of sovereignty and the emergence of 'multilayered governance' (Rosenau and Czempiel, 1992) that seems closely connected to the rise of the information society (Castells, 1996). Even critical theories of contemporary capitalism have profoundly reworked political concepts, to which the post-Marxist vision of Hardt and Negri (2000) testifies, returning to categories of flow and deterritorialization (Deleuze and Guattari, 1980) in order to comprehend transformations in the mode of production that characterizes the contemporary world. These critics substitute the antagonism between empire and multitude for the Marxist class struggle, whereas the social movements that rise against the surveillance society have neither common enemies nor common languages.

The focus in writings about global politics differs as to the institutions that govern society. According to their orientation, authors emphasize either the emergence of new political forms or the continuity of traditional structures. But the dominant point of view is a top-down vision, one that accents the rationality or the irrationality of centralized political processes in the context of globalization. This vision, however, is partial in both senses of the term: it does not take into account the main aspects of current changes at the global level and is indifferent to new representations of political action in contemporary societies. The time has come for reconsidering the statist conception that oriented most of the political studies and that

does not afford a relevant framework in order to understand globalization and its consequences.

Long time ago, in his Preface to *African Political Systems*, Radcliffe-Brown (1940) considered that the state, as a concept, was an epistemological obstacle. From an anthropological perspective, the focal issue is that of governing. Here, one can point out the convergence between political anthropology and the political theory articulated by Michel Foucault for whom reference to the state was much more a theoretical fetish than an efficient instrument used to better understand the operation of what he termed 'governmentality.'

As Foucault wrote: "The analysis in terms of power must not postulate as given the sovereignty of the State, the form of law or the total unit of domination; these are none other than the terminal forms" (1976: p. 120). If political anthropology has contributed to dereifying centralized political authority by showing the degree to which a focus on the form of state control has ultimately obscured the real diversity of figures of power, Foucault, for his part, has underlined the need to ask the question 'how' and to examine the exercise of power. To think of power in action, as a "way of acting upon actions" (2000b: p. 341), was also to challenge the traditional tools of political theory, which "had recourse only to ways of thinking about power based on legal models, that is: What legitimizes power? Or they had recourse to institutional models, that is: what is the state?" (2000b: p. 327). That meant delegalizing and deinstitutionalizing the approach to politics. Foucault made history play the role of the operator, allowing him to take a distant view of contemporary power and bring its constituent elements to light. He concluded that, in the seventeenth and eighteenth centuries, a new mechanism of power was invented that applied to people rather than to land. On the contrary, earlier theory of sovereignty was linked to a power exercised foremost on the land and its products, and not on the bodies of subjugated individuals. In the older vision, power aimed to seize goods and wealth – not working hours: "This theory makes it possible for power to be based on and around the physical existence of the sovereign and not at all on and around continuous and permanent surveillance systems" (2001: p. 186).

This brings us to one of Foucault's major themes: the art of governing, or in other words the practices by which we manage to 'structure the possible field of action of others' (2001: p. 341). One of the difficulties encountered by any anthropology of power comes from the permanent overlap that exists between 'the art of governing' – the concrete procedures that define the field of power – and a theory of sovereignty that claims both to establish the legitimacy of these procedures and to extend their meaning. Foucault introduces the concept of governmentality in order to better understand the modalities and the complexity of power in action. Political rationality and technologies of government lend specific substance to 'governmentality' that has been defined as "the composition and assembling of actors, flows, buildings, relations of authority into relatively durable associations mobilized to a greater or a lesser extent, toward the achievement of particular objectives by common means" (Rose, 2011: p. 148). From an anthropological perspective, the main value of the concept of governmentality consists in disentangling power from the discourse of sovereignty. Analysis in terms of governmentality privileges the dynamics of power, and their specific productivity, rupturing the traditional associations of power and sovereignty.

Foucault never conceptualized governmentality in terms of repression or bans. Criticizing Reich (1945) and Legendre (1974), he argues clearly that these works reduce the processes of power to a legal problematic rooted in sovereignty, which implies, among other things, a conception of power as negativity, or as a producer of the forbidden ("power is what says no") (2001: p. 423). The shift Foucault implements leads him to emphasize the plasticity of power relations, showing to what extent they are themselves caught up in other types of multiform and local relations, which leads to the further notion of micropowers. In a conference given at the Bahia university in 1976 (Foucault, 2001: p. 183), he criticized Lévi-Strauss and the Durkheimian school for focusing their conception of power on the notions of law and of prohibition. He opposed to this tradition the works of the French anthropologist Clastres as promoting a different vision in which power is considered as a specific 'technology' (2001: p. 184).

In some way, the genealogy of neoliberalism that the French philosopher developed in his *Collège de France lectures of 1978–1979* (2004) is an attempt to show that neoliberalism is not only a new age of the economy characterized by the triumph of financial capital and free-market values, but moreover a reconfiguration of governmentality, now focused on the *bios*. Beyond bringing to light the emergence of disciplinary techniques centered on the individual and his body, Foucault also shows how a new political technology, dealing with men in their multiplicity – the population – was put in place at the end of the eighteenth century. This technology, which he calls biopolitics, deals with the population as both a scientific and political issue. The interest taken in demography, the development of public hygiene, the institutions of assistance and insurance, and the taking into account of the relationship between man and the environment define a new configuration where the disciplinary dimension fades in favor of a project aiming to prolong life and regulate biological mechanisms. With the rise in the power of capitalism, these disciplinary work technologies acquired less direct procedures that allowed power to take over men as a group of living beings. Man was no longer just subjugated in his individual singularity, he was also controlled as a specimen of a population of living beings: the population as an undivided entity was the new subject of biopolitics.

The problematic of biopower has directly inspired recent research in anthropology. With respect to globalization, a great deal of commentary has been made about the ever-increasing imbrications of the local and the global. There have also been essential interrogations of the paradigm of the nation-state, which dominated the twentieth century but which was also rocked by the emergence of new modes of infra- and supranational governmentality. It is the model of territorial sovereignty that is at stake here, and the findings converge with Foucault's analyses, which suggest that power no longer has any particular affinity with territory, but is exerted above all on multiplicities. In the transnational space delineated by globalization, mobility and moving around, or what Appadurai (1996) designates as flows, play an essential

role. This explains the recent political focus on questions of immigration, humanitarian catastrophes, and the problem of refugees.

It is not surprising that anthropologists use concepts of governmentality and biopolitics to discuss the 'deterritorialization' of the contemporary world, because they provide a new and historical perspective on complex current situations. Research on diaspora bears witness to this. One example is the case of the Chinese from Hong Kong, as Aihwa Ong (1999) has analyzed them. At the time of British domination, they were considered to be British Dependent Territory Citizens (BDTC), and had the right to travel to Great Britain, but not to reside there. Since the retrocession of Hong Kong to China, they have British National Overseas status, and benefit from what Ong calls 'partial citizenship.' They have always been considered to be people of color, in the same way as Afro-Caribbeans, but are treated particularly well because of their economic dynamism. In 1990, the British government modified its immigration policy in order to confer citizenship upon certain Chinese residents from Hong Kong. This action was first of all a way of avoiding panic linked to the retrocession to China, which could have brought about a long-term destabilization of the market. It was in this way that 50 000 Chinese nationals and their families, belonging to the social elite in Hong Kong, could obtain British nationality. Those that had been selected were still young, between 30 and 40, had a high level of education, and through their activities and networks participated in flows of transnational capital. For Ong, the successive flip-flopping of the British authorities with regard to their citizenship reflects the implementation of a biopolitics that treats various population groups differently according to their place in the economic system.

Another line of research opened up by reflections on modern governmentality concerns the condition of populations deprived of any territorial claims or institutional status, those who are reduced to a 'bare life' condition, according to Agamben (1998), who extends Foucault's reflections on the matter. It is this bare life, disconnected from civic context, which constitutes the 'matter' of the masses of refugees, who are noncitizen residents established either permanently or temporarily in the territories of industrialized states, in the ghettos of the cities, or in refugee camps. According to Agamben (1998), this condition prefigures a form of community with no territory and no borders, in which residents' current refugee status prefigures their future possibilities. Legally vague, as far as identity and citizenship are concerned, the status of the refugee, which is being applied to growing masses of populations, dismantles the state–nation–territory triptych, inherited from the classical age, and puts into play a new definition of the subject/sovereignty relationship.

The *bare life* theme has been a subject of discussion in anthropology recently. This interest is due to the fact that from Kosovo to Rwanda researchers have generated ethnography of the effects of conflict and violence. The experience of suffering, destitution, and the denial of humanity in these extreme situations (Malkki, 1995; Kleinman et al., 1997; Agier, 2002), and also in the context of the marginalization that victimizes those abandoned at the margins of society (undocumented immigrants, victims of social exclusion, etc.), has led to reflections on the discourse and the practices through which contemporary governmentality deals with life (Fassin and Memmi, 2004).

Other research focuses on modes of intervention of the international community within the humanitarian domain. Here the idea is to reflect on the emergence of political actors whose relative autonomy results in new tensions, because their initiatives tend to continuously destabilize existing sovereignties. Likewise, on the sidelines of conflicts, 'nongovernmental' organizations, insisting that they are rooted in 'civil society,' perturb and add other levels of complexity (Abélès, 2006). One does not have to be an expert on the subject to see to what extent the opposition between politics and civil society is not very operational, because NGOs that claim to represent constituents are directly involved in a power relation of their own and thus attempt to exert power, or a 'mode of action on actions,' to repeat Foucault's formula (1984: p. 316). It can be stressed that many NGOS are not only involved in humanitarian intervention, but are also acting against the neoliberal logics that intensify poverty, suffering, inequalities between the North and the South, and so on.

The research carried out by anthropologists on armed conflicts and humanitarian interventions is exemplary of their political character. Concerning Kosovo, Pandolfi (2002) clearly shows how humanitarian interventions involve real investments by transnational organizations, which develop specific strategies and attempt to extend their control over local institutions. In theory, they are considered to be apolitical, but their actions fall within the scope of a power relation they contribute to shaping, with the more or less self-interested help of groups at the local level. We see then a configuration of power coming into existence that is more and more deterritorialized, leading the anthropologist to qualify it as 'migrant sovereignty,' insofar as the traditional nation-state apparatus no longer has a direct hold on the real elements being profoundly modified by the conflict.

Taking a more global point of view, what is at stake are processes that oblige us to profoundly modify a fixed conception of politics articulated around the concept of sovereignty. By proposing an approach that could be designated as pragmatic, in which governmentality is defined as a 'group of actions on possible actions' Foucault opened up a fruitful path of inquiry. It can provide a better understanding of the fact that state organization is now contrabalanced by other sources of power at the global and local levels. This phenomenon – the displacement of politics – has been commented on in terms of the weakening of the state, characterized as a rising 'governance without government' (Rosenau and Czempiel, 1992). Such governance involves not only formal political actors, but also other players from civil society. In a world increasingly threatened by various kinds of turbulence (Rosenau, 1990) – including terrorism, the proliferation of nuclear weapons, and mafia-orchestrated trafficking – the need for transnational governance is gradually becoming obvious. The increasing power of organizations aiming to provide types of regulation on a larger scale is a symptom that leads Badie (1999) to raise the prospect of a 'world without sovereignty.' This scholarship highlights the new power that supranational organizations have acquired, as well as the appearance of 'nongovernmental' organizations that have succeeded in occupying center stage

in contexts where authority was hitherto only held by the diplomacy of states. Scholars suggest that world of governance will be more and more a set of common rules imposed by an authority whose legitimacy will be recognized on a global scale. These observations reflect the development of a transnational level of governance. At the same time, one can question the functionalist perspective that informs these studies, as if the emergence of global politics could be interpreted as a complexification of institutions necessarily responding to a new globalized order.

The displacement of politics, that is, the fact that state is no longer the only protagonist and that the Hegelian dyad of state/civil society has lost its centrality at a conceptual level, is not only a matter of institutions and of sovereignty. It is not limited to the appearance of a new political scene in which old institutional powers have been replaced by newer ones, more able to handle the scalar change of globalization. Those who seek to extend political anthropology into the field of emerging transnational political forms find themselves facing the following conceptual dilemma: either they will be won over by the dominant institutionalist theory or they will realize the necessity of taking into account the specificity of these 'unidentified political objects,' as the former president of the European Commission Jacques Delors once described the European Union. There has been a global redefinition of the meaning and aims of political action. This redefinition is not simply cognitive. It also shows up in modes of action, in the constitution of organizational and institutional forms, in the selection of issues for public debate, and in the construction of epistemic spaces where this debate will happen. In other words, the redefinition is a matter of governmentality, in its original Foucaultian meaning.

If the focus of Foucault's work on biopower is the genealogy of the modern state, its analysis of governmentality provided a very powerful conceptual instrument and original interpretation of the rise of the welfare state. Introducing a dichotomy between an Oriental tradition of political philosophy and a Greek one, he considers the welfare state as the result of the convergence of these two traditions: here a "tricky adjustment between political power wielded over legal subjects and pastoral power wielded over live individuals" (2000a: p. 307) has been implemented. In one case, there is a perception of power or the relationship of the government to the governed, as illustrated through the metaphor of the shepherd leading his flock. This pastoral theme is prevalent in ancient oriental societies. The other tradition, illustrated notably by Plato in *The Statesman*, sees in the art of governing the capacity to achieve the unity of the city – to 'weave the finest of fabrics' (cited in Foucault, 2000a: p. 307). The welfare state was an unusual moment in history in that it managed to bring the two perceptions together in a single, cohesive whole – the state as 'centralized and centralizing power,' oriented toward the citizen, and 'pastoralism' as the 'individualizing power' (2000a: p. 300), aimed at subjects.

Foucault's analysis also helps us to better understand the displacement of politics that characterizes the contemporary forms of governmentality. In the era of globalization, the harmony between pastoral power and the centralizing state has disappeared. It is as though our perception and experience of politics is becoming bifurcated between a focus on the citizen/individual and a focus on the biopolitical subject. In fact, we can speak of a real transition, with a rise in preoccupations of life and *survival* at the heart of political action, while the issue of the Platonic city and the relationship of the individual to sovereignty, what I call *convivance*, is relegated to the background. The displacement can be summarized in the simple question that we constantly ask ourselves: 'what will our world consist of tomorrow?' It is this fundamental anxiety that not only alters our political representations but that also determines the space that can be allocated to political activity and the new places in which it is best exercised.

The place occupied today by transnational institutions such as the WTO, the UN, and the IMF is not due merely to their functionality in the global system. In fact, we are witnessing a reconfiguration of the political domain that does not content itself with developing a new mode of biopolitical governmentality. This is why theories traditionally based on the dialectic between state and civil society focus on changes in political scale and space in search of an impossible return to a harmonious civil order in which the forces of society could recover control over increasingly distant and mysterious powers and remediate their imbalances and inequalities. These theoretical and political expressions share a vision of harmonious living together, while the actual political practices we encounter are unfolding in the context of survival. To use Michel Foucault's distinction, these practices fall within the 'pastoral' dimension of power and apply less to 'civil subjects' than to 'living individuals.'

If one can speak of global governmentality in reference to this constellation of transnational institutions and organizations, it is not in the sense of an inclusive reordering, of a new power overlaying existing powers. This vision of things is part of the convivance perspective. Its final realization is 'empire,' a concept that falls within the same paradigm as the philosophical conception of the 'state' – a form intended to organize the communal life of subject citizens. Yet, in societies where threat is an integral dimension of the present, questions of life and survival now appear much more fundamental – thus, ideas of balance and order, justice and law only make sense in the context of a precaution-oriented viewpoint (Beck, 1992).

It is no accident that the theme of sustainability is now so intimately tied to theories concerning inequalities and the relationship between power and society. The separation of living individuals and civil subjects highlighted by Foucault has, in our own historical context, led to the emergence of the issue of survival and its political expressions which, while it connects to the traditional domain of 'international relations,' actually far exceeds it. There is now a dense, internalized space of anthropological anxiety regarding an assortment of threats, ranging from terrorism to the planet's ecological vulnerability – threats to which the political form of the nation-state form has been unable to give satisfactory answers or to guarantee human rights.

Global politics today thus casts us into a regime of anticipation and signals a lack of fulfillment. Global politics can neither be conceptually contained in terms of power balances nor considered as a superstate form. It is developing alongside the world of territorial division, without imposing itself on states and on traditional regimes of sovereignty as an overarching power. Between the nation-state and global politics,

we find the same polarity that I have described above as the tension between a regime of convivance, which places harmoniously living together as the highest aim of social beings, and a regime of survival. Today, global politics remains, to a high degree, dependent on the strategies of nation-states. At the same time, the specific development of new forms of transnational governmentality is modifying the distribution of power and this situation presents a new horizon for political anthropology.

See also: Anarchism; Citizenship: Political; Comparative Method in Anthropology; Cosmopolitanism; Factionalism; Human Rights, Anthropology of; Immigration, Citizenship, and Integration: Social Work Connections; Industrial Societies; Liberalism, Anthropology of; Modernity: Anthropological Aspects; Nongovernmental Organizations: Anthropological and Historical Aspects; Political Anthropology; Refugees in Anthropology.

Bibliography

Abélès, Marc, 2005. Anthropologie de l'État, second ed. Petite Bibliothèque Payot, Paris.

Abélès, Marc, 2006. The Politics of Survival. Duke University Press, Durham, 2010.

Agamben, Giorgio, 1998. Homo Sacer: Sovereign Power and Bare Life. Stanford University Press, Stanford, CA.

Agamben, Giorgio, 2000. Means without Ends: Notes on Politics (Vincenzo Binetti and Cesare Casarino, Trans.). University of Minnesota Press, Minneapolis.

Agier, Michel, 2002. Aux bords du monde, les réfugiés. Flammarion, Paris.

Althusser, Louis, 2006. Ideology and ideological state apparatuses. In: Aradhana, Gupta (Ed.), The Anthropology of the State: A Reader. Blackwell Publishing, Malden, MA.

Appadurai, Arjun, 1996. Modernity at Large: Cultural Dimensions of Globalization. University of Minnesota Press, Minneapolis.

Badie, Bertrand, 1999. Un Monde Sans Souveraineté. Fayard, Paris.

Beck, Ulrich, 1992. Risk Society: Towards a New Modernity. Sage, London.

Castells, Manuel, 1996. The Rise of the Network Society. Blackwell, Cambridge, MA; Oxford, UK.

Deleuze, Gilles, Guattari, Felix, 1980. Mille Plateaux. Capitalisme et Schyzophrénie. Editions de Minuit, Paris.

Fassin, Didier, Memmi, Dominique, 2004. Le Gouvernement des Corps. Editions de l'EHESS, Paris.

Foucault, Michel, 1976. La Volonté de savoir. Gallimard, Paris.

Foucault, Michel, 1984. Polemics, politics and problematizations. In: Rabinow, Paul (Ed.), The Foucault Reader. Pantheon Books, New York.

Foucault, Michel, 1997. Il faut défendre la société. In: Bertan, Mauro, Fontana, Alessandra (Eds.), Cours au Collège de France, 1976. Hautes Études, Gallimard, Seuil, Paris.

Foucault, Michel, 2000a. Omnes et singulatim. In: Faubion, James (Ed.), Power: Essential Works of Foucault, 1954–1984, vol. III. The New Press, New York.

Foucault, Michel, 2000b. The subject and power. In: Faubion, James (Ed.), Power: Essential Works of Foucault, 1954–1984, vol. III. The New Press, New York.

Foucault, Michel, 2001. Dits et écrits II 1976–1988. Gallimard, Paris.

Foucault, Michel, 2004. Naissance de la biopolitique. In: Senellart, Michel (Ed.), Cours au Collège de France, 1978–1979. Hautes Études, Gallimard, Seuil, Paris.

Hardt, Michael, Negri, Antonio, 2000. Empire. Harvard University Press, Cambridge, MA.

Hardt, Michael, Negri, Antonio, 2004. Multitudes: War and Democracy in the Age of Empire. Penguin Press, New York.

Kleinman, Arthur, Das, Veena, Lock, Margaret, 1997. Social Suffering. University of Californa Press, Berkeley.

Legendre, Pierre, 1974. L'Amour du censeur. Essai sur l'ordre dogmatique. Le Seuil, Paris.

Malkki, Lisa, 1995. Purity and Exile. Violence, Memory and Colonial Cosmology among the Hutu Refugees of Tanzania. University of Chicago Press, Chicago.

Ong, Aihwa, 1999. Flexible Citizenship: The Cultural Logics of Transnationality. Duke University Press, Durham.

Pandolfi, Mariella, 2002. 'Moral entrepreneurs,' souverainetés mouvantes et barbelés. Anthropologie et Sociétés 26 (1), 29–51.

Radcliffe-Brown, Alfred Reginald, 1940. Preface. In: Evans Pritchard, E.E., Fortes, M. (Eds.), African Political Systems. Oxford University Press, London.

Reich, Wilhelm, 1945. The Sexual Revolution (1930, Die Sexualität im Kulturkampf.). FSG, New York.

Rose, Nikolas, 2011. Governing 'advanced' liberal societies. In: Sharma, A., Gupta, A. (Eds.), 2006. The Anthropology of the State: A Reader. Blackwell Publishing, Malden, MA, pp. 144–162.

Rosenau, James, 1990. Turbulence in World Politics. Harvester, New York.

Rosenau, James, Czempiel, E.O., 1992. Governance without Government: Order and Change in World Politics. Cambridge University Press, Cambridge.

Sharma, Aradhana, Gupta, Akhil (Eds.), 2006. The Anthropology of the State: A Reader. Blackwell Publishing, Malden, MA.

Strange, Susan, 1996. The Retreat of the State: The Diffusion of Power in World Economy. Cambridge University Press, Cambridge.

Trouillot, Michel-Rolph, 2003. Global Transformations: Anthropology and the Modern World. Palgrave Macmillan, New York.

Governments

Jean Blondel, European University Institute, Florence, Italy

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 9, pp. 6321–6323, © 2001, Elsevier Ltd.

Abstract

All human organizations have a government, but the most visible governments are those of nations and of public bodies in general. They are focal points of political life. They have three key functions, policy initiation, coordination, and implementation, although they are always at least somewhat constrained by forces in the society. They are typically compact, but vary widely in composition, organization, duration, powers: one of the crucial ways in which they differ is by the role of their leaders, be they presidents, prime ministers or, in a few cases, monarchs. Governments can be 'autocratic,' oligarchical, or democratic, democratic governments being essentially a development of the twentieth century: with these, parties have become involved, obviously if they are represented formally in the executive, singly or in coalition, but even if they are not. All governments, democratic or not, typically initiate policies, often under pressure of their leaders, whatever role parties and the 'civil society' may also exercise.

Governments are universal. They exist in every social organization: there is always a body, normally relatively small, which has the task of running that organization; but the most visible governments are those of nations and of other public institutions. With the decolonization process, independent national governments have come to rule practically the whole of the planet and they have more than doubled in number, from well under a hundred in the early 1950s to nearly two hundred at the beginning of the third millennium.

Problems of Definition

Being at the center of political life, national governments tend to be rather small and, often at least, compact. Yet it is not possible to circumscribe them neatly as their boundaries are not precise. In many cases, indeed most, an official list exists, but this list is typically not exhaustive, as key personalities, who may formally only be 'advisers,' but are in fact highly influential, tend not to be included. There are also often two or more 'levels' of the government: thus in the UK and in countries which have followed closely the British model, for instance, the 'cabinet' constitutes only an inner core, with 'ministers of state' and 'parliamentary secretaries' being below this core. In the USA, the government is typically referred to as the 'administration' or as the 'executive': however, both of these expressions remain imprecise with respect to those whom they cover.

It is therefore more appropriate to look for a definition in terms of the functions which national governments fulfill. In general, these are said to 'run the affairs of the nation,' although they only do this up to a point since they are helped and advised by a multiplicity of individuals and of groups, and, in modern democratic countries in particular, by parties and legislatures, while everywhere the bureaucracy is at their 'disposal.' One can in reality distinguish among three functions of governments. First, they have to elaborate realistic policies which can be implemented: this is the function of conception. Second, they have a function of implementation, as they must find the means of ensuring that the policies which have been elaborated are indeed turned into practice. There can be tensions between these two functions. The skills required to undertake each of them successfully are different: to conceive means to 'imagine'; to implement means to 'manage.' Yet governments have also a third function, that of coordination: they must ensure that the policies which are elaborated and implemented do not pull in different directions but are complementary.

Types of Government, Liberal-Democratic and Authoritarian

A 'revolution' in the characteristics of governments has occurred since the eighteenth century, when 'modern' government began. This 'revolution' has had two main aspects. First, there has been a massive increase in the complexity of governmental activities, as states have come to be increasingly involved in the economic and social life of countries, especially from the late nineteenth century to the 1980s. Second, many governments have come to be (or have had to be) responsive to the demands of citizens, in particular where the representative system had been introduced: for the first time in human history, these governments had to abide by rules based on liberal democratic principles and enshrined in constitutions. Admittedly, Greek cities, the Roman Republic, and, during the Renaissance, some parts of Italy, had experienced forms of 'limited' government; but only from the end of the seventeenth century did representative government take root in a nation, England, which hitherto had been a traditional monarchy, while, almost a century later, American independence marked the emergence of the first liberal Republic.

England and America were indeed the originators of the two main forms of constitutional government which have existed to this day, the cabinet system, which was gradually adopted in almost every Western European country and in the Commonwealth, and the presidential system, which spread to Latin America, though with only limited success. Thus, at the beginning of the third millennium, approximately half the

national governments could be described as being broadly responsive to their people. Meanwhile, in the rest of the world, including in many countries of Africa, of Asia and, at least in the past, especially up to the mid-1980s, of Latin America, governments have tended to be authoritarian. The Communist form was particularly harsh, both in what was the Soviet Union and in Eastern Europe up to the late 1980s, as well as in China and a number of other Asian countries: it was based on the dominance of a single party whose leaders formed the real government and controlled the formal executive. Independence from colonial rule which started in the 1950s and gathered momentum in the 1960s, especially in Africa, also led to many dictatorships, sometimes based on single-party arrangements but relying even more on the 'charismatic' figure of a president, often drawn from the military and having come to power through a coup. Admittedly, since the 1980s, a move towards responsive governments occurred in many countries hitherto ruled by dictatorships, in Latin America, in what had been the Communist block, in East and Southeast Asia and in parts of Africa; but authoritarian rule continued to characterize many countries of the contemporary world, sometimes on the basis of old-fashioned monarchical governments but more often on 'modern' types of dictatorships, in many cases of the military type.

The Organization of Governments: Cabinet and Presidential Systems. The Role of Leaders

Partly as a result of these differences, governments vary widely in organization, duration, composition, and powers. The distinction between the cabinet system and the presidential system remains the main one among liberal democratic governments, although there are variations within each of these two models and some cases are intermediate. The cabinet system is based on the division between a symbolic head of state, often a monarch, and a government, responsible to parliament, headed by a prime minister: the prime minister has to operate in the context of a collegial system in which those members of the executive who are part of the cabinet (normally referred to as 'ministers') are entitled to participate fully and collectively in decision-making: in practice, this is far from being always the case and there are substantial differences from country to country and within each country. Some cabinets are controlled closely by their prime minister, while others are more egalitarian. By and large, the British government and the governments of many Commonwealth countries tend to be tilted toward prime ministerial dominance, while a number of continental European governments, at least those which are based on complex party coalitions, give more elbow room to ministers. Yet, even though cabinet members may be more or less collectively involved, there is a real sense in which the government constitutes a team. A variation on this model is provided by the French system, in which the picture is somewhat complicated by the fact that the president, elected directly by the people, plays a significant part in policy making, at least when the party of the president also dominates parliament: this model has been adopted by some of the post-Communist governments of Eastern Europe as well as in the developing world.

The presidential system, on the other hand, is based on the separation of powers principle, with both president and legislature elected directly by the people. The president, as head of the executive, has thus greater leeway to organize the government in the way he/she wishes: this is particularly true in the USA. In some Latin American countries, on the other hand, the president has to an extent to take into account the desire of party leaders to play a part in the appointment of members of the executive. However, by and large, presidential governments are both constituted more independently from the parties and in particular from the parties in the legislature. They are also more hierarchical in character: the members of the government (typically known as 'secretaries') are responsible to the president for the department of which they are in charge (although in some cases the legislature can also censure them individually). They are not collectively involved in decision-making.

Governments are fashioned by their leaders. The twentieth century was characterized by the emergence of 'charismatic' rulers, including and perhaps above all in Communist countries, despite the fact that the Communist ideology is based on the notion that the economic 'substructure' is held to be paramount. 'Charismatic' or personalized leadership can also be found in liberal regimes, whether of the cabinet or of the presidential type. This has been the case not just of wartime leaders, such as Churchill or de Gaulle, but of rulers who attracted the imagination of whole populations, such as Kennedy.

The Duration of Governments

The duration of governments varies. The longest are those of the few remaining traditional regimes, in effect absolute monarchies, and those of some of the more 'modern' dictatorships: Communist governments, their leaders in particular, often remained in power for over a decade; on the other hand, most other leaders are only in office for a few years. The other government members often stay for even shorter periods: the average duration of ministers or secretaries is in the region of three to four years, with a substantial proportion of them being truly transient and remaining in government for a year or less.

Career Patterns

The composition of governments also varies. Some of these, primarily where there is a cabinet system, include almost exclusively professional politicians, men and women who have been active in a political party, have then been elected to parliament and have joined the ranks of the executive after a few years on the 'backbenches.' Other government members are 'technicians,' more often drawn from the public sector than from the private sector: their career in politics tends to last a few years only and they return to other public or private jobs after their stint in government. Moves of this kind occur less frequently in cabinet systems than in presidential systems, where, as we noted, the head of the state has more freedom to pick and choose whom he or she wants to appoint; they also occur in many authoritarian governments, though, in these, military men are often side by side with civilian 'managers.'

Variations in career patterns are associated with differences in social background. In general, except where social democratic parties are frequently in power, as well as in the past Communist systems, men and women with a working class background are rare: lawyers, teachers, and managers form the bulk of the civilian members of governments. Only in the last decades of the twentieth century and in some countries, mainly Northern European, have women had more than a token share of governmental posts, especially in such key departments as those of foreign, economic, or internal affairs.

The Impact of Governments: An Unresolved Question

There is debate about the extent to which leaders and governments affect the countries which they rule. As their careers are often short, it is suggested that members of the higher civil service, together with the lobbies which gravitate around them, play a key part in decision making. As a matter of fact, it has not been possible so far to determine precisely the impact of leaders and members of governments even when they stayed in power many years: environmental factors, both internal and external to the country concerned, must be taken into account. Yet some leaders at least, whether 'harsh,' such as Lenin or Stalin, or 'good,' such as Roosevelt or Churchill, had a major influence on the destinies of the countries which they ruled and indeed of the world. What can only be said is that there are massive variations in impact, as there are major variations in the characteristics of governments and in particular in their responsiveness, organization, duration, composition, and complexity.

See also: Attitudes, Political: Authoritarianism and Tolerance; Liberalism: Historical Aspects; Liberalism: Political Doctrine and Impact on Social Science; Parliamentary Government; Parliaments, History of; Presidential Government.

Bibliography

Barber, J.D., 1977. The Presidential Character, second ed. Prentice-Hall, Englewood Cliffs, NJ.
Bass, E.M., 1981. Stogdill's Handbook on Leadership. Free Press, New York.
Blondel, J., Muller-Rommel, F. (Eds.), 1993. Governing Together. Macmillan, London, UK.
Blondel, J., 1982. The Organisation of Governments. Sage, London.
Blondel, J., 1987. Political Leadership. Sage, London.
Bunce, V., 1981. Do New Leaders Make a Difference. Princeton University Press, Princeton, NJ.
Burns, J.M., 1978. Leadership. Harper & Row, New York.
Cammack, P., Pool, D., Tordoff, W., 1988. Third World Politics. Macmillan, London.
Castles, F.G., Wildenmann, R. (Eds.), 1986. Visions and Realities of Party Government. De Gruyter, Berlin.
Castles, F.G. (Ed.), 1982. The Impact of Parties. Sage, London.
Cronin, T.E., 1975. The State of the Presidency. Brown, Boston: Little.
Finer, S.E., 1962. The Man on Horseback. Pall Mall, London.
Heclo, H., 1977. A Government of Strangers. Brookings Institution, Washington, DC.
Hook, S., 1955. The Hero in History. Beacon Press, Boston.
Kellerman, B. (Ed.), 1984. Leadership. Prentice-Hall, Englewood Cliffs, NJ.
Neustadt, R.E., 1960. Presidential Power. Wiley, New York.
Rejai, M., Phillips, K., 1983. World Revolutionary Leader. Sage, London.
Shugart, M.S., Carey, G.M., 1992. Presidents and Assemblies. Cambridge University Press, New York.
Strom, K., 1990. Minority Government and Majority Rule. Cambridge University Press, New York.
Weber, M., 1968. Economy and Society, 3 Vols. Bedminster Press, New York.

Grammatical Agreement

Dieter Wunderlich, University of Düsseldorf, Germany; and Centre for General Linguistics (ZAS) at Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In grammatical agreement, a governing predicate (verb), the members of a noun phrase, or an anaphorically dependent pronoun can bear similar ϕ-feature values (person, number, gender) as the controlling noun. Syntactic theories assume a narrow association of case and agreement (disproved in Indo-Aryan), while lexicon-based accounts work with referential indexes in the syntax–semantics interface. In long-distance agreement, a matrix verb is linked to an object noun occurring in a dependent clause; all theories assume some specific local mechanism for the matrix verb to have access to the dependent-clause noun. Finally, some constraints on agreement are considered.

Introduction

Grammatical agreement is a phenomenon in which co-occurring word forms have similar or the same values of number, gender, and person related to some entity. So all words in (1a) are singular (sg), while all words in (1b) are plural (pl); moreover, several words in (1a) are feminine (f), while the corresponding words in (1b) are masculine (m). (The values that a word form bears are indicated by superscripts.) The examples in (1) to (4) show some common agreement relations.

1. French
   a. La^{fsg} fillette^{fsg} est^{3sg} belle^{fsg}.
      'The girl is beautiful.'
   b. Les^{pl} garcons^{mpl} sont^{3pl} beaux^{mpl}.
      'The boys are handsome.'
2. Italian
   a. Tu^{2sg} l'^{3sg} hai^{2sg} vista^{fsg}.
      'You have seen her.'
   b. (io^{1sg}) li^{3mpl} ho^{1sg} visti^{mpl}.
      'I have seen them (masc).'
3. Finnish
   a. Taina löysi minun$^{1sg}{}_{GEN}$ kirja-ni^{1sgPoss}.
      'Taina found my book.'
   b. Taina löysi heidän$^{3pl}{}_{GEN}$ kirja-nsa^{3Poss}.
      'Taina found their book.'
4. German
   a. Ein kleines^{nsg} Mädchen^{n} sass^{3sg} auf der Treppe, und es^{nsg} lachte.
      'A little girl sat on the stairs, and she laughed.'
   b. Ein kleiner^{msg} Junge^{msg} stand^{3sg} daneben, und er^{msg} heulte.
      'A little boy stood beside, and he cried.'

In (1), the definite article agrees with the noun, and both the copula and the predicative adjective agree with the subject. If one item of (1a) is exchanged by the corresponding item of (1b), the sentence becomes ungrammatical (e.g., **la fillette sont belle*). (2) shows two agreement relations intermingled: The auxiliary agrees with the subject pronoun, while the participle agrees with the preceding object clitic. (3) is an example where the possessed (Poss) noun agrees with the genitive (GEN) possessor. In (4), the pronoun in the second clause agrees with the antecedent noun phrase in the first clause. Note that *Mädchen* 'girl' is a neuter (n) noun, triggering a neuter pronoun, although it refers to a female being; in contrast, English pronouns are always selected on semantic grounds.

Grammatical agreement comes as an interplay of morphological, syntactic, and semantic aspects (Anderson, 1998; Lapointe, 1988; Corbett, 1994, 1995, 1998; Pollard and Sag, 1994). Morphologically, some words are inflected for agreement. Syntactically, constituents stand in the agreement relation if the words they consist of bear corresponding information. Semantically, agreement serves to keep record of discourse referents: Only constituents that relate to the same referent can overtly agree with each other (Lehmann, 1982). The sentences in (1) are about a single person or a group of persons, while those in (2) are about two referents standing in a 'see'-relation.

Siewierska (1999) distinguishes three types: "(1) Anaphoric agreement markers are in complementary distribution with free nominal or pronominal arguments; (2) ambiguous agreement markers occur obligatorily both in the presence and absence of free nominal or pronominal arguments; (3) grammatical agreement markers are obligatory, like ambiguous markers, but, unlike ambiguous or anaphoric markers, must necessarily be accompanied by overt nominal or pronominal arguments." Accordingly, Italian shows ambiguous agreement, while German shows grammatical agreement.

Generally, agreement relations can be found among the following elements:

1. (noun phrase (NP)-/determiner phrase (DP)-internally) determiner and noun, attribute and noun, possessor and possessed noun;
2. (clause-internally) a predicate and one of its arguments: verb–subject, verb–object, preposition–object, predicative-argument;
3. (cross-clausally) a pronoun/anaphora and its antecedent.

The categories that appear in these relations are called ϕ-features, including person (first, second, third, inclusive = 1 + 2), number (singular sg, plural pl, dual), and gender (e.g., feminine f, masculine m, neuter n).

Nouns and pronouns (including pronominal affixes and clitics) function as the controller (source or probe) of an agreement relation in virtue of their features. Nouns are inherently classified for gender (or noun class), and sometimes also for number (cf *scissors*, *trousers*, *mice*, *men*, *brethren*), whereas pronouns are inherently classified for person and number, and sometimes also for gender. The inherent classification is often semantically

 http://dx.doi.org/10.1016/B978-0-08-097086-8.53034-7

based: Female beings are represented by feminine nouns but male beings by masculine nouns; aggregates are represented by plural nouns; addressees are represented by second-person pronouns, while speakers by first-person pronouns. Otherwise, feature assignment can be purely grammatical (obvious in genders such as French *lune*^{f}, *soleil*^{m} versus German *Mond*^{m}, *Sonne*^{f} (both 'moon, sun,' respectively)), and then even override the semantic default (as in German *Mädchen*^{n} 'girl').

The grammatical elements that agree with the noun or pronoun are said to be the agreement targets (=controllees); in (1a), the article *la* is the target of *fillette*, which is feminine, and both the copula *est* and the adjective *belle* are targets of the DP *la fillette*, which is fsg. The agreement features of controller and target do not have to be identical, but they must be compatible with each other (such as `3sg∪fsg⇒3fsg vs 3sg∪fpl⇒*`). In order for a complex DP to be the controller of an agreement relation, the features of the noun and the other DP-internal elements have to be unified. For short, ϕ-features index the referential elements of a discourse (pronouns and DPs), as well as the categories that predicate of these elements.

In the wide sense, one speaks about 'agreement' if, e.g., the verb bears information about arguments. Sometimes this information is clearly pronominal; sometimes it is less than that. In Italian, *ho*^{1sg} *vista*^{fsg} is possible alongside with *io*^{1sg} *l'*^{3fsg} *ho*^{1sg} *vista*^{fsg}, so that one can conclude the information is pronominal. In German, one cannot drop the pronouns; *habe*^{1sg} *gesehen* ('have seen') is clearly incomplete. Forms such as *ich*^{1sg} *habe*^{1sg}, *du*^{2sg} *hast*^{2sg} 'I have, you have,' however, are redundant because the information of the subject pronoun is doubled by the agreement morphology. Redundancy is one of the 19 criteria of canonical agreement offered by Corbett (2003a); in that sense, agreement in German is more canonical than agreement in Italian.

If the verb bears information about just one argument, it is usually about the nominative argument (the subject in nominative (NOM)–accusative (ACC) patterns, or the object in ergative (ERG)–absolutive (=nominative) patterns). That the Italian participle can bear information about the object has ultimately to do with the fact that these participles also occur in passive constructions.

The notion of 'agreement' is sometimes used in a still broader sense, including other types of correspondences as well, for instance, case concord in (5) and sortal correspondence in (6).

5. a. Latin: DP-internal case concord
   Illa^{fsg}-m$_{ACC}$ femina^{fsg}-m$_{ACC}$ bella^{fsg}-m$_{ACC}$ videbam.
   'I saw that beautiful lady'
   b. Finnish: DP-internal case concord
   tuo-n$_{GEN}$ nariseva-n$_{GEN}$ tuoli-n$_{GEN}$ alla
   that squeaky chair under
   'under that squeaky chair'
   c. German: Case concord between the predicative noun and its argument
   Man nannte ihn$_{ACC}$ einen Idioten$_{ACC}$; er$_{NOM}$ wurde ein Idiot$_{NOM}$ genannt.
   'He was called an idiot'
6. The horse neighs; the dog barks; the cock crows.

These relations, however, differ in many respects from agreement described above, and, therefore, are not included here. Case reflects the grammatical role of a constituent (subject or object), rather than its referential status, whereas sortal correspondence is based on much finer semantic distinctions than ϕ-features can offer.

Theoretical Accounts

Several theories have been launched to account for agreement phenomena. In a first systematic attempt, the constituents introduced by a phrase structure rule, such as S → DP + VP, are indexed with the same set of ϕ-feature values:

7. a. $S \rightarrow DP^{\phi i} + VP^{\phi i}$ (subject + verb phrase)
   b. $DP^{\phi j} \rightarrow D^{\phi j} + NP^{\phi j}$ (determiner + noun phrase)

This account does not allow the instantiation of different (although still unifying) feature values, as in the French example (1), in which one constituent is fsg and another one is 3sg, and it does not account for anaphora–antecedent agreement, because these elements do not originate from a common syntactic node. On the other hand, this account is not specific enough, because it would describe case concord with the same means. Moreover, the claim that ϕ-feature values distribute along with phrase structure rules seems rather arbitrary.

Notice that (7) is symmetric between controller and target. In contrast, general phrase structure grammar proposed an asymmetric analysis: Controller elements can inherit a feature structure AGR lexically, whereas target elements (mostly the functors) receive such a structure only by morphological affixation; the distribution of AGR-features within a clause is then checked by a general control-agreement principle (Gazdar et al., 1985). Again, this theory is both too general and too restricted with respect to agreement.

In the Chomskyan framework of *Principles and Parameter Theory*, later replaced by *Minimalist Syntax* (Chomsky, 1995), the bearer of agreement features are functional categories represented by independent syntactic projections (Laka, 1988; Pollock, 1989; Chomsky, 1991). Consider the schematic structure in (8). This structure describes subject–verb agreement when the AGR-node is identified with AGRS, which is specified for certain ϕ-feature values, as well as for case features (such as NOM). In order to get marked for subject agreement morphologically, the verb stem must move to AGRS, while the subject DP moves into SPEC.AGRS, where it receives nominative case (both movements leave a trace, which is coindexed in (8)). Similarly, AGRO is responsible for object–verb agreement as well as accusative case, and AGRPOSS is responsible for possessor–noun agreement as well as genitive case. Regardless of various modifications in detail, the basic idea remained the same: There is one syntactic configuration that determines both case and agreement. Agreement, as well as case, is checked between AGR and

SPEC.AGR (and both must be licensed by a particular argument role of the verb or noun).

8.

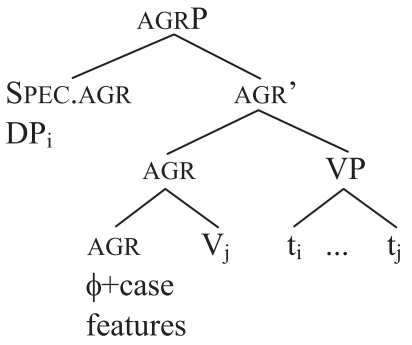

This configuration deals with agreement as an additional structure superimposed on argument structure (such as subject–verb, object–verb, or possessor–noun). It is unclear what additional structure would account for agreement with attributes, adverbs, secondary predicates, or dependent objects. More recently, Chomsky (2000, 2004) proposed to dispense with additional AGR structures and syntactic Spec–Head relations. Agreement is now assumed as a relation between a probing head and a target goal, which is in the probe's c-command domain. Note that the terminology has been shifted: Now the controller (say, a subject NP) is called target, and the controllee (the verb) is called probe. TENSE rather than arbitrary AGRS is the category responsible for subject agreement, and little *v* (which c-commands VP) for object agreement. The probing head bears unvalued ϕ-features that must find their values within the respective domain.

For most syntacticians, agreement is a local phenomenon and part of narrow syntax. The cross-sentential anaphora–antecedent relation certainly is nonlocal and cannot be captured syntactically. There are more instances of nonlocal ('long-distance') agreement between matrix and embedded clauses, where it is argued that they can be handled locally (see below). In view of these more complex phenomena, Bobaljik and Wurmbrand (2005) come to the conclusion that "agreement is with a (LF) phasemate" (where a phase is something like a clause and LF is 'logical form'); thus, for these authors, agreement is postsyntactic (see also Bobaljik, 2008), a position that is not so far from lexically based accounts.

Another problem for syntactic accounts comes from the assumption that agreement and case assignment are the two sides of the same coin. If a verb agrees with just one argument, more than often it is the nominative argument: In a nominative-accusative language, the verb agrees with the subject, and in an ergative-nominative ('absolutive') language with the object. Hindi is compatible with that correlation even if it is more complex in showing two types of differential realizations. Verbs select ergative for the subject only in the perfective, and accusative for the object only if that is definite or human specific. Thus, a transitive verb of Hindi can show ERG-ACC or ERG-NOM in the perfective, but NOM-ACC or NOM-NOM in the imperfective. The agreement rule is 'agree with the highest nominative' (which is the subject in the imperfect but the object in the ERG-NOM pattern); there is no agreement in the ERG-ACC option. Thus, Hindi very clearly exemplifies a narrow correlation between case and agreement.

But other Indo-Aryan languages (with millions of speakers) behave totally differently, even if they have similar case patterns. The Nepali verb agrees with the ergative subject in an ERG-NOM pattern (9a, Deo and Sharma, 2006), while in the opposite, the Kutchi-Gujarati perfective verb agrees with the focused accusative object in a NOM-ACC pattern (9b, Patel, 2007), i.e., contrary to what is expected, the verb agrees here with the marked rather than the unmarked argument in terms of case. There obviously is no case-agreement correlation in general. Wunderlich (2012) proposes an analysis where case and agreement, based on the same argument ranking, are determined by different constraint rankings (such as *ERG ≫ ERG! or ERG! ≫ *ERG; AGR(S) ≫ AGR(O) or *AGR(S)/*context* ≫ AGR(S) ≫ AGR(O), etc.). Nepali combines ERG! ≫ *ERG with AGR(S) ≫ AGR(O), while Kutchi-Gujarati combines *ERG ≫ ERG! with *AGR(S)/+perf ≫ AGR(S) ≫ AGR(O) (ERG! = realize ergative; *ERG = do not realize ergative, etc.).

9. Agreement with case-marked arguments

a.	Mai-le mero	luga:	dho-en.	*Nepali*
	I-ERG my	clothes.NOM	wash-PERF.1sg	
	'I washed my clothes'			
b.	Reena	chokra-ne	mar-ya.	*Kutchi-Gujarati*
	Reena.f.NOM	boys-ACC	hit-PERF.mpl	
	'Reena hit the boys'			

The relation between controller and controllee is asymmetric, as well as the (reversed) relation between probing head and goal. Yet, the agreement relation itself is symmetric: Agreement holds if the pieces of information on the controller and on the head are compatible. Thus, a unification-based approach might be promising.

Following the idea that ϕ-features keep record in what the individual referents are predicated of, agreement is seen as a mechanism that rests on the normal composition procedures; there is no need for a special syntax of agreement. In this spirit, agreement was analyzed as a syntax–semantics interface phenomenon in *Head-driven Phrase Structure Grammar* (HPSG; Pollard and Sag, 1994) as well as in *Lexical Decomposition Grammar* (LDG; Wunderlich, 1994, 1997). In HPSG, all grammatical information of lexical items or more complex forms is encoded by means of feature structure; the ϕ-features that are shared in an agreement relation (regardless of whether they are lexically inherent or morphologically introduced) form a substructure INDEX (as part of CONTENT), and when a clause is composed, the relevant INDEX structures are unified.

In LDG, all morphological and phrasal composition is determined by the theta structure of lexical items, constituted as a list of lambda abstractors related to the semantic form (a partial semantic representation). In this theory, ϕ-features are associated with (indexed to) individual variables under a binder, such as the λ-abstractor or the Russellian iota operator 'ι'; in the process of composition, these indexes are unified at the common binder of the variables in question. This is illustrated in (10), representing the composition of Latin *illa femina bella* 'that beautiful woman' from three individual entries; unification is represented by the sign '∪.'

10. D: *illa* — $\lambda P\ \iota z^{fsg}\ \{\text{DISTAL}(z)\ \&\ N(z)\}$
N: *femina* — $\lambda x^{fsg}\ \text{WOMAN}(x)$
A: *bella* — $\lambda y^{fsg}\ \text{BEAUTIFUL}(y)$
NP: $\lambda x^{fsg \cup fsg}\ \{\text{WOMAN}(x)\ \&\ \text{BEAUTIFUL}(x)\}$
DP: $\iota z^{fsg \cup fsg \cup fsg}\ \{\text{DISTAL}(z)\ \&\ \text{WOMAN}(x)\ \&\ \text{BEAUTIFUL}(x)\}$
'that entity (referred to by fsg) that is distal, a woman, and beautiful'

Long-Distance Agreement

In the Italian example (11), the subject agrees locally with the auxiliary, and nonlocally with the participle embedded in a dependent infinitive clause. The agreement facts thus indicate that 'have fear' and 'be arrived too late' predicate of the same individual referent. If one assumes that *ha paura* is a subject-control verb, the identity of arguments follows from the lexical entry of control verbs; by functional composition, the argument variable of 'be arrived too late' then fuses with the argument variable of *ha paura*, and the ϕ-features must unify, as shown in (11c).

11. LDA with Italian participles

a. Giovanni^{3msg} ha^{3sg} paura di essere arrivato^{msg} troppo tardi.
G. has fear to be arrived too late
'Giovanni fears to have arrived too late'

b. I ragazzi^{3mpl} hanno^{3pl} paura di essere arrivati^{mpl} troppo tardi.
'The boys fear to have arrived too late'

c. ha paura — $\lambda P\ \lambda x^{3sg}\ \text{FEAR}(x,P(x))$
di essere arrivato troppo tardi — $\lambda y^{msg}\ \text{ARRIVED_TOO_LATE}(y)$
$\lambda x^{3sg \cup msg}\ \text{FEAR}(x,\text{ARRIVED_TOO_LATE}(x))$
'the property (of 3msg entities) of fearing to have arrived too late'

Some of the elements may be underspecified for f or sg, but any other specification would lead to ungrammaticality (such as **illa femina bellus*^{msg}). Notice that the indexed information becomes relevant only when semantic composition takes place; therefore, this account indeed locates agreement at the syntax–semantics interface. Index unification itself is symmetric; all asymmetries result from the syntactic categories.

This approach can easily be extended to all other kinds of agreement, taking into account the specific mode of composition. If one forms the assertion *illa femina bella modesta est* 'that beautiful woman is modest,' one predicates 'be modest' ($\lambda u^{fsg}\ \text{MODEST}(u)$) of the result of (10), and if one continues with the pronoun *ea*^{fsg} 'she,' *illa femina bella* is a possible antecedent (referring to the same person) because the ϕ-features are compatible.

Another language well known for long-distance agreement (LDA) is Hindi (Mahajan, 1989; Wunderlich, 1994; Boeckx, 2004; Bhatt, 2005; Chandra, 2007). An imperfective control verb such as *caah-t-* 'want' agrees (together with a possible auxiliary) with the nominative subject (12a), while the perfective variant agrees with the embedded nominative object (12b) because ergative on the subject blocks agreement with the subject. There are two further possibilities, namely, that no agreement takes place in the perfect and the verbs get the default ending msg (12c), or that the embedded verb agrees with the object in the imperfect (12d, not available in all dialects). The respective acceptability rates are 95%, 88%, 78%, and 67% according to a study by Bornkessel-Schlesewsky et al. (2008).

12. LDA with Hindi control verbs

a.	Raam^{msg}	roTii^{fsg}	khaanaa^{msg}	caahtaa^{msg}	thaa^{msg}	
	Ram	bread	eat.INF	**want**.IMPF	**be**.PAST	
	'Ram habitually wanted to eat bread'					
b.	Raam ne	roTii^{fsg}	khaanii^{fsg}	caahii^{f}	thii^{fsg}.	LDA
	Ram ERG	**bread**	**eat**.INF	**want**.PERF	**be**.PAST	
	'Ram had wanted to eat bread'					
c.	Niinaa^{fsg} ne	roTii^{fsg}	khaanaa^{msg}	caahaa^{msg}	thaa^{msg}.	
	Ram	ERG bread	eat.INF	want.PERF	be.PAST	
	'Ram had wanted to eat bread'					
d.	Raam^{msg}	roTii^{fsg}	khaanii^{fsg}	caahtaa^{msg}	thaa$^{msg.}$	
	Ram	**bread**	**eat**.INF	**want**.IMPF	**be**.PAST	
	'Ram habitually wanted to eat bread'					

Under what condition the option b or c is selected is controversially discussed. According to Chandra (2007), LDA is possible only with specific objects, while Bhatt (2005) states that the embedded object can get scope over the matrix verb only when LDA takes place. Thus, 'Naim wanted to read every book' with no agreement only allows the reading *want* > ∀, while the LDA version b allows both readings *want* > ∀ and ∀ > *want* (with the latter being preferred). Bobaljik and Wurmbrand (2003) observed a similar pattern in Itelmen; see (16c) below.

This suggests that LDA is possible by reconstruction, where a verb complex is formed and the argument structure of the lowest verb is inherited to the result. The analysis of (12b) is sketched in (13). The suffix *-ii* in *khaan-ii* instantiates the feature complex fsg, which may be associated with either argument (the association with the subject 'u' being the default), whereas the perfect form *caah-ii* instantiates f, which cannot be associated with the argument role assigned ergative. In this context, only the second alternative of *khaanii* can survive because of unification.

13.

khaanii	*caahii*
$\lambda y\ \lambda u^{fsg}$ EAT(u,y)	
$\lambda y^{fsg}\ \lambda u$ EAT(u,y)	$\lambda P^{f}\ \lambda x$ WANT(x,P(x)) ERG

$\lambda y^{fsg \cup f}\ \lambda x$ WANT(x,EAT(x,y))
ERG

Other types of verbs in Hindi that make LDA possible by forming verb complexes are raising-to-subject verbs like 'seem' in (14a), raising-to-object verbs like 'see' in (14b), light verbs like 'let' in (14c), and modals with a lexically marked subject like 'should' in (14d). Here, again, ergative/accusative blocks subject agreement, and a verb complex is formed.

14. More LDA verbs in Hindi

a.	Raam ne	roTii^{fsg}	khaayii^{f}	lagtii^{f}	thii^{fsg}.	
	Ram-ERG	bread	eat.PERF	seem.IMPF	be.PAST	
	'Ram seemed to have eaten bread.'					
b.	Raam ne	billii^{fsg} aatii^{f}	dekhii^{f}	thii^{fsg}.		
	Ram-ERG	cat	come.IMPF	see.PERF	be.PAST	
	'Ram had seen a cat coming.'					
c.	Nadia-ne	sarosh-ko	gaaRii ^{fsg}	chalaa-ne	d-ii ^{fsg}	
	Nadia-ERG	Sarosh-ACC	*car-F.SG*	drive-GER.OBL	*let-PERF.F*	
	'Nadia let Sarosh drive car.'					
d.	Raam-ko	davaaii^{fsg}	khaa	lenii^{f}	chaahiye	thii^{fsg}.
	Ram-ACC	medicine	eat	take.INF	should	be.PAST
	Ram should have taken medicine.					

In principle, LDA is recursive: The following sentences, although not usual, are grammatical (Shravan Vasishth p.c.).

15. a. Raam ne larki^{fsg} aatii^{f} dekhnii^{fsg} chaahii^{f}.
Ram-ERG girl come.IMPF see.INF want.PERF
'Ram wanted to see a girl coming.'
b. Raam-ne har kitaab^{fsg} parhnii^{fsg} chaahii^{f} lagtii^{f} hai.
Ram-ERG every book read.INF want.PERF seem.IMPF be.PRES
'Ram seemed to have wanted to read every book.'

LDA is found in many more languages. In Basque, the auxiliary (which encodes up to three arguments) agrees with the number of the embedded object (16a, Etxepare, 2006). In Itelmen, a Chukotko-Kamchatkan language, the matrix verb agrees with person.number of the direct or indirect embedded object (16b,c, Bobaljik and Wurmbrand, 2003, 2005); (16c) shows the same scopal reading as found in Hindi. Note that the matrix verbs here are control verbs, i.e., restructuring verbs forming a verb complex.

16. LDA with restructuring verbs in Basque (a) and Itelmen (b,c)
a. [liburu-ak itzultze-ko] eskatu dizkidate
book-**pl**.NOM give-back asked have.3plERG.1sDAT.**3pl**NOM
'They asked me to give back the books.'
b. əntxa-βum = nIn kəma-nk nən^{l}č i zəl-es.
forget-**1sg** = 3.CL **me**-DAT fish INTERJ give-INF
'He forgot to give me (a) fish.'
c. t'-əntxa-ceʔn [miλ okno?-n sop-es]
1sg-forget-**3pl**.OBJ [all window-**pl** close-INF]
'I forgot to close all the windows.' (∀ > *forget*)

Another group of LDA languages includes the NE Caucasian Daghestan languages. In Tsez, the matrix verb agrees with the gender class of the embedded object (17a); Polinsky and Potsdam (2001) state the following conditions: "(i) The embedded clause must be in the absolutive position, (ii) the trigger must be the absolutive argument of the embedded clause, (iii) the absolutive NP must be the topic of the embedded clause." In Khwarshi, the matrix verb likewise agrees with the gender class of the embedded object, but under slightly different conditions: The embedded object must be either topic (17b) or focus (17c) and it must be moved into initial position (Khalilova, 2007). Note that 'know' is not a restructuring verb, because the embedded clause is finite. Just to illustrate what LF-'phasemate'-ship (Bobaljik and Wurmbrand, 2003) here means, (17a) might be paraphrased as 'the mother knows of the bread (=topic) that the boy ate.'

17. LDA with nonrestructuring verbs in Tsez (a) and Khwarshi (b,c)
a. eni-r [už –aa magalu b-aac' -ru-λi] **b**-iy-xo
mother-dat boy-erg bread.**III**.NOM III-eat-PASTPARTIC-NOML **III**-know-PRES
'The mother knows that the boy ate the bread.'
b. Zihe-n už-a-l b-iq'-še **b**-iti-xx-u.
cow(**III**)-TOP boy.OBL-LAT **III**-know-PRES III-divide-CAUS-PASTPARTIC
'The boy knows that the cow was stolen.'
c. K^{ʔ}aba zihe b-ot'uq'q'-u **b**-iq'-še už-a-l.
black cow(**III**) III-come-PASTPARTIC **III**-know-PRES boy.OBL-LAT
'The boy knows that the black cow has come.'

Despite theoretical differences, all accounts of LDA are agreed that apparent nonlocal agreement is in fact local. Either the matrix verb and the embedded verb form a verb complex in which subject and embedded object are clausemates (clause union in Hindi, Basque, Itelmen, etc.), or the embedded object is moved into a peripheral position (topic or focus) in which it is accessible to the next higher domain (in the spirit of Chomsky's (2000) edge condition) – which is the case in the Daghestan languages (17).

Since LDA in the sense of (16) always is lexically triggered by the matrix verb (note that Kutchi-Gujarati has just one lexical item of this sort, namely *par* 'have to,' according to Patel (2007)), lexically based accounts like HPSG and LDG seem especially suitable for the analysis of these constructions: There is something in these lexical items that makes them able to 'see into' a dependent clause. Furthermore, discourse categories like topic and focus (to be dealt with in discourse-semantics) seem to play an important role for connecting matrix and dependent clause. In contrast, syntactic accounts assuming that agreement is mediated by extra-syntactic AGR heads (like Mahajan, 1989) come into trouble to analyze LDA; therefore, the syntactic minimalism in Chomsky (2000) was a major step for approximating syntactic theory to lexical frameworks.

Constraints on Agreement

Most languages have the possibility of 'no agreement' in using a default verb form that otherwise could agree such as neuter, third person, or masculine singular. In these cases, the feature complex introduced by the verb form is not associated with an argument variable. Since in German only nominative forms can agree with the verb, agreement is possible in (18a), but is blocked in (18b), so that only the default form friert$^{\varnothing}$, lacking an index value, can appear.

18. a. Ich^{1sg} friere^{1sg} 'I am cold'
    b. Mich^{1sg} *friere^{1sg}/*friert^{3sg}/friert$^{\varnothing}$ 'I am cold'

Not all languages use agreement. Clearly, a language such as Chinese with nearly no morphology cannot have agreement, but also Malayalam (a Dravidian language spoken in Kerala, SW India), which shows rather rich derivational morphology and also morphological case, nevertheless does not have any agreement. Moreover, the extent to which agreement is used differs from language to language. Not only may the number of distinctions in the gender, number, and person categories, as well as the richness of inflectional affix inventories vary, it also depends on language-specific grounds in which contexts agreement is obligatory, optional, or even forbidden.

Hungarian marks plural on nouns, adjectives, and verbs. However, adjectives only agree with a noun when they are used predicatively (19a,b). Surprisingly, in the context of a numeral, which inherently contains the concept of plurality, the noun must not be marked for plural (19c). Furthermore, a predicative adjective or verb only agrees with a subject phrase that is overtly marked for plural, see (19d,e).

19. Hungarian plural agreement
    a. gyors hajók^{pl}/*gyorsak^{pl} hajók^{pl}
       'fast ships'
    b. Ezek^{pl} a hajók^{pl} gyorsak^{pl}/*gyors^{sg}.
       'These ships are fast.'
    c. öt hajó^{sg}/*öt hajók^{pl}
       'five ships'
    d. Az öt hajó^{sg} gyors^{sg}/*gyorsak^{pl}.
       'The five ships are fast.'
    e. Az öt nagynéni^{sg} sört iszik^{3sg}/*isznak^{3pl}.
       DET five aunt beer drinks
       'The five aunts drink beer.'

Since Hungarian has the plural forms *gyorsak* 'fast' and *hajók* 'ships,' these forms must somehow be blocked in (19a,c). Ortmann (2000) proposed that the agreement facts in (19) follow from the interaction of economy and expressivity principles. More specifically, if one assumes the economy constraint 'avoid multiple expression of plurality in NPs' to rank above the expressivity requirement 'mark plural on the noun!,' both *gyors hajó* and *gyorsak hajók* would violate one of these constraints, and *gyors hajók* turns out to be the optimal form in a context in which more than one ship is meant. In contrast, the singular form *öt hajó* is preferred because the numeral *öt* 'five' already expresses plurality; in this case, it is better to violate the lower-ranked constraint rather than the higher-ranked one. The tableau illustrates that the winner exhibits the fewest violations of higher-ranked constraints.

	Several fast ships are meant	*Avoid multiple expression of plurality in NPs!*	*Mark plural on the noun!*
a.	gyors hajó^{sg}		Violated
⇒	gyors hajók^{pl}		
	gyorsak^{pl} hajók^{pl}	Violated	
b. ⇒	öt^{5} hajó^{sg}		Violated
	öt^{5} hajók^{pl}	Violated	

If plural is unmarked on the noun, the noun phrase is sg and thus incompatible with a plural verb form, as shown in (19e).

As we have seen, Hungarian (unlike English and German) avoids NP internal agreement. There are also languages that avoid subject–verb agreement, even if they have the suitable affixes. Kurdish, for instance, marks plural only on the verb, see (20). In contrast, Georgian marks plural on the verb only if the subject refers to animate beings, see (21).

20. Kurdish
    Mirov/*mirovan^{pl} hat-in^{3pl}.
    man come.past-3pl
    'The men came.'

21. Georgian
    a. Burt-eb^{pl}-i goravs^{3}/*goraven$^{3pl.anim}$.
       ball-pl-NOM roll
       'The balls are rolling.'
    b. Knut-eb^{pl}-i goraven$^{3pl.anim}$/*goravs^{3}.
       kitten-pl-NOM roll
       'The kittens are rolling.'

These few examples show that the implementation of agreement considerably varies from language to language.

Additional Literature

Baker (2008), Boeckx (2006), Corbett (2003b, 2006), D'Alessandro et al. (2008).

See also: Anaphora; Case, Morphological: Linguistics; Control, Syntactic; Morphology in Linguistics; Semantics.

Bibliography

Anderson, S.R., 1998. Inflection. In: Hammond, M., Noonen, M. (Eds.), Theoretical Morphology. Approaches in Modern Linguistics. Academic Press, San Diego, CA, pp. 23–43.

Baker, M.C., 2008. The Syntax of Agreement and Concord. Cambridge University Press.

Bhatt, R., 2005. Long-distance agreement in Hindi-Urdu. Natural Language and Linguistic Theory 23, 757–807.

Bobaljik, J.D., Wurmbrand, S., 2003. Long distance object agreement, restructuring and antireconstruction. In: Kadowaki, M., Kawahara, S. (Eds.), Proceedings of NELS 33. GLSA, Amherst, pp. 67–86.

Bobaljik, J.D., Wurmbrand, S., 2005. The domain of agreement. Natural Language and Linguistic Theory 23, 809–865.

Bobaljik, J.D., 2008. Where's phi? Agreement as a post-syntactic operation. In: Harbour, D., Adger, D., Béjar, S. (Eds.), Phi-Theory: Phi Features across Interfaces and Modules. Oxford University Press, Oxford, pp. 295–328.

Boeckx, C., 2004. Long-distance agreement in Hindi: some theoretical implications. Studia Linguistica 58, 23–36.

Boeckx, C., 2006. Agreement Systems. John Benjamins, Amsterdam.

Bornkessel-Schlesewsky, I., et al., 2008. Bridging the gap between processing preferences and typological distributions: initial evidence from the online comprehension of control constructions in Hindi. In: Richards, M., Malchukov, A.L. (Eds.), *Scales* (=Linguistische ArbeitsBerichte 86). Institut für Linguistik, University of Leipzig, pp. 397–436.

Chandra, P., 2007. (Dis)Agree: Agreement and Movement Reconsidered (Ph.D. dissertation). University. of Maryland, College Park.

Chomsky, N., 1991. Some notes on economy of derivation and representation. In: Freidin, R. (Ed.), Principles and Parameters in Comparative Grammar. MIT Press, Cambridge, MA.

Chomsky, N., 1995. The Minimalist Program. MIT Press, Cambridge, MA.

Chomsky, N., 2000. Minimalist inquiries: the framework. In: Martin, R., Michaels, D., Uriagereka, J. (Eds.), Step by Step: Essays on Minimalist Syntax in Honor of Howard Lasnik. MIT Press, Cambridge, MA, pp. 89–155.
Chomsky, N., 2004. Beyond explanatory adequacy. In: Belletti, A. (Ed.), Structures and Beyond. The Cartography of Syntactic Structures, vol. 3. Oxford University Press, Oxford, pp. 104–131.
Corbett, G.G., 1991. Gender. Cambridge University Press, Cambridge, UK.
Corbett, G.G., 1994. Agreement. In: Asher, R.E. (Ed.), Encyclopedia of Language and Linguistics, vol. 1. Pergamon Press, Oxford, UK, pp. 54–60.
Corbett, G.G., 1995. Agreement. In: Jacobs, J., von Stechow, A., Sternefeld, W., Vennemann, T. (Eds.), Syntax: An International Handbook of Contemporary Research, vol. 2. De Gruyter, Berlin, pp. 1235–1244.
Corbett, G.G., 1998. Morphology and agreement. In: Spencer, A., Zwicky, A.M. (Eds.), The Handbook of Morphology. Blackwell, Oxford, UK, pp. 191–205.
Corbett, G.G., 2003a. Agreement: the range of the phenomenon and the principles of the Surrey database of Agreement. Transactions of the Philological Society 101, 155–202.
Corbett, G.G. (Ed.), 2003b. Agreement: A Typological Perspective. Special issue of Transactions of the Philological Society, 101, pp. 149–370.
Corbett, G.G., 2006. Agreement. Cambridge University Press, Cambridge.
D'Alessandro, R., Fischer, S., Hrafnbjargarson, G.H. (Eds.), 2008. Agreement Restrictions. De Gruyter, Berlin.
Deo, A., Sharma, D., 2006. Typological variation in the ergative morphology of Indo-Aryan languages. Linguistic Typology 10, 369–418.
Etxepare, R., 2006. Number long-distance agreement in (substandard) Basque. ASJU=International Journal of Basque Linguistics and Philology 40, 303–350.
Gazdar, G., Klein, E., Pullum, G., Sag, I., 1985. Generalized Phase Structure Grammar. Blackwell, Oxford, UK.
Khalilova, Z., 2007. Long-distance agreement in Khwarshi. In: Kokkonides, M. (Ed.), Proceedings of LingO. Faculty of Linguistics, University of Oxford, pp. 116–123.
Laka, I., 1988. Configurational heads in inflectional morphology: the structure of the inflected forms in Basque. International Journal of Basque Linguistics and Philology 22, 343–365.
Lapointe, S.G., 1988. Toward a unified theory of agreement. In: Barlow, M., Ferguson, C.A. (Eds.), Agreement in Natural Language: Approaches, Theories and Descriptions. CSLI, Stanford, CA, pp. 67–87.
Lehmann, C., 1982. Universal and typological aspects of agreement. In: Seiler, H., Stachowiak, F.J. (Eds.), Apprehension: Das sprachliche Erfassen von Gegenständen, vol. 2. Narr, Tübingen, Germany, pp. 201–267.
Mahajan, A.K., 1989. Agreement and agreement phrases. MIT Working Papers in Linguistics 10. Cambridge, MA, pp. 217–252.
Ortmann, A., 2000. Where plural refuses to agree: feature unification and morphological economy. Acta Linguistica Hungarica 47, 249–288.
Patel, P., 2007. Split Ergativity and the Nature of Agreement. University College London.
Polinsky, M., Potsdam, E., 2001. Long-distance agreement and topic in Tsez. Natural Language and Linguistic Theory 19, 583–646.
Pollard, C., Sag, I.A., 1994. Head-Driven Phrase Structure Grammar. The University of Chicago Press, Chicago.
Pollock, J.-Y., 1989. Verb movement, Universal Grammar, and the structure of IP. Linguistic Inquiry 20, 365–424.
Siewierska, A., 1999. From anaphoric pronoun to grammatical agreement marker: why objects don't make it. In: Corbett, G.G. (Ed.), Agreement. Special Issue of Folia Linguistica, vol. 33 (2), pp. 225–251.
Wunderlich, D., 1994. Towards a lexicon-based theory of agreement. Theoretical Linguistics 20, 1–35.
Wunderlich, D., 1997. Cause and the structure of verbs. Linguistic Inquiry 28, 27–68.
Wunderlich, D., Dec. 5, 2012. Case and Agreement Variation in Indo-Aryan. Ms, Leipzig. http://www.zas.gwz-berlin.de/wunderlich.html.

Relevant Websites

http://wals.info/ – The World Atlas of language structures online (2011) informs about the distribution of 168 structural features in a larger set of languages, including features concerning agreement such as verbal person marking (#102) and person marking on adpositions (#48).

http://www.smg.surrey.ac.uk/Agreement/index.aspx – The Surrey Database of Agreement records detailed information on agreement in fifteen genetically diverse languages.

http://www3.surrey.ac.uk/LIS/SMG/projects/agreement/agreement_bib_unicode.htm – The Surrey morphology group presents agreement: a bibliography (2001).

Grammatical Relations

Wei-wen Roger Liao, Institute of Linguistics, Academia Sinica, Taipei, Taiwan

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A.E. Kibrik, volume 9, pp. 6342–6348, © 2001, Elsevier Ltd.

Abstract

Grammatical relations refer to grammatical functions, such as subjects and direct/indirect objects, and are among the most extensively assumed concepts in linguistic analyses. Despite their wide applications, traditional grammar does not successfully define grammatical relations in a precise manner. Modern linguistics, however, has achieved such a task in the formal or the functional approaches to language. Grammatical relations are now understood as a cluster of properties involving (functional-)thematic roles, cases, and syntactic derivations regarding designated argument positions.

Grammatical relations (or grammatical functions), such as subject, direct object, indirect object, refer to nominal elements that appear in designated structural positions in relation to verbal elements, in particular between subjects and predicates, and between verbs and objects. Grammatical relations can be characterized in terms of meanings they express, in terms of their morphosyntactic properties, and/or in terms of their structural positions in formal analyses of language. Therefore, the term grammatical relations may refer to a cluster of different properties. For example, at least three types of arguments are typically referred to as 'subjects' in the linguistic literature: thematic subject, case subject, and structural/grammatical subject. Although these notions may overlap, it is important to identify each of them when studying grammatical relations. Once these notions are clearly distinguished, it seems theoretically possible to derive the notion of grammatical subject/object from the interaction of thematic roles and cases. The definitions of grammatical relations also vary drastically according to different linguistic theories. From a formal syntactic point of view, identity of the underlying structural relation serves as the main criterion for defining/deriving a grammatical function. From a functional point of view, on the other hand, one can say that grammatical relations rank-order the noun phrases (NPs) in a clause on the basis of the values of a number of parameters in accordance with the NPs' overall communicative significance in the concrete utterance. Whether grammatical relations are universal is also a highly theory-dependent question, but in terms of formal syntactic structures, universal principles of grammatical relations can indeed be found, while the structural positions of grammatical functions might be subject to language-specific parameters.

The linguistic concept *grammatical relation*, or *grammatical function*, belongs to the most important concepts of syntactic theory. Grammatical relations characterize the structural relation between a clausal/verbal element and its dependent nominal element(s), in particular between a predicate and its subject, and the verb and its object(s). Consider the following examples:

1. *Peter sent a message to Mary.*
2. *My friend sent a congratulatory telegram to his boss.*

In addition to the verbal predicate *sent*, one can identify three NPs linked to the predicate by specific grammatical relations. One of the NPs (*Peter/my friend*) is the subject, the second (*a message/a congratulatory telegram*) is the direct object, while the third (*Mary/his boss*) is the indirect object. Generally speaking, NPs that realize the same grammatical relation occupy a designated position in a clause. In (1), for example, the subject precedes the verb, the direct object strictly follows the verb, while the indirect object is introduced by a preposition *to*. Each of the NPs is structurally distinct from the other two and makes unique reference to a concrete participant (or a thematic role) in the event denoted by the given sentence.

Grammatical relations are very useful in linguistic analysis. First, they permit identification of the various participants in the event (i.e., the subject is identified as the agent or actor of the sending event, the object as the theme being sent, and the indirect object as the recipient or goal of the sending event). Second, they allow us to identify possible structural transformations among sentences/constructions. For example, one can detect that (3) and (4) are both structurally and semantically related to sentence (1) because the NPs *a message* and *Mary*, which play the same thematic roles (the theme being sent and the recipient, respectively), have dislocated from the direct object and the indirect object positions to the subject position, respectively, while the optional agent NP is then introduced by a preposition *by*. Such transformations identify the signature properties of the passive construction in English:

3. *A message was sent to Mary (by Peter).*
4. *Mary was sent a message (by Peter).*

In addition to NPs with distinct thematic roles, subjects (and objects) can also be clausal, as in (5). Some languages, like English, also allow (and even require) a dummy element to occupy the subject position, like *it* and *there* (the so-called expletive subjects) in (6) and (7):

5. [***That he won***] is fine.
6. ***It*** *is fine that he won. (cf *Is fine that he won.)*
7. ***There*** *comes a strange man. (cf *Comes a strange man.)*

However, this obligatory 'subject' requirement is not seen in the so-called 'null subject languages,' such as Chinese, Japanese, Korean, Italian, Spanish, etc. The following example illustrates a Chinese sentence without an apparent subject:

8. ø zhaodao ta le.
   find he LE
   'Someone has found him.'

 http://dx.doi.org/10.1016/B978-0-08-097086-8.52034-0

Furthermore, some languages, like Chinese and Japanese, also allow for null objects, where an apparent object is missing in a transitive predicate, as in (9):

9. Wo zhaodao ø le.
   I find LE
   'I found (it).'

What Do Grammatical Relations Mean?

Despite the widespread use of the terms subject and object, traditional grammar did not succeed in giving them coherent and precise definitions. Defining 'subjects' as the 'actors' or 'agents' of the sentences might at first blush be suitable for sentences (1) to (4), but such a definition is less transparent in (5), while the subjects in (6) and (7) clearly do not fall under a definition of this type.

In a series of investigations of the last decades (cf, for instance, the pioneering work of Keenan (1976)), it has been established that in the most elementary sentences subject NPs correlate statistically with the meanings of various autonomous semantic domains. First, subjects are linked to the semantic role properties of participants in the given situation. The subject usually encodes the agent, if there is one in the sentence. If there is no agent in the sentence, then the prime candidate for subject position is the patient. In the presence of an agent, the patient is the prime candidate for the position of direct object. Sentences with an agent and patient constitute the prototypical transitive clause with a subject and direct object. Sentences with participants in other roles are assimilated to varying extents to the structure of transitive or intransitive clauses, or use the position oblique object. The participant with the role of recipient/beneficiary usually occupies the position of indirect object.

Second, there are statistical correlations with the discourse (communicative) domain (*see* Information Structure in Linguistics). Thus, subjects often communicate known (given) information, while objects communicate new information; subjects often coincide with the focus of empathy, i.e., with that participant from whose point of view the situation is described, and with the topic (theme), while objects often coincide with the comment (rheme).

However, these distributional correlations are often probabilistic in nature, rather than deterministic, and exceptions to these generalizations can be abundant and systematic. The confusion is amplified by the fact that traditional grammar does not often distinguish different types of 'subjects' and 'objects.' Recent advances in formal syntax, however, identify at least three types of subjects: thematic subjects, case subjects, and structural/grammatical subjects, and the formal syntactic theories have rather successfully formalized the possible structural relations among them (for convenience, the discussion here is limited to subject, but similar notions may apply to objects as well). First, thematic subjects are related to thematic roles (*see* Valency and Argument Structure in Syntax), and thematic subjects, or external arguments, refer to agent/actor/experiencer roles, which are determined by the predicate as a whole (the thematic roles of X's are very different in X *killed a man* and X *killed time*, where the verb phrase as a whole determines the thematic role of X). Case subjects refer to NPs that are assigned a designated case, that is, the NP with nominative case in accusative languages (*see* Case, Morphological: Linguistics). Finally, structural/grammatical subjects refer to NPs that occur in a designated syntactic position in the hierarchical structure. Theoretically, it is possible to derive grammatical subject from the combination of thematic subject and case subject. Nevertheless, grammatical subjects are not equivalent to either thematic subjects or case subjects alone. In (6) and (7), the expletive subjects are not assigned any thematic roles. In some cases, grammatical subjects may also be caseless, such as the unpronounced grammatical subjects in infinitive clauses in English (*see* Control, Syntactic): *John tried* (*ø to fix the car*) (*ø = John*). The empty subject is a thematic subject and a grammatical subject, yet it occurs in a position that is not assigned case.

Structurally speaking, one of the distinctive properties of grammatical subject is that it occupies the most dominant position than any other arguments. For example, it has been shown that when a pronoun has a coreferring reading with (or is bound by) a universal quantified NP (*see* Quantifiers), the universal QP must be located in a structurally more dominant position than the 'bound' pronoun. Technically speaking, the universal QP must c-command the bound pronoun (*see* Generative Grammar):

10. a. everyone likes his mother. (*everyone* and *his* can co refer: x likes x's mother.)
    b. His mother likes everyone. (*everyone* and *his* cannot co refer: *x's mother likes x.)
11. a. Everyone sent a postcard to his mother. (*everyone* and *his* can co refer.)
    b. His mother sent a postcard to everyone. (*everyone* and *his* cannot co refer.)

Examples (10) and (11) show that there is an asymmetry between the structural position of subject and those of the other arguments. The subject is able to c-command the direct object and the indirect object, yet none of the arguments can c-command the subject. The subject thus occupies the most prominent structural position in the syntactic structure. The subject is also obligatory in some languages, in contrast to all other arguments (no languages impose an obligatory object requirement). Notably, there is often a unique structural position for the subject in a given language, but the exact position might vary cross-linguistically (see the section 'Grammatical Relations and the Functional Approach to Language').

How Are Grammatical Relations Encoded?

Languages may employ different means to encode grammatical relations. The best known morphological category specifically designated to encode grammatical relations is case (Morphological Case in Linguistics). Overt morphological case marking is present in many language families, e.g., in Latin, Greek, the Slavic and Dravidian languages, the Altaic languages, the languages of Australia, and many of the Austronesian languages. Under overt case marking, grammatical relations are expressed through various overt morphological forms of nouns (as in Latin and Russian) or through different forms of other structural constituents related to NPs (e.g., forms of articles and nouns in German). Alongside inflectional morphology, many languages (such as Japanese and Korean)

make use of grammatical words or particles (usually prepositions or postpositions), which combine with NPs to give rise to analytical elements that indicate grammatical relations (e.g., in English, the combination of preposition *to* and NP, such as *to Mary* in (1), encodes the indirect object). In some languages, overt morphological marking may be indicated not on NPs, but on the verb by means of agreement (*see* Grammatical Agreement): The form of the verb expresses information about the grammatical properties of that NP which stands in a particular grammatical relation to the verb (or auxiliaries). An example in English is agreement of the verb (or auxiliary) with a third-person singular subject in number. Some languages have polypersonal agreement of the verb with two or more NPs. Thus, in Abkhaz the verb has three agreement positions indicating agreement with specific syntactic types of NPs. The linear order of NPs in the sentence is the positional equivalent of inflectional or analytical marking (*see* Word Order) and is calculated according to the position of the NP relative to the verb, relative to the beginning or end of the sentence, or relative to other NPs. This strategy is often seen in languages without overt morphological case markings. Thus, in linguistic typology, it is common to characterize languages in terms of their unmarked word orders among subject, verb, and object. For example, one can distinguish between SVO languages (e.g., English and Chinese) and SOV languages (e.g., Japanese and Korean), the combination of which constitutes the largest share among world languages. It is noteworthy that the mainstream generative grammar embraces the hypothesis that case is universal in natural languages, and for those overt case markings are not observed, cases are assigned under specific structural configurations that are systematically mapped to certain linear word orders. If such a position is maintained, then the close association between word order and case assignment can be established through structural relations.

Grammatical relations are also strongly tied with grammatical voices. Take English for example. In a sentence with a transitive voice, the NP with the thematic role agent/actor emerges as the grammatical subject, while in a sentence with a passive voice, the NP with the thematic role patient/theme emerges as the grammatical subject. Some languages, especially the Austronesian languages, have other voice constructions (inflected on verbs) that allow NPs with various thematic roles (e.g., instrument, location, beneficiary, etc.) to surface as grammatical subjects.

Formal Analyses of Grammatical Relations

Modern formal theories of syntax may speak of grammatical relations directly as theoretical primitives (as in lexical function grammar and relational grammar) or indirectly as derivatives of theta roles and cases (as in generative grammar). Nevertheless, one common ground is that grammatical relations strongly correlate to verb/predicate types. Four types of predicates will be briefly introduced: accusative/transitive, unaccusative, ergative, and unergative.

A typical accusative/transitive verb takes one or two objects, in addition to the subject. An example of transitive verb is *like*, as in *He likes her*, and *send*, as in *John sent a letter to Mary*, the latter of which is called a ditransitive verb. In the former example, *he* is a nominative marked NP with an agent/actor role and occupies the structural subject position, while *her* is an unaccusative marked NP with a patient role and occupies the (direct) object position. In English, it is not possible to use expletive subject in transitive sentences (cf *there ate John some cake), but it is possible to find such sentences in Icelandic, as in (12) (from Bobaljik and Jonas, 1996: 196):

12. það borðuðu sennilega margir jólasveinar bjúgun
    there ate probably many Christmas. the.
    trolls sausages
    'Many Christmas trolls probably ate the sausages.'

Transitive sentences can also undergo passivization in which the agent NP is demoted to a less prominent position and the patient NP in turn occupies the structural subject position (and obtain the structural nominative case). Under a structural account, it is argued that, in active sentences, the agent NP is base-generated in a position closer to the grammatical subject position than the patient NP, and therefore, the agent NP is the NP that can move into the grammatical subject position, rather than the patient NP, as in (13):

13. [subject position: ___ [Agent [V Patient]]]

When a transitive verb undergoes passivization, on the other hand, the agent NP is demoted, and the verb no longer assigns an accusative case to the object position, and therefore, this forces the patient NP to move into the structural subject position instead:

14. [subject position: ___ [[V Patient] [*by* Agent]]

For unaccusative verbs, such as *arrive*, *appear*, and *come*, there are no agent arguments, and the only argument (a theme NP) may either stay in the object position or move to the subject position (for nominative case), as shown in the following examples:

15. a. There arrived three guests.
    b. Three guests arrived.

Unlike transitive predicates, the promotion of the theme argument to the subject position does not require a passive voice in this type of predicate, and an expletive may occupy the subject position. Under a structural account, the obligatory subject position is either fulfilled by an expletive *there* or triggers the theme argument to move into the subject position:

16. a. [subject position: *there* [V Theme]]
    b. [subject position: _____ [V Theme]]

In the null subject languages, where there is no obligatory subject requirement, such as Chinese, the subject can be left empty in this type of predicate:

17. a. ø Lai-le san ge keren.
    come-LE three Classifier guest
    'There came three guests.'
    b. San ge keren lai-le.
    'Three guests came.'

An ergative verb is minimally different from the transitive and the unaccusative verbs in that the agent argument is

optionally projected in the argument structure. Examples of ergative verbs in English are *roll*, *sink*, *open*, etc., as in (18) and (19). The structural configuration of ergative verb is displayed in (20):

18. a. John rolled the ball down the hill.
    b. The ball rolled down the hill.
19. a. Peter opened the door.
    b. The door opened.
20. [subject position: ___ [(Agent) [V Theme]]]

In the (a) examples, the cases assigned to the subject and the object are the same as the transitive predicates, where the subject position is assigned a nominative case and the object position an accusative case. In the (b) examples, on a par with the unaccusative predicates, the theme argument (as the only argument) must move to the subject position and is assigned a nominative case. Unlike unaccusative predicates, however, expletive subjects are not allowed in ergative predicates in English: **There rolled a ball down the hill*. Therefore, when the agent role does not appear, movement of theme NP to the subject position is obligatory. It has also been argued that the agent argument in this type of predicate is introduced by a silent causative verb, and therefore, in (18a) and (19a), the structures may involve an additional layer of null verb CAUSE:

21. [subject position: ___ [Agent/Causer CAUSE [V Theme/Causee]]]

Finally, an unergative verb, like *laugh*, *complain*, *lie*, etc., can only project an agent argument, which, being the only argument, moves to the subject position and is assigned a nominative case. There is no direct object in this type of predicate. However, an indirect object may appear if it is introduced by a preposition, as shown in (22). Formally, this type of predicate has the structural configuration in (23), where the only argument (the agent NP) ultimately moves into the subject position:

22. a. Tom laughed (at the girl).
    b. Harry complained (about his lunch).
23. [subject position: ____ [Agent V]]

It is therefore possible for the formal structural analyses to derive the less transparent notions of grammatical relations from more primitive and well-defined notions of theta roles (which are regulated by argument structures) and cases, and thus avoid the ambiguous uses of grammatical relations.

(Non)universality of Grammatical Relations

Whether grammatical relation is universal is a theory-internal question, and answers to the question vary according to different theoretical assumptions and syntactic tests. In descriptive linguistics and linguistic typology, where linear word orders and functional morphosyntactic alignments are concerned, one usually concludes that grammatical relations are not universal among languages, and some languages might even have no well-defined grammatical relations at all. The widely adopted morphosyntactic alignment test is often carried out by nuclear arguments with the symbols S, A, P (or O), where S is the nuclear argument of an intransitive clause, and A and P the agent-like and patient-like arguments, respectively, of a transitive clause, then in European languages the S/A arguments have identical encoding, grouping them together as the grammatical relation of subject, in opposition to the encoding of the P argument. Such a system is called accusative alignment. Alongside this possibility, many languages found in various continents have the ergative pattern (e.g., Basque, Dyirbal), uniting the S/P arguments in opposition to the A argument. Here the S/P arguments are usually morphologically unmarked, while the A argument is marked (ergative case or its analog). Examples (24) and (25) draw a comparison between an accusative language like Japanese and Dyirbal (from Aldridge 2008):

24. Japanese (NOM = Nominative, ACC = Accusative)
    a. Hanako-ga kita.
       Hanako-NOM came
       'Hanako came.'
    b. Hanako-ga Taro-o hihan-sita.
       Hanako-NOM Taro-ACC criticize-did
       'Hanako criticized Taro.'
25. Dyirbal (ABS = Absolutive, ERG = Ergative, NF = Nonfuture)
    a. yabu banaga-n^{y}u
       mother.ABS return-NF
       'Mother returned.'
    b. nguma yabu-nggu bura-n
       father.ABS mother-ERG see-NF
       'Mother saw father.'

The ergative alignment of nuclear arguments thus brings about an analytical challenge for theories of grammatical relations. Other patterns are also found in natural languages. One possibility is a pattern that distinguishes agent-like and patient-like arguments of both transitive and intransitive clauses (active alignment). Under active alignment one finds different expression of the single argument of verbs like 'run' and verbs like 'fall.' There is even the tripartite scheme, distinguishing all three arguments, S, A, and P. Finally, there is the neutral pattern, which fails to differentiate the S/A/P arguments by means of coding. Language of this type is often claimed to have no grammatical relations at all (Dixon, 1994; Van Valin and LaPolla, 1997).

From a structural point of view, on the other hand, where more complicated hierarchical relations and structural dependencies are taken into consideration, grammatical relations are generally considered to be universal and may be derived from more primitive and universal notions like theta roles and cases, as illustrated in the section 'Formal Analyses of Grammatical Relations.' For example, the universal principle of subjecthood is that the structural position of subject is structurally more prominent than all other argument positions, and a language always has a designated structural position for grammatical subject that is derivationally related to thematic and/or case subjects. While the principle holds, languages may still differ with respect to the structural position where a given grammatical relation is encoded. For example, it is often argued that the grammatical subject in English occupies a unique position licensed by the inflection or tense phrase (or the specifier of TP), which dominates the verb phrase (VP), while the object NP is base-generated as a complement or specifier of verb inside VP. On the other hand, it has been argued that the grammatical subject in Irish occupies a lower position (perhaps inside the VP) than

that in English. McCloskey (1996, et seq.), for example, has demonstrated that in Irish, the subject is elided along with the object in the VP-ellipsis construction:

26. Ní tháinig muid 'na bhaile anuraidh
not came we home last-year
ach tiocfaidh — i mbliana
but come [Future] this-year
'We did not come home last year, but we will
(~~come home~~) this year.'

Irish is a language with the VSO word order, where the verb is argued to move to a higher position out of VP (to the tense or inflection) and pass the structural position of subject. If the subject in Irish also moved out of VP to a position licensed by tense/inflection phrase (like the subject in English), we would expect that after VP-ellipsis, the subject should escape being elided within the VP. The fact that subject and object are both elided in the VP-ellipsis construction in Irish then suggests that the Irish subject occupies a position inside VP, on a par with the object. Conversely, in some Austronesian languages, it is often argued that the grammatical subject may occupy a structurally higher position than that in English. For example, Pearson (2005) has argued that the grammatical subject in Malagasy occupies a position typically associated with topic and *wh*-movement (*see* Movement Theory and Constraints in Syntax), which is a higher projection than TP. In this type of language, the structural requirements of topics outweigh those of grammatical subjects to an extent that the role of topic has replaced the grammatical function of subject. In language typology, languages of the Malagasy type are generally referred to as topic-prominent languages, in contrast to subject-prominent languages, such as English and French. Some languages, such as Japanese and Korean, show mixed characteristics, where subjects and topics may coexist and are both considered prominent in their discourse functions.

The controversial issue of ergative alignment, which is often used as an argument against the universality of grammatical relation, may also be neutralized once a more articulated structural account is adopted, especially when we consider grammatical relations as derivatives of theta roles and case assignments. Recent studies in Legate (2002, et seq.) and Aldridge (2004, et seq.), for example, both argue on independent grounds that the ergative case in ergative languages is tied to the thematic role of agent/actor (or the external argument), and the absolutive case is identified as the default unvarying form of structural cases (compared to nominative case and accusative case in accusative languages). If such approaches are on the right track, then the difference between ergative languages and accusative languages simply supports the idea that grammatical relations are not primitives in natural languages, but are derivatives of universal notions of theta roles and cases, as argued in the mainstream generative grammar (Williams, 1984).

Grammatical Relations and the Functional Approach to Language

The functional explanation of grammar (*see* Functional Approaches to Grammar) has permitted establishing the polyfunctional character of the elementary categories that form the clusters known as grammatical relations. The majority of these categories were themselves discovered and investigated in the framework of the functional approach (Kibrik, 1997).

In this connection, the question of the existence of different patterns of encoding S, A, and P arguments merits special attention. Proceeding from the now widely established assumption of the iconicity of the linguistic sign (*see* Iconicity in Linguistics), replacing the Saussurean assumption of the arbitrariness of the sign, one can allow for the existence of a functional motivation for the various coding patterns. These patterns unite arguments with different elementary roles in different ways. The basic prototypes of all these groupings (*see* Linguistics: Prototype Theory) are given by the agent and patient of the transitive construction. From this viewpoint, active alignment is a natural metonymic extension of the elementary roles of agent and patient, generalizing them to the hyperroles (macroroles) of actor (the participant in the event that performs, effects, instigates, or controls it) and undergoer (the participant in the event that does not perform, effect, instigate, or control it, but rather is affected by it in some way). Arguments with different elementary roles can be interpreted in terms of these hyperroles.

Accusative alignment is also motivated essentially not by grammatical relations, but by the hyperroles principal and patientive. The principal indicates the main participant, the 'hero' of the situation, who is primarily responsible for the fact that this situation takes place. In the transitive clause the principal is, of course, the agent-like argument, while in the intransitive clause the single nuclear argument has no competitors for the role of principal. The patientive is the most patient-like participant of a multiparticipant event. This role unites, in addition to the patient, arguments of other types.

Ergative alignment articulates the S, A, P arguments by means of other hyperroles, absolutive and agentive. The absolutive is the metonymic extension of the patient to intransitive verbs; it is the immediate, nearest, most involved, or affected participant in the situation. In the transitive clause the patient is most involved in the situation. In the intransitive clause there is no competition for the role of the closest and most involved participant in the situation. The absolutive in the transitive clause is in opposition to the agentive, the most agent-like participant of the multiparticipant event.

Tripartite alignment distinguishes agentive and patientive of the transitive clause, distinguishing them from the single nuclear argument of the intransitive clause.

All these means of generalizing the elementary roles are in principle motivated, but a language has the option of grammaticalizing just one of them as its basic pattern. In addition, a language may have different patterns in different contexts, and examples of this kind have been attested.

Grammatical Relations and the Cognitive Approach to Language

The 'statistical' correlations between concepts of different semantic domains noted above and which are grammaticalized in grammatical relations and the means of their encoding are not chance, rather they have a cognitive nature, reflecting

the human capacity for categorizing reality. Behind the limitless variety of whole grammatical systems lie the same stable cognitive structures of human thought and the pragmatic strategies (*see* Pragmatics, Linguistic) for their verbalization. The reconstruction of these structures and strategies on the basis of data from different languages is the task of cognitive typology (*see* Linguistic Typology).

See also: Case, Morphological: Linguistics; Grammatical Agreement; Grammaticalization.

Bibliography

Aldridge, E., 2004. Ergativity and Word Order in Austronesian Languages. Ph. D. dissertation. Cornell University.

Aldridge, E., 2008. Generative approaches to ergativity. Language and Linguistics Compass 2, 966–995.

Aissen, J., 1999. Markedness and subject choice in optimality theory. Natural Language and Linguistic Theory 17, 673–711.

Bobaljik, J.D., Jonas, D., 1996. Subject positions and the roles of TP. Linguistic Inquiry 27, 195–236.

Cole, P., Sadock, J.M. (Eds.), 1977. Grammatical Relations. Syntax and Semantics, vol. 8. Academic Press, New York.

Croft, W., 1991. Syntactic Categories and Grammatical Relations. The University of Chicago Press, Chicago.

Dik, S., 1997. The Theory of Functional Grammar. Mouton de Gruyter, Berlin.

Dixon, R.M.W., 1994. Ergativity. Cambridge University Press, Cambridge, UK.

Dixon, R.M.W., Aikhenvald, A.Y. (Eds.), 2000. Changing Valency: Case Studies in Transitivity. Cambridge University Press, Cambridge, UK.

Dryer, M., 1986. Primary objects, secondary objects, and antidative. Language 62, 808–845.

Dryer, M., 1997. Are grammatical relations universal? In: Bybee, J., Haiman, J., Thompson, S. (Eds.), Studies in Language Function and Language Type. Benjamins, Amsterdam, pp. 115–143.

Givon, T. (Ed.), 1997. Grammatical Relations: A Functional Perspective. Benjamins, Amsterdam.

Keenan, E., 1976. Towards a universal definition of 'subject'. In: Li Ch, N. (Ed.), Subject and Topic. Academic, New York, pp. 303–334.

Kibrik, A.E., 1997. Beyond subject and object: toward a comprehensive relational typology. Linguistic Typology 1, 279–346.

Legate, J.A., 2002. Warlpiri: Theoretical Implications. Ph. D. dissertation. MIT.

Legate, J.A., 2008. Morphological and abstract case. Linguistic Inquiry 39, 55–101.

Li Ch, N. (Ed.), 1976. Subject and Topic. Academic, New York.

McCloskey, J., 1996. Subjects and subject positions in Irish. In: Borsley, R.D., Roberts, I. (Eds.), The Syntax of the Celtic Languages: A Comparative Perspective. Cambridge University Press, Cambridge, pp. 241–283.

Pearson, M., 2005. The malagasy subject/topic as an A'-element. Natural Language & Linguistic Theory 23, 381–457.

Perlmutter, D.M. (Ed.), 1983. Studies in Relational Grammar. The University of Chicago Press, Chicago.

Plank, F. (Ed.), 1979. Ergativity: Toward a Theory of Grammatical Relations. Academic, London.

Radford, A., 2004. Minimalist Syntax: Exploring the Structure of English. Cambridge University Press, Cambridge.

Van Valin, R., LaPolla, R.J., 1997. Syntax: Structure, Meaning, and Function. Cambridge University Press, Cambridge, UK.

Williams, E., 1984. Grammatical relations. Linguistic Inquiry 15, 639–673.

Grammaticalization

Paolo Ramat, Center for Postgraduate Education and Research LETiSS, Istituto Universitario di Studi Superiori (IUSS) (Retired), Pavia, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

What do we mean by 'grammaticalization'? What are the criteria for establishing grammaticalization processes? After a short sketch of the past debate some 'canonical' examples of grammaticalization are discussed in this article. One of the main points in today's discussion concerns the nature and the causes of grammaticalization process: does it happen step by step along a continuum or is it an abrupt innovation? Examples are discussed and evidence is adduced in favor of the former alternative: there exists a diachronic dynamic of the grammaticalization processes, but at the end of a process we may have an abrupt transcategorization. Finally, this article discusses the much debated question of the unidirectionality of the grammaticalization phenomenon and concludes that grammaticalization and degrammaticalization (with lexicalization as possible output) are two complementary processes along a continuum which move into two opposite directions. Thereby, reanalysis plays an important role in the shifting from grammar to lexicon and vice versa.

Definition of Grammaticalization

Many definitions of grammaticalization have been suggested according to various theoretical approaches. One of the most diffused and clearest is Kuryłowicz' definition: "Grammaticalization consists in the increase of the range of a morpheme advancing from a lexical to a grammatical or from a less grammatical to a more grammatical status," Kuryłowicz (1975). Some examples of grammaticalization are Latin *habere* from a full verb meaning 'to grasp, seize' developed to an auxiliary in periphrastic verbal constructions for the expression of the future (*dicere habet* 'he/she has to say > shall say') or the past (*dictum habet* 'he/she has said'). The prepositional + relative phrase of the Old English *þa hwile þe* 'at the time that' evolved to the adversative and even concessive meaning of the Modern English conjunction *while* (see Traugott and König, 1991). The Latin demonstrative *ille* 'that,' with strong deictic function (:*ille homo* 'that man,' *homo ille* 'that notorious, famous man') gave rise to the obligatory article of the Romance languages (French *l'homme*, Spanish *el hombre*), where its function is simply to mark the noun phrase as determined (vs *un homme*, *un hombre*, indefinite). Grammaticalization is thus a process leading from lexemes to grammatical formatives: "a sign is grammaticalized to the extent that it is devoid of concrete lexical meaning and takes part in obligatory grammatical rules" (Lehmann, 1995: p. 6).

Sometimes the terms 'grammaticization' or 'grammatization' have been used instead of grammaticalization without any substantial difference.

As is shown by the example of Latin *ille*, grammaticalization may lead to the emergence of new grammatical categories (in the case in point, the article which did not exist in Latin). In all the adduced examples the grammatical output of the process has passed through a syntactic use of the lexeme: see the evolution path below. Note, however, that not all the steps of the path need be always present.

Lehmann (1995: p. 123) has suggested three relevant aspects of grammaticalization, which realize six parameters both at the paradigmatic and syntagmatic level (see **Table 1**).

While grammaticalization is a process, the six parameters are not processes but properties that linguistic forms possess in increasing or decreasing measure. A linguistic form shows integrity inasmuch as it maintains its phonological and semantic identity. Phonological erosion and semantic bleaching reduce the integrity of a form (Latin *ille* > Italian *il*, Spanish *el*, etc.). The scope is the structural size (the 'weight' in Lehmann's terms) of a construction: when a full verb is grammaticalized as auxiliary it makes up one phrase with the main verb and thereby reduces its scope to just the main verb (see the example of *habeo cultellum inventum* 'I have found a knife' in Section The Continuum of the Grammaticalization Process). By paradigmatic cohesion is meant the formal and semantic integration of a form into a paradigm: the suppletive forms *better*, *best* are less paradigm-integrated than *longer*, *longest*. Bondedness refers to the syntagmatic cohesion: the more a form is grammaticalized the more it will be syntagmatically constrained; French *le*, Spanish *el*, etc. are fully grammaticalized noun determiners which cannot appear in isolation. Paradigmatic variability refers, according to Lehmann, to the freedom with which the language user chooses a sign – and, again, variability decreases with increasing grammaticalization: for instance, if cases are weakly grammaticalized the case suffixes may vary with relative freedom:

(1) Hungarian

a gyerek az asztal-on	*jatszik*
The child the table-SUPERESSIVE	plays
'the child is playing on the table'	

but *asztal-nál* (ADESSIVE) and *asztallal* (<*asztal-val*, INSTRUMENTAL-COMITATIVE) would also be possible,

Table 1 Six aspects of grammaticalization

	Paradigmatic level	*Syntagmatic level*
Weight	Integrity	Scope
Cohesion	Paradigmaticity	Bondedness
Variability	Paradigmatic variability	Syntagmatic variability

 http://dx.doi.org/10.1016/B978-0-08-097086-8.53011-6

while keeping the same meaning. Another example of paradigmatic variability is as follows: the use of Latin *ille* was not compulsory in pre- or postnominal position and was semantically governed, but in Romance languages it is in most cases impossible to have a noun without the article. Finally, the syntagmatic variability is the ease with which a form may be shifted in the linear order of a sentence: the more a form has undergone a grammaticalization process the less variable it will be; **I have a knife found* or **I have a found knife* are impossible in Modern English, and the only grammatical form is *I have found a knife*.

It has to be noticed that the processes alluded to under the three labels of weight, cohesion, and variability cut across different levels of language structure; therefore the same items (e.g., *ille*) turn out to be appropriate examples of different parameters (integrity, scope, etc.). Grammaticalization is a pervasive phenomenon whose effects may be seen at all linguistic levels. Consequently, "There is a growing interest in integrating grammaticalization with theoretical work in descriptive and historical linguistics" (Giacalone Ramat and Hopper, 1998: p. 1).

A Historical Sketch of the Research

The classical examples of *dicere habet* and *l'homme* go back to Antoine Meillet (1916, repr. 1948), who probably used first the term grammaticalization – though the concept is implicitly referred to by some previous linguists such as Wilhelm von Humboldt (*Über das Entstehen der grammatischen Formen und Ihren Einfluß auf die Ideenentwicklung* (On the Origin of Grammatical Forms and their Influence on the Ideas' Development), 1825), William Dwight Whitney (*The Life and Growth of Language: An Outline of Linguistic Science*, 1875), Georg von der Gabelentz (*Die Sprachwissenschaft: Ihre Aufgaben, Methoden und bisherigen Ergebnisse* (Linguistics. Its tasks, methods, and up to now results), 1901), and others (see Lehmann, 1995: Chapter 1; Heine et al., 1991: pp. 5–11).

As can be seen from the titles just quoted, grammaticalization has from the very beginning been considered in a large theoretical frame as one of the most characteristic developments languages may undergo. This holds also for the linguists of the following generations. Edward Sapir (1921: p. 102), for one, observed that "it is possible for a concrete concept, represented by a simple word, to lose its material significance entirely and pass over directly into the relational sphere [i.e., the expression of abstract morphosyntactic relations, PR] without at the same time losing its independence as a word": forms live longer than their own conceptual content, which may be bleached (weakened) and eventually fade completely. Sapir's examples are the Chinese and Cambodian verbs for 'give' used in a very abstract sense to express the 'indirect object' relation (see also Section The Present State of the Discussion about Grammaticalization).

Though not using the term grammaticalization, the French linguist Emile Benveniste (1968) took up again the case of *habere*, and called it a conservative mutation whereby a new (periphrastic) form takes over the function of an older one in the grammatical paradigm (e.g., *habeo dictum* 'I have said' instead of *dixi* 'I said').

In the last 30 years or so grammaticalization studies have experienced a real boom, as part of a countermovement to the dominant formal approaches to linguistics (though generative research has also contributed to the study of grammaticalization) (e.g., Roberts and Roussou, 2003; van Gelderen, 2004; Batllori et al., 2005). After the pioneering studies by Givón (1979), Hopper (1987), Traugott and Heine (1991), Heine et al. (1991), Bybee et al. (1994), Lehmann (1982, repr. 1995), a real flood of monographies and articles was published. A critical assessment of grammaticalization was edited by Campbell in a special issue of 'Language Sciences,' 2001.

An important point in Campbell's contribution to the special issue is the question whether grammaticalization can be considered as a theory per se. Campbell's answer is 'No.' He argues that "grammaticalization does not have any independent status of its own, but rather is derivative of other kinds of language change" (p. 116). However, this epistemological question needs to be further discussed (cf Newmeyer, 1998; Fischer, 2000; among others).

In the mentioned special issue the introductory paper by Campbell and Janda gives an exhaustive review of the many definitions that have been given of 'grammaticalization' (Campbell and Janda, 2001). A new book series called 'New reflections on grammaticalization,' which collects the proceedings of international conferences with the same title, started in 2002. The fifth and most recent conference took place in Edinburgh, in July 2012. Reasons of space do not permit to provide here an exhaustive bibliography of the works on grammaticalization that deal with general as well as particular aspects of the phenomenon. Just to give an idea of the breadth of topics which go under the label 'grammaticalization' we can mention van Gelderen (2004) on grammaticalization as economy or Heine and Kuteva (2005) on contact-induced grammaticalization, and, among the most recent writings, Traugott and Trousdale (2010) on synchronic gradience and grammaticalization and Heine and Kyung-An Song (2011) on the grammaticalization of personal pronouns. Stathi et al. (2010) offer a panorama on current views and issues of grammaticalization. A short list of further readings, not directly quoted in text, has been added to the references.

As happens when terms become fashionable, the original meaning of 'Grammaticalization' – as results in Kuryłowicz' definition given at the beginning of this article – has been extended (and therefore bleached) to cover practically almost any phenomenon of language change. On (previous) research on grammaticalization from a broader perspective see Bisang et al's. 2004 book on grammaticalization which has the significant subtitle 'A look from the fringes and its components.' Heine and Kuteva have published a lexicon of grammaticalization as a reference book in an attempt to collect data from as many different languages as possible (Heine and Kuteva, 2002): they have listed more than 400 grammaticalization processes that may have resulted in new grammatical forms; see for instance, German *rings*, genitive of the noun *ring* 'ring' → preposition as in *rings um den Dom* 'round about the cathedral,' exactly as in Icelandic *í kring um húsin* 'around the houses' (lit. in the circle around the houses), or Latin *Terra se conuertit circum axem suum* 'the earth turns around its own axle,' where the preposition *circum* is actually the accusative of the noun *circus* 'circle' (Heine and Kuteva, 2002: pp. 67f).

Indefinite pronouns can originate from nouns meaning 'person, man' as in Albanian

(2) *s' pashë njeri*
NEG I saw someone$_{ACC}$ 'I haven't seen anybody,'
or in French
(3) *on parle français*
'French is spoken,'

where *njer* in (2) means 'man' and *on* in (3) derives from Lat. *homo* 'man.' The shift from the original nominal meaning to the impersonal pronoun did not happen suddenly: there are so-called bridging texts where both interpretations – noun and/or pronoun – are possible. The famous 'Oaths of Strasbourg' (842) have in the Romance version

(4) *si cum* ***om*** *per dreit son fradra salvar dift*
as man with right his brother save must
'as a man rightfully must save his own brother,' or 'as one rightfully' etc.,

The German version has

(5) *sōso* ***man*** *mit rehtu sīnan bruodher scal (scil. haldan)*
as man with right his brother must (save, protect).

Present-day German has kept the double meaning of *man*:

(6) *Ich habe einen* ***Mann*** *gesehen* 'I have seen a man' and
(7) ***Man*** (weakened form of Mann) *spricht Deutsch* 'German is spoken,'

whereas French distinguishes between *homme* 'man' (from the Latin accusative form *homine(m)*: see below) and *on* '(some) one.' We can say that French went a step further in the grammaticalization process of the original Lat. *homo* in that *on* keeps no trace whatsoever of its origin.

Recently, Narrog and Heine have edited a very large and comprehensive handbook of grammaticalization in the series 'The Oxford Handbooks in Linguistics' (Narrog and Heine, 2011).

The Present State of the Discussion about Grammaticalization

The Continuum of the Grammaticalization Process

Already Whitney and Benveniste had underlined the fact that linguistic changes do not happen abruptly: "Present possession often implies past action: *habeo cultellum inventum*, (…) *habeo digitum vulneratum* 'I possess my knife' (recovered after loss) [>'I have found the/my/a knife,' PR], (…) 'I have a finger wounded' [>'I have wounded my finger,' PR]." Thus, *I have found the knife* became the indicator of "a peculiar variety of past action contemplated as completed," (Whitney 1875: p. 91; cf Heine et al., 1991: pp. 7f). Benveniste gives an insightful philological description of how the periphrastic forms with *habere* came first to be used and later spread out.

As a matter of fact, grammaticalization, like all terms ending in *–ization*, alludes to a gradual process, not a state. Grammaticalization is different from grammar. Even what has recently been alluded to as 'synchronic grammaticalization' implies a process whose final stage cannot be simultaneous with the beginning moment. Grammaticalization occurs in gradual microsteps, as we have seen when speaking of 'bridging texts' with the example from the 'Oaths of Strasbourg' (see further the Ewe examples (10)–(12), below). This is not in keeping with the assumption by generative transformational grammar of rigidly separated grammatical categories: clines, continua, squishes are by definition ruled out. When generativists try to take account of the microsteps, they are bound to a proliferation of discrete grammatical categories (cf Vincent and Börjars, 2010). Actually, the 'new trend' of grammaticalization studies which has flourished in linguistics since the 1980s is basically oriented toward a functionalist approach, with very few exceptions. Grammatical categories are not seen as waterproof boxes; they permit, on the contrary, the gradual leaking of forms from one category to another. Just to give an example, the German particle *auch* 'also, too, even' was originally an imperative of a verb meaning 'add' (cf Lat. *augeo* 'augment'). The next paragraph discusses a classical case of transcategorization (on the term 'transcategorization' see below in this section). It may also happen that a grammaticalization process is interrupted and, so to say, reversed. Old Italian knew *(u)omo* as indefinite/impersonal pronoun:

(8) *co(n) ciò sia cosa che Dio sia inn o(n)gna luogo, inn ogna luogo può* ***l'uomo*** *servire Dio*
(Anonymous, *Trattati di Albertario da Brescia volgarizzati*, IV, 23: 1288).
since God is everywhere, man/one can serve God everywhere (cf Giacalone Ramat and Sansò, 2007).

Present-day Italian does not know this use of *uomo* any longer and *uomo* means simply 'man, person.' We can represent this interruption of the grammaticalization process as follows.

(9) A > A,B > A

where A represents the original phase (i.e., *(u)omo* meaning just 'man'), A,B the intermediate phase with the coexistence and a possible contrast between the A and B (=impersonal pronoun) meanings, and, finally, the come back to A and the disappearance of B which was substituted by another form (namely *si*, as impersonal pronoun). In a complete grammaticalization the third step would be B (i.e., A > A,B > B) as in French *on*, where A disappears and is substituted by another form (namely *homme* < Lat. accusative *hominem*). Remember also that Middle English knew a pronoun *man*, just as Germ. *man*, until the fifteenth century. But later this indefinite pronoun was abandoned.

To be functionally oriented in grammaticalization studies means that large attention is paid to the internal mechanisms, that govern linguistic structures, and the external psychological/cognitive causes of grammaticalization. Several linguists have hinted at the fact that grammaticalization very often originates in discourse strategies (i.e., in pragmatics), evolving to syntactic structures, and later obligatory morphological markers (up to a possible total erosion of the markers) along the cline:

Discourse > Syntax > Morphology > Morphophonemics > Zero (see Givón, 1979: p. 209, 2009; cf Lehmann, 1995: p. 13; Traugott and Heine, 1991: p. 3, Bybee et al., 1994: pp. 40f). A good example of this evolution is the so-called Jespersen's cycle (Jespersen, 1917) of the negative construction in French, and other languages. First, we had constructions as Lat. *Non uado passum* 'I don't go a step,' where *passum* (step$_{ACC}$)

is used as the object of the intrinsically intransitive verb 'to go' and the negation is expressed by *non*. The sentence has a strong emphasis at the discourse level, like *non comedo/manduco mica(m)* 'I don't eat a crumb,' *non bibo gutta(m)* 'I don't drink a drop,' etc. Thereafter, *pas(sum)*, *mica(m)*, *gutta(m)* became Negative Polarity Items (NPIs) that denoted a least quantity and were used, like *pas*, to reinforce the negation. Thus, we get the syntactic construction with double negation *je ne vais pas*. But, because of its semantic bleaching, *pas* turned out to a plain negation marker (even ousting other NPIs: (Old) French *mie* < Lat. *mica(m)* 'crumb,' *gote* (French *goutte*) < Lat. *gutta(m)* 'drop,' etc.) so that it became possible to say *je ne vois pas* 'I don't see,' *je ne comprends pas* 'I don't understand,' etc.: *pas* lost its original meaning ('step') and was used just as a marker of negation. Modern spoken/colloquial French has *je vois pas*, *je comprends pas* where *pas* is the only morphological negation marker and the first member of the double negation, i.e., *ne*, has been dropped to zero. Finally, in French-based creoles we may find *mo pa kup* (Louisiana, from French **moi pas coupe*) 'I don't cut,' *nu pavle* (Guadeloupe, from French **nous pas voulez*) 'We don't want' where *pa* (<French *pas*) is again in preverbal position as was *non* at the beginning of the cycle, since in the discourse strategy the verb is usually in the scope of the negation (cf Condoravdi and Kiparsky, 2005; Ramat, 2006).

The Latin example *habeo cultellum inventum* discussed above also originates at the discourse level: the speaker wants to underline that he does really possess (*habet*) a knife which has previously been found (*inventum*) and uses a construct where every word keeps its own value. This pragmatic, effect-loaded expression evolves to a standard analytic construction of the syntax (with inversion and no longer with syntagmatic variability (see Section Definition of Grammaticalization above): *habeo inventum cultellum*) and finally enters the morphological paradigm of the verb *invenire* 'to find' as its periphrastic past form (note that at this stage of the Romance languages it is no longer possible to have the word order *habeo cultellum inventum* (**j'ai le couteau trouvé*): the auxiliary has agglutinated to the main verb and no other word can be inserted between the auxiliary and the main verb). As an example of morphophonemic evolution one can quote the French future *dir-ai*, *dir-a*, etc. 'I shall say,' '(s)he shall say' or Catalan *cantaré*, *cantarà*, etc., from periphrastic future forms with *habeo*: *dicere habeo*, *dicere habet* and *cantare habeo*, *cantare habet*, respectively: the auxiliary has undergone phonetic attrition and has become an inflectional ending. Only in Portuguese it is possible to insert a pronoun in future and conditional forms: *comprá-lo-á* '(s)he will buy it,' *lavar-me-ei* 'I shall wash myself' and *lavar-se-iam* 'they would wash themselves,' along with the nowadays more usual *ela o comprará*, *me lavarei*, *se lavariam*.

As in the case of indefinite pronouns presented above, there is a continuum between the phases of the grammaticalization process: Heine et al. (1991: p. 1) quote the case of Ewe (Kwa branch of the Niger–Congo family):

(10) *me-ná ga kofí*
1Sg-give money Kofi
'I gave Kofi money.'

(11) *me-ple βɔtrú ná kofí*
1Sg-buy door give Kofi
a. 'I bought a door and gave it to Kofi.'
b. 'I bought a door for Kofi.'

(12) *me-wɔ dɔ́ vévié ná dodókpɔ́ lá*
1SG-do work hard give exam DEF
'I worked hard for the exam.'

In (10) *ná* has the function of a full verb, meaning 'give'; in (12) *ná* can be interpreted only as a preposition ('for'), whereas in (11) it is ambiguous and allows two different readings. Should one day the Ewe language admit for (11) only the second reading, then we would have a categorial shift (verb > preposition), much in the sense of the evolution of Old English *hwil* 'time' (noun) to *while* (conjunction) (see above, Section Definition of Grammaticalization). At the end of an uninterrupted process there is really a break and a category shift.

Shifts like verb > preposition, noun > conjunction are examples of so-called *transcategorization*. Items belonging to a main (and open) category (verb, noun) pass to a functional (and restricted) category (e.g., adposition, pre- or postposition according to the structure of the language in question, conjunction). These examples clearly hint at a move from a less grammatical to a more grammatical categorial status – it is however clear that the term *transcategorization* (or *recategorization*) is more appropriate than *decategor(ial)ization* used by Hopper (1991: pp. 30f) and others, since the word (e.g., Ewe *ná*) does not fall out of the language categories, but simply passes from one category to another (cf Ježek and Ramat, 2009).

The Dynamic of the Grammaticalization Processes

In such a dynamic typology of grammaticalization it is possible to account also for interrupted grammaticalization processes: in Old Italian it was possible to use movement verbs (*andare* 'go' and *venire* 'come') as auxiliaries in far more cases than in Modern Italian (cf Giacalone Ramat and Hopper, 1998: pp. 2f, with further references). Further, Hopper has drawn attention to what he calls 'emergent grammar,' i.e., to the incipient stages of grammaticalization at the level of discourse. As already shown in the previous section, the negative French adverb *pas* meant originally 'step'; therefore it was used first with movement verbs ('he does not go a step'), but later it spread to verbs such as 'eat,' 'see,' etc. and became, via analogical extension (cf Hopper and Traugott, 2003: pp. 65f), the general mark of negation (specialization principle). Thereby the original meaning of *pas* was completely lost (bleaching process).

As noted by Hopper himself, the principles he suggests (specialization, persistence, divergence, etc.) characterize linguistic change in general and are not distinctive for grammaticalization (Hopper, 1991: p. 21). This has raised the problem of defining the limits of grammaticalization versus other dynamic linguistic processes (cf Giacalone Ramat and Hopper, 1998).

Unidirectionality of Grammaticalization?

Unidirectionality is a much debated question in the literature on grammaticalization. If we define grammaticalization as the

process by which an item or a construction enter the grammar, unidirectionality is true by definition: it is a definitional property (cf Campbell, 2001: pp. 126f). In fact, many linguists maintain that the evolution toward grammar cannot be reversed, and there are no examples of degrammaticalization of forms that belong to the grammatical array of a language (Lehmann, 1995: p. 19; Traugott and Heine, 1991: pp. 4f; Heine et al., 1991: pp. 31f; Haspelmath, 1999, 2004; among others). This is true if by this it is meant that once a lexeme has entered the grammar, it will hardly come back and assume again a lexical value. But there are many instances of items or morphemes leaving the realm of grammar and assuming an autonomous value as lexical entries: Hungarian, Spanish, French, Italian, and many other languages have a large number of present participles used as nouns, e.g., Hungarian *költő* 'poet' (<*költ* 'to compose'), *hallgató* 'student' (<*hallgat* 'to listen to'), Spanish and Italian *cantante* 'singer' (<*cantar(e)* 'to sing'), calmante 'painkiller' (<*calmar(e)* 'to soothe'), French *revenant* 'ghost, spirit' (<*revenir* 'to come back'); nobody would recognize in French *seigneur* (>English *sir*) the accusative form (*seniorem*) of the comparative of Lat. *senex* 'old,' or in English *elder* 'official in the Presbyterian church' a comparative from Old English *ald* 'old.' Both were originally forms belonging to the grammatical paradigm of the adjectives, but later they developed an autonomous lexical meaning. Even syntactic constructs may develop to new lexemes: Dutch *misschien* 'perhaps,' (<(*het*) *mach schien* literally 'it may happen'), Italian *forse* 'perhaps' (<Latin *fors sit* (*an*) literally 'chance be (or)'), English *perhaps* literally 'per chances (plural).' Grammaticalization and lexicalization are two complementary processes (Ramat 1992; Moreno Cabrera 1998; Croft, 2003: pp. 64f), in his functional typology approach speaks of 'competing motivations.'

In Norde's words (2009: p. 120) degrammaticalization can be defined as follows: "Degrammaticalization is a composite change whereby a gram [i.e., a grammatical morpheme, or a construction: P.R.] in a specific context gains in autonomy or substance on more than one linguistic level (semantics, morphology, syntax, or phonology)." Some examples of degrammaticalized forms such as French *seigneur*, Engl. *elder*, have already been adduced above. Degrammaticalization processes are certainly less frequent than grammaticalization processes, but they exist and have to be explained. The causes for their existence are to be sought in the overall tendency to use labels as economic symbols, e.g., suffix morphemes like *ade* (from *orangeade*, *lemonade*, or even *gatorade*) can be separately manipulated and become autonomous lexemes: the label *ade* with the general meaning of 'fruit juice' is more economic than the phrase *fruit juice*, though not as transparent (see *bus*, in *school bus*, *autobus*, etc., as hypernym for 'public transportation vehicle' derived from *omnibus* 'transport vehicle for everyone'). As for constructions becoming degrammaticalized items, we can quote Lithuanian *gal* 'perhaps' along with *galbut* lit. '(it) may be' and Polish *może* lit. '(it) can' compared with Rus. *mozet* '*byt*' '(it) may be': *gal* and *może* are autonomous words meaning 'perhaps,' i.e., instances of grammatical forms which entered the lexicon as independent items. They are shortenings of small syntactic clauses (cf Ramat, 2001). Economic symbolism and noneconomic iconism which tend toward transparency are conflicting forces that are at work in every natural human language (cf Ramat, 1992: p. 557; Newmeyer, 1994: p. 69). If we had just unidirectional evolution toward grammar we would expect that languages become more and more grammaticalized, which by all evidence is not the case (see Moreno Cabrera, 1998: p. 224). On the contrary, we have a double evolution toward grammar as well as toward lexicon (see Kuryłowicz, 1975).

Phonogenesis is also a process which shows that loss of grammatical function may give rise to new lexemes (cf Hopper 1991, 1992). As the German grammatical prefix *ge-*, used to indicate collectives (see *Ge-beine* 'limbs,' plural, *Ge-dränge* 'crowd'), was no longer felt as a prefix, it underwent phonetic erosion to *g-* and gave rise to nouns such as *Glück* 'luck,' *Glied* 'limb,' *Gnade* 'grace' which synchronically are no longer analyzable: this degrammaticalized *g*-became phonetic junk, no longer recognizable as a functional morpheme, and produced new lexemes. In the same vein of reasoning we have to consider also remnants of disappeared paradigms: this applies to isolated dual forms which survived the loss of the dual declension: cf Latin *ambō* 'both' (perhaps also *octō* 'eight') with the old dual marker – *ō* we find in other ancient Indo-European languages (cf Greek *oktṓ*. But in ancient Greek the dual is still part both of the nominal and verbal flection: see *lykō* 'two wolves,' *nṓ* 'we two, the both of us,' *phileîton* 'you two/they two love').

Summing up, we may represent (see **Table 2**) the grammaticalization and degrammaticalization processes along a continuum which move into two opposite directions. At both poles of the scale we have – at the α as well at the β level – either clear lexemes (full words like *habet, perhaps*) or morphemes (inflectional – *á*, Italian and Catalan 3Sg. Fut., or Engl. – *ade* as derivational suffix). Since it is difficult to state when *habet* was delexicalized and the construct *dicere habet* became the norm to express the future, or *ade* was reanalyzed as a second member of a nominal compound, the forms *di(ce)re (h)a(be)t* and *lemon-ade* represent in the previous scheme the gradualness of the transition and are placed midway between (de)lexicalization and (de)morphologization. Transcategorization is a gradual process along a continuum, but its final output represents a break in the categorial status of the input (see Section The Continuum of the Grammaticalization Process).

In recent years the debate concerning grammaticalization has focused on the mechanisms which take part in the process: the role of metaphor and metonymy on the one hand and the role of reanalysis versus analogy on the other, have been particularly stressed: [*habeo*]$_{VB}$ [*inventum cultellum*]$_{NP}$ lit. 'I possess a found knife' was reanalyzed as [*habeo inventum*]$_{VP}$ [*cultellum*]$_{N}$ 'I have found a knife.' Analogical extension has

Table 2 The grammaticalization–degrammaticalization continuum; α represents the evolution toward grammar (grammaticalization) and β the opposite process of degrammaticalization

α.	lexicon ⟶	delexicalization ⟶	morphologization ⟶	morphology
	full words:			full morphemes:
	habet	*di(ce)re (h)a(be)t*		*dir-a*
β.	lexicon ⟵	lexicalization ⟵	demorphologization ⟵	morphology
	full words:			full morphemes:
	ade	*lemon-ade*		*– ade* (*lemon + ade*)
	perhaps			*per* (*+ haps*)

been exemplified above by the case of French *pas,* originally just an NPI, applying only to movement verbs. As for metaphors, a classical example is the English phrase *be going to* as a grammaticalized expression of future, while metonymy refers to semantic changes through contiguity. Hopper and Traugott (2003: pp. 90f) consider the evolution of the already quoted (Section Definition of Grammaticalization) Old English phrase *þa*$_{DAT}$ *hwile*$_{DAT}$ *þe* lit. 'at that time that' to the Modern Engl. conjunction *while* as a metonymical process conveyed by conversational inferences.

See also: Functional Approaches to Grammar; Internal Reconstruction in Linguistics; Parts of Speech.

Bibliography

Batllori, M., Iglesias, N., Martins, A.M., 2005. Sintaxi dels clitics pronominals en català medieval. Caplletra 38, 137–177.

Benveniste, E., 1968. Mutations of linguistic categories. In: Lehmann, W.P., Malkiel, Y. (Eds.), Directions for Historical Linguistics: A Symposium. University of Texas Press, Austin, TX, pp. 83–94.

Bisang, W., Himmelmann, N.P., Wiemer, B. (Eds.), 2004. What Makes Grammaticalization? A Look from Its Fringes and Its Components. Mouton de Gruyter, Berlin and New York.

Bybee, J., Perkins, R., Pagliuca, W., 1994. The Evolution of Grammar, Tense, Aspect, and Modality in the Languages of the World. University of Chicago Press, Chicago.

Campbell, L., 2001. What's wrong with grammaticalization? Language Sciences 23, 113–161.

Campbell, L., Janda, R., 2001. Introduction: conceptions of grammaticalization and their problems. Language Sciences 23, 93–112.

Condoravdi, C., Kiparsky, P., March 11, 2005. Jespersen's cycle: the argument phase. In: Stanford Semfest. Semfest.pdf (application/pdf Object).

Croft, W., 2003. Typology and Universals, second ed. Cambridge University Press, Cambridge.

Davidse, Kr., Vandelanotte, L., Cuykens, H. (Eds.), 2010. Subjectification, Intersubjectification and Grammaticalization. De Gruyter Mouton, Berlin/New York.

Fischer, O., 2000. Grammaticalisation: unidirectional, non-reversable? The case of to before the infinitive in English. In: Fischer, O., Rosenbach, A., Stein, D. (Eds.), Pathways of Change: Grammaticalization in English. Benjamins, Amsterdam, pp. 149–169.

Fischer, O., Rosenbach, A., Stein, D. (Eds.), 2000. Pathways of Change: Grammaticalization in English. Benjamins, Amsterdam/Philadelphia.

van Gelderen, E., 2004. Grammaticalization as Economy. Benjamins, Amsterdam.

Giacalone Ramat, A., Hopper, P. (Eds.), 1998. The Limits of Grammaticalization. Benjamins, Amsterdam.

Giacalone Ramat, A., Sansò, A., 2007. The indefinite usage of *uomo* ('man') in early Italo-Romance. Grammaticalization and areality. Archivio Glottologico Italiano 92, 65–111.

Givón, T., 1979. On Understanding Grammar. Academic Press, New York.

Givón, T., 2009. The Genesis of Syntactic Complexity. Benjamins, Amsterdam/Philadelphia.

Haspelmath, M., 1999. Why is grammaticalization irreversible? Linguistics 37, 1043–1068.

Haspelmath, M., 2004. On directionality in language change with particular reference to grammaticalization. In: Fischer, O., Norde, M., Perridon, H. (Eds.), Up and Down the Cline: The Nature of Grammaticalization. Benjamins, Amsterdam, pp. 17–44.

Heine, B., Claudi, U., Hünnemeier, F., 1991. Grammaticalization. A Conceptual Framework. University of Chicago Press, Chicago.

Heine, B., Kuteva, T., 2002. World Lexicon of Grammaticalization. Cambridge University Press, Cambridge.

Heine, B., Kuteva, T., 2005. Language Contact and Grammatical Change. Cambridge University Press, Cambridge.

Heine, B., Song, K.-A., 2011. On the grammaticalization of personal pronouns. Journal of Linguistics 47, 587–630.

Hopper, P., 1987. Emergent grammar. Berkeley Linguistic Society 13, 139–157.

Hopper, P., 1991. On some principles of grammaticization. In: Traugott, E., Heine, B. (Eds.), Approaches to Grammaticalization, vol. 1. Benjamins, Amsterdam, pp. 17–35.

Hopper, P., 1992. Phonogenesis. In: Pagliuca, P. (Ed.), Perspectives on Grammaticalization. Benjamins, Amsterdam, pp. 119–135.

Hopper, P., Traugott, E. Cl., 2003. Grammaticalization, second ed. Cambridge University Press, Cambridge.

Jespersen, O., 1917. Negation in English and Other Languages. Høst, Copenhagen.

Ježek, E., Ramat, P., 2009. On parts-of-speech transcategorization. Folia Linguistica 43, 391–416.

Kuryłowicz, J., 1975. The evolution of grammatical categories. In: Esquisses Linguistiques, vol. 2. Fink, Munich, Germany, pp. 38–54.

Lehmann, Chr, 1995. Thoughts on Grammaticalization. A Programmatic Sketch. Lincom Europa, Munich, Germany (First published in the Arbeiten des Kölner Universalien-Projekts, Vol. 48, 1982).

López-Couso, M.J., Seoane, E. (Eds.), 2008. Rethinking Grammaticalization. New Perspectives. Benjamins, Amsterdam/Philadelphia.

Meillet, A., 1948. L'évolution des formes grammaticales. In: Meillet, A. (Ed.), Linguistique Historique et Linguistique Générale. Champion, Paris, pp. 130–148 (First published in 1916).

Moreno Cabrera, J.C., 1998. In: Giacalone Ramat, A., Hopper, P. (Eds.), On the Relationships between Grammaticalization and Lexicalization, pp. 211–227.

Narrog, H., Heine, B., 2011. The Oxford Handbook of Grammaticalization. Oxford University Press, Oxford.

Newmeyer, Fr. J., 1994. Competing motivations and synchronic analysis. Sprachtypologie und Universalienforschung 47, 67–77.

Newmeyer, Fr. J., 1998. Language Form and Language Function. MIT Press, Cambridge, MA.

Norde, M., 2009. Degrammaticalization. Oxford University Press, Oxford.

Ramat, P., 1992. Thoughts on degrammaticalization. Linguistics 30, 549–560.

Ramat, P., 2001. Degrammaticalization or transcategorization? In: Schaner-Wolles, Chr., Rennison, J., Neubarth, Fr. (Eds.), Naturally! Linguistic Studies in Honour of Wolfgang Ulrich Dressler Presented on the Occasion of His 60th Birthday. Rosenberg & Sellier, Turin, pp. 393–401.

Ramat, P., 2006. Italian negatives from a typologycal/areal point of view. In: Grandi, N., Iannàccaro, G. (Eds.), Zhì. Scritti in onore di Emanuele Banfi in occasione del suo 60° compleanno. Caissa Italia editore, Cesena/Roma, pp. 355–370.

Roberts, J., Roussou, A., 2003. Syntactic Change: A Minimalist Approach to Grammaticalization. Cambridge University Press, Cambridge.

Sapir, E., 1921. Language. An Introduction to the Study of Speech. Harcourt. Brace, New York.

Stathi, K., Gehweiler, E., König, E. (Eds.), 2010. Grammaticalization. Current Views and Issues. Benjamins, Amsterdam/Philadelphia.

Traugott, E.C., Heine, B. (Eds.), 1991. Approaches to Grammaticalization. Benjamins, Amsterdam.

Traugott, E.C., König, E., 1991. On the expression of spatio-temporal relations. In: Greenberg, J., Ferguson, Ch., Moravcsik, E. (Eds.), Universals of Human Language, vol. 3. Stanford University Press, Stanford, CA, pp. 369–400.

Traugott, E.C., Trousdale, Gr (Eds.), 2010. Gradience, Gradualness and Grammaticalization. Benjamins, Amsterdam.

Van linden, A., Verstraete, J.-Chr., Davidse, Kr. (Eds., in collaboration with Hubert Cuyckens), 2010. Formal Evidence in Grammaticalization Research. Benjamins, Amsterdam/Philadelphia.

Verhoeven, E., Skopeteas, St., Shin, Y.-M., Helmbrecht, J. (Eds.), 2008. Studies in Grammaticalization. Mouton de Gruyter, Berlin/New York.

Vincent, N., Börjars, K., 2010. In: Traugott, E.C., Trousdale, Gr. (Eds.), Grammaticalization and Models of Language.

Whitney, W.D., 1875. The Life and Growth of Language. King & Co., London.

Grandparent–Grandchild Relationships

Teun Geurts, Ministry of Security and Justice, Research and Documentation Centre, The Hague, The Netherlands
Theo van Tilburg, VU University Amsterdam, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A brief review of major conceptual notions and empirical findings within the literature on grandparent–grandchild relationships is presented. Four major topics for understanding the intergenerational relationship are addressed: the historical context, the importance of the relationship, changes over individual time, and culture and variation. The focus is on grandparents and grandchildren from Western societies and who are biologically connected.

Historical Context

In the popular and academic literature it is frequently presumed that grandparent–grandchild relationships have become more important in Western societies than ever before. This idea is often motivated by greater availability of grandparents due to increased length of life and decreased fertility (e.g., Bengtson, 2001). Due to these demographic changes, the lives of grandparents and grandchildren overlap for a longer period of time and there are fewer grandchildren per grandparent than in previous times (Uhlenberg, 2009). The demographic changes allow for a more intense relationship between grandparents and grandchildren for a longer period of time (Bengtson, 2001).

Greater freedom among contemporary grandparents strengthened the idea of gained importance. The period of time in which older adults are in good health and unfettered by child care responsibilities and work obligations increased, due to better health care and the introduction of pensions funds in many countries during the 1970s and the 1980s (Laslett, 1991). Although current developments of pension reform aim to reduce opportunities for early retirement and to raise mandatory retirement age to deal with the economic pressures coming along with an aging population, the amount of free time for most people in old age in current Western societies is greater than ever before. New concepts were coined to capture and describe this distinct life phase for older adults, such as the 'third age' (Laslett, 1991). Although increased free time may have induced involvement in social roles that compete with extended family relationships (Silverstein and Long, 1998), having free time is often mentioned as one of the reasons for greater involvement with grandchildren (Herlofson and Hagestad, 2012).

The idea of greater importance of grandparent–grandchild bond is reinforced by several other economic and family transformations in the past decades (Herlofson and Hagestad, 2012). The most important economic transformation is increased employment among women. In earlier cohorts, mothers typically stayed at home to take care of children and household. Contemporary mothers of young children often continue to work at least part-time or return to the labor market after short breaks (Vlasblom and Schippers, 2006). Due to emancipation and a shift in gender roles, female labor participation is greater than before and, alongside with it, the need for child care from beyond the nuclear family increased. As grandparents can be called upon for assistance more than ever before, it is often assumed that their availability in combination with the parent's need for child care increased grandparents' involvement with grandchildren (e.g., Fergusson et al., 2008).

As to family transformations, the increase in divorce among parents is often referred to as the most important transformation affecting the importance of grandparent–grandchild relationships, predominantly because divorce increases the need for child care (Bengtson, 2001). In families of divorce, child care is not easily shared with the former partner, and for that reason, grandparents are more often needed as childminders, in particular when the mother is employed. Hank and Buber (2009) observed across 10 European countries that grandparents are more likely to provide child care when the parent is single than when the parent has a partner. Other family transformations that led to greater need for grandparental involvement is an increase in the number of poor functioning parents, in particular in the United States (Uhlenberg and Cheuk, 2010). In such families, grandparents may take over the parental responsibilities and act as surrogate parents due to drug abuse, illness, incarceration, or poverty in the middle generation. These grandparent-headed households – sometimes referred to as 'skipped-generation' households – modestly increased in the second half of the twentieth century (Uhlenberg and Cheuk, 2010). The restructuring of gender roles and increase in poor functioning parents created the need and opportunity for grandparents to become more involved in one of the most prominent functions of the family as an institute: socialization and nurturance of a child.

Despite the widespread belief of greater importance of the grandparent–grandchild relationship in family life, only few studies exist that examined trends over historical time. Data that allow for examining of changes over historical time, in particular grandparental involvement with grandchildren, are scarce. The few studies that do exist suggest both gained and reduced importance. On the one hand, a study from Finland suggests that older people from an earlier generation generally had more frequent contact with grandchildren than grandparents from more recent cohorts (Lyyra et al., 2010). Likewise, research from Silverstein and Long (1998) indicates that earlier cohorts of grandparents from the United States

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23219-4

had more frequent contact with their adult grandchildren than later cohorts. On the other hand, an increase in grandparental child care provision was observed in the Netherlands (Geurts et al., 2014): grandparental child care substantially increased between 1992 and 2006 and this was linked to increased needs on the part of adult daughters due to higher employment rates and a higher rate of single motherhood. The increase also appeared to reflect greater opportunities of grandparents to provide care as indicated by decreased travel time and less competition among grandchildren because recent grandparents have fewer of them. A similar trend of increased child care provision by other family members – most likely grandparents – was observed in Great Britain (Gray, 2005).

Although empirical support for the assumption of greater importance of grandparent–grandchild relationships in family life is scarce and results are mixed, an abundance of scholarly research show that the sheer availability and freedom of grandparents increased, that the need for grandparental child care increased, and that many of today's grandparents are involved with their grandchildren in terms of child care provision (Uhlenberg, 2009; Fuller-Thomson and Minkler, 2001). These observations provide good reasons to believe that, in many families of Western societies, the grandparent–grandchild relationships are more important than ever before – in particular during the grandchild's childhood.

Importance of the Relationship

The importance of the grandparent–grandchild relationship in contemporary Western societies is mainly given shape through the meaning grandparents attach to their role (Reitzes and Mutran, 2004). Depending on personal opportunities and preferences, grandparents give shape to their role, which reflects their own interpretation of what it means to be a grandparent. Hayslip et al. (2003) argue that the meaning of this role includes a sense of continuity (carrying on the family line, or the sense of living on through the lives of grandchildren), extension of the self (by feeling valued as an elderly person or by vicarious accomplishments through the grandchildren), and satisfaction (by contributing to their grandchild's well-being through help, advice, or indulgence). In addition, grandchildren can be a source of pride and grandparents can derive enjoyment and companionship from the relationship with their grandchildren. The grandparents' freedom in the interpretation of their role is often stressed in the academic literature by referring to what is called 'a roleless' role: a social status without clear cultural expectations and prescriptions. There is only one study that suggests that culture plays a significant role (Herlofson and Hagestad, 2012).

Because of unclear cultural norms about how this role should be enacted, there is a great variety in how grandparents enact their role and therefore also in the importance of the relationship. Earlier research suggested that grandparents can enact their roles by being a 'family watchdog' (Troll, 1983), nurturer, mentor, family historian, and role model (Kornhaber and Woodward, 1981). More recent research examined the content of the relationships between grandparents and grandchildren (contact frequency, activities, intimacy, instrumental help, and authority/discipline) and developed a typology of five grandparenting styles: influential, supportive, passive, authority oriented, and detached (Mueller and Elder, 2003).

The importance of the relationship is mainly manifested through its importance for the middle generation. That is, grandparents can play a significant role in child care provision and, in some cases, even the upbringing of their child. As time beyond working hours is often a luxury commodity, in particular for dual-income parents, grandparental child care provision is presumably of high value to parents, in particular because parents generally find child care provision from their parents more convenient, more beneficial to their child, more trustworthy, and less expensive than care from other child minders (Fergusson et al., 2008). Furthermore, it enables women to be employed outside of the house because it eases reconciliation of child care with work. It is well documented that the intergenerational relationship is important during grandchild's childhood (e.g., Fergusson et al., 2008; Fuller-Thomson and Minkler, 2001; Hank and Buber, 2009). Through their child care activities, grandparents contribute to their family, the society (by increasing employment opportunities for women), and the welfare state (by producing support functions that are absent or would otherwise be produced by the welfare state).

Little attention, however, has been given to the importance of the intergenerational relationship when grandchildren are adults. It is known that the frequency of contact between grandparents and grandchildren declines when grandchildren grow older (Silverstein and Long, 1998). Other studies examining the significance of the intergenerational relationship concluded that the relationship continues to be highly valued, personally meaningful, and potentially important (Kemp, 2005). Adult grandchildren may even contribute to their grandparents' well-being by providing emotional and practical support (Fruhauf et al., 2006). For example, adult grandchildren may introduce new technologies and instruct their grandparents in how to operate them. In this way, grandparents can learn about societal developments through contact with their adult grandchildren. Although such support can also be provided by other young people, for instance, in the context of intergenerational programs, adult grandchildren are more easily approached because many older people maintain contact with younger generations only within the family context. Furthermore, grandchildren, and in particular adult grandchildren, may assist a grandparent in need of care (e.g., Fruhauf et al., 2006), even though they are generally not the first in order of preferred care providers.

To understand the importance of the relationship across generations in Western societies, researchers commonly draw on the intergenerational solidarity framework developed by Bengtson and Roberts (1991). This model distinguishes between structural, consensual, functional, associational, affectual, and normative solidarity. Structural solidarity refers to factors that facilitate or hinder the opportunity for contact between generations. Consensual solidarity indicates the amount of agreement in beliefs and values. Functional solidarity refers to the amount of help and assistance within the intergenerational relationship. Associational solidarity concerns the frequency of contact and shared activities between

the generations. Affectual solidarity involves the amount of emotional closeness as perceived by both generations. Normative solidarity refers to obligations felt regarding the other party in the relationship and expectations regarding the content of the relationship.

Although the intergenerational solidarity model was originally developed to understand parent–adult child relationships (Bengtson and Roberts, 1991), it is applied to grandparent–grandchild relationships as well (e.g., Bengtson, 2001; Silverstein and Marenco, 2001). In particular research focusing on grandparent–adult grandchild relationships uses the model to understand the intergenerational relationship (e.g., Mills, 1999). The model is, however, less suited to understand connectedness during the grandchild's childhood. Because the focus is on dyadic relationships, it is unclear how intergenerational relationships should be understood when it is mediated by a third party. For instance, when grandparents provide child care, should this be understood in terms of solidarity between grandparent–adult child relationships, grandparent–grandchild relationships, or both?

The concern of applying the intergenerational solidarity model to grandparent–grandchild relationships underscores the importance of keeping in mind that the generation in between plays a key role in connecting grandparents and grandchildren. That is, they often serve as a lineage bridge between grandparents and grandchildren by either facilitating or hindering the contact. Likewise, parents shape the opportunity structure for intergenerational contact, for instance, through their choices regarding residential location, parental divorce and the quality of the parent–grandparent relationship (Uhlenberg and Hammill, 1998). Over individual time, however, the role of the middle generation as mediators between grandparents and grandchildren presumably declines in importance. Adult grandchildren can maintain contact with their grandparents independently from their parents and may reestablish the relationship on the basis of their own and their grandparents' terms (Kemp, 2005).

Changes over Individual Time

The grandparent–grandchild relationship is considered to be the most intense before the grandchildren have reached adolescence (Cherlin and Furstenberg, 1986). Although the relationship may continue to be personally meaningful and significant for grandchildren and grandparents in and after adolescence (Kemp, 2005), the relationship presumably becomes less intense when grandchildren grow older for two reasons. First, the initiative for maintaining contact is likely to shift from parents and grandparents to grandchildren. During childhood and early adolescence, parents are most important, as they initiate and facilitate contact with grandparents (Brown, 2003). When grandchildren enter adulthood, the parental influence on the grandparent–grandchild relationship is assumed to become less important and grandchildren may reestablish the relationship on their own terms. Although the grandparents' need for family contact may increase, as these contacts are considered to be more emotionally rewarding at an older age (Carstensen, 1992), grandparents also believe they should not interfere in the lives of younger generations and may be reluctant to contact their grandchildren (Kemp, 2005).

Second, grandchildren's priorities and opportunities for maintaining contact with grandparents decrease. Grandchildren in early adulthood are likely to prefer peer relationships over intergenerational relationships because they place more emphasis on the potential for information gain and future contact (Carstensen, 1992). Moreover, grandchildren face more time restrictions as they take up adult roles such as starting their own families or pursuing careers (Mills, 1999). Kemp (2005) observed that adult grandchildren frequently use their busy lives as a legitimate excuse for not contacting their grandparents, supporting the assumption that limited time restricts grandchildren to contact their grandparents. The reduced importance of parents and grandparents, combined with grandchildren's weaker preferences and fewer opportunities for intergenerational contact, is assumed to weaken the grandparent–grandchild relationship when grandchildren enter adulthood. Cherlin and Furstenberg (1986) even go so far as to say that it evolves into a relationship with limited meaning and little content. Most studies have focused on young grandchildren and showed a decline in contact frequency during adolescence (e.g., Silverstein and Marenco, 2001). The few studies that track grandchildren beyond adolescence also suggest a decline (e.g., Geurts et al., 2009; Mills, 1999; Silverstein and Long, 1998).

The development of the relationship into one of low intensity may at least partly be counteracted by an intense relationship during childhood. The relationship in childhood is of importance because discrepancies in developmental stages are likely to hinder continuation of the relationship when both parties age. Older generations strive to maintain continuity in the intergenerational relationship, while younger generations tend to exaggerate differences to facilitate separation from the family of origin (Harwood, 2001). A strong bond between the parties at an early stage of the relationship may hold back a decline in relationship intensity, even though discrepancies in the developmental stages become increasingly pronounced. Results from three studies support the idea that a strong bond established during childhood positively affects the bond in adulthood. Taylor et al. (2005) observed more positive perceptions and greater satisfaction in the intergenerational relationship if adult grandchildren had coresided with their grandparents during childhood. Brown's (2003) study observed that the quality of adult grandchild–grandparent relationships was higher when grandparents had intensively cared for their grandchildren during childhood. The presumed link between childhood intensity on later relationship outcomes was also supported by a study by Geurts et al. (2014), who observed that more overnight visits, contacts, and child care increased the likelihood that grandparents identify adult grandchildren as a personal and important contact.

Variation between Countries

Grandparent–grandchild relationships take shape within a social-cultural context. Economical and political regimes shape the social conditions that are assumed to play a major

role in variation in intergenerational relationships; it influences the allocation of care responsibilities among state, market, and family (Igel and Szydlik, 2011). Although nonparental child care is often dictated by the reconciliation of mothers' care responsibility with paid employment, families with children at risk can mobilize nonparental support as well. In such families, grandparents may take over some or even all of the parental responsibilities. In the following, variation in grandparental child care provision is briefly described by looking at Europe, China, South Korea, Russia, and the United States.

Across European countries, 58% of grandmothers and 49% of grandfathers took care of a grandchild aged 15 years or younger in 2004 (Hank and Buber, 2009). In European countries with extensive public child care arrangements, grandparental child care is often complementary to these public services. That is, grandparents are occasionally called upon for child care assistance but are generally not needed for daily child care. In European countries with limited public child care arrangements, however, grandparental child care provision is predominantly a substitute for absent services. In particular for full-time employed mothers, grandparents generally provide daily rather than occasional child care (Hank and Buber, 2009).

In China, 56% of grandparents provide child care (Ko and Hank, 2014). Labor migration plays a significant role: about 20% of Chinese elders in rural areas with high migration rates provide full care of their grandchildren (Silverstein et al., 2006). In South Korea, grandparental child care is less common (6% of grandparents; Ko and Hank, 2014). Because child care is widely considered as a primary task for the family, Korean mothers prefer to rely on relatives rather than on public or privately paid services (Ko and Hank, 2014). Russian grandparents – in particular grandmothers – play a significant role in the upbringing of children: One-third of the young adults born in the 1990s grew up with a grandparent in the household (Semenova and Thompson, 2004). Grandparents' significance increased due to increased numbers of full-time employed single mothers, increased poverty, and decreased state provision of child care (Lokshin et al., 2000).

In the United States, 24% of grandparents engage in caregiving activities between 10 and 29 h per week (Fuller-Thomson and Minkler, 2001). About half of the children with employed mothers who received grandparental child care received this child care combined with another arrangement (Uhlenberg and Cheuk, 2010). Because public-financed child care is limited, this grandparental child care is presumably complementary to private services. Next to grandparental child care services in relation to mothers' employment, a substantial number of grandparents also provide non-work-related care. See Hayslip and Kaminski (2005) for an overview on custodial grandparenting in the United States.

Finally, the authors stress that while the coverage in this article is restricted mainly to advanced Western societies, intergenerational relationships of the sort that is discussed are at the very core of extended family life in most of the rest of the world as well. Grandparents parenting their grandchildren are by no means restricted to low-income households in Western urban societies.

See also: Aging, Theories of; Emotions and Aging; Population Aging: Economic and Social Consequences.

Bibliography

Bengtson, V.L., Roberts, R.E.L., 1991. Intergenerational solidarity in aging families: an example of formal theory construction. Journal of Marriage and Family 53, 856–870.

Bengtson, V.L., 2001. Beyond the nuclear family: the increasing importance of multigenerational bonds. Journal of Marriage and Family 63, 1–16.

Brown, L.H., 2003. Research and evaluation intergenerational influences on perceptions of current relationships with grandparents. Journal of Intergenerational Relationships 1, 95–112.

Carstensen, L.L., 1992. Social and emotional patterns in adulthood: support for socioemotional selectivity theory. Psychology and Aging 7, 331–338.

Cherlin, A.J., Furstenberg, F.F., 1986. The New American Grandparent: A Place in the Family, a Life Apart. Basic Books, New York.

Fergusson, E., Maughan, B., Golding, J., 2008. Which children receive grandparental care and what effect does it have? Journal of Child Psychology and Psychiatry 49, 161–169.

Fruhauf, C.A., Jarrott, S.E., Allen, K.R., 2006. Grandchildren's perceptions of caring for grandparents. Journal of Family Issues 27, 887–911.

Fuller-Thomson, E., Minkler, M., 2001. American grandparents providing extensive child care to their grandchildren. Gerontologist 41, 201–209.

Geurts, T., Poortman, A., Van Tilburg, T.G., Dykstra, P.A., 2009. Contact between grandchildren and their grandparents in early adulthood. Journal of Family Issues 30, 1698–1713.

Geurts, T., Van Tilburg, T.G., Poortman, A., 2012. The grandparent-grandchild relationship in childhood and adulthood: a matter of continuation? Personal Relationships 19, 267–278.

Geurts, T., Van Tilburg, T.G., Poortman, A., Dykstra, P.A., 2014. Childcare by grandparents: changes between 1992 and 2006. Ageing & Society.

Gray, A., 2005. The changing availability of grandparents as carers and its implications for childcare policy in the UK. Journal of Social Policy 34, 557–577.

Hank, K., Buber, I., 2009. Grandparents caring for their grandchildren: findings from the 2004 survey of health, ageing, and retirement in Europe. Journal of Family Issues 30, 53–73.

Harwood, J., 2001. Comparing grandchildren's and grandparents' stake in their relationship. International Journal of Aging and Human Development 53, 195–210.

Hayslip, B., Kaminski, P.L., 2005. Grandparents raising their grandchildren: a review of the literature and suggestions for practice. Gerontologist 45, 262–269.

Hayslip, B., Henderson, C.E., Shore, R.J., 2003. The structure of grandparental role meaning. Journal of Adult Development 10, 1–11.

Herlofson, K., Hagestad, G.O., 2012. Transformations in the role of grandparents across welfare states. In: Arber, S., Timonen, V. (Eds.), Contemporary Grandparenting: Changing Family Relationships in Global Contexts. The Policy Press, Bristol, pp. 25–51.

Igel, C., Szydlik, M., 2011. Grandchild care and welfare state arrangements in Europe. Journal of European Social Policy 21, 210–224.

Ko, P.C., Hank, K., 2014. Grandparents caring for grandchildren in China and Korea: findings from CHARLS and KLoSA. The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences 69, 646–651.

Kornhaber, A., Woodward, K.L., 1981. Grandparents, Grandchildren: The Vital Connection. Anchor, New York.

Kemp, C.L., 2005. Dimensions of grandparent-adult grandchild relationships: from family ties to intergenerational friendships. Canadian Journal on Aging 24, 161–177.

Kreider, R.M., 2008. Living Arrangements of Children, 2004. US Department of Commerce, Economics and Statistics Administration, US Census Bureau.

Laslett, P., 1991. A Fresh Map of Life: The Emergence of the Third Age. Harvard University, Cambridge, MA.

Lokshin, M., Harris, K.M., Popkin, B.M., 2000. Single mothers in Russia: household strategies for coping with poverty. World Development 28, 2183–2198.

Lyyra, T.M., Lyyra, A.L., Lumme-Sandt, K., Tiikkainen, P., Heikkinen, R.L., 2010. Social relations in older adults: secular trends and longitudinal changes over a 16-year follow-up. Archives of Gerontology and Geriatrics 51, e133–e138.

Mills, T.L., 1999. When grandchildren grow up: role transition and family solidarity among baby boomer grandchildren and their grandparents. Journal of Aging Studies 13, 219–239.

Mueller, M.M., Elder, G.H., 2003. Family contingencies across the generations: grandparent–grandchild relationships in holistic perspective. Journal of Marriage and Family 65, 404–417.

Reitzes, D.C., Mutran, E.J., 2004. Grandparent identity, intergenerational family identity, and well-being. The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences 59B, S213–S219.

Semenova, V., Thompson, P., 2004. Family models and transgenerational influences: grandparents, parents and children in Moscow and Leningrad from the Soviet to the market era. In: Bertaux, D., Rotkirch, A., Thompson, P. (Eds.), On Living through Soviet Russia. Taylor & Francis, New York, pp. 120–145.

Silverstein, M., Cong, Z., Li, S., 2006. Intergenerational transfers and living arrangements of older people in rural China: consequences for psychological well-being. The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences 61, S256–S266.

Silverstein, M., Long, J.D., 1998. Trajectories of grandparents' perceived solidarity with adult grandchildren: a growth curve analysis over 23 years. Journal of Marriage and Family 60, 912–923.

Silverstein, M., Marenco, A., 2001. How Americans enact the grandparent role across the family life course. Journal of Family Issues 22, 493–522.

Taylor, A.C., Robila, M., Lee, H.S., 2005. Distance, contact, and intergenerational relationships: grandparents and adult grandchildren from an international perspective. Journal of Adult Development 12, 33–41.

Troll, L.E., 1983. Grandparents: the family watchdogs. In: Brubaker, T.H. (Ed.), Family Relationships in Later Life. Sage, Beverly Hills, pp. 63–74.

Uhlenberg, P., 2009. Children in an aging society. The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences 64B, 489–496.

Uhlenberg, P., Cheuk, M., 2010. The significance of grandparents to grandchildren: an international perspective. In: Dannefer, D., Phillipson, C. (Eds.), Sage Handbook of Social Gerontology. Sage, London, pp. 447–458.

Uhlenberg, P., Hammill, B.G., 1998. Frequency of grandparent contact with grandchild sets: six factors that make a difference. Gerontologist 38, 276–285.

Vlasblom, J.D., Schippers, J., 2006. Changing dynamics in female employment around childbirth: evidence from Germany, the Netherlands and the UK. Work, Employment and Society 20, 329–347.

Graphical Markov Models: Overview

Nanny Wermuth, Chalmers University of Technology, Gothenburg, Sweden and Medical Psychology and Medical Sociology, Gutenberg-University, Mainz, Germany
David R Cox, Nuffield College, Oxford University, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

We describe how graphical Markov models emerged in the past 40 years, based on three essential concepts that had been developed independently more than a century ago. Sequences of joint or single regressions and their regression graphs are singled out as being the subclass that is best suited for analyzing longitudinal data and for tracing developmental pathways, both in observational and in intervention studies. Interpretations are illustrated using two sets of data. Furthermore, some of the more recent, important results for sequences of regressions are summarized.

Some General and Historical Remarks on the Types of Model

Graphical models aim to describe in concise form the possibly complex interrelations between a set of variables so that key properties can be read directly off a graph. The central idea is that each variable is represented by a node in a graph. Any pair of nodes may become coupled, that is joined by an edge. Coupled nodes are also said to be adjacent. For many types of graph, a missing edge represents some form of conditional independence between the pair of variables and an edge present can be interpreted as a corresponding conditional dependence. Because the conditioning set may be empty, or may contain some or all of the other variables, a variety of types of graph have been developed and are used to represent different types of structure.

A particularly important distinction is between directed and undirected edges. In the former, an arrow indicates the direction of dependence of a response on an explanatory variable; the latter is also called a regressor. If, on the other hand, two variables are to be interpreted on an equal standing then the edge between them is typically undirected. For instance, systolic and diastolic blood pressure values are treated as being on equal standing because they are two aspects of a single phenomenon, namely a blood pressure wave.

Graphical Markov models started to be developed after 1970 as special subclasses of log–linear models for contingency tables and of joint Gaussian distributions, where conditional independence constraints are imposed such that conditioning is on all the other variables; see Darroch et al. (1980), Wermuth (1976, 1980). These models are typically represented by undirected graphs with edges that are full lines, called nowadays concentration graphs. The same type of graph were used by the physicist Gibbs (1902) to describe for two systems of particles having the same node sets, one as more complex whenever its nodes have more edges, that is larger numbers of 'nearest neighbors.'

The first extension was to situations in which the variables can be arranged recursively, that is in an ordered sequence, so that each variable turns into a single response to variables in its past and may be explanatory only to other variables in its future; see Wermuth (1980), Wermuth and Lauritzen (1983). This led for single responses to what are now called directed acyclic graphs and for sequences of joint or single responses to the so-called regression graphs, and furthermore to distributions said to be generated over graphs.

For such generated Gaussian distributions, these models include as a subclass the path analyses of the geneticist Wright (1923, 1934). Wright had studied, for his data, sequences of exclusively linear regressions as possible stepwise generating processes. In his graphs, each missing arrow corresponds to the vanishing of the partial correlation coefficient given all remaining variables in the past of a given response, which in Gaussian distributions is a conditional independence constraint. For other types of generated distributions, an important issue is to identify parametric consequences of conditional independences; as for instance for the so-called CG-regressions, see Lauritzen and Wermuth (1989) and Edwards and Lauritzen (2001).

In general, the vanishing of a correlation coefficient may coexist with a nonlinear dependence. Conversely, a substantial partial correlation can occur in spite of conditional independence; for an example, see Wermuth and Cox (1998). Currently, one knows also for jointly symmetric binary variables, generated over a directed acyclic graph, that an arrow vanishes in the graph if and only if there is a corresponding zero partial correlation; see Wermuth et al. (2009). Zero partial correlations given all other variables may reflect conditional independences more generally, provided the distribution is generated over even more specialized types of graph; see Loh and Wainwright (2013), Wermuth et al. (2014), and Wermuth and Marchetti (2014).

The next extensions were to variables of any type so that a missing edge corresponds to a conditional independence. This exploits a proposal by the probabilist Markov (1912): many types of seemingly complex joint distribution may be much simplified by conditional independences. Important issues are defining sets of independences for a graph and criteria to derive all implied independences so that an independence structure captured by the graph becomes well understood; see the section 'Questions Regarding Applications And Statistical Research.'

Further developments include different types of chain graph models; see Cox and Wermuth (1993) and Drton (2009).

The name 'chain graph' reflects a full ordering of the variables into a sequence of joint or single responses and possibly a set of remaining variables that is to capture the context of a study or properties of individuals given at the baseline and hence regarded as given. For parameterizations of Gaussian distributions with different types of chain graph see Wermuth et al. (2006).

Statistical monographs documenting the early development of what are now called graphical Markov models are by Whittaker (1990), Edwards (1995), Lauritzen (1996), and Cox and Wermuth (1996). For surveys of probabilistic aspects, see Pearl (1988) and Studený (2005). For graphical models used in expert systems, see Cowell et al. (1999) and for machine learning, see Wainwright and Jordan (2008).

Special theoretical results, not touched upon here, have been derived under the assumption that a given probability distribution satisfies all independence constraints represented by a graph and no more. Such distributions have been said to be 'faithful to the graph'; see Spirtes et al. (1993).

The existence of a family of faithful distributions means that one can choose a member at random, constrained just by the independences specified by a graph, and this distribution captures precisely the independence structure of the graph. But such an existence result is of no relevance when other constraints are known for subsets of parameters. For instance, the whole family of Gaussian distributions is faithful to concentration graphs, but there is a simple subfamily with additional parameter constraints that is always 'unfaithful' to concentration graphs, see Wermuth (2012: Section 2.5).

More importantly, for data analyses, in which it is crucial to respect known important dependencies, such an existence is also of no relevance. In most research settings, much is known about subsets of parameters because the direction and strength of some dependences derive from substantive theory or from results in previous empirical studies.

Sequences of Regressions and Data Generating Processes

Directed acyclic graphs are a common subclass of all currently known types of chain graph. Distributions generated over a directed acyclic graph result in a stepwise fashion in terms of univariate conditional distributions, also called recursive single-response regressions. These may be compatible with causal interpretations and, therefore, appear attractive, at first sight, for substantive research that is driven by causal hypotheses.

It is however well understood, from randomized clinical trials in particular, that real interventions often result in modifying more than a single response and possibly other features. For instance, a medication that is to reduce blood pressure, will typically affect both systolic and diastolic pressure values. Hence, models that are to be compatible with causal interpretations should contain response variables that may respond at the same time to a change in relevant explanatory variables. This is possible with joint response regressions.

Nevertheless, much of the current literature on causal modeling is based on directed acyclic graphs and on only virtual interventions; see Pearl (2009). With such a virtual intervention, those changes in single response are recorded that result by fixing the values of some variables. This may for instance be treatments. One removes arrows pointing to these variables in a given graph and estimates treatment effects given these additional independence constraints. Implicit is the assumption that a given joint family of distribution remains unchanged, otherwise.

These recorded changes in virtual interventions, even though they are often called 'causal effects,' may tell next to nothing about actual effects in real interventions with, for instance, completely randomized allocation of patients to treatments. In such studies, independences result by design and they lead to missing arrows in well-fitting graphs; see, for example, **Figure 9** below in the last section.

Sequences of regressions in joint and single responses are the most attractive chain graph models for observational as well as intervention studies. For these models, residual associations among the responses may often be regarded as secondary features. But, standard univariate estimation methods for regression coefficients may lead to distorted estimates whenever two associated joint responses on equal standing have disjoint subsets of regressors, a situation that has become known as seemingly unrelated regressions. When strong residual dependences remain after one has been regressing each response component separately on two different sets of regressors, then distorted parameter estimates may result, see Haavelmo (1943), Zellner (1962), and Drton and Richardson (2004).

For known types of chain graph other than regression graphs, recursive generating processes may not exist since response components are conditioned on other joint responses on equal standing. While substantive theory may occasionally support such models, this type of conditioning is typically counterintuitive whenever one main aim of an analysis is to identify consequences of interventions or early predictors for joint responses. But now, graphical criteria are available to decide whether the regression graph of a given sequence of joint or single regressions defines the same independence structure as another chain graph, that is whether they are Markov equivalent; see Wermuth and Sadeghi (2012: Section 7), and here the section 'Markov Equivalence of Regression Graphs.'

For applications, arguably the most important developments concern conditions under which the regression graph can be used to trace development, to predict implied dependences in addition to implied independences, as well as graphical criteria for possible confounding effects when some variables are ignored or subpopulations are studied.

Regression Graphs as Partial Summaries of Data Analyses

We give here two examples of data for which corresponding detailed statistical analyses have been published; see the Appendix in Wermuth and Sadeghi (2012) for the first and the Appendix in Wermuth and Cox (2013) for the second set of data.

In the regression graphs derived for both examples, the dashed lines capture remaining associations among two joint responses and full lines, for the context variables, reflect that conditioning is on all the context variables simultaneously. Estimation is for both data sets by local modeling of each response alone on its past and of response pairs on the union

of their directly explanatory variables. When there are only independence constraints concerning responses and no seemingly unrelated regressions, local maximum-likelihood (m.l.) estimates lead directly to proper m.l. estimates for the joint distribution; see Cox (2006) for such types of general principles.

In both the following examples, data analyses lead to traceable regressions; see Wermuth (2012). This means that for a given ordered sequence of single and joint responses, the resulting regression graph permits one not only to read off independences, but also more importantly for most applications, sequences of directed edges show pathways of development.

Tracing of paths is always possible for joint Gaussian distributions whenever each edge in its regression graph represents a substantial conditional dependence; see the section Questions Regarding Applications And Statistical Research. This may not be possible with graphs of some structural equation models; see Wermuth (1992). More importantly, sufficient conditions for traceability using just the graph, are now known for other types of distributions generated over regression graphs.

Data on Chronic Pain Treatment

For some medical data, Figure 1 shows a first ordering of the variables. This ordering was derived in discussions by physicians and statisticians, prior to the analyses.

We are grateful to Judith Kappesser, now Department of Psychology, Liebig University of Giessen, for letting us use her data. They are for 201 chronic pain patients who have been given a 3-weeks stationary treatment at a chronic pain clinic. Two main research questions are: which development is most influential for success or failure of treatment and is it necessary to include information on the patient's site of main pain?

The response of primary interest is self-reported success of treatment, measured 3 months after discharge by a score summarizing several aspects of the illness. Among the context variables, capturing features that cannot be modified, were age, gender, lower or higher level of formal schooling, and the number of other previous illnesses. Only those are shown in Figure 1 that had sizeable dependences to the responses of main interest.

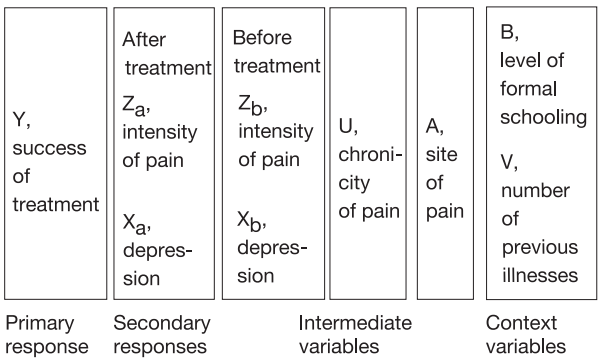

Figure 1 Initial ordering of variables in the chronic pain study for a sequence of regressions. Variables within a same box are treated on an equal standing. Variables in any one box are considered conditionally given all variables their past, listed within boxes to their right.

There are a number of intermediate variables. Before and after a 3-week stationary treatment, questionnaire scores are available of depression and of intensity of pain ranging from 'no pain' to 'pain as strong as imaginable.' The chronification score incorporates different aspects, such as time since onset of pain, spreading of pain, use of pain relievers, and the patient's pain treatment history. Main site of pain here has two categories: 'back pain' and pain on the 'head, face, or neck.'

The regression graph of Figure 2 summarizes some aspects of the statistical analyses. Throughout, symbol $\perp\!\!\!\perp$ means independence, symbol $\not\perp$ means dependence and symbol | is to be read as 'given.' The graph shows, in particular, which of the variables are needed so that for any given response, adding one more of the potentially explanatory variables does not improve prediction. Site of pain is an important intermediate variable since it is a node along a direction-preserving path of arrows pointing from level of schooling to treatment success, hence should be part of any future study of chronic pain.

Some of the directions and type of dependences, which cannot be read off the graph, are as follows. Patients with many years of formal schooling (13 years or more) are more likely to be head-pain patients, the others are more likely to be back-pain patients, possibly because more of them have jobs involving heavy physical work. Back-pain patients have higher scores of pain chronicity, reach higher stages of intensity of pain before treatment and report higher intensity of pain after treatment.

Never captured by the graph alone are nonlinear dependences, here of Y on Z_a. Treatment success, Y, is low whenever higher levels of intensity of pain remain after treatment, Z_a. But at relatively lower levels of Z_a, treatment is clearly the more successful when the intensity of pain at discharge is lower, Z_a. The model fits the data well since for each response taken separately, no indication was found that adding variables, further

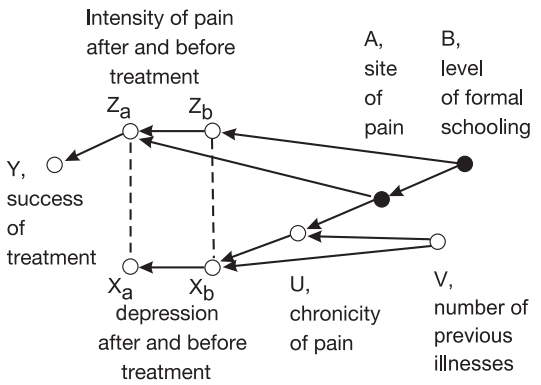

Figure 2 Well-fitting sequences of regressions which have a statistically significant relation for each edge present in the graph. Discrete variables are drawn as dots, continuous ones as circles. For instance, the two context variables are marginally independent, written as $B \perp\!\!\!\perp V$ and response Z_b is conditionally independent of V given A, B, U; $Z_b \perp\!\!\!\perp V \mid A, B, U$, also A is marginally dependent on B, written as $A \not\perp B$, while U depends on V given A; $U \not\perp V \mid A$. *Note*: Full arrows in regression graphs are sometimes also drawn as dashed-line arrows and dashed lines as so-called arcs, which are full lines with arrow-heads at both ends.

nonlinear or interactive effects would improve prediction and no strong dependences remained among the residuals of the joint responses on equal standing.

One important path of development is that patients with shorter formal schooling are more likely to get chronic back-pain and patients with chronic back-pain get help too late and respond less well to the type of treatment offered in the chronic pain clinic. This suggests as possible interventions to modify the type of treatment for the back-pain patients or, much more ambitiously, to raise the general level of formal schooling.

Data on Child Development

Here we use data of 347 families participating in the 'Mannheim study of children at risk.' We are grateful to Manfred Laucht, Central Institute of Mental Health Mannheim, for permitting a reanalysis of the data.

The study started with a random sample of more than 100 newborns from the general population of children born near Mannheim, Germany. This sample was completed to give roughly equal subsamples, in each of nine level combinations of the two types of adversity at birth, categorized to be at levels 'no, moderate, or high'; see Laucht et al. (1997). In other words, there was heavy oversampling of children at risk for motoric or cognitive deficits in later years and a random sample that serves as a control group, which provides, in particular, comparable norm values.

The recruitment of families stopped with 362 children. All measurements were reported in standardized form using the mean and standard deviation of the starting random sample. Of the 362 German-speaking families who entered the study when their first, single child was born without malformations or any other severe handicap, 347 families still participated when their child reached the age of 8 years.

There are joint responses of main and of secondary interest at age eight and at age four-and-a- half years. Each of these contains two components of possible deficits: cognitive or motoric; see the derived graph in Figure 3.

One main difference to previous analyses is that we averaged three different assessments of each of two types of risk: recorded at birth, at 3 months, and at 2 years. In both cases, this can be justified by the six observed pairwise correlations being all nearly equal; see for example, Wermuth (2013). These averaged scores are risks at 2 years for motoric deficits, labeled X_r, and for cognitive deficits, labeled Y_r. Two possible additional risks were identified at 3 months after birth, a score called unprotective environment E, and a binary variable H, which records whether the child had to be hospitalized during the first 3 months after birth.

The graph shows in particular for motoric deficits of the child at the age of 8 years, X_8, that no arrows are pointing to it from risks at 3 months, E, H, or from any variable for the cognitive side, Y_8, Y_4, Y_r, given information on the more recent motoric deficit, X_4 and the risk for motoric deficits, X_r. By contrast, cognitive deficits at the age of 8 years, Y_8 depend directly on unprotective environment at 3 months, on being hospitalized up to 3 months, as well as on motoric deficits at 2 years, given both more recent deficits, Y_4, X_4. A tentative interpretation is given later using Figure 7 in the next to last section.

The coding of variables implies that all dependences are positive so that deficits accumulate with each additional regressor. Some of the effects are accelerated compared to merely linear dependences. This is not reflected in the graph alone in Figure 4, but in Wilkinson's notation which adds to the graph sums of highest-order effects, as shown in Figure 3.

Questions Regarding Applications and Statistical Research

Here, we first list questions that arise in general when new statistical models are applied, and then give partial answers and references for sequences of regressions:

1. Are studies available in which the new models have been used fruitfully?
2. Can well-fitting models be derived from data and be tested as hypotheses formulated for future data?
3. What additional possibilities do the models offer to gain better insight into given research questions, especially to an understanding of development over time?

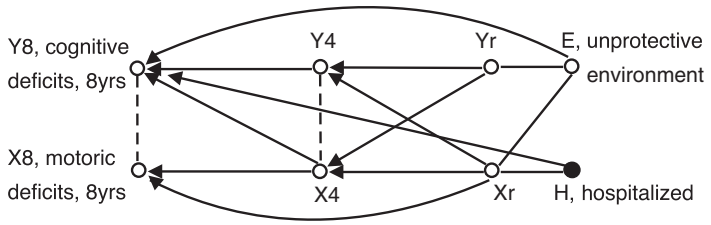

Figure 3 A well-fitting regression graph in the child development study; arrows point from regressors in the past to responses in the future; dashed lines are for dependent responses given their past; full lines are for dependent early risk factors given the remaining context variables.

$$
\begin{array}{ll}
Y_8 : Y_4 + X_4^2 + E + H & X_8 : X_4^2 + X_r \\
Y_4 : Y_r + X_r^2 & X_4 : Y_r + X_r^2 \\
Y_r : E^2 & X_r : E + H
\end{array}
$$

Note: In this notation supplementing the graph, every square term implies that also a main effect is included in a regression; see discussions of Wilkinson's notation in McCullagh and Nelder (1989).

As to (1), case studies with sequences of regressions, which include joint responses, continue to accumulate. In addition to the two examples summarized here in the previous section; see for instance Cox and Wermuth (1996), Cheung and Andersen (2003), Hardt et al. (2004), Smith (2009), Hardt et al. (2011), Marchetti et al. (2011), Schier et al. (in press), and Solis-Trapala et al. (in press).

As to (2), one most attractive feature of sequences of regressions is that their fitting requires typically no new estimation techniques. Standard methods are often available; for linear regressions, see Weisberg (2014), for categorical responses, see Cox (1972), McCullagh and Nelder (1989), Andersen and Skovgaard (2010). To screen for nonlinear or interactive effects, see Cox and Wermuth (1994). Regression graphs represent hypotheses about generating processes which can be tested in future studies.

When seemingly unrelated regressions are hypothesized, direct estimation methods are based for discrete variables on generalized linear models; see Marchetti and Lupparelli (2011), for joint Gaussian distributions, estimation is possible with structural equation models; see, for example, Bollen (1989). But as mentioned before, when strong residual dependences remain in Gaussian seemingly unrelated regressions, especially for small sample sizes, estimates may be far from the population values.

One may also regard seemingly unrelated regressions as so-called reduced models and embed them in larger so-called covering models for which estimation is again standard; see Cox and Wermuth (1990). Often m.l. estimates in the covering model are not far from those obtained by separate regressions and are then good approximations to the m.l. estimates in the reduced model.

As to (3), a number of new results have been obtained in the last years. We concentrate on the following few, discussed in the next three sections:

1. graphical conditions for the Markov equivalence of two regression graphs, to be used for possible alternative interpretations or for different types of fitting algorithm;
2. conditions and criteria for deriving implied independences and implied dependences from a given regression graph;
3. ways of distinguishing different types of confounding that are implied by a generating process when some variables are unobserved, that is marginalized over, or a population subgroup is studied; that is, when there is conditioning on the levels of some variables present in the generating process.

Markov Equivalence of Regression Graphs

Two different graphs are Markov equivalent if they define the same independence structure. The independence structure of a regression graph, with a given set of nodes and an ordering for the responses, is determined by its list of missing edges for node pairs i,k. Such a graph has possibly two types of undirected edges; full lines between context nodes and dashed lines between responses on equal standing, as well as arrows starting at a regressor and pointing to a response; see **Figure 1** to recall the meaning of these terms.

A missing full i,k-line between two context nodes means conditional independence given all remaining context nodes, and a missing dashed i,k-line between two response nodes on equal standing means conditional independence given other nodes in the past of i and k. If a possible i,k-arrow pointing from k to i is missing, this means conditional independence given all other nodes in the past of i except for k.

A V-configuration in such a graph, often called just a V, consists of three nodes with two uncoupled nodes, called the outer nodes, i,k. Both outer nodes are coupled to an inner node. There are two types of Vs in regression graphs, named collision and transmitting Vs. In each of the three possible types of collision Vs:

$$i\text{---}\circ\prec k, \qquad i\succ\circ\prec k, \qquad i\text{---}\circ\text{---}k\,,$$

the two edges are removed when the inner node is marginalized over. In each of the remaining transmitting Vs, both edges are removed by conditioning on the inner node.

Directed acyclic graphs consist exclusively of arrows so that there can be only the middle V above, where two uncoupled regressors point to a common response. Graphs of only full lines, often called concentration graphs, have only transmitting Vs, while graphs of only dashed lines, often called covariance graphs, have only collision Vs.

These three types of graph are subclasses of regression graphs. The names of the last two derive from parameters in joint Gaussian distributions. For such a distribution, zero off-diagonal elements, σ_{ik}, in the covariance matrix capture a marginal independence, and zero off-diagonal elements, σ^{ik}, in the concentration matrix, the inverse of the covariance matrix, capture a conditional independence given all the remaining nodes.

Two different types of regression graphs, which have the same node set and the same set of missing edges but different types of edge for some node pairs, are Markov equivalent if and only if their sets of collision Vs coincide, even though for any given V, there may be any of the three types of collision Vs above; for a proof, see Wermuth and Sadeghi (2012). One example is shown in **Figure 4**.

The graph on the left of **Figure 4**, resulted from a retrospective study, with more variables than reported here. Questions about their childhood were answered by 283 adult females when visiting a general practitioner for some minor health problems; see Hardt et al. (2008). No nonlinear or interactive effects were detected.

The well-fitting graph contains two binary variables, level of formal schooling and severe sexual abuse during childhood, as well as three quantitative measurements. Age in years is recorded directly, the other variables being derived from questionnaires. Family status indicates the recalled social standing of the family during early childhood. Family distress includes psychological disturbances and alcohol or drug problems of the parents.

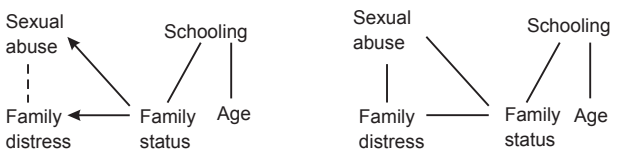

Figure 4 A regression graph that is Markov equivalent to a concentration graph.

From the Markov equivalence to the concentration graph on the right of **Figure 4**, one knows, for instance, directly that sexual abuse is independent of the level of formal schooling given knowledge about the family status. This implication may be derived from the defining independences of the graph on the left or be based on the following separation result in graph theory. If in an undirected graph with three disjoint subsets α, β, c of the given set of nodes, every path between α and β has a node in c, then α is separated from β by c, that is removal of set c leaves α and β disconnected. For concentration graphs, this separation implies $\alpha \perp\!\!\!\perp \beta | c$; see, for example, Lauritzen et al. (1990).

Figure 5 shows instead a regression graph on the left that is Markov equivalent to the covariance graph on the right. It is the simplest case of the seemingly unrelated regressions with regressors labeled 3 and 4 and responses labeled 1 and 2. For instance, it follows from the graph definitions that edge $1 - - - 2$ means $1 \not\perp\!\!\!\perp 2 | 3, 4$ on the left and $1 \not\perp\!\!\!\perp 2$ on the right.

As discussed in the next section, to transfer such equivalence results for graphs to a distribution generated over a regression graph, this distribution has to mimic some properties of joint Gaussian distributions; see also Wermuth and Sadeghi (2012).

Implied Independences in Directed Acyclic Graphs

Work on reading all implied independences off a directed acyclic graph started with a path-breaking, complex theorem, called d-separation; see Pearl (1988) and Geiger et al. (1990). For the reduction of d-separation to the above described separation criterion for undirected graphs; see Lauritzen et al. (1990).

A third criterion is based on matrix representations of graphs, named their edge matrices. Square edge matrices contain zeros for missing edges, ones for edges present, and ones along the diagonal to extend the graph theoretic notion of adjacency matrices in such a way that some sums or products of edge matrices describe transformations of graphs; see Wermuth et al. (2006). It gives the most direct criterion in that it leads to a matrix with the dimensions of two disjoint node subsets α and β: if this is a matrix of zeros only, then it is established that the graph implies a desired conditional independence.

In particular, let four disjoint subsets α, β, c, m partition the node set of a directed acyclic graph, let further $a = \alpha \cup m$, and $b = \beta \cup c$, then the transformation to the edge matrix $\mathcal{P}_{a|b}$ parallels the transformation of parameters in linear sequences of single-response regressions to the matrix $\Pi_{a|b}$ of least-squares regression coefficients with Y_a as response and Y_b as regressor. For joint Gaussian distributions, the submatrix of $\Pi_{a|b}$ for rows α and columns β is zero, $\Pi_{\alpha|\beta.c} = 0$, if and only if $\alpha \perp\!\!\!\perp \beta | c$. In general, $\mathcal{P}_{\alpha|\beta.c}$, the corresponding submatrix of $\mathcal{P}_{a|b}$ is zero if the given graph implies $\alpha \perp\!\!\!\perp \beta | c$.

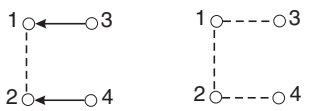

Figure 5 A regression graph that is Markov equivalent to a covariance graph.

For directed acyclic graphs, the three types of criteria have been proven to be equivalent; see Marchetti and Wermuth (2009). For combining independences in sequences of regressions, the generated distribution has to satisfy additional conditions that mimic some properties of Gaussian distributions.

Independence properties of joint Gaussian distributions have been nicely summarized in the information theory literature, together with other properties that hold for all probability distributions; see Lněnička and Matúš (2007: Definition 1). To combine pairwise conditional independences in general sequences regression just as in the Gaussian case, two properties are needed; see Sadeghi and Lauritzen (2014). We call the presence of the intersection property the 'downward combination' and the presence of the composition property the 'upward combination' of pairwise independences.

A graphical illustration of these two properties is given for just three nodes i, h, k in **Figure 6**. In this simplest case, upward and downward combination mean, respectively,

$$(i \perp\!\!\!\perp h \text{ and } i \perp\!\!\!\perp k) \Rightarrow (i \perp\!\!\!\perp h, k) \Rightarrow (i \perp\!\!\!\perp h | k \text{ and } i \perp\!\!\!\perp k | h)$$

$$(i \perp\!\!\!\perp h | k \text{ and } i \perp\!\!\!\perp k | h) \Rightarrow (i \perp\!\!\!\perp h, k) \Rightarrow (i \perp\!\!\!\perp h \text{ and } i \perp\!\!\!\perp k)$$

Essential for both properties in the above two equations is the first implication. The second implication holds in all probability distributions and just serves to motivate the chosen names. For instance, upward combination means above to move from two simple to two more complex conditional independence statements; expressed more generally: it means moving to increased conditioning sets.

On the left of **Figure 6** is a complete regression graph with joint response $\{h, k\}$, on the right a complete directed acyclic graph with single responses ordered as. (i, h, k)

A sufficient condition for the intersection property is to have a strictly positive distribution; see San Martin, Mouchard & Rolin (2005) for extensions. Sufficient for the composition property is that to every nonlinear effect there is also a linear one and that to every interactive effect there are also main effects. Such linear or main effects have to be substantial to be relevant for a model fit; see Németh and Rudas (2013) for an example when this property does not hold in a sociological context. In such instances, models more complex than graphical models are needed such as those included for discrete variables in the class of marginal models; see Bergsma and Rudas (2002).

Deciding on Implied Dependences

Edge matrix results for directed acyclic graphs have been extended to the more general types of regression graph and, more importantly, to inferring that a generating regression

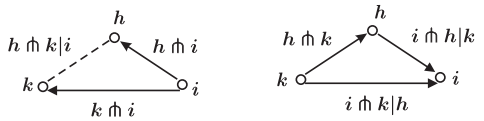

Figure 6 Two regression graphs where removal of the *hi*-edge and the *ki*-edge leads to $i \perp\!\!\!\perp h, k$, provided independences combine upward (left) and downward (right).

graph implies the conditional dependence $\alpha \not\perp\!\!\!\perp \beta | c$ if $\mathcal{P}_{\alpha|\beta.c} \neq 0$; see Wermuth (2012).

To permit conclusions about dependences implied by sequences of regressions, the generated distribution has to have, in addition to upward and downward combination of independences, nonvanishing parameters to each edge present in the graph and it has to be dependence-inducing, a property also called singleton transitivity. Distributions with all above mentioned properties have been called traceable, since their regression graph can be used to study pathways of development.

In any specific application, there may be special parametric constellations that lead to path cancellations and hence to preventing us from recognizing dependences that are implied by the given generating process for other parametric constellations. This cannot happen, when all variables are coded, as in the Mannheim study, to have higher values for stronger adversities, for higher risks and for increased deficits and only linear dependences are observed or dependences that are accelerated compared to the linear effects. In addition, prediction of the strength or direction of induced associations may vary with the measures of dependence used; see Jiang et al. (in press).

For linear dependences among three variables, there are already three types of possible parameters: covariances, concentrations, and regression coefficients. A detailed notation is needed to understand their relations:

$$\sigma_{12|3} = \sigma_{12} - \sigma_{13}\sigma_{23}/\sigma_{33} \qquad [1]$$

$$\sigma^{23.1} = \sigma^{23} - \sigma^{12}\sigma^{13} \qquad [2]$$

$$\beta_{1|3} = \beta_{1|3.2} + \beta_{1|2.3}\beta_{2|3} \qquad [3]$$

These three nice recursive properties were derived by different authors. Here, $\sigma_{12|3}$ denotes the conditional covariance of 1,2 given 3; see Anderson (1958), $\sigma^{23.1}$ the concentration of 2,3 marginalized over 1; see Dempster (1969), and $\beta_{1|3.2}$ the population coefficient of 2 when regressing 1 on both 2 and 3; see Cochran (1938) for the property.

The right-hand sides of the above equations contain, for trivariate Gaussian distributions, the parameters representing dependence in different graphs for three nodes. These parameters are in eqn [1] for a covariance graph, in eqn [2] for a concentration graph and in eqn [3] for a directed acyclic graph: with 1 as response to 2, 3 and 2 as a response to 3; see **Figure 8**.

The coefficient of 3 when regressing 1 only on 3, denoted by $\beta_{1|3}$ can be regarded as the result of two paths: of $\beta_{1|3.2}$, an arrow pointing from 3 to 1, and of a sequence of two arrows starting at 3 and pointing to 1 via 2 with effect $\beta_{1|2.3}\beta_{2|3}$. Sometimes, the first is called the direct effect of 3 on 1, the second the indirect effect, and $\beta_{1|3}$ the overall effect.

If for instance $0 = \sigma_{12} = \sigma_{13}$ in eqn [1], then $\sigma_{12|3} = 0$ and consequently also $\beta_{1|3.2} = 0$ in eqn [3] so that 1 has no dependence on 2 and 3 jointly. This is the Gaussian case of the more general notion of combining independences upward since here $(1 \perp\!\!\!\perp 2 \text{ and } 1 \perp\!\!\!\perp 3) \Rightarrow 1 \perp\!\!\!\perp 2,3$.

With more variables, eqns [1]–[3] generalize by adding everywhere a larger conditioning set in eqns [1] and [3], to get, for instance, $\sigma_{12|c}$, $\sigma_{12|3c}$ and $\beta_{1|3.c}$, $\beta_{1|3.2c}$, and by adding a larger marginalizing set in eqn [2] to get $\sigma^{23.m}$, $\sigma^{23.1m}$. The three measures of dependence relate for a node set consisting of 1, 2, 3, c, m as.

$$\beta_{1|3.c} = \sigma_{13|c}/\sigma_{33|c} = -\sigma^{13.m}/\sigma^{11.m}$$

derived via the sweep operator by Dempster (1969), and explaining why $0 = \sigma_{12|3} = \beta_{1|2.3}$ and $0 = \sigma^{23.1} = \beta_{2|3}$ above when all variances such as $\sigma_{33|c}$ and all precisions, such as $\sigma^{11.m}$, are nonzero in the two single-response regressions.

For joint Gaussian distributions, the dependence-inducing property can be proven in its simplest form using eqn [1]: for $1 \not\perp\!\!\!\perp 3$ and $2 \not\perp\!\!\!\perp 3$, one may have at most $\sigma_{12|3} = 0$ or $\sigma_{1|2} = 0$ but never both. In the case of $\sigma_{12|3} = 0$, the induced correlation, $\rho_{12} = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$, has also been named a 'spurious correlation,' since the dependence between the two variables 1 and 2 can be 'explained away' by conditioning on their common neighbor 3. For three variables, the dependence-inducing property can be written as:

$$(1 \not\perp\!\!\!\perp 3 \text{ and } 2 \not\perp\!\!\!\perp 3) \Rightarrow \text{at most } 1 \perp\!\!\!\perp 2|3 \text{ or } 1 \perp\!\!\!\perp 2 \text{ but never both}$$

It has been proven that binary variables are dependence-inducing; see Simpson (1951). But there are constellations of counts for which both independences $1 \perp\!\!\!\perp 2|3$ and $1 \perp\!\!\!\perp 2$ seem to hold and a decision would have to be based on outside information.

Independence-Predicting and Independence-Preserving Graphs

The results for deciding on implied independences and dependences, based on an induced edge matrix, have been extended to derive regression graphs implied by a given generating process when, for instance, the order of the responses or the conditioning sets, or the marginalizing sets are changed; see Wermuth and Cox (2004).

Then the so-called independence-predicting graphs, may be derived by using the partial closure operator (Wermuth et al., 2006) as well as sums and products of edge matrices, but each of these operations has also an intuitive translation into closing special types of paths in graphs; see Wermuth (2012).

For instance, when in a study, different from the Mannheim one, the same variables were available except for cognitive deficits after 4 years, Y_8,Y_4, one predicts, from the starting graph in **Figure 3** and an unchanged order of the remaining variables, the regression graph in **Figure 7**. In this case, it is just a subgraph of **Figure 3** obtained by removing nodes

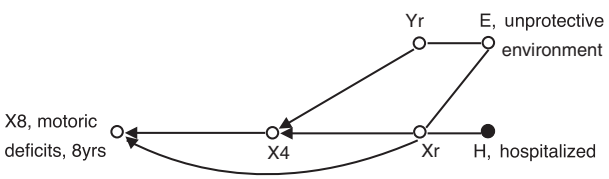

Figure 7 The regression graph induced by Figure 3 after marginalizing over Y_8 and Y_4.

and edges of Y_8 and Y_4, no additional dependences are induced.

By contrast, when X_8 and X_4 are not available, all remaining variables turn out to be directly explanatory for Y_8. One consequence of this finding is that it is more important for psychologists and psychiatrists to evaluate also motor development than it is for physicians to take also cognitive development into account.

Typically, one cannot use such an induced regression graph to condition on or marginalize over more nodes and still see what the starting graph would have implied for this case, while this is possible with the so-called independence-preserving graphs. Starting from directed acyclic graphs, three types of such graphs have been proposed, the MC-graphs by Koster (2002), the maximal ancestral graphs by Richardson and Spirtes (2002), and the summary graphs by Wermuth (2011). For their relations and proofs of Markov equivalence, see Sadeghi (2013). To construct these types of graph within the program environment R, see Sadeghi and Marchetti (2012).

Summary graphs may be derived for the more general regression graphs and, if obtained after marginalizing only, they provide graphical criteria to detect confounding.

Graphical Representations of Distortions

For traceable regressions, the regression graph represents an independence structure and a partly specified dependence structure. The graph can be viewed as a hypothesis of how sequences of single or joint responses generate a joint distribution when, for instance, corresponding point estimates of parameters are taken as the population values. For reliable interpretations of estimated effects, it is important to prevent, if at all possible, any important, sizable distortion of the actual population parameters.

Different types of distortions of a treatment effect may be avoided by using regression graphs and appropriate study designs. To see this, we show in **Figure 8** simple cases of under-conditioning, of over-conditioning, and of direct confounding by using a node crossed by two lines, $\not\emptyset$, to indicate that the variable is to be marginalized over and a node surrounded by a square, $\boxed{\bigcirc}$, to indicate that the variable is to be conditioned on.

For instance, by using label 3 for treatment and label 1 for the outcome or response to 2 and 3, the graph on the left of **Figure 8** shows one important intermediate variable, labeled 2, which could be, say, compliance of the patients to treatment.

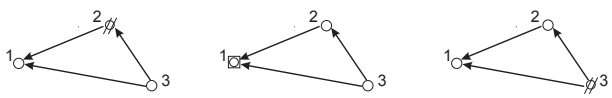

Figure 8 Left: Under-conditioning for $1 \perp\!\!\!\perp 3|2$ by ignoring an intermediate variable. Middle: Over-conditioning for $2 \perp\!\!\!\perp 3$ by selecting levels of the common response 1. Right: Direct confounding for $1 \perp\!\!\!\perp 2|3$ by ignoring the common explanatory variable 3 for 1 and 2; generating a double edge $1 \circ\!\!\!

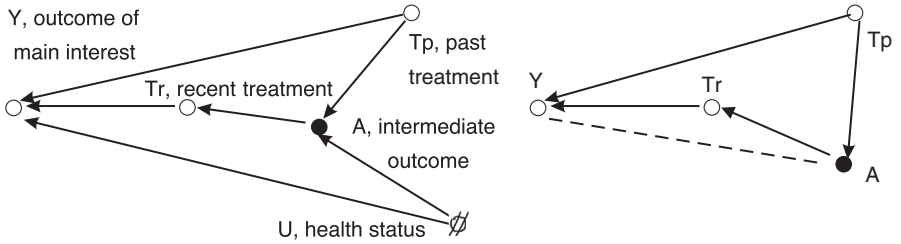

Figure 9 Left: full randomization for T_p leads to removal of $T_p \leftarrow U$, randomized allocation of individuals to selected levels of a more recent treatment, T_r, given results of the intermediate outcome A, leads to removal of $T_r \leftarrow U$ and of $T_r \leftarrow T_p$ and to the presence of $T_r \leftarrow A$; right: the summary graph obtained by marginalizing over U; path $Y \dashrightarrow A \prec\!\!\longleftarrow T_p$ explaining the indirect confounding of $Y \leftarrow T_p$ when response Y is regressed on only T_p and T_r.

See also: Binary Response Models and Logistic Regression; Causal Inference: Overview; Compliance; Longitudinal Data; Multivariate Analysis: Discrete Variables (Loglinear Models); Observational Studies: Overview; Simultaneous Equation Estimation: Overview.

Bibliography

Andersen, P.K., Skovgaard, L.T., 2010. Regression with Linear Predictors. Springer, New York.

Anderson, T.W., 1958. An Introduction to Multivariate Statistical Analysis, third ed., 2003. Wiley, New York.

Bergsma, W., Rudas, T., 2002. Marginal models for categorical data. Annals of Statistics 30, 140–159.

Bollen, K.A., 1989. Structural Equations with Latent Variables. Wiley, New York.

Cheung, S.Y., Andersen, R., 2003. Time to read: family resources and educational outcomes in Britain. Journal of Comparative Family Studies 34, 413–433.

Cochran, W.G., 1938. The omission or addition of an independent variate in multiple linear regression. Supplement of the Journal of the Royal Statistical Society 5, 171–176.

Cox, D.R., 1972. The analysis of multivariate binary data. Journal of the Royal Statistical Society, Series C 21, 113–120.

Cox, D.R., 2006. Principles of Statistical Inference. Cambridge University Press, Cambridge.

Cox, D.R., Wermuth, N., 1990. An approximation to maximum-likelihood estimates in reduced models. Biometrika 77, 747–761.

Cox, D.R., Wermuth, N., 1993. Linear dependencies represented by chain graphs (with discussion). Statistical Science 8, 204–218, 247–277.

Cox, D.R., Wermuth, N., 1994. Tests of linearity, multivariate normality and adequacy of linear scores. Journal of the Royal Statistical Society, Series C 43, 347–355.

Cox, D.R., Wermuth, N., 1996. Multivariate Dependencies: Models, Analysis, and Interpretation. Chapman and Hall (CRC), London.

Cowell, R.G., Dawid, A.P., Lauritzen, S.L., Spiegelhalter, D.J., 1999. Probabilistic Networks and Expert Systems. Springer, New York.

Darroch, J.N., Lauritzen, S.L., Speed, T.P., 1980. Markov fields and log-linear models for contingency tables. Annals of Statistics 8, 522–539.

Dempster, A.P., 1969. Elements of Continuous Multivariate Analysis. Addison-Wesley, Reading, MA.

Drton, M., 2009. Discrete chain graph models. Bernoulli 15, 736–753.

Drton, M., Richardsom, T.S., 2004. Multimodality of the likelihood in the bivariate seemingly unrelated regressions model. Biometrika 91, 383–392.

Edwards, D., 1995. Introduction to Graphical Modelling, second ed., 2000. Springer, New York.

Edwards, D., Lauritzen, S.L., 2001. The TM algorithm for maximising a conditional likelihood function. Biometrika 88, 961–972.

Geiger, D., Verma, T.S., Pearl, J., 1990. Identifying independence in Bayesian networks. Networks 20, 507–534.

Gibbs, J.W., 1902. Elementary Principles in Statistical Mechanics (Reprinted by Dover Publications, New York, 1960).

Haavelmo, T., 1943. The statistical implications of a system of simultaneous equations. Econometrica 11, 1–12.

Hardt, J., Petrak, F., Filipas, D., Egle, U.T., 2004. Adaption to life after surgical removal of the bladder – an application of graphical Markov models for analysing longitudinal data. Statistics in Medicine 23, 649–666.

Hardt, J., Sidor, A., Nickel, R., Kappis, B., Petrak, F., Egle, U.T., 2008. Childhood adversities and suicide attempts: a retrospective study. Journal of Family Violence 23, 713–718.

Hardt, J., Herke, M., Schier, K., 2011. Suicidal ideation, parent–child relationships, and adverse childhood experiences: a cross-validation study using a Graphical Markov Model. Child Psychiatry & Human Development 42, 119–133.

Jiang, Z., Ding, P., Geng, Z. Qualitative evaluation of associations by the transitivity of the association signs. Statistica Sinica, in press. http://dx.doi.org/10.5705/ss.2013.095.

Koster, J., 2002. Marginalising and conditioning in graphical models. Bernoulli 8, 817–840.

Laucht, M., Esser, G., Schmidt, M.H., 1997. Developmental outcome of infants born with biological and psychosocial risks. Journal of Child Psychology and Psychiatry 38, 843–853.

Lauritzen, S.L., 1996. Graphical Models. Oxford University Press, Oxford.

Lauritzen, S.L., Dawid, A.P., Larsen, B., Leimer, H.G., 1990. Independence properties of directed Markov fields. Networks 20, 491–505.

Lauritzen, S.L., Wermuth, N., 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. Annals of Statistics 17, 31–57.

Loh, P.L., Wainwright, M.J., 2013. Structure estimation for discrete graphical models: generalized covariance matrices and their inverses. Annals of Statistics 41, 3022–3049.

Lněnička, R., Matúš, F., 2007. On Gaussian conditional independence structures. Kybernetika 43, 323–342.

Marchetti, G.M., Lupparelli, M., 2011. Chain graph models of multivariate regression type for categorical data. Bernoulli 17, 827–844.

Marchetti, G.M., Vannini, I., Gottard, A., Vignoli, D., 2011. Regression graph models: an application to joint modelling of fertility intentions among childless couples. In: Conesa, D., Forte, A., Lopez-Quilez, A., Munoz, F. (Eds.), Proceedings of the 26th International Workshop on Statistical Modelling, pp. 358–363.

Marchetti, G.M., Wermuth, N., 2009. Matrix representations and independencies in directed acyclic graphs. Annals of Statistics 47, 961–978.

Markov, A.A., 1912. Wahrscheinlichkeitsrechnung (German translation of 2nd Russian edition: Markoff, A.A., 1908). Teubner, Leipzig.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman & Hall (now CRC), London.

Németh, R., Rudas, T., 2013. On the application of discrete marginal graphical models. Sociological Methodology 43, 70–100.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo.

Pearl, J., 2009. Causality: Models, Reasoning and Inference. Cambridge University Press, Cambridge.

Pearl, J. Trygve Haavelmo and the emergence of causal calculus. Econometric Theory, in press. http://dx.doi.org/10.1017/S0266466614000231.

Richardson, T.S., Robins, J.M., 2013. Single World Intervention graphs: A Primer. Sec. UAI Workshop Causal Struct. Learn., Bellevue, Washington.

Richardson, T.S., Spirtes, P., 2002. Ancestral Markov graphical models. Annals of Statistics 30, 962–1030.

Robins, J., Wasserman, L., 1997. Estimation of effects of sequential treatments by reparametrizing directed acyclic graphs. In: Geiger, D., Shenoy, O. (Eds.), Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence. Morgan and Kaufmann, San Francisco, pp. 409–420.

Sadeghi, K., 2013. Stable mixed graphs. Bernoulli 19, 2330–2358.
Sadeghi, K., Lauritzen, S.L., 2014. Markov properties for mixed graphs. Bernoulli 20, 676–696.
Sadeghi, K., Marchetti, G.M., 2012. Graphical Markov models with mixed graphs in R. The R Journal 4, 65–73.
San Martin, E., Mochart, M., Rolin, J.M., 2005. Ignorable common information, null sets and Basu's first theorem. Sankhya 67, 674–698.
Schier, K., Herke, M., Nickel, R., Egle, U.T., Hardt, J. Long-term sequelae of emotional parentification: a cross-validation study using sequences of regressions. Journal of Child and Family Studies, in press. http://dx.doi.org/10.1007/s10826-014-9938-z.
Simpson, E.H., 1951. The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society. Series B 13, 238–241.
Smith, R.B., 2009. Issues matter: a case study of factors influencing voting choices. Case Studies in Business, Industry and Government Statistics 2, 127–146.
Solis-Trapala, I., Ward K., Webb, R.E.B., Shoenmakers, I., Prentice, A., Goldberg G.R., The Role of Non-Mechanical Physiological Factors in Skeletal Phenotype using Sequences of Regressions, submitted.
Spirtes, P., Glymour, C., Scheines, R., 1993. Causation, Prediction and Search. Springer, New York.
Studený, M., 2005. Probabilistic Conditional Independence Structures. Springer, London.
Vandenbroucke, J.,P., 2002. The history of confounding. Social and Preventive Medicine 47, 216–224.
Wainwright, M.J., Jordan, M.I., 2008. Graphical models, exponential families, and variational inference. Foundations and Trends Machine Learning 1, 1–305.
Weisberg, S., 2014. Applied Linear Regression, fourth ed. Wiley, New York.
Wermuth, N., 1976. Analogies between multiplicative models for contingency tables and covariance selection. Biometrics 32, 95–108.
Wermuth, N., 1980. Linear recursive equations, covariance selection, and path analysis. Journal of the American Statistical Association 75, 963–997.
Wermuth, N., 1992. On block-recursive regression equations (with discussion). Brazilian Journal of Probability and Statistics 6, 1–56.
Wermuth, N., 2011. Probability models with summary graph structure. Bernoulli 17, 845–879.
Wermuth, N., 2012. Traceable regressions. International Statistical Review 80, 415–438.
Wermuth, N., 2013. Comment: are marginal models needed? Sociological Methodology 43, 114–119.
Wermuth, N., Cox, D.R., 1998. Statistical dependence and independence. In: Armitage, P., Colton, T. (Eds.), Encyclopedia of Biostatistics. Wiley, New York, pp. 4260–4264.
Wermuth, N., Cox, D.R., 2004. Joint response graphs and separation induced by triangular systems. Journal of the Royal Statistical Society, Series B 66, 687–717.
Wermuth, N., Cox, D.R., 2008. Distortions of effects caused by indirect confounding. Biometrika 95, 17–33.
Wermuth, N., Cox, D.R., 2013. Concepts and a case study for a flexible class of graphical Markov models. In: Becker, C., Fried, R., Kuhnt, S. (Eds.), Robustness and Complex Data Structures. Festschrift in Honour of Ursula Gather. Springer, Heidelberg, pp. 331–350. Also on ArXiv: 1303.1436.
Wermuth, N., Cox, D.R., Marchetti, G.M., 2009. Triangular systems for symmetric binary variables. Electronic Journal of Statistics 3, 932–955.
Wermuth, N., Lauritzen, S.L., 1983. Graphical and recursive models for contingency tables. Biometrika 70, 537–552.
Wermuth, N., Marchetti, G.M., 2014. Star graphs induce tetrad correlations: for Gaussian as well as for binary variables. Electronic Journal of Statistics 8, 253–273.
Wermuth, N., Marchetti, G.M., Zwiernik, P., 2014. Binary distributions of concentric rings. Journal of Multivariate Analysis 130, 252–260. Also on ArXiv: 1311.5655.
Wermuth, N., Sadeghi, K., 2012. Sequences of regressions and their independences (invited paper with discussion). TEST 21, 215–279.
Wermuth, N., Wiedenbeck, M., Cox, D.R., 2006. Partial inversion for linear systems and partial closure of independence graphs. BIT Numerical Mathematics 46, 883–901.
Whittaker, J., 1990. Graphical Models in Applied Multivariate Statistics. Wiley, Chichester.
Wright, S., 1923. The theory of path coefficients: a reply to Niles' criticism. Genetics 8, 239–255.
Wright, S., 1934. The method of path coefficients. Annals of Statistics 5, 161–215.
Zellner, A., 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. Journal of the American Statistical Association 57, 348–368.

Graphical Methods, Analytic

Antony Unwin, University of Augsburg, Augsburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Graphics are a powerful tool for representing and analyzing data. Interactive graphics are ideal for exploratory data analyses, for checking data quality, for investigating assumptions, and for uncovering information. Standard statistical displays (e.g., histograms and scatterplots) are empowered by the addition of interactive features such as direct querying, reformatting, and linking. Displays of complex multivariate data become workable in an interactive implementation. Mosaic plots for categorical data and parallel coordinate plots for continuous data are two recent examples.

Introduction

Everyone is accustomed to seeing graphical displays used for presentation purposes: in scientific articles, newspapers, and magazines. Yet graphical displays may be used for exploratory purposes too, for investigating data sets and for uncovering the information hidden in them. Exploratory graphics are different. The aim is not to present information, but to use the graphics to uncover information. Static printed graphics can be useful for exploratory work, but the real gain lies in using interactive graphics on a computer.

In all the following examples it is assumed that exploratory analyses are being carried out, that each graphic is part of a broader analysis and that they are drawn to aid the analyst, not to present the information to others. This is why there are few scales given. The analyst will know the approximate scaling and can always query the graphic on screen to get exact values, not only along axes, but for any point or object. In general, a single presentation graphic is drawn to convey information to a potentially large and varied audience, whereas a large and varied collection of exploratory graphics are drawn to extract information for a single person, the analyst.

Graphics are useful for checking basic assumptions and for getting a feel for data. Figure 1 shows a bar chart of gender from an insurance data set. Unusually there are 7 codes for gender, three made sense (firms are neither male nor female) and the rest were errors.

The histogram for policyholders' birthdates for the same data set was also revealing (Figure 2). The tall white column to the left represents missing values for the variable (mainly firms again). There is a small isolated column at 1900 (missings originally coded as '00'?) and there are some very recent dates of birth, surely errors for car-driver insurance data. This is a large data set (over 50 000 cases) and some errors may be expected, but it is extraordinary how often data sets that have been purportedly quality checked still have errors in them. Incidentally, the dip in the histogram is due to low birthrates during the war.

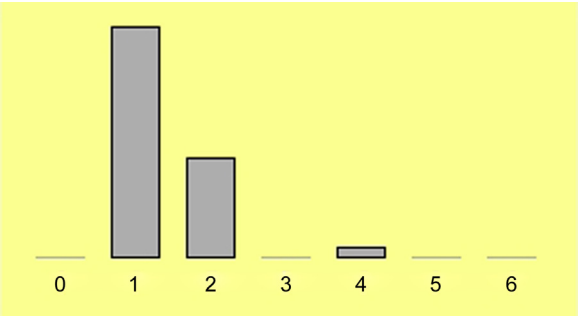

Figure 1 Bar chart of gender of car insurance policyholders.

Assuming major errors have been corrected or removed, graphics are valuable both for checking old ideas and for generating new ones. After the historic 1990 Irish Presidential election it was hypothesized that Mary Robinson had won because she got the women's vote. Data from the last MRBI survey before the election suggests otherwise. Figure 3 shows a bar chart for the preferences for the three candidates linked to a spine plot for gender (in a spine plot all column heights are the same and column widths represent the counts, so highlighted proportions are directly comparable). The proportion of women intending to vote for Mary Robinson was only slightly higher than the proportion of men. Linking to another spine plot shows that urban voters were more important for Mary Robinson's win.

Background

Presentation graphics (*see* Graphical Methods: Presentation) have been used for a long time. Exploratory graphics are newer, not because people did not appreciate their value, but because without a computer it was not efficient to draw many variations of many different displays. John Tukey led the way with his innovative and provocative book, *Exploratory Data Analysis* (Tukey, 1977) (*see* Exploratory Data Analysis: Univariate Methods) although, curiously enough, the book was written slightly too early, before personal computers were becoming commonplace. Thus he emphasizes pencil and paper methods like stem and leaf plots for continuous data. These are like histograms laid on their side with each bin having a width of a power of 10. The advantages are the simple construction by hand and the listing of each value individually. The disadvantages are the dependence on a very limited range of bin sizes and the unsuitability of the method for anything other than small data sets. Imagine a stem and leaf plot for the car insurance data set of Figure 2.

More recent computer-oriented books for EDA include Cook and Swayne (2007) and Theus and Urbanek (2007).

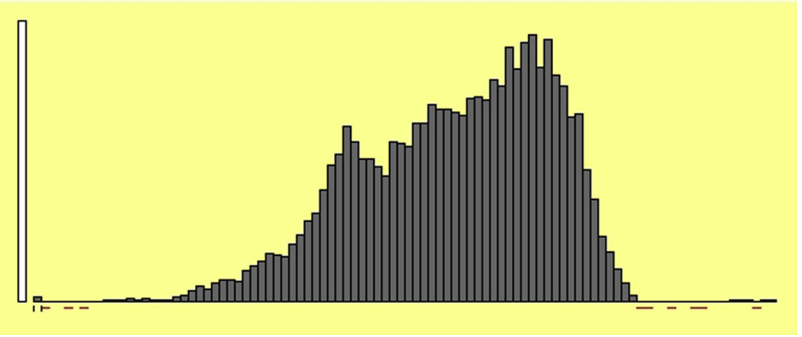

Figure 2 Histogram of years of birth of car insurance policyholders (from 1900 to 1995).

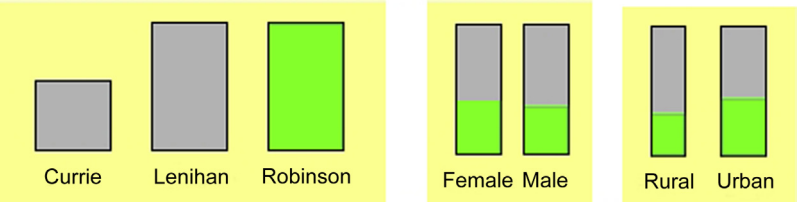

Figure 3 Bar chart of candidates' support linked to spine plots for gender and location. Results from the last MRBI survey before the 1990 Irish Presidential Election. Thanks to Jack Jones and the Market Research Bureau of Ireland for permission to use the data.

There has also been much related work outside statistics for which Spence (2007) and Ward et al. (2010) are good sources. In none of these works is a substantial formalization of exploratory graphics attempted. Considering that presentation graphics have been around much longer and yet Wilkinson's groundbreaking work (Wilkinson, 2005) first appeared in 1999, this is not surprising. Furthermore, exploratory analyses are by their nature interactive and dynamic and so the process is difficult to explain on the printed page. Some of the best sources of information and examples can be found on the web using the keywords of this article and/or author names as search parameters.

Interacting with Displays

Interactive Features

Graphics may mislead. For instance, bar charts will look similar whether they are based on 50 or on 5000 cases. (The survey in Figures 3 and 4 was based on 1000 interviews.) It is important to be able to check any impressions gained.

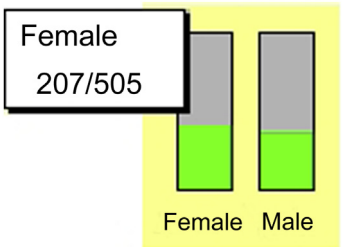

Figure 4 Direct querying of a bar. How many women supported Mary Robinson?

Sometimes this might mean carrying out the corresponding statistical test (e.g., a chi-square test for comparing bar chart proportions), but a first step is to directly query the graphic, as in Figure 4.

Direct querying is useful in all displays, especially when it is context sensitive. The same action may be used for each part of each plot and yet provide different information according to which part of which plot is queried. Other steps that may be carried out to gain more insight are to rescale the graphic (e.g., change the bin size in a histogram), to reformat the display (e.g., switch the axes of a scatterplot), to reorder the display (e.g., order the categories in a bar chart by count), to zoom in (e.g., concentrate on an outlying cluster of points in a scatterplot), or to select various groups of points and link them to their representation in other displays (Unwin, 2012).

Histogram bin sizes, which are small, pick out gaps and identify preferred values, while large bin sizes emphasize overall trends. Figure 5 shows two histograms of math marks for 126 Irish schoolchildren. Notice how one version picks out the typical gap below the pass mark of 40 and the other does not.

Scatterplots can be extremely informative tools for revealing the relationship between two variables, but the scales, the aspect ratio of the axes, the size and color of the points can all affect how they look. For large data sets with extensive overplotting (i.e., many points are drawn at the same location on the plot), some form of density estimation (*see* Probability Density Estimation) is necessary or there is a risk the data set may be misunderstood. Figure 6 shows a scatterplot of the birth weights of 4000 babies plotted against the estimated duration of pregnancy. A coarse density estimation has been used whereby black dots represent single cases and the brightest white ones at least 30 cases. It is clear where the majority of the data is concentrated.

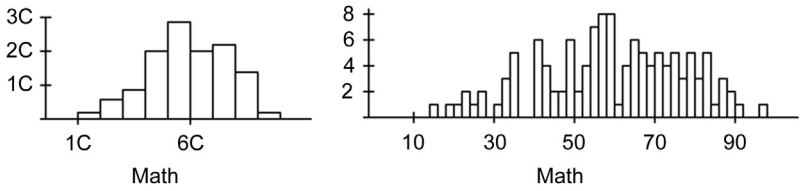

Figure 5 Histograms of math marks out of 100 for a group of Irish schoolchildren.

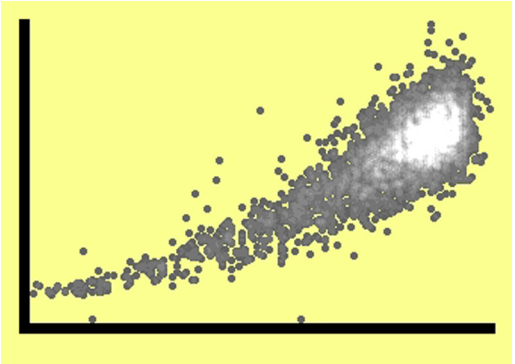

Figure 6 Babies' birth weights plotted against duration of pregnancy.

Neither drawing sunflowers at each point with the numbers of petals representing the numbers of points, nor jittering the data (adding small random components to each point to reduce the overlapping) are serious solutions for large data sets like this (although both suggestions are to be found in the literature). Sounder solutions have been suggested by several authors including Carr et al. (1987) and Hyndman (1996), but would be better implemented interactively. Single displays optimized on a particular criterion cannot reveal all the structure in the data. An interesting problem for interactive scatterplots is how to display both highlighted and unhighlighted distributions informatively.

Variations of scatterplots may be used to display subgroup distributions in parallel, e.g., income of those surveyed by job category. The resulting parallel dot plots are a valuable initial exploration tool for smaller data sets, but parallel box plots are more effective for large data sets as in Figure 7.

Bar charts are good for comparing absolute numbers, but are not good for displaying proportions. Simple pie charts are good for emphazising shares, though it is difficult to compare the sizes of individual sectors and too often they are drawn with irrelevant three-dimensional features and other distracting distortions. Spine plots are best for comparing proportions based on selections, but can only deal with one selection at a time (see Figure 3 above). Switching between bar chart and spine plot representations interactively is one solution.

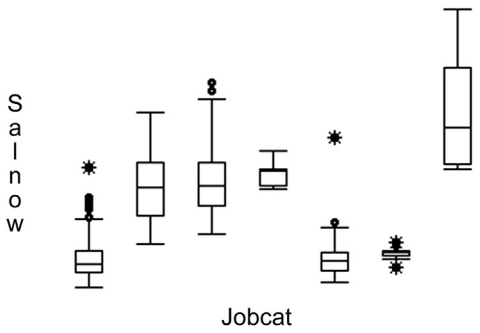

Figure 7 Box plots of salaries by job category in a bank.

Linking

Linking means that points selected in one display are highlighted in all other displays. Selecting a subgroup may identify an outlier within the group in another plot. There may be patterns within subgroups which mask each other when the data for all groups are shown together. Highlighting area-based plots is straightforward, an area corresponding to the size of the selected group is colored. Highlighting point displays is complicated by overplotting, as mentioned above in relation to scatterplots. The best solution for box plots is to be to draw two box plots, one for all the data in the background and second for the highlighted data on top. This idea is due to Wills (Unwin et al., 2006).

Linking is central to successful interactive graphics in the way it provides insight into multidimensional structure. Not just points in graphics may be linked, but cases, scales, orderings, colors may be linked in any kind of window. In this sense querying is just a form of linking.

Interfaces

All of these tools are exploratory in nature and only work effectively when they are interactive. If you have to wait for the screen to clear and a graph to redraw, you will have forgotten why you wanted the change in the first place. Interactivity implies an intuitive interface, flexible commands, and a fast response. Command-line interfaces are valuable for the precise control they offer and the ease with which they can be recorded. For exploratory work it is direct manipulation of the objects on screen which is important and it is rarely of interest to know how a conclusion was first discovered, as it is common to find much better ways of showing results after you have first discovered them. All components of a graphic display (bars, axes, points, legend, etc.) may ideally be queried, moved, rescaled, and reformatted. To avoid clutter on screen, it is helpful to use cues to hint at what is possible rather than to fill the screen with visible controls. To protect the user from possible unwanted side effects of their actions, warnings and undos are beneficial.

Dynamic Graphics

The term 'dynamic graphics' refers to graphical displays that move, such as rotating plots, grand tour displays, or animations controlled by sliders. (A slider is a control that allows the user to change the value of a parameter of a model in real time and see the affected displays update continuously. *See* Data Desk or Tierney, 1990.) While the resulting dynamic displays may be informative, it is the interactive control in the hands of the user, which is most useful, and in this sense they are a specialist subset of interactive exploratory graphics.

Multivariate Displays

Continuous Variables

Modern graphical tools are very good for exploring many variables at once. Scatterplot matrices (a matrix display of scatterplots of all possible pairs of variables) have often been suggested for continuous variables, though they are inefficient in their use of space. For up to about eight variables (the number depending on the screen's resolution and the analyst's eyesight) they can be used for exploratory purposes with linking.

Parallel coordinate displays, e.g., **Figure 8**, are primarily for large numbers of continuous variables. Each variable is represented by a vertical axis and individual cases are joined by lines. There is an attractive geometric duality, pointed out by Inselberg (2009), that a point in a scatterplot is represented by a line in a parallel coordinate display and vice versa. Interactive tools are essential for these plots for two reasons: the order of the axes affects what can be seen, so it has to be possible to change the order easily; the large numbers of lines for a data set of even moderate size makes it essential to be able to temporarily hide points and lines. Categorical variables can also be displayed on their own axes (Kosara et al., 2006). Since there is no natural order of categories for most such variables, being able to reorder categories interactively is also valuable. While there is, in principle, no limit to the number of axes, which may be displayed in a scrollable window, it is clear that these displays provide only limited insight to the full multidimensional structure.

Categorical Variables

Interest in investigating multivariate categorical variables has grown with the rise of data mining. The large data sets which banks, retailers, and other commercial organizations want to study often include many categorical variables. Survey data sets commonly do too, but it is rare to find visualizations of such data. Mosaic plots are one way of uncovering the multivariate structures. Although first discussed in general by Hartigan and Kleiner in 1981, it was only in the late 90s with the addition of interactive features that they revealed their true potential (Hofmann, 2000). A mosaic plot in its original form is constructed by dividing up the horizontal and vertical axes in turn by each new variable. **Figure 9** shows a diagram for the Titanic data set (Dawson, 1995). Initially the horizontal axis was divided up according to the numbers in each class, then the individual columns were divided vertically by age, then the resulting cells were each divided horizontally by gender. Finally, the numbers surviving in each group have been highlighted.

This mosaic plot allows the comparison of survival rates by gender within class and age group. The higher survival rates of the female adults compared to the males are obvious. Redrawing with a different order of variables (gender, age, class) permits direct comparisons of survival rates by class within each gender and age group. There are gaps drawn between the cells for two reasons: the display is easier to interpret and empty cells can be shown (for instance, there were no children in the crew of the Titanic).

Alternative versions of mosaic plots include: (1) the same bin size display, where each cell has the same size; this is good for identifying patterns of empty combinations and for comparing highlighted rates for all combinations; (2) fluctuation diagrams, where each cell's size is proportional to the count of the combinations which it represents, but the cell is drawn at a fixed position on the grid. Fluctuation diagrams have the advantage of maintaining a clearer structure, but can

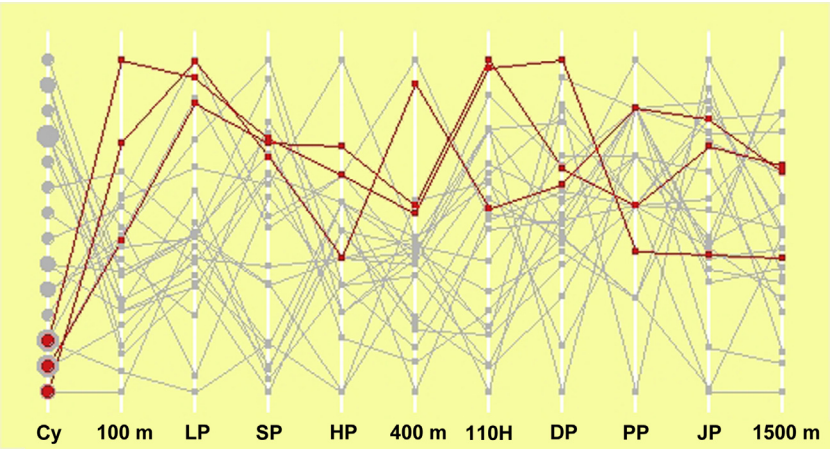

Figure 8 A parallel coordinate plot of the points by event for the mens' decathlon at the Sydney Olympics 2000. The medal winners have been highlighted.

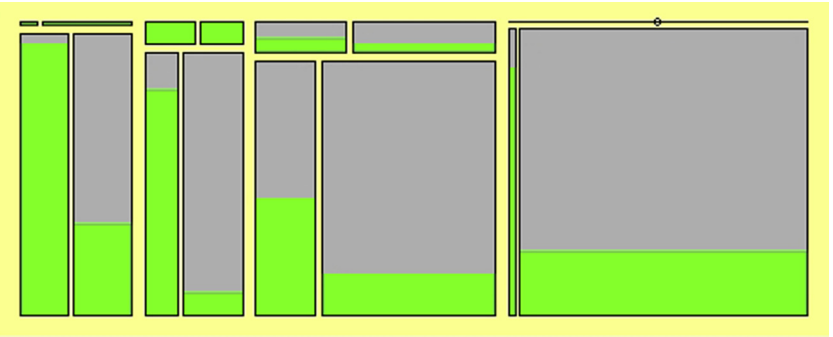

Figure 9 Mosaic plot of the Titanic data in the order class, age, gender with survivors highlighted.

be very inefficient in their use of space in comparison to the original format. The alternating of horizontal and vertical axes is an arbitrary restriction, and Hofmann has introduced double-decker plots, where only the horizontal axis is divided, as a particularly insightful way of looking at association rules.

One of the most useful approaches for analyzing multivariate data, especially with large data sets, is to study subsets of the data in parallel as in **Figure 7**. Trellis displays (Sarkar, 2008) are a more sophisticated version and draw individual displays for each subset in a structured panel display. As in a mosaic plot, the panels are defined by a hierarchical succession of variable splits. Unlike mosaic plots the panels are all the same size. Each panel may display any one of a number of basic statistical graphics, e.g., a box plot or a scatterplot. The art in a successful trellis display lies in the choices of panel plot and of how to split the data set into subgroups. For this reason it would be useful for them to be interactive, but at the time of writing trellis displays are only available in a static form.

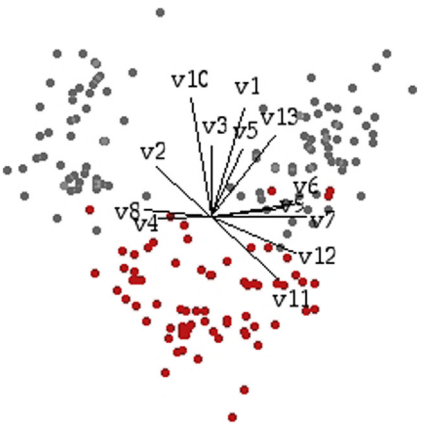

Figure 10 A biplot of the 13 variables and 178 cases of the UCI Wine data set.

Multivariate Projective Displays

The multivariate displays mentioned thus far show all the raw data. An alternative is to concentrate on projections of the data onto specially chosen lower dimensional spaces. Good projections reveal almost all the information in the data, albeit in the transformed space. Biplots are so-called because they show both the projected points and also the axes of the variables (for continuous variables) or the category level points (for categorical variables) in the projection space (Gower and Hand, 1996). **Figure 10** shows the UCI wine data set with the second group highlighted. This projection separates the groups fairly well. Long axes show that the corresponding variables are well represented in the projection space. Biplots have many attractive properties, but remain two-dimensional representations. The larger the number of variables, the less likely it is that a biplot can reveal the major features of a data set, but the addition of interactive tools enhances the capabilities of biplots substantially (Hofmann, 1998).

The first major innovation in exploratory graphics was none of these tools, it was the rotating plot (Fisherkeller et al., 1974 in Cleveland and McGill, 1988), in which a succession of two-dimensional images gives the impression of rotation in three dimensions. This is a natural extension of a scatterplot, is excellent for three-dimensional continuous data, and is fun to use, but for higher dimensional data sets it is not so practical. In that case it can be used in a different way, displaying three-dimensional projections of the data. Many methods of identifying 'interesting' projections have been discussed and the most effective implementation seems to be a directed grand tour. The grand tour was designed to rotate through all possible three-dimensional projections of a data set of continuous variables and the analyst would decide which of the resulting projections gave him insight into the data. Directed grand tours make use of projection pursuit indices to guide the user toward revealing projections (Cook and Swayne, 2007).

Specialist Displays

Missing Values

In the histogram in **Figure 2** there was a special column for recording the numbers of cases with no values for that variable. Missing values are a problem for large data sets and any initial overview of the data needs to check for numbers of missings and for any structure among them. Some statistical methods cannot handle cases with missings. A simple approach is the missing value plot as in **Figure 11**. There is a bar for each variable. The solid part of the bar on the left represents the cases with values, and the blank section on the right represents the missing values. Bars can be queried and selected to give a rough

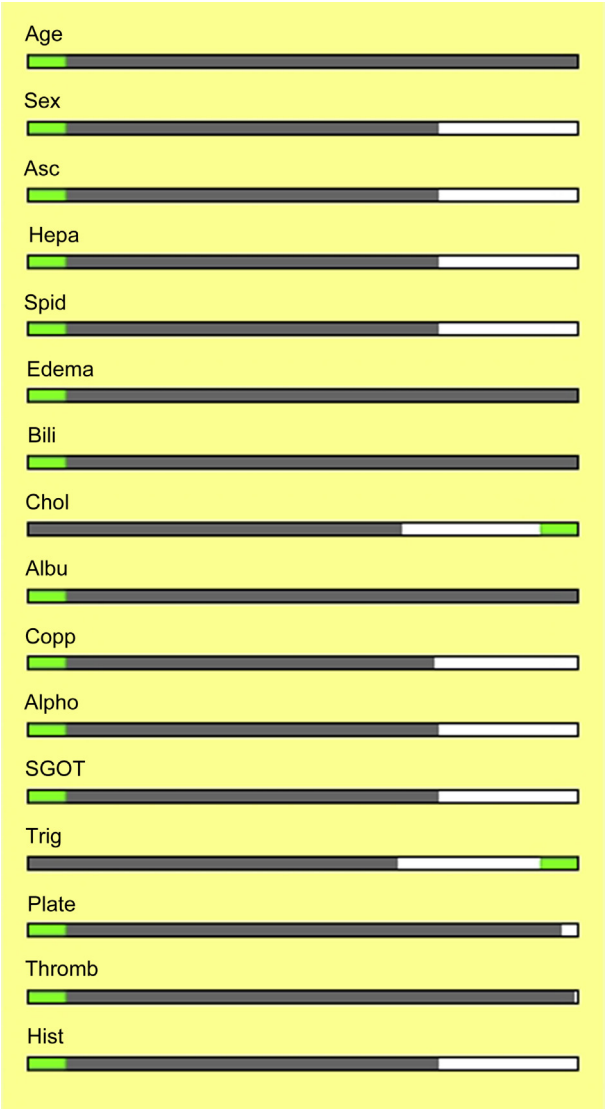

Figure 11 A missing value plot for variables from a patient survival study. A subgroup missing on only two variables is highlighted.

idea of associations. In the figure it is clear that many cases are missing on a large number of variables.

This kind of display cannot reveal more complex structure. Displays of the multivariate patterns of missing values are one promising approach, though interactive versions are not yet available.

Weighted Displays

In data sets arising from surveys or based on geographic areas it can often be the case that individual data points are of differing importance: the unemployment rate in London should have a greater weight than the unemployment rate in Stratford-upon-Avon. Some authors have suggested using bubbles instead of points in dot plots and scatterplots, where the size of the bubble is proportional to the weighting variable. This can be effective for small numbers of cases, but generally leads to overplotting problems. Area-based displays such as histograms, bar charts, or mosaic plots have no such problems and counts are simply replaced by sums of weights. This can lead to very striking and informative differences in displays. The left-hand histogram in Figure 12 just about shows a single county with the highest percentage of blacks among the over 400 counties in the five US Midwest States. Most counties have low percentages of blacks. The right-hand histogram has the same horizontal scaling, but the height of the bins is now proportional not to the number of counties, but to the aggregate black population in the bins. Most blacks live in counties with relatively higher percentages of blacks. It is, of course, necessary that the weighting variable is additive! Population is fine, population density is not (see Unwin et al., 2006).

Statistical Tools and Plots

Surprisingly, and disappointingly, statistical graphics rarely make use of statistical tools. There are a few exceptions. Some packages offer the option of displaying box plots with confidence intervals around the medians. Some packages offer the chance of plotting error bars round means. SAS have software

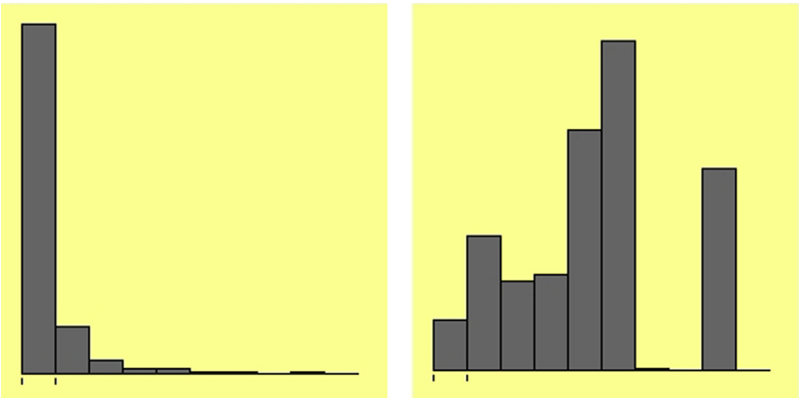

Figure 12 Histograms of black population percentage by US Midwest county, unweighted (left) and weighted by black population (right).

(JMP) where circles are drawn round subgroup means to enable graphical detection of whether differences are significant or not. However, the last example illustrates one of the problems. JMP offers several different options because there is no agreed optimal way of comparing several means simultaneously. The same problem would arise with bar charts and more seriously with mosaic plots – there are so many possible statistical tests, that it is not easy to specify a universally acceptable default. The solution will be to offer the various options interactively.

The Future

Advances in statistical graphics have not been due to improvements in screen resolution. The best exploratory graphics are simple displays. What makes them powerful is the speed and flexibility with which they can be manipulated and combined. It is the increase in computing power which has been crucial for modern interactive graphics. More powerful computing has also led to the development of computer-intensive analytic procedures and these have sometimes been combined successfully with graphics. Evaluation of projection pursuit indices and rotating plots is one example. Smoothing methods such as loess and spline models are another.

A major requirement for all statistical methods is whether they can be scaled up to deal with huge data sets. Area-based graphical tools such as histograms and bar charts can in principle be used for data sets of any size, but interaction is necessary to catch details and explore links between variables. Much more needs to be done to cope with displays for large numbers of variables, and interactive capabilities will be essential.

Statistical graphics has grown in fits and starts. There are many attractive ideas, some general principles, but little theory. Real progress will only come with a successful formalization, so that the future can build on the successes of the past. This is something to look forward to in all senses of the phrase.

Software

The displays in this article have been taken from screenshots. There has been no attempt to pretty them up, they illustrate exploratory analyses in progress. Most are from MANET (Unwin et al., 2006) and could be drawn in Mondrian (www.rosuda.org/Mondrian/), a couple are from Data Desk (www.datadesk.com). Other software also offers interactive capabilities, but the degree of interactivity varies considerably. For up to date information on software, check statistical Web sites.

See also: Correspondence Analysis; Exploratory Data Analysis: Univariate Methods; Graphical Methods: Presentation; Missing Data; Multidimensional Scaling I; Multivariate Analysis: Classification and Discrimination; Spatial Statistical Methods.

Bibliography

Carr, D.B., Littlefield, R.J., Nicholson, W.L., Littlefield, J.S., 1987. Scatterplot matrix techniques for large *N*. Journal of the Acoustical Society of America 82 (398), 424–436.
Cleveland, W.S., McGill, M.E., 1988. Dynamic Graphics for Statistics. Wadsworth & Brook's Cole, Pacific Grove, CA.
Cook, D., Swayne, D.F., 2007. Interactive and Dynamic Graphics for Data Analysis. Springer, New York.
Dawson, R.J.M., 1995. The 'unusual episode' data revisited. Journal of Statistics Education (Online) 3 (3).
Gower, J.C., Hand, D.J., 1996. Biplots. Chapman & Hall, London.
Hofmann, H., 1998. Interactive biplots. New Techniques & Technologies for Statistics 98 (1), 127–136. Sorrento: Eurostat.
Hofmann, H., 2000. Exploring categorical data: interactive mosaic plots. Metrika 51 (1), 11–26.
Hyndman, R.J., 1996. Computing and graphing highest density regions. American Statistician 50, 120–126.
Inselberg, A., 2009. Parallel Coordinates: Visual Multidimensional Geometry and Its Applications. Springer, New York.
Kosara, R., Bendix, F., Hauser, H., 2006. Parallel sets: visual analysis of categorical data. Transactions on Visualization and Computer Graphics 12, 558–568.
Sarkar, D., 2008. Multivariate Data Visualization with R. Springer, New York.
Spence, R., 2007. Information Visualization: Design for Interaction. Prentice Hall, New York.
Theus, M., Urbanek, S., 2007. Interactive Graphics for Data Analysis. CRC Press, London.
Tierney, L., 1990. Lisp-Stat. Wiley, New York.
Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, London.
Unwin, A.R., 2012. Oscars and interfaces. Journal of Statistical Software 49, 1–18.
Unwin, A.R., Theus, M., Hofmann, H., 2006. Graphics of Large Datasets. Springer, New York.
Ward, M., Grinstein, G., Keim, D., 2010. Interactive Data Visualization: Foundations, Techniques, and Applications. AK Peters/CRC Press, New York.
Wilkinson, L., 2005. The Grammar of Graphics, second ed. Springer, New York.

Graphical Methods: Presentation

Leland Wilkinson, SPSS Inc., Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 9, pp. 6368–6379, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

This article surveys briefly the history of presentation graphics, principles of usage, and applications. Many popular charts were invented by social scientists centuries ago, although social scientists in the twentieth century were more reluctant to use them than were natural scientists and business people. Recent developments in graphics theory and may change this pattern. Well-designed graphics summarize both models and data. They conserve space and highlight essential features with modern computer software that implements recent technical advances in graphics, graphical displays may prove to be one of the most powerful research tools available to social scientists.

Figure 1 shows a graphic of death rates against birth rates per 100 000 population for 27 selected countries in a 1990 UN databank (Wilkinson, 1999b). The contours in Figure 1 show two concentrations of countries. One, to the left, has relatively lower death rates and small-to-moderate birth rates. The second, toward the upper right, has high death rates and extraordinarily high birth rates. The curve in the middle of the contours shows that the overall relation between death and birth rates is curvilinear.

The single plot in Figure 1 contains three graphs: (a) a point graph (collection of points) whose labels show country names; (b) a set of contours based on a kernel density estimate (Silverman, 1986) that represents the concentration of the countries; and (c) a LOESS smoother (Cleveland and Devlin, 1988). The graphic also includes three guides for our understanding. The first is a general geometric object called a form that is, in this instance, a line delineating zero population growth. Countries to the left of this line tend to lose population, and countries to the right tend to gain. The other two are guides that demarcate axes for the represented space. Other examples of guides are legends and titles.

Each of the constituent graphical elements in Figure 1 has a long lineage in the history of statistical charts. Scatterplots of points and contours are several centuries old, and regression smoothers are at least a century old. The use of a rectangular coordinate system to locate graphs within a frame demarcated by axes predates by centuries Decartes' eponymous Cartesian coordinate system. We shall review briefly the history of statistical graphics and charts.

The lines, colors, typography, proportions, and other aspects of Figure 1 have been chosen to facilitate communication of the ideas and models underlying the chart. We do not need to resort to platitudes such as, 'A picture is worth a thousand words' to understand that this graphic comprises a relatively substantial analysis of the data. We shall review briefly the topic of graphics principles which can help us devise effective ways to display analyzed data.

The world of charts is enormous. Principles can guide us through that world, but there are many alternatives for representing one or more variables in a graphic. In the final section, we shall present some of these choices and use them to derive additional guidelines for developing effective presentations.

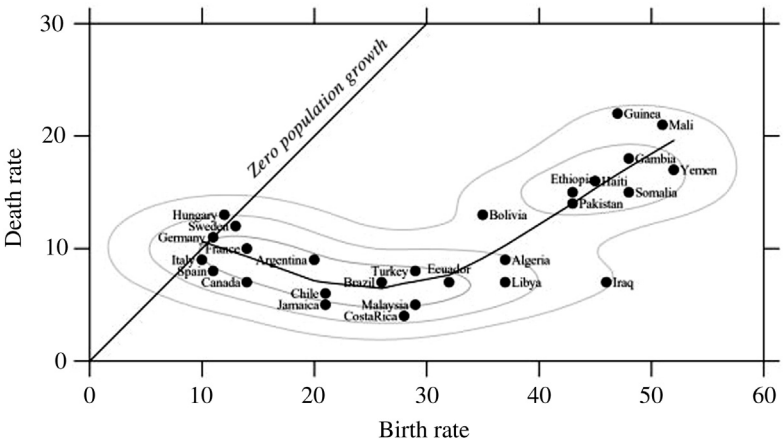

Figure 1 Plot of 1990 death rates against birth rates for selected countries.

History

The contours, or level curves, in **Figure 1** represent a map. This representation is based on a statistical function that maps birth rate and death rate onto a density. Any value on this density can be derived from a corresponding pair of birth rate and death rate values taken from the domain of values within the frame bounded by the two displayed scales. The geometric space in which this mapping is embedded is three-dimensional. The first two dimensions are birth rate and death rate, and the third dimension is density.

The technical mathematical terms in the previous paragraph (contours, level curves, map, function, onto, density, value, domain, frame, bounded, scales, geometric space, and dimensions) are derived from ordinary language, often ancient, describing the physical world. This correspondence between the world of charts and the physical world is not coincidental. Charts are maps of abstract worlds. The word chart and the word cartography have the same root (Latin *charta*, a piece of paper or papyrus). The Greek word for geometry means 'land measurement.' A scale (in Latin) is a ladder. We map in order to organize our world in our mind. However, our world is larger than the domain of geography. As Pinker (1997) has argued, abstract reasoning is built on metaphors for reality. We manipulate abstractions by making them analogous to concrete objects.

Statistical charts are maps, but it does not suffice to simply describe them as abstract geography. The history of statistical charts contains remarkable leaps from a world of continuous rectangular (or spherical) space to other worlds that are categorical (worlds with gaps), multidimensional (worlds beyond three dimensions), and topological (worlds in abstract coordinate systems). Funkhouser (1937), Tilling (1975), Beniger and Robyn (1978), Fienberg (1979), Robinson (1982), Stigler (1983), Tufte (1983, 1990, 1997), Collins (1993), Wainer (1997), and Wainer and Velleman (2001) document these leaps. These references also contain numerous examples of graphics, ancient and modern, that represent significant landmarks in this history. The remainder of this section will highlight a few.

Physical coordinates were used by Egyptian surveyors as early as the third millennium BCE for locating points on land (Beniger and Robyn, 1978). By the beginning of the second millennium CE coordinates were used to locate nongeographic entities. Funkhouser (1936) shows an example of a line graphic of planetary movements on a space–time grid produced by an unknown astronomer in the tenth or eleventh century. Not surprisingly, one of the 'planets' is the Sun.

Cartesian coordinates were devised by Descartes for locating points in abstract space, particularly for representing equations underlying his analytic geometry (Funkhouser, 1937). However, relatively few seventeenth or eighteenth century embeddings of observed data points in Cartesian coordinates have been found. (Funkhouser, 1937; Tilling, 1975). Scientists of the Enlightenment sometimes regarded graphics as dilettantish, and if they used them at all, it was for geometrical diagrams or maps of algebraic functions rather than for plots of raw data (Beniger and Robyn, 1978). Social scientists of this period focused on collecting tables of state statistics that could be perused in detail without resorting to graphical approximation (Beniger and Robyn, 1978, although see Stigler, 1983). Beniger and Robyn (1978) and Wainer and Velleman (2001) discuss notable exceptions to these observations (graphics by Wren, Halley, Huygens, Plot, Lister, Priestley) but conclude that the important varieties of what we now call the statistical chart did not arise until the work of Playfair in the late eighteenth century.

Playfair graphed the data of the state, the root of the word statistics. He invented the pie chart, the circle chart, the categorical bar chart, the area chart, and he adapted other geometric forms in novel ways to graph data over time and space. Wainer and Spence (1997) provide a brief biography of Playfair's fascinating life and the charts he created.

Other widely used graphical forms were developed in the nineteenth and early twentieth centuries. Beniger and Robyn (1978) and Stigler (1983) summarize a number of these. Density bars were used by Guerry in 1833 to display crime by age groupings. Pearson adopted this form for representing a frequency distribution and named it a histogram. Minard developed a number of schematic maps overlaying statistical data on geography. Florence Nightingale introduced polar area charts. The age–sex pyramid (a dual histogram) was introduced by F. A. Walker, Superintendent of the US Census in 1874. The sample cumulative distribution plot was introduced by Galton in 1875.

If charts are maps of abstract worlds, we might conclude that guiding principles for graphics usage could be derived from the psychology of perception and cognition in the physical world. Indeed, recent psychological research has tended to support the simple statement that the more graphics adhere to the rules of the physical world, the more likely we are to extract and process the information in them, reliably and accurately. Section Choices covers this area.

Principles

Over the centuries of their history, various charts have emphasized communication (e.g., Snow's map of cholera deaths in London, reproduced in Tufte, 1983), persuasion (Fletcher's map of the distribution of ignorance in nineteenth century England and Wales, reproduced in Wainer, 1997), and graphic design (e.g., Playfair's chart of imports and exports of England to and from North America, reproduced in Tufte, 1983). These criteria are, of course, not exclusive. Some charts incorporate all three (e.g., Minard's compound chart of Napoleon's troop losses in the Russian campaign, reproduced in Tufte (1983).

Not surprisingly, modern critics of charts have tended to align themselves along these same dimensions. Communicators (e.g., Cleveland, Kosslyn) have conducted experiments to determine rules that improve the transfer of information from chart-maker to chart-viewer. Persuaders (e.g., Holmes, 1991) have evaluated charts on the basis of their psychological impact. Designers (e.g., Herdeg, 1981) have applied aesthetic criteria. Others (e.g., Tufte, 1983) have adopted the criteria of the communicator, and the methods of the designer. None of these approaches can claim exclusive validity. Since scientists are more interested in rules that foster accurate communication

of replicable results, however, this section will focus on communication.

Evaluating graphical communication has been the province of psychology, in the subdisciplines of human factors, ergonomics, and applied cognitive science. For almost a century, human factor psychologists have developed methods and criteria for evaluating graphics. Military psychologists, for example, have conducted randomized experiments to test the effectiveness of aircraft cockpit displays.

Psychologists gave much less attention to the perception of statistical graphics until a statistician, William Cleveland, published a series of experiments designed to identity aspects of charts that helped or hindered accurate decoding of quantitative data (Cleveland and McGill, 1984). (Beniger and Robyn, 1978, cite a number of statisticians' informal studies of the effectiveness of popular charts in the 1920s, but these did not induce many psychologists to study the topic more formally.)

Cleveland used paper-and-pencil tests containing simple graphical elements – points, lines, angles, areas, colors – generated randomly by a computer. His subjects gave magnitude estimates of the values generating the instances of each element. From these estimates, he derived and analyzed error rates. **Figure 2** summarizes the main result of these experiments. The top of the figure (BEST) represents elements that resulted in fewer errors and the bottom (WORST) represents those with more.

Cleveland's hierarchy gives us a start in designing effective displays. All other things being equal, we might choose to use a bar chart instead of a pie chart because judgments of position on a common scale and judgments of length are both more accurate than judgments of angle. There are exceptions to these rules, however (see the following discussion of **Figure 3(b)**). A chart is more than the sum of its parts; elements interact in the context of a chart to suppress or enhance errors. Kosslyn (1994) summarizes relevant psychological studies in the extensive notes at the end of his book.

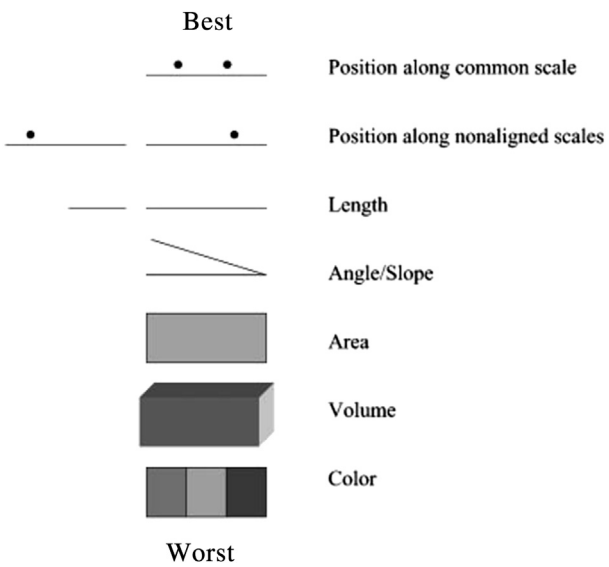

Figure 2 Cleveland (1994) graphical features hierarchy.

Others have provided numerous general rules for effective graphical communication. The following guidelines are derived from Bertin (1981), Tufte (1983), Cleveland (1994), and Kosslyn (1994).

Avoid clutter. Especially, do not clutter the data region inside the plotting frame demarcated by axes. Tufte constructs a data/ink ratio and urges us to maximize it by removing irrelevant detail. This exhortation has value as long as we do not remove redundant features that can reinforce an accurate perception. Most writers agree that one should avoid jazzy textures (especially stripes), gratuitous colors, excessive tick marks and grid lines, ornate fonts, and unnecessary embellishment. Avoid visual illusions due to pseudo-3D and other gratuitous use of color, angles, area, or volume in a 2D plotting world.

Seek simple geometric forms – straight lines, circles, triangles, squares. When it helps to simplify a relationship, use transformations (logs, square roots) to convert curves to straight lines, for example. Use polar coordinates when variables are cyclical (astronomical perceptual, directional).

Sort and organize. This is especially important for complex graphics. Usually, one should sort the values of a dependent variable (the range of the graph), but other sorts can help viewers navigate and perceive structure. Bertin (1981) has numerous examples showing the value of simple and multivariate data sorts. **Figure 12** shows an example.

Annotate extensively. Make legends comprehensive and informative. Figure captions should describe the source of the data and explain the relation of the data to the graphic. Note exceptions (see **Figure 6**).

Choices

Rules guide, but examples instruct. We need to consider specific choices for specific problems to understand exceptions to rules. The graphic representation problem is often expressed as one of choosing the proper chart type for a given set of data. This is an unhelpful way to express the problem, however. As Wilkinson (1999c) shows, different chart types with their own popular names are often mirror images or simple transformations of each other. Instead, we will examine several basic problems confronting the designer of a scientific graphic: geometry (choosing the type of graphical element for a chart), coordinates (choosing a coordinate system for the chart), and uncertainty (representing random error).

Geometry

The following examples were selected to illustrate principles, not exhaust possibilities. It is organized by the configuration and number of variables, categorical and continuous, that we wish to graph.

One variable

Figure 3 shows five different ways of representing values on one categorical variable. The data are from 1606 respondents to the 1993 General Social Survey (Davis et al., 1993). Respondents were asked, 'How many sex partners have you had in the last 12 months?' Those reporting more than four partners

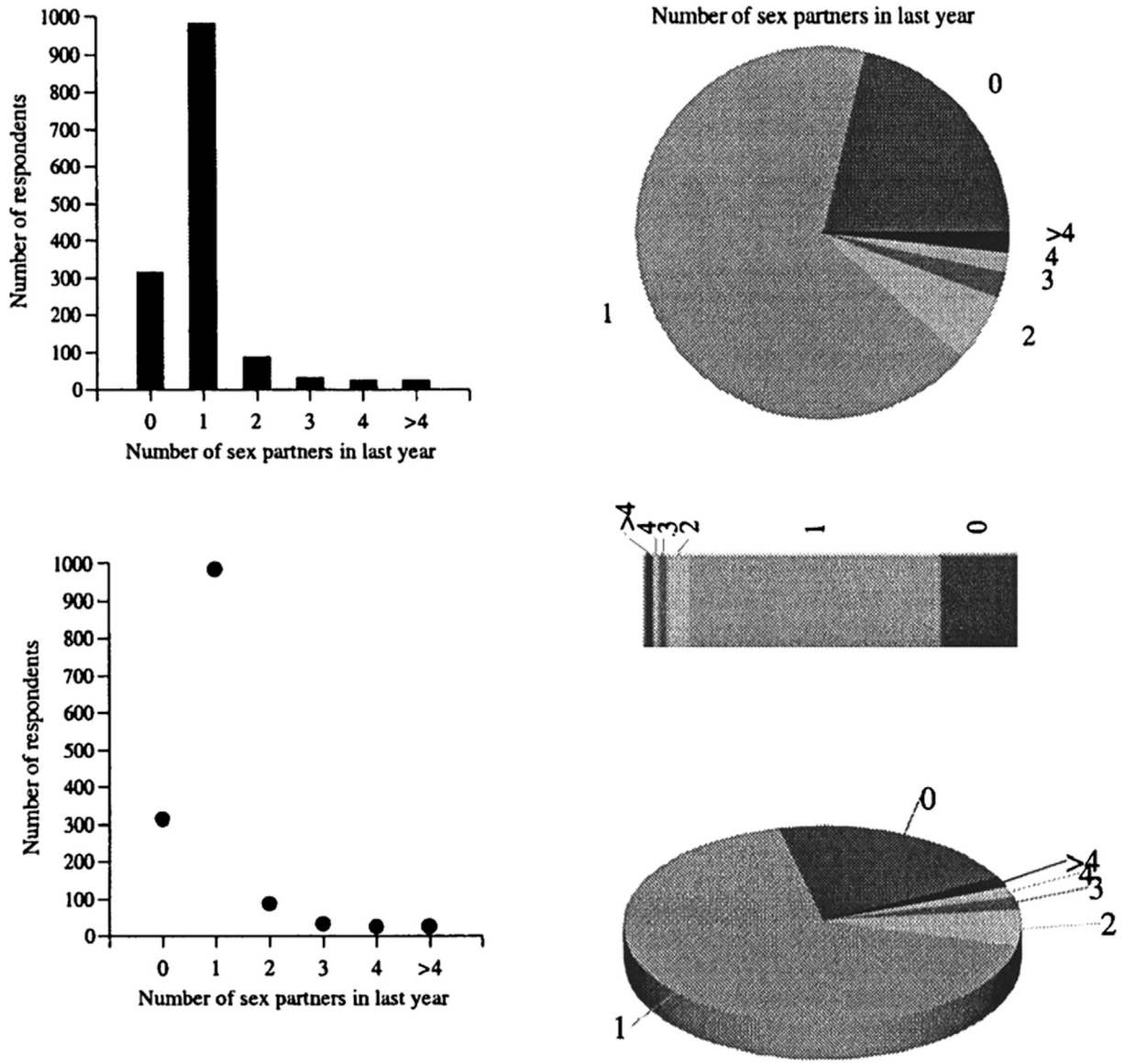

Figure 3 Reported number of sexual partners (General Social Survey, 1993).

(some reported up to 100) were consolidated into the last category. There were 1466 responses in the resulting six categories.

The upper left graphic is a bar chart of the frequencies of occurrence of each category. Bars are most suited for displaying either magnitudes referenced against zero (anchored bars, as in this figure) or a range of values on a continuous scale (range bars, as in Figure 7).

Anchoring bars at zero can sacrifice resolution when no data values exist near zero; this creates substantial white space in the plotting area (Cleveland, 1994). In these cases, dots are preferable to bars because we do not need to include zero on the scale. The upper right graphic shows a dot plot of the same variable. In this case, we would prefer the bar version because there are several values near zero and it is easier to anchor the bars visually on the partner categories. The locations of both the dots and the tops of the bars are easily decoded on the vertical scale.

The third graphic in Figure 3 is a pie chart of the proportions of total category frequencies. The general antipathy statisticians have toward pie charts is misinformed. Pie charts are not always bad, although the issues on when to use them are complex. In general, bar charts are more suited to absolute judgments and pie charts are more suited to judging proportions (Simkin and Hastie, 1987). Pie sectors near a fourth or half of a pie in size are judged especially accurately. When there are many sectors, however, pies are not usually appropriate. In this example, the bar chart would be more suitable.

The fourth graphic is a divided bar chart, which is nothing more than a pie chart in rectangular coordinates. Divided bars do worse than pies in studies of graphical perception

(e.g., Simkin and Hastie, 1987). Cleveland (1994) provides several examples to show why they are generally a bad choice.

The fifth graphic is a 3D pie chart. This is probably the worst of the popular ways to represent proportions. Three-dimensional graphics are not in themselves harmful. What creates the problem with the 3D pie chart is the distortion of angles that the perspective projection produces. In the 3D pie, the single partner category seems to cover almost 75%; the actual coverage is 68%.

Some users and computer programs apply line or area elements (see Figure 7) to categorical variables. Popular examples are line graphics of means in analysis of variance, or area graphics of profiles across categories. These elements should generally be avoided with categorical data. Spacings between categories (even ordered categories) are not fixed, so slopes of line segments or profiles between categories are meaningless.

Figure 4 shows five ways of representing one continuous variable. The data are per-capita consumption of spirits for each of the 50 US states (Bureau of the Census, 1986). The top graphic is a classic histogram overlaid with a kernel density estimate (Silverman, 1986). Histogram bars look like ordinary bars, but they measure areas rather than intervals and they rest on a continuous scale rather than on a set of categories. Histograms are more useful than kernels when we are interested in displaying frequencies within intervals, especially when the intervals themselves are bounded by meaningful, round units. Kernels are more useful than histograms when we are looking for a smooth estimate of a continuous distribution.

The next graphic down is a dot histogram (Chambers et al., 1983; Wilkinson, 1999a). This display represents each observation with a dot located at its scale value. If several values coincide on the scale, the dots are stacked vertically. This display is most suited for small samples, when every value is to be displayed.

The graphic below that is a stripe plot (Chambers et al., 1983). A vertical stripe occurs at each data value. This display can handle more cases than the dot plot, although it is not suitable for large datasets.

The graphic following is a jittered density (Chambers et al., 1983). Random error is added to each value so that the plotting symbols do not overlap. This display is designed to handle roughly the same conditions as the stripe plot.

The last graphic in Figure 4 is a box plot, or schematic plot (Tukey, 1977). As Tufte (1983) notes, Tukey's plot is based on an earlier display that represents the quartiles of a distribution. In that earlier plot, the whiskers cover the range and the box covers the midrange. Tukey improved on this design by scaling the whiskers to allow for outliers. The display in Figure 4 shows how important this feature can be. Two states (Nevada and New Hampshire) have unusually large values (see Figure 6), possibly due to gambling and cheap liquor, respectively. The box plot is a clean representation of the summary values we usually want to see – the median, the twenty-fifth and seventy-fifth centiles (hinges), the extreme values (usually bounding the range), and possible outliers (outside values).

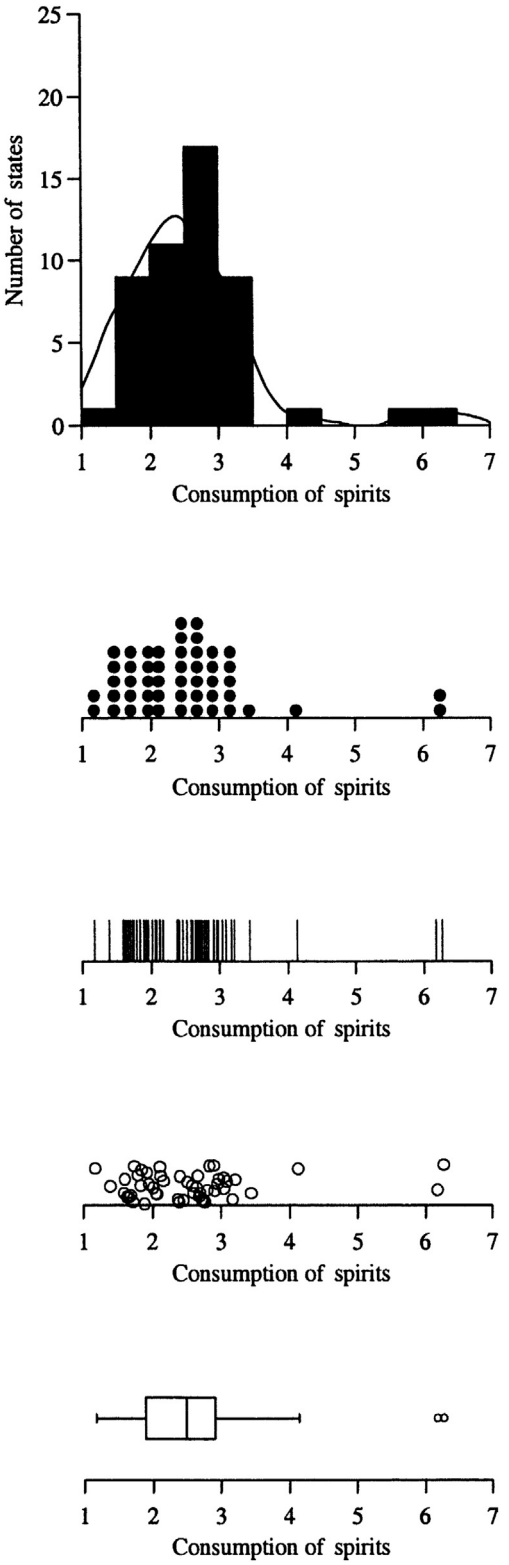

Figure 4 Consumption of spirits in gallons per capita, by US state, 1986.

Two Variables

Figure 5 shows four ways of representing two categorical variables crossed. The data are from the 1993 General Social

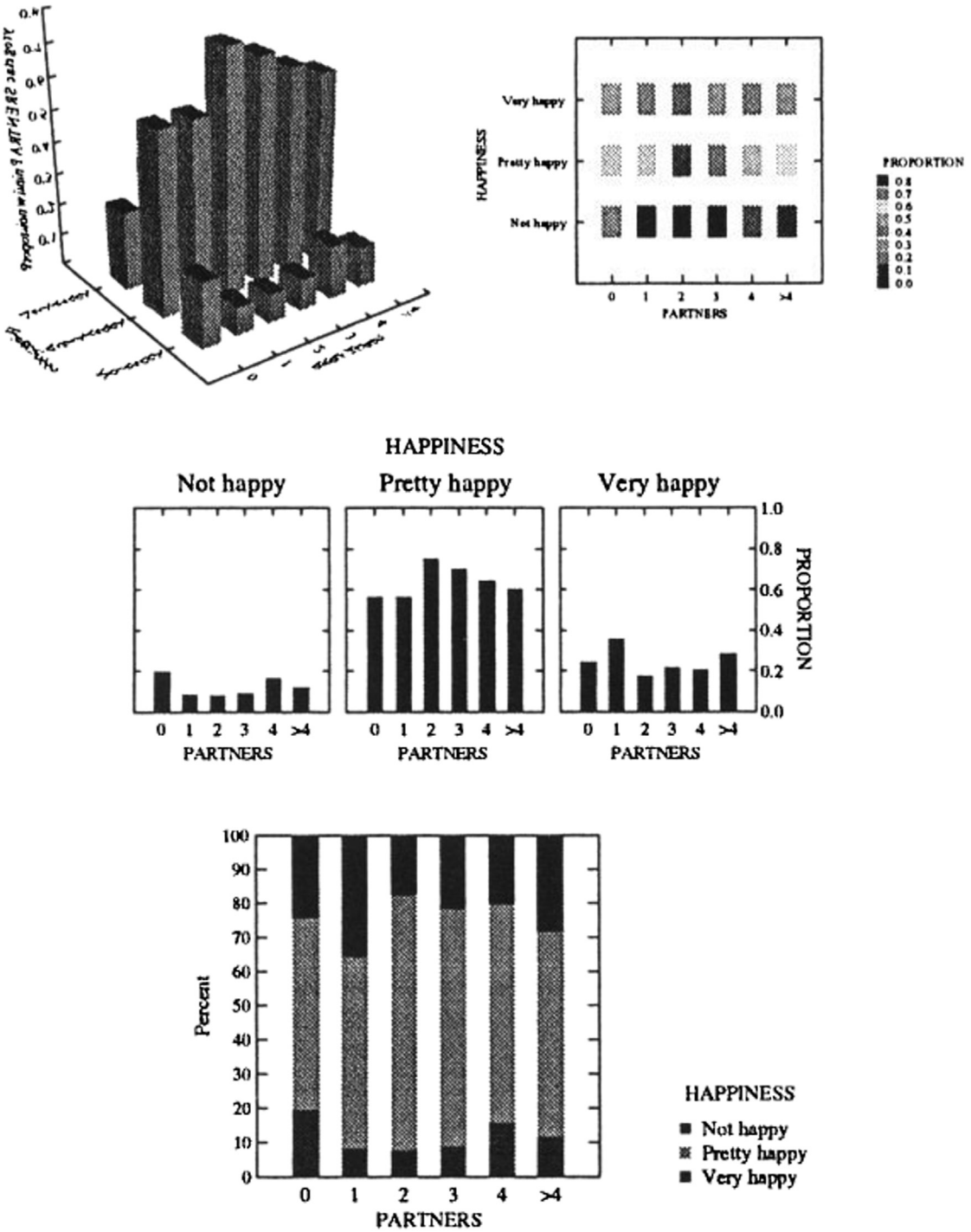

Figure 5 Number of reported sexual partners during previous year by reported level of happiness among 1462 respondents to 1993 General Social Survey. Third variable is proportion of respondents in each sexual partner category.

Survey used in Figure 3. The additional variable is general happiness, as measured by the response to the question, 'Taken all together, how would you say things are these days – would you say that you are very happy, pretty happy, or not too happy?'

The upper left plot shows a 3D bar chart of happiness against number of sexual partners. The vertical scale is the proportion of respondents within each partner category. This graphic has the disadvantage that short bars tend to be hidden by long ones. Generally it should be avoided.

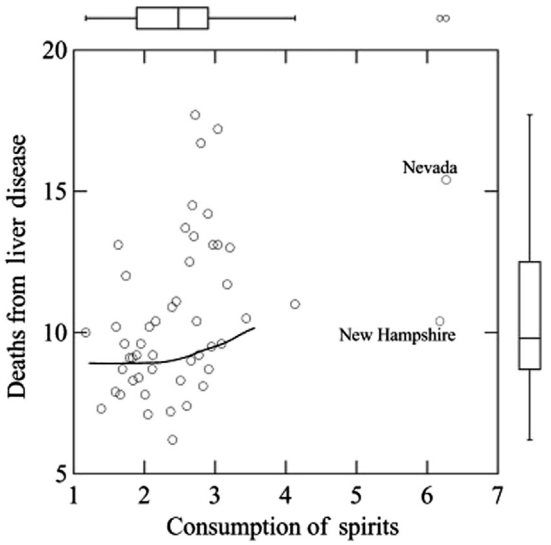

Figure 6 Consumption of spirits in gallons per capita by US state vs number of deaths from chronic liver disease and cirrhosis per 100 000 by US state, 1986.

The graphic to the right is a tiled bar chart. This graphic uses color or gray scale to indicate the relative frequency of each category combination. However, as Cleveland's research has shown, the use of color or gray scale to encode continuous values is problematic (Cleveland, 1994).

The next graphic below shows a paneled categorical plot. Instead of using an extra dimension in 3D, we use an extra categorical dimension in 2D to panel the plot. This is probably the most effective method for handling the extra dimensionality required for representing the frequencies, plus the two categories. The bottom panel shows a multiple divided bar chart. This layout has the same problems as the single divided bar chart (see **Figure 3**). Cleveland (1984) shows an example of this graphic and offers a more effective dot plot alternative.

Figure 6 shows the most popular way of representing two continuous variables crossed with each other – the scatterplot. The data are per capita consumption of spirits for each of the 50 US states, used in **Figure 4** (Bureau of the Census, 1986). The additional variable is number of deaths from chronic liver disease and cirrhosis per 100 000 people, by state. This scatterplot has been enhanced in a number of ways. The bordering

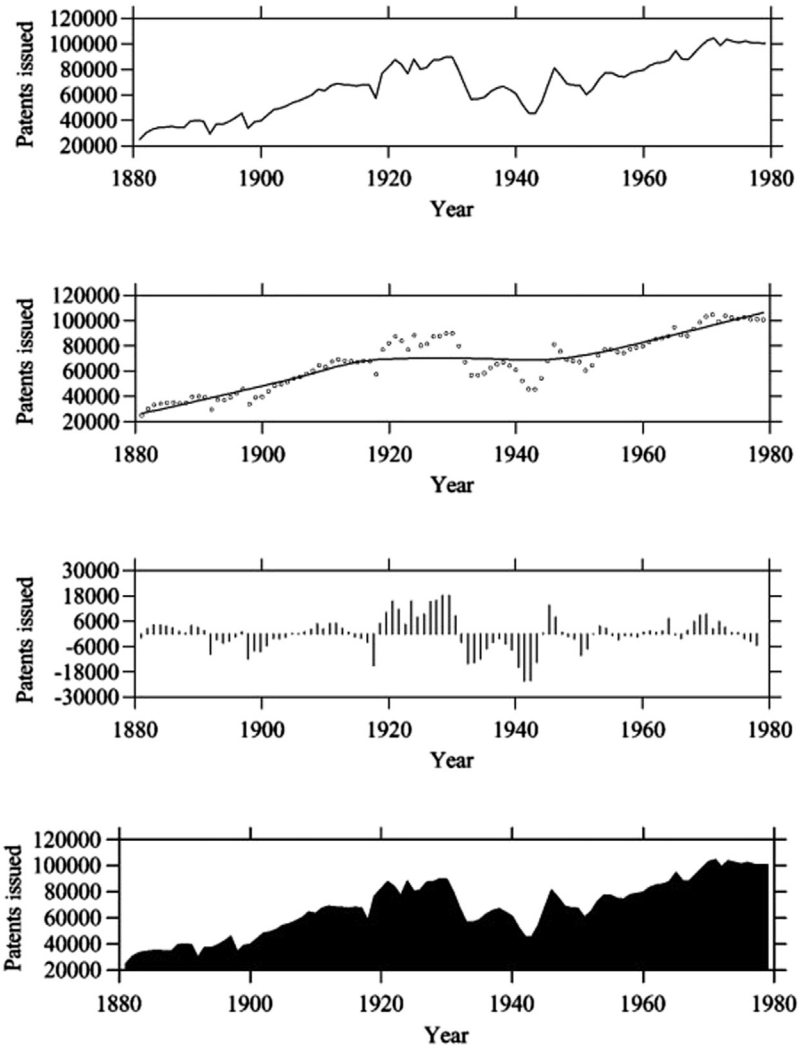

Figure 7 US patents issued from 1880 to 1980.

box plots help us to assess the marginal distributions of the variables and highlight the Nevada and New Hampshire outliers. The smoother shows the conditional mode of deaths (estimated mode of deaths given consumption). This was computed through kernel regression (Scott, 1992). This particular smoother limits itself to areas where there are relatively more data values, so it provides a conservative estimate of trend.

Time series data require special treatment. **Figure 7** shows four examples. The data are number of US patents issued in the century from 1880 to 1980 (Wilkinson et al., 1996). The top panel shows a line element representing the series. Lines are most useful when a series is relatively smooth. The next panel shows a point element with a superimposed LOESS smoother. Points are best when we supplement them with a smoother; they tend to obscure time order if used alone, unless the series is especially smooth. The third panel shows spikes. These are vertical lines best used to reveal deviation from a constant level. This panel displays the residuals from the LOESS smooth. The lowest panel shows an area chart for the raw series. This highlights trend but prevents the use of confidence intervals and other enhancements to the plot. Generally, it should be avoided. Cleveland (1994) discusses other graphical representations for time series data.

Figure 8 shows three ways of representing a categorical variable crossed with a continuous variable. The data are based on 80 graduate students over a 10-year period in a US psychology department (Wilkinson et al., 1996). Graduate Record Examination Advanced Psychology Test scores are plotted against whether or not the students eventually received their PhD.

The first graphic employs a dot for the mean and a range bar to represent a 95% confidence interval on the mean. The bars do not overlap, which suggests that the advanced test can help identify those who achieve a PhD. (Interestingly, neither the Verbal nor the Quantitative test scores predicted graduation in this sample.)

The second graphic is a notched box plot (McGill et al., 1978). This is a variation on the box plot that not only conveys more of the important data landmarks than the classic confidence interval plot but also provides an approximate confidence interval of its own. If the data are independent samples from identically distributed populations that are lumpy in the middle (approximately normal in their midrange), then comparing the notches yields an approximate 95% test of the null hypothesis that the true medians in the population are equal. In this example, the notches do not overlap, reinforcing what we concluded from the confidence intervals on the means. Because it relies on the median instead of the mean, the notched box plot procedure is more robust against outliers.

The third graphic is a dot plot (Wilkinson, 1999a). Used frequently in the medical literature, this plot shows every data point. In a small sample (e.g., clinical case study), a dot plot can be useful for readers who wish to consider every data value. However, it is of little use in making informal graphical inferences concerning group differences. It is best to think of grouped dot plots as unadorned one-dimensional scatterplots within categories.

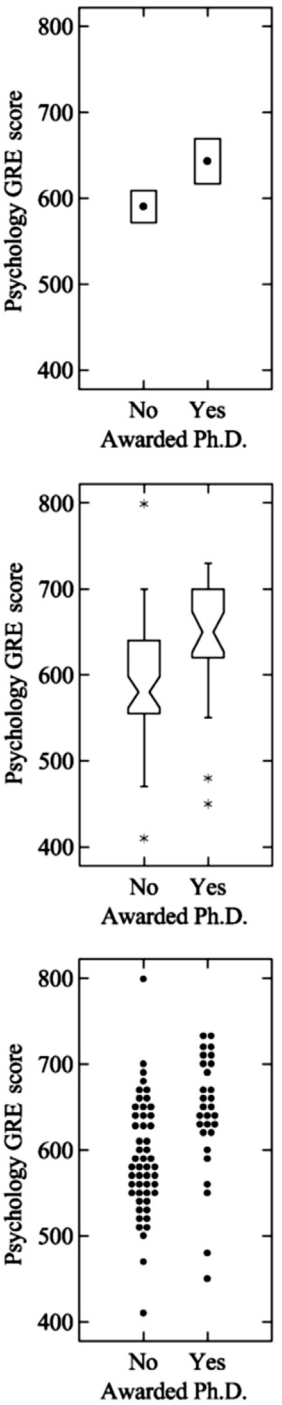

Figure 8 Advanced Psychology Graduate Record Examination (GRE) scores for 80 students taken from 10 years of a graduate psychology program. Students are grouped by whether or not they were awarded a PhD.

Three variables

Figure 9 shows a triple crossing of categorical variables. The data are from the 1993 General Social Survey used in **Figures 3** and **5**. The additional variable is gender (observed by the interviewer, not reported by the respondent). We ask whether

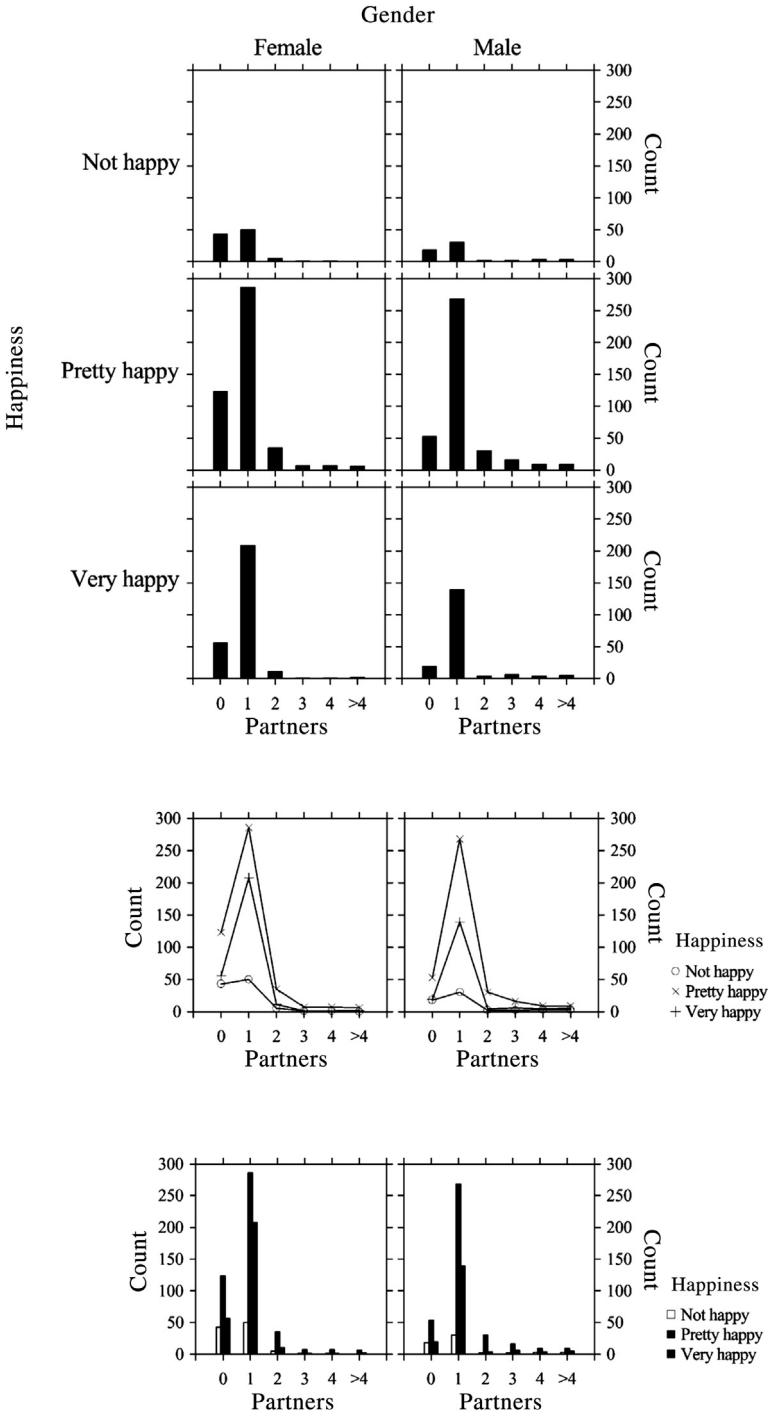

Figure 9 The number of reported sexual partners during last year by reported level of happiness among 1462 respondents to 1993 General Social Survey. The third variable is the proportion of respondents in each sexual partner category.

the relationship between reported happiness and number of sexual partners differs by gender.

The top plot in the figure shows a paneled graphic. The format is similar to a Trellis display (Becker et al., 1996), but the labeling of the paneling variables (happiness and gender) is placed outside the plotting area to improve readability.

The middle plot collapses happiness into a legend in order to reduce the number of panels. This is a popular method for saving space, particularly when representing factorial layouts in ANOVA and other designs. It has several problems. First, the line segments used to highlight trends have slopes that depend on the particular spacing and arrangement of partner categories. Unless we employ an external scaling procedure to

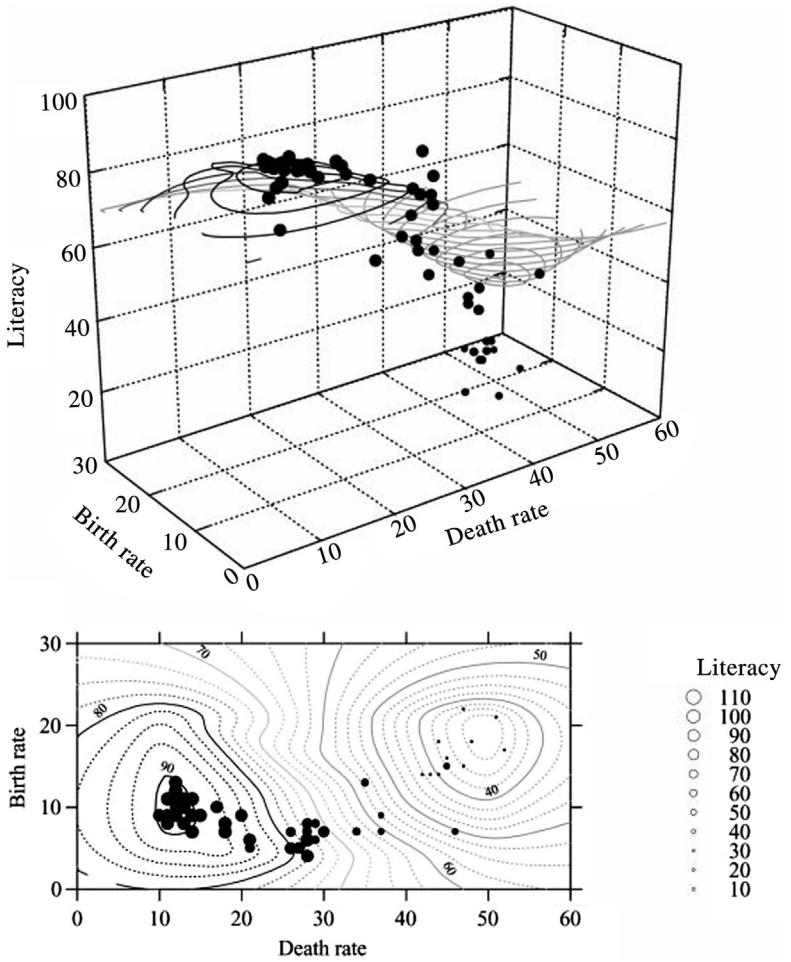

Figure 10 Birth rate and death rate vs literacy for selected countries.

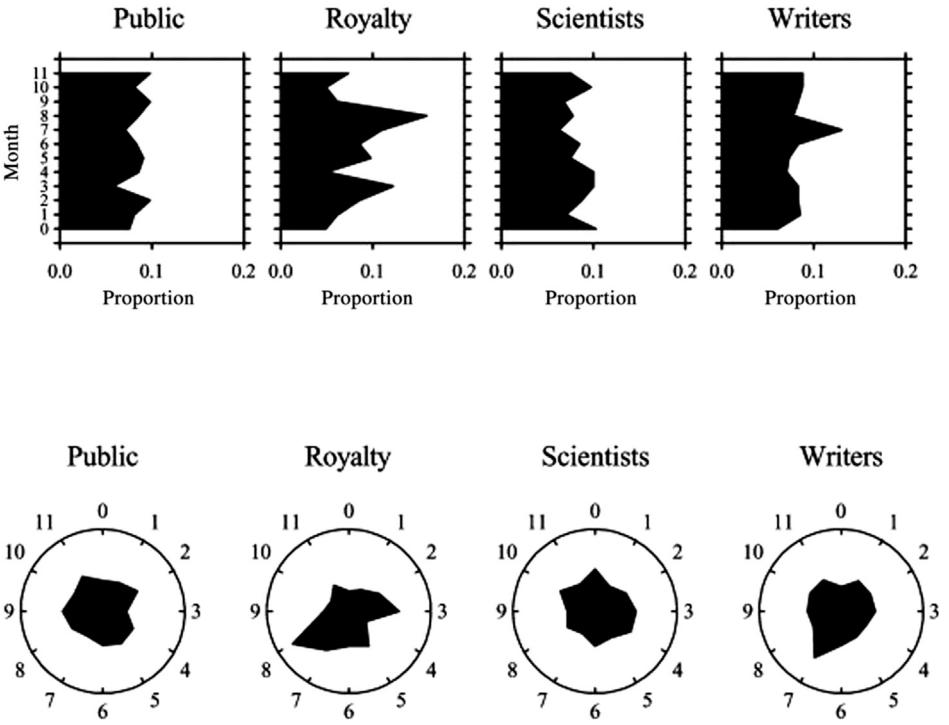

Figure 11 Difference between death month and birthday by occupational category.

determine the spacing of the partners categories, we have scant justification for using these lines. Second, the collapsing introduces a symbol choice problem. It is difficult to find symbols that are easily distinguishable for more than a few categories. The symbols collide, as well, at the upper end of the horizontal scales.

The lower plot introduces an even less appealing alternative. Clustered bar charts are used widely but they have several defects. First, it is difficult to discern separate patterns for the categories. One needs to focus on one set of bars to do this, but there is visual interference from the other bars in each cluster. Second, these bars can become quite thin with more than a few categories. Decoding this graphic is problematic.

Figure 10 shows a triple crossing of continuous variables. The data are from the UN databank used in Figure 1. The additional variable is literacy (percentage of the population that can read). We ask whether literacy is related to birth and death rates taken together.

The upper plot is a 3D frame that includes a contoured regression surface and a scatterplot. The contours (as opposed to a wire-frame or rendered surface) help viewers discern literacy levels in different elevations of the plot and keep the surface from hiding the data. Coloring the contours improves the coherence of the surface.

The lower plot is a 2D frame that includes the same contoured surface plus a scatterplot of symbols whose size is proportional to literacy. This plot makes it easier to decode specific *xyz* triplets.

Some critics tend to eschew 3D surface plots, but they have their uses. As Wilkinson (1999c) argues, surfaces elicit a holistic impression of a function. They are less useful for decoding individual values, however. Perhaps the best compromise is to present both displays simultaneously to provide different structural views. Another alternative is to categorize one or more of the variables and display the results in a paneled plot such as those in Figure 9.

Coordinates

We do not ordinarily think of charts being embedded in different coordinate systems, except for certain scientific graphics such as polar plots. Some chart types, such as radar charts, pie charts, and horizontal bar charts are nothing more than popular charts in unusual coordinate systems. Wilkinson (1999c) provides numerous examples.

Figure 11 shows two different graphics in different coordinate systems, based on data from Andrews and Herzberg (1985), contributed by C. O'Brien. The data are months since their last birthday during which members of different social groups died. The upper plot shows a paneled horizontal area chart produced by rotating a vertical area chart 90 degrees (exchanging x and y coordinates in this case). This graphic reminds us that horizontal bar charts are a coordinate transformation of a vertical bar chart, not a new chart type.

If it makes sense to consider a death one month before a birthday as being more similar to a death one month after, than to one six months after, then we might want to plot these variables in polar coordinates. The lower plot shows a profile plot of the same results.

Figure 12 contains a triangular coordinate plot. Mosteller and Tukey (1968), Bishop et al. (1975), and Wainer (1997) show examples of this type of plot. The data are from the UN dataset used in Figure 1. The variables represent expenditures in US dollars per person for health, education, and the military.

The outside tick marks signify how to read the scales. Grid lines collinear with ticks demarcate values on the three scales. Iraq and Pakistan, for example, allocated most to military spending and least to education and health. Costa Rica spent more than 40% on education, almost 50% on health, and less than 10% on the military. While triangular coordinates may be used most frequently in analysis of mixture models in

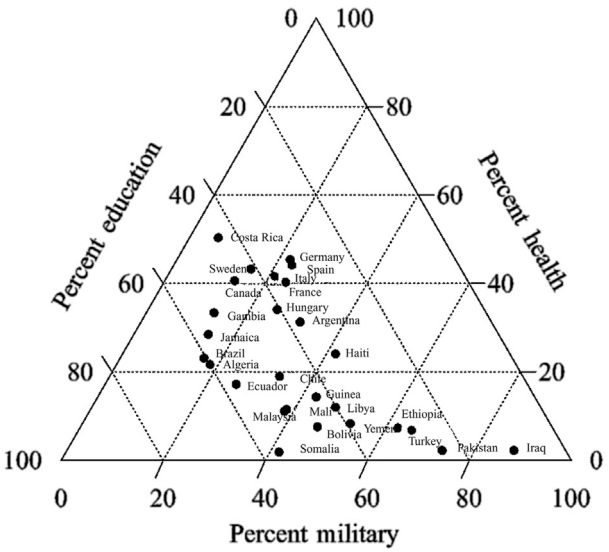

Figure 12 Percentage per-capita spending on education, military, and health.

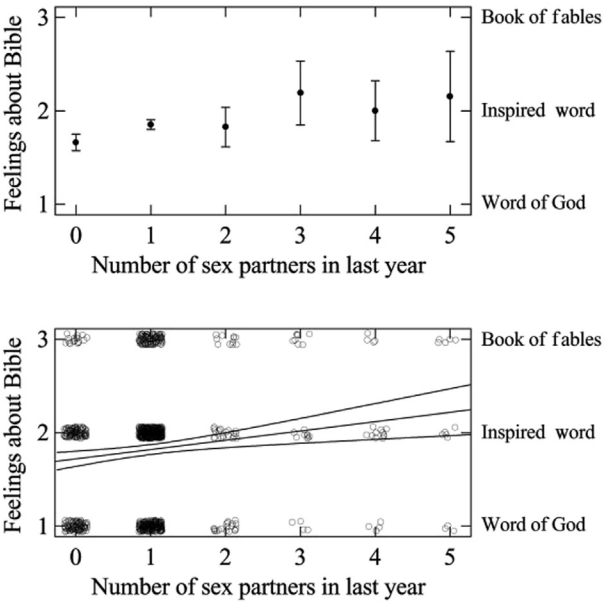

Figure 13 Sex partners vs feelings about the Bible (GSS survey responses).

industrial experiments, they can also be useful for social science models involving tradeoffs or compensatory mechanisms.

Uncertainty

There are two chief methods for representing error in a graphic. The first is sharp: error bounds are represented by clear edges, points, or lines. The second is fuzzy: error is represented by blurring an estimate.

Figure 13 shows two sharp ways of representing error in a graphic. The data are from the 1993 General Social Survey used in Figure 3. The additional variable is feelings about the Bible, as measured by the response to the question. 'Which of these statements comes closest to describing your feelings about the Bible?' The responses coded are 1 (Word of God), 2 (Inspired word), and 3 (Book of fables). We assume that the dependent variable is continuous (biblical absolutism vs relativism?) even though the responses are integers.

The upper plot shows error bars representing 95% confidence intervals on the means by category. Error bars are also used to represent one standard deviation or one standard error. It is important to make clear in accompanying titles or notes, which type is used. If the spacing of values on the independent variable is meaningful, then we can fit a linear regression model to our data. We shall further assume the number of sex partners to represent values of a continuous variable (promiscuity?). Assuming the assumptions for inference using a t distribution are appropriate, we can fit continuous confidence bounds on

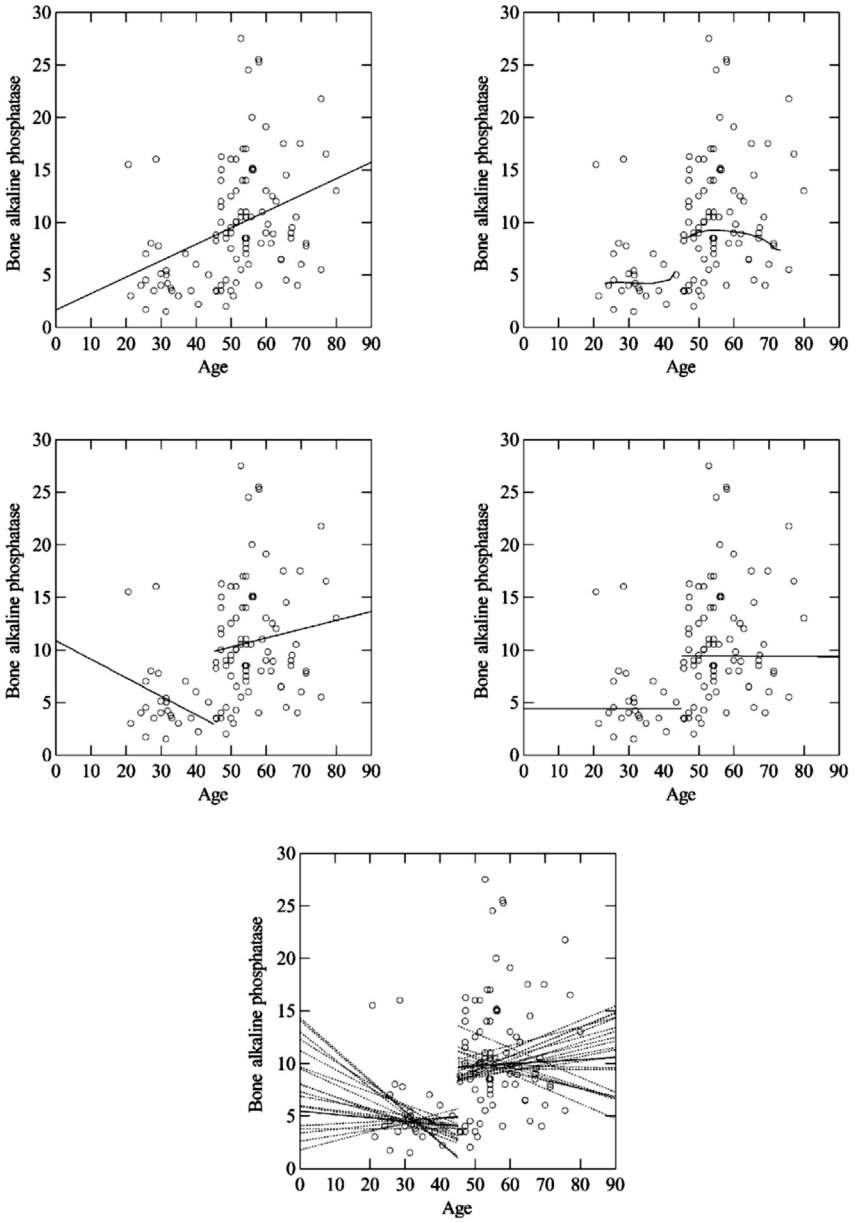

Figure 14 Bootstrapped data from Gonnelli, S., Cepollaro, C., Montagnani, A., Monaci, G., Campagna, M.S., Franci, M.B., Gennari, C., 1996. Bone alkaline phosphatase measured with a new immunoradiometric assay in patients with metabolic bone diseases. European Journal of Clinical Investigation 26, 391–396.

the regression line as in the lower panel of the figure. This plot uses jittering to reveal the concentration of data points where they are tied.

Figure 14 shows one way to represent error by fuzziness. The data are from Gonnelli et al. (1996). They represent concentration of bone alkaline phosphatase (BAP) in a sample of women of different ages. The authors fit a linear regression (shown in the upper left panel) to argue that BAP levels increase with age.

A modal regression in the upper right panel indicates that there is a discontinuity in this relationship at age 45 years or so, corresponding most likely to the onset of menopause. Accordingly, we fit separate linear regressions to the two subgroups split at age 45 years (third panel). These regressions appear to be sensitive to outliers, however, so we fit linear models using robust regression with t weighting (fourth panel). This fit indicates that a plausible model for predicting BAP from age involves level differences but no slope differences.

It is not easy to compute confidence intervals on the robust linear fits, so we resort to bootstrapping to provide an estimate of error. The bottom panel shows the result of 20 bootstrap robust fits displayed as faint dashed lines. The non-overlapping envelopes of the fits indicates that our level-change model is reasonable.

Conclusion

Figure 15 shows the principal result of a survey by Cleveland (1984) on the use of graphics in scientific articles. The horizontal axis represents the percentage of total page area devoted to graphs in 47 articles sampled from the 1980–81 volumes of selected journals. As Cleveland noted, the substantial differences among disciplines remain when one examines number of graphs per article instead of page area.

In Cleveland's survey, social science journals ranked lowest in the use of graphics. We might speculate that this is a legacy of the aversion toward graphics among eighteenth century social scientists that Beniger and Robyn (1978) describe. Whatever the reason, there is little justification for this to continue. Software packages such as S-Plus, SAS, SPSS, and SYSTAT (used for the graphics in this article) now offer statistical graphics and publishing standards such as Adobe PDF format now make it possible to display graphics as easily as text.

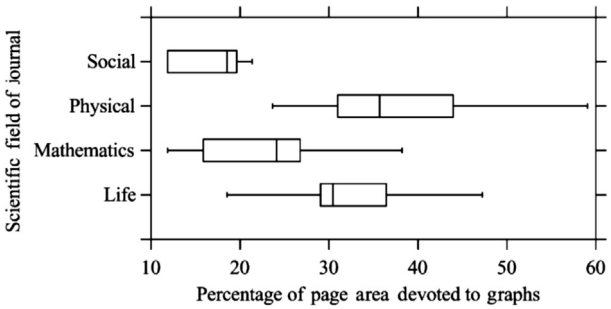

Figure 15 Use of graphics in journal articles (data from Cleveland, W.S., 1984. Graphs in scientific publications. The American Statistician 38, 19–26).

See also: Bayesian Graphical Models and Networks; Graphical Markov Models: Overview.

Bibliography

Andrews, D.F., Herzberg, A.M., 1985. Data: A Collection of Problems from Many Fields for the Student and Research Worker. Springer-Verlag, New York.
Becker, R.A., Cleveland, W.S., Shyu, M.J., 1996. The design and control of Trellis display. Journal of Computational and Graphical Statistics 5, 123–155.
Beniger, J.R., Robyn, D.L., 1978. Quantitative graphics in statistics: A brief history. The American Statistician 32, 1–11.
Bertin, J., 1967. Semiologie Graphique. Paris: Editions Gauthier-Villars (English translation Berg W J, Wainer H 1983 Semiology of GraphicsUniversity of Wisconsin Press, Madison, WI).
Bertin, J., 1981. Graphics and Graphic Information-Processing (English translation Berg W J, Scott P. Walter de Gruyter, New York).
Bishop, Y.M., Fienberg, S.E., Holland, P.W., 1975. Categorical Data Analysis. MIT Press, Cambridge, MA.
Bureau of the Census, 1986. State and Metropolitan Area Data Book. US Government Printing Office, Washington, DC.
Chambers, J.M., Cleveland, W.S., Kleiner, B., Tukey, P.A., 1983. Graphical Methods for Data Analysis. Wadsworth, Monterey, CA.
Cleveland, W.S., Devlin, S., 1988. Locally weighted regression analysis by local fitting. Journal of the American Statistical Association 83, 596–640.
Cleveland, W.S., McGill, R., 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American Statistical Association 79, 531–554.
Cleveland, W.S., 1984. Graphs in scientific publications. The American Statistician 38, 19–26.
Cleveland, W.S., 1993. A model for studying display methods of statistical graphics (with discussion). Journal of Computational and Graphical Statistics 2, 323–343.
Cleveland, W.S., 1994. The Elements of Graphing Data, revised ed. Hobart Press, Summit, NJ.
Cleveland, W.S., 1995. Visualizing Data. Hobart Press, Summit, NJ.
Collins, B.M., 1993. Data visualization – has it all been seen before? In: Earnshaw, R.A., Watson, D. (Eds.), Animation and Scientific Visualization: Tools and Applications. Academic Press, New York, pp. 3–28.
Davis, J.A., Smith, T.W., Marsden, P.V., 1993. The General Social Survey. National Opinion Research Center, Chicago.
Fayyad, Usama M., Wierse, Andreas, Grinstein, Georges G., 2002. Information visualization in data mining and knowledge discovery. Morgan Kaufmann, San Francisco, CA.
Fienberg, S., 1979. Graphical methods in statistics. The American Statistician 33, 165–178.
Funkhouser, H.G., 1936. Note on a tenth century graph. Osiris 1, 260–262.
Funkhouser, H.G., 1937. Historical development of the graphical representation of statistical data. Osiris 3, 269–404.
Gonnelli, S., Cepollaro, C., Montagnani, A., Monaci, G., Campagna, M.S., Franci, M.B., Gennari, C., 1996. Bone alkaline phosphatase measured with a new immunoradiometric assay in patients with metabolic bone diseases. European Journal of Clinical Investigation 26, 391–396.
Griethe, Henning, Schumann, Heidrun, 2006. The Visualization of Uncertain Data: Methods and Problems. Proceedings of SimVis, 143–156.
Herdeg, W., 1981. Graphis Diagrams: The Graphic Visualization of Abstract Data. Graphis Press, Zürich, Switzerland.
Holmes, N., 1991. Designer's Guide to Creating Charts and Diagrams. Watson-Guptill Publications, New York.
Kosslyn, S.M., 1994. Elements of Graph Design. W. H. Freeman, New York.
McGill, R., Tukey, J.W., Larsen, W.A., 1978. Variations of box plots. The American Statistician 32, 12–16.
Mosteller, F., Tukey, J.W., 1968. Data analysis, including statistics. In: Lindzey, G., Aronson, E. (Eds.), Handbook of Social Psychology, second ed. Addison-Wesley, Reading, MA, pp. 80–203.
Pinker, S., 1997. How the Mind Works. Norton, New York.
Robinson, A.H., 1982. Early Thematic Mapping in the History of Cartography. University of Chicago Press, Chicago.
Salanti, Georgia, Ades, A.E., Ioannidis, John, 2011. Graphical Methods and Numerical Summaries for Presenting Results from Multiple-Treatment Meta-Analysis: An Overview and Tutorial. Journal of clinical epidemiology 64 (2), 163–171.

Sarkar, Deepayan, 2008. Lattice: multivariate data visualization with R. Springer, Seattle, WA.
Scott, D.W., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley, New York.
Series, Statistical Concepts, 2002. Describing Data: Statistical and Graphical Methods. Radiology 225, 622–628.
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.
Simkin, D., Hastie, R., 1987. An information processing analysis of graph perception. Journal of the American Statistical Association 82, 454–465.
Stigler, S., 1983. The History of Statistics. Harvard University Press, Cambridge, MA.
Tilling, L., 1975. Early experimental graphics. The British Journal for the History of Science 8, 193–213.
Tufte, E.R., 1983. The Visual Display of Quantitative Information. Graphics Press, Cheshire, CT.
Tufte, E.R., 1990. Envisioning Data. Graphics Press, Cheshire, CT.
Tufte, E.R., 1997. Visual Explanations. Graphics Press, Cheshire, CT.
Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, MA.
Wainer, H., Spence, I., 1997. Who was Playfair? Chance 10, 35–37.
Wainer, H., Velleman, P.F., 2001. Statistical graphs: Mapping the pathways of science. The Annual Review of Psychology 52.
Wainer, H., 1997. Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot. Springer-Verlag, New York.
Wilkinson, L., Blank, G., Gruber, C., 1996. Desktop Data Analysis with SYSTAT. Prentice-Hall, Upper Saddle River, NJ.
Wilkinson, L., 1999a. Dot plots. The American Statistician 53, 276–281.
Wilkinson, L., 1999b. SYSTAT. SPSS, Chicago.
Wilkinson, L., 1999c. The Grammar of Graphics. Springer-Verlag, New York.

Green Jobs

Olivier Deschenes, University of California Santa Barbara, Santa Barbara, CA, USA; IZA, Bonn, Germany; and National Bureau of Economic Research (NBER), Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In recent years, the prospect of 'green jobs' or 'green growth' policies have become increasingly prominent, proposed to solve both the environmental challenges associated with global climate change and the persistent unemployment problems observed in many industrialized countries. This article reviews the evidence from the simulation-based studies that assess the impact of green policies on employment. Due to the sizable heterogeneity in the scope and assumptions of the existing studies, it is difficult to make a definitive conclusion about their likely impact. More careful and detailed empirical research is needed to assess the job creation potential of green job policies.

Introduction

In recent years the prospect of 'green jobs' or 'green growth' policies have become increasingly prominent, proposed to solve both the environmental challenges associated with global climate change and the persistent unemployment problems observed in many industrialized countries. In the United States, such proposals began in the Green Jobs Act of 2007 (as a component of the 2007 Energy Independence and Security Act) and continued in the American Recovery and Reinvestment Act of 2009. While the Green Jobs Act was never enacted as law, the American Recovery and Reinvestment Act of 2009 procured funding for green initiatives, primarily for energy efficiency and renewable energy investments, and for related Research and Development.

At the same time, similar policies were also proposed and implemented in many other countries, in many instances as part of broader stimulus packages. According to HSBC (2009), $430 billion (USD) in fiscal stimulus have been allocated globally to 'climate change investment themes,' with China and the United States being the two largest contributors. This corresponds to roughly 15% of global stimulus expenditures, with targeted investments primarily in low-emission transportation, grid expansion and efficiency, building energy efficiency, and renewable energy (especially in the United States for the latter).

The motivation for such targeted national investments is threefold: First, they reflect a desire by governments to stimulate the production and development of less carbon-intensive energy sources in order to reduce emissions of carbon. Second, green policies clearly have an explicit objective of boosting employment. For example, employment in sectors such as construction is expected to grow as the demand for energy-efficient homes and buildings is expected to increase as a result of these policies. Third, green policies are implemented with the expectation that they will lead to spillovers, economies of scale, and comparative advantage in the development of green technologies, and in the production of green goods and services. Consequently, these policies should be evaluated with three corresponding criteria: (1) whether they lead (or will lead) to improvement in environmental quality; (2) whether they create long-term employment and earnings growth for workers in green sectors; and (3) whether they contribute to a sustained competitive edge through learning-by-doing effects in the relevant industries. Another important criterion for evaluating targeted green policies is whether they achieve their goals in an efficient manner, relative to alternative approaches.

While many green policies are still ongoing, it is crucial to begin examining the evidence regarding the impacts of such massive investments. This short article will concentrate on the labor market impacts of green job policies. From a broader perspective, the question of whether or how green policies affect labor markets is hardly a new question. There is a long standing debate on the connection between environmental regulations and firm competitiveness, and ultimately employment. A major concern is that stricter environmental regulations lead to displacement of jobs, especially in energy-intensive and globally trading industries, although the empirical evidence regarding the magnitude of those displacement effects is mixed (see e.g., Jaffe et al., 1995; Berman and Bui, 2001; Greenstone, 2002; Walker, 2011).

The goal of this short article is threefold. First, I describe a conceptual framework to evaluate the possible employment effects of green policies. Second, I discuss the various definitional and measurement issues related to green jobs, and present descriptive evidence for the United States, based on recent data from the U.S. Bureau of Labor Statistics (BLS). Third, I review the evidence from the existing studies that attempt to assess the impact of green policies on employment. (Many of these studies also discuss possible green skills shortages associated with green growth policies. I do not address these considerations in this short article. See Bowen (2012) for a discussion of these issues.) These studies are almost entirely based on simulation models as opposed to direct empirical analysis of labor market data. Further, the majority of these studies have been published in reports as opposed to peer-reviewed journals. (An Econlit search on 'green job' reveals only two published articles in English peer-reviewed journals (accessed 25.03.13).) As a result there are substantial research gaps that need to be addressed in the coming years.

I draw two main conclusions from this exercise. First, my descriptive analysis of the U.S. BLS data on green jobs highlights that green jobs currently represent a small share of overall employment in the United States. For instance, the service sector, which accounts for half of total hours worked in the U.S.

economy is not a green sector: only 2% of employment in that sector is labeled 'green' according to the U.S. BLS definition. Published estimates from other countries also indicate that green employment only accounts for 2–3% of aggregate employment. Further, between 2001 and 2010, the share of hours worked in the 'greenest' sectors in the U.S. economy only accounted for one-third of total hours worked, and if anything the trend over time shows a slightly declining share. This very simple descriptive evidence appears to invalidate the premise that green sectors grow faster than 'brown' sectors. This premise is one of the primary justifications motivating the use of green job policies as engines of job growth and as long-term drivers of sustainable economic growth (UNEP, 2008). Thus more research needs to evaluate these claims with more rigorous empirical methods. Future research should also compare the net employment gain associated with a specific green job policy with the improvements in environmental quality it generates.

Second, due to the sizable heterogeneity in the scope and assumptions made in the existing simulation studies of the labor market impacts of green policies, it is difficult to make a definitive conclusion. This issue is only magnified by the fact that almost none of the existing studies have gone through the formal peer-review process associated with journal publications. Such process will help create and preserve continuity and progress across studies, and help streamline and homogenize the assumptions made in the simulation models. It will also better disseminate the results of this research to the economics profession. This will hopefully fuel an improved research agenda in the coming years.

In summary, much remains to be learned regarding the effect of green policies on employment creation, and whether these policies can be long-term determinants of sustainable economic growth. Further, there is little quantitative evidence on the environmental benefits associated with green job policies, and whether these benefits are sufficient to justify the costs of the green policy investments. There is a glaring lack of quality real-world microeconomic data on green jobs, and characteristics of 'green' workers. The preparation and dissemination of such data should be a priority for the relevant statistical agencies of countries investing in green growth programs.

Conceptual Framework

Targeted green policies are typically implemented through a combination of tax adjustments and publicly funded investments projects. (While obviously designed and implemented with green environmental objectives in mind, carbon-pricing policies such Europe's EU-ETS and California's AB32 are not considered as green job policies in this article.) Examples include tax credits or subsidized interest rates on loans for improving energy efficiency in buildings, various taxes, and standards promoting energy efficiency in vehicles, etc. Similarly, public funds are provided as R&D grants and subsidies for the development and generation of solar and other renewable energy, and for training programs targeted at workers in green sectors.

To begin conceptualizing the impact of such policies on labor markets, it is useful to consider the simple neoclassical labor demand model with an employment subsidy. For example, the U.S. Tax Relief and Job Creation Act of 2010 offers tax credits to manufacturers of energy-efficient dishwashers, clothes washing machines, and refrigerators, effectively reducing the costs of hiring for these firms. An employment subsidy will shift the labor demand curve in the sectors where the subsidy is granted, thus increasing both employment and wages in that sector in the short run. This effect is essentially the short-term effect described in the framework below.

Another component of the conceptual framework is the time horizon. Most existing studies consider a dynamic framework for evaluating the impacts of green policies over a varying time horizon. Sustained green policies will have effects on labor markets in the short, medium, and long run. The nature of these effects will differ over time and characterize the transition toward a less carbon-intensive economy. To illustrate this, consider this framework adapted in part from Fankhauser et al. (2008):

- *Short-term effects*: Employment grows in subsidized industries (e.g., green construction sector, low-carbon vehicle sector) and/or in sectors that produce goods and services that are complements with the green policy objectives. Employment may decline in sectors producing goods and services that are substitutes to the green policy objectives (e.g., low-efficiency vehicles). Labor and especially capital are less mobile in the short run so the possibilities of adjustment are more limited.
- *Medium-term effects*: Employment changes as the impact of the green policy diffuses through all sectors in the economy. Jobs are created and lost along the carbon-intensity dimension and due to complementarities and substitutability across sectors. (For example, a reduction in the coal-powered electricity generation could lead to reduced employment in mining and transportation industries.) Adjustments to the capital stocks begin to reflect the new economic environment.
- *Long-term effects*: Capital stocks are freely mobile. Innovation and the development of new technologies create new opportunities for investment. Net effect of targeted policies on employment attenuates over time as labor fully reallocates across sectors.

While useful as a starting point to evaluate green growth policies, this stylized framework ignores important issues that require attention. A central question not addressed is whether green jobs would emerge in the absence of governmental intervention. In other words, would consumer demand for green goods and services alone drive employment growth in these sectors? The literature remains mostly silent on this question. Another issue is whether green growth objectives could be achieved through price incentives as opposed to 'industrial policy' push, specifically by increasing the price of fossil-based energy (for example, through a carbon tax).

Regarding the transition from the short-term to the long-term effects of green policies, the main point is that the employment effects are likely to vary across the various time horizons. Employment losses are likely to be significant in carbon-intensive sectors in the short run, however, in the long run with full-labor mobility across sectors, the overall

employment effects may be null or even positive. An important question is what research methods can tell us about the short-term versus the long-term effects of green policies, and whether these policies lead to growth. While it is possible to identify the short-term effects using empirical methods, long-term effects where intersector reallocations are important will primarily be informed from model-based computable general equilibrium (CGE) analysis, although some empirical studies of long-term effects exist. (See, for example, Greenstone, 2002; Popp, 2002; Walker, 2011.)

Another important question ignored to date in the literature is whether green policy investments lead to actual improvements in environmental quality. Every job includes components of 'green' and 'brown' activity. And so the net effect of a green policy on environmental quality will depend on the types of jobs that are created and destroyed by the implementation of the policy, and on the 'green' and 'brown' nature of these jobs. A simple approach to answer this question would be to use data on carbon intensity or emissions (e.g., from life-cycle assessment models) associated with various jobs and calculate the net effect on aggregate carbon emissions of the net employment effect of a policy. As the share of green to brown jobs increases, we should expect a reduction in carbon emissions in the economy. This reduction can be monetized using estimates of the social cost of carbon (see e.g., Greenstone et al., 2011) and then compared to the size of the public investments in green job policies.

Finally, while most studies of green policies' impacts focus on employment changes as the indicator of costs or benefits associated with a policy, transitional costs associated with reallocating production and workers across sectors are likely to be significant. For example, Walker (2011, 2013) finds significant transitional costs for workers displaced by regulation implemented under the 1990 Clean Air Act Amendments. His analysis shows that 10 years after the change in regulation, employment in the impacted sectors remained 15% lower than before the change in regulation. Further, Walker also reports persistent earnings losses for workers displaced by the regulation, up to 5 years after the regulatory change, for a total earnings loss equivalent to 20% of their preregulatory earnings. Given that most studies of green job policies are based on CGE-type models that generally assume immediate reallocation, these potentially important transition costs have not been accounted for properly in the modeling literature.

Definitions and Measurement of Green Jobs

While broadly used as a notion of employment in environmentally conscious or clean energy sectors, there is no universally accepted definition of what constitutes a 'green job.' Most definitions used by statistical agencies and researchers reflect common themes, namely jobs related to the production and distribution of 'clean' energy, and to the production of environmental goods and services. Some researchers also define green jobs exclusively as the jobs created by green growth policies. Others, like the United Nations Environmental Program (UNEP), also add qualifications that green jobs must also offer adequate wages, and support workers' rights and safe working conditions.

The Green Goods and Services Survey (GGS) from the U.S. BLS considers two definitions of green jobs:

1. "Jobs in businesses that produce goods and provide services that benefit the environment or conserve natural resources," (Output-based definition).
2. "Jobs in which workers' duties involve making their establishment's production processes more environmentally friendly or use fewer natural resources," (Process-based definition).

For both definitions, the focus is on goods and services and/or production processes that relate to (1) Energy from renewable sources, (2) Energy efficiency, (3) Pollution reduction and removal, greenhouse gas reduction, and recycling and reuse, (4) Natural resource conservation, and (5) Environmental compliance, education and training, and public awareness.

Similar definitions are used in other countries. In Germany, the definition refers to "employees who produce environmental goods and services, or employees involved in environment-related activities." At the same time, other countries such as France do not have an official definition of green jobs, even though the country's statistical agency reports green employment counts. See the Supplement to the 2012 OECD Employment Outlook (OECD, 2012) for a presentation of country-specific definitions of green jobs. Based on these definitions, examples of green jobs include: manufacturing, installation, and maintenance of solar panels; production and services in sustainable agriculture, enforcement and monitoring of environmental regulations, collection and processing of recyclables, etc.

Combined with the relatively recent interest in green sector jobs, the absence of a clear definition explains why there is less data available on green jobs, except for aggregate counts. This lack of reliable microdata on characteristics of individuals working in green sectors, in turn, explains why there are no proper microeconometric studies of the employment effects of green policies.

Even after a definition of green jobs is accepted, its credible measurement remains challenging. To illustrate the problem, consider the output-based measure employed by the U.S. BLS: "jobs in businesses that produce goods and provide services that benefit the environment or conserve natural resources." The advantage of this definition is that it allows the agency to select fine industry classifications in which some or all of the employment can be considered green. The U.S. BLS then proceeds in two steps. First, the BLS has identified 333 detailed industries (6-digit NAICS, out of a possible 1192) in which green goods and services are potentially produced, after consulting with industry groups, governmental agencies, etc. (http://www.bls.gov/green/final_green_def_8242010_pub.pdf (accessed 25.03.13).) Perhaps this restriction is necessary from a practical viewpoint, but it is also a limitation of the BLS data: green goods and services are also likely to be produced in the other 859 industries.

Clearly, both green and nongreen employment will occur within and across establishments in these 333 industries. Consider the 'Commercial building construction' industry (NAICS 236220). Employment in the construction of Leadership in Energy and Environmental Design certified buildings clearly represents green employment, while employment in the

Table 1 Number and percentage of green jobs, various countries

	Number of green jobs	*Percent of total jobs 'green'*	*Data year*
Austria	162 986	3.99%	2008
Finland	5 888	0.24%	2009
France	–	1.60%	2010
Germany	–	4.50%	2006
Hungary	200 000	4.18%	2020
Japan	1 400 000	2.19%	2008
South Korea	604 400	2.56%	2008
Portugal	–	0.40%	2008
Spain	531 000	2.62%	2008
Switzerland	–	4.50%	2009
United States	3 100 000	2.43%	2010

Notes: Adapted from OECD, 2012. What green growth means for workers and labour market policies: an initial assessment. In: OECD, OECD Employment Outlook 2012. OECD Publishing. http://dx.doi.org/10.1787/empl_outlook-2012-5-en.

construction of standard buildings would not. To resolve this, the U.S. BLS conducts an establishment-level survey to learn about green and nongreen employment activities within establishments in a second step. (http://www.bls.gov/respondents/ggs/forms.htm (accessed 25.03.13).) Establishments participating in the survey are asked to provide a share of revenue for their green goods or services. In case of establishments with both green and nongreen employment, this revenue share was used to obtain the establishment's green employment. (In case of nonprofit or government agencies, the survey directly asks to provide a share of employment involved in the production of green goods or green services.)

Another limitation of the U.S. BLS approach of measuring the importance of green jobs is that it is based on the 2007 NAICS classifications (at least the existing 2010 U.S. BLS data is – future releases will be based on the 2012 NAICS). Any creation of new green industrial sectors post-2007 will be ignored and thus lead to undercounting. See Pollack (2012) for a more in depth discussion of the limitations of the U.S. BLS green job definition.

With these limitations in mind, it is nevertheless useful to look at the available data. **Table 1**, adapted from OECD (2012), reports the number (and percentage) of green jobs for 11 countries where such data are available. Despite slight variations in the definitions and measurement, the percentage of green jobs (relative to total employment) is fairly similar across countries, ranging from 2% to 3%. Finland has the lowest share (0.24%) and France and Germany have the highest (4.50%). The main point of **Table 1** is that green jobs account for a relatively small share of total employment in all countries where the data is available.

Table 2 examines the industrial composition, carbon intensity, and labor market importance of green jobs in the United States. The analysis is based on data from the 2010 U.S. BLS GGS, from a 2010 report by the U.S. Department of Commerce on sectoral carbon intensity, and from the individual-level data on hours worked and earnings from the 2010 American Community Survey (ACS). (Ruggles et al., 2010). The variables are linked across 27 broad industry classifications that can be linked across the different data bases. (Unfortunately, some sectors like government and public administration cannot be consistently linked across the data bases and therefore excluded from the analysis. The 27 industries considered account for 90% of the U.S. workforce.) Column (1) reports the 2008 direct carbon intensity, measured in million metric tons (mmt) of CO_2 per billion dollars (2000 USD). (The industry-specific carbon intensity measures are derived by applying CO_2 conversion factors to industry-specific energy use data. See Department of Commerce (2010) for more details.) This measure of carbon intensity provides an alternative metric of how green each sector is. In principle, data on carbon intensity could be used as a metric to compute changes in environmental quality associated with changes in employment across industrial sectors. Not surprisingly, sectors like transportation, primary metals, and nonmetallic mineral products have high carbon intensity, exceeding 1 mmt per billion (USD) of output. Across all industries, the hours-weighted average carbon intensity is 0.23. This reflects the fact that more than 50% of total hours worked in the United States in 2010 were in the relatively low-carbon intensity service sector.

Column (2) displays the BLS measure of percent green jobs (relative to total employment) in each industry. For example, these data indicate that 20% of employment in Forestry and Logging is 'green,' while essentially no jobs in Financial Activities, Education and Health Services, and Leisure and Hospitality are green. Across all industries, the hours-weighted percent of green jobs is about 3%. Further, the correlation between carbon intensity and percentage of green jobs across industries is relatively weak. It is evident that the sectors with the highest concentration of green jobs are the primary and manufacturing sectors, the sectors with the largest carbon intensity. At the same time, the concentration of green jobs in the service sector – one that accounts for roughly half of all hours worked in the U.S. economy – is relatively smaller. Nevertheless, a key point of column (2) is that it clearly establishes, at least in the United States, that green jobs are not only jobs associated with the specific sectors such as renewable energy industry (e.g., solar panel production and installation), but rather are represented in most of the major industrial sectors.

Columns (3) and (4) report the share of total hours worked and share of total wage bill (total hours worked times average hourly earnings) by industry, relative to the aggregate U.S. workforce. For both hour shares and wage bill shares, the data show a negative correlation, that is, the industries with higher concentrations of green jobs typically have lower shares of hours worked and wage bill earned. For example, primary metals has 18.5% of its employment falling under the GGS definition of the U.S. BLS, yet the sector accounts for only 0.5% of total hours worked and total wage bill in the United States in 2010. Similarly, 19.8% of employment in Forestry and Logging is green, but the sector accounts for only 0.1% of total hours worked and total wage bill in the United States in 2010. On the other hand, only 0.2% of employment in Education and Health Services contributes to the production of green goods and services, but it alone accounts for more than 25% of all hours and earnings in the U.S. economy.

Figure 1 displays the trends in the share of hours worked and total wage bill in the United States from 2001 to 2010 for

Table 2 Industrial distribution and characteristics of private sector green jobs in United States

	Carbon intensity	*Percentage of industry employment 'green' (2010)*	*Percentage of total hours worked (2010)*	*Percentage of total wage bill (2010)*
Industry				
Crop production	0.09	6.9	0.7	0.4
Animal production	0.25	3.7	0.4	0.2
Forestry and logging	0.05	19.8	0.1	0.1
Mining	0.09	–	0.8	0.9
Utilities	–	11.9	1.1	1.6
Construction	0.16	6.8	6.2	5.9
Manufacturing	–	–	12.4	14.2
Textile product mills	0.14	10.7	0.5	0.3
Furniture and wood products	0.20	7.4	0.8	0.6
Paper	0.61	7.8	0.9	0.9
Petroleum and coal products	1.21	2.8	2.6	3.4
Chemicals	0.62	3.7	1.2	1.7
Plastics and rubber products	0.28	0.9	0.6	0.7
Nonmetallic mineral products	1.61	9.6	0.4	0.4
Primary metal	1.05	18.5	0.5	0.5
Fabricated metal products	0.19	2.6	0.7	0.7
Machinery	0.12	7.1	1.6	1.9
Appliance and electronic products	0.08	11.1	2.8	3.5
Transportation equipment	0.10	2.9	1.9	2.4
Trade (Wholesale and retail)	0.16	1.0	15.6	12.8
Transportation and warehousing	1.22	6.2	4.8	4.5
Services	–	–	48.9	53.1
Information	0.03	1.4	0.2	0.3
Financial Activities	0.07	0.0	8.6	11.8
Prof., scientific and Tech. Services	0.01	4.7	6.7	10.4
Administrative and waste services	0.01	4.3	4.1	3.1
Education and health services	0.16	0.2	25.8	26.2
Leisure and hospitality	0.27	0.2	9.0	5.2
Other services	0.02	1.3	4.2	3.2

the top 50% 'greenest' industries in Table 2. (The industries with above median percent green jobs are: crop production, forestry and logging, utilities, construction, textile products, furniture and wood products, paper, nonmetallic mineral products, primary metals, machinery, appliances and electronics, transportation and warehousing, and professional, scientific, and technical services.) The trends clearly show little progress in employment and wages in these sectors during the 2000s, with both series indicating relatively constant green shares, at about 30% of aggregate hours and 32% of the aggregate wage bill. These shares also appear to dip slightly during the 2008–09 Great Recession period. This very simple descriptive evidence appears to invalidate the premise that green sectors grow faster than brown sectors. This premise underlies most of the argument for why green job policies can deliver job growth and can be long-term drivers of sustainable economic growth (UNEP, 2008). Clearly, more detailed empirical microeconomic research is needed to offer better empirical evidence and testing of the central hypotheses in green growth theory.

Review of the Existing Employment Studies

A significant number of studies, primarily based on computer simulations (CGE-type modeling) and input–output table analysis, have been conducted to evaluate the job growth potential of variously defined green job policies in Europe

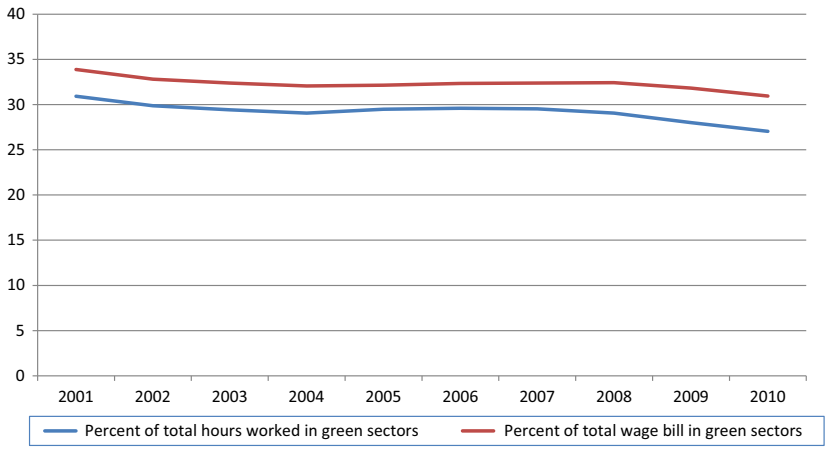

Figure 1 Percentage of aggregate hours worked and total wage bill in U.S. BLS 'greenest' industries in the United States, 2001–10. Notes: Author's calculations based on 2001–2010 American Community Surveys.

and in the United States (see e.g., GHK, 2009; GCN, 2010; Bowen, 2012 for extensive reviews of this literature). Notably, few of these studies have been published in peer-reviewed academic journals. Rather, they appear as reports or working papers. For example, none of 29 studies reviewed in GHK 2009 are published in peer-reviewed journals.

Given the considerable heterogeneity in the scope and assumptions of these studies and the absence of a formal peer-review process for most existing studies, it is very difficult to draw definitive and broad conclusions. (GHK (2009) note that the employment change estimates of green policy studies are not comparable due to many factors, including difference in the consideration of gross or net employment effects, differing assumptions concerning economic growth and the effect of existing business as usual policies, etc.) In addition, studies vary greatly in the scope of policies evaluated. Some consider single renewable energy policies while others evaluate multiple policies in combination, including renewable energy targets, carbon-pricing policies, etc. Others ignore the indirect effect of green policies on job destruction in brown sectors, or other general equilibrium effects. It should be noted that these limitations are in addition those embodied in CGE analysis by construction. In the context of green job policies, an especially important limitation of CGE models is that they typically assume inelastic labor supply and free labor mobility across sectors. This implies that all adjustments occur quickly, and only through wage changes, ruling out frictional unemployment. (See Wing (2004) for a clear a discussion of the main limitations of CGE models.) It is unclear how much the combination of such assumptions leads to systematic biases in the CGE analysis.

As stated earlier, the limited availability of labor microdata with information on green jobs explains why there is no direct 'program evaluation' evidence on green job policies. In fact, there are very few empirical studies altogether. This is in contrast with the literature focusing on employment effects of carbon-pricing policies in the United States, where empirical studies and modeling studies have been successfully conducted (see e.g., Deschenes, 2012; Kahn and Mansur, 2010; Morgenstern et al., 2008).

These important limitations notwithstanding, the consensus appears to be one where green policies can deliver employment growth. In his review, Bowen (2012) also notes this and writes that taken together the evidence suggests "that climate-change policies in general and renewable energy policies in particular can generate considerable extra employment," also adding that "environmental policies will lead to substantial job creation only if other inefficiencies—including those of labor markets—are tackled."

In my view, there are too many limitations associated with the existing studies to reach a firm consensus. Conclusions such as the one presented in Bowen (2012) may turn out to be correct, but there is not enough empirical evidence to support them. Two key limitations need to be addressed before a more definitive conclusion can be reached. First, statistical agencies need to accelerate the data collection process on green jobs and provide the appropriate microdata sets to the researchers. Researchers need access to large data bases with detailed worker level information on demographic attributes, job attributes (including measures of jobs' green content), hours worked, wages, etc. This process is slowly beginning in the United States, but more countries need to follow suit.

Second, more careful and detailed empirical work is required to evaluate the job creation potential of green policies, and to inform the simulation models employed in making such assessments. In particular, more labor economists need to begin producing research in this area. The rigorous evaluation of publicly funded job training programs in the United States and Europe done by labor economists has been important in improving the design of such programs, and in developing methodological tools for evaluation (see e.g., Ashenfelter, 1978; Heckman et al., 1999). If green policies are to be successful in solving unemployment and environmental problems, the same standards of evaluation should apply to them. These empirical studies should aim to credibly document the short-run effects of green growth policies on employment, wages, unemployment duration, as well as the

transitional costs associated with labor and production reallocation.

Final Remarks and Conclusions

Massive public investments have been made and proposed to promote green growth and green jobs, yet little is known about their impacts on labor market outcomes. This reflects the lack of readily available microdata, the focus on long-term simulation analysis as opposed to program evaluation-type empirical studies, and the lack of homogeneity in the study settings and policy scenarios considered. Moving beyond labor market impacts, the improvements in environmental quality associated with green policies also need to be carefully evaluated. There remains a large agenda for future research in all of these areas.

Acknowledgment

I thank Nico Pestel, Jing Lin, and Corey White for excellent suggestions and research assistance. I also thank IZA for research support.

See also: Climate Change and Global Poverty; Climate Change: Economics; Climate Change; Environmental Economics; Environmental Sociology; Sustainable Development: An Economic Perspective; Sustainable Production and Consumption.

Bibliography

Ashenfelter, O., 1978. Estimating the effect of training programs on earnings. The Review of Economics and Statistics 60 (1), 47–57.

Berman, E., Bui, L., 2001. Environmental regulation and labor demand: evidence from the South Coast air Basin. Journal of Public Economics 79, 265–295.

Bowen, A., 2012. "Green" growth, "Green" jobs and labor markets. World Bank. Policy Research Working Paper, 5990.

Deschenes, O., 2012. Climate policy and labor markets. In: Fullerton, Don, Wolfram, Catherine (Eds.), The Design and Implementation of U.S. Climate Policy. University of Chicago Press, 2012.

Fankhauser, S., Sehlleier, F., Stern, N., 2008. Climate change, innovation and jobs. Climate Policy 8, 421–429.

GCN, 2010. Low-Carbon Jobs in an Interconnected World. Global Carbon Network Discussion Paper No. 3 and Literature Review, March.

GHK, 2009. The Impacts of Climate Change on European Employment and Skills in the Short to Medium-Term. A Review of the Literature, Final Report to the European Commission Directorate for Employment, Social Affairs and Inclusion Restructuring Forum, vol. 2, May.

Greenstone, M., 2002. The impacts of environmental regulations on industrial activity: evidence from the 1970 and 1977 clean air act amendments and the census of manufactures. Journal of Political Economy 110 (6), 1175–1219.

Greenstone, M., Kopits, E., Wolverton, M., 23 March, 2011. Estimating the Social Cost of Carbon for Use in U.S. Federal Rulemakings: A Summary and Interpretation. MIT Department of Economics Working Paper No. 11-04. Available at SSRN: http://ssrn.com/abstract=1793366. or. http://dx.doi.org/10.2139/ssrn.1793366.

Heckman, J., LaLonde, R., Smith, J., 1999. The economics and econometrics of active labor market programs. Handbook of Labor Economics 3, 1865–2097.

HSBC, 2009. A Climate for Recovery: The Colour of Stimulus Goes Green. Website: www.research.hsbc.com (accessed 21.03.13.).

Ho, M., Morgenstern, R., Shih, J., 2008. Impact of Carbon Price Policies on U.S. Industry. Resources for the Future Discussion Paper 08–37.

Jaffe, A., Peterson, S., Portney, P., Stavins, R., 1995. Environmental regulation and the competitiveness of U.S. Manufacturing: what does the evidence tell us? Journal of Economic Literature 33 (1), 132–163.

Kahn, M.E., Mansur, E.T., 2010. How Do Energy Prices, and Labor and Environmental Regulations Affect Local Manufacturing Employment Dynamics? A Regression Discontinuity Approach. NBER Working Paper No. 16538.

OECD, 2012. What green growth means for workers and labour market policies: an initial assessment. In: OECD (Ed.), OECD Employment Outlook 2012. OECD Publishing. http://dx.doi.org/10.1787/empl_outlook-2012-5-en.

Pollack, E., 9 October 2012. Counting up to green: Assessing the green economy and its implications for growth and equity. Economic Policy Institute Briefing Paper No. 349.

Popp, D., 2002. Induced innovation and energy prices. American Economic Review 92, 160–180.

Ruggles, S., Alexander, T., Genadek, K., Goeken, R., Schroeder, M., Sobek, M., 2010. Integrated Public Use Microdata Series: Version 5.0 [Machine-readable Database]. University of Minnesota, Minneapolis.

UNEP, 2008. Green Jobs: Towards Decent Work in a Sustainable Low Carbon World. United Nations Environment Program.

U.S. Department of Commerce, 2010. U.S. Carbon Dioxide Emissions and Intensities Over Time: A Detailed Accounting of Industries, Government and Households (accessed 26.03.13.). http://www.esa.doc.gov/Reports/u.s.-carbon-dioxide.

Walker, R., 2011. Environmental regulation and labor reallocation. American Economic Review: Papers and Proceedings 101 (2).

Walker, R., 2013. The Transitional Costs of Sectoral Reallocation: Evidence from the Clean Air Act and the Workforce. Mimeographed, UC Berkeley.

Wing, Ian Sue, 2004. Computable general equilibrium models and their use in economy-wide policy analysis. In: MIT. Joint Program on the Science and Policy of Global Change 6.

Green Parties

Neil Carter, University of York, Heslington, York, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E. Bomberg, volume 9, pp. 6386–6389, © 2001, Elsevier Ltd.

Abstract

Green parties are well established in many liberal democracies. The green party family is relatively homogeneous with parties sharing common ideologies, similar origins, party structures, and constituencies of support. Electoral performance is variable – there are around 15 established successful green parties but success elsewhere is limited – and shaped by various structural factors and country-specific political opportunities. Electoral success has brought new pressures on green parties to compromise their radical policies and practices. Green parties have joined or supported government coalitions involving parties of left and right, achieving some policy impact and demonstrating that they are responsible coalition partners.

Introduction

Since their emergence in the late 1970s and early 1980s, green parties have become established political players in many industrialized democracies. Hence, today, green parties are represented in around 30 national parliaments (mostly in Western Europe), in the European Parliament, and in many more subnational and local assemblies. The German Greens are the largest and probably the most successful green party: Die Grünen first entered the federal parliament in 1983 and were sufficiently established by 1998 to form a governing red–green coalition with the Social Democrats, with Joschka Fischer assuming the roles of Foreign Minister and Vice-Chancellor for 7 years. Greens have also joined coalition governments in other European countries, including Belgium, the Czech Republic, Finland, France, Italy, and Luxembourg. In 1999, German Green Michaele Schreyer was the first green appointed to be a member of the European Commission, the European Union's (EU) powerful executive.

The steady electoral advance of the greens in advanced industrialized liberal democracies contrasts with near invisibility in the Global South, although green parties have recently secured some limited electoral successes in South America, and a smattering of greens have been elected in countries as diverse as Mongolia, Kenya, Vanuatu, Congo, and Mauritius. Moreover, electoral success brought strategic dilemmas for the greens, as they struggled to maintain their alternative 'green' credentials while entering parliaments and joining national governments. This article provides an overview of the origins, development, and current status of green parties. It examines various factors that have shaped green party electoral performance and assesses the contribution made by green parties within government coalitions and the impact of that experience on the parties themselves.

Emergence

The emergence of green parties needs to be understood in the context of the broader socioeconomic transformation of advanced industrial societies over the last 50 years that has contributed to major changes in political behavior associated with the rise of a 'new politics.' The 'old politics' consisted of frozen party systems and stable political cleavages based on class, religion, and region; political discourses were dominated by material issues such as economic growth, stable prices, unemployment, public order, national security, and the protection of traditional lifestyles. For younger, better-educated citizens in particular, the prosperity and rise in living standards enjoyed in the post-War era formed the backdrop for the emergence of new quality of life issues such as enhanced political participation, gender and racial equality, and environmental protection. Many citizens felt that these concerns were inadequately addressed by existing parliamentary structures. A proliferation of extra-parliamentary, nonpartisan citizens' movements – the new social movements (NSMs) – emerged in the 1960s and 1970s to protest about the issues of ecology, peace, nuclear power, and women's rights (Dalton, 2014).

Various explanations have been offered for this mushrooming of alternative values, issues, and forms of political behavior. Some observers identified a 'silent revolution' under way involving a shift from materialist to postmaterialist values (Inglehart, 1977). Others identified the emergence of a 'new middle class' of educated, professional service workers who were disproportionately represented in the NSMs (Cotgrove, 1982). Certainly, the massive expansion of higher education in the post-War era must have played a part. Whatever the cause, the changes in the values and forms of political activity associated with the new politics did not fit easily into the traditional left–right divide so that mainstream parties have often struggled to respond adequately to them. Yet, despite the predictions of some scholars, the new politics has not replaced the old politics, but it has disrupted and supplemented it.

However, a danger in looking to broad structural developments to explain the rise of green parties is that one critical factor in the rise of environmentalism gets ignored: the objective state of the environment. In the 1960s and 1970s, growing concern among the public was steadily heightened by continued environmental degradation and the increased media attention given to environmental issues. Toxic chemical spills in major rivers, air pollution in cities, and oil slicks in pristine coastal waters further increased public anxiety. Meanwhile, the accelerated construction of nuclear plants in the wake of the energy crisis, as well as the increased deployment of nuclear weapons, awakened or strengthened fears about the safety of

nuclear power, the problems of disposing of nuclear waste, and fears of nuclear annihilation. Subsequently, the emergence of new global problems, such as acid rain, ozone depletion, and climate change, reinforced these concerns. Thus, in searching for sophisticated social science explanations for the emergence of green parties, we should not sacrifice the most straightforward interpretation that in the 'risk society' (Beck, 1992), green politics may be fueled by a specific concern about the state of the environment, as much as it is a reflection of postmaterial values or class politics.

Whatever the explanation(s) for growing environmental concern, it is of great significance that the NSMs served as the antecedents for green parties because it meant that most green parties started out ideologically focused on new politics issues and committed to political practices that profoundly challenged conventional party politics.

The Green Party Family

Green parties make up one of only two distinct new party families to have emerged in the post-War era, the other being the radical right (Ennser, 2012). The green party family is unusually homogeneous as measured against all four common approaches to the identification of party families: party name; transnational links; origins and social base; and party policy and ideology (Mair and Mudde, 1998: p. 214–215).

First, the name of almost every party that might fall within the green party family contains the word 'Green' or a variation on 'Ecologist' (one exception is the Hungarian 'Politics Can Be Different' formed in 2009). Second, most parties belong to transnational green party federations such as the Global Greens. Membership of these transnational federations indicates that green parties see themselves as sharing common aims and principles with sister parties elsewhere. Indeed, the European Green Party even contests elections to the European Parliament on a common manifesto in every member state.

Third, the vast majority of green parties share a common ideology and program. One of the distinctive ideas underpinning green party ideology is the belief that the planet imposes limits to economic and population growth: the Earth's resources are finite and without a fundamental transformation in our economic, political, and social structures, we will exceed those limits relatively soon (Carter, 2007; Dobson, 2006). Greens offer a wholesale critique of advanced industrial society for its dependency on the inexorable pursuit of economic growth, globalization, mass consumerism, and worship of material goods. Instead, greens advocate a 'sustainable society' characterized by simpler lifestyles, less resource depletion, social equality, and local action to solve global problems. Thus, the core ideological distinctiveness of green parties rests on the priority they give to environmental issues, which underpins their rejection of the dominant growth-oriented economic paradigm and their opposition to nuclear power.

Despite this commitment to radical ecological thinking, most green parties, especially those emerging from NSM roots, were determined not to become a single-issue party. To be sure, most parties have always had factions that prioritize pure environmental policies, reject leftist ideology as too pro-industrial, and try to establish clear distance between their views and those of mainstream parties, left or right (Spoon, 2009; Talshir, 2002). Some small parties with a primarily ecological focus persist on the margins of green politics such as the Dutch De Groenen and German Ecological Democratic Party. But green party programs were typically modeled around the foundational 'four pillars' of the German Greens: ecological sustainability, grassroots democracy, social justice, and nonviolence. They adopted a range of liberal positions on social issues such as immigration, women's rights, and lifestyle choices; supported pacifist foreign policies; and took leftist positions on many core economic issues. Although frequently proclaiming that they were 'neither right nor left, but out in front,' in practice, green parties were widely regarded as left wing (often prompting the criticism from right-wing opponents that green parties resemble a watermelon – green on the outside but red on the inside). Indeed, Kitschelt (1988) characterized green parties as part of a wider left-libertarian party grouping, although this generally helpful label misses what is distinctive about green parties – their ecological orientation.

Over time, there has been considerable convergence, particularly among European green parties as a result of having formed a European green party and a common platform for European Parliament elections – gradually, those green parties opposed to EU membership as in Sweden or critical of deeper European integration as in England and Ireland have shifted to a pro-EU position. One comparative study of green party manifestos and expert survey data finds that green programs continue to prioritize environmental protection (albeit less so than in the 1980s), but they have developed a broader left-wing agenda combining radical libertarian positions with more traditional 'socialist' concerns such as expanding the welfare state and the education sector, while some of the earlier green-party priorities, such as antimilitarism and decentralization, have been downplayed, but not discarded (Carter, 2013).

The left-wing label does not so readily fit green parties outside Western Europe. One obvious anomaly is the Latvian Green Party, which has regularly contested elections with the right-wing nationalist-agrarian Farmers' Union in an alliance that has employed nationalist rhetoric and accepted significant corporate funding from the Ventspils Oil Transit Group, whose namesake is one of the oligarchs that Latvian anticorruption reformists oppose (Auers and Galbreith, 2009). The Mexican Greens (Ecological Green Party of Mexico, PVEM) have achieved some electoral success, but following their support for capital punishment, the European Greens withdrew recognition of PVEM as a legitimate green party. Subsequently, a series of homophobic statements by PVEM candidates seems to confirm that this party does not fit comfortably into the green party family.

If we exclude these maverick parties, we see that even some mainstream green parties outside Western Europe have positioned themselves differently from their longer-established sister parties. The Czech Greens achieved electoral success in 2006 by campaigning as a nonnationalistic, pro-European, center-right party, subsequently forming a right-wing coalition government with the Civic Democrats and Christian Democrats. Although the program of the Hungarian 'Politics Can be Different' is inspired by the original four pillars of the German Greens, it too is liberal on economic issues. This liberal positioning is shaped by the political context of former Soviet

bloc countries, where terms such as 'left wing' and 'socialist' often have reactionary, antidemocratic connotations that represent the very opposite of the radical and progressive politics to which green parties generally aspire. Nevertheless, the Czech, Hungarian, and also the Estonian green parties share the emphasis on environmental protection and libertarianism that is characteristic of the wider green party family. They clearly belong to the green party family, but they bring an element of heterogeneity to it.

Fourth, most green parties in Western Europe and Oceania share common origins in the NSM milieu of the 1970s and 1980s (Frankland et al., 2008; Richardson and Rootes, 1995). The broad coalition of environmental, peace, and leftist groups that formed the antinuclear movements was the catalyst for green party formation in Germany, France, Luxembourg, and Finland, and in Austria and Sweden, green parties emerged from referendum campaigns against nuclear power. Even where a green party had different origins – the first green party in the United Kingdom (originally called 'People,' later the 'Ecology Party') was formed by a group of conservative environmentalists – over time, they have taken on the characteristics of continental sister parties.

NSM activists have left a strong imprint on green parties: 'process' was critical, so the way greens expressed and pursued their beliefs was as important as the sustainable society that they were advocating. Above all, greens emphasize 'grassroots democracy': a decentralized, direct democracy within a radically participatory society. Policy decisions should be made at the lowest possible level and matched with varying local requirements. Diversity is not only tolerated but celebrated. Thus, the 'amateur-activist' organizational structure of most green parties was based on grassroots democracy, weak central organization, and a range of institutional mechanisms designed to resist the professionalization of the party, including unsalaried party officials, enforced job rotation, and collective leadership (Frankland et al., 2008). This crusade to be, in the words of Petra Kelly, an 'antiparty party' was an earnest attempt to resist the oligarchic tendencies that historically have emerged in most radical democratic parties. But green parties soon came under pressure to discard these unconventional practices, resulting in often bitter internal disagreements between social movement activists committed to the grassroots model and cadres who wanted a more 'professional' organization capable of achieving sustained electoral success. This tension was characterized by the conflicts within the German Greens between the 'fundamentalists' (*fundis*) and the 'realists' (*realos*), but it took similar forms elsewhere.

Two to three decades after their formation, green parties are undoubtedly more conventional than when they started out. Organizationally, they are much more professional and media savvy, illustrated by the willingness of some parties to sacrifice collective leadership principles in favor of a single leader (as in England and Wales and in Ireland) or to focus electoral campaigns on popular individuals such as Joschka Fischer. Yet, despite the many pressures to conform, all green parties still retain several features of the amateur-activist model. Although some of the more radical ideas such as the midterm rotation of MPs proved unworkable and were quickly dispensed with, several successful green parties (e.g., in Germany, New Zealand, and Sweden) retain a form of collective leadership with joint male and female spokespersons, while women play a prominent role and make up around half of most national green parliamentary groups. While green parties do outwardly appear more professional, the underlying party organization retains many of the amateur-activist features: rather like the mythological *centaur*, the upper part of the body has increasingly embraced a professional electoral logic, while the lower half retains its amateur-activist form (Rihoux and Frankland, 2008: p. 281).

The sociological profile of green voters and activists seems everywhere to be dominated still by members of the 'new middle class.' Dolezal's (2010) study of 12 European countries, using European Social Surveys conducted between 2002 and 2007, found that compared with supporters of other parties, greens were younger and better educated, less attached to the dominant Christian churches, and more likely to be female and an urban dweller. Dolezal identifies three categories of 'new middle class' – managers, technical experts, and 'sociocultural specialists' – and finds much higher green support among the latter. Sociocultural specialists, working outside the hierarchical structures of conventional organizations and possessing a strong orientation toward their profession and their clients, are most likely to hold libertarian values (although they are not necessarily economically left), which may draw them to the greens. Dolezal claims that there are important structural components to green voting – young, highly educated, sociocultural specialists or students, urban, and not very religious – that are connected to environmental, libertarian, and pro-immigration attitudes. Recent studies of green voters in Australia (Miragliotta, 2013), France (Beaudonnet and Vasilopoulos, 2013), and Germany (Rüdig, 2012) broadly support these demographic and attitudinal findings, although in Germany there is a 'graying' of the green voter, as habitual individual green voters have grown older.

Overall, the green party family is unusually homogeneous with most parties sharing common ideologies and programs, similar origins, party structures, and constituencies of support. However, this package has delivered variable electoral success.

Electoral Fortunes

Green parties have slowly become a stable electoral force as minor (rarely polling over 10% at national level) but serious actors in many advanced liberal democracies. They are most strongly established in Western Europe, where there is a core group of electorally 'successful' green parties. Following Müller-Rommel (1998) in defining a party as electorally successful if it contests at least two national elections and polls (on average) at least 3% of the national vote, and adding the reasonable requirement that a party is represented in a national parliament, then greens have become consistently successful in nine Western Europe countries over two decades or more: Austria, Belgium (Ecolo and Groen!), Finland, France, Germany, Luxembourg, The Netherlands, Switzerland, and Sweden. This list has been supplemented in recent years by successful green parties further afield in Australia, Canada, New Zealand, Hungary, Brazil, and Columbia (**Table 1**).

Table 1 Electoral performance of green parties in 18 countries

	First elected MP	Number of MPs[a]	Average % vote at national elections		
			1980s	1990s	2000s
Austria	1986	24	4.0	6.0	10.3
Belgium[b]	1981	12	5.9	10.9	8.0
Finland	1983	10	2.8	6.9	7.9
France	1997	17	1.1	7.2	4.5
Germany	1983	63	5.1	6.3	9.2
Ireland	1989	0	1.0	2.1	3.4
Italy	1987	0	2.5	2.7	2.1
Luxembourg	1984	6	6.8	8.9	11.7
The Netherlands	1990	4	–	5.0	5.1
Sweden	1988	24	2.9	4.3	6.0
Switzerland	1979	15	7.1	5.5	8.5
The United Kingdom	2010	1	0.3	0.5	0.9
The Czech Republic	2006	0	–	1.1	3.6
Estonia	2007	0	–	–	5.5
Hungary	2010	5	–	–	6.4
Australia	2010	1	–	2.3	7.9
Canada	1996	1	–	0.3	4.9
New Zealand	1996	14	–	6.1	7.5

[a]At most recent election before 18 September 2014.
[b]Combined results of Ecolo and Agalev/Groen!.

Elsewhere, green parties have met limited success. Green parties have a national presence characterized by limited or sporadic electoral successes in Britain, Greece, Ireland, Italy, Norway, Spain, the Czech Republic, Estonia, and Hungary. In the United States, no green has been elected to Congress or as governor, with just a handful of representatives elected to state legislatures and in mayoral contests, although the long-standing consumer campaigner Ralph Nader secured almost three million votes when standing on a green ticket in the 2000 Presidential election. Beyond Brazil and Columbia, green party success in the Global South is very limited.

Explaining Electoral Performance

The varied electoral fortunes of green parties can be explained as a product of wider structural factors, as well as more specific opportunity structures in which these parties operate. The macroeconomic environment in which parties develop is a good starting point for explaining green party success. It is widely recognized that green parties fare better in advanced industrialized countries with high levels of economic development and a strong welfare state, especially where long-standing material affluence seems to have contributed to the rise of postmaterial values. This broad structural difference helps explain the contrast between the development of green parties in the industrialized North with their struggle to emerge and survive in the Global South. But these features are less able to explain differences among developed industrialized states: Why do green parties fare better in Germany and New Zealand than in Norway and the United States? It is helpful to turn to more specific political opportunity structures to account for such variation.

Electoral laws are often cited (especially by green parties themselves) as critical in determining success. The experience of countries such as Belgium, Finland, Germany, and Switzerland suggests that green parties do better in electoral systems based on proportional representation (PR), where there are fewer barriers to the emergence of new parties. By contrast, the plurality system used in national elections in Canada, the United States, and the United Kingdom has made it difficult for any small party to succeed unless they are regionally concentrated such as the Bloc Québécois or the Scottish National Party. The experience of New Zealand, where the green party struggled to get MPs elected under the old plurality system and then flourished when PR was introduced in the mid-1990s provides a good illustration of the key role played by electoral rules.

PR is, however, no guarantee of electoral success. Green parties have failed to make any significant electoral breakthrough in several countries with PR systems, including Denmark, Norway, Greece, and Spain. While the weakness of green parties in southern Europe may be explained partly by lower levels of economic development and the presence of fewer postmaterialists, Denmark and Norway are affluent, developed economies with high levels of public environmental concern. Conversely, the presence of greens in the French National Assembly regularly since 1997 shows that the plurality system does not pose an insuperable barrier, although this success would probably not have happened without electoral pacts with the Socialists. Recent elections have also seen individual green MPs elected in Canada and the United Kingdom. Nevertheless, on balance, a facilitative electoral system will normally be a necessary, but not a sufficient, condition for green party success.

A federal structure offers green parties more access points to articulate their demands at subnational level and more opportunities to make an electoral breakthrough, where they can gain experience and publicity. Several of the most successful green parties are in federal states (Australia, Austria, Belgium, Germany, Switzerland), although, again, the weakness of the U.S. greens shows that federal structures are no guarantee of success, while green parties have also been successful in unitary states such as Finland, New Zealand, and Sweden.

Green parties have performed comparatively well in elections to the European Parliament, where low turnouts and widespread protest voting can often reward smaller parties. Moreover, there is evidence suggesting that in elections for this supranational body the greens benefit from the public recognition that many environmental problems require international solutions. Thus, 46 green Members of European Parliament (MEPs) from 12 countries were elected in 2009, and in combination with the European Free Alliance (EFA) (a disparate group of regionalists and nationalists) they formed the fourth largest political group in the European Parliament with 55 MEPs. The number of green MEPs elected slipped slightly to 38 in 2014, but they now represented 14 countries as part of the greens–EFA group in the 2014–19 parliament.

Like other small parties, the greens benefit from those political systems that provide state funding to parties securing a certain share of the vote or number of seats as in Germany, Australia, and elsewhere. However, dependence on these funds

means that when their electoral success wanes, so does their funding: for example, after the Irish Greens dropped below 2% of the vote in 2011, the party lost all its state funding.

In addition to the impact of electoral rules and institutional features of political systems, party competition also plays an important role in green electoral success. Much depends on how green parties position themselves relative to other parties in a country's party system and how those parties respond to the presence of a green party. Most green parties have used their left-libertarian programs to locate themselves to the left of the established center-left social democratic or labor party. When many mainstream center-left parties shifted to the center ground through the 1980s and 1990s, green parties tried to colonize this vacant political territory, successfully so in West Germany, Austria, and Belgium. Sometimes, however, this political territory was already occupied, which helps explain the weakness of green parties in both Norway and Denmark, where the long-established 'left-libertarian' Socialist Left Party and Socialist People's Party, respectively, had already secured the loyalties of potential new middle-class green voters and willingly embraced a green agenda. Green parties have also encountered competition on the left from radical socialist and ex-Communist parties, such as the Swedish Left Party, the Danish Red–Green Alliance, and Syriza in Greece, who have been prepared to adopt some environmental policies. Moreover, in Scandinavia, most parties have embraced a progressive environmental agenda, so, for example, social liberal parties such as the Danish Radical Party and the Swedish Center Party have competed strongly for the environmental vote, while in Norway, the liberal Venstre and the Conservative have notably 'green' agendas. The emergence of the Swiss Green Liberals, a social liberal party outside the green party family, has harmed the electoral performance of the long-established Swiss Greens.

Green party fortunes are also shaped by the extent to which the mainstream center-left and center-right parties have been willing to share or poach the ideas of greens and adopt pro-environment policies in deed or at least in word. Typically, mainstream center-left parties have been more willing to adopt 'green' policies, presumably because they are more threatened by the rise of a left-libertarian green party than their conservative or Christian Democrat rivals. However, contemporary conservative leaders such as Chirac, Sarkozy, Merkel, and Cameron have all made attempts to present their parties as progressive and serious about the environment when they identify rising public concern about the environment that might manifest itself in support for the green party.

Episodic factors, such as environmental disasters (e.g., the Chernobyl and Fukushima nuclear calamities or major oil tanker spillages), or the rise of new controversies, such as genetically modified foods, can be the catalyst for upsurges in environmental concern. Here, green parties also face competition from other green actors, notably environmental groups or protest movements, who absorb environmental demands without themselves seeking government office. Thus, environmental awareness in the United Kingdom and the United States is not noticeably lower than in countries with strong green parties, but this concern is expressed in swelling membership of environmental groups rather than by voting for green parties.

Finally, whatever the structural constraints within which they must perform, green parties shape their own fortunes by their (in)ability to develop and articulate a coherent party platform, run a professional electoral campaign, elect popular politicians, and avoid internal combustion through infighting and personality squabbles. In short, agency matters.

Taken together, these contending explanations provide a plausible, if nuanced, explanation of green party development and change.

Greens in Government

In several countries, electoral success has brought new governmental responsibilities. The Finnish Green League in 1995 was the first green party to join a national coalition government, and by 1999, green parties were also in center-left governments in France, Germany, and Italy. Subsequently, they have provided support for minority center-left governments in Sweden, New Zealand, and Australia (and a Scottish Nationalist government) and joined rainbow coalitions in Belgium, Finland, and Luxembourg and center-right governments in the Czech Republic, Finland, and Ireland.

As the junior coalition partner, green parties have only been able to exercise limited influence over government programs (Evrard, 2012; Lees, 2000; Poguntke, 2002). Perhaps, the most notable achievements saw the German and Belgian greens push through policies for the long-term phasing out of nuclear power, although the Finnish and French greens failed in their attempts to alter nuclear policy. Green coalition governments saw the introduction of new eco-taxes and stronger conservation policies. Greens have also supported a range of liberal reforms that included giving greater protection to asylum seekers, new rights to illegal immigrants, and legal status to gay and lesbian couples. During the late 1990s, when several greens held the post of environment minister, they were able to exercise considerable influence over environmental policy within the Council of Ministers (Bomberg and Carter, 2006).

By entering parliament, the nature of internal green party debate shifted from whether they should govern to how they should govern (Poguntke, 2002). In practice, green ministers have worked civilly with their cabinet colleagues from other parties, made tough decisions, and generally demonstrated that green politicians could be trusted to hold office. But this new experience inevitably threw up many challenging dilemmas for the greens, particularly over the compromises that the realities of government required them to make. Most famously, German foreign minister, Joschka Fischer, found himself supporting military policies such as NATO air strikes on Serbia, which flouted the cherished green principle of pacifism – a decision that led to him being flour-bombed by activists. Sometimes, greens decided that it was better to resign than to compromise on a core green issue: the Finnish Greens resigned from government in 2002, when the parliament supported the construction of a new nuclear power station and the New Zealand Greens withdrew support for the labor government over its support for genetically modified crops. By contrast, the sudden resignation in April 2014 of the two EELV ministers from President Hollande's embattled Socialist government, apparently in response to its austerity measures, looked more opportunistic than principled. But most green parties have stayed the course and have not defected from the government

coalition, and where they have (e.g., Ireland), the election was fast approaching (Little, 2014). Thus, if their overall policy impact in government has been quite modest, in the longer term, green parties have demonstrated that they are a serious political force and reliable coalition partners.

The electoral impact of green involvement in government is quite varied (Little, 2014; Rüdig, 2006). Green parties in Finland, France, and Germany have not been badly punished electorally for their actions in office, and the Swedish and New Zealand greens largely escaped electoral censure for their support of minority center-left governments. But the Czech, Irish, and Dutch greens have been badly damaged. They must hope to emulate Belgium, where both the Flemish-speaking Groen! (then called Agalev) and French-speaking Ecolo were severely punished at the ballot box in 2003 after four rather ignominious years in government – but they have since recovered to previous levels of support. Either way, the internal party postmortems following their departure from office nowhere seems to have resulted in a green party deciding to return permanently to the role of protest party. Instead, the dominant response has been to make (limited) programmatic and institutional changes to enable the Greens to return to power when a suitable opportunity arises. It seems that green parties have a strong office-seeking mentality, perhaps best illustrated by the willingness of several parties to join coalitions dominated by right-wing parties.

Conclusion

The green party family is remarkably homogeneous. Most green parties continue to emphasize environmental issues while embracing a broad agenda that is left-wing on economic and social justice issues and libertarian on a range of social issues. The environmental imperative and the libertarian agenda also characterize green parties from Central and Eastern Europe, although they tend to be centrist on the left/right dimension. The recent emergence of the center-right Swiss Green Liberals is the first party of its kind in Western Europe; if replicated elsewhere, it would represent an interesting challenge to the homogeneity of the green party family.

Green parties are now not only a familiar part of the political landscape, but due to their capacity to shape party competition, government formation, and government policy, in several countries they have joined the political establishment. Although most green parties are left-wing, they are more pragmatic than in their early days (Blühdorn, 2004): having tasted power (or seen sister parties do so), greens want to make a difference, hence their willingness to work with both left-wing and right-wing parties.

Green parties face several challenges. First, how to carve out a policy niche as established parties become wiser to green demands and as ecological concerns appear more mainstream. The failure of any green party to break through the 10% barrier on a regular basis means that they are likely to remain, at best, a junior coalition partner in national government – so in seeking to influence policy, they are engaged in a careful balancing act to avoid alienating their core constituency of left-libertarian, postmaterialist voters. Second, how to take green ideas beyond the confines of rich industrialized states into Eastern Europe and the developing world, where green parties remain marginal and environmental problems acute. Third, how to ensure that the broader role of green parties – as consciousness raisers and agitators – is not sacrificed on the altar of electoral success.

See also: Environmental Movements; Environmental Policy: Protection and Regulation; Environmental Sociology; Parliamentary Government; Party Systems.

Bibliography

Auers, D., Galbreith, D., 2009. Green, black and brown: uncovering Latvia's environmental politics. Journal of Baltic Studies 40, 333–349.
Beaudonnet, L., Vasilopoulos, P., 2013. Green parties in hard times: the case of EELV in the 2012 French presidential election. Party Politics 20, 275–285.
Beck, U., 1992. Risk Society. Sage, London.
Blühdorn, I., 2004. New Green' pragmatism in Germany – Green politics beyond the social democratic embrace? Government and Opposition 39, 564–586.
Bomberg, E., Carter, N., 2006. The Greens in Brussels: shaping or shaped? European Journal of Political Research 45, S99–S125.
Carter, N., 2007. The Politics of the Environment: Ideas, Activism, Policy. Cambridge University Press, Cambridge.
Carter, N., 2013. Greening the mainstream. Environmental Politics 22, 73–94.
Cotgrove, S., 1982. Catastrophe or Cornucopia. Wiley, Chichester.
Dalton, R., 2014. Citizen Politics, sixth ed. C.Q. Press, Washington.
Dobson, A., 2006. Green Political Thought, fourth ed. Routledge, London.
Dolezal, M., 2010. Exploring the stabilization of a political force: the social and attitudinal basis of green parties in the age of globalisation. West European Politics 33, 534–552.
Ennser, L., 2012. The homogeneity of West European party families: the radical right in comparative perspective. Party Politics 18, 151–171.
Evrard, A., 2012. Political parties and policy change: explaining the impact of French and German Greens on energy policy. Journal of Comparative Policy Analysis: Research and Practice 14, 275–291.
Frankland, E.G., Lucardie, P., Rihoux, B. (Eds.), 2008. Green Parties in Transition: The End of Grass-roots Democracy? Ashgate, Farnham.
Inglehart, R., 1977. The Silent Revolution. Princeton University Press, Princeton.
Kitschelt, H., 1988. Left-libertarian parties: explaining innovation in competitive party systems. World Politics 40, 194–234.
Lees, C., 2000. The Red-Green Coalition in Germany. Manchester University Press, Manchester.
Little, C., 2014. Politics on the Margins of Government: A Comparative Study of Green Parties in Governing Coalitions (Ph.D. thesis). European University Institute, Florence.
Mair, P., Mudde, C., 1998. The party family and its study. Annual Review of Political Science 1, 211–229.
Miragliotta, N., 2013. The Australian Greens: carving out space in a two-party system. Environmental Politics 22, 706–727.
Müller-Rommel, F., 1998. Explaining the electoral success of green parties: a cross-national analysis. Environmental Politics 7, 145–154.
Poguntke, T., 1993. Alternative Politics: The German Green Party. Edinburgh University Press, Edinburgh.
Poguntke, T., 2002. Green parties in national governments: from protest to acquiescence. Environmental Politics 11, 133–145.
Richardson, D., Rootes, C. (Eds.), 1995. The Green Challenge. Routledge, London.
Rihoux, B., Frankland, E.G., 2008. The metamorphosis of amateur-activist newborns into professional-activist centaurs. In: Frankland, E.G., Lucardie, P., Rihoux, B. (Eds.), Green Parties in Transition: The End of Grass-roots Democracy? Ashgate, Farnham, pp. 259–287.
Rüdig, W., 2006. Is government good for Greens? Comparing the electoral effects of government participation in Western and East-Central Europe. European Journal of Political Research 45, S127–S154.
Rüdig, W., 2012. The perennial success of the German Greens. Environmental Politics 21, 108–130.
Spoon, J.-J., 2009. Holding their own: explaining the persistence of green parties in France and the UK. Party Politics 15, 615–634.
Talshir, G., 2002. The Political Ideology of Green Parties. Palgrave, Basingstoke.

Green Social Work

Lena Dominelli, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Green social work is a holistic approach to social workers' involvement in ecological and environmental issues. It combines a structural analysis centered on social institutions and social relations with the duty to respond to the welfare needs of individuals, groups, and communities and care for the environment. By critiquing neoliberal capitalist models of social development and irresponsible exploitation of the earth's resources, it argues that the intergenerational interdependency between all living things and the physical environment is crucial to developing alternative sustainable models of living and being that do not conflict with the needs of the flora and fauna that share the earth with people. Solidarity built on this interdependency guarantees people's duty to care for others and the right to be looked after by them now and into the future.

Introduction

Ecosystem problems are a recent concern in social work, despite the profession's basic commitment to the 'person-in-the-environment,' which over time became limited primarily to the immediate social environment of the family and the social relationships associated with it. Ecological social work attempted to reinstate a consideration of the impact of the physical environment into social work in the early 1990s, but it was seen as a minority interest (see Park, 1996). As climate change and other ecological disasters that have affected social relationships have entered popular discourses with increasing frequency, interest in this intellectual arena has been renewed, and ecosystem concerns are becoming a growing element within the discipline.

This article discusses green social work, which differentiates itself from ecological social work (*see* Ecological Social Work: Shifting Paradigms in Environmental Practice) on several significant dimensions, and calls for the further development of this emerging part of the profession.

Defining Green Social Work

Green social work aims to promote the well-being of people and the planet's flora, fauna, and physical environment and ensure the equitable and sustainable distribution of the earth's bounty. Its vision of a more equitable, egalitarian world is rooted in a recognition of the 'politics of practice,' whereby social workers, as part of the societies they live within, have to transcend the political limitations of a normative worldview based on the inequitable distribution of the earth's resources by arguing that although this way of organizing social relations is currently dominant, this need not and should not be the case. Indeed, they argue that social development cannot continue in this vein if the world's endeavors to eliminate poverty and develop sustainable initiatives both now and in the long term are to come to fruition. In taking this approach, green social workers ground their practice in the profession's core values of human rights and social justice; an understanding of contemporary power relations and realities; and a critical theory stance toward existing dominant paradigms of education, theory, and practice in the profession.

Dominelli (2012a: p. 25) defines green social work as:

> a form of holistic professional social work practice that focuses on the: interdependencies amongst people, the social organization of relationships between people and the flora and fauna in their physical habitats; and the interactions between socio-economic and physical environmental crises and interpersonal behaviors that undermine the well-being of human beings and planet earth. It proposes to address these issues by arguing for a profound transformation in how people conceptualize the social basis of their society, their relationships with each other, living things and the inanimate world. It does so by: tackling structural inequalities including the unequal distribution of power and resources; eliminating poverty and various 'isms'; promoting global interdependencies, solidarity and egalitarian social relations; utilizing limited natural resources such as land, air, water, energy sources and minerals for the benefit of all rather than the privileged few; and protecting the earth's flora and fauna.

Consequently, green social workers seek to transform the sociopolitical and economic forces that have degraded the quality of life of poor and marginalized populations, and they demand the policy changes and social transformations needed to make this possible. The privileging of some people at the expense of others is a barrier to the right for all living things to enjoy a healthy and sustainable environment. There are 1153 billionaires enjoying the fruits of the earth and its luxuries while 3 billion people live on less than $2 per day (Kroll and Fass, 2011). These billionaires either own or manage the multinational corporations that are making decisions that impact the livelihoods and well-being of all living things through the choices they make about development options and the exploitation of natural resources and human labor power. Challenging such inequalities, which are in the main structural ones although they are experienced at the personal level, and promoting their elimination are integral ingredients in the environmental justice mix. This consideration, endorsed by those elements of the environmental movement that assert that the world's elites cannot continue with 'business as usual' or a retention of the status quo (TWN, 2012), is a component of green social work. Its practitioners address it by holding these powerful elites to account for the decisions they make and by demanding that they involve all those who are affected by their decisions in making them, to coproduce solutions to the consequences that might follow specific courses of action.

This political approach is independent of political parties, which it sees as part of the problem of governance that does not adequately protect the rights of its citizens to a healthy social development that safeguards all living things and planet earth. The centrality of this critical political approach, which questions the fundamental premises of the current organization of socioeconomic relations, differentiates green social work from ecological social work, including the 'deep ecology' movement. Moreover, this stance encompasses the whole of society and the terrain in which each person is embedded.

Additionally, green social workers use the nation-state's commitments to human rights through the Universal Declaration of Human Rights (UDHR) to argue for each person's environmental rights and the pursuit of environmental justice so that each of the earth's inhabitants can lay claim to a protected ecosystem that sustains life for the entire biosphere, which includes human beings now and into the future. This tenet is embedded in the social obligation to create a sustainable holistic environment and to maintain it, and it includes the ecosystems that their lives depend upon. In return, an individual's duty to care for others sustains his or her right to be looked after by others. This dependency upon one another becomes interdependency between people and the physical environment as well as among all peoples inhabiting planet earth. These ideals underpin the notions of solidarity and interdependency that are integral to an ethics of care that encompasses people and the physical environment within a green social worker's point of view.

All member states within the United Nations (UN) have ratified the UDHR, including Articles 22–27. These endorse people's welfare entitlements. However, each nation-state's commitment to the realization of the UDHR's principles is variable, and the UN does little to enforce compliance. The nation-state's abrogation of the responsibility to care for its peoples is a global phenomenon. The contemporary fiscal crises have further undermined these rights by demanding that individuals and families secure their own welfare needs. Countries located in the Global South plead lack of funds to legitimate their failure to provide for their citizens. Western states, especially those hard hit by the recent economic recession like Greece, Spain, and Portugal, have imposed draconian public expenditure cuts as part of an 'austerity' response that has decimated public services and hastened the loss of welfare state provisions. These cuts have brought with them political instability, a decline in social well-being, and resistance to the policies that governments have been imposing upon their peoples. Despite these protests, there has been little disruption to the dominant policy thrust that has devolved the provision of social welfare services onto individuals and families. And, although pauperization is progressing rapidly, the bankers who created these fiscal crises retain their high salaries and standards of living (Jordan and Drakeford, 2012). Part of the problem of enforcement is the UN Security Council's reluctance to trump its members' belief in the principle of the sovereignty of nations with that regarding their responsibility to care for their citizens. Thus, state violations of the duty to ensure that their citizens can secure food, clothing, shelter, education, and health services as required by the UDHR are seldom punished.

Rectifying the gross imbalances of power that this scenario represents is not easy, as the antiglobalization movements, the Occupy Wall Street Movement, and the Indignados among others have found out through their resistance to economic exigencies determining socioeconomic policies. Despite this lack of success, the vision of a better world as a realizable possibility must be retained if greater social and environmental justice is to become the norm. Green social workers promote environmental justice by resisting injustice and helping people mobilize and organize activities that protect their physical environment, promote sustainable socioeconomic development, and foster psychosocial well-being. In this sense, green social workers go beyond their civic duty of being concerned about the physical environment to conceptualize their practice as being embedded within it and being responsible for its care into the future, not just with how it is used by people in contemporary societies. Consequently, green social work practice affirms human rights and social justice to enhance the well-being of people and the environment for today and the future. Its concern to deal with forms of environmental degradation that undermine people's health and well-being provides the rationale that justifies social workers' involvement in environmental issues (Dominelli, 2011, 2012a,b).

Green social workers adopt a political stance in addressing environmental issues. They do so by recognizing that power relations shape human interactions. This includes a concern with gender, ethnicity, age, disability, and other social divisions that produce unequal outcomes for individual and collective well-being. Moreover, this approach is rooted in an ethics of care whereby people care for one another and the environment in sustainable ways to ensure that all living things will survive now and for generations to come. Green social workers are critical of the unsustainable forms of development encapsulated in contemporary neoliberal capitalist modes of production, reproduction, and consumption because these exploit the environment and people's labor and perpetuate social and environmental injustice. Green social workers' interventions are holistic, and they tackle structural forms of oppression, environmental degradation, and injustice to empower people and promote their well-being individually and collectively as well as that of planet earth. Moreover, their work intersects with and straddles the micro, meso, and macro levels of practice.

Commonalities between Ecological Social Work and Green Social Work

Social work practice has focused on the significance of the social environment as articulated through social institutions and analyzed by systems theory to define the ecological dimension of the profession. At best, its consideration of the physical environment extended to housing and transportation until recently (Shaw, 2011). Germain's (1979) edited work on this topic explicates this position comprehensively and draws attention to a range of other authors who follow this trajectory. While it shares concerns with green social work about the relationship between people and the social domain, ecological social work has traditionally ignored the impact of the wider physical environment on people's behavior except for housing and the buildings that house social institutions. This approach is currently replicated in Jack (2012).

In the United States, the concern with the physical environment, raised earlier by Park (1996), remained in the shadows until it was brought onto the agenda at a meeting of the National Association of Schools of Social Work in 2009. This was linked to sustainability and social justice in 2010, when the American educators' Council for Social Work Education produced a policy statement on the subject. A range of other initiatives followed, and Besthorn's (2002) earlier view that social work needed to be free of its focus on people as the basis of its endeavors if it were to embrace due care for the physical environment began to be openly discussed as a curriculum matter. Coates (2003) and Zapf (2009) in Canada have articulated similar points. Along with Besthorn and others, Coates sought to widen the appeal of environmental social work by incorporating the tenets of the 'deep ecology movement' into its considerations. In Europe, environmental social work became particularly strong in Belgium, where Jan Peteers (2010) and his colleagues highlighted the importance of addressing social development in sustainable terms and improving community resilience and stamina in avoiding natural disasters and responding to them. The determination to ensure sustainable social development and enhance community resilience is shared by green social work. Additionally, green social workers argue for the mainstreaming of environmental issues in research, education, and practice (Dominelli, 2012a; McKinnon, 2008).

The desire to move social work away from its anthropomorphic orientations is common to both ecological social work and green social work. Gitterman and Germain (2008) have attempted to go beyond earlier ecological writings by ascribing environmental issues to three conditions that engage practitioners in their daily work. These are crises linked to life transitions, dysfunctional interpersonal relations, and environmental pressures. While these have been integrated into the ecological model, green social work has a different stance in that it sees environmental issues as something that go beyond the life course and an individual's transitions through each separate stage. Instead, green social work views the life course as a thread that links current and future generations of people and the entire ecosphere; eschews the use of stigmatizing labels like 'dysfunctional' in referring to social relations that have gone awry; does not disassociate environmental pressures from the demands of neoliberal capitalism's profit imperative; and demands the coproduction of solutions to socio-economic and environmental catastrophes so as to engage all stakeholders. Thus, the ecological and green social work models share a concern with the 'senseless' exploitation of the earth's resources (Coates, 2003: p. 24), but the ecological approach lacks a sustained critique of neoliberal social relations and the socioeconomic and political systems associated with these. As articulated by Ungar (2002), both models are embedded in the idea that social justice cannot occur without the realization of environmental justice. Resonance between environmental justice and social justice is also crucial in the territories claimed by indigenous peoples who remain custodians of significant portions of the earth's biosphere and have waged significant struggles against environmental exploitation and the loss of attachment with the land (Hart, 2010; also *see* Indigenous Social Work).

Environmental Degradation Perpetuates Environmental Injustice

Environmental degradation perpetuates environmental injustice by making the damage and losses suffered by poor and marginalized people worse than would otherwise be the case. This damage can occur as the loss of health, life, and/or property during so-called natural disasters and human-made ones. Moreover, both poverty and inadequate social development contribute to and are caused by environmental degradation and injustice. Consequently, while 1 billion people overconsume products, 1 billion people die of hunger and a further 1 billion are malnourished. Yet, food prices have been increasing rapidly, and, in 2008, these led to riots across the globe. Additionally, 1 billion people lack clean drinking water in their homes, 2.6 billion people are without sanitation, and 1.6 billion people have no access to electricity. Furthermore, 15 million people lose their lives annually because they cannot afford drugs or access them cheaply (Newsweek, 4–11 June 2012: p. 49). Each of these factors exacerbates the hardships encountered during a disaster, regardless of its cause. Green social workers declare this situation unacceptable and work toward eliminating the inequalities that give rise to these avoidable outcomes.

These statistics are likely to get worse as environmental crises, including those initiated through climate change, increase in frequency and intensity (Oven et al., 2011). Environmental degradation can be caused by human-made disasters, including industrial pollution or malfunctioning nuclear energy plants, and/or natural disasters like tsunamis, volcanic eruptions, and earthquakes. Moreover, rising greenhouse gas emissions based on burning fossil fuels for energy and industrial production purposes are contributing to climate change. With climate change, extreme weather events associated with heat waves, cold snaps, and heavy precipitation can exacerbate the environmental damage and health problems caused by so-called natural events such as flooding and drought, which are exacerbated by inappropriate urbanization practices such as building housing on flood plains (Pearce, 2012). Today, this inequality of outcomes is evident in the figures that reveal that 76% of the damages caused by disasters are sustained in the Global South, 92% of people affected by disasters live in countries located there, and 65% of the economic losses caused by disasters are incurred in low-income countries (IFRC, 2009).

Questioning the links between industrial pollution and damage to the environment has been obvious to a number of thinkers for some time. Bullard (1987), for example, coined the term 'environmental racism' to identify and resist corporate America's use of African-American communities as dumping grounds for toxic wastes. Indigenous people have decried the destruction of pristine hinterlands in North America, South America, and Australasia because their exploitation for minerals and other commodities has left them with few jobs, fewer resources, and lost autonomy as their national governments collaborate with multinational companies based elsewhere to determine their development trajectories (Escobar, 1998). These usually disrespect indigenous peoples' priorities and concerns, thereby promoting environmental injustice. Social workers often get involved in raising consciousness about the issues and resisting these measures alongside local residents protecting their communities (Dominelli, 2012a). Moreover,

social workers are demanding a place at the table in key decision-making fora such as the climate change talks held under the auspices of the UN (Dominelli, 2011).

Environmental degradation intensifies the vulnerabilities of poor and marginalized people to disaster and its impact upon their lives and livelihoods. Increased vulnerabilities lead to increased risk, so reducing vulnerabilities becomes an important part of risk reduction strategies (RRS) and, as Peeters (2010) argues, 'sustainable' social development. The Brundtland Commission (1987) defines sustainability as "a form of development that meets the needs of the present without compromising the ability of future generations to meet their own needs," thus giving sustainability a generational twist. Social workers assist communities in formulating and implementing RRS and sustainable forms of social development. Social workers are also involved in delivering humanitarian aid to those affected by disasters and taking care of people who reach humanitarian aid camps in search of safety.

RRS are more effective if proposed activities engage people through their everyday life routines instead of being treated as special events. Involving people in making decisions about what happens in reducing risks and enhancing resilience for themselves and their communities facilitates the coproduction of knowledge and formation of partnerships that cross disciplinary divides and allow for contact between local people and subject specialists.

Green Social Work Practice

The struggle for environmental justice is an important one and focuses green social workers' activities on redressing a complex range of issues. These include the following:

- Human rights violations
- Environmental degradation
- An economic system that perpetuates structural inequalities and lacks corporate accountability
- Conflicts over the distribution of 'natural' resources
- The neglect of cultural diversity and the nonaffirmation of aboriginal and indigenous lifestyles
- Lack of people-friendly localities and supportive community relationships, especially those associated with unsustainable economic developments
- Lack of provision for health and social services that promote the well-being of people
- Lack of care for the physical environment
- Lack of recognition of the interdependency among peoples and between people and the biosphere–ecosystem
- Lack of resilient built infrastructures, resources, and communities (Dominelli, 2012a)

When working for environmental justice, practitioners uphold the duty to care for and be cared by others, including the duty to care for planet earth. In conducting these tasks, social workers (Dominelli, 2012a):

- Support people in affirming their human, social, and environmental rights
- Support people in protecting the environment, enhancing the well-being of humans and the ecosphere, and securing environmental justice
- Mobilize people in various partnerships and alliances to promote residents' and the earth's well-being
- Empower marginalized individuals, communities, and groups, especially in influencing commercial and other institutional routines as well as policy and decision-makers

Green social workers support people who are physically and psychosocially injured during disasters and assist in repairing damaged environments with the aim of facilitating longer term transformative initiatives that develop environmentally just, sustainable, life-enhancing forms of being and doing. In the course of completing these activities, green social workers carry out tasks that include:

- Assessing need
- Coordinating and delivering goods and services
- Assisting in family reunification
- Supporting individuals and communities in rebuilding their lives, developing resilience, and building capacity to minimize future risks
- Advocating, lobbying, and mobilizing for changes that safeguard the environment and prevent future disasters

Green social workers engage service users in holistic approaches that bring together people to protect their physical, social, political, economic, and cultural environments. These dimensions are part of one whole, and not separate from it. To pursue the cause of environmental justice and undertake tasks effectively, green social work practitioners ought to understand the complex power dynamics and relationships that underpin the entrenched views that people hold about their involvement in the elimination of environmental injustice. These roles, which are relevant at the individual, group, and community levels, include those of:

- Facilitator
- Coordinator
- Community mobilizer (of people and systems)
- Mobilizer of resources
- Negotiator or broker between communities and different levels of government
- Mediator between conflicting interests and groups, including gendered relations
- Consultant to government and other agencies
- Advocate for people's rights and entitlements
- Educator, giving out information about how to access relief aid and avoid diseases that can erupt following a disaster
- Trainer, particularly in how to respond effectively and in mobilizing local resources when disaster strikes
- Cultural interpreter
- Interdisciplinary translator to enable laypeople to understand scientific messages
- Therapist, helping people deal with the emotional consequences of disaster
- Protector of the ecosystem and physical environment (Dominelli, 2009, 2010, 2012a)

Given green social work's rootedness in everyday life practice, no aspect of life lies outside its remit. Moreover, its interventions involve the integration of complex understandings and diverse concerns across the social and physical terrains that green social workers articulate in their practice. These are depicted in **Figure 1**.

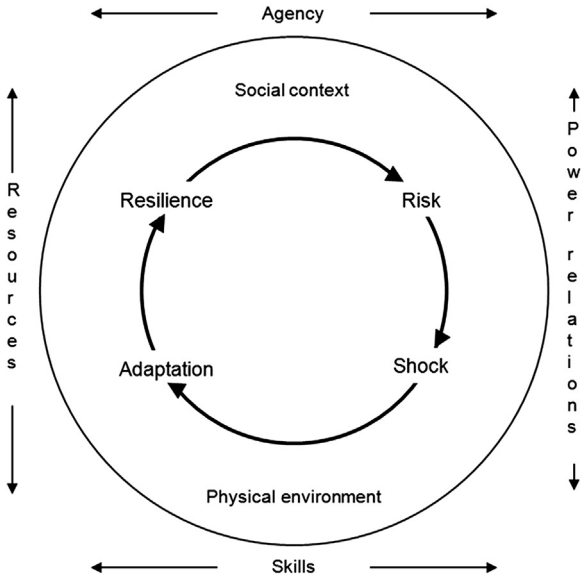

Figure 1 Green social work interventions. Source: Adapted from Lena Dominelli, lecture given when receiving the Katherine A Kendall Award, Stockholm, Sweden, 11 July 2012.

Practice Examples

Case Study 1: Securing Affordable Food

Tom was a green social worker concerned about the lack of fresh food among working families in an inner-city estate. Few people could afford to buy fresh fruit and vegetables. This matter was made worse by the expensive bus trip to get to a market where these items were reasonably cheap. Tom convened a meeting of interested residents, and they decided that they would try to establish a food cooperative on the estate, and that they would seek the local councilor's support in getting a plot of wasteland in the middle of the estate set aside for allotments (garden plots) so that residents could grow some of their own food to decrease reliance on the supermarkets and grow food sustainably and organically.

Tom assisted a small group of residents in discussions with farmers who were interested in supplying the cooperative with fresh fruit and vegetables at affordable prices. The farmers discovered that they could be paid more per unit of food by bringing the goods directly to the cooperative than they were able to extract from well-known supermarket chains, and the prices charged to the residents were lower than in the nearby shops. The councilor agreed to support the estate's wish to use the wasteland for allotments.

The residents' dream of having allotments proved more difficult to realize than anticipated because the large development company that owned the land was concerned that in allowing residents to use it for allotments, its plans for developing the land in future might be thwarted, especially if the government became serious about promoting 'green' initiatives of which growing one's own food could become one expression of them. The residents refused to give up. Initially, they offered to pay the company the going price for the land, but the company director refused to sell unless it were sold as land with development potential, even though he had not submitted any plans for its development.

The ensuing stalemate was shifted only when the councilor successfully convinced the company that there would be little likelihood of such permission being given for at least the foreseeable future, and challenged them to submit plans to the local council for consideration. The company refused to do this, arguing that the current economic situation mitigated against it putting forward a serious proposition. Eventually, a compromise was suggested of a sum above the current price, but with the company retaining the possibility of receiving a further amount if the land's development potential was realized. Although most of the residents were satisfied with this option, several were not. They asked Tom what could be done to protect their allotments from being taken over in the future for development purposes. Tom agreed to help them find out and took them to meet planners and lawyers for expert advice on the issues. He also found some specialists on organic farming who would provide the estate with free advice on how to grow crops sustainably. The inputs from these experts enabled the residents to put forward an alternative proposition that the company accepted.

By becoming organized, the residents had turned the issue from one about the lack of fresh food on the estate to the right to food on their terms. And, with the assistance of the green social worker, they were able to mobilize human and other resources on their estate and elsewhere to challenge a large company's definition of the land as holding future development potential that entitled them to a deferred profit. Moreover, they sought to make their gains sustainable by seeking ways of preventing the large company that had already profited by selling the land above the current market value from making further money from this deal in future by denying its interest in the matter if the land were to acquire development potential. In this way, they attempted to formulate alternative visions of the world and the relationships between them, the land, and the economic system in which they lived. To achieve this goal, the residents and Tom had to rely on expertise from several professions and obtain the support of their local representative. This indicates that green social work has to work holistically across the micro and meso levels, and engage with and navigate through competing discourses to find a solution that is acceptable to all the stakeholders involved. This case also indicates that finding local solutions to global problems – such as food insecurity, including food shortages and environmentally unfriendly forms of food production – can make a small contribution toward the development of alternatives that could be replicated elsewhere if people were to adopt them.

Case Study 2: Multilevel Interventions during Extreme Weather Events

Alice lived in a remote rural village that was prone to flooding. The villagers were a strong cohesive group that supported each other during difficult times. However, they were finding the failure of the local authority to take appropriate measures after each flooding event troublesome. In the last 5 years, their lives had been disrupted three times with major flooding requiring the evacuation of the village. Moreover, as the repairs after each event were substantial, people had barely moved back into their refurbished houses before having to leave them again and seek long-term accommodation with friends or relatives or in rented housing. A few people got so disgruntled that they left the village, and it was not long before they lost contact with friends and neighbors there.

After the third event, Alice decided that it was time that the village organized to demand that substantial flood alleviation work be undertaken. She was not sure where to look for help and support, but she talked to one of the villagers who had recently moved into the area and worked at a nearby university and asked what they could do to help. It turned out that the university had a center undertaking specific research on flooding, especially that linked to climate change, and several of the staff there agreed to help. Among this group were mathematical modelers, engineers, geographers, and

community development workers. They all attended a public meeting organized by Alice.

Jack, one of the community development workers who was interested in green issues, suggested that the local representatives for the village, the emergency planners, the fire services, and the police were also invited to attend, which they did. By having them there, Jack told Alice, they would be able to offer advice based on previous experiences of intervening in flooding situations, but they could also ensure that any proposals that came forward would get formal attention from their respective institutions and that this might result in a sustainable solution being found.

A flooding action group was formed at the meeting. Moreover, the residents made it clear that they had views about their village and what they wanted done in it, and ideas about how to minimize flooding risk that the experts from the university endorsed. The villagers were fascinated with how the mathematical modelers and geographers turned their anecdotal suggestions into models and hypotheses that could be tested through computer modeling and that these could assist policy-makers in difficult decisions when it came to prioritizing needs to make the best use of scarce resources.

After the meeting, Alice took Jack aside and informed him of how pleased she had been with the results, which included a series of further meetings to develop concrete action plans and take the matter forward. However, she said she was concerned that there had been no discussion about the links between formal carers and the informal carers who had to pick up the pieces once the emergency services departed. She added that the support of informal carers was often required for extensive periods of time, which often strained relationships between those providing informal care and those needing this help. While she did not want to discourage volunteers from coming to assist in these situations, she did not think it fair that they became the mainstay of provisions.

Jack suggested that this was an issue that could be addressed at a subsequent meeting of the flooding action group. Furthermore, he thought it also might be worthwhile to take up the issue with the local authority because it was responsible for ensuring that emergency responses were adequate for the needs of the people affected. He was sure that this matter had not been a priority in the council's earlier deliberations on flooding in the area, because he had been an observer at them, and this issue had not been raised. He also volunteered to accompany Alice and a group of residents to the council to discuss this matter if Alice organized the meeting. He suggested that the local councilor could help her set up this meeting. Alice thought this was a great idea and said she would take it forward.

As the work progressed and the long-term flood defenses began to materialize, the local media reported on the developments. The local representative later reported this work to a national meeting that Alice also attended, and it sparked national interest in the approach that the village utilized of coproducing solutions to problems. Additionally, Jack went to an international conference and gave a presentation on their work, thereby giving the initiative an international airing.

In this scenario, Alice is a villager concerned with addressing the impact of an extreme weather event – flooding in her village, which is likely to increase in frequency and intensity as a result of climate change. The multidisciplinary team, which came to assist her attempts to change the local authority's response to the villagers' desire to reduce the risk of flooding taking place and minimize the damage this might inflict, operated on a number of levels – the local, the national, and the international. Additionally, Alice raised an issue that is seldom considered in disaster interventions, including those caused by climate change – service providers' overreliance on unpaid carers for substantial periods of time, which Dominelli's (2012c) study discovered was built into the planning cycle as part of the local authorities' cost-cutting measures.

Conclusions

Environmental issues are becoming more embedded in social work education, theory, and practice. However, there is much work to be done to build on the foundations that the ecological movement, ecological social workers, and green social workers have provided. Green social workers recognize the political nature of ecological interventions that challenge the exploitation of the earth's peoples, flora, fauna, and physical ecosystems. Green social workers' analyses and actions are centered on an extensive critique of neoliberal capitalism that prioritizes profit making over the biosphere. Many vested interests will seek to condemn green social workers' involvement in ecological matters and portray their doing so as a political concern that remains outside the bounds of professional practice.

See also: Ecological Social Work: Shifting Paradigms in Environmental practice; Sustainable Development and Social Work.

Bibliography

Besthorn, F., 2002. Radical environmentalism and the ecological self. Journal of Progressive Human Services 13 (1), 53–72.

Bullard, D., 1990. Dumping in Dixie. Race, Class and Environmental Quality. Westview Press, New York.

Brundtland, G., 1987. The World Commission on the Environment and Development. Our Common Future. Oxford University Press, Oxford.

Coates, J., 2003. Ecology and Social Work toward a New Paradigm. Fernwood Publishing, Black Point, NS.

Dominelli, L., 2009. Introduction to Social Work. Polity Press, Cambridge.

Dominelli, L., 2010. Social Work in a Globalising World. Polity Press, Cambridge.

Dominelli, L., 2011. Climate change: a social work perspective. International Journal of Social Welfare 20 (4), 430–438.

Dominelli, L., 2012a. Green Social Work. Polity Press, Cambridge.

Dominelli, L., 2012b. Gendering climate change: implications for debates, policies and practices. In: Alston, M., Whittenbury, K. (Eds.), Research, Action and Policy: Addressing the Gendered Impact of Climate Change. Springer, Sydney, pp. 77–94.

Dominelli, L., 2012c. Mind the gap: built infrastructures, sustainable caring relations and resilient communities in extreme weather events. Australian Social Work. http://dx.doi.org/10.1080/0312407X.2012.708764.

Escobar, A., 1998. Whose knowledge, whose nature? Biodiversity, conservation, and the political ecology of social movements. Journal of Political Ecology 5, 53–82.

Germain, C. (Ed.), 1979. Social Work Practice: People and Environments: An Ecological Perspective. Columbia University Press, New York.

Gitterman, A., Germain, C.B., 2008. The Life Model of Social Work Practice, third ed. Columbia University Press, New York.

Hart, M., 2010. Indigenous worldviews, knowledge and research: the development of an indigenous research paradigm. Journal of Indigenous Voices in Social Work 1 (1), 1–16.

International Federation of the Red Cross and Red Crescent Societies (IFRC), 2009. World Disasters Report: Focus on Displacement and Migration. IFRC, Geneva.

Jack, G., 2012. Ecological perspective. In: Gray, M., Midgley, J., Webb, S. (Eds.), Sage Handbook of Social Work. Sage, London.

Jordan, B., Drakeford, W., 2012. Social Work and Social Policy under Austerity. Palgrave Macmillan, London.

Kroll, L., Fass, A., 2011. The World's Billionaires. Forbes Magazine, Special Report, 8 March.

McKinnon, J., 2008. Exploring the nexus between social work and the environment. Australian Social Work 61 (3), 256–268.

Oven, K., Curtis, S., Reaney, S., Riva, M., Ohlemüller, R., Dunn, C.E., Nodwell, S., Dominelli, L., Holden, R., 2011. Climate change and health and social care: defining future hazard, vulnerability and risk for infrastructure systems supporting older people's health care in England. Journal of Applied Geography. http://dx.doi.org/10.1016/j.apgeog.2011.05.012.

Park, K., 1996. The personal is ecological: environmentalism of social work. Social Work 41, 320–323.
Pearce, F., 5 May 2012. Could a climate change set off volcanoes and quakes? Environment 360: Opinion, Analysis, Reporting and Debates. http://e360.yale.edu/feature/could_a_changing_climate_set_off_volcanoes_and_quakes/2525/ (Accessed 20 Dec 2012).
Peeters, J. (Ed.), 2010. A Resilient Society: Social Work and Sustainable Development. EPO Publishing Group, Antwerp, Belgium.
Shaw, T.V., 2011. Is social work a green profession? An examination of environmental beliefs. Journal of Social Work. http://dx.doi.org/10.1177/1468017311407555.
TWN (Third World Network) (2012). Doha News Update, No. 27, 19 December.
Ungar, M., 2002. A deeper, more ecological social work practice. Social Services Review 76 (3), 480–497.
War Child, 2012. Congo's Conflict Minerals. Retrieved from http://www.warchild.org.uk/issues/congo-conflict-minerals?gclid=CLS44 NegqrQCFaI7Q godhDUAPQ.
Zapf, M.K., 2009. Social Work and the Environment: Understanding People and Place. Canadian Scholars' Press, Toronto, ON.

Greening of Business

John A Mathews, MGSM Macquarie University, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The global industrial system has been amazingly successful in raising productivity and lifting incomes for hundreds of millions of people. But as it expands to come up against global limits, it gives rise to serious, if not existential, problems of which global warming is the headline case. In response, an expanding literature is contrasting two kinds of overall business trajectories, one that continues with fossil fuels and resource extraction (known as business as usual) and an alternative based on renewables and resource circulation, known as greening of business. The greening option poses a challenge for business in that it departs from the convention that business categories are considered 'content-neutral.'

Introduction

The theory and practice of global business turns on a number of widely accepted abstractions – revenues, costs, profit, inputs, outputs, efficiency, industrial dynamics, and entrepreneurial dynamics – all of which are assumed to operate without regard to underlying physical realities. The firm is held to operate in a position of 'solvency' where revenues exceed costs, and where the physical realities of these categories are not usually specified. Profit opportunities are discovered within a system of abstractions and conventions that have no direct link to the physicality of inputs and outputs. But of course the reality is that 'inputs' are always physical – in terms of certain kinds of energy and resource choices – as are outputs. Under a variety of pressures – including looming uncertainties in the supplies of these physical commodities and energy sources (or their 'peaking'), and concerns over energy and resource security – more and more firms and the countries where they are based are coming to view the underlying physical realities as being of more than passing interest.

This leads to the identification of two dominant business trends at the current time – one where the physical realities are ignored (or rather the existing fossil fuels and resource mix is assumed to continue indefinitely), termed 'business as usual' (BAU), and one where there are serious efforts made to bring business/economic activities into closer alignment with planetary boundaries and biophysical cycles. This latter is termed 'sustainable business' or 'global sustainability' (e.g., Hart and Milstein, 1999) or, more simply, 'greening of business' (GoB).

The BAU pathway is certainly the dominant trend, backed by the full weight of the global fossil fuels and resources industries together with their adjuncts and related activities such as the automotive and petrochemical industries. The majority of capital investment channeled both through the world's stock markets and the bond markets is invested in BAU activities, giving rise to a condition of 'carbon lock-in' (Unruh, 2002). The enormous productivity gains unleashed by fossil fuels and global resources extraction brought economic wealth and political dominance to the countries that first perfected this system – states such as Britain, France and ultimately Germany, the United States, and in the twentieth century, Japan – giving rise to the present drive on the part of China, India, and other emerging giants to share in the fruits of industrialization.

But as the system spreads around the globe, through the twin forces of globalization and industrialization, bringing countries such as China, India, and Brazil into the orbit of the industrialized world, so the pressures of the BAU system on the physical and ecological realities of the planet become increasingly conspicuous. These pressures are felt as resource scarcity and the 'peaking' of numerous commodities previously thought to be in near 'infinite' supply; in the growing pollution levels in air, water, and soil; in species loss and drastic reduction in biodiversity; and ultimately in global problems such as global warming that threaten extinction of the entire industrial system.

The 'greening' business variant arises as a response to these pressures and trends. It does not propose a replacement of the system of global industrial capitalism – in the way that socialism in the nineteenth and twentieth centuries proposed a socially owned alternative to the private ownership of industrial capitalism. But it does propose a radical alternative in the rules by which capitalist firms make decisions, driving their operations into closer alignment with physical and ecological realities through the operation of the financial system, fiscal system, and government regulation. While in the West, these pressures are commonly viewed as constituting a moral imperative (this is what we *should* do, to provide a legacy for our grandchildren); in China and the other emerging industrial giants, the pressure to green the economy is being viewed more as an economic imperative. The traditional 'BAU' model of industrialization simply will not scale when put into practice by the new global giants. They have to find a new and sustainable alternative – and are doing so (Hu, 2014). Whether they will do so in time is a matter that can only be subject to speculation.

In what follows the emergence of these 'greening' forces that are transforming business, and the obstacles encountered, are reviewed, and the main lines of development are traced. Note that the emphasis is on the process or trend of 'greening' and not on some putative end state of a 'green capitalism' – if such can ever be achieved.

The Nonsustainability of 'BAU'

It is the vast scale of the changes to the natural world unleashed by industrial capitalism that is the core issue for sustainability.

The workings of markets, finance, and securitization have allowed an economic system to flourish that is divorced from its ecological moorings – to the short-term benefit of all who enjoy such a system. The mode of operation, however, is probing pathologies such as 'overshoot and collapse' in one market after another – peaking of oil and other fossil fuels being the most obvious symptoms. Global warming and its consequences constitute the end result. Financialization is taking its toll (i.e., finance becoming the prime driver of business), not just in the finance sector itself but in the commodities markets as well, undermining the operation of the entire industrial system. The issue is to reestablish the system's moorings by reframing capitalism as a system embedded in an ecological matrix.

The consequences of scaling up the industrial system to its planetary limits are by now clear enough and well known in the scientific literature. The impact of humans on the terrestrial cycles of the earth is now so profound that it has spawned a new term for a new geological epoch – the *Anthropocene* (Steffen et al., 2007). Successive reviews have pointed to the seriousness of the changes unleashed – culminating in the threat and reality of global warming (Garnaut, 2008; Stern, 2007). But air and water pollution are simply the most visible of the problems created as the industrial system expands to become global. Of more immediate concern is the pressure imposed on existing supplies of fossil fuels and resources, with the prospect of political and ultimately military clashes between countries over the remaining resources. As the peak is passed in one commodity after another (particularly for conventional oil and gas) so the industry searches for more extreme resources (Klare, 2012) – meaning that they become extremely difficult to find, to produce, and entail extreme levels of danger (as evidenced in the Deepwater Horizon oil spill experienced by BP in the Gulf of Mexico in 2010). Energy security concerns, and resources security more generally, loom large as immediate problems generated by the BAU trajectory.

On top of the physical problems, there are problems facing firms that elect to continue with BAU in the form of investment in assets such as coal-fired power stations or deepwater drilling that are likely to lose value as pressure to find an alternative pathway intensifies. These 'stranded assets' that could grow to mean billions or even trillions of dollars constitute a uniquely business response to the sustainability issues raised by the existing industrial system (CTI, 2012).

Elements of Greening of Business

The main alternative to the problems created by BAU is to reestablish the industrial system's moorings by reframing of capitalism as a system embedded in an ecological matrix – calling for a new ecologic for capitalism (Mathews, 2011, 2014). This is, in essence, what is meant by the GoB. The Western model of industrial capitalism has developed through three kinds of markets. Changes are already observable in the major markets – in energy systems (from fossil fuels to renewables); in resources and commodity markets (from a linear to a circular economy); and in finance (from generic finance to eco-targeted finance). These markets are interconnected, and changes in one sector trigger changes in another, and so the process of change can be anticipated to proceed as in a chain reaction, driven by capitalist competitive dynamics.

Energy

The generation of power through harvesting renewable energy supplies constitutes a completely different energy foundation from one based on exploiting fossil fuels drilled or mined from the earth. Renewable energy sources provide a regenerative foundation for the industrial economy that calls for sophisticated technology and smart companies to develop and disseminate it; all renewable generating devices are products of manufacturing, and hence potentially available to all (Mathews and Reinert, 2014). In place of international conflict being fueled by geographical accidents where oil is found in some countries but not others, all countries will have the option of building the technologies and industries needed to harvest renewable power. China has emerged as world leader in this historic transition, creating markets, building industries, and investing in clean technologies at the maximum available rate. Renewables then emerge as the default option for the energy system of a global greening of capitalism.

Resources

There is a complementary shift under way from the *linear economy*, with its wasteful practices, to a *circular economy*. This is meant by a shift in commodity/resource flows. The prevailing economic model is based on a straight-line process starting at one end with capturing materials from nature and, at the other end, dumping the wastes again in the sink called 'nature.' But it is now known that the biosphere works according to great cycles or recirculation of materials, where wastes from one process become nutrients for another. The capitalist version of this involves eco-linkages, where one firm's waste products become another firm's inputs, and entrepreneurial activity is focused on discovering and implementing such eco-initiatives. The Chinese call this a 'circular economy' and it is already embodied in the country's 12th Five Year Plan (Mathews and Tan, 2011). The circular economy provides a template for the green economy – and the default option for the material system of a greening capitalism.

Finance

Third, there are the complementary changes in finance needed to drive through these changes in the energy and materials markets. Apart from the toxic effects of many financial derivatives, introduced without regulation and wreaking destruction, the banking and finance sector has also developed sophisticated instruments for creating credit and channeling finance to companies looking to invest in new processes. Green economy projects are looking increasingly attractive and are to be serviced by targeted finance (green banks, climate bonds). Green development projects call for high capital investment, but then have low running costs because they do not need fuel. The 'gap' between expenditure and rising income can be covered by a financial instrument such as bond, suitably

certified to ensure that funds raised are expended on a climate-friendly project. As they become adopted by institutional investors such as pension funds, these green credit instruments will aggregate projects and lower the cost of finance, thus making green economy projects even more attractive. At the last count, the OECD estimated the total as $71 trillion in pension and superannuation funds, insurance funds, hedge funds, and sovereign wealth funds (Kaminker and Stewart, 2012). If properly marshalled, these funds would transform the energy and resources world and do something serious to dent prospects of global climate change.

It is the interaction between these changes that is having the most profound effects, with each one stimulating developments in the others. The 'greening' of business is actually consistent with a worldwide trend toward living within our resource limits and makes abundant business sense – once the drivers of greening are understood. So what are the drivers of greening?

Drivers of Greening

As a business proposition, the GoB may be evaluated at different levels. At the level of individual firms, there is the prospect of new business opportunities created by greening, and the profits to be won when firms make these choices. This framing of strategic choices by firms to green their operations may be captured in the notion of 'corporate and social responsibility' (CSR) – where environmental responsibility or more broadly 'eco-responsibility' is the key idea. While it is undoubtedly the case that some firms (and some entrepreneurs, such as Elon Musk in California or Wang Chuanfu in China) have reaped substantial profits from their greening strategies, it is also the case that many firms are holding back and continuing their BAU operations in a context of carbon lock-in; these are the 'bad' and 'ugly' aspects of CSR (Devinney, 2009).

Greening initiatives clearly cover a broad span of responses, from minimal changes and those that are branded as progressive such as 'creating shared value' (Porter and Kramer, 2011; for critique see Crane et al., 2014) to genuinely comprehensive shifts in behavior and values. The 'framing' of such issues is all-important (Hoffman, 2011). A broader perspective, in which greening implies something more than an attitude on the part of firms, generates greater insight; as Siegel (2009) has argued, there has to be an objective dimension to the issue.

But the literature on greening locates broader drivers than CSR. The GoB literature today traces its origins to two important articles that appeared in the 1990s. One is the article by Hart (1997; as elaborated in Hart and Milstein, 1999) which argues that greening constitutes a trend that threatens destruction of existing industries, and greening strategies therefore represent survival choices for firms (and the planet). The other is the article by Porter and van der Linde (1995) which argued that tighter regulation on environmental matters could enhance firms' competitiveness, by forcing them to be cleaner and tighter in their management of resources (rather than simply adding to their costs, as in the conventional view). Both articles have spawned a large literature (and the Porter hypothesis in particular has been subjected to deep historical evaluation, e.g., Desrochers and Haight, in press), but as yet there has been little interaction between these two streams of work.

A way can be found to bring the two streams together, by demonstrating how they are both consistent with the greening imposed by the planetary diffusion of industrialization. In the case of Hart and Milstein, the greening pressures arise from an economic imperative (as experienced by China et al.) rather than a moral imperative as originally stated – but the emphasis on creative destruction of 'old' nongreen industries will remain. In the case of Porter and van der Linde, the regulatory pressure they were concerned with (and which has sustained a huge secondary literature, e.g., Lanoie et al., 2011) can be viewed as being generalized by the economic imperatives facing Chinese firms and driving them to change the rules of the game for everyone else. But their insistence that green pressures are procompetitive remains intact.

The strongest argument for greening of the business system is that it constitutes an economic imperative for China, India et al. as they industrialize. This puts the process of greening in the context of the diffusion of industrialization, and frames it as an economic imperative for countries such as Brazil, India and China as they seek to raise their living standards and become parties to the global business system (Spence, 2011; Hu, 2014). From this perspective the principal driver is the diffusion of industrialization to acquire a planetary scale, and the choices that this is imposing on China in the first instance, but on India, Brazil, etc. as well.

In a nutshell, China can be seen to be driving a green revolution because the country has no alternative. The transformation of China achieved over the past 30 years and more is attributed to its rapid industrialization, powered (as in all previous countries' cases) by access largely to fossil fuels and seemingly unlimited access to raw materials. But in China's case, the scale of the transformation involved means that it is rapidly becoming the primary, or largest consumer of fuels (world #1 in coal, and in imports of oil) and in this sense it is highly exposed to threats of peaking of these fuels, but more immediately to geopolitical threats and the risks of war (see e.g., Mathews and Tan, 2014). In such a setting, building up renewable energy sources as a serious and scalable alternative to fossil fuels makes abundant sense, because renewables are manufactured – and manufacturing can be managed as a domestic enterprise, unlike extraction of fossil fuels on the other side of the world and their transport through hostile waters. So there is an energy security argument for China's rapid buildup of renewables (as will be demonstrated through the data) and not just a climate change argument.

Likewise a case can be made that China cannot afford (economically, politically) to infringe too far on other countries' territory for its raw materials, and will have to develop a recycling economy as a national development priority – and this is exactly what it is doing with its circular economy developmental goal. The circular economy idea goes beyond recycling to envisage outputs of one industrial process being fed as inputs to related industrial processes – on the model of the great cycles of the biosphere – and in the case of China has been elevated to a national development goal (Mathews and Tan, 2011).

Hence it may be argued that these choices having to be made by China are driving its own industrialization and economy in a green direction, and that by capitalist competition and emulation other countries will follow suit – so that greening becomes a global trend. Germany is already the most important case in Europe (e.g., Davidson, 2012), while India and Brazil are emerging as the fastest followers among developing countries. This provides a fresh context for discussing the 'GoB' – characterized as the overall direction in which business is moving worldwide in the twenty-first century. Strategies are being fashioned for a world where renewables will be more prominent than fossil fuels, and recycling more common than virgin extraction. If these trends are sound, then 'greening' of business is in fact a survival strategy, for businesses themselves as well as for the planet.

This viewpoint provides a perspective at variance with the mainstream economics literature. One can take issue with the standard economic argument that climate policies constitute a 'cost' that represents a burden on economies (e.g., Stern, 2007; Garnaut, 2008). By contrast, it can be argued that greening is not a cost but a business opportunity – which is the appropriate perspective for the business literature. In fact greening is the biggest business opportunity of the century, given the aspirations for industrial development held by China, India, Brazil, and many more countries. Several business billionaires such as Elon Musk (Tesla), Masayoshi Son (Softbank and founder of Japan Renewable Energy Foundation), and Wang Chuanfu (BYD) already demonstrate that green issues represent an enormous business opportunity.

This business perspective spills over to an argument about policy. The economists argue that climate change calls for pricing of carbon emissions, which are perceived as a cost – whereas a business perspective would look for economic incentives to be put in place, rather than disincentives (i.e., taxes). A good example of a market-based approach to changing economic behavior is the Chinese Sloping Lands Conversion Program, where the terrible problem of loss of soil in China's river valleys has been tackled over the past decade by a program that rewards farmers for avoiding growing crops on steeply sloping land, and penalizes those who continue to do so. Feed-in tariffs likewise constitute a market-friendly incentive for the uptake of renewable energies by businesses and households, without impinging on the tax system.

In this way it may be argued that GoB is not a marginal 'add-on' for corporate strategy but can in fact be viewed as the central issue with which business will have to engage in the coming century – an engagement with energy and resources certainly, but also with water supplies (e.g., Sedlak, 2014), agriculture, construction, and almost all industrial activities. The emerging industrial economies are the primary drivers of the greening process (as opposed to other perspectives such as those that view carbon markets as drivers).

If this GoB is increasingly a worldwide phenomenon, then businesses will have to adapt their strategies to accommodate this trend – and to seek to lead it rather than just follow. In this sense, the argument is one that transcends CSR arguments. The perspective adopted engages with analyses that point to the emergence of new business models based on sustainability (e.g., Girotra and Netessine, 2013) where the emphasis is on economic imperatives rather than moral imperatives. Controversial issues abound, such as the rise of carbon markets (Dhanda and Murphy, 2011) or the rise of carbon indices on stock exchanges (Orlitzky, 2013) – all jostling for attention. These measures need to be contrasted as drivers of greening when compared with the fundamental changes driving energy and resource systems.

Transition Dynamics

The greening of markets for energy, commodities, and capital can be expected to propagate to encompass the entire economy, through multiple interfirm connections and driven by competitive forces. Green products will call for new value chains that will propagate via intermediate suppliers and aggregators back to ultimate commodity suppliers, where the greening of commodity markets will exert their effects downstream. The barriers that stand in the way of this emergent system are formidable, from the protection of vested interests and continuation of subsidies to fossil fuels, to the clash of sectional interests and the continuing financialization of markets that divorce them ever further from nature. Ultimately it is the strong states that drive fundamental change – and this was Polanyi's point in relation to the first *Great Transformation* (Mathews, 2014). It is the state intervention that creates the space where entrepreneurial initiative in building new systems can flourish. Recent work indicates, for example, the powerful role of the US state in driving enterprise and innovation, where state agencies take risks that the private sector cannot or will not accommodate (Weiss, 2014; Block and Keller, 2011) – particularly motivated by national security concerns, as in promotion of renewable energy systems by the US military. Now the role of the state in driving industrial systems onto a new, broader, greening trajectory is clear – with the case of China being the preeminent example.

From this perspective, it is clear that China's greening is already underway (as counterpart to its black development), and there is clear evidence to support this (e.g., cost reductions in solar photovoltaic industry, as demonstrated by Bazilian et al., 2013) – and that this constitutes the principal driver of greening of the economy worldwide. Basic business processes of competition and emulation will exert their effects.

As greening trends are taken up by developing countries, they can be expected to pursue the latecomer and fast-follower strategies already perfected by the East Asian countries that successfully industrialized in the twentieth century. Targets for catch-up can be identified in the renewable energy industries (Mathews et al., 2011) and in the industrial ecology initiatives now being diffused through China. It is not innovation as such that will drive this process, but market expansion and the cost reductions that follow.

Concluding Remarks

There is a dawning realization within business that the greening economy provides an alternative conceptualization of an economy that can coexist with the great ecological cycles of

nature. The kind of industrial capitalism that is emerging in the twenty-first century is one where capitalist firms operate in 'naturalized' markets, where the firm itself can grow but the system as a whole remains within its ecological limits. Boulding's image of Spaceship Earth will remain the reigning image for our time – until a new form of industrial capitalism has been put in place, one that exhibits greening trends that supersede BAU. In place of the 'content-neutral' categories of BAU capitalism, commodities are likely to carry comprehensive records of their origins (e.g., through the Internet of Things) – with future contracts on commodities markets likely to be the first instances of a new, ecologically aware industrial capitalism (Mathews, 2008). There will be growth and innovation in this system, but not the extensive growth that has characterized industrial capitalism to date. It is likely to emerge first in China because it is in China that the leadership now recognize that the promise of Western capitalism – that it could be replicated around the world – is an empty promise; this kind of 'wild capitalism' will not scale to allow billions of people to enjoy higher living standards (Nolan, 2009). And it is in China that the state is powerful enough to drive through the changes needed, backed by a commitment to sustainable enterprise as a way of legitimizing the regime. Other developing countries can be expected to emulate this Chinese model, and then the developed world itself is likely to change its ways (as is already happening in Germany and Japan). This next industrial capitalism is one that will not cost the earth. But will it appear in time? The greening of capitalism is a natural experiment whose outcome is anything but predetermined.

See also: Business and Society: Social Accounting; International Business; Interorganizational Relationships and Networks; Management: General; Multinational Corporations; Organization: Overview; Special Economic Zones and Cluster Dynamics: China.

Bibliography

Bazilian, M., et al., 2013. Re-considering the economics of photovoltaic power. Renewable Energy 53, 329–338.

Block, F.L., Keller, M.R., 2011. State of Innovation: The U.S. Government's Role in Technology Development. Paradigm Publishers, Boulder, CO.

Crane, A., Palazzo, G., Spence, L.J., Matten, D., 2014. Contesting the value of "Creating shared value". California Management Review 56 (2), 130–153.

CTI, 2012. Unburnable Carbon. Report from Carbon Tracker Initiative, London at: http://www.carbontracker.org/site/carbonbubble.

Davidson, O.G., 2012. Clean Break: The Story of Germany's Energy Transformation and what Americans Can Learn from it. InsideClimate News, Brooklyn, NY.

Desrochers, P., Haight, C.E. Squandered profit opportunities? Some historical perspective on industrial waste and the Porter Hypothesis, Resources, Conservation and Recycling, in press.

Devinney, T.M., 2009. Is the socially responsible corporation a myth? The good, the bad, and the ugly of corporate social responsibility. Academy of Management Perspectives 23 (2), 44–56.

Dhanda, K.K., Murphy, P.J., 2011. The new wild west is green: carbon offset markets, transactions, and providers. Academy of Management Perspectives 25 (4), 37–49.

Garnaut, R., 2008. The Garnaut Climate Change Review: Final Report. Cambridge University Press, Cambridge.

Girotra, K., Netessine, S., 2013. Business model innovation for sustainability. Manufacturing & Service Operations Management 15 (4), 537–544.

Hart, S.L., January–February 1997. Beyond greening: strategies for a sustainable world. Harvard Business Review 66–76.

Hart, S.L., Milstein, M.B., 1999. Global sustainability and the creative destruction of industries. Sloan Management Review 41 (1), 23–33.

Hoffman, A.J., 2011. Talking past each other? Cultural framing of skeptical and convinced logics in the climate change debate. Organization & Environment 24 (1), 3–33.

Hu, A., 2014. China: Innovative Green Development. Springer, Berlin, Germany.

Kaminker, C., Stewart, F., 2012. The Role of Institutional Investors in Financing Clean Energy. OECD Working Papers on Finance, Insurance and Private Pensions, No. 23. OECD, Paris.

Klare, M.T., 2012. The Race for What's Left: The Global Scramble for the World's Last Resources. Metropolitan Books, New York.

Lanoie, P., Laurent-Luchetti, J., Johnstone, N., Ambec, S., 2011. Environmental policy, innovation and performance: new insights on the Porter hypothesis. Journal of Economics & Management Strategy 20 (3), 803–842.

Mathews, J.A., 2008. Towards a sustainably certifiable futures contract for biofuels. Energy Policy 36, 1577–1583.

Mathews, J.A., 2011. Naturalizing capitalism: the next great transformation. Futures 43, 868–879.

Mathews, J.A., 2014. Greening of Capitalism: How Asia Is Driving the Next Great Transformation. Stanford University Press, Redwood City, CA.

Mathews, J.A., Hu, M.-C., Wu, C.-W., 2011. Fast-follower industrial dynamics: the case of Taiwan's solar PV industry. Industry and Innovation 18 (2), 177–202.

Mathews, J.A., Reinert, E.S., September 2014. Renewables, manufacturing and green growth: Energy strategies based on capturing increasing returns. Futures 61, 13–22.

Mathews, J.A., Tan, H., 2011. Progress towards a circular economy in China: drivers (and inhibitors) of eco-industrial initiative. Journal of Industrial Ecology 15 (3), 435–457.

Mathews, J.A., Tan, H., 2014. China's energy industrial revolution. Carbon Management 5 (1), 1–3.

Nolan, P., 2009. Crossroads: The End of Wild Capitalism. Marshall Cavendish, London.

Orlitzky, M., 2013. Corporate social responsibility, noise, and stock market volatility. Academy of Management Perspectives 27 (3), 238–254.

Porter, M.E., Kramer, M.R., 2011. Creating shared value. Harvard Business Review 89 (1/2), 62–77.

Porter, M.E., van der Linde, C., 1995. Toward a new conception of the environment-competitiveness relationship. Journal of Economic Perspectives 9 (4), 97–118.

Sedlak, D.L., 2014. Water 4.0: The Past, Present and Future of the World's Most Vital Resource. Yale University Press, New Haven, CT.

Siegel, D.S., 2009. Green management matters only if it yields more green: an economic/strategic perspective. Academy of Management Perspectives 23 (3), 5–16.

Spence, M., June 1, 2011. Asia's New Growth Model, Project Syndicate. Available at: http://www.project-syndicate.org/commentary/spence23/English.

Steffen, W., Crutzen, P.J., McNeill, J.R., 2007. The anthropocene: are humans now overwhelming the great forces of nature? Ambio 36, 614–621.

Stern, N., 2007. The Stern Review of the Economics of Climate Change. H.M. Treasury, London. Available at: http://www.hm-treasury.gov.uk/sternreview_index.htm.

Unruh, G.C., 2002. Escaping carbon lock-in. Energy Policy 30 (4), 317–325.

Weiss, L., 2014. America Inc? Innovation and Enterprise in the National Security State. Cornell University Press, Ithaca, NY.

Grit

Angela L Duckworth and Lauren Eskreis-Winkler, The University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Grit is a personality trait defined as passion and perseverance for long-term goals. Grit entails committing to challenging projects and working diligently toward their achievement over extremely long stretches of time. Grit predicts objectively measured achievement outcomes, including performance in the National Spelling Bee, retention at West Point military academy, and graduation from the Chicago Public Schools. Consistent with speculation by early twentieth-century psychologists that effort and talent are independent, grit demonstrates incremental predictive validity for achievement outcomes above and beyond measures of talent (e.g., general intelligence, physical aptitude), to which grit is not positively related.

Grit is passion and perseverance for long-term goals. In the face of adversity, failure, and plateaus in progress, gritty individuals hold steadfast to their commitments – not just over weeks or months, but over years and decades. With respect to the Big Five taxonomy of personality traits, grit loads most strongly on conscientiousness, a broader construct that also includes the facets of orderliness, dependability, self-control, and industriousness (Roberts et al., 2005). Like other facets of conscientiousness, grit denotes the tendency to prefer labor over leisure, to hold high standards of behavior, and to fulfill obligations rather than shirk responsibility. Unlike other facets, however, grit further denotes extreme stamina in terms of particular interests and applied effort toward these interests. Crucially, grit is not just about working hard on tasks at hand but, rather, working diligently toward the same higher-order goals for extremely long stretches of time.

The notion that sustained hard work and focused interests are distinct from talent but equally vital to success has been discussed in the psychological literature for well over a century. Perhaps in the earliest systematic inquiry into the psychological determinants of high achievement, Galton (1892) reviewed biographical information on prominent individuals in a broad array of disciplines, including science, poetry, music, art, and the law. Intellect, Galton proposed, was insufficient for eminent achievement. Rather, the most eminent individuals displayed intellect combined with 'zeal,' and the 'capacity for hard labor' (p. 38). Likewise, in a biographical study of 301 geniuses, Cox (1926) concluded that "youths who achieve eminence are characterized not only by high intellectual traits, but also by persistence of motive and effort" (p. 218). In fact, Cox speculated that persistence might be ever slightly more important: "High but not the highest intelligence combined with the greatest degree of persistence will achieve greater eminence than the highest degree of intelligence with somewhat less persistence" (p. 187). In an early review of studies on the determinants of success, Hull (1928) likewise remarked that even a person possessing the greatest natural capacity would not accomplish anything without "a certain amount of industry" (p. 185).

More recently, prospective longitudinal studies have confirmed that the personality trait of grit predicts objectively measured achievement outcomes, even when controlling for measures of talent (Duckworth et al., 2007; Duckworth and Quinn, 2009). This article discusses prior research on closely related constructs before summarizing contemporary research on grit, including known correlates and consequences, psychological antecedents, and cognitive and behavioral mechanisms. Throughout, we emphasize limitations of extant research that suggest promising directions for future research.

Prior Research on Persistence

The early twentieth century witnessed several attempts to assess persistence as a relatively stable, domain-general trait using questionnaires similar in form to those used in modern personality research. For example, under the supervision of Charles Spearman, Edward Webb (1915) studied 194 students at a teacher training institute with the aim of finding "evidence of the same person exhibiting the same characteristics generally, that is to say, under varied conditions" (p. 52). Each student was rated by two older students on 39 items, including the "degree in which he works with distant objects in view (as opposed to living 'from hand to mouth')," the 'tendency *not* to abandon tasks from mere changeability,' and the 'tendency *not* to abandon tasks in the face of obstacles.' While Webb interpreted these items to mark a single general factor of character he termed *will*, reanalysis of his data demonstrates that these items, as well as several others assessing interest in the goodness or wickedness of actions, trustworthiness, interest in religious beliefs, and the desire to excel at performances, load on a personality factor readily identifiable as Big Five conscientiousness (Deary, 1996). Subsequent researchers noted a reliable cross-sectional association between persistence ratings by informants and academic performance. For example, Hughes (1924) found that honors students were rated by their teachers as more persistent than less accomplished students, and Flemming (1925) found that teachers' ratings of high school students' persistence and their academic grades were strongly correlated.

At around the same time, psychologists began developing performance tasks to assess trait-level persistence more objectively. Typically, measurement paradigms entailed a very difficult or impossible task (or, more rarely, a facile but tedious

task) and an unlimited amount of time to work on it, with persistence indexed either as the number of attempts or the duration of voluntary engagement before giving up. Among the most extensive of these endeavors were the studies of character by Hartshorne et al. (1929), who devised physically and mentally demanding tasks to assess trait-level persistence in school-age children. For example, their Story Resistance Task brought students to the pinnacle of suspense in a written story, at which point the typeset became increasingly muddled, necessitating that students "plow through material … with capitals and small letters misplaced and words improperly spaced" (p. 329). Correlations among persistence tasks designed to be administered in groups, though small in size, were consistently positive, as were correlations between these tasks and teacher ratings of persistence. In contrast, tasks designed to be administered individually (e.g., standing on one foot for as long as possible, whistling after eating a cracker) demonstrated inconsistent associations with each other and with teacher ratings of persistence.

Several years after Hartshorne and May, Ryans (1938b) created a similar battery of physical and mental tasks for older adolescents. For instance, participants were timed on how long they could keep themselves from blinking while reading a long passage. In another task, the patellar reflex inhibition task, "the subject was informed that he should attempt to hold back or inhibit the response when…[his patellar] tendon was tapped" (p. 340). Composite scores of persistence on such tasks correlated strongly with academic success in two samples of high school students (Ryans, 1938a). Likewise, MacArthur (1955) administered a battery of persistence tasks, several borrowed from Hartshorne, May, and Maller, to adolescent boys. The underlying persistence factor accounting for positive correlations among tasks correlated moderately with academic grades ($r = 0.3$), an association that was only mildly diminished when controlling for intelligence.

By the mid-twentieth century, the study of persistence as a trait variable waned as psychologists became more interested in the study of situational influences on persistent behavior (Peterson and Seligman, 2004). More generally, the assumption that individual differences in behavior were reasonably consistent across situations came under scrutiny, with many psychologists determining that situational factors were vastly more important than traits (Funder, 2009; Mischel, 1968; Ross and Nisbett, 1991). Laboratory studies of both humans and other animals, most of which were transacted without explicit reference to earlier personality research on trait-level persistence, yielded important insights into situational factors that influence persistence (Pittenger, 2002). Among the most important of these discoveries was the partial reinforcement extinction effect (PREE), which describes the paradoxical phenomenon of increased persistence in the absence of reward after an initial training period in which behavior is rewarded intermittently, compared to a training period in which behavior is consistently rewarded (see Robbins, 1971 for a review).

More recently, Eisenberger (1992) has explained the PREE as a special case of *learned industriousness*, in which the essentially aversive sensation of effort acquires positive secondary reinforcement value by virtue of being paired with reward. For instance, Eisenberger et al. (1979) found that rats randomly assigned to a high effort, infrequent reinforcement condition (requiring the rats to shuttle up a runway five times to get a food pellet) as opposed to a low effort, frequent reinforcement condition (requiring the rats to shuttle up a runway only once to get a food pellet) showed greater persistence on a subsequent persistence task requiring a different behavior (pressing a lever for food). Similarly, Nation et al. (1979) rewarded college students on each trial or only half the trials for moving a peg back and forth; all students then experienced extinction trials in which moving the peg was not rewarded. Later, all students were asked to complete a button-pressing task in which performance was initially rewarded but, subsequently, unrewarded during extinction trials. Students who had been rewarded for only half the trials in the peg task persisted longer in the button-pressing task than students who had been rewarded on all of the peg task trials. These experimental findings, among others, led Eisenberger (1992) to speculate that stable, domain-general differences in persistence of the sort studied in Webb (1915) may originate in the extent to which individuals have experienced reward for low or high effort.

Contemporary understanding of behavior recognizes both trait-level and situation-specific determinants, a nuanced position nicely illustrated by Eisenberger's (1992) work and, indeed, augured by at least a subset of early scholars of persistence (e.g., Feather, 1962). To date, grit has primarily been studied as a relatively stable, domain-general inclination to maintain consistent, focused interest and effort for personally important and challenging goals. However, research on grit as an aspect of personality does not deny the importance of situational factors on behavior, nor does it assume that grit is entirely fixed over time. On the contrary, as detailed below, the conditions that encourage the expression of grit and its development over time constitute promising avenues for future research.

Measuring Grit

To date, most research on grit has relied on self-report questionnaires. The Grit Scale (Duckworth et al., 2007), originally containing 12 items, was later shortened to an 8-item version called the Grit-S (Duckworth and Quinn, 2009; see **Table 1**).

Table 1 Grit-S items

Grit scale
1. New ideas and projects sometimes distract me from previous ones.*
2. Setbacks do not discourage me.
3. I have been obsessed with a certain idea or project for a short time but later lost interest.*
4. I am a hard worker.
5. I often set a goal but later choose to pursue a different one.*
6. I have difficulty maintaining my focus on projects that take more than a few months to complete.*
7. I finish whatever I begin.
8. I am diligent.

Note. Starred items are reverse coded. All items are rated by participants on a scale from 1 to 5 ranging from 1 = *Not like me at all* to 5 = *Very much like me.*

Both scales demonstrate adequate psychometric properties, including internal reliability, test-retest stability, and convergent and discriminant validity (Duckworth et al., 2007; Duckworth and Quinn, 2009). An informant-report version of the Grit-S demonstrates similarly positive psychometric properties and medium-to-large associations with the self-report Grit-S (Duckworth and Quinn, 2009). Both the Grit and Grit-S scales demonstrate an oblique, two-factor structure, with half of the items describing consistency of interests over time and the other half describing sustained hard work.

Due to the well-known shortcomings of self-report measures, recent efforts have focused on developing novel ways to measure grit. One method, introduced by Robertson-Kraft and Duckworth (2012), entails coding resumes for biographical evidence of sustained commitment over time. This approach was inspired by Willingham (1985), who assessed "evidence of purposeful, continuous commitment to certain types of activities versus sporadic efforts in diverse areas" (p. 213) from high school resumes, specifically by identifying the extent to which students evinced continuous, multiyear involvement in activities and objective evidence of advancement therein. Grit ratings assigned via an analogous resume coding procedure applied to college extracurricular and work activities prospectively predicted retention and teaching performance among novice teachers 1 year later (Robertson-Kraft and Duckworth, 2012).

Correlates and Consequences of Grit

In cross-sectional research, grit has been positively associated with lifetime educational attainment among adults as well as with GPA among undergraduates at an elite university (Duckworth et al., 2007). In addition to completing more formal education, grittier individuals tend to make fewer major career changes (Duckworth et al., 2007). In large samples, grit demonstrates a monotonic, positive association with age, suggesting the possibility of normative mean-level increases in grit over the life course (Duckworth et al., 2007). However, the cross-sectional nature of such data is also consistent with cohort effects; it is possible that cultural influences more strongly encouraged grit in decades past than in contemporary society.

Stronger evidence that grit contributes to achievement comes from prospective longitudinal studies. For instance, cadets at West Point military academy completed grit questionnaires on their second day of arrival to campus; grit scores predicted which cadets would persist through the summer training, known colloquially as Beast Barracks for its exceptional rigors both mental and physical (Duckworth et al., 2007: Study 4). In contrast, a composite measure used by West Point admissions called the Whole Candidate Score, comprising a weighted average of SAT score, high school rank, ratings of leadership potential, and performance on a physical aptitude test failed to predict summer retention. Moreover, grit was not related to Whole Candidate Score or to any of its components. Nearly identical findings were obtained in two subsequent replication studies (Duckworth et al., 2007: Study 5; Duckworth and Quinn, 2009: Study 5).

Likewise, among competitors in the National Spelling Bee, grit assessed using either the Grit or Grit-S self-report questionnaires predicts superior performance in final competition months later (Duckworth et al., 2007: Study 6; Duckworth and Quinn, 2009: Study 6). Consistent with the resume coding study by Robertson-Kraft and Duckworth (2012), grittier teachers, assessed using the Grit self-report questionnaire 1 month before setting foot in the classroom, produced superior academic learning gains in their students over the course of the following school year. More recently, prospective longitudinal studies have shown that grit predicts graduation from the Chicago Public Schools, controlling for standardized achievement test scores, as well as completion of a grueling Army Special Operations course among soldiers, controlling for general intelligence and physical fitness (Eskreis-Winkler et al., manuscript in preparation).

Importantly, several studies confirm that grit predicts incremental variance for life outcomes above and beyond that predicted by Big Five conscientiousness, including educational attainment, (fewer) career changes, retention at West Point, sales job retention, and marital stability in men (Duckworth et al., 2007; Eskreis-Winkler et al., manuscript in preparation). In addition, grit is more strongly associated with the tendency to maintain a regular exercise regimen than conscientiousness (Reed et al., 2013). Collectively, both cross-sectional and longitudinal evidence point to grit as an important determinant of achievement outcomes, especially in domains where giving up is common.

Mechanisms Linking Grit to Outcomes

Why do grittier individuals excel in situations where less gritty individuals do not? To date, only two studies have identified mechanisms explaining associations between grit and achievement outcomes. The first of these studies investigated the possibility that grittier individuals dedicate more hours of deliberate practice to their chosen endeavor. Deliberate practice refers to practice activities designed to improve specific aspects of performance, performed repetitively with feedback, at levels such that challenge exceeds current skill (Ericsson et al., 1993). In a study of National Spelling Bee finalists, Duckworth et al. (2010) operationally defined deliberate practice as studying vocabulary words and word origins in solitude and distinguished these activities from being quizzed by another person, reading, and word games, which were rated by spellers as more enjoyable and less effortful than deliberate practice (Duckworth et al., 2010). Although deliberate practice was rated by all study participants as the hardest and least enjoyable form of practice, gritty contestants consistently did more of it. Moreover, hours of deliberate practice fully mediated the prospective association between grit and performance in final competition.

Counterfactual thinking is another mechanism that explains why gritty people are more likely to excel in their commitments. Counterfactual thinking refers to how individuals imagine alternatives to their present realities. Downward counterfactual thinking entails imagining a worse alternative, whereas upward counterfactual thinking entails imagining something better (Roese, 1997). In a replication study

predicting retention among cadets at West Point military academy (Kwok et al., manuscript in preparation), cadets were asked to imagine what they might be doing if they were not currently cadets at West Point. Grittier cadets, who were less likely to drop out, spontaneously imagined an alternative they rated as worse than being at West Point or reported that they could not imagine an alternative at all. By contrast, less gritty cadets, who were most likely to drop out, spontaneously imagined an alternative which they rated as better than or about the same as being a cadet at West Point. Downward counterfactual thinking fully mediated the prospective association between grit and summer retention. If replicated, this finding suggests that grittier individuals may generate lower estimates of the opportunity cost (e.g., the next best alternative) of their present commitments, thereby rendering their present commitments relatively more attractive.

Psychological Antecedents of Grit

Recent work has begun to explore the relation between grit and three motivational orientations to happiness identified by Peterson et al. (2005): pleasure, engagement, and meaning. According to this framework, individuals who pursue happiness through pleasure seek positive hedonic experiences, individuals who pursue happiness through engagement seek attention-absorbing activities, and individuals who pursue happiness through meaning seek activities that serve a greater altruistic purpose. In two samples, grittier individuals were most motivated by engagement and, next, by meaning (Von Culin et al., 2014). In contrast, grittier individuals were less likely than others to pursue happiness through pleasure, a finding consistent with the positive association between grit and deliberate practice (Duckworth et al., 2010).

Other attributes known to lead to persistence, such as optimistic explanatory style (Seligman and Schulman, 1986), provide promising avenues for future research. Optimistic explanatory style is the tendency to attribute bad events to specific, changeable causes; by contrast, pessimistic explanatory style is the tendency to attribute bad events to stable, global causes. Given that individuals with pessimistic explanatory style are less likely to persevere in the aftermath of failure (Peterson and Park, 1998; Peterson, 2000), it is unsurprising that optimistic explanatory style shows medium-sized correlations with grit, $r = 0.32$, $p < .001$ and that, in at least one study, grit statistically mediates the predictive association between optimistic explanatory style and performance (Duckworth et al., 2009). Relatedly, self-report data in a national sample of high school seniors (Duckworth, unpublished raw data) reveal a medium-sized correlation between grit and growth mind-set, the implicit belief that intellectual endowments can be improved with effort (Dweck, 2006).

Conclusion

Grit is often referred to as an exemplar 'noncognitive' characteristic, particularly in work by economists and policymakers (e.g., Almlund et al., 2011). The term noncognitive is less often used by psychologists, who recognize that all psychological processes are cognitive in the sense that they involve the brain and the processing of information (Duckworth et al., 2009). However, there is a sense in which grit is noncognitive: in most samples, grit is either orthogonal or slightly inversely related to measures of cognitive ability (Duckworth et al., 2007; Duckworth and Quinn, 2009; Eskreis-Winkler et al., 2014). Thus, the moniker 'noncognitive' is appropriately applied to grit if narrowly interpreted to denote independence from the capacity to learn. Moreover, the growing evidentiary base on grit supports earlier speculations based entirely on qualitative insights (e.g., Cox, 1926; Galton, 1892) that talent, intellectual or otherwise, is no guarantee of effort. Hence, both talent and grit would seem essential to outstanding achievement.

In 1907, William James speculated, "We are making use of only a small part of our possible mental resources…men the world over possess amounts of resource, which only exceptional individuals push to their extremes of use" (James, 1907, p. 322–323). Condemning the 'habit of inferiority to our full self,' (p. 331) James advocated that psychologists turn to explore not only the range of human abilities, but also the diverse means through which these abilities are unleashed. A little more than a century after James' pronouncement, relatively little is known about the traits that unleash the abilities in which psychology is now well versed. Research on grit has begun to address this imbalance.

See also: Academic Achievement Motivation, Development of; Academic Engagement; Academic Motivation and Performance: Task Value Interventions; Affect-Regulation Motivation; Avoidance and Approach Motivation: A Brief History; Flow in Motivational Psychology; Gender and Academic Motivation; Interest, Psychology of; Leisure Activities Choices among Adolescents; Mastery Learning; Motivation in Youth Sport and Physical Activity: Developmental Perspectives; Motivation, Learning, and Instruction; Optimism and Pessimism; Passion and Motivation; Personal Projects; Personality and Adaptive Behaviors; School Achievement: Motivational Determinants and Processes; Self-Determination Theory; Self-Regulated Learning: Theories, Measures, and Outcomes; Self-Regulation in Adulthood; Test Anxiety and Academic Achievement; Wanting vs Needing.

Bibliography

Almlund, M., Duckworth, A.L., Heckman, J., Kautz, T., 2011. Personality psychology and economics. In: Hanushek, E.A., Machin, S., Woessmann, L. (Eds.), Handbook of the Economics of Education. Elsevier, Amsterdam, pp. 1–181.

Cox, C.M., 1926. Genetic Studies of Genius: Vol. 2. The Early Mental Traits of Three Hundred Geniuses. Stanford University Press, Stanford, CA.

Deary, I.J., 1996. A (latent) Big Five personality model in 1915? A reanalysis of Webb's data. Journal of Personality and Social Psychology 71 (5), 992–1005.

Duckworth, A.L. Predictors of college persistence. unpublished raw data.

Duckworth, A.L., Kirby, T., Tsukayama, E., Berstein, H., Ericsson, K.A., 2010. Deliberate practice spells success: why grittier competitors triumph at the National Spelling Bee. Social Psychological and Personality Science 2, 174–181.

Duckworth, A.L., Peterson, C., Matthews, M.D., Kelly, D.R., 2007. Grit: perseverance and passion for long-term goals. Journal of Personality and Social Psychology 92, 1087–1101.

Duckworth, A.L., Quinn, P.D., 2009. Development and validation of the short grit scale (Grit-S). Journal of Personality Assessment 91, 166–174.
Duckworth, A.L., Quinn, P.D., Seligman, M.E.P., 2009. Positive predictors of teacher effectiveness. Journal of Positive Psychology 19, 540–547.
Dweck, C.S., 2006. Mindset: the new psychology of success. Random House, New York.
Eisenberger, R., 1992. Learned industriousness. Psychological Review 99, 248–267.
Eisenberger, R., Terborg, R., Carlson, J., 1979. Transfer of persistence across reinforced behaviors. Animal Learning & Behavior 7 (4), 493–498.
Ericsson, K.A., Krampe, R.T., Tesch-Römer, C., 1993. The role of deliberate practice in the acquisition of expert performance. Psychological Review 100 (3), 363.
Eskreis-Winkler, L., Duckworth, A.L., Shulman, E.P., Beale, S., 2014. The grit effect: predicting retention in the military, the workplace, school and marriage. Frontiers in Personality Science and Individual Differences 5, Article 36. Retrieved from http://journal.frontiersin.org/Journal/10.3389/fpsyg.2014.00036/full.
Feather, N.T., 1962. The study of persistence. Psychological Bulletin 59, 94–115.
Flemming, C.W., 1925. A detailed analysis of achievement in the high school. Teachers' College Contribution to Education 196, 35–47.
Funder, D.C., 2009. Persons, behaviors and situations: an agenda for personality psychology in the postwar era. Journal of Research in Personality 43 (2), 120–126.
Galton, F., 1892. Hereditary Genius: An Inquiry into its Laws and Consequences. Macmillan, London.
Hughes, W.H., 1924. Organized personnel research and its bearing on high school problems. Journal of Educational Research 10, 386–398.
Hull, C.L., 1928. Aptitude Testing. World Book, Oxford.
Hartshorne, M., May, M., Maller, J.B., 1929. Studies in the Nature of Character: Studies in the Service of Self-control. Macmillan, New York.
James, W., March 1, 1907. The energies of men. Science 25, 321–332.
Kwok, D., Buchl, S., Duckworth, A.L. Passion, perseverance, and perturbation: the relationship between grit and counterfactual thinking, manuscript in preparation.
MacArthur, M.S., 1955. An experimental investigation of persistence in secondary school boys. Canadian Journal of Psychology 9 (1), 42–54.
Mischel, W., 1968. Personality and Assessment. Wiley, New York.
Nation, J.R., Cooney, J.B., Gartrell, K.E., 1979. Durability and generalizability of persistence training. Journal of Abnormal Psychology 88, 121–136.
Peterson, C., 2000. The future of optimism. American psychologist 55 (1), 44–55.
Peterson, C., Park, C., 1998. Learned helplessness and explanatory style. In: Barone, D.F., Van Hasselt, V.B., Hersen, M. (Eds.), Advanced Personality. Plenum, New York, pp. 287–310.
Peterson, C., Park, N., Seligman, M.E.P., 2005. Orientations to happiness and life satisfaction: the full life versus the empty life. Journal of Happiness Studies 6 (1), 25–41.
Peterson, C., Seligman, M.E.P., 2004. Character Strengths and Virtues: A Classification and Handbook. Oxford University Press: American Psychological Association, New York. Washington, DC.
Pittenger, D.J., 2002. The two paradigms of persistence. Genetic, Social, and General Psychology Monographs 128 (3), 237–268.
Reed, J., Pritschet, B., Cutton, D.M., 2013. Grit, conscientiousness, and the transtheoretical model of change for exercise behavior. Journal of Health Psychology 18 (5), 612–619. http://dx.doi.org/10.1177/1359105312451866. Advance online publication.
Robbins, D., 1971. Partial reinforcement: a selective review of the alleyway literature since 1960. Psychological Bulletin 76 (6), 415–431.
Roberts, B.W., Chernyshenko, O.S., Stark, S., Goldberg, L.R., 2005. The structure of conscientiousness: an empirical investigation based on seven major personality questionnaires. Personnel Psychology 58 (1), 103–139.
Robertson-Kraft, C., Duckworth, A.L. True grit: trait-level perseverance and passion for long-term goals, manuscript submitted for publication.
Roese, N.J., 1997. Counterfactual thinking. Psychological Bulletin 121 (1), 133.
Ross, L., Nisbett, R., 1991. The Person and the Situation: Perspectives of Social Psychology. McGraw-Hill, New York.
Ryans, D.G., 1938a. A study of the observed relationship between persistence test results, intelligence indices, and academic success. Journal of Educational Psychology 29, 573–580.
Ryans, D.G., 1938b. An experimental attempt to analyze persistent behavior: measuring traits presumed to involve 'persistence'. Journal of General Psychology 19, 333–353.
Seligman, M.E., Schulman, P., 1986. Explanatory style as a predictor of productivity and quitting among life insurance sales agents. Journal of Personality and Social Psychology 50 (4), 832–838.
Von Culin, K.R., Tsukayama, E., Duckworth, A.L., 2014. Unpacking grit: motivational correlates of perseverance and passion for long-term goals. The Journal of Positive Psychology 9 (4), 306–312.
Webb, E., 1915. Character and intelligence. British Journal of Psychology 1 (Suppl. 3).
Willingham, W.W., 1985. Success in College: The Role of Personal Qualities and Academic Ability. College Entrance Examination Board, New York.

Grounded Theory: Methodology and Theory Construction

Kathy Charmaz, Sonoma State University, Rohnert Park, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article focuses on the grounded theory method of qualitative analysis and outlines its strategies for theory construction. The method begins with inductive inquiry but builds a series of checks into its strategies that increase the robustness of the findings. The intellectual heritage and development of grounded theory is presented followed by a depiction of its distinctive strategies. The concluding discussion describes the current and potential contributions of this method.

Introduction

The term 'grounded theory' refers to both the product of inquiry and the method of its construction, which this article features. Grounded theory is an inductive, iterative, interactive, and comparative method that provides systematic guidelines for gathering, synthesizing, analyzing, and conceptualizing qualitative data for the purpose of theory construction.

Grounded theory begins with inductive data that the researcher analyzes from the very beginning of data collection. It is an iterative method because early analysis informs further data collection, and grounded theorists go back and forth between data collection and analysis. This strategy helps grounded theorists to refine their emerging theoretical categories through successive levels of increasingly focused data collection and analytic development. Grounded theory is an interactive method because its system of checking categories and analytic progression keeps researchers interacting with their data, emerging analyses, participants, and later, current disciplinary conversations about their topics and analytic conceptions of these topics. Throughout the whole process, grounded theory is a comparative method. Grounded theorists engage in comparing data with data, data with codes of data, codes with codes, codes with theoretical categories, and these categories with categories. When taken to its logical extension, the entire process aids the grounded theorist to move the analysis to more abstract and theoretical levels.

This article offers an explanation of the method followed by a description of its intellectual heritage. Grounded theory is the most common form of qualitative analysis that researchers identify as being used (Bryant and Charmaz, 2007). However, what constitutes a grounded theory study and which versions represent the method remain contested. The author outlines major debates about the method and presents an overview of its strategies about which researchers grounded theorists would agree with. Current emphases are presented with note of the increasing internationalization of grounded theory.

Explanation of the Method

Grounded theory assumes that researchers can construct substantive and formal theories from systematic analysis of empirical data. By beginning with inductive data and successively developing abstract concepts and their relationships, grounded theory unites data and theoretical analysis. This method consists of systematic strategies to guide qualitative inquiry, particularly, data analysis. Grounded theory favors theory construction over description, collective patterns over individual narratives, developing fresh concepts over applying received theory, and theorizing processes over assuming stable structures.

Grounded theory has two crucial characteristics that enhance its capability for theory construction: flexibility and focus. The specific strategies of grounded theory allow researchers to adapt the method to the particular exigencies of their respective research problems. Simultaneously, the method fosters sustaining increasing focus on theoretical construction. The objectives of the method include (1) demonstrating that qualitative research can generate theory, (2) explicating and providing systematic strategies for collecting and analyzing data, (3) developing a method for studying processes, and (4) democratizing the practice of theorizing. The iterative approach of grounded theory provides tools for interrogating the data with increasingly conceptual questions, correcting errors and omissions, and refining analytic ideas. Grounded theory relies on constant comparative methods. Through conducting comparisons throughout inquiry, grounded theorists can identify, explicate, and analyze the properties of their emerging categories. Grounded theory has a unique feature of providing strategies for studying basic social and social psychological processes in natural settings.

Grounded theory has gained purchase throughout many disciplines ranging from its home in sociology to moving across the social sciences. Nursing adopted grounded theory as it professionalized at the doctoral level and pursued empirical research. The method has permeated other disciplines and professions as diverse as management, education, criminology, psychology, mathematics education, cultural studies, and kinesiology. Grounded theory has produced award-winning books and noted articles in many fields in which it has been adopted.

Many researchers identify grounded theory as their analytic approach, but do most engage in theory construction? No, relatively few researchers have taken grounded theory into explicit theory construction despite numerous claims to having done so. More accurately, grounded theory has been invoked to explain conducting qualitative research. Some researchers align themselves with grounded theory because they engage in inductive inquiry. Other researchers have treated citing the method as a convenient means of legitimizing their studies. Not only has the method been poorly understood but also its

 http://dx.doi.org/10.1016/B978-0-08-097086-8.44029-8

major proponents have disagreed about what stands as erosion or evolution of the method. Simultaneously, grounded theory has become a general method of analysis, and its strategies have been diluted and generalized as they have become routine practices in qualitative research. To further complicate this matter, meanings about what constitutes theory are varied and contested within and between disciplines. Despite some researchers' limited or superficial use of grounded theory, its strategies can help qualitative researchers increase the analytic power of their studies. Thus, this method still holds enormous potential for theory construction.

Intellectual Heritage of the Method

The founders of grounded theory, Glaser and Strauss (1967), offered the first systematic statement for analyzing qualitative data in their book, *The Discovery of Grounded Theory*. Prior to publication of their book, almost all discussions of qualitative research focused on data collection and attended to field researchers' membership roles. Grounded theory derives from the intellectual traditions of each of its founders. Strauss brought Chicago School pragmatism, symbolic interactionism, and field research to grounded theory (*see* Symbolic Interaction: Methodology; Field Observational Research in Anthropology and Sociology; Participant Observation). The writings of pragmatist philosophers impressed Strauss and permeated his mode of inquiry. Grounded theory underscored and strengthened several hallmarks of Chicago school sociology. Glaser and Strauss emphasized (1) conducting research in natural settings, (2) focusing on action and processes, and (3) assuming that actions and processes are emergent and somewhat indeterminate. In addition, Glaser and Strauss intended that grounded theory studies would offer their respective research participants a useful conceptual handle on their realities. Consistent with pragmatist notions of usefulness, Glaser and Strauss saw the potential usefulness of grounded theory research as a major contribution of the method.

Strauss's imprint on grounded theory can be found in its inductive beginnings, comparative methods, processual analyses, abductive reasoning, and emergent outcomes. In brief, grounded theorists invoke abductive reasoning to account for surprises, anomalies, or puzzles that they discover in the data. Abduction entails going back and forth between these empirical data and developing theoretical explanations of them. Abductive reasoning leads researchers to consider and test all conceivable theoretical explanations for the surprising findings. Abduction makes grounded theory more than solely an inductive method; it involves hypothesis testing as well as induction. Thus, the process of conducting research can take unanticipated directions.

Glaser's rigorous training in quantitative research at Columbia University gave grounded theory its systematic approach, positivist leanings, and procedural terms. Glaser aimed to codify steps of analysis of qualitative data in an analogous way as his mentor, Paul Lazarsfeld (Lazarsfeld and Rosenberg, 1956), had codified quantitative research. Glaser's efforts culminated in an explicit statement of how to handle data analysis (Glaser, 1978; 1998). Glaser's approach assumed a knowable world waiting to be discovered, unbiased observers, and a view of grounded theory categories as arising from the data. A cornerstone of Glaser's approach requires the researcher to avoid using preconceived theories and extant research literature to shape a grounded theory study. As a result of its divergent origins, grounded theory contains elements of both positivism and constructionism.

Grounded theory methods reflect the disciplinary arguments that had arisen in the 1950s and early 1960s. At that time, quantification had gained hegemony throughout the social sciences. Government granting agencies supported quantitative studies, and universities created research institutes for obtaining funds and pursuing quantitative research. Theory and research had become separate pursuits. The quest for quantified research findings overshadowed qualitative studies. Except for an occasional classic study, most quantitative methodologists ignored qualitative research and relegated it to disciplinary sidelines, or treated it only as a precursor to quantitative research. Quantitative researchers evaluated qualitative studies by the criteria for their research, which qualitative researchers could not meet. From these criteria, qualitative research was impressionistic, anecdotal, and idiosyncratic. Quantitative researchers often viewed the inability of their qualitative colleagues to generate verifiable facts and replicable studies as evidence of lack of rigor. Meanwhile qualitative researchers attempted to answer these criticisms by emphasizing the length of time they spent in the field, the depth of their knowledge of it, and their detailed descriptions of patterned relationships within the field.

In contrast, Glaser and Strauss (1967) argued that qualitative research had its own canons – and these must be applied to judge qualitative studies. Applying the canons of quantitative research imposed inappropriate criteria and failed to acknowledge contributions of qualitative inquiry to knowledge. Glaser and Strauss contended that qualitative research could stand as science in its own right, demonstrate rigor, and generate theory. Certainly, their works on the organization of dying buttressed these claims. Furthermore, Glaser and Strauss proposed a remarkably democratic view of theory construction by bringing it into the purview of working researchers. A sharp divide between theorists and researchers had existed in 1950s and 1960s sociology and only a few elite sociologists constructed theory. They developed theories consisting of abstract conjectures about the structure of whole societies rather than building on studies of empirical problems. Glaser and Strauss made a strong case for constructing theories that focused on the empirical world and thus linked theory and research.

Glaser and Strauss's methods appealed to social scientists who wished to conduct qualitative research but lacked tools for doing it. In the late 1960s and throughout the 1970s, grounded theory methodological rationales contributed significantly to reestablishing the legitimacy of qualitative research. This method provided researchers with ready justifications for conducting qualitative studies and strong rationales for claiming that their methodological approach was systematic. Not surprisingly, diverse qualitative researchers still claim to use grounded theory to establish their credibility and the legitimacy of their research enterprise. However, their claims of using grounded theory may rest on having adopted some form of qualitative research or on only following the initial steps of the grounded theory method.

The Discovery of Grounded Theory laid out the principles of grounded theory with remarkable success. It succeeded less as a manual for handling the concrete realities of conducting research. Glaser viewed the book as a major methodological statement; Strauss, whose lifelong connections to the discipline of sociology remained much stronger than Glaser's, appreciated its rhetorical power but was not as committed to central ideas such as entering the research as a naïve learner. Neither author foresaw the vast influence the book would have across disciplines.

Developments and Changes in the Method

With publication of Strauss's (1987) manual, *Qualitative Analysis for Social Scientists*, Glaser and Strauss's divergent directions for grounded theory began to surface. Strauss's approach maintained the inductive, iterative characteristics of their original statement but took a looser form than Glaser' (1978) had articulated in *Theoretical Sensitivity*. In this book, Glaser set forth his concept-indicator model of grounded theory and coding processes. However, the book was densely written, unindexed, and assumed prior knowledge of grounded theory that many aspiring qualitative researchers lacked. Strauss's (1987) book and his later volumes with Juliet Corbin (Strauss and Corbin, 1990; 1998) brought their version of the method to large audiences across the globe. The dual and not altogether complementary foundations of the method indicated differences from its beginnings that grew increasingly more incongruent.

Strauss and Corbin's (1990) book explicated how to conduct a grounded theory study but introduced new procedures to be applied to the research. Despite the authors' efforts to produce an accessible manual for novices, this book changed grounded theory in fundamental ways because it (1) reduced the method to procedures and thus minimized its flexibility, (2) introduced new preconceived ways of handling data that skewed the resulting analysis, and (3) deemphasized the development of emergent theoretical categories. Glaser (1992) had justifiable criticisms of the book, but his acrimonious response divided proponents of grounded theory into two camps, Glaser's objectivist method and Strauss and Corbin's postpositivist approach.

Charmaz's (2000) analysis of objectivist and constructivist grounded theory identified the positivist strains in both versions of grounded theory and proposed a new approach, constructivist grounded theory. Her approach not only builds on social constructionism of the 1970s and 1980s, but, unlike earlier forms of social constructionism, also addresses the construction of the research project and takes into account the subjectivity of the researcher as well as other research participants (*see* Constructivism/Constructionism: Methodology). Constructivist grounded theory maintains the useful methodological strategies of earlier versions of the method theory but frames them in a twenty-first century ontology and epistemology and addresses methodological developments that had occurred since the 1960s. As a result, it brought into purview relationships between the viewer and the viewed, fact and value, and the conditions of research and its products. The constructivist version treats grounded theory strategies as tools for researchers to use and adapt to fit their empirical problems. Hence, this version is less prescriptive than either earlier version of grounded theory. The constructivist revision challenges earlier notions that researchers can enter their studies uninfluenced by earlier studies and their own backgrounds and interests. The constructivist emphasis on interaction acknowledges that conceptualizations of data result from researchers' interpretive frames as well as what happens during their research and what they learn about their topics. Thus, constructivists view their analyses as emergent interpretive constructions of data rather than inhering within data.

Before the constructivist turn, grounded theorists like many qualitative researchers, had not examined their assumptions about data collection, themselves as observers and writers, and the situated nature of their studies. Adele Clarke (2005) has made these concerns central to her explication of situational analysis, which she developed to advance grounded theory. Situational analysis shares assumptions with constructivist grounded theory and builds on both Strauss's sociology and poststructuralist critiques. Before constructivist grounded theory and situational analysis, earlier grounded theorists had attempted to remain objective, neutral, and passive. They did not realize that this approach constituted a particular standpoint rather than a means of ensuring that their values remained separate from their data. Examining their own assumptions – personal and political – as well as those about the topic and research participants had not become a prerequisite for conducting grounded theory. Thus all kinds of unquestioned assumptions, such as those concerning race, gender, social class, age, and culture, along with tacit personal and professional views of the research problem and participants could enter the research.

Constructivist grounded theory and situational analysis brought new concerns into purview such as reflexivity representation of research participants, and silences into purview as central methodological considerations that have implications for substantive and formal theoretical construction. Clarke (2005) constructs ways that researchers can discover organizational silences and thus helps to move grounded theory beyond the study of individuals. Earlier grounded theory approaches had emphasized overt actions and statements rather than covert actions and implicit meanings. In recent statements about her methodological development, Juliet Corbin (Corbin, 2009; Corbin and Straus, 2008) discusses how her reexamination of method made her previous assumptions about conducting research problematic. Subsequently, she has arrived at a position in many ways consistent with that of constructivist grounded theory.

Grounded theory is gaining new spokespersons. The second generation of spokespersons for grounded theory (Morse et al., 2009) was trained by its originators, but each has adapted the method in his or her own way. For example, Phyllis Stern (2009) follows Glaserian grounded theory but brings feminist and cultural sensibilities to it. Now international researchers are joining the effort to articulate grounded theory for their colleagues. In nursing, Melanie Birks and Jane Mills (2011) offer a basic text that attends to indigenous research. In education, Robert Thornberg (2012) provides sophisticated arguments for using extant literature and advancing theoretical construction.

Conducting Grounded Theory Research

Simultaneous data collection and analysis shape grounded theory studies. Early qualitative data coding and categorizing leads grounded theorists to gather material that fits the empirical world and raises questions about it, the substantive area of inquiry and the theoretical directions of the study. By adopting Glaser's (1978) questions concerning what the data suggest that this study is about and which theoretical category a specific datum indicates, researchers begin to think theoretically about their data. Invoking analytic and theoretical questions throughout the research process focuses inquiry and gives researchers analytic leads to pursue and directions for finding data that can inform these leads.

Three major strategies of the grounded theory method are introduced here: coding, memo writing, and theoretical sampling. These strategies make grounded theory distinctive from other kinds of qualitative analysis, although they have become more diffuse and generalized as they have entered the common lexicon of qualitative inquiry. Most qualitative researchers engage in coding and memo writing but do not adopt the distinctive features of these grounded theory strategies that enable researchers to advance their work into theory construction.

Coding Qualitative Data

Grounded theorists begin coding data from the start of data collection. Original conceptions of grounded theory required researchers to avoid applying theories and ideas to their data from relevant literature. Constructivist grounded theorists view this position as untenable and naïve because both novice and seasoned researchers already have a fund of knowledge and experience. Hence, constructivist grounded theorists are less troubled by researchers being aware of ideas about the research topic before entering the field than their objectivist counterparts. However, constructivists argue that researchers must become aware of their standpoints and starting points and subject them to rigorous scrutiny.

Constructivists often begin with sensitizing concepts (Blumer, 1969). These concepts are loosely defined, general concepts, rather than those that have specific, precise definitions. Constructivist grounded theorists invoke sensitizing concepts as a beginning lens rather than a means of application. However, consistent with earlier versions of grounded theory, constructivists study their data and ask, 'What is happening here?' Researchers can use conducting line-by-line coding as a heuristic device to define what they see in their data. Using gerunds for coding helps researchers to see actions and to begin to theorize processes.

A major strength of grounded theory is that its strategies foster analyzing processes. Initial line-by-line coding also helps researchers to identify implicit meanings and actions early in the research. Charmaz (2006) advises researchers to make their codes short, simple, active, and analytic. Early line-by-line coding can enable researchers to begin conceptualizing basic processes occurring in their data by breaking bits of data into specific types of action. From studying the segments, grounded theorists may conceptualize a major process. The identification of a process allows the researcher to begin to theorize its properties and the conditions under which it develops, is maintained, and changes. Line-by-line coding gives the researcher some codes to pursue further. The researcher can use the most frequent and/or significant codes in focused coding to test whether and to what extent these codes account for large amounts of data.

Glaser (1978) first advocated line-by-line coding but more recently favors incident-by-incident coding in which he compares various incidents in the data with each other. Nonetheless, both approaches are comparative and focus on analytic development.

Memo Writing

Memo writing is an intermediate stage of writing that bridges coding data with drafting the theoretical analysis. Early memos typically begin with examination of telling codes. Grounded theorists use these codes to title their memos. Memo writing is a way for researchers to take their codes apart and study what constitutes them. Memo writing is a crucial part of the process of theoretical construction because grounded theorists define major aspects of theorizing through constant comparisons of data with other data, data with codes, codes with codes, codes with categories, and categories with categories. Memos record and preserve the comparative process and the discoveries the researcher makes from them. As researchers engage with the data and their nascent analyses, new analytic questions arise that spawn further scrutiny of data and more memos.

Throughout memo writing, researchers bring raw data into the memo and explicate how these data inform the emerging analysis. By integrating data and analysis, researchers construct sections of their reports as they proceed. Hence, they can keep fine-tuning their memos to give them greater analytic power.

Several steps move the research forward analytically. First grounded theorists define codes by their empirical properties. Distilling a code to its essential properties simultaneously condenses its description and allows the researcher to treat it analytically. Second, grounded theorists begin to ask under which conditions these properties are present or absent. Third, grounded theorists treat codes as tentative categories and interrogate their analytics significance through memo writing. Third, memo writing helps to identify gaps in the data. Subsequently the researcher can return to the field and gather the missing data. Fourth, through memo writing, researchers can ascertain the relationship between codes and categories and subsequently gain a sense of how the analytic pieces fit together. Fifth, memos also give constructivist grounded theorists a place to examine their own assumptions and implicit meanings, as well as those of their research participants. This examination may reveal liminal meanings related to the codes and tentative categories. Sixth, in addition to delineating and analyzing codes and categories, memo writing both sharpens the analysis and helps the researcher to correct problems such as making overgeneralizations and leaving his or her assertions unchecked. In short, memo writing is the field in which grounded theory discoveries germinate, take root, and grow.

Theoretical Sampling

Theoretical sampling is a pivotal part of theory construction in grounded theory but one of its most misunderstood strategies. The term confuses some researchers because the term 'sampling' ordinarily refers to specific populations that researchers intend to investigate. In contrast, theoretical sampling means sampling to develop, refine, or fill out the properties of tentative theoretical categories. Theoretical sampling makes these categories more robust and precise. What and who the researcher needs to sample cannot be known before the research begins. To engage in theoretical sampling, researchers must have a tentative category or categories. Thus, theoretical sampling follows focused coding, and researchers often discover the need for it through memo writing. Theoretical sampling simultaneously builds analytic complexity into the researcher's developing theory and keeps it grounded in data.

When engaging in theoretical sampling, researchers gather data to answer specific questions about the properties of their theoretical categories. Researchers may also use theoretical sampling to ascertain the range of variation of a category or process. Theoretical sampling increases the conceptual level of the analysis, specifies its theoretical reach, and extends its usefulness. Theoretical sampling is related to abductive reasoning. When researchers discover a surprising finding, they need to account for it. Consistent with the logic of abductive reasoning, grounded theorists attempt to entertain all possible theoretical explanations for this finding. Subsequently, they need to check and confirm the most plausible theoretical explanation for this surprising finding and thus engage in theoretical sampling.

One of the difficulties of theoretical sampling is that it can entail conducting comparisons across groups or research settings for which the researcher must justify with human subjects review committees and gatekeepers for these groups and field sites.

Theoretical sampling allows researchers to 'saturate' their theoretical categories. Saturation occurs when researchers find no new properties of these categories. At this point grounded theorists stop collecting data. The definition of saturation leaves ample room for claims-making among researchers as the criteria for saturated properties is at best ambiguous. However, saturation of theoretical categories involves a considerably more exacting engagement with data and theoretical construction than saturation of themes or data, with which it is sometimes confused. In the latter case, researchers report that same themes reoccur, or research participants report similar views and experiences. This type of saturation contrasts with the saturation in grounded theory that researchers accomplish through theoretical sampling.

Current Emphases and Future Directions

Over the past decade, four emphases have emerged that bode well for further developments in grounded theory. First, the method continues to make major contributions to qualitative research in the social sciences and beyond. The constructivist and postpositivist versions of grounded theory have earned credibility across the globe, which recent translations of their key works have buttressed. Their major spokespersons view their versions as having distinctive grounded theory characteristics but as complementing and extending trends in qualitative research. In addition, both these versions acknowledge the symbolic interactionist and pragmatist roots of the method and the legacy of Anselm Strauss. Protagonists of objectivist grounded theory, in contrast, treat the origins and development of the method as wholly Glaserian and emphasize their differences from constructivist and postpositivist grounded theory – as well as from qualitative inquiry.

Second, renewed interest in pragmatism and the symbolic interactionist underpinnings of grounded theory have sparked new discussions of the role of abduction in the method. As more researchers aim toward theory construction, engaging in abductive reasoning becomes a crucial element in constructing fresh theoretical analyses of the studied empirical problems. The most helpful way to use abduction is to expand the range of theoretical possibilities and subsequently confirm which hypothesis offers the best theoretical account for the data. However, grounded theorists must not let the existing range of theories inhibit constructing new theories or initiating new directions of inquiry.

Third, grounded theory is taking an increasingly significant role in mixed methods research. Grounded theory methods fit both the logic and practice of mixed methods. Until recently, the definition of mixed methods assumed that studies included both qualitative and quantitative components. A mixed methods study may be driven by either component, although quantitatively driven studies are the most common. In both types of studies, methods are likely to be added rather than mixed in actual research practice. For quantitatively driven studies, grounded theory can offer ways to (1) fine-tune instrument design; (2) check quantitative findings; (3) deepen empirical knowledge; (4) address participants' views and experiences; (5) provide a conceptual framework for the study, as well as meeting such practical concerns to improve clinical trials; (6) identify and reduce investigator biases; and (7) fulfill funding agencies' goals. In qualitatively driven grounded theory studies, the method has always fostered gathering the kinds of data pertinent to the research problem and the merging questions it raises, although few grounded theorists have defined themselves as mixed methods researchers.

Fourth, the constructivist turn has supported research in social justice. The constructivist emphasis on language and meaning heightens awareness of different definitions and the resulting actions such definitions encourage. Similarly the constructivist awareness of multiplicity of perspectives and multiple realities sensitizes grounded theorists to look beyond the surface and to study what happens among diverse research participants. All these concerns fit researching the effects of ideologies, hierarchies, and uses of power. Grounded theorists are poised to construct useful theories in social justice inquiry that offer conceptual handles on specific issues and may generate persuasive changes in policy arenas.

Last, the constructivist turn placed grounded theory on new epistemological and ontological foundations. It also developed a clear statement of grounded theory strategies that preserves the original grounded theory emphasis on emergent theory construction. Grounded theory has always focused on analytic

strategies. Now Charmaz (in press) addresses data collection and its implications for developing grounded theory studies. This direction will help new and seasoned researchers to integrate all aspects of their developing grounded theory studies. Such efforts are likely to be taken up by other researchers who see the advantages of grounded theory strategies and wish to contribute to strengthening them.

See also: Constructivism and Constructionism: Methodology; Content Analysis; Interpretive Methods: Micromethods; Qualitative Methods, History of; Symbolic Interaction: Methodology; Triangulation: Methodology.

Bibliography

Birks, Melanie, Mills, Jane, 2011. Grounded Theory: A Practical Guide. Sage, London.

Blumer, Herbert, 1969. Symbolic Interactionism. Prentice-Hall, Englewood Cliffs, NJ.

Bryant, Antony, Charmaz, Kathy, 2007. Introduction. In: Bryant, Antony, Charmaz, Kathy (Eds.), Handbook of Grounded Theory. Sage, London, pp. 1–28.

Charmaz, Kathy, 2000. Grounded theory: constructivist and objectivist methods. In: Denzin, Norman K, Lincoln, Yvonna S (Eds.), Handbook of Qualitative Research, second ed. Sage, Thousand Oaks, CA, pp. 509–535.

Charmaz, Kathy, 2006. Constructing Grounded Theory: A Practical Guide through Qualitative Analysis. Sage, London.

Charmaz, Kathy, (in press). Constructing Grounded Theory: A Practical Guide through Qualitative Analysis, second ed. Sage, London.

Clarke, Adele E, 2005. Situational Analysis: Grounded Theory after the Postmodern Turn. Sage, Thousand Oaks, CA.

Corbin, Juliet, 2009. Taking an analytic journey. In: Morse, Janice M., Stern, Phyllis N., Corbin, Juliet, Charmaz, Kathy, Clarke, Adele E. (Eds.), Developing Grounded Theory: The Second Generation. Left Coast Press, Walnut Creek, CA, pp. 35–53.

Corbin, Juliet, Strauss, Anselm, 2008. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, third ed. Sage, Thousand Oaks, CA.

Glaser, Barney G, 1978. Theoretical Sensitivity. Sociology Press, Mill Valley, CA.

Glaser, Barney G, 1992. Emergence vs. Forcing: Basics of Grounded Theory Analysis. Sociology Press, Mill Valley, CA.

Glaser, Barney G., Strauss, Anselm L, 1967. The Discovery of Grounded Theory. Aldine, Chicago. Sage, London.

Glaser, Barney G., 1998. Doing grounded theory: Issues and discussions. Mill Valley, CA: Sociology Press.

Lazarsfeld, Paul F., Rosenberg, Morris (Eds.), 1956. The Language of Social Research. Free Press, Glencoe, IL.

Morse, Janice, Stern, Phyllis, Corbin, Juliet, Bowers, Barbara, Charmaz, Kathy Clarke, Adele E. 2009. Developing Grounded Theory: The Second Generation. Left Coast Press, Walnut Creek, CA

Stern, Phyllis. 2009. Glaserian grounded theory. In: Morse, Janice, Stern, Phyllis, Corbin, Juliet, Bowers, Barbara, Charmaz, Kathy, Clarke, Adele E., Developing Grounded Theory: The Second Generation. Left Coast Press, Walnut Creek, CA.

Strauss, Anselm L, 1987. Qualitative Analysis for Social Scientists. Cambridge University Press, New York.

Strauss, Anselm L., Corbin, Juliet, 1990. Basics of Qualitative Research: Grounded Theory Procedures and Techniques. Sage, Newbury Park, CA.

Strauss, Anselm L., Corbin, Juliet, 1998. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory Procedures and Techniques, second ed. Sage, Thousand Oaks, CA.

Thornberg, Robert, 2012. Informed grounded theory. Scandinavian Journal of Educational Research 56 (3), 243–259.

Group Processes in Organizations

Susanne Braun, Ludwig-Maximilians-Universität München, Munich, Germany; and Technische Universität München, Munich, Germany
Dieter Frey and Felix Claus Brodbeck, Ludwig-Maximilians-Universität München, Munich, Germany
Tanja Hentschel, Technische Universität München, Munich, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D. Frey, F. Brodbeck, volume 9, pp. 6407–6413, © 2001, Elsevier Ltd.

Abstract

The article refers to group processes as they influence organizational functioning. The main features of groups (e.g., norms, cohesion, roles), and factors contributing to their performance are characterized. We detail well-established concepts like groupthink and social identity theory. Further, we refer to important current topics in organizational practice and research related to group functioning, like team mental models, diversity in groups, and counterproductive work behavior. Finally, hints at the practical relevance in terms of the implementation of these concepts in organizational contexts are provided (e.g., organizational change, leadership in multinational corporations, and shared leadership).

General Introduction

Success and failure in organizations depend not only on individual knowledge, skills, and abilities but also on the quality of the interaction and communication within and between organizational groups. Group processes are intertwined with many parameters at individual levels (e.g., motivation and commitment) and organizational levels (e.g., mergers and acquisitions, and error prevention). Group processes are therefore of high relevance to organizational functioning. Moreover, the study of groups is a central subject of social psychology as well as organizational psychology. During the 100-year history of group research, a substantial amount of literature has been amassed. In the following, we concentrate on a selection of topics, which we believe are of particular relevance to the understanding of group processes in organizations. The topics are in part traditional and long established, and in part innovative and future-oriented.

Definition and Structural Dimensions of Groups

Groups are defined as a number of individuals who are oriented toward the same or similar goals and who interact and communicate in order to achieve these goals. All groups share at least three structural principles: norms and values, a certain degree of cohesion, and role differentiation.

Norms and Values

Norms can be defined as the shared expectations of group members about how to act in certain situations. Values comprise a set of shared beliefs, which represent the ethical ideals of the group, like interpersonal fairness and openness. Both norms and values guide the behavior of group members, make their behavior more predictable, and thus help to clarify social relationships.

Degree of Cohesion

When group membership is highly attractive – because individuals assume that goal achievement and need satisfaction are more likely in the group than individually – group members are motivated to sustain their contact with the group. The sum of all forces that increase group attractiveness is called cohesion.

Role and Status

Role differentiation in groups is the social-psychological basis for joint work. Roles are defined as the expectations of group members about appropriate behavior in a certain position. Role-conformant behavior is sanctioned positively, whereas role deviation is sanctioned negatively. Status means that group members differ in the hierarchy of prestige, power, influence, and privileges. High-status group members as compared to low-status group members experience fewer restrictions when attempting to deviate from group norms (idiosyncrasy credit).

Conformity and Dissent

Conformity and dissent are central to group functioning and goal achievement. Behaving in accordance with group norms is required to interact at low levels of conflict. Usually, group pressure aligns group members' behaviors and attitudes with group norms, which can be functional for goal achievement. However, normative behavior can also be dysfunctional, as has been shown by Solomon Asch's (1951) studies in which group members expressed clearly wrong and irrational perceptions due to group pressure toward conformity.

Groupthink

Groupthink describes the phenomenon by which existing information that is relevant to a certain decision is not processed by group members in a rational and systematic way. Instead, information processing is biased in the direction of a preferred decision alternative within the group. This bias increases the risk of faulty decisions. Analyzing historical political disasters following group decisions, such as the American invasion of Cuba, Janis (1982) specified symptoms of groupthink: (1) Illusions of invulnerability; that is, high

conformity of group members' views and thinking styles makes them overconfident. (2) Willingness to take high risks leads group members to not take negative feedback seriously; that is, group members avoid questioning their own basic beliefs when facing a challenging decision. (3) Stereotyping of members with a different opinion by labeling them as incompetent. (4) Pressure toward conformity exerted by the formal leader or the majority of the group; that is, subsequent self-censorship processes make group members (falsely) assume that silence means approval. The risk of groupthink is amplified when groups are homogeneous (see Diversity in Groups), when group structures are markedly hierarchical, and when a strong commitment to particular decision alternatives exists before group discussion begins. Considerable empirical evidence deals with prevention and intervention strategies to counteract groupthink (for an overview see Frey et al., 1996). For example, effects of groupthink are reduced if (1) the group leader acts as an impartial coordinator, (2) self-confident minority members are encouraged to speak up, (3) opinions of external experts are integrated into the group decision, (4) artificial conflict is implemented in the group decision process (e.g., devil's advocacy or dialectical inquiry techniques), (5) each group member is held individually responsible for the decision process and its outcome, and (6) different subgroups are created to separately engage in decision making. Symptoms and consequences of groupthink are observable in many applied contexts, such as work in project teams or decision making on organizational boards. Therefore, the concept of groupthink has been frequently applied to explain corporate or political failure (e.g., accounting fraud; Scharff, 2005).

Team Mental Models

Team mental models are defined as organized mental representations, which group members share about key elements in their environment (for an overview see Mohammed et al., 2010). These mental representations refer to task-related contents (i.e., which tasks need to be completed) as well as team-related contents (i.e., how task completion is best achieved through cooperation). Two major aspects further characterize team mental models: (1) their similarity among group members and (2) their accuracy. High levels of similarity imply that group members hold compatible mental models, which in turn results in common expectations for task completion and interpersonal cooperation. Accuracy of team mental models is defined by the degree to which shared knowledge and belief structures reflect existing conditions in the organizational environment. Team mental models have been studied predominantly under the assumption that they contribute to organizational effectiveness. However, overly shared knowledge and belief structures may result in phenomena comparable to groupthink (see Groupthink) when group members oppose revising outdated or incorrect mental models because the group shares them.

Minority Influence

Conformity and homogeneity in groups may come at the expense of divergent thinking. Due to conformity pressure, the default information processing style in groups is consent oriented. Minority influence, however, often generates productive dissent. According to Moscovici (1976), minority members are able to oppose the pressure toward conformity in groups under certain conditions – and are thereby able to influence attitudes and behaviors of the majority. In doing so, minority members are required to present their point of view in a consistent manner, though not rigidly, while signaling their willingness to compromise. Thereby, minority members trigger an internal cognitive conflict for majority members. This conflict in turn leads to majority members' intensified information processing and is accompanied by an increase in divergent and creative thinking styles, less biased information search, and greater openness for the minority position. In contrast, majorities exert a pressure to conform, which produces social conflict between individual group members or subgroups rather than cognitive conflict. This pressure leads to more superficial processing of information and is accompanied by conventional and conformist styles of thinking. Minority influence in organizations is a precondition for scrutinizing common patterns of thought and practice, and thus represents an important condition to organizational learning, change, and innovation. Recent research shows that the impact of minority opinions in organizations is significantly moderated by motivational and structural factors, like social motives and decision rules underlying negotiations (e.g., Velden et al., 2007).

Diversity in Groups

Diversity is defined as an actual or perceived difference in salient attributes of a given group. Work groups have become increasingly diverse over the last decades: The retirement age went up in several countries, more women engage in full-time jobs, and individuals from different countries are often part of the same work group. Groups with diverse members offer advantages, but also face more difficulties than work groups composed of members who are similar to each other.

Gains and Losses of Diversity

In their thorough review of 40 years of diversity research, Williams and O'Reilly (1998) discuss the impact of demographic diversity on group processes and performance. They present three main theoretical perspectives: (1) The similarity/attraction paradigm; that is, similarity between individuals fosters social attraction. Hence, groups with similar members will collaborate better than groups with diverse members. (2) The social categorization perspective; if individuals in a group are different or perceive themselves to be different from other group members, categorization will occur. The subsequent formation of subgroups will lead to more conflict and less cohesion. However, following the (3) information/decision-making perspective, heterogeneous groups will integrate diverse perspectives, experiences, and task-relevant information, which in turn enhance creativity and task performance. Overall, as research shows, if and when diversity leads to desired outcomes is significantly influenced by contextual variables and diversity dimensions (for an overview see van Knippenberg and Schippers, 2007).

Diversity Dimensions

Group members may differ from each other on several dimensions. The most prominent dimensions are group tenure, age, educational background, gender, and ethnicity. In work groups with high tenure diversity, members with either lowest or highest tenure as compared to other group members are more likely than others to leave the group, often because they experience more group conflict. Hence, these group members will take their knowledge, skills, and experiences along. In contrast, age diversity is hardly associated with an increase of relationship conflict within the group, but can still have negative consequences for turnover. Furthermore, groups that consist of members with functionally diverse educational backgrounds experience more conflict about the task at hand than homogeneous groups. However, this type of conflict as compared to relationship conflict eventually leads to better problem solving. When group members are diverse in terms of gender, it depends on the proportion of men and women in the group whether gender diversity leads to positive or negative performance results: In groups with high numbers of men and only few women, group outcomes are generally lower than in gender homogeneous groups. Yet, the presence of few women in top positions paves the way for the advancement of other women to leadership positions. In groups with comparable numbers of men and women, and in groups with many women and few men, diversity does generally not lead to negative team outcomes. Ethnic diversity has similar effects: When there is a clear minority, ethnic diversity will lead to more conflict or turnover. However, when the group is very heterogeneous, the salience of differences in ethnicity diminishes and group performance increases. In addition to demographic diversity, so-called deep-level differences of group members, such as diversity in values or work motivation, become pivotal for group outcomes over time. When group members work together for a longer period of time, their deep-level diversity leads to negative performance outcomes, whereas the impact of demographic diversity wears off.

Creativity and Innovation

In many cases, diversity is not advantageous for groups when they need to solve simple and routine tasks. However, diversity is beneficial when innovation and creativity are required for problem solving because members of heterogeneous groups are more likely to engage in in-depth information processing (see Minority Influence). Moreover, when group members hold positive as compared to negative beliefs about diversity as well as when they *perceive* themselves to be similar to each other, diversity is more likely to lead to desired outcomes (Hentschel et al., 2013). Thus, organizations are required to emphasize a culture that values its diverse work force when striving for creativity and innovation. Constructive communication in heterogeneous groups is another important factor to stimulate creativity. Open-mindedness toward diverging opinions and experiences enhances innovation, idea transfer, and continuous optimization in groups. West (2002) describes group members' need to experience a feeling of psychological safety to express diverse views. Additionally, he specifies the importance of managing task conflict adequately and reflecting on team processes, goals, and strategies as major parameters to advancing innovation. Team reflection serves to raise group members' awareness of their group's strength as well as weaknesses and potential for improvement. Behaviors of the group's leader are also of high importance: Clear goals and supervisory support are profitable for idea generation, whereas too much control impedes creative thinking.

Integration of Individual Contributions

Organizational successes rely heavily on group performance, that is, the integration of individual contributions in a way such that production losses (e.g., uncoordinated action, demotivation, interpersonal conflict) are minimized and procedural gains or synergies (e.g., transfer of motivation, mutual learning, knowledge sharing) are maximized. Theory and research on group performance have undergone significant developments, leading to new trends like research on team mental models, collective emotions, and group faultlines in addition to traditional predictors of group performance (for an overview see Kerr and Tindale, 2004). We first address the concept of synergy in groups, then turn to antecedents of high-level group performance, and shed light on counteracting influences.

Synergy in Groups

The term group synergy suggests that a group is more than the sum of its parts, or more precisely, that a given group generally performs better than its most talented single member. Yet, Stroebe et al. (1992) argued that the common belief in group synergy is grounded in individuals' tendency to overestimate their success in group contexts. For example, where brainstorming is concerned, individuals usually believe that they are more productive when working in groups – probably because they feel 'enriched' by the ideas of others. However, individual group members are inclined to claim other group members' ideas as their own body of thought and, hence, to overestimate their own contribution. Yet, other phenomena have been described in the literature, which speak in favor of process gains and group synergy rather than process losses in groups. First, group learning will occur over time, if several individuals work interactively on the same task. That is, skills and knowledge will be transferred between individual group members, from individuals to the group, and vice versa, as well as at the group level. Second, joint problem solving in groups holds the potential for motivation gains through social competition (i.e., group members seeking to outperform each other), social compensation (i.e., stronger group members compensating for weaker group members' performance), or increased feelings of responsibility (e.g., weaker group members working harder than they would do individually to avoid being held responsible for a weak group performance, the so-called Köhler Effect). Third, initial dissent reduces the probability that group members prematurely agree on suboptimal decisions because conflicting opinions intensify and 'debias' discussions about existing, but hidden information. Overall, synergy in groups requires systematic activities of group performance management (for details, see Schulz-Hardt and Brodbeck, 2012).

Group Performance and Type of Task

A common approach to evaluate group performance is to compare actual group performance to potential group performance (or group potential). Group potential is the performance that would have occurred if the members of a group had worked individually on a given task. When actual group performance diverges from group potential, this is due to process losses and process gains. This relationship is expressed in the following formula (Hackman and Morris, 1975): Actual group performance = group potential − process losses + process gains. In order to determine whether and to what extent a group is working below, above, or at its full potential, actual group performance must be obtained and compared to the potential level of performance. The difference reflects the extent to which group-induced process gains and losses have occurred.

Further, different types of group tasks have been distinguished (Steiner, 1972), which hold significant implications for group potential. The structure of a task may be additive, in which case the potential group performance is the sum of all individual performances, for example, in brainstorming. If a task is structured in a conjunctive way, it requires all group members to be successful for the group to complete the task. In this case, the group potential is determined by the individual performance of the group's weakest member, for example, when climbing a mountain in a rope team. If a task is disjunctive, the group has to decide on one particular solution to a problem. The group potential is the result of the best individual performance, for example, in mathematical calculations.

Top Performance and Organizational Culture

Despite antagonistic forces such as conformity, groupthink, and motivation or coordination losses, top performance in organizations does occur. In particular, organizational cultures of excellence promote top performance based on the following values (Peus and Frey, 2009):

- Problem solving for continuous learning and improvement, for instance, through self- and other-reflection (i.e., what is good and should be sustained, what is bad and should be improved)
- Mistakes as learning opportunities with an orientation toward best practice ideas and benchmarking (i.e., comparing the group with best practices in the organization, in the branch, with national and international competitors)
- Constructive confrontation and conflict solving (i.e., the demand for and encouragement of thinking outside the box, constructive stubbornness, speaking up)
- High commitment to excellent and innovative output (i.e., rules and regulations are minimized, and the organization is willing to take risks in order to motivate visionary thinking)

The homogeneity in norms, values, and assumptions creates organizational cultures, which enable the systematic integration of individual contributions from group members equipped with a diverse set of talents, experiences, and educational backgrounds (Schein, 1990).

Counterproductive Work Behavior

As opposed to the factors just presented, certain aspects in groups significantly hinder organizational performance, among them partly involuntary actions like the inability to detect and correct human errors, and but also deliberately harmful behavior in organizational settings.

With regard to counterproductive work behavior (CWB), O'Boyle et al. (2011) proposed an integrative multilevel model of individual, group, and organizational antecedents based on a review of earlier theory and research. CWB captures a wide range of deliberate actions that harm the organization (e.g., theft, sabotage) or its members (e.g., bullying) by undermining interpersonal relationships and performance. These actions may vary in severity as well as underlying motivation (e.g., anger, revenge, narcissism). With reference to work groups, the following predictors of CWB are suggested:

- Supervision (e.g., abusive or laissez-faire)
- Group norms (e.g., acceptance of drinking at work or absenteeism)
- Group processes (e.g., relational conflicts resulting in hostile atmosphere)
- Dissimilarity (e.g., dissimilar moral values resulting in reduced compliance to ethical norms)

Due to the severe consequences of CWB for organizational functioning, it is advisable for practitioners and researchers alike not only to focus on the presence or absence of positive predictors for group performance but also to detect active detrimental influences.

Group Dynamics

Interaction in organizational groups is significantly shaped by the development of interpersonal relationships and roles. Such dynamic processes haven been considered from different theoretical perspectives. First, from the perspective of socioemotional relations, cohesion, role formation, communication, and building of trust are relevant. Linear models predict the development of socioemotional relations in groups over time. Second, from a perspective of goal-directed and output-oriented group actions, goal setting, role differentiation, leadership, decision making, and problem solving are necessary for positive group dynamics. In the following, we present stages of group development, which are related to both, interpersonal and task characteristics, and introduce central concepts related to intragroup conflicts as well as group processes in intercultural settings.

Group Stages

Typically, groups in organizations undergo several phases of development from their initial formation to a joint functioning. Established models propose five stages of group development (Tuckman and Jensen, 1977):

- Forming: Group members discover norms of interpersonal behavior and structures of interdependence, and identify their group tasks.
- Storming: Intragroup conflict arises because group members express their individuality and resist formation of stable group structures.

- Norming: Individuals accept group membership, and group cohesion develops. The group becomes an entity protected by the development of group norms. Task conflicts are avoided in favor of group harmony.
- Performing: Interpersonal relationships have been established, and the group has developed into a problem-solving unit. Members take over specific roles to enhance the group's task functioning, and to solve task-related problems.
- Adjourning: As a project terminates, dissolution of the group takes place, wherein tasks are completed, roles are given up, and interdependence of group members is reduced.

A more recent model by Hackman and Wageman (2005) proposes three group stages along a task cycle from its beginnings (i.e., the group launch), to the midpoints (i.e., after making first experiences with group and task), and to the ends (i.e., when the task or relevant subtasks have been accomplished). The model posits that coaching interventions adapted to each of these stages will enhance team effectiveness. Effort-related (motivational) interventions apply at the beginning, strategy-related (consultative) interventions are effective near the midpoint, and (educational) interventions, which address knowledge, skills, and lessons learned, apply at the end of a task cycle.

Intragroup Conflicts

In early stages of theory and research, conflicts in groups have been described as detrimental to organizational functioning. Yet, within the last decades, the field of intragroup conflict has developed considerably, such that it is now providing more comprehensive insights into the different types of conflict and their consequences (for an overview see De Wit et al., 2012). The underlying goal of this string of research is to identify conditions under which intragroup conflict is destructive or productive, and to derive recommendations for improved organizational functioning. While relationship conflict in groups generally decreases satisfaction and performance, task conflict holds advantages for performance on nonroutine tasks. As this finding indicates, not only the type of conflict but also the type of task at hand is crucial for the impact of intragroup conflict. In the face of task conflict, group members engage in deeper and more deliberate information processing. For complex, nonroutine tasks, such critical thinking processes lead to enhanced creativity and effective problem solving. In contrast, for routine tasks with established procedures, task conflict is rather detrimental to performance because it interferes with existing routines. Overall, the negative picture of intragroup conflict has shifted. Therefore, organizations are advised to design and structure collaboration in work groups in a way such that relationship conflict between group members decreases, but that task conflict informs decision making and problem solving.

Culture and Group Dynamics

Since globalized work settings are more and more common, it is important to understand intercultural group processes and their impact on group dynamics. For example, group members' surface-level dissimilarity, which is in part shaped by ethnic diversity, holds the potential to negatively influence group dynamics (e.g., social integration in groups), which in turn relates to outcomes of work groups in organizations (Guillaume et al., 2012). Further, certain conditions moderate these relationships, like interdependence in groups or perspectives toward cultural diversity at work. That is, intercultural group processes can contribute positively to group dynamics and outcomes when cultural diversity is interpreted as a resource for learning and change. In contrast, intercultural group processes can also entail risk of within-group alignment. Group faultlines are formed when group members split up into subgroups based on individual attributes (e.g., race, ethnicity). Group faultlines have been shown to affect group dynamics (e.g., relationship and task conflict, cohesion), and to explain variance above and beyond the effects of team diversity on group outcomes (for an overview see Thatcher and Patel, 2011). Finally, according to group members' cultural orientations and value preferences, resulting norms for social interaction (e.g., open communication as a viable approach to conflict management in Western but not in Asian cultures) can negatively impact group dynamics. Overall, in order to be effective, intercultural group processes require high levels of awareness of structural influences and interpersonal differences.

Social Identity: In-Group and Out-Group Processes

The above-described phenomena refer to processes within organizational groups as opposed to processes between groups in organizations, an important issue to which we turn next. Social identity theory (SIT) is the most prominent psychological theory in this regard. Since this theory is rooted within psychology, it addresses individuals' social cognition with reference to group processes, as opposed to sociological theories, which cover entire collectives such as organizations. The following predictions derived from SIT are among the best demonstrated in social psychology. Individuals identify with the groups they belong to (i.e., the in-group) and derogate others who do not belong to their group. The resulting degree of identification with the in-group is at the cost of respect for and cooperation with individuals belonging to other groups (i.e., the out-groups). SIT emphasizes that all individuals are motivated to develop a positive social identity. The underlying processes are as follows: (1) Categorization; within their social environment, individuals categorize on the basis of who belongs to the in-group and who does not (in-group and out-group distinction). (2) Social comparison; individuals compare their in-group to out-groups, preferably in such a way that the dimension on which the comparison is performed produces positive distinctiveness (i.e., a feeling of positive uniqueness). If the in-group compares badly on a given dimension (e.g., group performance), it is argued that other dimensions are more important (e.g., to have a strong moral character rather than success), thereby protecting the positive distinctiveness of the in-group. Tajfel (1978) showed that even a minimal group context through a more or less arbitrary distinction between two groups (e.g., admirers of Vasily Kandinsky vs admirers of Paul Klee), suffices to produce in-group favoritism and out-group discrimination. Individuals from the

out-group as compared to in-group members are discriminated against more – for example, concerning the evaluation of positive attributes, interpersonal liking, and distribution of material rewards.

These processes are highly relevant in organizational contexts because subgroups within organizations predominantly think in terms of in-groups and out-groups – instead of perceiving the overall organization as their in-group. Therefore, organizational strategy is challenged. On the one hand, increasing in-group perceptions fosters a sense of togetherness, which in turn promotes engagement. On the other hand, overly strong in-group perceptions will hinder contact and cooperation between departments. Effective responses to these challenges following SIT encompass (1) communicating a common goal, which can only be achieved by joint action and cooperation of all groups in the organization; (2) providing members of each group with background information about tasks of the other groups; (3) rotating positions between groups; and (4) encouraging informal activities and personal relationships between members of different groups. Thereby, group borders remain permeable, and groups are open to each other's influence.

Groups and Organizational Change

Organizational change takes place frequently, often in terms of mergers and acquisitions or restructuring and downsizing of organizations. Leadership and group processes play a central role for the outcomes of organizational change (Peus et al., 2009) because change processes challenge existing structures on levels of individuals, groups, and the entire organization, and create a sense of uncertainty.

Steps of Organizational Transformation

In order to achieve the desired outcomes of organizational change, despite counteracting forces as described above, a number of basic principles and steps apply, which take into account psychological findings about the acceptance of change processes (Kotter, 1995):

- Creating a sense of urgency and inevitability of the change process
- Forming a group of powerful leaders who guide organizational members through the change process and act as positive role models
- Creating a clear vision and strategies to achieve the vision (i.e., giving direction and meaning to the change process)
- Communicating the vision through leaders and other channels in the organization (e.g., websites, newsletters)
- Empowering organizational members to actively engage in the change process, and giving them a voice in important decisions
- Creating short-term wins through visible improvements, and rewarding organizational members for their contributions
- Consolidating initial changes, and proceeding in further efforts
- Institutionalizing the new approach

Clarity, predictability, and procedural fairness play major roles in successful organizational change initiatives. Yet, in many cases, those who are responsible for the initiation and implementation of change do not reflect upon and adhere to these principles. Consequently, the success of many organizational change processes is put at risk due to the neglect of processes within and between organizational groups. Underlying is lack of consideration with regard to (1) organizational groups and their members' fears, hopes, and expectations in change processes; and (2) transparent communication about the whys and hows of organizational change. Taking the above-listed steps into account will foster successful change processes in organizations.

Change Through Non-Western Group Processes

Today's organizations are confronted with national as well as international competitors. In the car industry, for example, leading manufacturers like Toyota, General Motors, or Volkswagen are required to adapt production and distribution of nationally generated models to international markets in order to ensure their profitability. Thereby, a non-Western perspective to group processes has evolved, and a deeper understanding of such processes is desirable. One prominent example, the approach of lean manufacturing as introduced in the Toyota Production System, has influenced change processes in organizations worldwide. Lean manufacturing focuses on the fastest and most efficient way of producing and delivering high-quality goods to customers by eliminating all influences on the production that do not provide added value to reach these goals. The Toyota Production System specifically builds on the two concepts of just-in-time (i.e., continuous flow of production in which each step produces only what is needed in the next step) and *jidoka* (i.e., immediate stop to production when problems occur to prevent defective products). Similarly, the Japanese *kaizen* philosophy has been applied to business purposes with the aim of continuously improving organizational processes and standards by eliminating all redundant and unnecessary waste from the system (Spear and Bowen, 1999). Research and practice suggest that these principles contribute to organizational productivity. Further, it is important to consider how to best transfer these approaches from one culture into another. Management support, organizational communications, and enhanced learning opportunities are main pillars of successful implementation.

Group Processes and Leadership

The relevance of leadership for group processes in organizations has been widely acknowledged. We refer to theory and research in the areas of leadership in multinational corporations and shared leadership in groups because they represent important recent trends in literature.

Leadership in Multinational Corporations

The rise of multinational corporations is a result of increasingly globalized markets. Notwithstanding the advantages of

globalization, leading these kinds of organizations is particularly challenging to the extent that leadership takes places across and within national boundaries – including different cultural values and interpretations of effective leadership. As shown in the comprehensive Global Leadership and Organizational Behavior Effectiveness(GLOBE) studies, cultural values in part relate to culture-specific prototypes of leadership. Six second-order factors were empirically extracted: (1) charismatic/value-based, (2) team-oriented, (3) humane, (4) participative, (5) autonomous, and (6) self-protective leadership. While charismatic/value-based and team-oriented leadership were endorsed as part of effective leadership across cultures, humane and participative leadership were not fully universally endorsed, and perceptions of autonomous and self-protective leadership varied across cultures (Dorfman et al., 2004). Overall, leadership in multinational corporations is confronted with the challenge to integrate universally shared values (e.g., fairness, appreciation, motivation) with the specific value system of each individual culture. Promoting integration is a key contributor to the success of today's globalized organizations through effective leadership and group processes.

Shared Leadership in Groups

As an alternative to common, hierarchical approaches to leading individuals and organizations, pluralistic forms of leadership have gained increasing attention over the last two decades. Shared leadership is an example of pluralistic leadership that comprises an innovative integration of leadership and group processes, wherein group members lead each other toward goal achievement (Pearce and Conger, 2003). Many characterizations evolved, taking into account that shared leadership (1) relies on dynamic and interactive group processes integrated with top-down and bottom-up influence, (2) stretches along a continuum of few to many different sources of leadership in a group, and (3) comprehends shifting leadership roles over time in response to situational or contextual demands. The emergence of shared leadership – as opposed to hierarchical leader–follower relations – goes hand in hand with a change of perspectives on organizational structures (e.g., systems thinking, decentralization, and self-management in work groups). Research paradigms have been broadened such that the study of shared leadership is predominantly based on (1) ratings of entire teams with established leadership questionnaires extended to the group level; (2) ratings of the extent to which each individual team member shows certain leadership behaviors, and subsequent computation of group-level indices; and (3) empirical modeling of multilevel data. Hitherto empirical results from organizational contexts suggest that shared leadership positively influences desired outcomes, like team effectiveness, top management team performance, and group processes in virtual teams, exceeding the positive effects of hierarchical leadership. Yet, structural aspects of group functioning (e.g., group members' demographic and/or deep-level diversity) may hinder shared leadership. Therefore, further theory and research are needed to explore the premises to as well as advantages and disadvantages of leadership through group processes for organizational functioning in greater detail.

Conclusion and Perspectives

As the presented theoretical and empirical insights demonstrate, group processes are a typical example of the need for close interdisciplinary cooperation (e.g., between social psychologists and organizational psychologists) to fully capture the mechanisms of organizational functioning. Moreover, it was our goal to show the connections between social psychology, organizational psychology, and practice in organizations. Indeed, it appears that nothing is as practical as a good theory. Of course, no all-encompassing theory of group processes exists, but as we have shown, several small group theories are directly applicable to manifold organizational phenomena.

Future research should be devoted to (1) compare, integrate, and differentiate predictions based on existing group theories; (2) pursue new methodological pathways toward the study of group processes (e.g., longitudinal designs; measurement and analysis at individual, group, and organizational levels simultaneously; integration of social psychology and neuroscience); and (3) develop and evaluate evidence-based training and development programs for enhanced group functioning in organizations.

See also: Attitudes and Behavior; Conflict and Conflict Resolution, Social Psychology of; Conflict: Organizational; Corporate Culture; Decision Making, Psychology of; Development: Organizational; Group Processes, Social Psychology of; Human Resource Management, Psychology of; Leadership in Organizations, Sociology of; Management: General; Network Analysis; Obedience: Social Psychological Perspectives; Organizational Behavior, Psychology of; Organizational Climate; Organizational Decision Making; Organizations: Authority and Power; People in Organizations; Social Business; Social Comparison, Psychology of.

Bibliography

Asch, S.E., 1951. Effects of group pressure upon the modification and distortion of judgments. In: Guetzkow, H. (Ed.), Groups, Leadership, and Men. Carnegie Press, Pittsburgh, PA, pp. 177–190.

De Wit, F.R.C., Greer, L.L., Jehn, K.A., 2012. The paradox of intragroup conflict: a meta-analysis. Journal of Applied Psychology, 97, 360–390.

Dorfman, P.W., Hanges, P.J., Brodbeck, F.C., 2004. Leadership and cultural variation. In: House, R.J., Hanges, P.J., Javidan, M., Dorfman, P.W., Gupta, V. (Eds.), Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies. Sage Publications, Thousand Oaks, CA, pp. 669–719.

Frey, D., Schulz-Hardt, S., Stahlberg, D., 1996. Information seeking among individuals and groups and possible consequences for decision making in business and politics. In: Witte, E.H., Davis, J.H. (Eds.), Understanding Group Behavior. Small Group Processes and Interpersonal Relations, vol. 2. Lawrence Erlbaum, Hillsdale, NJ, pp. 211–225.

Guillaume, Y.R.F., Brodbeck, F.C., Riketta, M., 2012. Surface- and deep-level dissimilarity effects on social integration and individual effectiveness related outcomes in work groups: a meta-analytic integration. Journal of Occupational and Organizational Psychology 85, 80–115.

Hackman, J.R., Morris, C.G., 1975. Group tasks, group interaction process, and group performance effectiveness: a review and proposed integration. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 8. Academic Press, New York, NY.

Hackman, J.R., Wageman, R., 2005. A theory of team coaching. Academy of Management Review 30, 269–287.

Hentschel, T., Shemla, M., Wegge, J., Kearney, E., 2013. Perceived diversity and team functioning: the role of diversity beliefs and affect. Small Group Research 44, 33–61.
Janis, I.L., 1982. Victims of Groupthink: A Psychological Study of Foreign-policy Decisions and Fiascoes, second ed. Houghton Mifflin, Oxford, UK.
Kerr, N.L., Tindale, R.S., 2004. Group performance and decision making. The Annual Review of Psychology 55, 623–655.
Kotter, J.P., 1995. March–April. Leading change: why transformation efforts fail. Harvard Business Review, 59–67.
Mohammed, S., Ferzandi, L., Hamilton, K., 2010. Metaphor no more: a 15-year review of the team mental model construct. Journal of Management 36, 876–910.
Moscovici, S., 1976. Social Influence and Social Change. Published in cooperation with European Association of Experimental Social Psychology by Academic Press, London, UK.
O'Boyle, E.H., Forsyth, D.R., O'Boyle, A.S., 2011. Bad apples or bad barrels: an examination of group- and organizational-level effects in the study of counter-productive work behavior. Group & Organization Management 36, 39–69.
Pearce, C.L., Conger, J.A. (Eds.), 2003. Shared Leadership. Reframing the Hows and Whys of Leadership. Sage, Thousand Oaks, CA.
Peus, C., Frey, D., 2009. Humanism at work: crucial organizational cultures and leadership principles. In: Spitzeck, H., Pirson, M., Amann, W., Khan, S., von Kimakowitz, E. (Eds.), Humanism in Business. Perspectives on Responsible Business in Society. Cambridge University Press, Cambridge, UK, pp. 260–277.
Peus, C., Frey, D., Gerkhardt, M., Fischer, P., Traut-Mattausch, E., 2009. Leading and managing organizational change initiatives. Management Revue 20, 158–175.
Scharff, M.M., 2005. Understanding WorldCom's accounting fraud: did groupthink play a role? Journal of Leadership & Organizational Studies 11, 109–118.
Schein, E.H., 1990. Organizational culture. American Psychologist 45, 109–119.
Schulz-Hardt, S., Brodbeck, F.C., 2012. Group performance and leadership. In: Hewstone, M., Stroebe, W., Jonas, K. (Eds.), An Introduction to Social Psychology, fifth ed., Blackwell, Oxford, UK, pp. 415–448.
Spear, S., Bowen, H.K., 1999. Decoding the DNA of the Toyota production system. Harvard Business Review 77, 96–106.
Steiner, I.D., 1972. Group Process and Productivity. Academic Press, New York, NY.
Stroebe, W., Diehl, M., Abakoumkin, G., 1992. The illusion of group effectivity. Personality and Social Psychology Bulletin 18, 643–650.
Tajfel, H., 1978. Differentiation between Social Groups: Studies in the Social Psychology of Intergroup Relations. Academic Press, London, UK.
Thatcher, S.M.B., Patel, P.C., 2011. Demographic faultlines: a meta-analysis of the literature. Journal of Applied Psychology 96, 1119–1139.
Tuckman, B.W., Jensen, M.A., 1977. Stages of small-group development revisited. Group & Organization Studies 2, 419–427.
van Knippenberg, D., Schippers, M.C., 2007. Work group diversity. Annual Review of Psychology 58, 515–541.
Velden, F.S.T., Beersma, B., De Dreu, C.K.W., 2007. Majority and minority influence in group negotiation: the moderating effects of social motivation and decision rules. Journal of Applied Psychology 92, 259–268.
West, M.A., 2002. Sparkling fountains or stagnant ponds: an integrative model of creativity and innovation implementation in work groups. Applied Psychology 51, 355–387.
Williams, K.Y., O'Reilly, C.A., 1998. Demography and diversity in organizations: a review of 40 years of research. Research in Organizational Behavior 20, 77.

Relevant Websites

www.richmond.edu – Forsyth's Group Dynamics Resources Page.
lean.mit.edu – MIT Lean Advancement Initiative.
www.harvard.edu – Team Diagnostic Survey.
www.toyota-global.com – Toyota Production System.
www.tuftsgloballeadership.org – Tufts University, The Institute for Global Leadership.
www.socialpsychology.org – Social Psychology Network.

Group Processes in the Classroom

Angela M O'Donnell, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Group processes in the classroom occur when peers engage in a common task and can include a variety of both social and cognitive processes. The focus of this article is on cooperative or collaborative groups of peers. Thus, group processes occur during peer learning. The article describes a number of theoretical orientations toward peer learning that suggest different explanatory mechanisms for effective peer learning. These orientations have different implications for important decisions in the classroom. The relationship between what researchers know about effective group processes and the use of this knowledge by teachers is examined. Two models for considering group processes are described which attempt to integrate the varying theoretical orientations to peer learning. These models may make it possible to provide a better linkage between empirical research and classroom practice. Finally, directions for future research are outlined.

Interest in group learning and group processes has a very long history and remains a key focus of current efforts to improve education. Group processes in the classroom occur when peers engage in a common task and can include a variety of both social and cognitive processes. Cooperative and collaborative learning in the classroom has been associated with a variety of positive and social outcomes. Cooperative learning has been shown to be more effective than individualistic learning or competitive learning (Johnson and Johnson, 1989; Slavin, 2012). Students who participate in cooperative learning activities develop more positive relationships with peers including academically handicapped peers and have more positive attitudes toward subject matter (Johnson and Johnson, 1989). Furthermore, the social standing of students with learning disabilities (Fuchs et al., 2002) improves as a result of participating in peer learning activities.

There is no single method for arranging group interactions in the classroom. The focus of this article is on cooperative and collaborative groups of peers and the various theoretical approaches that attempt to explain group processes in these kinds of groups. These approaches propose varying mechanisms for the effectiveness of group learning. For example, some approaches emphasize within-group cooperation and between-group competition that result in the attainment of rewards. Other approaches do not rely on the provision of external rewards. Determining which method of fostering group interaction is most effective for producing academic achievement or positive social outcomes is complicated. Comparisons of empirical findings across studies conducted from different theoretical perspectives are very difficult because of the wide variation in tasks used, duration of studies, roles of teacher and students, outcome measures, and other variables such as group composition, group size, or the use of individual accountability. One consequence of the variation in theoretical approaches is the fact that these approaches require different decisions with regard to key variables such as group size or composition and the use of rewards. Group processes can be best understood within particular theoretical approaches. This article will not address the influences of specific variables but will instead focus on the delineation of the various theoretical perspectives and the linkage to practice.

Theoretical Approaches to Group Processes

Different theoretical approaches to peer learning propose different underlying mechanisms or processes that come into play when a group works together (O'Donnell and O'Kelly, 1994). Slavin (1996) described two main classes of orientation toward peer learning: a social-behavioral orientation that includes motivational and social cohesion approaches, and a cognitive orientation that includes elaboration and developmental theories.

Social-Behavioral Perspective

The concept of interdependence among group members is fundamental to the motivational and social cohesion perspectives on cooperative learning, although they differ in how interdependence is created and sustained. When group members are positively interdependent, the success of one member can occur only if other members are also successful. An example of a positively interdependent group is a tug-of-war team. In contrast, negative interdependence occurs in competitive situations in which one person's success requires that another person fail. An example of a negatively interdependent event is a track race. Only one person can win which means others must fail to meet their goals.

In the social motivational approach, interdependence is created by the use of group rewards. The principle underlying the use of rewards is that students will be motivated to help their peers to gain rewards. In Student Teams Achievement Divisions (Slavin, 1995), students work in small groups to master material that the teacher has presented using the teacher's preferred teaching method. Each student takes an individual quiz on the material, and points are assigned to the group based on each individual's improvement score. The improvement score is calculated based on the student's prior performance. Students are accountable for improving their achievement over past performance. Team points are awarded based on levels of improvement. The advantage of using improvement points is that learners of varying abilities can contribute equally to the total group score. High achieving students are less likely to feel resentful in working with less

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92078-6

capable students who in turn will be less likely to feel negative peer pressure. Group rewards are assigned based on the team's number of points. A key component of this approach is the reliance on individual accountability. Evidence of the effectiveness of this approach comes from a large number of studies (Slavin, 1996, 2012). The kind of learning task used in this approach, however, tends to focus on basic skills or factual knowledge and it is unclear whether this approach is effective in promoting higher-order skills. Students quiz one another using fact sheets that include the answers on the back of the sheet. The cognitive processes addressed are ones of rehearsal and some elaboration.

Social Cohesion Perspective

In the social cohesion approach, interdependence derives from a motive of care and concern experienced by group members for one another (Johnson and Johnson, 1994). Group members help one another because they are concerned about their peers. Teaching social skills that will support this kind of interdependence is an important emphasis in a social cohesion approach. Depending on their prior experiences, students may be unskilled at productive interaction. Classroom teachers need to spend time to encourage positive social norms, modeling helping behavior, and providing feedback on students' social interactions.

The Johnson's Learning Together technique (Johnson and Johnson, 1994) illustrates the social cohesion approach. The empirical evidence supporting this orientation toward collaborative learning as a method for enhancing academic achievement is much weaker than that supporting the motivational perspective. Slavin (1996) notes that cooperative learning methods that emphasize team building and group process, but do not provide group rewards, are not effective but when they are combined with the use of rewards or require accountability, student achievement from group learning exceeds individualistic learning. Slavin concluded that individual accountability and group rewards are critical to the success of cooperative learning methods.

Not everyone subscribes to the idea that rewards are necessary to create cohesion or interdependence. Cohen (1994) believes that cooperative learning that uses group rewards may be useful for lower-level skills but may not either be necessary or helpful for promoting higher-order skills. Thus, decisions about whether to adopt a motivational or social cohesion approach to peer learning may depend on the desired outcomes from group interaction.

Cohen (1994) emphasizes a different source of social cohesion, suggesting that the interest value of the task in which students engage will create the kind of interdependence necessary for students to succeed. Such tasks are likely to be authentic and engaging. Social cohesion is promoted by reducing the influence of status differences on students' interactions. When a student has high status in a group, he or she participates more, provides explanations, asks questions, and is generally active in group work (Cohen, 1982). A student with low status often does not have access to the materials being used by the group, participates less, and is often in the position of help-seeker rather than an active participant in the group. When tasks are used that are interesting to students and when positive social and cognitive interaction is supported, difficulties associated with status differences among students are ameliorated. When students have low status in a group, they participate less and in ways that do not promote achievement (Cohen, 1982). Explicit rewards in this context may only exacerbate the problems created by differential status in the group. Cohen and her colleagues have successfully promoted social cohesion in groups by reducing status differences among group members so that they are oriented to accept that all students *can* contribute to task accomplishment. Complex and interesting tasks are necessary for this level of inclusion to occur (Lotan, 2003). Cohen and her colleagues have been very successful in promoting student achievement using Complex Instruction (Cohen, 1994; Cohen and Lotan, 1995). Results from studies by Cohen and her colleagues suggest that the nature of the task may have important implications for the kind of accountability that is needed to promote effective group learning by improving group processes within the classroom (Cohen, 1994).

Cognitive/Elaboration Perspectives

The third perspective identified by Slavin is a cognitive/elaboration approach and is based on information processing theory. The role of group interaction from this perspective is to increase an individual's active processing of information by providing opportunities for restructuring and elaborating knowledge. The function of the group is to provide an individual with a mirror to his/her own cognition. Group learning techniques influenced by this perspective tend to provide some structure to the interaction of students by emphasizing the cognitive processes in which students must engage. Examples of work influenced by this perspective include the use of scripted cooperation by O'Donnell, Dansereau, and colleagues (O'Donnell and Dansereau, 1992), reciprocal peer questioning (King, 1999), reciprocal teaching (Palinscar and Brown, 1984), and the work of Noreen Webb (1992). The empirical evidence supporting the use of these techniques is strong.

The quality of discourse in group interaction is critical from this perspective and will have an influence on the outcomes from group interaction. Webb (1992) and Webb and Palinscar (1996) have shown that the quality of explanations that are provided in a group is linked to achievement. Giving explanations is more strongly associated with achievement than receiving explanations. King (1999) has also shown that instructional efforts to improve the quality of discourse by teaching students to ask questions that facilitate knowledge construction is very effective.

Developmental Perspectives

A fourth approach to understanding cooperative or collaborative groups relies on two different theories of development. From a Piagetian perspective (De Lisi and Golbeck, 1999), peers can learn from one another because they can provide opportunities for cognitive conflict and conflict resolution. Groups that experience mutuality of influence and power are most likely to provide a context within which cognitive development can occur because they are more likely to include

differences of viewpoint. Children who have not attained the principle of conservation often develop and maintain conservation concepts when they have worked with a child who has already developed this concept (De Lisi and Golbeck, 1999). Much of the work on conceptual change teaching in science education has been influenced by this perspective (e.g., Sinatra and Pintrich, 2003; Southerland and Sinatra, 2003).

From a Vygotskian perspective (Hogan and Tudge, 1999), children acquire knowledge and skills that are first modeled in the community and are subsequently internalized by the individual. When cooperative learning is anchored in Vygotskian theory, it is concerned with the larger community of practice in which individual learning is accomplished. Key processes include the scaffolding of the learner's performance by a more expert individual. The difference between what an individual can accomplish alone and with the assistance of a more knowledgeable adult or more capable peer is defined as the zone of proximal development.

Sociocultural and Combined Theoretical Perspectives

A fifth perspective that undergirds such group learning techniques such as Communities of Learners (Brown and Campione, 1990) may be identified. It includes the developmental perspective of Vygotsky, with its emphasis on the sociocultural basis of learning, and a social cohesion approach in which such learning is motivated by care and concern. Other forms of cooperative learning borrow from multiple perspectives in making choices about the selection of tasks and the organization of learning (Van Meter and Stevens, 2000).

The importance of the varied perspectives on collaborative and cooperative learning can be understood from the perspective of a classroom teacher who must make a variety of decisions about how to organize group learning in the classroom and what kinds of processes he or she wishes to promote. Table 1 delineates key differences among those perspectives as they relate to instructional decisions, problems that may arise from such decisions, and strategies for averting such problems. The goal is to promote positive group processes (e.g., helping) while minimizing negative group processes (e.g., evaluation of persons). Table 2 illustrates how four of these perspectives conceptualize the mechanisms that underlie effective cooperative and collaborative learning.

Table 2 Mechanisms for promoting learning from a variety of theoretical perspectives

Theoretical perspective	*Mechanism for promoting learning*
Social motivational	Interdependence created through motivation
Social cohesion	Interdependence created through care and concern
Cognitive elaboration	Deeper processing of information
Cognitive development–Piaget	Adaptation as a result of cognitive conflict
Cognitive development–Vygotsky	Modeled behavior that is scaffolded and internalized

Source: O'Donnell, A.M., Reeve, J., Smith, J.K., 2012. Educational Psychology: Reflection for Action. John Wiley, Hoboken, NJ, p. 303 © John Wiley, reproduced with permission.

Influences on Group Processes in the Classroom

A number of factors related to classroom practices influence group processes in the classroom. Three key dimensions of classroom practices are (1) the quality of students' discourse, (2) the role of the teacher in the collaborative classroom, and (3) the tasks selected.

Quality of Discourse among Participants

Simply being in a group does not necessarily mean that students will derive academic or social benefits from the group interaction. Negative processes can also occur in a group. An example of such a process is that of social loafing

Table 1

	Theoretical perspectives				
	Social-behavioral		*Cognitive*		
				Developmental	
Considerations	*Motivation*	*Social cohesion*	*Elaboration*	*Piagetian*	*Vygotskian*
Goals/incentives	Rewards essential	Rewards nonessential	Rewards nonessential	Rewards nonessential	Rewards nonessential
Group size	Large (4–6)	Large (4–6)	Small (2–4)	Small	Dyads
Group composition	Heterogeneous	Heterogeneous	Heterogeneous	Homogeneous	Heterogeneous
Tasks	Rehearsal	Rehearsal/integrative	Rehearsal/integrative	Exploratory	Skills
Teacher role	Director	Facilitator	Facilitator	Facilitator	Model/guide
Potential problems	Use of reward	Social loafing Cognitive loafing	Unequal participation	Inactive No cognitive conflict	Poor help-giving
Averting problems	Improvement scores	Team building	Direct instruction in help-giving	Structuring controversy	Providing adequate time/dialogue
	Simpler tasks	Conflict resolution strategies	Model help-giving		Direct instruction in help-giving
	Social cohesion	Discuss group processes	Scripted interaction		Modeling help-giving

Source: O'Donnell, A.M., O'Kelly, J., 1994. Learning from peers: beyond the rhetoric of positive results. Educational Psychological Review 6, p. 327, © Springer Publishers, reproduced with permission.

(Latané et al., 1979) in which an individual or individuals reduce their efforts in a group in comparison to how they would have performed if not in a group. The kinds of tasks on which students work will influence the quality of discourse within a group. The quality and depth of the discourse among students significantly affects the outcomes from group interaction (Webb and Farivar, 1999).

Task instructions can also influence the quality of discourse within a group. In one study by Chinn et al. (2000), pairs of students worked together to generate and evaluate conclusions from experiments they conducted with electrical circuits. Half of the pairs were asked to decide if each of three conclusions were 'okay' or 'not okay.' The other half of the students were asked to examine the same set of three conclusions and determine which of the conclusions was the best conclusion. Results of the study showed that the level of argumentation was much richer among pairs of students who were asked to choose the best conclusion.

Even when high achieving students work together, outcomes from collaborative learning depend on the quality of interaction (Barron, 2003). Barron examined the interaction of 12 triads of 6th graders engaged in problem-solving. Features of the dialogue between group members were predictive of problem-solving outcomes. When correct ideas were proposed and accepted, group outcomes improved. Less successful groups ignored or rejected correct ideas or did not link new ideas to previous ideas. Prior achievement was not predictive of successful outcomes.

The heterogeneity of intellectual ability and social skills within the group will also influence the quality of discourse within a group. Students who vary in gender, race, ethnicity, and ability or who experience special needs contribute to heterogeneity in the classroom. The variables may also have an influence on the effectiveness of the group. These differences can enhance learning as students with different experiences and background may deepen the quality of discourse by the introduction of varied perspectives and/or by the use of questions.

If one adopts a Vygotskian view of peer learning, intellectual skills must be modeled before a child can internalize them. The advantage of having a heterogeneous group is that there is an increased likelihood that someone in the group will be able to model higher-level thinking skills and may also be able to scaffold the efforts of less skilled individuals. Most students need support in generating the quality of discourse that is associated with learning. Teachers can promote high levels of cognitive activity by asking questions that vary in complexity (King, 1991; King et al., 1998).

The Role of the Teacher

Teachers' roles are key components of effective peer learning. These roles can be very complex and include such varied roles as those of community builder, preparing of learning activities, model of problem solving and other intellectual skills, coordinator of activities, and evaluator of students' work. All of these roles contribute to the promotion of effective group processes. Different peer learning activities require different stances with respect to students, tasks, and outcomes. Teachers must take into account both the social context in which learning occurs and the cognitive processes that are either supported or hindered in that context. They need to analyze their classrooms to determine whether there may be obstacles that limit the cognitive opportunities available to students or prevent them from making use of those opportunities. Teachers can adopt many roles in relation to the use of peer learning.

The quality of discourse from students is not the only discourse that is important. The example provided by the teacher in terms of discourse is also important. Webb et al. (2006) found that students reproduce the discourse of teachers in the context of a semester-long program of cooperative learning in middle school mathematics classes. They concluded that the models of discourse presented by the teacher were imitated by students. When the teacher modeled low levels of cognitive activity, the students followed suit. Teachers experienced difficulty in changing their behaviors to ones in which they allowed students authority for discourse. It requires a lot of professional development for teachers to effectively use group processes in the classroom.

When more than one person is engaged in a learning task, social processes are involved. In using cooperative groups or other forms of peer learning, the teacher also needs to provide feedback on students' interactions. They may provide direct instruction in the use of appropriate social skills or strategies for disagreeing with peers without being disrespectful. Cohen and colleagues have shown that when students are given criteria for an acceptable group product and use those to evaluate their work, their interactions improve and their learning increases (Cohen et al., 2002).

The Role of the Task

Teachers need to set goals for any task they assign to groups of students. Teachers may select tasks with the goal of motivating or engaging students, providing a context for social interaction with a student who is having particular difficulties, getting students to learn factual material, or getting them to engage in higher-order reasoning and thinking skills. Depending on the goals of the task, the kind of peer learning employed will vary.

Classroom tasks can be classified along a number of dimensions. The first dimension is complexity. Tasks will vary in the demands they make on students' cognitive, affective, social, and metacognitive processes. Complex tasks work best in conditions where reward contingencies are not high. Simple tasks may fail to interest students who may respond by engaging in social or cognitive loafing. Tasks can be simple and require simple cognitive strategies, such as rehearsal. An example of a simple task is that of memorizing the names of the capitals of states which requires rehearsal. Tasks require different kinds of knowledge. Those that involve basic knowledge acquisition are those in which information is encoded into the student's memory for later recall. Strategies for helping students master such tasks include teaching them memorization and rehearsal strategies, such as the use of mnemonics, summarization, and elaboration strategies.

Other tasks require more complex cognitive skills. An example of a complex task is asking students in small groups to design a zoo. To perform this task, students need to do research on the kinds of areas that are included in a zoo, learn about the kinds of animals that might inhabit the zoo, and learn about how to care for the animals. Tasks like these require

students to pose their own questions, explore alternatives, gather information, organize the information, generate evidence or arguments in support of ideas, and draw conclusions. The techniques that support basic knowledge acquisition and rehearsal are not sufficient to promote these kinds of higher-order thinking. Tasks that are more open-ended or require more collaborative knowledge construction call for cooperative techniques that are less structured in order to permit the kinds of interactions that promote the cognitive complex that will result in the desired outcomes.

A second dimension along which tasks vary is the degree to which group interaction can promote positive learning outcomes. Cohen (1994) distinguished between individual and group tasks. This distinction differentiates between tasks that are better done alone and those that will be best performed by a group. According to Cohen, a group task is one that "requires resources (information, knowledge, heuristic problem-solving strategies, materials, and skills) that no single individual possesses so that no single individual is likely to solve the problems or accomplish the task objectives without at least some input from others" (1994, p. 8). Such tasks are open-ended or less structured and require problem solving. Students can approach complex tasks from a number of vantage points and have multiple opportunities to demonstrate and further develop their intellectual competence. These tasks which are concerned with important content require positive interdependence among group members and the individual accountability of all members. Such tasks also include clear criteria for how the group's product will be evaluated (Lotan, 2003). Lotan (2003) refers to these kinds of tasks as group-worthy tasks. Deploying these kinds of tasks to promote learning requires a great deal of time and expertise on the part of the teacher.

There are some risks associated with open-ended tasks. Students may be drawn into negative social interactions; some may be easily frustrated if the task is ambiguous; and some will not participate fully. A combination of high risk of failure and task ambiguity may result in excessive demands on students' social and affective skills resulting in impairments in the use of cognitive or metacognitive strategies. Various strategies, such as the social skills training advocated by the Johnsons, the status interventions proposed by Cohen, and close monitoring of groups, can limit these negative outcomes.

It is important to note that not all students will be ready to assume complex roles and strategies. Depending on their history of working in groups, students may not be sufficiently independent to work effectively in groups. Students who have not worked in cooperative groups will need training and gradual experience in taking more responsibility for their learning. Likewise, students who have not valued educational achievement in the past will not suddenly become passionate about learning without a major effort to change their values regarding achievement. In such circumstances, a teacher will need to work to establish a culture of valuing academic achievement.

Research and Practice

Despite the burgeoning research findings that might assist teachers in making choices about grouping for instructional purposes, teachers do not seem to rely on the empirical research for guidance when making such choices (Antil et al., 1998). However, few studies have examined teachers' beliefs about group processes in the classroom.

Teachers' Use of Cooperative Learning

The research findings from any of the perspectives described seem to have little systematic influence on teachers' choices about groups. Although Slavin (1996) is convinced that empirical evidence supports the use of group rewards, he noted that there was quite a reluctance on the part of classroom teachers to use them.

Antil et al. (1998) conducted a study of elementary school teachers' conceptualization and utilization of cooperative learning. Eighty-five teachers responded to an initial survey. Ninety-three percent of the respondents reported that they used cooperative learning for both academic and social reasons. In interviews with 21 of these teachers, Antil et al. found that the majority of the teachers structured tasks to create positive interdependence and taught students skills for working in groups. Few, however, employed recognized forms of cooperative learning and, in particular, did not link individual accountability to group goals. No data were presented, however, that linked classroom practices related to cooperative learning to student achievement. It is clear from this study that teachers do not implement the particular practices of cooperative learning that research has shown to be effective. However, when teachers receive extensive professional development, the effects on student achievement can be powerful (Schlutz, 2000).

Another interesting feature of teacher behavior with respect to collaborative learning is the quality of discourse modeled by the teacher. Webb et al. (2006) have shown that the quality of discourse by students in groups parallels that of the teacher. When teachers modeled low-level discourse such as recitations or assumed the responsibility for problem solving, students tended to follow suit and confine their contributions to lower-level cognitive activities such as providing terminal answers or asking for simple levels of help.

Usable Models of Group Process

The availability of multiple perspectives on peer learning and the proliferation of specific techniques may contribute to the disconnection between the available research findings and their use by classroom teachers. The kind of classification system represented in **Table 1** is rarely presented to either preservice or in-service teachers. Usable models of per learning that can guide practical decisions are needed. Such models need to clearly link expected outcomes to classroom processes and teachers' actions or decisions.

Slavin (1996) proposed a model of effective peer learning with motivation at its core. Motivation is necessary in order for students to engage in the active processing (e.g., explanations) that is linked to achievement. The keys to motivating students, according to Slavin, are group rewards and individual accountability. Teachers can act to create and sustain children's motivation that will allow them to link will and skill in pursuit of academic achievement.

Van Meter and Stevens (2000) emphasize the quality of discourse within the group as the key variable in effective group learning. Other instructional choices should be such that the quality of discourse is enhanced. Thus, task instructions and group composition are only interesting insofar as they influence the quality of discourse among peers.

These two models provide a simple set of goals for teachers using group learning. Coherence among classroom decisions about fundamental aspects of arranging instruction (e.g., group size, group composition, and type of task) is more likely to occur if there is a clear goal in place (e.g., motivate students to engage in active learning). Teachers in the Antil et al. (1998) study reported that they did not subscribe to particular cooperative techniques because they wanted to maintain their decision-making autonomy. The two models that are briefly described here with their emphases on motivation and discourse provide a set of superordinate goals that can accommodate teachers' need for autonomy, while preserving a principled approach to decision making about group processes.

Future Directions

Given that teachers seem reluctant to use group rewards (Slavin, 1996), it is important to specify the conditions under which alternative methods of creating interdependence among group members can be effective. What are the kinds of tasks for which extrinsic rewards are not necessary? Answers to this question lie in the analysis of the kind of processes in which students engage when they are involved in interesting and challenging tasks. A clearer understanding of effective processing of complex tasks may make it possible to find alternatives to rewards.

The tension between theories of group processes and classroom practices of group learning presents the richest source of future research questions. Teachers' beliefs about learning lie at the fulcrum of their abilities to implement effective group learning in the classroom. The relationships between teacher beliefs, implement practices, and student achievement remain untouched as research questions. Research on this set of relationships would provide valuable insights that could inform teacher preparation and in-service professional development and also improve classroom practice of group learning.

See also: Classroom Level in Educational Effectiveness Research; Communities of Practice; Teacher Behaviours and Student Outcomes.

Bibliography

Antil, L.R., Jenkins, J.R., Wayne, S.K., Vadasy, P.F., 1998. Cooperative learning: prevalence, conceptualizations, and the relationship between research and practice. American Educational Research Journal 3, 419–455.

Barron, B., 2003. When smart groups fail. Journal of the Learning Sciences 12, 307–359. http://dx.doi.org/10.1207/S15327809JLS1203_1.

Brown, A.L., Campione, J.C., 1990. Communities of learning and thinking, or a context by any other name. Human Development 21, 108–126.

Chinn, C.A., O'Donnell, A.M., Jinks, T.S., 2000. The structure of discourse in collaborative learning. The Journal of Experimental Education 69, 77–97.

Cohen, E.G., 1982. Expectation states and interracial interaction in school settings. Annual Review of Sociology 8, 209–235.

Cohen, E.G., 1994. Restructuring the classroom: conditions for productive small groups. Review of Educational Research 64, 1–35.

Cohen, E.G., Lotan, R.A., 1995. Producing equal-status interaction in the heterogeneous classroom. American Educational Research Journal 32, 99–120.

Cohen, E.G., Lotan, R.A., Abram, P.L., Scarloss, B.A., Schultz, S.E., 2002. Can groups learn? Teaches College Record 104, 1045–1068.

De Lisi, R., Golbeck, S., 1999. Implications of Piagetian theory for peer learning. In: O'Donnell, A.M., King, A. (Eds.), Cognitive Perspectives on Peer Learning. Lawrence Erlbaum, Mahwah, NJ, pp. 3–37.

Fuchs, D., Fuchs, L.S., Mathes, P.G., Martinez, E.A., 2002. Preliminary evidence on the social standing of students with learning disabilities in PALS and no-PALS classrooms. Learning Disabilities Research 17 (4), 205–215.

Hogan, D., Tudge, J.H.R., 1999. Implications of Vygotskian theory for peer learning. In: O'Donnell, A.M., King, A. (Eds.), Cognitive Perspectives on Peer Learning. Lawrence Erlbaum, Mahwah, NJ, pp. 39–65.

Johnson, D.W., Johnson, R.T., 1989. Cooperation and Competition: Theory and Research. Interaction Books, Edina, MN.

Johnson, D.W., Johnson, R.T., 1994. Learning Together and Alone: Cooperative, Competitive, and Individualistic Learning, fourth ed. Allyn and Bacon, Boston.

King, A., 1991. Effects of training in strategic questioning on children's problem-solving performance. Journal of Educational Psychology 83, 307–317.

King, A., 1999. Discourse patterns for mediating peer learning. In: O'Donnell, A.M., King, A. (Eds.), Cognitive Perspectives on Peer Learning. Lawrence Erlbaum, Mahwah, NJ, pp. 87–115.

King, A., Staffieri, A., Adelgais, A., 1998. Mutual peer tutoring: effects of structuring tutorial interaction to scaffold peer learning. Journal of Educational Psychology 90, 134–152. http://dx.doi.org/10.1037/0022-0663.90.1.134.

Latané, B., Williams, K., Harkins, S., 1979. Many hands make light the work: the causes and consequences of social loafing. Journal of Personality and Social Psychology 37, 822–832. http://dx.doi.org/10.1037/0022-3514.37.6.822.

Lotan, R.A., 2003. Group worthy tasks. Educational Leadership 6, 72–75.

O'Donnell, A.M., Dansereau, D.F., 1992. Scripted cooperation in student dyads: a method for analyzing and enhancing academic learning and performance. In: Lazarowitz, R., Miller, N. (Eds.), Interaction in Cooperative Groups: The Theoretical Anatomy of Group Learning. Cambridge University Press, New York, pp. 120–141.

O'Donnell, A.M., O'Kelly, J., 1994. Learning from peers: beyond the rhetoric of positive results. Educational Psychological Review 6, 321–349.

O'Donnell, A.M., Reeve, J., Smith, J.K., 2012. Educational Psychology: Reflection for Action. John Wiley, Hoboken, NJ.

Palinscar, A.S., Brown, A.L., 1984. Reciprocal teaching of comprehension monitoring activities. Cognition and Instruction 2, 117–175.

Schlutz, S.E., 2000. Are four heads better than one: effects of group work on students' achievement in science. In: Paper Presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Sinatra, G.M., Pintrich, P., 2003. Intentional Conceptual Change. Lawrence Erlbaum, Mahwah, NJ.

Slavin, R.E., 1995. Cooperative Learning: Theory, Research, and Practice, second ed. Allyn and Bacon, Boston.

Slavin, R.E., 1996. Research on cooperative learning and achievement what we know, what we need to know. Contemporary Educational Psychology 21, 43–69.

Slavin, R.E., 2012. Classroom applications of cooperative learning. In: Harris, K., Graham, S., Urdan, T. (Eds.), Educational Psychology Handbook. American Psychological Association, Washington, DC, pp. 359–378.

Southerland, S.A., Sinatra, G.M., 2003. Learning about biological evolution: a special case of intentional conceptual change. In: Sinatra, G.M., Pintrich, P. (Eds.), Intentional Conceptual Change, pp. 315–342.

Van Meter, P., Stevens, R.J., 2000. The role of theory in the study of peer collaboration. Journal of Experimental Education 69, 113–127.

Webb, N.M., 1992. Testing a theoretical model of student interaction and learning in small groups. In: Lazarowitz, R., Miller, N. (Eds.), Interaction in Cooperative Groups: The Theoretical Anatomy of Group Learning. Cambridge University Press, New York, pp. 102–119.

Webb, N.M., Farivar, S., 1999. Developing productive group interaction in middle school mathematics. In: O'Donnell, A.M., King, A. (Eds.), Cognitive Perspectives on Peer Learning. Erlbaum, Hillsdale, NJ, pp. 117–149.

Webb, N.M., Nemer, K.M., Ing, M., 2006. Small-group reflections: parallels between teacher discourse and student behavior in peer-directed groups. Journal of the Learning Sciences 15, 63–119.

Webb, N.M., Palinscar, A.S., 1996. Group processes in the classroom. In: Berliner, D., Calfee, R.C. (Eds.), Handbook of Educational Psychology. Macmillan, NewYork.

Group Processes, Social Psychology of

Monique H Matelski and Michael A Hogg, Claremont Graduate University, Claremont, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Groups lie at the core of the human experience. Group processes encompass such a wide range of phenomena that the social psychological study of group processes has many different focuses. The study of group processes, in general, concerns the cognitive and social causes and consequences of human aggregation. This article provides a history and review of group processes research including information on group formation, group performance, and group decision making; of how people are socialized into groups and roles within groups; and of how groups influence their members through persuasion, leadership, and norms.

What Is a Group?

Groups occupy much of our day-to-day life and have a profound effect on our actions, thoughts, and feelings. People work in groups, socialize in groups, and play in groups. Furthermore, people represent their views and attitudes through groups. The variety of groups we may belong to and identify with is vast – it includes groups based on ethnicity, ability, gender, nationality, religion, political ideology, profession, hobbies, and so on. Groups can be described in various ways emphasizing specific characteristics such as mutual dependence, cohesion, communication, entitativity, influence, structure, and shared identity. The tendency to join with others in a group is one of the most fundamental characteristics of humans, and the processes that unfold within groups leave a lasting imprint on both their members as well as society.

Numerous definitions of groups have been proposed in the literature. There is, however, little consensus on a single social psychological definition of a group. The dominant view characterizes a group in two ways. First, a group consists of a collection of individuals who interact with one another. Second, in all groups there are processes that are interpersonal processes among a number of people. Alternative perspectives contend that groups possess 'emergent properties' that arise from interactions among members and therefore are qualitatively distinct from members' individual characteristics.

Early observers of group behavior (including Durkheim, McDougall, and Mead), more contemporary theorists (including Sherif and Tajfel), and those who espouse the social identity and intergroup approach to studying groups argue that the analysis of groups in terms only of interpersonal processes is reductionist – it does not adequately address questions to do with group processes, but instead deals only with interpersonal processes among more than two people. Although, the issue of 'what is a group' continues to be unresolved, there have been more recent attempts to integrate small group perspectives, intergroup perspectives, and cognitive perspectives.

Studying Group Processes

Interest in, and study of, group processes has ebbed and flowed. The 1930s marked the beginning of systematic research on group dynamics by social psychologists. Prior to World War II a number of classic studies were conducted including Sherif's experiments on the development of social norms under uncertainty. The late 1940s and early 1950s marked the golden age of groups research, a time when it gathered momentum and helped set experimental social psychology off on its current expansion. It was during this time, after World War II, that researchers began to investigate issues of leadership and conformity, as possible avenues for understanding how and why human atrocities associated with the Holocaust could occur.

Group processes research thrived until about the mid-1960s when it suffered a fairly rapid decline in popularity due to an increased interest in individual-level phenomena, attribution theory, and social cognition. Groups research went through a relatively quiet period until the early 1980s when it started to show strong signs of recovery, mainly attributed to increased interest in cognitive perspectives and European approaches to studying groups (i.e., social identity and minority/majority influence). There have been a number of explanations offered for the apparent demise of traditional groups research such as the financial and time constraints of conducting research on interactive groups and the fact that data from group research are rarely neat, tidy, and easy to explain (Levine and Moreland, 2012).

Group Formation

Why do people join and identify with groups, and why do they prefer some groups over others? The answer to these questions depends on a number of factors such as how the group is defined, how people come to join the group, and the degree of free choice in the decision. However, there are a number of circumstances, goals, and motives that tend to lead people to form, join, and remain in groups.

People are more likely to associate and form groups with certain people such as those who are in close proximity, those who express similar attitudes and values, and those who respond positively to them. Groups exist for a reason; they make it easier to attain goals that cannot be attained alone. In groups, people can coordinate their actions and cooperate to achieve goals that they could not attain on their own.

Motivations associated with group affiliation are basic and central to our existence. People have a basic need to belong and this causes them to affiliate with others. According to this perspective, self-esteem is seen as a meter of successful inclusion and group belonging. A sense of belonging produces a sense of self-esteem and self-worth. Some social psychologists believe this perspective to be too simplistic and argue that it does not account for membership in large-scale social categories or for the identity functions of groups. An alternative theory, terror management theory (Greenberg et al., 1997), posits that fear of one's own death is the main motivating force that leads people to join groups. Group affiliation provides people with symbolic immortality through social validation of their cultural worldviews that outlive them, and thus reduces existential anxiety and raises self-esteem. A third perspective, uncertainty–identity theory, proposes that the main motivation responsible for belonging and identifying with groups has less to do with self-esteem and anxiety reduction and more to do with individuals' desire to reduce uncertainty about themselves, their attributes, and their place in the world (Hogg, 2012).

Group Development, Cohesion, and Commitment

Groups are dynamic structures – new members join, old members leave, members are socialized into the group, and the group in turn is imprinted with the contribution of individuals. Most theorists propose that small groups go through five stages of development: orientation (forming), conflict (storming), structure (norming), work (performing), and dissolution (adjourning). Cohesion, in most cases, is a consequence of a period of group development. It is a term used to characterize the psychological process that transforms an aggregate of unrelated individuals into a group. Cohesion, a group-level construct, is one of the most theoretically important concepts in the study of group processes. Cohesiveness refers to the overall solidarity of the group – the property of a group that affectively binds people, as group members, to one another and to the group as a whole giving them a sense of oneness.

For many years, cohesion was studied in terms of interpersonal attraction. This individualistic conception of groups emphasizes individuals' attachment to individual other group members and overlooks their feelings about and attachment to the group as a whole. Subsequently, however, cohesion has been studied from a social identity perspective. From this viewpoint, cohesion is a consequence of the cognitive process of identifying with the group (i.e., self-categorization in terms of the group's defining features) and feelings about other group members are determined by how well those members match the group's prototype. A group is cohesive when its members are committed to the group and regard one another as good/prototypical exemplars of the group.

Group Influence

Groups influence the thoughts, feelings, and behavior of their members in many different ways.

Direct Influence and Persuasion

Groups tend to be aggregations of like-minded people and so those group members who do not go along with the majority of the group are subject to direct influence attempts. Groups use direct influence to try and persuade members to come into line with the group. Persuasion attempts may range from relatively polite requests all the way up to direct orders reinforced by threat and potential force. Persuasion is generally more effective if the source of influence has a degree of legitimate power or authority within the group (e.g., expert source or leader; Hogg, 2010). Direct influence attempts usually produce compliance, a superficial, public yielding to influence, which is not accompanied by underlying cognitive or attitudinal change. Individuals go along with the group for instrumental reasons which may include a desire to obtain a reward (i.e., social approval) or to avoid punishment.

Indirect Influence and Group Norms

Groups are powerful agents that can also influence members indirectly through group norms – shared beliefs about what is appropriate conduct for a group member. Norms mark group boundaries and describe and prescribe how group members should behave. Groups conform to norms in the absence of direct social pressure. Research on conformity suggests that there are three main processes of social influence: 'normative influence,' 'informational influence,' and 'referent informational influence.'

Normative influence is an influence to conform to the positive expectations of others. People comply with the norms of the group for instrumental reasons such as to gain social approval and acceptance, avoid disapproval, or achieve specific goals. Normative influence is only effective when people believe they are under surveillance by the group; therefore this form of influence creates surface-level, situation-specific compliance rather than underlying cognitive change.

Informational influence refers to an influence to accept information from others as evidence of reality. Individuals desire to feel that their perceptions, feelings, and beliefs are correct and appropriate in a given situation. When people feel uncertain about how they should think or behave (e.g., because of ambiguous stimuli or social disagreement) they initially make objective tests against reality. But if this is not possible, they will make social comparisons to gauge the correctness of their behavior (Festinger, 1954). Effective informational influence causes true cognitive change which may lead to behavioral change.

Social identity theory (Turner et al., 1987) argues that normative influence and informational influence represent a dual-process perspective that has focused on interpersonal dependency and underemphasized the role of group norms and group belongingness. This perspective proposed a third form of social influence associated with identification-based conformity, referent informational influence. There are three stages in which conformity to the group norm evolves: (1) individuals identify as a group member, (2) a contextually-appropriate norm (or prototype) is constructed or identified, and (3) group members internalize the norm, assimilate themselves to the norm, and use it as a guide for their own group behavior.

Group Structure

Interactions among group members are rarely random. Instead, they tend to follow regular patterns that reflect the underlying structure of the group. Group structure reflected in the roles, status relations, and communication networks, define and organize the group, and shape group processes.

Positions, Roles, and Status

Not everyone in a group performs the same function. They have differing *positions*, and people perform different tasks that they are expected to accomplish for the group. That is, people or subgroups of people occupy roles. Roles are coherent sets of behaviors that people occupying specific positions within a group are expected to perform. Some roles are assigned and others are acquired over time. Regardless of how people come to occupy roles, they emerge to help the group function as an integrated entity and to facilitate goal attainment. Levine and Moreland (1994) developed the 'group socialization model' which describes the ways roles are allocated to individuals and the ways in which members move in and out of roles within a group as a function of their commitment to the group.

In theory, roles could be defined simply in terms of task and need not imply differences in status; however roles are rarely equal – some are more desirable than other, and thus have more power and status attached to them. Social psychologists suggest that with increasing role differentiation also comes greater differentiation in status for those roles and thus hierarchy. It is the group members' expectations that lead to this status differentiation. Status tends to be given to those group members who have specific status characteristics (skills and abilities directly related to the group's function) and/or diffuse status characteristics (general qualities that are thought be indicators of competence). Groups do not always allocate status fairly. Group members let general status (i.e., diffuse status) characteristics influence their expectations; leading individuals who have achieved or been ascribed authority or prestige in one context enjoy a higher status in other, unrelated contexts.

Communication Structures

Another feature of group structure is the patterning of information exchange and ways in which group members communicate with each other. Communication structures tend to become more complex and varied as groups increase in size. Centralized communication structures channel all communication through a single, central communication hub (often a leader), while decentralized communication structures allow freer communication among all members because no one position is more central than another. For relatively simple tasks, centralized structures improve group performance; the hub person can receive, process, and disseminate information effectively and efficiently allowing other group members to focus on their roles. For tasks that are more complex and require greater information exchange and processing, decentralized structures are superior. This type of communication structure prevents any one member from being overloaded and overwhelmed with complex information, and thus improves overall group performance.

The nature of the communication structure also affects group functioning, overall commitment, and satisfaction with membership in the group. For example, because centralized networks have lower levels of closeness, the overall level of member satisfaction in such groups tends to be low. Communication networks can be traced and classified using an 'interaction process analysis' method, which documents the underlying recurring patterns of interpersonal behavior in a group. However, with the explosion of computer-mediated communication (CMC), methodologies for studying communication networks will have to adapt to changes in the nature of interaction between individuals.

Leadership

Groups normally require guidance and support as they strive to meet their goals, and the individual who coordinates and motivates the group can profoundly shape the group's future. This individual, the leader, occupies one of the most important and certainly the highest status role in the group. Leadership can be defined as a process of social influence through which an individual guides others in the attainment of a collective goal, often by mobilizing, directing, coordinating, supporting, and motivating their efforts (Hogg, 2010). There is much theory and research on leadership – particularly in the organizational and management sciences (Yukl, 2010). Some of this work focuses on the emergence and acceptance of leaders, but most of it focuses on their effectiveness. Effective leadership is often operationalized as being able to successfully define and set collective goals and influence others to achieve this vision for the group.

Effective leadership is thought to be a function of a match between leadership style or individual personality and the demands posed by the task or situational. These contingency theories of leadership recognize that the effectiveness of a particular leadership style is dependent on the properties of the leadership situation – some styles are better suited to some situations than are others.

Social cognitive theories of leadership suggest that leaders are more favorably appraised and thus more effective if followers perceive them as closely matching general or task-specific leadership stereotypes. However, when a particular social group becomes salient and is important to member's self-concept, leadership may be viewed as an identity process in which group prototypicality becomes important. Under these circumstances group prototypical leaders are more favorably evaluated, more strongly supported, and thus more effective.

It is because leadership is a role that leadership dynamics also involve followers – followers must follow in order for leaders to lead. Transactional leadership perspectives emphasize the dynamic nature of leader–follower interaction, an interaction in which leaders and followers provide mutual support and gratification. Generally, followers tend to endorse leaders who have good leadership qualities, who have the abilities to benefit and promote the group as a whole, and who are good group members or have been good group members in the past. Transformational leadership theories view group transformation as the key function of leadership and focus on the charismatic properties or behavior of leaders that allow them transform groups.

Deviance

Most groups are structured in terms of the central or marginal group membership of specific members. Those members that embody the group's attributes often have influence over the group and occupy leadership roles whereas those who significantly deviate from the majority are often treated as 'black sheep.' There are a number of variables that affect how groups respond to deviants such as the extremity and content of the deviant's position, the degree to which the deviant interferes with group goal attainment, and the group's openness to alternative perspectives. Research suggests that in general groups will try hard to socialize and include marginal members – marginalizing and rejecting members only as a last resort. Deviation from the group's normative practices may raise questions about what the group stands for, thus posing a threat to the group and its identity.

Research on reactions to deviance has also focused on the intergroup context. Much of this work focuses on the 'black sheep effect' – dissenting ingroup members are rejected and evaluated more negatively than a similar member who belongs to an outgroup. However, a different process occurs when a member not only is a deviant member but also actively criticizes the group norms. In this case, it is the outgroup member that is rejected by the ingroup (Hornsey, 2005), or a collection of such 'rejects' may schism from the parent group (Sani, 2005).

Group Performance

Groups often exist to accomplish a specific task. Group performance is the process and outcome of members' efforts to create a joint product or complete a task in which they all have a stake. When people combine their energies and coordinate their efforts, they can accomplish goals that would overwhelm an individual, however, people working together can also encounter problems coordinating and maximizing their efforts. An important question to ask is whether groups perform better or worse than individuals. The answer to this question depends on the specific nature of the task – social facilitation tends to occur when tasks are simple, whereas social interference occurs for tasks that are complex.

Social Loafing and Social Dilemmas

Although group interactions can improve overall group productivity, most research on group performance has focused on social interference in the form of process loss – reduction in performance caused by factors (e.g., reduced effort, coordination problems, and faulty group processes) that prevent the group from reaching its full potential. Two sources of process loss are coordination loss, which occurs when members are not able to organize their actions in an ideal way within the group, and motivation loss, which occurs when group members do not exert maximal effort on a task.

Social interference research has generally focused on 'social loafing' which occurs when people exert less effort while working together than while working alone. 'Social impact theory' and the 'collective effort model (CEM)' suggest that the social loafing effect is partially due to a reduction in perceived personal identifiability and accountability, task attractiveness, and responsibility within the group. Most research on social loafing has focused on eliminating or reversing its effects by, among other things, increasing task importance, relevance, enjoyment, group cohesiveness, and personal identifiability. Similar to loafing, 'free-riding' is a tendency for group members to contribute less to a group task when they believe that other group members will compensate for this lack of effort.

Group life by its very nature creates 'social dilemmas,' situations in which individuals can maximize their own short-term gains by acting in one way, with little regard for the long-term greater good of the collective. Because social dilemmas are so common in everyday life and group members often respond to them by seeking to maximize their short-term individual welfare, researchers have searched for ways to increase the likelihood that members will cooperate rather than compete. Evidence indicates, for example, that social dilemmas can be (partly) avoided when group members view their contributions as critical, when they communicate with one another, and when the group has a system for punishing people who refuse to cooperate. Additionally, leadership can serve to reduce the negative impact of social dilemmas; that is, leaders can transform selfish individual goals into shared group goals by building a sense of common identity, shared fate, and responsibility for the collective good.

The Crowd, Deindividuation, and Collective Behavior

Group dynamics research is based on the core assumption that people in a group act collectively. Although collective action usually takes place in small groups, people sometimes join larger collectives such as crowds and social movements. A crowd is a specific type of collective in which individuals share a common focus of interest and are concentrated in a single location (i.e., an audience, a mob, or a riot). Beginning with LeBon in the late nineteenth century, early theorists and researchers largely assumed that crowds are unique social aggregations where antisocial, violent, and impulsive behavior is the norm. These theorists argued that crowds are governed by a primitive collective mind that turns people into unreasonable, extreme, and impulse-driven followers. Early social psychological perspectives emphasized the power of the group and assumed that being a member of a crowd lead to 'deindividuation,' a psychological state characterized by reduced self-awareness and personal identity salience that had negative consequences for behavior.

Many contemporary theorists, rather than assuming that collectives are atypical groups that are out of control and normless, argue that collectives are at their core, groups and so the processes that shape group behavior also shape 'collective behavior.' Normative interpretations of collective behavior suggest that like any other group behavior, members of a collective adhere to norms (that could either be prosocial or antisocial). Additionally, social identity perspectives suggest that collectives are intergroup phenomena and that people in collectives do not experience a loss of identity rather their social identity is augmented in these situations. Collective behavior, similar to other group behavior, is a rational attempt by

a group to seek change through unified action. Research suggests that people engage in social movements and associated crowd events because they feel a sense of injustice, have a higher sense of efficacy, and identify with the group and its purpose.

Group Decision Making

One of the most common and important activities that groups perform is making decisions. Groups often make better decisions than individuals because in a group there is a larger pool of expertise, viewpoints, and so forth. Groups may come to a decision slowly; however, they are likely to make use of group procedures that enhance the way they gather, analyze, and weigh information. According to the functional theory of group decision making, groups generally follow the same processes or steps when they make a decision. The group first identifies the problem, sets goals, and develops a strategy in the orientation phase. Next, during the discussion phase, the group gathers information and considers the options. In the decision phase, the group chooses its solution by coming to a consensus, voting, or using another social decision process. Finally, in the implementation stage, the decision is put into action and the impact of the decision is assessed. Although groups have greater informational resources and the potential to outperform individuals, during the group discussion phase they are also subject to an array of group processes that may impair decision making such as conflict between group members, biased leadership, and conformity pressures.

Groupthink and Group Polarization

Groups sometimes employ deficient decision-making procedures that produce poor decisions. 'Groupthink' is one such poor decision-making process in which the desire for unanimity overrides the motivation to consider rational alternative courses of action. Several factors lead to groupthink and its corresponding symptoms including high group cohesiveness, impartial leadership, insulation from external information, and stress from external threat. Groupthink can have disastrous ramifications, and has been identified as the cause of many poor crisis decisions made by governments; for example, the 1961 Bay of Pigs fiasco (the failed attempt by the United States to invade Cuba and overthrow Fidel Castro's regime thought to be the result of a flawed decision made by President Kennedy and his advisors). There is only partial empirical support for groupthink and more recently researchers have proposed alternative groupthink models. The group centrism approach claims that groups whose members strive to maintain unity and have a high need for closure are more likely to make poor decisions. Another model proposes that negative decisional features previously thought to cause groupthink combined with shared social identity, restrictive norms, and lack of confidence will trigger groupthink.

Another potential bias in group decision making is 'group polarization,' a tendency for group discussion to cause the group and its members to shift to a more extreme position than initially held by the individual members. Polarization occurs when the prediscussion group preference tends to favor one pole of the decision-making dimension. The group decision is then more extreme in this direction. Social comparison (normative influence) processes, exposure to persuasive arguments (informational influence), and social identity processes are offered as explanations for group polarization.

Group Remembering

Another important component of group decision making involves the collective recollection of information – group remembering. Groups generally recall more than individuals on simple tasks because groups draw on more memories that contain different types of information. But groups tend to remember less than individuals on complex tasks, likely because of process loss. Group remembering is more than a simple collaborative recall, rather it is a constructive process whereby social comparison processes among group members shape collective memory. Thus, group memory is subject to social influence processes which may lead to group decision making bias (i.e., some individuals' memories will contribute to the developing consensus while others' will not). Additionally, group memory may be weakened by social loafing and by the complexity of the group setting which disrupts group members' ability to organize and retrieve information.

Another feature of group remembering is 'transactive memory.' Groups, like individuals, have memory systems that collect, encode, store, and retrieve information. Because groups generally have to process large amounts of information when making decisions, groups share the memory load so that each individual is responsible for remembering only part of what the group needs to know, but all members know who specializes in what memories. Transactive memory involves uneven distribution of memory in a group, and although it helps to improve overall group functioning and decision making, one potential problem is that when a group member leaves there is a temporary loss or reduction in group memory.

Intergroup Decision Making

Decision-making groups may contain individuals with opposing views, or can be considered to be interactive decision-making contexts in which opposing groups or their representatives come together to reach an agreement. In this situation, the group is a context for intergroup decision making. In this context, individuals act as group representatives rather than individuals. Thus they tend to bargain more fiercely and are less compromising than if they were bargaining for themselves. Intergroup decision making often ends in deadlock. Third party mediation may help resolve a stalemate but only if the mediator has power and is seen by both groups to be impartial.

The Intergroup Dimension

Historically, the study of group processes focused on what happens among members of small interactive groups and did not fully recognize how the relations between groups impacted intragroup processes. More recently, scholars have begun to incorporate intergroup perspectives in their understanding of intragroup processes. For instance, group polarization may be

an intergroup phenomenon in which a relatively extreme ingroup differentiates itself from a less extreme outgroup by becoming even more extreme. Intergroup processes can also affect leadership. For example, leaders not only lead their groups, but in some ways they lead their group against other groups. They must also bridge group divisions by building social harmony, a common purpose, and a shared identity. Also, extreme behavior in crowds may be best explained in terms of intergroup conflict – the crowd is a group with a common purpose which is directly in opposition to an outgroup.

Future Directions

There has been a marked resurgence of interest in social psychology in group processes. The rise in popularity of group research has been attributed to an integration of small group, intergroup, and cognitive perspectives in social psychology that sets the groundwork for promising future developments. The increase in popularity of online communities and public interest groups presents new and unique avenues for valuable research and insight. In the future, online group researchers will have to integrate their findings with theory and data from traditional face-to-face groups. Additionally, small groups research seems to have been taken over by organizational behavior programs in business schools and industrial–organizational psychology programs. Thus group research on understanding and improving the performance of work teams will likely remain strong in the foreseeable future.

See also: Collective Behavior, Social Psychology of; Decision Making, Psychology of; Deindividuation, Psychology of; Intergroup Relations; Leadership; Obedience: Social Psychological Perspectives; Social Cognition; Social Comparison, Psychology of; Social Dilemmas, Psychology of; Social Identity in Social Psychology; Social Movements: A Social Psychological Perspective; Social Psychology.

Bibliography

Chemers, M.M., 2001. Leadership effectiveness: an integrative review. In: Hogg, M.A., Tindale, R.S. (Eds.), Blackwell Handbook of Social Psychology: Group Processes. Blackwell, Oxford, UK, pp. 376–399.

Festinger, L., 1954. A theory of social comparison processes. Human Relations 7, 117–140.

Forsyth, D.R., 2009. Group Dynamics, fifth ed. Brooks/Cole, Pacific Grove, CA.

Greenberg, J., Solomon, S., Pyszczynski, T., 1997. Terror management theory of self-esteem and cultural worldviews: empirical assessments and conceptual refinements. In: Zanna, M.P. (Ed.), Advances in Experimental Social Psychology, vol. 29. Academic Press, New York, pp. 61–139.

Hackman, J.R., Katz, N., 2010. Group behavior and performance. In: Gilbert, D., Fiske, S.T., Lindzey, G. (Eds.), Handbook of Social Psychology, vol. 2. McGraw-Hill, New York, pp. 1208–1251.

Hogg, M.A., 2012. Uncertainty-identity theory. In: Van Lange, P.A., Kruglanski, A.W., Higgins, T.E. (Eds.), Handbook of Theories of Social Psychology, vol. 2. Sage, Thousand Oaks, CA, pp. 62–80.

Hogg, M.A., 2010. Influence and leadership. In: Gilbert, D., Fiske, S.T., Lindzey, G. (Eds.), Handbook of Social Psychology, vol. 2. McGraw-Hill, New York, pp. 1166–1207.

Hogg, M.A., Cooper, J. (Eds.), 2007. The Sage Handbook of Social Psychology: Concise Student Edition. Sage, London.

Hogg, M.A., Tindale, R.S. (Eds.), 2001. Blackwell Handbook of Social Psychology: Group Processes. Blackwell, Oxford, UK.

Hogg, M.A., Vaughan, G.M. (Eds.), 2011. Social Psychology, sixth ed. Pearson Education, London.

Hornsey, M.J., 2005. Why being right is not enough: predicting defensiveness in the face of group criticism. European Review of Social Psychology 16, 301–334.

Levine, J.M., Moreland, R.L., 1994. Group socialization: theory and research. European Review of Social Psychology 5, 305–336.

Levine, J.M., Moreland, R.L. (Eds.), 2006. Small Groups. Psychology Press, Philadelphia, PA.

Levine, J.M., Moreland, R.L., 2012. A history of small group research. In: Kruglanski, A.W., Stroebe, W. (Eds.), Handbook of the History of Social Psychology. Psychology Press, New York, pp. 383–405.

Moreland, R.L., Levine, J.M., 2000. Socialization in organizations and work groups. In: Turner, M. (Ed.), Groups at Work: Theory and Research. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 69–112.

Sani, F., 2005. When subgroups secede: extending and refining the social psychological model of schisms in groups. Personality and Social Psychology Bulletin 31, 1074–1086.

Turner, J.C., Hogg, M.A., Oakes, P.J., Reicher, S.D., Wetherell, M.S., 1987. Rediscovering the Social Group: A Self-categorization Theory. Blackwell, Oxford, UK.

Yukl, G., 2010. Leadership in Organizations, seventh ed. Prentice Hall, Upper Saddle River, NJ.

Group Psychotherapy, Clinical Psychology of

Donelson R Forsyth, Jepson School of Leadership Studies, University of Richmond, Richmond, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Group psychotherapy is the treatment of psychological, behavioral, emotional, and interpersonal problems by a mental health professional who meets regularly with clients in a group rather than a one-on-one context. Group psychotherapy is used to treat a variety of disorders, including addiction, cognitive disturbances, depression, eating disorders, characterological disorders, and many more, using methods that are as varied as those used in individual approaches. Multiple group-level factors, including instilling hope, social learning, social support, and self-insight, operate to promote adjustment in psychotherapy groups, and research suggests that this treatment technique is as effective as more individualized methods when properly administered.

Like any other form of therapy, group psychotherapy is designed to help clients overcome debilitating psychological and interpersonal problems. But unlike individual therapies, group therapies involve treating individuals in groups, with the group and its members providing resources for achieving positive personal change. When this treatment modality was initially proposed, many questioned the wisdom of promoting mental health by encouraging guided interactions between individuals suffering from psychological problems, but experience and research affirmed its efficacy. Note, however, that group psychotherapy is a treatment for individuals rather than groups. Group therapists are mindful of the interpersonal processes that operate within the group and may deal with the group as a whole, but they are not seeking to treat the group per se: They seek to promote the adjustment of the individuals in the group and with the support of the group.

Approaches to Group Psychotherapy

Group psychotherapy emerged as a means of helping people deal with psychological problems in the early years of the twentieth century. Clinical psychology itself was developing during this period, but most early psychotherapists worked with patients in one-on-one sessions only. Some innovators, however, applied their basic change principles in group settings, as did some therapists who turned to groups for didactic purposes or because their patient load demanded a massed approach to treatment. These group applications were often confined to a particular psychiatric facility or clinic, but in time they became more dispersed as other mental health professionals heard of and made use of group methods. These early psychotherapy sessions involved treating individuals "in groups, with the group itself constituting an important element in the therapeutic process" (Slavson, 1950: p. 42).

Group psychotherapy is used currently to treat a variety of problems, including major depressive disorder, bipolar disorder, panic disorder, posttraumatic stress disorder, social phobia, obsessive-compulsive disorder, bulimia nervosa, binge-eating disorder, substance use disorder, schizophrenia, borderline personality disorder, and general personality disorder (Burlingame et al., 2004). Group psychotherapists are usually mental health professionals, and their patients tend to be suffering from diagnosed clinical conditions. The methods used by group psychotherapists are as varied as individual approaches to therapy. Some therapists, for example, are primarily psychoanalytic in orientation, for they focus on pathological personality, characterological disorders, and deep-seated psychological conflicts. Others, in contrast, adopt a more interpersonal perspective that stresses the exploration of the social processes that unfold in the group. Other approaches are more behavioral; they teach group members specific behaviors they can use to cope with the problems they are facing. Most therapists, however, are eclectic: They base their work on a melding of many theoretical perspectives (DeLucia-Waack and Kalodner, 2005).

Psychoanalytic Group Psychotherapy

Psychoanalytic group psychotherapy draws on Freud's (1922) conceptualization of the causes of dysfunction and his recommendations regarding the most effective way to treat those dysfunctions (*see* Psychoanalysis in Clinical Psychology). The analyst, after establishing an enduring, supportive relationship with the patient, uses questioning, free association, and interpretation to stimulate self-insight, transference, and the cathartic release of unconscious anxieties and characterological conflicts (Shedler, 2010). The psychoanalytic group therapist, however, takes advantage of the group setting to help patients deal with hostilities, needs, and wants that are repressed in everyday interactions but surface in the group setting. As Freud's (1922) replacement hypothesis suggests, the group becomes a surrogate family with members serving as symbolic siblings and the group therapist acting as the primal authority figure. As transference develops, the group provides the therapist with the means of exploring the childhood roots of current adult anxieties. Members may react to one another inappropriately, but in ways that parallel their responses to people they know in their everyday lives. Some therapists are more fully Freudian in their orientation than others, but rare is the therapist who does not deal with transference processes, familial tensions, and other latent conflicts.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.21066-0

Psychoanalytic group psychotherapists diverge on one key aspect of the process. Some stress the importance of the individual in the group. This approach argues that the therapist should stress individual-level processes in treatment, and avoid spending treatment time exploring the dynamics of the therapeutic group. Others, in contrast, integrate the treatment of the individual with the analysis of the group itself. The group-as-a-whole approach, which can be traced to the British psychotherapists at the Northfield Military Hospital in England, is a sterling example of this approach (Harrison, 2000). Although group-as-a-whole therapists embrace psychoanalytic assumptions of unconscious motivations, personality conflicts, and transference, they make use of the tension between the group and the individual to promote growth and development. Bion (1959), for example, stressed the analysis of defense mechanisms used by groups to cope with ego threats, uncertainties, and anxieties. He believed that these strategies are rooted in dependence, for the group members engage in collective projective identification as they struggle to avoid blame for their problems by shifting responsibility from themselves to the therapist. Bion assumed that rapid and lasting adaptive change could be achieved by helping group members understand both individual and collective defensive processes that unfold in their group session, and his work provided the basis for the Tavistock Institute of Human Relations (Conci, 2011).

Cognitive–Behavioral Therapy Groups

Cognitive and cognitive–behavioral psychotherapists, rather than searching for the cause of the problematic behavior in unseen unconscious conflicts or interpersonal transactions, base their interventions on principles derived from learning theories (*see* Concepts and Methods of Cognitive Therapies). They identify and quantify desirable cognitions and behaviors that will be encouraged and undesirable cognitions and behaviors that will be extinguished, and then modify these behaviors through cognitive restructuring, self-instructional training, and problem-solving instruction.

Cognitive–behavioral therapy groups use these principles with two or more individuals (Hollon and Beck, 2004). Cognitive–behavioral group therapy is generally both structured and problem focused. Prior to treatment, the therapist observes the reactions of each member to the group to index the degree of functioning and reviews the theories and procedures that sustain the intervention to clarify expectations and goals. Therapists can also use public commitment to these goals to enhance the binding strengths of a behavioral contract that describes in objective terms the goals that the group members are trying to achieve. During the therapeutic sessions themselves, the cognitive–behavioral group therapist can capitalize on the presence of multiple actors to magnify the effects of modeling, rehearsal, and feedback. Members of the group can practice interpersonal exchanges with other members, and other members can act as observers. These sessions can be recorded and reviewed so that the participants can see what they are doing correctly and what aspects of their behavior need improvement. During this feedback phase, the leader offers reassurance and praise, and members add their support and encouragement.

Interpersonal Group Psychotherapy

Many group psychotherapists not only exploit the social processes that occur in groups during treatment but also trace psychological disturbances back to social sources – particularly interactions with friends, relatives, and acquaintances. Interpersonal theorists assume that dysfunctional behaviors result from individuals' failures to take note of, and correct, cognitive and behavioral tendencies that disrupt their relationships with other people.

Many group therapists, recognizing the interpersonal bases of psychological problems, use the group setting to help members examine their interpersonal behavior. Yalom's interpersonal group psychotherapy (also called interactive group psychotherapy), for example, stresses the "here-and-now" of the group's interaction over the analysis of events and experiences that occur outside of the group (the "then-and-there"; Slavin, 1993). Members may spend some time, especially early in the group's life span, discussing problems they are facing at home or work, but Yalom feels that much of this discussion will be futile since members characterize such events inaccurately. Yalom instead recommends that members focus on interactions within the group, with each member eventually responding to one another as they do to others in their lives. The group thus becomes a "social microcosm," giving members true insight into the causes of the personal problems (Yalom and Leszcz, 2005).

Existential and Humanistic Psychotherapeutic Approaches

Another set of perspectives within the broad field of group psychotherapy draws on clinical and personality theories that stress the existential, self-reflective roots of adjustment and disorder. Humanistic, third-force theories such as person-centered theory, psychodrama, Gestalt, and human potential groups assume that individuals are, at their core, searching for ways to improve their current functioning and maximize their potentialities. Such approaches are often phenomenological and experiential, and they underscore the importance of a strong, positive, and empathic therapeutic environment for facilitating personal and collective growth.

Perls (1969), for example, frequently conducted Gestalt therapy sessions in groups. Members observed one another's 'work,' and they engaged in interpersonal exchanges with one another as Perls orchestrated the experience. In Gestalt groups, therapists frequently use unstructured interpersonal activities to stimulate members' emotional understanding (Gaffney, 2012). For example, a member will sit apart from the others and spend time publicly working through his or her understanding of an emotionally difficult experience (the 'hot seat'), or imagine that another person or a part of him- or herself is sitting in an empty chair and then carry on a dialogue with the person (the 'empty chair'). These techniques, when properly applied, can trigger strong emotional reactions, but Gestalt therapists resist offering direct interpretations to their patients (Greve, 1993).

Psychodrama, developed by Moreno (1953), also makes use of exercises to stimulate emotional experiences. Moreno conducted therapeutic groups perhaps as early as 1910, and he used the term "group therapy" in print in 1932 (Hare and

Hare, 1996). Moreno believed that the interpersonal relations that develop in groups provide the therapist with unique insights into members' personalities and proclivities, and that by taking on roles the members become more flexible in their behavioral orientations. He made his sessions more experiential by developing psychodrama techniques. When role playing, for example, members would take on the identity of someone else, and then act as he or she would in a simulated social situation. Moreno believed that psychodrama's emphasis on physical action was more involving than passive discussion, and that the drama itself helped members overcome their reluctance to discuss critical issues.

Therapeutic Factors in Groups

Group methods were used, initially, because they are efficient: Services could be rendered to many rather than just a few. But group psychotherapy, as a practice, is also consistent with psychologist Kurt Lewin's law of psychological change: "It is easier to change individuals formed into a group than to change any of them separately" (Lewin, 1951: p. 228). But what aspects of the group promote positive change? A common-factors perspective on group therapy suggests that various types of group psychotherapies, despite their many differences, share some common properties, and these commonalities may be responsible for their effectiveness (Forsyth, 2010). A psychodynamic therapist uses different techniques than a cognitive–behavioral or interpersonal therapist, but these various formats share mechanisms that promote change and sustain well-being. Yalom and Leszcz (2005) term these shared qualities curative factors, and their list includes the installation of hope, universality; the imparting of information, altruism; the corrective recapitulation of the primary family group; and the development of socializing techniques, imitative behavior, interpersonal learning, group cohesiveness, catharsis, and existential factors.

Social Motivation in Group Psychotherapy

Therapeutic groups are an informational resource for their members. The lone individual, experiencing some psychological difficulty, may misunderstand the sources of the problem and generate only a limited number of solutions or alternatives. But individuals who actively communicate with others experiencing similar difficulties can avail themselves of the relatively rich informational resources of the group. When assembled in a therapeutic group, members who initially felt they were uniquely challenged realize that the situation they face is shared by others. Yalom calls this realization universality, for patients realize that they have not been singled out to bear some extraordinary misfortune; rather, the problem they face is a universal one (Yalom and Leszcz, 2005).

Therapeutic groups also capitalize on social comparison processes to stimulate growth and adjustment. Social comparison occurs when individuals spontaneously compare themselves to other people around them. This comparative process does not occur in individualistic therapies, for patients can only imagine how they are coping relative to others. In groups, in contrast, a patient learns both directly and indirectly about one another's functioning. These comparative data yield two useful consequences: First, comparison provides patients with a baseline to calibrate their own level of functioning and determine where they stand along the coping continuum. They may learn, through downward social comparison – comparing themselves to those in the group who are performing poorly – that their own situation is not as bleak as they originally assumed. Second, this comparison can increase their sense of optimism about reaching an improved level of functioning. Yalom terms this process the installation of hope. Whereas individuals facing unpleasant circumstances alone may feel discouraged and pessimistic, when group members recognize that others are improving their motivational levels rise (Yalom and Leszcz, 2005).

Interpersonal Learning

Much of the learning that occurs during psychotherapy is social in nature. A skilled therapist can shape clients' behaviors through feedback and selective reinforcement, offer support, and model appropriate cognitive and behavioral responses to difficult situations. But when therapy is delivered in a group context, the sources of social learning are multiplied. Group psychotherapy necessarily involves patients interacting with other patients, and so it sets up opportunities for interpersonal learning among patients. Some therapists purposely minimize the impact of the group on the individual patients, but most group therapists exploit these social processes to promote change.

Interpersonal therapies assume that much of the curative power of a group springs from such social processes as imparting information from one person to another (guidance), the development of socializing techniques, imitative behavior, and interpersonal learning (Forsyth, 2010). For example, the exchange of information among participants was one of the first purposes of groups; these group therapies were primarily didactic in nature, with psychotherapists lecturing to large groups of patients about mental and physical health. Interpersonal learning also occurs indirectly, as members monitor their impact on the other people within their group, and the psychotherapist helps them draw conclusions about their maladaptive interpersonal tendencies. Members also gain simply by observing others dealing with their difficulties. By watching others' emotional displays, problem-solving behaviors, coping strategies, and interpersonal maneuverings, patients acquire adaptive social skills.

Groups also provide their members with feedback about interpersonal effectiveness and shortcomings. Members often exchange corrective, and in some cases negative, feedback, but such exchanges are usually counterbalanced by substantial doses of congratulatory, positive evaluations. In interpersonal therapy, in particular, patients often act as they do in other social contexts, so much so that the behaviors that caused them interpersonal difficulties erode their relationships within the group. When their fellow group members and therapist call their attention to the parallel, patients gain insights into their own motives and emotions. Members who report a strong level of identification with the group are more likely to accept the group's influence.

Group Cohesion and Development

As therapeutic groups develop over time, members become closely connected to each other emotionally, developing bonds of friendship that are as strong as any they experience in nontherapeutic relationships. A cohesive therapeutic group offers members something that groups, in general, offer their members: social support. When people experience challenging, traumatic events, the risks of psychological and physical illness are elevated, but if they are members of supportive groups, they are often protected against these negative consequences (Wills and Ainette, 2012). Most people turn to commonplace groups such as friendship cliques and family for their social support, but members of therapeutic groups frequently help one another by providing advice, guidance, direction, and emotional support. A cohesive group accepts its members, and this feeling of acceptance is considered by many group members to be the most central benefit of group membership (Kivlighan and Tarrant, 2001).

The group's level of cohesion depends, in part, on its stage of development (Forsyth and Diederich, 2014). Early in the group, interactions tend to be superficial, but in time conflicts tend to surface. In most groups, these conflicts between members and the group therapist give way to the development of cohesion (Burlingame et al., 2011). This progression of emotional experience, although characterized in a variety of ways, generally follows the forming, storming, norming, performing, and adjourning pattern identified by Tuckman (1965). During the forming stage, members are uncertain how to behave, and they do not trust one another enough to disclose personal information or to admit they need help. During the storming stage, group members often find themselves in conflict over the purpose of the group, and they may withdraw from the group and the therapist. In groups where the leader takes a nondirective orientation, this conflict stage focuses on the leader's failure to 'take charge' of the interactions. During the next phase, norming, the group becomes more cohesive, as norms and standards emerge that dictate how each session will unfold. This stage makes the performing stage possible, where group members engage earnestly in the exchange of information that will help them achieve their therapeutic goals. In time, members must deal with issues of termination as the group adjourns. This movement from one stage to another is therapeutic, so much so that a group that does not move through these stages may not benefit its members (Yalom and Leszcz, 2005).

Corrective Emotional Experiences

Therapeutic change results not only from the dispassionate analysis of one's psychological problems, but also from the direct experience of emotionally charged situations and their resolution. When a therapeutic session triggers a strong emotional reaction, such as a heated argument between group members, an unexpected expression of dislike or hostility for an idea or an individual, or an unreserved flow of grief and sadness, members can work through this 'corrective emotional experience' to better understand the source of their emotional reactions and how they can better handle them.

Group settings provide members with a range of opportunities to express emotions in ways that promote, rather than undermine, their well-being. As the concept of catharsis suggests, holding in strong negative emotions, such as anger and frustration, is unhealthy, and group sessions provide opportunities to 'blow off steam'; in psychologically and socially favorable circumstances (Ormont, 1984). Release of emotions, however, is not the curative benefit, but rather the analysis of the emotion-generating experience and the constructive communication of emotions to others. Group sessions provide members with more effective ways to disclose their emotions and to discriminate the effects that their emotional expressions have on others (Ormont, 1999).

Self Processes in Therapeutic Groups

All psychotherapies stress the value of self-understanding and self-acceptance, and group psychotherapies are no exception. The guided interaction of therapy groups provides the members with a rich source of information about aspects of themselves, for when members observe the way they interact with others – say, by dominating the group interaction, constantly criticizing others, displaying strong emotions, or expressing concern for others' well-being – they come to realize that they are dominant, critical, emotional, or caring, respectively. Members may be concerned that the group may see them for what they are – and that this accurate appraisal may not match their own sense of self – but self-discovery is a key goal of most individuals seeking analysis (Kivlighan et al., 2010).

Groups not only encourage the growth of self-knowledge but also promote self-acceptance. A number of psychological problems, including depression, anxiety, alcohol abuse, masochism, and eating disorders such as bulimia, are rooted in a devalued self, but inclusion in a stable, supportive group tends to promote self-esteem (Leary et al., 1995). As Baumeister and Leary's belongingness hypothesis argues, all "human beings have a pervasive drive to form and maintain at least a minimum quantity of lasting, positive, and impactful interpersonal relationships" (1995: p. 497). Just as exclusion from a group can trigger a loss in self-esteem, so can inclusion in a group contribute to more positive feelings of self-worth. Therapeutic groups also sustain self-esteem by giving members an opportunity to help others (altruism).

The Effectiveness of Group Psychotherapy

When therapeutic groups were initially proposed, skeptics questioned the wisdom of putting people who are suffering from psychological problems together in one group (Lazell, 1945). How, they asked, could troubled individuals be expected to cope in a group when they had failed individually? And how could the therapist guide the therapeutic process in a group, given the subtle nuances that pervade one-to-one therapies? In time, however, the method gained widespread acceptance as therapists added group sessions, either to supplement or to completely replace their individual sessions. Group psychotherapy is currently used to treat all types of psychiatric problems, including addictions, thought disorders,

depression, eating disorders, posttraumatic stress disorder, and personality disorders (Burlingame et al., 2011).

But do members leave the group better adjusted than when they entered? Do groups help patients explore and gain insight into sources of anxiety, help them identify and practice behaviors that they can use to cope with sources of stress, and provide them with a way to develop the skills they need to interact successfully with others? Meta-analytic reviews, including those that take into consideration methodological rigor and therapist orientation, suggest that group approaches are as effective as individual methods. Burlingame and his colleagues, for example, statistically combined the results of hundreds of experimental, quasiexperimental, and correlational studies before concluding that researchers had been more diligent in their studies of cognitive–behavioral therapies, and their review indicated that such methods were superior to untreated control groups and equivalent in effectiveness to individual approaches (Fuhriman and Burlingame, 1994). These findings were generally sustained when Burlingame et al. (2004) summarized findings from 107 studies and 14 meta-analyses of the use of group methods for treating a range of disorders, including anxiety, eating disorders, substance abuse, and personality disturbances. These analyses indicated that group methods are better for some problems than others – particularly social phobia and substance abuse. Other reviews indicate that groups that generate higher levels of member-to-member interpersonal feedback and stronger working relationships among members were more effective than other methods. The use of pregroup preparation and providing a higher level of structure during early group sessions also improved outcomes (Barlow, 2013).

Burlingame et al. (2004: p. 652) concluded that "group psychotherapy is potent enough to be the sole or primary treatment for patients suffering from a psychiatric disorder," but this positive conclusion requires some qualification. First, the empirical evidence is not definitive (Bednar and Kaul, 1994). Whereas a number of reviews are positive, the "use of varied and undocumented therapeutic methods, with different types of clients, by therapists who differ in skills and experiences, in studies that too frequently lack valid measures and adequate controls, make[s] it difficult to draw firm conclusions" (Forsyth and Corazzini, 2000: p. 325).

Second, group approaches work better for some disorders than for others (Burlingame et al., 2004). In particular, in both outpatient and inpatient settings, individuals experiencing mood disorders (e.g., anxiety and depression) respond better to group psychotherapies than individuals experiencing other types of disorders (e.g., thought and dissociative disorders).

Third, too little is known about the psychological and social processes that sustain the changes that are produced by group therapy. Do group-level processes such as social influence (informational, normative, and interpersonal), group cohesion, group norms, and social networking operate to shape the structure and function of therapeutic groups? If so, how do these group-level processes combine to influence outcome? Which group-level processes are most responsible for attitudinal and behavioral change, and which are less critical? Researchers have only begun to answer these questions pertaining to the factors that mediate the treatment–outcome relationship.

Last, the changes brought about by group experiences may be more perceptual than behavioral. Group members frequently report achieving major insights into their behavior and are positive about the gains they have achieved, but objective indices of outcome are more contradictory. Self-reported insight is also related to actual improvements in adjustment. One review, for instance, identified 26 controlled studies of personal growth groups that (1) used both pretest and posttest measures, (2) met for at least 10 h, and (3) had a long-term follow-up (at least 1 month after termination). Summarizing these methodologically superior studies, the reviewers concluded that group treatments did result in enduring positive changes, but primarily on self-report questionnaire data and not on behavioral measures (Berman and Zimpfer, 1980). These and other findings suggest that groups are most useful in promoting changes when members themselves appraise the gains they have achieved through participation.

These significant caveats aside, the available evidence prompts a guardedly optimistic conclusion about the therapeutic use of groups. Groups often exert an unrelenting influence on their members. Nearly all human societies are organized around small groups, and these groups shape their members' psychological adjustment and dysfunction. Given their ubiquity, people generally respond positively when presented the opportunity to work in a group to achieve mental health goals. Indeed, the increasing popularity of self-help groups signals the need for treatments that take advantage of groups for promoting adjustment. Although the use of groups in business and educational settings was once considered innovative, individuals now expect to learn and work in group contexts.

Group psychotherapy is also an efficient therapy, for a single therapist can treat more than one individual at a time – a characteristic that is considered very valuable in healthcare systems that seek to minimize costs without cutting quality. Far more research is needed to analyze the nature of the therapeutic group and its impact on members, but given the powerful self-processes and interpersonal processes that such groups instigate and the positive findings already reported by researchers, groups should be considered a treatment of choice rather than a radical alternative.

See also: Occupational Health; Psychodynamic Therapy: Short-term; Therapist–Patient Relationship.

Bibliography

Barlow, S.H., 2013. Specialty Competencies in Group Psychology. Oxford University Press, New York.

Baumeister, R.F., Leary, M.R., 1995. The need to belong: desire for interpersonal attachments as a fundamental human motivation. Psychological Bulletin 117, 497–529.

Bednar, R.L., Kaul, T., 1994. Experiential group research: can the canon fire? In: Garfield, S.L., Bergin, A.E. (Eds.), Handbook of Psychotherapy and Behavior Change, fourth ed. Wiley, New York, pp. 631–663.

Berman, J.J., Zimpfer, D.G., 1980. Growth groups: do the outcomes really last? Review of Educational Research 50, 505–524.

Bion, W.R., 1959. Experience in Groups and Other Papers. Basic Books, New York.

Burlingame, G.M., MacKenzie, K.R., Strauss, B., 2004. Small group treatment: evidence for effectiveness and mechanisms of change. In: Lambert, M.J. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, fifth ed. Wiley & Sons, New York, pp. 647–696.

Burlingame, G.M., McClendon, D.T., Alonso, J., 2011. Cohesion in group therapy. Psychotherapy 48, 34–42.

Conci, M., 2011. Bion and his first analyst, John Rickman (1891–1951): a revisitation of their relationship in the light of Rickman's personality and scientific production and of Bion's letters to him (1939–1951). International Forum of Psychoanalysis 20, 68–86.

DeLucia-Waack, J.L., Kalodner, C.R., 2005. Contemporary issues in group practice. In: Wheelan, S.A. (Ed.), The Handbook of Group Research and Practice. Sage, Thousand Oaks, CA, pp. 65–84.

Forsyth, D.R., 2010. Group process and group psychotherapy: social psychological foundations of change in therapeutic groups. In: Maddux, J.E., Tangney, J.P. (Eds.), Social Psychological Foundations of Clinical Psychology. Guilford, New York, pp. 497–513.

Forsyth, D.R., Corazzini, J.G., 2000. Groups as change agents. In: Snyder, C.R., Ingram, R.E. (Eds.), Handbook of Psychological Change. Wiley, New York, pp. 309–336.

Forsyth, D.R., Diederich, L.T., 2014. Group dynamics and development. In: DeLucia-Waack, J., Kalodner, C., Riva, M. (Eds.), The Handbook of Group Counseling and Psychotherapy, second ed. Sage, Thousand Oaks, CA.

Freud, S., 1922. Group Psychology and the Analysis of the Ego (J. Strachey, Trans.). Hogarth Press and the Institute of Psycho-analysis, London.

Fuhriman, A., Burlingame, G.M., 1994. Group psychotherapy: research and practice. In: Fuhriman, A., Burlingame, G.M. (Eds.), Handbook of Group Psychotherapy: An Empirical and Clinical Synthesis. Wiley, New York, pp. 3–40.

Gaffney, S., 2012. A neo-Lewinian perspective on gestalt group facilitation. In: Levine, T.B.Y. (Ed.), Gestalt Therapy: Advances in Theory and Practice. Routledge/Taylor & Francis Group, New York, pp. 149–159.

Greve, D.W., 1993. Gestalt group psychotherapy. In: Kaplan, H.I., Sadock, M.J. (Eds.), Comprehensive Group Psychotherapy, third ed. Williams & Wilkins, Baltimore, MD, pp. 228–235.

Hare, A.P., Hare, J.R., 1996. J.L. Moreno. Sage, Thousand Oaks, CA.

Harrison, T., 2000. Bion, Rickman, Foulkes, and the Northfield Experiments. Jessica Kingsley Publishers, London.

Hollon, S.D., Beck, A.T., 2004. Cognitive and cognitive behavioral therapies. In: Lambert, M.J. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, fifth ed. Wiley & Sons, New York, pp. 447–492.

Kivlighan Jr., D.M., Tarrant, J.M., 2001. Does group climate mediate the group leadership–group member outcome relationship? A test of Yalom's hypotheses about leadership priorities. Group Dynamics: Theory, Research, and Practice 5, 220–234.

Kivlighan Jr., D.M., Miles, J.R., Paquin, J.D., 2010. Therapeutic factors in group-counseling: asking new questions. In: Conyne, R.K. (Ed.), The Oxford Handbook of Group Counseling. Oxford University Press, New York, pp. 121–136.

Lazell, E.W., 1945. Group psychotherapy. Sociometry 8 (3/4), 101–107.

Leary, M.R., Schreindorfer, L.S., Haupt, A.L., 1995. The role of low self-esteem in emotional and behavioral problems: why is low self-esteem dysfunctional? Journal of Social and Clinical Psychology 14, 297–314.

Lewin, K., 1951. Field Theory in Social Science. Harper, New York.

Moreno, J.L., 1953. How Kurt Lewin's "Research Center for Group Dynamics" started. Sociometry 16, 101–104.

Ormont, L.R., 1984. The leader's role in dealing with aggression in groups. International Journal of Group Psychotherapy 34, 553–572.

Ormont, L.R., 1999. Progressive emotional communication: criteria for a well-functioning group. Group Analysis 32, 139–150.

Perls, F., 1969. Gestalt Therapy Verbatim. Real People Press, Lafayette, CA.

Shedler, J., 2010. The efficacy of psychodynamic psychotherapy. American Psychologist 65, 98–109.

Slavin, R.L., 1993. The significance of here-and-now disclosure in promoting cohesion in group psychotherapy. Group 17, 143–150.

Slavson, S.R., 1950. Group psychotherapy. Scientific American 183, 42–45.

Tuckman, B.W., 1965. Developmental sequences in small groups. Psychological Bulletin 63, 384–399.

Wills, T.A., Ainette, M.G., 2012. Social networks and social support. In: Baum, A., Revenson, T.A., Singer, J. (Eds.), Handbook of Health Psychology, second ed. Taylor & Francis Group, New York, pp. 465–491.

Yalom, I.D., Leszcz, M., 2005. The Theory and Practice of Group Psychotherapy, fifth ed. Basic Books, New York.

Groups, Sociology of

Siegwart Lindenberg, University of Groningen, Groningen, The Netherlands; and Tilburg University, Tilburg, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A 'group' refers to a bounded collection of interacting individuals who are functionally, cognitively, and structurally interdependent to various degrees. Groups also differ in the degree to which they are task- or sentiment-oriented and in which the members are unique or substitutable. Thus, friendship groups, clubs, crowd-sourcing groups, and teams are paradigmatically different types of groups. The sociological study of groups is dispersed in various university departments (ranging from social psychology and sociology to business and communications), but there is an important common theme: the collective phenomena that emerge in groups. The most important collective phenomena studies are 'social embedding' (including shared knowledge and reality, social identity, collective memory, belonging), norms, collectively induced motivations, and status structures. In turn, these phenomena have considerable influence on behavior, but this influence is more frequently studied by psychologists.

The Sociological Approach to Groups

In sociology, a group refers to a bounded collection of interacting individuals who are interdependent in various ways. The most prominent interdependencies that have been identified are functional, cognitive, and structural interdependencies (see Lindenberg, 1997). In a group, individuals are functionally interdependent with regard to the realization of one or more goals. For example, group members need each other to reach a common goal (task interdependence) and may also affect each others' outcome (cooperative outcome interdependence). They are cognitively interdependent with regard to their collective identity, memory knowledge, sense of reality, expectations, and perceptions. For example, members influence each other in the formation of expectations and depend on each other with their collective identity. They are structurally interdependent with regard to the effects of the ties between them. For example, each member is affected by the constellation of informal ties between other members or by different links due to division of labor among group members. The three interdependencies influence each other, but sociologists often differ in the degree to which they focus on one or more of these interdependencies at the same time.

Groups differ in important respects. Traditionally, sociologists distinguished between primary and secondary groups, mostly based on the distinction of relationships seen as constitutive of the group (close relationships for primary groups, impersonal relations for secondary groups). By now, it seems more useful to distinguish groups by their place in a two-dimensional space (see Figure 1). One dimension is the degree to which the specific identity of a group member matters (is the group member unique or, in principle, substitutable?). A second dimension is the degree to which the common goal is related to task fulfillment or sentiment. For example, in a friendship group, each member's personal identity is important and people cannot easily be substituted. In addition, the group is not linked to specific common tasks but mainly to the maintenance of common sentiment. Members in a social club (such as a faculty club) jointly create common sentiments and are not tied to specific common tasks. However, they may stay or leave and are thus fairly substitutable. A team (such as a research team) is oriented toward common tasks with close coordination of activities, so that the identity of each team member also matters. By contrast, in a crowd-sourcing group (such as the makers of open source software), there are common tasks but contributors can come and go, so that identity matters little. For all kinds of groups, the 'groupishness' of groups (also called 'entativity' or 'cohesion' and mainly a matter of cognitive interdependency) can differ, but the uniqueness of members is likely to be linked to greater groupishness than the substitutability of members. Maybe because of this diversity of types of groups, sociology has not developed a field of research dealing with groups as such. The departments in which groups are studied from sociological points of view are also varied (e.g., sociology, social psychology, management, business, organization studies, social movements, communications).

Individuals have needs that are satisfied by group membership, such as the need to belong (Baumeister and Leary, 1995), the need for meaning (Heine et al., 2006), and the need for

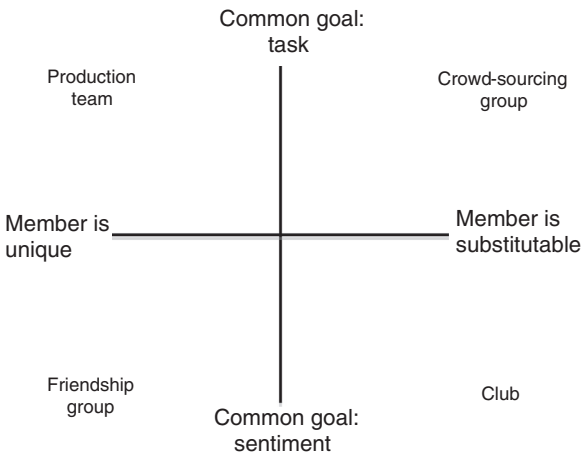

Figure 1 Two-dimensional group space with prototypical kinds of groups.

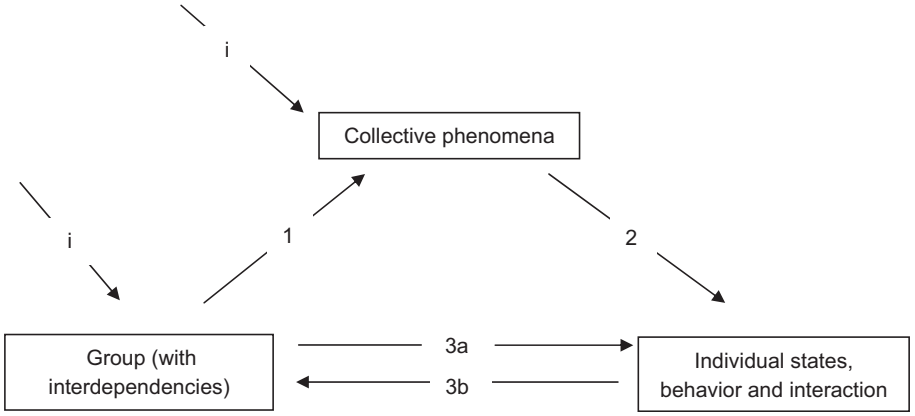

Figure 2 Schematic depiction of the study of phenomena related to groups.

status (Steverink and Lindenberg, 2006). However, the sociological study of groups focuses not so much on why people join groups but on the collective aspects related to such needs. For example, for sociologists, belonging is not an individual state but a joint product of interacting individuals just as, for them, memory is mostly a collective phenomenon. Thus, in the work of sociologists, all kinds of groups are seen as giving rise to durable or semidurable collective phenomena (such as social embedding, social norms, patterns of motivation, status structures). These collective phenomena, in turn, influence the behavior of individuals and their interactions, such as performance, conflict behavior, creativity, openness to information, etc. (see **Figure 2**). Most of the sociological work deals with relation 1, some deal with relation 2 and some deal with relationships 1 and 2 (as well as other combinations). Relation 3b allows explicit feedback mechanisms involving both levels (as in expectation states theory, see below). Psychological studies of groups focus more often on the direct influence of interaction (or the presence of others) and on attitude change, productivity, satisfaction, self-esteem, depression, etc., occasionally allowing for feedback mechanisms (3a and 3b). The extra arrows (i) indicate exogenous factors (such as genetic dispositions, rules, established categories, and externally established status differences).

Historical Background of Small Group Research

The scientifically serious investigation of groups in the sociological tradition started with anthropological studies of tribes, clans, and families in the nineteenth and early twentieth centuries. Sumner, Durkheim, Simmel, and Cooley were early codifiers of this research. The most important turn in this development occurred in the 1930s. Within the span of a few years, five great books appeared, which had two very important messages in common: that groups are constituted by a cognitively mediated functional interdependence between individuals, that is, people who are dependent on each other for achieving some goal (including common sentiment) interact and something important happens to their minds and the way their minds operate when they interact. In addition, the actual situation in which their action is placed matters most for behavior (as opposed to traits or other cross-situational constants). This point was elaborated in various ways by a number of important books. In 1932, Alfred Schütz published *Der sinnhafte Aufbau der sozialen Welt* (*The Phenomenology of the Social World*). In 1934, Mead's (then already posthumous work) *Mind, Self and Society* appeared. Two years later, Lewin's *Principles of Topological Psychology* and Sherif's *The Psychology of Social Norms* came out. A year later, Parsons' *The Structure of Social Action* made its appearance. Merton's early work on reference groups was also quite influential in this regard. From this time, group studies continued in two traditions: experimental and field studies; both traditions bore the mark of this convergent conception of groups. Under the leadership of Lewin and Sherif, 'group dynamics,' as the experimental study of groups was then called, achieved great heights. Field studies had Mayo's Hawthorne studies and Whyte's famous *Street Corner Society* as models. Both traditions owe their success to the simultaneous consideration of the three fundamental interdependencies in groups: functional (related to goal achievement), cognitive (related to common categorizations, shared beliefs, and knowledge), and structural (related to the configurations of social relations). See Lindenberg (1997) for more detail on these interdependencies. However, the precise nature of the interdependencies was not yet well investigated and in the late 1950 and early 1960s, specialized groups of studies on each of these interdependencies became dominant, with interesting crossovers between psychology and sociology. Homans (1950, 1958), a sociologist, combined both field and experimental traditions in the study of functional interdependencies in exchange relations. Inspired by Homans, two psychologists, Thibaut and Kelley (1959), developed the study of small groups as the study of functional interdependencies. Their book is to this day one of the best on the subject. It influenced Homans' further work on the subject (Homans, 1961). Cognitive interdependencies were elaborated by Sherif and turned into a major strand of group research called 'Social Identity Theory' by Tajfel (1982) and Turner (1982), both psychologists but very influential among sociologists and in turn influenced by sociological symbolic interactionism. In a related development, Moscovici, a psychologist,

elaborated Durkheim's theory of collective representations (see Flick, 1998). Berger and Luckmann (1967) elaborated the social construction of reality and Goffman (1961, 1974) showed how this construction might take place in social interaction. Structural interdependencies were more formally elaborated by the psychologist Cartwright and the mathematician Harary and turned into a full-blown network approach by sociologist White and his followers (see White et al., 1976). This approach is now often combined with the exchange approach.

Collective Phenomena That Emerge in Groups

Because for sociologists the emerging collective phenomena are the most important aspects of groups, the remainder of this overview will discuss such collective phenomena. Studies of social dilemmas are covered elsewhere. Due to space restrictions, the reader will be referred mainly to 'collections' rather than individual works.

Groups and Social Embedding

Often, just differences in labeling groups of strangers will induce an instant tendency to favor the 'in-group' (Durham et al., 2011). Individuals need to belong. For sociologists, a particularly interesting aspect is that belonging is not just an individual need and its satisfaction, but a product of processes of social embedding. There is also an epistemological need for shared reality and groups provide such a reality. An individual experience remains ephemeral until it is recognized, shared, and mutually confirmed (Kruglanski et al., 2006). As groups develop, they will engage in activities that make the group recognizable and special, supplying a shared reality (including shared knowledge and collective memory) and collective identity, and often setting the 'in-group' off against 'out-groups.' An important instrument for bringing this embedding about is what Fine (1987), in a famous study on Little League baseball teams, has called 'idioculture.' It consists of shared knowledge, beliefs, behaviors, and customs, such as jokes, nicknames, slang, superstitions, rituals, and taboos. An important part of this process is the emergence of the collective memory, together with ceremonies and commemorative objects (Misztal, 2003; Olick et al., 2011). Sociologists have studied processes that make belonging, shared reality, knowledge, and memory joint products in all sorts of groups: music bands, work groups and teams, sports clubs and teams, primary care units, educational groups (e.g., school classes), and more loosely structured groups, such as social worlds and Internet groups (see, e.g., Engestrom and Middleton, 1998; Lim et al., 2011). Much of group-related ethnographic research in sociology is focused on these processes and much of it is said to be based on 'grounded theory' (see Charmaz and Bryant, 2007). Grounded theory means many things but in this context it refers mainly to four interrelated features: firsthand data collection (as in ethnography); a high degree of contextualizing (i.e., the constant search for social embeddings); varying the (sub)categories used for coding observations in order to facilitate discovery; and moving back and forth between data, analysis, and emerging theory.

There are also quantitative approaches to shared cognitions. For example, the study of 'shared mental models' quantitatively maps group members' similar mental models of content (say of a common task), of structure (say, networked associations of components of the task), and the degree to which the members' models are truly shared in content, structure, and realization of collectiveness (see DeChurch and Mesmer-Magnus, 2010). Yet another approach to social embedding focuses on coorientation (Scheff, 1967) or what is now often called 'interactive belief systems.' Here the emphasis is on the intertwining of (higher orders of) knowledge: "I know that you know that I know," etc. For the coordination of behavior in interacting groups and the development of trust (as a collective phenomenon), the development of coorientation has been identified as vital (Moldoveanu and Baum, 2011). What is taken for granted in game theory (there referred to as 'common knowledge'), is for sociologists an important group product all by itself. Related studies deal with certain aspects of collective knowledge and memory referred to as 'transactive memory' (Hecker, 2012; Lewis and Hemdon, 2011). The idea is that group members can mutually reinforce each other's memories of facts, events, and expectations that are collectively deemed important, which, in turn, increases coorientation and thus communication, trust, and coordination. Another kind of transactive memory refers to distributed memory whereby group members individually remember different things so that the total memory capacity of the group is greatly enhanced (Gupta and Hollinghead, 2010).

The development of social embedding depends in part on the heterogeneity of groups. A growing number of studies focus on such preconditions of the development of collective group phenomena (see, e.g., Chatman and Flynn, 2001; Flache and Mäs, 2008). Such studies are important complements to the studies that focus on social embedding.

Groups and Norms

Groups are the primary producers of social norms (Hechter and Opp, 2001). There are four important kinds of norms. First, norms develop for social embedding in groups. They are specific to the group and help create a social (i.e., group) identity, increasing groupishness and making each group special and different from other groups (see above). Examples are norms about attitudes toward certain out-groups; norms about religious beliefs; norms about what is and is not funny; norms about clothes, hairstyle, and form of greeting. Second, there are norms about positive and negative externalities, such as helping and not hurting others. In contrast to the embedding norms, such norms are often quite general in content across groups (e.g., about helping in need) but specific with regard to the circumstance of application (what constitutes need) and with regard to legitimately expected sacrifice for conforming to the norm (see Lindenberg, 1994). Third, there are norms that deal with the fact that groups have common goals and that attainment of these goals must be facilitated by consensual rules and norms. For example, for teams, there are norms that regulate roles, cooperation,

coordination, and conflict (Mathieu and Rapp, 2009). What are the different roles and responsibilities? How are they related to the common task? When is it necessary to consult others? Who is responsible for completing certain activities? How often and when will individuals meet to discuss issues relating to reaching their common goals? How should conflicts be resolved? Fourth, there are 'metanorms.' They regulate the applicability of and conformity to all sort of norms (such as norms on externalities and norms about the severity of deviance and about desired reactions to deviance, Lindenberg, 1994); and, importantly, they regulate collective memory: (what needs to be remembered, what forgotten). Norms on collective memory also pertain to all three groups of norms (e.g., to remembering commemorative events, feasts, and losses) but also to externalities (when should what effect on others be remembered) and to reaching common goals (such as details of the work design and distribution of responsibilities in teams).

Sociologists have also dealt with 'ghost norms' that are based on a false consensus, in the sense that everybody thinks they apply, whereas everybody believes himself or herself to be the only one who does not subscribe to them. For example, in groups, many members might believe that heavy drinking is expected by the others even though they do not support such a norm themselves. The driving force behind the stability of such ghost norms is a perverse product of an important feature of the stability of norms in general: the metanorm that if you believe that something is an important norm, then you should give social approval to others who follow the norm and disapproval to those who don't, even if you don't subscribe to the norm yourself. In this way, the others interpret your behavior as endorsement of the norm (see Willer et al., 2009).

Groups and Motivation

The motivation to follow group norms and to be oriented toward group goals has been an important subject in group research. Traditionally, in sociology, norm conformity has been linked to socialization processes, in which persons learn to want to do what they have to do. By contrast, modern rational choice explanations maintain that people conform to a norm when rewarded for doing so and punished for not doing so. However, the cognitive interdependence that has long been central to the study of groups leads to a different view – one that has been elaborated in social identity theory, and in cognitive sociology and social psychology. Even though rewards and punishments are important, it is people's ability to identify with group goals ('We' rather than 'I') and the activation of this orientation through cues in the environment that are the major drivers of norm conformity (especially where people are not monitored). Thus, studies that are concerned with the processes of identity formation in groups (Tyler and Blader, 2003; Postmes et al., 2005) and studies that deal with the activation of group orientation (see Lindenberg and Foss, 2011 for overview) have become most important for the explanation of norm conformity in groups. Processes that increase people's social identity will increase their sensitivity to demands relating to group goals and norms (Simpson, 2006). The same will be true for cues that increase a 'normative goal frame' (Lindenberg and Foss, 2011). Much group-related research thus focuses on the conditions that lead to a strengthened group orientation. For example, fairness of procedure and outcomes will strengthen social identity (Tyler and Blader, 2003). Common rituals can serve the same purpose (Islam and Zyphur, 2009). True to the collective aspect of norm conformity, cues that others show respect/disrespect for norms are highly effective in strengthening or weakening a collective orientation (Keizer et al., 2008). This cue effect derives its power from the fact that it is not just limited to imitation (I don't help if others don't help), but that it 'jumps' across norms (e.g., observing somebody else not helping may induce stealing), called a 'cross-norm effect.' Another important influence on norm conformity is 'transformational' group leadership that does not motivate by contingent rewards but by vision, example, and care. Group leaders can develop and articulate a shared vision, they can serve as role models, and they can attend to the needs of group members, thereby increasing trust in the group's collectivity and cohesion (Wang et al., 2011).

For task groups, there is an extra kind of motivation that relates to the joint production of a valued result. The classical example is collective barn raising in villages where all chip in to get the project done. Seemingly, we are hardwired to be especially dedicated to what we perceive to be a common task with individual responsibilities (Lindenberg and Foss, 2011). This contrasts with sentiment groups ('community'), in which there are no individual responsibilities and thus there are the possibilities of depersonalization and social loafing when common tasks suddenly appear (Williams and Karau, 1991). For the special motivation for joint production to occur, the collective orientation must be high (as for sentiment groups). However, in addition, the task must be clear and well understood in order to create a clear picture of one's own responsibilities in relation to those of others. For this reason, increasing the understanding of not just one's own role but also that of others ('cross understanding') is an important contributor to the motivation for joint production (see Huber and Lewis, 2010).

Groups and Status Hierarchies

One of the best-studied collective phenomena with regard to groups is the formation and maintenance of status structures. There are two main approaches that should be distinguished and both apply particularly to task groups. One is based on expected differences in the ability to contribute to a common task (expectation state theory); the other is based on inequalities of resources in social exchange. Expectation state theory (see literature in Berger and Zelditch, 1998) has a long-standing research tradition in sociology. In this theory, following the tradition of Bales in the 1950s, factual power and prestige are conflated and their combined structure is conceived in terms of behavioral aspects (opportunities and influence) rather than the distribution of scarce resources. The higher a person is on the power and prestige structure, the more opportunities to act he or she receives, the more often he or she takes advantage of these

opportunities, and the more often he or she wins disagreements, if and when they occur. However, the basic mechanism that creates this structure lies with cognitive interdependencies: the pair-wise interaction (between, say, Ego and Alter) that is governed by performance expectations with regard to the common task. It is the (unobservable) structure of these performance expectations that keeps generating the observable power and prestige structure. When Ego has positive performance expectations regarding Alter, he or she will give more action opportunities to Alter; and Alter is more likely to take up each opportunity offered. Alter has influence over Ego. It is assumed that there is an evaluation bias. Ego is more likely to judge Alter's performance to be positive if the performance expectation was positive and vice versa. And if Ego nonetheless voices disagreement, he or she is more likely to eventually yield to Alter if the performance expectation was positive to begin with. Such evaluation biases greatly enhance the stability of the power and prestige structure in task groups (provided the task endures), without necessarily undermining the groupishness of the task group.

The link between functional (aspects of the tasks) and cognitive (performance expectations) interdependencies is made even stronger through extensions of the theory to include external status characteristics (see literature in Berger and Zelditch, 1998). Task groups are embedded in a wider culture in which such status and competence beliefs exist (such as 'men are better at mathematics'). People seem to follow the maxim "when in doubt, assume that a salient characteristic is task relevant (especially if it is diffuse, such as race or gender)" (Simpson and Walker, 2002). With regard to gender, this favors men in general tasks as well as in 'masculine' tasks, and it favors women only in stereotypically 'feminine' tasks (see also Webster and Rachotte, 2010). Newer developments are mainly concerned with the ways in which status beliefs arise in group contexts. Seemingly, performance expectations are not only generated by clear perceptions of competence or diffuse status characteristics imported from the outside, but also by persistent inequalities in resources (say rich or poor) that are mis attributed as causally linked to group success. In this way, nominal characteristics can become status characteristics (Ridgeway et al., 2009). A classical variant of these status-building processes (outside expectation state theory but compatible with it) has been suggested by Coleman (1961). He highlights the importance of intergroup rivalry for internal status hierarchies. For example, competitive sports teams in schools develop status hierarchies due to the players' contributions to success in intercollegiate competition.

Dependency Networks, Groupishness, and Individual Differences

Status differences in groups can also come about by a process that is independent of expectation states. As Blau (1964) has shown in a pioneering study, asymmetric exchanges among group members can lead to status structures in task groups. The central idea is that people cherish status and thus also cherish being called superior by others. This is likely to happen if a group member needs something from another group member but has nothing else to offer than to admit his or her inferiority and the superiority of the other (who cherishes this gesture and, in return, is thus willing to give up information or other valued resources). If exchanges are made among members who are unequal in resources, this can lead to inequality in exchange which, in turn, can lead to people claiming (and others accepting) superior status. Newer versions of this theory (in combination with the exchange theory by Emerson, 1976) deal mainly with power-dependency differences rather than status structures, even though some theorists link the two by using status effects to explain power-dependency effects (Thye et al., 2006). In contrast to status research, power-dependence research emphasizes functional and structural dependencies (especially dependencies due to possible exclusion) and de-emphasizes attention to cognitive interdependencies, as mutual perceptions play no important role (see Cook et al., 2006; Willer, 1999). Thereby they focus less on the group embeddedness of social networks than approaches that include cognitive interdependencies. Still, the structural emphasis has its own advantage in that it shows that the network structures (that are possibly embedded in groups) have their own considerable influence that may undermine groupishness in favor of gain-related and/or affect-related individual outcomes (Burt, 2002); and that this influence is mostly due to dependencies (Bonacich and Bienenstrock, 2009). The structural focus also shows that seemingly individual characteristics (such as being powerful, being successful, being a failure, having high or low self-esteem, etc.) can be collective phenomena generated by social structures. It is also evident that for fostering groups, pure power asymmetries do not work well, compared to generalized exchange structures (Willer et al., 2012). With generalized exchange, cognitive interdependencies (in the form of group identification) are more likely to emerge. As we have seen above, when there are power differences, it is the powerful members' contribution to group goals (best aided by transformational leadership) that can neutralize the negative effects of power differentials and stress positive cognitive interdependencies, thereby fostering groupishness (Wang et al., 2011). Finally, there is also research that deals with the reverse process: certain cognitive interdependencies (in the form of reflected appraisals of one's own position) can be the source rather than the consequence of power differentials within groups (Friedkin, 2011) that may or may not support groupishness.

In short, groups give rise to social embeddings, to norms, to collectively induced motivations, and to status structures. However, embedded dependency networks within groups and the resulting power structures may undermine groupishness to various degrees, producing instead structured differences in individual characteristics. Contributions to group goals by the powerful may neutralize this potentially negative effect of dependency structures.

Evaluation

Sociological research on groups mainly deals with collective phenomena that emerge in groups, such as social embedding (including belonging and identity, shared knowledge and

reality, collective memory), social norms, collectively produced motivation, and status structures. Ideally, the study of groups include attention to functional, cognitive, and structural interdependencies, but often the focus is only on one or two of these three structures. There is much productive research going on but, unfortunately, it is spread over a wide variety of university departments and journals, ranging from social psychology, sociology, management, business, to social movements, communication, and simulation journals. Groups themselves are presently not a 'field' in sociology, even though they are the wellspring of society's most important collective fabrics. A common field would make sense, given the communality among the various types of groups that are now dealt with as separate topics in the literature. It is the articles in encyclopedias that keep at least a semblance of a field called 'groups' alive.

See also: Collective Memory, Social Psychology of; Conformity: Sociological Aspects; Cooperation and Competition; Cooperation: Sociological Aspects; Deindividuation, Psychology of; Evolutionary Group Dynamics; Exchange: Social; Gangs, Sociology of; Group Processes in Organizations; Group Processes in the Classroom; Group Processes, Social Psychology of; Identity and Identification, Social Psychology of; Social Dilemmas, Psychology of; Social Identity in Sociology; Teamwork and Team Performance Measurement; Tribe; Youth Gangs.

Bibliography

Baumeister, R.F., Leary, M.R., 1995. The need to belong: desire for interpersonal attachments as a fundamental human motivation. Psychological Bulletin 117, 497–529.

Berger, P., Luckmann, T., 1967. The Social Construction of Reality. Anchor Books, Garden City, NY.

Berger, J., Zelditch, M., 1998. Status, Power, and Legitimacy: Strategies and Theories. Transaction, New Brunswick, NJ.

Blau, P., 1964. Exchange and Power in Social Life. Wiley, New York.

Bonacich, P., Bienenstock, E.J., 2009. Structural implications of reciprocal exchange: a power-dependence approach. Social Psychology Quarterly 72, 226–240.

Burt, R.S., 2002. The social capital of structural holes. In: Guillen, M.F., Collins, R., England, P., Meyer, M. (Eds.), The New Economic Sociology: Developments in an Emerging Field. Russell Sage Foundation, New York.

Charmaz, K., Bryant, A., 2007. Sage Handbook of Grounded Theory. Sage, London.

Chatman, J.A., Flynn, F., 2001. The influence of demographic heterogeneity on the emergence and consequences of cooperative norms in work teams. Academy of Management Journal 44, 956–974.

Coleman, J.S., 1961. The Adolescent Society. Free Press, Glencoe IL.

Cook, K.S., Cheshire, C., Gerbasi, A., 2006. Power, dependence, and social exchange. In: Burke, P. (Ed.), Contemporary Social Psychological Theories. Stanford University Press, Stanford, CA, pp. 194–216.

DeChurch, L.A., Mesmer-Magnus, J.R., 2010. Measuring shared team mental models: a meta-analysis. Group Dynamics: Theory, Research, and Practice 14, 1–14.

Durham, Y., Baron, A.S., Carey, S., 2011. Consequences of 'minimal' group affiliations in children. Child Development 82, 793–811.

Emerson, R.M., 1976. Social exchange theory. Annual Review of Sociology 2, 335–362.

Engestrom, Y., Middleton, D., 1998. Cognition and Communication at Work. Cambridge University Press, Cambridge.

Fine, G.A., 1987. With the Boys. University of Chicago Press, Chicago, IL.

Flache, A., Mäs, M., 2008. Why do faultlines matter? A computational model of how strong demographic faultlines undermine team cohesion. Simulation Modelling Practice and Theory 16, 175–191.

Flick, U. (Ed.), 1998. The Psychology of the Social. Cambridge University Press, Cambridge.

Friedkin, N.E., 2011. A formal theory of reflected appraisals in the evolution of power. Administrative Science Quarterly 56, 501–529.

Goffman, E., 1974. Frame Analysis. Free Press, New York.

Goffman, I., 1961. Encounters. Bobbs-Merrill, Indianapolis, IN.

Gupta, N., Hollinghead, A.B., 2010. Differentiated versus integrated transactive memory effectiveness: it depends on the task. Group Dynamics: Theory, Research, and Practice 14, 384–398.

Hechter, M., Opp, K.D. (Eds.), 2001. Social Norms. Russell Sage, New York.

Hecker, A., 2012. Knowledge beyond the individual? Making sense of a notion of collective knowledge in organizations. Organization Studies 33, 423–445.

Heine, S., Proulx, T., Vohs, K., 2006. The meaning maintenance model: on the coherence of social motivations. Personality and Social Psychology Review 10, 88–110.

Homans, G.C., 1950. The Human Group. Harcourt, Brace & World, New York.

Homans, G.C., 1958. Social behavior as exchange. American Journal of Sociology 63, 597–606.

Homans, G.C., 1961. Social Behavior: Its Elementary Forms. Harcourt, Brace & World, New York.

Huber, G.P., Lewis, K., 2010. Cross-understanding: implications for group cognition and performance. Academy of Management Review 35, 6–26.

Islam, G., Zyphur, M.J., 2009. Rituals in organizations: a review and expansion of current theory. Group Organization Management 34, 114–139.

Keizer, K., Lindenberg, S., Steg, L., 2008. The spreading of disorder. Science 322, 1681–1685.

Kruglanski, A.W., Pierro, A., Mannetti, L., De Grada, E., 2006. Groups as epistemic providers: need for closure and the unfolding of group-centrism. Psychological Review 113, 84–100.

Lewis, K., Herndon, B., 2011. Transactive memory systems: current issues and future directions. Organization Science 22, 1254–1265.

Lim, S., Cha, S.Y., Park, C., Lee, I., Kim, J., 2011. Idioculture in crowd computing: a focus on group interaction in an event-driven social media system. International Journal of Human-Computer Studies 69, 632–646.

Lindenberg, S., 1994. Norms and the power of loss: Ellickson's theory and beyond. Journal of Institutional and Theoretical Economics 150, 101–113.

Lindenberg, S., 1997. Grounding groups in theory: functional, cognitive, and structural interdependencies. Advances in Group Processes 14, 281–331.

Lindenberg, S., Foss, N., 2011. Managing joint production motivation: the role of goal-framing and governance mechanisms. Academy of Management Review 36, 500–525.

Mathieu, J.E., Rapp, T.L., 2009. Laying the foundation for successful team performance trajectories: the roles of team charters and performance strategies. Journal of Applied Psychology 94, 90–103.

Misztal, B., 2003. Theories of Social Remembering. Open University Press, Maidenhead, PA.

Moldoveanu, M.C., Baum, J.A., 2011. "I Think You Think I Think You're Lying": the interactive epistemology of trust in social networks. Management Science 57, 393–412.

Olick, J.K., Vinitzky-Seroussi, V., Levy, D., 2011. The Collective Memory Reader. Oxford University Press, Oxford.

Postmes, T., Haslam, S.A., Swaab, R.I., 2005. Social influence in small groups. European Review of Social Psychology 16, 1–42.

Ridgeway, C.L., Backor, K., Li, Y.E., Tinkler, J.E., Erickson, K.G., 2009. How easily does a social difference become a status distinction: gender matters. American Sociological Review 74, 44–62.

Scheff, T.J., 1967. Towards a sociological model of consensus. American Sociological Review 32, 32–46.

Simpson, B., 2006. Social identity and cooperation in social dilemmas. Rationality and Society 18, 444–470.

Simpson, B., Walker, H.A., 2002. Status characteristics and performance expectations: a reformulation. Sociological Theory 20, 24–40.

Steverink, N., Lindenberg, S., 2006. Which social needs are important for subjective well-being? What happens to them with aging? Psychology and Aging 21, 281–290.

Tajfel, H., 1982. Instrumentality, identity and social comparisons. In: Tajfel, H. (Ed.), Social Identity and Intergroup Relations. Cambridge University Press, Cambridge, pp. 483–507.

Thibaut, J., Kelley, H.H., 1959. The Social Psychology of Groups. Wiley, New York.

Thye, S.R., Willer, D., Markovsky, B., 2006. From status to power: new models at the intersection of two theories. Social Forces 84, 1471–1495.
Turner, J.C., 1982. Toward a cognitive redefinition of the social group. In: Tajfel, H. (Ed.), Social Identity and Intergroup Relations. Cambridge University Press, Cambridge, pp. 15–40.
Tyler, T.R., Blader, S., 2003. Procedural justice, social identity, and cooperative behavior. Personality and Social Psychology Review 7, 349–361.
Wang, G., Oh, I., Courtright, S.H., Colbert, A.E., 2011. Transformational leadership and performance across criteria and levels: a meta-analytic review of 25 years of research. Group & Organization Management 36, 223–270.
Webster Jr., M., Rashotte, L.S., 2010. Behavior, expectations and status. Social Forces 88, 1021–1049.
White, H.C., Boorman, S.A., Breiger, R.L., 1976. Social structure from multiple networks: I. Blockmodels of roles and positions. American Journal of Sociology 81, 730–780.
Willer, D., 1999. Network Exchange Theory. Praeger, Westport CT.
Willer, R., Flynn, F.J., Zak, S., 2012. Structure, identity, and solidarity: a comparative field study of generalized and direct exchange. Administrative Science Quarterly 57, 119–155.
Willer, R., Kuwabara, K., Macy, M.W., 2009. The false enforcement of unpopular norms. American Journal of Sociology 115, 451–490.
Williams, K., Karau, S.J., 1991. Social loafing and social compensation: the effects of expectations of co-worker performance. Journal of Personality and Social Psychology 61, 570–581.

Growth Curve Analysis

John J McArdle, University of Virginia, Charlottesville, VA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 9, pp. 6439–6445, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article begins with an introduction of the intellectual history of the development of growth curve analysis since the 1750s. It then introduces the prevailing linear and nonlinear models for growth curve analysis for univariate as well as multivariable situations. It ends with a discussion of issues with growth curve modeling in the recent literature, especially how it relates to structural equation modeling, more general differential equation models, and Bayesian estimation.

The phrase 'growth curve analysis' denotes the processes of describing, testing hypotheses, and making scientific inferences about the growth and change patterns in a wide range of time-related phenomena. The term 'growth curve' was originally used to describe a graphic display of physical stature (e.g., the height or weight) of an individual over consecutive ages, and some early data collections are presented here. Techniques for the analysis of growth curves were initiated in the physical sciences, more fully developed in the biological sciences, and used in studies of the size and health of plants, animals, and humans. In the behavioral sciences, growth curve analyses have been applied routinely to a wide range of phenomena – from experimental learning curves, to the growth and decline of intellectual abilities and academic achievements, and to changes in other psychological traits over the life span. Formal models for the analysis of growth data are described in a nontechnical fashion including linear and latent growth models, nonlinear growth models, and multivariate growth models. This article includes recent historical developments, contemporary issues, and future directions.

In contrast to some longitudinal data, growth curves have unique features: (1) the same entities are repeatedly observed, (2) the same procedures of measurement and scaling of observations are used, and (3) the timing of the observations is known. Formal models for the analysis of growth curves were developed in many different substantive domains with a common goal – to examine and uncover a fundamental set of regularity conditions or 'basic functions' responsible for the manifest growth and change. Techniques for the analysis of growth curves were initiated in the physical sciences, more fully developed in the biological sciences, and then used in studies of the size and health of plants, animals, and humans. In the behavioral sciences, growth curve analyses have been applied routinely to a wide range of phenomena – from experimental learning curves, to the growth and decline of intellectual abilities and academic achievements, and to changes in other psychological traits over the life span.

Growth curve analyses are among the most widely studied and well-developed mathematical and statistical techniques in all scientific research. These analyses have roots in the seventeenth- and eighteenth-century calculus of Newton and probability of Pascal, but this article is limited to more recent historical developments and other related techniques, such as time series and dynamical systems analyses are not covered.

Growth Curve Data Collections

The first reported growth curve data were collected by the French Count de Montbeillard between 1759 and 1777. These data consist of semiannual measurements on the growth of the height of his son over nearly 18 years. The data were plotted by Scammon (1927) and are presented here in the upper curve of **Figure 1**. As Scammon reported,

> it will be noted that the curve shows the typical four phases which most modern students have observed in the postnatal growth in stature of man, and which are characteristic of the growth of so many parts of the body.
>
> p. 331

The first analysis of these data was by the naturalist Buffon in 1799 and

> should be given full credit for the discovery of seasonal differences in growth a full hundred years before the modern investigation of this work.
>
> p. 334

The lower growth curve in **Figure 1** is based on group averages obtained by Variot in 1908 and

> it is interesting to note that, while the absolute values of the two series are quite different, the general form of the curve is essentially the same in both instances.
>
> p. 335

These early growth curves were the precursor to an enormous collection of biological data on growth and change. A more recent illustration comes from the important work of Tanner and his colleagues (see Tanner, 1960). In the individual plots of height in **Figure 2**: (1) the 'velocity' or rate of change is plotted at each age and (2) these curves are plotted against their

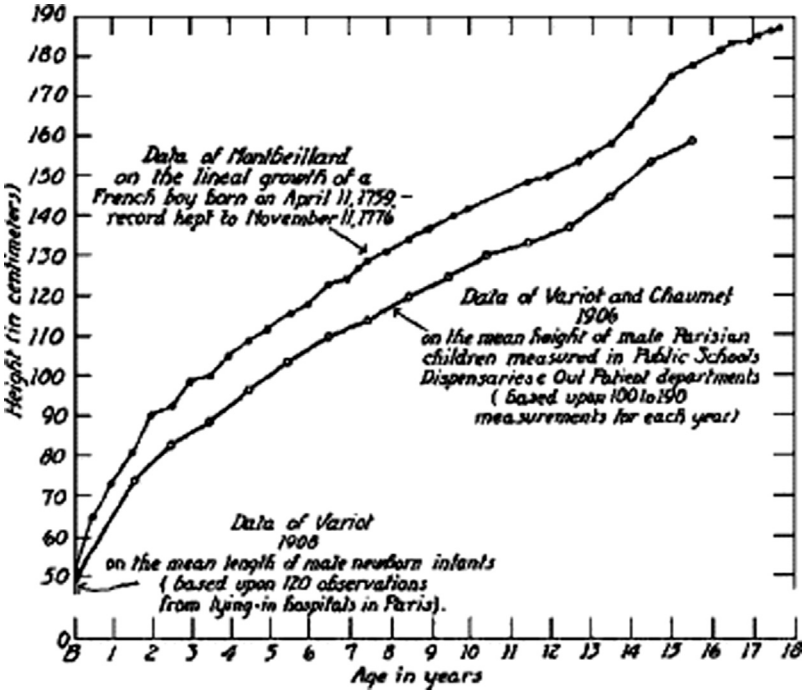

Figure 1 The initial 'growth curves' of human height data from Scammon, R.E., 1927. The first seriatim study of human growth. American Journal of Physical Anthropology 10, 334. Note: The vertical (*y*) axis represents the height in centimeters and the horizontal (*x*) axis represents the age in years from birth.

own 'highest peak velocity.' This display demonstrated two interesting features of physical growth: (1) persons who start growth at the earliest ages also grow the most but (2) all individuals share a remarkably similar shape in their 'adolescent growth spurt.' Extensions of these basic findings on physical growth systems and the contrast of chronological time versus biological time are still used in contemporary research.

Experimental psychologists have routinely collected different kinds of growth curves. Among the first here were the classical 'forgetting' curves collected by Ebbinghaus in the

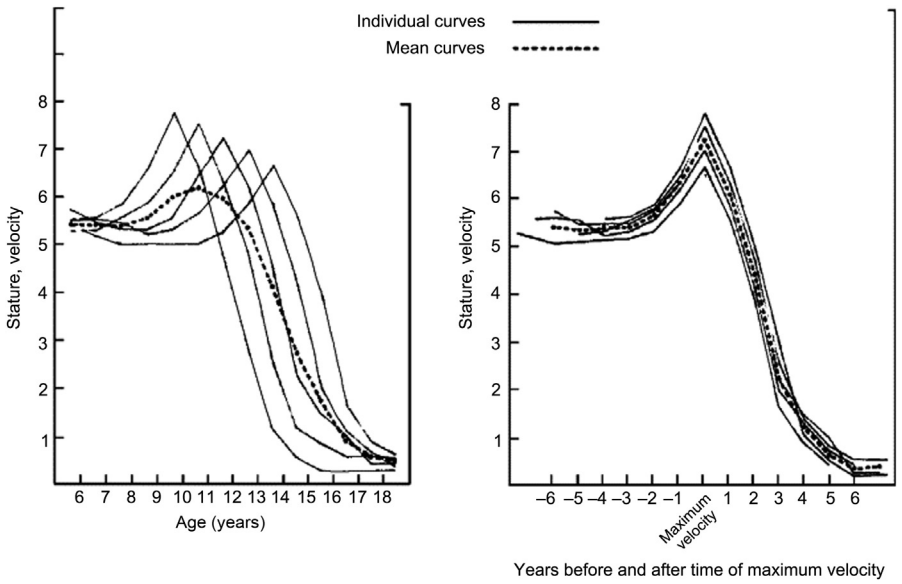

Figure 2 Alternative 'velocity curves' of physical stature from Tanner, J.M. (Ed.), 1960. Human Growth. Pergamon Press, New York, p. 22. Note: The left-hand graph shows the change in height as a function of the change in time (*y*) vs the age in years (*x*), while the right-hand graph shows the same curves (*y*) plotted around the time of maximum change in scores (*x*) for each individual.

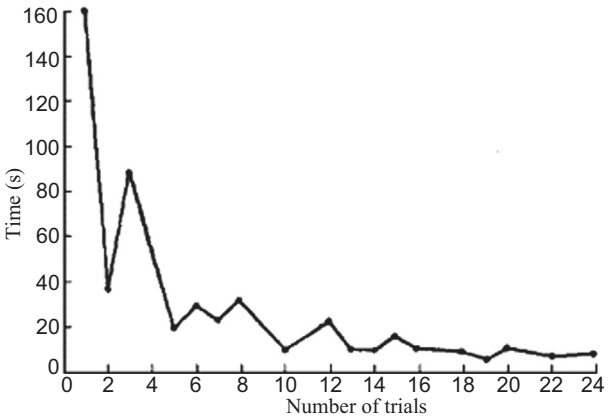

Figure 3 The 'learning curve' of a cat from Thorndike, E.L., 1911. Animal Intelligence. MacMillan, New York; as reported by Garrett, H.E., 1951. Great Experiments in Psychology. Appleton-Century-Crofts, New York, p. 44. Note: The response time (*y*) required for a cat to escape from a closed box by pulling on a wire as a function of consecutive number of trials (*x*).

1880s, whose work introduced the use of quantitative methods in the study of learning and memory and stimulated many experimental data collections. Included here are the 'curves of improvement' in studies on telegraphic language by Bryan and Harter in 1897. These principles were also used in the animal learning curve experiments of Thorndike in 1911, and one example of such data appears in **Figure 3**. This plot illustrates an example of 'trial and error' learning: learning was objectively defined by decreasing response time and the lack of smooth function over trials was considered as error. Thorndike used these growth curves to illustrate several classical principles of learning, including the 'law of exercise' and the 'law of effect' (see Garrett, 1951).

Differential psychologists have also contributed growth data in many different substantive areas. One good example of this tradition is given in the plots of **Figure 4** from work of Bayley (1956). Individual growth curves of mental abilities from birth to 25 years are plotted for a selected set of boys and girls from the well-known Berkeley Growth Study. Because mental ability was not easily measured in exactly the same way at each age, these individual curves were created by adjusting the means and standard deviations of different mental ability tests (i.e., Stanford–Binet, Terman–Merrill) at different ages into a common metric. As Bayley says,

> they are not in 'absolute' units, but they do give a general picture of growth relative to the status of this group at 16 years … These curves, too, are less regular than the height curves, but perhaps no less regular than the weight curves. One gets the impression both of differences in rates of maturing and of differences in inherent capacity.
>
> 1956: p. 66

This use of 'linked' measurement scales created a novel set of growth data and permitted the use of growth analyses initially derived in other scientific areas.

Linear Models for Growth Curve Analyses

The simplest growth models are based on the simplest mathematical functions. The fitting of a straight-line to a set of measurements, a standard procedure in scientific research, was an initial model used by Wishart (1938) in a classic paper on growth analysis. These linear models were quite practical because the calculations used standard statistical assumptions (e.g., independence and normality of the residuals), standard forms of estimation (e.g., Gauss–Markov least squares, maximum likelihood), and standard statistical tests of goodness-of-fit (*F*-tests, chi-square tests, and likelihood ratio tests).

Another key idea was also introduced by Wishart (1938) when he showed how power polynomials could be used to

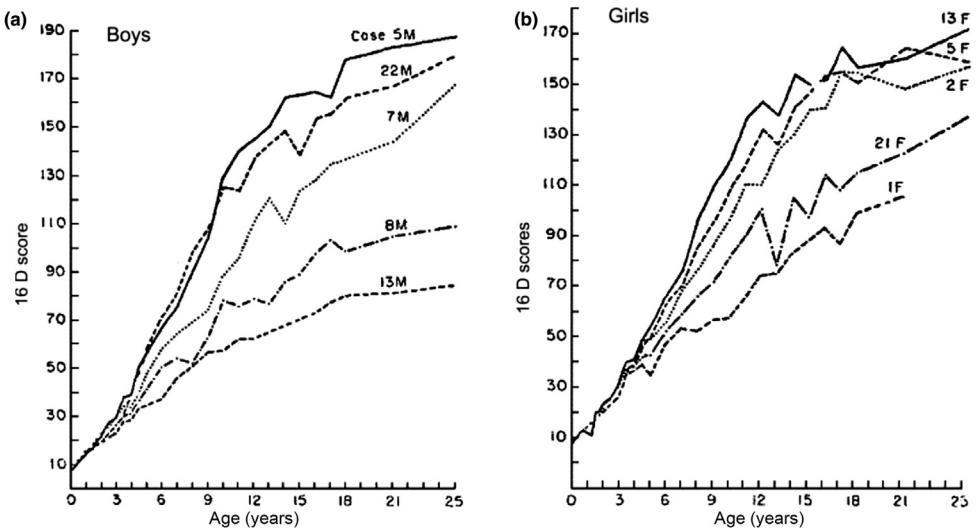

Figure 4 Growth curves of intellectual abilities in selected boys and girls from Bayley, N., 1956. Individual patterns of development. Child Development 27 (1), 67. Note: Age 16 D scores (*y*) plotted as a function of age at measurement (*x*) for (a) five boys and (b) five girls.

fit the nonlinear curvatures apparent in growth data. The individual growth curve (consisting of $t = 1$, T occasions) is summarized into a small set of linear orthogonal polynomial coefficients based on a power series of time (t, t^2, t^3, … , t^p) describing the general nonlinear shape of the growth curve. A model of growth data might require a second-order (quadratic), third-order (cubic), or even higher-order polynomial model fitted to the data, but the basic shape of each individual curve could be captured with a small number of parameters.

The organization of individual differences is an important consideration in any contemporary growth model. In the linear model, each individual is assumed to be following a straight-line over time, describable by an intercept term and a slope term. The average of these individual scores can be used to describe the overall growth curve for the group, and the individual differences can be considered in the context of other substantive questions, e.g., do the groups under observation differ in intercepts or slopes? In the more complex polynomial models, Wishart suggested a second stage of analysis where these polynomial coefficients describing each individual curve were considered as the underlying scores that covary among different individuals. In a quadratic polynomial, for example, a person's trajectory over time could be summarized using an individual intercept, linear slope, acceleration, and a residual term, and secondary questions could be addressed, e.g., do the groups under observation also differ in acceleration?

This 'curve-fitting' or 'two-stage' approach was relatively easy to implement using a standard statistical basis for hypothesis testing (e.g., Bock, 1975). The simplicity of this approach allowed linear growth models to enjoy widespread use in behavioral studies and become a useful starting point for understanding individual differences in change (e.g., Rogosa and Willett, 1985). The persistent critiques of this polynomial approach include some key issues:

1. The individual parameters can be difficult to estimate, especially for higher-order terms and incomplete data.
2. Even high-order polynomials (i.e., $p > 4$) cannot capture 'seasonality' or 'phase-shifts.'
3. The substantive meaning of these individual parameters is often in doubt.
4. The individual level parameters can be transformed (i.e., rotated) to yield alternative meaning.

These standard critiques are appropriate, and they still apply to many other contemporary forms of growth curve analyses.

More complex linear models have been used to represent growth. A linear model applied to different age segments but with connected critical 'knot points' permitted a 'conjoined' or linear or polynomial 'spline' model (e.g., Bryk and Raudenbush, 1992). This kind of spline model is not often used, partly because it is atheoretical, and partly because it may require a relatively large number of parameters to achieve adequate fit. A more parsimonious organization of individual differences was proposed in the form of 'latent curves,' independently defined by both Rao and Tucker in 1958 (cited in Meredith and Tisak, 1990). Principal components of the repeated observations were used to reorganize these data into the sum of a small number of unspecified linear functions that are fitted to growth data, the shape of the latent curves is determined by the component loadings, and the individual curve parameters are the component scores. The summation of latent curves has roots in the classical nineteenth-century work of Fourier as now widely used. In a recent innovation, Meredith and Tisak (1990) showed how these 'Tuckerized curve' models could be represented using confirmatory common factor analysis models, used to represent a wide range of alternative models (including most of the models discussed herein), and fitted using structural equation modeling techniques (also see McArdle, 1988; McArdle and Bell, 2000; McArdle and Woodcock, 1997).

Nonlinear Models for Growth Curve Analysis

The linear models can be used to describe a variety of nonlinear shapes, but many growth curve models for specific substantive problems have included nonlinear functions of the parameters. For example, Ebbinghaus in 1880 described his forgetting curves using a form of the classic 'exponential growth' model. Here the rate of change is a linear function of the percentage of initial size (e.g., compound interest), and individual trajectories rise or fall in nonlinear fashion toward some asymptotic values. In contrast, the nineteenth-century Velhurst curve of population growth, an S-shaped 'logistic curve,' was used in the early twentieth century by Pearl for many forms of growth. At about the same time, Peters advocated an 'ogival' curve of growth in ideational learning, Thurstone found that a hyperbolic curve of learning is best to fit the norms of many different tests, and both Ettlinger and Valentine demonstrated direct relationships among these functions (see Seber and Wild, 1989; Bock and Thissen, 1980).

More complex nonlinear models have been used extensively. In about 1825, Gompertz described the growth curve in terms of two exponential accumulations of different rates toward different asymptotes. This flexible model was studied in the first half of the twentieth century by Winsor, used by Medwar to study the growth of chicken hearts, and by Deming for human physical growth and in studies of learning curves (see Browne and du Toit, 1991). Another popular nonlinear model, introduced around 1938 by von Bertalanffy, suggested that an individual's change in weight was the direct result of the difference in the forces of anabolism and catabolism, with the exact relationship specified by known empirical properties. In subsequent work, Richards challenged the empirical basis of this model, and formally showed how many prior models can be seen as specific solutions of a 'family' of deterministic differential equations; specific restrictions led to the exponential, logistic, Gompertz, and von Bertalanffy equations. This work was more recently extended by Nelder and Sandland and McGilchrist (1979). (See also Seber and Wild, 1989; Zeger and Harlow, 1987; Tsoularis and Wallace, 2002.)

Attempts to fit a single growth model to observations over a wide range of ages with a minimal set of parameters led researchers to combine aspects of other models. Jenss and Bayley in 1937 proposed a simple composite where a linear part was used to fit the early rapid growth phase and this was added to an exponential part to fit negative acceleration of the later slowing-down phase. A more complex composite model, the sum of multiple logistic curves, was suggested by Robertson

in 1908 and Burt in 1937 but only made practical and fully developed by Bock and his colleagues (see Bock and Thissen, 1980). In each segment, the rate of growth (1) increases early, (2) reaches a maximum (peak growth velocity), (3) decreases toward the asymptote, and (4) the final value of one segment is used as the starting value of the next segment. A novel composite model based on multiple logistic functions was developed by Preece and Baines in 1978 and here the model parameters were supposed to be linearly related to meaningful biological parameters. Simple versions of this new family of curves have proven useful in contemporary work (see Hauspie et al., 1991).

The mathematical elegance of these nonlinear formulations often make it seem that the parameters estimated are related directly to the underlying phenomena and there is a tendency to treat these estimates with direct substantive interpretation. Prior research has shown this to be a difficult goal and now it is known that all models permit alternative formulations that may fit equally well. Another important property of growth curves comes when the fixed parameters of the nonlinear curves, the 'parameters of the means' also represent the 'means of the parameters' of the individual curves. However, classical research by Merrill in 1931, Estes in 1957, Tucker in 1966, and Keats in 1983 have shown this form of 'dynamic consistency' does not always hold, even for some of the most popular growth models (e.g., a group exponential). Contemporary statistical models are now constructed carefully to clarify the unique substantive interpretation of nonlinear parameters (see Browne and du Toit, 1991; Bock and Thissen, 1980).

Multivariate Growth Curve Analyses

The collection of multiple variables at each occasion of measurement leads naturally to questions about relationships among growth processes and multivariate growth models. In biological research, early efforts were directed at characterizing the parallel properties of different growth variables. Models were developed originally to deal with the size of two different organs, and an early nineteenth-century work was used by Huxley to form a classical 'allometric' model where two variables have a constant ratio of growth rates throughout the growth period. Many physical processes were subsequently found to grow in parallel, or in an ordered time-sequence (e.g., Tanner, 1960). This work led to more sophisticated models based on systems of differential equations for the size of multiple variables. In one comprehensive multivariate model, Turner in 1978 extended the simple growth principles to more variables and permitted an examination of biologically important interactions based on the size and sign of the estimated parameters (see Griffiths and Sandland, 1984).

Multivariate research in the behavioral sciences has relied on advanced versions of the linear growth models formalized by Rao, and Pothoff and Roy (see Bock, 1975). The structural models developed within this framework have added measurement models and this includes work on multivariate models of latent curves (McArdle, 1988). These structural models have proven useful in testing hypotheses about parallel growth processes across time in multiple psychological variables. In recent examples, McArdle and Woodcock (1997) showed how latent growth of a single common factor of intellectual abilities is a poor representation of multiple abilities data and McArdle (2000) showed how a multivariate dynamic sequence could be fitted using standard structural equations. Current research in this area is focused on an improved understanding of the time-sequences and interactions among multiple behavioral and biological variables (e.g., Li et al., 2000).

Recent Issues in Growth Curve Analysis

The prior developments in growth curve analysis represent an important but limited class of longitudinal data analyses (e.g., Nesselroade and Baltes, 1979; Bock, 1975; Pinherio and Bates, 2000). These developments have been distinguished by features of the models themselves: (1) the degree of mathematical specification, (2) the way these statistical models are fitted, and (3) the clarity and substantive meaning of the results. In future work, most of these problems are likely to remain under active investigation.

Most contemporary growth curve models can be written in a common symbolic form (Seber and Wild, 1989). A general model for a change in the scores over time (often using derivatives dy/dt or differences $\Delta y/\Delta t$) can be based on some mathematical functional form ($f\{x\}$) with unobserved scores ($x[t]$) and with unknown parameters ($\beta[t]$) to be estimated. Additional statistical forms not discussed here can be included, including autoregressive residual error structures, Markov chains, and Poisson processes. The growth functions described here are not an exhaustive listing and future extensions to more generalized functions, and to dynamic and chaotic forms are likely.

Recent research in statistical model fitting is important here too. It is now clear that growth curve models of arbitrary complexity can be fitted to the observed trajectory over time (i.e., the integral) and the unknown parameters can be estimated to minimize some statistical function (e.g., weighted least squares, maximum likelihood), and parameter estimates obtained using nonlinear programming. This generic statistical approach avoids some problems in older techniques, such as fitting a model to a log scale or directly to the velocity or the analysis of difference score data (**Figure 2**). These new techniques make it possible to address the critical problems of forecasting future observations, and further research on Bayesian estimation is the focus of current research here (for details, see Seber and Wild, 1989).

This generic model-fitting approach also permits a wide range of new possibilities for dealing directly with unbalanced, incomplete, or missing data. In early work, linear polynomials were used extensively to deal with these kinds of problems. But recent work on linear and nonlinear mixed and multilevel models have surfaced, making it possible to estimate growth curves and test hypotheses by collecting only small segments of data on each individual (Bryk and Raudenbush, 1992; Pinherio and Bates, 2000; Verbeke and Molenberghs, 2000; Preacher et al., 2006). These statistical models are being used in many longitudinal studies to deal with self-selection and subject attrition, multivariate changes in dynamic patterns of development, and the trade-offs between statistical power and costs

of person and variable sampling. The statistical power questions of the future are not 'how *many* occasions do we need?' but 'how *few* occasions are adequate?' and 'which persons should we measure at which occasions?' (McArdle and Woodcock, 1997; McArdle and Bell, 2000; Wu et al., 2009).

Some of the most difficult problems for future work on growth curves are related to the elusive meaning of the growth model parameters. Growth curve analyses also rely on basic measurement properties including (1) interval scaling, (2) age-and-time equivalence, and (3) substantively meaningful. These issues are key problems in the behavioral sciences and future research is needed to address these fundamental concerns. Multivariate models may be useful here, especially as these models are further extended to a fully dynamic time-dependent form. In any case, empirical information and experimentation will be needed to judge the utility of any growth curve model.

Given the long history of elegant formulations from mathematics and statistics in this area, it is somewhat humbling to note that many of the most insightful growth curve analyses have come from careful visual inspection of the growth curves (e.g., in **Figures 1–4**). The insight gained from visual inspection of a set of growth curves is not in dispute now and in fact obvious visual features should be highlighted and emphasized in future research (e.g., Pinherio and Bates, 2000). As in the past, the best future growth curve analyses are likely to be the ones we can see most clearly.

See also: Panel Surveys: Uses and Applications; Reliability: Measurement; Structural Equation Modeling.

Bibliography

Bayley, N., 1956. Individual patterns of development. Child Development 27 (1), 45–74.

Bock, R.D., 1975. Multivariate Statistical Methods in Behavioral Research. McGraw-Hill, New York.

Bock, R.D., Thissen, D., 1980. Statistical problems of fitting individual growth curves. In: Johnston, F.E., Roche, A.F., Susanne, C. (Eds.), Human Physical Growth and Maturation: Methodologies and Factors. Plenum, New York, pp. 265–290.

Browne, M., du Toit, S.H.C., 1991. Models for learning data. In: Collins, L., Horn, J.L. (Eds.), Best Methods for the Analysis of Change. APA Press, Washington, DC, pp. 47–68.

Bryk, A.S., Raudenbush, S.W., 1992. Hierarchical Linear Models: Applications and Data Analysis Methods. Newbury Park, CA.

Fischer, G.H., Molenaar, I.W., 1995. Rasch Models: Foundations, Recent Developments, and Applications. Springer, New York.

Garrett, H.E., 1951. Great Experiments in Psychology. Appleton-Century-Crofts, New York.

Griffiths, D., Sandland, R., 1984. Fitting generalized allometric models to multivariate growth data. Biometrics 40, 139–150.

Hauspie, R.C., Lindgren, G.W., Tanner, J.M., Chrzastek-Spruch, H., 1991. Modeling individual and average growth data from childhood to adulthood. In: Magnusson, D., Bergman, L.R., Rudinger, G., Törestad, B. (Eds.), Problems and Methods in Longitudinal Research: Stability and Change. Cambridge University Press, Cambridge, GB.

Li, F., Duncan, T.E., Acock, A., 2000. Modeling interaction effects in latent growth curve models. Structural Equation Modeling 7, 497–533.

McArdle, J.J., 1988. Dynamic but structural equation modeling of repeated measures data. In: Nesselroade, J.R., Cattell, R.B. (Eds.), The Handbook of Multivariate Experimental Psychology, vol. 2. Plenum Press, New York, pp. 561–614.

McArdle, J.J., 2000. A latent difference score approach to longitudinal dynamic structural analyses. In: Cudeck, R., du Toit, S., Sorbom, D. (Eds.), Structural Equation Modeling: Present and Future. Scientific Software International, Lincolnwood, IL, pp. 342–380.

McArdle, J.J., Bell, R.Q., 2000. Recent trends in modeling longitudinal data by latent growth curve methods. In: Little, T.D., Schnabel, K.U., Baumert, J. (Eds.), Modeling Longitudinal and Multiple-Group Data: Practical Issues, Applied Approaches, and Scientific Examples. Erlbaum, Mahwah, NJ, pp. 69–107.

McArdle, J.J., Woodcock, J.R., 1997. Expanding test-rest designs to include developmental time-lag components. Psychological Methods 2 (4), 403–435.

Meredith, W., Tisak, J., 1990. Latent curve analysis. Psychometrika 55, 107–122.

Nesselroade, J.R., Baltes, P.B. (Eds.), 1979. Longitudinal Research in the Study of Behavior and Development. Academic Press, New York.

Pinherio, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-PLUS. Springer, New York.

Preacher, K.J., Curran, P.J., Bauer, D.J., 2006. Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. Journal of Educational and Behavioral Statistics 31, 437–448.

Rogosa, D., Willett, J.B., 1985. Understanding correlates of change by modeling individual differences in growth. Psychometrika 50, 203–228.

Sandland, R.L., McGilchrist, C.A., 1979. Stochastic growth curve analysis. Biometrics 35, 255–271.

Scammon, R.E., 1927. The first seriatim study of human growth. American Journal of Physical Anthropology 10, 329–336.

Seber, G.A.F., Wild, C.J., 1989. Nonlinear Models. Wiley, New York.

Tanner, J.M. (Ed.), 1960. Human Growth. Pergamon Press, New York.

Thorndike, E.L., 1911. Animal Intelligence. MacMillan, New York.

Tsoularis, A., Wallace, J., 2002. Analysis of logistic growth models. Mathematical biosciences 179, 21–55.

Verbeke, G., Molenberghs, G., 2000. Linear Mixed Models for Longitudinal Data. Springer, New York.

Wishart, J., 1938. Growth rate determinations in nutrition studies with the bacon pig, and their analyses. Biometrika 30, 16–28.

Wu, W., West, S.G., Taylor, A.B., 2009. Evaluating model fit for growth curve models: integration of fit indices from SEM and MLM frameworks. Psychological methods 14, 183.

Zeger, S.L., Harlow, S.D., 1987. Mathematical models from laws of growth to tools for biologic analysis: fifty years of growth. Growth 51, 1–21.

Growth Management and Urban Planning in the United States

Arthur C Nelson, University of Arizona, Tucson, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Growth management is a particular kind of comprehensive land use and facility planning used in parts of the United States to achieve a preferred form of the built and natural environment in a way that accommodates all projected growth. When done properly, growth management has been shown to elevate economic performance, raise the quality of life, reduce fiscal burdens, and improve environmental conditions relative to trend outcomes. This article defines growth management and its early history; outlines growth management's primary elements; differentiates growth management from related concepts; reviews growth management efforts at the state, regional, and local levels of government; and summarizes key benefits.

Growth Management Defined and Early History

Growth management involves projecting future development and land use needs and accommodating them so as to achieve goals relating to (1) preserving public goods, (2) maximizing positive land use interactions while minimizing negative ones, (3) minimizing fiscal burdens, (4) fairly distributing the benefits and burdens of growth and change, and (5) elevating the quality of life. It entails identifying appropriate locations for development to occur and to be steered away from. It also requires choreographing public facilities and their financing to achieve efficient land use patterns consistent with meeting growth management.

The Earliest Known Mention of a Growth Management-Like Effort Might Have Been Recorded in the Old Testament

> The Lord said to Moses … Command the people of Israel, that they give to the Levites … cities to dwell in; and pasture lands round about the cities … The pasture lands of the cities … shall reach from the wall of the city outward a thousand cubits all around. (Numbers 35: p. 1–4.)

This command is clearly aimed at containing urban development within a boundary so that farming can occur around it. Britain's 1947 Planning Act extended this concept by nationalizing greenbelt land and directing local government to contain urban development within stoplines.

The American experience is less draconian. The earliest growth management effort appears to have come out of Milford, Connecticut, in the early 1950s, which mapped high- and low-priority development areas. Development in low-priority areas needed to demonstrate that adequate public facilities existed to serve it without prematurely directing development away from existing urban areas. Moreover, approved projects could only be constructed in increments of 20% annually.

More famous is the plan devised by the Ramapo township, New York, in the mid-1960s. It provided for township build-out over an 18-year period. A capital improvements program helped to pace build-out during that period. Developers had to demonstrate that their project was consistent with the build-out plan (through a point system) and, if not, either wait for infrastructure to reach their site or install it themselves at their own cost.

A version of growth management, called 'growth control,' came from Petaluma, California, in the early 1970s, which capped development at 500 dwelling units annually. The cap was suballocated to different housing types to assure reasonable diversity of housing opportunities.

Generally, growth management aims to influence the location, magnitude, and timing of urban development through land use and capital facility planning, and development permitting. Unlike conventional planning whereby land is designated for particular land uses and facilities are extended when market demand dictates, growth management is a more conscious effort to shape development patterns during discrete intervals of time based on explicit urban form objectives.

Primary Elements

Growth management planning is composed of six elements: (1) establishing a vision for the desired form of land use patterns that meets growth management goals; (2) guiding growth toward achieving that vision; (3) using infrastructure to facilitate the desired development patterns; (4) preserving open spaces from development; (5) enhancing community quality of life by improving accessibility to jobs, services, shopping, and other destinations as well as providing stable housing markets and neighborhoods, among other things; and (6) coordinating actions between different government units to achieve the overall vision. Each of these elements is summarized in the following discussion.

Overall Form: Creating the Vision

A first step in the growth management planning process is projecting population, housing, employment, and nonresidential space needs in a horizon year (often 20 years out but it could be more or less), and comparing them to existing conditions to determine net new development needs. This is followed by identifying alternative development scenarios to meet those needs. Typical scenarios include development based on historic patterns, which in the United States is typically the perpetuation of low-density urban sprawl where land

uses are disconnected and the single-occupant vehicle is the principle if not the only mode option; development based on accommodating all future development on existing urbanized land, which usually means relying on infill and redevelopment to meet future land use needs; and one or two scenarios in between. Each scenario is compared along numerous dimensions determined by the scenario planning leaders but often includes housing mix and affordability, jobs-to-housing balance, infrastructure needs and their costs, amount of open space consumed, air and water impacts, and transportation modes needed to connect land uses within each scenario.

Often, the outcome of a visioning process is the conclusion that more compact, integrated development patterns served by multiple transportation modes consume less land (thereby preserving more open space), provide important economies in infrastructure delivery, create synergies between land uses, enhance residential choice, reduce vehicle miles traveled per person, reduce air pollution and water consumption, and have other benefits compared to more sprawling, disconnected land use patterns with fewer transportation mode options. To maximize total benefits, the preferred scenario usually includes urban growth boundaries, British-style development stoplines, or urban service limits. They are called collectively 'urban containment boundaries' (UCBs) and are complemented by urban service plans that control the location, size, and timing of infrastructure improvements. Land within the UCB may be subdivided into different categories for the phasing of services, depending on development suitability, proximity to existing public facilities, contiguity with existing development, and other factors. Sufficient land must be included within a UCB to accommodate market demand within the planning horizon. If too little land is included, land prices may increase beyond the ability of the market to accommodate development needs. Too much land defeats the purposes of encouraging compact, contiguous development and usually results in excessive and premature capital outlays for public services.

Growth Guidance

Growth guidance relates to the timing, phasing, and dimensioning of development within UCBs. Within UCBs, there may be intermediate containment boundaries which, as their name suggests, are short-term development boundaries within the long-term UCB. In contrast, outside UCBs, there may be urban development reserves where future urban development will be targeted once the UCB is filled in.

Growth management usually results in 'upzoning' land within UCBs and 'downzoning' land outside. The reason is simple. Land intended for urban development should accommodate more intense activity and land not intended for urban development should be prevented from urban development.

The 'upzoning/downzoning' requirement permeates through other land use policies. For example, conventional planning uses transitional zoning to buffer land use activities from each other and to facilitate a concentric ring urban form wherein uses farther away are at lower densities than those closer in. Transitional zoning is not generally compatible with growth management, however. The result of transitional zoning is that low-density and very low density land uses expand into rural areas and into resource lands, thereby perpetuating urban sprawl and undermining the long-term economic viability of resource production. Instead, growth management programs use mixed use, nontransitional zoning strategies, which, among other things (1) establish moderate- to high-density and moderate- to high-intensity land use categories throughout much of the urban area, (2) direct infill and redevelopment into urban and suburban commercial areas, (3) where they exist, steer development into transit development-oriented areas, (4) reduce the scale of low and very low urban densities within urban areas, (5) eliminate low-density and very low density development in areas that are predominantly resource lands or other environmentally sensitive lands, and (6) restrict the size and location of low-density and very low density development outside UCBs and away from resource or environmentally sensitive lands.

Growth management also aims to eliminate exclusivity. Many communities use zoning strategies to exclude certain classes of people, especially those occupying low-income housing. A typical strategy is to zone the community for large lot sizes and large minimum floor areas. Low-income households will thus be unable to afford to live in the community. Growth management recognizes the need to provide housing to all income levels. Many growth management communities have 'inclusionary zoning' policies that require large-scale projects to have housing opportunities affordable to a wide range of households.

Finally, it is one thing to designate land for moderate- to high-density development and give developers the impression that they have flexibility in meeting market demands but it is quite another to actually allow such densities. Unwary developers who propose high-density developments can get stalled in decision making, often because existing residents oppose such developments. Housing density is usually reduced but this defeats the purpose of growth management. The solution is 'minimum density zoning.' Miami-Dade County, for instance, requires communities to achieve an average of at least 75% of the maximum allowable density for residential zoning categories over near-term planning periods. If a community has a zoning category that allows up to 10 units per acre, the average of new projects built in that category can be no less than 7.5 units per acre.

Infrastructure Provision

Growth management cannot succeed without infrastructure planning and financing that makes urban development more attractive than development in the countryside. On the other hand, development must not exceed the ability of existing infrastructure to accommodate it. 'Adequate public facility' standards assure that, in addition to meeting applicable zoning and subdivision standards, facilities and services are able to serve it. 'Concurrency,' whereby facilities needed to serve new development are in place or committed concurrent with the impacts of new development, is another version of the same principle.

Under growth management, infrastructure is often a shared responsibility between government and developers. One form of share responsibility is 'development exactions' in which the developer provides a certain range of infrastructure

at her or his cost. Infrastructure within the development is normally the developer's responsibility but the responsibility for financing offsite roads, water, sewer, schools, parks, and other facilities is complex. Offsite infrastructure costs are often financed through ad hoc exactions or impact fees – or sometimes a combination. Ad hoc exactions are concessions made by the developer to local government to mitigate the impact of development. Traffic lights and water and sewer extensions are examples. Impact fee are one-time assessments on new development and to pay for new or expanded facilities needed to serve new development, help cover a share of the offsite costs. The fees are calibrated to cover the costs of one or more specific facilities apportioned to the development, such as parks or roads.

Protecting Open Spaces

What happens outside UCBs? Growth management is not effective unless it preserves open spaces and manages rural density development. Some states, such as Hawaii and Oregon, are very aggressive in preventing any development of open spaces. In these states, 'exclusive farm/forest use' regulations require owners of such land to manage it for farming and forestry if they wish to live on it.

Economic mechanisms can be used to encourage development inside UCBs and prevent development outside. 'Purchase of Development Rights' programs involve public acquisition of the development rights of open spaces. 'Transfer of Development Rights' programs involve developers acquiring development rights from open space owners in rural areas for transfer to areas within UCBs at higher density.

Not all land outside UCBs is appropriate for farming, forestry, or other open space uses. Indeed, there is some demand for rural living that if not accommodated would result in urban households acquiring farms and then maintaining only the yard they wish, typically one-quarter to one-half acre. 'Clustered' development in rural areas is one solution. For example, a 100-acre farm may be subdivided into 10 parcels of which nine are 1 acre and one is 91 acres, which in many parts of the country is sufficiently large to accommodate reasonable commercial farming operations.

Enhancing Community Life

Growth management changes existing neighborhoods. Oddly, many neighborhoods view growth management as anathema to their quality of life especially if infill and redevelopment activities locate there. 'Design review processes' help assure that new development in stable neighborhoods enhances neighborhood character. 'Supplemental housing' such as mother-in-law, overgarage, basement, or attic living units allowed in established neighborhoods improve property values, increase safety (more 'eyes-on-the-street'), make more efficient use of infrastructure, and generally assure neighborhood stability thereby increasing neighborhood vitality.

Coordination between Governments

Growth management is not effective in isolation. Because a number of governmental units interact with each other, effective growth management needs all relevant governmental units to share the same vision for a preferred built and natural environment. Intergovernmental coordination is necessary. 'Cross-acceptance is one way in which this is achieved.' Conceptually, this means that all units of government affecting a particular community – including state, regional, local, and special district governments (such as school districts, water districts, fire protection districts, and the like) – agree to each other's plans and they all agree to share the vision of the region's future built and natural environment. 'Extraterritorial jurisdiction' allows municipalities to decide on the efficacy of development proposed outside its jurisdiction within some distance, usually the equivalent of 5–10 km. 'Annexation policies' enable municipalities to assure that development outside their boundaries meet certain minimum requirements prior to annexation.

'Tax-base sharing' removes the incentive of individual local governments to offer lucrative incentives to particular kinds of development that may offer unique property tax or sales tax benefits. A shopping center locating in one community would have a portion of its property and sales tax revenue shared with nearby communities. This removes competition between communities for tax base, thereby preventing costly concessions by host communities and compensating all affected communities for potential impacts.

Growth Management Differentiated from Related Concepts

Growth management is not to be confused with growth controls (also known as 'development control' in the United Kingdom), sustainability, or smart growth, although some elements may be common across all three. Growth management focuses on accommodating anticipated growth in efficient and equitable land use arranges. Growth control, in contrast, reduces the amount of growth that is allowed. This can lead to exclusionary outcomes where only affluent households can afford to live in growth-controlled communities. On the other hand, natural constraints may require growth restrictions such as is found in the Florida Keys, Sanibel Island, Rocky Mountain resort towns, and elsewhere. Even in these situations, growth control plans often include efforts to meet development needs elsewhere proactively, such as affordable housing requirements imposed on all new development in several Rocky Mountain resort towns.

At its heart, sustainability is about making decisions today that do not foreclose options for future generations. While it can address social and economic concerns, it is mostly an environmental and ecological approach to planning. One could argue that all land use planning approaches including growth management are a subset of sustainability but there is no consensus.

In recent years, 'smart growth' has emerged as the popular characterization of what growth management tries to accomplish. There is a certain appeal to avoiding the term 'growth management' as it connotes controlling growth, which it does not, while the term 'smart growth' implies that doing otherwise would result in 'dumb' growth. Yet, smart growth has no

overall goals, while growth management does. Its goals include protecting public goods, minimizing taxpayer costs and maximizing long-term economic benefits, minimizing adverse and maximizing positive land use interactions, equitably distributing the benefits and burdens of development, and elevate the quality of life.

State, Regional, and Local Applications

Examples of growth management at different levels of government are reviewed here.

State Growth Management

Only a few American states engage in growth management on a statewide level. Five are summarized here: Hawaii, Oregon, Washington, Maryland, and Florida.

Hawaii, the nation's newest state, has the nation's oldest statewide growth management and for good reason: it is composed of islands for which there are very few alternatives to sharing development needs with agricultural and environmental needs. Through efforts by a state planning agency pursuant to the Hawaiian Land Use Law of 1961, all land on major islands is designated for urban, agriculture, low-density rural, and conservation land uses. Counties (individual islands) have very little say about those allocations.

The state with the longest history of statewide growth management among the contiguous 48 states is Oregon. The Land Conservation and Development Act of 1973 established a state commission and an administrative department. Like Hawaii, land in the state is designated essentially for urban, agriculture/forestry, low-density rural, and conservation uses. Unlike Hawaii, local governments have a significant say in how those allocations are made. Today, Oregon's land use pattern is a system of urban centers contained within UGBs surrounded by open space.

Oregon's program has four important features that are not found in other growth management states. First, it requires all local governments to project and then accommodate future development needs. Although other states also require projections, Oregon is the only one that requires local governments to actually designate land (through zoning) targeted to accommodate development needs; this reduces the opportunity for opposition in the future to thwart the kind of development that is needed to accommodate development needs.

Second, Oregon requires clear and objective review standards. Such vague review terms as 'minimizes drainage' are replaced with engineering standards that are clearly understandable by developers. Those standards must also be 'reasonable.'

Third, Oregon sets a time limit for reviewing development proposals that are 'quasijudicial' in nature (i.e., not requiring policy making such as a plan amendment). Local governments have 120 days from the time the proposal is formally and completely submitted to make a decision to approve, deny, or approve the proposal with conditions. If the 120-day limit is missed, the development can be granted automatic approval by a local judge. (Not a few local planners have lost their jobs by missing this deadline.)

Last, Oregon is unique in the United States in having a Land Use Board of Appeals (LUBA) that decides land use appeals at the state level. Appeals are to be decided within 120 days of filing. Moreover, appeals from LUBA's decisions must be decided by the Oregon Court of Appeal within 120 days of LUBA's decision. In all, under Oregon law, nonlegislative land use decisions must be decided within about a year. In contrast, court systems in other states can take years and sometimes decades to decide.

In the early 1990s, the state of Washington adopted its Growth Management Act (GMA), which has many of the same features as Oregon, including UGBs and open space preservation. Key differences are that the law applies only to counties meeting certain urban criteria and others that wish to voluntarily come under the GMA. Another difference is that it does not go as far as Oregon in requiring local governments to designate land to accommodate development needs. And, although it has an appeal process, it is through a system of Growth Management Hearing Boards at the county level.

Florida's growth management efforts, stemming from the 1970s and revamped in the mid-1980s, are relatively weak in containing urban development or preserving open space from low-density rural development. Its 'concurrency' provisions, however, which limit new development to only those areas where adequate public facilities exist or are planned, appear to have shifted development away from many open spaces to infill and redevelopment sites. But the Florida approach does not appear to have managed development overall in recent years.

A cornerstone of Florida's effort is the separation of urban and rural land uses. This is seen as a principal way in which to prevent urban sprawl. To accomplish this, local governments have used UGBs, urban service limits, phased urban extension planning, and open space preservation.

As initially conceived in the mid-1980s, a state–local partnership would use various state and local tools to match housing supply with demand to avoid overproduction that led the state into recessions historically. For a while, the GMA worked. In 1990, before the GMA took hold fully, the statewide housing vacancy rate stood at 15.3%, while the median home value to median household income ratio was 2.9. In 2000, a decade into the GMA, the vacancy rate dropped to 12.8%, while the home value to income ratio dropped to 2.7. But the 2000s saw a succession of Republican governors dismantle the state–local partnership allowing local governments to approve developments far in excess of demand. By 2010, Florida's vacancy rate shot up to 17.4%, the home value to income ratio rose to 3.0, and Florida led the nation in foreclosures. Overproduction of housing once again led Florida's economy into recession.

Maryland's growth management initiative was launched in 1997 through the Smart Growth and Neighborhood Conservation Act. Its goals include enhancing existing communities and other locally designated growth areas, protect valuable farmland and other natural landscapes, and reduce the cost of infrastructure. The state's Smart Growth Program finances infrastructure projects in designated Priority Funding Areas and provides inducements to protect landscapes outside those areas. It was the first state to explicitly use state funds to direct growth into targeted areas. Assessments indicate that since the

middle 2000s, the effort has increased the statewide share of development occurring in Priority Funding Areas and reduced the share of development outside them.

Regional Growth Management

In this context, regions can include multiple counties or large counties with multiple cities within them. Three efforts are highlighted: the Twin Cities Metropolitan Council serving the Minneapolis-St Paul metropolitan area, the New Jersey Pinelands Commission; and the Baltimore County, Maryland.

One of the earliest and best known regional growth management efforts is that of the seven-county Twin Cities (Minneapolis-St Paul) metropolitan area established by state law in 1967. This example of regional growth management was accelerated through another state law in 1976 that led to a four-step regional planning process. First, Twin Cities Metro Council staff would project future growth demands for housing, employment, and other land uses. Second, vacant land that was already 'serviced' with sewers and roads was identified along with land having environmental constraints; this led to an inventory of land available for development. Third, the Metropolitan Urban Service Area (MUSA) plan required to provide enough land within each subarea of the region to accommodate realistic growth needs over the next 10 years while minimizing the public infrastructure investments. The last step was to present alternative density scenarios for the MUSA along with estimates of public investments needed to implement each. A planning process would then adopt one scenario or a hybrid to guide development. Once the 10-year scenario is adopted and implemented, it is reviewed every 5 years to make adjustments as needed.

The New Jersey Pinelands is another example of regional growth management. The Pinelands Comprehensive Management Plan manages development to preserve the area's unique natural, ecological, agricultural, archeological, historical, scenic, cultural, and recreational resources. The effort is managed jointly by the state of New Jersey and the US Department of the Interior through the New Jersey Pinelands Commission whose jurisdiction includes parts of seven counties.

A cornerstone of the plan is the combination of urban containment and transfer of development rights. Landowners in preservation areas are granted a certain number of development 'credits' that they can sell to developers to use in 'regional growth areas.' Once sold, landowners within preservation areas are subject to conservation easements, restricting them to nonresidential uses only. Developers may use each acquired credit to build four more residential units than otherwise allowed by the local zoning code. The program has protected more than 12 000 acres of sensitive land while directing development to targeted regional growth areas.

In 1967, Baltimore County, Maryland, created its 'urban–rural demarcation line' for the purpose of preventing urban development in the county's rural areas. Tools used to protect rural land from urbanization include agricultural districts, a land trust established under the state's Rural Legacy program, development clustering, property tax incentives for conservation, and acquisition of conservation easements. The growth management plan is implemented by seven major local governments including Baltimore city. A *de facto* UGB is administered by the Metropolitan District of Baltimore County, the county's public sewerage agency, through its sewer service area boundary. This does not prevent large-lot subdivisions using onsite sewerage treatment facilities ('septic systems') from encroaching into rural areas, however.

Local Growth Management

Every state has at least one example of local government efforts. Three very different ones are reviewed here: Sioux Falls, South Dakota; Boulder, Colorado; and Lincoln–Lancaster County, Nebraska.

Sioux Falls' initial growth management policy was adopted in 1979. It was an outcome of poor planning combined with development pressures that threatened the Big Sioux River with pollution from failing septic systems. The outcome was a set of public facility policies that limited development both inside and outside the city limits to sewer systems. To manage the outward expansion of development, a plan was crafted for seven districts. Within each district, the plan notes the timing and extent of investment in future urban infrastructure. Thus, by phasing infrastructure extensions, the city manages the timing and pattern of future growth. Land outside the urban area is subject to agricultural zoning that minimizes rural residential development.

Boulder, Colorado, has one of the United States' longest histories of growth management, dating to 1958 when the city limited water extensions above 5750 ft. This was to protect the Rocky Mountain foothills viewshed for Boulder residents. In 1967, the city became the nation's first to purchase open space for viewshed protection purposes; it was financed in part through an increment to the local option sales tax. In 1976, voters approved the Growth Limitation Ordinance, which controlled growth through capital facilities planning within an UGB. To prevent development from outflanking the Boulder UGB, there is an intergovernmental agreement with Boulder County to restrict land uses outside the UGB to rural, mostly residential and open space, uses for several miles around it.

In 1994, the city of Lincoln and Lancaster County, Nebraska, adopted a comprehensive plan restricting the extension of water and sewer services outside the city limits. The plan includes an assessment of existing and future infrastructure needs over a 20-year planning period. The city then manages the location and timing of future growth through its infrastructure extension policies. Once served, new development is annexed in to the city. Moreover, the state of Nebraska provides Lincoln with extraterritorial zoning authority over land within 3 miles of its boundaries. This allows the city to prevent premature annexation of urban-scale development that occurs outside its boundaries. The intergovernmental agreements combined with state-granted powers provide Lincoln with the ability to manage growth through the planned extension of its infrastructure and subsequent annexation.

Benefits

For a long time, the benefits of growth management in the United States seemed illusory. Evidence is emerging that when

growth management plans include policies composed of the overall growth management goals noted above, important benefits result. They include more housing choices and higher supplies of affordable housing, more transportation options, more open space preservation combined with larger supplies of 'working' landscapes (such as farms and forests), more jobs resulting in lower unemployment and higher wages, more economic opportunities for minorities, lower public facility costs, and greater fiscal capacity principally because economics are more efficient and resilient.

See also: Suburbanization and Suburbanism; Urban Internal Spatial Structure; Urban Sprawl.

Bibliography

DeGrove, John M., 2005. Planning Policy and Politics: Smart Growth and the States. Lincoln Institute of Land Policy, Cambridge, MA.

Freilich, Robert H., Sitkowski Robert J., Mennillo Seth D., 2010. American Bar Association, Chicago, IL.

Ingram, Gregory K., Carbonell, Armando, Hong, YuHung, Flint, Anthony, 2009. Smart Growth Policies: An Evaluation of Programs and Outcomes. Lincoln Institute of Land Policy, Cambridge, MA.

Johnson, William C., 1998. Growth Management in the Twin Cities Region: The Politics and Performance of the Metropolitan Council. Center for Urban and Regional Affair, University of Minnesota, Minneapolis, MN.

Juergensmeyer, Julian Conrad, Roberts, Thomas, 2013. Urban Planning and Land Development Control Law, third ed. West Publishing Co, St. Paul, MN.

Knaap, Gerrit J., Nelson, Arthur C., 1992. The Regulated Landscape. Lincoln Institute of Land Policy, Cambridge, MA.

Nelson, Arthur C., Duncan, James B., 1995. Growth Management Principles and Practices. American Planning Association, Chicago, IL.

Nelson, Arthur C., 2004. Planner's Estimating Guide: Projecting Land Use and Facility Needs. American Planning Association, Chicago, IL.

Nelson, Arthur C., Dawkins, Casey J., 2004. Urban Containment in the United States: History, Models, and Techniques for Regional and Metropolitan Growth Management. American Planning Association, Chicago, IL.

Nelson, Arthur C., Dawkins, Casey J., Sanchez, Thomas W., 2007. The Social Impacts of Urban Containment. Ashgate, London, UK.

Porter, Douglas R., 2000. The Practice of Sustainable Development. Urban Land Institute, Washington, DC.

Porter, Douglas R., 2007. Managing Growth in America's Communities, second ed. Island Press, Washington, DC.

Randolph, John, 2004. Environmental Land Use Planning and Management. Island Press, Washington, DC.

Guilt

Amy Farrell, School of Criminology and Criminal Justice, Northeastern University, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by N. Harris, volume 9, pp. 6445–6448, © 2001, Elsevier Ltd.

Abstract

In US criminal law, a person is guilty of a crime when a court of law determines beyond a reasonable doubt that they have violated a criminal law or performed acts that meet the elements of a criminal offense that is established through statute. The determination of guilt through criminal law reflects the expectation of society that individuals should be held accountable for their judgments and actions resulting from their exercise of free will. This article defines the concept of guilt in US criminal law and explores how criminal cases arrive at a guilty verdict. Additionally, the article examines errors of guilt and the meaning of an acquittal from guilt.

Definition of Guilt

A crime is an intentional act that is committed without excuse or justification. The act has been determined by the state to cause harm that results in punishment through criminal prosecution. A person is determined to be guilty of a crime when a court of law finds beyond a reasonable doubt that person has violated a criminal law or performed acts that meet the elements of a criminal offense that is established through statute.

The determination of guilt through criminal law reflects the expectation of society that individuals have free will and should be held accountable for their judgments and actions resulting from the exercise of that will. Law in the United States was heavily influenced by English common law (judge-made law). One of the most important principles of common law is the notion that crimes require both *actus reas* or a 'voluntary and guilty act' and *mens rea* or a 'guilty mind' to occur simultaneously. As a result, the state must prove that a person voluntarily engaged in behavior that constitutes a criminal act and had the intent to violate the law. There is no crime without a bad act, although failure to act or do something required by law may also constitute a crime. For example, parents have a legal duty to protect their children from harm and failure to do so may constitute a crime of neglect.

To be found guilty of a crime, the state must prove beyond a reasonable doubt that a person who committed a criminal act intended to commit a crime. Intent can be general or specific and the legal standards to prove intent differ from crime to crime. For example, to be guilty of a crime of murder, a person must demonstrate a state of mind to do harm (malice) before the commission of a homicide. Homicide without 'malice aforethought' is manslaughter not murder. There are some situations where people who commit criminal acts cannot form the intent necessary to be found guilty of a crime. In William Blackstone's *Commentaries on the Laws of England* (1790), he suggests children under the age of seven are 'conclusively presumed' not to be able to form criminal intent. Between the ages of 7 and 14, children are presumed not to be able to form the intent to commit a crime but that presumption can be refuted through the presentation of evidence. Common law has historically presumed that children who were 14 years and older were responsible for criminal acts. In the United States, state laws have increased or decreased the age of criminal responsibility overtime. All states now have juvenile courts where criminal conduct of minors is addressed through juvenile rather than criminal proceedings. Juveniles are adjudicated as delinquent as opposed to being adjudicated as guilty. Blackstone (1790) also argued that the insane cannot form the intent necessary to commit a criminal act.

Criminal acts can also be justified. In certain circumstances, the law also excuses people from crimes committed in defense of themselves, other people, or property. A defendant who claims self-defense must prove that he or she did not provoke the threatening events, faced an immediate threat to which there was no other way to escape harm and genuinely believed that there was no alternative to the use of force. Since 2005, 18 states and the federal government have adopted 'stand your ground' provision within their criminal law, allowing a person to defend themselves, others, or their property without the requirement to retreat to safety (McClellan and Tekin, 2012) (these states include Alabama, Arizona, Florida, Georgia, Indiana, Kansas, Kentucky, Louisiana, Michigan, Mississippi, Montana, New Hampshire, Oklahoma, South Carolina, South Dakota, Tennessee, Texas, and West Virginia). Those acting under duress or being compelled by another to engage in criminal acts that he or she would otherwise not commit may also be excused from guilt.

If a person is charged with a violation of a criminal offense, either through indictment by grand jury or through a prosecutor filing an information, their case can be disposed of in three main ways. First, person can plead guilty to the offenses charged or bargain with the prosecutor to plead guilty to a lesser charge and have the disposition of their sentence be determined by a judge. In some US jurisdictions, a person accused of a crime has the option of pleading *nolo contendere*, which essentially means the defendant does not affirm or deny the charges. Courts must consent to accept a *nolo contendere* plea. The effect of the plea is similar to a guilty plea in that once the plea is rendered a judge determines the appropriate sentence. Second, a person can plead not guilty to the formal charge and elect to have a judge or a jury of their peers hear evidence and determine their guilt through a criminal trial, the result of which could be a finding of guilt, an acquittal, or a hung jury. Third,

a prosecutor can elect to dismiss charges against a person before a plea of guilt or a court of law determines guilt through trial.

To be found guilty of a crime by jury trial or trial before a judge without a jury, the prosecution must present evidence proving beyond a reasonable doubt that the defendant committed the accused criminal act. The prosecution must establish that all elements of the crime that are required by law occurred in the alleged offense and present evidence of the defendant's intentional misconduct and blameworthiness.

Patterns of Guilt

The trial by jury or judge in states where a defendant has the right to waive a jury trial is the judicial process that determines a person's guilt or innocence. Until the early twenty-first century, most criminal cases in the US were disposed of through trial. Today, however, the typical outcome of a criminal case is a conviction based upon the defendant's entry of a guilty plea.

Prior to the early part of the twentieth century, criminal trials were conducted swiftly and most defendants elected to have their cases heard before a judge or jury (Alschuler, 1979; Friedman, 1979). Courts often resolved multiple criminal trails in a single day because cases were commonly not complex and there were fewer procedural rules or rules of evidence. Defendants often did not have lawyers acting upon their defense. The professionalization of the criminal justice system in the form of state prosecutors and professional police meant screening of criminal offenders shifted from primarily the responsibility of judges and juries to police and prosecutors. The rising authority and autonomy of police and prosecutors corresponded with widespread corruption among criminal justice officials in the early twenty-first century. Anecdotal evidence from this time also suggests some prosecutors were willing to make concessions to criminal defendants to reduce charges in exchange for fees (Alschuler, 1979). By the 1920s, criminal caseloads in major US cities had increased significantly and a majority of cases were resolved through plea bargains since cases could be processed relatively quickly and efficiently (Alschuler, 1968). In 1971, the US Supreme Court upheld the constitutionality of plea bargaining in Santobello v. New York (404 U.S. 257).

State Courts

Of criminal cases disposed in the largest urban counties in the United States in 2006, the year of most recent available statistics (see Table 1), 68% of the individuals charged with felony offenses were actually convicted of felony or misdemeanor crimes. Of those who were convicted, the overwhelming majority (approximately 96%) pled guilty. Prosecutors decided to dismiss cases before adjudication in another 23% of all cases. Eight percent of all cases were either disposed through alternatives to traditional criminal punishment or resolved in the following year. Acquittals accounted for only 1% of all case outcomes. This pattern has been very consistent over time.

According to the Bureau of Justice Statistics data, only 3% of all criminal cases were actually resolved through trials in which judges or juries rendered verdicts of guilt or acquittal. While acquittals represent only a tiny fraction of all criminal dispositions, they make up a much larger proportion of those cases that do go to trial. Approximately one third of all criminal cases resolved through trial result in the acquittal of the defendant.

Federal Courts

The dominance of plea bargaining as the primary method of disposing of criminal cases is even more evident in the federal criminal justice system. Of the 98 489 defendants in cases that were closed by federal prosecutors in 2010, an astonishing 92% were convicted (Bureau of Justice Statistics, 2013). Just under 8% of defendants had their cases dismissed by prosecutors or were given pretrial diversion. Less than 1% of all federal defendants were acquitted by either federal juries or federal judges. In the federal criminal justice system, the guilt is determined predominately by the prosecutor. Highly structured federal sentencing laws have limited judicial discretion and incentivized defendants accepting responsibility through a guilty plea and cooperating with the government as a mechanism to reduce tough sentences (Etienne, 2003). Of the federal defendants whose cases were prosecuted to conclusion (i.e., they were not dismissed or resolved through pretrial diversion), 97% plead guilty as the result of a plea bargain. The proportion of cases resulting in a plea bargain has risen substantially since 1990 when only 80% of federal cases prosecuted to conclusion resulted in a guilty plea (Brown et al., 1984).

Table 1 Adjudication outcomes for felony defendants in large urban counties (2006)

			Convicted of felonies			*Convicted of misdemeanors*			*Not convicted*		
Most serious arrest charge	*No. of defendants*	*% Convicted*	*Total*	*% Plea*	*% Trial*	*Total*	*% Plea*	*% Trial*	*% Dismissed*	*% Acquitted*	*% Other outcome*[a]
All offenses	51 922	68	57	55	2	11	10	1	23	1	8
Violent offenses	11 303	61	50	45	4	11	10	1	33	1	5
Property offenses	15 351	70	57	56	1	13	12	1	22	1	8
Drug offenses	19 295	68	59	56	1	10	9	1	20	–	12
Public-order offenses	5973	72	60	58	1	12	11	1	21	1	6

– denotes less than 0.05%.
[a]Includes diversion and deferred adjudication. Murder defendants were followed for an additional year.
Note: Eleven percent of all cases were still pending adjudication at the end of the 1-year study period and are excluded from the table. Murder defendants were followed for an additional year. Data on adjudication outcomes were available for over 99.9% of those cases that had been adjudicated. Detail may not sum to total because of rounding.
Source: Cohen, T., Kyckelhahn, T., 2010. Felony Defendants in Large Urban Counties, 2006. Statistical Table 11. Bureau of Justice Statistics, Washington, DC, p. 11.

Mistakes of Guilt

The US criminal justice system has a two-stage process for determining who is guilty of a crime following arrest. In the first stage, prosecutors screen criminal cases to eliminate those cases where there is insufficient evidence to prove a person's guilt. Prosecutors must either present information to a grand jury that there is a probable cause to believe a defendant committed a crime sufficient to warrant a trial or file an information, a document presenting evidence of a defendant's commission of a crime. Even after charges are filed, prosecutors regularly assess the strength of their case against a defendant. In advance of trial, they may discover that key witnesses will not agree to testify or evidence-supporting prosecution is weaker than they originally believed. Prosecutors may decide to dismiss a case because they do not think they can convince a jury of the defendant's guilt and the defendant cannot be persuaded to plead guilty. Prosecutors have significant and unreviewable discretion to pursue or dismiss criminal charges and are commonly considered the "gateway to the criminal justice system" (Kerstetter, 1990: p. 268). In the second stage, courts adjudicate guilt in the hope that judge or jury fact finders will prevent falsely accused defendants from being unjustly convicted of a crime. The US criminal justice system is designed to err on the side of acquitting the guilty. At trial, juries are informed that defendants are to be presumed innocent and it is the burden of the government to prove defendant guilt beyond a reasonable doubt.

Any fear about innocent defendants being erroneously convicted is alleviated by a widespread belief that innocent people have been effectively screened out before criminal charges are filed (Givelber, 2005). As Edwin Meese III, United States Attorney General under President Ronald Reagan, once explained, "… you don't have many suspects who are innocent of a crime. That's contradictory. If a person is innocent of crime, then he is not a suspect" (Justice Under Reagan, 1985). As a result, errors in the adjudicatory process should only result in the acquittal of the guilty. But mistakes do happen.

While the true extent of wrongful convictions is unknown, since DNA evidence was first used to help exonerate defendants from erroneous convictions in 1989 (DNA evidence was first used in 1989 to exonerate Gary Dotson who was wrongfully convicted of rape and incarcerated in Illinois for 10 years), the courts have increasingly used new forensic methods to establish innocence. Analyzing hundreds of exonerations occurring since the late 1980s, Gross et al. (2005) estimated the number of mistaken convictions to "be in the thousands, perhaps the tens of thousands" (551). To date, just over 300 people have been exonerated postconviction through forensic evidence that established their innocence (Innocence Project, 2013). Landmark research conducted by Bedau and Radelet (2004) analyzing hundreds of cases of wrongful convictions occurring in capital cases suggests that mistakes of guilt happen even in the most serious cases.

Wrongful convictions occur for a number of different reasons. Eyewitnesses misidentify suspects (Garrett, 2011), informants provide false statements to prosecutors and perjure themselves during testimony, and prosecutors make errors (Gross and Shaffer, 2012). Defendants may also be persuaded to falsely confess to crimes they did not commit, usually under psychological coercion by the police (Ofshe and Leo, 1997). Defendants provided false confessions in nearly two-thirds of the homicide cases exonerated by DNA evidence (White, 2003). Often defendants confess to crimes they did not commit and agree to accept a plea bargain because they do not understand the evidence against them. When a defendant agrees to accept a plea bargain he or she gives up the right to see evidence from the prosecution that may support his or her claim of innocence. As a result, people may plead guilty to a crime even when the state has evidence that indicates their innocence.

It is nearly impossible for a defendant to establish their innocence once they have been convicted of a crime. On appeal, the defendant must persuade the court that no reasonable judge or jurors could have concluded that they were guilty of the accused crime (House v. Bell, 2006).

The Meaning of an Acquittal

A person who is legally acquitted of a crime has claimed innocence of the criminal charge and had their case heard at trial. The finding of 'not guilty' by a judge or jury reflects that the state has failed to provide evidence sufficient to establish the defendant's guilt beyond a reasonable doubt. It is possible that the state was correct that the defendant committed the crime but simply failed to persuade a judge or a jury that this was the case. The criminal justice system is replete with examples of guilty people who were erroneously acquitted. In these cases, the prosecutors or the police produce evidence such as confessions or statements by codefendants that could not be presented to the jury due to due process considerations. To ensure adherence to the mandates of the Fourth Amendment to the United States Constitution prohibits unlawful searches and seizures, the Supreme Court has determined that evidence seized in violation of these provisions cannot be presented to judges or juries to determine a defendants' guilt (Weeks v. U.S., 1914; Mapp v. Ohio, 1961). Interviews with jurors following a not guilty verdict also demonstrate mistakes jurors made in applying the law. Double jeopardy prevents the state from appealing acquittals, so there are no appellate court rulings clarifying when an acquittal is appropriate. The legal inquiry ends when the jury or judge returns a not guilty verdict.

Actual innocence is nearly impossible to prove. While guilt rests upon a conclusion about the defendant's behavior, an acquittal reflects only an absence of proof or a finding that guilt should be set aside due to legal error. Judges or juries do not address the question of whether or not a defendant is truly innocent. Once they decide that the prosecutor has not made a persuasive case that the defendant is guilty their work is done. Actual innocence is a question that judges and juries are neither asked nor expected to answer. Claims of actual innocence require judgments beyond those that are made by criminal courts.

Absent definitive answers about whether or not people who are acquitted of crimes are actually guilty, empirical research has examined the degree to which acquittals correspond to factors that are consistent with actual innocence. Kalven and Zeisel's (1966) seminal research in *The American Jury* relied on surveys from hundreds of judges to determine the accuracy

of jury verdicts. They found that judges agreed with the jury verdicts in the majority of cases. When judges and juries disagreed, juries were more likely to acquit when judges would have convicted (19% of all cases) than they were to convict when the judges would have acquitted (3% of all cases). Kalven and Zeisel (1966) sought to explain why juries acquitted when judges would have convicted suggesting that in cases that were close on the evidence juries were 'liberated' to consider non-evidentiary factors such as personal and community values in deciding to acquit a defendant whom the judge would have convicted. For the past 50 years, the 'liberation hypothesis' has the dominant theory to explain acquittals. New research looking at data on jury decision making in four US cities, challenges this perspective. Using survey data from both judges and juries, Givelber and Farrell (2012) find that jury acquittals actually reflect the evidence presented at trial. Juries acquit in the face of unpersuasive presentations of evidence by the prosecution, more persuasive presentations of evidence by the defense, or both. Despite the best efforts of prosecutors to screen for guilt, the jury does its job when the innocent are prosecuted and acquitted and fails in its task when the innocent are prosecuted and erroneously convicted.

See also: Criminal Justice, Sociology of; Misdemeanor offenses; Property Crime: Victimization and Trends; Punishment: Comparative Politics of; Punishment; Wrongful Convictions.

Bibliography

Alschuler, A., 1968. The prosecutor's role in plea bargaining. The University of Chicago Law Review 36, 50–112.

Alschuler, A., 1979. Plea bargaining and its history. Law and Society Review 12, 211–245.

Bedau, H.A., Radelet, M.L., 2004. Convicting the innocent in capital cases: criteria, evidence, and inference. Drake Law Review 52, 587–603.

Blackstone, W., 1790. Commentaries on the Laws of England. Clarendon Press, Oxford, UK.

Brown, E.J., Flanagan, T.J., McLeod, M. (Eds.), 1984. Sourcebook of Criminal Justice Statistics – 1983. U.S. Department of Justice, Bureau of Justice Statistics. U.S. Government Printing Office, Washington, DC.

Bureau of Justice Statistics Federal Justice Statistics Program website. http://bjs.ojp.usdoj.gov/fjsrc/ (accessed 10.04.13.).

Cohen, T., Kyckelhahn, T., 2010. Felony Defendants in Large Urban Counties, 2006. Statistical Table 11. Bureau of Justice Statistics, Washington, DC, p. 11.

Etienne, M., 2003. Acceptance of responsibility and plea bargaining under the feeney amendment. Federal Sentencing Reporter 16 (2), 109–113.

Friedman, L.M., 1979. Plea bargaining in historical perspective. Law and Society Review 13, 247–259.

Garrett, B.L., 2011. Convicting the Innocent: Where Criminal Prosecutions Go Wrong. Harvard University Press, Cambridge, MA.

Givelber, D., Farrell, A., 2012. Not Guilty: Are the Acquitted Innocent? New York University Press, New York.

Givelber, D., 2005. Lost innocence: speculation and data about the acquitted. American Criminal Law Review 42, 1167–1199.

Gross, S.R., Jacoby, K., Matheson, D.J., Montgomery, N., Patil, S., 2005. Exonerations in the United States 1989 through 2003. Journal of Criminal Law and Criminology 95, 523–560.

Gross, S.R., Shaffer, M., 2012. Exonerations in the United States, 1989-2012. National Registry of Exonerations. http://www.law.umich.edu/special/exoneration/Documents/exonerations_us_1989_2012_full_report.pdf (accessed 10.04.13.).

House v. Bell, 547 U.S. 518, 2006.

Innocence Project, 2013. http://www.innocenceproject.org/know/.

Justice Under Reagan: Reagan seeks judges with 'traditional approach', 1985. U.S. News and World Report, 99, October 14, p. 67.

Kalven, H., Zeisel, H., 1966. The American Jury. Little Brown, Boston, MA.

Kerstetter, W., 1990. Gateway to justice: police and prosecutorial response to sexual assaults against women. The Journal of Criminal Law and Criminology 81, 267–313.

Mapp v. Ohio, 367 U.S. 643, 1961

McClellan, C.B., Tekin, E., 2012. Stand Your Ground Laws, Homicides and Injuries. National Bureau of Economic Research. NBER Working Paper Series. http://www.nber.org/papers/w18187 (accessed 10.04.13.).

Ofshe, R.J., Leo, R.A., 1997. The social psychology of police interrogation: the theory and classification of true and false confessions. Studies in Law, Politics, & Society 16, 189–251.

Santobello v. New York, 404 U.S. 257, 1971.

Weeks v. U.S. 232 U.S. 383, 1914.

White, W.S., 2003. Miranda's Waning Protections: Police Interrogation Practices after Dickerson. University of Michigan Press, Ann Arbor, MI.

Guns and Criminals

James D Wright, University of Central Florida, Orlando, FL, USA
Rachel L Rayburn, Indiana University Purdue University – Fort Wayne, Fort Wayne, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The first systematic effort to understand where, how, and why criminals acquire, carry, and use guns was by Wright and Rossi (1986). Key findings were that criminals rarely acquired firearms through conventional retail markets, that much criminal gun possession was self-defensive and not specifically for the purpose of committing crimes, and that criminals were for various reasons adept at finding ways to circumvent stricter gun controls. This article reviews subsequent research on these and a range of related topics.

In 1986, Wright and Rossi published *Armed and Considered Dangerous: A Survey of Felons and Their Firearms* (ACD), the most comprehensive survey and analysis then available (or since) of where, how, and why criminals acquire, carry, and use guns. A second edition was published in 1994 and a third in 2008; in both cases, there were new prefaces pointing out that the main themes and findings had held up remarkably well in the face of new research and evidence. Despite its advanced age, "no research has satisfactorily disentangled" the effects of gun acquisition any better than the 1986 study (Hauser and Kleck, 2013). But what of even newer research on gun-related topics? The purpose of the present article is to review the literature pertaining to the criminal acquisition and use of firearms produced in the years since ACD was initially published, with an emphasis on the post-2008 literature.

The Criminal Gun Market

Major replications of ACD's findings concerning firearms acquisitions by imprisoned offenders were carried out in the Surveys of Inmates in State and Federal Correctional Facilities (SISCF and SIFCF) in 2004. The 2004 SISCF was conducted for The Bureau of Justice Statistics (BJS) by the US Census Bureau, which also conducted the SIFCF for the Federal Bureau of Prisons (BJS, 2004). Both surveys include information about current offense and criminal history; family background; personal characteristics; prior drug and alcohol treatment; gun possession; and prison treatment, programs, and services. Unlike the ACD survey, which sampled only male inmates in state prisons, the BJS studies added federal prisoners to the sampling frame and sampled both male and female inmates. The sample sizes were also an order of magnitude larger (about 2000 in ACD vs about 18 000 in the BJS surveys). A final important difference is that the ACD questionnaire was largely self-administered, whereas the BJS surveys utilized computer-assisted personal interviews.

Methodological differences aside, the findings from the 2004 BJS surveys are very similar to those reported in ACD. The BJS surveys found that approximately 16% of state prison inmates and 18% of federal prison inmates who used, carried, or possessed a firearm when they committed the crime they were currently serving a sentence. During the offense that landed them in prison, only 2–3% carried a military-style weapon. The criminal preference for conventional handguns, which compromised 13–16% of the firearm possessions, over the so-called 'assault weapons' was again confirmed.

Perhaps the most commonly cited finding from ACD was that criminals rarely obtained guns through conventional retail channels. (Thus, by implication, controls imposed at the point of retail sale largely miss the majority of criminal firearms transactions.) In the 2004 survey of state inmates, among those who possessed a gun at the time of their arrest, "fewer than 2% bought their firearm at a flea market or gun show, about 10% from a retail store or pawnshop, 37% obtained it from family or friends, and another 40% obtained it from an illegal source" (BJS, 2004: 13). This closely resembles the findings in ACD (p. 185) that "one out of six of the men in this sample acquired his most recent handgun through means and sources likely to be concerned about the legality of the transaction; five out of six did not." ACD showed and subsequent studies have confirmed that "the illicit firearms market … [is]…clearly and heavily dominated by informal, off-the-record transactions, either with family and friends or with various gray- and black-market sources" (p. 190).

The BJS findings are of particular interest because in the decades since ACD was published, much legislation and commentary has focused on 'military-style assault weapons' and on the illegal acquisition of firearms at gun shows and flea markets. Recently, Dan Gross (President of the Brady Campaign to Prevent Gun Violence) stated, "most Americans would be outraged to learn the number of loopholes that allow felons and other dangerous people to get their hands on machine guns and other military-grade firearms" (2013). This campaign, as well as both houses of Congress, has lobbied and passed legislation attempting to control firearm sales at gun shows. One legislator claimed, "70 percent of guns used in crimes come from gun shows" (Kopel, 2000). However, no study, and certainly none with any sort of methodological credibility, has ever found that more than 2–3% of criminals' guns are obtained at these venues (Wright and Rayburn, 2012).

One active line of research whose origins can be traced directly to ACD is patterns of acquisition of firearms used in well-publicized school shootings – Sandy Hook, Newtown, Northern Illinois University, Virginia Tech, and many others. After each of these incidents, parents, teachers, and the media

always exclaim with evident exasperation, "Where did they get their guns?" The answer is that they are obtained largely through the same methods and outlets as adult offenders, i.e., via informal purchases, swaps, and trades in secondary gun market, or often enough, directly from family members. For a review of legislative responses to school shootings, see Schildkraut and Hernandez (2014).

Crime, Crime Rates, and Guns

Rates of crime in general and violent crime in particular trended upward from the 1960s to about 1994, when a downward slide occurred that has generally continued since. No one has produced the conclusive account of why violent crime rates dropped, but the decline was steep and nearly universal across the United States (Rosenfeld, 2005). Some suggest that the generally good economy of the mid- to late 1990s contributed (Hannon and DeFina, 2010), although the relationship between crime and the economy is not straightforward. Others have suggested an aging population (White, 2014) and enhanced law enforcement practices (Rosenfeld et al., 2014), yet the most salient factor seems to be a reduced demand for crack cocaine (Blumstein, 2006).

The reasons for the decline in crack usage are also not well understood. It seems that as the original crack generation grew older, stopped using, or died off, their generational 'replacements' preferred marijuana and alcohol. "Following a lag of a year or two, rates of gun violence also began to fall off" (Rosenfeld, 2004: 87). When considering 5- and 10-year trends in violent crime, that decline has continued. The 2012 violent crime total was roughly 13% below the 2008 level. Further, the 2008 level was approximately 12% that of 2003. Over recent years, it seems that the availability and use of hard-core drugs has been more consequential for crime rates than anything having to do with the availability of guns, which probably did not slacken significantly even as crime rates were plummeting.

That said, guns are still involved in an enormous number of crimes and acts of violence in this country. Last year, the BJS compiled a report on firearm violence for the years 1993–2011 (BJS, 2013). In that span, the nation suffered approximately 242 000 firearm-related fatalities and a much larger rate of nonfatal firearm victimizations. In 2011, 478 400 fatal and nonfatal violent crimes were committed with a firearm. Homicides made up approximately 2% of all firearm-related crimes, down by 39% from a high of 18 253 in 1993. Deaths by suicide with a firearm continue to outnumber homicidal deaths, with fatal firearms accidents a distant third. Of the nonfatal injuries, 31% occurred during the course of an aggravated assault and 26% occurred during the course of a robbery.

Firearm usage has been found to be exceptionally high in crime-precipitated homicides, defined as cases where the victim was killed while engaging in illegal behavior such as predatory crimes, vice crimes, or drug dealing (Copes et al., 2002). In an analysis of homicides in Chicago, Copes and colleagues found that about 64% of all crime-precipitated homicides involved a firearm. Of these, 77% of the homicides were committed in connection with predatory crime involving firearms, compared to 49% of the homicides connected to vice crimes and 89% of murders connected to drug transactions.

ACD described predatory gun criminals as relatively young nonwhite men from troubled backgrounds, usually living in violent neighborhoods. This description has been confirmed in subsequent studies. Stell (2004) found most perpetrators of homicide and other gun-related crimes to be male and younger than 30 years. Furthermore, 70–80% of firearm offenders have criminal records with an average of four felony arrests prior to the commission of a homicide.

Regarding homicide victimization, researchers have found that criminal offending increases the chances of becoming a homicide victim (Dobrin and Bursk, 2003; Dobrin et al., 2005). Faria (2005) analyzed FBI data and found that approximately 75% of violent crimes in any locality are committed by a very small minority of criminals (6%) who are the most active repeat offenders (less than 0.2% of crimes committed with firearms are committed by citizens licensed to carry a concealed weapon). Criminal activity involving firearms is disproportionately committed by and against a small subset of the population (Cook et al., 2005).

Firearms Behaviors and Criminological Theory

The number of criminals who are also victims of crime, including homicides, drives home the important theoretical point that, as ACD puts it, "these men inhabit a violent and generally hostile world…these men came to possess and carry [guns] as much to aid in surviving life on the streets as to use in committing crime" (p. 139). ACD was itself largely descriptive, but since initial publication, researchers in the field have made some theoretical developments on the subject of criminal activity and firearms.

Routine activities theory was originally developed by Cohen and Felson (1979) and argues that the opportunity to commit crime happens with the combination of a motivated offender, a suitable target, and the absence of capable guardians. Lifestyles and routine activities help control the risk of victimization. In other words, how people spend their time, when they work or leave home, how they dress, what defensive measures they take, who they spend time with, the presence or absence of guardians, etc., all impact upon their suitability as a target.

ACD drew attention to guns and gun carrying as part and parcel of the routine activities of the criminal class. Several factors affect the decision for the 'bad guys' to carry firearms routinely, including a fear or history of victimization, involvement with delinquent peer groups, exposure to violence, and a deviant or criminal lifestyle. The key 'decision' of criminals was not to acquire one or even to use a gun in a particular offense, but rather the decision to carry a gun on a routine basis. This was largely in response to a perceived dangerous and hostile environment, where a gun becomes useful in a threatening world. Once a firearm is being carried routinely, it is then available for use in crime.

Recently, researchers have been examining routine activities theory by showing how gang membership and carrying a gun can be categorized as both risk and protective factors in high-poverty areas (Spano et al., 2008). More specifically, researchers found that carrying a gun was a significant risk

factor for violent victimization. However, these effects were interceded by some lifestyles such as family relationships and employment status.

These findings have been cited as one justification for law enforcement emphasis on the detection and intervention of illegal gun carrying in crime 'hot spots,' for example, the Kansas City Gun Experiment in the early 1990s, Project Safe Neighborhoods implemented in Chicago and a number of other US cities, Project Exile in Richmond, and a number of others. A recent case study of one such program, Youth ALIVE!, found reductions in gun violence in California (Calhoun, 2013). This not-for-profit organization trained youth in California by training them to become peer educators to reduce the supply and demand for guns. These gun buyback programs, now fairly common in the United States (Goering, 2010), may take years to affect crime rates and are a topic in need of more research (Phillips et al., 2013).

Studies such as these elucidate how routine activities theory suggests a degree of rationality in criminal behavior (McCarthy, 2002). From the offender's perspective, it indicates cost–benefit calculations are being made, and that public policy solutions can thus contribute to a reduction of gun homicides. When individuals see people around them carrying guns, the act becomes a very rational act. As one respondent in ACD put it, "You can't never afford to be the only guy on the street who ain't got him a piece."

In asking ACD respondents why they carried a gun, the single most common answer among was, "if you carry a gun, your victim doesn't put up a fight and that way you don't have to hurt them" (mentioned as 'very important' by 57% of the sample). The 'bad guys' carry guns as a means of circumventing, not perpetrating, violence. Other life situations that might rationalize the routine carrying of guns would be harassment by others with guns, engaging armed victims, or having an encounter with police.

Firearms Laws and the Reduction of Gun Violence

A commonplace argument in debates about gun control is that compared to other countries, the United States has fairly relaxed firearms laws, lots and lots of guns, and more gun violence than most other countries. All of this is true. What is less obvious is the conclusion that lenient laws and ease of obtaining a firearm are the reasons why we suffer so much gun violence. In other words, would increasing laws and restrictions on firearm ownership reduce gun violence? One set of researchers (Hahn et al., 2005) attempted to answer this question by examining dozens of attempts to reduce gun violence. In their study, they considered the following firearm legislation and violence prevention attempts: firearm registration, bans on specific guns, or ammunition; 'shall issue' statutes making it easier to carry concealed weapons; restrictions on acquiring guns; acquisition waiting periods; gun licensing, laws that require childproof storage of guns; and zero-tolerance approaches to firearms in schools. Across all available studies and multiple indicators of violence, it was concluded that, "the evidence is insufficient to determine whether the degree or intensity of firearms regulation is associated with increased (or decreased) violence" (2005: 59).

In the sequel to ACD focused on juvenile offenders (Sheley and Wright, 1995), it was pointed out that since most of the methods through which criminals acquire guns are already against the law, one wonders whether new, more strict, or aggressive enforcement of existing laws will make much of a difference. The failure of existing studies to detect the presumed violence-reductive effects of more and stricter gun laws also implies that even if such effects exist, they cannot be very strong (Legault and Hendrickson, 2009). As intimated time and again throughout ACD, those who seek to reduce the level of violence in the United States by putting further and more burdensome restrictions on the rights of the general public to own firearms are probably barking up the wrong tree.

Perhaps the most well-known firearms law to have been passed since the 1980s was the Brady Act, which mandated a federal 5-day waiting period for gun acquisition and a mandatory criminal background check. The law has since been replaced with the FBI's National Instant Criminal Background Check System, itself controversial (Price and Norris, 2010). Numerous evaluations of Brady and similar legislative initiatives showed they have little to no effect in reducing firearms violence (Moorhouse and Wanner, 2006; McGarrell, 2005; Stell, 2004; Levitt, 2004; Lott and Whitley, 2001; Brezina and Wright, 2000; Ludwig and Cook, 2000). The large majority of illegal firearms transactions do not involve conventional over-the-counter retail transactions; therefore, restrictions imposed at the point of retail sale do not have much of an effect.

More generally, efforts to control violent crime by controlling the supply of firearms have been generally unsuccessful, at least partly because the vast number of guns already in circulation (several hundreds of millions) implies that actual reductions in the available supply are almost impossible to achieve (Brezina and Wright, 2000). Guns already in circulation represent a virtually inexhaustible source of supply to the informal market in firearms. And, of course, there are always corrupt 'legitimate' firearm dealers engaged in trafficking to supplement the supply of pre-owned guns.

Studies in the last 5–10 years have shown mixed results for the effects of different gun laws. A study using state data from 2007 to 2010 found that states with more gun laws had fewer firearm homicides and firearm suicides (Fleeger et al., 2013). By using five categories of firearm legislation across states, a 'legislative strength score' was created for each state. The average firearm fatality rates among states varied from a low of 2.9 fatalities (Hawaii) to a high of 17.9 (Louisiana). Firearm legislation scores ranged from a low of 0 (Utah) to a high of 24 (Massachusetts). A major limitation of the study is that the researchers could not determine causation, calling for further research on the topic.

Another study examining all 50 states (2004) gathered data on concealed carry and 'shall issue' laws (Hepburn et al., 2004). Researchers found no significant association between changes in concealed carry laws and homicide rates. Consistent with previous studies, a citizen legitimately carrying a concealed weapon does not seem to have any impact on increases or decreases in violent crime rates. This study, however, only examined data from 1979 to 1998. An array of firearm legislation since this time could show different results.

Law Enforcement Strategies and the Reduction of Gun Violence

Because of the above results, more recent efforts to control violence have looked to law enforcement rather than new legislation (Ludwig, 2005) as the potential solution. Within recent years, numerous cities have initiated 'gun emphasis' programs, directing law enforcement to target gun possession by persons with previous felony convictions. These approaches move to more aggressive enforcement of existing laws instead of enacting new laws. In general, there is little to no evidence that the number of guns legitimately owned by private citizens has a discernible effect on crime (Stolzenberg and D'Alessio, 2000). There is some indication, however, that reducing the use of firearms in crime would reduce the rate of firearm-related injuries and fatalities (Varano et al., 2004; Weaver et al., 2004).

Differing law enforcement approaches to getting guns off the streets share common thematic elements. For example, each program focuses on identifying 'hot spots' of criminal gun activity and intervening. Most of these programs also have community outreach and partnership components, and some degree of interagency cooperation efforts spanning local, state, and federal levels. A common element is the 'pulling levers' strategy, comprised essentially of 'throwing the book' at perpetrators of violence through increased focus and cooperation from law enforcement officials, prosecutors, and the courts. The general theory of 'pulling levers' is that law enforcement can deter violent behavior by reaching out to gang members and setting standards for their behavior (Corsaro et al., 2013). Law enforcement agencies then 'pull every lever' for punishment when those criterions are violated.

Although theoretically promising, the evidence is mixed on whether these programs curb gun violence. Braga and Pierce (2005) found significant reductions in firearm violence for Boston's Operation Ceasefire, but a reanalysis by Rosenfeld et al. (2005) did not confirm these. Fagin et al. (1998) found that Compstat had a lowered violent crime, but again Rosenfeld et al. (2005) disputed the result. Ironically, Rosenfeld et al. (2005) found support for the violence-reductive effects of Project Exile in Richmond, but the analysis by Raphael and Ludwig (2003) did not. The efficacy of Operation Ceasefire in Los Angeles and Atlanta's Project PACT is also questionable (Tita et al., 2005; Kellermann et al., 2006). Researchers indicated that The Indianapolis 'Directed Patrol Project' was a fairly cost-effective deterrent to firearm crime (McGarrell et al., 2002). The St. Louis Consent-to-Search program started strong, but the effect dwindled after policy changes altered the nature of the program (Decker and Rosenfeld, 2004). A mixed-methods evaluation of a pulling-levers approach in Rockford, Illinois was associated with a significant reduction in crime and drug use (Corsaro et al., 2013).

The National Institute of Justice published annual reviews on these gun enforcement strategies from 2001 to 2006, and the best that can be said is that some programs seem to have been at least marginally effective in some years under some conditions, whereas others were not.

It is hard to come to an omnibus conclusion in the face of these inconsistent results. It is difficult to control exogenous factors – such as declining crime and violent crime rates in most cities from the mid-1990s forward, employment trends, demographic changes, policing strategies, and political environments (Rosenfeld et al., 2005; Raphael and Ludwig, 2003) – in these kinds of evaluation studies.

More recently, a pulling-levers strategy adopted as part of the DOJ-sponsored Project Safe Neighborhoods was evaluated in Lowell, Massachusetts. This strategy focused on a small number of gang members who were assessed as playing a large role in citywide gun violence. This initiative included the involvement of the local police; Bureau of Alcohol, Tobacco, and Firearms (ATF); FBI; county prosecutors; probation and parole officers; and juvenile corrections caseworkers. The program combined public outreach and intervention programs with increased law enforcement, prosecutorial, and correctional attention. As a result, the intervention plan was found to have significantly decreased the amount of gun-related violence in Lowell (Braga et al., 2008).

Based on various evaluations, several things can be said regarding the efficacy of gun enforcement programs. First, problem-oriented policing (such as focusing resources heavily on a particular problem such as gun crime by youth) seems to show some potential. Also, programs with specialized directed patrol techniques ('hot spots') appear to be more effective than those with more general patrol policies. Interagency cooperation is clearly integral to gun enforcement programs. Further, 'pulling levers' only works when these agencies are all on board with the strategy. The lesson is apparently that reactive policing (where officers simply respond to calls for service) will not suffice. It is also apparent that programs work best when support is gathered from the public, and when realistic and meaningful alternatives for those involved in violent activity are provided. Unfortunately, when violent crime spikes, most people want immediate action, and things like community partnerships, community involvement, or 'community policing' take a backseat.

See also: Crime Trends and Debates; Crime and Class; Crime and Criminal Justice: Applied Research from Routine Monitoring to Evidence-Based Practices; Crime, Race and Ethnicity; Crime, Sociology of; Data Bases and Statistical Systems: Crime and Delinquency; Homicide; Routine Activities, Criminal Opportunities, Crime and Crime Prevention; Youth Gangs.

Bibliography

Blumstein, A., 2006. The crime drop in America: an exploration of some recent crime trends. Journal of Scandinavian Studies in Criminology & Crime Prevention 7, 17–35.

Brady Campaign to Prevent Gun Violence, 2013. Obama Administration Action Closes One Loophole Around Current Background Checks, Limits Availability of Military-Grade Weapons (Press Release). Retrieved from: http://www.bradycampaign.org/inthenews/obama-administration-action-closes-one-loophole-around-current-background-checks-limits.

Braga, A.A., Pierce, G., 2005. Disrupting illegal firearms markets in Boston: the effects of operation ceasefire on the supply of new handguns to criminals. Criminology & Public Policy 4, 717–748.

Braga, A.A., Pierce, G.L., McDevitt, J., Bond, B.J., Cronin, S., 2008. The strategic prevention of gun violence among gang-involved offenders. Justice Quarterly 25, 132–162.

Brezina, T., Wright, J.D., 2000. Going armed in the school zone. Forum for Applied Research and Public Policy 15, 82–86.
Bureau of Justice Statistics, 2004. Survey of Inmates in State Correctional Facilities (Siscf). US Department of Justice.
Bureau of Justice Statistics, 2013. Firearm Violence, 1993–2011. US Department of Justice.
Calhoun, D., 2013. Decreasing the supply of and demand for guns: Oakland's youth advocacy project. Journal of Urban Health: Bulletin of the New York Academy of Medicine 91, 72–83.
Cohen, L., Felson, M., 1979. Social change and crime rate trends: a routine activity approach. American Sociological Review 44, 588–608.
Cook, P., Ludwig, J., Braga, A., 2005. Homicide offending and prior criminal record. Journal of the American Medical Association 294, 598–601.
Copes, H., Kerley, K.R., Carroll, A., 2002. Killed in the act: a descriptive analysis of crime-precipitated homicide. Homicide Studies 6, 240–257.
Corsaro, N., Brunson, R.K., McGarrell, E.F., 2013. Problem-oriented policing and open-air drug markets: examining the Rockford pulling levers deterrence strategy. Crime & Delinquency 59, 1085–1107.
Decker, S.H., Rosenfeld, R., 2004. Reducing Gun Violence: The St. Louis Consent-to-Search Program – Research Report. National Institute of Justice, Washington, DC.
Dobrin, A., Bursk, J., 2003. The risk of offending on homicide victimization: a public health concern. American Journal of Health Behavior 27, 603–612.
Dobrin, A., Lee, D., Price, J., 2005. Neighborhood structure differences between homicide victims and non-victims. Journal of Criminal Justice 33, 137–143.
Fagin, J., Zimring, F., Kim, J., 1998. Declining homicide in New York city: a tale of two trends. Journal of Criminal Law and Criminology 88, 1277–1323.
Faria, M., 2005. Gun violence is not a public health crisis. In: Egendorf, L. (Ed.), Guns and Violence. Thomson Gale, Detroit, pp. 39–53.
Fleeger, E.,W., Lee, L.K., Monuteaux, M.C., Hemenway, D., Mannix, R., 2013. Firearm legislation and firearm-related fatalities in the United States. JAMA-Internal Medicine 173, 732–740.
Goering, G.E., 2010. Gun buybacks and firm behavior: do buyback programs really reduce the number of guns? Review of Economics and Finance. Retrieved 03.09.14. from: http://www.bapress.ca/journal-1.html.
Hahn, R.A., Bilukha, O., Crosby, A., Fullilove, M.T., Liberman, A., Moscicki, E., Snyder, S., Tuma, F., Briss, P.A., 2005. Firearms laws and the reduction of violence: a systematic review. American Journal of Preventative Medicine 28, 40–71.
Hannon, L., DeFina, R., 2010. The state of the economy and the relationship between prisoner reentry and crime. Social Problems 57, 611–629.
Hauser, W., Kleck, G., 2013. Guns and fear: a one-way street? Crime & Delinquency 59, 271–291.
Hepburn, L., Miller, M., Azrael, D., Hemenway, D., 2004. The effect of nondiscretionary concealed weapon carrying laws on homicide. Journal of Trauma 5, 676–681.
Kellermann, A.L., Fuqua-Whitley, D., Parramore, C.S., 2006. Reducing Gun Violence: Community Problem Solving in Atlanta – Research Report. National Institute of Justice, Washington, DC.
Kopel, David, 2000. The facts about gun shows. The Cato Institute. Available at: http://www.cato.org/dailys/01-10-00.html.
Legault, R.L., Hendrickson, J.C., 2009. Weapon choice and American political violence. Criminology & Public Policy 8, 531–559.
Levitt, S.D., 2004. Understanding why crime fell in the 1990s: four factors that explain the decline and six that do not. Journal of Economic Perspectives 18, 163–190.
Lott, J.R., Whitley, J.E., 2001. Safe-storage gun laws: accidental deaths, suicides, and crime. Journal of Law and Economics XLIV, 659–689.
Ludwig, J., Cook, P., 2000. Homicide and suicide rates associated with implementation of the Brady Handgun violence prevention act. Journal of the American Medical Association 284, 585–591.
Ludwig, J., 2005. Better gun enforcement, less crime. Criminology & Public Policy 4, 677–716.
McCarthy, B., 2002. New economics of sociological criminology. Annual Review of Sociology 28, 417–442.
McGarrell, E.F., Chermak, S., Weiss, A., 2002. Reducing Gun Violence: Evaluation of the Indianapolis Police Department's Directed Patrol Project – Research Report. National Institute of Justice, Washington, DC.
McGarrell, E., 2005. Gun control does not reduce gun violence. In: Egendorf, L. (Ed.), Guns and Violence. Thomson gale, Detroit, pp. 88–92.
Moorhouse, J., Wanner, B., 2006. Does gun control reduce crime or does crime increase gun control? Cato Journal 26, 103–124.
Phillips, S.S., Kim, D., Sobol, J.J., 2013. An evaluation of a multiyear gun buy-back programme: re-examining the impact on violent crimes. International Journal of Police Science & Management 15, 246–261.
Price, M., Norris, D.M., 2010. Firearm laws: a primer for psychiatrists. Harvard Review of Psychology 18, 326–335.
Raphael, S., Ludwig, J., 2003. Prison sentence enhancements: the case of project exile. In: Ludwig, J., Cook, P. (Eds.), Evaluating Gun Policy: Effects on Crime and Violence. Brookings Institution, Washington, DC.
Rosenfeld, R., 2004. The case of the unsolved crime decline. Scientific American 290, 70–77.
Rosenfeld, R., 2005. Firearms research and the crime drop. Criminology & Public Policy 4, 799–806.
Rosenfeld, R., Fornango, R., Baumer, E., 2005. Did ceasefire, compstat, and exile reduce homicide? Criminology & Public Policy 4, 419–450.
Rosenfeld, R., Deckard, M.J., Blackburn, E., 2014. The effects of directed patrol and self-initiated enforcement on firearm violence: a randomized controlled study of hot spot policing. Criminology 52, 428–449.
Schildkraut, J., Hernandez, T., 2014. Laws that bit the bullet: a review of legislative responses to school shootings. American Journal of Criminal Justice 39, 358–374.
Sheley, J., Wright, J.D., 1993. Gun Acquisition and Possession in Selected Juvenile Samples – Research in Brief. National Institute of Justice, Washington, DC.
Sheley, J., Wright, J.D., 1995. In the Line of Fire: Youth, Guns, and Violence in Urban America. Aldine de Gruyter, New York.
Sheley, J., Wright, J.D., 1998. High School Youths, Weapons and Violence: A National Survey – Research in Brief. National Institute of Justice, Washington, DC.
Spano, R., Freilich, J., Bolland, J., 2008. Gang membership, gun carrying, and employment: applying routine activities theory to explain violent victimization among inner city, minority youth living in extreme poverty. Justice Quarterly 25, 381–410.
Stell, L., 2004. The production of criminal violence in America: is strict gun control the solution? Journal of Law, Medicine, & Ethics 32, 38–46.
Stolzenberg, L., D'Alessio, S.J., 2000. Gun availability and violent crime: new evidence from the national incident-based reporting system. Social Forces 78, 1461–1482.
Tita, G.E., Riley, K.J., Ridgeway, G., Greenwood, P.W., 2005. Reducing Gun Violence: Operation Ceasefire in Los Angeles – Research Report. National Institute of Justice, Washington, DC.
Varano, S., McCluskey, J., Patchin, J., Bynum, T., 2004. Exploring the drugs-homicide connection. Journal of Contemporary Criminal Justice 20, 369–392.
Weaver, G.S., Wittekind, J.E.C., Huff-Corzine, L., Corzine, J., Petee, T.A., Jarvis, J.P., 2004. Violent encounters: a criminal event analysis of lethal and nonlethal outcomes. Journal of Contemporary Criminal Justice 20, 348–368.
Webster, D.W., Gainer, P.S., Champion, H.R., 1993. Weapon carrying among inner-city junior high school students: defensive behavior vs. aggressive delinquency. American Journal of Public Health 83, 1604–1608.
White, M.D., 2014. The New York City Police Department, its crime control strategies and organizational changes, 1970–2009. Justice Quarterly 31, 110–131.
Wright, J.D., Sheley, J.F., 1993. Motivations for gun possession and carrying among serious juvenile offenders. Behavioral Sciences and the Law 11, 375–388.
Wright, J.D., Rayburn, R.L., 2012. Felons and gun control. In: Guns in American Society: An Encyclopedia of History, Politics, Culture, and the Law, second ed. ABC-CLIO, Santa Barbara, Calif.
Wright, J.D., Rossi, P., 1986. Armed and Considered Dangerous, a Survey of Felons and Their Firearms. Aldine de Gruyter, New York.

Gynecological Health: Psychosocial Aspects

Pearl La Marca-Ghaemmaghami and Ulrike Ehlert, Institute of Psychology, University of Zurich, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C. Lee, volume 9, pp. 6448–6451, © 2001, Elsevier Ltd.

Abstract

Gynecological and obstetric health is closely linked to female psychological well-being and calls for a holistic bio-psychosocial and life span approach. The first section of the current contribution offers a short overview of selected gynecological disorders related to the menstrual cycle including menstrual cycle disorders, premenstrual syndrome/premenstrual dysphoric disorder, endometriosis, polycystic ovary syndrome, and infertility. Sexual dysfunctions and urinary incontinence are presented in the section on gynecological disorders not related to the menstrual cycle. Finally, psychosocial aspects of obstetric health and menopause are discussed.

Introduction

Gynecology is a medical discipline dedicated to female health care concerns including the development, diagnosis, prevention, and therapy of disorders and diseases distinct to the female reproductive system. The medical care of a women and her child before, during, and after birth is attributed to the field of obstetrics (Cunningham et al., 2010). Gynecology and obstetrics are related disciplines concerned with female reproductive health.

Across life span, women experience multiple events associated with reproductive development such as menarche, menstruation, pregnancy, maternity, and menopause. Compared to developmental events in male reproduction, these developmental events in female reproduction induce more drastic physiological alterations (e.g., menstrual bleeding, physiological changes in pregnancy, breast feeding, menopausal hormone fluctuations), more significant psychological changes and more complex psychosocial consequences for women. Though most women adapt well to these reproductive events, under certain conditions, they may develop an increased vulnerability for psychological maladjustment. Psychological problems, in turn, might exacerbate physiological symptoms associated with reproductive conditions (e.g., the influence of stress on menstrual cycle disorders). This interconnectedness between psychological and gynecological factors is reflected in findings reporting prevalence rates for mental disorders of up to 45.3% in gynecological outpatients (Leithner et al., 2009).

Female reproductive health is thus closely associated with psychological well-being and, therefore, a range of psychosocial, psychosexual, psychobiological, behavioral, and developmental psychological aspects. Consequently, gynecological and obstetric issues must be approached in a bio-psychosocial framework of health and illness that encompasses women's development and experiences across life span.

Psychosocial Aspects of Gynecological Health

Gynecological Disorders Related to the Menstrual Cycle

While menstrual cycle-related disorders can be secondary to medical conditions (e.g., cancer, hormonal dysfunctions), psychosocial (e.g., stress) and behavioral factors (e.g., excessive dieting, athletic training, obesity) may also trigger, sustain, or exacerbate these disorders. A basic biological knowledge of the female menstrual cycle is essential to understand menstrual cycle-related disorders. A short introduction is presented in the following section.

The Female Menstrual Cycle

The menstrual cycle begins with the first day of menstrual bleeding, lasts 28 days (plus or minus 3 days), and is characterized by cyclic changes within the ovary and endometrium (inside lining of the uterus).

The ovarian cycle is divided as follows:

1. the follicular phase, during which the ovarian follicle containing the immature egg cell (oocyte) develops,
2. ovulation, during which the mature egg cell (ovum) is released into the fallopian tube, and
3. the luteal phase, during which the remaining follicle cells in the ovary convert into the corpus luteum.

The corpus luteum is a transient endocrine gland that produces high levels of progesterone and moderate levels of estrogen to prepare the endometrium for potential implantation of an embryo (Cunningham et al., 2010). If no fertilization occurs, the corpus luteum withers and the concomitant decline in progesterone levels results in the shedding of the endometrium and menstrual bleeding (Henriet et al., 2012).

The endometrial cycle is divided as follows:

1. the menstrual phase, during which the top layer of the endometrium breaks down and flows out through the vagina,
2. the proliferative phase, during which the endometrium thickens, and
3. the secretory phase, during which the endometrium differentiates for preparation of implantation.

The complex processes of the female menstrual cycle are controlled by the hormones of the hypothalamic-pituitary-ovarian (HPO) axis. The hypothalamus releases gonadotropin-releasing-hormone (GnRH) in a pulsatile fashion, and induces a pulsatile secretion of follicle-stimulating hormone (FSH) and luteinizing hormone (LH) from the anterior pituitary into the blood stream. In the ovaries, FSH and LH stimulate the

production of progesterone and estrogen, which exert their influences on the endometrium and other end organs and send negative and positive feedback signals to the pituitary and hypothalamus (Hoffman et al., 2012). For example, during the follicular phase, FSH stimulates the growth of follicles (Beshay and Carr, 2013). In response, cells surrounding the follicle start producing estrogen to further advance follicular development. The increasing estrogen concentrations trigger a LH surge, which ultimately leads to ovulation (Hawkins and Matzuk, 2008). Thus, during this phase of the menstrual cycle a positive feedback loop is established.

While estrogen production by the ovaries dominates during the follicular phase, progesterone production dominates during the luteal phase (Snowdon and Ziegler, 2007).

The HPO axis activity is influenced by numerous hormones including those of the hypothalamic-pituitary-adrenal (HPA) axis, which are released in response to psychosocial stress (Snowdon and Ziegler, 2007). For instance, corticotrophin-releasing hormone (CRH) and cortisol have an inhibiting effect on GnRH and LH secretion. This is thought to explain, at least in part, how stress might negatively affect gynecological health.

Menarche

Menarche is the first menstrual period in adolescent girls and occurs at an average age of 13.5 years (Thomas et al., 2001). It marks the beginning of the reproductive phase in the life of a woman.

The degree to which young girls are prepared for menstruation varies. Often they receive information from their mothers, female peers, and the media (van den Akker, 2012). While, the hygienic aspect generally tends to be emphasized, biological, practical, and psychological information seem to be neglected (e.g., menstrual cycle as something normal and natural; as a method to monitor gynecological health). Accordingly, many young girls tend to have a negatively biased or ambivalent attitude toward menarche and menstruation (Stubbs, 2008).

The onset of menarche is largely influenced by genetic factors. However, body fat composition, eating behavior, physical training, chemical pollutants (i.e., endocrine disrupting chemicals), and psychosocial stress can also impact on the timing of menarche (Allison and Hyde, 2011). For example, obesity and stressful life events prior to menarche including parental divorce, absence of a father, maternal mood disorders, presence of a stepfather, or an unrelated man at home are factors associated with earlier menarche. In contrast, excessive dieting, physical training, and thinness are related to a delayed onset (Snowdon and Ziegler, 2007; van den Akker, 2012).

Girls reaching menarche earlier than their peers tend to worry more about menstruation, experience menstruation less positively, experience depression, become more likely victims of peer sexual harassment, and are more likely involved in deviant behavior (Allison and Hyde, 2011). Therefore, further systematic research on the prevention of these adverse outcomes in young girls is necessary.

Menstrual Cycle Disorders

Menstrual cycle disorders are divided in cycle, duration, flow, and pain abnormalities (van den Akker, 2012). Cycle abnormalities include oligomenorrhea (cycle intervals exceeding 35 days), polymenorrhea (cycle intervals of less than 21 days), primary amenorrhea (absence of menarche by the age of 16), and secondary amenorrhea (absence of menstruation for more than 3 months). Normally, menstrual bleeding lasts 3–7 days. Polymenorrhagia (menstrual bleeding exceeding 7 days) and metrorrhagia (bleeding in between menstrual periods) are abnormalities regarding the duration of menstrual bleeding. Flow abnormalities include hypomenorrhea (diminished blood flow) and menorrhagia (very heavy blood flow). Dysmenorrhea is the term used for excessive pelvic pain during menstrual bleeding.

Menstrual cycle disorders are common. Dysmenorrhea is reported in approximately 75–80% of adolescent and in 20–50% of adult women (Harlow and Campbell, 2004; van den Akker, 2012). Oligo- or polymenorrhea have been described by 1.1–42% of women depending on the study and the country of investigation (Harlow and Campbell, 2004).

Menstrual cycle disorders can be caused by structural abnormalities of the endometrium (e.g., tumors), endometrial infections (e.g., chronic endometritis), and by disorders affecting the HPO axis (e.g., renal or thyroid diseases) (Hoffman et al., 2012). Psychosocial stress is also known to impact on the stability of the menstrual cycle and influence the secretion of the HPO axis hormones. For example, a population-based prospective study found that women with high-daily stress levels during one cycle had a twofold increased risk of experiencing dysmenorrhea in the subsequent cycle (Wang et al., 2004). Studies also point toward increased menstrual cycle irregularities after a natural disaster (Li et al., 2011) or a period of war (Hannoun et al., 2007).

Presently, menstrual cycle disorders are treated mainly with painkillers, estrogen therapy, or contraceptive pills. In some severe cases, the uterus is surgically removed (hysterectomy). Few studies have examined the effectiveness of psychotherapeutic interventions. A recent study suggests that cognitive behavioral therapy (CBT) may contribute to the recovery of ovarian function in women with functional hypothalamic amenorrhea (a reversible condition involving absent menstruation with no identifiable organic or anatomical disease) (Michopoulos et al., 2013). Relaxation techniques and physical exercise may alleviate symptoms of dysmenorrhea (Ehlert and Fischbacher, 2012).

Premenstrual Syndrome and Premenstrual Dysphoric Disorder

Premenstrual syndrome (PMS) is characterized by diverse physical and psychological symptoms that occur during the last week of the luteal phase and subside with, or a few days after menstruation (Hoffman et al., 2012). Premenstrual dysphoric disorder (PMDD) – a more severe form of PMS – is defined by the following core psychological symptoms according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-V; American Psychiatric Association, APA, 2013): emotional liability, irritability, depressed mood, and anxiety. At least one additional behavioral (e.g., social withdrawal) or physiological symptom (e.g., breast tenderness) must be present to meet the diagnostic criteria of PMDD. Moreover, symptoms need to have been present during most of the menstrual cycles of the preceding year and involve clinically significant distress, or result in impairment of relationships or work. Symptom-free

periods during the follicular phase are essential to the diagnosis. Therefore, prospective daily symptom ratings are required for a minimum of two cycles to confirm diagnosis.

A recent population-based study found that most women (91%) experienced at least one premenstrual symptom, while 10.3% had PMS, and 3.1% PMDD (Tschudin et al., 2010).

The underlying cause of PMS and PMDD is unclear. Numerous (sometimes contradicting) biological and psychological theories have been proposed. Biological theories focus on the potential influence of estrogen and progesterone and on the neurotransmitters gamma amino butyric acid and serotonin (Hoffman et al., 2012). Genetic factors such as serotonin receptor gene polymorphisms may be involved as well (Dhingra et al., 2007). Psychological models highlight the significance of stress, interpersonal trauma, sociocultural attitudes toward female sexuality, and gender roles in the etiology of PMS/PMDD (APA, 2013). For example, women with PMDD seem more likely to have a history of sexual or physical abuse than women with no PMDD (Girdler et al., 2003). Also, some studies have found marital dissatisfaction to predict PMS/PMDD symptom severity (cf van den Akker, 2012).

Oral contraceptives, selective serotonin reuptake inhibitors, vitamins (e.g., vitamins B_6 and E), minerals (e.g., magnesium, calcium), herbs (e.g., vitex agnus castus), nutrition (e.g., restricted sugar intake), exercise, relaxation, stress management, and psychotherapy are implemented to alleviate PMS/PMDD symptoms (Ehlert and Fischbacher, 2012; Hoffman et al., 2012; van den Akker, 2012). Certain treatment methods are, however, contested (e.g., nutrition, oral contraceptives) and further research, also regarding the etiology of PMS and PMDD, is needed to establish more effective treatments.

Polycystic Ovary Syndrome

Polycystic ovary syndrome (PCOS) is an endocrine disorder of the HPO-axis with heterogeneous symptoms. The chief diagnostic criteria include oligoovulation (infrequent, irregular ovulation) and/or anovulation (no ovulation), hyperandrogenism (excessive androgen secretion), which becomes clinically visible in acne and hirsutism (excessive body and facial hair) and polycystic ovaries identified by ultrasound (Hoffman et al., 2012).

Oligo- or anovulation can lead to menstrual cycle disorders such as oligo- or amenorrhea and consequently women with PCOS can be confronted with fertility problems. Further frequent short-term consequences of PCOS are obesity, abnormal blood lipid concentrations, and glucose intolerance. Long-term consequences include an increased risk of developing type 2 diabetes, high blood pressure, and cardiovascular disease. PCOS is one of the most common endocrine disorders affecting women of reproductive age (Ehlert and Fischbacher, 2012). Prevalence rates range between 6.1 and 19.9% (Yildiz et al., 2012).

The causes of PCOS remain elusive. Genetic factors involved in androgen production and insulin resistance may play a role. The main etiological hypotheses include dysregulation in pulsatile GnRH secretion, insulin resistance, and abnormal ovarian androgen production.

Compared to healthy control women, women with PCOS experience significant psychosocial distress, decreased satisfaction with physical appearance and sexuality (Elsenbruch et al., 2003), higher symptoms of depression and anxiety, and lower degrees of quality-of-life (QoL) (Veltman-Verhulst et al., 2012). Therefore, in addition to medical treatment, women with PCOS may benefit from psychotherapeutic interventions.

Endometriosis

Endometriosis is a frequent, often painful, chronic gynecological condition where endometrial glands and tissues are found outside the uterus (e.g., on the ovaries, the fallopian tubes, the outer surface of the uterus) (Hoffman et al., 2012). These misplaced tissues respond to the hormonal changes of the menstrual cycle. They grow and breakdown monthly, as does the endometrium. However, since the blood cannot flow out, blood-filled cysts develop. This ultimately causes (cyclic or chronic) pelvic pain.

In addition to pelvic pain, symptoms may include dysmenorrhea, pain during sexual intercourse (dyspareunia), urination (dysuria), and defecation, and fatigue and fertility problems (Culley et al., 2013). The severity of symptoms varies largely. Some women with endometriosis are even asymptomatic (Hoffman et al., 2012).

The diagnosis of endometriosis is confirmed by laparoscopy. But, the sites where endometriosis cysts are surgically identified often do not correspond to the sites where pain is felt (Ehlert and Fischbacher, 2012). This makes symptom treatment difficult. It is important to note that, on average, 5–8.9 years can pass from symptom onset to diagnosis. Reasons for such delays can stem both from women not seeking medical help and doctors failing to identify symptoms correctly (Culley et al., 2013). Assessing the prevalence of endometriosis is similarly challenging. Rates lie between 6 and 10% (Giudice and Kao, 2004).

The etiology of endometriosis remains unclear. While dysregulation in estrogen synthesis and metabolism seems to play a key role, it has also been suggested that endometrial tissues are carried away either by retrograde menstruation through the fallopian tubes, or spread via lymphatic or vascular circulation. Furthermore, evidence exists that the immune system may not respond adequately to remove the misplaced endometrial tissue. Heritability, environmental toxins, and anatomical anomalies may contribute to an increased risk of developing endometriosis as well (Giudice and Kao, 2004; Hoffman et al., 2012).

The strong, reoccurring experience of pain can substantially affect psychological well-being and social life. Women with endometriosis frequently report significant impairment in work performance, physical functioning, daily and social activities, reduction in sexual and general QoL, and increased emotional distress, anxiety, and depressive symptoms (Culley et al., 2013). Infertility may additionally cause feelings of inadequacy and negatively affect partner relationship.

The medical, but also psychological treatment depends on symptom specificity and severity. Pain is usually treated with medication (e.g., anti-inflammatory painkillers, oral contraceptives) and in more severe cases with surgery. Less research is available on psychotherapeutic interventions. Endometriosis support groups (including online platforms) have been found to be appreciated by affected women (Culley et al., 2013).

Infertility

The clinical definition of infertility is the inability to become pregnant after 12 months of regular, unprotected intercourse. According to the epidemiological definition, infertility is classified after 2 years (WHO, 2006). The term encompasses the inability to carry a pregnancy to a life birth. Primary infertility is diagnosed if a couple has never conceived before, and secondary infertility if a couple has difficulties conceiving a second time. In many cases, pregnancy is achieved after 2 years also without fertility treatment (Hoffman et al., 2012).

It is estimated that 24% of couples do not conceive within 12 months and 11% within 24 months (Slama et al., 2012).

Causes for infertility are manifold and include behavioral (e.g., advanced age by delaying to have children, obesity), genetic (e.g., genetic disorders), endocrine/hormonal (e.g., hormonal imbalances), and environmental factors (environmental toxins) as well as factors related to diseases (e.g., sexually transmitted diseases) and cancer treatment during child or adulthood (van den Akker, 2012).

Whether psychosocial stress is of etiological significance for infertility, is difficult to assess, as most studies are cross-sectional in design rendering it impossible to disentangle cause and effect, since the diagnosis of infertility itself induces significant distress. Psychosocial stress is, however, known to affect the menstrual cycle as discussed above. Stress has also been associated with an increased risk of spontaneous abortions (involuntary pregnancy loss before the 20th week of pregnancy) and reduced semen quality in men (Nakamura et al., 2008).

The diagnosis of infertility seems to affect women more negatively than men (Hämmerli et al., 2009). Yet, both men and women commonly experience a loss in self-esteem and reduced marital and sexual satisfaction. Active coping and information seeking are associated with less distress, while self-blame, avoidance, and social withdrawal are linked to increased suffering (van den Akker, 2012).

Treatment includes lifestyle changes (e.g., smoking cessation, weight reduction), hormone stimulation, operation of anatomical abnormalities, and assisted reproductive techniques (e.g., insemination, *in vitro* fertilization) (Hoffman et al., 2012). Adoption is an alternative for overcoming involuntary childlessness. Psychological interventions to assist coping with infertility can reduce mental distress and may possibly increase pregnancy rate (De Liz and Strauss, 2005; Hämmerli et al., 2009). However, methodological limitations and conflicting findings call for more systematic research.

Gynecological Disorders Not Related to the Menstrual Cycle

Sexual Dysfunctions

The APA (2013) classifies female sexual dysfunctions according to the following categories: orgasmic disorders, disorders regarding arousal and sexual interest, and pain during intercourse. The problem may either exist from the time of the first sexual experience (lifelong sexual disorder) or develop with time (acquired sexual dysfunction), occur only in specific circumstances (situational) or in all situations (generalized).

Prevalence rates vary depending on the investigated criteria, the applied research methods, the assessed age group or cultural setting. Orgasmic problems have been reported by 10–42% of women, low or no sexual desire by 64%, and pain during intercourse by 15–26% (APA, 2013; Hayes et al., 2006). Interestingly, only a proportion of women experiencing sexual problems seem to feel troubled by it. However, if distress exists, intimate relationship and QoL are likely to be negatively affected (van den Akker, 2012).

Factors important to diagnosis, etiology, and treatment of sexual dysfunctions relate to the partner (e.g., the partner has sexual difficulties too), the relationship (e.g., communication problems), individual vulnerability (e.g., childhood sexual abuse), comorbid mental disorders and stress (e.g., work stress), culture and religion (e.g., tabooing sexual education), and medical factors (e.g., medication) (APA, 2013).

Women most frequently cite relationship conflict as cause of sexual problems (Wylie, 2007). Psychological interventions include sexual education, sensate focus, individual and couple psychotherapy, communication training, relaxation techniques, and biofeedback.

Urinary Incontinence

Urinary incontinence is the involuntary loss of urine (Riss and Kargl, 2011). There are three common forms of incontinence:

1. Stress incontinence is the involuntary loss of urine caused by an increased pressure on the bladder, for example, during coughing, laughing, or sneezing. This is the most common form of incontinence.
2. Urge incontinence is the involuntary loss of urine prior to a sudden urge to urinate, which cannot be controlled.
3. In some cases, stress and urge incontinence are both present and termed mixed urinary incontinence.

Prevalence rates range between 25 and 45% depending on the definition of incontinence and the implemented methods, and increase with age (Buckley et al., 2010). Many women do not seek help out of embarrassment.

Causes for urinary incontinence include hereditary factors, damaged nerves following surgery or trauma, injury of the pelvic muscles following childbirth, or increased intra-abdominal pressure caused by obesity (Hoffman et al., 2012). Incontinence can lead to significant psychological distress and negatively impact women's QoL, daily work, social activities, and sexuality (Riss and Kargl, 2011).

Exercises to strengthen pelvic floor muscles, biofeedback, and scheduled voiding are nonsurgical and nonmedical treatment methods often implemented in psychotherapeutic interventions.

Psychosocial Aspects of Obstetric Health

Pregnancy

Pregnancy (gestation) begins with conception, ends with birth, and spans approximately 38 weeks. But, since the exact time of conception is often difficult to determine, except in pregnancy conceived by *in vitro* fertilization, pregnancy duration is calculated beginning from the first day of the last menstrual cycle – approximately 2 weeks prior to ovulation. This adds up to a total of 40 weeks (Cunningham et al., 2010).

During pregnancy, the female body undergoes massive changes, and the placenta develops into an additional

endocrine organ that produces and secretes hormones in increasing amounts, including progesterone and estrogen (Cunningham et al., 2010). Progesterone and estrogen exert negative feedback on FSH and LH, and thus egg maturation in the ovaries is suppressed during pregnancy.

The placenta also synthesizes CRH and adrenocorticotropin in increasing amounts. This causes an upregulation of the maternal HPA axis and an increase in maternal cortisol levels. Cortisol can cross the placenta and reach the fetus. To a certain extent it furthers fetal lung and brain development (La Marca-Ghaemmaghami and Ehlert, in press). Fetal overexposure to maternal cortisol, however, has been associated with adverse pregnancy outcome including preterm birth and low birth weight. It is further assumed that cortisol excess *in utero* may have programming effects on fetal organ development rendering the child vulnerable to disease in later life, including hypertension, type 2 diabetes, and cardiovascular disease (Harris and Seckl, 2011).

The underlying biological mechanisms of prenatal stress are just beginning to be understood. Findings suggest that when a certain level of placental CRH is attained, processes are triggered that ultimately lead to birth (McLean and Smith, 1999). Excessive maternal stress may prematurely increase placental CRH concentrations via maternal cortisol which unlike hypothalamic CRH stimulates placental CRH production.

A further mechanism by which stress may affect the fetus involves a placental enzyme named 11β-hydroxysteroid dehydrogenase type 2 (11β-HDS2). This enzyme controls the amount of maternal cortisol crossing the placenta by inactivating cortisol to cortisone. It is assumed that acute stress induces an upregulation of 11β-HDS2 activity, whereas chronic stress impairs its function exposing the fetus to increased cortisol concentrations (Welberg et al., 2005).

Although pregnancy is usually a happy event, the many physiological, hormonal, psychological, social, and socioeconomic changes that are brought on by pregnancy can be sufficiently challenging to provoke psychological maladjustment. Some findings even suggest that mental health problems are higher during pregnancy than postpartum (van den Akker, 2012). The prevalence rates for prenatal depression range between 7 and 20% and are estimated at 6.6% for prenatal anxiety disorders. Anxiety symptoms have been found in 54% of pregnant women and depressive symptoms in 37.1% (Lee et al., 2007). Psychological problems during pregnancy often remain unidentified, undiagnosed, and untreated, and pose a risk to maternal mental health after birth (Ghaemmaghami and Ehlert, 2012).

Pregnancy per se does not induce psychological problems and women's experiences of pregnancy vary substantially (Stanton et al., 2002). Factors associated with increased prenatal stress and psychological maladjustment include socioeconomic inequalities, poverty, racism, unwanted or adolescent pregnancy, lack of a partner and social support in general, problematic in-law relationship, single motherhood, first pregnancy, previous pregnancy loss, pregnancy complications, pregnancy by rape, domestic abuse, critical life events, chronic stress, and pregnancy-specific anxiety (Dunkel Schetter and Glynn, 2011; WHO, 2008).

Systematic psychological intervention studies for the alleviation of stress and the treatment of depression and anxiety during pregnancy are rare. A few have found a beneficial effect of psychotherapy and relaxation techniques on psychological symptoms and levels of stress hormones (Dennis et al., 2007; La Marca-Ghaemmaghami and Ehlert, in press). However, with increased knowledge about how prenatal psychological problems can impact on postpartum maternal well-being and child development, more studies in the field of prevention and therapy are warranted.

The Postpartum Period

Despite the joy the birth of a child brings, new responsibilities and demands of motherhood, such as breast feeding and lack of sleep can cause substantial psychological strain and, under certain circumstances, raise the vulnerability for psychological maladjustment. Possible mental health problems during the postpartum period (period following childbirth) include: (1) postpartum blues, also known as maternity or baby blues, which is characterized by increased negative emotional reactivity in the first days following childbirth (prevalence: 55.2%), (2) maternal postpartum depression, which is a major depression following childbirth (prevalence: 6.5–22%), (3) postpartum post-traumatic stress disorder (PTSD), which can develop in response to a traumatic birth experience (prevalence: 6–14%, subsyndromal symptoms in 28%), (4) postpartum obsessive-compulsive disorder (OCD), which is characterized by maternal obsessive-compulsive thoughts of harming the child with or without intention (prevalence: 2.7–3.9%), (5) peri- or postnatal panic disorder with excessive fear concerning infant well-being (prevalence: 1.3–2%), and (6) postpartum psychosis, which is a sudden psychotic or bipolar depressive disorder with an increased risk for suicide or infanticide (prevalence: 0.003%) (van den Akker, 2012; Cunningham et al., 2010; Reck et al., 2009; Zaers et al., 2008). Fathers can develop postpartum depression as well (prevalence: 10.4%) (Paulson and Bazemore, 2010).

Besides postpartum psychosis, a general increased risk of maternal suicide during the perinatal period has also been reported in several studies. In fact, in the United Kingdom and Australia maternal suicide is the leading cause of death during the perinatal period (WHO, 2008).

The etiology of postpartum mental disorders is poorly understood, and the hormonal alterations after birth explain symptoms only partially. A prior history of mental health problems, medical complications during pregnancy or birth (e.g., preterm birth), low socioeconomic status, financial problems, relationship dissatisfaction, low social support, increased stress, and a history of sexual abuse can increase the risk of developing mental problems postpartum (van den Akker, 2012; Cunningham et al., 2010).

Receiving psychological treatment as soon as possible is crucial, because both maternal and paternal postpartum disorders can significantly affect infant emotional and behavioral development. In some cases, antidepressants are required, although psychotherapeutic treatment should receive preference.

Psychosocial Aspect of Menopause

Menopause marks the end of female reproductivity and is defined as the permanent cessation of menstruation. It is

determined retrospectively, after 12 months of absent periods and occurs internationally at an average age of 49.2 years (Drobnjak, 2012; Thomas et al., 2001). Menopause is embedded within a phase of hormonal changes associated with natural aging in women. This phase is termed menopausal transition or perimenopause. It begins at an average age of 47 years and lasts between 4 and 7 years (Hoffman et al., 2012). The phase following menopause is termed postmenopause.

The end of menstruation is caused by a gradual decline in ovarian function due to the age-related decrease in follicle numbers. Ultimately, the ovaries stop producing estrogen and progesterone, and the negative feedback that these hormones usually exert on the hypothalamus and pituitary is lost. Hence, secretion of GnRH from the hypothalamus and FSH and LH from the pituitary increases. The equilibrium of the HPO axis is temporarily disturbed until the body readjusts.

The physiological changes associated with menopause can induce diverse symptoms including hot flashes, fatigue and sleep disruptions, bone mass loss, cardiovascular changes and an increased risk for cardiovascular disease, weight gain, urinary incontinence, sexual problems, particularly pain during intercourse due to vaginal dryness, and psychological symptoms such as poor concentration, forgetfulness, and in some cases depressed mood (Drobnjak, 2012; Hoffman et al., 2012).

For up to 85% of menopausal women these symptoms are considered rather unpleasant, and QoL is reduced for more than 50% (van den Akker, 2012). Epidemiological studies indicate an increased risk of developing depression during menopausal transition by about 30% as compared to premenopause. This risk has been found to be five times higher in women with a prior history of depression (Freeman, 2010). It is nevertheless important to emphasize that menopause per se does not cause depression and that most women experience the years of menopausal transition positively.

Although hormonal changes in estrogen, FSH, and LH concentrations (e.g., increased FSH and LH and increased variability in estrogen, FSH, and LH concentrations during menopausal transition) have been suggested to affect women's mood, this hypothesis has not been clearly established (Freeman, 2010). The elevated risk of developing depression during menopause seems to be multifactorial and likely to be influenced by psychosocial and behavioral aspects as well. These may include factors such as the appraisal of and attitude toward menopause, aging, and the end of reproductivity, the influence of cultural values, critical life events, financial problems, low social support, single relationship status, lower educational status, increased BMI, sleep problems, and smoking (van den Akker, 2012; Freeman, 2010).

Treatment of menopausal symptoms ranges from lifestyle changes (e.g., physical activity) and antidepressants to hormone replacement therapy. The latter has been associated with considerable side effects including breast cancer and poor compliance. A few studies examined the effect of psychological interventions and CBT, and found symptom alleviations and improved coping (van den Akker, 2012).

See also: Adoption, Demography of; Family Planning: Program Effects; Fertility of Single and Cohabiting Women; Genetics, Disease, and Reproduction; Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of; Hypothalamus; Infertility: Demographic Aspects; Neonatal Temperament: From Phenomenology to Mechanism; Post-Traumatic Stress Disorder; Prenatal and Infant Development: Overview, Current Trends, Future Directions; Reproduction and Cultural Anthropology; Reproductive Rights in Affluent Nations; Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective; Sex Selective Abortion; Sexual Risk Behaviors; Sexuality: Cultural Aspects; Stress: Effects on Neonatal and Child Development.

Bibliography

van den Akker, O.B.A., 2012. Reproductive Health Psychology. Wiley-Blackwell, Chichester, West Sussex.

Allison, C.M., Hyde, J.S., 2011. Early menarche: confluence of biological and contextual factors. Sex Roles 68, 55–64.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Publishing, Arlington, VA.

Beshay, V.E., Carr, B.R., 2013. Hypothalamic-pituitary-ovarian axis and control of the menstrual cycle. In: Falcone, T., Hurds, W.W. (Eds.), Clinical Reproductive Medicine and Surgery: A Practical Guide. Springer, New York, pp. 31–42.

Buckley, B.S., Lapitan, M.C., Epidemiology Committee of the Fourth International Consultation on Incontinence, Paris, 2008, 2010. Prevalence of urinary incontinence in men, women, and children–current evidence: findings of the fourth international consultation on incontinence. Urology 76, 265–270.

Culley, L., Law, C., Hudson, N., Denny, E., Mitchell, H., Baumgarten, M., Raine-Fenning, N., 2013. The social and psychological impact of endometriosis on women's lives: a critical narrative review. Human Reproduction Update 19, 625–639.

Cunningham, F.G., Leveno, K.J.L., Bloom, S.L., Hauth, J.C., Rouse, D.J., Spong, C.Y., 2010. Williams Obstetrics, 23rd ed. McGraw-Hill, New York.

De Liz, T.M., Strauss, B., 2005. Differential efficacy of group and individual/couple psychotherapy with infertile patients. Human Reproduction 20, 1324–1332.

Dennis, C.L., Ross, L.E., Grigoriadis, S., 2007. Psychosocial and psychological interventions for treating antenatal depression. Cochrane Database of Systematic Reviews 18. CD006309.

Dhingra, V., Magnay, J.L., O'Brien, P.M., Chapman, G., Fryer, A.A., Ismail, K.M., 2007. Serotonin receptor 1A C(-1019)G polymorphism associated with premenstrual dysphoric disorder. Obstetrics and Gynecology 110, 788–792.

Drobnjak, S., 2012. Menopause. In: Gellman, M.G., Turner, J.R. (Eds.), Encyclopedia of Behavioral Medicine. Springer, New York, pp. 1220–1221.

Dunkel Schetter, C., Glynn, L., 2011. Stress in pregnancy: empirical evidence and theoretical issues to guide interdisciplinary researchers. In: Contrada, R., Baum, A. (Eds.), Handbook of Stress. Springer, New York, pp. 321–343.

Ehlert, U., Fischbacher, S., 2012. Reproductive health. In: Gellman, M.G., Turner, J.R. (Eds.), Encyclopedia of Behavioral Medicine. Springer, New York, pp. 1658–1665.

Elsenbruch, S., Hahn, S., Kowalsky, D., Offner, A.H., Schedlowski, M., Mann, K., Janssen, O.E., 2003. Quality of life, psychosocial well-being, and sexual satisfaction in women with polycystic ovary syndrome. Journal of Clinical Endocrinology and Metabolism 88, 5801–5807.

Freeman, E.W., 2010. Associations of depression with the transition to menopause. Menopause 17, 823–827.

Ghaemmaghami, P., Ehlert, U., 2012. In: Gellman, M.G., Turner, J.R. (Eds.), Encyclopedia of Behavioral Medicine. Springer, New York, pp. 861–865.

Girdler, S.S., Sherwood, A., Hinderliter, A.L., Leserman, J., Costello, N.L., Straneva, P.A., Pedersen, C.A., Light, K.C., 2003. Biological correlates of abuse in women with premenstrual dysphoric disorder and healthy controls. Psychosomatic Medicine 65, 849–856.

Giudice, L.C., Kao, L.C., 2004. Endometriosis. Lancet 364, 1789–1799.

Hämmerli, K., Znoj, H., Barth, J., 2009. The efficacy of psychological interventions for infertile patients: a meta-analysis examining mental health and pregnancy rate. Human Reproduction Update 15, 279–295.

Hannoun, A.B., Nassar, A.H., Usta, I.M., Zreik, T.G., Abu Musa, A.A., 2007. Effect of war on the menstrual cycle. Obstetrics and Gynecology 109, 929–932.

Harlow, S.D., Campbell, O.M.R., 2004. Epdemiology of menstrual disorders in developing countries: a systematic review. BJOG: An International Journal of Obstetrics and Gynaecology 111, 6–16.

Harris, A., Seckl, J., 2011. Glucocorticoids, prenatal stress and the programming of disease. Hormones and Behavior 59, 279–289.

Hawkins, S.M., Matzuk, M.M., 2008. The menstrual cycle: basic biology. Annals of the New York Academy of Sciences 1135, 10–18.

Hayes, R.D., Bennett, C.M., Fairley, C.K., Dennerstein, L., 2006. What can prevalence studies tell us about female sexual difficulty and dysfunction? Journal of Sexual Medicine 3, 589–595.

Henriet, P., Gaide Chevronnay, H.P., Marbaix, E., 2012. The endocrine and paracrine control of menstruation. Molecular and Cellular Endocrinology 358, 197–207.

Hoffman, B.L., Schorge, J.O., Schaffer, J.I., Halvorson, L.M., Bradshaw, K.D., Cunningham, F.G., 2012. Williams Gynecology, second ed. McGraw Hill, New York.

Li, X.H., Qin, L., Hu, H., Luo, S., Li, L., Fan, W., Xiao, Z., Li, Y.X., Li, S.W., 2011. Influence of the Wenchuan earthquake on self-reported irregular menstrual cycles in surviving women. Gynecological Endocrinology 27, 706–710.

Lee, A.M., Lam, S.K., Sze Mun Lau, S.M., Chong, C.S., Chui, H.W., Fong, D.Y., 2007. Prevalence, course, and risk factors for antenatal anxiety and depression. Obstetrics and Gynecology 110, 1102–1112.

Leithner, K., Assem-Hilger, E., Fischer-Kern, M., Loeffler-Stastka, H., Sam, C., Ponocny-Seliger, E., 2009. Psychiatric morbidity in gynecological and otorhinolaryngological outpatients: a comparative study. General Hospital Psychiatry 31, 233–239.

La Marca-Ghaemmaghami, P., Ehlert, U. Experienced Stress, Stress Hormones, and Stress Protective Factors during Pregnancy, manuscript in press.

McLean, M., Smith, R., 1999. Corticotropin-releasing hormone in human pregnancy and parturition. Trends in Endocrinology and Metabolism 10, 174–178.

Michopoulos, V., Mancini, F., Loucks, T.L., Berga, S.L., 2013. Neuroendocrine recovery initiated by cognitive behavioral therapy in women with functional hypothalamic amenorrhea: a randomized, controlled trial. Fertility and Sterility 99, 2084–2091.e1.

Nakamura, K., Sheps, S., Arck, P.C., 2008. Stress and reproductive failure: past notions, present insights and future directions. Journal of Assisted Reproduction and Genetics 25, 47–62.

Paulson, J.F., Bazemore, S.D., 2010. Prenatal and postpartum depression in fathers and its association with maternal depression: a meta-analysis. JAMA 303, 1961–1969.

Reck, C., Stehle, E., Reinig, K., Mundt, C., 2009. Maternity blues as a predictor of DSM-IV depression and anxiety disorders in the first three months postpartum. Journal of Affective Disorders 113, 77–87.

Riss, P., Kargl, J., 2011. Quality of life and urinary incontinence in women. Maturitas 68, 137–142.

Slama, R., Hansen, O.K., Ducot, B., Bohet, A., Sorensen, D., Giorgis Allemand, L., Eijkemans, M.J., Rosetta, L., Thalabard, J.C., Keiding, N., Bouyer, J., 2012. Estimation of the frequency of involuntary infertility on a nation-wide basis. Human Reproduction 27, 1489–1498.

Snowdon, C.G., Ziegler, T.E., 2007. Reproductive hormones. In: Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (Eds.), Handbook of Psychophysiology, third ed. Cambridge University Press, Cambridge, pp. 319–346.

Stanton, A.L., Lobel, M., Sears, S., DeLuca, R.S., 2002. Psychosocial aspects of selected issues in women's reproductive health: current status and future directions. Journal of Consulting and Clinical Psychology 70, 751–770.

Stubbs, M.L., 2008. Cultural perceptions and practices around menarche and adolescent menstruation in the United States. Annals of the New York Academy of Sciences 1135, 58–66.

Thomas, F., Renaud, F., Benefice, E., de Meeüs, T., Guegan, J.F., 2001. International variability of ages at menarche and menopause: patterns and main determinants. Human Biology 73, 271–290.

Tschudin, S., Bertea, P.C., Zemp, E., 2010. Prevalence and predictors of premenstrual syndrome and premenstrual dysphoric disorder in a population-based sample. Archives of Women's Mental Health 13, 485–494.

Veltman-Verhulst, S.M., Boivin, J., Eijkemans, M.J., Fauser, B.J., 2012. Emotional distress is a common risk in women with polycystic ovary syndrome: a systematic review and meta-analysis of 28 studies. Human Reproduction Update 18, 638–651.

Wang, L., Wang, X., Wang, W., Chen, C., Ronnennberg, A.G., Guang, W., Huang, A., Fang, Z., Zang, T., Wang, L., Xu, X., 2004. Stress and dysmenorrhoea: a population based prospective study. Occupational and Environmental Medicine 61, 1021–1026.

Welberg, L.A., Thrivikraman, K.V., Plotsky, P.M., 2005. Chronic maternal stress inhibits the capacity to up-regulate placental 11beta-hydroxysteroid dehydrogenase type 2 activity. Journal of Endocrinology 186, R7–R12.

World Health Organization, 2006. Reproductive Health Indicators for Global Monitoring: Guidelines for Their Generation, Interpretation and Analysis for Global Monitoring. World Health Organization, Geneva.

World Health Organization and the United Nations Populations Fund, 2008. Maternal Mental Health and Child Health and Development in Low and Middle Income Countries. World Health Organization, Geneva.

Wylie, K., 2007. Assessment and management of sexual problems in women. Journal of the Royal Society of Medicine 100, 547–550.

Yildiz, B.O., Bozdag, G., Yapici, Z., Esinler, I., Yarali, H., 2012. Prevalence, phenotype and cardiometabolic risk of polycystic ovary syndrome under different diagnostic criteria. Human Reproduction 27, 3067–3073.

Zaers, S., Waschke, M., Ehlert, U., 2008. Depressive symptoms and symptoms of post-traumatic stress disorder in women after childbirth. Journal of Psychosomatic Obstetrics and Gynecology 29, 61–71.

H

Habermas, Jürgen (1929–)

William Outhwaite, Newcastle University, Tyne and Wear, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The philosopher and sociologist Jürgen Habermas (1929–), remains deeply influenced by his teenage experiences under and immediately after the Nazi regime. As he has said, the theme of democracy and rational communication runs like a red thread through his work. His critical theory of society is in many ways a continuation of the earlier critical theory developed in the Frankfurt Institut für Sozialforschung.

Habermas site: www.habermasforum.dk/.

Jürgen Habermas, born in 1929 and brought up in the small northwest German town of Gummersbach, remains deeply influenced by his teenage experiences under and immediately after the Nazi regime. As he has said, the theme of democracy and communication runs like a red thread through his work. His critical theory of society is in many ways a continuation of the earlier critical theory developed in the Frankfurt Institut für Sozialforschung until 1933, then in exile and again from 1950 in Frankfurt.

Unlike the experience of these 'permanent exiles' (Jay, 1985), Habermas's life has been relatively tranquil. He studied in Göttingen, Zürich, and Bonn, where he was drawn into the first of many interventions as a public intellectual by the philosopher Martin Heidegger's republication without comment, in 1953, of an article of 1935, which included a complimentary reference to the Nazi movement. Invited by Theodor W. Adorno to work as his assistant at the Institut für Sozialforschung, he participated in a survey of student political opinion and began work on his habilitation thesis on the public sphere. Driven out of the Institute by Max Horkheimer, for whom he was too radical, he completed this work at Marburg with the Marxist political scientist Wolfgang Abendroth. After teaching for a time at Heidelberg, he took up Horkheimer's chair in philosophy and sociology at Frankfurt in 1964. An active supporter but increasingly severe critic of the student movement of the late 1960s, he left Frankfurt in 1971 for a 10-year period as codirector of a research institute in Starnberg, where he completed his *Theory of Communicative Action* (1981). After 12 years back in Frankfurt, he retired in 1994 to Starnberg, where he has continued an astonishingly productive program of writing and worldwide lecturing.

Two of the main ways in which Habermas's work differs from the earlier critical theory of Adorno, Horkheimer, and Herbert Marcuse are a more distanced attitude to Marxism and a closer engagement with contemporary sociological research. Philosophy, in Habermas's conception, serves as a 'placeholder and interpreter' for the social sciences: it can raise issues for further analysis by the social sciences and it can generalize and reflect on their results (Habermas, [1981]1983).

Methodology: Reconstructive Science

Habermas's methodological critique of positivism in social science went along with his criticism of technocratic politics in the early 1960s, initially in his contribution to the 'positivism dispute' (Adorno [1969]1976), and then in major works of the late 1960s. Here, Habermas developed an approach that rejected positivism for its reduction of epistemology to methodology and its failure to appreciate the differences between the natural and social sciences. At the same time, he criticized what he called the 'hermeneutic claim to universality' and argued that understanding (Verstehen) needed to be complemented by causal explanation in a model of 'emancipatory' social science exemplified by Freudian psychoanalysis and the (originally Marxist) critique of ideology. Theories of this kind combined the two methods in an attempt to identify causal obstacles to the understanding of personal and social problems, which a purely interpretive approach could not capture. However, as he wrote later in a new preface to this work, the problem of Verstehen remains important because mere observation does not give access to a 'symbolically structured reality' and because "a participant's understanding is not as easy to monitor as an observer's perception" (Habermas, 1982a: 549).

Habermas's critique of positivism went along with a cautious appropriation of elements of the Marxist tradition – the theory of ideology in particular. Critical theory's response to what Habermas, following Horkheimer, calls 'traditional theory' parallels Marx's critique of idealist philosophy (Habermas, [1968] 1986: 212).

In the 1970s, Habermas came to feel that this epistemological approach, a model of knowledge grounded in 'cognitive interests' (in prediction and control, as in positivist conceptions of natural science, in understanding, as in hermeneutics,

and in 'emancipation' from unnecessary and unwanted constraints) was too rigid to capture the diversity of social science. He turned instead to seek an alternative foundation for critical social science in an analysis of human communication and, more broadly, 'communicative action.' When we make statements about states of affairs in the world, about our feelings or about what we consider right, we commit ourselves to offering grounds for these utterances to someone who might disagree. In this sense, our speech acts, if they are more than just orders or attempts to manipulate others, are oriented toward possible agreement. Communicative action in this sense is distinguished from, and prior to, 'strategic action,' as conceptualized in economic and rational choice approaches, 'normatively regulated action,' as theorized by Talcott Parsons and other functionalist sociologists, and 'dramaturgical action' as in Erving Goffman's 'presentation of self in everyday life.' For Habermas, the teleological and strategic, the normative and the dramaturgical model are 'one sided' or 'simplistic': "Only the communicative model of action presupposes language as a medium of uncurtailed communication …" (Habermas, 1987a: 142).

> The concept of *communicative action* refers to the interaction of at least two subjects capable of speech and action who establish interpersonal relations (whether by verbal or by extra-verbal means). The actors seek to reach an understanding about the action situation and their plans of action in order to coordinate their actions by way of agreement.
> (Habermas, 1987a: 86)

The theory of communicative action is not 'a continuation of the theory of knowledge by other means' (Habermas, 1987a: xxxix) but a 'reconstructive science,' which, like linguistics, explicates the rules and principle underlying our everyday speech and action (Habermas, 1987a: 18–19). As he said in an interview,

> I never say that people *want* to act communicatively, but that they *have to*. When parents bring up their children, when the living appropriate the transmitted wisdom of preceding generations, when individuals and groups cooperate, […] they all have to act communicatively. There are elementary social functions that can only be satisfied by means of communicative action.
> (Habermas, 1994: 111)

Habermas presented this model in his major work in 1981, through a detailed discussion of the history of social thought and methodology and a reflection on the history of Western modernity. Very briefly, the Enlightenment meant rational and critical reflection on tradition, but societal rationalization also brought with it the decoupling of economic and administrative systems from rational scrutiny, with markets and state bureaucracies following their own logics. Habermas had addressed these questions in his early book, *Structural Transformation of the Public Sphere* (1962), which examined the emergence in late seventeenth- and eighteenth-century Europe of a critical public, reading newspapers and discussing current issues in coffee houses and salons. This was based on "post-traditional legal and moral representations" (Habermas, 1987b: 166).

Max Weber continually points to the paradoxes of rationalization, such as the irrationality of a work ethic pursued to extremes, whereas Habermas poses the more precise question of an 'imbalanced' rationalization process, in which "the capitalist economy and modern administration expand at the expense of other domains of life that are structurally disposed to moral-practical and expressive forms of rationality and squeeze them into forms of economic or administrative rationality" (Habermas, 1987b: 183).

Here, Habermas refers to Marx and more precisely to Georg Lukács' concept of reification (Verdinglichung), which blends Weber's rationalization with Marx's concept of alienation or estrangement. The theoretical framework, then, is that of Western Marxism, in particular the early critical theory of Horkheimer, Marcuse, Adorno, and others. But where Max Weber, as he sees it, conceives rationalization in cognitive and individualistic terms and Adorno and Horkheimer lapse into an abstract, pessimistic, and ultimately unpromising critique of instrumental reason, Habermas grounds his critique of capitalism in a theory of communication. The specific pathologies of modernity are above all the reduction of the possibilities of communicative action through the 'colonization of the lifeworld.'

The counterfactual question then became whether we could have had the first form of rationalization without the second; the practical question whether we can reestablish democratic control over these processes. He ended the book with a statement of what he saw as the task of a critical theory of society: "explaining those pathologies of modernity that other approaches pass right by for methodological reasons" (Habermas, 1987b: 378). Here he reformulated some of his earlier analyses of crisis tendencies in advanced capitalist societies. There he had argued that crises of capitalist reproduction were increasingly displaced into rationality crises in state policy, legitimation crises, and crises of individual motivation. He restates this theme in the *Theory of Communicative Action*: that the "new conflicts" are located less "in domains of material reproduction" and rather "in domains of cultural reproduction, social integration and socialization"(Habermas, 1987b: 349).

The basic categories of Habermas's theory are, then, those of a broadly conceived sociological theory of action, which, however, also incorporates social historical and system-theoretical, as well as structuralist elements. Although he borrows some concepts from Talcott Parsons (Holmwood, 2009) and often mentions Niklas Luhmann, his conception is closer to those of Pierre Bourdieu and Anthony Giddens (Outhwaite, 2015).

What Habermas (1987b: 553) stresses is that system theories and theories of action 'isolate and overgeneralize' aspects of modernity (system and lifeworld, respectively). When, here and elsewhere, he emphasizes the role of language, he does not intend to reduce "social *action* to the interpretive accomplishments of participants in communication … assimilating action to speech, interaction to conversation" (Habermas, 1987a: 143). This is rather the way in which action is 'coordinated.'

The one-sidedness of the other action models can also be illustrated sociologically. I can act strategically but do not, I think, always do so; I can act thoroughly dramaturgically, for

example, by 'power dressing,' or not bother; I can orient myself to norms or ignore them, like tourists who behave as they would at home. The reference above to mechanisms of action coordination (Habermas, 1987a: 143) shows that Habermas is going beyond a narrowly conceived theory of action. An alternative form of action coordination is through markets and bureaucratic power.

Social Change: Rationalization, Increased Complexity, and the Colonization of the Lifeworld

Habermas borrows the concept of the lifeworld from phenomenological philosophy (Husserl) and its sociological adaptation by Alfred Schütz. But for Habermas it has a rather different meaning: the realm of informal social interaction, the informal coordination of action by language and, more broadly, by communicative (and also strategic) action. Relations between friends and family members are usually regulated informally rather than by monetary or legal transactions.

Habermas welcomes the explicit formulation and reflexive examination of norms in modernity, and sees the coordination of action through money and markets as a necessary increase in complexity that goes along with societal development. These formal media of nonlinguistic interaction, (I can buy a paper from a machine as well as from a friendly newsagent) only become problematic for Habermas when they intrude damagingly into the normative organization of the lifeworld, leading individuals to think of their whole lives in categories appropriate to the work process (Iser, 2009: 329).

Habermas has from time to time distanced himself from this model of colonization, but he would probably still endorse the principle he formulated in relation to the beginnings of the process of social differentiation: money and power "need to be institutionally and motivationally anchored in the lifeworld" (Habermas, [1981]1987b: 342). In 1981, the concept of juridification ('Verrechtlichung') had pejorative connotations, whereas with the development of his theory of law in the following decade it appears in a more positive (or, as some have suggested, uncritical) light.

The Democratic Constitutional State

In its developed form, Habermas's critical theory of modern society is no longer framed as a philosophy of history, as in the early 1960s (*Theory and Practice*), or in epistemological terms (*Knowledge and Human Interests*), but on the basis of a reconstruction of the implicit rules of communicative action. This yields not only a social theory but also a 'discourse ethic' or, more precisely, a discursive foundation for morality, in which to say something is right is to say that all those affected could come to a rational agreement about its rightness. This in turn leads to a theory of law and democracy.

Discourse ethics, based on the same reconstructive principles of universal pragmatics as the theory of communicative action, is also reconstructive in that, rather than prescribing a basic moral principle in abstract terms, it derives it from our daily communicative behavior. "The universalization principle can be understood […] as a reconstruction of the everyday intuitions underlying the impartial judgement of moral conflicts of action" (Habermas, 1990: 127). It "stands or falls with two assumptions: (a) that normative claims to validity have cognitive meaning and can be treated *like* claims to truth and (b) that the justification of norms and commands requires that a real discourse be carried out and therefore cannot occur in a strictly monological form, i.e. in the form of a hypothetical process of argumentation occurring in the individual mind" (Habermas, 1990: 68). As Thomas McCarthy (1978: 326) put it, "The emphasis shifts from what each can will without contradiction to be a universal law to what all can will in agreement to be a universal norm."

Habermas's moral principle, like the Kantian principle to which McCarthy alluded, is a formal one: the content comes from human communities. The basis for Habermas is what he calls the discourse principle: "Just those action norms are valid to which all possibly affected persons could agree as participants in rational discourses" (Habermas, 1996: 107). Habermas draws a distinction between fundamental moral issues and ethical ones, where local conventions play a bigger part. (This is of course somewhat problematic: abortion, for example, may be treated as a moral issue, but abortion legislation was notably more liberal in communist East Germany, where almost all Christians were Protestants, than in West Germany, with a Catholic majority.)

The discourse principle can also form the basis for legal norms, where again ethical and pragmatic considerations play a role. This yields, for Habermas, a very close connection between law and democracy: under modern conditions, legal principles can only be justified by democratic procedures. The last sentence of the 'Tanner Lectures' of 1986, in which Habermas first presented this theory, reads: "No autonomous law without the realization of democracy" (Habermas, 1986: 279). In *Between Facts and Norms*, Habermas (1996: xlii) repeats this claim, that "in a sphere of fully secularized politics the Rechtsstaat cannot be achieved and preserved without radical democracy." This aspect of Habermas's thought might seem a departure from social science toward normative questions of moral and political philosophy, but Habermas ([1992]1996: xl; 5-8) would resist this suggestion, stressing the parallels between legal and social theory.

As he often does, Habermas presents his conception of democracy as a synthesis between two conflicting alternatives: Discourse theory's 'deliberative' model of democracy incorporates elements both of the liberal conception of politics as the mediation of private interests and of the republican conception of a self-organizing ethical community (Habermas, [1992] 1996: 296). Solidarity, a very important concept for Habermas (Dews, 1992), is a driving force of this model of democracy and "… can also hold its own, in the medium of law, against the two other mechanisms of social integration: money and administrative power" (Habermas, [1992]1996: 299).

The critical response to *Between Facts and Norms* has focused on three issues. First, his optimistic conception of law, which has particularly disappointed readers associated with 'critical legal studies.' Second, and relatedly, his scaled-down expectations about the prospects for socialism. In his reflections on the anticommunist revolutions of 1989, Habermas, who had always condemned the state socialist dictatorships, declared (Habermas, [1990]1992) that he no longer saw a state-

controlled economy as realistic and that he now understood socialism as 'radical democracy' (Scheuerman, 1999).

The third question concerns the way Habermas framed his account of the constitutional state in relation to the national state and failed to address the implications of globalization, which had become a major focus of attention since early in the 1990s. David Held, Mary Kaldor, and others were already developing conceptions of 'cosmopolitan democracy,' which Habermas was quick to embrace.

In a postscript to his interview volume *The Past as Future* (Habermas, [1991]1994), he referred to Paul Kennedy's criticism of the national state: "For some problems, it is too large to operate effectively; for others, it is too small" (Kennedy, 1993: 131).

By 1996, he was asking 'Has the nation-state a future?'; and in *The Postnational Constellation*, he poses the question (Habermas, 2001: 90) "whether political communities can form a *collective identity* beyond national borders, and thus whether they can meet the legitimacy conditions for a postnational democracy."

Since then, Habermas has been centrally concerned with the prospects of the European Union, speaking eloquently for federalism, though now with increasing pessimism. "Executive federalism" threatens to develop into "intergovernmental rule by the European Council," in which "the first transnational democracy would be transformed into an arrangement for exercising a kind of post-democratic, bureaucratic rule. The alternative is to continue the democratic juridification [Verrechtlichung] of the European Union in a consistent way" (Habermas, [2011]2012: 52–53).

Habermas's conception of 'postnational' democracy includes a recognition of the persistence of 'still nationally shaped life projects'; the post-, preface indicates persistence as well as transformation. In the EU context, this means that people "must learn to distinguish between the role of a member of a 'European people' and that of a 'citizen of the Union'" (Habermas, [2011]2012: 29), with this idea of a double mandate as a German or Briton and as a European serving as a guide to further constitutional innovations.

Another area of Habermas's recent work is shaped around what he admits is the controversial concept of a postsecular society (Habermas, 2001a; 2001b; [2005]2008; Habermas and Ratzinger, 2005). Habermas does not mean by this that the Enlightenment critique of religion has run out of steam and that religion is back after all. It is rather that modernization has led to a secularized and ideologically plural society in which religions have to 'reflexively' confront other belief systems (Habermas, 2001a: 176f.). Since 2001, Habermas has addressed questions of religious fundamentalism and returned to the theme of multiculturalism that he had analyzed in *The Inclusion of the Other* (1996), notably in essays on religious tolerance in *Between Naturalism and Religion*; see also Joas, 2005; Anderson, 2007. (His discussion with Cardinal Ratzinger attracted additional publicity when Ratzinger became Pope; Habermas has stressed that he has not himself become religious.)

The term naturalism refers to a further theme of Habermas's recent work (Habermas, 2003): the implications of new biomedical techniques (genetic engineering, cloning, etc.) for what he calls 'liberal eugenics' and more broadly for our ideas of human identity – issues that had of course also attracted the attention of religious leaders (Habermas (2001b: 9; 2005: 7)).

All this raises again the question whether Habermas should be seen more as a philosopher than a sociologist, as his old friend Ralf Dahrendorf suggested. It is probably more useful to see him mediating between philosophy and sociology, just as, in philosophical terms, he moves between Kantian and Hegelian impulses, notably in his ethical philosophy, and in political terms between liberalism, especially in relation to human rights, and socialism.

Critiques and Continuations

One of the most striking features of Habermas's work is his engagement with a vast range of specialized fields, so it is notable that he had something of a blind spot in relation to postmodern theory and to major contemporary figures such as Michel Foucault, Jean-François Lyotard, and Jacques Derrida; also to Pierre Bourdieu and perhaps even Niklas Luhmann, despite his substantial exchange with him published in 1971. Others have pointed to affinities between these approaches and Habermas's.

A more serious line of criticism suggests that Habermas's theories are structurally blind to certain perspectives. Feminist critics, for example (Nancy Fraser, Seyla Benhabib, Amy Allen, Diana Coole), have pointed to patriarchal assumptions behind his models of the public sphere and politics, and to the fact that modernity may even have reinforced the exploitation and exclusion of women.

In other areas of social theory, many critics have suggested that he takes too much from Talcott Parsons, notably his system theory and his stress on norms. Anthony Giddens pointed to this: "it is an error to suppose that communicative action can be examined solely on the level of norms" (Giddens, 1982a: 157–158; see also Giddens, 1982b). Hans Joas (1991) referred more broadly to an 'unhappy marriage' of hermeneutics and functionalism and Thomas McCarthy (1991) to the 'seducements of system theory.' More recently, John Holmwood has criticized his acceptance of Parsons' assumption "that a general theoretical framework is a necessary precondition for social enquiry and that it can be grounded in terms of a generalized problem of social order" (Holmwood, 2009: 70).

Other critics argue that Habermas has abandoned too much of Marx, again ascribing too much independence to moral and legal principles that Marx saw as ideological. What is clear is that from an early stage in his work, Habermas, partly influenced by Max Weber and early critical theory and partly by system theory, developed an analysis of societal modernization which was quite far from orthodox Marxism (Habermas, 1971: 98).

On the other hand, Habermas stressed in an interview in 1984 that

> … from the outset my theoretical interests have been consistently determined by those philosophical and socio-theoretical problems which arise out of the movement of thought from Kant through to Marx. My intentions were given their stamp by Western Marxism in the mid-fifties, through a coming-to-terms with Lukács, Korsch and Bloch, Sartre and Merleau-Ponty, and of course with

Horkheimer, Adorno and Marcuse. Everything else which I have made my own has only acquired its significance in connection with the project of a renewal of the theory of society grounded in this tradition.

Habermas is not offering an alternative to such a broad theory as historical materialism, but rather a theoretical framework that can be applied (as he does himself in outline) to historical processes such as the development of Western modernity. In other words, the theory of communicative action can only be part of a critical theory of society. What Habermas formulates is a logic of development from processes of mutual understanding in the lifeworld all the way to the democratic constitutional state, but the forms of these developments and countermovements require further exploration.

Habermas's concept of modernity is of course contentious, as Anthony Giddens (1982b: 335) noted in an early review. Since then, postcolonial critics have fundamentally problematized the concept of modernity (Bhambra, 2007; G.R. et al., 2010; see also Wagner, 2012). Habermas, like Max Weber, may seem too fixated on the 'European miracle,' but Klaus Eder (2009: 66–68) has defended his evolutionary conception of modernity as a process of communicative sociation and a universalistic culture that happened to emerge at particular times and places. Postcolonial critics might still want to suggest that Habermas and others overemphasize the endogenous character of European and North American modernity.

A Habermasian approach has been carried forward by what can be called a third generation of critical theorists, notably Axel Honneth and Seyla Benhabib. Honneth, who took up Habermas' chair in Frankfurt on his retirement in 1996 and followed Helmut Dubiel as Director of the Institute for Social Research, has developed a theory focusing on concrete struggles for recognition driven not just by exploitation but by broader experiences of 'disrespect' and based on "empirically effective forms of morality to which critical theory can connect" (Honneth, 2000: 110). This feeds into conceptions of social pathology and a revival of the notion of 'diagnoses of the times.'

Honneth is in a sense returning to an earlier focus of critical theory on social movements. In a similar way, Seyla Benhabib, Nancy Fraser, and others in the United States have developed the links between critical theory and feminist theory. Benhabib (2004, 2006) has more recently turned in particular to questions of migration, ethnicity, and citizenship in Europe, arguing that structures developed to address cultural pluralism and minority rights are, "quite compatible with a universalist deliberative democracy model" (Benhabib, 2002:19). The development of theories of deliberative democracy by John Dryzek and others is a further important continuation of Habermasian theory.

See also: Action, Theories of Social; Capitalism; Critical Theory; Democracy, History of; Frankfurt School: Institute for Social Research; Logical Positivism and Logical Empiricism; Marxism in Contemporary Sociology; Modernity; Nationalism, Sociology of; Positivism, Sociological; Sociological Theory.

Bibliography

Adorno, T.W., Albert, H., Dahrendorf, R., Habermas, J., Pilot, H., Popper, K.R., 1976 [1969]. The Positivist Dispute in German Sociology. Heinemann, London.

Anderson, J., 2007. Introduction. Free will, neuroscience, and the participant perspective. Philosophical Explorations 10 (1), 3–11.

Benhabib, S., 1980. Critique, Norm and Utopia. Columbia University Press, New York.

Benhabib, S., 2002. The Claims of Culture. Equality and Diversity in the Global Era. Princeton University Press, Princeton.

Benhabib, S., 2004. The Rights of Others. Cambridge University Press, Cambridge.

Benhabib, S., 2006. Another Cosmopolitanism. Oxford University Press, New York.

Bhambra, G.K., 2007. Rethinking Modernity. Postcolonialism and the Sociological Imagination. Palgrave Macmillan, Basingstoke.

Dews, P., 1992. Jürgen Habermas: Autonomy and Solidarity, second ed. Verso, London.

Eder, K., 2009. Evolutionstheorie. In: Brunkhorst, H., Kreide, R., Lafont, C. (Eds.), Habermas-Handbuch. Metzler, Stuttgart and Weimar, pp. 65–68.

Giddens, A., 1982a. Labour and interaction. In: Thompson, J.B., Held, D. (Eds.), Habermas: Critical Debates. Macmillan, London.

Giddens, A., 1982b. Reason without revolution? Habermas' Theorie des kommunikativen Handelns. Praxis International 3, 318–338.

G.R., E., Boatca, M., Costa, C. (Eds.), 2010. Decolonizing European Sociology: Transdisciplinary Approaches. Ashgate, Farnham and Burlington.

Habermas, J., 1989[1962]. Structural Transformation of the Public Sphere. Polity, Cambridge.

Habermas, J., 1986[1963]. Theory and Practice. Polity, Cambridge.

Habermas, J., 1992. Final remarks. In: Calhoun, Craig (Ed.), Habermas and the Public Sphere. MIT Press, Cambridge, MA.

Habermas, J., 1971. Toward a Rational Society. Heinemann, London.

Habermas, J., 1986. Law and Morality. The Tanner Lectures on Human Values. http://tannerlectures.utah.edu/_documents/a-to-z/h/habermas88.pdf.

Habermas, J., 1986[1968]. Knowledge and Human Interests. Polity, Cambridge.

Habermas, J., 1976[1973]. Legitimation Crisis. Heinemann, London.

Habermas, J., 1987a[1981]. Theory of Communicative Action, vol. 1. Polity, Cambridge.

Habermas, J., 1987b[1981]. Theory of Communicative Action, vol. 2. Polity, Cambridge.

Habermas, J., 1983[1981]. Philosophical-Political Profiles. Heinemann, London.

Habermas, J., 1982a. Objektivismus in den Sozialwissenschaften. In: Habermas (Ed.), Zur Logik der Sozialwissenschaften, fifth ed. Suhrkamp, Frankfurt.

Habermas, J., 1982b. Reply to my critics. In: Thompson, J.B., Held, D. (Eds.), Habermas: Critical Debates. Macmillan, London, pp. 219–283.

Habermas, J., 1983[1981]. Moral Consciousness and Communicative Action. Polity, Cambridge.

Habermas, J., 1990. Die Nachholende Revolution. Suhrkamp, Frankfurt.

Habermas, J., 1992[1990]. Further reflections on the public sphere. In: Calhoun, C. (Ed.), Habermas and the Public Sphere. MIT Press, Cambridge, MA., pp. 421–461.

Habermas, J., 1996[1992]. Between Facts and Norms. Polity, Cambridge.

Habermas, J., 1994[1991]. In: Heller, Michael (Ed.), The Past as Future. University of Nebraska Press, Lincoln.

Habermas, J., 1998[1996]. The Inclusion of the Other. Polity, Cambridge.

Habermas, J., 2001[1998]. The Postnational Constellation. Polity, Cambridge.

Habermas, J., 2001a. Der demokratische Rechtsstaat – eine paradoxe Verbindung widersprüchlicher Prinzipien? In: Habermas (Ed.), Zeit der Übergänge. Suhrkamp, Frankfurt, pp. 133–151.

Habermas, J., 2001b. Glauben und Wissen. Suhrkamp, Frankfurt.

Habermas, J., 2003[2001]. The Future of Human Nature. Polity, Cambridge.

Habermas, J., 2005[2008]. Zwischen Naturalismus und Religion. Frankfurt, Suhrkamp (Between Naturalism and Religion, 2008, Trans). Polity, Cambridge.

Habermas, J., 2012[2011]. The Crisis of the European Union. A Response. Polity, Cambridge.

Habermas, J., Ratzinger, J., 2005. Dialektik der Säkularisierung. Űber Vernunft und Religion. Herder, Freiburg.

Holmwood, J., 2009. From 1968 to 1951: how Habermas transformed Marx into Parsons. In: Bhambra, G.K., Demir, I. (Eds.), 1968 in Retrospect: History, Theory, Alterity. Palgrave, Basingstoke, pp. 59–72.

Honnet, A., Joas, H. (Eds.), 1991[1988]. Communicative Action. Polity, Cambridge.

Honneth, A., 2007[2000]. Disrespect. The Normative Foundations of Critical Theory. Polity, Cambridge.

Iser, M., 2009. Kolonialisierung. In: Brunkhorst, H., Kreide, R., Lafont, C. (Eds.), Habermas-Handbuch. Metzler, Stuttgart and Weimar, pp. 328–330.

Jay, M., 1985. Permanent Exiles. Essays on the Intellectual Migration from Germany to America. Columbia University Press, New York.

Joas, H., 1991. The unhappy marriage of hermeneutics and functionalism. In: Honneth, A., Joas, H. (Eds.), Communicative Action. Polity, Cambridge, pp. 97–118.
Joas, H., 2005. Die Religion der Moderne. Die Zeit 42, p. 94.
Kennedy, P., 1993. Preparing for the Twenty-First Century. Vintage, New York.
McCarthy, T., 1978. The Critical Theory of Jürgen Habermas. Polity, Cambridge.
McCarthy, T., 1991. Complexity and democracy: or the seducements of systems theory. In: Honneth, A., Joas, H. (Eds.), Communicative Action. Polity, Cambridge, pp. 119–139.
Outhwaite, W., 1994. Habermas. A Critical Introduction, second ed. 2009. Polity, Cambridge.
Outhwaite, W., 2015. Sociological theory. In: Holborn, M. (Ed.), Contemporary Sociology. Polity, Cambridge.
Rapic, S. (Ed.), 2013. Habermas und der historische Materialismus. Alber, Freiburg.
Scheuerman, W., 1999. Between radicalism and resignation: democratic theory in Habermas's between facts and norms. In: Dews, P. (Ed.), Habermas. A Critical Reader. Blackwell, Oxford, pp. 153–177.
Specter, M., 2010. Habermas. An Intellectual Biography. Cambridge University Press, New York.
Wagner, P., 2012. Modernity. Understanding the Present. Polity, Cambridge.

Habit: History of the Concept

Charles Camic, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines aspects of the history of the concept of habit in Western social thought. As an expression, the concept of habit has generally referred to a disposition to act as one has previously acted, whether in regard to simple behaviors, more complex forms of conduct, or broader characterological tendencies (habitus). Among social thinkers who have used the concept, habit has generally been applied to recurrent forms of moral, economic, political, and religious conduct insofar as these occur more or less automatically, thus differing from reflective forms of human action that entail the deliberative selection of means and ends by normative standards. While the history of the concept of habit has been roughly coextensive with the entirety of Western intellectual history, four phases are distinguished. The first phase, running from antiquity to the early 1800s, was marked by the frequent discussion of the role and significance of habit by thinkers ranging from Aristotle to the philosophers of the Enlightenment. The second phase, extending from the mid-nineteenth to the early twentieth century, was characterized by two contending developments. One of these was the invocation of the concept of habit by a great many European and American social theorists, including Emile Durkheim and Max Weber, and the use of the concept, in tandem with models of reflective action, to analyze human conduct in the social world. The second development in this period was the effort on the part of natural scientists to restrict habit to more elementary human and subhuman behaviors. The third phase, beginning roughly in the late 1910s, saw the successful appropriation of habit, narrowly conceived, by behaviorist psychologists and, in a strong countermovement, the abandonment of the concept by a majority of European and American social thinkers, who for the next half century conceptualized all human conduct exclusively in reflective terms. The fourth phase, extending from the early 1980s to the present time, has witnessed steps by social theorists and empirical researchers to revive and elaborate the concept of habit and to examine the relationship between habitual and reflective forms of human action.

Found in ordinary languages the world over, 'habit' and various cognate expressions have functioned, for over two millennia, as concepts in diverse intellectual traditions in the social sciences, the humanities, and the natural sciences. As used in reference to various recurrent and relatively automatic forms of human conduct in the social world, "the discourse of habit is one of the great, albeit largely unacknowledged, cultural legacies of Western civilization" (Thomas, 1993: p. 8). Although the history of the discourse has been roughly coextensive with Western intellectual history itself, specific use of the concept, and explicit recognition of the aspects of human action to which it refers, have waxed and waned during different periods in the development of the social sciences.

The Concept of Habit

In the course of its lengthy history, habit has led multiple lives. Its origins lie in the unreflective, 'indiscriminate, quotidian use of habit' in various popular discourses, which are 'littered with offhand references' to habit (or to what is habitual, or done habitually) and which have historically furnished a vocabulary that has been imported into more rarified discussions of philosophical and scientific subjects. Students of biological life and the physical world have often used habit with reference to "non-human animals, the dynamics of populations, the growth pattern of plants, and the tendencies of systems." (For these quotations, see Sparrow and Hutchison, 2013: p. 3.) Applied to human conduct, 'habit' has generally designated a tendency or disposition for a person to act in ways that she or he has acted previously in similar contexts, insofar as that disposition is "activated automatically by the context cues that co-occurred" with the relevant action in the previous situations (quoting Neal et al., 2006: p. 198; see also Pollard, 2003: p. 411).

When these repeated actions are differentiated from one another in terms of their content, distinctions then arise between moral, cognitive, emotional, and motor habits. Of greater historical importance, however, has been another basis of differentiation. The actions whose repetition is designated as habit may range from simple and circumscribed to complex and generalized. Among the social thinkers who have used the concept, habit has most frequently been applied to recurrent patterns of action of moderate complexity and generality, i.e., to habits of economic, political, religious, and domestic conduct; habits of obedience to rules and to rulers; habits of sacrifice, disinterestedness, and restraint; etc. At various moments in the history of the concept, however, habit has taken on meanings both narrower and broader: sometimes, it has referred to simpler, or at least more elemental, behavioral regularities, such as habits of talking, walking, eating, dressing, routine problem-solving, etc.; at other times, it has been applied to the generalized and durable mode of action that a person may exhibit in one or more wider spheres of social life, such as work, politics, and interpersonal relations (or in all of these). When used in this more extended sense, the concept of habit has been largely synonymous with 'character' or 'habitus' – the latter is a term that traces back historically to the same linguistic roots as habit. The fact that habit has borne this entire range of connotations serves as a precaution against conflating particular notions of habit with habit as such. The assumption, prevalent in the middle decades of the twentieth century, that habit is a fixed, mechanical reaction to discrete stimuli and thus devoid of subjective meaning is, for example,

a view that has historically been associated mainly with habit in the sense of simple behavioral regularities, not with more complex habitual forms of action.

At no point has the vocabulary of habit been an autonomous discourse. To a significant degree, its usage has shifted over time because of the different ways in which habit has been combined with, or distinguished from, other concepts. At one important period in its history, habit functioned as the antithesis of instinct, such that all noninstinctual behavior was regarded as habitual. Prior to and subsequent to this period, however, habit was ordinarily set over against purposive, reflective types of action characterized by the deliberate selection of means and ends according to various cognitive and moral standards; in this case, habitual action served as the converse of those types of action foregrounded in reflective models of human conduct, such as those favored by many early- and mid-twentieth-century social scientists. At still other (albeit briefer) moments, however, habit has been taken not as the opposite of, but as the precondition for, reflexive action. The same flexibility of association has appeared with regard to moral conduct, which has sometimes been contrasted with habitual action, other times viewed as dependent upon it. Conceptual variability has been apparent as well at the collective level, where the common habits present throughout a social group have sometime been equated with 'custom,' whereas other usages define custom in terms of practices that result from adherence to established social prescriptions and are, therefore, distinguishable from actions that have become automatic owing to repetition or other habit-formation processes. (For a fuller treatment of the issues in this section, see Camic, 1986; Funke, 1958; Sparrow and Hutchison, 2013.)

Historical Development

Antiquity to the Early 1800s

The everyday discourse of habit was appropriated and rendered esoteric by Greek philosophers as they opened pathways for later social thinkers. In the *Nicomachean Ethics*, Aristotle laid it down that "none of the moral virtues arise in us by nature," that "we are adapted by nature to receive them and are made perfect by habit" (Book 2, Chapter 1) (Aristotle, 1973; for discussion, see Lockwood, 2013). The same conceptual vocabulary was incorporated into the philosophical and social teachings of Plato, Plotinus, Quintilian, Aquinas, William of Ockham, Luther, Montaigne, Pascal, Locke, and Wolff, among a great number of other philosophers, theologians, and men of letters; at the same time that generations of natural philosophers applied the concept in studies of physical and biological phenomena (Funke, 1958; Sparrow and Hutchison, 2013).

In the Age of the Enlightenment, as part and parcel of the emerging movement to create a 'science of man,' the concept of habit received renewed attention from a range of social thinkers who shared the belief that habit is a widespread and potent phenomenon in social life, even as they characterized this phenomenon differently. In France, Claude Adrien Helvetius (1715–71) celebrated habit as the wellspring of public and private morality and accepted it as a "principle by which [humans everywhere] are actuated" (Helvetius, [1758]1807: p. 57); in Scotland, Adam Ferguson (1723–1816) declared habit as the force "by which the good or bad actions of men remain with them and become part of their characters" (quoting Hill, 1996: p. 210); while from Prussia Immanuel Kant (1724–1804) held that "all acquired habits are objectionable" – that "virtue is moral strength in the pursuit of one's duty, which should never be a matter of habit" (Kant, [1798] 1978: pp. 32–34). David Hume, Thomas Reid, George Berkeley, Etienne Condillac, Jean-Jacques Rousseau, Johann Fichte, Friedrich Schelling, and G.W.F. Hegel were among the many thinkers who took part in this wide-ranging discussion (on the thinkers named here and others, see Camic, 1986; Funke, 1958; Sparrow and Hutchison, 2013).

An Age of Ferment

From the middle of the nineteenth century until the early decades of the twentieth century, the language of habit was an even more prominent intellectual presence. This was true in European as well as in American philosophical, natural scientific, and social scientific thought, sociological thought in particular.

In England, the concept of habit entered, with varying degrees of explicitness, into the work of a diverse range of thinkers that included Jeremy Bentham, James and John Stuart Mill, Herbert Spencer, Charles Darwin, Walter Bagehot, and Alfred Marshall; in France, into the writings of Jean-Baptiste Lamarck, August Comte, Alexis de Tocqueville, Frederic LePlay, and Henri Bergson; and in Germany, into the work of Friedrich Nietzsche, Ferdinand Tönnies, Georg Simmel, and Edmund Husserl, among many others. In the United States, the vocabulary of habit, which had long played significant roles in the teachings of popular social reformers, churchmen, and educators, was mobilized in new ways by the founders of philosophical pragmatism, C.S. Peirce, William James, John Dewey, and George Herbert Mead (see Shook and Margolis, 2009), by Darwinian natural scientists like Charles Whitman and Jacques Loeb, and by the first generation of professional, university-based social scientists. Among the latter were leading economists such as Francis Walker, Laurence Laughlin, and Thorstein Veblen, as well as all of the period's major sociologists, including Lester Ward, William Graham Sumner, Franklin Giddings, Albion Small, and W.I. Thomas.

In this respect, the doctrines of contemporary American sociologists accorded with those of the two towering European sociological thinkers of the same era, Emile Durkheim (1858–1917) and Max Weber (1864–1920). Durkheim's lifelong position, for example, was that "habits…are the real forces which govern us" (Durkheim, [1905–06]1956: p. 152). For him, as for the American pragmatists, human reflection is a process that generally "overloads or paralyzes action"; only "when habit is disrupted," and the individual or collective "being is … at a cross-roads situation [facing] a whole range of possible solutions," does reflection emerge in order to restore "equilibrium" and allow "habits of all kinds" to reassert themselves (Durkheim, [1913–14]1983: pp. 38, 79–80). Given this understanding, Durkheim constructed a program for the moral regeneration of modern society that relied heavily on the force of habit: on developing in children "the habit of self control and restraint" and "habits of group life"; on cultivating in occupational specialists "an absolute regularity in habits"

and "habits [of social interaction which], as they grow in strength, are transformed into rules of conduct"; and on instilling in society's leaders "a habitus of moral being" (Durkheim, 1902–03[1961]: pp. 149, 249, 1893[1984]: pp. 187, 302, 1904–05[1977]: p. 29).

Although engaged with very different intellectual questions, Max Weber voiced similar views. Concerned with the analysis of various forms of reflective action, Weber nevertheless insisted on the importance of habit, holding that "the further we go back in history, the more we find that conduct, and particularly social action, is determined in an ever more comprehensive sphere exclusively by the disposition toward the purely habitual"; and that "individuals are still markedly influenced" by this force "even today," when "the great bulk of all everyday action [continues to occur in the form of an] almost automatic reaction to habitual stimuli which guide behavior in a course which has been repeatedly followed" (Weber, [1922]1978: pp. 320, 337, 25 (modified translation)). Understood in this way, the concept of habit figured into Weber's treatment of economic, political, and religious traditionalism, into his writings on economic, political, legal, and communal action in the modern world, and even into his analysis of the process by which ascetic Protestantism promoted the rational spirit of modern capitalism. Regarding "the spirit of capitalism" as a "particular habitus," Weber credited the early Calvinists with creating, "out of their religiously conditioned family traditions and from the religiously influenced life-style of the environment, [a distinctive] 'habitus' among individuals which prepared them in specific ways to live up to the specific demands of early modern capitalism" (Weber, [1910]1968, [1910]1978: p. 1124).

As Weber, Durkheim, and their sociological contemporaries used the concept of habit to refer to forms of action of moderate as well as broader generality, however, natural scientists of the period increasingly associated habit with recurrent human and subhuman activities of much simpler and more mechanical kinds. This, for instance, was the practice of Jean-Baptiste Lamarck (1744–1829), and Charles Darwin (1809–82) following the example, regularly spoke of the flowering habits of plants, the feeding habits of insects, and the flying habits of pigeons (Darwin, [1859] 1964: pp. 11, 183; [1872]1975: pp. 29–31). This was also the practice of many European physiologists and psychologists as they turned attention to the study of the neural foundations of simple animal and human reflexes to external stimuli, invoking habit to describe neural pathways that were noninstinctual yet well established (see Fearing, 1930). This circumscribed usage gained wide currency among late-nineteenth and early-twentieth-century American psychologists, who shared with William James (1842–1910) the view that even "the most complex habits [are] nothing but concatenated discharges in the nerve-centres, due to the presence there of systems of reflex paths" (James, [1890]1950: p. 108). Taking up this same formulation and substituting it for various broader sociological usages, early twentieth-century American social scientists came to define the concept of habit in terms of "the physiology of the nervous system," a formulation that led them to argue that "education and social control are largely dependent upon our ability to establish habits" – habits thus conceptualized – "in ourselves and others" (quoting Park, [1915]1969; see also Camic, 1986).

The Demise of Habit

This important shift in usage was part of a broader social organizational change that occurred as the subject of habit became the object of ownership battles among university-based scholars situated in specialized – and competing – academic disciplines. Indeed, as members of different disciplines engaged in turf wars to differentiate their emerging academic fields from one another, habit assumed the role of a conceptual pawn, a role unknown to it in earlier pre-professional writings where proprietary rights to the concept were not yet in dispute.

The watershed in this development was the behaviorist movement, which engulfed American psychology in the early and middle decades of the twentieth century. Launched by John B. Watson (1878–1958) in an effort to invest the discipline of psychology with the prestige that then inhered in the application of concepts drawn from the natural sciences, behaviorism claimed the realm of habit as its core topic of investigation, asserting that *every* "mode of acting … not belonging to man's hereditary equipment must be looked upon as a habit" – habit understood in physiological terms as "muscular and glandular changes which follow upon a given stimulus." For Watson and his disciples, the corollary of this argument was that a mechanistic habit-psychology provided the way "to guide society…toward the control of group and individual behavior" (Watson, 1919: pp. 270, 14, 2–3) – from which corollary it followed that the academic discipline required for the genuine scientific study of the social world was "not [a social science like] sociology, but psychology," as it derives from "biology, chemistry, and the other natural sciences" (Allport, 1927: pp. 167–168).

Because these sweeping ownership claims directly challenged the bases of their own academic disciplines, social scientists (as well as humanists) from the period almost immediately recoiled from the language of habit, ceasing their use of a concept that had been central to their analysis of human conduct since antiquity. Eager to stop the behaviorist campaign to eliminate, as a valid object of scientific investigation, the reflective forms of action that had previously commanded recognition alongside habitual forms, sociologists of the era undertook an aggressive countercampaign to rescue purposive human conduct from assault. To accomplish this, they deliberately abandoned their prior use of the discourse of habit, thereby abruptly purging habitual phenomena from their conceptualization of human conduct in the social world. This purge was inaugurated by (a one-time proponent of the concept of habit) W.I. Thomas (1863–1947) and Florian Znaniecki (1882–1958), whose work attacked the "behavioristic school" for its "indistinct [application] of the term 'habit' to [all] uniformities of behavior" and insisted that 'habit' … be restricted to the biological field, [since] it involves no conscious, purposeful regulation of [conduct], but merely … is unreflective, [whereas] the uniformity [characteristic of social life] is not a uniformity of organic

habits but of consciously followed rules (Thomas and Znaniecki, [1918]1958: pp. 1849–1852).

Within a generation, this formulation became social scientific dogma, eliminating all sociological talk of habit and replacing it with the language of 'attitude,' which Thomas and Znaniecki had simultaneously ushered onto center stage and which subsequently furnished American social scientists with a new vocabulary for describing recurrent tendencies in human conduct but without sacrificing reflective action models as the behaviorists had done (Camic, 1986; Fleming, 1967). Analogous movements away from the concept of habit occurred in European social thought, as Durkheim and Weber themselves subtracted the concept from their work; and they occurred as well as in European and American humanistic literatures (Fisher, 1973; Thomas, 1993).

Following this conceptual sea-change, the discourse of habit lay dormant in the social sciences from the late 1910s to the early 1980s, a few isolated exceptions notwithstanding. Shortly after the start of the behaviorist revolution in psychology, John Dewey (1859–1952) insistently harked back to a broader conception of habit (1922), as Norbert Elias, Martin Heidegger, Jean Piaget, and Maurice Merleau-Ponty would subsequently attempt to do in Europe (Kestenbaum, 1977; Sparrow and Hutchison, 2013; and, for Elias, [1939]1994; see also Berger and Luckmann, 1966). But these philosophically oriented interventions failed to stem the behaviorist tide or the social-scientific rejection of the concept of habit that had emerged in reaction. Only among psychologists working in the behaviorist tradition, and among a few economists concerned with consumer behavior, did habitual phenomena remain the focus of empirical research during the long mid-decades of the twentieth century (Deaton and Muelbaur, 1980; Smith, 1986). Otherwise, social scientists simply dispensed with the concept, opting to conceive all human action as though it were a reflective process of selecting means to ends in accord with normative standards.

Recent Developments

The period from the 1980s to the 2010s has brought the onset of another reversal, however. Behind this recent development has been a change in the social-organizational conditions that led to the demise of social-scientific interest in habit during the previous historical era. By the closing years of the twentieth century, the intellectual turf warfare among disciplines that had marked the early to middle decades of that century was long ended, and its waning rendered increasingly anachronistic the behaviorists' proprietary claims on habit. The obsolescence of these claims, and of the counterclaims that they had called forth, freed up the concept of habit to reassume some of the broader connotations it once possessed and to reenter, with scarcely any resistance, the social-scientific analysis of human conduct.

To be sure, this revival of the language of habit has thus far been spotty and gradual, and it has left untouched many areas of the social sciences where the language might arguably be put to use. Even so, references to concept of habit and a renewed appreciation of the significance of the phenomenon have begun to reappear throughout the social sciences and the humanities, notably in the work of contemporary philosophers (Ostrow, 1990; Pollard, 2003; Shook and Margolis, 2009), social psychologists (Charng et al., 1988; Neal et al., 2006; Ouellette and Wood, 1998), and economists (Campbell and Cochrane, 1995; Carroll et al., 2000; Dynan, 2000; Fuhrer, 2000; Kahneman, 2003; Li, 2005; Shachar, 2003).

Moreover, just as they had previously spearheaded the assault on the discourse of habit, so sociologists have been particularly active in the current resurrection of the discourse. To varying degrees, this about-face has been evident in the writings of sociological theorists of several different persuasions, all of them of the view that habitual forms of action warrant inclusion along with reflective forms (Archer, 2012; Calhoun et al., 2012; Camic, 1986; Emirbayer and Mische, 1998; Giddens, 2009; Gronow, 2011; Joas, [1992] 1996; Schatzki et al., 2001; Martin, 2011; Sewell, 2005; Turner, 2010). Both explicitly and implicitly, this same view has come to inform research in the specialty areas of economic sociology (Biggart and Beamish, 2003; White, 2004), the sociology of organizations (Adler, 2010), the sociology of science (Collins, 2012), the sociology of morality (Hitlin and Vaisey, 2012), and the sociology of the body (Shilling, 2008). In a further development, sociologist Pierre Bourdieu (1930–2002) drawing on a previously neglected strand in the work of Weber and Durkheim, has resuscitated and elaborated the notion of habitus ([1980] 1990), setting it in place as one of the foundation stones of his oeuvre. Among scholars influenced by Bourdieu's writings, habitus has also crystallized as a major locus of theoretical inquiry and empirical investigation (Gorski, 2013; Hillier and Rooksby, 2005).

Although these numerous lines of recent work remain generally disconnected from one another after the first decade of the twenty-first century, their simultaneous emergence and likely future cross-pollination augur the return of the concept of habit to a position of centrality in social-scientific theory and research, and, accompanying this development, the transformation of exclusively reflective models of action into analytical approaches concerned with the complex and changing relationship between reflective and habitual forms of human conduct in the social world.

See also: Attitude: History of Concept; Conceptual History: Begriffsgeschichte; Discipline Formation in the Social Sciences; Enlightenment: Impact on the Social Sciences; Habitus: History of a Concept; Scientific Disciplines, History of; Sociology, History of; Weberian Social Thought, History of.

Bibliography

Adler, P.S., 2010. The Oxford Handbook of Sociology and Organization Studies: Classical Foundations. Oxford University Press, Oxford, UK.

Allport, F.H., 1927. The nature of institutions. Social Forces 6, 167–179.

Archer, M.S., 2012. The Reflexive Imperative in Late Modernity. Cambridge University Press, Cambridge, UK.

Aristotle, 1973. In: McKeon, R. (Ed.), Introduction to Aristotle, second ed. University of Chicago Press, Chicago.

Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality. Doubleday, New York.
Biggart, N.W., Beamish, T.D., 2003. The economic sociology of conventions: habit, custom, practice, and routine in market order. Annual Review of Sociology 29, 443–464.
Bourdieu, P., [1980]1990. The Logic of Practice. Stanford University Press, Stanford, CA.
Calhoun, C., Gerteis, J., Moody, J., Pfaff, S., Virk, I., 2012. Classical Sociological Theory. Wiley-Blackwell, New York.
Camic, C., 1986. The matter of habit. American Journal of Sociology 91, 1039–1087.
Campbell, J.Y., Cochrane, J.H., 1995. By Force of Habit. National Bureau of Economic Research, Cambridge, MA.
Carroll, C., Overland, J., Neal, D.N., 2000. Saving and growth with habit formation. American Economic Review 90, 341–355.
Charng, H.-W., Piliavin, J., Callero, P.L., 1988. Role identity and reasoned action in the prediction of repeated behavior. Social Psychology Quarterly 51, 303–317.
Collins, H., 2012. Tacit and Explicit Knowledge. University of Chicago Press, Chicago.
Darwin, C., [1859]1964. On the Origin of Species. Harvard University Press, Cambridge, MA.
Darwin, C., [1872]1975. The Expression of Emotions in Man and Animals. University of Chicago Press, Chicago.
Deaton, A., Muelbaur, J., 1980. Economics and Consumer Behavior. Cambridge University Press, New York.
Dewey, J., 1922. Human Nature and Conduct. Henry Holt, New York.
Durkheim, E., [1893]1984. The Division of Labor in Society. Free Press, New York.
Durkheim, E., [1902–03]1961. Moral Education. Free Press, New York.
Durkheim, E., [1904–05]1977. The Evolution of Educational Thought. Routledge and Kegan Paul, London.
Durkheim, E., [1905–06]1956. Education and Society. Free Press, Glencoe, IL.
Durkheim, E., [1913–14]1983. Pragmatism and Sociology. Cambridge University Press, Cambridge, UK.
Dynan, K., 2000. Habit formation and consumer preferences. American Economic Review 90, 319–406.
Elias, N., [1939]1994. The Civilizing Process (Edmund Jephcott, Trans.). Oxford: Blackwell.
Emirbayer, M., Mische, A., 1998. What is agency? American Journal of Sociology 103, 962–1023.
Fearing, F., 1930. Reflex Action. Williams and Williams, Baltimore, MD.
Fisher, P., 1973. The failure of habit. In: Engel, M. (Ed.), Uses of Literature. Harvard University Press, Cambridge, MA, pp. 3–18.
Fleming, D., 1967. Attitude: history of a concept. Perspectives in American History 1, 287–365.
Fuhrer, J.C., 2000. Habit formation in consumption and its implication for monetary policy models. American Economic Review 90, 367–390.
Funke, G., 1958. Gewohnheit. Archive fur Begriffsgeschichte 3, 1–606.
Giddens, A., 2009. Sociology. Polity Press, Cambridge, UK.
Gorski, P.S. (Ed.), 2013. Bourdieu and Historical Anlaysis. Duke University Press, Durham, North Carolina.
Gronow, A., 2011. From Habits to Social Structures. Peter Lang, New York.
Helvetius, C.A., [1758]1807. De l'Esprit. Jones, London.
Hill, L., 1996. Anticipations of nineteenth and twentieth century social thought in the work of Adam Ferguson. Archives europennes de sociologie 37, 203–228.
Hillier, J., Rooksby, E. (Eds.), 2005. Habitus: A Sense of Place, second ed. Ashgate, Aldershot, UK.
Hitlin, S., Vaisey, S. (Eds.), 2012. Handbook of the Sociology of Morality. Springer, New York.
James, W., [1890]1950. The Principles of Psychology. Dover, New York.
Joas, H., [1992] 1996. The Creativity of Action. University of Chicago Press, Chicago.
Kahneman, D., 2003. A psychological perspective on economics. American Economic Review 93, 162–168.
Kant, I., [1798]1978. Anthropology from a Pragmatic Point of View. Southern Illinois University Press, Carbondale, IL.
Kestenbaum, V., 1977. The Phenomenological Sense of John Dewey. Humanities Press, Atlantic Highlands, NJ.
Li, Y., 2005. The wealth-consumption ratio and the consumption-habit ratio. Journal of Business and Economic Statistics 23, 226–241.
Lockwood, T.C., 2013. Habituation, habit, and character in Aristotle's Nicomachean Ethics. In: Sparrow, T., Hutchison, A. (Eds.), A History of Habit: From Aristotle to Bourdieu. Lexington Books, Lanham, MD, pp. 19–36.
Martin, J.L., 2011. The Explanation of Social Action. Oxford University Press, New York.
Neal, D.T., Wood, W., Quinn, J.M., 2006. Habits: a repeat performance. Current Directions in Psychological Science 15, 198–202.
Ostrow, J.M., 1990. Social Sensitivity. State University of New York Press, Albany, NY.
Ouellette, J.A., Wood, W., 1998. Habit and intention in everyday life. Psychological Bulletin 124, 54–74.
Park, R.E., [1915]1969. Man not born human. In: Park, R.E, Burgess, E.W. (Eds.), Introduction to the Science of Sociology. University of Chicago Press, Chicago, pp. 82–84.
Pollard, B., 2003. Can virtuous actions be both habitual and rational? Ethical Theory and Moral Practice 6, 411–425.
Schatzki, T.R., Knorr Cetina, K., von Savigny, E., 2001. The Practice Turn in Contemporary Theory. Routledge, London.
Sewell Jr., W.H., 2005. Logics of History: Social Theory and Social Transformation. University of Chicago Press, Chicago, IL.
Shachar, R., 2003. Party loyalty as habit formation. Journal of Applied Economics 18, 251–269.
Shilling, C., 2008. Changing Bodies: Habit, Crisis and Creativity. Sage, London.
Shook, J.R., Margolis, J. (Eds.), 2009. A Companion to Pragmatism. Wiley-Blackwell, Boston, MA.
Smith, L.D., 1986. Behaviorism and Logical Positivism. Stanford University Press, Stanford, CA.
Sparrow, T., Hutchison, A. (Eds.), 2013. A History of Habit: From Aristotle to Bourdieu. Lexington Books, Lanham, MD.
Thomas, J.M., 1993. Figures of habit in William James. The New England Quarterly 66, 3–26.
Thomas, W.I., Znaniecki F., [1918]1958. The Polish Peasant in Europe and America. University of Chicago Press, Chicago.
Turner, S., 2010. Explaining the Normative. Polity, Cambridge, UK.
Watson, J.B., 1919. Psychology from the Standpoint of a Behaviorist. Lippincott, Philadelphia.
Weber, M., [1910]1968. Antikritisches zum 'Geist' des Kapitalismus. In: Winckelmann, J. (Ed.), Die Protestantische Ethik II. Siebenstern, Munich, pp. 149–187.
Weber, M., [1910]1978. Anticritical last word on 'the spirit of capitalism'. American Journal of Sociology 83, 1105–1131.
Weber, M., [1922]1978. Economy and Society. University of California Press, Berkeley, CA.
White, H.C., 2004. Markets from Networks: Socioeconomic Models of Production. Princeton University Press, Princeton, NJ.

Habituation

RF Thompson, University of Southern California, Los Angeles, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 6458–6462, © 2001, Elsevier Ltd.

Abstract

Habituation is defined as a decrement in response as a result of repeated stimulation not due to peripheral processes like receptor adaptation or muscular fatigue. It is a process occurring within the nervous system (in animals with nervous systems). Habituation is definedin more detail by a number of parametric properties, involving such factors as stimulus frequency and intensity, pontaneous recovery of the habituated response, etc. Sensitization is defined as an increase in response as a result of (usually strong) stimulation.

Habituation primarily refers to stimulus-response systems that are 'hard-wired,' i.e., the stimulus elicits the response in the absence of any training, as in reflexes, fixed action patterns, arousal, etc. In neural systems where analysis has been possible, habituation is due to synaptic depression; a decrease in probability of transmitter release at appropriate synapses within the habituating circuit, and sensitization is an increase in the probability of transmitter release, due either to intrinsic or extrinsic processes. In vertebrates, at least, dishabituation, the disruption of habituation due to a strong extra stimulus, appears to be an independent superimposed process of sensitization. Habituation (and sensitization) appear to be ubiquitous in the animal kingdom and are extremely adaptive processes.

If a drop of water falls on the surface of the sea just over the flower-like disc of a sea anemone, the whole animal contracts vigorously. If, then, a second drop falls within a few minutes of the first, there is less contraction, and finally, on the third or fourth drop, the response disappears altogether. Here in this marine polyp with the primitive nerve net is clearly exhibited one of the most pervasive phenomena of the animal kingdom – decrement of response with repeated stimulation. Almost every species studied, from amoeba to man, exhibits some form of response decrement when the stimulus is frequently repeated or constantly applied. The ubiquity of the phenomenon plus its obvious survival value suggests that this kind of plasticity must be one of the most fundamental properties of animal behaviour. (Sharpless and Jasper, 1956, p. 655)

With these words, Sharpless and Jasper introduced their now classic study of habituation of the electroencephalogram (EEG) arousal response. Their study had a profound impact on behavioral neuroscience and served to kindle the interest of neural scientists in brain mechanisms of habituation. *Habituation* is a decrease in response to repeated stimulation, and *sensitization* is an increase in response as a result of stimulation. Virtually all behavioral responses, as well as autonomic and EEG responses and signs of arousal, habituate rapidly and to a profound degree. The processes responsible for habituation of behavioral responses probably occur largely after the primary sensory pathways of the brain and before the motor neurons that move muscles – somewhere between initial processing of sensory information and the execution of decisions to respond. Sensitization, the increase in response that often occurs when strong stimuli are given, may be similarly localized, although some evidence indicates that sensitization may occur in motor systems and motor neurons.

In early work on brain processes of habituation, it was thought that virtually all measures of neural activity might habituate, ranging from first-order sensory nuclei to the cerebral cortex. An early experiment to report evoked response habituation was that of Hernández-Peon et al. (1957); they recorded responses to click stimulation at several levels of the auditory system. Trains of clicks were delivered once every two seconds for long periods of time and evoked responses of the cochlear nucleus (the first relay of the auditory system) were reported to habituate.

Careful studies by Worden and associates (see Marsh et al., 1962) demonstrated that click-evoked responses at the cochlear nucleus do not show habituation. Instead, amplitudes of responses at this first relay nucleus in the auditory system are rigidly controlled by the physical properties of the sound stimulus. Because of acoustic factors, the intensity of a sound is often weaker at the floor of a test cage. If an animal gradually became bored and rested his head on the floor, the cochlear nucleus-evoked response to click would decrease because of reduced sound intensity.

In an extensive and careful study, Wickelgren (1968) found that habituation of evoked responses to auditory stimuli was not prominent below the level of the thalamus and cortex in the auditory system. This appears to be a general finding in all sensory systems – the lower ascending relay nuclei do not exhibit habituation. It would be most surprising if there were marked habituation in sensory systems. If our sensory systems habituated to sensory stimuli, we would rapidly cease to have sensation – a most unadaptive state of affairs. However, behavioral responses and phenomena that seem to be related to the probability of behavioral responding, such as EEG arousal and autonomic signs of orienting, show marked habituation. It is very adaptive to cease responding to arousing stimuli that have no particular significance.

Historically, such processes as habituation and sensitization were viewed simply as reflexes and not as phenomena of learning and memory. However, the broad definition of learning as a change in behavior as a result of experience certainly encompasses both phenomena. When this was recognized, analysis of neural mechanisms became 'legitimized' (Sharpless and Jasper, 1956; Thompson and Spencer,

1966). Habituation and sensitization are now seen as examples of nonassociative learning and memory. Greater progress has been made in analysis of the basic neuronal mechanisms underlying habituation than in more complex forms of behavioral plasticity such as classical and instrumental conditioning. Major reasons for this progress lie in the widespread occurrence of habituation, in its simplicity, and in the reliability with which the behavioral properties of habituation can be specified and determined. Habituation occurs both in spinal reflexes of the vertebrate (Sherrington, 1906) and for responses of simple animals, as well as in complex human behavior. Consequently, neural mechanisms of habituation can be studies in simplified systems.

As noted earlier, behavioral habituation is simply response decrement as a result of repeated stimulation. For example, if a moderate electric shock is delivered to the foot of a normal but restrained animal for the first time, the animal will show marked struggling and leg flexion movements. However, if the stimulus is repeated a few times, the responses decrease considerably and may even disappear. Habituation is commonly differentiated from response decrements due to trauma, growth, aging, etc. In the example above, if the shock to the foot is withheld for a period of time after habituation and then given again, the response will reappear with normal strength. This spontaneous recovery is a common control used to show that habituation is a *reversible* response decrement.

Response decrements due to altered sensitivity of receptors are often termed *receptor adaptation*, and response decrements due to decremental effects occurring at the neuromuscular junctions of muscles are often termed *muscular fatigue*. Neither of these processes is considered to be habituation; it is a central rather than a peripheral process. Almost all behavioral responses of the intact organism initially elicited by a given stimulus, ranging in complexity from leg flexion to exploratory and play behavior, show habituation with repeated stimulation.

Response sensitization may be defined simply as an increased response strength to a given stimulus as a result of some other, usually strong, stimulus. The galvanic skin reflex (GSR) is a good case in point. Suppose we were to measure the GSR activity of sweat glands in the palm of the hand in response to a sound of moderate intensity. The response would probably be of relatively low amplitude and somewhat irregular in occurrence. If a strong shock were delivered to the skin, the sound alone now would evoke a much larger GSR response; the response would have been sensitized. The response increase due to sensitization is distinguished from conditioning in that no pairing of shock and sound is necessary to produce the effect. Sensitization itself typically habituates; after a number of shock presentations the response to the sound decreases to the initial weak level. In fact, it seems to decay spontaneously following a sensitizing stimulus.

The study of habituation has a long history; some of its basic properties were described in classic works by Jennings (1906), Humphrey (1933), Prosser and Hunter (1936), and Harris (1943). Thompson and Spencer (1966) surveyed the by then very extensive behavioral literature on habituation and identified some nine basic parametric properties or characteristics exhibited by behavioral habituation.

Properties of Habituation

1. Given that a particular stimulus elicits a response, repeated applications of the stimulus result in decreased response (habituation). The decrease is usually a negative exponential function of the number of stimulus presentations.
2. If the stimulus is withheld, the response tends to recover over time (spontaneous recovery).
3. If repeated series of habituation training and spontaneous recovery are given, habituation becomes successively more rapid (this might be called potentiation of habituation).
4. Other things being equal, the more rapid the frequency of stimulation, the more rapid and/or more pronounced is habituation.
5. The weaker the stimulus, the more rapid and/or more pronounced is habituation. Strong stimuli may yield no significant habituation.
6. The effects of habituation training may proceed beyond zero or asymptotic response level.
7. Habituation of response to a given stimulus exhibits stimulus generalization to other stimuli.
8. Presentation of another (usually strong) sensitizing stimulus results in recovery of the habituated response (dishabituation or sensitization).
9. Upon repeated application of the sensitizing stimulus, the amount of dishabituation produced habituates (this might be called habituation of sensitization).

It is important to emphasize that these characteristics of habituation are simply the empirically observed properties in a wide variety of behavioral studies on many species and as such do not imply particular theories or inferences. These properties have come to serve as the detailed definition of habituation. Thus, any particular stimulus-response system is said to exhibit habituation if the response decrement to repeated stimulation exhibits these properties.

The most potent variables determining the course of habituation are stimulus frequency and intensity. The frequency range must be behaviorally relevant. Very high frequencies can yield fatigue, potentiation, and other processes. The range of stimulus frequencies over which habituation occurs, i.e., behavioral response decrements exhibiting the properties noted above, is an empirical question dependent upon the organism, the stimulus, and the response system engaged. For the orienting or arousal response, the effective range may be from minutes to hours to days; for the spinal flexion reflex, the range is over seconds.

The fact that rate and degree of habituation are inversely related to stimulus intensity (property (5)) is true if the same intensity is used for test and habituating stimuli (the test stimulus is given sufficiently infrequently that it does not by itself cause habituation). However, if the test and habituating stimuli differ in intensity, more complex effects can occur (Davis and Wagner, 1968).

Dishabituation vs. Sensitization

Thompson and Spencer (1966) employed habituation of the spinal flexion reflex, first used as a model system to study habituation by Prosser and Hunter (1936), to characterize

the detailed properties of habituation and analyze possible neuronal mechanisms. They discovered that the process of 'dishabituation' caused by presentation of a strong extra stimuli was not in fact dishabituation. That is, it did not disrupt the process of habituation at all. Instead, it added a superimposed, independent process of increased excitability, namely sensitization. The fact that dishabituation, per se, does not appear to exist as such, but rather reflects a separate process of sensitization, appears to hold for all vertebrate preparations studied, ranging from reflexes to the startle response to EEG arousal and the orienting response. One exception to this rule has been reported for a monosynaptic system in the invertebrate *Aplysia* (Marcus et al., 1988). However, the neuronal mechanisms underlying the processes of sensitization/dishabituation appear to differ in vertebrates vs. the *Aplysia* circuit. In the latter, 'dishabituation' appears to be a presynaptic process, occurring at the sensory neuron terminals on the motor neurons (see Sect. 3 and Kandel, 1976); whereas, at least in spinal reflexes, sensitization can involve increased excitability in motor neurons (Thompson and Spencer, 1966). Actually, recent evidence suggests that changes may occur in motor neurons in the *Aplysia* circuit as well (Murphy and Glanzman, 1997).

The fact that dishabituation does not exist as such, but rather reflects a separate independent process of sensitization in vertebrate preparations led Groves and Thompson (1970) to develop the dual-process theory of habituation: the outcome of repeated stimulus presentations is due to the net interaction of the inferred processes of habituation and sensitization. This theory is able to account for an astonishingly wide range of behavioral phenomena (see, e.g., Peeke and Herz, 1973).

A somewhat differently construed and widely influential theory of habituation is the stimulus-model comparator theory developed in Russia by Evgeny Sokolov (1963), based largely on data from the human orienting response. In this view, repetition of a given stimulus results in the development of a neuronal 'model' of the stimulus in forebrain structures. If a repetitive stimulus matches the model, then there is little output, only the habituated response, which could be no response, i.e., no EEG arousal. However, if there is a mismatch, then the difference is amplified via the reticular formation and a larger response results.

The dual-process theory is more successful in accounting for basic phenomena of habituation and sensitization in a wide range of species; the modelcomparator theory can more easily account for more complex phenomena like the 'missing stimulus' effect (increased response to the absence of one presentation of the habituating stimulus, a phenomenon sometimes seen in EEG arousal but not in reflex systems). But note that a process of synaptic depression could underlie the formation of the neuronal model of the stimulus in the model-comparator theory; the theories are at different levels and not incompatible. A view somewhat similar to that of Sokolov was developed by Wagner (1979) to account for long-term habituation. Thus, the central model of the habituating stimulus is considered as a relatively long-term memory trace, yielding the basic properties of a learned response. In their classic study of human EEG arousal habituation, Sharpless and Jasper (1956) stressed the distinction between short-term and long-term habituation.

As noted earlier, primary sensory afferent system neurons do not exhibit habituation. This makes perfect sense; in order to 'decide' how much to respond to a stimulus it is necessary to know what the stimulus is. Virtually all behavioral responses elicited by stimuli show habituation to repeated stimulation under at least some conditions. Central brain processes that are behavior-related like cortical EEG arousal show clear habituation, as do brain motor systems. Davis and associates completed a comprehensive analysis of the startle circuit and startle habituation in the rat (Davis, 1984). The circuit includes auditory relay nuclei: the cochlear nuclei and the nuclei of the lateral lemiscus; a nucleus in the brainstem reticular system (nucleus reticularis pontis caudalis) and its projections to brainstem and spinal cord premotor and motor nuclei. Evidence suggests that habituation to the acoustic startle stimulus occurred before the motor systems, i.e., at the nucleus reticularis pontis caudalis (Davis et al., 1982). The reticular nucleus is a 'bridge' nucleus between sensory and motor systems.

Neural Mechanisms

Unlike associative learning, in habituation the stimulus-response neuronal circuit is hardwired. In systems where some degree of analysis of neuronal mechanisms is possible, the neural substrate of habituation occurs within the stimulus-response circuit itself and is not due to any extrinsic circuit action like external inhibition acting on the habituating circuit.

Greatest progress in analyzing neural mechanisms of habituation and sensitization has come from simplified neuronal systems. Two of these are the gill withdrawal reflex of the mollusk *Aplysia* and vertebrate spinal reflexes. Kandel (1976) and his many associates completed a detailed analysis of the processes of habituation and sensitization in the *Aplysia* monosynaptic circuit. They were able to show that habituation was a presynaptic process involving a decrease in the probability of neurotransmitter release as a result of repeated activation. Similarly, a process of synaptic depression is responsible for habituation of the tail-flip escape response in the crayfish, another very productive invertebrate model system (Krasne, 1969). In the *Aplysia* model, synaptic decrement is due to a persisting decrease in calcium ion influx into the terminals. Sensitization, elicited for example by strong shock to the tail of the animal, was due to activation of a polysynaptic circuit that acted presynpatically on the sensory neuron terminals, the same terminals that habituated. The effects on calcium channels are due in turn to complex and not fully understood biochemical cascades within the neurons. Kandel and associates have also described a long-term sensitization effect involving structural changes in the synapses that they have used as a model for long-term memory storage (Hawkins and Kandel, 1984).

In the vertebrate spinal cord, the flexion reflex is polysynaptic. Thompson and Spencer (1966) were able to rule out any changes in excitability in the sensory neuron terminals and in the motor neurons themselves during habituation. However, sensitization did involve increases in motor neuron excitability, as noted above. They suggested, but could not prove, that a process of low-frequency depression in the interneurons was responsible

for habituation, a process analogous to the presynaptic depression process demonstrated by Kandel and associates in *Aplysia*. Using a monosynaptic pathway in the isolated frog spinal cord, Farel and Thompson (1976) demonstrated a presynaptic process for habituation. The story concerning neuronal mechanisms of habituation is particularly satisfying because the same process of synaptic depression appears to underlie habituation in both invertebrate and vertebrate systems. This simple process of synaptic decrement can account for a wide range of behavioral phenomena of habituation.

The study of habituation continues to be a most active field of research, ranging from responses of nematode C. *elegans* (Wicks and Rankin, 1996) to the startle response in mammals (Pilz and Schnitzler, 1996) to behaviors in the human fetus and infant (Leader, 1995, Kaplan et al., 1990).

Bibliography

Davis, M., 1984. The mammalian startle response. In: Eaton, R.C. (Ed.), Neural Mechanisms of Startle Behavior. Plenum, New York, pp. 287–351.

Davis, M., Parisi, T., Gendelman, D.S., Tischler, M.D., Kehne, J.H., 1982. Habituation and sensitization of electrically elicited 'startle' reflexes. Science 218, 688–689.

Davis, M., Wagner, A.R., 1968. Startle responsiveness after habi-tuation to different intensities of tone. Psychonomic Science 12, 337–338.

Farel, P.B., Thompson, R.F., 1976. Habituation of a monosynaptic response in frog and spinal cord: evidence for a presynaptic mechanism. Journal of Neurophysiology 39, 661–666.

Groves, P.M., Thompson, R.F., 1970. Habituation: a dual-process theory. Psychological Review 77, 419–450.

Harris, J.D., 1943. Habituatory response decrement in the intact organism. Psychological Bulletin 40, 358–422.

Hawkins, R.D., Kandel, E.R., 1984. Is there a cell-biological alphabet for simple forms of learning? Psychological Review 91, 373–391.

Hernandez-Peon, R., Jouvet, M., Scherrer, H., 1957. Auditory potentials at cochlear nucleus during acoustic habituation. Acta neurologica latinoamericana 3, 144–156.

Humphrey, G., 1933. The Nature of Learning. Harcourt, Brace, New York.

Kandel, E.R., 1976. Cellular Basis of Behavior: An Introduction of Behavioral Biology. W. H. Freeman, San Francisco.

Kaplan, P.S., Werner, J.S., Rudy, J.W., 1990. Habituation, sensitization and infant visual attention. In: Rovee-Collier, C., Lipsitt, L.P. (Eds.), Advances in Infancy Research. Ablex Publishing, Norwood, NJ, vol. 6, pp. 61–109.

Krasne, F.B., 1969. Excitation and habituation of the crayfish escape reflex: the depolarization response in lateral giant fibers of the isolated abdomen. Journal of Experimental Biology 50, 29–46.

Leader, L.R., 1995. The potential value of habituation in the prenate. In: Lecanuet, J.-P., Fifer, W.P., Krasnegor, N.A., Smotherman, W.P. (Eds.), Fetal Development: A Psychobiological Perspective. Lawrence Erlbaum, Hillsdale, NJ, pp. 383–404.

Marcus, E.A., Nolen, T.G., Rankin, C.H., Carew, T.J., 1988. Behavioral dissociation of dishabituation. sensitization, and inhibition of Aplysia. Science 241, 210–213.

Marsh, J.T., Worden, F.G., Hicks, L., 1962. Some effects of room acoustics on evoked auditory potentials. Science 137, 280–282.

Murphy, G.G., Glanzman, D.L., 1997. Mediation of classical conditioning in *Aplysia Califórnica* by LTP of sensorimotor synpases. Science 278, 467–471.

Peeke, H.V.S., Herz, M.J. (Eds.), 1973. Habituation, Volume I: Behavioral Studies, Volume II: Physiological Studies. Academic Press, New York.

Pilz, P.D., Schnitzler, H.-U., 1996. Habituation and sensitization of the acoustic startle response in rats: Amplitude, threshold, and latency measures. Neurobiology of Learning and Memory 66, 67–79.

Prosser, C.L., Hunter, W.S., 1936. The extinction of startle responses and spinal reflexes in the white rat. American Journal of Physiology 41, 609–618.

Sharpless, S., Jasper, H.H., 1956. Habituation of the arousal reaction. Brain 79, 655–680.

Sherrington, C.S., 1906. The Integrative Action of the Nervous System. Yale University Press, New Haven, CT.

Sokolov, Y.N., 1963. Perception and the Conditioned Reflex. (Waydenfeld S.W. Trans.). Pergamon Press, Oxford, UK.

Thompson, R.F., Spencer, W.A., 1966. Habituation: a model phenomenon for the study of neuronal substrates of behavior. Psychological Review 173, 16–43.

Wagner, A.R., 1979. Habituation and memory. In: Dickinson, A., Boakes, R.A. (Eds.), Mechanisms of Learning and Motivation: A Memorial Volume for Jerry Konorski. Lawrence Erlbaum, Hillsdale, NJ, pp. 53–82.

Wickelgren, W.O., 1968. Effects of acoustic habituation on click-evolved responses in rats. Journal of Neurophysiology 31, 777–784.

Wicks, S.R., Rankin, C.H., 1996. Recovery from habituation in *Caenorhabditis elegans* is dependent on interstimulus interval and not habituation kinetics. Behavioral Neuroscience 110, 840–844.

Habitus: History of a Concept

Gisèle Sapiro, École des hautes études en sciences sociales and Research, Paris, France; and CNRS (Centre européen de sociologie et de science politique), Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Borrowed from the Aristotelian–Thomist tradition, the concept of habitus was occasionally used in sociology and philosophy to designate an acquired and stable disposition. While phenomenologists (Husserl namely) were interested in its role in the construction of the self (ego) through past experience, it served sociologists such as Weber and Elias in their reflection on the collective dimension of attitudes and behaviors shared by a social group, be it religious (Weber) or national (Elias). Drawing from these two traditions of thought (phenomenology and sociology), Bourdieu theorized this concept, relating it to his theory of practice, which emphasizes the process by which individuals interiorize habits in their body through education. Bourdieu applies the concept of habitus to social classes and their differing lifestyles.

Designating an acquired disposition characterized by its stability, the concept of 'habitus' originally derives from the Aristotelian–Thomist tradition. Whereas medicine appropriated it to signify ways of being, and sickly states in particular, by their external manifestations (general physical and facial demeanors), the emerging field of sociology (notably Weber, Mauss, and Elias) designed it into a tool to conceptualize the incorporated form of collective habits, which vary from one society to another. Phenomenology, and Husserl in particular, employed it to think about the characteristics of the self constituted by past perceptions and decisions, which have settled into a prereflexive state of consciousness. These two traditions were synthesized by Pierre Bourdieu, who developed the 'habitus' into a chief notion of his sociological theory, linking it with his theory of practice.

From Scholastic Philosophy to Phenomenology and Sociology

The concept of habitus, derived from the Latin verb, *habere*, first appeared in scholastic philosophy in the sixth century in the work of Boèce, as a translation of Aristotle's 'hexis,' which signifies the positive dimension of 'the habit,' which is guided by reason and constitutes a virtue, unlike the ethos, which refers to the passive repetition of acts. It is reappropriated in the thirteenth century by Thomas Aquinas, particularly as he examines the ways of acquiring faith in the *Summa Theologiae*.

Max Weber employs this concept in the context of his sociology of religions to designate the psychological state that the good of salvation bestows upon the religious person, and which, notwithstanding the projection in the afterlife that it implicates, is a "habitus of the present" (Weber, 1996, p. 346). In his introduction to *The Ethical Economy of World Religions* (1915–20), Weber describes the "affective habitus" of the ascetic engendered through the specifically religious (or magical) act, through asceticism or contemplation (Weber, 1996, p. 347). He sets the active religious habitus that characterizes the West against Eastern traditions, which he describes as more contemplative. As he explains in the chapter of *Economy and Society* dedicated to religion, the rationalization of the appropriation of the goods of salvation consisted in eliminating the contradiction between the daily religious habitus and the extraordinary religious habitus, and in thereby assuring the continuity of the religious habitus (Weber, 1996, p. 189). The continuous and conscious religious habitus has become, in the religions of Salvation and Deliverance, a method of religious salvation that seeks to mitigate the inebriety attained through orgy (Weber, 1996, pp. 183–189). Thus, as the sorcerer requires rapture and a habitus defined by permanent charisma, for the prophet, orgiastic inebriety runs contrary to an ethical way of life. Instead of a sacred state reached through inebriety, asceticism, or contemplation, it is precisely the 'sacred habitus' that will ensure the salvation of individuals. If, in *The Protestant Ethic and the Spirit of Capitalism* (begun in 1904 and completed in 1920), the term does not appear more than twice, and always with respect to the religious habitus (Weber, 2008), Weber uses it in his response to criticisms of the work in order to name that which drove protestants to adopt behaviors in accordance with primitive capitalism, and which found its origins in their religious life, family tradition, or lifestyle imbued with religiosity (Weber, 1996, p. 155). He also wields this concept on several occasions in *Hinduism and Buddhism* (1915–20) to designate certain states leading to salvation, and particularly that of nirvana, the habitus attained once one's ties to the world are severed.

The Weberian definition is imbued with the concept's original Thomist meaning, in so far as it refers to a virtuous way of life based on deliberately adopted habits.

During the interwar period, the concept acquired a more encompassing meaning, which was no longer restrained to its religious dimension, and which came to also include habits. While phenomenology emphasized individual consciousness in its relationship to the surrounding world, sociology invested its interest in collective consciousness and transmission.

Preferring the concept of habitus to that of habits as a heuristic tool wielded to reflect upon the properties of the 'ego' ('Ich'), Husserl uses it in *Erfahrung und Urteil* (1939), translated in English as *Experience and Judgment* (1973), to refer to the way in which perception remains intentionally attached to an object beyond the first experience that it has of that object, transforming itself into latent knowledge capable of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03085-3

reactivating itself each time it encounters the said object. This concept thus simultaneously disagrees with empiricism and inneism, for it supposes that perception is structured by an acquired knowledge about objects situated "*on its horizon* of typical familiarity and precognizance" (Husserl, 1939/1973, p. 122). This is why Schütz translated the term habitus as employed in Husserl's work with that of 'habitual knowledge.' The ego is a 'substrate of habitualities,' explains Husserl in *Cartesian Meditations* (Méditations cartésiennes, originally published in French in 1931): the term habitus, which appears twice in the work (in sections 27 and 32), extends its scope to include decisions taken in the past and transformed into beliefs, convictions, which inform me, constitute myself as a stable and abiding ego, forming one's "personal character," according to Husserl's expression, or one's individual "style," to cite a term he also uses (Moran, 2011, p. 61). In this last work, as in *Ideen II* (1952), Husserl insists on the double dimension, passive and active, of the habitus: On the one hand, there is prereflexive daily action, and on the other hand, there is critical reflexive activity, which has a grip on the former, for instance the decision one may take to retain a belief (the second sort of habitus underpins the 'theoretical habitus,' which Husserl attributes to the scholar and the philosopher).

This double dimension is also central in the (nonsystematic) usage that Merleau-Ponty (1945/1976) makes of this concept. It was above all against behaviorism, which reduces action to simple reactions to stimuli (reflexes), that the French phenomenologist, nourished by his readings of Husserl, developed his reflection on the habitus as a disposition, comparable to a productive synthesis rather than the simple residue of past experiences. However, by contrast to Husserl, the author of *The Structure of Behavior* (*La Structure du comportement*, originally published in 1942) situates this knowledge in the body and no longer in consciousness. Neither automatism nor interiorized knowledge, the habit is "in the hands" (Merleau-Ponty, 1945/1976, p. 168). It is, as in a dance, through the kinesthetic grip that learning is accomplished, it is the body that 'understands.'

Foregoing a systematic usage of the concept, Durkheim and Mauss integrate it into their reflection on the process of 'socialization,' or, in other words, the inculcation on young generations through their education of 'systems of ideas, sentiments, and practices' specific to the groups to which they belong, such as "religious beliefs, moral practices and beliefs, national or professional traditions, and collective opinions of every kind," as Durkheim writes in 1911 in an article on "education, its nature, and its role" (Durkheim, 1922, p. 51; English trans. 1956, p. 72). Explaining the idea of conversion as invented by Christianity in *The Evolution of Pedagogy in France* (*L'Évolution pédagogique en France*, 1904–05), Durkheim uses the concept of habitus to qualify a "general disposition of the mind and will": a "habitus of moral being" (1938, p. 37; English trans. 1977, pp. 28–29). In his 1937 article on the "techniques of the body," Marcel Mauss also introduces the bodily dimension of the concept. According to Mauss, the term 'habitus' translates better than 'habit,' Aristotle's notions of the 'existing,' the 'acquired,' and the 'faculty.' Contrary to the theory of Gabriel Tarde, for whom the shared practices result from imitation, it is, according to Mauss, through education, and particularly through the training, the molding of the body, that are internalized and incorporated, not only social representations, but also rules of conduct and practices.

The term 'habitus' appears occasionally under the pen of Max Scheler, who, in *Formalism in Ethics* (1911), describes the 'habitus' of the chauvinist as being identical in every country, manifesting the same discourses and the same attitudes, and thus debunking the idea that the individuation of nations implies a different sort of nationalism. It is not certain whether this usage, which is closer to the common understanding of the term in German (manner, attitude) than it is to scholarly usages of the concept as developed in the work of Weber or in the Aristotelian–Thomist tradition, was a source for Norbert Elias' reflection on the national habitus. In his work, *Uber den Prozess der Zivilisation* (1937), translated into English as *The Civilizing Process*, Elias evokes the 'psychological habitus of civilized peoples,' with respect to the discussion of notions of culture and civilization, through which French and German customs and traditions came to distinguish themselves from one another. In *The Society of Individuals* (1939), Elias employs the term "habitus" to designate the way in which a particular form of "behavior-control" crystallized into a "character" or a "psychological individual habitus" (Elias, 2010, p. 41). Elias went on to systematize his usage of the concepts of 'social habitus' and 'national habitus' in the 1980s, after the notion was developed and popularized in the works of Bourdieu. The social habitus shared by the individuals of a same group is the expression of the particular forms of civilization characteristic of that group, as Elias expresses in an essay on the notion of civilization, written in 1986 (Elias, 2010, p. 5). In a text entitled, "Changes in the We–I balance" (1987), in which he describes how he relies on the concept of "social habitus" – that he sometimes likens to Kardiner and Linton's concept of the 'personality structure' – in order to overcome the opposition between individual and society, he explains how in differentiated societies, this social habitus is composed of several strata. In constituted nation-states, the national habitus takes precedent over regional belongings, while in modern states still in the process of construction, the contrary occurs. That national habitus hampers European unification (Elias, 2010, pp. 164, 188, 197). In *Studien über die Deutschen* (1989), he seeks to comprehend German nationalism in order to understand the rise of Nazism and what he calls the process of 'decivilization,' all while taking into account differences between classes and social groups. Far from being fixed, this national habitus, which Elias distinguishes from the essentialist notion of 'national character,' transforms itself through the process of intergenerational transmission. According to him, the advent of Nazism can be explained by a long experience of submission to autocratic regimes, closely associated with a code of obedience within a hierarchical structure: "The anchoring of an autocratic form of rule in the habitus of individual people kept on creating a strong desire for a social structure corresponding to this personality structure" (Elias, 1989/1996, p. 69).

Religious habitus, group habitus, national habitus, all of these sociological usages have in common the idea that certain durable dispositions fundamentally characterize groups of individuals within which it is transmitted, and which vary across space and/or time. Bourdieu would go on to apply this concept to class differences.

Pierre Bourdieu's Theory of Practice

The concept of habitus constitutes a key element in the sociological theory of Pierre Bourdieu, in that it founds his conception of action as much as his notion of the perception of the world. It is also at the heart of his analysis of social relations and lifestyles that structure the social space. This concept and the theory of practice that he associates with it were elaborated during a conjuncture in the social sciences defined by a critical paradigm shift. Although a bit younger, Pierre Bourdieu indeed belongs, alongside Louis Althusser and Michel Foucault, to the generation that constituted and affirmed itself against the branch of existentialism that then predominated in the intellectual field, pitting against a 'philosophy of the subject,' a 'philosophy without subject,' challenging the reigning subjectivism with the objectivism of the structures, setting against existentialist or 'personalist' humanism, what was then referred to as an 'antihumanism.' Against the rationalist tradition that, from Descartes to Sartre, posits consciousness' transparency to itself, the emerging social sciences, in the wake of psychoanalysis and Durkheimian sociology, honed in upon the gaps between actions agents undertake and the consciousness they have of them, which supposes an unconscious whereby the individual is acted upon more than he or she is acting, without necessarily reducing human behavior to a series of reactions to stimuli, as does behaviorism.

Nourished by the studies he undertook in Algeria and in Béarn between 1958 and 1960, then in France in the 1960s, as well as by his readings of Sartre, Merleau-Ponty, Husserl, Heidegger, Durkheim, Mauss, Saussure, Lévi-Strauss, Marx, Weber, Panofsky, Chomsky, Wittgenstein, and so on, the theory of 'habitus' progressively elaborated by Pierre Bourdieu during that period, participates in these questionings, to which it offers an original response: Seeking to overcome the contradictions between objectivism and subjectivism, mechanism and finalism, structures and individuals, etc., it reintegrates the experience of agents as partaking in social reality without searching therein the full explanation of their actions, the entire meaning of which is bound up in the system of hierarchical and structured relations within which they take place. Above all, Bourdieu's 'habitus,' following in the tradition of Mauss and Merleau-Ponty, places the body at the core of the individuals' capacity to produce a series of responses more or less adjusted to various situations.

The notion of 'habitus' appeared for the first time under Bourdieu's pen in a 1962 article on celibacy in Béarn, in which he referred to Mauss' article on the 'techniques of the body': describing a ball where peasants stand on the sidelines while the girls dance with the boys from the city, he explains that their habitus, that is to say, their physical properties – physical force, solidity, attachment to the earth – which served to their advantage in the precapitalistic era, came to constitute a handicap for the process of reproduction, as demonstrated in their inability to dance. The concept reappeared in 1967 in his postface to the French translation of Erwin Panofsky's *Gothic Architecture and Scholasticism*, which Bourdieu published in the series *le sens commun* of which he was the editor at Les Éditions de Minuit. According to him, Panofsky's approach, which illuminates the shared structural foundations of gothic architecture and scholastic thinking, allows the reader to grasp "the collectivity at the very heart of individuality, in the form of culture," that is to say, "the habitus by which the creator participates in and of his collectivity and his epoch, and which orients and directs, in spite of him, his acts of creation that appear to be the most unique" (Bourdieu, 1967, p. 142; free translation). Bourdieu directly draws upon Panofsky's concept, which its author himself derived from Thomas Aquinas, *The Sum* serving as his model of scholasticism. Bourdieu defines the habitus as a grammar generative of the thinking, perceptions, and behaviors that characterize a culture, by analogy with Noam Chomsky's notion of generative grammar in linguistics. However, for Bourdieu's habitus, this generative capacity is not innate, as it is for Chomsky's grammar, but, rather, is socially acquired through education. In this regard, Bourdieu refers to Durkheim and Mauss' classic article on "*The Primitive Forms of Classification*" (1903/2009). The site of this habitus' transmission and inculcation, which helps to explain the structural homologies between activities as different as the construction of cathedrals and scholastic philosophy, is school. Panofsky's approach invites us to seek for the 'modus operandi' (method of operation) of practices, how they produce the work, rather than the mere decoding of the 'opus operatum' (the work wrought).

It is in *Outline of a Theory of Practice* (*Esquisse d'une théorie de la pratique*, 1972) that Bourdieu first delivered a complete elaboration of his theory of practice, which he went on to reformulate in more systematic terms in *The Practical Sense* (1980a). His reflection on practice is anchored in a double criticism of objectivism and subjectivism. Against the mechanistic visions of structuralism and Marxism, which posit agents as mere conveyors of structures or executors of rules or roles, Pierre Bourdieu reintroduces the finalistic conception dear to phenomenology, which describes behaviors and modes of conduct as oriented with respect to objectives and underlines their meaning, just as he reintroduces the subjective point of view of agents, which, if it does not fully account for their conduct, definitely guides and inflects it. From this line of approach emerged the notion of strategy, another key concept of Bourdieu's theory, which helps to reinvest agents with a certain amount of inventiveness and improvisation, and to explain their attitudes in relation to their perception of the possibilities offered to them and, in particular, to the extent to which they perceive such possibilities as probable, as well as to their subjective hopes, more or less adjusted to their objective opportunities (Bourdieu, 1987, p. 75sq). Bourdieu reproaches structuralism for its 'juridicism,' which, according to him, constitutes an intellectual bias: The researcher tends to project the model he himself has constructed into the minds of agents without problematizing the distance that exists inherently between theoretical reconstruction and practice. This confusion can also be observed in the polysemous use that structural anthropology makes of the notion of 'rule,' between a norm that agents would consciously follow, a principle that would unconsciously guide them, and a model constructed by the researcher to explain the regularity of practices. Similarly, the structuralist theory appears insufficient to Bourdieu as a means of explaining the gift and countergift exchange, most notably because of the temporal gap that ensures that the giver can never fully know what the countergift will be. This temporal gap, which also occurs in the challenge and to

the riposte, opens a space for agents to develop strategies more or less adjusted to situations, following the rules of honor.

However, although trained in phenomenology, Bourdieu never intended upon returning to the ultrasubjectivist finalism of Sartrian philosophy, or to an intersubjectivity that ignores the objective conditions of existence and the weight of the structures within which agents are taken. He rejects subjectivist approaches, which, from Sartrian existentialism to the theory of the rational actor, suppose that action is the product of the will of a conscious subject capable of projecting himself/herself into the future as he/she anticipates the consequences of his/her actions. He simultaneously contests the idea of total freedom of choice and that of unbridled projection into the future. Drawing on Husserl's notion of 'protention,' he distinguishes between long-term rational planning constructed around a future project (the education of children, for example), which characterizes capitalistic societies, from "practical aiming at a yet-to-come inscribed in the present, thus apprehended as already there and endowed with doxic modality of the present," which prevails in traditional societies, organized around a cyclical calendar (Bourdieu, 1987/1990, p. 12).

The concept of habitus, conceived of as the generative principle of individual practices, designates the capacity of agents to orient themselves in the social world and to adopt modes of conduct adapted to their objective conditions of existence without explicitly adhering to a rule, to improvise attitudes adjusted to the situation at hand, just as a speaker is capable of producing an infinite number of sentences, with only a few rules as his/her guiding framework, and without speech being reducible to language. This concept is also what permits us to overcome the alternative between mechanistic approaches and finalism by lending an inventive dimension to structures that have been incorporated in the form of dispositions. In large part determined by past experiences, the habitus implies a degree of incertitude, of openness, and a capacity for adapting to change. It is also linked to the logic of practice, and thus to a certain vagueness and fluidity.

In Chapter 3 of *The Practical Sense* (*Le Sens pratique*, 1980a), entitled "*Structure, Habitus, Practice*," Pierre Bourdieu proposes that the reader thinks about practices and representations as constituting the product of a process undergone by individuals whereby they interiorize the social structures in the form of dispositions. These systems of dispositions, which compose the habitus, are 'structured structures' therewith rendered structuring structures. These structuring structures are schemes of perception, action, and evaluation that orient the agent's vision of the world and individual modes of conduct in conformity with the groups to which they belong (Bourdieu, 1980a, pp. 88–89). Thus, the modes of conduct of individuals belonging to a particular group are regulated, orchestrated, without necessarily following a particular rule, or being explicitly concerted.

The conditions whereby such dispositions are acquired through education explain the adjustment that occurs between incorporated structures (dispositions, habitus) and objective structures (things, institutions). Following in the line of Durkheim, Bourdieu conceives of social facts as ways of being, thinking, and acting that are more or less instituted, like in the law, and more or less solidified in things. There is thus not a difference of essence between incorporated structures and objective structures, two forms through which social life manifests itsef (Bourdieu, 1980b, p. 7). Furthermore, the most ritualized or objective forms of social life are nothing else than codified practices that have been institutionalized.

The illusion of immediate comprehension as evident, which characterizes the practical experience of the world, its being taken for granted, is the fruit of that adjustment between objective structures and incorporated structures. It also founds the practical sense, that is, the capacity to adopt conducts that are objectively adapted to situations without mechanically obeying any specific rule. This theory allows one to thereby explain the correlation that tends to be observed between objective probabilities and subjective hopes, but also the disparities that can occur between these in cases of dissonance between dispositions and structures, in periods of crisis, social change, or confrontation with another culture: Bourdieu evokes in this regard the inertia effect – or hysteresis – of the habitus, which resists change, and which often brings about sanctions as illustrated in the emblematic figure of Don Quichotte. The capacity to adjust to new structures differs according to social groups and the dispositions of individuals: Bourdieu (1977) compares, for instance, the attitudes of Algerian subproletariat workers who, unlike workers benefiting from the stability of an employment and thus more apt to adapt themselves to a capitalist temporality, are incapable of projecting themselves into a rational future and tend, as a result, to have recourse to a magical interpretation of the world. The illusion of an immediate comprehension of the world is also at the origin of the blindness that tends to prevail with respect to the principle that founds this very illusion and of a certain resistance to history, that is to say, to the study of its sociogenesis, which is nevertheless the only approach that can denaturalize it.

The inertia of the past makes itself felt through habits embedded in the body. The socialized body thinks, it is a 'habituated' body, as Bourdieu explains in *Outline of a Theory of Practice* (Bourdieu, 2002, p. 297). The body is temporally structured by the work of pedagogy, which consists in teaching it to defer pleasures, to domesticate its impulses (to transform hunger into appetite, etc.). However, the body is also the site of practice, invention, and improvisation. This faculty for improvisation remains nonetheless limited by the agents' conditions of socialization, the habits they have incorporated that often pose resistance – a manifestation of their inertia – to the most conscious efforts they make to acquire mastery over their bodies. All while taking into account this faculty for improvisation, which need not be grasped by reflexive consciousness to operate in an efficacious manner, the concept of habitus thus traces the limits of the freedom of action and of the symbolic mastering of practice.

The concept of habitus allows us to move on from 'practice' as a concept to practices and to their principles of differentiation. In class-based societies, groups of individuals are differentiated by their practices – alimentary habits, practices of consumption, ways of dressing, cultural practices, voting, etc. These practices are strongly linked among themselves and form lifestyles, manifestations of class-based habitus, but they are also differentiated according to the predominance of economic or cultural capital, as Bourdieu demonstrated in *Distinction* (1979). These differences assume their full meaning in a system of cultural oppositions that found principles

of classification that agents enact in their daily judgments of others' behaviors (for example, distinguished/vulgar, elegant/rude, etc.). Far from being invariable, the practices that characterize groups can evolve and even acquire 'nobility' when they are adopted by the dominant classes (like the punk style that has been introduced into high fashion) or become banal when they are vastly diffused (such as the practice of tennis).

Bourdieu distinguishes between the primary habitus, acquired by the child in the household, and the secondary habitus, acquired at school. The schemes of perception, action, and evaluation transmitted by the educational institution are closer to the culture of the dominant classes than to the culture of the dominated classes. The contradictory dispositions that form the primary and the secondary habitus of individuals belonging to the dominated class may produce what he calls a 'cleft habitus.' Such a cleavage within the individual habitus can also result from migration (cleavage between national habitus in Elias' terms) or from contradictory habitus within the family, for instance, between the values of the dominant fractions of the dominant class and those of its dominated fractions, as in the case of the French painter Edouard Manet to which Bourdieu devoted a study (Bourdieu, 2013).

As a 'sense of the game,' the 'practical sense' is also what is at work in the encounter between a habitus and a field (Bourdieu, 1980a). Indeed, far from concerning only ritualized practices, the logic of practice also prevails in intellectual and creative activities. Practical learning plays a major role in the preparation undertaken for artistic professions (art, music, dance), alongside a theoretical training that, indeed, does not always provide the tools for symbolically mastering practice. This is why the 'creative project,' an extreme expression of intentionalist theories, does not suffice in attempting to account for an artist's work, as Sartre would have it. The encounter between the habitus of a cultural producer and the space of possibilities that his or her field offers him or her is a much more powerful explicative principle, which allows one to make the connection between internal and external analyses of works, on the basis of a theory of practice that emphasizes the modus operandi rather than the opus operatum, even for an art as little taught as literature (Bourdieu, 1994, pp. 59–99). This theory of practice thus provides the basis for a general theory of action that extends from the most ritualized to the most inventive forms of action without a break in continuity.

The concept of habitus, which was largely discussed and appropriated by a number of researchers of different disciplines, earned Bourdieu the accusation of determinism, pronounced by certain critics who opposed him with the idea of 'agency,' without so much as noting that this capacity to act was in fact at the very core of the concept of habitus and manifested itself precisely in the form of strategies. The notion of habitus is more sociologically precise than that of agency because it allows for the definition of types (more or less conscious) of agency, its variation according to groups and individuals, according to their dispositions and to their trajectories, as well as the limitations that this capacity comes up against (interior, as well as exterior). At odds with rational choice theory, the concept of habitus enables us to understand how individuals are able to conduct their lives without always being reflexive, how they remain to a certain extent caught in their past experiences, and how they are – more or less – able to adapt to change.

See also: Bourdieu, Pierre (1930–2002); Chomsky, Noam (1928–); Durkheim, Emile (1858–1917); Elias, Norbert (1897–1990); Field Theory; Foucault, Michel (1926–84); Husserl, Edmund (1859–1938); Levi-Strauss, Claude (1908–2009); Weber, Max (1864–1920).

Bibliography

Bourdieu, P., 1967. Postface. In: Panofsky, E. (Ed.), Architecture gothique et pensée scolastique. Minuit, Paris.

Bourdieu, P., 1972. Esquisse d'une théorie de la pratique. Minuit, Paris. Ed. paperback 2000, Seuil 'Points', Paris. English trans.: 1977, Outline of a Theory of a Practice, trans. Nice, R., Polity Press, Cambridge.

Bourdieu, P., 1977. Algérie 60. Structures économiques et structures temporelles. Minuit, Paris. English trans.: 1979, Algeria 1960, trans. Nice, R. Polity Press-Éditions de la Maison des sciences de l'homme, Cambridge-Paris.

Bourdieu, P., 1979. La Distinction. Critique sociale du jugement. Minuit, Paris. English trans. 1984, Distinction, trans. Nice, R. Harvard UP, Cambridge.

Bourdieu, P., 1980a. Le Sens pratique. Minuit, Paris. English trans.: 1990, The Logic of Practice, trans. Nice, R. Polity Press, Cambridge.

Bourdieu, P., 1980b. Le mort saisit le vif: les relations entre l'histoire réifiée et l'histoire incorporée. Actes de la recherche en sciences sociales, 32–33.

Bourdieu, P., 1986. Habitus, code et codification. Actes de la recherche en sciences sociales 64.

Bourdieu, P., 1987. Fieldwork in Philosophy. Choses dites. Minuit, Paris. Bourdieu, P., 1987, Choses dites, Minuit, Paris. English trans.: 1990. In: Others Words. Essays Towards a Reflexive Sociology, Polity Press, Cambridge.

Bourdieu, P., 1994. Raisons pratiques. Seuil, Paris. English trans.: 1998, Practical Reasons. On the Theory of Action, trans. Johnson, R., Polity Press, Cambridge.

Bourdieu, P., 1997. Méditations Pascaliennes. Seuil, Paris. Ed. paperback 2003, Seuil 'Points', Paris. English trans.: 2000, Pascalian Meditations, trans. Nice, R. Polity Press, Cambridge.

Bourdieu, P., 2013. Manet. Une révolution symbolique. Seuil-Raisons d'agir, Paris.

Camic, C., 1986. The matter of habit. American journal of sociology 91 (5), 1039–1087.

Durkheim, E., 1922. L'Éducation, sa nature et son rôle. (1911), Éducation et sociologie. PUF, Paris. English trans. 1956. Education, Its Nature and Its Role. Education and Sociology, The Free Press, Glencoe, pp. 61–90.

Durkheim, E., 1937. L'Évolution pédagogique en France. PUF, Paris (1st published 1904–1905); English trans.: 1977. The Evolution of Educational Thought. Collins, P. trans.: Routledge & Kegan Paul, London.

Durkheim, E., Mauss, M., 1903/2009. English Transl. De quelques formes de classification – contribution à l'étude des représentations collectives. Année sociologique, 6; 'Primitive Classification", trans. Rodney Needham, London: COHEN & WEST. Second ed., 1969.

Elias, N., 1939/2000. Uber den Prozeß der Zivilisation. Soziogenetische und psychogenetische Untersuchungen. Band 1: Wandlungen des Verhaltens in den weltlichen Oberschichten des Abendlandes (LXXXI, 333 S.)/Band 2: Wandlungen der Gesellschaft: Entwurf zu einer Theorie der Zivilisation (491 S.), Verlag Haus zum Falken, Basel. English transl. The Civilizing Process. Sociogenetic and Psychogenetic Investigations. Revised edition of 1994. Basil Blackwell, Oxford.

Elias, N., 1989/1996. Studien über die Deutschen. Machtkämpfe und Habitusentwicklung im 19. und 20. Jahrhundert. Suhrkamp, Frankfurt am Main. English transl. The Germans. Power struggles and the development of habitus in the 19th and 20th centuries, Polity Press, Cambridge.

Elias, N., 1987/1991/2010. Die Gesellschaft der Individuen. Suhrkamp, Frankfurt am Main (first published 1939). English transl. The Society of Individuals, Blackwell, Oxford; new publication in the complete works of Norbert Elias, ed. van Krieken, R. UCD Press, Dublin.

Héran, F., 1987. La seconde nature de l'habitus. Tradition philosophique et sens commun dans le langage sociologique. Revue française de sociologie 28 (3), 385–416.

Husserl, E., 1931/1960. Méditations cartésiennes. English trans. by Cairns, D., Kluwer, Dordrecht.

Husserl, E., 1939/1973. Erfahrung und Urteil. Untersuchungen zur Genealogie der Logik. Claassen, Hamburg, 1948 (first published: 1939). English trans. Experience and Judgement. Northwestern UP, Evanston, Illinois.

Husserl, E., (French transl.) 1952/1989. Ideen II: Phänomenologische Untersuchungen zur Konstitution. English trans. by Rojcewicz, R., Schuwer, A., Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy – Second Book: Studies in the Phenomenology of Constitution. Kluwer, Dordrecht.

Mauss, M., 1950. Les techniques du corps. In: Sociologie et anthropologie. PUF, Paris.

Merleau-Ponty, M., 1942/1976. La Structure du comportement. PUF/Quadrige, Paris reprint 1970. English transl. Fisher., A. L., Beacon Press, Boston, 1963.

Merleau-Ponty, M., 1945/1976/1962. Phénoménologie de la perception. Tel 1976/1990. Gallimard, Paris. English transl. by Landes, D.A., Phenomenology of perception. Routledge, London.

Moran, D., 2011. Edmund Husserl's phenomenology of habituality and habitus. Journal of the British Society for Phenomenology 42 (1), 53–77.

Scheler, M., 1973. Formalism in Ethics and Non-formal Ethics of Values. English trans.: Frings, M.S., Funk, R.L.. Northwestern UP, Evanston, Illinois (first published in 1913–1916).

Sapiro, G., 2004. Une liberté contrainte. La formation de la théorie de l'habitus. In: Pinot, L., Sapiro, G., Champagne, P. (Eds.), Pierre Bourdieu, Sociologue. Fayard, Paris.

Wacquant, L., 2004. Habitus. In: International Encyclopedia of Economic Sociology. Routledge, London.

Weber, M., 1921/1968/1978. In: Roth, Guenter, Wittich, Claus (Eds.), Economy and Society. University of California Press, Berkeley and Los Angeles.

Weber, M., 1998. Gesamtausgabe. Studienausgabe/Schriften und Reden/Die Wirtschaftsethik der Weltreligionen. Hinduismus und Buddhismus: 1916–1920: Gemeinschaften (Max Weber-Studienausgabe) Taschenbuch – 1. von H. Schmidt-Glintzer (Herausgeber), K.H. Golzio (Herausgeber), & 1 mehr.

Weber, M., 1996. Sociologie des religions French transl. Grossein ,J.-P., Gallimard, Paris.

Weber, M., 2008. English trans. The Protestant Ethic and the Spirit of Capitalism. W.W. Norton & Company.

Halévy, Élie (1870–1937)

Michael Hill, Victoria University, Wellington, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Halévy was a French scholar whose original background in philosophy yielded a mode of analysis based on the Platonic dialectic which, together with his intense liberal perspective, provided a consistent core to his entire work. He examined the English Utilitarians, counterposing the laissez faire fundamentals of their economic theory with their panoptic emphasis in the judicial sphere. His subsequent study of Methodism's political moderation identified an underlying contradiction between Wesley's doctrine of egalitarian salvation and Methodism's hierarchical and disciplined structure. Finally, his study of European socialism identified the antinomy between the goal of liberty and the imperative of a compulsory organization of society. As he contemplated the onset of an 'era of tyrannies' in Europe, his liberalism was increasingly tinged with pessimism.

Élie Halévy was a French historian and philosopher who made three principal contributions to the social sciences: a study of the English Utilitarians; a historical account of English society in the eighteenth and nineteenth centuries, which is noteworthy for the role it attributes to Methodism and Evangelicalism in maintaining political stability; and an analysis of the development of European socialism culminating in what he designated an 'era of tyrannies.'

Biography

Élie Halévy was born at Étretat in Northwestern France to wealthy, artistic parents on 6 September 1870, a time which marked a turbulent period in French history, with the origins of the Third Republic, France's defeat by Prussia, and the collapse of the Paris Commune. Halévy's grandfather Léon Halévy, who was of Jewish descent, had been a disciple of Saint-Simon and succeeded Auguste Comte as his secretary for a time and he collaborated with Saint-Simon on his final works before Saint-Simon's death in 1825. Halévy's father was a successful librettist for such composers as Bizet and Offenbach, and his mother was of Swiss Calvinist origins. Both Élie and his brother Daniel were brought up as Protestants under her strong moral influence and like Halévy's contemporary Max Weber, there was a strong familial influence on the study of religious – especially Christian Protestant – ideas. Though the family had important connections in French political life, Élie remained distant from, but also keenly concerned with, political affairs.

While still a student at the Lycée Condorcet, he planned with his brother and other associates, the *Révue de Métaphysique et de Morale*, a journal with which he was associated for the whole of his life. Celestin Bouglé collaborated on the *Révue* and later joined Émile Durkheim as editor of the *Année Sociologique*. Halévy entered the École Normale Supérieure in 1889, where he studied philosophy and completed a thesis on Platonism. The dialectical method he discerned in Plato was to form the analytical core of all his subsequent work. Already his political and moral principles had become established as liberal, individualist, rationalist, and secular – principles which were to characterize him for the rest of his career. He attributed these principles to the prevailing ethos of the École, which during his time as a student, was strongly liberal and *Dreyfusard*. In later life he mused that if he had attended the École 5 or 6 years later, he would have been a socialist. In 1898 Halévy became a professor at the École Libre des Sciences Politiques, which had been founded after France's defeat in 1871, with the goal of diagnosing the causes of French political instability; and he remained there all his professional life. He was initially appointed to teach a course of lectures on the evolution of political ideas in England but in the academic year 1901–02 he launched another course on the history of European socialism; and he alternated between the two for the following 35 years. The École Libre was much admired in Britain by the Fabians and their associates who founded the London School of Economics and Political Science; and Halévy came to know them well as he moved in British intellectual circles.

The Dreyfus affair and the defence of the Third Republic were issues which united Halévy with Émile Durkheim, though on a philosophical level he rejected Durkheim's positivism and collectivism. On the other hand, he attributed his initial interest in Socialism, to which he devoted half of his scholarly life, to Durkheim's lectures on Socialism at Bourdeaux in 1895. Another scholarly characteristic which Halévy and Durkheim shared was the remarkable continuity and consistency in their intellectual development. The affinity that Halévy perceived between Communist and Fascist régimes – since he viewed 1914 as the start of an 'era of tyrannies' – was an unpopular stance in the intellectual climate of the interwar years, but he persisted in his opinion until his death at Sucy-en-Brie in the Paris suburbs on 21 August 1937. The enthusiasm of the Webbs and Shaw for Fascist or Communist experiments in no way surprised him and he ended his life firmly committed to the liberal, rational, democratic ideals of his youth.

Intellectual Contributions

In his thesis, published in 1896 as *La théorie platonicienne des sciences*, Halévy explicated a dialectical approach, which he labeled Platonic, in which the contradictions in a subject were

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61041-3

stated as systematic dichotomies or antinomies. His objective was to systematize the seeming contradictions in Plato between what he termed the 'regressive dialectic' and the 'positive dialectic,' constructing from these his distinctive mode of analysis. Thus Halévy stated his aim as that of interpreting the dialectic "both as a critical method, with the purpose of destroying in us the illusion of science, and as a positive method, with the purpose of justifying all the sciences" (in Gillespie, 1950: p. 234; my translation). By exposing the core assumptions and practical implications behind a doctrine or belief system that usually involved logical conflicts, Halévy's own analysis of it could proceed. This fundamental principle of resolving contradictions into systematic dichotomies was deployed in his subsequent studies of Bentham's thought, Wesleyan doctrine and Socialism. His choice of England as the site of his first two areas of research was motivated by two main considerations: first, because it offered a striking location for the study of his central interest, the problem of liberty; and second, because the problematic relationship of ideas and beliefs to material circumstances in historical causality were clearly identifiable there.

In his study of the English Utilitarians he found a fundamental contradiction between Bentham's juridical and political theory on the one hand and economic theory on the other, and he argued that "the contradiction was continually breaking out in the current formulae of Benthamism" (Halévy, 1952: p. 489). In his political theory, Bentham postulated the necessity of a legislator who would reconcile conflicts of interest and thus secure the greatest good of the greatest number: this he called the artificial identity of interests. In his economic theory, however, Bentham assumed, with Adam Smith and his successors, that there was a spontaneous identity of interests and that individual interests were naturally harmonious. Halévy's book is a reminder that the Bentham who designed the *Panopticon* – "a new application of the principle of the artificial identification of interests, the idea of which he had found in Helvetius" (Halévy, 1952: p. 82) – also propagated a belief in laissez faire and the 'hidden hand.' Halévy suggests that because Bentham and his disciples focused on different aspects of their philosophy at different periods in its development, they were unaware of the contradictory dualism which underpinned their work. On the broader influence of the Utilitarians, Halévy argued that while they proposed no fundamentally new ideas – drawing on the ideas of Locke and their systematic development in French thought, particularly that of Montesquieu – their achievement lay in popular and practical agendas, which promoted free trade, legal reform, and the democratization of parliamentary government.

Soon after the completion of *The Growth of Philosophic Radicalism*, Halévy's interest turned to Methodism in eighteenth and nineteenth century England. Methodism was sometimes regarded as the complete antithesis of Utilitarianism, but Halévy regarded Methodism as an ally of Utilitarianism in shaping the social morality of modern England; indeed, he went on to argue that there was a strong affinity between Utilitarianism and Evangelicalism in the emphasis that both placed on self-restraint and self-discipline. In 1905 he traveled to England to work on material relating to the early history of Methodism and it was at this time that he expressed his intention of writing a history of England in the nineteenth century. His research on Methodism had a more immediate outcome, however, in two articles on 'La Naissance du Méthodisme en Angleterre' in the *Revue de Paris* in 1906. With very few exceptions these articles went unnoticed by historians until their translation into English in 1971 (Halévy, 1971).

The fundamentals of Halévy's interpretation of the role of Methodism in 'La Naissance…' correspond to his later treatment in *A History of the English People in 1815* (Halévy, 1924) – the first volume of his five-volume *A History of the English People*, which was published in French between 1912 and 1926 and in English between 1912 and 1932; however, the focus and detail of the earlier work provides a particularly strong statement of the 'Halévy thesis.' Since the thesis has become so indelibly associated with him, it is important to preface an account of it by noting that he was not the first to highlight Methodism's purported impact in averting revolution in England. Such an influence had been noted as early as the 1830s, and in 1878 William Lecky had published a particularly strong version of this interpretation in the second volume of his *History of England in the Eighteenth Century*. Halévy took as his point of departure the earliest development of manufacturing industry in the first three decades of the eighteenth century and pointed out the growth of pauperism in the industrial centers. By the late 1730s this was compounded by a commercial crisis of overproduction accompanied by working class agitation and it was this that stimulated the 'eruption,' in 1739, of Methodism. He argues that working class despair was transformed by Methodist doctrine and discipline into a religious and mystical movement of a conservative nature. In this process Halévy gave a key role to the bourgeoisie, a class that he saw as an essential component of leadership in harnessing the enthusiasm of the proletariat. Had the bourgeoisie in 1739 been fired by revolutionary ideals there might well have been conditions for social revolution; but the English bourgeoisie retained a strong element of Puritanism and a revulsion against the excesses of the seventeenth century Civil War. In Halévy's words, "There was the raw material for a general revolt. But popular discontent took the form the discontent of the bourgeoisie wanted to give it: a religious and conservative form" (Halévy, 1971: p. 75).

On this foundation is constructed the broader thesis with which Halévy is more generally known. At the outset it is noteworthy that his historical interpretation is specifically advanced as a critique of the materialist theory of history: "If the materialist theory of history is to be trusted, if economic facts explain the course taken by the human race in its progress, the England of the nineteenth century was surely, above all other countries, destined to revolution, both politically and religiously" (in Gillespie, 1950: p. 239). In *A History of the English People in 1815*, he stated the central problem as that of explaining "the extraordinary stability which English Society was destined to enjoy throughout a period of revolutions and crises; what we may truly term the miracle of modern England, anarchist but orderly, practical and businesslike, but religious, and even pietist" (Halévy, 1924: p. 339). The explanation he gives is complex and is not focused exclusively on Methodism nor solely on the working class. He traces the impact of Methodism first on Nonconformity, then on the Evangelical wing of the Church of England, and through them to the ruling class and finally to secular thought. In contrast to the view of religion as an 'opiate,' which Marxian critics of

Halévy such as E.P. Thompson have employed, the thesis is better interpreted as an exposition of upward social mobility through a process of *embourgeoisement*. Halévy did not attribute the influence of Methodism to the depths of the working class but saw it as having an impact on upwardly mobile workers: early in his *History* he refers to the "independent Churches of the lower middle class." Methodism, originating in the Church of England, occupied an interstitial position between that Church and Nonconformity, offering a means of passage between the two: "Puritan nonconformity thus tended to become a transitional creed, a stage in the history of an English family. The unskilled labourer becomes in turn a skilled workman, an artisan, the head of a small business, a business man possessed of a modest capital, and as he rises out of the barbarism in which the working class was plunged, he becomes a Nonconformist. If he himself rises still higher on the social ladder, or if his children rise after his death, he, or they go over to the Church of England" (Halévy, 1924: p. 371). Wesleyan Methodism, he claimed, rendered the transition imperceptible. Paradoxically, as the religious institutions facilitated such upward social mobility they also served to sharpen status distinctions; whereas the middle-class Nonconformist might have been despised by members of the Church of England (to which his family might one day belong), compensation for the precarious social position in which such upwardly mobile individuals found themselves could be identified in the contempt shown toward the working class from which the family had come. Hence Halévy's opinion that "the élite of the working class, the hard-working and capable bourgeois" had been instilled with values which did not threaten the existing order of society.

Wesleyan Methodism had originated in the Church of England among a group of conservative High Church clergy. In preaching an egalitarian doctrine of salvation for all, it had successfully mobilized an important segment of the rising working class. Here is an excellent example of Halévy's utilization of systematic dichotomy at the core of a belief system. He goes on to demonstrate the influence which Methodism had on establishment opinion by proposing that Wesley had left a 'rear-guard' in the Church of England in the form of the Evangelical movement, which sought to realize his original goal of church regeneration: "If the Wesleyan sect, with its hierarchic constitution, and frank political conservatism, constituted the High Church of Nonconformity, the new Low Church or evangelical party was a species of Anglican Methodism" (Halévy, 1924: p. 379). With powerful representatives at the University of Cambridge, and with wealthy and influential advocates in the 'Clapham sect' (which was comprised principally of Members of Parliament like William Wilberforce), the Evangelicals were in a position not only to influence ecclesiastical reform but also to campaign for a wider reform of morals – including the morals of the ruling class. In these several ways, argued Halévy, the influence of Wesleyan Methodism had permeated throughout English society and helped to establish the temper of Victorian England. The dominant features of English life in the nineteenth century were moral discipline, self-restraint, and the ethos of individualism which derived from Methodism and the wider Evangelical movement. Furthermore, Evangelical religion at the level of the individual, advocated philanthropy for the successful and resignation coupled with self-help for those who did not succeed. By contrast, no comparable institution, creed, or religious culture had developed in eighteenth century France.

Halévy's *Histoire du socialisme européen* (1948) and *L'ère des tyrannies* (1938) are the most cited of his works in France. The *Histoire* … was based on Halévy's own abundant lecture notes and those of some of his students. It was edited by Celestin Bouglé and Raymond Aron and has not been translated into English. *L'ère* appeared in English translation in 1966 as *The Era of Tyrannies: Essays on Socialism and War*. These works further extend the central theme of Halévy's work on English ideas, where the contrast between reform and revolution provided his comparative focus. Just as the contradiction between natural and artificial interests had been identified in Utilitarianism, and that between egalitarian salvation and hierarchical discipline in Wesleyanism, so the contradiction in socialism derived from the conjoint goals of human emancipation and the zeal to organize society. His history of socialism focused especially on the economic doctrine of Saint-Simon and his followers and a few years after he began his lectures on socialism he identified Saint-Simon as the pioneer or 'great precursor' of Socialism (Jones, 2002: p. 60) – a reminder that his grandfather Léon had succeeded Auguste Comte as Saint-Simon's secretary. The First World War was the decisive event in Halévy's view of Socialism, which was revised from a skeptical liberalism to a critical pessimism. The growth of the power of the state and the spread of bureaucracy were seen not only as threats to parliamentary democracy but also as evidence of the dominance of the organizational over the emancipatory principle in Socialism.

Halévy's central thesis is clearly stated: "Since its beginnings, in the early years of the nineteenth century, socialism has suffered from an internal contradiction. On the one hand, its partisans often present it as the outcome and fulfilment of the Revolution of 1789, a revolution of liberty, a liberation from the last remaining subjection after all the others have been destroyed: the subjection of labor by capital. But, on the other hand, it is also a reaction against individualism and liberalism; it proposes a new compulsory organization in place of the outworn institutions destroyed by the Revolution …" (Halévy, 1966: p. 265). It is known that the pessimistic conclusions he drew from this diagnosis surprised many in his intellectual circle at the Société Française de Philosophie and arrangements were being made for a further debate on his propositions, when he died in 1937.

Impact and Controversy

Halévy's work continues to be a focus of research and debate among social scientists. Two biographies have appeared, the first relating him to the French and European contexts of his time and the second emphasizing the major themes of his work: both characterize the 40-year span of his writing with the same term, 'unity' (Bramsen, 1978: p. 348; Chase, 1980: p. 3). In Chase's work there is a particularly clear exposition of the 'Halévy thesis' which gives effective treatment to aspects of his argument which have sometimes been obscured in subsequent debate, such as the broader significance of Nonconformity and his portrayal of the "hard-working and capable bourgeois" as the "élite of the working class." The republication in 1995 of

a French edition of his book on English Utilitarianism has stimulated critique of a work which had been widely accepted as a definitive analysis of its subject. Thus Halévy has been taken to task for allegedly interpreting the principle of utility or the pursuit of pleasure, as both a positive and a normative statement – as both an 'is' and an 'ought' (Vergara, 1998).

The parallels between Halévy and Max Weber are unmistakable. First, they are both associated with an eponymous 'thesis,' the core of which already existed as "the germ of a hypothesis" (Hill, 1973: p. 103). As indicated earlier, the claim that Methodism had been influential in preventing a revolution in England had been made as early as the 1830s. In a similar manner, the link between Protestantism and the rise of Capitalism which is enshrined in the 'Weber thesis,' had been proposed as early as 1829 by the poet Southey. Secondly, and more substantially, both adopted a comparative methodology in their accounts of the development of the modern state while retaining a sensitive awareness of the importance of historical contingency. A third equivalence in the work of Halévy and Weber can be found in their exploration of the link between ideal and material factors in the processes of social change. As Gillespie notes, Halévy's work "leaves the impression that the chief practical influence of ideas comes after they have hardened into dogmas and (though this makes the idealist interpretation of history a little difficult) that what often converts ideas into dogmas is their applicability to real interests in definite circumstances" (Gillespie, 1950: p. 236). This 'difficulty' is precisely what Weber addressed in his treatment of the dialectic between 'ideal' and 'material' interests associated with the twin concepts of 'elective affinity' and 'routinization.' Finally, the work of both scholars is characterized by a marked pessimism over the future progress of European societies, which they saw as destined to be subject to ever-encroaching bureaucratic control.

Halévy has exercised an enormous impact on the analysis of England's religious history and extensive attention has been paid to his claim that the Methodist and Evangelical movements were responsible for England's remarkable stability during a period of political ferment in Europe. Shortly after the publication of the first volume of *A History of the English People in the Nineteenth Century*, the Hammonds offered their portrayal of Methodism as having a repressive and intolerant stance toward the working class, and their interpretation questions Halévy's portrayal of its liberalizing influence on the exercise of judicial authority. In 1957 in *History Today* E.J. Hobsbawm took issue with Halévy's appraisal of the extent to which Methodism penetrated the working class and offered an alternative explanation that England's political stability was derived from the ability of the ruling class to retain control. E.P. Thompson's *The Making of the English Working Class* in 1963/1968, however, attributed even more influence to Methodism in subduing the working class than Halévy. Emphasizing the duality of Methodist emotionalism in worship and self-discipline in the workplace, he succinctly links Halévy with Weber when Thompson (1968: p. 406) notes: "Sabbath orgasms of feeling made more possible the single-minded weekday direction of those energies to the consummation of productive labour." Clark provides a reminder that Wesleyan Methodism was not a static movement when he states that "If the Methodism of the 1740s is to be understood in part as an Anglican exercise in social control, the evolution of Methodism in the direction of Dissent may have limited the movement's ability to provide an underpinning for society" (Clark, 2009: p. 22). Finally, in a 'Postscript on the Halévy Thesis' in *Victorian Minds*, Gertrude Himmelfarb (1968) commented on the curious status of this thesis in current historical literature. Whereas it has been often assumed that the thesis had been discredited or greatly modified, in reality there has been no substantial exploration of it. After reviewing the main features of the debate she concluded by paraphrasing Halévy: "Will Halévy defeat Marx as the interpreter of this crucial period in English history?" (Gertrude Himmelfarb, 1968: p. 299).

Himmelfarb's postscript became the preface to two decades of research into the 'Halévy thesis.' In 1971 Bernard Semmel's translation of Halévy's 1906 articles, *The Birth of Methodism in England*, was published with an introduction which noted that "the full dimensions of the thesis, as Halévy stated them, have been imperfectly understood" (Gertrude Himmelfarb, 1968: p. 2). This was because the 1906 articles had gone virtually unread. The thesis was further explored in *The Methodist Revolution* (Semmel, 1973), which began by noting that in major discussions of the 'Halévy thesis,' the role of Methodist theology had been afforded little significance, in contrast to social scientific discussions of seventeenth-century Calvinism. Semmel devoted several chapters to the elements which contributed to the synthesis of Methodist beliefs, restating Halévy's observation of an underlying contradiction between the revolutionary and counterrevolutionary elements of Wesleyan doctrine.

The thesis continued to be the subject of discussion throughout the 1970s and attracted the attention of sociologists examining religion and social change. In this context the 'Halévy thesis' has been described as "the political corollary, applied in a more restricted historical and geographical context, of the broader thesis put forward by Weber" (Hill, 1973: p. 183). As closer attention has been paid to Halévy's work, some of his specific historical observations – and their interpretation by more recent writers – have provided further stimulus for debate. One such area has been the role of foreign missions, and specifically of the Wesleyan Methodist Missionary Society. Did such activity provide a safety valve for social and religious tensions, rechanneling potentially disruptive enthusiasm by exporting it, or was there equal concern for home missions? These questions have been addressed in a revisiting of Halévy's original account (Piggin, 1980). Piggin challenges the 'Semmel thesis' which argued that foreign missions provided a mechanism of social control by rechanneling potentially disruptive evangelical enthusiasm, and especially Semmel's claim that the 'mission party' within Wesleyan Methodism "deliberately redirected resources and energies into the safe and enervating task of converting the rest of the world" (Piggin, 1980: p. 18). Instead, he thinks, foreign missions revitalized the home churches and made available enhanced resources for home missions.

In a wide-ranging review of the debate, Itzkin (like Himmelfarb before her) has noted that many of the contentious issues remain unresolved. While Halévy's antagonists have succeeded in revising some of his claims, the central thesis has remained viable and continues to constitute a 'working historical hypothesis' (Itzkin, 1975: p. 56). In the work of

revision, attention has focused on the numerical strength of Methodism together with its geographical distribution; on the nature and political stances of the variety of bodies which came within the overall rubric of Methodism; and on the way in which moderate political deviance can make "an important contribution to public order and political stability" (Gilbert, 1979: p. 392).

With the culmination of two decades of lively debate over the 'Halévy thesis' a collection of most of the important statements was published in Olsen (1990). Extracts from Halévy, from historians of Methodism and Evangelicalism such as Maldwyn Edwards, John Walsh, and Kiernan, and from critics like the Hammonds, Hobsbawm, and Thompson, provide a striking testimony to the continuing vitality of a thesis first advanced in 1906. As has been indicated, the parallel with the 'Weber thesis,' first published in 1904–05, is too close to ignore.

Halévy's lectures on the history of Socialism, reconstructed from his students' notes and published posthumously as *Histoire du Socialisme Européen* in 1948, remains untranslated into English. *The Era of Tyrannies*, first published posthumously in French in 1938, has however been translated and published in English (Halévy, 1966). After its first publication in France, his argument stimulated interest among those with whom Halévy maintained links at the London School of Economics, including Freidrich Hayek, who was editor of the LSE house magazine *Economica* in which Halévy's piece was republished. It has been suggested that his influence is identifiable in Hayek's *The Counter-Revolution of Science* and *The Road to Serfdom*, both of which date from the same period in the early 1940s (Jones, 2002). In *The Era of Tyrannies* the author provides not only a summary account of socialism but also a treatment of the 'era of tyrannies' – using the term in the classical Greek sense of a group of armed men who seize the state – under Stalin, Hitler, and Mussolini. At the end of World War I Halévy had written: "…I recognize that socialism holds the key to the secret of the future. But I cannot decipher this secret nor can I say whether socialism will lead us to a Swiss-type, universalized Republic or to European Caesarism" (in Aron, 1984: p. 407). Halévy's emphasis on the concrete influence of World War I rather than the influences of Marxism or other collectivist doctrines in the advance of socialism has been seen as qualifying somewhat his otherwise idealist interpretation of historical change.

In the last years of his life Halévy took an increasingly pessimistic view of Europe's future in the event of another war, which he thought likely. For him, this would represent the final blow to liberal traditions; but for this judgment Aron has criticized him. While cautioning against overconfidence, Aron thought Halévy had "sinned through over-pessimism" (Aron, 1984: p. 422). With the benefit of hindsight Aron's criticism may be justified, but it should be recalled that in the era of the 1930s Halévy was not alone among social scientists who believed that the capitalist élite were becoming 'highly placed bureaucrats.'

See also: Durkheim, Emile (1858–1917); Liberalism, Anthropology of; Marx, Karl (1818–83); Pestalozzi, Johann Heinrich (1746–1827); Socialism; Weber, Max (1864–1920).

Bibliography

Aron, R., 1984. Élie Halévy. Government and Opposition 19 (4), 407–422.
Bramsen, M.B., 1978. Portrait D'Élie Halévy. B.R. Grüner, Amsterdam.
Chase, M., 1980. Élie Halévy: An Intellectual Biography. Columbia University Press, New York.
Clark, J.C.D., 2009. The eighteenth-century context. In: Abraham, William J., Kirby, James E. (Eds.), The Oxford Handbook of Methodist Studies. Oxford University Press, Oxford.
Frobert, L., 2007. Élie Halévy's first lectures on the history of European socialism. Journal of the History of Ideas 68 (2), 329–353.
Gilbert, A.D., 1979. Methodism, dissent and political stability in early industrial England. Journal of Religious History 10 (4), 381–399.
Gillespie, C.C., 1950. The work of Élie Halévy: a critical appreciation. Journal of Modern History 22 (3), 232–249.
Halévy, É., 1924. A History of the English People in 1815 (E.I. Watkin, D.A. Barker, Trans.). T. Fisher Unwin, London.
Halévy, É., 1952. The Growth of Philosophic Radicalism (Translated by M. Morris with a Preface by A.D. Lindsay). Faber and Faber, London.
Halévy, É., 1966. The Era of Tyrannies (Translated by R.K. Webb with a Note by F. Stern). University Press, New York.
Halévy, É., 1971. The Birth of Methodism in England (B. Semmel, Ed. and Trans.). University of Chicago Press, Chicago.
Hill, M., 1973. A Sociology of Religion. Heinemann Educational, London.
Himmelfarb, G., 1968. Victorian Minds. Weidenfeld and Nicolson, London.
Itzkin, E., 1975. The Halévy thesis – a working hypothesis? Church History 44 (1), 47–56.
Jones, H.S., 2002. The era of tyrannies: Élie Halévy and Friedrich Von Hayek on socialism. European Journal of Political Theory 1 (1), 53–69.
Lecky, W.E.H., 1883. A History of England in the Eighteenth Century, third ed., Revised. Longmans, Green, and Co., London.
Olsen, G.W. (Ed.), 1990. Religion and Revolution in Early-Industrial England: The Halévy Thesis and Its Critics. University Press of America, Lanham.
Piggin, S., 1980. Halévy revisited: the origins of the Wesleyan methodist missionary society. Journal of Imperial and Commonwealth History 9 (1), 17–37.
Pombeni, P., 2005. Political models and political transfer in the shaping of Europe. European Review of History 12 (2), 223–238.
Semmel, B., 1973. The Methodist Revolution. Basic Books, New York.
Smith, C.H., 1942. Élie Halévy (1870–1937). In: Schmitt, Bernadotte E. (Ed.), Some Historians of Modern Europe. Chicago University Press, Chicago.
Thompson, E.P., 1968. The Making of the English Working Class. Penguin Books, Harmondsworth.
Vergara, F., 1998. A Critique of Élie Halévy: Refutation of an important distortion of British moral philosophy. Philosophy 73 (283), 97–111.

Hall, Granville Stanley (1844–1924)

Lester F Goodchild, University of Denver, CO, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 6466–6469, © 2001, Elsevier Ltd.

Abstract

Granville Stanley Hall's (1844–1924) lifetime intellectual accomplishment was the creation of the American discipline of psychology and the science of education out of the then broader study of philosophy. He developed these academic specializations through his long faculty and presidential career, numerous books, and more than 400 other publications, creation of several research journals, and extensive professional conference presentations. Accordingly, after earning the first American doctorate in psychology from Harvard in 1878, a brilliant academic career followed. Hall began as a lecturer in philosophy at the new Johns Hopkins University that led to his first appointed position there as professor of psychology and pedagogics in 1884. He established the *American Journal of Psychology* there in 1887. The following year, Hall became president of research-oriented Clark University where he continued his professional and scholarly activities. In 1892, Hall founded the American Psychological Association. He then furthered the scientific study of education through his lectures and became the national leader of the American child study movement. To promote such research, the president created the educational research journal, *The Pedagogical Seminary*. Through these and other activities, Hall played the initial pivotal and defining role in founding psychology and education as social and behavioral sciences in American higher education.

The age of the modern research university began with the founding of the Humboldt University in Berlin in 1810. One of the 9000 Americans who studied at this and other German universities during the nineteenth century was Granville Stanley Hall. He, like they, sought to gain more advanced knowledge and research skills than was available elsewhere. This educational opportunity aided him in becoming the first professor of psychology and pedagogics at the new Johns Hopkins University, and later the first president of research-oriented at Clark University in the USA. During his academic career, Hall launched the American discipline of psychology (Ross, 1972, 1979) and the science of education (White, 1992), out of the then broader study of philosophy. He achieved these intellectual accomplishments through his long faculty and presidential career, numerous books and more than 400 other publications, creation of research journals, and extensive professional conference presentations. In this way, Hall played the initial pivotal and defining role in founding psychology and education as social and behavioral sciences in American higher education (Mitchell and Haro, 1999, p. 50; Lagemann, 2000).

Hall's Early Life, Career, and Intellectual Pursuits Until 1880

Born in Ashfield, Massachusetts on 1 February 1844 to Granville Bascom Hall and Abigail Beals Hall, Stanley was the oldest of nine siblings. His upbringing on the family farm offered his inquisitive mind opportunities to observe nature intimately, which would so powerfully shape his future intellectual outlook. Eschewing agrarian life, Hall completed college preparatory studies at Williston Seminary in Easthampton, Massachusetts in 1862 and entered Williams College in the following year. There he studied under two famous American educators and philosophers, Mark Hopkins and John Bascom. Their influences shaped his interest in philosophy and theology, respectively. Finishing his bachelor's degree in 1867, Hall next decided to study for the ministry and entered Union Theological Seminary in New York. Yet, two years of study, some pastoral experience, and a frank discussion about his intellectual interests with the imposing American theologian Henry Ward Beecher resulted in Hall's first trip to Germany to study philosophy at the University of Berlin in 1869. In Friedrich Adolf Trendelenburg's seminary, he became an avowed Hegelian and contemplated becoming a philosopher. With the outbreak of the Frank-Prussian war, Hall returned to New York and finished his bachelor's degree in divinity in 1871. Yet, his intellectual and philosophy study there led him to embrace the evolutionary ideas of Charles Darwin and Herbert Spencer (Hall, 1923).

Forsaking the ministry then, Hall sought the academic life. He became a professor at Antioch College in Ohio, where he taught English, French, and German from 1872 to 1876. Becoming increasingly disenchanted with Hegel intellectually during this time, Hall openly became an enthusiast for Darwin, Spencer, and Huxley and positivistic science (Hall, 1923, p. 199, Ross, 1972, pp. 58–59). The turning point in his intellectual life occurred with his reading of Wilhelm Wundt's (1873) first volume of *Physiological Psychology*, after which he decided to specialize in psychology. Leaving Antioch, Hall completed a Ph.D at Harvard University in 1878 under William James; it was the first doctorate in psychology awarded in the USA. His dissertation explored the physiological relationships between muscles, nerves, and consciousness from a genetic perspective (Ross, 1972, pp. 70–75; Wilson, 1914, p. 64). Hall's research showed his fascination with experimental psychology and its links to Darwinian and Spencerian evolutionary thought. To bolster his research in the study of psychology, and because he could not obtain a suitable professorial appointment at a major university, Hall again went to Germany to study under Hermann von Helmholtz and William Wundt at the universities of Berlin and Leipzig, respectively (Goodchild, 1996, pp. 72–73).

Hall's Faculty Appointments at Johns Hopkins University from 1882 to 1887

After returning to the USA in 1880, Hall lectured at Harvard at the invitation of President Charles W. Eliot and then at

Hopkins by courtesy of President Daniel Coit Gilman. Later the Hopkins president appointed him a lecturer in 1882 and then Professor of Psychology and Pedagogics in 1884. From these positions, Hall furthered the study of education and psychology through a series of major articles exploring their relationships to human development. First, he surveyed pedagogical study in the USA and then explored the moral and religious training of both children and adolescents in one of his first discussions of child development. In analyzing the needs of education later, the professor stated boldly that 'we must study and follow the child's nature as it actually is' (Hall, 1883a, p. 286). Hall further proclaimed his new concept of 'natural' methods of education where education and psychology may both be utilized to understand the relationship between intelligence and human development in his important article, New Departures in Education (Hall, 1885a, pp. 147–148). Finally, he and one of his doctoral students outlined the entire field of education in their 1886 book, *Bibliography of Education* (Hall and Mansfield, 1886).

Second, in psychology, Hall quickly created one of the first experimental psychological laboratories in America at Hopkins and authored many of its early articles in experimental psychology (Boring, 1950; Hawkins, 1960, pp. 202–203). One of his important contributions was the publication of 'The new psychology' (1885b) which outlined this developing field. In 1887, he created a new research journal, the *American Journal of Psychology*, to publish psychological research. It was pivotal in establishing this new discipline. Third, his pedagogical and psychological ideas had already coalesced a few years earlier in his first major empirical studies using questionnaires on child development. 'The contents of children's minds' (1883b) and *The Study of Children* (1883c) described 'the real nature' of children, used the reports of actual teachers' observations from the questionnaires, and suggested appropriate pedagogical techniques. They played a major role in launching Hall's national reputation and his central position as the founder of the child study movement in the USA (White, 1990). Some of his doctoral students followed him as his career evolved; others, particularly John Dewey and James McKeen Cattell, were not favored and moved onto their own significant accomplishments (Ross, 1972, pp. 144–147). Thus Hall's work at Hopkins set the stage for the remainder of his scholarly career and publications as he pursued research in education, psychology, and their mutual implications for understanding human development.

Hall's Presidency of Clark University from 1888 to 1920

In Worcester, Massachusetts, wealthy industrialist Jonas Clark founded an institution of higher learning with a $1 million gift in 1887. He intended it to have a research mission for older male students and be a 'useful' college for local boys(Koelsch, 1987, pp. 9–12). Hall was named its first president in April 1888. He was determined to create a new graduate research university comparable to Hopkins, following the German ideal. The first three years saw considerable fulfillment of his vision. However, conflict with the founder over the college and loss of many faculty members to the new University of Chicago severely constrained the president. The struggling and much smaller institution still embraced the primacy of research as the reform ideal for the mission of all American universities (Goodchild, 1996; Goodchild and Miller, 1997). Eventually, Hall allowed an undergraduate college to be created when funds from the founder's estate became available. The president remained highly productive in his research and continued to oversee the institution until his retirement in 1920.

Hall's Prolific Scholarly Achievements during his Presidency

The president's accomplishments in psychology, education, and their implications for a new genetic interpretation of human development established the American prominence of these new scientific fields of study. In psychology, Hall founded the American Psychological Association in 1892, became its first president, and anchored his *American Journal of Psychology* to the organization. It remained the pre-eminent national association for this growing academic discipline. In 1904, Hall's most famous psychological study, *Adolescence: Its Psychology and Its Relations to Physiology, Anthropology, Sociology, Sex, Crime, Religion and Education*, appeared. This definitive two-volume work on adolescence comprised many of his earlier ideas on the child's next stage of development, established this stage as a scientific concept, related this developmental period to various human behaviors, and discussed its meaning for other emerging academic fields, especially education. Hall's interest in the sexual development of youth in this work was ground breaking. It led him to invite Sigmund Freud, Carl Jung, and many other leading psychologists to the famous 1909 twentieth Anniversary Conference of the founding of Clark University (Koelsch, 1987, pp. 246–247). As if to secure his role in this developing discipline, Hall published the valuable book, *Founders of Modern Psychology*, in 1912. After World War I, the scholarly president turned to his lifelong interest in religion and its psychological meaning. He authored a major reinterpretation of Christianity's founder in his *Jesus the Christ, in the Light of Psychology* (1917) and its application in the advocacy of a new life ethic in his *Morale: The Supreme Standard of Life and Conduct* (1920a) in which Hall espoused a holistic understanding of health and well-being as the rule for human conduct. With retirement from Clark's presidency in 1920, Hall reflected on his accomplishments in his *Recreations of a Psychologist* (1920b), and *Life and Confessions of a Psychologist* (1923). Hall achieved another scholarly triumph by completing his work on the last stage of human development with the publication of *Senescence: The Last Half of Life* (1922). In all this pioneering work, he made some significant missteps, especially claiming child development recapitulates human evolution (White, 1990, p. 31) and a too pervasive Darwinian and Spencerian 'evolutionism' (Ross, 1972, p. 92). On the other hand, he awarded the first doctorate in psychology to an African-American and encouraged Asian students to study psychology at Clark (Goodchild, 1996).

By his death on 24 April 1924, Hall had established psychology as an American discipline and study, almost entirely graduated the next generation of psychologists (Goodchild and Miller, 1997), and secured the role of behavioral and genetic psychology as being an integral part. To insure the

continuation of his work, Hall's estate established at Clark an endowment for a genetic psychology professorship and for his educational journal that was redirected and renamed *The Journal of Genetic Psychology*.

Similarly, his work in education commanded national attention. Hall believed that education was applied psychology. He followed identical approaches in establishing the importance of education as a field of study, as he had done in psychology. Hall offered a series of university lectures on kindergarten, elementary, secondary, and higher education (Goodchild, 1991). He next founded a research journal, *The Pedagogical Seminary: An International Record of Educational Literature, Institutions, and Progress,* in 1891. It provided a new research outlet from studies intended to reform education from American and international perspectives. Further, he extracted his pedagogical ideas from his books on adolescence and released them as *Youth: Its Education, Regimen, and Hygiene* (1906) to serve educators and teachers who were implementing his ideas in the classroom. Five years later, Hall presented his definitive study on education with another extensive two-volume work, *Educational Problems* (1911). He claimed to have fomented an American educational revolution where 'now the school, including its buildings, all its matter and method, revolve about the child' (p. 1: vi). This 'natural' education was to be individualized to the student's developmental stage, sex, and race. Thus pedagogy was to assist in the intellectual development of every person, although higher education was intended for the most able. At Clark, Hall had been instrumental in creating the study of education – particularly in conducting empirical research on children to develop better pedagogy, offering the first courses in higher education, and espousing genetic psychology as the basis for educational practice. His extensive publications in higher education focused on the mission of the new research university and its faculty (Goodchild, 1996). Similarly, missteps occurred in education as Hall advocated separate education for women, decried 'identical' coeducation, and espoused elitism in higher education (Ross, 1972, pp. 311–312). Nevertheless, the president greatly assisted in the arrival of education as an academic field of study (Cremin, 1961).

Hall's most significant contribution may be the creation of developmental psychology and its relationship to education (White, 1992). His research on child study, adolescence, and old age enabled him to argue for five stages of human development: (1) childhood; (2) adolescence; (3) middle life, from 25 to 45; (4) senescence, after 40; and (5) senectitude, approximately near the end of life (Hall, 1922, p. vii). His psychological and educational ideas and practices were focused on what he believed, would further the human race. His Darwinian and Spencerian concepts of evolution pervaded his understanding of human genetic development. He consistently maintained these ideals, even in the face of demands for coeducation in schools, growing behaviorism in psychology, and expansive progressivism in education. Hall was a man of his times, his scientific worldview centered on the mid- to late nineteenth century ideas and their implications for professional practice.

Finally, Hall's most important achievement when assessing the rise of the social and behavioral sciences was to engender the development of psychology and education from their philosophical progenitor. His patterns consisted of a critical assessment of the condition of the field first, next an assertion of a new position for the field, then a new research journal to publish these findings, and finally definitive works exploring the field of study. Hall thus played an essential role in the establishment of psychology and education as new academic fields of study in the USA. In 1992, the centenary meeting of the American

Psychological Association pointed to his pivotal role in these founding moments (Goodchild, 1992; White, 1992).

See also: Helmholtz, Hermann Ludwig Ferdinand von (1821–94); James, William (1842–1910); Life Course: Sociological Aspects; Lifespan Development, Theory of; Rousseau, Jean-Jacques (1712–78); Spencer, Herbert (1820–1903); Wisdom, Psychology of; Wundt, Wilhelm Maximilian (1832–1920); Youth Culture, Sociology of.

Bibliography

Boring, E.C., 1950. A History of Experimental Psychology, second ed. Appleton-Century-Crofts, New York.
Cremin, L.A., 1961. The Transformation of the School: Progressivism in American Education, 1876–1957. Vintage Books, New York.
Goodchild, L.F., 1991. Higher education as a field of study: its origins, programs, and purposes. In: Fife, J.D., Goodchild, L.F. (Eds.), Administration as a Profession. New Directions for Higher Education, no. 76. Jossey-Bass, San Francisco, pp. 1893–1960.
Goodchild, L.F., 1992. G. Stanley Hall's Psychological Avocation: Founding the Study of Higher Education. Research paper given at the 100th American Psychological Association annual meeting in Washington, DC.
Goodchild, L.F., 1996. G. Stanley Hall and the study of higher education. The Review of Higher Education 20, 69–99.
Goodchild, L.F., Miller, M.M., 1997. The American doctorate and dissertation: Six developmental stages. In: Goodchild, L.F., Green, K.E., Katz, E.L., Kluever, R.C. (Eds.), Rethinking the Dissertation Process: Tackling Personal and Institutional Obstacles. New Directions for Higher Education, no. 99. Jossey-Bass, San Francisco.
Hall, G.S., 1883a. Educational needs. North American Review 236, 284–290.
Hall, G.S., 1883b. The content of children's minds. Princeton Review 11, 249–272.
Hall, G.S., 1883c. The Study of Children. Printed privately.
Hall, G.S., 1885a. New departures in education. North American Review 140, 144–152.
Hall, G.S., 1885b. The new psychology. Andover Review 3, 120–136, 239–48.
Hall, G.S., Mansfield, J.M., 1886. Hints Toward a Select and Descriptive Bibliography of Education. Heath, Boston.
Hall, G.S., 1904. Adolescence: Its Psychology and Its Relations to Physiology, Anthropology, Sociology, Sex, Crime, Religion and Education, vol. 2. Appleton, New York.
Hall, G.S., 1906. Youth: Its Education, Regimen, and Hygiene. Appleton, New York.
Hall, G.S., 1911. Educational Problems, vol. 2. Appleton, New York.
Hall, G.S., 1912. Founders of Modern Psychology. Appleton, New York.
Hall, G.S., 1917. Jesus, The Christ, in the Light of Psychology. Doubleday Page, New York.
Hall, G.D., 1920a. Morale: The Supreme Standard of Life and Conduct. Appleton, New York.
Hall, G.S., 1920b. Recreations of a Psychologist. Appleton, New York.
Hall, G.S., 1922. Senescence: The Last Half of Life. Appleton, New York.
Hall, G.S., 1923. Life and Confessions of a Psychologist. Appleton, New York.
Hawkins, H., 1960. Pioneer: A History of Johns Hopkins University, 1874–1889. Cornell University Press, Ithaca, NY.
Koelsch, W.A., 1987. A Narrative History. Clark University Press, 1887–1987. Clark University, Worcester, MA.
Lagemann, E.C., 2000. An Elusive Science: The Troubling History of Educational Research. University of Chicago Press, Chicago.
Mitchell, T.R., Haro, A., 1999. Poles apart: Reconciling the dichotomies in educational research. In: Lagemann, E.C., Shulmlan, L.S. (Eds.), Issues in Educational Research: Problems and Possibilities. Jossey-Bass, San Francisco.
Ross, D., Stanley Hall, G., 1972. The Psychologist as Prophet. University of Chicago Press, Chicago.

Ross, D., 1979. The development of the social sciences. In: Oleson, A., Voss, J. (Eds.), The Organization of Knowledge in Modern America. Johns Hopkins University Press, Baltimore, MD, pp. 1860–1920.

White, S.H., 1990. Child study at Clark University: 1894–1904. Journal of the History of the Behavioral Sciences 26, 31–50.

White, S.H., 1992. G. Stanley Hall: From philosophy to developmental psychology. Developmental Psychology 28, 25–34.

Wilson, L.N., 1914. G. Stanley Hall: A Sketch. Stechert, New York.

Wundt, W., 1873. Principles of Physiological Psychology. Engelmann, Leipzig, Germany.

Hallucinations, Neural Basis of

Rachael Goldsworthy, Central Michigan University, Mt. Pleasant, MI, USA
Harry Whitaker, Northern Michigan University, Marquette, MI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R.E. Hoffman, volume 10, pp. 6469–6473, © 2001, Elsevier Ltd.

Abstract

Hallucinations are perceptions that occur in the absence of external stimuli, and have sufficient realistic quality to resemble actual perception of external phenomena. This article presents an overview of several proposed neural mechanisms for the production of hallucinations, such as deafferentation, release theory, misinterpretations of internal input, and deficits in top-down processing. The most common types of hallucinations are reviewed in detail (i.e., visual and auditory phenomena). Hallucinogenic drugs, such as lysergic acid diethylamide, mescaline, and psilocybin, are known to produce visual and auditory distortions; however, withdrawal from certain pharmacological substances may also produce hallucinations. Psychiatric illnesses and the associated hallucinatory phenomena are also discussed.

Hallucinations are perceptions that occur in the absence of external stimuli, and have sufficient realistic quality to resemble actual perception of external phenomena. These perceptions are not under voluntary control. Hallucinations can occur in any sensory modality, and differ from delusions. Although hallucinations occur in various psychiatric conditions, around 6% of the general population reported having hallucinations at least once a month (Ohayon, 2000). The phenomenon can be distinguished from that of *illusions*, which are misinterpretations of stimuli that are actually present. Gilles Fénelon (2013) categorizes the mechanisms of hallucinations into four groups. The hallucinations occurring in epilepsy and migraines result from spontaneous overactivation in sensory areas. Deafferentation (the loss or destruction of sensory input) facilitates hallucinatory experiences by leading to a chronic hyperexcitable state in cortical areas. A third mechanism is attributed to the dissociation of the dream and sleep state. Lastly, hallucinations can arise from dysfunction in bottom-up processing and top-down mechanisms. *Pseudo-hallucinations* refer to hallucinations with preserved insight, or the knowledge that one is hallucinating (Fénelon, 2013: 59). The major types of visual hallucinations are described below.

Visual Hallucinations

Elementary visual hallucinations are more common than complex shapes and human figures (Billock and Tsou, 2012). There are three types of elementary hallucinations. *Phosphenes* are the simplest form, and consist of small spots of light. *Geometric figures* are the second type, and consist of lattices, pinwheels, and spiral images. *Fortifications* are bright flashing arc patterns that may appear zigzagged (Billock and Tsou, 2012). Complex hallucinations consist of images such as objects, people, and landscapes. During *palinopsia*, an image remains in view after the perceived stimulus is no longer present. The colors of these persistent images can be the same as the original stimulus, inverted, or can appear to have 'trails.' Palinopsia has been localized in the parietal projections to the occipital cortex (Smith et al., 2003). A person is aware that visual hallucinations are not real during *peduncular hallucinosis*, which consists of brief, vivid, colorful representations of geometric patterns, people, animals, and scenes. These hallucinations do not have symbolic significance and are not associated with delusional thinking. They have commonly been associated with lesions in the midbrain and the reticular formation (Parisis et al., 2003). *Hypnagogic* hallucinations occur during the wake–sleep phase, whereas *hypnopompic* hallucinations occur during the sleep–wake phase; both are frequent in nonclinical populations (Fénelon, 2013: 59). These hallucinations have been associated with attentional biases occurring in the anterior cingulate gyrus (Fénelon, 2013: 59). Body misperceptions may consist of both out-of-body experiences and *autoscopic* hallucinations. Autoscopic hallucinations involve seeing an image of one's self in external space from his or her physical point of view. Out-of-body experiences have been associated with the right temporo-parietal cortex, whereas autoscopic hallucinations are associated with both temporo-occipital cortices. *Heautoscopy* refers to seeing a double of oneself in external space, but having difficulty differentiating in which body the self is situated (Fénelon, 2013: 59). *Heautoscopy* has been associated with the left temporo-parietal cortex.

Charles Bonnet syndrome (CBS) refers to visual hallucinations that occur throughout the course of an eye disease; these visual hallucinations include palinopsia (the appearance of a trailing object) and *polyopia*, in which a stimulus becomes multiplied (Ffytche and Howard, 1999). Elementary hallucinations are more common, but complex visual hallucinations can include people, animals, and complex scenes. Interestingly, visual hallucinations reported by persons with CBS include *lilliputian figures*, or small, doll-like people (Ffytche, 2007). Although the type of eye disease is not relevant, the amount of cognitive impairment appears to contribute to the presence of visual hallucinations. Traditionally, postulated neural mechanisms for Charles Bonnet syndrome are that a loss of visual stimuli leads to cortical inhibition, which creates visual hallucinations. More recent theories by Burke (2002) stress the importance of the molecular changes that take place in deafferented neurons which create the bursts of activity, not necessarily an interruption of inhibitory processes.

Auditory Hallucinations

Auditory hallucinations can be simple sounds (e.g., knocks, rings, and tinnitus), or can consist of complex sounds, such as music or voices. When these hallucinations occur outside of schizophrenia, they are rarely threatening or shaming in nature (Fénelon, 2013: 59). Auditory hallucinations can vary in location (coming from inside or outside of the head), voice familiarity, loudness, and content (Stephane, 2013). Analogous to visual hallucinations, auditory hallucinations may arise from an impairment in hearing. Hallucinations associated with hearing impairment are typically elementary sounds, such as tinnitus, but may also be voices or music (Cole et al., 2002). The musical hallucinations that occur in deaf individuals are considered to be an equivalent of the visual hallucinations reported in CBS. Musical hallucinations are usually reported in acquired deafness, but have also been associated with depression, obsessive–compulsive disorder, schizophrenia, and lesions in the brainstem or cerebral hemispheres (Stewart et al., 2006). The physiology of auditory hallucinations associated with deafness is less clear; PET scans have indicated that the same neural networks are activated during hallucinations as in normal musical perception (Stewart et al., 2006).

A number of variables are associated with different neural bases for auditory hallucinations, such as *anosognosia* (lack of awareness of the hallucination), content, spatial location of the hallucination (inside or outside of the head), and voice familiarity. Anosognosia is associated with lesions of the frontal lobes, the visual associative cortex, and specific neural substrates (Stephane, 2013). The right temporoparietal junction has been associated with the spatial location of hallucinations (Plaze et al., 2011). Larøi and Woodward (2007) have theorized that auditory hallucinations may result from deficits in understanding the boundaries between self–other and inner space–outer space. Stephane et al. (2006) has found that abnormal laterality in motor areas of the brain may be attributable to the failure to discriminate whether speech was generated internally or externally. Activation in Wernicke's area may also account for the actual perception of auditory hallucinations (Stephane et al., 2006).

Hallucinations in Other Sensory Modalities

Although they are less common than auditory and visual hallucinations, hallucinations can consist of olfactory, gustatory, and somatosensory content. Olfactory hallucinations are usually reported during temporal lobe seizures or migraine auras, typically as a result of damage to the olfactory bulb. The smells usually comprise unpleasant odors, such as the smell of burning material, decaying flesh, or other rancid odors, although smells of perfume, flowers, and food have also been reported (Stevenson et al., 2011). Olfactory hallucinations have been noted more often in late-onset schizophrenia, and may influence delusional beliefs that the schizophrenic individual smells of the foul odor (Hoffman, 2003). Interestingly, olfactory hallucinations are not well explained by theories of visual and auditory hallucinations. Visual and auditory phenomena are explained by misattributions of internal versus external world phenomena due to impairment in attentional and metacognitive processing (Bentall, 1990). The same theory does not apply to olfactory hallucinations; it is difficult to imagine forming a mental image of a certain smell and misattributing it to an external stimulus (Stevenson et al., 2011).

Tactile hallucinations may consist of sensations that one's skin is crawling with insects, that is, *formication*. The phantom-limb hallucination, or the feeling that an amputated limb is still intact (and possibly painful), is a common somatosensory hallucination following surgery, although the experience of phantom-limb pain may occur years after amputation. Gustatory hallucinations have been reported in patients experiencing mood disorders, drug ingestion, and sleep deprivation (Ohayon, 2000). The experience of olfactory and gustatory hallucinations occurs as often together as they do separately (Lewandowski et al., 2009). Similar to olfactory hallucinations, gustatory hallucinations typically involve the sensation of tasting something unpleasant.

Psychiatric Illnesses Associated with Hallucinations

Elementary hallucinations are more commonly associated with the effects of migraine auras, epilepsy, hypoglycemia, and eye diseases (e.g., CBS). Complex hallucinations are more common in partial epilepsy, Parkinson's disease, Alzheimer's disease, and dementia with Lewy bodies (DLB) (Fénelon, 2013: 59). Hallucinations are generally associated with the positive symptoms of *schizophrenia*. Auditory hallucinations associated with schizophrenia are often voices, and may contain shaming or derogatory content. The hallucinations may vary from single words to complex sentences, although single words are reported more frequently (Larøi et al., 2012). These voices can sound familiar to the way other voices sound, or may be perceived as unreal (Larøi et al., 2012). Some individuals may find it difficult to differentiate whether the voice is coming from within the head or outside of the body. Functional connectivity studies of auditory hallucinations in schizophrenic populations suggest that the mechanism involves a complex pathway between Wernicke's area, the left inferior frontal cortex, and the putamen (Hoffman and Hampson, 2012). Interestingly, the putamen is involved both in determining whether auditory stimuli are consciously perceived and in the production of spontaneous language (Hoffman and Hampson, 2012). While functional neuroimaging studies report different activation areas, it has generally been found that speech generation and perception areas in both the left hemisphere and their corresponding right hemisphere are involved. The initiation of such hallucinations is theorized to be attributed to activity in the parahippocampal gyrus (Aleman and Vercammen, 2012: 267). Visual hallucinations in schizophrenic patients are associated with increased hippocampal activity (Larøi et al., 2012).

In partial *epilepsy*, the following hallucinations are most common, in decreasing order of prevalence: complex visual, somatosensory, auditory, gustatory, and olfactory hallucinations (Mauguiére, 1999). Autoscopic visual hallucinations and out-of-body experiences are often reported in epilepsy. The majority of hallucinations associated with occipital lobe epilepsy are elementary hallucinations, such as rotating lights

or flashing colors (Bauer and Hamer, 2012: 241). Complex visual hallucinations have also been reported, consisting of seeing people, objects, or some visual alteration of one's environment (Guimond et al., 2009). Hallucinations reported in parietal lobe epilepsy include both elementary and complex visual content. Temporal lobe seizures, on the other hand, have typically been associated with gustatory and olfactory hallucinations as well as with visual hallucinations. Phosphenes, sparks, and other simple visual hallucinations are reported in patients who have experienced occipital lobe seizures (Lhatoo and Sander, 2003: 207). Ictal auditory hallucinations are both simple and complex; simple hallucinations have been associated with activity in the primary auditory cortex (Fénelon, 2013: 59). It has been shown that patients experiencing verbal auditory hallucinations have greater left-hemispheric activation, whereas nonverbal auditory ictal hallucinations have more right-hemispheric activation. This has been demonstrated regardless of whether the hallucinations are heard in one or both ears (Guimond et al., 2009). Hallucinations experienced may be a combination of more than one sensory modality; however, olfactory and gustatory sensations are not common (Fénelon, 2013: 59).

A 'classic migraine' is accompanied by an *aura*, which is a term for the accompanying neurological symptoms that either precede or appear congruently with the migraine attack (Silberstein, 2003: 174). To meet criteria for a classic migraine, an individual must experience a migraine during an aura or within 60 min after the aura has begun; in addition, reversible visual distortions or hallucinations are present. Elementary hallucinations are more common in migraine auras, followed by somatosensory hallucinations (Pietrobon and Striessnig, 2003). Common elementary hallucinations include geometric patterns, phosphenes, and fortifications (Billock and Tsou, 2012). Complex visual distortions may include Alice in Wonderland-like perceptions (more frequently seen in children), and *teichopsia* (an arc of light that appears as a zigzag pattern, similar to the appearance of fortifications seen on castles) (Silberstein, 2003: 174). Olfactory hallucinations are also a common aura symptom. Compared with epileptic seizures, which only last several seconds, aura symptoms may last for many minutes. The length of hallucinatory experiences provides valuable differential diagnostic information between epilepsy and migraines (Billock and Tsou, 2012).

In *Parkinson's disease*, the following hallucinations are most common, in decreasing order of prevalence: complex visual, auditory, tactile, and olfactory hallucinations (Fénelon et al., 2011). Complex hallucinations consist of people, objects, or animals; they typically last for only seconds, are usually superimposed onto the actual environment, and may be relatively stereotyped for the individual (Fénelon, 2013: 59). In the majority of cases, auditory hallucinations are heard externally and do not contain affective components or derogatory content. When multimodal hallucinations are present (e.g., both auditory and visual hallucinations), there is typically no relationship between their content (Inzelberg et al., 1998). Models of hallucinations in Parkinson's disease comprise deficits in both bottom-up and top-down processing of sensory information, as well as intrusions of internal phenomena (e.g., dreams) during consciousness (Fénelon, 2013: 59). It has been speculated that some hallucinations associated with Lewy body disorders may result from dissociation between the dream and the wake state, or the intrusion of internal phenomena misinterpreted as external stimuli (Fénelon, 2013: 59). Hallucinations typically occur several years after onset of Parkinson's disease.

In *DLB*, the following hallucinations are most common, in decreasing order of prevalence: complex visual, auditory, tactile, and olfactory hallucinations (McKeith, 2002). The diagnosis of DLB is based upon the presence of Lewy bodies, which are intracellular hyaline inclusions that are found in some patients suffering from Parkinson's disease (Goral et al., 2006: 16). Lewy bodies are found in cortical and subcortical regions of the brain. Characteristics of DLB include a decline in cognitive functioning, visual hallucinations, and spontaneous parkinsonism (Goral et al., 2006: 16). Hallucinations associated with DLB are similar to those observed in Parkinson's disease, although they typically have a shorter timecourse for onset (Fénelon, 2013: 59). Recurrent visual hallucinations in both DLB and Parkinson's disease tend to be complex rather than elementary; this may reflect a mechanism of hallucinations that is disease dependent and affected by cortical areas, specifically the dorsal and ventral pathways (Mosimann et al., 2006). Visual hallucinations are more complex in DLB than that experienced in Alzheimer's disease; the complexity provides a useful diagnostic distinction between the two (Swanberg and Cummings, 2003: 207). The hallucinations are more vivid and contain complex images of objects, animals, or people that are normal in size (Swanberg and Cummings, 2003: 207). In *Alzheimer's disease*, the following hallucinations are most common, in decreasing order of prevalence: complex visual, auditory, tactile, and olfactory hallucinations (Ropacki and Jeste, 2005). Visual hallucinations typically include people, animals, and objects. Auditory hallucinations are less common, and olfactory hallucinations are quite rare (Fénelon, 2013: 59). Delusions and hallucinations, especially visual hallucinations, have been shown to be related to greater cognitive impairment and rates of cognitive decline in Alzheimer's patients (Wilson et al., 2006). The supposed mechanisms are still speculative, but it is theorized that hallucinations in dementia of the Alzheimer's type result from deficits in top-down processing and visual sensory processing (Fénelon, 2013: 59).

Diagnostic criteria for *delirium* include a disturbance in consciousness that is accompanied by several cognitive impairments; delusions and hallucinations are also common. Visual hallucinations are most commonly reported, followed by auditory and somatosensory misperceptions (Folstein and Caplan, 2003: 207). Severe *mood disorders*, such as a major depression, can be accompanied by hallucinations. Auditory misperceptions are the most commonly reported hallucination, with single words and short phrases heard. Patients with bipolar disorder may experience hallucinations that contain both verbal and nonverbal auditory contents; these hallucinations may be inconsistent throughout mood phases (Larøi and Woodward, 2007). These hallucinations are often accompanied by delusions and may contain themes of grandiosity, a hallmark symptom of a manic episode. The prevalence of hallucinations in the context of *borderline personality disorder* is less clear; it is suggested that the hallucinations are short lasting compared to schizophrenic hallucinations, and are often

perceived as coming from within one's head, but out of one's control (Larøi and Woodward, 2007).

Pharmacology of Hallucinations; Drug-Induced States

Hallucinogenic drugs, or psychedelics, are chemicals that produce the psychotic experiences of hallucinations, alterations in cognition, and grandiosity or paranoia (Goetz, 2003: 174). The user experiences misperceptions that arise from the dissociation between internal and external phenomena. The most common hallucinogen in the United States is *lysergic acid diethylamide* (LSD), which produces powerful effects that may persist for several hours. Initial feelings of grandiosity and pleasure may be replaced by feelings of anxiety or panic as the drug's effect progresses. Hallucinations associated with LSD include reports of intense, vivid colors, and elementary hallucinations (Goetz, 2003: 174). In general, effects caused by hallucinogens are more likely to be visual misperceptions, although auditory tactile hallucinations have been reported (Larøi and Woodward, 2007). Mescaline (peyote cactus) and psilocybin (mushrooms) are hallucinogens that have been used as part of Native American religious ceremonies (Goetz, 2003: 174).

Cocaine, cyclic antidepressants, cannabis (tetrahydrocannabinol, THC), MDMA (methylenedioxymethamphetamine), carbamazepine, and salicylates have been known to produce hallucinatory experiences (Von Rosen and Brust, 2003: 207). In high doses, anticholinergic drugs may cause hallucinations and delirium, which may be misattributed to the presence of a psychiatric disorder (Fann, 2006: 354). Hallucinations may also be triggered by withdrawal from substances. Patients with chronic alcoholism may experience hallucinations during withdrawal (Hoffman, 2003: 500). Withdrawal from cocaine, amphetamines, cannabis, and LSD may also be accompanied by hallucinations (Larøi and Woodward, 2007). Feelings of formication (i.e., the sensation of one's skin crawling with insects) have been associated with withdrawal from alcohol, amphetamine, and cocaine (Larøi and Woodward, 2007).

Proposed Mechanisms of Hallucinations

The remainder of the article serves as an overview of the different neural mechanisms of hallucinations. Hallucinations associated with seizures or migraines with auras are assumed to be the product of overactivation that spontaneously results in sensory cortical areas (Fénelon, 2013: 59). *Ictal* hallucinations refer to hallucinations that occur during an aura or partial epilepsy; however, the same effect can be achieved by cortical stimulation. For example, stimulation of the V1 area of the cortex has been known to produce simple visual hallucinations in the form of colored phosphenes; stimulation outside of this area leads to the experience of colorless phosphenes (Billock and Tsou, 2012). In addition, the type of psychological hallucination has been shown to be dependent on which hemisphere is activated. It is theorized that ictal hallucinations are caused by an activation of brain tissue near that area; as such, ictal phenomena may provide a route to further examine hemispheric specialization (Guimond et al., 2009).

Deafferentation, also known as the loss/disruption of sensory input, may also produce hallucinations. Deafferentation in any sensory modality has the ability to facilitate hallucinations in that sense (Fénelon, 2013: 59). This type of mechanism produces the hallucinations seen in CBS and, more classically, phantom-limb phenomena. When a limb is amputated, it is usually followed by a feeling that the limb is still intact, a sensation known as 'phantom-limb awareness'; this sensation is reported by almost all amputees (Flor et al., 2006). The sensation may be rare and short lived, or may be a chronic sensation of pain during which the individual perceives the missing limb. It is theorized that phantom limb sensations are experienced due to a restructuring of somatosensory maps in the cortex (Flor et al., 2006).

When a region of the brain is deafferented, the cells within this region are more excitable; this can occur in any region associated with sensory perception. The deafferentated neurons undergo biological changes that cause hyperexcitability (Burke, 2002). These neurons experience periods of inactivity interspersed with bursts of excitability that resemble seizures (Burke, 2002). The *release* theory of hallucinations by West (1962) states that hallucinations arise due to an impairment of a sensory modality or its connecting pathway, coupled with neuronal memories of past perceptions that emerge into consciousness. When the release theory is applied to visual hallucinations that result from vision loss/impairment, it has been hypothesized that bottom-up processing (which is usually driven by visual stimuli) inhibits spontaneous activity. This bottom-up processing simultaneously facilitates the release of these stored images, producing a visual hallucination (Cogan, 1973). It has also been argued that the excitation is simply a result of a lack of input to these neurons due to deafferentation (Fénelon, 2013: 59).

Disturbances in *top-down processing* have also been theorized to produce hallucinatory phenomena. Adaptive resonance theory states that hyperexcitability in top-down processes, specifically processes involving attention, contribute to the perception of hallucinations by occasionally activating images (Grossberg, 2000). The Perception and Attention Deficit (PAD) model assumes that hallucinations (specifically recurrent complex visual hallucinations) are caused by impairments in attention and perception combined with scene representations that are typical for an environment, but are not actually present (Collerton et al., 2005). According to PAD theory, biases in perception allow the experience of a hallucination because the individual assumes an object to be part of an environment, poor sensory activation, and a functioning scene representation.

Hallucinations may also result from the *misinterpretation of internal input* as coming from an external source. Internal input, such as a dream, may be experienced and misinterpreted during the waking state, causing hallucinatory phenomena (Fénelon, 2013: 59). The relationship between waking state, REM sleep, and non-REM sleep may be less distinct than once thought, which provides further evidence for the occurrence of dream-like hallucinations during consciousness (Mahowald and Schenck, 2001). Hallucinations may occur during the transition between falling asleep and waking, and have been reported in nonclinical populations (Fénelon, 2013: 59). During sleep, dreams are experienced because the brain is attending to

internal activity. In some circumstances, such as the sleep–wake transition state, sleep deprivation, or drug-induced states, the brain attends to both internal and external information, resulting in hallucinations (Mahowald et al., 1998).

See also: Aging and Memory; Alzheimer's Disease, Neural Basis of; Epilepsy; Phantom Limbs; Prosopagnosia.

Bibliography

Aleman, A., Vercammen, A., 2012. Functional neuroimaging of hallucinations. In: Blom, J.D., Sommer, I.E.C. (Eds.), Hallucinations: Research and Practice. Springer, New York, pp. 267–281.

Bauer, S., Hamer, H.M., 2012. Extratemporal epilepsies. In: Handbook of Clinical Neurology: Epilepsy, Part I. Elsevier Sciences, USA, pp. 241–256.

Bentall, R.P., 1990. The illusion of reality: a review and integration of psychological research on hallucinations. Psychological Bulletin 107 (1), 82–95.

Billock, V.A., Tsou, B.H., 2012. Elementary visual hallucinations and their relationships to neural pattern-forming mechanisms. Psychological Bulletin [online] 138 (4), 744–774. Available from: www.proquest.com (10.10.13.).

Burke, W., 2002. The neural basis of Charles Bonnet hallucinations: a hypothesis. Journal of Neurology, Neurosurgery and Psychiatry 73 (5), 535–542.

Cogan, D.G., 1973. Visual hallucinations as release phenomena. Graeffe's Archive for clinical and experimental ophthalmology 188, 139–150.

Cole, M.G., Dowson, L., Dendukuri, N., Belzile, E., 2002. The prevalence and phenomenology of auditory hallucinations among elderly subjects attending an auditory clinic. International Journal of Geriatric Psychiatry 17, 444–452.

Collerton, D., Perry, E., McKeith, I., 2005. Why people see things that are not there: a novel perception and attention deficit model for recurrent complex visual hallucinations. Behavioral and Brain Sciences 28, 737–794.

Fann, W.E., 2006. Pharmacology. In: Encyclopedia of Gerontology, second ed. Elsevier, Oxford, pp. 354–365.

Fénelon, G., 2013. Hallucinations associated with neurological disorders and sensory loss. In: Jardri, R., Cachia, A., Thomas, P., Pins, D. (Eds.), The Neuroscience of Hallucinations. Springer, New York, pp. 59–83.

Fénelon, G., Soulas, T., Cleret de Langavant, L., Trinkler, I., Bachoud-Lévi, A.C., 2011. Feeling of presence in Parkinson's disease. Journal of Neurology, Neurosurgery, and Psychiatry 82 (11), 1219–1224.

Ffytche, D.H., Howard, R.J., 1999. The perceptual consequences of visual loss: positive pathologies of vision. Brain 122, 1248–1260.

Ffytche, D.H., 2007. Visual hallucinatory syndromes: past, present and future. Dialogues in Clinical Neuroscience 9, 173–189.

Flor, H., Nikolajsen, L., Jensen, T.S., 2006. Phantom limb pain: a case of maladaptive CNS plasticity? Neuroscience: Nature Reviews 7, 873–893.

Folstein, M., Caplan, L.R., 2003. Delirium. In: Brandt, T., Caplan, L.R., Dichgans, J., Diener, H.C., Kennard, C. (Eds.), Neurological Disorders: Course and Treatment, second ed. Elsevier Science, USA, pp. 207–234.

Goetz, C.G., 2003. Hallucinogens. In: Encyclopedia of the Neurological Sciences, first ed. Elsevier Science, USA, pp. 174–178.

Goral, M., Clark-Cotton, M.R., Albert, M.L., 2006. Language disorders: general. In: Encyclopedia of Gerontology, second ed. Elsevier, Oxford, pp. 16–24.

Grossberg, S., 2000. How hallucinations may arise from brain mechanisms of learning, attention, and volition. Journal of the International Neuropsychological Society 6, 583–592.

Guimond, A., Braun, C.M.J., Daigneault, R., 2009. Ictal hallucination and hemispheric specialization: findings from 217 cases with a unilateral epileptic focus. Biological Psychology 82, 133–148.

Hoffman, R.E., 2003. Hallucinations. In: Encyclopedia of the Neurological Sciences, first ed. Elsevier Science, USA, pp. 500–501.

Hoffman, R.E., Hampson, M., 2012. Functional connectivity studies of patients with auditory verbal hallucinations. Frontiers in Human Neuroscience 6, 1–7.

Inzelberg, R., Kipervasser, S., Korczyn, A.D., 1998. Auditory hallucinations in Parkinson's disease. Journal of Neurology, Neurosurgery and Psychiatry 64, 533–535.

Lewandowski, K.E., DePaola, J., Camsari, G.B., Cohen, B.M., Öngür, D., 2009. Tactile, olfactory, and gustatory hallucinations in psychotic disorders: a descriptive study. Annals of the Academy of Medicine, Singapore 38 (5), 383–388.

Larøi, F., Sommer, I.E., Blom, J.D., Fernyhough, C., ffytche, D.H., Hugdahl, K., Johns, L.C., McCarthy-Jones, S., Preti, A., Raballo, A., Slotema, C.W., Stephane, M., Waters, F., 2012. The characteristic features of auditory verbal hallucinations in clinical and nonclinical groups: state-of-the-art overview and future directions. Schizophrenia Bulletin 38 (4), 724–733.

Larøi, F., Woodward, T.S., 2007. Hallucinations from a cognitive perspective. Harvard Review of Psychiatry 15 (3), 109–117.

Lhatoo, S.D., Sander, J.W.A.S., 2003. The epilepsies. In: Brandt, T., Caplan, L.R., Dichgans, J., Diener, H.C., Kennard, C. (Eds.), Neurological Disorders: Course and Treatment, second ed. Elsevier Science, USA, pp. 207–234.

Mahowald, M.W., Schenck, C.H., 2001. Evolving concepts of human state dissociation. Archives Italiennes de Biologie 139, 269–300.

Mahowald, M.W., Woods, S.R., Schenck, C.H., 1998. Sleeping dreams, waking hallucinations, and the central pervous system. Dreaming 8 (2), 89–103.

Mauguiére, F., 1999. Scope and presumed mechanisms of hallucinations in partial epileptic seizures. Epileptic Disorders 1, 81–91.

McKeith, I.G., 2002. Dementia with Lewy bodies. The British Journal of Psychiatry 180, 144–147.

Mosimann, U.P., Rowan, E.N., Partington, C.E., Collerton, D., Littlewood, E., O'Brien, J.T., Burn, D.J., McKeith, I.G., 2006. Characteristics of visual hallucinations in Parkinson disease dementia and dementia with Lewy bodies. Journal of Geriatric Psychiatry 14 (2), 153–161.

Ohayon, M.M., 2000. Prevalence of hallucinations and their pathological associations in the general population. Psychiatry Research 97, 153–164.

Parisis, D., Poulios, I., Karkavelas, G., Drevelengas, A., Artemis, N., Karacostas, D., 2003. Peduncular hallucinosis secondary to brainstem compression by cerebellar metastases. European Neurology 50 (2), 107–110.

Pietrobon, D., Striessnig, J., 2003. Neurobiology of migraine. Neuroscience: Nature Reviews 4, 386–398.

Plaze, M., Paillére-Martinot, M.L., Penttilä, J., Januel, D., de Beaurepaire, R., Bellivier, F., Andoh, J., Galinowski, A., Gallarda, T., Artiges, E., Olié, J.P., Mangin, J.F., Martinot, J.L., Cachia, A., 2011. "Where do auditory hallucinations come from?" A brain morphometry study of schizophrenia patients with inner or outer space hallucinations. Schizophrenia Bulletin 37 (1), 212–221.

Ropacki, S.A., Jeste, D.V., 2005. Epidemiology of and risk factors for psychosis of Alzheimer's disease: a review of 55 studies published from 1990 to 2003. The American Journal of Psychiatry 162, 2022–2030.

Silberstein, S.D., 2003. Migraine with aura. In: Encyclopedia of the Neurological Sciences, first ed. Elsevier Science, USA, pp. 174–178.

Smith, P.E.M., Shah, P., Sharpe, J., Todd, A., Goringe, A.P., 2003. Case report: palinopsia. The Lancet 361, 9363.

Stephane, M., 2013. Auditory verbal hallucinations result from combinatoric associations of multiple neural events. Frontiers in Human Neuroscience 7, 1–8.

Stephane, M., Hagen, M.C., Lee, J.T., Uecker, J., Pardo, P.J., Kuskowski, M.A., Pardo, J.V., 2006. About the mechanisms of auditory verbal hallucinations: a positron emission tomographic study. Journal of Psychiatry and Neuroscience 31 (6), 396–406.

Stevenson, R.J., Langdon, R., McGuire, J., 2011. Olfactory hallucinations in schizophrenia and schizoaffective disorder: a phenomenological survey. Psychiatry Research 185, 321–327.

Stewart, L., von Kriegstein, K., Warren, J.D., Griffiths, T.D., 2006. Music and the brain: disorders of musical listening. Brain 129, 2533–2553.

Swanberg, M.M., Cummings, J.L., 2003. Diagnosis and treatment of dementia. In: Brandt, T., Caplan, L.R., Dichgans, J., Diener, H.C., Kennard, C. (Eds.), Neurological Disorders: Course and Treatment, second ed. Elsevier Science, USA, pp. 207–234.

Von Rosen, F., Brust, J.C.M., 2003. Acute intoxications. In: Brandt, T., Caplan, L.R., Dichgans, J., Diener, H.C., Kennard, C. (Eds.), Neurological Disorders: Course and Treatment, second ed. Elsevier Science, USA, pp. 207–234.

West, L.J., 1962. Hallucinations. Grune & Stratton, New York.

Wilson, R.S., Tang, Y., Aggarwal, N.T., Gilley, D.W., McCann, J.J., Bienias, J.L., Evans, D.A., 2006. Hallucinations, cognitive decline, and death in Alzheimer's disease. Neuroepidemiology 26, 68–75.

Handedness and Language, Relationship between

Brigitte Stemmer, Université de Montréal, Montreal, QC, Canada, Brock University, St. Catherines, ON, Canada and CRIUGM, Montreal, QC, Canada

Published by Elsevier Ltd.

Abstract

Most people are right handed with the left hemisphere controlling hand movement. This coincides with aspects of language being lateralized to the left hemisphere. Does this mean that handedness and language representation in the brain are in some way linked? How did handedness develop in evolutionary terms and developmentally? Is there a handedness gene, maybe even common for handedness and language? Is there a relation between the emergence of language and development of language before and after birth? There is a common belief that left-handers have language represented in the right brain. Is this myth or truth? More males are left-handed than females. Is this in any way related to the way language is lateralized in males and females? Left handedness has been associated with certain language disorders. What is the scientific evidence? Looking for an answer to these questions is not – as one might assume – as easy as it seems.

Introduction

Most people would probably agree that human beings look pretty much symmetrical. A closer look, however, reveals that symmetry can be quite deceiving. An important question to ask is *what* is symmetrical – the anatomy, the function, or both? Although our arms and legs may look anatomically symmetrical, one of our members may in fact be shorter than the other (anatomical asymmetry) and most people have a preferred arm or leg to throw or kick a ball with (functional asymmetry). A similar thing can be said about the brain. Although the two hemispheres may look pretty much the same to the inexperienced naked eye, they differ anatomically as well as functionally. Genetically determined anatomical brain asymmetry may be the reason for functional asymmetries observed such as language. When asked which hemisphere is dominant for language, most people would say the left hemisphere – but again, things are not that simple. When distinguishing between different speech sounds, the left hemisphere is dominantly involved while judging whether a sentence is grammatically correct involves both hemispheres. Many people would also say that the left hemisphere is dominant for language in right-handers while it is the right hemisphere for left-handers. However, this is a common misconception. Although reports vary somewhat on the exact number, research has consistently shown that about 90% of people are right-handed and about 10% are left-handed, and in the majority of right- *and* left-handers certain aspects of language are represented and/or processed in the left hemisphere. Although research supports the notion that there is some kind of relation between handedness, anatomical brain asymmetries, and the side of the brain to which language is lateralized (also referred to as (functional) language lateralization), it is far from clear what exactly this relation is.

Broca's View on Handedness and Language Dominance

In the nineteenth century, the French neurologist Paul Broca published his seminal paper in which he stated that the two hemispheres of the brain were (anatomically) perfectly similar with incidental variations from person to person but that they had different 'attributes' with language being represented and processed in the left hemisphere (Broca, 1865 translated in Berker et al., 1986). At the time, this contradicted the dominant view that the two hemispheres were functionally similar. To argue his case, Broca wanted to show that functional asymmetry of something as complex as language was, indeed, not such an outrageous exception and used right handedness as an example for a motor function dominated by the left hemisphere. He observed that most people were right-handed and only few left-handed. Further, functions that required special training such as writing, drawing, or playing a musical instrument required a division of labor between both hands with the most difficult part, "that which requires the greatest ability, precision, expression, or strength … is very generally assigned to the right hand. (…) if we consider only those that we carry out spontaneously, such as throwing a ball, hitting with our fists, handling a stick, or lifting a weight, we find that with the exception of a very small number of persons, designated as left-handed, everybody naturally uses their right hand to which the left hand is nothing but the auxiliary" (Broca, 1865 translated in Berker et al., 1986, p. 1067). Although Broca thought that education and imitation contributed to handedness, he assigned a strong 'organic predisposition' to handedness. In relation to the brain, he thought that "most men are naturally 'left-brained' and that, by exception, some of them, those called left-handed, are, on the contrary, right-brained" (Broca, 1865 translated in Berker et al., 1986, p. 1068). Broca then cited findings which suggested that the convolutions of the left hemisphere developed earlier than those of the right hemisphere in the embryo. It was thus natural for the newborn and young infant to show preference for the extremities with the more perfect innervation and become right-handed. This made humans prone to execute complex manual and intellectual actions such as articulate speech with the left hemisphere.

Broca's belief that left-handers are right-brained for articulate speech dominated research for years and is still more often than not a common belief of the layperson. Today we know that this is, in fact, a relatively rare incidence. Of the 10% of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54040-9

people who are left-handed, about 30–35% show right hemisphere language dominance. There are also right-handers who show right hemisphere dominance although this number is much smaller with about 5–6% of the 90% right-handers (McManus, 2009). In other words, one in three left-handers and one in 20 right-handers have language functions that involve processing in the right hemisphere (McManus et al., 2013; Knecht et al., 2000a,b).

The high incidence of left hemisphere language representation (at least at the word level) in right- and left-handers has meanwhile been shown in many studies. The stronger right-handedness is expressed in a person, the more likely it is that language is represented in the left hemisphere. Although being left-handed raises the chances that language is represented in the right hemisphere, this does not mean that left-handedness is a precondition or a necessary consequence of language lateralization to the right hemisphere.

Measuring Handedness and Language Dominance

Although people often determine handedness by the hand with which a person writes, this is – as also noted by Broca – a rather imprecise measure as societal pressure may have influenced hand writing preferences. In research, there is no consensus how handedness should be established. Frequently questionnaires are used such as the Edinburgh Handedness Inventory (Oldfield, 1971) which identifies the hand used for manipulating tools. These measures are, however, poorly related to manual proficiency (for a summary see Michel et al., 2013). Other measures used are quantitatively oriented such as the peg-board task, or, more rarely, observation of hand use. Another possibility is to use functional neuroimaging (such functional magnetic resonance imaging, fMRI) to measure the neural correlates of handedness during unilateral and bilateral finger movements. Compared to the previously mentioned measures, this is, however, a rather time-consuming and expensive enterprise.

Another issue to consider is what right- or left-handedness means. While classically handedness is referred to in directional terms such as left-handed, right-handed, or mixed, it is now acknowledged that the degree of handedness may be a better indicator of cerebral organization. Degree of handedness refers to how consistently a person prefers to use one hand over the other over a wide variety of tasks and magnitude. (Arning et al., 2013; Tabiowo and Forrester, 2013). Measuring and establishing handedness is thus less clear-cut as one might think.

Compared to handedness, language dominance is even more difficult to establish. Many studies rely on handedness measures and simply assume that there is a relationship between handedness and language dominance. At a group level, a correlation between handedness and language dominance has indeed been shown. At the individual level, however, handedness may not be a very good predictor. More reliable measures of language lateralization are obtained with functional brain imaging methods (such as fMRI, functional transcranial Doppler ultrasound (fTCD)), the Wada test, or the dichotic listening technique. Among these techniques, the fTCD technique is relatively low cost and easy to use. It has been validated against fMRI and the Wada test showing good correlations (Knecht et al., 1998; Somers et al., 2011).

Is There a Handedness Gene?

One question that has received considerable attention is whether handedness is heritable. Broca observed that handedness seems to run in families and attributed some heritable trait to handedness. Indeed, research has provided data showing a positive correlation between the handedness of a child and its biological parents while no such correlation is found with adoptive parents. Further, identical twins are more likely to share hand preference than nonidentical twins (for a summary see McManus et al., 2013; Ocklenburg et al., 2013). Although most researchers agree that genetic factors are involved in handedness, it has been extremely difficult to identify those factors. Two genetic models suggested that a single gene is responsible for handedness (Annett, 1998; McManus, 1985). For example, the 'right shift gene' model assumes that there is a gene that predisposes us to right-handedness. Strong right-handedness emerges if there are two copies of the gene and weak to moderate right-handedness if there is one. It may also happen that we do not get any copy of that gene. In this case, handedness will be a random outcome of chance factors and half of us will be left and the other half right-handed. Single-gene models have, however, been challenged.

While initially it was thought that handedness is a monogenic trait, recent views, mostly based on genome-wide association studies (that is statistical searches in the DNA of a large population), favor a polygenetic view. These studies suggest that at least 30–40 or even up to 100 genes may play a role in handedness and most of them are still unidentified (McManus et al., 2013). One line of research has linked handedness with a network of genes involved in establishing left–right asymmetry in the developing embryo such as inner organs and also the brain with the latter possibly influencing the development of handedness (Brandler et al., 2013). The gene variant that is most strongly associated with handedness is gene PCSK6. Disrupting this gene in mice (and other organisms) causes defects of left–right asymmetry such as abnormal positioning of organs in the body. For example, these mice may have the heart on their right and the liver on the left side. It has thus been suggested that the same biological mechanisms associated with the development of left–right asymmetry is involved in determining handedness (Brandler et al., 2013).

Generally, there is still controversy around the question whether or to what extent multiple genetic and environmental factors are involved in the ontogenesis of handedness. Some researchers have reported that more than 74% of the observed variance in handedness for writing is accounted for by environmental factors (Vuoksimaa et al., 2009). It seems that handedness may be derived from a mixture of genes, environmental factors, and cultural pressures.

Although no direct link between specific genes and handedness or forms of functional hemispheric asymmetries such as language has firmly been established, it is nevertheless assumed

that handedness and language dominance are in some way linked. Support comes from functional imaging studies that have reported an association between hemispheric lateralization of language in children and adults and handedness (Tzourio-Mazoyer et al., 2010; Szaflarski et al., 2012). If handedness and language lateralization are linked, then the gene or genes underlying handedness may also influence language dominance – at least this is how the argument goes. Hence, the study of handedness is viewed by some researchers as an indirect way of studying language dominance (McManus, 2009).

Historical and Geographical Variation of Handedness

As discussed previously, there are indications that handedness is under genetic control. Multiple forms of a single gene (genetic polymorphism) have been shown in biological systems, and biological systems vary in history and in different geographical locations (for a summary see McManus, 2009; McManus et al., 2013). Handedness also has polymorph properties, and hence the question has been raised whether handedness also varies in history and in different geographical locations. Further, as there is a correlation between handedness and language dominance, it may be that language dominance varies geographically and historically although this is difficult to prove. For handedness, however, there is indeed some support for historical and geographical variation. It seems that our primitive great ape ancestors were more or less bimanual. At some point in history (1 million years before present), so the story goes, all humans became right-handed (McManus, 2009). Subsequently, at some stage, there was some occurrence of left-handedness, about 9–10%. These numbers remained constant until between 1800 and 1950 when a dip occurred and left-handedness dropped to about 3% before the numbers rose again to 10–11% after 1950.

Concerning geographical variation, while right-handedness is more frequent in all human societies, *rates* of left-handedness vary between different continents, countries, and even within countries. For example, studies using large sample sizes found that the rate of left-handedness was 9.8% in Canada compared to 5.2% in India and 4.7% in Japan (reported in McManus, 2009). These historical and geographical differences have been explained in terms of gene frequency rather than social pressure. Assuming that handedness and cerebral lateralization are related, and using a specific mathematical model, it has been suggested that predictions can be made concerning language lateralization and disease prevalence such as dyslexia, stuttering, autism, or schizophrenia. This, however, is still open to verification.

Evolutionary Views on Handedness and Language

Observations that handedness and language evolved along the human lineage, and that the use of the right hand and language representation are dominated by the left hemisphere have led numerous researchers to postulate an evolutionary link between language and handedness (for a summary see Tabiowo and Forrester, 2013). One evolutionary theory suggests that the two hemispheres shared labor for critical survival processes. In this view, the right hemisphere evolved as dominant for controlling arousal levels (e.g., important for fight or flight actions) and the left hemisphere for processing routine behaviors such as structured sequences of actions (e.g., feeding behavior). Parallels have been seen between such structured motor activity and hierarchical structured language processes such as syntax. Some researchers have tried to establish a link between tool manufacture or tool use and linguistic processes. Right-handed tool use may be the link between left hemisphere dominant action sequences and language processes. Researchers have thus turned to animal models, and specifically great apes, to investigate the evolutionary link. Although great apes do not possess human-like language, they show neuroanatomical brain regions that overlap and show similar cytoarchitecture with Broca's and Wernicke's region in humans (Striedter, 2005, p. 327). Further, studies of *bimanual* object manipulation have reported right-handed population biases in chimpanzees, bonobos, gorillas, and orangutans (Tabiowo and Forrester, 2013). It has thus been speculated that language emerged from regions in the left hemisphere that originally evolved to cope with the hierarchical sequences of actions inherent in tool use (Pulvermüller and Fadiga, 2010). Critics counter the 'tool-use hypothesis' with the argument that tool use is not directly related to hemispheric dominance for language. As mentioned previously, the majority of left-handers do not show right hemisphere dominance for language.

Others have argued that gestural communication has played a major role in the evolution of hand preferences and cerebral asymmetries. One influential theory proposes that the region in charge of manual gestures was modified in humans such that human language emerged from gestural communication and vocalizations were gradually incorporated into the gestural system (Corballis, 2003, 2012) (see **Figure 1**). In this view, language evolved from the mirror system in primates. In humans, the mirror system was adapted for communication as a result of bipedalism. Hands were now freed and could be used to produce elaborate gestures. Over time, possibly due to environmental pressures, forelimbs became increasingly involved in other activities such as tool manufacture, use of weapons, or carrying. This scenario may have prompted a shift from manual to facial gestures in communication. Vocal sound may have been added to distinguish between otherwise identical gestures. Eventually, vocal speech replaced gestures of the face and hands and became the dominant language medium. It is assumed that the appearance of Broca's area (in the left prefrontal cerebral cortex) played an important role in the process of adding sounds to gestures.

Not all researchers agree with these views and challenge the interpretation that vocalization preceded right-handedness, the role assigned to Broca's area or the mirror system hypothesis in lateralization (for example, Annett, 2003; Beaton, 2003; Bradshaw, 2003; Dale et al., 2003; Johnson-Frey, 2003). Regardless of whether the evolutionary precursors of human language involve mainly gestures or a combination of gestures and vocalizations, primate studies as well as developmental and neuroimaging data point to the importance of gestures in social interaction.

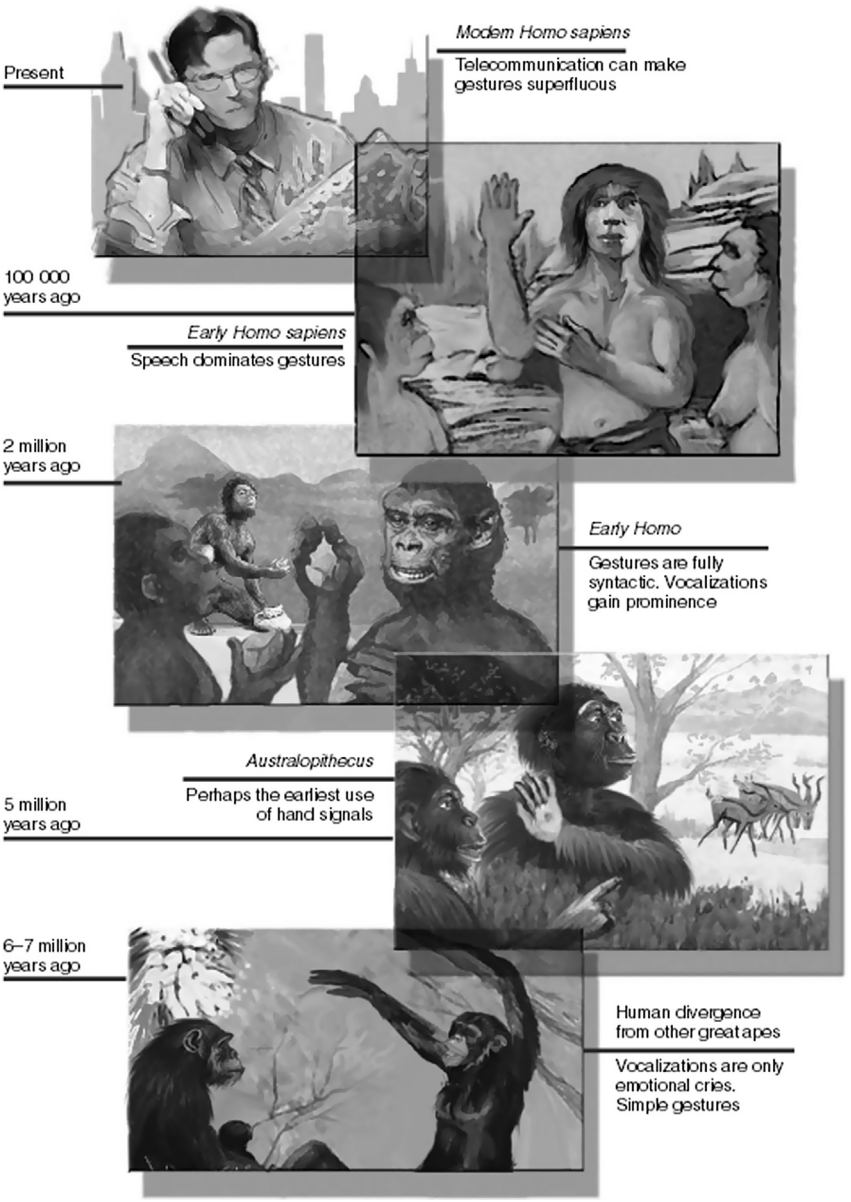

Figure 1 When did autonomous speech emerge? Figure taken from Corballis, 2007. Taking up arms. In: Cohen, M.H., Stemmer, B. (Eds.), Fragments of Mind and Brain Fragments of the Mind and Brain. Consciousness and Cognition, Elsevier, p. 25.

Development of Handedness and Language Lateralization

How and when does handedness develop? There is no definitive answer to this question. There are some suggestions that handedness is already determined prenatally. One suggestion is that limb use *in utero* such as frequency of left- or right-arm movement or right thumb sucking will determine handedness. This, however, was not confirmed by systematic studies looking at the relationship between fetal lateralization and handedness (de Vries et al., 2001). Another suggestion views fetal posture and position influencing asymmetrical biases. For example, *in utero* head orientation preference leads to asymmetrical motor actions including hand and arm movements and there is some indication that this is predictive of later handedness for reaching (summary in Michel et al., 2013). There is more agreement on postnatal handedness development. After birth manual abilities change and handedness in infancy and childhood is considered unreliable and unstable before 6–10 years of age (for a summary see Michel et al., 2013).

A question that emerges is whether the development of handedness and language lateralization are linked. There is some indication that around the time when handedness becomes more stable in childhood, lexical–semantic language processing is increasingly lateralized to the left hemisphere in distinct regions of the brain (Ressel et al., 2008; Szaflarski et al., 2006a,b). Although handedness

correlates with (lexical–semantic) language lateralization for left-handed and right-handed children and adults, handedness has not been found to be a dependable predictor of hemispheric language dominance, especially in left-handers (Szaflarski et al., 2012) (see **Figure 2**). The increasing left lateralization of language with age in left- and right-handed children in specific brain regions may indicate similar developmental processes that are, so it seems, independent of handedness. However, the last words on these relationships may not have been spoken as the interpretation of some of the studies is hampered by methodological limits (such as the way handedness was established, the way performance of the task was measured in the scanner, the cultural diversity of the subjects investigated, or the fact that most data are based on cross-sectional instead of longitudinal studies).

Handedness, Language Lateralization, and Sex

Most studies report that left-handedness and ambiguous handedness is more common in males than in females, and this has been shown for children, adults, twins, and triplets (for a summary see meta-analysis by Papadatou-Pastou et al., 2008; Vuoksimaa et al., 2009; Vuoksimaa and Kaprio, 2010). Meta-analyses suggest that there is a 25% higher prevalence of non-right-handedness in males (Sommer et al., 2008). Researchers have thus pursued the question of whether gender differences observed with handedness has an effect on language lateralization.

Trying to establish a link between gender distribution of handedness and language lateralization is complicated by several issues. One is that cultural differences seem to offset the balance of the distribution of handedness of males and

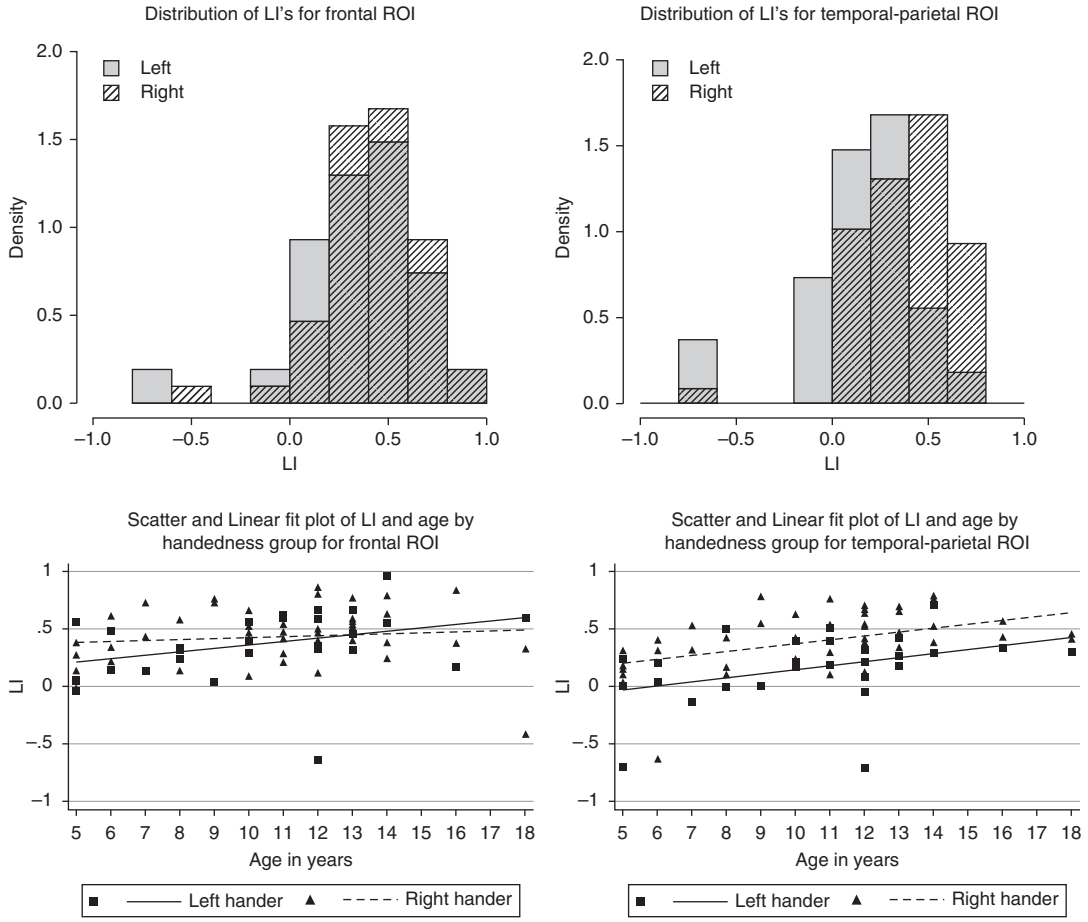

Figure 2 Szaflarski et al. (2012) investigated the development of language lateralization in left- and right-handed children. Brain scans (using the fMRI technique) were obtained from 54 right-handed and 27 left-handed children while they performed a verb generation task. The task reflects lexico-semantic processing. The authors did not find an interaction between handedness and age. The top panel shows the distribution of the lateralization index (LI) (a measure of functional hemispheric lateralization) in frontal brain regions and temporal–parietal brain regions for right- and left-handed children. As can be seen in response to the word generation task, the frontal as well as the temporal–parietal regions in the left hemisphere ($LI > 0.01$) were predominantly activated in children. The bottom panel shows the lateralization growth curves for the frontal and temporal–parietal regions (frontal and temporal–parietal) for the left-handed and the right-handed group at different ages. While age did not have a clear effect on frontal lateralization (left bottom panel), temporal–parietal lateralization increased with age in both left-handers and right-handers. Figure taken from Szaflarski et al. (2012), p. 89. Legend for figure by B. Stemmer.

females (Sommer et al., 2008) (Figure 3). Another is that it is currently not clear whether language lateralization (regardless of handedness) is influenced by gender in the first place. A large meta-analysis investigated studies according to the method with which sex differences were measured. It was concluded that there was no effect of sex on language lateralization, regardless of whether dichotic listening techniques or neuroimaging techniques were used (for a review see Sommer et al., 2008) (Figures 4 and 5). Numerous individual studies and literature reviews, however, report other findings. While the relationship between sex and language lateralization remains controversial, current data suggest that handedness and sex are in some way linked.

Other studies have tried to establish a link between brain regions that have been associated with language such as the

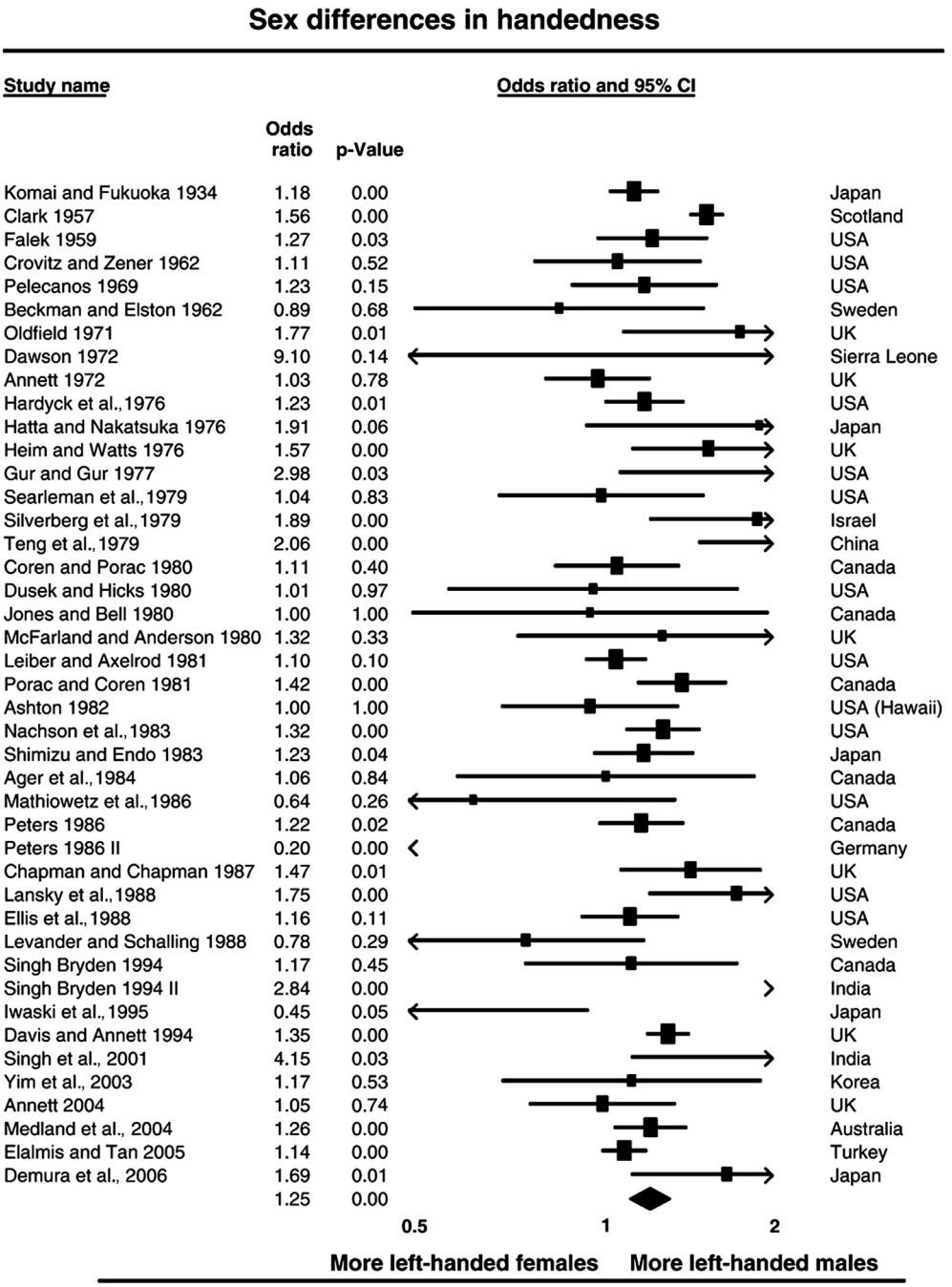

Figure 3 Findings from studies in different countries investigating sex differences in handedness. Taken from Sommer, I.E., Aleman, A., Somers, M., Boks, M.P., Kahn, R.S., 2008. Sex differences in handedness, asymmetry of the planum temporale and functional language lateralization. Brain Research 1206, p. 77.

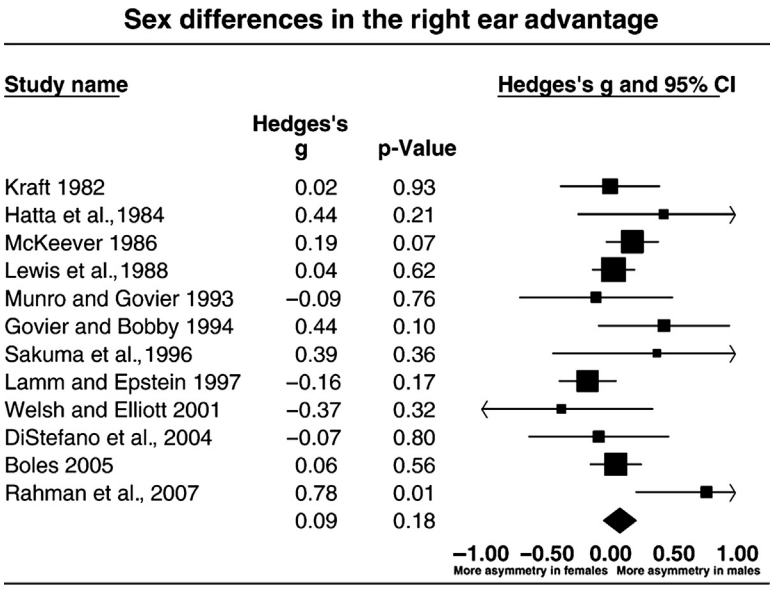

Figure 4 Studies investigating sex differences in language lateralization using the dichotic listening technique. Negative values indicate more asymmetry in females and positive values indicate more asymmetry in males. Taken from Sommer, I.E., Aleman, A., Somers, M., Boks, M.P., Kahn, R.S., 2008. Sex differences in handedness, asymmetry of the planum temporale and functional language lateralization. Brain Research 1206, p. 79.

planum temporale or Broca's area and sex and handedness. The planum temporale is a region adjacent to the primary auditory area in the temporal lobe. Decreased planum temporale asymmetry has been associated with non-right-handedness (Steinmetz, 1996; Hervé et al., 2006). Some studies have reported more asymmetry of the planum temporale in men, while others reported more asymmetry in women. A study with a large number of subjects and a meta-analysis could not find any relation between handedness and the planum temporale, or an interaction with sex (Good et al., 2001; Sommer et al., 2008) (**Figure 6**).

Classically, Broca's area has been subdivided into two subregions, the pars triangularis and the pars opercularis (areas 44 and 45 in **Figure 7**). Studies have related the volume of the pars opercularis to handedness, while no such asymmetry has been described for the pars triangularis (Foundas et al., 1995; Foundas et al., 1998; Powell et al., 2012). When trying to establish a direct link between Broca's region, sex, and handedness, no interaction was, however, found.

Why do we find such discrepant results and seem unable to come up with a clear statement? Much has to do with differences in methodological approaches and statistical power. For example, measurement techniques of the planum temporale in individual studies differ greatly. With regard to Broca's region, research has shown that the anatomical correlates of language processing go well beyond Broca's classical region (see **Figure 7**). Further, there is also high anatomical and functional variability in individual brains. In addition, considering that the number of left-handers is relatively small, a large number of subjects is necessary to find any effects, and this number is usually too high for costly neuroimaging studies.

Embodied Cognition, Handedness, and Language

Some researchers consider asymmetries of communicative gestures, manual skills, as well as concepts expressed through language as activities involving bodily actions. For example, to comprehend a concept like *jump* it must be represented in some way in our brain. Traditionally concepts like *jump* have been viewed as symbolic representations independent of modality. In theories of embodied cognition, the representation of concepts is viewed as dependent on related bodily actions such as perceptual or motor processes. For example, the representation of the concept *jump* depends on motor processes related to the action of motion. Cognition – and language as part of cognition – is considered 'grounded' in – and thus inseparable from – bodily action (for a summary see Willems and Hagoort, 2007). This implies, for example, that an action verb such as *jump* should, at least partially, be represented in the cortical motor system. Ultimately, there should be an overlap of brain processes supporting language and action, and, in order to comprehend the concept *jump*, the motor system should be activated. Indeed, neuroimaging and brain stimulation studies have shown activation of brain regions that support perception and action in response to action-related language including gestures that accompany speech (for reviews see Michel et al., 2013; Willems and Hagoort, 2007). It is assumed that an implicit mental simulation of the referent (in our example, the word *jump*) is created in these brain regions. Several neuroimaging studies have reported left-lateralized activations of cortical motor areas in response to manual-action verbs. This is not surprising considering that the left hemisphere is dominant for certain aspects of

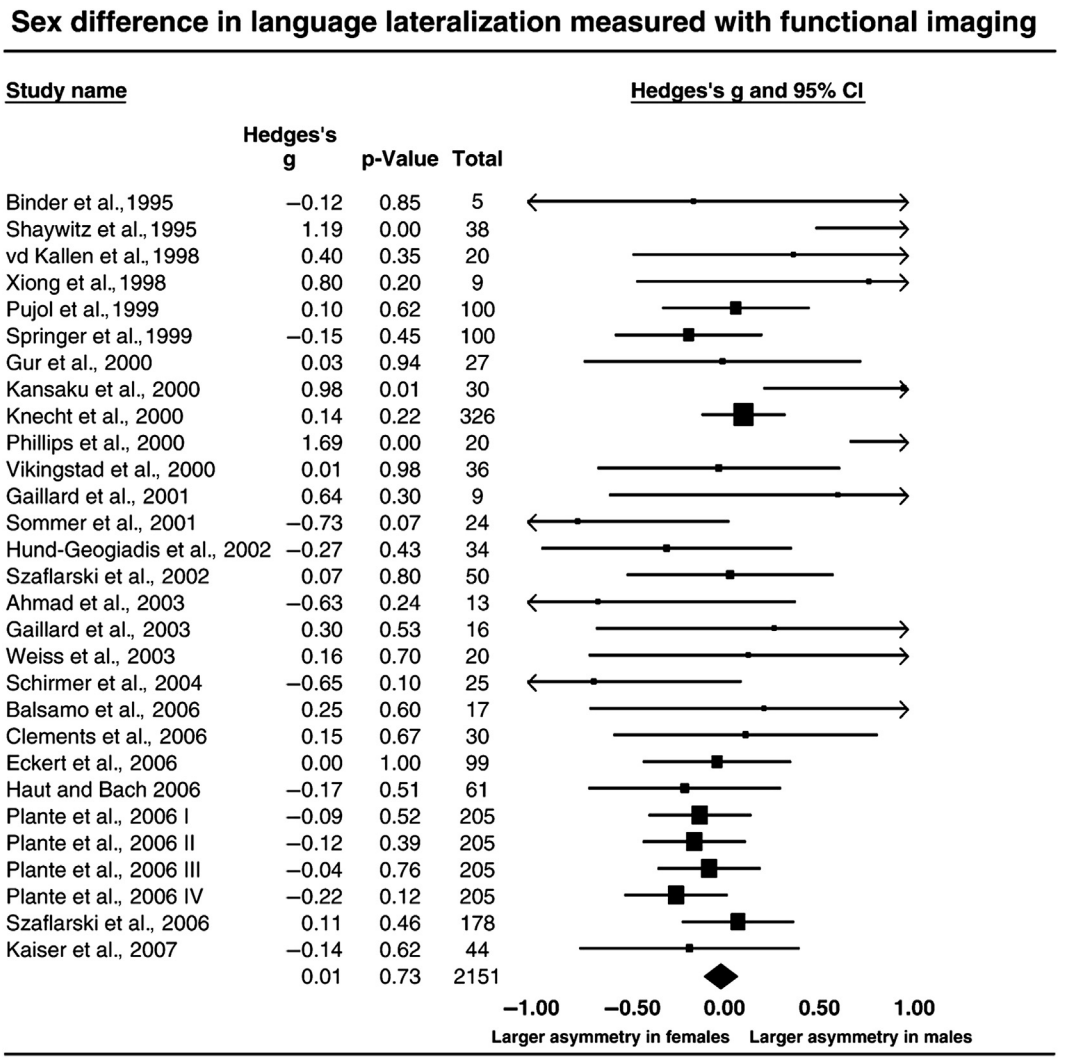

Figure 5 Studies investigating sex differences in asymmetry of language activation. Negative values indicate larger asymmetry in females and positive values indicate larger asymmetry in males. Taken from Sommer, I.E., Aleman, A., Somers, M., Boks, M.P., Kahn, R.S., 2008. Sex differences in handedness, asymmetry of the planum temporale and functional language lateralization. Brain Research 1206, p. 80.

language. Some researchers wondered whether these results were influenced by research that only used right-handed subjects and thus investigated the influence of handedness on the reading of manual-action verbs such as *grasp* and *throw* (Willems et al., 2010). Right-handers showed stronger activations in premotor areas in the left hemisphere and left-handers in the right hemisphere in response to reading manual-action verbs compared to reading non-manual-action verbs (such as *kneel* or *giggle*). It thus seems that actions we perform with our preferred hand have different hemispheric representations in right- and left-handed individuals.

Concern has been raised on the interpretation of some of these studies as it cannot be excluded that imagery of the action word played a role in activating the motor region. Caution has also been raised to not interpret the link between action and language as meaning that activation of the motor cortex is *necessary* for word comprehension.

Handedness, Language Lateralization, and Language Pathology

Observations that left-handedness is over-represented in some psychiatric and developmental language disorders and an observed tendency toward decreased brain asymmetries in language-related regions has raised the question of whether there is a relationship between pathological language disorders, language lateralization, and handedness.

Support for an over-representation of left-handedness in psychiatric disorders comes from a meta-analysis of over 40 studies. Left-handedness was more prevalent in psychiatric illness (Dragovic and Hammond, 2005). Abnormal lateralization of regions associated with language function have been described in the schizophrenia literature such as decreased density of gray and white matter in left temporal regions (including Wernicke's area), reduced leftward asymmetry in the planum temporale, or rightward

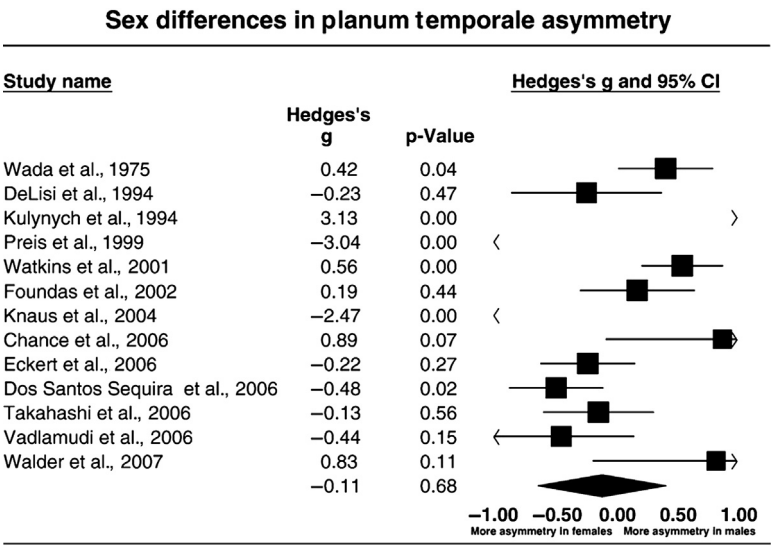

Figure 6 Findings from a meta-analysis of sex differences in asymmetry of the planum temporale. Figure taken from Sommer, I.E., Aleman, A., Somers, M., Boks, M.P., Kahn, R.S., 2008. Sex differences in handedness, asymmetry of the planum temporale and functional language lateralization. Brain Research 1206, p. 77.

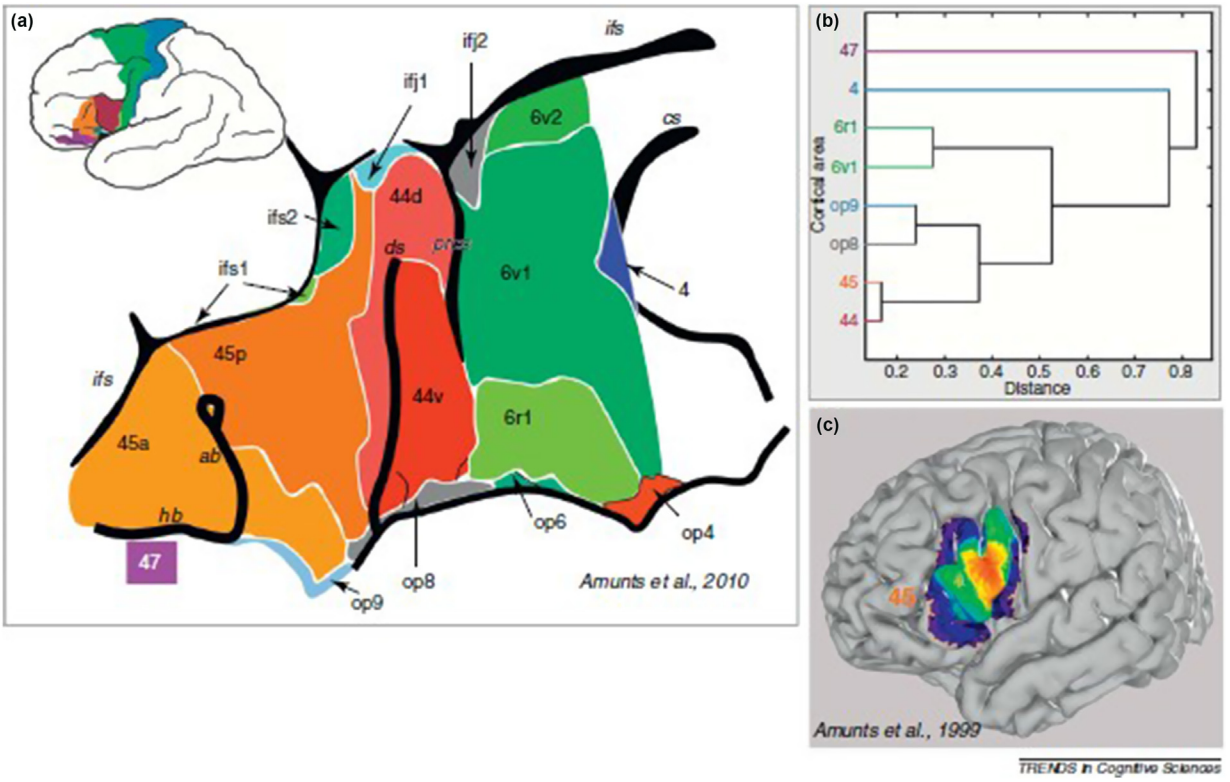

Figure 7 New topography of Broca's region. (a) The map of Broca's region based on the distribution of receptors of neurotransmitters and modulators shows a complex segregation of the inferior frontal gyrus and adjacent cortices (Amunts et al., 2010). In particular, it shows the presence of a series of new areas in the inferior frontal sulcus (ifs) labeled as areas ifs1, ifs2, ifj1, and ifj2, as well in the frontal operculum ('op'-areas). The areas of this region can be clustered based on their receptor architectonic pattern as shown in (b). Transmitter receptors are key molecules for neurotransmission, and therefore, indicative of function. Areas 44 and 45 of the classical Broca region are most similar to each other and they cluster with areas op8 and op9 of the frontal operculum. Interestingly, area 47 is the most different of all analyzed areas, suggesting distinct function (Amunts et al., 2010). (c) Cytoarchitectonic probabilistic map of right area 45 (Amunts et al., 1999) in the standard reference space of the single subject template from the MNI (Collins et al., 1994). The color indicates the overlap of individually mapped area 45 in common reference space. Note the high intersubject variability (labeled in blue and green) of area 45 in regions outside the free surface of the triangular part (orange and red). Figure taken from Amunts Zilles (2012): Figure 4 on page 424.

asymmetry in the pars triangularis (a subregion of Broca's area) (for a review see Li et al., 2009). In a sample of more than 100 patients with depression or a psychotic disorder (schizoaffective disorder), the prevalence of left-handedness was 40% for those with psychotic disorders compared to 11% for mood disorder (Webb et al., 2013). Although these observations are suggestive of a relation between the structural asymmetries and left-handedness, caution is necessary before drawing firm conclusions as (1) a *direct* link between these asymmetries and handedness in schizophrenia has not been established, and (2) there are also schizophrenia patients who are right-handed.

Psychosis (such as schizophrenia) is frequently accompanied by language-associated phenomena such as auditory verbal hallucinations. These hallucinations, however, are not specific to schizophrenia and also occur in about 10–15% of the general nonpsychotic population. Although the majority of these individuals does not have a clinically relevant psychotic disorder, there are indicators that suggest that they may be vulnerable for schizophrenia (Sommer et al., 2010). One difference between these two populations with auditory verbal hallucinations is that in nonpsychotic individuals language is normally lateralized while this does not seem to be the case in psychotic individuals who show decreased language lateralization (Diederen et al., 2010). It is currently not known whether auditory hallucinations are especially prevalent in psychotic patients who are left-handed and whether there is an interaction between functional language asymmetry, psychosis, and handedness.

In addition to psychiatric disorders, developmental disorders have also been associated with abnormal language lateralization and some studies report that left-handedness is more frequent in this population. Specific language impairment (SLI) is a disorder in children that is characterized by problems learning the native language without any apparent reason. SLI is frequently accompanied by problems learning to read (developmental dyslexia). Behavioral observations have suggested that the typical pattern of left language lateralization is disrupted in children with SLI and developmental dyslexia and neuroimaging studies have supported these observations (for a summary see Bishop, 2013). Despite the common belief, however, that left-handedness is associated with dyslexia, no association between handedness and language and literacy impairment was found in a review of the literature (Bishop, 2013). This has been supported by genome-wide association studies. A genetic marker located on gene PCSK6 has linked cerebral asymmetry and dyslexia. In addition, a *trend* toward greater relative right-hand skill compared to the nondyslexic population was observed (Scerri et al., 2011). However, further investigations could not establish a consistent association between hand preference or hand skill in individuals with developmental dyslexia (Brandler et al., 2013).

Autistic spectrum disorders also belong to the group of developmental disorders. Most but not all children with autistic spectrum disorders have receptive and expressive language impairments. Although the data are inconsistent, there is growing evidence that the left hemisphere structural asymmetry is atypically reduced or organized (Lindell and Hudry, 2013). Anatomical regions that have been described as asymmetrically organized include left inferior frontal brain regions (including Broca's area), posterior superior temporal regions (including Wernicke's area), the left planum temporale, and white matter tracts connecting language-relevant regions. Although heterogeneous findings with regard to decreased functional lateralization are reported by neuroimaging studies, most of the studies suggest a pattern of atypical lateralization (Lindell and Hudry, 2013). Most studies also report a higher incidence of left and mixed handedness in individuals with autistic spectrum disorder compared to the general population. The data thus appear to suggest an association between atypical handedness, atypical lateralization, and autistic spectrum disorder. As already pointed out previously, however, caution is required to draw premature conclusions as long as no direct relationship has been established between handedness, lateralization, and disease.

See also: Asymmetry of Body and Brain; Lateralization of Language as Demonstrated by Brain Imaging Procedures.

Bibliography

Amunts, K., Zilles, K., 2012. Architecture and organizational principles of Broca's region. Trends in Cognitive Sciences 16 (8), 418–426.

Annett, M., 1998. Handedness and cerebral dominance: the right shift theory. The Journal of Neuropsychiatry and Clinical Neurosciences 10, 459–469.

Annett, M., 2003. Myths of first cause and asymmetries in human evolution. Behavioral and Brain Sciences 26 (2), 208–209.

Arning, L., Ocklenburg, S., Schulz, S., Ness, V., Gerding, W.M., Hengstler, J.G., et al., 2013. VNTR polymorphism is associated with degree of handedness but not direction of handedness. PLoS One 8 (6), e67251.

Amunts, K., et al., 1999. Broca's region revisited: cytoarchitecture and intersubject variability. Journal of Comparative Neurology 412, 319–341.

Amunts, K., et al., 2010. Broca's region: novel organizational principles and multiple receptor mapping. PLoS Biology 8, e1000489.

Beaton, A.A., 2003. Going for Broca? I wouldn't bet on it! Behavioral and Brain Sciences 26 (2), 212–213.

Berker, E., Berker, A., Smith, A., 1986. Translation of Broca's 1865 report: localization of speech in the third left frontal convolution. Archives of Neurology 43 (10), 1065–1072.

Bishop, D.V., 2013. Cerebral asymmetry and language development: cause, correlate, or consequence? Science 340 (6138), 1230531.

Bradshaw, J.L., 2003. Gesture in language evolution: could I but raise my hand to it! Behavioral and Brain Sciences 26 (2), 213–214.

Brandler, W.M., Morris, A.P., Evans, D.M., Scerri, T.S., Kemp, J.P., Timpson, N.J., et al., 2013. Common variants in left/right asymmetry genes and pathways are associated with relative hand skill. PLoS Genetics 9 (9), e1003751.

Broca, P., 1865. Sur le siège de la faculté du langage articulé. Bulletin de la Société d'Anthropologie 6, 337–393.

Corballis, M.C., 2003. From mouth to hand: gesture, speech, and the evolution of right-handedness. Behavioral and Brain Sciences 26 (2), 199–208.

Corballis, M.C., 2012. Lateralization of the human brain. Progress in Brain Research 195, 103–121.

Collins, D.L., et al., 1994. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. Journal of Computer Assisted Tomography 18, 192–205.

Dale, R., Richardson, D.C., Owren, M.J., 2003. Pumping for gestural origins: the well may be rather dry. Behavioral and Brain Sciences 26 (2), 218–219.

Diederen, K.M., De Weijer, A.D., Daalman, K., Blom, J.D., Neggers, S.F., Kahn, R.S., et al., 2010. Decreased language lateralization is characteristic of psychosis, not auditory hallucinations. Brain 133 (Pt 12), 3734–3744.

Dragovic, M., Hammond, G., 2005. Handedness in schizophrenia: a quantitative review of evidence. Acta Psychiatrica Scandinavica 111 (6), 410–419.

Foundas, A.L., Eure, K.F., Luevano, L.F., Weinberger, D.R., 1998. MRI asymmetries of Broca's area: the pars triangularis and pars opercularis. Brain and Language 64 (3), 282–296.

Foundas, A.L., Leonard, C.M., Heilman, K.M., 1995. Morphologic cerebral asymmetries and handedness. The pars triangularis and planum temporale. Archives of Neurology 52 (5), 501–508.

Good, C.D., Johnsrude, I., Ashburner, J., Henson, R.N., Friston, K.J., Frackowiak, R.S., 2001. Cerebral asymmetry and the effects of sex and handedness on brain structure: a voxel-based morphometric analysis of 465 normal adult human brains. Neuroimage 14 (3), 685–700.

Hervé, P.-Y., Crivello, F., Perchey, G., Mazoyer, B., Tzourio-Mazoyer, N., 2006. Handedness and cerebral anatomical asymmetries in young adult males. Neuroimage 29 (4), 1066–1079.

Johnson-Frey, S.H., 2003. Mirror neurons, Broca's area and language: reflecting on the evidence. Behavioral and Brain Sciences 26 (2), 226–227.

Knecht, S., Deppe, M., Dräger, B., Bobe, L., Lohmann, H., Ringelstein, E.-B., et al., 2000a. Language lateralization in healthy right-handers. Brain 123 (1), 74–81.

Knecht, S., Deppe, M., Ebner, A., Henningsen, H., Huber, T., Jokeit, H., et al., 1998. Noninvasive determination of language lateralization by functional transcranial Doppler sonography: a comparison with the Wada test. Stroke 29 (1), 82–86.

Knecht, S., Dräger, B., Deppe, M., Bobe, L., Lohmann, H., Flöel, A., et al., 2000b. Handedness and hemispheric language dominance in healthy humans. Brain 123 (12), 2512–2518.

Li, X., Branch, C.A., DeLisi, L.E., 2009. Language pathway abnormalities in schizophrenia: a review of fMRI and other imaging studies. Current Opinion in Psychiatry 22 (2), 131–139.

Lindell, A.K., Hudry, K., 2013. Atypicalities in cortical structure, handedness, and functional lateralization for language in autism spectrum disorders. Neuropsychology Review 23 (3), 257–270.

McManus, I.C., 1985. Handedness, language dominance and aphasia: a genetic model. Psychological Medicine, Monograph Supplement No.8. Cambridge University Press.

McManus, I.C., 2009. The history and geography of human handedness. In: Sommer, I.E.C., Kahn, R.S. (Eds.), Language Lateralization and Psychosis. Cambridge University Press, Cambridge, MA, pp. 37–57.

McManus, I.C., Davison, A., Armour, J.A.L., 2013. Multilocus genetic models of handedness closely resemble single-locus models in explaining family data and are compatible with genome-wide association studies. Annals of the New York Academy of Sciences 1288 (1), 48–58.

Michel, G.F., Nelson, E.L., Babik, I., Campbell, J.M., Marcinowski, E.C., 2013. Chapter nine – multiple trajectories in the developmental psychobiology of human handedness. In: Richard, M.L., Janette, B.B. (Eds.), Advances in Child Development and Behavior. JAI, pp. 227–260.

Ocklenburg, S., Beste, C., Güntürkün, O., 2013. Handedness: a neurogenetic shift of perspective. Neuroscience and Biobehavioral Reviews 37, 2788–2793. http://dx.doi.org/10.1016/j.neubiorev.2013.09.014.

Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9 (1), 97–113.

Papadatou-Pastou, M., Martin, M., Munafo, M.R., Jones, G.V., 2008. Sex differences in left-handedness: a meta-analysis of 144 studies. Psychological Bulletin 134 (5), 677–699.

Powell, J.L., Kemp, G.J., Roberts, N., García-Finaña, M., 2012. Sulcal morphology and volume of Broca's area linked to handedness and sex. Brain and Language 121 (3), 206–218.

Pulvermüller, F., Fadiga, L., 2010. Active perception: sensorimotor circuits as a cortical basis for language. Nature Reviews Neuroscience 11 (5), 351–360.

Ressel, V., Wilke, M., Lidzba, K., Lutzenberger, W., Krägeloh-Mann, I., 2008. Increases in language lateralization in normal children as observed using magnetoencephalography. Brain and Language 106 (3), 167–176.

Scerri, T.S., Brandler, W.M., Paracchini, S., Morris, A.P., Ring, S.M., Richardson, A.J., et al., 2011. PCSK6 is associated with handedness in individuals with dyslexia. Human Molecular Genetics 20 (3), 608–614.

Somers, M., Neggers, S.F.W., Diederen, K.M., Boks, M.P., Kahn, R.S., Sommer, I.E., 2011. The measurement of language lateralization with functional transcranial doppler and functional MRI. A critical evaluation. Frontiers in Human Neuroscience 5. http://dx.doi.org/10.3389/fnhum.2011.00031.

Sommer, I.E., Aleman, A., Somers, M., Boks, M.P., Kahn, R.S., 2008. Sex differences in handedness, asymmetry of the planum temporale and functional language lateralization. Brain Research 1206, 76–88.

Sommer, I.E., Daalman, K., Rietkerk, T., Diederen, K.M., Bakker, S., Wijkstra, J., et al., 2010. Healthy individuals with auditory verbal hallucinations; who are they? Psychiatric assessments of a selected sample of 103 subjects. Schizophrenia Bulletin 36 (3), 633–641.

Steinmetz, H., 1996. Structure, function and cerebral asymmetry: in vivo morphometry of the planum temporale. Neuroscience and Biobehavioral Reviews 20 (4), 587–591.

Striedter, G.F., 2005. Principles of Brain Evolution. Sinaur Associates, Sunderland, MA.

Szaflarski, J.P., Holland, S.K., Schmithorst, V.J., Byars, A.W., 2006a. fMRI study of language lateralization in children and adults. Human Brain Mapping 27 (3), 202–212.

Szaflarski, J.P., Rajagopal, A., Altaye, M., Byars, A.W., Jacola, L., Schmithorst, V.J., et al., 2012. Left-handedness and language lateralization in children. Brain Research 1433, 85–97.

Szaflarski, J.P., Schmithorst, V.J., Altaye, M., Byars, A.W., Ret, J., Plante, E., et al., 2006b. A longitudinal functional magnetic resonance imaging study of language development in children 5 to 11 years old. Annals of Neurology 59 (5), 796–807.

Tabiowo, E., Forrester, G.S., 2013. Structured bimanual actions and hand transfers reveal population-level right-handedness in captive gorillas. Animal Behaviour 86 (5), 1049–1057.

Tzourio-Mazoyer, N., Petit, L., Razafimandimby, A., Crivello, F., Zago, L., Jobard, G., et al., 2010. Left hemisphere lateralization for language in right-handers is controlled in part by familial sinistrality, manual preference strength, and head size. Journal of Neuroscience 30 (40), 13314–13318.

de Vries, J.I., Wimmers, R.H., Ververs, I.A., Hopkins, B., Savelsbergh, G.J., van Geijn, H.P., 2001. Fetal handedness and head position preference: a developmental study. Developmental Psychobiology 39 (3), 171–178.

Vuoksimaa, E., Kaprio, J., 2010. Sex differences in left-handedness are also evident in Scandinavia and in twins: comment on Papadatou-Pastou, Martin, Munafo, and Jones (2008). Psychological Bulletin 136 (3), 344–347.

Vuoksimaa, E., Koskenvuo, M., Rose, R.J., Kaprio, J., 2009. Origins of handedness: a nationwide study of 30,161 adults. Neuropsychologia 47 (5), 1294–1301.

Webb, J.R., Schroeder, M.I., Chee, C., Dial, D., Hana, R., Jefee, H., et al., 2013. Left-handedness among a community sample of psychiatric outpatients suffering from mood and psychotic disorders. SAGE Open 3 (4).

Willems, R.M., Hagoort, P., 2007. Neural evidence for the interplay between language, gesture, and action: a review. Brain and Language 101 (3), 278–289.

Willems, R.M., Hagoort, P., Casasanto, D., 2010. Body-specific representations of action verbs: neural evidence from right- and left-handers. Psychological Science 21 (1), 67–74.

Happiness and Work

Annabelle Krause, Institute for the Study of Labor (IZA), Bonn, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The relationship between happiness and work is subject to an evergrowing empirical literature in economics. The analyses are mostly based on large-scale survey data to measure subjective well-being. Whereas one large strand of research investigates the effect of job loss and becoming unemployed, another field of study focuses on the determinants of job satisfaction evolving around employment conditions, self-employment, and potential public sector satisfaction premiums. A smaller part of the literature investigates potential driving effects of happiness on labor market outcomes. This article will give an overview about the most significant subareas of research and the empirical literature in economics to date.

Introduction

The economic research community has experienced a steep rise in the interest of analyzing individual happiness over the past two decades. Also frequently termed (life) satisfaction or subjective well-being, this field of research aims to detect information about subjective utility. Thereby, it constitutes a complementary approach to the common way in economics of measuring utility through revealed preferences or in other words, observable choices made by the individual (Frey and Stutzer, 2002). (I use the terms happiness, subjective well-being and life satisfaction interchangeably in this article, as with most economists, see, e.g., Frey and Stutzer (2002). However, as they are separable constructs, I will use the exact term when citing empirical research.) The pioneering work of Easterlin (1974) challenged the standard focus on income as the sole measure for human utility by detecting the fact that while people in the same country with higher incomes are happier than those with less income, increasing national income per person over time is not accompanied with a growing average level of happiness, at least in countries where basic needs are met. Next to this growing attention of academic researchers, policy makers are also eager to find out whether gross domestic product (GDP) alone still represents a reasonable measure of societal progress. The Stiglitz–Sen–Fitoussi Commission on the Measurement of Economic Performance and Social Progress initiated by the former French president Nicolas Sarkozy in 2008 is a prominent example of this recent development (Stiglitz et al., 2009). The Commission explicitly recommended to take into account both objective and subjective dimensions of well-being given that there appears to be an increasing gap between what people actually find important for their well-being and the information contained in the GDP level.

Parallel to this relationship between happiness and income, the question of whether and how work is related to individual happiness takes up another large fraction of the literature on happiness economics. This may not be surprising as work constitutes a significant part of people's (everyday) life. However, new classical economic theory constitutes work as a necessary evil and unemployment is seen as voluntary. According to this theory, individuals choose to be unemployed because they prefer receiving unemployment benefits and enjoy leisure time as compared to a dissatisfying wage and the time spent working. They choose to work if the offered wage exceeds a certain threshold, which enables an increase in consumption with the higher income received. In contrast, social psychologists claim employment possesses non-pecuniary values as well. As early as the 1930s, the classic study 'Marienthal,' which observed an unemployed community in Austria, revealed unemployment to have detrimental effects for individuals leading to passive resignation (Jahoda et al., 2009) by cutting their social relationships, removing their structured daily routine and lacking the feeling of contributing to society. Likewise, Layard (2011) identifies work to be the third most important factor (out of seven) affecting happiness and claims that while work generates utility gains through increased income and the related consumption possibilities, most people also attach nonpecuniary meanings to work.

When trying to understand the relationship between happiness and work, two key questions evolve: first, is having a job important at all and second, what factors contribute to job satisfaction in the case that individuals do work? While the first question may have been already touched upon in the former paragraph, there are several subquestions related to it, which refer to what actually happens when people lose their job and how this may differ according to certain conditions. While the subsequent section will address these exact questions, Section With Work: What's Important? will cover the literature related to job conditions and job satisfaction. Section Satisfaction as a Driver of Labor Market Outcomes turns to the small literature on how happiness is affecting certain labor market outcomes, thereby acknowledging the fact that happiness not only constitutes an outcome of behavior. This article concludes with Section Conclusions. (I acknowledge that there are far more studies related to the topics covered in this article, and probably also other more detailed questions are investigated, but space is limited and I intend on giving an overview of the most prominent notions and literature strands in economics.)

Without Work: Is that Bad and Why?

To detect whether being without work is an issue for individuals, a large empirical literature in economics has investigated

the relationship between unemployment and subjective well-being. (Darity and Goldsmith (1996) provide an overview on studies reporting harmful effects of unemployment from the (social) psychological literature.) These studies are usually based on large-scale survey data from different countries that measure how satisfied the individual is with his or her life, how happy the individual is currently or provides an index of mental distress by using the General Health Questionnaire (GHQ) which consists of 12 questions related to mental health. These sorts of measures pose some challenges for an accurate analysis. Unobserved individual specific factors may have an influence on answering behavior, thus omitted variable bias may arise. If these unobserved factors are time invariant however, the use of longitudinal or panel data may help to overcome these problems. The focus would then lie on *intra*-person and not *inter*person comparisons. Moreover, there is evidence of measures such as interview ratings, peer reports, and the average daily ratio of pleasant to unpleasant moods that show a strong convergence to self-reports (e.g., Diener and Lucas, 2000). Other objective validity has been shown through, e.g., brain-science data (Urry et al., 2004).

Individual and General Unemployment

The wide consensus that emerged from the literature on the effect of unemployment on subjective well-being is the detection of a detrimental effect across several data sets and countries. One of the early economic studies includes for example Clark and Oswald (1994). They use one cross-section of the British Household Panel Study (BHPS) to estimate ordered probit regressions using the GHQ mental health scale. Their findings indicate a strong negative, statistically significant correlation between unemployment and mental health, which appears larger than between any other characteristic and mental health. Whereas these results are prone to be biased with respect to the true direction of causality, they set the stage for a number of studies dedicated to a similar question. Numerous studies have used the German Socio-Economic Panel (GSOEP) for the analysis of the question of the effect of unemployment on happiness. Two early ones include Gerlach and Stephan (1996) and Winkelmann and Winkelmann (1998) that both use several waves to detect a causal effect of unemployment on life satisfaction. By using fixed effects regressions, it is possible to look at the *change* in life satisfaction scores if an individual enters unemployment as opposed to simply comparing employed and unemployed individuals given that there could also be selection of rather unhappy individuals into unemployment. Both studies confirm the negative impact of an unemployment incidence on life satisfaction, even after controlling for income. The non-pecuniary costs of unemployment appear to be large (Winkelmann and Winkelmann, 1998). These results provide some evidence against the hypothesis that it is a selection of the unhappy individuals into unemployment. Kassenboehmer and Haisken-DeNew (2009) provide more recent evidence based on GSOEP data and focus on exogenous entries into unemployment to detect a clearer causal effect of involuntary unemployment. They approximate the latter by using company closings as an exogenous unemployment shock and investigate whether these company closings, voluntary unemployment such as quitting and wanting to look for another job, as well as being fired exhibit differential effects on life satisfaction. The findings indicate strong negative effects, specifically for women, of company closures, well above the overall effect of unemployment in general, in the year of entry into unemployment. The authors interpret these results as prima facie evidence of reduced outside work options, large investments in firm-specific human capital or a family constraint. For men, the reason for unemployment does not translate into a differential impact on life satisfaction.

A study that extends the literature to more detailed evidence on the effect of unemployment on happiness includes Knabe et al. (2010). They use the Day Reconstruction Method to combine time use data and measures of emotional affect. Thereby, one is able to detect experienced utility, or in other words to get an approximation of the emotions and feelings individuals experience while performing an activity. Around 1000 employed and unemployed individuals were interviewed and asked to keep a diary on how they use their time during a specific day, their emotions during all these activities, their general life satisfaction and life circumstances. Unemployed and employed individuals can then be compared with respect to all these different domains. The study confirms previous results concerning the lower life satisfaction levels of unemployed individuals. Moreover, there is evidence that the unemployed feel less happy while performing similar activities to the employed. However, given that the unemployed have more time to perform more enjoyable activities compared to being at work, which is associated with rather negative feelings, they are able to compensate this gap in experienced utility by the amount of time the employed do not have. Therefore, average experienced utility does not differ between the unemployed and the employed.

While most studies focus on the incidence on individual unemployment, Di Tella et al. (2001) investigate how individuals' life satisfaction levels vary with their country's inflation and general unemployment rate. For the analysis, they use several waves of cross-sections from the Eurobarometer for 12 different European countries. A two-step procedure is conducted by first estimating microeconometric life satisfaction regressions for each country and calculating the mean residual life satisfaction for each nation in each year. The latter then serves as the dependent variable in a second step to represent a life satisfaction measure that is not explained by personal characteristics. This measure is then regressed on 3-year moving averages of inflation and unemployment rates for each country. Both factors exhibit negative statistically significant effects on life satisfaction. Moreover, unemployment has an even stronger effect, whereas individuals would trade off a 1-percentage-point increase in the unemployment rate for a 1.7-percentage-point increase in the inflation rate.

No Adaptation to Unemployment

Maybe becoming unemployed represents a brief shock for the individual, but over time people get used to it. For example, Clark et al. (2008) show that individuals adapt to many major life events, observed in terms of life satisfaction scores. These include events such as marriage, divorce, widowhood, and the birth of a child. Even though these events partly exert strong

effects on life satisfaction – widowhood in particular – individuals do adapt to all of them. Interestingly, unemployment represents an exception. Even after 5 years of being unemployed, life satisfaction scores have not recovered (however, a little for females). The results of several fixed-effects regressions investigating anticipation and adaptation effects with GSOEP data confirm the detrimental effect of unemployment. Importantly, this effect is not a short-term shock, but a long-lasting issue.

Scarring Effects of Unemployment

If individuals do not adapt to being unemployed, do those people who are employed, but have experienced unemployment in the past, have different well-being levels from those who have never experienced unemployment? Using GSOEP data, Clark et al. (2001) investigate this hypothesis. Fixed effects conditional logit regressions show that indeed, employed individuals with past unemployment experience report lower life satisfaction scores than those without unemployment experience. This finding is interpreted as the 'scarring' effect of unemployment. It appears to be an experience that individuals do not forget. Knabe and Rätzel (2011) examine the reasons for this finding. Again, the rich data set of the GSOEP is used, and fixed effects ordered logit estimations are performed. Interestingly, they show that past unemployment mostly exhibits an effect on current well-being before controlling for future employment prospects and individual fixed effects, including time invariant personality traits. The findings suggest only weak evidence of an effect of past unemployment on the life satisfaction levels of the employed and the unemployed when future employment prospects are taken into account. However, the fear of becoming unemployed in the future appears to have a strong negative effect on life satisfaction. Whereas on the one hand, low job security exhibits a negative effect for the employed, on the other hand insecure reemployment prospects exert a decrease in life satisfaction for the unemployed. A large part of the *scarring* effect of unemployment may therefore also be described as a *scaring* effect.

Unemployment and the Social Norm

Individuals usually do not react only upon their own situation but also compare themselves to the people around. Fulfilling certain expectations is an important factor in this regard, which leads to the question of how social norms may influence the effect of unemployment on happiness. Working can be seen as an essential social norm in many societies for individuals who are of working age. One could therefore assume that individuals who become unemployed in an environment with many other unemployed individuals are not as negatively affected as those who turn into a sort of 'outsider.' A study using BHPS data tests this hypothesis (Clark, 2003). The findings from fixed effects regressions on mental well-being scores from the GHQ-12 reveal a confirmation of the hypothesis: unemployed individuals, especially men, suffer less when the regional unemployment rate is higher, when the partner is also unemployed and when the unemployment rate of other adults in the household is higher. Whereas the results make intuitive sense at first glance, in particular the findings on the household level may have gone the exact other way. If the partner already is unemployed, then the individual's unemployment exerts even higher – especially financial – pressure on the household and may actually even make those individuals suffer more. However, these results seem to confirm a rather robust channel of a social norm effect.

Unemployment and Identity

It has been shown that unemployment causes a drop in well-being that goes beyond the material loss of income. Moreover, the creation of unemployment benefit systems in modern welfare states already contributes to moderate the income loss. In this context, a small strand of the literature investigates the relationship between unemployment and a person's identity as a channel for the drop in well-being. In other words, the individual's perception of 'self' may be threatened upon entry into unemployment. Hetschko et al. (2014) provide an example of an analysis investigating changes in life satisfaction when individuals move from unemployment to retirement. This study uses GSOEP data and performs a difference-in-difference analysis by comparing both unemployed and employed individuals moving into retirement and their potential changes in life satisfaction scores. The findings suggest a clear increase in life satisfaction for those who move to retirement from unemployment, whereas this effect is not found for those who were employed before retirement. The interpretation of this finding relates to the reinstallation of the social norm of not working when being retired compared to not working when being unemployed. The authors refer to this as positive identity effects upon retirement, thus individuals change their feeling of belonging to the new category of 'retirement age.'

With Work: What's Important?

Employment Conditions

While there seems to be a clear positive effect of being employed on subjective well-being, one should furthermore understand how certain job characteristics impact job satisfaction given the individual is actually employed. It is highly probable that while employment itself is important for life satisfaction, different employment conditions may exert differential effects on the specific satisfaction with the job. Cornelißen (2009) investigates exactly this relationship between job characteristics and job satisfaction and moreover, the impact of low job satisfaction scores on job search efforts. Using GSOEP data and conducting fixed effects regressions, relations with colleagues and supervisors, task diversity and job security are identified as main determinants of job satisfaction. Moreover, fringe benefits, the net wage, independence, influence, learning opportunities, and the subjective probability of promotion are also positive, statistically significant determinants of job satisfaction. Further regressions show that job satisfaction is a strong predictor of job search. Job search efforts increase with lower job satisfaction, which makes intuitive sense. However, this influence appears weaker with higher tenure, at a higher age, and in the public sector. Finally, the

higher job search probability is shown to be an important determinant of the probability of a job change.

Promotions

Most careers do not remain static, but rather workers receive certain promotions from time to time. The a priori hypothesis about the direction of the effect of promotions on workers' well-being is unclear. On the one hand, workers may enjoy the higher wage, greater authority and control, more challenging tasks and other privileges related to the new job status and therefore experience an improvement in their well-being. On the other hand, a promotion may be accompanied by more stress through added responsibility and longer working hours. The net effect of promotions on well-being is therefore an empirical question. A relatively recent article examines this effect of job promotions on workers' health and happiness. While several studies have focused on the effect of promotions on job satisfaction, Johnston and Lee (2013) additionally investigate the effect on certain job attributes including job satisfaction as well as on (mental) health and life satisfaction. Moreover, by introducing dummy variables indicating certain time periods before and after the occurrence of a promotion, the study allows for anticipation and adaptation effects. Potential anticipation and selection effects include: changes in well-being resulting from changes in behavior to receive a promotion such as working longer hours, from workers becoming informed about a future promotion, as well as healthier workers having higher chances of being promoted. Using several waves of the Household, Income and Labour Dynamics in Australia survey, the authors estimate a number of regressions on different subjective dependent variables. While they do not find any evidence for anticipation effects, workers indicate increased security, control, a fairer perception of pay, but also more stress and longer working hours 1 or 2 years after the promotion. Moreover, job satisfaction is significantly higher after the promotion. However, 3 years after the promotion, and despite still earning a higher wage and having more control in the job, the positive feelings vanished. The study did not find any significant effects on general health or life satisfaction, but negative effects on mental health 2 years after the promotion, which are primarily driven by anxiety-type feelings. In sum, the authors inquire as to why workers strive to be promoted as their well-being is not largely influenced and because there appears to be a negative effect on mental health. A missing argument at this point could be individuals' higher aspirations and the desire for change even though the resulting effect on well-being is unknown and may not be clearly positive.

Self-employment

A so far neglected here, but important characteristic of a job is the type of employment the worker is engaged in. More specifically, are there differences in job satisfaction depending on whether the worker is self-employed or a regular employee? A large literature is dedicated to this question. One relatively robust result implies that self-employed workers report higher job and life satisfaction (see, e.g., Blanchflower and Oswald (1998) for one of the early contributions in the economics literature). The conventional interpretation of this finding is the greater autonomy that the self-employed enjoy, or in other words 'being one's own boss,' resulting in higher job satisfaction. However, as Blanchflower and Oswald (1998) note, the selection of happier, more optimistic or cheerful people into self-employment may present another interpretation, especially when considering cross-sectional data.

Up to the current date, there have been a number of contributions in the literature that focus on more specific questions regarding self-employment and happiness, e.g., trying to shed light on the apparent satisfaction gap between the self-employed and the regular employed. For example, Millán et al. (2013) differentiate between the impact of self-employment on job satisfaction in terms of type of work and job satisfaction in terms of job security compared to paid employees. They use several waves from the ECHP to estimate generalized ordered logit models. Their results suggest that while the self-employed are more satisfied with the type of work they do, paid employees are more satisfied in terms of job security. These results appear intuitive. However, it seems important to investigate changes in satisfaction measures depending on the status before the transition into self-employment. Individuals who move to self-employment out of unemployment may differ from those who move to self-employment out of regular employment. While the former may see self-employment as a necessity, the latter may rather view it as an opportunity. Binder and Coad (2013) investigate these questions with data from the BHPS using a matching approach to better identify treatment and control groups. Their main results show that those individuals who exit regular employment to become self-employed experience a significant increase in their life satisfaction compared to those staying in regular employment, with the effect even increasing between year 1 and 2 of self-employment. When comparing those moving from unemployment to self-employment with those moving from unemployment to regular employment, no difference in life satisfaction scores were found.

Hanglberger and Merz (2011) find contrasting results to the ones presented so far. They offer a further explanation of the difference between self-employed and regular employed by focusing on anticipation and adaptation effects. They use data from the GSOEP to estimate several fixed-effects regressions. The results show large negative anticipation effects preceding the change from regular employment to self-employment and a large improvement in the level of job satisfaction upon becoming self-employed. However, this increase diminishes after 3 years, in other words individuals adapt to being self-employed. Furthermore, the negative anticipation effects were found for any job change, in particular changing jobs within regular employment as well. The authors conclude that in the long run, the self-employed are not more satisfied than regular employees. These results of course challenge many of the previous findings in the literature and show how important the data structure and quality as well as the type of methodology are for the analysis. It will surely be necessary to perform similar analyses with different data sets and for different countries. Moreover, as many authors note, the group of self-employed is very heterogeneous, including for example farmers and free-lancers. Heterogeneous subgroup effects are therefore important to take into account.

Public and Private Sector

Besides employment-specific conditions, promotions and the importance of the distinction between regular and self-employment, sector-specific conditions are another essential source for differences regarding job satisfaction. Particular attention in this regard has been paid to the difference between working in the public or private sector. Several studies found employees in the public sector on average to be more satisfied than those in the private sector (see, e.g., Luechinger et al., 2008 with a focus on life satisfaction). A recent article by Danzer (2013) aims to disentangle the channels of worker sorting and sector-specific job characteristics. In order to do that, she uses a natural experiment, namely the massive privatization process in post-Soviet countries, in particular in Ukraine. The instrumental variable approach is based on the exogenous probability of working in the private sector in post-Soviet Ukraine based on pretransition job characteristics. The data for the analysis stem from the Ukrainian Longitudinal Monitoring Survey, a nationally representative panel data set. The results suggest a positive public sector job satisfaction gap even after correcting for self-selection. Moreover, a certain part of this public sector premium with respect to satisfaction can be explained by differences in fringe benefits. However, a significant share of the satisfaction gap still remains unexplained, which may be related to job and time flexibility, differences in wage compression across sectors, and perceived job and income uncertainty.

Satisfaction as a Driver of Labor Market Outcomes

While the majority of studies in the area of happiness and work treat the specific well-being variables as an outcome, there is also a small strand emerging that aims to detect potential driving effects of happiness. Examples in the economics literature evolve mainly around the connection of happiness, job search, and labor market outcomes. For example, as already discussed above, Cornelißen (2009), among other things, investigates the effect of job satisfaction on job search.

Furthermore, Clark et al. (2008) use 20 years of GSOEP data to investigate the anticipation and adaptation effects of certain life events on well-being. They include individual fixed effects in the estimations, which help to pick up actual anticipation effects rather than selection. Their findings indicate that future unemployment reduces current well-being, which they interpret as a lead or anticipation effect. Clark (2003) and Mavridis (2010) both use BHPS data to investigate how the drop in mental well-being (GHQ score) affects the future unemployment probability. While Clark (2003) uses a simple probit model, Mavridis (2010) estimates a duration model. The findings suggest that those individuals with a higher drop in mental well-being upon becoming unemployed are less likely to remain unemployed 1 year later. Moreover, they are more likely to have a shorter unemployment duration. These results relate to the fact that those individuals who are hurt more severely from unemployment are the ones who exert more effort to exit unemployment. Clark (2003) moreover finds that those with a higher drop in their mental well-being score were more likely to have looked for work over the past week. In contrast, Gielen and van Ours (2014) find that the drop in life satisfaction upon unemployment entry does not stimulate job finding. They use GSOEP data and also estimate a hazard model. Three factors may induce the different findings: first, Clark (2003) and Mavridis (2010) do not allow the job finding rate to be influenced by unobserved heterogeneity; second, Mavridis (2010) pools women and men; and third, the subjective well-being measures are different. Moreover, Gielen and van Ours (2014) do not find the drop in life satisfaction to affect post-unemployment job quality. Finally, Krause (2013) examines the effect of an unemployed individuals' life satisfaction on future labor market outcomes. The study uses rich survey data from 2007 to 2009 of entrants into unemployment in Germany (the IZA Evaluation Dataset S) to calculate residual happiness. This residual displays higher (or lower) satisfaction levels than would be predicted by a number of demographic and socioeconomic characteristics. Subsequently, the effect of residual happiness on labor market outcomes is analyzed. Moreover, by accounting for the individual's labor market history and information about future job prospects, worries about reverse causality bias can be reduced. The findings indicate a statistically significant inverted U-shaped effect of residual happiness on the future reemployment probability and reentry wage, where the effect on wages even seems to have a cubic shape. Further investigation offers three mechanisms. First, happiness is mainly a predictor for exit into self-employment; second, only the male unemployed experience an effect of happiness upon reemployment; and third, the concept of 'locus of control' and the personality traits of neuroticism and extraversion are main drivers of the baseline effect on regular reemployment and are able to explain the effect on reemployment for males. The nonlinear effects on wages and self-employment are robust to the inclusion of personality traits. These results may give rise to the notion of an optimal level of happiness, which is not necessarily the highest, at least in terms of achievement outcomes.

Even though endogeneity concerns with respect to happiness make it difficult to detect causal effects of happiness as a driver, these few studies show that with the aid of panel data and the constructions of a measure of the drop in happiness upon some event as well as the possibility to control for several sources of bias can help to reduce potential worries. Moreover, it is important to show that happiness may in some cases exhibit a potential source of motivation for behavior. One study circumvents the endogeneity issue by the conduction of an experiment. Here, happiness increases are exogenously induced for one randomly chosen group but not the other (Oswald et al., 2009). The treatment consists of short comedy movies, whereas the control group either watches no movie or a neutral movie. The results show that improved happiness (or positive affect) increases productivity for a paid task.

Conclusions

An ever growing literature is engaged in questions that revolve around the relationship between work and subjective well-being. One large strand of research investigates the extensive margin by analyzing the effect of losing the job and becoming unemployed. Another important field of study focuses on the

determinants of job satisfaction revolving around employment conditions, self-employment, and potential public sector satisfaction premiums. A smaller part in the literature investigates potential driving effects of happiness with respect to labor market outcomes.

Work proves to be a very important determinant for people's well-being. The most obvious evidence for this statement is the detrimental negative effect that individuals experience upon becoming unemployed, which goes beyond a pure income loss effect. Losing one's identity, purpose in society, and contacts to others are identified as main reasons for this nonpecuniary cost of unemployment. Whereas these determinants are essential for the overall satisfaction with life, it has been shown that people experience rather negative emotional affect while working. Moreover, the unemployed are able to make up for the stronger negative emotional affect experienced during an activity compared to the employed with the time the unemployed have for enjoyable activities throughout the day. These results contribute to a more sophisticated view of the unemployed individual's subjective well-being. Importantly, the overall life satisfaction gap between the unemployed and the employed is not questioned by these latter results and remains crucial.

Regarding job satisfaction determinants, five patterns emerge from the literature: (1) relations with colleagues and supervisors, task diversity and job security are main determinants of job satisfaction; (2) job satisfaction predicts job search effort; (3) a promotion affects job satisfaction positively only in the short-term; (4) the self-employed seem to report higher job satisfaction scores which is probably related to greater autonomy; and (5) there seems to be a satisfaction premium in the public sector. Even though some studies lack a clear test of causality and a few others challenge these existing views, the majority of evidence seems to confirm these five patterns. Following these findings as well as Stiglitz et al. (2009), it altogether appears as if happiness research constitutes a valuable complementary source of utility measurement for both academia and policy.

Future research using high quality data structures and advanced methodologies may be able to shed light on certain contradictory results. Panel data emerge as important to take into account difficulties when working with subjective well-being data, such as individual unobserved heterogeneity. Although identifying a clear natural experiment that relates to subjective well-being is not an easy task, a more frequent use of natural experiments or other external shocks such as in, e.g., Danzer (2013) may constitute another potential direction for future research to overcome certain obstacles when comparing subjective well-being data between different groups.

See also: Emotions and Work; Happiness: History of the Concept; Job Satisfaction; Personality and Values at Work; Subjective Wellbeing and Culture; Subjective Wellbeing, Psychology of; Workplace Stress.

Bibliography

Binder, M., Coad, A., 2013. Life satisfaction and self-employment: a matching approach. Small Business Economics 40 (4), 1009–1033.

Blanchflower, D.G., Oswald, A.J., 1998. What makes an entreprenuer? Journal of Labor Economics 16 (1), 26–60.

Clark, A., 2003. Unemployment as a social norm: psychological evidence from panel data. Journal of Labor Economics 21 (2), 323–351.

Clark, A., Diener, E., Georgellis, Y., Lucas, R.E., 2008. Lags and leads in life satisfaction: a test of the baseline hypothesis. The Economic Journal 118, F222–F243.

Clark, A., Georgellis, Y., Sanfey, P., 2001. Scarring: the psychological impact of past unemployment. Economica 68 (270), 221–241.

Clark, A., Oswald, A., 1994. Unhappiness and unemployment. The Economic Journal 104, 648–659.

Cornelißen, T., 2009. The interaction of job satisfaction, job search, and job changes. An empirical investigation with German panel data. Journal of Happiness Studies 10 (3), 367–384.

Darity, W.J., Goldsmith, A.H., 1996. Social psychology, unemployment and macroeconomics. Journal of Economic Perspectives 10 (1), 121–140.

Danzer, N., 2013. Job Satisfaction and Self-selection into the Public or Private Sector: Evidence from a Natural Experiment. IZA Discussion Paper No. 7644, Institute for the Study of Labor (IZA), Bonn.

Di Tella, R., MacCulloch, R.J., Oswald, A., 2001. Preferences over inflation and unemployment: evidence from surveys of happiness. American Economic Review 91, 335–341.

Diener, E., Lucas, R., 2000. Subjective emotional well-being. In: Lewis, M., Haviland, J. (Eds.), Handbook of Emotions, second ed. Guilford, New York, pp. 325–337.

Easterlin, R., 1974. Does economic growth improve the human lot? Some empirical evidence. In: David, P., Reder, M. (Eds.), Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz. Academic Press, New York, London, pp. 89–125.

Frey, B.S., Stutzer, A., 2002. What can economists learn from happiness research? Journal of Economic Literature 40 (2), 402–435.

Gerlach, K., Stephan, G., 1996. A paper on unhappiness and unemployment in Germany. Economics Letters 52, 325–330.

Gielen, A.C., van Ours, J.C., 2014. Unhappiness and job finding. Economica 81 (323), 544–565.

Hanglberger, D., Merz, J., 2011. Are Self-employed Really Happier than Employees? – an Approach Modelling Adaptation and Anticipation Effects to Self-employment and General Job Changes. IZA Discussion Paper No. 5629, Institute for the Study of Labor (IZA), Bonn.

Hetschko, C., Knabe, A., Schöb, R., 2014. Changing identity: retiring from unemployment. The Economic Journal 124 (575), 149–166. http://dx.doi.org/10.1111/ecoj.12046.

Jahoda, M., Lazarsfeld, P.F., Zeisel, H., 2009. Marienthal – the Sociography of an Unemployed Community, fourth printing ed. Transaction Publishers, New Brunswick.

Johnston, D.W., Lee, W.-S., 2013. Extra status and extra stress: are promotions good for us? Industrial and Labor Relations Review 66 (1), 32–54.

Kassenboehmer, S., Haisken-DeNew, J., 2009. You're fired! The causal negative effect of unemployment on life satisfaction. The Economic Journal 119, 448–462.

Knabe, A., Rätzel, S., 2011. Scarring or scaring? The psychological impact of past unemployment and future unemployment risk. Economica 78 (310), 283–293.

Knabe, A., Rätzel, S., Schöb, R., Weimann, J., 2010. Dissatisfied with life but having a good day: time-use and well-being of the unemployed. The Economic Journal 120, 867–889.

Krause, A., 2013. Don't worry, be happy? Happiness and reemployment. Journal of Economic Behavior and Organization 96, 1–20.

Layard, R., 2011. Happiness – Lessons from a New Science, second ed. Penguin, London.

Luechinger, S., Meier, S., Stutzer, A., 2008. Bureaucratic rents and life satisfaction. The Journal of Law, Economics, & Organization 24 (2), 476–488.

Mavridis, D., 2010. Can Subjective Well-being Predict Unemployment Length? Policy Research Working Paper 5293, The World Bank.

Millán, J.M., Hessels, J., Thurik, R., Aguado, R., 2013. Determinants of job satisfaction: a European comparison of self-employed and paid employees. Small Business Economics 40 (3), 651–670.

Oswald, A., Proto, E., Sgroi, D., 2009. Happiness and Productivity. IZA Discussion Paper 4645, Institute for the Study of Labor (IZA), Bonn.

Stiglitz, J.E., Sen, A., Fitoussi, J.-P., 2009. Report by the Commission on the Measurement of Economic Performance and Social Progress. www.stiglitz-sen-fitoussi.fr.

Urry, H.L., Nitschke, J.B., Dolski, I., Jackson, D.C., Dalton, K.M., Mueller, C.J., Rosenkranz, M.A., Ryff, C.D., Singer, B.H., Davidson, R.J., 2004. Making a life worth living – neural correlates of well-being. Psychological Science 15 (6), 367–372.

Winkelmann, L., Winkelmann, R., 1998. Why are the unemployed so unhappy? Evidence from panel data. Economica 65, 1–15.

Happiness: History of the Concept

Ruut Veenhoven, Erasmus University Rotterdam, Rotterdam, The Netherlands; and North-West University, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Notions of a good life have not changed much over time. All the concepts known today can be found in early writings. What has changed is the prominence attached to the aspects of the good life. Another long-term change is that conceptualizations became more specific and that empirical research has revealed their reality links, which made increasingly clear that there is no such thing as 'true happiness.'

Philosophy of the Good Life

As humans are conscious beings, they have always given thought about the quality of their lives. This thought became more systematic after the invention of scripture and the development of professional scholarship. Different notions of quality of life crystallized, often called by the same name of 'happiness.' A long-standing discussion emerged on the relative importance of these and on what constitutes 'true happiness.'

These views on the good life have been described at length in several books on the 'philosophy of happiness,' such as recently in the monumental review by Darrin McMahon (2006) entitled *Happiness: A History*. An overview of this literature is available in the 'Bibliography of Happiness' (Veenhoven, 2013a; subject sections Tb and Tc).

These historical accounts compare schools of thought over time, typically beginning with ancient Greek philosophers, such as Aristotle and the Stoics, and ending with the twentieth-century postmaterialists. The focus is on the ideas as such rather than on the social forces that shaped these. Information is drawn from writings that have stood the centuries, which involve some selection. These historiographies describe views on the good life in scholarly circles and do not inform us about views held in the general public.

Approach of This Article

In this article, I will follow a different approach. I start from a classification of notions of the good life and indicate which of these appeals most to particular parties in the debate. On that base, I make some informed guesses about prominence among scholars in the past. Next, I discuss the effect of recent empirical research on the conception of quality of life.

Notions of the Good Life

The word 'happiness' is used in various ways. In the widest sense, it is an umbrella term for all that is good. In this meaning, it is often used interchangeably with terms like 'well-being' or 'quality of life,' and denotes both individual and social welfare. This use of words suggests that there is no conflict between individuals and society. It further suggests that the quality of life could be measured unidimensionally.

Quality-of-life concepts can be sorted using two distinctions, which together provide a fourfold matrix. I have proposed that classification in an earlier attempt to bring order in the many measures used in contemporary quality-of-life research (Veenhoven, 2000). Here I use it for ordering historical views on the good life.

The first distinction is between chances and outcomes, that is, the difference between opportunities for a good life and the good life itself. The second difference is between outer and inner qualities of life, in other words between 'external' and 'internal' features. In the first case, the quality is in the environment; whereas in the latter, it is in the individual. Robert Lane (1994) made this distinction clear by distinguishing 'quality of society' from 'quality of persons.' The combination of these two dichotomies yields a fourfold matrix. This classification is presented in **Table 1**.

Livability of the Environment

The left-top quadrant denotes the meaning of good living conditions, shortly called 'livability.'

Ecologists see livability in the natural environment and describe it in terms of pollution, global warming, and degradation of nature. Currently, they associate livability typically with preservation of the environment. City planners see livability in the built environment and associate it with such things as sewer systems, traffic jams, and ghetto formation. Here the good life is seen as a fruit of human intervention. In the sociological view, society is central. Livability is associated with the quality of society as a whole and also with the position one has in society.

Life-Ability of the Person

The right-top quadrant denotes inner life chances. That is, how well we are equipped to cope with the problems of life. Amartya Sen (1992) calls this quality of life variant 'capability.' I prefer the simple term 'life-ability,' which contrasts elegantly with 'livability.'

Table 1 Four qualities of life

	Outer qualities	*Inner qualities*
Life chances	Livability of environment	Life-ability of the person
Life results	Usefulness of life	Satisfaction with life

Source: Veenhoven, R., 2000. The four qualities of life: ordering concepts and measures of the good life. Journal of Happiness Studies 1 (1), 1–39.

The most common depiction of this quality of life is absence of functional defects. This is 'health' in the limited sense, sometimes referred to as 'negative health.' Next to absence of disease, one can consider excellence of function. This is referred to as 'positive health' and associated with energy and resilience. A further step is to evaluate capability in a developmental perspective and to include acquisition of new skills for living. This is commonly denoted by the term 'self-actualization.' From this point of view, a middle-aged man is not 'well' if he behaves like an adolescent, even if he functions without problems at this level. Since abilities do not develop alongside idleness, this quality of life is close to the 'activity' in Aristotle's concept of 'eudemonia.' Lastly, the term 'art of living' denotes special life-abilities; in most contexts, this quality is distinguished from mental health and sometimes even attributed to slightly disturbed persons. Art of living is associated with refined tastes, an ability to enjoy life, and an original style of life.

Utility of Life

The left-bottom quadrant represents the notion that a good life must be good for something more than itself. This assumes some higher values. There is no current generic for these external outcomes of life. Gerson (1976: 795) refers to these effects as 'transcendental' conceptions of quality of life. Another appellation is 'meaning of life,' which then denotes 'true' significance instead of mere subjective sense of meaning. I prefer the simpler 'utility of life,' while admitting that this label may also give rise to misunderstanding.

When evaluating the external effects of a life, one can consider its functionality for the environment. In this context, doctors stress how essential a patient's life is to his or her intimates. At a higher level, quality of life is seen in contributions to society. Historians see quality as the addition an individual can make to human culture, and rate, for example, the lives of great inventors higher than those of anonymous peasants. Moralists see quality in the preservation of the moral order, and would deem the life of a saint to be better than that of a sinner. As an individual's life can have many environmental effects, the number of such utilities is almost infinite.

Apart from its functional utility, life is also judged on its moral or aesthetic value. For instance, most of us would attribute more quality to the life of Florence Nightingale than that of a drunk, even if it appeared that her good works had a negative result in the end.

Subjective Enjoyment of Life

Finally, the bottom-right quadrant represents the inner outcomes of life. That is, the quality in the eye of the beholder. As we deal with conscious humans, this quality boils down to subjective appreciation of life. This is commonly referred to by terms such as 'subjective well-being,' 'life satisfaction,' and 'happiness' in a limited sense of the word.

Humans are capable of evaluating their life in different ways. We have in common with all higher animals that we can appraise our situation affectively. We feel good or bad about particular things and our mood level signals overall adaptation. As in animals these affective appraisals are automatic, but unlike other animals it is known that humans can reflect on this experience. We have an idea of how we have felt over the last year, while a cat does not. Humans can also judge life cognitively by comparing life as it is with notions of how it should be.

Most human evaluations are based on both sources of information, that is, intuitive affective appraisal and cognitively guided evaluation. The mix depends mainly on the object. Tangible things such as our income are typically evaluated by comparison; intangible matters such as sexual attractiveness are evaluated by how it feels. This dual evaluation system probably makes the human experiential repertoire richer than that of our fellow creatures.

In evaluating our life, we typically summarize this rich experience in overall appraisals. For instance, we appreciate several domains of life. When asked how we feel about our work or our marriage, we will mostly have an opinion. Likewise, most people form ideas about separate qualities of their life, for instance, how challenging their life is and whether there is any meaning in it. Such judgments are made in different time perspectives, in the past, present, and future. Mostly such judgments are not very salient in our consciousness. Now and then, they pop to mind spontaneously. Though not in the forefront of consciousness, all the time estimates of subjective enjoyment of life can be recalled and refreshed when needed. This makes these appraisals measurable in principle.

Such a subjective evaluation can also concern one's life as a whole. Jeremy Bentham (1789) referred to such appraisal as 'the sum of pleasures and pains' and called it 'happiness.'

Universal Topics

All the above-mentioned notions of the good life figure in classic thought, clearly because they all are of relevance in the human condition. A few examples suffice to illustrate their timeliness.

Livability

Since our life depends very much on external conditions, livability issues will always pop up in reflections on the good life. We see this, for instance, in imageries of paradise, such as described by Alois Hahn (1976) and also in Utopist blueprints of an ideal society on earth, such as by Thomas Moore (1869). These designs of a livable environment typically reflect the problems of the times in which they emerged. Contemporary notions of the good society focus very much on the degree to which they have solved problems of the past, such as poverty, inequality, and oppression.

Life-Ability

Quality of life depends evidently on what we make of the given living conditions, so capability issues have always figured in notions of the good life. In this context, different interest groups have emphasized different competences. 'Bravery' and 'fighting skill' were praised in the warrior classes, spokesmen for which have long dominated in the discourse. The spokesmen themselves, typically moral philosophers, rather emphasized 'wisdom.' Intellectual development was cherished

by bureaucrats, such as the Chinese Mandarins among which the teachings of Confucius were well received. 'Perseverance' and 'rationality' were emphasized in the business classes, such as in the sixteenth-century Protestant Ethic (Weber, 1905). Personal 'autonomy' is more central among contemporary psychologists and pedagogues, who see the good life more in terms of 'mental health' than of 'morality.'

Utility of Life

Since we are thinking animals, hardwired to pick up causal relations, we cannot evade wondering what our life is good for. Consequently, issues of meaning figure typically in discussions of the quality of life. All religions provide answers to that question, which often involve a promise of paradise, entrance to which is typically not free.

In the above-mentioned case of the Protestant Ethic, the meaning of life is seen in being a good steward in God's garden during one's lifetime. In some radical religions, meaning is found in the eradication of heretics. Meaning is also central in secular notions of the good life and is typically seen in furthering social justice and saving mother earth. Postmodern agnostic philosophers also emphasize meaning, but cannot decide what for.

Subjective Enjoyment of Life

Human thinking capacity has not replaced the evolutionary older affect system. Like other animals, we can feel good or bad and tend to approach situations where we feel good and avoid conditions that feel bad. Facial reflections of feeling happy or unhappy are recognized all over the world (Ekman and Friesen, 1975). As a result, hedonic appraisals are always on the scene.

In Christian mythology, Adam and Eva are assumed to have felt good before being expelled from paradise and to have suffered much after in the harsh conditions of real life. Pictorial representations of earthly paradises typically show smiling faces, such as in Chinese communist propaganda (e.g., Laing, 2004: Chapter 10).

Subjective enjoyment of life is also a recurrent topic in philosophy. Among the early Greeks, 'hedonists' such as Democritus thought of the good life as a life one likes. This view is also found in the eighteenth-century utilitarianism, such as professed by Bentham (1789).

Subjective enjoyment of life is also prominent in real-life experience. Most modern people think of it every day (Freedman, 1978). The salience of this notion appears also in the promptness of responses to survey questions about life satisfaction and a low percentage 'do not know' answers; typically less than 1% (Veenhoven, 2013b).

Historical Variations

Still these topics have not always been equally prominent in conceptualization of the good life.

Classic Focus on Morality

Virtue is central in much classic philosophy, probably because most philosophers made their living as moral advisors. In that context, personal capabilities such as honesty and faith are emphasized (right-top quadrant in **Table 1**) and also manifestations of utility, such as martyrdom (left-bottom quadrant). This emphasis on moral behavior seems to have been more pronounced in historical conditions where morality was at its weakest.

Some classic philosophers have also given thought about what makes for a good society, such as Plato (380 BC) in his 'Politeia' and in the writings of Confucius (Veenhoven and Guoquing, 2008). This emphasis is typical for developed states and is therefore more prominent in contemporary nation-states than it has ever been in the past. Today, all modern states monitor the quality of the living conditions they provide, using sophisticated systems of 'social indicators,' which become increasingly internationalized.

Moral philosophers were typically mixed about the worth of subjective enjoyment of life. Most accept it as a by-product of living a good life, rather than a manifestation of the good life as such. One reason for this reservation is in their professional involvement with moral disciplining. Another reason is probably in the poor quality of life in agrarian societies, which appears in historically high rates of homicide, poor health, and consequently a short lifetime. This historical deep point in quality of life is depicted in **Figure 1**. This misery is likely to have called for the glorifying of suffering and in the projection of happiness in afterlife, a demand to which the churches of these days have responded.

Modern Emphasis on Subjective Well-Being

During the Middle Ages, it was widely believed that happiness is not possible in earthly life and that the basis of morality is in the word of God. These views were contested in the eighteenth-century 'Enlightenment'; happiness came to be seen as attainable and morality was regarded as artificial. A lively discussion

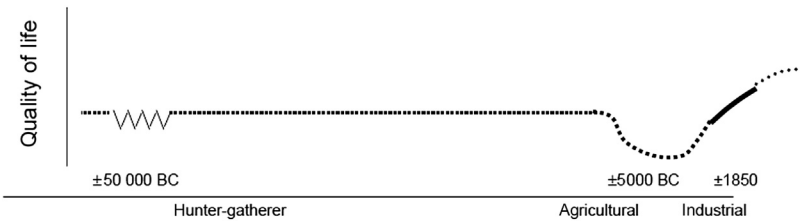

Figure 1 Years lived in good health over human history. Source: Veenhoven, R., 2010. Life is getting better: societal evolution and fit with human nature. Social Indicators Research 97, 105–122.

of the relation between happiness and morality emerged (Mauzi, 1960; Buijs, 2007) and in that climate an instrumental view on morality appeared, in which ethical codes are seen as ways of securing a happy life.

Much of that enlightened thought is reflected in Bentham's (1789) 'Introduction to morals and legislation.' Bentham argues that the moral quality of action should be judged by its consequences on human happiness and in that line he claims that we should aim at the 'greatest happiness for the greatest number.' Bentham defined happiness in terms of psychological experience, as 'the sum of pleasures and pains,' that is, in the right-bottom quadrant of **Table 1**. His philosophy is known as 'utilitarianism,' because of its emphasis on the utility of behavioral consequences.

Though welcomed in enlightened circles in the eighteenth century, this view was rejected by the dominating ideologies of the nineteenth century and the first half of the twentieth century. The strongest opposition came from the churches, which saw little value in earthly happiness and professed a principalistic morality based on the 10 commandments. The liberals of that time also had reservations about the greatest happiness principle; in their power struggle with the aristocracy they rather emphasized freedom. Likewise, the socialists who entered the scene in the late nineteenth century prioritized social equality. Nationalism dominated in the first half of the twentieth century when the two World Wars took place, and nationalists were more interested in national glory than in individual happiness.

As a result, interest in happiness declined and one of the indications is a sharp drop in the use of the word in book titles after 1800 (Buijs, 2007). When I took a course in social philosophy in the 1960s, I encountered Bentham's greatest happiness principle as a historical topic. Yet, a revival took off at that time. Ever more books on 'How to be happy' appeared in the bookshops and happiness also became a topic in the political agenda.

This renewed interest in greater happiness was driven by several factors. One of these is that many pressing ills had been overcome, at least in the Western world. The second half of the twentieth century was characterized by peace, democracy, and an unprecedented rise in the standard of living. This gave way to more positive goals, such as health and happiness. Another factor was the development of a multiple-choice society in which individuals can choose how to live their life and therefore get interested in what way of life will be most satisfying. The rise of happiness in the political agenda was also facilitated by the weakening of the earlier ideological opposition mentioned above. The churches had declined, the liberals and the socialists got their deal, and nationalism had lost much of its appeal.

The effect of these long-term ideological shifts was much amplified by the emergence of empirical research on quality of life.

Empirical Research on Quality of Life

Quality of life became a subject of empirical research in the second half of the twentieth century, when modern welfare states developed and social engineering came to require systematic information about the well-being of the general population in order to identify needs and track progress in meeting them. Specialized research institutes were established in all developed nations and periodical welfare surveys were started. This 'social indicator revolution' (Bauer, 1967) has affected the thought about the good life in several ways.

Empirical research gave rise to greater conceptual differentiation than armchair theorizing had done in the past, both because measurement pressed to greater precision and findings revealed unexpected differences between aspects of the good life. For instance, the classic notion of 'wisdom' has crumbled in a set of rather loosely related traits (Bergsma and Ardelt, 2012).

The new quality-of-life research has also augmented the growing interest in subjective appreciation of life, that is, in the quality of life denoted in the right-bottom quadrant of **Table 1**. Life satisfaction appeared to be easily measurable in survey research, and in fact better measurable than most of the other qualities of life mentioned in scheme 1 (Veenhoven, 2000). As a result, subjective happiness became more tangible a topic; research showed how happy we are in this sense and also indicated how happy we can realistically be. Findings of this kind are well covered by the media, which has also augmented the rising prominence of subjective well-being. Moreover, social scientists have surpassed philosophers as quality-of-life specialists and also for that reason the above-mentioned philosophical reservations have lost appeal.

Rise of Research on Life Satisfaction

Empirical research on life satisfaction took off as a topic in 'social indicators research,' which emerged in the 1970s. The number of papers on this subject in the journal 'Social Indicators Research' grew so much that the specialized 'Journal of Happiness Studies' was split off in 2000. Subjective enjoyment of life became a side topic in medical 'quality-of-life research' that developed in the 1980s and in both 'positive psychology' and 'happiness economics,' which appeared around the year 2000. The number of scientific publications on happiness has grown steeply since the 1960s, as can be seen in **Figure 2**. In all these research, the leading question is what makes people subjectively happy with the major motivation being to advance this kind of happiness. Results of this strand of research are gathered in the World Database of Happiness (2013), which currently involves some 20 000 findings, some 5000 'distributional' findings on how much people like the life they live, and about 15 000 'correlational' findings on things that go together with subjective enjoyment of life.

Observed Links with Other Qualities of Life

Many of the correlational findings on life satisfaction concern aspects of the good life denoted in **Table 1**, in particular with the environmental conditions of the top-left quadrant and the personal capabilities in the top-right quadrant.

The data confirm many of the educated guesses philosophers have made in the past, such as that average life satisfaction is typically higher in democratic nations and that happy individuals tend to abide more to moral principles. Yet, the data also show much variation, not all citizens flourish equally

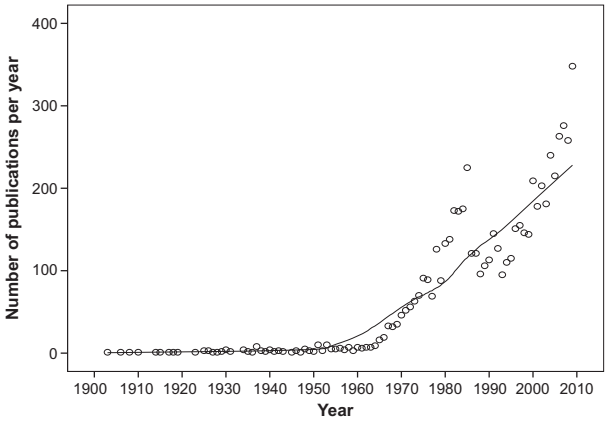

Figure 2 Scientific publications on life satisfaction since 1900. Source: Veenhoven, R., 2013a. Bibliography of Happiness. World Database of Happiness. Erasmus University Rotterdam, The Netherlands. Available at: http://wor;ddatabasepofhappiness.eur.nl/hap_bib/bib_fp.php.

well under democracy and morality is no sure ticket to happiness.

Some of the presumed links between qualities of life appear not to exist. For instance, average happiness in nations is unrelated to equality of incomes and negatively related to religiousness (Berg and Veenhoven, 2009; Veenhoven & Berg, 2010). Likewise, studies at the individual level found no relation with happiness and IQ (Choi and Veenhoven, 2012).

This means that the different varieties of the good life do not always go together. There is no such thing as 'true' happiness that covers all qualities, but rather quality configurations, in which some qualities strengthen each other, while others conflict. A challenge of future research is to chart configurations that are optimal in given historical conditions.

Conclusion

Notions of the good life are largely given by the human condition and pop up at all times in the great civilizations, though typically in the context of current intellectual discourses. Relative prominence varies across times and cultures. Subjective life satisfaction has gained prominence in contemporary Western society. This is partly due to the development of multiple-choice society, and another reason is that empirical research is more possible on the aspect of this good life.

See also: Applied Social Research, History of; Emotion: History of the Concept; Empirical Social Research, History of; Enlightenment: Impact on the Social Sciences; Psychology, History of (Twentieth Century); Sociology, History of; Utilitarian Social Thought, History of.

Bibliography

Bauer, R.A., 1967. Social Indicators. MIT Press, Cambridge, USA.
Berg, M., Veenhoven, R., 2009. Geloof en geluk in landen (Religion and happiness in nations). Ethiek en Maatschappij 12, 49–69.
Bergsma, A., Ardelt, M., 2012. Wisdom and happiness: an empirical investigation. Journal of Happiness Studies 13, 481–499.
Bentham, J., 1789. Introduction to the Principles of Morals and Legislation. Payne, London.
Buijs, P.M., 2007. De eeuw van het geluk: Nederlandse opvattingen over geluk ten tijde van de verlichting 1658–1835 (The Age of Happiness: Dutch Opinions on Happiness During the Enlightenment (Ph.D. dissertation)). University of Utrecht, Netherlands.
Choi, Y., Veenhoven, R., 2012. Does intelligence boost happiness? Smartness of all pays more than being smarter. International Journal of Happiness and Development 1, 5–27.
Ekman, P., Friesen, P.W., 1975. Unmasking the Face. Prentice Hall, Englewood Cliffs.
Freedman, J.L., 1978. Happy People. Harcourt Brace Janovitz, New York, USA.
Gerson, E.M., 1976. On quality of life. American Sociological Review 41, 793–806.
Hahn, A., 1976. Soziologie der Paradiesvorstellungen (Sociology of Imagery of Paradise). NCO Verlag, Trier, Germany.
Lane, R., 1994. Quality of life and quality of persons: a new role for government? Political Theory 22, 219–255.
Laing, E.J., 2004. Selling Happiness: Calendars, Posters and Visual Culture in Early 20th Century Shanghai. University of Hawaii Press, USA.
Mauzi, M.R., 1960. L'idée du bonheur dans la literature et la pensee Francaise au 18e sciecle (The idea of happiness in French literature and thought in the 18th century). Librarie Armand Colin, Paris, France.
McMahon, D.M., 2006. Happiness: A History. Atlantic Monthly Press.
Moore, T., 1869. Utopia. London, England.
Plato ±380 BC. Politeia. The Republic (A. Lindsey, Trans.). Dover Publications, UK, 2000.
Sen, A., 1992. Capability and wellbeing. In: Nussbaum, M., Sen, A. (Eds.), The Quality of Life. Clarendon, Oxford.
Veenhoven, R., 2000. The four qualities of life: ordering concepts and measures of the good life. Journal of Happiness Studies 1–1, 1–39.
Veenhoven, R., 2010. Life is getting better: societal evolution and fit with human nature. Social Indicators Research 97, 105–122.
Veenhoven, R., 2012. Happiness in Nations. World Database of Happiness. Erasmus University Rotterdam. http://worlddatabaseofhappiness.eur.nl/ha-_nat_nat_fp.htm (assessed on 01.10.11.).
Veenhoven, R., 2013a. Bibliography of Happiness. World Database of Happiness. Erasmus University Rotterdam, The Netherlands. Available at: http://wor;ddatabasepofhappiness.eur.nl/hap_bib/bib_fp.php.
Veenhoven, R., 2013b. World Database of Happiness: Archive of Research Findings on Subjective Enjoyment of Life. Erasmus University Rotterdam, Netherlands. Available at: http://worlddatabaseofhappiness.eur.nl.
Veenhoven, R., Berg, M., 2010. Income inequality and happiness in 119 nations. In: Greve, Bent (Ed.), Social Policy and Happiness in Europe. Edgar Elgar, UK, pp. 175–194.
Veenhoven, R., Guoquing, Z., 2008. Ancient Chinese philosophical advice: can it help us find happiness today? Journal of Happiness Studies 9, 425–443.
Weber, M., 1905. The Protestant Ethic and the Spirit of Capitalism (Stephen Kalberg, New Trans.). Routledge, New York, USA, 2012.

Harlow, Harry Frederick (1905–81)

Stephen J Suomi, National Institute of Health, Bethesda, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Harry F. Harlow, an American psychologist remembered for his studies of the development of the mother–infant attachment bond in monkeys, made numerous other seminal empirical contributions including developing a standardized procedure for testing the cognitive capabilities of nonhuman primates, demonstrating the capacity for abstract learning in rhesus monkeys, characterizing the long-term effects of different early social experiences, and reversing social deficits associated with early isolation rearing. Harlow's findings not only influenced researchers, theorists, and clinicians who focused on human development, but also provided the basis for several new areas of investigation.

Overview

Although Harry F. Harlow is most often remembered for his pioneering studies of the development of the mother–infant attachment bond in monkeys, he made numerous other seminal empirical contributions over his 44-year career at the University of Wisconsin. These include developing a standardized procedure for testing the cognitive capabilities of nonhuman primates, demonstrating the capacity for abstract learning in rhesus monkeys, characterizing the long-term effects of different early social experiences, and reversing social deficits associated with early isolation rearing. Harlow's findings from these and other studies with monkeys not only had major influences on contemporary researchers, theorists, and clinicians focusing on issues concerning human development but also provided the basis for several new areas of investigation. These areas of investigation continue to generate active research with both humans and nonhuman primates almost four decades after his retirement in 1974.

Indeed open any Introductory Psychology textbook published since 1960; start thumbing through the pages, and more likely than not you will come across a photograph of a rhesus monkey infant clinging to a cloth-covered dummy adjacent to another dummy, this one with a wire surface and a protruding milk-filled bottle. Along with the photo will be a description of Harry Harlow's classic 'surrogate mother' experiments, in which he dramatically demonstrated that tactile contact was more important than feeding in establishing a monkey infant's bond to its surrogate. Harlow's conclusions from these studies that "there is more to mother love than mere milk" (Harlow, 1958), had important and lasting consequences. They effectively demolished the prevailing theoretical consensus that an infant's primary social relationship is formed as a direct consequence of the feeding process, be it through oral gratification, as advocated by orthodox psychoanalysts, or by its reinforcement history, the contemporary behaviorist view. More importantly, Harlow's findings also provided compelling empirical support for specific biological foundations underlying Bowlby's attachment theory, which went on to become the predominant view of social and emotional development over the latter one-third of the twentieth century (Suomi et al., 2008). But Harlow's contributions to, and influences on, both psychology and a remarkably diverse range of other disciplines went far beyond his legacy as the father of the surrogate mother. His numerous ground breaking discoveries over a lifetime of research with monkeys have not only stood the test of time and scientific debate, but indeed provided the very basis for many new areas of scientific inquiry that still remain active.

Background, Career, and Accomplishments

Harry F. Harlow was born as Harry F. Israel on 31 October 1905 in Fairfield, Iowa, where he spent his childhood years. He left this small Midwestern town to go to college, initially matriculating at Reed College for a year (as an English major) before transferring to Stanford University, where he changed his major and then earned both his BA (1927) and PhD (1930) in psychology. His graduate advisors at Stanford, concerned that his family name might restrict his academic employment opportunities during a time of economic depression and latent anti-Semitism, suggested a name change. Harlow obliged, and later in life he often quipped that they had given him not only his degree but his very name as well. In the fall of 1930, he was hired as an Assistant Professor in the Department of Psychology at the University of Wisconsin in Madison, where he remained for the rest of his professional career. Following his retirement in 1974 Harlow moved to Tucson, Arizona, where he served as an Emeritus Professor of Psychology at the University of Arizona until his death on 3 December 1981.

Although Harlow's graduate training in comparative psychology had been largely limited to research with rodents, he began working with nonhuman primates shortly after his arrival at Wisconsin, carrying out his initial studies with a variety of monkeys and apes then residing in Madison's public Villas Park Zoo. He quickly discovered that the standard types of conditioning and memory tests then being used in studies with rats or pigeons were simply not adequate to tap the considerable cognitive capabilities of his nonhuman primate subjects. He and his students therefore began creating new tasks and test batteries that were substantially more challenging. Shortly after moving into his own laboratory building on the university campus, Harlow designed and built the first prototype of an apparatus that would profoundly alter the course of research in primate cognition and that continues to be used in primate laboratories around the world. This testing device was named the Wisconsin General Test Apparatus; it enabled an experimenter to present a large number of discrete learning and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61043-7

memory tasks in a rapid and highly standardized fashion to subject after subject (Harlow and Bromer, 1938). At the same time, he began directing most of his efforts away from broad cross-species comparisons to more focused studies using a single species – the rhesus monkey (*Macaca mulatta*), an unusually hardy and adaptive species that thrived in captivity.

Armed with a laboratory specifically designed to house and test primates, in which individual monkeys could be maintained in good physical health for year after year, and a testing apparatus that enabled the same monkey to be tested repeatedly under standardized conditions, Harlow began to change the field of animal learning. In the 1940s, he and his students created a battery of learning and memory tasks that provided a standardized 'intelligence test' for his monkeys. He then proceeded to study cortical localization of learning capabilities by lesioning different brain areas and recording subsequent differential patterns of deficits across the various components of the test battery. This research established his credentials as a physiological psychologist, and it opened up areas of inquiry and an experimental methodology that continue to be pursued in neuroscience research.

In 1949, Harlow made a major conceptual and methodological breakthrough with his discovery of 'learning sets' in rhesus monkeys. He found that monkeys presented with a long series of six-trial two-choice discrimination problems, soon learned to achieve near-perfect performance on the second and all subsequent trials of each problem. This demonstrated unequivocally that the monkeys had acquired a general strategy for problem solution, i.e., an abstract concept (in his words, 'learning to learn'), rather than a product of simple associative learning (Harlow, 1949). This finding rendered untenable neo-Decartes characterization of animals as 'reflex machines,' and it ultimately has led to increasingly more sophisticated studies of primate cognitive capabilities that foster direct comparisons with those of young human subjects. Largely because of this discovery, Harlow was elected to the National Academy of Sciences in 1951.

Harlow became interested in the ontogeny of rhesus monkey cognitive capabilities in the early 1950s, when he developed a program of research designed to precisely establish when specific cognitive capabilities first emerged during development and when they subsequently achieved adult levels. He demonstrated that infant monkeys could be both classically and operantly conditioned and performed associative learning tests at adult levels as early in life as such tests could be administered. In contrast, monkeys did not achieve adult levels of performance on more abstract problems such as learning set and oddity until their juvenile and adolescent years, respectively (Harlow, 1959). In the process of carrying out these studies, he also observed that young monkeys would solve problems and otherwise engage their physical environment in the absence of appetite rewards. This led him to conclude that in many circumstances curiosity was a more powerful motivator than a reduction of any primary drive (Harlow, 1953). These observations provided the conceptual foundation for the 'surrogate mother' studies cited above.

In the 1960s, Harlow shifted the major focus of his research from primate cognition to the development of social behavior, both normal and abnormal. He developed the concept of 'affectional systems,' in which he characterized emerging relationships formed by monkey infants with their mothers, peers, and others in the social group as differing in qualitative fashion and serving complementary roles in the socialization process (Harlow and Harlow, 1965). He and his students also studied the consequences of blocking the formation of different affectional systems via various degrees and duration of social deprivation initiated at birth (Harlow and Harlow, 1962, 1969). These studies established convincingly the importance of early social experiences for the development of species-normative adult social activities, including reproduction and maternal behavior, and they provided the conceptual and methodological basis for current studies of early experience effects on brain as well as behavior. In recognition for this research Harlow was awarded the National Medal of Science, America's highest official scientific award, in 1967.

Harlow continued his investigations of abnormal behavior during his final years at Wisconsin, generating two more areas of research in the process. First, he and his students demonstrated that the devastating behavioral consequences of total social isolation for the first 6 months or one year of life, previously thought to be permanent, could be largely reversed by exposing the monkeys upon emergence from isolation to conspecific 'therapists' deliberately chosen to not only be socially normal but chronologically younger than the isolates themselves (Suomi and Harlow, 1972; Novak and Harlow, 1975). These findings directly refuted a strict 'critical period' interpretation of social deprivation effects, providing compelling evidence of developmental plasticity and once again opening up a new area of study that remains at the forefront of developmental neuroscience research today. At the same time, Harlow set out to create a rhesus monkey model of human affective disorders, relying on earlier observations by Rene Spitz and Bowlby that certain forms of social deprivation were associated with infant and childhood depression. He and his students used a variety of social separation procedures to precipitate analogous behavioral states in young monkeys and then employed both social and pharmacological interventions to reverse to behavioral disorders (Harlow and Suomi, 1974; Suomi and Harlow, 1977). In recognition of these efforts, Harlow received the Kittay Scientific Award in 1975, the most prestigious award in the field of psychiatry at that time.

Harlow's influence on the field of psychology extended far beyond his empirical contributions, which resulted in over 323 original scientific publications over his 40-year career (a complete bibliography of Harlow's works may be found in Suomi and LeRoy, 1982). He served as Editor for the *Journal of Comparative and Physiological Psychology* for over a quarter century. He also took an active role in many scientific societies, including the National Academy of Sciences, the Society for Experimental Psychology, and the American Psychological Association, for which he was elected President in 1958. Harlow served as the Chief Psychologist for the US Army during the Korean War, and he subsequently was appointed to numerous US government planning and review committees. He played a major role in the establishment of the US Regional Primate Research Center System in the early 1960s and was instrumental in establishing the University of Wisconsin as the site for one of the seven original centers, which he directed from 1964 to 1971.

Despite all of his research activities and administrative responsibilities, Harlow maintained an active teaching role throughout his career. Not only did he train numerous graduate students and postdoctoral fellows (a total of 35 students completed their PhD dissertations under his tutelage), but he was also legendary as an undergraduate instructor, especially for Introductory Psychology, which he taught nearly every year during his tenure at Wisconsin. In addition, Harlow worked hand-in-hand with veterinary and technical personnel to establish numerous basic advances in primate husbandry, including the development of the first self-sufficient primate colony, the first large-scale primate neonatal nursery, and various quarantine and screening procedures that are still used.

Harlow was a colorful character who enjoyed dealing with controversy. He delighted in generating empirical findings that directly challenged existing theoretical dogma. In addition to his considerable skills as a writer and editor, he was a gifted public speaker, who for years was one of the most sought-after scientific lecturers in the country. As a result, he was able to present his findings not only to psychologists around the world but also to representatives of numerous other scientific disciplines, as well as to lay audiences. Indeed, 30 years after his death Harlow's scientific contributions are still cited not only by comparative, physiological, developmental, cognitive, and clinical psychologists but also by others in fields ranging from anthropology to zoology. Clearly, he was much more than the 'father of the cloth mother.'

See also: Cognitive Neuroscience; Learning to Learn; Psychologism, History of; Social Learning and Culture in Nonhuman Organisms.

Bibliography

Harlow, H.F., 1949. The formation of learning sets. Psychological Review 56, 51–65.
Harlow, H.F., 1953. Mice, monkeys, men, and motives. Psychological Review 60, 23–32.
Harlow, H.F., 1958. The nature of love. American Psychologist 13, 673–685.
Harlow, H.F., 1959. The development of learning in the rhesus monkey. American Scientist 47, 459–479.
Harlow, H.F., Bromer, J., 1938. A test-apparatus for monkeys. Psychological Record 2, 434–436.
Harlow, H.F., Harlow, M.K., 1962. Social deprivation in monkeys. Scientific American 207, 136–146.
Harlow, H.F., Harlow, M.K., 1965. The affectional systems. In: Schrier, A.J., Harlow, H.F., Stollnitz, F. (Eds.), Behavior of Nonhuman Primates, vol. II. Academic Press, New York, pp. 287–334.
Harlow, H.F., Harlow, M.K., 1969. Effects of various mother-infant relationships on rhesus monkeys behaviors. In: Foss, B.M. (Ed.), Determinants of Infant Behaviour, vol. IV. Methuen, London, pp. 15–36.
Harlow, H.F., Suomi, S.J., 1974. Induced depression in monkeys. Behavioral Biology 12, 273–296.
Novak, M.A., Harlow, H.F., 1975. Social recovery of monkeys isolated for the first year of life: 1. Rehabilitation and therapy. Developmental Psychology 11, 453–465.
Suomi, S.J., Harlow, H.F., 1972. Social rehabilitation of isolate-reared monkeys. Developmental Psychology 6, 487–496.
Suomi, S.J., Harlow, H.F., 1977. Production and alleviation of depressive behaviors in monkeys. In: Maser, J., Seligman, M.E.P. (Eds.), Psychopathology: Experimental Models. W H Freeman, San Francisco, CA, pp. 131–173.
Suomi, S.J., LeRoy, H.A., 1982. In memoriam: harry F. Harlow (1905–1981). American Journal of Primatology 2, 319–342.
Suomi, S.J., van der Horst, F.C., van der Veer, R., 2008. Rigorous experiments on monkey love: an account of Harry F. Harlow's role in the history of attachment Theory. Intergrative Psychological and Behavioral Science 42, 354–369.

Harmony and Social Work Practice

Terry Tse Fong Leung, The Chinese University of Hong Kong, Hong Kong, People's Republic of China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although harmony is not a popular catchphrase in the social work literature, the philosophical tenets associated with the notion of harmony underpin the mission and vision of the social work profession. This article will identify the philosophical tenets in the notion of harmony, and the ways that the relational, personal, and ecocentric arena of harmony is manifested in social work practice.

Harmony – Its Philosophical Tenets

Originating from the Latin word 'harmonia,' harmony literally means 'concord of sounds,' suggesting the coming together of different things (Li, 2008: p. 90). In Greek philosophy, the inevitable tension and struggles between opposites are recognized as the universal and creative force of harmony (Guthrie, 1962).

> The things which were like and related needed no harmony; but the things which were unlike and unrelated and unequally arranged are necessarily fastened together by such a harmony, through which they are destined to endure in the universe. (Freeman, 1983: p. 74)

Since all things are subject to internal tension that emerges from differences, harmony is manifested as 'the opposites in concert' (Freeman, 1983: p. 25), the 'accordance of the discordant,' 'a unity of many mixed,' and 'an agreement between the disagreeing' (Freeman, 1983: p. 75).

> From things that differ comes the most beautiful harmony. (Freeman, 1983: p. 25)

Harmony is probably the most cherished ideal in Chinese culture (Li, 2006). Ancient Chinese philosophers describe harmony as "various sounds responding to one another in a mutually promoting, mutually complementing, and mutually stabilizing way" (Li, 2006: p. 584).

> A single sound is nothing to hear, a single color does not make a pattern, a single taste does not satisfy the stomach, and a single item does not harmonize. (Shi Bo, a pre-Confucian scholar-minister, translated by Li, 2006: p. 584)

Like their counterparts in the West, ancient Chinese philosophers conceive diversity as a precondition of unity in the notion of harmony. Harmony presupposes differences and precludes the 'overpresence of sameness' (Li, 2006: p. 591). Confucius captures from this emphasis on the necessity of differences, an ethical and relational nexus in the ideal of harmony. In the *Analects of Confucius* (collection of ideas by Confucius and his contemporaries), Confucius preaches that

> The noble man is in harmony but does not follow the crowd; the inferior man follows the crowd but is not in harmony. (*Analects* 13.23, translated by Muller, 2012)

Ability to stand up for diversity in the quest for harmony is an ethical standard Confucius sets for the noble person. Hence, harmony is different from mindless concordance with the others. In relational terms, it is manifested in respect for different opinions and ability to work with different people. Li (2006) identifies a pragmatic attitude in addressing relational issues in the Confucian ideal of harmony.

> The attitude of harmony has a strategic significance. It makes us more willing to engage in negotiation, more willing to compromise, and less willing to resort to confrontation and conquest. It enables us to take into consideration the whole picture when considering an issue and to give each party its due. It is thus more conducive to peaceful solutions to the world's problems. (Li, 2006: p. 595)

In Taoist philosophy, harmony is a state of dynamic equilibrium that is achieved through a ceaseless process of balancing and rebalancing (Lee et al., 2009; Leung et al., 2009; Wang and Zou, 2011). Taoism underlines contradictory and opposing forces in the natural world, and metaphorically represents them as 'yin' and 'yang.' In ancient Chinese metaphysics, 'yin' and 'yang' refers to two cosmological forces manifested in many natural and social dualities such as "dark and light, female and male, below and above, cold and hot, earth and heaven, grassroots and elites" (Wang and Zou, 2011: p. 39). If well balanced or harmonized, these seemingly dualistic opposites can "exist within each other in a symbiotic relationship and interact with each other to form a holistic, dynamic and cyclical unity" (Wang and Zou, 2011: p. 41). By emphasizing the dialectic interplay of 'yin' and 'yang,' philosophical Taoism embraces dualistic holism and highlights mutuality and interdependence for achieving harmony within the person as well as between human beings and nature (Lee et al., 2009).

In sum, harmony in the philosophical sense captures the spirit of an ideal way of life, which can take place in three

different levels: (1) between individuals – at the relational level in the family, the community, the nation, and the world, whereby a harmonious relationship presupposes different perspectives and views among different social and ethnic groups; (2) within the individual – where various parts of one's body and various pursuits in life can be integrated into a well-functioning, organic whole; and (3) between human beings and the natural universe – in a healthy interplay where each gets its due (Li, 2006: p. 588). As such, harmony is a fundamental attitude with which people handle the various problems confronting life. As a profession destined to improve human well-being and the conditions of life, social work has a role to play in the harmonization processes of all three arenas.

The Relational Arena of Harmony in Social Work Practice

Social work has always celebrated the vision of a harmonious society in which human and community well-being can be secured amid differences and diversified interests. 'Promoting collective social harmony and social stability' is included as a core purpose of social work by the International Association of Schools of Social Work (IASSW) and International Federation of Social Workers (IFSW) in their global standards for social work practice (International Association of Schools of Social Work, 2004; International Federation of Social Workers, 2012). In the social work mission of maintaining social harmony, attention to differences with and among social work clients is always in place. Maintaining the social work value of 'starting where the client is' requires conscious vigilance of the individualized circumstance and cultural diversity of clients with whom the social workers work (Hugman, 1996). Since early days of the profession, individual differences among clients have also been recognized and respected in the important social work value of acceptance (Biestek, 1957; Reynolds, 1932). Acceptance is about accepting the person but not necessarily accepting their behaviors. Feelings of affiliation with the clients can coexist with respect for their autonomy, without imposing on them heavy-handed criticism or pressure to change (Berlin, 2005). By encouraging social workers consciously to restrain from attempting to mold clients to fit their own sensibilities, acceptance tempers the controlling and demeaning aspects of compassion (Berlin, 2005). The acceptance principle in social work, therefore, echoes the relational ideal of harmony to 'agree between the disagreeing.'

Meanwhile, social work represents social intervention to improve balance and harmony in society by mediating between the mainstream and dispossessed populations. As stated in the Global Standards of the IFSW and IASSW, the purpose of social work practice is to eliminate the social exclusion of marginalized, dispossessed, vulnerable, and at-risk groups at the points where people interact with their environments (International Association of Schools of Social Work, 2004; International Federation of Social Workers, 2012). In mediating between diversified groups for social harmony, social work is, however, caught in constant argument about its role in the balance between unity and differences (e.g., Bell, 2012; Berlin, 2005; Borden, 2010; Dominelli, 2002; Fook, 2002; Jones and Truell, 2012; Jose, 2010; Webb, 2009). Since early days of the profession, social work has been emphasizing its socialization function, which supports unity more than differences. Marston and McDonald (2012) describe social work as part of the modernist impulse to create a better world conducive to the normative definition of wellness. Central to the modernist project of social work is a positivist paradigm, which often carries a dominant narrative that accommodates itself to prevalent cultural values and prevailing institutional goals (Reisch and Jani, 2012). Feminist social work academics and practitioners are critical of the controlling function of social work as expressed in the positivist paradigm (Stanley and Wise, 1993; Dominelli and McLeod, 1989; White, 2006). This modernist impulse was, however, salient in the early days of the social work profession, when middle-class friendly visitors involved in creating helping as a career option subscribed to a moral view that clients could significantly reduce their problems by submitting to their advice and example (Berlin, 2005). In the 'tough love' offered to get clients work hard for what was defined by early social workers as the 'clients' best interests,' conformity to an existing social order was forged, and differences in ways that clients registered their experiences became a secondary concern (Berlin, 2005; Jordan, 2001).

This imperative to compromise individual differences for conformance to normative social expectations is perpetuated in contemporary social work practice. Emphasizing the socialization of individuals and their adjustment to their social environments, this strategy of achieving social harmony is best exemplified by what Dominelli (2002) refers to as 'therapeutic approaches' to social work, in which clients are assisted to better understand themselves and their relationships with others in the course of adjusting to the normative social order. These clinical and psychotherapeutic models in social work practice have greater concern for settling clients in existing social realities than handling the complexities of negotiating different and competing claims (Gibelman, 1999; Reisch and Jani, 2012). As acquiescence to prevalent social realities is accepted as the guiding principle of practice, difference is captured in pathological terms as inferior or a deficit. The 'othering' process in which people different from the mainstream majority are cast in a subordinate status treats clients as dependent beings, who are expected to subscribe to the technique of benign control and superior knowledge of the social workers in acquiring the survival strategies for coping with their predicaments (Dominelli, 1998, 2002, 2005; Johnson et al., 2004). Differences as such are objects for eradication through assimilation of the minorities into the dominant culture (Hugman, 1996).

In emphasizing the adjustment of clients to social realities, the maintenance-oriented view of social work and the movement toward clinical and psychotherapeutic models in social work practice depoliticalizes issues such as unemployment and poverty within society (Davis, 1994; Dominelli, 2002; Strier and Binyamin, 2010). The focus on unity and conformity to the structural order in this depoliticalization process reinforces the political function of social work in forging social and political stability. The recent inauguration of social work in the People's Republic of China

exemplifies the political role of social work, as the Chinese Government sets out to develop a strong social work force for 'constructing a harmonious socialist society' (Central Committee of the Communist Party, 2006). Promising to implant the techniques of responsible citizenship and supplant failing morals with morals of service and dedication under the tutelage of expertise, social work claims a sophisticated technology for governing at a distance, which makes it the right candidate to ease the social tensions that the Chinese Government can no longer manage with an authoritative and hierarchical governance system (Leung et al., 2012; Miller and Rose, 2008). The social harmony that the Chinese Government attempts to actualize is a state of stability in which unity under the 'socialist order' is emphasized. By socializing individuals to conform to preexisting structural order, sameness displaces diversity in this approach to achieve social harmony. Harmony as such is a state of unity achieved by silencing differences and sustaining control of the 'others' by existing power relations. Coined in this conformist sense, harmony can be a disguise for oppression.

In the critical tradition of social work, the 'othering' of difference is, however, received with much discomfort by critics who reject the use of binary opposites in defining social work outcomes. Proponents of antiracist and feminist social work underpin the unity–diversity dialectics in social work, acknowledging similarities and common causes among social work clients amid their considerable differences and uniqueness (Dominelli, 2008; Dominelli and McLeod, 1989; Phillipson, 1992; White, 2006). Practicing equality and social justice in critical social work practices 'requires the practitioner to value difference in lifestyles and identities' instead of settling for 'a uniform conformity in others' (Dominelli, 2002: p. 9). From a postmodern perspective, Fook (2002) calls for "a wealth of diverse meanings, experiences and identities" to be represented in social work discourse (Fook, 2002: p. 13). 'Difference' to be sought and embraced in the complexities, contradictions, and ambiguities underpinning everyday practices plays a key part in the ontology of an alternative paradigm for social work (Bell, 2012; Borden, 2010). Recognizing the damaging consequences in 'aspects of globalization that standardize and marginalize people,' Jones and Truell (2012) also embrace differences across regions as they consider a global agenda for social work, and support the diversity of interpretation of broader global standards for social work in regional contexts (Jones and Truell, 2012: p. 458). As the fluid, heterogeneous, and dynamic realities in practice are increasingly emphasized by proponents of progressive practices popularly known as emancipatory or antioppressive practice, contemporary social workers become more sensitive to differences within the population they serve (Bell, 2012; Dominelli, 2005; Strier and Binyamin, 2010).

Representing interactive spaces within which people construct their lives, community as an important platform of social work intervention is also fused with differences and diversities (Dominelli, 2007). Conventional conception of community as 'a unified constellation of people with shared goals' does not adequately capture plurality and contradictions that prevail in the community (Dominelli, 1999, 2005). At the same time, while collective interests of members in community create bonds of solidarity and symbolic identification to make it 'the warm glow of inclusion,' community can also be the 'cold icicles of exclusion' when differences from other communities separate insiders from outsiders (Dominelli, 1999: p. 442). The 'othering' processes that privilege insiders and foster the penetration of hegemonic discourses can also occur within the community or in the name of it (Dominelli, 2007).

The escalating call to embrace differences in social work practice encourages a broader view of politics, which challenges the normative boundary and the 'othering' processes that deny the human, social, and political rights of clients (Dominelli, 2002). In recognition of the inevitable power relationships between the social workers and their clients, achieving harmony among diversities requires social workers to reflect upon and examine how their 'biases, assumptions and cultural world-views' affect the ways they 'perceive difference and power dynamics' (Sakamoto and Pitner, 2005: p. 441). Critical consciousness and courage to challenge power differentials becomes a renewed professional ethic for social workers. Dominelli (2008) highlights in her practice guidelines for social work practice aiming to address diversities that social workers should 'think of the world as organized holistically with interdependent social relations' that acknowledge differences rather than submerge differences through 'binary dyads rooted in relations of superiority and inferiority'; and 'create new narratives' and 'form alliances' across social divides to 'ensure that differences are considered' and 'equality across differences' established (Dominelli, 2008: pp. 225–226). The social harmony for which an emancipatory approach to social work is advocating is a collaborative process for integrating 'different forces in a fluid yet dynamic world,' rather than a 'conformist harmony' that seeks to conform to a firm, preset and rational order imposed onto the world (Li, 2008: p. 91).

> If harmony is founded on a linear, logical, inherent structure of the world, we need to discover this reality and conform to it; if harmony is a multidimensional, inclusive, and dynamic creative process, we need to participate in a proactive way to generate it through collaboration. (Li, 2008: p. 95)

Such reflexive collaboration with clients realizes the radical notion of harmony, which Li (2008) describes as 'deep harmony.'

The Personal Arena of Harmony in Social Work Practice

Harmony is also a state of wellness within individuals – "when in harmony, people thrive, are resilient beyond expectation, and contributes in a synergistic manner to those around them with their energy" (Hodge et al., 2009: p. 216). Committed to enhancing human well-being, social work has always been playing an active role in maintaining harmony within the person. The bio-psycho-social construction of the self and human existence in conventional social work practice, however, are increasingly considered inadequate for promoting resilience and peace cherished in the ideal state of

intrapersonal harmony (Lee et al., 2009). Lee and colleagues criticize contemporary clinical social work practice as employing dualist thinking to conceptualize the person as one "who has distinctively separate and interrelated cognitive, behavioral, and affective domains, but not as a holistic and connected being" (Lee et al., 2009: p. xxvii). In the tendency to compartmentalize the person in terms of cognition, behavior, and affect, different therapeutic interventions assume the linearity of change enabled by specific therapeutic techniques suggested by respective theories – e.g., the emphasis on intrapsychic forces in psychoanalysis, thoughts in cognitive theories, and observable behaviors in behavioral theories (Chan et al., 2002; Lee et al., 2009; Leung et al., 2009). Informed by Eastern philosophies and postmodern thinking, alternative helping strategies are gaining momentum in the purview of clinical social work practices. These alternative approaches view the mind, the body, and the spirit as interconnected entities, and take a holistic and preventive view of a person's problem in the project of helping (Chan et al., 2002; Henderson, 2000; Lee et al., 2009; Leung et al., 2009; Weick, 1987).

The influence of Eastern philosophies is apparent in this holistic approach to achieve harmony within the person. Transcending the Western body–mind duality, the yin-yang perspective in Taoism sees personal wellness as a harmonious interplay between the body–mind–spirit components of a person (Chan et al., 2002; Lee et al., 2009; Leung et al., 2009). Believing that seemingly opposite forces are in fact complementary, the criterion for well-being is not the possession of any one attribute or characteristic, but the presence of a dynamic balance between diverse forces (Lee et al., 2009). In this integrative body–mind–spirit model of clinical social work practice, 'human existence is a manifestation of physical, emotional, cognitive, social, and spiritual being' interconnected into a whole, and wellness is therefore the 'harmonious dynamic equilibrium within the self' (Ng et al., 2008: p. 68). As clients going through their predicaments may experience disruption in one or more of the body–mind–spirit domains, a harmonizing intervention in social work is to reorganize and revitalize the persons for a new, fluid, and balanced state of total being (Chan et al., 2002; Lee et al., 2009; Leung et al., 2009; Ng et al., 2008). Considering spirituality, life values, and personal meaning as the foundation of the mind and body, the goal of social work intervention in this integrative body–mind–spirit approach is to encourage spiritual transformation in the clients (Lee et al., 2009). Translating these philosophical premises into practice principles, social workers are required to look at the whole and understand the dynamic balance within it, while appreciating change as a process of moving toward new balance (Leung et al., 2009).

The body–mind–spirit approach to clinical social work practice advocates for a harmonizing process to regain balance and harmony within the person (Lee et al., 2009; Ng et al., 2008). This harmonizing process involves four stages: (1) awareness of disharmony and imbalance of the system; (2) strengthening the system to restore balance; (3) harmonization of body and mind by facilitating recognition of the interconnectedness of body and mind; and (4) spiritual transformation to experience the course of life by redefining a goal-oriented pathway of life (Ng et al., 2008). Novel intervention techniques emerge from this holistic perspective in restoring wellness for the clients, including Tai Chi and Qigong exercises, body scan, breathing meditation, mindful tea drinking and eating, and cognitive reconstruction by acknowledging and accepting impermanence and unpredictability to enable clients to regain self-control by letting go of control (Lee et al., 2009; Ng et al., 2008). Although this alternative and holistic approach to clinical social work practice is largely derived from Eastern philosophies, a growing interest in a spiritual approach for balance and harmony within the person is not confined to the Chinese societies (Chan et al., 2002). In Henderson's survey of the use of alternative therapeutic methods among social work practitioners in America, mind–body techniques were provided to clients by most practitioners (Henderson, 2000). The spiritual side of clinical social work practice also receives increasing attention in the West, especially in cases of bereavement, terminal illness, and immigration where national communities are invoked (Canda and Furman, 2010; Seinfeld, 2012). A 'community of others' (Lax, 1992) comprising individuals who share similar experiences is important for supporting spiritual transformation in clients (Lee et al., 2009). Social work groups that harness collectives of support and mutual aid provide relevant medium for engendering the harmonious equilibrium within the self (Ng et al., 2008).

The Ecocentric Arena of Harmony in Social Work Practice

Harmony as balance, interrelationship, and complementarities also applies to the relationship of human beings with nature. Ancient Chinese philosophy supports the human ideal of achieving harmony not only within the human community but also with the rest of the cosmos, to allow 'myriad things to flourish' in a human–nature integration (Lee et al., 2009; Li, 2006). Addressing human's interaction with the ecological environment is, however, not a popular purview of social work practice. Although social work claims to espouse an ecological perspective and ground its practice in the person-in-environment assessment framework, 'the environment' in the social work literature refers primarily to the sociocultural or psychosocial environment, not the ecological environment (McKinnon, 2008; Norton, 2012; Zapf, 2005). The unexamined assumption of the superiority of human beings over the natural world, which allows human wants to dominate nonhuman entities, is a cause of concern to critics who lament the absence of ecological consciousness in social work practice (Besthorn, 2012; Coates, 2003; Dominelli, 2012; Gray and Coates, 2012; Norton, 2012). Conceiving ecological devastations as rooted in 'blind obsession with markets and profit, and materialistic and consumerist lifestyles' (Coates, 2003: p. 8), critics ascribe the cause of the planetary crisis to the 'competitive-individualistic-rational' belief system and 'modernism's innate attitude of superiority' (Coates, 2005: p. 33) that fails to appreciate our connection to other life forms and nature (Coates, 2003, 2005; Coates et al., 2006; Gray et al., 2008; Hart, 2002; Hodge et al., 2009; Zapf, 1999, 2005).

The role of social work in maintaining harmony between the natural universe and human beings has been attracting attention of the social work community lately. Environmental sustainability is included as a priority area for international effort in the global agenda for social work and social development promulgated jointly in 2012 by the IFSW, the IASSW, and the International Council on Social Welfare (IFSW, IAASW and ICSW, 2012). Some social work scholars appeal to expand the definition of client's life spaces and revise the configuration of the person-in-environment framework to include the natural world (McKinnon, 2008; Norton, 2012; Zapf, 2005). Critics are also calling for a new paradigm of social work practice that is premised on connectedness between human beings and nature, to facilitate a nondestructive and mutually beneficial human/nature relationship (Besthorn, 2012; Coates, 2003; Dominelli, 2012; Norton, 2012). This escalated concern for the relationship of the human community with nature is attributed to increasing awareness of the strong link between environmental problems and social problems, amid relentless industrial and technological progress and expansion of capitalism (Coates, 2003; Kemp, 2011; Norton, 2012). Directly or indirectly, the burdens of ecological devastations fall most heavily on poor, marginalized, and vulnerable populations concentrated in geographically undesirable and spatially isolated communities; and this legitimizes the ecosocial emphasis of social work in its quest for social justice (Gray and Coates, 2012; Kemp, 2011; Norton, 2012).

Dominelli (2012) describes social work with an ecocentric emphasis on holistic practices as focusing on the "interactions between socio-economic and physical environmental crises and interpersonal behaviors that undermine the well-being of human beings and Planet Earth" (Dominelli, 2012: p. 25). This highlights the political nature of what she terms 'green social work,' which addresses the sociopolitical and economic forces underlying the unequal distribution of power and natural resources. Dominelli (2012) translates this political and moral project of social work in the ecocentric arena into the following practice activities: (1) raising awareness of different disaster scenarios including poverty, and their outcomes; (2) maintaining dialogues about polices with policy makers, practitioners, and service users to encourage the sharing of resources and green technologies across the world; (3) lobbying for preventive measures locally, nationally, and internationally to create egalitarian and sustainable models of development; (4) mobilizing and developing resilient responses to disasters; (5) researching issues and problems to formulate new ways of looking at the world, and of living within it; and (6) training people to value and respect differences, and observe a duty of care toward the environment (Dominelli, 2012: pp. 201–2).

Apart from tackling ecological problems that affect human lives, proponents of ecocentric social work practice are also interested in the spiritual and philosophical insight obtained from the "deeper connection with, and more sensitive interplay, between humans and the non-human world" (Besthorn, 2012: p. 250). At the core of the philosophy of ecology is a 'person-with-environment' perspective, which calls for deep questioning of prevailing knowledge, values, and social institutions that maintain human dominance over nature, and recognizes the fundamentally spiritual nature of the ecological challenge (Besthorn, 2012; Coates, 2003). Underpinning social work's concern for harmony with Mother Nature is a moral and spiritual ethos that sets out to question the arrogance of seeing 'humankind as in control of and superior to other species' (Coates, 2003: p. 76). A spiritual change engendered by global consciousness is expected to promote deep respect for and harmony with nature. In practice terms, Coates (2003) advocates that social work can play an important role in helping people break free of a sense of superiority and separateness, and "recognize the need for a shift in consciousness away from the passion to consume, the centrality of economic growth and the drive for more" (Coates, 2003: p. 76).

Conclusion

As a philosophical premise, harmony is a state of balance, complementarities, and interdependence in the inevitable presence of differences. Manifested in the relational, personal, and ecological arena, harmony emphasizes acceptance of diversity rather than the 'othering' of differences, supports holistic integration rather than compartmentalization, and is achieved through cooperation rather than assimilation. Social work has long played a significant role in handling the unity–diversity dialectics in the relational arena. In mediating among diverse social groups with differing backgrounds and interests, intervention approaches emphasizing conformance and socialization of the minorities coexist with progressive practices that encourage critical consciousness of the social workers to embrace diversities. Essentially, social work contributes to social harmony by upholding and encouraging the value of acceptance in the face of difference. Confronting the 'othering' process by reflexive collaboration with clients also promises a state of 'deep harmony' that social work practice in a fluid yet dynamic world calls for. In the intrapersonal arena, the yin-yang perspective and its emphasis on the harmonious interplay between the body–mind–spirit components of a person encourages growing interest in an integrative and holistic approach to clinical social work practice. A 'social community of others' (Lax, 1992) comprising those who share similar experiences often plays an important role in facilitating the spiritual transformation required of harmony in the intrapersonal arena (Lee et al., 2009). While spiritual transformation is important for achieving the harmonious dynamic equilibrium within the self (Ng et al., 2008), it is also a necessary process to maintain harmony between human community and the natural world, amid modernist thinking that fuels human arrogance in rejecting peoples' embeddedness in nature, and blind obsession with materialistic and consumerist lifestyles (Coates, 2005). Maintaining harmony with the natural universe is also a political project, in which the sociopolitical and economic forces underlying the unequal distribution of power and natural resources are addressed (Dominelli, 2012). Committing to equilibrium and balance in the dynamic interplay among differences, the philosophical notion of harmony renders social work practice in various arenas a political and spiritual project.

See also: Antisocial Behavior; Ethics and Values; Green Social work; Risk and Uncertainty in the Social Sciences: Implications for Social Work Theory and Practice; Social Work in China; Spirituality.

Bibliography

Bell, Karen, 2012. Towards a post-conventional philosophical base for social work. British Journal of Social Work 42, 408–423.

Berlin, S.B., 2005. The value of acceptance in social work direct practice: a historical and contemporary view. Social Service Review 79 (3), 482–510.

Besthorn, Fred H., 2012. Deep ecology's contributions to social work: a ten-year retrospective. International Journal of Social Welfare 21, 248–259.

Biestek, Felix P., 1957. The Casework Relationship. Loyola University Press, Chicago.

Borden, William, 2010. Taking multiplicity seriously: pluralism, pragmatism and integrative perspectives in clinical social work. In: Borden, William (Ed.), Reshaping Theory in Contemporary Social Work: Toward a Critical Pluralism in Clinical Practice. Columbia University Press, New York, pp. 3–30.

Canda, Edward R., Furman, Leola D., 2010. Spiritual Diversity in Social Work Practice: The Heart of Helping, second ed. Oxford University Press, Oxford; New York.

Central Committee of the Communist Party, 2006. Important Decision about Building a Socialist Harmonious Society by the Central Committee of the Communist Party. Retrieved on June 14, 2011 at: http://news.xinhuanet.com/politics/2006-10/18/content_5218639.htm (In Chinese).

Chan, Cecilia, Ho, Petula Sik Ying, Chow, Esther, 2002. A body–mind–spirit model in health. Social Work in Health Care 34 (3–4), 261–282.

Coates, John, 2003. Ecology and Social Work: Toward a New Paradigm. Fernwood Press, Halifax.

Coates, John, 2005. The environmental crisis. Journal of Progressive Human Services 16 (1), 25–49.

Coates, John, Gray, Mel, Hetherington, Tiani, 2006. An 'ecospiritual' perspective: finally, a place for indigenous approaches. British Journal of Social Work 36, 381–399.

Davis, Martin, 1994. The Essential Social Worker: An Introduction to Professional Practice in the 1990s. Arena, Aldershot, Hants.

Dominelli, Lena, 1999. Community, citizenship and empowerment. Sociology 33 (2), 441–446.

Dominelli, Lena, 1998. The changing face of social work: globalisation, privatisation and the technocratisation of professional practice. In: Lesnik, Bogdan (Ed.), Change in Social Work. Ashgate, Aldershot, UK, pp. 13–26.

Dominelli, Lena, 2002. Anti-oppressive practice in context. In: Adams, Robert, Dominelli, Lena, Payne, Malcolm (Eds.), Social Work: Themes, Issues and Critical Debates, second ed. Palgrave, New York, pp. 3–19.

Dominelli, Lena, 2005. Community development across borders: avoiding dangerous practices in a globalizing world. International Social Work 48 (6), 702–713.

Dominelli, Lena (Ed.), 2007. Revitalising Communities in a Globalising World. Ashgate Publishing Limited, Aldershot; Burlington.

Dominelli, Lena, 2008. Anti-racist Social Work, third ed. Palgrave, New York.

Dominelli, Lena, 2012. Green Social Work: From Environment Crises to Environmental Justice. Polity, Cambridge.

Dominelli, Lena, McLeod, Eileen, 1989. Feminist Social Work. Macmillan Education, Basingstoke.

Fook, J., 2002. Social Work: Critical Theory and Practice. SAGE, London.

Freeman, Kathleen, 1983. Ancilla to the Pre-socratic Philosophers: A Complete Translation of the Fragments in Diels, Fragmente der Vorsokratiker. Harvard University Press, Cambridge, MA.

Gibelman, Margaret, 1999. The search for identity: defining social work – past, present, future. Social Work 44 (4), 298–310.

Gray, Mel, Coates, John, 2012. Environmental ethics for social work: social work's responsibility to the non-human world. International Journal of Social Welfare 21, 239–247.

Gray, Mel, Coates, John, Yellow Bird, Michael (Eds.), 2008. Indigenous Social Work around the World: Towards Culturally Relevant Education and Practice. Ashgate, Aldershot, England; Burlington, VT.

Guthrie, Wiliam K.C., 1962. A History of Greek Philosophy, vol. I. Cambridge University Press, Cambridge.

Hart, Michael, 2002. Seeking Mino-Pimatisiwin: An Aboriginal Approach to Helping. Fernwood Press, Halifax.

Henderson, Lorrie, 2000. The knowledge and use of alternative therapeutic techniques by social work practitioners: a descriptive study. Social Work in Health Care 30 (3), 55–71.

Hodge, David R., Limb, Gordon E., Cross, Terry L., 2009. Moving from colonization toward balance and harmony: a native American perspective on wellness. Social Work 54 (3), 211–219.

Hugman, Richard, 1996. Professionalization in social work: the challenge of diversity. International Social Work 39 (2), 131–147.

International Association of Schools of Social Work, 2004. Global Standards for the Education and Training of the Social Work Profession. Retrieved on June 11, 2013 at: http://cdn.ifsw.org/assets/ifsw_65044-3.pdf.

International Federation of Social Workers, 2012. Global Standards. Retrieved on January 15, 2013 at: http://ifsw.org/policies/global-standards/.

International Federation of Social Workers, International Association of Schools of Social Work, International Council on Social Welfare, 2012. The Global Agenda for Social Work and Social Development: Commitment to Action. Retrieved on January 26, 2013 at: http://cdn.ifsw.org/assets/globalagenda2012.pdf.

Johnson, Joy L., Bottorff, Joan L., Brown, Annette J., Grewal, Sukhdev, Hilton, B. Ann, Clarke, Heather, 2004. Othering and being othered in the context of health care services. Health Communication 16 (2), 255–271.

Jones, David N., Truell, Rory, 2012. The global agenda for social work and social development: a place to link together and be effective in a globalized world. International Social Work 55 (4), 454–472.

Jordan, Bill, 2001. Tough love: social work, social exclusion and the third way. British Journal of Social Work 31, 527–546.

Jose, Jim, 2010. Rethinking social work ethics: what is the real question? Responding to Stephen Webb's 'against difference and diversity in social work'. International Journal of Social Welfare 19, 246–252.

Kemp, Susan P., 2011. Recentring environment in social work practice: necessity, opportunity, challenge. British Journal of Social Work 41, 1198–1210.

Lax, William D., 1992. Postmodern thinking in a clinical practice. In: McNamee,-Sheila, Gergen, Kenneth J. (Eds.), Therapy as Social Construction. Sage Publications, London; Newbury Park.

Lee, Mo Yee, Ng, Siu Man, Leung, Pamela Pui Yu, Chan, Cecilia Lai Wan, 2009. Integrative Body–Mind–Spirit Social Work: An Empirically Based Approach to Assessment and Treatment. Oxford University Press, Oxford; New York.

Leung, Pamela Pui Yu, Chan, Cecilia Lai Wan, Ng, Siu Man, Lee, Mo Yee, 2009. Towards body–mind–spirit integration: East meets West in clinical social work practice. Clinical Social Work Journal 37, 303–311.

Leung, Terry Tse Fong, Yip, Ngai Ming, Huang, Ronggui, Wu, Ying, 2012. Governmentality and the politicization of social work in China. British Journal of Social Work 42 (6), 1039–1059.

Li, Chenyang, 2006. The confucian ideal of harmony. Philosophy East and West 56 (4), 583–603.

Li, Chenyang, 2008. The ideal of harmony in ancient Chinese and Greek philosophy. Dao 7, 81–98.

Marston, Greg, McDonald, Catherine, 2012. Getting beyond 'heroic agency' in conceptualizing social workers as policy actors in the twenty-first century. British Journal of Social Work 42, 1022–1038.

McKinnon, Jennifer, 2008. Exploring the nexus between social work and the environment. Australian Social Work 61 (3), 256–268.

Miller, Peter, Rose, Nikolas, 2008. Governing the Present: Administering Economic, Social and Personal Life. Polity Press, Cambridge.

Ng, Siu Man, Chan, Cecilia L.W., Leung, Pamela P.Y., Chan, Celia H.Y., Yau, Josephine K.Y., 2008. Beyond survivorship: achieving a harmonious dynamic equilibrium using a Chinese medicine framework in health and mental health. Social Work in Mental Health 7 (1–3), 62–81.

Norton, Christine L., 2012. Social work and the environment: an ecosocial approach. International Journal of Social Welfare 21, 299–308.

Phillipson, Julia, 1992. Practising Equality: Women, Men and Social Work. CCETSW, London.

Reisch, Michael, Jani, Jayshree S., 2012. The new politics of social work practice: understanding context to promote change. British Journal of Social Work 42, 1132–1150.

Reynolds, Bertha C., 1932. A changing psychology in social casework—after one year. Family 13 (4), 107–111.

Sakamoto, Izumi, Pitner, Ronald O., 2005. Use of critical consciousness in anti-oppressive social work practice: disentangling power dynamics at personal and structural levels. British Journal of Social Work 35, 435–452.

Seinfeld, Jeffrey, 2012. Spirituality in social work practice. Clinical Social Work Journal 40 (2), 240–244.

Stanley, Liz, Wise, Sue, 1993. Breaking Out Again: Feminist Ontology and Epistemology, second ed. Routledge, London and New York.

Strier, Roni, Binyamin, Sharon, 2010. Developing anti-oppressive services for the poor: a theoretical and organizational rationale. British Journal of Social Work 40, 1908–1926.

The Analects of Confucius, 2012 (A.C. Muller, Trans.). Retrieved on January 17, 2013 at: http://www.acmuller.net/con-dao/analects.html.

Wang, Ningchuan, Zou, Yuze, 2011. Yin-yang theory and globalization. Studies in Sociology of Science 2 (2), 38–49.

Webb, Stephen A., 2009. Against difference and diversity in social work: the case of human rights. International Journal of Social Welfare 18, 307–316.

Weick, Ann, 1987. Reconceptualizing the philosophical perspective of social work. Social Service Review 61 (2), 218–230.

White, Vicky, 2006. The State of Feminist Social Work. Routledge, London; New York.

Zapf, Michael Kim, 1999. Barriers to the acceptance of indigenous knowledge: lessons from social work in Canada's North. Rural Social Work 5, 4–12.

Zapf, Michael Kim, 2005. Spiritual dimension of person and environment: perspectives from social work and traditional knowledge. International Social Work 48 (5), 633–642.

Hart, Herbert Lionel Adolphus (1907–92)

Nicola Lacey, All Souls College, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Herbert Lionel Adolphus Hart (1907–92) revived the tradition of analytical jurisprudence, exploring both the concept of law in general and a number of specific concepts – causation, rights, obligation, responsibility, rules – that structure legal reasoning and doctrines. His theory of law lies in the positivist tradition of Bentham and John Austin, which asserts that law is a social artifact separate from morality. Hart's innovation was to draw on linguistic philosophy to elucidate the relevant concepts by examining their operation within the practical contexts in which they are used, and to theorize law as a system of primary and secondary rules with both internal and external aspects. Hart also contributed to moral and liberal political philosophy, making influential interventions in normative and policy debates about matters such as capital punishment, abortion, and the regulation of homosexuality. His treatment of these issues was located within a tradition stretching back to Hobbes, Hume, Locke, and, most obviously, J. S. Mill, to which Hart contributed, among other things, an elaboration of principles limiting the proper scope of state action, a novel account of the justification of punishment, and an analysis of the relationship between utility and rights.

Introduction

Herbert Lionel Adolphus Hart was born on 18 July 1907 in the spa town of Harrogate, England, where his family, which was Jewish, ran a furrier's business. Hart's was the first generation of the family to receive a university education: he studied at New College, Oxford, where he took an outstanding first class degree in Greats (ancient history and philosophy) in 1929. He then qualified as a barrister, and enjoyed a successful practice at the Chancery Bar throughout the 1930s. In 1940, he became a civil servant working in military intelligence; in 1941 he married Jenifer Fischer Williams, a civil servant in the Home Office, who was later herself to become an academic.

During the war, Hart's keen interest in philosophy was stimulated by regular conversations with Stuart Hampshire and Gilbert Ryle, who were working in a related department, and who introduced him to the new movements in linguistic philosophy. In 1945, after some hesitation, but encouraged by his close friend Isaiah Berlin, Hart accepted a fellowship in philosophy at New College. Despite his 16 years outside the academy, he established himself quickly as a leading member of the influential group of analytic philosophers, including Ryle, J. L. Austin, G. A. Paul, and Friedrich Waissman, working in Oxford at the time. Such was his reputation that, despite having published only three papers, he was in 1952 elected professor of Jurisprudence at University College, Oxford. During his tenure of the Chair, he revived what had become an intellectually sterile subject, bringing to jurisprudence the rigor of a philosophical method of which it had been deprived since the days of Bentham and John Austin, and introducing the fresh insights of Wittgenstein's and, primarily, J. L. Austin's philosophy of language.

Outside the United Kingdom, his reputation was especially strong in the United States, where he spent several periods, notably at Harvard in 1956–57, and in Israel, which he first visited in 1964. He was elected President of the Aristotelian Society in 1959 and Fellow of the British Academy in 1962. In 1968 he resigned from the Chair, taking up a senior research fellowship at University College, Oxford, in order to devote himself to the editing and analysis of the works of Bentham. From 1967–73 he served as a member of the Monopolies Commission; he also chaired a committee that made important recommendations for the reform of Oxford's governance in the light of student unrest in the late 1960s. From 1973 to 1978 he was Principal of Brasenose College; following his retirement he returned to University College as an honorary fellow. He died in Oxford on 19 December 1992.

Hart's Contribution to Legal, Moral, and Political Philosophy

Hart's intellectual contribution lies in two distinct but related fields. First, he revived the tradition of analytical jurisprudence, exploring both the concept of law in general and a number of specific concepts – causation, rights, obligation, responsibility, rules – that structure legal reasoning and doctrines. In this work, Hart emphasized J. L. Austin's assertion that one may use a "sharpened awareness of words to sharpen our perception of the phenomena" (Hart 1961: v). He insisted, however, that the task was not one of simple definition: rather, the aim was to build a theory or elucidation of the relevant concepts by examining their operation within the practical contexts in which they are used, and paying heed to the 'open texture' of language, with its core meaning surrounded by an unsettled penumbra (Hart, 1961; see also Hart, 1982, 1983; and for commentary, Bayles, 1992; Cane, 2010; Coleman, 2001; Gavison, 1987; Hacker et al., 1997; Kramer et al., 2008; Lacey, 2004; MacCormick, 2008; Martin, 1987; NYU Law Review, 2008; Simpson, 2011). Perhaps the most outstanding example of these methodological commitments lies in Hart and Honoré's monumental *Causation in the Law*, first published in 1959, which sets its conceptual analysis of causation within the context of hundreds of legal cases, and which draws criteria for the identification of causes out of the common sense understandings that underlie everyday usage in the relevant contexts Hart and Honoré (1959).

Hart's second contribution was in the fields of normative legal, moral, and political philosophy (Hart, 1955, 1963, 1965, 1968). Building out from his analytic conceptions, he made highly influential interventions in normative and policy debates about matters such as capital punishment, abortion, and the regulation of homosexuality. His treatment of these issues was located within a tradition stretching back to Hobbes, Hume, Locke, and, most obviously, J. S. Mill, to which Hart contributed, among other things, an elaboration of principles limiting the proper scope of state action, a novel account of the justification of punishment, and an analysis of the relationship between utility and rights.

To understand why Hart's work in analytical jurisprudence had such an extraordinary impact, one has to understand something about the development of jurisprudence in the century preceding his election to the Oxford Chair. From the late eighteenth to the midnineteenth century, the utilitarian thinkers Jeremy Bentham and his disciple John Austin had developed an influential theory of law known as 'legal positivism.' Rejecting the idea that law derives its authority from God, or from some metaphysical conception of nature or reason – so-called 'natural law' – Bentham and Austin argued that law is essentially man-made: it is a command, backed by sanctions, issued by a political superior or sovereign, to whom the populace is in a habit of obedience. This positivist theory had, however, been little developed since Austin's death. The one exception was the brilliant Austrian jurist Hans Kelsen, whose theory of law as a system of norms was, however, mainly influential on the continent of Europe. In early twentieth century Britain, after the decline of historical jurisprudence, the slow development of legal education in universities contributed to the impoverishment of the tradition of analytical jurisprudence, which concerned itself primarily with the decontextualized analysis of legal terms such as ownership, possession, or the corporation. This analysis was not enriched by any systematic exchange with philosophy, nor was there any attempt to link it to any broader idea of the nature of law, or to consider how technical legal concepts assisted law to serve its various social functions. In the United States, by contrast, jurisprudence had moved in a sociological direction that arguably emphasized the empirical at the expense of a sustained examination of the conceptual dimensions necessary to an analysis of law's distinctive normativity. Prescriptive questions about what purposes law ought to pursue were left to the attention of moral and political philosophy – the latter itself a field that was relatively stagnant at this time.

Hart's first single-authored book, *The Concept of Law* (1961), which stands as his most important work, claimed to provide a general, descriptive theory of law that was at once a contribution to 'analytical jurisprudence' and to 'descriptive sociology.' In other words, he sought to elucidate a concept of law that would be relevant to an understanding of all forms of law, wherever or whenever they arose. In pursuing this project, Hart returned to the insights of Austin and Bentham, but – in a crucial innovation – combined their methods with those of the new linguistic philosophy represented by J. L. Austin and Ludwig Wittgenstein.

The nub of Hart's theory is the startlingly simple idea that law is a system of social rules structurally similar to the rules of games such as chess or cricket. The rules are of different kinds, with complementary functions. Some – 'primary rules' such as criminal laws – directly govern behavior; others – 'secondary rules' – provide for the identification, interpretation, and alteration of the former. Distinctive among these secondary rules is the 'rule of recognition.' This rule provides criteria by which officials identify valid rules belonging to the legal system. It, therefore, also serves to identify the content of one particular legal system. Yet it is not, Hart argued, itself a valid rule of the system. Rather, it exists in the constellation of official attitudes and practices. The existence of a legal system is hence, fundamentally, a complex matter of social fact.

Like his nineteenth-century counterparts, Hart therefore insisted that law is a social, human invention: though legal rules generate genuine obligations, they are not moral rules. Their authority derives not from their content but from their source, which lies in a distinctively institutionalized system of social recognition. For example, the rule that we should drive on the left is authoritative not because there is any intrinsic value to driving on the left. Rather, it is because the rule can be identified in accordance with an agreed set of criteria for recognition, such as parliamentary enactment or judicial precedent. Precisely the same is true, moreover, of legal rules that overlap with moral standards: the legal prohibition on murder is not the same as, and derives its validity in a different way from, the moral injunction against killing. Hart also teased out, however, what he saw as important grains of truth in the natural law tradition. While he insisted that the connection between law and morality was a contingent one, he also pointed out that a rule-based legal order brings with it a distinctive conception of formal justice. He also argued that, given certain features of human beings – our physical vulnerability and limited altruism, and the limited resources of the natural world, for example – legal systems, given their purpose of facilitating human survival, would have to include a core set of rules governing – justly or unjustly – matters such as physical integrity and property. This he termed, perhaps misleadingly, "the minimum content of natural law" (Hart 1961: 189–95).

Hart's account of how legal rules are recognized as valid served, on the other hand, to distinguish law from a mere system of force, or 'orders backed by threats.' For – crucially – according to Hart, legal rules have not only an external but also an internal aspect: we know that a rule is in existence not only because it is regularly observed, but also because those subject to it use it as a reason or standard for behavior, criticizing themselves or others for breaches of the rules. Here the linguistic philosopher's focus on the understandings embedded in ordinary usage becomes important. For example, Hart explores the distinction between habitual behavior (going to a café on Sunday morning) and rule-governed behavior (going to church on Sunday morning); between being obliged to do something (handing over money because someone threatens to kill you if you refuse) and having an obligation to do it (paying your taxes). Of crucial importance to his status as a positivist is his insistence that the decision to adopt an 'internal attitude' to legal rules can be based on a variety of reasons, ranging from moral commitment to mere pragmatism: indeed, for a legal order to exist, it is theoretically only necessary for the officials of the

system to adopt an internal attitude to its rules. In his elucidation of the internal attitude, Hart claimed to have detached the concept of law from that of sanction: the normativity of law derives from the internal aspect of rules and not from the threat of punishment for their breach.

By moving from the early positivist notion of law as a sovereign command to the notion of law as a system of rules, Hart produced a theory that resonated with the impersonal idea of authority embedded in modern democracies, providing a remarkably widely applicable rationalization of the nature of law in a secular and pluralistic world. He also offered an account of legal validity that purported to explain the (limited) sense in which citizens have an obligation to obey the law. Unlike the 'command' theory, Hart's concept of law accounts not only for laws that impose duties but also the wide variety of laws that confer powers upon or create normative facilities for citizens: the law of property, contract, marriage, or succession, for example.

Hart also succeeded in reviving the idea that the project of jurisprudence should be not only to provide a clear, descriptive concept of law; but also – though as a separate project – to elaborate a framework for systematic criticism of law and for law reform. Hart's work became increasingly concerned with these prescriptive questions: from the late 1950s on he published highly influential papers dealing with issues such as the punishment of those who had committed what would now be regarded as serious crimes in Nazi Germany: capital punishment, the decriminalization of homosexuality, and the legalization of abortion. In these works, and in his widely publicized debate with the conservative judge Patrick Devlin, he regenerated a liberalism that finds its roots in J. S. Mills' *On Liberty* (1859). In *Law, Liberty and Morality*, Hart (1963) argued, as against Devlin, that democratic states are not entitled to enforce moral standards for their own sake. With the exception of special cases where paternalistic legislation can be justified, or where grave public offense is at issue, the state should respect individual freedom, intervening only to punish the commission of tangible harms. This books stands, over 30 years after its publication, as the resounding late twentieth-century statement of the most influential genre of liberal social policy. The philosophical ideas on which it rests are articulated closely with those expressed in Hart's conception of rights as founded in respect for individual human liberty and choice (Are there any natural rights?), and in his argument that consequentialist aims of punishment should be tempered by considerations of individual fairness in the distribution and quantum of punishment (Hart, 1968).

Hart's Importance for the Social Sciences

Since the mid-1960s, Hart's work has been subject to intense scrutiny and criticism. The most significant criticism has come from two different directions. First, from the direction of natural law theory, it has been argued that, ultimately, the reason why those subject to legal rules adopt an internal attitude to them has to do with moral factors, and that consequently Hart's restricted sense of legal obligation, and with it his distinction between law and morality, collapses. Second, from the direction of the social sciences, it has been argued that although the techniques of analytic philosophy may establish that law is simply one form of social rule, the further question of just what is distinctive about legal as opposed to other social rules can only be understood in terms of historical and social facts in which Hart shows surprisingly little interest. Notwithstanding these criticisms, there is strong reason to think that Hart's work has real significance for the social sciences.

First, in reviving the tradition of descriptive legal theory – the attempt to provide a systematic elucidation of law – Hart illuminated an institution whose distinctive importance in the organization of modern societies has become a fundamental preoccupation of the social sciences. In spite of his universalistic philosophical ambitions, and of the undoubted tension in his work between his analytic and his descriptively sociological methods, his theory can be argued to provide a persuasive account of the distinctive nature of legal authority – law's normativity – while maintaining that law's existence is, ultimately, a matter of social fact and hence susceptible to empirical inquiry.

Second, therefore, in elucidating the nature of law as a social institution, Hart arguably provides precisely the sort of legal theory which social scientists need in incorporating law into broader social analyzes (Sugarman, 2005). In his insistence on the structural continuity between legal and other social rules yet on law's distinctive mode of institutionalization, Hart provides a vision that is at once analytically well delineated yet relatively modest in terms of its substantive assumptions. Hence, the feature of his theory that might at first sight seem least sympathetic from a social science point of view – its pretension to provide a universal theory of law – may in fact be argued to be one of its strengths, for it provides a clear model for the purposes of testing hypotheses about, for example, the relationship between legal and other forms of social authority. This idea of his concept of law as, at root, a social science model is backed up by his own insistence, in terms reminiscent of Weber (though, in a significant ambiguity, also of Aristotle), that he is delineating a central case with reference to which other, penumbral cases can be identified and analyzed rather than banished to another discipline.

Third, in his insistence on the internal aspect of rules, and hence on the idea that the social phenomenon of law can only be understood by reference to the attitudes and understandings of the agents who are its subjects, Hart's thinking meshed with a cluster of theoretical developments in disciplines such as sociology and anthropology. In this respect, his theory is a vast step forward from the cruder empiricism of the nineteenth-century positivists. Although it is not something with which Hart himself was centrally concerned, this feature of his method undoubtedly opens up the possibility of an analysis of the symbolic aspects of legal power and authority such as would be central to any adequate social science account of legal practices. In his emphasis on the importance of linguistic usage, moreover, his ideas relate, albeit obliquely, to the so-called 'linguistic turn' in much social theory of the late twentieth century.

Finally, in his revival of the utilitarian tradition of systematic critique of law and argument for its reform, Hart provided an example of the way in which conceptual analysis can be marshaled in aid of social policy. Clear thinking and

reason – Hart's deepest intellectual commitments – are used, in an apparently modest and sometimes painstaking way, to clear the confusion in social and political debate, opening up in often unexpected ways paths forward in the development of policy and the construction of social consensus. Though Hart's own work is surprisingly little concerned with empirical – social, psychological, economic – evidence, his arguments often serve to clarify the relevance of just this sort of social science data. Though the fulcrum of his work lay firmly at the analytical-philosophic rather than the descriptive-sociological end of the spectrum, he can, therefore, justly claim a place among those whose thinking has made a significant contribution to the social and behavioral sciences.

Bibliography

Bayles, M.D., 1992. Hart's Legal Philosophy. Kluwer, Dordrecht, The Netherlands.

Cane, P. (Ed.), 2010. The Hart-Fuller Debate in the Twenty-first Century. Hart Publishing, Oxford, UK and Portland, Oregon.

Coleman, J. (Ed.), 2001. Hart's Postscript: Essays on the Postscript to the Concept of Law. Oxford University Press, New York and Oxford, UK.

Gavison, R. (Ed.), 1987. Issues in Contemporary Legal Philosophy: The Influence of H. L. A. Hart. Clarendon Press, Oxford, UK.

Hacker, P.M.S., Raz, J. (Eds.), 1997. Law, Morality and Society: Essays in Honour of H. L. A. Hart. Clarendon Press, Oxford, UK.

Hart, H.L.A., 1955. Are there any natural rights? The Philosophical Review 64, 175–191.

Hart, H.L.A., 1961. The Concept of Law (second ed., with postscript, 1994). Clarendon Press, Oxford, UK.

Hart, H.L.A., 1963. Law, Liberty and Morality. Stanford University Press, Stanford, CA.

Hart, H.L.A., 1965. The Morality of the Criminal Law. Oxford University Press, Oxford, UK.

Hart, H.L.A., 1968. Punishment and Responsibility (second ed., with introduction by John Gardner, 2008). Clarendon Press, Oxford, UK.

Hart, H.L.A., 1982. Essays on Bentham. Clarendon Press, Oxford, UK.

Hart, H.L.A., 1983. Essays in Jurisprudence and Philosophy. Clarendon Press, Oxford, UK.

Hart, H.L.A., Honoré, A.M., 1959. Causation in the Law (second ed., 1985). Clarendon Press, Oxford, UK.

Kramer, M., Grant, C., Colburn, B., Hatzistavrou, A. (Eds.), 2008. The Legacy of H.L.A. Hart: Legal, Political and Moral Philosophy. Oxford University Press, New York and Oxford, UK.

Lacey, N., 2004. A Life of H.L.A. Hart: The Nightmare and the Noble Dream. Oxford University Press, Oxford, UK.

MacCormick, N., 2008. H. L. A. Hart. Stanford University Press, Palo Alto, CA (2nd Ed.: 1st Ed. London, Edward Arnold 1981).

Martin, M., 1987. In: The Legal Philosophy of H. L. A. Hart. Temple University Press, Philadelphia, PA.

Mill, J.S., 1859/1974. On Liberty. Penguin, Harmondsworth, UK.

New York University Law Review. 2008. Symposium: The Hart-Fuller Debate at Fifty. vol. 83, Number 4.

Simpson, A.W. B., 2011. Reflections on the Concept of Law. Oxford University Press, Oxford, UK.

Sugarman, D., 2005 June. Hart interviewed: H.L.A. Hart in conversation with David Sugarman. Journal of Law and Society 32 (2), 267–293. http://www3.interscience.wiley.com/cgi-bin/fulltext/118651809/PDFSTARThttp://eprints.lancs.ac.uk/26217/.

Hate Crimes

Jack Levin and Jack McDevitt, Northeastern University, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E.M. Beck, volume 10, pp. 6479–6482, © 2001, Elsevier Ltd.

Abstract

The term 'hate crime' has been frequently applied not only by media and government representatives, but also by social scientists, to characterize attacks based on race and religion, but also on sexual orientation, national origin, disability status, and gender. Hate crime statutes vary from state to state. Since 2009, a federal law lends some consistency to the protected groups covered by hate crime laws. The majority of hate crimes are committed by dabblers – individuals who attack on a part-time basis and lack any affiliation with organized hate groups. Some commit hate offenses for thrill; others believe they are being defensive. Still others retaliate for an offense initially perpetrated by a member of another groups. Few are on a mission to rid the world of evil, as they see it. Hate crimes against particular groups tend to rise and fall in response to the presence of perceived threats.

Hate crimes are criminal offenses motivated at least in part by the fact or perception that a victim is different from the perpetrator. The term has been employed since the late 1980s, when it was first applied by journalists and politicians to characterize a racial incident in the Howard Beach section of New York City. In this incident, an African-American man was killed when he attempted to evade a violent mob of teenagers shouting racial epithets. By the early 1990s, the hate crime designation was being applied not only by media and government representatives, but also by social scientists, to characterize attacks based not only on race and religion, but also on sexual orientation, national origin, disability status, and gender (Herek and Berrill, 1992; Levin and McDevitt, 1993).

As used by the FBI and a number of other law enforcement agencies across the United States, the term hate crime involves a behavior that has already been defined as illegal in state or federal statutes. Thus, the majority of hate crime laws do not criminalize any new behavior; instead, they increase the penalty for behaviors that are already against the law. In addition, this definition specifies the motivation for committing the offense; it requires that a racial, religious, ethnic, or some other identified characteristic of the victim plays at least some role with respect to inspiring an attack.

Although widely used by the federal government, the media, and researchers in the field, the term 'hate crimes' is somewhat misleading because it suggests incorrectly that hatred is invariably a distinguishing characteristic of this type of offense (Perry, 2003). While it is true that many hate crimes involve intense animosity toward the victim, many others do not. Conversely, many crimes involving hatred between the offender and the victim are not hate crimes in the sense intended here. For example, an assault that arises out of a dispute between two coworkers who compete for a promotion might involve intense hatred, even though not based on any racial or religious differences between them. Similarly, a love triangle resulting in manslaughter may provoke intense emotion, but may have nothing at all to do with race or religion. After decades, however, the term hate crime has become a widely recognized and accepted part of our vernacular as well as the social scientific literature and legislation across the country. Notwithstanding its problematic aspects, those who study the phenomenon have continued to use the term as a basis for what appears to be an ever-expanding literature. In the early 1990s, for example, only a book or two were entitled 'Hate Crimes' (including a work written by the authors of this article); by the turn of the century, there were at least 30 scholarly volumes bearing this title and many more on the way.

Legal Distinctions

In legal terms, the groups protected by hate crime laws have traditionally differed from state to state. Almost every state – at least 45 out of 50 – protects victims who are targeted because of their race, ethnicity, or religion, but far fewer cover sexual orientation (25 out of 50), disability status (22 out of 50), or gender (21 out of 50). The implication of this lack of uniformity is that members of a particular group may be protected by a hate crime statute in one community but not protected in a neighboring community in an adjacent state (Smith and Foley, 2010).

One of the early pieces of federal legislation regarding hate crimes, the Hate Crime Sentencing Enhancement Act, was passed in 1994. This statute increased sentencing penalties in federal cases with proof of victim targeting based on race, color, religion, national origin, ethnicity, gender, disability, or sexual orientation (Cogan, 2002; Levin, 2002). Also in 1994, the Violence Against Women Act was passed allowing for punishment and compensation in relation to federal crimes motivated by gender and promoting the inclusion of gender in the definition of hate crimes in a number of states (Grattet and Jenness, 2001).

In 2009, a greater degree of consistency among the 50 states was achieved by the expansion of a federal hate crimes statute known as the Matthew Shepard and James Byrd, Jr. Hate Crime Prevention Act, which supported federal prosecution of offenses based not only on race, religion, and national origin

but also on sexual orientation, gender, gender identity, and disability status. Based on the 2009 federal law, even where a particular category is not protected at the state level, federal officials can now move to support the prosecution of a hate crime charge. It also makes grants available to communities in order to train law enforcement officers or assist in their investigations and prosecutions of bias-motivated offenses.

A second area of legal distinction involves the penalty structure of the statute. In some states, a separate statute exists that prohibits hate crime behavior. In other states the hate crime statute is a 'penalty enhancement.' This means that if an existing crime is committed and it is motivated by bias, the penalty on the existing crime may be increased. Penalty enhancements have been applied not only to hate crimes, but to other offenses as well. For example, they have been enacted in connection with crimes committed with a gun, crimes committed by individuals with long criminal histories, and crimes committed against vulnerable victims such as children.

A number of critics of hate crime legislation have suggested that the kinds of legislation described above are an unconstitutional infringement on our freedom of speech. The US Supreme Court has been careful to protect free speech; but in 1993 (Wisconsin vs Mitchell), the Court decided that statutes that enhance the criminal penalty due to the bias motivation of the offender were constitutional.

Do Hate Crimes Really Differ from Other Offenses?

The question is frequently posed as to why we need additional penalties in response to hate offenses. After all, hate crimes by definition involve behavior – for example, assaults, threats, and vandalism – that is already prohibited by state or federal statute. It is therefore important to ask: Are hate crimes truly different?

We believe that a number of characteristics of hate incidents make them different from other types of offenses. First, hate crimes are directed symbolically at large groups of people, not only at a single individual. If youths decide that they do not want Asian-Americans living on their block, they may decide to throw a rock through the window of a home owned by a new neighbor who is of Asian descent. Their intention is to send a message not just to that neighbor but to all Asians, informing them that their presence in the neighborhood will not be tolerated.

Thus, not unlike acts of terrorism, hate crimes are frequently about transmitting messages. Offenders use a criminal event to put the members of an entire group on notice, by example, that they are not welcome in a community, in a workplace, on a college campus, or at school. By contrast, if a window is broken in a simple act of vandalism, the offenders typically have no desire to communicate anything in particular to the property owner; in general, they are totally unfamiliar with the victim they have targeted.

Another characteristic that differentiates hate crimes from most other offenses is that the victim characteristic motivating the attack (e.g., race or ethnicity) is in most cases ascribed and immutable. A person cannot simply choose to modify her or his race, ethnicity, age, or disability status. Even a religious identity or a sexual orientation cannot be changed without causing an individual to make dramatic and painful modifications in lifestyle. Consequently, if a woman is attacked because she is a Latina, there is little that individual can do to become 'de-Latinized' and thus reduce the likelihood of her future victimization. This is also true of perceived characteristics. If a man becomes a hate crime victim because he is perceived by a group of youths to be gay, he is also powerless to change the offenders' perception of him. The feeling on the part of victims that they lack control over the characteristic that motivated their victimization causes most hate crime victims to feel extremely vulnerable to future bias-motivated attacks.

A third characteristic of hate crimes that makes them different from many other offenses is that the individual victim typically did nothing in particular to precipitate the attack and is therefore interchangeable, at least from the perpetrator's standpoint. To a group of youths waiting outside a gay bar to attack someone whom they believe might be gay, it does not matter which individual is next to pass through the door. Whoever comes out is likely to become a victim, because all bar patrons are identical in the mind of the perpetrator.

Indeed, the interchangeability of victims tends to apply as well across various groups of victims. If offenders cannot locate the members of one racial group to terrorize, they are likely to target members of another racial group. This aspect of hate crimes suggests that they are often motivated by an offender's psychological need to feel superiority at the expense of his victims. Particular characteristics of victims are apparently less important conceptually than the motivation of offenders who, out of a profound sense of powerlessness, decide to attack victims who have been stereotypically defined in the dominant culture as inferior or evil.

A fourth characteristic that separates hate crimes from many other offenses involves the severity of their outcome. Research suggests that victims of hate crimes are more likely to experience intrusive thoughts, depression, anxiety, and Posttraumatic Stress Disorder (PTSD) when compared to victims of nonbias crimes (McDevitt et al., 2001).

Problems Collecting Hate Crime Data

During the 1980s researchers, journalists, and government officials increasingly turned their attention to questions regarding the extent and nature of hate crimes in the United States. At that time there were no national data gathered on the incidence and character of such offenses. The FBI, the agency responsible for collecting and reporting national data on crime, did not collect information separately for offenses motivated by bias. For example, a hate-motivated assault reported by the local police to the FBI would be grouped together with other assaults, regardless of their motivation. As a result, it was impossible to distinguish which offenses were hate-motivated and which were not.

In April 1990, Congress passed and the President signed the Hate Crime Statistics Act (28 USC 534). Although not criminalizing any particular behavior, the Act required the Attorney General of the United States to publish an annual report about crimes that 'manifest evidence of bias based upon race, religion, sexual orientation, or ethnicity.' This act authorized the first national data collection effort undertaken by any public

agency to be targeted specifically at hate crimes. The Uniform Crime Reporting (UCR) Section of the FBI was designated as the federal agency empowered to collect and tabulate these data. In September 1994, the Hate Crime Statistics Act was amended to include bias against persons with disabilities by passing the Violent Crime Control and Law Enforcement Act of 1994. The FBI began collecting data for the additional bias type on 1 January 1997.

Although mandating the Attorney General with the responsibility for gathering information and publishing an annual report about hate crimes, the Hate Crime Statistics Act did not require that local law enforcement agencies report to the FBI. Historically, crime reporting has been voluntary with most but not all major law enforcement agencies agreeing to participate.

Hate crime reporting was initially a different story. Many agencies had not separated hate crime data from other offenses in their collection procedures. As a result, retrieving these data was not an easy task. In addition, many law enforcement agencies resisted the new national emphasis on hate crime investigation. For these reasons, few agencies participated in the original effort. In 1991, 2771 agencies submitted hate crime data to the FBI, which represented only 20% of the agencies participating in the UCR program. As a result of significant training efforts of the UCR Section of the FBI and effective support from local advocacy groups, however, the participation rate by 2011 had increased to 14 575 agencies representing approximately 80% of all agencies submitting information to the FBI.

These improved participation figures may not, however, be as positive as they appear. According to the Department of Justice, two out of three hate crimes go unreported. Estimates of the prevalence of hate crimes vary substantially from 7000 to 8000 annually, as reported by police jurisdictions to the FBI, to some 200 000 yearly, as reported anonymously by victims to the Department of Justice. Part of the reason for such a wide disparity involves the lack of responsiveness to hate crime victims by numerous police agencies, where hate attacks are simply unreported. Moreover, relatively few hate-motivated offenses are ever prosecuted. Even fewer result in a conviction. One of the problems moving from arrest to conviction involves the large number of cases that are dismissed based on the lack of cooperation of victims and/or eyewitnesses, or cases where the prosecutors refuse to indict because they feel they do not have adequate evidence of bias motivation.

The FBI program allows agencies to submit zero as their total count of hate crimes. According to a recent FBI report, nearly 87% of participating agencies report no hate crimes occurring in their jurisdiction during the previous 12 months. While reporting zero hate crimes may accurately reflect the number of hate crimes in many jurisdictions, scholars suggest that some agencies, particularly in larger and more diverse communities, are not fully and accurately collecting information on and reporting hate crimes (McDevitt et al., 2003). In 2011, for example, three states reported fewer than six hate crimes: Mississippi (1), Wyoming (2), and Louisiana (5). Similarly, a number of major cities reported surprisingly small numbers of hate offenses, such as Detroit, Michigan (7); Cleveland, Ohio (4); Atlanta, Georgia (2); St. Louis, Missouri (7); and Houston, Texas (13). Over the same period, Boston reported some 206 hate crimes (FBI, 2012).

Also contributing to what appears to be a vast underestimate in FBI statistics is the unwillingness of victims to make official reports. They may see law enforcement as 'the enemy of occupation,' based on traditional antagonisms between the police and people of color or gay and lesbian residents. Or, if they have migrated from certain other countries, they may consider law enforcement as associated with a tyrannical regime. People with psychiatric or intellectual disabilities may not totally understand the process of reporting hate-motivated attacks and may instead suffer in silence.

Another problem involves indicators for identifying the motivation of an offender. In many cases, the most obvious measures consist of slurs, epithets, or graffiti left by the perpetrator at the crime scene. In addition, police investigators may look for a suspect's criminal record, a location where previous hate attacks have occurred, and evidence that the suspect holds membership in an organized hate group. Obviously, not all offenders leave evidence of their motivation; many are first offenders who lack a criminal record and are dabblers in hate. In many cases, the use of slurs is the only indicator available, causing a court battle of 'he said, she said' between victims and offenders.

Types of Hate Crimes

In terms of offender motivation, hate crimes have been categorized into four major types (Levin and McDevitt, 2002; McDevitt et al., 2002): (1) *thrill*, which are almost recreational attacks committed by youngsters – usually groups of teenagers or young adults – who seek excitement as well as 'bragging rights' with their friends; (2) *defensive*, which are designed to protect an individual's neighborhood, workplace, school, or women from those who are considered to be outsiders or intruders; (3) *retaliatory*, which are motivated by an individual's need for revenge as a result of a hate attack directed against his or her own group members; and (4) *mission*, which are rare and usually committed by the members of an organized hate group or an offender who suffers from a severe mental illness (e.g., psychosis). This typology has become part of the National Hate Crime Training Curriculum and is taught in the FBI Training Academy in Quantico, Virginia.

The majority of hate crimes are committed by dabblers – individuals who attack on a part-time basis and lack any affiliation with organized hate groups. They might go out with their friends on a Saturday night looking for a vulnerable victim to bash. Or, they might send a threatening message to a new resident of their neighborhood, informing them that 'their kind' is not welcome. Actually, no more than 5% of all hate crimes nationally are committed by the members of organizations like the Ku Klux Klan, Aryan Nations, or the Hammerhead Skins. Yet through their presence on the Internet, organized hate groups continue behind the scenes to support much larger numbers of violent offenses committed by nonmembers who may be unsophisticated with respect to the ideology of hate, but are looking to feel important and a sense of belonging. Not only do such groups provide angry and isolated youngsters with effective propaganda but they

also generate chat rooms and discussion boards, where recruits develop newfound friendships with like-minded hatemongers.

A brutal example of the influence of hate groups on the Internet involves the January 2009 murder of two Cape Verdean individuals and the rape of a third in Brockton, Massachusetts, by 22-year-old Keith Luke who was not a member of any organized hate group but was clearly looking for a scapegoat on whom he could blame all of his personal miseries. Unemployed, friendless, and suffering from chronic mental illness, Luke spent day after day logging into hate Web sites, where he found racist propaganda and numerous opportunities to chat with fellow hatemongers. A few hours after President Obama's first inauguration speech, Luke decided that he had had enough from non-white Americans and went on his bloody rampage. His intended victims also included a nearby synagogue, where he had planned to open fire and kill as many Jews as possible. Instead, he was taken into custody and eventually stood trial. On 30 May 2013, Luke was convicted of first-degree murder and sentenced to life without parole eligibility.

Are Hate Crimes on the Rise?

Though limited to those crimes that are reported to the police, national hate crime statistics provide a critical measure of the prevalence and distribution of hate crimes throughout the country and over time. Between 2000 and 2011, with the exception of a spurt to almost 10 000 hate incidents in the year 2001, the FBI reports the total number of hate-motivated crimes reflected in the national statistics remained relatively constant, ranging from slightly more than 6000 in 2011 to almost 8000 in 2008 (**Figure 1**).

Rather than yield a strong trend over time, hate crimes against particular groups tend to rise and fall in response to the presence of perceived threats (Quillian, 1995; Levin, 2014). In Green's (2001) defended neighborhood hypothesis, hate crimes occur as a defense against neighborhood in-migration of ethnic or racial 'intruders.' This hypothesis can be expanded to include perceived threats to a group's self-esteem, economic well-being, religious or cultural values, sexual identity, or physical survival. In 2001, for example, hate-motivated assaults against Muslim/Arab Americans soared by some 1600%, as many Americans became anxious about their safety from the impact of international terrorism.

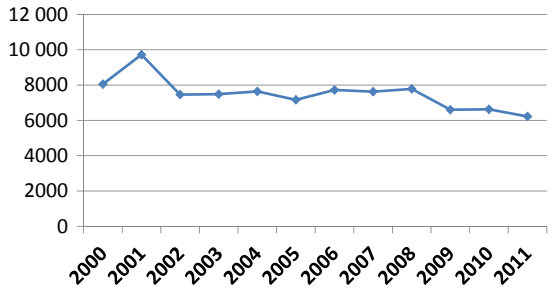

Figure 1 Hate incidents by year, 2000–2011.

Responses to Hate Crimes

Interviews with victims of hate violence indicate that the aftermath of the victimization is characterized by a pervasive feeling of fear. As indicated earlier, the victims of these incidents generally did nothing to bring this violence upon them and thus do not know what to do to reduce their chance of future victimization. Their fear may be based on threats by the offender or friends of the offender but often it is simply based on the random nature of the crime (McDevitt et al., 2001). In addition, the fear is felt by many others who may share characteristics with the victim.

In order to reduce fear in victims of hate crimes, it is important to offer them some form of protection from future violence and a degree of reassurance that they are valued members of the community. First of all, community leaders must speak out, condemning the attack. This is important because it sends two essential messages: to the victims, that local residents want them to remain members of that community and, to the offenders, that most people in the community do not support their illegal behavior. This speaking out should be done on two levels: one, a public event such as a vigil; and second, individual conversations with the victim and the victim's family where questions can be answered and personal assurances be conveyed.

Interviews with hate crime offenders indicate that they frequently fear that most of the community shares their desire to eliminate the 'outsider.' The offenders often see themselves as heroes or at least as 'cool' in the eyes of their friends, because they have the courage to act on what they believe to be commonly held beliefs. Public statements by local community leaders challenge this idea and send a message to offenders that their actions are not supported or encouraged.

Many hate crimes are perpetrated by young people who do not yet have a profound commitment to bigotry and therefore may be dissuaded from repeating their offense. It is important, therefore, to apprehend youthful hate crime perpetrators at this point, especially in light of the possibility that many property offenders who go undetected may later graduate to more violent hate crimes directed against people. Because what the perpetrators derive from committing such crimes is so minimal in a practical sense, they may be very deterred by a strong statement from society at large that demonstrates that this type of behavior will not be tolerated and a fear that they could be arrested and brought into the justice system. There is evidence that the police and the courts (not to mention public opinion) may have become much tougher on youthful hate crime offenders during the past several years.

The local police play an essential role in responding to hate crimes. Advocacy groups can offer support and encouragement, and political leaders can offer reassurance, but the police represent the only group that can legitimately promise to protect the victim from any future attack. Since so many victims are extremely fearful of future attacks, this offer of protection is so important. Before law enforcement personnel can effectively offer this protection to victims of hate crimes, however, they must be trained to identify and investigate these difficult cases. The UCR Section of the FBI has developed an outstanding training program for local law enforcement that teaches officers how to identify and effectively prosecute hate crimes.

In order for police to effectively support hate crime victims, as indicated above, it is helpful if they have training in what hate crimes are and how victims generally respond. Additionally, for many victim groups (e.g., Muslim or Middle Eastern victims or members of the Lesbian, Gay, Bi-sexual, Transgender, Questioning (LGBTQ) community) it is very important that police make outreach efforts before a hate crime occurs and begin to develop relationships in the community so that victims will feel comfortable coming forward to the police and asking for help. Hate crime victims are often ambivalent about the value of law enforcement and do not necessarily see a police officer as someone they can go to for help.

Finally, the most important response to a hate crime, as reported by its victims, is the reaction of those closest to them – their neighbors, coworkers, or fellow students. When a hate crime occurs, victims quite realistically wonder just how widespread is the hatred directed toward them. Do all their neighbors agree with the person who attacked them? Do all of their classmates want them to leave school? Therefore, the most significant reaction for most victims is when members of the perpetrator's group come forward to assure the victim that they do not agree with the offenders and to urge the victims to remain in the neighborhood, the workplace, or the campus.

See also: Afrocentric Approaches to Social Work; Ethnic Cleansing, History of; Heterosexism and Homophobia; Immigration and Crime in the United States; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Prejudice in Society: Psychological Perspectives; Prejudice in Society: Sociological Perspectives; Religiosity and Prejudice against Minorities; Terrorism; Violence, Criminalization, and the Homeless; Wisdom, Psychology of.

Bibliography

Anti-Defamation League, 1988. "JAP Baiting": When Sexism and Anti-semitism Meet. Special Education of ADL Periodic Update. ADL, New York.
Anti-Defamation League, 1995. The Skinhead International. ADL, New York.
Anti-Defamation League, 1997. Vigilante Justice: Militias and Common Law Courts Wage War against the Government. ADL, New York.
Berrill, K., 1992. Anti-gay violence and victimization in the United States: an overview. In: Herek, G., Berrill, K. (Eds.), Hate Crimes: Confronting Violence against Lesbians and Gay Men. Sage Publications, Newbury Park, CA.
Bureau of Justice Assistance, 1997. A Policy Makers Guide to Hate Crimes. U.S. Government Printing Office, Washington, DC.
Cogan, J.C., 2002. Hate crime as a crime category worthy of attention. American Behavioral Scientist 46 (1), 173–185.
Delgado, L., 31 July 1992. Black Family in North End Is Met with Racist Graffiti. Globe, Boston, MA, p. 17.
Ehrlich, H., 1972. The Social Psychology of Prejudice. Wiley, New York.
Ehrlich, H., 1990. Ethno-violence on College Campuses. National Institute Against Prejudice and Violence, Baltimore, MD.
Federal Bureau of Investigation, 2011. Hate Crime in the United States. Washington, DC.
Federal Bureau of Investigation, 2012. Hate Crimes in the United States. Washington, DC.
Federal Office for the Protection of the Constitution, 2000. Annual Report of the Office for Protection of the Constitution. Washington, DC.
Fox, J., Levin, J., 1996. Overkill: Mass Murder and Serial Killing Exposed. Dell, New York.
Freeman, S., 1993. Hate crime laws: punishment which fits the crime. Annual Survey of American Law, 581–585. New York University School of Law, New York.
Gambino, R., 1977. Vendetta. Doubleday, New York.
Gordon, D., Pardo, N., September 1997. Hate Crimes Strike Changing Suburbs. Reporter, University of Chicago, Chicago, IL, p. 1.
Grattet, R., Jenness, V., 2001. The birth and maturation of hate crime policy in the United States. American Behavioral Scientist 45 (4), 668–696.
Green, D.P., McFalls, Laurence H., Lawrence, H., Smith, Jennifer K., 2001. Hate crime. Annual Review of Sociology 27, 479–504. Reprinted in Perry, Barbara. 2003. Hate and Bias Crime: A Reader. Routledge, New York.
Green, D.P., Strolovitch, D.Z., Wong, J.S. Defended Neighborhoods, Integration, and Hate Crime. Institution for Social and Policy Studies, Yale University, Unpublished manuscript.
Halpern, T., Levin, B., 1996. The Limits of Dissent: The Constitutional Status of Armed Civilian Militias. Aletheia Press, Amherst, MA.
Hamm, M.S., 1994. Hate Crime: International Perspectives on Causes and Control. Anderson, Cincinnati, OH.
Herek, G.M., Berrill, K.T. (Eds.), 1992. Hate Crimes: Confronting Violence against Lesbians and Gay Men. Sage Publications, Newbury Park, CA.
June 6, 1997. Gunshots Aimed at Home of Black Family. Associated Press.
Jacobs, J.B., Henry, J.S., 1996, Winter. The social construction of a hate crime epidemic. The Journal of Criminal Law and Criminology 82, 366–391.
Jacobs, J.,B., Potter, K.A., 1997. In: Tonry, M. (Ed.), Hate crimes: a critical perspective. Crime and Justice: A Review of Research. University of Chicago Press, Chicago, IL, pp. 1–50.
Jenness, V., Broad, K., 1997. Hate Crimes: New Social Movements and the Politics of Violence. Aldine De Gruyter, New York.
Karl, J., 1995. The Right to Hear Arms. Harper, New York.
Keen, S., 1988. Faces of the Enemy. Harper and Row, New York.
Keough, W., 1990. Punchlines. Paragon House, New York.
Klanwatch Intelligence Report, 1997, Spring. Two Years after: The Patriot Movement since Oklahoma City, 18–20.
Lamy, P., 1996. Millennium Rage. Plenum Press, New York.
Langer, E., 16/23 July 1990. The American neo-nazi movement today. The Nation, 82–107.
Levin, B., 1992–1993, Winter. Bias crimes: a theoretical and practical overview. Stanford Law and Policy Review, 165–171.
Levin, B., 2002. From slavery to hate crime laws: the emergence of race and status-based protection in American criminal law. Journal of Social Issues 58 (2), 227–245.
Levin, J., 1 March 1997. Visit to a Patriot Potluck. USA Today, 6A.
Levin, J., 12 December 1996. Teenagers Burn Churches. Globe, Focus Section, Boston, MA.
Levin, J., 13 July 1997. N. Irish Racialize "The Troubles." Herald, Boston, MA.
Levin, J., 2014. From thrill to defensive hate crimes: the impact of 9/11. Hate Studies, in press.
Levin, J., Levin, W.C., 1982. The Functions of Discrimination and Prejudice. Harper and Row, New York.
Levin, J., McDevitt, J., 1993. Hate Crimes: The Rising Tide of Bigotry and Bloodshed. Plenum, New York.
Levin, J., McDevitt, J., 4 August 1995a. The research needed to understand hate crime. Chronicle of Higher Education, BI–2.
Levin, J., McDevitt, J., August, 1995b. Landmark study reveals hate crimes vary significantly by offender motivation. Klanwatch Intelligence Report, 7–9.
Levin, J., McDevitt, J., 2002. Hate Crimes Revisited: America's War on Those Who Are Different. Westview Press, Boulder, CO.
Martin, S., 1995. A cross-burning is not just an arson: police social construction of hate crimes in Baltimore county. Criminology 33 (3), 303–330.
McDevitt, J., Balboni, J., Garcia, L., Gu, J., 2001. Consequences for victims: a comparison of bias- and non-bias-motivated assaults. American Behavioral Scientist 45 (4), 697–713.
McDevitt, J., Balboni, J., Bennett, S., Weiss, J., Orchowsky, S., Walbolt, L., 2003. Improving the quality and accuracy of bias crime statistics nationally. In: Hate and Bias Crime: A Reader. Routledge Press, New York, pp. 77–89.
McDevitt, J., Levin, J., Bennett, S., 2002. Hate crime offenders an expanded typology. Journal of Social Issues 58 (2), 303–317.
Not in Our Town II, 1996. Public Broadcasting System. Office of Juvenile justice and Delinquency Prevention, 1996. Report to Congress on Juvenile Hate Crime. US Government Printing Office, Washington, DC.
Olzak, S., Shanahan, S., McEneaney, E.H., August 1996. Poverty, segregation, and race riots: 1960 to 1993. American Sociological Review, 590–613.
Paulsen, M., 6 July 1996. Fires come in different shades of hate. The State, 1A.
Pearce, D.M., February 1979. Gatekeepers and homeseekers: institutional patterns in racial steering. Social Problems 26 (3), 325–342.

Perry, B., 2003. Hate and Bias Crime a Reader. Routledge Press, New York, NY.
Potok, M., 2003, Fall. Against the wall. SPLC Intelligence Report, 12.
Quillian, L., 1995. Prejudice as a response to perceived group threat: population composition and anti-immigrant and racial prejudice in Europe. American Sociological Review 60 (4), 586–611.
Schevitz, T., 26 November 2002. FBI Sees Leap in Anti-Muslim Hate Crimes. San Francisco Chronicle, p. 1.
Schuman, H., Steeh, C., Lawrence, B.L., 1988. Racial Attitudes in America. Harvard University Press, Cambridge, MA.
Sherif, M., Sherif, C., 1961. Intergroup Conflict and Cooperation: The Robbers Cave Experiment. University of Oklahoma, Norman, OK.
Smith, A.M., Foley, C.L., 28 September 2010. State Statutes Governing Hate Crimes. Congressional Research Service. http://www.fas.org/sgp/crs/misc/RL33099.pd.
Stossen, N., May 1993. Special penalties should apply to hate crimes. American Bar Association Journal 44.
Watts, M., 1997. Xenophobia in United Germany. St. Martin's Press, New York.
Weisburd, S.B., Levin, B., 1994, Spring. "On the basis of sex": recognizing gender-based bias crimes. Stanford Law and Policy Review 5 (2), 21–43.
Weiss, J., Ehrlich, H., Barbara, L.B., 1991/1992, Winter. Ethnoviolence at work. Journal of Intergroup Relations, 21–33.

Hazards and Disaster Research in Contemporary Anthropology

Anthony Oliver-Smith, University of Florida, Gainesville, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Disasters are highly complex, multidimensional processes and events that take place in both the material and social worlds. It is in this multidimensionality of disasters that the salience of anthropological theory and application manifests itself most clearly. The holistic perspective of anthropology is ideally suited to address the multifaceted realities of disasters. As disasters develop and occur, all dimensions of a social structural formation and the totality of its relations with the total environment become involved, affected, and focused in patterns of vulnerability and resilience over time. In disasters are expressed continuity and change, cooperation and conflict, power and resistance, all articulated through the operation of physical, biological, and social systems and their interaction among populations, institutions, and practices. Mining this rich vein of thought and behavior that disasters bring into intense focus, presents not only a challenge but an important opportunity to probe the multiple linkages and relationships they reveal for sociocultural theory and, by the same token, establish what anthropology can offer to the study of disaster both in the realm of research as well as the practice of disaster management.

Introduction

Disasters, in their origins and outcomes, are highly complex, multidimensional processes and events that take place in both the material and social worlds. In that sense, disaster research is inherently multidisciplinary and anthropology shares many perspectives with other social sciences, in particular geography and sociology. However, the holistic approach of anthropology is ideally suited to address the multifaceted realities of disasters. From this perspective, a disaster is a process/event involving the combination of a potentially destructive agent(s) from the natural, modified, and/or constructed environment and a population in a socially and economically produced condition of vulnerability, resulting in a perceived disruption of the customary relative satisfactions of individual and social needs for physical survival, social order, and cultural coherence.

A disaster is a collectivity of intersecting processes and events, environmental, social, cultural, political, economic, physical, technological, transpiring over varying lengths of time. Disasters vary from sudden to rapid onset events such as earthquakes, tornadoes, hurricanes, tsunamis, floods, technological accidents, and terrorist attacks to slower onset processes such as droughts, heat waves, epidemics, famines, toxic exposures, and frosts. Many of these events/processes will also vary, sometimes according to the available predictive technology, as to the degree of warning and preparation a society may have before onset. However, as they occur and unfold, all dimensions of a society and the totality of its relations with the environment may become involved, affected, and focused, expressing consistency and inconsistency, coherence and contradiction, cooperation and conflict, hegemony and resistance, expressed through the operation of physical, biological, and social systems and their interaction among populations, groups, institutions, and practices. Like few other phenomena the multidimensionality of disasters forces us to confront the many and shifting faces of socially constructed reality(ies).

As events and processes, hazards and disasters are totalizing phenomena, subsuming culture, society, and environment together. Anthropology as a discipline embraces in its substantive platform the physical, biological, and social aspects of human existence. It includes developmental and comparative perspectives. Anthropology frames specifics against larger wholes and examines issues of social change and evolution. With the holistic perspective that is the hallmark of the field, anthropology employs a theoretical framework that can encompass the entire scope of disaster causation and impact, even for the analysis of the essentially novel conditions that emerged in human–environment relations in the latter half of the twentieth century. In addition, disasters present anthropology with important contexts to test key theoretical formulations regarding such themes as social organization, the moral economy–political economy question, and human–environment relations.

Although anthropologists occasionally studied disasters since its post–World War II emergence as a research field, only in the last 25 years have disasters actually become a significant focus for anthropological research. However, a number of anthropologists were among the earliest contributors when the field emerged in the 1950s. Early anthropological studies generally occurred because anthropologists were examining other issues in places that had experienced disasters. Moreover, prior to 1970 these studies constituted the only disaster research, anthropological or otherwise, in the developing world. The field of disaster research in general was entirely focused on European and North American events, where anthropologists did not generally study, with the notable exception of the lasting contribution of Anthony F.C. Wallace's research on a tornado in Worcester, Massachusetts tornado from which he developed his spatial/temporal framework for disaster events (1956) that still informs much disaster research today.

As social scientific disaster research began to expand in the 1950s, anthropological participation remained low, due largely to a particularist and functionalist emphasis on the construction of cultural profiles based on the ethnography of "normal, daily" life that precluded addressing the issues of disruption and change that disasters represented. Much disaster research, including anthropological studies along with other fields such

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12188-9

as sociology and geography, tended to focus around the event in both space and time.

The reemergence of interest in sociocultural change, development (and underdevelopment), and cultural ecology in the 1960s, however, created a set of theoretical fields that would eventually encourage and enable anthropologists to engage with the issue of disasters in the 1970s. In effect, these emerging theoretical concerns addressing issues of human environmental adaptation in the past or present required a consideration of the interaction of the social, technological, and natural processes of change and human welfare that revealed the importance of hazards and disasters (Torry, 1979).

A Paradigm Shift

The decade from the mid 1970s to the mid 1980s marks a watershed in disaster research generally, and in the anthropological study of disasters specifically, with the emergence of two concepts that drove a paradigmatic change. One of these concepts, vulnerability, emerged from disaster research itself and the other, resilience, saw its origins in the field of ecology, which in turn had drawn from engineering and physics. Both concepts required greater scrutiny of the underlying causes of both disaster loss and damage and are also complexly intertwined with the questions of adaptation and development. Indeed, both concepts are foundational for important advances in our understanding of disasters in anthropology and in the other social sciences. Importantly, the shifts in perspective toward vulnerability analysis occurred in the same time frame as a number of large disasters in the developing world, such as the Peruvian earthquake (1970), the Guatemalan earthquake (1976), and the Bangladesh flood (1987) and increased media coverage of these events. Furthermore, there was greater appreciation by anthropologists that major events were interspersed with other innumerable deadly events of sometimes only slightly lesser magnitude that were occurring in the nations and regions where they were working. Depending on the definition of disaster, many anthropologists also began to discover that the societies they were studying were actually, in one form or another, postdisaster societies (Oliver-Smith and Hoffman, 1998).

The concept of vulnerability focused attention on those aspects of society that reduce or exacerbate the impact of a hazard. Basically, anthropologists and cultural geographers working in the developing world found that mainstream disaster studies brought little of value to the task of analyzing disasters in Third World contexts. These researchers criticized the essentially passive role prior investigators had assigned to society in risk etiology and the scant attention paid to local, national, and international factors in creating or exacerbating both risk and impact. They called for a rethinking of disasters from a political economic perspective, based on the high correlation between disaster proneness, chronic malnutrition, low income, and famine potential, leading to the conclusion that the root causes of disasters lay more in socially constructed conditions of vulnerability than in nature (O'Keefe et al., 1976; Hewitt, 1983). The work of the Network for Social Research on Disaster Prevention in Latin America, known as La RED, particularly contributed to the expansion of the vulnerability-risk perspective beyond the Euro-American context (Lavell, 1994). Vulnerability, soon generally called social vulnerability, refers to those characteristics of a society that render its members subject to harm and loss in a disaster. Vulnerability expresses the multidimensionality of disasters by focusing attention on the totality of relationships in a given social situation into the formation of a condition that, in combination with environmental forces, produces a disaster.

Moreover, because vulnerability was most intense in the developing world, where most anthropologists worked, they and other social scientists began to draw an association between disasters and development, or underdevelopment. The enormous disparities in deaths and losses between disasters in the developed and developing worlds thus identified high disaster mortality and destruction as features of underdevelopment. This is, in essence, a political ecology of disasters. Factors such as the lack of access to resources, lack of political power and representation, inadequate housing and infrastructure, poor sanitary conditions, lack of access to education and health care, and density of settlement, all these characteristics of underdevelopment, have been cited as most frequently associated with high vulnerability (Wisner et al., 2004).

At roughly the same time, ecological anthropologists concerned with advancing the field of cultural ecology urged researchers to focus on hazards from an ecological and social organizational perspective (Vayda and McKay, 1975). Ecological science itself also began moving away from the idea of ecosystems as static entities in equilibrium toward systems in which change is normal, asserting that disturbances actually increase the chances that a given ecological system will not return to a former state, but will nonetheless persist. In other words, it is resilient (Holling, 1973). The concept of resilience refers to "the capacity of an eco-system to undergo disturbance and maintain its functions and controls and may be measured by the magnitude of disturbance the system can tolerate and still persist" (Wallington et al., 2005: 15).

Social scientists soon saw the application of the concept to society, framing resilience as the ability of social groups or individuals to bear or absorb sudden or slow changes and variation without collapsing is social resilience (Holling and Meffe, 1996). Indeed, Holling argues that this stream of research "… has the most natural connection to related ones in the social sciences that are historical, analytical, and integrative. It is also the stream that is most relevant for the needs of policy and politics" (1993: 553). Like vulnerability, resilience is a function of social organizational features that govern the interaction of human systems and the natural world. Basically, resilience refers to the ability to prepare and plan for, absorb, recover from or more successfully adapt to actual or potential adverse events (National Research Council, 2012). Rather than the opposite of vulnerability, resilience interacts dialectically with vulnerability in the form of dynamic adaptations to environmental and social processes in moving from disaster risk reduction to sustainability (Aguirre, 2007).

In this view, community resilience is a constantly evolving process of learning and adaptation to new knowledge,

information that is dialectically related to disaster risks and emerging vulnerabilities. Disaster resilience may also be framed as both an outcome and a process where there are inherent vulnerabilities and resiliencies in the community prior to an event and these influence the ability of the community to absorb the impacts, which result in either rapid or slow post-disaster recovery. Therefore there are constant feedbacks between vulnerabilities and resiliencies such as preparedness and disaster mitigation activities, as well as social learning that occur as a means for enhancing resilience over longer time frames (Cutter and Morath, 2013).

Both vulnerability and resilience to hazards are complex and multidimensional as are the specific social contexts and conditions they express and reflect. As the resilience and the social vulnerability perspectives have developed, both generated similar policy relevant paths. Both fields of inquiry internalized and "routinized" risk and disturbance. Both emphasized the centrality of interactions across temporal and spatial scales, particularly the importance of historical analysis in assessing current conditions. Finally, both vulnerability and resilience required a new perspective on the relationship between society and environment. In effect, with the concepts of vulnerability and resilience both the social and the natural sciences moved from environment and society as a duality to a more mutually constitutive relationship. This mutual constitution becomes abundantly clear in the context of disasters and even more so with climate change. Thus, in this understanding, people are not just vulnerable to hazards, but hazards are increasingly the result of human activity, not just technologically, but in terms of human alteration, or construction of the environment. Disasters more than ever now express and energize most clearly that mutuality (Oliver-Smith, 2002).

This more complex understanding of vulnerability enables researchers to conceptualize how social systems generate the conditions that place different kinds of people, often differentiated along axes of class, race, ethnic, gender, or age, at different levels of risk from the same hazard and suffering from the same event. By incorporating specific political, economic, and social variables in combination with specific environmental features, the cause of disasters is situated in the interface of society and environment.

Disasters, no matter how large, are experienced first at the local level. Even an enormous disaster affecting great areas and legions of people, while it may result from a single climatological, geological, or technological phenomenon, ultimately comes down to a compendium of local but related disasters experienced throughout the region and numbers affected. Not all communities experience a disaster in the same way or to the same degree. Each community undergoes a catastrophe in the context of its culture, social structure, and profile of vulnerability. The fact that vulnerability is unequally distributed means that the same disaster agent will show great variation in both patterns of destruction as well as interpretation of cause, effect, and responsibility. It sharpens our understanding of disasters as the outcomes of specific features of systems, social and otherwise. In colloquial terms, it makes it clear that disasters just do not happen. They are features, characteristics of systems of social and material relations, albeit very complex systems.

Complexity, Risk, and Disaster in Contemporary Anthropological Perspective

As we enter the second decade of the twenty-first century, anthropological disaster research, and the field of disaster research in general, is in flux, grappling with increasing levels of vulnerability among increasing populations in many areas of the globe. Debates are active around such issues as the role of culture in disaster vulnerability and resilience, the concepts of coping and adaptation, the role of development in driving disaster risk, and the interface between disasters and climate change. While natural and technological sciences are advancing understanding of hazards significantly, the major challenges for disaster risk reduction are found in the domains addressed by the social sciences. Anthropological disaster research is confronting these challenges, sharing a variety of perspectives and methods with other social sciences. While anthropology in general has adopted some quantitative methods, there is now a broad social scientific recognition that ethnographic methods are extremely effective in capturing the processual dimensions of disaster vulnerability, disaster events, and postdisaster processes of recovery.

The salient characteristics of anthropological research on disasters continue to be longitudinal and processual, oriented by a recognition that disasters involve far more than catastrophic events. They are recognized as processes that unfold through time and their beginnings are deeply embedded in societal history. These concerns relate directly to the relationship between humans and nature and possible strategies for more sustainable forms of development. The anthropological focus has contributed to a shift from an exclusive focus on the disaster site to greater attention to the multiple sites where both policies and practices are developed as well as outcomes play themselves out. Therefore, anthropological disaster research focuses not solely on space and time, but on the social construction of risk and vulnerability, perhaps driven by nonlocal forces, that historically preceded the event, the event(s), and the operative social relations among individuals and groups, both formal and informal, through the various stages of impact, emergency, relief, and reconstruction. There is also an emphasis on the role of social relations and cultural orientations, knowledge, values, norms, and beliefs as they are reestablished, reinvented, renegotiated, discarded, or renewed in the process of recovery. Today other social sciences have adopted theoretical approaches and methodologies that are at least partially informed by the anthropological perspectives on culture, society, and the environment.

A number of current trends would appear to indicate that, indeed, an increase in impact and scope of hazards and resulting disasters, already complex and multidimensional, is taking place through combined effects of economic, social, demographic, ideological, and technological factors, that are compounding their complexity many times over. Greater numbers of people are more vulnerable to natural hazards than ever before, due in part to increases in population, but more so, to their location in dangerous areas. We have also created new forms of disaster agents and new vulnerabilities. Indeed, each new advance in technology brings with it a set of risks and vulnerabilities framed by its potential failure. Today, the creation of completely new technologies, whose mere

implementation, regardless of potential or actual malfunction, has had profound environmental and, in some cases, catastrophic impacts. Moreover, global climate change is now recognized as a major threat to human welfare and security that must be addressed both locally and globally. Essentially, it is now predicted that if we continue to use fossil fuels as usual and do not reduce carbon emissions, the eventual outcome may be something similar to that which earth experienced during the time of mass extinctions.

Although its relevance to daily realities in the developing world is debatable, Beck (1992) has referred to the cultural milieu emerging from the recognition of this potential as a globalized 'risk society,' largely to characterize a western society preoccupied with environmental degradation and the threats it represents for terrestrial life.

Anthropological Research on Hazards and Disasters in the Twenty-First Century

Anthropological research encompasses the full spectrum of social research on disaster complexity, vulnerability, warning, impacts, humanitarian aid, recovery, and reconstruction. Certainly, the research tradition of the years prior to the paradigm shift, focusing on individual and organizational behavioral responses, power relationships, economic impacts, and patterns of persistence and change, has continued. Since the paradigm shift anthropological research has acquired a number of different trajectories. However, to give full scope to current anthropological research on disasters is well beyond the available space for this article. Nonetheless, several research streams bear mention for their recent contributions.

In the fields of archeology, ethnohistory and historical archeology the anthropological concern with longitudinal analyses has advanced understanding of disasters in many cases deep into history and prehistory. Archeologists have traditionally been contributors to anthropological analyses of disasters. Recent work in the archeology of disasters has shed light on such issues as the capacities of differently organized prehistoric societies to recover, environmental thresholds of past decision-making, paleo-environmental interaction and previous environmental capacities and human effects on environments, cultural evolution and human management of different environments and the role of the past in creating the ethnographic present (Bawden and Reycraft, 2001; Garcia Acosta, 2002). Both prehistoric and historical archeologists, historical ecologists, historians, cultural anthropologists and other historically oriented anthropologists who are concerned with the long term effects of disasters are exploring human ecodynamics operating in different areas and on different time scales. Their objective is to get a better comparative handle on the interaction of climate change, human environmental impacts, and human–human interaction through a long term perspective on sustainability, adaptation, resilience, spread of pathogens, and threshold crossing changes (www.gheahome.org).

The pursuit of root causes over extended periods of time has led anthropology into more intensive analysis of historically embedded and spatially distant forces behind disaster risk and vulnerability, seeking drivers in broader regional and global systems. In a related vein and continuing the trend of political economic analyses of disaster vulnerability and impact, recent anthropological and other publications document the ways in which crises provide opportunities for the penetration of corporate capitalism and the restructuring of societies. These analyses document such resource transfers disguised as disaster reconstruction or mitigation as the post-tsunami relocation of fishermen and the sale of their land to tourist enterprises and the permanent closing of public housing and the transfer of these facilities to private development in New Orleans after Katrina (Gunewardena and Schuller, 2008). This strain of anthropological research focuses on how people are revictimized by general reconstruction initiatives and specific project design that in effect advance elite interests. These studies point to the implementation of the neoliberal project (privatization, trade liberalization, structural adjustment, decreased social spending, etc.) in disaster settings. Indeed, this research discloses the not always hidden subtexts that are conveyed in interactions between aiders and victims, resulting in greater integration into the state and the market, which some would define as fundamental to standard development practice, which is frequently implicated in the construction of disaster risk.

Over the last 10 or 15 years there has been a greater emphasis in anthropology on developing more situated and sophisticated analyses of the complex relations between humanitarian and development assistance actors and the communities they intend to help. These ethnographies focusing on the micro-sociology of relief and reconstruction follow the complex processes of decision-making about project design and implementation among the diverse social actors from the transnational sector and local populations Benadusi et al., 2011). In effect, these ethnographies foreground the analysis of interactions between the world of practitioners, and the values and norms that guide them, and the world of local populations who must contend not only with disasters, but with the sometimes disruptive impacts of assistance programs, not to mention their frequent failure. The uneven record of disaster preparedness, relief, and reconstruction can in some measure be attributed to failure to take into account local contexts, social relations, and values in and through which disasters are experienced and interpreted. Disasters, no matter how large, are experienced first at local level and effective disaster assistance whether through humanitarian emergency relief or longer-term reconstruction must be implemented in ways that empower local people, rather than restoring or exacerbating local inequalities and vulnerabilities.

Among the vulnerable groups that have been identified, the impact on and the role of women in disasters have taken a central role in research and analysis. Prior to 1995, apart from a few pioneer studies, there was little attention to the issue of women, much less gender in general in disasters. Along with the links between disaster and economic or racial/ethnic stratification, researchers also tended to ignore gender, despite the fact that gender divisions pervade all societies and channel access to resources, generally toward men and away from women, making them more subject to hazard risk and more vulnerability to disaster impact (Enarson and Hearn-Morrow, 1998). Because they are often naturalized, gender relations remain implicit. Men and women for the most part act

according to socially defined expectations, which in the context of human security, generally produces a gender-related social vulnerability (Bari, 1998; Hoffman, 1999). The interrelation between environmental risk and social vulnerability foregrounds a focus on gender as an aspect of social vulnerability. Anthropological studies of gender relations in disaster reveal how they impinge on conditions of risk, vulnerability, impact, and the experience of recovery and reconstruction. For example, disaster management is gendered, shaping organization and dynamics of assistance agencies and operations such as the fact that delivery of assistance is often designed for and delivered to male household heads. Research has also focused on women's capacities as well as vulnerabilities, highlighting the fact that the gendered division of labor often makes women both first responders in impact and long-term caregivers in disaster stricken households. However, women are not the only vulnerable group to be considered by anthropological disaster research. The risks and vulnerability of the elderly and children as well as ethnic minorities has also become an important research stream.

Reconstruction remains, along with disaster risk and vulnerability reduction, our most pressing problem in disaster studies. It is also the least studied by any social science discipline. Why that is so is rooted in the challenges that reconstruction presents to researchers, challenges that anthropologists are uniquely prepared to confront. The discipline's foundational research methodology, ethnography, is particularly appropriate for the process of reconstruction because it is a lengthy process in which people must traverse the difficult path between restoration and change (Ensor, 2009). This process of negotiation between what has been destroyed, what is to be reconstituted and what is to be replaced generally involves tensions among diverse interest groups and cultural norms and values. Methods that privilege narrative and observation with researchers present and in dialogue with participants to gather local knowledge and information are far more appropriate for exploring the process of adjustment and recovery than are more synchronic forms of research.

Postdisaster reconstruction is most often approached as a material transfer process rather than a process of social reconstitution. Disasters are complex phenomena and their complexity is refracted into the process of disaster reconstruction. Social recovery is therefore inextricably linked to the recovery of structures and infrastructural elements, ecosystems, organizations and institutions, economic activity, and culture, making recovery a truly holistic process. From the perspective of applied anthropology, the fundamental question to be asked is how reconstruction can address the complex of environmental, economic, and social variables that produce disasters in a way that will be sustainable, reduce vulnerability, and enable people at the household and community level to survive. In other words, the challenge of reconstruction is how to reconstruct and recover without reinstalling the same system that generated the vulnerability in the first place. Reconstruction becomes a test of the capacity of the system to respond to a clear demonstration that human and environmental destruction is deeply rooted in the social and environmental history. Anthropologists have recognized that reconstruction, to be truly transformative, has to address that complex of factors that makes disasters into the horrors that they become. In other words, reconstruction has to recognize and reflect the complexity, and in its policies and practices address the mutuality between society and nature that resides at the core of any disaster. Moreover, there is increasing cross-fertilization between research and practice on disaster induced resettlement and development forced resettlement based on the recognition that both involve a process of reconstruction of community (Cernea and McDowell, 2000).

Nowhere is that mutuality more clearly expressed today than in the challenges presented by global climate change. The reports from the Intergovernmental Panel on Climate Change (IPCC, 2007a,b) affirm that human induced factors are responsible for generating significant increases in temperatures around the world. Among the consequences of this rise in temperature are increases in the rate of sea level rise, increases in glacial, permafrost, Arctic, and Antarctic ice melt, more rainfall in specific regions of the world and worldwide, more severe droughts in tropical and subtropical zones, increases in heat waves, changing ranges and incidences of diseases, and more intense hurricane and cyclone activity. The changes will disrupt lives and destroy property and livelihoods at the local level, principally through loss of ecosystem services, including food, water, fuel, and nutrition, as well as those cultural elements, largely spiritual and/or aesthetic, that sustain communities through expressive links to natural features. In addition, loss of land due to desertification, sea level rise, and displacement will increase local vulnerability and deprive people of viable livelihoods. The increased intensity and frequency of climate based natural disasters also threaten local communities in many coastal regions around the world. With the increase in surface temperature over the last 50 years, the levels of damage from extreme weather events have increased. On a global scale, losses from natural disasters have increased dramatically over the last half century, particularly so since the middle of the 1980s (Munich Re, 1999: 16). Will the impacts of global climate change qualify as disasters? Regardless of how analysts view these processes, it is fairly clear that their outcomes will be seen and felt as disasters by the affected populations (Crate and Nuttall, 2009).

Moreover, loss of ecosystem services, loss of land, and increased and intensified climate events can be said to interact with each other to compound and intensify the effects of all of them. Global climate changes, in addition, also may combine with other factors, such as environmental contamination, to increase food and water stress, enhance hazard risk, and drive people from their homes. The long tradition of anthropological research on local adaptation to environments generally and climates specifically across cultures and over time has lent particular salience to the field for understanding and managing current challenges presented by global climate change (Strauss and Orlove, 2003). All the previously discussed anthropological perspectives and methodologies, including archeology, ethnohistory and ethnography, that bring insight to specific disaster related topics across temporal and spatial scales are equally relevant to understanding climate change driven processes in specific prehistoric, historic, or contemporary communities. Moreover, while the effects of climate change will be manifested globally, the most serious and damaging impacts will be felt in those poorer and lesser developed countries and regions of the world that have least contributed

to the problem and have fewer resources with which to confront the challenges.

From an applied perspective, anthropology can assist local communities to understand the risks associated with climate change impacts, integrate such risks with local knowledge to develop adaptive responses, and also defend their rights and make their needs known in national policy contexts. In effect, climate change will in most cases simply exaggerate the effects and frequencies of existing hazards, which are largely the outcome of reigning patterns of local vulnerability. Indeed, climate change effects will also increase the vulnerability of people to geological and other hazards not related to climate change. Disaster risk reduction must continue to focus on local, national, and international forces that make people vulnerable to events and processes made even more extreme by climate change effects. Thus, both climate change adaptation and disaster risk reduction must be framed and designed to address those social and economic features that render people vulnerable to environmental hazards in general. Climate change adaptation and disaster risk reduction then are related undertakings, and both must address systemic vulnerabilities as well as hazards posed by specific climate change effects as fundamentally development related phenomena (Kelman and Gaillard, 2010; Lavell, 2011). In that sense, the most effective overall policy for both climate change adaptation and disaster risk reduction is addressing the systemic vulnerabilities and reducing inequality. This policy in effect is also entirely consistent with a policy of sustainable development.

As the fundamental concept of the discipline, culture plays a significant role in the anthropological framing of disaster processes in general. As that complex of thought and behavior has enabled human beings to transmit and accumulate knowledge through time, adapt to myriad environments and generate the enormous diversity of the species, culture is the context in which disaster is experienced. Culture is the grid or lens though which we experience and interpret the material world and the world of our experience through time. It is through cultural knowledge, belief, attitudes, that we generate behavior or act. All the material, economic structures that channel our behaviors are ultimately based on, derived from and mediated by culturally based values about life and the world, values about what the relationships should be between human beings and between human beings and the material world. In other words, even physical forces like earthquakes and hurricanes have to be interpreted before they can be responded to. As Hornberg has recently noted, culture is crucial to socioecological processes and relations. It shapes our way of thinking about ecology, economy, and technology and generates our specific kinds of fetishism and consumption, many of which specifically channel the way we inhabit and use our environment (2009: 255).

Although consistently a subtheme of disaster research, the role of culture in disasters has recently gained greater attention among anthropologists, particularly in the way people experience and interpret disaster threat, warning, impact and recovery. A focus on culture reminds us that not all reconstruction is infrastructural nor are all resources material. The role of memory, not only of predisaster systems, but also of public recognition of significance of loss, is also engaged to reveal its links to hegemonic socioeconomic structures. This research tells us that much of recovery is posited on the restoration, not necessarily of a predisaster system but of a system that allows for predictability of the outcome of actions. Recovery then is less about the process of material reconstruction, and more about the way a sense of logic and predictability, always one of the fundamental aspects of culture, is restored. They also tell us that it is crucial to restore a sense of subjective agency and control, not only in confronting the impact of the disaster, but also in dealing with the uncertain living conditions that characterize life prior to the disaster, particularly the lives of the poor in the developing world in general. There has also been a tradition of contributions on the role of religion in the explanation of loss and tragedy, in the development of cultural resilience, and in the reformulation of meaning in the aftermath of disasters (Bode, 1977, 1989; Carter, 2010). In all these studies the importance of ritual and symbol play central roles in the restoration of meaning to life in the aftermath of disasters.

Current approaches frame culture as an array of discourses – essentially a conversation about how life is lived and how it should be lived. It is culture as a dynamic process of production of understandings, as constantly being produced, interpreted, reinterpreted, imposed, and resisted. Society is not generally the organization of completely equitable distribution of resources, but a dynamic arena of contesting interpretations and interests organized along some lines of differentiation, whether they be only age and gender or in more complex situations, color, class, caste, religion, ethnicity, kinship, or any of the other myriad ways that humans have of grouping themselves and distinguishing others as different.

From this perspective, culture is not a single voice, but multiple voices sharing some perspectives, differing in others, all articulating their particular interpretations. Some voices, however, can by virtue of different kinds of power and wealth, come to dominate the conversation. Nowhere is this more apparent than in the conditions of vulnerability and the aftermath of disasters. Everyone in a disaster has a particular perspective and a particular set of interests, perhaps shared with others with whom they share certain characteristics. Thus, disasters are both socially constructed and experienced differently by different groups and individuals, generating multiple interpretations of a complex event/process. A single disaster can fragment into different and conflicting sets of circumstances and interpretations according to the experience and identity of those affected (Hoffman and Oliver-Smith, 2002). The disaster process, evolving through various stages, that while difficult to demarcate exactly for and experienced differently by every affected individual and group, displays characteristic social relations, informed by situated knowledge forms and value systems, among the affected population as well as between the affected population and humanitarian and reconstruction aid agents. That is, different forms of knowledge and different moralities become activated as the character of social relations changes in the evolution of the disaster process.

Disasters almost always bring to the fore tensions in the local cultural scene, the contestation of the different voices in the conversation about what the society should be, between those who want to return to the status quo and those who want

change, between the need to restore the known, the understandable, the familiar and those who want to use the disaster as an opportunity to change society, and many positions along the spectrum between those two perspectives. In effect, disasters can challenge cultural coherence and the integration of society. In some contexts, people come together around shared cultural understandings and goals in a kind of postdisaster social solidarity that nurtures the formation of resilience, and in others, particularly where powerful interests work to reinstate an inequitable status quo, the society is left often more vulnerable than before the disaster. In the final analysis, both vulnerability and resilience are expressions of cultural values about human social relations and how those relations relate to material circumstances.

Conclusion

Disasters focus with great intensity the widest possible variety of intersecting and interpenetrating processes and events of social, environmental, cultural, political, physical, and technological natures, requiring an integrated or holistic analytical perspective for full understanding. In disasters, all dimensions of a society and its relations with the total environment become involved, affected, and magnified, articulated through the operation of physical, biological, and social systems and their interaction among populations, institutions, and practices. There are few contexts in which the mutual constitutionality of the physical and the social are so starkly displayed as in a disaster. Contemporary events have dramatically displayed both long-standing threats and emerging risks and vulnerabilities that need to be confronted. In addition to climate change, research trends now beginning to appear in current anthropological analyses include the interface between critical infrastructures (energy, water, etc) and risk construction, the links between disaster, displacement, migration and resettlement, and the gap between disaster research and practice.

Although the multidimensionality of disasters presents a complex challenge to both researchers and practitioners, anthropologists have found an important opportunity to ask synthetic questions. Mining this rich vein that disasters bring into intense focus, presents not only a challenge but an important opportunity to social science theory as well as practice. Disaster researchers and practitioners need to tackle this multidimensionality and probe the multiple linkages and relationships they reveal for sociocultural theory and, by the same token, establish what anthropology can offer to the study of disaster both in the realm of research as well as practice. It is in this multidimensionality of disasters that the salience of anthropological theory and application manifests itself. It is both appropriate and necessary that theoretical and policy oriented projects in anthropology and the social sciences in general be closely linked. If policies are not based on a solid understanding of human behavior in general and cultural behavior specifically, their success in terms of how they respond to human needs, is jeopardized. By the same token, policy and practice can form the testing ground for theory. In broader terms, then, disaster research constitutes a major theme to address both theory and practice in anthropology.

See also: Community Resilience to Disasters; Demographic Impact of Disasters; Genetics and Disaster Victim Identification; Mass Trauma: Psychopathological Effects across the Life Span; Natural Disasters: Health-Related Aspects; Social Work and Disasters.

Bibliography

Aguirre, B.E., 2007. Dialectics of vulnerability and resilience. Georgetown Journal of Poverty Law and Policy 14 (1). https://articleworks.cadmus.com/geolaw/z5800107.html.

Bari, F., 1998. Gender, disaster, and empowerment: a case study from Pakistan. In: Enarson, E., Hearn-Morrow, B. (Eds.), The Gendered Terrain of Disaster: Through Women's Eyes. Praeger, London, pp. 1–8.

Bawden, G., Reycraft, R., 2001. Environmental Disaster and the Archaeology of Human Response. Anthropological Papers No. 7. Maxwell Museum of Anthropology and University of New Mexico, Albuquerque, NM.

Beck, U., 1992. Risk Society: Towards a New Modernity. Sage, Thousand Oaks, CA.

Benadusi, M., Brambilla, C., Riccio, B., 2011. Disasters, Development and Humanitarian Aid: New Challenges for Anthropology. Guaraldi, Rimini.

Bode, B., 1977. Disaster, social structure and myth in the peruvian andes: The genesis of an explanation. Annals of the New York Academy of Sciences 293, 246–274.

Bode, B., 1989. No Bells to Toll: Destruction and Creation in the Andes. Scribners, New York.

Carter, R., 2010. The Blessed Placemakers: Violent Crime, Morality and Urban Redevelopment in Post Katrina New Orleans. Unpublished PhD dissertation, University of Michigan, Department of Anthropology, Ann Arbor, MI.

Cernea, Michael M., McDowell, Christopher, 2000. Risk and Reconstruction: Experiences of Settlers and Refugees. The World Bank, Washington, DC.

Crate, S., Nuttall, M., 2009. Anthropology and Climate Change: From Encounters to Actions. Left Coast Press, Walnut Creek, CA.

Cutter, S.L., Morath, D.P., 2013. The evolution of the social vulnerability index. In: Birkmann, J. (Ed.), Measuring Vulnerability to Natural Hazards, second ed. United Nations University Press, New York and Tokyo.

Enarson, E., Hearn-Morrow, B., 1998. In: Enarson, E., Hearn-Morrow, B. (Eds.), The Gendered Terrain of Disaster: Through Women's Eyes. Praeger, London.

Ensor, M.J., 2009. The Legacy of Mitch: Lessons from Post-Disaster Reconstruction in Honduras. The University of Arizona Press, Tempe, AZ.

Garcia Acosta, V., 2002. Historical disaster research. In: Hoffman, S., Oliver-Smith, A. (Eds.), Catastrophe and Culture: The Anthropology of Disasters. The School of American Research Press, Santa Fe, NM.

Gunewardena, N., Mark, S. (Eds.), 2008. Capitalizing on Catastrophe: The Globalization of Disaster Assistance WalnutCreek, AltaMira, CA.

Hewitt, K., 1983. Interpretations of Calamity. Allen & Unwin, Inc., Boston, MA.

Hoffman, S., 1999. The regenesis of traditional gender patterns in the wake of disaster. In: Oliver-Smith, A., Hoffman, S.M. (Eds.), The Angry Earth: Disaster in Anthropological Perspective. Routledge, New York.

Hoffman, S., Oliver-Smith, A. (Eds.), 2002. Catastrophe and Culture: The Anthropology of Disasters. The School of American Research Press, Santa Fe, NM.

Holling, C.S., 1973. Resilience and stability of ecological systems. Annual Review of Ecology and Systematics 4, 1–23.

Holling, C.S., Meffe, G.K., 1996. Command and control and the pathology of natural resource management. Conservation Biology 10 (2), 328–337.

Hornborg, A., 2009. Zero sum world: challenges in conceptualizing environmental load displacement and ecologically unequal exchange in the world-system. International Journal of Comparative Sociology 50, 237. http://dx.doi.org/10.1177/0020715209105141.

Intergovernmental Panel on Climate Change (IPCC), 2007. Summary for policymakers. In: Soomon, S., Qin, C., Manning, M., Cher, Z., Marquis, M., Avery, K.B., Tignor, M., Miller, H.L. (Eds.), Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York.

Kelman, I., Gaillard, J.C., 2010. Embedding climate change adaptation within disaster risk reduction. In: Shaw, R., Pulhin, J.M., Pereira, J.J. (Eds.), Climate Change Adaptation and Disaster Risk Reduction: Issues and Challenges. Community, Environment and Disaster Risk Management. Emerald Group Publishing Limited, Bingley, UK, pp. 23–46 (Chapter 2).

Lavell, A., 1994. Viviendo en Riesgo: Comunidades vulnerables y prevencion de desastres en America Latina. FLACSO/La RED/CEPREDENAC/Tercer Mundo Editores, Bogota.

Lavell, A., 2011. Unpacking Climate Change Adaptation and Disaster Risk Management: Searching for the Links and the Differences: A Conceptual and Epistemological Critique and Proposal. IUCN-FLACSO Project on Climate Change Adaptation and Disaster Risk Reduction.

Munich Re, 1999. Topics: Natural Catastrophes. The Current Position. Munich Reinsurance Company, Munich.

National Research Council, 2012. Disaster Resilience: A National Imperative. National Academies Press, Washington, DC.

O'Keefe, P., Westgate, K., Wisner, B., April 1976. Taking the naturalness out of natural disasters. Nature 260, 556–557.

Oliver-Smith, A., 2002. Theorizing disasters: nature, culture, power. In: Hoffman, S.M., Oliver-Smith, Anthony (Eds.), Culture and Catastrophe: The Anthropology of Disaster. The School of American Research Press, Santa Fe, NM.

Oliver-Smith, A., Hoffman, S.M. (Eds.), 1998. The Angry Earth: Disaster in Anthropological Perspective. Routledge, New York.

Strauss, S., Orlove, B.S., 2003. Weather, Climate, Culture. Berg Publishers, London.

Torry, W., 1979. Anthropological studies in hazardous environments: past trends and new horizons. Current Anthropology 20, 517–541.

Vayda, A., McKay, B., 1975. New directions in ecology and ecological anthropology. Annual Review of Anthropology 4, 293–306.

Wallace, A.F.C., 1956. Tornado in Worcester: An Exploratory Study of Individual and Community Behavior in an Extreme Situation. NAS-NRC Disaster Study #3. National Academy of Sciences-National Research Council, Washington, DC.

Wallington, T., Hobbs, R.J., Moore, S.A., 2005. Implications of current ecological thinking for biodiversity conservation: a review of the salient issues. Ecology and Society 10 (1), 15. http://www.ecologyandsociety.org/vol10/iss/art15.

Wisner, Ben, Blaikie, P., Cannon, T., Davis, I., 2004. At Risk: Natural Hazards, Peoples' Vulnerability and Disasters, second ed. Routledge, London.

Hazing in Fraternities and Sororities

Hank Nuwer, Franklin College, Franklin, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Hazing in U.S. university fraternal groups is a pernicious and sometimes even deadly practice that dates back to the founding of the first collegiate fraternities in the nineteenth century. In spite of growing antihazing sentiment, laws against hazing in 44 states and campus crackdowns, hazing persists as a rite of passage and tradition in colleges and universities across the United States and around the world. This article examines the roots and manifestations of hazing in university-based fraternities and sororities in the United States. It includes a discussion of university policies and practices to curtail it by national fraternal organizations and antihazing activists. Best practices include the encouragement of bystander intervention to halt hazing as it happens, along with the mandating of alcohol-free (dry) fraternity houses and an end to the pledge period.

Defining Hazing

The number of serious researchers devoted to studying and highlighting hazing is still limited, but the subject is attracting increasing study from social scientists and other academics concerned with trying to end the unabated string of deaths every year from 1970 to 2013 in fraternal chapters, as well as in other clubs and sports teams.

What exactly is hazing? Researcher Aldo Cimino defines the term as "the generation of induction costs (i.e., elements of the experiences necessary to be acknowledged as a 'legitimate' group member) that appear unattributable to group relevant assessments, preparation, or chance." He cites calisthenics being required by a fraternity as an example (Cimino, 2013). Cimino cites automatic accrual theory to explain why higher status fraternities with more and better benefits for a pledge can demand far more severe tests of hazing than a chapter with less status and fewer benefits can expect (Cimino, 2013).

Most higher-education institutions in the United States use the definition of the practice of hazing devised by the Fraternity Executives Association and endorsed by the Fraternal Information and Programming Group, the leading fraternal organization addressing risk-management issues. Hazing is "any action taken or situation created intentionally, whether on or off fraternity premises, to produce mental or physical discomfort, embarrassment, harassment or ridicule" (www.fipg.org). All higher-education institutions and all national and international fraternities and sororities in the United States publish strict policies forbidding hazing. Hazing also meets a definition of bullying if fraternal members put newcomers through tortures intended to make the pledges, neophytes, or associate members quit a chapter, as opposed to the usual hazing methods of putting newcomers through stressful conditions in order to welcome them eventually into the group (Nuwer, 2004).

Early work by Lionel Tiger of Rutgers University promoted the term 'male bonding' and made a case for the evolutionary aspects of hazing from his neo-Darwinian perspective. In 1969, in his book *Men in Groups*, revised in 2004 by Tiger as 'Males Courting Males' in *The Hazing Reader* by Hank Nuwer, he focused on hazing in male fraternities as a means of addressing and explaining some hazing activities he witnessed firsthand at McGill University, a Canadian institution of higher learning. Tiger states that men attracted to higher-status groups, actual or perceived, are attracted in part because they define their own self-image from attractive, higher-status males. In groups such as fraternities that haze (and not all do haze), Tiger states in 'Males Courting Males' that the males staunchly try to defend the status quo by incorporating new members willing to show they are a suitable 'fit' by performing and successfully participating in whatever hazing practices the group deems appropriate and suitable. These practices at the chapter level can be crude, even demeaning, and run counter to what the national or international umbrella fraternal groups may approve. Thus, all the over-the-top practices in a chapter get hidden away in hazing episodes done clandestinely behind closed doors out of sight of advisors and national executive officers.

Tiger, in his 'Males Courting Males' essay, stressed that the ability to consume copious amounts of alcohol in the presence of other fraternal values is a common bonding practice practiced by some fraternities and other high-status male groups such as athletes. In 'Males Courting Males,' he also observes that many fraternal chapters that haze require members to participate in activities that appear homoerotic in nature, requiring pledges alone or in the pledge group to be fully or partially nude during one or more hazing sessions and occasionally assuming submissive poses in front of the poses of dominant senior members. Once initiated, the pledge is given the opportunity to buy a costly ring with the fraternity's Greek letters on the stone, and the ring is worn on the finger traditionally reserved for a wedding ring. There also are pins with the fraternal insignia reserved for females that members find attractive.

Tiger's observations on the need for new fraternal members to achieve a 'clubby' bonding with senior members have echoes in the well-established 'groupthink' theory of researcher Irving Janis (1972). Nuwer (1978, 1999a) adapted Janis' term to 'Greekthink.' The basis for 'groupthink' and 'Greekthink' is that the clubby nature of groups noted by Tiger reflects the overarching quest for camaraderie and a desire for harmonious relationships seen in high-status groups such as fraternities, even rising to a U.S. President's inner circle of advisers and cabinet members, athletes, and so on. Which may not be so

surprising given that the Center for the Study of College Fraternity and Inside Greek U., author Alan DeSantis counts 18 U.S. Presidents since 1877 that claimed fraternity affiliations as collegians.

Janis' term, groupthink neatly sums up fraternal organizations, and their compulsive need and appreciation for camaraderie and the approval of current members and even much older alums whose presence at several hazing incidents in which a pledge died was noted by Nuwer. Groupthink captures the behavior of members performing acts of hazing in a group that prior to joining, they as individuals likely would have dismissed as deplorable. Groupthink also explains why fraternal members can display a delusional belief in their own invincibility while encouraging or ordering hazing activities in order to foster group unanimity. The groupthink theory espoused by Janis also explains the deception, denial, and dishonesty that hazing chapters often demonstrate after a risky practice leading to serious injury or death and an associated investigation. Janis' extensive study in group dynamics further points out that the group's excessive need for solidarity colors the judgment of its members, leading them to approve and carry out activities such as hazing and alcohol overconsumption that, if exposed, can lead to that hazing chapter losing its charter. The cost of losing that charter includes the consequence that the university and its national organization order individual members to disband for anywhere from a year to forever and lose the privileges and status of membership during that ban. Yet, in spite of so much to lose, hazing continues and also continues to attract media coverage and legislative attention.

Because actual acts of hazing are widely different, ranging from beatings to drinking extravaganzas to nude groping and lining up to take verbal abuse, it is difficult to point to a single behavioral cause of hazing. Most hazing researchers refer to 'The Effect of Severity of Initiation on Liking for a Group' by Elliott Aronson and Judson Mills as of crucial value. The two researchers established through experimentation that hazing leads to tighter group solidarity because severe initiation practices cause initiates to appreciate and like the group more than do initiates who are simply invited to join a group without hazinglike demands and sacrifices.

Another often-quoted expert on hazing behavior is James C. Arnold who applied the work on cults by Margaret Thaler Singer, coauthor with Janja Lalich of *Cults in Our Midst*, to fraternities (1995). Arnold asserts in his long essay 'Hazing and Alcohol in a College Fraternity' (in Nuwer's *The Hazing Reader*) that many fraternities fall under the category of 'addictive organizations.' Such groups quite convincingly are compared to individuals obsessed with alcohol, for example, and are similar to a dysfunctional family unit. In short, the chapters that haze use cultlike systematic manipulation and coercion to effect psychological and social influence. In particular, cults and hazing fraternities alike purposely engender enforced dependency by instituting ways and means to make potential members and new members spend the majority of waking and sleeping hours in the company of members and their fellow newcomers. Such organizations promise the new initiates enduring abuse that if they only persist all will be well in the end, and they will experience many rewards once awarded membership when the time of trial expires.

Arnold undertook close participant observation as a doctoral student over a long period of time under the supervision of well-known Indiana University Higher Education Professor George Kuh, also the Director of National Institute for Learning Outcomes Assessment. Arnold on many occasions observed his study chapter perform acts of hazing and alcohol abuse while he performed research for his dissertation in Higher Education at Indiana University that was published in Nuwer's *The Hazing Reader*. He was allowed by the chapter, its national organization, and Indiana University to have nearly complete access to such fraternity customs as rush in which attractive newcomers get invited to the house for inspection under party conditions and then are given a 'bid' if enough members agree they possess the 'right stuff' for membership.

Sometimes, the bid results in a good fit, and sometimes it results in conflict when an independent thinker among the new class of pledges raises concerns about the illicit or even illegal behaviors that the newcomers must brush away. The desired outcome of hazing is to achieve 100% pledge class unity – that class fully and wholly subservient to the wishes of the veteran members. Ironically, to get the 'trust' of the pledges, the senior members may lie, deceive, and deny all responsibility, notes Arnold. What occurs during the stress and chaos of a pledgeship marked by hazing is what the ethnographer John van Maanen refers to as 'cultural learning' for a group.

In its barest form, the pledgeship involves full indoctrination of the new members and total obeisance on the part of pledges as a socialization practice. Members justify their poor treatment of pledges and forcing them to endure often-squalid living conditions for weeks under the justification that everything asked of the pledges serves some important bonding purpose, noted Arnold. The chapter he lived with even used such terms as 'responsible hazing' and 'responsible drinking' that were oxymoronic since from rush until initiation, the chapter's behavior was flagrantly irresponsible and reckless. In the end, the outward result of an entire chapter as a model of control is a hazing chapter that Arnold stresses (p. 92) in Nuwer's *The Hazing Reader* must live with 'denial, dishonesty, self-centeredness, confusion, and so on.' All this is done by a group that projects the 'illusion of control' (Nuwer, 2004: p. 93). And a hazing chapter strives to admit members who respect the status quo and work to maintain a homogenous chapter; nothing incurs the collective wrath of a hazing group's members more than a pledge, who refuses to cower, or who actually quits pledging and rejects that group. Arnold establishes that the group treats the unhappy quitter with disdain, anger, and even threats. Conversely, pledges – away from the security of family and home – that buy into the addictive chapter mentality then get rewarded by the senior members with a supply of alcohol, parties, and access to desirable women who tolerate or even support the practices of the addictive chapter. Arnold says addictive groups and members have a dualistic approach to perceiving the world. Everything is all right or all wrong, and pledges are told it is 'our way or the highway.' Whatever good philanthropic activities the chapter supports are, in part, an attempt to cover up the confused, addictive nature of the chapter, Arnold concludes, citing the theories on the pervasive and

persistent nature of addiction noted by *The Addictive Organization* authors Anne Wilson Schaef and Diane Fassel that the national organizations and Greek Affairs staffs at colleges can make a difference if "they can turn away from a model of control and continued participation in an unhealthy system and through their actions demonstrate a healthier way of being" (Nuwer, 2004).

Stephen Sweet uses a frame of symbolic interactionism to explain how some undergraduates attracted to hard-hazing groups will do whatever it takes to belong. Herbert Blumer's symbolic interactionist construct, cited with regard to fraternal groups by Sweet, declares that human beings react to events because the events have meaning for them based on the social interaction they have with one another and then interpret those events with some selectivity – perhaps even outright justification and rationalization for activities banned by society. Blumer's theory in particular fits when applied to fraternal group members whose concept of self is at the time of pledging being crucially shaped. Sweet notes that the fraternal rings, T-shirts, and paddles represent part of the identity pledges, are accepting as they enter a crucial stage of their lives. In turn, the social relationships with those who already wear the colors and paraphernalia of the group they aspire to join can be a powerful lure to those young people who see enduring hazing as a necessary price to pay to get into an organization that has value. Conversely, says Sweet, those who quit pledging rather than endure not only suffer guilt because their pledge brothers will be asked to carry an additional burden, but they also lose a tie and part of their identity by forgoing all connection to the chapter they already may have sacrificed much for during early hazing. The research by T.A. Leemon, who lived with a fraternity chapter for one entire pledging period to observe it, demonstrates that fraternity chapters very deliberately in the hazing process set out to manipulate and to alter each pledge's concept of a self.

Ricky L. Jones also examines hazing in his book *Black Haze: Violence, Sacrifice, and Manhood in Black Greek-Letter Fraternities* (2004). Jones postulates that veteran fraternity members in hazing chapters love their pledges even if they put them through beatings with canes, fists, and paddles. While not every African-American Greek chapter hazes, the heaviest hazing chapters, according to Jones, demand the ability for pledges to withstand such sacrifices to demonstrate their readiness and willingness to maintain the group. Significantly, physical hazing at the chapter level is fundamentally the exact opposite of the relatively tame requirements for initiation that the national organizations put out in pledging manuals. Jones persuasively argues that hazing deaths in Asian fraternities at the University of Texas and University of California, Irvine, for example, demonstrate a requirement to endure physical challenges and a demonstration of sacrifice on the part of initiates. While the physical hazing may be a part of the black membership process in groups that violate their national organization's mandate to be hazing-free, it is also very much present in the rites of even some integrated chapters or all-Caucasian chapters that have seen new members hospitalized with physical injuries. Jones, himself a member of a black fraternity, has become a nationally known antihazing activist, particularly outspoken in his objections to hazing of a severe physical nature.

Psychologist Susan Lipkins, in her book *Preventing Hazing: How Parents, Teachers, and Coaches Can Stop the Violence, Harassment, and Humiliation* (2006), maintains that even short-term hazing can lead to long-term, even lifelong, psychological, and physical consequences for hazing victims. Short of death, psychological trauma due to hazing may be an even more pernicious problem than physical injuries for the latter nearly always heal, while victims report an inability to get on with their lives in normal fashion due to the afflictions of hazing, according to Lipkins.

Hazing in U.S. Universities

Hazing in fraternities and sororities in the United States involves expectations of senior members that potential new members – often called pledges or associate members – will perform stunts that demean, degrade, or endanger them. Those that endure the group's humiliation prove their willingness to conform to a perceived status quo. The members who insist newcomers endure such barbaric initiations claim they are merely protecting established tradition and ensuring group solidarity. They maintain that hazing weeds out potential members who lack the wherewithal to do anything the fraternal chapter asks of those who seek membership. It also establishes a senior pecking order in which new members learn the rules of precedence and a need to respect both senior members and alumni as well.

Thus, hazing is common, albeit forbidden, among fraternities and sororities whose reason for being at educational institutions is to instill a lifelong commitment to leadership, community service, and the worth of lifelong friendships and mentoring. This general 'mission' coincides with functionalist theories of rites of passage extending back to Arnold van Gennep (*Les rites de passage* (*The Rites of Passage*), 1909).

Hazing in university and college fraternities and sororities is an illicit rite of passage (against the rules of the institution) that provides opportunities for veteran members to include or exclude new members on the basis of a variety of trials that involve varying degrees of physical and psychological stress. In all too many cases, the theory of social exchange put forth by psychologists Harold H. Kelley and John W. Thibaut in *The Social Psychology of Groups* (Tiger, 1969) describes the recurring reality. According to this theory, newcomers submit to whatever torture or trials are asked of them, comforted by the assurance that all abuse will cease after the organization declares them full members. During the next cycle, the new veterans get to turn the tables and haze the next set of newcomers.

All new pledge classes experience much the same brutal hazing their predecessors endured. Hazing can increase in intensity if members are inebriated or inclined toward sadism (Nuwer, 1990). After initiation, the hazed become the hazers and return the ordeals they experienced in equal or greater intensity and measure to the next pledge class (Nuwer, 1990, 1999a).

Deaths of pledges in Greek organizations have occurred under bizarre circumstances as the database of hazing deaths kept by Nuwer in *Wrongs of Passage* and online (http://www.hanknuwer.com/hazingdeaths.html) attests. Some have consumed lethal amounts of alcohol, being required by members to fill garbage pails with the collective vomit of new members. Some have perished after being dropped off far from campus (with victims dying in falls, auto accidents, and by drowning). Some have endured hours of strenuous exercise in steam rooms with fatal results. Some have died after being on the receiving end of beatings with a paddle, fists, or cane, according to Nuwer's list of hazing deaths.

History of University Hazing in the United States

In his 1990 book *Broken Pledges: The Deadly Rite of Hazing*, Nuwer demonstrates that the rampant hazing in European medieval universities such as the fifteenth-century practice of pennalism developed over the next 200 years into sometimes savage hazing practices quite similar to the aptly named Hell Nights of today's hazing fraternities and sororities. In France and Germany, the hazed pledges wore a cap similar to the beanies worn by pledges until all national groups began outlawing them as symbols of hazing around 1979, according to Nuwer's *Broken Pledges: The Deadly Rite of Hazing*. At Cambridge and Oxford in England, physical beatings and enforced servitude among even the aristocracy was carried out in a custom known as fagging where new boys on campus became the manservants of older students.

Hazing deaths were not always attributed mainly to fraternities historically. Rather, they were a greater problem in the so-called 'class hazing' that pitted first-year students against upperclassmen. (While female class hazing existed, not once did a death occur to either a first- or second-year female, but a female cook did perish in 1894 when an undergraduate piped chlorine gas into a class party to disrupt it.) Eighteen of the 25 hazing deaths of males that occurred on U.S. college campuses from 1838 through 1927 were a result of freshman-sophomore class hazing excesses during battle royals, pranks, and organized fights to capture class flags, but not because of fraternity hazing, which claimed the lives of seven young men, according to Nuwer's *Broken Pledges* appendix of deaths. During the late 1920s, thanks to more awareness on campus and an outpouring of condemnation of hazing by students themselves, hazing of first-year students no longer included battles for class flags and class pennants that had seen so much violence. From 1928 through March 2014, only two first-year hazing deaths (Cheyney State and the University of Richmond) occurred, the rest entirely committed by hazers in fraternities, sororities, athletic teams, and one band (Florida A & M), according to Nuwer's list of hazing deaths and his appendix in *Wrongs of Passage*.

Fraternities and sororities often trace their origins back to literary and honor societies that flourished back in 1776. The first fraternity per se was the Kappa Alpha Society founded in 1825 at Union College in New York State. The first fraternities and sororities emulated early literary societies in their educational values, and their founders were students themselves who wanted to engage in intellectual debate as well as to socialize. In addition to establishing the solemn initiation rituals that often borrowed symbolism from Masonic initiations, the early fraternity chapters also hazed new members on occasion in usually ordeals known as stunts. Some stunts included riding a real or mechanical goat or tossing a newcomer in the air on a blanket. Injuries occasionally occurred as a result of youths getting carried away. Nuwer's *Broken Pledges* documents injuries to two University of Georgia Zeta Chi pledges that jumped out a window when an older member whipped out a pistol as a prank.

The first U.S. fraternity death was Kappa Sigma Society member Mortimer Leggett in 1873 at Cornell University. At least one newcomer has perished while enduring hazing every year from 1970 to 2013, according to Nuwer's *Wrongs of Passage* and the Internet list of verifiable hazing deaths. The last one was an alcohol-related death at Baruch College in New York State in December 2013 (http://www.hanknuwer.com/hazingdeaths.html), one of five hazing deaths in 2013.

Thus, some years far more deaths than a single death have occurred in Greek organizations. In 2012, five deaths of fraternity men in U.S. colleges occurred by hazing that stemmed from alcohol overdoses; a sixth death occurred by drowning: a male pledge was asked to swim in a reservoir at the University of Idaho (http://hanknuwer.com/hazingdeaths.html). Because some law enforcement officers lack a clear understanding of what hazing actually is, it is plausible that the number of hazing deaths on the Nuwer's list is lower than might actually be the higher total. Similarly, neither universities nor Greek national organizations are eager to claim a death as related to hazing lest that admission result in damages during civil litigation.

Hazing deaths are not limited to on campus to fraternities and sororities, but those two groups have experienced the majority of incidents. In the United States, the vast majority of the 180 hazing deaths from 1838 to 2013 have involved newcomers in fraternities and, to a lesser extent, in sororities. Far more males than females have perished in Greek organizations because of hazing, and experts uniformly agree that fraternity hazing deaths have proven far more common than sorority hazing deaths. According to a University of Maine national survey, about half of all students acknowledge that they belong to groups that haze. Thus, while hazing in sororities is often reported, it only occasionally rises to the level of an actual crime and is more likely to be addressed by the hazers' institution than by a criminal prosecution. It is important to note that deaths in local, unrecognized, or suspended fraternities have resulted in deaths at several institutions, including the University of Nevada, SUNY Plattsburgh, Chico State, and SUNY Geneseo.

Deaths by hazing also have been reported in countries such as the Philippines, India, Japan, Canada, and Great Britain (Nuwer, 2004).

Hazing and Alcohol Abuse

Alcohol was consumed by some fraternal chapters in the nineteenth century through 1939, but such factors as alumni, faculty, and administration supervision of chapters resulted in far less tolerance for drunkenness or using alcohol as a sort of litmus test of new member readiness during hazing. Significantly, fraternities likely did not experience a single reported

hazing death due to alcohol until 1940. That year, Hubert L. Spake, Jr. died following a mandatory drinking session at the University of Missouri while pledging Theta Nu Epsilon, according to Nuwer (1990). Clearly, with the escalating number of hazing deaths from 1970 to December 2013, according to Nuwer's list of deaths, a case can be made that the presence of alcohol has made hazing rituals far more extreme and life threatening than they were prior to 1969.

Research published in Nuwer's online list of hazing deaths demonstrates that the majority of hazing deaths in Greek groups from 1970 to 2013 involve an overdose of alcohol. In the period 2012–13, six of the last eight fraternity hazing deaths involved alcohol. Other victims from 1873 to 2012 have died from beatings, car wrecks, pedestrian accidents, falls from buildings or high elevations, burns, drowning, choking, ingestion of many gallons of water or other liquids, and other causes. A small number of suicides have occurred immediately following hazing activities and are listed on Nuwer's list of hazing deaths.

Here are three examples of hazing deaths due to alcohol from the Internet 'List of Hazing Deaths' maintained by Hank Nuwer.

- In 1980, University of South Carolina Sigma Nu pledge Barry Ballou choked to death after passing out at a ritualized drinking session attended by an alumnus and members.
- Pi Kappa Alpha pledge David Bogenberger, a Northern Illinois University freshman, died in 2012 after being asked to chug copious amounts of hard alcohol. Police charged 22 individuals with crimes, including five felony counts.
- In 2012, Fresno State University Theta Chi pledge Philip Dhanens, a 350-pound former football player, died following a weekend binge. He died at a hospital where he had been taken for assistance.

U.S. Laws against Hazing

Hazing may be illicit – against the rules of a collegiate institution – or forbidden by state law. As of 2012, hazing was illegal by statute in 44 states, but as of March 2014, there is no federal law regulating hazing; although, activists such as Lianne Kowiak, mother of a Theta Chi pledge killed at Lenoir Rhyne University, advocates for such a law, according to Bloomberg.com (24 July 2013). In most states, hazing is most commonly a misdemeanor offense. Florida has the strictest possible penalty among all U.S. states, allowing a sentence of 5 years to be imposed upon conviction for felony hazing. Nonetheless, the courts have traditionally imposed light jail sentences, often requiring merely community service or fining individual hazers and/or their chapters. Hazing can be hard to establish since these illegal actions often take place behind closed doors. In addition, after a death or serious incident, uncooperative members refuse to share details of what took place when investigators interview them. This is why prosecutors for the state often file more easily proven charges such as assault or serving alcohol to a minor because hazing is a more difficult crime to prove in court. Felony convictions for repeated hazing beatings were imposed on a Florida A & M fraternity, and a Chico State fraternity member was convicted of a felony following the death of a pledge required to drink so much water than his body chemistry was altered.

University Hazing Practices and Deaths

The increasing numbers of prospective members injured or killed by hazing reveals that they stand ready to do whatever is necessary for acceptance into an organization. When a fraternity or sorority pledge fails to complete the pledge process, it can be a traumatic experience for individuals who fail to cross the liminal space, particularly if they feel less than full-fledged adults and must ask either their parents or school officials to intervene on their behalf.

Sororities have been connected to eight deaths of females and one death of a male in alleged hazing incidents, according to Nuwer's list of hazing deaths. In 1970, an Eastern Illinois sorority member became the first to die after she perished from effects of a head injury incurred in a vehicular accident after she had resisted being 'kidnapped' by pledges that intended to abandon her in the country as a prank, according to Nuwer's *Wrongs of Passage*. Three female deaths have been attributed to drowning in dangerous ceremonial initiations, one at Virginia State and two at California State, Los Angeles, according to Nuwer's list.

A 2012 lawsuit by the mother of a deceased sorority pledge at East Carolina University maintained that the 2010 deaths of her daughter and a second pledge were directly caused by sleep deprivation due to hazing. According to that lawsuit, East Carolina State University Delta Sigma Theta pledges Victoria Carter, 20, and Briana Latrice Gather, 20, died in a car accident. In addition to the aforementioned two sorority deaths at Eastern Carolina University, one additional death occurred at Plymouth State University during an automobile accident while pledges allegedly were made to lie flat on the floor of an SUV that then overturned. None of these deaths involved alcohol. However, in 2008, a total of 12 Chi Omega sorority and Sigma Nu fraternity members at Utah State University were charged with crimes following the alcohol-related hazing of Michael Starks, 18, conducted by both chapters, according to Nuwer's list. In addition, in March 2014, an alleged hazing incident involving Kappa Kappa Gamma sorority and Sigma Alpha Epsilon (SAE) fraternity at the University of Connecticut resulted in the hospitalization of a sorority member admitted with a life-threatening blood-alcohol account, according to widely published media accounts and an email from the victim to Hank Nuwer sent on 12 March 2014.

In addition to pledge deaths, the so-called pledge 'sneaks' routinely occur in which pledges haze senior members to retaliate for what they themselves have gone through and endured, and these have resulted in deaths at schools such as Eastern Illinois University, the University of Texas, University of Georgia, and Frostburg State University, according to the appendix listed in Nuwer's *Wrongs of Passage* (1999b). Such a practice is often encouraged by the membership as a means of forging pledge solidarity, according to the same source. The latest example of the reverse hazing of members by pledges has been the death of SAE member George Desdunes who was 'encouraged' by pledges to drink what proved to be a deadly amount of alcohol in a question-and-answer session at his

Cornell University chapter house in 2011, according to *The New York Times* of 12 April 2012.

As mentioned, hazing can also have serious consequences even without a death. Many male and female students have been hospitalized for acute alcohol intoxication, for example. Greek chapters have been punished for such behaviors as activities of a sexual nature, physical pummeling, verbal abuse, overwork, sleep deprivation, branding with chemicals, and a host of bizarre pranks intended to amuse veteran members, according to the list of examples published in the appendices of Nuwer's *Broken Pledges* and *Wrongs of Passage*.

To be sure, hazing can also involve activities that do not qualify as deadly or particularly demeaning such as requiring a pledge to sing a song or carry a veteran's book. Nonetheless all national and international fraternal organizations ban these activities as well since the possibility exists that a rogue member or members might get creative with a relatively noninjurious act of hazing and take the activity to a potentially harmful level. In addition, what may seem like innocuous hazing to a perpetrator may be perceived as threatening to a newcomer. Fraternal leaders such as Associate Dean of Students Travis Apgar of Cornell University (http://www.hazingprevention.org/hazing-information/hidden-harm-of-hazing.html) refer to resultant psychological trauma as a type of 'hidden harm.' Even the most stable of new recruits can break down when subjected to night after night of verbal abuse, subjugation, and sleep deprivation, particularly when forced to sleep on a floor in confined quarters and to leap up on the instant to carry out some inane errand that a veteran member has conjured up. Nuwer, in the *Chronicle of Higher Education* (26 November 1999a), termed these latter examples as examples of cultlike systematic hazing.

Prosecuting and Preventing Hazing

From the first fraternity hazing death in 1873 to deaths in 1973, there appear to have been few arrests and no convictions for fraternity and sorority hazing deaths, according to Nuwer's *Broken Pledges* (1990: pp. 286–298). And while the adviser in the 1974 shooting death for a hazing at Bluefield State was convicted and served time for killing a Tau Kappa Epsilon pledge, and a North Carolina A & T University in 1987 received a 2-year sentence for pledge beatings, it was not until 1999 that tightened state laws and increasing media attention led to stricter enforcement and a small number of arrests resulting in convictions. Most notable was the 2007 sentencing of nearly 2 years in prison for two Florida A & M fraternity men who pleaded no contest to felony charges of beating a Kappa Alpha Psi pledge; they served jail time but on appeal had the charges expunged from their records.

By 2010, more arrests and a handful of convictions for serious hazing cases began to get the attention of the press. For example, two young men were convicted of misdemeanor hazing in the 2008 death of SAE pledge Carson Starkey and received jail sentences of 30 days each, according to widely published media accounts such as CBS Sunday Morning (5 February 2012). On the other hand, in the 2010 death of a Radford University fraternity pledge, the court punishment was merely $1000 each and no jail time for five Tau Kappa Epsilon members in a plea deal, according to the *Orlando Sentinel* (11 July 2012).

Perhaps more effective than laws curtailing hazing have been attempts by international organization such as SAE to ban or shorten the pledge period. In March 2014, according to the Web site of the parent SAE organization, all pledging now has ceased due to the deaths of 10 SAE pledges, the most of any national fraternity according to Bloomberg News (8 March 2014) and Nuwer's Internet list of hazing deaths (15 March 2014).

Another effective best practice was the dry house movement followed most conspicuously by all chapters of the large national fraternity Phi Delta Theta that in 2000 removed alcohol from the premises of chapter houses, thereby eliminating the substance that has caused the most hazing deaths after 1970, according to Nuwer (2001). On 14 March 2014, executive Vice President of Phi Delta Theta, Bob Biggs told the *Washington Post* that in addition to a dramatic decrease in hazing claims, undergraduate membership increased by 25%.

Hazing Judgments by the Courts

Civil lawsuits in the aftermath of an injury or death blamed on hazing have become quite common. The parents of Chad Meredith, a University of Miami pledge who drowned when forced to swim across a lake by older Kappa Sigma members, resulted in a $14 million judgment against those young men. The highest settlement ever paid by a university was $6 million in 2002 by Massachusetts Institute of Technology to the parents of Scott Krueger who died of alcohol intoxication following a Phi Gamma Delta pledge party.

Prevention of Hazing

As might be expected, since deaths by hazing largely occur in fraternities and to a lesser extent in sororities, the main anti-hazing activists are undergraduate members, alumni, and national officers of fraternities and sororities. The two most influential organizations are the Web-based Stophazing.org created by University of Maine hazing researcher Elizabeth Allan and the national group HazingPrevention.Org (HPO).

Stophazing's Allan is a professor at the University of Maine in Higher Education Leadership, one of the few academic programs offering a hazing studies concentration. She and her Maine colleague Mary Madden completed in 2011 a 3-year research project on hazing with responses from 11 482 college students enrolled in 54 colleges. The survey response rate was 12%.

Slightly more than half of all respondents acknowledged participating in acts of hazing as members of student groups such as fraternities and sororities. In her essay for *The Hazing Reader* titled 'Hazing and Gender: Analyzing the Obvious,' Allan notes that empirical research on gender differences is scant, but while a national survey distributed to Greek professionals found that 44% had to deal with reported hazing cases, the number of violent and otherwise physical hazings were far fewer than those committed by male students.

HPO is an activist organization dedicated to empower people to prevent hazing in college and university student groups. HPO's board of directors and numerous volunteers and small staff provide antihazing education through conferences, Webinars, and training sessions. HPO sponsors its annual

Hank Nuwer Anti-Hazing Hero Award, which to date has honored mainly fraternity and sorority undergraduates who effected positive changes in their own chapters to curb hazing practices. The award was named for the author of this article to commemorate his research on hazing that has resulted in four scholarly books on the topic through 2012 (*Broken Pledges: The Deadly Rite of Hazing, High School Hazing, Wrongs of Passage, The Hazing Reader*). In January 2014, Nuwer published *Sons of the Dawn: A Basque Odyssey*, a novel that Kirkus Reviews praised online (10 January 2014) for its treatment of hazing and bullying in the Old West.

In addition to these two antihazing groups, there is a college archive that attempts to put all hazing scholarship under a single collection. Buffalo State College's (BSC) Butler Library has since 2006 sponsored an extensive Hazing Collection available without charge to hazing researchers and undergraduate students alike (http://bsclib01.buffalostate.edu/library/archives/). Under the direction of archivist Daniel Di Landro and BSC Vice President Maryruth Glogowski, the Hazing Collection in Buffalo, New York, has endeavored to house every hazing-related scholarly article, doctoral dissertation, master's thesis, book, and miscellaneous research items for free use by researchers. Before the collection was established, one difficulty that researchers face was that the relatively small literature related to hazing scattered at widely separated libraries and institutions, making research an expensive, time-consuming challenge for researchers, according to author John Nicoletti in *Violence Goes to College* (2009: p. 200).

While a number of best practices to curtail hazing have been championed by a number of educators, no single solution has been found to stamp out hazing. Nuwer, in *Broken Pledges* (1990), along with other authors, has argued that bystander intervention might have prevented specific hazing incidents from leading to serious injuries and deaths in a number of fraternity tragedies. Bystander intervention refers to the practice of emboldening those who witness crimes or potential criminal behavior to step in and take action to halt the behavior before it escalates, according to *The New York Times* of 9 February 2014. The writings and seminars by hazing intervention expert Alan Berkowitz very specifically target bystander intervention as an important means of intervention before a hazing activity gets to the point of no return. In addition, a great many U.S. educational institutions use social norms data in an attempt to discourage binge drinking and hazing by stressing to undergraduates that such unacceptable practices are anything but normal in an attempt to sway behavior changes through marketing campaigns such as posters in college classroom buildings and residence halls.

Many other education experts, including Christopher Bollinger, coauthor of *Violence Goes to College*, have argued for the value of substituting useful, benign group-building activities in place of hazing practices. HPO seminar leader Kim Novak offers Greek life professionals hands-on suggestions for activities such as ropes courses that encourage group unity to achieve a common goal.

In 2013, the VIV Family Outreach Foundation, founded by the surviving families of a 2007 Virginia Tech (Virginia Polytechnic Institute and State University) school shooting massacre, brought together nine experts (including in hazing prevention) to create a new project termed the 32 National Campus Safety Initiative that has identified best practices for institutions to employ that counter violent campus behaviors such as hazing and sexual assault. A pilot program was initiated with a number of institutions to provide a voluntary rating system so that an individual institution not only can judge how it measures up in terms of combating hazing and other behavioral maladies, but also take steps to shore up areas where deficiencies appear to occur.

As the number of deaths associated with hazing continues to climb, a trend of late has been the zero tolerance toward the practice practiced by a number of institutions. In addition to the expulsion of chapters caught hazing, a number of universities also have created hazing task forces composed of faculty, staff and, in some cases, outside experts.

For example, following the hazing death of George Desdunes, 19, Cornell University President David Skorton not only wrote a widely quoted Op-Ed article for *The New York Times* (23 August 2011) titled 'A Pledge to End Fraternity Hazing,' but his Greek Affairs staff at the venerable New York institution issued a clear warning to fraternal groups that the new hazing policy was zero tolerance. The no-nonsense new policy resulted in nine fraternities cited and/or punished by Cornell for hazing from 2012 to March 2014, according to that institution's Web site. In addition, Cornell established the Recruitment-Acceptance-Retention-Education Task Force with one of its main agenda items being to find and to research best practices for curtailing hazing in fraternities and other student groups.

In California, following the alcohol-related birthday death of a Chico State University pledge, which its President Paul Zingg termed hazing-related, all campus fraternities and sororities were suspended in November 2012. The unprecedented step was taken because of a rash of nonhazing-related deaths of late, as well as the fact that this was at least the fourth hazing death at Chico State. Likewise, Central Florida University, which had a near fatality during pledging for Kappa Alpha Psi, put an end to all pledging and fraternal social activities in 2013, concerned by reports that pledging activities had in some cases escalated into hazing. The University of Virginia ordered all pledging to cease immediately in 2013 or the offending chapters would be shut down. The University also started the antihazing activist group called Hoos Against Hazing (http://www.virginia.edu/hazing/).

If these hard-line steps fail to work, no doubt a small number of universities will follow the example of Alfred University. In 2002, following a number of hazing deaths in local and national fraternities, the Alfred Board of Trustees decided regulation of Greek groups was not working. Consequently, the Board moved to shut down all Greek houses permanently. However, even though media outlets such as the Bloomberg News (Editorial, 7 January 2014) clamor for colleges to shut down permanently all fraternal systems, Alfred was the last to end a long-standing Greek system, and it unlikely many more colleges will risk alumni outrage by banning fraternities and sororities. National fraternities and sororities vociferously defend the right of public universities to maintain a freedom of association, making it unlikely that any educational institutions (besides the occasional private college) will ever ban an entire Greek system. At present, besides Alfred, the handful of private colleges in the historic past to close their Greek systems

Box 1 Chronology of fraternity and sorority hazing in the United States

1873 Cornell first-year student Mortimer N. Leggett died in a fall into a gorge while members were taking him on a traditional walk in the country.

1928 National Interfraternity Conference leaders voted to end all hazing.

1940 The first alcohol-related fraternity hazing death occurred in the United States. It was at the University of Missouri.

1970 A female Eastern Illinois sorority member died in a vehicle accident as pledges tried to abandon her in the country. There would be a death every year after (and some years many deaths) in Greek life until 2013.

1978 A death of a pledge named Chuck Stenzel at Alfred University marked the beginning of the antihazing movement in the United States led by Eileen Stevens, a New York housewife and the pledge's mother. She called her organization CHUCK, the Committee to Halt Useless College Killings.

1978 The movie Animal House was released, mocking the practice of fraternity hazing.

1994 The beating death of Michael Davis for a Southeast Missouri State black fraternity chapter created outrage over physical hazing. He died in the back seat of a member's car after being beaten by members who had abducted him from the library.

2000 As deaths mount, the National Interfraternity Conference and National Panhellenic Conference plan measures to control hazing and binge drinking. Phi Delta Theta officially ordered its chapter houses to become alcohol-free in an attempt to end hazing and binge drinking deaths.

2002 Hank Nuwer, the author of *Wrongs of Passage*, issues a challenge to fraternity and sorority leaders to begin a hazing task force to address the issue nationally. Subsequently, a fraternity antihazing activist named Dan Bureau takes the challenge and convenes the first-ever task force on hazing.

2013 A hazing death in December at Baruch College in New York extended the unbroken stretch of hazing deaths 43 years from 1970 to 2013.

2014 Following the deaths of 10 pledges, the largest national fraternity in the world, Sigma Alpha Epsilon, bans all pledging.

included Middlebury College (1990) (1990: p. 147), Williams College (1962) (1990: p. 242), and Colby College (1983) (1990: p. 242). A few private colleges such as Bates College in Maine never have permitted fraternities. The last public university to ban pledging for a semester was the University of Alabama in November 2012, according to *Fraternal Law Newsletter* (**Box 1**).

See also: Adolescence, Sociology of; Conflict: Education and Youth; Disposable Youth in America in the Age of Neoliberalism; Gender Differences in Personality and Social Behavior; Masculinity, Scripts of; Personal Identity: Philosophical Aspects; Responsibility: Philosophical Aspects; Social Competence During Adolescence Across Cultures; Social Identity in Sociology; Teens, Gender, and Self-Presentation in Social Media; Transition to Adulthood; Youth Culture, Anthropology of; Youth Culture, Sociology of.

Bibliography

Allan, E., Madden, M., March 2012. The nature and extent of college student hazing. International Journal of Adolescent Medicine and Health 24 (1), 83–90.

Allen, E., 2004. Hazing and gender: analyzing the obvious. In: Nuwer, H. (Ed.), The Hazing Reader. Indiana University Press, Bloomington, IN, pp. 275–294.

Aronson, E., Mills, J., September 1959. The effect of severity of initiation on liking for a group. Journal of Abnormal and Social Psychology 59 (2), 177–181.

Arnold, J., 1995. Alcohol and the Chosen Few: Organizational Reproduction in an Addictive System (Ph.D. dissertation), Indiana University, Bloomington, IN. Reprinted in condensed form as 'Hazing and alcohol in a college fraternity'. In: Nuwer, H. (Ed.), The Hazing Reader. Indiana University Press, Bloomington, IN, 2004.

Berkowitz, A., 2009. Response Ability: A Complete Guide to Bystander Intervention. Beck & Company, Chicago.

Bollinger, C., Nuwer, H., 2009. Hazing. In: Nicolletti, J., et al. (Eds.), Violence Goes to College. Charles C. Thomas (Publisher), Springfield, IL, pp. 185–203.

Blumer, H., 1969. Symbolic Interactionism. Prentice Hall, Englewood Cliffs, NJ.

Cimino, A., 2013. Predictors of hazing motivation in a representative sample of the United States. Evolution and Human Behavior 34, 446–452.

DeSantis, A., 2007. Inside Greek U.: Fraternities, Sororities and the Pursuit of Pleasure. University of Kentucky Press, Lexington, KY.

Flanagan, C., 14 March 2014. What Sigma Alpha Epsilon's Pledging Ban Won't Fix. Washington Post.

Girard, R., 1977. Violence and the Sacred. The Johns Hopkins University Press, Baltimore.

Janis, I., 1972. Victims of Groupthink. Houghton Mifflin, Boston.

Jones, R., 2004. Black Haze: Violence, Sacrifice, and Manhood in Black Greek-Letter Fraternities. State University of New York Press, Albany, NY. Revised as an essay as 'Examining violence in black fraternity pledging'. In: Nuwer, H. (Ed.), The Hazing Reader. Indiana University Press, Bloomington, IN, 2004.

Kelly, H., Thibaut, J., 1969. The Social Psychology of Groups. Wiley, New York.

Kimmel, M., 2008. Guyland: The Perilous World Where Boys Become Men. HarperCollins, New York.

Leemon, T., 1972. The Rites of Passage in a Student Culture. Teachers College Press, New York.

Lipkins, S., 2006. Preventing Hazing: How Parents, Teachers, and Coaches Can Stop the Violence, Harassment, and Humiliation. Jossey-Bass, San Francisco.

McCarthy, D., November 2012. Alabama suspends all pledging. Fraternal Law 122.

Nuwer, H., October 1978. Dead souls of hell week. Human Behavior 7 (10), 50–56.

Nuwer, H., 26 November 1999a. Greek letters don't justify cult-like hazing of pledges. Chronicle of Higher Education.

Nuwer, H., 1990. Broken Pledges: The Deadly Rite of Hazing. Longstreet Press, Atlanta.

Nuwer, H., 1999b. Revised 2001. Wrongs of Passage: Fraternities, Sororities, Hazing and Binge Drinking. Indiana University Press, Bloomington, IN.

Nuwer, H., 2001. Wrongs of Passage, second ed. Indiana University Press, Bloomington, IN.

Nuwer, H., 2004. Exterminating the frat rats. In: Nuwer, H. (Ed.), The Hazing Reader. Indiana University Press, Bloomington, IN, pp. XIII–XXIV.

Olmert, M., September 1983. Points of origin. Smithsonian, 150–154.

Rashdall, H., 1895. Reprint 1936. The Universities of Europe in the Middle Ages, 2 vols. Oxford University Press, Oxford.

Robbins, A., 2005. Pledged: The Secret Life of Sororities. Hyperion, New York.

Singer, M., Lalich, J., 1995. Cults in Our Midst. Jossey-Bass, San Francisco.

Skorton, D., 23 August 2011. A Pledge to End Fraternity Hazing. The New York Times.

Sweet, S., 2001. College and Society: An Introduction to the Sociological Imagination. Pearson, Upper Saddle River, NJ. Revised as an essay as 'Understanding fraternity hazing'. In: Nuwer, H. (Ed.), The Hazing Reader. Indiana University Press, Bloomington, IN, 2004.

Tiger, L., 1969. Men in Groups. Random House, New York. Revised as an essay as 'Males courting males'. In: Nuwer, H. (Ed.), The Hazing Reader. Indiana University Press, Bloomington, IN, 2004.

Van Gennep, A., 1908. The Rites of Passage. Reprint 1960. University of Chicago Press, Chicago.

Van Maanen Van Maanen, J., Schein, E., 1979. Toward a theory of organizational socialization. Research in Organizational Behavior 1, 209–264.

Relevant Websites

http://stophazing.org – Stop Hazing.
http://hazingprevention.org – Hazing Prevention.
http://hanknuwer.com/blog – Hank Nuwer Blog.
http://hanknuwer.com – Hank Nuwer.
http://umaine.edu/hazingresearch/ – The University of Maine Hazing Research and Prevention.

Healing Practices: Alternative and Complementary

Adrian Furnham, University College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes the rise in complementary and alternative medicine (CAM). It also looks at the typical user. It describes the range of treatments available and how they differ from orthodox medicine (OM). This article considers some of the controversies in this area specifically which (and why) people choose CAM over OM. Secondly, it looks at the evidence for the efficacy of CAM.

The Rise of Complementary and Alternative Medicine and Healing Practices

Complementary and alternative medicine (CAM) has gained wide recognition and CAM is now big business (Ernst and Furnham, 2000). CAM seems also to be favorably perceived by many general practitioners (GPs) (Easthope et al., 2000). The rise of CAM has led to a UK House of Lords inquiry into six aspects of CAM: evidence, information, research, training, regulation and risk, and British National Health Service (NHS) provision (Ernst, 2000).

The figures of before the millennium showed the beginning of this rise. In 1993, 34% of all Americans visited a CAM therapist, more than visits to all US primary care physicians put together. Expenditure was estimated at $13.7 billion per annum. By 1998, 47% of all Americans were estimated to visit a CAM practitioner. Annual visits rose from 427 million in 1990 to 629 million in 1997 (Eisenberg et al., 1998).

Individuals spend £1.6 million per annum on CAM therapies, the NHS about £40, and £500 million is spent on products (White and Ernst, 2000). Some estimates suggested that this number has doubled but it is difficult to get reliable statistics.

There are many reports of how much people spent (Nahin et al., 2009).

A 2012 *Journal of the American Medical Association* report noted that in 2012, £130 million was spent on CAM research (Offit, 2012).

All this attests to the increasing popularity of CAM and presumably the belief by the general public that they are getting value for money, in the sense that it 'works for them.'

Controversy

Since the turn of the twenty-first century, there has been unprecedented and sour debate as to the efficacy of alternative medicine (Vincent and Furnham, 2001). Complementary medicine has been a wildly controversial topic over recent years, as techniques have been rigorously scientifically tested and various fraudsters revealed. Singh and Ernst (2008) in the book *Trick or Treatment* evaluated 40 complementary techniques and were scathing in their findings. They found that scientific evidence for the efficacy of these methods was lacking, and that the methods were not just ineffective, but positively dangerous in some cases. For example, methods such as homeopathy are labeled merely as placebos lacking in credible, repeatable effects. Singh was later sued by the British Chiropractic Association for libel for comments associated to the book, although they later dropped the case. Other widely available books such as *Suckers: How Alternative Medicine Makes Fools of Us All* by Shapiro (2008) follow a similar line.

However, many CAM practitioners are wary of, and hostile to, the scientific testing of their claims (Leach and Gillham, 2011). An Australian study found that, although 92% of practitioners believed evidence-based, scientific practice was useful, only a small amount actually practiced it in the correct manner. Their suggestion is that, to push through this inertia, training of clinicians is necessary to reinforce how evidence-based practice is useful and works. Stoneman et al. (2013), however, see the problem being in the public consciousness and support of alternative medicines and therapies, saying that this is the most difficult thing to change.

Critics have lampooned CAM. A good example is that of Prof. David Colquhoun (http://www.dcscience.net/, http://www.ucl.ac.uk/npp/research/dc) who has a Web site dedicated to CAM quackery and fraud. In one entry, he gives the following funny and cynical definitions:

- Homeopathy: giving patients medicines that contain no medicine whatsoever.
- Herbal medicine: giving patients an unknown dose of an ill-defined drug, of unknown effectiveness and unknown safety.
- Acupuncture: a rather theatrical placebo, with no real therapeutic benefit in most if not all cases.
- Chiropractic: an invention of a nineteenth-century salesman, based on nonsensical principles, and shown to be no more effective than other manipulative therapies, but less safe.
- Reflexology: plain old foot massage, overlaid with utter nonsense about nonexistent connections between your feet and your thyroid gland.
- Nutritional therapy: self-styled 'nutritionists' making untrue claims about diet in order to sell you unnecessary supplements.
- Spiritual healing: tea and sympathy, accompanied by arm-waving.
- Reiki: ditto.
- Angelic Reiki: The same but with added "Angels, Ascended Masters and Galactic Healers." Excellent for advanced fantasists.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.14004-8

- Colonic irrigation: a rectal obsession that fails to rid you of toxins that you did not have in the first place.
- Anthroposophical medicine: followers of the mystic barmpot, Rudolf Steiner, for whom nothing whatsoever seems to strain credulity.
- Alternative diagnosis: kinesiology, iridology, vega test, etc., various forms of fraud, designed to sell you cures that do not work for problems you have not got.
- Any alternative 'therapist' who claims to cure acquired immunodeficiency syndrome or malaria: agent of culpable homicide.
- Libel: A very expensive remedy, to be used only when you have no evidence. Appeals to alternative practitioners because truth is irrelevant.

What Are Complementary and Alternative Medicines?

The great range of CAM inevitably means there is a considerable diversity of therapies and their attendant theories and philosophies. Yet there are common themes in the philosophies of CAM. Aakster (1986) believes that they differ from orthodox medicine (OM) in five ways:

1. Health: Conventional medicine sees health as an absence of disease, whereas alternative medicine frequently mentions a balance of opposing forces (both external and internal).
2. Disease: Conventional medical professionals see disease as a specific, locally defined deviation in organ or tissue structure. CAM practitioners stress body-wide signs, such as body language indicating disruptive forces and/or restorative processes.
3. Diagnosis: Regular medicine stresses morphological classification based on location and etiology, while alternative interpretations often consider problems of functionality to be diagnostically useful.
4. Therapy: Conventional medicine often claims to destroy, demolish, or suppress the sickening forces, while alternative therapies often aim to strengthen the vitalizing, health-promoting forces. CAM therapists seem particularly hostile to chemical therapies and surgery.
5. The Patient: In much conventional medicine, the patient is the passive recipient of external solutions, while in CAM, the patient is an active participant in regaining health.

Aakster described three main models of medical thinking: The pharmaceutical model is a demonstrable deviation of function or structure than can be diagnosed by careful observation. The causes of disease are mainly germlike and the application of therapeutic technology is all-important. The integrational model resulted from technicians attempting to reintegrate the body. This approach is not afraid of allowing for psychological and social causes to be specified in the etiology of illness. The third model has been labeled holistic and does not distinguish between soma, psyche, and social. It stresses total therapy and holds up the idea of a natural way of living.

Furnham (2000) used factor analysis to see how over 500 members of the general British public classified 39 different types of CAM on whether they had heard of it, know how it worked, whether they had tried it, and whether they believed it worked. A pattern emerged with art therapies (e.g., music, dance), talk therapies (i.e., counseling), and 'foreign techniques' (e.g., Reiki, Shiatsu). The 'big six' most established therapies (acupuncture, chiropractic, homeopathy, medical herbalism, naturopathy, and osteopathy) were often grouped together, presumably because they are seen as most established and regulated, despite the fact that they are based on very different methods and philosophies. In essence, he showed that some CAM are very well known and trusted while others are very little known.

Gray (1998) argued there are currently four quite different perspectives on CAM:

1. The biomedical perspective: This is concerned with curing of disease and control of symptoms where the physician-scientist is a technician applying high-level skills to the patient. This perspective asserts (1) that the natural order is autonomous from human consciousness, culture, morality, psychology, and the supernatural; (2) that truth or reality resides in the accurate explanation of material (as opposed to spiritual, psychological, or political) reality; (3) that the individual is the social unit of primary importance (as opposed to society); and (4) that a dualistic framework (e.g., mind/body) is most appropriate for describing reality. This approach is antagonistic toward and skeptical of CAM, believing many claims to be fraudulent and many practitioners unscrupulous.
2. The complementary perspective: Though extremely varied, those with this perspective do share certain fundamental assumptions: (1) believing in the importance of domains other than 'the physical' for understanding health, (2) viewing diseases as symptomatic of underlying systematic problems, (3) a reliance on clinical experience to guide practice, and (4) a cogent critique of the limits of the biomedical approach. Interventions at the psychological, social, and spiritual level are all thought to be relevant and important, supporting the idea of a biopsychosocial model. Many advocates are critical of biomedicine's harsh and often unsuccessful treatments, and point out the paradox of biomedicine often not being based on 'solid scientific evidence.'
3. The progressive perspective: Proponents of this perspective are prepared to support either of the above, depending entirely on the scientific evidence. They are hardened empiricists who believe it is possible to integrate the best of biomedicine and unconventional approaches. Like all other health care professionals, their approach is not value-free – the advocates of this approach welcome the scientific testing of all sorts of unconventional therapies.
4. The postmodern perspective: This approach enjoys challenging those with absolute faith in science, reason, and technology, and deconstructing traditional ideas of progress. Followers are distrustful of, and cynical toward, science, medicine, the legal system and institutionalized religion, and even parliamentary democracy. Postmodernists see truth as a socially and politically constructed idea and believe orthodox practitioners to be totalitarian persecutors of unconventional medicine. Proponents of this position argue that (1) to have a complementary perspective in any debate is healthy, (2) CAM practitioners are also connected to particular economic and theoretical interests,

(3) a variety of values and criteria for assessing success is beneficial, and (4) the ill people themselves should be the final arbiters of the success of the therapy.

There is more diversity than unity within CAM. While there have been calls to find regulatory bodies to oversee all CAM practices, this has proved very difficult because of the theoretical, historical, and political differences between the various CAM specialties.

Research in CAM

The popular interest in CAM has been matched by a relatively sudden and dramatic increase in research into the two central questions in this area:

1. Does it work?
   Is there good evidence from double-blind, placebo-controlled, randomized studies that the therapy 'cures illness' as it says it does? That is, is there any indisputable scientific evidence that documented findings of success are due to anything more than a placebo effect? Properly designed and executed studies are complex, very expensive, and similar to the research effort to determine the efficacy of psychotherapy. Indeed, it is the extensive research into the placebo effect that makes psychological input particularly valuable (Vincent and Furnham, 2001). The answer to the question is that either very little or no good evidence (from the gold-standard randomized, controlled, double-blind studies) is available for the efficacy of most CAM, with the possible exception of herbalism. However, as more and more sophisticated meta-analyses are published, there does seem to be clear evidence for small but robust positive effects of specific CAM treatments (Ernst and Pittler, 1998).
2. Why choose it?
   If the evidence is limited, equivocal, and indeed often points to lack of efficacy, the central question must be why do patients choose at their own expense to visit a CAM practitioner? What do they get from the treatment? Why do they persist? This is where there have been many psychological studies (Furnham, 2002; Vincent and Furnham, 2001). They concern the often mixed motives that patients have in shopping for health treatments.

   The principal reason for individuals beginning any CAM treatment appears to be that they regard it as more natural and effective, and it allows a more active role for them. The second reason is the failure of OM to provide relief for specific (usually chronic) complaints. The adverse effects of OM, and a more positive patient–practitioner relationship, are also important for many patients. There is little to support the widely held view that CAM-seeking patients are especially gullible or naïve, or have unusual (neurotic) personalities or (bizarre) value or belief systems. However, comparisons of users and nonusers of CAM have shown evidence of different beliefs about health and disease in general. To some extent, CAM patients are as much 'pushed away' from OM, which does not deliver what it promises as 'pulled toward' CAM by its promises and different treatments.

CAM and the Medical Profession

There have been a limited number of studies of practicing doctors' (usually general practitioners') attitudes to CAM. In an early study, Wharton and Lewith (1986) found high level of interest in CAM among GPs. In total, 38% of the 145 respondents had received some training in CAM, and 15% wished to arrange training. In the previous year, 76% had referred patients for alternative treatment to a doctor, 72% had referred patients to nonmedically qualified practitioners, and 70% thought that the more acceptable techniques (e.g., hypnosis, acupuncture, spinal manipulation, and homoeopathy) should be available through the NHS. It should be noted that homoeopathic treatment has been available through the NHS from its inception.

Although early studies were relatively small scale, they yielded comparable results, showing a generally high level of interest in CAM among GPs. Critics have noted many quite orthodox techniques (e.g., manipulation, counseling) were wrongly classified as complementary, so exaggerating the interest in truly complementary techniques (Vincent and Furnham, 2001). GPs may also not be representative of the orthodox medical profession. Visser and Peters (1990) suggested that GPs have a more pragmatic attitude to medicine than their hospital colleagues, which makes them more willing to accept CAM, although they may remain skeptical of its efficacy (Furnham et al., 1995).

The consistent message from these studies is that most GPs seemed favorably disposed toward CAM therapies, particularly the better established therapies like acupuncture, chiropractic, and osteopathy, with homoeopathy being the most popular of the CAM therapies.

There is also a growing literature on the attitudes, beliefs, and expectations of medical students, as opposed to qualified practitioners, toward CAM (Easthope et al., 2000). Complementary medicine has even been introduced to the curriculum of some medical schools. Furnham et al. (1999) went on to examine preclinical medical students' attitudes specifically to the following five complementary therapies: herbalism, homoeopathy, hypnosis (hypnotherapy), osteopathy, and acupuncture. They found very few statistically significant differences in the students' attitudes to the five therapies. Furnham and colleagues compared the attitudes to CAM of both medical and social science students. They found few differences between them and that both took a strongly 'psychological' approach to health, believing that psychological factors were often the cause of illness. Both were supportive, and skeptical, of some aspects of OM. However, they had more positive attitudes to CAM. Compared to GPs, they believed that CAM practitioners were more sympathetic, had more time to listen, were more sensitive on emotional issues, and were better at explaining treatments. Both student groups seemed particularly positive toward dietary and exercise treatments.

Differences in the Consultation between CAM and Orthodox Medicine

Is the popularity of CAM and its powerful placebo effect due to the often fundamental and dramatic differences between the

stereotypic CAM and OM consultation? There is only a very limited amount of research in this area.

Is it possible to generalize? Is there indeed such a thing as a typical consultation? Is the variation within each group (i.e., CAM vs OM) different from/greater than the variation between each group? The differences between the consultations of an aromatherapist compared to an osteopath or a psychiatrist compared with an orthopedic surgeon consults are considerable. Indeed, there may be different 'schools of thought' that result in different types/styles of consultations within each CAM or OM specialty. Then, there may be differences depending on the biography, demographics, and training of the individual practitioner. A typical consultation may be hard to define.

Consultations are so varied that any generalizations about the difference are only stereotypical, possibly misleading or meaningless.

Yet patients and practitioners acknowledge, even celebrate, the particular and peculiar approach to the consultation. Wyatt and Furnham (2010) compared 16 different criteria on which these two may differ and may in part explain the popularity of CAM. GPs all too often have too little time, may be perceived as patronizing, and may not examine (touch) the patient. Further, patients are often not asked the full set of questions they expect to be asked for a 'full' diagnosis. In short, they are not treated like a modern adult consumer. CAM practitioners have longer consultations and appreciate patients' need to talk, to be examined/touched, etc. The question is how the traditional/average CAM consultation is different from (better than) the traditional orthodox consultation. It is possible to compare and contrast the typical GP and CAM consultation across a number of variables (history-taking approach, language used, patient role, decision-making process, bedside manner, etc.) to show how different they are, which may account for the popularity of the latter. Research is needed to confirm the view that it is very much the nature of the consultation that both differentiates CAM from OM and makes it attractive.

Why Do People Use CAM?

There is some evidence that frequent CAM users are more health conscious and believe more strongly that they can influence their own state of health, both by lifestyle and through maintaining a psychological equilibrium. CAM patients appear to have less faith in 'provider control,' that is, in the ability of medicine (specifically orthodox doctors) to resolve problems of ill health. Patients with cancer using CAM were more likely to believe that cancer was preventable through diet, stress reduction, and environmental changes and that patients should take an active role in their own health.

1. People shop for health. They want to use all possible (and affordable) options in health care. People are not loyal to a brand, to OM, or any particular therapy. They shop, try out, and experiment. CAM is therefore to many just another product/service. The question is how the particular brand offers something quite different that no other product or service offers. This raises the question as to what makes an individual 'brand loyal'; that is loyal to a therapy, a therapist, or indeed a place of treatment.
2. People want a cure without side effects or pain. This may offer a very strong, unique selling point for homeopathy over herbalism, acupuncture, etc. because of the scare stories about poisoning with herbs or minerals, and pain/infection with acupuncture. It is, for instance, the 'gentleness' of homeopathy and its dilutions that may be particularly attractive to people. The possible contradiction between being harmless and effective is often not confronted.
3. Because they have chronic illnesses or conditions they have difficulty living with. Many patients with chronic painful conditions or addictions have tried many other cures. They turn to CAM sometimes as a last hope. Some therapies have a powerful psychological component, particularly those associated with touch (i.e., massage, reflexology).
4. Because they are disappointed by the traditional orthodox consultation. There are many reasons for patients' disappointment with OM, but it seems that the nature and style of the consultation is the primary explanation of this.
5. Because patients want to learn more about self-care (fitness, wellness, and prevention). OM is seen as narrow and disease (complaint) orientated, which aims to destroy, demolish, or suppress 'sickening forces,' through things such as chemical therapies and surgery. But many people prefer an emphasis on natural restorative processes on how to strengthen the vitalizing health-promoting forces. The emphasis is quite different – illness vs. wellness. Psychologists have long recognized this. CAM is often seen as restorative, balancing, natural, and preventative, which fits in with the particular zeitgeist.
6. Patients believe in the 'holistic' message. It seems obvious to most patients that lifestyle, personal relationships, and work operate together and simultaneously to have an impact on health. Equally they believe that there are many and manifold signs of wellness and illness from digestion, sleep patterns, and body appearance to more subtle nonverbal signs associated with gait, balance, and body odor. The implication is that the diagnostic interview needs to have questions about all aspects of the person's life not only their physical symptoms.

Who Seeks CAM?

It is foolhardy to talk about the 'typical' patients as CAM rejoices in differences, individuality, and the uniqueness of people's lives. However, there are patterns and a higher probability that particular types of people are likely to seek out, use, and benefit from CAM.

Demography: CAM patients are more likely to be women rather than men, aged 30–40 years rather than older or younger, middle rather than working class, well educated, and urban rather than rural.

Medical History: First, patients seeking CAM have chronic problems rather than acute. Second, their health problems are often nonspecific or have a psychological component. Third, many patients have a 'thick file' in the sense that their interest in health issues has led them to seek out various remedies from many different sources.

Beliefs, Attitudes, and Values: To a large extent, it seems the CAM practitioners have tapped into the zeitgeist of people's beliefs about health. Many patients seem to be sympathetic with green issues, ideas, and understanding. These include environmentalism, one-world-ism, and antimaterialism. Pro-CAM beliefs may also include issues of inequality, alienation, and social exclusion. CAM patients also seem to be interested in general consumer affairs and may even belong to bodies that lobby in favor of a certain position. They appear to be sensitive to consumer rights, bad practice, and poor treatment. CAM patients appear to be particularly interested in the 'life of the mind.' They certainly believe the maxim of "a healthy mind in a healthy body." CAM patients are, because of their own medical condition, likely to be very empathic to the plight of others, and hostile to the 'uncaring' attitude of certain specialists (e.g., surgeon).

See also: Chronic Illness, Psychosocial Coping with; Culture as a Determinant of Mental Health; Doctor–Patient Interaction in the West: Psychosocial Aspects; Environmental Health and Safety: Social Aspects; Evidence-Based Medicine; Race: Ethnicity and Health; Socioeconomic Status and Health Care.

Bibliography

Aakster, C., 1986. Concepts in alternative medicine. Social Science and Medicine 22, 265–273.

Bishop, F., Prescott, P., Chan, Y.-K., Saville, J., von Elm, E., Lewith, G., 2010. Prevalence of complementary medicine use in pediatric cancer: a systematic review. Pediatrics 125, 768–776.

Easthope, G., Tranter, B., Gill, G., 2000. General practitioners' attitudes toward complementary therapies. Social Science and Medicine 51, 1555–1561.

Eisenberg, D., Davis, R., Ettner, S., Appel, A., Wilkey, A., Van Rompay, M., Kessler, R., 1998. Trends in alternative medicine use in the national survey. Journal of the American Medical Association 11, 1569–1575.

Ernst, E., Wider, B., 2008. Oxford Handbook of Complementary Medicine. Oxford University Press, Oxford.

Ernst, E., 2000. The british house of lords enquiry into complementary and alternative medicine. Focus on Alternative and Complementary Medicine 5, 3–5.

Ernst, E., Furnham, A., 2000. BMWs and complementary/alternative medicine. Focus on Alternative and Complementary Therapies 5, 253–254.

Ernst, E., Pittler, M., 1998. The effectiveness of acupuncture in treating acute dental pain: a systematic review. British Dental Journal 184, 443–447.

Furnham, A., Hanna, D., Vincent, C., 1995. Medical students' attitudes to complementary medicine. Complementary Therapies in Medicine 3, 212–219.

Furnham, A., 2000. How the public classify complementary medicine: a factor analytic study. Complementary Therapies in Medicine 8, 82–87.

Furnham, A., 2002. Complementary and alternative medicine. The Psychologist 15, 228–231.

Furnham, A., Yardley, L., Fahmy, S., Jamie, A., 1999. Health beliefs and preferences for medical treatment: a comparison between medical and social science students. Journal of Alternative and Complementary Medicine 7, 101–109.

Gray, R., 1998. Four perspectives on unconventional therapy. Health 2, 55–74.

Leach, M., Gillham, D., 2011. Are complementary medicine practitioners implementing evidence based practice? Complementary Therapies in Medicine 19, 128–136.

Nahin, R., Barnes, P., Stussman, B., Bloom, B., 2009. Cost of Complementary and Alternative Medicine (CAM) and Frequency of Visits to CAM Practitioners. United States, 2007. NCHS, Maryland.

Offit, P., 2012. Studying complementary and alternative therapies. Journal of the American Medical Association 307, 1803–1804.

Shapiro, R., 2008. Suckers: How Alternative Medicine Makes Fools of Us All. Vintage Books, London.

Singh, S., Ernst, E., 2008. Trick or Treatment. Bantam Press, London.

Stoneman, P., Sturgis, P., Allum, N., Sibley, E., Sampson, M., 2013. Incommensurable worldviews? is public use of complementary and alternative medicines incompatible with support for science and conventional medicine? PLoS ONE 8, e53174.

Vincent, C., Furnham, A., 2001. Complementary Medicine. Wiley, Chichester.

Visser, G., Peters, L., 1990. Alternative medicine and general practitioners in the Netherlands. Family Medicine 7, 227–232.

Wharton, R., Lewith, E., 1986. Complementary medicine and the general practitioner. Lancet 292, 1498–1500.

White, A., Ernst, E., 2000. Economic analysis of complementary medicine: a systematic review. Complementary Therapies in Medicine 8, 111–118.

Wyatt, B., Furnham, A., 2010. The medical consultation: what do users and non-users of complementary medicine value? Focus on Alternative and Complementary Medicine 15, 88–96.

Health and Illness: Mental Representations in Different Cultures

Ross G White, Institute of Health and Well-being, University of Glasgow, Glasgow, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G.D. Bishop, volume 10, pp. 6489–6491, © 2001, Elsevier Ltd.

Abstract

Mental representations of *health* and *illness* are thought processes that guide people's response to the threats to their physical and mental well-being. This article will explore how people living in different cultural contexts understand health and illness, and the implications this has for how they subsequently behave. Biomedical approaches will be presented as the dominant model for understanding health and illness in the West. However, comparisons will be made with other frameworks of understanding that are popular in other cultural settings. In this age of globalization, focus is also allocated to considering the costs and benefits to people of having competing ways of understanding health and illness available at onetime.

At the outset it is important to highlight that contemporary understanding about notions of health and illness in the West owe much to the contributions of prominent sociologists working in the mid-to-late twentieth Century. These individuals were particularly interested in how the age of modernity had impacted on the conceptualization of health and illness. Talcott Parsons was a functional sociologist who defined illness as:

> … a state of disturbance in the 'normal' functioning of the total human individual, including both the states of the organism as a biological system and of his personal and social adjustments. It is thus partly biological and partly socially defined.
>
> Parsons, 1951: 431

Parsons viewed illness as a form of social deviance. Parsons coined the phrase *sick role* to capture how those who experience illness are assigned a particular status in human societies so that the social system's stability and cohesion can be maintained. Health on the other hand was defined as:

> State of optimum capacity of an individual for the effective performance of the roles and tasks for which he has been socialized. It is thus defined with reference to the individual's participation in the social system.
>
> Parsons, 1964: 274

Michel Foucault proposed an alternative form of functionalism, which suggested that the medical profession had emerged to replace religion-based discourses that had dominated preceding centuries. He suggested that modernity had resulted in social discipline and influence being driven by secular science rather than religion. Foucault's position was that the medicalization of the body could be understood as a process of social control (Foucault, 1975). Foucault claimed that definitions of normality (health) and deviance (illness) provided people with the vocabulary through which their medical needs were defined. He suggested that the ultimate beneficiary of this discourse was the medical profession (Bilton et al., 2002).

More recently, Turner (1987) has proposed an integrated sociological analysis of health, illness, and medicine by distinguishing between three separate levels of potential focus: the individual (i.e., the experience of illness and disease), the social (i.e., institutional dynamics – Parsons), and the societal (i.e., wider macro or systemic structures – Foucault). The contributions of Parsons, Foucault, and Turner were all made in the context of the biomedical model being the dominant framework for understanding health and illness in the West. In recent years there have been concerted efforts made to globalize biomedical health systems and services. But with this globalization agenda comes the danger of globalizing systems of understanding that clash with local understanding. This has highlighted the importance of understanding health and illness across different cultures.

Cultural Differences in Mental Representations of Health and Illness

Culture has been defined as:

> a set of institutional settings, formal and informal practices, explicit and tacit rules, ways of making sense and presenting one's experience in forms that will influence others.
>
> Kirmayer, 2006: 133

Cultural factors have been highlighted as influencing the decisions that individuals and groups make about health and illness (Angels and Toits, 1987). In this way 'illnesses' can be distinguished from 'diseases' which are defined in terms of biological processes and are culturally invariant (Kleinman, 1980; Good and Good, 1981). Research has indicated that different cultural groups can employ distinctive explanations to account for health and illness. Kleinman (1980) introduced the term *explanatory model* (EM) to capture understanding about the origins of illness endorsed by particular cultural groups. Angels and Toits (1987) state that the structured narratives about health and illness that can emerge from these EMs can become 'overlearned' to such an extent that they are assumed to be objective reality. These narratives, irrespective of their veracity, can exert a strong influence at a societal level on how people think they should behave (Hacking, 1995). A range of

interview measures exist for eliciting narratives and exploring EMs in different cultural contexts e.g., Explanatory Model Interview Catalogue (EMIC; Weiss, 1997), Short Explanatory Model Interview (SEMI: Lloyd et al., 1998), McGill Illness Narrative Interview – MINI (Groleau et al., 2006), and Barts Explanatory Model Inventory (BEMI: Rüdell et al., 2009). The BEMI builds on the strengths of the previous measures and utilizes mixed qualitative and quantitative methods to facilitate the exploration of perceptions of illnesses across different cultural groups.

When considering the potential interplay between culture and EMs, it is important to guard against the misconception that understanding about health and illness will be uniformly similar across a particular cultural group. Pronounced levels of heterogeneity in beliefs/practices can exist within cultures. For example, Angels and Toits (1987) claimed that within particular cultural groups social factors can influence whether 'symptoms' are viewed as atypical. Social changes such as migration or economic downturns can have cultural repercussions that can impact on narratives about health and illness (Angels and Toits, 1987). So, in considering cross-cultural perspectives on health and illness it will be important to be aware of the risk of assuming that beliefs and practices are homogenous across particular cultural groupings.

Reflecting on the multitude of diverse perspectives on health and illness, Phillips and Verhasselt (1994: p. 316) stated:

> … there is evidently neither a single, universally applicable health system nor an agreed view of health as a concept.

As such, the EMs employed by the biomedical model constitute what has been termed an *ethnomedicine*; a way of understanding health and illness unique to a particular cultural grouping. Other major forms of ethnomedicine include: Chinese traditional medicine, Indian medicines (*ayurveda*, *unani*, and *siddha*), and African traditional healing approaches. Chinese medicine is based on the idea that a life force ('qi') flowing through the body on particular meridians maintains health and vitality (Micozzi, 2005). The philosophical underpinnings of ayurvedic medicine propose that 'prana' gives life to the physical body, and that the body, the mind, and the spirit are interconnected (Pole, 2006). Similarly, traditional healing approaches in Africa generally regard the mind, body, and spirit as one (Graham, 2002), with disruptions in this harmony causing illness. Balance is restored by reintegrating the person with his or her community, the earth, and the spiritual world. Whereas biomedical education teaches medical students to eliminate contextual issues from clinical interviews (Good, 1994), these alternative ethnomedicines appear to be more open to the influence of sociocultural factors.

Implications of Cultural Differences in Mental Representations of Health and Illness

Traditional African healing approaches differ from the Western approach in two main ways (Uys, 1989). First, it focuses on the *who* that caused the illness e.g., was it angered ancestors or did a particular individual engage in witchcraft? Secondly, healing is concerned with the total person, which also incorporates close and distant social relationships. The amount of focus afforded to the social milieu of health and illness is likely to be related to the importance that a particular society allocates to the individual compared to the group i.e., individualistic societies versus collectivist societies (Hofstede, 1980). An example of how important social factors can be in the identity of people living in sub-Saharan Africa is provided by the isiZulu concept of *Ubuntu*, which refers to the notion that *a person is a person through other persons*. Liddell et al. (2005) pointed out that Sub-Saharan African societies tend to categorize illness into three different types:

1. Illnesses which are not linked to particular moral or social triggers (e.g., colds and rashes). These types of illness are believed to be caused by chance, and reasons are not pursued as to why the illness has happened (Yamba, 1997).
2. Illnesses which affect people across the globe and were brought to Africa as a result of colonial intrusions onto the continent (Green, 1994).
3. Illnesses which are exclusive to African people irrespective of tribal identity (Ngubane, 1977).

Liddell et al. (2005) also differentiated between two types of causes that people might seek for the occurrence of illness in Africa:

1. A proximate cause, which describes the *method* by which an illness is contracted (Yamba, 1997). For example, believing that an illness is caused by contamination from pollutants is an example of a proximate cause (Ingstad, 1990).
2. An ultimate cause, which describes *why* a particular person contracted the illness. Examples of ultimate causes for illness include witchcraft and sorcery, ancestral vengeance or punishment (Liddell et al., 2005; Ross, 2010).

Pretorius's (2001) work from South Africa, illustrates the types of support that can be available from traditional healers:

1. Herbalists who are referred to locally as *Izinyangas*
2. *Sangomas* who operate within a religious supernatural context.
3. The faith healers who integrate Christian ritual and traditional practices.

Traditional healers tend to be more geographically accessible to people than the centers offering Western biomedicine (Freeman and Mots, 1992). Although some have argued that traditional healers are a viable alternative when primary health care is unavailable or costly (Puckree et al., 2002; Aitken, 2003), these forms of traditional healing remain poorly regulated (Morris, 2001). In addition, there are risks of serious biomedical complications e.g., rectal mucosal shedding from the use of certain emetics (Ellis, 1996), toxicity from ingestion of substances (Jacobs, 1993), and septicemia and/or death stemming from the use of unsterilized knives for scarification and circumcision (Kale, 1995; Kuse, 1997). Advice from traditional healers may also contradict that given by proponents of biomedical models of illness. For example, Ross (2010) points out that some traditional healers advise people living with HIV/AIDS to take only traditional herbs rather than

antiretroviral medication. However, there is also scope for integrating traditional understanding about health and illness with biomedical approaches. For example, resistance to condom use in some parts of Africa has been addressed by practitioners describing HIV infection as the work of an evil spirit which can be tricked by condom use (Treichler, 1999).

Pluralism is a term used to describe the adoption of multiple medical systems by a particular population. Factors that impact on whether someone opts for biomedicine or traditional healing include: treatment cost (Broom et al., 2009), concerns about how other people will perceive their treatment choice (Crandon-Malamud, 1991), and whether each interaction with a particular type of practitioner is perceived to provide a fresh opportunity to find a solution to the problem (Orr, 2012). However, pluralism is not necessarily always a positive development. Broom et al. (2009) concluded that medical pluralism in India can actually serve to perpetuate restrictive and discriminatory ideas about human value. Whereas in the West, biomedicine is regarded as transcending politics and cultural factors, the practice of Indian medicine is strongly influenced by sociocultural factors such as caste, class, gender, and religious dynamics (Arnold, 1993; Khan, 2006; Pati and Harrison, 2001). Broom et al. (2009) claimed that the politics of aging, along with the patriarchal nature of Indian society, contribute to biomedical interventions being prioritized for younger people and for males. So pluralism may serve to perpetuate and amplify inequality and inequities in health provision to particular sections of the population.

In recent times there has been a proliferation in the availability of information about health and illness through the Internet, which has served to enhance opportunities for medical pluralism. The abundance of information (e.g., self-health books) and sign-posting for additional sources of information (e.g., support groups) means that people can engage in a process of constructing and reconstructing narratives which may serve to maintain or alter their sense of self-identity (Giddens, 1991; Hardey, 1999). Bilton et al. (2002) point out that whereas historically doctors relied on people's descriptions to identify illness, this has now been passed over to clinical instruments, or assessment tools. Indeed, people are now in a position to self-diagnose by using books, or Web-based resources. Questions remain about how reliable and valid these assessment tools are, and how appropriate and relevant they are for use across borders of language, beliefs, and practice.

Conclusions

It is clear that conceptualization of health and illness have varied across time and geographies. These conceptualizations have been influenced by the cultural beliefs and practices of people living in these times and geographies. Social change can bring shifts in what is considered illness. As Feierman (1985: p. 73) pointed out:

> … the evolution of health cannot be separated from the broader story of social change.

Social constructionist perspectives have variously viewed notions of health and illness as being related to themes of conformity, deviance, and/or maintaining power differentials between medical professionals and the public. Today, people across the world draw on diverse EMs to understand health and illness. Culturally defined differences in beliefs and practices about notions of health and illness have led to the development of particular ethnomedicines including biomedicine, ayurveda, Chinese traditional medicine, and a range of African traditional healing approaches. It is important to appreciate that even within particular cultures there will be great heterogeneity in how people understand concepts of health and illness. We live in increasingly cultural diverse populations, and with this come demands for greater pluralism in the services offered to support health and illness. The choice of support that people might seek will be influenced by the: EMs that people employ, cost, perceived effectiveness, and concerns about what others might think of their choice. Pluralism offers people choices about how they understand illnesses and seek support for these illnesses. However, in some contexts, pluralism can be used to perpetuate political agendas and lead to a tiered system of support that serves to discriminate against sections of the population. The Internet is serving to enhance opportunities for pluralism by providing individuals with a wealth of perspectives and information on health and illness. This is giving rise to increasingly eclectic narratives about mental health, narratives that have the potential to, in turn, dynamically shape expectations about how people should think, feel, and behave when healthy and when ill.

See also: African Studies: Health; Culture as a Determinant of Mental Health; Explanatory Style and Health; Foucault, Michel (1926–84); Globalization and Health; Health: Anthropological Aspects; Health: Definitions; Illness Behavior and Care-Seeking; Medicalization: Sociological and Anthropological Perspectives; Parsons, Talcott (1902–79); Race: Ethnicity and Health.

Bibliography

Aitken, S., 2003. Biodiversity and Health: Conserving Our Healing Resources. http://www.tc-biodiversity.org.

Angel, R., Thoits, P., 1987. The impact of culture on the cognitive structure of illness. Culture, Medicine and Psychiatry 11, 465–494.

Arnold, D., 1993. Colonizing the Body. Oxford University Press, Delhi.

Bilton, T., Bonnett, K., Jones, P., Lawson, T., Skinner, D., Stanworth, M., Webster, A., 2002. Introduction to Sociology. Palgrave MacMillan, Hampshire.

Broom, A., Doron, A., Tovey, P., 2009. The inequalities of medical pluralism: hierarchies of health, the politics of tradition and the economies of acre in Indian oncology. Social Sciences and Medicine 69, 698–705.

Crandon-Malamud, L., 1991. From the fat of our souls: social change, political process, and medical pluralism in Bolivia (Vol. 26). University of California Press.

Ellis, C., 1996. Ukufa Ukwaba. South African Family Practice 10, 105–106.

Feierman, S., 1985. Struggles for control: the social roots of health and healing in modern Africa. African Studies Review 28, 73–147.

Foucault, M., 1975. The Birth of the Clinic: An Archaeology of Medical Perception. Vintage, New York.

Freeman, M., 1992. Recognition and Registration of Traditional Healers: Possibilities and Problems. University of Witwatersrand, Johannesburg, South Africa.

Freeman, M., Motsei, M., 1992. Planning health care in South Africa – is there a role for traditional healers? Social Science & Medicine 34, 1183–1190.

Gbodosson, E., 2000. Benin's Traditional Medicine Promotion Group. In: Twelfth International Conference on AIDS and Sexually Transmitted Diseases in Africa, held in Burkina Faso, Ouagadougou, 2000. The Kaiser Daily HIV/AIDS Report, the National Journal Group Inc., and Kaiser Foundation, USA.

Giddens, A., 1991. Modernity and Self-identity. Polity Press, Cambridge.

Good, B.S., Good, M.D., 1981. The meaning of symptoms: a cultural hermeneutic model for clinical practice. In: Eisenberg, L., Kleinman, A. (Eds.), The Relevance of Social Science for Medicine. D. Reidel, Dordrecht, Holland, pp. 165–196.

Good, B.J., 1994. Medicine, Rationality and Experience. An anthropological Perspective. Cambridge University Press, Cambridge.

Graham, M., 2002. Social Work and African-Centred Worldviews. Venture, Birmingham.

Green, E.C., 1994. AIDS and STDs in Africa. Westview Press, Boulder, CO.

Groleau, D., Young, A., Kirmayer, L.J., 2006. The McGill Illness Narrative Interview (MINI): an interview schedule to elicit meanings and modes of reasoning related to illness experience. Transcultural Psychiatry 43, 671–691.

Hacking, I., 1995. The looping effects of human kinds. In: Sperber, D., Premack, D., Premack, A.J. (Eds.), Causal Cognition: A Multidisciplinary Approach. Clarendon Press, Oxford, pp. 351–383.

Hardey, M., 1999. Doctor in the house: the Internet as a source of lay health knowledge and the challenge to expertise. Sociology of Health and Illness 21, 820–835.

Hofstede, G., 1980. Culture's Consequences: International Differences in Work-Related Values. Sage, Beverly Hills, CA.

Ingstad, B., 1990. The cultural construction of AIDS and its consequences for prevention in Botswana. Medical Anthropology Quarterly 7, 28–40.

Jacobs, M., 1993. The role of traditional medicine in a changing South Africa. Critical Health 43, 73–76.

Kale, R., 1995. Traditional healers in South Africa: a parallel health care system. BMJ 310, 1182–1185.

Khan, S., 2006. Systems of medicine and nationalist discourse in India. Social Science and Medicine 62, 2786–2797.

Kirmayer, L.J., 2006. Beyond the "new cross-cultural psychiatry": cultural biology, discursive psychology and the ironies of globalization. Transcultural Psychiatry 43, 126–144.

Kleinman, A., 1980. Patients and Healers in the Context of Culture: An Exploration of the Borderland between Anthropology, Medicine, and Psychiatry. University of California Press, Berkeley.

Kuse, T.T., 1997. The need for indigenisation of the education of social workers in South Africa. In: De Kock, D., Hall, T., Hayden, W., et al. (Eds.), Indigenisation in Social and Community Work Education. Human Sciences Research Council, Pretoria, pp. 19–23.

Lloyd, K.R., Jacob, K.S., Patel, V., et al., 1998. The development of the Short Explanatory Model Interview (SEMI) and its use among primary-care attenders with common mental disorders. Psychological Medicine 28, 1231–1237.

Liddell, C., Barrett, L., Bydawell, M., 2005. Indigenous representations of illness and AIDS in sub-Saharan Africa. Social Science and Medicine 60, 691–700.

Morris, K., 2001. Treating HIV in South Africa: a tale of two systems. The Lancet 357, 1190.

Micozzi, M.S., 2005. Fundamentals of Complementary and Integrative Medicine. Saunders, Elsevier, Missouri.

Ngubane, H., 1977. Body and Mind in Zulu Medicine. Academic Press, London.

Orr, D.M.R., 2012. Patterns of persistence amidst medical pluralism: pathways toward cure in the Southern Peruvian Andes. Medical Anthropology 31, 514–530.

Parsons, T., 1951. The Social System. Routledge & Kegan Paul, London.

Parsons, T., 1964. Social Structure and Personality. Free Press, London.

Pati, B., Harrison, M., 2001. Health, Medicine, and Empire: Perspectives on Colonial India. Orient Longman, New Delhi.

Pole, S., 2006. Ayurvedic Medicine: The Principles of Traditional Practice. Elsevier Ltd, Philadelphia.

Phillips, D.R., Verhasselt, Y., 1994. Health and development: retrospect and prospect. In: Phillips, D.R., Verhasselt, Y. (Eds.), Health and Development. Routledge, New York, pp. 301–318.

Pretorius, E., 2001. Traditional healers. In: South African Health Review. Health Systems Trust, Pretoria, South Africa, pp. 249–256.

Puckree, T., Mkhize, M., Mgobhozi, Z., Lin, J., 2002. African traditional healers: what health care professionals need to know. International Journal of Rehabilitation Research 25, 247–251.

Ross, E., 2010. Inaugural lecture: African spirituality, ethics, and traditional healing – implications for indigenous South African social work education and practice. SAJBL 3, 44–51.

Rüdell, K., Bhui, K., Priebe, S., 2009. Concepts, development and application of a new mixed method assessment of cultural variations in illness perceptions: Barts explanatory model inventory. Journal of Health Psychology 14, 336–347.

Turner, B.S., 1987. Medical Power and Social Knowledge. Sage, London.

Treichler, P.A., 1999. How to Have Theory in an Epidemic. Duke University Press, Durham, NC.

Uys, L., 1989. Psychiatric care in the South African context. Nursing RSA Verpleging 4, 16–18.

Workneh, F., 1980. Traditional healing in Ethiopia. Excerpta Medica 2, 63–66.

Weiss, M., 1997. Explanatory Model Interview Catalogue (EMIC): framework for comparative study of illness. Transcultural Psychiatry 34, 235–263.

Yamba, C.B., 1997. Cosmologies in turmoil: witch finding and AIDS in Chiawa, Zambia. Africa 67, 200–221.

Health: Anthropological Aspects

Adriana Petryna, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Kleinman, A. Petryna, volume 10, pp. 6495–6499, © 2001, Elsevier Ltd.

Abstract

Images of health, illness, and healing are key media through which societies and groups struggle with central values and beliefs and negotiate social and political progress. The invention and implementation of techniques of health improvement have taken on many forms, including epidemic control, disease eradication, and combinations of environmental and individual behavioral change and medical advance. Yet definitions of health depend on normative and oftentimes conflicting values and contexts, which themselves are influenced by various transformations in economy and polity, science and technology, and culture. Such variables shape the conditions of health and are in turn affected by it. In attending to such reciprocal processes, anthropologists have demonstrated that health as an experience is irreducible to a set of strict criteria of physiological and mental competence or to a set of cultural differences. Only through concrete understandings of particular worlds of well-being and suffering and the way they are shaped by local histories and political economies can we possibly come to terms with the complex dimensions that protect or undermine human health. Health, seen in this way, is a construction as well as a contested way of being in the world.

Health and health practices are part of the inmost complexities of social existence, permeating the domains of politics, economics, and religion and almost always connected with dimensions that go beyond the body, such as interpersonal, family, and community relationships. Health is at once an end and a means. Explanations of the causes and patterns of health and disease often convey value judgments, senses of right and wrong and of accountability and blame, as well as reveal what is morally at stake in definitions of health and its failures. Medical anthropological studies have built upon historical and cultural analyses over a long period of time that indicate an array of difference with respect to the metaphors and meanings that signify health. These studies have emphasized cultural fabrics that give coherence and depth to these meanings such that cosmology and ethical traditions come to define the body in states of sickness and well-being.

Anthropological studies focus on local contexts where health and illness are recognized and responded to. Such studies also trace the effects of global flows of commodities, information, finance, images, people, and pathogens on such worlds. Anthropologists have been as interested in the social roots and consequences of health (and illness) as in their cultural representations. But what most particularly characterizes the anthropological perspective is the use of ethnography to understand health, illness, and health care. Anthropological perspectives on health bring together individual and collective realities in the way they are organized, narrated, contested, and in every sense lived as social trajectories (Kleinman, 1995).

What anthropologists have learned through research in different societies is a story of diversity in experiences of health and health care giving, as well as of the ways social relationships and emotional processes interfuse in the mediation of health-relevant experiences. The meanings of health range from holistic and organic interpretations to hyperindividualistic ones. While biomedicine has treated health as a separate domain, popular meanings and interpretations of healthy states are often inseparable from other moral, political, and economic domains. Anthropologists have been particularly concerned with the transformations in health as those transformations express differences in power, social position, and social inequality, particularly as experienced by marginal groups and individuals. But in recent years, equal interest has been devoted to studies of health professionals, bureaucrats, and scientists who contribute to the public definition of health and health problems. Culture affects health as much through the culture of biomedicine and bureaucratic institutions as through the popular culture; indeed in the age of wide-ranging social media the two are closely related.

Applied medical anthropology addresses issues such as the ones raised and addresses them at the policy level. In heterogeneous societies such as North America and Western Europe, cultural differences continue to matter in the health field in large part because of differences in health conceptions which affect the panoply of health-relevant practices, from popular assessments of risks and responses to them to the uses and misuses of health care resources. Those differences contribute to different health outcomes for people; but they also contribute to different health systems, which are themselves undergoing transformation due to both local and global economic and social changes.

Images of Health and Diverse Experiences of Health

The contemporary world contains impressive variations of images of health, modes of health maintenance, efforts at prevention, and illness experiences. Much of this variation is subsumed under what demographers and epidemiologists have termed the 'health transition.' This phrase refers to an increase in life expectancy, decline in mortality, and overall decrease in infectious diseases in the twentieth century by means of interventions such as vaccinations and dietary improvements. Health status measures have been introduced and are continually revised to develop policy, planning, and evaluation; these measures shape ideas of health and health improvement according to specific indices that promote interventions and

sociobehavioral change. The health transition has also been accompanied by major changes in patterns of morbidity, increases in chronic degenerative diseases, old and new epidemics, and behavioral and mental pathologies and is thus more complex than its original conceptualization (Chen et al., 1994). The very economic, technical, and industrial development that has benefited so many coexists with ongoing massive societal disruptions. Developing countries, for example, are experiencing the social and health costs of rapid urbanization and economic restructuring, costs that are confirmed in statistical increases of preventable disease and chronic illness, and borne in terms of health problems such as infant mortality, life-threatening dehydration, urban violence, and hunger (Scheper-Hughes, 1992; Bourgois, 2002). Such acts as war, political persecution, and homicide contribute to a deliberate production of suffering and death. The era of infectious diseases is by no means past as the spread of HIV/AIDS, multidrug resistant tuberculosis, and other infectious diseases suggest. The collapse of public health measures and treatment regimens, demographic changes, migration, and other human actions intensify the flows and resistance of pathogens. It is through such mechanisms that Paul Farmer has shown the ways social forces alter disease distribution and contribute to the persistence of new microbial conditions; indeed, the way "social forces and processes come to be embodied as biological events" (1999: p. 5). The appearance of these disorders, closely correlated with poverty and social inequality, challenges the notion of a single biological etiology for disease and illustrates that these infectious diseases are mediated by multiple interacting causes that require anthropological inquiry and policy intervention.

The burden of chronic diseases such as heart disease, cancer, stroke, diabetes, and asthma continue to be distributed unevenly across minority groups. Yet statistics also disclose impressive variations. In the United States, for example, there is the so-called Latino Epidemiological Paradox: Latinos, who generally experience a low socioeconomic status, nonetheless, have health outcomes that are either equal or better than non-Hispanic whites. African Americans, by comparison, experience significantly higher rates of infant mortality, death, diabetes, HIV/AIDS, heart attacks, strokes, and homicides. Differences in levels of institutionalized racism, dietary patterns, and the influence of the Spanish language as a protective barrier have been pointed to as possible explanations for this paradox. Notably, health indices of Latino immigrants worsen from generation to generation in the United States, owing to increased assimilation, a loss of balanced diet, chronic insecurity, as well as limited accessibility of health care services (Hernandez and Charney, 1998; Galvez, 2011; Waldstein, 2010; Holmes, 2013).

The latest cluster of disorders to receive attention in the health transition is the rising rates of substance abuse, violence, psychological trauma, depression, suicide, and related mental health problems (Desjarlais et al., 1995). This group of health problems has increased to epidemic proportions both in developed and developing societies. Anthropological studies indicate that these problems often cluster together in settings of deep poverty and societal breakdown. They are compounded by the fact that services tend to be limited, and health care financing and resources are inadequate. Many psychiatric hospitals in China, Russia, and Latin America and elsewhere continue to rely on a model of care involving prolonged stays in locked wards with antipsychotic medication and maltreatment (Phillips et al., 2013; Gostin and Gable, 2004). Anthropological interventions have gone beyond explication of the 'soft' variables underpinning mental health problems and focus on the ways social, economic, and political contexts shape the formation, distribution, and manifestation of disorders. Thus in the anthropological perspective, health problems and social problems are inseparable; health policies and social policies are viewed as necessary complements in defining criteria and techniques of successful rehabilitation.

Local Variations in Illness and Health

In response to emerging sets of health issues, anthropologists have identified the need for cross-cultural studies of health, healing, and morbidity in local contexts. One thread linking these studies is the importance of understanding patterns of everyday life experience, local knowledge, and social networks that influence personal agency and access to health care, as well as biosocial modes of transmission of diseases. From this perspective, health is seen as not simply originating in individuals themselves or deriving strictly from social forces, but as being engendered in complex interactions between individuals and their social contexts (Leslie and Young, 1992; Good et al., 2010; Inhorn and Wentzell, 2012; Lock and Nguyen, 2010). Such lived responses may occur in contexts of medical pluralism where several therapeutic systems coexist in one cultural setting; or in circumstances in which symptoms are subject to several distinct classifications, diagnoses, and hence, therapeutic courses of action or, once again, where different moral and religious values shape health-promoting and disease-preventing activities related to diet and life style practices and support. By stressing contingencies and by detailing local histories of particular places and peoples undergoing many kinds of social change, ethnographies of health and medicine highlight the way that particular worlds define health and conceive of health equity, while also illuminating the ways in which risks to disease and illness interrelate.

It is not surprising that ethnographers have focused on illness whose event structure is relatively defined. Emphasized is the experience of illness as an interpersonal phenomenon as it is narrated through stories of illness. Ethnographies disclose how suffering is transformed through language and cultural meanings into distinctive illness, narratives whose shared metaphors, voices, and genres represent distinctive experiences and political processes. This transformation, in turn, shapes the lived experience of illness through embodiment of cultural meanings and social relations into body-self processes.

Ethnographers have also focused on different mediating processes – ritual, devotional practices, music, and poetics – that structure a sociosomatic continuum between moral and emotional states. This approach relates to the concept of 'local biologies' by which is meant the different means of relating social worlds and biological processes such that those processes become particularized. Examples include

Margaret Lock's demonstration of fundamental differences in the experiences of menopausal symptoms and the aging process among women in Japan and North America (1993); Lawrence Cohen's contrast of distinctive interpersonal experiences of dementia in India and the West (1998); Rayna Rapp's description of how experiences of class in the United States and technologies of the female reproductive life cycle come to alter experiences of pregnancy, infertility, abortion, and delivery (1999); and Adriana Petryna's analysis of biological citizenship and the deep vulnerabilities left behind by state interventions and failures to intervene in a nuclear crisis (2013).

Ethnographers have also analyzed the trajectory and outcome of illness experiences on the social course of illness. Thus Benedicte Ingstad and Susan Reynolds Whyte (2007) and others have shown how disability is as much a product of social definitions and restrictions and opportunities of access as it is a physical or functional limitation or disease, thus turning disability into an experience of potential wholeness and health. Other anthropologists have investigated societal responses to highly stigmatized disorders that respond to feelings of threat with practices aimed at marginalizing sufferers. Still others have studied experiences of chronic conditions like chronic fatigue syndrome, chronic pain syndrome, and fibromyalgia – disorders that are in a limbo of partial or minimal authorization where patients feel discredited or discreditable because of the lack of social legitimacy. Research here shows how patients often resist biomedical responses that are perceived as undermining and delegitimating.

More recently, this area of research has expanded well beyond biomedically defined conditions to include a variety of forms of violence – structural, political, domestic – that represent significant threats to health. Here researchers of violence have shown how different societal idioms of those institutions authorized to deal with social problems – legal, welfare, religious, and of course medical – construct distinctive objects of programs and policies that not only differ greatly, but that coerce different courses of health and disease and outcomes of these conditions (Biehl, 2013; Das, 2006; Farmer, 2004; Han, 2012; James, 2010; Bourgois and Schonberg, 2009). Perhaps the largest field of anthropological inquiry has examined the varieties of community healing systems that parallel state institutions. Studies of herbalists, shamans, and other folk practitioners (and practices) illustrate how experience is manipulated, how meaning is negotiated, and how illness comes to serve some societal ends such as social control, cultural remembering, and political resistance. The point being that the collective significance of illnesses goes beyond individual and even familial concerns and plays a role in how societies remember, how social reality is legitimated, and how social worlds change (Adams, 2013; Biehl, 2007; Dilger and Luig, 2010; Hyde, 2007; Livingston, 2012; Kohrman, 2005; Langwick, 2010; Nguyen, 2010; Rivkin-Fish, 2005; Wendland, 2010).

The anthropology of health and medicine largely began as an applied field. There continues to be important segments of the field that seek application in public health and biomedical settings and in the policy arena. Such applied work classically demonstrated that health beliefs could either facilitate or become a substantial barrier to implementing policies and progress (Paul, 1955). More recently, anthropologists have engaged the nongovernmental, administrative, and political dimensions of international health agencies and the policy development process indicating how community organization can positively inform health and social programs in arenas such as HIV/AIDS prevention, cancer or infectious disease control, and family planning (Hahn, 1999). Hence, anthropology has moved from a stance of critique of application to a stance of generating novel applications combining anthropological knowledge and public health technologies to enhance health.

Nowhere is this better seen than in the recent attention to global health initiatives. Such initiatives and programs are booming in the United States and have begun also to displace earlier framings of the field (such as 'international health'). Anthropologists are engaged in empirical and theoretical investigations of global health-related programs and epistemologies and are concerned with the actual impacts of these initiatives on medical care, citizenship, and governance (Farmer et al., 2013; Biehl and Petryna, 2013; Hahn and Inhorn, 2008; Singer and Derrick Hodge, 2010; Janes and Corbett, 2009; Nichter, 2008; Rylko-Bauer et al., 2009). Multiple theoretical ideas, methods, and epistemic frameworks cohabit within the broad enterprise of global health. Anthropologists engaging this field have pointed to the tensions between private markets and public interests in the design of collective and individual health. They attend to the breakdowns in public health systems and to the myriad social determinants of health (such as education, water, sanitation, vector control, air pollution, and violence prevention) that make people vulnerable to disease in the first place, as well as the humanitarian, legal, and experimental settings that make up new infrastructures of humanitarianism, survival, and global patient life – and where law and medicine become deeply personal (Redfield, 2013; Petryna, 2009; Fassin, 2009; Ticktin, 2011). Through intervention into these fields and analysis of how global health programs are implemented, received, or improvised, anthropologists have pointed to the limits of traditional evidence-making practices and models of care delivery, as well as to the moral values underpinning large-scale health designs. In measuring the value that interventions have for people, anthropologists have illustrated both intended and unintended consequences of these interventions on those in need and how they are 'contested or resisted at different social scales,' thus challenging the certainties of planners and implementers alike.

Culture of Biomedicine

Distinctions between health and disease or illness are patterned according to norms. Norms of health and disease often emerge through a combination of statistics patterns, public consensus around desirable health outcomes, and social definitions and responses to deviance (Mechanic, 1997). The extent to which biomedical norms have achieved a position of preeminence in health care systems and definitions of disease is undeniable. Biomedicine's success in eradicating certain diseases and prolonging life shows that norms, however value-laden and socially constructed they may be, can indeed be health-enhancing (DelVecchio Good, 1995). Yet anthropological

studies have questioned biomedicine's reductionism, primary value of efficiency, and authority in mediating interpretations of health, illness, and suffering (Good, 1994; Kleinman, 1995); they point to the fact that physiological measures alone fail to capture the subjective, social, and even racialized dimensions of health (Rouse, 2009; Mattingly, 2010; Montoya, 2011; Fulwiley, 2011; Lee et al., 2008). The growth of biomedicine, its characteristic quality of redefining social problems as health problems needing biomedical intervention, and its global cultural reach have also been analyzed. That biomedicine has been closely allied with the growth of industrial capitalism has led many ethnographers to examine biomedicine as a deep conduit of social regulation and discipline influencing all sectors of life, from definitions of disability to norms of occupational health, to definitions of mental illness and fitness (Turner, 1995; Martin, 2009). This process has led to strong critiques regarding the necessity and indeed at times harmfulness of biomedical interventions, particularly with respect to the unpredictable side effects of technical procedures and drugs used to diagnose and treat patients. Public concerns over medically induced adverse effects (known as iatrogenesis) are reflected in current debates over the consumption of psychiatric drugs by preschool children in the United States, for example. Cross-cultural studies reveal the extent to which this integration with local experience is contested or resisted by highlighting: the many discrepancies between biomedical knowledge and subjective accounts of illness and health; the moral and religious tensions underlying choices of medical systems; and the degree to which distinctions between health and disease are rooted (or uprooted) by global ideologies and beliefs.

Attention to the international political economic context in which biomedicine operates is central to understanding how ideas of health modify, as well as to understanding the modes through which they are integrated as a set of values and practices in local contexts. As with nonbiomedical healing traditions, pharmaceuticals too are a compelling theme in the anthropology of health and medicine because they link individual understandings of disease and therapeutic practice with broader scientific and political–economic regimes of health (Whyte et al., 2002; Petryna et al., 2006; Ecks, 2013). The pharmaceutical industry is the major producer of today's therapeutic armamentarium. In more than 70% of all medical doctor visits in the United States, doctors will prescribe or continue to provide a medication to a patient. On average, two drugs or 'drug mentions' have been documented per medical visit. Sustaining this pill-taking, life is a complex globalized system of drug testing and drug development. In capturing the trajectory of pharmaceuticals (including their regulatory dimensions, clinical testing, industrial production, and mass consumption), anthropologists have made critical inquiries into notions of 'best evidence' as well as changing boundaries between the normal and the pathological in worlds of commodified health. From local and national ethnographic perspectives, they engage the regulatory and international politics of patents and intellectual property and evolving health-as-human-right paradigms and infrastructures, and have assessed the impact of market-driven medicine on cultures of care, public health, and ethics.

Biotechnology and Health

In response to the large influence that public health has achieved in our times, many due to the advances of biomedicine and advances in technology, the power of state action in the public domain, and market forces, health has increasingly become a specialized domain. The history of public health is punctuated by major advances in population-level health in the nineteenth and early twentieth centuries, through sanitary and workplace improvements, housing, and nutrition. In recent years, public health priorities are shifting from social medicine (with its attention to the social determinants of health) to individually targeted health promotion. The launching and completion of the largest biology project ever conceived – the decoding and mapping of the human genome – has already accelerated this process of individualizing health vis-a-vis the genetic profiling of diseases and increased prognostic information. For example, silicon chips embedded with hybrid bits of DNA can test for various genetic conditions in individuals. Such 'self-quantifying' diagnostic tools are refocusing medical practice away from intervention to risk factor analyses and disease prevention.

The diseases individuals might be predisposed to, how they might react to particular medicines, and what health-promoting behaviors they might adopt to prevent future illness are all encompassed in this revolution (Rabinow, 1996; Eisenburg, 1999; Hamdy, 2012; Sanal, 2011; Roberts, 2012; Edmonds, 2010; Brotherton, 2012; Taussig, 2009; Fischer, 2009; Sharp, 2006; Dumit, 2012). Genomic information and the multitude of genomic sequences represent novel, but as yet largely unrealized, targets for therapeutic intervention. In short, the 'life sciences industry' – knowledge of the genetic code and how to manipulate it and therapeutic promises – is transforming diagnostics and intensifying screening for genetic disorders, often state-mandated (Timmermans and Buchbinder, 2012), but it has yet to usher in a new age of personalized treatment. As consumers embrace incremental self-improvement via genomic knowledge, the prospect of widespread genetic testing has prompted concerns about how such information is being used to discriminate against those labeled as predisposed to genetic risks. Historically, such discriminations took the form of restricting or denying insurance coverage or employment. Inherent inequalities of access to health care can also limit the effectiveness of new technologies.

The success of biotechnology highlights the growing demand for health as a market sector. It attests to the power of biomedicine as a rationality in everyday life (in the organization of health care and delivery, policy analysis, legal procedures) that comes up against other forms of knowledge such as religion, and that coexists with older power structures related to family, gender, ethnicity, and class. Health as commodity, to be objectified, bought, and sold in the marketplace has led to new freedoms to pursue health as an individual project, but also raises dilemmas and creates new risks: who is caretaker of genomic information? Who will have access to genetic therapies? These questions highlight some of the dangers regarding the social uses of genetic knowledge in the genomic and postgenomic era; they make it difficult to say what counts as

success in this drive toward health and profits, a drive which has both sacred and secular sources and consequences.

The trade off between benefits to individuals and risks to populations of any given new technology must also be carefully assessed. For example, over a decade ago xenotransplantation promised to end the so-called shortage in donor organs by engineering transplantable animal organs. While breakthrough technologies might save the life of a few, they pose dangers to populations in terms of the resistant viral strains and antibodies that might develop as a result of such transplant procedures. Another example of the way scientific advancement produces new social risks is found in the case of the traffic in human organs. Such traffic has been afforded by the availability of cyclosporine, among other things. This availability, combined with demand for human organs, has turned vulnerable populations in Brazil, South Africa, and India into potential organ suppliers. Inconsistencies across the globe regarding the rights of donors as well as the lack of follow-up care for many of them have lead to prolonged suffering or premature deaths, or both. Processes related to the globalization of the life sciences industry demand ethnographic attention to technical advances in terms of their development and diffusion; patterns of accessibility and differences of improved health outcomes across socioeconomic divides; and new forms of health and vulnerability these advances engender.

As the life sciences redraw human experience and redefine life chances, anthropologist have turned their gaze once again to social relations, namely, to the moral dimensions and necessities of care and caregiving, particularly in the context of chronic disease such as dementia, diabetes, cancer, and addiction and global health and end-of-life care (McLean, 2006; Norwood, 2009; Mol, 2008; Taylor, 2008; Garcia, 2012; Kaufman, 2006; Manderson and Smith-Morris, 2010). Care is fundamental to the managing of illness but it is often taken for granted or undervalued within professional medical practice and broader political and moral economies of health care. Anthropologists have described caregiving as "the close experiential language of actually doing it" (Kleinman, 2009: p. 292). The elderly, disabled, or chronically or terminally ill run the risk of being 'unmade' without the intimate practices of care that protect them as recognized persons. Anthropological studies of caregiving in cancer wards, hospitals, nursing homes, and hospice settings have pointed to the unexpected kin and social networks that sustain openness toward the humanity of the care recipient that an overly technical biomedical approach might foreclose.

Images of health, illness, and healing have become key media, a 'fragile salvation,' through which societies and groups struggle with central values and beliefs, and negotiate social and political progress. The invention and implementation of techniques of health improvement have taken on many forms, including epidemic control, disease eradication, and combinations of environmental and individual behavioral change and medical advance. Yet definitions of health depend on normative and oftentimes conflictual values.

Health as an individual goal coexists with new threats to health worldwide. The spread of advertising for cigarettes or the dumping of expired pharmaceuticals in low-income countries are all part of the globalization of disease. Under such conditions, what was once health, now becomes disease. Health maintenance and promotion, once so influential a concept, can be undermined owing to new political and economic realities.

Contexts of health are themselves undergoing changes and are influenced by transformations in economy and polity, in science and technology, and in culture. These contexts shape the conditions of health and are in turn affected by it. These processes demonstrate that health as an experience is irreducible to a set of norms of physiological and mental activity, or strictly to a set of cultural differences. Only through concrete understandings of particular worlds of suffering and the way they are shaped by local histories and political economies can we possibly come to terms with the complex human dimensions that protect or undermine health. Health, seen in this way, is a construction as well as a contested way of being in the world.

See also: Body Image; Body: Anthropological Aspects; Comparative Health Systems; Culture as a Determinant of Mental Health; Disability: Sociological Aspects; Drinking, Anthropology of; Ethnography in Applied Social Research; Globalization and Health; Globalization, Anthropology of; Health in Low-Income Countries; Religion and Health: Sociological Considerations; Reproduction and Cultural Anthropology; Science and Technology, Anthropology of.

Bibliography

Adams, Vincanne, 2013. Markets of Sorrow, Labors of Faith: New Orleans in the Wake of Katrina. Duke University Press, Durham.

Biehl, João, 2013. Vita: Life in a Zone of Social Abandonment. Updated with a New Afterword and Photo Essay. University of California Press, Berkeley.

Biehl, João, 2007. Will to Live: AIDS Therapies and the Politics of Survival. Princeton University Press, Princeton.

Biehl, João, Petryna, Adriana (Eds.), 2013. When People Come First: Critical Studies in Global Health. Princeton University Press, Princeton.

Biehl, João, Petryna, Adriana, 2011. Bodies of rights and therapeutic markets. Social Research 78 (2), 359–394. Summer.

Bourgois, Philippe, 2002. In Search of Respect: Selling Crack in El Barrio. Cambridge University Press, Cambridge.

Bourgois, Philippe, Schonberg, Jeffrey, 2009. Righteous Dopefiend. University of California Press, Berkeley.

Brotherton, Sean P., 2012. Revolutionary Medicine: Health and the Body in Post-Soviet Cuba. Duke University Press, Durham.

Chen, L., Kleinman, A., Ware, N., 1994. Health and Social Change in International Perspective. Harvard University Press, Cambridge.

Cohen, Lawrence, 1998. No Aging in India: Alzheimer's, the Bad Family, and Other Modern Things. University of California Press, Berkeley.

Das, Veena, 2006. Life and Words: Violence and the Descent into the Ordinary. University of California Press, Berkeley.

Desjarlais, R., Eisenberg, L., Good, B., Kleinman, A., 1995. World Mental Health: Problems, and Priorities in Low-Income Countries. Oxford University Press, New York.

Dilger, Hansjörg, Luig, Ute (Eds.), 2010. Morality, Hope and Grief: Anthropologies of AIDS in Africa. Berghahn, New York.

Dumit, Joseph, 2012. Drugs for Life: How Pharmaceutical Companies Define Our Health. Duke University Press, Durham.

Ecks, Stefan, 2013. Eating Drugs: Psychopharmaceutical Pluralism in India. NYU Press, New York.

Edmonds, Alex, 2010. Pretty Modern: Beauty, Sex, and Plastic Surgery in Brazil. Duke University Press, Durham.

Eisenburg, Leon, 1999. The Implications of Genetics for Health Professional Education. Josiah Macy, Jr Foundation, New York, p. 10.

Farmer, Paul, 1999. Infections and Inequalities: The Modern Plagues. University of California Press, Berkeley.
Farmer, Paul, 2004. Pathologies of Power: Health, Human Rights, and the New War on the Poor. University of California Press, Berkeley.
Farmer, Paul, Kleinman, Arthur, Yong Kim, Jim, Basilico, Matthew (Eds.), 2013. Reimagining Global Health an Introduction. University of California Press, Berkeley.
Fassin, Didier, 2009. The Empire of Trauma: An Inquiry into the Condition of Victimhood. Princeton University Press, Princeton.
Fischer, Michael M.J., 2009. Anthropological Futures. Duke University Press, Durham.
Fulwiley, Duana, 2011. The Enculturated Gene: Sickle Cell Health Politics and Biological Difference in West Africa. Princeton University Press, Princeton.
Galvez, Alyshia, 2011. Patient citizens, immigrant mothers: Mexican women, public prenatal care, and the Birth weight paradox. Rutgers University Press.
Garcia, Angela, 2012. The Pastoral Clinic: Addiction and Dispossession along the Rio Grande. University of California Press, Berkeley.
Good, B., 1994. Medicine, Rationality, and Experience: An Anthropological Perspective. Cambridge University Press, Cambridge and New York.
Good, Mary-Jo DelVecchio, 1995. American Medicine: The Quest for Competence. University of California Press, Berkeley.
Good, Byron, Fischer, Michael M.J., Willen, Sarah S., Good, Mary-Jo DelVecchio (Eds.), 2010. A Reader in Medical Anthropology: Theoretical Trajectories, Emergent Realities. Wiley-Blackwell.
Gostin, Lawrence O., Gable, Lance, 2004. The human rights of persons with mental disabilities: a global perspective on the application of human rights principles to mental health. Maryland Law Review 63, 20–121.
Hahn, R., 1999. Anthropology in Public Health: Bridging Differences in Culture and Society. Oxford University Press, New York.
Hahn, Robert, Inhorn, Marcia, 2008. Anthropology and Public Health: Bridging Differences in Culture and Society. Oxford University Press, New York.
Hamdy, Sherine, 2012. Our Bodies Belong to God: Organ Transplants, Islam, and the Struggle for Human Dignity in Egypt. University of California Press, Berkeley.
Han, Clara, 2012. Life in Debt: Times of Care and Violence in Neoliberal Chile. University of California Press, Berkeley.
Hernandez, D., Charney, E., 1998. From Generation to Generation: The Health and Well-being of Children in Immigrant Families. National Academy Press, Washington, DC.
Holmes, Seth, 2013. Fresh fruit, broken bodies: migrant farmworkers in the United States. University of Cailfornia Press, Berkeley.
Hyde, Sandra T., 2007. Eating Spring Rice: The Cultural Politics of AIDS in Southwest China. University of California Press, Berkeley and Los Angeles.
Inhorn, Marcia, Wentzell, Emily, 2012. Medical Anthropology at the Intersections: Histories, Activisms, and Futures. Duke University Press, Durham.
Ingstad, Benedicte, Reynolds Whyte, S., 2007. Disability and Culture. University of California Press, Berkeley.
James, Erica, 2010. Democratic Insecurities: Violence, Trauma, and Intervention in Haiti. University of California Press, Berkeley.
Janes, Craig R., Corbett, Kitty K., 2009. Anthropology and global health. In: A Reader in Medical.
Kaufman, Sharon, 2006. And a Time to Die: How American Hospitals Shape the End of Life. University of Chicago Press, Chicago.
Kleinman, Arthur, 1995. Writing at the Margin: Discourse between Anthropology and Medicine. University of California Press, Berkeley.
Kleinman, Arthur, 2009. Caregiving: the odyssey of becoming more human. The Lancet 373 (9660), 292–293.
Kohrman, Matthew, 2005. Bodies of Difference: Experiences of Disability and Institutional Advocacy in the Making of Modern China. University of California Press, Berkeley.
Langwick, Stacey, 2010. Bodies, Politics, and African Healing: The Matter of Maladies in Tanzania. University of California Press, Berkeley.
Lee, Sandra, Koenig, Barbara A., Richardson, Sarah S. (Eds.), 2008. Revisiting Race in a Genomic Age. Rutgers University Press, New Brunswick.
Leslie, C., Young, A., 1992. Paths to Asian Medical Knowledge. University of California Press, Berkeley.
Livingston, Julie, 2012. Improvising Medicine: An African Oncology Ward in an Emerging Cancer Epidemic. Duke University Press, Durham.
Lock, M., 1993. Encounters with Aging: Mythologies of Menopause in Japan and North America. University of California Press, Berkeley.
Lock, Margaret, Nguyen, Vinh-Kim, 2010. An Anthropology of Biomedicine. Wiley-Blackwell.
Manderson, Lenore, Smith-Morris, Carolyn, 2010. Chronic Conditions, Fluid States: Chronicity and the Anthropology of Illness. Rutgers University Press, New Brunswick, NJ.
Martin, Emily, 2009. Bipolar Expeditions: Mania and Depression in American Culture. Princeton University Press, Princeton.
Mattingly, Cheryl, 2010. The Paradox of Hope: Journeys through a Clinical Borderland. University of California Press, Berkeley.
McLean, Athena, 2006. The Person in Dementia: A Study of Nursing Home Care in the US. University of Toronto Press.
Mechanic, D., 1997. The social context of health and disease and choices among health interventions. In: Brandt, A., Rozin, P. (Eds.), Morality and Health. Routledge, New York.
Mol, Annemarie, 2008. The Logic of Care: Health and the Problem of Patient Choice. Routledge.
Montoya, Michael, 2011. Making the Mexican Diabetic: Race, Science, and the Genetics of Inequality. University of California Press, Berkeley.
Nguyen, Vinh-Kim, 2010. Republic of Therapy: Triage and Sovereignty in West Africa's Time of AIDS. Duke University Press, Durham.
Nichter, Mark, 2008. Global Health: Why Cultural Perceptions, Social Representations and Biopolitics Matter. University of Arizona Press, Tucson, AZ.
Norwood, Frances, 2009. The Maintenance of Life: Preventing Social Death through Euthanasia Talk and End-of-Life Care - Lessons from the Netherlands. Carolina Academic Press.
Paul, Benjamin, 1955. Health, Culture, and Community: Case Studies of Public Reactions to Health Programs. Russell Sage Foundation, New York.
Petryna, Adriana, Lakoff, Andrew, Kleinman, Arthur, 2006. Global Pharmaceuticals: Ethics, Markets, Practices. Duke University Press, Durham.
Petryna, Adriana, 2009. When Experiments Travel: Clinical Trials and the Global Search for Human Subjects. Princeton University Press, Princeton.
Petryna, Adriana, 2013. Life Exposed: Biological Citizens after Chernobyl (with New Introduction). Princeton University Press, Princeton.
Phillips, Michael R., Chen, Hanhui, Diesfeld, Kate, Xie, Bin, Cheng, Hui G., Mellsop, Graham, Liu, Xiehe, 2013. China's new mental health law: reframing involuntary treatment. American Journal of Psychiatry 170, 588–591.
Rabinow, Paul, 1996. Essays on the Anthropology of Reason. Princeton University Press, Princeton, Princeton.
Rapp, Rayna, 1999. Testing Women, Testing the Fetus: The Social Impact of Amniocentesis in America. Routledge, New York.
Redfield, Peter, 2013. Life in Crisis: The Ethical Journey of Doctors without Borders. University of California Press, Berkeley.
Rivkin-Fish, Michele, 2005. Women's Health in Post-Soviet Russia: The Politics of Intervention. Indiana University Press, Bloomington.
Roberts, Elizabeth F.S., 2012. God's Laboratory: Assisted Reproduction in the Andes. University of California Press, Berkeley.
Rouse, Carolyn, 2009. Uncertain Suffering: Racial Health Care Disparities and Sickle Cell Disease. University of California Press, Berkeley.
Rylko-Bauer, Barbara, Whiteford, Linda, Farmer, Paul (Eds.), 2009. Global Health in Times of Violence. School for Advanced Research Press, Santa Fe, NM.
Sanal, Aslihan, 2011. New Organs Within Us: Transplants and the Moral Economy. Duke University Press, Durham.
Sharp, Lesley, 2006. Strange Harvest: Organ Transplants, Denatured Bodies, and the Transformed Self. University of California Press, Berkeley.
Scheper-Hughes, Nancy, 1992. Death without Weeping: The Violence of Everyday Life in Brazil. University of California Press, Berkeley.
Singer, Merrill, Derrick Hodge, G. (Eds.), 2010. The War Machine and Global Health. AltaMira Press, Lanham.
Taussig, Karen-Sue, 2009. Ordinary Genomes: Science, Citizenship, and Genetic Identities. Duke University Press, Durham.
Taylor, Janelle, 2008. On recognition, caring, and dementia. Medical Anthropology Quarterly 22 (4), 313–335.
Ticktin, Miriam, 2011. Casualties of Care: Immigration and the Politics of Humanitarianism in France. University of California Press, Berkeley.
Timmermans, Stefan, Buchbinder, Mara, 2012. Saving Babies? The Consequences of Newborn Genetic Screening. University of Chicago Press, Chicago.
Turner, Bryan, 1995. Medical Power and Social Knowledge, second ed. Sage, London.
Waldstein, Anna, 2010. Popular medical knowledge and self-care in a Mexican migrant community: an explanation for an epidemiological paradox. Medical Anthropology 29 (1), 71–107.
Wendland, Claire, 2010. A Heart for the Work: Journeys through an African Medical School. University of Chicago Press, Chicago.
Whyte, Susan R., Geest, Sjaak van der, Hardon, Anita, 2002. Social Lives of Medicines. Cambridge University Press, Cambridge.

Health Behavior, Psychosocial Theories of

Stephen Sutton, Institute of Public Health, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A number of psychosocial theories have been developed to predict, explain, and change health behaviors. In this article, six widely used theories are discussed in turn and then compared. They include four continuum theories (the health belief model; protection motivation theory; the theory of planned behavior; and social cognitive theory), one stage theory (the transtheoretical model), and one theory that has both stage and continuum versions (the health action process approach). Common criticisms of these theories are briefly addressed.

The term 'health behavior' is used very broadly in this article to refer to any behavior that influences, or is believed to influence, physical health outcomes, either by increasing or decreasing their risk or severity. A number of psychosocial theories (or models – the two terms are used interchangeably in this article) have been developed to predict, explain, and change such health behaviors. There are two main types: continuum theories and stage theories (Sutton, 2005; Weinstein et al., 1998). Stage theories assume that behavior change involves movement through a sequence of discrete, qualitatively distinct, stages, and that different factors are important at different stages. Continuum theories have a simpler structure. They specify a single prediction equation for behavior; for example, the health belief model (HBM) specifies that behavior is influenced by perceived susceptibility, severity, benefits, and barriers. The difference in structure of these two types of theories has implications for how they are tested and used as the basis for health behavior interventions. An intervention based on a continuum theory should target the determinants of behavior that are specified by the theory, whereas, for a stage theory, different interventions are required for people in different stages to encourage or help them to move to the next stage.

In this article, six theories are discussed in turn and then compared. They include four continuum theories (the HBM; protection motivation theory (PMT); the theory of planned behavior (TPB); and social cognitive theory (SCT)), one stage theory (the transtheoretical model (TTM)), and one theory that has both stage and continuum versions (the health action process approach (HAPA)). Common criticisms of these theories are briefly addressed.

The Health Belief Model

The HBM (Becker, 1974) was developed in the 1950s by a group of social psychologists working in the field of public health who were seeking to explain why some people do not use health services such as immunization and screening. The model is still in common use. There are four core constructs: the first two refer to a particular disease whereas the second two refer to a possible course of action that may reduce the risk or severity of that disease. Perceived susceptibility (or perceived vulnerability) is the individual's perceived risk of contracting the disease if he or she were to continue with the current course of action. Perceived severity refers to the seriousness of the disease and its consequences as perceived by the individual. Perceived benefits refers to the perceived advantages of the alternative course of action including the extent to which it reduces the risk of the disease or the severity of its consequences. Perceived barriers (or perceived costs) refers to the perceived disadvantages of adopting the recommended action as well as perceived obstacles that may prevent or hinder its successful performance. These factors are commonly assumed to combine additively to influence the likelihood of performing the behavior. Thus, high susceptibility, high severity, high benefits, and low barriers are assumed to lead to a high probability of adopting the recommended action. Another factor that is frequently mentioned in connection with the HBM is cues to action (events that trigger behavior), but little empirical work has been conducted on this construct.

There have been several meta-analyses (quantitative reviews) of research using the HBM. The most recent of these (Carpenter, 2010) was the most rigorous and focused on 18 studies that used a longitudinal design in which the HBM variables were measured at time 1 and behavior at time 2. Benefits and barriers were consistently the strongest predictors of behavior, with mean correlations of 0.27 and 0.30, compared with 0.15 for severity and only 0.05 for susceptibility. Several moderators were identified, including the length of time between the measurement of beliefs and behavior (the longer the time, the lower the correlation) and prevention versus treatment behaviors (higher correlations for prevention). Carpenter recommends that researchers should abandon the simple four-variable additive model and instead examine possible mediation and moderation among the variables.

Protection Motivation Theory

PMT (Rogers, 1983; Rogers and Prentice-Dunn, 1997) was originally developed to explain how people respond to fear-arousing health threat communications or 'fear appeals.' It can be regarded as an adaptation of the HBM. Protection motivation refers to the motivation to protect oneself against a health threat; it is usually defined operationally as the intention to adopt the recommended action. Of the determinants of intention specified by the model, the four that have received the most

empirical attention are vulnerability and severity (equivalent to perceived susceptibility and severity in the HBM), response efficacy (the belief that the recommended action is effective in reducing the threat), and perceived self-efficacy (the belief that one can successfully perform the recommended action; Bandura, 1997). Thus, a person will be more motivated to protect himself or herself (i.e., have a stronger intention to adopt the recommended action) to the extent that he or she believes that the threat is likely if the current course of action is continued, that the consequences will be serious if the threat occurs, that the recommended action is effective in reducing the likelihood or the severity of the threat, and that he or she is able to carry out the recommended action.

In many studies using this model (e.g., Wurtele and Maddux, 1987), specific PMT variables are experimentally manipulated in a factorial design and their effects on intention (and sometimes behavior) are measured. In fact, PMT is unique among health behavior theories with respect to the relatively large number of experimental tests that have been conducted.

To date, two meta-analyses of PMT studies have been conducted (Floyd et al., 2000; Milne et al., 2000). The analyses used different study inclusion criteria and different effect size measures. Floyd et al. analyzed 65 studies with about 30 000 research participants, whereas Milne and colleagues included 27 studies with about 8000 participants. There were only 12 studies in common. Both analyses found support for each of the main PMT variables as predictors of intentions and/or behavior. Self-efficacy had the strongest, most consistent, and most robust effect.

The Theory of Planned Behavior

The TPB (Ajzen, 1991, 2006b), an extension of the theory of reasoned action (TRA), developed out of social–psychological research on attitudes and the attitude–behavior relationship. According to the theory, behavior is determined by the strength of the person's intention to perform that behavior and the amount of actual control that the person has over performing the behavior. Perceived behavioral control, similar to Bandura's construct of self-efficacy, refers to the person's perceptions of their ability to perform the behavior and is assumed to reflect actual behavioral control more or less accurately. To the extent that perceived behavioral control is an accurate reflection of actual behavioral control, it can, together with intention, be used to predict behavior.

The strength of a person's intention is determined by three factors: their attitude toward the behavior (their overall evaluation of performing the behavior); their subjective norm (the extent to which they think that important others would want them to perform it); and their perceived behavioral control. Attitude, subjective norm, and perceived behavioral control are each held to be determined by sets of salient (readily accessible) beliefs about the behavior. Attitude toward the behavior is determined by behavioral beliefs about the personal consequences of performing the behavior; subjective norm by normative beliefs about the views of important others; and perceived behavioral control by control beliefs about the presence of factors that may facilitate or impede performance of the behavior.

The TPB has been widely used in observational studies to predict and explain intentions and behavior. Meta-analyses of these studies show that, on average, the TPB explains between 35 and 50% of the variance in intentions and between 26 and 35% of the variance in behavior (Sutton, 2004). To date, the theory has not been widely used to develop behavior change interventions (Sutton, 2010).

Social Cognitive Theory

According to SCT (Bandura, 1986, 1997), behavior is influenced directly by goals and self-efficacy expectations and indirectly by self-efficacy, outcome expectations, and sociostructural factors. Goals are similar to intentions in other theories of health behavior; they determine the amount of effort that the individual invests in changing their behavior and serve as guides to action. Self-efficacy refers to a person's belief in their ability to perform a specific action in a particular situation. Outcome expectations are beliefs regarding the consequences (positive and negative) of performing the behavior. The theory distinguishes between different kinds of outcome expectation. For example, social outcome expectations might include anticipated approval or disapproval from one's partner, whereas self-evaluative outcome expectations refer to anticipated feelings arising from internal standards such as pride in having achieved a change in one's behavior. Finally, the theory incorporates perceived opportunities and barriers, which are assumed to influence goals.

Although SCT includes a number of constructs, the majority of empirical applications of the theory focus on self-efficacy. There is a large body of evidence from observational studies showing that self-efficacy expectations consistently predict behavior (Luszczynska and Schwarzer, 2005). The theory has also been used as the basis for behavior change interventions, again focusing almost exclusively on self-efficacy. According to Bandura (1997), there are several ways of enhancing self-efficacy. These include mastery experiences, in which the person gains confidence by achieving a modest goal; observation of someone similar to themselves successfully performing a behavior, known as modeling or vicarious learning; and verbal persuasion.

The Health Action Process Approach

The HAPA (Schwarzer, 2008) exists in both continuum and stage versions and is sometimes described as a 'hybrid' model. The model incorporates several constructs from other theories of health behavior, including risk perception, outcome expectancies, self-efficacy, and intention. Unlike other theories, it specifies different types of self-efficacy. For example, maintenance (or coping) self-efficacy refers to optimistic beliefs about one's ability to deal with barriers that arise during the maintenance period, and recovery self-efficacy refers to self-efficacy to get back on track after experiencing a failure, lapse, or setback. Another HAPA construct that does not occur in the other theories discussed in this article is action planning (or action plans), which refers to having

made a detailed plan regarding when, where, and how to carry out the target behavior.

Different versions of the model have been tested in longitudinal observational studies. In a typical application of the HAPA model, the constructs are measured on two or three occasions and structural equation modeling is used to test a prespecified causal model. For example, Schwarzer and Luszczynska (2008; Study 2) measured risk perception, positive outcome expectancies, pre-action self-efficacy and intention at time 1, and maintenance self-efficacy, planning, recovery self-efficacy, and behavior (high-fat diet) at time 2, 2 months later. Among other links, the model specified that risk perception, positive outcome expectancies, and pre-action self-efficacy directly influenced intention to reduce fatty foods; planning mediated the intention–behavior relationship; and planning and recovery self-efficacy directly influenced behavior.

These models, and the findings from such model tests, are often interpreted in terms of phases. For example, the part of the model that deals with determinants of intention is referred to as the motivational phase and the mediation of the intention–behavior link is referred to as the postintentional or volitional phase. However, the HAPA as specified and tested in such studies is a continuum theory, similar in structure to the TPB, for example (Sutton, 2005, 2008).

Several recent studies have specified the HAPA as an explicit stage theory. For example, in a study of fruit and vegetable intake, Wiedemann et al. (2009) used a staging algorithm to categorize participants into three stages (pre-intenders, intenders, and actors) and examined predictors of stage transitions between two time points.

The Transtheoretical Model

The TTM (Prochaska et al., 1992) is the dominant stage theory in the health behavior field. It developed from studies of the processes of change in psychotherapy and smoking cessation. Smoking still accounts for the majority of applications of the model but it has been applied to a wide range of other health behaviors, including condom use, exercise, sunscreen use, and healthy eating (Prochaska and Velicer, 1997). Although it is often referred to simply as the stages of change model, the TTM includes 15 different theoretical constructs drawn from different theories of behavior change. These include the stages of change (which provide the basic organizing principle), the 10 processes of change, the perceived pros and cons of changing, and self-efficacy and temptation. The TTM was an attempt to integrate these different constructs in a single comprehensive framework – hence the name transtheoretical.

The version of the TTM used most widely in recent years specifies five stages: pre-contemplation, contemplation, preparation, action, and maintenance (DiClemente et al., 1991). Using a 'staging algorithm,' participants are classified into stages on the basis of their responses to a small number of questionnaire items. Prochaska et al. (1992) represented the stages of change as a spiral. People start at the bottom in pre-contemplation. They then move through the stages in order (contemplation, preparation, action, maintenance) but may relapse to an earlier stage. They may cycle and recycle through the stages several times before achieving successful long-term behavior change.

Of the other constructs in the TTM, the 10 processes of change refer to the things that people think and do to help them move through the stages, the pros and cons are the perceived advantages and disadvantages of changing, self-efficacy is borrowed from Bandura's SCT, and temptation is a related concept that refers to the degree to which a person expects to feel tempted to lapse in different situations.

Although the TTM is still enormously popular, it has attracted a lot of criticism in recent years, culminating in a call for the model to be abandoned (West, 2005). The problems include lack of standardization of measures, particularly of the central construct of stages of change; logical flaws in current staging algorithms; inadequate specification of the causal relationships among the different constructs; misinterpretation of cross-sectional data on stages of change; and confusion concerning the nature of stage models and how they should be tested. (Herzog, 2008; Sutton, 2005). However, although the TTM cannot be recommended in its present form, the notion that behavior change involves movement through a sequence of qualitative stages is an important idea that deserves further consideration.

Comparison of Theories

Aside from the distinction between stage and continuum theories, the theories outlined earlier show a number of important similarities and differences. Some constructs are common to more than one theory. For example, perceived susceptibility or perceived vulnerability occurs in the HBM, PMT, and the HAPA. Other constructs appear to be very similar, for example, perceived behavioral control and self-efficacy. Resolution of current controversies concerning the extent of overlap between such constructs requires the development of clear definitions, so that similar constructs can be distinguished on conceptual grounds, and more frequent tests of discriminant validity to investigate whether sets of apparently similar measures are tapping the same or different constructs.

The theories also differ in the extent to which the constructs and the causal relationships among them are clearly specified. Of the four, the TPB is the most highly specified (though there are still some ambiguities; Sutton, 2004). Another advantage of the TPB is that there exist clear recommendations for how the constructs should be operationalized (Ajzen, 2006a). In addition, Ajzen emphasizes the principle of compatibility. Put simply, this states that for maximum prediction the measures of all the constructs in the model should use the same wording to define the target behavior. This principle is not widely applied in research using the other theories. The other theories lack clarity in important respects. The TTM, for example, lacks a clear specification of the variables that influence each stage transition.

The theories also differ with regard to their scope of application. Key constructs in the HBM and PMT include perceived susceptibility and perceived severity with respect to a given health threat. Although these components can be extended to non-health-related events, for example, the risk of financial

loss, the scope of these theories is necessarily limited by the nature of these two constructs. By contrast, SCT, the TPB, and the TMM are general theories that can be applied to any domain or behavior, for example, voting, career, and purchasing decisions. Stated differently, these theories regard health behaviors (or stage transitions in the case of the TTM) as having the same proximal determinants as other kinds of behavior. Thus, they offer the potential benefit of parsimony.

The theories differ in the degree to which they specify the content of the cognitions they identify. With the TPB, for example, once the behavior of interest has been defined, it is possible to generate questionnaire items for intention and for the direct measures of attitude, subjective norm, and perceived behavioral control. However, in order to generate items for the underlying beliefs, it is recommended that researchers gather information on salient beliefs from members of the target population. These constructs remain 'content-free' until such information is obtained.

Theories of health behavior are often criticized for offering an unrealistically rational account of how people form intentions and make decisions. However, the theories do not imply that individuals always deliberate carefully and always make optimal decisions. People may not be aware of all the options available to them and of all the consequences that may follow from their actions. They may hold incorrect beliefs about the outcomes. They may make rapid decisions based on a few salient considerations. Having made a considered decision (e.g., to go jogging every Sunday morning), they do not necessarily have to weigh up the pros and cons again unless circumstances change; they may simply retrieve their previously formed intention from long-term memory and act on it. Thus, theories of health behavior imply a more limited rationality than is sometimes suggested by their critics.

Theories of health behavior, particularly continuum theories, are also sometimes criticized for being static. This criticism is unfounded: the theories summarize dynamic causal processes. In SCT, for example, changes in self-efficacy and outcome expectations are assumed to produce changes in goals which in turn lead to changes in behavior. Theories of health behavior should specify the time lags involved in these causal processes but most of them do not. However, it is often assumed implicitly that effects on intention are almost instantaneous whereas effects on behavior may be delayed.

Concluding Comments

There is a plethora of theories of health behavior and new theories are continually being developed. Although this can be interpreted as a sign of a field rich in conceptual and theoretical development, it makes it increasingly difficult to accumulate research findings into a coherent body of knowledge. The field would benefit from clearer definition of concepts, greater standardization of measures, more tests of convergent and discriminant validity, greater concentration on a small number of theories, and more empirical comparisons of theories. The vast majority of empirical studies use nonexperimental between-individuals designs. Obtaining repeated measures on the same individuals would allow models to be tested at the individual level. Experimental manipulation of constructs, as has been done in many studies of PMT, would provide stronger evidence of causal effects (Sutton, 2002).

See also: Control Beliefs: Health Perspectives; Health Behaviors, Assessment of; Health Behaviors; Health Psychology; Health Risk Appraisal and Optimistic Bias; Health Risk Perception; Illness Behavior and Care-Seeking; Patient Adherence to Health Care Regimens; Self-Efficacy and Health.

Bibliography

Ajzen, I., 1991. The theory of planned behavior. Organizational Behavior and Human Decision Processes 50, 179–211.

Ajzen, I., 2006a. Constructing a TPB Questionnaire. http://www.people.umass.edu/aizen/tpb.html (accessed 01.07.13.).

Ajzen, I., 2006b. Theory of Planned Behavior. http://www.people.umass.edu/aizen/tpb.html (accessed 01.07.13.).

Bandura, A., 1986. Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice-Hall, Englewood Cliffs, NJ.

Bandura, A., 1997. Self-efficacy: The Exercise of Control. Freeman, New York.

Becker, M.H. (Ed.), 1974. The Health Belief Model and Personal Health Behavior. Slack, Thorofare, NJ.

Carpenter, C.J., 2010. A meta-analysis of the effectiveness of the health belief model variables in predicting behavior. Health Communication 25, 661–669.

DiClemente, C.C., Prochaska, J.O., Fairhurst, S.K., Velicer, W.F., Velasquez, M.M., Rossi, J.S., 1991. The process of smoking cessation: an analysis of pre-contemplation, contemplation, and preparation stages of change. Journal of Consulting and Clinical Psychology 59, 295–304.

Floyd, D.L., Prentice-Dunn, S., Rogers, R.W., 2000. A meta-analysis of research on protection motivation theory. Journal of Applied Social Psychology 30, 407–429.

Herzog, T.A., 2008. Analyzing the transtheoretical model using the framework of Weinstein, Rothman, and Sutton (1998): the example of smoking cessation. Health Psychology 27, 548–556.

Luszczynska, A., Schwarzer, R., 2005. Social cognitive theory. In: Conner, M., Norman, P. (Eds.), Predicting Health Behaviour: Research and Practice with Social Cognition Models, second ed. Open University Press, Maidenhead, UK, pp. 127–169.

Milne, S., Sheeran, P., Orbell, S., 2000. Prediction and intervention in health-related behavior: a meta-analytic review of protection motivation theory. Journal of Applied Social Psychology 30, 106–143.

Prochaska, J.O., DiClemente, C.C., Norcross, J.C., 1992. In search of how people change: applications to addictive behaviors. American Psychologist 47, 1102–1114.

Prochaska, J.O., Velicer, W.F., 1997. The transtheoretical model of health behavior change. American Journal of Health Promotion 12, 38–48.

Rogers, R.W., 1983. Cognitive and physiological processes in fear appeals and attitude change: a revised theory of protection motivation. In: Cacioppo, J.T., Petty, R.E., Shapiro, D. (Eds.), Social Psychophysiology: A Sourcebook. Guilford Press, New York, pp. 153–176.

Rogers, R.W., Prentice-Dunn, S., 1997. Protection motivation theory. In: Gochman, D. (Ed.), Handbook of Health Behavior Research, Determinants of Heath Behavior: Personal and Social, vol. 1. Plenum, New York, pp. 113–132.

Schwarzer, R., 2008. Modeling health behavior change: how to predict and modify the adoption and maintenance of health behaviors. Applied Psychology: An International Review 57, 1–29.

Schwarzer, R., Luszczynska, A., 2008. How to overcome health-compromising behaviors: the health action process approach. European Psychologist 13, 141–151.

Sutton, S., 2002. Testing attitude-behaviour theories using non-experimental data: an examination of some hidden assumptions. European Review of Social Psychology 13, 293–323.

Sutton, S., 2004. Determinants of health-related behaviours: theoretical and methodological issues. In: Sutton, S., Baum, A., Johnston, M. (Eds.), The Sage Handbook of Health Psychology. Sage, London, pp. 94–126.

Sutton, S., 2005. Stage theories of health behaviour. In: Conner, M., Norman, P. (Eds.), Predicting Health Behaviour: Research and Practice with Social Cognition Models, second ed. Open University Press, Maidenhead, UK, pp. 223–275.

Sutton, S., 2008. How does the Health Action Process Approach (HAPA) bridge the intention–behavior gap? An examination of the model's causal structure. Applied Psychology: An International Review 57, 66–74.

Sutton, S., 2010. Using social cognition models to develop health behaviour interventions: the theory of planned behaviour as an example. In: French, D., Vedhara, K., Kaptein, A.A., Weinman, J. (Eds.), Health Psychology, second ed. BPS Blackwell, Oxford, pp. 122–134.

Weinstein, N.D., Rothman, A.J., Sutton, S.R., 1998. Stage theories of health behavior: conceptual and methodological issues. Health Psychology 17, 290–299.

West, R., 2005. Time for a change: putting the transtheoretical (stages of change) model to rest. Addiction 100, 1036–1039.

Wiedemann, A.U., Lippke, S., Reuter, T., Schüz, B., Ziegelmann, J.P., Schwarzer, R., 2009. Prediction of stage transitions in fruit and vegetable intake. Health Education Research 24, 596–607.

Wurtele, S.K., Maddux, J.E., 1987. Relative contributions of protection motivation theory components in predicting exercise intentions and behavior. Health Psychology 6, 453–466.

Health Behaviors

Mark Conner, Institute of Psychological Sciences, University of Leeds, Leeds, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article defines health behaviors and then overviews data on key health behaviors (smoking, diet, exercise/physical activity, health screening, sexual behaviors, and alcohol use). Variations in these behaviors by sociodemographic factors are then considered. The next section addresses the psychological determinants of health behaviors and considers key models, such as the theory of planned behavior, health belief model, and social cognitive theory, and how these models might be integrated. The final section discusses how health behaviors might be changed to improve health outcomes.

Introduction

Much of the interest in behaviors that have important impacts on our health and well-being is based upon two assumptions: that a significant proportion of the mortality from the leading causes of death is caused by the behavior of individuals, and that such behavior is modifiable (Conner and Norman, 2005). Behavior is held to exert its influence on health in three basic ways: by producing direct biological changes; by conveying health risks or protecting against them; or by leading to the early detection or treatment of disease (Baum and Posluszny, 1999).

The Definition of Health Behaviors

Health behaviors have been defined in a variety of ways. For example, Conner and Norman (2005) define them as any activity undertaken for the purpose of preventing or detecting disease, or for improving health and well-being. Gochman (1997) in the *Handbook of Health Behavior Research* defines them as "… behavior patterns, actions and habits that relate to health maintenance, to health restoration and to health improvement" (vol. 1: p. 3). Behaviors within these definitions include medical service usage (e.g., physician visits, vaccination, screening), compliance with medical regimens (e.g., dietary, diabetic, antihypertensive regimens), and self-directed health behaviors (e.g., diet, exercise, smoking, alcohol consumption). Each has received considerable attention from social and behavioral researchers and we now have a growing understanding of the factors determining engagement in such behaviors and ways in which such behavior can be changed.

In describing health behaviors it is common to distinguish health-enhancing from health-impairing behaviors. Health-impairing behaviors have harmful effects on health or otherwise predispose individuals to disease and include smoking, excessive alcohol consumption, and high dietary fat consumption. In contrast, engagement in health-enhancing behaviors convey health benefits or otherwise protect individuals from disease and include physical activity and exercise, fruit and vegetable consumption, and condom use in response to the threat of sexually transmitted diseases. A third set of health behaviors focus on detecting potential health problems and include behaviors such as health screening and testicular/breast self-examination.

Key Health Behaviors

Numerous studies have examined the health-behavior–health-outcome relationship. One of the first such studies identified seven features of lifestyle which were associated with lower morbidity and higher subsequent long-term survival: not smoking, moderate alcohol intake, sleeping 7–8 h per night, exercising regularly, maintaining a desirable body weight, avoiding snacks, and eating breakfast regularly (Belloc and Breslow, 1972). Health behaviors also impact upon individuals' quality of life by delaying the onset of chronic disease and extending healthy life span. Smoking, alcohol consumption, diet, gaps in primary care services, and low screening uptake are all significant determinants of poor health, and changing such behaviors should lead to improved health. Health recommendations in countries across the Western world currently emphasize the importance of an increase in fruit and vegetable consumption, a reduction in dietary fat consumption, increased physical activity, and reductions in tobacco, alcohol, and drug use as important for health promotion and disease prevention.

Smoking

Smoking is the health behavior most closely linked with long-term negative health outcomes. Morbidity and mortality from coronary heart disease (CHD) are increased among smokers (Doll et al., 1994). Moreover, there is a strong positive relationship between the number of cigarettes smoked per day and the incidence of CHD (Friedman et al., 1979). Smoking has also been linked to a number of cancers including cancer of the lung, throat, stomach, and bowel as well as a number of more immediate negative health effects such as reduced lung capacity and bronchitis (Royal College of Physicians, 1983). Despite the array of negative health outcomes, smokers often report positive mood effects from smoking and the use of smoking as a strategy for coping with stress.

A variety of studies support the idea that the vast majority of smokers take up this habit as adolescents with 40% of adult smokers having started before they reached 16 years of age (Royal College of Physicians, 1992). The number of people smoking in the USA and UK has shown a steady decline over the past 20 years from over 30% in 1999 to below 20% now (Center for Diseases Control, 2008). Smoking is higher among less educated, lower income, and minority groups (Rigotti,

1989). Those who quit smoking reduce the risk to their health, particularly if they quit before 35 years of age (Doll et al., 1994).

Diet

The impacts of diet upon morbidity and mortality are well established (e.g., USDHHS, 1988). In the Third World, the problems related to diet and health are ones of undernutrition; in the First World, the problems are predominantly linked to overconsumption of food. In Western industrialized countries, excessive fat consumption and insufficient fiber, fruit, and vegetable consumption are related to health problems. In addition, excess consumption of calories combined with insufficient exercise has made obesity a major health problem. Diet has been implicated in cardiovascular diseases (CVDs), stroke and high blood pressure, cancer, diabetes, obesity, osteoporosis, and dental disease.

It is generally agreed that elevated blood cholesterol level is a major risk factor for the development of CVD (Consensus Development Conference on Lowering Blood Cholesterol to Prevent Heart Disease, 1985) with significant proportions of the population in countries in the Western world having cholesterol levels higher than the level considered to be healthy. For example, in the USA, it is estimated that 50% of the adult population is at risk of CHD due to elevated blood cholesterol levels (Sampos et al., 1989). The reduction of blood cholesterol via dietary change is now widely accepted as an important way of tackling CHD. Dietary recommendations include reducing fat in the diet and increasing soluble fiber intake. However, their impact upon cholesterol levels may be limited.

Physical Activity/Exercise

The potential health benefits of engaging in physical activity and regular exercise include reduced cardiovascular morbidity and mortality, lowered blood pressure, and the increased metabolism of carbohydrates and fats, as well as a range of psychological benefits such as improved self-esteem, positive mood states, and reduced life stress and anxiety. Nevertheless, many adults in Europe (Office for National Statistics, 2006) and the US (USDHHS, 1996) fail to meet physical activity recommendations. For example, the General Household Survey (1989) indicated that only one in three men and one in five women in the UK participate in any sport or recreational physical activity. Moreover, the Allied Dunbar Fitness Survey (1992) of 6000 English adults reported that only one in six adults had engaged in any physical activity that would have been likely to benefit their health (i.e., for 20 min or more at a moderate or vigorous level) in the previous 4 weeks. Participation in regular physical activity and exercise is strongly related to a number of sociodemographic variables. In particular, young people, males, and those from higher socioeconomic status groups are more likely to engage in regular physical activity and exercise.

Health Screening

Individuals may seek to protect their health by participating in various screening programs, which attempt to detect disease at an early, or asymptomatic, stage. In the UK, screening programs have been set up for various diseases including anemia, diabetes, bronchitis, and various cancers (e.g., cervical, bowel, and breast cancers). Taking the example of cervical screening (or PAP testing), it is estimated that if women were screened every 3 years, the cervical cancer mortality rate could be reduced by 70–95% (Greenwald and Sondik, 1986). Research from the UK shows that between 10 and 32% of eligible women remain unscreened (National Audit Office, 1998), while in the USA the number of unscreened women ranges from 13 to 30% (Ruchlin, 1997). Participation tends to be negatively related to age, and positively related to education level and socioeconomic status.

Sexual Behaviors

Sexual behaviors are considered health behaviors because of their impact upon the spread of sexually transmitted infections such as gonorrhea and syphilis. More recently, the role of sexual behaviors in the spread of the human immunodeficiency virus (HIV) has been a focus of attention. While early health education campaigns emphasized the need to reduce the number of sexual partners or avoid particular sexual practices (e.g., anal sex, penetrative sex), more recently the focus has been upon the use of condoms during penetrative sex to reduce the risk of HIV transmission. Condom use is particularly recommended for those with multiple partners or those who do not know their partners' sexual history. For these reasons, much of the health advice concerning condom use has been focused on young people.

There seems to be considerable variation in the use of condoms in response to the threat of HIV/AIDS, although systematic condom use is not widespread even among single heterosexuals. For example, 78% of respondents in an American study declared they did not always use a condom during sexual intercourse (Choi and Catania, 1996). A Canadian study among a population aged 15 years and older reported similar findings with 28% of respondents reporting not having used a condom during their last sexual intercourse with an occasional partner (Health Canada, 1998). In a European study, among respondents who had more than one partner in the past year, only 52% of the men and 41% of the women declared having used a condom at least once (Guiguet et al., 1994).

Alcohol Use

Moderate alcohol consumption has been linked to positive health outcomes. However, high alcohol consumption has been linked to a range of negative health outcomes including high blood pressure, heart disease, and cirrhosis of the liver. High levels of alcohol consumption have also been associated with accidents, injuries, suicides, crime, domestic violence, rape, murder, and unsafe sex (British Medical Journal, 1982). While many of the adverse effects of high alcohol consumption are due to continued heavy drinking (e.g., cirrhosis of the liver, heart disease), others are more specifically related to excessive alcohol consumption in a single drinking session (e.g., accidents, violence).

The General Household Survey (1992) reported that the average weekly consumption of alcohol in the UK was 15.9

units (1 unit = 1 glass of wine or 1 measure of spirits or 0.5 pints of beer) for men and 5.4 for women. In addition, 27% of men and 11% of women were drinking more than the recommended weekly sensible limits (21 units for men, 14 units for women). Heavy drinking is more likely among younger age groups. In a survey of 12 000 Welsh adults, Moore et al. (1994) reported that 31.1% of drinkers aged 18–24 engaged in binge-drinking (i.e., drinking half the recommended weekly consumption of alcohol in a single session) at least once a week.

Relationship of Health Behaviors to Sociodemographic Factors

A clearer understanding of why individuals perform health behaviors might assist in the development of interventions to help individuals gain health benefits. A variety of factors have been found to account for individual differences in the performance of health behaviors. Demographic variables show reliable associations with the performance of health behaviors (e.g., age, gender, ethnic status). For example, there is a nonlinear relationship between many health behaviors and age, with high incidences of many health-risking behaviors such as smoking in young adults and much lower incidences in children and older adults (Blaxter, 1990). Such behaviors also vary by gender, with females being generally less likely to smoke, consume large amounts of alcohol, engage in regular exercise, but more likely to monitor their diet, take vitamins, and engage in dental care (Waldron, 1988). Differences by socioeconomic status are also apparent for behaviors such as diet, exercise, alcohol consumption, and smoking (e.g., Blaxter, 1990) with increases in health-risking behaviors and decreases in health-protective behaviors in lower socioeconomic status groups.

Generally speaking, younger, wealthier, and better-educated individuals are more likely to practice health protective behaviors. Access to medical care has been found to influence the use of such health services (e.g., Black Report, 1988) and may explain some socioeconomic status differences in health behaviors. These factors may interact with other influences such as levels of stress and social support. Higher levels of stress and/or fewer resources are associated with health-risking behaviors such as smoking and alcohol abuse (Adler and Matthews, 1994). Social factors seem to be important in establishing health behaviors in childhood. Parent, sibling, and peer influences are important, for example, in the initiation of smoking. Cultural values also have a major impact, for instance, in determining the number of women exercising in a particular culture. Steptoe and Wardle (1992) report that between 34 and 95% of women in their European student sample had exercised in the past 14 days.

Understanding the Distribution/Prevalence of Health Behaviors

A considerable body of social and behavioral research has examined individual variables explaining sociodemographic differences in performance of health behaviors. An assumption is that these variables may be more readily modifiable in attempts to change behavior. For example, personality variables such as conscientiousness have been related to mortality. Conscientiousness refers to the ability to control one's behavior and to complete tasks. Highly conscientious individuals are more organized, careful, dependable, self-disciplined, and achievement-oriented than those low in conscientiousness. There is considerable evidence showing conscientiousness to impact on health behaviors, health outcomes, and even mortality. Friedman et al. (1993) reported that conscientiousness was significantly associated with lower mortality in later life with those high in conscientiousness likely to live longer (by about 2 years), compared to those low in conscientiousness. An important mechanism by which conscientiousness may influence health is through health behaviors. Friedman et al. (1995) showed that the impact of conscientiousness on longevity was partly accounted for by its effect on reducing smoking and alcohol use. A review of work on the relationship between conscientiousness and behavior (Bogg and Roberts, 2004) showed conscientiousness to be positively related to a range of protective health behaviors (e.g., exercise) and negatively related to a range of risky health behaviors (e.g., smoking).

Thoughts and feelings about a health behavior (or health cognitions) also determine whether or not an individual practices health behaviors and may explain how other factors influence behavior. Knowledge about behavior–health links is an important factor in an informed choice concerning health behaviors. Various health cognitions have been studied including perceptions of health risk, efficacy of behaviors in influencing this risk, social pressures to perform the behavior, and control over performance of the behavior. The relative importance of various health cognitions in determining who performs various health behaviors constitutes the basis of several different models. Such models have been labeled social cognition models (SCMs) because of their focus on health cognitions as the primary determinant of individual social behaviors. These SCMs provide a basis for understanding the determinants of behavior and behavior change. Each of these models emphasizes the rationality of human behavior and assumes that behavior is based upon elaborate, but subjective, cost/benefit analysis of the likely outcomes of differing courses of action. It is assumed that individuals generally aim to maximize benefits and minimize costs in selecting a behavior. The effects of sociodemographic variables on health behavior are mediated by health cognitions in these models, although direct and moderated relationships are reported (Conner et al., 2013).

Health Belief Model

The health belief model (HBM) outlines two types of health beliefs that make a behavior in response to illness more or less likely (Abraham and Sheeran, 2005): perceptions of the threat of illness and evaluation of the effectiveness of behaviors to counteract this threat. Threat perceptions depend upon the perceived susceptibility to the illness and the perceived severity of the consequences of the illness. Together these variables determine the likelihood of the individual following a health-related action, although their effect is modified by

demographic variables, social pressure, and personality. The particular action undertaken is determined by the evaluation of the possible alternatives. This behavioral evaluation depends upon beliefs concerning the benefits or efficacy of the health behavior and the perceived costs or barriers to performing the behavior. Hence, individuals are likely to follow a particular health behavior if they believe themselves to be susceptible to a particular condition or illness, which they consider to be serious, and believe the benefits of the behavior undertaken to counteract the condition or illness outweigh the costs. It is assumed that this whole process is set in motion by cues to action. Cues to action include a diverse range of triggers to the individual taking action and are commonly divided into factors that are internal (e.g., physical symptoms) or external (e.g., mass media campaigns, advice from others) to the individual. Other influences upon the performance of health behaviors, such as demographic factors or psychological characteristics (e.g., personality, peer pressure, perceived control over behavior), are assumed to exert their effect via changes in the components of the HBM.

Theory of Planned Behavior

The theory of planned behavior (TPB) was developed by social psychologists and has been widely employed to health behaviors (Ajzen, 1991; McEachan et al., 2011). The TPB specifies the influences that determine the individual's decision to follow a particular behavior. Within the TPB, the determinants of behavior are intentions to engage in that behavior and perceived behavioral control (PBC) over that behavior. Intentions represent a person's motivation. The construct is conceptualized as an individual's conscious plan or decision to exert effort in order to engage in a particular behavior. PBC is a person's expectancy that performance of the behavior is within his or her control or confidence that he or she can perform the behavior. Intentions are determined by three variables. The first is attitudes, which are an individual's overall evaluation of the behavior. The second is subjective norms, which consist of a person's beliefs about whether significant others think he or she should engage in the behavior. The third is PBC. Recent revisions to the TPB have suggested that attitudes may divide into instrumental and affective components and subjective norms may divide into injunctive and descriptive norms while PBC may divide into perceived confidence and perceived control (Conner and Sparks, 2005).

The attitude, subjective norm, and PBC components are determined by underlying beliefs. Attitude is a function of a person's salient behavioral beliefs, which represent perceived likely consequences of the behavior (e.g., taking exercise will reduce my risk of heart disease). Subjective norm is a function of normative beliefs, which represent perceptions of specific salient others' preferences about whether one should or should not engage in a behavior (e.g., my family think I should take exercise). PBC is based on beliefs concerning access to the necessary resources and opportunities to perform the behavior successfully (e.g., I have easy access to a place where I can exercise). So, according to the TPB, individuals are likely to engage in a health behavior if they believe that the behavior will lead to particular outcomes which they value, if they believe that people whose views they value think they should carry out the behavior, and if they feel that they have the necessary resources and opportunities to perform the behavior.

Social Cognitive Theory

In social cognitive theory (SCT; Bandura, 1982), behavior is held to be determined by four factors: goals, outcome expectancies, self-efficacy, and sociostructural variables. Goals are plans to act and can be conceived of as intentions to perform the behavior (see Luszczynska and Schwarzer, 2005). Outcome expectancies are similar to behavioral beliefs in the TPB but here are split into physical, social, and self-evaluative depending on the nature of the outcomes considered. Self-efficacy is the belief that a behavior is or is not within an individual's control and is usually assessed as the degree of confidence the individual has that he or she could still perform the behavior in the face of various obstacles (and is similar to PBC in the TPB; e.g., "I am confident that I can refrain from smoking, even if someone offers me a cigarette"). Sociostructural variables are factors assumed to facilitate or inhibit the performance of a behavior and affect behavior via changing goals (e.g., impediments or opportunities associated with particular living conditions, health systems, political, economic, or environmental systems). They are assumed to inform goal setting and be influenced by self-efficacy.

Self-efficacy is one of the most powerful predictors of health behavior (Bandura, 1997). Individuals with a strong sense of self-efficacy are believed to develop stronger intentions to act, to expend more effort to achieve their goals, and to persist longer in the face of barriers and impediments. Perceived self-efficacy is therefore believed to play a crucial role in the determination of health behavior.

Integrated Models of Health Behavior

Models like the HBM, TPB, and SCT show several similarities but also a number of key differences. One approach to combining these models is to consider developing an integrated model. This may be valuable given the overlap in cognitions considered by different models. For example, it is widely accepted that the key cognitions influencing behavior are intention, self-efficacy, and outcome expectancies (or attitudes). One attempt at integration suggested that there were eight key variables determining behavior (Fishbein et al., 2001). The variables are organized into two groups. The first group includes variables viewed as necessary and sufficient determinants of behavior. For behavior to occur an individual must:

- have a strong intention to perform the behavior;
- have the necessary skills to perform the behavior; and
- experience an absence of environmental constraints that could prevent enactment of the behavior.

The second group of variables primarily influences behavior through intention. A strong intention to act is likely to occur when an individual:

- perceives the advantages (or benefits) of performing the behavior to outweigh the perceived disadvantages (or costs);

- perceives the social (normative) pressure to perform the behavior to be greater than that not to perform the behavior;
- believes that the behavior is consistent with his or her self-image;
- anticipates the emotional reaction to performing the behavior to be more positive than negative; and
- has high levels of self-efficacy.

Changing Health Behaviors

The above models detail the key social cognitive determinants of health behavior. To the extent that these models outline the key determinants, interventions, which target these variables should lead to associated changes in behavior. For example, enhancing feelings of self-efficacy could be one means to encourage health behavior change. As Bandura (1997) outlines, there are four main sources of self-efficacy, each of which could be addressed in interventions. First, individuals can develop feelings of self-efficacy from personal mastery experience (e.g., it may be possible to split a behavior into various subgoals, such that the easiest subgoals are achieved before more difficult tasks are attempted). Second, individuals may develop feelings of self-efficacy through observing other people succeed on a task (i.e., vicarious experience). Third, it is possible to use standard persuasive techniques to try to instill feelings of self-efficacy. Finally, one's physiological state may be used as a source of information, such that high levels of arousal or anxiety may indicate to the individual that he or she is not capable of performing a given action (e.g., relaxation techniques may be employed to help maintain feelings of self-efficacy). Alternatively interventions might target changing attitudes as a means to change behavior. This might be achieved using persuasive messages designed, for example, to target behavioral beliefs (e.g., Brubaker and Fowler, 1990). A further approach would be to target the anticipated emotional reactions to a behavior again using persuasive messages (e.g., Conner et al., 2011).

Another interesting approach has focused directly on the immediate determinant of behavior in several of these models, namely intentions. Where individuals do have an intention to engage in a health behavior (goal intentions), but are having trouble implementing their intention, forming a specific plan about where and when to act has been found helpful (Gollwitzer and Sheeran, 2006). For example, Conner and Higgins (2010) got adolescents who did not want to become smokers to repeatedly form plans about how to refuse the offer of a cigarette. This was shown to significantly reduce rates of smoking initiation over the next 4 years.

Over the past few years, research has also sought to develop a taxonomy of the different behavior change techniques (BCTs) that have been used in attempts to change health behavior and relate them to existing theory. For example, Abraham and Michie (2008) set out a total of 26 distinct BCTs used in relation to a range of behaviors. A good number of these relate directly to SCMs described earlier (e.g., provide information about the consequences of the behavior) although others set out alternative means to changing behavior (e.g., motivational interviewing).

Conclusions

Health behaviors have important consequences for both the quality and length of life partly via influencing disease outcomes. Nevertheless, there is still considerable variation in who performs these behaviors. SCMs provide one approach to understanding the variation in who performs health behaviors. These models are also useful because they suggest ways to change health behaviors in order to improve health. Research on BCTs has provided useful information on how to change health behaviors and further insights into why individuals perform these behaviors. The next big challenge for social and behavioral research on health behaviors is to demonstrate how theory-based interventions can produce effective and long-lasting behavior change that results in real benefits in terms of morbidity and mortality.

See also: Adolescent Health and Health Behaviors; Alcohol Use among Young People; Cancer Screening; Control Beliefs: Health Perspectives; Coronary Heart Disease: Psychosocial Aspects; Gender and Health Care; Gender and Physical Health; Health Behavior, Psychosocial Theories of; Health Behaviors, Assessment of; Health Psychology; Health Risk Perception; Personality and Health; Physical Activity and Health; Self-Efficacy and Health; Sexual Risk Behaviors; Smoking and Health; Socioeconomic Status and Health Care; Socioeconomic Status and Health; Weight Regulation.

Bibliography

Abraham, C., Michie, S., 2008. A taxonomy of behavior change techniques used in interventions. Health Psychology 27, 379–387.

Abraham, C., Sheeran, P., 2005. The health belief model. In: Conner, M., Norman, P. (Eds.), Predicting Health Behaviour, second ed. Open University Press, Buckingham, pp. 28–80.

Adler, N., Matthews, K., 1994. Health psychology: why do some people get sick and some stay well? Annual Review of Psychology 45, 229–259.

Ajzen, I., 1991. The theory of planned behavior. Organizational Behavior and Human Decision Processes 50, 179–211.

Allied Dunbar Fitness Survey, 1992. Health Education Authority, London.

Bandura, A., 1982. Self-efficacy mechanism in human agency. American Psychologist 37, 122–147.

Bandura, A., 1997. Self-Efficacy: The Exercise of Control. Freeman, New York.

Baum, A., Posluszny, D.M., 1999. Health psychology: mapping biobehavioral contributions to health and illness. Annual Review of Psychology 50, 137–163.

Belloc, N.B., Breslow, L., 1972. Relationship of physical health status and health practices. Preventive Medicine 9, 409–421.

Black Report, 1988. Inequalities in Health: The Black Report and the Health Divide. DHSS, London.

Blaxter, M., 1990. Health and Lifestyles. Tavistock, London.

Bogg, T., Roberts, B.W., 2004. Conscientiousness and health-related behaviors: a meta-analysis of the leading behavioral contributors to mortality. Psychological Bulletin 130, 887–919.

British Medical Journal, 1982. Alcohol Problems. BMJ Publications, London.

Brubaker, R.G., Fowler, C., 1990. Encouraging males to perform testicular self-examination: evaluation of a persuasive message based on the revised theory of action. Journal of Applied Social Psychology 17, 1411–1422.

Center for Diseases Control, 2008. Smoking-attributable mortality, years of potential life lost, and productivity losses United States, 2000–2004. Morbidity and Mortality Weekly Reports 57, 1226–1228.

Choi, K.H., Catania, J.A., 1996. Changes in multiple sexual partnerships, HIV testing and condom use among US heterosexuals 18 to 49 years of age, 1990 and 1992. American Journal of Public Health 86, 554–556.

Conner, M., Higgins, 2010. Long-term effects of implementation intentions on prevention of smoking uptake among adolescents: a cluster randomized controlled trial. Health Psychology 29, 529–538.

Conner, M., McEachan, R., Jackson, C., McMillan, B., Woolridge, M., Lawton, R., 2013. Moderating effect of socioeconomic status on the relationship between health cognitions and behaviors. Annals of Behavioral Medicine 46, 19–30.

Conner, M., Norman, P. (Eds.), 2005. Predicting Health Behaviour, second ed. Open University Press, Buckingham.

Conner, M., Rhodes, R., Morris, B., McEachan, R., Lawton, R., 2011. Changing exercise through targeting affective or cognitive attitudes. Psychology & Health 26, 133–149.

Conner, M., Sparks, P., 2005. The theory of planned behaviour and health behaviours. In: Conner, M., Norman, P. (Eds.), Predicting Health Behaviour, second ed. Open University Press, Buckingham, pp. 170–222.

Consensus Development Conference on Lowering Blood Cholesterol to Prevent Heart Disease, 1985. Lowering blood cholesterol to prevent heart disease. Journal of the Australian Medical Association 253, 2080–2086.

Doll, R., Peto, R., Wheatley, K., Gray, R., Sutherland, I., 1994. Mortality in relation to smoking: 40 years' observations on male British doctors. British Medical Journal 309, 901–911.

Fishbein, M., Triandis, H.C., Kanfer, F.H., Becker, M., Middlestadt, S.E., Eichler, A., 2001. Factors influencing behavior and behavior change. In: Baum, A., Revenson, T.A., Singer, J.E. (Eds.), Handbook of Health Psychology. Lawerence Erlbaum Associates, Mahwah, NJ, pp. 3–17.

Friedman, G.D., Dales, L.G., Ury, H.K., 1979. Mortality in middle-aged smokers and non-smokers. New England Journal of Medicine 300, 213–217.

Friedman, H.S., Tucker, J.S., Schwartz, J.E., Martin, L.R., Tomlinson-Keasay, C., Wingard, D.L., Criqui, M.H., 1995. Childhood conscientiousness and longevity: health behaviors and cause of death. Journal of Personality and Social Psychology 68, 696–703.

Friedman, H.S., Tucker, J.S., Tomlinson-Keasay, C., Schwartz, J.E., Wingard, D.L., Criqui, M.H., 1993. Does childhood personality predict longevity? Journal of Personality and Social Psychology 65, 176–185.

General Household Survey, 1989. OPCS, London.

Gochman, D.S. (Ed.), 1997. Handbook of Health Behavior Research, vols. 1–4. Plenum, New York.

Gollwitzer, P.M., Sheeran, P., 2006. Implementation intentions and goal achievement: a meta-analysis of effects and processes. Advances in Experimental Social Psychology 38, 249–268.

Guiguet, M., Lepont, F., Retel, O., Valleron, A.J., 1994. Changes in HIV behavior among French heterosexuals: patterns of sexual monogamy and condom use between 1988 and 1991. Journal of Acquired Immune Deficiency Syndromes 7, 1290–1291.

Greenwald, P., Sondik, E.J., 1986. Cancer Control Objectives for the Nation: 1985–2000. National Cancer Institute, Bethesda, MD.

Health Canada, 1998. Sexual Risk Behaviour of Canadians. Bureau of HIV/AIDS, STD and TB. Update Series, Centre for Prevention and Control. Canada, Ottawa.

Luszczynska, A., Schwarzer, R., 2005. Social cognitive theory. In: Conner, M., Norman, P. (Eds.), Predicting Health Behaviour, second ed. Open University Press, Buckingham, pp. 127–169.

McEachan, R.R.C., Conner, M., Taylor, N.J., Lawton, R.J., 2011. Prospective prediction of health-related behaviors with the theory of planned behavior: a meta-analysis. Health Psychology Review 5, 97–144.

Moore, L., Smith, C., Catford, J., 1994. Binge drinking: prevalence, patterns and policy. Health Education Research 9, 497–505.

National Audit Office, 1998. The Performance of the National Health Service Cervical Screening Programme in England. HMSO, London.

Office for National Statistics, 2006. Eating & Exercise. http://www.statistics.gov.uk/cci/nugget.asp?id=1329 (retrieved on 03.01.11.).

Rigotti, N.A., 1989. Cigarette smoking and body weight. New England Journal of Medicine 320, 931–933.

Royal College of Physicians, 1983. Health or Smoking: Follow-up Report. Pitman, London.

Royal College of Physicians, 1992. Smoking and the Young. Lavenham, Suffolk.

Ruchlin, H.S., 1997. Prevalence and correlates of breast and cervical screening among older women. Obstetrics and Gynecology 90, 16–21.

Sampos, C., Fulwood, R., Haines, C., et al., 1989. The prevalence of high blood cholesterol levels among adults in the United States. Journal of the American Medical Association 262, 45–52.

Steptoe, A., Wardle, J., 1992. Cognitive predictors of health behaviour in contrasting regions of Europe. British Journal of Clinical Psychology 31, 485–502.

U.S. Department of Health and Human Services, 1988. The Surgeon General's Report on Nutrition and Health. US Government Printing Office, Washington, DC.

U.S. Department of Health and Human Services, 1996. Physical activity and health: a report of the surgeon general. Journal of the American Medical Association 276, 552.

Waldron, I., 1988. Why do women live longer than men? Journal of Human Stress 2, 2–13.

Health Behaviors, Assessment of

Britta Renner, Verena Klusmann, and Gudrun Sproesser, University of Konstanz, Konstanz, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Health-related behaviors, such as physical activity, smoking, and healthy eating differ along numerous dimensions (e.g., frequency, volition), requiring various analysis methods. Physiological methods (e.g., salivary cotinine as a biomarker for smoking) offer high accuracy, and low demand characteristics but are comparably cost-intensive. Moreover, behavior-specific biomarkers are rare, preventing routine application. Behavioral observation methods using electronic devices (e.g., smartphones, accelerometers) offer accurate, valid, and effective possibilities for assessing everyday health behaviors. Self-reports on past behavioral frequencies are the most common and least resource demanding; however, irrelevant or biased questions seriously decrease the validity of a survey.

Introduction: Health Behaviors as a Multidimensional Construct

In the field of health psychology, health behaviors are most often defined as behaviors that people engage in to maintain or improve their current health and to avoid illness. These definitions focus on subjective health beliefs as they include any behavior performed to protect, promote, or maintain health, whether or not such behaviors are objectively effective toward that end. In contrast, most empirical public health studies and epidemiological research focus on behaviors which are defined by experts as having a substantial relationship to health. Therefore, the term health behavior has two different connotations. First, a subjective one which focuses on which behaviors laypeople believe might affect their health, and second, an objective one which focuses on what experts evaluate and define as health-affecting behavior. Importantly, lay and expert views of which behaviors are related to health may differ systematically. In addition, studies have typically found only moderate-to-low interrelations between health behaviors (e.g., Cohen-Mansfield and Kivity, 2011; Héroux et al., 2012). For example, drinking and physical activity are often not associated while the former is often related to smoking. One reason why health behaviors and habits are not consistently interrelated is the fact that they differ on a number of dimensions (see **Table 1**) and may be influenced by different factors. Considering the modest link between different health behaviors and the high variability in individual health behavior profiles, health behaviors should not be conceptualized as unidimensional but rather as multidimensional in character (Stefansdottir and Vihjalmsson, 2007).

Furthermore, health behaviors could be examined as antecedents or causes of health status. Such an approach is taken by social epidemiologists. In contrast, health educators are interested in health behaviors as objective dimensions for systematic interventions directed at producing behavioral changes. Finally, as subjects of scientific study, health behaviors are analyzed as consequences or outcomes, as well as correlates of individual factors. For an accurate and reliable measurement of health behaviors it is, therefore, essential to distinguish between their different dimensions and clearly define the topic to be investigated. There are various methods for assessing health behaviors, ranging from self-report to biological indicators.

What People Say: Self-Report Measures

Questionnaires

Questionnaires which assess past behavior frequency for a specific behavior such as food consumption or general behavior patterns such as eating styles are the most commonly used method for measuring health behaviors. There are numerous questionnaires which ask for the average or typical frequency and quantity of alcohol consumption (see Sobell and Sobell, 1995), food intake (see Kristal et al., 1992), or physical activity (see Pereira et al., 1997). The analytic strategy is the same across all *quantity and frequency measures* (QF estimates). The frequency of the reported behavior is multiplied by the quantity (or intensity) of the behavior, and these are summed up to obtain behavior totals. For example, the International Physical Activity Questionnaires (IPAQ), developed under the auspices of the World Health Organization, demonstrate a good reliability among 18- to 65-year-old adults in diverse settings and countries (Craig et al., 2003) and can be administered by telephone interview or self-administration. The long versions of the IPAQ assess sedentary behavior (sitting) as well as physical activity in three different domains (work, travel/transportation, home/garden recreation) for three different intensity levels (vigorous,

Table 1 Dimensions of health behaviors

Avoidance of negative outcomes (e.g., eating a healthy diet for preventing obesity; flu vaccination)	Approach of positive outcomes (e.g., increased well-being through physical activity)
Voluntary; conscious	Nonvoluntarily; unconscious
Performed without medical assistance	Requires professional medical assistance
Vital (e.g., eating)	Nonvital (e.g., smoking)
Delayed consequences (pay-offs)	Immediate consequences (pay-offs)
Infrequent	Frequent
Occasional	Habitual
Simple; low skill or resource demand	Complex; high skill or resource demand
Monofaceted (e.g., smoking)	Multifaceted (e.g., healthy eating behavior)

 http://dx.doi.org/10.1016/B978-0-08-097086-8.14155-8

moderate, and walking) for the last 7 days or a typical week. The total weekly physical activity can be estimated by weighting the reported minutes per week within each activity category using a Metabolic Equivalent of Task (MET) energy expenditure estimate. In terms of criterion validity, the self-reported IPAQ data show a fair-to-moderate agreement with accelerometers (pooled $\rho = 0.33$ for the long versions and pooled $\rho = 0.30$ for the brief versions, Craig et al., 2003). Likewise, food frequency questionnaires (FFQ) such as the European Prospective Investigation into Cancer and Nutrition (EPIC) questionnaires comprise up to 148 food items eaten during the past year (frequency and portion size) and allow calculations for nutrient intake via computerized software programs that multiply the reported frequency of each food by the amount of nutrient in a serving of that food (Nagel et al., 2007; Bohlscheid-Thomas et al., 1997).

Questionnaires such as quantity and frequency measures (QF estimates) assume that respondents are able and willing to provide accurate information. However, asking questions such as "How many eggs have you eaten during the last year?," or "How many times have you had beer in the past month?" involves a complex answering process (see Schwarz and Oyserman, 2001) requiring that respondents (1) understand the question, (2) identify the relevant behavior episode, (3) retrieve all relevant behavior instances, (4) correctly add up the behavior instances, and (5) map the behavior instances onto the response format.

From this perspective, asking clear questions about meaningful behaviors that allow candid answers is key to obtaining reliable and valid behavioral data (Schwarz and Oyserman, 2001). Thus, asking leading questions, which suggest an expected answer, or loaded questions, which have emotionally laden connotations, can quickly invalidate a survey. In particular, sensitive questions about topics such as illicit drug use or sexual behavior commonly elicit misreporting that it is largely situational. The extent of misreporting depends on whether the respondent has anything embarrassing to report and the design of the questionnaire. Cumulating evidence also indicates that misreporting sensitive topics is a motivated process in which respondents edit the information they report to avoid repercussions from third parties or to avoid embarrassing themselves in the presence of an interviewer (see Tourangeau and Yan, 2007).

In addition to social desirability concerns, lack in shared meaning can seriously invalidate self-reports. The term 'sexual partner,' for instance, may be interpreted in many different ways by the respondents, ranging from describing someone with whom one has experienced orgasm to someone with whom one has shared a bed while naked. Therefore, the target population should be carefully selected because health behaviors have different meanings in different cultures or age groups and the use of questions, which have been designed and validated for example in younger populations, might be inappropriate for older people.

Furthermore, by their very nature, questions about past behaviors assume an accurate memory of events as well as the willingness to report these to a researcher. However, respondents may not recall the actual events, employing instead various cognitive heuristics (rules of thumb) to estimate frequencies. This could result in certain biases. People use different strategies to answer frequency questions over different time spans. Episodic enumeration (recalling and counting individual incidents) is more likely to be used when shorter or more recent time spans are involved in frequency reports, whereas rate-based estimation (projecting the typical rate over the length of the recall period) is more likely to be used when longer time or more distant spans are involved. Reported behavioral frequencies for a year are generally less than 12 times the equivalent frequencies for 1 month. This phenomenon has been coined the time span effect (Hoorens and Harris, 1998). Use of different time spans over or within studies may lead to inconsistent or even misleading results.

In order to support respondents in their recall of relevant behavioral episodes over a designated time period, Sobell and Sobell (1995) developed the *Timeline Followback Method* (TLFB). Respondents are required to provide retrospective estimates of their daily behavior by using a calendar over a certain time period, which can range up to 12 months prior to the interview. In particular, starting with questions about the previous day and then moving further back into the past facilitates more accurate responses (Heesch et al., 2010). With this method, the pattern, variability, and level of physical activity, drinking, or smoking can be profiled, which is especially useful when relatively precise estimates are required or when researchers wish to evaluate specific changes in health behaviors before, during, and after interventions. Alternatively, 24 h recalls (24-HRs) can be used to decrease the recall burden for respondents. For example, the Day Reconstruction Method comprises the systematic reconstruction of activities and experiences of the preceding day with procedures designed to reduce recall biases (Kahneman et al., 2004). Similarly, the Automated Multiple-Pass Method (AMPM) from the US Department of Agriculture is a computer-based dietary intake assessment which uses multiple memory cues with standardized wording to help the respondent remember all of the food consumed in the last 24 hours (Moshfegh et al., 2008). Cumulated evidence shows that 24-HRs outperform FFQs in accurately assessing individual food intake (Illner et al., 2012).

In addition to memory-based information, respondents may use context-based information to make frequency judgments. One context-based heuristic involves the use of information provided by response alternatives in questionnaires. For example, respondents may use the range of the response alternatives provided to them as a frame of reference for estimating their own behavioral frequency (see Schwarz and Oyserman, 2001 for an overview).

Taken together, questionnaires are a very efficient and cost-effective way to collect data on health behavior. Although self-reports are often criticized for evoking the overreporting of health behavior and the underreporting of risk behavior due to social desirability concerns, a respective design of the survey and anonymity can substantially reduce systematic biases (Tourangeau and Yan, 2007). Moreover, questionnaires are suitable for most populations while placing a comparably low burden on respondents and allowing for comparisons between and within populations over time; thus, questionnaires are inherently suited to surveillance systems and large-sample epidemiological surveys. Given that questions are clear and straightforward, answers have been shown to be valid and reliable. They can also be summarized and transformed into

global indices such as the standardized Metabolic Equivalent of Task (MET). Computerized and internet-based questionnaires offer new possibilities for large-scale studies as they are flexible in terms of time and place of administration and simultaneous participation. Moreover, they offer comprehensive possibilities for tailoring questionnaires to subsamples through additional prompts, explanations, and branches with filter questions. They also save time through immediate data entry, reduce coding errors, and offer possibilities for scoring, reporting, and interpreting results. Furthermore, respondents feel more anonymous as no written record exists, and reduced social desirability concerns may allow for increased reporting of sensitive information (see Warren et al., 2010).

Interviews

Unstructured or semistructured interviews are useful qualitative techniques for researchers interested in understanding individuals' cognitive and conceptual models of health behavior and the frames of reference that individuals use to organize health behaviors. Therefore, qualitative methods are mainly concerned with the exploration and analysis of health behavior because they allow the interviewee room to address the issues which he or she feels relevant to the topics being raised by the investigator. For instance, detailed interviews and ethnography could provide insight into subjective and experiential dimensions of those behaviors related to substance misuse. One major disadvantage of qualitative methods is the fact that generality is, by definition, not quantifiable. Furthermore, since anonymity is not possible, the self-reports elicited by interviews may be affected by social desirability biases, resulting in overreporting of socially desirable behavior and underreporting of socially undesirable behavior. The results from the National Household Survey of Drug Abuse (NHSDA) showed that the percentage of respondents reporting marijuana use was higher when questions were self-administered than when they were administered by an interviewer (see Schaeffer, 2000).

Alternatively, one could introduce a 'reality check' to the respondent's account by using the *method of focus groups* (Morgan, 1997). A focus group is a collective, discussion-based interview orchestrated by a moderator who uses the small group setting to achieve a group dynamic to construct a supportive social context that allows respondents to reveal thoughts and behaviors they would normally be reluctant to disclose. While focus groups are mostly conducted in face-to-face settings, computer-based chat programs offer an alternative platform with greater anonymity and where less socially desirable behavior bias affects discussions. Overall, focus groups could be useful for generating information about behaviors that are rarely talked about (e.g., sexual behaviors and drug use).

Behavior Record Methods

Diary log techniques are data collection strategies which gather information in relation to the passage of time. As such, the yielded information is temporally ordered, providing a sequence of events and temporal profile of actions. Diary techniques can be particularly useful when data for one individual are required over a considerable period of time and/or very frequently, such as providing a general estimate of an individual's consumption for assessing smoking behavior, alcohol consumption, or dietary habits. For example, alcohol consumption diaries often include questions about the frequency of consumption, the type of drink taken, and the quantity consumed on each occasion. In comparison to questionnaires, the diary log format minimizes recall biases associated with retrospective reporting, but daily reporting may be more reactive. In addition, diaries could be valuable for providing access to so-called 'intimate' information (e.g., sexual behavior).

Stone and Shiffman (1994) have labeled strategies for collecting respondents' self-reports for their momentary or current state as *Ecological Momentary Assessment* (EMA). EMA studies usually consist of the repeated assessment of participants' momentary state as they go about the tasks of daily life in their natural environment. There are several ways to approach this sampling task. *Interval-contingent* assessments require assessment at regular intervals. One example is the method of *interactive voice response* (IVR), where people with alcohol use disorder are asked to call in on a regular basis to report their drinking to interviewers. Another way is asking individuals to record every episode of smoking, eating, or some other behavior of interest. This *event-contingent* approach, however, may not lead to a representative sample of the participant's general state and requires a clear definition of the triggering event. In contrast, *signal-contingent* sampling supplies participants with an external signal cue, which is usually timed to be emitted at random, to prompt subjects to complete a written assessment or an electronic diary. Beepers, electronic watches, and palmtop computers have been used as signal devices. Smartphones and tablet PCs offer the possibility of texting participants a personalized online questionnaire link (see, for example, www.surveysignal.com) at random or fixed time intervals.

Comparing global retrospective behavior estimates (QF estimates) with the Time-Line Follow-Back (TLFB) and EMA-based behavioral reports shows only a small difference between global and TLFB estimates and favors the validity of EMA data. Shiffman (2009) for example showed that EMA reports averaged about 2.5 cigarettes less than TLFB for the same period while demonstrating a higher interrelation with the biomarker salivary cotinine. Thus, EMA is a method which assesses recent health behaviors more precisely because it minimizes biases of recall. Instead it relies on respondents' experience reports at the moment of inquiry. However, the burden of frequent reporting may limit the application of EMA for certain behaviors or populations.

Image-based assessment methods (Kerr et al., 2012) or barcode scanning (Boissy et al., 2006) using smartphones or Personal Digital Assistants (PDAs) offer less strenuous opportunities to record health behaviors. Boushey and colleagues developed a dietary assessment application on a mobile telephone where participants take images of their foods. These images are sent to a server where food items are identified and the volume of each food is estimated automatically. The data are then indexed with a nutrient database to compute the nutrient content of the foods consumed (see Six et al., 2010; www.tadaproject.org). The analysis of naturalistic images or

scanning of barcodes provides more 'real-world' information on participants' consumption patterns. However, this information is still provided by participants and shares some critical factors with other self-report methods.

Nonreactive (Objective) Measures

Behavior Observation Techniques

Behavioral observation is a widely used method of behavioral assessment. Unlike self-reports or behavior records, which rely on respondents' perceptions of behavior, behavioral observation involves watching and recording the behavior of a person in laboratory or natural environments (e.g., workplace). Therefore, it is commonly assumed that observational data are more objective than are self-reports.

Most methods of behavioral observation provide a wealth of quantitative data, which capture both the current behavioral status and the dynamics of behavior across time and situation. Depending on the nature of the behavior, observers may be interested in any one or combination of several related characteristics. The most common characteristic observed is behavior frequency. Other characteristics include duration and intensity of the behavioral episode. A behavior change agent might be interested in reducing the frequency, intensity, or duration of a problem behavior. Regardless of which characteristic is observed, it is important to measure that characteristic consistently throughout the observation process.

Behavior observation techniques can be applied in the laboratory as well as in the field. In the field, a wealth of indicators can be used to track the target behavior. Observers can simultaneously assess the time people spent eating in restaurants (Rozin et al., 2003) or technological aids like smart cards can be used to record data on eating behavior in cafeterias. Moreover, individual transactions can be recorded by an electronic audit (Lambert et al., 2005) and participants' food intake can automatically be recorded via a wearable miniature camera (Sun et al., 2010). Behavior signatures can also be used to deduce people's behavior by examining their rubbish (e.g., Rathje and Murphy, 2001). Particular to the domain of physical activity, well-established technical tools such as pedometers and accelerometers are available for observation in natural environments. Pedometers are only suitable to measure steps taken while walking, whereas accelerometers provide an objective measure of time spent in different activities of varying intensities and indicate frequency and duration of activities. These tools are easy to wear and allow unobstructed movements while enabling reliable and comprehensive data collection (see Warren et al., 2010). However, standardized measurement protocols (e.g., number of measurement days for determining habitual activity or cut-off points for classifying activity as sedentary, moderate, or vigorous) still need to be developed. Moreover, wearing an accelerometer appears to increase physical activity (Ainsworth, 2009), thus, putting its use as a diagnostic tool into question.

In the laboratory, behavior might be observed as a response to seminaturalistic settings. For example, eating behavior can be triggered through a bogus taste test and the amount of consumed food can be recorded (e.g., Sproesser et al., 2014). In addition, the microstructure of eating, which is nonassessable through self-report, can be assessed by recording the cumulative food intake with an eating monitor (e.g., Kissileff and Guss, 2001) or chewing, swallowing, and arm movements using on-body sensors (see Illner et al., 2012). Likewise, healthy and unhealthy foods might be offered in choice paradigms. Using authentic food replicas ensures a standardization of presented foods while food preparation costs are greatly reduced (Bucher et al., 2011). For certain behaviors, other seminaturalistic measures such as video-based simulation, which involves participants viewing a video film of various situations relating to the task at hand and measuring responses to that task (e.g., a video-simulation measure of drivers' speed choice), might be more feasible. Another possibility for the seminaturalistic measuring of health behaviors has been proposed by Rothman et al. (1993), who offered participants the opportunity to order a free sample of sunscreen and to indicate their preferred protection factor rather than asking for behavioral intention. However, these techniques can only be approximations of 'reality' because they can be confounded with other variables. In the study of Rothman et al. (1993), one could argue that the participants only ordered the sunscreen because it was free.

While behavioral observation techniques could provide ecologically valid measures, which simultaneously satisfy the demands of internal and external validity, they are often impractical for broader samples or complex and habitual health behaviors. In addition, behavioral observation is not always superior to self-report since, for physical activity for example, objective measures alone do not allow for assessment using different qualitative domains. Therefore, for surveys and large-sample epidemiological studies, self-reports are a crucial – and often the only – source of information about health behaviors.

Physiological Measures

Physiological methods can be used on rare occasions. This is the most accurate way to measure alcohol consumption (via blood or urine sampling), drug consumption (via immunoassay, hair or sweat bioassay procedures), habitual dietary intakes (via biochemical markers), or physical activity (via doubly labeled water, through indirect or direct calorimetry or heart rate monitoring).

Bioassay methods are only required when a high level of accuracy about recent health behavior is needed (e.g., for workplace drug testing) or used in addition to self-reported data to provide a means of validating self-reported information (e.g., about recent drug use). However, in some circumstances respondents' belief that their behaviors can be measured physiologically may be sufficient to reduce misreporting.

Not only do physiological methods often place a moderate-to-high burden on respondents but they are comparably expensive and personally intense and therefore impractical in large-sample studies. Despite providing precise measures (e.g., doubly labeled water for total energy expenditure calculated via CO_2 production) they do not provide detailed information on health behavior patterns (e.g., the intensity, frequency and duration of physical activity) or settings (e.g., frequency during leisure time or work hours). Moreover, biomarkers and physiological measures often are not behavior-specific. Doubly

labeled water, for example, does capture all forms of energy expenditure, including physical activity and the thermogenic effects of food. Doubly labeled water therefore overestimates physical activity unless corrections are made (Prince et al., 2008). In a similar vein, heart rate monitors are efficient tools for assessing total energy expenditure, but heart rate is influenced by various factors such as age, body size, mode of exercise, and physical fitness. Also, heart rate lags behind changes in movement and thus intermittent activity patterns might be masked (e.g., in children). As with accelerometry, there is no standard for the heart flex point that allows discrimination between active and inactive phases, which is why heart rate monitoring should only be used to assess time spent in moderate and vigorous activity (see Warren et al., 2010). Due to their complementary qualities, accelerometers and heart rate monitoring are often combined: At lower levels of activity, heart rate monitoring is less accurate at estimating energy expenditure and accelerometers are at their most accurate. Nonwearing times in accelerometers are readily apparent in measured heart rate (see Warren et al., 2010).

Conclusion

If high accuracy is required, health behavior should be assessed by physiological measures. If these are not available, behavioral observation, EMA, or diary techniques are useful. However, they are resource-consuming and often not applicable to field research on large samples or complex, habitual health behaviors. Although memory or motivational biases lead to decreased validity and reliability in self-reports, these can provide very useful information for scientific studies on the consequences of health behaviors. To ascertain the accuracy of self-reports, researchers should consider the different connotations of the questions between different samples and construct the questions carefully in order to reduce autobiographical memory biases and social desirability influences. The selection of the most appropriate measurement tools must be based on well-thought research questions and an understanding of the dimensions of health behavior that are of interest. For intervention studies the chosen method must be sensitive to potential treatment effects. All of the methods presented in this article are useful tools for repeated assessments seeking to uncover changes in health behaviors, e.g., due to intervention. However, short-term fluctuations, for example in substance use, due to environmental (e.g., social settings) and psychological (e.g., mood or stress) influences may affect the psychometric properties of the measures used. Hence, seasonal effects and short-term fluctuations may lead to superficial behavioral changes, which could be misinterpreted by researchers as genuine changes.

As this article demonstrates, the perfect assessment method does not exist and the solution often lies in an adapted combination of different methods.

See also: Adolescent Health and Health Behaviors; Alcohol Use among Young People; Health Behavior, Psychosocial Theories of; Health Behaviors; Health Outcomes, Assessment of; Health Research, Qualitative; Health Risk Appraisal and Optimistic Bias; Health Surveys; Illness Behavior and Care-Seeking.

Bibliography

Ainsworth, B.E., 2009. How do I measure physical activity in my patients? Questionnaires and objective methods. British Journal of Sports Medicine 43, 6–9.

Bohlscheid-Thomas, S., Hoting, I., Boeing, H., Wahrendorf, J., 1997. Reproducibility and relative validity of food group intake in a food frequency questionnaire developed for the German part of the EPIC project. International Journal of Epidemiology 26, 59–70.

Boissy, P., Jacobs, K., Roy, S.H., 2006. Usability of a barcode scanning system as a means of data entry on a PDA for self-report health outcome questionnaires: a pilot study in individuals over 60 years of age. BMC Medical Informatics and Decision Making 6, 42.

Bucher, T., van der Horst, K., Siegrist, M., 2011. Improvement of meal composition by vegetable variety. Public Health Nutrition 14, 1357–1363.

Cohen-Mansfield, J., Kivity, Y., 2011. The relationships among health behaviors in older persons. Journal of Aging and Health 23, 822–842.

Craig, C.L., Marshall, A.L., Sjostrom, M., Bauman, A.E., Booth, M.L., Ainsworth, B.E., Pratt, M., Ekelund, U., Yngve, A., Sallis, J.F., Oja, P., 2003. International physical activity questionnaire: 12-country reliability and validity. Medicine and Science in Sports and Exercise 35, 1381–1395.

Heesch, K.C., van Uffelen, J.G.Z., Hill, R.L., Brown, W.J., 2010. What do IPAQ questions mean to older adults? Lessons from cognitive interviews. International Journal of Behavioral Nutrition and Physical Activity 7, 35.

Héroux, M., Janssen, I., Lee, D.-C., Sui, X., Hebert, J.R., Blair, S.N., 2012. Clustering of unhealthy behaviors in the Aerobics Center Longitudinal Study. Prevention Science 13, 183–195.

Hoorens, V., Harris, P., 1998. Distortions in reports of health behaviors: the time span effect and illusory superiority. Psychology & Health 13, 451–466.

Illner, A.K., Freisling, H., Boeing, H., Huybrechts, I., Crispim, S.P., Slimani, N., 2012. Review and evaluation of innovative technologies for measuring diet in nutritional epidemiology. International Journal of Epidemiology 41, 1187–1203.

Kahneman, D., Krueger, A.B., Schkade, D.A., Schwarz, N., Stone, A.A., 2004. A survey method for characterizing daily life experience: the day reconstruction method. Science 306, 1776–1780.

Kerr, D.A., Pollard, C.M., Howat, P., Delp, E.J., Pickering, M., Kerr, K.R., Dhaliwal, S.S., Pratt, I.S., Wright, J., Boushey, C.J., 2012. Connecting Health and Technology (CHAT): protocol of a randomized controlled trial to improve nutrition behaviours using mobile devices and tailored text messaging in young adults. BMC Public Health 12, 477.

Kissileff, H.R., Guss, J.L., 2001. Microstructure of eating behavior in humans. Appetite 36, 70–78.

Kristal, A.R., Shattuck, A.L., Williams, A.E., 1992. Food frequency questionnaires for diet intervention research. In: Proceedings of the 17th National Nutrient Databank Conference, June 7–10, 1992, Baltimore, Maryland. International Life Sciences Institute Press, 1994, Washington, DC, pp. 110–125.

Lambert, N., Plumb, J., Looise, B., Johnson, I.T., Harvey, I., Wheeler, C., Robinson, M., Rolfe, P., 2005. Using smart card technology to monitor the eating habits of children in a school cafeteria: 1. Developing and validating the methodology. Journal of Human Nutrition and Dietetics 18, 243–254.

Morgan, D.L., 1997. Focus Groups as Qualitative Research, second ed. Sage, Thousand Oaks, CA.

Moshfegh, A.J., Rhodes, D.G., Baer, D.J., Murayi, T., Clemens, J.C., Rumpler, W.V., Paul, D.R., Sebastian, R.S., Kuczynski, K.J., Ingwersen, L.A., Staples, R.C., Cleveland, L.E., 2008. The US Department of Agriculture Automated Multiple-Pass Method reduces bias in the collection of energy intakes. American Journal of Clinical Nutrition 88, 324–332.

Nagel, G., Zoller, D., Ruf, T., Rohrmann, S., Linseisen, J., 2007. Long-term reproducibility of a food-frequency questionnaire and dietary changes in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Heidelberg cohort. British Journal of Nutrition 98, 194–200.

Pereira, M.A., FitzerGerald, S.J., Gregg, E.W., Joswiak, M.L., Ryan, W.J., Suminsk, R.R., Utter, A.C., Zmuda, J.M., 1997. A collection of physical activity questionnaires for health-related research. Medicine and Science in Sports and Exercise 29, 1–205.

Prince, S.A., Adamo, K.B., Hamel, M.E., Hardt, J., Gorber, S.C., Tremblay, M., 2008. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. International Journal of Behavioral Nutrition and Physical Activity 5, 56.

Rathje, W.L., Murphy, C., 2001. Rubbish! The Archaeology of Garbage. University of Arizona Press, Tucson, Arizona.
Rothman, A.J., Salovey, P., Antone, C., Keough, K., Drake, M.C., 1993. The influence of message framing on intentions to perform health behaviors. Journal of Experimental and Social Psychology 29, 408–433.
Rozin, P., Kabnick, K., Pete, E., Fischler, C., Shields, C., 2003. The ecology of eating: smaller portion sizes in France than in the United States help explain the French paradox. Psychological Science 14, 450–454.
Schaeffer, N.C., 2000. Asking questions about threatening topics: a selective overview. In: Stone, A.A., Turkkan, J.S., Bachrach, C.A., Jobe, J.B., Kurtzman, H.S., Cain, V.S. (Eds.), The Science of Self-report: Implications for Research and Practice. Lawrence Erlbaum Associates, Inc., Mahwah, pp. 105–121.
Schwarz, N., Oyserman, D., 2001. Asking questions about behavior: cognition, communication, and questionnaire construction. American Journal of Evaluation 22, 127–160.
Shiffman, S., 2009. How many cigarettes did you smoke? Assessing cigarette consumption by global report, time-line follow-back, and ecological momentary assessment. Health Psychology 28, 519–526.
Six, B.L., Schap, T.E., Zhu, F.M., Mariappan, A., Bosch, M., Delp, E.J., Ebert, D.S., Kerr, D.A., Boushey, C.J., 2010. Evidence-based development of a mobile telephone food record. Journal of the American Dietetic Association 110, 74–79.
Sobell, L.C., Sobell, M.B., 1995. Alcohol consumption measures. In: Allen, J.P., Columbus, M. (Eds.), Assessing Alcohol Problems: A Guide for Clinicians and Researchers. NIAAA Treatment Handbook Series 4. National Institutes of Health, Bethesda, MD, pp. 55–73.
Sproesser, G., Schupp, H., Renner, B., 2014. The bright side of stress-induced eating: eating more, when stressed but less when pleased. Psychological Science 25, 58–65.
Stefansdottir, I.K., Vihjalmsson, R., 2007. Dimensions of health-related lifestyle in young adulthood: results from a national population survey. Scandinavian Journal of Caring Sciences 21, 321–328.
Stone, A.A., Shiffman, S., 1994. Ecological momentary assessment (EMA) in behavioral medicine. Annals of Behavioral Medicine 16, 199–202.
Sun, M., Fernstrom, J.D., Jia, W., Hackworth, S.A., Yao, N., Li, Y., Li, C., Fernstrom, M.H., Sclabassi, R.J., 2010. A wearable electronic system for objective dietary assessment. Journal of the American Dietetic Association 110, 45–47.
Tourangeau, R., Yan, T., 2007. Sensitive questions in surveys. Psychological Bulletin 133, 859–883.
Warren, J.M., Ekelund, U., Besson, H., Mezzani, A., Geladas, N., Vanhees, L., Experts, P., 2010. Assessment of physical activity – a review of methodologies with reference to epidemiological research: a report of the exercise physiology section of the European Association of Cardiovascular Prevention and Rehabilitation. European Journal of Cardiovascular Prevention & Rehabilitation 17, 127–139.

Relevant Websites

https://sites.google.com/site/compendiumofphysicalactivities/ – Compendium of Physical Activities.
www.surveysignal.com – Smartphone Survey Management.
www.tadaproject.org – Technology Assisted Dietary Assessment.

Health Care: Legal Issues (United States)

Nicole Huberfeld, University of Kentucky, Lexington, KY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P.D. Jacobson, volume 10, pp. 6520–6523, © 2001, Elsevier Ltd.

Abstract

Medicine and the law are inextricably intertwined. This article summarizes key issues in the realm of health care litigation, ranging from liability of insurers and health care providers to the role of courts in national health care policy-making.

In the United States, health care is primarily a private enterprise, run by nongovernmental organizations. The major actors in the system are private health insurers, private physicians, private medical institutions, and private management companies, none of which historically have had cause to align their interests, particularly in the context of legal liability. Even so, the role of governmental actors in US health care has been growing over the past several decades. The governmental actors involved in US health care are predominantly public insurers, most prominently the programs called Medicare (social health insurance for the aged and disabled) and Medicaid (welfare-based medical insurance for the poor). Medicare has been a stable source of public health insurance for people over the age of 65 years since its inception in 1965. Medicaid, on the other hand, is a much more dynamic program, often growing and changing to accommodate other areas of the law, such as changes in welfare rules, but always covering only the poorest citizens. The federal governmental role is poised to increase even more in coming years due to major health care reform embodied in the Patient Protection and Affordable Care Act (the Affordable Care Act (ACA)), which was signed into law in 2010. This law strives to achieve what many other industrialized nations accomplished decades ago: universal health insurance coverage.

The relationship between law and health care in the United States is multilayered and constantly in flux; health care is ensconced in layers of the law. Private common law concepts overlap with public law and constitutional law theories; health care providers are subject to both federal and state regulation of their activities, especially when public insurance is taken as payment for medical services; the federal and state actors often disagree over which governmental body is properly responsible for and thus has power over health care; each type of health care provider is separately regulated by both state and federal agencies and each is subject to differing mechanisms of liability for negligence, fraud, and the like; and costs rose, while access diminished, all of which contributed to a cacophonous health care system.

This article encapsulates the relationship between law and health care in the United States, with a focus on how courts have responded to challenges to managed care's cost containment innovations and the policy changes that litigation may be creating. One source of important litigation change in the past two decades was both a response to, and a reflection of the transition to, a managed care-influenced health care system. How the courts and legislatures responded to these changes in part determined the limitations of the ambitious cost containment initiatives represented by the heavy administrative role of managed care organizations. The move to managed care also included greater enrollment costs and aggressive strategies to exclude undesirable individuals from insurance policies, leading to high rates of uninsurance and underinsurance. Thus, after discussing legal aspects of managed care, the move toward policy-based litigation that appears to be the wave of the future for health care litigation will be considered.

Law and Health Care: Framing the Issues

The legal and medical systems are inextricably intertwined and have long intersected at many points. Since the early 1800s, medical jurisprudence, where physicians are called as expert witnesses and where the courts act to regulate the quality of medical care, has been an integral part of the growth of the health care enterprise. As Mohr noted, "… medical jurisprudence … figured prominently in early American medical history …; neither subject can be fully understood in its modern context without taking the history of legal medicine into account" (Mohr, 1993: p. 251). As this quote suggests, courts have long influenced health care delivery and payment. Even so, the relationship between law and health care has become ever more manifest and more controversial, especially with the high-profile litigation surrounding the ACA.

Historical View

Acknowledging the ebbs and flows of the relationship between law and medicine, it is arguable that the relationship was essentially stable from the mid-nineteenth century until the mid-1960s (Stryker, 1932). As recently as the late 1960s and the early 1970s, health care attorneys and health care executives had to be concerned with a relatively easy set of legal rules. Both sides could predict the nature and scope of liability litigation because the essential rules establishing the standard of care and the possible types of litigation varied little during this period. A typical court case involved one patient suing one physician, guided by liability rules that reflected judicial deference to physicians in setting the standard of care.

This relative stability was greatly disrupted by two major changes in the health care system: the advent of the managed

care organization and its growth, set off by the Health Maintenance Organization Act of 1973, and the establishment of Medicare and Medicaid in 1965. At both the state and federal levels, all three branches of the legal system – legislatures, executive regulatory agencies, and courts – have persistently been involved in setting health care policy and monitoring health care delivery since the mid-1960s, and the regulatory overlay on the prior, stable, and relatively simple mechanisms of liability has created new meanings for legal issues in health care.

Litigation Context

The framing of the legal issues in the modern managed care era reflects the dramatic changes in the health care system since the mid-1980s and the resulting policy conflicts that have emerged. For many years, courts have actively monitored quality of care through determining the standard of care for medical liability (Jacobson, 1989; Furrow, 1997) and interpreting contractual definitions of medical necessity (Eddy, 1996; Hall and Anderson, 1992). Courts have also been called on to interpret the vast array of state and federal legislation regulating health care delivery, and that set of regulations has grown both wider and deeper as the major federal programs have aged.

In responding to litigation challenging managed care cost containment practices, courts have had to confront the question of what role the judiciary should play in monitoring the health care system relative to legislatures. Courts face the tension of applying common law liability principles to new situations or deferring to elected representatives to set potentially bad or inequitable public policy. In the modern health care environment, courts are asked to distinguish the economic aspects of managed care that order relationships between health plans, physicians, and patients, from the monetary incentives that contribute to below-standard care. The former are more traditionally legislative prerogatives, while the latter are traditionally within the judicial purview. For those cases decided in the courts, judges determine whether the issues should be resolved by contract or tort (civil wrongs such as negligence) law. The difference is that recovery under contract is more difficult and damage awards are lower. What complicates matters is that in managed care, the insurance (i.e., contract) and medical care (i.e., potential negligence) are combined in one entity.

The same questions regarding the role of the judiciary, deference to legislatures, and public policy concerns arise in the context of public insurance mechanisms, but the questions become even more complex. This is because the federal government and the states fight over whose power is primary when making decisions about Medicaid, which is funded and structured by the federal government but administered and supplementally funded by the states, as well as Medicare, which accounts for a large portion of the national budget and thus draws political and economic scrutiny. This means that health care providers get caught in the fight, vying often for fair payment, and also being forced to choose a position among these governmental power struggles. In addition, any misuse of the public insurance funds leads to steep penalties at both the federal and the state levels, leading potentially to both civil and criminal liability.

Liability

Common law liability doctrine has traditionally emerged incrementally to adapt to changing circumstances, taking time for the actual contours of a doctrine of liability to be assessed. Nevertheless, certain themes have emerged in modern health care litigation that are likely to continue to evolve so long as the United States maintains its fragmented health care system.

Despite variation in states' case law, the trend has been that courts apply traditional liability doctrine to managed care entities (Furrow, 1997). For early models of Health Maintenance Organizations, the most restrictive form of managed care, the law had already developed holding them liable for torts (civil wrongs) committed by their employees. The courts have also applied indirect liability doctrines, such as vicarious liability and agency principles, to other managed care organizational forms, which exercise less intense administrative oversight, for malpractice committed by an independent physician.

In determining vicarious liability, courts look to the health insurer's control over the physician or how the entity markets its physicians. The greater the indications of control and the more the organization markets the quality of its physician panel, the greater is the likelihood that the managed care organization will be held vicariously liable. In a leading case, *Boyd v. Albert Einstein Medical Center*, 547 A.2d 1229 (Pa. Super. 1988), the court held that an independent practice association (IPA) could be liable for an individual physician's medical malpractice because the patient reasonably believed that the IPA controlled the physician. The court rejected the IPA's argument that it was not liable because the physician was an independent contractor, holding that the physician could still be an agent with respect to the patient. A key element to managed care liability is whether the managed care organization has sufficient control to override a physician's clinical decisions. This occurs through such mechanisms as utilization review, which allows the insurer to evaluate a physician's medical decisions and deny payment for such care.

Challenges to Cost Containment Practices

Health care accounts for a large percentage of the United States' gross domestic product as well as federal and state budgets, and so cost containment is often part of the national conversation. Of course, this is not so different from other countries with more centralized health care systems, but the spending is harder to control because of the significant role of private actors and their distinct interests. The conflict between population-based cost containment and access for individual health insurance subscribers has created large quantities of case law, all of it complicated by the multiplicity of actors in a given case and the evolving nature of the organizational structures in the health care system. Just as the courts had to adapt medical liability principles to the emergence of hospitals as central to health care delivery in the 1960s, the same process is now underway with regard to the ever-morphing organizational forms that form the umbrella concept of managed care. In addition, many health care providers are aligning their interests, making provider liability more complex as well. Sorting out which party is legally responsible for delayed or denied

care, and the acceptable level of care resulting from cost containment strategies, remains a challenge for both state and federal courts. The idea of rationing is still a tough one for legislatures to swallow, and they have *de facto* deferred rationing to nongovernmental organizations.

Nonetheless, courts have not impeded the implementation of cost containment initiatives exemplified by health maintenance organizations and other forms of managed care (Jacobson, 1999). Courts have upheld cost containment programs, except when such programs are seen as limiting access to life-sustaining treatment. For example, state and federal courts have generally rejected challenges to financial incentives (although some courts have allowed such challenges to proceed to trial), utilization management programs, and contractual challenges to benefit denials.

Few courts have been willing to usurp legislative choices in formulating health care policy or to obstruct the market in organizing and delivering health care services, and some courts specifically deflect to legislature's allegations that managed care incentives violate public policy. One reason for this is that courts have interpreted narrowly a federal law called the Employee Retirement Income Security Act to block state court challenges to cost containment programs that deny health benefit coverage. Another reason is that courts look to contract law to determine the extent of litigating parties' obligations and responsibilities, especially in physicians' litigation against managed care organizations. Although the trend toward contract is neither unlimited nor uniform, it represents a potentially profound shift from the early 1970s, when courts were perceived as protectors of individuals, to the modern era when courts act more as supporters of market-based changes.

Physician Autonomy

Courts have ruled that people who are dissatisfied with managed care, including physicians, should take their case to legislatures. This has represented a dramatic departure from the era when courts deferred to physician dominance. And, physicians had a reasonable expectation that courts would protect the physician–patient relationship from interference by managed care plans. On balance, the essence of the health care system lies with the physician–patient relationship, and courts consistently protected physician dominance in the past (Kapp, 1985; Hall, 1988). But, courts have not done so in the managed care era. To the extent that physicians have been able to secure protections through the courts, it has largely been through fair process requirements and retaining the legal standard of professional care.

The question is why courts are not deferring to physicians; what has changed over time to cause the courts' shift from protecting physician dominance to reinforcing managed care's dominance? One potential answer is that courts respond to changes in the underlying environment. As the nature of health care delivery has changed, it stands to reason that courts would begin to incorporate those changes. Slowly but surely, courts have internalized cost containment goals in developing legal doctrine, rather than ensuring the primacy of the physician–patient relationship.

Nevertheless, state legislatures have responded to physicians' complaints by implementing protective regulations, which are now also reflected in the federal ACA. State laws have limited some of the cost containment practices that were on the rise in the late 1980s and 1990s, leading to turnabout for managed care organizations. Even if courts were not willing to rein in the financially dominant managed care organizations, some state legislatures have been able to do so. The federal government has followed these states' lead in the ACA by ending certain practices that had led, over time, to deep inequities in access to and level of coverage by health insurance.

Future Challenges

During the next few years, the legal system will face many questions regarding health care both as implemented and in terms of pure public policy. Major policy questions, such as how to balance the health care access and coverage needs of individual patients with the cost containment goals of the health care system and how to reconcile conflicting social policy goals, seem to more often be decided through judicial processes. This section considers some current and future trends in the overlap between law and medicine.

Balancing Access with Coverage

No easy solution exists for balancing the individual patient's access to health care with conserving health insurers' assets for the entire patient population. Private health insurers in the United States often opt for protecting their economic well-being and that of their shareholders, and patients seek care regardless of its cost – creating constant tension that can lead to life or death litigation.

Despite its name, the ACA leans toward health care access rather than toward cost management. The Act focused on creating a health insurance home for every US citizen because health insurance is the gatekeeper for health care in the United States. To facilitate health insurance coverage, the Act put an end to certain practices, such as policy rescission and preexisting coverage exclusions, which disproportionately excluded the chronically ill, the poor, people of color, and women from health insurance coverage. To ensure a large risk pool to cover the expense of the additional covered lives, the Congress required that every person (with some exceptions) obtain health insurance coverage or pay a tax penalty for failure to carry insurance. In a highly publicized litigation, a private business organization and half of the states challenged the constitutionality of the 'individual mandate' as well as the expansion of the Medicaid program, which helps to provide health coverage to the poor who could not afford private insurance policies, even with federal tax subsidies. These aspects of the ACA were upheld by the US Supreme Court in *National Federation of Independent Business v. Sebelius*, 132 S. Ct. 2566 (2012), although the Medicaid expansion was rendered optional for states when it had been created as a mandatory Medicaid eligibility expansion. But, the win for federal lawmaking may be short lived.

This type of policy-challenging litigation may become the wave of the future. Already, additional challenges to the ACA's health insurance rules are percolating in lower federal courts. Once again, the proper role of the judiciary in health care policy

choices has come to the fore, but now on a much broader scale, as the judiciary has been asked to consider not the decisions of private health executives, but those of the federal legislative and executive branches, which were striving to ameliorate a nationwide, and growing, problem by enacting the ACA.

Social Policy Conflicts

A plethora of legislation has placed health care providers, institutions, and executives in the policy-making cross fire. Some regulations, such as antifraud and abuse measures, are specifically designed to reduce health care costs. Others, such as the Emergency Medical Treatment and Labor Act, are designed to improve access through a point of rescue in the emergency room, regardless of the cost. Taken alone, each individual law is defensible. Taken together, however, the social policy goals of cost containment, improving access to health care, and improving quality of care conflict across statutes and regulations. Complying with all the federal and state laws that apply to health care providers and health insurers may be impossible. And, the multiple layers of law impacting medicine reflect disjointed social policy.

The fundamental policy tensions that underlie such laws are arguably better left to the political branches of government, but courts are mired in litigation regarding health care both on a macro- and a microlevel. Courts' involvement ranges from simple medical liability for negligence cases to major questions about access to courts for Medicaid beneficiaries and the constitutionality of major health care legislation. Each of these cases, whether adjudicating interpersonal conflicts or wholesale social policy, has a ripple effect in health care.

Implications

Other countries may not learn much from the managed care model, as it has created a long stream of litigation and other strife in the United States, resulting in a significant toning down of the original (health maintenance organization) model throughout the early 2000s. On the other hand, it seems likely that most countries will face the need to impose cost constraints on their national health care systems. These choices are already evident in the British and Canadian systems, which are facing the dual pressures of governmental budget concerns and the private sector performing alongside the public system. Reconciling cost containment with institutional stakeholders' objectives, patients' interests, and limited governmental resources takes compromise, careful planning, and often contentious debate. Even then, it can be difficult to ensure that changes are implemented fairly and equitably, and lawmaking by litigation can derail even the most meticulous efforts.

See also: Comparative Health Systems; Doctor–Patient Interaction in the West: Psychosocial Aspects; Health Care Markets: Theory and Empirical Evidence; Health Care, Rationing of; Health Insurance: Economic and Risk Aspects; Litigation; Managed Care and Psychotherapy in the United States; Socioeconomic Status and Health Care.

Bibliography

Eddy, D.M., 1996. Benefit language: criteria that will improve quality while reducing costs. Journal of the American Medical Association 275, 650–657.

Elhauge, E. (Ed.), 2010. The Fragmentation of U.S. Health Care. Oxford University Press, New York.

Furrow, B.R., 1997. Managed care organizations and patient injury: rethinking liability. Georgia Law Review 31, 419–509.

Hall, M.A., 1988. Institutional control of physician behavior: legal barriers to health care cost containment. University of Pennsylvania Law Review 137, 431–536.

Hall, M.A., Anderson, G.F., 1992. Health insurers' assessment of medical necessity. University of Pennsylvania Law Review 140, 1637–1712.

Jacobson, P.D., 1989. Medical malpractice and the tort system. Journal of the American Medical Association 262, 2230–2237.

Jacobson, P.D., 1999. Legal challenges to managed care cost containment programs: an initial assessment. Health Affairs 18 (4), 69–85.

Jacobson, P.D., Pomfret, S.D., 1999. Establishing new legal doctrine: a model of judicial response to industrial change. University of Michigan Journal of Law Reform 32, 813–816.

Kapp, M.B., 1985. Medicine and law: a symbiotic relationship? American Journal of Medicine 78, 903–907.

Mohr, J.C., 1993. Doctors and the Law: Medical Jurisprudence in Nineteenth-century America. Oxford University Press, New York.

Rosenbaum, S., Frankford, D.M., Moore, B., Borzi, P., 1999. Who should determine when health care is medically necessary? New England Journal of Medicine 340, 229–332.

Starr, P., 1982. The Social Transformation of American Medicine. Basic Books, New York.

Stevens, R., Stevens, R., 1974. Welfare Medicine in America. The Free Press, New York.

Stryker, L.P., 1932. Courts and Doctors. Macmillan, New York.

Health Care Markets: Theory and Empirical Evidence

Stefan K Lhachimi, University of Bremen, Bremen, Germany; and Leibniz Institute for Preventive Research and Epidemiology, Bremen, Germany
Johannes Siegrist, University of Duesseldorf, Duesseldorf, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D.W. Light, volume 10, pp. 6523–6527, © 2001, Elsevier Ltd.

Abstract

The notion of health care markets emerged several decades ago in response to rising costs and limited resources available in health care. Although providing health care to some extent conflicts with established economic principles of functioning markets, attempts were made to strengthen competition, quality, coverage, and choice in the organization of health care. The main current approaches, their promises, and their problems are briefly described, with special emphasis on the United States, Great Britain, and the Netherlands. Robust empirical evidence is still lacking on whether competitive health care markets always result in providing better health outcomes than regulated markets. At the level of research, an econometrically driven health care research might profit from incorporating sociological insights into – and analytical concepts of – organizational functioning.

Origins

'Health care market' is a comparatively new concept of analysis. Health is not an ordinary commodity and therefore was long considered not to be suitable for classical economic analysis. In the 1970s, however, the crisis of the welfare state became apparent, and economical, structural, and social forces started to scrutinize those segments of welfare provision where rising costs were faced with stagnating revenues. The health care sector was no exception to this, in particular as this time period was labeled by some as a 'medical arms race.' Patients were fully covered by (public) insurance, which paid on a fee-for-service basis, and medical providers were motivated to supply more and increasingly costly medical services. Although the empirical evidence of a substantial causal increase of health care costs remains scarce, it goes unsaid that the overall spending for health care continued (and still continues) to rise. Moreover, it became obvious that the increase in the volume of treatment did not fully correspond with an increase in overall health, casting doubt on the efficiency of the health care system's performance. Hence, the concept of 'health care markets' was developed to apply the principles of competition to health care in order to increase efficiency by containing costs while improving quality. A major goal was to incentivize providers to define treatment in terms of overall health improvements rather than in terms of single medical services. This approach started to challenge the professional dominance of physicians who lost some control over the definition of appropriate medical treatments, and it strengthened the role of the payer – who was not identical with the patient – in the health care sector.

Problems of Market Failure

Classical economic thinking for a long time avoided applying the theory of price competition to the health care arena. The exchange between patient and doctor was judged as lacking the basic qualities of efficient markets. These qualities are, among others, a large number of market participants, free market entry and exit, marketability of all products, symmetric information, and zero search costs. In particular, the asymmetry between patients and clinicians (e.g., evaluation of necessity of treatment) on the one hand and the asymmetry between insurer and insurance seeker (e.g., higher likelihood of unhealthy individuals to buy insurance) on the other hand were identified as conditions that distort competition to a substantial extent. Moreover, the structure of the market with regard to health differs from established procedures as there are usually only a few or even only one local health care provider, and the nature of products, that is, every single medical treatment, is virtually a differentiated service that is not retradable. Given these conditions, unregulated competition may or may not lead to a socially desirable outcome. For instance, fixing prices by regulation may result in an overdifferentiation of unimportant product characteristics, as is the case if patients decide to prefer treatment in hospitals with hotellike amenities, although this poorly reflects treatment quality (Gaynor and Vogt, 2000).

Managed Competition as a Possible Solution

To overcome the market failure resulting from the incongruities between economic competition and clinical medicine, Enthoven (1988) developed the theory of managed competition. 'Managed competition' is split into (1) competition between payers for enrollees and (2) competition between hospitals for patients. The payers have now an incentive to contract selectively with hospitals to lower prices and increase quality for their enrollees. Elements of managed care of include (1) government provision of comparative information on health plan quality, (2) definition of minimum package of guaranteed treatments with a means-tested subsidy for low-income individuals, and (3) mandatory risk adjustment schemes between health plans to avoid 'cream skimming' by focusing on young/healthy individuals and avoiding high-cost patients (Cookson et al., 2013). Managed competition was received very well and it was considered by many as the only viable solution to curb health care costs without sacrificing

quality. Nevertheless, critics pointed out several flaws of this concept. Most notably, it implicitly assumes that providers (physicians) are inherently untrustworthy, whereas private payers (managers) are more trustful and, more importantly, more capable of dealing with uncertainty, information asymmetry, and other sources of market failure. Furthermore, managed competition increases marketing costs – that are now necessary to attract patients/enrollees – and it requires a new layer of administration to implement and control services and bills or (changing) clinical behavior. Similar to the concept of managed competition, the concept of 'quasimarkets' was proposed. Quasimarkets developed in health care systems where the government is the only payer and where all citizens are covered by a single (comprehensive) benefit package. Competition between hospitals is often induced by introducing the notion of 'payers' competition,' i.e., by artificially creating (geographically defined) third-party payers, e.g., local trusts, to contract selectively with hospitals. 'Patient-driven competition,' i.e., the principle of allowing patients to freely choose between different hospitals, is thought to motivate hospitals to compete for better quality. However, these two elements may interfere with each other as the payers' ability of contracting selectively is contradicted by the patient's freedom to choose any hospital (Cookson et al., 2013).

Proliferation and Variations

To illustrate, different approaches to health care market's recent developments of health care systems in three countries, the United States, the United Kingdom, and the Netherlands, with markedly distinguished traditions and backgrounds are briefly described.

In the United States, during the early 1980s, Medicare (as a government payer) introduced diagnosis-related groups (DRGs), i.e., prospective payer schemes that reimburse hospitals according to the average cost of a treatment, instead of reimbursing the actual costs of a particular patient. If actual costs are higher than the reimbursed average cost, the hospital realizes an economic loss with a patient. This offers hospitals the incentives to lower average treatment costs and, at the same time to avoid patients with high-cost treatment. Private insurers started to develop similar reimbursement rules. Since the late 1980s, private payers also started to negotiate (discounted) prices with preferred provider organizations, and they registered patients in health maintenance organizations (HMO). HMOs aim to lower prices while trying to ensure the quality of care provided. Despite this promising approach, many HMOs put restrictions on the choice of their providers, and they declined certain treatments. As a result, their public appeal declined. More recently, the Affordable Care Act ('Obamacare') in the United States has passed legislation in 2012, with most provision coming into effect in 2014. This highly debated innovation, considered by some as a groundbreaking structural reform of the world's largest health care market, consists mainly of three elements. First, insurers are not allowed to cap benefits on annual or lifetime basis, and essential benefit packages are defined by the government. Moreover, insurers are not allowed to deny coverage or to charge higher premiums to most individuals with preexisting conditions or adverse health status. Second, the reform stipulates an individual mandate where every individual has to have insurance coverage, either individually or via an employer. Public insurance exchanges are established that rate insurance plans according to the treatments covered in order to increase the quality of information available to the individual buyers. Third, Medicaid is being expanded to cover a higher proportion of individuals. Moreover, insurance contracts will be subsided depending on income.

In 1989, the United Kingdom reformed the centrally administered National Health Service, the world's largest health care organization, by introducing elements of quasimarkets. Internal markets were created, separating providers of health care from the purchasers of health care (Propper et al., 2004). Local bodies contracted an annual fixed amount of treatments from local hospitals for their patient populations – with the intention of maximizing quality for the lowest price. Most elements of the internal markets were dismantled in 1997, however, by the newly elected Labour government, except for the separation between purchases and providers (Cooper et al., 2011). Starting in 2003, a series of market-oriented reforms introduced several new elements. On the payment side, hospitals were reimbursed in a DRG-like system, setting a fixed ex ante price. On the demand side, the concept of patient choice was introduced: instead of being referred by the treating physician, patients were given the choice among at least five different hospitals. This improved patient choice has been, and continues to be, supported by extensive, publicly available comparative information on several quality measures of hospital performance, such as health outcomes or waiting times. The innovation was intended to give hospitals an incentive to lower average treatment costs and yet attract patients through good and further improved quality (Cookson et al., 2013).

In 2006, the Dutch government introduced a mandatory private insurance market. Earlier on, the majority of the population was insured by a public nonprofit system. With this reform, every individual was required to buy private insurance, and private insurers had to accept every applicant. The government subsidizes individuals with low income and organizes a risk adjustment scheme between insures. Insurers are now engaged in mutual competition, and they negotiate with (nonprofit) hospitals' prices through selective contracting (Schut and van de Ven, 2011). This system meets the main criteria proposed by the model of managed competition that Enthoven outlined in the 1970s.

Other countries too transformed or changed their health care sector by introducing elements of 'managed competition,' such as Germany, Switzerland, New Zealand, France, or Sweden.

Empirical Evidence

The following paragraphs can only provide some selective evidence of the role of competition in health care markets, specifically in the current context of increased market competition. The majority of health economic research publications originates from the United States (about one-third of the vast number of published papers (Wagstaff and Culyer, 2012)).

Substantially less research findings are available from Canada, Australia, and the Netherlands, which, nevertheless, document prominent developments in these countries. Recent comprehensive reviews of the empirical evidence include Gaynor (2006), Gaynor and Town (2011), and Propper and Leckie (2011).

Market Structure

The last two decades witnessed a significant structural shift of the US hospital markets. Market structure on the supply side matters, as a low number of competitors in a given market increases their bargaining power, which usually results in overall higher prices (Dranove, 2011). In 2006, 250 of 332 Metropolitan Statistical Areas were classified as 'Highly Concentrated' according to criteria used by the Federal Trade Commission. Mergers and acquisitions were identified as the main driving forces behind this concentration (Town et al., 2006). As for the underlying cause of this consolidation wave among hospitals, many authors point to the rise of managed care and the increase in number of HMOs, among others, and the pressure to improve bargaining power of hospitals (Chernew, 1995; Cutler et al., 2000; Dranove et al., 2002). The causal link, however, may not be as straightforward, and the observed mergers may be a reaction to the anticipated price pressure rather than an actual consequence (Town et al., 2007). Similarly, in the Netherlands the number of hospitals declined between 1997 and 2010 by almost 20%, leading to a high market concentration where half of all Dutch hospitals are now in markets that consist of three hospitals or less. The evidence in England is somewhat mixed: after undergoing a concentration in hospital markets in the late 1990s during the first wave of procompetition reforms, this trend was slightly reversed between 2003 and 2006 as a result of the second reform wave.

A second important area of health care markets beside hospitals concerns physicians and the local organization of their practice. So far, surprisingly little systematic research exists about the structure of local physician markets. Liebhaber and Grossmann (Liebhaber and Grossman, 2007) report that in the United States between 1996 and 1997 and 2004 and 2005 the proportion of solo-/two-physician practices declined from 40.7 to 32.5%. In the same period, the proportion of practicing physicians with an ownership stake declined. Hence, physicians are increasingly working jointly in practices where the number varies between 6 and 50 physicians. These practices, however, are mainly dealing with a single specialty rather than providing multispecialty services, although this latter organizational model may be better suited to support care coordination and quality improvement. In an analysis of the Californian market structure of physician organizations, i.e., the number of groups of organized physicians that contracted with health plans at the county level, it was found that 80% of the 46 counties included in the analysis exhibited a high market concentration (Schneider et al., 2008).

Effect of Competition on Quality

In the past decade, an increasing number of empirical studies investigated the effects of competition on the quality of health care. Usually two price-setting mechanisms are distinguished in these studies: regulated prices and market-determined prices.

In regulated price regimes – such as in England or in the United States for Medicare patients – payers reimburse hospitals with a fixed amount that should be set in such a way that the reimbursement rate is above the cost for an average patient. Several studies in the United States used a measure based on overall market share or on number of hospitals in a local market as a proxy for the degree of competition. Furthermore, these studies included predominantly mortality as a measure of quality of care, whether defined as cause-specific (e.g., coronary heart disease) or all-cause mortality. A majority of these studies found a positive effect of competition on quality (Kessler and Geppert, 2005; Shen, 2003; Tay, 2003), that is, lower mortality in less concentrated markets. This relationship was also observed in England for the changes initiated in 2003 (Cooper et al., 2011, Gaynor et al., 2013). However, some studies find no effect (Mukamel et al., 2001; Shen, 2003) or a reverse relationship for certain local markets (Gowrisankaran and Town, 2003).

In regimes with market-determined prices, little competition took place in the 1960s and the 1970s when consumers were fully covered by insurance and where hospitals were reimbursed for their full cost. This situation changed in the early 1990s with the implementation of policies that strengthened competition, in particular in the US health care system. Studies analyzing these changes showed variable findings: The number of studies reporting an increase in quality (usually lower mortality) when competition is stronger (Abraham et al., 2007; Howard, 2005; Rogowski et al., 2007) was matched by an almost equal number of studies that found a negative effect (Encinosa and Bernard, 2005; Mukamel et al., 2002; Propper et al., 2008).

For the Netherlands, preliminary analyses show that private insurers could curb the price growth for treatments with negotiated prices, i.e., market-determined prices, as compared to regulated prices. However, despite a range of publicly available quality measures, the effects of these developments on the quality of health care are still under study (Schut and van de Ven, 2011).

In the case of regulated prices, economic theory predicts an increase in quality in markets with more competition, as long as prices allow an overall profit. This relationship was confirmed in most studies conducted in England and the United States. For regimes with market-determined prices, it may seem surprising that some studies find an increase in quality, while others find the opposite. However, this is in line with economic theory, which maintains that quality may improve or deteriorate when prices and qualities are determined under market conditions.

Concluding Remarks

It turns out that in competitive markets quality can be higher, while costs are lower. However, this depends on a number of institutional arrangements. For instance, it is important to know who chooses the provider, how prices are set, and to what extent information on quality is available. The last point is crucial in order to overcome the information asymmetry

between provider and patients that is inherent in health care markets. So far, no universally agreed-upon set of indicators measuring overall health care quality is available, e.g., whether to base it on morbidity, mortality, quality of life, or treatment length. A further challenge concerns the definition of best ways of communicating quality information to patients, e.g., through league tables or report cards. It will be important to ensure that patients understand such information and that they incorporate it into their decision making. To rely solely on payer choice in interpreting information on quality measure may not be a good solution, given the vested interests of payers, which may bias their decision. Despite the developments described above, there is still no definite evidence available that competitive health care markets achieve the same or better cost and quality results than regulated health care markets. Against this background, it is difficult to predict potential effects of the recent American health care reform ('Obamacare'), which by its magnitude may challenge established approaches of health care research. Nevertheless, it is of interest to note that the Dutch experience, which shares several elements of with the American health care reform, was confronted with some major problems: both insurers and patients are requiring more meaningful information than currently provided in order to be able to assess the quality of providers. In addition, the amount of competition occurring between insurers might turn out to be weaker than expected, both in terms of attracting enrollees and in terms of bargaining with providers. Most importantly, the Dutch system so far did not achieve total coverage of the population, a main goal of the American health care reform. In the Netherlands, a significant number of people are still uninsured or are defaulting on their insurance premiums (Schut and van de Ven, 2011). However, the interplay between institutions and regulations in the United States is more complex than is the case in the Netherlands as Medicaid – as a single payer system – and the number of privately insured individuals are expected to expand simultaneously.

In a broader perspective, it might be a promising future research approach to health care markets to apply the perspective of economic sociology. The econometrical approaches used to analyze the working of health care markets are very limited as they rarely take into account the individual interactions in health care, such as the relative frequency, complexity, specificity, size, and risk of transactions; the relative quality of information; or the degree of social relationships developing over time. This gap of knowledge is most aptly demonstrated by almost nonexisting research on physician markets. Moreover, empirical studies often ignore the sociological dimensions of the relationships between parties in an economic transaction. These dimensions are typically described in terms of degree of autonomy vs mutual obligation, degree of proximity or distance between partners of a network or organization, size of the organizations or networks, and structure of power relationships. At the macrolevel, the underlying causes of the proliferation and variation of different concepts of health care markets across countries are not well understood. It seems that the prominence of the concept of health care market in part was due to promotion efforts developed by international consulting firms and global institutions, such as the World Bank or the International Monetary Fund. As it is now an international policy movement, many countries are motivated to initiate partial or complete reforms of their health care system, relying on elements of this approach. This will probably remain an ongoing process. The quality and cost outcomes of health markets strongly depend on the institutional structure in which they are embedded and on their delivery procedures whose specific effects are still poorly understood.

See also: Comparative Health Systems; Doctor–Patient Interaction in the West: Psychosocial Aspects; Globalization and Health; Health Care, Rationing of; Health Insurance: Economic and Risk Aspects; Health Outcomes, Assessment of.

Bibliography

Abraham, J.M., Gaynor, M., Vogt, W.B., 2007. Entry and competition in local hospital markets. Journal of Industrial Economics 55, 265–288.

Chernew, M., 1995. The impact of non-IPA HMOs on the number of hospitals and hospital capacity. Inquiry 32, 143–154.

Cookson, R., Laudicella, M., Li Donni, P., 2013. Does hospital competition harm equity? Evidence from the English National Health Service. Journal of Health Economics 32, 410–422.

Cooper, Z., Gibbons, S., Jones, S., McGuire, A., 2011. Does hospital competition save lives? Evidence from the English NHS patient choice reforms. Economic Journal 121, F228–F260.

Cutler, D.M., McClellan, M., Newhouse, J.P., 2000. How does managed care do it? RAND Journal of Economics 31, 526–548.

Dranove, D., 2011. Health care markets, regulators, and certifiers. In: Pauly, Mark V., McGuire, T.G., Pedro, P.B. (Eds.), Handbook of Health Economics. Elsevier, pp. 639–690.

Dranove, D., Simon, C.J., White, W.D., 2002. Is managed care leading to consolidation in health-care markets? Health Services Research 37, 573–594.

Encinosa, W.E., Bernard, D.M., 2005. Hospital finances and patient safety outcomes. Inquiry 42, 60–72.

Enthoven, A., 1988. Theory and Practice of Managed Competition in Health Care Finance. North-Holland, Amsterdam.

Gaynor, M., 2006. Competition and quality in health care markets. Foundations and Trends in Microeconomics 2, 441–508.

Gaynor, M., Moreno-Serra, R., Propper, C., 2013. Death by market power: reform, competition, and patient outcomes in the national health service. American Economic Journal: Economic Policy 5, 134–166.

Gaynor, M., Town, R.J., 2011. Competition in health care markets. In: Pauly, Mark V., McGuire, T.G., Pedro, P.B. (Eds.), Handbook of Health Economics. Elsevier, pp. 499–637.

Gaynor, M., Vogt, W.B., 2000. Antitrust and competition in health care markets (Chapter 27). In: Anthony, J.C., Joseph, P.N. (Eds.), Handbook of Health Economics. Elsevier, pp. 1405–1487.

Gowrisankaran, G., Town, R.J., 2003. Competition, payers, and hospital quality. Health Services Research 38, 1403–1421.

Howard, D.H., 2005. Quality and consumer choice in healthcare: evidence from kidney transplantation. Topics in Economic Analysis and Policy 5 (1), 1–20.

Kessler, D.P., Geppert, J.J., 2005. The effects of competition on variation in the quality and cost of medical care. Journal of Economics and Management Strategy 14, 575–589.

Liebhaber, A., Grossman, J.M., 2007. Physicians moving to mid-sized, single-specialty practices. Tracking Report, 1–5.

Mukamel, D.B., Zwanziger, J., Bamezai, A., 2002. Hospital competition, resource allocation and quality of care. BMC Health Services Research 2, 1–9.

Mukamel, D.B., Zwanziger, J., Tomaszewski, K.J., 2001. HMO penetration, competition, and risk-adjusted hospital mortality. Health Services Research 36, 1019–1035.

Propper, C., Burgess, S., Gossage, D., 2008. Competition and quality: evidence from the NHS internal market 1991–9. Economic Journal 118, 138–170.

Propper, C., Burgess, S., Green, K., 2004. Does competition between hospitals improve the quality of care? Hospital death rates and the NHS internal market. Journal of Public Economics 88, 1247–1272.

Propper, C., Leckie, G., 2011. Increasing competition between providers in health care markets: the economic evidence. In: Glied, S., Smith, P. (Eds.), The Oxford Handbook of Health Economics. Oxford University Press, London, pp. 164–188.

Rogowski, J., Jain, A.K., Escarce, J.J., 2007. Hospital competition, managed care, and mortality after hospitalization for medical conditions in California. Health Services Research 42, 682–705.

Schneider, J.E., Miller, T.R., Ohsfeldt, R.L., Morrisey, M.A., Zelner, B.A., Li, P., 2008. The economics of specialty hospitals. Medical Care Research and Review 65, 531–553.

Schut, F.T., van de Ven, W.P., 2011. Effects of purchaser competition in the Dutch health system: is the glass half full or half empty? Health Economics Policy and Law 6, 109–123.

Shen, Y.C., 2003. The effect of financial pressure on the quality of care in hospitals. Journal of Health Economics 22, 243–269.

Tay, A., 2003. Assessing competition in hospital care markets: the importance of accounting for quality differentiation. RAND Journal of Economics 34, 786–814.

Town, R., Wholey, D., Feldman, R., Burns, L., 2006. The Welfare Consequences of Hospital Mergers. NBER Working Paper.

Town, R.J., Wholey, D., Feldman, R., Burns, L.R., 2007. Revisiting the relationship between managed care and hospital consolidation. Health Services Research 42, 219–238.

Wagstaff, A., Culyer, A.J., 2012. Four decades of health economics through a bibliometric lens. Journal of Health Economics 31, 406–439.

Health Care, Rationing of

David Mechanic, Rutgers, The State University of New Jersey, Institute for Health, Health Care Policy and Aging Research, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Health care rationing is ubiquitous. It varies from formal approaches designed to allocate limited organs for transplantation to more common strategies and processes that constrain demand for care. Many countries seek to explicitly formulate priorities and translate these through benefit design and other explicit managerial decisions but much rationing is implicit, lacks transparency, and occurs in the routine provision of care. Given unlimited possibilities for expenditure, rationing is inevitable but remains a source of ethical and political controversy.

The term 'rationing' is used with different meanings in health care discussions (Mechanic, 1979). Most generally, it refers to processes of allocation when demand for care exceeds the available supply or the willingness of payers to increase it. In its more narrow application, it refers to how a scarce medical resource, such as hearts, kidneys, or livers, is allocated among competing recipients, only some of whom can receive it. Medical care has always been rationed in both senses, but the concept itself is associated with much affect and denials that rationing occurs.

To the extent that medical care is treated substantially as a commodity, as in the USA, then the market distributes services. Individuals with more comprehensive insurance or the ability to pay out-of-pocket command a greater quantity of and more sophisticated services. But even in market systems, some life and death services in scarce supply transcend the market and are governed by different rationing schemes. In publicly funded medical systems, where most of the population has health care entitlement, whether through national health insurance or a national health service, the availability of services depends on decisions about what services will be financed and their comprehensiveness and intensity (Klein et al., 1996). Some such systems, for example the English National Health Service (NHS), allow competing private systems of care. In contrast, Canada prohibits private provision of services covered by its health insurance program.

Rationing processes and decisions are implemented at different levels of government, by sick funds, large insurance plans, provider organizations such as hospitals, clinics, and health maintenance organizations (HMOs), and at the point of service delivery between clinician and patient. Disagreements are common on where in the hierarchy of health care organization varying types of rationing decisions should be made (macro, meso, or micro) and on appropriate processes for making such decisions (Mechanic, 1997). At each level, rationing may be more or less explicit.

Explicit versus Implicit Rationing

Explicit rationing refers to the detailed specification of how resources will be allocated by a rationing authority. Explicit rationing includes specifying eligibility rules, the specific services to be covered, the amount of coverage for recipients, intervals for receiving services, copayment requirements, the proportion of the budget to be allocated to varying services, the geographic allocation of services, the specific form and location of many services, and appliances and drugs that may or may not be allowed.

Implicit rationing, in contrast, involves setting budgetary limits but allowing the processes of allocation to be worked out at lower organizational levels (Mechanic, 1979). Implicit rationing includes prospective reimbursements or capitation payments. In each case the authority sets the financial limits but then allows the budgeted or capitated entity to establish the processes for decision-making. Total responsibility is sometimes delegated to the lower unit but more typically there are some explicit guidelines and constraints over discretion. All medical systems combine explicit and implicit processes but the balance may differ considerably from one nation to another and within varying national contexts.

Klein et al. (1996) describe a range of strategies for implicitly limiting resources allocated to individuals. These include denial as beneficiaries or for the services claimed; selection of beneficiaries or services on the basis of need or likely success, or demographic characteristics such as age; deflection of care to other sectors or providers; deterrence through bureaucratic barriers such as waiting lists and other delays; diluting services by reducing intensity or restricting specialty care; and premature closure or early discharge.

Concerns About Rationing

Despite general acceptance of the inevitability of rationing, major issues of controversy include who should determine rationing strategies, the openness of decision processes, and whether decisions should be made by managers or by clinicians (New, 1997). Common criticisms are that the implicit rationing processes used are not transparent, often favor the privileged, and result in unequal care for persons of lesser status, education, and sophistication, and with less skill in managing bureaucratic systems (Churchill, 1987). Critics also argue that implicit rationing gives doctors free reign to act on their prejudices. The lack of openness makes it difficult to assess the fairness of decision-making or to allow aggrieved

individuals to successfully appeal exclusions or dilution of care. Implicit rationing is also seen as undermining democratic processes through which communities can bring their preferences to bear in the allocation of important public resources.

A significant difficulty with explicit rationing, in contrast, is that it mobilizes interest groups who seek to affect decisions about coverage. Bringing medical decisions into the public arena energizes a political process and instability in the medical care system (Mechanic, 1995). Using the political process, individuals and/or groups may insist on services that are known to be ineffective or have little value. In contrast, while a particular treatment may be of little value in general, there may be instances, difficult to specify beforehand, where it is appropriate. Explicit rationing cannot, in most instances, accommodate the complex contingencies associated with the use of a particular treatment for an individual patient.

Questions about transparency apply at all levels of decision-making. The criteria used when a national government divides a health budget among regions can have significant consequences. Funds can be allocated politically, on the basis of historical funding, and on the basis of algorithms that take account of population characteristics and dispersion, and varying concepts of health status and need. The availability of facilities and specialists differs widely among population areas and accounts for large variations in expenditures among varying population groups. Nations can seek to bring greater equity among regions, types of services, and even individuals by how they decide to distribute resources.

Within organizations, decisions are made to allocate funding among primary and specialty services, among specialty units, and between medical and related health services. These decisions can have large consequences for the likelihood that a person with a particular need can receive responsive care. At the clinical level, decisions about waiting time, access to specialized services, availability of specialized treatments, or home or hospital treatment can profoundly affect patterns of care. Debate persists as to the extent such decisions should be governed by public discussion and decision-making or left to the discretion of clinicians based on their own professional judgment and experience (Symposium on the Law and Policy of Health Care Rationing, 1992).

Concerns about fairness and facilitating public discussion of priorities would support making practices explicit but there are significant competing concerns. Despite increasing interest in knowledge-based medical practice, much remains uncertain. There is no consensus, in most areas, on standards that apply to a diverse patient population with varying needs and inclinations and unique medical histories. Medical knowledge and practices are rapidly changing and it is uncertain whether explicit standards, once established, can be modified to keep pace with new knowledge. Moreover, medical care is an iterative process with assessment and disease management developing over time. As a complex process of discovery and negotiation, clinical practice is not easily subjected to definitive technical rules.

The notion of explicit decision-making at the level of direct patient care implies that individuals with comparable conditions have similar needs and preferences. But treatment must be flexible and take account of varying social contexts, cultural inclinations, and personal preferences. Treatments of comparable medical effectiveness may have vastly different implications for work and family life, or for comfort under varying circumstances. To treat medical care simply as a technical process, devoid of meanings, preferences, and differential consequences, contributes to dehumanizing medicine, and is unacceptable to many professionals and patients.

Rationing of Specific Resources

Rationing becomes especially salient when problems result, not simply because of the unwillingness to provide more resources or allocate them fairly over some population (which may result in increased illness and death in a statistical sense), but when there is an inability to provide a life-saving resource to all identified individuals who seek it. Such situations involving the allocation of hearts, kidneys, livers, and lungs have been called 'tragic choices' because there are no correct or generally accepted rules for making the decisions as to who receives the resource and who dies (Calabresi and Bobbitt, 1978). A variety of criteria have been applied from one occasion to another, such as allocation by age, by community contribution, by the largest opportunity to benefit, by family circumstances, by ability to cooperate in the treatment process, and many more, but all raise serious ethical conflicts (Elster, 1992). Age is commonly suggested as a suitable criterion (Callahan, 1990) and has been used in making tragic choices (Aaron and Schwartz, 1984) but has been strongly contested. The most satisfactory criteria are medical ones based on likely technical success. But even these criteria do not adequately discriminate among potential beneficiaries, and social judgments and biases inevitably intrude (Halper, 1989). For example, judgments about who can best cooperate in treatment typically give preference to patients with more education, higher intelligence, attractive personality characteristics, and strong social attachments. Such judgments may appear intuitively correct but may be empirically incorrect.

Several analysts have concluded that because the equity issues cannot be resolved intellectually, the fairest mode of allocation is a lottery where persons with comparable medical qualifications have equal chances to be selected. There is little professional or public support for such selections because most people believe that consensual judgments can be made that would improve on any random process. Many rationally constructed strategies of allocation fail also because of public perceptions. The public places greater value on rescuing identified lives than on saving statistical lives (Hadorn, 1991).

Specific Rationing Policies/Practices

As evidence of inequalities in care increases, motivation to distribute limited health resources more equitably has led several countries to address rationing and seek explicit allocation principles and rules. The Oregon rationing program under Medicaid (the USA's federal-state partnership to provide health care to poor and disabled subgroups) was followed with great interest in Europe and perhaps represents the most extreme example of explicit rationing (Leichter, 1999). This effort sought to increase enrollees under its Medicaid program by

selecting to rank treatments consistent with judgments of efficacy and other community values. Through an elaborate process of evaluating condition–treatment pairs, a list was developed that prioritized these pairs. The basic idea was to establish a funding line, given available funds, that excluded from coveraging the lowest ranking condition–treatment pairs. The formal evaluation and ranking process and the use of community meetings to involve consumers and professionals in defining the values for performing the rankings excited the imagination of many outside observers. The process was seen as a rational and fair way to make difficult public decisions about what would and would not be paid for. The Oregon program, however, was not quite what it appeared to be. It was a very public and explicit format for designing the benefit structure. The Oregon Medicaid program is predominantly a managed care program in which HMOs are capitated to provide all funded services. Once capitated, HMOs are free to provide whatever services they deem necessary and need not conform to the list. Thus, the list was more an instrument for bargaining over rates than an explicit set of rules for rationing health care. Early implementation was aided by increased state funding and a favorable economy, which allowed the establishment of acceptable capitation rates. The political sustainability was challenged by fiscal crises that led to large restrictions in coverage (Oberlander, 2007). Oregon is presently rebuilding its Medicaid program with the help of the implementation of national health reform (Affordable Care Act).

The English NHS exemplifies implicit rationing. There are no explicit guaranteed benefits in the NHS but the system seeks to provide all necessary care within the resources available. Care is rationed by the availability of professionals and treatment facilities, by professional decisions about intensity of care and referrals, and by waiting lists for various types of services (Klein et al., 1996; Aaron and Schwartz, 1984). The extraordinary professional discretion within the NHS has led to efforts for greater equity and accountability in resource use. Under recent reforms, health authorities were given purchasing responsibilities and could limit the treatments covered. The political difficulty of doing so explicitly in any but marginal aspects of medical care became evident.

Ethical and Political Issues with Rationing Policies/Practices

A major ethical issue concerns the appropriateness of having physicians, acting as agents of their patients, also responsible for allocation decisions. Traditionally, doctors have been seen as sole agents of their patients, trying to get them what they need. Some believe it is unethical to dilute such agency with allocation responsibilities at the bedside. Others believe that physicians must have a dual responsibility for both their patients and the 'commons.' With the development of HMOs and managed care strategies in the USA, financial incentives are increasingly used to encourage doctors to allocate resources efficiently. There are worries that such incentives create ethical conflicts between the doctor's financial interests and patient care. Such conflicts are also common in unmanaged care when physicians have financial interest in diagnostic and treatment facilities to which they make referrals or in other practices for which they receive financial incentives. Studies have consistently found that such arrangements increase the use of ancillary services (Rodwin, 1993).

There is widespread agreement that it is appropriate and necessary for some governing authority to establish financial constraints on health care expenditures. All national systems have such constraints in place for at least public expenditures. Few believe that it is any longer possible, even in the richest of nations, to provide open-ended financing, allowing physicians to provide all they deem useful. A more-contested issue is the extent to which such an authority, once having established such limits, should specify how the budget should be divided among different types of facilities, modes of care, professional groups, types of patient needs, and treatments and procedures, and by what criteria. The extent of authority given to the nation, as compared with regions, health districts, states, or even smaller population entities, elicits tensions in many contexts. Issues also arise as to whether organizational rationing decisions are best made by management, specialty services, or professionals exercising their best individual judgment.

A possible model that blends explicit and implicit rationing is one emphasizing clinical advocacy within a structure of procedural justice. Such a model would explicitly constrain some clinician and patient decisions through clear rules that are reasonably justified, while also allowing doctors and patients to challenge decisions and argue for revision through dispute resolution mechanisms. Such compromise models offer opportunity for achieving more consensus.

Public views of rationing processes are important because any rationing process that lacks political legitimacy becomes unstable. One of the attractions of the Oregon program was its openness and the participation of large numbers of people. In contrast, in many areas of the USA, the implementation of utilization review (a form of explicit rationing) by managed care organizations overruling physician and patient definition of need has resulted in a significant political backlash. Managed care organizations are finding that transferring risk to professional groups (implicit rationing) kindles less friction than imposing explicit decisions on professionals and patients.

Priority Setting in Modern Nations

Increasingly most modern nations are seeking approaches to better establish priorities as they confront impressive but highly expensive innovations in health technologies and health care and constraints on public budgets. Scandinavia and other European countries have established formal governmental commissions on priorities that have engaged in useful but broad discussion about abstract principles of equity and efficiency but distant from the engaged politics of the many interest groups that affect decisions at the service level. Assessments of these efforts find that actual policy and practice are divorced from the recommendations made by these commissions and other discussions (Sabik and Lie, 2008).

One of the more successful ventures has been the National Institute for Health and Clinical Excellence (NICE) in England, first established in 1999. NICE recommendations determine publicly funded benefits in the NHS. NICE evaluates new health

technologies, pharmaceuticals and interventions, and develops clinical guidelines, taking cost-effectiveness and public and professional input into account. Its analyses and recommendations are focused on new products and approaches, but not long-established but untested ones, and, contrary to expectations, its decisions have not significantly influenced rising health care costs. Yet, NICE's evaluations, not only of efficacy but also cost-effectiveness, remain controversial in many quarters, and new reforms in 2004 introduce the alternative policy encouragement of value-based purchasing (Faden and Chalkidou, 2011).

In the USA, one of the most controversial sections of the Patient Protection and Affordable Care Act of 2010, the extensive health reform known as 'Obama Care,' was the establishment of a new 15-member Independent Payment Advisory Board (IPAD) required to recommend proposals to reduce Medicare spending if expenditures exceed target growth rates, with initial implementation in 2015. These proposals will have force unless the U.S. Congress adopts alternative proposals having an equivalent level of savings or if the President vetoes Congressional action and the veto is sustained. The tension involving this mechanism is illustrated by the prohibition that the Board may not submit proposals that would ration care, increase taxes, change Medicare benefits or eligibility, increase beneficiary premiums and cost-sharing requirements, or reduce low-income subsidies under prescription coverage. Although now law, there is great opposition to the IPAD, particularly among right wing Republicans. The toxicity in the USA of considering rationing explicitly is apparent. Nevertheless, rationing is ubiquitous and will increase in the future in the USA as well as elsewhere around the world.

It will take decades for rationing issues to be sorted out. Health care systems and facilities are experimenting with a range of strategies that may improve decisions and help achieve legitimacy for rationing. These include community representation and involvement in decision-making, surveys to elicit community preferences, focus on evidence-based services, use of treatment guidelines and pathways, involvement of patients in treatment decisions, and implementation of disease management approaches. Rationing and the ability to sustain whatever system is preferred depends on achieving a sufficient level of value consensus so that the difficult choices are widely perceived as fair and responsive to individual need.

See also: Health Care Markets: Theory and Empirical Evidence; Health Economics; Health Politics and Policy; Occupational Health.

Bibliography

Aaron, H.J., Schwartz, W.B., 1984. The Painful Prescription: Rationing Hospital Care. The Brookings Institution, Washington, DC.

Calabresi, G., Bobbitt, P., 1978. Tragic Choices. W.W. Norton and Company, New York.

Callahan, D., 1990. What Kind of Life: the Limits of Medical Progress. Simon and Schuster, New York.

Churchill, L.R., 1987. Rationing Health Care in America: Perceptions and Principles of Justice. University of Notre Dame Press, Notre Dame, IN.

Elster, J., 1992. Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens. Cambridge University Press, Cambridge.

Faden, R.R., Chalkidou, K., 2011. Determining the value of drugs – the evolving British experience. New England Journal of Medicine 364, 1289–1291.

Hadorn, D.C., 1991. Setting health care priorities in Oregon. Journal of American Medical Association 265, 2218–2225.

Halper, T., 1989. The Misfortunes of Others: End-stage Renal Disease in the United Kingdom. Cambridge University Press, Cambridge.

Klein, R., Day, P., Redmayne, S., 1996. Managing Scarcity: Priority Setting and Rationing in the National Health Service. Open University Press, Buckingham.

Leichter, H.M., 1999. Oregon's bold experiment: whatever happened to rationing? Journal of Health Politics, Policy and Law 24, 147–160.

Mechanic, D., 1979. Future Issues in Health Care: Social Policy and the Rationing of Medical Services. Free Press, New York.

Mechanic, D., 1995. Dilemmas in rationing health care services: the case for implicit rationing. British Medical Journal 310, 1655–1659.

Mechanic, D., 1997. Muddling through elegantly: finding the proper balance in rationing. Health Affairs 16, 83–92.

New, B. (Ed.), 1997. Rationing: Talk and Action in Health Care. BMJ Publishing Group, London.

Oberlander, J., 2007. Health reform interrupted: the unraveling of the Oregon health plan. Health Affairs 26, w96–w105.

Rodwin, M., 1993. Medicine, Money, and Morals: Physicians' Conflicts of Interest. Oxford University Press, New York.

Sabik, L.M., Lie, R.K., 2008. Priority setting in health care: lessons from the experiences of eight countries. International Journal of Equity Health 7, 4–13.

Symposium of the law and policy of health care rationing. University of Pennsylvania Law Review 140, 1992, 1505–2095.

Health: Definitions

Machteld Huber, Louis Bolk Institute, Driebergen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Health etymologically means being whole. Ancient philosophies describe health as a balance between qualities. Hippocrates' view on health, as a balanced mixture of four humors, remained dominant till the sixteenth century. Aristotle stressed development of eudaimonia. From the sixteenth century the biomedical view emerged among physicians, whereas philosophers still stressed the larger context. The World Wars of the twentieth century stimulated the creation of international organizations on health, and the UN preparations led to the WHO. The minutes of these preparations express the idealism with which the WHO was established, as expressed in the WHO definition of health. The WHO aims at monitoring, so a family of classifications was designed, first mainly on diseases, later on functions. The WHO definition met criticism for its absoluteness and being static, and different definitions were proposed by physicians as well as philosophers. Three components are consistent in these proposals: capacity to perform, individual fulfillment, and relation to the social environment.

Etymology

Etymologically the English word 'health' means wholeness, being whole, complete, sound and well. To 'heal' literally means to make whole. Both words go back to the old English word 'hal' and the old German word 'heil.' The ancient Greek word for health was 'euexia,' which means to be in a good condition.

Historical Definitions of Health

Ever since human documentation existed, expressions are found where man differentiates between human states of well-being and 'illnesses,' the latter considered to be abnormal, undesirable and to be healed, if possible. In different ancient philosophies (e.g., Chinese, Indian, Egyptian, or Greek) concepts of illnesses and a state of 'good health' are described, and also the ways how man can achieve good health (Lindeboom, 1982). In these concepts, good health is conceptualized as a state of balance between different 'elements,' which represent different qualities, e.g., the five elements (wood, fire, earth, metal, and water) and the Yin and Yang qualities in Chinese philosophy (Taoism), or the four humors in Greek medicine (black and yellow bile, phlegm, and blood), which originated from knowledge of ancient Egypt and Mesopotamia.

Health is described to be achievable and influenced by lifestyle habits and in certain circumstances, if fate allows, to be influenced by god(s), as mediated by priests in temples. 'Hygeia' was the name of the goddess of health in ancient Greece, the daughter of Asclepius, who represented a good way of living.

Physicians and Philosophers

Over the course of history, physicians as well as philosophers have developed views on the nature of health. Hippocrates (approximately 460–370 BC) (London, 1984), is still considered to be 'the father of Western medicine,' as he was the first to emphasize the importance of a good physical examination of the patient; he created a new paradigm, discriminating in humans four humors: black bile (*melan chole*), yellow bile (*chole*), phlegm (*phlegma*), and blood (*haima*), each corresponding to one of the traditional four elements and temperaments. A healthy situation was the balance between the four humors, and Hippocrates considered diet to be the most influential lifestyle factor in changing the relative proportions between the four, in addition to influences from climate, winds, quality of the water, and the soil.

Historically even more influential Greek-Roman doctor Galenus (131–216) followed and refined the teachings of Hippocrates and described six lifestyle factors, '*res non naturalia*,' which should be in balance in order to maintain health: food and drink (*cibus et potus*), sleep and being awake (*somnus et vigilia)*, light and air (*aer*), secretions and excretions (*secreta et excreta*), work and relaxation (*motus et quies*), and emotions (*affectus animi*).

The 'father of Western philosophy,' Aristotle (384–322 BC) does not discuss 'health' as such, but states that extremes in bodily condition should be avoided and maintaining a proper balance (the mean) is a virtue (Osborne, 2011). Yet, in his ethics he considers '*eudaimonia*' or well-being to be the final goal and good for man. Eudaimonia is an important concept as it is rediscovered in modern views on health. Eudaimonia literally means 'the state of having a good indwelling spirit, a good genius'; in modern translations, it is described as 'happiness' or 'personal well-being.' Eudaimonia is not a static state but a continuous process of development, to fulfill one's potentials and to flourish. This implies self-realization, by which man achieves this happiness and personal well-being.

These views of Hippocrates, Galenus, and Aristotle, about the human potential to be in a state of balance and the virtue of developing oneself, remained highly influential in Western medicine and thinking over fifteen centuries.

From the Sixteenth Century – A New Paradigm Dawns in Medicine

The Physicians' View

A different way of interpreting human health and disease started emerging in medicine when Vesalius (1543) published his studies on the anatomy of the human body. These were based on his own observations by the dissection of bodies, which was considered sacrilege until then. A century later William Harvey (1628) described the system of blood circulation, which was until then thought to be a tidal movement, like ebb and flow. The Galenic humoralism was definitely displaced in 1858, when Rudolph Virchow published his theory on cellular pathology (Virchow, 1858, 1862). In the same period, Semmelweiss discovered the protective effect of disinfective measures against deathly puerperal fever; soon thereafter Pasteur described the existence of bacteria, which he could see in the microscope, and the ways to disarm these (e.g., by pasteurization). From then the paradigms of cell physiology, microbiology, and pathological anatomy became the dominant views in medicine. Diseases were no longer understood as caused by misbalances of qualities, but were to be searched and understood in the physical body, in organs, in cells and in microbes, and should be treated there. Health became the absence of disease (Shorter, 2009).

Parallel to this transition in medical thinking, public health developed in the fast growing cities due to the industrial revolution and by sanitary measures, clean water provision, garbage disposal as well as first vaccinations, infectious diseases like cholera, typhus, and small pox were eradicated and became extinct. The new medical paradigm proved to be effective.

Contemporary Philosophers in the Eighteenth and Nineteenth Century

In the period of this paradigm shift in medical thinking, philosophers still had a broader perspective on health, than as absence of disease. From Goethe (1749–1832) comes the quote "What is the highest virtue on earth? To be healthy? No: to become healthy" (Hygiene Museum Dresden). Nietzsche (1844–1900) discriminated between 'small health' and 'big health' (Verduin, 1998). While 'small health' in his view is concerned with the daily derangements and illnesses with which doctors deal, attaining 'big health' means saying a full YES toward life, with all its insecurities, tragedies and finally death. Big health includes small health, but is connected to growth far beyond that. Rilke (1875–1926) stated "Illness is the means by which my organism frees itself from unfamiliarity. I need just to support my body in being ill, because that is my development" (www.rilke.de/briefe/120804).

Health in the Twentieth Century and the Foundation of the WHO

From the twentieth century, public health started to be organized on a national and international scale, operated by big health organizations (Borowy, 1988). In 1902, the Pan American (international) Sanitar Bureau was founded in Washington DC, while in 1907 the Office International d'Hygiène Publique followed, with headquarters in Paris. In 1913, the Rockefeller Foundation as a first nongovernmental organization was initiated with the aim to enhance the well-being of all mankind.

After World War I, the League of Nations Health Organisation was established, which stood out as the first official body with a global perspective and the broadest work spectrum. In 1919, the League of Red Cross Societies was organized, which developed health programs with an international outlook.

In 1943, a Health Division of the United Nations Relief and Rehabilitation Administration was initiated by the U.S. President Roosevelt "for the relief of victims of war in any area under the control of any of the United Nations through the provision of food, fuel, clothing, shelter and other basic necessities, medical and other essential services." The term 'United Nations' referred to the Allies of World War II.

World War II led to the wish to establish a world organization like the present United Nations, to preserve world peace toward the future. When in 1945 in San Francisco 50 nations met to draft the Charter of the future United Nations, among them there were three medical doctors from Norway, Brazil, and China. They met 1 day for a 'medical lunch,' where the Norwegian Karl Evang suggested "Why don't we start a new health organization?" (WHO, 1988). The idea was to unite the different existing organizations in one single health organization. The men succeeded in having the word 'health' inserted in the Charter of the United Nations and they recommended the general conference to establish an international health organization. This was approved unanimously and this was the very beginning in 1945 of the WHO.

A Technical Preparatory Committee was established to prepare a draft Constitution of the WHO. This committee of 16 men and advisors met in Paris 22 times between March and June 1946. Their work resulted in the draft Constitution of the WHO, which was signed on 22 July 1946 by the representatives of 61 states at the International Health Conference in New York. Thereafter all countries had to ratify the Constitution, which was completed in 1948.

Between 1946 and 1948, an Interim Committee, presided by Andrija Stampar from Yugoslavia, ensured continuity. This Committee decided in favor of Geneva as seat for the headquarters of the WHO, and it was in Geneva at the first World Health Assembly that the Constitution entered into force on 7 April 1948.

The Constitution of the WHO

The Constitution of the WHO starts with the Preamble, stating the principles upon which the whole document is based (WHO, 1948):

> The States parties to this Constitution declare, in conformity with the Charter of the United Nations, that the following principles are basic to the happiness, harmonious relations and security of all peoples:
>
> Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity.
>
> The enjoyment of the highest attainable standard of health is one of the fundamental rights of every human being without

distinction of race, religion, political belief, economic or social condition.
The health of all peoples is fundamental to the attainment of peace and security and is dependent upon the fullest co-operation of individuals and States.
… Etcetera.

The definition of health has a central position in the Preamble Constitution of the WHO. There had been some discussion whether a definition of health should be in the Constitution. Szeming Sze, a public health doctor from China, one of the three who initiated the idea for the WHO and a member of the Technical Preparatory Committee, describes how the phrasing of the definition of health came about (WHO, 1988). Some detail on this process is also to be found in the minutes of the Technical Preparatory Committee (WHO, 1947). A subcommittee of four physicians, Canadian psychiatrist Brock Chisholm (later to become the first Director-General), Argentinian psychiatrist Gregorio Bermann, Joseph Cancik from Czechoslovakia, and Sze himself had 'some pleasant academic discussions' and decided to emphasize mental health as well as the preventive side of health and so they came to the wording of health as not merely the absence of disease. The minutes of the Technical Preparatory Committee describe several preversions of the definition, with terms like 'physical fitness' and 'positive health.'

Idealistic Intentions

The WHO definition of health is, like the Constitution as a whole, phrased very idealistically. One should realize these are formulated in a reaction to the horrors of World War II. The atmosphere in the Preparatory Committee may be characterized by a phrase in the minutes of this Committee where Chisholm is cited with the following phrase: "The world was sick, and the ills from which it was suffering were mainly due to the perversion of man, his inability to live at peace with himself. It was in man himself that the cause of present evils should be sought. The microbe was no longer the main enemy: science was sufficiently advanced to be able to cope with it admirably, if it were not for such barriers as superstition, ignorance, religious intolerance, misery and poverty. These psychological evils must be understood in order that a remedy might be prescribed, and the scope of the task before the Committee, therefore, knew no bounds. For that reason Dr. Chisholm associated himself with the 'visionaries.' What was taking place in these meetings would be of great historical importance if all members aimed at universal and worldwide achievement. To do this, they might find it necessary to bite off more than they could chew, but the alternative was complete chaos. They should do their utmost to bring all the peoples of the world together in the service of physical, social and emotional health."

Chisholm refers to penicillin, which became available for general use, after Fleming had discovered it in 1928. In a time where infectious diseases were still dominant, this discovery nourished great expectations to ban these diseases. Yet, the broadness of the definition of health, beyond the absence of disease, was at the time ground breaking and is owed to this special group of physicians, who directed this process of establishment.

In several charters, declarations and statements the idealistic aims of the WHO were renewed every few years since the inception of the definition of health, (www.who.int/healthpromotion/conferences/en/; www.who.int/trade/glossary/story039/en/). Of these, the best known is the 'Alma Ata Declaration,' which emerged from the UNICEF/WHO Conference on Primary Health Care of 1978, addressing health as a fundamental human right, health inequalities and the connected need for economic and social development; the 'Health for All by the year 2000' declaration of 1981, as target for the WHO; and especially the condensed and strongly formulated 'Ottawa Charter for Health Promotion' of 1986. There the WHO definition was described in connection to the conditions by which this state could be met: "To reach a state of complete physical mental and social wellbeing, an individual or group must be able to identify and to realize aspirations, to satisfy needs, and to change or cope with the environment. Health is, therefore, seen as a resource for everyday life, not the objective of living. Health is a positive concept emphasizing social and personal resources, as well as physical capacities. Therefore, health promotion is not just the responsibility of the health sector, but goes beyond healthy lifestyles to wellbeing." As fundamental "prerequisites for Health are mentioned: peace, shelter, education, food, income, a stable co-system, sustainable resources, social justice and equity." The charter calls for 'building healthy public policy' and the 'creation of supportive environments.'

During the 2013, Helsinki 8th Global Conference on Health Promotion the tone was a somewhat less idealistic and the 'implementation gap' was addressed. It was concluded that the 'Health in All Policies' approach was one of the key domains of shared governance toward health in the twenty-first century. In all these texts, the original definition of health of the WHO was maintained.

Operationalization of Health by the WHO since 1948

The WHO Classification Systems

The WHO intends to monitor population health around the world and is thus in need of health indicators (Chatterji et al., 2002). Yet, the WHO itself recognizes that the definition is not sufficient to develop operational indicators of health. The definition places health in the broad context of well-being in general, whereas health in itself is separate of well-being, yet instrumental for it. Health is first of all an attribute of an individual person and comprises states or conditions of the human body and mind; attempts to measure health must thus include measures of functions of body and mind.

There is general acceptance of describing health states of individuals, in terms of multiple domains and in self-report instruments. For measurement purposes of health status valid, reliable, and comparable indicators are needed. According to the WHO, a measurement instrument requires:

- a classification of health status domains,
- specification of a set of domains necessary and sufficient to describe health states for measurement purposes,

- specification of what is being measured in each domain,
- a common understanding of what is full health versus exceptional talent in any given domain, and
- if it is required, to construct summary measures of average level of population health, a method to place a single cardinal value on the overall level of health associated with a health state defined on multiple domains (Chatterji et al., 2002).

Since 1948, the WHO designed several classification systems, together described as the WHO Family of International Classifications (www.who.int/classifications). At first, the focus was on causes of mortality, diagnoses, and causes for morbidity, for which the International Classification of Diseases (ICD) and Related Health Problems was developed, which is the international standard for epidemiological purposes.

However, gradually thinking about health states evolved and in order not to just evaluate mortality and diseases, but also disabilities, functioning and health states, the International Classification of Functioning, Disability and Health (ICF) (previously the International Classification of Impairments, Disabilities and Handicaps (ICIDH)) was elaborated (**Figure 1**). The ICF measures six domains of activity and participation, by a person's capacity, as well as performance. The gap between capacity and performance is considered to reflect the impact of the actual environment. The ICF goes beyond health condition, functions, activity and participation, and also takes environmental factors into consideration. The domain of personal factors is least elaborated, in fear of 'blaming the victim,' e.g., in cases of health damage through inequalities.

Quality of Life and Well-Being and Health-Related Quality of Life

Objective indicators of health and functioning as specified in the ICF need to be distinguished from a person's subjective appraisal of well-being and quality of life (QoL). It can very well be that a person with impairments still experiences a good QoL; this is called the 'disability paradox' (Faber von et al., 2001). In this context, the WHO identifies 'QoL' as "an individual's perception of their position in life in the context of the culture and value systems in which they live, and in relation to their goals, expectations, standards, and concerns" (Chatterji et al., 2002).

The WHO developed several instruments for measuring well-being and QoL: the WHOQoL-100 (www.who.int/mental_health) with 100 items covering six domains and the WHOQoL-BREF (www.who.int/substance_abuse) with 26 questions in four domains, which are worldwide one of the most widely used instruments for assessing QoL. The short WHO-5 well-being index (www.sayhelp.net) connects well-being mainly to mental health.

General QoL should be distinguished from Health-related QoL (HR-QoL), i.e., QoL perceived in relation to symptoms, impairments, and functions.

For HR-QoL, the best-known measurement instruments (from outside the WHO) are the SF-36, the Health Utility Index (HUI), the Nottingham Health Profile (NHP), and the EQ-5D.

Many more measurement approaches have been developed, like the measurement of QALY's (Quality-adjusted Life Years) or DALY's (Disability-adjusted Life Years), but these will not be described here.

Contemporary Philosophers of the Twentieth and Twenty-First Century on Health

The WHO was designed and further developed mainly by members of the medical profession. Yet 'health' is not a topic that concerns just physicians – it concerns humanity as a whole and during the existence of the WHO also

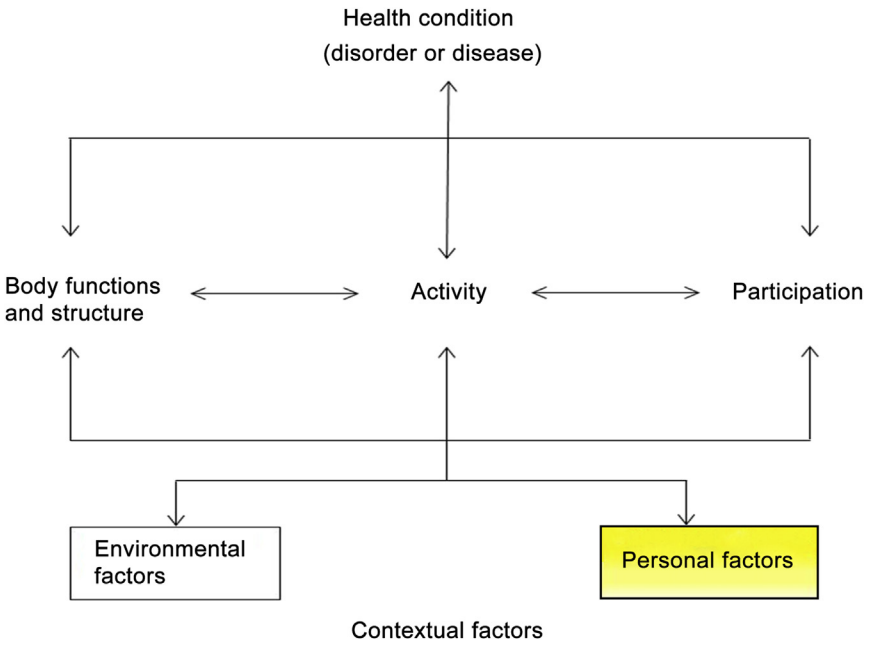

Figure 1 The different domains, included in the international classification of functioning, disability and health.

philosophers have thought about health. Some influential views should be mentioned.

In 1975, Christopher Boorse presented his Bio-Statistical Theory (BST) (Boorse, 1975, 1977), theorizing that 'health' is a state within a range of species typical normal functioning and thresholds in the organism; consequently disease is abnormal functioning. Health is thus the absence of disease, with a strict demarcation. Normal functions can empirically and statistically be discovered by the natural sciences and are thus objective and value-free, according to Boorse. The need is to develop enough and appropriate reference classes, connected to functions. Despite diverse criticism, the appeal to reference classes is commonly accepted in health care and the theory reached the status of background theory of health and medicine (Venkatapuram, 2013).

Out of discontentment with the prevailing view of health as the absence of disease, the Swedish philosopher Lennart Nordenfelt published in 1987 (revised in 1995) his theory of 'health as the ability to achieve vital goals.' Nordenfelt described a person's vital goal as "a state of affairs that is such that it is a necessary condition for the person's minimal happiness in the long run." The stipulation 'in the long run' is aimed at avoiding the definition of health to be centered on immediate pleasure and instead, be more in line with long-term happiness, thriving, or flourishing such as Aristotle's *eudaimonia*. More recently (1995), he rephrased the definition of vital goals as "a state of affairs which is either a component of or otherwise necessary for the person's living a minimally decent life. This includes more than survival" (Nordenfelt, 2001; Venkatapuram, 2013).

In 2013, Venkatapuram proposed a modification of Nordenfelt's conception of health (Venkatapuram, 2013). He criticizes Nordenfelt's definition in that it conceptualizes vital goals as social, cultural and ethically relative and, thus, subjective and not normative; in addition, Nordenfelt does not specify a set of vital goals. Venkatapuram aims to describe a human species-wide conception of basic vital goals or 'central human capabilities and functionings,' as inspired by the capabilities approach of Martha Nussbaum, which sets a norm.

Venkatapuram reasons that 'health is a person's ability to achieve or exercise a cluster of basic human activities.' He refers to Nussbaum's 10 capabilities, which consist in a life worthy of human dignity: (1) being able to live a normal length of life span; (2) having good health; (3) maintain bodily integrity; (4) being able to use senses, imagination, and think; (5) having emotions and emotional attachments; (6) possess practical reason to form a conception of the good; (7) have social affiliations that are meaningful and respectful; (8) express concern for other species; (9) able to play; and (10) have control over one's material and political environment.

Venkatapuram stresses that these 10 capabilities, as moral entitlements emanating from a person's human dignity, become the source of political principles for a liberal pluralistic society; ensuring that each member achieves a threshold level of these 10 central capabilities become primary political goals. He expects different societies to define thresholds levels of each capability depending on their history and resources, and thus defining health as a minimal conception of human well-being (Venkatapuram, 2013).

A clear ethical statement is being made, which in its intention is similar to the ideals of the WHO. It is remarkable that a philosophy such as the one by Boorse reaches the medical establishment, whereas the philosophies of Nordenfelt and Venkatapuram do not cross the apparent gap between philosophy and medicine. Or are new developments promising for more exchange and integration?

More Health Concepts Outside the WHO in the Twentieth and Twenty-First Century

In the twentieth and twenty-first centuries also nonphilosophical professionals outside the WHO thought about the definition of health, some of whom are mentioned here.

In 1941 Henry Sigerist, a Swiss born American medical historian, analyzing the relevance of health for human welfare, stated that "A healthy individual is a man who is well balanced bodily and mentally, and well-adjusted to his physical and social environment. He is in full control of his physical and mental faculties, can adapt to environmental changes, so long as they do not exceed normal limits, and contributes to the welfare of society according to his ability. Health therefore is not simply the absence of disease; it is something positive, a joyful attitude towards life, and a cheerful acceptance of the responsibilities that life puts upon the individual" (Sigerist, 1941).

René Dubos, a French-born American microbiologist and humanist (well known by his 'Think global, act local') stated in 1968 that in his view "Health is a *modus vivendi* enabling imperfect men to achieve a rewarding and not too painful existence while they cope with an imperfect world… Health and vigour can be achieved in the absence of modern sanitation and without the help of Western medicine. Man has in his nature the potentiality to reach a high level of physical and mental wellbeing without nutritional abundance or physical comfort" (Dubos, 1968).

Ivan Illich, Austrian philosopher, Roman Catholic priest and critic of the institutions of contemporary Western culture, described health in 1975 as "the ability to adapt to a changing environment, to growing up and to aging, to healing when damaged, to suffering and to the peaceful expectation of death. Health embraces the future as well, and therefore includes anguish and the inner resources to live with it" (Illich, 1975).

Stokes, Noren, and Shindell described in 1982 health as "A state characterized by anatomical, physiological, and psychological integrity; an ability to perform personally valued family, work, and community roles; an ability to deal with physical, biological, and psychological stress; a feeling of well-being; and freedom from the risk of disease and untimely death" (Stokes et al., 1982).

In 1995 Last described health as "A state of equilibrium between humans and the physical, biological, and social environment, compatible with full functional activity" (Last, 1997).

In 1996, Tarlov analyzed the definitions of health of the past half century and concluded that there is a remarkable consistency of three conceptual components, being:

1. The capacity to perform (relative to potential).
2. The achievement of individual fulfillment; the pursuit of values, tasks, needs, aspirations, and potential.
3. In a social environment, good health provides the potential to negotiate demands of the social environment.

Salutogenesis

In addition to these definitions of health, the American-Israeli medical sociologist Aaron Antonovsky introduced in 1979 a new term on health, which he called salutogenesis (Antonovsky, 1979), i.e., the study of generating health, as opposed to pathogenesis, the study of diseases. Antonovsky found during his qualitative studies in Israel that a certain group of women, who had survived concentration camps, did not have negative health outcomes in response to their experiences and could manage stress and stay well. Apparently, some people achieve health despite their exposure to potentially disabling stress factors. Antonovsky studied the characteristics of these resilient people and found three personality traits, which in combination he called the Sense of Coherence (SOC).

SOC consists of comprehensibility, manageability, and meaningfulness. Antonovsky described meaningfulness to be the strongest factor. The importance of meaningfulness is in concordance with the findings of psychiatrist Viktor Frankl, based on his own observations among survivors of concentration camps (Frankl, 2006).

SOC reflects two of the three previously described components in Tarlov's analysis of definitions of health: The component of capacity to perform reflects manageability, whereas the achievement of individual fulfillment reflects meaningfulness. Where Antonovsky adds comprehensibility, Tarlov describes the importance of the social environment.

Antonovsky rejects the view of health and disease as sheer opposites; rather he describes the two as extremes on a continuum and stresses the possibilities to strengthen SOC, in order to enhance health. His views are further elaborated, among others, by Lindström and Eriksson (2010).

Increasing Criticism of the WHO Definition of Health in the Twenty-First Century

In 2008, Alex Jadad from Canada initiated a global conversation using a Web-based discussion forum (blog) for which he invited participants in the British Medical Journal (Jadad, 2008). This blog was like a reenactment of the discussion of 60 years ago, which led to the WHO definition, with the major difference of using the power of social media. Shortly thereafter, an editorial published in The Lancet was entitled 'What is health? The ability to adapt' (Lancet, 2009).

At that time in the Netherlands also the urgency was felt for redefining health in a more dynamic and operational way. At the end of 2009, a 2-day international expert conference was organized about 'health.' First limitations of the WHO definition were discussed and summarized as follows: (1) the absoluteness of the word 'complete,' which is utopian and stimulates medicalization; (2) the present demography of diseases, which shows no longer infectious, but mainly chronic diseases; and (3) the difficulty with operationalization, as 'complete' is not measurable. There was broad support for moving from the present static formulation of the WHO definition of health, toward a more dynamic description based on resilience and the human capacity to cope and to maintain and restore one's integrity, equilibrium, and sense of well-being.

Participants questioned whether a new formulation should be called a definition, because this implies set boundaries and trying to arrive at a precise meaning. They preferred the definition should be replaced by a concept or conceptual framework of health. A general concept, according to sociologist Blumer, represents the characterization of a generally agreed direction in which to look, as a reference. But operational definitions are also needed for practical life, such as for measurement purposes. The preferred description for a dynamic general concept was "Health as the ability to adapt and to self-manage, in the face of social, physical, and emotional challenges." This was published in the British Medical Journal (Huber et al., 2011).

In 2012–13, this general concept was further elaborated by Huber into the broad operational concept of 'positive health,' integrating objective medical and functional facts on the one hand, and SOC and meaningfulness on the other hand, thus approaching the ideals of Nordenfelt and Venkatapuram, and presenting optional concepts for 'personal factors' in the ICF. At present, (2014) a scientific publication is in preparation.

See also: Functioning, Disability and Health, International Classification of; Health Education and Health Promotion; Health Outcomes, Assessment of; Human Development and Health; Resilience.

Bibliography

Antonovsky, A., 1979. Health, Stress and Coping. Jossey-Bass Publishers, San Francisco, ISBN 978-0-671-02337-9.
Boorse, C., 1975. On the distinction between disease and illness. Philosophy & Public Affairs 5, 49–68.
Boorse, C., 1977. Health as a theoretical concept. Philosophy of Science 44, 542–573.
Borowy, I., 1988. The league of nations health organization: from European to global health concerns? In: Andresen, A., Hubbard, W., Ryymin, T. (Eds.), International and Local Approaches to Health and Health Care. University of Bergen 2010, pp. 11–30.
Chatterji, S., Ustün, B.L., Sadana, R., et al., 2002. The conceptual basis for measuring and reporting on health. WHO. In: Global Programme on Evidence for Health Policy Discussion Paper No. 45.
Dubos, R., 1968. Man, Medicine, and Environment. Praeger, New York, p. 69.
Faber von, M., Bootsma-van der Wiel, A., van Exel, E., et al., 2001. Successful aging in the oldest old: who can be characterized as successfully aged? Archives of Internal Medicine 161, 2694–2700.
Frankl, V., 2006. Man's Search for Meaning, Part One, 'Experiences in a Concentration Camp'. Pocket Books, ISBN 978-0-671-02337-9.
Goethe; quote on glass plate at the entrance of the Hygiene Museum in Dresden.
Harvey, W., 1628. De motu cordis. Level 12 Spec Coll Hunterian Ac.4.18, Frankfurt.
Hippocrates, 1984. Nature of Man. Humours. Aphorisms, vol. IV. Loeb Classical Library, No 150, London.
Huber, M., Knottnerus, J.A., Green, L., et al., 2011. How should we define health? British Medical Journal 343, d4163.
Illich, I., 1975. Medical Nemesis: The Expropriation of Health. Marion Boyars, London, p. 167.
Jadad, A.R., 2008. How should health be defined? British Medical Journal 337 (a2900), 1361–1364.
The Lancet 373 (9666), 7 March 2009, 781.
Last, J.M., 1997. Public Health and Human Ecology. Conn. Appleton & Lange, Stanford.
Lindeboom, G.A., 1982. Begrippen in de Geneeskunde. Editions Rodopi, Amsterdam, 33–41.
Lindström, B., Eriksson, M., 2010. The Hitchhiker's Guide to Salutogenesis. Salutogenic Pathways to Health Promotion. Folkhälsan, Helsinki.

Nordenfelt, L., 1995. On the Nature of Health: An Action-Theoretic Approach, second ed. Kluwer Academic, Dordrecht.
Nordenfelt, L., et al., 2001. Health, Science, and Ordinary Language. Rodopi, Amsterdam.
Osborne, R., 2011. The History Written on the Classical Greek Body. Cambridge University Press, Cambridge, New York.
Shorter, E., 2009. Doctors and Their Patients, a Social History, sixth ed. Transaction Publishers, New Brunswick, NJ.
Sigerist, H.E., 1941. Medicine and Human Welfare. Yale University Press, New Haven.
Stokes, J., Noren, J.J., Shindell, S., 1982. Definition of terms and concepts applicable to clinical preventive medicine. Journal of Community Health 8, 33–41.
Tarlov, A.R., 1996. Social determinants of health: the sociobiological translation. In: Blane, D., Brunner, E., Wilkinson, R. (Eds.), Health and Social Organization. Routledge, New York, pp. 71–93.
Venkatapuram, S., 2013. Health, vital goals, and central human capabilities. Bioethics 27 (5), 271–279.
Verduin, P., 1998. Thesis: De vraag naar het lichaam. Elsevier/De Tijdstroom, 25–35.
Vesalius, A., 1543. De humani corporis fabrica libri septem, Basel Level 12 Spec Coll Hunterian Z.1.8.
Virchow, R., 1858. Vorlesungen über Pathologie. A. Hirschwald, Berlin.
Virchow, R., 1862. Die Cellularpathologie in ihrer Begründung auf physiologische und pathologische Gewebelehre. A. Hirschwald, Berlin.
WHO, 1947. Minutes of the Technical Preparatory Committee for the International Health Conference Held in Paris from 18 March to 5 April 1946. Citation Chisholm, p. 13.
WHO, 1948. www.who.int/governance/eb/who_constitution_en.pdf.
WHO, 1988. From small beginnings. Interview with szeming sze. World Health Forum 9.

Relevant Websites

www.rilke.de/briefe/120804.htm – Rilke: Franz Xaver Kappus.
www.sayhelp.net/pdfs/1345285728_WHO-5%20well%20being%20index.pdf – World Health Organization Well-Being Index.
www.who.int/healthpromotion/conferences/en/ – World Health Organization Global Conferences on Health Promotion.
www.who.int/trade/glossary/story039/en/ – World Health Organization Trade, Foreign Policy, Diplomacy and Health.
www.who.int/classifications – World Health Organization Classifications.
www.who.int/mental_health/publications/whoqol/en/ – World Health Organization Mental Health.
www.who.int/substance_abuse/research_tools/whoqolbref/en/ – World Health Organization Management of Substance Abuse.

Health Economics

Åke G Blomqvist, University of Western Ontario, ON, Canada; and Carleton University, Canada; Ottawa, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Advancing medical and pharmaceutical technology has allowed us to lengthen human life spans and reduce suffering, but has also increased the health sector's use of scarce economic resources. Health economics studies the evaluation techniques and institutional arrangements that can be employed to promote efficient use of health resources and to control aggregate health care costs, as well as the social policy objective of guaranteeing that sick people have access to needed health care regardless of ability to pay.

Since the turn of the century, health care has continued to take up an increasing share of total economic resources in most countries in the world. In 1997, health care spending as a share of GDP fell in the range 6.7–13.6% for the G7 countries; by 2010, the range was 9.5–17.6%. The health economics literature has expanded apace. In 2011–12 alone, just over 10 years after the appearance of the first two-volume handbook in the field (Culyer and Newhouse, 2000), three new major survey works were published: Glied and Smith, 2011; Danzon and Nicholson, 2011; Pauly et al., 2012. Readers looking for recent references to any of the topics discussed in this article should look to these volumes in the first instance.

Health care remains one of the economic sectors in which views on the proper role of government intervention, as opposed to reliance on markets and competition, continue to be hotly debated. In some European countries, government has assumed responsibility for both the funding and production of health services; in others, funding is through government plans but production of services is mostly by private agents, while in the U.S. and a few other countries private insurance plays a large role on the funding side as well.

Much of the early literature in health economics focused on the reasons why unregulated private markets are unlikely to perform well in health care, including uncertainty and imperfect information. The demand for health services is state-contingent (most health services are demanded only when people have fallen ill), and while markets for private health insurance offer consumers an opportunity to protect themselves against the financial risk of costly health care, they accomplish this only imperfectly. In his classic article which some characterize as the origin of health economics, Arrow (1963) argues that inadequate medical knowledge and imperfect state-contingent markets put consumers in a very weak position relative to providers when they need to buy health services. As medical and pharmaceutical technology has advanced and more and more resources have been devoted to health care, economic analysis has contributed to the design of regulatory approaches to deal with this problem, and has interacted with the development of new institutional arrangements in private health care and insurance markets that have evolved as a response to it.

Insurance Protection versus Moral Hazard Effects

Under ideal conditions, efficient health insurance would consist of state-contingent lump-sum payments that would depend only on the nature of the consumer's health problem (Goddeeris, 1984b). In practice, however, it is not possible to write insurance contracts that cover all possible health problems, so benefits under conventional insurance contracts typically are specified instead as a share of the cost of the health services that the insured has used. Implicitly, such contracts therefore act as a subsidy on ill consumers' utilization of health services. As with other subsidies, this tends to make them use more than the economically efficient quantity (Pauly, 1968). This effect (referred to in the literature as 'moral hazard') implies some degree of waste in the conventional microeconomic sense, in an amount that is larger the higher is the share of cost that is covered by the insurer. The resulting trade-off between the objectives of more complete insurance coverage and avoidance of overutilization of health services has been a central topic in much of the health economics literature, beginning with Zeckhauser (1970) and in the multimillion dollar-controlled trial of differing degrees of insurance carried out by the Rand Corporation in the early 1970s (Newhouse and The Insurance Experiment Group, 1993). While the issue of patient cost-sharing ('user fees') remains controversial, the results of the Rand study supported the idea that some degree of patient cost-sharing could lead to substantial payoffs in the form of reduced overutilization of health services. Although there are exceptions in which patients pay nothing out of pocket for hospital and physician services (the U.K. National Health Service (U.K. NHS), Canada's provincial insurance plans), most private and government insurance plans today require patients to pay at least a share of the cost, subject to an annual maximum of some kind.

An approach to the moral hazard problem that has become popular in places like Singapore and China, as well as in the U.S., is that of Medical Savings Accounts (MSAs). When an insurance plan includes MSAs, enrolees must, in addition to paying the regular insurance premium, also contribute an amount to an individual account (the MSA), the balance in which they can only use to pay for their own health services in the future. In the U.S., the popularity of MSAs is due largely to the fact that contributions to MSAs are considered a nontaxable

benefit, so designating part of employees' compensation as MSA contributions reduces the income tax they must pay. The rationale for compulsory MSA contributions in other places (such as Singapore) is less easy to understand, since their only effect is to restrict the way citizens can use some of their income. However, the growing evidence from behavioral economics that individuals do not appear to make rational decisions in the face of risk and uncertainty can perhaps also be used to justify the gentle 'nudging' that is inherent in compulsory MSA contributions (Thaler and Sunstein, 2009).

Information Asymmetry, Agency, and Supplier-Induced Demand (SID)

Imperfect information on the part of insurance plans regarding the exact nature and severity of the consumers' health problems is the reason why health insurance typically pays all or part of actual health care costs, rather than lump-sum indemnities that depend on the consumers' diagnoses. But imperfect information is also a problem for patients. Sick individuals often do not know what is wrong with them, or to what extent their pain and suffering will be alleviated by different treatment approaches or drugs. Although patients today are becoming more adept at finding information about their treatment options, they still have to rely to a large extent on the advice of their doctors in deciding what services to buy. Much of the literature in health economics has focused on this asymmetry of information between doctors and patients: Coupled with the fact that it is difficult for sick people who need care urgently to compare the terms offered by competing sellers, consumers are in a weak bargaining position relative to providers in health services markets.

If doctors can influence patients' demand for their services through the advice they provide, as suggested in the literature on SID, and competition among providers is weak, the result may be both high prices and larger-than-efficient volumes of services in markets where doctors can set their own fees. The empirical evidence on the extent to which SID exists remains fragmentary and ambiguous. However, markets of this kind have become much less common over time. Levels of medical fees in most countries with government funding of health care are not market-determined but regulated or negotiated between governments and provider groups. In the U.S., they are increasingly determined in advance of treatment under contracts between providers and insurance plans who obviously have more bargaining power than individual patients. Moreover, other methods of paying providers than fee for service are becoming more common; often these alternative payment methods are structured in such a way that providers have less incentive to supply large volumes of services. These trends are consistent with the suggestion that SID is a significant problem under the conventional market forms.

As the focus has shifted from demand-side incentives (patient cost-sharing) to those that are inherent in the methods used to compensate providers (supply-side incentives), health economics has drawn on agency theory, which analyzes efficient contracts in situations where an agent with better information (in this case, the doctor) is paid to make decisions that affect the welfare of a principal (the patient) with less information. An early application of agency theory related to compensation of corporate managers; in that case the principals are the shareholders, who have less information than the managers who act as their agents. In health care, however, a key feature of the principal–agency relationship is the fact that most of the costs are paid by third parties, so that the doctors' decisions affect not only the well-being of the patients, but also of the collective that bears most of the cost (taxpayers, members of private insurance plans). To be efficient, the contracts under which the agents provide health care must give them incentives to take account both of the interests of the individual patients, and the costs borne by the collective (Blomqvist, 1991).

Compensation methods that imply enhanced incentives on providers to take cost into account include capitation in primary care (the backbone of the U.K. NHS), and episode-based payments (such as those based on Diagnosis-Related Groups (DRG)) for hospitals. Both are examples of *prospective* funding models, meaning that the payment is fixed in advance of treatment and does not depend on what volume of services the patient actually received. Under capitation (which is often combined with partial fee for service in 'blended' systems), primary care providers agree to supply services as needed for all patients on their roster, in return for a fixed monthly amount per patient. In the DRG model (which originally was developed for the U.S. Medicare plan, but now is used in many countries), the principle is that a hospital's revenue from treating given patients will depend only on the classification of their health problems, not on their length of stay or what drugs and services were actually used in treating them.

Prospective funding methods have been used not only in government plans, but also by private insurers in the U.S., in conjunction with expanded use of 'selective contracting and patient steering.' In plans that use this approach, the insured are only fully covered if they get their care from providers who have signed a contract in which they agree to be compensated for their services according to terms set by the plan, and often also to abide by restrictions in the way they will treat the plan's patients. For example, in some managed-care plans, the participating doctors have to agree to choose prescription drugs from a formulary provided by the plan, or to a second opinion before a patient can receive major elective surgery. Under selective contracting, fees and other terms according to which services are supplied are no longer determined through transactions between providers and individual patients at the point of service; instead, insurance plans negotiate the terms in advance on behalf of their insured clients. The trend toward more use of selective contracting has strengthened the market power on the buyers' side in the markets for health services and drugs, and made them more competitive in comparison with the earlier model in which the buyers were the individual patients.

Private Health Insurance: Competition, Cream Skimming, and Adverse Selection

Even though government has taken over most of the responsibility for pooling the financial risk associated with health care costs in almost all advanced countries other than

the U.S. and a few others, theoretical and empirical research on private health insurance markets continues to be a prominent part of the health economics literature. This research is motivated in part by the importance of private health insurance in the U.S. system, but also by the fact that private insurance continues to play a significant role in several other countries, as provider of supplementary coverage for items excluded from government plans (such as outpatient pharmaceuticals and dental care), or complementary coverage when the public plan contains deductibles or patient cost-sharing (as in France, or in the basic U.S. Medicare plan for retirees).

An active line of research, on risk adjustment, relates to attempts in some countries (especially the Netherlands) to allow private insurers a more prominent role in government-organized universal social insurance systems in a way that is both efficient and consistent with the equity objective. In the absence of regulation, competition among private insurance plans will cause them to charge higher premiums, or deny coverage, to individuals who can be predicted to have high expected health care costs, based on factors such as their own or their families' illness history. In order to create more equitable access to insurance, regulators can simply prohibit premium discrimination and require plans to cover everyone who is willing to pay the quoted premium. However, the theoretical and empirical literature on risk selection shows that there are ways in which insurers can design their plans so as to make them more attractive to low-risk individuals ('cream skimming'), undermining the equity objective to some extent. In response, regulators in the Netherlands and elsewhere have tried to reduce insurers' incentive to engage in risk selection by paying a risk-adjusted premium subsidy for individuals with high expected expenditure, and by restricting insurers' ability to cream-skim by limiting the extent to which they can vary the content of their coverage. Every plan must also accept any applicant willing to pay its net-of-subsidy premium (which must be the same for everyone).

Much of the early literature on private health insurance focused on the case where the information about factors leading to differences in expected costs was private (known to the buyers) and not readily observable by the insurers. As Rothschild and Stiglitz showed in their classic article on adverse selection (1976), competitive insurance markets could then break down or lead to inefficient outcomes. A somewhat surprising prediction of their analysis was that plans with generous benefits would tend to survive in a competitive market, though they would be very expensive and used only by individuals whose risk of loss was high. In the context of health insurance, the empirical evidence suggests that another consequence may be 'death spirals' under which plans with generous benefits will become so expensive that they will disappear, leading to outcomes in which surviving plans will offer only limited benefits (Cutler and Reber, 1998). The problem of adverse selection can be at least partially overcome when health insurance is in the form of a common plan that is purchased by a group of individuals who are related in some way that is independent of their expected health care costs, for example, all employees in a firm; the fact that most private health insurance in the U.S. is through employment is partly a reflection of this (Diamond, 1992).

The Industrial Organization of the Health Insurance and Health Services Industries

Selective contracting and patient steering is a means by which insurance plans can take a more active role on the buying side in the markets for health services, creating a better balance in market power between buyers and sellers. But closer integration between insurers and producers of health services can also be interpreted as a model for supplying these services that is more like the way they would be supplied in an idealized model of state-contingent contracts.

Considerable attention has been paid in the literature to Health Maintenance Organizations (HMOs), a type of firm that was considered a promising model in the U.S. health care industry in the 1970s and 1980s (Luft, 1981). In their simplest form, HMOs were insurance plans whose enrolees paid a fixed monthly premium, in return for being entitled to health services and drugs as needed, but only in hospitals owned by the HMO, and from doctors who were its salaried employees. Because the doctors were paid by salary, and the hospitals' revenues did not directly depend on the services they supplied, they had no direct incentive to supply large volumes of services. Moreover, because HMOs were responsible for the full range of health services (primary care, services of specialists and hospitals) and drugs, they could ask their providers to follow rules that would combine these inputs efficiently. Empirical evidence in the Rand study suggested that HMOs could supply care to given populations at lower cost than under the conventional model, with similar health outcomes on average (Newhouse and The Insurance Experiment Group, 1993). The fact that the U.K. health care system seems to be able to provide good care to its population at a low cost relative to others is consistent with this evidence, since the U.K. NHS can be loosely interpreted as a nationwide HMO.

HMOs and similar managed-care plans increased their share of the U.S. health insurance market to a considerable extent in the 1980s and early 1990s, and health economists like Enthoven envisaged reforms that would create an efficient system of universal insurance coverage through one of many competing plans (Enthoven, 1993). However, the 1993 reform proposals of the first Clinton administration which drew on these ideas met with vigorous opposition, and in the second half of the 1990s, there was what has been described as a 'backlash' against managed care (Havighurst, 2002). While the principle of selective contracting has survived, plans with fewer restrictions on providers and less integration between the different kinds of health services now dominate the market. In the Obamacare legislation, there are incentives encouraging different providers who treat patients under the Medicare plan to form Accountable Care Organizations for the purpose of saving resources by integrating care more closely. These provisions can be interpreted as a sign that integration of care remains an unresolved issue in much of the U.S. system (McClellan et al., 2010).

Enthoven's ideas of 'managed competition' have also been an important influence on the process of health financing reform that has been underway in the Netherlands over several decades (van de Ven and Schut, 2008). Under the legislation governing the system, the competing insurance plans from which citizens can choose are allowed to engage in selective

contracting and patient steering, although so far they have only done so to a limited extent.

Health as Capital; Communicable Disease

Part of the work in health economics has been concerned with quantifying the concept of health. In the model analyzed by Grossman (1972), health is interpreted as a form of capital. Each person is born with a certain quantity, and can build up his or her health capital by using health services produced by others, by eating well, exercising, and so on. But health capital is also subject to depreciation, as people fall ill and grow old. Health capital yields benefits in the form of a flow of healthy time which is both valuable in itself (the consumption motive for investing in health), and because it can be used to earn an income through working (the investment motive).

Empirical work on the production of health capital obviously is difficult, because health can only be imperfectly observed and measured. Nevertheless, research inspired by Grossman's analysis has yielded important insights and evidence. In particular, it has shown that childhood health, and the health services used to improve it, including prenatal care, are important predictors of adult health and well-being (Currie, 2009). It has also focused on the long-term benefits of factors such as exercise and weight control, as well as on the relation between addiction to tobacco and alcohol and health. The theory of 'rational addiction' proposed by Becker and Murphy (1988) is similar to Grossman's analysis in that both seek to reconcile the evidence on the long-term consequences of behaviors that influence health with the hypothesis of rationality in human decision making. However, the evidence from behavioral economics is casting more doubt on the helpfulness of this hypothesis in contexts that involve long time horizons and small risks of major adverse outcomes (Rabin and Thaler, 2001). Both these features are characteristic of many decisions that affect health in the long run.

In the analysis of population health, there is a critical distinction between communicable and noncommunicable diseases, with the former being much more prevalent in poor countries while the latter now dominate in rich countries. In the early health economics literature, the focus was implicitly on noncommunicable diseases and on decisions by individuals that only affect their own health. For communicable diseases, there are significant spillover effects as actions (or lack of actions) by individuals influence the health risks of others. In epidemiological models of communicable diseases, these actions have typically been taken as given. Since the 1990s, however, there is a branch of the literature that draws on both epidemiology and economics and looks at how economic incentives can influence individual decisions that affect the spread of contagious diseases (Kremer, 1996). This research has partly been motivated by the spread of HIV/AIDS, but also by the increasing attention being paid to health care and health policy in developing countries. Analysis that combines the modeling of individual incentives and spillover effects will also be relevant as more resources are devoted to the growing problem of infections that become resistant to existing drugs.

Valuing Health

Government decision makers are often called upon to evaluate projects or regulations that affect safety, for example, in the transportation sector, leading to the question whether there is a meaningful way to estimate the value of saving human lives. Following the work of Schelling (1968) and others, it is widely accepted that it is reasonable to use estimates of the value of saving a 'statistical life' (that is, the life of a person whose identity is not known in advance) on the basis of evidence regarding how much people are willing to pay for reducing the risk that they will die. Such evidence can be derived from contingent valuation (answers to hypothetical questions), or from revealed preference, for example, based on observed wage differentials between risky and safe jobs (Viscusi, 1993). Estimated values of reducing mortality are obviously relevant to health economics, since much of what is spent on health care goes toward drugs and interventions that increase the chances of survival among patients with different kinds of illness. Murphy and Topel (2006) argue that conventional GDP statistics have seriously underestimated secular growth in human welfare because they fail to account for the value of increased life expectancy, which has occurred at least partly in response to the development of new drugs and medical technology.

Death, however, is only the most extreme form of ill health; the main purpose of many types of health care is to improve patients' quality of life, not just to reduce the risk of death. An active branch of health economics research has concerned itself with the development of a generalized health index, quality-adjusted life expectancy, that accounts both for improvements in the quality of patients' lives and longer life expectancy. In calculating the expected number of quality-adjusted life years (QALYs) among members in a given population, a year lived with a disabling or painful health problem is counted as the equivalent of a fraction of a year lived in normal health. Much of the research in this area has been about simple methods to estimate what quality-of-life weights should be assigned to years lived with different kinds of health problems, based on patients' answers to hypothetical questions.

Cost-effectiveness evaluations of different treatment approaches or drugs that compare their costs with their ability to produce additional QALYs in given patient populations are increasingly common in the literature; the term 'cost-utility analysis' is often used to refer to this methodology. In countries where most health care is directly paid for by government, such as the U.K., certain procedures or drugs may not be approved for coverage under the public plan if their cost-effectiveness is unfavorable (the cost per incremental QALY is too high), but even in the U.S. where coverage restrictions are opposed by many who object to 'rationing' of patients' access to care, economic evaluations can play a role in establishing clinical guidelines or drug formularies that are used by many insurance plans.

Although cost-utility analysis is widely used, certain conceptual issues remain controversial, for example, whether future QALYs should be discounted, and when it is appropriate for an evaluation to include only costs to the government as opposed to costs borne by all members of society, including private individuals. Or, in calculating the costs to society of

using a new technology that reduces mortality, should one include an allowance for the fact that additional survivors will mean higher health care costs in the future, as well as more resources required for their consumption of other goods and services?

The importance of providing information and incentives to make cost-effective choices in medical care has been indirectly illustrated by studies showing large variation in practice patterns across different locations; this line of research has been particularly active in the U.S. Data on the per capita cost of medical care in similar populations (for example, U.S. retirees covered by the Medicare plan) show differences of a hundred percent or more in comparisons across major urban areas, and while they can be attributed in part to different unit prices for various services, they remain large even after correcting for this. Studies of particular procedures also show wide variation in the rates at which they are applied in comparable populations. These findings lend credence to the widely held view that large gains could be realized (either in the form of lower cost or better health outcomes on average) if practice patterns everywhere could be made to align more closely with the findings of cost-effectiveness studies.

Pharmaceuticals and Medical Technology

While other factors, such as the high labor intensity of health care production, have contributed to the increase in health care spending as a share of GDP in almost all countries, the main driver appears to have been the continuing development of new and often costly types of medical and pharmaceutical technology. The costs of developing and using the new technologies have been high, but their benefits, in the form of longer life expectancy and better quality of life have been large and, on average, have more than justified the costs (Murphy and Topel, 2006). New technology will continue to come on stream, partly driven by advances in molecular biology and research on the human genome, but several issues regarding the cost of developing and using new technology remain controversial.

Financing of global R&D is to a large extent through the patent system under which an incentive to develop new technology is provided through legal rules that give the patent owner a time-limited monopoly in the market for any products that use the patented technology. The most important health-related industry in which R&D is financed in this way is the pharmaceutical industry. Although most countries are bound by international agreements that place certain restrictions on their patent laws, national governments still retain considerable discretion with respect to the detailed rules regarding matters such as the definition of what constitutes a patentable invention, the length of the patent period, and what rules apply to competing producers (such as manufacturers of generic drugs) when the patent expires for a brand-name drug.

The application of the patent system in the pharmaceutical industry involves the same trade-off as in other sectors, between the need to create an incentive for future R&D spending, on the one hand, and the desire to quickly make new technology available to consumers at a reasonable price, on the other. However, in the market for drugs, the trade-off is complicated by the fact that they typically are paid for by third parties (public or private insurance plans), not by the consumers themselves. Prices of patented drugs in many countries therefore tend to be regulated by government, or established in bilateral negotiations between central agencies and the patent owner, with evidence from health economic evaluations playing a role. Rules with respect to drug patents have been important elements in international trade negotiations. In these negotiations, countries or jurisdictions where the large and research-intensive pharmaceutical firms are located and generate jobs and tax revenue (the U.S., the E.U., Japan) tend to push for policies that lead to high drug prices and revenue for these firms. Smaller and poorer countries that mostly use drugs that have been developed elsewhere prefer rules that shorten the period of patent protection, lower the prices of patented drugs, and allow quick entry of lower-cost generic drugs after the patents have expired.

The nature of R&D in the pharmaceutical and medical devices industries in the future may also be affected by the changing roles of insurers and of physicians as agents of both patients and third parties in many health care systems. In important early contributions, Goddeeris (1984a,b) introduced the idea that firms engaged in R&D face a choice between efforts to develop cost-saving technology (that would reduce the cost of existing treatment methods), or technology that leads to increased spending because it enables providers to effectively treat new patients with problems for which current methods do not work. His insight was that when doctors act exclusively in their role as agents for their patients, they would look relatively more favorably on the latter type of innovations, since most of the additional costs will be paid by third parties, not by the patients. As more doctors work under contracts that also give them an incentive to pay attention to the costs of the third party, the implicit bias against cost-reducing innovations will be reduced, at least to some extent.

Health Economics, Health Policy, and Equity

Mandatory social insurance programs or direct government funding of health services are crucial elements of the welfare state in many countries that strive to attain a more equal distribution of income. The correlation between health and income or social status has been known for a long time, and continues to be explored in the literature. The question to what extent low income causes ill health or the other way around (or whether both are correlated with some third factor) remains open, but the correlation persists even in systems where universal government programs has largely eliminated income as a determinant of access to health care.

Most of the literature on economic inequality concerns itself with the distribution of measures of real income, whether in a given year or over individuals' lifetime; less attention is paid to the question how inequality is reflected in different rates of consumption of specific goods and services. In health economics, however, there is a substantial literature on inequality in the utilization of health care, and whether it is correlated with income, partly reflecting the view that society should be less tolerant of inequality in health care than for other kinds of consumption (Tobin, 1970). In some countries,

for example Canada, there is widespread support for the idea that utilization of health services should be independent of income, and various policies are used to discourage purchases of health services outside of the government plans even by people with high income who are willing to pay for them. An important question for the future, especially in countries with universal coverage through government plans, will be to what extent their health financing systems will allow choice for individuals among insurance plans with different treatment strategies and hence expected costs (Havighurst, 2002). This question will become more urgent as medical and pharmaceutical technology continues to progress and introduce a wide range of new drugs and methods that can be used to improve expected health outcomes, but often at very high cost. The aging of the population in many countries is relevant to this issue, since a large proportion of health care costs typically are incurred by older patients who are nearing the end of their lives. Rising government expenditure to pay for the health care costs of the elderly therefore contribute in a major way to the increasing burden that the relatively small younger generations of working-age taxpayers face in countries where the proportion of retirees will rise over the next several decades.

Health care and health insurance are not internationally traded to any significant extent, so there is little competition among providers and insurance systems across countries. While much of the research in health economics remains country-specific, the literature comparing health care systems and policies in different countries has become increasingly important over time, especially in the twenty-first century. A country whose health policy is attracting great interest is China, where it is a critical element in the regime's efforts to reduce inequality, and health care reform is accorded a high priority because the current market-oriented model compares unfavorably, from the viewpoint of equity, with the one that existed before the 1980s when the experiment with less centralized economic management began (Qian and Blomqvist, 2014).

See also: Aging and Health in Old Age; Behavioral Economics; Comparative Health Systems; Cost–Benefit Analysis; Cost-Benefit and Cost-Effectiveness Analyses in Evaluation Research; Economic Evaluations and Electoral Consequences; End of Life Choices; Ethical Dilemmas, Research, and Access to Treatment; Family Theory: Economics of Intergenerational Relations; Health Care Markets: Theory and Empirical Evidence; Health Care, Rationing of; Health Care: Legal Issues (United States); Health Insurance: Economic and Risk Aspects; Health Politics and Policy; Health in Low-Income Countries; Insurance; Market and Nonmarket Allocation; Mortality, Epidemiological, and Health Transitions; Pharmaceutical Industry: Political Economies of Drugs and Knowledge; Population Aging: Economic and Social Consequences; Retirement, Economics of; Social Insurance: Legal Aspects; Socioeconomic Status and Health Care; Value of Life, Economics of; Welfare State, History of.

Bibliography

Arrow, K., 1963. Uncertainty and the welfare economics of medical care. American Economic Review 53, 941–973.

Becker, Gary S., Murphy, Kevin M., 1988. A theory of rational addiction. Journal of Political Economy 96, 675–700.

Blomqvist, A.G., 1991. The doctor as double agent: information asymmetry, health insurance, and medical care. Journal of Health Economics 10, 411–432.

Culyer, Anthony J., Newhouse, Joseph P. (Eds.), 2000. Handbook of Health Economics, vol. 1A and 1B. North-Holland, Amsterdam.

Currie, Janet, 2009. Healthy, wealthy, and wise: socioeconomic status, poor health in childhood, and human capital development. Journal of Economic Literature 47, 87–122.

Cutler, David M., Reber, S.J., 1998. Paying for health insurance: the tradeoff between competition and adverse selection. Quarterly Journal of Economics 113, 433–466.

Danzon, Patricia M., Nicholson, Sean (Eds.), 2011. The Oxford Handbook of the Economics of the Biopharmaceutical Industry. Oxford University Press, New York.

Diamond, Peter A., 1992. Organizing the health insurance market. Econometrica 60, 1233–1254.

Enthoven, Alain C., 1993. The history and principles of managed competition. Health Affairs 12 (Suppl.), 24–48.

Glied, Sherry, Smith, Peter C. (Eds.), 2011. The Oxford Handbook of Health Economics. Oxford University Press, Oxford.

Goddeeris, J., 1984a. Insurance and incentives for innovation in medical care. Southern Economic Journal 51, 530–539.

Goddeeris, J., 1984b. Medical insurance, technological change, and welfare. Economic Inquiry 22, 56–67.

Grossman, Michael, 1972. On the concept of health capital and the demand for health. Journal of Political Economy 80, 223–255.

Havighurst, Clark C., 2002. How the health care revolution fell short. Law and Contemporary Problems 65 (4), 55–101.

Kremer, Michael, 1996. Integrating behavioral choice into epidemiological models of AIDS. Quarterly Journal of Economics 111, 549–573.

Luft, Harold S., 1981. Health Maintenance Organizations: Dimensions of Performance. Wiley, New York.

McClellan, Mark, McKethan, Aaron N., Lewis, Julie L., Roski, Joachim, Fisher, Elliott S., 2010. A national strategy to put accountable care into practice. Health Affairs 29, 982–990.

Murphy, Kevin M., Topel, Robert H., 2006. The value of health and longevity. Journal of Political Economy 114, 871–904.

Newhouse, Joseph P., The Insurance Experiment Group, 1993. Free for All? Lessons from the Rand Health Insurance Experiment. Harvard University Press, Cambridge, MA.

Pauly, Mark V., 1968. The economics of moral hazard. American Economic Review 58, 531–537.

Pauly, Mark V., Mcguire, Thomas G., Barros, Pedro P. (Eds.), 2012. Handbook of Health Economics, vol. 2. North Holland.

Qian, Jiwei, Blomqvist, Åke, 2014. Health Policy Reform in China: A Comparative Perspective. World Scientific.

Rabin, Matthew, Thaler, Richard H., 2001. Anomalies: risk aversion. Journal of Economic Perspectives 15, 219–232.

Rothschild, Michael, Stiglitz, Joseph, 1976. Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. Quarterly Journal of Economics 90, 281–305.

Schelling, Thomas, 1968. The life you save may be your own. In: Chase, Samuel B. (Ed.), Problems in Public Expenditure Analysis. The Brookings Institution, pp. 127–162.

Thaler, Richard H., Sunstein, Cass R., 2009. Nudge: Improving Decisions about Health, Wealth, and Happiness. Penguin Books, New York.

Tobin, James, 1970. On limiting the domain of economic inequality. Journal of Law and Economics 13, 264–278.

van de Ven, Wynand P.M.M., Schut, Frederik T., 2008. Universal mandatory health insurance in the Netherlands: a model for the United States? Health Affairs 27, 771–781.

Viscusi, W. Kip, 1993. The value of risks to life and health. Journal of Economic Literature 31, 1912–1946.

Zeckhauser, Richard, 1970. Medical insurance: a case study of the tradeoff between risk spreading and appropriate incentives. Journal of Economic Theory 2, 10–26.

Health Education and Health Promotion

Gerjo Kok and Nanne K de Vries, Maastricht University, Maastricht, Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Kok, volume 10, pp. 6557–6563, © 2001, Elsevier Ltd.

Abstract

Health promotion is any planned combination of educational, political, regulatory, and organizational supports for actions and conditions of living conducive to the health of individuals, groups, or communities. Health education is one type of health promotion intervention. Health education is a planned activity, stimulating learning through communication, to promote health behavior. In this article we give definitions and examples of health promotion, health education, and prevention. The PRECEDE-PROCEED model is explained and illustrated and the importance of systematically planning interventions is supported with evidence. Theories are the basis for evidence-based planning. We present Intervention Mapping as a protocol for developing theory- and evidence-based health promotion interventions. Appropriate theories for determinants analysis, intervention development, and anticipation of implementation are listed with practical applications.

Health Promotion, Health Education, and Prevention: Definitions and Perspectives

Definitions

Health promotion is any planned combination of educational, political, regulatory, and organizational supports for actions and conditions of living conducive to the health of individuals, groups, or communities (Green and Kreuter, 2005: pG4). Three types of prevention are the goals of health promotion: (1) primary prevention; (2) early detection and treatment; and (3) patient care and support. Alongside prevention, which is disease oriented, health promotion also aims for enhancing participation, improving living circumstances, and optimizing health; it is the process of enabling people to increase control over and improve their health focusing on the determinants of health (WHO, 1986). Different from marketing and advertising, the goals of health promotion are derived from consensus declarations about health and quality of life, by organizations such as the World Health Organization (WHO) or the United Nations.

Health education is one type of health promotion intervention. Health education is a planned activity, stimulating learning through communication, to promote health behavior. Other health promotion instruments are facilities or resources and regulation. Health education is based on voluntary change, while regulation is based on forced compliance and will only be effective in combination with control and sanctions.

In general, interventions which use a variety of health promotion instruments will be more effective. An example of this last statement is the prevention of drinking and driving. There is regulation: most countries have laws against driving under the influence of alcohol. Often there is control: drivers are stopped by the police and may be tested, although countries differ in their commitment to these control activities. There are facilities: public transport is available and, especially in the weekends and during the night, cheap taxis for adolescents. There is education: information is provided about the rules and possible sanctions, about the facilities, and about drunk driving itself, i.e., the consequences of drunk driving, but also on ways to prevent getting into that situation. The combination of control, facilities, and education is more effective than any one of these elements alone.

Health promotion as an intervention for patient care and support is usually called patient education instead of health promotion. Patient education is defined as a process of assisting people to learn health-related behaviors, so that they can incorporate those behaviors into daily life (Bastable, 2006: p. 11). Sometimes patient education supports patients who recover, for instance, after surgery. Sometimes the focus is on chronic patients, such as patients with asthma or diabetes; health promotion seeks to support disease management and self-management. Patient education also includes slowing down deterioration as well as support for people who are dying, for instance, in the case of terminal cancer. Patient education is not only directed at the patient but also at the patient's family, at health professionals, and at the environment. Health workers such as physicians and nurses are the primary providers of patient education and they need to be motivated and skilled to do so. Frequently, many different health workers are involved in the treatment of the same patient. Patient education by each of these professionals is not always attuned to the needs of the patient. This calls for the development of patient education programs that coordinate the educational activities of the different health professionals involved, for example, by developing protocols for continuity of care.

Perspectives

Current health promotion research and practice is characterized by three perspectives: an ecological approach, participation of all stakeholders, and the use of theories and evidence.

People's behavior is not only determined by personal factors but also by environmental conditions: social networks, organizations, communities, and society. We call that the ecological approach (Kok et al., 2008). The key to understanding environmental conditions is the recognition of agents in the environment, also called decision-makers or gatekeepers. At every environmental level there are agents who decide about an environmental condition. To change those conditions, we

have to understand the behavior of the agents and plan interventions to influence the agents' behavior.

Participation is another essential element in health promotion; participation of the target population, relevant environmental agents, and potential implementers of the intervention (Belansky et al., 2011). Principles of participation include: build upon community strengths, facilitate equitable decision-making in an iterative process, foster colearning among partners, local problems, and multiple determinants of health.

Careful intervention planning makes health promotion an evidence-based practice. Evidence can be found in empirical research but also in theory. Most health promotion interventions are based on theories and research in an iterative process of brainstorming, reviewing the literature, applying theories, and executing additional research (Bartholomew et al., 2011; Buunk and van Vugt, 2013). This is also a continuous process; evidence-based health promotion means that effective interventions may be adapted for implementation in other communities, again based on theories and research.

The Planning of Health Promotion

Green's PRECEDE-PROCEED Model

Health promotion is a planned activity. The best known and most often used planning model in health promotion is Green's PRECEDE-PROCEED model (Green and Kreuter, 2005). In short, Green starts with the social assessment: what is the quality of life of a certain group, community, or country, and the epidemiological assessment; is health relevant for the quality of life and if it is, what are the most serious health problems? With the epidemiological assessment, the causes of health in terms of behavior/lifestyle, genetics, and environment can be analyzed. With the educational and ecological assessment, an analysis can be made of the determinants of the relevant behaviors, in terms of predisposing, reinforcing, and enabling factors, and the variables can be selected that one wishes to influence. With intervention alignment and administrative and policy assessment, the possible usefulness of health education and other potential interventions, of facilities and regulations, can be analyzed. Health promotion interventions are developed and then implemented. Finally, evaluation is made of the process, impact, and outcome of these interventions, resulting in feedback and improvement of the interventions.

An Example of PRECEDE-PROCEED

For example, in most Western countries, the quality of life in general is high. People value health as one of the most important aspects of their quality of life, especially when it is threatened. What are the most important health problems? In terms of mortality, cardiovascular disease (CVD) and cancer fulfill this role. What are the causes of CVD and cancer? In terms of environment, these are, for instance, industrialization and unhealthy working environments. In terms of behavior and lifestyle, smoking and distorted energy balance (diet and physical activity) are major causes. Let us select smoking and, in particular, the prevention of the onset of smoking, as a relevant behavior that we might want to influence. Why do adolescents start smoking? Not because they like it and not because they do not know about the dangers. One reason is social pressure, mostly from peers, family, or mass media. In terms of predisposing factors, adolescents often do not know how to resist social pressure.

In terms of reinforcing factors, belonging to a peer group is very important for adolescents. In terms of enabling factors, cigarettes are easy to obtain and sanctions against smoking are weak. What can we do? In terms of health education, we might want to improve adolescents' self-efficacy in resisting social pressure. There are a number of techniques that help adolescents to learn how to resist pressure to smoke, for instance, through positive role modeling (see Bartholomew et al., 2011). In terms of facilities and regulation we want to develop anti-smoking policies, for instance, in schools, or through community action against vendors who sell cigarettes to adolescents under legal age. The implementation of such a health promotion program is organized in cooperation with schools, parents, and students. Finally, the effectiveness of the program is measured. In terms of process evaluation, the central question is: was the program implemented as planned and was the supposed self-efficacy improvement actually realized? In terms of impact evaluation: did fewer students in our program group start smoking than in a comparison group? It may be clear that an improvement in terms of outcomes, reduction of CVD and cancer, and quality of life, cannot be expected for several years. For some behaviors there are predictive models to estimate these improvements, often enabling cost-effectiveness analysis.

Quality of the Planning

The process of planning health promotion programs is a cumulative and iterative process. On the one hand, we need answers on earlier planning questions to decide about later phases. On the other hand, the process is not rigid but flexible and we go back and forth through the model. During the planning process we make use of existing knowledge that is systematized in theories, and which is available as empirical data. Why does it have to be so complex? Certainly not because health promoters admire complicated models, but because careful planning is essential if we strive for effect.

Theory-Based Interventions

A very important early study in this respect is a meta-analysis by Mullen et al. (1985) on 70 patient education studies that were methodologically sound and which included knowledge and compliance or adherence as effect variables. Several diseases and several types of interventions were represented. Mullen et al. estimated the quality of the interventions using six criteria that they derived from literature and that can also be seen as guidelines for intervention development:

1. Consonance: the degree of fit between the program and the program objectives.
2. Relevance: the tailoring of the program to knowledge, beliefs, circumstances, and prior experiences of the learner, as assessed by pilot testing or other means.

3. Individualization: the provision of opportunities for learners to have personal questions answered or instructions paced according to their individual progress.
4. Feedback: information given to the learner regarding the extent to which learning is being accomplished (e.g., blood pressure reading).
5. Reinforcement: any component of the intervention that is designed to reward the behavior (other than feedback) after the behavior has been enacted (e.g., social support).
6. Facilitation: the provision of means for the learner to take action and/or means to reduce barriers to action (e.g., subsidies).

Mullen et al. show that the best predictor of success was the rating score for the quality of the educational intervention. The choice for a specific educational technique was not related to effectiveness, again demonstrating that there is no such thing as a 'magic bullet.' Thus, the effectiveness of a health promotion intervention is determined by the quality of the planning process. In the six criteria that Mullen et al. used, we recognize different concepts from Green's planning model: the choice of relevant behavioral objectives, tailoring to determinants of behavior, learning (health education), reinforcement, support, and facilitation (health promotion). There are also some additional concepts that have not been described in the PRECEDE-PROCEED model, but which are selected from behavioral science theories, especially Social Cognitive Theory (McAlister et al., 2008).

Reviews and Meta-Analyses

Recent reviews and meta-analyses also demonstrate a positive relationship between the use of theory-based intervention methods and the effectiveness of the intervention (Albarracín et al., 2005; de Bruin et al., 2009; Fisher et al., 2002; Peters et al., 2009; van Achterberg et al., 2010). In these reviews, theory-based methods were coded from intervention descriptions, after which the authors show an association between the number of theory-based methods used and the effectiveness of interventions. Even while it was impossible to evaluate the appropriateness of the way the methods were used (which is crucial in the correct application of methods, see Peters et al., ahead-of-print), still the conclusion is that the quality of the planning process is a key for intervention success.

Intervention Mapping

Intervention Mapping is a protocol for developing theory- and evidence-based health promotion programs (Bartholomew et al., 2011). It describes the process of health promotion program planning in six steps: (1) the needs assessment based on the PRECEDE-PROCEED model; (2) the definition of performance and change objectives based upon scientific analyses of health problems and problem causing factors; (3) the selection of theory-based intervention methods and practical applications to change (determinants of) health-related behavior; (4) the production of program components, design and production; (5) the anticipation of program adoption, implementation and sustainability; and (6) the anticipation of process and effective evaluation. Intervention Mapping recognizes the earlier mentioned perspectives: an ecological approach, participation of all stakeholders, and the use of theories and evidence. Although Intervention Mapping is presented as a series of steps, Bartholomew et al. (2011) see the planning process as iterative rather than linear. Program planners move back and forth between tasks and steps. The process is also cumulative: each step is based on previous steps, and inattention to a particular step may lead to mistakes and inadequate decisions.

An Example: Long Live Love

An example of the use of Intervention Mapping in program development is the Long Live Love HIV/STI-prevention program for Dutch adolescents (Schaalma and Kok, 2011):

1. Needs assessment: In the needs assessment, the desired behavioral and environmental outcomes are identified. HIV is a serious health problem and young people today face increased risk of infection. In addition health promotion programs are most effective when the population has not yet formed risky behavior pattern and habits. The most relevant outcome related to HIV/STI prevention for adolescence is a behavioral outcome: condom use.
2. Performance and change objectives: Performance objectives are the specific behaviors that we want the target group (or the environmental agents) to 'do,' as a result of the program. For example, in the case of HIV prevention; buy condoms; have them with you; negotiate with your partner; use them correctly; and keep using them. The next task is to identify the determinants of the performance objectives: factors that have been shown to be associated with the behavior. Change objectives are the combination of target group, performance objectives, and determinants: they describe what the target individual or agent has to learn (determinant change) in order to do the performance objective. One change objective for an HIV-prevention program in schools would be: "adolescents (target population) express their confidence (determinant) in successfully negotiating with the partner about condom use (performance objective)."
3. A theoretical method is a general technique or process derived from theory and/or research, for influencing changes in the determinants of behavior; applications are specific techniques for practical use of a method in ways that fit the target group and the setting. For instance, one method for self-efficacy improvement (for negotiating) could be modeling, and the application for adolescents in school could be peer modeling by video. Every health promotion intervention exists of many applications of many methods. Bartholomew et al. (2011) give a systematic description of methods for each determinant and each level. In the literature there is an increasing interest in systematic descriptions of methods, or behavior change techniques at the individual level (Abraham and Michie, 2008) as well as the environmental levels (Khan et al., 2009). However, there is more to methods than just definitions if the planner wants to select appropriate methods for an intervention (Peters et al., ahead-of-print). An important task in this step is to identify the conditions or parameters that limit the

effectiveness of theoretical models, such as identification, observability, and reinforcement as necessary conditions for the effectiveness of learning by modeling. Bartholomew et al. (2011) also provide the parameters for use for a large number of methods.

4. The actual designing of the program involves organizing the applications into a deliverable program and producing and pilot testing the materials. In the example of the 'Long Live Love' HIV-prevention in schools, the program comprised five lessons, an interactive video, a brochure for students, and a workbook for teachers.
5. The first activity to anticipate implementation, actually at the start of intervention development, is the development of a linkage system, linking program developers with program users. Then, a specific intervention is developed that aims to promote adoption and implementation of the program by the intended program users, analyzing performance and change objectives, and specifying methods and applications.
6. Finally, Intervention Mapping step 6 anticipates process and effects evaluation. Again, this is relevant from the start. For instance, "adolescents express their confidence in successfully negotiating with the partner about condom use" is an objective, but is also a measure of that objective, that can be asked in pre- and postinterviews with experimental and control group subjects.

Using Theories in Health Promotion: Determinants

At different phases in the planning process different theories will be most relevant (Glanz et al., 2008; Simons-Morton et al., 2012). The first phases rely on epidemiological theories; the later phases, when the intervention is developed, rely on theories from the behavioral sciences, especially social psychology. In this and the next two paragraphs we will describe theories that can be applied in the three phases: determinants, intervention, and implementation. In the phase of looking at determinants we try to understand why people behave as they do. We distinguish ecological, social cognitive, and automatic determinants of behavior.

Ecological Determinants

People's behavior is not only determined by personal factors but also by environmental conditions: social networks, organizations, communities, and society (Kok et al., 2008). A changed, facilitating environment without barriers makes the health promoting behavior the easiest behavior. At each environmental level there are agents who decide about an environmental condition, for example: parents, managing boards, newspaper editors, legislators. Environment-oriented theories, explaining ecological determinants of behavior and suggesting methods to change those, have become more popular in health promotion, for example: systems theories, empowerment theories, Stakeholder Theory, and Agenda-Building Theory. At each ecological level there are agents who can change the system to facilitate health. Most environment-oriented theories focus on changing (the behavior of) agents; see the next section.

Social Cognitive Determinants

Current social psychological models indicate three types of social cognitive determinants of behavior (Fishbein and Ajzen, 2010: Reasoned Action Approach; Montano and Kasprzyk, 2008: Integrated Behavioral Model; McAlister et al., 2008: Social Cognitive Theory):

1. Attitude: beliefs about advantages and disadvantages of behavior resulting in an attitude about the behavior, also described as outcome expectations.
2. Social influence: beliefs about social norms (subjective norms, injunctive norms) and about behavior of others (descriptive norms, vicarious learning, modeling).
3. Self-efficacy: beliefs about perceived control, self-efficacy expectations.

In health promotion, theories that are more specific for risk behavior are also applied. They indicate four basic types of risk-related social cognitive determinants (Champion and Skinner, 2008: Health Belief Model; Ruiter et al., 2014: Protection Motivation Theory and Extended Parallel Process Model):

1. Seriousness: how serious is the consequence of the risky behavior?
2. Susceptibility or vulnerability: is that going to happen to me?

Together these two form the threat appraisal.

3. Response-efficacy: what can I do to avoid those consequences?
4. Self-efficacy: am I confident that I can do that?

Together these last two form the coping appraisal. If people see a threat, they try to control it. If the recommended action is thought to be effective and feasible, people will try to perform the recommended action: danger control. However, if the recommended action is thought to be ineffective or impossible, then the threat will result in defensive reactions, denial or avoidance coping: fear control.

Models about determinants of behavior do not imply a one-directional influence; attitudes, social influence, self-efficacy, and risk appraisals can be antecedents as well as consequences of behavior. In interventions we try to change determinants in order to change behavior, but we also use techniques that influence behavior rather directly, such as reinforcement and having people publicly commit themselves to the desired behavior. Positive experiences with behavior, in turn, may change psychosocial determinants of behavior, thus creating reciprocal determinism.

Automatic Determinants

Much of health promotion research and theory is based on the assumption that people are consciously and systematically processing information in order to construe and interpret their world and engage in action. However, recently, there has been greater focus on automatic, impulsive, or habitual behavior (Hassin et al., 2005). Habits are learned sequences of acts in response to environmental cues that become automatic: for example, taking the elevator instead of the stairs (Danner et al., 2008). However, Fishbein and Ajzen (2010: p. 53) argue that

habits are still predicted by intentions. Goal-directed behavior may start consciously but become automatic over time; for example, driving a vehicle or drinking water with meals.

Reflective behavior is the result of deliberation and conscious control; impulsive behavior is directed by learned associations with pleasure and pain; for example, consuming snack foods even when the intention is to eat more healthy (Hofmann et al., 2009: dual systems theories). While social cognitive determinants can be identified through self-reports, impulsive reactions can be measured by implicit measures; for example, the Implicit Association Test (Greenwald et al., 1998). Automatic and impulsive behaviors are often guided by external stimuli, and are more likely when the individual has reduced cognitive resources or lacks motivation to allocate resources. Impulsive behavior may also be a reason why adolescents and young adults take more risks. Research in neuroscience shows a relatively slow maturation of the cognitive control system in puberty (Steinberg, 2007).

Using Theories in Health Promotion: Interventions

Stage Models

Some current general social and health psychological models on behavior change distinguish steps, phases, or stages of change. Within those steps, a number of different specific theories can be applied. One general framework for theories on behavior change is provided by McGuire's (2001) persuasion–communication model. He describes different steps that people take, from the initial response to the message to, hopefully, a continuous change of behavior in the desired direction. Simplified, the first steps indicate successful communication, the next steps involve changes of determinants and behavior, and the last step is concerned with maintenance of that behavior change. Going through these steps, McGuire argues that the educational interventions should change with each step. The choices that have to be made about the message, the target group, the channel, and the source, will be different, or may even be conflicting, depending on the particular step that is addressed: for example, mass media may reach many people but are not adequate for self-efficacy improvement. At various steps, people may have different tendencies to process the information with either high or low elaboration likelihood, also called the central or the peripheral route (Petty et al., 2009: Elaboration Likelihood Theory). Information processing may be disturbed by lack of motivation and lack of abilities, caused by information complexity and external factors.

The Transtheoretical Model or Stages of Change Model (Prochaska et al., 2008) distinguishes stages of change within the person: precontemplation, contemplation, preparing for action, action, and maintenance or relapse. An important contribution of the stage model is the specific tailoring of health promotion efforts to groups of people in different stages. Interventions based on the model usually have different methods or applications for each stage, the so-called processes of change. There is considerable debate on the validity of the specific stages; some authors suggest that there are only two stages: motivation and action (Brug et al., 2005; see Fisher et al., 2009). A frequently used application of the theory is the motivational interviewing counseling method (Miller and Rollnick, 2012).

Learning, Goals, and Self-Management

Many theories provide possibilities for intervention methods directed at behavior change. Learning theories assume that experiences shape behavior, through classical and operant conditioning (Kazdin, 2008). Methods derived from learning theories include: conditioning, modeling, feedback, and reinforcement. Theories of information processing describe how information is perceived, stored, and retrieved (Kools, 2012a,b). Preexisting schemes may function as facilitators or barriers to information processing. Methods derived from information processing theories include chunking, advance organizers, imagery, and providing cues.

Theories about risk behavior (see 'Determinants') recognize perceived threat in relation to perceived coping (Ruiter et al., 2014). Methods derived from risk perception theories include: framing, self-affirmation, coping skills, and fear appeals (the last only under appropriate, rare conditions). Theories about social cognitive determinants, such as the Reasoned Action Approach, do not provide change methods, but identify outcome, social and control beliefs that need to be influenced in order to change the behavior (Fishbein and Ajzen, 2010). The Social Cognitive Theory (McAlister et al., 2008) suggests methods such as modeling, guided practice, and feedback.

Goal Setting Theory assumes that goal setting leads to better performance because people with goals exert themselves to a greater extent, persevere in their task, concentrate more, and more often develop applications (Latham and Locke, 2007). Goal setting is most effective when the goal is challenging but feasible. Theories of goal-directed behavior assume that people hold personal goals, which represent desired outcomes. One barrier for achieving one's goal is habitual behavior (see 'Determinants'). Change methods derived from goal theories include implementation intentions (Gollwitzer and Sheeran, 2006), cue altering, and planning coping responses.

Theories about attribution and relapse prevention suggest that people differ in making stable versus unstable attributions for successes and failures (Marlatt and Donovan, 2005). People who make stable attributions for failure are inclined to give up their change goals. Methods derived from these theories include reattribution training, relapse prevention, and planning coping responses. Self-regulation or self-management is an active, iterative process of observing oneself, making judgments based on observation, setting goals, using strategies, reacting accordingly, and revising one's strategy (Clark, 2003). Methods derived from self-regulation theories include self-monitoring, feedback, and methods to change self-efficacy and skills.

E-Health

A relatively recent development in interventions is the use of information and communication technology (ICT) in care, shortly e-health. The application of ICT started with the use of telephones and automated calls (McBride and Rimer, 1999), and by creating written messages generated by a computer

(Ruiter et al., 2006), but e-health as a method has exploded with the availability of the Internet and broadband connections. In health care, the creation of electronic patient records makes information available wherever it is needed; telemedicine applications can enter information into these records. This also enables quick support in self-management, varying from supplying immediate feedback and advice after measurement, e.g., of blood glucose, to sending text messages to remind patients to take their medicine. Also in prevention, such possibilities can facilitate the provision of quick and relevant services or messages.

Computer software has been developed that can adapt health communication to specific characteristics of respondents, for instance, tailored to specific determinants of health behaviors such as knowledge, motivation, attitudes, risk perceptions, intentions, concrete plans, perceived control, etc. Depending on answers to questionnaire items, a person might be supported in developing self-efficacy or alternatively receive information concerning the advantages and disadvantages of adopting a certain diet. The flexibility of computer-delivered messages also makes it possible to personalize messages and to adapt communication styles, e.g., providing written text or a spoken message according to the preferences and literacy of the receiver.

Effective computer tailoring makes use of determinant theories and intervention theories as described in the last paragraphs to create a diagnostic instrument, for instance, focusing on specific beliefs or matching to stages and processes of changes as described in the Transtheoretical Model. After taking this instrument, a rule base is used to compose messages consisting of relevant elements.

Tailored messages have been shown to be more effective than generic messages in changing health behavior (Krebs et al., 2010), among others because they attract and keep the attention of the receiver (Ruiter et al., 2006).

Changing Environmental Conditions

An individual is embedded within social networks, organizations, community, and society; and each lower level is embedded within higher levels. Next to changing the individual, the health promoter wants to change unhealthy environments and the agents that control or influence the environment, or more precise: the determinants of those agents. These determinants are personal but also environmental. And a personal determinant of, for example, a manager is not a personal health threat, but a threat for the organization. And perceived self-efficacy is not related to changes in the manager's personal life, but confidence in bringing about organizational changes. All methods at the individual level may be used for changes at the higher environmental levels. But at those higher levels, there are more theoretical insights that may be applied.

Some theories can be applied to any environmental level: systems theories, theories of power, and empowerment theories. Systems theories are used to define and address all levels of the environment as interrelated social systems (National Cancer Institute, 2007). Methods derived from systems theories include systems change, participatory problem solving, and feedback. Power theories identify various forms of power, and agents at each level with power are able to exert their influence to make environmental changes (Turner, 2005). Methods include persuasion, coercion, community organizing, and agenda setting. Empowerment theories address the process by which community members become involved in their communities and take action to gain power and control (Minkler et al., 2008). Methods include participation, enactive mastery experiences, feedback, and modeling.

Interpersonal level theories deal with social networks and social support: family members, friends, peers, neighbors, associates, and service providers. A social network is an analytic framework for understanding relationships among members of social systems (Heaney and Israel, 2008). Social networks provide emotional, instrumental, informational, and appraisal support. Methods for change include enhancement of existing networks, use of lay health workers, participatory problem solving, and linking members to new networks.

Organizational level theories focus on the development of health promoting organizational structures and culture (Butterfoss et al., 2008; Cummings and Worley, 2014). Methods at the organizational level include sense making, participatory problem solving, organizational diagnosis and feedback, and technical assistance. Stakeholder Theory provides guidance to health promoters, working through their own organization or through the community to change an organization's policies and practices (de Bakker and den Hond, 2008), using methods such as alliances, coalitions, and social action.

Community-level theories relate to coalitions, social capital, social norms, conscientization, and community organization. A coalition is a method for influencing community and policy environmental outcomes (Butterfoss and Kegler, 2009). Social capital is defined as the processes and conditions among people and organizations that lead to accomplishing a goal of mutual social benefit (Green and Kreuter, 2005). Social capital is associated with good health. Community members and opinion leaders at various levels demonstrate and support health promoting behavior, resulting in health promoting social norms (Kelly, 2004). Methods to influence social norms include entertainment education and behavioral journalism. Conscientization, or critical consciousness, links individual- and community-level empowerment and is a key method for strength-based community organization (Wallerstein et al., 2004). Methods include reflection-action-reflection and nonjudgmental small-group discussion. Community organization theories distinguish three types of community organization (Minkler et al., 2008; Rothman, 2004). Locality development raises consciousness about the underlying causes of problems and identifies strategies for action. Social planning uses information derived from empirical research. Social action is based on coercive change and seeks to redress imbalances of power. Methods to promote community capacity include participation, using lay health workers, and media advocacy. The community, in turn, may use methods such as participatory problem solving, forming coalitions, and advocacy to promote a healthy environment.

Theories of public policy identify policy makers as the agents for the environmental outcomes of health promoting policies (Clavier and de Leeuw, 2013). These agents enact

legislation, regulation, and policy (Longest, 2006). Agenda-Building Theory and Multiple Streams Theory both describe agenda setting as the process of moving an issue to a higher agenda status (Kingdon, 2003). In policy advocacy, timing is essential; and opportunity knocks for those advocates who are well prepared at the moment that the 'windows' between politics, problems, and policies are open. Methods at the societal level include policy advocacy, agenda setting, and timing for policy windows. While quite difficult to study, there are convincing examples of scientific research and evaluation at environmental levels, including the policy level, that result in better informed policy and practice (see: www.CampbellCollaboration.org; Chen, 2014).

These environmental changes relate to the definition of settings for health promotion that the WHO has formulated in the Ottawa Charter, focusing on the contexts in which people live, work, learn, and relax (WHO, 1986). The setting approach emphasizes the potentials for health in daily life rather than disease, using natural environments like the workplace, the school, the hospital, the community, and the city. Since settings are mutually embedded (a school in a neighborhood in a city) this approach adopts an ecological view.

Using Theories in Health Promotion: Implementation

Development of an educational intervention includes making plans for program implementation, because no intervention can have any effect when it is not implemented. Implementation ensures that the intervention reaches the intended population. The adoption of innovations is a systematic process, following the pattern of diffusion of innovation (Oldenburg and Glanz, 2008; Rogers, 2003). Implementers can be distinguished by adopter category, such as innovators, early adopters, early majority, late majority, and laggards. Relevant performance objectives are adoption, implementation, and sustainability of the intervention. Especially sustainability may require an organizational change in the implementing organization (Durlak and DuPre, 2008). A number of characteristics of the intervention are related to faster adoption: compatibility, relative advantage, flexibility, observability, 'trialability' (can the innovation be tried before making a decision?), low risk, reversibility, and low complexity.

Currently, diffusion researchers stress the importance of carefully developing the linkage between the change agent system and the target group system (Fleuren et al., 2004; Oldenburg and Glanz, 2008). In, for example, smoking prevention in schools, it is important to involve teachers and students in the development of the program, as well as principals and parents. In, for example, cooperation among different health professionals to achieve continuity of care, it is important to involve all those groups in the development of the protocol.

Epilogue: Planned Health Promotion

Health promotion is any planned combination of educational, political, regulatory, and organizational supports for actions and conditions of living conducive to the health of individuals, groups, or communities. Health education is a planned activity, stimulating learning through communication in order to promote health behavior.

Health promotion planning involves a series of phases where quality of life, health problems, health behavior, determinants of behavior, and possible interventions are successively analyzed, followed by the development and implementation of the intervention and evaluation of the process, impact, and outcome. Careful planning is important, since quality of planning has been shown to determine effectiveness of the intervention.

Intervention Mapping provides a helpful protocol for theory- and evidence-based health promotion planning. A wide range of theories from the behavioral sciences may contribute fruitfully to the analysis of determinants and to the development and implementation of health promotion interventions.

See also: Health Psychology; Social Psychology.

Bibliography

Abraham, C., Michie, S., 2008. A taxonomy of behavior change techniques used in interventions. Health Psychology 27, 379–387.

Albarracín, D., Gillette, J.C., Earl, A.N., Glasman, L.R., Durantini, M.R., Ho, M.-H., 2005. A test of major assumptions about behavior change: a comprehensive look at the effects of passive and active HIV-prevention interventions since the beginning of the epidemic. Psychological Bulletin 131, 856–897.

van Achterberg, T., Huisman-de Waal, G.G.J., Ketelaar, N.A.B.M., Oostendorp, R.A., Jacobs, J.E., Wollersheim, H.C.H., 2010. How to promote healthy behavior in patients? A review of systematic reviews. Health Promotion International. http://dx.doi.org/10.1093/heapro/daq050.

Bartholomew, L.K., Parcel, G.S., Kok, G., Gottlieb, N.H., Fernández, M.E., 2011. Planning Health Promotion Programs; an Intervention Mapping Approach, third ed. Jossey-Bass, San Francisco, CA.

Bastable, S.B., 2006. Essentials of Patient Education. Jones & Bartlett, Burlington, MA.

Belansky, E.S., Cutforth, N., Chavez, R.A., Waters, E., Bartlett-Horch, K., 2011. An adapted version of intervention mapping (AIM) is a tool for conducting community-based participatory research. Health Promotion Practice 12, 440.

Brug, J., Conner, M., Harre, N., Kremers, S., McKellar, S., Whitelaw, S., 2005. The transtheoretical model and stages of change: a critique. Health Education Research 20, 244–258.

Butterfoss, F.D., Kegler, M.C., 2009. The community coalition action theory. In: DiClemente, R.J., Crosby, R.A., Kegler, M.C. (Eds.), Emerging Theories in Health Promotion Practice and Research, second ed. Jossey-Bass, San Francisco, pp. 237–276.

Butterfoss, F.D., Kegler, M.C., Francisco, V.T., 2008. Mobilizing organizations for health promotion: theories of organizational change. In: Glanz, K., Rimer, B.K., Viswanath, K. (Eds.), Health Behavior and Health Education: Theory, Research, and Practice, fourth ed. Jossey-Bass, San Francisco, pp. 335–362.

Buunk, A.P., van Vugt, M., 2013. Applying Social Psychology: From Problems to Solutions. Sage, London.

de Bakker, F.G.A., den Hond, F., 2008. Introducing the politics of stakeholder influence: a review essay. Business & Society 47, 8–20.

de Bruin, M., Viechtbauer, W., Hospers, H.J., Schaalma, H.P., Kok, G., 2009. Standard care quality determines treatment outcomes in control groups of HAART adherence intervention studies: implications for the interpretation and comparison of intervention effects. Health Psychology 28, 668–674.

Champion, V.L., Skinner, C.S., 2008. The health belief model. In: Glanz, K., Rimer, B.K., Viswanath, K. (Eds.), Health Behavior and Health Education: Theory, Research, and Practice, fourth ed. Jossey-Bass, San Francisco, pp. 45–65.

Chen, H.T., 2014. Practical Program Evaluation; Theory-driven Evaluation and the Integrated Evaluation Perspective, second ed. Sage, Thousand Oakes, CA.

Clark, N.M., 2003. Management of chronic disease by patients. Annual Review of Public Health 24, 289–313.

Cummings, T., Worley, C., 2014. Organization Development and Change. Cengage Learning.
Clavier, C., de Leeuw, E. (Eds.), 2013. Health Promotion and the Policy Process. Oxford University Press.
Danner, U.N., Aarts, H., de Vries, N.K., 2008. Habit vs. intention in the prediction of future behavior: the role of frequency, context stability and mental accessibility of past behavior. British Journal of Social Psychology 47, 245–265.
DiClemente, R.J., Salazar, L.F., Crosby, R.A., 2012. Health Behavior Theory for Public Health. Jones & Bartlett, Burlington, MA.
Durlak, J.A., DuPre, E.P., 2008. Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. American Journal of Community Psychology 41, 327–350.
Fishbein, M., Ajzen, I., 2010. Predicting and Changing Behavior: The Reasoned Action Approach. Taylor & Francis, New York.
Fisher, J.D., Fisher, W.A., Bryan, A.D., Misovich, S.J., 2002. Information-motivation-behavioral skills model-based HIV risk behavior change intervention for inner-city high school youth. Health Psychology 21, 177–186.
Fisher, J.D., Fisher, W.A., Shuper, P.A., 2009. The information-motivation-behavior skills model of HIV preventive behavior. In: DiClemente, R.J., Crosby, R.A., Kegler, M. (Eds.), Emerging Theories in Health Promotion Practice and Research, second ed. Jossey-Bass, San Francisco, pp. 21–64.
Fleuren, M., Wiefferink, K., Paulussen, T., 2004. Determinants of innovation within health care organizations: literature review and Delphi study. International Journal for Quality in Health Care 16, 107–123.
Glanz, K., Rimer, B.K., Viswanath, K. (Eds.), 2008. Health Behavior and Health Education: Theory, Research, and Practice, fourth ed. Jossey-Bass, San Francisco, CA.
Gollwitzer, P.M., Sheeran, P., 2006. Implementation intentions and goal achievement: a meta-analysis of effects and processes. In: Zanna, M.P. (Ed.), Advances in Experimental Social Psychology, vol. 38. Academic Press, San Diego, pp. 69–119.
Green, L.W., Kreuter, M.W., 2005. Health Program Planning: An Educational and Ecological Approach, fourth ed. McGraw Hill Professional, New York.
Greenwald, A.G., McGhee, D.E., Schwartz, J.L.K., 1998. Measuring individual differences in implicit cognition: the implicit association test. Journal of Personality and Social Psychology 74, 1464–1480.
Hassin, R.R., Uleman, J.S., Bargh, J.A. (Eds.), 2005. The New Unconscious. Oxford University Press, New York.
Heaney, C.A., Israel, B.A., 2008. Social networks and social support. In: Glanz, K., Rimer, B.K., Viswanath, K. (Eds.), Health Behavior and Health Education: Theory, Research, and Practice, fourth ed. Jossey-Bass, San Francisco, pp. 189–210.
Hofmann, W., Friese, M., Strack, F., 2009. Impulse and self-control from a dual-systems perspective. Perspectives on Psychological Sciences 4, 162–176.
Kazdin, A.E., 2008. Behavior Modification in Applied Settings, sixth ed. Waveland Press, Long Grove, IL.
Kelly, J.A., 2004. Popular opinion leaders and HIV prevention peer education: resolving discrepant findings, and implications for the development of effective community programmes. AIDS Care 16, 139–150.
Khan, L.K., Sobush, K., Keener, D., Goodman, K., Lowry, A., Kakietek, J., et al., 2009. Recommended community strategies and measurements to prevent obesity in the United States. Morbidity and Mortality Weekly Report 58 (RR07), 1–26.
Kingdon, J.W., 2003. Agendas, Alternatives, and Public Policies, second ed. Longman, New York.
Kok, G., Gottlieb, N.H., Commers, M., Smerecnik, C., 2008. The ecological approach in health promotion programs: a decade later. American Journal of Health Promotion 22, 437–442.
Kools, M., 2012a. Making written materials easy to understand. In: Abraham, C.A., Kools, M. (Eds.), Writing Health Communication: An Evidence-based Guide. Sage, London, pp. 23–42.
Kools, M., 2012b. Making written materials easy to use. In: Abraham, C.A., Kools, M. (Eds.), Writing Health Communication: An Evidence-based Guide. Sage, London, pp. 43–62.
Krebs, P., Prochaska, J.O., Rossi, J.S., 2010. A meta-analysis of computer-tailored interventions for health behavior change. Preventive Medicine 51, 214–221.
Latham, G.P., Locke, E.A., 2007. New developments in and directions for goal-setting research. European Psychologist 12, 290–300.
Longest, B.B., 2006. Health Policymaking in the United States, fourth ed. Health Administration Press, Chicago.
Marlatt, G.A., Donovan, D.M. (Eds.), 2005. Relapse Prevention; Maintenance Strategies in the Treatment of Addictive Behaviors, second ed. Guilford, New York.
McAlister, A.L., Perry, C.L., Parcel, G.S., 2008. How individuals, environments, and health behaviors interact: social cognitive Theory. In: Glanz, K., Rimer, B.K., Viswanath, K. (Eds.), Health Behavior and Health Education, fourth ed. Jossey-Bass, San Francisco, pp. 169–188.
McBride, C.M., Rimer, B.K., 1999. Using the telephone to improve health behavior and health service delivery. Patient Education and Counseling 37, 3–18.
McGuire, W.J., 2001. Input and output variables currently promising for constructing persuasive communications. In: Rice, R.E., Atkin, C.K. (Eds.), Public Communication Campaigns, third ed. Sage, Thousand Oaks, CA, pp. 22–48.
Miller, W.R., Rollnick, S., 2012. Motivational Interviewing: Helping People Change. Guilford Press.
Minkler, M., Wallerstein, N., Wilson, N., 2008. Improving health through community organization and community building. In: Glanz, K., Rimer, B.K., Viswanath, K. (Eds.), Health Behavior and Health Education: Theory, Research, and Practice, fourth ed. Jossey-Bass, San Francisco, pp. 287–312.
Montano, D.E., Kasprzyk, D., 2008. Theory of reasoned action, theory of planned behavior, and the integrated behavioral model. In: Glanz, K., Rimer, B.K., K. Viswanath (Eds.), Health Behavior and Health Education: Theory, Research, and Practice, fourth ed. Jossey-Bass, San Francisco, pp. 67–96.
Mullen, P.D., Green, L.W., Persinger, G.S., 1985. Clinical trials of patient education for chronic conditions: a comparative meta-analysis of intervention types. Preventive Medicine 14, 753–781.
National Cancer Institute, 2007. Greater than the Sum: Systems Thinking in Tobacco Control (Publication No. 06–6085, Tobacco Control Monograph No. 18). US Department of Health and Human Services, National Institutes of Health, National Cancer Institute, Bethesda, MD.
Oldenburg, B., Glanz, K., 2008. Diffusion of innovations. In: Glanz, K., Rimer, B.K., Lewis, F.M. (Eds.), Health Behavior and Health Education: Theory, Research, and Practice, fourth ed. Jossey-Bass, San Francisco, pp. 313–333.
Peters, G.J.Y., de Bruin, M., Crutzen, R., Everything should be as simple as possible, but no simpler: towards a protocol for accumulating evidence regarding the active content of health behaviour change interventions. Health Psychology Review 1–14. http://dx.doi.org/10.1080/17437199.2013.848409, ahead-of-print.
Peters, L.H.W., Kok, G., Ten Dam, G.T.M., Buijs, G.J., Paulussen, T.G.W.M., 2009. Effective elements of school health promotion across behavioral domains: a systematic review of reviews. BMC Public Health 9, 182. http://dx.doi.org/10.1186/1471-2458-9-182.
Petty, R.E., Barden, J., Wheeler, S.C., 2009. The elaboration likelihood model of persuasion: developing health promotions for sustained behavioral change. In: DiClemente, R.J., Crosby, R.A., Kegler, M. (Eds.), Emerging Theories in Health Promotion Practice and Research, second ed. Jossey-Bass, San Francisco, pp. 185–214.
Prochaska, J.O., Redding, C.A., Evers, K.E., 2008. The transtheoretical model and stages of change. In: Glanz, K., Rimer, B.K., Viswanath, K. (Eds.), Health Behavior and Health Education: Theory, Research, and Practice, fourth ed. Jossey-Bass, San Francisco, pp. 97–121.
Rogers, E.M., 2003. Diffusion of Innovations, fifth ed. Free Press, New York.
Rothman, J., 2004. Three models of community organization practice, their mixing and phasing. In: Cox, F.M., Erlich, J.L., Rothman, J., Tropman, J.E. (Eds.), Strategies of Community Organization: A Book of Readings, third ed. F.E. Peacock, Itasca, IL.
Ruiter, R.A.C., Kessels, L.T.E., Jansma, B.M., Brug, J., 2006. Increased attention for computer-tailored health communications: an event-related potential study. Health Psychology 25, 300–306.
Ruiter, R.A., Kessels, L.T., Peters, G.J.Y., Kok, G., 2014. Sixty years of fear appeal research: current state of the evidence. International Journal of Psychology 49 (2), 63–70.
Schaalma, H.P., Kok, G., 2011. A School HIV-prevention Program in the Netherlands. Case study 3 at the Wiley companion site for Bartholomew et al., 2011. http://bcs.wiley.com/he-bcs/Books?action=index&itemId=0470528516&bcsId=6246.
Simons-Morton, B., McLeroy, K.R., Wendel, M.L. (Eds.), 2012. Behavior Theory in Health Promotion Practice and Research. Jones & Bartlett, Burlington, MA.
Steinberg, L., 2007. Risk taking in adolescence: new perspectives from brain and behavioral science. Current Directions in Psychological Science 16, 55–59.
Turner, J.C., 2005. Explaining the nature of power: a three-process theory. European Journal of Social Psychology 35, 1–22.
Wallerstein, N., Sanchez, V., Velarde, L., 2004. Freirian praxis in health education and community organizing: a case study of an adolescent prevention program. In: Minkler, M. (Ed.), Community Organizing and Community Building for Health, second ed. Rutgers University Press, Piscataway, NJ, pp. 218–236.
World Health Organization. The Ottawa Charter for Health Promotion. Adopted on November 21, 1986.

Health, Geography of

Jennifer O'Brien, School of Environment, Education and Development, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Health geography emerged from the traditions of positivist medical geography with the cultural turn in geography in the 1990s and the incorporation of cultural and social theory into interpretations of health in place. Health geographers utilize a range of theoretical frameworks, methods, and concepts to probe the dynamic linkages among power, space, place, and time at different spatial scales. Using different examples to highlight the new approach to understanding health, this section will explain how health geographers have contributed to health discussions, often through policy, to alleviate health inequalities and improve health.

Introduction

Health geography emerged from the traditions of positivist medical geography with the cultural turn in geography in the 1990s and the incorporation of cultural and social theory into interpretations of health in place. Understanding health is complex. As set in its Constitution in 1948, the World Health Organization defines health as "the complete state of physical, emotional and social well-being and not merely the absence of disease or infirmity." This definition suggests that good health is utopian; under this understanding of health, an Olympic athlete who is physically fit and free from disease might technically be unhealthy if she is worried about, for example, her personal development or even a forthcoming bill. As health is complex, it has to be interpreted in a manner that endeavors to understand its nuances. Health geographers utilize a range of theoretical frameworks, methods, and concepts to probe the dynamic linkages among power, space, place, and time at different spatial scales of the globe, nation, region, home, and body (Atkinson et al., 2011).

Health geographers aim to understand questions of health, healing, and well-being through three key approaches. First, health geographers move beyond a spatial analysis of disease to consider the role that place plays in health. Second, the incorporation of cultural theory enables nuanced understandings of the way cultural beliefs and practices structure the place of health and healthcare provision. Third, health geographers borrowed from critical theory to provide insights into the way that people and places can be negatively affected by large-scale forces such as global firms or changing political frameworks (Cutchin, 2007), conjuring questions of responsibility at various scales. Moving beyond multilevel modeling or epidemiological mapping, health geographers utilize a range of methodologies including interviews, focus groups, and ethnographic engagement, often in combination. With new insights and approach, health geographers have retheorized previously marginalized health spaces, which has facilitated a reinterpretation of health in place in a manner that can contribute to health understanding, activism, and policy.

This article will first consider how health geographers used cultural theory to reconceptualize space and place, which provided new lines of critical enquiry and methodologies that enabled them to contribute to key debates and policies surrounding health and healthcare. These changes will be illustrated across varying spatial scales, beginning with the way that health geographers used social and cultural theory to reconsider the space of the human immunodeficiency virus (HIV) and its transmission. At the regional scale, the article will consider the debate surrounding the context and composition of place in health, highlighting key questions of structural power and responsibility. At the local scale, the place of home will illustrate health geographers' growing interest in care through the transformation of the home and the body as sites of caring.

Space and Place in Health

Geographers have always been interested in questions of space and place and the processes, interactions, and practices that occur within them. The 'cultural turn' of the mid-1990s shifted geographical focus away from utilitarian concepts of space as simple area descriptors, toward a more reflective understanding of the meaning, utility, and construction of space and place. Both medical and health geographers aim to alleviate health inequalities through an understanding of health and illness in place, but the two take different methodological approaches. Stemming from a culturally informed understanding of place, health geographers began to consider the roles that place, and the processes occurring within place, play in the making of and experiences of health (Gattrell, 2002; Williams, 1998). Unlike positivist medical geography, which conceived space as a blank canvas upon which medical subjects could be mapped, health geographers understand place as a complex of material, sociological, experimental, and philosophical phenomena that inform and affect health, healthcare, and health behaviors. There is still an appreciation of the physical attributes of places, as will be further discussed in the next section, but the experience of health in place is recognized through the social, cultural, economic, or political processes that simultaneously construct and affect the health space. Indeed, drawing upon the aforementioned World Health Organization's definition, far from being the mere absence of disease, health is an entirely personal, temporal, and fluid concept. As health knowledges are multiple and socially constructed in place, health geographers argue that such knowledges are not always best suited to an empiricist or positivist interpretation. This new socially constructed

understanding of health, and the role of place in health, enabled health geographers to rethink existing places of health.

Gesler's (1992) notion of therapeutic landscapes is perhaps the seminal example of this evolution in thinking. Used as a way to critically analyze the complex dynamics among place, well-being, and healthcare, therapeutic landscapes are places that are seen to be conducive to healing. Gesler's original thesis recognized that certain physical places, or geographical locations, have potential for physical, mental, and spiritual health promotion and for the process of healing such as Lourdes, France. With further development of the thesis, public health interest moved away from physical places of well-being (Lourdes, etc.) to considering the potential therapeutic or healing properties of other spaces such as gyms, prisons, parks, and wilderness (see Smyth, 2005). Within those spaces of healing, such as a park, the physical, social, and symbolic organization of the space is much more significant than its geographical location or physical constitution. The social impact of an individual appreciating the sense of peace she experiences through sunshine-strewn wilderness may have greater therapeutic benefit than the physical trees and fresh air; her temporal place-based experience constructs a place of healing. A health geographer's approach recognizes how the space promotes health, rather than just where it is or how it is physically constructed. Developing the thesis, such spaces of health dramatically vary in scale from a vast open wilderness to Grandfather's chair which, positioned by the window, perhaps enhances his health and well-being much more than a hospital could as a space of therapy (Gattrell, 2002). In health geography, therefore, place is more than a specific geographic location, empty environment, or a 'backdrop' for health interactions (Dyck, 1999: 247) but a place of process, practice, and meaning that is situated and socially produced in particular contexts; place thereby reflects and informs understandings of health, which, in turn, affects the interpretation of place. A good example is the aforementioned wilderness that brought such therapy to one individual, but could easily be a space of fear for another.

A constructivist understanding of place and health suggests that neither can exist independent of the actions or factors that constitute and are contained within it. As Gesler (2003: 1) put it, "healing and place are inseparable." Health actions are embedded in broader spatiotemporal social, historical, cultural, economic, political, and often power-bound contexts (Smyth, 2005) which are informed by, and constitute, place. Power impacts place and the health actions within them at various scales through broader structural processes in play such as governance by large corporations or legislation and also through, for example, individual power to assert health changes over the body. The following sections will apply the new theoretical approach to understanding health in place and highlight how health geographers have contributed to the discussion.

Inequalities and HIV

Since the World Health Organization issued its Commission on the Social Determinants of Health, 'Closing the gap' in 2008, questions of global health inequalities have risen significantly in importance as both a subject for academic research and a focus of international policy development. While global levels of poverty are decreasing, inequalities are increasing with significant implications for health. Curtis (2004), among many others, argues that the geographies of health inequalities are a long-established and central tenet of research in medical geography but one that has significantly grown in opportunity through the new health geography. A health geographer's understanding of HIV provides a good global example.

HIV is transmitted through the exchange of bodily fluids, blood, semen, vaginal secretions, and breast milk. The virus attacks the immune system of infected individuals which impairs their capacity to resist infection. When the infected individual's immune system has a CD4 cell count of below 200 cells mm^{-1}, the person is clinically diagnosed as having the acquired immune deficiency syndrome. According to the World Health Organization, since the beginning of the epidemic almost 75 million people have been infected with the HIV virus and about 36 million people have died of HIV. HIV is largely associated with areas of the developing world, such as sub-Saharan Africa where inequalities in healthcare access exacerbate the rate of HIV-related death. With the first outbreak of the epidemic, epidemiologists and medical geographers mapped the incidence of disease transmission which showed the spatial distribution but did little to explain transmission.

In many societies, particularly during the 1990s and early twenty-first century, HIV was a disease of stigma and shame due to its association with drug use and unprotected sexual intercourse. Fear of ostracism or discrimination prevented many people from being treated for the virus. The stigma then became a barrier to knowledge to prevent HIV transmission and could prevent access to appropriate treatment. Despite the expected low recording of transmission rates, the number of people living with, or dying from HIV was growing rapidly and exacerbating socioeconomic inequalities. Among others, Craddock (2000) stressed the need to move beyond a structural analysis of HIV in geographical location to incorporate social theory to examine the power-bound social health *place* of HIV and why it was that, for example, some groups were more vulnerable to HIV transmission, or might be less likely to access treatment or prevention knowledge. Highlighting a health geographer's use of theory, a feminist analysis enabled contemplation of the female body that in rural Uganda had few sexual rights and thus became a vulnerable place of disease transmission. As a good example, developing the health geographer's approach, Kuhanan (2010) considered the place of high HIV incidence in Rakai, Uganda. Rather than just mapping disease incidence, Kuhanen recognized that the long history of the epidemic needed to be understood as a social and economic problem, as well as an epidemiological one and considered sexualized public spaces as key places of HIV transmission. Kuhanen reflected upon the dramatic move away from the traditional colonial control over female movement and space and the growth of rural-to-urban migration, which instigated informal sex economies as women struggled to earn a living. While reconstructing the power-bound identity of 'women' in place, Kuhanan (2010) identified that the shift in roles of women and the growth of the informal sector

constructed places of HIV transmission and in turn created places of transmission which informed behavior. This analysis also highlights the key role that geographic scale plays in health geography. Global distributions of HIV might be decreasing in figures but do not reveal the growing rates of inequality in standards of living, or explain rates of transmission.

Seeing the site of HIV transmission as more than simply a geographical location but also as a power-bound space that is constructed by both health behavior and constructs, it reveals a very different landscape of transmission than does a purely epidemiological approach. Health geographers have thus been able to contribute to HIV discussions to aid policy: Craddock (2007), for example, considered the role of pharmaceutical companies in the geographies of HIV in Africa; Ingram (2008) considered the political construct of HIV as it was framed as a biosecurity emergency in global governance strategies that was key in the intersection of HIV, immigration, national security, and notions of refugee and asylum; and Gwanzura-Ottemoller and Kesby (2005) considered the place of HIV from the perspective of young people and their negotiation of safe sexual practices.

The Regional: Context, Composition, and Air Quality

While understanding that place and healing are inseparable, a regional example demonstrates that place itself matters in a true understanding of health. Macintyre et al. (2002) discuss the debate that arose in the early 1990s among health geographers over the relative importance of people or place characteristics in shaping people's health experiences. Although health spaces are constructed through behavior and certainly are not mere physical containers of healthcare interaction, places do affect health. That impact can be through the physical attributes of place, known as contextual effects, which are the direct effects of air quality and landscape but also indirect impacts through, for example, the availability of healthy food stuffs or recreational facilities. Compositional factors of a place can also shape health; these factors include the people themselves and key attributes such as gender, ethnicity, socioeconomic, and employment status. This long-standing and, as Smyth (2008) would argue, uneasy, debate questions how geographical inequalities in health can be explained by the composition of a population and the context of the place. The varying impact of context and composition of place to explain spatial variations in health has implications for who has the responsibility to change variables that are impacting health.

More recently, the context versus compositional debate has become less polarized as the nuanced approach of health geographers considers health in place and utilizes a variety of methodologies to understand questions of health inequality such as access to health services, obesity, and pollution levels (Macintyre et al., 2002). Day (2007), for example, considered how place is significant in the experience of environmental hazards but particularly considered the ways in which inequalities are exacerbated through exposure or risk. Applying the mixed methodology of a health geographer to consider place in health, Day studied four different neighborhood areas in the London Borough of Barnet which had a wide variation in air pollution levels and indeed privilege. The contextual understanding of health in place highlighted the high levels of air pollution in the boroughs but, conducting the study over 2 years and using semistructured interviews, Day identified a disjuncture between peoples' judgments of air quality (as gleaned from the interviews) and meteorological measures of air quality. The study showed that air pollution was understood and experienced both through, and as part of, multidimensional place and provided crucial insights into the production of some inequality of experience. Day's approach recognized the need to pay more attention to the broader understanding of place and consider both the contextual and compositional factors in trying to improve living environments and alleviate inequality. Furthermore, Cairns (2012) utilized data that challenged the context versus compositional debate by suggesting that some economically deprived areas in England defied the trend of a positive and graded association between area deprivation and health outcomes in Britain. Their study revealed that certain areas with a high ethnic density had lower than expected mortality rates due to positive social support networks or 'health resilience' which they found to have a positive impact on health.

Cutchin (2007) furthered this argument by highlighting the need to bridge the epidemiological understanding of health in place with a more social understanding but added the necessity to consider how a health geographer's utilization of critical theory provides insights into the way that people and places can be negatively affected by larger scale forces, such as the state or international organizations. Cutchin moved beyond the classic therapeutic landscape literature and applied a number of health geography approaches including cultural ecology (which considers the notions of culture and environment interaction) to the situation of Texas City where he was considering how the health of local residents was affected by the stress response created by the petrochemical industry and other stressors. The study drew on context and composition by contemplating the physical environment, particularly petrochemical production, as well as the status, social ecology, and social relationships of the population. Developing the therapeutic landscape concept, Cutchin (2007) discussed how culture can become material through the landscape, which in turn can be read as an often power-bound materialization of that culture. The example here was how the memory of a terrible, large-scale disaster where a ship caught fire and exploded was physically inscribed on the landscape through memorials and parks; in turn the landscape could be read as a space of disaster with resultant impacts on residents. Cutchin also highlighted the insight of a political ecology approach where traditional ecological views are expanded within a political economy perspective to consider the structures or agencies that shape environments and health risk for populations, for example, the availability of green spaces within a council borough. Applied to the Texas City situation, this approach returned to the power-bound question of responsibility for health and how, in this case, the petrochemical industry had a responsibility for the impact that it was having on the health of the population both directly and indirectly. Cutchin's study highlights how different approaches within health geography enable a more holistic perspective of place and process on the question of health and provide evidence of the interplay of values and power from various geographic scales.

Caring, Home, and Self

Since early in 2000, health geographers have begun to consider the intricacies of the interrelationships between people, places, and materialities in different spaces of care in an effort to critique the unproblematic conceptualization of care as it is understood within a medical model (see Atkinson et al., 2011). Lawson (2007) explains how we live in times defined by relentless right-wing extension of market relations into all spheres of our existence. A consumer- or client-related view of healthcare, university education, care for the elderly, support for the working poor, or environmental protection – generated by the market-orientated approach – challenges not only how such services are perceived and delivered, but where they are delivered. The reduction in public provision of care and social support, for example, coupled with a neoliberal ideology based on individual over collective responsibility, has challenged more traditional understandings of care delivery, particularly in care of the elderly (Cutchin, 2003). By considering the role that place plays in the health and healthcare delivery, health geographers have questioned notions of responsibility in caring for the elderly, where care should be delivered, and the impact that those changes have had. Addressing these questions of caring in place illustrates how health geographers have been able to traverse the theoretical, empirical, and policy gulf and use their nuanced understanding to actively influence policy, engage with the public, and make an impact.

In many developed countries, from the latter half of the twentieth century, the implementation of policy focused on aging and community care in place, re-centered the main space of care for vulnerable people from institutional place (hospitals or residential homes) into domestic space, and by default into the realm of women (Milligan et al., 2007). With the relocation of care the boundaries between a place of healthcare delivery and a domestic space of home blur, which in turn challenges existing knowledge systems and (re)constructs the identity of the carers in relation to the care recipient. Such challenges can alter relationships and perceptions of the meanings of home. For example, if an infirm grandmother moves into her daughter's family home, the relationship between mother and daughter can change into one of dependent but informal care delivery; in turn the family home could be transformed into a place of informal care delivery through medical or caring paraphernalia and renegotiated relational identities (Dyck et al., 2005). The home can potentially become a place of marginalization and oppression in which carers, particularly women caring for long-term health needs, can feel isolated and disconnected from the outside world. McDowell (2004) discusses how that caring can be quantified; for example, care homes make substantial profits by creating spaces which simulate home caring for the growing elderly population; the conversion into a commodity of the day-to-day care that is so taken for granted or is often poorly paid work. Health geographers' interpretations of care delivery for the elderly and how that is impacted by place, and in turn affects place, have contributed to healthcare delivery in a variety of outputs. Wiles and Wild (2011), for example, have made policy recommendations about the ideal place(s) to grow older based on research conducted on two very different communities in Auckland and South Waikato. Dyck's 2014 research on migration, health, and well-being was used to produce a DVD that is being used by two mental health counseling groups in the Borough of Greenwich, London, and by community organizations. Moreover within the discipline, the necessity or moral obligation to care through ethical and impactful research has become both a driving force and a measure of success (see Lawson, 2007).

Health geographers' understanding of care has cast considerable critical insight into retheorizing the space of the body as a place of health and healthcare. The body itself is far from a new focus of analysis; literature is abundant across a host of disciplines from sociology to sports science addressing questions such as diet and fitness, cosmetic modification, sexual health and its expression, drug use in sport, and an increasing awareness of risk discourses surrounding the body (see Lupton, 2003), much of it stemming from feminist literature of the 1980s (see Haraway, 1987). The body is a space of health interaction, experience, and expression, which, beyond blood and bone, is constructed by cultural, political, historical, and indeed personal process and discourse. Bodies are intrinsically bound up with our sense of identity; they are cognitive, subconscious, and emotional as well as being experienced through physical flesh. A poststructural analysis of a body considers it as a metaphor for social organization and social anxieties which generate questions of responsibility. Bodies get sick, they can get strong, they are potentially dirty or leaky, they are gendered, sexualized, a space of production and reproduction, ultimately a site for death. Bodies thus are conceptualized as spaces in need of regulation and control, whether by an individual or a third party such as the state, hospitals, beauticians, or schools, and can be read as personal spaces that represent global processes (Atkinson, 2011). As reflective of those processes, the body is always an ongoing malleable project that is never finished.

Health geographers recognized that, with constructionist understandings of place and health and multiple power-bound knowledges, the body could no longer be considered a given – and often taken for granted – reality, but a product of certain kinds of knowledge and discourses that are subject to change. Bodies are shaped by social relations but also play a role in the construction of those politically, historically, and culturally bound relationships and can thus be read as a landscape of those power-bound forces. Health geographers have contributed to the broader body discussion focusing on the gendered (e.g., Binnie, 2011) or disabled body (see Gleeson, 1999), which has developed with understandings of the commodification of care (Green and Lawson, 2011) to consider caring for the body, care through the internet (Atkinson and Ayers, 2010), or aesthetic surgery (Atkinson, 2011).

Conclusion

Prompted by the cultural turn in geography, health geographers moved beyond the epidemiological understandings of the spatial analysis of disease to recognize the role that place plays in health and its understandings. While the physical attributes (the context and composition) of place were still recognized, the new health geography retheorized place as a complex material, sociological, experimental, and

philosophical phenomenon that both constructs and is constructed by health. By incorporating more cultural theory into the place-based analysis, health geographers were better positioned to understand the nuanced processes and practices that were shaping health behavior and inequalities – as illustrated here by the feminist landscape of HIV transmission. Further engagement with critical literature enabled health geographers to expand the analysis of health from the individual scale and consider the broader structural processes and powers that construct a potentially negative landscape of health, such as the petrochemical plant in Texas, which raises key questions of roles and responsibility in maintaining health. The new approach enabled health geographers to retheorize previously neglected or sensitive spaces of health, such as the home, and unveil complex questions of power and identity in place. Kearns and Moon (2002) argue that health geographers still have some way to go to be recognized as real contributors to health policy compared to their medical counterparts. Yet, the broader theoretical shift not only has witnessed the marked separation of health geographers from medical geographers, but also has opened up new lines of critical enquiry that, as illustrated earlier, have enabled health geographers to make some significant contributions to the policy agenda from global spaces to the personal health place of the body.

See also: Aging and Health in Old Age; Aging and Older People, Geography of; Environment and Health; Health in Low-Income Countries; Immigration: Health and Mortality; Social Geography.

Bibliography

Atkinson, S., 2011. Scales of care and responsibility: debating the surgically globalised body. Social and Cultural Geography 12 (6), 623–636.
Atkinson, S., Lawson, V., Wiles, J., 2011. Care of the body: spaces of practice. Social and Cultural Geography 12 (6), 563–572.
Atkinson, S., Ayers, A., 2010. The potential of the internet for alternative caring practices for health. Anthropology and Medicine 17, 75–86.
Binnie, J., 2011. Class, sexuality and space: a comment. Sexualities 14 (1), 21–26.
Cairns, J.M., 2012. Defying deprivation: a cross-sectional analysis of area level health resilience in England. Health and Place 18, 928–933.
Craddock, S., 2007. Market incentives, human lives, and AIDS vaccines. Social Science and Medicine 64, 1042–1056.
Craddock, S., 2000. Disease, social identity and risk: rethinking the geography of AIDS. Transactions of the Institute of British Geographers 25, 153–200.
Curtis, S., 2004. Health and Inequality: Geographical Perspectives. Sage, London.
Cutchin, M., 2003. The process of mediated aging-in-place: a theoretically and empirically based model. Social Science and Medicine 57 (6), 1077–1090.
Cutchin, M.P., 2007. The need for the 'new health geography' in epidemiologic studies of environment and health. Health and Place 13, 725–742.
Day, R., 2007. Place and the experience of air quality. Health and Place 13, 249–260.
Dyck, I., 2014. Public Engagement (Online) available at: http://www.geog.qmul.ac.uk/staff/dycki.html (accessed 07.03.14).
Dyck, I., 1999. Using qualitative methods in medical geography: deconstructive moments in a subdiscipline? Professional Geographer 51, 243–253.
Dyck, I., Kantos, P., Angus, J., McKeever, P., 2005. The home as site for long-term care: meanings and management of bodies and space. Health and Place 11, 173–185.
Gattrell, A.C., 2002. Geographies of Health: An Introduction. Blackwell, London.
Gesler, W., 2003. Healing Places. Rowman and Littlefield Publishers Inc., New York.
Gesler, W., 1992. Therapeutic landscapes: medical geographic research in the light of the new cultural geography. Social Science and Medicine 34, 735–746.
Gleeson, B., 1999. Geographies of Disability. Psychology Press.
Green, M., Lawson, V., 2011. Recentring care: interrogating the commodification of care. Social and Cultural Geography 12 (6), 639–654.
Gwanzura-Ottemoller, F.P., Kesby, M.G., 2005. 'Lets talk about sex, baby': conversing with Zimbabwean children about HIV. Children's Geographies 3 (2), 201–218.
Halpern, D., Nazaroo, J., 2000. The ethnic density effect: results from a national community survey of England and Wales. International Journal of Social Psychiatry 46, 34–36.
Haraway, D., 1987. A manifesto for cyborgs: science, technology and socialist feminism in the 1980s. Australian Feminist Studies 2 (4), 1–42.
Ingram, A., 2008. Domopolitics and the international response to HIV/AIDS: governmentality, globalism and security. Area 42, 293–301.
Kearns, R., Moon, G., 2002. From medical to health geography; novelty, place and theory after a decade of change. Progress in Human Geography 26 (5), 605–625.
Kuhanan, J., 2010. Sexualised space, sexual networking and the emergence of AIDS in Rakai, Uganda. Health and Place 26 (2), 226–235.
Lawson, V., 2007. Geographies of care and responsibility. Annals of the Association of American Geographers 97 (1), 1–11.
Lupton, D., 2003. Medicine as Culture: Illness, Disease and the Body in Western Societies, second ed. Sage Publications, London.
Macintyre, S., Ellaway, A., Cummins, S., 2002. Place effects on health: how can we conceptualise, operationalise and measure them? Social Science and Medicine 55, 125–139.
McDowell, L., 2004. Work, workforce, work/life balance and an ethic of care. Progress in Human Geography 28 (2), 145–163.
Milligan, C., Atkinson, A., Skinner, M., Wiles, J., 2007. Geographies of care: a commentary. New Zealand Geographer 63, 135–140.
Milligan, C., Wiles, J., 2010. Landscapes of care. Progress in Human Geography, 1–19.
Smyth, F., 2008. Medical geography: understanding health inequalities. Progress in Human Geography 32 (1), 119–127.
Smyth, F., 2005. Medical geography: therapeutic places, spaces and networks. Progress in Human Geography 29, 488–495.
Wiles, J.L., Wild, K., 2011. Resilient Ageing in Place: Cook Island Report & Recommendations (Cook Island & English Versions).
Williams, A., 1998. Therapeutic landscapes in holistic medicine. Social Science and Medicine 46 (9), 1193–1203.

Health in Low-Income Countries

Solomon R Benatar, University of Cape Town, Observatory, South Africa; and University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the 12 years since publication of the previous issue of this volume, the world has changed significantly. In addition to further widening of many of the disparities in health previously described, the impact and implications of the still evolving global economic crisis and of climate change have led to increasing recognition of how the 'development' agenda has failed, and that global health is one of the major challenges for the twenty-first century (Benatar and Brock, 2011). World population has increased to 7 billion people, half of whom continue to live in severe poverty (on less than US$2–3 a day), most in constant hunger and under squalid conditions, with little prospect of significant improvement unless new solutions are envisaged. This article discusses the health of citizens of very poor countries and the global interdependence of the health of all peoples, including the privileged.

Global Health Trends since the 1940s

Human health and well-being worldwide are profoundly affected by complex interactions between economic, social, and political forces. Some historical facts about tuberculosis in rapidly industrializing countries provide an illuminating example. With better nutrition and living conditions, the annual mortality rate from tuberculosis in the United Kingdom fell from 500 per 100 000 people in the early 1700s to 50 per 100 000 in the early 1940s before antituberculosis medications were discovered. Health in general also improved significantly during this time and life expectancy at birth increased from 37 years in 1750 to 48 years in 1940 (McKeown and Record, 1962; McKeown, 1976).

Between 1945 and 1970, another consolidating wave of advances in health status was associated with reconstruction of a global economy that had been severely compromised by the Great Depression and two devastating World Wars. Growth of the middle class was promoted by a spirit of solidarity forged through access to jobs, universal access to education, and health care and the provision of other social services funded by progressive taxation within well-regulated economies. This is the context within which modern medical advances could be effectively applied to improve health and facilitate the eradication of many epidemic infectious diseases, as exemplified by further reduction in the annual death rate from tuberculosis to about two per 100 000 in the UK, and increase in life expectancy to about 75 years. During this time, health and life expectancy also increased in many poor and colonized countries (World Bank, 1993).

With the introduction of financial deregulation, privatization, and liberalization of global trade in the late 1970s, the general trend in health globally followed two diverging pathways. One was associated with boosted economic growth and the application of ongoing advances in medicine for the benefit of the top 20% of the global population (about 1 billion people) living predominantly, but not exclusively, in wealthy countries. The other was characterized by impediments to economic growth for a poor majority, deterioration in their living conditions, and curtailment of public health services (Birn, 2011; Labonte and Schrecker, 2011)

Relationship between Health and Wealth

The Global Political Economy and Health

While the twentieth century was considered by many to be the most successful in world history, with spectacular achievements in science, technology, and health care it is necessary to focus on the dark side of progress – as reflected in ever-widening disparities in economic and health status (Labonte and Schrecker, 2011), the extent of expenditure on weapons of war (Strategy Page, 2009), and the social construction of systems that abuse human rights (Farmer, 2003; Benatar and Doyal, 2009; Benatar, 2011). In the past 150 years, world population has grown more than sevenfold and human consumption has increased more than 50-fold. The USA, with 5% of the world's population, accounts for about 24% of annual global energy consumption compared to China with 20% of the world's population using 10% of global energy in the early 2000s, rising to 20% by 2012 and predicted to approach 30% by 2020 if China's growth continues at the same pace (Sharma, 2012 cited by Zakaria, 2012). Similarly, within health care, even a nation that spends over $7000 per person per year on health care is unable to provide access for all citizens to modern medical care (Iglehart, 1999).

The world at the beginning of the twenty-first century is thus characterized by widening disparities in economic and health status (within and between countries), and by suffering, conflict, and alienation associated with pervasive social forces (Birn, 2011; Labonte and Schrecker, 2011; Benatar et al., 2011). Erosion of the economy of many poor countries, under the impact of globalization and the 'debt trade', underpinned by neoliberal economic policies, has obstructed the introduction of effective forms of modern medicine into many poor countries and thus prevented achievement of widespread access to even basic health care for billions of people (Rowden, 2009). Third world debt requires special mention. Insistence on repayment is unconscionable when one considers the ways in which such debt has been created, that third world debt is a small component of total world debt and can never be repaid, and that in sub-Saharan Africa four times as much is spent on debt repayment each year than on health and education combined, (Pettifor, 1996; Rudin and Sanders, 2011). The term structural violence is used to

describe how the ideology lying behind such social changes pervasively affects life and health adversely (Farmer, 2003).

Globally, privileged people (20% of the world who have benefitted most from progress since the early 1900s) have become less aware of the social determinants of health, and their focus has shifted to a biomedical and increasingly (expensive) technological approach to health (Daar and Singer, 2011). Health care has become a marketable commodity within a so-called free market system, where everything is calculated, planned, and delivered within an economic and regimented managerial mindset (Callahan and Wassuna, 2006; Relman, 2007). Over time, western biomedicine has in many ways shifted from being a caring, social function, provided through universal access as a social duty to all citizens of equal moral worth, to a profit-driven commercial enterprise run by businesses in the private sector, and poorly managed, cost-containment exercises in overwhelmed public sectors.

The emergence of new diseases, most prominently HIV infection in 1981 and the Ebola virus in 1976 with its recent major epidemic that afflict predominantly those marginalized by poverty (almost 80% of HIV positive persons and almost all infected and dying from Ebola live in some of the poorest countries in the world (Fauci, 2014)), has been attributed to the social and environmental niches created by the nature of the global political economy and its ideology (Lee and Zwi, 1996; Quammen, 2012).

The changing spatial dimensions of human relations resulting from urbanization, migration, travel, ethnic conflict, and displacement of people and refugees, facilitates the spread of such diseases from which no one should feel immune. Failure to appreciate such associations will result not only in failure to control such diseases as HIV/AIDS (and tuberculosis), but more importantly will probably favor the emergence of new infectious diseases in the future (Benatar, 2001).

Modern consumption patterns and health care practices are thus neither universalizable nor sustainable. Their effects on the environment and climate are adverse for both high- and low-income countries and can no longer be ignored. The World Health Organization (WHO) estimates that globally, for the year 2000, there were 150 000 deaths and 5.5 million disability-adjusted life years (DALYs) lost due to climate change (WHO, 2013). Global environmental change is predicted to cause 500 times more loss of healthy life years in low-income African countries than in Europe (McMichael, 2012).

Increasing Inequalities in Wealth and Health

While it is now well established that there is a definite relationship between wealth/poverty and health/disease, these relationships are complex and not linear. For example, one of the wealthiest nations in the world (the USA) has worse health statistics (infant mortality and longevity) than some other highly industrialized countries, and Kerala, a particularly poor state in India, has achieved lower infant mortality rates and greater longevity than many wealthier countries (Labonte and Schrecker, 2011; Birn, 2011). Both absolute and relative wealth affect health. Among the developed countries it is not necessarily the richest societies that have the best health, but rather those with the smallest income differentials between rich and poor (Wilkinson and Pickett, 2009).

The social circumstances influencing women's and children's health begin with the physical, mental, and nutritional states of women during pregnancy and childbirth and continue throughout life (Bhutta and Black, 2013). Care of infants, education and nurturing of children, opportunities for further study during adolescence, and access to rewarding work are all enhanced by maternal literacy, as demonstrated in some relatively poor countries with high levels of female literacy. Access to adequate nutrition, clean water, sanitation, housing, and basic health care is essential.

Poverty (defined as lack of economic resources, lack of education, lack of access to basic life resources such as food water and sanitation, and lack of control over the reproductive process) directly accounts for about 18 million, or one-third of all human deaths each year (Pogge, 2005). Poverty leads to poor health, which in turn aggravates poverty – through the following:

- A bad start in life, for example, undergoing brain or other damage *in utero* through maternal deprivation or abuse, poor antenatal care or inadequate access to safe delivery
- Physical and emotional deprivation during childhood and adolescence
- Inadequate education and lack of exposure to the positive social and environmental forces required to promote effective adolescent maturation
- Inadequate access to the basic living conditions required throughout life for good health (clean water, safe sanitation, adequate nutrition, housing, and clothing)
- Lack of training for work that will allow development of the capacity to become an independent adult
- Lack of a sense of belonging as a valued citizen

All of the above, in different combinations, adversely affect the lives of almost 4 billion people who live below what has been defined as the 'ethical poverty line' of $2.80–3.00 per person per day. These vulnerability-promoting features, while long-understood, have only recently been formally acknowledged by the World Health Organization in its report on the Social Determinants of Health (WHO, 2009). They are largely determined by socioeconomic, politicocultural, and organizational influences sustained by legally enforceable processes that have been locked into place through international legal mechanisms, such as trade laws and tax regulations, since the second half of the twentieth century under the influence of neoliberal economic policies (Rowden, 2009; Gill and Bakker, 2011; Birn, 2011).

Continued use of the terms 'developed' and 'developing' countries, implies a linearity of progress toward a condition that has been achieved by some countries but not others. It is an increasingly inappropriate term, as it misguidedly equates development with economic status in terms of average per capita gross national product (GNP) or the level of industrialization. The more sophisticated model of development (e.g., that used in the United Nations Human Development Report, 1991) emphasizing achievements in health, education, female literacy, gender equality and political liberty, more accurately reflects the real human development required. However, it should be acknowledged that even such development indices are incomplete, as they leave out other important aspects of human flourishing.

On the one hand, high-income, or so-called developed, countries may not have what many would consider very

important aspects of development. For example, they may lack a commitment to reducing the impact of consumption on ecology, concern for the adverse effects of opulent and aggressive lifestyles on billions of people in other countries, and fail to provide universal access to health care.

On the other hand, the achievements made by some so-called developing countries despite their low-economic status is impressive. For example, in Kerala, India the levels of health achieved beyond that expected from economic status alone illustrates the striking impact of improving human capital resources such as female literacy and female empowerment in tandem with a focus on primary health care (Government of Kerala, 2005; Birn et al., 2009).

Sustaining and amplifying such examples is a major challenge, but unlikely to be met without new ways of thinking about health and the global political economy. Such considerations reveal the inadequacy of the terms 'developed' and 'developing' countries, and beg the question of how the idea of sustainable development could be changed toward the more reasonable idea of developing sustainability (Bensimon and Benatar, 2006).

Economic Inequalities

At the beginning of the twentieth century, the wealthiest 20% of the world's population were nine times richer than the poorest 20% of people (BMJ, 1999). This ratio increased progressively – to 30 times by 1960, 60 times in 1990 and to well over 100 times by the early 2000s. The Gini coefficient (a measure of economic inequality) increased from 0.25 to 0.4 in the UK in the past three decades, from 0.38 in 1968 to 0.47 in 2010 in the USA and from 0.6 in 1995 to almost 0.7 in 2009 in South Africa. As of 2011, almost half the world's population lives on less that $1000 per year, and more than a quarter of the world's population lives (on less than $1.5 per day) under conditions of absolute poverty (Bhatt, 2011). In the middle 2000s, approximately 1.75 billion people in the 104 countries covered by the Multidimensional Poverty Index lived in multidimensional poverty that is characterized by at least 30% of the indicators reflecting acute deprivation in health, education, and standard of living (Human Development Report, 2012).

Widening disparities within and between countries in wealth and health during the twentieth century and into the twenty-first century were aggravated by the implementation of neoliberal economic policies in the late 1970s that led into the most severe economic crisis since 1929. As a result, the price of food increased, and an additional 100 million people were thrust into extreme poverty since 2008. As of 2009 data, a billion people live under such conditions (Benatar et al., 2011; Blas, 2009). While tens of millions of people, including many children, die every year of starvation – in a world with enough food to feed all – greater attention is being directed toward providing antiretroviral drugs to all who need these than to saving the lives of those who are starving. The very positive impact of widespread availability of antiretroviral drugs starting in the late 1980s is a testament to the success of the biomedical approach to health that dominates in wealthy countries where the social determinants of health are largely taken for granted. Yet this successful use of antiretroviral treatment eclipses the fact that in low-income countries improved health is generally more dependent on social forces than on the application of modern health care methods.

Health Inequalities

In the early twenty-first century, patterns of disease and longevity vary markedly across the world (Labonte and Schrecker, 2011, **Figure 1**, **Figure 2**). Low-income countries bear over 80% of the global burden of disease in DALYs. This burden is likely to increase as the epidemiological transition progresses, with added disability and suffering from non-communicable diseases such as vascular disease, malignant neoplasms (especially of the lung associated with smoking), neuropsychiatric disease, accidents, and trauma (Lancet, 2012). The lifetime risk of dying from complications of pregnancy and childbirth ranges from 1: 11 000 in Canada to 1: 7 in Niger, and to 1: 76 in the developing world as a whole. Under 5-year-old child, mortality rates range from less than two per 1000 live births in high-income countries to over 170: 1000 in Africa. Life expectancy at birth (LEB) in countries ravaged by HIV/AIDS has fallen to 45 years (and continues to fall) while in high-income countries LEB is over 80 years and rising. It is estimated that in 2010 about 34 million people were living with HIV/AIDS (about 3.5 million of these children), that 1.7 million died of HIV/AIDS that year and about 2.5 million became newly infected each year (Avert, 2012).

Changes in patterns of disease with widening disparities in health and life expectancy are also taking place within high-income countries. Reduction in the frequency of infectious diseases combined with excessive consumption of salty and high-calorie food and less exercise is leading to growing rates of morbidity (sickness/illness) and early deaths from diabetes, stroke and cardiovascular diseases. These features illustrate the degree to which many of the contradictory aspects of global health are part of a wider pattern connected to the inequality arising from the structures and values of modern Western societies (Wilkinson and Pickett, 2009). Changing ratios of young-to-aged people and the health care implications of care for large numbers of old and frail persons, add yet another demographic health challenge for high-income countries in need of more carers, and for low-income countries from which health care workers are actively recruited (Eckenwiler, 2012). In addition, climate change will disproportionately affect the health of the poor (Friel et al., 2011).

Even in high-income countries, improvements in physical health since the mid-twentieth century have not been matched by improvements in mental health. In the late twentieth century, mental illness accounted for 8% of the global burden of disease measured in disability life years (Desjarlais et al., 1995). An updated assessment suggests that mental, neurological and substance abuse disorders account for 17% of the global burden of disease (Collins et al., 2011, *Nature*). The inadequacy of the knowledge base about mental disease (compiled almost exclusively in North America and Europe) to adequately explain and effectively treat psychiatric diseases in other parts of the world poses formidable challenges for progress (Summerfield, 2008)

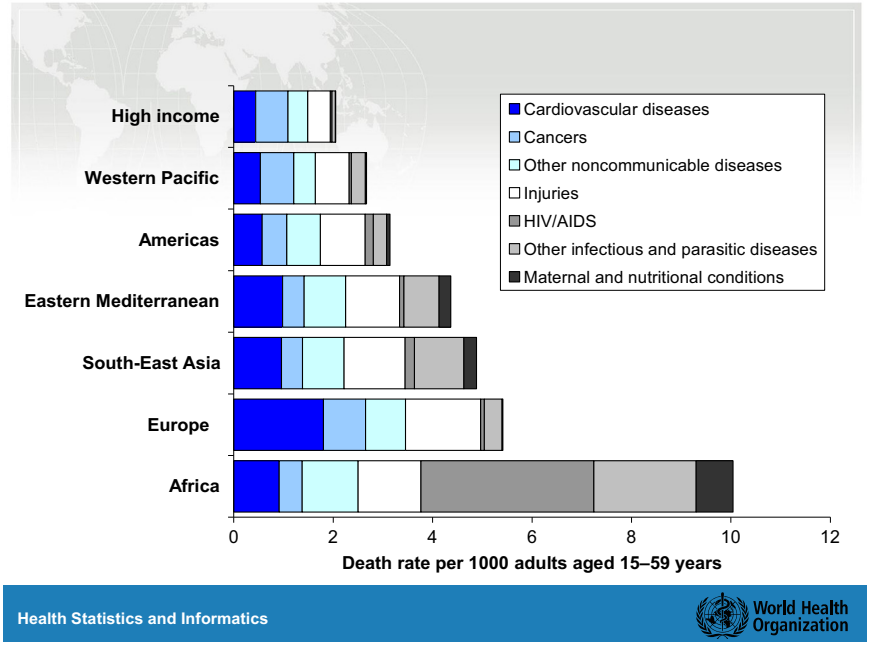

Figure 1 Adult deaths per 1000 adults (15–91) by World Health Organization region. Source: http://www.who.int/entity/healthinfo/global_burden_disease/GBD2004ReportFigures.ppt.

Policy Issues

Militarization and the Social Construction of Human Rights Abuses

Annual global defense spending fell to less than $1 trillion after the Cold War ended in 1991, but has grown by almost 50% (to over $1.3 trillion) in the past decade (about 2.5% of global Gross domestic product), with the greatest increase in expenditures (up 62% in the last decade) in the Middle East (Strategy Page, 2009). Industrialized countries spend on average 5.3% of their GNP on the military but only 0.3% on economic aid to developing countries (Sivard, 1996), well below the target of 0.7% of Gross National Income (GNI) of the Development Assistance Committee countries (Zwi, 2011). The arms trade, with its link to economic aid, has contributed significantly to the escalating number of wars, conflict, and

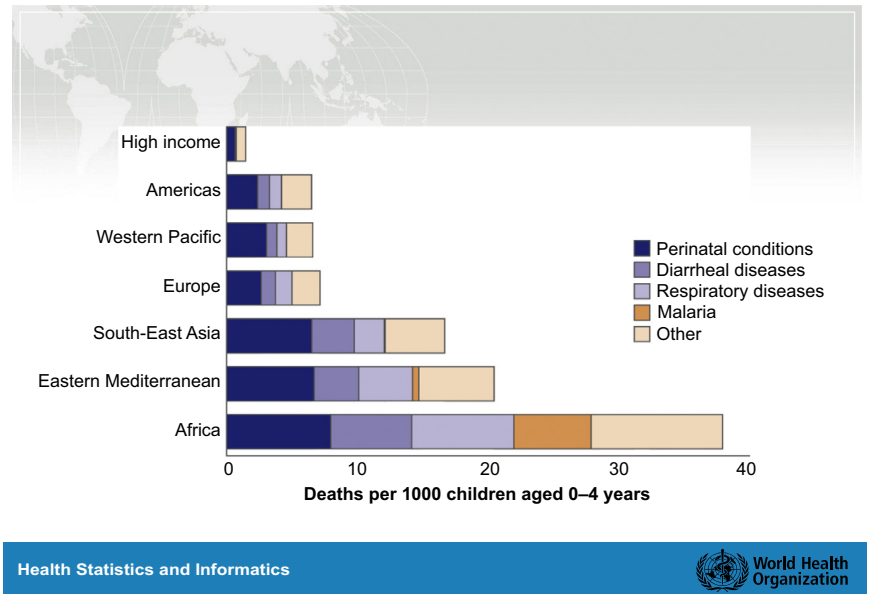

Figure 2 Child mortality rates by cause and region 2004. Source: http://www.who.int/entity/healthinfo/global_burden_disease/GBD2004ReportFigures.ppt.

widespread torture since the Second World War, and it has displaced over 25 million people. The war in the Democratic Republic of Congo that started in 1993 resulted in 3.3 million deaths of which about 200 000 were directly through violence, and the remainder from the indirect effects of war on the country (Mahmudi-Azer, 2011). Some years ago, it was estimated that the annual budget of the WHO was equivalent to 2.5 h of global military expenditure.

Health Care Expenditure and Health Outcomes

Since the 1960s, major advances in medicine and technology have been associated with escalating expenditure on their application in health care – most of this in highly industrialized countries. In recent decades about 90% of annual world expenditure on health care was spent on about 16% of the world's population who bear approximately 7% of the global burden of disease (in DALYs) and less than 10% of medical research expenditure was on health problems of the developing world (Iglehart, 1999). In the first decade of the twenty-first century, annual per capita expenditure on health care ranged from more than $8000 in the US to less than $15 in the poorest countries. The poorer WHO regions like the Africa Region and the South-East Asian Region comprising 38% of the world's population, account for the largest share of the global burden of disease (over 50% of global DALYs lost), and spend only 2.5% of global health resources (WHO, 2011).

These are examples of global injustice that we should surely find intolerable if we have a real commitment to universal human rights and human dignity. Vaccine development programs and drug donation programs for poor countries, admirably promoted and supported by pharmaceutical companies, are necessary but insufficient responses to such injustices. However, modest increments in economic status, combined with provision of a basic health care package, have the potential to improve health in poor countries significantly. Economist Jeffrey Sachs has estimated that a tax of one cent on every $10 earned applied to the wealthiest 1% of the world's population would generate the $35 billion required to increase annual per capita expenditure on health from $15 to $50 per year for the poorest billion people in the world (Sachs, 2009).

Even though some parasitic diseases (for example, onchocerciasis, schistosomasis, and lymphatic filariasis) could be controlled by mass vertical treatment campaigns using inexpensive drugs, the infrastructure required to provide such coverage is inadequate. Moreover, effective vertically organized services for these specific diseases and others like HIV/AIDS can pose the danger of impairing the development of sustainable, horizontally organized and well integrated health delivery systems. Effective treatment of diseases such as tuberculosis, leprosy, and HIV/AIDS requires more complex infrastructures than can be afforded on current health care budgets in poor countries. However, the achievement of improved health is not merely a resource issue even in rich countries – as exemplified by comparative health statistics around the world (Wilkinson and Pickett, 2009; Labonte and Schrecker, 2011; Simpson, 2012), and by formal acknowledgment of the greater impact of well organized health services (McKee, 2011) and other social and societal forces on health (Birn, 2011).

Access to Health Care, and the Nature of Health Care Systems

Many countries consider access to at least a minimum package of health care as a basic human right that nation-states should be committed to providing for their citizens. Some form of equitable access to health care is provided in all Western European nations and in Canada. Medical care, delivered as a marketable commodity, albeit with considerable state assistance for the poor and the aged, is the example that has been set by some of the wealthiest nations in the world. However, many citizens have inadequate access to health care. While recognizing that private/public alliances cannot be avoided and indeed may be beneficial, Buchanan has provided cogent arguments for the nine elements he considers essential for a just health care system. He argues further that achieving a just health care system as a cooperative between the public and private sectors requires an organization, such as the state, to act as a coordinator and guarantor of equitable access to care and fair distribution of costs. A balance between private and public medicine with appropriate coordination could overcome the deficiencies of both systems (Buchanan, 1995).

The effects of intensive privatization, aided and abetted by the structural adjustment programs promoted by the International Monetary Fund and the World Bank, have been much more marked in low-income countries and have had severe adverse effects on health and access to health care. Acknowledging the need for, and the right to, universal access to a basic health care package and achieving this goal pose challenges for the future. The WHO's renewal strategy for health-for-all places emphasis on equity, solidarity, and appropriate technical, political, and economic strategies that could promote health and sustainable health care systems as central to development. Here too an appropriate balance needs to be struck between public and privately funded medicine. How this balance can be struck and how health can be coordinated within an effective strategic alliance between many parties remain hotly debated subjects. The example of making antiretroviral drugs available to millions of people in poorer countries is encouraging (UNAIDS, 2011).

Traditional Healers

Despite the pervasiveness of Western influences, traditional healers have survived and play an increasing role in some countries. For example, while there are about 25 000 Western trained medical practitioners in South Africa, and Western medicine has a powerful foothold in this middle-income country, over 200 000 traditional healers carry out a thriving trade. In part, this may be attributed to the emphasis in traditional medicine on the disruptive societal behavior of individuals as an explanatory causal factor in disease. Increasing contact and collaboration between Western medicine and traditional healers as well as other complementary and alternative healers is to some extent being encouraged. It has been recommended that the potential impact of their contributions to health in the broadest sense (Pretorius, 2012), especially in association with a primary health care approach should be brought into an ongoing pluralistic health care system rather than being ignored (Engelbrecht and Van Rensbug, 2012).

Prospects for the Future

The still unfolding global economic crisis (Benatar et al., 2011) together with the very ominous implications of climate change (Oreskes and Conway, 2013) are stark reminders of the unpredictability of progress and the threats to much that is valued in Western civilization. Many hope that exponential increments in genetic information and other new biotechnological advances could be applied to medical practice and improve health. However, we need to seriously contemplate whether faith in such new power is exaggerated and, even if realized whether it will be used for the benefit of all humankind or whether it will be reserved for a privileged minority.

Advances in human health globally in the twenty-first century and achieving real improvement in the lives of all people throughout the world will be dependent on acknowledging many adverse features of modern life. Acquiring a deeper understanding of the forces promoting health and diseases in developing countries needs to be coupled to a commitment to address these forces and how they could be reshaped to achieve health and human flourishing on a wider scale (Gill and Bakker, 2011; Gill, 2012; Schweickart, 2012).

New paradigms of thinking and consequent innovative policies and action are also required to promote widespread human flourishing. This will entail development of new social structures and using available resources more effectively. Within health care reducing some of the massive wastage associated with delivery of health care has been shown to be very effective (Drumm, 2012). New approaches will also need to be applied to many other activities in modern life. In brief, there is a need to find ways of doing better with less! (Benatar, 2013)

See also: African Studies: Health; Health Economics; Infant and Child Mortality in the Less Developed World.

Bibliography

Avert. Worldwide HIV and AIDS statistics. http://www.avert.org/worldstats.htm (accessed 18.10.12.).

Benatar, S., Brock, G., 2011. Global Health and Global Health Ethics. Cambridge University Press, Cambridge.

Benatar, S.R., 1998. Global disparities in health and human rights. American Journal of Public Health 88, 295–300.

Benatar, S.R., 2011. Global Health and Human Rights: Working with the 20th Century Legacy. Annual Human Rights Lecture. University of Alberta. http://www.uofaweb.ualberta.ca/humanrightslecture/ (accessed 31.10.12.).

Benatar, S.R., 2013. Needs, obligations and international relations for global health in the 21st Century. In: Coggon, J., Gola, S. (Eds.), Global Health and International Community. Bloomsbury, London and New York, pp. 63–80.

Benatar, S.R., Doyal, L., 2009. Human rights abuses: balancing two perspectives. International Journal of Health Services 39 (1), 139–159.

Benatar, S.R., Gill, S., Bakker, I., 2009. Making progress in global health: the need for a new paradigm. International Affairs 85 (2), 347–371.

Benatar, S.R., Gill, S., Bakker, I., 2011. Global health and the global economic crisis. American Journal of Public Health 101, 646–653 (see reference 62).

Bensimon, C.A., Benatar, S.R., 2006. Developing sustainability: a new metaphor for progress. Theoretical Medicine and Bioethics 27 (1), 59–79.

Bhatt, K., Sunday, 29 May 2011. Feature Interview. Thomas Pogge on the Past, Present and Future of Global Poverty Truthout. Available at: http://truth-out.org/news/item/792:thomas-pogge-on-the-past-present-and-future-of-global-poverty (accessed 06.12.13.).

Bhutta, Z.A., Black, R.E., 2013. Global maternal, newborn, and child health: so near and yet so far. The New England Journal of Medicine 369, 2226–2235.

Birn, A.-E., 2011. Addressing the societal determinants of health: the key global health ethics imperative. In: Benatar, S., Brock, G. (Eds.), Global Health and Global Health Ethics. Cambridge University Press, Cambridge, pp. 37–52.

Birn, A.-E., Pillay, Y., Holtz, T.H., 2009. Textbook of International Health, third ed. Oxford University Press, New York.

Blas, J., 6 April 2009. Relentless Tide of Global Hunger Engulfs 1bn. Financial Times. Available at: http://www.ft.com/cms/s/0/d520cf02-22e0-11de-9c99-00144feabdc0.html#axzz1Aa8kozJ7. accessed January 10th December 2013.

BMJ, 10 April 1999. The champagne glass of world poverty. BMJ 318 (7189), 0.

Buchanan, A., 1995. Privatization and just health care. Bioethics 9, 220–239.

Callahan, D., Wassuna, A., 2006. Medicine and the Market. Johns Hopkins University Press, Cambridge, Baltimore.

Collins, P.Y., Patel, V., Joestl, S.S., March, D., Insel, T.R., Daar, A.S., 2011. Grand challenges in global mental health. Nature 473, 27–30.

Daar, A.A., Singer, P.A., 2011. The Grandest Challenge. Random House, Canada.

Desjarlais, R., Eisenberg, L., Good, B., Kleinman, A., 1995. World Mental Health: Problems and Priorities in Low-Income Countries. Oxford University Press, New York.

Drumm, B., 2011. The challenge of change: putting patients before providers. Orpen Press, Dublin.

Eckenwiler, L., 2012. Long-term Care, Globalization, and Justice. Johns Hopkins University Press, Baltimore, MD.

Engelbrecht, M.C., Van Rensburg, H.C.J., 2012. Primary health care: nature and state in South Africa. In: Van Rensburg, H.C.J. (Ed.), Health and Health Care in South Africa, second ed. Van Schaik, Pretoria, pp. 483–534.

Falk, R., 1999. Predatory Globalization: A Critique. Polity Press, Cambridge, UK.

Farmer, P., 2003. Pathologies of Power. University of California Press, Berkeley, CA.

Fauci, A.S., 2014. Ebola: Underscoring the Global Disparities in Health Care Resources. New England Journal of Medicine 371 (12), 1084–1086.

Friel, S., Butler, C., McMichael, A., 2011. Climate change and health risks and inequities. In: Benatar, S., Brock, G. (Eds.), Global Health and Global Health Ethics. Cambridge University Press, Cambridge, pp. 198–219.

Gill, S. (Ed.), 2012. Global Crises and the Crisis of Global Leadership. Cambridge University Press, Cambridge.

Gill, S., Bakker, I.C., 2011. In: Benatar, S., Brock, G. (Eds.), The Global Crisis and Global Health. Cambridge University Press, Cambridge, pp. 221–228.

Government of Kerala, 2005. Human Development Report 2005 (Kerala).

Human Development Report, 1991. United Nations. http://hdr.undp.org/en/reports/global/hdr1991/ (accessed 31.10.12.).

Human Development Report, 2012. United Nations. http://hdr.undp.org/en/media/HDR_2010_EN_Complete_reprint.pdf (accessed 10.12.13.).

Iglehart, J., 1999. The American health care system: expenditures. New England Journal of Medicine 340, 70–76.

Labonte, R., Schrecker, S., 2011. The state of global health in a radically unequal world: patterns and prospects. In: Benatar, S., Brock, G. (Eds.), Global Health and Global Health Ethics. Cambridge University Press, Cambridge, pp. 24–36.

Lancet, 2012. The Global Burden of Disease Study 2010. Available at: http://www.thelancet.com/themed/global-burden-of-disease (accessed 07.12.13.).

Lee, K., Zwi, A.B., 1996. A global political economy approach to AIDS. New Political Economy 1, 355–373.

Mahmudi-Azer, S., 2011. In: Benatar, S., Brock, G. (Eds.), The International Arms Trade and Global Health. Cambridge University Press, Cambridge, pp. 166–172.

McKee, M., 2011. Health systems and health. In: Benatar, S., Brock, G. (Eds.), Health and Global Health Ethics. Cambridge University Press, New York, pp. 63–73.

McKeown, T., 1976. The Role of Medicine: Dream, Mirage, or Nemesis? Nuffield Provincial Hospitals Trust, London, England.

McKeown, T., Record, R.G., 1962. Reasons for the decline of mortality in England and Wales during the nineteenth century. Population Studies 16, 94–122.

McMichael, T., 2012. Health risks, present and future, from global climate change. British Medical Journal 344, e1359.

Oreskes, N., Conway, E.M., 2013. The collapse of western civilization: a view from the future. Daedalus 142 (1), 40–58.

Pettifor, A., 1996. Debt: The Most Potent Form of Slavery. Christian Aid, London (Debt Crisis Network).

Pogge, T., 2005. World poverty and human rights. Ethics & International Affairs 19 (1) (Spring) Available at: http://www.carnegiecouncil.org/publications/journal/19_1/symposium/5109.html.

Pretorius, E., 2012. Complementary and alternative medicine and traditional health care in South Africa. In: Van Rensburg, H.C.J. (Ed.), Health and Health Care in South Africa, second ed. Van Schaik, Pretoria, pp. 593–652.

Quammen, D., 2012. Spillover: Animal Infections and the Next Human Pandemic. W. W Norton & Co, New York.
Relman, A.R., 2007. A Second Opinion: Rescuing America's Health Care. MA Perseus Books, Cambridge.
Rowden, R., 2009. The Deadly Ideas of Neoliberalism: How the IMF Undermined Public Health & the Fight against AIDS. Zed Books, London and New York.
Rudin, J., Sanders, D.D., 2011. Structural adjustment and health. In: Benatar, S., Brock, G. (Eds.), Global Health and Global Health Ethics. Cambridge University Press, Cambridge, UK, pp. 155–165.
Sachs, J., October 2009. A statement made by Jeffrey Sachs in a video conference presentation to the Canadian International Health Conference in Ottawa.
Schweickart, D., 2012. After Capitalism. Rowman & Littlefield, Lanham, MD.
Sharma, R., 2012. Breakout Nations: In Pursuit of the Next Economic Miracles. W. W. Norton, London.
Simpson, J., 2012. Chronic Condition: Why Canada's health care system needs to be dragged into the 21st century. Allen Lane Publishing, Toronto.
Sivard, R.L., 1996. World Military and Social Expenditures, sixteenth ed. WMSE, Leesburg, VA.
Strategy Page, 28 April 2009. Doing More with Less. http://www.strategypage.com/htmw/htlead/20090428.aspx (accessed 18.11.12.).
Summerfield, D., 2008. How scientifically valid is the knowledge base of global mental health? British Medical Journal 336, 992–994.
UNAIDS, 2011. Nearly 50% of people who are eligible for antiretroviral therapy now have access to lifesaving treatment. http://www.unaids.org/en/resources/presscentre/pressreleaseandstatementarchive/2011/november/20111121wad2011report/.
Van Rensburg, H.C.J., Benatar, S.R., 2001. History of medicine in Africa. In: Last, J., Lock, S. (Eds.), The Oxford Illustrated Companion to Medicine. Oxford University Press, Oxford.
WHO, 2013. Climate change and Health. Fact Sheet number 266. Available at: http://www.who.int/mediacentre/factsheets/fs266/en/ (accessed 06.12.13.).
WHO, 2011. Global Health Expenditure Atlas. World Health Organization, Geneva. Available at: http://www.who.int/nha/atlas.pdf (accessed 06.12.13).
Wilkinson, R.G., 1996. Unhealthy Societies: The Afflictions of Inequality. Routledge, London.
Wilkinson, R.G., Pickett, K., 2009. The Spirit Level: Why More Equal Societies Almost Always Do Better. Allen Lane, London.
World Bank, 1993. World Development Report 1993: Investing in Health. Oxford University Press, New York.
World Health Organisation, 2009. Closing the Gap in a Generation. Health Equity through Action on the Social Determinants of Health. http://www.who.int/social_determinants/thecommission/finalreport/en/index.html (accessed 31.10.12.).
World Health Organization, 1995. Poverty and Health in Developing Countries. No. 16 in the WHO Series Macroeconomics, Health and Development. World Health Organization, Geneva.
World Health Report, 1992. Data citied in Climate change and Human Health. Available at: http://www.who.int/globalchange/summary/en/index6.html (accessed 05.12.13.).
Zakaria, F., 23 May 2012. China's Economic Crisis. The Washington Post.
Zwi, A., 2011. International aid and global health. In: Benatar, S., Brock, G. (Eds.), Global health and global health ethics, first ed. Cambridge University Press, Cambridge, pp. 184–197.

Relevant Websites

Global Burden of Disease, 2004. Update: Selected Figures and Tables. http://www.who.int/entity/healthinfo/global_burden_disease/GBD2004ReportFigures.ppt (accessed 18.11.12.).
Social Determinants of Health. Commission on Social Determinants of Health – Final Report World Health Organisation. http://www.who.int/social_determinants/thecommission/finalreport/en/index.html (accessed 31.10.12.).

Health Insurance: Economic and Risk Aspects

Conchita D'Ambrosio, Université du Luxembourg, Walferdange, Luxembourg
Simone Ghislandi, Università Bocconi, Milan, Italy

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C.G. McLaughlin, M.E. Chernew, volume 10, pp. 6570–6576, © 2001, Elsevier Ltd.

Abstract

Insurance can be defined as a contract where the subscriber makes regular payments to the insurer in exchange of the promise of indemnity against specified risks. In health insurance, the specific risks covered are the financial costs of the treatments needed after an illness or an accident. The economics of health insurance is well developed and identifies a series of challenges, which can seriously jeopardize the proper functioning of health insurance markets based on private contracts. It is indeed rare to observe health systems that do not rely on some degree of government involvement. In this article, the theoretical and empirical challenges to health insurance markets will be described and discussed. Subsequently, we will provide a short description of the characteristics of the most significant health systems around the world, linking these with the issues emerging from the economic analysis.

Introduction

Insurance can be defined as a contract where the subscriber makes regular payments to the insurer in exchange of the promise of indemnity against specified risks. In health insurance, the specific risks covered are the financial costs of the treatments needed after an illness or an accident. The health insurance contract can be signed by either the individual subscriber or a sponsor, for example, the employer or a community organization. The insurer can be a private for-profit organization, a private not-for-profit organization, or the government (Claxton, 2002).

A health insurance agreement usually specifies the value and the form of the following:

- Premium: the monthly or annual regular payment to be made by the subscriber;
- Coverage: the type and the limits of the medical expenses that can be reimbursed;
- Deductibles: the limit below which medical expenses are not reimbursed. The insurance pays the difference between the medical bill and the deductible;
- Copayment or Coinsurance: the out-of-pocket payment to access an insured medical service. In copayments, this amount is fixed; in coinsurance, it is a percentage of the medical bill; and
- Network providers: the list of accredited providers (for example, hospitals, clinics, or general practitioners) selected by the insurer. In order to get coverage, medical services must be provided by the institutions included in the network.

The economics of health insurance is well developed and identifies a series of challenges, which can seriously jeopardize the proper functioning of health insurance markets based on private contracts. It is indeed rare to observe health systems that do not rely on some degree of government involvement. In the following, the theoretical and empirical challenges to health insurance markets will be described and discussed. Subsequently, we will provide a short description of the characteristics of the most significant health systems around the world, linking these with the issues emerging from the economic analysis.

The Market for Health Insurance: The Basic Framework

Why do people buy health insurance? When it comes to important issues like health and health-related expenditure, people are mainly risk averse (Arrow, 1970). Formally, a rational individual is risk averse if she prefers to gain X for sure rather than participating in a lottery whose expected value is X. In other words, people tend not to like risks and are willing to give up part of a certain income in order to avoid it.

The definition of risk aversion applies directly to the demand for health insurance. Imagining the future health situation of a subject as a lottery, health insurance requires that a person should be willing to give up part of her own income (insurance premium) in the present in exchange of a (partial) insulation from the risk of getting ill and incur monetary, physical, and psychological losses in the future. This can be a chosen option only if the person is risk averse.

In the baseline economic model of health insurance markets, risk averse individuals are assumed to be perfectly informed about the parameters driving their own choice. They know about their probability of getting sick and they know what losses they will incur in that case. Indeed, the assumption of perfect information of the insurance subscriber is common in many studies (Phelps, 2010). Similarly, the baseline model implicitly assumes that risks are correctly perceived. Not only do people have information about the risks, but they also interpret it correctly. In other words, in the baseline framework, problems of too much or too little confidence or any other source of misperception are ruled out a priori.

Although rational individuals choose to insure against health risks, market outcomes and equilibria depend also on the supply side. Economic theory has clarified that three main assumptions are needed for the provision of insurance through private markets to be efficient. First, administrative costs or any other cost outside the agreed reimbursements (these costs are also known as *loading factors*) should be small. Second, firms should be risk neutral. Last but not least, firms should compete on premiums (Einav and Finkelstein, 2011). This in turn assures that premiums are *actuarially fair*, i.e., they reflect the actual risk profiles of the insured subject. In a world where all

these three assumptions hold true, firms cannot charge unnecessarily high (unfair) premiums, because other firms would undercut them, gaining market shares and profits. At the same time, even in competitive markets, too many loading factors might force firms to charge unfair premiums, which could turn out to be too high for many potential subscribers.

The optimal insurance coverage is what emerges from a bargaining process where all the above-mentioned assumptions hold true (Phelps, 2010). It is thus possible to show that risk averse people, perfectly informed about their private risks, will choose *full coverage* (i.e., a reimbursement of all the medical expenses they will incur if they get ill) in exchange of the *fair premium* offered by perfectly informed firms working in a competitive environment. In this baseline model, markets are efficient and people get their fair deal.

Health Coverage and Insurance Markets

The baseline model is a useful benchmark. It provides a precise definition of the assumptions needed for health insurance markets to be efficient. However, it is realistic to expect that in many real world situations some of these assumptions would fail. Since the 1970s, the economic literature has gained substantial understanding on the functioning of health insurance markets when more realistic frameworks are used.

This literature has become very influential in the last decade and has set the basis for many policy interventions in the field of health systems design and health insurance regulation. In the following, the most important issues emerging from this literature are briefly explained.

Preexisting Conditions and Risk Selection

One of the most pervasive problems related to a voluntary market for health insurance comes from the simple observation that the premium should reflect the individual risk. Age-related illnesses or chronic diseases (preexisting conditions) can become very difficult to cover in an insurance market, since the premium required to insure against an almost sure event (i.e., needing some sort of health treatment) is too high. This situation can create two related types of problems (Cutler and Zeckhauser, 2000).

First, the market can fail altogether. In the most extreme cases, the probability of getting ill is so close to one that the fair premium is equal to the reimbursement. This means that the insurance contract does not exist. In practice, this failure translates into the exclusion of people with preexisting conditions from the market for health insurance (see also Section Supply Side below). In this sense, some have suggested that the choice of insurance is mainly intertemporal and that people should buy health for their lifetime (always *before* a health condition appears) instead of annually (Cochrane, 1995). However, these types of contract are not easily found in real world markets. Second, it should be noted that the insurance choice is always limited by resources availability. A poor is less likely to buy health insurance and access health care than a person with high income (e.g., Clark et al., 2013; Hsiao, 1995; Preker et al., 2007). Moreover, budget constraints are more likely to 'bite,' excluding from the market the high-risk people, which are also poor.

It is important to stress that the high risks are the people that most need access to health care and health treatments. Hence, the system fails exactly those it is supposed to serve. The problems raised by preexisting conditions and aging in health insurance markets are then particularly important and are well-recognized in most health systems around the world (see also Section Health Insurance around the Globe on this).

Asymmetric Information

The baseline model assumes that agents have perfect information about virtually everything. However, in reality one party can have more information than the other, i.e., information can be asymmetric. The literature on asymmetric information, started in the 1970s, has developed a great body of theoretical and empirical evidence. Here we define the most common typologies of asymmetric information problems commonly associated to health insurance markets.

Adverse and Advantageous Selection

In health insurance, adverse selection occurs whenever the asymmetric distribution of information causes markets to screen out the good risks, selecting only the 'bad' ones. Adverse selection is a phenomenon intrinsic to any insurance system. It can happen mainly for two reasons: the insured knows something that the insurance does not or they both have the same information but the insurance company is not allowed to use it in setting the premium, due to antidiscrimination laws or to regulatory limits to the rating system.

The problem of adverse selection was firstly identified in markets with asymmetric information on products' quality (Akerlof, 1970). Subsequently, the analysis was extended to the health sector in a very influential article that set up the basic formalization of the issue (Rothschild and Stiglitz, 1976), which proceed as follows: imagine that there are now two types of persons, one high and one low risk, but that the insurer cannot distinguish one from the other. For example, subscribers can be either smokers or nonsmokers, with the insurer unable to observe their true type. In this case, the problem is that if the insurer offers two different contracts, one for the low risk (nonsmokers) and the other for the high risks (smokers), the high risk can pretend to be low risk and get a better deal. Since the insurer cannot tell one from the other, this situation is not sustainable.

One possible solution could be to offer some form of average premium to everyone (pooled premium). However, for low risks this premium is unfair. They are thus less likely to accept the deal and some of them will exit the market. Since the remaining pool will include relatively more high risks than before, the insurance will be forced to increase further the average premium in order to balance the exits. This will force some other low risk types to turn down the offer. The process could potentially go on until only high risks remain in the market, i.e., adverse selection. This phenomenon is also known as the 'death spiral' in health insurance markets (Cutler and Zeckhauser, 1998).

Under certain conditions, the failure of pooling contracts could be overcome by offering contracts through which buyers

self-select themselves (i.e., separating equilibrium). That is, the insurer can offer a combination of premium-coverage that is the best option only for the high risk and another that is optimal only for the low risk. By picking one option or the other, subscribers reveal their true risk profile. This could potentially represent an equilibrium. It turns out that these types of contract imply offering to the low risks less than the full coverage in exchange of a fair premium (Rothschild and Stiglitz, 1976). That is, low risk profiles agree to give up part of the coverage in exchange of a fair premium, while the high risks prefer to pay a higher premium in order to be fully reimbursed. With respect to the full information scenario, the low risks would be worse off, since they pay the same but they get a lower coverage. This is the loss associated to adverse selection when the market does not fail. In addition, however, it is also possible that if the proportion of low risk buyers is higher than a certain threshold, no separating equilibrium exists either, so that only the high-risk profiles remain in the market.

The existing empirical evidence on adverse selection is contradictory (Cohen and Siegelman, 2010). According to the method and the data analyzed, some studies find a positive relation between risk profiles and insurance coverage (Sapelli and Vial, 2003; Spenkuch, 2012), others do not observe such a correlation (Cardon and Hendel, 2001; Fang et al., 2008; Finkelstein and McGarry, 2006). Evidence of separating equilibria through underprovision of insurance to low risks has also been found (e.g., Buchmueller and DiNardo, 2002).

There are different responses to the challenges posed by adverse selection. Two are worth mentioning here. First, if the insurance company can acquire the missing information after the claim has been filled, then an *ex-post* control can be imposed and defined in the contract. In this case, however, the company must be allowed to use the information about the individual claimant. This applies, for example, to smoking. Second, the regulator can introduce an individual mandate on health insurance. This is the approach taken by many countries such as the Netherlands or Switzerland and is one of the main pillars of the Obama health care reform known as the Affordable Care Act (ACA) in the US (US Congress, 2010). In this case, since all the residents are obliged to subscribe to some form of insurance, good risks are automatically included into the system and average premiums might reduce (e.g., Hackmann et al., 2013).

In recent years a literature has emerged pointing out an opposite phenomenon to adverse selection: advantageous selection (De Meza and Webb, 2001; Hemenway, 1990). Advantageous selection arises whenever those who are more risk averse both buy more insurance coverage and have lower risks. In other words, when cautious and risk averse individuals are also more likely to sign up for insurance. Advantageous selection shifts the attention from the actual individual risk profiles to their risk aversion, which is obviously not known to the insurer. From the point of view of the insurance company, however, this type of selection is profitable. Although the empirical evidence so far seems to confirm advantageous selection in the US, for example, in long-term care insurance (Finkelstein and McGarry, 2006) or in the Medigap (Fang et al., 2008), it is important to note that adverse and advantageous selection can, and probably do, coexist in these markets (Einav and Finkelstein, 2011).

Moral Hazard

Moral hazard occurs whenever people take more risks or demand more health care just because they signed up for health insurance (Arrow, 1963; Folland et al., 2006). Broadly speaking, moral hazard can arise before (ex-ante) or after (ex-post) a claim is filed. In the former case, people would reduce their preventive effort (for example, paying less attention or doing less physical activity) because they feel 'safe,' i.e., insulated from the losses they would face in case the negative event materialized. In the ex-post moral hazard, the claimants would increase their demand for health care services beyond what they would be willing to buy if they were not reimbursed. In other words, they demand excessive medical services just because they are (partially) free. Note that moral hazard is based on the assumption that the insurer cannot observe the insured's actions. This is why it is often referred to as *hidden action*.

Ex-post moral hazard is considered a serious problem for voluntary private insurance (Pauly, 1968; Zeckhauser, 1970). Indeed, when confronted with an excessive demand of reimbursements, insurance companies might raise their premiums above what is considered fair by some subscribers. Government-based health systems also suffer ex-post moral hazard, often in the form of too high demand for pharmaceuticals and medical diagnostics. Although ex-post moral hazard is a pervasive phenomenon in health insurance markets, solutions do exist and are widely applied. Since the problem originates in the free access to health care services, the solution is simply to make the patient share part of the medical costs. This is usually done through deductibles, copayments, and coinsurance, as defined above (Folland et al., 2006).

Empirically, moral hazard has been tested in one of the first large-scale randomized study in social science, the RAND Health Insurance Experiment (Brook et al., 1983; Manning et al., 1987). Between 1974 and 1982 a total of 5809 people were randomly assigned to different insurance plans: free health care, individual deductible plan (family pays 95% of costs under $150), a series of intermediate coinsurance plans (coinsurance of either 25, 50, or 95% until the threshold limit of $1000 per family). These subjects were then followed for the whole observation period. Results showed that free-care insured were more likely to go to the doctor and to be hospitalized. At the same time, health consequences were negligible. This can be seen as a strong support for the theory of moral hazard.

These results have been very influential throughout the 1990s in the health policy debate. Serious criticisms directed toward both the method and the interpretation of the results have, however, been advanced in the literature. Two such criticisms are of particular importance.

First of all, what might look like opportunistic behavior (demanding more medical care when it is freely accessible) could actually be the consequence of the income transfer from the healthy to the sick implicit in any health insurance mechanism (Nyman, 2003, 2001, 1999). In other words, one person might ask for new medical treatments not because medical care is free or cheaper, but simply because without insurance the same treatment would not be affordable. Indeed, increasing accessibility by making treatment affordable should be one of the main purposes of health insurance. Also, medical treatments are always related to personal costs that are increasing

with the complexity and the invasiveness of the intervention. In these cases, people would not ask for medical treatments if they did not actually need them. It is thus important to distinguish between different types of services. For important and expensive treatments, an increase in usage following insurance coverage might signal an improvement in accessibility to health care rather than opportunistic behavior.

A second criticism is related to health care access for preventive reasons (Ellis and Manning, 2007; Seog, 2012). Consistent with the evidence, higher copayments imply lower demand for diagnostic and preventive medical services. This means that people might refrain from seeing the doctor or run diagnostic tests if the cost is too high. Studies generally focus on the savings for the insurance from this contraction in demand and interpret this phenomenon as a reduction of moral hazard. However, for patients who have a condition, this would represent just a postponement of treatment. In general, high copayments might inhibit early detection of illness. Since typically it is cheaper and safer to treat conditions earlier rather than later, what might appear as a saving might turn out to be an increase in the overall expenditure. For example, analyzing 172 Medicare plans in the US, 18 of which raised copayments for ambulatory care during the observation period, a study found an average reduction (between 2001 and 2006) in the number of ambulatory visits for the group whose copayment increased (Trivedi et al., 2010). This is consistent with moral hazard. On the other hand, for the same group the hospital admissions increased by 2.2%. Overall, although copayment saved $7150 per capita in reduced demand for ambulatory care, it increased by $24 000 the expenditure for inpatient care.

Overall, moral hazard is unavoidable in health insurance. It affects both private and public systems. In theory, the problem can be ameliorated by using deductibles and copayments. However, caution should be taken in interpreting any 'excess' demand as a sign of opportunistic behavior, possibly avoiding overreactions and extreme policy implications.

Supply Side

The main assumption for the supply side in the baseline model is that firms compete on prices. In other words, they try to get people to subscribe to their policies by offering the best price for any given coverage. This assumption is particularly problematic in reality and often requires some sort of government intervention at various levels.

The first issue to be considered is risk selection, also known in the economic literature as *cream skimming*: the selection by the insurance providers of those consumers expected to be more profitable given their risk profiles (Barros, 2003; Ellis, 1998). Within the same premium-risk group, profitability is inversely related to the risk profile of the subscribers: good risks are profitable and firms will try to attract them. Even when they are not allowed to directly deny access to insurance to the bad risks, for example, because they are obliged by law to accept subscribers independently from their risk profile, they still can set up contracts and offer services that are typically more appealing for the good risk profiles. Competition in this case can be problematic to manage. One solution is risk adjustment (Shen and Ellis, 2002; Van de ven and Ellis, 2000). Information can be used to subsidize firms for each high risk in the pool of the enrollees. Hence, pools with proportionally more high risks will receive more subsidies. Subsidization can come from either the government or other firms with advantageous risk profiles among their pool. If correctly set, these subsidies can significantly reduce the incentive for risk selection.

A second issue arising from competitive health insurance markets is related to the presence of administrative costs and other loading factors. In these cases, firms would need to increase premiums above the fair level in order to recover the additional costs. This would alter significantly the picture emerging from the baseline model (Einav and Finkelstein, 2011). Administrative costs in competitive markets imply a duplication of resources and procedures, which reduces the efficiency of a free market health insurance system.

Competition might even be difficult to achieve. The literature in industrial economics is full of examples of markets where firms either do not compete with each other because they can collude or do not find it profitable to start price wars (Martin, 1994; Tirole, 1988). Much depends on how easily new and small firms could enter the market and gain market shares by offering low premiums. For example, the presence of switching costs, i.e., fixed and psychological costs people have to pay when they change insurer, do indeed represent an obstacle for competition in health insurance markets (Farrell and Klemperer, 2007; Frank and Lamiraud, 2009).

Health Insurance around the Globe

Both the theory and the empirical observations have clarified that universal health coverage, i.e., universal access to health care, cannot be achieved through market mechanisms only. Information and economic problems are too pervasive to the system. Government interventions, whether in the form of regulation, redistribution, or direct provision of insurance and health care, characterize the organization of health systems everywhere around the world.

Countries generally have developed different and complex systems to provide health care to their citizens. Health care systems are often classified in many different ways depending on the types and numbers of factors researchers might find relevant to consider (Beckfield et al., 2013). A health system, however, is determined by the interaction of different actors, such as insurers, patients, and medical providers. Hence, a precise taxonomy of health systems goes beyond the purpose of the present article, which focuses on insurance only. In this sense, a common classification of health systems from the point of view of the insurance dates back to the early 1970s (Anderson, 1972) and identifies three general approaches: the Bismarck model, the Beveridge model, and the voluntary/private insurance model.

The *voluntary health insurance* model is essentially the one described so far, where people decide whether to buy health insurance and companies supply a variety of contracts suited for all the different risk profiles. It thus sees health insurance as a good to be exchanged in the market, as any other good. To the extent that universal health coverage is a concern within a country, government interventions in this model usually take the form of the introduction of some form of public health

Table 1 Bismarck and Beveridge approaches to health insurance

	Bismarck model	*Beveridge model*
Universal health coverage	Not necessarily	Yes
Entitlement basis	Contribution	Citizenship
Source of finance	Social insurance	Taxes
Number of insurers	Multiple	One
Ownership of insurance providers	Multiple	Public
Ownership of services providers	Multiple	Multiple
Possible patient choice	Of insurers and providers	Of providers
Access to specialists	Can be direct	Through consent from a general practitioner (gatekeeper)
Examples	Germany, the Netherlands, Switzerland, Japan, Luxembourg, France	The United Kingdom, Italy (quasi-market version in Lombardy), Spain, Canada, South Korea

insurance for those high risks for which the system might fail. The typical examples in this sense are Medicare (public insurance for those over 65) and Medicaid (public health insurance for the poor) programs in the US. In addition, more recent reforms, such as the ACA, have tried to tackle adverse selection and the exclusion of preexisting conditions by imposing an individual mandate for health insurance for every US citizen and by requiring insurance providers to compulsorily accept high risks (Austin and Wetle, 2012).

Interestingly, politicians have grasped the weaknesses implicit in the voluntary health insurance mechanisms long before the economic theory could formally explain them. In particular, the two most common names quoted when describing the alternative health systems are Bismarck and Beveridge (Lassey et al., 1997).

Otto von Bismarck (1815–1898), a Prussian statesman and Chancellor, first understood the importance of welfare programs for maintaining order and status by gaining the support of parts of the working class. Often described as a conservative and nationalistic politician, he introduced, with the support of the German industry, old age pensions, unemployment insurance, and health insurance. In *Bismarck* model, health coverage is distributed through employers related to Social Health Insurances (SHIs) (Burau and Blank, 2006). Hence, health coverage can change substantially according to the position in the labor market. SHIs are financed through payroll contributions levied on employers and employees and are in competition with each other. The private sector can thus have an important role within the system, although coordination among different players is a key for success. In particular, competition is usually heavily regulated and the government can keep a strong redistributive role. In general, workers and patients can maintain the power to choose among different insurance providers.

William Beveridge (1879–1963) was a British economist and politician who, in 1942, wrote a very influential report on welfare and unemployment. The so-called 'Beveridge report' set the scene for the establishment of the National Health Services under the Labor government of Clement Atlee in 1945. According to Beveridge, health and social insurance in general should be financed through a contribution by all working people in the country, something that now is referred to as general taxation. Originally, in these systems health care was meant to be provided entirely by public hospitals. Indeed, sometimes in the international literature the Beveridge model is viewed in this purest version. However, as clarified by the *quasi-market* reforms introduced by Margaret Thatcher at the end of the 1980s, Beveridge systems can separate the purchasing authority, which must remain the taxed-based public insurance, from the providers, which can be either private or public.

In Table 1 the main characteristics of the Bismarck and Beveridge models are summarized. Examples of countries are provided. Note that although in principle the distinction between the two types of systems is clear, in most Bismarckian countries the government actually intervenes in various ways to assure universal health coverage and sometimes to fix directly the reimbursement rates of medical services. Indeed, according to some scholars, the two systems are meant to converge in many different dimensions (e.g., Stevens, 2009).

See also: Comparative Health Systems; Health Care Markets: Theory and Empirical Evidence; Health Care, Rationing of; Socioeconomic Status and Health Care.

Bibliography

Akerlof, G.A., 1970. The market for "Lemons": quality uncertainty and the market mechanism. Quarterly Journal of Economics 84, 488–500.

Anderson, O.W., 1972. Health Care: Can There Be Equity?; the United States, Sweden, and England. John Wiley & Sons Inc., New York.

Arrow, 1963. Uncertainty and the welfare economics of health care. American Economic Review 53, 947–973.

Arrow, K.J., 1970. Essays in the Theory of Risk-bearing. North-Holland, Amsterdam.

Austin, A., Wetle, V.L., 2012. The United States Health Care System: Combining Business, Health, and Delivery. Pearson, Upper Saddle River.

Barros, P.P., 2003. Cream-skimming, incentives for efficiency and payment system. Journal of Health Economics 22, 419–443.

Beckfield, J., Olafsdottir, S., Sosnaud, B., 2013. Healthcare systems in Comparative perspective: classification, convergence, institutions, inequalities, and five missed turns. Annual Review of Sociology 39, 127–146.

Brook, R.H., Ware, J.E., Rogers, W.H., Keeler, E.B., Davies, A.R., Donald, C.A., Goldberg, G.A., Lohr, K.N., Masthay, P.C., Newhouse, J.P., 1983. Does free care improve adults' health? New England Journal of Medicine 309, 1426–1434.

Buchmueller, T., DiNardo, J., 2002. Did community rating induce an adverse selection death spiral? Evidence from New York, Pennsylvania, and Connecticut. American Economic Review 92, 280–294.

Burau, V., Blank, R.H., 2006. Comparing health policy: an assessment of typologies of health systems. Journal of Comparative Policy Analysis 8, 63–76.

Cardon, J.H., Hendel, I., 2001. Asymmetric information in health insurance: evidence from the national medical expenditure survey. Rand Journal of Economics 32, 408–427.

Clark, C.R., Ommerborn, M.J., Coull, B.A., Pham, D.Q., Haas, J., 2013. State medicaid eligibility and care delayed because of cost. New England Journal of Medicine 368, 1263–1265.

Claxton, G., 2002. How Private Insurance Works: A Primer. Kaiser Foundation, Georgetown City University Institution for Health Care Research and Policy, Washington, DC.

Cochrane, J.H., 1995. Time-consistent health insurance. Journal of Political Economy 103, 445–473.

Cohen, A., Siegelman, P., 2010. Testing for adverse selection in insurance markets. Journal of Risk and Insurance 77, 39–84.

Cutler, D.M., Zeckhauser, R.J., 1998. Adverse selection in health insurance. National Bureau of Economic Research, 1–32.

Cutler, D.M., Zeckhauser, R.J., 2000. The anatomy of health insurance. In: Culyer, A.J., Newhouse, J.P. (Eds.), Handbook of Health Economics. Elsevier, Amsterdam, pp. 563–643.

De Meza, D., Webb, D.C., 2001. Advantageous selection in insurance markets. Rand Journal of Economics 32, 249–262.

Einav, L., Finkelstein, A., 2011. Selection in insurance markets: theory and empirics in pictures. Journal of Economic Perspectives 25, 115–138.

Ellis, R.P., 1998. Creaming, skimping and dumping: provider competition on the intensive and extensive margins. Journal of Health Economics 17, 537–555.

Ellis, R.P., Manning, W.G., 2007. Optimal health insurance for prevention and treatment. Journal of Health Economics 26, 1128–1150.

Fang, H., Keane, M.P., Silverman, D., 2008. Sources of advantageous selection: evidence from the medigap insurance market. Journal of Political Economy 116, 303–350.

Farrell, J., Klemperer, P., 2007. Coordination and lock-in: competition with switching costs and network effects. In: Armstrong, M., Porter, R. (Eds.), Handbook of Industrial Organization. Elsevier, Amsterdam, pp. 1967–2072.

Finkelstein, A., McGarry, K., 2006. Multiple dimensions of private information: evidence from the long-term care insurance market. American Economic Review 96, 938–958.

Folland, S., Goodman, A., Stano, M., 2006. Economics of Health and Health Care, fifth ed. Prentice Hall, Upper Saddle River.

Frank, R.G., Lamiraud, K., 2009. Choice, Price competition and complexity in markets for health insurance. Journal of Economic Behavior and Organization 71, 550–562.

Hackmann, M.B., Kolstad, J.T., Kowalski, A.E., 2013. Adverse Selection and an Individual Mandate: When Theory Meets Practice (Working Paper No. 19149). NBER.

Hemenway, D., 1990. Propitious selection. Quarterly Journal of Economics 105, 1063–1069.

Hsiao, W.C., 1995. Abnormal economics in the health sector. Health Policy 32, 125–139.

Lassey, M.L., Lassey, W.R., Jinks, M.J., 1997. Health Care Systems Around the World: Characteristics, Issues, Reforms. Prentice Hall, Upper Saddle River.

Manning, W.G., Newhouse, J.P., Duan, N., Keeler, E.B., Leibowitz, A., 1987. Health insurance and the demand for medical care: evidence from a randomized Experiment. American Economic Review 77, 251–277.

Martin, S., 1994. Industrial Economics: Economic Analysis and Public Policy, second ed. Macmillan, New York.

Nyman, J.A., 1999. The economics of moral hazard revisited. Journal of Health Economics 18, 811–824.

Nyman, J.A., 2001. The income Transfer Effect, the access value of insurance and the Rand health insurance Experiment. Journal of Health Economics 20, 295–298.

Nyman, J.A., 2003. The Theory of Demand for Health Insurance. Stanford University Press, Stanford.

Pauly, M.V., 1968. The economics of moral hazard: comment. American Economic Review 58, 531–537.

Phelps, C.E., 2010. Health Economics, fourth ed. Addison-Wesley, New York.

Preker, A.S., Scheffler, R.M., Bassett, M.C., 2007. Private Voluntary Health Insurance in Development: Friend or Foe? World Bank Publications, Washington, DC, pp. 115–146.

Rothschild, M., Stiglitz, J., 1976. Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. Quarterly Journal of Economics 90, 629–649.

Sapelli, C., Vial, B., 2003. Self-selection and moral Hazard in Chilean health insurance. Journal of Health Economics 22, 459–476.

Seog, S.H., 2012. Moral Hazard and health insurance when treatment is preventive. Journal of Risk and Insurance 79, 1017–1038.

Shen, Y., Ellis, R.P., 2002. How profitable is risk selection? A comparison of four risk adjustment models. Health Economics 11, 165–174.

Spenkuch, J.L., 2012. Moral Hazard and selection among the poor: evidence from a randomized Experiment. Journal of Health Economics 31, 72–85.

Stevens, F., 2009. The convergence and divergence of modern health care systems. In: President, W.C.C. of S (Ed.), The New Blackwell Companion to Medical Sociology. Wiley-Blackwell, Hoboken, pp. 434–454.

Tirole, J., 1988. The Theory of Industrial Organization. MIT Press, Cambridge, Mass.

Trivedi, A.N., Moloo, H., Mor, V., 2010. Increased ambulatory care copayments and hospitalizations among the elderly. New England Journal of Medicine 362, 320–328.

US Congress, 2010. Pub.L. 111-148. 111th US Congress, Washington, DC.

Van de ven, W.P.M.M., Ellis, R.P., 2000. Risk adjustment in competitive health plan markets. In: Culyer, A.J., Newhouse, J.P. (Eds.), Handbook of Health Economics. Elsevier, Amsterdam, pp. 755–845.

Zeckhauser, R.J., 1970. Medical insurance: a case study of the tradeoff between risk spreading and appropriate incentives. Journal of Economic Theory 2, 10–26.

Health Interventions, Community-based

John W Farquhar, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes the optimum mix of education, organization, and regulation needed for success in community-based programs for changing health habits underlying preventable chronic disease. Historically, success occurred only when social cognitive theory's principles were embedded within a social marketing framework. A paradigm shift toward greater effects and broader global dissemination could occur were digital health methods incorporated into the current model, particularly if supportive regulatory public health policy were also present.

Recent decades of comprehensive multilevel community-based health interventions have yielded major successes in health outcomes and, in some instances, transfer of novel health education technologies to the health, media, and education infrastructures of communities. The likelihood for success benefited from leadership involving activism, advocacy, and coalition building; and use of social marketing in timing, intensity, frequency, and duration of the messages delivered. Educational content relying on social cognitive theory embedded in a mass communication vehicle produces satisfactory change, even for ordinarily resistant behaviors, such as eating habits and use of alcohol or tobacco. Such a model leads to empowerment, an increase in social capital, and a 'healthy' community, which has achieved 'community efficacy.' The model has had widespread international diffusion and has changed health policies toward primary prevention in many nations. However, there are challenging future obstacles to its adoption, including poverty, social inequity, environmental and educational deficiencies, and, unfortunately, lack of recognition of the excellent scientific basis for the model's efficacy. Achieving future success will therefore require: (1) a conviction that cost-effective technologies for community-based health interventions are ready for adoption and dissemination; (2) policy changes based on a 'prevention pays' philosophy; and (3) a wise allocation of intervention resources, a step needed to overcome barriers and avoid widening existing gaps in health status among population subgroups and among nations varying in their economic development.

This article considers theories and methods underlying successful community-based health interventions, emphasizing six exemplars. Three exemplars focused on cardiovascular disease (CVD) prevention and relied heavily on modern means of mass communication and social marketing, relatively low-cost approaches with the potential to reach and change lifestyle behaviors of entire populations, in contrast to traditional individual or group counseling. One exemplar, studying alcohol-involved trauma, used community organizing for adherence to existing laws, rather than public education for behavior change. Another relied primarily on changes in a city ordinance on a harmful food ingredient, initiated by its public health department. In this instance, the public had become aware of the health issue, largely from a few years of local and national media attention. The sixth exemplar, the nationwide tobacco control program, spanning 1964 to the present, represents how a long-term effort involving multiple communities and states led to a national program, thus demonstrating *national* success from combining educational and regulatory measures used in the five other exemplars. The article also describes barriers to change and lessons learned for future extensions of such endeavors, including the need for a paradigm shift in the methods of education itself.

By definition, a community-based program is organized locally, achieving its goals through the community's institutions and communication networks, often assisted by public health and academic agencies extrinsic to the community. A community is defined here as a residential area with legally defined geographic boundaries, where a governmental system regulates certain local aspects of schools, businesses, transportation, policing, and recreational activities. A community is ordinarily last in a nation's regulatory chain, where education must ultimately occur, although for rural areas in the US, the *county* becomes the governing agent for many services.

The relevance and intellectual contexts of community-based health education are addressed through these questions: (1) What is the rationale for community-wide interventions rather than in more limited locales such as clinics, hospitals, worksites, or schools? (2) What theories, methods, and resources underlie success? (3) How does a community become committed to a novel health education campaign? (4) How can barriers to success be overcome in such interventions? (5) What roles do community activism, laws, and regulations play? (6) What is the future of community-based health education?

Rationale

Health education in schools, worksites, and clinics have a long history of individual success, but synergistic interactive effects can occur when they are embedded in a total community campaign that adds inherently cost-effective mass media and environmental change (Farquhar et al., 1991; Schooler et al., 1997). Culturally induced lifestyle changes, so powerfully influenced by custom and by the media in developed countries, cannot be countered effectively through simple means. Community-wide approaches fit the public health model because the traditional medical model cannot prevent most chronic disease, reach the entire population in need, or modify the environmental and social factors that underlie many

harmful health habits. The community provides influence through locally produced electronic and print mass media, when available, and through worksites, schools, medical and hospital settings, families, and all social networks and social gatherings. It offers opportunities for health-promoting regulations, such as providing opportunities for physical activity for all, school fitness classes, healthful school lunches, alcohol sales limits, and preventing tobacco product marketing to children (*see* Health Promotion in Schools).

Studies at the Stanford Prevention Research Center (SPRC) have shown that only multiple influences produce meaningful changes in the dietary, exercise, or tobacco-use behavior of adults, adolescents, and children. For adolescents, always resistant to health behavior change, the following influenced success: parent and teacher roles; school environments; amount and quality of school-based education on tobacco, fitness, and nutrition; peer influence; and amount and quality of community-wide health education (Perry, 2000). Also, certain personal characteristics were influential, for example, self-efficacy toward one's behavior-change abilities (Fortmann et al., 1995).

Theoretical Formulations

Community-based health education done by SPRC and by some analogous projects has been guided by three major theoretical and management models (Flora et al., 1989; Farquhar et al., 1991).

1. The community organization framework provides methods of identifying the health problem (and resources needed or available), mobilizing the community's opinion leaders and organizations, gaining populace support, forming coalitions, launching and maintaining education programs, achieving regulatory changes, and empowering communities to reach and maintain their goals (Green and Kreuter, 2005).
2. Social marketing provides designs and sequences of messages for population subgroups based on health needs, cultural attributes, social networks, media habits, and readiness for change (Flora et al., 1997).
3. The mass communication–behavior-change model describes theories underlying educational content, such as social cognitive theory (Bandura, 1986), and media-based methods of reaching and influencing the total population in multiple ways (*see* Self-Efficacy and Health).

Intervention Principles, Methods, and Resources Needed for Success

Acceptance

Both relevant organizations and political and opinion leaders must be recruited into a coalition that sanctions any novel health education project. Many projects are initiated by public health and academic leaders extrinsic to the community, but require full acceptance and cooperation within a coalition of the *community's* leaders and organizations. This coalition must identify the health problem (needs assessment), provide a resource inventory, obtain populace support, and plan the intervention, with support derived from both internal and extrinsic sources.

Comprehensive

Interventions must go beyond attempting to change knowledge, the usual educational goal, by providing training in behavior-change skills. They must also go well beyond the individual, enlisting multiple community organizations, including local media outlets, in campaigns for change and seeking changes in the social environment and in regulations that promote access to the facilities and resources needed for healthful practices. Multiple communication channels are needed to reach different audiences, recognizing their preferences, stage of knowledge, and readiness for change. A comprehensive intervention should involve local media organizations, schools, worksites, churches, and sport, recreation, and health facilities. These organizations and others can serve as education conduits, with the community's electronic and print mass-media organizations assisting in message design and delivery. The Internet, mobile, and other digital health devices provide new channels, whose community education role is as yet poorly defined. Interactive computer learning can be designed for large groups – an emerging mass-media variant. The advent of these digital health technologies has now furnished the potential for new and exciting intervention methods (King et al., 2008), which have not yet been studied in a comprehensive multilevel community-based intervention setting.

Variety, Integration, and Timing

Comprehensiveness requires variety. For example, in tobacco control components of Stanford's CVD campaigns, local medical clinics, dental offices, pharmacies, libraries, and voluntary health associations distributed a low-cost skills-training Quit Kit; a local smoking cessation class appeared on television; many newspaper articles and columns were printed; local businesses supported costs of a smoking cessation contest; and all newspapers and electronic media 'cross-advertised' activities designed for mass audiences. Success also requires an appropriate mix and sequence of programs delivered through varied channels. This integration, with goals set in advance and goal changes based on early results, is analogous to a commercial marketing campaign, hence the term 'social marketing.' Timing dictates that differently themed 'mini-campaigns,' often of 3–6 months duration, be used to create novelty in an overall campaign of more than 2 years duration, which is often a time for evaluation of major results.

Message Design and Audience Segmentation

Social marketing requires message tailoring for each subgroup's needs and preferences, recognizing cultural differences, learning styles, and preferred learning sites. Message sequencing should increase awareness, then, increase knowledge, and last, increase motivation and provide training in the skills needed for adoption and maintenance of a new behavior (Bandura, 1986). Electronic media can carry out the sequence's

first two parts and stimulate use of the more information-dense print media of newspapers and booklets, which are inherently more effective in skills training than are electronic media (Flora et al., 1997), although newer digital health technologies can now excel in skills training (King et al., 2008). Messages must be clear, focused, and salient. Salience requires broad-reaching media, arousing interest, and awareness – topics must break through passive indifference engendered by the information overload of many societies and become 'on the public agenda.' Given the large advertising budgets of today's mass media, health agencies' messages must be of sufficient production quality to compete for the public's attention (Farquhar et al., 1990).

The Message Must Be Effective

Behavioral research confirms the rather self-evident rule that 'learning by doing' is more effective than 'learning from observing' (modeling a behavior), and both are more effective than an 'information-only' approach that changes knowledge alone. These principles are contained in Bandura's social cognitive theory, which posits that guided practice in a new behavior can lead to increased self-efficacy and greater behavior change (Bandura, 1986). Thus, 'knowledge-only' campaigns have been found less effective than those that apply Bandura's recommendations (*see* Self-Efficacy and Health).

Intervention Dose

The dose needed depends on many factors: lesser amounts for smaller communities, at earlier stages in a country's adoption of a 'health innovation,' and when advocated behavior changes are reasonably simple (such as mammography, hypertension screening, and immunization campaigns). Clearly, more complex changes are needed in individuals and in society's norms to alter eating or exercise patterns or to control tobacco use. Complexity in respect to nutrition arises from many sources, including long-standing cultural beliefs and practices; entrenched agricultural methods, including those of food production and retailing; advertising of 'unhealthful' foods; and the advent of widespread fast-food chains dominated by commercial interests unresponsive to local demands and needs. Few projects have measured intervention dose with any precision. One excellent method, used in Stanford's Five-City Project (FCP), records the total number and duration of messages distributed over all channels for a defined time period, with message quality estimated from formative evaluation (Farquhar et al., 1990). However, were this method widely adopted, public health practice would benefit greatly.

Role of Activism and Advocacy

Successful community-based health education requires effective leaders, community activists with the courage and charisma to advocate health innovations. Waves of advocacy filtering down from international, national, state, or provincial sources can wash up on the shore of an inert community to provide a local activist leader with the popular support to 'fight city hall' or other entrenched bureaucrats who defend the status quo. Tobacco control in Australia and the US provide examples. National and state advocacy groups with access to mass media created a strong mass movement for change, allowing advocates to enlist popular support for local tobacco control measures (Levy et al., 2012) – (see exemplar six). In both California and Australia's State of Victoria this popular support led to statewide increases in tobacco taxes, with some retained for education against tobacco, a measure that had been resisted by state legislators who had long been influenced by tobacco lobbyists – an example of community activism filtering up the political chain to provide pressure at a higher political level (Catalonia Declaration, 1996; Rogers, 2010).

Role of Policy, Laws, and Regulations

Local actions can affect alcohol and tobacco sales and create environments that improve nutrition and enhance physical activity. However, national, state, or provincial actions can magnify local policy, laws, and regulations, and education efforts on topics, such as tobacco taxation, automobile seat belt laws, food and drug safety, school nutrition, school physical activity policies, and (in the United States) laws on firearms. As described in Section Community Projects in Other Health Topics on tobacco control, widespread popular attitude changes in numerous communities can also affect the political process at the state or federal level.

History of Comprehensive Community-based Health Education for CVD Prevention

The history of CVD risk factor trials since the 1970s is restricted largely to 13 formal research projects (Schooler et al., 1997). They involved entire populations of at least one education community, compared with at least one control community. Three of the following five exemplars are drawn from these 13 projects. Dissemination into practical applications of community organizing and mass communication technologies occurred throughout many parts of the world, derived in part from these research projects.

The First Two Exemplars: The First Decade

The first and second exemplars, Stanford Three-community Study (TCS), in three small agricultural marketing towns in California (total population 45 000), and the North Karelia Study (NKS), in two adjoining predominately rural Finnish counties (North Karelia population about 180 000), each began in 1972. TCS, the first Stanford project, was carried out from 1972 to 1975 in both English and Spanish, comparing effects of mass media alone in one community and mass media plus 10-session risk reduction classes for high-risk adults in a second, with a third as a control (Schooler et al., 1997). Groups exposed to varied education amounts showed a dose–response change in knowledge, smoking, blood pressure, and blood cholesterol, with a proportionately larger effect in the Spanish-speaking residents than in the Anglo majority. A composite CVD risk reduction of about 23% and 30% occurred in the mass media only and mass media plus

classroom conditions, respectively. Thus, a relatively modest amount of mass media (about 30 television and radio 'spots,' weekly newspaper columns on heart health, and 4 separate mass mailings of booklets) was sufficient to change the population's body weight, cholesterol and blood pressure levels, and smoking prevalence. The education followed Bandura's principles but adapted to the mass media (as was also the case with exemplars two and three).

NKS evaluated two, matched, rural Finnish counties containing many villages with farming and lumbering as the main occupations. North Karelia received an education campaign beginning in 1972, continuing for over 20 years. After 5 years, CVD risk factor changes comparable to the TCS and significant net reductions in CVD events occurred (Schooler et al., 1997). Continued intervention, assisted by Finland's excellent public health system, showed progressively larger decreases in coronary heart disease and stroke mortality, and after 20 years, a reduction in North Karelian male lung cancer mortality related to successful smoking cessation (Puska, 1995a). This study employed extensive community organizing, resulting in strong partnerships with residents and their organizations. The NKS influence on its country's policies was unparalleled among the CVD projects, providing its most important lesson – that a well-executed project led by respected scientists can move an *entire* country. As examples, Finland's food and agricultural industries made large changes: fertilizers were supplemented with selenium (a substance needed for health, but low in Finland's soil), milk pricing was changed (based on protein instead of fat content), some dairy farms were replaced with berry farms, a new canola industry replaced jobs lost in dairying, and increased production of low-saturated fat foods occurred (Puska, 1995b).

In 1972, population-wide nutrition change and smoking cessation interventions were an innovation internationally, which may partly explain the success of these two pioneering programs (TCS and NKS).

Early International Diffusion, 1977–83

Studies in small communities patterned after the TCS were done in Italy, Australia, Switzerland, and South Africa, all with significant risk factor changes, adding evidence for the effectiveness of the TCS CVD prevention model (Farquhar et al., 1991; Schooler et al., 1997).

The Second and Third Decades: Projects Begun in the 1980s and 1990s

The third exemplar is the Stanford FCP (USA). Its intervention phase from 1980 to 1986 extended TCS methods to larger populations (total population of about 360 000) with multifactor CVD prevention directed at two northern California cities. There were three control cities (Farquhar et al., 1990). It differed from TCS in greater use of community organizing and collaboration with the communities' health, media, and education organizations in planning and implementing programs. It was similar in generous use of mass media, both print and electronic (63% television and 36% print). It was unique in measuring the total dose of education (about 5 h per year and about 100 episodes of exposure per year to all forms of media and classroom education) (Farquhar et al., 1990).

FCP's initial year of television messages (defined as 'high reach/low involvement') stimulated the public's print media use, which supplied more effective skills training than television (Flora et al., 1997). Results were comparable with the TCS (about a 15% fall in composite risk of CVD), with a major impact on blood pressure and smoking (Farquhar et al., 1990). Health bureaucracies, usually timid, should gain courage from the FCP's 'David and Goliath' demonstration that only 3 h per year of high-quality television health education can counteract the public's exposure to about 100 h per year of television advertising devoted to unhealthy nutrition. Although all preceding CVD studies showed effects in small towns and/or rural districts, FCP showed benefit in cities (with populations as high as 100 000). Also, these effects occurred despite the advent (at least in the State of California) in the 1980s of dual working families and increasing public use of fast food, factors that made achieving nutrition behavior change more difficult.

FCP's modest resources and greater effects, as compared with some comparable projects, support the benefits of the skills-training aspects of Bandura's social cognitive theory and in extensive use of social marketing. Perhaps the most practical lesson to policy makers is that adult residents saved 30 times more money from their decreased cigarette purchases ($120 per adult per year) than the cost of the campaign ($4 per adult per year) – savings retained by the community. Lastly, the communities adopted FCP's technologies, later applying them to seat belt promotion, violence prevention, and adolescent pregnancy prevention.

Other successful CVD projects, both large and small, occurred in these decades in the US, Sweden, Denmark, Canada, the Czech Republic, and China. One notable study, the Minnesota Heart Health Project, developed many innovative methods, including pioneering the use of a community-wide smoking cessation contest (Luepker et al., 1994) and pioneering health education involving multiple schools in a community (Perry, 2000). Also, World Health Organization-sponsored projects began in about 23 other countries (Schooler et al., 1997). In many instances, they borrowed heavily from the experiences of the three exemplars, often receiving training from either Stanford or North Karelia groups.

Community Projects in Other Health Topics

HIV/AIDS trials were launched in the 1990s, using many methods developed in the earlier CVD trials (Coates et al., 2008), giving beginning evidence of the model's effectiveness in developing countries. Interventions on alcohol, mammography, immunization, family planning, mobile X-ray screening, tobacco control, motor vehicle injuries, and, recently, obesity control are prominent examples of other community-based projects, but with fewer well-controlled studies than in CVD. One of these, Preventing Alcohol Trauma (PAT), is the fourth exemplar. This 5-year study of three US communities showed a 10% decrease in alcohol-related traffic injuries and a 50% decrease in adolescent alcohol use (Holder et al., 1997). Coalition building and organizational behavior change among the

police, alcohol sales outlets, and alcohol servers were the main interventions. The public responded to a fear, instilled through publicity, of penalties from greater enforcement of existing laws on underage alcohol purchases or drinking and driving. Therefore, in contrast to the three CVD exemplars, PAT showed that major public education is not needed for large public behavior changes. PAT estimated a cost saving of $2.88 for every dollar invested, a number close to that found in many worksite health promotion (WHP) studies. PAT found their communities had the required infrastructure for the campaigns, requiring only training provided by one indigenous community coordinator, a part-time clerk, an imaginative plan, and the will to proceed.

TCS, the first, and PAT, the fourth exemplar, are opposites: TCS had a maximum of mass-media education and a minimum of organizing, whereas PAT the opposite. Thus, either model works, but to change complex behaviors, one needs sophisticated behavior-change methods, such as those adapted from Bandura (*see* Self-Efficacy and Health) and embedded within a social marketing framework (Flora et al., 1989). If fear of arrest for breaking existing laws on alcohol use suffices to change personal drinking and driving habits, then the education process is simpler.

The fifth exemplar, done recently in New York City, demonstrated that a City's Public Health Department ordinance alone was very successful in eliminating trans fats, which raise cholesterol levels, from the city's restaurants and food markets, whereas a voluntary ban did not. However, an additional feature of this success is that considerable public support existed, engendered by a few years of public attention to the health impact of these fats (Angell et al., 2009). The lesson fits with the aforementioned successes in the national antitobacco campaign's use of laws and regulations in fighting a public health menace.

How Community and State Actions Can Change National Policy

The sixth exemplar, the national tobacco control policy, formally launched in 1964 by the US Surgeon General's first report on Smoking and Health (Smoking and Health, PHS publication 1103, 1964), was assisted throughout by actions at community and state levels. It is often heralded as a singular public health success, appropriately demonstrating the impetus for policy change *filtering up* from an informed, and now concerned, electorate, a process termed 'denormalization' of tobacco use. This emerges as an example of 'multiple communities' banding together, producing a mass effect. The national media, including of course its local media outlets, was responsible for some of the public's transformation, as were the excise tax policies of the States of California and Massachusetts, which acted as harbingers of changes yet to come (Chaloupka et al., 2002). Accordingly, the impetus for public health policy changes at all government levels will increase as economic burdens from preventable chronic disease become clearer. For example, major reductions in health-care costs are predicted from smoking cessation trends in the State of California (Max et al., 2012).

Lessons Lead to a 'Master Plan'

A five-step approach emerges, particularly from the CVD prevention trials, although the educational and organizational elements pertain in principle to all community-based education programs.

1. Define the problem. Does the community need a health education campaign? A decision can emanate from national, state, or local survey data.
2. Organize the community, creating a campaign that includes education, continued organizing, training of community organizations, empowerment of the public and the community's organizations, and future maintenance of programs developed during an initial phase of about 2 years.
3. Implement an intervention, delivering 3–5 h of education exposure per year for 2 years, using initially about 50% television to arouse interest, followed by about 25% from more specific print media (emphasizing newspapers, if available), and 25% from community events and programs (such as health fairs, contests, and classes). Future integration of digital health methods will be needed, following needed research.
4. Institutionalize programs. The community becomes a demonstration project, with its 'empowered' organizations functioning as health promotion resource centers for a wider region.
5. Use the new community resources and its residents' potential power to advocate for local, regional, and national governmental regulations and laws that will increase local intervention effects and extend them beyond the community.

Any individual or organization that wishes to engage in community-based health education should gain courage from the words of Margaret Mead, as quoted in the closing passage of the Catalonia Declaration (1996: p. 75), "Never doubt the capacity of a few dedicated individuals to change the world, in fact, it is the only way it ever has."

Barriers and Challenges

A healthy city has been defined by the absence of crime, crowding, and poverty and by the presence of educated residents and enlightened (and trained) organizations. Together these lead to a community empowered to solve its social problems (i.e., to increase its social capital) (Travers, 1997).

Considering barriers, MacIntyre (2000) found Glasgow's environmental factors to be major barriers to healthful exercise behavior. Certain macroeconomic factors inherent in globalization, such as capital flight and increased wealth and income gaps, have been described as barriers to planned change (Bezruchka, 2000; Cahill, 1983). All such barriers threaten community stability, inhibiting greatly the success of health promotion attempts. However, wise compensatory resource allocation can overcome many barriers, as was shown in the Hispanic minority of the TCS. Therefore, the challenge for the future is: Responsibility for success in community-based health interventions lies with the interventionist, not with the

community's residents! A major potential barrier is omnipresent given the diversity of cultures and levels of economic development worldwide. However, some assurance of commonality of responsiveness is present given the success of some HIV/AIDS programs in developing countries that have used media and organizational approaches common to some of the most successful CVD prevention programs (Coates et al., 2008).

New Opportunities for Greater Cost-Effectiveness

Although, satisfactory cost-effectiveness has been achieved in a few carefully executed comprehensive, multilevel CVD prevention trials, their full potential is greater. Examples: (1) Although, worksite health promotion (WHP) in large US companies has achieved over a 3/1 return on investment in health-care cost reduction (Baicker et al., 2009), WHP has rarely been applied to small worksites, which predominate worldwide, nor has a community-wide education campaign ever targeted *all* the community's worksites; (2) Similarly, school-based health education has never been done for multiple sequential years in all the community's schools during a community-wide health education campaign; (3) Social marketing has rarely been applied well; and, lastly (4) Digital health technologies, including their recent applications (King et al., 2008), have never been embedded within a comprehensive community-based campaign. Topol (2012) persuasively outlines some of digital health's promise – which forecasts a paradigm shift in community-based health education.

Conclusions

Recent decades of the comprehensive, multilevel community-based health education approach strongly support that, at relatively low cost, achieving transfer of public education technologies to a community's infrastructure (public health, media, schools, etc.) can be achieved, resulting in significant changes in health habits of populations. Although most studies derive from small communities, recent successes in New York City, a city of >8 million population (two exemplars examined populations of >100 000), suggest that the model also works in large populations (Schooler et al., 1997). Organizing and educating communities requires advocacy, activism, coalition building, and leadership; success is enhanced by regulatory change. Theory matters: when the population gains self-efficacy through education, the result – community efficacy – enhances capacity to change institutional policy and practice, thus maintaining community change. A paradigm shift toward greater effects and broader global dissemination is an exciting prospect were digital health methods added to the existing model. Science cannot serve society if its evidence for educational benefit is ignored. This is not a new concept. As written over 2000 years ago, "If a virtuous and learned scholar aims to influence the people as a whole, one must first educate the people" (Legge, 1983).

See also: Drugs: Illicit Use and Prevention; Health Behaviors; Health Education and Health Promotion; Physical Activity and Health; Public Health as Social Science; Self-Efficacy and Health; Smoking and Health; Socioeconomic Status and Health Care; Sun Exposure and Skin Cancer Prevention.

Bibliography

Angell, S.Y., Silver, L.D., Goldstein, G.P., Angell, S.Y., Sillver, L.D., Goldstein, G.P., et al., 2009. Cholesterol control beyond the clinic: New York City's trans fat restriction. Annals of Internal Medicine 151, 129–134.

Baicker, K., Cutler, D., Song, Z., 2009. Workplace wellness programs can generate savings. Health Affairs 29 (2), 304–311.

Baker, L., Wagner, T.H., Singer, S., Bundorf, M.K., 2003. Use of the internet and e-mail for health care information: results from a national survey. Journal of the American Medical Association 289, 2400–2406.

Bandura, A., 1986. Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice-Hall, Englewood Cliffs, NJ.

Bezruchka, S., 2000. Is globalization dangerous to our health? Western Journal of Medicine 172, 332–334.

Cahill, J., 1983. Structural characteristics of the macroeconomy and mental health: implications for primary prevention research. American Journal of Community Psychology 11, 553–571.

The catalonia declaration: investing in heart health. In: Declaration of the Advisory Board, Second International Heart Health Conference 1995, 1996. Department of Health and Social Security of the Autonomous Government of Catalonia, Barcelona, Spain.

Chaloupka, F.J., Cummings, K.M., Morley, C., Horan, J., 2002. Tax, price and cigarette smoking: evidence from the tobacco documents and implications for tobacco company marketing strategies. Tobacco Control 11, i62–i72.

Coates, T.J., Richter, L., Caceres, C.F., 2008. Behavioural strategies for HIV prevention: how to make them work better. The Lancet 372, 669–684.

Farquhar, J.W., Fortmann, S.P., Flora, J.A., Taylor, C.B., Haskell, W.L., Williams, P.T., Maccoby, N., Wood, P.D., 1990. The Stanford five-city project: effects of community-wide education on cardiovascular disease risk factors. Journal of the American Medical Association 264, 359–365.

Farquhar, J.W., Fortmann, S.P., Flora, J.A., Maccoby, N., 1991. Methods of communication to influence behaviour. In: Holland, W.W., Detels, R., Knox, G. (Eds.), Oxford Textbook of Public Health, second ed., vol. 2. Oxford University Press, Oxford.

Flora, J.A., Maccoby, N., Farquhar, J.W., 1989. Communication campaigns to prevent cardiovascular disease: the Stanford studies. In: Rice, R., Atkin, C. (Eds.), Public Communication Campaigns. Sage, Newbury Park, CA, pp. 233–252.

Flora, J.A., Saphir, M.N., Schooler, C., Rimal, R.N., 1997. Toward a framework for intervention channels: reach, involvement and impact. Annals of Epidemiology S7, S104–S112.

Fortmann, S.P., Flora, J.A., Winkleby, M.A., Schooler, C., Taylor, C.B., Farquhar, J.W., 1995. Community intervention trials: reflections on the Stanford five-city project. American Journal of Epidemiology 142 (6), 576–586.

Green, L.W., Kreuter, M.W., 2005. Health Program Planning: an educational and ecological approach, 4th edition. McGraw-Hill Higher Education, New York.

Holder, H.D., Salz, R.F., Grube, J.W., Treno, A.J., Reynolds, R.I., Voas, R.B., Gruenewald, P.J., 1997. Summing up: lessons from a comprehensive community prevention trial. Addiction 92 (Suppl. 2), S293–S301.

King, A.C., Ahn, D.K., Oliveira, B.M., et al., 2008. Promoting physical activity through hand-held computer technology. American Journal of Preventive Medicine 34, 138–142.

Lando, H.A., Loken, B., Howard Pitney, B., et al., 1990. Community impact of a localized smoking cessation contest. American Journal of Public Health 80, 601–603.

Legge, J., 1983. The Chinese Classics: Confucian Analects, the Great Learning, the Doctrine of the Mean. Southern Materials Center, Taipei, Republic of China.

Levy, D., Gallus, S., Blackman, K., et al., August 2012. Italy SimSmoke: the effect of tobacco control policies on smoking prevalence and smoking attributable deaths in Italy. BMC Public Health 29 (12), 709.

Luepker, R.V., Murray, D.M., Jacobs, D.R., et al., 1994. Community education for cardiovascular disease prevention: risk factor changes in the Minnesota Heart Health Program. American Journal of Public Health 84, 1383–1393.

MacIntyre, S., 2000. The social patterning of exercise behaviours: the role of personal and local resources. British Journal of Sports Medicine 34 (1), 6.

Max, W., Sung, H.-Y., Lightwood, J., January 17, 2012. The impact of changes in tobacco control funding on healthcare expenditures in California, 2012–2016. Tobacco Control. E-pub ahead of print.

Perry, C., 2000. Commentary: school as a setting for health promotion. In: Poland, B.D., Green, L.W., Rootman, I. (Eds.), Settings for Health Promotion: Linking Theory to Practice. Sage, Thousand Oaks, CA.

Puska, P., 1995a. Mortality changes. In: Puska, P., Tuomilehto, J., Nissinen, A., Vartianinen, E. (Eds.), The North Karelia Project: 20 Year Results and Experiences. Helsinki University Printing House, Helsinki, pp. 159–167.

Puska, P., 1995b. Experience with major subprogrammes and examples of innovative interventions. In: Puska, P., Tuomilehto, J., Nissinen, A., Vartianinen, E. (Eds.), The North Karelia Project: 20 Year Results and Experiences. Helsinki University Printing House, Helsinki, pp. 273–288.

Rogers, T., 2010. The California tobacco control Program: introduction to the 20-year retrospective. The Quarter That Changed the World: Celebrating 20 Years of California Tobacco Control Program, Special Issue of Tobacco Control 19 (Suppl. 1), i1–i2.

Schooler, C., Farquhar, J.W., Fortmann, S.P., Flora, J.A., 1997. Synthesis of findings and issues from community prevention trials. Annals of Epidemiology S7, S54–S68.

Smoking and Health. Report of the Advisory Committee to the Surgeon General of the Public Health Service, 1964. US Dept of Health, Education and Welfare, Public Health Service, PHS Publication 1103, Washington, DC.

Topol, E., 2012. The Creative Destruction of Medicine: How the Digital Revolution Will Create Better Health Care. Basic Books, New York.

Travers, K.D., 1997. Reducing inequities through participatory research and community empowerment. Health Education and Behavior 24, 344–356.

Health Literacy

Kristine Sørensen, CAPHRI, Maastricht University, Maastricht, Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The health literacy field has evolved since the 1970s. It is associated to empowerment and refers to the capacities of people to meet the complex demands of health in a modern society. Regarding health literacy as a social determinant for health, common grounds are established to define the term; measure the concept; and apply it in research, policy, and practice. Growing evidence shows that limited health literacy is a public health challenge, which can be targeted through people-centered, cross-disciplinary, and intersectorial approaches.

The term 'health literacy' was introduced in the 1970s in relation to health education and social policy (Simonds, 1974). Since the 1990s the concept has received increasingly attention as part of the capacities of people to meet the complex demands of health in a modern society (Kickbusch and Maag, 2008). Ever since, the composite notion of health literacy has been scrutinized in terms of how to define and conceptualize it; how to measure it; and how to apply it in research, policy, and practice.

Defining Health Literacy

Health literacy is linked to literacy entailing people's knowledge, motivation and competencies to access, understand, appraise, and apply information to form judgment and take decisions in everyday life in terms of health care, disease prevention, and health promotion to maintain and improve quality of life (Sorensen et al., 2012). As such it is "a shared function of social and individual factors, which emerges from the interaction of the skills of individuals and the demands of social systems" (Nielsen-Bohlman et al., 2004). In turn, a distinction can be made toward public health literacy, which refers to "the degree to which individuals and groups can obtain process, understand, evaluate, and act upon information needed to make public health decisions that benefit the community" (Freedman et al., 2009). Whereas several definitions exist, there is a common understanding that health literacy represents the cognitive and social skills, which determine the motivation and ability of individuals to gain access to, understand, and use information in ways, which promote and maintain good health (Nutbeam, 1998). In addition, the importance of shaping health literate-friendly services and systems has been highlighted to enhance a better interaction between users and health professionals to optimize people-centered care (Brach et al., 2012).

Conceptualizing Health Literacy

The European Health Literacy Consortium (Sorensen et al., 2012) developed a conceptual framework to illustrate how health literacy can be understood as a social determinant for health. It focuses on developing health literacy through being

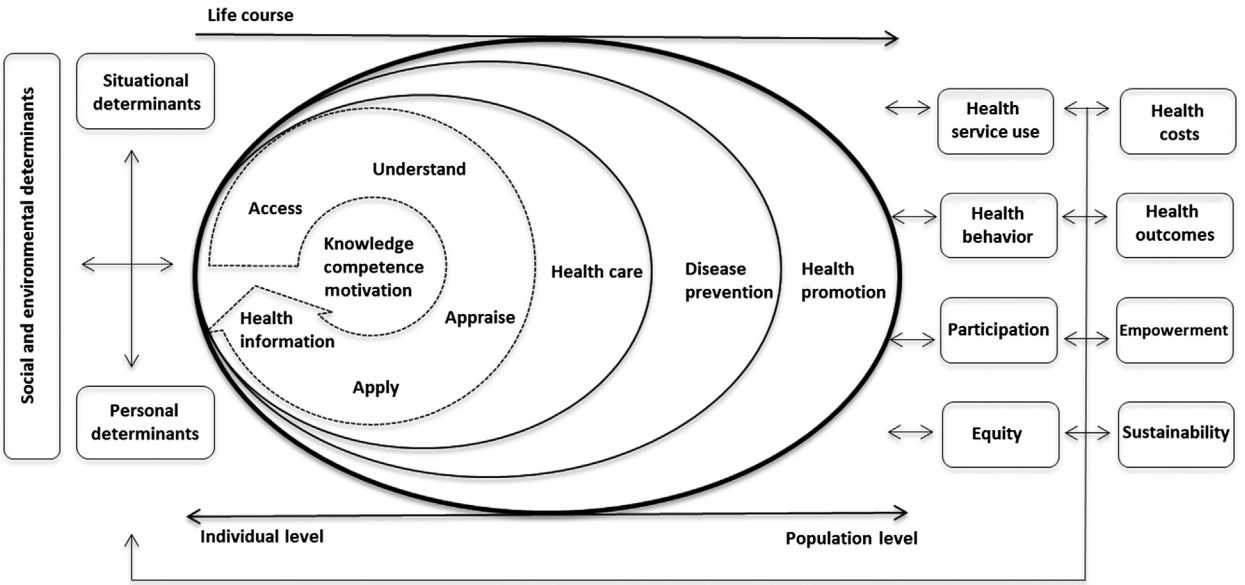

able to access and obtain information; to understand information; to appraise and evaluate information; as well as applying information in all forms when being ill, when trying to avoid risks and when trying to stay healthy during course of life. Health literacy is influenced by personal, situational, and broader societal and environmental factors as outlined in the Ottawa charter (WHO Regional Office for Europe, 1986) and elaborated on by the Commission on Social Determinants of Health (Commission on Social Determinants of Health, 2008). In turn, it is illustrated how health literacy has an impact on health service use (Paasche-Orlow et al., 2005), health behavior (Paasche-Orlow and Wolf, 2007), active participation in society as well as for socioeconomic sustainability (Kickbusch, 2007). The conceptual model takes into account the individual as well as the system perspective, hence recognizing that health literacy relates to the individual capacity to make sound health decisions in the context of everyday life – at home, in the community, at the workplace, in the health care system, in the market place, and in the political arena (Kickbusch, 2008) and health literacy is closely associated to the design of organizations and systems and how they are managed in order to respond to the needs and demands of people (Brach et al., 2012; Parker and Ratzan, 2010).

Previously, widespread health literacy research has been conducted, which was grounded in educational theory based on Nutbeam's three qualities of health literacy: functional health literacy, interactive health literacy, and critical health literacy (Nutbeam, 2000). Yet, recent research suggests that health literacy encompasses even more distinct attributes associated to feeling understood and supported by health care providers, having sufficient information to manage own health, being able to actively managing one's own health, social support for health, appraisal of health information, active engagement with health care providers, navigating the health care system, ability to find good health information and understanding health information well enough to know what to do (Osborne et al., 2013).

Measuring Health Literacy

Numerous tools have been developed over the course of the past 20 years in an effort to measure health literacy in various contexts. Some of these tools were developed as screening tests and others were developed as more comprehensive assessments. Accurate measurement is a critical component of success to identify topics and populations most in need of support, to help tailor interventions, and to provide metrics to evaluate progress (McCormack et al., 2013). The plethora of health literacy measures vary in the dimensions they measure, their mode of administration, the time and resources needed to use them, and level of psychometric rigor (Haun et al., in press). The most well-known measurements include the Rapid Estimate of Adult Literacy in Medicine (Davis et al., 1991, 1993), the Test of Functional Health Literacy (Parker et al., 1995), and the Newest Vital Sign (Weiss et al., 2005). As they are often used in clinical settings, these tools are necessarily short, quick, and easy to use. Other tools aim at measuring a broader concept of health literacy, with a view to provide an in-depth assessment of the dimensions of health literacy, or to explore its relationships with social determinants, health behavior, health status, or healthy service use such as the Critical Health Competence Test (Steckelberg et al., 2009), the European Health Literacy Questionnaire (Sorensen et al., 2013), and the Health Literacy Questionnaire (Osborne et al., 2013).

During the last two decades, the focus of the health literacy measurements has changed considerably. Previously, tools were predominantly designed to measure functional health literacy in clinical settings such as, understanding medical terms (R.M. Parker et al., 1995), whereas later measurements have strived to encompass health literacy as a multidimensional term focusing on a variety of aspects related to health literacy (Osborne et al., 2013; Sorensen et al., 2013). So far, more than 50 tools measuring health literacy have been identified in a recent inventory and more are under development. With the wide range of tools emerging, it is advisable to evaluate the measurements for a conceptual and practical match, when choosing a health literacy instrument for research and screening (Haun et al., in press).

Health Literacy: A Contemporary Public Health Challenge

Health literacy studies across the world has shown limited health literacy to be a risk factor in modern societies (Barber et al., 2009; HLS-EU Consortium, 2012; Nielsen-Bohlman et al., 2004). The growing evidence reveals that limited health literacy is a risk, not only for vulnerable groups such as elderly, the less educated and, among people with low socioeconomic status, but also in the general population per se, when levels of limited health literacy are up to 60%. In this perspective, health literacy can be regarded a contemporary public health challenge that needs to be addressed to ensure a more appropriate use of resources and a better health for all (HLS-EU Consortium, 2012; Kickbusch et al., 2013). Several governments have, therefore, developed strategies to advance health literacy at national levels. The United States has launched a national action plan to improve health literacy (U.S. Department of Health and Human Services., 2010); Canada has opted for an intersectoral approach to improve health literacy for Canadians; in Scotland a national action plan is under development (The Scottish Government, n.d.); and in Wales, the advancement of health literacy is associated to the efforts of overcoming inequalities (Puntoni, 2012). Austria has adopted the investment in health literacy as one of the 10 national goals (Bundesministerium für Gesundheit, 2013). In Australia, a health literacy response framework is being developed that provides tools for a whole-societal approach to build the capacity of citizens as well as of staff involved in organization of services and decision-makers in strategic planning (Osborne, n.d.). Notably, also the supranational organizations such as the European Commission and the World Health Organization have included health literacy as priority areas in their strategies (European Commission, 2007; World Health Organisation's European Office, 2012).

Characteristics of the Evolving Health Literacy Field

While health literacy is gaining more and more attention on the global agenda, the health literacy field itself is maturing. Whereas health literacy in the 1990s was promoted by individual stakeholders believing in its value for health care, disease prevention, and health promotion, the amount of stakeholders in the field has now reached a critical mass that allows for the organization of national and regional conferences as well as development of national, regional, and international networks. With these newly developed platforms for knowledge exchange, it has been possible to share ideas and evidence across the world regions. In the beginning of the century, the published peer-reviewed articles on health literacy could be counted in hundreds, yet within a few years it has raised exponentially to be counted in thousands. Due to the rapid developments and vast amount of information; the role of health professionals as knowledge brokers are as equally important as using and boosting the potential of health literacy embedded in the population (Sorensen, 2013). The advanced conceptualization of health literacy invites professionals from many different disciplines to incorporate health literacy as a priority area into their specific field of work. As a result, the notion of health literacy is increasingly being specified according to the specific topic or target group it involves, e.g., diabetes health literacy, mental health literacy, patient health literacy, aids literacy, health insurance literacy, etc.

Investing in Health Literacy

Nutbeam highlights that the solutions to the health literacy gap will differ whether health literacy is regarded a risk or an asset (Nutbeam, 2008). Generally, limited health literacy should be framed not as a problem of patients and citizens, rather as a challenge to health care providers and health systems to reach out and more effectively communicate with patients, citizens, and families (Kickbusch et al., 2013). Priority areas entail to

- include health literacy as part of school curriculum and life-long learning programs,
- survey and monitor health literacy in the general population as well as in subgroups to target interventions with a participatory approach,
- integrate the health literacy approach in management strategies at all levels in health-associated organizations,
- develop the health literacy of interdisciplinary staff involved in health care, disease prevention, and health promotion,
- identify the cost-effectiveness of health literacy-associated actions, and
- highlight health literacy as an ethical virtue that enhances people-centeredness in health.

The increased focus on cocreation of health between citizens and health providers requires a combination of personal and social responsibility not only from the individuals themselves, but even more so from societal institutions to promote choice, empowerment, self-management, responsiveness, and active and self-directed participation in health and well-being (Kickbusch et al., 2013). The people-centered approach that health literacy is enhancing stimulates self-definition, active realization, and value-driven decision making based on a dual responsibility of the individual and society, which in turn requires freedom of choice and acceptance of diverse outcomes (Sorensen, 2013). The increased global focus on health literacy is expected to qualify health care, disease prevention, and health promotion at large. Yet, to have a profound impact in the future it will require power, priority, and leadership from all involved stakeholders to change the status quo.

See also: Doctor–Patient Interaction in the West: Psychosocial Aspects; Evidence-Based Medicine; Health Education and Health Promotion; Health Promotion in Schools; Public Health as Social Science; Socioeconomic Status and Health.

Bibliography

Barber, M.N., Staples, M., Osborne, R.H., Clerehan, R., Elder, C., Buchbinder, R., 2009. Up to a quarter of the Australian population may have suboptimal health literacy depending upon the measurement tool: results from a population-based survey. Health Promotion International 24 (3), 252–261. http://dx.doi.org/10.1093/heapro/dap022.

Brach, C., Dreyer, B.P., Schyve, P., Hernandez, L.M., Baur, C., Lemerise, A.J., Parker, R.M., 2012. Attributes of a Health Literate Organization. Institute of Medicine.

Bundesministerium für Gesundheit, 2013. Rahmen-Gesundheitsziele Richtungsweisende Vorschläge für ein gesünderes. Österreich, Wien.

Commission on Social Determinants of Health, 2008. Closing the Gap in a Generation: Health Equity through Action on the Social Determinants of Health. Final Report of the Commission on Social Determinants of Health, Geneva.

Davis, T.C., Crouch, M.A., Long, S.W., Jackson, R., Bates, P., George, R.B., Bairnsfather, L.E., 1991. Rapid assessment of literacy levels of adult primary care patients. Family Medicine 23 (6), 433–435.

Davis, T.C., Long, S., Jackson, R., Mayeaux, E., George, R., Murphy, P., Crouch, M., 1993. Rapid estimate of adult literacy in medicine: a shortened screening instrument. Family Medicine 25 (6), 391–395.

European Commission, 2007. Together for Health: A Strategic Approach for the EU 2008–2013. Brussels.

Freedman, D.a, Bess, K.D., Tucker, H. a, Boyd, D.L., Tuchman, A.M., Wallston, K. a, 2009. Public health literacy defined. American Journal of Preventive Medicine 36 (5), 446–451. http://dx.doi.org/10.1016/j.amepre.2009.02.001.

Haun, J., McCormack, L., Valerio, M., Sorensen, K. Health literacy measurement: health literacy measurement: an inventory and descriptive summary of 52 instruments. Journal of Health Communication, in press.

HLS-EU Consortium, 2012. Comparative Report on Health Literacy in Eight EU Member States. The European Health Literacy Project 2009–2012, Vienna.

Kickbusch, I., 2007. Responding to the health society. Health Promotion International 22 (2), 89–91. http://dx.doi.org/10.1093/heapro/dam014.

Kickbusch, I., 2008. Health literacy: an essential skill for the twenty-first century. Health Education 108 (2), 101–104. http://dx.doi.org/10.1108/09654280810855559.

Kickbusch, I., Maag, D., 2008. Health literacy. In: Heggenhougen, K., Quah, S. (Eds.), International Encyclopedia of Public Health, vol. 3. Academic Press, San Diego, pp. 204–211.

Kickbusch, I., Pelikan, J., Apfel, F., Agis, T., 2013. Health Literacy: The Solid Facts, first ed. World Health Organisation Regional Office for Europe, Copenhagen.

McCormack, L., Haun, J., Sørensen, K., Valerio, M., 2013. Recommendations for advancing health literacy measurement. Journal of Health Communication 18 (Suppl. 1), 9–14. http://dx.doi.org/10.1080/10810730.2013.829892.

Nielsen-Bohlman, L.T., Panzer, A.M., Kindig, D.A., 2004. Health Literacy: A Prescription to End Confusion. The National Academies Press, San Fransisco.

Nutbeam, D., 1998. Health promotion glossary. Health Promotion International 13 (4).

Nutbeam, D., 2000. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. Health Promotion International 15 (3), 259–267. http://dx.doi.org/10.1093/heapro/15.3.259.

Nutbeam, D., 2008. The evolving concept of health literacy. Social Science & Medicine (1982) 67 (12), 2072–2078. http://dx.doi.org/10.1016/j.socscimed.2008.09.050.

Osborne, R.H., n.d. OPHELIA. Retrieved from: http://www.ophelia.net.au/.

Osborne, R.H., Batterham, R.W., Elsworth, G.R., Hawkins, M., Buchbinder, R., 2013. The grounded psychometric development and initial validation of the Health Literacy Questionnaire (HLQ). BMC Public Health 13 (1), 658. http://dx.doi.org/10.1186/1471-2458-13-658.

Paasche-Orlow, M.K., Parker, R.M., Gazmararian, J.A., Nielsen-Bohlman, L.T., Rudd, R.R., 2005. The prevalence of limited health literacy. Journal of General Internal Medicine 20 (2), 175–184. http://dx.doi.org/10.1111/j.1525-1497.2005.40245.x.

Paasche-Orlow, M.K., Wolf, M.S., 2007. The causal pathways linking health literacy to health outcomes. American Journal of Health Behavior 31 (Suppl. 1), S19–S26.

Parker, R.M., Baker, D.W., Williams, M.V., Nurss, J.R., 1995. The test of functional health literacy in adults: a new instrument for measuring patients' literacy skills. Journal of General Internal Medicine 10 (10), 537–541.

Parker, R.M., Ratzan, S.C., 2010. Health literacy: a second decade of distinction for Americans. Journal of Health Communication 15 (Suppl. 2), 20–33. http://dx.doi.org/10.1080/10810730.2010.501094.

Puntoni, S., 2012. Health Literacy in Wales. A Scoping Document for Wales. Public Health Wales; Welsh Assembly Government.

Simonds, S., 1974. Health education as sociol policy. Health Education Monograph 2, 1–25.

Sorensen, K., 2013. Health Literacy: The Neglected European Public Health Disparity. Maastricht University, the Netherlands.

Sorensen, K., Van den Broucke, S., Fullam, J., Doyle, G., Pelikan, J., Slonska, Z., Brand, H., 2012. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health 12 (1), 80. http://dx.doi.org/10.1186/1471-2458-12-80.

Sorensen, K., Van den Broucke, S., Pelikan, J., Fullam, J., Doyle, G., Slonska, Z., Brand, H., 2013. Measuring health literacy in populations: illuminating the design and development process of HLS-EU-Q. BMC Public Health 13 (948). http://dx.doi.org/10.1186/1471-2458-13-948.

Steckelberg, A., Hülfenhaus, C., Kasper, J., Rost, J., Mühlhauser, I., 2009. How to measure critical health competences: development and validation of the Critical Health Competence Test (CHC Test). Advances in health sciences education. Theory and Practice 14 (1), 11–22. http://dx.doi.org/10.1007/s10459-007-9083-1.

The Scottish Government, n.d. Health Literacy. Retrieved from: http://www.scotland.gov.uk/Topics/Health/Support-Social-Care/Health-Literacy.

U.S. Department of Health and Human Services, 2010. Action Plan to Improve Health Literacy. Washington, DC.

Weiss, B.D., Mays, M.Z., Martz, W., Castro, K.M., DeWalt, D.A., Pignone, M.P., Hale, F.A., 2005. Quick assessment of literacy in primary care: the newest vital sign. Annals of Family Medicine 3 (6), 514–522.

WHO Regional Office for Europe, 1986. Ottawa Charter for Health Promotion.

World Health Organisation's European Office, 2012. Health 2020: A European Policy Framework Supporting Action across Government and Society for Health and Well-Being. Copenhagen.

Relevant Websites

http://ahls-asia.org/ – Asian Health Literacy Association.

http://www.gezondheidsvaardigheden.nl/english/ – The Dutch Health Literacy Alliance.

https://www.health-literacy.eu – The European Health Literacy Project.

http://www.healthliteracyeurope.org – Health Literacy Europe.

http://www.iuhpe.org/index.php/en/global-working-groups-gwgs/gwg-on-health-literacy – The International Union of Health Promotion and Education Working Group on Health Literacy.

http://www.healthliteracy.org.uk/ – The UK Health Literacy Network.

https://www.wun.ac.uk/research/wun-global-health-literacy-network – The World University Network Working Group on Health Literacy.

http://www.youtube.com/watch?v=fzMA9TIPJUk – Infographic concerning the European Health Literacy Project: HLS-EU The Movie.

Health Outcomes, Assessment of

Robert M Kaplan, University of California, San Diego, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 6581–6586, © 2001, Elsevier Ltd, with revisions made by the Editor in Chief.

Abstract

Measurement of health-related quality of life has become common in health services research and in medical practice. Measures of health-related quality of life can be divided into generic and disease-targeted categories. Generic measures are used to evaluate outcomes from any illness while disease-targeted measures are focused on a particular health condition. Generic measures can be subdivided into psychometric or utility-based. Psychometric methods typically produce a profile of outcomes in which each respondent is given a score on multiple dimensions. Utility-based measures usually summarize health outcomes using a single score that ranges between 0.0 (for death) and 1.0 (for optimum health). Utility-based measures are required for cost/effectiveness analysis. Examples of psychometric measures include the Medical Outcomes Study, the 36-item Short Form, and the Nottingham Health Profile. Decision theory approaches include the EQ-5D, the Health Utilities Index, and the Quality of Well-Being Scale. Examples of disease-targeted measures include the Arthritis Impact Measurement Scale and the University of California, San Diego, Shortness of Breath Questionnaire. Expansion of the number of measures and their applications is expected over the next few years.

Health is probably the most valued human asset. In fact, studies on the preference for different states of being show that virtually everyone rates health as most important. Despite the perceived importance of health, health status has remained difficult to define. There are two common themes in definitions of health. First, premature death is undesirable, so one aspect of health is the avoidance of mortality. The health status of nations is often evaluated in terms of overall mortality rates or infant mortality rates. Second, quality of life is also very important. Disease and disability are of concern because they affect either life expectancy or life quality. For example, cancer and heart disease are the two major causes of premature death in the United States. In addition, disease or disability can make life less desirable. A person with heart disease may face restrictions in daily living activities and may be unable to work or participate in social activities. Even relatively minor diseases and disabilities affect quality of life. A cold, for example, may interfere with the ability to concentrate, work, or attend school. The cold, however, lasts only a short time. A chronic disease, such as arthritis, may affect the quality of life for a long time (Brown et al., 2000).

Within the last few years, medical scientists have come to realize the importance of measuring patient-reported health. Many major diseases, including arthritis, heart disease, and diabetes, or digestive problems are evaluated in terms of the degree to which they affect life quality and life expectancy (Koloski et al., 2000). One can also evaluate treatments for these conditions by the amount of improvement they produce in everyday life. The Food and Drug Administration now considers patient-reported outcomes in their evaluations of new products and nearly all major clinical trials in medicine use quality of life assessment measures. Several approaches to health outcome measurement will be reviewed.

Historical Perspective

For many years, researchers and clinicians have struggled to reach consensus on the measurement of health and health outcomes. Different definitions emphasize related, but divergent, concepts. Some terms used to describe health are positive concepts, such as wellness or normality. Other definitions emphasize negative concepts, such as disability and illness. Much of the debate has centered on whether health is a continuum. Are disability and illness distinct from health, or are they opposite ends of the same continuum? (Patrick and Erickson, 1993).

Because the concept of health lacked conceptual clarity, the World Health Organization (WHO) proposed a comprehensive definition of health in their charter document. Health was defined as "… a state of complete physical, mental, and social well-being and not merely the absence of infirmity" (World Health Organization, 1948). Since the introduction of this definition there has been some convergence of thought. It is now widely recognized that health has multiple dimensions. Further, it is now accepted that measures of mortality alone cannot summarize the health status of populations (Field and Gold, 1998). However, there is still considerable debate about what constitutes health and health status.

Some of the debate about the definition of health originates from sociologists. From the sociological perspective, illness represents deviation from society's standards for physical and mental well-being. Often, this deviation is recognized when individuals cannot perform activities usual for their social roles. In addition to observable functioning, illness may cause reports of symptoms of pain and these deviate from society's norms on wellness. Parsons defined illness as "… a state of disturbance in normal functioning of the total human individual including both the state of the organism as a biological system, and/or his personal and social adjustments" (Parsons, 1951: p. 431). For Parsons, health represented the capacity to perform valued tasks.

The traditional biomedical model is predicated on finding a specific biologic problem and repairing it. In contrast, measures of health outcome consider benefit from the patient's perspective. Successful treatments are those that improve the quality of life or extend life. Most surgical interventions are

justified on the basis of saving lives. However, the great majority of surgical procedures have no effect on life expectancy. For example, four of the most common indications for surgery in the United States are transurethral resection prostatectomy (TURP) for benign hyperplasia of the prostate, extractions of cataracts, joint replacement for severe osteoarthritis of the knee or hip, and hysterectomy for diseases of the uterus. Although it is often argued that these procedures prevent deaths, reviews of the evidence suggest that these procedures are rarely performed to extend life expectancy. In each case, however, the surgery may lead to improvements in functioning or the reporting of symptoms. For example, TURP may reduce the symptom of urinary frequency in men. Cataract extraction may help older patients function better with reading, night driving, and other activities of daily living. Joint arthroplasty may increase mobility for older patients. For most of these cases, traditional physiological measures are limited as outcome assessment. Patients seek these procedures because they want relief of symptoms. Outcomes assessment must consider whether the symptoms are, indeed, relieved. Ultimately, measures of health outcome can be used to evaluate the benefits of treatment from the patient's perspective.

Assessment of Health Outcomes

Methods of assessment of health-related quality of life represent at least two different conceptual traditions. One grows out of the tradition of health status measurement. Several efforts to develop measures of health status were launched in the late 1960s and early 1970s. All the projects were guided by the WHO's definition of health status. The projects resulted in a variety of assessment tools, including the sickness impact profile (SIP) (Bergner et al., 1981), the Quality of Well-Being Scale (QWB) (Kaplan et al., 1995), the 36-item Short Form (SF-36) (Ware and Gandek, 1998), and the Nottingham Health Profile (NHP) (Lowe et al., 1990). Many of the measures examined the effect of disease or disability on the performance of social roles, the ability to interact in the community, and physical functioning. Some of the systems have separate components for the measurement of physical, social, and mental health. The measures also differ in the extent to which they consider subjective aspects of life quality (Brown et al., 2000).

Perhaps the most important distinction between methods to assess quality of life is the contrast between psychometric and decision theory approaches. The psychometric approach attempts to provide separate measures for the many different dimensions of quality of life. Perhaps the best known example of the psychometric tradition is the SIP. The SIP is a 136-item measure that yields 12 different scores displayed in a format similar to an MMPI profile (Bergner et al., 1981).

The decision theory approach attempts to weight the different dimensions of health in order to provide a single expression of health status. Supporters of this approach argue that psychometric methods fail to consider that different health problems are not of equal concern. A runny nose is not the same as severe chest pain. In an experimental trial using the psychometric approach, one will often find that some aspects of quality of life improve while others get worse. For example, a medication might reduce high blood pressure but might also produce headaches and impotence. Many argue that the quality of life notion is the subjective evaluation of observable or objective health states. The decision theory approach attempts to provide an overall measure of quality of life that integrates subjective function states, preferences for these states, morbidity, and mortality.

Common Methods for the Measurement of Quality of Life

A variety of methods have been proposed to measure quality of life, but we cannot review and critique them all here. Instead, we present some of the most widely used psychometric and decision theory-based methods. Readers interested in more detailed reviews should consult Walker and Rosser (1993) and McDowell and Newell (1996).

Psychometric Methods

SF-36

Perhaps the most commonly used outcome measure in the world today is the Medical Outcomes Study (MOS) SF-36. The SF-36 grew out of work by the RAND Corporation and the Medical Outcomes Study (MOS) (Ware and Gandek, 1998). Originally, it was based on the measurement strategy from the RAND Health Insurance Study. The MOS attempted to develop a very short, 20-item instrument known as the Short Form-20 or SF-20. However, the SF-20 did not have appropriate reliability for some dimensions. The SF-36 includes eight health concepts: physical functioning, role-physical, bodily pain, general health perceptions, vitality, social functioning, role-emotional, and mental health (Kosinski et al., 1999). The SF-36 can be either administered by a trained interviewer or self-administered. It has many advantages. For example, it is brief, and there is substantial evidence for its reliability and validity. The SF-36 can be machine scored and has been evaluated in large population studies. The reliability and validity of the SF-36 are well documented (Keller et al., 1999; Scott-Lennox et al., 1999; Stewart and Ware, 1992; Ware and Gandek, 1998).

Despite its many advantages, the SF-36 also presents some disadvantages. For example, it does not have age-specific questions and one cannot clearly determine whether it is equally appropriate at each level of the age continuum. The items for older retired individuals are the same as those for children (Stewart and Ware, 1992). Nevertheless, the SF-36 has become the most commonly used behavioral measure in contemporary medicine.

Nottingham Health Profile

The NHP is another profile approach that has been used widely in Europe. One of the important features of the NHP is that the items were originally generated on the basis of extensive discussions with patients. The NHP has two parts. The first includes 38 items divided into six categories: sleep, physical mobility, energy, pain, emotional reactions, and social isolation. Items within each of these sections are rated in terms of relative importance. Items are rescaled in order to allow them to vary between 0 and 100 within each section. The second part

of the NHP includes seven statements related to the areas of life most affected by health: employment, household activities, social life, home life, sex life, hobbies and interests, and holidays. The respondent indicates whether or not a health condition has affected his or her life in these areas. Used in a substantial number of studies, the NHP has considerable evidence for its reliability and validity.

An important strength of the NHP is that it is based on consumer definitions of health derived from individuals in the community. The language in the NHP is simple and the scale requires only a low level of reading ability. Psychometric properties of the NHP have been evaluated in a substantial number of studies. However, the NHP, like most profile measures, does not provide relative-importance weightings across dimensions. As a result, it is difficult to compare the dimensions directly with one another (Lowe et al., 1990).

Decision Theory Approaches

Within the last few years, interest has grown in using quality of life data to help evaluate the cost/utility or cost-effectiveness of health care programs. Cost studies have gained in popularity because health care costs have grown so rapidly in recent years. Not all health care interventions equally return benefit for the expended dollar. Objective cost studies might guide policy makers toward an optimal and equitable distribution of scarce resources. Cost-effectiveness analysis typically quantifies the benefits of a health care intervention in terms of years of life, or quality-adjusted life years (QALYs). Cost/utility is a special case of cost-effectiveness that weights observable health states by preferences or utility judgments of quality. In cost/utility analysis, the benefits of medical care, behavioral interventions, or preventive programs are expressed in terms of years of life adjusted for reduced quality of life or QALYs.

If a man dies of heart disease at 50 years of age and we expected him to live to 75 years of age, we might conclude that the disease precipitated 25 lost life-years. If 100 men died at 50 years of age (and also had a life expectancy of 75 years), we might conclude that 2500 (100 men $\times$ 25 years) life-years had been lost. Yet death is not the only relevant outcome of heart disease. Many adults suffer myocardial infarctions that leave them somewhat disabled for a long time. Though still alive, they suffer diminished quality of life. QALYs take into consideration such consequences. For example, a disease that reduces quality of life by one-half will take away 0.5 QALY over the course of each year. If the disease affects two people, it will take away 1 year (2×0.5) over each year. A medical treatment that improves quality of life by 0.2 for each of five individuals will result in the equivalent of 1 QALY if the benefit persists for 1 year. This system has the advantage of considering both benefits and side effects of programs in terms of the common QALY units.

The need to integrate mortality and quality of life information is clearly apparent in studies of heart disease. Consider hypertension. People with high blood pressure may live shorter lives if untreated, longer if treated. Thus, one benefit of treatment is to add years to life. However, for most patients, high blood pressure does not produce symptoms for many years. Conversely, the treatment for high blood pressure may cause negative side effects. If one evaluates a treatment only in terms of changes in life expectancy, the benefits of the program will be overestimated because one has not taken side effects into consideration. On the other hand, considering only current quality of life will underestimate the treatment benefits, because information on mortality (death) is excluded. In fact, considering only current function might make the treatment look harmful because the side effects of the treatment might be worse than the symptoms of hypertension. A comprehensive measurement system takes into consideration side effects and benefits and provides an overall estimate of the benefit of treatment (Russell, 1986).

Of the several different approaches for obtaining QALYs, most are similar. The three most commonly used methods are the EQ-5D, the Health Utilities Index (HUI), and the QWB.

EQ-5D

The approach most commonly used in the European community in the late twentieth century was the EQ-5D. This method was developed by a collaborative group from Western Europe, led by Paul Kind and associates, known as the EuroQol group (Kind, 1997). The intention of this effort was to develop a generic currency for health that could be used commonly across Europe. The concept of a common EuroQol was stimulated by the desire for a common European currency – the Euro dollar. The original version of the EuroQol had 14 health states in six different domains. In addition, respondents placed their health on a continuum ranging from death (0.0) to perfect health (1.0). The method was validated in postal surveys in England, Sweden, and the Netherlands. More recent versions of the EuroQol, known as the EQ-5D, are now in use in a substantial number of clinical and population studies (Gudex et al., 1996; Hurst et al., 1997). Although the EQ-5D is easy to use and comprehensive, there have been some problems with ceiling effects. Substantial numbers of people obtain the highest possible score.

Health Utilities Index

Another approach has been developed in Canada by Torrance et al. (1996). This method, known as the HUI, is derived from microeconomic theory. There have been several versions of the measure, typically identified by 'Mark.' The HUI Mark I was developed for studies in the neonatal intensive care unit. The measure had 960 unique health states. In 1992, the HUI Mark II was developed and included 24 000 unique health states. The HUI Mark III, released in 1995, had 972 000 health states. Eight components of the HUI Mark III include vision (six levels), hearing (six levels), speech (five levels), ambulation (six levels), dexterity (six levels), emotion (five levels), cognition (six levels), and pain (five levels). Multiplying the number of levels across the eight dimensions gives the 972 000 states.

Using multiattribute utility scaling methods, judges evaluate levels of wellness associated with each level of each domain. A multiattribute model is used to map preference for the 972 000 possible states onto the 0.0 to 1.0 continuum. The HUI has been used in many population and clinical studies. For overall health status, men obtain higher scores early in the life cycle. However, after about the 45 years of age, women obtain higher scores and this difference grows systematically through the remainder of the life span (Kaplan and Erickson, 2000).

Quality of Well-Being Scale

A third method, known as the QWB, integrates several components into a single score. First, patients are classified according to objective levels of functioning. These levels are represented by the scales of mobility, physical activity, and social activity. Once observable behavioral levels of functioning have been classified, each individual is placed on the 0.0 to 1.0 scale of wellness, which describes where a person lies on the continuum between optimum function and death.

To accomplish this, the observable health states are weighted by quality ratings for the desirability of these conditions. Human value studies have been conducted to place the observable states onto a preference continuum, with an anchor of 0.0 for death and 1.0 for completely well. Studies have shown that the weights are highly stable over a 1-year period and that they are consistent across diverse groups of raters. Finally, one must consider the duration of stay in various health states. Having a cough or a headache for 1 day is not the same as having the problem for 1 year. A health measure must take these durations into consideration. Using this information, one can describe health-related quality of life in terms similar to years of life. For example, 1 year in a state assigned the weight of 0.5 is equivalent to 0.5 of a QALY.

Cost-Effectiveness Evaluations

Utility-based measures of quality of life are often used to evaluate the cost-effectiveness of health care programs (Kaplan et al., 1997). For example, one study using the QWB showed a new medication for patients with arthritis produced an average of 0.023 QALY per year, whereas a new medication for AIDS produced nearly 0.46 of these units per year. However, the benefit of the arthritis medication may last as long as 20 years, ultimately producing 0.023×20 years $= 0.46$ year. The AIDS treatment produced a benefit for only 1 year, so its total effect was 0.46×1 year $= 0.46$ year. In other words, the general system allows the full potential benefits of these two very different treatments to be compared (Kaplan et al., 1995).

QALYs are generic measures of life expectancy with adjustments for quality of life (Gold, 1996). QALYs consider both benefits and side effects of programs in terms of the common outcome units. Although QALYs are typically assessed for patients, they can also be measured for others, including care givers who are placed at risk because they experience stressful life events. The Institute of Medicine recommended that population health metrics be used to evaluate public programs and to assist the decision-making process (Field and Gold, 1998).

In addition to health benefits, programs also have costs. Resources are limited and good policy requires the allocation of resources to maximize life expectancy and health-related quality of life. Thus, in addition to measuring health outcomes, costs must also be considered. Methodologies for estimating costs have now become standardized (Gold, 1996). From an administrative perspective, cost estimates include all costs of treatment and costs associated with caring for any side effects of treatment. From a social perspective, costs are broader and may include the costs of family members staying off work to provide care. Comparing programs for a given population with a given medical condition, cost-effectiveness is measured as the change in costs of care for the program compared with the existing therapy or program, relative to the change in health measured in a standardized unit such as the QALY. The difference in costs over the difference in effectiveness is the *incremental cost-effectiveness* and is usually expressed as the cost/QALY. Since the objective of all programs is to produce QALYs, the cost/QALY ratio can be used to show the relative efficiency of different programs (Gold, 1996).

Some traditional interventions, such as bypass surgery for one-vessel coronary heart disease, may cost as much as $700 000 to produce a QALY. Screening programs, such as mammography, may also require many resources to produce a QALY. On the other hand, public health programs such as pneumonia vaccines for the elderly or laws requiring children to be in infant seats and adults to be in seat belts in motorcars may produce a QALY at a very low cost. It might be argued that programs to the left of the pay line should be funded, but those with cost/QALY ratios to the right of the line should be examined more carefully.

Disease-Specific Measures

This article has focused on generic measures of health outcome. These measures can usually be applied to any age, gender, or ethnic group. In most cases age- and gender-related norms are available. However, there are a significantly larger number of situation-specific measures. Some of these measures are appropriate for a particular age, gender, or ethnic group. For example, several measures are designed specifically to assess health outcomes in children. Most often, the measures are disease-specific, meaning that they are designed to assess outcomes of a particular health problem. Validated measures of health outcome are available for illnesses such as arthritis, diabetes, heart disease, kidney failure, and virtually every other major health condition. Two examples are the Arthritis Impact Measurement Scale (AIMS) (Meenan, 1982) and the UCSD Shortness of Breath Questionnaire (SOBQ) used in studies of patients with emphysema (Eakin et al., 1998).

The AIMS is a health index designed at the Multipurpose Arthritis Center at Boston University. It is intended to measure physical health and social well-being for patients with rheumatoid arthritis (Meenan, 1982). The resultant scale includes 67 items, with questions about functioning, health perceptions, morbidity, and demographics. The AIMS contains scales for mobility, physical activity, social activity, activities of daily living, depression and anxiety, and arthritis-related symptoms. In effect, it is an adaptation of an early version of the QWB (Eakin et al., 1998), with a series of items designed to tap more specifically the effect of arthritis upon functioning and the quality of life. Factor analysis of the AIMS has produced three subscales: physical function, psychological function, and pain. Most current applications of the AIMS use composite scores for these three areas.

The University of California, San Diego, SOBQ includes 25 items that evaluate self-reported shortness of breath during the performance of various activities of daily living. Evaluations of the measure show it highly correlated with other quality of life measures such as the QWB and the Center for Epidemiologic Studies Depression Scale. The measure has high internal consistency (alpha $= 0.96$) and is significantly correlated with

performance measures such as the distance that can be walked in 6 min (Eakin et al., 1998).

Summary

The assessment of patient-reported experience has become a standard component of health outcome evaluation. There are several important distinctions between commonly used approaches. Generic measures are used to evaluate the health outcome for any illness or disease of state. Generic measures can typically be classified as derived from psychometric or decision theory. Psychometric approaches include the SIP, the SF-36, and the NHP. Decision theory approaches are used to estimate outcomes in terms of QALYs. Methods used for this purpose include the EQ-5D, the HUI, and the QWB. Disease-specific measures are available for a wide variety of health conditions. Although disease-specific measures may be more sensitive for outcomes of a particular condition, they cannot be used for cross-illness comparisons or for cost-effectiveness analysis. The assessment of health outcome is a rapidly developing field and we anticipate major new developments over the next decade.

Editors Note

Since the original publication of Dr. Kaplan's article, the literature on assessment of health outcomes has continued to evolve. One significant development has been the development of the Patient-Reported Outcomes Measurement Information System (PROMIS), a standardized data reporting and database system specifically to keep track of patient-reported health outcomes (see David Cella et al., 2010). A useful assessment of this and related assessment tools in health settings can be found in Jill Dawson et al. (2010) (see also Claire F. Snyder et al., 2012).

The search for causes, correlates, and predictors of untoward health outcomes continues apace. Two factors that have figured prominently in recent literature are health literacy and sedentary life styles. Health literacy is an issue because if patients do not understand their doctors or medical regimens, they are far less compliant than otherwise. A useful review of the many studies reaching this conclusion is Nancy D. Berkman et al. (2011). Likewise, those whose sedentary lives bespeak their general indifference to their own well-being also do not report or achieve promising health outcomes; for a review of this voluminous literature, see Karin I. Proper et al. (2011).

The general concern with patient-reported outcomes signals a shift in contemporary medical care toward so-called 'patient-centered' or 'goal-oriented' patient care. Useful descriptions of this larger health care 'movement' are Ronald M. Epstein and Richard L. Street (2011) and David B. Reuben and Mary E. Tinetti (2012).

See also: Subjective Wellbeing, Psychology of.

Bibliography

Bergner, M., Bobbitt, R.A., Carter, W.B., Gilson, B.S., 1981. The sickness impact profile: development and final revision of a health status measure. Medical Care 19 (8), 787–805.

Brown, M., Gordon, W.A., Haddad, L., 2000. Models for predicting subjective quality of life in individuals with traumatic brain injury. Brain Injury 14 (1), 5–19.

Berkman, Nancy D., Sheridan, Stacey L., Donahue, Katrina E., Halpern, David J., Crotty, Karen, 2011. Low health literacy and health outcomes: an updated systematic review. Annals of Internal Medicine 155 (2), 97–107.

Cella, David, Riley, William, Stone, Arthur, et al., 2010. PROMIS developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. Journal of Clinical Epidemiology 63 (11), 1179–1194.

Dawson, Jill, Doll, Helen, Fitzpatrick, Ray, et al., 2010. The routine use of patient reported outcome measures in healthcare settings. British Medical Journal 340, c186.

Eakin, E.G., Resnikoff, P.M., Prewitt, L.M., Ries, A.L., Kaplan, R.M., 1998. Validation of a new dyspnea measure: the UCSD shortness of breath questionnaire. University of California, San Diego. Chest 113 (3), 619–624.

Epstein, Ronald M., Street Jr., Richard L., 2011. The values and value of patient-centered care. Annals of Family Medicine 9 (2), 100–103.

Feeny, D., Furlong, W., Mulhern, R.K., Barr, R.D., Hudson, M., 1999. A framework for assessing health-related quality of life among children with cancer. International Journal of Cancer Supplement 12 (5), 2–9.

Field, M.J., Gold, M.R., 1998. Summarizing Population Health. Institute of Medicine. National Academy Press, Washington, DC.

Gold, M.R., 1996. Cost-effectiveness in Health and Medicine. Oxford University Press, New York.

Gudex, C., Dolan, P., Kind, P., Williams, A., 1996. Health state valuations from the general public using the visual analogue scale. Quality of Life Research 5 (6), 521–531.

Hurst, N.P., Kind, P., Ruta, D., Hunter, M., Stubbings, A., 1997. Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol EQ-5D. British Journal of Rheumatology 36 (5), 551–559.

Kaplan, R.M., Anderson, J.P., Patterson, T.L., McCutchan, J.A., Weinrich, J.D., Heaton, R.K., Atkinson, J.H., Thal, L., Chandler, J., Grant, I., 1995. Validity of the quality of well-being scale for persons with human immunodeficiency virus infection. HNRC Group HIV Neurobehavioral Research Center. Psychosomatic Medicine 57 (2), 138–147.

Kaplan, R.M., Erickson, P., 2000. Gender differences in quality-adjusted survival using a health-utilities index. American Journal of Preventive Medicine 18 (1), 77–82.

Kaplan, R.M., Sieber, W.J., Ganiats, T.G., 1997. The quality of well-being scale: comparison of the interviewer-administered version with a self-administered questionnaire. Psychology & Health 12 (6).

Keller, S.D., Ware Jr., J.E., Hatoum, H.T., Kong, S.X., 1999. The SF-36 arthritis-specific health index ASHI: II tests of validity in four clinical trials. Medical Care (Suppl. 5), MS51–60.

Kind, P., 1997. The performance characteristics of EQ-5D, a measure of health related quality of life for use in technology assessment [abstract]. Annual Meeting of International Society of Technology Assessment in Health Care 13 (5), 81.

Koloski, N.A., Talley, N.J., Boyce, P.M., 2000. The impact of functional gastrointestinal disorders on quality of life. American Journal of Gastroenterology 95 (1), 67–71.

Kosinski, M., Keller, S.D., Ware Jr., J.E., Hatoum, H.T., Kong, S.X., 1999. The SF-36 health survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: relative validity of scales in relation to clinical measures of arthritis severity. Medical Care 37 (Suppl. 5), MS23–39.

Lowe, D., O'Grady, J.G., McEwen, J., Williams, R., 1990. Quality of life following liver transplantation: a preliminary report. Journal of the Royal College of Physicians of London 24 (1), 43–46.

McDowell, I., Newell, C., 1996. Measuring Health: A Guide to Rating Scales and Questionnaires, second ed. Oxford University Press, New York.

Meenan, R.F., 1982. The AIMS approach to health status measurement: conceptual background and measurement properties. Journal of Rheumatology 9 (5), 785–788.

Parsons, T., 1951. The Social System. Free Press, Glencoe, IL.

Patrick, D.L., Erickson, P., 1993. Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resource Allocation. Oxford University Press, Boston, MA.

Proper, Karin I., Singh, Amika S., Mechelen, Willem van, Chinapaw, Mai J.M., 2011. Sedentary behaviors and health outcomes among adults: a systematic review of prospective studies. American Journal of Preventive Medicine 40 (2), 174–182.

Russell, L.B., 1986. Is Prevention Better than Cure? Brookings Institution, Washington, DC.

Reuben, David B., Tinetti, Mary E., 2012. Goal-oriented patient care – an alternative health outcomes paradigm. New England Journal of Medicine 366, 777–779.

Snyder, Claire F., Aaronson, Neil K., Choucair, Ali K., et al., 2012. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. Quality of Life Research 21 (8), 1305–1314.

Scott-Lennox, J.A., Wu, A.W., Boyer, J.G., Ware Jr., J.E., 1999. Reliability and validity of French, German, Italian, Dutch, and UK English translations of the medical outcomes study HIV health survey. Medical Care 37 (9), 908–925.

Stewart, A.L., Ware, J.E., 1992. Measuring Functioning and Well-being: The Medical Outcomes Study Approach. Duke University Press, Durham.

Torrance, G.W., Feeny, D.H., Furlong, W.J., Barr, R.D., Zhang, Y., Wang, Q., 1996. Multiattribute utility function for a comprehensive health status classification system. Health utilities index mark 2. Medical Care 34 (7), 702–722.

Walker, S.R., Rosser, R., 1993. Quality of Life Assessment: Key Issues in the 1990s. Kluwer Academic Publishers, Dordrecht.

Ware Jr., J.E., Gandek, B., 1998. Overview of the SF-36 health survey and the international quality of life assessment IQOLA project. Journal of Clinical Epidemiology 51 (11), 903–912.

Health Politics and Policy

Claus Wendt, University of Siegen, Siegen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Healthcare politics and policy cover two main areas. Healthcare politics refers to how healthcare systems are regulated and governed and maintains a focus on the influence of actors and political institutions. Healthcare policy covers healthcare system characteristics and thus how these systems function. This article, first, provides information about actors and political institutions as well as modes of governance in the field of healthcare. Second, it introduces main characteristics of healthcare systems with respect to healthcare expenditure, financing, service provision, and institutional regulations.

Introduction

Health is considered the most important 'good' of modern society. Due to its great value, attempts have long been made to provide security in the case of sickness. Throughout the nineteenth century, voluntary sickness funds, friendly societies, trade unions, and other organizations began providing coverage in the case of sickness, mainly offering a sickness benefit to protect income when the family's main breadwinner could no longer work because of illness. Partly related to the fact that voluntary organizations could not cover the parts of the population most in need, the state began to step in by the end of the nineteenth century. First, Social Health Insurance (SHI) systems were established (in 1883 in Germany, followed by Austria in 1888, Denmark in 1892, and Great Britain in 1911). One main reason for implementing social insurance in Germany was that the national government had no resources (taxes) to cover the costs and therefore decided to finance public healthcare on the basis of social insurance contributions. Only later did the state assume an even greater role and, in addition to providing the regulatory framework, financed healthcare through taxation starting with the British National Health Service (NHS) in 1946/48 (Klein, 2010).

The term health policy generally describes public policy in the healthcare arena; the role of health policy actors and different modes of governance; the three main dimensions of 'financing,' 'service provision,' and 'regulation'; as well as how healthcare systems are organized. The healthcare arena is characterized by high costs (with some countries spending more than 10% of their GDP on healthcare) and high provider density (with some public healthcare systems being the largest employer in their respective country). Countries have found various forms to organize healthcare. Many of them still resemble either an SHI type of system and an NHS type. A closer look, however, identifies important differences within the respective type. SHI systems from Central and Eastern Europe, for instance, differ from Western SHI systems due to weak corporate actors and a stronger role of the state. Compared with the British NHS, Scandinavian NHS systems are characterized by a stronger influence of regional and local state actors, and Southern European NHS systems suffer from weak administrative capacities.

Data

This article covers OECD countries (Organization for Economic Cooperation and Development). In addition to national sources, the OECD Health Data and the World Health Organization's Health for All Database represent the best modern data sources for international comparison. Increasingly up-to-date country information is given by the European Observatory on Health Systems and Policies covering healthcare systems and reforms both in Europe as well as in some non-European countries (for databases, see web sites below; for international comparison, *see* Welfare Policy Comparisons in this volume).

Actors

The healthcare arena is characterized by more actors (including organized interests) than most other areas of the welfare state. One reason for the involvement of multiple actors is that healthcare is 'service heavy' in contrast to other social policy fields that concentrate on financial transfers (e.g., pensions and unemployment benefits; e.g., *see* Pension Policy in this volume). Actors are not only involved in the financing but also in the delivery of care, and this higher level of complexity (e.g., negotiations among financing agencies and providers) has led to the involvement of multiple actors.

The healthcare arena can be divided into three sections: financing, healthcare provision, and regulation. Three different groups of actors can be involved in each of the three sections: the state, private not-for-profit actors, and private for-profit actors (Rothgang et al., 2010). State actors are present at the national, regional, and local levels. In the British NHS, decision-making in healthcare is still top–down, and the greatest influence is exerted by the Department of Health. However, even in the hierarchical British system, other forms of governance have become increasingly influential, such as the implementation of a purchaser–provider split and other market elements. In other countries, the regional level is of great relevance, as is the case in Denmark, whose regions are responsible for the hospital sector, and Germany, where the Bundesländer cover hospitals' investment costs. In primary healthcare and prevention, local-level actors such as the Danish

municipalities are often responsible. Characterizing private not-for-profit actors is somewhat more difficult. They can be defined as nonstate actors that have no for-profit motives. SHI funds fall into this category and have, by state regulation, received quasi-governmental responsibilities but maintain autonomy from direct state intervention. Welfare organizations that traditionally run hospitals in a number of countries offer another example. Although they may be major players in a country's hospital market, they are not allowed to make profits. Private for-profit actors represent a third group. Private hospitals, private health insurance funds, self-employed doctors, and other self-employed healthcare providers fall under this category.

Governance

In her comparison of France, Sweden, and Switzerland, Immergut (1992) focused on the veto potential of interest groups in different settings and showed that veto opportunities arise out of the specific features of political institutions. Her study suggests that until the late 1980s, the impact of physicians' organizations on health policy may have been much less decisive than was generally believed. Immergut concluded that veto points in the political system are of greater importance than veto groups within society.

Tuohy (1999) provided the most theoretical approach to models of governance. She proposed that the dominant model of accountability in healthcare changed from agency to contract models. In the older, trust-based, principal–agent relationship, the state (the principal) delegated authority for the regulation and distribution of healthcare to the medical profession (the agent). As a result of better access to information, the medical profession dominated this model of governance until the end of the twentieth century. New information technologies have since provided governments with better information, reducing the information asymmetry between the state and the medical profession. In a contract model, the state's role is to purchase rather than to simply finance healthcare. The medical profession's loss of autonomy is due less to cost pressures than to the way in which the provision and verification of detailed medical information strengthen governments' power.

In his concept of 'the health care state,' Moran (1999) classified healthcare systems according to three governing concepts: consumption, provision, and production. By consumption, Moran refers to patients' basis of eligibility for access to healthcare and to financing mechanisms. The provision dimension encompasses the control of hospitals and doctors, and the production dimension encompasses mechanisms that regulate medical innovation. Based on these dimensions, Moran constructed four types of healthcare states: the 'entrenched command-and-control state' (Scandinavian countries, Great Britain), the 'supply state' (United States), the 'corporatist state' (Germany), and the 'insecure command-and-control state' (Greece, Portugal). 'Supply states' are dominated by provider interests in all three dimensions, and it is this domination of suppliers, not the operation of market principles, that Moran considers the primary problem with American healthcare. In 'command-and-control states,' on the other hand, the state is distinctive in all three governing areas. Despite providing universal healthcare in legal terms, Southern Europe's 'insecure command-and-control states' lack the administrative capacity for guaranteeing universal coverage and equal access to care – principles that have characterized the Nordic countries and Great Britain for many decades. In 'corporatist healthcare states,' finally, public law bodies and doctors' associations are dominant.

Financing

Healthcare can be financed by state taxes, social insurance contributions, private insurance contributions, and private copayments. The first two modes of financing represent the idea that access to medical care should not depend on the size of the patients' bank account. In private health insurance, in contrast, insurance contributions are not income related, and redistribution among social groups is low. Private copayments force patients to participate directly in the costs of their sickness and thus represent a particular burden for those in poor health and for lower income groups. Despite much debate about privatization in healthcare, there is currently only a slight retreat of the state at the aggregate level. However, private healthcare financing has become dominant in certain sectors. In most OECD countries, the private share for the financing of pharmaceuticals is above 40% and far above 50% for the financing of dentures (OECD Health Data, 2012).

Among OECD countries, the share of public healthcare financing (taxes and social insurance contributions) displays a modest decrease from more than 76% in the early 1980s to 73% in the 1990s (Rothgang et al., 2010). Countries such as Britain (81.9%), Sweden (81.6%), and Denmark (79.3%) have remained primarily tax financed. France, Germany, and Belgium have continued to rely to a vast extent on social insurance contributions, with shares of 74.3, 67.3, and 61.6%, respectively. A few countries, including the United States (55.6%) and South Korea (47.9%), are financed predominantly by private sources (private insurance and private out-of-pocket). In these countries, however, the private share has decreased since the early 1990s. Moreover, there are some mixed financing systems. In Switzerland and Israel, the shares of both private and social insurance financing are about the same, and Turkey has similar shares of tax, social insurance, and private financing (see **Table 1**).

Provision

In healthcare provision, the role of the state has always been less pronounced than in healthcare financing. In the outpatient sector, healthcare services have mainly been offered by private for-profit actors. Although healthcare providers' access to the healthcare market is strictly regulated (through access to medical schools and through service providers' access to the healthcare market), providers can be defined as private for-profit actors due to their for-profit orientation. In inpatient healthcare, in contrast, we can distinguish between public, private not-for-profit, and private for-profit hospitals. The role of the state (or the public–private mix) in healthcare provision therefore depends on the share of the outpatient sector as well as on the

Table 1 Main sources of healthcare financing: taxes, social insurance, and private financing

	Taxes	*Social insurance*	*Private financing*
Mainly tax financed			
Great Britain	81.9	0.0	18.1
Sweden	81.6	0.0	18.4
Denmark	79.3	0.0	16.3
Ireland	76.1	0.5	23.4
Italy	76.1	0.1	23.8
Portugal	71.0	0.8	28.2
Norway	69.8	13.8	16.5
New Zealand	69.0	8.1	22.9
Canada	68.9	1.4	29.7
Australia	66.9	0.0	33.1
Spain	65.3	5.2	29.4
Finland	58.8	14.6	26.5
Iceland	54.2	27.2	18.6
Mainly social insurance financed			
Czech Republic	8.6	78.7	12.7
France	4.9	74.3	20.7
Luxembourg	19.4	70.8	9.8
Germany	9.5	67.3	23.2
Slovenia	4.6	67.3	28.1
Estonia	10.5	66.2	23.0
Slovak Republic	9.2	65.2	25.6
Japan	16.1	64.3	17.3
Belgium	13.7	61.6	24.6
Hungary	11.3	61.0	27.7
Poland	11.4	57.9	30.6
The Netherlands	3.6	56.9	31.4
Austria	31.0	45.1	23.9
Switzerland	16.7	42.7	40.5
Israel	16.1	40.5	39.0
Turkey	28.2	39.6	32.2
Mainly privately financed			
Greece	30.3	29.8	39.9
United States	31.9	12.6	55.6
Korea	11.8	40.3	47.9

Reproduced from OECD Health Data, 2012. A Comparative Analysis of 30 Countries. OECD.

share of public, private not-for-profit, and private for-profit hospitals. The size of the outpatient healthcare sector depends on whether specialists have the right to open up their own practice or not. In countries such as Denmark and Great Britain, most specialists work in hospitals, and only a few specialists are allowed to become self-employed and open up their own practice or outpatient clinic. The outpatient sector is therefore much smaller compared with that of the United States and Germany, where both general practitioners and specialists are allowed to provide ambulatory healthcare outside hospitals.

Some countries' hospital sector is almost exclusively public (Denmark, Finland, Great Britain, Sweden), while a few rely primarily on private for-profit providers (e.g., South Korea). In the United States, most hospitals are private but maintain nonprofit status. Here, as in Germany, Austria, and France, inpatient care beds are provided through public, private nonprofit, and private for-profit organizations. In Western Europe, there is a long tradition of private nonprofit hospitals. Only in a few countries is the private for-profit hospital sector growing to a great extent, most prominently in Germany, where the share of private for-profit hospital beds as a percentage of total beds has more than tripled, from 4% in 1991 to 15% in 2007 (Rothgang et al., 2010).

Taking all healthcare sectors together, OECD countries display a huge variety in the state's role in service provision, ranging from a public share below 10% in countries such as the United States and the Netherlands to a share of more than 60% in countries such as Finland, Iceland, and Great Britain. As a rule, private (non- and for-profit) provision prevails in market-based and SHI systems, while NHS systems rely more often on public service delivery. However, the public share of healthcare service provision in almost all countries is currently declining, while private (particularly for-profit) service provision is growing.

International comparison thus shows a slight retreat of the state in healthcare financing and a somewhat more pronounced retreat of the state in direct service provision. Both dimensions – financing and service provision – still show vast country variations with respect to the public–private mix.

Regulation

Healthcare can be directly regulated by state actors, by corporate actors in a system of self-regulation, or by market mechanisms. Regulation focuses on the relationship between three sets of actors, that is, healthcare providers, financing bodies, and beneficiaries of the healthcare system. As shown in Table 2, regulation concerns six areas.

Some healthcare systems, most prominently the British NHS, have a long tradition of hierarchical state intervention. Most healthcare policy issues have been regulated at the national level. The growth of market mechanisms in the British NHS that began with the introduction of the internal market in the early 1990s is well known (Klein, 2010). The purchaser–provider split doubtlessly triggered profound organizational changes. However, at the same time, evidence from the initial years of the internal market reveals that little measurable change occurred. About 10 years later, the Healthcare Commission became responsible for both the public and

Table 2 Relations between financing agencies, service providers, and beneficiaries

Between (potential) beneficiaries and financing agencies:
1. Coverage: the inclusion of (parts of) the population in public and/or private healthcare systems
2. System of financing: the financing of healthcare by public (taxes, social insurance contributions) and/or private (private insurance contributions, out-of-pocket payments) sources

Between financing agencies and service providers:
3. Remuneration of service providers: the specific system of provider compensation
4. Access of (potential) providers to healthcare markets: access to financing agencies

Between service providers and (potential) beneficiaries:
5. Access of patients to service providers
6. Benefit package: the content and range of services offered to patients

Reproduced from Rothgang, H., Cacace, M., Frisina, L., Grimmeisen, S., Schmid, A., Wendt C., 2010. The State and Healthcare. Comparing OECD Countries. Palgrave Macmillan, Basingstoke, p. 14.

private sectors and began setting standards as well as engaging in the inspection and assessment of performance, clinical governance, and finance in addition to promoting improvements in healthcare provision. Thus, the market and state regulation in Britain has become more powerful over the past decades (Rothgang et al., 2010).

The state is not only active at the national level. Other examples of strong state intervention include Denmark and Sweden, which demonstrate a major role of the state in both financing and service provision. In contrast to Britain, regulation does not take place at the national but primarily at the regional and local levels. In Denmark, for instance, regions are responsible for the hospital sector while municipalities are in charge of primary healthcare provided by general practitioners, school nurses, and salaried dentists for children in public dental clinics. As in Britain, Sweden has experimented with introducing a purchaser–provider split, and Denmark and Sweden have sought to strengthen competition by giving more room to private hospitals. Freeman and Moran (2000) have argued that 'competition,' which was on the political agenda of a number of healthcare systems in the 1990s, quickly turned into collaboration among larger units of healthcare provision: "In health care we encounter the wider paradox of public policies that try to strengthen market forces: markets need states, and strong markets need strong states" (Freeman and Moran, 2000: 56). Again, Britain is a good example, with Primary Care Trusts, that is, larger collaborative healthcare units, which control about 75% of the NHS budget (Klein, 2010).

Alongside state regulation, a number of countries are characterized by self-regulation, in which corporate actors are responsible for areas of the healthcare system such as financing, remuneration, and the definition of the benefit package (see areas of regulation in **Table 2**). In Germany, sickness funds are self-administered by a professional board and a supervisory board, which generally consist of employees and employers' representatives. Sickness funds, much like panel doctor associations as their negotiating partners in the provision of ambulatory medical care, are public corporations. Since the early 1990s, the self-regulatory core of the German system has come under pressure. Free choice of sickness funds for patients (and thus competition among funds) was introduced in 1993. Competition and other market instruments have since been complemented by a 'risk structure equalization scheme' with financial redistribution to compensate for the different risk structures of funds related to age, gender, number of dependent family members, disability, income, and (since 2007) also differences in the morbidity structure. Furthermore, since 2009, the contribution rates to sickness funds have been set by the national government instead of by the individual sickness funds, as was previously the case. Thus, more competition and more direct state intervention have entered the German SHI at the expense of self-regulation. Within the system of self-regulation, sickness funds have become stronger while doctors' associations have lost part of their former power. In Austria, sickness funds and doctors' associations are also autonomous actors. However, state intervention has always been stronger than in Germany. For instance, contribution rates were for many years fixed by the state as opposed to the sickness funds, resulting in much lower contribution rates in Austria compared with Germany.

The US healthcare system gives greater scope to markets than do other OECD healthcare systems. However, far from being a pure 'private market model,' it combines social insurance elements (Medicare), means-tested programs (Medicaid), private insurance, and direct private out-of-pocket payments. Private health insurance (often in the form of company-based plans), which is the first line of protection against the risk of medical costs for most Americans, is heavily regulated by the state and today provides collective coverage for about 70% of the US population. Since the 1960s, public programs such as Medicare and Medicaid have become increasingly important and today cover about 25% of the population. Since private insurance is also heavily tax subsidized, about half of total healthcare expenditure is financed via public funds (Rothgang et al., 2010). Private out-of-pocket financing was reduced from 48.5% of the total healthcare expenditure in 1960 to 13.2% in 2005 (OECD Health Data, 2012). At the same time, more than 17% of the US population below the age of 65 is without coverage by private or public schemes in the case of sickness (White, 2007). Thus, policies with strong emphasis on private solutions leave room for the exclusion of larger groups of society from access to healthcare, even if healthcare services are financed to a large degree out of the public purse.

Healthcare Systems

Comparative healthcare policy captures the financing, provision, and regulation of medical care. There are several studies that individually consider financing, provision, and regulation (for an overview, see Marmor and Wendt, 2012); however, few comparisons capture all three dimensions (with the notable exception of the textbooks by Blank and Burau (2010) and Gauld (2009)). Such comparison is important for improving our knowledge of how healthcare systems function and what they produce. In particular, combining information from healthcare systems and outcomes such as health status, health inequalities, access to healthcare, and patients' perceptions of medical care arrangements help to better inform health policy decision-making. Healthcare systems can be captured using the dimensions of expenditure, financing, healthcare provision, and regulation.

Expenditure is generally measured either in monetary units per head of population (often in US$ or Euros by using purchasing power parities, or ppp) or as a percentage of countries' gross domestic product (GDP). Healthcare expenditure as a percentage of GDP indicates the amount a society is able or willing to spend on healthcare. Healthcare expenditure in monetary units per head indicates the actual amount a society invests in its members' health on average. Wealthier societies often invest a higher share than do less affluent ones. However, in recent decades, cost containment in healthcare has become a global issue, and the level of resources spent on healthcare has become related less to the level of GDP and more to institutional characteristics of the respective healthcare system. Today, there is a strong negative correlation between the share of public financing and the share of GDP invested in healthcare (Wendt and Kohl, 2010). Countries with a higher share of public financing have been more successful at controlling costs, whereas the United States, which gives

a higher scope to the market and to private health insurance than do most other OECD healthcare systems, spends the highest amount of money on healthcare worldwide.

The second dimension for comparing healthcare systems, namely financing, has already been discussed above. The public share of healthcare financing (particularly tax financing) indicates the interventional power of the state in healthcare. For patients, the mode of financing (taxes, social health insurance, private health insurance, and private copayments) signals coverage, solidarity, and according to Jacob Hacker, the privatization of social risks in the case of sickness (see Marmor and Wendt, 2011). In tax-financed NHS systems, coverage is generally provided on the basis of citizenship; in SHI systems, the provision is on the basis of contribution payments, which often cover dependent family members without own contribution payments. As a rule, both types of financing allow for a high level of redistribution and thus solidarity among their respective populations. Taxes and social insurance contributions are generally related to income, and we find a high redistribution from higher to lower income groups. However, countries such as the Netherlands have introduced flat-rate premiums with the effect of lowering redistribution among certain income groups. Furthermore, in tax- and social insurance-financed systems, we find redistribution from younger to older age groups due to a higher take-up rate of healthcare by older people, from single households to family households with children (if the children are covered without own contributions), and from those in good health to those in poor health since taxes and social insurance premiums are not related to health status. In private health insurance, in contrast, premiums are set irrespective of income and are therefore a greater burden for lower income groups, children are covered by own insurance premiums and there is thus no redistribution toward families with children, and contributions are calculated on the basis of age and health and as a consequence, redistribution among health- and age groups is lower than in tax- and social insurance-financed systems. In the case of direct private out-of-pocket payments, patients participate directly in the costs of healthcare. The higher the share of private out-of-pocket financing, the greater the privatization of risk in the case of sickness and therefore, especially for lower-income groups, the greater the barriers to accessing healthcare services.

The third dimension, healthcare provision, is more difficult to measure than expenditure and financing. Total healthcare employment includes different categories of providers and qualification levels. The main categories of healthcare providers are medical doctors (general practitioners and specialists), nurses, dentists, pharmacists, midwifes, physiotherapists, and doctor's assistants. On average, the number of practicing physicians in OECD countries has increased from about 2.6 per 1000 inhabitants in 1995 to 3.0 per 1000 inhabitants in 2008. We find the highest density of physicians in Austria (4.7 in 2009) and the lowest density in Korea (1.9 in 2009). Other provider groups have experienced similar developments and variations. Today, we find the highest numbers of practicing nurses in Denmark, Finland, Iceland, Norway, and Switzerland, with levels above 14 per 1000 inhabitants, and the lowest numbers in Greece, Korea, and Mexico, with levels below 5 per 1000 inhabitants. While the number of healthcare personnel has increased on average, most OECD countries have opted for a reduction of hospital beds and the average length of stay. The average length of acute hospital stays was reduced from around 8.5 days in 1995 to around 6.5 days in 2008 (OECD mean), with a current range of 19 days in Japan to less than 4 days in Denmark. The average number of acute care beds per 1000 population was reduced from 4.6 in 1995 to 3.6 in 2008 (OECD mean), with a range of 8.1 beds in Japan to 1.6 beds in Mexico (OECD Health Data, 2012). The reductions of hospital beds and the length of hospital stays are partly related to the financing of hospitals on the basis of diagnosis-related groups, which set incentives for shorter hospital stays. However, the Scandinavian countries demonstrate that shorter hospital stays require intensive collaboration between hospitals and social-care units at the local level. Without well-established social care, the quality of service provision may decrease when patients are cared for outside the hospital and, as the US example has demonstrated, overall costs may even increase (Reinhardt, 2011).

In regulation, finally, comparative studies focus on issues such as patients' access to healthcare and doctors' remuneration. While scholars of politics typically ask *who* is responsible for governance and regulation, policy studies tend to focus on *how* healthcare is organized and what the consequences of its structure are (Marmor and Wendt, 2012). Access regulation captures whether patients have free choice of medical doctors or whether they have to sign up on a GP's list for a longer period of time. Here, we find a broad spectrum of systems, including strict access regulation in countries such as Denmark, the Netherlands, and Great Britain, where patients have to sign up on a GP's list and need a referral to access specialist healthcare. Today, only a few countries (e.g., Switzerland) do not require registration with a GP and allow free choice and direct access to specialists. Most OECD countries have introduced some kind of access regulation and require either a referral from a GP or additional copayments when visiting a specialist without a referral. Belgium and France, social insurance countries with traditionally unregulated access to specialists, have introduced gatekeeping elements through financial incentives. In the United States, private insurance conventional indemnity coverage is still an example of free choice, while US managed care (e.g., Health Maintenance Organizations/HMOs, Preferred Provider Organizations/PPOs) restrict patients' choice of both GPs and specialists (Reibling and Wendt, 2012).

The issue of remuneration is important since the remuneration method indicates the degree of doctors' autonomy from state control (Moran, 1999) and presumably also influences the level of healthcare provided. There are four kinds of remuneration of medical doctors, and they can be paid on the basis of fee-for-service, per case, per capita (the number of patients on a doctor's list), or by a salary. Control over doctors' income is highest when paying a salary and lowest under a fee-for-service scheme. The reimbursement method also affects the way doctors alter their workload, and as a rule, fee-for-service methods result in both more active treatment by and higher income for doctors. A fee-for-service payment may set an incentive for the doctor to see his or her patients as often as possible, while reimbursement per capita or a fixed salary may set an incentive for reducing the workload. Countries often combine different remuneration methods, for example,

per-capita payment with either a salary or flat-rate payment (Denmark, Great Britain) or fee-for-service with per-case reimbursement (Germany). When focusing on the main method of GP remuneration, about the same number of countries within the OECD show a preference for fee-for-service (e.g., Austria, Belgium, France, Germany, Japan, Switzerland, United States) and for capitation payment (e.g., Australia, Denmark, the Netherlands, Poland, Great Britain), while fewer countries show a preference for fixed-salary components (e.g., Finland, Israel, Spain, Sweden) (see the European Observatory of Health Systems and Policies).

The healthcare system characteristics presented here have been used for cross-country comparison (see Marmor and Wendt, 2012, for an overview). A meaningful selection of dimensions and characteristics, however, depends on the research question. Access to healthcare providers and the remuneration of medical doctors can be used for analyzing the effects of institutional characteristics on inequalities of healthcare utilization. Additional questions require other healthcare system indicators, for example, for measuring effects on health status and health inequalities and for analyzing healthcare system transformation and related social dynamics (Kuhlmann and Annendale, 2012).

Conclusion

Retrenchment in healthcare is only in its beginning stages, and access to affordable healthcare for the entire population will remain on the political agenda for some time to come. In particular, healthcare costs are expected to increase as a result of demographic changes. Since healthcare is 'service heavy,' efficiency gains are more difficult to achieve than in other areas of the welfare state. Healthcare systems with a high share of public financing have proven more successful in controlling healthcare costs than have private ones. Market-based US healthcare remains the most expensive healthcare system worldwide without also being able to achieve a leading position in healthcare provision. Increasing direct private payments in OECD countries would not only increase the privatization of social risks in the field of healthcare, but would also abolish the necessity of effective cost–control mechanisms. The governance of patients' access to healthcare providers, rather than financial burdens, may lead to a reduced but more equal distribution of necessary healthcare services. Over the past decades, the degree of regulation concerning patients' access to healthcare providers has increased, and this process can be expected to continue. At the same time, the issue of patients' choice is becoming more prominent, particularly in countries with a traditionally high level of access regulation. Patients' choice among larger units of healthcare delivery may eventually combine the goals of cost control, guidance for patients' through the healthcare system, and patients' choice.

See also: Health Care, Rationing of; Health Care: Legal Issues (United States); Health Education and Health Promotion; Health Insurance: Economic and Risk Aspects; Pension Policy; Public Health as Social Science; Socioeconomic Status and Health; Welfare Policy Comparisons.

Bibliography

Blank, R.H., Burau, V., 2010. Comparative Health Policy, third ed. Palgrave Macmillan, Basingstoke.
Freeman, R., Moran, M., 2000. Reforming health care in Europe. West European Politics 23, 35–58.
Gauld, R., 2009. The New Health Policy. Open University Press, Maidenhead.
Immergut, E.M., 1992. Health Politics. Interests and Institutions in Western Europe. Cambridge University Press, Cambridge.
Klein, R., 2010. The New Politics of the NHS. From Creation to Reinvention, sixth ed. Radcliffe Publishing, Abington.
Kuhlmann, E., Annandale, E., 2012. Transforming Health Services and Policy: New International Experiences. Current Sociology (Special Issue). SAGE Publications, Thousand Oaks.
Marmor, T.R., Wendt, C. (Eds.), 2011. Reforming Healthcare Systems, 2 vols. Edward Elgar Publishing, Northampton, MA, USA.
Marmor, T.R., Wendt, C., 2012. Conceptual frameworks for comparing healthcare politics and policy. Health Policy 107, 11–20.
Moran, M., 1999. Governing the Health Care State. A Comparative Study of the United Kingdom, the United States and Germany. Manchester University Press, Manchester.
OECD Health Data, 2012. A Comparative Analysis of 30 Countries. OECD.
Reibling, N., Wendt, C., 2012. Gatekeeping and provider choice in OECD healthcare systems. In: Annandale, E., Kuhlmann, E. (Eds.), Current Sociology, vol. 60, pp. 489–505 (special issue).
Reinhardt, U.E., 2011. Spending more through cost control: our obsessive quest to gut the hospital. In: Marmor, T.R., Wendt, C. (Eds.), Reforming Healthcare Systems, 2 vol. Edward Elgar Publishing, Northampton, MA, USA, pp. 614–623.
Rothgang, H., Cacace, M., Frisina, L., Grimmeisen, S., Schmid, A., Wendt, C., 2010. The State and Healthcare. Comparing OECD Countries. Palgrave Macmillan, Basingstoke.
Tuohy, C.H., 1999. Accidental Logics: The Dynamics of Change in the Health Care Arena in the United States, Britain, and Canada. Oxford University Press, New York.
Wendt, C., Kohl, J., 2010. Translating monetary inputs into health care provision: a comparative analysis of the impact of different modes of public policy. Journal of Comparative Policy Analysis 12, 11–31.
White, J., 2007. Markets and medical care: the United States, 1993–2005. The Milbank Quarterly 85, 395–448.

Relevant Websites

http://www.euro.who.int/en/who-we-are/partners/observatory – European Observatory on Health Systems and Policies.
http://www.euro.who.int/en/what-we-do/data-and-evidence/databases/european-health-for-all-database-hfa-db2 – European Health for All Database (HFA-DB).
http://www.oecd.org/health/healthdata – OECD Health Data.

Health Promotion in Schools

Inge Seiffge-Krenke and Annika Krick, Johannes Gutenberg-Universität, Mainz, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the past two decades, an increasing number of school-based mental health programs and services have been developed and implemented. In this overview, quantitative reviews of prevention and intervention programs targeting adolescent obesity, drug use, smoking, alcohol consumption, and aggressive and violent behaviors such as bullying are presented. Research into the effectiveness of comprehensive school health programs provides evidence of the positive influence of these programs on student health behaviors. However, conditions that initiate and maintain effective school health promotion programs depend on several school-based organizational and program implementation factors. Therefore, implications for school health programs and personnel are given.

Several World Health Organization (WHO) reports have reported an increase in obesity, diabetes, and cardiovascular and other diseases among children and adolescents and outlined the high prevalence of unhealthy diets, alcohol and tobacco use, and inactive lifestyles as causal factors among the youth (Jepson et al., 2010). Therefore, preventing unhealthy lifestyles and promoting health-enhancing behaviors as early as possible are important objectives. For this reason, health promotion in schools is a topic of both practical and empirical interest.

Schools that provide comprehensive school health promotion programs are more effective in encouraging children to adopt health-enhancing behaviors and reducing health-compromising behaviors than schools that provide only health education. In this overview, quantitative reviews of prevention and intervention programs targeting adolescent obesity, drug use, smoking, alcohol consumption, and aggressive and violent behaviors such as bullying are presented. Research into the effectiveness of comprehensive school health programs provides evidence of the positive influence of these programs on student health behaviors. However, conditions that initiate and maintain effective school health promotion programs depend on several schools-related organizational and program implementation factors.

Changing Concepts of Health Promotion

According to the WHO, health is defined as more than avoiding illness. Similarly, health promotion involves more than steering people away from unhealthy lifestyles. Health promotion is viewed, according to the Ottawa Charter for Health Promotion (1986), as the process of helping people to increase control over and to improve their health. To reach a state of complete physical, mental, and social well-being, individuals or groups must be able to identify and realize their aspirations, to satisfy needs, and to change or cope with their environment. The Jakarta Declaration on Health Promotion in the 21st Century (1997) presented the principles of equity, empowerment, and participation as central to health promotion practice and outlined priorities for promoting social responsibility for health, increasing investments for health development, increasing community capacity, empowering the individual, and securing an infrastructure for health promotion. However, working with values and principles of health promotion is difficult, because none of these key documents report how to apply and implement these principles in practice. In addition, even though many health promotion models were designed and implemented to evaluate health promotion programs using principles such as empowerment and participation, the realization of these values and principles were not the focus. Currently, the ever-widening gap between rich and poor is an issue of great concern, and compelling evidence confirms that societal and material exclusion is highly linked to health problems. Health inequalities exist in the levels of mortality and morbidity, accidents and injuries, risk behavior, mental health, school achievement, and family violence. Given that American and European children and adolescents in high-risk settings (e.g., low income families, high unemployment of the caregivers, poor neighborhoods) were in poorer health than those living in wealthier settings, the implementation of the principles of equity, empowerment, and participation remains a continuous challenge for future health promotion in schools.

Health-Promoting Programs

In the past two decades, an increasing number of school-based mental health programs and services have been developed and implemented. The development and assessment of school-based health programs is an emerging field with many issues that need to be addressed (Paternite, 2005).

These programs vary with respect to age range, comprehensiveness, theoretical background, and the importance of evaluation. Although some studies comprised primarily survey data, others examined the effect of intervention programs. Several papers incorporate a substantial body of epidemiological and demographic data as basis for their conclusions. Most discuss policy options and recommendations. Thus, substantial variation exists with respect to such factors as level of detail, use of research data, intended audience, and political viewpoint.

In the following sections, some examples of comprehensive programs will be provided, prominent trends will be presented,

and finally frequent domains in health promotion programs will be evaluated.

A prime example is the Child and Adolescent Trial for Cardiovascular Health (CATCH), which is the most comprehensive multilevel program that has been extensively evaluated and disseminated since its development (Hoelscher et al., 2001). The interventions targeted changes at the school level (food services offerings and minutes of class time devoted to physical activity) and at the student level (psychosocial factors, dietary behaviors, physical activity levels, smoking intentions, physiological indicators). CATCH interventions consisted of (1) Eat Smart, a school food service program to provide children with meals reduced in fat and sodium; (2) CATCH physical education, to increase the amount of time students spent in physical activity at school and for other times of the day; (3) Hearty Heart and Friends, 15 classroom lessons to teach dietary and physical activity information through a series of stories; and (4) Family Fun Nights, during which pupils and family members participated in 90 min of games, fitness-related activities, and healthy snacks. The trial demonstrated positive changes in the school food service and physical education program, as well as in students' cardiovascular health behaviors.

Generally, health promotion programs have been confined to highly selective programs designed to identify specific health risks, such as tobacco smoking, alcohol consumption, or drug use, and to train health-promoting behaviors, for example, coping strategies. Most of these programs are designed for late childhood and early adolescence and pursue multiple strategies for preventing health damage, including more traditional strategies (e.g., teaching drug information and general communication and problem-solving skills) and approaches based on the social influence model (which emphasizes strategies for identifying and resisting social pressure to use drugs or alcohol). Time reserved for cognitive instruction was usually brief, with most curricula ranging between 5 and 15 lessons. Although the overall cognitive progress in most programs was impressive, the time provided was not sufficient to produce corresponding reduction in alcohol, drug or nicotine use. Mostly, the effects were stronger and longer lasting for girls than for boys. Frequently, developmentally relevant variables were incorporated; for example, increasing peer pressure or a risky lifestyle during adolescence. Resistance-Skill Training, however, was not always found to be an effective strategy for preventing the onset of alcohol, drug or tobacco use; sometimes counterproductive effects were found.

Health promotion in middle schools is a neglected area, although the middle school years are a time of increasing risk for negative health-related outcomes. Health promotion in the middle school mostly focuses on a social decision approach (e.g., teaching strategies and skills that help children to make favorable decisions when confronted with health damaging situations and offers) and social skills training (in which children are prompted, coached, and guided to practice using the skills). These approaches are based on the observation that children are unlikely to retain and use information conveyed solely through formal lessons; they need opportunities for reinforcement and practice.

Additionally, in recent years, health promotion or health education programs have been developed for elementary school children. For example, innovative methods of engaging and increasing physical activity in children in elementary school have been implemented, examining the effectiveness of an unstructured, moderate to vigorous before-school physical activity program on academic performance (Tompkins et al., 2012).

Recently, the interaction with families of high-risk children became a major focus, particularly in settings outside school. The Triple P Program (Sanders, 2008), for example, aims at helping parents of kindergarten children with sleeping and eating disorders to cope with these challenges by developing positive parenting attitudes and behaviors.

Results of the Healthy Youth Survey in Washington State that assessed school-based health interventions and academic achievement showed strong associations between health risks and academic achievement. Based on 13 key physical and mental health risk factors, the association between health risks and academic performance was bidirectional. Students who performed poorly at school had more health risk factors, which negatively influenced their performance and in turn led to further health risks (Dilley, 2009).

To summarize, health promotion and education programs target early childhood to adolescent students and assess problematic behaviors (that are highly interrelated such as drug and alcohol use) and unhealthy lifestyles (with respect to nutrition and physical activity). The number of programs preventing bullying, violence, and aggression or stress increased during the last decade. Most programs are implemented in educational environments, but preventive approaches sometimes reach out to families with school children at risk.

Evaluation of School-Based Prevention and Intervention Programs

In assessing program outcomes, the following five criteria should be met: (1) the use of appropriate comparison groups; (2) the use of appropriate pretests, posttests, and follow-up assessments; (3) examining research attrition; (4) the validity of dependent variable assessments; and (5) the use of statistical procedures appropriate for program evaluation research. In the following sections, quantitative reviews of prevention and intervention programs targeting adolescent drug use, smoking, alcohol consumption, physical activity, bullying, violence and aggression, and stress are presented. They represent a new approach, because program classification schemes were used to categorize outcomes, which were then employed in a quantitative metaanalysis of study effect sizes based on the criteria listed above.

Metaanalyses of Smoking Prevention Programs

Current estimations show that about one of four adolescents smokes cigarettes on a regular basis. Tobacco-use prevention interventions for adolescents that have been studied can be classified as community-based, school-based, or health care provider-based programs (Krowchuk, 2005). In school environments, interventions can be administered as part of the normal curricula with access to almost all pupils. The frameworks used in implementing these interventions include

inform-and-instruct models, social competence models, social influence models, and combined social competence/social influence models. Rigorous analytic studies confirmed that interventions based on the social influence model showed the greatest effects (Thomas, 2002). Programs based solely on the social competence model are also quite effective in preventing teens from smoking. Increasing knowledge of the adverse effects of smoking does not prevent adolescents from smoking. No evidence could be found that the combination of methods is more effective than social influence alone (Skara and Sussman, 2003). For example, a metaanalysis conducted by Thomas et al. (2013) revealed an average 12% reduction in starting smoking for pure prevention cohorts compared with control groups. Studies reporting change in smoking behavior over time showed no overall effects, but at an intervention level positive findings were found for social competence and combined social competence and social influences interventions.

Effectiveness of Drug Abuse Resistance Education

School-based drug prevention programs have been an integral part of the US and European antidrug campaign. Among the different programs proliferated, Project DARE (Drug Abuse Resistance Education) is the most widely implemented school-based, drug use prevention program in the US, which has been adopted by approximately 50% of school districts nationwide. However, results of a metaanalysis showed that DARE's short-term effectiveness for reducing or preventing drug use behavior is small (Pan and Bai, 2009). Considering the enormous amount of public financial support for DARE, the short-term effectiveness of DARE cannot be regarded as satisfactory (West and O'Neal, 2004).

Effectiveness of Bullying Prevention Programs

Studies showed that bullying in school is seen as a severe problem that concerns one in five school-aged children throughout the whole world. Bullying has recently become an issue of public concern and has also reached the attention of the major media networks. Articles have been featured in several top newspapers and magazines describing cases of children who committed (or attempted) suicide because of bullying. Consequently, school bullying has increasingly become the focus of intervention and evaluation programs in school. Recent metaanalyses have revealed only modest effects of antibullying campaigns (Smith, 2011).

These results are evidence that single-level interventions are not effective. In contrast, multidisciplinary whole-school interventions have been shown to be the most effective in reducing bullying behavior with rate reductions of up to 20% (Ttofi and Farrington, 2012). Baldry and Farrington's (2007) assessment of bullying studies is noteworthy. They reported that about half of the studies had positive outcomes and the other half showed mixed, small, or undesirable effects.

The most effective strategies target those who are bullied, those who bully others, and bystanders to the bullying. Universal intervention programs have that potential (Michaud, 2009). The metaanalysis by Ttofi and Farrington (2012) revealed that the most effective program components for reduction of bullying prevalence rates were videos, disciplinary methods, parent training/meetings, and cooperative group work between professionals.

Effectiveness of Physical Activity Health Promotion Programs

Data from the WHO revealed that 1.9 million deaths worldwide are associated with physical inactivity, and at least 2.6 million deaths are caused by excessive weight or obesity (WHO, 2004). Children and adolescents spend a substantial amount of time in school and lack the free time to engage in physical activity to support their well-being. Consequently, public health agencies must invest in the promotion of physical activity (WHO, 2010).

A Cochrane review by Dobbins et al. (2013) analyzed randomized controlled trials of school-based programs for physical activity and found that the interventions had small effects on physical activity rates, systolic and diastolic blood pressure, body mass index, and pulse rate. In total, school-based physical activity interventions increased the number of children involved in moderate to vigorous physical activity during school hours. In addition, children and adolescents who took part in the program spent more time in moderate to vigorous physical activity (with results across studies ranging from 5 to 45 min more) and spent less time watching television (results range from 5 to 60 min less per day).

Effectiveness of Violence and Aggression Prevention Programs

Violence among adolescents is an important public health problem in the US; the rate of violent crimes averages about 4.2 for every 100 youths between the ages of 12 and 20. Reports from victims revealed that the rate of violent and aggressive acts by youth perpetrators is higher than in any other age group (Maguire and Pastore, 2003).

The Surgeon General's 2001 report on youth violence (U.S. Department of Health and Human Services, 2001) reported a list of school-based violence prevention programs as model or encouraging programs, such as the Seattle Social Development Project, Families and Schools Together (FAST) Track, the PATHS (Promoting Alternative THinking Strategies) curriculum, and I Can Solve Problems. Evaluated school programs focused on topics or skills to reduce aggressive or violent behavior such as emotional self-awareness, emotional control, self-esteem, positive social skills, social problem-solving skills, conflict resolution, and teamwork.

In elementary and middle school, the intervention focus is on disruptive and antisocial behavior and shifts to general violence and specific forms of violence, including dating violence in middle and high school. Results revealed that all strategies (e.g., informational, cognitive/affective, and social skills building) showed a reduction in violent behavior. Universal school-based programs also have additional benefits, such as significant improvements in school attendance and school success. A metaanalysis of school-based programs confirmed and supplemented the *Community Guide* findings (Hahn et al., 2007). However, even single approaches had a mild positive effect on reducing aggressive and violent behavior (Park-Higgerson et al., 2008).

Effectiveness of Stress Prevention Programs

Studies have indicated that children are constantly affected by daily stressors in their daily routines. Their capacity and skills to cope with these stressors were significantly associated with their psychological well-being (Pincus and Friedman, 2004).

Various studies have revealed that children can use coping skills as a buffer or moderator against negative life stressors to maintain psychological well-being. Dysfunctional coping is a highly relevant issue for the development and maintenance of depression, considering that depression is ranked as the fourth leading cause of disease, accounting for almost 12% of disabilities worldwide (Üstün et al., 2004), and its association with dysfunctional psychosocial adaptation over the life span (Myron, 2008). Therefore, the development of adequate coping skills and especially school-based interventions is important. Programs are developed with the aim to teach children specific problem-focused skills to enhance coping. Fewer programs focus on emotion-focused coping strategies, such as how to modify thoughts or feelings, or learning how to manage the demands of a situation (Pincus and Friedman, 2004).

Two programs are presented below as examples of stress and coping prevention intervention that have had positive effects.

The *Rochester Child Resilience Project*, a preventative program, was developed to support coping strategies with the goal of enhancing resilience in urban children (Cowen et al., 1995). In grades 4–6, urban children were taught various skills for coping with life stress. This program consists of 12 sessions. The five variables found that divide between stress-resistant and stress-affected children were perceived global self-worth, realistic control attributions, empathy, social problem-solving skills, and self-esteem. Two sessions consist of teaching children the distinction between solvable and unsolvable problems and the training of specific strategies for coping with unsolvable problems. The third session focuses on identifying feelings in oneself and in others and on teaching the relationship between feelings and life experiences.

Outcome data revealed several positive effects. Learning problems and task-orientation significantly improved. Evidence of additional benefits was also found for overall perceived self-efficacy and for dealing with new and difficult problematic situations. Furthermore, participants had more realistic control expectations for coping with controllable and uncontrollable problems and also reported reduced anxiety (Pincus and Friedman, 2004).

The *I Can Do Program* was developed as a short-term classroom-based program to teach children in general problem-solving and coping skills, to give children practice in utilizing these skills to five specific stressors, to increase efficacy for effective implementation of the coping strategies for each stressor, to enhance children's problem-solving skills in relation to each stressor, and to increase children's knowledge about the stressors. Overall, program results have been positive. Children significantly improved in the effectiveness of their coping solutions even at 5-month follow-up. This program provides promising evidence that supporting children in utilizing specific coping skills can improve efficacy of coping (Pincus and Friedman, 2004).

Implications for School Health Programs and Personnel

The School Health Education Evaluation indicated that for school health education to promote health in children, schools should devote at least 50 h of classroom time each year to health instruction. However, in many industrialized countries, particularly in southern and eastern Europe, these standards are not met (Takanishi and Hamburg, 1997). Limited resources, including curricular time, make it unlikely that health promotion in schools will realistically increase in strength. Thus, developing more efficient ways to promote children's health is essential; for example, by adding complementary interventions to school health education from other sources.

Metaanalyses of more recent school-based studies indicate that carefully evaluating the programs is a necessary part of the improvement process. Overall, improvements of methodology and conceptualization of intervention programs are noticeable, but additional empirical work needs to be conducted to upgrade substandard procedures regarding experimental attrition and to establish the validity of dependent variable measures. Furthermore, the need still exists for developing program intervention models that clearly and fully follow tenets of the guiding theory. Prevention and intervention programs that have been already designed and evaluated are largely confined to certain age groups (often to adolescents) and are limited to the treatment of certain problematic behaviors (such as drug, alcohol, or tobacco use). The health needs of younger ages, for example, preschool children or children from grades 1 to 5, are largely neglected.

Based on the recent developments of health promotion values and principles detailed in the Jakarta Declaration, health promotion in schools should acknowledge societal factors such as inequalities and poverty as important determinants of health in children and adolescents. Health promotion in schools, therefore, must incorporate life quality at the community level when designing programs. School health personnel needs to manage factors that influence children and adolescent health but are beyond the traditional focus of health education and behavior change. They need to assess the quality of life as it is perceived by children and adolescents and identify the specific needs of individual children. Then, community-based action can be developed to address these issues.

See also: Health Education and Health Promotion; Occupational Health; Social Exclusion, Social Deprivation and Health.

Bibliography

Baldry, A.C., Farrington, D.P., 2007. Effectiveness of programs to prevent school bullying. Victims and Offenders 2, 183–204.

Cowen, E.L., Wyman, P.A., Work, W.C., Iker, M.R., 1995. A preventive intervention for enhancing resilience among highly stressed urban children. The Journal of Primary Prevention 15 (3), 247–260.

Dilley, J., 2009. Research Review: School-Based Health Interventions and Academic Achievement. Retrieved from: http://sboh.wa.gov/Pubs/docs/Health&AA.pdf.

Dobbins, M., Husson, H., DeCorby, K., LaRocca, L.R., 2013. School-based physical activity programs for promoting physical activity and fitness in children and adolescents aged 6 to 18 (Review). The Cochrane Database of Systematic Reviews (2).

Hahn, R., Fuqua-Whitley, D., Wethington, H., Lowy, J., Crosby, A., Fullilove, M., Dahlberg, L., 2007. Effectiveness of universal school-based programs to prevent violent and aggressive behavior – a systematic review. American Journal of Preventive Medicine 33 (2S), S114–S129.

Hoelscher, D.M., Kelder, S.H., Murray, N., 2001. Dissemination and adoption of the Child and Adolescent Trial for Cardiovascular Health (CATCH): a case study in Texas. Journal of Public Health Management and Practice: JPHMP 7 (2), 90–100.

Jakarta Declaration on Health Promotion into the 21st Century, 1997. World Health Organization, Geneva.

Jepson, R.G., Harris, F.M., Platt, S., Tannahill, C., 2010. The effectiveness of interventions to change six health behaviours: a review of reviews. BMC Public Health 10, 538.

Krowchuk, H.V., 2005. Effectiveness of adolescent smoking prevention strategies. MCN The American Journal of Maternal/Child Nursing 30 (6), 366–372.

Maguire, K., Pastore, A.L., 2003. In: Pastore, A.L. (Ed.), Sourcebook of Criminal Justice Statistics 2003. U.S. Department of Justice, Bureau of Justice Statistics: 2005, Washington, DC. Retrieved from: http://www.albany.edu/sourcebook/ (accessed 21.05.13.).

Michaud, P., 2009. Bullying: we need to increase our efforts and broaden our focus. Journal of Adolescent Health 45, 323–325.

Myron, L.B., 2008. Child and adolescent mental disorders: the magnitude of the problem across the globe. Journal of Child Psychology and Psychiatry 49, 226–236.

Ottawa Charter for Health Promotion, 1986. World Health Organization, Geneva.

Pan, W., Bai, H., 2009. A multivariate approach to a meta-analytic review of the effectiveness of the D.A.R.E. program. International Journal of Environmental Research and Public Health 6, 267–277.

Park-Higgerson, H.K., Perumean-Chaney, S.E., Bartolucci, A.A., Grimley, D.M., Singh, K.P., 2008. The evaluation of school-based violence prevention programs: a meta-analysis. Journal of School Health 78 (9), 465–479.

Paternite, C.E., 2005. School-based mental health programs and services: overview and introduction to the special issue. Journal of Abnormal Child Psychology 33 (6), 657–663.

Pincus, D.B., Friedman, A.G., 2004. Improving children's coping with everyday stress: transporting treatment interventions to the school setting. Clinical Child and Family Psychology Review 7 (4), 223–240.

Sanders, M.R., 2008. Triple P-Positive Parenting Program as a public health approach to strengthening parenting. Journal of Family Psychology 22, 506–517.

Skara, S., Sussman, S., 2003. A review of 25 long-term adolescent tobacco and other drug use prevention program evaluations. Preventive Medicine 37, 451–474.

Smith, P.K., 2011. Why interventions to reduce bullying and violence in schools may (or may not) succeed: comments on this special section. International Journal of Behavioral Development 35 (5), 419–423.

Takanishi, R., Hamburg, D.A. (Eds.), 1997. Preparing Adolescents for the Twenty-First Century: Challenges Facing Europe and the United States. Cambridge University Press, New York, pp. 82–107.

Thomas, R., 2002. School-based programmes for preventing smoking. The Cochrane Database of Systematic Reviews (2). http://dx.doi.org/10.1002/14651858.CD001293.

Thomas, R.E., McLellan, J., Perera, R., 2013. School-based programmes for preventing smoking. The Cochrane Database of Systematic Reviews (4).

Tompkins, C.L., Hopkins, J., Goddard, L., Brock, D.W., 2012. The effect of an unstructured, moderate to vigorous, before-school physical activity program in elementary school children on academics, behavior, and health. BMC Public Health 12. Article Nr. 300.

Ttofi, M.M., Farrington, D.P., 2012. Risk and protective factors, longitudinal research, and bullying prevention. New Directions for Youth Development, 85–98. http://dx.doi.org/10.1002/yd.20009.

U.S. Department of Health and Human Services, 2001. Youth Violence: a Report of the Surgeon General. U.S. Department of Health and Human Services, Washington, DC.

Üstün, T.B., Ayuso-Mateos, J.L., Chatterji, S., Mathers, C., Murray, C.J., 2004. Global burden of depressive disorders in the year 2000. British Journal of Psychiatry 184, 386–392.

West, S.L., O'Neal, K.K., 2004. Project D.A.R.E. outcome effectiveness revisited. American Journal of Public Health 94, 1027–1029.

World Health Organization (WHO), 2004. Global Strategy on Diet, Physical Activity and Health. Retrieved from: www.who.int/dietphysicalactivity/strategy/eb11344/en/.

World Health Organization (WHO), 2010. Global Recommendations on Physical Activity for Health. WHO, Geneva, Switzerland.

Health Psychology

Denise de Ridder, Utrecht University, Utrecht, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by T.W. Smith, volume 10, pp. 6602–6608, © 2001, Elsevier Ltd.

Abstract

Health psychology is an area of applied psychological research and a profession. Health psychology research examines the determinants and consequences of physical health, showing that health behaviors are critical to health. Four core topics in health psychology are discussed in this article: (1) the promotion and maintenance of health, (2) prevention and treatment of (chronic) illness, (3) the etiologic and diagnostic correlates of health and illness, and (4) patient behavior in the health care system.

Health psychology is concerned with the study of psychological processes in health, illness, and health care. Health psychologists study behavioral factors associated with staying healthy, and they examine how patients with health problems can be helped to improve their quality of life and well-being. They also examine individuals in experimental research settings, work with groups in real-life settings, or observe interactions taking place between health professionals and patients in a health care setting. Findings from basic research are applied by health psychologists in a variety of settings, ranging from primary care and hospitals to municipal health services and health policy bodies, employing different intervention techniques, such as health education, training, and cognitive-behavioral therapy. Health psychology is a fast growing subdiscipline in psychology and employs theories from personality psychology, social psychology, developmental psychology, and biological psychology to understand and explain how psychological factors impact on health. Specifically, social psychological theories on behavioral regulation have proven important in advancing the field and contributing to why and how people adopt health goals and engage in strategies to achieve these goals (Mann et al., 2013). Similarly, self-regulation models, which have been developed to explain adaptive responses to threats, have been derived in specific formats to explain responses to threatened health goals (De Ridder and De Wit, 2006). These theories draw attention to the difficulties that may arise when individuals have to deal with long-term goals while being occupied with immediate needs that can be at odds with long-term investments in health. These more specific theories have replaced the generic biopsychosocial model (Schwartz, 1982) that was used in earlier work but has proven difficult to inform research and practice.

The term 'health psychology' was first used by Matarazzo in 1980, defining it as "the aggregate of the specific educational, scientific, and professional contributions of the discipline of psychology to the promotion and maintenance of health, the prevention and treatment of illness, and the identification of etiologic and diagnostic correlates of health, illness, and related dysfunction" (p. 815). Health psychology is distinguished from related disciplines such as clinical psychology and behavioral medicine with its focus on physical (rather than mental) health and its orientation on behavioral (rather than medical) aspects of health. Following Matarazzo's definition, the core areas of health psychology are commonly defined as (1) the promotion and maintenance of health, (2) prevention and treatment of (chronic) illness, and (3) the etiologic and diagnostic correlates of health and illness. Another important core area in health psychology has been defined as (4) patient behavior in the health care system.

Promotion and Maintenance of Health

Lifestyle factors that have been identified as relevant for health, such as the holy five of eating, smoking, alcohol use, exercise, and sleep, are an important topic in health psychology. These health behaviors are estimated to account for half of the premature death from the 10 leading causes (most prominently, cardiovascular disease and cancers) in the developed world. While infectious diseases were the leading causes of morbidity and mortality until the first half of the twentieth century, lifestyle-related diseases are now regarded as most contributing to morbidity and mortality. Although many other factors are involved in the development of these diseases, it is recognized that healthier lifestyles are the key to reducing the incidence of these conditions. The concept of health behavior as behaviors that are relevant for health either because they pose a risk to health (e.g., smoking) or because they promote health (e.g., sufficient sleep) was established in the 1960s in a large epidemiological study among 7000 adults living in Alameda County (California, USA), showing that behaviors such as not being physically active, skipping breakfast, or regularly drinking more than five units of alcohol at a single session were predictive of mortality (Belloc, 1973). Health behavior has been defined as "any activity undertaken by a person believing himself to be healthy for the purpose of preventing disease or detecting it at an asymptomatic stage" (Kasl and Cobb, 1966). Importantly, activities that are relevant for health from a medical point of view may not be regarded as such by individuals. Individuals smoke because they consider it relaxing or they eat fatty foods because they appreciate their taste – they do not necessarily engage in these behaviors because they want to behave unhealthily. In a similar vein,

people may engage in health-promoting activities for other reasons than a concern about their health and, for example, exercise because they enjoy the fun of being physically active. One particularly important issue in health behavior research is to better understand the factors associated with people's nonengagement in behaviors that are beneficial to their health. Even when people have a genuine desire to stay healthy, they often do not display the behavior that is required to achieve their health goals. The 'intention behavior gap' (Sheeran, 2001) has become a widely known concept to explain why people may have the intent to be healthy but most of the time face serious difficulties in engaging in that behavior for a variety of reasons, such as underestimating the effort that is associated with acting upon one's intentions, lack of planning the activities that are required for acting, or forgetting to act when one is occupied with other activities. It should be noted that the concept of health behavior is fluid and the activities that are included may vary as a function of medical knowledge. For example, the use of a condom when having sex with a person whose sexual history is unknown only emerged as a preventive health behavior after the human immunodeficiency virus/acquired immune deficiency syndrome (AIDS) epidemic in the 1980s. Apart from the classic five health behaviors named previously (eating, smoking, alcohol use, exercise, and sleep), nowadays safe sexual behavior, using sunscreen, driving sensibly, using a seat belt, and dental behaviors are also regarded as relevant for health, but this may change as medical knowledge evolves.

Health behaviors have been studied mainly from the perspective of social-cognitive models, such as the health belief model and the theory of planned behavior, highlighting the beliefs that individuals hold about the outcomes of their action as key determinants of health behavior (Armitage and Conner, 2000). These beliefs are thought of as relatively stable characteristics that have the potential to influence behavior while being amenable to change. The models that have been used are based on the expectancy–value principle: the expectancy that a given action will lead to a specific outcome and the value that is attached to that outcome. These models thus highlight that individuals engage in deliberate rational decision making about several options. The most widely used social-cognitive model is the theory of planned behavior that has gained considerable empirical support, although explaining intention to engage in health behavior (39% of the variance in intention) has proven easier than predicting actual health behavior (21% in case of objective reports and 31% in case of subjective reports) (Armitage and Conner, 2001). Research in social-cognitive models has resulted in interventions (e.g., education and training in primary care setting) employing persuasive communication strategies to change health beliefs, mostly with disappointing results. Several effective intervention strategies exist that help people to formulate intentions to change their health behavior, whereas strategies aimed at supporting people to enact their intentions have witnessed modest results, even when information was tailored to the specific needs of people depending on the stage of change they were in (Weinstein et al., 1998). A particular successful exception are interventions that employ implementation intentions (Gollwitzer, 1999) that specify when, where, and how the intended behavior should be performed by employing an if–then plan, increasing the chance that people will recognize the opportunities for action, which they may otherwise forget when they experience a momentary drop in their motivation for action. New generations of models of health behavior have departed from social cognition models and reflect that health behavior is not always rational and deliberate but rather governed by habits, impulses, or emotions (Hofmann et al., 2008). To date, these novel approaches recognize that interventions aimed at improving health behavior should take account of the very nature of health behavior and implement strategies that do not require a lot of effort to support people who do not have sufficient resources to change their behavior in the desired direction (Marteau et al., 2012). Regardless of the different approaches that are typically used, health promotion in behavioral models pertains to primary prevention of any type of illness that may result from engaging in these risky health behaviors and thus pertains to "maneuvers that reduce the chances that a health problem will ever develop" (Kaplan, 2000: p. 382).

Prevention and Treatment of (Chronic) Illness

Prevention and treatment of specific illness has some overlap with the first area but obviously focuses on people who have been identified to be at risk for disease (for example, those who are at risk for breast cancer as the result of screening programs involving mammography) or who have been diagnosed with chronic illness (for example, rheumatoid arthritis). Here, the aim is to detect risk of disease at an early enough stage in order to eliminate or slow down its development or, in case of already diagnosed illness, to promote adjustment to disease.

With regard to the former, screening programs (secondary prevention) take the form of health checks, such as measuring weight, blood pressure, carrying out cervical smears, and offering genetic tests for illness such as Huntington's disease and some forms of cancer. Three broad types of screening exist: opportunistic screening (using the time when a patient visits medical services for a particular reason), population screening (which involves the systematic examination of potential health problems in large community surveys), and self-screening (encouraging people to practice self-examination of, for example, breasts and testes, or buy over-the-counter kits to measure blood sugar levels or cholesterol). The aim of all screening programs is to detect a problem at the asymptomatic stage in order to offer medical treatment as early as possible. Although there exists enthusiasm for screening as a prevention strategy, there are ethical considerations as well as issues of cost-effectiveness and the psychological consequences of screening for individuals (Shaw et al., 1999). For health psychology, the effects of screening on the psychological state of individuals are most relevant. For example, receiving a negative result (i.e., not having the condition tested for) may create a false sense of reassurance while receiving a positive result (i.e., having the condition tested for) may lead to elevated levels of anxiety. Most studies in this area have shown that stress and negative effect resulting from receiving a diagnosis disappear after a short while. In addition, health psychologists study the factors that predict uptake of screening among the population, which has demonstrated considerable variety relating to health beliefs, risk perception, and feelings of

being able to do something about one's own health. Studies on perceived risk and related concepts such as perceived vulnerability refer to subjective estimates of the likelihood of personally contracting a disease and have proven a more powerful predictor of screening uptake than the perceived (serious) consequences of disease (Weinstein and Klein, 1995). Message framing (i.e., presenting health risk information as gains vs presenting it as losses) has proven an important moderator of how the general public responds to testing for health risks (Rothman and Salovey, 1997).

There is an extensive literature on the theoretical and empirical foundations of coping and adjustment in chronic illness. It is estimated that at any given moment, about 50% of the population suffers from chronic illness requiring some form of medical intervention (Taylor and Aspinwall, 1996). These patients face adaptive tasks that require them to come to terms with the disease-related changes, such as dealing with pain and incapacity, preserving an emotional balance, sustaining relationships with family and friends, and preparing for an uncertain future (Bensing et al., 2002). Coping is a response to the experience of such adaptive tasks and consists of three stages (Lazarus and Folkman, 1984): an event appraisal of what is at stake, an appraisal of one's personal and social resources to deal with the task, and the behavioral or emotional coping response aimed at either resolving the problem (problem-focused coping) or one's emotional reaction to the problem (emotion-focused coping). Research on coping strategies has met with some methodological challenges and effects of coping as studied by omnibus questionnaires have been relatively modest. Nevertheless, the coping concept is immensely popular among health professional involved in the design of interventions to help people adjust to their chronic condition (De Ridder and Schreurs, 2001). While most research has focused on problem-focused coping as a means to adjust to (chronic) illness, emotion-focused coping has demonstrated favorable results on adjustment. Specifically, expression (rather than inhibition) of emotions and benefit finding in adverse events as specific forms of emotion-focused coping have shown impressive results (De Ridder et al., 2008; Stanton et al., 2000). Another issue deserving discussion has to do with the kind of outcome measures that health psychologists employ as dependent variables in the studies examining effects of psychological interventions. A typical measure derived from medical practice is to determine the level of adjustment by health-related quality of life, defined as measures of daily functioning and assessed by the widely used Rand 36-Item Health Survey (Van der Zee et al., 1996). Such functional measures of mental, physical, and social functioning are relevant in the context of medical treatment but have proven to be unrelated to psychological measures of adjustment, such as satisfaction with life or well-being (Taylor and Aspinwall, 1996). The basic assumption of psychological theories is that the successful pursuit of meaningful goals plays an important role in well-being (Ditto et al., 1996). Therefore, several researchers in the field of quality-of-life studies have emphasized the role of personal goals and related constructs in patient perception of quality of life. Highlighting this perspective may also account for the finding that many patients with chronic illness report positive adjustment (De Ridder et al., 2008).

Etiologic and Diagnostic Correlates of Health and Illness

Early research in the 1950s on psychological factors in chronic illness has been characterized by a strong desire to identify personality factors, which would explain why some individuals run a greater risk of becoming ill than others. In an attempt to link psychological factors to the onset of disease, research was based on the assumption that a specific psychological makeup would be predictive of disease. A typical example is the type A personality with a proneness to coronary artery disease whose behavior is characterized by competitive drive, impatience, and hostility (Booth-Kewley and Friedman, 1987). Nowadays, such strong views on the role of personality and psychological factors in the etiology of (specific) disease are no longer maintained. Rather, recent research emphasizes the role of psychological factors in (dealing with) stress and emotions. The experience of psychophysiological arousal (stress) and related negative mood is the result of exposure to stressors that have been defined as any event or condition that is demanding or challenging. Typical categories of stressors are calamitous events (e.g., natural disasters), major life events (e.g., bereavement), daily hassles, and chronic stress conditions (e.g., occupational stress) that are evaluated as threatening and potentially harmful and thus requiring some adaptive response. Acute stressors may challenge the cardiovascular, neuroendocrine, and immunological systems, whereas acute stress in general does not strain these systems as does chronic stress. The most commonly held view of the link between stress and illness suggests that stress leads to disease due to a prolonged interaction of physiological, behavioral, and psychological factors. The 'stress–diathesis' model (Levi, 1974) posits that the wear and tear caused by stress can explain the accumulative damage to the cardiovascular system and other systems of the body. Stress has been mostly studied in the context of coronary heart disease, but there are also studies exploring links between stress and illnesses such as cancer, diabetes, rheumatoid arthritis, and AIDS. Research has been devoted to studying the impact of stress on sympathetic activation (e.g., increased blood pressure and irregular heartbeats via the prolonged production of (nor)adrenaline), which may lead to increased chances of heart disease and infectious diseases, and hypothalamic–pituitary–adrenocortical activation (e.g., cortisol), which may lead to decreased immune function. In addition to these direct effects of stress on health, research has also explored the indirect impact via health-threatening behaviors such as smoking and alcohol abuse, which are also related to the experience of stress. Current stress models consider a wide variety of responses to stress, including mood, behavior, problem solving, motivation to achieve (health) goals, and engagement in health-protective behavior. These wide-ranging effects of stress illustrate the importance of examining the effects of stress on the whole organism rather than focusing on one system only. Responses across all systems work in concert to help the individual adapt by either altering the situation or accommodating its effects. However, chronic activation of these response systems may make individuals more susceptible to negative physical health outcomes at any point in the disease process: onset, progression, treatment, recovery, or recurrence. Despite the impressive effects of stress

on health, there is considerable individual variability in the stress–illness link caused by individual differences in stress reactivity and stress recovery, as well as ways of coping with stress. Negative effects of stress on health become manifest when coping and other resources are not sufficient to overcome stressful conditions. Uncontrollable stress appears more difficult to handle than controllable and predictable periods of threat or demand. An important moderator of the impact of stress on health is social support. In addition to health-protective effects of being integrated in social networks (i.e., through marriage, children, or friends), social support research has identified stress-buffering effects of supportive relationships especially when they provide emotional support. A consistent finding reported in the literature is that social support is related to better cardiovascular functioning and better immune responses (Uchino et al., 1996). Also, strategies for dealing with negative emotions have attracted considerable attention as moderators of the stress–illness relationship. There is accumulating evidence that repression or inhibition of negative emotions may be a risk factor in the onset and course of disease. Although the exact psychophysiological mechanisms involved are still unclear, it has been suggested that the inhibition of emotions may result in increased physiological activity. Denial and nonexpression of emotions can be a useful initial coping strategy to deal with the stress of disease; failure to acknowledge and express emotions can leave these emotions unresolved with chronic raised activity of the sympathetic nervous system as a result. The inhibition of emotions can also delay help-seeking behavior or compromise communication with health care providers. In contrast, the regulated (rather than uncontrolled) expression of negative emotions has demonstrated impressive effects on health. A typical example is the disclosure of emotions paradigm by talking or writing about one's emotions (Smyth, 1998). Talking or writing about emotions is believed to make the emotional experience less invasive, increase insight into why emotions are experienced, and restore psychophysiological balance (De Ridder et al., 2008).

Patient Behavior in the Health Care System

Once an individual gets ill, he or she enters the health care system. Health psychology examines a variety of topics that are related to patients using health services, such as symptom recognition and the decision to seek help, the interaction between patients and health professionals, and adherence to treatment regimens. It is estimated that over 50% of the population experiences symptoms of ill health over any 2-week period. Only a minority of these people seek help from health professionals, depending on how symptoms are perceived and interpreted. Research on symptom perception has been informed by the seminal work of Leventhal (Leventhal et al., 1984), distinguishing between five beliefs about symptoms, which influence subsequent action (identity, cause, consequences, timeline, and control and treatment effectiveness). This and other research suggest that beliefs about symptoms and emotional responses to symptoms more strongly affect the decision to consult medical professionals than the mere experience of symptoms themselves. Depending on the beliefs individuals hold, they may either delay help seeking for reasons of wanting to avoid potentially negative outcomes (blunting) or, alternatively, may use health services for reasons of wanting to be reassured (monitoring). Although beliefs of patients play an important role in shaping their expectancies when seeking medical help, in the past decades, the role of medical professionals has also proven to be important in affecting patient behavior. In general, studies on patient–provider communication show that doctors' communication skills have the capacity to affect patient understanding and satisfaction, adherence, and health outcomes. Consequently, the success and cost-effectiveness of health care services depends critically on the way doctors interact with patients. Models of doctor–patient relationships have identified several important factors that contribute to successful interaction, such as doctors appreciating patients' dual need to know and understand and to feel known and understood (Engel, 1977). Doctor–patient interaction models have promoted new approaches to medical decision making, such as shared decision making, involving the patient to a greater extent in what kind of treatment is required. These new approaches have contributed to better patient understanding of their medical condition and better patient recall of the information given during medical encounters. Both factors are important in increasing patient adherence to treatment regimens, which is generally low with estimates of about 50% in medication adherence and even lower in behavioral recommendations. Not following advice given by health care professionals has even been observed when nonadherence can be life-threatening (e.g., medication use after organ transplantation). Low adherence has wide implications for patients' health (e.g., unnecessary complications), population health (e.g., antimicrobial resistance in case of interrupted use of antibiotics), and health care cost-effectiveness (e.g., nonused medicines). Despite the magnitude of the problem, it has proven difficult to pinpoint the causes of nonadherence in terms of patient or provider characteristics or health care system characteristics. New promising avenues for adherence research relate to the identification of patterns of adherence, distinguishing between failure to participate in the regimen altogether, to variability in regimen conduct, or to timing errors (e.g., because of cognitive impairments in prospective memory). Moreover, it has been proposed that medical professionals should seek collaboration with patients about the best way of treatment rather than instructing them what to do. It has even been proposed that the term 'adherence' should be replaced by 'concordance' as an expression of mutual agreement about treatment between patient and health care provider. These developments align with the growing importance of self-management in illness treatment, highlighting the need for health professionals to motivate and support patients to take care of health themselves by, for example, feeling responsible for treatment on a day-to-day basis (De Ridder et al., 2008). Psychological research examines in what way doctors can communicate with patients in ways that help them to manage their illness by empowering patients to actively participate in the consultation. The importance of the patient's role is also demonstrated in research on placebo effects (health gains observed following administration of inert interventions), which can be explained by patient expectations and

anxiety reduction. Psychological research has demonstrated that stress reduction based on positive expectations of treatment efficacy and trust in a competent physician provides a powerful explanation of placebo effects.

Concluding Remarks

Health psychology research can reveal psychological processes underpinning health and illness and thereby suggest changes in health care practices designed to optimize effectiveness. Central to health psychology is the concept of health behavior, which plays a major role in health promotion, prevention of illness, illness onset, and patient behavior in a treatment setting via physiological and psychological pathways. New models of health behavior and health behavior change are being developed, moving beyond the traditional models emphasizing rational deliberate decision making about engaging in health behavior, raising several new important questions about its impulsive and automatic nature that need to be answered in future research. One particular issue deserving attention from this novel perspective is how people maintain their initial changes for health behavior and how environmental changes can be implemented that support individuals in behaving healthily. In addition, societal changes pose new questions about health behavior change, such as health behavior in the elderly and cultural issues related to health behavior in an ethnically diverse population. Answers to these questions should be informed by rigorous scientific evidence and academic debate among health psychologists working in a variety of health care practices.

See also: Chronic Illness, Psychosocial Coping with; Doctor–Patient Interaction in the West: Psychosocial Aspects; Health Behavior, Psychosocial Theories of; Health Behaviors, Assessment of; Health Behaviors; Health Promotion in Schools; Health Risk Appraisal and Optimistic Bias; Health Risk Perception; Health Self-Regulation, Motivational and Volitional Aspects of; Health Social Work; Illness Behavior and Care-Seeking; Patient Adherence to Health Care Regimens; Personalized Medicine; Social Support and Recovery from Disease and Medical Procedures; Stress, Coping and Health; Water, Health and Social Inequality.

Bibliography

Armitage, C.J., Conner, M., 2000. Social cognition models and health behavior: a structured review. Psychology and Health 15, 173–189.
Armitage, C.J., Conner, M., 2001. Efficacy of the theory of planned behavior: a meta-analytic review. British Journal of Social Psychology 40, 471–499.
Belloc, N.B., 1973. Relationship of health practices and mortality. Preventive Medicine 2, 67–81.
Bensing, J.M., Schreurs, K.M.G., De Ridder, D.T.D., Hulsman, R., 2002. Adaptive tasks in multiple sclerosis: development of an instrument to identify the focus of patients' coping efforts. Psychology and Health 17, 475–488.
Booth-Kewley, S., Friedman, H.S., 1987. Psychological predictors of heart disease: a quantitative review. Psychological Bulletin 102, 343–362.
De Ridder, D.T.D., De Wit, J., 2006. Self-regulation in Health Behavior. Wiley, Chichester, UK.
De Ridder, D.T.D., Geenen, R., Kuijer, R., Van Middendorp, H., 2008. Psychological adjustment to chronic disease. The Lancet 372, 246–255.
De Ridder, D.T.D., Schreurs, K.M.G., 2001. Developing interventions for the chronically ill: is coping a helpful concept? Clinical Psychology Review 21, 205–240.
Ditto, P.H., Druley, J.A., Moore, K.A., Danks, H.J., Smucker, W.D., 1996. Fates worse than death: the role of valued life activities in health state evaluations. Health Psychology 15, 332–343.
Engel, G.L., 1977. The need for a new medical model: a challenge for biomedicine. Science 196, 129–136.
Gollwitzer, P., 1999. Implementation intentions. Strong effects of simple plans. American Psychologists 54, 493–503.
Hofmann, W., Friese, M., Wiers, R.W., 2008. Impulsive versus reflective influences on health behavior: a theoretical framework and empirical review. Health Psychology Review 2, 111–137.
Kaplan, R.M., 2000. Two pathways to prevention. American Psychologist 55, 382–396.
Kasl, S.V., Cobb, S., 1966. Health behavior, illness behavior, and sick role behavior, I: health and illness behavior. Archives of Environmental Health 12, 246–266.
Lazarus, R.S., Folkman, S., 1984. Stress, Appraisal, and Coping. Springer, New York.
Leventhal, H., Nerenz, D.R., Steele, D.J., 1984. Illness representation and coping with health threats. In: Baum, A., Taylor, S.E., Singer, J.E. (Eds.), Handbook of Psychology and Health. Erlbaum, Hillsdale, NJ, pp. 219–252.
Levi, L., 1974. Psychosocial stress and disease: a conceptual model. In: Gunderson, E.K., Rahe, R.H. (Eds.), Life, Stress, and Illness. Thomas, Springfield, IL, pp. 8–33.
Mann, T., De Ridder, D.T.D., Fujita, K., 2013. Self-regulation of health behavior: social psychological processes to goal setting and goal striving. Health Psychology 32, 487–498.
Marteau, T.M., Hollands, G.J., Fletcher, P.C., 2012. Changing human behavior to prevent disease: the importance of targeting automatic processes. Science 337, 1492–1495.
Matarazzo, J.D., 1980. Behavioral health and behavioral medicine: frontiers for a new health psychology. American Psychologist 35, 807–817.
Rothman, A.J., Salovey, P., 1997. Shaping perceptions to motivate healthy behavior: the role of message framing. Psychological Bulletin 121, 3–19.
Schwartz, G., 1982. Testing the biopsychosocial model: the ultimate challenge facing behavioral medicine? Journal of Consulting and Clinical Psychology 50, 1040–1053.
Shaw, C., Abrams, K., Marteau, T.M., 1999. Psychological impact of predicting individuals' risk of illness: a systematic review. Social Science and Medicine 49, 1571–1598.
Sheeran, P., 2001. Intention-behavior relations: a conceptual and empirical review. In: Stroebe, W., Hewstone, M. (Eds.), European Review of Social Psychology, vol. 12. Wiley, Chichester, UK.
Smyth, J.M., 1998. Written emotional expression: effect sizes, outcome types, and moderating variables. Journal of Consulting and Clinical Psychology 66, 174–184.
Stanton, A.L., Kirk, S.B., Cameron, C.L., Danoff-Burg, S., 2000. Coping through emotional approach: scale construction and validation. Journal of Personality and Social Psychology 78, 1150–1169.
Taylor, S.E., Aspinwall, L.G., 1996. Psychosocial aspects of chronic illness. In: Costa, P.T., Vandenbos, G.R. (Eds.), Psychological Aspects of Serious Illness: Chronic Conditions, Fatal Diseases, and Clinical Care. APA, Washington, DC, pp. 7–60.
Uchino, B.N., Cacioppo, J.T., Kiecolt-Glaser, J.K., 1996. The relationship between social support and psychophysiological processes: a review with emphasis on underlying mechanisms and implications for health. Psychological Bulletin 119, 488–531.
Van der Zee, K.I., Sanderman, R., Heyink, J.W., De Haes, H., 1996. Psychometric qualities of the RAND 36-item health survey1.0: a multidimensional measure of general health status. International Journal of Behavioral Medicine 3, 104–122.
Weinstein, N.D., Klein, W.M., 1995. Resistance of personal risk perceptions to debiasing interventions. Health Psychology 14, 132–140.
Weinstein, N.D., Rothman, A.J., Sutton, S.R., 1998. Stage theories of health behavior: conceptual and methodological issues. Health Psychology 17, 290–299.

Relevant Websites

www.eaohp.org – European Academy of Occupational Health Psychology.
www.ehps.net – European Health Psychology Society.
www.health-psych.org/ – American Psychological Association Division 38: Health Psychology.
www.healthpsychology.net – The Health Psychology Network.
www.selfregulationlab.nl – Self-Regulation Laboratory, Utrecht University.

Health Psychology in Childhood

Petra Warschburger, University of Potsdam, Potsdam, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Within the field of Health Psychology, children have been rather neglected for quite sometime. Nevertheless, childhood is seen as a crucial phase of life for the establishment of stable health-relevant behaviors, and accordingly many prevention programs focus on this age group. The necessity as well as the urgency of establishing the field of 'Child Health Psychology' is emphasized even more by the fact that children's health risks have increased over the past decades.

Child Health Psychology includes an explicit consideration of developmental aspects, as not only health threats and everyday challenges change with age, but also the scope of action and the determinants of health behavior. Development-oriented Health Psychology needs to consider biological, psychological, and social changes to develop appropriate concepts for prevention and intervention.

Child Health Psychology – An Emerging New Interdisciplinary Field

Child Health Psychology is a young developing field within Health Psychology. Maddux and colleagues are considered pioneers in the area, who in 1986 called for a developmental approach to Health Psychology, especially in childhood and adolescence. This demand is justified by many factors, including the following most important ones:

- Childhood is characterized by various pronounced cognitive, physical, emotional, and social developmental processes. These processes influence a child's understanding of health and health behavior, as well as the possibilities for assuming personal responsibility.
- Parents play a pivotal role in establishing health behaviors. Child health psychology is therefore inseparably intertwined with health-related cognitions and behaviors of parents in a number of ways: First, parents serve as models for their children; second, they build the framework for the child's health beliefs; and third, they often decide for their children. With increasing age, the child's microsystem widens and other protagonists (like peers and teachers) gain in influence.
- Health Psychology focuses explicitly on prevention and health promotion. Many health-relevant behaviors are formed in childhood and remain stable into adulthood. To prevent unfavorable developments and promote adequate behaviors, intervention efforts must take place very early and include the relevant actors (depending on the child's age) to achieve sustainable effects.
- Health challenges change over the course of development and certain diseases manifest in childhood for the first time. Manifestations of chronic health problems may hinder or accelerate the development process. Moreover, health-related behavior itself is an ongoing process of change (from initiating a new behavior to stabilizing that behavior).
- For ethical reasons it is increasingly accepted that children are entitled to an appropriate explanation of treatment options and should be actively included in decision-making processes.

Hence, Child Health Psychology is an interdisciplinary basic and applied research field that addresses questions of health and illness. In this context, the focus is inevitably on the role of the family. In basic research, the determinants of child health behavior from an ecosystem/familial perspective, the psychological consequences of chronic illnesses, and the development of concepts of health and illness are examined. Applied research is based on two particularities: On the one hand, the need for development-appropriate concepts (in relation to didactics as well as to content, inclusion of the family); on the other hand, a strong focus on prevention. In this context, Child Health Psychology overlaps with the fields of Clinical Child Psychology and Pediatric Psychology. Nevertheless, while Clinical Child Psychology focuses mainly on mental disorders and their treatment, Pediatric Psychology is mainly concerned with chronic physical diseases and secondary prevention. In addition, commonalities can be found with the interdisciplinary field of developmental psychopathology that studies the onset and course of mental disorders over a lifetime. One last distinction needs to be made: The term 'Child Health Psychology' is often used for all minors, including adolescents and young adults (up to 18/21 years of age). As issues of adolescence are addressed elsewhere in this volume, this article will, in accordance with established classifications in developmental psychology, focus on the age range from birth to 12 years. However, this is not a universal, inflexible age limit: As a result of the increasingly more common acceleration of especially physical maturation processes, so-called adolescent risk behaviors (e.g., smoking, drinking, and dieting) can be found increasingly more often among younger children as well.

Developmental Tasks and Competences in Childhood

No other period of life is associated with changes as obvious as childhood. These changes are not only physical, but also emotional, social, and cognitive in nature. The concept of 'developmental tasks' describes universal, normative requirements that certain age groups face and the competences they gain. The developmental context (e.g., culture) influences rather the precise degree and timing of the developmental tasks.

Table 1 Developmental milestones according to Havighurst (1971)

Age	*Developmental milestones*
Birth to 6 years	● *Body*: Sitting – crawling – standing up – walking – learning to control the body and its wastes ● *Food*: Making the transition from liquid to solid food and from being fed to eating independently ● *Cognition*: Forming concepts about the social and physical reality (e.g., sex differences, physical laws, etc.) – preparing to learn reading ● *Interaction*: Talking – perspective taking – joint attention
6–12 years	● *Body*: Refinement of physical skills – developing healthy behaviors ● *Cognition*: Learning social roles – developing values, moral attitudes and conscience – refinement of concepts ● *Interaction*: Establishing friendships – becoming more and more independent from parents – developing positions toward social groups and institutions

Havighurst (1971) distinguishes between six major phases of development. For childhood, the first two phases, 'infancy and early childhood' (birth to 6 years) and 'middle childhood' (6–12 years) are relevant. **Table 1** lists the respective requirements and competences. The successful accomplishment of these requirements is crucial for further development. The establishment of health-related behaviors is listed as one of the developmental tasks for the 6- to 12-year age group.

Not only do the requirements change but also the possibilities to deal with these requirements. Coping in this context is understood as the use of self-regulatory competences (emotional, motivational, attentional, behavioral, cognitive, social processes) under stress and how these systems work together (Skinner and Zimmer-Gembeck, 2007). There is a broad research consensus on normal age-graded shifts in coping strategies: In infancy, mainly reflexes can be observed, while at preschool age direct actions are important; in middle childhood, cognitive strategies occur. It seems as though it is not the precise function of strategies that changes but rather their phenomenology. Thus, children of different ages mostly use the search for *support*, *problem solving* (including action), *escape*, and *distraction* coping strategies to deal with difficult situations. As the child grows older the strategies become more differentiated and their use more context-specific. Coping strategies are shaped not only by the child's individual characteristics (e.g., developmental stage, temperament), but also by the social context, which means, in childhood, mostly the parents or legal guardians/carers. Modeling/observation, parental practices and the emotional climate in the family are considered important influences.

Children's Concepts of Health and Illness

Impressive developmental changes can be observed, for example, in the context of concepts of health and illness. Piaget's approach to cognitive development is mostly used for the explanation of this idea, which postulates an association between general cognitive development and the development of health- and illness-related concepts. While this approach has met with some opposition, there is general consensus that beliefs about health and illness become more and more differentiated with children growing older, in relation to symptoms as well as to their interaction, the cause and course of illness, and treatment options (**Table 2**). In addition to children's developmental stage, their experiences with (chronic) illnesses as well as social influences (e.g., parental illness concepts) are discussed as distinctive influences on children's illness concepts (Lohaus, 2013).

Concepts of illness play an important role in coping with a disease. According to Leventhal's Common-Sense Model (Leventhal et al., 1980), people develop an organized set of subjective beliefs that guide their further actions when dealing with health threats. Especially control beliefs in relation to

Table 2 Developmental stages of illness concepts according to Biface and Walsh (1980; see Lohaus, 2013)

Stage	*Description*
I. Prelogical explanations (2–6 years)	
1. Phenomenistic	● The cause of illness is an external phenomenon which can cooccur while being temporally and spatially remote.
2. Contagion	● The cause of illness lies in things/people around the child that are close to the child but do not touch him/her. The connection happens via proximity or magic.
II. Concrete logical explanations (7–10 years)	
3. Contamination	● Distinction between cause and effect: The cause is a person, object or action outside of the child that is harmful for the body. ● The illness is caused by touching the person or object or by a dangerous action of the child.
4. Internalization	● The disease is located inside the body while the real cause (person, object) may be outside of it. The connection happens via internalization (e.g., inhaling the object). ● Illness is still a vague concept, inaccuracies occur (e.g., in relation to organs and functions).
III. Formal logical explanations (≥11 years)	
5. Physiological	● The cause of illness is seen in specific internal physiological structures and functions and can be evoked by external incidents. The cause is described as a malfunction of an inner organ or as a process of internal events culminating in illness.
6. Psychophysiological	● Illness is seen as an internal physiological process, but additional or alternative psychological causes are appreciated: It is acknowledged that inner processes like thoughts and emotions can influence bodily functions.

treatment show a consistent association with a child's self-management (Law et al., 2012). Prevention and intervention strategies should, therefore, take into account children's concepts of health and illness (and particularly their perceived controllability).

Theories of Health-Related Behaviors in Childhood and Adolescence

In Health Psychology, a plethora of models exists to explain and predict individuals' health behaviors. Most of these models are adopted from social and motivational approaches and focus on personal beliefs, intentions, and motives. These models have been developed and tested for adults and then transferred to childhood. However, to use models for children that have been developed for adults seems to be inappropriate as children are dependent on their parents' beliefs, behaviors, and decisions in many health issues (e.g., vaccination, consultations, but also health-protective behaviors like dental care). Therefore, studies often investigate parental beliefs and their influence on decisions (e.g., participation in vaccinations). In recent years, self-regulation has been demonstrated to play an important role in Health Psychology. Many health-related behaviors like eating or physical activity imply resisting habits or temptations to achieve a goal in the future. Given these considerable barriers, self-regulation theory focuses on the ways in which individuals direct and monitor their emotions and actions to achieve their (long-term) goals. Self-regulation competences involve different skills, such as regulating one's emotions, inhibiting impulses, maintaining attention, and setting or pursuing long-term goals. Self-regulation is a genuine developmental concept, since self-regulatory competences differ not only interindividually but also intraindividually. From a developmental perspective, the first 5 years of life are seen as the most important period in acquiring self-regulatory skills and the growth of these skills is a cornerstone of early child development. Social influences such as parent–child interactions are important for shaping these skills as well (Best and Miller, 2010). Studies have illustrated not only the relevance of self-regulatory skills for current health behaviors but also their significance as a predictor of morbidity and mortality in adulthood (e.g., Moffitt et al., 2011). The emergence of self-regulation theories in health psychology opens up new perspectives for development-sensitive health theories, especially for younger children.

For a comprehensive consideration of children's social context, the ecological systemic approach is often used as a theoretical framework for the design of prevention and intervention programs. Bronfenbrenner (1994) views development in terms of different social contexts and their interaction with the individual. He divides the ecological system surrounding a person into five different systems: On the microsystem level (i.e., the level of the immediate environment and its interaction with the child), parents in particular (including parenting style, health beliefs, beliefs about the child's knowledge and behavior) have to be considered. Other important settings are school and child care. On the mesosystem level, interactions between these systems are described (e.g., parents and school). The exosystem includes contexts in which the child is not directly involved but which influence the child's development nevertheless. Here, the parents' place of work or circle of friends are examples, as they codetermine, for instance, the parents' temporal resources although the child has no direct influence on them. The macrosystem refers to the norms, traditions, rules, and laws of a society. At this level the media affect, for example, the popularity of foods via advertisements. To allow for developmental changes, Bronfenbrenner introduces the chronosystem describing the temporal dimension of development with normative (e.g., school enrollment) and nonnormative (e.g., chronic illness) events and their influence on a child's development. Refining his model, and in addition to the aforementioned contexts, Bronfenbrenner later addresses the issues of personal variables (a child's characteristics, e.g., age, gender, preferences, genetic predispositions) as well as of proximal (e.g., parent–child interaction) and distal (e.g., family structure, social, and ethnic background) processes as engines of development.

For Child Health Psychology, mainly the social context and its interaction with the child and proximal processes like the parent–child interaction are important. In the sense of a 'meta-health theory,' that is, the concept that parents have an organized set of beliefs, thoughts, and feelings about their own and their child's health and health-relevant behaviors, the interdependencies between children and parents should be considered. For example, the parental perception of the health risks associated with being overweight influences their willingness to participate in prevention programs (Warschburger and Kröller, 2012). In summary, models integrating the child's socioeconomic context (and here mainly the parental influences) along with their developmental stage are still lacking.

Figure 1 illustrates the draft of an integrative approach considering the most important institutions of socialization in childhood as well as a development-oriented view.

Children's Health Status

Childhood is seen as the period of life with the best health status. Nearly 80% of all children between 11 and 15 years rate their health as excellent or very good. With increasing age and lower socioeconomic status, the percentage of children rating their health negatively increases. Additionally, gender differences become more pronounced with increasing age and especially older girls rate their health more negatively and report more health issues than boys of the same age (Currie et al., 2012). This is in accordance with data on health-related quality of life of children. The importance of the family environment can be observed as well: children from families with higher educational background and better financial resources report a higher quality of life (von Rueden et al., 2006).

In childhood, acute illnesses still top the list of health problems with manifold causes of illness. Most common are respiratory infections (like episodes of the common cold, bronchitis, and tonsillitis) and gastrointestinal infections. Communicable childhood diseases like pertussis, measles, mumps, rubella, or scarlet fever play a minor role. Chronic physical illnesses can be found less often in children than in adults, yet still 15–20% of all children are concerned (Kamtsiuris et al., 2007). In a German representative sample, the most prevalent chronic conditions were overweight and obesity (14.8%), obstructive bronchitis

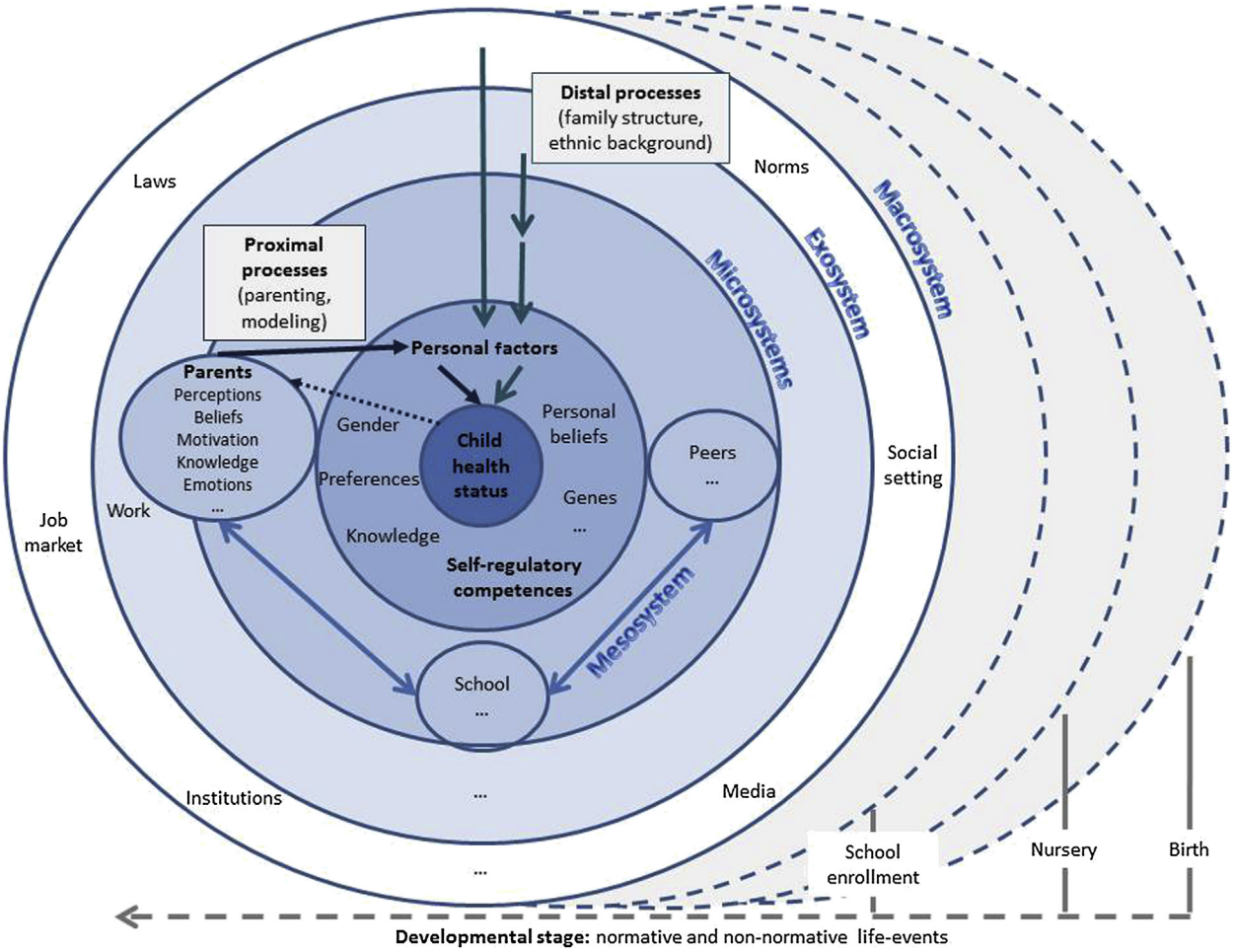

Figure 1 Child health model on the basis of Bronfenbrenner's (1994) ecological systemic approach. Bronfenbrenner, U., 1994. Ecological models of human development. In: Postlethwaite, T., Husen, T. (Eds.), The International Encyclopedia of Education, second ed. Pergamon, Oxford, pp. 37–43.

(13.3%), atopic dermatitis (13.2%), hay fever (10.7%), scoliosis (5.2%), and asthma (4.7%) (Kamtsiuris et al., 2007; Kurth and Schaffrath-Rosario, 2010). For some chronic physical illnesses, a pronounced increase in prevalence rates over the past decades has been observed (Roberts, 2003). In addition to physical conditions, the comparatively high rate of mental problems in childhood must be noted, with attention-deficit hyperactivity disorder, mood disorders, and conduct disorders ranking first (Merikangas et al., 2010).

The increasing importance of chronic illnesses can also be observed when looking at mortality data (cf **Table 3**). Given the improved treatment methods of the past century, acute infections play a minor role, while external causes (e.g., traffic accidents, suicides) as well as chronic illnesses (e.g., cancer) come to the fore.

Prevention and Health Promotion

Childhood is considered the phase of life in which health-relevant behaviors are established and then often maintained into adulthood. A balanced diet and regular physical activity, for example, not only contribute to increased well-being, but also help to prevent various diseases (e.g., diabetes and obesity). These important health behaviors are, therefore, discussed in more detail in the following.

Children's Dietary Behavior

The course for the formation of food preferences and healthy nutrition is set already during pregnancy and breast-feeding. The composition of the mother's diet affects the baby's taste preferences (Mennella et al., 2001). Furthermore, certain taste preferences like those for sweetness (as a sign of energy-dense food) and the aversion to bitter tastes (as a potential sign for toxic substances) are congenital. While it is assumed that babies dispose of an excellent physiological control of food intake, learning processes become more and more important with increasing age. For example, children develop taste preferences via repeated exposure as well as via model (parents, peers, media) or associative learning (e.g., food as a gratification strategy). Not only taste preferences but also knowledge, social norms, and attitudes regulating dietary

Table 3 Causes of death for children between 1 and 15 years old in Germany in 2010 (Ellsäßer, 2012)

Rank	*1–5 years*	*5–10 years*	*10–15 years*
1	Congenital and chromosomal abnormalities (18.6%)	Tumors (26.7%)	External causes (29.6%) ● Transportation accidents (11.4%) ● Deliberate self-harm (5.3%) ● Drowning and assault (1.8%) ● Falls and smoke, fire, flames (1.0%)
2	External causes (17.9%) ● Transportation accidents (4.5%) ● Drowning (2.8%) ● Falls (1.9%) ● Smoke, fire, flames (1.3%) ● Assault (0.9%)	Malignant tumors (24.4%)	Benign tumors (26.8%)
3	Benign tumors (15.1%)	External causes (17.7%) ● Transportation accidents (8.2%) ● Drowning and assault (1.9%) ● Smoke, fire, flames (1.6%)	Malignant tumors (25.1%)
4	Malignant tumors (14.3%)	Congenital and chromosomal abnormalities (14.8%)	Diseases of the nervous system (10.6%)
5	Diseases of the nervous system (12.7%)	Diseases of the nervous system (12.9%)	Congenital and chromosomal abnormalities (8.6%)

intake are shaped on the basis of positive sensory and affective experiences. Here, psychological processes play an important role (Ogden, 2010). Parents try to influence their child's dietary behavior in many different ways. On the one hand, in a very direct way by offering or prohibiting certain foods themselves; on the other hand, in a more indirect way by eating certain foods. In this context, home availability and accessibility, parental modeling and parents' own intake as well as family rules and parental encouragement proved to be positively associated with children's fruit and vegetable consumption. These associations were especially strong in childhood up to 12 years, while the parental influence decreased in adolescence. Moreover, the entire parenting context is important as well: children whose parents used an authoritative parenting style consumed more healthy foods (Warschburger, 2011).

In terms of food intake, a trend toward less healthy eating choices (with the exception of sweets) can be noticed with increasing age: fruits and vegetables are consumed less often while soft drinks, meat, and fast food are consumed more often. While 60% of all children from 3 to 6 years of age eat fresh fruit at least on a daily basis, this proportion decreases to 40% (boys) and 50% (girls) at the age of 11–13 years. These numbers are considerably lower than nutrition guidelines advise and emphasize the deficiencies in children's dietary behaviors already present (Mensink et al., 2007). Not only the choice of foods changes over time, but also specific eating behaviors (e.g., emotional or restrictive eating). With increasing age, the percentage of emotional eating behaviors and the prevalence of weight and figure concerns increases as well. While in infancy mostly feeding disorders can be observed, school children show avoidant, selective, and restrictive eating behaviors and in adolescence 'classical eating disorders' like anorexia and bulimia nervosa come to the fore. Recent longitudinal studies show that, despite the changes in the phenomenology of problematic eating behaviors, a considerable stability can be assumed (Ogden, 2010).

Physical Activity

Sufficient physical activity is considered essential for a child's development and physical and mental health. Physical activity is defined as any sort of bodily movement generated by muscular activity above the normal energy demand. This includes leisure activities (like rollicking or catch games), daily activities (like helping in the garden or on the way to school) but also exercise with performance and competition aspects. The World Health Organization (2011) recommends 60 min of daily physical activity at moderate or vigorous intensity for children over 5 years of age. For younger children, up to 2 h of daily physical activity in the form of active games are recommended.

Traditionally, childhood is seen as a phase of life with a high level of physical activity. However, the precise level of physical activity in children is difficult to measure due to rapid and frequent changes between activity and inactivity. According to national representative data, three-quarters of all children 3–10 years of age get daily exercise and one-third at least three times a week – without considering the physical activity occurring in daily life. However, this quite positive picture changes for the worse with increasing age, especially when entering adolescence. Only 15.3% of all children between 4 and 17 years fulfill the WHO recommendations for physical activity; at 29.2% the prevalence among children 4–10 years old is much higher compared with the 8.9% among 11- to 13-year-old children (Krug et al., 2012). The low prevalence of physical activity is of high relevance for health issues, as a high rank stability of the degree of physical activity from childhood to adulthood can be found (Craggs et al., 2011).

What factors affect the amount of physical activity children engage in? The correlates are manifold and the results inconclusive. As already discussed for dietary behavior, parents play a key role in children up to 12 years with parental model behaviors and parental support influencing the child's activity level moderately; from adolescence on, this influence is

negligible. In addition, the school environment, neighborhood characteristics, and traffic safety prove to be relevant too (de Vet et al., 2011). At an individual level, self-efficacy is correlated with a lesser decline of physical activity with increasing age (Craggs et al., 2011).

Prevention Programs

These results emphasize the fact that already in childhood a high percentage of children show unhealthy dietary and physical activity patterns and, therefore, carry a higher risk for the development of health problems. Especially the stability of these behaviors highlight the fact that childhood is a very favorable time to establish primary prevention programs and to prevent stable undesirable progress. Universal approaches to promoting health in the general population take place in settings like school or day care. As indexed (e.g., children of overweight parents) or selective (e.g., overweight children) approaches, they address selected risk groups. The line up of methods is broad and ranges from structural/environmental (e.g., providing drinking water, rearranging the school yard) to behavioral prevention (e.g., food education). Reviews and meta-analyses emphasize the effectiveness of prevention concepts related to nutrition, physical activity, or both in preventing childhood obesity. In addition, it is important to include the school context (e.g., availability and supply of food in vending machines) as well as parental influence. Furthermore, the bidirectional influences in the parent–child relationship when realizing a healthy way of life should be considered (Waters et al., 2012). In other health areas like prevention of substance abuse or mental disorders, positive effects of prevention programs in childhood and adolescents can be achieved, while success rates vary depending on the extent of the program, the target behavior, the integrated methods, and the age of the children included (Carr, 2002). Key elements for effective prevention programs include a comprehensive approach considering all relevant contexts and starting points as well as an extensive, didactically multifaceted, and age-appropriate program with a focus on strengthening positive parent–child and peer relationships (Nation et al., 2003).

Coping with Chronic Illnesses

Despite the overall excellent health status of children, a considerable percentage of children suffer from chronic disease. A number of chronic illnesses have their incidence peak in childhood (e.g., atopic dermatitis, asthma, tumors of the central nervous system, and functional abdominal pain). Diseases are considered chronic if they last for at least 1 year and lead to functional impairments like repeated consultations or hospitalizations, dependence on regular treatments or impairments in daily activities. Hence, coping with the disease and treatment demands become important (biopsychosocial view).

Chronic childhood illness is often considered a nonnormative, additional stressor in the living environment of the whole family. Thus, chronic illness can not only negatively affect the accomplishment of childhood developmental tasks (e.g., becoming more independent from parents, building relationships to peers), but can also impose – as already mentioned – disease-specific demands like adhering to preventive and curative therapeutic regimes (e.g., avoiding certain environments, regular treatment), confrontation with a potentially reduced life expectancy and the integration of the disease into the self. Chronic illnesses, therefore, pose grave demands on the coping competences of children and their families. While results from many empirical studies suggest that children and families cope astonishingly well with these additional demands, chronic illnesses can nevertheless be considered a risk factor for the development of mental disorders, with an assumed two- to threefold increased risk for affected children. Notwithstanding the importance of mental disorders, their exclusive consideration is not sufficiently sensitive to the cumulative strains experienced in daily life. Hence, attention to health-related quality of life assesses the subjective view of different everyday functions like emotional well-being, social aspects, and physical activity. It appears that chronic illnesses are associated with a reduced quality of life (Varni et al., 2007). Overall, boys, children with migration background, and those with a lower socioeconomic status more often show negative consequences of chronic (mental and physical) illnesses (Mohler-Kuo and Dey, 2012).

In parents, different sources of strain must be mentioned as well, like the question of causes/possible guilt, worries about the child and financial impairments, but also the additional organizational, temporal, and financial efforts. In treatment, furthermore, they are the main contact for the medical staff as children are not allowed to independently decide on their treatment options. Empirical studies emphasize the strong interdependence of parental and child coping and adaptation reactions (cf Wallander et al., 2009).

Patient education programs – in childhood better called family education programs – have been established to support those coping with chronic illnesses. Their goal is not only to increase adherence to treatment regimes, but also to reduce symptoms and psychological strain. Here, children (and their parents) learn how to handle their disease autonomously. Especially children should – according to their age – be actively integrated into treatment (participative approach). Programs incorporate knowledge about the disease and therapeutic skills, but also about building emotional and social competences to deal with the disease. It has been demonstrated in various studies that the integration of psychological methods into the treatment of chronic illnesses is a very promising approach with medium to large effect sizes, depending on the disease (Graves et al., 2009; Kirk et al., 2011). Once again, the involvement of parents seems to be important. In obesity treatment for children up to 12 years, behavioral approaches that include the parents are more promising than those that do not. Relevant aspects in working with parents in the context of weight reduction programs prove to be self-management strategies (e.g., self-monitoring, self-instruction, goal setting) but also the identification of barriers to the implementation of new behaviors at home, the transfer of knowledge about gratification strategies, and the practice of new behaviors. Especially with younger children, first results suggest that parental trainings alone are even more successful than combined parent–child trainings (Warschburger, 2011).

Conclusion

In many basic research fields of Psychology, developmental approaches have been established. These findings should be better integrated into Health Psychology. So far, a development-oriented perspective has not been realized/implemented in Health Psychology. However, Child Health Psychology can be considered an emerging field within the scope of Health Psychology. Childhood is an important stage in life for the establishment of health-relevant behaviors and, hence, for prevention and intervention programs. These programs have to be tailored to the specific characteristics of childhood to unfold their full potential. This contribution discusses the challenges of a development-oriented approach.

See also: Caregiver Burden; Chronic Illness, Psychosocial Coping with; Health Education and Health Promotion; Health Psychology; Physical Activity and Health; Self-Efficacy and Health.

Bibliography

Best, J.R., Miller, P.H., 2010. A developmental perspective on executive function. Child Development 81 (6), 1641–1660.

Bronfenbrenner, U., 1994. Ecological models of human development. In: Postlethwaite, T., Husen, T. (Eds.), The International Encyclopedia of Education, second ed. Pergamon, Oxford, pp. 37–43.

Carr, A. (Ed.), 2002. Prevention: What Works with Children and Adolescents?: A Critical Review of Psychological Prevention Programmes for Children, Adolescents and Their Families. Taylor & Francis, New York.

Craggs, C., van Sluijs, E.M.F., Corder, K., Panter, J.R., Jones, A.P., Griffin, S.J., 2011. Do children's individual correlates of physical activity differ by home setting? Health & Place 17 (5), 1105–1112.

Currie, C., Zanotti, C., Antony, M., Currie, D., de Looze, M., Roberts, C., Samdal, O., Smith, O., Barnekow, V., 2012. Social Determinants of Health and Well-being among Young People: Health Behaviour in School-Aged Children (HBSC) Study: International Report from the 2009/2010 Survey. World Health Organization, Geneva.

Ellsäßer, G., 2012. Unfälle, Gewalt, Selbstverletzung bei Kindern und Jugendlichen 2012. Ergebnisse der amtlichen Statistik zum Verletzungsgeschehen 2010. Fachbericht. Statistisches Bundesamt, Wiesbaden.

Graves, M.M., Roberts, M.C., Rapoff, M., Boyer, A., 2009. The efficacy of adherence interventions for chronically ill children: a meta-analytic review. Journal of Pediatric Psychology 35 (4), 368–382.

Havighurst, R.J., 1971. Developmental Tasks and Education, third ed. Longman, New York.

Kamtsiuris, P., Atzpodien, K., Ellert, U., Schlack, R., Schlaud, M., 2007. Prävalenz von somatischen Erkrankungen bei Kindern und Jugendlichen in Deutschland. Ergebnisse des Kinder- und Jugendgesundheitssurveys (KiGGS). Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz 50, 686–700.

Kirk, S.F.L., Penney, T.L., McHugh, T.L., Sharma, A.M., 2011. Effective weight management practice: a review of the lifestyle intervention evidence. International Journal of Obesity 36 (2), 178–185.

Krug, S., Jekauc, D., Poethko-Muller, C., Woll, A., Schlaud, M., 2012. Zum Zusammenhang zwischen körperlicher Aktivität und Gesundheit bei Kindern und Jugendlichen. Bundesgesundheitsblatt. Gesundheitsforschung. Gesundheitsschutz 55 (1), 111–120.

Kurth, B.-M., Schaffrath-Rosario, A., 2010. Übergewicht und Adipositas bei Kindern und Jugendlichen in Deutschland. Bundesgesundheitsblatt 53, 643–652.

Law, G.U., Tolgyesi, C.S., Howard, R.A., 2012. Illness beliefs and self-management in children and young people with chronic illness: a systematic review. Health Psychology Review, 1–19.

Leventhal, H., Meyer, D., Nerenz, D., 1980. The common sense representation of illness danger. In: Rachman, S. (Ed.), Medical Psychology. Pergamon Press, New York, pp. 7–30.

Lohaus, A., 2013. Kindliche Krankheitskonzepte. In: Pinquart, M. (Ed.), Wenn Kinder und Jugendliche körperlich chronisch krank sind. Psychische und soziale Entwicklung, Prävention, Intervention. Springer, Berlin, Heidelberg, pp. 17–32.

Maddux, J.E., Roberts, M.C., Sledden, E.A., Wright, L., 1986. Developmental issues in child health psychology. American Psychologist 41 (1), 25–34.

Mennella, J.A., Jagnow, C.P., Beauchamp, G.K., 2001. Prenatal and postnatal flavor learning by human infants. Pediatrics 107 (6), e88.

Mensink, G.B., Kleiser, C., Richter, A., 2007. Lebensmittelverzehr bei Kindern und Jugendlichen in Deutschland. Ergebnisse des Kinder- und Jugendgesundheitssurveys (KIGGS). Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz 50, 609–623.

Merikangas, K.R., He, J.P., Brody, D., Fisher, P.W., Bourdon, K., Koretz, D., 2010. Prevalence and treatment of mental disorders among US children in the 2001–2004 NHANES. Pediatrics 125 (1), 75–81.

Moffitt, T.E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R.J., Harrington, H.L., Houts, R., Poulton, R., Roberts, B., Ross, S., Sears, M., Thomson, M., Caspi, A., 2011. A gradient of childhood self-control predicts health, wealth, and public safety. Proceedings of the National Academy of Sciences of the United States of America 108, 2693–2698.

Mohler-Kuo, M., Dey, M., 2012. A comparison of health-related quality of life between children with versus without special health care needs, and children requiring versus not requiring psychiatric services. Quality of Life Research 21 (9), 1577–1586.

Morris, A.S., Silk, J.S., Steinberg, L., Myers, S.S., Robinson, L.R., 2007. The role of the family context in the development of emotion regulation. Social Development 16 (2), 361–388.

Nation, M., Crusto, C., Wandersman, A., Kumpfer, K.L., Seybolt, D., Morrissey-Kane, E., Davino, K., 2003. What works in prevention. Principles of effective prevention programs. American Psychology 58 (6–7), 449–456.

Ogden, J., 2010. The Psychology of Eating, second ed. Wiley, Chichester.

Roberts, M.C., 2003. Handbook of Pediatric Psychology, third ed. Guilford Press, New York.

von Rueden, U., Gosch, A., Rajmil, L., Bisegger, C., Ravens-Sieberer, U., the European KIDSCREEN Group, 2006. Socioeconomic determinants of health related quality of life in childhood an adolescence: results from a European study. Journal of Epidemiology and Community Health 60, 130–135.

Skinner, E.A., Zimmer-Gembeck, M.J., 2007. The development of coping. Annual Review of Psychology 58, 119–144.

Varni, J.W., Limbers, C.A., Burwinkle, T.M., 2007. Impaired health-related quality of life in children and adolescents with chronic conditions: a comparative analysis of 10 disease clusters and 33 disease categories/severities utilizing the PedsQL™ 4.0 Generic Core Scales. Health and Quality of Life Outcomes 5, 43–57.

de Vet, E., de Ridder, D.T.D., de Wit, J.B.F., 2011. Environmental correlates of physical activity and dietary behaviours among young people: a systematic review of reviews. Obesity Reviews 12 (5), e130–e142.

Wallander, J.L., Taylor, W.C., Grunbaum, J.A., Franklin, F.A., Harrison, G.G., Kelder, S.H., Schuster, M.A., 2009. Weight status, quality of life, and self-concept in African American, Hispanic, and white fifth-grade children. Obesity (Silver Spring) 17 (7), 1363–1368.

Warschburger, P., 2011. Psychologische Aspekte der Adipositas. Konsequenzen für Therapieinhalte, Therapieindikation und Therapieerfolg. Bundesgesundheitsblatt 54, 562–569.

Warschburger, P., Kroeller, K., 2012. Childhood overweight and obesity: maternal perceptions of the time for engaging in child weight management. BMC Public Health (Online Journal) 12 (1), 295.

Waters, E., de Silva Sanigorski, A., Hall, B.J., Brown, T., Campbell, K.J., Gao, Y., Armstrong, R., Prosser, L., Summerbell, C.D., 2012. Interventions for preventing obesity in children (review). Cochrane Collaboration 12, 1–212.

World Health Organization, 2011. Global Recommendations on Physical Activity for Health. World Health Organization, Geneva.

Relevant Websites

http://www.apadivisions.org/division-54/ – American Psychological Association Division 54: Society of Pediatric Psychology.

http://www.apa.org/pi/families/children-mental-health.aspx – American Psychological Association Children's Mental Health.

http://www.healthychildrenhealthyfutures.org/ – Healthy Directions: Healthy Children Healthy Futures Programs.

http://www.health-psych.org/index.cfm – American Psychological Association Division 38: Health Psychology.

Health-Related Support and Self-help Mutual Aid Groups

Benjamin H Gottlieb and Karla D Stroud, University of Guelph, Guelph, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Face-to-face self-help and support groups provide opportunities to meet and interact with individuals who share a common affliction, disorder, life circumstance, noxious habit, or other adversity, and represent a highly specialized personal community that can mitigate feelings of loneliness, stigma, and stress. The past two decades have seen a dramatic increase in the number of online arrangements for the exchange of support and information about health issues, including chat rooms, bulletin boards, forums, newsletters, and groups on the Internet. This article spotlights recent research on the processes, outcomes, and limitations of these peer support systems.

In the 15 or so years since the original article to this encyclopedia was written, the pace of research on health-related support and self-help groups has quickened. This is largely attributable to the proliferation of online groups, a topic that received only a single paragraph at the end of the original article. Recent years have seen countless people turn to their computers to meet other arthritis sufferers, stroke and heart attack victims, substance abusers, and people with such mental health problems as anxiety and eating disorders (Finn and Steele, 2010). MDjunction.com lists 102 online support groups (OSGs) for mental health topics, including highly specialized groups, like those for people interested in bipolar disorders; there are online venues for teenagers affected by that disorder, for those with alcoholism and anxiety, parents of children with the disorder, pregnant ladies, and those dealing with PTSD concurrently.

Of course, the same medium has been exploited for more nefarious and disturbing purposes, such as sites that promote suicide (Golant et al., 2003), exploit people recovering from sexual offenses (Kernsmith and Kernsmith, 2008), and those that encourage eating disorders (pro-ana and pro-mia). For example, the latter sites depict eating disorders as desirable, describe crash dieting techniques and recipes, coach users on using socially acceptable pretexts for refusing food, such as veganism, hold weight loss competitions, advise on how to best induce vomiting and use laxatives and emetics, and share tips on hiding weight loss from parents and doctors (http://en.wikipedia.org/wiki/Pro-ana). It would not be overreaching to declare that, functionally, these sites offer a support group or network to their devotees. Moreover, the many names that are used to market social support on the Web make it hard to know whether to trust the information and advice offered, much less whether the support being offered has been vetted by any authority who knows the difference between fact and myth and prosocial vs antisocial messages. Hence, as discussed later, there is an urgent need for some form of certification of Websites or quality standards to help consumers seeking support to distinguish between legitimately moderated sites and self-styled 'fly by site' URLs.

Online Support Groups

A recent national US survey of more than 3000 individuals conducted by the prestigious Pew Research Center's Internet and American Life Project (2011) testifies to the growing use of the Internet for peer information and support related to health matters. Although the survey did not inquire specifically about participation in groups, the findings illuminate the kinds of resources sought from peers online and the unique attractions of this medium. They reveal that peers are sought mainly for emotional support, tips on coping and how to communicate with health care professionals, and when a speedy solution for an everyday health issue is needed. The survey also spotlighted the very high use of the Internet among those with rare conditions and chronic disorders (Murray et al., 2005). For example, of those consulting the Internet for a chronic health disease, almost one-fourth were people living with high blood pressure, diabetes, heart conditions, lung conditions, cancer, or another chronic ailment. An even greater proportion of online users with a rare condition accessed peers for advice, counsel, and support; they simply had to reach beyond their existing social ties and even their associates' ties to find people in the same boat or shipwreck. Readers seeking the most comprehensive listings of OSG studies are referred to a chapter by Finn and Steele (2010) and an excellent online bibliography by Azy Barak (http://construct.haifa.ac.il/~azy/refsupp.htm).

The proliferation of Internet social resources for physical and mental health disabilities and concerns has also resulted in the founding of new journals, making it easier for researchers to learn how to conduct methodologically and ethically sound Internet intervention studies. As one example, the *Journal of Medical Internet Research* has published a systematic review of OSGs for depression, noting that this psychiatric disorder is common among people with serious, life-threatening, and chronic diseases, not just people experiencing psychosocial stressors (Griffiths et al., 2009). As most literature reviews observe methodological weaknesses, such as the absence of a control or comparison group and failure to control for cointerventions in a multicomponent program, precluded any firm conclusions about the effectiveness of the group support. Even the strongest conclusion drawn on the basis of the 31 reports (whose characteristics and outcomes are displayed in two tables) was couched in highly tentative language: "Overall, studies of breast cancer ISGs were more likely to report a reduction in depressive symptoms than studies of other ISG types (Fisher $P = 0.02$), but it is possible that this finding was due to confounding design factors rather than the nature of the ISG" (p. e30).

More troublesome, however, is the inclusion of diverse structures and formats of group support in the review. Specifically, it includes such diverse online arrangements as a 12-week structured 'newsgroup' for women with breast cancer, moderated by a psychologist (Winzelberg et al., 2003), a 16-week 'chatroom' for women with breast cancer that was hosted by a professional and also offered continuous access to a bulletin board but lacked a control group (Lieberman et al., 2003), a strictly dyadic intervention involving 4–8 weeks of online chat sessions with the same anonymous partner (Shaw and Gant, 2002), and participation in public listservs/bulletin boards for at least a 12-month period (Houston and Allison, 2002). Still other Internet support group (ISG) formats were mailing list/newsgroups, instant messaging, and other combinations of Internet social contacts. Clearly, such polymorphous entities as those included in this review underscore the need for a common nomenclature for the diverse 'online communities' that marshal, augment, or specialize social support. Although there are ISGs that mimic face-to-face groups, such as the ISG implemented by Winzelberg et al. (2003) that is presented later, there is great variety in online groups and communities, with few guideposts for novice consumers.

More generally, there is a need for clearer standards or guidelines for conducting and reporting Internet intervention research. Several recent documents have aimed to provide such quality control (Rippen and Risk, 2000; Risk and Dzenowagis, 2001). One sterling example is Proudfoot et al.'s (2011) contribution that begins with a 'mapping sentence' that lays out the domain of inquiry and then sets out the 'facets' of the topic being mapped as well as their 'elements,' and a guideline arising from each facet. The sentence employed for online intervention research is somewhat tortuous but inclusive: "An Internet-based intervention is a self-guided or human-assisted program for health promotion, disease and emotional distress prevention, risk factor management, treatment, or relapse prevention that is executed by means of a prescriptive online intervention and/or communication operated through the Internet and used by consumers and patients seeking health- and mental health-related assistance" (p. 86). Examples of the 12 facets produced from this sentence include the target population, the model of change, the type and dose of the intervention, and the program interactivity and degree of synchronicity. An example of a guideline for the latter facet is: "The synchronicity of support and feedback are clearly reported within the program as well as the duration between program sessions, where relevant" (p. 90). These and other guidelines yet to be written for best practices in online intervention research await the views and potential endorsement of individual investigators and professional associations.

Eysenbach et al. (2004) also conducted a review of peer-to-peer online support initiatives, defining their object of study more clearly than the preceding review but nevertheless leaving much ambiguity about the phenomena of interest. They state that they searched for interventions involving "a virtual community (or had a virtual community component), defined for the purpose of this study as a group of individuals with similar or common health related interests and predominantly non-professional backgrounds (patients, healthy consumers, or informal caregivers) who interact and communicate publicly through a computer communication network such as the internet, or through any other computer based tool (including non-text based systems such as voice bulletin board systems), allowing social networks to build over a distance" (p. 1166). Of the 45 publications that met their inclusion criteria, they report that only 6 exclusively involved moderated peer-to-peer interaction over time, the rest being multicomponent interventions. The findings for the six 'pure' ISGs are virtually identical to the review by Griffiths et al. (2009), namely that shortcomings in the study designs, small numbers of participants, and lack of information about the clinical significance of small group differences render the evidence inconclusive. Happily, the review found none of the ill-effects reported for ISGs elsewhere (Houghton et al., 2001; Kelly et al., 2002).

Granted that there is much definitional slippage surrounding the use of the interchangeable ISG and OSG labels, there are nonetheless a few examples of tightly designed Internet studies that bear a strong similarity to face-to-face support groups, except that interaction is not synchronous. One example is Winzelberg et al.'s (2003) study of three 12-week, structured and professionally moderated ISGs composed of 72 women with primary breast cancer. Like a typical face-to-face support group, a mental health professional introduces different topics each week, calling on the participants to share their thoughts and swap their experiences, with time set aside for each group member to 'check in' whenever she logged on, to share a personal profile of herself, and to maintain a private diary of her own reflections. Members of this ISG called Bosom Buddies could also see when their 'buddies' logged on, and were encouraged by the facilitator to exchange support and stay on topic. The findings proved gratifying; at posttest, the effect of the immediate ISGs was to significantly reduce participants' scores on the Center for Epidemiological Studies Depression scale, and to lower levels of perceived stress, and cancer-related trauma, with all effect sizes in the moderate range. Equally important, despite their inability to see and communicate in real time with their 'buddies,' the participants formed close relationships with one another that involved sharing family events and ceremonial occasions (e.g., birthdays and anniversaries), and even expressed concern when they had not heard from a member for longer than usual. Some participants developed off-line relationships, presumably integrating their buddies into their natural social network, thereby further sustaining their mutual access to support. Perhaps future ISGs could aspire to such beneficial social side effects when the participants are geographically proximal enough to affiliate with one another.

Yet questions arise about the external validity of this particular ISG and about the filtering that occurred at the time of study intake. On the latter score, we learn that, of the 151 women who replied to the advertisement for the ISG, 79 did not enroll, of whom 39 or approximately 50% decided not to participate once they were told the details of the study. Access to computers was not an issue because they were provided to those who needed one. So the question is why so many did not enroll, and the answers would be instructive for future trials of ISGs. Was it fear of the unknown or perhaps fear that the group would raise subjects that the women preferred not to consider, or that it would attract members who pose new threats by virtue of their worse prognoses or superior coping behaviors? As it turns out, those who did enroll were a highly educated

sample; 64% had at least a baccalaureate degree and another 28% had attended some college. In future reports of these initiatives, it would be highly informative if the investigators were to collect information about the characteristics of those who decline participation and solicit their reasons for opting out. Given the flexibility of asynchronous ISGs that allow users to go online whenever it is convenient or when the need for contact with others arises, time constraints are not as likely to be cited as an obstacle to group participation as they are for scheduled, face-to-face group meetings.

More generally, questions are being raised about the reach and representativeness of people who participate in online support programs and other types of computer-mediated public health programs, leading Bennett and Glasgow (2009) to recommend targeting high-risk groups such as older adults, people in rural areas, and people with chronic conditions that require careful self-surveillance. In addition, by way of standard setting, they recommend that Internet interventionists report the size of the target population, the size of those who were invited to participate, the proportion who are eligible, and of those, the proportion who participate, and finally, how representative the final sample is. To this must be added the attrition rate because there is growing evidence that a significant proportion of those who enter online trials lose interest and quit (Eysenbach, 2005).

Another relatively recent development in the evaluation of online support is the inclusion of measures that assess consumers' psychological sense of control over their own health care, health-related decision making, and ultimately their own health status. Investigation of the empowering effects of self-administered programs has yielded insights into the virtues of being accepted and validated by peers, being useful to others, gaining benefit from the very process of writing about one's experience and processing the accompanying effect, and reducing self-stigma by virtually meeting people from diverse cultures and geographic locations who share and have overcome similar plights (Barak et al., 2010). A fascinating report of both empowering and disempowering processes that unfold in OSGs not only substantiates the preceding empowering processes, but also documents the perceived outcomes reported in interviews with ISG participants (van Uden-Kraan et al., 2008). Specifically, after contacting the Webmasters of Dutch ISGs to ask them if the members could be informed of a study about their ISG experience, 32 participants in groups for people with breast cancer, fibromyalgia, or arthritis consented and were interviewed. Based on a thematic analysis, the outcomes they collectively attributed to their group experience include: "being better informed; feeling confident in the relationship with their physician, their treatment, and their social environment; improved acceptance of the disease; increased optimism and control; enhanced self-esteem and social well-being; and collective action" (p. 405). Moreover, a study of empowerment among people living with HIV/AIDS employed structural equation modeling to examine the downstream effects of empowerment processes involving the receipt and provision of informational and emotional support, along with the positive meaning gleaned from the group experience (Moa and Coulson, 2012). The findings show that these processes selectively lead to both a higher level of self-care self-efficacy (confidence in one's ability to look after one's own health) and more adaptive coping, which lead in turn to higher life quality.

A final observation about online support resources is that their sheer abundance and variety make it very difficult to distinguish professionally moderated groups from mutual aid self-help (MASH) groups. Indeed, as the authors of this encyclopedia article, we were often challenged by insufficient detail about the format, leadership, eligibility criteria, and duration of the online groups and communities that were the subject of study to classify them as support groups or self-help groups. And if we could not tease the two apart, imagine the quandary facing the naïve Internet user who is searching for a supportive personal community. It follows that, in the future, it will be necessary to provide greater detail and clearer definitions of these online resources, with sites hosted by spokespersons who can answer questions about what each site does and does not offer and how it goes about providing support.

MASH Groups

The principal distinction that is made when support groups and self-help groups are compared is that the latter do not employ professional facilitators, do not limit the length of participation, and generally rely on mutual aid and a structured program of activities or steps to help their members. In Borkman's (1976) terms, self-help groups draw on their members' experiential knowledge, not expert information or training. Recent developments in research on self-help groups are reviewed below, beginning with selected studies of take-up and attrition, followed by investigations of the process and outcomes of MASH groups.

Attraction and Attrition

Few studies have investigated attrition and retention factors within the context of self-help groups. Although self-help groups have an indefinite potential duration, not all groups survive over the course of months and years. Given the potential benefits of self-help groups, it is important to identify the factors that attract members and those that influence attrition. Randomly selecting self-help groups addressing a variety of concerns (except for anonymous groups such as Alcoholics Anonymous) from the registration database of the Self-Help Network of Kansas, Wituk et al. (2002) identified characteristics that distinguished between active and disbanded groups. When compared to groups that had disbanded, active groups had a significantly higher number of new members within the past year, a significantly higher average meeting attendance, had been in existence for a significantly longer period of time, had a greater number of people participating in leadership roles in groups, had membership recruitment and maintenance techniques, and received funding from local agencies.

If a higher number of new members within the past year contributes to the survival of self-help groups, recruitment should concentrate on people who match the characteristics of those who join self-help groups. What limited research on this topic exists reveals some of the social and psychological characteristics of cancer patients who choose to join online

self-help groups (Hobye et al., 2010). Individuals taking part in a cancer rehabilitation course were invited to participate in an online self-help group. Those who chose to participate in the online self-help group were generally younger, with a higher socioeconomic status, and married or cohabitating. There were no significant gender differences, and users and nonusers did not differ significantly on measures of psychological well-being and quality of life. Of special interest, cognitive avoidance as a coping style also discriminated between users and nonusers, with nonusers endorsing the use of cognitive avoidance significantly more often. This suggests that cancer patients who tend to employ forms of coping that reflect disengagement from stressors are unlikely to both join and benefit from groups that involve disclosure of threatening feelings and thoughts. However, as this was a study of an Internet self-help group, it is not yet determined whether these distinguishing characteristics are generalizable to face-to-face self-help groups.

Double Trouble in Recovery (DTR) groups are dual-focus 12-step self-help groups for those with substance addiction and a comorbid mental health disorder. They have been the subject of much evaluation, including a study of predictors of retention of members recruited from 24 DTR meetings over a 1-year period (Laudet et al., 2004). Significant predictors included: "older age, greater number of lifetime arrests, abstinence from drugs and alcohol in the preceding year, not taking psychiatric medication, more psychiatric symptoms during the pre-baseline year, being more troubled by substance abuse than by mental health, greater level of self-efficacy for recovery, residing in supported housing and being enrolled in outpatient recovery at follow-up" (p. 290). These results, portraying a more mature, motivated, and service-linked profile, are likely specific to these unique dual-focus groups.

Process

Though few studies have delved into the processes and mechanisms through which MASH groups exert an impact on health outcomes, some authors have attempted to identify the mediators. In their investigation of the DTR groups, Magura et al. (2003) found that the degree of involvement in DTR groups and their sponsoring organization (frequency of attendance, sharing at meetings, chairing meetings, speaking with other group members about issues) was significantly positively associated with drug and alcohol abstinence and associated health maintenance behaviors (medication adherence, keeping doctor's appointments, self-care). Furthermore, higher internal locus of control and degree of participant sociability mediated the relationship between DTR involvement and abstinence. Taking into account the relationship between DTR involvement and health maintenance behavior, internal locus of control, sociability, installation of hope, and spirituality all mediated the relationship. These findings are not surprising given that internal locus of control increases individuals' self-efficacy for recovery and sociability facilitates the formation of meaningful relationships and mutual aid in the recovery process (Magura et al., 2003).

The formation of close social ties in self-help groups has also been found to promote member retention and benefit. Drawing on a sample of military veterans who participated in MASH groups for drug and alcohol addiction, Humphreys et al. (1999) found that enhanced friendship networks at 1 year follow-up partially mediated the relationship between group participation and abstinence. It stands to reason that increased extragroup affiliation with other group members who also endorse abstinence is implicated. Even the formation of one good friendship with a group member can spur continued group attendance (Kelly and Moos, 2003). In a separate study of a DTR group, by the first year follow-up, Laudet et al. (2004) found that longer attendance predicted drug and alcohol abstinence, and this relationship was partially mediated by higher levels of social support, presumably based on the development of friendships. In the addictions field, it seems that group programs like these supplement or supplant the negative influence of associates with network members who promote a healthier way of life.

In a fascinating study of 'Clubhouse Communities' for people with mental illness, Pernice-Duca and Onaga (2009) report similar findings about the value of supportive peer ties for people affected by mental illness. Clubhouse programs are much like psychosocial rehabilitation centers, offering support mainly from peers and egalitarian relationship with a small cadre of staff. The 'members' engage in 'work-ordered day units' that require interdependence among the unit members, giving them reciprocal and meaningful social roles, which reflect a recovery model of rehabilitation. Selecting 221 members of 15 clubhouses, the researchers conducted two interviews with each member, separated by an interval of 12–15 months. They measured the number of hours and work-ordered units members participated in, along with several network structural properties (e.g., size) and support-related variables (e.g., perceived support), examining their relative and combined strength in predicting a validated measure of recovery. The findings at both the first and second interviews revealed that, after inserting appropriate statistical controls (e.g., baseline level of functioning), three support network variables made significant and relatively large contributions to the members' recovery: the size of the network, the extent of support, and reciprocity.

Given that more intensive group affiliation has been implicated in positive outcomes for 12-step group members, it should be an area of focus for intervention research. Intensive referrals (those referrals in which a counselor contacts a current self-help group member to accompany a new client to a meeting) have resulted in more frequent attendance, group engagement, and longer periods of abstinence for group attendees compared to clients who received standard referrals among veterans with substance use disorders (Timko and DeBenedetti, 2007). In short, in addition to referring people to 12-step programs, health care professionals should link them to other group members they can befriend and integrate in their natural social network. Clearly, one important limitation of online self-help groups attended by people who are at a geographic distance from one another is the barrier to their formation of a face-to-face friendship in the larger personal communities they inhabit.

Outcomes

With respect to the impact of MASH groups on participants, research continues to focus mainly on 12-step self-help groups,

with the main focus on the participants' psychological functioning. In a study comparing drug and alcohol abstinence among low-income veterans who attended a 12-step self-help program and those given cognitive behavioral therapy, Humphreys and Moos (2007) found comparable short-term psychiatric improvement, but the 12-step program yielded significantly higher abstinence rates at the 2-year follow-up. Pistrang et al. (2008) provide an excellent review of the effects of 12-step self-help groups for individuals with three types of mental health problems, namely chronic mental illness, depression/anxiety, and bereavement. Focusing only on methodologically strong studies containing quantitative evaluations that allowed for some degree of causal inference about the effectiveness of self-help groups, Pistrang et al. (2008) conclude that seven of the studies show positive and significant improvement, including two randomized controlled trials (RCTs). The other five did not result in any changes, and none of the evaluations reported adverse effects. Outcomes included use of and adherence to psychiatric medication and a variety of standardized self-report measures of psychiatric symptomatology. In this regard, the authors suggest that future studies should consider other relevant outcomes that assess life quality, such as reduced isolation, increased confidence, changes in self-identity, and a sense of empowerment.

Promising results have also been reported from other RCTs of self-help groups. In an evaluation of a self-help group for Chinese people with a family member affected by schizophrenia, Chien et al. (2008) reported that participants experienced a significant decrease in caregiver burden, an improvement in family functioning, and an increase in support received compared to those in the wait-listed control group. In addition, the length of hospitalization of the family member with schizophrenia was reduced compared to those in the control group. Regrettably, because this group did not adhere to a clearly defined treatment protocol amenable to replication, its generalizability to other samples of caregivers is unknown. To optimize the potential for knowledge transfer, self-help initiatives such as this one need to be manualized and implemented in ways that are faithful to the blueprint.

More than any other illness, depression has commanded the attention of self-help group evaluators. In a recent meta-analysis of the effects of 'peer support groups' on recovery from depression, Pfeiffer et al. (2011) included nine rigorously designed RCTs that compared the effectiveness of peer support groups vs treatment as usual (TAU), and in a separate meta-analysis, compared the effectiveness of peer support groups vs cognitive behavior therapy. They found a significantly greater reduction in depressive symptomatology and treatment cost among participants in peer support groups compared to TAU, but no significant differences between those support groups and those receiving group cognitive behavioral therapy. However, it is not clear whether the peer support groups were strictly self-help groups or whether they were professionally led support groups; the authors state that "Peer support groups could be professionally led; however, in order to distinguish these from group psychotherapy (i.e., cognitive behavioral therapy), group interventions needed to either be described as peer support (or mutual support or self-help) or to be organized so that participants determined the majority of the topics and content of discussion" (p. 30). Regardless of their significantly lower cost, without knowing exactly how professionals function in these groups, and how active they are, it is impossible to credit peer support for an impact on depression equivalent to professionally administered cognitive behavior therapy.

Conclusion

OSGs are a welcome addition to the armamentarium of supportive resources available to the public. For people who are loathe to approach a professional practitioner directly or attend a self-help or support group meeting, and for people who live in remote areas or whose state of health precludes attendance at face-to-face meetings, online resources can be a boon. However, at this early stage in the evolution of these online support arrangements there is an urgent need for greater Internet literacy so that they can identify legitimate and credible resources and distinguish among the many modes of communication available to them. Quality standards are being developed, both to protect users from sites of dubious quality and intent, and to guide the increasing number of individuals and organizations that want to design ethically sound and user-friendly Web-based supportive resources. Although the Internet's interactive technology has brought about a monumental change in what we can communicate and with whom, it will only fulfill its democratizing and empowering vision when people of all ages, ethnicities, and income levels contribute, thereby fostering greater social integration and mutual aid.

See also: Choice, Empowerment, and Social Work; Healing Practices: Alternative and Complementary; Health Psychology; Health Social Work; Social Integration, Social Networks, and Health; Social Support and Recovery from Disease and Medical Procedures.

Bibliography

Barak, A. Online Support Groups & Self-Help bibliography. http://construct.haifa.ac.il/~azy/refsupp.htm.

Barak, A., Boneh, O., Dolev-Cohen, M., 2010. Factors underlying participants' gains in online support groups. In: Blachnio, A., Przepiórka, A., Rowiński, T. (Eds.), Internet in Psychological Research. Cardinal Stefan Wyszyński University Press, Warsaw, pp. 17–38.

Bennett, G.G., Glasgow, R.E., 2009. The delivery of public health interventions via the internet: actualizing their potential. Annual Review of Public Health 30, 273–292.

Borkman, T., 1976. Experiential knowledge: a new concept for the analysis of self-help groups. Social Service Review 50 (3), 445–456.

Chien, W., Thompson, D., Norman, I., 2008. Evaluation of a peer led mutual support group for Chinese families of people with schizophrenia. American Journal of Community Psychology 42, 122–134.

Eysenbach, G., 2005. The law of attrition. Journal of Medical Internet Research 7 (1), e. 11.

Eysenbach, G., Powell, J., Englesakis, M., Rizo, C., Stern, A., 2004. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. British Medical Journal 328, 1166–1172.

Finn, J., Steele, T., 2010. Online self-help/mutual aid groups in mental health practice. In: Brown, L.D., Wituk, S. (Eds.), Mental Health Self-help: Consumer and Family Initiatives. Springer, New York, pp. 87–105.

Golant, M., Winzelberg, A., Lieberman, M., States, M., Berman, H., Levy, S., et al., 2003. Discussions about suicide in internet delivered cancer support groups. Psychooncology 12, s. 96.

Griffiths, K.M., Calear, A.L., Banfield, M., 2009. Systematic review on Internet support groups (ISGs) and depression (1): do ISGs reduce depressive symptoms? Journal of Medical Internet Research 11 (3), e. 40.

Hobye, M., Dalton, S., Christensen, J., Ross, L., Kuhn, K., Johansen, C., 2010. Social and psychological determinants of participation in internet based cancer support groups. Support Care Cancer 18, 553–560.

Houghton, A.R., Wolstenholme, C.R., Hudson, I., 2001. What questions do patients ask in a cardiology newsgroup and how reliable are the answers? British Journal of Cardiology 8, 165–168.

Humphreys, K., Moos, R.H., 2007. Encouraging posttreatment self-help group involvement to reduce demand for continuing care services: two-year clinical and utilization outcomes. Alcoholism: Clinical and Experimental Research 31 (1), 64–68.

Humphreys, K., Mankowski, E.S., Moos, R.H., Finney, J.W., 1999. Do enhanced friendship networks and active coping mediate the effect of self-help groups on substance abuse? Annals of Behavioural Medicine 21 (1), 54–60.

Houston, T.K., Allison, J.J., 2002. Users of Internet health information: differences by health status. Journal of Medical Internet Research 4 (2), e. 7.

Humphreys, K., Wing, S., McCarty, D., Chappel, J., Gallant, L., Haberle, B., Horvath, A.T., Kaskutas, L.A., Kirk, T., Kivlahan, D., Laudet, A., McCrady, B.S., McLellan, A.T., Morgenstern, J., Townsend, M., Weiss, R.J., 2004. Self help organization for alcohol and drug problems: toward evidence based practice and policy. Substance Abuse Treatment 26 (3), 159–165.

Kelly, J.F., Moos, R., 2003. Dropout from 12 step self help groups: prevalence, predictors and counteracting treatment influences. Journal of Substance Abuse Treatment 24, 241–250.

Kelly, W., Connolly, V., Bilous, R., Stewart, A., Nag, S., Bowes, D., Loader, B.L., Muncer, S., 2002. Reader beware: diabetes advice on the web. Practical Diabetes International 19 (4), 108–110.

Kernsmith, P.D., Kernsmith, R.M., 2008. A safe place for predators: online treatment of recovering sex offenders. Journal of Technology in Human Services 26 (2/4), 223–238.

Laudet, A., Magura, S., Cleland, C.M., Vogel, H.S., Knight, E.L., 2004. Predictors of retention in dual focus self help groups. Community Mental Health Journal 39 (4), 281–297.

Lieberman, M.A., Golant, M., Giese-Davis, J., Winzlenberg, A., Benjamin, H., Humphreys, K., Kronenwetter, C., Russo, S., Spiegel, D., 2003. Electronic support groups for breast carinoma. Cancer 97 (4), 920–925.

Magura, S., Knight, E.L., Vogel, H.S., Mahmood, D., Laudet, A.B., Rosenblum, A., 2003. Mediators of effectiveness in dual focus self help groups. The American Journal of Drug and Alcohol Abuse 29 (2), 301–322.

Moa, P.K.H., Coulson, N.S., 2012. Developing a model for online support group use, empowering processes and psychosocial outcomes for individuals living with HIV/AIDS. Psychology and Health 27 (4), 445–459.

Murray, E., Burns, J., See, T.S., Lai, R., Nazareth, I., 2005. Interactive health communication: applications for people with chronic disease. Cochrane Database Systematic Reviews 4, CD004274.

Pernice-Duca, F., Onaga, E., 2009. Examining the contribution of social network support to the recovery process among clubhouse members. American Journal of Psychiatric Rehabilitation 12, 1–30.

Pfeiffer, P., Heisler, M., Piette, J.D., Rogers, M.A.M., Valenstein, M., 2011. Efficacy of peer support interventions for depression: a meta-analysis. General Hospital Psychiatry 33, 29–36.

Pistrang, N., Barker, C., Humphreys, K., 2008. Mutual help groups for mental health problems: a review of effectiveness studies. American Journal of Community Psychology 42, 110–121.

Proudfoot, J., Klein, B., Barak, A., Carlbring, P., Cuijpers, P., Lange, A., Ritterband, L., Andersson, G., 2011. Establishing guidelines for executing and reporting internet intervention research. Cognitive Behaviour Therapy 40 (2), 82–97.

Rippen, H., Risk, A., 2000. E-Health code of ethics. Journal of Medical Internet Research 2 (2), 9.

Risk, A., Dzenowagis, J., 2001. Review of internet health information quality initiatives. Journal of Medical Internet Research 3 (4), 28.

Shaw, L.H., Gant, L.M., 2002. In defense of the Internet: the relationship between Internet communication and depression, loneliness, self-esteem, and perceived social support. CyberPsychology and Behaviour 5 (2), 157–171.

Timko, C., DeBenedetti, A., 2007. A randomized control trial of intensive referral to 12 step self help groups: one year outcomes. Drug and Alcohol Dependence 90, 270–279.

van Uden-Kraan, C.F., Drossaert, C.H.C., Taal, E., Shaw, B.R., Seydel, E.R., van de Laar, M.A., 2008. Empowering processes and outcomes of participation in online support groups for patients with breast cancer, arthritis, or fibromyalgia. Qualitative Health Research 18 (3), 405–417.

Wituk, S., Shepherd, M.D., Warren, M., Meissen, G., 2002. Factors contributing to the survival of self help groups. American Journal of Community Psychology 30 (3), 349–366.

Winzelberg, A.J., Classen, C., Alpers, G.W., Roberts, H., Koopman, C., Adams, R.E., Ernst, H., Dev., Bar Taylor, C., 2003. Evaluation of an internet support group for women with primary breast cancer. Cancer 97 (5), 1164–1173.

Health Research, Qualitative

Thorsten Meyer, Institute for Epidemiology, Social Medicine, and Health Systems Research, Hannover Medical School, Hannover, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P. Conrad, volume 10, pp. 6608–6612, © 2001, Elsevier Ltd.

Abstract

In qualitative health research qualitative social research methods are applied to study problems at the interface of health or illness and social life. Qualitative research is an umbrella term of research characterized by openness, subjectivity, process orientation, and interpretation of meaning. Interviews, group discussions, observations, and document analysis are the main modes of enquiry. Critiques of qualitative health research that is predominantly related to issues of sample size and generalizability ignore central concepts of qualitative methodology. Qualitative health research has been discussed within the premises of an evidence-based medicine. It has made valuable contributions to the field of health research.

Qualitative health research refers to the broad area of health research applying predominantly qualitative social research methods to study problems at the interface of health or illness and related social issues such as health care. Qualitative research is an umbrella term for a wide spectrum of different social research methods and methodologies. The term implies a clear distinction from 'quantitative' research – a *term* that is not prominent in the health sciences, although health sciences are dominated by quantitative research methods. Rather, distinctions are made with regard to the perspective of enquiry, for example, public health, epidemiology, clinical research; the domain of research, for example, oncology, rheumatology; the health strategy applied, for example, prevention, or rehabilitation; or the developmental phase of an intervention, for example, from basic research via translational research via clinical research to health services research. Qualitative health research comes into play, the more important social aspects of health, illness, and care become – that is, people and their behaviors, decisions, preconceptions, attitudes, interactions as well as social groups including institutions, systems, rules, structures. Qualitative health research is especially suited to deal with aspects of health interventions and health care, as well as social conditions or consequences of illness.

Like 'quantitative' research, qualitative research comprises a plethora of different ideas, approaches, methods, methodologies, and concepts. Different qualitative approaches even resort to different epistemologies. However, there are some core characteristics of qualitative research that make it particular, especially when they result in quality criteria that contradict health researchers' thinking that has been trained in biomedical research traditions.

Characteristics of Qualitative Research

Qualitative health research draws from traditions from different fields of research, including the social sciences/sociology, ethnology, psychology, and linguistics. Common to all qualitative approaches is a focus on persons or people within social relations. The spectrum of different methodological approaches ranges from the description, reconstruction, and analysis of subjective perspectives, attitudes, behaviors, and social interactions, including the reconstruction of structures in social fields, and the interpretation of basic or latent meanings of actions (Meyer and Flick, 2010).

Openness is a core characteristic of qualitative research with very different facets. In terms of research designs, qualitative research is set up to allow what is unknown or unexpected to emerge during the conduct of the study. It also means that researchers are able to adopt their design or methods during the course of a study, if there are good reasons to do so and could not be anticipated at the outset of the study. Qualitative researcher might need to adopt the research question if it turned out to be flawed, to be imprecise, or to make too restrictive assumptions. The sampling criteria of participants might change or develop during the course of the study (e.g., theoretical sampling in grounded theory; Glaser and Strauss, 1967). Methods of collecting data should be set up to allow for openness, for example, interview questions that stimulate interviewees to talk openly and provide rich data. Openness also refers to a specific stance toward research subjects, for example, by providing conditions – a quiet and undisturbed room, a trustful and accepting interviewer, taking care that the participation of the interviewee is indeed voluntary – in which a person is able to talk freely. Openness also refers to the data format, that is, in contrast to fixed or closed response formats. Usually data in qualitative enquiries are synonymous to texts that need further interpretation. Openness can also refer to an important characteristic during analysis, where it is important to be open to very different interpretations of data to provide a deeper and more comprehensive understanding of the phenomenon under study.

Subjectivity is another core characteristic of qualitative research. Often qualitative research aims to reconstruct the (spectrum of) subjective perspectives and attitudes of different people, for example, patients, professionals, or other stakeholders, both with regard to individual persons and in social situations. Also, the researcher as a subject is acknowledged as part of the research process. Researchers' interests, motivations, expectations, values, personal experiences, theoretical stances, and knowledge are known to influence every step of the research process. Instead of pursuing the impossible task of

making the researcher invisible or 'objective,' researchers have to be very aware of how their characteristics influence the research process and to take this into account during the different phases of a project. In other words, the subjectivity of the researcher is acknowledged and harnessed. Subjectivity also relates to an epistemological idea that research does not result in objectively valid findings, but rather in subjective 'readings' of a social phenomenon. That means, qualitative research does not strive to discover the one and only objective truth, but to provide statements that are well founded in the data and developed with methodological rigor.

Process and iteration: A linear research process is seen as an ideal in biomedical research traditions. In qualitative research, research processes are rather circular and iterative. For example, data acquisition, analysis, and theoretical interpretation are not different research phases but can be deeply interrelated. In a qualitative interview, analysis should start as soon (or late) as the interviewee responds to the first question. In grounded theory, analysis and theoretical considerations form the basis of further sampling. Text is not read and coded once, but it is reread and revisited repeatedly. This procedure stimulates different interpretations and allows to develop categories and theoretical assumptions that are consistent with the data. The term *process* also refers to the fit of qualitative research method with the object (or subject!) of research, which requires recurrent reflection. Orientation toward processes also means that inductive reasoning has an important part in the research process. Principles of induction are used to develop new theories grounded in the data without ignoring the value of deductive or possible pitfalls of inductive reasoning. Distant from the domains of qualitative research, the dominance of hypothetico-deductive reasoning in scientific epistemology runs the risk of neglecting the process of hypothesis development. Qualitative research offers well-founded approaches for the development of hypotheses or theories. Process orientation, together with openness and subjectivity, points to another characteristic of qualitative research, that is, the need for transparency in terms of rigor in documenting and disclosing methodological decisions made during the course of a study.

Interpretation of meanings is also at the core of qualitative research. Doing qualitative research is doing interpretation work. While it could be argued that interpretation of meaning is a general task of research, qualitative research is particular in this regard because here interpretation is an integral part of the analysis (and even research) process. It is not confined to the final discussion phase of study results (Meyer et al., 2012).

Methods of Qualitative Inquiry

Interviews

In health research a prominent method of inquiry is doing interviews (Pope and Mays, 2009; Karbach et al., 2012). Interviews are characterized by a direct, personal interaction of an interviewer and an interviewee. It is also possible to have more than one interviewer. Also, it is not uncommon to have more than one interviewee present, for example, family members or caregivers in phases of special care needs. Usually, interviews are conducted by a face-to-face interaction, although other modes are also practiced, for example, telephone interviewing or use of interactive media. Very different interview forms exist depending on the purpose of the interviews. For example, Kvale (2007) distinguishes conceptual interviews, narrative interviews, and discursive interviews. He also refers to factual interviews and confrontational interviews, which can be thought of as rather active interview forms that do not always comply with the core characteristics of qualitative research introduced above, for example, the characteristics of openness, orientation toward the subject, or the need for interpretation. Also, there are special subject groups who require special forms of interview, for example, palliative care patients, children, persons with different levels of language competence as in certain immigrant groups, elites/professionals, and persons with intellectual disabilities. Interviews differ according to the extent of standardization or reduced openness in the conduct of the interview. On the one side there are standardized interviews that have fixed questions to be asked in a defined order. On the other side there are narrative interviews or informal conversational interviews, which provide the interviewee with an incentive to talk about a certain topic, without making further restrictions. The more open the type of interview, the more important is the task of the interviewer to ensure the flow of communication. One part of the analysis is then to give explanations why the interviewee has chosen the contributions presented in the interview. Between structured and narrative interviews the popular form of semistructured interview (also called the interview guide approach (Patton, 2002)) can be located. Here, an interview guide is developed containing the important topics the interview should cover, some pivotal questions, and possible further questions. In this type of interview it is important that the interviewee is stimulated to talk about each of the topics deemed important, without necessarily adhering to the specific wordings of the questions or their order. Choice of the extent of standardization in interviews should depend on the knowledge of researchers in the field of interest.

Group Discussions/Focus Groups

Group discussions or focus groups are sometimes referred to as a special form of interviews (Kvale, 2007; Patton, 2002). Focus groups have found their way to health research via marketing research, where it has long been a prominent tool (Krueger and Casey, 2009). Present use of focus groups in health research seems to primarily resort to this pragmatic approach. However, there have been traditions in other fields and countries that have provided, besides others, substantial methodological advancements, for example, Morley (1980) within the framework of cultural studies in the United Kingdom, or the work of the Frankfurt Institute for Social Research in Germany shaped by the early work of Pollock (1955) and its successors, including the documentary method developed by Bohnsack (2010).

Besides similarities between individual interviews and group approaches, it should be acknowledged that group discussions share potentials that make them qualitatively distinct from individual interviews. Most importantly, interactions among participants can be used as an additional resource in interpretations. They render focus groups/group discussions especially suitable for gathering opinions at the

group level, because participants tend to check and balance each other. Also, there are sensitive topics, for example, relating to stigmatization of persons, that are more easily conveyed in group settings with persons deemed as similar to oneself than in individual interviews. A discussion of limitations of focus groups can be found in Patton (2002).

Observations

Observations, especially field observations, represent a pivotal method of qualitative inquiry. However, it is underrepresented in current qualitative health research (Pope and Mays, 2009; Karbach et al., 2012). Scientific observations are different from everyday life observations in that they serve a certain purpose, are planned and documented systematically, and are scrutinized with regard to their methodological rigor. There are different forms of observations: they can be open or covert, that is, persons either knowing or ignoring that they are observed, or knowing or ignoring the reasons for being observed. However, covert observations are highly sensitive to ethical considerations and not compatible with the openness characteristic of qualitative enquiry. The degrees of participation of the researcher may vary. For example, researchers can become part of the setting, both in terms of a patient and professional. The study by Rosenhan (1973), in which persons were admitted to psychiatric wards after having presented some fake symptoms and behaving as normal as possible afterward, is a famous example of the use of participant observation. Alternatively, researchers participate by 'shadowing' persons in their daily activities, that is, they closely follow a person over an extended period of time (McDonald, 2005). In medical anthropology, participant observation has been successful, for example, in improving our understanding of the effects of public needle exchange programs in drug users to prevent HIV infection programs which turned out to be positively related in epidemiological studies (cf Bourgois and Bruneau, 2000). Nonparticipatory observations can be found, for example, in studies using video recordings of patient–physician interactions (cf Themessl-Huber et al., 2008). Observations can be made either in the field, that is, where the phenomenon of interest occurs, or in the laboratory. While the latter can be used to control for possible biases, the former has the advantage of being a representation of people's lived experiences. The psychological tradition of self-observation (introspection) is hardly applied in qualitative health research today. Observations are often combined with other modes of inquiry, as had been done by a famous sociological work related to the analysis of the consequences of unemployment, the 'Marienthal' study by Jahoda and coworkers in the early 1930s in Austria (Jahoda et al., 1971). In this study, participatory research had been combined with very different research methods, including nonreactive techniques such as documentary analysis, analysis of official statistics, or elements of action research, surveys, and tests.

Documents

Retrieval of documents can be viewed as a nonreactive method of inquiry. Its primary advantage is that these documents have not been set up with the knowledge of being analyzed in a research project. This reduces an important social desirability effect in qualitative research. However, researchers have to be aware of other social processes pertinent to the production of the document. For example, filling in a medical record should be related to the purpose of the documentation and is based on specific social codes. Another example are minutes of a meeting that may be used to analyze different social roles of the participants. Document analysis, therefore, has always to take into account the context in which these documents are generated. It is a challenge of document analysis to develop a clear-cut rationale of how these behavioral traces relate to the theoretical construct of interest. Digitalization of social life – especially so in relation to internet activities – offers a renaissance to the analysis of documents within qualitative research (example in Willis et al., 2012).

Methods of Qualitative Analysis

There is a considerable gap between a plethora of available methods of qualitative analysis and the proficiency of qualitative health researchers in doing different types of analysis. Sackett's and Wennberg's dictum of choosing appropriate methods for research questions (Sackett and Wennberg, 1997) – and not vice versa – holds true for qualitative analysis as for any analysis in empirical research. In practice, this knowledge is difficult to obtain because each method of analysis needs to be learned by experience and experienced teacher and colleagues. Also, contrary to statistical analysis, different traditions of analysis have emerged in different regions confined to cultural and language boundaries. Qualitative research requires a high level of language competence and a thorough understanding of the sociocultural context of the study. These challenges might have contributed to a lack of convergence in this field from an international perspective.

The choice of method of analysis is related to the type of phenomenon under study. If a researcher is interested in manifest views of subjects or stakeholders, then a content analysis approach that summarizes what is being said deems appropriate. There is a tradition for this type of analysis, especially in eliciting special patient groups' views, in documenting their experiences with illness and care (cf Greenhalgh and Hurwitz, 1998). These can be single case studies or – relating to groups of patients – descriptive studies (cf the taxonomy of qualitative study evidence by Daly et al., 2007). The starting point of analysis can be a thorough analysis of single cases, and this could result in the development of different types of cases. Another starting point of analysis could focus on categories beyond the single case and could try to develop common themes.

The strength of qualitative analysis is to move beyond the apparent or manifest by elaborating latent characteristics or mechanisms through the researcher's interpretation. From a psychological perspective this relates to all those important aspects of people's lives that are situated beyond their immediate grasp and, thus, have to be inferred from the data. From a sociological perspective social phenomena have to be reconstructed to identify their underlying conditions, such as beliefs, rules, norms, values, structures, power relationships, or societal discourses.

Beyond the manifest–latent distinction and the type of explanations the researcher is targeting at, it is also important to know how far-reaching the aim of a research project is. For example a qualitative health researcher could be interested in a local theory on why people do not respond to regular screening invitations in a national screening program for colorectal cancer (e.g., Hall et al., 2013). In a further step one would try to develop a general theory on reasons for (non) participation in public health programs. Generalizable theories are at the top of the evidence hierarchy in the taxonomy of Daly et al. (2007). The most elaborated methodology to develop such theories is the 'grounded theory approach' introduced by Glaser and Strauss (1967), and further developed by their colleagues and scholars (e.g., Corbin and Strauss, 2007; Charmaz, 2006). Grounded theory integrates both a case perspective and an overarching categorical perspective, especially so by applying the method of constant comparisons. It posits that theoretical statements should be developed from data and that their conceptual assumption should be rigorously checked against the data. The grounded theory approach still represents a landmark in qualitative health research.

Critiques of Qualitative Health Research: Sampling, Sample Size, and Generalizability

Qualitative research has ever since been confronted with the view that sample sizes are too small to secure representativeness and therefore generalizability. Rather, this research is useful in exploring a new field of interest that should be investigated by other research methods later on. The emergence of grounded theory contradicts this assumption. Sample size is still a lively debated issue (Patton, 2002). The specification of a necessary number of subjects should always be related to the aim and scope of the research question. It should be kept in mind that there is a trade-off between the number of subjects that are included in the analysis and the depth or thoroughness by which each subject can be analyzed. A well-founded, targeted sampling of research subjects is at the heart of qualitative inquiry.

Qualitative analysis does not strive toward statistical generalization. However, other models of generalization are relevant to qualitative enquiry, such as analytic generalization and transferability (Polit and Beck, 2010). In analytical generalization, researchers aim to develop their theory by an interplay of inductive and deductive (confirmatory) research processes. An analysis of this kind should be able to represent all relevant aspects of a field of interest. It should be able to capture the diversities, that is, the complete range of phenomena in the field of interest. The idea of data saturation of grounded theory is related to this type of generalization – albeit difficult to argue for in practice. This approach toward generalization has implications for the sampling procedure. For example, one such strategy of sampling, the maximum variation, aims at including extreme or deviant cases, from which a lot can be learned. On the other hand, convenience sampling (i.e., sampling those that can be recruited easily) is the least desirable strategy (Patton, 2002).

Qualitative Health Research in an Era of Evidence-Based Medicine

Evidence-based medicine (ebm) has been defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients." (Sackett et al., 1996: p. 71). The ebm approach evolved from a perspective of clinical epidemiology. The use of hierarchies of study designs, the use of critical appraisal of original studies, and the development of systematic reviews have been hallmarks of this movement. It has spread over to other health-related fields such as nursing, physiotherapy, or rehabilitation. Although ebm has been recognized as a threat to qualitative health research (Nelson, 2008), it has moved the field toward new directions and fostered new methodological reflections (cf Morse et al., 2001).

Qualitative studies can have their own place as evidence in clinical decision making. The naturalism of qualitative studies, that is, understanding health-related behavior in its everyday context, has been acknowledged. Or, understanding problems in the implementation of evidence-based medicine into practice is another important domain of qualitative research (Grenn and Britten, 1998). Although the idea of a hierarchy of study designs in answering questions of clinical practice (Daly et al., 2007) has not been widely adopted, the interest in qualitative studies has increased remarkably (Nelson, 2008). However, the appraisal of qualitative studies is still a much debated and contested area (Dixon-Woods et al., 2004). This is exemplified by a number of substantially different approaches, including checklists, explication of general criteria, appraisal within the usage of a specific approach (e.g., discourse analysis), among others. So far, there is no consensus about these issues in the qualitative research community. Quality criteria for qualitative studies (e.g., Cohen and Crabtree, 2008), as well as criteria concerning the reporting of qualitative studies (e.g., Tong et al., 2007), have found a broader audience. However, important issues of evaluating qualitative studies have not been resolved and, consequently, the use of checklists to appraise the quality of qualitative studies has not been thoroughly adopted in the qualitative research community (Barbour, 2001). It is of interest to see that prominent clinical journals, such as the British Medical Journal (e.g., Jones, 1995; Greenhalgh and Taylor, 1997; Mays and Pope, 2000), the Journal of the American Medical Association, or the Lancet (Malterud, 2001), have published contributions on qualitative methods of inquiry (e.g., Giacomini and Cook, 2000), paving the way to subsequent original qualitative studies. Moreover, the need to integrate qualitative studies in systematic reviews has been acknowledged quite some time ago, for example, in the field of health technology assessment by Murphy et al. (1998) and clinical research (Campbell et al., 2003), which in turn brought another impetus into the quality debate (e.g., Barnett-Page and Thomas, 2009). This new proximity of qualitative inquiry to biomedical research is also apparent in a new movement toward so-called mixed method studies (Song et al., 2010), which relates mainly to a pragmatic research tradition (although critically reviewed by some researchers, e.g., Miller and Crabtree, 2005: p. 610).

Contributions of Qualitative Health Research

It is impossible to summarize the contributions of qualitative health research to the field of health research in a single paragraph. Therefore, a few lines of enquiry will be outlined. There are famous and seminal qualitative studies on mental illness. Qualitative research in psychiatry is strongly related to institutionalization and deinstitutionalization of persons with mental illness. For example, a classical work of Erving Goffmann (1961), Asylums, deals with the psychiatric institution as a total institution, a work with wide recognition beyond the field of psychiatry. David Rosenhan (1973) dealt with the question of how the assignment of the 'patient role' contributes to the way of how patients with mental illness are treated within psychiatric institutions. Denise Jodelet (1991) conducted a study on social mechanisms of how mentally healthy and mentally ill people construct their social world in a small French town where formerly hospitalized patients lived within the homes of the villagers. From an anthropological perspective Sue Estroff (1985) investigated the life of persons with chronic mental illness who were discharged to their community. Another example of qualitative enquiry is the work by Ulrike Hoffmann-Richter from Switzerland, who thoroughly analyzed how schizophrenia is covered in print media and what this means for the stigmatization of persons with the diagnosis of schizophrenia (Hoffmann-Richter, 2000). Outstanding contributions to qualitative health research have been made by one of its founding fathers, medical sociologist Anselm Strauss, and his research affiliates. In the academic environment of the University of Chicago, Strauss learned to examine illness from the perspective of the patient. His early work related to the process of dying (together with Barney Glaser; Glaser and Strauss, 1965, 1968), a work that still resonates in the field. Inherent in this work was the concept of illness trajectory which became important later on in research on chronic illness. Together with Juliet Corbin he analyzed the role of family members in the care of chronic patients, as expressed by the concept of *kin work* (Corbin and Strauss, 1988). His methodological work on the development of the grounded theory approach has become seminal to the broad field of qualitative health research (Corbin and Strauss, 2007). An important line of research is related to the integration of the perspective of the ill persons by means of illness narratives (Kleinman, 1988; Greenhalgh and Hurwitz, 1998; Herxheimer et al., 2000) which can provide the basis for further in-depth analysis. Qualitative research methods are also central to feminist health research and medical ethics. Here, qualitative research is instrumental in identifying tacit assumptions like beliefs, values, or societal powers/power relationships. While qualitative health research still appears to be prominent in the Anglo-Saxian and European research literature, threads of qualitative enquiry have also developed in Latin America and in African countries.

See also: Bereavement; Doctor–Patient Interaction in the West: Psychosocial Aspects; Evidence-Based Medicine; Field Observational Research in Anthropology and Sociology; Gender and Health Care; Health Outcomes, Assessment of; Health Surveys; Health and Illness: Mental Representations in Different Cultures; Health: Anthropological Aspects; Organizations and Culture; Public Health as Social Science; Qualitative Methods, History of.

Bibliography

Barbour, R.S., 2001. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? British Medical Journal 322, 1115–1117.
Barnett-Page, E., Thomas, J., 2009. Methods for the synthesis of qualitative research: a critical review. BMC Medical Research Methodology 9, 59.
Bohnsack, R., 2010. Documentary method and group discussions. In: Bohnsack, R., Pfaff, N., Weller, W. (Eds.), Qualitative Analysis and Documentary Method in International Educational Research. Opladen & Farmington Hills, Budrich, pp. 99–124.
Bourgois, P., Bruneau, J., 2000. Needle exchange, HIV infection, and the politics of science: confronting Canada's cocaine injection epidemic with participant observation. Medical Anthropology. Cross-Cultural Studies in Health and Illness 18, 325–350.
Campbell, R., Pound, P., Pope, C., Britten, N., Pill, R., Morgan, M., Donovan, J., 2003. Evaluating meta-ethnography: a synthesis of qualitative research on lay experiences of diabetes and diabetes care. Social Science & Medicine 56, 671–684.
Charmaz, K., 2006. Constructing Grounded Theory. A Practical Guide through Qualitative Analysis. Sage, London.
Cohen, D.J., Crabtree, B.F., 2008. Evaluative criteria for qualitative research in health care: controversies and recommendations. Annals of Family Medicine 6, 331–339.
Corbin, J., Strauss, A.L., 1988. Unending Work and Care: Managing Chronic Illness at Home. Jossey-Bass, San Francisco.
Corbin, J., Strauss, A.L., 2007. Basics of Qualitative Research. Techniques and Procedures For Developing Grounded Theory, third ed. Sage, Thousand Oaks, CA.
Daly, J., et al., 2007. A hierarchy of evidence for assessing qualitative health research. Journal of Clinical Epidemiology 60, 43–49.
Dixon-Woods, M., Agarwal, S., Young, B., Jones, D., Sutton, A., 2004. Integrative Approaches to Qualitative and Quantitative Evidence. Health Development Agency, NHS, London.
Estroff, S.E., 1985. Making It Crazy: An Ethnography of Psychiatric Clients in an American Community. University of California Press, Berkeley.
Giacomini, M.K., Cook, D.J., for the Evidence-Based Medicine Working Group, 2000. Users' guide to the medical literature. XXIII. Qualitative research in health care. A. Are the results of the study valid? Journal of the American Medical Association 284, 357–362.
Glaser, B.G., Strauss, A.L., 1965. Awareness of Dying. Aldine Publishing, Chicago.
Glaser, B.G., Strauss, A.L., 1967. The Discovery of Grounded Theory: Strategies for Qualitative Research. de Gruyter, New York, USA.
Glaser, B.G., Strauss, A.L., 1968. Time for Dying. Aldine Publishing, Chicago.
Goffman, E., 1961. Asylums: Essays on the Social Situation of Mental Patients and Other Inmates. Doubleday, New York.
Greenhalgh, T., Hurwitz, B. (Eds.), 1998. Narrative Based Medicine. BMJ Books, London.
Greenhalgh, T., Taylor, R., 1997. How to read a paper: papers that go beyond numbers (qualitative research). British Medical Journal 315, 740–743.
Grenn, J., Britten, N., 1998. Qualitative research and evidence based medicine. British Medical Journal 316, 1230–1232.
Hall, N.J., et al., 2013. Attitudes and beliefs of non-participants in a population-based screening programme for colorectal cancer. Health Expect. http://dx.doi.org/10.1111/hex.12157. Epub ahead of print.
Herxheimer, A., et al., 2000. Database of patients' experiences (DIPEx): a multimedia approach to sharing experiences and information. Lancet 355, 1540–1543.
Hoffmann-Richter, U., 2000. Psychiatrie in der Zeitung. Urteile und Vorurteile [Psychiatry in the Newspaper. Judements and Prejudices.]. Psychiatrie-Verlag, Bonn, Germany.
Jahoda, M., Lazarsfeld, P.F., Zeisel, H., 1971. Marienthal. The Sociography of an Unemployed Community (Translation from the German by the authors with John Reginall and Thomas Elsaesser). Aldine Atherton, Chicago, USA.
Jodelet, D., 1991. Madness and Social Representations: Living with the Mad in One French Community. University of California Press, Berkeley.
Jones, R., 1995. Why do qualitative research? British Medical Journal 311, 2.

Karbach, U., et al. für die Arbeitsgruppe Qualitative Methoden im Deutschen Netzwerk für Versorgungsforschung e.V. (DNVF), 2012. Qualitative Studien in der Versorgungsforschung, Diskussionspapier, Teil 2. [Qualitative studies in health services research, discussion paper 2] Stand Qualitative Versorgungsforschung in Deutschland – ein exemplarischer Überblick. Gesundheitswesen 7, 516–525.
Kleinman, A., 1988. Illness Narratives. Basic Books, New York.
Kvale, S., 2007. Doing Interviews. Sage, London.
Krueger, R.A., Casey, M.A., 2009. Focus Groups- a Practical Guide for Applied Research, fourth ed. Sage Publications, Thousand Oaks, CA.
Malterud, K., 2001. Qualitative research: standards, challenges, and guidelines. Lancet 358, 397–400.
Mays, N., Pope, C., 2000. Assessing quality in qualitative research. British Medical Journal 320, 50–52.
McDonald, S., 2005. Studying actions in context: a qualitative shadowing method for organizational research. Qualitative Research 5, 455–473.
Meyer, T., Flick, U., 2010. Methoden der qualitativen Forschung [Methods in qualitative research]. In: Pfaff, H., Neugebauer, E.A.M., Glaeske, G., Schrappe, M. (Eds.), Lehrbuch für Versorgungsforschung. Systematik, Methodik, Anwendung. Schattauer, Stuttgart, pp. 296–302.
Meyer, T., et al., 2012. für die Arbeitsgruppe Qualitative Methoden des DNVF. Qualitative Studien in der Versorgungsforschung – Diskussionspapier, Teil 1 [Qualitative studies in health services research, discussion paper 1]. Gegenstandsbestimmung. Gesundheitswesen 74, 510–515.
Miller, W.L., Crabtree, B.F., 2005. Clinical research. In: Denzin, N.K., Lincoln, Y.S. (Eds.), The Sage Handbook of Qualitative Research, third ed. Sage, Thousand Oaks, USA, pp. 605–639.
Morley, D., 1980. The Nationwied Audience. Structure and Decoding. BFI, London.
Morse, J.M., Swanson, J.M., Kuzel, A.J., 2001. The Nature of Qualitative Evidence. Sage Publications, Thousand Oaks, CA.
Murphy, E., et al., 1998. Qualitative research methods in health technology assessment: a review of the literature. Health Technology Assessment 2 (16).
Nelson, A.M., 2008. Addressing the threat of evidence-based practice to qualitative inquiry through increasing attention to quality: a discussion paper. International Journal of Nursing Studies 45, 316–322.
Patton, M.Q., 2002. Qualitative Research and Evaluation Methods, third ed. Sage Publications, Thousand Oaks, CA.
Polit, D.F., Beck, C.T., 2010. Generalization in quantitative and qualitative research: myths and strategies. International Journal of Nursing Studies 47, 1451–1458.
Pollock, F., 1955. In: Gruppenexperiment. Ein Studienbericht. Franfurter Beiträgte zur Soziologie, Bd. 2. Europäische Verlagsanstalt, Frankfurt/M.
Pope, C., Mays, N., 2009. Critical reflections on the rise of qualitative research. British Medical Journal 339, 737–739.
Rosenhan, D.L., 1973. On being sane in insane places. Science 179, 250–258.
Sackett, D.L., et al., 1996. Evidence based medicine: what it is and what it isn't. British Medical Journal 312, 71.
Sackett, D.L., Wennberg, J.E., 1997. Choosing the best research design for each question. British Medical Journal 315, 1636.
Song, M.-K., Sandelowski, M., Happ, M.B., 2010. Current practices and emerging trends in conducting mixed methods intervention studies in the health sciences. In: Tashakkori, A., Teddlie, C. (Eds.), Mixed Methods in Social & Behavioral Research, second ed. Sage Publications, Thousand Oaks, CA, pp. 725–747.
Themessl-Huber, M., et al., 2008. Audio-visual recording of patient-GP consultations for research purposes: a literature review on recruiting rates and strategies. Patient Education and Counseling 71, 157–168.
Tong, A., Sainsbury, P., Craig, J., 2007. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. International Journal for Quality in Health Care 19, 349–357.
Willis, K., Small, R., Brown, S., 2012. Using documents to investigate links between implementation and sustainability in a complex community intervention: the PRISM study. Social Science & Medicine 75, 1222–1229.

Relevant Websites

http://cqim.cochrane.org/ – Cochrane Qualitative and Implementation Methods Group.
http://www.qualitative-research.net/index.php/fqs – Forum Qualitative Sozialforschung/ Forum Qualitative Social Research.
http://global-qhr.org/ – Global Congress for Qualitative Health Research.
http://healthtalkonline.org/ – Healthtalk.
http://qhr.sagepub.com/ – Qualitative Health Research.
http://www.slu.edu/organizations/qrc/QRjournals.html – Qualitative Research Journals.

Health Risk Appraisal and Optimistic Bias

Neil D Weinstein, Rutgers University, Tucson, AZ, USA
William MP Klein, National Cancer Institute, NIH, Bethesda, MD, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by N.D. Weinstein, volume 10, pp. 6612–6615, © 2001, Elsevier Ltd.

Abstract

People's beliefs about the hazards they face, their 'risk perceptions,' can affect their actions and emotions. These perceptions are often quite different from risk assessments made by experts. Experts emphasize the numerical probabilities of hazards, a metric that is poorly understood by laypeople. Just as hazards have many dimensions, risk appraisals are influenced by many factors and are subject to a variety of errors and biases. One in particular is the optimistic bias (unrealistic optimism), a tendency to believe that our own vulnerability to hazards is less than that of our peers.

One reason for expert–layperson disagreements is the public's difficulty in understanding numerical presentations of probabilities and odds. For example, people tend to overweight the representativeness of single cases and small samples (Tversky and Kahneman, 1974), and exaggerate the probability of rare events (Slovic, 1987). Another reason is an optimistic bias (or unrealistic optimism) about personal risk, a tendency to believe that one's own vulnerability is less than that of one's peers. Depending on the context, underestimation of risk may be harmful or beneficial.

A third reason for disagreement is that people's judgments of risk often depend on their emotions, which can either facilitate or hinder their understanding of risk information (Loewenstein et al., 2001).

Importance of Health Risk Perceptions

Perceptions of health risks can influence the actions people take to prevent illness and injury, the decisions they make to seek medical care, their choices among treatment options, and their adherence to treatment requirements. Recent work shows that this may be particularly true for affective and intuitive perceptions of risk (i.e., feeling at risk) rather than deliberative perceptions of risk (e.g., ratings on a percentage likelihood scale) (Weinstein et al., 2007). Public controversies over the risks created by government and corporate action are common. During these controversies, scientists and members of the public passionately disagree, demonstrating that laypeople do not passively accept the risk judgments of experts.

The controversy and complexity of risk appraisals suggest that it is important to understand how these appraisals are formed, how accurate they are, and how they influence decisions and behavior. Furthermore, as research continues to identify new risks and 'risk factors' (attributes of individuals or environments that alter a person's likelihood of harm), the need grows to explain this information clearly so that individuals can make better decisions about their health. Such education about risk is the goal of the specialty called 'risk communication.'

Defining 'Risk' and 'Risk Perceptions'

Before pursuing this topic it should be mentioned that the term 'risk' has several meanings. Sometimes 'risk' is used synonymously with the word 'hazard,' denoting a specific undesirable outcome. In this sense, cancer, the common cold, and automobile accidents are all 'risks.' 'Risk' is also used as a synonym for 'probability,' referring to the probability (or likelihood) of the undesirable outcome. Thus, one can speak of the risk that a person will develop cancer. 'Risk' is frequently used in still another way, as a synonym for 'danger' or 'threat' (for example, 'How risky is smoking?'). The term 'risk perceptions,' then, is a general term that refers to the various beliefs that people hold about health and safety hazards.

How Do People Think about Risk?

Dimensions of Hazards

Hazards have many dimensions. According to classical theories of decision-making, the key attributes of a hazard are the probability that it will occur and the severity of the undesirable consequences should it occur.

Pioneering research by Lichtenstein et al. (1978) demonstrated that laypeople are also influenced by hazard attributes that are not so obviously important. When scientists are asked about the riskiness of different illnesses and technologies, their judgments strongly reflect the number of people who die annually from each one, thus reflecting both seriousness (deaths) and likelihood (number of events in a particular population). Lay judgments, however, are only moderately related to the number of fatalities. In addition, laypeople respond to whether the risk is familiar or unfamiliar, manufactured or natural, voluntary or involuntary, and other factors. Summarizing the many attributes that influence risk perceptions, researchers have suggested that lay judgments of risk reflect two main dimensions, the extent to which the hazard is perceived to be known or unknown and the extent to which it is dreaded or not (Slovic, 1987). Because of the dread associated with significant hazards such as nuclear accidents, people often misperceive that there is a negative correlation between the

magnitude of risk and the extent of benefit, when in fact the correlation is positive (Slovic et al., 2004).

Leventhal and colleagues suggest that each person copes with illness threats in accord with their own 'common sense' model of that illness (*see* Illness Behavior and Care-Seeking). These models involve five attributes: the identity of the risk, its timecourse (e.g., chronic vs acute), its consequences, its causes, and its potential cure or control (Leventhal et al., 1992).

The preceding perspectives emphasize various hazard attributes. Another approach emphasizes the interconnections among the hazard properties. Beliefs about the cause of an illness, its prevalence, routes of exposure, and personal risk are all important, but it is also important to know how they are related to one another. Misunderstandings about the causes and prevalence of hazard, for example, may result in misperceptions about the likelihood of victimization. These different aspects of knowledge and belief are viewed as a 'mental model' that describes both the separate pieces and the connections between these pieces (Bostrom et al., 1992).

Hazard Probabilities

Much attention has been given to how people think about hazard probabilities. (Apparently researchers think that hazard severity is a simpler issue, for this topic has received little attention.) Although proportions (e.g., '0.25'), percentages (e.g., '25%'), frequencies (e.g., 25 out of 100), and odds (e.g., '1 in 4') are the units used by scientists to measure probability; there is abundant evidence that laypeople often have great difficulty understanding and using such information. For instance, people sometimes give a higher numerical estimate to the likelihood of *dying* of breast cancer than to the *likelihood* of developing breast cancer. Thus, the ability to recite the correct percentage, rate, or odds should not be used as the criterion to decide whether people truly understand the probability. Except for percentages given in weather forecasts, the public is rarely exposed to probability statistics, so it is not surprising that many people find numerical information about hazard probabilities unhelpful.

Some evidence indicates that people find verbal designations of probability (e.g., 'small chance of happening') to be easier to use and to better reflect their own views than numerical risk measures (Diefenbach et al., 1993), but people vary greatly in the numbers they associate with particular verbal labels, making the conversion of numbers to words highly subjective. Expressions of probability in terms of odds (e.g., 1 in 50 chances) are especially disliked by laypeople.

Because it seems impossible to navigate an uncertain world without forming estimates of the likelihood that particular events will occur, we tend to assume that people must form and encode judgments of hazard probabilities. In fact, it is possible that people rarely do encode such information. Instead they may generate risk judgments from past experiences and from their own implicit theories as needed. It is also possible that people store in memory only very crude probability information (e.g., no chance, small chance, large chance, certain) and never use the finer distinctions that preoccupy scientists and policy makers.

Because social comparison is a central feature of human existence, in everyday life people may find it more natural to make decisions on the basis of risk comparisons than on beliefs about absolute probability (the likelihood that an event will occur). Thus, a woman's interest in preventing heart disease may depend less on her beliefs about her absolute likelihood of a heart attack (e.g., 40% chance) than on her belief that her risk is greater than the risk of her friends, or even that her heart disease risk is less than her risk for other diseases.

Goals of Risk Perceptions

Obviously, accurate risk perceptions can help people make appropriate decisions about actions to avoid illness or injury. However, accuracy is only one of the goals guiding people as they formulate judgments about themselves, such as risk estimates (Kunda, 1990). Risk perception is not an unbiased appraisal of information, but an attempt to seek the most comforting view of one's personal vulnerability that fits within the bounds of the evidence. Beliefs about our vulnerability to harm can lead to positive or negative emotions (pleasure about successfully avoiding harm, anxiety about potential illness or disability). Perceptions of vulnerability can also affect our feelings of self-esteem. Because we want to believe that we are intelligent, competent, and foresighted, to admit that our actions place us at greater risk than our peers would be a threat to our self-esteem. Indeed, people are more likely to acknowledge their risk when they have a chance to reflect on their important values, thereby separating egoistic concerns from objective self-judgment (Harris and Epton, 2009). Thus, minimizing negative affect and maximizing self-esteem are also goals of risk perceptions, and these goals may be at variance with the goal to be accurate.

Optimistic Bias

Bias in risk perceptions is revealed especially clearly when people are asked to compare their risks with the risk of the rest of the population. This is typically done by asking people to indicate on a Likert-type scale how their risk compares with that of similar others (e.g., same age and sex), usually on a scale ranging from 'below average' to 'above average' with average as the midpoint. Sometimes people are asked on separate scales to estimate their own risk (e.g., from 'not at all likely' to 'extremely likely') and their peers' risk on separate scales, and a difference score is then computed; optimistic bias is present when the mean difference is less than zero. Using both types of measures, investigators have demonstrated that people often think their risk is lower than that of others (Weinstein, 1987). For example, when people are asked how their risk of a particular illness compares with the risk of others of the same age and sex, many claim to have below average risk and few admit to having above average risk. Across a variety of hazards, it is common to find 40–70% of a group asserting below average risk; another 30–50% saying that their risk is average; and less than 10% acknowledging that their risk is above average. Obviously, this distribution cannot be correct. The 'average' person has, by definition, an 'average' risk. So when the people who claim below average risk greatly outnumber those who say their risk is higher than average, implying that the average person is below average in risk, something must be wrong with their risk

judgments. It is interesting to note that people are often unrealistically pessimistic when estimating their absolute probability of experiencing hazards – for example, women greatly overestimate their risk of breast cancer – while being unrealistically optimistic about their comparative risk (e.g., Woloshin et al., 1999). Nevertheless, several (though not all) studies show that comparative risk perceptions may be more predictive than absolute risk perceptions of behavior, demonstrating that they are important to study (Klein, 1997).

Optimistic bias is not just denial, a blanket defense against anxiety. If it were, we would expect to find the greatest optimism among hazards, such as cancer, that generate the most fear. The amount of optimistic bias (that is, the extent to which people claiming below average risk exceed those who say their risk is above average) does vary from hazard to hazard (Weinstein, 1987), but it is not related to the severity of the hazard. Instead it is greatest for events that are believed to be preventable by individual action and events with which people have little personal experience. Bias is especially great if people think that vulnerability appears early, leading them to conclude that if they have not experienced the problem so far, they are free from future risk.

The belief that we are less at risk than others is apparently reassuring, since people resist attempts to correct their misperceptions. Most interventions designed to eliminate bias in comparative risk judgments – for example, having people think of the ways in which the negative event *could* happen or think of the precautions that they have *not* taken – have proved unsuccessful (Weinstein and Klein, 1995).

It may appear obvious that underestimating risk will expose people to unnecessary harm. Indeed, people who are unrealistically optimistic are less likely to worry about a hazard; learn about how to prevent it; and take adequate precautions, with downstream implications. For example, one study showed that people who where unrealistically optimistic about their risk of heart disease had higher levels of atherosclerosis (Ferrer et al., 2012) and another study showed the smokers who were unrealistically optimistic about their lung cancer risk had less interest in quitting (Dillard et al., 2006). However, a different view of this issue, known by the label 'positive illusions,' suggests that optimistic biases are really beneficial (Taylor and Brown, 1988). It points out that people who are overly optimistic about their ability to fight off the AIDS virus, for example, may be encouraged to maintain a healthy lifestyle and to bear with burdensome treatments. Thus, even if the perception of resistance is false, it may support other actions that are themselves beneficial. Although definitive evidence is lacking, it seems likely that underestimations of personal vulnerability decrease interest in prevention, especially when people see themselves as members of a group that does not need to be concerned about the risk, but optimism about surviving a life-threatening illness can help patients avoid depression and hopelessness (Wood et al., 1985).

Effects of Risk Perceptions on Health-Related Behaviors

Perceived probability and/or perceived severity occupy an important role in nearly all theories that are used to explain the health-related behavior of individuals (*see* Psychosocial Theories of Health Behavior). The literature on health behavior generally supports the idea that increases in perceived risk increase precautionary behavior (Brewer et al., 2007), but there are also numerous instances in which risk perceptions and behavior turn out to be unrelated.

This should not be surprising. There are many factors other than risk perceptions that influence behavior, and people who think that they have a high risk may nevertheless doubt their ability to carry out precautions or the efficacy of the precautions. In addition, most health behavior studies have used correlational research designs, which are problematic for examining risk perception–risk behavior linkages. Finally, as mentioned earlier, experimental interventions may fail to alter peoples' perceptions of their invulnerability.

The link between perceptions and action may depend on the type of action. Precautions that prevent a health problem from occurring (such as vaccinations) should reduce both perceived risk and worry about that risk. In contrast, being tested for the presence of an existing health problem (as in cancer screening) offers the immediate potential of extremely distressing information, especially if the problem is difficult to treat, so high perceived risk may be associated with high fear. The links between risk perceptions and different types of action still need to be elucidated.

See also: Attitudes and Behavior; HIV Risk Interventions; Health Risk Perception; Optimism and Pessimism; Personality and Risk Taking; Sexual Risk Behaviors.

Bibliography

Bostrom, A., Fischhoff, B., Morgan, M.G., 1992. Characterizing mental models of hazardous processes: a methodology and an application to radon. Journal of Social Issues 48, 85–100.

Brewer, N.T., Chapman, G.B., Gibbons, F.X., Gerard, M., McCaul, K.D., Weinstein, N.D., 2007. A meta-analysis of the relationship between risk perception and health behavior: the example of vaccination. Health Psychology 26, 136–145.

Diefenbach, M.A., Weinstein, N.D., O'Reilly, J., 1993. Scales for assessing perceptions of health-hazard susceptibility. Health Education Research 8, 181–192.

Dillard, A.J., McCaul, K.D., Klein, W.M.P., 2006. Unrealistic optimism in smokers: implications for smoking myth endorsement and self-protective motivation. Journal of Health Communication 11, 93–102.

Ferrer, R.A., Klein, W.M.P., Zajac, L.E., Sutton-Tyrrell, K.C., Muldoon, M.F., Kamarck, T.W., 2012. Unrealistic optimism is associated with subclinical atherosclerosis. Health Psychology 31, 815–820.

Harris, P.R., Epton, T., 2009. The impact of self-affirmation on health cognition, health behavior, and other health-related responses: a narrative review. Social and Personality Psychology Compass 3, 962–978.

Klein, W.M., 1997. Objective standards are not enough: affective, self-evaluative, and behavioral responses to social comparison information. Journal of Personality and Social Psychology 72, 763–774.

Kunda, Z., 1990. The case for motivated reasoning. Psychological Bulletin 108, 480–498.

Leventhal, H., Diefenbach, M.A., Leventhal, E.A., 1992. Illness cognition: using common sense to understand treatment adherence and affect cognition interactions. Cognitive Therapy and Research 16, 143–163.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., Combs, B., 1978. Judged frequency of lethal events. Journal of Experimental Psychology: Human Learning and Memory 4, 551–578.

Loewenstein, G.F., Weber, E.U., Hsee, C.K., Welch, N., 2001. Risk as feelings. Psychological Bulletin 127, 267–286.

Slovic, P., 1987. Perception of risk. Science 236, 280–285.
Slovic, P., Finucane, M.L., Peters, E., MacGregor, D.G., 2004. Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. Risk Analysis 24 (2), 311–322.
Taylor, S.E., Brown, J.D., 1988. Illusion and well-being: a social psychological perspective on mental health. Psychological Bulletin 103, 193–210.
Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. Science 185, 1124–1131.
Weinstein, N.D., 1987. Unrealistic optimism about susceptibility to health problems: conclusions from a community-wide sample. Journal of Behavioral Medicine 10, 481–500.
Weinstein, N.D., Klein, W.M., 1995. Resistance of personal risk perceptions to debiasing interventions. Health Psychology 14, 132–140.
Weinstein, N.D., Kwitel, A., McCaul, K.D., Magnan, R.E., Gerrard, M., Gibbons, F.X., 2007. Risk perceptions: assessment and relationship to influenza vaccination. Health Psychology 26 (2), 146–151.
Woloshin, S., Schwartz, L.M., Black, W.C., Welch, H.G., 1999. Women's perceptions of breast cancer risk: how you ask matters. Medical Decision Making. 19, 221–229.
Wood, J.V., Taylor, S.E., Lichtman, R.R., 1985. Social comparison in adjustment to breast cancer. Journal of Personality and Social Psychology 49, 1169–1183.

Health Risk Perception

Britta Renner, Martina Gamp, Ralf Schmälzle, and Harald T Schupp, University of Konstanz, Konstanz, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Perceptions of health-related risks are a prerequisite for taking protective action, adopting a healthier lifestyle, attending health screenings, and adhering to medical care. It seems inherently plausible that the greater the perceived risk for one's own health is the greater the motivation for protective action. Accordingly, it is important to understand how people perceive health risks, how accurate these perceptions are, and how information about one's own health risk is received. This article focuses on general and personal risk perceptions, discusses the role of intuition in personal risk perception, and presents findings regarding reactions to individualized feedback about risk.

Introduction: Why Study Risk Perception?

Risk perception, i.e., how individuals think and feel about the risks they face, is an important determinant of protective behavior. For instance, responses to the A/H1N1 virus or 'swine flu' outbreak in 2009 showed that the success of public health intervention programs is largely dependent on individual risk perception. Although many governments launched large-scale vaccination campaigns, vaccine uptake in the general population was very low. In Germany, for example, only 6.8% of the general population followed the recommendation to be vaccinated, even though vaccination is the most effective intervention for preventing influenza and A/H1N1 (Renner and Reuter, 2012).

Accordingly, it seems inherently plausible that people not only need to be aware of an existing health risk ('Many people are becoming infected with a new type of influenza.'), but they also need to feel personally at risk ('I might catch the new influenza myself.') in order to take protective action. It is thus critical to understand (1) how people construe and evaluate health risks in general (general risk perception), (2) how they gauge their own personal risk (personal risk perception), and (3) how they react to information indicating that they are personally at risk.

General Risk Perception

Risk Assessment and the Numbers of Risk

Risk is commonly defined as a multiplicative combination of the probability of a hazardous event occurring (e.g., smoking) and the severity of the resulting negative consequences (e.g., lung cancer). This definition of risk, as 'probability × severity,' implies that greater probability and greater severity result in greater overall risk (Slovic, 2000).

Numerically determining the probability of adverse events and possible adverse consequences is one of the main challenges for scientific *risk assessment* (Slovic, 2000). Risks to health can be assessed on the basis of different metrics, such as morbidity and mortality rates (number of cases of disease or deaths per year). For example, there were 2 426 264 deaths in the United States in 2006. While heart diseases led to 631 636 deaths in the United States in 2006 (26%), accidents were responsible for 121 599 deaths (5%) (Heron et al., 2009). Thus, heart diseases posed a five times greater risk in comparison to accidents. Moreover, the magnitude of risk can be framed in many numeric formats, such as natural frequencies (e.g., 5 out of 100), percentages (e.g., a 5% chance of contraction), or probabilities (e.g., 0.05), and in many different visual formats (e.g., bar charts, pie graphs, or icons displaying the number of individuals affected). All of these numeric formats are equally valid representations of 'the numbers of risk' in the realm of risk assessment.

In contrast to 'risk assessment,' 'risk perception' considers how people, in particular laypeople, construe risk. Wright et al. (2002) examined risk judgments made by experts (life insurance underwriters) and laypeople (business students) for 31 hazards (smoking, asthma, etc.) and compared them to the actual risk level (annual fatalities). The correlation between risk judgments and factual risk for fatalities was $r = 0.66$ for the experts and $r = 0.73$ for the laypeople. Thus, both laypeople and experts were capable of ordering the hazards by the absolute risk level in the metric of annual fatalities. However, the accuracy of lay perceptions commonly decreases when it comes to estimating more complex numbers of risk, and systematic deviations between risk assessment and lay risk perception occur (see for further discussion Gigerenzer et al., 2008; Renner and Schupp, 2011).

How well risk statistics are understood depends on numeracy, i.e., the ability to process basic numerical and probability concepts (Reyna et al., 2009). As expected, highly numerate adults are more likely to extract and use appropriate numerical principles and thereby have more complete and complex information than less numerate individuals. Furthermore, the issue of which type of numeric and visual formats facilitates accurate perceptions of risks has been addressed by many studies (Gigerenzer et al., 2008). One of the most striking effects emerged from the comparison of relative and absolute numeric risk formats. Relative risk formats describe risk in conditional probabilities, percentages, or ratios without defining the baseline or absolute risk level. Thus, a risk factor may be described as conferring '50% more risk' than another, even when the absolute increase is only from 2 to 3%. Without information on the baseline risk level, people tend to overestimate not only the impact of risk factors but also the impact of precautionary measures. Furthermore, conveying risk in

natural frequencies (e.g., 5 out of 1000 instead of 0.5%) facilitated the understanding of risk within laypeople (cf also Ferguson and Starmer, 2013; Gigerenzer et al., 2008). Thus, the pessimistic view that the public is not capable of understanding the numbers of risk may at least partly be the result of using suboptimal formats when communicating risk. Sensible risk communication should ideally present an array of risk information, i.e., absolute and relative risk information, mortality rates, and survival rates, as well as natural frequencies, in order to increase risk literacy in laypeople and experts over time.

How People Construe Risk: The Multidimensionality of General Risk Perception

Risk assessment can be seen as an objective approach to determining risk, which commonly considers two core components: the severity of the negative consequences and the probability of an occurrence of a hazardous event. However, numerous studies suggest that risk perception in the layperson is more complex, and that it is influenced by risk characteristics other than 'probability' and 'severity.' The issue has been most thoroughly examined with respect to the psychometric paradigm, which provides a systematic attempt to determine the characteristics associated with risk.

The Psychometric Paradigm

A pioneering study by Fischhoff et al. (1978) examined risk perceptions of various hazards, which were evaluated in terms of several risk attributes. Factor analysis was used to condense ratings of risk attributes into two orthogonal dimensions of 'dread' and 'unknown risk.' The factor 'dread' captures aspects such as perceived control over exposure to the risk, the degree of catastrophic consequences, or global ramifications. The factor 'unknown risk' refers to the degree to which a risk is predictable, observable, and understood. The psychometric paradigm plays a prominent role in risk perception research and numerous studies have confirmed the two factors, 'dread' and 'unknown risk' (see Sjöberg et al., 2004, for a review). Overall, despite noted shortcomings (e.g., Sjöberg et al., 2004), this work constitutes a landmark in risk perception research by showing that characteristics beyond the 'probability' by 'severity' calculus influence public perception and attitudes toward risk.

Intuition and Feelings in Risk Perception

The factor 'dread,' identified by research using the psychometric paradigm, clearly alludes to emotional characteristics. However, emotions were seldom discussed explicitly in earlier theories of risk perception. This has now changed and emotion has become a key area of risk perception research. For instance, Loewenstein et al. (2001) propose that emotional feelings can affect risk perceptions independently from numerical cognitions (probability, severity) or other attributes of risks (e.g., controllability, catastrophic potential). Similarly, Slovic and Peters (2006) distinguish between an intuitive mode of risk perception based on feelings and the rational, deliberative analysis of risk. Both approaches share the notion that feelings of risk stem from rapid and largely automatic routines that reflect an association-based, parallel mode of thinking, which is often referred to as 'intuition.' In contrast, the so-called 'rational' mode of risk analysis is presumed to depend on resource-dependent, effortful, and serial processes (see also Pachur et al., 2012).

One challenge for examining intuitive risk perception is that rapid and largely automatic processing is hard to capture. Neuroimaging methods, such as functional magnetic resonance imaging (fMRI) and event-related potentials (ERPs), allow measuring of rapid, introspectively opaque, and affectively charged processes and may accordingly be useful in examining the intuitive processes involved in risk perception. A series of ERP studies on HIV risk perception (e.g., Renner et al., 2012; Schmälzle et al., 2011, 2012) provides an example of this novel approach. Previous research using focus groups and retrospective interviews with HIV-positive people had revealed that people have spontaneous impressions of risk – i.e., they often reported that they were convinced that their partners were safe. These impressions about a potential partner's HIV risk may make people prone to rely on 'illusory control strategies,' such as selecting 'safe-looking' partners. To uncover the mechanisms behind these intuitive HIV risk perceptions, neural responses were measured while participants spontaneously evaluated the HIV risk of unacquainted individuals depicted in photographs. Demonstrating the fast and frugal risk perception mode, ERPs reliably differentiated between individuals perceived as risky and safe early in the processing stream (e.g., Schmälzle et al., 2011). The early onset of this electrocortical differentiation (<300 ms) precedes systematic reasoning about health risks and supports the notion of intuitive as opposed to analytic processing. Furthermore, the late positive potential (LPP), a specific ERP component that has been linked to affective evaluation processes, was also associated with HIV risk perception. In particular, individuals perceived as risky elicited larger LPPs compared to individuals perceived as safe (see **Figure 1**; Schmälzle et al., 2011). These findings were corroborated by fMRI measurements, which revealed that individuals perceived as risky activated regions of the saliency network, i.e., the anterior insulae and medial frontal cortex, which are also engaged by threatening and negative-affect-related stimuli (Häcker et al., 2014). Based on these findings, when people report that they 'just know' the risk posed by a certain individual, their feeling of risk may reflect the implicit assessment of personal characteristics related to a HIV risk stereotype. Extrapolating from this line of research, the integration of experimental methods, i.e., self-reports, behavioral observation, and psychophysiology, seems highly promising for uncovering the intuitive sensing of health risks.

Personal Risk Perceptions

At first glance, it seems obvious that perceiving a health threat is a prerequisite for the motivation to change a risk behavior. However, from a psychological perspective, the process is far more complex: people not only need to know about the existence of a health risk (general risk perception; e.g., 'Many people are becoming infected with a new type of influenza.'), but they need to feel personally at risk (personal risk

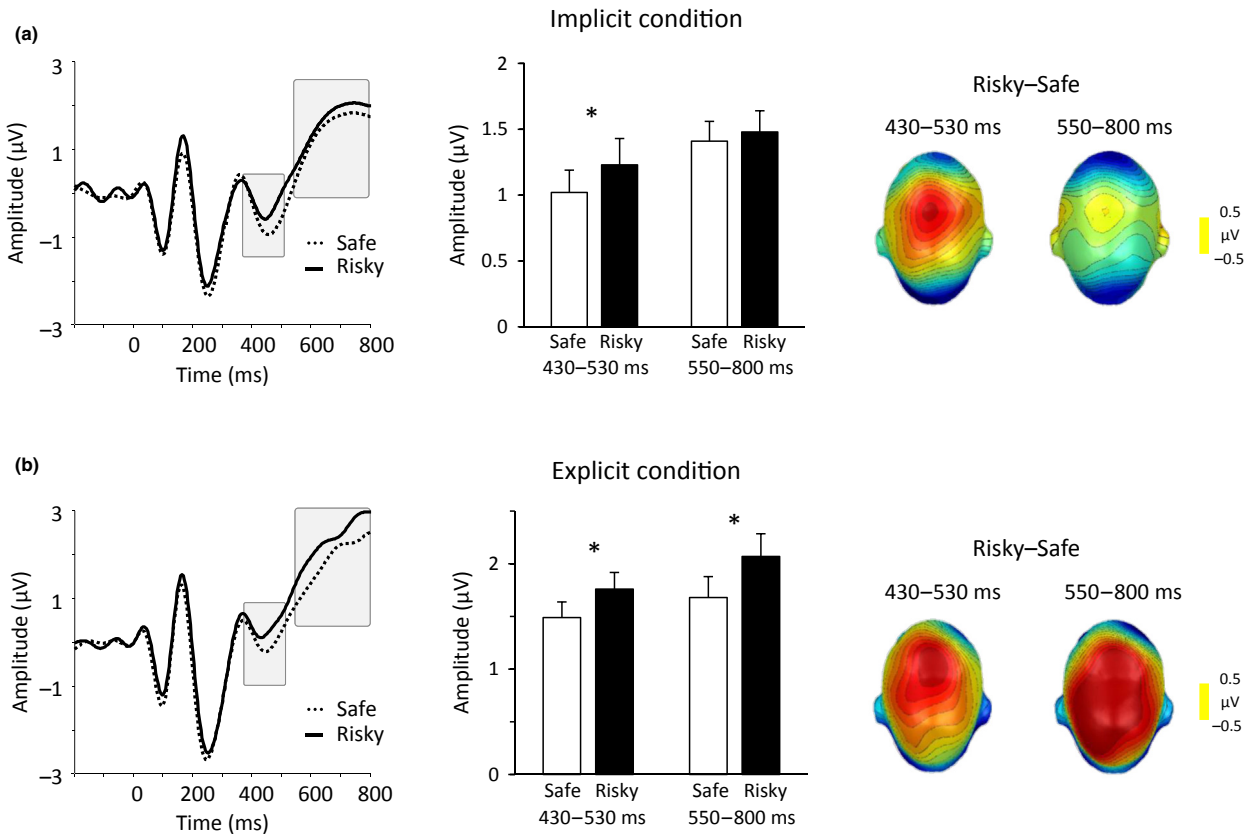

Figure 1 Neural signature of (a) implicit and (b) explicit HIV risk perceptions. Reproduced from Schmälzle, R., Schupp, H.T., Barth, A., Renner, B., 2011. Implicit and explicit processes in risk perception: neural antecedents of perceived HIV risk. Frontiers in Human Neuroscience 5, 43. doi:10.3389/fnhum.2011.00043.

perception, e.g., 'I might catch the new influenza myself.') in order to take protective action. In general, two different types of personal risk perceptions are distinguished: (1) absolute personal risk perceptions and (2) comparative personal risk perceptions (Renner and Schupp, 2011; Shepperd et al., 2013; see also Table 1).

Absolute Personal Risk Perceptions

One of the most common ways of assessing people's perception of their own health risks is to ask them to estimate their personal absolute risk in numerical form ('My risk of catching the new influenza this winter is 70%.'). Comparing personal absolute risk perception with epidemiological or actual risk estimates shows that they often greatly deviate. Similar to general risk perceptions, the likelihood of contracting a disease is often drastically overestimated. Lipkus et al. (2001), for example, asked adult women between the ages of 45 and 54 years to estimate their lifetime breast cancer risk. In addition, they measured the actual breast cancer risk using the Gail score, which calculates breast cancer risk on the basis of seven risk factors (e.g., age, race, age at menarche, age at first live birth, number of first-degree relatives with breast cancer). On average, compared to their actual lifetime risk (8%), the women greatly overestimated their lifetime breast cancer risk (34%), indicating an unrealistic pessimistic bias in absolute personal risk perception at the group level (see Table 1).

Comparative Personal Risk Perceptions

Although people often overestimate their personal absolute risk, they are often convinced that their risk is lower than that of other people. The same women who overestimated their personal absolute breast cancer risk in the study by Lipkus et al. (2001) were convinced that their comparative risk was below average (see Table 1). Numerous studies have shown that when people are asked to rate their chances of experiencing certain illnesses, and other problems, most of them report their risk is below average (see Shepperd et al., 2013). However, if people believe on average that their risk is below average, they are systematically underestimating their personal risk. This bias in comparative personal risk perception has been labeled *unrealistic optimism* or *optimistic bias* (Perloff and Fetzer, 1986; Weinstein, 1980). Unrealistic optimism has been demonstrated for men and women across age-groups and education levels (see Renner and Schupp, 2011; Weinstein, 2003).

It is important to note that the 'classical unrealistic optimism' only indicates a bias in comparative personal risk perceptions at the level of the group (see Table 1). A woman who says that her breast cancer risk is 'below average' might

Table 1 Types of personal risk perceptions and examples for optimistic/pessimistic biases

	Type of actual risk standard	
Type of risk perception	*Individual-based (e.g., individual health risk assessment (HRA) such as an individual Gail score or CVD Framingham Risk Score)*	*Group-based (e.g., annual disease incidence)*
Absolute personal risk perception (absolute unrealistic pessimistic/optimistic bias)	1. Pessimistic A woman believes her breast cancer risk is 34% but the objective risk estimate (e.g., Gail score) indicates that her risk is 8%. 2. Optimistic A woman believes her breast cancer risk is 2% but the objective risk estimate (e.g., Gail score) indicates that her risk is 8%.	1. Pessimistic A representative group of women believes their breast cancer risk is on average 34% but the annual incidence rate is 8%. 2. Optimistic A representative group of women believes their breast cancer risk is on average 2% but the annual incidence rate is 8%.
Comparative personal risk perception (comparative unrealistic pessimistic/optimistic bias)	1. Pessimistic A woman believes her breast cancer risk is higher than the risk of an average woman of her age (average peer) but the objective risk estimate (e.g., Gail score) indicates that her risk is lower than the average peer risk. 2. Optimistic A woman believes her breast cancer risk is lower than the risk of an average woman of her age (average peer) but the objective risk estimate (e.g., Gail score) indicates that her risk is higher than the average peer risk.	1. Pessimistic A representative group of women believes on average their breast cancer risk is higher than the risk of an average woman of their age (average peer). 2. Optimistic A representative group of women believes on average their breast cancer risk is lower than the risk of an average woman of their age (average peer). =‘classical unrealistic optimism’

Reproduced from Renner, B., Schupp, H., 2011. The perception of health risks. In: Friedman, H.S. (Ed.), Oxford Handbook of Health Psychology. Oxford University Press, New York, pp. 637–665 and Shepperd, J.A., Klein, W.M.P., Waters, E.A., Weinstein, N.D., 2013. Taking stock of unrealistic optimism. Perspectives on Psychological Science 8 (4), 395–411. doi:10.1177/1745691613485247.

give a somewhat optimistic risk rating, but if she has a low risk factor profile, her optimistic rating may actually be a realistic perception of her risk status (cf Weinstein, 2003). To assess the optimistic bias at the individual level, objective risk measures are required such as ‘health risk appraisals’ that estimate a person’s actual risk based on characteristics such as medical history, blood pressure, smoking habits, and epidemiological data (e.g., CVD Framingham Risk Score at http://cvdrisk.nhlbi.nih.gov).

Why do people feel less at risk than others? Research has posited various explanations for this intriguing phenomenon, involving both motivational and cognitive processes. Motivational accounts postulate that the optimistic bias is fueled by the motivation either to self-enhance or protect, and maintain a positive view of one’s health. However, a purely motivational account cannot sufficiently explain this phenomenon. If people predominantly bias comparative risk perceptions because they want to maintain a positive self-view, serious diseases should provoke an increased degree of optimistic bias because they are particularly threatening to the self. Contrary to the idea that biased comparative risk perceptions are a defense against feelings of threat and anxiety, unrealistic optimism is generally no greater for serious, life-threatening hazards than for more mundane problems (Harris et al., 2008).

Interestingly, some hazards appear to be associated with a strong and reliable comparative optimistic bias (e.g., STDs, alcohol problems), whereas other hazards commonly elicit only a weak or no bias (e.g., colds, cancer). A seminal study by Neil Weinstein (1980) observed that unrealistic optimism was pronounced when the hazard was perceived as controllable and associated with a vivid victim stereotype. People find it easier to picture stereotypical victims of controllable events (Hahn and Renner, 1998; Perloff and Fetzer, 1986; Thornton et al., 2002). For example, lung cancer is perceived as a controllable and behavior-dependent disease, and people can easily picture a typical victim (Hahn and Renner, 1998). This may explain why lung cancer typically elicits a pronounced unrealistic optimism, whereas cancer in general does not. The less similar a person believes he or she is to the high-risk stereotype, the safer the person will perceive him- or herself to be (Perloff and Fetzer, 1986; see Renner and Schupp, 2011).

One intriguing conclusion from this research is that public health education campaigns may often facilitate, instead of reducing, unrealistic optimistic risk perceptions. The guiding principles proposed by campaigns for health risk reduction usually include vivid presentations of risk factors and show high-risk persons. This may foster risk stereotypes that are perceived as dissimilar and thereby may lead to an underestimation of personal risk if individuals think that many risk factors do not apply to them. Consequently, risk communication that only provides information about general risk may make people aware of a risk (‘Smoking causes coronary heart disease.’) but also lead to an underestimation of self-risk (‘It is unlikely that this will happen to me.’) by creating high-risk stereotypes. Accordingly, providing information about an individual’s

personal risk status should be less ambiguous and more effective for facilitating realistic personal risk perception and protective behavior.

Reactions toward Personalized Risk Feedback

In recent years, health risk assessment tools have been developed to assess individual risk for particular diseases. For example, the National Institute of Health in the United States offers an online tool for estimating the 10-year risk of having a heart attack (http://cvdrisk.nhlbi.nih.gov/). The increasing availability of such services over the Internet and the rise of over-the-counter testing kits underscore the need for a better understanding of how individuals react to such feedback. In an agenda-setting study, Jemmott et al. (1986) devised an experimental paradigm that enables the study of reactions toward personalized health risk feedback. Participants who were led to believe that they suffered from a hypothetical thioamine acetylase (TAA) deficiency perceived their test result as less accurate and rated TAA deficiency as a less-serious health threat when compared to the non-TAA group (for a review, see Croyle et al., 1997). The reduced acceptance of negative health risk information compared to positive one is a very robust phenomenon, evident across a wide range of diseases and samples (for an overview see Ditto, 2009; Renner and Schupp, 2011).

The Motivated Reasoning Perspective

Why is bad news less accepted than good news? The lower acceptance of negative health risk feedback than positive one is commonly seen as clear-cut evidence for 'self-defensive denial' or 'motivated reasoning' (e.g., Croyle et al., 1997; Helzer and Dunning, 2012; Kunda, 1990; McQueen et al., 2013). People who are informed having an elevated risk of disease appear to derogate the validity of the risk factor test in order to maintain a favorable sense of their health. At first glance, the motivated reasoning account, which proposes self-defensive denial in information processing, poses a straightforward explanation for the frequently observed asymmetrical acceptance pattern: people do not like bad news; consequently, they try to derogate it with superficial rational strategies. However, several limitations, such as the 'adaptive paradox,' have been noted (cf Ditto and Lopez, 1992; Renner and Schupp, 2011; but see McQueen et al., 2013). If the predominant response to threatening information is self-defensive denial, recipients should not see a need for adaptive action and protection, ultimately resulting in serious harm. In contrast, studies assessing intentions and behavior change commonly report adaptive responses instead of the predicted unresponsiveness to information that signals a health threat (e.g., Shiloh et al., 2013; see for a review Sheeran et al., 2013).

The Quantity of Processing Perspective

In an attempt to explain the asymmetrical acceptance of health risk information, Ditto and Lopez build upon the principle that negative information generally has greater impact than neutral or positive information (1992; for a review see Ditto, 2009). Their quantity of processing (QOP) view draws upon the fact that negative information triggers more elaborate cognitive analyses than positive information does. The more deeply people think about negative information, the more likely they are to consider plausible alternative explanations, producing greater uncertainty regarding the validity of the information. As a consequence, negative information is less likely to be accepted than positive information. Supporting this notion, participants who received negative TAA feedback accepted low-quality feedback less readily than high-quality TAA feedback. Moreover, while positive feedback was highly accepted regardless of whether it was of high or low quality, participants only accepted negative information to a similar degree when they believed the information was high quality (Ditto et al., 1998). These results are in line with the QOP view that negative risk information triggers more elaborate processing than positive risk information does, and that negative information is, therefore, less likely to be accepted than positive information (see also Mata et al., 2013). However, from a motivational reasoning perspective, one could argue that people receiving highly valid, negative TAA risk feedback only accepted it because there was virtually no leeway for derogating the information. Thus, it remains unclear whether the unambiguous acceptance of good news or differential acceptance of bad news indicates an adaptive or self-defensive denial reaction pattern.

The Cue-Adaptive Reasoning Perspective

It is commonly assumed that positive health risk feedback such as "Your lifetime risk for developing colorectal cancer is 10:10,000" is met with great acceptance and reassurance (Wardle et al., 2003). Curiously, this seems not always to be the case. Weinstein et al. (2004) examined primary care patients between 40 and 70 years of age who received computerized risk assessment feedback about their colon cancer risk. For most patients, the risk feedback indicted a lower risk status than they had expected (average actual risk: 10:10.000 vs average perceived pretest risk: 146:10.000). According to the motivated reasoning and QOP views, receiving such good news should have resulted in pronounced acceptance of the feedback. However, an opposite pattern of result emerged: post-feedback-perceived risk was in many cases still substantially higher than the actual risk, indicating that people often distrusted their personalized risk feedback. Similar resistance toward positive feedback or 'lack of reassurance' has been observed in various empirical studies (e.g., Dillard et al., 2006; Linnenbringer et al., 2010). One might argue that the personalized risk feedback was difficult for the recipients to understand and thus, the findings do not indicate a lack of reassurance but rather a lack of understanding. However, in an experimental study using personalized risk feedback about a (fictitious) fatigue syndrome, Gamp and Renner (manuscript submitted for publication) found a similar lack of reassurance after positive information in post-feedback-perceived risk. Conversely, post-feedback risk for an average peer was accurately adapted to the given peer risk feedback score. Hence, people are generally capable of

understanding risk information and adapting their risk estimations accordingly. This raises the question, when and why do people resist self-related positive health risk feedback?

The cue-adaptive reasoning account (CARA) assumes that both negative feedback and risk feedback, which conflicts with preexisting risk perceptions, serve as cues that draw attentional resources for more elaborate processing (Panzer and Renner, 2009; Renner, 2004). If people spend more cognitive resources on negative or unexpected risk information, plausible alternative explanations are more likely taken into account. This, in turn, increases the likelihood that negative or unexpected risk information is perceived as being less valid and as a consequence it is less likely to be accepted than expected positive risk information. One important implication of this reasoning is that not only unexpected bad news, but also unexpected good news is received with greater reluctance. Data from a community cholesterol screening (Renner, 2004; see also Gamp and Renner, manuscript submitted for publication) showed that participants who received unexpected or expected negative risk feedback were sensitive to the quality of the given feedback. Indicating more elaborate information processing, feedback acceptance was significantly higher for high-quality risk feedback than low-quality (see **Figure 2**). Importantly, participants receiving unexpected positive risk feedback also showed sensitivity to feedback quality, and low-quality feedback was rated as less accurate than high-quality. The only group that was not sensitive to the quality of the feedback was the group that expected and received positive risk feedback information. This group accepted the given feedback independently of the quality. A similar pattern was also found for behavioral-proximal measures.

The assumption that unexpected good news also receives deeper processing might first seem counterintuitive. However, especially in the context of personally consequential feedback, it represents an adaptive response because accepting false 'all-clear' health risk feedback could result in the refusal or cessation of prophylactic measures. In this case, resources previously invested in preventive actions would be expended, while severe harm would not be prevented by protective action. Accordingly, the CARA view conceptualizes both, the lack of reassurance following unexpected positive health risk feedback and the asymmetrical acceptance of negative versus positive risk feedback as an unintentional by-product of the adaptive allocation of cognitive processing resources.

Conclusion

It is widely agreed that people's understanding of health risks is of central relevance for effective risk communication and health interventions. Being aware of health risks is certainly a precondition for accurate risk perception. But even when people are aware of a risk (general risk perception), this does not mean that they regard themselves as personally at risk (personal risk perception). Public health interventions portraying people at high risk may foster high-risk stereotypes and, as a consequence, induce a paradoxical effect: people might be well informed about the magnitude of the risk but at the same time they might be convinced that this misfortune will not happen to them because they feel highly dissimilar to the high-risk stereotype. Thus, although general risk perceptions might be substantial, personal risk perceptions might be low and optimistically biased. Research on personal risk perception has focused on absolute as well as comparative aspects of risk perception, and recent approaches adopted neuroimaging in order to shed light on the mechanisms of risk perception, whose nature is presumably less numerical and analytic, but rather more intuitive-experiential and affective. To cross the Rubicon between general and personal risk perceptions,

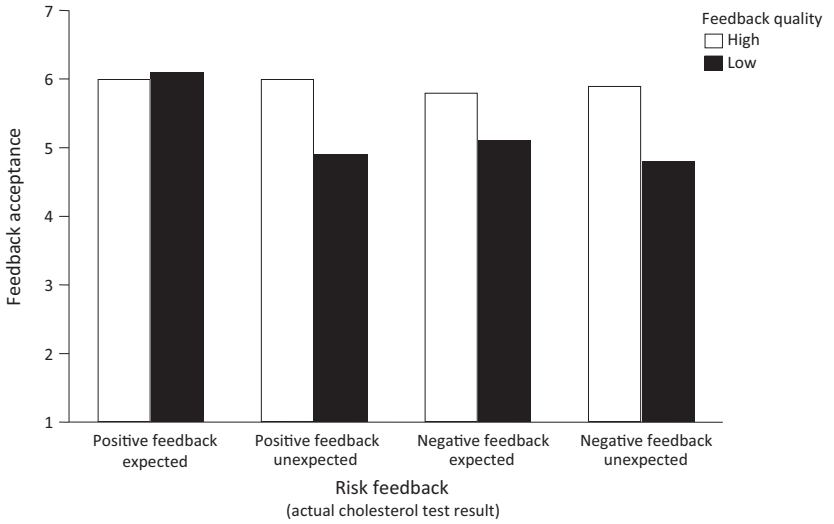

Figure 2 Feedback acceptance as a function of risk feedback valence, risk expectancy, and feedback quality. Reproduced from Renner, B., 2004. Biased reasoning: adaptive responses to health risk feedback. Personality & Social Psychology Bulletin 30 (3), 384–396. doi:10.1177/0146167203261296.

recipients should receive personalized risk information. However, personalized risk information is filtered through preexisting risk perceptions, and conflicting information is carefully scrutinized before it is accepted. Interestingly, this not only holds for negative risk feedback, but also applies to unexpected positive risk feedback. Therefore, to avoid increasing concern among participants or the need for retesting, an 'all-clear' nonelevated risk status should be as carefully explained and discussed as an elevated risk status health.

See also: Cancer Screening; Control Beliefs: Health Perspectives; Depression, Pessimism, and Health; Doctor–Patient Interaction in the West: Psychosocial Aspects; Explanatory Style and Health; Health Behavior, Psychosocial Theories of; Health Behaviors; Health Education and Health Promotion; Health Interventions, Community-based; Health Literacy; Health Promotion in Schools; Health Risk Appraisal and Optimistic Bias; Personalized Medicine; Sexual Risk Behaviors; Smoking and Health.

Bibliography

Croyle, R., Sun, Y., Hart, M., 1997. Processing risk factor information: defensive biases in health-related judgments and memory. In: Petrie, K., Weinman, J. (Eds.), Perceptions of Health and Illness: Current Research and Applications. Harwood Academic Publishers, Amsterdam, pp. 267–290.

Dillard, A.J., McCaul, K.D., Kelso, P.D., Klein, W.M.P., 2006. Resisting good news: reactions to breast cancer risk communication. Health Communication 19 (2), 115–123. http://dx.doi.org/10.1207/s15327027hc1902_3.

Ditto, P.H., 2009. Passion, reason, and necessity: a quantity-of-processing view of motivated reasoning. In: Bayne, T., Fernández, J. (Eds.), Delusion and Self-deception: Affective and Motivational Influences on Belief Formation. Psychology Press, New York, pp. 23–53.

Ditto, P.H., Lopez, D.F., 1992. Motivated skepticism: use of differential decision criteria for preferred and nonpreferred conclusions. Journal of Personality and Social Psychology 63 (4), 568–584. http://dx.doi.org/10.1037/0022-3514.63.4.568.

Ditto, P.H., Scepansky, J.A., Munro, G.D., Apanovitch, A.M., Lockhart, L.K., 1998. Motivated sensitivity to preference-inconsistent information. Journal of Personality and Social Psychology 75 (1), 53–69. http://dx.doi.org/10.1037//0022-3514.75.1.53.

Ferguson, E., Starmer, C., 2013. Incentives, expertise, and medical decisions: testing the robustness of natural frequency framing. Health Psychology 32 (9), 967–977. http://dx.doi.org/10.1037/a0033720.

Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S., Combs, B., 1978. How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. Policy Sciences 9 (2), 127–152. http://dx.doi.org/10.1007/BF00143739.

Gamp, M., Renner, B. Lack of absolute and comparative reassurance: how negative expectancies mitigate the effects of positive health risk feedback, manuscript submitted for publication.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L.M., Woloshin, S., 2008. Helping doctors and patients make sense of health statistics. Psychological Science in the Public Interest 8 (2), 53–96. http://dx.doi.org/10.1111/j.1539-6053.2008.00033.x.

Häcker, F., Schmälzle, R., Renner, b., Schupp, H.T., 2014. Neural correlates of HIV risk feelings. Social. Cognitive, and Affective Neuroscience. http://dx.doi.org/10.1093/scan/nsu093.

Hahn, A., Renner, B., 1998. Perception of health risks: how smoker status affects defensive optimism. Anxiety, Stress & Coping: An International Journal 11 (2), 93–112. http://dx.doi.org/10.1080/10615809808248307.

Harris, P.R., Griffin, D.W., Murray, S., 2008. Testing the limits of optimistic bias: event and person moderators in a multilevel framework. Journal of Personality and Social Psychology 95 (5), 1225–1237. http://dx.doi.org/10.1037/a0013315.

Helzer, E., Dunning, D., 2012. On motivated reasoning and self-belief. In: Vazire, S., Wilson, T.D. (Eds.), Handbook of Self-knowledge. Guilford Press, New York, pp. 379–396.

Heron, M., Hoyert, D., Murphy, S., Xu, J., Kochanek, K., Tejada-Vera, B., 2009. Deaths: final data for 2006. National Vital Statistics Reports 57 (14), 1–134.

Jemmott, J.B., Ditto, P.H., Croyle, R.T., 1986. Judging health status: effects of perceived prevalence and personal relevance. Journal of Personality and Social Psychology 50 (5), 899–905. http://dx.doi.org/10.1037/0022-3514.50.5.899.

Kunda, Z., 1990. The case for motivated reasoning. Psychological Bulletin 108 (3), 480–498. http://dx.doi.org/10.1037/0033-2909.108.3.480.

Linnenbringer, E., Roberts, J.S., Hiraki, S., Cupples, L.A., Green, R.C., 2010. "I know what you told me, but this is what I think:" perceived risk of Alzheimer disease among individuals who accurately recall their genetics-based risk estimate. Genetics in Medicine 12 (4), 219–227. http://dx.doi.org/10.1097/GIM.0b013e3181cef9e1.

Lipkus, I.M., Biradavolu, M., Fenn, K., Keller, P., Rimer, B.K., 2001. Informing women about their breast cancer risks: truth and consequences. Health Communication 13 (2), 205–226. http://dx.doi.org/10.1207/S15327027HC1302_5.

Loewenstein, G.F., Weber, E.U., Hsee, C.K., Welch, N., 2001. Risk as feelings. Psychological Bulletin 127 (2), 267–286. http://dx.doi.org/10.1037//0033-2909.127.2.267.

Mata, A., Ferreira, M.B., Sherman, S.J., 2013. Flexibility in motivated reasoning: strategic shifts of reasoning modes in covariation judgment. Social Cognition 31 (4), 465–481. http://dx.doi.org/10.1521/soco_2012_1004.

McQueen, A., Vernon, S.W., Swank, P.R., 2013. Construct definition and scale development for defensive information processing: an application to colorectal cancer screening. Health Psychology 32 (2), 190–202. http://dx.doi.org/10.1037/a0027311.

Pachur, T., Hertwig, R., Steinmann, F., 2012. How do people judge risks: availability heuristic, affect heuristic, or both? Journal Of Experimental Psychology: Applied 18 (3), 314–330. http://dx.doi.org/10.1037/a0028279.

Panzer, M., Renner, B., 2009. Spontaneous reactions to health risk feedback: a network perspective. Journal of Behavioral Medicine 32 (4), 317–327. http://dx.doi.org/10.1007/s10865-009-9206-7.

Perloff, L.S., Fetzer, B.K., 1986. Self-other judgments and perceived vulnerability to victimization. Journal of Personality and Social Psychology 50 (3), 502–510. http://dx.doi.org/10.1037//0022-3514.50.3.502.

Renner, B., 2004. Biased reasoning: adaptive responses to health risk feedback. Personality & Social Psychology Bulletin 30 (3), 384–396. http://dx.doi.org/10.1177/0146167203261296.

Renner, B., Reuter, T., 2012. Predicting vaccination using numerical and affective risk perceptions: the case of A/H1N1 influenza. Vaccine 30 (49), 7019–7026. http://dx.doi.org/10.1016/j.vaccine.2012.09.064.

Renner, B., Schmälzle, R., Schupp, H.T., 2012. First impressions of HIV risk: It takes only milliseconds to scan a stranger. PLoS ONE 7 (1), e30460. http://dx.doi.org/10.1371/journal.pone.0030460.

Renner, B., Schupp, H., 2011. The perception of health risks. In: Friedman, H.S. (Ed.), Oxford Handbook of Health Psychology. Oxford University Press, New York, pp. 637–665.

Reyna, V.F., Nelson, W.L., Han, P.K., Dieckmann, N.F., 2009. How numeracy influences risk comprehension and medical decision making. Psychological Bulletin 135 (6), 943–973. http://dx.doi.org/10.1037/a0017327.

Schmälzle, R., Schupp, H.T., Barth, A., Renner, B., 2011. Implicit and explicit processes in risk perception: neural antecedents of perceived HIV risk. Frontiers in Human Neuroscience 5, 43. http://dx.doi.org/10.3389/fnhum.2011.00043.

Schmälzle, R., Renner, B., Schupp, H.T., 2012. Neural correlates of perceived risk: the case of HIV. Social Cognitive and Affective Neuroscience 7 (6), 667–676.

Sheeran, P., Harris, P.R., Epton, T., 2013. Does heightening risk appraisals change people's intentions and behavior? A meta-analysis of experimental studies. Psychological Bulletin 140 (2), 511–543. http://dx.doi.org/10.1037/a0033065.

Shepperd, J.A., Klein, W.M.P., Waters, E.A., Weinstein, N.D., 2013. Taking stock of unrealistic optimism. Perspectives on Psychological Science 8 (4), 395–411. http://dx.doi.org/10.1177/1745691613485247.

Shiloh, S., Dagan, E., Friedman, I., Blank, N., Friedman, E., 2013. A follow-up study on men tested for BRCA1/BRCA2 mutations: impacts and coping processes. Psycho-oncology 22 (2), 417–425. http://dx.doi.org/10.1002/pon.2106.

Sjöberg, L., Moen, B., Rundmo, T., 2004. Explaining Risk Perception. An Evaluation of the Psychometric Paradigm in Risk Perception Research. Rotunde publikasjoner Rotunde no. 84.

Slovic, P.E., 2000. The Perception of Risk. Earthscan Publications, London.

Slovic, P.E., Peters, E., 2006. Risk perception and affect. Current Directions in Psychological Science 15 (6), 322–325. http://dx.doi.org/10.1111/j.1467-8721.2006.00461.x.

Thornton, B., Gibbons, F.X., Gerrard, M., 2002. Risk perception and prototype perception: Independent processes predicting risk behavior. Personality and Social Psychology Bulletin 28 (7), 986–999. http://dx.doi.org/10.1177/01467202028007011.

Wardle, J., Williamson, S., Sutton, S., Biran, A., McCaffery, K., Cuzick, J., Atkin, W., 2003. Psychological impact of colorectal cancer screening. Health Psychology 22 (1), 54–59. http://dx.doi.org/10.1037/0278-6133.22.1.54.

Weinstein, N.D., 1980. Unrealistic optimism about future life events. Journal of Personality and Social Psychology 39 (5), 806–820. http://dx.doi.org/10.1037//0022-3514.39.5.806.

Weinstein, N.D., 2003. Exploring the links between risk perceptions and preventive health behavior. In: Suls, J., Wallston, K.A. (Eds.), Social Psychological Foundations of Health and Illness. Blackwell Publishing, Malden, pp. 22–53.

Weinstein, N.D., Atwood, K., Puleo, E., Fletcher, R., Colditz, G., Emmons, K.M., 2004. Colon cancer: risk perceptions and risk communication. Journal of Health Communication 9 (1), 53–65. http://dx.doi.org/10.1080/10810730490271647.

Wright, G., Bolger, F., Rowe, G., 2002. An empirical test of the relative validity of expert and lay judgments of risk. Risk Analysis 22 (6), 1107–1122. http://dx.doi.org/10.1111/1539-6924.00276.

Health Self-Regulation, Motivational and Volitional Aspects of

Ralf Schwarzer, Freie Universität Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Health self-regulation refers to the motivational, volitional, and behavioral processes of abandoning health-compromising behaviors in favor of adopting and maintaining health-enhancing behaviors. The article addresses self-regulatory constructs and self-control models that examine how people resist temptations, persist at tasks, cope with adverse emotions, make choices, or overcome undesired impulses. A process model is described that suggests a distinction between preintentional motivation processes that lead to a behavioral intention (goal-setting phase), and post-intentional volition processes that lead to actual behavior (goal pursuit phase).

Introduction

Many health conditions are caused by risk behaviors, such as problem drinking, substance use, smoking, reckless driving, overeating, or unprotected sexual intercourse. The key question in health behavior research is how to predict and modify the adoption and maintenance of health behaviors. Fortunately, human beings have, in principle, control over their conduct. Health-compromising behaviors can be eliminated by self-regulatory efforts, and health-enhancing behaviors can be adopted instead, such as physical exercise, weight control, preventive nutrition, dental hygiene, condom use, or accident prevention. Health self-regulation refers to the motivational, volitional, and behavioral processes of abandoning such health-compromising behaviors in favor of adopting and maintaining health-enhancing behaviors (Mann et al., 2012).

Self-regulation, however, is a broader term going beyond the application to health behaviors and refers to all contexts where goal setting and goal striving are at stake. It is related to self-control, self-discipline, or willpower and applies to various domains of human functioning. In a general sense, self-regulation is a process that includes setting goals for oneself, what courses of action one chooses to pursue, how much effort one invests in given endeavors, and how long one perseveres in the face of barriers and setbacks. It encompasses reflective as well as automatic processes that guide goal-directed thoughts and behaviors in a more or less adaptive way.

Social–emotional–cognitive process models often distinguish phases of self-regulation and allocate persons in terms of their individual position within these phases. A useful distinction is the one between motivation and volition (Heckhausen and Gollwitzer, 1987). In the motivation phase, individuals are in a deliberative mindset while setting a goal (intention), whereas in the volitional phase, they are in an implementation mindset while pursuing their goal. Some health self-regulation theories are in line with such a process perspective (e.g., Bandura, 1997; Baumeister et al., 1994; Carver and Scheier, 1998; Kuhl, 2001; Mischel et al., 1996; Schwarzer, 2008). In such models, self-regulation is viewed as a multiphase process that involves a set of beliefs, expectancies, emotions, or actions.

As an alternative to the process view of self-regulation, some authors have identified specific constructs that may help to investigate how people engage in goal-directed actions, how they bring about desirable outcomes, or how they disengage from goals that appear unattainable. Whereas the social–cognitive self-regulation models are process based, perceived self-efficacy is a capacity-based belief and self-control is a capacity-based volitional strategy. Such constructs are viewed as personal resources that allow resisting temptations, modifying habits, and overriding impulses. With this capacity, individuals exert a conscious effort to alter their dominant responses to internal or environmental challenges.

In this article, we first discuss self-efficacy and self-control as self-regulatory constructs, and second we will describe a process-oriented self-regulation model that is mainly applied to health behavior change.

Self-Regulatory Constructs and Models

Perceived Self-Efficacy

Perceived self-efficacy portrays individuals' beliefs in their capabilities to exercise control over challenging demands (Bandura, 1997). It involves the regulation of thought processes, affective states, motivation, behavior, or changing environmental conditions. These beliefs are critical in approaching novel or difficult situations, or in adopting a strenuous self-regimen. People make an internal attribution in terms of personal competence when forecasting their behavior (e.g., 'I am certain that I can quit smoking even if my friend continues to smoke'). Self-efficacy influences the challenges that people take on as well as how high they set their goals (e.g., 'I intend to reduce my smoking,' or 'I intend to quit smoking altogether'). Some people harbor self-doubts and cannot motivate themselves. They see little point in even setting a goal if they believe they do not have what it takes to succeed. Thus, the intention to change a habit that affects health depends to some degree on a firm belief in one's capability to exercise control over that habit.

Perceived self-efficacy has been found to be important at all stages of health behavior change (Bandura, 1997), but it does not always constitute exactly the same construct. Its meaning depends on the particular situation of individuals who may be more or less advanced in the change process. A distinction between action self-efficacy, maintenance or coping self-efficacy, and recovery self-efficacy is useful, because, over the course of health behavior change, different tasks have to be

 http://dx.doi.org/10.1016/B978-0-08-097086-8.26031-5

mastered, and different self-efficacy beliefs are required to do so successfully. For example, a person might be confident in his or her capability to be physically active in general (i.e., high action self-efficacy), but might not be very confident to resume physical activity after a setback (low recovery self-efficacy).

Action self-efficacy (also called 'preaction or task self-efficacy') refers to the first phase of the process in which an individual does not yet act, but develops a motivation to do so. It is an optimistic belief during the preactional goal-setting phase. Individuals high in action self-efficacy imagine success, anticipate potential outcomes of diverse strategies, and are more likely to initiate a new behavior. Those with less self-efficacy imagine failure, harbor self-doubts, and tend to procrastinate.

Whereas (pre)action self-efficacy is instrumental in the motivation phase, the two following constructs are instrumental in the subsequent volition phase and can, therefore, also by summarized under the heading of 'volitional self-efficacy' that supports goal pursuit. *Maintenance* (*coping*) *self-efficacy* represents optimistic beliefs about one's capability to cope with barriers that arise during the maintenance period. A new health behavior might turn out to be much more difficult to adhere to than expected, but a self-efficacious person responds confidently with more adaptive strategies, more effort, and prolonged persistence to overcome such hurdles.

Recovery self-efficacy addresses the experience of failure and recovery from setbacks. If a lapse occurs, individuals may attribute their lapse to internal, stable, and global causes; dramatize the event; or interpret it as a full-blown relapse. High self-efficacious individuals, however, avoid this effect by attributing the lapse to an external high-risk situation and by finding ways to control the damage and to restore hope. Recovery self-efficacy pertains to one's conviction to get back on track after being derailed.

There is a functional difference between these three self-efficacy constructs, whereas their temporal sequence is less important. Different phase-specific self-efficacy beliefs may be harbored at the same point in time. The assumption is that they operate in a different manner. For example, recovery self-efficacy is most functional when it comes to resuming an interrupted chain of action, whereas action self-efficacy is most functional when facing a novel challenging demand. This distinction between phase-specific self-efficacy beliefs has proven useful in various domains of behavior change (Luszczynska and Schwarzer, 2003; Ochsner et al., 2013; Renner and Schwarzer, 2000). Action self-efficacy tends to predict intentions, whereas maintenance (coping) self-efficacy tends to predict behaviors. Individuals who have recovered from a setback obviously need different self-beliefs than those who have maintained their levels of activity.

Self-Control

Self-control is a capacity-based volitional strategy that is related to self-efficacy beliefs. People are more inclined to resist temptations, modify habits, and override impulses if they believe they can do it. With this capacity, individuals exert a deliberate effort to alter their behavioral responses such as emotions, thoughts, or impulses. More than four decades ago, Walter Mischel and collaborators studied self-control in children which became the groundwork for today's research (Mischel et al., 1996). The children were offered sweets either right now or twice the amount later. Being able to wait for the more generous reward was taken as an indicator for self-regulatory success. Children who earned more by waiting turned out to be better adapted a decade later in school and at home. Thus, deliberate delay of gratification has been shown to be a powerful strategy to cope with impulsive tendencies. However, self-control is not merely waiting passively for a delayed gratification. It is an active problem-solving strategy, as children have found diverse ways of distraction from the tempting sweets in front of them. Sacrificing immediate pleasures in the service of attaining long-term desirable outcomes is a basic paradigm of self-regulation. Mischel and coworkers have developed a 'hot-and-cool' framework to account for individual differences in self-control. The cool system is reflective in nature where people contemplate their perceptions, goals, and actions. This represents the executive function. On the other hand, the hot system is emotional and impulsive, allowing people to respond intuitively and rapidly to tempting or challenging environments. If a hot stimulus overrides the cool system, immediate impulsive action follows, and self-control fails. There are individual differences in the susceptibility to hot stimuli, and such differences are relatively stable. In other words, the 'hot-and-cool' framework serves not only an experimental paradigm for self-regulation but is also related to a dispositional view of self-control.

Such dual-process models have been put forward by a number of authors. For example, Strack and Deutsch (2004) have suggested a reflective–impulsive model proposing that behavior is a joint function of reflective and impulsive processes. The reflective system activates behavioral schemata based on cognition such as deliberate balancing of pros and cons, awareness of personal goals, and social evaluation. In contrast, the impulsive system is a faster, less conscious, or automatic one, that activates an effortless motivation to approach or avoidance when confronted with a demanding situation. A conflict between the reflective and the impulsive systems may occur, for example when there is a spontaneous (impulsive) approach toward sweets that becomes overruled by a separate goal-consistent controlled (reflective) approach toward a carrot. Temptations are governed by impulsive processes whereas inhibiting temptations is governed by reflective processes. The former operate quickly, effortless, and automatically whereas the latter operate slowly and deliberately, requiring sufficient cognitive resources. The impulsive system is more oriented toward the present needs whereas the reflective system is less situation-dependent and rather guided by long-term personal goals (see also the temporal self-regulation theory by Hall and Fong, 2007). These processes are resource-based which means that low resources lead to an increased likelihood for impulsive processes to prevail because such processes can operate outside of awareness (Hofmann et al., 2009). Being successful or not in fighting off the tempting impulses depends on one's attentional resources and conscious coping abilities. As long as the necessary resources are available, the reflective system can overrule the tempting impulses.

The strength model of self-control (Baumeister et al., 1994) regards self-control as a finite resource that can be exhausted if

expended or overused. According to this view, self-regulation requires strength that can be equated with what is commonly called willpower. This self-control energy allows people to resist temptations, persist at tasks, cope with adverse emotions, make choices, or overcome undesired impulses. However, when they do so, they deplete their reservoir of limited resources. When this happens, failure on a subsequent, unrelated task requiring self-control is more likely. This drop in willpower has been called 'ego depletion': Self-control is seen as analogous to a muscle that also requires energy to exert force over a period of time (Muraven and Baumeister, 2000). Effortfully inhibiting tempting impulses depletes cognitive and motivational resources. Once depleted, subsequent efforts to resist temptation are less successful because self-control is a global, limited resource that relates to different domains of human functioning. Depletion is not just a matter of fatigue but various beliefs, mood, and attitudes might affect one's ability to resist temptation and control one's desired path of action. Experimental research has attempted to identify factors that manage to prevent ego depletion including manipulating levels of fatigue, changing expectancies for task performance, forming action plans, or stimulating the organism by glucose supplementation. A meta-analysis of 83 studies examined the effects of ego depletion on various kinds of outcomes, investigated moderators of such effects as well as alternative explanations, providing support for the strength model of self-control (Hagger et al., 2010).

Self-regulation of attention and emotion might also be seen as a stable personal disposition, an individual's characteristic that enables habitual control over recurrent actions, as well as in the process of behavior change (Mischel et al., 1996). As the reflective system can transcend tempting situations only if control resources are available, then these resources may have a dispositional basis. In line with this assumption, psychometric tools were developed that tap the trait dimension of self-control. For example, the self-control scale (Tangney et al., 2004) assesses people's ability to override or change inner responses (e.g., 'I get carried away by my feelings'; reversed) and to interrupt undesired behavioral tendencies and refrain from acting on them (e.g., 'I am good at resisting temptations'). The self-regulation scale by Schwarzer et al. (1999) refers to post-intentional self-regulation when individuals are in the phase of goal pursuit, and face difficulties in maintaining their action. In such a maintenance situation it is required to focus attention on the task at hand and to keep a favorable emotional balance. Thus, attention regulation and emotion regulation are reflected in these scale items 'I stay focused on my goal and do not allow anything to distract me from my plan of action,' 'If an activity arouses my feelings too much, I can calm myself down so that I can continue with the activity soon' (Luszczynska et al., 2004).

A Process-Oriented Self-Regulation Model

In the following, a process model is briefly described that covers the goal setting as well as the goal striving phase of health self-regulation, called the health action process approach (HAPA; Schwarzer, 1992, 2008). Other social–cognitive models have been criticized mainly because of the so-called intention–behavior gap (referring to the frequent failure of intention to predict behavior). As a comprehensive self-regulation model, HAPA suggests a distinction between preintentional motivation processes that lead to a behavioral intention, and post-intentional volition processes that lead to actual behavior.

Motivation: The Goal-Setting Phase

Before changing their habits, people need to become motivated. This is seen as a process toward an explicit goal or intention (e.g., 'I intend to quit smoking this weekend'). Three constructs are considered to play a major role in this process: (1) risk perception, (2) outcome expectancies, and (3) self-beliefs.

Risk Perception

Perceiving a health threat seems to be the most obvious prerequisite for the motivation to replace a risk behavior. If one is not aware at all of the risky nature of one's actions, motivation would hardly develop. Usually, people are aware of some level of risk although the accuracy of their perception may be biased. When it comes to a comparison with similar others, one's view of the risk is somewhat distorted ('Compared to others of my age and sex my risk of getting lung cancer is low/medium/high'). This has become known as the 'optimistic bias.' Nevertheless, persons also acknowledge some degree of risk when confronted with objective data (health risk appraisal). There is a realistic component that keeps the positive illusions in leash. For example, smokers not only know that smoking can cause adverse health in others, they also perceive that they themselves are more at risk for lung cancer and other diseases than nonsmokers. Risk perception has two aspects: perceived severity of a health condition and personal vulnerability toward it. The first refers to the amount of harm that might occur, and the second pertains to the subjective probability that one could fall victim to that condition. Thus, it has been recommended that people should be informed about the existence of a health risk, and moreover, that they should imagine themselves as possible victims if they do not take the necessary precautions. Scaring people into health behaviors, however, has not been shown to be effective. In general, initial risk perception seems to be advantageous to put people on track for developing a motivation to change, but later on other variables are more influential in the self-regulation process, such as outcome expectancies, self-efficacy, and behavioral intentions. Risk perception in itself is insufficient to enable a person to form an intention. Rather, it may set the stage for a contemplation process and further elaboration of thoughts about consequences and competencies.

Outcome Expectancies

People not only need to be aware of the existence of a health threat, they also need to know how to regulate their behavior by understanding the contingencies between their actions and subsequent outcomes. These outcome expectancies are influential beliefs in the motivation to change (Bandura, 1997). A smoker may find more good reasons to quit than good reasons to continue smoking ('If I quit smoking then I will save money'). Similarly, positive outcome expectancies (e.g., 'If I exercise five times per week, I will reduce my cardiovascular

risk') are chiefly seen as being important in the motivation phase, when a person balances the pros and cons of certain behavioral outcomes. This imbalance does not lead directly to action but it can help to form an intention to quit. The pros and cons represent positive and negative outcome expectancies that are typical in rational decision-making. However, such contingencies between actions and outcomes need not be explicitly worded and evaluated; they can also be rather diffuse mental representations, loaded with emotions. Short-term outcome expectancies ('eating this chocolate now improves my mood') are in conflict with long-term ones ('not eating it helps me to stay slim'). Outcome expectancies are prospective whereas outcome experience is retrospective. The latter can generate the former when people develop habits, based upon good experience that further stimulates positive expectancies.

Self-Beliefs

Further, one needs to believe in one's capability to perform a desired action (perceived self-efficacy; e.g., 'I am capable of adhering to my exercise schedule in spite of the temptation to watch TV'). Perceived self-efficacy (as described above) operates in concert with positive outcome expectancies, both of which contribute to forming an intention. Both beliefs are needed for forming intentions to adopt difficult behaviors such as regular physical exercise or a dietary regimen. This is in line with the capacity-based approach to self-regulation. Whereas other indicators of capacity (e.g., willpower, dispositional self-control) are often understood as rather general constructs, self-efficacy pertains to a behavior-specific task, and it may fluctuate rapidly when facing imminent self-regulatory demands.

Volition: The Goal Striving Phase

Changing one's health behavior is considered to be a difficult self-regulation process. After people have become committed to a goal they need to prepare action and, later, maintain the changes in the face of obstacles and failures. Thus, goal setting and goal pursuit can be understood as two distinct processes that require self-regulatory effort. After forming an intention, the volitional phase is entered. When a person is inclined to adopt a particular health behavior, the 'good intention' has to be transformed into detailed instructions on how to perform the desired action. Once an action has been initiated, it has to be maintained. This involves self-regulatory beliefs, skills, and strategies such as planning, coping self-efficacy, and recovery self-efficacy. Additional volitional constructs, sometimes included in HAPA research, are action control (monitoring) and social support that may shield one's goal pursuit from distracting or tempting situations.

The volitional process can be subdivided into sequences such as planning, initiation, maintenance, relapse management, and disengagement. The adoption and maintenance of the health behavior is not achieved through an act of will but involves the development of self-regulatory skills and strategies. This embraces various means to influence one's own motivation and behaviors such as the setting of attainable proximal subgoals, creating incentives, drawing from an array of coping options, and mobilizing social support. Plans specify the when, where, and how of a desired action, and persons take initiative when the critical situation arises and perform preparatory behaviors. Action control processes include focusing one's attention on the task, while avoiding attention to distractors, resisting temptations, and managing unpleasant emotions. Perceived self-efficacy is required to overcome barriers and to stimulate self-motivation repeatedly. Once an action has been taken, individuals with high maintenance self-efficacy invest more effort and persist longer than those who are less self-efficacious.

Competent relapse management is needed to recover from setbacks. Some people rapidly abandon their newly adopted behavior when they fail to get quick results. When entering high-risk situations (e.g., a place where others smoke), they cannot resist due to a lack of self-efficacy. The competence to recover is different from the competence to initiate an action. Restoration, harm reduction, and renewal of motivation are serviceable strategies within the context of health self-regulation.

Disengagement from the goal can be evidence of lack of persistence and, thus, indication of self-regulatory failure (Carver and Scheier, 1998). But in the case of repeated failures, disengagement or scaling back the goal might become an option which can be adaptive, depending on the circumstances. For example, if the goal was set too high or if the situation has changed and become more difficult than before, it is often not worthwhile to continue the struggle. In the case of health-compromising behaviors, giving up is not a tenable option. Better self-regulatory skills have to be developed and unique approaches to the problem need to be taken. The experience of failure can be a useful learning experience to build up more competence, on the condition that the individual makes a beneficial causal attribution of the episode, and practices constructive self-talk to renew the motivation.

A large body of literature has emerged that supports the HAPA as a useful health self-regulation model for various health behaviors, as documented by the most recent research on sunscreen use (Craciun et al., 2012), vaccination (Payaprom et al., 2011), condom use (Carvalho et al., 2014; Teng and Mak, 2011), smoking cessation (Radtke et al., 2011), alcohol consumption (Murgraff et al., 2003), physical activity (Gellert et al., 2014; Parschau et al., 2014), oral self-care (Schwarzer et al., 2014), and healthy eating (Godinho et al., 2014; **Figure 1**).

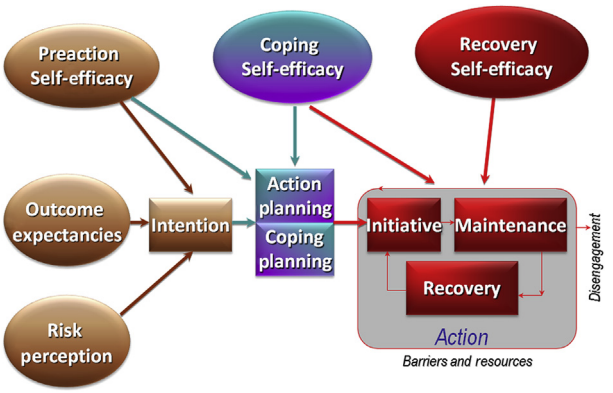

Figure 1 Diagram of the Health Action Process Approach (Schwarzer, 2008).

The HAPA is designed as an open architecture framework that is based on the following principles rather than on a set of specific hypotheses. Developed in 1988 (Schwarzer, 1992) as a health self-regulation model, it was an attempt to integrate the theory of action phases (Heckhausen and Gollwitzer, 1987) with social–cognitive theory (Bandura, 1997), based on five principles:

Principle 1: Motivation and volition. The health behavior change process is divided into two phases. There is a switch of mindsets when people move from deliberation to action. First comes the motivation phase in which people develop their intentions. Afterward, they enter the volition phase.

Principle 2: Two volitional phases. In the volition phase, there are two groups of people: those who did not yet translate their intentions into action, and those who did. Thus, there are inactive as well as active persons in this phase. In other words, in the volitional phase one finds intenders as well as actors who are characterized by different psychological states. In addition to health behavior change as a continuous process, one can also create three categories of people with different mindsets, depending on their current location within the course of behavior change: preintenders, intenders, and actors.

Principle 3: Post-intentional planning. Intenders who are in the volitional preactional stage are motivated to change, but they do not act because they might lack the right skills to translate their intention into action. Planning is a key strategy at this point. It serves as an operative mediator between intentions and behavior.

Principle 4: Two kinds of mental simulation. Planning can be divided into action planning and coping planning. Action planning pertains to the when, where, and how of intended action. Coping planning includes the anticipation of barriers and the design of alternative actions that help to attain one's goal in spite of impediments.

Principle 5: Phase-specific self-efficacy. Perceived self-efficacy is required throughout the entire process. However, the nature of self-efficacy differs from phase to phase. This is because there are different challenges as people progress from one phase to the next one. Goal setting, planning, initiative, action, and maintenance all pose challenges that are not of the same nature. Therefore, (pre)action self-efficacy, maintenance (coping) self-efficacy, and recovery self-efficacy are being distinguished.

An overview of motivational and volitional processes in health behavior change is presented in the video that supplements this article.

Appendix A. Supplementary Data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/B978-0-08-097086-8.26031-5.

See also: Motivation and Actions, Psychology of; Self-Regulated Learning: Theories, Measures, and Outcomes.

Bibliography

Bandura, Albert, 1997. Self-efficacy: The Exercise of Control. Freeman, New York.

Baumeister, Roy F., Heatherton, T.F., Tice, D.M., 1994. Losing Control: How and Why People Fail at Self-regulation. Academic Press, San Diego, CA.

Carvalho, Telma, Alvarez, Maria J., Barz, Milena, Schwarzer, Ralf, 2014. Preparatory behavior for condom use among heterosexual young men: a longitudinal mediation model. Health Education and Behavior. http://dx.doi.org/10.1177/1090198114537066. Published online: 02 July 2014.

Carver, Charles S., Scheier, Michael F., 1998. On the Self-regulation of Behavior. Cambridge University Press, New York.

Craciun, Catrinel, Schüz, Natalie, Lippke, Sonia, Schwarzer, Ralf, 2012. A mediator model of sunscreen use: a longitudinal analysis of social-cognitive predictors and mediators. International Journal of Behavioral Medicine 19, 65–72. http://dx.doi.org/10.1007/s12529-011-9153-x.

Godinho, Cristina, Alvarez, Maria J., Lima, Maria L., Schwarzer, Ralf, 2014. Will is not enough: coping planning and action control as mediators in the prediction of fruit and vegetable intake. British Journal of Health Psychology 19 (4), 856–870. http://dx.doi.org/10.1111/bjhp.12084.

Gellert, Paul, Simon, Krupka, Ziegelmann, Jochen P., Knoll, Nina, Schwarzer, Ralf, 2014. An age-tailored intervention sustains physical activity changes in older adults: a randomized controlled trial. International Journal of Behavioral Medicine 21 (3), 519–528. http://dx.doi.org/10.1007/s12529-013-9330-1. Online 17 July 2013.

Hall, P.A., Fong, G.T., 2007. Temporal self-regulation theory: a model for individual health behavior. Health Psychology Review 1, 6–52.

Heckhausen, Heinz, Gollwitzer, Peter M., 1987. Thought contents and cognitive functioning in motivational vs volitional states of mind. Motivation and Emotion 11, 101–120.

Hagger, Martin S., Wood, C., Stiff, C., Chatzisarantis, N.L.D., 2010. Ego depletion and the strength model of self-control: a meta-analysis. Psychological Bulletin 136, 495–525.

Hofmann, Wilhelm, Friese, Malte, Strack, Fritz, 2009. Impulse and self-control from a dual-systems perspective. Psychological Science 4, 162–176.

Kuhl, Julius, 2001. Motivation und Persönlichkeit. Interaktionen psychischer Systeme [Motivation and Personality: Interactions of Mental Systems]. Hogrefe, Göttingen, Germany.

Luszczynska, Aleksandra, Schwarzer, Ralf, 2003. Planning and self-efficacy in the adoption and maintenance of breast self-examination: a longitudinal study on self-regulatory cognitions. Psychology & Health 18, 93–108.

Luszczynska, Aleksandra, Diehl, Manfred, Gutiérrez-Doña, Benicio, Kuusinen, P., Schwarzer, Ralf, 2004. Measuring one component of dispositional self-regulation: attention control in goal pursuit. Personality and Individual Differences 37, 555–566.

Mann, Traci, De Ridder, Denise, Fujita, Kentaro, 2012. Self-regulation of health behavior: social psychological approaches to goal setting and goal striving. Health Psychology 5, 487–498.

Mischel, Walter, Cantor, Nancy, Feldman, S., 1996. Principles of self-regulation: the nature of willpower and self-control. In: Higgins, E.T., Kruglanski, A. (Eds.), Social Psychology: Handbook of Basic Principles. Guilford, New York, pp. 329–360.

Muraven, M., Baumeister, Roy F., 2000. Self-regulation and depletion of limited resources: does self-control resemble a muscle? Psychological Bulletin 126, 247–259.

Murgraff, V., McDermott, M.R., Walsh, J., 2003. Self-efficacy and behavioral enactment: the application of Schwarzer's Health Action Process Approach to the prediction of low-risk, single-occasion drinking. Journal of Applied Social Psychology 33 (2), 339–361.

Ochsner, Sybille, Scholz, Urte, Hornung, Rainer, 2013. Testing phase-specific self-efficacy beliefs in the context of dietary behaviour change. Applied Psychology: Health and Well-Being 5, 99–117. http://dx.doi.org/10.1111/j.1758-0854.2012.01079.x.

Parschau, Linda, Barz, Milena, Richert, Jana, Knoll, Nina, Lippke, Sonia, Schwarzer, Ralf, 2014. Physical activity among adults with obesity: testing the health action process approach. Rehabilitation Psychology 59, 42–49. http://dx.doi.org/10.1037/a0035290.

Payaprom, Y., Bennett, P., Alabaster, E., Tantipong, H., 2011. Using the Health Action Process Approach and Implementation Intention to increase flu vaccination uptake in high risk Thai individuals: a controlled before-after trial. Health Psychology 30, 1–10. http://dx.doi.org/10.1037/a0023580.

Radtke, T., Scholz, Urte, Keller, R., Hornung, R., 2011. Smoking is ok as long as I eat healthily: compensatory health beliefs and their role for intentions and smoking within the Health Action Process Approach. Psychology and Health 27 (2), 91–107.

Schwarzer, Ralf, 1992. Self-efficacy in the adoption and maintenance of health behaviors: theoretical approaches and a new model. In: Schwarzer, Ralf (Ed.), Self-efficacy: Thought Control of Action. Hemisphere, Washington, DC, pp. 217–243.

Schwarzer, Ralf, 2008. Modeling health behavior change: how to predict and modify the adoption and maintenance of health behaviors. Applied Psychology: An International Review 57, 1–29.

Schwarzer, R., Antoniuk, A., Gholami, M., 2014. A brief intervention changing oral self-care, self-efficacy, and self-monitoring. British Journal of Health Psychology. http://dx.doi.org/10.1111/bjhp.12091.

Schwarzer, Ralf, Diehl, Manfred, Schmitz, Gerdamarie, 1999. Self-Regulation Scale. http://userpage.fu-berlin.de/∼health/selfreg_e.htm.

Strack, Fritz, Deutsch, R., 2004. Reflective and impulsive determinants of social behavior. Personality and Social Psychology Review 8, 220–247.

Tangney, J., Baumeister, Roy F., Boone, A.L., 2004. High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. Journal of Personality 72, 271–324.

Teng, Y., Mak, W.W.S., 2011. The role of planning and self-efficacy in condom use among men who have sex with men: an application of the Health Action Process Approach model. Health Psychology 30, 119–128. http://dx.doi.org/10.1037/a0022023.

Relevant Websites

http://www.psy.miami.edu/faculty/ccarver/TOCofSRbook.html – Classic book by Charles Carver and Michael Scheier on Self-Regulation.

http://www.selfregulationlab.nl/index.php/home – The Dutch Self-regulation Lab.

www.hapa-model.de – The Health Action Process Approach Resource Pages.

http://userpage.fu-berlin.de/∼health/selfreg_e.htm – The Self-Regulation Scale by Schwarzer, Diehl, and Schmitz, 1999.

Health Social Work

Cecilia Lai Wan Chan and Bo Yan, The University of Hong Kong, Hong Kong

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Health social work has been historically associated with environmental, sociopolitical, and economic context of the society. With the industrialization, urbanization, and global warming, health social workers are combating poverty, environmental degradation, natural disasters, and health inequity. Health social work in this article is defined as a subcategory of social work profession that includes medical social work and public health social work. Health social workers serve to humanize the health and social care system, emphasizing the importance of the holistic psycho-social-spiritual needs of individuals and groups when they experience distresses of illness, trauma, death, or disabilities across the life span. Health social workers practice in hospitals, communities, health departments, nongovernmental organizations, and international bodies. Advocacy, self-help, psychosocial education, and community organization are commonly adopted strategies of patient empowerment for health social workers. International organizations are active in promoting reform in the global environment so as to make the world a better place.

Overview of Health Social Work

Introduction

The development of health social work was closely linked to global development of social work in the Europe and North America through history. With the industrialization, urbanization, and global warming, health social workers are combating issues such as poverty, environmental degradation, natural disasters, and health inequity. They are increasingly specialized in providing support and services to specific disease groups and treatment interventions such as cancer and organ transplant in hospitals and clinics. Table 1 summarizes the three major categories in the field of health-related social work.

Definition

Health social work is defined as a specialization of social work that includes medical social work, rehabilitation social work, and public health social work. As rehabilitation social work can also be grouped under medical social work, this article will mainly discuss health social work under the context of medical social work and public health social work. The former primarily uses professional activities to help individuals, families, groups, or communities related to certain medical conditions to "restore, retain, and enhance" (Council on Social Work Education, 2009: 7) their well-being, while the later adopts "an epidemiological approach" (Practice Standards Committee, 2005: 4) to address social problems affecting health and social functioning.

Development and Transformation of Health Social Work

The Nineteenth and Early Twentieth Century

Health social work was initially developed in the nineteenth century in the United Kingdom (UK) and the United States (US) in response to the social needs. The first professional health care social workers were employed as 'hospital almoners' in UK. They took care of patients with disease-related, social, and family issues under the London Charity Organization Society (COS). In the US, the first social worker was hired in Massachusetts General Hospital in 1905 (Gehlert and Browne, 2006: 6–7). As the profession developed and matured, hospitals have remained a key source of employment for social workers (Table 2).

One of the major public concerns during the nineteenth century was urban poverty and related issues such as infectious

Table 1 Typology of heath-related social work

Medical social work	*Rehabilitation social work*	*Public health social work*
Social workers by settings (clinic, intensive care unit, emergency room, hospice, homes) Social worker by disease types (oncology, nephrology, cardiology, pediatrics, gynecology, geriatrics) Social workers by tasks (organ transplant coordinators, infection control, infertility and reproductive technology, bone-marrow transplant)	Physical rehabilitation (stroke, amputation, organ failure, pain, cognitive deficits) Emotional rehabilitation (grief and bereavement, depression, insomnia, mood disorders) Behavioral rehabilitation (addictive behavior of drugs/alcohol/gambling/sex/shopping) Social rehabilitation (social inclusion and antidiscrimination for persons with HIV/AIDS, mental illness, offenders)	Disease prevention (smoke cessation, healthy lifestyle, education, and promotion of regular exercise and balanced diet in prevention of obesity in the industrialized world and malnutrition in the industrializing world) Global health (policy analysis and advocacy against injustice, trafficking and forced labor, environmental degradation, disaster and pollution, promotion of environmental protection, fair trade, gender equality, social economy)

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28039-2

Table 2 Development and transformation of Health Social Work

	Health social work	*World events and public health*
1910 and before	• The first hospital social worker, called 'a hospital almoner,' was hired by the Royal Free Hospital in London in 1895 • Trainings for almoners started by British Hospital Almoner's Council; medical social workers were introduced by Dr Richard Cabot at Massachusetts General Hospital in 1905; community-based hospice care with social work involved was marked by the opening of St. Joseph's Hospice in UK in 1905	• Rely on local traditional charities and lineage mutual help organizations • The London Charity Organizations of Society in 1870s; settlement house movement as community charity work in UK (1884), the US (1906), and Canada (1902); fire and infectious diseases, poverty, and child labor become global concerns; public health social work for the control and prevention of outbreaks of infectious diseases such as syphilis and tuberculosis
1910–1950	• First hospital social work conference in the US; first psychiatric SW training course in the US • The first community health clinic in Canada was founded in 1920, which integrated social work and public health • First psychiatric social work training in UK; Association of Psychiatric Social Workers (UK) and the British Federation of Social Workers; definition of medical social work in the US • CASW advocacy for national health insurance; clinical treatment orientation in the US; National Committee of Schools of Social Work established; almoners in hospitals in Hong Kong, Singapore, and other British colonies; Institute of Almoners in Britain was established (1945)	• First SW training program in Canada in University of Toronto (1914); veteran resettlement in Canada after World War I; social workers played a role in disaster relief in the Halifax Explosion (1917) • Canadian Association of Social Workers (CASW) founded in 1927; first national conference in 1928; CASW advocacy for national health insurance; Clinical International Conference on Social Work (1928) • The Great Depression; program in support of unemployed and persons in poverty • End of World War II; the formation of United Nations; World Health Organization (1948); the end of the poor law (UK) with new development of individualized systems of care; establishment of the Hong Kong Council of Social Services and national associations
1950–1999	• Development of private practice clinical social workers in the US; community mental health and public health movement in the US; the first US half-way house for mental health clients was established in Brookline, Massachusetts (1954) • Social work trainings in UK particularly for working with elderly, disabled, and youth offenders; the Society for Hospital Social Work Directors (US); Institute of almoners renamed as Institute of Medical Social Work in 1964 • NASW Journal *Health and Social Work*; hospice social work in the US; hospital social work in Germany • First HIV case report in the US • Trend of cut-back and privatization for services in Canada UK and the US; call for enhancing community social work for de-institutionalization	• The National Association of Social Workers (NASW, US); Singapore social work as almoners in local hospitals to handle social disruption caused by WW-II • Radical social work in Canada; international migration and rapid industrialization • 'Death in America' project that started the movement of palliative care social work • Institute of Medical Social Work re-constituted to form the British Association of Social Workers (BASW); increased interest in community work in UK; shift of focus from 'poor relief' to 'social care' in Nordic countries; community work and advocacy based on radical social work in Britain • Antidiscriminatory practice, empowerment theories; community health centers were promoted by the Canadian government in 1970s
2000 and beyond	• Higher level of specialization in hospital social workers based on advances in medical sciences • Development of social work and public health degree programs; trend of prevention and elimination of the root problems • Development of holistic bio-psycho-social-spiritual care through integrative interventions to promote holistic health and wellness • Advances in medical sciences and genetic engineering brought about ethical issues of human reproduction, life, and death	• Major international natural and human disasters leading to the development of environmental and disaster social work • Participation of IFSW, ICSW in the UN consensus of MDG (2015) on antipoverty and social development goals • Founding of the Society for Palliative Social Work by Grace Christ and her associates • Health social work moving into the international arena of global injustice and fair trade

Sources: Hick, S.F., 2002. Social Work in Canada: An Introduction. Thompson Educational Pub, Toronto; Jennissen, T., Lundy, C., 2011. One Hundred Years of Social Work: A History of the Profession in English Canada, 1900–2000. Wilfrid Laurier University Press, Waterloo, ON; Mizrahi, T., Davis, L.E. (Eds.), 2008. Encyclopedia of Social Work (2th ed.). NASW Press; Oxford University Press. Washington, DC; Oxford, New York. Payne, M., 2005. The Origins of Social Work: Continuity and Change. Palgrave Macmillan, Houndmills, England; New York.

diseases. While the health almoners under COS focused on helping the poor, a group of settlement house workers were more active in promoting social reform, initiating the Settlement Movement (Hick, 2002: 41–42). The Settlement Movement first begun in London to assist young well-educated people to serve in poor neighborhoods and to provide moral models for social improvement (Payne, 2005: 37–39). It had quickly spread to countries such as Canada and the US and brought a broad range of development in vital aspects including education, political action, and youth work (Payne,

2005: 46–47). Some social workers became involved in initiatives to influence policy and legislation as they found the coverage of the Poor Laws in Britain to be insufficient in combating poverty during the Great Depression. At approximately the same period of time, there is also a need for the US to address issues such as poverty and unemployment, where community development programs naturally grew.

Occupational Social Work

With rapid industrialization and economic development, new workforce including women, minorities, and people with disabilities emerged in the nineteenth century. Occupational social work, also referred as industrial social work, started to develop in the 1970s, catering to the needs and protecting the rights of the working populations (Kurzman, 2008: 314). Programs were introduced to ensure needs of daycare, occupational safety, equal opportunity, and quality of life of workers are protected. Early occupational social work programs targeted at alcoholism, with funding largely coming in from the National Institute on Alcohol Abuse and Alcoholism. Nowadays, as the field gets more mature, occupational social workers respond to more needs, often in union counseling and advocacy programs in the US (Kurzman, 2008: 312). Their duties cover individual counseling with employees in need, training relevant staff for referrals to social services, initiating programs to promote health, and advising on affirmative action plan, especially when vulnerable populations are involved.

Global Scene

With the passage of time, health social work has gradually developed across the globe, in European as well as Asian countries (see **Table 2** for a summary of health social work development and world events). Professional social work organizations and training institutions began to emerge across countries such as UK, Europe, and North America. The colonization movements of these countries, together with the establishment of international philanthropic nongovernmental organizations (NGOs), brought health social work to other countries including Australia, India, Singapore, China, Taiwan, Africa, South America, as well as their other colonies in different parts of the world. The rapid development of health social work in countries such as Bulgaria, China, and Estonia slowed down, when the communist parties came to power in 1940s. The tasks of social workers to provide welfare and address social problems were taken over by socialist states (Payne, 2005: 72–75) and social development was prioritized over social work. Health social work did not regain its role and significance until the breakup of the USSR and Eastern Europe as well as the opening up of China and other communist regimes in the late 1980s to harmony, and integration of individuals and families in societies (Payne, 2005: 74).

Health Social Work and Environmental Challenges

At the dawn of the twenty-first century, social work profession increasingly realized the importance of environmental protection when large-scale natural disasters such as flood, forest fire, earthquakes, and volcano eruptions were seen everywhere across the globe as a result of climate changes and environmental degradation (Gray et al., 2013: 4). As early as in the 1970s, social workers started to take initiatives in environmental issues and showed interests that range from ecologic perspective to environmental justice movement (Gray et al., 2013: 11). More recently, social workers advocate for best practice and policies to reduce carbon footprint, preserving sources of clean water and forests, and raise global awareness about environment protection. The three leading social work organizations – International Association of Schools of Social Work (IASSW), the International Federation of Social Workers (IFSW), and International Council on Social Welfare (ICSW) – explicitly announced their support toward the extension of the UN Kyoto Protocol to protect the environment and to reduce the effect of climate change. They have also presented in UN conferences, such as the Climate Change Conference in Copenhagen in 2009 (Dominelli, 2012: 85–86).

Health Disparity and the Roles of Health Social Work

At the turn of the twenty-first century, health disparities are becoming an international problem for industrialized and industrializing societies. The World Bank recognized poverty and social disparity as major threats to social development in the contemporary era (World Bank, 2012). Low-income populations are less accessible to health care and social care. Industrialized countries such as the US, Canada, UK, and Finland invest more on health compared with industrializing countries such as Niger, Nepal, Kenya, and China (Karki et al., 2013: 214–215). The situation is clearly indicated by health indicators including tuberculosis (TB) prevalence, human immunodeficiency virus (HIV), and relation of life expectancy to health expenditure.

Poverty and the insufficient investment in health and social care have lead to a variety of health problems in industrializing countries. Over 2 billion people live on less than US$2 per day and they cannot even afford for the basic medical care. Malnutrition, infant deaths, neonatal deaths, and stillbirths are taking more than one-third of the lives of global newborns and putting 20 million pregnancies at risks (IFRC, 2011: 13). Battles against communicable diseases such as TB, hepatitis, malaria, simple bacterial infections, and diarrhea are on-going. HIV/AIDS, as a global epidemic, has brought tremendous challenges for children, youth, adults, and their families. The estimated population who are living with HIV worldwide is 40 million according to UNAIDS (Newman, 2008: 357).

Facing the challenging situations in health care, health social workers can take active roles in leading and coordinating the interdisciplinary healthcare team. For example, health social workers can assist in reducing infant deaths and pregnancy risk through mobilizing community health workers, midwives, and doctors to provide home visits and cost-effective interventions such as immunization, micronutrients, treatment for diarrhea, malaria, and acute respiratory infections, as well as improved prenatal and delivery care in the rural areas (IFRC, 2011: 13). They can help to fight against communicable diseases with the promotion of environmental cleanliness, health, poverty reduction, and gender equality. Moreover, they may also contribute to HIV testing, early intervention,

treatment adherence, case management, and care coordination, as well as research and community engagement (Newman, 2008: 360–361).

Health social work has been historically associated with environmental, sociopolitical, and economic context of the society. While psychosocial and environmental aspects of deprivation and trauma are often closely related to illness and death, physical illnesses can also elicit emotional distress, and socioeconomic relationship breakdowns such as divorce and bankruptcy may result in somatic problems. Health social workers play vital roles in providing direct services, as well as promoting advocacy in research and policy practices.

Major Areas of Practice

As shown in Table 1, medical social workers, rehabilitation social workers, and public health social workers are the three main categories of social workers working in the healthcare system in serving individuals, families, communities, and societies. The following discussion will be focused on the two of the three categories: medical social work and public health social work. Rehabilitation social work will be excluded from the following discussions as there is a section on social work and disability; besides it can also be regarded as one of the subcategories under medical social work.

Medical Social Work

Health social workers practice in hospitals, communities, health departments, NGOs, and international bodies. Health social workers serve to humanize the health and social care system, emphasizing the importance of the holistic psycho-social-spiritual needs of individuals and groups when they experience distresses of illness, trauma, death, or disabilities across the life span. Those serving in hospitals and clinics are known as medical social workers. In face of diseases and illnesses, patients often experience distress including anxieties about the tests before the diagnosis, worries about loss of income or functioning due to the disabling effects of illness and/or treatment, grief over the loss of health and sense of control over one's life, frustration over social discrimination due to physical disfigurement or intellectual limitations, and disorientation with the world and feeling overwhelmed by the sense of vulnerability.

Medical social workers work flexibly as case managers, service coordinators, patient advocates, health educators, counselors, family therapists, and pastoral and spiritual care providers in helping patients throughout their course of illness. They serve patients through assessment and enhancement of their bio-psycho-social functioning, discharge planning, case management, or group intervention to facilitate their adjustment to illness and loss, strengthening their support networks, connecting them with necessary community resources for rehabilitation and long-term adjustment to chronic illness or disabilities, and building of self-help and survivors groups. They also mobilize volunteers to serve as patients' advocates, peer counselors, and expert patients.

In addition to the multiple roles, medical social workers in hospitals often work closely with the medical teams and are also becoming increasingly specialized. Some of them specialized according to disease groups and target patient groups such as oncology, cardiology, gynecology, geriatrics, and pediatrics and some specialized by settings such as social workers in clinics, intensive care unit, hospice and palliative unit. Others specialized by tasks including prenatal diagnosis, genetic counseling, case management, organ transplant, and infection control.

There is increasing research evidence supporting the importance of health social work in the healthcare system. For instance, Egan (2010) summarized research evidence of effective social work interventions for five diseases and found that health social work brings multiple positive effects including improving patients' knowledge, adherence, and competence in disease management, while enhancing their quality of life and self-esteem. More recent evidence sheds light on the integrative approach toward health and social work practices, suggesting the use of Eastern health models of body–mind–spirit (IBMS) intervention such as Taichi and qigong exercises, massage, meditation, mindful practices, yoga, meaning-making, resilience training, and spirituality of suffering and growth through pain to enhance holistic wellness (Lee et al., 2009: 217). Evidence has also been accumulating for positive- and strength-oriented intervention programs on forgiveness, gratitude, generosity, compassion, social support, and transformation (John Templeton Foundation, 2010).

Medical social workers participate not only in providing direct services to individuals and families, they also advocate for social and legislature change. An example would be the social workers in oncology and palliative care. Apart from playing a vital role in meeting the needs of clients with life-limiting illnesses and their family members, they also engage in education, navigating through and advocating the medical and social systems (Blacker and Christ, 2011: 21–23).

Public Health Social Work

Public health has emerged to promote health by preventing, limiting, and stopping the spread of disease (Keefe and Evans, 2013). As one of the members of the multidisciplinary settings, public health social workers address health from a broader perspective (Practice Standards Committee, 2005: 4), focusing on promoting healthy lifestyle to individuals, families, and groups; preventing illness; and reducing unhealthy behaviors such as smoking, the use of drugs and alcohol, physical or sexual violence, promiscuity, gambling, drunk driving, self-harm, and suicide. In countries such as Canada, social workers are hired as 'health promoters' (Raphael, 2000: 356–357) to identify key factors for health, promote activities to improve individuals' health conditions, reduce high-risk behaviors, and evaluate program effectiveness.

Public health social workers adopt a community development approach to mobilize resources, cultivate collective problem-solving, as well as improve health and resilience among underprivileged groups. Raising awareness for disease prevention, health promotion, improvement in nutrition, and access to immunization can enhance social development of the society. The short-term voluntary program for at-risk mothers and newborns offered by a Philadelphia

interdepartmental collaborative in the US would be an excellent example. The program was established in response to the amendment for the Federal Child Abuse Prevention and Treatment Act, which mandates a system for the mothers who are identified to have used illegal drug during pregnancy. The Philadelphia Department of Human Services and Philadelphia Department of Public Health, Division of Maternal, Child and Family Health expected law changes at the state level, and they developed this home-based program to serve the mothers who have used illegal drugs during pregnancy to ensure their health and a healthy environment for the development of their newborn babies. Safety of children is the top priority in service delivery. With a clear protocol, social workers, behavioral health specialists, and nurses work with the mothers closely for assessment, brief intervention plans, and referrals as needed. Encouraging outcomes have been reported in terms of mothers' action plan and capability of parenting (Claxton et al., 2010: 291–303).

With the increasing concerns over environmental issues, public health social work also extends to green social work or environmental social work, to promote a healthy or harmonious relationship between people and environment. Social workers in this field seek solutions to protect the environment and to achieve sustainability (Dominelli, 2012: 15–16). Interdisciplinary interventions at multiple levels (Macro, Mezzo, and Micro) curbing the sources of environmental challenges, including but not limited to industrial pollution, overuse of natural resources, global warming, waste disposal, and disasters are essential. Gray et al. (2013) even proposed the needs to better understand the interactions between human society and the natural environment and to develop case studies on how models of effective practice can be established in this new area of health social work (Gray et al., 2013: 13–16).

Public Health Social Worker and Disaster

In addition to the 'preventive' environmental protection work, public health social workers also take part in managing natural disasters such as drought, frost, tsunami, landslide, extreme heat waves, and cold spells as well as 'man-made' disasters such as wars, terrorist attacks, as well as nuclear accidents and oil spills. Corruption and the socioeconomic–political oppression of women, religious, or ethnic minorities are creating large refugee populations as they escape from homes ruined by political instability and brutality. These disasters, often accompanied by physical, sexual, and spiritual violence can put large populations under post-traumatic disorders that may bear life-long consequences. Public health social workers are thus playing an important role in the humanitarian actions along with others in the disaster management team (Mathbor and Bourassa, 2012: 294). Social workers serving in disastrous areas are taking on rescue, support, and rehabilitation tasks to minimize the impact of disasters as well as political and societal conflicts on individuals and society. Witnessing forms of human suffering and emotional trauma, they also involved in the development of positive- and strength-oriented approaches such as reconciliation, restorative justice, and forgiveness and hope interventions to assist disaster-affected populations in moving on with life. One example would be the medical social workers in US Veteran Administration hospitals. While providing direct services in meeting the increasing needs of war veterans, they were also actively involved in the interdisciplinary teams in the areas of research and public health (Keefe and Evans, 2013: 11).

Public health social workers play pivotal roles in facilitating communication of rescue information and help victims to identify family members and in managing volunteers and rescue workers in collaboration with the disaster victims on temporary shelters while establishing daily routines and mutual help teams in transient resettlement sites. For example, during the mega Sichuan Earthquake in 2008, Chinese social workers, along with other professionals and volunteers, helped in organizing social and recreational activities for victims displaced from their home and work, mobilizing care for vulnerable populations such as disoriented older adults and persons with disabilities, re-establishing kindergarten and schools for children as well as play groups, facilitating economic activities of shops and cook food stalls plus other necessary services in temporary shelters so as to maintain a sense of normalcy. Self-management teams were also established for hygiene and sanitation. The involvement of social workers does not only provide immediate assistance in disasters management, but also empowered the community with the capacity and hope toward rebuilding their homes in the long run.

Ineffective management of natural or man-made disasters can lead to further disasters, such as an outburst of infectious diseases due to poor management of dead bodies. When large populations are moved into temporary housing or refugee camps, their access to a safe environment, clean water and food, sanitary facilities of waste, and refuse management become crucial. For example, the most common causes for death in Africa are diarrheal diseases, measles, acute respiratory infections, and malaria. Effective prevention includes protection from violence, disease control, immunization, case management, and maternal and child health care (Toole and Waldman, 1997: 283).

Disaster services have grown from treating Post-Traumatic Stress Disorder (PTSD) to "psychosocial capacity building" (Miller, 2012: 12). Public health social workers engage local partners so that culturally relevant and effective interventions can be developed with input from people and resources in disaster communities (Miller, 2012: 50). Evidence-based practice is highly recommended for application of previous lessons learned from emergent contexts, and evaluation processes are crucial to generate evidence for future reference. Handbooks in this field can be found to be referenced in Mathbor and Bourassa's article and they are particularly helpful in addressing the issues around collaboration in psychosocial responses in emergency settings.

Core Competencies

Despite the differences in duties and focuses, there are shared core competencies involved in medical social work (Council on Social Work Education, 2009) and public health social work (Practice Standards Committee, 2005). As in **Figure 1**, it is clear that medical social work and public health social work

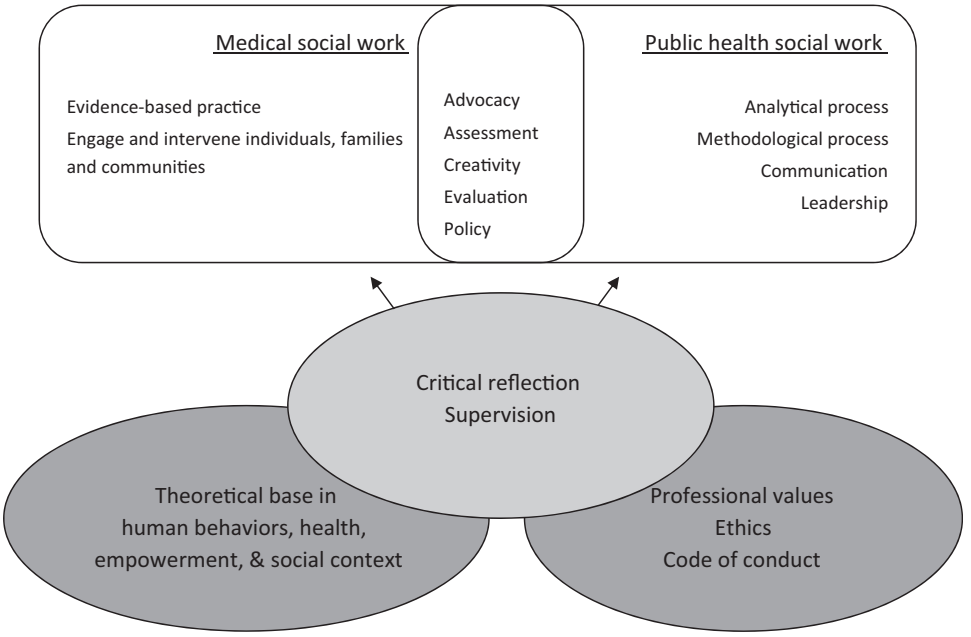

Figure 1 Core competencies of health social work. Sources: Clinical Social Work Competencies (Council on Social Work Education, 2009) and Practice Standards Committee, 2005. Public Health Social Work Standards and Competencies. http://oce.sph.unc.edu/cetac/phswcompetencies_may05.pdf.

are rooted in the same theoretical base and grounded by the same set of social work professional ethics, values, and code of conduct. For instance, a public health social worker may assist a community development project responding to a collective health issue affecting all the members, especially the vulnerable (Practice Standards Committee, 2005: 12), and a medical social worker may conduct psycho-education with a patient's family to reduce possible discrimination against the patient's diseases such as HIV/AIDS. The underlying values for the different services are the same: promoting social and economic justice. Health social workers have to interpret and apply those theoretical foundations and ethics flexibly across different practical contexts, and, therefore, critical thinking and reflection is crucial during their practice. Professional supervision also serves as an on-going support for social workers, helping them to better understand, review, and reflect on their practice.

The two ovals on the top are core competencies, respectively, summarized for medical social work and public health social work. Their domain-specific professional knowledge and skills are listed. It is clear that their emphases are different (see **Figure 1**). Meanwhile, core components such as assessment, evaluations, creativity, research, advocacy, and policy practice are shared. Health social workers often advocate for access of services, clients' rights, and justice at multiple levels for the purpose of reducing health disparity (Council on Social Work Education, 2009: 4–7). In policy practice (Practice Standards Committee, 2005: 21–22; Council on Social Work Education, 2009: 8), health social workers actively:

- engage in policy practice regarding agency, regional, or national policy, law, and regulations;
- identify gaps, collect data relevant to particular policy, build coalitions, and plan for action; and
- inform legislators and administrators, and collaborate with others for effective policy actions at the organizational and societal level.

Challenges and Opportunities

Challenges in Health Social Work Practice

There has been increasing pressure on the justification of professional social worker positions. Naturally, a shift of focus flows from macrolevel social work such as community development to more tangible areas in direct services. Therefore, the productivity requirement for frontline social work duties has been raised drastically. Community-based social workers may have to prioritize their time allocated to advocacy as their major duty is specifically tied to the hourly 'fee-for-service,' for instance, assessment on whether an adult is eligible for certain benefits. While the waiting list is getting longer, other daily functions of these social workers can be overwhelming. It is difficult, if not impossible, to devote time for research, continuing education, and service improvement, which are vital for creating 'knowledge-in-practice' (D'Cruz, 2009: 69–71).

Another challenge is the demand for cost-efficiency and accountability. The call for 'pragmatic evaluation' for practice, promoted by the managerialist approach, lead to the use of simplistic performance indicators to guide practice (Payne, 2005: 219, 241). Priorities are given to productivity, data input, documentation, and funding applications, which may 'squeeze' the quality of direct client contact to the side. For

example, one of the major complaints from mental health social workers working with public medical reimbursement system in the US is their struggle on a weekly basis to meet the demanding paperwork submission.

There are debates and concerns on the use of existing scales and assessment protocols to measure risk and needs, make diagnosis, and determine priority and access to services. For example, despite the fact that DSM-IV-TR and ICD-10 are widely used in the US, some social workers find these mental health assessments disconnected with the clients' holistic bio-psycho-social wellness guided by social work professional values (Dziegielewski, 2004: 235). In addition, current social work assessment in health care settings is often under the pressure for moving clients through the system quickly through discharge planning, to save medical costs by reducing period of stay in hospital. Thus, health social workers will need to develop and utilize more comprehensive psychosocial measures to support intervention plans (Dziegielewski, 2004: 211).

Challenges in Health Policy

As social problems of health disparities, poverty, and ill-health are closely linked, health social workers will inevitably adopt grassroots mobilization and collective community actions to fight system injustice and global inequity. However, it is probably much easier to obtain funding for microsocial work compared to macropractices, such as advocacy for system changes. Under heavy workloads, clinical and medical social workers normally find it very difficult to engage in additional activities, such as policy practice.

Many of the industrializing countries are so poor that there are few affordable and accessible primary health clinics and hospitals services available to the population. Social workers, if any, may be hired only by local or international NGOs. Community leaders and grassroot workers can be trained by social workers to provide disaster relief, refugee camps, community support to reduce maternal and infant mortality, enhance access to clean water, promote self-help nutrition programs, healthy sexual behavior, immunization, and women's development projects. Unfortunately, funding for community health in these nations, especially war-affected areas, is often insufficient and unstable. Despite, the fact that there are programs that offer joint master degrees in social work and public health (MSW-MPA), the uptake is usually small with few employment opportunities.

Public health social workers will have to be creative and innovative so to engage the community in social development and poverty alleviation projects. It is essential for social work educators and researchers to generate knowledge in community health promotion, disease prevention, and especially to serve disadvantaged populations who are socially excluded and marginalized. They will have to develop competency models of working with international organizations, mass and social media, as well as to train students in systems change, legislative reform, social entrepreneurship, organization, and project management. There are youth and community social workers who actively participate in promoting a green awareness and lifestyle, as well as engaging themselves in the advocacy to preserve the environment. Such global challenges cannot be addressed by the social work profession alone, as multiple levels of intervention with the involvement of other professions and other sectors will be necessary to tackle such environmental issues (Gray et al., 2013: 12–16).

Challenges in Ethical Issues

Health social workers are often in the midst of ethical debates in their practices. Advancements in medicine and genetics bring along controversial issues in genetic engineering – preserving oocytes and sperms after the death of individuals, gender selection, organ cloning, and transplant. Ethical debates also arise concerning the provision of services and support during life donor organ transplant, adjuvant surgery (such as mastectomy, colectomy) after diagnosis of cancer-prone genes, surrogate motherhood in reproductive technologies, termination of pregnancies when fetus are diagnosed with higher risk of defects, serving persons with different sexual orientations (LGBT) on issues of marriage and child bearing, and serving in disciplinary forces such as prison or refugee camps on management of victims and/or perpetrators of war/violence/rape who are trapped in social discrimination and political oppression. The dilemma of pro-life versus free-will of woman to decide on abortion appears every time in the US presidential election debates. While medical doctors and nurses can refuse to be involved in abortion due to religious reasons, these moral, ethical, and spiritual issues are less commonly discussed or debated among health social workers.

In resolving ethical dilemmas and debates, compliance and adherence to high ethical standards and code of conduct as well as engagement in advocacy are essential for professional social work organizations. The Association of Oncology Social Workers actively engages its members into advocacy and provides them with toolboxes in lobbying political leaders in the fight for cancer in the US. International social work bodies, IASSW and IFSW, state ethical practice principles together. They explicitly encourage social workers around the world to debate and reflect on the dilemmas and to reflect on and make 'ethically informed decisions' in their specific situations.

Opportunities in Education and Knowledge Development

With the importance of health social work being increasing recognized internationally, related social work trainings are offered in many advanced education programs. Most of these programs generally provide specialized courses in medical social work and psychiatric social work. In industrialized countries such as the US, Canada, and Australia where hospital social work is more established, continuing education programs for social workers to be cognized as social work specialist in oncology, pediatrics, gerontology, nephrology, cardiology, palliative care, bereavement, and more, are available. While addressing the needs for cultural sensitive practices, most of these specializations are originated and organized under the Western Medicine framework.

Though conventional training has been based on Western Medicine framework, increasing attention points to the indigenous healing wisdom and practices of the east and south on health social work interventions. There is an increasing proportion of people using Complementary Alternative

Medicine in the treatment of pain, insomnia, diabetes, and autoimmune diseases internationally. Based on Eastern philosophical thoughts, Traditional Chinese Medicine, and meridian concepts, Chan and her colleagues established an integrative IBMS social work empowerment intervention in health care since the 1990s (Lee et al., 2009; Chan et al., 2012). Following years of practice and research, evidence for the IBMS model has been accumulating across various populations including cancer patients, women with health issues, and depressed populations (Lee et al., 2009; Chan et al., 2012). A number of randomized controlled studies have highlighted the effectiveness of the IBMS model on improving the well-being of participants (Chan et al., 2006; Hsiao et al., 2011; Liu et al., 2008; Yang et al., 2009). The team has also started to test the efficacy of traditional Chinese movement exercises, such as Qigong on persons with chronic fatigue syndrome and Taichi on persons with schizophrenia, and evidence of holistic improvement through psychosocial measures and physiological outcome markers was found (Ho et al., 2012; Ho et al., 2013). Given the increasing evidence, there is an urgent need for social work research and practice on how traditional healing wisdom from Asia, Africa, South America, and the Pacific Islands can be integrated and utilized by health social workers.

Conclusion

As a specialization in professional social work, health social work has been historically closely related to the environmental, sociopolitical, and economic context of the society. Health social work can roughly be divided into the medical stream and the public health stream. While medical social workers provide clinical services and support to patients, public health social workers work on global injustice to build a more socially inclusive and harmonious society. Contemporary changes on sociopolitical environment, medical technology, and poverty have led to new challenges in terms of practice, policy, and ethical issues. Continuing education programs, research for knowledge development, and social and policy advocacy are key in meeting the new challenges in promoting general well-being of the populations, as well as the health, harmony, and well-being of society.

See also: Afrocentric Approaches to Social Work; Death and Dying, Sociology of; Early Life Influences on Health and Mortality in Adulthood; Environment and Health; Environmental Stress and Health; Global Agenda for Social Work and Social Development; Health-Related Support and Self-help Mutual Aid Groups; Human Development and Health; Marital Quality and Health; Natural Disasters: Health-Related Aspects; Public Health as Social Science; Public Psychiatry and Public Health Mental Systems; Urban Life and Health.

Bibliography

Blacker, S., Christ, G.H., 2011. Defining social work's role and leadership contributions in palliative care. In: Altilio, T., Otis-Green, S. (Eds.), Oxford Textbook of Palliative Social Work. Oxford University Press, New York, pp. 21–30.

Chan, C.L.W., Tso, I.F., Ho, R.T.H., Ng, S.M., Chan, C.H.Y., Chan, J.C.N., Lai, J.C.L., Evans, P.D., 2006. The effect of a one-hour eastern stress management session on salivary cortisol. Stress and Health 22 (1), 45–49.

Chan, C.L.W., Wang, C.-W., Ho, R.T.H., Ho, A.H.Y., Ziea, E.T.C., Wong, T.V.C.W., Ng, S.M., 2012. A systematic review of the effectiveness of qigong exercise in cardiac rehabilitation. The American Journal of Chinese Medicine 40 (02), 255–267.

Claxton, N., Lamparska, B., Maus, K.C., 2010. The child and parent team approach: a Philadelphia interdepartmental collaborative to provide community-based services to at-risk mothers and newborns. In: Kerson, T.S., McCoyd, J.L.M., & Associates (Eds.), Social Work in Health Settings: Practice in Context, third ed. Routledge, Abingdon, Oxon; New York, pp. 291–303.

Council on Social Work Education, 2009. Advanced Social Work Practice in Clinical Social Work. Alexandria, Virginia.

D'Cruz, H., 2009. Social work knowledge-in-practice. In: D'Cruz, H., Jacobs, S., Schoo, A.M.M. (Eds.), Knowledge-in-practice in the Caring Professions: Multidisciplinary Perspectives. Ashgate, Farnham, pp. 69–92.

Dominelli, L., 2012. Green Social Work: From Environmental Crises to Environmental Justice. Malden Polity Press, Cambridge.

Dziegielewski, S.F., 2004. The Changing Face of Health Care Social Work: Professional Practice in Managed Behavioral Health Care, second ed. Springer Pub. Co, New York.

Egan, M., 2010. Evidence-Based Interventions for Social Work in Health Care. Routledge, New York.

Gehlert, S., Browne, T.A., 2006. Handbook of Health Social Work. Wiley, Hoboken.

Gray, M., Coates, J., Hetherington, T., 2013. Environmental Social Work. Routledge, Abingdon, Oxon; New York.

Hick, S.F., 2002. Social Work in Canada: An Introduction. Thompson Educational Pub., Toronto.

Ho, R.T.H., Chan, J.S.M., Wang, C.W., Lau, B.W.M., So, K.F., Yuen, L.P., Sham, J.S.T., Chan, C.L.W., 2012. A randomized controlled trial of qigong exercise on fatigue symptoms, functioning, and telomerase activity in persons with chronic fatigue or chronic fatigue syndrome. Annals of Behavioral Medicine 44 (2), 160–170.

Ho, A.H.Y., Chan, C.L.W., Leung, P.P.R., Chochinov, M., Neimeyer, R.A., Pang, S.M.C., Tse, D.M.W., 2013. Living and dying with dignity in Chinese society: perspectives of older palliative care patients in Hong Kong. Age and Ageing 424, 455–461.

Hsiao, F.H., Jow, G.M., Lai, Y.M., Chen, Y.T., Wang, K.C., Ng, S.M., Ho, R.T.H., Chan, C.L.W., Yang, T.T., 2011. The long-term effects of psychotherapy added to pharmacotherapy on morning to evening diurnal cortisol patterns in outpatients with major depression. Psychotherapy and Psychosomatics 80 (3), 166–172.

Jennissen, T., Lundy, C., 2011. One Hundred Years of Social Work: A History of the Profession in English Canada, 1900–2000. Wilfrid Laurier University Press, Waterloo, ON.

Karki, K., Miah, M., Jurkowski, E.T., 2013. Global health. In: Keefe, R.H., Jurkowski, E.T., American Public Health Association Public Health Social Work Section (Eds.), Handbook for Public Health Social Work. Springer Pub, New York, pp. 213–229.

Keefe, R.H., Evans, T.A., 2013. Introduction to public health social work. In: Keefe, R.H., Jurkowski, E.T., American Public Health Association Public Health Social Work Section (Eds.), Handbook for Public Health Social Work. Springer Pub, New York, pp. 1–20.

Kurzman, P., 2008. Occupational social work. In: Mizrahi, T., Davis, L.E. (Eds.), Encyclopedia of Social Work, twentieth ed., vol. 3. NASW Press; Oxford University Press, Washington, DC; Oxford; New York, pp. 311–319.

Lee, M.Y., Ng, S.M., Leung, P.Y., Chan, C.L.W., 2009. Integrative Body–Mind–Spirit Social Work: An Empirically Based Approach to Assessment and Treatment. Oxford University Press, New York.

Liu, C.-J., Hsiung, P.-C., Chang, K.-J., Liu, Y.-F., et al., 2008. A study on the efficacy of body–mind–spirit group therapy for patients with breast cancer. Journal of Clinical Nursing 17 (19), 2539–2549.

Mathbor, G.M., Bourassa, J.A., 2012. Disaster management and humanitarian action. In: Lyons, K.H., Hokenstad, T., Pawar, M., Huegler, N., Hall, N. (Eds.), The SAGE Handbook of International Social Work. SAGE, London, pp. 294–310.

Miller, J., 2012. Psychosocial Capacity Building in Response to Disasters. Columbia University Press, New York.

Mizrahi, T., Davis, L.E. (Eds.), 2008. Encyclopedia of Social Work (2th ed.). NASW Press/Oxford University Press, Washington, DC/New York.

Newman, P., 2008. HIV/AIDS. In: Mizrahi, T., Davis, L.E. (Eds.), Encyclopedia of Social Work, twentieth ed., vol. 2. NASW Press; Oxford University Press, Washington, DC; Oxford; New York, pp. 357–363.

Payne, M., 2005. The Origins of Social Work: Continuity and Change. Palgrave Macmillan, Houndmills, England; New York.

Raphael, D., 2000. The question of evidence in health promotion. Health Promotion International 15 (4), 355–367.

Toole, M.J., Waldman, R.J., 1997. The public health aspects of complex emergencies and refugee situations. Annual Review of Public Health 18, 283–312.

Yang, T.T., Hsiao, F.H., Wang, K.C., Ng, S.M., Ho, R.T., Chan, C.L., Lai, Y.M., Chen, Y.T., 2009. The effect of psychotherapy added to pharmacotherapy on cortisol responses in outpatients with major depressive disorder. Journal of Nervous and Mental Disease 197 (6), 401–406.

Relevant Websites

www.aapcsw.org – American Association for Psychoanalysis in Clinical Social Work (AAPCSW).

www.apha.org/membergroups/sections/aphasections/socialwork/about/ – American Public Health Association® © Social Work Section.

www.aosw.org – Association of Oncology Social Work.

www.aposw.org – Association of Pediatric Oncology Social Worker.

www.cansw.org – Canadian Association of Nephrology Social Workers.

www.cswe.org – Council on Social Work Education (CSWE).

www.globalhealth.org – Global Health.

http://ifsw.org/news/ifsw-supports-kyoto-extension/ – IFSW, 2012. IFSW Supports Kyoto Extension.

http://ifsw.org/policies/statement-of-ethical-principles/ – IFSW, 2012. IASSW and IFSW Statement of Ethical Principles.

http://www.iassw-aiets.org/ – International Association of Schools of Social Work.

http://www.icsw.org/ – International Council on Social Welfare.

http://www.ifrc.org – International Federation of Red Cross and Red Crescent Societies (IFRC, Geneva), 2011. Eliminating Health Inequities – Every Woman and Every Child Counts.

http://www.templeton.org/ – The John Templeton Foundation.

http://www.un.org/millenniumgoals/ – Millennium Development Goals (MDGs).

www.socialwork.org – National Association of Social Workers.

http://www.pathfinder.org/our-work/how-we-work/ – Pathfinder (reproductive health).

http://oce.sph.unc.edu/cetac/phswcompetencies_may05.pdf – Practice Standards Committee, 2005. Public Health Social Work Standards and Competencies.

www.swhpn.org – Social Work Hospice & Palliative Care Network.

www.socialworkersinhealthbc.org – The Social Workers in Health Society of British Columbia.

www.sswlhc.org – The Society for Social Work Leadership in Health Care.

http://siteresources.worldbank.org/INTPAH/Resources/Publications/459843-1195594469249/HealthEquityFINAL.pdf – World Bank, 2012.

Health Surveys

Jennifer H Madans and Kassi M Webster, National Center for Health Statistics, Hyattsville, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Information on the health of a population can be obtained through a variety of data collection systems: health surveys represent one mechanism. Major topics to be covered include aspects of survey design, conceptualization and measurement of health status, determinants of health, summary measures of health, health care utilization, and health behaviors. The effect of recent technological advances on data collection methods are discussed as are the growing opportunities for data linkage and the future challenges of providing access to data while protecting confidentiality and privacy.

According to the most recent data available for the United States, 10.1% of the population are in fair or poor health; 7.8% of children under 18 years of age have no health insurance coverage; 68.5% of children 19–35 months old are up-to-date on immunizations; 19.3% of adults smoke cigarettes; and 31.9% are hypertensive based on measured blood pressure or use of blood pressure medication (National Center for Health Statistics, 2012). Such information is used by public health officials to plan, conduct, and evaluate programs; by elected officials to inform policies and legislation; and by researchers to expand knowledge about health and health care. Information on the health of a population can be obtained through a variety of data collection systems; health surveys represent one mechanism.

While health surveys are sponsored and carried out by many entities, core information is generally obtained through relatively large, multipurpose, government-sponsored data collection systems. These types of surveys will be the focus here, although most of the issues discussed pertain to all health surveys. Information from health surveys are used to profile the health of the population, to track demographic shifts and trends, and to monitor change in health and health care. In addition, the data are also used to investigate epidemiologic relationships between risk factors and health outcomes and to provide information for prevention, evaluation, research, planning, and policy.

The multidimensionality of health increases the complexity of health surveys and requires the use of different data collection methods. Of particular importance is how measures of health status are translated into data collection instruments. Measures of determinants of health, health care utilization, and health behaviors are also needed. Recent technological advances in data collection methods are changing how information is being obtained, including growing opportunities for data linkage, but these advances also present challenges for providing access to data while protecting confidentiality and privacy. As nations include general-purpose health surveys in their statistical systems, it is tempting to directly compare the results of these surveys; but cross-national cultural differences and variation in the nature of health care systems as well as survey designs, must be taken into account when making international comparisons.

Definition and Operationalization of Health

The World Health Organization's definition of health as a state of physical, mental, and social well-being and not merely the absence of disease or infirmity, provides the starting point for health surveys (WHO, 1948). However, because this definition is broad, complex, and multidimensional, it is difficult for any single survey to collect all the needed information. Depending on the specific objectives of a particular survey, there will be variation in the universes from which the samples are drawn and in the health and health-related information obtained. Since health is both a biologic and a social construct, health surveys must reflect this duality. In fact, these interrelationships between the biological and social aspects of health are often best approached through health surveys; and advances in technology have made it increasingly possible to incorporate direct measures of the biologic aspects of health into more traditional surveys.

General Design Issues

Health surveys can obtain information directly from individuals (the subject or an informant), can rely on administrative records, or use a combination of both approaches. A wide range of sampling methodologies, sampling frames, and data collection strategies are used, depending on survey objectives and the characteristics of the area in which the survey is done. Countries that have universal health care with centralized administrative systems can utilize designs that would be inappropriate in countries with less well-developed or decentralized health care systems. The simplest sampling scheme is a simple random sample of a population obtained from a complete and up-to-date list, such as the census of the population or a centralized health care system. In the case of surveys of persons, if such lists do not exist, area sampling can be used with clustering to reduce costs. In order to be able to report information on subpopulations of interest as defined by geography and/or demographic, socioeconomic, or health characteristics, designs often include oversampling of these groups. It is often necessary to incorporate screening into designs to achieve this objective. Information can be obtained either by mail, phone, in person interview, examination through the web, or by some combination of modes depending on the ability of the researchers to contact the population using these means. The latter three modes can also incorporate computer-assisted technologies, which enhance data quality and timeliness of data release. The rapid adoption of cell phones has affected phone surveys, generally by reducing

already-low response rates (Marsden and Wright, 2010). To reduce potential bias resulting from differences between the wireless-only and landline-only populations, many phone surveys have chosen to include wireless numbers in the sampling frame; however, response rates for wireless numbers are extremely low (Brick et al., 2006; Blumberg and Luke, 2012). The Internet is growing in popularity as a timely, low-cost data collection mode. However, web-based sampling frames present challenges in obtaining a representative sample. (Marsden and Wright, 2010).

In general, obtaining information directly from the survey subject provides more reliable and valid health data. However, this is not universally true but depends on the nature of the information sought and the characteristics of the subject (Moore, 1988). For example, while accurate information on risk behaviors among adolescents can probably only be obtained from the adolescent in a private environment, adolescents are not good reporters of health care utilization or household income. Even more problematic is the situation where a health condition does not permit a subject to respond for himself or herself. Since eliminating the subject from the survey would seriously bias the results, a proxy respondent is often used.

Although necessary, the use of proxy respondents does introduce a source of error into health surveys, particularly longitudinal surveys. Information on the relationship of the proxy to the subject as well as the conditions leading to the use of the proxy can reduce this error. In other cases, 'subject reporting' is fraught with a great deal of measurement error. For example, information on the characteristics of an insurance plan is usually better obtained from the insurer rather than the person holding the insurance. Finally, reliable assessment of specific types of morbidity, risk factors, and mortality requires direct measurement through physical examination or access to administrative records from health care providers.

Many health surveys limit their populations of interest. For example, it is common for health surveys to only include the noninstitutionalized population. The dramatic differences in the living conditions of the institutionalized and noninstitutionalized populations make it difficult to design survey methodologies that would apply in all situations. The military population is also often excluded from general health surveys. The scope of the universe needs to be clearly defined, especially if populations that differ in their health status – such as persons residing in nursing homes – are omitted.

When obtaining information directly from the population is inefficient or would result in either poor data quality or high costs or both, an alternative is to sample from existing records that were created for other purposes. Information on health care utilization for specific conditions is more easily obtained from hospital or health care provider records than from a population sample, given the relative rarity of the phenomena of interest. The vital statistics system can also be considered a survey of administrative/legal records for the purpose of describing the health of the population. Administrative systems, in addition to providing basic health information, are often used as sampling frames for population-based surveys. They are very attractive for this purpose as they eliminate the need to identify members of the population who have the characteristics of interest.

Health Status

Health status is a multidimensional concept, requiring multiple indicators and multiple methodologies for adequate measurement. Several different indicators of health status are usually included in health surveys, including single summarizing measures; questions relating to disease incidence and prevalence; and questions relating to functioning (physical, cognitive, emotional, and social) or disability (Stewart and Ware, 1992). Health status measures also vary depending on whether they are based on objective information obtained from standardized examinations or medical records or from information obtained from the individual or a proxy. To be most effective, individual health surveys should capture a variety of aspects of health status so as to provide a more comprehensive and complete assessment of health status than would be possible from any single strategy. Measures of health status should also be constructed to be useful in epidemiologic analyses of risk factors as well as for monitoring trends.

There is often interest in using a single question to measure health status in health surveys, as well as in surveys of other areas, such as employment, income, or education. The subject's self-assessment of his/her health status as excellent, very good, good, fair, or poor is a popular summary indicator of health status. This measure has been shown to be highly correlated with other measures of health status and is predictive of mortality and admission to long-term-care facilities (Idler and Angel, 1990). However, this measure can be problematic when used to monitor change over time. The means by which individuals evaluate various aspects of health have been shown to be affected by contextual parameters. Implicit in self-perceived health status is the individual's evaluation of his/her health status against some unstated standard. Societal norms act to define the standard but the norms and the standards change in response to a variety of conditions.

Medical Conditions

A variety of approaches can be used to measure disease incidence and prevalence including diagnosis by a health care provider; symptomatology (if an appropriate symptom battery exits); or medication use. Additional information of interest related to conditions includes date of onset, the use of health care services, and the impact of the condition on the individual's ability to function. Severity of condition is often inferred from the type of medications or health care services used or the nature of the impact of the condition on functional abilities. In examination surveys, diagnostic testing adds to the information available on the prevalence of conditions. Most important, direct physical exams allow for the measurement of previously undiagnosed conditions.

Major issues in the collection of 'condition data' are the number of conditions to be considered and whether the condition needed to be diagnosed by a health care provider. Condition checklists can be used to obtain information on a great number of conditions, but in most cases, only a minimal amount of information on any given condition is collected. In addition, long checklists are difficult to administer and result in reporting errors. However, the use of

comprehensive lists does allow for the ascertainment of comorbidities. An alternative is to obtain more detailed information on a select group of conditions. Criteria such as the prevalence of the condition or its potential impact on the individual and the health care system are used to select the target condition. Access to the health care system also affects whether a condition has been diagnosed and can be reported by a respondent. This affects reported prevalence rates and makes it difficult to investigate the social correlates of health status, since the definition itself incorporates a social dimension. The collection of information on symptoms takes on greater importance when investigating chronic conditions that may or may not have been diagnosed. Diagnostic examination or testing is necessary for some important conditions, which may be asymptomatic in their early or less-severe stages.

Self-reported information on self-perceived health status, disease symptomatology, physician diagnosis, and level of functioning is also obtained through interviews with participants; but the results of physical examination and of the blood- and other tests can be used to measure health status objectively. For example, information on diagnosed diabetes is obtained from the interview, and information on undiagnosed diabetes is obtained from blood tests. Information on functional limitations related to diabetes (e.g., vision and mobility problems) coupled with similar information available from the other components of a survey allows for a detailed characterization of an individual's health status. The information is also used for the investigation of risk factors for disease and for monitoring of functional outcomes. Since the 1960s, the use of examination surveys has grown worldwide, ranging from comprehensive examination surveys to the addition of at least some physical and biological measures to health-interview surveys (Tolonen et al., 2008b). Surveys differ in the extent of physical measurements and the medical and laboratory tests conducted and may or may not include nutritional assessment. In an effort to standardize protocols and information collected across countries, the European Union reviewed methods and developed recommendations for member countries (Tolonen et al., 2008a, 2008b). Examination surveys utilize various sites for collection of physical measurements and laboratory testing, including mobile clinics, health centers, and the participant's residence, which in turn affects the types of measurements and testing that can be done. One of the oldest and most comprehensive examination studies, the National Health and Nutrition Examination Survey, is conducted in the United States and combines health and nutrition interviews with physical examinations and laboratory tests administered by trained health professionals in a specially designed and equipped Mobile Examination Center (MEC) (http://www.cdc.gov/nchs/nhanes.htm). Canada also uses MECs for their Canadian Health Measures Survey (www.statcan.gc.ca). The United Kingdom's Health Survey for England (http://www.natcen.ac.uk/series/health-survey-for-england) conducts health interviews and, for those that agree, a nurse obtains body measurements and blood pressure in the participant's residence.

Information on conditions can also be obtained from the surveys of health care providers (hospitals, private physicians, hospital outpatient departments, emergency departments, ambulatory surgery centers, and long-term care). Sampling frames of providers must first be developed. Random samples of discharges, stays, or visits are then selected from the records of a random sample of providers. Information on conditions is routinely found in medical records in addition to other descriptive information concerning the patient and the encounter. Such survey designs are generally used to obtain information on health care utilization and the health care system, but the information available can also be used to study conditions that are associated with medical care.

Health care or health provider surveys offer independent assessment of the prevalence of disease by approaching the problem from the provider's side rather than the patient's side. Health care surveys are particularly important since they obtain information on rare conditions that would not be picked up in population-based surveys. Since data are obtained from existing patient records, they also allow for larger sample sizes than would be obtained from population-based surveys given the same resources. While information is collected on all conditions for which health care is obtained, the nature of the information available is limited to that which would routinely be obtained from records. In addition, the event (e.g., the discharge) rather than the person, is the unit of analysis. For example, while it is possible to estimate the number of hospital stays associated with hip fracture, it is not possible to estimate the number of persons with hip fracture, even if one could assume that all hip fractures were admitted to hospitals. An enhancement to health care surveys would be to contact patients after the health care encounter and obtain information from them about their health and health care prior to the encounter and after it.

'Cause of death' statistics obtained from a vital statistics system also provide essential information on medical conditions. While not error free, a great deal of effort is put forth to standardize the collection and coding of cause of death through international classification systems (http://www.who.int/classifications/icd/en/) as well as training for persons completing death certificates. The adoption of electronic death record systems will improve the quality of 'cause of death' reporting.

Functioning

An important component of health status concerns its functioning – sensory, physical, cognitive, emotional, and social. While 'conditions' relate more to the physiologic aspects of health, 'functioning' captures the pathways through which health conditions impact social participation. The International Classification of Functioning, Disability and Health (ICF) developed by the World Health Organization provides a framework for measuring health and disability at both individual and population levels (http://www.who.int/classifications/icf/en/). The ICF also contains a detailed classification system that can be used to code various aspects of functioning and disability.

Information on functional status can be ascertained through objective tests but is most often obtained through direct reporting by an individual or a proxy. Information on functional status encompasses a wide range of activities from the very basic, such as raising an arm, to the more complex, such as working and going to school. Obtaining information

on functioning is complicated by the fact that performance is affected not only by physiologic abilities but by the environment as well. A person's ability to walk may be impaired by a condition that leads to weakened muscles, but with the use of braces or a wheel chair along with appropriate accommodations in the physical environment (e.g., ramps), that individual is mobile and able to carry out appropriate social roles. For some purposes, information is needed on functional ability without the use of assistance of any type. In other cases, functioning ability as measured by usual performance is important. In any case, questions concerning how performance should be measured must be specific. A large number of question batteries have been developed to measure function (McDowell and Newell, 1996). Some use only a few questions, focus on the more complex social roles, and do not differentiate between aspects of functioning (physical, cognitive, emotional, or behavioral). Others are quite lengthy and obtain detailed information on a wide range of activities both with and without assistance.

Standard question sets for the measurement of functioning and disability have been developed for use on censuses and surveys by the Washington Group on Disability Statistics, a United Nations (UN) city group that was tasked by the UN Statistical Commission to address issues on noncomparability in disability statistics. The Group has developed and tested a set of six questions covering a core set of functional domains for use on censuses and an extended set of questions for surveys that addresses a wider range of domains and obtains more information for each domain. A set of questions for children is currently being tested. The ICF was used as the framework for developing the question sets and the adoption of these standard, tested questions will enhance the quality and comparability of disability statistics worldwide (Washington Group on Disability Statistics, 2009, 2011).

Many health surveys are limited to the noninstitutionalized population and therefore do not include those segments of the population in poorest health or with the highest level of functional limitations, such as those in long-term care settings. Special studies of these populations can be carried out and there should be some attempt to combine results from the noninstitutionalized and long-term care populations. Changes in the delivery of long-term care with the expansion of transitional establishments have made this an even more important issue for survey designers.

Determinants of Health

Health surveys generally include measures of risk factors, health behaviors, and nonhealth determinants or correlates of health, such as socioeconomic status. The range of measures that can be included is wide and varies by survey. Basic demographic and socioeconomic variables are usually included (i.e., age, race/ethnicity, education, income, urbanization, regionality). Tobacco use, alcohol use, diet, and physical exercise are common health behaviors. In order to identify causal relationships between determinants and outcomes, longitudinal designs are needed; however, cross-sectional surveys can be used to monitor changes in the prevalence of risk factors and other health behaviors. Changes in risk factors can be viewed as early signs of future trends in health status. Such factors can be considered leading health indicators. For example, observed declines in smoking levels signal future changes in lung cancer and other conditions caused by smoking. These leading health indicators can be important tools for planning and evaluating public health programs and policies.

Summary Measures of Health

There has been considerable interest in the development of summary measures of health in order to more easily characterize and monitor the health status of populations and for use in cost-benefit and cost-effectiveness analyses. There is also increasing interest in developing basic health status indicators comparable to the economic indicators currently in use. These indicators would be used to monitor the effects of public policy and to identify areas where interventions are needed. An overall health measure is very appealing in this context. Many health surveys include the information necessary to calculate the summary measures that have been developed. One such approach is to use a measure of morbidity or health status to transform life expectancy into 'healthy' life expectancy. Examples include parameters such as Disability Free Life Years, Healthy Life Expectancy, and Years of Healthy Life. Constructs such as these combine measures of different aspects of health into a single number using a conceptual model that views health as a continuum. The health states along the continuum are each assigned a number that represents the value that either society as a whole or the individual places on the particular health status. Various methods have been used to determine the values of the different health states. (Erickson et al., 1989).

Health Care Utilization

The use of the health care system is a major dimension of health. Information on utilization and related characteristics such as health insurance is collected from person-based surveys and through surveys of administrative records. In person-based surveys, respondents are asked to recall the number of contacts they have had with various health care providers and the nature of the provider and the contact. Information on the receipt of particular services, such as immunizations or mammography, is also obtained. This is often difficult for respondents. Research has shown that reporting of visits to the doctor drops off considerably when the recall period extends beyond 2 weeks (Edwards et al., 1996). To address this recall problem, some surveys ask respondents to keep detailed diaries of their medical care contacts. They obtain the names of the providers and then contact the providers (with the subject's permission) to verify the self-reported information (Monheit et al., 1999). Information on utilization can also be obtained from surveys of providers, and these usually provide more accurate information on the nature of the contact, including the reason for contact and the services provided. Surveys of providers can also yield valuable information on the structure of the health care system itself

and how the characteristics of the system affect access to care and the cost of care.

Longitudinal Designs

Most health surveys use a cross-sectional design which provides data for the calculation of population averages and is most appropriate for monitoring changes in health over time at the population level. A cross-sectional design is not appropriate for investigating causal patterns or for documenting transition probabilities from one health state to another where change at the individual level needs to be measured. A solution would be longitudinal studies which are expensive and more complicated to operate. One alternative is to transform cross-sectional studies into longitudinal studies by linking information from mortality, hospital, and other administrative records to the survey data. To do this, variables such as name, birth date, address, and other identification numbers are linked across unrelated datasets (Jutte et al., 2011). These data linkages allow researchers to examine factors such as health care utilization, patient outcomes, disease progression, and mortality. Population survey data are also linked to environmental exposures such as air monitoring data (http://www.cdc.gov/nchs/data_access/data_linkage_activities.htm). In Sweden and Finland, where all hospital discharges and deaths are registered in national databases, researchers have linked morbidity data to mortality data to study the incidence and progression of cardiovascular disease (Mähönen et al., 2000; Hammar et al., 2001).

Technological Advances

Advances in technology have affected the conduct of all surveys. Computer-assisted telephone and personal interviews have allowed for increasing levels of complexity in survey administration. Errors associated with keying and coding have been replaced by errors associated with programing; however, improvements in the development of survey tools continue. The use of these technologies has resulted in the need for longer lead times prior to implementing a study but has reduced the time from the end of data collection to the availability of results.

The increase in the use of electronic (as opposed to paper) files has greatly expanded the opportunities for linking data from various sources. Ecological or contextual information about the geographic area in which survey respondents live can now be linked to the survey data. This has the advantage of allowing researchers to address different levels of analysis. Survey data can also be augmented with administrative data from mortality files and from health care providers. In addition to reducing survey costs, this often leads to data of improved quality.

Methodological Research and Analytic Techniques

The quality of health survey data has been greatly increased by methodological research on survey techniques (Bradburn and Sudman, 1979; Fowler, 1995; Lessler and Kalsbeek, 1992). A major advancement has been the development and application of methods for questionnaire construction and evaluation (Sirken et al., 1999). Cognitive research laboratories have been established and can investigate how respondents perceive and answer questions. Another fruitful area of research involves study of the behavior of both interviewers and respondents within the social context of the interview situation (Cannel et al., 1977). A more structured approach to the errors found in health surveys will enhance data quality. The results of this research will lead to improved design that minimizes error but will also lead to ways to adjust for known errors in the analysis and interpretation of results (Jabine, 1987; Madans et al., 2011).

Advances have also been made in the development of appropriate analytic tools for use with health survey data. These include methods to account for such issues as survey nonresponse and complex survey design. As it becomes possible to collect data of increasing complexity, it will be necessary to develop the analytic approaches appropriate for the collection methods used (Korn and Graubard, 1999; Biemer et al., 1991; Groves and Couper, 1998; Rubin, 1987).

Privacy and Confidentiality

For ethical reasons and in order to obtain high response rates and valid information, most health surveys closely guard the information provided by respondents. In some cases, the requirement to protect confidentiality is legislatively mandated. This is particularly important for health surveys, given the personal nature of the information collected. Confidentiality can also be protected by not releasing data that could identify a respondent. Most survey sponsors subject files that are to be released for public use to rigorous disclosure review. The risk for inadvertent disclosure has increased in recent years, given the ability to easily download survey data and the availability of increasing amounts of information on the web which can be used to identify individuals in the survey data files. It is the potential for linking survey data to these external databases that increases the risk of disclosure, and the risk increases with the amount of information that is available.

The ability to link external data to survey responses presents an analytic breakthrough because the utility of survey data can be greatly increased through appropriate linkages; however, this ability also greatly increases the risk of disclosure especially when the linkages are done in an uncontrolled and inappropriate way. The increased sensitivity to issues of privacy in many countries, especially as related to health care, is affecting the manner in which confidential data are being protected. There has been a decrease in the amount of data that can be released as public use files. Other mechanisms are being developed so that access to data can be maximized while protecting confidentiality. The use of special use agreements, licensing, and research data centers are examples of these approaches (Journal of Official Statistics, 1998).

See also: Data Bases and Statistical Systems: Health; Health Behaviors, Assessment of; Health Outcomes, Assessment of; Occupational Health; Public Health as Social Science.

Bibliography

Biemer, P.P., Groves, R.M., Hyberg, L., Mathiowetz, N., Sudman, S. (Eds.), 1991. Measurement Errors in Surveys. Wiley, New York.

Blumberg, S.J., Luke, J.V., Dec 2012. Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, January–June 2012. National Center for Health Statistics. National Health Interview Survey Early Release Program.

Bradburn, N.M., Sudman, S., 1979. Improving Interview Method and Questionnaire Design. Jossey-Bass, San Francisco.

Brick, J.M., Dipko, S., Presser, S., Tucker, C., Yuan, Y., 2006. Nonresponse bias in a dual frame sample of cell and landline numbers. Public Opinion Quarterly 70 (5), 780–793.

Cannel, C.F., Monquis, K.H., Launent, A., 1977. A summary of studies of interviewing methodology. Vital and Health Statistics 1 (69).

Disclosure limitation methods for protecting the confidentiality of statistical data. Journal of Official Statistics 14, 1998, 337–573 (special issue).

Edwards, W.S., Winn, D.M., Collins, J.G., 1996. Evaluation of 2-week doctor visit reporting in the National Health Interview Survey. National Center for Health Statistics. Vital and Health Statistics 2 (122).

Erickson, P., Kendall, E.A., Anderson, J.P., Kaplan, R.M., 1989. Using composite health status measures to assess the nation's health. Medical Care 27 (3), 566–577.

Fowler, F.J., 1995. Improving Survey Questions. Sage, Thousand Oaks, CA.

Groves, R.M., Couper, M.P., 1998. Nonresponse in Household Interview Surveys. Wiley, New York.

Hammar, N., Alfredsson, L., Rosén, M., Spetz, C.L., Kahan, T., Ysberg, A.S., 2001. A national record linkage to study acute myocardial infraction incidence and case fatality in Sweden. International Journal of Epidemiology 30, S30–S34.

Idler, E.L., Angel, R.J., 1990. Self-rated health and mortality in the NHANES I. Epidemiologic follow-up study. American Journal of Public Health 80 (4), 446–452.

Jabine, T.B., 1987. Reporting chronic conditions in the National Health Interview Survey: a review of tendencies from evaluation studies and methodological tests. National Center for Health Statistics. Vital and Health Statistics 2 (105).

Jutte, D.P., Roos, L.L., Brownell, M.D., 2011. Administrative record linkage as a tool for public health research. Annual Review of Public Health 32, 91–108.

Korn, E.L., Graubard, B.I., 1999. Analysis of Health Surveys. Wiley, New York.

Lessler, J.T., Kalsbeek, W.T., 1992. Nonsampling Error in Surveys. Wiley, New York.

Madans, J.M., Miller, K.S., Maitland, A., Willis, G.B. (Eds.), 2011. Question Evaluation Methods: Contributing to the Science of Data Quality. Wiley, Hoboken, NJ.

Mähönen, M., Salomaa, V., Keskimäki, I., Moltchanov, V., 2000. The feasibility of routine mortality and morbidity register data linkage to study the occurrence of acute coronary heart disease events in Finland: the Finnish Cardiovascular Diseases Registers (CVDR) Project. European Journal of Epidemiology 16, 701–711.

Marsden, P.V., Wright, J.D. (Eds.), 2010. Handbook of Survey Research. Emerald Group Publishing, Bingley, UK.

McDowell, I., Newell, C., 1996. Measuring Health: a Guide to Rating Scales and Questionnaires. Oxford University Press, Oxford, UK.

Monheit, A.C., Wilson, R., Arnett, R.H. (Eds.), 1999. Informing American Health Care Policy: the Dynamics of Medical Expenditure and Insurance Surveys. Jossey-Bass, San Francisco, pp. 1977–1996.

Moore, J., 1988. Self/proxy response status and survey response quality: a review of the literature. Journal of Official Statistics 4 (2), 155–172.

National Center for Health Statistics, 2012. Health, United States 2011. National Center for Health Statistics, Hyattsville, MD.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

Sirken, M.G., Herrmann, D.J., Schecter, S., Tourangeau, R. (Eds.), 1999. Cognition and Survey Research. Wiley, New York.

Stewart, A.L., Ware, J.E. (Eds.), 1992. Measuring Functioning and Well-Being: the Medical Outcomes Study Approach. Duke University Press, Durham, NC.

Tolonen, H., Koponen, P., Aromaa, A., et al. (Eds.), 2008a. Recommendations for the Health Examination Surveys in Europe B21/2008. National Public Health Institute, Helsinki, Finland.

Tolonen, H., Koponen, P., Aromaa, A., et al. (Eds.), 2008b. Review of Health Examination Surveys in Europe B18/2008. National Public Health Institute, Helsinki, Finland.

Washington Group on Disability Statistics, 2009. Understanding and Interpreting Disability as Measured Using the WG Short Set of Questions. Retrieved from: http://www.cdc.gov/nchs/data/washington_group/meeting8/interpreting_disability.pdf.

Washington Group on Disability Statistics, 2011. Washington Group – Extended Question Set on Functioning. Retrieved from: http://www.cdc.gov/nchs/data/washington_group/WG_Extended_Question_Set_on_Functioning.pdf.

World Health Organization, 1948. World Health Organization Constitution in Basic Documents. World Health Organization, Geneva.

Relevant Websites

www.cdc.gov/nchs – Centers for Disease Control and Prevention, National Center for Health Statistics.

http://www.cdc.gov/nchs/nhanes.htm – Centers for Disease Control and Prevention, National Center for Health Statistics: National Health and Nutrition Examination Survey.

www.statcan.gc.ca – Statistics Canada.

http://www.thl.fi/fehes/ – Feasibility of a European Health Examination Survey.

http://www.natcen.ac.uk/series/health-survey-for-england – NatCen: Health Survey for England.

http://www.who.int/classifications/icd/en/ – World Health Organization International Classification of Diseases Classifications.

http://www.who.int/classifications/icf/en/ – World Health Organization International Classification of Functioning, Disability and Health Classifications.

http://www.cdc.gov/nchs/ndi.htm – Centers for Disease Control and Prevention, National Death Index.

http://www.cdc.gov/nchs/data_access/data_linkage_activities.htm – Centers for Disease Control and Prevention, NCHS Data Linkage Activities.

Hebb, Donald Olding (1904–1985)

Peter M Milner

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Hebb was very well grounded in the psychological theories of the early twentieth century by his graduate training at McGill, Chicago, and Harvard before his almost unique experience of measuring the mental capacities of human patients with surgical brain lesions. This impressed upon him the importance of the brain for determining behavior. He recognized that, although the proponents of the mainstream stimulus–response theories claimed to reject neural explanations, they in fact based their theories on a ridiculously simplified brain in which sensory input went directly to motor output. He therefore developed a behavior theory that involved a more realistic model of the brain and in doing so, brought about a profound change in the course taken by psychology.

Introduction

Donald Hebb was responsible in no small part for the great changes in the outlook of psychologists that took place in the 1950s. Few, if any, books have changed the course of psychology as radically as his major work, *The Organization of Behavior* (1949). The cognitive theory based on the association of neurons representing ideas that he advanced in that monograph was an important factor in the demise of neo-Pavlovian behaviorism, the attempt to explain behavior in terms of conditioned stimulus–response associations that was very influential up to that time. But more important, on the positive side, Hebb overcame the widespread aversion among psychologists to explanations of behavior in terms of brain activity. What was known about the brain at the time seemed irrelevant to the commonly espoused behavior models, so that attempts to introduce physiological interpretations were often regarded as waste of time and effort.

Mainstream behaviorists also rejected mental explanations of behavior, believing that thoughts and other mental processes were beyond the bounds of physical measurement, hence resorting to them might be regarded as vitalism. This forced them to ignore phenomena such as attention, expectancy, and ideas that Hebb regarded as essential to behavior theory. Hebb's mentor, Karl Lashley, postulated that concepts were represented in the cortex by patterns of neural activity that he called engrams; Hebb argued that if ideas can be represented by physical processes in the brain, they should be just as capable of forming associations with each other as were the behaviorist's stimuli and responses. Hebb's model incorporating association of ideas was more versatile than stimulus – response models, so it was enthusiastically adopted by the many psychologists who had found the established theories lacking in scope.

The Organization of Behavior presented the first theory of behavior based on the nervous system to attain widespread acceptance among psychologists. There is ample evidence that the consequent interest and participation of psychologists in brain research helped to fuel the vigorous growth of neuroscience that took off during the last half of the twentieth century.

Early Life and Education

Hebb, the eldest of four siblings, was born on 22 July 1904 in Chester, Nova Scotia. Both his parents were physicians. He was educated at home until he was 8 years old, by his mother, who used Montessori methods of teaching. He then attended the local school, which he found very boring.

Hebb's ambition as a young man was to become a novelist. He graduated from Dalhousie University (Halifax, Nova Scotia) in 1925 with a BA in English and Philosophy, though his father had urged him to specialize in science and mathematics, in which he obtained higher grades. While awaiting inspiration for the 'Great Canadian Novel,' he returned to his village school as a teacher. Some of the students remembered him as a former student and he was not a great success. Hebb recalls (Hebb, 1980a) that when the Provincial Inspector visited the school he attempted to cheer him up by telling him "I wasn't worth a damn in my first year either." By the time Hebb joined the teaching staff at McGill in 1947 he had become an excellent teacher, both of very large undergraduate classes and small graduate seminars.

Hebb's novel made little headway so, in search of wider experience, he spent the next summer as an itinerant farm worker in western Canada and then tried, unsuccessfully, to get a job as a deckhand in Vancouver.

In 1927 Hebb was back East, teaching in a Montreal high school. As a would-be novelist he had become interested in Sigmund Freud's ideas and applied to McGill University to study psychology part time. Professor Tait, the chairman of the psychology department at the time, was also from Nova Scotia and had known Hebb's mother in college, so he took it for granted that Hebb had to be intelligent. He was admitted to McGill in 1929 as a part-time graduate student in psychology. At the same time, Hebb was also appointed Principal of an elementary school in a working-class suburb of Montreal and, with assistance from members of the McGill psychology department, enthusiastically set about instituting and testing novel teaching methods intended to make learning more attractive to students (perhaps based on fond memories of his Montessori experience).

In 1930 Hebb was bedridden for a year with a tubercular hip, during which time he read Pavlov, Sherrington, and Ariëns Kappers (on neurobiotaxis) and demonstrated his innate empiricism by conceiving a theory of how spinal reflexes might be learned *in utero*. This formed the basis of an MA thesis titled *Conditioned and Unconditioned Reflexes and Inhibition*; Hebb later described it as nonsense (Hebb, 1980a). About 60 years later, I could not find the thesis in the McGill library and it was presumed to be lost. The idea that spinal reflexes are learned may well be nonsense, but as Richard Brown, who eventually unearthed the thesis, points out (Brown and Milner, 2002, 2003), the neural mechanism Hebb proposed to bring it about was not. It later formed the basis for the learning synapse that became known as the Hebb Synapse.

In 1932 Hebb began doctoral research, still part-time, at McGill. The experiments, which involved Pavlovian conditioning of dogs, were supervised by Leonid Andreyev, a former student of Pavlov. Hebb was not impressed either by the technique or by its interpretation, and when his first wife was killed in a motor accident, he abandoned the part-time research. Hebb enjoyed teaching and administration and was tempted to make a career of being a Headmaster. His enthusiasm waned however when the Quebec educational authorities clamped down on his more radical teaching innovations and he finally chose to continue his psychological studies.

In 1934, Hebb was accepted at Yale to work for a graduate degree with Robert Yerkes, a well-known psychologist who studied intelligence in humans and other primates, but he was dissuaded from accepting this arrangement by Boris Babkin. Babkin was another former student of Pavlov who had moved to the McGill Department of Physiology, and he was one of the examiners of Hebb's Master's thesis. He urged Hebb to go to the University of Chicago and work with Karl Lashley, whose book, published a few years earlier (Lashley, 1929), had established him as the preeminent physiological psychologist of the period, Hebb followed Babkin's advice.

This choice was significant as it is very unlikely that *The Organization of Behavior* would ever have been written had Hebb gone to Yale where Clark Hull, the leading exponent of a neo-Pavlovian behaviorist school, held sway. Yerkes was mainly interested in intelligence testing in humans and other primates. He was of the opinion that foreigners were less intelligent than American citizens because they obtained lower scores in his tests (which incidentally had many culturally loaded questions), a degree of chauvinism that might have been a factor in Babkin's advice.

Hebb spent a year in Chicago where, in addition to his close association with Lashley, he attended lectures on Gestalt psychology by Wolfgang Köhler, on statistics by Thurstone, and on physiology by Nathanial Kleitman the authority on sleep, whose lectures, according to Hebb, were appropriately soporific. Lashley then accepted the offer of a professorship at Harvard and arranged for Hebb to go with him to continue his graduate research as a Harvard scholar. The transfer interrupted Hebb's research on the effect of cortical lesions on spatial learning in the rat. So Lashley suggested that he study the vision of dark-reared adult rats to ascertain the effects of being reared in darkness – an experiment that could be completed in the time available.

Hebb, the empiricist, might have been expected to believe that vision would be impaired in dark- reared rats. Instead, perhaps influenced by Lashley's predilections, he found that dark-reared groups of rats could learn to discriminate brightness and shape just as well as normally reared groups. However, he managed to ignore the fact that it took the dark-reared rats six times as many trials as the control rats to acquire the skill. His thesis was accepted and he was awarded a Harvard PhD in 1936.

Montreal Neurological Institute

In the 1930s, physiological psychology was not taken seriously by most psychologists and jobs were scarce. Lashley gave Hebb a postdoctoral year of research so that he could complete the work on spatial learning that he had started in Chicago. The next year, 1938, Boris Babkin again played a role in shaping Hebb's career, though on this occasion an indirect one. Hebb's sister Catherine was doing research at McGill for her PhD in physiology with Babkin, when she heard that Wilder Penfield, the neurosurgeon who had recently founded the Montreal Neurological Institute (Neuro), was looking for somebody to study mental changes resulting from brain operations in epileptic patients. She relayed this information to her brother who applied for the job and was offered a two-year Fellowship. Penfield's operations were mostly frontal and temporal lobectomies performed to remove epileptic scar tissue.

Hebb did not rely entirely on standard tests; he devised several ingenious tests for special purposes including a series of ill-drawn scenes containing incongruous elements, which the patient was asked to point out. Hebb found a patient with a right temporal lobe removal who showed a deficit on these visual tests; an early suggestion that the temporal lobe might be involved in vision. His most surprising finding was that a large removal of the frontal cortex had little effect on intelligence as measured by a standard Stanford–Binet test (Hebb and Penfield, 1940). The frontal cortex, being relatively larger in humans than in other animals, had been thought to be very important for intelligence. At first Hebb was disappointed at his finding, but then he realized he had uncovered an important phenomenon that required explanation. Much of Hebb's subsequent research was based on his investigation of this problem.

Hebb maintained his association with Penfield and other members of the Neuro staff. When he returned to McGill in 1947 he sent his graduate students to seminars at the Neuro given by Herbert Jasper in neurophysiology and by Francis McNaughton in neuroanatomy. Hebb also obtained permission from Penfield to install one of his graduate students at the Neuro, in order to study patients before and after surgery. This had a far-reaching effect on our knowledge of the effects of brain lesions in human patients, mainly because one of the first students Hebb sent to occupy this position was Brenda Milner. She is still there, more than 60 years later.

Primate Research with Lashley

After his Neuro Fellowship, Hebb obtained a post as Lecturer at Queen's University (Kingston, Ontario). There, in addition to

teaching four courses he, with his student K. Williams, devised a rat-maze (the Hebb–Williams maze) with movable partitions so that he could continue his brain-lesion research using rats. In 1942 Lashley offered him a 5 year Fellowship to study the effects of brain lesions in chimpanzees at the Yerkes Laboratories of Primate Biology in Orange Park, Florida, where Lashley had taken over after Yerkes retired. Hebb would have preferred to study the effect of suppressing learning in puppies to test his frontal-lobe theory, but his heavy teaching load at Queens made that difficult and so he settled for the chimpanzees.

The Orange Park project was to study the effects of brain lesions on intelligence and emotion in chimpanzees. Lashley was to work on intelligence tests and Hebb on tests of emotion. Lashley was surprised to discover that it was more difficult to measure learning in chimpanzees than in rats, with the result that no brain lesions were made before 1947 when Hebb's Fellowship ended. Hebb did not waste his time however; he had the opportunity to observe the personalities of a group of 30 chimpanzees and recognized many of the foibles to be found among a random group of people. In the process he assembled a good stock of chimpanzee stories that contributed to the popularity of subsequent lectures. More important, with the intellectually stimulating group of psychologists working with Lashley in Orange Park he also found incentives and time to write a great deal of his most influential publication, *The Organization of Behavior*.

McGill Once More

During the 1939–45 war, the McGill psychology department had dwindled to one part-time child psychologist, so in 1946 Prof. Robert McLeod was appointed Chairman with the mandate to rebuild the department. Hebb's fellowship at Orange Park ended in 1947, so he was an obvious candidate for the rebuilding team. He moved to McGill, bringing the manuscript of his book with him. He taught a graduate seminar and distributed copies of the chapters to the students for discussion. In those days, with a departmental teaching staff of only four or five, graduate seminars were held in the evenings. The students in Hebb's first seminar included Mortimer Mishkin, Bernard Hymovitch, Donald Forgays, Lila Ghent (Brain), Herbert Lansdell, Woodburn Heron, and Brenda Milner. Brenda was a graduate of Cambridge University who had a full time job teaching psychology, in French, at the Université de Montréal. She was interested in McLeod's phenomenological approach, and the previous year she had obtained permission to attend his evening graduate seminars at McGill. The next year, when Hebb gave his 'first-year' seminars, she also went to those.

Brenda was in Montreal because I had been sent from England during the war as a member of a group of scientists to work on nuclear energy, and she had come with me. Just before the end of the war, I was transferred from Montreal to Chalk River, a sparsely populated site in Ontario where the nuclear reactor was to be built. Brenda wrote to me very excitedly about Hebb's seminars and sent me copies of chapters that Hebb had handed out. Here was someone trying to figure out how neural circuits could explain behavior.

My speciality was electronic circuitry and in 1947 I was one of the multitude trying to decide on a peacetime career. I had many options but for several reasons I decided to stay in Canada and go back to school. I knew some of the McGill physicists, as they sometimes visited Chalk River. Prof. John Foster was building a cyclotron at McGill and I could have had the job of designing an ion source for it. Reading Hebb's chapters convinced me that I would enjoy studying with him more than with the physicists, so the next time I came to Montreal I made an appointment to talk to him about it.

I do not remember whether I knew it at the time, but when I met Hebb he had just been appointed chairman of the psychology department, as McLeod was moving to Cornell. I was duly impressed by Hebb and my interview with him went well. I discovered years later, when I read his autobiography, that he gave me more or less the same treatment he had received when he was in a similar situation 20 years earlier. He sent me away for a year to learn some psychology with the promise of admission to graduate school if I qualified.

The next year I experienced the famous Hebb Seminar first-hand. My first assignment was an article in the Journal of Neurophysiology by Lorente de Nó, a difficult writer, about the feedback loops in the cerebral cortex.

Having made a highly successful foray into the world of behavioral neuroscience (physiological psychology as it was then called), Hebb returned to teaching and administration with enthusiastic vigor. He taught the introductory course to overflow classes for 27 years, and wrote three editions of a highly respected introductory textbook. He also conducted a seminar for all incoming psychology graduate students (including clinical), instilling a thorough understanding of a range of theoretical issues and availing himself of the opportunity to coach the participants in the presentation of succinct and closely timed talks (his stopwatch and tapping pencil were famous far and wide, as were many of the seminars' graduates).

To test some predictions of his theory, he supervised research on the effects of early environment on adult behavior in rats and dogs, and the effects of sensory deprivation on human thought processes. He discovered that well-paid volunteers, subjected to extreme sensory deprivation, could not stand it for more than a few days (Heron, 1957). Hebb was unusual in that, despite being the principal investigator and final editor of ensuing publications of the work he supervised, he did not put his name on it unless he had collected some of the data himself. The deprivation experiments caused trouble later when, as discussed by Brown (2007), Hebb was accused of providing government agencies with a method of torture.

Hebb's Contributions to Psychological Theory

During his years as a graduate student, Hebb had learned the hard way what was wrong with contemporary psychology. At McGill, as is evident from the topic of his Masters thesis, he flirted with the extreme empiricism then rampant in North American psychology. Influenced by Boris Babkin and Leonid Andreyev, former students of Pavlov, he passed through a Pavlovian phase. When he moved to Chicago he entered the

Gestalt enclave of Wolfgang Köhler and Karl Lashley; and at Harvard he encountered the historian, Edwin Boring, who taught him what Psychology had been like before Behaviorism. Sooner or later, for reasons that are cogently presented in the first few chapters of *The Organization of Behavior* (Hebb, 1949), Hebb found flaws in all the established schools of psychology.

Because of his research on cortical lesions and his neural theories, Hebb is usually labeled a physiological psychologist, but his interests were really much broader. *The Organization of Behavior* was an important catalyst, aiding the revival of cognitive psychology, and although he was a zealous proponent of the idea that the brain is the organ of mind, Hebb's interest in neurophysiology took second place to his interest in thinking and other aspects of behavior. In his hands, neurologizing was a tool to restore respectability to the study of mental processes.

As a psychologist, Hebb was frustrated by the behaviorists' refusal to acknowledge such important processes as attention, expectancy, and thought, lest they should be accused of vitalism; hence his eagerness to show that those mental processes could be performed by a material brain.

Hebb based his neural model as far as possible on established physiological principles but neuroscience was on the brink of a revolution, and many of those principles were soon to be revised. For example, in 1951 Eccles, whose strenuous defense of the electrical theory of synaptic transmission Hebb had accepted, observed excitatory and inhibitory chemical synapses and recanted (Eccles, 1953). In addition, the behavioral importance of subcortical structures such as the reticular arousal system, the amygdala and the hippocampus became clearer; and improved staining techniques revealed a much more complex visual system than was known to Hebb when he was writing *The Organization of Behavior*. The fact that these revelations had almost no effect on the impact of Hebb's book shows to what extent the psychological implications of his neurologizing outweighed the underlying physiology. By proposing a material basis for concepts, Hebb demolished barriers to the study of cognitive processes that the behaviorists had erected a generation earlier.

In his Presidential address to the Experimental Psychology Division of the American Psychological Association in 1954, Hebb (1955) introduced the reticular activating system into his model to extend his explanation of emotion; but apart from that he did little to revise the theory in the light of new neuroscientific data. In spite of this, Hebb made a valuable contribution to neuroscience. As pointed out earlier, many psychologists, introduced to the nervous system by Hebb's theorizing, contributed useful behavioral techniques and ideas to investigations of neural processes.

Hebb's most famous personal contribution to neuroscience was not based on neurophysiological research; it was inferred from behavior. Most physiologists thought that learning synapses were strengthened by use, but Hebb, along with other psychologists, recognized that association involved at least two active paths. Thus he postulated that learning changes at synapses require simultaneous firing of both presynaptic and postsynaptic neurons. As pointed out by Brown, synapses with properties similar to those postulated were subsequently identified experimentally and were given the name 'Hebb synapses.'

Hebb held strong views about the 'nature–nurture' question. After his early gyrations, he adopted the unassailable position that both inheritance and experience are totally necessary for intelligent behavior, using the analogy of area (the product of length and breadth). Of course, one may still be interested in whether an area is square, long and narrow, or short and fat. In spite of Hebb's bipartisanship, his concept of the development of vision was long in the 'nurture' dimension; in fact, Hebb's visual cortex bore a strong resemblance to Locke's *tabula rasa*. Thus, Hebb assumed that neurons are randomly connected in the newborn visual cortex. With experience, synapses are strengthened between neurons activated simultaneously by sensory input, producing groups of connected neurons (*cell assemblies*) that then represent the input. Hebb further postulated that a complex figure elicits a series of eye movements, thereby activating a succession of associated cell assemblies that Hebb called a *phase sequence*.

Anatomical diagrams of the cortex suggested to Hebb that cell assemblies consist of networks of closed loops. This led him to speculate that once a cell assembly is activated, impulses continue to circulate around the loops for some time (reverberation), constituting short-term memory. In this way Hebb explained the persistence of an idea or mental image in the absence of the corresponding object. Today, short-term memory may be explained more plausibly by persistent chemical traces, but Hebb never abandoned his reverberation theory, which dominated the thinking of many neural-model builders in the latter part of the last century.

According to Hebb, the impulses reverberating around the cortical loops of an activated cell assembly produce synaptic changes that bind the neurons together more permanently. This explains long-term memory. Ideas arouse other ideas by association, thus cell assemblies must acquire associations with each other whenever they are active at about the same time. Hebb never satisfactorily explained how, after linking in this way, they avoid merging to become bigger cell assemblies.

The capacity to explain intention, attention, and other mental processes was an important goal of Hebb's theorizing but one that, in *The Organization of Behavior*, he was content to achieve in principle rather than by rigorous neural modeling. Short-term expectancy and attention were explained in terms of the phase sequence getting ahead of the sensory input. Thus a previously experienced sequence of stimuli sets up a central process that facilitates cell assemblies in advance of the sensory input they represent. Hebb postulated higher-order central processes that facilitate more extended sequences. In much of his writing Hebb treats the cell assembly as equivalent to an idea; its neural basis is tailored to fit behavioral observation.

In later years Hebb received many honors; he was President of both the Canadian (1952) and the American (1960) Psychological Associations. He was elected a Fellow of the Royal Society of London (1966), a Foreign Associate of the National Academy of Sciences (USA), and an Honorary Member of the American Academy of Arts and Sciences. He was elected posthumously to the Canadian Medical Hall of Fame. In1970 he was appointed Chancellor of McGill University. In 1974 Hebb retired to his family home in Chester Basin and became active in the psychology department of his *alma mater*,

Dalhousie University, where he was appointed Honorary Professor.

He published his last monograph, *Essay on Mind* (1980a). This is really a multitude of short essays, for the most part defending, though not dogmatically, his conviction that the mind is a biological function. Hebb, as usual, deploys his arguments with enviable and devastating reasonableness.

Summary

The Organization of Behavior is now a classic. Many psychologists regard it as the herald of the era of cognitive research that blossomed in the 1950s. Hebb's notion of the neural representation of concepts and their associations with each other is still common currency among psychologists, even those who prefer to ignore the brain; but devotees of neural networks and neurophysiologists studying sensory systems are those who most often refer explicitly to his theory.

After Hebb left Lashley's 'think-tank' in Florida and took up his duties as chairman of the McGill Psychology Department, most of his energies turned to building and running a strong department, with emphasis on teaching. Although Hebb took every opportunity to promote neurologizing in psychology, he showed little further inclination to keep pace with the rapid advances in neuroscience or to make any major changes to his neural theory. He was well aware that changes were required but he was content with the knowledge that his book had turned psychological theory toward the brain.

Apart from his own characteristically stylish and witty account (Hebb, 1980b), there have been several more recent commentaries that complement this entry for those who wish to learn more about Hebb's life and work. These include nine essays in his honor, written by former students and edited by Jusczyk and Klein (1980) and a delightful chapter by Stephen Glickman (1996), which admirably portrays the man and his achievements. Fentress (1999) and Klein (1999) edited a special issue of the *Canadian Journal of Experimental Psychology* that contains a further collection of papers on Hebb and on research inspired by his ideas. Orbach, a student of both Hebb and Lashley, published a book (1998) in which the contributions of his two mentors are discussed and compared.

In 2002 *The Organization of Behavior* was out of print and difficult to find, so Richard Brown had it reprinted. A foreword to this recent edition by Brown and Milner contains more information about Hebb. Lists of his publications and biographical obituaries are also included. Brown and Milner (2003) also published an article, *The legacy of Donald O. Hebb, more than the Hebb Synapse*. Other recent biographies of Hebb include Brown (2007) and Glickman and Brown (2008).

See also: Brain Stimulation Reward; Cognitive Neuropsychology, Methods of; Cognitive Psychology: History; Learning and Memory, Neural Basis of; Long-Term Potentiation and Long-Term Depression; Synaptic Mechanisms and Protein Synthesis in Memory.

Bibliography

Brown, R.E., 2007. Alfred McCoy, Hebb, the CIA and torture. Journal of the History of the Behavioral Sciences 43, 205–213.

Brown, R.E., Milner, P.M., 2003. The legacy of Donald O. Hebb; more than the Hebb Synapse. Nature Reviews; Neuroscience 4, 1013–1019.

Brown, R.E., Milner, P.M., 2002. Foreword to the Erlbaum Edition: Hebb, D.O., (2002). The Organization of Behavior. Lawrence Erlbaum Associates, Mahwah, NJ. F1–F14.

Eccles, J.C., 1953. The Neurophysiological Basis of Mind. Oxford University Press, Oxford.

Fentress, J.C., 1999. The organization of behavior revisited. Canadian Journal of Experimental Psychology 53, 8–20.

Glickman, S.E., Brown, R.E., 2008. Hebb, Donald Olding. In: Koertge, Noretta (Ed.), New Dictionary of Scientific Biography. Scribner, New York.

Glickman, S.E., 1996. Donald Olding Hebb: returning the nervous system to psychology. In: Kimble, G.A., Boneau, C., Wertheimer, M. (Eds.), Portraits of Pioneers in Psychology, vol. 2. Erlbaum Press, Mahwah, NJ, pp. 227–244.

Hebb, D.O., 1949. The Organization of Behavior: A Neuropsychological Theory. Wiley, New York, 1–335.

Hebb, D.O., 1955. Drives and the C.N.S. (conceptual nervous system). Psychology Review 62, 243–254.

Hebb, D.O., 1980a. Essay on Mind. Erlbaum Press, Hillsdale, NJ.

Hebb, D.O., 1980b. D.O. Hebb. In: Lindzey, G. (Ed.), A History of Psychology in Autobiography, vol. 7. Freeman, San Francisco, CA, pp. 273–303.

Hebb, D.O., Penfield, W., 1940. Human behavior after extensive bilateral removal from the frontal lobes. Archives of Neurology and Psychiatry 44, 421–438.

Heron, W., 1957. The pathology of boredom. Scientific American 196, 52–56.

Jusczyk, P.W., Klein, R.M. (Eds.), 1980. The Nature of Thought. Essays in Honor of D.O. Hebb. Erlbaum Press, Hillsdale, NJ.

Klein, R.M., 1999. The Hebb legacy. Canadian Journal of Experimental Psychology 53, 1–7.

Lashley, K.S., 1929. Brain Mechanisms and Intelligence. University of Chicago Press, Chicago, IL, pp. 1–186.

Orbach, J., 1998. The Neuropsychological Theories of Lashley and Hebb. University Press of America, Lanham, MD.

Hegel, Georg Wilhelm Friedrich (1770–1831)

Miriam Wildenauer, University of Heidelberg, Heidelberg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

G.W.F. Hegel (1770–1831) was one of the last great philosophers, who developed a system of all philosophical sciences: Logic, Philosophy of Nature, and Philosophy of Spirit. This article gives an outline about his life, philosophy, and influence on the social sciences. Hegel was born in Stuttgart (Germany), educated at the University at Tübingen, and worked as a private tutor in patrician families in Bern and Frankfurt. He began his academic career as an Associate Professor in Jena, was then the editor of a daily newspaper of Bamberg, and then principal of a secondary school in Nuremberg for 8 years. He then taught as a regular professor at the University of Heidelberg (1816) and at the University of Berlin (1818–31). His main works are *Phenomenology of Spirit* (1807), *Science of Logic* (1812–16), *Encyclopedia of the Philosophical Sciences in Outline* (1817, 1827, and 1831), and *Elements on the Philosophy of Right* (1821). In the development of the social sciences, Hegel's philosophy influenced not only Karl Marx, but also, what is less known, Lorenz von Stein, the St Louis Hegelians, Wilhelm Wundt, G.H. Mead, and Juergen Habermas.

Biographical Data

Hegel, Georg Wilhelm Friedrich the philosopher was born on 27 August 1770, in Stuttgart (Germany), and died on 14 November 1831, in Berlin. In 1788, Hegel, son of a medium-grade civil servant, entered the Protestant Seminary in Tübingen, where he won the friendship of his roommates Hoelderlin and Schelling. He held a ducal scholarship for studying Theology and Philosophy at the University of Tübingen. After graduating in 1793, he earned a living as a private tutor in patrician families, first in Bern, Switzerland (1793–96), and then in Frankfurt (1797–1800), where he regularly met Hoelderlin. At the beginning of 1801, Hegel joined Schelling in Jena and qualified for lecturing with his habilitation thesis *On the Orbits of the Planets*. While at Jena (1801–06), Hegel published his *The Difference Between Fichte's and Schelling's System of Philosophy* (1801). In the years 1802 and 1803, his essays *The Relationship of Skepticism to Philosophy, Faith and Knowledge*, and *On the Scientific Ways of Dealing with Natural Law* came out in the *Critical Journal of Philosophy*, which Schelling copublished with Hegel. The major work of this period was, however, the *Phenomenology of Spirit (PS)* (1807), which was intended to be the first part of Hegel's system of philosophical sciences.

After the Battle of Jena, Hegel, having no income, took over as editor of a newspaper in Bamberg in 1807 and 1 year later, he became the principal of a secondary school in Nuremberg. There, Hegel married Marie von Tucher with whom he had two sons. During the Nuremberg period, Hegel published his *Science of Logic* in three volumes (1812, 1813, and 1816), the first of which, the *Doctrine of Being*, reappeared in heavily revised edition in 1832. In 1816, Hegel was finally granted the privilege of earning a secured income as a full professor in Heidelberg. In 1818, owing to the personal intervention of Minister Altenstein, he became Fichte's successor in Berlin.

From 1816 on, Hegel published – apart from several minor writings – the *Encyclopedia of Philosophical Science* (1817 in Heidelberg, and in two greatly enlarged editions 1827 and 1830) as well as *Elements on the Philosophy of Right (PR)* (1821). His lectures on Esthetics, the History of Philosophy, and the Philosophy of History and Religion were attended not only by students but also by members of the Berlin society hungry for education. Until the mid-1940s, philosophy in Germany could not be exercised without taking a position on Hegel's philosophy. Moreover, his philosophy influenced the formation of those special sciences, which Dilthey later called '*Geisteswissenschaften*.' Soon after Hegel's unexpected death, his school split into different wings as a consequence of philosophical disputes on religion. In 1837, David Friedrich Strauss classified this splitting into the left (e.g., Karl Marx), the right (e.g., Friedrich Wilhelm Hinrichs), and the central (e.g., Karl Rosenkranz) (*see* Jaeschke, 2010; Pinkard, 2001).

Aspects of His Philosophy

Interests of the Young Hegel

The young Hegel studied the ancient authors, the literature of the Enlightenment, and above all Kant's critical philosophy, Rousseau's works on cultural criticism and theory of politics, as well as the dispute on Spinoza, which had been initiated by F.H. Jacobi. Furthermore, he enthusiastically paid close attention to the course of the French Revolution. At that time, he mainly worked on three questions. Firstly, how must a religious education of a people be conceived for it to strengthen the motivation for ethical and political action? This question had strongly suggested itself after Rousseau had failed to come to a convincing result concerning the question: which type of religion can be legitimated from the political point of view of a rational community? In his works on religion and ethics, Kant had then shown that a conflict is avoidable between religious and ethical duties on the one hand and religious and political duties on the other. Therefore, what mattered was to translate Kant's idea of such a union of all those duties into an educational program. Hegel, Hoelderlin, and Schelling devoted themselves to finding a corresponding way of life, which they called 'invisible church.'

Related to these interests is the second, historical question: what explains the fact that Christian faith developed into a 'positive,' authoritarian religion, which became a tool for

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61047-4

despotism to misuse? It was the strategic reaction of the established scientific community to the French Revolution and Kant's philosophy that gave rise to the third question; in the mid-1990s, the theological Orthodoxy of Tübingen began to misuse Kantian philosophy in an attempt to justify antiliberal church dogmas in a new manner. This led to the third question: how can Kant's critical philosophy be protected argumentatively from such a dogmatic misuse? From Hegel's penetrating critique, first of Kant's doctrine of postulates in the *Critique of Practical Reason*, emerged stepwise a system of philosophy whose position may be called 'speculative idealism.' But it did not replace his original interests; it was designed to answer them.

Phenomenology of Spirit

The PS exhibits the 'path of despair,' which 'natural consciousness' has to go in order to try to justify its knowledge claims under the conditions set forth by the contrast constitutive of itself: its own subjectivity – made into a first principle – in opposition to the objectivity not reducible to it. On the way, it becomes clear that the knowledge intended by various shapes of natural consciousness does not satisfy its own, immanent truth criteria. For example, the first shape of natural consciousness, called sense certainty, experiences its inability to refer immediately, i.e., without concepts, to supposedly *ontic* independent singulars by means of indexicals alone. This self-examination is the first step of the procedure. In a second step, '*we*,' who contemplate consciousness phenomenologically, have to identify this negative result – taken positively – with a new shape of consciousness. Out of the perspective of consciousness its process of self-examination will terminate only if it has produced a knowledge claim that fulfills its own truth criteria. Hegel calls the ultimate shape 'absolute knowing,' which is detached (Latin *ab-solvere*) from the contrast of subjectivity and objectivity constitutive of consciousness as such, so that it is not longer a shape of consciousness. Therefore, it presents the decisive step of the 'self-consummating skepticism' (Section 78) the PS exhibits – which may be summarized with Socrates' "I (a conscious experiencer) know that I know nothing."

Apart from the much discussed chapter 'Mastery and Servitude' – which exhibits how two people self-consciously struggling for recognition end up in the unstable – for recognitionless – shape of intersubjectivity of mastery and servitude – the chapters on practical reason, spirit, and religion of the PS should be of some interest to the social scientist. The chapter on Reason exhibits that any effort of self-actualization – i.e., gaining pleasure, making inner laws universally valid, and spreading virtue – must necessarily fail. Even a self-conscious practical reason cannot reach its goal of self-actualization if it merely tries to advance the concealed good of the world instead of rearranging the world in the first place.

The chapter on Spirit contains the doctrine of the basic convictions of a society concerning the purpose of society or its central elements. The internal problems of occidental societies – exhibited in their historical order – transformed them into one another, so that – in them – the Spirit reached new basic convictions about itself. The relevant problems, however, are not those that an objective observer of a later stage in history may discover, but those that the members of the societies in question could experience themselves as contemporaries.

In the chapter on *Religion*, Hegel carries out a similar analysis for the religions in history. As Max Weber after him, Hegel believed that the religions above all – *qua* their internally developed convictions – promoted a process of rationalization in world history, a process that Hegel comprehended as an advance in the consciousness of freedom. For him, Christianity was the religion that first gave effect to the principle of subjectivity and thereby pushed through the consciousness that *all* men as men are free. At the end of the PS, it is supposedly possible to expose a concept of cognition based on the historical experiences that natural consciousness has made in the course of history of occidental societies and religions. This concept, which Hegel called 'absolute knowledge,' has the process of these experiences as its contents and the type of rules concerning the transformation of the shapes of consciousness as its form. It is the purpose of this knowledge to promote humanity, whose existence 'lies only in the commonality of consciousness' brought about by the process of driving men forward 'to agreement with one another' (*see* PS, Section 69).

The knowledge arrived at in the PS, however, so far only concerns the historical genesis of the present as an epoch that still has to bring about humanity. In order to satisfy this demand, knowledge of what emphatically is has to be acquired, too (Preface to the PR). Since the PS has shown that certain fundamental concepts are the central constituents of each shape of consciousness, Hegel conceives of a doctrine of such 'determinations of thinking' in his *Science of Logic (SL)*: it has to examine whether these determinations are true for themselves.

Science of Logic

To logically unfold the socratic result of the PS is what Hegel declared to be the only 'proper metaphysics.' It is its task to develop an adequate concept of philosophical cognition and thereby of itself ('absolute Idea'). To fulfill this task, it primarily criticizes the aims at knowledge of pre-Kantian metaphysics and Kant's transcendental philosophy. In the SL, which is this metaphysics, the traditional concepts of Metaphysics and Logic incapable of defining a concept of the Absolute are improved in an iterated process: abstract understanding begins with trying to determine the traditional concepts in a way that makes them as distinct as possible from other concepts and hence suitable to define the Absolute. In a second step, dialectical reasoning demonstrates that no concept, however well determined by understanding, can be distinguished even from its opposite, and demands to skeptically suspend any judgment in the face of this negative result. But then, speculative reason identifies this negative result with a new term of the history of philosophy, which the understanding tries to determine again. Since the results of the first two steps are included in the new concept, the concepts of the SL become richer in content in the process. Once dialectical reason cannot even find an opposite concept anymore, the process reaches its end. Hegel calls this ultimate concept 'absolute Idea,' which is at the same time the concept of the philosophical cognition carried out in the SL (Fulda, 2003).

In the final step, this concept is arrived at by overcoming the dualism of our theoretical epistemic attitudes ('Idea of Truth')

and our practical ones ('Idea of the Good'). In these attitudes of finite cognition – paradigmatically conceived of by Kant as theoretical and practical reason – we regard the world not only as being a well-structured touchstone of our knowledge but also as insignificant and worthless in the face of the Good, which lies purportedly only in the will of the perceiver. The concept of the absolute idea is intended to free us from those conflicting references to the world. But in order not to give rise to a new conflict between the finite knowledge (prescientific and of the special sciences) and philosophical cognition, Hegel requires translation rules mediating between both types of attitudes and obliged philosophical cognition of the Real to demonstrate the correspondence of its results to 'empirical appearances' (Wildenauer, 2004). According to Hegel, a Philosophy of the Real depends therefore on integrating new expressions from other sciences to provide for a denotation of the philosophically developed concepts. To that extent, recognizing the Real is underdetermined by the SL, and a Philosophy of the Real as well as Logic itself depends on the developing status quo of the sciences.

Philosophy of the Real

The *Encyclopedia of the Philosophical Science in Outline* was meant to instruct the students who attended Hegel's lectures, and therefore only contains the rudiments and basic concepts of all philosophical sciences. The logic outlined above makes up the first part of the encyclopedia, the second and third parts being the Philosophy of Nature and Spirit. It is primarily through these two parts that we have access to Hegel's Philosophy of the Real.

Philosophy of Nature

In his Philosophy of Nature, Hegel tries to answer the question of how we can overcome the duality of our theoretical and practical interests concerning Nature and thereby satisfy our needs without turning Nature – and thereby aspects of ourselves – into the Other per se, which would merely become controlled by technology. This philosophy looks for a third way between the modern objectification and quantification of phenomena in the natural sciences on the one hand and a romantic subjectification of Nature on the other. In doing so it does not, however, regard itself as competing with the natural sciences, but rather tries – on the basis of the results of scientific research – to conceive of Nature as an idea in which we can finally recognize ourselves.

Philosophy of Spirit

In so far as Hegel gains his concept of Spirit from his doctrine of living natural beings, which contrasts the sphere of Spirit from that of Nature, he tries to overcome Descartes' dualism between *res cogitans* and *res extensa* in his whole Philosophy of Spirit, without, however, reducing the phenomena of one sphere to those of the other.

In particular, Hegel teaches in his *Doctrine of the Subjective Spirit* how the subjective spirit of individuals rises step by step from its embodiment (Anthropology) and its object-boundedness (Phenomenology) toward higher cognitive and voluntative competencies, which culminate in the knowledge of and will to its own freedom (Psychology). In the first part, the *Anthropology*, the phenomena to be comprehended are, e.g., dreaming, habits, aging, gestures, and mental disorders, while the *Phenomenology* treats the different levels of consciousness of objects. Sensory consciousness, perceptual consciousness, and comprehending consciousness are all subsumed under the heading 'consciousness,' since these forms of consciousness do not imply any knowledge the subject has of itself. Under the heading 'self-consciousness' are reviewed those forms into which some knowledge of the subject of consciousness of itself necessarily enters (desire and social recognition ('*Anerkennung*')). Finally, 'reason' is the consciousness of objects, in which it is comprehended that knowledge of general objects is only possible when the subject understands that the functions of his subjectivity constitute these objects in the first place. In conclusion, the higher cognitive performances (intuition, recollection, imagination, memory, and thinking) and voluntative ones (practical emotions, instincts, arbitrary acts, and striving for happiness) are discussed in *Psychology*. Within the 'free spirit,' the highest form of the subjective spirit, the attitude is reached and comprehended in which an individual knows of itself that it is only free and can only realize its freedom when it has overcome the dualism of interests that are merely of a theoretical or practical nature.

The *Doctrine of the Objective Spirit*, which Hegel has also worked out as *Elements of the* PR, comprehends the objective institutions in themselves, which men who know of their freedom give themselves – right in a narrower sense (property, contract, punishment of felonies), moral norms of acting individuals who strive toward the individual or collective good ('*Moralität*'), and the forms of social life in family, civil society, state, communities of states, and world history ('*Sittlichkeit*'). In order to accomplish this project, Hegel did not try to set up as large a set of consistent principles as possible, but instead conceived of concepts adequate to the subjects in question. The adequacy of the concepts is to be achieved by penetrating the fundamental conceptual material of effective norms and existing convictions of common sense and thus correcting the concepts.

Hegel's analysis of civil society has been especially influential (*see* Marx, Karl (1818–83)). According to this analysis, modern society tends to accumulate great wealth in the hands of a few, alienate workers in the working processes from their products and from themselves, press large parts of the population below the subsistence level, undergo times of over- and underproduction, form global markets, and colonize other countries. But at the same time, civil society implements institutions to administer justice, provide for the common welfare, and unite people in trade associations and municipalities. These institutions of society and others of the state make comprehensible how civil societies can develop and stabilize themselves to some, although not sufficient, degree in spite of their inherent tendency to crises. In his doctrine of state, Hegel tried to comprehend institutions that represent the general interests of a rational community and enforce them in the face of the special needs of estates and individuals of the civil society. So the state becomes the sphere of freedom that alone can guarantee the 'concrete freedom' of the individual of self-determination.

According to Hegel, world history, the doctrine of which makes up the final part of the PR, is constituted by the state. The principle organizing world history is the 'world-spirit,' the

process of the developing consciousness that *all* men as such are free. In this history, several states have made – to Hegel's mind – decisive contributions to this progress at certain times. In these periods, Hegel thought, they have the right – and indeed, the power – to politically and organizationally establish and expand a new principle. France was an example of Hegel's time: a country that had carried out a progress in the consciousness of freedom in its revolution and had expanded this new order through the Napoleonic Wars in great parts of Europe. But since the doctrine of the objective Spirit did not reach the highest form conceivable of freedom, Hegel's philosophy does not end with it, but rather with the doctrine of the absolute spirit.

Hegel's *Doctrine of the Absolute Spirit* is a doctrine of the three leading forms of spirituality in which men have at all times tried to understand what is significantly true and who they are within the whole of actuality that depends on such a 'Truth.' According to Hegel, these forms are art, religion, and philosophy. However, they have not always had the same importance since the time they came into being. Each form rather has its particular period in which it becomes the model form to exhibit the Truth. Hegel believed that for the ancient Greeks, art was the medium through which they best understood themselves and were able to express what they cherished most. For the Christians, religion was the best form for men to represent their relation to the Truth and the Truth itself. Both forms, however, have the drawback that their means of explication are at most somewhat discursive – or, indeed, not at all. It is the purpose of philosophy as a science to conceive the content of Christianity's true representations in an argumentative discourse and thereby to gain insight into the deepest reasons for which men come to agreement with each other and afford a stable basis for their life with each other. Therefore, Section Philosophy of the Real closes with a concept of philosophy, which is also to justify in retrospect the recognition carried out in the system of the philosophical sciences.

Some Lines of Reception in the Social Sciences

Lorenz von Stein (1815–90)

Lorenz von Stein was one of the most politically influential Hegelians among the theoreticians of society in the nineteenth century. With his diverse writings on modern society, he influenced central economic terms of Marx's theory (e.g., surplus value and accumulation) on the one hand, and considerably influenced Bismarck's social policy on the other. Up to 1933 he was studied by theorists of public and constitutional law (e.g., Georg Jellinek and Hermann Heller), a discussion that was continued in the German debates on the principle of the welfare state after the Second World War (Stein, 1972). Stein took up the question posed in Hegel's doctrine of '*Sittlichkeit*': how must a state be structured for the principles of the French Revolution to be realized considering the conditions of the industrializing society? He regarded the political constellation of his times as an unavoidable antagonism between the civil society based on the satisfaction of individual needs and the state aiming at general freedom. To his mind, this antagonism leads the propertied to try to occupy all state functions through unequal elections, in order to prevent the socioeconomical rise of the nonpropertied. Therefore, Stein thinks it to be the duty of a legitimate state and its institutions (monarch, officials, army) to implement social reforms resulting in a rise of the nonpropertied to the economical level necessary for them to participate in the process of forming a public will. In addition to this, the state should counteract the inequalities of the civil society by introducing universal and equal suffrage and guaranteeing basic rights. To this extent, Stein's doctrine of the state – like Hegel's – aims at more than just the guarantees of the liberal constitutional state, because it does not confine itself to the task of protecting the freedom of the individual from the state, but instead goes further to protecting the political freedom of the individual against the inequalities of the civil society through the state.

Karl Marx (1818–83)

In 1836, Karl Marx arrived in Berlin as a student 5 years after Hegel's death and had contact with members of the left wing of the Hegelians (*see* Marx, Karl (1818–83)). As Lorenz von Stein Marx started his reception of Hegel's PR by receiving and developing Hegel's arguments for a general and equal electoral law as a solution to the growing social question of his times (MEW, Bd. 1: pp. 320–333). This is one of the reasons why he confines his 'Critique of Hegel's Philosophy of Right' (1843) to a critique of Hegel's philosophy of state, which might have found some sympathies by Hegel as the latter had even publicly stated in one of his lectures about the philosophy of history, that there "still subsists a main one-sidedness, namely, that the *volonté générale* has to become also the *empirically* general will, i.e., that the individuals as such should reign or take part in reigning. […] It is this collision, this knot, this problem at which history stands [now] and which history has to solve in future times" (own translation: 12/534). According to Engels, Marx incorporated into his own Political Economy without any substantial critique Hegel's doctrine of the inherent instability of the development of a civil society if it is dominated only by pure laws of economics (MEW, Bd. 16: p. 326) as well as Hegel's dialectical method – cleansed from an allegedly Hegelian mysticism – into his own Historical Materialism (MEW, Bd. 40: S. 568–588; Bd. 21: S. 292; Bd. 23: S. 27 f.). Exemplarily for the first point, compare Marx's positive reception of Hegel's claim, that work contracts are illegitimate, if the employer demands of the employee all of his time of productivity, because in such a case the employee becomes the property of somebody else (PR Section 67; MEW, Bd. 23: footnote 164). But probably because of the factual consequences of the legislation concerning the poor in England and Germany as well as the political developments in Prussia since 1840, Marx underestimates the stabilizing potential of a combination of democracy and welfare state, which Hegel hinted at and Lorenz von Stein developed some steps forward. Exemplarily for this assessment, Marx seemed to have overlooked article 21 of the French Constitution of 1793 as well as Hegel's morally grounded and therefore overpositive, and also juridical right for a satisfaction of basic needs in a modern society (Enc.: Section 533). These might be some of

the reasons why Marx's ideas about a postrevolutionary society leave the field of Hegel's philosophy.

St Louis Hegelians

This American philosophical school was founded by William Torrey Harris (1835–1909) and Hans Conrad Brokmeyer (1828–1906). After the Civil War, this school was institutionalized by the foundation of the *St Louis Philosophical Society* (1866) and the *Journal of Speculative Philosophy* (1867). In contrast to the Transcendentalists at Yale and many of the German intellectuals, questions concerning religion did not play an important role for the St Louis Hegelians. Instead, questions concerning social justice in a modern society were central for them. In addition to this, another one of Hegel's basic interests became important again, on which, however, Hegel had only remarked in passing in his systematic works: how must pedagogy be in order to make the individuals growing up be able to competently participate in forming a free society? Besides being influenced by Hegel, Denton Snider (1841–1925) and Thomas Davidson (1840–1900), two St Louis Hegelians, were also influenced by the pedagogy of the German Friedrich Froebel (1782–1852) when they realized several educational projects: the *Communal University* in Chicago, the *Chicago Kindergarten College*, the *Goethe School* in Milwaukee, and the *Breadwinner's College* in New York City, which was devoted to the education of the working class. In 1889, Harris became the United States Commissioner of Education. Brokmeyer played a key role in the Constitutional Convention of Missouri in 1875, which guaranteed the right to education for all aged 6–20. Besides these important roles of the St Louis Hegelians in the formation of the modern educational system of the United States, the *Journal of Speculative Philosophy* offered a platform for the publications of important American philosophers such as John Dewey, William James, Charles S. Peirce, and Josiah Royce. Beyond offering this platform, they successfully animated the study of Hegel's philosophy (Flower and Murphy, 1977).

Wilhelm Wundt (1832–1920)

Wundt is commonly regarded as the founding father of Psychology, which established itself as a science around the beginning of the twentieth century. In 1879, Wundt founded the first psychological laboratory of the world in Leipzig, Germany, where he mainly studied sensations and feelings by employing experimental methods. This foundation also became influential outside Germany due to the many visits of foreigners – especially American students and scientists. Although Wundt often mocks the methodology of Hegel's philosophy, he explicitly praises Hegel's achievement of "denying the right to exist to all 'things in themselves,' whether they are to count as metaphysical borderline concepts or as practical postulates" (Wundt, 1911: p. 736). Without this Hegelian basis, it would have been impossible for Wundt to pursue psychology as an empirical special science, because the soul would have been withdrawn from the applicability of methods borrowed from the natural sciences, as long as it were conceived of as an independent substance removable from the body (Descartes) or were postulated as such on practical grounds (Kant). Hegel, on the other hand, "has regarded the individual soul as the immediate relationship of the experiences of the soul, as a piece of that infinite reality of the world spirit, which still draws only on its very own actual reality to obtain all that it means" (Wundt, 1911: p. 736). So Wundt arrives at the conclusion that Hegel "had also declared the liberty of psychology when it proclaimed for all that all mental coming into being – and, therefore, all happening in the soul as well – is actuality, is immediate experienced reality, and that essence and appearance of the Spirit are one and the same, and only mean something different in so far as we comprehend the essence as the relationship of the appearances which is correctly recognized" (Wundt, 1911: p. 737; own translations).

George Herbert Mead (1863–1931)

The posthumously published book *Philosophy of the Present* (1959) primarily promoted an interpretation of Mead's thinking that considers the last phase of it as on the way toward a speculative metaphysics of Hegel's kind. This is surely exaggerated. It is, nevertheless, possible to prove that Mead named Hegel as an inspirer of his own thoughts throughout his working period. Hegel's philosophy especially stimulated the scientific program of the young Mead. His early outline of a program of philosophical disciplines puts forward two main tasks of psychology, one of which was inspired by Hegel's PS: psychology is to give "an analysis of the objective world in terms of the consciousness of the objective individual" (Mead, 1900: p. 12). In pursuing this task, Hegel – according to Mead's view – e.g., proved that the organization of different characters, which constitute 'social objects,' comes from 'our mutual relations' (Mead, 1903: p. 106). In contrast to the explaining and describing psychology, logic is a normative science (Mead, 1903: p. 107), into which 'the general theory of the intelligent act as a whole would fall,' and which is supposed to be as such of Hegel's kind (Mead, 1900: p. 2). However, apart from this praise, Mead also continued to criticize Hegel's doctrines throughout his life: e.g., he falsely subsumed Hegel under the philosophy of romanticism, which, to his mind, only discussed the rationalization of bygone attitudes – quite contrary to the orientation toward the future of the American society and philosophy. Nevertheless, Mead praises Hegel's philosophy of history as "[t]he most grandiose attempt to combine the three great motives of speculation (that of Greek contemplation, that of the church's plan of salvation, and that of experimental science)" (Mead, 1964: p. 505).

Juergen Habermas (1929–)

In his essay *Can Complex Societies Form a Rational Identity?* (Habermas, 1976), which Habermas chose as the basis of his speech on the occasion of accepting the Hegel Award of the city of Stuttgart, Germany, in 1974, he explicitly classifies his proposal as standing in the tradition of Hegel's and Marx's thinking (p. 118). Although his interpretation of Hegel's answer to the question posed in the title of his essay is affirmative in many points, Habermas mainly criticizes – firstly – that for Hegel, 'only the idea of freedom' is to be conceived as a constituent of modern consciousness. In addition to this, 'the unrestricted objectifying thinking and a radical orientation toward the future' are also constitutive for it

(p. 105). Therefore, Hegel did not succeed in abolishing the threefold division between the modern subjectivity and external nature, society, and inner nature (p. 102) in all respects. Habermas objects, second, that modern society did not gain its rational identity in the constitutional state; he objects, third, that philosophy is unable to exhibit this identity as rational for all individuals – as religion has done in the past (p. 106 ff.). It is Habermas' opinion instead that modern societies could gain their rational identity in "the consciousness of the general and equal opportunities of participation in learning processes of forming values and norms" (p. 119) by realizing communicative structures 'by themselves,' in whose 'autonomy' alone a new collective identity can crystallize 'today' (p. 120; own translations). Critical remembrance of tradition, science, philosophy, and art propose rival identity projections (p. 121). However, later on, Habermas seems to have partially taken back his criticism of Hegel with his concept of 'constitutional patriotism' (Habermas, 1990). In addition to this, his interpretation of one of the rational functions of Right, namely, to relieve us from moral discourses (Habermas, 1996), can be understood as a continuation of Hegel's thoughts on the relationship of '*Moralität*' and '*Sittlichkeit*.' In 2008, Habermas has described his own philosophical development as 'my way from Hegel-Marxism to Kantian Pragmatism' (p. 78).

See also: Civil Society and the Public Sphere: History of the Concept; Civil Society, Concept and History of; Hall, Granville Stanley (1844–1924); Kant, Immanuel (1724–1804); Marx, Karl (1818–83); Phenomenology: Philosophical Aspects; Rousseau, Jean-Jacques (1712–78); Weber, Max (1864–1920); Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Flower, E., Murphy, M.G., 1977. The absolute immigrates to America: the St. Louis Hegelians. In: A History of Philosophy in America, vol. 2. Putnam, New York.

Fulda, H.F., 2003. G.W.F. Hegel. Beck, München.

Habermas, J., 1976. Können komplexe Gesellschaften eine vernünftige Identität ausbilden? In: Zur Rekonstruktion des Historischen Materialismus. Suhrkamp, Frankfurt a. M., Germany.

Habermas, J., 1990. Die Nachholende Revolution. Suhrkamp, Frankfurt a.M., Germany.

Habermas, J., 1996. Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy. Polity Press, Cambridge, UK.

Habermas, J., 2008. Ach, Europa. Suhrkamp, Frankfurt a.M., Germany.

Hegel, 1968ff. Gesammelte Werke. Hrsg. In: Verbindung mit der Deutschen Forschungsgemeinschaft von der Nordrhein-Westfälischen Akademie der Wissenschaften. Felix Meiner, Hamburg, Germany.

Hegel, G.W.F., 1969a. Science of Logic. (W.H. Johnston, Trans. (int. al.)). Allen and Unwin, London.

Hegel, G.W.F., 1969bff. Werke. In: Moldenhauer, K.M., Michel (Eds.), 20 Bände und ein Registerband. Suhrkamp, Frankfurt a.M., Germany.

Hegel, G.W.F., 1979. Phenomenology of Spirit. (A.V. Miller, Trans. and introduced by J.N. Findlay). Clarendon Press, Oxford.

Hegel, G.W.F., 1983ff. Vorlesungen: Ausgewählte Nachschriften und Manuskripte. Patent, Hamburg, Germany.

Hegel, G.W.F., 1988. Aesthetics. Lectures on Fine Art. (T.M. Knox, Trans.). Clarendon Press, Oxford.

Hegel, G.W.F., 1990. In: Behler, A.W. (Ed.), Encyclopedia of the Philosophical Science in Outline. Continuum, New York.

Hegel, G.W.F., 1991. In: Wood (Ed.), Elements on the Philosophy of Right. Cambridge University Press, Cambridge.

Hegel, G.W.F., 1995a. Lectures on the History of Philosophy. (E.S. Haldane, Trans.). University of Nebraska Press, Lincoln, NE.

Hegel, G.W.F., 1995bff. In: Hodgson, G.H. (Ed.), Lectures on the Philosophy of Religion. University of California Press, Berkeley, CA.

Jaeschke, W., 2010. Hegel-Handbuch. Leben-Werk-Schule. Metzler, Stuttgart und Weimar. Second Edition.

Marx, K., Engels, F., 1956ff. Marx-Engels-Werke (MEW), 43 Bände. Ost-Berlin, Dietz Verlag.

Mead, G.H., 1900. Suggestions toward a theory of the philosophical disciplines. Philosophical Review 9, 1–17.

Mead, G.H., 1903. The Definition of the Psychical. In: First Series, vol. III. Decennial Publications of the University of Chicago, Chicago, pp. 77–112.

Mead, G.H., 1959. In: Murphy, A.E. (Ed.), The Philosophy of the Present. Open Court Publishers, La Salle.

Mead, G.H., 1964. In: Morris, C.W. (Ed.), Philosophy of the Act. The University of Chicago Press, Chicago.

Pinkard, T., 2001. Hegel. A Biography. Cambridge University Press, Cambridge.

Stein, v.L., 1972. In: Forsthoff, E. (Ed.), Gesellschaft-Staat-Recht. Propyläen Verlag, Frankfurt a.M, Germany.

Wildenauer, M., 2004. Epistemologie freien Denkens. Die logische Idee in Hegels Philosophie des endlichen Geistes. Felix Meiner, Hamburg.

Wundt, W., 1911. Grundzüge der physiologischen Psychologie. Band 3. Leipzig (Principles of Physiological Psychology. In: Titchener, E.B. (Ed.), Sonnenschein and Macmillan, London 1904).

Hegemony and Cultural Resistance

Sujeong Kim, University of California, San Diego, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 6645–6650, © 2001, Elsevier Ltd.

Abstract

Gramsci's concept of 'hegemony' has provided one of the most influential theoretical frameworks in contemporary cultural and media studies, and Marxist intellectual work in particular. Stated briefly, hegemony refers to a historical process in which the dominant group exercises 'moral and intellectual leadership' throughout society by winning the 'consent' from people. This concept has enabled cultural and media studies to explain not only the way in which the domination of a power group in modern capitalist societies is secured, but also how the dominant order can be challenged and contested through popular resistance directed in various cultural forms and practices. The applications of the concept of hegemony in cultural and media studies can be addressed in four distinct areas: the analysis of historical moments, the analysis of media texts, the analysis of subculture/popular culture, the analysis of audiences.

The concept of 'hegemony' has provided one of the most influential theoretical frameworks in contemporary cultural and media studies, and Marxist intellectual work in particular. While imprisoned by Mussolini's Fascist regime in the 1920s and 1930s, the Italian Marxist Antonio Gramsci developed the concept of hegemony to address the question of why the working class failed to make a communist revolution in the advanced capitalist countries despite Marx's prophecy. By 'hegemony', Gramsci (1971) intended to conceptualize the pivotal role of cultural persuasion in acquiring, exercising, maintaining, or challenging power without resorting to physical coercion. Stated briefly, the concept of hegemony refers to a historical process in which a dominant group exercises 'moral and intellectual leadership' throughout society by winning the voluntary 'consent' of popular masses.

The Concept of Hegemony

The concept of hegemony can be more effectively understood when it is compared to Gramsci's notion of 'domination,' as each term represents a different kind of power. Whereas domination rests on a dominant group's coercion of the less powerful, hegemony relies on a dominant group's ability to procure subordinate groups' consent to the existing social order. Domination belongs to the state, which Gramsci called 'political society,' and is equipped with instruments of coercion such as the police and armed forces. In contrast, hegemony takes place in the realm of private institutions – including those outside of the spheres of economic production and the state, such as families, churches, trade unions, and the media – which Gramsci called 'civil society.' Gramsci attributed the political stability and obedience of the working class in Western European countries to the hegemony built into the vigorous civil societies of those countries. Consent and coercion always coexist, but typically, dominant groups in contemporary capitalist societies rule through hegemony unless severe social and political crises require directly coercive interventions.

The Concepts of Consensus and Historical Bloc

The essence of the concept of hegemony lies in the idea of subordinates' consent. According to Gramsci, the consent of subordinates is constructed when a dominant group's particular world view is accepted by the subordinate groups as their world view, and thus the dominant group's interests appear to be the interests of the society at large. As a result, the dominant group is seen as entitled to its ethical and intellectual leadership and arises as a hegemonic group in a particular historical period. Moral and political authority legitimated by popular consent characterizes hegemony as power from below, distinguished from power wielded by coercion or by manipulation from the top down.

However, the paradox of subordinates' consent to their subordination in the long run raises suspicions about the nature of popular consent (Lears, 1985). Indeed, some critics of the concept of hegemony have questioned how subordinates can give consent to what will adversely affect their interests unless they are self-deceived in false consciousness. The very idea of 'false consciousness' threatens to undermine the concept of hegemony in at least two ways. First, if one accepts the idea of false consciousness as the basis of popular consent, subordinates' consent is the outcome of sheer manipulation, and Gramsci might be accused of elitism for his implicit assumption that the popular masses are dupes. Alternatively, if one rejects the idea of false consciousness by arguing that people know the purpose of their actions and consistently base their actions on rationally determined self-interest, then the consensus may represent genuine or true consensus. Then, the characteristic of hegemony as dominance over subordinates would be invalidated, and their actual grievance or resistance may slip from view.

Both Jackson Lears (1985) and Stuart Hall (1996a) responded to this dilemma by rejecting the binary opposition of false and true consciousness in understanding the nature of consent. They instead embraced Gramsci's idea of 'contradictory consciousness' grounded in 'common sense.' According to Gramsci, common sense as everybody's spontaneously arrived at philosophy constitutes the popular masses' practical consciousness guiding their actions. However, 'common sense' is fragmented and contradictory; the human conception of the world is 'disjointed and episodic' rather than 'critical and coherent' (Gramsci, 1971, p. 324). Gramsci's conception of 'contradictory consciousness' as the condition of humanity makes it possible to understand that human action does not

 http://dx.doi.org/10.1016/B978-0-08-097086-8.95055-4

always result from fully conscious free choices based on rational calculations. Individuals' subtle subjective meanings and psychology also lead human action. So, without subscribing to the idea of false consciousness, one can understand popular consent in this broader picture of human beings whose consciousness may be 'contradictory,' but not 'false' (Lears, 1985; Hall, 1996a).

The formative process of hegemony, for Gramsci, also involves a 'historical bloc.' According to Gramsci, a historical bloc means an alliance of disparate groups or class fractions which is bound not only by shared economic interests but also by a particular world view, understood as ideology which cements varied classes and groups into a unity of collective will. An instance of the successful formation of a hegemonic historical bloc is found in the French Revolution of the eighteenth century. According to Gramsci, the French revolution was possible because the Jacobins were able to bring the French bourgeoisie class fractions and the great popular masses together under discourses of 'equality, fraternity, and liberty.' These discourses were not just the bourgeois class's language circulating only within its own circles, but also the popular masses' words reflecting the exigencies of their time in the French cultural tradition. By identifying the interests of the bourgeois class with the general interests of all the national forces in French society, the bourgeois class was able to form a historical bloc, and to exercise the intellectual and moral authority to lead the masses toward the revolutionary struggles. This case also makes clear that ideology plays a crucial role in building a historical bloc through both economic and spiritual solidarity.

In this vein, hegemony seems to represent the most effective and formidable means of exercising power. Steven Lukes (1974) argued that a significant layer of power is found in the ability of powerholders to affect and modify the desires and beliefs of the powerless to the extent that resistance from the powerless cannot easily take place. With regard to the concept of hegemony, this view of power highlights the cultural dimension of power and offers insights into the acquiescence of subordinates under hegemonic consensus (Gaventa, 1980). However, this view may unwittingly generate the impression that hegemony is such an unconquerable, overarching form of power that it can prevent subordinates' possible resistance in the first place by saturating their consciousness with hegemonic ideologies. An important question follows from this impression: 'Is popular resistance really possible, and if so, what makes it possible?' This question reflects Gramsci's foremost concern. In explicating the concept of hegemony, Gramsci constantly emphasized the instability of hegemony, and enthusiastically proposed the possibility that subordinate and popular groups may form 'counterhegemonies.' The idea of counterhegemony suggests Gramsci's vision of social change embodied in the concept of hegemony itself.

Hegemony, Cultural Resistance, and Social Change

The question of where power lies is closely related to an interest in the source of change and in the ways of making a difference in history for practical and political purpose (Lukes, 1974). Gramsci understood popular consent as the locus of, as well as a means of, subverting hegemony, and as a strategy for historical social change. Hegemony is not a complete state or system but a historical process which needs "continually to be renewed, recreated, defended" and is also "continually resisted, limited, altered, challenged by pressures" (Williams, 1977, p. 112).

What, then, may be the grounds for the possibility of alternative or counterhegemony? For Gramsci, common sense makes it possible for subordinates to come to occasional understandings of the conditions of their subordination and exploitation. By combining common sense with the coherent philosophy of the intellectual class, subordinates can constitute themselves as historical agents. Then, common sense turns out to be a strategically important terrain of ideological struggles. Since popular masses make sense of their world and themselves through common sense, various groups and classes compete over moral and intellectual leadership in this terrain.

The concept of a historical bloc also opens up the possibility of forming a national-popular alliance led by subordinate groups for an alternative or a counterhegemony. Discussions about a historical bloc in counterhegemony have been extended and elaborated by Ernesto Laclau and Chantal Mouffe (1985) and Stuart Hall (1996b). According to them, the principle underlying hegemony is 'articulation,' which refers to the practices of linking disparate groups, disparate ideologies, or various forces into a unity. In other words, a historical bloc is the very outcome of articulations.

Utilizing the concept of articulation, Laclau and Mouffe argue that there is no necessary, intrinsic tie between ideologies and classes, as assumed in such terms as, 'bourgeois ideology' or 'working class ideology.' Rather, certain ideologies come to serve certain groups' or classes' interests, and represent the world view of members of a class as a result of their efforts to articulate ideologies according to their own interests. In the French Revolution, for example, the Jacobins succeeded in articulating the ideology of 'liberty and equality' to the bourgeois class. Therefore, Laclau and Mouffe believe that subordinate groups such as women, people of color, ethnic minorities, and others can also enlist the ideology of liberty and equality to carry out a new, radical, democratic project in contemporary capitalism. This project implies that the working class is not exclusively responsible for the task of social and political transformation. That is, in contemporary modern society, relations of dominance and subordination do not hinge solely on the axis of class; the concept of a historical bloc as a strategic alliance can be applied to the strategies of 'all' classes and groups (Laclau and Mouffe, 1985; Hall, 1996b, p. 425). Moreover, individual human subjectivity and class identity are not a priori fixed and unified. Rather, individuals and groups negotiate and construct their own identities in the articulatory practices of hegemony.

Laclau and Mouffe and Hall have not so much diverged the theory of hegemony as they have returned to Gramsci's argument that people, as the ensemble of social relations and thus as contradictory subject, 'create their own personality' (Gramsci, 1971, p. 360). In the concept of hegemony, the popular masses' ideological struggles and cultural resistance to form a collective will against the existing hegemony constitute the essential base of social and political change.

The Implications for Theories of Culture

The concept of hegemony marks a groundbreaking departure from classical Marxist understandings of culture, which stand on the 'base-superstructure' model. According to this model, the economic foundation of society, called the 'base,' determines the social, political, cultural life processes called the 'superstructures.' Related to this model is the proposition that the dominant ideas in a given society are the ideas of the economic ruling class. Upon these propositions, the complexity of culture is conceived of as a pale reflection of the economic imperatives of capitalism, and as an expression of the interests of the ruling class. In this line of thinking, culture is understood in terms of 'ideology' which was identified with 'false consciousness.' Therefore, cultural theories in the Marxist tradition have been underdeveloped and often limited to critiques of ideology.

In this context, Gramsci's concept of hegemony has brought remarkable changes to the understanding of culture. In the concept of hegemony, culture takes on a fundamentally important, formative role. Culture is grasped as a complicated, contradictory, multilayered sphere in which ideas and values are materialized in the forms of varied institutions. With the concept of hegemony, Gramsci significantly reformulated the relations between base and superstructure as dialectical interactions rather than a one-way causal determination, and emphasized cultural struggles. Therefore, he could establish the understanding of the relative autonomy of culture from the economic base, and effectively overcame the classical Marxist tendency toward both economic and class reductionism of superstructures.

In this regard, Raymond Williams, an influential Marxist literary critic, convincingly argued that, by conceptualizing culture as 'a whole social process' and by connecting it to the power relations of domination and subordination, the concept of hegemony dialectically develops two powerful earlier concepts, 'culture' and 'ideology' (Williams, 1977, p. 108). The concept of counterhegemony also emphasizes the relative autonomy of politics from the base, and facilitates the recognition of nonelite classes as agents of historical and political change.

Applications of the Concept of Hegemony in Cultural and Media Studies

Cultural and media studies have applied the concept of hegemony to understand and analyze the roles and meanings of various cultural and media forms and practices. Several of these applications are briefly described in the following four areas: 'journalism,' 'textual analysis,' 'subcultures,' and 'audiences.'

The Analysis of Journalism

A seminal work relying on Gramsci's concept of hegemony emerged from British cultural studies. Stuart Hall et al. (1978) argued that the radical right in the 1970s was able to reintegrate the crisis of the social–democratic consensus into a new 'law and order' framework and smoothly usher in Thatcherism. Underlying the reconstructed popular consent was a moral panic against social crime like mugging. The moral panic, created through the work of the mass media as well as judiciary practices, was interwoven with racist bias, further mobilizing popular anxiety. This work demonstrates the ideological work of mass media as well as that of the state in recovering hegemony (see British Cultural Studies).

In the USA, Todd Gitlin (1980) took up the concept of hegemony as a theoretical framework to explain the relationship between the US news media and social movements in contest with broad popular consent in the 1960s. He illuminated how commercial news 'framed' the antiwar movement in ways that 'trivialized' and 'marginalized' it, and consequently how the New Left finally failed to control the public cultural space. According to Gitlin, news is a way of interpreting, organizing, and presenting the news, and it is a routine journalistic practice rooted in journalists' professional beliefs. Through this framing practice, the news media tend to implicitly define dissident social actors as illegitimate, and to deflect, absorb, and domesticate oppositional or conflicting ideas in ways that reinforce the hegemony. For instance, journalists reported on antiwar student leaders as they would report on celebrities, since this 'celebrity frame' met their need for spokespeople, based on the code of objectivity and concern for the newsworthiness of the story. Furthermore, this frame could also satisfy the commercial media's preference for the human interest angle. However, the impact of this frame on the student leaders was devastating. Subject to the glamor and pressures of popularity under this frame, the leaders became estranged from the movement base, and the movement itself gradually lost control of its own leaders. As a result, despite their original intention to use the media to reach the great popular masses, the antiwar movement failed to obtain popular consent (see Media and Social Movements).

If news framing is a routine practice, then how can news frames be changed? If commercial news frames maintain hegemony, as Gitlin concluded, one must look elsewhere for possible conditions for social change. Daniel C. Hallin's (1989, 1994) work on US coverage of the Vietnam war and of the conflicts in Central America offers valuable insights. By tracing shifts in news framing, Hallin identified dynamic relations between the news media and the dominant political group, which underlay the competition of different frames for a hegemonic presence. According to Hallin, depending on the degree of consensus among the political elite, news coverage of war took on different frames. Hallin stressed that journalistic routines relying on official sources (e.g., press conference) tend to constrain the media's consensus over the frame to the boundaries of the hegemonic view. But he also convincingly argued that when the hegemonic group reveals cracks in the political power bloc, the media become more critical and independent, and journalist routines break down. As a result, dissident views and groups can be framed more positively, until the hegemonic group, recovering its unity, actively engages in repairing the boundaries of general consensus in society. These analyses of journalistic routines and news framing successfully connect the microlevel sociology of journalism with the macrolevel politics of hegemony, and simultaneously suggest the possibility of social change.

Tamar Liebes (1997) shed light on another dimension of journalists' voluntary consent to the hegemonic social order. According to her work on Israeli media coverage of the Arab–Israeli conflict, Israeli journalists genuinely believe in the necessity of censorship for national security and, as a result, they actively compromise their ideals of free speech and professional autonomy by colluding with military censorship. Thus, journalistic 'self-censorship' of news routinely takes place before any forceful, formal official censorship actually occurs. Despite the fact that Israeli journalists self-consciously perform self-censorship, interestingly enough, they are rarely seen as either manipulators of public opinion or conspirators with hegemonic power. Rather, as Liebes showed, the journalists' consent to Zionism as a dominant ideology was formed and internalized throughout their life experiences in the Israeli culture and history; that is, the journalists are both part of the community and members of the elite group. The discussions of Gitlin, Hallin, and Liebes on the practices of journalism are distinguished from the instrumental view which considers the news media as a tool for the privilege group. For example, Edward Herman and Noam Chomsky (1988) argued that the media are a propaganda system which is controlled and directly manipulated by the elite so as to benefit the powerful and the state. Their approach ignores the decentered hegemonic process which Gitlin, Hallin, and Liebes argued by revealing the dimension of journalist acculturation as well as socialization in news organization.

The Analysis of Texts

The concept of hegemony has influenced the theories and practices of 'textual analysis' by conceptualizing texts as site of struggles over meaning, not as means of inculcating dominant ideologies. Textual analysis in cultural studies has been applied to various cultural forms and practices, in particular film and television in media studies. It has significantly contributed to the understanding of ways in which dominant meanings are encoded in cultural products so that readers take the established social order for granted. In many textual analyses, however, audiences were often assumed to be passive and vulnerable to dominant ideologies; texts, especially the mass media, were easily credited with powerful ideological effects. Many textual analyses tend to read the ruling class's or state's successful hegemonic process directly from the text, to privilege the researchers' perspective, or to easily assume subordinates' consent (Condit, 1994). These limitations may become serious in cases where rearchers endorse language-centrism to the extent that only language and its structure determine human subjects and power struggles, apart from texts' meanings in peoples' real social life process which Gramsci relentlessly emphasized.

Studies of Subculture and Popular Culture

The concept of hegemony provided a key theoretical framework to British subcultural studies in the 1970s. Subcultural studies inquired into the everyday life of the young working class and the meanings of their cultural consumption in a framework of power inequality (Hall and Jefferson, 1976; Hebdige, 1979). The researchers were primarily concerned with the styles of youth subcultures in the UK such as mods, rockers, teddy boys, and punks. According to them, these teenagers actively create their styles by appropriating various symbols out of consumerism to construct and express their identity and to resist the hegemonic social order. In this view, styles are not only sites for negotiating meanings but also forms of cultural resistance built in lived experience, although not all of their responses prove potentially counter-hegemonic. This body of work on youth cultures takes popular culture seriously as the terrain of symbolic struggle, which is not necessarily bound to the economic class struggle (see British Cultural Studies).

Research on long-term consequences of youth subculture on power relations by Paul Willis (1977) fascinatingly demonstrated how working-class boys' counterschool culture prepares them for reproducing themselves as laborers. Although Willis did not employ Gramscian terminology, this work may deepen the understanding of hegemony. In this ethnographic work, Willis observed that the schoolboys' resistant practices, intertwined with the dominant values of masculinity as patriarchal ideology, ultimately produced their concession to the existing power relations. In such a case, any simple dichotomy of 'resistance' and 'hegemony' becomes problematized, and the line between them blurred. Willis showed that hegemonic consensus contains very contradictory and uneven processes with unintended consequences, which refutes any necessary correspondences between consciousness, practices, and material conditions.

Recent applications of the concept of hegemony to subculture and popular culture seem to be located in a broader cultural matrix than the earlier British subcultural studies embedded in the context of 'English culture.' Focusing on race, gender, ethnicity, class, age, and other axes of cultural subordination, studies of subculture and popular culture have examined the complicated process of cultural politics. For example, rap music is not simply a music for the young. It is an important form of young black cultural resistive expressions in the USA. Black urban male rap artists strenuously attempt to delegitimate powerful social order by criticizing racism and police brutality on blacks; but at the same time their lyrics often reflect sexist, homophobic, or anti-Semitic ideas (Rose, 1994). Cultural politics hardly proceed without contradiction.

The Analysis of Audience

After the influential Marxist linguist Mikhail Bakhtin argued compellingly that language innately bears a multiplicity of voice in conversation, the understanding of the monolithic and predetermined ideological role of texts has come under criticism. Following this, abundant interest in audiences' interpretations of texts has flourished in cultural and media studies. With regard to the concept of hegemony, this interest seems promising to provide substantial empirical evidence about popular consent, culled from audiences' actual interpretations of symbolic work.

A theoretically significant model sparking the emergence of critical audience studies was proposed by Stuart Hall's (1980) 'Encoding/decoding.' He suggested three decoding positions, i.e., 'dominant-hegemonic,' 'negotiated,' and 'oppositional' positions in which the accommodation of hegemony and popular struggles over meaning can be simultaneously considered. These reading positions help to identify whether audiences read a text according to the way that the message is 'dominantly' encoded, or in an 'oppositional' way although they would understand the dominant meaning preferred by the producers of the message; or, they generally accept the dominant meaning of the text but partially read against it in a 'negotiated' way of reflecting their own interests (see British Cultural Studies; Audiences, Media). Building on this model, David Morley (1980) launched a landmark work on audience reception of a British news magazine television program. In this work, Morley, who was primarily interested in workingclass audience interpretation of the program, concluded that neither text nor audiences' social class directly determine audience readings. After his work, audience research proliferated and has provided evidence that audiences are capable of interpreting texts oppositionally or alternatively based on their own social experiences, and of using texts for their own purpose.

The shift of focus from the text to the audience and from ideology to cultural resistance runs parallel to increasing interest in counterhegemony in the context of everyday life, although some research has been criticized for its tendency to celebrate audience autonomy without evidencing any substantial social consequences (see Audiences, Media).

Cultural and media studies since the 1980s tend to be theoretically more synthetic, including Gramsci's concept of hegemony together with other postmodern theories, and have drawn more attention to issues of identity politics. Noticeably, scholars outside of the Anglo-American context have increased their presence in English-speaking forums, and have enriched cultural and media studies. For example, deeply engaged with the concept of hegemony, Néstor García Canclini (1995) and Jesús Martín-Barbero (1993) have offered valuable elaboration on the development and complexity of popular culture, and on the active struggles for the formation of new cultural identities in Latin America. Case studies of Mexican film, Brazilian popular music, and the *telenovela* illuminate how the melodramatic and cultural imaginations of the popular classes in everyday life are articulated popular with cultural memories as a way to form new cultural identities (Martín-Barbero, 1993).

Gramsci's concept of hegemony has been continually reinterpreted and applied to different areas of cultural and media studies, depending on the immediate social and cultural questions of theorists in specific social and historical contexts. However, all research relying on the concept of hegemony shares a common pursuit in criticizing culture as a source of control, and, simultaneously, in envisioning it as a source of resistance.

Bibliography

Condit, C.M., 1994. Hegemony in a mass-mediated society. Critical Studies in Mass Communication 11 (3), 205–230.

García Canclini, N., 1995. Hybrid Cultures: Strategies for Entering and Leaving Modernity. University of Minnesota Press, Minneapolis, MN.

Gaventa, J., 1980. Power and Powerless: Quiescence and Rebellion in an Appalachian Valley. University of Illinois Press, Urbana, IL.

Gitlin, T., 1980. The Whole World Is Watching. University of California Press, Berkeley, CA.

Gramsci, A., 1971. Selections from the Prison Notebooks. Lawrence & Wishart, London.

Hall, S., 1980. Encoding/decoding. In: Hall, S., Hobson, D., Lowe, A., Willis, P. (Eds.), Culture, Media, Language. Hutchinson, London, pp. 128–138.

Hall, S., 1996a. The problem of ideology: Marxism without guarantees. In: Morley, D., Chen, K. (Eds.), Stuart Hall. Routledge, London, pp. 25–46.

Hall, S., 1996b. Gramsci's relevance for the study of race and ethnicity. In: Morley, D., Chen, K. (Eds.), Stuart Hall. Routledge, London, pp. 411–440.

Hall, S., Critcher, C., Jefferson, T., Clark, J., Roberts, B., 1978. Policing the Crisis. Macmillan, London.

Hall, S., Jefferson, T. (Eds.), 1976. Resistance Through Rituals. Hutchinson, London.

Hallin, D.C., 1989. The Uncensored War. University of California Press, Berkeley, CA.

Hallin, D.C., 1994. We Keep America on Top of the World. Routledge, London.

Hebdige, D., 1979. Subculture: The Meaning of Style. Methuen, London.

Herman, E.S., Chomsky, N., 1988. Manufacturing Consent. Pantheon, New York.

Laclau, E., Mouffe, C., 1985. Hegemony and Socialist Strategy. Verso, London.

Lears, T.J.J., 1985. The concept of cultural hegemony: problems and possibilities. American Historical Review 90, 567–593.

Liebes, T., 1997. Reporting the Arab-Israeli Conflict. Routledge, London.

Lukes, S., 1974. Power: A Radical View. Macmillan, London.

Martín-Barbero, J., 1993. Communication, Culture, and Hegemony: From the Media to Mediations. Sage, London.

Morley, D., 1980. The 'Nationwide' Audience. British Film Institute, London.

Rose, T., 1994. Black Noise: Rap Music and Black Culture in Contemporary America. University Press of New England, Hanover, NH.

Williams, R., 1977. Marxism and Literature. Oxford University Press, Oxford.

Willis, P.E., 1977. Learning to Labor. Columbia University Press, New York.

Hegemony: Anthropological Aspects

Donald V Kurtz, University of Wisconsin, Milwaukee, WI, USA; University of Texas, San Antonio, TX, USA; and Southwest Texas State University, San Marcos, TX, USA

© 2001 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, Volume 10, pp. 6642–6645, © 2001, Elsevier Ltd.

Abstract

Anthropologists commonly cite Antonio Gramsci when they refer to hegemony and the related ideas that make up Gramsci's neo-Marxist theory of hegemony, such as culture, common sense, and the intellectuals. However, most anthropologists are inspired by interpretations of hegemony other than Gramsci's. To understand how anthropologists understand and use the idea of hegemony, it is essential to know how Gramsci's idea of hegemony as a moral and intellectual leadership was revised by non-anthropologists and equated with ideas of culture, tradition and ideology. By the time anthropologists discovered the idea of hegemony in the late 1980s it barely resembled Gramsci's. Because of the various interpretations and different applications of hegemony, hegemony has for many anthropologists been only a trendy buzz word that negates its political intent. Yet other anthropologists have seriously applied the idea of hegemony in diverse, novel, and creative ways in order to explore primarily two related concerns: the historical consequences of state and colonial policies for contemporary post-colonial societies, and relations of resistance and political domination. After briefly considering how Gramsci conceived hegemony and how others revised it, these two concerns provide the focus here for the anthropological use of hegemony.

Introduction

By the time anthropologists discovered Antonio Gramsci's idea of hegemony in the late nineteen eighties, it had been reworked by others so that it barely resembled Gramsci's. Although anthropologists cite Gramsci's *Prison Notebooks* (1971) when they refer to hegemony, most are inspired by revisionist interpretations of hegemony.

Hegemony

Hegemony is the central idea among many by which Antonio Gramsci established a humanistic, neo-Marxist approach to revolutionary change. Instead of subordinating the superstructure of ideas to the force of the economic base, Gramsci empowered the influence of ideas over Marxism's economic determinism. At the core of Gramsci's methodology is the dialectical relationship between hegemony (consent-support) and domination (coercion-force). His primary theoretical concern was how to resolve this contradiction, for in political practice one is always tempered by the other. He tried to reach resolution in part by redefining hegemony.

Lenin defined hegemony as domination. Gramsci redefined it as an *intellectual and moral leadership* directed by contradictory political and cultural agents and organizations he called *organic* and *traditional* intellectuals. Organic intellectuals represented the interests of working classes. One of their tasks was to make the proletariat aware of capitalist domination and exploitation. Traditional intellectuals represented the interests of the bourgeoisie. One of their tasks was to quell counter-hegemonic resistance to capitalism.

More than other Marxist thinkers, Gramsci recognized the significance of culture in revolutionary practice. Gramsci defined culture as 'the exercise of thought, the acquisition of general ideas *the habit of connecting cause and effect* … enlivened by organization' (Kurtz 1996a, 1996b). Culture was central to Gramsci's political strategy. He argued that culture was a political product of intellectuals, for another of their tasks was to develop the people's culture through education and ideological indoctrination. Traditional intellectuals worked to convince people of the virtues of capitalism and obtain their consent to capitalist rule. Organic intellectuals worked to help people develop an alternative to a capitalist culture. Among other strategies, Gramsci suggested how organic intellectuals might use popular cultural forms and signs to marshal working class revolutionary energies, depose capitalism, and construct a unitary proletarian culture.

Revisionists

Of the many works that explore Gramsci's thinking (Kurtz 1996a), anthropologists rely largely on those of Williams (1977) and Laclau and Mouffe (1985). Williams is concerned with the relationship between culture and hegemony. He argues that hegemony goes beyond culture, is synonymous with culture, serves as a form of 'cultural domination,' and, as a structure of cultural forms, provides a repository for traditional values. But, by equating hegemony with tradition, Williams divests it of Gramsci's political specificity.

Laclau and Mouffe are concerned with the heterogenous social organization of contemporary capitalist nations. They equate hegemony with ideology grounded in discourses by which oppressed categories in capitalist societies – women, gays, ethnic categories, and the like – can be made aware of their oppression. For Laclau and Mouffe hegemony (read ideology) provides the 'organic cement' of political formations that will transform western capitalism into a 'radical socialist democracy.'

These authors appropriated Gramsci's idea of hegemony and made themselves its authority, yet little that they say corresponds to Gramsci's idea of hegemony. They also ignore

the role of intellectuals in political practice. Still, the beauty and bane of their interpretations are the extent to which hegemony becomes amenable to diverse interpretations by anthropologists, but sometimes at the cost of the political intent with which Gramsci imbued it.

Anthropology and Hegemony

Anthropologists were attracted to the idea of hegemony for several reasons. It provided a neo-Marxist approach that complemented the idealist orientation among anthropologists who emphasize the power of ideas over material forces; it provided a new way to explore political relationships; it enabled reconsideration of old ideas, such as culture. Above all, it was thought provoking. Consequently there is little consistency regarding how anthropologists use it.

Most often anthropologists use hegemony as a trendy buzz-word that dilutes the concept's power. The idea of hegemony becomes more complicated when anthropologists use it seriously to analyze political relations. Only rarely do anthropologists use Gramsci's definition of hegemony to explore relations of domination and subordination. Instead, most anthropologists cite Gramsci, if only in passing, as the source of the idea of hegemony. Then they relegate his idea of hegemony to a secondary status and make it a point of departure to use or develop other interpretations of hegemony (Kaplan and Kelly 1994, Kurtz 1996a).

As a tool for analysis, anthropologists most often equate hegemony with a political force in society which usually emanates from those in power. Some anthropologists use revisionist ideas of hegemony to focus on other of Gramsci's ideas, such as *common sense* (peoples' awareness of their condition) (*American Ethnologist* 1993). However they approach hegemony, anthropologists infer that a hegemony of the proletariat to counter domination is rare (*American Ethnologist* 1993, Kurtz 1996b).

When anthropologists explore the relationship between hegemony and culture, it is Williams' ideas of hegemony and culture, not Gramsci's, that informs these explorations. Others use Laclau and Mouffe's idea of hegemony as ideology in conjunction with Williams' ideas of hegemony as culture and tradition to rework Gramsci's ideas. In these instances anthropologists become the source and authority of new ideas of hegemony. Almost no anthropologist places either hegemony and domination or traditional and organic intellectuals in the dialectical context that Gramsci thought to be important in hegemonic encounters.

This diversity of application of hegemony by anthropologists can be explained by their frequent argument that Gramsci did not define hegemony clearly. It is true that Gramsci's *Prison Notebooks* are fragmentary and open to interpretation. But anthropologists also do not read the *Notebooks'* informative footnotes, Gramsci's preprison writings, or the authoritative exegetes of his work. As a result they privilege others' interpretations of hegemony. Regardless how anthropologists define and locate hegemony theoretically, it is most applicable to two related concerns: the consequences of state and colonial policies and political relations of domination and subordination.

State-Colonial Policies

Only a few anthropologists apply Gramsci's formulation of hegemony to explore the impact of colonial policies. Carstens (1991) used Gramsci's idea of hegemony to account for how unequal relations between dialectically opposed intellectuals, British missionaries and Okanagan Indian chiefs, changed Okanagan culture and relegated the Indians to an inferior economic status in contemporary Canada. Kurtz (1996b) extrapolated from Gramsci's idea of hegemony to develop the idea of *hegemonic culturation* to explain why political leaders strive to instill an ideology of work in their political communities and why resistance to these efforts is rare.

More commonly, analyses rely on Williams' idea of hegemony as cultural domination. Fox (1989) used Williams' idea of hegemony as culture to explore how cultural meanings arose in colonial India and compelled colonizer and colonized to conform to certain behaviors. Contrary to Williams, hegemonic agents are central to Fox's idea of hegemony. He demonstrates how intellectuals continually experiment with existing cultural meanings which both enable and constrain resistance to colonial rule. But hegemony is not a source of revolutionary change for Fox. Instead, as Williams' cultural domination, hegemony continues to subordinate behaviors to the meanings it encodes.

In analyzing cultural confrontations in colonial South Africa, the Comaroffs (1991) emphasize Williams' ideas of hegemony as tradition, and Laclau and Mouffe's idea that discourse has the power to transform indigenous culture. For the Comaroffs, hegemony represents a tradition which, through discourses, acquires an ideology that helped to reorganize indigenous South African societies in accord with colonial rule. Ideology, not hegemony, provided the force of change. As tradition, hegemony had power only to shape actions, perceptions, and changes to the extent that it provided the source of new ideological challenges to existing traditions. As a process, hegemony (read tradition) is reproduced by its interplay with ideology into a functionalist 'hegemonic order' that provides a repository from which cultural forms may be created through the discursive power of ideology.

Domination and Subordination

Anthropologists who analyze the impact of state and colonial policies on anthropological subjects do not ignore the relationship between resistance and domination. Domination and subordination are implicated in almost all application of hegemony. But anthropologists who emphasize these relations explore them largely in post colonial and authoritarian societies, such as Sri Lanka and Bolivia, through ideas of hegemony that are largely of their own making. They agree only that hegemony is always a process.

Among these explorations hegemony is perceived as a *synthesizing principle* for welding class interests into a coherent world view (McDonogh 1991). In another context hegemony is a *social process* embedded in discourses aimed at instilling an ideology of nationalism (Brow 1988). The persistence of patron-client relations is maintained by hegemony identified as a *force* that maintains a social structure through the cultural shaping of experience (*American Ethnologist* 1993).

Hegemony, conceived as a *historical moment*, explains how a ruling class shapes perceptions of ideas, such as female beauty, to legitimate their rule over the working class (*American Ethnologist* 1993). Some anthropologists load the idea of hegemony with an array of theoretical considerations. In one instance, hegemony functions as the source of *education* for Gramsci's idea of 'common sense' (Woost 1993). In another it represents an ongoing *formative process* that may promote resistance to domination (*American Ethnologist* 1993).

Most anthropologists who consider the role of agents in these relations emphasize the role of elites, and argue that proletarian agents are not very effective in developing counter-hegemonies. Even those few anthropologists who emphasize the role of working class intellectuals in confrontations with elites conclude that proletarian intellectuals have little impact on changing the sources of rule and domination. As Laclau and Mouffe suggest, this is because the proletariat is too alienated and submerged under a dominant culture and rule. A logical consequence of this tendency is to privilege hegemony as tradition and render it apolitical.

Although anthropologists use hegemony as understood by others than Gramsci, and often in more convoluted fashion, they tend to agree on what could only be disheartening to Gramsci: the hegemonic impact of colonial and some contemporary governments has sufficiently restructured the culture and society of subordinate peoples that counter-hegemonies to domination are difficult to attain. Hegemony, however conceived, seems to accommodate best the interests of the elites.

The Future of Hegemony in Anthropology

The concern of anthropologists with the idea of hegemony peaked in the early 1990s. Now hegemony is firmly ensconced in anthropological discourse and theoretical orientations concerned with culture, structuralism, a poststructuralism *cum* postmodernism, even a revised historical materialism that objectifies ideas as material forces. Regardless how anthropologists conceive hegemony, it holds a privileged place among a cadre of non-materialist, neo-Marxist anthropologists and provides them a humanistic methodology to explore novel cultural and political relations. The collapse of the Western European and, more recently, Soviet empires and the energy of the expanding world capitalist order have produced global disjunctions in social, cultural, economic and political relations. To explain these epiphenomena, anthropologists are likely to continue to develop creatively the explanatory power of the idea of hegemony.

See also: Civil Society and the Public Sphere: History of the Concept; Class Consciousness; Hegemony and Cultural Resistance; Marxism in Contemporary Sociology.

Bibliography

American Ethnologist 20 (1), 1993.
Brow, J., 1988. In pursuit of hegemony: representations of authority and justice in a Sri Lankan Village. American Ethnologist 15, 311–327.
Carstens, P., 1991. The Queen's People. University of Toronto Press, Toronto.
Comaroff, J., Comaroff, J.L., 1991. Of Revelation and Revolution. University of Chicago Press, Chicago.
Fox, R.G., 1989. Gandhian Utopia. Beacon Press, Boston.
Gramsci, A., 1971. Selections from the Prison Notebooks. Lawrence and Wishart, London.
Kaplan, M., Kelly, J.D., 1994. Rethinking resistance: Dialogics of 'disaffection' in colonial Fiji. American Ethnologist 21, 123–151.
Kurtz, D., 1996a. Hegemony and anthropology: Gramsci, exegeses, reinterpretations. Critique of Anthropology 16, 103–135.
Kurtz, D., 1996b. Hegemonic culturation and work in state formations. In: Claessen, H.J.M., Oosten, H.G. (Eds.), Ideology and the Formation of Early States. Brill, Leiden, The Netherlands.
Laclau, E., Mouffe, C., 1985. Hegemony and Socialist Strategy. Verso, London.
McDonogh, G.W., 1991. Culture and categorization in a turn-of-the-century Barcelona elite. Cultural Anthropology 6, 323–345.
Williams, R., 1977. Marxism and Literature. Oxford University Press, Oxford.
Woost, M.D., 1993. Nationalizing the local past in Sri Lanka: Histories of nation and development in a Sinhalese village. American Ethnologist 20, 502–521.

Heider, Fritz (1896–1988)

Bernard Weiner, University of California Los Angeles, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Fritz Heider is the originator of two dominant theories that have impacted a number of research areas within the social sciences, particularly psychology. These conceptions, labeled balance theory and attribution theory, were most visible during the 1960–1980 decades, following the publication of his groundbreaking book *The Psychology of Interpersonal Relations*. This article traces Heider's life, discusses his contributions to social science, and reviews some of the fields of research influenced by his ideas.

Fritz Heider was born in Vienna on 19 February 1986, the younger of two sons, and spent his childhood in Graz in the province of Styria (Steiermark), Austria. There he received a classical education at the state official Gymnasium. Unable to serve in the army during World War I because of the loss of vision in one eye due to an injury during childhood, he entered the University of Graz (where ironically at one point he studied depth perception). He did not have a professional goal in mind and in the European tradition attended a wide range of lectures available to all university students. His studies eventually centered on philosophy and psychology (in opposition to the pursuit of architecture or law desired by his more practical father, who urged him to have his other interests as hobbies rather than a life focus). Heider remained a student at Graz for his entire university training.

While at the University of Graz, Heider made contact with Alexius Meinong (1853–1920), an influential figure in European philosophy. Meinong was a student of Franz Brentano and came to Graz as Professor of Philosophy in 1882. He taught courses in experimental psychology and established the first psychological laboratory in Austria (in 1894) as well as the Graz school of Gestalt psychology. Heider also was influenced by the psychologist Vittorio Benussi (1846–1928), a student of Brentano and Meinong and among the first to conduct and publish experiments in the field of gestalt perception.

It was Meinong who suggested the topic of Heider's (1920) doctoral thesis with the obscure question: "Why do we say we see the house rather than the sun, for the reflected rays of sunlight are the stimuli that actually strike the eye?" Understanding this question, and the analysis and answer provided by Heider, offers insights into attribution theory, one of Heider's contributions to the social sciences. Heider's solution to Minong's question was to provide a variation to the so-called causal-representational theory of perception, which was known at that time. Heider accepted that objects of perception are not directly perceived but are inferred as causes of the sensations they produce. He contended that aspects of the environment differ in terms of stability and causal potency. Things, or distal stimuli, are relatively stable centers of causality; they are objective structures in the environment. Media, on the other hand, or proximal stimuli, are more variable parts of the environment providing information about things. These ambiguous symbols require decoding and vary depending on environmental conditions. Heider proposed in his dissertation, entitled *Thing and Medium*, that thing perception is a complex causal inference process, reconstructing the relatively constant environmental centers of causality from sensations by filtering out the causal effects of the media. He believed in a 'real world,' with the task of the observer to come to know this reality. Heider later carried these ideas into his analysis of person perception and dispositional inferences, thereby unifying some aspects of object and person perception.

After finishing his dissertation, written in 1919, Heider worked for a year as an applied vocational psychologist. But he soon became restless and moved to Berlin, where his uncle worked as a zoologist. At the suggestion of Benussi, he contacted the psychologists at the University of Berlin, the center of Gestalt psychology (a school of thought based on the principle that perception is dependent upon the environment in which the object is embedded and perceptual experiences cannot be explained by the mere sum of the component parts). An air of excitement permeated the Berlin department of psychology. The lectures in perception, especially those of Max Wertheimer, were popular among students and intellectuals from widely differing fields. There was the feeling that gestalt research and theory would have an impact on the development of psychology. Heider attended the lectures of Wertheimer as well as those of Wolfgang Kohler and Kurt Koffka. He also formed a lasting intellectual and personal friendship with Kurt Lewin, a younger member of the Berlin faculty. Lewin, in turn, had been influenced by Ernst Cassirer, one of Germany's rising young philosophers. Cassirer left Berlin for the University of Hamburg in 1919, prior to Heider's arrival, but he indirectly affected Heider via Lewin and more so later when Heider also moved to Hamburg.

After Berlin came the *Wanderjahre*, a period when Heider spent a good deal of time meandering while reading philosophy, especially Spinoza (who guided his subsequent reasoning about emotions) and Nietzsche, as well as literature and psychology. He also worked in an orphanage and later was supported by a caring uncle. In 1927, after he had returned to Berlin, Heider became an assistant to Wilhelm Stern (1871–1938) at the University of Hamburg, whose primary contribution to psychology was in intelligence testing (he coined the term IQ). During this time, Heider also came to know Heinz Werner as well as Cassirer.

Heider then experienced a major shift in his life and career. Kurt Koffka was on an extended fellowship conducting research

at Smith College in Northampton, Massachusetts, and needed an assistant. Stern recommended Heider, who accepted what he regarded as a 1-year position in the United States in 1930. Heider's principle appointment was at the Clark School for the Deaf, where Grace Moore was also a member of Koffka's research group and was one of the first persons Heider met in America. They were married shortly thereafter and Heider remained in the USA.

In the years that followed, Heider combined research at the School for the Deaf with teaching at Smith College, where he was an Assistant Professor. He published two monographs jointly with Grace Moore on the psychology of the deaf. The methods he devised to study deaf children inspired a well-known silent film Heider later made with Marianne Simmel that captured perceptions of personal causality. The Heiders also raised three sons during this period: Karl, who became an anthropologist; John, a clinical psychologist; and Stephan, a physicist. Then, in 1947, now more than 50 years old and still an Assistant Professor, Heider made another academic change, moving to the University of Kansas in Lawrence, Kansas.

This appointment was at the instigation of Roger Barker, one of the founders of environmental or ecological psychology. Barker was trained at Stanford University and completed a 2-year postdoctoral position with Kurt Lewin while Lewin was at the Iowa Child Welfare Research Station (later renamed the Institute of Child Behavior and Development, closed in 1974). Barker was asked to chair the Department of Psychology at the University of Kansas, which like many other educational institutions was growing at the completion of World War II. Barker brought with him a group of other Gestalt psychologists, including Martin Scheerer, trained at the University of Hamburg; Herbert Wright, who also worked with Lewin in Iowa; and Heider. Heider remained in Lawrence, Kansas for the remainder of his life.

While at Kansas, Heider (1958) completed his groundbreaking book on social perception and social cognition, *The Psychology of Interpersonal Relations* (PIR). Chapters of PIR had been available earlier in mimeographed form but were known to only a few psychologists. The book was completed with the considerable aid of Beatrice Wright, who provided order to Heider's thoughts. A number of awards followed, including the Distinguished Scientific Contribution from the American Psychological Association.

Heider's subsequent enormous influence is all the more remarkable in view of the fact that it is largely based on this single, nonempirical publication. Furthermore, he worked at a rather unknown institution, without research assistants and little funding, with almost the total absence of doctoral students, and the book was published when he was 62 years old – only 7 years from retirement in 1965. For the most part, he was removed from the mainstream of American psychology.

Fritz Heider died in Lawrence, Kansas, on 2 January 1988 at the age of 92. He was a gentle man with a sly twinkle in his eyes, shy and introspective, with a love of art and literature and a wide-ranging intellectual curiosity. Among other personal descriptors, he was tall and thin; spoke with a soft, heavily accented voice; was modest and unpretentious; and exhibited little care about the material world around him, as opposed to the psychological world that dominated his thoughts.

Scientific Contributions

Heider applied the laws of the perception of physical objects and events developed by Gestalt psychologists to the perception of persons and their actions. Two conceptual contributions guided by this unification were presented in PIR, known as balance theory and attribution theory. These areas of thought were dominant in social psychology and related fields for two decades, 1960–80, and remain current. In addition and of lesser influence was his folk psychology approach to scientific thinking and methodology, which was displayed in his analysis of balance and attribution.

Balance Theory

Gestaltists contended that there is a tendency toward simplicity in the perceptual system such that simple structures are preferred. For example, faces are perceived as symmetrical although that is not exactly the case; and even when a circular form is not quite closed, it is nonetheless perceived as a circle. Furthermore, parts of a system become organized and the groupings form a simple structure in an uncontrollable and inevitable manner. For example, two similar objects in close proximity to one another are joined to form a unit, as in the example of ** ##. In the perceptual field, particular arrangements are imposed.

Heider reasoned that object configurations, including social objects, could be classified into those that are in a balanced or imbalanced state. In social perception, balanced and harmonious configurations are preferred for they represent 'good' figures or simple structures. In determining the balance of a social structure, the sentiment and unit relations between the elements play a role. Sentiment relations are attitudinal or evaluative and are either positive ('like,' 'approve') or negative ('dislike,' 'disapprove'). Unit relations refer to belonging and, like sentiments, can either be positive ('similar,' 'close,' 'belongs to') or negative ('dissimilar,' 'far from,' 'does not belong to'). Heider drew upon Wertheimer's grouping principles of similarity, proximity, and covariation when formulating his ideas about unit determination. Furthermore, everyday concepts such as 'liking' and 'belongs to' formed the heart of his scientific method of mining folk psychology for the development of a scientific psychology of interpersonal relationships.

In a two-person system, if 'a' likes 'b' while 'b' likes 'a' (sentiment relations), then the system is in balance. But if 'a' likes 'b' while 'b' does not like 'a,' then there is disequilibrium; imbalance exists and there is a force toward change in the direction of balance. There is a simple mathematical rule determining balance: if the multiplication of the relations results in a positive sign, there is balance; a negative result indicates imbalance. Balance in this two-person system can be brought about by changing one's attitude toward the other, misperceiving the sentiment of the other, altering the attitude of the other, and the like. However, not all balanced systems are pleasant or sought. For example, if 'a' and 'b' dislike one another, then the system is in balance (since the multiplication of two negatives result in a positive sign) but is less satisfying than when 'a' and 'b' like one another.

Configurations with three entities, such as those involving two persons and an object, are more complex. If, for example,

'a' likes 'b' as well as object x (e.g., the elected president), then if 'b' does not like x, the situation is imbalanced and a force will be created to bring the system into a state of equilibrium by, for example, a change in one's attitude toward 'b' or toward x. In Heider's manner of integrating everyday observations, literary examples, and scientific analysis, he noted that many literary tragedies make use of balance principles (e.g., Romeo loves Juliet, Juliet loves Romeo, Juliet loves and obeys her parents but they do not accept Romeo). An example of 'latent' imbalance that contributed to the dramatic tension in Ibsen's *Wild Duck* occurs when a scrupulous husband loves his wife who, unbeknownst to him, has engaged in a dishonest act for his benefit.

Balance principles have been extended to larger number of elements and widely applied in the study of interpersonal relations, attitude change, social cognition, and other area of research. It was reported, for example, that allied coalitions before and during World War II could be predicted with the use of balance theory. In addition, a variety of cognitive consistency theories and associated ideas, including the widely recognized concept of cognitive dissonance, are closely related to Heider's notions of balance.

Balance theory provided a distinct approach to motivation because force toward change is derived from the configuration of cognitive elements rather than from psychogenic or viscerogenic need states, which Lewin and others deemed necessary for motivated action. In accordance with Gestalt principles, for Heider the dynamics of action arise out of the entire pattern of a complex structure. This documented an essential position of Gestalt psychology, namely, the whole is other than (or different from) the sum of the parts.

Attribution Theory

Causal attributions, or beliefs regarding the causes of events, were a second major focus of PIR. Attribution theory was even more impactful on the field of psychology than were Heider's beliefs about balance and played a major role in the history of social psychology.

In his analysis of causal inferences, Heider applied his causal theory of thing perception to the perception of persons, while recognizing that persons, being agents with intentions, are more complex than inanimate objects. Hence, in addition to determining when a person is the cause of an event, the perceiver needs to understand or draw inferences about the mind of that individual. As in thing perception, the perceiver is confronted with proximal stimuli and draws inferences about more fundamental distal causes. In precursory research, Heider and Simmel showed participants three moving objects on a screen (a large and small triangle and a circle). Interpretations of the film were usually made in terms of interactions between persons with feelings, with impacts perceived as fights generated by jealousy, joint movements as evidence of liking and 'belonging together,' and so on. Perceived causality thus organized perceptions and eliminated the distinction between the physical and a person.

The motivational instigator for causal attributions is the assumption that individuals desire to understand and to master their environment. This belief is in part derived from Buddhist philosophy as well as from functionalism in that understanding the causes of events is adaptive for future behavior. For example, if one's car does not start, then knowledge regarding the cause (shortage of gas, damaged battery) is sought and causal beliefs determine future behavior. Thus, attribution theory is included within the study of cognitive functionalism.

For Heider, causal beliefs, which answer 'why' questions, are elicited by physical events such as the unexpected appearance of water on the floor or a crack in the ceiling; by outcomes that apply to the self, such as failing an exam or being rejected for marriage; and by events and outcomes that refer to others, including another person failing an exam or being rejected for marriage. However, attribution search tends to be restricted to unexpected, negative, and/or important events inasmuch as causal understanding requires cognitive work.

In his analysis of causal inferences, Heider made use of some principles from balance theory. For example, individuals tend to attribute success to themselves and failure to the external environment. This has been labeled a 'hedonic bias' that maximizes positive affect and self-regard. Inasmuch as most individuals feel positively about themselves, ascribing positive outcomes to the self (a positive unit) and negative outcomes to external factors (a negative unit) results in a balanced system and therefore this bias may be a result of cognitive rather than (in addition to) motivational mechanisms.

In addition and in accordance with Heider's interpretation of thing perception, observers tend to ascribe events to the disposition of the actor rather than to the 'pull' of the environment. For example, in the situation where Jim hits John, the cause is likely to be perceived as Jim's aggressive disposition rather than to a provocation by John. The overestimation of person causality and the underestimation of environmental effects have been labeled 'the fundamental attribution error.' According to Heider, in a manner similar to the perception of a house rather than the rays of the sun, this perceptual error is fueled by the tendency to locate causality in enduring distal causes (the disposition and/or intent of the person) rather than to changing environmental stimuli (the proximal cause or media that provides information about distal causes).

Contemporary Attribution-Related Research

As previously indicated, attribution theory was a dominant conception in social psychology from approximately 1960–80. However, its influence did not end there. It remains a presence in social psychology and has made deep inroads in the fields of clinical psychology, motivation and emotion, cognitive psychology, educational psychology, personality, and political psychology. It is almost impossible to discuss interpersonal relationships without recognizing the concepts of balance and/or attribution. When assessing the contemporary impact of Heider, it must be recognized that his influence is now carried by more contemporary figures who interpreted and elaborated his ideas, while adding their own insights. Among the main contributors one generation removed from Heider are Harold Kelley, who made Heider's thoughts known through an influential book review and because of his analysis of covariation information as depicted in the 'Kelley cube'; Edward Jones, who conducted the initial studies regarding dispositional

biasing; and Bernard Weiner, who brought attribution principles to the study of achievement motivation and emotion. These three were among the authors of the book: *Attribution: Perceiving the Causes of Behavior* (1972), which helped lift attribution to a prominent role in psychology. More recently, other researchers two generations removed from Heider have, in turn, replaced these figures as leaders in their field.

What follows is a sampling of some of the attribution-related research areas that can be traced to the foundation laid by Heider. Rather than assessing the validity of the various hypotheses that have been suggested and tested, these research directions are merely presented here to show some of the intellectual activity stimulated by Heider's thoughts.

Clinical Psychology

There is substantial research examining whether causal ascriptions are precursors to depression and/or follow depression onset. Some have contended that when the causes of aversive outcomes are perceived as unchanging, unchangeable, and/or implicate the self, then a state of learned helplessness or hopelessness is experienced and depression may follow. That is, depression is caused by dysfunctional causal reasoning.

In addition to the adverse consequences of self-related attributions, inferences about the causes of symptoms displayed by others can contribute to the course of mental illness. Critical and hostile emotional expression from caregivers, spawned by the causal belief that the other is responsible for his or her illness, increases the recidivism rates of schizophrenics as well as depressives. Mutual blaming for negative events also contributes to (and follows from) dissolving marriages by giving rise to resentment, anger, and discord.

Aggression and Prejudice

A number of different expressions of aggression have been subject to attribution analyses. Some types of child abuse, wife battering, and marital hostility can be traced to the aggressor ascribing intentional misbehavior to the other ("She purposively made me a bad meal."). In general, aggressive individuals tend to overattribute intentionality to others when experiencing a negative outcome. Hostility also is displayed toward the stigmatized when those individuals are perceived to be responsible for their plights (e.g., obese because of overeating), as opposed to being nonresponsible victims (e.g., obese because of a thyroid problem). Prejudice also is a product of (or may elicit) different attributions for in-group and out-group members. For example, aggression by an out-group member tends to be perceived as controllable by others and intentional, thus producing further out-group derogation.

Coping

Following a negative life stressor, such as rape, unwanted pregnancy, or illness, there often is a search for causality ("Why me?"). Some individuals ascribe the aversive event or outcome to their character ("I am too impulsive."), whereas others place fault on a particular behavior ("I was at the wrong place at the wrong time."). These contrasting causal beliefs elicit different coping mechanisms. The more adaptive behavioral blame results in active coping, whereas the less adaptive characterological blame, which implicates stable aspects of the self, gives rise to dysfunctional reactions. In general, coping depends on the taxonomic placement of the perceived cause of a problem into three fundamental categories: due to the person or the environment (labeled causal locus); an enduring or temporary cause (causal stability); and a controllable or not controllable cause (which in good part determines responsibility and blame). Self-attributions to internal, stable, and uncontrollable causes particularly elicit poor coping strategies.

Motivation and Emotion

Attributions for achievement success and failure have important consequences on subsequent striving. Guided by a can/try distinction articulated by Heider, it has been documented that attributions for failure to lack of aptitude (an internal, stable, and uncontrollable cause) are particularly debilitating, in contrast to ascriptions to insufficient effort (internal, unstable, and controllable causality), which promote positive motivation. In a similar manner, the belief or mindset that ability is changeable is more adaptive than the belief that one's ability level is a fixed entity.

Reactions of others also are guided by inferences about the causes of success and failure. Nonattainment of a goal because of lack of aptitude elicits sympathy and help from others, whereas failure due to low effort gives rise to anger and antisocial reactions. Furthermore, self-ascription of success to high aptitude elicits negative reactions from others, who regard the successful person as arrogant. On the other hand, self-ascription of success to effort or luck results in others labeling this individual as modest.

Appraisal approaches to emotion have documented that what one thinks, particularly about causes, determines how one feels. For example, guilt is typically preceded by the belief that one "could and should have done otherwise," whereas shame often follows from the inference that the causes of personal failure cannot be changed and mark one as different from others. Anger, regret, gratitude, and many other emotions follow from specific causal ascriptions.

Cognitive Psychology and Psycholinguistics

Many researchers have examined the antecedents of causal beliefs, ranging from individual differences in traits such as optimism – pessimism to the role of covariation information and rules regarding necessary versus sufficient causality. Of course, not all attempts to understand the contents and process of causal inferences can be traced back to Heider, inasmuch as phenomenal causality has been a long-standing interest of philosophers. Nonetheless, some of the proposals and thoughts do find their origin in Heider's thinking. For example, linguists have pointed out that verb forms affect causal reasoning. The phrase "I admire him" results in causality residing in the admired person, whereas the action sequence "Mary dominates Joan" places causality in the person committing the act. Those interested in gender issues have contended that journalists influence lay causal inferences when they state: "The flirtatious woman was raped by him," as

opposed to "He raped the flirtatious woman," with the former more promoting blame of the victim.

Political Psychology

Causal ascriptions are central components of political ideology and differ along the person – environment continuum pointed out by Heider. Conservatives, or those on the right of the political spectrum, tend to place causality for problems such as poverty on the person (laziness, lack of thrift, addicted, and so on) while maintaining it is not the fault of the government. In this manner, their causal beliefs are consistent with their desire to maintain the status quo and justify the negative stance of Republicans toward social welfare. On the other hand, liberals, or those on the left of the political spectrum, tend to see the causes of problems as residing in the environment (the government does not provide education or jobs, low wages). These attributions are consistent with their desire to change society and justify the positive attitude toward social welfare held by Liberals. In sum, causal beliefs are linked with broader political identifications.

Summary

It is evident from this very brief and selective article that Heider's ideas about perceptions of causality, represented in what is known as attribution theory, have a broad influence in psychology. In combination with balance theory, Heider left a flourishing psychological legacy.

Scientific Method

The sources of Heider's ideas were largely observational and literary. He relied on 'thought experiments' in which interpersonal situations were imagined and altered in a systematic manner (e.g., assume Romeo does not like Juliet, or that Juliet does not care about her parents). The psychological consequences of these new cognitive patterns were then inferred and related to configuration balance. This method may seem atavistic given the movement of psychology toward an experimental science but it proved fruitful for Heider, who advocated conceptual explication rather than empirical research.

For Heider, the psychology of social relations requires an understanding of the concepts used by the layperson, who was regarded as a naïve Newtonian relying on conceptual or logical inferences. The understanding of lay concepts does not necessitate, according to Heider, empirical methods. For example, if one leaves the house and turns right at the block corners, then that individual will arrive back home and this conclusion is not in need of empirical proof. Heider used this method of analysis to explicate concepts such as want, wish, benefit, harm, suffer, ought, can, try, and so on. He contended that reconstruction of the meaning of these terms reveals basic principles of social cognition. That is, the general conceptual structure embedded in everyday language provides the framework for describing psychological phenomena. In sum, Heider believed that a scientific psychology could be built on the common-sense psychology and language of everyday life and that a nonempirical conceptual analysis of everyday concepts provides psychologists with the methodology for the building of a social psychology of interpersonal relationships. Because of these beliefs and his mental experiments, Heider was labeled a 'naïve psychologist.' However, he did not advocate the elimination of empirical methods as opposed to an increase in conceptual treatments. And since his theoretical analysis embraced abstract principles, it provided laws that a layperson could not articulate. Thus, the 'naïve' characterization is not correct (although it contains a kernel of truth).

This contribution of Heider, however, met with little success, even among attribution and balance theorists. There were many reasons for the aversion of social psychologists to an analysis of common sense, ranging from the irrational (fear of not being regarded as a scientific endeavor) to the rational (the belief that science must develop its own language systems). However, the common-sense label applied to attribution theory played an important role in the history of attribution thinking. Attribution theory took hold after cognitive dissonance ruled psychology in the 1960s. At the core of this theory was the belief expressed by Leon Festinger, most responsible for this conception, that theory should contradict common sense. Subsequently, the waning of attribution theory toward the end of the 1980s coincided with (was caused by) a refocus on unconscious determinants of behavior and the growth of neuropsychology in cognition. Thus, attribution theory was framed within two contrasting Zeitgeists.

See also: Causation: Physical, Mental, and Social; Intentionality and Rationality; Lewin, Kurt (1890–1947); Mental Causation; Person Perception, Accuracy of; Phenomenology: Philosophical Aspects; Responsibility: Philosophical Aspects.

Bibliography

Heider, F., 1927. Ding und Medium. Symposium, Philosophische Zeitschrift fur Forschung und Ausprache 1, 109–157.
Heider, F., 1944. Social perception and phenomenal causality. Psychological Review 51, 358–374.
Heider, F., 1946. Attitudes and cognitive organization. Journal of Psychology 21, 107–112.
Heider, F., 1958. The Psychology of Interpersonal Relations. Wiley, New York.
Heider, F., 1976. A conversation with Fritz Heider. In: Harvey, J.H., Ickes, W.J., Kidd, R.F. (Eds.), New Directions in Attribution Research, vol. 1. Erlbaum, Hillsdale, NJ, pp. 3–21.
Heider, F., 1983. The Life of a Psychologist: An Autobiography. University of Kansas Press, Lawrence, KS.
Heider, F., 1988. The Notebooks of Fritz Heider. In: Benesh-Weiner, M. (Ed.), Psychologie Verlags Union, vols. 1–6. Munchen-Weinheim, Germany.
Heider, F., Simmel, M., 1944. An experimental study of apparent behavior. American Journal of Psychology 57, 243–259.
Jones, E.E., Kanouse, D.E., Kelley, H.H., Nisbett, R.E., Valins, S., Weiner, B., 1972. Attribution: Perceiving the Causes of Behavior. General Learning Press, Morristown, NJ.
Rudolph, U., Reisenzein, R. (Eds.), 2008. 50 Years of Attribution Research. Social Psychology, vol. 39. (special issue).
Weiner, B., 1995. Judgments of Responsibility: A Foundation for a Theory of Social Conduct. Guilford, New York.

Helmholtz, Hermann Ludwig Ferdinand von (1821–94)

David Cahan, University of Nebraska-Lincoln, Lincoln, NE, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article outlines the life and career of Hermann von Helmholtz in order to illuminate his influence on the nascent social sciences, above all on psychology. In doing so, it indicates how Helmholtz's major contribution to developing the Law of Conservation of Energy helped inspire social and economic thinkers to conceive of individual human bodies and society at large in terms of energy. Moreover, it indicates how his work on determining the velocity of nerve impulses, other measuring work in physiological acoustics and optics, and his leading role in the nativist–empiricist controversy, contributed to the rise of experimental psychology. It stresses his use of instruments and the measuring process in making the sciences (here especially physiology and psychology) ever more rigorous disciplines. Finally, it indicates how his work and name were admired (if not always emulated) by several young luminaries of social science around the turn of the twentieth century.

Life

Hermann Ludwig Ferdinand von Helmholtz was born in Potsdam, Germany on 31 August 1821. His father taught humanistic subjects at the local gymnasium and he encouraged Hermann's keen interests in music and the arts, in enjoying nature, and in philosophy. At the gymnasium Helmholtz distinguished himself in every subject of study. Although his great intellectual love was physics, his father persuaded him that there was little call for professional physicists and that, in any case, he could not afford to send him to university. Helmholtz thus enrolled (in 1838) at the Royal Medical-Surgical Frederick-William Institute (Königliches medicinisch-chirurgisches Friedrich-Wilhelms-Institut) in Berlin, the Prussian military medical school, which provided full support to its students in exchange for 8 years of military service.

After graduating in 1842 and serving for a year as an intern at Berlin's Charité hospital, Helmholtz served nearly 5 years (1843–48) in Potsdam as an army doctor. The position afforded him much free time and he spent it conducting research in physiology and physics, publishing several papers in physiology and a seminal essay in physics. Then in 1848–49, he served as assistant to his former teacher Johannes Müller, a leading comparative anatomist and physiologist, and as an anatomy instructor in Berlin, before being given an early release from his remaining military obligations in order to begin an academic career as Extraordinary Professor of Physiology at the University of Königsberg. His career was launched.

While at Königsberg, Helmholtz invented the ophthalmoscope (1850). It soon revolutionized ophthalmology by allowing doctors to examine the living human retina, and thus to diagnose the eye far more thoroughly and treat it for disease. It also led to Helmholtz's becoming a full professor, boosting his reputation among the Prussian educational authorities, and giving him a pan-European reputation. So too did his related work in creating an ophthalmometer and in developing a theory of the eye's accommodation. In 1855, he transferred to the University of Bonn, where he spent the next 3 years as Professor of Anatomy and Physiology. Then, in 1858, he accepted Baden's offer to become Professor of Physiology at the University of Heidelberg. A year after his arrival there, his wife Olga died, leaving him a widower with two young children to raise. He was soon remarried, to Anna von Mohl (with whom he had three children), who made their home into a social center of Heidelberg intellectual life. The university built him (as part of a larger science complex) a new institute of physiology; he served as the university's Prorector (1862–63); and, more generally, became one of its stellar figures during what became its golden era.

With the establishment of the German Reich in 1871, the Prussian authorities were determined to reacquire Helmholtz as part of their effort to turn the University of Berlin into the centerpiece of German academic life. Helmholtz now became Professor of Physics at Berlin and Director of its Physics Institute. Prussia built him an entirely new physics building, the largest such in Germany; and his institute became a mecca for German and foreign students alike. In going to Berlin, Helmholtz also left physiology almost entirely and devoted himself almost exclusively to physics. From time to time he also pursued work in other areas: independently inventing non-Euclidean geometry; conducting pathbreaking work in chemical thermodynamics and theoretical chemistry; and conducting work in theoretical meteorology. His contributions to each of these areas were considered foundational. Still, it was mostly in the area of physics, in particular theoretical physics, that he worked from the late 1860s until his death in 1894. In Berlin, Helmholtz became an even more prominent figure in German academic and intellectual life than in Heidelberg. Since the early 1850s, he had regularly been called upon to give popular lectures on science, which he continued to do throughout Germany and in Britain. Thanks to his wife Anna, his home became one of Berlin's leading *salons*. He regularly served the Prussian and Reich governments as a representative at national and international scientific meetings, and in 1883 the German Kaiser ennobled him for his services. Four years later, he was appointed the founding President of the Imperial Institute of Physics and Technology (Physikalisch-Technische Reichsanstalt, PTR), a nonacademic government institute devoted to research in physics and technology, including the setting of physical standards and testing. He spent the last 7 years of his life directing the PTR and pursuing theoretical physics. He died in Berlin on 8 September 1894.

Scientific Contributions Pertinent to the Social Sciences

Helmholtz produced well over 150 original scientific papers, which were reproduced in his three-volume *Scientific Papers* (*Wissenschaftliche Abhandlungen*, 1882–1895); a three-part *Handbook of physiological Optics* (*Handbuch der physiologischen Optik*, 1856–1867) that analyzed and synthesized the results of physiological optics; a pathbreaking treatise on the scientific foundations and aspects of music, *On the Sensations of Tone as a Physiological Basis for the Theory of Music* (*Die Lehre von den Tonempfindungen, als physiologische Grundlage der Theorie der Musik*, 1863); and a two-volume collection of some 25 popular and philosophical essays (*Vorträge und Reden*, fifth edn., 1903) on various aspects of science. The hallmarks of Helmholtz's efforts in science included a gift for generalizing and synthesizing results; using the methods or results of one discipline to advance another; showing the importance of instrumentation and measurement for advancing science and doing so in a rigorous manner; turning philosophical problems into scientifically amenable questions; analyzing the epistemological foundations of science; relating science and art to one another; and popularizing science. Although Helmholtz was himself not a social scientist – the English term itself being an anachronism during much of his active career – several highly selective aspects of his work in physics and physiology merit attention for their pertinence to the social sciences, above all psychology.

The Conservation of Energy and Socioeconomic Thought

In the mid-1840s, Helmholtz and several other young physiologists (Carl Ludwig, Emil du Bois-Reymond, and Ernst Brücke) sought to establish an 'organic physics,' i.e., to explain life processes exclusively on the basis of the laws of physics and chemistry, thereby ridding physiology of the vague notion of 'life force' (*Lebenskraft*). Helmholtz centered his early scientific efforts on the problem of heat and its mechanical equivalent. In 1847, he published his epoch-making concept of the principle of conservation of force. Although he was only one of several scientists in the 1840s who conceived of the idea of force conservation in physical or physiological systems, his presentation offered by far the most systematic, mathematical, and general account to date, thereby helping to connect various disparate branches of physics. The work was part of his larger program of providing a fully mechanical interpretation of nature, organic as well as inorganic, though Helmholtz neither then nor later, maintained that human beings, whom he believed to possess free will, could be understood solely in material and mechanical terms.

During the 1850s and the 1860s several British physicists (especially William Thomson) transformed Helmholtz's principle of force conservation into that of 'energy conservation' and thus into the First Law of Thermodynamics. This new energy-based physics and the associated concept of 'work,' generally inspired a number of economic and social thinkers in the second half of the century. Thanks in good measure to Helmholtz's work, physical scientists and social thinkers alike now unequivocally recognized the impossibility of perpetual motion. Social and economic thinkers began to see both the individual human body and socioeconomic systems in general, in terms of individual or societal energy production, consumption, transformation, and output. Helmholtz helped give economists a greater sense of the importance of the conservation and general circulation of energy. His work in thermodynamics also helped deepen understanding of the human body as a mechanical system, a human motor. The new thermodynamic viewpoint helped give new meaning to the terms 'work,' 'labor,' and 'value.'

Nerve–Impulse Measurements

Helmholtz's work on the propagation velocity of the nerve impulse spoke directly (if unintentionally) to the interests of an emerging set of experimental psychologists. In 1849–50, Helmholtz conducted a series of experiments on the propagation velocity in frogs, demonstrating that the impulse was propagated at a finite and measurable velocity. He and others saw that the reflex nerve action could also be measured in humans, and that there was a small, noninstantaneous (finite) time interval between human thought and bodily response.

As a result, the traditional mind–body problem took on a new shape and a vigorous program of stimulus–response studies in physiological psychology began. Helmholtz's work provided (by implication) a measure of a subject's reaction time to sensory stimulation, e.g., to visual or auditory cues, to voluntary movements, and to the association of ideas. Wilhelm Wundt and others built on Helmholtz's work by constructing methods for measuring the time intervals in thought processes. Helmholtz's work, along with Ernst Heinrich Weber's and Gustav Theodor Fechner's psychophysical results (including the latter's stimulus–response law and his *Elements of Psychophysics* (*Elemente der Psychophik*, 1860)); Adolph Hirsch's chronometric studies on 'physiological time'; Wundt's physiological psychology; the work of Franciscus Cornelius Donders and Johan Jacob de Jaagers on the speed of thought, among others, marked important milestones in the emergence of experimental psychology as a separate discipline in the 1850s and 1860s. To be sure, through the last third of the nineteenth century the term 'experimental psychology' assumed a wide variety of meanings; yet it also (if not mainly) meant that some psychological phenomena could be measured quantitatively and so become the subject of laboratory research.

Physiological Acoustics and Optics

Beyond neurophysiology proper, Helmholtz's importance to the emergence of experimental psychology lay generally in his work in physiological acoustics and optics, especially in his two great books, the *Handbook* and *On the Sensations of Tone*. These two masterpieces became the standard reference works – and, for many, points of scientific departure and philosophical debate – for psychologists and others. Both had a similar structure: they first gave analyses of the physical aspects of their respective subjects, physiological optics and acoustics; then came the physiological analyses of the respective sensations of light and sound; and, finally, arguments for the importance of the psychological factor in mentally integrating sensations into experienced perceptions or sounds. In particular, Helmholtz believed that 'unconscious inferences' transformed the physical and physiological sensations received by the sensory organs and

associated nervous systems into perceptions. From birth onwards, he argued, the individual learned through empirical experience that certain types of behavior produced certain types of responses, which, in the course of time, the mind transformed into unconscious inferences, and from there into actions.

Helmholtz's work in physiological acoustics and optics began with his theory (borrowed from Müller) of specific sense (or nerve) energies. Müller had maintained that the stimulation occurring in each type of sensory organ was responsible for a specific sensory experience; each type of sensory nervous system was qualitatively different from the others. Helmholtz extended this theory for vision by claiming, as had Thomas Young long before him, that there were three sets of optical nerve fibers, each of which provided its own specific nerve energy that corresponded to the primary colors. Similarly, he extended Müller's theory for hearing by maintaining there were several thousand specific auditory energies. He presented his so-called resonance theory of hearing in *On the Sensations of Tone*, arguing that the ear is a set of resonators that analyzes homogeneous sounds into their components. Although he thought that the sensory processes of hearing and seeing were in good part mechanical ones, he did not think that sensations were copies, let alone mirror images, of the external world; rather, they were the 'signs' of it. Although his theory of specific energies was modified or rejected by some, it came to constitute part of the foundations of experimental psychology.

The Nativist–Empiricist Controversy

Helmholtz was not simply one of the two principal antagonists (Ewald Hering being the other) in the nativist–empiricist controversy over the proper foundations of vision research. He coined the very labels of these two fundamental positions and did much to shape the controversy between himself and Hering and their respective followers. Although their differences were many, the two sides mainly disagreed on whether the eye possessed and required a mind to see. The controversy began in the 1860s and lasted until the 1920s; indeed, in several ways it has continued down to the present.

In the *Handbook*, Helmholtz labeled Hering and his followers 'nativists,' since they claimed that spatial perception was innate to newborns. By contrast, he called himself and his followers 'empiricists,' since they claimed that spatial perception was an acquired phenomenon, one that each individual gradually learns through daily experience and which requires the mediation of unconscious inferences. The two sides also disagreed about the receptor mechanisms affecting color vision and about explanations for contrast phenomena, adaptation, optical illusions, eye movements, the constancy of color, retinal correspondence, and so on.

The controversy helped shape the intellectual substance and disciplinary formation of physiology and psychology. Experimentally minded psychologists like Carl Stumpf, Georg Elias Müller, and Franz Brentano opposed the empiricist view insofar as it employed unconscious inferences and so harkened back to psychology's roots in philosophy. On the other hand, nativism's stress on the biological foundation and explanation of perception threatened to replace psychology with physiology. Yet nativism also emphasized the role of phenomenological experience, and this appealed greatly to a number of experimental psychologists, including Stumpf, Hermann Ebbinghaus, Oswald Külpe, Erich Jaensch, and David Katz.

The Broad Impact of Helmholtz's Work on Psychology and Beyond

Unlike Fechner and Wundt, Helmholtz had no desire to create a new, institutionally independent discipline of psychology. His contributions and overall importance to the discipline came largely (and indirectly) from his work in physiology and physics and from his general prestige. His contributions to the 'New Psychology' – a broad, hardly uniform movement that first arose in Germany between 1860 and 1910 and was associated with such names as Wundt, Ebbinghaus, Külpe, Stumpf, Hugo Münsterberg, and others – consisted largely in the observational and experimental study of, and speculation about, sensations and perceptions. The New Psychology in good measure replaced the older philosophical psychology and in so doing adopted, adapted, or extended many of its methods and results from sensory physiology and psychophysics, not least those introduced by Helmholtz.

Helmholtz's work deeply influenced and must be seen against the emergence and course of the New Psychology. During the 1860s and 1870s, a number of psychologists aligned themselves with (sensory) physiology and it was during these decades that Helmholtz had his greatest impact on the field. For example, Wundt, who was Helmholtz's assistant at Heidelberg, shared Helmholtz's empiricist theory of perception, and Helmholtz later drew on Wundt's work on vision for his *Handbook* and supported him professionally. As advocates of psychology as an experimental discipline, they were allies. Yet there were also differences between the two: Wundt later rejected Helmholtz's mechanistic approach to sensation, the two men had a priority dispute concerning the theory of unconscious inferences, and Helmholtz criticized several of Wundt's views. The young Ebbinghaus was similarly influenced and supported by Helmholtz, whose work on the measurement of the velocity of nerve impulse served as a starting point for Ebbinghaus' famous work on measuring memory. Helmholtz was also one of Ebbinghaus' examiners for the right to teach at Berlin, where Ebbinghaus lectured in psychology (1880–94) and where Helmholtz supported him in other ways.

In 1890, when Ebbinghaus and Arthur König founded the *Zeitschrift für Psychologie und Physiologie der Sinnesorgane*, they included Helmholtz on their editorial board, perhaps more to honor him as a venerable figure than as an active practitioner of physiological psychology. The new journal played a leading role and constituted an important milestone in physiological psychology. Yet the discipline of psychology was itself changing. As it gradually separated from physiology during the last third of the century, a number of its practitioners (especially after 1920) gradually abandoned its elementalist, sensory-based approach in favor of the phenomenalistic approach of the Gestalt psychologists, who rejected Helmholtz's mechanistic approach to psychology and, like the nativists, his theory of perception. In short, from the 1890s psychology moved away from Helmholtz's views and gained a new sense of intellectual and institutional autonomy.

Helmholtz also had a strong impact on other young psychologists, two of whom warrant special mention. As a medical student and young medical researcher, Sigmund Freud became especially enamored with the program of 'organic physics.' (He studied principally with Brücke in Vienna and once hoped to study with Helmholtz in Berlin, though nothing came of this.) The young Freud sought to link neuropsychological to physiological phenomena, and his early psychological language is the language of physiology (e.g., that of 'energy'). He called Helmholtz 'one of my idols.'

Likewise the young William James had intended to study with Helmholtz, but never did so. He called Helmholtz 'immortal' and said that he was "perhaps the first scientific genius now [in 1868] above ground." Helmholtz's work and views on physiological psychology constituted an important source for James. As a young scientist, he assiduously studied Helmholtz's *Handbook*, *On the Sensations of Tone*, and the popular and philosophical essays. Helmholtz was one of the most-frequently cited authorities in James's *Principles of Psychology*. Although James often cited him positively (especially as a source of empirical evidence) he rejected Helmholtz's notion of unconscious inference and his sensation-based theory of perception; and he found Helmholtz's mechanistic and mathematical predilections in science alien to his own. James wrote of the *Handbook*: "Can I find fault with a book which, on the whole, I imagine to be one of the four or five greatest monuments of human genius in the scientific line? If truth impels I must fain try, and take the risks. It seems to me that Helmholtz's genius moves most securely when it keeps close to particular facts."

Beyond psychology, Helmholtz also affected several of the future leaders of the next generation of social science in ways that are perhaps impossible to pin down precisely. The young Franz Boas had originally intended to become a physicist and to study with Helmholtz but he never did, soon turning away from physics and toward geography before settling on anthropology. (Helmholtz did, however, serve as one of Boas's examiners when he, successfully, sought to obtain the right to teach at the University of Berlin.) The young Wilhelm Dilthey was a frequenter visitor to the Helmholtz home in Berlin. Although Dilthey's holistic and dynamic theory of consciousness stood in sharp contrast to Helmholtz's 'sensationalist' approach, he studied Helmholtz's writings carefully; considered Helmholtz's popular scientific essays as the best such set that had ever appeared in German; engaged him in friendly discussions over their philosophical differences; and graced his own *Introduction to the Human Sciences* (*Einleitung in die Geisteswissenschaften*, 1883) with an opening quote from Helmholtz.

By the 1920s Helmholtz's substantive achievements and outlook had undoubtedly lost some of their attraction for some psychologists. Nonetheless, for many in Germany and abroad, Helmholtz still stood, as he did for Max Weber in 'Science as a Vocation' (1919), as a model scientific researcher. Moreover, in developing his general views and analyses of the sociology of music, Weber turned to Helmholtz's work in the physiological foundations of music. In *On the Sensations of Tone*, Helmholtz had stressed the historical nature of musical development and he distinguished between natural laws and aesthetics. His book effectively laid the basis for the sociology of music. Weber used Helmholtz's book as one of his principal scientific analyses of music and music history guides in his own *The Rational and Sociological Foundations of Music* (*Die rationale und soziologischen Grundlagen der Musik*, 1921).

Helmholtz's general impact on the rise of psychology (and other areas of the social sciences) from the mid-nineteenth century onwards was thus broad and fundamental. Edwin G. Boring, in the second edition of his monumental *A History of Experimental Psychology* (1950), ranked Helmholtz, along with Darwin, James, and Freud, as one of the four most important figures in the history of psychology.

See also: Boas, Franz (1858–1942); Experimentation in Psychology, History of; Freud, Sigmund (1856–1939); James, William (1842–1910); Psychology, History of (Twentieth Century); Psychophysical Theory and Laws, History of; Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Ash, M., 1995. Gestalt Psychology in German Culture, 1890–1967. Holism and the Quest for Objectivity. Cambridge University Press, Cambridge.
Blaukopf, K., 1982. Musical Life in a Changing Society: Aspects of Music Sociology (D. Marinelli, Trans.). Amadeus Press, Portland, OR.
Boring, E., 1950. A History of Experimental Psychology, second ed. Appleton Century Crofts, New York.
Braun, C., 1992. Max Webers 'Musiksozologie'. Laaber-Verlag, Laaber.
Cahan, D., 1994a. Helmholtz and the civilizing power of science. In: Cahan, D. (Ed.), Hermann von Helmholtz and the Foundations of Nineteenth-Century Science. University of California Press, Berkeley, CA, pp. 559–601.
Cahan, D. (Ed.), 1994b. Hermann von Helmholtz and the Foundations of Nineteenth-Century Science. University of California Press, Berkeley, CA.
Cahan, D., 1995. Helmholtz als Kulturträger. PTB-Mitteilungen 105 (4), 249–254.
Canales, J., 2001. Exit the frog, enter the human: physiology and experimental psychology in nineteenth-century astronomy. British Journal for the History of Science 34, 173–197.
Conrat, F., 1904. Hermann von Helmholtz' psychologische Anschauungen. In: Erdmann, B. (Ed.), Abhandlungen zur Philosophie und Ihrer Geschichte, Heft 18. Max Niemeyer, Halle a. d. S.
Heidelberger, M., 2004. Nature from within: Gustav Theodor Fechner and His Psychophysical Worldview. University of Pittsburgh Press, Pittsburgh.
Helmholtz, H., 1856–1867. Handbuch der physiologischen Optik. Voss, Leipzig.
Helmholtz, H., 1863. Die Lehre von den Tonempfindungen, als physiologische Grundlagen für die Theorie der Musik. Friedrich Vieweg und Sohn, Braunschweig.
Helmholtz, H., 1882–1895. Wissenschaftliche Abhandlungen, 3 vols. Barth, Leipzig.
Helmholtz, H., 1903. Vorträge und Reden, fifth ed., 2 vols. Friedrich Vieweg und Sohn, Braunschweig.
Helmholtz, H., 1995. On the interaction of the natural forces. In: Cahan, D. (Ed.), Science and Culture: Popular and Philosophical Essays. University of Chicago Press, Chicago.
Koenigsberger, L., 1902–1903. Hermann von Helmholtz, 3 vols. Friedrich Vieweg und Sohn, Braunschweig.
Mirowski, P., 1989. More Heat than Light: Economics as Social Physics, Physics as Nature's Economics. Cambridge University Press, Cambridge, UK.
Rabinbach, A., 1992. The Human Motor: Energy, Fatigue, and the Origins of Modernity. University of California Press, Berkeley, CA.
Schmidgen, H., 2002. Of frogs and men: the origins of psychophysiological time experiments, 1850–1865. Endeavour 26, 142–148.
Schmidgen, H., 2003. Time and noise: the stable surroundings of reaction experiments, 1860–1890. Studies in History and Philosophy of Biological and Biomedical Sciences 34, 237–275.
Schmidgen, H., 2005. Repetitions and differences: psychophysiological time machines, 1850–1865. In: Zielinski, S., Wagnermaier, S.M. (Eds.), Variantology 1: On Deep Time Relations of Arts, Sciences and Technologies. Walther König, Cologne.
Turner, R.S., 1994. In the Eye's Mind: Vision and the Helmholtz-Hering Controversy. Princeton University Press, Princeton, NJ.
Weber, M., 1919. Science as a vocation. In: Gerth, H.H., Mills, C.W. (Eds.), From Max Weber: Essays in Sociology. Oxford University Press, New York, pp. 129–156.
Weber, M., 1921. Die rationale und soziologischen Grundlagen der Musik. Drei Masken, Munich.

Helping Intergroup Relations

Arie Nadler, Tel Aviv University, Ramat-Aviv, Israel
Samer Halabi, School of Behavioral Sciences, Tel Aviv – Jaffa Academic College, Jaffa, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article begins with research on help giving across racial boundaries, and continues to consider the different motivations that underlie the giving of help to the out-group (e.g., secure rewarding relations with the out-group or repairing the in-group's damaged image). The article continues to describe the effects of group members' self-categorizations on their willingness to help the out-group. It continues to the intergroup helping as status relations model, which rests on social identity theory that analyzes how groups create, maintain, or challenge existing status relations between them through the readiness to give, seek, or receive help to and from the out-group.

Since the 1960s, the phenomenon of helping has been a growing area of social psychological research (Penner et al., 2005). Much of this research centered on interpersonal helping and sought to answer the questions of *who*, *whom*, *when*, and *why* people help (Dovidio et al., 2006). The question of 'who helps' centered on developmental processes and personality traits that characterize the helpful individual. The question of 'whom' one helps centered on the effects of the identity of the person in need on the readiness to help him or her. And the question of 'when help' takes place focused on situational determinants of helping. Research on the helper's thoughts and feelings as factors that explain the giving of help is the focus of the 'why help' question. A parallel line of research centered on the recipient of help. It sought to uncover the variables and processes that predict and explain people's decision to seek help and their reactions to receiving it (Nadler, in press). This research has demonstrated that being helped is a double-edge sword. On the positive side, receiving help often alleviates difficulties and is seen by the recipient as a reflection of the helper's genuine care. On the other hand, dependency on others conflicts with the deeply ingrained value of self-reliance and may result in loss of prestige in the eyes of others and be detrimental to the recipient's self-esteem. Under these conditions, people often prefer to avoid seeking help or receiving needed help (Nadler, 2012).

The present article on *intergroup helping relations* departs from this traditional emphasis on interpersonal helping. It considers the variables that affect, and the psychological processes that explain, why and when people are willing to stretch a helping hand across the in-group boundary. The article begins with the research on *cross-racial help giving*, moves to consider the *motivations* that explain giving to an in-group or out-group, and continues to research on the role of *social identity* concerns in the giving, seeking, and receiving of help across group boundaries.

Cross-Racial Helping

Early research on intergroup helping focused on the readiness of White Americans to assist a White or an African-American. This research was driven by the concern that not helping a Black person constitutes the White bystander's expression of subtle racial prejudice. The decision of a White American to help an African-American reflects two opposing psychological tendencies. While prejudicial attitudes may inhibit the readiness to help, the norm against displaying such blatant discrimination will encourage help giving regardless of the needy's group affiliation. One manner in which people solve the conflict between prejudicial tendencies toward discriminatory nonhelping and the normative sanctions against expressing them is 'aversive racism' (Dovidio and Gaertner, 2004). This concept suggests that prejudices will be expressed when the costs of expressing them are relatively low and will not be expressed when these costs are relatively high. The aversive racist "discriminates primarily when features of the situation allow him or her to maintain a positive (that is, nonprejudiced) view of the self" (Hodson et al., 2010: p. 120). Applied to the context of cross-racial helping, this suggests that if nonhelping can be attributed to external circumstances rather than to the helper's prejudicial tendencies, the costs of refusing help are relatively low and discriminatory nonhelping is more likely.

This idea has received considerable empirical support. An early study found that when White participants had been under the impression that other bystanders are present, they discriminated against African-Americans by not helping them more than when they had thought that only they and the victim were present (Gaertner and Dovidio, 1977). What accounts for this finding is that when others are present, responsibility for intervention is distributed among all bystanders present, thereby lowering the cost of nonintervention (Latané and Darley, 1970). Two literature reviews of cross-racial help giving, published 25 years apart, have similarly noted that cross-racial helping is dominated by the phenomenon of aversive racism. The first concluded that a White helper helps Blacks and Whites equally only when the interaction between them is a face-to-face interaction (Crosby et al., 1980). This is because the cost of discrimination in a face-to-face interaction is higher than in an impersonal interaction. The second review indicates that discriminatory nonhelping was likely only when not helping could be rationalized by the helpers (e.g., help involved high risk) (Saucier et al., 2005). Finally, consistent

with the logic of aversive racism, White individuals who had failed to help a Black person defended against the negative implications of this discriminatory nonhelping by assessing the event as being less of an emergency than when they had failed to help a White person (Kunstman and Plant, 2008).

Although this research has focused on racial relations in American society, its implications extend more generally to help giving by advantaged to disadvantaged or stigmatized group. The findings suggest that advantaged groups will express their prejudices toward stigmatized groups by not helping them only when external rationalizations decrease the costs of a blatant discriminatory nonhelping. This principle will predict the readiness of the majority to assist a stigmatized group within society (e.g., AIDS victims) and the willingness of rich nations in the Western world to assist poorer, and often stigmatized, nations in the developing world.

Motivations for Cross-Group Helping

Giving help across group boundaries is multicausal. It is often driven by the helper's genuine desire to assist members of disadvantaged or persecuted out-groups. The example of gentiles who had rescued Jews in Nazi-occupied Europe is a shining example of such genuine altruism (Oliner and Oliner, 1988). But, the motivation for cross-group helping may be less noble. It may reflect in-group members' desire for rewards such as gaining a reputation of generosity for their group. In this section we consider the multicausal nature of intergroup helping.

Empathy for the Other's Need? Giving to In-Group versus Out-Group Members

The motivational mechanisms that drive help to out-group members are different from those driving help giving to in-group members. People perceive others in their group as similar to them, and this interpersonal similarity is an antecedent of feelings of empathy with the other's plight, which in turn drives giving help to them (Stotland, 1969). In comparison, giving to others from the out-group is based on cost–reward considerations; it is driven by the liking that one feels for the person in need and a concomitant assessment of the degree to which a future relationship with him or her would be personally rewarding (Stürmer and Snyder, 2009). Consistent with this logic, the personal attractiveness of the person in need predicted helping directed toward an out-group member more than it did help toward an in-group member (Stürmer et al., 2005), and empathy with the needy's plight predicted readiness to help an in-group member more than it did an out-group member (Stürmer et al., 2006). In support of the role of empathy as driving help giving to in-group members, Penner and Finkelstein (1998) reported that dispositional empathy predicted volunteering for AIDS victims among homosexual men, but not among heterosexual women.

A similar conclusion concerning the different motivational paths that determine giving to in-group and out-group members comes from theory and research on the principles that underlie social interactions in close and distant relational systems. Clark and Mills (1993) have distinguished between communal and exchange relations. In communal relations giving is determined by the other's need, and in exchange relations it is dependent on rewards that one had previously received from the person in need. A similar analysis of relations systems indicates that it is only in emotionally significant groups (e.g., family), labeled as communal sharing relational systems, that people give to others according to their need and not because of other considerations (e.g., equal sharing between all group members) (Fiske, 1991). Although the three distinctions (i.e., in-group vs out-group, communal vs exchange relations, and communal sharing vs other relational systems) do not overlap, they convey a similar message regarding the motivational underpinnings of giving to an in-group or to an out-group member. In closer and psychologically significant relationships that are more common within the group, people give on the basis of empathy with others' need. In relations that are more distant and formal and less psychologically significant – attributes that are more characteristic of relations with out-groups – the considerations that dominate are more instrumental (e.g., whether or not one had received similar rewards) and unrelated to empathy for the other's need.

Helping as a Mechanism of Image Repair: Strategic Helping

Because helping out-group members is more diagnostic of the helper's generosity than helping in-group members, giving to the out-group is more common when it is driven by the in-group members' desire to gain a reputation of generosity for the in-group. This motivation to achieve positive reputation through out-group helping – a form of strategic help giving – is likely to be higher when in-group members experience a threat to their group's reputation or prestige (van Leeuwen and Tauber, 2009). In support of this prediction, Dutch students who had read about the disintegration of Dutch national identity (i.e., high threat to national prestige) were later more supportive of the suggestion that the Netherlands should assist the victims of the 2004 tsunami in Asia than Dutch participants whose national reputation had not been threatened (van Leeuwen, 2007). The greater readiness to assist the out-group was particularly pronounced in an area central to Dutch national identity (i.e., managing water resources; van Leeuwen, 2007: Study 2). In another experimental demonstration of strategic help giving, compared to a no-threat control condition, Scottish participants whose group image had been threatened by being told that the English view Scots as mean gave more to an out-group than to the in-group (i.e., a Welsh and a Scottish organization, respectively) (Hopkins et al., 2007).

The research reviewed in the preceding sections addressed helping relations between 'real groups.' These were racial groups in the USA, or Dutch and Scottish national groups. The focus was on giving to another collective of individuals that made up a distinctly different group. In the remainder of this article, the dominant concept will be people's social identity. This shifts the focus from the 'real group' to the 'psychological group' and puts the limelight of theory and research on individual group members' feelings and cognitions associated with their group membership.

The Social Identity Perspective on Intergroup Helping

The Psychological Group: The Social Identity Perspective

Early social psychological research on intergroup relations has focused on relations between two, or more, *real groups* made up of congregations of individuals. The social identity perspective, introduced in the 1970s, asserts that an important anchor of people's self-concept is the group(s) to which they belong, and that when group affiliation is salient, an interaction with an out-group member is experienced as an intergroup interaction even when it is an interpersonal interaction (Turner and Reynolds, 2001). The social identity perspective consists of two lines of research and theory: self-categorization and social identity theories (Turner et al., 1987; Tajfel, 1982, respectively). *Self-categorization theory* elaborates the cognitive mechanisms that cause people to categorize themselves in terms of a particular social group. It proposes that such self-categorizations are malleable. While people are affiliated with numerous groups (e.g., citizen, student, football fan), situational conditions cause one affiliation to be more salient than others and to drive social behavior. In the context of helping, research has examined how processes of self-categorization affect in-group members' willingness to assist members of the out-group. *Social identity theory* has focused on the comparative element in intergroup relations. It asserts that group members are motivated to enhance the value of their in-group relative to the out-group because this enhances their own self-concept. Intergroup helping research anchored in social identity theory has centered on the way in which giving, seeking, and receiving help affects and is affected by intergroup status hierarchy.

Self-Categorization and Intergroup Helping

People's readiness to intervene on behalf of a stranger is affected by the intentions and attitudes of fellow in-group members. Participants who had watched a violent clip on CCTV were more likely to express, privately, intentions to intervene to help a stranger if the in-group members had publically indicated that they would do so and less likely to intervene if the in-group members had indicated that they would not. The public expressions of the out-group members had no effects on participants' intentions to intervene (Levine et al., 2002). The attitudes that other in-group members hold about the out-group also affect people's readiness to intervene across group boundaries. White participants who had been informed that fellow in-group members hold positive attitudes toward African-Americans were more helpful toward them than those who had not received similar information (Sechrist and Milford, 2007). Together these findings demonstrate that only the behavioral intentions and attitudes of other in-group members affect readiness to help across group boundaries.

An experimental investigation of the helper's self-categorization on his or her readiness to help an in-group and out-group person in need used the context of the rivalry between British supporters of different football clubs (Levine et al., 2005). The first experiment examined the hypothesis that when the identity of the in-group is salient, people would be ready to help only other in-group members. To induce a salient in-group identity, supporters of Manchester United had been asked to write a short essay on the importance for them of being a supporter. In a second phase, as they were leaving the room, they encountered a person in an apparent need for assistance. In the in-group condition the person was wearing a shirt with a Manchester United insignia, in the out-group condition his shirt had an insignia of the rival Liverpool club, and in the control condition his shirt had no insignia. The findings indicate that more participants stopped to help in the in-group than in the out-group or control conditions. Thus, increasing the salience of in-group membership lowers the readiness to help across group boundaries.

How can this discriminatory nonhelping be overcome? One way, suggested by self-categorization research, is the induction of a common group identity that unites the in-group and the out-group under an overarching identity, while allowing the two groups to maintain their separate identities (Gaertner and Dovidio, 2000). Thus, for example, if professors of psychology and professors of economics are induced to think about themselves as 'social science professors,' they are encouraged to use a broad category of representation that includes both the in-group and the out-group under the same overarching identity. Consistent with research showing the beneficial effects of common identity on attitudes toward the out-group, it was expected that the induction of common identity will lower discriminatory nonhelping (Levine et al., 2005: Exp. 2). The experimental context was the same as the one used in the first experiment. To induce common identity, participants, fans of Manchester United, had been asked to write about the importance of being a 'lover of football.' This manipulation created a common identity of 'football lovers' that united the separate identities of supporters of Manchester United and Liverpool under the broader category of 'lovers of football.' Under these conditions, the number of participants who intervened to assist a member of the out-group was the same as those who intervened to assist a member of the in-group, and higher than the number of those who intervened to assist a person in the control group. Similar findings are reported in other studies. White students who had been induced to view Black students as fellow team members, working on a common task, complied with their requests more than those who viewed them as individual Black students (Nier et al., 2001). Finally, Dovidio et al. (1997) obtained similar findings with ad hoc groups created in the laboratory. This replication is of particular importance because obtaining the same results with groups who have no prior history increases the internal validity of this phenomenon.

A real-world illustration of the ameliorative effects of common group identity on intergroup helping comes from research on individuals who rescued Jews during the World War II. By risking their own and their families' lives, these individuals demonstrated extreme altruism toward an out-group (i.e., Jews). In their research on the 'altruistic personality,' Oliner and Oliner (1988) compared a large group of rescuers to a comparable group of nonrescuers. Of special relevance for the present discussion of the effects of common group identity on out-group helping is the finding that rescuers had higher 'extensivity' scores than nonrescuers. Extensivity denotes people's "propensity toward inclusiveness with respect to the diversity of individuals and groups to whom one will

assume obligations" (Oliner and Oliner, 1992: p. 370). Consistent with the view of extensivity as a general personal disposition, rescuers were also found to be more likely than nonrescuers to have positive feelings toward other out-groups like Turks and Gypsies (Oliner, 2004). The meaning of extensivity is close to that of common group identity. It means that the person views the in-group and the out-groups as sharing an overarching and inclusive identity of human beings.

Another illustration of the effects of common identity on intervening in behalf of the out-group during extreme circumstances is Reicher et al.'s (2006) analysis of the saving of Bulgarian Jews from Nazi deportation in 1941. Based on an analysis of pronouncements made by the Bulgarian leadership, Reicher et al. conclude that a key reason for this rescue was that Bulgarian leaders emphasized that Bulgarian Jews share a common identity with non-Jewish Bulgarians (Reicher et al., 2006). Todor Polakov, a Bulgarian parliamentarian of the time, is quoted as saying, "Bulgarian Jews … speak Bulgarian … they sing Bulgarian songs and tell Bulgarian stories…" (Reicher et al., 2006: p. 58). Finally, recent research has substantiated the links between 'extensivity' and readiness to help the out-groups in more mundane daily interactions as well (Einolf, 2010). High scorers on an 'extensivity' scale built especially for this purpose evidenced equal readiness to help family members and strangers, whereas low scorers exhibited greater readiness to help family members and friends than strangers.

This discussion of self-categorization and intergroup helping has focused on cognitive mechanisms that affect the representations of the out-group and the in-group, but has not attended to the status relations that exist between these groups before and after the helping interaction. Yet many, if not all, intergroup relations occur between unequal groups, i.e., high- and low-status groups. With the aid of social identity theory, the next section analyzes intergroup helping between unequal groups.

Social Identity Processes: Helping Relations between Unequal Groups

Providing to others implies relative strength and resourcefulness, whereas being dependent on their help marks relative weakness and inferiority (Nadler, in press). In evolutionary theory, 'costly signaling' theory holds that organisms incur short-term losses in order to advertise their fitness as mates and allies (Zahavi and Zahavi, 1997). Giving to others constitutes a 'costly signal.' For example, by feeding genetically unrelated chicks, birds engage in wasteful short-term behavior that establishes their longer-term desirability as mating partners (e.g., McDonald et al., 2008). In human groups, people gain status and prestige by giving to others and lose it when they habitually depend on others' assistance (Blau, 1963; Flynn, 2003). Because generosity is linked to increased intragroup status, individuals often compete on who is more generous ('competitive altruism'; van Vugt and van Lange, 2006). From the recipient's perspective, being dependent on the help of others is associated with lower status. Consequently, people in need often refrain from seeking needed help and respond negatively to its receipt (Nadler and Halabi, in press).

Intergroup relations are also characterized by links between giving and higher status on the one hand, and dependency and lower status on the other. In his seminal essay "The Gift," Marcel Mauss (1925) described the custom of Potlatch, common in tribal societies of the Pacific Northwest, where a clan leader who is motivated to signal the clan's superiority gives excessive gifts to other clan leaders. Similarly, in the Melanesian Islands, the more lavish the gift given by the family of the deceased to other families, the higher its status (Smith and Bleige-Bird, 2000). In his classic essay *Democracy in America*, Alexis de Tocqueville remarked that members of the ruling class maintain their dominance by assisting disadvantaged groups. He wrote that receiving such assistance is "servitude of the regular, quiet and gentle kind" (de Tocqueville, 1835/1956: p. 303).

These observations amplify the fact that intergroup helping often occurs within the context of structural inequality. The intergroup helping as status relations model (IHSR, Nadler, 2002) proposes that intergroup helping relations between structurally unequal groups depend on (1) structural characteristics of the intergroup hierarchy, (2) the autonomy or dependency orientation of the help, and (3) characteristics of individual group members.

Regarding *structural characteristics*, social identity theory suggests that a status hierarchy can be structurally secure or insecure (Tajfel and Turner, 1979). The security or insecurity of the status hierarchy is determined by whether or not it is viewed as stable and/or legitimate. In a secure hierarchy, inequality is viewed as stable and/or legitimate. In insecure hierarchy, it is viewed as unstable and/or illegitimate. When status hierarchy is secure, neither the high- nor the low-status group is motivated to change the unequal status quo. When the hierarchy is insecure, members of the low-status group who perceive an opportunity to change their disadvantaged position are said to be motivated to move toward greater equality.

Gender relations offer an illustrative example of secure and insecure status hierarchies. In past centuries, the superiority of men over women was considered as both legitimate and stable. In more recent times such inequality is regarded as illegitimate and in need of change. The IHSR proposes that when the status hierarchy is secure, the advantaged group is expected to cater to the needs of the disadvantaged in return for the latter's gratitude, loyalty, and legitimization of the existing inequality from which the high-status group benefits. When the status hierarchy is insecure, dependency on the high-status group is inconsistent with the low-status group's desire for equality. Under these conditions, members of the low-status group are expected to reject help from the high-status group.

Regarding characteristics of the help, its *autonomy or dependency orientation* determines the dynamics of helping relations between structurally unequal groups. *Dependency-oriented help* consists of full solutions and *autonomy-oriented* assistance consists of tools with which the needy can solve the problem on their own (Nadler, 1998). Dependency-oriented help underscores the recipient's reliance on the helper, whereas autonomy-oriented help allows the recipient to maintain perceptions of self-reliance. Therefore, in secure hierarchies where dependency of the low-status group is consensual, help is likely to be dependency-oriented, whereas

in insecure hierarchies, members of the low-status groups who strive for equality are expected to be receptive only to autonomy-oriented assistance. Finally, *group members' characteristics*, such as in-group identification, are also expected to affect these helping dynamics. These hypotheses were examined in research on low-status group members' readiness to seek and receive help from the high-status out-group, and the high-status group's readiness to help members of the low-status out-group.

Help Seeking and Receiving: The Case of the Low-Status Group

Compared to African-Americans who had not received help, those who had been assisted by a White American suffered a decrease in competence-related self-esteem (Schneider et al., 1996). In a different context, Israeli Arabs (i.e., a disadvantaged group in Israeli society) had more negative reactions to help from an Israeli Jew than to help from an Israeli Arab (Halabi et al., 2011). Although these studies had not measured or manipulated the perceived security of the status hierarchy, it seems likely that in both societies, African-Americans and Israeli Arabs viewed their group's relatively lower social status as illegitimate and in need of change. Therefore, being dependent on the higher-status group was threatening to the social identity of the disadvantaged group in both of the above-mentioned societal settings. A more direct assessment of the role of the security of structural inequality indicates that when members of a low-status group had been induced to view status relations as unstable, they were less receptive to help from the high-status group than when inequality was perceived as stable. Importantly, this was true with both ad hoc, experimentally created groups (Nadler and Halabi, 2006: Exp. 1) and real-world low- and high-status groups (i.e., Israeli Arabs and Israeli Jews, respectively) (Nadler and Halabi, 2006: Exp. 2). In the same vein, the more Israeli Arabs viewed status relations with the advantaged group of Israeli Jews as unstable, the more suspicious they were of offers for help from the latter (Halabi et al., 2012). Regarding the role of individual characteristics, Arab Israelis characterized as high in-group identifiers reacted more negatively to help from an Israeli Jew than low identifiers (Nadler and Halabi, 2006: Exp. 3).

A final experiment in this research program assessed the combined effects of (1) security of structural inequality, (2) group members' characteristics, and (3) help characteristics on the readiness of members of a low-status group to seek help from the high-status group (Nadler and Halabi, 2006: Exp. 4). Participants were high school students who could seek needed help from a student in a prestigious high school in their region. Half of the participants had, or had not, been induced to identify with their school, and had been told that the achievement gap between their school and the prestigious school had either narrowed or remained the same over the past 10 years (i.e., *unstable vs stable hierarchy*, respectively). The findings indicate that the participants who had been induced to identify with their school and believed that the achievement gap between the schools is narrowing (i.e., unstable hierarchy) were unwilling to seek dependency-oriented help from a member of the high-status group. This finding illustrates the joint effects of *social structure*, the *kind of help*, and the *person in need* on low-status group members' receptivity to help, as predicted by the IHSR.

Giving by the High- to the Low-Status Group: Defensive Helping

When the advantaged position of the high-status group is threatened, its members may defend against this threat by negative means such as arresting the forward movement of the low status (Ellemers, 2002). They defend against the threat by the more positive means of giving the low-status group much dependency-oriented help, thereby underscoring its dependence and relative inferiority. Such helping, intended to defend against incoming threats to the advantaged group's position, was labeled 'defensive helping' (Nadler et al., 2009). In support of this rationale, veteran Canadians who had been told that new immigrants to Canada present an economic threat to veteran Canadian citizens recommended giving new immigrants dependency-oriented help in the form of financial assistance more than autonomy-oriented help in the form of teaching them new working skills (Jackson and Esses, 2000). Similarly, Australians who had been informed that status relations between veteran Australians and newcomers to Australia are unstable (i.e., high threat) recommended less autonomy-oriented help to newcomers to Australia than to veteran Australians (Cunningham and Platow, 2007). The defensive nature of such help in maintaining the existing unequal status quo is highlighted by its links with the belief that social inequality is necessary and just (i.e., social dominance orientation (SDO); Sidanius and Pratto, 1999). In fact, defensive helping is more characteristic of high-SDO people than low-SDO people. In both Canadian (Jackson and Esses, 2000; study 2) and Israeli societies (Halabi et al., 2008), high-SDO scorers in the advantaged group (i.e., veteran Canadians and Israeli Jews, respectively) gave less autonomy-oriented help to members of the low-status group (i.e., newcomers to Canada and Israeli Arabs, respectively).

Self-categorization processes shape the phenomenon of defensive helping. When a common identity had been induced by telling participants from a high-status school that their own and the lower-status school share the common identity of schools that represent the same region, they did not react to the threatening information that the low-status school is closing the achievement gap by giving much dependency-oriented help (Nadler et al., 2009: Exp. 3). Instead, under these conditions of common identity, students from the high-status group gave predominantly autonomy-oriented help to students from the low-status group who had been described as closing the achievement gap with them. These findings indicate that categorization processes determine the meaning of 'threatening' information and the consequent kind of help giving. When the out-group shares no commonality with the in-group, intergroup reality is construed as a win–lose reality. The advancement of the out-group is perceived as undermining the in-group's social position. These perceptions give rise to defensive helping intended to ward off the threat to social identity. When the two groups share a common identity, however, intergroup reality is no longer experienced as a win–lose situation, and the other's progress does not pose a threat to the in-group identity. Rather, it is regarded as evidence that

the low-status group has the potential to move forward, and autonomy-oriented help is dispensed to bring about greater intergroup equality.

Conclusions and Implications

The common view suggests that we care for, and are cared for by, those who are similar and psychologically close to us. The findings that similarity breeds empathy reinforce this view and suggest that helping relations typically occur within group boundaries. The present article, however, indicates that this is not always the case. Under certain conditions, people take risks and sacrifice their resources in offering a helping hand across group boundaries. Thus, for example, when members of disadvantaged groups (e.g., White Americans) cannot avoid responsibility to help a distressed individual from the disadvantaged group (e.g., Black Americans), they help in-group and out-group members equally. Also, when people view the in-group and the out-group as possessing a common identity of shared humanity, they take great risks to save members of stigmatized groups from mortal danger (e.g., rescuers of Jews in Nazi-occupied Europe).

Intergroup helping relations are multicausal. Giving to the out-group can be driven by self-serving motivations. People may assist an out-group in order to enhance or repair the reputation of their group and/or to change or maintain an existing social hierarchy between unequal groups. In secure status hierarchies, the unidirectional downward flow of dependency-oriented help institutionalizes the unequal status quo. In insecure hierarchies, the exchange of autonomy-oriented assistance builds a ladder of social change toward greater equality. Self-categorization processes shape these dynamics of intergroup helping. When the intergroup boundary is clear and the separateness of the two groups is salient, intergroup helping relations between unequal groups are viewed from a competitive zero-sum perspective. Under these conditions, the advantaged group engages in defensive helping in the form of dependency-oriented help to maintain its relative superiority. When the disadvantaged and advantaged groups, however, share a broader common identity, a perspective of intergroup equality dominates, and intergroup helping takes the form of empowering the disadvantaged group toward forward movement.

The findings and processes discussed here have important implications. They enrich our understanding of basic dynamics of intergroup and helping relations. On the more applied level, the findings and their conceptualization are relevant to numerous real-world contexts such as efforts by advantaged groups to promote self-reliance among the members of disadvantaged groups. These processes are relevant to contexts in which rival groups seek to build a base of cooperation between them, or when richer nations build assistance programs to help poorer nations. In all these contexts, intergroup helping relations are multicausal and can have positive or negative consequences for the helper and the recipient. The key question concerns the mechanisms whereby intergroup helping is a vehicle to promote values of cooperation, mutual respect, and equality, rather than to perpetuate a competitive, win–lose perspective based on social inequality.

See also: Cooperation and Competition; Intergroup Relations; Prosocial Behavior and Empathy; Prosocial Behavior, Cultural Differences in; Racism: Social Psychological Perspectives; Self-Categorization Theory; Social Dominance Orientation; Social Identity in Social Psychology; Social Psychology.

Bibliography

Blau, P.M., 1963. The Dynamics of Bureaucracy: A Study of Interpersonal Exchange in Two Government Agencies. University of Chicago Press, Chicago.

Clark, M.S., Mills, J.R., 1993. The difference between communal and exchange relationships: what it is and is not. Personality and Social Psychology Bulletin 19, 684–691.

Crosby, F., Bromley, S., Saxe, L., 1980. Recent unobtrusive studies of black and white discrimination and prejudice: a literature review. Psychological Bulletin 87, 546–563.

Cunningham, E., Platow, M.J., 2007. On helping lower status outgroups: the nature of the help and the stability of the intergroup status hierarchy. Asian Journal of Social Psychology 10 (4), 258–264.

Dovidio, J.F., Gaertner, S.L., 2004. Aversive racism. In: Zanna, M.P. (Ed.), Advances in Experimental Social Psychology, vol. 36. Elsevier Academic Press, San Diego, CA, pp. 1–52.

Dovidio, J.F., Gaertner, S.L., Validzic, A., Matoka, K., Johnson, B., Frazier, S., 1997. Extending the benefits of recategorization: evaluations, self-disclosure and helping. Journal of Experimental Social Psychology 33, 401–420.

Dovidio, J.F., Penner, S.L., 2004. Aversive racism. In: Zanna, M.P. (Ed.), Advances in Experimental Social Psychology, vol. 36, pp. 1–52.

Dovidio, J.F., Piliavin, J.A., Schroeder, D.A., Penner, L.A., 2006. The Social Psychology of Prosocial Behavior. Lawrence Erlbaum, Mahwah, NJ.

Einolf, C.J., 2010. Does extensivity form part of the altruistic personality? An empirical test of Oliner & Oliner's theory. Social Science Research 39, 142–151.

Ellemers, N., 2002. Self and social identity. Annual Review of Psychology 53, 161–186.

Fiske, A.P., 1991. Structures of Social Life: The Four Elementary Forms of Human Relations: Communal Sharing, Authority Ranking, Equality Matching, Market Pricing. Free Press, New York.

Flynn, F.J., 2003. How much should I give and how often? Academy of Management Journal 46, 539–553.

Gaertner, S.L., Dovidio, J.F., 1977. The subtlety of white racism, arousal, and helping behavior. Journal of Personality and Social Psychology 35, 691–707.

Gaertner, S.L., Dovidio, J.F., 2000. Reducing Intergroup Bias: The Common Ingroup Identity Model. Psychology Press, Philadelphia.

Halabi, S., Dovidio, J.F., Nadler, A., 2008. When and how do high status group members offer help: effects of social dominance orientation and status threat. Political Psychology 29, 841–858.

Halabi, S., Dovidio, J., Nadler, A., 2012. Responses to intergroup helping: effects of perceived stability and legitimacy of intergroup relations on Israeli Arabs' reactions to assistance by Israeli Jews. International Journal of Intercultural Relations 36, 295–301.

Halabi, S., Nadler, A., Dovidio, J.F., 2011. Reactions to receiving assumptive help: the moderating effects of group membership and perceived need for help. Journal of Applied Social Psychology 41, 2793–2815.

Hodson, G., Dovidio, J.F., Gaertner, S.L., 2010. The aversive form of racism. In: Chin, J.L. (Ed.), The Psychology of Prejudice and Discrimination: Racism in America, vol. 1. Praeger, Westport, CT, pp. 119–135.

Hopkins, N., Reicher, S., Harrison, K., Cassidy, C., Bull, R., Levine, M., 2007. Helping to improve the group stereotype: on the strategic dimension of prosocial behavior. Personality and Social Psychology Bulletin 33, 776–788.

Jackson, L.M., Esses, V.M., 2000. Effects of perceived economic competition on people's willingness to help empower immigrants. Group Processes and Intergroup Relations 3, 419–435.

Kunstman, J.W., Plant, E.A., 2008. Racing to help: racial bias in high emergency helping situations. Journal of Personality and Social Psychology 65, 1499–1510.

Latané, B., Darley, J.M., 1970. The Unresponsive Bystander: Why Doesn't He Help? Appleton-Century-Crofts, New York.

van Leeuwen, E., 2007. Restoring identity through outgroup helping: beliefs about international aid in response to the December 2004 tsunami. European Journal of Social Psychology 37, 661–671.

van Leeuwen, E., Tauber, S., 2009. The strategic side of out-group helping. In: Stürmer, S., Snyder, M. (Eds.), The Psychology of Prosocial Behavior: Group Processes, Intergroup Relations, and Helping. Wiley-Blackwell, Oxford, England, pp. 81–99.

Levine, M., Cassidy, C., Brazier, G., Reicher, S.D., 2002. Self-categorization and bystander non-intervention: two experimental studies. Journal of Applied Social Psychology 32, 1452–1463.

Levine, M., Prosser, A., Evans, D., Reicher, S., 2005. Identity and emergency intervention: how social group membership and inclusiveness of group boundaries shape helping behavior. Personality and Social Psychology Bulletin 31, 443–453.

Mauss, M., 1925. The Gift: Forms and Functions of Exchange in Archaic Societies (I. Cunnison, Trans.). Norton, New York.

McDonald, P.G., Kazem, A.J.N., Clarke, M.F., Wright, J., 2008. Helping as a signal: does removal of potential audiences alter helper behavior in the bell miner? Behavioral Ecology 19, 1047–1055.

Nadler, A., 1998. Relationship, esteem, and achievement perspectives on autonomous and dependent help seeking. In: Karabenick, S.A. (Ed.), Strategic Help Seeking: Implications for Learning and Teaching. Lawrence Erlbaum, Mahwah, NJ, pp. 61–93.

Nadler, A., 2002. Inter-group helping relations as power relations: maintaining or challenging social dominance between groups through helping. Journal of Social Issues 58, 487–502.

Nadler, A., 2012. From help-giving to helping-relations: belongingness and independence in social relations. In: Snyder, M., Deaux, K. (Eds.), The Handbook of Personality and Social Psychology. Oxford University Press, New York, pp. 394–419.

Nadler, A. The other side of helping: seeking and receiving help. In: Schroeder, D., Graziano, W. (Eds.), The Oxford Handbook of Prosocial Behavior. Oxford University Press, New York, in press.

Nadler, A., Halabi, S., 2006. Intergroup helping as status relations: effects of status stability, identification, and type of help on receptivity to high-status group's help. Journal of Personality and Social Psychology 91, 97–110.

Nadler, A., Halabi, S. Helping relations and inequality between individuals and groups. In: Dovidio, J.F., Simpson, J. (Eds.), The APA Handbook of Personality and Social Psychology. APA Press, Washington, DC, in press.

Nadler, A., Harpaz-Gorodeisky, G., Ben-David, Y., 2009. Defensive helping: threat to group identity, ingroup identification, status stability, and common group identity as determinants of intergroup help-giving. Journal of Personality and Social Psychology 97, 823–834.

Nier, J.A., Gaertner, S.L., Dovidio, J.F., Banker, B.S., Ward, C.M., Rust, M.C., 2001. Changing interracial evaluations and behavior: the effects of common group identity. Group Processes & Intergroup Relations 4, 299–316.

Oliner, P.M., 2004. Saving the Forsaken: Religious Culture and the Rescue of Jews in Nazi Europe. Yale University Press, New Haven, CT.

Oliner, S.P., Oliner, P.M., 1988. The Altruistic Personality: Rescuers of Jews in Nazi Europe. Free Press, New York.

Oliner, P.M., Oliner, S.P., 1992. Promoting extensive altruistic bonds: a conceptual elaboration and some pragmatic implications. In: Oliner, P.M., et al. (Eds.), Embracing the Other: Philosophical, Psychological and Historical Perspectives. NYU Press, New York, pp. 369–390.

Penner, L.A., Dovidio, J.F., Piliavin, J.A., Schroeder, D.A., 2005. Prosocial behavior: multilevel perspectives. Annual Review of Psychology 56, 365–392.

Penner, L.A., Finkelstein, M.A., 1998. Dispositional and structural determinants of volunteerism. Journal of Personality and Social Psychology 74, 525–537.

Reicher, S.D., Cassidy, C., Wolpert, I., Hopkins, N., Levine, M., 2006. Saving Bulgaria's Jews: an analysis of social identity and the mobilisation of social solidarity. European Journal of Social Psychology 36, 49–72.

Saucier, D.A., Miller, C.T., Doucet, N., 2005. Differences in helping Whites and Blacks: a meta-analysis. Personality and Social Psychology Bulletin 9, 2–16.

Schneider, M.E., Major, B., Luhtanen, R., Crocker, J., 1996. Social stigma and the potential costs of assumptive help. Personality and Social Psychology Bulletin 22, 201–209.

Sechrist, G., Milford, L., 2007. The influence of social consensus information on intergroup helping behavior. Basic and Applied Social Psychology 29, 365–374.

Sidanius, J., Pratto, F., 1999. Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression. Cambridge University Press, New York.

Smith, E.A., Bleige-Bird, R.L., 2000. Turtle hunting and tombstone opening: public generosity as costly signaling. Evolution and Human Behavior 21, 245–261.

Stotland, E., 1969. Exploratory investigations of empathy. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 4. Academic Press, New York, pp. 271–313.

Stürmer, S., Snyder, M., 2009. Helping "us" versus "them": towards a group-level theory of helping and altruism within and across group boundaries. In: Stürmer, S., Snyder, M. (Eds.), The Psychology of Prosocial Behavior: Group Processes, Intergroup Relations, and Helping. Wiley-Blackwell, Oxford, England, pp. 33–58.

Stürmer, S., Snyder, M., Omoto, A.M., 2005. Prosocial emotions and helping: the moderating role of group membership. Journal of Personality and Social Psychology 85, 532–546.

Stürmer, S., Snyder, M., Kropp, A., Siem, B., 2006. Empathy-motivated helping: the moderating role of group membership. Personality and Social Psychology Bulletin 32, 943–956.

Tajfel, H., 1982. Social psychology of intergroup relations. Annual Review of Psychology 33, 1–39.

Tajfel, H., Turner, J.C., 1979. An integrative theory of intergroup conflict. In: Austin, W.G., Worchel, S. (Eds.), The Social Psychology of Intergroup Relations. Brooks/Cole, Monterey, CA, pp. 33–47.

de Tocqueville, A., 1956. Democracy in America. In: Heffner, R.D. (Ed.), Democracy in America: Mentor Book Edition. New American Library, New York.

Turner, J.C., Reynolds, K.J., 2001. The social identity perspective in inter-group relations: theories, themes and controversies. In: Brown, R., Gaertner, S. (Eds.), Intergroup Processes. Blackwell, Oxford, England, pp. 133–153.

Turner, J.C., Hogg, M., Oakes, P., Reicher, S., Wetherell, M., 1987. Rediscovering the Social Group: A Self-categorization Theory. Basil Blackwell, Oxford, England.

van Vugt, M., van Lange, P.A.M., 2006. Psychological adaptations for prosocial behavior: the altruism puzzle. In: Schaller, M., Simpson, J.A., Kenrick, D.T. (Eds.), Evolution and Social Psychology. Psychology Press, New York, pp. 237–261.

Zahavi, A., Zahavi, A., 1997. The Handicap Principle: A Missing Piece of Darwin's Puzzle. Oxford University Press, Oxford, England.

Hemispatial Neglect, Neural Basis of

Marlene Behrmann, Carnegie Mellon University, Pittsburgh, PA, USA
Sarah Shomstein, George Washington University, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M. Behrmann, volume 15, pp. 10467–10473, © 2001, Elsevier Ltd.

Abstract

Patients with hemispatial neglect, a neurobehavioral deficit that manifests after a lesion that usually affects the right hemisphere, fail to process or attend to information appearing on the contralateral side of space. This deficit can arise in multiple modalities (visual, auditory, and somatosensory) and can adversely affect eye and hand movements. In addition to describing the phenomenon and its anatomical concomitant, the underlying mechanisms giving rise to this deficit are explored. The advent of cutting-edge neuroimaging and the recent development of new assessment and rehabilitation methods have also offered novel insights into the disorder and these are summarized here, too.

The Term 'Hemispatial Neglect'

One of the most dramatic deficits in spatial perception and action is hemispatial neglect ('neglect' for short), a neurobehavioral disorder in which a patient is unable to acknowledge, report, or otherwise make explicit use of information falling in the visual hemispace contralateral to the lesion (Bartolomeo and Chokron, 2001; Harvey and Rossit, 2012; Kerkhoff and Schenk, 2012). Neglect occurs more commonly and with greater severity after right than left hemisphere lesions in humans (Suchan and Karnath, 2011) and this asymmetry is consistent with the differential involvement of the right hemisphere in processes of attention and spatial representation, as revealed in many recent neuroimaging studies (Chica et al., 2011; Ptak and Schnider, 2011). Because of this hemispheric asymmetry, in this review, we will refer to neglect as 'left-sided' and to 'contralesional' as a lesion to the right hemisphere.

The Phenomenon of Hemispatial Neglect

Neglect is not an uncommon disorder and every year about 3–5 million patients present with neglect after sustaining a stroke (Corbetta et al., 2005). Moreover, about a third of these patients manifest a chronic form of the impairment, which can persist for more than a year postonset (Karnath et al., 2011; Rengachary et al., 2011). Neglect can be greatly debilitating in everyday life; individuals may fail to notice objects on the left of a scene, may ignore words on the left of a page or food on the left of a plate, may drive their wheelchair into the side of a doorframe, and may dress and/or shave only one side of their body. Neglect can also interfere with rehabilitation efforts by adversely affecting mobility, discharge destination, length of hospital stay, meal preparation, and independence in self-care skills (Barer, 1990; Bernspang et al., 1987) and, understandably, then, neglect can affect the outcome of rehabilitation and long-term prognosis.

Neglect is not restricted to the visual domain; patients may not orient to contralateral auditory information, may not detect a contralateral somatosensory stimulus and may even ignore contralateral odors (for review, see Kerkhoff et al., 2012). Neglect is also not restricted to sensory input; patients with neglect may plan and even execute movements more poorly to the contralateral than ipsilateral side and this holds both for upper limb movements as well as for eye movements. The reduction in the speed and the precision of movements to the left appears not to be solely a downstream consequence of the sensory impairment and thus, problems in computing the parameters of contralateral movements is yet another manifestation of neglect.

The failure to process information on the contralateral side cannot be attributed to either a primary motor impairment or to a hemianopia; even when information is presented to the intact visual field of a neglect patient, information in the contralateral hemispace is still neglected. Patients may even show neglect in the absence of visual input; for example, as revealed in a particularly elegant and dramatic experiment, patients may ignore the left side of a scene constructed entirely in visual mental imagery (Bisiach and Luzzatti, 1978) and may fail to search the left of a dark room (Hornak, 1992). Finally, patients can be cued to process information on the contralesional side through verbal or other visual instructions although, left to their own devices, they tend not to orient contralesionally. Taken together, these findings all rule out a primary sensory deficit as being causally linked to neglect.

Aside from the general characteristics of neglect as set out above, the exact manifestation of neglect across patients is not monolithic. Thus, some patients demonstrate neglect for the contralesional side of the body (personal neglect as in dressing only one side of their body) but show no deficits on clinical tests requiring scanning across space (for example, on drawing or cancellation tasks). Other patients can show extrapersonal or allocentric neglect, where they fail to respond to external, contralesional stimuli, and when they do, the neglect can be for information appearing in near space (peripersonal neglect) or far space (or upper or lower space, see below). Neglect occurs most frequently along the horizontal plane although 'altitudinal neglect,' in which information appearing in the upper rather than lower portion of an array (or vice versa) is ignored, may also occur (Pitzalis et al., 2001). Neglect may also occur along the radial dimension although this pattern is thought to be related to the need for a motor response and is less commonly seen when perception alone is tested.

Finally, in some individuals, the severity of the neglect can be modulated by the nature of the information being processed; thus, the left side of items that 'belong' together or are connected perceptually is less subject to neglect than disconnected items (Behrmann et al., 1990; Tian et al., 2011; Ward et al., 1994). Whether different patterns of neglect are associated with different lesion sites is still controversial and this issue is discussed in greater detail in Section Different Forms of Neglect in Relation to Different Lesion Sites.

The Assessment and Diagnosis of Neglect

The diagnosis of neglect is generally rather straightforward – neglect is, in a standard manner, elicited in tasks in which the patient must either cross out items that appear randomly scattered on a piece of paper or must mark the midpoint of a single horizontal line and the neglect disorder can be elicited at the patient's bedside. Neglect is also typically observed when a patient is asked to draw or copy a simple picture such as a daisy or a clock and details on the contralesional side are generally omitted. A particularly popular test for hemispatial neglect, the Behavioral Inattention Test (Wilson et al., 1987), which consists of six subtests including drawing and line bisection, continues to be a measure of choice for many clinicians and researchers. The Birmingham University Cognitive Screen (BUCS; http://www.bucs.bham.ac.uk) is also useful and can be helpful in differentiating between different forms of neglect such as allocentric and/or egocentric neglect (Riddoch et al., 2010).

There have been other recent attempts to increase the sensitivity and reliability of assessment measures. The Sunnybrook Neglect Assessment Procedure, is the result of one such recent effort, and includes a small battery of tests that are quickly and easily administered and have high diagnostic differentiation (Leibovitch et al., 2012). Some bedside tests of hemispatial neglect, however, are still somewhat insensitive to subtle (but clinically relevant) forms of the disorder and considerable effort has been devoted to developing tools to address this issue. For example, the virtual reality lateralized attention test (Buxbaum et al., 2012) is an easy-to-administer computerized measure of hemispatial neglect that mimics the attentional demands of real-world tasks. In this test, participants name objects as they navigate (or are navigated) along a winding virtual path. Similarly, using virtual reality procedures, assessment techniques have been developed in which targets are placed at multiple locations and participants are required to locate the spatially disparate targets (Dvorkin et al., 2012). These latter, more dynamic, three-dimensional assessment tools offer a new approach, which may be valuable in identifying more subtle forms of the disorder and in tracking changes associated with interventions.

Underlying Mechanisms

A number of general interpretations have been advanced to account for the varied manifestations of neglect. For example, some have suggested that neglect arises from a deficit in orienting spatially to the contralesional side with a gradient of attention biasing attention to the ipsilateral side. Others have argued that the problem is one in which there is a rotation or translation of the egocentric frame of reference toward the ipsilesional side. Yet a third suggestion, the premotor theory of neglect, is that there is an imbalance between the spatial representations that control motor programs, and that this imbalance gives rise to an attentional deficit. Although described as alternatives, these differing interpretations may not be mutually exclusive and may refer to different aspects of a single deficit. What is clear at present is that, although they may diverge in detail, these explanations share a common perspective and that is that neglect is a higherorder, rather than sensory, spatial impairment.

One view of neglect, formalized by (Pouget and Driver, 2000) is that neglect reflects a partial loss or dysfunction of cells in the parietal cortex and other related neural areas. Because the number of neurons for representing information in different lateral positions is limited, damage to these neurons gives rise to a pathological gradient with a smooth decline for increasingly leftward locations and an off-center peak reflecting the peak activation on the ipsilateral side. Consistent with this idea, a spatial gradient has been incorporated into several neural network models that simulate neglect, and these theoretical models have successfully captured many aspects of neglect including extinction, performance on line-bisection and line cancellation tasks, and neglect dyslexia as manifest in various reading tasks (Mozer and Behrmann, 1990; Pouget and Sejnowski, 2001).

Alternatively, it has been argued that neglect behavior is best explained by the altered balance between the functioning of distributed default attentional cortical networks rather than by cell death, or structural damage, of specific regions. Recent studies investigating attentional control in neurologically intact individuals have demonstrated that attentional control is largely supported by two parallel networks, one dorsal and one ventral, each playing a specific role in attentional control (Corbetta and Shulman, 2011; Shomstein et al., 2010a,b; Vandenberghe et al., 2012). Corbetta and Shulman (2011) have argued that more ventral lesions to right parietal, temporal, and frontal cortices *directly* impair nonspatial functions (mediated by a ventral frontoparietal attention network) while also *indirectly* impacting the structurally intact more dorsal attentional network that is directly responsible for spatial attention. The fact that the behavioral gradient, described earlier, is known to be modulated by arousal and task instruction is then taken as further evidence that the underlying neural mechanisms are modulated by signals from other regions that are intact (areas within the dorsal network) rather than completely destroyed by structural damage. Perhaps related is the observation that neglect is accompanied by hypoperfusion (insufficient levels of oxygenation) of areas in the right angular and supramarginal gyrus and/or right superior temporal gyrus (Hillis et al., 2005; Khurshid et al., 2012), and that reperfusion, or rTMS, of these areas is associated with improvement of neglect (Cazzoli et al., 2010; Khurshid et al., 2012).

Lesion

As evident from the above discussion, much recent research has been concerned with demarcating the lesions that give rise to

neglect and the relationship between the lesions and the neural correlates that have been identified with attentional networks in nonneurological individuals in recent functional imaging studies. Traditionally, visual neglect has been associated with lesions of posterior parietal cortex but more recently, there has been increasing disagreement concerning the neural concomitant of the disorder.

Recent Lesion Mapping in Patients with Neglect

Recent studies exploiting high-resolution structural imaging, in tandem with voxel-based mapping and lesion overlap analyses, have confirmed the standard brain-behavior relationship, pinpointing the temporoparietal junction and the angular gyrus as the neural basis of neglect. Several other studies, however, have identified additional regions such as the intraparietal sulcus as being critical for neglect – a lesion to this region affects the reorientation of the spatial focus of attention and the ability to select between competing stimuli (Gillebert et al., 2011; Vandenberghe et al., 2012). Karnath and colleagues, for example, have reported that neglect may be associated with damage to more anterior areas, including the superior temporal gyrus and insula (Karnath and Rorden, 2012). Others have argued for the involvement of several other critical regions in addition to damage to posterior parietal regions, including, for example, the frontal eye fields and cingulate cortex. Yet others have proposed that the site of the lesion is tightly coupled to the type of neglect, with more dorsal lesions resulting in deficits of top-down guidance of attention and more ventral lesions resulting in a neglect deficit affecting bottom-up attentional guidance. Finally, much recent work exploiting diffusion tensor imaging to map white matter integrity in lesion patients has revealed that visual neglect can result from, and be correlated with, damage to frontoparietal connections in the right hemisphere, and that these white matter changes adversely affect large-scale cortical networks that are important for orientation of spatial attention, arousal, and spatial working memory (Bartolomeo et al., 2012; Karnath and Rorden, 2012; Urbanski et al., 2011).

Different Forms of Neglect in Relation to Different Lesion Sites

Many patients with left neglect following right hemisphere damage have neglect that manifests both egocentrically (i.e., show neglect with left defined with respect to the midline of coordinates centered on the body) and allocentrically (i.e., show neglect with left defined with respect to the midline of coordinates centered extrapersonally). In one recent study, roughly two thirds of individuals with neglect displayed both allocentric and egocentric neglect, while the remaining patients showed pure egocentric neglect. Consistent with the findings reported in Section Recent Lesion Mapping in Patients with Neglect above, the lesions in the inferior frontal gyrus, precentral gyrus, postcentral gyrus, superior temporal gyrus (STG), middle temporal gyrus (MTG), insula, and surrounding white matters were more frequent in the neglect group than in the control group. Of greater interest here is that there is a differentiation in lesion site depending on the pattern of neglect; thus, lesions in the right STG, MTG, lenticular nucleus, and the surrounding white matter were more frequent in the group displaying both allocentric and egocentric neglect than in those with pure egocentric neglect (Yue et al., 2012). The exact neural underpinnings of the different manifestations of neglect is still controversial, however. For example, Riddoch and colleagues have argued that egocentric neglect was associated with more anterior lesions, including the superior temporal gyrus, the supramarginal gyrus, and middle frontal gyrus whereas, allocentric neglect was linked to more posterior lesions, including the middle occipital and temporal gyri and the angular gyrus (Riddoch et al., 2010). Note, however, that Vallar and colleagues have alternately proposed that allocentric neglect is associated with more ventral, occipitotemporal lesions (Vallar et al., 2003).

It is also interesting that reperfusing specific areas has differential impact on different forms of neglect: for example, reperfusion of dorsal frontoparietal cortex is associated with improvement in viewer-centered (egocentric) neglect, whereas reperfusion of a more ventral temporooccipital cortex, is related to improvement in allocentric, stimulus-centered neglect (Cazzoli et al., 2010; Khurshid et al., 2012).

Implicit Visual Information Processing

One of the most provocative claims in the research on neglect is that, despite the loss of phenomenal awareness of the contralateral information, patients are able to process a considerable amount of contralateral information implicitly and this information can be used to guide their behavior (Driver and Vuilleumier, 2001). An early observation suggested that these patients 'knew' more than their overt responses suggested (Kinsbourne and Warrington, 1962). Thus, for centrally presented words, patients with neglect dyslexia often retained word length in their erroneous responses, for example, producing MILKMAN for MAILMAN. The influence of the unattended left-sided information has also been observed in other lexical tasks, in which patients with neglect produced faster responses to a stimulus in the right field when it had been preceded by a brief presentation of an associated word in the neglected left field (Ladavas et al., 1997a,b).

Other innovative paradigms have been developed to explore this issue in greater detail, allowing for the probing the integrity of more rudimentary perceptual processes on the neglected (contralesional) side. Specifically, studies have shown some degree of processing of the contralesional stimulus when it can be grouped with the ipsilesional stimulus on the basis of bottom-up factors such as color and proximity (Driver and Halligan, 1991), or by brightness or collinear edges (Gilchrist et al., 1996; Rorden et al., 1997). This is also true when the contralesional information is grouped with the right information by a global outline (Farah et al., 1993), and when the contralesional information forms the left side of an illusory contour (Kanizsa-type figure) of a partially occluded figure (Mattingley et al., 1997), or of any well-configured object or whole (Gilchrist et al., 1996; Humphreys and Riddoch, 1994).

Although the findings reported above are provocative and of great interest, many of the studies employed designs in which stimuli presented to the neglected field were associated with

stimuli presented on the right side of space by virtue of, for example, perceptual filling in, grouping by similarity, lexical relatedness, or semantic category. It remains possible, therefore, that the activation and processing of information on the neglected side occurs by virtue of the priming of the left-sided information by the intact right-sided information rather than being activated on its own merit (Boutsen and Humphreys, 2000; Brooks et al., 2005; Driver et al., 1992; Kumada and Humphreys, 2001; Mattingley et al., 1997; McGlinchey-Berroth, 1997). A recent study by Shomstein et al. (2010) employing semantically unrelated and task-irrelevant stimuli that avoids the possibility of priming information from the right to the left side, has nevertheless confirmed the processing of nonattended contralesional information. In this study, patients with left-sided neglect (and matched controls) judged whether successive complex checkerboard stimuli (targets), presented entirely to the intact right side of space, were the same or different. Concurrent with this demanding ipsilateral task, irrelevant distractor elements appeared on the unattended side and either changed or retained their perceptual grouping on successive displays, independently of changes in the ipsilesional, task-relevant, target. Changes in the grouping of the unattended task-irrelevant distractor produced congruency effects on the judgments of the ipsilateral attended targets to the same extent in neglect patients as in the control participants and this was true even in those patients with severe attentional deficits. These finding provide confirmatory evidence for the residual processing, at least to the level of perceptual grouping, of the neglected information. Much remains to be done to uncover the neural mechanisms that enable the processing of unattended information that is not available for overt report.

Perspectives on Rehabilitation

A diverse set of strategies has been adopted in studies designed to rehabilitate patients with neglect. While the intervention methods are numerous, there is not yet a single agreed-upon method that has proved to be the most effective. In a comprehensive literature review of cognitive rehabilitation following hemispatial neglect, Bowen and Lincoln examined the findings from 12 randomized controlled trials with 306 participants. Although the conclusion reached is that the overall effect of intervention was not statistically significant, cognitive rehabilitation did improve performance on some standardized neglect tests, including a reduction in the number of cancellation errors and an improvement in the ability to bisect the midpoint of a line. In a similar vein, an inclusive review of cognitive rehabilitation studies performed from 1998 to 2002 and then updated in 2005 (Cicerone et al., 2000, 2005) acknowledges the efficacy of cognitive rehabilitation, but also found that there is limited or insufficient evidence for the duration of treatment effects or relevance to everyday functioning (see also Rohling et al., 2009).

One of the most widely used rehabilitative methods, employed in clinical settings, is visual scanning or visual exploration therapy (VST or VET) in which patients are presented with an array of items and are asked to find an item embedded somewhere in the visual display. Of note is that this type of therapy is rather time-consuming with a reported minimum of 40 treatment sessions needed in order to achieve stable results (Antonucci et al., 1995; Kerkhoff, 1998). Empirical studies aimed at evaluating VST/VET's efficacy have yielded mixed results finding that, at worst, patients do not benefit from the method (Antonucci et al., 1995) or, at best, show very specific improvements (Kerkhoff, 1998).

It has been suggested that one of the central deficits in neglect is an orientation bias to the right, and therefore, several therapeutic techniques have been aimed at weakening the rightward bias and shifting the orientation bias centrally. Caloric vestibular stimulation in which iced water is introduced to the left ear and stimulates the vestibular apparatus has been shown to lead to a dramatic, albeit, short-term improvement in the disorder. This procedure has been shown to improve neglect-related disturbances of body perception, unawareness of hemiplegia, and postural imbalances, as well as the perception of the subjective straight ahead that is usually shifted to the right (Karnath, 1994; Rode et al., 1992, 1998). Vibration of the left neck muscles (which produces an afferent that is similar to twisting the trunk midline leftward) has also been shown to be successful (Johannsen et al., 2003; Schindler et al., 2002). Another approach to restoring interhemispheric imbalance that is currently being explored is rTMS and, although this is relatively new, some promising data are already available (see Cazzoli et al., 2010, for review).

Most recent rehabilitative efforts have been focused on prism adaptation as a technique showing potential for long-lasting effects. Patients wear glasses with lenses that shift visual information several degrees to the right (anywhere from 10° to 20°). When goggles are first placed on the patient, the patient performs motor actions, typically consisting of reaching out and touching a visual target that is placed randomly in right and left visual fields. As a consequence, patients will make erroneous reaches proportional to the degree of visual shift. If, however, patients are allowed to observe the last part of the action, they quickly learn to correct for the discrepancy. Following the adaptation period, the goggles are removed and patients are asked to make the same type of reaching movements again. Just as during the adaptation, patients make erroneous reaches only this time shifted in the opposite direction from the target (i.e., to the left of the actual target). Studies using this technique have shown that, following a brief (in some studies as short as 15 min) prism adaptation procedure, there is an increase in attentive behaviors toward the neglected side. Patients also showed increased exploratory behaviors toward the previously neglected side and this benefit persists even after the glasses have been removed (Ladavas et al., 2011; Serino et al., 2006, 2007, 2009; Turton et al., 2010). What sets this intervention apart from the previous attempts to alleviate unilateral neglect is that the observed amelioration generalizes across several different clinical measures, including wheelchair navigation, postural control, and neglect of mental imagery. In addition, the effects can persist for as long as 4 days (in rare cases 5 weeks) after a single adaptation procedure (Frassinetti et al., 2002; Pisella et al., 2002). The neural basis of the improvements remains poorly understood and may involve plasticity in some brain areas, although the parietal lobe is notoriously nonplastic

(Kolb and Cioe, 1998). It is also possible, however, that whatever change is observed may arise from involvement of other postural alterations (such as the shift of the viewer-midline more leftward).

Many other forms of treatment are also being explored. For example, in recognition of the fact that patients with neglect often have deficits in nonspatial forms of attention, too, some intervention approaches have focused on improving vigilance or sustained attention. One recent successful study of this type had patients monitoring long streams of novel visual scenes and, by virtue of this tonic and phasic alertness training, the researchers were able to demonstrate improvement in spatial as well as nonspatial forms of attention (Van Vleet and Degutis, 2012).

Concluding Remarks and Future Directions

As is clear from this review, hemispatial neglect is a heterogeneous and complicated disorder that defies simple explanation, and there remains a pressing need to understand what psychological and neural alterations give rise to this disorder. A further issue that requires additional investigation is the relationship of the spatial and nonspatial deficits in patients with neglect and their underlying neural correlates (Corbetta and Shulman, 2011). Much also remains to be determined regarding the underlying spatial map or frame of reference that defines neglect and whether the different manifestations of neglect arise from different underlying lesions and require different types of intervention.

An additional pressing issue concerns the fate of the neglected information; whether this information is processed normally and in a rich and detailed fashion but is somehow not available for conscious or explicit report continues to be hotly debated and highly controversial (Esterman et al., 2000). Promising studies using functional magnetic resonance imaging have shown differences in brain activation in the striate and extrastriate regions of the contralateral hemisphere for stimuli that have been extinguished compared with those that are reported (Driver and Vuilleumier, 2001; Marzi et al., 2000; Rees et al., 2000; Vuilleumier and Rafal, 2000) and further explication of the neural dynamics that support representations of nonattended information remains to be undertaken.

A final issue to resolve concerns the role of the parietal/temporal cortex in relation to other cortical regions that are involved in spatial processing. The parietal/temporal cortex does not have a monopoly on spatial processing and regions such as the frontal cortex and parahippocampal regions also contribute importantly to spatial representations. How these different areas work separately and/or together to give rise to our unitary visual experience remains unclear but neuropsychological studies with patients suffering from hemispatial neglect allow us to observe some of the representations and processes that ultimately contribute to our unified spatial experience. These findings together with those obtained from functional imaging studies with normal subjects, single-unit recordings from nonhuman primates and computational analyses continue to provide new insights into this perplexing disorder.

See also: Attention, Neural Basis of; Cognitive Neuropsychology, Methods of; Cognitive Neuroscience; Parietal Lobe; Perception of Action Space: Using Multiple Frames of Reference; Visual Streams: Dorsal and Ventral.

Bibliography

Antonucci, G.A., Guariglia, C., Judica, A., Magnotti, L., Paolucci, S., Pizzamiglio, L., et al., 1995. Effectiveness of neglect rehabilitation in a randomized group study. Journal of Clinical and Experimental Neuropsychology 17, 383–389.

Barer, D.H., 1990. The influence of visual and tactile inattention on predictions for recovery from acute stroke. The Quarterly Journal of Medicine 74 (273), 21–32.

Bartolomeo, P., Chokron, S., 2001. Levels of impairment in unilateral neglect. In: Boller, F., Grafman, J. (Eds.), Handbook of Neuropsychology, vol. 4. Elsevier Science, North-Holland, pp. 67–98.

Bartolomeo, P., Thiebaut de Schotten, M., Chica, A.B., 2012. Brain networks of visuospatial attention and their disruption in visual neglect. Front Hum Neurosci 6, 110. http://dx.doi.org/10.3389/fnhum.2012.00110.

Behrmann, M., Moscovitch, M., Black, S.E., Mozer, M., 1990. Perceptual and conceptual factors in neglect dyslexia: two contrasting case studies. Brain 113 (4), 1163–1183.

Bernspang, B., Asplund, K., Eriksson, S., Fugl-Meyer, A.R., 1987. Motor and perceptual impairments in acute stroke patients: effects on self-care ability. Stroke: A Journal of Cerebral Circulation 18 (6), 1081–1086.

Bisiach, E., Luzzatti, C., 1978. Unilateral neglect of representational space. Cortex 14, 129–133.

Boutsen, L., Humphreys, G.W., 2000. Axis-based grouping reduces visual extinction. Neuropsychologia 38 (6), 896–905.

Brooks, J.L., Wong, Y., Robertson, L.C., 2005. Crossing the midline: reducing attentional deficits via interhemispheric interactions. Neuropsychologia 43 (4), 572–582.

Buxbaum, L.J., Dawson, A.M., Linsley, D., 2012. Reliability and validity of the virtual reality lateralized attention test in assessing hemispatial neglect in right-hemisphere stroke. Neuropsychology.

Cazzoli, D., Muri, R.M., Hess, C.W., Nyffeler, T., 2010. Treatment of hemispatial neglect by means of rTMS–a review. Restorative Neurology and Neuroscience 28 (4), 499–510. http://dx.doi.org/10.3233/RNN-2010-0560.

Chica, A.B., Bartolomeo, P., Valero-Cabre, A., 2011. Dorsal and ventral parietal contributions to spatial orienting in the human brain. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 31 (22), 8143–8149.

Cicerone, K.D., Dahlberg, C., Kalmar, K., Langenbahn, D.M., Malec, J.F., Bergquist, T.F., Morse, P.A., 2000. Evidence-based cognitive rehabilitation: recommendations for clinical practice [Review]. Archives of Physical Medicine and Rehabilitation 81 (12), 1596–1615.

Cicerone, K.D., Dahlberg, C., Malec, J.F., Langenbahn, D.M., Felicetti, T., Kneipp, S., Catanese, J., 2005. Evidence-based cognitive rehabilitation: updated review of the literature from 1998 through 2002 [Review]. Archives of Physical Medicine and Rehabilitation 86 (8), 1681–1692.

Corbetta, M., Kincade, M.J., Lewis, C., Snyder, A.Z., Sapir, A., 2005. Neural basis and recovery of spatial attention deficits in spatial neglect. Nature Neuroscience 8 (11), 1603–1610.

Corbetta, M., Shulman, G.L., 2011. Spatial neglect and attention networks. Annual Review of Neuroscience 34, 569–599.

Driver, J., Baylis, G.C., Rafal, R.D., 1992. Preserved figure-ground segregation and symmetry perception in visual neglect. Nature 360 (6399), 73–75.

Driver, J., Halligan, P.W., 1991. Can visual neglect operate in object-centered co-ordinates? An affirmative single-case study. Cognitive Neuropsychology 8 (6), 475–496.

Driver, J., Vuilleumier, P., 2001. Perceptual awareness and its loss in unilateral extinction. Cognition 79, 39–88.

Dvorkin, A.Y., Bogey, R.A., Harvey, R.L., Patton, J.L., 2012. Mapping the neglected space: gradients of detection revealed by virtual reality. Neurorehabilitation Neural Repair 26 (2), 120–131.

Esterman, M., McGlinchey-Berroth, R., Verfaellie, M., Grande, L., Milberg, W., 2000. Aware and unaware perception in hemispatial neglect: evidence from a stem completion priming task. Neuropsychologia.

Farah, M.J., Wallace, M.A., Vecera, S.P., 1993. "What" and "where" in visual attention: evidence from the neglect syndrome. In: Robertson, I.A., Marshall, J.C. (Eds.), Unilateral Neglect: Clinical and Experimental Studies. Erlbaum, Hove, UK, pp. 123–137.

Frassinetti, F., Angeli, V., Meneghello, F., Avanzi, S., Ladavas, E., 2002. Long-lasting amelioration of visuospatial neglect by prism adaptation. Brain 125 (3), 608–623.

Gilchrist, I., Humphreys, G.W., Riddoch, M.J., 1996. Grouping and extinction: evidence for low-level modulation of selection. Cognitive Neuropsychology 13, 1223–1256.

Gillebert, C.R., Mantini, D., Thijs, V., Sunaert, S., Dupont, P., Vandenberghe, R., 2011. Lesion evidence for the critical role of the intraparietal sulcus in spatial attention. Brain: A Journal of Neurology 134 (6), 1694–1709.

Harvey, M., Rossit, S., 2012. Visuospatial neglect in action. Neuropsychologia 50 (6), 1018–1028.

Hillis, A.E., Newhart, M., Heidler, J., Barker, P.B., Herskovits, E.H., Degaonkar, M., 2005. Anatomy of spatial attention: insights from perfusion imaging and hemispatial neglect in acute stroke. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 25 (12), 3161–3167.

Hornak, J., 1992. Ocular exploration in the dark by patients with visual neglect. Neuropsychologia 30 (2), 547–552.

Humphreys, G.W., Riddoch, M.J., 1994. Attention to within-object and between-object spatial representations: multiple side for visual selection. Cognitive Neuropsychology 11 (2), 207–241.

Johannsen, L., Ackermann, H., Karnath, H.O., 2003. Lasting amelioration of spatial neglect by treatment with neck muscle vibration even without concurrent training. Journal of Rehabilitation Medicine: Official Journal of the UEMS European Board of Physical and Rehabilitation Medicine 35 (6), 249–253.

Karnath, H.O., 1994. Subjective body orientation in neglect and the interactive contribution of neck muscle proprioception and vestibular stimulation. Brain: A Journal of Neurology 117 (5), 1001–1012.

Karnath, H.O., Rennig, J., Johannsen, L., Rorden, C., 2011. The anatomy underlying acute versus chronic spatial neglect: a longitudinal study. Brain 134 (Pt 3), 903–912.

Karnath, H.O., Rorden, C., 2012. The anatomy of spatial neglect. Neuropsychologia 50 (6), 1010–1017.

Kerkhoff, G., 1998. Cognitive neurovisual rehabilitation: a cross-over study in patients with neglect and hemianopia. European Journal of Neuroscience 10, 375.

Kerkhoff, G., Keller, I., Artinger, F., Hildebrandt, H., Marquardt, C., Reinhart, S., Ziegler, W., 2012. Recovery from auditory and visual neglect after optokinetic stimulation with pursuit eye movements–transient modulation and enduring treatment effects. Neuropsychologia 50 (6), 1164–1177.

Kerkhoff, G., Schenk, T., 2012. Rehabilitation of neglect: an update. Neuropsychologia 50 (6), 1072–1079.

Khurshid, S., Trupe, L.A., Newhart, M., Davis, C., Molitoris, J.J., Medina, J., Hillis, A.E., 2012. Reperfusion of specific cortical areas is associated with improvement in distinct forms of hemispatial neglect [Research Support, N.I.H., Extramural]. Cortex; a Journal Devoted to the Study of the Nervous System and Behavior 48 (5), 530–539.

Kinsbourne, M., Warrington, E.K., 1962. A variety of reading disability associated with right hemisphere lesions. Journal of Neurology, Neurosurgery and Psychiatry 25, 339–344.

Kolb, B., Cioe, J., 1998. Absence of recovery or dendritic organization after neonatal posterior parietal lesions. Psychobiology 26 (2), 134–142.

Kumada, T., Humphreys, G.W., 2001. Lexical recovery from extinction: interactions between visual form and stored knowledge modulate visual selection. Cognitive Neuropsychology 18 (5), 465–478.

Ladavas, E., Bonifazi, S., Catena, L., Serino, A., 2011. Neglect rehabilitation by prism adaptation: different procedures have different impacts. Neuropsychologia 49 (5), 1136–1145.

Ladavas, E., Shallice, T., Zanella, M.T., 1997a. Preserved semantic access in neglect dyslexia. Neuropsychologia 35 (3), 257–270.

Ladavas, E., Umilta, C., Mapelli, D., 1997b. Lexical and semantic processing in the absence of word reading: evidence from neglect dyslexia. Neuropsychologia 35 (8), 1075–1085.

Leibovitch, F.S., Vasquez, B.P., Ebert, P.L., Beresford, K.L., Black, S.E., 2012. A short bedside battery for visuoconstructive hemispatial neglect: Sunnybrook Neglect Assessment Procedure (SNAP). Journal of Clinical and Experimental Neuropsychology 34 (4), 359–368.

Marzi, M., Girelli, M., Mimussi, C., Smania, N., Maravita, A., 2000. Electrophysiological correlates of conscious vision: evidence from unilateral extinction. Journal of Cognitive Neuroscience 12 (5), 869–877.

Mattingley, J.B., Davis, G., Driver, J., 1997. Preattentive filling-in of visual surfaces in parietal extinction. Science 275 (5300), 671–674.

McGlinchey-Berroth, R., 1997. Visual information processing in hemispatial neglect. Trends in Cognitive Sciences 1, 91–97.

Mozer, M.C., Behrmann, M., 1990. On the interaction of selective attention and lexical knowledge: a connectionist account of neglect dyslexia. Journal of Cognitive Neuroscience 2 (2), 96–123.

Pisella, L., Rode, G., Farne, A., Boisson, D., Rossetti, Y., 2002. Dissociated long lasting improvements of straight-ahead pointing and line bisection tasks in two hemineglect patients. Neuropsychologia 40 (3), 327–334.

Pitzalis, S., Di Russo, F., Spinelli, D., Zoccolotti, P., 2001. Influence of the radial and vertical dimensions on lateral neglect. Experimental Brain Research 136 (3), 281–294.

Pouget, A., Driver, J., 2000. Relating unilateral neglect to the neural coding of space. Current Opinion in Neurobiology 10, 242–249.

Pouget, A., Sejnowski, T.J., 2001. Simulating a lesion in a basis function model of spatial representation: comparison with hemispatial neglect. Psychological Review 108, 653–673.

Ptak, R., Schnider, A., 2011. The attention network of the human brain: relating structural damage associated with spatial neglect to functional imaging correlates of spatial attention. Neuropsychologia 49 (11), 3063–3070.

Rees, G., Wojciulik, E., Clarkje, K., Husain, M., Frith, C., Driver, J., 2000. Unconscious activation of visual cortex in the damaged right hemisphere of a parietal patient with extinction. Brain 123, 1624–1633.

Rengachary, J., He, B.J., Shulman, G.L., Corbetta, M., 2011. A behavioral analysis of spatial neglect and its recovery after stroke. Front Hum Neurosci 5, 29.

Riddoch, M.J., Chechlacz, M., Mevorach, C., Mavritsaki, E., Allen, H., Humphreys, G.W., 2010. The neural mechanisms of visual selection: the view from neuropsychology [Review]. Annals of the New York Academy of Sciences 1191, 156–181.

Rode, G., Charles, N., Perenin, M.T., Vighetto, A., Trillet, M., Aimard, G., 1992. Partial remission of hemiplegia and somatoparaphrenia through vestibular stimulation in a case of unilateral neglect. Cortex; a Journal Devoted to the Study of the Nervous System and Behavior 28 (2), 203–208.

Rode, G., Perenin, M.T., Honore, J., Boisson, D., 1998. Improvement of the motor deficit of neglect patients through vestibular stimulation: evidence for a motor neglect component. Cortex; a Journal Devoted to the Study of the Nervous System and Behavior 34 (2), 253–261.

Rohling, M.L., Faust, M.E., Beverly, B., Demakis, G., 2009. Effectiveness of cognitive rehabilitation following acquired brain injury: a meta-analytic re-examination of Cicerone et al.'s (2000, 2005) systematic review. Neuropsychology 23 (1), 20–39.

Rorden, C., Mattingley, J.B., Karnath, H.O., Driver, J., 1997. Visual extinction and prior entry: Impaired perception of temporal order with intact motion perception after unilateral parietal damage. Neuropsychologia 35, 421–423.

Schindler, I., Kerkhoff, G., Karnath, H.O., Keller, I., Goldenberg, G., 2002. Neck muscle vibration induces lasting recovery in spatial neglect. Journal of Neurology, Neurosurgery, and Psychiatry 73 (4), 412–419.

Serino, A., Angeli, V., Frassinetti, F., Ladavas, E., 2006. Mechanisms underlying neglect recovery after prism adaptation. Neuropsychologia 44 (7), 1068–1078.

Serino, A., Barbiani, M., Rinaldesi, M.L., Ladavas, E., 2009. Effectiveness of prism adaptation in neglect rehabilitation: a controlled trial study. Stroke 40 (4), 1392–1398.

Serino, A., Bonifazi, S., Pierfederici, L., Ladavas, E., 2007. Neglect treatment by prism adaptation: what recovers and for how long. Neuropsychological Rehabilitation 17 (6), 657–687.

Shomstein, S., Kimchi, R., Hammer, M., Behrmann, M., 2010a. Perceptual grouping operates independently of attentional selection: evidence from hemispatial neglect. Attention, Perception, and Psychophysics 72 (3), 607–618.

Shomstein, S., Lee, J., Behrmann, M., 2010b. Top-down and bottom-up attentional guidance: investigating the role of the dorsal and ventral parietal cortices. Experimental Brain Research 206 (2), 197–208.

Suchan, J., Karnath, H.O., 2011. Spatial orienting by left hemisphere language areas: a relict from the past? Brain 134 (10), 3059–3070 http://dx.doi.org/10.1093/brain/awr120.

Tian, Y., Huang, Y., Zhou, K., Humphreys, G.W., Riddoch, M.J., Wang, K., 2011. When connectedness increases hemispatial neglect. PLoS One 6 (9), e24760.

Turton, A.J., O'Leary, K., Gabb, J., Woodward, R., Gilchrist, I.D., 2010. A single blinded randomised controlled pilot trial of prism adaptation for improving self-care in stroke patients with neglect. Neuropsychological Rehabilitation 20 (2), 180–196.

Urbanski, M., Thiebaut de Schotten, M., Rodrigo, S., Oppenheim, C., Touze, E., Meder, J.F., Bartolomeo, P., 2011. DTI-MR tractography of white matter damage in stroke patients with neglect. Experimental Brain Research. 208 (4), 491–505.

Vallar, G., Bottini, G., Sterzi, R., 2003. Anosognosia for left-sided motor and sensory deficits, motor neglect, and sensory hemiinattention: is there a relationship? Progress in Brain Research 142, 289–301.

Van Vleet, T.M., Degutis, J.M., 2012. Cross-training in hemispatial neglect: auditory sustained attention training ameliorates visual attention deficits. Cortex; a Journal Devoted to the Study of the Nervous System and Behavior.

Vandenberghe, R., Molenberghs, P., Gillebert, C.R., 2012. Spatial attention deficits in humans: the critical role of superior compared to inferior parietal lesions. Neuropsychologia 50 (6), 1092–1103.

Vuilleumier, P.O., Rafal, R.D., 2000. A systematic study of visual extinction. Between- and within-field deficits of attention in hemispatial neglect. Brain 123 (6), 1263–1279.

Ward, R., Goodrich, S., Driver, J., 1994. Grouping reduces visual extinction: neuropsychological evidence for weight-linkage in visual selection. Visual Cognition 1 (1), 101–129.

Wilson, B., Cockburn, J., Halligan, P., 1987. Development of a behavioral test of visuospatial neglect. Archives of Physical Medicine and Rehabilitation 68 (2), 98–102.

Yue, Y., Song, W., Huo, S., Wang, M., 2012. Study on the occurrence and neural bases of hemispatial neglect with different reference frames. Archives of Physical Medicine and Rehabilitation 93 (1), 156–162.

Hempel, Carl Gustav (1905–97)

Gereon Wolters, Universität Konstanz, Konstanz, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Carl Gustav Hempel was a leading figure of the logicoempiricist movement, originating in the 1920s in Vienna and Berlin. He is particularly famous for his work on scientific explanation that led to the 'Hempel–Oppenheim model' of scientific explanation and his work on concept formation in empirical science. Hempel, furthermore, connected his methodological ideas with the thesis of the unity of science, according to which there holds the same methodology both in the natural and in the social sciences. In his later work, Hempel opened up to T. S. Kuhn's historic-pragmatic conception of science.

Carl Gustav Hempel was a leading figure of the logicoempiricist movement that dominated much of American philosophy in the second third of the twentieth century and has been a principal source of present day analytical philosophy. In the spirit of logical empiricism, Hempel aimed at a scientific philosophy with the principal objective to arrive at a purely logical methodology of the empirical sciences, including the social sciences. His most important contribution in this respect is what can be regarded as the first theory of scientific explanation (Salmon, 1989). This theory is referred to as 'the Hempel–Oppenheim model,' the 'covering-law model,' or the 'standard model of scientific explanation.' In his later years, Hempel gave up the idea of a purely logical (i.e., syntactical and/or semantical) 'rational reconstruction' or 'explication' of basic methodological concepts like explanation, confirmation, law of nature, probability, type, etc. Both the internal failure of this program and the influence of Thomas S. Kuhn caused Hempel to acknowledge for a satisfying reconstruction of methodological concepts also 'pragmatic' aspects of science in the sense of social and historical circumstances of its development.

Biography

Hempel was born on 8 January 1905, in Oranienburg (near Berlin, Germany). He studied mathematics, physics, and philosophy at the universities of Göttingen, Heidelberg, and Berlin from 1923 to 1929. After graduation (*Staatsexamen*) that qualified him to teach in secondary schools (*Gymnasien*), Hempel went to Vienna to study with Moritz Schlick and, particularly, Rudolf Carnap. Both, together with the mathematician Hans Hahn, the economist Otto Neurath, and others, had founded in 1923 the 'Vienna Circle' (*Wiener Kreis*), which was to become the nucleus of logical empiricism. After his return to Berlin in 1930, Hempel entered teacher's training and teaching and worked at the same time with Hans Reichenbach, head of the 'Berlin Group' of logical empiricism, on a philosophical dissertation on the concept of probability. After graduation in 1934, Hempel preferred to leave Nazi Germany and went to Brussels to assist Paul Oppenheim, a private scholar, in carrying out philosophical research that related principally to concept formation in empirical science. In 1937–38, Hempel worked for a year as research associate with Carnap in Chicago and again went back to Brussels until he found a post of instructor at City College, New York, in 1939. The academic year 1940–41 saw him at Queens College, New York, from where he went as associate professor to Yale in 1948. In 1955, Hempel became Stuart Professor of Philosophy at Princeton. After retirement at Princeton in 1973, he worked as university professor at Pittsburgh University from 1977 until 1985. He died at Princeton on 9 November 1997.

Major Contributions

Hempel has given important contributions to (1) the rational reconstruction of methodological concepts, particularly the concept of scientific explanation; (2) the 'universality thesis,' i.e., the thesis that the social sciences do not, and should not, use methods that are essentially different from the methods of empirical science; and (3) concept formation and taxonomy in empirical science.

Scientific Explanation

Hempel reconstructs scientific explanations of concrete events in nature, history, and society as arguments, i.e., as logical deductions containing as premises (*explanans*) at least one (general) scientific law L_1, as well as initial and boundary conditions C_r and as conclusion (*explanandum*) the statement E describing the event in question. We thus obtain the following model of 'deductive-nomological (D-N) explanation':

$L_1, L_2, \ldots, L_n$	the *explanans*
$C_1, C_2, \ldots, C_r$	[entails logically]
E	the *explanandum*

In case the *explanandum* contains essentially a statistical law (e.g., for all *x*: the probability of *G*, if *F* is given, equals *r* and as boundary condition that *F* is given for some *b*), the explanatory argument does not deliver a deduction of the *explanandum* statement E but rather gives inductive support for E, i.e., the *explanans* shows that E was to be expected with a certain degree *r* of probability. In simplified form (one statistical law, one

boundary condition), we thus obtain the following model of 'inductive-statistical (I-S) explanation':

$p(G, F) = r$	the *explanans*
$F(b)$	[lends inductive support of probability r to]
$G(b)$	the *explanandum*

In the case of I-S explanations, one must further assume that the *explanans* is maximally specified, i.e., the *explanans* must contain all available information that is possibly relevant for the explanation of the *explanandum* E. In other words, the *explanandum* must contain all pertinent statistical laws and those special facts that can be connected with the E event by statistical laws. I-S explanations thus relate to the body of scientific knowledge at a given time, i.e., they become 'epistemically relative.' The postulate of maximum specificity avoids the 'explanatory ambiguity,' i.e., the possibility to give acceptable I-S explanations based on true premises for mutually exclusive *explananda*.

In case one deduces as *explanandum* a general or a statistical law of lesser generality than the one(s) in the *explanandum* one obtains the reduction of a law or theory in the first case (e.g., explaining the law of free fall by means of the law of gravitation), and a deductive-statistical or D-S explanations of a statistical law by a more comprehensive statistical law in the second.

Originally, Hempel claimed the structural identity of explanation and prediction. Later he had to admit that, although every adequate explanation is (under pragmatically changed circumstances) also a prediction, not every adequate prediction yields also an adequate explanation, for symptoms are often adequate for a prediction but not for the corresponding explanation. So is, for example, a sudden fall of the barometer (together with suitable laws) sufficient for predicting a thunderstorm, but it does not explain it, since the fall of the barometer is not the cause of the thunderstorm, but rather a symptom of its arrival (Hempel, 1965).

Unity of Science

Hempel claims the methodological unity of empirical science. With respect to scientific explanation this amounts to a universality thesis of scientific explanation: There is no essential difference between explanations in the natural sciences and those in psychology, the social sciences, or history. Explanations in these fields likewise aim at showing that the event in question was to be expected on the basis of antecedent and boundary conditions and general laws. In history, those general laws are usually taken from psychology, sociology, economics, the natural sciences, etc. Hempel leaves the question open whether genuine historical laws exist. Normally, explanations in history and sociology fail to include an explicit statement of the laws they presuppose. This failure is due to the fact that those laws are part of folk psychology and seem to be tacitly taken for granted. Apart from that, it is difficult to formulate in a sufficiently exact way the underlying general assumptions about, e.g., the outbreak and course of revolutions. Furthermore, most of the regularities in the social sciences are statistical. Therefore, one can expect only explanation sketches in these fields.

The method of empathic understanding that is often claimed to distinguish the social from the natural sciences is neither sufficient nor necessary for explanations in these fields. It is not *sufficient*, because the individual act of understanding is only a heuristic device in order to suggest general psychological hypotheses that might subsequently serve as explanatory principles in the case under consideration. It is not *necessary*, because a historian might well be able to explain the deeds of a paranoiac historic personality by reference to the principles of abnormal psychology without being able to arrive at an emphatic understanding of such a personality.

Concept Formation in Empirical Science

Hempel's work on concept formation in empirical science originates in his early cooperation with Paul Oppenheim and consists basically in the application of principles of logical empiricism to the formation of scientific concepts (Hempel and Oppenheim, 1936). Hempel distinguishes (1) classificatory, (2) comparative, and (3) quantitative concepts. These distinctions normally characterize at the same time the historical development of a science from its initial descriptive 'natural history' stages to its mature theoretical stages. Classification divides a given set or class of objects (e.g., human beings) into subclasses (e.g., beings that suffer from certain mental disorders). Each of these subclasses is defined by means of a certain concept (e.g., schizophrenia) that represents the complex of characteristics essential for the membership in that subclass. Hempel speaks also of 'classificatory types.' The elements of these subclasses are those individuals to whom the respective concept applies (e.g., schizophrenics). The characteristics that form the defining concept are ascertainable fairly directly by observation. Classificatory concepts as well as scientific concepts in general have to fulfill two requirements. They have to (1) be objective and (2) have systematic import. Objectivity is attained first by using operational criteria and (often only partial) operational definitions in a large sense that includes observation as operational and second by eliminating criteria with valuation overtones. Scientific concepts have systematic import if they lend themselves to the formulation of general laws or theoretical principles. Classificatory concepts are a yes-or-no affair, i.e., a certain individual belongs to some subclass or not, whereas comparative concepts admit of more or less with respect to a certain trait (e.g., x is warmer than y with respect to cold–hot). They lead to (quasi) linear orderings as soon as one includes the relation of coincidence (e.g., x is equally warm as y). Hempel calls the extremes of such orderings 'ordering types.' The next step would consist in giving these linear orderings a metric and thus arriving at quantitative concepts (e.g., temperature).

For Hempel the 'ideal types' that have become popular in the social sciences since Max Weber are neither classificatory nor ordering types. Ideal types are interpretative or explanatory schemas that contain a set of empirical generalizations that establish subjectively meaningful connections between different aspects of some kind of phenomenon, e.g., purely rational economic behavior, a capitalist society, and a religious sect. Accordingly, ideal types must be reconstructed not as concepts but as theoretical systems that are intended to provide explanations and therefore must contain testable hypotheses. Despite their lack of clarity and precision, there

thus are also with respect to ideal types no essential methodological differences between psychology or the social sciences and the corresponding methods in the natural sciences (Hempel, 1952, 1967).

Influence, Impact, and Current Significance

Hempel's ideas constitute the mature form of logicoempiricist thinking (Fetzer, 2001). His work on scientific explanation 'stands out as the benchmark to which all subsequent studies of this topic must be referred' (Salmon, 1998, p. vii). His reconstruction of concept formation in empirical science has become a sort of standard for objective procedures of concept formation. It has greatly influenced subsequent editions of the *Diagnostic and Statistical Manual* of the American Psychiatric Association (Houts, 2001). His later work reflects the openness of logical empiricism to new developments in the philosophy of science (particularly T.S. Kuhn's work) that put a strong emphasis on the historical and sociological (pragmatic) factors of the development of science (Hempel, 1979, 2000; Wolters, 2003).

See also: Empiricism, History of; Explanation: Conceptions in the Social Sciences; Idealization, Abstraction, and Ideal Types; Kuhn, Thomas S. (1922–96); Logical Positivism and Logical Empiricism; Mathematical Models in Philosophy of Science.

Bibliography

Fetzer, J.H. (Ed.), 2001. The Philosophy of Carl G. Hempel: Studies in Science, Explanation, and Rationality. Oxford University Press, Oxford.

Hempel, C.G., 1952. Fundamentals of Concept Formation in Empirical Science. University of Chicago Press, Chicago (International Encyclopedia of Unified Science. Vol. II, no. 7).

Hempel, C.G., 1965. Aspects of Scientific Explanation and Other Essays in the Philosophy of Science. The Free Press, New York.

Hempel, C.G., 1967. Philosophy of Natural Science. Prentice-Hall, Englewood Cliffs, NJ.

Hempel, C.G., 1979. Scientific rationality: analytic vs. pragmatic perspectives. In: Geraets, T.F. (Ed.), Rationality To-day – La Rationalité Aujourd'hui. University of Ottawa Press, Ottawa, pp. 46–66.

Hempel, C.G., 2000. Selected Philosophical Essays. In: Jeffrey, R. (Ed.). Cambridge University Press, Cambridge.

Hempel, C.G., Oppenheim, P., 1936. Der Typusbegriff im Lichte der Neuen Logik. Wissenschaftstheoretische Untersuchungen zur Konstitutionsforschung und Psychologie. A. W. Sijthoff, Leiden.

Houts, A.C., 2001. The Diagnostic and Statistical Manual's new white coat and circularity of plausible dysfunctions: response to Wakefield. Behaviour Research and Therapy 38, 315–345.

Salmon, W.C., 1989. Four Decades of Scientific Explanation. University of Minnesota Press, Minneapolis, MN.

Salmon, W.C., 1998. Causality and Explanation. Oxford University Press, Oxford.

Wolters, G., 2003. Carl Gustav Hempel – pragmatic empiricist. In: Parrini, P., Salmon, M.H., Salmon, W.C. (Eds.), Analytical and Continental Aspects of Logical Empiricism. University of Pittsburgh Press, Pittsburgh, pp. 109–122.

Henry, Louis (1911–91)

Etienne van de Walle[†] University of Pennsylvania, Philadelphia, PA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 6660–6661, © 2001, Elsevier Ltd.

Abstract

Louis Henry, a French demographer, made significant contributions to the analysis and measurement of demographic phenomena and to population modeling, particularly in the area of fertility, where he focused on the physiological factors and the determinants of natural fertility. He is credited with developing the basic methods used today in historical demography, including family reconstitution.

Henry was born on 19 February 1911 and was 35 years old when he joined the newly created French Institut National d'Études Démographiques (INED) in 1946; he retired in 1973 and died on 30 December 1991. Most of his work was published under the auspices of INED, either in the journal *Population*, or in the Institute's monograph series. He was one of several retired artillery officers hired by INED because of their rigorous mathematical education at the national *Ecole Polytechnique*; others were Jean Bourgeois-Pichat and Sully Lederman, who would also leave their mark in the field, although they had no formal training in the social or historical sciences. Henry soon joined forces with Paul Vincent in the study of fertility, and this became his area of specialization. He made an early contribution to its measurement by inventing a new index, the parity progression ratio (Henry, 1953). This is defined as the proportion of women with a particular number of children, or parity, who have gone on to have another child, and is computed on the basis of the number of children ever born recorded at a particular age.

Henry intended to go beyond the mere measurement of phenomena, and undertook to decompose reproductive processes into their component factors, what later became known as the proximate determinants of fertility. He decided to study the 'physiological' level of fertility, that is, the fertility that would prevail among married women or couples in the absence of deliberate efforts to limit their number of births; this he called 'natural' fertility. In a contemporary population, however, contraception and abortion that are subject to human decisions are widespread, and it becomes difficult to investigate physiological factors. In his quest to find reliable data on populations where family limitation was not practiced, Henry studied records from the European past, at a time when the recording of vital events was complete and accurate. He first used a series of genealogies on the ruling classes of Geneva, going all the way to the sixteenth century (Henry, 1956); later he analyzed the parish records for a village of Normandy, Crulai, that had been extensively compiled by a local historian, Etienne Gautier (Gautier and Henry, 1958). The records for Geneva demonstrated the sharp contrast between the earliest marriage cohorts, where marital fertility by age seemed to be independent from age at marriage, and the later cohorts where the behavior of couples was clearly modified by the duration of their marriage and the number of children they had attained. It could be shown conclusively that the bourgeoisie of Geneva had adopted family limitation as early as the second half of the seventeenth century. In contrast until the end of the eighteenth century, Crulai, the first parish analyzed by the new method of family reconstitution that Henry had invented, showed the features of a regime of natural fertility, where parity does not affect a woman's subsequent behavior.

In a manual written with Michel Fleury (Fleury and Henry 1956), he laid out – to the minutest details – the rules for the processing of parish data. These rules have been followed faithfully by demographic historians, not only in France but also in other European countries such as England, where they guided the work of the Cambridge Group for the History of Population and Social Structure, and in Canada. Henry would publish another textbook on historical demography (Henry, 1967). Moreover, applying the vast experience accumulated in his early parish studies, he initiated a study of rural France, using a representative sample of French parishes during the eighteenth century. He analyzed this material by aggregative methods, and for a smaller sample of parishes used nominal information to reconstitute families. The project provided the first national series of estimates of fertility, mortality, and nuptiality for a Western European country in the eighteenth century. This body of work makes Louis Henry the uncontested father of historical demography as it is practiced today.

Mining the past for data on populations that did not consciously limit their family size was initially only a tool towards measuring the physiological components of human reproduction. Henry identified the factors to be considered as fecondability, the probability that a conception would result in a live birth; the delay in the resumption of ovulation after a delivery or nonsusceptible period, and total or partial sterility. He provided estimates of the value of these parameters on the basis of his own historical studies of population, but also of various accurate measurements made on other selected populations. His work was picked up by English and American demographers and biometricians working on fertility models, and among them by Mendel Sheps and Jane Menken who edited a collection of translations from Henry's most important articles (Henry, 1972). The seminal importance of Henry's work was acknowledged by the Population Association of America

[†] Deceased.

when it made him the first recipient of the Irene B. Taeuber Award for Excellence in Demographic Research in 1977. Just before his death in 1991, the International Union for the Scientific Study of Population had named him its first Laureate.

The concept of natural fertility is often considered as one of Henry's most significant contributions to population studies. He did not invent the term, but he defined it rigorously and provided means to identify natural fertility regimes. It refers specifically to the behavior of married couples, and is therefore not affected by variations in the proportions married and the age at marriage. Natural fertility is compatible with behavior that may affect the levels of natural fertility (frequency of intercourse, length of breast feeding) or even with attempts to space birth that are not depending on parity. Although the notion remains controversial, it has been used widely in the study of the fertility transition, both in the West and in today's developing countries.

See also: Demography, Early History of; Demography: History Since 1900; Historical Demography.

Bibliography

Fleury, M., Henry, L., 1956. Des registres paroissiaux ä l'histoire de la population. Manuel de dépouillement et d'exploitation de l'état civil ancien. INED, Paris.

Gautier, E., Henry, L., 1958. La population de Crulai, paroisse normande. Étude historique. Presses universitaires de France, Paris.

Henry, L., 1953. Fécondité des mariages. Nouvelle méthode de mesure. Presses universitaires de France, Paris.

Henry, L., 1956. Anciennes familles genevoises. Étude démographique, XVIe–XXe siecle. Presses universitaires de France, Paris.

Henry, L., 1967. Manuel de démographie historique. Droz, Geneva, Switzerland.

Henry, L., 1972. On the Measurement of Human Fertility. Selected Writings of Louis Henry. Elsevier, New York.

Hermeneutics

Oliver R Scholz, Westfälische Wilhelms-Universität Münster, Münster, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The methodology of interpretation has traditionally been called 'hermeneutics.' This article begins with elucidation of the central concept of hermeneutics and distinguishes it from other usages of the term. The subsequent sections clarify the relevant notion of methodology and the leading concepts of hermeneutics, *understanding* and *interpretation*. A glance at the history of hermeneutics focuses on turning points in the historical development of methodologies of interpretation. The concluding section seeks to bring out the main issues in the methods controversies of the nineteenth and twentieth centuries and to suggest strategies to resolve these controversies.

Hermeneutics: Concepts and Tasks

Hermeneutics is a discipline that sets itself upon the task of specifying and justifying *a methodology of interpretation*, originally of texts, but by extension of many other interpretanda. It has also been called the science or art of interpretation.

The Latin half-neologism 'hermeneutica' was introduced into scientific terminology shortly before 1630 by the German philosopher and theologian Johann Conrad Dannhauer (1603–66) as an equivalent of the old Greek term 'hermé-neutike [techné]' (=hermeneutic [art]) that in turn derives from the verb 'hermeneúein,' meaning originally 'to express' or 'to translate,' but since the days of Plato also 'to interpret' (cf Jaeger, 1974). Dannhauer's *Idea boni interpretis et malitiosi calumniatoris* was the first textbook of the new discipline called 'hermeneutica generalis' (Dannhauer, 1630).

It has been a common practice to distinguish between (1) a *general* hermeneutics covering all interpreting disciplines and (2) several *domain- or discipline-specific* subspecies of hermeneutics including, particularly, *hermeneutica sacra*, that is, bible hermeneutics (cf von Reventlow, 1990–2001), and *hermeneutica iuris*, that is, legal hermeneutics (cf Schröder, 2001), and also archaeological (Robert, 1919), historical (Bernheim, 1908: Chapter 5), literary (cf Szondi, 1975; Weimar, 1975), art-historical (Bätschmann, 1984), and musical hermeneutics (Dahlhaus, 1975). While general hermeneutics has traditionally been conceived as a philosophical discipline, commonly as part of logic in a wide sense (including methodology and parts of what is nowadays called epistemology), the different kinds of special hermeneutics have usually been assigned to their respective disciplines (bible hermeneutics to theology, literary hermeneutics to literary studies, and so on), although some philosophers preferred to regard the domain-specific arts of interpretation as mere applications of general hermeneutics to a special object.

Since interpretation is a means to understanding, *general hermeneutics* has two leading concepts, understanding and interpretation, and should, accordingly, consist of two main parts: (1) a theory, or at least a conception, of understanding and (2) a methodology of interpretation. As long as the objects of understanding were restricted to texts, theories of understanding accordingly amounted to conceptions of textual meaning. When the scope is broadened to include other objects of understanding (e.g., persons, actions, and artifacts), more comprehensive theories of understanding are needed. A methodology of interpretation should contain (1) a conception of the aims of interpretation (cf Bühler, 2003; Tepe, 2007) and (2) a system of principles of interpretation, including rules for the critical testing of interpretational hypotheses (cf Bühler, 2003; Scholz, 1999, 2014).

In more recent times, the technical term 'philosophical hermeneutics' has been applied to a philosophical program associated with the names of Martin Heidegger (1889–1976) and Hans-Georg Gadamer (1900–2002) that differs considerably from traditional hermeneutics. Early Heidegger had in mind a phenomenology or existential analytic of *Dasein* (i.e., human existence) that should replace traditional ontology (Heidegger, 1962). Gadamer's project was the more modest one of a phenomenology of hermeneutic experience, with special emphasis on the acquisition of the classical tradition (Gadamer, 1960). While Heidegger boldly appropriated the venerable term 'hermeneutics' for the peculiar brand of proto-anthropology he outlined in *Sein und Zeit*, Gadamer, much more cautiously, tried to use some of Heidegger's ideas to the benefit of philosophical esthetics and the humane studies. In his later years, Gadamer preferred to call his overall project 'hermeneutic philosophy' to avoid confusion with classical hermeneutics. Whatever merits Heidegger's and Gadamer's contributions might have, they should not be praised for 'deepening' the traditional project of hermeneutics; rather, they should be taken to have addressed, by and large, different questions and pursued different goals. Most importantly, they showed no interest in epistemological and methodological questions concerning interpretation. (The very title of Gadamer's *opus magnum* – *Truth and Method* – is meant to hint at its central thesis that truth in art and in the humanities lies outside the control of scientific methods.)

Today, 'hermeneutics' is frequently used to allude to a loose family of antinaturalist approaches in humanities and social sciences. Besides 'hermeneutics,' many social scientists employ terms like 'interpretive social science,' 'interpretive anthropology,' and 'interpretive theory of culture' (Geertz, 1973) or refer, more generally, to an 'interpretive turn' or 'interpretivism' in social sciences (Rabinow and Sullivan, 1987). In American sociology, it has been common practice to use *Verstehen* (Abel, 1929, 1948) or derivative terms such as 'the doctrine of

Verstehen' (Martin, 1969, 1970: Chapter 8) to refer to a technical use of 'understanding' in the work of Wilhelm Dilthey (1833–1911), namely 'empathic understanding,' and especially to the conception of a 'verstehende Soziologie'; that is, interpretive sociology, that was inaugurated by Max Weber (1864–1920) and soon became influential in German and American sociology (cf the critical discussions in Truzzi, 1974 and Martin, 2000). In addition, well-known German scholars following Dilthey – inter alia Eduard Spranger (1882–1963) and Karl Jaspers (1883–1969) – called for a 'verstehende' (interpretive) or 'geisteswissenschaftliche' (humanistic) psychology to counterbalance the dominant research program of an explanatory psychology (cf the critical discussion in Störring, 1928).

The present article focuses on general hermeneutics in the traditional methodological sense and its impact on the humanities and the social sciences. Other so-called 'hermeneutical' or 'interpretive' approaches are taken into account only insofar as they promise to be relevant for the methodological issues.

Methods and Methodology

A *methodology* for a field of activities is a specification of the basic methods that are recommended to be followed in this field. A *method* in turn is a teachable step-by-step procedure that is suited to solve specific practical or theoretical problems. It is useful to distinguish between *special scientific methods* that are specific for particular sciences or a well-defined subgroup of sciences and *general scientific methods* that are applied in all empirical sciences (cf Kraft, 1925). To be sure, a reasonable method need not and in most cases will not provide an algorithm for solving problems in a specified domain; but, insofar as hermeneutics claims scientific status, it should employ accepted scientific methods.

Hermeneutics may be conceived as either a descriptive or a normative discipline. In a *descriptive* vein, it aims at a rational reconstruction of an already existing practice of interpretation that presumably is less than perfect. In a *normative* vein, the goal is to direct, control, and, hopefully, to optimize this imperfect practice. Both projects, the descriptive and the normative, do seem not only legitimate, but also worthwhile.

Understanding and Interpretation

As noted above, the two leading concepts of hermeneutics are interpretation and understanding. Interpretation is a rational activity directed toward the cognitive aim of correct or adequate understanding. (The product of this activity is also called 'an interpretation.')

'To interpret' and 'to understand' are transitive verbs; understanding, like interpretation, has an object. One understands something (x), or one fails to understand it. Sometimes we understand something without effort, simply because we have acquired the requisite ability, and nothing more is needed. Thus, for example, we understand the linguistic utterance 'Piss off!,' since we have learned the English language. But, often, understanding requires work; this is revealed in phrases like 'trying to understand' or 'struggling to understand' (Vendler, 1994). When we cannot understand something immediately, we may nevertheless make an effort to understand it. Interpretation of x is necessary whenever there are special difficulties in understanding x. Such difficulties may be of different sorts (cf Rosenberg, 1981): (1) Thus, difficulty of understanding x is sometimes due to the intrinsic complexity of x; what is needed in such cases is some sort of analysis or decomposition (Ziff, 1972; Moravcsik, 1979). (2) In other cases, x may be difficult to understand because it does not fit coherently in a larger context; here what is required is rather some kind of synthesis or of embedding x in a network of connections. (3) In still other cases, difficulty of understanding x may be a question of the indistinctness of x, so that a careful tracing of boundaries or a better articulation might help.

What kinds of items are possible objects of understanding and interpretation? Traditional hermeneutics concentrated on the interpretation of texts plus, occasionally, things that were supposed to be sufficiently text-like. Reading in a book had been the great paradigm of understanding and interpretation the two most eminent cases being: reading in the book of nature (*liber naturae*) and reading in the Holy Bible, the *liber supernaturalis* (cf Danneberg, 2003).

For centuries, interpreters aimed at understanding and teaching others to understand canonical texts, inter alia Homer's and Hesiod's epics and myths, the works of Plato and Aristotle, Virgil's *Aeneid*, Ovid's *Metamorphoses*, the Holy Bible and the Corpus Iuris. Some of them maintained that in order to understand texts, we should also try to understand their authors. From a more systematic point of view, the objects of understanding may be divided into the following classes (cf Föllesdal, 1981; Scholz, 2010): (1) *persons* and collectives of persons; (2) intentional *attitudes* of persons, especially their cognitive and conative attitudes; (3) individual and collective *actions*; (4) certain *products of actions*; (5) *situations*; and (6) *rules* and rule-constituted social entities, for example, institutions (Wittgenstein, 1958; Nemirow, 1995). With regard to subclass (4), that is, products of actions, it should be emphasized that it encompasses a great variety of items: signs, signals, symbols, words, sentences, texts, arguments, proofs, paintings, maps, diagrams, sculptures, works of architecture, instruments, machines, and other artifacts. The understanding of languages and, more generally, symbol systems (cf Goodman, 1968) has sometimes been treated as a special case of (4) and sometimes as a special case of (6).

Common usage and common sense would certainly add one more item to the list, namely: (7) *natural phenomena*. At any rate, scientists and other curious people have always talked about understanding the universe and its many properties, structures, and mechanisms (cf Cooper, 1994); thus, we want to understand, for example, what osmosis is and how it works, or why there are black holes, and what their typical causes and effects are. But, although many representatives of classical semiotics and hermeneutics (e.g., Meier, 1757) had included natural signs in their domain, in more recent times, class (7) has been contested as a proper object of a theory of understanding (see below).

In addition to the direct object constructions, we use other grammatical structures, for example, 'understanding that,' and, more importantly, indirect question constructions such as 'understanding why' or 'understanding how' (cf Bromberger,

1962, 1992; von Kutschera, 1981: Chapter 2; Vendler, 1984, 1994). Attempts to understand *x*, say an action, may still aim at answering various different questions about *x*, for example, what was done, why it was done, and for what purpose. Thus, attending to the indirect questions that may follow the verb 'to understand' provides us with a more fine-grained classification of types of understanding (e.g., 'understanding what was done' and 'understanding why *x* was done') than the focus on the direct object constructions (e.g., 'understanding action *x*').

Is there something that all objects of understanding have in common? A plausible although still rather vague suggestion is that they all have *structure* (cf Ziff, 1972; Moravcsik, 1979). A bit more precisely, understanding in all cases involves cognition of internal or external relations and interconnections (cf Martin, 1970; Franklin, 1995), especially causal nexuses and nexuses of reference and meaning (cf Mantzavinos, 2005), and also logical relations (as in understanding a proof or an argument), means–end relations (as in understanding actions and artifacts), nomological relations (as in understanding events) and convention-based relations (as in understanding institutions) etc. (Scholz, 2014).

A Glance at the History of Hermeneutics

Hermeneutics has a long and rich history (cf Scholz, 1999: Pt. I; Schönert and Vollhardt, 2005; Detel, 2011; Böhl et al., 2013). The historiography of hermeneutics has been deeply influenced by the insufficient historical studies of Wilhelm Dilthey (1833–1911) and Joachim Wach (1898–1955), and as a result of that has been highly selective and often biased (cf Dilthey, 1900, 1966; Wach, 1926–33). Especially, the elaborate systems of general hermeneutics that emerged long before Friedrich Daniel Ernst Schleiermacher (1768–1834), Dilthey's great hero, have been unduly neglected. Another source of error is the temptation to invent whiggish narratives of a history that leads in an inexorable march of progress from alleged primitive beginnings to the alleged masters of present-day hermeneutics.

Prehistory

While the history of hermeneutics, strictly speaking, begins in the early modern period, its prehistory dates back to Greek antiquity where there is already a highly developed practice and experience of interpreting, inter alia, oracles, dreams, and other signs, myths and epics, philosophical and poetical works, the deeds and works of famous human beings as well as laws, contracts, and testaments. First attempts at explicitly formulating and systematizing rules of interpretation can be found in ancient grammar and rhetoric (cf Pfeiffer, 1968; Eden, 1997), in Galenus and the Hellenistic commentators of Plato and Aristotle (Jaeger, 1974) as well as in Rabbinic and Early Christian exegesis (von Reventlow, 1990–2001).

The most salient and important tendency of ancient exegesis was *allegoresis*, a method of nonliteral interpretation of authoritative texts that contained statements that prima facie seemed either (1) morally or theologically inappropriate, or (2) false. In order to 'rescue' or defend these texts, interpreters were urged to find a deeper sense (*hypónoia*; 'underlying meaning') or allegorical meaning (*allegoría*, from *alla agoreuein*, 'saying something different') and to ascribe it to the offensive passages such that the wording can be maintained, but the alleged imperfections of the text disappear. Allegoresis was practiced widely from the sixth century BC to the Stoic and the Neoplatonistic schools and beyond. It was, however, rejected by others, most notably by Plato, by the Epicureans (with the exception of Lucretius), and by the Alexandrian scholars (Eratosthenes, Aristarchus).

In late antiquity and the Middle Ages, the ancient techniques of allegorical interpretation developed into complex systems of threefold or fourfold written meaning that were applied mainly to the Bible, and also to natural things (cf von Dobschütz, 1921; de Lubac, 1959–64; Ohly, 1977; Brinkmann, 1980).

In the tradition of the eminent commentators, the preliminary questions of interpretation were summarized in the so-called *accessus ad auctores* that preceded commentaries, translations, and editions of Aristotle, Porphyry, Virgil, and other classical authors. Three major traditions, literary, rhetorical, and philosophical, lead to overlapping, but somewhat different typologies of the hermeneutical tasks. The literary tradition (Servius) emphasized questions about the biography of the poet (*vita poetae*), the title of the work (*titulus operis*), the quality of the poem (*qualitas carminis*), the author's intention (*intentio scribentis*), and the division of the text (*numerus librorum*, *ordo librorum*). In the rhetorical tradition (from Cicero to Erasmus), the tasks were derived from an analysis of the *circumstantiae orationis*, in modern terminology from features of the context of a speech or written text (especially place, time, and author) which were usually indicated by the corresponding questions (*quis*, *quid*, *cur*, *quomodo*, *quando*, *ubi*, *quibus facultatibus*). In the philosophical tradition (Alexander Aphrodisias, Ammonius, Simplicius et al.), the commentators also asked questions about the title and the author of the work; a good interpreter, however, should not only concentrate on the purpose or intention of the work (*skopós*; *intentio operis*), its theme, structure, and internal connections, but also take into consideration its utility and ask himself to which part of philosophy it belongs. The philosophical *accessus ad auctores* could be systematized in the Middle ages in terms of Aristotle's doctrine of four causes identifying the author as efficient cause, the themes, and sources of the text as material cause, the proceeding of the author and the form of the text as the formal cause, and the purpose of the work as final cause.

Early History

In the early modern period, the 'rationes' and 'media interpretationis' known from antiquity were assembled and integrated into a new conception of education and methodology (Piccart, 1605a,b; cf Jaeger, 1974; Kessler, 2002). Several Aristotelian and eclectic philosophers, among them Clemens Timpler (1567–1624), Bartholomaeus Keckermann (1571–1609), and Michael Piccart (1574–1620), called for a new discipline that would teach us to be good interpreters and to avoid intentional and unintentional misinterpretations (cf Jaeger, 1974; Sdzuj, 1997). Such a discipline was badly needed for several reasons: (1) Without it, a rich potential source of knowledge, namely all oral and written testimony, was in danger of being spoiled and ruined. (2) It promised to provide a neutral ground from which many theological and confessional controversies could be settled. This new

self-consciousness culminated in the *Idea boni interpretis et malitiosi calumniatoris* (1630) by Johann Conrad Dannhauer (Jaeger, 1974; Sdzuj, 1997). Shortly before he presented the first systematic textbook on general hermeneutics, he had introduced the Latin half-neologism *hermeneutica* as the title of a new *modus sciendi* intended to supplement the Aristotelian *Organon*. It is general in that it is common to all sciences ('omniis scientiis communis') and useful for scholars of all faculties ('omnium facultatum studiosis perquam utilis').

After Dannhauer had established the systematic *locus* of hermeneutics within logic (Dannhauer, 1630, 1634), eminent philosophers such as Johann Clauberg (1622–65), Christian Thomasius (1655–1728), and Christian Wolff (1679–1754) devoted integral parts of their logical writings to general hermeneutics (cf Bühler, 1994; Schönert and Vollhardt, 2005). Thus, from Dannhauer to Bernard Bolzano (1781–1848), it became a regular component of logic textbooks or logic-based treatises, either as an integral part of Aristotelian logic or as a separate practical logic (Scholz, 1999).

The idea of a general hermeneutics evoked the widest interest in the seventeenth and eighteenth centuries, and there emerged a number of alternative outlines in the different philosophical schools (Alexander, 1993; Bühler, 1994; Sdzuj, 1997). I can mention only two conceptions which seem to me both of historical and of intrinsic interest: (1) On the one hand, there is the Rationalist project of a systematic general hermeneutics as envisaged in the Cartesian and Wolffian schools with their methodological ideal of an axiomatic–deductive system. Thus Clauberg, in his sophisticated hermeneutics, made distinctions between the rules of interpretation with regard to their generality, from the most general principles (Clauberg, 1654: §§. 86–89) to the special rules of interpretation (Clauberg, 1654: §§ 90–104). Among the most general principles are what may be called principles of benevolent interpretation (cf Künne, 1990), for example, 'in dubiis benigniora eligere,' that is, in cases of doubt to choose the more benign interpretations, or 'dubium in meliorem partem esse accipiendum,' that is, the doubtful is to be understood for the better (Clauberg, 1654). Or to take an extreme example, Wolffians Alexander Gottlieb Baumgarten (1717–62) and Georg Friedrich Meier (1718–77) emphasized the principle of hermeneutic equity (*aequitas hermeneutica*) – according to which the interpreter should interpret any sign in such a way that his interpretation tallies with the presumption of the perfection of the sign and its author (until the opposite has been proven) – as *the* first principle of general hermeneutics (Meier, 1757; cf Scholz, 1999). (2) On the other hand, there is the empiricist idea of a general hermeneutics as a theory of hermeneutic probability that was elaborated by the school of Christian Thomasius (1655–1728) and Andreas Rüdiger (1673–1731). Thus, Christian August Crusius (1715–75), in his excellent hermeneutics, enumerated the most general presumptions of hermeneutic probability, among them, defeasible presumptions of logical consistency and means–ends rationality (Crusius, 1747; cf Scholz, 1999).

It should be clear by now that when the theologian Friedrich Daniel Ernst Schleiermacher (1768–1834) propagated, with great emphasis, a 'general hermeneutics' at the beginning of the nineteenth century, he was (albeit in his own distinctive way) reinventing the wheel. While his conception of general hermeneutics may be interesting in its own right, his claim that such a discipline did not already exist was simply false. Moreover, most of the rules he recommended were taken over from contemporary textbooks on bible hermeneutics. Any text or speech is to be understood, on the one hand, as part of a language and, on the other hand, as part of an individual life (Schleiermacher, 1977). Schleiermacher's conception of psychological interpretation and his emphasis on the individuality and originality of the author had a great influence on August Boeckh (1785–1867), Wilhelm Dilthey (1833–1911), and other modern theoreticians of hermeneutics.

Dilthey took an interest in the history of hermeneutics for two reasons: in the first place, because he was working on a comprehensive biography of Schleiermacher (cf Dilthey, 1966) and, secondly, because he thought hermeneutics developed his own project: a laying of the foundations of the new *Geisteswissenschaften* (history, national economy, jurisprudence, political science, religious studies, literary studies, art history, musicology, and psychology (!)) which he tried to demarcate strictly from the natural sciences (Dilthey, 1927). The common object of these human sciences is mankind and its history, the 'human–social–historical reality' or, in Hegelian terms, the world of objective spirit. The objects of the humanities are 'constituted' in a way that differs fundamentally from the way the objects of the natural sciences are 'constituted.' To reconstruct the constitution of the historical world is the central task of a 'logic' of the *Geisteswissenschaften*. According to Dilthey, the leading concepts of the human sciences are life or lived experience ('Erlebnis'), expression, and understanding. While we are only able to explain the connections of the physical world with the help of abstract concepts and theoretical extrapolations, our access to the mental and spiritual world is immediate: we are able to understand manifestations of life because we know the connection between the outward expression and something inner that is thereby expressed from our own case. In the elementary forms of understanding, we infer (by some kind of analogical reasoning) from a single manifestation of life that it is the expression of a certain inner content; in the higher forms of understanding, we try to reconstruct the totality of an individual life, for example, the life of Martin Luther, via a special form of induction that is, unfortunately, not analyzed in detail by Dilthey. It is in this context that he underlined the importance of reproduction ('Nachbildung') and re-living ('Nacherleben') for understanding other minds and their works. While the higher forms of understanding are initially based on the personal genius of the interpreter, for the purposes of historical sciences they have to be developed into a technique, the art of interpretation. According to Dilthey, the science of this art is hermeneutics (Dilthey, 1927).

Recent Developments

The Methods Controversies

One consequence of the emergence of the 'Geisteswissenschaften' (moral sciences; humanities) or 'Kulturwissenschaften' (cultural studies) was a long series of methods controversies ('Methodenstreite') that continue till today. Time and again, self-styled antipositivists and antinaturalists reacted to so-called positivists

and naturalists. (Since 'positivism' and 'naturalism' are mainly used in a pejorative sense, I prefer the more neutral terms 'Methodological Monism' and 'Methodological Dualism' to designate the main issue between the two camps: whether there is a methodological unity or a strict methodological separation.)

In the early stages of the controversy, Auguste Comte (1798–1857), John Stuart Mill (1806–73), and the historian Henry Thomas Buckle (1821–62) counted as the arch-positivists. In Germany, there was, in addition, the Leipzig Circle that included the experimental psychologist Wilhelm Wundt (1832–1920) and the historian Karl Gotthard Lamprecht (1856–1915). Comte, Mill, and their followers pleaded for applying the general 'logic' of scientific research to the humanities and social sciences. Mill expressed the hope that the desideratum of a 'Logic of the Moral Sciences' could be met by generalizing and adjusting the methodology that was developed for the natural sciences (Mill, 1843: Book VI, Chapter I, p. 1). In other words, the 'positivists' pleaded for the methodological unity of the sciences.

Their opponents insisted on a special methodological status of the *Geisteswissenschaften*, and in particular on a high degree of methodological autonomy of, say, history. Despite this fundamental consensus, they offered rather different suggestions for demarcating the two groups of sciences (cf Becher, 1921). Johann Gustav Droysen (1808–84) and Wilhelm Dilthey (1833–1911) famously opposed the method of 'Erklären' (explanation) as against the method of 'Verstehen' (understanding) (Droysen, 1977; Dilthey, 1924). The neo-Kantian philosophers Wilhelm Windelband (1848–1915) and Heinrich Rickert (1863–1936) focused on the respective aims of each group of sciences and especially on the methods of concept formation and judgment; according to them, the logic of history and, more generally, the logic of the *Kulturwissenschaften* is characterized by an idiographic interest in singular judgments about the past (Windelband, 1894) or in an individualizing form of concept formation (Rickert, 1929). In addition, they underlined that the objects of cultural studies are essentially related to shared cultural values ('wertbezogen').

While Dilthey and Rickert in their most exaggerated statements pleaded for a strict separation of the *Geistes-* or *Kulturwissenschaften* from the natural sciences, some contemporaries argued for more subtly differentiated positions. Thus, Max Weber (1864–1920) emphasized, on the one hand, that the sociologist must first try to understand human action in terms of the categories of the actors, their goals, and beliefs. On the other hand, he stressed that understanding of this kind has to be controlled by the usual methods of causal attribution (Weber, 1922). Thus, he avoided claiming an opposition between understanding and explanation. On the contrary, his 'verstehende Soziologie' (interpretive sociology) leads from interpretive understanding to a causal analysis of the actual processes (Ringer, 1997). His causalist methodology makes it possible to explain the conduct of actors that are quite unlike us. Against an overestimation of Diltheyan understanding qua congenial reliving, Weber, echoing Georg Simmel (1858–1918), reminded us that one need not be Caesar to understand Caesar (Weber, 1922; cf Simmel, 1905). Nevertheless, since Weber limits the subject matter of sociology to subjectively meaningful action, his intermediate position in the methods controversies is still too restrictive for the whole spectrum of social scientific understanding. Social scientists should try to explain behavior, for example, traditional behavior, that may no longer be subjectively meaningful; and they should be allowed to draw on regularities discovered by scientists from other disciplines in order to explain individual and collective behavior (Martin, 2000).

To the already mentioned alleged dichotomies (explanation versus understanding; nomothetic versus idiographic sciences), more recent methods that dualists have added: construing human actions as mere events from outside versus reenacting the thoughts of the agent in the historian's mind (Collingwood, 1946); covering law explanations versus rationalizing explanations (Dray, 1957), causes versus reasons (various ordinary language philosophers); causal explanations versus teleological explanations (von Wright, 1971); and so on.

Against Separation

One strategy to avoid the separation of disciplines implied by Methodological Dualism is to question the alleged dichotomies and their theoretical presuppositions. Thus, it has been argued that understanding is not a method in itself that can be contrasted with other methods, but the end at which all methods aim (Lundberg, 1939), or that the reasons for acting can at the same time be the causes of the action (Davidson, 1963). In addition, it has been argued that special mental operations aiming at understanding, for example, operations of reliving, reenacting, or empathic understanding, may be useful heuristic techniques, but cannot replace reliable methods for validating hypotheses (Hempel, 1942, 1965; Abel, 1948; Rudner, 1966). The most important objection to the idea that human action has to be understood 'from inside' or 'from the perspective of the agent' is that it would unduly restrict historical and social scientific understanding (Martin, 2000).

Another strategy to defend a fundamental methodological unity of the sciences is to focus on the most general scientific methods (cf Kraft, 1925), especially the methods for the critical testing and validating of hypotheses. At least since the nineteenth century most philosophers of science have favored some version of the Hypothetico-Deductive Method: Begin by forming hypotheses (including hypotheses about unobservable entities, processes, and forces), then deduce consequences about observable phenomena from these hypotheses, and finally examine whether these consequences fit in with the experiential data. In view of the generality and wide applicability of this method, it is not surprising that it has been suggested as the method of hermeneutics (cf Hirsch, 1967; Göttner, 1973; Föllesdal, 1979; Mantzavinos, 2005). And, it can hardly be denied that many interpreters have made use of this general kind of reasoning in their interpretive efforts. In my view, the Method of Inference to the Best Explanation is even more widely applicable: Ask yourself what would explain the available evidence and take the ability of a hypothesis to explain the relevant data as an indicator that the hypothesis is probably correct (cf Lipton, 2004). Applications of this general method are ubiquitous in all forms of inquiry. Moreover, it promises to be free from the serious logical and epistemological

problems that emerge when we try to make the Hypothetico-Deductive Method precise (Scholz, 2014).

A third strategy to overcome the method's controversies is to opt for an eclectic methodological pluralism that encourages all methodologies and theoretical approaches that seem to be relevant to the questions being asked and the aims pursued (cf Little, 1991; Martin, 2000). This strategy accepts the diversity of methods without separating the disciplines by an unbridgeable gulf.

See also: Dilthey, Wilhelm (1833–1911); Hermeneutics, History of; Interpretation in Anthropology.

Bibliography

Abel, T., 1929. Systematic Sociology in Germany. Columbia University Press, New York.

Abel, T., 1948. The operation called *Verstehen*. American Journal of Sociology 54, 211–218.

Albert, H., 1994. Kritik der reinen Hermeneutik. Mohr (Siebeck), Tübingen.

Alexander, W., 1993. Hermeneutica Generalis. Zur Konzeption und Entwicklung der allgemeinen Verstehenslehre im 17. und 18. Jahrhundert, Stuttgart.

Bätschmann, O., 1984. Einführung in die kunstgeschichtliche Hermeneutik. Die Auslegung von Bildern. Wissenschaftliche Buchgesellschaft, Darmstadt.

Becher, E., 1921. Geisteswissenschaften und Naturwissenschaften. Untersuchungen zur Theorie und Einteilung der Realwissenschaften. Duncker & Humblot, Munich and Leipzig.

Bernheim, E., 1908. Lehrbuch der historischen Methode und der Geschichtsphilosophie, fifth and sixth ed. Duncker & Humblot, Leipzig.

Böhl, M., Reinhard, W., Walter, P. (Eds.), 2013. Hermeneutik. Die Geschichte der abendländischen Textauslegung von der Antike bis zur Gegenwart (Dichtung, Bibel, Recht, Geschichte, Philosophie). Wien – Köln – Weimar: Böhlau Verlag.

Brinkmann, H., 1980. Mittelalterliche Hermeneutik. Niemeyer, Tübingen.

Bromberger, S., 1962. An approach to explanation. In: Butler, R.S. (Ed.), Analytical Philosophy – Second Series. Blackwell, Oxford, pp. 72–105.

Bromberger, S., 1992. On What We Know We Don't Know. Explanation, Theory, Linguistics, and How Questions Shape Them. The University of Chicago Press, Chicago and London.

Bühler, A., 1999. Die Vielfalt des Interpretierens. Analyse & Kritik 21, 117–137.

Bühler, A. (Ed.), 1994. Unzeitgemäße Hermeneutik. Verstehen und Interpretation im Denken der Aufklärung. Vittorio Klostermann, Frankfurt am Main.

Bühler, A. (Ed.), 2003. Hermeneutik. Basistexte zur Einführung in die wissenschaftstheoretischen Grundlagen von Verstehen und Interpretation. Synchron, Heidelberg.

Cataldi Madonna, Luigi (Ed.), 2013. Naturalistische Hermeneutik. Ein neues Paradigma des Verstehens und Interpretierens. Könighausen & Neumann, Würzburg.

Chladenius, J.M., 1742. Einleitung zur richtigen Auslegung vernünftiger Reden und Schriften. Lankisch, Leipzig.

Clauberg, J., 1654. Logica vetus et nova. Elsevier, Amsterdam. Reprinted in: Clauberg, J., 1691. In: Schalbruch, J.T. (Ed.), Opera omnia philosophica. P. & I. Blaev, Amsterdam.

Collingwood, R.G., 1946. The Idea of History. Clarendon Press, Oxford.

Cooper, N., 1994. Understanding. Proceedings of the Aristotelian Society LXVIII (Suppl.), 1–26.

Cooper, N., 1995. The epistemology of understanding. Inquiry 38, 205–215.

Crusius, C.A., 1747. Weg zur Gewißheit und Zuverlässigkeit der menschlichen Erkenntnis. J.F. Gleditsch, Leipzig.

Dahlhaus, C. (Ed.), 1975. Beiträge zur musikalischen Hermeneutik. Bosse, Regensburg.

Dallmayr, F.R., McCarthy, T.A. (Eds.), 1977. Understanding and Social Inquiry. University of Notre Dame Press, Notre Dame.

Danneberg, L., 2003. Die Anatomie des Text-Körpers und Natur-Körpers: das Lesen im liber naturalis und supernaturalis. de Gruyter, Berlin and New York.

Danneberg, L., 2005. Vom' grammaticus' und' logicus' über den' analyticus' zum' hermeneuticus'. In: Schoenert, J., Vollhardt, F. (Eds.), Geschichte der Hermeneutik und die Methodik der textinterpretierenden Disziplinen. de Gruyter, Berlin and New York, pp. 281–363.

Dannhauer, J.C., 1630. Idea boni interpretis et malitiosi calumniatoris. Wilhelm Christian Glaser, Straßburg. Reprint of the Fourth Edition Straßburg, 1652. Olms, Hildesheim; Zürich; and New York (2004).

Dannhauer, J.C., 1634. Epitome dialectica. Wilhelm Christian Glaser, Straßburg.

Dannhauer, J.C., 1654. Hermeneutica sacra sive methodus exponendarum s. literarum proposita et vindicata. Wilhelm Christian Glaser, Straßburg.

Davidson, D., 1963. Actions, reasons, and causes. Journal of Philosophy 60, 685–700. Reprinted in: Davidson, D. Essays on Actions and Events. Clarendon Press, Oxford, 3–19 (1980).

Davidson, D., 1984. Inquiries into Truth and Interpretation. Clarendon Press, Oxford.

Detel, W., 2011. Geist und Verstehen. Historische Grundlagen einer modernen Hermeneutik. Klostermann, Frankfurt am Main.

Dilthey, W., 1900. Die Entstehung der Hermeneutik. In: Philosophische Abhandlungen – Christoph Sigwart zu seinem siebzigsten Geburtstag gewidmet. Mohr, Tübingen, pp. 185–202.

Dilthey, W., 1924. Gesammelte Schriften V: Die geistige Welt. Teubner, Leipzig and Berlin.

Dilthey, W., 1927. Gesammelte Schriften VII: Der Aufbau der geschichtlichen Welt in den Geisteswissenschaften. Teubner, Leipzig and Berlin.

Dilthey, W., 1966. Gesammelte Schriften XIV/2: Leben Schleiermachers. de Gruyter, Berlin.

von Dobschütz, E., 1921. Vom vierfachen Schriftsinn. Die Geschichte einer Theorie. In: Harnack-Ehrung. Beiträge zur Kirchengeschichte ihrem Lehrer Adolf von Harnack zu seinem siebzigsten Geburtstag dargebracht von seinen Schülern, Leipzig, pp. 1–13.

Dray, W., 1957. Laws and Explanation in History. Clarendon Press, Oxford.

Droysen, J.G., 1977. Historik. Frommann-Holzboog, Stuttgart-Bad Cannstatt.

Eden, K., 1997. Hermeneutics and the Rhetorical Tradition. Yale University Press, New Haven and London.

Föllesdal, D., 1979. Hermeneutics and the hypothetico-deductive method. Dialectica 33, 319–336.

Föllesdal, D., 1981. Understanding and rationality. In: Parret, H., Bouveresse, J. (Eds.), Meaning and Understanding. de Gruyter, Berlin and New York, pp. 155–168.

Frank, G., Meier-Oeser, S. (Eds.), 2011. Hermeneutik, Methodenlehre, Exegese. Zur Theorie der Interpretation in der Frühen Neuzeit. Frommann-Holzboog, Stuttgart-Bad Cannstatt.

Franklin, R.L., 1995. The Search for Understanding. Peter Lang, New York.

Gadamer, H.-G., 1960. Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik, fourth ed. Mohr, Tübingen. (1975).

Gadamer, H.-G., 2003. Truth and Method (J. Weinsheimer and D.G. Marshall, Trans.), second rev. ed. Continuum, New York.

Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.

Göttner, H., 1973. Logik der Interpretation. Wilhelm Fink, München.

Gomperz, H., 1929. Über Sinn und Sinngebilde. Verstehen und Erklären. Mohr (Siebeck), Tübingen.

Gomperz, H., 1939. Interpretation: Logical Analysis of a Method of Historical Research. van Stockum & Zoon, The Hague.

Goodman, N., 1968. Languages of Art. An Approach to a Theory of Symbols. Hackett, Indianapolis.

Heidegger, M., 1962. Being and Time (J. Macquarrie and E. Robinson, Trans.). Harper & Row, New York.

Hempel, C.G., 1942. The function of general laws in history. Journal of Philosophy 39, 35–48.

Hempel, C.G., 1965. Aspects of Scientific Explanation and Other Essays in the Philosophy of Science. The Free Press, New York.

Hirsch, E.D., 1967. Validity in Interpretation. Yale University Press, New Haven and London.

Hübener, W., 1985. Schleiermacher und die hermeneutische Tradition. In: Selge, K.V. (Ed.), Schleiermacher-Archiv, Internationaler Schleiermacher-Kongress Berlin 1984, Band 1. de Gruyter, Berlin and New York, pp. 561–574.

Jaeger, H.-E.H., 1974. Studien zur Frühgeschichte der Hermeneutik. Archiv für Begriffsgeschichte 18, 35–84.

Kessler, E., 2002. Logica universalis und Hermeneutica universalis. In: Piaia, G. (Ed.), La presenza dell'aristotelismo padovano nella filosofia della prima modernità. Editrice Antenore, Rome and Padua, pp. 133–171.

Kincaid, H., 1996. Philosophical Foundations of the Social Sciences. Cambridge University Press, Cambridge.

Kraft, V., 1925. Die Grundformen der wissenschaftlichen Methoden, second rev. ed. Wien & Leipzig, Hölder-Pichler-Tempsky. Verlag der Österreichischen Akademie der Wissenschaften, Wien (1973).

Künne, W., 1990. Prinzipien der wohlwollenden Interpretation. In: Forum für Philosophie Bad Homburg (Ed.), Intentionalität und Verstehen. Suhrkamp, Frankfurt am Main, pp. 212–236.

von Kutschera, F., 1981. Grundfragen der Erkenntnistheorie. de Gruyter, Berlin and New York.

Lewis, D., 1974. Radical interpretation. Synthese 27, 331–344.

Little, D., 1991. Varieties of Social Explanation. Westview Press, Boulder, CO.

Lipton, P., 2004. Inference to the Best Explanation, Second edition. Routledge, London and New York.
de Lubac, H. de, 1959–64. Exégèse médiévale. Les quatre sens de l'écriture. Aubier, Paris.
Lundberg, George A., 1939. Foundations of Sociology. Macmillan, New York.
Mantzavinos, C., 2005. Naturalistic Hermeneutics. Cambridge University Press, Cambridge.
Martin, J.R., 1969. Another look at the doctrine of Verstehen. The British Journal for the Philosophy of Science 20, 53–67.
Martin, J.R., 1970. Explaining, Understanding and Teaching. McGraw Hill, New York.
Martin, M., 2000. Verstehen. The Uses of Understanding in Social Science. Transaction Publishers, New Brunswick and London.
Meier, G.F., 1757. Versuch einer allgemeinen Auslegungskunst. Hemmerde, Halle.
Mill, J. St, 1843. A System of Logic, Ratiocinative and Inductive. Longmans, London.
Moravcsik, J.M., 1979. Understanding. Dialectica 33, 201–216.
Nemirow, L.E., 1995. Understanding rules. The Journal of Philosophy 92, 28–43.
O'Hear, A. (Ed.), 1996. Verstehen and Humane Understanding. Cambridge University Press, Cambridge.
Ohly, F., 1977. Schriften zur mittelalterlichen Bedeutungsforschung. Wissenschaftliche Buchgesellschaft, Darmstadt.
Parret, H., Bouveresse, J. (Eds.), 1981. Meaning and Understanding. de Gruyter, Berlin and New York.
Pfeiffer, R., 1968. History of Classical Scholarship from the Beginnings to the End of the Hellenistic Age. Clarendon Press, Oxford.
Piccart, M., 1605a. Oratio de optima ratione interpretandi. Sebastian Körber, Nuremberg.
Piccart, M., 1605b. Oratio de ratione interpretandi. In: Piccart, M. (Ed.), Isagoge in lectionem Aristotelis, hoc est: Hypotyposis totius philosophiae Aristotelis. Sebastian Körber, Nuremberg.
Rabinow, P., Sullivan, P.M. (Eds.), 1987. Interpretive Social Science: A Second Look. University of California Press, Berkeley.
von Reventlow, H., 1990–2001. Epochen der Bibelauslegung, 4 vols. C.H. Beck, Munich.
Rickert, H., 1929. Die Grenzen der naturwissenschaftlichen Begriffsbildung. Eine logische Einleitung in die historischen Wissenschaften, fifth ed. Mohr (Siebeck), Tübingen.
Ringer, F., 1997. Max Weber's Methodology: The Unification of the Cultural and Social Sciences. Harvard University Press, Cambridge, MA and London.
Robert, C., 1919. Archäologische Hermeneutik – Anleitung zur Deutung klassischer Bildwerke. Weidmannsche Buchhandlung, Berlin.
Rosenberg, J.F., 1981. On understanding the difficulty in understanding understanding. In: Parret, H., Bouveresse, J. (Eds.), Meaning and Understanding. de Gruyter, Berlin and New York, pp. 29–43.
Rudner, R.S., 1966. Philosophy of Social Science. Prentice-Hall, Englewood Cliffs.
Ryle, G., 1949. The Concept of Mind. Hutchinson, London.
Schleiermacher, F.D.E., 1977. In: With an Introduction by Frank, M. (Ed.), Hermeneutik und Kritik. Suhrkamp, Frankfurt am Main.
Schönert, J., Vollhardt, F. (Eds.), 2005. Geschichte der Hermeneutik und die Methodik der textinterpretierenden Disziplinen. de Gruyter, Berlin and New York.
Scholz, O.R., 1999. Verstehen und Rationalität. Untersuchungen zu den Grundlagen von Hermeneutik und Sprachphilosophie, second ed. Vittorio Klostermann, Frankfurt am Main. (2001).
Scholz, O.R., 2010. Verstehen. In: Sandkühler, H.J. (Ed.), Enzyklopädie Philosophie, second ed. Verlag Felix Meiner, Hamburg, pp. 2905–2909.
Scholz, O.R., 2014. Texte interpretieren – Daten, Hypothesen und Methoden. In: Borkowski, Jan, Descher, Stefan, Ferder, Felicitas, Heine, Philipp David (Eds.), Literatur interpretieren. Mentis, Münster.
Schröder, J., 2001. Recht als Wissenschaft. Geschichte der juristischen Methode vom Humanismus bis zur historischen Schule. C.H. Beck, München.
Schurz, G., 2004. Erklären und Verstehen: Tradition, Transformation und Aktualität einer klassischen Kontroverse. In: Handbuch der Kulturwissenschaften, Band 2. Metzler, Stuttgart and Weimar, pp. 156–174.
Sdzuj, R., 1997. Historische Studien zur Interpretationsmethodologie der Frühen Neuzeit. Könighausen & Neumann, Würzburg.
Simmel, G., 1905. Die Probleme der Geschichtsphilosophie. Eine erkenntnistheoretische Studie, second rev. ed. Duncker & Humblot, Leipzig.
Störring, G., 1928. Die Frage der geisteswissenschaftlichen und verstehenden Psychologie. Eine Streitschrift. Akademische Verlagsgesellschaft, Leipzig.
Szondi, P., 1975. Einführung in die literarische Hermeneutik. Suhrkamp, Frankfurt am Main.
Tepe, P., 2007. Kognitive Hermeneutik. Könighausen & Neumann, Würzburg.
Truzzi, M. (Ed.), 1974. Verstehen: Subjective Understanding in the Social Sciences. Addison-Wesley Publishing Company, Reading, MA.
Vendler, Z., 1984. Understanding people. In: Shweder, R.A., LeVine, R.A. (Eds.), Culture Theory. Cambridge University Press, Cambridge, pp. 200–213.
Vendler, Z., 1994. Understanding misunderstanding. In: Language, Mind and Art. Essays in Appreciation and Analysis, in Honor of Paul Ziff. Reidel, Dordrecht, pp. 9–21.
Wach, J., 1926–33. Das Verstehen. Grundzüge einer Geschichte der hermeneutischen Theorien im 19. Jahrhundert, 3 vols. Mohr, Tübingen.
Weber, M., 1922. Gesammelte Aufsätze zur Wissenschaftslehre, fourth ed. Mohr (Siebeck), Tübingen. (1973).
Weimar, K., 1975. Historische Einleitung zur literaturwissenschaftlichen Hermeneutik. Mohr (Siebeck), Tübingen.
Winch, P., 1958. The Idea of a Social Science and Its Relation to Philosophy. Routledge, London.
Windelband, W., 1894. Geschichte und Naturwissenschaft. In: Windelband, W. (Ed.), 1915. Präludien. Aufsätze und Reden zur Philosophie und ihrer Geschichte, vol. 2, fifth expanded ed. Mohr, Tübingen, pp. 136–160.
Wittgenstein, L., 1958. Philosophical Investigations. Basil Blackwell, Oxford.
von Wright, G.H., 1971. Explanation and Understanding. Cornell University Press, Ithaca, NY.
Ziff, P., 1972. Understanding Understanding. Cornell University Press, Ithaca and London.

Hermeneutics, History of

William Outhwaite, Newcastle University, Newcastle upon Tyne, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term 'hermeneutics' refers to the science, art, or technique of interpretation, of written texts and, by extension, of human actions and other social phenomena. In this extended sense, it has come to denote a variety of approaches in the social and behavioral sciences ('*verstehende*' sociology, symbolic interactionism, social phenomenology, social constructionism, ethnomethodology). Hermeneutic or phenomenological approaches now coexist and are even combined with more structural conceptions of social science or ones modeled more closely on the natural sciences. Hermeneutics in this broader sense continues to exist as a major research tradition in the humanities, as well as a minority one in the social and behavioral sciences.

The term 'hermeneutics' refers to the science, art, or technique of interpretation, paradigmatically of written texts but by extension of human actions and other social phenomena. In this extended sense, it has come to denote a variety of approaches in the social and behavioral sciences. The word refers to the Greek messenger God Hermes, and reflection on the problems of interpretation and criteria for the truth, validity, or adequacy of interpretations, goes back to ancient Greek thought in the European tradition and a similar historical distance in the other world civilizations. We can consider this as the prehistory of hermeneutics, revived in the philological criticism of classical texts in the Renaissance, the interpretation of Roman law, the interpretation of the Bible in (especially Protestant) Christianity, and the philosophical analysis of texts.

The systematization of hermeneutics occurred largely in German-speaking Europe in the course of the nineteenth century, although two earlier thinkers deserve particular mention. Johann Martin Chladenius (1710–59) published in 1742 a general account of interpretation, which brought together traditional Aristotelian rhetoric and more recent Enlightenment themes, developing in particular the notion of point of view or perspective (*Sehe-Punkt*), crucial to later philosophy of history and social theory through to Karl Mannheim (1893–1947) and beyond (Mueller-Vollmer, 1986). Second, Giambattista Vico (1688–1744) formulated the basic principle that our knowledge of what we ourselves have made (individually or collectively) is different from what we have not made. The world of human society and culture is in some sense 'our' product.

This distinction, crucial to later hermeneutics and 'hermeneutic' social science, between what we know from the inside and what we know because we have learned about it is taken up in a rather more speculative form in Hegel's differentiation between reason (*Vernunft*) and the understanding (*Verstand*), although the maker here is the world spirit coming to recognize its own productions (including, ultimately, the world itself) and its learning processes as rational, in contrast to the essentially contingent states and relations found in nature and described by the mathematical and natural sciences. Something more like Vico's idea returns again in the second half of the nineteenth century, when Hegel's concept of objective mind is extracted from its surrounding developmental 'grand narrative' and treated more as a descriptive category.

It was Friedrich Schleiermacher (1768–1834), following Friedrich Schlegel, who consolidated hermeneutics in a systematic form, establishing the term understanding (*Verstehen*), which had been used by Chladenius and which has survived as standard usage in English-language social science discourse, and making it central to interpretation – understood as a more systematic activity. Interpretation here involves both the linguistic understanding of meaning and a psychological understanding of the author's intention. Schleiermacher introduced the distinction between technical and more developed or speculative forms of interpretation and formulated the much-discussed principle that it should be possible to understand an author better than he understood himself. Schleiermacher's contribution was made the central and culminating point of the account of 'the rise of hermeneutics' given by Wilhelm Dilthey (1833–1911); with Dilthey, hermeneutics becomes central to the self-definition of what he called the human sciences or *Geisteswissenschaften*. In these sciences, as Dilthey put it, the mental activity of humans and of some other animals, and its products, can be understood.

Dilthey and his contemporary, the philosopher of history J. Gustav Droysen (1808–84), developed what we would now call a research program for history and the other human sciences based on the distinctiveness of human psychic expressions and the understanding of those expressions. In a move that was to become a definitional feature of later interpretive social science, Dilthey, like Schleiermacher, emphasized the continuity between everyday understanding and more formal processes of interpretation. His distinction between the natural and human sciences was developed in large part in opposition to Comtean positivism, which had become influential, even in the German-speaking countries, by the middle of the nineteenth century. In a parallel, but more methodological formulation, two other neo-Kantian thinkers, Wilhelm Windelband (1848–1915) and Heinrich Rickert (1863–1936), argued that the study of culture is essentially concerned with individual processes and relating them to shared human values, whereas the natural sciences are concerned with general laws about objects that are essentially remote from questions of value. We are interested, for example, in the French Revolution, not just as a member of a class of revolutions exhibiting certain common features (this would

be, for Rickert, a natural-scientific mode of approaching it), but as a unique event embodying, and perhaps also violating, certain crucial human values.

This opposition between positivism and methodological dualism, and more particularly between causal explanation, analyzed in terms of universal regularities, and 'understanding' comes to structure the emergent human or social sciences, as the term 'culture' increasingly gives way to 'society' or 'sociation' (*Vergesellschaftung*). There is a fairly strong line of influence from Rickert to Max Weber, both directly and also through his friend Georg Simmel, who discusses our knowledge of the social world in terms that foreshadow social constructionist theory (*see* Constructivism/Constructionism: Methodology).

Max Weber's personal development from economic and legal historian to sociologist continues the encounter between hermeneutics and the social sciences, which had been a dominant feature of the late nineteenth century, and he also develops a middle position on the issue of methodological dualism, as he had earlier on the related opposition in economic theory (the *Methodenstreit*) between the pursuit of a systematic law and a more interdisciplinary and historical approach. For Weber, as he put it in an early essay, "the course of human action and every sort of human expressions are susceptible to meaningful interpretation; in the case of other objects, an analogous interpretation would only be possible in the sphere of metaphysics" (Weber, 2012: p. 10). He therefore later defines sociology, in the first sentence of his major work *Economy and Society*, as a science that aims at an interpretative understanding of action in order thereby to understand its course and effects. Whether by this Weber means that explanatory understanding is itself a form of causal explanation, or merely complementary to it, the crucial point for him is that explanations of social phenomena must be both 'causally adequate' and 'meaningfully adequate.'

For Weber, then, our access to knowledge of the social world is importantly different from our knowledge of nature. However, it is not, in his view, any less objective. He heroically attempts to hold together Rickert's principle that our perspectives on cultural phenomena, and our knowledge of them, are shaped by values (based for Weber on ultimately ungroundable existential choices), with the idea that the social sciences can attain a bedrock of solid and 'value-free' knowledge, which would have to be accepted, as he sometimes curiously puts it, 'even by a Chinese (Weber, 2012: p. 105)'. Weber insists that it is the intentions and purposes (subjective meanings) of human actors that define their actions and therefore have to be understood by the historian or sociologist, but he moves rapidly to the construction of a system of ideal types of action orientation less limited than that found in economic theory but still substantially dependent on it.

Weber's synthesis was pulled apart from both sides in the decades following his early death in 1920. At one pole, there was now a more stridently naturalistic and indeed reductionist variant of positivism: the logical empiricism of the Vienna Circle, in whose 'unified science' the statements of all sciences should be ultimately reducible to material–object language or to statements in physics; *Verstehen* was of no more importance, in Otto Neurath's vigorous formulation, than a good cup of coffee, which sustains the social scientist (Neurath, 1973: p. 357).

From the other direction, Alfred Schutz initiated the tradition of social phenomenology, and what is often loosely called 'hermeneutics' in the context of the social sciences, with a book published in Vienna in 1932 with the title *The Meaningful Constitution of the Social World*. Schutz felt that the problem with Weber's ideal types was not that they were insufficiently scientific, but precisely the opposite: Weber was too quick to impose them on the phenomena he described, paying insufficient attention to their grounding in acts of typification performed by ordinary members of society. For Schutz, the social scientist is merely constructing second-order typifications based on those already carried out in the lifeworld. This theme was taken up by the phenomenologist Aron Gurwitsch in the United States, where Schutz had also settled, and by Peter Berger and Thomas Luckmann (1967), who finally put the term 'social construction' on the map, offering in the guise of a sociology of knowledge a paradigm more directly adapted to use in social research. By the time Schutz's first book was republished, in Austria in 1960 and in the United States in 1967, the way was prepared by his own later work and, in the English-speaking world, by Berger and Luckmann (1967).

The middle decades of the twentieth century replay many of the debates, which had dominated the second half of the nineteenth century. Droysen's complaint, in a letter of 1852, about the rise of what he called 'crass positivism,' and his crusade against it in a course, which he taught from 1857 onward, is echoed by antinaturalist social science just over a hundred years later. Social phenomenology also of course had affinities with the well-established minority tradition in North American social science of symbolic interactionism; this also experienced a certain resurgence in the 1960s with, for example, the republication of the work of G.H. Mead and that of Herbert Blumer, and also the publication of various studies by Erving Goffman (Helle, 2005). But where the interactionists tended not to spend time on formal critiques of empiricism, Harold Garfinkel's ethnomethodology (Garfinkel, 1967) was more aggressive. A former student of the structural functionalist Talcott Parsons, Garfinkel was led by his study of the deliberations of a trial jury to emphasize, like Schutz, the importance of practical reasoning in everyday situations. He showed that our drive to interpret leads us to impute a meaning even to random processes in the social world, as in one of his later experiments in which an interviewer playing the role of a counselor replied to the clients' questions by a random 'yes' or 'no,' leading them into more and more contorted interpretative efforts. The production of meaning is at the same time the production of social order – Parsons' major concern. Unlike his former teacher, however, Garfinkel insisted that social actors are not simply bearers of their social roles ('cultural dopes'), but active subjects obliged to practice social analysis in order to function in everyday society.

Social phenomenology found a somewhat unexpected ally in a convergent move within analytic philosophy. Having been close to the Vienna Circle, Ludwig Wittgenstein came to abandon the simple conception of a picturing relation between propositions and the world, and was drawn into a more sensitive and holistic analysis of the practicalities of 'language games,' based on implicit rules and embedded in what he enigmatically called 'forms of life.' An important book by the Wittgensteinian philosopher Peter Winch (1958) drew the

consequences for social theory, using Max Weber, as Schutz had done, as one of the foils for his argument. For Winch, knowing a society meant learning the way it is conceptualized by its members. He thus revived the central principle of nineteenth century German historicism, according to which every age must be understood in its own terms. Winch directly identified himself with the German idealist tradition by insisting that social relations are 'like' logical relations between propositions (1958: p. 126) and encouraged an ethnographic fieldwork approach, which often includes language learning in the research process.

Analytic philosophy of language was also drawn on by the anthropologist Clifford Geertz (1973) in his very influential defense of what, following the English philosopher Gilbert Ryle (1949), he called 'thick description' (Luhrmann). For Ryle, a thick description is one, which can, for example, differentiate between the twitch of an eyelid and a wink, with all its complex contextual and cultural connotations. Geertz' classic account of cockfighting in Bali exemplifies this approach, moving from a sensitive description of the practice to a broader consideration of its cultural significance. This relatively informal research strategy, which has affinities with what Glaser and Strauss (1967) called 'grounded theory,' can be contrasted with more formal approaches such as functionalism and structuralism. It is sometimes framed in terms of the use of 'emic' (actors') as opposed to 'etic' (observers') categories, terms based on the linguistic distinction between phonemic and phonetic (Pike, 1966; article on cultural ecology). Geertz himself deprecated this usage, as well as that of 'verstehen': an 'actor-oriented' anthropological interpretation is usually too casually referred to as 'seeing things from the actor's point of view,' too bookishly as 'the *verstehen approach*,' or too technically as 'emic analysis…' (Geertz, 1973: p. 14).

Hermeneutic theory itself took a new turn with the 'philosophical hermeneutics' of Hans-Georg Gadamer (1900–2002), whose *Truth and Method*, published in 1960 and translated into English in 1975, insists, in opposition to historicist hermeneutics, on the practical dimension of interpretation, conceived in Heidegger's sense of an 'encounter' between the 'horizon' of the interpreter and that of the text itself. Gadamer's philosophical hermeneutics is thus set in opposition to the methodological emphasis of traditional hermeneutic theories and their concern with the accuracy of interpretation. Gadamer's aim is to describe the underlying process, an existential encounter between two perspectives or horizons of expectation, which makes interpretation possible in the first place. Understanding is not just a matter of immersing oneself imaginatively in the world of the historical actor or text, but a more reflective and practical process that operates with an awareness of the temporal and conceptual distance between text and interpreter and of the ways in which the text has been and continues to be reinterpreted and to exercise an influence over us. This effective history (*Wirkungsgeschichte*), which traditional historicist hermeneutics tends to see as an obstacle, is for Gadamer an essential element that links us to the text. Our prejudgments or prejudices are what make understanding possible.

Although Gadamer has often stressed that his philosophical hermeneutics, with its origin in Heidegger's hermeneutic ontology, is distinct from hermeneutics as a technique of interpretation, his approach clearly poses a challenge to more traditional conceptions of hermeneutics. These differences are brought out in particular in Gadamer's exchanges in the 1960s with Emilio Betti. The alternative conception of the human sciences or *Geisteswissenschaften* put forward in Gadamer's work also made it central to Jürgen Habermas' reformulation of the 'logic of the social sciences' (Habermas, 1988[1967]). Habermas welcomed Gadamer's critique of hermeneutic objectivism, which he saw as the equivalent of positivism in the philosophy of the natural sciences, and also his stress on the totalizing character of understanding. For Habermas, however, Gadamer's stress on the fundamental nature of language, as expressed in his claim that 'Being that can be understood is language,' amounted to a form of linguistic idealism. Together with Gadamer's stress on the importance of tradition and his rehabilitation of the category of prejudice, this suggested an ultimately conservative approach, which was unable to deal with the 'systematic distortion' of communicative processes by relations of power and domination. Habermas and Gadamer debated these issues in the late 1960s and early 1970s (Bleicher, 1980); more recent theorists have tended to stress the compatibility of hermeneutics and critical theory in a conception of critical hermeneutics (Thompson, 1981; Giddens, 1984b; Outhwaite, 1987; Kögler, 1996a, 1996b). Gadamer also engaged briefly with the French deconstructionist philosopher Jacques Derrida, whose conception of interpretation was more skeptical (Silverman and Ihde, 1985; Michelfelder and Palmer, 1989; Bernstein, 2008). Another major French philosopher, Paul Ricoeur (1913–2005), engaged more directly with hermeneutic issues in the social sciences, in his book on Freud (Ricoeur, 1970[1965]), in which he formulated his concept of a 'hermeneutics of suspicion,' in his response to the Habermas–Gadamer debate, and in a large number of other essays (Ricoeur, 1981a,b, 1991, 2010). He addressed in particular the question of analyzing action in textual terms (Ricoeur, 1973).

Habermas' version of 'critical theory' can be seen, along with critical realism and Anthony Giddens' structuration theory, as one of the three particularly influential attempts in the final third of the twentieth century to reconcile, as Max Weber had done at the beginning of the century, the rival claims of explanation and understanding in the social sciences. Habermas, along with his close collaborator Karl-Otto Apel, argued for a complementarity between an empirical–analytic approach oriented to the explanation, prediction, and control of objectified processes and a hermeneutic approach concerned with the extension of understanding; these combined in an emancipatory model of critical social science, instantiated by psychoanalysis and the Marxist critique of ideology, which aims at the removal of causal blocks on understanding. Much of this remains in Habermas's more recent theories of reconstructive science and communicative action.

Critical theory tends toward a dualism of natural and social science, but in an increasingly muted form. The assumption that opposition to positivism also entailed dualism or anti-naturalism was, however, also put in question in the late 1960s and early 1970s by the realist metatheory of science developed by Rom Harré and Roy Bhaskar (1975). Both Harré and Bhaskar, like Habermas, were substantially motivated by the desire to undermine positivistic theories and approaches in the

social sciences. Harré and Secord (1972) developed a philosophy for social psychology based on the work of the later Wittgenstein and the analytic philosophy of language practiced at Oxford by John L. Austin (1911–60). Ordinary language, they argued, is better suited to the description of the mental processes of social actors than apparently more scientific artificial terminology, and they drew attention to models of research practice of this kind in the work of Goffman, Garfinkel, and others.

Harré and Bhaskar were in any case interested in giving a more adequate account of science as a whole, in a world composed of relatively enduring structures and mechanisms. Some of these could be isolated in scientific experimentation, given the contingent existence of *homo sapiens* and *homo scientificus*. An important aspect of the realist program developed by Harré, Bhaskar, and others was a conception of explanation as involving not an essentially semantic reduction of causal statements to general laws but a reference to the causal *powers* of entities, structures, and mechanisms. Causal tendencies might or might not be outweighed by countervailing tendencies, and two causal tendencies may neutralize one another, as do the centrifugal force of the Earth's rotation and its gravitational attraction, with the convenient consequence that human beings and other animals are safely anchored to the Earth's surface.

This and other features of realism meant that the whole issue of naturalism could be rethought. Human beings could be seen as having causal powers and liabilities, just like other entities; it no longer mattered so much that their relations rarely sustained any universal generalizations of an interesting kind, but only sets of tendencies regular enough to be worth exploring. The fact that many of the entities accorded causal force in social scientific explanations were necessarily unobservable was not, as it was for empiricism, a problem of principle. And the understanding of meaning could, as Bhaskar (1979) put it, be seen as in some ways equivalent to measurement in the natural sciences. Finally, it seemed natural to include among the causes of human action the agents' reasons for acting – reasons that must be understood as far as possible.

The realist critique of traditional epistemology found an echo in social theory, notably in the work of Anthony Giddens, who had become similarly impatient with the residues of positivist social science as well as the more radical contentions of social constructionism. Giddens' conception of the 'duality of structure' was designed to replace the traditional dichotomies between theories of social structure and social change, the micro–macro divide, and between interpretive and more structural approaches. Approaches like these, which aim to mediate between pure hermeneutics and more naturalistic conceptions of social science, coexist, in the early twenty-first century, with more explicitly hermeneutic or phenomenological conceptions (see, e.g., Oevermann et al., 1987; Soeffner, 1989; Müller-Doohm and Jung, 1993; Luhrmann). Hermeneutics in a broader sense continues to exist as a major research tradition in the humanities, including the history of political thought (Richter, 2010) and intermediate fields such as 'critical discourse analysis' (Wodak and Meyer, 2009), as well as a minority one in the social sciences (Shapiro and Sica, 1984). More importantly perhaps, social scientists who would not sign up to an explicitly hermeneutic program have at least accepted the importance of hermeneutic issues.

See also: Ethnomethodology, General; Interactionism, Symbolic; Phenomenology in Sociology; Phenomenology: Philosophical Aspects; Schütz, Alfred (1899–1959); Semiotics; Weber, Max (1864–1920); Wittgenstein, Ludwig (1889–1951).

Bibliography

Apel, K.-O., Bubner, R., Gadamer, H.-G., Giegel, H.J., Habermas, J., 1971. Hermeneutik und Ideologiekritik. Suhrkamp, Frankfurt.
Berger, P.L., Luckmann, T., 1967. The Social Construction of Reality. Allen Lane, London.
Bernstein, R.J., 1983. Beyond Objectivism and Relativism. Science, Hermeneutics, and Praxis. Blackwell, Oxford.
Bernstein, R.J., 2008. The conversation that never happened (Gadamer/Derrida). Review of Metaphysics 61 (3), 577–603.
Bhaskar, R., 1975. A Realist Theory of Science. Leeds Books, Leeds, UK.
Bhaskar, R., 1979. The Possibility of Naturalism. Harvester, Brighton.
Bleicher, J., 1980. Contemporary Hermeneutics. Routledge & Kegan Paul, London.
Chladenius, J.M., 1742. Einleitung zur richtigen Auslegung vernünftiger Reden und Schiften. Mit einer Einleitung von Lutz Geldsetzer. F Lanckisch, Leipzig (1969. Stern-Verlag Janssen, Düsseldorf).
Dallmayr, F.R., McCarthy, T.A., 1977. Understanding and Social Inquiry. University of Notre Dame Press, Notre Dame, IN.
Dilthey, W., 1964[1924]. Die Geistige Welt: Einleitung in die Philosophie des Lebens. Teubner, Stuttgart (1989. Introduction to the Human Sciences. Princeton University Press, Princeton).
Dostal, R.J. (Ed.), 2002. The Cambridge Companion to Gadamer. Cambridge University Press, Cambridge.
Droysen, J.G., 1974[1868]. Grundriss der Historik: Vorlesungen über Enzyklopädie und Methodologie der Geschichte. Oldenbourg, Munich.
Gadamer, H.-G., 1960. Wahrheit und Methode. Mohr, Tübingen (1975. Truth and Method. Sheed and Ward, New York).
Garfinkel, H., 1967. Studies in Ethnomethodology. Prentice-Hall, Englewood Cliffs, NJ.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Giddens, A., 1976. New Rules of Sociological Method. Hutchinson, London.
Giddens, A., 1984a. The Constitution of Society. Polity, Cambridge, UK.
Giddens, A., 1984b. Hermeneutics and social theory. In: Shapiro and Sica. Reprinted in Giddens, Profiles and Critiques in Social Theory. University of California Press, Berkeley and Los Angeles, 1987, pp. 215–238.
Glaser, B.G., Strauss, A.L., 1967. The Discovery of Grounded Theory. Aldine, Chicago.
Goffman, E., 1959. The Presentation of Self in Everyday Life. Doubleday, Garden City, NY.
Habermas, J., 1978[1968]. Knowledge and Human Interests. Heinemann, London.
Habermas, J., 1987[1981]. The Theory of Communicative Action, vol. 1. Polity, Cambridge, UK, 1984, vol. 2: 1987.
Habermas, J., 1988[1967]. On the Logic of the Social Sciences. Polity, Cambridge, UK.
Harré, R., Secord, P., 1972. The Explanation of Social Behaviour. Blackwell, Oxford, UK.
Helle, Horst J., 2005. Symbolic Interaction and Verstehen. Peter Lang, Frankfurt am Main, New York.
Kögler, H.-H., 1996a. The Power of Dialog. Critical Hermeneutics after Gadamer and Foucault. MIT Press, Cambridge, MA and London.
Kögler, H.-H., 1996b. The self-empowered subject: Habermas, Foucault and hermeneutic reflexivity. Philosophy and Social Criticism 22 (4), 13–44.
Luhrmann, T.M., 2001. Thick description: methodology. In: International Encyclopedia of the Social and Behavioral Sciences. Elsevier, New York, NY.
Michelfelder, D.P., Palmer, R., 1989. Dialogue and Deconstruction: The Gadamer-Derrida Encounter. SUNY Press, Albany.
Müller-Doohm, S., Jung, T. (Eds.), 1993. Wirklichkeit im Deutungsprozeß. Suhrkamp, Frankfurt.
Mueller-Vollmer, K., 1986. The Hermeneutics Reader. Blackwell, Oxford, UK.
Neurath, O., 1973. In: Neurath, M., Cohen, R.S. (Eds.), Empiricism and Sociology. Reidel, Dordrecht.

Oevermann, U., Tilman, A., Konau, E., Krambeck, J., 1987. Structures of meaning and objective hermeneutics. In: Meja, V., Misgeld, D., Stehr, N. (Eds.), Modern German Sociology. Columbia University Press, New York, pp. 436–447.
Outhwaite, W., 1987. New Philosophies of Social Science. Macmillan, London.
Pike, K.L., 1966. Etic and emic standpoints for the description of behavior. In: Smith, A.G. (Ed.), Communication and Culture. Holt, Rinehart and Winston, New York.
Richter, M., 1995. The History of Political and Social Concepts. A Critical Introduction. Oxford University Press, New York.
Richter, M., 2010. Towards a Lexicon of European political and legal concepts: a comparison of Begriffsgeschichte and the "Cambridge School". Critical Review of International Social and Political Philosophy 6 (2), 91–120.
Ricoeur, P., 1970[1965]. Freud and Philosophy: An Essay on Interpretation. Yale University Press, New Haven.
Ricoeur, P., 1973. The model of the text: meaningful action considered as a text. New Literary History 5 (No. 1). What is Literature? (Autumn), pp. 91–117.
Ricoeur, P., 1975. The Conflict of Interpretations: Essays in Hermeneutics. Northwestern University Press, Evanston.
Ricoeur, P., 1981a. Hermeneutics and the Human Sciences. Cambridge University Press, Cambridge.
Ricoeur, P., 1981b. Lectures on Ideology and Utopia. Cambridge University Press, New York.
Ricoeur, P., 1991. From Text to Action: Essays in Hermeneutics II. Northwestern University Press, Evanston.
Ricoeur, P., 2010. Écrits et conférences 2: Herméneutique. Seuil, Paris.
Ryle, G., 1949. The Concept of Mind. Hutchinson, London.
Schutz, A., 1932. Der sinnhafte Aufbau der sozialen Welt. Springer, Vienna (1972. The Phenomenology of the Social World. Heinemann, London).
Shapiro, G., Sica, A., 1984. Hermeneutics: Questions and Prospects. University of Massachusetts Press, Amherst, MA.
Silverman, H., Ihde, D. (Eds.), 1985. Hermeneutics and Deconstruction. State University of New York Press, Albany.
Soeffner, H.-G., 1989. Auslegung des Alltags – Der Alltag der Auslegung. Suhrkamp, Frankfurt.
Thompson, J.B., 1981. Critical Hermeneutics. Cambridge University Press, Cambridge, UK.
Weber, M., 1975. Roscher and Knies. Free Press, New York.
Weber, M., 2012. In: Bruun, H.H., Whimster, S. (Eds.), Collected Methodological Writings. Routledge, Milton Park.
Winch, P., 1958. The Idea of a Social Science and its Relation to Philosophy. Routledge, London.
Wodak, R., Meyer, M. (Eds.), 2009. Methods for Critical Discourse Analysis, second ed. Sage, London.

Heterosexism and Homophobia

Jodi O'Brien, Seattle University, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Acceptance of homosexuality is increasing in Western cultures, but homophobic negative, fearful, or hateful attitudes and behavior toward gays, lesbians, and bisexuals are still pervasive in those same cultures. Social science studies emphasize that homophobia is deeply rooted in cultural beliefs about gender and supported by heterosexist assumptions that only heterosexual relationships are natural and normal. Homophobia is institutionalized through the actions of law enforcement, the judiciary, politicians, and religious leaders, as well as through public support for continued homophobic discrimination. Challenges to homophobic social practices are rising, but action to insure the rights of all sexual minorities remains essential.

Homophobiais a form of cultural prejudice that includes negative, fearful, or hateful attitudes and behavior toward gays, lesbians, and bisexuals. One reason why lesbians, gay men, and bisexuals are the object of so much cultural fear and loathing is that homosexual sex does not conform to the gendered or reproductive expectations of heterosexuality. The extent to which homophobia is a US institution is indicated in the decisions of law enforcement officials and judges who consider homosexuality sufficient grounds for justifying acts of prejudice, discrimination, and violence. This article will cover theories of homophobia and the normative expectations of heterosexuality, manifestations of homophobia, and a brief history of social movements related to homophobia.

Terminology

Before proceeding, it is important to note that there is debate about the use of the word 'homophobia' and the closely related word 'heterosexism'. Although homophobia is a commonly used term to indicate antilesbian and antigay prejudice, many social scientists (e.g., Adam, 1998; Plummer, 1992; Stein and Plummer, 1994) find it useful to distinguish analytically and conceptually between heterosexism, which they define as the societal prejudice located in the structures, institutions, and material conditions of society, and homophobia, which they define as the manifestation of the individual-level attitudes and behavior reflective of heterosexism. In Popular media, however, 'homophobia' commonly refers to both.

Heterosexism and homophobia are mutually constitutive: individuals feel that their antigay attitudes and behaviors are legitimate to the extent that these prejudices are entrenched in social institutions. Individuals are more likely to act out their prejudices when they consider them to be culturally legitimate and shared by others (Berrill, 1992). Expressions of antigay prejudice reinforce ideologies and practices of intolerance and hostility that suggest that homosexuality is something that should be despised and feared and should remain hidden.

Attitudes and Expressions of Homophobia

One common stereotype is that homosexuality is the same as gender-role nonconformity. It is frequently assumed that lesbians manifest characteristics defined as masculine and gay men exhibit behavior associated with feminine characteristics. For many people, this presumed gender nonconformity is reason enough to be suspicious and intolerant of homosexuals. Gay men are also often assumed to be hypersexual, animalistic, promiscuous, overly visible, and conspiratorial. Lesbians are assumed to be man haters, aggressive, socially nonconforming, and sexually cold (Herek, 1995). Until 1973, the American Psychiatric Association considered homosexuality to be a form of mental illness. Many people continue to hold this belief and also believe homosexuality to be immoral and antisocial.

Expressions of homophobia include verbal assaults and derisive joking, negative stereotypical media representations, discrimination in employment, education, housing, medical research, legal defense, and custody rights, physical attack, and tacit approval of all forms of discrimination, including gay bashing. One of the most widely publicized cases of homophobia-related violence is that of Matthew Shepherd, a college student in Laramie, Wyoming. One evening in 1998 two men offered Shepherd a ride home from a gay bar and then took him to a field where they beat and tortured him then tied him to a fence. He died a few days later. The case raised national awareness of gay-targeted beatings and murders.

Various studies, including surveys of gays and lesbians based on probability samples, indicate that nearly half of all respondents have experienced some form of discrimination in employment or housing based on their sexual orientation. The federal government has collected data on hate crimes since the 1990s. This information shows that, per capita, gays and lesbians report the greatest number of hate crimes and these crimes are more likely to be attacks on them personally, rather than attacks on property which are more common for other groups, such as African-Americans and Jews, who are also frequent targets of hate crimes (Rubenstein, 2003). Reports also show that gay men and transgender women prisoners are at greater risk of sexual assault than other prisoners while incarcerated (Jenness et al., 2011). In recent years, bullying of gender nonconforming students in schools has become a national issue. These students may or may not self-identify as gay, lesbian, or transgender, but their gender nonconformity is seen as deviant and punishable.

Institutional Heterosexism

Pervasive negative stereotypes regarding lesbians and gay men suggest that these attitudes and beliefs are rooted in cultural ideologies and social practices. All of these attitudes and expressions reflect the prevailing cultural belief that sexual desire for the opposite sex is expected and normal; sexual desire for someone of the same sex is abnormal and problematic. The lingering belief that homosexuality is dysfunctional and that it constitutes a state of anguish for the individual can be traced to the medicalization of homosexuality and its treatment as an illness. Many medical practitioners and clinicians continue to offer treatments intended to 'cure' lesbians and gays of their 'affliction'. These practices reflect the cultural belief that individuals will be unhappy and maladjusted if they do not experience sexual desire for persons of the opposite sex. Medical assessments that treat homosexuality as a form of disease reinscribe the notion that sexuality is primarily physiological and that heterosexuality is the healthy manifestation of human physiology.

Most contemporary religions also treat homosexuality as problematic. In some cases a distinction is made between homosexual feelings and homosexual acts (e.g., Catholic Catechism, 1989). The experience of homosexual inclinations may be considered an affliction – a 'cross to bear' – but to act on one's homosexual feelings is considered a sin. Many religious leaders advocate celibacy for those 'afflicted' with homosexual tendencies. Some religious organizations sponsor intensive therapy programs intended to help the individual who is 'suffering' from homosexuality to recover. Similar to medical institutions, religious institutions take heterosexuality for granted as the normal and desired form of sexual organization and treat homosexuality as deviant, undesirable, dysfunctional, and additionally, sinful.

Legal practices also reflect a cultural ideology that supports heterosexuality as normal and denigrates homosexuality as abnormal.

Containing Sexuality

Several theorists have observed that homophobia is an expression of a culture that is confused and inconsistent about sexual behavior. Attitudes and actions reflect both obsession with and repression of sexual desire. One basis of tacit agreement in the culture is that sexuality is something that must be contained. Homosexuality, along with adultery, prostitution, and pederasty, mark the boundaries by existing outside culturally approved forms of sexuality – what poet Rich (1980) refers to as compulsory heterosexuality. Social historians (e.g., Boswell, 1980; D'Emilio and Freedman, 1988; Katz, 1976) note that the emergence of a sexual binary in the nineteenth century posits heterosexuality as natural and socially desirable by contrasting it with homosexuality, which is characterized as unnatural and immoral. This binary represents the general tendency in modern Western thought to conflate morality and nature. According to this cultural equation, sexual desire is a manifestation of a natural drive to reproduce. The moral, disciplined individual contains this desire within the bounds of heterosexual monogamy. These attitudes are reflected in clinical, social, and religious definitions of the homosexual as cursed, perverted, sexually obsessed, and evil and, until recently, mentally and physically sick.

Internalized Homophobia

Though recent trends indicate that acceptance of gays is increasing for those in the urban middle class (Barrett and Pollack, 2005), homophobia and heterosexism are so culturally entrenched that many lesbians and gay men continue to live in a state of fear, and this is particularly true for minorities and those living in more culturally conservative regions. This entrenchment of heterosexist expectations and the corresponding active prejudice against homosexuality also induces feelings of 'internalized homophobia'. Internalized homophobia is a form of self-disgust and hatred experienced by many lesbians and gays. The phrase 'internalized homophobia' suggests that individuals who are trying to come to terms with their homosexual desires absorb or take in attitudes of cultural prejudice and use these attitudes as a basis of self-judgment. When the prevalent cultural story line is one of intolerance and debasement, then individual lesbians and gays are likely to have to deal with this in formulating their own sense of themselves.

In a culture that presumes heterosexuality, gays and lesbians must decide continually whether to make evident their deviant sexuality. 'Coming out' is an ongoing process. This process involves wrestling with and reconciling family, peer, and religious expectations with personal feelings and experiences. Many lesbians and gay men fear negative consequences for revealing their sexuality, and experience rejection or isolation by their families, friends, and religious and working communities when they reveal their homosexuality. Fear of rejection and a lack of positive role models can lead to confusion and self-hatred. Gay identified youth are one of the highest risk groups for suicide (Russell and Joyner, 2001). Recent studies compiled by the US Department of Health and Human Services and the University of California, Los Angeles (UCLA)-based Williams Institute indicate that lesbian and gay youth are up to five times more likely to attempt suicide than their heterosexual peers. More promising are the religious and community organizations that endeavor to understand and support gay and lesbian youth. These groups have been successful in reconciling family and social conflict and in integrating gay, lesbian, and transgender identified youth into their schools and communities.

Explanations for Homophobia/Heterosexism

Adam (1998) delineates three theoretical approaches to understanding homophobia. Each approach is distinguished by a conceptual term denoting a specific theoretical orientation to antihomosexual prejudice: Homophobia, Heteronormativity, and Heterosexism. Adam links these concepts to specific disciplines: psychology, cultural studies (literary), and sociology, respectively. The distinguishing aspect of each approach is the analytical unit: personality, discourse, and social structure. Explanations can be grouped accordingly.

Within the field of psychology, Herek (1995) defines the functional explanations that people hold and express because they derive psychological benefit from them. Herek proposes

three principle functions underlying attitudes about homosexuality: experiential, social identity, and ego defense. A person's attitude may be a consequence of personal experiences (positive or negative); a prejudiced attitude may be intended as an expression of identification with a particular social group; or it may be a defensive response to personal psychological conflicts (such as gender expectations). He emphasizes that it is important to recognize the nexus between these functional attitudes and cultural heterosexism. The source and expression of these attitudes directly reflects a person's perceptions of cultural expectations.

The development of literary cultural studies in the 1990s has generated several interrelated theories that come together under the term, 'queer theory'. Queer theorists (e.g., Butler, 1999; Sedgewick, 1990; Seidman, 1997) emphasize discursive practices that promote certain lines of thought and action, and discourage or punish others. Queer theorists stress that heteronormativity is one of the most prevalent discourses in contemporary Western culture, and focus on the manifestations of heterosexuality in art, media, education, legal discourse, religion, medicine, and so forth. According to these theorists, the dominant cultural discourse is saturated with images and ideals that promote heterosexuality and condemn homosexuality (or anything that does not fit the expected monogamous, heterosexual, family idea). Queer theorists have focused particularly on the experience of the 'closet' – physical and discursive practices intended to keep homosexuality hidden and contained in separate spheres.

Adam (1998) subsumes the psychological and cultural studies approaches within the sociological studies. His approach emphasizes social structure, particularly the ways in which psychological attitudes and the values revealed in cultural discourses are manifest and maintained through institutional structures and practices. Change, he suggests, will be a result not only of education against prejudiced attitudes and engaging in transgressive acts intended as critiques of taken-for-granted cultural practices, but will depend on the extent to which this education and alternative expression permeates and unsettles existing institutional arrangements (such as the family, law, and religion).

Gender Transgression

Understanding homophobia and heterosexism is rooted in the strong link between gender and sexuality in culture. Gender is the basic criterion by which we select sexual partners. The violation of expected gender behavior elicits taunting, prejudice, and physical and social abuse. Frequent targets of this prejudice include the mannish lesbian and the effeminate gay male. Both evoke stereotypes that represent violations of gender expectations and hint at sexual transgressions. Transgender persons are frequent targets of violence (*see* Transgender, Transexual, and Gender Variant Individuals). It is difficult to know if attackers of transgender persons are provoked by what they assume to be a sexual aberration or a gender violation. Given the tight connection between gender and sexuality, it is likely that any deviation is perceived as a threat to heterosexuality.

In the 1990s Lamba Legal Defense documented several cases in the US of murdered transgender persons in which the murderers were acquitted by the courts. The rationale offered in each case was that to arouse sexual desire in another under 'false gender pretense' was to invite violent 'self-defense' when the pretense was revealed (Singer and Deschamps, 1994). This rationale can be interpreted as a manifestation of homophobia; in effect the message is that individuals who do not conform to the expected norms of gender and sexuality do not deserve social justice.

Institutional Homophobia and Permissible Prejudice

The extent to which homophobia is institutionalized is indicated in the decisions of law enforcement officials, judges, politicians, and religious leaders who consider homosexuality sufficient grounds for justifying acts of prejudice, discrimination, and violence. This implicit approval constitutes a form of permissible prejudice. Persons presume their discriminatory attitudes are legitimate and permissible when these prejudices are supported by people in positions of authority.

The most prevalent indication of institutional homophobia is the widespread lack of legal protection from antigay/lesbian discrimination. Lesbians and gays who experience discrimination in employment, housing, education, and immigration often find it difficult to obtain justice. In some US states it is still legal to ban lesbians and gays from various public institutions. In 1993 a legal statute was implemented that prohibited harassment of closeted gay men and lesbians in the military, but also prohibited them from service if they were openly gay. This "don't ask don't tell" policy was overturned in 2011 (at a time when the US military was desperate to increase troops), but openly gay military personnel continue to report high rates of discrimination.

Another indication of institutional homophobia is public resistance to attempts by lesbians and gay men to reduce discrimination and to seek justice. In the 1990s US citizens in several states voted against ballot initiatives that would ensure equal legal protection for lesbians and gays in employment, housing, other public accommodations, or relationships. Opponents of nondiscrimination initiatives insisted that lesbians and gays were seeking 'special rights and privileges' The logic underlying this argument was that the US constitution already guarantees every individual equal protection under the law. This is true in theory, but advocates of the equality initiatives point out that in practice this protection is often not extended to lesbians and gays. In fact, many social officials, including employers, housing agencies, and health care providers actively refuse to grant lesbians and gays equal rights and openly cite deviant sexual orientation as the basis for this discrimination. US courts have routinely upheld these discriminatory actions.

Not only does the law fail to protect the rights of lesbians and gays, but it has been enforced differentially as a means of punishing lesbians and gays. One example of differential enforcement is the enactment of sodomy laws. Until recently, oral or anal sex was still prohibited in 14 states. Technically, these 'sodomy' laws refer to the behavior and not to the gender of the participants; anyone, straight or gay, can legally be punished for this behavior. However, the only recorded arrests in the past several decades based on these archaic laws have

been gays and lesbians. When challenged, several states (Kentucky, Louisiana, Michigan, and Texas) have ruled sodomy laws as unconstitutional. However, in 1986, in the case of Bowers v. Hardwick, the US Supreme Court upheld a Georgia sodomy law in a narrow vote of 5 to 4. The majority opinion, written by Chief Justice Warren Burger expresses the general legal failure to separate the expression of private individual sexual behaviors from perceived cultural expectations and focuses specifically on the gendered aspect of the behavior. In his brief he writes: "To uphold that *homosexual* sodomy is somehow protected as a fundamental right would be to cast aside millennia of moral teaching" (Bowers v. Hardwick, 1986, emphasis mine).

Court decisions such as Bowers v. Hardwick galvanized many US citizens who were concerned with the implications of court-mandated moral behavior and invasion of privacy. As a consequence, in the 1990s lesbians and gays fought actively against discrimination. Several gains were made, especially in areas such as housing and employment discrimination, but these efforts were matched in intensity by the rise antigay and antilesbian initiatives and legislation. The intent of initiatives such as Colorado's Amendment 2 (enacted in 1992 and overturned in 1996 by the Supreme Court in Romer v. Evans) was to officially prohibit lesbians and gays from seeking legal protection for discrimination based on their status as homosexuals. This preemptive legislation actively denied the pursuit of civil rights based on a single specific category – homosexuality. In an address to Congress regarding legislation intended to protect persons from discriminatory violence (hate crimes), Senator Dannemeyer makes clear this categorical prejudice. He notes: "(I)t is a Federal responsibility to ensure the equal protection of all citizens regardless of their race, religion, or ethnic origin. It is not a Federal obligation to protect citizens in their sexual orientation" (US Congress, 1988: p. 12).

Comments made by those with institutional authority such as Chief Justice Burger and Senator Dannemeyer buttress individual feelings of homophobia and send the message that discrimination and violence will be tolerated.

Social Movements

The social protests and political activism associated with lesbian and gay history can be interpreted as a response to widespread, organized manifestations of homophobia. Several observers have suggested that the struggle for gay and lesbian rights was *the* civil rights issue of the 1990s. This history represents a fight for rights that are constitutionally guaranteed to all US citizens but frequently withheld from gays, lesbians, bisexuals, and transgender people. Many local and national lesbian, gay, bisexual, and transgender movements have been organized as a direct response to political initiatives designed to legalize discrimination against homosexuals.

In the early 1990s various forms of antigay/lesbian legislation appeared on the ballots of 41 states and the District of Columbia. This sort of legislation is indicative of the extent to which homophobia is culturally legitimated as an institutionalized form of prejudice. Critics who suggest that antidiscrimination laws are 'special rights' fail to comprehend the extensive reach of homophobic actions into the everyday lives of lesbians, gays, bisexuals, and transgender persons. The Defense of Marriage Act introduced into US federal legislation in 1996 was one such case. The urgency with which several states have pursued legislation to restrict marriage to a man and a woman and to prohibit same-sex marriages is a display of mass homophobia.

Recent federal level acts as well as various social movements give the impression that institutional level homophobia is on the decline: In 2003 the Supreme Court overturned the state level sodomy laws in the decision of Lawrence v. Texas; In 2009 President Obama signed into law the Matthew Shepherd and James Byrd, Jr. Hate Crimes Act (Byrd was a young black man who was offered a ride and then beaten and dragged three miles to his death chained to the truck by three white men in Texas), the bill had lingered on the desk of the Bush administration for several years prior (Bush stated publicly that if passed in the House and Senate, he would veto the bill when it reached his office). Recently (2013) the US Supreme Court overturned Defense of Marriage Act (DOMA). This widely celebrated action provides federal recognition of the marriages of same gender partners who are legally married in states that allow same-sex marriage. The marriage equality movement and numerous antibullying campaigns (e.g., 'It Gets Better') have been central issues over the past decade. Many celebrities, religious leaders, and politicians courting votes have jumped on the bandwagon to support these particular causes.

Problematizing Homophobia

Is the tide turning on culturally sanctioned homophobia? A 2013 Pew Research study found that one in seven respondents had changed their minds from being opposed to same-sex marriage to support for it. Most people cited getting to know someone who was gay or lesbian and the reason for changing their minds. In the past decade, the presence of gay, lesbian, and transgender icons has risen in the media and, at least in some states, the public sentiment appears to be turning from extreme prejudice and discrimination to support for legal recognition and protection. To many observers, homophobia-based exclusion seems to be giving way to a new era of inclusion. How accurate is this perception and what does it imply for the future?

Many critics have challenged the progressive potential of both the marriage equality movement and the currently prolific representations of gays and lesbians in the media. They note that both the emphasis on marriage and the inclusion of gays and lesbians as ordinary, everyday people eclipses the full range of queer experiences. Feminist legal scholars (e.g., Polikoff, 2009) note that marriage has always been an exclusionary institution and predict that same-sex marriage will only serve reinforce traditional ideas of the 'good citizen' by punishing those who cannot or would not conform to traditional nuclear family ideals.

Homonormativity

In their 1997 book, *Homoeconomics*, Amy Gluckman and Betsy Reed describe the 'gay media marketing moment' as a period in the early 1990s in the US during which advertisers recognized that a certain class of people, namely urban gay white men,

were likely to have more discretionary income than the average married man. This recognition ushered in an era of direct marketing to gay men and lesbians and included imagery intended to reflect the interests and lifestyles of these potential consumers. This marketing focus probably did more for securing cultural citizenship for open and self-identified lesbians and gay men than any other form of political movement. In consumer-based capitalism, purchasing power is one of the hallmarks of citizenship. Representation in marketing campaigns is a manifestation of this 'cultural belonging'. One obvious implication of this form of cultural belonging is that it is limited to those who have the purchasing power.

Contemporary manifestations of 'gay chic' and related cultural expressions of acceptability regarding homosexuality (included in phrases such as 'queer in the streets, straight in the sheets', and 'metrosexual', etc.) reflect an increasing tolerance for homosexuality in certain regions of the world. Nonetheless, and again, ironically, although overt homophobia may be less allowable in some mainstream cultural venues, this so-called tolerance may come at a price. The price may be the increased exclusion and oppression of those who cannot or choose not to participate in these gay mainstreaming expressions. Furthermore, the ways in which this exclusion operates may be even more intractable and insidious than the forms of homophobia with which we are generally familiar (e.g., Bryant, 2008).

Historian Lisa Duggan refers to this trend as 'homonormativity' which is the proliferation of a culturally specific way of being gay that is enough in sync with existing gender, class, racial, and cultural norms as to be considered 'acceptable'. Expressions of homonormativity are the flip side of uncritically deployed homophobia. Together, both function to reinscribe culturally acceptable forms of gayness that render its associations with whiteness, gender appropriate behavior and (US influenced) middle-class status invisible and unproblematic.

Just as homophobia is a culturally learned prejudice, people learn to recognize the 'acceptable' homosexual based on cultural information – specifically portrayals in popular culture. What many people in Western countries, especially the US, think of when they consider homosexuality is the way in which it is portrayed in popular television shows and movies. The current era might be referred to as the post 'Will and Grace' or post 'L Word' or post 'Queer as Folk' era. In-depth analyses of these programs reveal an underlying set of rules about who and what is acceptable and these (homonormative) rules look very much like the rules for heteronormativity. This trend is reflected sloganistically in the shift away from the historically provocative call to action, "we're here, we're queer, get used to it," toward the more banal and assimilationist, "we're here, we're queer, let's go to Ikea."

Throughout the 1990s institutional homophobia may have served as a rallying cry in response to entrenched social injustice, and, as a result of political action, culturally sanctioned homophobia is on the decline. But paradoxically, this has resulted in a form of homonormativity that will likely underscore and increase prejudice toward those whose lives do not reflect Western, middle-class, mainstream, nuclear family values (whether in same or differently gendered marriages). This potentially excluded group includes not only nonnormative queers, but single- and extended-family households, as well as the alternative communal arrangements for support that many people cultivate for themselves and their chosen kin.

Homophobia is a culturally learned prejudice. It stands to reason, therefore, that the manifestations and experiences of homophobia will differ across cultural contexts and situational circumstances. Future activism and scholarship should focus on the multiple factors associated with varying forms of homophobia and also demonstrate the complex ways in which these factors intersect with other forms of difference, inequality, and cultural status. Recent scholars have begun to highlight the complexities of homophobia by intersecting it with informed, critical analyses of institutions such as religion, family, and migration status. These studies reveal a complicated terrain marked by tensions and contradictions, none of which can be reduced to a monolithic concept of homophobia operating outside the parameters of cultural and historical context. In short, we must complicate the idea of 'gayness' masquerading as a master status that presumably transcends other forms of difference; as a kind of 'one size fits all' form of discrimination that is too often applied universally and uncritically.

See also: Feminist Theory: Radical Lesbian; Foucault, Michel (1926–84); Gay and Lesbian Movements; Gay, Lesbian, and Bisexual Youth; Identity Movements; Lesbians: Social and Economic Situation; Queer Theory and Intersectionality; Sexual Harassment: United States and Beyond; Sexual Orientation and the Law; Sexual Orientation: Biological Influences; Sexual Orientation: Historical and Social Construction; Sexuality and Social Work; Sexuality, Theories of; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Adam, B., 1998. Theorizing homophobia. Sexualities 1, 387–404.

Barrett, Donald C., Pollack, Lance M., 2005. Whose gay community? Social class, sexual identity, and gay community involvement. The Sociological Quarterly 46 (3), 437–456.

Berrill, K.T., 1992. Anti-gay violence and victimization in the United States: an overview. In: Herek, G., Berrill, K. (Eds.), Hate Crimes: Confronting Violence against Lesbian and Gay Men. Sage, Newbury Park, CA, pp. 19–45.

Boswell, J., 1980. Christianity, Social Tolerance, and Homosexuality. University of Chicago Press, Chicago.

Bowers v. Hardwick, 1986. Burger, C.J. Concurring opinion. 478 U.S. 186, 106 S. Ct. 2841, 92 L. Ed. 2d 140.

Bryant, Karl, 2008. In defense of gay children? 'Progay' homophobia and the production of homonormativity. Sexualities 11, 445–455.

Butler, J., 1999. Gender Trouble. Routledge, New York.

Catechism of the Catholic Church, 1989. In: Article Six: Offenses against Chastity. Paulist Press Mahwah, NJ, pp. 566–567.

D'Emilio, J., Freedman, E., 1988. Intimate Matters: A History of Sexuality in the United States. Harper and Row, New York.

Herek, G., 1995. Psychological heterosexism in the United States. In: D'Augelli, A., Patterson, C. (Eds.), Lesbian, Gay and Bisexual Identities over the Lifespan. Oxford University Press, Oxford, UK, pp. 321–345.

Jenness, V., Sexton, L., Sumner, J., 2011. Transgender Inmates in California's Prisons: An Empirical Study of a Vulnerable Population. Report submitted to the California Department of Corrections and Rehabilitation, State of California. Sacramento, California.

Katz, J., 1976. Gay American History. Crowell, New York.

National Institute of Justice, US Department of Justice, 1987. The Response of the Criminal Justice System to Bias Crime: An Exploratory Review. Abt Associates, Cambridge, MA.

Plummer, K., 1992. Speaking its name: inventing a lesbian and gay studies. In: Plummer, K. (Ed.), Modern Homosexualities. Routledge, London, pp. 3–28.
Polikoff, N., 2009. Beyond (Gay and Straight) Marriage: Valuing All Families under the Law. Beacon, Boston, MA.
Rich, A., 1980. Compulsory heterosexuality and lesbian existence. Signs: Journal of Women in Culture and Society 5, 631–660.
Rubenstein, W., 2003. The Real Story of U.S. Hate Crime Statistics: An Empirical Analysis. The Williams Institute.
Russell, S., Joyner, K., 2001. Adolescent sexual orientation and suicide risk: evidence from a national study. American Journal of Public Health 91 (8), 1276–1281.
Sedgewick, E.K., 1990. Epistemology of the Closet. University of California Press, Berkeley, CA.
Seidman, S., 1997. Difference Troubles: Queering Social Theory and Sexual Politics. Cambridge University Press, Cambridge, UK.
Seidman, S., Meeks, C., Traschen, F., 1999. Beyond the closet? the changing social meaning of homosexuality in the United States. Sexualities 2, 9–34.
Singer, B.L., Deschamps, D., 1994. Gay and Lesbian Stats. New Press, New York.
Stein, A., Plummer, K., 1994. I can't even think straight: queer theory and the missing sexual revolution in sociology. Sociological Theory 12, 178–187.
US Congress, 1988. Hate Crimes Statistics Act. Report from the Committee on the Judiciary, 100th Congress, 2nd Session. Government Printing Office, Washington, DC.
US Department of Health and Human Services, 1989. Report of the Secretary's Task Force on Youth Suicide. Department of Health & Human Services, Rockville, MD.

Heterosexual Expression at Older Ages

Amelia Karraker, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews recent research on the roles of marital status and dyadic characteristics, physical and mental health, attitudes and beliefs, and cohort effects in sexual expression. It also focuses on gender differences in the roles of these factors in sexual expression and suggests directions for future research.

Introduction

Sexual expression is an integral component of human experience throughout life, including older ages. Although scholarship on sexual expression has focused primarily on populations of reproductive age or younger, interest in sexual expression over the life course has increased in recent years, in part because of the aging of the large cohort of Baby Boomers into older age (*see* Sexuality over the Life Course, this volume). The potential consequences of sexual expression are numerous. Among other potential benefits, sexual activity at older ages is associated positively with physical health (Lindau et al., 2007), well-being (Ganong and Larson, 2011), and relationship quality (Karraker and DeLamater, 2013), and surveys of attitudes suggest that sexual expression remains important to many older adults (Waite et al., 2009). This article begins by defining and describing sexual activity at older ages, defined as approximately age 55 and older. As sexual expression occurs at the intersection of biological, psychological, social, and cultural factors, the article reviews recent research on the roles of marital status and dyadic characteristics, physical and mental health, attitudes and beliefs, and cohort effects in sexual expression. Throughout, it focuses on gender differences in the correlates of sexual expression. The article concludes by outlining avenues and opportunities for future research.

Defining and Describing Sexual Activity at Older Ages

Before examining the factors of sexual expression at older ages, an important first step is to establish what 'sex' is and how its definition may vary by age. Landmark studies of sexuality at various periods of the life course have utilized relatively expansive definitions of sexual activity. For example, the National Social Life, Health, and Aging Project (NSHAP), a nationally representative study of older Americans (ages 57–85 years) which begun in 2005, defines 'sex' or 'sexual activity' as "any mutually voluntary activity with another person that involves sexual contact, whether or not intercourse or orgasm occurs." (Lindau et al., 2007: p. 763) Such sexual activity may include a variety of specific acts, such as hugging, kissing, and other sexual touching (i.e., 'foreplay'); giving and receiving oral sex; and vaginal intercourse, among other behaviors. Consistent with findings of younger people taken from the National Health and Social Life Survey (Laumann et al., 1994), the majority of older sexually active men and women report that vaginal intercourse (74–91%, depending on age and gender) and 'kissing, hugging, or other sexual touching' (nearly 90% or higher for all age-gender groups) 'usually' or 'always' occurs during sex (Waite et al., 2009: p. i59). However, the percentage either giving or receiving any oral sex drops markedly by age; this may reflect cohort effects, which are discussed in a subsequent section. The likelihood of sexual activity and sexual frequency varies greatly by gender and age (Lindau et al., 2007; Karraker et al., 2011). Younger people are more likely to be sexually active than older people, and men are more likely to be active than women at all ages (Lindau et al., 2007). For example, among those age 57–64 years, 84% of men reported being sexually active in the past year, compared with 62% of women; by ages 75–85 years, these percentages drop to 39% of men and 17% of women (Lindau et al., 2007).

Heterosexual expression pertains both to partnered activities, such as vaginal intercourse, as well as masturbation. For both women and men, reports of masturbation within the past year decline with age for both genders, although as for partnered activity, masturbation is more common for men than women at all ages (Waite et al., 2009). Interestingly, rates of masturbation are similar for partnered (52% of men, 25% of women) and unpartnered older adults (55% of men, 23% of women) (Lindau et al., 2007). This finding of little differences in masturbation by partner status is consistent with research on younger populations emphasizing masturbation as either complementary to partnered sexual activity or compensatory to a lack of partnered activity (Das, 2007; Kontula and Haavio-Mannila, 2003).

Sexual problems are quite common at older ages. Using data from NSHAP, Waite and colleagues (2009) found that among men who were sexually active in the past year, the most common sexual problems were erectile difficulties (30–44%, depending on age-group), premature climax (21–29%), and lack of interest in sex (24–29%). Among women sexually active in the past year, the most common sexual problems were lack of interest in sex (38–49%), lubrication problems (36–44%), and inability to climax (33–38%). The authors also found that both men and women experiencing sexual problems reported avoiding sex due to these problems. This finding, combined with the fact that these questions often are asked only of those who are currently or recently sexually active, suggests that these estimates may underestimate sexual problems if they result in cessation of sexual activity.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.35004-8

Qualitative work also adds important nuance to these findings from quantitative studies. In the case of diminished physical function, some elders (both male and female) report expanded definitions of what 'sex' is, including increased emphasis on physical closeness and foreplay rather than on vaginal intercourse or the presence of erections (Vares et al., 2007; Potts et al., 2006). Furthermore, although quantitative studies document declines in the prevalence of a variety of sexual activities (Waite et al., 2009) and their frequency (Karraker et al., 2011), qualitative work has found that for some individuals, sexual satisfaction actually increased with age (Lodge and Umberson, 2012; Vares et al., 2007; Potts et al., 2006), and for others, positive sexual experiences earlier in life shaped continued pleasure in sex (Loe, 2012).

Determinants of Sexual Activity

Marital Status and Dyadic Characteristics

One of the strongest predictors of sexual activity at older ages is partner status (Lindau et al., 2007; Karraker et al., 2011), as availability of a partner is a necessary, although not sufficient, condition for many sexual behaviors. For many, but not all individuals, partner availability coincides with marital status. The distributions of marital status by gender, however, diverge with increases in age, yielding an increased share of older unpartnered women. The largest gender differences are in the proportion of older people widowed and married. Before age 40 years, widowhood is extremely rare for both men and women, but by ages 55–64 years, 2.5% of men are widowed compared with 8.1% of women, and by ages 75 years and older, 21% of men are widowed whereas almost 57% of women are widowed (U.S. Census Bureau, 2010). In contrast, among those ages 55–64 years, about 74% of men are married, compared with about 67% of women, and by age 75 years and older, nearly 70% of men are married, compared with about 32% of women (U.S. Census Bureau, 2010).

Two primary processes are driving this phenomenon of increasing partner scarcity for women with age. First, women live longer than men (Austad, 2006). Women begin to outnumber men around age 40 years, and by age 85 years, there are more than two women for every man (U.S. Census Bureau, 2006). The increasing scarcity of men with age is exacerbated by a second process: gendered marriage markets yielding age hypergamy. Age hypergamy is the tendency of men to marry women younger than themselves. Scholars argue that cultural standards that privilege youthful appearance and reproductive fitness for women, as well as economic position for men, may all be social forces contributing to age hypergamy (England and McClintock, 2009). Women partnered to same-age men already would expect to outlive their husbands on average given greater female longevity, but age hypergamy increases the length of time by which wives are likely to outlive their husbands, as well as to have a spouse in poorer health. Recent work using the NSHAP data found that, on average, men are 3 years older than their female partners (Lindau et al., 2007). Thus, although a 70-year-old man may be partnered with a 67-year-old woman, a 70-year-old woman may be partnered with a 73-year-old man. Although a 3-year age difference may not at first appear large, increasing morbidity may generate large gender asymmetries in partner health as individuals age.

Opportunities for repartnering also differ by gender. Women also are more likely to be widowed as they age (and outlive their husbands), but less likely to remarry (Buckle et al., 1996). Furthermore, there is evidence that the characteristics of older married women and men's spouses diverge with age, especially for remarriages. England and McClintock (2009) found that the older the age of men at marriage (through their 60s), the younger his wife is likely to be. Those older men who are able to attract younger female partners are likely in better physical health (and likely have other characteristics associated with sexual activity) than those who cannot. This 'double standard of aging' (England and McClintock, 2009: p. 797), may partially explain why even within marriage, older married men are more likely to report being sexually active than older married women (Lindau et al., 2007; Karraker and DeLamater, 2013).

For those individuals with partners, the characteristics of the dyadic relationship also are associated with sexual expression (see also Marital Quality and Health, this volume). Partnered sexual expression is an important component of intimate relationships for many couples and thus enhances and strengthens the relationship in many circumstances (Michael, 2004). Not surprisingly, as at younger ages, marital satisfaction remains positively associated with sexual activity at older ages (Karraker and DeLamater, 2013). The absolute differences between sexually active and inactive couples are small, however. This finding is contrasted with work on longtime married couples in middle age (those in their 40s and 50s), which has found that high initial levels of sexual satisfaction increased marital quality and decreased subsequent marital instability (Yeh et al., 2006). Taken together, these findings suggest possible differences in the relationship between marital quality and sexual activity by cohort or age. In addition, the 'aging' of relationships also is associated with sexual activity. A recent study using the NSHAP data found that marital duration was positively (though weakly) associated with sexual inactivity and that this relationship was independent of the age of both spouses and well as their physical and mental health (Karraker and DeLamater, 2013).

Physical Health and Psychological Well-Being

Few factors related to sexual expression have received as much attention as physical health. Indeed, poor physical health, which often accompanies aging, is associated with sexual expression. Physical conditions, such as cardiovascular disease, cancer, diabetes, and arthritis, can compromise sexual expression through a variety of pathways (DeLamater and Sill, 2005). Some physical illnesses can interfere with the mechanics of sexual function (see also Epidemiology of Sexual Dysfunction, this volume). For example, cardiovascular disease and diabetes can both interfere with blood flow, leading to erectile dysfunction among men. Conditions like arthritis and obesity also may create barriers to sexual activity by impairing functional health. Hormonal changes associated with biological aging may have consequences for sexual function. Declines in various hormone levels during

menopause, for example, may result in reduced vaginal lubrication, making vaginal intercourse uncomfortable or painful in the absence of additional lubrication and may result in decreased sex drive (Dennerstein et al., 2003).

As for marital status, the impacts of physical illness on sexual activity differ by gender. Prior research has found that male partner's physical health is a stronger predictor of partnered sexual activity than female partner's (Karraker and DeLamater, 2013; Lindau et al., 2007). This likely reflects gender differences in specific disease burden, as well as differences in the impact of a given condition on sexual expression. For example, although high blood pressure and vascular problems may impede a man's erectile function to the extent that vaginal intercourse is not physically possible, the same condition in a woman may not interfere substantially with sexual function.

Psychological well-being and mental health also are associated with sexual expression. In fact, although less often studied than physical health, mental health may be as strongly or more strongly associated with sexual expression than physical health. Psychological factors are deeply implicated in sexual desire (see also Sexual Motivation, this volume), which can be defined as the 'interest in being sexual' (Kingsberg, 2002: p. 434). A variety of psychological conditions are association with sexual activity, including depression, anxiety, stress, and self-rated mental health (Laumann et al., 2008; Kingsberg, 2002). Furthermore, recent work suggests that mental health conditions are associated more consistently with women's sexual problems than with men's (Laumann et al., 2008).

Medications are tied to sexual function. The medications used to treat several conditions, such as high blood pressure and depression, may create barriers to sexual activity by compromising the mechanics of sex, or the desire to engage in activity, respectively (DeLamater and Sill, 2005). Of course, the most notable medication tied to sexual activity is Viagra. Used to treat erectile dysfunction, the introduction of Viagra to the US market in 1998 represented an important shift in the ways in which Americans thought, talked, and acted about sexuality in later life (Loe, 2004) by recasting a former threat to masculinity, 'impotence,' as erectile dysfunction, a 'normal' part of aging with a biological cause, and pharmacological solution (Marshall, 2012). Using the NSHAP data, researchers estimated that 14% of men ages 57–85 years reported using either medications or supplements to improve sexual function (Lindau et al., 2007). The initial enthusiasm about these medications, however, may have been premature in some instances. One study of men taking Viagra found that 30% discontinued use (i.e., did not refill their prescription within 6 months) even though they reported that the drug was effective at treating erectile dysfunction (Klotz et al., 2005). Common reasons for discontinuation were a lack of opportunity for sex and lack of interest (both for self and partner). These findings suggest that an overemphasis on the biological mechanics of specific sexual activities (such as penetrative intercourse) for one partner in the intimate relationship is ill-advised.

Some scholars have cited the development and marketing of Viagra and other sexual dysfunction drugs reflects the growing medicalization of sexuality at older ages (Loe, 2004; Marshall, 2012). Conrad (1992: p. 211) defined medicalization as 'defining a problem in medical terms, using medical language to describe a problem, adopting a medical framework to understand a problem, or using a medical intervention to "treat" it.' The medicalization of sexuality at older ages may help to liberate sexuality among older people from taboos and stereotypes of asexual older ages or 'dirty old men,' as well as to facilitate treatment for sexual problems like erectile dysfunction, which may be indicative of underlying serious health problems. Some argue, however, that sexuality has become 'overmedicalized,' citing 'excessive medical diagnostic, workups, excessive use of medical and surgical treatments, and excessive claims about the medical causes of sexual problems' (Tiefer, 1996: p. 252). Critics posit that excessive medicalization of sexuality privileges particular forms of sexual expression as more 'valid' (such as penetrative vaginal intercourse) and equates sexual health at older ages with conformity to youthful standards of sexual function (Marshall, 2012). They argue that the study of sexual expression at older ages instead should acknowledge the diverse forms that sexual expression at older ages may take, as well as the complex interplay of biological, psychological, social, and cultural factors that influence both men and women's sexual expression as individuals and in intimate relationships.

Attitudes and Beliefs

Personal attitudes and beliefs both reflect and influence sexual expression (Waite et al., 2009). Not surprisingly, negative beliefs about sex are associated with lower sexual desire, which may result in lower levels and poorer quality of sexual activity (DeLamater and Sill, 2005). As for many behaviors, beliefs about sex vary by gender and age-group among older adults. Women tend to hold more conservative beliefs about sex than men, and older people have more conservative beliefs than younger people, especially in the case of agreeing or strongly agreeing that sexual ability decreases with age (Waite et al., 2009).

Beliefs about masculinity and femininity also may shape sexual expression (*see* Femininities and Scripts of Masculinity, this volume). Prior work has argued that heterosexual expression is another way in which men and women 'do gender' (Connell, 1987; West and Zimmerman, 1987) and recent empirical work suggests that this continues into older ages (Cornwell and Laumann, 2011; Lodge and Umberson, 2012). The ways in which men and women engage in sexual expression may reflect and reinforce ideas of men as strong, virile, and dominant over women, and of women as passive recipients of male sexual desire. In particular, this perspective emphasizes penile performance and vaginal intercourse (Burgess, 2004). Changes in sexual function may threaten these notions, particularly in light of messages that sexual expression in older ages should be resemble that in younger ages (Marshall, 2012). The distressing nature of sexual problems has been highlighted in both quantitative and qualitative studies (Hirayama and Walker, 2011; Lodge and Umberson, 2012). Recent work has found that a partner's sexual problem is associated positively with depression for both men and women (Hirayama and Walker, 2011). Qualitative studies also highlight that sexual problems are

distressing for both men and women in gendered ways. For example, in study of older married couples, men reported embarrassment when faced with erectile dysfunction, whereas their wives reported blaming themselves for not being physically attractive enough for their husbands (Lodge and Umberson, 2012). The authors, however, found that these concerns were confined largely to younger couples in the survey (ages 50–69 years), whereas older couples (ages 70 years and older) emphasized emotional rather than sexual intimacy. Although prior work has not examined specific individual beliefs and affinities to notions of masculinity and femininity, it is likely that these also influence sexual expression across older ages.

Finally, religious beliefs and associated behavior may shape sexual expression. Religious beliefs and affiliations shape the nature and context of 'appropriate' sexual expression. Although little work has explicitly examined religious beliefs at older ages, there is evidence that the relationship between religious beliefs and behaviors vary by partner status. One recent study found that religious attachment is not related strongly to sexual frequency or satisfaction among the married, but rather is an important predictor of these same outcomes for women (but not men) without partners (McFarland et al., 2010). Another recent study examining sexual inactivity found that Catholic identity is positively associated with sexual inactivity for women but not men (Karraker and DeLamater, 2013), which may reflect Church doctrine about the procreative purpose of sexual activity and stronger ties between religious affiliation and behavior for women than men. Indeed, work examining younger samples that religion is more influential for women's sexual behavior than men's (Rostosky et al., 2004).

Cohort Effects

Sexual expression also is shaped by cohort experiences. A cohort perspective stipulates that individuals born in the same time period and experience certain historical, cultural, institutional, and technological events together carry these experiences with them across their lives (Elder and O'Rand, 1994). The aging of Baby Boomers (those born between 1946 and 1964) often is cited as a motivation for the study of sexuality in older age (Lindau et al., 2007; Karraker and DeLamater, 2013), but what does the future hold for this cohort? Compared with older cohorts, this cohort has experienced better medical care throughout the life span, which likely translates into healthier older ages, facilitating continued sexual activity. Many of the Boomers also experienced the important technological innovation of oral contraceptives around the formative ages of adolescence and young adulthood, disconnecting sexual activity from reproduction. Expansions in education likely contributed to more liberal attitudes about sex. These changes may result in higher levels of sexual activity in the postreproductive years compared with older cohorts. Furthermore, the rise of cohabitation as an alternative to marriage (Brown et al., 2006) also suggests changing contexts for sexual activity for more recent birth cohorts. Changing relationship contexts for sexual activity, including the extent to which they are monogamous, in turn may increase risks for sexually transmitted infections at older ages. In addition, if more recent cohorts place greater emphasis on continued sexual activity in relationships, sexual inactivity may be an increasingly important mechanism selecting individuals into and out of relationships. Finally, greater female agency as a result of experiencing the sexual revolution in the critical periods of adolescence and young adulthood (Sassler, 2010), combined with narrowing life expectancies between men and women, may result in narrower gender differentials in at least some aspects of sexual expression for younger cohorts.

This article has focused on a variety of correlates of sexual expression, but what is the relationship between age itself and sexual expression? It has been notoriously difficult to untangle the relationships among age, cohort, and period (such as the introduction of Viagra in 1998) in most studies because few studies contain information on multiple cohorts as they pass through specific historic time periods and ages. Some evidence exists that younger cohorts born in more sexually open times carry sexual repertoires and attitudes into older ages (Kontula and Haavio-Mannila, 2003). Evidence also suggests that the aging process yields more conservative attitudes about such issues as premarital sex within cohorts (Harding and Jencks, 2003), which may yield decreases in sexual expression as cohorts age, but preserve intercohort differences. One avenue for future research is to disentangle the independent forces of age, cohort, and period for sexual expression.

Directions for Future Research

Major advances have been made in understanding sexual expression among older people and its determinants in recent years, but much work remains to be done. In addition to several specific issues already discussed, this article concludes with two suggestions for future research. First, several analytic and methodological issues remain to be resolved. To what extent are associations between such factors as health with sexual expression causal, and in what direction does causation flow? Most of the prior work on the determinants of sexual expression has established associations between factors and sexual expression, but cannot establish causal order. For example, poor health may be an antecedent to sexual inactivity, its consequence, or both (Karraker and DeLamater, 2013). Health problems may create sexual problems that in turn may strain intimate relationships, further compromising health. The collection and analysis of longitudinal data documenting changes in sexual activity and its correlates can help untangle the relative importance of pathways flowing from sexual expression to health and vice versa. In addition, because much (but not all) sexual expression takes place in an intimate relationship, more work is needed on how each individual within the sexual dyad, as well as the dyad itself, influences expression. Prior quantitative and qualitative work has identified gender differences in the determinants of partnered behavior, but usually has not used data reported by each partner. Analysis of both partners characteristics and experiences may provide more direct evidence of a gendered sexual life course. More broadly, the incorporation of life-course theories often employed by scholars of aging would be productive additions to the study of sexuality at older ages.

Second, much work (including this article) has focused on variation by gender in the determinants and experiences of sexual expression at older ages; far less has focused on the role of race and class beyond including controls for such factors as race or ethnicity and education. Race and elements of class shape sexual expression earlier in the life course (Laumann et al., 1994), and it is likely that these forces persist in later life. Race and class also pattern other determinants of sexual activity, such as health (Williams, 1999; Herd et al., 2007). Furthermore, women of color and poor women (who are more likely to be women of color) are less likely to marry than those who are white or more economically advantaged (Goldstein and Kenney, 2001). This suggests that the contexts in which sexual activity occur as well as its determinants may vary jointly by class, gender, and race. One strategy to incorporate these empirical facts is to utilize an intersectional approach, which contends that researchers must account for the complex, multiple, and interdependent social positions individuals inhabit that shape opportunities and their lives (McCall, 2005). Likewise, the study of sexual expression at older ages presents exciting opportunities for social scientists at the crossroads of biological, psychological, social, and cultural forces.

See also: Aging and Health in Old Age; Emotions and Aging; Life Course: Sociological Aspects; Motivation: Life Course and Sociological Perspectives; Sexuality: Cultural Aspects.

Bibliography

Austad, S.N., 2006. Why women live longer than men: sex differences in longevity. Gender Medicine 3, 79–92.

Brown, S.L., Lee, G.R., Bulanda, J.R., 2006. Cohabitation among older adults: a national portrait. Journals of Gerontology: Social Sciences 61, S71–S79.

Buckle, L., Gallup, G.G., Rodd, Z.A., 1996. Marriage as a reproductive contract: patterns of marriage, divorce, and remarriage. Ethology and Sociobiology 17, 363–377.

Burgess, E.O., 2004. Sexuality in midlife and later life couples. In: Harvey, J., Wenzel, A., Sprecher, S. (Eds.), The Handbook of Sexuality in Close Relationships. Erlbaum, Mahwah, NJ, pp. 437–454.

Connell, R.W., 1987. Gender and Power. Stanford University Press, Stanford, CA.

Conrad, P., 1992. Medicalization and social control. Annual Review of Sociology 18, 209–232.

Cornwell, B., Laumann, E.O., 2011. Network position and sexual dysfunction: Implications of partner betweenness for men. American Journal of Sociology 117, 172–208.

Das, A., 2007. Masturbation in the United States. Journal of Sex and Marital Therapy 33, 301–317.

DeLamater, J., Sill, M., 2005. Sexual desire in later life. Journal of Sex Research 42, 138–149.

Dennerstein, L., Alexander, J.L., Kotz, K., 2003. The menopause and sexual functioning: a review of the population-based studies. Annual Review of Sex Research 14, 64–82.

Elder, G.H., O'Rand, A., 1994. Adult lives in a changing society. In: Cook, K.S., Fine, G.A., House, J.S. (Eds.), Sociological Perspectives on Social Psychology. Allyn & Bacon, Needham Heights, MA, pp. 452–475.

England, P., McClintock, E., 2009. The gendered double standard of aging in U.S. marriage markets. Population and Development Review 35, 797–816.

Ganong, K., Larson, E., 2011. Intimacy and belonging: the association between sexual activity and depression among older adults. Society and Mental Health 1, 153–172.

Goldstein, J.R., Kenney, C.T., 2001. Marriage delayed or marriage forgone? New cohort forecasts of first marriage for U.S. women. American Sociological Review 66, 506–519.

Harding, D.J., Jencks, C., 2003. Changing attitudes toward premarital sex: cohort, period, and aging effects. Public Opinion Quarterly 67, 211–226.

Herd, P., Goesling, B., House, J.S., 2007. Socioeconomic position and health: the differential effects of education and income on the onset versus progression of health problems. Journal of Health and Social Behavior 48, 223–238.

Hirayama, R., Walker, A.J., 2011. When a partner has a sexual problem: gendered implications for psychological well-being in later life. Journals of Gerontology: Social Sciences 66, 804–813.

Karraker, A., DeLamater, J., Schwartz, C.R., 2011. Sexual frequency decline from midlife to later life. Journals of Gerontology: Social Sciences 66, 502–512.

Karraker, A., DeLamater, D., 2013. Past-year sexual inactivity among older married persons and their partners. Journal of Marriage and Family 75, 142–163.

Kingsberg, S.A., 2002. The impact of aging on sexual function in women and their partners. Archives of Sexual Behavior 31, 431–437.

Klotz, T., Mathers, M., Klotz, R., Sommer, F., 2005. Why do patients with erectile dysfunction abandon effective treatment with sildenafil (Viagra®)? International Journal of Impotence Research 17, 2–4.

Kontula, O., Haavio-Mannila, E., 2003. Masturbation in a generational perspective. Journal of Psychology and Human Sexuality 14, 49–83.

Laumann, E.O., Gagnon, J.H., Michael, R.T., Michaels, S., 1994. The Social Organization of Sexuality: Sexual Practices in the United States. University of Chicago Press, Chicago.

Laumann, E.O., Das, A., Waite, L.J., 2008. Sexual dysfunction among older adults: prevalence and risk factors from a nationally representative U.S. probability sample of men and women 57–85 years of age. Journal of Sexual Medicine 5, 2300–2311.

Lindau, S.T., Schumm, L.P., Laumann, E.O., Levinson, W., O'Muircheartaigh, C.A., Waite, L.J., 2007. A study of sexuality and health among older adults in the United States. New England Journal of Medicine 357, 762–774.

Lodge, A., Umberson, D., 2012. All shook up: sexuality of mid to later life married couples. Journal of Marriage and Family 74, 428–443.

Loe, M., 2004. Sex and the senior woman: pleasure and danger in the Viagra era. Sexualities 7, 303–326.

Loe, M., 2012. Pleasure in old age. In: Carpenter, L., DeLamater, D. (Eds.), Sex for Life: From Virginity to Viagra, How Sexuality Changes Throughout Our Lives. NYU Press, New York, pp. 278–298.

Marshall, B.L., 2012. Medicalization and the refashioning of age-related limits on sexuality. Journal of Sex Research 49, 337–343.

McCall, L., 2005. The complexity of intersectionality. Signs 30, 1771–1800.

McFarland, M.J., Uecker, J.E., Regnerus, M.D., 2010. The role of religion in shaping sexual frequency and satisfaction: evidence from married and unmarried older adults. Journal of Sex Research 47, 1–12.

Michael, R.T., 2004. Sexual capital: an extension of Grossman's concept of health capital. Journal of Health Economics 23, 643–652.

Potts, A., Grace, V.M., Vares, T., Gavey, N., 2006. 'Sex for life'? Men's counter-stories on 'erectile dysfunction', male sexuality and ageing. Sociology of Health and Illness 28, 306–329.

Rostosky, S.S., Wilcox, B.L., Wright, M.L.C., Randall, B.A., 2004. The impact of religiosity on adolescent sexual behavior: a review of the evidence. Journal of Adolescent Research 19, 677–697.

Sassler, S., 2010. Partnering across the life course: sex, relationships, and mate selection. Journal of Marriage and Family 72, 557–575.

Tiefer, L., 1996. The medicalization of sexuality: conceptual, normative, and professional issues. Annual Review of Sex Research 7, 252.

U.S. Census Bureau, 2006. Women's History Month: March 2006. U.S. Census Press Release.

U.S. Census Bureau. 2010. America's Families and Living Arrangements: 2010, Table 1A. Marital Status of People 15 Years and Over, by Age, Sex, Personal Earnings, Race, and Hispanic Origin: 2010.

Vares, T., Potts, A., Gavey, N., Grace, V.M., 2007. Reconceptualizing cultural narratives of mature women's sexuality in the Viagra era. Journal of Aging Studies 21, 153–164.

Waite, L.J., Laumann, E.O., Das, A., Schumm, L.P., 2009. Sexuality: measures of partnerships, practices, attitudes, and problems in the National Social Life, Health, and Aging Study. Journals of Gerontology: Social Sciences 64, i56–i66.

West, C., Zimmerman, D.H., 1987. Doing gender. Gender and Society 1, 125–151.

Williams, D.R., 1999. Race, socioeconomic status, and health: the added effects of racism and discrimination. Annals of the New York Academy of Sciences 896, 173–188.

Yeh, H.C., Lorenz, F.O., Wickrama, K.A.S., Conger, R.D., Elder, G.H., 2006. Relationships among sexual satisfaction, marital quality, and marital instability at midlife. Journal of Family Psychology 20, 339–343.

Heterosexuality

James Iveniuk, NORC PRC, Chicago, IL, USA
Edward O Laumann, University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, we provide a broad overview of research findings on heterosexuality, focusing on sexual practices, attitudes and sexual dysfunctions. We briefly review the theory of sexual scripts as an introduction to interpreting a wide range of findings from numerous nationally representative surveys and several smaller studies, concentrating our review on American datasets, but providing the reader with cross-cultural findings where possible.

In 1948, entomologist-turned-sexologist Alfred Kinsey published the first half of his two-part study of sexuality, *Sexual Behavior of the Human Male*. The companion volume, *Sexual Behavior of the Human Female*, came out in 1953, and readers today will probably find it difficult to recreate the shock felt by members of the American public at the time. Kinsey's work was not simply an exposé about a deviant subculture, but purportedly included a portrait of average Americans. In light of the variety of sexual expression that Kinsey discovered, he would argue that sexual orientation was a continuous spectrum rather than discrete categories, which came to be known as 'the Kinsey scale' (Kinsey et al., 1948, 1953). While Kinsey would make strong claims about the natural variability of human sexuality, he conceived of the categories that Americans used to sort people into 'heterosexuals' and 'homosexuals' as being more notional than discrete, mutually exclusive categories. Thus his work set generations of sex researchers on the path to studying social processes that define, constrain and enable heterosexual expression. In this article, we summarize the sociologically informed research on heterosexuality that developed after Kinsey, touching upon theory where it is appropriate to interpret these findings.

While Kinsey created a fundamental foundation for method and theory, many problems that beset his research still trouble sexologists today. First, eliciting reliable data that truly measure unobserved sexual practices is challenging for any sex research, in part because people may be reluctant to disclose this information to a stranger (Das and Laumann, 2010). Second, securing funding for sex research remained problematic, however, much researchers were motivated to provide an objective characterization of sexual expression, instead of portraying sexuality as one or another advocacy group would like to speak of it (Laumann and Michael, 2001). Third and finally, in order to make statements as broad and sweeping as Kinsey's, researchers require representative samples from a specified population, e.g., adult males in the United States. This was by far the greatest limitation on Kinsey's findings. While an outside committee of statisticians found his work to be outstanding compared with the scant sex research available at the time, the lack of probability sampling made the results little more than a loose assemblage of case studies (Cochran et al., 1953). Kinsey's data could not be used to make statistical inferences about America as a whole, and so while the scope of his conclusions would set a standard, his methods were lacking in statistical sophistication and rigor.

A study that comprehensively addressed this problem would not appear until the early 1990s. In the intervening period, sex studies tended to be convenience samples of idiosyncratically specified populations, which were unable to test Kinsey's claims about sexual expression in America as a whole (Ericksen and Steffen, 1994). The so-called sexual revolution of the late 1960s also arrived, further transforming America's national discourse on sexuality. However, it was not clear whether this change in discourse created any real change in sexual practices, or whether the American population was characterized by free, open, sexual experimentation, or conventional, less exploratory sexual expression. The National Health and Social Life Survey (NHSLS) was designed to answer these questions, and others, in a nationally representative sample. However, it was nearly not fielded; the American Senate voted 66–34 in September 1991 not to fund the NHSLS in favor of a 'say-no-to-sex' campaign (Laumann and Michael, 2001). The study, unfortunately at a much reduced sample size, was funded by a consortium of seven private foundations, and eventually gathered the data. The first field interviewers began interviews in February of 1992, and the investigator team eventually released the data to the public (Laumann et al., 1994).

Since the publication of the NHSLS's main findings, subsequent datasets have become available which cover countries around the world. The speed with which these datasets were produced means that many of the results can be collated to form a global picture of heterosexual behavior in our current era. Thus, although this article will focus largely on America, it will make comparisons to other countries where possible. This article will also focus on quantitative studies more than ethnographic or interview case studies, in order to provide wide-ranging descriptive information. Where no nationally representative information is available, we cite smaller studies.

Defining 'Heterosexual'

Kinsey theorized sexual orientation as a continuum, and accordingly created a scale that was calculated based on overt

sociosexual experience in adulthood. His calculations produced a score which ranged from '0' to '6' where '6' was exclusively homosexual and '0' was exclusively heterosexual (Kinsey et al., 1948, 1953; Van Wyk and Geist, 1984). Kinsey's main finding was that most people were in the middle of this continuum, meaning that 'heterosexuals' in the unambiguous sense were not the overwhelming majority in America (Kinsey et al., 1948, 1953; McConaghy, 1987; McConaghy et al., 1994). In light of his tabulations, Kinsey declared that if 'heterosexual' meant someone with exclusively heterosexual partners and desires, then nearly half of the American male population was not heterosexual (1948: p. 617).

Some would dispute these findings in subsequent studies, saying that very few Americans fall between the values 2 and 4 on the scale (Van Wyk and Geist, 1984). Others would take issue with the scale being unidimensional (McWhirter et al., 1990). Heterosexual and homosexual desire might be two separate axes, and asexuals could be individuals who were low on both dimensions, just as bisexuals were people who ranked highly on both (Storms, 1980). While these may seem like solely methodological issues, they explicitly address the nature of sexual orientation by trying to set down an acceptable way to measure it.

In the 1994 NHSLS, Laumann and colleagues proposed another way of conceptualizing sexual orientation, to show that the question 'what percent of the American population is heterosexual?' did not have a simple or singular answer. Rather than choosing among self-definition, actual sexual practice, and sexual impulses, the NHSLS team conceived of orientation as a combination of all three, that is: (1) sexual identity, or identifying oneself as 'heterosexual,' 'homosexual' or something else; (2) sexual practice, or actually engaging in genital-stimulating contact with members of the same and/or the other gender; and (3) sexual desire, or having erotic feelings toward and fantasies about members of the same and/or the other gender (Laumann et al., 1994). This is not only a 'multidimensional' definition of heterosexuality, as some have proposed (Storms, 1980), but it also acknowledges overlapping, separate domains of sexuality which intersect to create what we colloquially term 'being heterosexual.' **Tables 1** and **2** show the breakdown of these three domains using data from the NHSLS, augmented by questions from the General Social Survey (GSS) collected that same year. More than 85% of men and women reported exclusively opposite-gender partners in the past 12 months, and if one measures partners since puberty, one finds that over 90% of men and women have only had opposite-gender partners. More than 96% of men and 98% of women identified as heterosexual, and 93.8% of men and 95.6% of women reported only opposite-gender attraction. These numbers show that for the majority of Americans at the time the NHSLS was fielded, their behavior, identity, and desire was overwhelmingly heterosexual. It is difficult to say which of these criteria deserve more weightage for determining who is and is not heterosexual. More reasonably, one might argue that the label is inherently ambiguous, and therefore one cannot count the number of heterosexuals in any society.

According to the framework which was used to analyze the NSHLS, heterosexual expression is informed by what Simon and Gagnon call *sexual scripts*. Sexual scripts are the often improvised and mutually shared narratives that sexual actors use to orchestrate sexual expression, and attribute meaning to their experiences (Gagnon and Simon, 1973). Sexual scripts operate at three levels: cultural, interactional and interpersonal, and intrapsychic. Cultural scripts are collective scenarios which provide the broadest picture of acceptable sexual expression. Interpersonal scripts are shared expectations about the chain of behaviors which will occur in a particular setting and span of time. Intrapsychic scripts are motivational elements that allow an individual to grasp the meaning of their actions, and become committed to those actions. At the cultural level, scripts may appear hegemonic, but individuals are constantly renegotiating and changing sexual scripts in order to create distinctive

Table 2 Prevalence of sexual identity and sexual attraction by gender (percentages)

Sexual identity	*Men*	*Women*	*Sexual attraction*	*Men*	*Women*
Other	0.3	0.1	Only opposite gender	93.8	95.6
Heterosexual	96.9	98.6	Mostly opposite gender	2.6	2.7
Bisexual	0.8	0.5	Both genders	0.6	0.8
Homosexual	2.0	0.9	Mostly same gender	0.7	6
			Only same gender	2.4	0.3
Total N	1401	1732	Total N	1404	1731

Table 1 Prevalence of same-gender and opposite-gender partners (Ps) (percentages)

	Ps in last year		*Ps in past 5 years*		*Ps since age 18*		*Ps since puberty*	
	Men	*Women*	*Men*	*Women*	*Men*	*Women*	*Men*	*Women*
No partners	10.5	13.3	5.9	7.1	3.8	3.4	3.3	2.2
Opposite gender only	86.6	85.4	90.0	90.7	91.3	92.5	90.3	94.3
Both men and women	0.7	0.3	2.1	1.4	4.0	3.7	5.8	3.3
Same gender only	2.0	1.0	2.0	0.8	0.9	0.4	0.6	0.2
Any same-gender sex								
Both men and women	25.3	25.0	51.6	62.9	81.6	89.9	90.7	94.9
Same gender only	74.7	75.0	48.4	37.1	18.4	10.1	9.3	5.1
Total N	3494	4376	2224	2838	3073	3853	1334	1678

Note: Partner variables (last year, past 5 years, and since 18) are from combined GSS and NHSLS data. Partners since puberty are based on age of first vaginal intercourse and age of first same-gender partner. From the NHSLS.

interpretations of particular sexual situations, especially at the interactional level (Masters et al., 2012 (Epub ahead of print)).

Heterosexual behavior is scripted to the degree that actors use scripts to decide what it is they will do in a sexual encounter, how they feel about it, and what they expect to lead up to it and follow from it. Rather than speaking of heterosexual *behavior*, scripts prompt one to speak about heterosexual *expression* as performative action, where scripts intertwine with biological impulses and feelings in generic cultural repertoires. As an example of how sexual scripts shape sexual practice, consider the order of sexual practices which are normatively appropriate in a given sexual encounter. In a nationally representative sample of over 19 000 Australians in 2001–02, researchers found that 13 combinations of sexual practices accounted for 93% of all sexual encounters – half of all encounters were intercourse accompanied by manual stimulation of the man or woman's genitals (Richters et al., 2006). Such extraordinary consistency is probably not the result of random chance, but rather testifies to the power of social expectations that both partners bring to otherwise private, intimate moments.

Also, consider the length of time between meeting a potential sexual partner, and first having sex with that person. Research from the NHSLS showed that it is very rare for this interval to be short, with only 10% of marriages and 37% of more short-term relationships having an interval of less than a month (Laumann et al., 1994: p. 239). Only 1.4% of marriages involved sex within 2 days of meeting (239), testifying to the long negotiations which usually occur between becoming acquainted and first sex, despite reigning cultural myths that sex is the result of spontaneous outpourings of desire (Dune and Shuttleworth, 2009). Courtship negotiations are, in fact, often heavily scripted, which is especially evident when scripting breaks down. For instance, gendered differences in scripts can lead to conflict between men and women over how to understand their relationship to each other at some stage of the courtship process (Rose and Frieze, 1993; Bartoli and Clark, 2006). Courtship also becomes a staging ground for intrapsychic conflicts between dating scripts; for example, American heterosexual men sometimes experience strong tensions between cultural expectations that they become emotionally invested in their female partner, while at the same time performing masculine scripts of courtship-as-conquest (Seal and Ehrhardt, 2003).

As a final note, intrapsychic sexual scripts are related to sexual attitudes, although sexual attitudes are more connected to the individual's morality. Sexual attitudes can be understood as the actor's normative orientation toward sexual conduct – their ideas about where, how, and with whom sex should take place. Like scripts, sexual attitudes also cluster within master status groups, (*Master Status* is a socially defined position, such as age, gender, or martial status, occupied by a person in society that is important in shaping his or her self-concept and life choices (Laumann et al., 1994: p. 31).), since status groups often condition group members to have relatively similar sets of attitudes and normative beliefs. Most obviously, membership in religious master status groups seems to be associated with different attitudes toward sex and procreation; Catholics are far more likely to endorse the view that sex is only appropriate for procreation, and that accordingly, sex which does not result in, or is not intended to result in children, is wrong (de Visser et al., 2007; Laumann et al., 1994). Homosexuality is only one subset of such behavior; premarital sex, oral and anal sex all fall within this class of sexual practices as well. This does not mean that Protestants will necessarily be more likely to condone homosexual practices; conservative protestant groups such as Pentacostals have continued to grow in size and influence as the Catholic church loses membership throughout the globe, and many of these churches have hard-line positions against homosexuality (Adamczyk and Pitt, 2009). By contrast, individuals with no religious affiliation seem more likely to endorse a relational view of sex, where sexuality is best-expressed in committed and loving relationships as a part of two peoples' emotional commitment to each other, or a recreational view of sex, where any kind of sex is acceptable as long as it is between consenting adults (Laumann et al., 1994, 2006). Note that it is just as possible for one's normative orientation to prohibit categorical approval *or* disapproval of certain sexual practices. Compared with any other country in the world, people in Japan are more likely to say that premarital sex is only sometimes wrong, and to avoid condemning or condoning it outright; this may be because Japanese people tend to view premarital sex as acceptable, as long as one is discreet about it (Widmer et al., 1998).

Heterosexual Expression over the Life Course

The theory of sexual scripts also holds that not every script is available to every person at any time. Actors will consider certain scripts to be acceptable for some people rather than others, and thus sexual scripts will adhere to a number of *master statuses* especially gender, race/ethnicity, age, and class (Laumann and Michael, 2001; Laumann et al., 1994). Age organizes acceptable expression because actors refer to age in order to understand what stage of life a person currently inhabits, and what would be normatively appropriate, given this stage (Mahay and Laumann, 2004). Accordingly, where a person falls in their perceived *life course* is a key factor for the social organization of heterosexual behavior.

While there are biological constraints on how soon in her life a female can bear a child, and a male can father one, cross-culturally there is no fixed interval from the onset of sexual readiness to sexual debut. In some cultures, years may pass between menarche and first child, as a matter of a woman's typical life-course trajectory, while in others, the gap between menarche and first child is fleetingly brief (Udry and Cliquet, 1982). The case of China is illustrative here, since social transformations following the cultural revolution narrowed the interval between sexual readiness and sexual debut. Premarital sex became more common for cohorts born after 1950, and the history of previous forms of social organization can be glimpsed across age cohorts of the current Chinese population.

As an illustration of how drastic this change has been, observe **Figure 1**, adapted from Parish et al. (2007). The figures show trends in premarital sex for four cohorts of urban men and women in China, which can be read in the light of China's changing political order. For men reaching age 20 between 1965

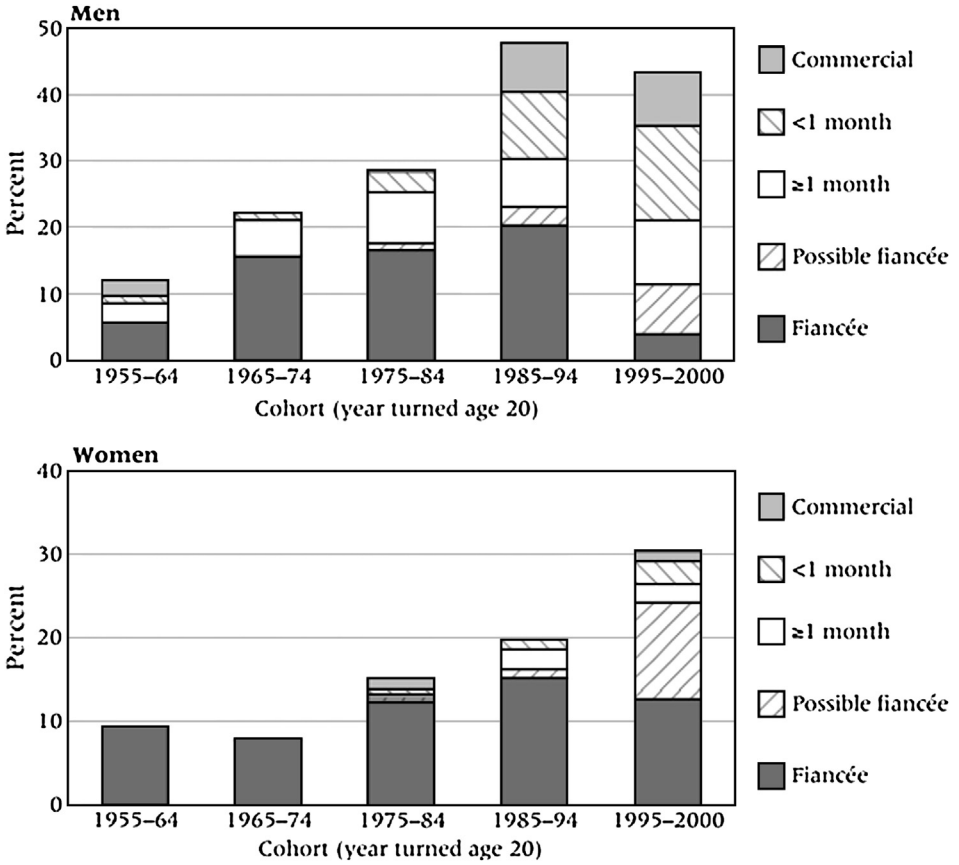

Figure 1 Trends in premarital sex (intercourse) among four cohorts of men and women in urban China. Note: Commercial = commercial sex worker or other paid sex; <1 month = relationship lasting less than a month not followed by marriage; ≥ 1 month = relationship lasting at least a month not followed by marriage; Possible fiancé = current relationship that might lead to marriage; Fiancé = premarital relationship that led to marriage. For people with multiple types of premarital sex, only their shortest-term relationship was counted.

and 1985, use of prostitutes almost vanished, due to Mao's heavy persecution of this practice. However, use of commercial sex workers reappears after 1985, making up almost a fifth of all premarital sexual experiences for men. For women in the 1955 cohort, premarital sex was almost exclusively with one's betrothed, but this proportion declines to about a third of all premarital sexual experience for the 1995 cohort. The authors point out that for all cohorts, men encountered a greater variety of sexual partners than women, indicating the reemergence or persistence of a gendered double standard for sexual behavior.

As this cultural transformation was underway, improvements in childhood nutrition and socioeconomic development affected Chinese ontogeny. Chinese diets improved in quality, quantity, and variety for both urban and rural dwellers (Popkin, 1994; Popkin et al., 2001), and age at start of puberty in China fell from 15.9 to 13.6 years (Parish et al., 2007). These developments paralleled trends seen elsewhere and at other times. In Europe, from 1795 to 1981, the age at menarche fell by 2.3 months per decade; in the United States, age at menarche fell from 14.75 years in 1877 to 12.8 years in 1947, eventually leveling off to remain stable at this age (Wyshak and Frisch, 1982).

The United States is also continuing to undergo transformations in age-appropriate sexual scripts, long after the so-called sexual revolution apparently slowed. As years of educational training become continually lengthened, more and more of Americans' lives are spent in college and postgraduate training (Morgan, 2005), leading to the emergence of new stages in the American life course (Arnett, 2000). Regarding sexual expression, one of the cultural innovations occasioned by this expansion of American higher education system is the so-called hookup culture. Hookups are defined by several key characteristics. First, hookups are not part of a long-term relationship; typically neither partner enters into the sexual union with any promise that this encounter will lead to some future romance. Second, hookups take place without any guarantee about which sexual practices will take place, or what communication will continue afterward (Bogle, 2008; Epstein et al., 2009; Garcia and Reiber, 2008; Owen et al., 2010). Third, hookups generally take place within the setting of college parties, and usually involve the influence of alcohol (Garcia and Reiber, 2008). Currently, no nationally representative studies of colleges have attempted to grasp the social dynamics of hookups, but several smaller studies have provided suggestive results (Garcia and Reiber, 2008; Owen and Fincham, 2011; Armstrong et al., 2010; England et al., 2007). For instance, hookups only seem to include sexual intercourse about a third of the time, and individuals who participate in hookups are not necessarily promiscuous – in one study

by senior year 40% of students who had ever 'hooked up' had experienced three or fewer hookups (Armstrong et al., 2010; England et al., 2007). Male and female respondents who report seeking or experiencing a hookup generally say that they do not want a 'serious' relationship, because they know that after college they will leave their current residence and will have to seek employment elsewhere. They know that their sexual partners are in the same situation, and the practical difficulties of solidifying a long-term sexual relationship, such as a marriage, make hookups a more appealing alternative as an institutionalized way to experience sexual companionship, without forming strong bonds which may eventually be strained by the exigencies of the labor market (Bogle, 2008). Hookup culture is also fomented by the fact that many colleges are relatively small and homogeneous communities, making assortative mating and the decoding of sexual scripts easy (Bogle, 2008). Hookups therefore represent a new concatenation of sexual scripts in the service of sexual expression, ushered into existence by changes in many Americans' life courses.

However, in both China and the United States, it is not yet clear whether or not these cultural transformations will permanently undermine traditional forms of sexual expression. The notion of youth as a period of sexual experimentation is often joined to a normative expectation that young people will subsequently settle into the conventional roles which culturally characterize a later stage of their life course. In the United States, 9% of men report that their first sexual encounter was with another male, but the number of men who settle into long-term homosexual partnerships is far lower (Laumann et al., 1994), indicating the power of dominant scripts to constrain and channel sexual desire into dominant, normative structures.

In general, heterosexual expression in the United States is characterized by punctuated equilibria of short periods of partner-seeking and partner turnover, followed by longer periods of stable sexual partnerships with no turnover and limited or no concurrency of sexual partners (Laumann and Michael, 2001; Das et al., 2011). In other words, for most of their lives, social actors are not active participants in a sexual marketplace.

Tables 3–7 show the lifetime number of sexual partners for each gender in the United States, with estimates derived from the NHSLS. American women report a median of only two lifetime sexual partners, and men report six. The disparities between men and women could be the result of higher sexual partner turnover in homosexual circles, larger pools of potential partners for men due to their greater geographical mobility, or even different definitions of sex between men and women (men are slightly more likely to consider manual stimulation to be sex (Sanders and Reinisch, 1999). The 2002 National Survey of Family Growth (NFSG) found that men over 30 report six to eight lifetime partners, and that women report only four (Mosher et al., 2005). Neither estimate confirms the stereotype that Americans are promiscuous, nor does it suggest that America, following the sexual revolution, completely eliminated scripted constraints on sexual expression.

Furthermore, the vast majority of people in the United States population reports only one sex partner a year. More men than women report two to four sex partners annually, but only 5% of men and less than 2% of women report more than five. Furthermore, it seems the modal frequency of sex for both men and women is only a few times a month, and sexual duration is only usually about 15–60 min per event. Thus the role of partnered sex in American respondents' lives is actually relatively small, as measured by time spent. Sexual partnership, of course, plays a much larger role in other respects, but sex is so heavily organized by sexual scripts that the real probability of a 'free-for-all' of sexual expression is surely very small. In fact, spikes in the number of sexual partners for American men almost always takes place outside of marriage, and higher rates of coitus among African-Americans is likely the result of lower marriage rates (Mosher et al., 2005).

Table 3 The vast majority of the US population aged 18–59 has only one sex partner within a year

	Number of sex partners in the past 12 months (%)			
	0	*1*	*2–4*	*5+*
All	11.9	71.1	13.7	3.2
Men	9.9	66.7	18.3	5.1
Women	13.6	74.7	10.0	1.7

Table 4 The median number of sex partners since age 18 is 2 for women and 6 for men

	Number of sex partners since age 18						
	0	*1*	*2–4*	*5–10*	*11–20*	*21+*	*Median*
All	2.9	26.1	29.5	21.7	10.6	9.2	3
Men	3.4	19.5	20.9	23.3	16.3	16.6	6
Women	2.5	31.5	36.4	20.4	6.0	3.2	2

Heterosexual Practice, and the Spread of Sexually Transmitted Diseases

The human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome (AIDS) crisis was the principal reason that sex research again achieved national prominence in America. If

Table 5 Americans do not have sex all that frequently, and not for that long a duration when they do have sex

	Frequency of sex in the past year				
	Not at all	*A few times a year*	*A few times a month*	*2–3 times a week*	*4+ times a week*
Men	9.8	17.6	35.5	29.5	7.7
Women	13.6	16.1	37.2	26.3	6.7

	Duration of last sexual event (%)		
	Less than 15 min	*15–60 min*	*60 or more minutes*
Men	11.0	69.3	19.7
Women	14.7	70.5	14.8

Table 6 A large majority of Americans do not find many sex practices very appealing

	Percentage saying practice is 'very appealing'		*Percentage engaged in practice in last event*		*Percentage engaged in practice ever*	
	Men	*Women*	*Men*	*Women*	*Men*	*Women*
Vaginal intercourse	83.8	76.8	94.6	95.6	95.0	97.0
Watching partner undress	47.8	26.8				
Receiving oral sex	45.0	28.8	27.5	19.9	78.7	73.1
Giving oral sex	33.5	16.5	26.8	18.8	76.6	67.7
Group sex	13.3	1.1				
Using a dildo or vibrator	4.4	2.9				
Watching others have sex	5.3	1.5				
Same-gender sex	3.2	2.9	2.7[a]	1.3[a]	4.9[b]	4.1[b]
Having sex with a stranger	4.1	0.9				
Anal intercourse	2.8	1.0	2.3	1.2	25.6	20.4
Forcing someone to have sex	0.3	0.2			2.8	1.5
Being forced to have sex	0.1	0.1			3.9	22.8

[a]Last 12 months, not last event.
[b]Since age 18, not ever.

Table 7 These same men and women report themselves to be quite happy with their sex lives

	Satisfaction with sex life (%)		
	Extremely or very happy	*Generally satisfied*	*Unhappy*
All	57.7	29.2	13.1
By number of sex partners in past 12 months			
0	40.7	35.4	23.9
1	63.9	27.2	9.4
2–4	44.9	32.7	22.4
5+	47.2	37.1	15.7
By frequency of sex in past 12 months			
None	39.5	36.7	23.8
1–2 times per year	45.6	32.8	21.6
2–3 times per month of once/week	59.2	29.3	11.5
2–3 times per week	69.0	25.1	5.9
4+ times per week	64.5	22.0	13.6

Kinsey's picture of a continuum between heterosexual and homosexual behavior was the case, then it was plausible that HIV could readily jump from the stigmatized 'high-risk' homosexual population to the heterosexual mainstream. As a side-effect of research into sexual practice related to sexually transmitted disease (STD) transmission, research into a wider variety of sexual practices became more politically legitimate, and sex researchers were able to develop a broader understanding of sexual practices.

Delineating what is and is not 'heterosexual' practice is difficult, in part, because of practices which some heterosexual couples find pleasurable in conjunction with intercourse, but which have nothing to do with genital stimulation, such as kissing or hugging. Masturbation might also be counted as a kind of heterosexual expression, if it takes place in the context of heterosexual fantasies, or in the presence of a heterosexual partner. (According to the NHSLS, 58% of women report that they do not masturbate at all, compared to only 36% of men (Laumann et al., 1994: p. 83), numbers which are still remarkably high given several social movements in American society aimed at reducing the incidence of masturbation, including circumcision (in a case of historical irony, circumcised males are actually more likely to masturbate; Laumann et al., 1997).) For the sake of parsimony, we will focus on only four types of heterosexual practice here: vaginal-penile intercourse, cunnilingus, fellatio, and anal sex. The ways in which cultural scripts orchestrate these practices, shape the enjoyment actors derive from them, and the implications of the social organization of sexual practice for the transmission of STDs will be briefly reviewed in this section.

By and large, the same picture of conventional heterosexual expression in America holds for sexual practices. Three-quarters of American men and women report some lifetime experience with oral sex, but only one-quarter report that it took place during their last heterosexual coupling (Laumann et al., 1994: p. 101). It is probably worth resuming a brief historical discussion here. Based on a cohort analysis from the NHSLS, oral sex seems to become widely practiced in the United States in the 1920s, long before the sexual revolution. It appears that nationwide rates remained relatively low (less than 70% of the sexually mature population) until the early 1950s, when rates of lifetime experience with oral sex climb to 90% (Laumann et al., 1994: p. 104). For cohorts who reached sexual maturity after the early 1950s, oral sex became a readily available and relatively unstigmatized technique for obtaining sexual pleasure.

Rates of anal sex are even lower in America, with 26% of men aged 18–59 reporting any anal sex with a woman in their lifetimes, and only 9% experiencing heterosexual anal sex in the past year (Laumann et al., 1994: p. 107). Later studies employing the 2002 NFSG report higher rates: approximately 38% aged 18–44, which could be due to the exclusion of older respondents (Mosher et al., 2005). The rate of heterosexual anal sex seems to diverge according to master

status group membership as well: highly educated persons, respondents with no religious affiliation, and Hispanic men all report higher rates of anal sex (Laumann et al., 1994: p. 109). Regardless of subdivisions of the sample, the statistics suggest that anal sex is much rarer than oral sex among American heterosexuals.

Compare the numbers above with rates of penile-vaginal sex in America. According to the NHSLS, over 95% of men and women in the United States report some lifetime experience with heterosexual intercourse (Laumann et al., 1994: p. 97). Even more noteworthy: this rate did not appear to vary by master status, except younger age cohorts which have not had as much time to accumulate partners. Numbers from the NFSG are similar, reporting 97% of men and 98% of women between the ages of 25–44 ever having had penile-vaginal intercourse (Mosher et al., 2005). At last sexual contact, individuals with opposite-sex partners were also overwhelmingly likely to say they engaged in penile-vaginal penetrative sex (95% of men and 96% of women; Laumann et al., 1994: p. 100). This finding is also stable across all master statuses, which helps us to understand why the word 'sex' is often used by English speakers as if it were synonymous with this specific sexual practice.

One might expect that sexual expression of all kinds gradually declines as people age sexual expression of all kinds gradually declines and sexual impulses weaken. Recent analyses of the sexual life of America's older population suggest that this may not be the case. The National Social Life Health and Aging Project (NSHAP), which fielded its first wave in 2005, found that over 60% of men between the ages of 66 and 70 who report being in good health have had sex in the past year, and that more than 35% of women in this age range who are in good health report sex in the past year (Lindau et al., 2007). The lowest rates were among older persons (81–85), in poor or fair health, and even then, more than 10% reported sex in the past year. Partnership was also strongly associated with sex in the past year: among those who were not in a romantic relationship, only 22% of men and 4% of women reported being sexually active in the past 12 months. This may help to explain why older women were overall less likely to have had sex in the past year, since women were less likely to be partnered. Recall that if most sex takes place in the context of monogamous relationships, as long as physical function continues to permit sexual performance, then sexual expression will continue, and perhaps even thrive.

Paradoxically, for all of this apparently conventional sexual expression in mutually exclusive sexual relationships, STDs are common in America. **Table 8** shows the prevalence of STDs by gender and number of sex partners for the American population using estimates derived from the NHSLS. Nine percent of men and 5% of women report gonorrhea in their lifetimes, and about 6% of women report genital warts. Taken together, about 16% of men and 18% of women have ever had a sexually transmitted infection, and the incidence of STDs seems to be a function of the number of sexual partners that they have (Laumann et al., 1994: p. 425). It is noteworthy that for every level of risk as indexed by numbers of lifetime partners, women are almost twice as likely to report having had an STD. This disparity in likelihood of contracting STDs by gender may be explained, in part, by the fact that women's partners are more risky in the sense of their having had more sex partners than men's partners who, being women, have a lower likelihood of having had multiple partners. There may also be physiological reasons which leave women more vulnerable to STD infection than men.

Table 8 The Prevalence of STDs in the United States and the risk factors for those diseases; one in every six men and women reported having an STD at some point in their lives, and those with more partners are dramatically more likely to have a bacterial or viral infection

	Respondent's gender (%)		*Respondent's number of sex partners since age 18 (cases per 1000)*					
	Male	*Female*	*0*	*1*	*2–4*	*5–10*	*11–20*	*21+*
Bacterial								
Gonorrhea	9.0	4.7	0	11	30	83	148	206
Syphilis	.9	.7	0	1	4	13	9	21
Chlamydia	1.9	4.4	0	14	20	44	64	70
NGU[a]	1.9		0	0	4	16	31	62
PID[b]		2.2	0	9	12	40	77	54
Any bacterial	12.1	10.6	0	28	58	150	241	300
Viral								
Genital warts	3.3	5.9	0	11	37	70	91	98
Herpes	1.2	2.9	0	5	10	19	51	80
Hepatitis	1.3	0.9	11	5	5	12	24	25
HIV/AIDS	0.2	0.1	0	0	0	2	3	7
Any viral	5.4	9.0	11	17	50	98	159	177
Any STD	15.9	17.8	11	43	104	234	338	404

[a]Nongonococcal urethritis.
[b]Pelvic inflammatory disease.

These high rates of STD infection have also led some to posit a 'core' and 'periphery' model of sexual networks. At any given time, there may be some individuals who are highly connected by sex to other, similarly highly connected individuals, who form a core of sexual contact with high risk of STD transmission. Beyond this core is a periphery of disconnected persons who have relatively few sex partners, and whose sex partners are rarely connected to anyone else. By one account, higher rates of STD infection among African-Americans may be due to the fact that the African-American core is not as separated from the periphery, and thus the probability of a new tie forming between core and peripheral members is more likely for blacks than for whites, making it easier for STDs to escape the core (Laumann and Youm, 1999). A comparable network-based study of sexuality in China found that Chinese women (well over 90% of whom have had only one partner in their lifetime) have a higher rate of Chlamydia infections than men, even though men report more partners and make much more use of commercial sex workers (Parish et al., 2003). Connecting women to their male partners' sex networks sheds light on this issue.

The NHSLS also shows that master statuses are powerfully predictive of partner choice. Observe **Table 9**, which provides a breakdown of homophilous (same-status) partnering according to the self-reports of respondents for four relationship categories, in descending order of purported commitment: marriages, cohabitations, longer-term partnerships, and short-term partnerships. Partner choice requires a certain amount of shared sexual scripts and outlooks on life in order for both actors to communicate what they expect from a partner, and

Table 9 Percentage homophilous by relationship type (only marriages and cohabitational relationships that began in the last 10 years)

Type of homophily	Type of relationship			
	Marriages	Cohabitation	Long-term partnerships	Short-term partnerships
Racial/ethnic	93[a]	88	89	91
Age[b]	78	75	76	83
Educational[c]	82	87	83	87
Religious[d]	72	53	56	60

[a]Indicates weighted percentage of marriages that began within the past 10 years and involved two people of the same race/ethnicity. Cases in which either partner was reported as Native American or other or had missing data were omitted.
[b]Age homophily is defined as a difference of no more than 5 years between partners' ages.
[c]Educational homophily is defined as a difference of no more than one educational category. The categories are less than high school, high school, vocational training, 4-year collegeor more.
[d]Cases in which either partner was reported as 'other' or had missing cases are omitted.

to tacitly understand that there is some sexual attraction between them, and as we have already argued above, these scripts are organized around master status categories. Partner choice also often occurs in the context of other, non-partner confidants and acquaintances, who may ask the respondent to explain why it is that they are seeking to initiate a relationship outside their status group (Laumann et al., 1994). Thus we might predict that lower-investment relationships would show less status-homophily than higher-investment relationships, since longer contact with a partner often means introducing that new person into the respondent's network of kin and friends, which are also, often, homophilous (McPherson et al., 2001). A quick look at the numbers reveals that this is not the case. Partnerships at every level are highly homophilous on race and education, with about 90% of partnerships of any kind being homophilous with regards to race. More recently collected data from the United States census confirms that less than 10% of African-Americans marry outside their race (Fryer, 2007). Coupled with the findings on African-Americans' sexual networks above, and taking into account the intersection of race, education, religion and age, the possibility of a sexually transmitted disease outbreak in America that crosses into and permeates every master status group is fairly low.

Sexual Dysfunction and Sexual Well-Being

Data from the NHSLS and NSHAP show that sexual dysfunctions are common among younger and older adults (Lindau et al., 2007; Laumann et al., 2008). In **Table 10**, we can see that 37% of American men over the age of 57 suffer from some degree of erectile dysfunction, and 28% experience premature ejaculation. Among women, the most common form of sexual problem is lacking interest in sex (43%) followed by having problems with lubrication (39%), and not achieving orgasm (34%). NHSLS data also showed that rates of sexual dysfunction are higher for women than men at any age (Laumann et al., 1999).

Table 10 Prevalence of sexual dysfunctions among older Americans, aged 57–85

Variable	Respondents (number)	Age group (%)		
		57–64 years	65-74 years	75-85 years
Lack of interest in sex				
Men	881	28.2	28.5	24.2
Women	504	44.2	38.4	49.3
Difficulty getting or maintaining an erection				
Men	874	30.7	44.6	43.5
Difficulty with lubrication				
Women	495	35.9	43.2	43.6
Climaxing too quickly				
Men	860	29.5	28.1	21.3
Inability to climax				
Men	866	16.2	22.7	33.2
Women	479	34.0	32.8	38.2
Anxious about performance				
Men	873	25.1	28.9	29.3
Women	500	10.4	12.5	9.9

Data gathered from the Global Study of Sexual Attitudes and Behaviors (GSSAB) also suggest that sexual dysfunction is common around the globe (Laumann et al., 2005). In Korea, 33% of male respondents aged 40–80, reported problems with premature ejaculation, and 32% had trouble getting or maintaining an erection. Thirty-seven percent of Korean women reported a lack of pleasure in their sexual experiences, and 32% reported an inability to reach orgasm (Moreira et al., 2006). Other data from the same study found that 43% of men and 49% of women throughout the world report at least one sexual problem (Moreira et al., 2005; Laumann et al., 2005).

Just as sexual scripts help social actors understand sexual performance, scripts inform the meaning of sexual dysfunction. While some respondents consider a lack of sexual interest to be 'a problem,' others, who understand losing interest in sex to be a natural, and perhaps even desirable part of aging, may not feel as troubled. Recall the low frequency of sex even within marriage: the overall good functioning of a long-term partnered relationship does not hinge on coital frequency or satisfaction alone, and other benefits such as companionship and social support may keep both partners happy in their current state of affairs, without a strong sexual component to their relationship.

Consider these GSSAB data from Korea. While both America and Korea report similarly high levels of sexual dysfunction at older ages, Korean men and women are less likely to report their conditions as bothersome. Furthermore, only 5% of men and 3% of women have been asked by their doctor about their sexual lives (Moreira et al., 2006). Globally, East Asian men are the least likely to seek help for their sexual lives (6.5%) while South-East Asian men are most likely (22.5%; Moreira et al., 2005). This difference in rates seems to be attributable to ideas about

the relationship between sexual function and aging: that sexual ability does decline as one ages, and so there is nothing 'abnormal' about being unable to have sex, or being uninterested in it.

Data from the GSSAB also provide us with a window into global rates of satisfaction with one's sex life. Laumann and colleagues performed a cluster analysis on individual-level data gathered from 29 countries, and clustered those respondents according to their satisfaction with their sexual lives, grouping them by country (Laumann et al., 2006). Three clusters emerged which could be ranked ordinally by the average amount of satisfaction with their sex lives. Table 11 shows the results. We can see that the 'high satisfaction' cluster contains mostly Western, industrialized countries such as Australia, Canada, and Spain, while the 'low' cluster comprises mostly East Asian countries. Again, we should be cautious about interpreting these results since while East Asian peoples may be less satisfied with their sexual lives, this does not have to reflect on their happiness with their spouses. There may be other cultural factors which make long-term heterosexual partnership in East Asia particularly burdensome for one gender or the other, but these data do not speak to that issue.

Conclusion

In this article, we have not focused on biological or psychological factors that may underpin the frequency or patterning of heterosexual expression, but that does not mean that we believe that sociological factors simply override them. Neither do we consider biological or psychological accounts of sexuality to be overly simplistic, but rather the opposite, that they involve

Table 11 Four aspects of subjective sexual well-being arranged by country and gender vertically, arranged by cluster membership horizontally; cells contain percent of persons responding positively in each of the four domains

	Physical pleasure[a]		*Emotional pleasure*[b]		*Satisfaction with sexual function*[c]		*Importance of sex*[d]	
Country	*Men*	*Women*	*Men*	*Women*	*Men*	*Women*	*Men*	*Women*
1. Western European and European-linked Western countries (high levels of sexual satisfaction) N = 14503 (Male = 7224, Female = 7279)								
Australia	67.7	62.9	72.3	68	82.1	82.7	43.6	29.7
Austria	79.7	63	83.4	70.6	91.4	91.1	56.3	35.5
Belgium	66.2	65.1	72.3	67.3	86.7	78.1	48.5	32.3
Canada	71.2	67.8	73.7	62.5	87.4	87.6	44.6	33.8
France	61.2	51.7	59.9	53.2	85.2	69.6	55.8	39.4
Mexico	69.1	57.4	71.8	62.3	85.2	64.4	59	36.6
New Zealand	71	62.5	76.9	61.6	70.6	64.1	35.7	21.6
South Africa	60.6	46.9	64.7	46.7	77.7	65.5	54.5	26.3
Spain	72.9	67.5	76	69.1	90.2	78.4	64.1	33.8
Sweden	58.2	55.9	57.8	59.7	89.4	84.5	46.4	31.7
Germany	62.5	56.8	69	59.9	87.7	85.9	41.1	33.1
United Kingdom	60.9	56	71.3	67.5	73.9	74.6	44.8	29.6
USA	72.9	65	77.1	68	83.9	82.4	36.6	27.7
Mean	67.3	59.9	71.2	62.8	83.9	77.6	48.6	31.6
2. Islamic and selected Asian and European countries (middle levels of sexual satisfaction) N = 8997 (Male = 4394, Female = 4603)								
Algeria	21.1	33.7	46.3	24.6	70.2	47.6	53.9	33.8
Brazil	59.3	39.7	60.8	41.3	88.2	69.6	74.6	43.1
Egypt	36.2	23.6	44.4	27.6	73.2	52.2	52.8	40.4
Israel	51.8	47.7	59.3	54.8	76	75.1	64.2	53.6
Italy	42.8	31.6	48.8	40.6	90.9	82.1	58	32.7
Korea	47.3	36.2	54	47.2	60.9	46.3	73.1	46.4
Malaysia	45.6	50.2	50.9	58	60.1	44.5	45.8	34.9
Morocco	47.4	37	53.7	45	82.2	77.3	73.3	42.6
Philippines	46.5	40.2	48.7	38.5	71.3	56.8	56.9	30.2
Singapore	46.8	34	50.6	39.2	65.1	55.1	43.4	25.9
Turkey	50.9	25.4	55.6	33	88.6	56.7	69.5	28.6
Mean								
3. East Asian countries (low levels of sexual satisfaction) N = 3500 (Male = 1750, Female = 1750)								
China	25.2	24.4	36	32.8	68.7	45.5	28.8	17.5
Indonesia	22	24.7	18.5	19.9	73.8	61	35.4	16.6
Japan	17.6	9.8	23.6	15.5	60.3	39.7	27.8	12
Taiwan	21.5	19.4	28.7	25.4	60	42.3	24.7	7.3
Thailand	38.1	21	42.6	22.8	67.6	61.3	23.8	10.1
Mean	24.9	19.8	29.9	23.3	66.1	50	28.1	12.7

[a]Extremely/very pleasurable as opposed to moderately/slightly/not at all pleasurable.
[b]Extremely/very satisfying as opposed to moderately/slightly/not at all satisfying.
[c]Very/somewhat satisfied as opposed to neither satisfied nor dissatisfied or somewhat/very dissatisfied.
[d]Extremely/very important as opposed to moderately/slightly/not at all important.

systems of nearly overwhelming complexity. Sexual preferences, desires, and resilience to sexual infection or dysfunction are all apparently enmeshed in biological and psychological developments which can originate in the uterine environment (Udry, 2000; Udry and Talbert, 1988), and individually based experiences throughout childhood, adolescence, and adulthood. Going forward, we need to broaden our theories to understand more thoroughly where and how the biological, the psychological, the social, and the cultural intersect and interact to create the observed patterns of sexual expression.

See also: Adolescent Sexual Risk; Catholicism and Gender; Gender Development: Cultural Differences; Heterosexism and Homophobia; Judaism and Gender; Sex Role Development and Education; Sex Selective Abortion; Sexism; Sexual Orientation and the Law; Sexual Orientation: Biological Influences; Sexual Orientation: Historical and Social Construction; Sexual Politics in the Arab World.

Bibliography

Adamczyk, A., Pitt, C., 2009. Shaping attitudes about homosexuality: the role of religion and cultural context. Social Science Research 38, 338–351.

Armstrong, E., Hamilton, L., England, P., 2010. Is hooking up bad for young women? Contexts 9, 22–27.

Arnett, J.J., 2000. Emerging adulthood: a theory of development from the late teens through the twenties. American Psychologist 55, 469–480.

Bartoli, A.M., Clark, M.D., 2006. The dating game: similarities and differences in dating scripts among college students. Sexuality and Culture 10, 54–80.

Bogle, K.A., 2008. Hooking up: Sex, Dating and Relationships on Campus. NYU Press, New York.

Cochran, W.G., Mosteller, F., Tukey, J.W., 1953. Statistical problems of the Kinsey report. Journal of the American Statistical Association 264, 673–716.

Das, A., Laumann, E.O., 2010. How to get valid answers from survey questions: what we learned from asking about sexual behavior and the measurement of sexuality. In: Walford, G., Viswanathan, M., Tucker, E. (Eds.), The SAGE Handbook of Measurement. Sage, Thousand Oaks, CA.

Das, A., Laumann, E.O., Waite, L.J., 2011. Sexual expression over the life course: results from three landmark surveys. In: Delamater, J., Carpenter, L. (Eds.), Sexualities over the Life Course: Emerging Perspectives. NYU Press, New York.

Dune, T.M., Shuttleworth, R.P., 2009. 'It's just supposed to happen': the myth of sexual spontaneity and the sexually marginalized. Sexuality and Disability 27, 97–108.

England, P., Shafer, E.F., Fogarty, A.C.K., 2007. Hooking up and forming romantic relationships on today's college campuses. In: Kimmel, M. (Ed.), The Gendered Society Reader. Oxford University Press, New York.

Epstein, M., Calzo, J.P., Smiler, A.P., Ward, L.M., 2009. "Anything from making out to having sex": Men's negotiations of hooking up and friends with benefits scripts. Journal of Sex Research 46, 414–424.

Ericksen, J., Steffen, S.A., 1994. Sexual behavior surveys before Kinsey. In: Annual Meeting of the Society for the Scientific Study of Sex. Miami, FL.

Fryer, R.G., 2007. Guess who's been coming to dinner? Trends in interracial marriage over the 20th century. Journal of Economic Perspectives 21, 71–90.

Gagnon, J.H., Simon, W., 1973. Sexual Conduct: The Social Sources of Human Sexuality. Aldine Transaction, New York.

Garcia, J.R., Reiber, C., 2008. Hook-up behavior: a biopsychosocial perspective. Journal of Social, Evolutionary, and Cultural Psychology 2, 192–208.

Kinsey, A.C., Pomeroy, W.B., Martin, C.E., 1948. Sexual Behavior of the Human Male. Indiana University Press, Bloomington, IN.

Kinsey, A.C., Pomeroy, W.B., Martin, C.E., Gebhard, P.H., 1953. Sexual Behavior in the Human Female. Indiana University Press, Bloomington, IN.

Laumann, E.O., Das, A., Waite, L.J., 2008. Sexual dysfunction among older adults: prevalence and risk factors from a nationally representative U.S. probability sample of men and women 57–85 years of age. Journal of Sexual Medicine 5, 2300–2311.

Laumann, E.O., Gagnon, J.H., Michael, R.T., Michaels, S., 1994. The Social Organization of Sexuality: Sexual Practices in the United States. University of Chicago Press, Chicago.

Laumann, E.O., Masi, C.M., Zuckerman, E.W., 1997. Circumcision in the United States. Journal of the American Medical Association 277, 1052–1057.

Laumann, E.O., Michael, R.T., 2001. Introduction: setting the scene. In: Laumann, E.O., Michael, R.T. (Eds.), Sex, Love and Health in America: Private Choices and Public Policies. University of Chicago Press, Chicago, IL.

Laumann, E.O., Nicolosi, A., Glasser, D.G., Paik, A., Gingell, C., Moreira, E., Wang, T., 2005. Sexual problems among women and men aged 40–80 y: prevalence and correlates identified in the Global Study of Sexual Attitudes and Behaviors. International Journal of Impotence Research 17, 39–57.

Laumann, E.O., Paik, A., B.Glasser, D., Kang, J.-H., Wang, T., Levinson, B., Moreira, E.D., Nicolosi, A., Gingell, C., 2006. A cross-national study of subjective sexual well-being among older women and men: findings from the Global Study of Sexual Attitudes and Behaviors. Archives of Sexual Behavior 35, 143–159.

Laumann, E.O., Paik, A., Rosen, R.C., 1999. Sexual dysfunction in the United States: prevalence and predictors. Journal of the American Medical Association 281, 537–544.

Laumann, E.O., Youm, Y., 1999. Racial/ethnic group differences in the prevalence of sexually transmitted diseases in the United States: a network explanation. Sexually Transmitted Diseases 26, 250–261.

Lindau, S.T., Schumm, L.P., Laumann, E.O., Levinson, W., O'muircheartaigh, C.A., Waite, L.J., 2007. A study of sexuality and health among older adults in the United States. The New England Journal of Medicine 357, 762–774.

Mahay, J., Laumann, E.O., 2004. Meeting and mating over the life course. In: Laumann, E.O., Ellingson, S., Mahay, J., Paik, A., Youm, Y. (Eds.), The Sexual Organization of the City. University of Chicago Press, Chicago, IL.

Masters, N.T., Casey, E., Wells, E.A., Morrison, D.M., 2012. Sexual scripts among young heterosexually active men and women: continuity and change. Journal of Sex Research [Epub ahead of print].

McConaghy, N., 1987. Heterosexuality/homosexuality: dichotomy or continuum? Archives of Sexual Behavior 16, 411–424.

McConaghy, N., Buhrich, N., Silove, D., 1994. Opposite sex-linked behaviors and homosexual feelings in the predominantly heterosexual male majority. Archives of Sexual Behavior 23, 565–577.

McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a feather: homophily in social networks. Annual Review of Sociology 27, 415–444.

McWhirter, D.P., Sanders, S.A., Reinsch, J.M., 1990. Homosexuality/Heterosexuality: Concepts of Sexual Orientation. Oxford University Press, New York.

Moreira, E.D., Brock, G., Glasser, D.B., Nicolosi, A., Laumann, E.O., Paik, A., Wang, T., Gingell, C., 2005. Help-seeking behaviour for sexual problems: the Global Study of Sexual Attitudes and Behaviors. International Journal of Clinical Practice 59, 6–16.

Moreira, E.D., Kim, S.C., Glasser, D., Gingell, C., 2006. Sexual activity, prevalence of sexual problems, and associated help-seeking patterns in men and women aged 40–80 years in Korea: data from the Global Study of Sexual Attitudes and Behaviors (GSSAB). The Journal of Sexual Medicine 3, 201–211.

Morgan, S., 2005. On the Edge of Commitment: Educational Attainment and Race in the United States. Stanford University Press, Stanford, CA.

Mosher, W.D., Chandra, A., Jones, J., 2005. Sexual behavior and selected health measures: men and Women 15–44 years of age, United States, 2002. Advance Data from Vital and Health Statistics 362, 1–55.

Owen, J., Fincham, F.D., 2011. Young adults' emotional reactions after hooking up encounters. Archives of Sexual Behavior 40, 321–330.

Owen, J.J., Rhoades, G.K., Stanley, S.M., Fincham, F.D., 2010. "Hooking up" among college students: demographic and psychosocial correlates. Archives of Sexual Behavior 39, 653–663.

Parish, W.L., Laumann, E.O., Cohen, M.S., Pan, S., Zheng, H., Hoffman, I., Wang, T., Ng, K.H., 2003. Population-based study of chlamydial infection in China. Journal of the American Medical Association 289, 1265–1273.

Parish, W.L., Laumann, E.O., Mojola, S.A., 2007. Sexual behavior in China: trends and comparisons. Population and Development Review 33, 729–756.

Popkin, B.M., 1994. The nutrition transition in low income countries: an emerging crisis. Nutrition Reviews 52, 285–298.

Popkin, B.M., Horton, S., Kim, S., Mahal, A., Suigao, J., 2001. Trends in diet, nutritional status and deit related non-communicable diseases in Chinca and India: the economic costs of the nutrition transition. Nutrition Reviews 59, 379–390.

Richters, J., Visser, R.D., Rissel, C., Smith, A., 2006. Sexual practices at last heterosexual encounter and occurrence of orgasm in a national survey. Journal of Sex Research 43, 217–226.

Rose, S., Frieze, I.H., 1993. Young singles' contemporary dating scripts. Sex Roles 28, 499–509.

Sanders, S.A., Reinisch, J.M., 1999. Would you say you "Had Sex" if…? Journal of the American Medical Association 281, 275–277.

Seal, D.W., Ehrhardt, A.A., 2003. Masculinity and urban men: perceived scripts for courtship, romantic and sexual interactions with women. Culture, Health and Sexuality 5, 295–319.

Storms, M.D., 1980. Theories of sexual orientation. Journal of Personality and Social Psychology 38, 783–792.

Udry, J.R., 2000. Biological limits of gender construction. American Sociological Review 65, 443–457.

Udry, J.R., Cliquet, R.L., 1982. A cross-cultural examination of the relationship between ages at menarche, marriage and first birth. Demography 19, 53–63.

Udry, J.R., Talbert, L.M., 1988. Sex hormone effects on personality at puberty. Journal of Personality and Social Psychology 54, 291–295.

Van Wyk, P.H., Geist, C.S., 1984. Psychosocial development of heterosexual, bisexual and homosexual behavior. Archives of Sexual Behavior 13, 505–544.

de Visser, R.O., Smith, A.M.A., Richters, J., Rissel, C.E., 2007. Associations between religiosity and sexuality in a representative sample of Australian adults. Archives of Sexual Behavior 36, 36–46.

Widmer, E.D., Treas, J., Newcomb, R., 1998. Attitudes toward nonmarital sex in 24 countries. The Journal of Sex Research 35, 349–358.

Wyshak, G., Frisch, R.E., 1982. Evidence for a secular trend in age of menarche. The New England Journal of Medicine 306, 1033–1035.

Heterosexuality in Africa

Signe Arnfred, Centre for Gender, Power and Diversity, Roskilde University, Roskilde, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ways of understanding sexuality in Africa have deep colonial roots. In classical anthropology – itself an aspect of colonialism – 'African sexuality' was studied as an aspect of marriage, family, and kinship. With the onset of the HIV/AIDS pandemic sexuality in itself emerged as a field of study, however, closely linked to risk, disease, and danger. Critical of these assumptions, postcolonial scholarship has introduced broader approaches to studies of sexuality. These approaches include a new focus on sexual pleasure as part of 'being African the modern way,' for men as well as for women.

One cannot talk about sexuality in Africa without talking also about colonialism. Ideas and conceptualizations of sexuality in Africa were conceived in the minds of Western missionaries and colonialists, colored by their ideas of morality and marriage. From the very first encounters of European travelers, missionaries and (later) colonial civil servants with Africans, 'African sexuality,' male and female, was understood to be 'naturally heterosexual.'

In the late nineteenth century, the hey-day of colonialism when Africa was divided between European powers (cf the Berlin Conference 1884–85) homosexuality was generally perceived as a product of an overdeveloped, decadent urban society, and thus not applicable to Africa. Heterosexuality, on the other hand, was understood as the healthy and natural state of affairs. At the time of colonization, according to Marc Epprecht, Africans figured significantly in debates within Europe about the meaning of civilization and of virility versus decadence. "Since the prevailing prejudice was that Africans were uncivilized and close to nature, by definition they could not be decadent or exhibit social traits and behaviors that were assumed to come with a sophisticated level of culture" (Epprecht, 2013: 59). Due to African peoples' close relationship with nature, they were presumed to be 'ruled by instinct and culturally unsophisticated' for which reason they "had to be heterosexual, [their] sexual energies and outlets devoted exclusively to their 'natural' purpose: biological reproduction" (Murray and Roscoe, 1998: xi). This colonial construction of Africans as naturally heterosexual has repercussions even today; actually it is only in the very last decade of the twentieth century that these colonial ideas of sexuality in Africa have started to be challenged.

Because of the massive impact of colonial ideas on the ways in which sexualities in Africa have been and still are conceptualized, by Western development agencies as well as by Africans themselves, in the first section of this article some of these ideas – their background as well as their characteristics – will be presented. With the HIV/AIDS pandemic a new wave of research on sexuality in Africa was set in motion, quite a lot of it rooted in colonial ideas, as further discussed below. This new focus on sexuality/sexualities in Africa has, however, also been a breeding ground for postcolonial critique and alternative analytical approaches, based less on colonial assumptions, and more on investigations into sexual relations as experienced by African men and women. In the second section of the article some findings from this new research will be presented, still with a focus on heterosexuality, even if a basic finding of this line of research is that a clear dividing line between heterosexuality and homosexuality cannot possibly be drawn. Even men and women, who practice and enjoy same-sex relationships, will very often also be involved in a reproductive heterosexual relationship: as a human being you should contribute to the continuation of your lineage (Murray and Roscoe, 1998).

Section I: Colonial Ideas of 'African Sexuality'

In classical anthropology – itself an aspect of colonialism, in as far as British, French, and Portuguese anthropologists were posted in the colonized territories more or less on behalf of the colonizing powers, and sharing their mindset regarding notions of morality and marriage – sexuality was studied as an aspect of marriage, family, and kinship (Spronk, 2014). Sexuality was, as already noted by Murray and Roscoe (1998), perceived as a means to reproduction. Generally speaking, sexuality – always and without question seen as heterosexuality – was synonymous with marriage, colonial and missionary campaigns being concerned about forms of marriage, i.e., that marriage should be monogamous and permanent, as prescribed by Christianity. Aberrations like polygamy, frequent divorce, bride price, and early marriage were condemned. Missionaries fought an everlasting battle defending a Christian, European conception of family and marriage, a two-generation family based on monogamous marriage (father–mother–child). This in spite of the fact that – as has been explained by postcolonial scholars such as Niara Sudarkasa (1996) – African family forms are often differently constructed, more along lines of consanguinity over several generations (grandparent–parent–child relations, and relations between siblings) than along lines of conjugality (husband–wife relations). At the same time it should be noted that this consanguinity is not necessarily biological, it may also be ascribed (Nzegwu, 2004).

Sexuality, Morality, and HIV/AIDS

Missionaries and colonialists defended a Christian morality, linking sexuality to marriage in ways which were not

 http://dx.doi.org/10.1016/B978-0-08-097086-8.35039-5

necessarily in tune with African mores. As early as 1987 Nici Nelson noted, based on fieldwork among women beer-brewers in Mathare Valley in Nairobi, that "Christian philosophy has always made a separation between body and spirit: things of the body (sexuality?) are valued less than things of the spirit. … [The] fear of women's unbridled sexuality has continued through the Christian era … In Africa this division between the flesh and the spirit into immoral and moral does not seem to exist" (Nelson, 1987: 235).

It was only with the onset of the HIV/AIDS pandemic in the last part of the twentieth century that sexuality in itself emerged as a field of study. As said by Sylvia Tamale, it is the HIV/AIDS pandemic which has 'flung open the doors on sexuality' (Tamale, 2005a). From an African point of view "it has forced into the open the myths and secrets in relationships and identities that are often silenced or taken for granted" (Tamale, 2005a). In more general terms, the global onslaught of HIV/AIDS with Africa as its epicenter has "engendered a profound re-medicalization of African sexualities" (Tamale, 2011: 17).

Western family ideals and Christian morality are however at play, even in medicalized and demographically oriented approaches, as, for instance, in the influential and oft-cited study by demographers John Caldwell et al. (1989), where they come to the conclusion that the basic evil behind HIV/AIDS is female promiscuity, or 'permissiveness,' indicating that "in Africa having sex is as simple and straightforward as eating and drinking" (Caldwell et al., 1989: 195). This statement has been criticized, and rightly so: very often in African societies sexuality is strictly regulated. But since these restrictions are likely to be different from the ones which Caldwell et al. expect to find, they presume restrictions do not exist. What Caldwell et al., most of all, see missing is the ideal of female chastity, a core value of Christianity from Virgin Mary onwards. Virgin Mary even managed to become a mother without having sex; second best is to be a mother having sex only with your husband: sex (for women) as a means to procreation. Female sexual activity beyond marriage and procreation is not supposed to take place. If, nevertheless, it does take place – as happens in Africa, according to the studies on which Caldwell et al. base their argument – this (by the colonial mindset) is seen as the root of evil. In the nineteenth century the evil let loose by sexual promiscuity – as pointed out by Sander Gilman (1989) – was often syphilis, a gruesome disease caused by excessive and immoral sexuality. In the present day, in the context of HIV/AIDS (which is the context in which Caldwell et al. write their paper) a new gruesome disease caused by sexual lasciviousness and lack of sexual morality has taken over.

A further consequence of this view of sexuality – in direct continuation of colonial approaches – is that rather than a source of pleasure, sexuality is seen as connected to risk and danger, closely linked to disease and death. This risk-and-danger approach to sexuality is still often found in development contexts – one needs only to think of development acronyms connected to sexuality, such as GBV (Gender Based Violence), FGM (Female Genital Mutilation), and HTP (Harmful Traditional Practices).

Colonial Constructions of Masculinities and Feminities

"Researching human sexuality without looking at gender is like cooking pepper soup without pepper" (Tamale, 2011: 11). Sexuality and gender are closely interlinked. Thus part of colonial ideas of African sexuality are colonial ideas of masculinity and femininity. Just as colonial ideas of sexual morality were extensions of sexual morality in the colonial mother countries, colonial ideas of masculinity and femininity were based on European middle class ideals of a male breadwinner and a female housewife. The woman as housewife is, however, a fairly recent European creation. In the Middle Ages European women were active in urban crafts, they could be members of guilds, and in agriculture women always had central roles. According to Veronika Bennholdt-Thomsen "Housework as performed by housewives does not represent a set of tasks which women have always carried out because of some natural predisposition, but is the product of a particular history. … [The housewife] was created – by the church, through legislation, medicine and the organization of the workforce (protective legislation, the 'family wage')" (Bennholdt-Thomsen, 1988: 159).

'Housewifization' in Europe was the result of a specific historical process. Nevertheless, this European model was now the model to be imposed on Africa: in colonial schools girls were instructed in cooking and sewing – in order to become appropriate wives for the new class of educated African men, trained as clergymen or clerks. In Mozambique (like elsewhere in Africa) this pattern of male breadwinner/female housewife was continued, but also subverted. Continued in terms of men being sent off to wage work (often on a migratory basis, far from home), subverted in terms of women – officially housewives – in actual fact feeding themselves and their children from their work on the land. The idea of a 'family wage' – wife and children living from the salary of the male breadwinner – was never applied.

In Mozambique for the Portuguese colonial power inducing (or forcing) the men to work was part of the civilizing mission. The image of the lazy African is the point of departure; work as such has a civilizing effect, and the repression of idleness is a goal in itself 'dignifying and uplifting the native through the work' (*Governador Geral de Moçambique*, in *Boletim Geral das Colónias* 1948). This line of argument seems only valid for men, however. As for women, the opposite seems to be the case. For women a heavy workload is an indication of subordination, the civilizing mission being to lift it off her shoulders, in order to enable her to devote herself to housework. "Development goes in the direction of leaving the bulk of agricultural work to the man. Also in Africa the woman shall become the queen of the home: *a Raínha do lar*" (da Silva Rego, 1960: 26). Civilization thus also included the well-known division between public and private spheres, with the husband 'the natural head of the family' as worker and breadwinner, and the wife and mother as 'queen of the home.' This is how things worked – or were supposed to work – at the level of ideology. In practical terms Portuguese colonialism was indeed based on women's productive work in family agriculture, feeding the family while the men were away on forced or contract labor. "Colonial exploitation of Mozambican wage labor was

conditioned by salaries being kept very low, as male workers' families were supported through women's work on the land" (Arnfred, 2011: 123).

An additional problem with the 'queen of the home' ideology is that it naturalized women's procreative capacity, and restricted their role as mothers to the private sphere. In this way, the carpet was pulled from under the feet of women, particularly in matrilineal northern Mozambique, who took pride in their role as mothers, not in the private context of nuclear male-headed families, but in powerful positions as lineage elders. But power based on maternity and fertility is not recognized in civilized Christian contexts, where fertility is trivialized as a function of nature, and maternity is reduced to the education of children in the seclusion of the patriarchal family.

Yet another aspect of women's lives, also not accepted in its own right in Christian/colonial thinking, is sexuality. The general attitude of the Portuguese colonial power was a strong condemnation of female initiation rituals, because of their focus on the education of female sexuality; it should be noted here that female initiation rituals mark the transition from young girl to adult woman, and that female genital cutting was never part of initiation rituals in Mozambique. Nevertheless, the rituals of female initiation were considered by the Portuguese colonialists as well as by the Catholic church as "immoral and offensive to the human nature" (Medeiros, 1995: 5), and also the Protestant mission saw customs relating to female sexuality as so 'vile and immoral' that they could only be described in Latin as a kind of medical journal for only doctors and professionals to read (Junod, 1912/1974: 176).

African Sexuality?

The term 'African sexuality' has heavy colonial overtones; it rings of stereotypes and prejudice, often presented as science, like in the case of Caldwell et al. discussed above. 'African sexuality' is a broad generalization with negative connotations. Nevertheless, postcolonial feminist scholars have started to take it on board, aiming to turn its negative connotations into something else: Sylvia Tamale's (2011) volume of writings on sexualities in Africa, almost exclusively by Africa scholars (and poets) is titled: 'African Sexualities.' In this book Tamale sets up a program for "rewriting and rerighting African Sexualities" (Tamale, 2011: 22), and the book certainly is a rewriting: dozens of articles rooted in experience of African men and women, and analytical concepts tuned to African social realities – much unlike the colonial approaches outlined above. Importantly, in this context, from these points of view (experiences of Africa men and women) sexuality is no more reduced to risk and danger: pleasure is part of the picture, as well as identity and power. And sexuality is not reduced to heterosexuality.

But what about the generalization? Africa is a huge continent – is it possible to talk of 'African sexualities' even in a positive sense? The position of Marc Epprecht is that yes, "the idea of African sexuality contain important grains of truth" (Epprecht, 2013: 65). In spite of early accounts of 'African sexuality' by colonial travelers and traders being "frequently exaggerated, sensationalized, judgmental and imbued by error … they nonetheless documented a basic truth: African societies tended to place an extremely high value on heterosexual marriage and reproduction" (Epprecht, 2013: 57). Epprecht however also points out that far from being non-existent or 'un-African' as claimed by certain African presidents, same-sex practices in a variety of forms always have been part of social and sexual lives in Africa, and thus should be embraced by any notion of 'African sexualities.'

Section II: Postcolonial Critique and Alternative Analytical Approaches

In her introduction to the 2011 volume on 'Africa Sexualities,' Sylvia Tamale picks up from Gayatri Spivak the concept of 'unlearning.' For many of us, she says, "our learning processes grossly neglect to instruct us in the important concept of unlearning. Without this skill, it is extremely difficult for us to think critically and to question unjust frameworks or challenge the established order. … Unlearning literally requires us to discard our old eyes and acquire a new set with which to see the world. It requires us to jettison assumptions and prejudices that are so deep-seated and internalized that they have become normal and appear to be natural" (Tamale, 2011: 3).

Step one in the process of unlearning is to identify those deep-seated assumptions and prejudices that have been part of our beliefs for so long that they seem natural and normal. Section I of this article focused on such identification. Step two is then to look at the world anew, from different vantage points, with different tools of analysis, all of this in order to relearn in ways that are better tuned to the phenomena we wish to understand. "We need to pry our minds open to fresh ideas, absorb new knowledge and apply our intellect, knowledge and experience to develop a critical analysis of the issue at hand. Opening our minds means to accept differences, to see the world through the eyes of others, to open our ears to diverse viewpoints and to venture beyond our familiar horizons" (Tamale, 2011: 4).

Sex as Distinct from Morality

When studying sexualities, one important aspect of unlearning is to make a distinction between sex and morality. In a Western context, for millennia under heavy influence from Christian conceptualizations, sex, morality, and marriage have been closely interlinked. Moral sex takes place in marriage, or at least in a stable heterosexual relationship; sexual relations of other kinds are in the danger zone of being deemed immoral, even more so if one of the sexual partners is in a stable, marriage-like relationship with somebody else. Morality and sex are so closely connected that it takes effort to think of sex in different ways – without going to the opposite extreme, as was seen in the analysis of the Caldwell et al.'s paper above, of only seeing 'promiscuity' and unrestricted sex.

Certainly African societies often impose restrictions on sexuality, but in different ways. It also depends on which kinds of sex we are talking about. Heterosexuality takes place in different contexts and settings. According to fieldwork in rural Mozambique, in some respects sexuality may be rather free,

particularly where there is no risk of pregnancy (prepuberty and/or non-penetrative sex); postpuberty, however, due to the risk of pregnancy, sexuality is more restricted. Ideally initiation rituals take place at the time of the first menstruation, and ideally a young woman would go from initiation rituals directly into marriage. In some places in Africa initiation rituals and marriage were conflated, such as in Namibia as described by Heike Becker (Becker, 2004). Marriage thus may be understood as an institution for control of fertility, more than control of sexuality per se. "What mattered in kinship contexts was control of sex for procreation. Sex in itself was much less severely controlled, as long as it did not result in pregnancies. Historical sources from Kwa-Zulu Natal describe how *usukoma* – non-penetrative thigh-sex – made it possible for young men and women to engage in sex before marriage without fearing for the consequences in terms of pregnancies. … In matrilineal northern Mozambique pre-puberty/pre-marriage sexual relations were distinctly encouraged in order that the women should be properly educated for adulthood. The marker of adulthood was the initiation ritual, immediately after which young women were expected to marry" (Arnfred, 2004: 16).

The general point to be made is that control of sexuality is/was guided by pragmatic concerns (procreative sex may result in children, and children must know where they belong) more than by morality. At the same time there are also norms, which ascribe a child born by a married woman to her husband's lineage, even if it is more or less known that the biological father of this particular child is somebody else. Nzegwu points to patterns of what she calls 'transmarital sex' of Igbo wives (Nigeria) confirming that according to Igbo norms "the question of which pater constituted the line of descent was irrelevant, because the collective view was that a husband of a bride is the father of the wife's children, regardless of how they were begotten" (Nzegwu, 2004: 7). In matrilineal contexts, the sexual autonomy of wives is/has been even greater, the children anyhow belonging to their mother's lineage. Matriliny accordingly makes for temporal marriages: if there is no more sexual attraction, the man might as well move on. In matrilineal contexts, marriage is a matter of getting new blood and new children into the matrilineage, it is not a question of founding new families. According to Karla Poewe, based on her research among matrilineal Luapulans in Zambia, "a man's marital role is to sexually satisfy and impregnate the woman and to provide for her during her pregnancy. The man should not in any way be an object of exclusive emotional investment, nor the focus of attention" (Poewe, 1981: 67). "Ties of intimate dependency between spouses are discouraged in many more or less subtle ways. While sexual enjoyment is valued, it is not limited to one specific partner" (Poewe, 1981: 56).

Women as Active Sexual Beings

In Poewe's research women in matrilineal Zambia come forward as active sexual beings. It is the duty of young women through the sexual powers to "engulf male strangers and convert them into kin" (Poewe, 1981: 68). In matrilineal northern Mozambique, Christian Geffray applies a similar analysis to his findings: men are progenitors, women are seducers. Young women are expected to be proficient in erotic arts (they are checked in this discipline during the initiation rituals) and in return for their role as progenitors the young men "at least have the right of expecting their wives to seduce them and to give them sexual pleasure" (Geffray, 1990: 129).

Research on female rituals of initiation in northern Mozambique reveals how sexual proficiency is considered a pride and a privilege by women (Arnfred, 2011). Young women are instructed in the art of lovemaking during the initiation rituals, but like other areas of women's arts, the learning process starts well before puberty. In parts of northern Mozambique, in bygone days, prepuberty sexual relations between girls and boys were accepted and it was even believed that intercourse and caressing was important for making the girl's breasts grow and for menstruation to emerge. Also well before puberty the girls would be instructed regarding how to pull the small labia of the vagina in order to make them longer. This pulling of labia (*ithuna*) was – and still is – considered an important aspect of a young girls' preparation for lovemaking, and for giving and receiving sexual pleasure. Similar findings are reported by Sylvia Tamale from her investigations in Kampala; she found the practice of labia-elongation anything but dead in the capital of Uganda. Tamale's findings also show the practice of elongation to be perceived as enhancing the erotic experience of the man as well as the woman: "This practice, encoded within the *Ssenga* institution, has enhanced sexual pleasure for women and expanded their perceptions of themselves as active sexual beings," Tamale asserts (2005b: 12). *Ssenga* is originally the name for 'paternal aunt,' having turned into an institution. 'The phenomenon of *Ssenga*,' Tamale says, "represents one of the most powerful cultural inscribers on women's bodies among the Baganda" (Tamale, 2005b: 12). *Ssenga's* are tasked with the sexual and cultural education of their nieces (and other young girls); in Tamale's description of *Ssenga's* tutorials you find a whole array of erotic gadgets and capacities: "Needless to say, sexuality featured prominently in *Ssenga's* tutorials, which would focus on erotic skills, sexual paraphernalia and aphrodisiacs in the form of herbal perfumes, sensual oils, sexual beads (*obutiti*) and so on" (Tamale, 2005b: 17).

The image of women as active sexual beings is confirmed by research from other parts of Africa too. In Senegal the women of a particular ethnic group, the Lawbe, are particularly known as specialized in erotic paraphernalia and knowledge, among which scents and perfumes have a special standing. According to Abdoulaye Ly "the eroticism of the Lawbe is most vividly expressed by the consummate art of their women to concoct a variety of scents ranging from the incense to the pearl belts soaked in a cocktail of diverse perfumes" (Ly, 1999: 47). Tshikala Biaya talks of "a double girdle of shimmering, multicolored pearls, *bine bine*, which evoke local forms of erotic play (…) This is the same type of pearls that is worn hidden under the *drianké's* loose-fitting *boubou*, from where it emanates intoxicating perfumes" (Biaya and Rendall, 2000: 716). The Senegalese figure of the *drianké*, according to Biaya, is "a titillating, plump and mature woman expert at *thiuraye*" (Biaya and Rendall, 2000: 715). *Thiuraye* is a term for a series of erotic practices, including "scented undergarments and girdles of fragrant pearls"

(Biaya and Rendall, 2000: 715). Francis Nyamnjoh describes *driankés* as "reputedly good at business, well-to-do and generally in a position to flatter younger men with gifts and money. They are also experts at the art of seduction, with an impressive mastery of the traditional kit for seduction and eroticism in Senegal" (Nyamnjoh, 2005: 300). The *diskette*, on the other hand, is more 'Barbie-like' (Nyamnjoh, 2005: 301). Biaya describes the *diskette* as "a young Senegalese woman with the slender body of a fashion model, who frequents the nightlife of urban discotheques and bars. Less expert than the mature, full-figured *drianké*, the *diskette* nonetheless carries a double erotic charge: a body type with global erotic purchase, stamped with the *thiuraye* seal of erotic sophistication and craft" (Biaya and Rendall, 2000: 715).

Sex an Asset, Not a Liability – From Women's Points of View

Generally in Western contexts, women's sexuality is perceived as a source of vulnerability and victimization. Initial studies of sexuality in Africa proceeded along similar lines, but post-colonial feminist scholars suggest different approaches. Research in northern Mozambique likewise confirms that sexuality is often seen as a strength and an asset for women, a source of identity and power; sex is a highly valued female capacity. Women, particularly older women, are masters of sexuality. Rituals of female initiation are occasions for the older women to pass on these capacities to younger generations. Very unlike Western contexts where sexual encounters often take place on the terms of men, fieldwork findings from northern Mozambique indicate that here men are expecting women to take the lead (Arnfred, 2011).

Nkiru Nzegwu frames the difference in this respect between Africa and the West as a matter of two different logics: "[The] sexualized gender hierarchy of the West eroticizes male dominance and female subjugation as sexual," she says, "making eroticism and sexuality a basis for women's moral anguish, conflict and downfall" (Nzegwu, 2011: 255). In African cultures, in contrast, "sexual pleasure and fulfillment were equally expected for both women and men, and sensuality was neither pornographic nor the basis of women's subjugation and domination, as was the case in Europe" (Nzegwu, 2011: 254). 'African sexual ontology' Nzegwu asserts, "is not Western sexual ontology in a black face; … it has its own defined history; … it is in fact quite progressive and a resilient stronghold of female power" (Nzegwu, 2011: 254).

Also Bibi Bakare-Yusuf sees sexuality as a potential stronghold of female power, in Africa and elsewhere. A potential which has, in her reading, been undermined by Christianity. In the Bible Eve represents female desire, however with the well known consequence of being expelled from the Garden of Eden. There is a tension at the heart of Christianity, Bakare-Yusuf says, in as far as "desire manifests itself at the outset and is immediately cast out. At the moment of desire's coming to presence, it leads to expulsion. Fertility must be excised from any relation to an originary female desire or action" (Bakare-Yusuf, 2013: 34). There is a point to keeping female desire subdued and invisible; the point is patriarchal power. "Positioning women as weak or damaged subjects gives renewed legitimacy to patriarchally motivated discourses of control and protection" (Bakare-Yusuf, 2013: 30). This is what took place under colonialism, with the construction of women as victims to be rescued by valiant colonial officers. "White men saving brown women from brown men" as Gayatri Spivak has put it (Spivak, 1988: 296).

Bakare-Yusuf sees the influence of Christianity in the present subordination of women. And furthermore, in her analysis, by quenching female sexual desire, and by defining active female sexuality as immoral, a core of women's potential for resistance has also been taken away. "For societies that have been subsumed under the panoptical gaze of Christian colonialism," she says, "occluding female originary desire … helps to reduce the opportunities for resistance. In other words, if female sexual pleasure and desire is highlighted, the supremacy of male desire, power and control is called into question" (Bakare-Yusuf, 2013: 35). In this optic, maintaining a focus on female sexual pleasure and desire has important political implications in terms of resistance to patriarchal power. The suggested strategy is counter-narratives to the dominant narratives of sexuality as risk-and-danger approach, the "hegemonic discourse of sexual terrorism," as Bakare-Yusuf calls it (Bakare-Yusuf, 2013: 30). Telling stories of female sexual power is what is needed, acknowledging female sexual pleasure and desire.

The Power of Pleasure

For a long time a focus on pleasure was totally absent from any account of sexuality in Africa. Sexuality in Africa was constructed as a troubled area, a problem for everybody: colonialists, African presidents, development agencies, even if they might construe the problem in different ways. Presumably, among African men and women, there has always been a focus on pleasure – this is what sexuality to a large extent is about, in Africa and elsewhere – but this was not highlighted neither in research, nor in politics, or in initiatives of development. This situation is now changing – at least in the context of research. The quoted article by Bakare-Yusuf is published in a 2013 volume titled: 'The Political Power of Pleasure,' with articles from many parts of the previously colonized world (not just Africa). Sylvia Tamale's (2011) volume on 'African Sexualities,' grand in content as well as in size (656 pp.), is another contribution to a different view on sexualities in Africa. And articles are emerging from young scholars: Rachel Spronk (2012, 2014) and Christian Groes-Green (2012, 2013) among them. The focus on pleasure throws new light on heterosexual practices in ways that highlight male as well as female sexual enjoyment.

From her interviews with young urban professionals in Nairobi, Kenya, Spronk reports that men as well as women see fulfilling sexual lives as important aspects of their identities as professionals 'being African the modern way.' Regarding the young women of her material Spronk says: "Enjoying sex is considered to be part of healthy self-awareness and they desire fulfilling relationships in which mutual sexual pleasure is the norm. Sexual pleasure, then, has become a positive self-identifier for being a contemporary independent woman" (Spronk, 2014: 14). And regarding the men Spronk says: "For

the men in my study, besides the need to work and/or study hard and have a career to become socially successful, sex was also constitutive of their sense of manhood. … [They] claimed that they felt obliged to be knowledgeable about women's sexual pleasure … [and] they took pride in having sexual skills which they had learnt from watching porn or having read booklets or internet sources" (Spronk, 2014: 15). For men and women together Spronck concludes that "in their definition of sexuality female sexuality becomes redefined as pleasure and not exclusively as procreation, and male sexuality becomes partly redefined in relation to female pleasure" (Spronk, 2014: 17). These are important findings, signaling remarkable changes – at least in certain layers of society – compared to what was supposed to be the case in previous generations.

Christian Groes-Green reports similar findings from his work among young men and women in Maputo, Mozambique. Groes-Green's young people are less well-off and less established than Spronk's young professionals. The women are typically *curtidoras*, young women, who cultivate relationships with often older, well-off men, *patrocinadores* in Maputo slang, in order to gain a living and/or to pay for their studies; at the same time they also have a boyfriend, a *namorado* or a *bom pico*, with whom they have pleasurable sex. The *bom pico* is a good lover, a young man who sexually is tuned to women's pleasure. 'Sex is all about the woman,' as one young man says. "My pleasure is her pleasure" (Groes-Green, 2012: 100). 'The *bom pico* notion,' Groes-Green sums up, "could be seen as an expression of an emerging male ideal which provides young men with an alternative to more dominant and misogynist versions of manhood" (Groes-Green, 2012: 105). In Groes-Green's analysis, however, the *bom pico* notion is also linked to marginality: young men who in order to compensate for lack of financial resources, cultivate their capacities as good and caring lovers. With the *bom pico* notion and the concept of philogynous masculinties, Groes-Green wants to "open a theoretical landscape that is able to embrace contextual variations of men's and women's own enactments and notions of gender, power and equality, rather than a priori assuming universal gender inequalities across diverse cultural settings" (Groes-Green, 2012: 107).

In another paper Groes-Green elaborates on the concept of female sexual power, showing how young Maputo *curtidoras* manage their sexual relationships with older men through close relations to female kin, from whom they learn the tricks of the trade regarding not only how to seduce men, but also how to keep them enchanted. "The erotic knowledge that female elders transmit to daughters and nieces is transformed into bodily capacities that women use to seduce *patrocinadores*. … Erotic powers are reproduced within a female space where secret knowledge is passed on from older to younger women" (Groes-Green, 2013: 106). In this article, Groes-Green makes the connection between on the one hand 'traditional' sexual capacities of rural women, and on the other hand 'modern' eroticism and female powers of seduction, a connection which is also pointed out by Sylvia Tamale in her writings on the Baganda institution of *Ssenga* (Tamale, 2005b). The *Ssengas* as sexual advisors to young girls seem to be an institutionalized version of the Mozambican female elders.

Conclusion

Ways of understanding sexuality in Africa today have deep colonial roots. Widespread homophobia – 38 out of 54 countries on the continent criminalize homosexuality – links directly back to colonial constructions of Africa as heterosexual. African populations and African presidents, nowadays, identify with this colonial construction – even if contemporary research, e.g., the work of Tamale (2011), Epprecht (2004, 2008, 2013), and many others, points to very different realities.

Researchers and presidents nevertheless agree that "African societies tend to place an extremely high value on heterosexual marriage and reproduction" (Epprecht, 2013: 57). Studying heterosexuality in Africa, it is however important to realize that sexuality and morality are not necessarily interlinked in the same way as in Western societies. In Africa regulations of sexuality seem to be more pragmatic (and less moralistic), marriage controlling fertility more than controlling sexuality per se.

Postcolonial scholarship is presently in the process of reconceptualizing sexualities in Africa. A new focus on pleasure is gaining ground, along with a broader approach to studies of sexuality, until recently dominated by notions of risk and danger. The risk-and-danger approach is particularly frequent in contexts of development interventions, where sexuality is perceived as linked to violence and mutilation, disease and death. This is not to say that risk and danger are not around, they surely are. The point is, rather, conceptualizations should be more inclusive, in order to capture also other aspects of sexuality. Recent research on sexual mores of young urban men and women show interesting results in terms of caring, attentive sexual relationships with focus on mutual pleasure and enjoyment being perceived as part of 'being African the modern way' for men as well as for women.

Bibliography

Arnfred, S., 2004. Re-thinking Sexualities in Africa. The Nordic Africa Institute, Uppsala.

Arnfred, S., 2011. Sexuality and Gender Politics in Mozambique. James Currey, Woodbridge.

Bakare-Yusuf, B., 2013. Thinking with pleasure: danger, sexuality and agency. In: Jolly, S., Cornwall, A., Hawkins, K. (Eds.), Women, Sexuality and the Political Power of Pleasure. Zed Books, London and New York.

Bennholdt-Thomsen, V., 1988. Why do housewives continue to be created in the third world too? In: Mies, M., Bennholdt-Thomsen, V., von Verlhof, C. (Eds.), Women: The Last Colony. Zed Books, London.

Becker, Heike, 2004. Efundula: women's initiation, gender and sexual identities in colonial and post-colonial Northern Namibia. In: Arnfred, S. (Ed.), Re-thinking Sexualities in Africa. The Nordic Africa Institute, Uppsala.

Biaya, Tshikala K., Rendall, Steven, 2000. "Crushing the pistachio": eroticism in Senegal and the art of Ousmane Ndiaye Dago. Public Culture 12 (3).

Caldwell, J., Caldwell, P., Quiggin, P., 1989. The special context of AIDS in Sub-Saharan Africa. Population and Development Review 15 (2), 185–234.

Epprecht, M., 1998. The 'unsaying' of indigenous homosexualities in Zimbabwe: mapping a blindspot in African masculinity. Journal of Southern African Studies 24 (4), 631–651.

Epprecht, M., 2004. Hungochani. The History of a Dissident Sexuality in Southern Africa. McGill-Queens University Press, Montreal and Kingston.

Epprecht, M., 2008. Heterosexual Africa? Ohio University Press, Athens, OH. University of KwaZulu-Natal Press, Scottsville, South Africa.

Epprecht, M., 2013. The making of 'African sexuality': early sources, current debates. In: Nyeck, S.N., Epprecht, M. (Eds.), Sexual Diversity in Africa. McGill-Queens University Press, Montreal and Kingston.

Geffray, C., 1990. Ni père, nu mère. Critique de la parenté: le cas Makhuwa. Seuil, Paris.

Gilman, S., 1989. Sexuality: An Illustrated History. John Wiley and Sons, New York.

Groes-Green, C., 2012. Philogynous masculinities: contextualizing alternative manhood in Mozambique. Men and Masculinities 15 (2), 91–111.

Groes-Green, C., 2013. 'To put men in a bottle'. Eroticism, kinship, female power and transactional sex in Maputo, Mozambique. American Ethnologist 40 (1), 102–117.

Junod, H.-A., 1912/1974. Usos e Costumes dos Bantos. Imprensa Nacional, Lourenço Marques.

Ly, Abdoulaye, 1999. Brief notes on eroticism among the Lawbe, Senegal. CODESRIA Bulletin 3&4. CODESRIA, Dakar.

Medeiros, E., 1995. Os Senhores da Floresta. Universidade de Coimbra, Coimbra.

Murray, S.O., Roscoe, W. (Eds.), 1998. Boy-Wives and Female Husbands. Studies of African Homosexualities. St Martins Press, New York.

Nelson, N., 1987. 'Selling her kiosk': Kikuyu notions of sexuality and sex for sale in Mathare Valley, Kenya. In: Kaplan, P. (Ed.), The Cultural Construction of Sexuality. Routledge, London and New York.

Nyamnjoh, Francis, 2005. Fishing in troubled waters: disquettes and thiofs in Dakar. Africa 75 (3).

Nzegwu, N., 2004. The epistemological challenge of motherhood to patriliny. Jenda, A Journal of Culture and African Women's Studies 5.

Nzegwu, N., 2006. Family Matters. State University of New York Press, New York.

Nzegwu, N., 2011. 'Osunality' (or African eroticism). In: Tamale, S. (Ed.), African Sexualities. Pambazuka Press, Cape Town; Dakar; Nairobi; and Oxford.

Poewe, Karla, 1981. Matrilineal Ideology. Male-Female Dynamics in Luapula, Zambia. Academic Press, Maryland.

da Silva Rego, A., 1960. Alguns problemas socio-missionários da Africa negra. Junta da Investigações da Ultramar, Lisboa.

Spivak, G., 1988. Can the subaltern speak? In: Nelson, C., Grossberg, L. (Eds.), Marxism and the Interpretation of Culture. University of Illinois Press, Champaign.

Spronk, R., 2012. Ambiguous Pleasures. Sexuality & Middle Class Self-perceptions in Nairobi. Berghan Books, New York and Oxford.

Spronk, R., 2014. Sexuality and subjectivity; erotic practices and the question of bodily sensations. Social Anthropology 22 (1), 3–21.

Sudarkasa, N., 1996. The Strength of Our Mothers. Africa World Press, Asmara, Eritrea.

Tamale, S., 2005a. Gendered Bodies, Sexuality and Negotiating Power in Uganda, Paper Presented at Conference Organized by the Centre for Global Gender Studies. University of Gothenburg, Sweden.

Tamale, S., 2005b. Eroticism, sexuality and women's secrets among the Baganda. Feminist Africa 5, 9–36.

Tamale, S., 2011. African Sexualities. Pambazuka Press, Cape Town; Dakar; Nairobi; and Oxford.

Heterosexuality in China

Emil Man-Lun Ng, Academy of Family Therapy, Hong Kong SAR, China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

China came from the fertility cult that took heterosexuality and reproduction as its primary duty. Being a large country consisting of numerous tribes with different sexual practices, within the limits of patriarchy, it has been sexually open and accommodating to variant sexual life styles for most periods of its history. Modernization and an open-door policy since the last century have posed many sexual problems that China has yet to solve, overcome, or adjust to.

The Chinese as a Valuable Resource for Sexuality Studies

Modern sexology originates from the West, but Chinese sexuality, with its long reliable historical record and diversified practices covering up to one-quarter of the world population, has long been a valuable source for international reference. The recent rapid socioeconomic changes in China also have raised many new sexual challenges to the country. It would be of global interest to know what these challenges are and how they are being coped with. This article will summarize the key issues in Chinese sexuality since early times, widening to the more horizontal social perspective coming to the recent period.

The Legendary Heterosexual Gods and Universe

Similar to many ancient cultures, the Chinese primitive culture is best described as a fertility cult that practiced a system of nature worship involving rites and ceremonies believed to ensure productiveness and fertility. According to Chinese legend, man was created by the Goddess Nuwa who also invented the marital system when she took her brother Fu-Xi as husband. The most ancient Chinese philosophy, the Yin-Yang theory holds that the cosmos is composed of the female (Yin) and male (Yang) elements, and their interaction is the cause of all things, matters or nonmatters.

With their Gods and universe all heterosexual, the Chinese has never stopped celebrating sexuality as the Tao (way) of nature, to be followed, imitated, and nourished. The ancient Chinese medical text *Su Nu Jing* (*Book of the Plain Woman*), which written around AD 200–500 or earlier, gave a comprehensive account of the medical observation and study of sexual intercourse and, based on this account, made intelligent recommendations on how healthy and pleasurable sexual intercourse can be achieved and sexual problems or dysfunction remedied. Subsequent important texts on the same topic have continued in Chinese medicine, improving the findings and recommendations, and signifying the importance of this aspect of life in the Chinese mind (Ng and Lau, 1990).

Consolidation of Patriarchal Heterosexuality (AD *c.*1000)

As the largest of the Chinese tribes, the Hans, became increasingly more 'civilized' and patriarchal, heterosexual concepts and practices leaned toward the male side. The patriarchal family and polygyny marital system already had set in when reliable historical records of Chinese history began from around 2000 BC, but the negation of female sexuality at the expense of male superiority and dominance only reached its peak after the Tang dynasty (AD *c.*1000) reflected by the prosperity of bound feet and female prostitution. The national obsession for the fetishistic bound-feet damaged the health of Chinese women from their toddler days for more than a 1000 years. Brothels were not only legal but also taken as elite cultural circles for scholars and the higher social class.

The male-dominated heterosexuality, however, did little to affect other sexual lifestyles in ancient Chinese, as long they were by tradition or voluntary choice. Homosexuality, transgenderism, pedophilia, incest, bestiality, and many other types of sexual interests that modern sexologists name as paraphilia or sexual variations were accepted or tolerated within a wide range. So, too, were the different types of marital systems or practices that we now call marriage by free choice, arranged marriage, polygamy, polyandry, group marriage, temporary marriage, trial marriage, 'running marriage', sex partnership, pre- or extramarital sex, husband–wife swapping, and more. This acceptance owes to the fact that the China is ethnologically a hotchpotch of more than 50 tribes with quite independent cultural, anthropological, and historical origins. Mutual acceptance, tolerance, or respect of different sexual lifestyles was the only way to keep the country united and harmonious (Wu and Ng, 2000).

Sexual Restructuring in the Twentieth Century

The situation underwent a big change by the end of the nineteenth century (Deng and Cheng, 2004). In its first open-door policy started in 1899, China had to take up a lot of the Western culture to survive and maintain its global competitiveness. Western sexual influences also were strengthened by the Western powers, which, in their colonization of Asia, also imposed much of their culture and lifestyle on the people they governed. When Dr Sun Yat Sin, a Christian and Western-trained medical doctor overthrew the last Chinese Empire in 1911, he led Chinese sexuality toward a nearly completely Western system: Monogamy was encouraged, bound feet and prostitution were condemned, gender equality was promoted, and homosexuality and many other types of sexual variations were considered problematic. After the Communist Party took over China in 1949 (as communism is basically an ideology of

Western origin), much of the Western conservative sexual mentality remained and was further strengthened. Monogamy was made the only legitimate form of marriage in China and prostitution was illegal until the early 1960s when the Communist Party embarked on a series of campaigns that purportedly eradicated prostitution and hence venereal diseases in the country (Li, 2010). The result was that by 1964, all venereal disease research institutes were closed, and the treatment for the disease was practically not taught in the medical schools. At the same time, the existence of homosexuals in China was officially denied.

The pressure of overpopulation in China prompted the government to apply the one-child policy in 1978 (Zhang, 2007). This was the first time China ran a policy that was incompatible with what is basically a fertility culture.

Re-Enlightenment in the New Millennium

After the Cultural Revolution (1966–76) and with the implementation of the 'second' open-door policy, sexuality in China gradually began to take a more enlightened and practical turn. In 1982, the *Textbook of Sexual Medicine* written by Robert C. Kolodny, William H Masters, and Virginia E Johnson was translated by urologist Wu Jia-Ping. It revived interests in the promotion of sex education sexual health in the public as well as the academic sector. Prostitution and venereal diseases were recognized as serious problems again. Laws against prostitution were refined and implemented mainly in the form of periodic antiprostitution campaigns by the law enforcers.

The first sex education research association in China was established in Shanghai in 1986. Similar associations in other major cities in China soon followed. In 1988, the Ministry of Education, the Ministry of Health, and the National Population and Family Planning Commission jointly announced the launch of sex education in schools. Sex education has since been included officially in middle school curricula. Despite difficulties and variation in different regions, it can be said that nearly all students in China now take at least one sexual physiology class before they finish the third year of junior middle school.

The China Association of Sexology was established in 1994. It oversees and advises on the nation's sex education and sexology studies and runs annual conferences on sexology topics in different cities of China. It also publishes monthly the *Chinese Journal of Human Sexuality*, which is the most authoritative sexology journal in the country.

Parallel with these positive changes in social policy, sex became more easily expressed, discussed, and taken care of openly in public (Dong and Li, 2010). The number of clinics dedicated to the prevention, treatment, counseling, and rehabilitation of sexually transmitted diseases (including AIDS) and sexual dysfunction increased, applying knowledge and skills from both Western and Chinese traditional medicine in different combinations (Yan, 2008).

Sex Shops and Sex Aid Industries

The first sex (aid) shop, with the Christian name 'Adam and Eve', was opened in 1993 in Beijing and, since then, sex shops have opened in nearly every town and city of China. Contrary to those in many other countries, these shops visible publicly, as independent shops or as part of a pharmacy or a department store, marked with a notice 'For age 18 or above'. The sales attendants, who wear white gowns, could be serving as sex counselors at the same time. Some major cities in China also have sex shopping malls. For example, in Guangzhou, an entire building or a floor can be dedicated to 100 or more sex shops. The items for sale are produced mainly by factories in China, which also produce more than half of the adult products sold globally, accounting for 70% of the world market share. A 2008 report estimated that the sale of adult products in China that year had reached 100 billion Yuan and was anticipated to maintain an annual growth rate of 20% (Chinese American News, 2012).

Sex Law Changes

In 1991, the local police in Anhui consulted the national Ministry of Public Security in the handling of a case of lesbian cohabitation. The Ministry's conclusion was that it was not a case for law enforcement because Chinese law does not expressly provide any rules, and it was not a case of criminal hooliganism. As the first official judicial interpretation of homosexuality, this judgment set a historic landmark, a reference point for future police to deal with similar problems, confirming that China always has had a nondiscriminating attitude toward homosexuality. As for homosexuality being a mental illness, the Chinese Classification and Diagnostic Manual for Mental Disorders (CCMD), modeled from the US Diagnostic and Statistical Manual of Mental Disorders (DSM), originally had included homosexuality as a mental illness. In its third edition published in 2001, however, the condition was removed after public consultation and advice from mental health experts.

The first documented sex change surgery was a male-to-female procedure performed in Beijing Medical University Hospital in 1983. The surgery has since been performed successfully in different places of China. One well-known male-to-female is world-renowned modern dancer Jin Xing who underwent the operation in 1996 and performed for the Prime Minister of Britain Tony Blair during his official visit to China in 1998. In 2000, the Ministry of Public Security for the first time officially evaluated the transgender issue and decided that sex or gender selection was a basic right of citizens. Currently, Jin Xing is living in China, married and the mother of three adopted children. It is estimated that more than 1000 transsexuals have undergone sex-change operations in China.

In 2003, a female teacher in Wuhan complained of having been sexually harassed by her supervisor. The case was brought to court. After a nearly 2-year hearing, the court sentenced the supervisor to make an apology to the teacher and awarded compensation of 2000 Yuan for moral damages. This was the first sexual harassment case judged in the country (Yan, 2008).

In 2006, a representative of the People's Congress of China proposed legalization of prostitution to ensure health checks on prostitutes and to prevent the spread of sexually transmitted diseases and AIDS. The proposal was discussed but not passed (Dougherty, 2006).

Sex Cultural Festivals

Another major sign of advance in sexual openness and enlightenment in China is its innovative Sex Cultural Festival. The festival has been held annually in various cities since the start of the millennium. The precursor to the festival was the Sex Aids Trade Fair, which ran annually in Guangzhou for the sex product industries from various parts of China to showcase their latest products, receive orders, and exchange information on recent market trends and opportunities. As the fair grew larger every year and was partly open to the public, the government collaborated with the fair organizer to make it an annual public sex education event consisting of sex education shows, dramas, forums, exhibitions, and street-side sex counseling stalls and any other activities that could bring public attention to important sexuality issues (Ng, 2006). The event was so successful that many other cities began to follow suit and run their own fairs, including Shanghai, Beijing, Foshan, Shenzhen, Zhengzhou, Hong Kong, Macau, Xian, Nanjing, Hangzhou, and many others. The degree of government participation in these festivals varies, but its involvement suggests that China is probably the only country in the world that the government would directly and openly provide priority funding, manpower, and time to support sex education (Ruan, 2008).

Sexual Permissiveness in Youth

Sexual openness and enlightenment brought improvement to some aspects of sex life in China but also presented new problems. As modernization and the development of the Internet facilitated interpersonal and sexual contact, more and more Chinese, especially the young, have begun to advocate sexual permissiveness and sexual freedom. The 1989 nationwide sexuality survey (Liu et al., 1992) reported an average of only about 0.5% college students above age 21 years and about 0.4% of married people between ages 26 and 45 years had five or more sexual partners. In 2004, however, a condom company published an anonymous Internet survey that suggested that the Chinese had an average number of 19.5 sexual partners, higher than the world average of 10.5 and the averages of the Americans and the French. The reliability of this survey might be questioned, but other surveys between 2000 and 2006 also showed a dramatic increase in the numbers of premarital or extramarital sex partners in the subjects studied. For example, in 2000, 32.9% of the subjects said they had had premarital sex, and 16% of subjects said they had more than two sex partners (Pan et al., 2004). The figures increased to 43.5 and 25%, respectively, in 2006 (Jeffreys, 2006).

In the face of the increase in one-night stands, premarital sex, and extramarital sex, some Chinese university students began to move to the opposite direction. In 2005, seven female students from two colleges in Chengdu of Sichuan signed a pact to remain virgins until they married. The pact was posted online and soon was joined by several hundred female college students across the country (Yan, 2008). On 8 April 2008, it was reported online that Zhejiang University had instituted an optional course that aimed to persuade students to preserve their virginity before marriage (Deng, 2008). This trend, however, did not seem to spread. The official policy has remained largely on promoting safe sex, sex and marital education, and sex counseling for youths and rethinking the legal, ethical, and moral implication of sex.

Low Fertility Rate in an Aging Population

Overpopulation has been a main problem of China since the middle of last century (China National Bureau of Statistics, 2010). Its first population boom after 1949 raised its population to a new height of more than 600 million in 1952. The second boom from 1962 to 1973 brought the figure rapidly up to about half a billion. The one-child policy in 1978 slowed down the population growth somewhat, but the total figure still reached 1.3 billion by 2006, more than doubling in less than three decades. The fertility rate, however, already has reached a low value of 1.5, well below the world recognized replacement value of 2.1. This low value, if held unchecked, would mean a rapidly aging population in 20 years, which is unacceptable to the proper functioning of any society. The country therefore has reached a dilemma point as to whether to continue with the one-child policy at the risk of a further aging population or to encourage reproduction for a younger but a continuously expanding population. The official policy, however, thus far has had no plan to loosen the one-child policy at least in the foreseeable 10 years (Zhang, 2007).

Gender Imbalance and Marriage Squeeze

Gender imbalance and marriage squeeze is another serious social complication caused by the one-child-policy in China that has come to the surface gradually at the beginning of this millennium. Because China has a long history of patriarchal tradition, when families were forced to give birth to one child per couple, they wanted their only child to be a male as far as possible. Many families use various means to achieve this, including preconception gender selection, sex-selective abortion, or female infanticide, which are all illegal.

In the 1960s and 1970s, the gender ratio (number of female to 100 male) at birth was already 106, very close to the upper acceptable limit of 107, which the United Nations recommends. Since the mid-1980s, it has been rapidly increasing, with 114.2 in 1992, 116.3 in 1994, and up to 118 in 1999 (Chan et al., 2002). In 2007, a sample survey of the China National Bureau of Statistics reported that the Chinese national average of gender sex ratio at birth was 119.58 and was 122.85 in the rural areas. Hence, China has now the world's highest sex ratio at birth and for the longest duration.

This long duration of gender birth imbalance is precipitating an unprecedented state of 'marriage squeeze' in China. It is estimated that when the imbalance reaches 2020, there will be 30 million more men than women in the marriageable age between 20 and 45 years old. The marriage squeeze theory states that when a gender-imbalanced generation moves into the marital age, the majority gender (the male in this case) will be forced to solve their mate-pairing problem by slightly

postponing their age of marriage or expanding the age difference with their partners. If the gender imbalance lasts too long, however, the sex and age structure will make these simple solutions impossible, inevitably forcing an increasing proportion of the less competitive males to remain single. When this happens, the effect of the squeeze will impinge onto the fertility rate, the confidence and quality of sex, love, marital and family relationships, child rearing, and finally social stability.

Delayed Marriage and Fertility

Statistics are appearing to suggest that the possible negative effects of the marriage squeeze have arrived. China is beginning to see a low fertility rate of below 1.0 in the cities that can no longer be explained by the one-child policy alone. It was 0.61 in Beijing in 2004; 0.86 on average in the cities in 2005; 0.71 in Shanghai in 2010; and in one district of the Jiamusi City (a city in Heilongjiang at the most north-eastern part of China), in 2010, it was only 0.43, the lowest in the world (China National Bureau of Statistics, 2010).

An increasing number of adults in China have chosen or have been forced to remain single, and even if married, to have no children. A survey (National Population and Family Planning Commission Training and Exchange, 2013) conducted from 2010 to 2013 reported that there are in China 249 million unmarried adults over 18 years of age. The data are not sufficient to show how fast this unmarried population is growing, generally, but statistics from individual cities show that growth has been rapid. For example, the 2010 sixth census (China National Bureau of Statistics, 2010) reported that in Shenyang, the unmarried population was about 200 000, three times that of 10 years before; and in Urumqi, the figure was more than 60 million in 2010, which was more than double of that in 2005.

The marriage drop appeared to happen most prominently in the younger age-group. From 2005 to 2010, in the Chinese couples registered for marriage, the proportion of 20- to 24-year-olds has dropped from 47 to 37%, but that of couples over 40 had grown from 3.9 to 12.2% (China National Bureau of Statistics, 2010). Thus, the marriage age of the Chinese people are being pushed back progressively. It was speculated that the causes for this marriage decline, besides the marriages squeeze, have been the high cost for setting up a family, the difficulty in looking for suitable partners, and the loss of confidence in maintaining long-term sex, love, or marital relationships in a busily modernizing society.

Low Sexual Activity and Satisfaction in Couples

Infrequent sexual activities and low sexual satisfaction also may contribute to the low fertility and poor marital quality. By the end of the last century, it was reported (Liu et al., 1992) that the mean frequency of sexual intercourse of married couples was about five times per month, lower than the international average by about half, and only about 50% of were satisfied with their sex life. By early this century, the most extensive report on the frequency of sexual intercourse and extent of sexual satisfaction has been the yearly Global Sex Index Survey performed and publicized by the condom production company Durex (2005–2010). Its reliability and comparability with other sexuality surveys may be questionable, but it is a convenient source of reference. Its coverage and continuity allow for the detection of time changes in different cultures or regions. In its year 2000 report, the sexual intercourse frequency of Chinese adults was 69 per year, only a little more than once per week, below the global average of 96.

The same survey asked about sexual satisfaction for the first time in 2007 when it reported that only 24% of China's sexually active population was satisfied with their sex life. The percentage rose to 42% in 2009, but was still below the global average of 45%, and worse in the modernized cities such as Hong Kong, which was only 29%.

A sexuality survey done by the Renmin University of China in 2002 also gave similar results (Zhang, 2002). In a random sample of 3824 adult males with regular sex partners in different parts of China, their average frequency of sexual intercourse was about once a week, and only 27.1% of them were very satisfied with their sex life. The subjects took life pressure, such as family chores, as the main reason for their low sexual satisfaction. The other reasons raised included poor sex education; relationship problems, such as indifferent attitude with the partner; and extramarital affairs.

Changing Attitude toward Love and Couple Relationships

To look further into the reasons for the poor sexual activity and satisfaction in couples, a nationwide Internet survey was conducted in 2010 on the love and marriage attitudes of Chinese men and women (National Population and Family Planning Training and Exchange Center, 2013). The survey found that the respondents generally believed that intelligence in love ('love quotient') is very important in the process to successful coupling or marriage and that some love education could help them to improve their initiative, optimism, and effectiveness in solving or adjusting to their love problems, through reflection, learning, and growth. A large majority (92%) of the respondents also held an accepting attitude to arranged dating or intimate marriage–couple matching parties.

Television programs and Internet sites were the first to respond to this public need for rapid and arranged dating and match-making. Since the beginning of this decade, more than 400 television match-making programs have appeared across China for single gentlemen (or ladies). These men and women appear before the camera to state their criteria of an ideal love partner and, on the spot, meet a dozen or more ladies (or gentlemen) to make a mutual match and begin dating. These long-series programs have high ratings and serve different age-groups and social classes. On the Internet, match-making companies have good business. Some of them hold yearly large-size on-the-spot meeting and information exchange gatherings in public parks. Each gathering could last for a few days with up to 10 000 participants (Gong, 2013).

The Future of Chinese Sexuality

Chinese sexuality has come a long way from her mainly fertility-based culture, with some periods of regression and conservatism. Coming to this millennium, new and serious problems have arisen because of China's attempt to control population growth and its process to modernize. Many other industrial countries have similar problems, too, but the Chinese, with their pluralistic culture and long history of practical and open-minded philosophy on sex, love, and marriage, seem to be taking innovative steps to overcome or adjust to those problems, attracting a lot of international attention. Much of this attention might have started from a universal concern regarding China's sexual rights, such as poor sexual information provision and censorship, lack of freedom in sexual choices, limited women's rights, child and women trafficking for sexual purposes, sexual violence, and sexually transmitted diseases and AIDS, but this attention gradually has spread to the positive aspects of Chinese sexuality, as reflected in the works of Gulik (1961), Bullough (1976), and many others. Hatfield and Rapson (1993), in a cross-cultural review on passionate love and sexual desire, concluded that the tide of history is in the direction of "an increasing acceptance of passionate love and sexual desire as legitimate, expressible feelings, and a blurring of the differences between the western and eastern sexual cultures".

Ruan (2008), after reviewing the two 'peaks' of sexology development in the last century through the women's liberation movement and the invention of effective contraception, asserted that a third peak will occur this century and be led by China. Undoubtedly, modern sexology started from the West, but there is also an increasingly strong opposition from the Western culture and religion that appears to be getting an upper hand, bringing Western sexological development to a dead-end. Experiences in the last century show that preconditions for successful sexual 'revolutions' include social change and an economic–technological boom, and it is well expected that Asia, especially China and India, will replace the West as they lead the next boom of the twenty-first century (Sheth, 2008). Because sexual revolutions arose along with the social economic booms of the West in the last century, similar revolutions are likely to occur in Asia in this century. With lessons learned from the West, however, combined with the characteristics of Eastern cultural heritage, these revolutions hopefully could be more peaceful, in the mode of 'reforms', in a faster pace, more close to sexological ideals, and with fewer obstacles. Moreover, except for the few regions that had been colonized by the West, Western Christian influences are low in China, especially for fundamentalism. Without these obstacles, and along with social economic changes and a fading of colonial remnants, the original naturalistic Chinese sexual philosophy could return more easily. China has the most numerous and diversified tribes and ethnic groups of the world, practicing very different sexual lifestyles concurrently and harmoniously for thousands of years. Thus, achieving some sexual rights that the West has had the greatest difficulties to attain – such as aspects of sexual freedom, autonomy, and sexual pleasure – would be merely a homecoming for China to its fertility-based culture.

See also: Adolescent Sexual Risk; Family Size Preferences; Fertility Transition: China; Fertility and Culture: Anthropological Insights; Heterosexual Expression at Older Ages; Heterosexuality; Population Change in the Former Communist States of Europe and Asia; Satisfaction and Commitment in Long-Term Heterosexual Relationships; Sexual Behavior and Social Networks; Sexuality Education; Sexuality Over the Life Course; Sexuality: Cultural Aspects.

Bibliography

Bullough, V.L., 1976. Sexual Variance in Society and History (Chapter 11: Sexual Theory and Attitudes in Ancient China). University of Chicago Press, Chicago.

Chan, C.L.W., Yip, P.S.F., Ng, E.H.Y., Ho, P.C., Chan, C.H.Y., Jade, S.K.A., 2002. Gender selection in china: its meanings and implications. Journal of Assisted Reproduction and Genetics 19 (9), 532–549, 11:53.

China National Bureau of Statistics, 2010. Compilation of Sixty Years Statistical Information on China (国家统计局：中国六十年统计资料汇编). China Statistics Publications, Beijing.

China National Bureau of Statistics, 2005. Communiqué on Major Data of 1% National Population Sample Survey. http://www.stats.gov.cn/was40/gjtjj_en_detail.jsp?channelid=4920&record=40 (retrieved 15.04.13.).

Chinese American News, 2012. Sex Aid Market in Mainland China (美中新闻网：大陆性用品市场). http://www.canews.com/News_Center/News_2012/News_release_2012_0408002_03.html (retrieved 15.04.13.).

Zhejiang University offers pre-marital chastity courses, 2008. In: Deng, M.Y. (Ed.), Chinese Sexual Health Times 7 (华人性健康报:浙江大学开设婚前守贞课程) http://blog.sina.com.tw/iacmsp/category.php?categ_page=2&categoryid=176643 (retrieved 15.04.13.).

Deng, T.T., Cheng, F., 2004. General History of Chinese Medicine, the Modern Era (邓铁涛，程之范.中国医学通史-近代卷). People Health Publication, Beijing.

Dong, Y.Z., Li, D.H., 2010. Thirty years of reform and open up: attitudes to marriage, family and sex of Guangdong youths, exploration and analysis of an evolving trajectory. In: Zhang, F. (Ed.), 2010 Sexology Report of Guangdong. (董玉整，黎德化. 改革开放 30 年：广东青年婚姻家庭与性观念发展变化的轨迹探析. 见：张枫编，2010 广东性学报告). Guangzhou Publishing, Guangdong, pp. 141–192.

Dougherty, N., 2006. Prostitution in Contemporary China -The Case of Shanghai Jiading. http://zh.scribd.com/doc/3186890/Prostitution-in-China (retrieved 15.04.13.).

Durex, 2005–2010. Face of Global Sex Report. http://www.durexnetwork.org/en-GB/research/faceofglobalsex/Pages/Home.aspx (retrieved 15.04.13.).

Gong, M.Y., 2013. Guanyin Hill Blind Date Enrollment over 5000 (龚名扬:观音山相亲会报名人数超). http://news.sina.com.tw/article/20130416/9383453.html (retrieved 15.04.13.).

Gulik, R., 1961. Sexual Life in Ancient China: A Preliminary Survey of Chinese Sex and Society from Ca. 1500 BC till 1644 AD. E.J. Brill, Leiden.

Hatfield, E., Rapson, R.L., 1993. Love, Sex, and Intimacy: Their Psychology, Biology, and History. HarperCollins, New York.

Jeffreys, E. (Ed.), 2006. Sex and Sexuality in China. Routledge, London and New York.

Li, J. (Ed.), 2010. Considerations in the Treatment of Syphilis. (李俊：梅毒治疗应注意的事项). http://big5.china.com/gate/big5/health.china.com/html/urinary/treatment/201010/15-83249.html (retrieved 15.04.13.).

Liu, D.L., Ng, M.L., Zhou, L.P., Haeberle, E., 1992. Sexual Behavior in Modern China. Continuum, New York.

National Population and Family Planning Commission Training and Exchange Center, 2013. Pink Paper Report on Marriage Attitude Survey 2010–2013 (国家人口计生委培训交流中心，中国社科院及全国多家主流媒体:中国男女婚恋观调查报告粉皮书). http://ishare.iask.sina.com.cn/f/13442555.html (retrieved 15.04.13.).

Ng, M.L., 2006. The transformation of a sexual culture, the Chinese paradigm. Sexual and Relationship Therapy 21 (2), 137–141.

Ng, M.L., Lau, M.P., 1990. Sexual attitudes in the Chinese. Archives of Sexual Behavior 19 (4), 273–388.

Pan, S.M., Bai, W., Wang, A.L., Lao, M., 2004. Sexual Behavior and Sexual Relation in Contemporary China. (潘绥铭，白维廉，王爱丽，劳曼：当前中国人的性行为与性关系). China Social Sciences Academic Press, Beijing.

Ruan, F.F., 2008. To build and receive the third wave of human sexology. (阮芳赋：缔造和迎接人类性学的第三个高峰). Chinese Sexuality Research 1, 16–18. http://www.wacshome.net/china/chases_1.pdf (retrieved 15.04.13.).

Sheth, J.N., 2008. Chindia Rising: How China and India Will Benefit Your Business. McGraw-Hill, New York.

Wu, J.P., Ng, M.L., 2000. History of sexology in China. In: Ng, M.L., Borras-Valls, J.J., Perez-Conchillo, M., Coleman, E. (Eds.), Sexuality in the New Millennium. Editrice Compositori, Bologna, pp. 11–14.

Yan, P., 2008. Sexual Revolution in China. http://www.china.org.cn/china/reform-opening-up/2008-12/19/content_16979847.htm (retrieved 15.04.13.).

Zhang, W.Q., 2007. Strategy in Stock of Human Capital: Alternatives in Quantity and Quality Population Building. (张维庆：建设人口质量替代数量的人力资本储备战略 中国政府网). http://www.gov.cn/gzdt/2007-07/04/content_672652.htm (retrieved 15.04.13.).

Zhang, A.J., 2002. About the Sexless Marriage Survey. (张爱敬：关于无性婚姻的调查). http://www.people.com.cn/BIG5/shenghuo/79/111/20021031/855279.html (retrieved 15.04.13.).

Heuristics in Social Cognition

Sascha Topolinski, University of Cologne, Cologne, Germany
Fritz Strack, University of Wuerzburg, Wuerzburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Heuristics are strategies of simplifying judgments that allow individuals to make decisions under suboptimal circumstances. The research on heuristics had a profound and lasting impact on modern psychology, particularly on social psychology. Three classical examples are outlined in detail, the availability, the representativeness, and the anchoring heuristic. Also, heuristic judgments using feelings as information as well as processing fluency are discussed.

Heuristics afford decisions under suboptimal conditions. In this article, the psychological mechanisms underlying such heuristics will be identified. After exploring their role in simplifying judgments under uncertainty, special emphasis is given to the availability, the representativeness, and the anchoring heuristic and to their importance in different social contexts. In addition, other judgmental strategies that serve the same purpose are discussed.

'Heuristics' are strategies of simplifying judgments that allow individuals to make decisions under suboptimal circumstances. The discovery of heuristics has had a profound impact on social psychology (*see* Social Psychology), especially in the field of 'social cognition,' which studies the attempts by individuals to make sense of others. This article describes the general idea of heuristic information processing, looks at specific heuristics, and examines the importance of heuristics for social psychology.

Judgments under Suboptimal Conditions

Whereas traditional theories of human decision making (*see* Decision Making, Psychology of) have focused on the normative aspects of valid judgments while neglecting the context in which they occur, the 'heuristics' approach has directed its attention to the psychological processes that enable individuals to make judgments and decisions under situational, motivational, and cognitive conditions that are less than optimal, such as under distraction or time pressure. Because decisions typically are made in a state of uncertainty, it has proven fruitful to identify the strategies that individuals actually use to arrive at solutions that do not represent the best possible outcome but simply meet specific criteria. The study of these simplifying rules of thumb, called 'judgmental heuristics,' has been spearheaded chiefly by Tversky and Kahneman (1974). Because the concept of heuristics is associated closely with these authors, the application of this term is often limited to the specific mechanisms they have identified. However, the underlying logic of *heuristic processing* also applies to other strategies that serve the same purpose, that of simplifying human judgments and making them feasible under suboptimal conditions (Kahneman, 2003).

Because generally it is not possible to identify the use of a heuristic procedure from the outcome of a judgment, researchers have had to create specific conditions under which heuristics lead to errors. As a result, heuristics have often been associated with the supposed irrationality of human reasoning (e.g., Nisbett and Ross, 1980). However, just as visual illusions do not testify to the deficiency of human perception, the errors produced by judgmental heuristics are not proof of the inadequacy of human judgment in general; instead, they point to mechanisms that allow individuals to make acceptable decisions under natural constraints. These mechanisms, however, have no common psychological characteristics that would make them distinct from systematic processing as their status of a heuristic depends solely on the existence of a more systematic way of processing. That means, heuristic processing could be characterized as 'systematic' if there was a way of further simplifying the heuristic procedure. Still, heuristics are important in shedding light on new ways of simplifying judgments and thereby revealing psychological procedures that have been unknown.

The discussion that follows will introduce the three most prominent heuristics.

The Availability Heuristic

In assessing the frequency or probability of an event (or the co-occurrence of several events), individuals often employ a strategy that is based on the ease with which bits of information can be retrieved or generated from memory. Employers wishing to gauge the rate of unemployment in their community may go to the trouble of obtaining the relevant information from official sources. But if they are not motivated or able to do that, they can try to think of unemployed friends or acquaintances. The more easily they are able to do so, the higher will be their estimate of the rate of unemployment. Tversky and Kahneman (1973) called this judgment strategy the 'availability heuristic.'

A variety of judgments in the social domain have been demonstrated to be based on availability, such as self-perception (Schwarz et al., 1991). One well-researched example is judgments of risks, the assessment of which depends on the frequency with which a type of event occurs. For instance, riding a motor bicycle is risky to the extent that accidents occur frequently. However, the actual frequency of an event and the ease with which it comes to mind may be biased

by several factors, such as the frequency with which events are reported in the media. For example, Lichtenstein et al. (1978) found that causes of death frequently reported in the press were greatly overestimated in terms of their frequency.

Importantly, 'availability' has two psychological components that usually are confounded: the content that comes to mind and the ease (or effort) experienced while retrieving the information from memory (Schwarz et al., 1991). Several studies have shown that these two components can be dissociated. In their classic demonstration, Schwarz et al. (1991) asked participants to recall either 6 (which feels easy) or 12 (which feels hard) biographical examples in which they had shown assertive behaviors. Then, participants were asked to rate their own assertiveness. If participants would base their judgment on the retrieved content as informational evidence, they would rate themselves as more assertive in the 12 than in the 6 examples conditions. ('There are a lot of examples of assertive behaviors in my life, so I must be an assertive person.') In contrast, if they would base their judgment on the experienced ease with which they had retrieved memory evidence, they would rate themselves as more assertive in the 6 than in the 12 examples conditions. ('It was easy to recall assertive episodes from my life, so I must be an assertive person.') The result was that participants based their judgments on the experienced ease of information retrieval, not on the number of retrieved examples.

Beyond this basic finding, which maps the default use of the availability heuristic, several further studies have identified factors that moderate which of the two components, experience or content of retrieval, is used, with subjective ease more likely used under suboptimal judgmental conditions such as low motivation or low cognitive capacity (Schwarz, 1998). For instance, individuals use subjective ease more likely when judging out-group members (Rothman and Hardin, 1997), or under positive mood (Ruder and Bless, 2003). Thus, the availability heuristic, such as many other heuristics (see below, for Section Processing Fluency), is no fixed link, but rather a flexible judgmental tool.

Apart from these examples of judgments and affective reactions on the basis of the perceived ease of cognitive operations, the 'availability principle' in its general form, that is, the finding that increased accessibility of any cognitive representation influences judgment and cognitive performance, has stimulated a host of research in social psychology, such as categorization of persons, causal attribution, or the constancy of opinions after they have been discredited. Finally, the more general notion *processing fluency* (e.g., Reber et al., 2004) can be derived from the availability principle and has become a powerful conceptual tool to explain the causal undercurrents of a variety of judgments under uncertainty (see below).

The Representativeness Heuristic

The representativeness heuristic refers to people's tendency to simplify categorical judgments by relying solely or excessively on similarity. For example, a person who wants to determine the profession of another person may use the target person's similarity to the typical member of this profession. Consequently, a student might decide that a fellow student at the next table in a university cafeteria is a business administration major he or she displays the characteristics of a typical MBA student, such as reading the business section of the paper and talking about the stock exchange. Although this strategy has a logical basis, it tends to neglect other types of relevant information, such as base rates. For example, this university may have many more students of law than of business administration. Information about such a frequency distribution should affect the judgment in the same way as information about the target person. However, research by Tversky and Kahneman (1982) has demonstrated that such base rates largely are neglected even if the individuating information is not very diagnostic.

This neglect of base rates is strikingly similar to findings from attribution research. A number of studies discovered that when people identify the causes of an observed behavior, they attribute it to characteristics of the person and neglect situational influences. This response tendency implies that judges underestimate 'consensus information' (Kelley, 1967) that reflects the power of situations. 'Consensus information' describes how other people behave under the same circumstances and serves as a basic determinant of causal attributions. Specifically, if everybody shows the same behavior in a given context, there is little reason to attribute an action to the unique characteristics of the actor. The tendency to neglect such base rate information in social judgments and to give it less weight than it should have under normative considerations can be understood as a manifestation of the representativeness heuristic because it consists of drawing inferences on the basis of similarity ('a person who behaves aggressively is aggressive') at the expense of base rate information (*see* Person Perception, Accuracy of).

The relationship between this attributional bias and heuristic processing was demonstrated subsequently by varying the conduciveness of the judgmental situation (for a review, see Gilbert and Malone, 1995). Researchers demonstrated that the neglect of situational influences depended on judges' cognitive resources, such that this bias was more likely to occur when people were distracted.

The use of the representativeness heuristic may also result in violations of logical principles. In one study by Tversky and Kahneman (1983), judges had to assess the probability of the joint occurrence of two characteristics of a target (a liberated woman). One feature was highly representative ('She is a feminist'), while the other was not ('She is a bank teller'). In line with the representativeness heuristic, judges considered it less likely that the target person was a bank teller than that she was a feminist bank teller, which is, of course, logically impossible.

Other effects of the representativeness heuristic concern the misperception of chance and the neglect of sample size (Tversky and Kahneman, 1974).

Anchoring and Adjustment

'Anchoring and adjustment' describes the phenomenon that judgments are assimilated toward a value that was initially considered. For example, a person who has to estimate the

proportion of African nations in the United Nations may arrive at a higher percentage if he or she has been exposed previously to a high rather than a low standard of comparison. While such an influence would be hardly surprising if this 'anchor' is offered as a piece of information that is relevant to the judgment in question, Tversky and Kahneman (1974) were able to show that the resulting assimilation effect occurred even if relevance was ruled out by presenting the anchor as the outcome of a probabilistic process.

The psychological process that was suggested by Tversky and Kahneman (1974) for this phenomenon is an insufficient adjustment of the final judgment. Alternatively, Strack and Mussweiler (1997) have proposed a mechanism that is related to the availability heuristic, the so-called selective accessibility model. Here, the anchoring effect involves two stages. In the first stage, judges engage in biased hypothesis testing when they consider the anchor as a possible value of the target. In this process, semantic information consistent with the anchor is activated. As a consequence of this selective activation, consistent information will be more accessible in a second stage when information for the final judgment will be retrieved. In this perspective, the anchoring effect is not the result of a numeric influence or a mere effect of insufficient adjustment but caused by a mechanism of semantic priming that has been demonstrated in many studies in social cognition (e.g., Higgins et al., 1977).

Perhaps the most relevant social domain is that of social comparisons (Mussweiler, 2003; *see* Psychology of Social Comparison). Here, work on the anchoring heuristic suggests a mechanism that explains why and under what conditions comparing ourselves to another person will make us appear similar to that person (Mussweiler and Strack, 2000).

Other Heuristics in Social Psychology: Feelings and Fluency

In the previous sections, three basic judgmental strategies were introduced that were identified by Kahneman and Tversky as devices employed to form sufficiently accurate judgments under adverse conditions and with little cognitive effort, although under certain circumstances, these strategies can lead to systematic distortions. The idea of heuristic information processing has been expanded subsequently to other heuristic cues, with lasting influence on theory formation in broad areas of social psychology.

Feelings as Information

A prominent example is the use of *feelings as information* (for a recent review, see Greifeneder et al., 2011). Specifically, it has been argued that under suboptimal circumstances, individuals use their subjective experiences to judge some other criterion or dimension. The most prominent example surely is the classical study by Schwarz and Clore (1983) who showed that judgments of overall life satisfaction are based on current transient feelings, such as good or bad moods due to sunny or rainy weather – which are irrelevant for general life satisfaction from a normative perspective. The impact of feelings on judgments has been demonstrated for both affective (e.g., Schwarz and Clore, 1983; *see* Emotions, Psychological Structure of) and nonaffective feelings (e.g., Strack and Neumann, 2000). For instance, Strack and Neumann (2000) found that furrowing the brow, which evokes the feeling of mental effort, decreases the likelihood that individuals judge some target person as being familiar. Here, the nonaffective feeling of mental effort was used as a heuristic to guide familiarity judgments (cf, Topolinski and Reber, 2010).

Moreover, a next phase of psychological research does not only show that judgments can be based on (irrelevant) feelings in general but also identifies the boundary conditions and moderators of the use of feelings of information (as already addressed in Schwarz and Clore, 1983). Specifically, calling the informational value of current affect into question using a reattribution paradigm – for instance, by telling participants that their current feelings are caused by an external source – substantially reduces or even neutralizes the judgmental impact of feelings because they are reattributed to that external transient source and thus discounted from the judgment at hand (e.g., Schwarz and Clore, 1983; Schwarz et al., 1991). This not only applies to judgments for which the affective cue is irrelevant and thus only a judgmental bias, such as in the studies on mood and life satisfaction (Schwarz and Clore, 1983), but also for judgments for which current affect is a reliable cue to the judgmental criterion. For instance, while current affect – the gut feeling – is indeed a reliable predictor in intuitions on coherence and hidden structure, reattribution procedures decrease intuitive accuracy and thus sabotage intuition (Topolinski and Strack, 2009).

Finally, beyond this instructed invalidation, other psychological moderators of the use of feelings as information exist. As reviewed by Greifeneder et al. (2011), feelings are more likely to be used as judgmental cues when they are salient, representative, and relevant for the judgmental criterion; when the judgment itself is susceptible for affective information; and when level of processing is low (e.g., under low motivation).

Processing Fluency

A common underlying mechanism of many heuristic judgments is processing fluency being the ease and efficiency of information processing (Reber et al., 2004). Specifically, fluency can pertain to any mental operation, such as perceptual, semantic, or motor processing (Reber et al., 2004). As already discussed, the availability heuristic is based on the fluency of memory retrieval. Relatively high fluency triggers a positive experience (e.g., Topolinski et al., 2009), which has been shown to be used as an experiential cue in many heuristic judgments. For instance, fluency drives spontaneous preference (Reber and Schwarz, 1999), the feeling of truth (Unkelbach, 2007), or intuitions in general (Topolinski and Strack, 2009).

Interestingly, several works have shown that the judgmental use of fluency is flexible in the way that fluency-triggered experiences can be relabeled, reattributed, and discounted. For instance, Unkelbach (2007) has shown that although high fluency is generally used as a cue signaling the truth of a statement, only a brief period of relearning this fluency–truth link prompts individuals to use high fluency as a cue to

incorrectness. Also, when the informational value of fluency as a judgmental cue is discredited – such as in the reattribution paradigms already mentioned – individuals can again flexibly discount their fluency experiences from their judgments (Schwarz et al., 1991; Topolinski and Strack, 2009).

Finally, the idea of heuristics has become effective in a second line of research in social psychology. The study of persuasive communication, in particular, has profited from the distinction between heuristic and systematic processing (Chaiken and Trope, 1999). In this domain, the reliance on peripheral cues (such as the expertise of the communicator or the length of the message) instead of on central features (i.e., the strength of the arguments) was shown to occur under suboptimal conditions. Furthermore, the distinction between heuristic and systematic processing has eventually stimulated the development of dual-process models of the human mind (e.g., Strack and Deutsch, 2004), which map the twofold underlying representational modes of human judgment and behavior.

Conclusion and Future Directions

Over the years, the concept of heuristics has generated many insights into the cognitive dynamics of social behavior. At the same time, however, the idea of heuristic processing has gone far beyond the three mechanisms described by Kahneman and Tversky. In a more general perspective, a universal psychological principle has emerged that sheds light on the flexibility of human information processing under divergent epistemic goals (e.g., Kahneman, 2003). However, the work of finding the laws of this flexibility and integrating it into psychological theorizing, as well as isolating the undercurrent causal mechanisms of heuristic judgments, is still an ongoing challenge.

See also: Decision Making, Psychology of; Implicit Social Cognition; Social Categorization; Social Comparison, Psychology of.

Bibliography

Chaiken, S., Trope, Y. (Eds.), 1999. Dual-Process Theories in Social Psychology (Reader). Guilford Press, New York.

Gilbert, D.T., Malone, P.S., 1995. The correspondence bias. Psychological Bulletin 117, 21–38.

Greifeneder, R., Bless, H., Pham, M.T., 2011. When do people rely on affective and cognitive feelings in judgment? A review. Personality and Social Psychology Review 15, 107–141.

Higgins, E.T., Rholes, W.S., Jones, C.R., 1977. Category accessibility and impression formation. Journal of Experimental Social Psychology 13, 141–154.

Kahneman, D., 2003. A perspective on judgment and choice: mapping bounded rationality. American Psychologist 58 (9), 697–720.

Kelley, H.H., 1967. Attribution theory in social psychology. In: Levine, D. (Ed.), Nebraska Symposium on Motivation, vol. 15. University of Nebraska Press, Lincoln, NE, pp. 192–240.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., Combs, B., 1978. Judged frequency of lethal events. Journal of Experimental Psychology: Human Learning and Memory 4, 551–578.

Mussweiler, T., Strack, F., 2000. Consequences of social comparison: selective accessibility, assimilation, and contrast. In: Suls, J.M., Wheeler, L. (Eds.), Handbook of Social Comparison: Theory and Research. Plenum, New York, pp. 253–270.

Mussweiler, T., 2003. Comparison processes in social judgment: mechanisms and consequences. Psychological Review 110, 472–489.

Nisbett, R.E., Ross, L., 1980. Human Inference: Strategies and Shortcomings of Social Judgment. Prentice-Hall, Englewood Cliffs, NJ.

Reber, R., Schwarz, N., 1999. Effects of perceptual fluency on judgements of truth. Consciousness and Cognition 8, 338–342.

Reber, R., Schwarz, N., Winkielman, P., 2004. Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? Personality and Social Psychology Review 8, 364–382.

Rothman, A.J., Hardin, C., 1997. Differential use of the availability heuristic in social judgment. Personality and Social Psychology Bulletin 23, 123–138.

Ruder, M., Bless, H., 2003. Mood and the reliance on the ease of retrieval heuristic. Journal of Personality and Social Psychology 85 (1), 20.

Schwarz, N., Clore, G.L., 1983. Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. Journal of Personality and Social Psychology 45, 513–523.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., Simons, A., 1991. Ease of retrieval as information: another look at the availability heuristic. Journal of Personality and Social Psychology 61, 195–202.

Schwarz, N., 1998. Accessible content and accessibility experiences: the interplay of declarative and experiential information in judgment. Personality and Social Psychology Review 2, 87–99.

Strack, F., Deutsch, R., 2004. Reflective and impulsive determinants of social behavior. Personality and Social Psychology Review 8, 220–247.

Strack, F., Mussweiler, T., 1997. Explaining the enigmatic anchoring effect: mechanisms of selective accessibility. Journal of Personality and Social Psychology 73, 437–446.

Strack, F., Neumann, R., 2000. Furrowing the brow may undermine perceived fame: the role of facial feedback in judgments of celebrity. Personality and Social Psychology Bulletin 26, 762–768.

Topolinski, S., Reber, R., 2010. Gaining insight into the "Aha" – experience. Current Directions in Psychological Science 19 (6), 402–405.

Topolinski, S., Strack, F., 2009. The architecture of intuition: fluency and affect determine intuitive judgments of semantic and visual coherence, and of grammaticality in artificial grammar learning. Journal of Experimental Psychology: General 138, 39–63.

Topolinski, S., Likowski, K.U., Weyers, P., Strack, F., 2009. The face of fluency: semantic coherence automatically elicits a specific pattern of facial muscle reactions. Cognition and Emotion 23 (2), 260–271.

Tversky, A., Kahneman, D., 1973. Availability: a heuristic for judging frequency and probability. Cognitive Psychology 54, 207–232.

Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. Science 185, 1124–1131.

Tversky, A., Kahneman, D., 1982. Judgments of and by representativeness. In: Kahneman, D., Slovic, P., Tversky, A. (Eds.), Judgments under Uncertainty: Heuristics and Biases. Cambridge University Press, New York, pp. 84–98.

Tversky, A., Kahneman, D., 1983. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. Psychological Review 90, 293–315.

Unkelbach, C., 2007. Reversing the truth effect: learning the interpretation of processing fluency in judgments of truth. Journal of Experimental Psychology: Learning, Memory, and Cognition 33 (1), 219.

Heuristics, History of

Ralph Hertwig and Thorsten Pachur, Max Planck Institute for Human Development, Center for Adaptive Rationality, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of heuristics is an interdisciplinary and time-honored enterprise, with heuristics being examined across a wide range of fields, some focusing on professionals' decision-making. In psychology, two influential research traditions have investigated which cognitive heuristics people use across various tasks. One tradition, which originated in the 1970s, has focused on the circumstances under which heuristics cause deviations from classical norms of rationality. The other, starting in the 1990s, has emphasized that heuristics are neither good nor bad per se, but that their success depends on how they are matched to environmental structures. This view suggests that heuristics, rather than leading to irrationality, enable ecological rationality.

What Is a Heuristic?

A heuristic is a simple but useful method for problem solving, decision-making, and discovery. The origin of the term goes back to the Ancient Greek verb *heuriskein*, which means 'to find out' or 'to discover.' Heuristics are sometimes also referred to as 'mental shortcuts' or 'rules of thumb.' One of their key functions is to reduce the complexity of a problem by ignoring part of the available information or searching only a subset of all possible solutions. Traditionally, heuristics have been regarded as necessary and efficient tools, but ones that produce only second-best solutions. However, more recent psychological research has shown that – under conditions that are ubiquitous in the real world, namely, limited knowledge and uncertainty – heuristics can in fact outperform more complex strategies.

Heuristic methods have been considered in various scientific disciplines (though what exactly is meant by a heuristic varies across fields). Two key treatments of heuristic methods can be distinguished. In philosophy, mathematics, operations research, and artificial intelligence (AI), heuristics have primarily been investigated as prescriptive procedures specifying how a reasonable solution can be found given constraints such as computational intractability and limited time. In biology and psychology, by contrast, heuristic principles have also been used as descriptive models, that is, as models that describe how people and other animals sample information from the external and internal (memory) world, and how they make decisions based on that information. The next section gives a historical overview of the discussion of heuristics within these two contexts.

A Short History of Heuristics

Heuristics as Prescriptive Procedures

Heuristic methods were first developed in philosophy and mathematics as a solution to the problems of algorithmic approaches to complex problems. To illustrate the algorithmic tradition, let us take the mechanical device developed by the Catalan philosopher Raimundus Lullus (1232–1315) in the thirteenth century. The device was able to automatically generate all combinations of religious and philosophical attributes that could be used in a debate (Figure 1). It consisted of six concentric discs representing the basic classes of arguments; each of these classes in turn had nine further attributes. Rotating the discs against each other produced every possible combination of attributes, thus automatically generating different arguments and potentially producing new ones. The German philosopher Gottfried Wilhelm Leibniz (1646–1716) had a similar, though even grander goal in the seventeenth century, namely to develop an algorithm for solving any conceivable problem, using a universal language that would allow every possible problem to be represented. It became increasingly clear, however, that such algorithmic approaches could easily lead to combinatorial explosion. Subsequent scholars therefore explored heuristics as means to bring

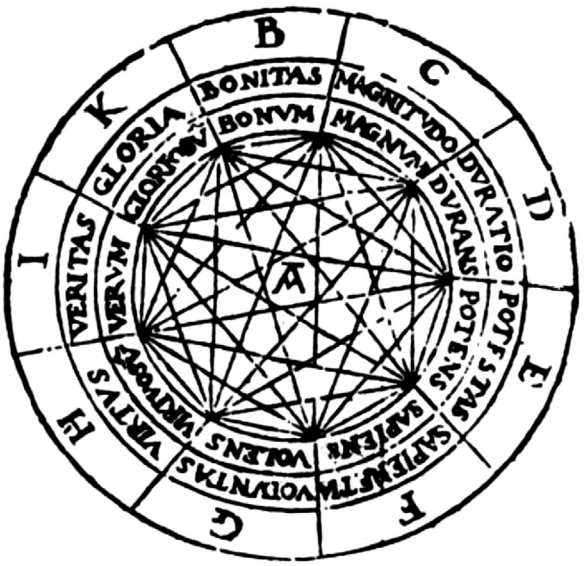

Figure 1 A sketch of one of the discs of Lullus' system for deriving all combinations of arguments. Source: http://www.medienkunstnetz.de/works/ars-magna/.

problem solvers to a solution without the need to explore the space of possibilities exhaustively.

René Descartes (1596–1650), the French polymath of the seventeenth century, had formulated simple rules that were supposed to guide the problem solver toward the relevant aspects, rather than to all parts of a problem. In the nineteenth century, the mathematician and philosopher Bernard Bolzano (1781–1848), who proposed various heuristics for problem solving in his renowned *Theory of Science* (*Wissenschaftslehre*), further developed this approach for epistemic agents (e.g., the method of attempting to find truths by means of something that is not yet known to be true, as opposed to deducing truth from known truths). These scholars' heuristics consisted of rather general procedures that were described in relatively vague terms. Their main goals were to offer a road to knowledge beyond proper deduction, to avoid an arbitrary solution process, and to foster creative thinking. For instance, many heuristics aimed at finding useful problem representations by way of analogy or metaphor.

Heuristic methods for problem solving and discovery received wider recognition in modern mathematics through the work of the mathematician George Pólya (1887–1985). Pólya's (1945) procedures consisted of simple rules, such as dividing the process toward a solution into simple steps by, for instance, finding an analogy to a problem, finding a more specialized problem, or decomposing and recombining the problem. Procedures inspired by Pólya later also informed the new field of AI. Relatedly, heuristics proved to be of great practical value in operations research, an applied field of mathematics. Here, they have been implemented as computer-based tools for planning and aiding decision-making in industry.

Methods for simplifying decision-making have also been discussed in economics. Many decisions are made under uncertainty, with key aspects of the problem remaining unknown. In the absence of complete knowledge of the option space and the probability distributions, normative principles of rational choice are difficult or even impossible to apply. Heuristics for making decisions under uncertainty have therefore been proposed (Savage, 1954). The maximin rule (Coombs et al., 1970), for example, selects from a set of options the one that would yield the most attractive outcome under the worst-case scenario, ignoring both the outcomes to be expected under better conditions and the (unknown) probability distribution. The satisficing principle is a heuristic for sequential decision-making (Simon, 1956). Instead of aspiring to identify the globally best option, this principle evaluates each option according to whether it meets a certain minimum aspiration level, and the first option encountered that satisfies that requirement is chosen.

Another research field of economics in which simple, heuristic decision strategies have received attention is game theory. One of the most frequently studied strategies is tit-for-tat. This strategy, which can be applied in repeated games (e.g., the iterated prisoner's dilemma), determines whether or not a player cooperates with his/her opponent. At the first encounter with the other player, tit-for-tat cooperates; in subsequent encounters, it simply copies the other's behavior (cooperate vs defect) in the previous encounter. In a classical computer simulation tournament (Axelrod, 1984), tit-for-tat proved to be the most successful – though simplest – of the strategies submitted. As well as informing attempts to improve conflict resolution, the study of simple strategies for cooperation within game theory offers insights into how cooperative societies might have evolved in the first place.

The advent of the computer as a computational tool and metaphor of the mind in the 1950s spawned attempts to simulate intelligent behavior in machines. The goal of the new field of AI has been to develop computer programs that can perform tasks such as playing chess, proving logical theorems, or understanding language – and a common challenge is, again, to find ways of limiting boundless and therefore prohibitively expensive search of the problem space. In contrast to previous treatments of heuristics, which often portrayed them in rather vague terms, heuristic rules in AI research have been formulated precisely, often in terms of computational models. A prominent heuristic method for limiting search is means–ends analysis, which was developed in the context of Alan Newell and Herbert Simon's (1972) General Problem Solver system. To move from a current state in the problem space toward a goal state (representing the solution), means–ends analysis reduces the distance between the two by first addressing the most important dimension on which the two differ, followed by the second most important dimension, and so on.

Heuristics as Descriptive Models of the Mind

On the assumption that heuristics, though not perfect, are often effective tools for dealing with a complex and uncertain world, it seems reasonable to suppose that the mind employs them naturally and spontaneously. Indeed, the psychologist Karl Duncker (1903–40) argued that human problem-solving strategies rest on heuristic principles (Duncker, 1935). Similarly, Gestalt psychologists conceived of perception in terms of heuristics. Max Wertheimer (1880–1943), for instance, identified a set of simple principles that organize sensory input to yield object perception – such as the 'laws' of proximity, closure, and similarity (Wertheimer, 1923/1938). As in other contexts, these principles of perceptual organization represent best guesses that usually can be trusted, although they do not work all of the time. They are useful because they are adapted to certain regularities in the environment; for example, elements belonging to the same object are typically in close vicinity to each other (the law or heuristic of proximity; see **Figure 2**).

Presently, cognitive heuristics are most commonly studied in research on how people make decisions. Inspired by Herbert Simon's (1916–2001) seminal work on 'bounded rationality' – the study of how people reason and make decisions with limited computational and informational resources, and when

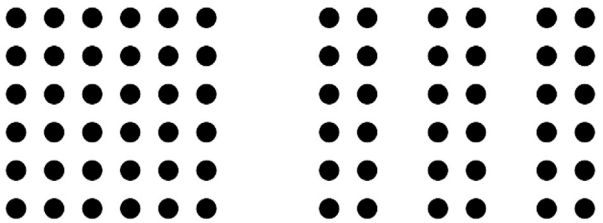

Figure 2 The Gestalt 'heuristic' of proximity: objects positioned close together are perceived as belonging together.

the conditions for rationality postulated by the model of neoclassical economics are not met (Simon, 1956, 1989, 1990) – several research programs have set out to identify and test simple decision strategies. Daniel Kahneman (1934 to present) and Amos Tversky (1937–96) showed that people assessing the frequency or probability of an event often use simple mental shortcuts, such as the availability heuristic or the representativeness heuristic, according to which a judgment is based on ease of retrieval (availability) or on the similarity of an event to a given class of events (representativeness) (Kahneman et al., 1982). What these heuristics have in common is that the cognitive system, when required to make a computationally complex judgment (that necessitates application of the laws of logic or probability theory), substitutes a less complex heuristic attribute, such as similarity (Griffin et al., 2002). Empirical support for the use of these heuristics was provided by creating conditions in which the use of heuristics resulted in systematic errors or lapses of reasoning ('fallacies,' 'cognitive illusions'; Griffin et al., 2002), suggesting human irrationality. From the 1970s, heuristics therefore came to be seen in cognitive and social psychology, and subsequently in behavioral economics, as fallible cognitive shortcuts that people – constrained by cognitive limitations – fall back on in situations in which they would be better off using logic and probability theory. The price of relying on these shortcuts is the risk of committing systematic and severe fallacies of reasoning.

The heuristics examined in Kahneman and Tversky's heuristics-and-biases program were typically vaguely defined verbal descriptions of mental processes. In the research program on the 'adaptive decision-maker,' in contrast, Payne et al. (1993) proposed heuristics for multi-attribute preference tasks that were described in terms of process models. Using computer simulations, Payne et al. tested the accuracy of various preference strategies, both simple and complex, under conditions of time pressure or when the inputs to the strategies, the options' attributes, were highly correlated. The results were then compared with people's actual strategy selection under the same conditions. One key insight was that people are adaptive decision-makers, switching between strategies based on a trade-off between a strategy's prospective accuracy and the time and cognitive effort required to implement it.

From the 1990s, a new research program on cognitive heuristics emerged that challenged both the idea that heuristics are error-prone mental tools (Kahneman et al., 1982) and the notion of a general accuracy–effort trade-off (Payne et al., 1993). Gerd Gigerenzer et al.'s (1999, 2011) research program on 'ecological rationality' proposed that simple heuristics represent prescriptive models of the human mind, thereby connecting the science of heuristics in psychology to earlier research (*see* Heuristics as Prescriptive Procedures). This approach also suggested that heuristics take advantage of environmental structures, and thus picked up on ideas first introduced by Gestalt psychologists (see above).

The new approach to heuristics thus emphasizes that the real world imposes circumstances on the cognitive system that render optimization difficult or impossible (e.g., computational intractability; incomplete knowledge of alternatives, consequences, and probabilities; and changing preference structures). People therefore rely on heuristics not primarily because of their cognitive limitations – although it is indisputable that the human cognitive system is subject to constraints – but because of inescapable properties of real-world problems.

How Well Do Heuristics Work?

The new conception of heuristics in terms of ecologically rational strategies (Gigerenzer et al., 2011) has also challenged the deeply entrenched beliefs that heuristics are doomed to produce second-best results and that optimization is always better. It has been shown that, under some circumstances, heuristics perform as well as or even better than complex strategies involving extensive estimations and computations. For instance, consider heuristics for inferring which of two (or more) alternatives (e.g., cities, companies, sports teams) has a higher value on a quantitative criterion (e.g., homelessness rate, revenue, final score) on the basis of known pieces of information (cues). The normative approach involves weighting and combining all the available cues using complex algorithms such as multiple regression, neural networks, Bayesian networks, or classification and regression trees (see Gigerenzer and Brighton, 2009). Although each of these algorithms processes cues in very different ways, they are all based on the shared premise that all available information should be taken into account.

Models of cognitive heuristics dispense with this premise. They can shortcut the demanding chore of processing and integrating all available information in at least three ways. First, rather than assigning an 'optimal' weight to each cue individually, heuristics can simply tally cues for or against each alternative (tallying heuristic), with each cue receiving the same weight (Dawes, 1979). Second, instead of using all available cues, they can consider just a subset of cues and ignore dependencies between cues. Such 'noncompensatory' heuristics use the first cue that permits a decision to be made and ignore all other cues. This class of heuristics assumes some kind of cue ordering and a stopping rule that specifies when search for information is to be terminated. Cue ordering can, for instance, be based on some measure of cue diagnosticity, such as validity (i.e., the probability that the cue correctly predicts the target variable). The take-the-best heuristic assumes that cues are processed in order of validity, and it compares both alternatives on a single cue, one at a time, until a cue is found that distinguishes between the alternatives (Gigerenzer et al., 2011). Third, some heuristics radically winnow down the number of cues inspected to just one. The recognition heuristic, for instance, only ever considers the 'recognition' cue: when one alternative is recognized and the other is not, it selects the recognized one (Goldstein and Gigerenzer, 2002). The fluency heuristic makes an inference based on the 'fluency' cue: when both alternatives are recognized but one is recognized faster, it selects the one that is recognized faster (Hertwig et al., 2008).

How well do heuristics that do not consider all information perform? And when and why can they perform well? Katsikopoulos et al. (2010) examined the accuracy of the take-the-best, tallying, and minimalist heuristics across real-world data sets drawn from domains such as biology, economics, and sociology. For example, one inference problem required

the strategies to predict which of two cities (e.g., Chicago or Los Angeles) has a higher homelessness rate (criterion value) on the basis of cues such as the cities' average temperature. Two more complex strategies were included as benchmarks: first, linear regression, which weights and sums cues and estimates dependencies between cues; second, naive Bayes, which selects the alternative with the higher probability of having the higher criterion value, given the alternatives' entire cue profile. To compare the accuracy of each of these models, the authors split each data set into two parts and estimated the parameters of each model (e.g., cue validity, cue weight) on one part, the 'training set.' These parameter estimates were used, for each model, to make inferences on the other part, the 'test set.' This is called cross validation, a standard procedure for assessing the predictive performance of models of inductive inference. The size of the training set ranged from just 2–10 objects in the population to 50% of the objects.

Figure 3 plots mean predictive accuracy, defined as a strategy's proportion of correct inferences in the test set, as a function of the training set size. Strategies are differentially influenced by the information that small training sets provide. The accuracy of linear regression and naive Bayes is compromised, presumably because very small training sets provide too few data points to reliably estimate the cue weights. The heuristics, in contrast, are capable of managing with very limited information: Tallying outperforms the other strategies when the training set contains only two objects; the take-the-best heuristic is the most accurate model for training sets of between 3 and 10 objects. The results in Figure 3 appear to be the rule rather than the exception. Simple heuristics such as take-the-best have also been shown to perform well in cross validation when compared with such complex and resource-intensive strategies as linear perception and tree-induction algorithms (Gigerenzer and Brighton, 2009).

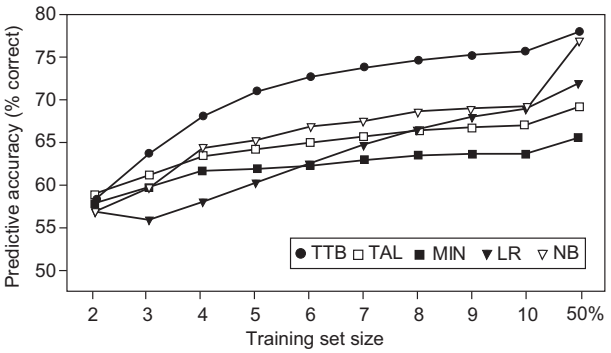

Figure 3 Mean predictive accuracy (averaged across 19 data sets) of three heuristics (take-the-best (TTB), tallying (TAL), and minimalist (MIN)) and two benchmark models (linear regression (LR) and naive Bayes (NB)) as a function of the size of the training set (2–10 objects and 50% of the objects in the population). For training samples as small as 3–10 alternatives, the take-the-best heuristic is more accurate than are the computationally complex models (or the other heuristics; the minimalist heuristic randomly selects one cue until it finds one that discriminates); here, the strategies' best performing variants are plotted. Adapted from Katsikopoulos, K.V., Schooler, L.J., Hertwig, R., 2010. The robust beauty of ordinary information. Psychological Review 117, 1259–1266.

These results suggest that more information, more computation, and more time need not always result in better performance, and that the accuracy–effort trade-off does not generally hold. A better understanding has recently emerged as to why and when heuristics are more (or less) accurate than strategies that require more information and computation.

Why Heuristics Work

There is more than one reason for the surprising success of heuristics. Two of the most important are their ability to exploit sophisticated evolved capacities, and the bias–variance dilemma.

Evolved Capacities

Heuristics can exploit evolved human capacities. For instance, the recognition heuristic and the fluency heuristic (see above for definitions) take advantage of the evolved capacities for recognition memory and systematic forgetting (i.e., pushing aside information that is unlikely to be needed again; Schooler and Hertwig, 2005). The simple tit-for-tat heuristic for making decisions in social exchange situations – cooperate first and then imitate your counterpart's behavior (Axelrod, 1984) – takes advantage of a suite of evolved capacities, including memory (players must remember the exchanged value and the accrued debt), numerical discrimination (to evaluate whether an exchange was equitable), and temporal discounting (as reciprocity involves paying an immediate cost for future benefits, those future benefits must be time-discounted appropriately; Stevens and Hauser, 2004).

Ecological Rationality

According to Simon's adaptive view of 'bounded rationality,' human rational behavior "is shaped by a scissors whose two blades are the structure of task environments and the computational capabilities of the actor" (Simon, 1990: 7). This view suggests that any cognitive strategy, heuristic in nature or otherwise, is not inherently good or bad, successful or unsuccessful (Gigerenzer et al., 1999). Rather Simon's scissors metaphor raises the following question: In which environment will a given strategy succeed, and when and where will it fail? Investigations that aim to uncover the environmental structures that foster or are detrimental to a specific heuristic's success have been referred to as studying 'ecological rationality' (Todd et al., 2012). Ecological rationality brings a prescriptive dimension to the psychological investigation of heuristics. One key insight stems from the bias–variance dilemma (Gigerenzer and Brighton, 2009).

The bias–variance dilemma is a well-known phenomenon in machine learning (Hastie et al., 2001). The term 'bias' does not refer to the kinds of deviations from classical norms of rationality that have been documented in the heuristics-and-biases research tradition (Kahneman, 2011). Rather, it refers to the theoretical bet that all inductive strategies, including heuristics, make about the structure of the environment. Bias is one of three components of the

prediction error produced by any inductive strategy. The total amount of error is the sum of these three components:

$$\text{error} = (\text{bias})^2 + \text{variance} + \text{noise}$$

The bias and variance components of the total prediction error can be understood by assuming an underlying (true) function that a given induction algorithm aims to learn. The algorithm's window on the true function is a (potentially noisy) data sample generated by this function. Averaged across all possible data samples of a given size (see the size of a training set in Figure 3), the algorithm's bias equals the difference between the true function and the mean function inferred by the algorithm on the basis of various data samples. If the mean function hits the mark perfectly (i.e., the true function), bias will be zero. Variance, in contrast, quantifies the algorithm's responsiveness to the specific contents of the samples. It is the sum-squared difference between the mean function and the individual functions extracted from each of the samples.

Ignoring noise, the total error of any induction algorithm depends on these two components, and on how the algorithm trades them off. Even an unbiased algorithm may suffer from substantial errors to the extent that the individual functions, possibly derived from few observations (see Figure 3), involve a great deal of variance. Therefore, seeking low bias will not always be adaptive for an organism. By means of illustration, Figure 4(a) plots the mean daily temperature for London in the year 2000, as well as two polynomial models aiming to capture the true temperature function underlying the data: a relatively simple degree-3 and a more complex degree-12 polynomial model. Figure 4(b) plots the mean error resulting from inducing the true function based on random samples of 30 daily observations (training set) as a function of the degree of the polynomial model. A simple relationship emerged: The higher the degree of the polynomial model, the better the model 'fits,' that is, the better it describes the mean function from the training set. Here, more is better. In predicting the observations not included in the training set, however, more is not always better. Instead, the relationship between the degree of the polynomial model and the models' predictive accuracy is U-shaped, with high-degree (>9) and low-degree (1) polynomials suffering from large errors (due to variance). The best predicting polynomial model has degree 4.

The bias–variance dilemma is one framework that provides insights into when and why simple heuristics are successful in some environments, but not in others (Todd et al., 2012). The notion of ecological rationality and Simon's scissors metaphor emphasizes the importance of the match between the structures of simple heuristics and the structures of the environment. Heuristics can be surprisingly accurate relative to computationally more complex strategies when their structure matches that of the environment, and the bias component of error is thus successfully kept at bay; the bias component will not, however, be within acceptable limits when heuristics and environment mismatch. The variance component is at least as important for the success of heuristics. Variance is likely to be the leading source of error when the available observations about the environment are sparse. Controlling this source of error is therefore crucial for

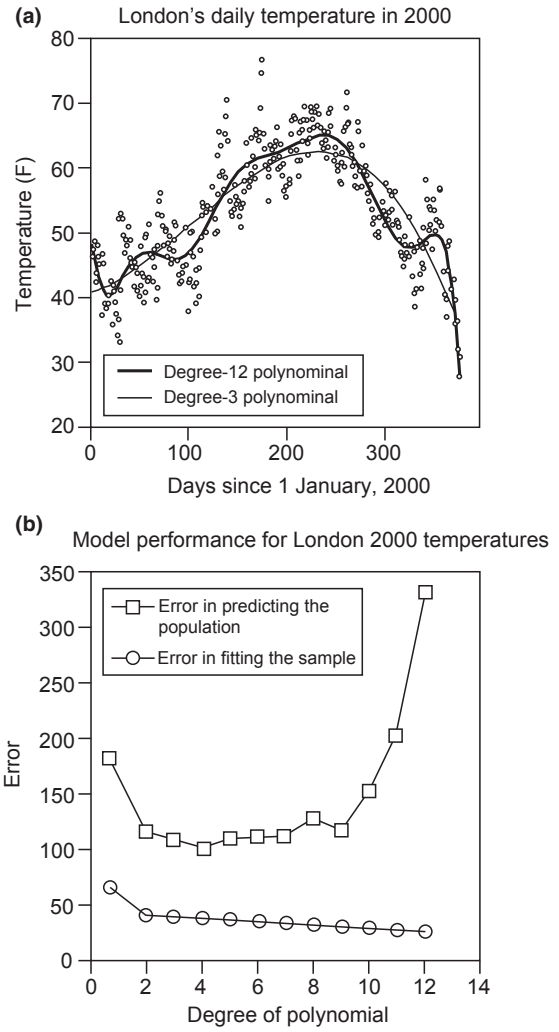

Figure 4 (a) London's mean daily temperature in the year 2000, along with a degree-3 polynomial and a degree-12 polynomial. The two polynomial models were fitted using the least squares method. (b) The mean error in fitting samples of 30 observations and the mean prediction error of the same polynomial models, as a function of degree of polynomial. Adapted from Gigerenzer, G., Brighton, H., 2009. Homo heuristicus: why biased minds make better inferences. Top Cognitive Science 1,107–143.

any inductive algorithm. An algorithm with many adjustable parameters – and thus equipped with great flexibility to accommodate myriad true underlying functions and keep the bias component at bay – is likely to incur substantial prediction error when the window on the environment (the data sample) is fogged up. With sparse observations, flexibility via parameters can become a burden and cause 'overfitting' of the function suggested by the small data samples. In contrast, by forgoing high flexibility and not extracting as much information as is possible from a limited sample of data, simple heuristics control the variance component.

These analyses have suggested a new answer to the question of why humans have learned to use simple heuristics that ignore part of the information and forgo complex processing: In

uncertain worlds that often afford organisms only limited and costly chances to make observations, heuristics appear to achieve robust inferences.

Taxonomies of Cognitive Heuristics

There are various ways to map the mind's inventory of heuristics, the building blocks of which they are composed, and the evolved capacities that they co-opt. One taxonomic dimension is the specific task a heuristic is designed to solve. Across different research traditions in psychology, cognitive heuristics have been proposed for tasks as diverse as probability and frequency judgments in situations of uncertainty (Griffin et al., 2002), preferential choices under risk and uncertainty (Payne et al., 1993), inferential choices and estimates (Gigerenzer et al., 1999), sequential search (Gigerenzer et al., 1999), moral judgments and behavior (Hertwig et al., 2013), and resource allocations in social games (Hertwig et al., 2013), as well as for a wide variety of professional tasks, ranging from predicting criminal offenders' whereabouts to customer loyalty (Gigerenzer et al., 2011).

Another way to carve up the mind's inventory of heuristics is in terms of the domains in which they can operate. Hertwig et al. (2013) proposed that two broad domains be distinguished: 'games against nature' and 'social games.' Games against nature refer to situations in which a person needs to predict or outwit nature (e.g., identifying a good enough location to set up camp or predicting tomorrow's weather). The outcome a person experiences in these games is determined jointly by his/her decision and the true state of nature. In social games, in contrast, how well a person performs depends not on a dispassionate other such as nature, but on the decisions of other players who pursue their own interests. Hertwig et al. (2013) distinguished heuristics according to whether they are bound to games against nature (e.g., the ancient rule of thumb, "Red sky at night, sailors' delight; red sky in the morning, sailors' warning"), to social games (e.g., tit-for-tat), or are able to travel across social and nonsocial worlds (e.g., take-the-best and the N^{-1} heuristic).

Still another taxonomic dimension is the route through which a heuristic becomes part of a person's repertoire of cognitive strategies. Heuristics can be learned individually (through trial and error), socially (e.g., from parents and peers), and through evolution (e.g., possibly sequential search heuristics in the context of mate choice; Gigerenzer et al., 1999). They can also be explicitly designed as cognitive tools to help professionals make better decisions in the real world. For instance, fast-and-frugal trees that follow the template of heuristics such as take-the-best (Gigerenzer et al., 1999) can be constructed to foster accurate medical diagnoses (Jenny et al., 2013). Due to their simplicity, such trees can be easily memorized and understood and could be taught in medical schools.

Further Directions

The study of heuristics is an interdisciplinary and time-honored enterprise, with heuristics being examined in a wide range of domains, including applied fields such as operations research and marketing. In psychology, two influential research traditions have investigated the cognitive heuristics that people use to make inferences, choices, and judgments in social and physical environments. One tradition has focused on the circumstances under which heuristics cause people to deviate from classical norms of rationality (Kahneman et al., 1982). Another tradition has emphasized that heuristics are neither good nor bad per se, nor are they always inferior to more complex strategies. Instead, the performance of a heuristic depends on its match with the structure of the environment ('ecological rationality'; Gigerenzer et al., 1999) and the amount of knowledge a decision-maker has about that environment (bias–variance dilemma). Future research should continue to explore the structures of decision environments and choice architectures that foster the usefulness of heuristics. Relatedly, the psychological issue of how decision-makers choose between different heuristics needs to be better understood. Finally, the insights into the interplay between the design of environments and heuristics should be implemented to help professionals in the real world – who often operate under conditions of time pressure and incomplete information – to make better decisions.

See also: Decision Making, Psychology of; Decision Making: Nonrational Theories; Decision and Choice: Bounded Rationality; Decision and Choice: Heuristics; Heuristics in Social Cognition; Moral Reasoning in Psychology; Risk and Uncertainty in the Social Sciences: Implications for Social Work Theory and Practice.

Bibliography

Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York.
Coombs, C.H., Dawes, R.M., Tversky, A., 1970. Mathematical Psychology: An Elementary Introduction. Prentice-Hall, Englewood Cliffs.
Dawes, R.M., 1979. The robust beauty of improper linear models in decision making. American Psychologist 34, 571–582.
Duncker, K., 1935. Zur Psychologie des produktiven Denkens [The Psychology of Productive Thought]. Springer, Berlin.
Gigerenzer, G., Brighton, H., 2009. Homo heuristicus: why biased minds make better inferences. Top Cognitive Science 1, 107–143.
Gigerenzer, G., Hertwig, R., Pachur, T. (Eds.), 2011. Heuristics: The Foundations of Adaptive Behavior. Oxford University Press, Oxford.
Gigerenzer, G., Todd, P.M., ABC Research Group, 1999. Simple Heuristics That Make Us Smart. Oxford University Press, New York.
Goldstein, D.G., Gigerenzer, G., 2002. Models of ecological rationality: the recognition heuristic. Psychological Review 109, 75–90.
Griffin, D.W., Gilovich, T., Kahneman, D., 2002. Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge University Press, Cambridge.
Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York.
Hertwig, R., Herzog, S.M., Schooler, L.J., Reimer, T., 2008. Fluency heuristic: a model of how the mind exploits a by-product of information retrieval. Journal of Experimental Psychology, Learning 34, 1191–1206.
Hertwig, R., Hoffrage, U., The ABC Research Group, 2013. Simple Heuristics in a Social World. Oxford University Press, New York.
Jenny, M.A., Pachur, T., Williams, S.L., Becker, E., Margraf, J., 2013. Simple rules for detecting depression. Journal of Applied Research in Memory and Cognition 2, 149–157.
Kahneman, D., 2011. Thinking, Fast and Slow. Farrar, Straus & Giroux, New York.
Kahneman, D., Slovic, P., Tversky, A. (Eds.), 1982. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, New York.

Katsikopoulos, K.V., Schooler, L.J., Hertwig, R., 2010. The robust beauty of ordinary information. Psychological Review 117, 1259–1266.
Newell, A., Simon, H.A., 1972. Human Problem Solving. Prentice-Hall, Englewood Cliffs.
Payne, J.W., Bettman, J.R., Johnson, E.J., 1993. The Adaptive Decision Maker. Cambridge University Press, Cambridge.
Pólya, G., 1945. How to Solve It. Princeton University Press, Princeton.
Savage, L.J., 1954. The Foundations of Statistics. John Wiley & Sons, New York.
Schooler, L.J., Hertwig, R., 2005. How forgetting aids heuristic inference. Psychological Review 112, 610–628.
Simon, H.A., 1956. Rational choice and the structure of the environment. Psychological Review 63, 129–138.
Simon, H.A., 1989. The scientist as problem solver. In: Klahr, D., Kotovsky, K. (Eds.), Complex Information Processing: The Impact of Herbert A. Simon. Erlbaum, Hillsdale, pp. 373–398.
Simon, H.A., 1990. Invariants of human behavior. Annual Review of Psychology 41, 1–19.
Stevens, J.R., Hauser, M.D., 2004. Why be nice? Psychological constraints on the evolution of cooperation. Trends in Cognitive Sciences 8, 60–65.
Todd, P.M., Gigerenzer, G., The ABC Research Group, 2012. Ecological Rationality: Intelligence in the World. Oxford University Press, New York.
Wertheimer, M., 1923/1938. Untersuchungen zur Lehre von der Gestalt II [Studies on the theory of Gestalt II]. Psychologische Forschung 4, 301–350.

Hierarchical Models: Random and Fixed Effects

Dennis V Lindley[†], Minehead, Somerset, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 6685–6690, © 2001, Elsevier Ltd.

Abstract

The article describes probability models that are in the form of a hierarchy; the basic quantities in a situation having their probabilities described in terms of parameters, which themselves are described in terms of other parameters, and so on. This is an arrangement which provides great flexibility in probability modeling. Distinction is made between random effects, to which are attached probabilities, and fixed effects, which have probabilities only in the Bayesian paradigm. Examples of hierarchical models are given, together with illustrations of calculations needed for their resolution. Alternative descriptions in terms of equations are described.

The Basic Statistical Model and Its Probability Interpretation

The model that has dominated statistical practice recently supposes that data x have, for each value ψ of a parameter, a known probability distribution dependent on the unknown ψ. The distribution may be described by the probability density

$$p(x|\psi) \qquad [1]$$

where p is a generic symbol for a density of the quantity, here x, before the bar; the symbol after the bar describing the conditions applying to the distribution, here the supposed value ψ of the parameter. Statistical inference then consists in making some statement about the parameter on the basis of the data; such as a point estimate for ψ, or a test that ψ assumes specified values. An important, special case has x, a vector $(x_1, x_2, \ldots, x_n)$, of n components, and ψ chosen such that, given ψ, the components are independent and identically distributed (i.i.d), when eqn [1] assumes the special form

$$\prod_{i=1}^{i=n} p(x_i|\psi) \qquad [2]$$

The Bayesian approach would include, in addition to the density eqn [1], a probability distribution for ψ expressed through a density $p(\psi)$. Statistical inference then consists of calculating $p(\psi|x)$ by Bayes's theorem, inverting the roles of x and ψ in eqn [1].

A distinction is often made between the nature of the probabilities involved in eqn [1], and more particularly eqn [2], and those in $p(\psi)$. In view of the repetitions under similar circumstances of $x_1, x_2, \ldots, x_n$, their probability is associated with frequency and the x_i thought of as successive realizations of a single random variable. Alternatively, $(x_1, x_2, \ldots, x_n)$ is viewed as a random sample from a population. In contrast, ψ is a fixed, but unknown, quantity and, when introduced, $p(\psi)$ expresses beliefs about the values it might assume, without reference to frequency or populations. The disparity between random, for x, and fixed, for ψ, has played an important role in the development of hierarchical models, leading to different inferences depending on whether the effect is supposed fixed or random. In the Bayesian approach there is no distinction and all probabilities express beliefs; the special case eqn [2] reflecting the belief in exchangeability among the observations x_i.

The Hierarchical Model

To illustrate the need for a hierarchical model, consider data consisting of repeated observations divided into groups. Let x_{ij} denote the jth observation in group i ($i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$), with n observations in each of m groups, and suppose that within each group the values are conditionally iid with distributions that may vary from group to group. A popular, special case has

$$x_{ij} \sim N(\theta_i, \sigma^2) \qquad [3]$$

with the x_{ij} being independent given $\psi = (\theta_1, \theta_2, \ldots, \theta_n, \sigma^2)$, so that they are normally distributed with common variance σ^2 but with means that depend on the group. This model is the basis for one-way analysis of variance where inference may well concentrate on whether the means are truly different and, if they are, of what these differences consist. As it stands, this model is just a special case of eqn [1] with x the set of x_{ij} and ψ as above; but now contrast two cases that differ in respect of the nature of the groups. In the first, the study is one investigating the potency of a drug and the groups correspond to dilution levels. In the second, the milk yields of cows are measured and the groups correspond to different sires. In both cases, it may be reasonable to think of the duplicate measurements within a group as i.i.d, forming a random sample from a possible population, in agreement with eqn [3]. But there are differences between the nature of the groups; for in the dilution experiment, the levels are fixed, whereas, with the sires, those selected may be thought of as a sample from a population of sires. In the language of the last paragraph, the θ_i are fixed in the dilution experiment, but random with the cows. In the latter, it might therefore be reasonable to suppose that the means are capable of a frequency description and a probability distribution. A possible, simple case is to suppose that, in addition to eqn [3],

$$\theta_i \sim N(\mu, \tau^2) \qquad [4]$$

[†] Deceased

with the θ_i being independent given μ and τ^2. The frequency interpretation is inappropriate in the dilution experiment, where the levels have been organized, for example at levels 1, 1/2, 1/4 etc. It is also possible that inferences in the two experiments could be directed to different ends. Thus the variability among sires expressed through τ^2 is meaningful, but not in the dilution experiment. There are many other possibilities besides eqn [4] for the second-stage distribution. For example, one group may be a control and the remainder correspond to new treatments. In this case, the distribution for the control mean may differ from that for the others. Outside the Bayesian view, only the latter may have a frequency interpretation. It is this flexibility that helps make hierarchical modeling so attractive.

The probability judgements in eqns [3] and [4] together form a hierarchical model with two stages to the hierarchy, the first expressing the dependence of the data on the group means in eqn [3], the second describing the variability of those means in eqn [4]. The hierarchical model has been thought by some to be appropriate in studies of milk yield, but not in dilution studies, because the means are random in the former but fixed in the latter. The general form of a two-stage, hierarchical model has data x dependent on parameters (θ_1, ϕ_1) expressed through

$$P(x|\theta_1, \phi_1) \qquad [5]$$

at the first stage, with θ_1 having a distribution

$$p(\theta_1|\theta_2, \phi_2) \qquad [6]$$

at the second stage. The parameters of the second stage, θ_2, ϕ_2 are often called hyperparameters. Notice that in this formulation some of the parameters at the first stage are fixed, ϕ_1, whilst others, θ_1, are random and have their probability structure expressed in the second stage in terms of the fixed hyperparameters. In the normal example, eqns [3] and [4], θ_1 here corresponds to the group means there, and ϕ_1 to the common variance. The hyperparameters are $\theta_2 = \mu$, $\phi_2 = \tau^2$, the values σ^2 and τ^2 describing the within- and between-groups variances. There may not be any fixed parameters at the first stage, as when the within-variances are supposed different, at σ_i^2, and given a distribution at the second stage, as with the means. It is possible to extend the number of stages in a hierarchical model, for example by adding a third stage to eqns [5] and [6] by supposing θ_2 depends on (θ_3, ϕ_3) expressed through

$$p(\theta_2|\theta_3, \phi_3) \qquad [7]$$

Such an extension would be appropriate for a two-way analysis of variance; the first stage describing the variation within cells of the two-way layout, the second concerning the row and column means (the main effects), the third involving the interactions between the main effects. Any Bayesian model is, in a sense, hierarchical, the two stages being $p(x|\psi)$ and $p(\psi)$ respectively, but the terminology is usually confined to the models just described. A further extension is to include regression effects. For example, eqn [5] may be extended to

$$p(x|z, \theta_1, \phi_1)$$

where z is an observed regressor variable. Similar regression effects may be introduced at other stages in the hierarchy.

Returning to the two-stage model in eqns [5] and [6], notice that x and θ_1 are random, whereas ϕ_1, θ_2, and ϕ_2 are fixed, and the stages describe the joint distribution of the random quantities, given the fixed ones, θ_2 and ϕ_2 having no effect on x except through θ_1. This lack of effect can be expressed in the form of a directed graph, or influence diagram, which clarifies the hierarchical form. The two-stage model is represented in the figure, the lack of links from x to θ_2 and ϕ_2 expressing the feature. The extension to more stages is easily apparent.

As with the basic model in eqn [1], the Bayesian approach would add a probability distribution $p(\phi_1, \theta_2, \phi_2)$ for the fixed parameters and hyperparameters. A joint distribution for all the quantities, random and fixed, is not thereby specified unless the additional assumption is made that, given (θ_1, ϕ_1), x is independent of (θ_2, ϕ_2). This assumption is usually made, often tacitly. The independence assumption is also reflected in the directed graph by the lack of links already mentioned. There may be additional links unless the quantities $(\theta_1, \theta_2, \phi_2)$ are independent. Since the probability structure is complete, it follows that every quantity has its probability structure defined, under any conditions expressed in terms of the quantities of the problem, which probability can be found following the usual rules of the probability calculus. In particular, the posterior distribution of any of the parameters or hyperparameters, given the data, may be determined.

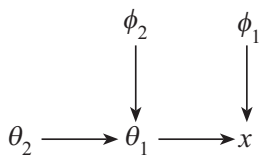

Inference in Hierarchical Models

We now return to the simple case where the observations are divided into m groups, each of size n, and possess the normal, independent distributions described by eqn [3]. Suppose that the common variance within groups, σ^2, is known. The extension to unknown variance presents complications that do not affect the general thrust of what follows. If the means $\{\theta_i\}$ are fixed, then eqn [3] completes the specification of the model and inference about them is ordinarily performed either by least-squares, or by maximum likelihood, the two methods here being equivalent, with the result that θ_i is estimated by the group mean

$$x_{i\cdot} = \sum_{j=1}^{n} x_{ij}/n \qquad [8]$$

with variance σ^2/n. Suppose, in contrast, the means are random and this randomness is expressed by the second-stage assumption eqn [4], where again τ^2, the between-group variance, is supposed known. Applying either of the inferential methods here gives another estimate for θ_i, namely

$$\left(\frac{nx_{i\cdot}}{\sigma^2} + \frac{x_{\cdot\cdot}}{\tau^2}\right) \Big/ \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right) \qquad [9]$$

with variance $(n/\sigma^2 + 1/\tau^2)^{-1}$, in the form of a weighted average of the group mean $x_{i\cdot}$, which was the estimate in the fixed case, and the overall mean $x_{\cdot\cdot} = \Sigma x_{i\cdot}/m$, the weights being inversely proportional to the variances, σ^2/n and τ^2 respectively. It is often said that the effect of randomness in the

group means is to shrink the fixed estimates towards the overall mean. An important feature of the shrinkage is that the apparently best (worst) groups with highest (lowest) values of $x_{i\cdot}$ will have their estimates decreased (increased) and, as a result, the group differences will be diminished. Notice that if τ^2 is large, the shrunk means approach the usual ones. The shrinkage feature is important in comparisons of the θ_i, for example in the construction of league tables (Goldstein and Spiegelhalter, 1996), or in census returns for different communities (Ericksen et al., 1989).

In the Bayesian paradigm all the unknown quantities will have a probability distribution, in particular $(\theta_1, \theta_2, \ldots, \theta_n, \sigma^2)$ in the normal, one-way case in eqn [3], irrespective of whether the means are fixed or random. One possibility is for the θ_i to be provided with the second-stage distribution eqn [4] and finally (μ, σ^2, τ^2) given a distribution. If, as above, the variances are known, only the distribution of μ needs additionally to be specified. If it too is supposed normal and the variance taken to be large in comparison with σ^2 and τ^2, eqn [9] is the mean of θ_i, given the data (and σ^2, τ^2). Thus the Bayesian approach favors the shrunk estimate whenever the second-stage distribution is thought to be of the form in eqn [4], which might be appropriate for sires but not with the dilution experiment. For a Bayesian, there is no real distinction between the fixed and random effects; what differences there are depends on the specific form of the second-stage distribution, and perhaps on the quantities of interest. For example, interest may centre on τ^2, the variance between means, rather than on the individual values θ_i. The situation is in marked contrast to other views where, as we have seen, the estimates of the means depend on whether or not the means are fixed or random; also the sample variance-components in the analysis of variance have different interpretations, in the fixed case being a measure of the differences between the specific means, in the random case referring to the population variance τ^2.

Returning to the general, statistical model in eqn [1] with data x and parameters ψ, it is often hard, especially when the data are numerous, to specify the required density $p(x|\psi)$. In general, whenever there are a number of quantities $u_1, u_2, \ldots, u_s$, it is difficult to describe a joint density $p(u_1, u_2, \ldots, u_s)$; an easier course often being to think about the quantities in some order and structure a series of conditional distributions $p(u_1)$, $p(u_2|u_1)$, $p(u_3|u_1, u_2)$ etc. This is essentially a hierarchical model except that, as explained earlier, it is usual to have the distribution of one quantity dependent only on the immediately previous one and not on all the previous ones; $p(u_3|u_2)$ without reference to u_1. In general, the order in which the quantities are introduced may be arbitrary, but in the hierarchical case this does not arise since the order is naturally dictated by the situation. For these reasons, hierarchical models have proved very popular in the description of complicated cases, enabling them to be broken down into stages, each of which is simple, and yet where they combine together to form a whole that reasonably describes complicated reality.

Types of Hierarchical Models

There are few hierarchical models in which it is possible to obtain explicit, analytical results for the inference desired. An exception is the normal, linear model of which eqns [3] and [4] provide a special case. Generally, at the first stage

$$x \sim N(A_1\theta_1, C_1) \qquad [10]$$

where $x(n \times 1)$ and $\theta_1(s \times 1)$ are vectors, and $A_1(n \times s)$ and $C_1(n \times n)$ are matrices. A_1 is known, θ_1 is unknown. Usually C_1 is known, the simplest case, or known up to an unknown multiplier, though it is possible to describe the distribution of its elements in a second stage. Equation [10] expresses the usual linear dependence of x_1 on θ_1 and is basic to least squares, where θ_1 is estimated by $(A_1^{\mathrm{T}}C_1^{-1}A_1)^{-1}A_1^{\mathrm{T}}C_1^{-1}x$. The second stage imposes a similar structure on θ_1,

$$\theta_1 \sim N(A_2\theta_2, C_2)$$

with obvious extensions to further stages as required. Explicit results are provided by Lindley and Smith (1972) using Bayesian ideas, though many of the results can be interpreted sensibly without invoking a distribution on the final hyperparameters, as was seen in the special case.

Similar ideas had previously been used, starting in the 1960s, by Henderson in genetics and specifically in animal breeding; see Henderson (1984). These were outside the usual statistical paradigms and remained unappreciated in broader contexts. An admirable statistical survey of Henderson's and related work is provided by Robinson (1991).

Another field in which explicit results have been obtained, though usually of an asymptotic form, is that of Empirical Bayes. This is not always considered a hierarchical model but does fit into the general framework above. We illustrate with a simple case again involving several groups and data $\{x_{ij}\}$ but this time with independent, Poisson distributions of means μ_i, the same for all observations in group i. This describes the first stage of the hierarchy. At the second stage, the μ_i are supposed iid with some unknown density $p(\mu)$. The model might be suitable when the groups refer to runs in an industrial process. A distinction from the model above is that hyperparameters are avoided and a nonparametric approach adopted at the second stage; no restrictions, other than of a mild nature like continuity, being placed on $p(\mu)$. The observed distribution of the group means $x_{i\cdot}$ provides information about $p(\mu)$. Despite the nomenclature, the analysis is ordinarily not Bayesian. A good account is provided by Maritz and Lwin (1989). The term has also been used in connection with the equation-based approach mentioned later in Section Equation-Based Approach.

Hierarchical models have also been successfully applied to stochastic processes. These may be spatial but here an exposition in the more usual temporal terms will be used. It is often reasonable to suppose that there is an underlying development process taking place over time, but that one's observation of it is disturbed by error, these two features providing the two stages of the hierarchy. A simple example is to suppose an observation x_t at time t depends on the state θ_t of the process then, with a superimposed error ε_t,

$$x_t = \theta_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, v_t) \qquad [11]$$

whereas the process is linear, autoregressive with

$$\theta_t = a_{t-1}\theta_{t-1} + \eta_t, \quad \eta_t \sim N(0, w_t) \qquad [12]$$

$t = 1, 2, \ldots$. Equation [11] is termed the observational equation, eqn [12] the system equation; the two sets of error terms

being supposed individually independent and independent of each other. The approach is usually Bayesian with $\{v_t\}$, $\{w_t\}$, and $\{a_{t-1}\}$ supposed known, at least in the simpler cases, and an initial distribution supposed for θ_0. If this last is normal, it is straightforward to update by Bayes to obtain successively the distributions of θ_t given $x^{(t)} = (x_1, x_2, \ldots, x_t)$, the observations up to time t. Often the primary purpose of the study is to forecast the future state of the process, θ_{t+1}, given the current data $x^{(t)}$,

$$p\left(\theta_{t+1}|x^{(t)}\right) = \int p(\theta_{t+1}|\theta_t)p\left(\theta_t|x^{(t)}\right)\mathrm{d}\theta_t$$

involving the system equation and the posterior distribution of θ_t. A full account is provided by West and Harrison (1989).

As previously mentioned, hierarchical models are especially useful in describing complicated, but realistic, situations, where the individual stages in the hierarchy are relatively easy to assess. Nevertheless, apart from the cases outlined here, analytical results are difficult to obtain because the integrals that arise in the probability calculations are not expressible in standard forms, so that resort has to be made to numerical methods. In one sense this is no hardship, since the data are already numeric and in any application one only requires answers for that specific data set. On the other hand, the integrals are typically in high dimensions where numerical results are hard to come by. An important, recent advance has been to use Markov-chain, Monte Carlo (MCMC) methods which themselves depend, like the hierarchical models, on conditional distributions. There has therefore been a recent explosion of papers using MCMC to handle these models. A summary is given by Brooks (1998).

Equation-Based Approaches

Hierarchical models are often described in an alternative terminology as multilevel, our stages being replaced by levels. An example is provided by students, at the first level, being in classes at the second, which form schools at the third. It is also common to alter the emphasis and not talk in terms of probability, as here, but in terms of equations which incorporate a probabilistic element, so that eqn [3] would be written

$$x_{ij} = \theta_i + \varepsilon_{ij} \qquad [13]$$

with ε_{ij} being an error term, of zero mean. Complete adherence to eqn [3] would in addition require the error terms to be independent, normal and of constant variance. It is usual to employ matrix notation, as in eqn [10], and a popular model is

$$y = X\beta + Zu + \varepsilon \qquad [14]$$

Here y, replacing x, is the observed data, dependent on observed covariates X and Z; whilst β is a set of unknown, fixed parameters, u a set of random quantities and ε the error term. Thus ε is at the first level and u at the second level. A further complication, that X is only observed with error, is sometimes included, so that

$$X = x + \eta$$

in terms of the unknown, true value x and error η at another level. This generalizes the problem of regression with both quantities subject to error. We shall refer to this whole approach as equation based, rather than probability oriented, though it is possible to pass from one to the other, as between eqns [3] and [13]. These ideas have been used in longitudinal studies where several observations are made on an individual over time, forming one level of variation, another level being that between individuals.

The distinction between the approaches is not confined to the model description but includes the analysis, where manipulations of equations replace the probability calculus. Thus eqn [14] leads to estimates $\widehat{\beta}, \widehat{u}\ \beta, u$ given by

$$X^TX\widehat{\beta} + X^TZ\widehat{u} = X^Ty$$

and

$$Z^TX\widehat{\beta} + \left(Z^TZ + I\right)\widehat{u} = Z^Ty$$

under simple assumptions about u and ε. Some of the methods that are used do not employ the full probability structure but only refer to the first and second moments, a popular variant of least squares being quasi-likelihood. Other methods are based more closely on likelihood and variants thereof and often embrace ideas associated with the general, linear model. Lee and Nelder (1996) provide a good account with a lively discussion in which R.L. Smith suggests disagreement between their results and those produced by the fully Bayesian approach. This is not surprising since, as has been shown in Section Inference in Hierarchical Models, Bayesian procedures employ more shrinkage than some other methods.

An issue of the *Journal of the Royal Statistical Society A* (162, 1999, 269–448) is devoted to articles on "Applications of random effects/multilevel models to categorical data in social science and medicine" and provides many examples of the ideas here briefly described. The book by Goldstein (1995) is relevant; he and his colleagues have provided a computer package (1998) to handle many models. Similar ideas have been put forward by Bryk and Raudenbush (1992). Best et al. (1996) present a Bayesian approach, whilst Spiegelhalter et al. (1998) supply some computer facilities.

Bibliography

Best, N.G., Spiegelhalter, D.J., Thomas, A., Brayne, C.E.G., 1996. Bayesian analysis of realistically complex models. Journal of the Royal Statistical Society A 159, 323–342.

Brooks, S.P., 1998. Markov chain Monte Carlo method and its application. The Statistician 47, 69–100.

Bryk, A.S., Raudenbush, S.W., 1992. Hierarchical Linear Models: Applications and Data Analysis Methods. Sage, Newbury Park, CA.

Ericksen, E.P., Kadane, J.B., Tukey, J.W., 1989. Adjusting the 1980 census of population and housing. Journal of the American Statistical Association 84, 927–944.

Goldstein, H., 1995. Multilevel Statistical Models. Arnold, London.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., Healy, M., 1998. A User's Guide to MLwiN. Institute of Education, London.

Goldstein, H., Spiegelhalter, D.J., 1996. League tables and their limitations: statistical issues in comparisons of institutional performance. Journal of the Royal Statistical Society A 159, 385–443.

Henderson, C.R., 1984. Applications of Linear Models in Animal Breeding. University of Guelph, Ontario, Canada.
Lee, Y., Nelder, J.A., 1996. Hierarchical generalized linear models. Journal of the Royal Statistical Society B 58, 619–678.
Lindley, D.V., Smith, A.F.M., 1972. Bayes estimates for the linear model. Journal of the Royal Statistical Society B 34, 1–41.
Maritz, J.S., Lwin, T., 1989. Empirical Bayes Methods. Chapman and Hall, London.
Robinson, G.K., 1991. That BLUP is a good thing: the estimation of random effects. Statistical Science 6, 15–51.
Spiegelhalter, D.J., Thomas, A., Best, N.G., 1998. WinBUGS User Manual, Version 1.1.1. Medical Research Council Biostatistics Unit, Cambridge, UK.
West, M., Harrison, J., 1989. Bayesian Forecasting and Dynamic Models. Springer-Verlag, New York.

Hierarchies and Markets

Oliver E Williamson, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Hierarchies and markets are herein described as alternative modes of governance, the main purpose of which is to economize on transaction costs. It originates with path-breaking contributions by Karl Llewellyn (1931) on contract law; John R. Commons (1932) on the need to come to terms with the three conditions of conflict, mutuality and order; Ronald Coase (1937) on the need to make provision for positive transaction costs; and Chester Barnard (1938) on the neglected importance of coordinated adaptation. These fundamental ideas were joined and operationalized in the 1970s as Transaction Cost Economics took shape and has been a work in progress since. Not only does a predictive theory emerge, but a large and growing empirical research literature that is broadly supportive has taken shape. Public policy ramifications proliferate.

Whether the first market entailed the exchange of nuts for berries on the edge of the forest or the first hierarchy involved leadership for the hunt of a wild boar, it is clear that markets and hierarchies go back to antiquity. It is also clear that organizational innovations in both markets and hierarchies have contributed significantly to economic progress.

As between the two, markets have the better reputation among economists. That goes back to Adam Smith (1776) and continues today. Thus, whereas Smith celebrated autonomy, specialization, and the marvels of the market, he had grave concerns over the separation of ownership and control in large joint stock companies and the excesses of managerial discretion that accrued thereto. Indeed, but for grudging acceptance of economies of scale due to technology, hierarchy since has been under a cloud. Only recently has a symmetrical approach to organizations of all kinds – markets, hierarchies, bureaus – begun to emerge.

Foundations

The neoclassical theory of the firm that had taken shape by the 1930s described the firm in technological terms – as a production function – to which a profit-maximization purpose was ascribed. The adverse consequences of monopoly, monopsony, and monopolistic competition were all readily displayed with, this abstract apparatus, but issues of organization – hierarchy – were scanted by this black-box construction.

That, however, began to change with the appearance of radical new ideas about organization by, economists (Coase, 1937; Commons, 1932), lawyers (Llewellyn, 1931), and organization theorists (Barnard, 1938). Coase advanced the proposition that firms were not preexisting technological entities but that firms and markets were alternative modes of organization. Whether the firm should make or buy was, thus, not, technologically determined (although the condition of technological separability between stages was relevant) but needed to be derived. Coase further averred that the choice between market and hierarchy was decided principally by transaction cost differences. The canonical problem, then and since, for the theory of markets and hierarchies, was that of vertical integration.

In principle, this choice was easy: holding technology constant, the firm should produce to its own needs when the transaction costs in the firm were less than in the market, and it should buy in the market when this difference was reversed. Awaiting, however, a predictive theory of when transaction costs went one way rather than another, that remained a tautological construction. Thus, although after-the-fact choices could always be rationalized in transaction cost terms, that were undisciplined, and such ad hoc explanations were understandably met with grave skepticism. The twin ideas of treating firms and markets as alternatives and of appealing to transaction costs to explain the allocation of activity between them nevertheless turned out to be very fruitful.

Coase went on, moreover, to pose two deep puzzles. If the market is a marvel, why do we see so much activity organized in firms? And if the firm enjoys management advantages over markets, why is not everything organized in one large firm? Any would-be theory of firms and markets needed to come to terms with both of these issues.

Commons' approach to economic organization was different than but also complementary to that of Coase. Commons was intensely aware of the tensions that beset organization and averred that "The ultimate unit of activity … must contain in itself the three principles of conflict, mutuality, and order. This unit is a transaction" (1932). Not only does the triple to which Commons referred – conflict, mutuality, and order – prefigure the idea of governance, in that governance is a means by which to achieve order, thereby to mitigate conflict and realize mutual gains, but viewing the transaction as the unit of analysis turns out to be very instructive.

Llewellyn took exception with the legalistic way in which both lawyers and economists viewed contract. Rather than interpret contract in the prevailing legal rules tradition, Llewellyn advanced the more elastic concept of contract as framework, according to which a contract between two parties "almost never accurately indicates real working relations, but … affords a rough indication around which such relations vary, an occasional guide in cases of doubt, and a norm of ultimate appeal when the relations cease in fact to work" (1931). Llewellyn's ideas prefigure the concepts of incomplete contracting (as against the complete contracting tradition in both law and economics) and of private ordering (as against court ordering).

Barnard brought his experience as a businessman and his instincts as a social scientist to the study of complex organization. Interestingly, both the economist Hayek and the organization theorist Barnard agreed that adaptation was the central problem of economic organization. But whereas the adaptations to which Hayek (1945) referred were autonomous adaptations in the market, the adaptations to which Barnard referred entailed "the kind of cooperation among men that is conscious, deliberate, purposeful" (1938). These two work out of different mechanisms. Autonomous adaptations of a spontaneous kind are made in response to changes in relative prices signaled by the market; cooperative adaptations of an intentional kind are implemented through administrative processes within firms. Adaptations of both kinds being crucial to a high performance economy, there is a need to treat hierarchies and markets in a combined way.

This collection of deep insights – from economics, law, and organization theory – provided an interdisciplinary, foundation for a science of organization. Aside, however, from work at Carnegie (by Simon (1947, 1957) and by Cyert and March (1963)), much of which was inspired by Barnard, such a program was slow to take shape. Instead, many of the prescient insights from the 1930s remained fallow for the better part of the next 35 years. The pressing but unmet needs were (1) to combine and (2) to operationalize these profound insights. Law, economics, and organization theory needed to inform one another, thereafter to be joined in a predictive theory of economic organization.

Operationalization

Work of a combined kind got under way in the 1970s (Williamson, 1971, 1975, 1979; Alchian and Demsetz, 1972; Jensen and Meckling, 1976; Klein et al., 1978) and has grown exponentially since. Six key moves in the progressive operationalization of the markets and hierarchies research program are sketched here.

Human Actors

According to Simon (1985), "Nothing is more fundamental in setting our research agenda and informing, our research methods than our view of the nature of the human beings whose behavior we are studying." Both the cognitive and self-interestedness features of human actors are crucial to the economics of organization.

Bounded rationality, according to which behavior is intendedly rational but only limitedly so, gradually has supplanted hyper-rationality as the basic cognitive condition. Opportunism, whereby economic actors are sometimes given to strategic behavior, has come into common use to describe self-interestedness. Bounds on rationality notwithstanding, economic actors are also presumed to possess foresight, which is crucial to the Commons triple. Taken together, these three attributes lead to the following: (1) all complex contracts are unavoidably incomplete (bounded rationality), (2) contract as mere promise unsupported by credible commitments poses hazards (opportunism), and (3) parties who are alert to prospective hazards will work out the ramifications and factor these into the organizational design (foresight).

Transactions

If the transaction is the basic unit of analysis, then the key dimensions with respect to which transactions differ need to be named and the ramifications worked out. Among the main attributes of transactions are: (1) the frequency with which transactions recur, (2) the uncertainty (disturbances to which transactions are subject), and (3) the condition of asset specificity. This last attribute has reference to the degree to which the assets that are used in support of a transaction can be redeployed to alternative uses and users without loss of productive value. Asset specificity takes several forms, of which physical asset specificity (specialized dies), human asset specificity (firm-specific learning), site specificity (locational), dedicated assets, brand name capital, and temporal specificity are leading types. A condition of bilateral dependency develops when significant investments in transaction-specific assets are made. Inter-temporal governance complications arise as a consequence.

Organization

Alternative modes of organization are described not in marginal but in discrete structural terms, which is to say that each mode of organization – market, hybrid, firm, and bureau – is described as a syndrome of internally consistent attributes. Key dimensions for describing alternative modes of governance include (1) incentive, intensity, (2) administrative controls, and (3) the contract law regime (Williamson, 1991). (The market mode – with strong incentives, limited controls, and legalistic dispute settlement – favors autonomous adaptation; the firm mode – with weak incentives, extensive controls, and internal dispute resolution – favors cooperative adaptation.) The upshot is that alternative modes have distinctive strengths and weaknesses. They differ from one another in kind rather than in degree.

Whereas standard economic theory describes the firm as 'a point or at any rate a black box,' the economics of organization maintains that 'firms are palpably not points.' They have internal structure. This structure arises for a reason (Arrow, 1999). The markets and hierarchies research paradigm describes the firm as a complex hierarchy in which both governance and specialization purposes are served. Cognitive and behavioral variance among the human actors in the population presents opportunities for specialization; and transactions for which cooperative adaptations are needed (namely, those for which assets are highly specific and need to be adapted to successive disturbances, given that contracts are incomplete) are ones for which added governance supports are needed. The concept of firm-as-governance structure – as a means by which to induce order, relieve conflict, and realize mutual gains for transactions that are in need of cooperative adaptations – thus comes to the fore.

Note that both with respect to the unit of analysis (the transaction) and the concept of the firm-as-governance structure, the study of economic organization operates at a more microanalytic level of analysis than had been customary in

economics (Williamson, 2008) – broadly in the spirit of Simon's observation that (1984: p. 40):

> In the physical sciences when errors of measurement and other noise are found to be of the same order of magnitude as the phenomena under study, the response is not to try to squeeze more information out of the data by statistical means; it is instead to find techniques for observing phenomena at a higher level of resolution. The corresponding strategy in economics is obvious: to secure new kinds of data at the micro level.

Refutable Implications

Most of the refutable implications pertaining to governance are derived from the discriminating alignment hypothesis, according to which transactions, which differ in their attributes, are aligned with governance structures, which differ in their cost and competence, so as to effect a (mainly) transaction cost economizing result. Vertical integration in intermediate product markets is the canonical case, in relation to which a large number of variations on a theme have been worked out – for labor markets, capital markets, vertical market restrictions (such as franchising), regulation (and deregulation), economic reform (especially privatization), and public bureaus. More generally, any problem that arises as or can be posed as a contracting problem can be examined to advantage in transaction cost economizing terms.

A second class of predictions arises in conjunction with specialization. Upon making provision for variance among human actors (in cognitive and self-interestedness respects), many opportunities to realize economies by matching the attributes of human agents to the needs of a task present themselves. This second class of predictions has been less thoroughly worked out but constitutes a promising research opportunity.

Empirical

Some theories of economic organization make little effort to advance refutable implications. Among those that do, few are empirically tested. Simon (1991) evidently believes that transaction cost economics is remiss in empirical respects: awaiting empirical testing, the new institutional economics and related approaches are acts of faith, or perhaps of piety.

Coase had registered similar concerns about the dearth of empirical work on contract and organization 20 years earlier, but that was before the operationalization of transaction cost economics had begun and predicted alignments were advanced. Empirical applications of transaction cost economics got under way in the 1980s and have since grown exponentially. It could have been otherwise, but the theory and evidence display remarkable congruity (Masten, 1995). As Joskow (1991) puts it, "this empirical work is in much better shape than much of the empirical work in industrial organization generally."

Combining refutable implications with empirical testing has been denoted as the fourth precept of pragmatic methodology (Williamson, 2009), broadly in the spirit of Nicholas Georgescu-Roegen's (1971) statement that although 'The purpose of science in general is not prediction but knowledge for its own sake' prediction is nevertheless 'the touchstone of scientific knowledge.' As most economists know in their bones, theories that are more congruent with the data are also more influential.

Full Formalization

The 1930s was the decade when the informal ideas for a new science of organization took shape, the 1970s was the decade when these ideas were worked up in a preformal way, the 1980s was the decade of semiformal analysis, and the 1990s witnessed a burgeoning of work of a fully-formal kind. The article by Grossman and Hart (1986) on *The Costs and Benefits of Ownership* was the first fully formal model of an incomplete contracting kind. Hart's book on *Firms, Contracts, and Financial Structure* (1995) elaborates upon and develops these ideas. Additional work of this kind has been published since and more is in progress. As, matters stand presently, however, there is a huge gap between the semiformal and fully-formal stages.

A leading reason for the break is that the fully formal analysis of incomplete contracts relies entirely on property rights assignments to assess comparative efficiency. Not only are the discrete structural differences in organization among alternative modes of governance denied, but the proposition that adaptation is the central problem of organization (of which two types are distinguished) is ignored. These differences are consequential, whence Kreps (1999) concludes that "If Markets and Hierarchies has been translated into game theory using notions of information economics, it is a very poor translation … In particular, mathematical-based theory still lacks the language needed to capture the essential ideas of bounded rationality, which are central to … transaction costs and contractual form." Kreps (1999) nevertheless remains sanguine that formal efforts will succeed eventually, possibly with the benefit of new mathematical apparatus. Full formalism efforts to accomplish this purpose are works-in-progress (Tadelis and Williamson, 2013).

Conclusions

Whereas once it was customary to talk about markets or hierarchies, where the former denoted decentralization and the latter central planning, that has changed. Now it is more customary to talk about, markets and hierarchies. That entails much more than a replacement of the conjunction 'or' by 'and.' It represents a veritable sea change in the way economic organization is studied. The combined study of law, economics, and organization that is herein described (1) involves comparative institutional analysis in which economizing is the main case and the action resides in the attributes of human actors, transactions, and governance structures, (2) generates numerous refutable implications, (3) invites empirical analysis, (4) has many public policy ramifications, and (5) invites progressive formalization. Many economists are now persuaded that the study of economic organization needs to be studied in a combined way in which law, economics, and organization are joined. Legal scholars have

incorporated many of these ideas into contract law, corporate governance, antitrust enforcement, and regulation. And organization theorists have come to appreciate the pervasive reach of economizing reasoning.

As discussed by Williamson (2010), new problems and challenges appear to be unending. Productive tension between transaction cost economics and each of the parts on which it stands – law, economics, and organization – can be projected safely into the future.

See also: Administration in Organizations; Llewellyn, Karl Nickerson (1893–1962); Organization: Overview; Organizational Control; Organizations, Sociology of; Organizations: Authority and Power; Transaction Costs and Property Rights.

Bibliography

Alchian, A., Demsetz, H., 1972. Production, information costs, and economic organization. American Economic Review 62 (December), 777–795.

Arrow, K., 1999. Forward. In: Carroll, G., Teece, D. (Eds.), Firms, Markets and Hierarchies. Oxford University Press, New York, pp. vii–viii.

Barnard, C., 1938(1982). The Functions of the Executive, fifteenth ed. Harvard University Press, Cambridge, MA.

Coase, R., 1937. The nature of the firm. Economica 4, 386–405 (Reprinted in: Williamson, O.E., Winter, S. (Eds.), 1991. The Nature of the Firm: Origins, Evolution, Development. Oxford University Press, New York, pp. 18–33.).

Commons, J.R., 1932. The problem of correlating law, economics, and ethics. Wisconsin Law Review 8, 3–26.

Cyert, R.M., March, J.G., 1963. A Behavioral Theory of the Firm. Prentice-Hall, Englewood Cliffs, NJ.

Georgescu-Roegen, N., 1971. The Entropy Law and Economic Process. Harvard University Press, Cambridge, MA.

Grossman, S.J., Hart, O.D., 1986. The costs and benefits of ownership: a theory of vertical and lateral integration. Journal of Political Economy 94 (August), 691–719.

Hart, O., 1995. Firms, Contracts, and Financial Structure. Oxford University Press, New York.

Hayek, F., 1945. The use of knowledge in society. American Economic Review 35 (September), 519–530.

Jensen, M., Meckling, W., 1976. Theory of the firm: managerial behavior, agency costs, and capital structure. Journal of Financial Economics 3 (October), 305–360.

Joskow, P., 1991. The role of transaction cost economics in antitrust and public utility regulatory policies. Journal of Law, Economics, and Organization 7 (special issue), 53–83.

Klein, B., Crawford, R.A., Alchian, A.A., 1978. Vertical integration, appropriable rents, and the competitive contracting process. Journal of Law and Economics 21 (October), 297–326.

Kreps, D.M., 1999. Markets and hierarchies and (mathematical) economic theory. In: Carroll, G., Teece, D. (Eds.), Firms, Markets, and Hierarchies. Oxford University Press, New York, pp. 121–155.

Llewellyn, K.N., 1931. What price contract? an essay in perspective. Yale Law Journal 40, 704–751.

Masten, S., 1995. Introduction to vol. II. In: Williamson, O., Masten, S. (Eds.), Transaction Cost Economics. Edward Elgar, Brookfield, VT.

Simon, H., 1947. Administrative Behavior. Macmillan, New York.

Simon, H., 1957. Models of Man. Wiley, New York.

Simon, H., 1984. On the behavioral and rational foundations of economics. Journal of Economic Behavior and Organization 5, 35–56.

Simon, H., 1985. Human nature in politics: the dialogue of psychology with political science. American Political Science Review 79, 293–304.

Simon, H.A., 1991. Organizations and markets. Journal of Economic Perspectives 5 (Spring), 25–44.

Smith, A., 1776 (1922). The Wealth of Nations. Dent, London.

Tadelis, S., Williamson, O., 2013. Transaction cost economics. In: Gibbons, R., Roberts, J. (Eds.), The Handbook of Organizational Economics. Princeton University Press, Princeton, NJ, pp. 159–189.

Williamson, O.E., 1971. The vertical integration of production: market failure considerations. American Economic Review 61 (May), 112–123.

Williamson, O.E., 1975. Markets and Hierarchies: Analysis and Antitrust Implications. Free Press, New York.

Williamson, O.E., 1979. Transaction-cost economics: the governance of contractual relations. Journal of Law and Economics 22 (October), 233–261.

Williamson, O.E., 1991. Comparative economic organization: the analysis of discrete structural alternatives. Administrative Science Quarterly 36 (June), 269–296.

Williamson, O.E., 2008. Transaction cost economics: the precursors. Economic Affairs 28 (3), 7–14.

Williamson, O.E., 2009. Pragmatic methodology: a sketch, with applications to transaction cost economics. Journal of Economic Methodology 16 (2), 145–157.

Williamson, O.E., 2010. Transaction cost economics: the natural progression. American Economic Review 100, 673–690 (American Economic Association).

Higher Education Management

Juergen Enders, School of Management, University of Bath, Bath, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Higher education management concerns the complex set of actors and processes responsible for decisions within universities. Higher education management is a historically recent phenomenon and linked to changes of the institutional form of the university. Traditionally, universities have been governed or administrated while not 'management.' The focus of this article is on four main issues: (1) the study of the university as a special organization; (2) the managerial revolution in higher education; (3) management roles and effectiveness; and (4) the further study of management in higher education.

Higher education management concerns the complex set of actors and processes responsible for decisions within universities and other higher education organizations about their goals and means (human and physical resources). The role of management is a multifunctional one: strategic leadership and institutional entrepreneurship, decision making, and negotiation, day-to-day management, custody of institutional values, mediation between the organization and its environment. It is a complex one, internally embedded into a technostructure of specialized knowledge and participation in decision making, and into a more or less shared responsibility with the academic professionals; externally bound to the complex environment of higher education organizations: governmental policies for the funding and regulation of universities (in many countries including their decision-making structures and the roles and responsibilities of the leadership and management itself); the behavior and strategies of other universities, of research funding organization, of students, parents, and other stakeholders; and wider societal expectations on the role of higher education.

Higher education management is a historically recent phenomenon and intimately tight to changes of the institutional form of the university as an organization. Traditionally, universities have been governed or administrated while 'management' has been thought of as a term and practice from the business world alien to them. The study of higher education management is thus also a relatively recent field of scholarly interest. The focus of this article is on the rise of the managerial function in higher education, and four main issues will be discussed: (1) the study of the university as a special organization; (2) the managerial revolution in higher education; (3) management roles and leadership effectiveness; and (4) the further study of management in higher education.

The University as a Special Organization

For decades, students of organizations have characterized universities as a specific kind of organization and have developed highly influential concepts addressing the limited rationality of organizational behavior based on the study of the university, such as 'organized anarchy,' 'garbage-can decision-making,' 'loose coupling,' 'professional bureaucracy,' 'leadership and ambiguity' (Cohen et al., 1972; Cohen and March, 1974; Mintzberg, 1993; Pfeffer, 1977; Weick, 1976). Organizational management played a minor or ambiguous role in this perspective. The study of the university highlighted its fragmented nature within decentralized structures, the important role of the academic collegium in its self-governance, weak capacities to govern decision-making processes from the top, the professional value of academic freedom; an organizational form that seemed coherent with the unclear and ambiguous core technologies of the university and the consensus-based nature of scholarly communities within universities. While many properties of the university can be found in other organizations (schools, knowledge-intense firms, professional public sector organizations), universities provide a unique mix of characteristics that are a challenge for leadership and management.

The traditional university resembles the characteristics of the 'arena,' an organization guided by public authorities and academic professionals, rather than by goals and decisions generated at the apex of the organization. In arenas the members perform their activities relatively free from the control of the leadership; there is a blurred hierarchy or more than one, such as a hierarchy of professionals for academic affairs and a hierarchy of bureaucrats for financial and administrative issues. In consequence, universities are characterized as being bottom-heavy, fragmented into numerous academic communities with low potency for collective action. Universities are held together by bureaucratic-administrative rules and procedures, and the values and norms of academic communities. Organizational leadership and management are weak compared to other organizations. Organizational change takes place mainly through continuous local adjustments, while major change is difficult to achieve. Central organizational policies are often responsive rather than proactive, and interventions on this basis may have only minor, local effects.

This perspective is gradually changing reflecting an ongoing sociopolitical struggle about the institutional form of the university as a public agency, a cultural institution, or a knowledge business (Bleiklie, 1998). The rise of management in higher education reflects re-definitions of the socio-economic role of universities, changes in their regulation and funding as well as wider processes of the diffusion of business-inspired organizational forms into other domains of society.

A Managerial Revolution

Traditionally, the term 'management' was not part of the vocabulary of universities except for pointing at ways of running business organizations alien to them. The university was governed and administered but not managed (Lockwood, 2011). Since the 1970s, the development is characterized by the emergence and spread of 'management' as an increasingly accepted though contested principle of organizing and leading the modern university. One indication for this trend are changes in the role of top-level and mid-level academic leadership that is expected to take over a more managerial role as well as changes in the administrative pillar of universities toward managerial roles and practices. Another indication concerns the rise of 'management talk' in universities, e.g., about line management, targets and accountability, full costing, strategic planning. Finally, concepts and practices derived from New Public Management approaches and the rise of 'managerialism' point at changing practices in the regulation and organization of universities based upon a philosophy that favors managerial authority over professional self-governance and an executive leadership style over symbolic leadership.

One of the first articles pointing at a managerial revolution (Rourke and Brooks, 1964) investigated changes in US universities; universities which had a long though contest tradition of strong presidential leadership but few practices derived from management science. Their case studies among a sample of state universities illustrate the advent of management practices indicated by the rise of institutional research, increasing gathering of quantitative data for internal policy analyses, the use of computers for administration and planning, the rationalization of the use of resources, and a new more public style of administration (internal transparency and external accountability). The influential book of Blau (1973) was among the first studies focusing on the administrative structure of universities and colleges in the American higher education analyzing the problem of reconciling the conflict between the administrative hierarchy and its requirements and the academic pursuit of scholarship. A more widespread political talk about the need for a more professional management in higher education emerged in the 1970s, namely in the UK that subsequently became a front runner and exporter of new management ideas and practices for higher education (Ferlie et al., 2008). In the 1980s, governments across continental Europe and other parts of the world began to reconsider the role of higher education and public research for society and economy, the governmental regulation and funding of universities as well as their internal administration (Amaral et al., 2003). Governmental attempts at withdrawing from a juridical and bureaucratic policy style and strengthening the strategic capabilities of universities as organizations, including their leadership and management, became a hallmark of higher education policy. New policy tools were implemented, such as political target setting, steering through indicator-based funding, contracts between governments and universities, the establishment of evaluation systems to monitor the quantity and quality of outputs, and strengthening management for internal control and strategic responsiveness.

The quality management of teaching and learning became one of the internationally most widespread developments that comprises in most cases the management of external quality requirements and related reporting, in turn calling for some processes of internal management of quality. Another widespread development concerns the rise of management of financial resources becoming the responsibility not only of central leadership offices and financial officers but of deans, head of departments, group leaders or chairs. They all gain in responsibility and use more time in understanding the resource situation and – depending on context – on resource decisions to be made in staffing, equipment, space allocation, student marketing, grant applications, and the time of the academic staff. Human resource management and performance management, namely in the context of evaluations, entered the university by introducing organizational peer reviews, staff appraisals and measurable quantitative indicators opening academic activities to internal and external scrutiny. Next to that, organizational policies of setting priority areas for research and teaching touch upon the academic heartland of universities. Such attempts to control academics' work are a considerable source of tension between their professional values and priorities, and the leadership and management of universities. Estate management also evolved over time – especially when universities already held or were taking over ownership for their campuses and facilities, and when physical interfaces with their environment emerged, such as joint public/private laboratories science parks, innovation or entrepreneurial centers. Technology management, brand management, and risk management entered universities as well (for an overview on the spread of management in universities, see Lockwood, 2011). Over time, management sediments into the organization vertically to the faculty, the department, the group, the individual; and horizontally across the structures and activities of the university.

In the broadest sense, organizational responses to environmental change have been identified as the main drivers for this spread of management in universities. In one influential school of thinking, the functional need for management arises – just like in business, even though much later – from the increasing scale and complexity of higher education and research (Cameron and Ulrich, 1986). Educational expansion leads to massification, a growth and diversification of the student body as well as of the higher education institutions hosting them. Research experiences its own expansion in terms of 'big science' with its permanent need for new and costly infrastructure. Technology also becomes more important for teaching inside and outside the class room. Internationalization in its various forms (of the student body, of academic staff, of curricula, of the campus) creates a need for more planning and management. Financial efficiency, difficult choices about the use of scarce resources and pressures for effective accounting, fuel the rise of university management and related offices for fundraising, auditing, financial reporting, and management.

In a second school of thinking, management or rather 'managerialism' forms part and parcel of an ideologically motivated political project of New Public Management (Reed, 2002). Political reforms aim at increasing the efficiency and effectiveness of the higher education sector by reinforcing the autonomy of universities in their operations (especially in countries where universities were embedded into a tight regime

of state-bureaucratic control), enabling managerial discretion and internal control, making universities more accountable for their outputs through evaluation systems. At the system level, the introduction of competition and performance-based resource allocation are meant to increase performance under the assumption that a virtuous cycle exists between competition, managerial choices, organizational specialization, and efficacy (Braun and Merrien, 1999).

Third, neoinstitutionalists interpret such developments as deeply rooted in beliefs of modern societies that create their own 'myths of rationality.' 'Organizations' and 'management' form important principles in organizing modern societies, they penetrate and transform all sorts of social domains (Drori et al., 2006). The late rationalization of the university, in the Weberian sense of increasing formalization and standardization calling for authority and coordination (Weber, 1968), is inspired by wider powerful culture scripts about the appropriate means and ends for organizing things. Diffusion processes bring such models and scripts from other sectors of society also to the university. New Public Management inspired reform agendas are just one carrier. Other carriers include international organizations diffusing the idea of the 'managed university,' consultants legitimizing the need for better managed universities, universities offering (executive) training programs for higher education management, and higher education managers diffusing the logics of the successfully managed modern university through growing professional associations.

From a resource dependency point of view, three major rationales can thus be mobilized for the expanded role of management in higher education: Organizational survival in changing and turbulent environments depends on active management attuned to changes in resource dependencies, to meeting changing environmental demands, and to mobilizing new resources to reduce existing dependencies. Further, in a managerial and entrepreneurial climate, universities gain legitimacy by developing business-like properties (or symbolically managing their images accordingly) to show their responsiveness to the rise of the idea of the university as a knowledge firm and of modern management. Finally, managers can mobilize these arguments to claim a stronger role in representing the university as a whole to the environment and to expand their role and responsibility internally (Gumport, 2000).

While the rise of management in higher education has become an international theme, the timing as well as the breadth and depth of strengthening management differ, however, substantially across national settings. Overall, leadership and management have been promoted most strongly in higher education systems characterized by growing competition for resources on quasimarkets, rising expectations of accountability and evaluations, and governmental interventions inspired by New Public Management ideas.

Management Roles and Leadership Effectiveness

The study of management roles and leadership effectiveness is a growing while still rather limited area of scholarly investigation. Reviews of the literature (De Boer et al., 2010; Bryman, 2007; Lumby, 2012) show that most research has been undertaken in Anglo-Saxon higher education systems, most notably the United States, while increasingly so in Australia and the UK, and has been focused on those in formal leadership and management positions. On this backdrop, a trend has been observed away from academic leaders and managers being first among equals toward more overt managerial roles at the presidential/vice-chancellor level and the dean/head of department level, a move from elected to appointed offices, the recruitment of nonexecutive managers from outside of universities, and a reduction in the representation and influence of faculty in decision-making bodies. At the same time, leadership of universities and their academic units is usually still awarded to reputed academics, and their terms of office tend to be fixed-term calling attention for, both, their management responsibility as well as their further academic career and return to their peers.

Most studies on management roles rely on perception data gathered among the managers and 'their' academics and tend to fall into two camps (for a literature review and further references, see Bryman, 2007; Lumby, 2012). On the one hand, studies show, and sometimes normatively embrace, an emphasis on the import of ideas and practices from business based on the assumption that the latter are superior to the old academic leadership and public administration of universities. They highlight the emerging new roles, for example of deans, in strategic management of the university and their faculty, in operational management, human resource management and performance evaluations, the academic management of teaching and research programs, and external stakeholder relationship management. On the other hand, studies show, and sometimes normatively embrace, an emphases on the detrimental effects of the emerging management of universities on the commitment of academics; undermining a sense of collegiality, values of academic freedom and participation in decisions. They highlight the gap between academic and managerial norms and values, and the rise of management as a substitute for the fading faith into professionalism and professional self-regulation.

Some studies provide a look into the recent politics of universities as organizations and their negotiated order. Leadership authority, even in a supposedly powerful office such as the US university president, is constrained by many external and internal factors (Birnbaum and Eckel, 2005). Hierarchical power is not always exploited by formal instruments of command and control; academic management also stimulates and coordinates by means of soft steering, legitimacy and prestige. Further, the rise of leadership and management in higher education is contested by the responses of academics resisting, negotiating, or ignoring managerial practices and coshaping the exercise of hierarchical power within the organizations (Trowler, 1998). Management in universities – like in many complex modern organizations – is a multilayered phenomenon comprising not only the few in formal management offices but the management responsibilities of the many academic professionals, and the silent work of nonexecutive management staff; a perspective also highlighted in the more normative writing of Clark (1998). Management itself, especially at the academic mid-level, can be reluctant to embrace new managerial roles and practices, defending academic interests rather than managerial ones, and

buffering academic work (Deem and Brehony, 2005). In addition, in the past universities provided barely any training for academic leadership and management (Kort-Krieger and Schmidt, 1982) and 'learning by doing' tends to be the dominant pattern until today.

The latter perspective is supported by rare studies on leadership effectiveness in universities that are methodologically limited by measuring the perception of leadership effectiveness rather than the mediating role of leadership for organizational performance (for a recent overview, see Bryman, 2007; Lumby, 2012). Leadership behavior associated with effectiveness comprises many aspects that can be found in the wider leadership literature, such as personal integrity and trust, consideration, and sense of direction. Other aspects refer to an intense association of leadership effectiveness with behavior that fosters academic participation in decision making, a collegial atmosphere, the setting and defending of academic values. Establishing a vision or wider goal is another item identified for effective leadership behavior. Members of staff tend, however, to report day-to-day management as the dominant activity of leadership rather than visionary and strategic leadership. In contrast, leadership literature reporting self-perceptions of academic leaders especially at the top-level of the universities tends to be biased toward narratives of heroic leadership and successful self-representation. Overall, the literature on leadership effectiveness does, however, imply that leadership can make a difference for the academic effectiveness of a university or its subunits. Under which conditions which type of leadership would make a difference for which outputs remains, however, an open question.

Toward a Loosely Coupled Tight Organization?

The task of integrating academics and their basic units into a single organization and to align them with organizational goals, structures, and processes, has been and remains complex and difficult. There is evidence that universities are moving from being administrated to being managed organizations, that changes strengthening the strategic actorhood of the university as well as its leadership and management are taking place. Some students of higher education have argued that this move has so far been toward the 'managed professional' model of the university (De Boer et al., 2007) that follows managerial principles while retaining traditional professional values and practices. Limitations in the rise of the managerial revolution have been ascribed to the inherent technological uncertainty of their core activities; their deeply embedded fragmentation; and competing sources of coordination. Institutional forms are, however, not permanent; technical properties of outputs are socially defined; and environments and organizations can redefine their products, services, and technologies. The study of higher education management has thus a lot to offer for enhancing our understanding of the changing 'nature' of the university as an organization and the classical issue of understanding processes of organizational innovation in day-by-day management (March and Simon, 1958).

One promising line of future research concerns the diffusion, adoption and rejection of concepts, policies and practices of higher education management in the context of globalization. On the one hand, distinct national institutional arrangements as well as ways of running universities persist. On the other hand, the increasingly interconnected world of higher education fosters the diffusion of a managerial culture in higher education. It is an open question to what extent a more business-like type of management travels around the world, leads to convergence along dominant templates or a hybridization of management cultures. Another promising line of research concerns organizational in-depth studies, ethnography and longitudinal studies of leadership and management within the negotiated organizational order of universities. So far, the dominance of perception data and self-reporting is limiting our insight into the contested enactment of management in higher education, the actual managerial practices and their limitations, including the management by the 'few as well as by the many.' Finally, and given the widespread talk about management for efficiency and effectiveness in higher education, studies addressing the impact of leadership and management on the actual performance of universities and their academic units are a promising area of future research.

See also: Educational Effectiveness, The Field of; Globalization in Education; Higher Education Market; Higher Education Research.

Bibliography

Amaral, Alberto, Meek, Lynn V., Larsen, Ingrid (Eds.), 2003. The Higher Education Managerial Revolution? Kluwer Academic Publishers, Dordrecht/Boston/London.

Birnbaum, Robert, Eckel, Peter D., 2005. The dilemma of presidential leadership. In: Altbach, Philip G., Berdahl, Robert R., Gumport, Patricia J. (Eds.), American Higher Education in the Twenty-First Century. Social, Political and Economic Challenges, second ed. Johns Hopkins University Press, Baltimore, pp. 340–368.

Blau, Peter M., 1973. The Organisation of Academic Work. Wiley, New York and London.

Bleiklie, Ivar, 1998. Justifying the evaluative state: new public management ideals in higher education. European Journal of Education 33 (3), 299–316.

Braun, Dietmar, Merrien, François-X. (Eds.), 1999. Towards a New Model of Governance for Universities? A Comparative View. Jessica Kingsley Publishers, London.

Bryman, Alan, 2007. Effective leadership in higher education: a literature review. Studies in Higher Education 32 (6), 693–710.

Cameron, Kim S., Ulrich, David O., 1986. Transformational leadership in colleges and universities. In: Smart, James D. (Ed.), Higher Education: Handbook of Theory and Research, vol. 2. Agathon Press, New York, pp. 1–42.

Clark, Burton R., 1998. Creating Entrepreneurial Universities: Organizational Pathways of Transformation. Pergamon Press, Oxford.

Cohen, Michael D., March, James G., Olsen, Johan P., 1972. A Garbage can model of organizational choice. Administrative Science Quarterly 17 (1), 1–25.

Cohen, Michael D., March, James G., 1974. Leadership and Ambiguity: The American College President. Harvard Business School Press, Boston.

De Boer, Harry F., Goedegebuure, L., Meek, Lynn V., 2010. The changing nature of higher education middle management. A framework for analyses. In: Meek, Lynn V., Goedegebuure, L., Santiago, R., Carvalho, T. (Eds.), The Changing Dynamics of Higher Education Middle Management. Springer, Dordrecht, pp. 229–241.

De Boer, Harry F., Enders, Jürgen, Leysite, Liudvika, 2007. Public sector reform in Dutch higher education: the organizational transformation of the university. Public Administration 85 (1), 27–46.

Deem, Rosemary, Brehony, Kevin J., 2005. Management as ideology: the case of 'new managerialism' in higher education. Oxford Review of Education 31 (2), 217–235.

Drori, Gili S., Meyer, John F., Hwang, H., 2006. Globalization and Organization. World Society and Organizational Change. Oxford University Press, Oxford.

Ferlie, Ewan, Musselin, Christine, Andresani, Gianluca, 2008. The steering of higher education systems: a public management perspective. Higher Education 56 (3), 325–348.

Gumport, Patricia J., 2000. Academic restructuring: organizational change and institutional imperatives. Higher Education 39 (1), 67–91.

Kort-Krieger, Ute, Schmidt, Peter, 1982. Participation and legitimacy conflict at West German universities. Organization Studies 3 (4), 297–319.

Lockwood, Geoffrey, 2011. Management and resources. In: Ruegg, Walter (Ed.), A History of the University in Europe, Universities since 1945, vol. IV. Cambridge University Press, Cambridge, pp. 124–161.

Lumby, Jacky, 2012. What Do We Know about Leadership in Higher Education? Review Paper. Leadership Foundation for Higher Education, London. www.lfhe.ac.uk/en/research-resources/publications/index.cfm/RP%20-%2001.

March, James G., Simon, Herbert A., 1958. Organizations. Wiley, New York.

Mintzberg, Henry, 1993. Structure in Fives: Designing Effective Organizations. Prentice-Hall, Englewood Cliffs, NJ.

Pfeffer, J., 1977. The ambiguity of leadership. The Academy of Management Review 2 (1), 104–112.

Reed, Michael, 2002. New managerialism, professional power and organisational governance in UK universities: a review and assessment. In: Amaral, Alberto, Jones, Glen A., Karseth, Berith (Eds.), Governing Higher Education: National Perspectives on Institutional Governance. Kluwer Academic Publishers, Dordrecht, pp. 163–186.

Rourke, Francis E., Brooks, Glen E., 1964. The "managerial revolution" in higher education. Administrative Science Quarterly 9 (2), 154–181.

Trowler, Paul, 1998. Academics Responding to Change: New Education Frameworks and Academic Cultures. Open University Press, Buckingham.

Weber, Max, 1968. Economy and Society: An Outline of Interpretive Sociology. Bedminster Press, New York.

Weick, Karl E., 1976. Educational organizations as loosely coupled systems. Administrative Science Quarterly 21 (1), 1–19.

Higher Education Market

Felix Maringe, Division of Education Leadership and Policy Studies, Faculty of Humanities, School of Education, University of the Witwatersrand, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The idea of markets in higher education (HE) invokes mixed reactions among academics and policy makers. Those in the global north generally associate markets with the commodification of HE and how this strips the educational enterprise of its inherent value. Issues around the creation of unevenness and the strengthening of inequality between the HE sectors across the global divide are generally not highlighted in the literature. Using theoretical underpinnings and global mobility data, this article explores factors that promote the strengthening of HE markets in the global north and those that weaken similar markets in the global south. This article argues that markets in HE tend to be exacerbating the unevenness of global HE terrains and promoting rather than removing a wide range of inequality fueled by an increasing competitiveness where survivors need to win the battle for the import of intellectual capital.

Introduction

The idea of markets in higher education (HE) is a fairly recent and highly contentious phenomenon. It is also an extraordinarily broad but under-researched field of study which includes, among others, student mobility markets; the HE funding markets; knowledge and curricula markets in HE; HE qualifications market; the public HE market; the private for profit HE market; and technology and pedagogical markets in HE. Each of these could indeed be the subject of a separate article. This article focuses on some detail on the first of these markets, as more than any other; it provides both direct and indirect evidence of the performance of other markets in HE across the globe.

Primarily, markets exist to serve the needs of buyers and sellers and, especially in the context of HE, to be platforms for the exchange of value. Essentially, sellers are motivated by a profit imperative, which tends to be unrelentingly exploitative while buyers generally look out for a bargain as they seek to meet their needs in the market while maximizing their investment. In an ideal world, markets are characterized by choice, where buyers make purchasing decisions based on a wide range of factors including product price, availability, and quality among many others. On the other hand, sellers of products and services make decisions about operating in a market based on both perceived and real need, the potential to make a profit and the presence or absence of competitors. Yet it is precisely the use of such language to analyze the new educational terrains that many people detest. 'Everything for sale,' screamed the title of a powerful critique of the marketization of UK HE by Brown and Carasso (2013) in which they make a strong argument against the growing consumerism in HE which they claim has seriously eroded the fundamental value of education.

In this article, I argue that, despite their assumed value, markets in HE tend to exacerbate polarization and unevenness across the terrains of the global HE sectors, creating a form of intellectual imperialism through which the markets of the global north cement their dominance, while the struggling markets of the global south propel rapidly into dysfunction. In pursuance of this argument, I deploy a range of theoretical perspectives, and begin with an overview of the classic market theory; the investment theory of education; the reproductive theory of education, the neoliberal theory, and the postcolonial theory. This is followed by a conceptual discussion of the key ideas which relate to the markets in HE including globalization, internationalization, and the notion of HE markets itself. A discussion of major actors in the global student mobility market will then be undertaken focusing on global mobility patterns from a point of view of both the recipient and sending countries. The article concludes with a critical discussion of the affordances and constraints associated with HE markets. It is hoped that this article will provide a different perspective to the traditional analysis of education markets we have become accustomed to.

Theorizing the HE Markets

Four broad theories will be reviewed here which shed light on the nature and performance of HE markets.

The Classical Market Theory

The classical theory exemplified in the works of Adam Smith, Jean-Baptiste Say, and Thomas Malthus and John Stuart-Mill is based on the assumption of the existence of a free market which always tends toward creating economic equilibrium. In this view, markets if left alone will maximize economic productivity. In the context of HE, if markets are allowed to operate freely, then students would make free and unfettered choices of where they want to study; institutions would decide freely what they want to teach and how, and more importantly make decisions about the price tag for their products. If, for example, enrollment in HE falls, then the institutions would respond by reducing the price. This would stimulate more interest and establish the enrollment equilibrium again. This was captured in Say's law of markets which simply states that supply creates its own demand. In other words, if higher education institutions (HEIs) create educational opportunities, then this would stimulate demand for the degree

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92118-4

programs. On the other hand, demand is also seen as an effective driver of supply. Thus, demand and supply are the cornerstones of markets and they exist in a dynamic which tends toward equilibrium. That is to say, as long as universities provide the opportunities, enrollment to the programs will be maximized. Unlike the Keynesian theory, the classical theory thus sees flexibility in price as the key stabilizing factor for markets. In reality, all these claims are only partially true. First, there is little or no real political willingness among governments in the world to allow the idea of a free market to operate in HE. If this happens, the governments cede control over issues of social justice, fairness, and quality to the markets and that often becomes a recipe for loss of political control. In reality, the governments tend to show a strong hand in issues related to participation and especially to widening access to previously disadvantaged groups not only as an entirely social justice imperative, but also as a mechanism for controlling the labor markets and the creation of both educational and employment opportunities for people with little or no history and tradition of participation in HE. The establishment of structures for monitoring access, fair admissions, price control, and quality of HE programs have become an integral part of the external control of HE in many parts of the world. As such, it can be argued that the free markets theory has no place in HE. Secondly, there is insufficient evidence that price alone creates stability in HE markets. In the UK, for example, attempts to democratize the pricing of HE have led to increased volatility in the market. Indeed, the majority of universities have fixed their prices within very narrow limits of difference charging around 9000 pounds per annum (previously set at about 3000 pounds per annum). The impact of this has been to depress rather than to stimulate enrollments, at least in the short term, across the different types of universities with prospective students voting with their feet as they sought to delay the burden of increased educational costs. However, price is not the only, nor is it the most important determiner of decisions in HE participation. Maringe et al. (2009) found that prospective students were not deterred by the new fees increases announced in the UK for the academic year 2009–10 and indeed vowed to "…survive on jam sandwiches for the next three years…" for their opportunity to participate in HE. In any case, pricing appears to have an inverse relationship with decisions to purchase, in HE. Lowly priced products and educational programs are generally associated with poor quality and tend to be less preferred than highly priced programs (Foskett and Hemsley-Brown, 2001). There is more to the value of HE that transcends its pricing. For example, Maringe and Gibbs (2009: 121), observe that:

> A cheap, unaccredited internet degree programme gives you just that; a cheap and low quality product, experience, and poor financial returns, whereas a Harvard experience gives high quality and excellent returns.

The Investment Theory of HE

Human Capital Theory (e.g., Schultz, 1961; Becker, 1993; Mincer, 1974) presents HE as both a public and private investment decision for governments and its people. As a public investment, the fundamental argument is that investing in education leads to economic growth through increased productivity, social stability, and healthier lifestyles. On the other hand, as a private investment choice, investing in education leads to increased lifetime earnings for those with more years of schooling; access to better paying jobs; reduced time spent in the unemployment market; and speedier transitions to enhanced career prospects (Wahrenburg and Weldi, 2007). In the context of contemporary HE markets, evidence does not always lend full support to these claims. For example, there are imbalances and inconsistencies related to returns to investment in HE. Psacharopoulos (1994), for example, identifies such imbalances and inconsistencies in areas of gender, degrees, and subjects studied, and levels of qualifications obtained among others. In addition, the employment market does not always behave in a fair way. For example, although there is some evidence which shows that those with higher qualifications and those who have spent more years in education tend to accumulate higher earnings across the life course, there is much less evidence which shows that investing in a highly priced degree program offers higher financial returns to the graduates. Financial returns in the professions are more closely related to the reputation of the university attended. For the same job, a Harvard graduate is likely to start off at a higher salary than one from a less well known university. In addition, graduates from community colleges in the USA tend to have limited career entry opportunities compared to their counterparts from the more reputable 4-year degree offering institutions. Community college graduates also tend to earn lower aggregate salaries over the life course compared to graduates from other more prestigious universities (see, e.g., Nagler, 2007). Others have concluded that institutions in the lower ranks of prestige and reputation in HE markets only serve to reproduce social and economic inequalities in the society (e.g., Brint and karabel, 1989; Dougherty, 1994; Karabel, 1986; Weis, 1985).

Reproductive Theory in HE

Since Karl Marx wrote that every social process of production is at the same time a process of reproduction, HE has increasingly come under the sociological spotlight as a market that reproduces both success and failure across different socioeconomic groups in society. HE has been described as being more closely inclined with the values, norms, practices, and cultures of upper and middle classes rather than with those of the working-class families (see, e.g., Giroux, 1983; Bowles et al., 2001, Apple, 1978; Willis, 1983). The central thesis behind these perspectives is that education does not change society, that is, education including HE reproduces the inequalities inherent in society. Based on this theory, it can be argued that the markets in HE create and reproduce inequality. For example, there now exists a steep intellectual divide between the HE environments in the global north compared to those in the global south which in large part has been driven by the unidirectional flows of talent from the south to the north. Further, it has been observed that various attempts at creating partnerships between institutions across this divide have

widened rather than narrowed the divide despite the good intentions to the contrary. Evidence shows very clearly, for example, that talent mobility flows are unidirectional resulting in the ever increasing phenomenon of brain drain in the universities of the global south. Equally there is evidence which shows that partnerships developed from a need to correct imbalances in intellectual capital have tended to result in little gains to the host institutions in the south and have also tended to the creation of new employment opportunities for academics in universities in the global north, thereby exacerbating the brain drain phenomenon (e.g., Mohamedbai, 2013; Ishengoma, 2011).

Neoliberal Theory and Higher Education

Harvey (2005) has argued that neoliberalism has become a dominant hegemony of HE across much of the world. As an ideology, neoliberalism is based on four fundamental assumptions: the benevolence of the free market; minimal state intervention; regulation of the economy, and the individual as a rational economic actor. On many counts, in many parts of the world, these assumptions do not stand the test of contemporary HE environments. For example, in the previous section, we have seen how governments are not prepared to loosen their grip on the mechanisms and processes associated with the delivery and quality of HE. That alone negates the assumptions of a free market and of minimal government involvement. While it may be true that competition has the potential to drive improvement and quality, there is a sense in which when the competition is between or among unequal competitors, with unequal access to power and resources, then that competition may become a seedbed for the growth of inequality. Neoliberalism is about deregulation and reduced control to allow markets to establish their own equilibriums. The assumption that the establishment of equilibrium is in itself a good thing needs to be challenged. For example, an equilibrium point which accommodates fee increases in HE does little to equalize educational opportunities across the socioeconomic groups in society. Allowing institutions to determine their fee levels with no government control creates a fee market in HE which has the potential to increase rather than reduce inequalities across institutions in terms of access to resources and to program quality issues. Finally, the assumption that HE choosers are rational economic actors is, at best, ill informed and at worst reckless imagination. The theory of rational choice only works in an environment supported by valid, accurate, and accessible information. Evidence suggests that such information to support educational choices is often not available to choosers in the market (Foskett and Hemsley-Brown, 2001).

Postcolonial Theory and HE Markets

Postcolonial theory, despite the numerous contestations surrounding its definition and stature, focuses on addressing the legacy of colonialism imposed by western nations in their attempts to dominate the world through conquest and cultural influence and to create markets for their burgeoning economies back home. In general, the three southern continents of Africa, South America, and Asia have provided the intellectual space for the growth of postcolonial theory. Its definition is difficult to capture because some argue that colonialism never ends, it only transforms (Young, 2001). However, it offers a powerful critique of the underlying purposes of colonialism and its effects on the colonized. A recurring theme in these critiques is that the postcolonial nations ought to offer resistance to all forms of control and to develop a heightened consciousness about possibilities of continued and subtle control in the postcolonial era. The theories privilege the development of local indigenous knowledge systems, use and development of local languages, and the deliberate engagement of diasporas in processes that are designed to shift and transform developmental agendas. In the context of HE markets in particular, they offer a powerful critique of the deployment of partnerships as a strategy to bridge knowledge gaps between institutions in the north and those in the south. Zeleza (2009), for example, suggests that although north–south partnerships have become a strategy of choice in attempts to equalize opportunities between the global north and south, the net impact is that, because they work on donor funding models, they tend to further the interests of the donor agencies at the expense of the uplifting objectives they were created to address. Moreover, as Tickly (1999) observes, issues relating to colonial history, economic power, culture, language, gender, class, race, and ethnicity are invariably involved, but seldom addressed or openly discussed in these partnerships. Even south–south partnerships come under severe criticism using the postcolonial lens. While they are heralded for stimulating development among like-minded nations with geographic and often cultural 'samenesses,' they tend to be sponsored by organizations and institutions in the north with a typical global north perspective of development. It is thus prudent to argue that as yet, we do not have real south–south partnerships in HE, but north–south arrangements which increase the complexity of the role and nature of international knowledge hierarchies, their production, dissemination, and management. Evidence drawn from such partnership arrangements suggests that these partnership arrangements reproduce and sometimes intensify the asymmetries between north and south global actors in the HE markets (e.g., Appiah, 2004; Ishengoma, 2011; Samoff, 2004). Thus, even though many such partnerships work well, the perpetuation of structural, cultural, intellectual, and economic asymmetries and dependencies in areas of knowledge production, application, utilization, control, and legitimation continue to characterize the relationships between HE markets in the global north and south (e.g., King, 1990; Escobar, 1995; 2005; Samoff and Carrol, 2004).

Conceptual Framework for Understanding HE Markets

In this section of the article, I seek to engage critically, albeit briefly with the ideas of markets, marketization, internationalization, and globalization and the very notion of HE itself. The analysis will by no means be exhaustive given the limitations of space, but will help situate and position the article in the context of the argument it seeks to develop.

The Notion of Markets in HE

Popularized by Adam Smith as the 'invisible hand,' a market is a system through which buyers and sellers try to outsmart each other for the best returns from their investment through exchanging value. In the context of HE, one can assume that students are the buyers and institutions the sellers while the product of exchange is education. But this characterization of HE markets is incomplete, inaccurate, and problematic in a number of ways. According to Jongbloed (2003), providers or sellers in a market should have four freedoms: freedom of entry, freedom to decide the nature of their product, freedom to use available resources, and freedom to determine prices. But as we know, entry into the HE market is highly regulated and restrictive primarily because of the high stakes involved, given their role as knowledge producers in the global context. In many ways, knowledge has replaced capital as the key resource for development in a global knowledge driven society (Carnoy and Castells, 1999; Drucker, 1993; Hargreaves, 2002; Vallima and Hoffman, 2008; Harvey, 2010). The control of knowledge and its flow across the globe including its utilization and application in the world have thus replaced the flows of money as the key weapon for human development. Governments, regional bodies, and international organizations have assumed a heightened interest in what goes on in universities and in the process have substantially transformed the freedom universities have had to teach what they want, to whom they want, and in what ways they want. While the seller's freedoms can no longer be fully guaranteed in the new global climate, the buyers' freedoms have also been radically transformed.

On the other hand, as consumers, buyers in a market should enjoy four fundamental freedoms related to choice of provider, of product, and to access to full information on prices and quality (Jongbloed, 2003). Available evidence clearly shows that such freedoms do not fully exist in the HE context. For example, a recurring theme in studies on consumer choice in HE is that choice in the market is highly constrained by lack of information, insufficient, and sometimes misleading information (Foskett and Hemsley-Brown, 2001). Despite an increasing marketization in universities, HEIs have not and seem broadly reluctant, for a whole range of genuine reasons, to brand themselves as supermarkets selling degrees to anyone who brings along the money. They maintain guarded protectionism and provide restricted entry only to those deemed deserving and qualified. For these and other reasons, markets in HE are at best quasi markets. Le Grand and Bartlett (1993) for example, argue that because of external regulatory frameworks and how universities are financed, there is little that reflects a free market in the HE context.

Equally it is anomalous to speak of a single HE market. Effectively, even within the same country or system, there exist a range of different markets in HE. For example, the entry degree market differs in many ways from the postgraduate degree market in as much as these differ from distance education markets in HE. Such differences include, for example, the broad age groups targeted by these different markets; the pedagogical approaches utilized in different markets; the product structure and quality; the quality of staff associated with these different markets; the study patterns associated with program delivery among others including whether they are public or private sector institutions, and fee paying or non-fee paying and other factors which differentiate these HE markets. On a global scale, we can speak of the receiving and sending markets depending on the direction of flows of human capital based on the net flows of human capital from the global south to the global north. Similarly, the markets in different places are undergoing significant transformation. For example, the Asian HE markets, led by China, India, Japan, and South Korea are increasingly becoming attractive destinations for human capital flows in the global HE context due to significant infrastructural capital investments, less stringent human migration laws, increasing political stability, economic progress, and increasingly maturing democratic governance systems (Dahlman and Jean-Eric, 2008). On the other hand, due to a variety of push and pull factors, and driven by largely poorly performing economies, unstable governments, wars, and poor management of natural disasters, countries in the global south tend to operate in sending markets which experience severe brain drain, a phenomenon seriously eroding the value of HE in the global south. The picture of markets in HE is thus very complex and variable.

Marketization

This is a term that came into increasing use in the early 1980s, coinciding with the dawn of liberalization in HE. Essentially, the term is used to describe the processes associated with the need for increasing efficiency and effectiveness of universities in delivering their missions. The increasing competition for resources, staff, and students and the growing importance of university league tables of performance all contributed to the marketization of HE. Broadly, the marketization idea is used to describe the effects emerging from three areas.

The first is the notion of expansion of HE systems across the world, which has increased diversity on university campuses and HE systems generally. HE systems across the world have expanded significantly, partly due to industrialization and the growth of postindustrial societies which requires higher level skills to service them; partly due to the worldwide adoption of the principle of education as a fundamental human right which has been at the heart of the Education for All philosophy that has defined global educational policies; and partly due to a global social justice imperative which has led to the idea of widening and increasing participation in HE. Worldwide, it is estimated that participation in HE has risen by approximately 40% in the last 10 years (University World News, 2012). The importance of direct marketing approaches to attract more and more people to HE have thus increased and so contributing to the marketization of HE environments.

The second factor emerges directly from the phenomenon of expansion. With expansion came diversity characterized by recruitment from a wider cross-section of people and the increasing mobility of talent from different parts of the world. The HE terrains have generally become more diverse than they have been for a long time. In many countries, the diversity of campus populations has increased tenfold over the last 15 years. For example, institutional statistics in UK and American universities indicate that the percentage of international students has been doubling annually over the last 5 years (HESA UK2010/2011 and the National Center for Educational

Statistics USA, 2012). Such diversity has brought with it both challenges and opportunities to different HE markets. For example, in terms of challenges, universities have been grappling with issues of appropriate pedagogical approaches and new instructional methodologies more suited to teaching and learning in increasingly culturally and demographically diverse HE classes. In market terms, how should the educational product be more appropriately packaged to meet the diverse needs of a broader based consumer group? On the other hand, new opportunities have also emerged. For example, those created by a need to reflect upon the nature of graduate attributes in a globalizing HE environment.

The third is a more global influence which recognizes the importance of transforming HE from one driven by 'national systems that educate citizens for local employment and national service to one that is increasingly defined as an international service to be regulated through the marketplace and through international trade agreements (Bassett, 2006: 4). In pursuance of this goal, the World Trade Organization (WTO), working within the framework of the General Agreement on Trade in Services (GATS), of which HE was a significant part, proposed that HE be viewed as part of the Free Trade Agreements, which aim, among other things, to reduce or eliminate government measures that prevent services from being provided across national borders; provide the legal framework for interrogating barriers to trade in education and finally to provide a vehicle that opens markets around the world and progressively liberalize conditions which promote and support open trade in education (Popovich, 2002). The marketization of HE environments thus became a choice strategy for enhanced institutional performance through increased effectiveness and efficiency of these institutions in the new environment. However, the performances of markets in different places are at varying levels of maturity. For example, while much has been achieved to rationalize the markets of the European Union (EU), through for example, the harmonization of qualifications and years of study, quality in different parts of Europe vary quite considerably with the relatively mature markets in Western Europe having higher quality higher education than many countries in the rest of the expanded European Union.

Hence, there is a sense in which marketization has been the product of expansion, diversification, and the international consensus around liberalizing HE services in an age of increasing globalization and internationalization.

Globalization and Internationalization in HE

I have chosen to discuss these two concepts under the same heading because they are so closely interconnected although they do not mean the same. Globalization is a term that has gradually superseded the idea of postmodernism and provides a lens which helps us understand the increasing interconnectedness of the world in terms of its politics, economics, sociocultural achievements, and broad ideological dimensions of human existence. It is based on a world development theory that privileges the view of an accelerated interconnectedness and sameness facilitated through information technologies, the Internet and the World Wide Web (WWW), and the sheer power of financial and ideological might that resides within countries and institutions in the global north especially. This implies that globalization can be imposed from above, but equally that it can be contested and configured from below. However, the homogenization theory has been hotly contested. Some people think that rather than becoming more integrated, the world is increasingly showing signs of division along lines of religion and on issues of nationhood. England, for example, despite being a full-fledged member, exists on the fringes of the European Union in terms of its economic ideology and especially with regard to joining the common European Union currency. Similarly, while there seems to be much convergence on capitalism as a practical economic theory, it seems that the ideology works better in countries where the majority of the population belong to the middle class while the benefits of this economic ideology seem to be less evenly spread in countries where the working class comprises the majority of the population. Such countries tend to experience greater political instability. This unstableness in turn, tends to be replicated in the institutions of the nations including those in HE. Globalization is thus a very contentious phenomenon, which, on one hand, defines the coming together of nations in pursuit of highly connected political, economic, cultural, social, and technological goals, creating an increasingly borderless world. On the other hand, it is equally widely seen as a postmodern project aimed at cementing the hegemonies of the rich global north through the creation of highly extractive relationships with the poor nations of global south. Discourses which highlight a strong relationship between globalization and poverty (Zedillo, 2007), inequality (Robert and Lajtha (2002), increased control by rich countries (Richardson, 2002; Zedillo, 2007), and increased growth of north south extractive partnerships and dominance through soft power approaches (e.g., Appadurai, 2001; Zeleza, 2003; Mudimbe, 1994; Chakrabarty, 2000; Rodney, 1983) do now occupy a permanent space in discussions and debates related to globalization. The growth of markets in the broad context of globalization and debate around their impact on HE are thus shaped by voices on both sides of fence.

Internationalization on the other hand is sometimes seen as having preceded globalization, at least in its present form. In HE, universities have always had an international character and intent both in terms of the staff who comprise it and in the sense of the knowledge that defines their central mission. However, as globalization intensified in the last few decades, universities have in turn intensified their internationalization agendas. Knight (2004) is credited for providing the most widely used definition of internationalization as the process through which universities seek to integrate international and intercultural dimensions into the teaching, research, and service elements of the institution. Maringe et al. (2013) observe that internationalization in universities tends to be built around motives which seek to create stronger strategic and symbolic value; knowledge eminence in the global context; social and cultural prominence; stronger positioning in the pecking orders of global comparators and league tables; and the creation of economic value in an environment in which national governments are increasingly reducing their funding levels to HE. Marketing is widely seen as an integral

part of this new agenda for universities' development and the entire marketization processes tend to be developed around these new pillars of growth. Key strategies for internationalization in universities have tended to be built around increasing cross border student recruitment, the hunt for distinguished scholars to enhance the research eminence, performance profiles, and prestige of the universities; increasing opportunities for students to engage with international dimensions as part of their degree programs through joint degrees, student exchange programs and other partnership arrangements; and through curriculum transformation processes. In their recent study, Maringe and Foskett (2010) found that the priority given to these different approaches to internationalization differed between countries and tended to reproduce and cement their positions across the global divides.

The Notion of HE and HE Markets

Stoke (1966) has argued that there is little that unifies the idea of HE as it tends to come in numerous shapes and sizes in different parts of the world. However, the one universal notion that underpins the idea of HE is the aspect of knowledge. Although knowledge itself is hard to define, there seems general consensus that systems and institutions of HE rally around a call to produce, reproduce, apply, and disseminate knowledge as the key to all forms of human endeavor. A widely accepted way of defining HE is to see it as any education offered beyond the traditional school system which often results in the award of a certificate, diploma, or degree at the end of a prescribed period of study. The term tertiary education (beyond secondary schooling) is sometimes used to mean the same as HE. The problem is that there is a lot of slippage between what is sometimes offered in secondary schools and in some colleges. For example, some colleges offer secondary school subjects leading to traditional secondary level certificates and qualifications, while some universities offer secondary level studies to students in their transitional year from school to university. For example, in Zambia, students typically do their sixth year of secondary school in a university as a preparation for university studies. This makes defining HE a very complicated issue. In addition, different countries use different nomenclature to define the idea of HE. The use of the idea of HE markets thus becomes elusive in this somewhat unsystematic and confusing terminological jungle. A review of a few HE systems helps to illustrate the complexities.

A way to gain some understanding of HE markets across the world is to engage with the notion of international student mobility. This not only provides a window into a key dimension in the behavior of markets, but also sheds light on how markets in HE are perceived globally. Five key dimensions of HE markets will be explored in the context of student mobility in the next section. These include: facts and figures relating to global mobility trends in HE; an overview of the world favorite study destinations; a review of the world's biggest exporters in the student mobility market; a discussion of issues associated with emerging markets in the global south and finally, a prospective view of what I see as the future of HE markets in the next decades to come.

HE International Student Mobility Markets

The international student mobility market has become a massive financial enterprise worth more than 500 billion US dollars per year across the world. In 2012, the OECD released the following figures in relation to international student mobility markets in HE which are corroborated by figures from the UN Statistical Data on HE:

1. In 2010, more than 4.1 million students were enrolled outside their countries of origin. Based on current rates of annual growth, it is estimated that by 2025 there will be about 25 million international students studying away from their home nations.
2. Between 2001 and 2011, the number of students enrolled outside their home countries increased threefold from 1.3 million to about 4.3 million representing an annual growth rate of about 6%.
3. The global economic crisis has not resulted in a decrease of international student mobility, contrary to expectations. If anything, the figures show a slight increase during the initial years of the crisis.
4. Global student mobility mirrors to a large extent inter- and intra-regional mobility patterns. For example, the majority of international students studying in top EU destinations such as the UK, Germany, and France tend to come from European countries.
5. The majority of international students tend to study in English speaking countries. For example, the UK, the USA, and Australia together host approximately 40% of all international students worldwide.
6. Over the last few years, new players in the international student market have emerged. These include Canada, Japan, Russia, and China. Consequently, the market share of traditional international study destinations has been eroded. For example, the US market share has dropped by about 6% between 2000 and 2011.
7. The HE Fee structures send mixed messages in terms of their influence on study decisions of international students. Despite charging higher fees for international students, the USA, the UK, and Australia command pride of position in attracting the highest numbers of international students. Conversely, despite that they do not charge any fees for international students, Finland and Norway are nowhere near the top of key destinations for international HE study. **Table 1** shows FEE practices in a range of OECD countries.
8. Asian students, mainly from India and China, make up approximately 51% of all international students studying abroad.
9. Approximately 77% of all students studying abroad do so in OECD countries.
10. The top 10 nations by % of international students enrolled are summarized in **Table 2**; the USA does not feature in this list. It is also important to note that with the exception of Australia, New Zealand, and Canada, the rest of the countries with the highest percentages of international students belong to the EU.

The following provide snippets of the top three recruiting or recipient nations in the global student mobility markets.

Table 1 Fee systems in OECD countries

Fee systems for international students	*Countries*
Countries charging higher fees for international students relative to domestic students	Australia, Austria[a], Belgium[a], Canada, the Czech Republic[a], Denmark[a], Estonia[a], Iceland, Ireland[a], the Netherlands[a], New Zealand, Poland[a], the Russian Federation[a], Sweden[a], Turkey[a], the UK[a] and the USA
Countries charging same fees for both international and home students	France, Germany, Italy, Japan, Korea, Mexico, Spain, Switzerland
Countries which do not charge any fees	Finland, Norway

[a]Applies for Non-European citizens.
Adapted from Education Indicators in Focus OECD, 2013. Organisation for Economic Co-operation and Development.

Table 2 Countries with the highest percentages of international students

Country	*% of international students enrolled in HE*
Australia	21
UK	15
Austria	15
Switzerland	15
New Zealand	14
Belgium	9
Ireland	8
Canada	6
Sweden	6
Denmark	5
Iceland	4

Adapted from OECD data Education at a glance 2011.

The US HE Market

Acknowledgment of Source

All statistical figures and information quoted in this section have been drawn from the UNESCO Institute of Statistics 2012 data and the OECD statistical data on HE markets.

In the USA, HE refers to education offered beyond school level and includes a system of public and private universities, community colleges, and liberal arts colleges. The system is probably the most well funded, and its quality has been relatively high in comparison with systems elsewhere in the world. The result has been that American universities have consistently been evaluated as the most prestigious institutions. Various league tables including the highly respected Shanghai Jiao Tong University's Academic Ranking of World Universities tend to place about 40 American universities in the world' top 50 based on awards and research outputs. Consequently, American Universities tend to be the most attractive destinations for study for the ever increasing numbers of students who choose to study away from their home nations. Almost 20% of all migrant students in HE across the world choose to study in American universities making the USA the top destination for HE study. The bulk of foreign students studying at American universities are drawn from China, India, South Korea, Japan, Canada, and Saudi Arabia. Interestingly, American students who opt to study abroad do not prioritize any of these sending countries. Rather, the five top destinations for American students seeking to study abroad are: the UK; Canada; Germany; France; and Australia. The US market is a favored destination of study for a wide variety of reasons including the prestige of the US degree brand; flexibility in pathways to degree qualifications; believability of the US qualifications in global labor markets; and the liberal social and cultural environment (Douglass et al., 2011). Closely connected with the US brand is the Canadian HE market. A big draw with the Canadian HE system is that it is cheaper than the USA and yet offers comparable quality and prestige. However, the weather and relatively smaller capacity offer drawbacks in the recruitment market.

The UK He Market

Like the US, the UK HE market comprises many different types of institutions including universities, Further Education Colleges, and some publicly designated and autonomous institutions. There is only one private university in the UK (University of Buckingham), while the rest are established by the Royal Charter as public universities and are wholly or partly funded by the government. The department for Business, Innovation and Skills (2012) website lists about 750 Further Education Colleges and degree awarding institutions together with about 63 Universities. Four of its universities are consistently in the world top 10 ranked universities for reputation, research output, and awards. Cambridge and Oxford together have affiliations with nearly 50% of all Nobel Prize winners in the world. The UK is noted for its significant contribution to English literature and for its world class inventions such as the steam engine which has been central to the successes of the industrial revolution. For these and other reasons, the UK HE system is considered very highly across the world. The UK HE brand attracts the second largest contingent of foreign students after the USA. Key exporters of foreign students to the UK HE system are China, India, Nigeria, Ireland, Germany, and the USA. The UK students tend to seek foreign study experience in the USA, France, Ireland, Australia, Germany, and Canada. Like the US market, there is little reciprocal mobility across sending and receiving nations in the UK HE market.

Student mobility to the UK HE market is stimulated primarily by widespread positive perceptions of high standards, the acceptability of the UK HE qualification in the global employment market, perceptions of a stable economy and political environment; and the use of English as a medium of instruction. However, like the USA, the UK is increasingly being perceived as exerting a hegemonic dominance across the world and its unusually high fees set at 9000 pounds for domestic students and approximately 15 000 for international students are increasingly eroding the potential of the UK as a zone of possibility for those seeking to undertake their studies overseas.

The Australian HE Market

The Australian HE market is geographically positioned close to countries with rapidly expanding economies which view talent growth as central to their developmental ambitions. It is widely seen as the most developed in the Asia–Pacific region. The fact that Australia delivers its university curricula in English is a huge advantage to those who do not wish to move too far from the region to receive a good English education. The country has approximately 50 universities, eight of which consistently appear in the top 50 for research, quality, and international profiles. Like the USA, the Australian HE market has a substantial Further Education sector which accommodates a large number of students who otherwise would not enter the conventional and highly competitive university sector. Factors that appear to erode the Australian HE brand competitiveness include: relatively poor standards of accommodation for foreign students; large distances for travel from the key metropolis in the global north and very restrictive employment and immigration laws for foreign students (e.g., the Knight Review 2011).

Key Features of Dominant HE Markets

Table 3 is based on statistical data obtained from the UNESCO Institute of Statistics and the OECD Educational Statistics data. It provides a summary of data for the top study destinations in the world.

A number of important points deserve to be highlighted in respect of these HE markets:

1. Language of instruction (in this case, English) appears to be a major factor in international students' decisions to study outside of their home nations.
2. English as a language of instruction or as lingua-franca seems to exert a very significant influence in study abroad decisions (e.g., Jenkins, 2013)
3. Perceptions of the quality of HE is a very important determinant of study abroad decisions (e.g., Mazzarol and Soutar, 2002; Cubillo et al., 2006).
4. Geographical and historical factors appear to play an important role in study aboard decisions (e.g., Chia-Ling and Kot, 2011).
5. Employment prospects also play a significant role in study abroad decisions (e.g., Lin, 2007)
6. The major importing nations are perceived to offer better opportunities for employment both during the times of study and after completion of the degrees (e.g., Mpinganjira and Rugimbana, 2009).

The Market Factors Characterizing the Biggest Sending Countries

The preceding data and discussion clearly show that China and India are the world' biggest exporters of international students. Three key factors combine to contribute to this strong HE export orientation in both countries.

Population Size

According to the Wall Street Journal India (2013), the two countries are the world's most populous nations. China's population currently stands at 1 339 724 852 while that of India is 1 210 193 422, a difference of about 130 million. However, because China has a one child per family policy, its rate of population growth is becoming smaller than that of India. As such, India is expected to be the most populous nation by 2050. Of these figures, the two countries currently have the largest sizes in the youth population groups. For example, about 144 million of India's population is currently between the ages 18 and 23 years. That means the two countries are host to the largest groups of young people who will be expected to engage with HE both presently and in the future (e.g., Gupta and Gupta, 2012).

Prevailing Perception of Low Quality Across the Sector

Although both India and China have large numbers of universities and other forms of HEIs and even a lower of student per university ratio compared to the USA and the UK, there are very few institutions in both countries which compete on an equal footing with some of the leading universities in the world. For example, according to The Times HE Universities ranking, China only has three universities listed on the top 200. Peking University is the only Chinese university frequently ranked in the top 50 universities in the world. The Indian Institute of Technology at Kharagpur is the top University in India featuring between 200 and 250 on different ranking systems. Speaking at the valedictory function of the golden jubilee celebrations of the Dayanand Education Society in Latur, President Pranab Mukherjee noted: "We have over 650 degree awarding institutions and over 33,000 colleges. But we are short of good quality academic institutions, on account of which many bright students prefer going abroad for higher studies." Students with strong government and parental support thus make the choice to study abroad. For both China and India, such students prefer studying either in the USA or the UK.

Strong Government Commitment to Economic Growth

Both China and India are gradually acknowledged as the world's fastest developing economies. China is predicted to overtake the USA as the world's largest economy by 2020. There is realization that such growth can only be maintained and propelled by an equally aggressive human resources development. As such, both countries, through excellent scholarship programs send some of their brightest youth to study in US and UK top universities in critical degree programs in science, engineering, technology, medicine, pharmacy, and management. There is also very strong parental support in both countries especially among the middle classes for sending children overseas for their degree programs.

Table 3 Market performance indicators of the top six study destination countries

Top host nations	*Number of international students hosted*	*% of global quota*	*In/out bound rates*	*Top inbound courtiers*	*Top outbound destinations*	*Indicators of market advantage*	*Indicators of market constraints*
USA	684 714	19.2	3.4%/0.4%	China; India; South Korea; Japan; Canada; and Saudi Arabia	UK; Canada; Germany; France; and Australia	Prestige and quality of the US HE brand; diverse HE curricula and pathways; prestige of the US HE qualification in global labor markets; liberal sociocultural and political environment; English as a medium of instruction	The rise of the Asian HE market; negative global political image of the USA; reduced government spending on HE; increasing unemployment rates; target of acts of terrorism
UK	389 958	10.9	15.4%/0.9%	China; India; Nigeria; Ireland; Germany; and the USA	USA; France; Ireland; Australia; Germany; and Australia	Prestige and quality of the UK HE brand; prestige of the UK qualification in global labor markets; liberal sociocultural political environment; English as a medium of instruction	Increasing global competition, especially from the Asian HE sector; High fees structure for both domestic and international students; declining government support for HE; stringent immigration and employment laws for international students
Australia	271 231	7.6	21%/0.8%	China; India; Malaysia; Hong Kong; Indonesia; Singapore; and Vietnam	USA; New Zealand; UK; Germany; France; and Japan	Most developed HE in the Asia Pacific region; High quality of Education; focus on student experience and satisfaction; prospects in the employment sector; English as a medium of instruction	Stringent immigration and employment laws for immigrant students; regional differences in learning and teaching cultures; substandard housing for foreign students
France	259 935	7.3	11.5%/2.4%	Morocco; China; Algeria; Tunisia; Senegal; and Germany	UK; USA; Switzerland; Canada; Germany; and Belgium	Equal treatment of international and domestic students (i.e., same fee structure); the quality associated with French HE; the opportunity to learn French or extend knowledge in it; Vibrancy of contemporary French culture;	The language of instruction as a barrier;
Germany	200 862	5.6	Unavailable	China; Turkey; Russia; Poland; Austria; and Ukraine	Austria; Netherlands; UK; Switzerland; USA; and France	Nation with Europe's largest economy; equality between foreign and local students in terms of fees where these are paid; largely government subsidized HE with no upfront fees by all students; good employment opportunities; with BMW, Audi and Mercedes brands, the country is a favorite destination for engineering students especially	The language of instruction despite an increasing usage of English as a medium of instruction
Japan	141 599	4	3.7%/1.1%	China; south Korea; Vietnam; Thailand; Malaysia; and USA	USA; UK; Australia; Canada; and Germany	High standards of education; affordable tuition fees (almost half the cost of tuition in US universities; increasing use of English as a medium of instruction; strong and expanding economy with reputable technological brands; availability of part time jobs	The language of instruction despite an increasing use of English as a medium of instruction

Emerging Patterns of Growth in Global HE Markets of the South

The African HE market as a representative of the global south deserves space in this discussion largely because Africa still bears the scars of colonialism and perhaps more than any other continent is the best case study for understanding the contemporary dynamics of the postcolonial legacy in the development of HE.

Africa is the future of the world in terms of development. It is estimated that by 2050, Africa alone will be host to more than 60% of the world's reserves of untapped natural resources in the form of minerals, oil, and gas. Moreover, it will be the world's largest reservoir of young economically and educationally active population with approximately 25% of the entire world's young people aged 16–25 years residing on the continent. In addition, the UN Population forecasts for 2050 indicate that the largest population increases will take place in developing regions especially in Africa where the fertility rates remain higher than in developed nations and where efforts to deal with the ravages of HIV and Aids are showing significant positive impacts. This suggests that the demand for HE is likely to be very high on the continent and that a new scramble for Africa will be reinvented, both for its resources and for the growth of its intellectual talent. Current trends in the mobility of human talent in Africa suggest that this scenario will continue to accelerate both in the medium to long term. The following is a review from the Southern African Development Community (SADC) group of countries in Southern Africa.

Based on data from the OECD, the UNESCO Institute of Statistics (UIS), EUROSTAT, and the Higher Education Statistics Agency of South Africa (HESA), the following trends in student mobility markets in the SADC region have been noted:

- As a region, despite the growth of and supply for HE participation, the SADC nations have the highest outward mobility rates in the world at 6%, more than China and India.
- Despite the unprecedented growth in HE in the sector, the region has the lowest participation rates with only six of every 100 potential HE entrants being enrolled in the HE institutions of the region compared to, for example, 72% in Europe, and about 48% in Asia.
- Unlike their counterparts in other parts of Africa who prefer to migrate to the global north in search of HE opportunities, more than 70% of SADC students tend to seek their opportunities in the region and especially in South African institutions of HE.
- This internal regional mobility parallels that of the students in the EU the majority of who seek opportunities within more developed nations of the region in the UK, France, and Germany, for example. It also mirrors the SADC vision for greater regional integration and growth. This is supported by the Mwalimu Julius Nyerere African scholarship fund administered by the African Union (AU) for SADC students who want to study science and technology based degrees in African universities.
- The quality of HE in the region has also been noted to be on a downward decline largely due to inadequate funding and the migration of talent due to a wide variety of push factors (e.g., Pillay, 2010; Maringe and Carter, 2007).
- While academic mobility has positive outcomes in terms of talent development, there is evidence showing that large numbers of students from the sub-Saharan region do not return to their home countries following completion of study, contributing to a growing brain drain which has negative financial and quality implications for the region. For example, it has been estimated that almost $4billion is used to support approximately 100 000 western expatriates who are rolled in to replace the lost talent. Brain drain thus creates further employment opportunities for academics in the global north while reducing the growth of local indigenous talent to spearhead local development (e.g., Teichler and Yağcõ, 2009).
- Mauritius, South Africa, Botswana, and Zimbabwe, in that order have the highest outward bound mobility ratios in the region. Conversely, South Africa (ranks 11th as the most favored destination in the world), Mauritius and Botswana have the highest inbound ratios for international students. Reasons associated with preference for these countries include: country stability measures; economic development indicators; the quality of HE in those countries; and recognition of qualifications in the employment sectors of the region and for South Africa especially, being host to globally recognized universities which feature prominently in the top echelons of global performance tables (e.g., McGregor, 2007).
- The most preferred subjects of study by inbound students are Business Studies, Health Sciences, Social sciences, and law; Science, engineering, and Humanities in that order (HESA, 2009).
- Outside of Africa, SADC students also show preference for the UK and the USA due mainly to the reasons already explored elsewhere in this article, but chiefly because of past historical, quality, and language factors.

The North South Divide in HE Markets: Polarization, Unevenness, and Inequality

The evidence discussed here bears out the claim made in the introduction section that the HE terrains of the global north and south are increasingly becoming uneven, widening the quality and equity gap and increasingly diminishing the prospects for the knowledge bases of the south to take their place in the global knowledge hierarchies. This unevenness, polarization, and inequity are clearly demonstrated by the following:

- Net migration of talent from the global south to the global north metropolis.
- A relentless erosion of the human resources in the global south to grow the indigenous knowledge bases.
- Tendency to replace migrant talent with western talent likely to maintain the status quo in terms of a western presence and intellectual domination.
- In absolute terms, the majority of students who migrate to the global north universities for study do not return to their home nations following completion of study. It is therefore

logical to surmise that the flow of global talent is unevenly balanced in favor of the global north HE markets.

- Due to this northward bound migration, and also due to poor government support for HE at the behest of World Bank (WB) and IMF advice which favors higher investment at lower levels of schooling than at HE, the main pillars of quality in HE in the global south now stand on very shaky ground.
- The erosion of quality has fueled a down spiral which drives greater desire for migrating to the more secure institutions in the global north.
- The phenomenon of regional mobility based on the growth of regional institutions seeking greater regional integration at multiple levels is making some contribution to the stabilization of talent in the global south. However, with the growing tendency for these formations to be supported by western money, it is likely that the knowledge hegemonies of the west will continue to dominate the new partnerships formations of the global south.
- It has become almost impossible to elevate the status of indigenous languages as vehicles for HE learning in the global south because of the dominance of English both as lingua franca and as medium of instruction. The very basis for privileging indigenous knowledge systems in the global south has thus suffered a knockout blow in the race for cultural and intellectual domination.
- Other factors and conditions which contribute to poor economic performance of global south markets, such as civil wars, disease, and poor governance among others are increasingly hypothesized as being linked to hidden western agendas which see opportunities for continued control over resources of institutions and organizations in the global south. In 'How Europe underdeveloped Africa,' for example, Rodney (1972) showed that, despite shipping huge quantities of resources to Europe from Africa over five centuries, no meaningful development has occurred on the continent. In addition, the fabric of human consciousness, the African condition and culture was plucked away through systematic destruction of core cultural values and the adoption of western forms of knowledge and values to which the local ones were to either exist as the poor second cousin or be completely decimated from the arena of their lives.
- The dominance of the global north in the control of knowledge depositories and disseminators suggests that knowledge hierarchies based on the north will continue to dominate those in the global south.

Conclusions

Discourses of HE markets have tended to privilege global north perspectives and have not sufficiently engaged with post/anticolonial epistemic and oppositional perspectives which represent more accurately the development of HE in the global south.

The discourses in the global north tend to critique HE markets in the context of the commoditization/commodification in an increasingly consumer-led HE environment. Little space is given to postcolonial and anticolonial theory as lenses through which HE markets can be analyzed. Such analysis as shown in this article uncovers new understandings which help to bring the global south into the spotlight. Beyond the responsibility for guarding against the excessive commodification of HE, placing the global south into perspective uncovers a broader range of issues which may be increasing the global inequalities in HE. Mobility patterns in HE markets clearly demonstrate a strengthening and weakening of the global north and global south HE markets, respectively. This exacerbates the unevenness that characterizes markets in these places creating an issue of intellectual dominance in the way these markets operate.

This article deliberately does not offer solutions to these problems as this has not been the primary intention. However, if it has made any contribution, it is that the north–south divide, which western discourses generally silence, signals a source of tension beyond the conventional arguments about markets and the commodification of knowledge. Such tensions cannot be ignored any longer even though the solutions may lie beyond the HE sectors themselves.

See also: Educational Institutions, History of; Globalization in Education; Higher Education Management; Higher Education Research.

Bibliography

Anand, V.K., Tuesday, July 23 2013. Quantity outpaces quality in growth of universities. The Tribune Online Edition. http://www.tribuneindia.com/2013/20130723/edu.ht accessed 17 March 2014.
Appadurai, A., 2001. Deep democracy: urban governmentality and the horizon of politics. Environment and Urbanization 13, 23–43.
Appiah, K.A., 2004. Language, race, and the legacies of the British Empire. Black Experience and the Empire 387–408.
Apple, M.W., 1978. Ideology, reproduction, and educational reform. Comparative Education Review 22, 367–387.
Bartlett, W., Le Grand, J., 1993. The Theory of Quasi-Markets. Macmillan Press.
Bassett, R.M., 2006. The WTO and the University: Globalization, GATS, and American Higher Education. CRC Press.
Becker, G.S., 1993 [1964]. Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. University of Chicago Press, Chicago.
Bowles, S., Gintis, H., Osborne, M., 2001. The determinants of earnings: a behavioral approach. Journal of Economic Literature 39, 1137–1176.
Brint, S., Karabel, J., 1989. The Diverted Dream: Community Colleges and the Promise of Educational Opportunity in America, 1900–1985. ERIC.
Brown, R., Carasso, H., 2013. Everything for Sale? The Marketisation of UK Higher Education. Taylor & Francis.
Camila, V., 2012. Students Demand Education Not Segregation. Available from: universityworldnews.com/article.php 2013.
Carnoy, M., Castells, M., 1999. Sustainable Flexibility: Work, Family and Community in the Information Age. Cambridge University Press, New York.
Chakrabarty, D., 2000. Rethinking Working-Class History: Bengal, 1890–1940. Princeton University Press.
Chia-Ling, Kot, 2011.
Cubillo, J.M., Sánchez, J., Cerviño, J., 2006. International students' decision-making process. International Journal of Educational Management 20, 101–115.
Dahlman, C.J., Aubert, J.-E.C., 2008. The Knowledge Economy: Seizing the 21st Century. WBI Development Studies. World Bank Publications.
Dougherty, K.J., 1994. The Contradictory College: The Conflicting Origins, Impacts, and Futures of the Community College. SUNY Press.
Douglass, J.A., Edelstein, R., Hoareau, C., 2011. A GLOBAL TALENT MAGNET: How a San Francisco/Bay Area Higher Education Hub Could Advance California's Comparative Advantage in Attracting International Talent and Further Build US Economic Competitiveness. University of California, Berkeley.
Drucker, P.F., 1993. Managing in Turbulent Times. Routledge.
Escobar, A., 1995. Anthropology and the future: new technologies and the reinvention of culture. Futures 27, 409–421.

Escobar, A., 2005. Economics and the space of modernity: tales of market, production and labour. Cultural Studies 19, 139–175.
Foskett, N., Hemsley-Brown, J., 2001. Choosing Futures. RoutledgeFalmer, London.
Foskett, N., Maringe, F., 2010. Globalization and Internationalization in Higher Education: Theoretical, Strategic and Management Perspectives. Continuum, Great Britain.
Giroux, H.A., 1983. Theories of reproduction and resistance in the new sociology of education: a critical analysis. Harvard Educational Review 53, 257–293.
Gupta, D., Gupta, N., 2012. Higher education in India: structure, statistics and challenges. Journal of Education and Practice 3, 17–24.
Hargreaves, A., 2002. Teaching and betrayal. Teachers and Teaching: Theory and Practice 8, 393–407.
Harvey, D., 2005. A Brief History of Neoliberalism. Oxford University Press.
Harvey, D., 2010. The enigma of capital and the crisis this time; presented at the American Sociological Association meeting, Atlanta, August 16 2010 at http//davidharvey.org/2010/08/the enigma-of-capital-and-the-crisis-this-time (accessed 15.03.14.).
HESA, 2009. Students in Higher Education Institutions. HESA, Cheltenham.
HESA, 2012. Statistics – Staff Employed at UK Higher Education Institutions.
Ishengoma, J.M., 2011. Investing in Africa's high-level human resource: the challenges and paradoxes of implementing cost sharing. In: Setting of New Social Science Research Agendas for Africa in the 21st Century, p. 179.
Jenkins, R., 2013. Transnational Corporations and Uneven Development (RLE International Business): The Internationalization of Capital and the Third World. Routledge.
Jongbloed, B., 2003. Marketisation in higher education, Clark's triangle and the essential ingredients of markets. Higher Education Quarterly 57, 110–135.
Karabel, J., 1986. Community colleges and social stratification in the 1980s. New Directions for Community Colleges 13–30.
Kelo, M., Teichler, U., Wächter, B., 2006. EURODATA: Student Mobility in European Higher Education. Lemmens.
Keynes, J.M., 2006 [1883–1946]. General Theory of Employment, Interest and Money. Atlantic Books.
King, A., 1990. Architecture, capital and the globalization of culture. Theory, Culture and Society 7, 397–411.
Knight, J., 2004. Crossborder education in a trade environment: complexities and policy implications. In: The Implications of WTO/GATS for Higher Education in Africa, p. 59.
Knight, M., 2011. Strategic Review of the Student Visa Program. DIAC.
Le Grand, J., Bartlett, W. (Eds.), 1993. Quasi Markets and Social Policy. Macmillan Press, London.
Lin, J.Y., 2007. Creating a market-oriented research and education institution in a transition economy. Capacity Building 65.
Malthus, T.R., 1815. An Inquiry into the Nature and Progress of Rent, and the Principles by Which It Is Regulated. J. Murray, and J. Johnson and Company.
Maringe, F., Carter, S., 2007. International students' motivations for studying in UK HE: insights into the choice and decision making of African students. International Journal of Educational Management 21, 459–475.
Maringe, F., Foskett, N., Roberts, D., 2009. "I can survive on jam sandwiches for the next three years": the impact of the new fees regime on students' attitudes to HE and debt. International Journal of Educational Management 23, 145–160.
Maringe, F., Foskett, N., Woodfield, S., 2013. Emerging internationalisation models in an uneven global terrain: findings from a global survey. Compare: A Journal of Comparative and International Education 43, 9–36.
Maringe, F., Gibbs, P., 2009. Marketing Higher Education: Theory and Practice. McGraw-Hill International.
Marx, K., 1977. A Contribution to the Critique of Political Economy. Progress Publishers, Moscow.
Mazzarol, T., Soutar, G.N., 2002. "Push-pull" factors influencing international student destination choice. International Journal of Educational Management 16, 82–90.
Mcgregor, J.A., 2007. Researching Human Wellbeing: From Concepts to Methodology.
Mill, J.S., 1991[1812–1848]. Collected Works of John Stuart Mill: Indexes, vol. 33. University of Toronto Press.
Mincer, J.A., 1974. Schooling and Earnings. Schooling, Experience, and Earnings. Columbia University Press.
Mohamedbhai, G., 2013. Transnational Education: What Impact on Local Institutions? A Blog from the Center for International Higher Education (Online). Available from: insidehighered.com/blogs/world-view/transnational-education-what-impact-local-institutions.
Mpinganjira, M., Rugimbana, R., 2009. Understanding International Student Mobility: What Motivates African Student Choices? ANZMAC.
Mudimbe, V.Y., 1994. The Idea of Africa. Indiana University Press.
Nagler, A., 2007. The Influences of Institutional Reputation on the Labor Market Outcomes of Education and Training: A Case Study of Community College Nursing Programs. ProQuest.
NCES, 2012. The National Center for Education Statistics (Online). Washington, DC. Available: nces.ed.gov/ý (accessed 10.08.13.).
OECD, 2011. Education at a Glance 2011. OECD Publishing.
OECD, 2012. Education at a Glance 2012: Highlights. OECD Publishing.
OECD, 2013. Education Indicators in Focus. Organisation for Economic Co-operation and Development.
Parry, G., Callender, C., Scott, P., Temple, P., 2012. Understanding Higher Education in Further Education Colleges (BIS Research Paper Number 69). Department for Business, Innovation and Skills, London.
Pillay, P., 2010. Linking Higher Education and Economic Development: Implications for Africa from Three Successful Systems. African Minds.
Popovich, N.G., 2002. On being a faculty member. Journal of Pharmacy Teaching 10, 81–94.
Psacharopoulos, G., 1994. Returns to investment in education: a global update. World Development 22, 1325–1343.
Rapoza, K., 2011. By 2020, China No. 1, US No. 2. Forbes, 5/26/@ 11:18PM.
Richardson, J., 2002. European Union: Power and Policy-Making. Routledge.
Robert, B., Lajtha, C., 2002. A new approach to crisis management. Journal of Contingencies and Crisis Management 10, 181–191.
Rodney, W., 1972. How Europe Underdeveloped Africa. Bogle-L'Ouverture, London.
Rodney, W., 1983. How Europe Underdeveloped Africa. Bogle-L'Overture, London.
Samoff, J., 2004. From Funding Projects to Supporting Sectors. Observation on the Aid.
Samoff, J., Carrol, B., 2004. The promise of partnership and continuities of dependence: external support to higher education in Africa. African Studies Review 67–199.
Say, J.-B., [1803]1826. Traite D'economie Politique. Rapilly, Paris.
Schultz, T.W., 1961. Investment in human capital. The American Economic Review 51, 1–17.
Smith, A., [1776]2003. The Wealth of Nations. Bantam, New York.
Stoke, H.W., 1966. Emerging concepts in higher education. The Phi Delta Kappan 47, 402–405.
Teichler, U., Yağci, Y., 2009. Changing challenges of academic work: concepts and observations. In: Higher Education, Research and Innovation: Changing Dynamics, p. 85.
Tickly, L., 1999. Postcolonialism and comparative education. In: Soudien, C., Kallaway, P., Breier, M. (Eds.), Education, Equity and Transformation. Special double issue of International Review of Education.
TSLE, 2013. The Times Higher Education World University Rankings. Thomson Reuters, UK.
University World News, September 2012. Massification continues to transform higher education, Issue No. 237.
Välimaa, J., Hoffman, D., 2008. Knowledge society discourse and higher education. Higher Education 56, 265–285.
Wahrenburg, M., Weldi, M., 2007. Return on Investment in Higher Education–Evidence for Different Subjects, Degrees and Gender in Germany. University of Frankfurt.
Weis, L., 1985. Without dependence on welfare for life: black women in the community college. The Urban Review 17, 233–255.
Willis, E., 1983. Medical Dominance: The Division of Labour in Australian Health Care. George Allen & Unwin.
WSJ., 2013. Wall Street Journal (Online). India. Available: http://online.wsj.com/home-page (accessed 10.08.13.).
Young, R.J.C., 2001. White Mythologies, Writing History and the West. Routledge, London.
Zedillo, E., 2007. The Future of Globalization: Explorations in Light of Recent Turbulence. Routledge.
Zeleza, P.T., 2003. Rethinking Africa's Globalisation. In: The Intellectual Challenges, Vol. 1. Africa World Press, Trenton, NJ.
Zeleza, P.T., 2009. Diaspora dialogues: engagements between africa and its diasporas. In: The New African Diaspora, pp. 31–58.

Higher Education Research

Ulrich Teichler, International Centre of Higher Education Research Kassel (INCHER-Kassel), University of Kassel, Kassel, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Higher education research is a small field as compared to the size of the higher education system. As a theme-based field, it is expected to strive both for academic quality and practical relevance. Many scholars involved address this area only occasionally, and more research is linked to administrations or external stakeholders than based on academics. The borderlines are fuzzy between researchers and practitioners active in gathering information as a basis of reflection and improvement. It is a growing field, though uneven across countries, with a widening thematic scope: Notably, comparative studies got momentum and internationalization became a major theme.

Introduction

All efforts in the past to characterize the situation of higher education research have come to similar conclusions:

- It is a small field of research. When looking at the substantial size of higher education systems all over the world and their important role in society, we have to come to the conclusion that systematic knowledge on higher education is not held in high esteem even today.
- It is a theme-based field that is expected on the one hand to strive for quality as far as theories, methods, and field knowledge are concerned as the established disciplines and on the other hand to legitimize its existence through being highly relevant for higher education policy and practice.
- Most scholars undertaking – occasionally or frequently – research in this area do not consider themselves to be higher education researchers, but rather as belonging to a certain discipline: education, psychology, sociology, political science, economics and business studies, law, history, etc. For them, higher education is just a thematic area, not a core element of their own academic identity.
- Higher education is a field of expertise where the borderlines are very fuzzy between research on the one hand and consultancy, evaluation, and other search for evidence on the other hand. The borderlines are fuzzy as well regarding the respective persons: researchers on the one hand and on the other hand consultants, evaluators, and 'reflective' policy makers and practitioners. For example, most journals in this domain provide a platform for both scholars and other 'experts.'
- Most higher education research are national in focus or concentrate on individual institutions or their subunits. Research with a worldwide focus or international comparative research in this field has been by no means frequent in the past. In recent years, however, a broader perspective is on the way: Internationalization of higher education has become a major theme of analysis, comparative studies gained momentum, and we note an increase of international cooperation and communication among the respective researchers.
- The conditions for this field of research vary substantially across countries. There are enormous differences in the size of this field: In some countries, higher education research is a noteworthy field (e.g., in the United States of America and China), while in other countries it is marginal or nonexisting. In some countries, we note a clear dominance of one or two types of institutional bases, e.g., notably faculties of education and 'institutional research,' while in other countries research is widely scattered.

We should add that it is a growing field: Developments such as the growth of the higher education system, the increasing strategic orientation of higher education management and policy, the growing emphasis on evaluation- and evidence-based action, the professionalization trends within higher education, and the widening of communication beyond regional and national horizons call for improved basis of systematic knowledge.

The aim of this article is to provide an overview primarily on the setting of higher education research: The characteristics of a theme-based field, the diversity of institutional bases and persons involved, and the thematic range. This will be intertwined, however, with glances at the substance and major findings of this field.

A Theme-Focused and Practically Relevant Field

The respective key literature underscores that higher education research is a theme-focused area of research: It addresses higher education (or tertiary education), its functions, institutional settings, and social context. Here, much is in common with other areas of behavioral and social sciences that gained attention and support for its thematic focus. Actually, some activities in this domain touch upon neighbor areas of thematic research, e.g., science of science, youth research, labor market research, or research on competences.

Higher education research does not have the typical basis of legitimacy as classical disciplines within universities and other institutions of higher education. It is not viewed as having emerged from the logic of the knowledge system, but rather from practical demands for knowledge. As a consequence, it has to find ways constantly to be both, respectable as a discipline in theoretical depth, methodological quality, and breadth of field knowledge and as a provider of a valuable knowledge base for policy and practice.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92120-2

Higher education research draws from a multitude of disciplines, notably education, psychology, sociology, political science, economics and business studies, law, and history.

It is linked to many interdisciplinary areas as well. For example, Clark (1984) included in his overview on research fields 'the political view,' 'the organizational conception,' and the 'the policy perspective.' In 1992, an encyclopedia on higher education names among 'disciplinary perspectives on higher education' also 'anthropology,' 'linguistics and rhetorical studies,' 'literature,' 'science studies,' and 'women's studies' (Clark and Neave, 1992).

Actually, various key thematic areas of higher education gain from ties to a variety of disciplines or interdisciplinary areas; e.g., research on the relationships between higher education and the world of work benefits from the knowledge of economics, sociology, and psychology as well as from interdisciplinary research on labor markets, professions, personnel management, competences, etc.

Higher education research, as many thematically defined fields as well, is disadvantaged vis-à-vis typical disciplines in various respects:

- The institutional basis within the academic domain of universities often is feeble and unstable.
- There are widespread claims that the theoretical and methodological advancements of behavioral and social sciences are stronger, if they are discipline-based, and that multidisciplinary areas just borrow from the disciplines and lack academic rigor.
- Often, theme-based fields do not have a visible educational role in terms of separate study programs and doctoral training activities. This makes the field vulnerable as far as institutional stability and 'self-reproduction' of future generation of scholars are concerned.

The development of higher education research is clearly supported by practical needs. A first trend report published in the early 1970s (Nitsch and Weller, 1970–1973) showed that it had been a marginal field in most countries of the world until the 1960s. It began to grow as a consequence of three developments: First, the growing belief fueled by concepts of economics education that an increase of postcompulsory education was essential for economic growth; second, a spreading conviction – reinforced by the student protests in the 1960s – that higher education was not functioning properly through incremental self-regulation and improvements, but rather faced substantial features of crisis that needed strategic responses; and third, the actual rapid expansion of higher education called many traditions into question.

The strong impact of policy and practice on higher education research is mirrored in the fact that the majority of research activities in this domain are undertaken outside academia. Moreover, research undertaken in the academic domain often is funded for practical purposes, and changes quickly thematic priorities in tune with the public discourse: Higher education research is likely to grow in areas, when respective problem awareness surface and reform efforts gain momentum, and is likely to grow overall, when sweeping broad-ranging reforms are underway, e.g., in Europe early in the twenty-first century in the wake of the so-called 'Bologna Process.'

The practical interest in knowledge about higher education, however, does not turn out to assure consistently a solid support for research in this domain:

- Academics and other representatives of higher education tend to believe generally that progress and innovation depend on the generation and interpretation of systematic knowledge. When it comes to higher education policy and practice as such, however, the belief is widespread, however, that firsthand experience and reflection on the part of the knowledgeable actors themselves are sufficient basis for wise strategic action.
- Policy makers and practitioners often advocate a tailor-made information gathering expected to serve decisions ('evidence-based' policy based on 'policy-based' research). As a consequence, many of the respective analyses lack a thorough theoretical basis and methodological soundness as well as a critical questioning of conventional wisdom.
- The rapid pace of changing political issues causes difficulties for higher education research in serving the reflection of these issues. Often, information is lacking when an issue surfaces; as a consequence, often poorly founded systematic information is put before. In turn, support for research often fades away when the respective theme is not anymore in the limelight of public debate, even though systematic information remains relevant as a basis of continuous improvement.

Certainly, higher education research constantly has to find ways to strike a balance between theoretical, methodological, and field knowledge consolidation like any discipline and to serve the needs of problem identification and problem solving like any applied field. The growing popularity of terms, such as 'mode 2 research,' 'targeted research,' etc., indicates that this is a typical challenge for many areas in the behavioral and social sciences.

The Variety of Institutional Bases and Roles of Those Undertaking Research

Trend reports and other overviews (Nitsch and Weller, 1970–1973; Fulton, 1992; Sadlak and Altbach, 1997; Teichler and Sadlak, 2000; Teichler, 2005; Altbach, 2009; Davis et al., 2010) show three striking features of higher education research as regards the institutional basis and the persons involved.

- An enormous heterogeneity according to country: While higher education research is fairly widespread in some countries, it is almost nonexistent in others.
- A diversity of institutional bases, whereby more research seem to be undertaken outside the typical domain of academic units at higher education institutions.
- A wide range of persons involved – from full-time research to occasional involvement, from an emphasis on academic values to practice- and service-oriented information delivery.

As regards country, the US is often been named as the one with the largest number of persons and research activities in the field of higher education. Actually, however, the respective

absolute figures in China seem to have become even higher in the wake of recent educational expansion. Relative to the population, the size of higher education as a field of research is more impressive in some small countries, e.g., Australia and Finland. In contrast, research in this domain is scarce in many southern European countries as well as in many low- and middle-income countries, even some improvement can be observed there as well (see, e.g., Brunner, 2009; Horta and Jung, 2013).

The scarcity of higher education research in low- and middle-income countries can be demonstrated by an example. In the context of a major project undertaken in the first decade of the twenty-first century under the umbrella of UNESCO, issues of higher education, research, and innovation were analyzed worldwide. While the project coordinators succeeded in forming a Europe–North America committee composed almost exclusively of respective researchers, policy makers, and practitioners clearly dominated as a consequence of scarce research potentials in the other regional committees (Meek et al., 2009).

As regards institutional basis, analyses available (cf Schwarz and Teichler, 2000) opted for different classifications, because not a single classification is commonly agreed among the respective experts and the ways of institutionalization vary from country to country.

El-Khawas (2000) identified three separate bases of higher education research in the US:

- 'Research': Academically based higher education research often is institutionalized in university units of teaching and research. Academics strive for a strong theoretical basis and the pursuit of knowledge for its own sake.
- 'Policy': Higher education policy research or information units are established by or linked to governments or other suprainstitutional actors. They aim at enriching the policy process through information and policy-driven interpretations.
- 'Practice': Various universities establish research and development units linked to the institutional management. This 'institutional research' is expected to provide information and action-oriented interpretations for the actors within the individual institutions – similarly as policy research serves macrosocietal actors.

Teichler (2005) names six types of institutional bases:

- Higher education research alimented by teaching, i.e., visibly institutionalized in units in charge of teaching and research – in some countries as a branch field in faculties of education. Notably in the US, a substantial number of master and doctoral programs focus on higher education and are expected to serve both the reproduction of scholars and the training professionals in charge of management support and services at higher education institutions: "The larger and more research-oriented programs include the University of California at Los Angeles, the University of Southern California, Pennsylvania State University, the University of Michigan, Michigan State University, the University of Pennsylvania, and several others" (Altbach et al., 2007: p. 6). Similarly, institutes of higher education with a relatively long tradition in doctoral education in China are those at Xia Men University, Peking University, East China Normal University, and Huazhong University of Science and Technology. Similarly, the Centre for Higher Education at the Institute of Education, University of London, might be named for the UK.
- University-based centers for higher education research: Some universities have opted for separate research units that can draw from a broad range of disciplines, concentrate fully on research, and aim to strike a balance between the quality criteria of academic research and relevance for policy and practice: e.g., the Research Institute for Higher Education of the Hiroshima University (Japan), the International Centre for Higher Education Research of the University of Kassel (Germany), the Center for Higher Education Policy Studies of the Twente University (the Netherlands), and the Center for the Study of Higher Education of the University of Melbourne (Australia).
- Units for research and practice, combined at institutions of higher education, e.g., in the area of 'staff development,' 'teaching and learning,' or 'didactics of higher education.' They are expected to train teaching staff, assist curricular innovation, etc., and, concurrently, to conduct research to assure a sound knowledge base in their area of practice.
- Institutional research – a term having emerged in the US for research units within the administration of individual higher education institutions: They collect systematic information as a basis for the reflection of institutional policies. This idea spread worldwide – however with different connotations. For example, the European Association for Institutional Research (EAIR) did hardly become a training ground for similar institutionalization, but rather served the dialogue between higher education researchers and higher education policy makers and practice (see Begg, 2003).
- Nationally sponsored research institutes for higher education: In various countries, institutes or other units were established outside higher education institutions. They might be sponsored by governments, foundations, or other external stakeholders. A typical example is the Nordic Institute for Studies in Innovation, Research and Education in Oslo (Norway), or a large research institute in Hannover (Germany), recently renamed German Institute for Higher Education Research.
- Higher education research as subunit or subfield of an institute with a broader thematic range: e.g., in the Institute de Recherche sur l'Economie de l'Education, Université de Bourgogne in Dijon (France).

One might add as a newly spreading mode institutes jointly supported by several universities, e.g., the Centre for Research in Higher Education Policies in Portugal and the National Higher Education Research Institute in Malaysia.

In looking at the persons involved in research on higher education, we can observe an even broader range. In addition to researchers at the various types of institutions named above, we note on the one hand a significant number of scholars of various disciplines not considering themselves to be higher education researchers, but undertake research in this domain occasionally. On the other hand, many policy makers and practitioners only embark occasionally in activities that can be

named research on higher education (Teichler, 2005; cf also Teichler, 1996) suggested grouping persons involved six categories. They are:

- The discipline-department-based occasional researchers on higher education: They devote part of their time or stages of their career to the thematic area, but see their identity in a discipline and consider research on a variety of thematic areas as natural. They contribute to theoretical and methodological enhancement of research in this field, but often lack the breadth of the field knowledge of higher education. Well-known examples of analyses based on such a background have been 'The Organization of Academic Work' (Blau, 1973) and 'The American University' (Parsons and Platt, 1983).
- The discipline-based continuous higher education researchers: Being based in disciplinary units they devote most of their research to this thematic field. They might be strong in theory, methodology, and field knowledge, but they often care little about higher education policy and practice. For example, references in major encyclopedias suggest Clark (1983), Bourdieu (Bourdieu and Passeron, 1997; Bourdieu, 1984), as well as Becher and Kogan, (1990) had a substantial impact on the conceptual discourses on higher education.
- The scholars based in higher education research units: They often excel as regards field knowledge, they try to strike a balance between the quality criteria of academia and relevance expectations of policy and practice, and they frequently draw from various disciplinary areas. Again a look at encyclopedias and overview books suggests that Trow (1973) and, subsequently, Geiger (1986), Teichler (1988, 2009) as well as Neave and van Vught (1991), and others played a role in agenda setting.
- The applied researchers: These are persons active in units of 'institutional research' or similarly, in institutes located outside higher education institutions (e.g., government-sponsored research institutes). The prime emphasis is to collect systematic information relevant as 'evidence' for decision making.
- The consultants (including evaluators) could be viewed as a mix between the applied researchers and the reflective practitioners. They draw from all kinds of systematic knowledge or practical experience and generate here and there systematic knowledge to advise practitioners whereby the fit of information and advice is a major concern.
- The reflective, occasionally researching practitioners: These might be managers (e.g., university presidents), 'higher education professionals' (experts in charge of service or management support functions), experts in special agencies (e.g., evaluation and accreditation agencies), etc., for whom secondary analysis of research, systematic reflection, and occasional studies might be part of their role and professional self-understanding.

The breadth of institutional bases and of persons with different professional roles involved in higher education research certainly can be viewed as a wealth for the overall knowledge base of the research field. However, most authors of overview reports perceive higher education research as fragmented, because most institutions and persons involved communicate and cooperate or generally pay attention primarily to their respective peers rather than to the field of higher education research as a whole.

The Thematic Range

Overviews on higher education research present varied classifications as well of the themes addressed. Some examples might be presented here (cf Teichler, 2005; Horta and Jung, 2013):

- Some analyses underscore a divide between research addressing the educational domain (teaching and learning, curricula, competences, etc.) and the policy domain (organization, management, funding, etc.): Separate 'continents,' 'archipelagos,' or similarly (see Macfarlane, 2012);
- Teichler (1996) notes four distinct areas of expertise among higher education researchers: quantitative-structural, knowledge- and subject-related, person-related as well as teaching- and research-related, and finally organization- and governance-related aspects.
- The major encyclopedia (Clark and Neave, 1992) is divided into eight areas: national systems of higher education; higher education and society; the institutional fabric of higher education; governance, administration, and finance; faculty and students; teaching, learning, and research; disciplinary perspective on higher education; and academic disciplines.
- Tight (2003) presented eight different areas in academic publications: teaching and learning, course design, student experience, quality, system policy, institutional management, academic work, and knowledge and research.
- Finally, the journal *Research into Higher Education* Abstracts names eight categories as well: national systems and comparative studies, institutional management, curriculum, research, students, staff, finance and physical resources, and contributory studies and research approaches.

These classifications aim to be useful across countries and over time. This does not mean, however, that research in this domain is similarly spread across countries or over time.

Thematic priorities by country certainly reflect major issues of higher education policy and practice. For example, macro-societal issues seem to play a stronger role in countries where governments play a strong steering role. Research is also influenced by its institutional setting: Educational issues of higher education are more likely priority areas in countries in which academic higher education research is predominantly located in faculties of education. However, many themes are popular across countries – may it be as a consequence of similar problems or as a consequence of the activities of international organization to stimulate international debates on common issues.

Thematic priorities seem to change over time along policy priorities. Overview reports point out that certain themes that were in the limelight of the public discourse at a certain period, also drew attention of higher education researchers at that time and were similarly popular across economically advanced countries and to some extent as well across other countries. Certain thematic priorities are frequently named.

- In the 1960s, higher education expansion, the relationships between expansion and economic growth as well as issues of equality of opportunity, etc.
- In the 1970s, students, teaching and learning, institutional diversification, higher education planning, graduate employment, etc.
- In the 1980s, evaluation, management, quality of research, etc.
- In the 1990s, system governance and institutional management, incentive steering, internationalization, etc.
- In the twenty-first century, globalization, 'rankings,' 'employability,' 'learning outcomes,' 'knowledge society'/ 'knowledge economy,' competition and market steering, professionalization, etc.

No thorough overview reports are available on higher education research in middle-income and low-income countries. Certainly, in addition to the themes named above, policy agendas of the World Bank and to some extent of UNESCO play a role: education and income, privatization of higher education, the role of tuition fees, financial matters, and finally the role of higher education in economic and social development.

Bibliographic analyses have been employed to get a more precise overview on the thematic development of higher education research. The value of such analyses, however, is limited.

First, the publication strategies and activities are heterogeneous among the institutions and persons involved. The academically based researchers on higher education spread their key publication across various types of publications: disciplinary journals and books; specialized publications on higher education; publications serving both, academics as well as policy makers and practitioners; sources published by governments, international organizations, and other stakeholders; and publications specifically addressing policy makers and practitioners. Some applied researchers and practitioners occasionally involved might choose the same sources; however, many of their texts are available as institutional reports, 'gray literature' or are viewed as 'internal' information.

Second, many publications in this domain do not fit into divides that prevail in bibliographic analyses. For example, only few journals on higher education can be defined as purely academic. Among those in the English language with an international authorship and readership, Higher Education and Studies in Higher Education (the highest rated ones by the European Science Foundation) can be named genuinely academic. Other international journals, in contrast, have fuzzy borderlines between research and policy/practice; they at least occasionally publish texts written by policy makers and practitioners (e.g., Higher Education Policy, Tertiary Education and Management, Journal of Studies in International Education, or European Journal of Higher Education).

Third, bibliographic studies on the international scene of higher education research mostly focus exclusively on publications in the English language. In most countries, however, higher education research has kept a strong national focus, and publications in the native language dominate: They certainly comprise nine-tenths among the major publications in China, France, Germany, Italy, Japan, Spain, and Central and Eastern European countries. As a consequence, e.g., an analysis of higher education research in Asia based on journals in the English language names almost twice as many articles from Hong Kong, where English dominates, as from Japan and five times as many as from Korea (Jung und Horta, 2013). In Latin America, the common dominant use of the Spanish language obviously facilitates regional comparisons. Only higher education researchers from a few small countries, e.g., Finland, the Netherlands, and Norway, publish so much in English that publications in English provide a sufficient insight into the national scene.

This language bias is further reinforced by the fact that some of these analyses are based on journals in the English language among which only some can be viewed as genuinely international while others primarily cater scholars from a single or a few Anglo-Saxon countries. As a consequence of such a mix of sources, Tight (2003) notes that as many as one-third of authors in 'specialist higher education academic journals' published outside North America are British; according to the numbers of authors frequently named in the references, Tight (2012) seems even to suggest that half of the leading higher education researchers are from the UK.

There is a wealth of documents on the state of higher education research, e.g., publications on the activities of international associations – such as the Consortium of Higher Education Researchers (CHER) (Kehm and Musselin, 2013) and EAIR – as well as reports on the national scene of higher education research – e.g., on Canada (Jones, 2012), Germany (Gunkel et al., 2003), and Japan (Kôtô kyôiku kenkyû no seidoka to kadai tokushû, 2013), but secondary analyses would be needed in order to draw conclusions of trends as far as the institutional basis, the types of persons involved as well as the concepts, methods, and themes are concerned.

Finally, one feature that is hardly discussed in the analyses of the state of research should not be overlooked: As a rule, higher education research does not discuss the substance of the knowledge system, e.g., that of physics or youth research, and higher education researchers, as a rule, are no experts on the substance of the knowledge system beyond their own disciplinary background. One could argue that higher education research focuses on the nonessential issues of the knowledge system, i.e., on the framework in which knowledge operates, and is incomplete by definition. If it wants to be complete, collaboration with specialists of the respective disciplines is needed.

The Growing Supranational Reference of Higher Education Research

Higher education has a strong national focus, and many studies pay attention to a few institutions or even single cases, even though it is viewed as surpassing borders in many respects as a matter of procedure: Knowledge is universal in many disciplines as well as regarding the logic of science. Academics like to cooperate internationally, international academic reputation is viewed as the most desirable, and cosmopolitan values are widespread in academia. However, higher education is shaped nationally, regionally, or even locally as regards supervision,

funding, organizations, personnel policies, institutional patterns and diversity, curricula, and its various links to society.

In three respects, research on higher education clearly surpasses national frameworks. And its supranational reference has increased substantially in recent years.

First, international comparative analyses (see Teichler, 1996; Välimaa, 2008; Kosmützky and Nokkola, 2014) play an increasing role. Information collection across countries is often characterized as a 'gold mine' to free oneself from believing that features of higher education in the vicinity are typical and rational for higher education per se. Macrosocietal phenomena, such as institutional or program diversity, cannot be analyzed without international comparison. Moreover, supranational perspectives play a key role in the link between higher education research and policy: We observe a constant search for better solutions, and benchmarking is the basic methodology. Supranational organizations played a significant role for the establishment and improvement of higher education research since about the 1960s. While many experts believe to grasp higher education in their home country without a systematic research basis, understanding the international variety of higher education obviously does not seem to be possible without gathering information systematically. Also supranational agencies, such as the OECD, the World Bank, UNESCO, the Council of Europe, or the European Community, played a major role in stimulating comparative studies and discourses of the most 'modern' or the most 'successful' models of higher education (see Sadlak and Hüfner, 2002).

Second, higher education research increasingly addresses border-crossing phenomena. The trends of 'internationalization' and 'globalization' of higher education became major themes since about the 1990s, and they are paid more and more attention every year. Frequent themes are 'mobility of students and academic staff,' 'mutual influences of higher education systems on each other,' 'internationalization of the substance of teaching, learning, and research,' 'institutional strategies of internationalization,' 'knowledge transfer,' 'cooperation and competition,' and 'national and supranational policies regarding the international dimension of higher education' (Kehm und Teichler, 2007). In recent years, the term 'globalization' got momentum – i.e., a term implying a declining relevance of borders amid increasing border-crossing activities. For example, the individual universities are increasingly viewed as acting primarily on a global arena, whereby the popular 'rankings' of 'world-class universities' are often referred to as the most visible phenomenon of this kind.

Third, communication and collaboration among researchers and other experts on higher education have grown substantially. This helps overcome the limitations of conceptual frameworks shaped by nationally bound experiences, strengthen the knowledge basis of comparative studies as well as do genuine comparative analyses through the development of joint conceptual and methodological frameworks, and make the field of higher education research, often relatively small within individual countries, more visible.

Supranational organizations built up expert groups from different countries and stimulated international expert studies. Higher education researchers themselves were also active in establishing international associations. In Europe, e.g., the – British-based – Society for Research into Higher Education, established in the 1960s, was internationally active from the outset. Cooperation spread in the framework of supranational associations, e.g., the European Association for Research and Development of Higher Education, founded in the 1970s and later substituted by the International Consortium for Educational Development. EAIR emerged in the late 1970s as a branch of the US-based AIR, but became an independent association stimulating communication between scholars and practitioners. CHER was founded in the late 1980s to strengthen the quality and the comparative thrusts of higher education research. Though these associations themselves are not carriers of comparative research, they certainly contributed to a growing number of international research teams undertaking jointly comparative studies, e.g., on the student life and study, graduate employment and work, the academic profession, governance and management of higher education, etc. (cf Brennan and Teichler, 2008). The growing strength of comparative analysis can be demonstrated, e.g., through a comparison of the major encyclopedias on higher education (Clark and Neave, 1992; Forest and Altbach, 2006).

The increasing international communication among higher education researchers is linked with a more widespread use of English as the Lingua Franca. For example, the most highly reputed international academic journal of this domain, i.e., Higher Education, still had 63% of its authors between 1993 and 1997 from the four major Anglo-Saxon countries (21% from the UK), 13% from other European countries, and 24% from other countries or jointly from different countries. In the years 2001–04, already more than half of the articles were from other countries than the four Anglo-Saxon countries, among them 29% were from other European countries, while only 12% were from the UK (Teichler, 2005), and this trend toward a broader international coverage has gone further in the mean time. Yet, most publications on higher education remain to be published in home-country languages. The spread of the English language hitherto has led to less concern of collecting information with the help of different languages, and information on higher education available in the English language remains enormously uneven.

Higher Education – A Small Academic Field and Abundant Knowledge Production by Diverse 'Experts'

If the field of higher education is viewed merely as an academic one, we might find a few thousand academics all over the world, for whom this thematic area is the core of research or one of the major themes and for whom research in this area is the source of professional identity. Their neighbors are scholars either considering themselves as representatives of disciplines, who address higher education occasionally, or as specialists of neighbor thematic research areas, e.g., science of science or sociology of the professions. This academic field can be described, as other academic areas, with reference to specialized institutional settings, research (and possibly teaching) activities, academic associations, and academic publications.

Of course, academically based higher education researchers are linked to the real world of higher education by analyzing its current state and changes; among others, they try to predict future developments of higher education in order to identify challenges for research on higher education (see Brennan and Teichler, 2008).

But higher education research as a field with a thematic focus is not legitimized merely by its inner academic life, but concurrently by its delivery of relevant systematic information that helps policy makers and practitioners to reflect the state of affairs in higher education and generate ideas for improvement. Such reflections, however, draw from research or other activities of generating knowledge more or less systematically on the part of hundreds or thousands of persons. Some clearly are researchers as well: In units linked to the administration of individual institutions of higher education (often called 'institutional research'), state institutes, and units linked to external stakeholders, consulting agencies, etc. The total number of higher education researchers in those settings certainly is larger than those in academic settings. Moreover, there are large numbers of higher education professionals involved in systematic data gathering – evaluators, international officers, and program coordinators who undertake surveys of the experiences of the persons for whom they have to take care in their practical assignments, etc. Moreover, there are many university leaders, academics of various disciplines, consultants, or other experts who embark in systematic knowledge production in one way or other.

The borderlines between respectable systematic and not so systematic information, between academically sound and not well-based information, between information on the one hand and advertisement and propaganda on the other are fluent. In some way or other, all these researchers and other experts compete in gathering and distributing information to those in charge of higher education. Two examples might illustrate this.

First, prior to the 2012 follow-up ministerial conference in the framework of the Bologna Process, the hosting Romanian government and others responsible to coordinate the activities between the conferences invited more than 60 experts to summarize the available information relevant for the overall assessment of the Bologna Process (see Curaij et al., 2012). According to those responsible, this activity "aimed at bringing the researchers' voice into higher education international level policy making," and it gave "an unprecedented opportunity for researchers dealing with higher education matters to interact and contribute to the political process shaping the European Higher Education Area, as well as national policy agendas" (Curaij et al., 2012: p. v). Among the actual authors of this collection of analysis calling for a high level of systematic knowledge, less than one-third can be viewed as typical higher education researchers, while a higher proportion of them are scholars who at most occasionally address higher education, and an almost as high proportion are experts in various policy and coordination agencies.

Second, the 'rankings' of 'world-class universities' are a politically very powerful tool of systematic information in the form of indicators on the 'quality' of universities. They are the product of 'experts.' Some of the higher education researchers with an academic base are directly involved in the production of the rankings, others embark on in-depth discussions of methodological improvement while accepting the political steering role of the rankings, other develop alternative measures, and finally others criticize both, the concepts and methods as well as their ideological underpinnings. This involvement of academics in various respects indicates that higher education research is not an academic field that can live more or less independently from other somewhat systematic knowledge production on higher education.

Thus, higher education as an academic field of research is not clearly separated and embedded in a multitude of other scholars and experts. It constantly has to strike a balance between academic quality and relevance as well as between strengthening its own academic territory and playing a convincing role for policy and practice among the multitude of knowledge producers.

Major Findings

Higher education research covers such a multitude of issues that one can hardly sketch its major findings. The growth of student numbers certainly has been consistently viewed as caused by rising features of educational meritocracy as thus by a growing demand of individuals, but it remained to be viewed as controversial as regards the human resource demands of the world of work: Notions of 'overeducation' competed with claims that societal modernization and economic growth call for an expansion of higher education. Higher education is viewed as serving society through intellectual enhancement, cultural development, stimulation of technological innovation, provision of trained human resources, and contribution to critical thinking, but neither the respective needs of society nor the contributions of higher education are predictable: Most studies are based on the assumption that some kind of a balance is needed between academic freedom to ensure room for unpredictable developments of the knowledge and soft pressures to safeguard that relevance is on the agenda. Many studies as well identify a vulnerable balance between academia being self-controlled by the experts of systematic knowledge on the one hand and on the other the demands of relevance and efficiency: As a consequence, many studies addressed the strength and weaknesses of steering on the part of various actors, e.g., academics, governments, university management, and external 'stakeholders,' or various mechanisms, e.g., legislation, evaluation, and incentive steering. Most recent research on higher education notes a growing exposure of higher education to an increasing range of diverse expectations in the wake of what is called 'knowledge society' as well as increasing weight of border-crossing activities and even crumbling of borders toward a world of higher education system and a global society. Consequences for higher education as such hardly can be identified anymore because a growing diversity hardly leaves any common features. The preponderance of search for improved organizational solutions obviously calls for more integrated views on higher education: "How do changes in the organisation of higher education institutions relate to changes in intellectual programmes, agendas and advances?" (Brennan, 2008: p. 192)

Bibliography

Altbach, P.G., 2009. Higher education: an emerging field of research and policy. In: Bessett, R.M., Maldonado-Maldonado, A. (Eds.), International Organizations and Higher Education Policy. Routledge, New York, pp. 9–26.

Altbach, P.G., Bozeman, L.A., Janashia, N., Rumbley, L.A. (Eds.), 2007. Higher Education: A Worldwide Inventory of Centers and Programs. Sense Publishers, Rotterdam.

Becher, T., Kogan, M., 1990. Process and Structure of Higher Education. Routledge, London.

Begg, R. (Ed.), 2003. The Dialogue between Higher Education Research and Practice. Kluwer Academic Publishers, Dordrecht.

Blau, P.M., 1973. The Organization of Academic Work. Wiley, New York.

Bourdieu, P., 1984. Homo Academicus. Les Editions de Minuit, Paris, 1988. Polity Press, Cambrigde.

Bourdieu, P., Passeron, J.C., 1977. Reproduction in Education, Society and Culture. Sage, New York.

Brennan, J., 2008. Higher education and social change. Higher Education 56 (3), 381–393.

Brennan, J., Teichler, U. (Eds.), 2008. Special Issue: The Future of Higher Education and the Future of Higher Education Research. Higher Education 56 (3).

Brunner, J.J., 2009. Sociologia de educacao superior nos contextos internacional, regional e local. Dados Revista de Ciencias Sociais 52 (3), 621–658.

Clark, B.R., 1983. The Higher Education System. University of California Press, Berkeley, CA.

Clark, B.R. (Ed.), 1984. Perspectives on Higher Education: Eight Disciplinary and Comparative Views. University of California Press, Berkeley, CA.

Curaj, A., Scott, P., Vlasceanu, L., Wilson, L. (Eds.), 2012. European Higher Education at the Crossroads: Between the Bologna Process and National Reforms, vol. 2. Springer, Dordrecht.

Davis, W.M., Devlin, M., Tight, M. (Eds.), 2010. Interdisciplinary Higher Education: Perspectives and Practices. Emerald, Bingley.

El-Khawas, E., 2000. Research, policy and practice: assessing their actual and potential linkages. In: Teichler, U., Sadlak, J. (Eds.), Higher Education Research: Its Relationships to Policy and Practice. Pergamon press, Oxford, pp. 37–44.

Forest, J.F., Altbach, P.G. (Eds.), 2006. International handbook of higher education, vol. 2. Springer, Dordrecht.

Fulton, O., 1992. Higher education studies. In: Clark, B.R., Neave, G.R. (Eds.), The Encyclopedia of Higher Education, vol. 4. Pregamon Press, Oxford, pp. 1810–1821.

Geiger, R., 1986. Private Sectors in Higher Education: Structure, Function and Change in Eight Countries. University of Michigan Press, Ann Arbor, MI.

Gunkel, S., Freidank, G., Teichler, U. (Eds.), 2003. Directory der Hochschulforschung (Directory of higher education research). Hochschulrektorenkonferenz, Bonn.

Horta, H., Jung, J., 2013. Higher education research in Asia: an archipelago, two continents or merely atomization? Higher Education 67.

Jones, G.A., 2012. Reflections on the evolution of higher education as a field of study in Canada. Higher Education Research and Development 31 (5), 711–722.

Jung, J., Horta, H., 2013. Higher education research in Asia: a publication and co-publication analysis. Higher Education Quarterly 67 (4), 398–419.

Kehm, B.M., Musselin, C. (Eds.), 2013. The Development of Higher Education Research in Europe: 25 Years of CHER. Sense Publishers, Rotterdam.

Kehm, B.M., Teichler, U., 2007. Research on internationalization in higher education. Journal of Studies in International Education 11 (3–4), 260–273.

Kosmützky, A., Nokkola, T. (Eds.), 2014. Special Issue on Challenges and Trends in Comparative Higher Education. Higher Education 67.

Kôtô kyôiku kenkyû no seidoka to kadai tokushû (Special issue: higher education research: institutionalization and its challenges). Kôtôkyôikukenkyû (Japanese Journal of Higher Education Research) 16, 2013.

Macfarlane, B., 2012. The higher education research archipelago. Higher Education Research and Development 31 (1), 129–131.

Meek, V.L., Teichler, U., Kearney, M.-L. (Eds.), 2009. Higher Education, Research and Innovation: Changing Dynamics. INCHER-Kassel, Kassel.

Neave, G., van Vught, F. (Eds.), 1991. Prometheus Bound: The Changing Relationship between Government and Higher Education in Western Europe. Pergamon Press, Oxford.

Nitsch, W., Weller, W., 1970–1973. Social Science Research on Higher Education and Universities. Mouton, The Hague and Paris.

Parsons, T., Platt, G.M., 1983. The American University. Harvard University Press, Cambridge, MA.

Sadlak, J., Altbach, P.G. (Eds.), 1997. Higher Education Research at the Turn of the New Century. New York and London: Garland Publishing, Paris: UNESCO.

Sadlak, J., Hüfner, K., 2002. International governmental organisations and research on higher education. In: Enders, J., Fulton, O. (Eds.), Higher Education in a Globalising World. Kluwer Academic Publishers, Dordrecht, pp. 87–100.

Schwarz, S., Teichler, U. (Eds.), 2000. The Institutional Basis of Higher Education Research: Experiences and Perspectives. Kluwer Academic Publishers, Dordrecht.

Teichler, U., 1988. Changing Patterns of the Higher Education System. Jessica Kingsley Publishers, London.

Teichler, U., 1996. Comparative higher education: potentials and limits. Higher Education 32 (4), 431–465.

Teichler, U., 2005. Research on higher education in Europe. European Journal of Education 40 (4), 431–465.

Teichler, U., 2009. Higher Education and the World of Work. Sense Publishers, Rotterdam and Taipei.

Teichler, U., Sadlak, J. (Eds.), 2000. Higher Education Research: Its Relationships to Policy and Practice. Pergamon Press, Oxford.

Tight, M., 2003. Researching Higher Education. Open University Press, Buckingham.

Tight, M., 2012. Researching Higher Education, second ed. Open University Press, Maidenhead.

Trow, M., 1973. Problems in the Transition from Elite to Mass Higher Education. Carnegie Commission on Higher Education, Berkeley, CA.

Välimaa, J., 2008. On comparative research in higher education. In: Amaral, A., Bleiklie, I., Musselin, C. (Eds.), From Governance to Identity – a Festschrift for Mary Henkel. Springer, Dordrecht, pp. 141–155.

Hilberg, Raul (1926–2007)

Detlev Claussen, Institute for Sociology, Leibniz University Hannover, Hannover, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Raul Hilberg, born in Vienna, was an Austrian-American political scientist most famous for his massive volume, *The Destruction of the European Jews*. The volume, derived from his dissertation, used archival material to document the role of German bureaucracies under National Socialism in the Holocaust. Though subsequently considered to be the definitive work on the Holocaust, public unwillingness to scientifically analyze the horrors created difficulties in publishing and distributing the work. After its publication, Hilberg analyzed the role of the railroads in the Holocaust and then the relationships between perpetrators, victims, and bystanders during the Holocaust. His later publications reflect on the interpretation of archival material related to the Holocaust.

Raul Hilberg's reputation rests on a single, albeit monumental work: *The Destruction of the European Jews*. When the first version appeared in English in 1961 no one could have foreseen that it would become a standard work in the social sciences. At that time, the 'Holocaust' did not exist as a concept either in academic discourse or the mass media. Nevertheless, with his first publication Hilberg found he had plunged into the turbulent waters of public debate both in Germany and in the USA. The *New Yorker*'s publication in installments of Hannah Arendt's reportage *Eichmann in Jerusalem* had created a sensation in the United States because it held the Jewish Councils partly responsible for the efficient functioning of the machinery of mass murder under the Nazis. Her study of 'the banality of evil' drew extensively on the findings in Raul Hilberg's book which had only recently appeared and was still largely unknown. It is an irony of history that when Hannah Arendt had acted as an anonymous reviewer for the Princeton University Press, she had recommended rejecting the book (See Raul Hilberg, 1996). For a long time Hilberg's work was thought to have shared the controversial views adopted by Hannah Arendt. This did nothing to promote the reception of his book either in the United States, Israel, or Germany.

Hilberg's first supervisor for his doctorate was Franz Neumann (1900–54), who had lost his life in a car accident in 1954. Neumann had warned him of the risks involved in his chosen topic but finally agreed to sponsor it. (*The Politics of Memory*: 66.) In 1948, when Hilberg submitted his master's dissertation to him, Neumann, the author of *Behemoth*, the pioneering study of National Socialism, objected to only one passage in which Hilberg had maintained that "administratively the Germans had relied on the Jews to follow directives (and), that the Jews had cooperated in their own destruction. Neumann did not say that this finding was contradicted by any facts; he did not say that it was under-researched. He said only, 'This is too much to take – cut it out.'" (*The Politics of Memory*: 66. Neumann added: "It's your funeral.") Neumann knew what he was talking about. After his escape from Germany, he had found a bolthole in the Institute of Social Research, which Max Horkheimer had succeeded in transferring from Frankfurt to New York. During the War he had worked on 'analyses of the enemy' for the Office of Strategic Services and had done preparatory work on war crimes trials. Returning to Columbia University, he lectured on political science. Neumann realized that the 'destruction of the European Jews' was a subject that would burst the boundaries separating the academic disciplines and that choosing such a topic would do nothing to advance Hilberg's career.

Neumann himself was a lawyer with sociological training. His *magnum opus*, *Behemoth*, which had appeared in 1942, took decades to gain recognition as a standard work of the social sciences, although from the very outset C. Wright Mills had at once judged it to be "a definitive analysis of the German Reich and a basic contribution to the social sciences"(C. Wright Mills, 1942). But the start of the Cold War spelled the end of research on National Socialism and its crimes. Neumann was personally affected by this – not just in academe but in attitudes in the Western world more generally. Neumann had encouraged Hilberg in his desire to examine the bureaucratic procedures involved in the mass murder of the Jews – an approach he had himself pioneered in his own account of National Socialism. This approach implied a close study of the records from which it would then be possible to reconstruct the course of events. As a trained lawyer, Neumann was familiar with such a procedure. Hilberg still had to learn to do this and to refine his own methods of interpreting the seemingly lifeless sources. Hilberg's initial experience of Nazi perpetrators had come as a young man with the US Army in Germany where he had been sent as a 'scout observer' at the end of the war. Hilberg came from a Jewish family in Austria and had left Vienna as a child in 1939. Having taken out American citizenship, he volunteered for the army. After being demobbed he returned to New York and while studying in Brooklyn College he came across Hans Rosenberg, an émigré German-Jewish historian who drew his attention to the part played by the Prussian civil service in the modernization of German society.

It was but a short step from there to meet up with Neumann who was teaching at Columbia. In 1951, Neumann found Hilberg a temporary job with the War Documentation Project in Alexandria (Virginia), where there were '28 000 linear feet' of records that had been captured from the Nazis. This is where Hilberg learned how to interpret documents. Following Neumann's unexpected death in 1954, he was forced to take up a teaching post in Burlington, Vermont, a job he retained until his retirement in 1991. This post enabled him to complete

his unique project 'The Destruction of the European Jews' for his dissertation and to save up enough money to pay toward the publication of his monumental, 1000-page work. His manuscript's bizarre journey through the publishing landscape of the United States, Germany, and Israel gives us a harsh insight into the public's unwillingness to face up to the scientifically grounded knowledge of crimes committed on such a massive scale that they threaten to overstep the limits of human understanding. This general climate did not change until 1978, when the American TV series 'Holocaust' was broadcast worldwide. In 1979, Hilberg succeeding in publishing *The Diary of Adam Czerniakow* and this was followed in 1982 with the first edition of *The Destruction of the European Jews* in German. An Israeli edition did not appear until 2012, 5 years after Hilberg's death.

The resistance to the results of Hilberg's research in different societies can be explained by their differing ways of coming to terms with the past. The 'Holocaust' TV series in 1978 marked a turning point. Until then the term was a rare word, used mainly by theologians. The realization that Jewish history in Europe could be regarded as one of the roots of immigration to the United States implied that the memory of the Holocaust had to be seen as an integral component of Americans' understanding of their own history. To Hilberg's surprise, the fundamental thesis of the book, that the process of destruction was one gigantic piece of bureaucracy, was scarcely contested initially. (Hilberg, *The Politics of Memory*: 124–125.) However, in the United States, thanks to the Hannah Arendt's positive endorsement, the book found itself caught in the cross fire of the debates about the Jewish Councils (Anson G. Rabinbach, 2000). Hilberg even encountered unexpected difficulties during the actual printing. His cooperation with Yad Vashem fell apart in 1958. Hilberg himself blames this on Yad Vashem's claim to be Israel's 'Remembrance Authority of the Disaster and Heroism,' to quote its official title (Hilberg, *The Politics of Memory*: 109). For his part, Gordon Hubel, the representative of Princeton University Press, fortified by Hannah Arendt's opinion on the manuscript, pointed out to him that there were already a number of 'readily available' books with a broader appeal in existence, and that his book was of interest only to "a very few specialists." (Hilberg, *The Politics of Memory*: 114)

The publication history of Hilberg's great work can only be grasped if we remind ourselves of the lack of clarity in the attitudes of the different nations toward the National Socialists' mass destruction of Jews of Europe. The division of labor in the different university disciplines prevented people from recognizing that this was a legitimate subject of research. Poliakov's *Harvest of Hate* (1951), with a preface by François Mauriac, was more like a contemporary report than an analysis. In 1953, Gerald Reitlinger, the English Orientalist, produced a succinct summation of the process of the so-called Final Solution in general terms. H.G. Adler, in contrast, presented the indifferent German public with an insight into the workings of a single concentration camp, viz. Theresienstadt. Raul Hilberg's magnum opus should be seen as neither a specialist study nor a general survey but as the meticulous empirical reconstruction of the entire process of the destruction of the Jews of Europe, based on the records of the perpetrators. As such it remains unsurpassed to this day, as Christopher Browning, an internationally influential expert on National Socialism, has confirmed (Christopher Browning, 1986).

Hannah Arendt identified a weakness in Hilberg, one that she described in drastic terms in her correspondence with Karl Jaspers, the philosopher of Existentialism. Hilberg "is talking some nonsense about the 'death wish' of the Jews. His book is really excellent, but only because he just reports facts in it. His introductory chapter that deals with general and historical matter wouldn't pass muster in a pigpen" (Lotte Kohler and Hans Sauer, 1992). Such talk of a 'death wish' is not to be found in Hilberg; that is her own invention, but it is true enough that his lengthy introductory chapter does appear to conflict with what he goes on to do in the main part of the book. Hilberg lacks a theoretical framework that might enable us to reflect properly on his empirical findings. The upshot is that the process of annihilation appears as a kind of logical consequence of a millennium and a half of anti-Semitic persecution. On this point Hannah Arendt, as a theorist, feels superior to Raul Hilberg, the political scientist. The resulting controversies point to a more general problem of Holocaust research, namely that the mere act of description can confer rationality on the irrational. (On this point, see Dan Diner, *Beyond the conceivable. Studies on Germany, Nazism, and the Holocaust*, University of California, Berkeley, CA. 2000, Chapter 7: 'Historical understanding and Counterrationality. The Judenrat as Epistemological Vantage [Point],': 130–137; see also the chapter 'Historical Experience. Juxtaposing Perspectives on National Socialism' for its comments on Neumann: 164–167.)

The unprotected nature of the empirical material meant that in the debate between the 'functionalists' and the 'intentionalists' that arose in Germany in the years after 1968 both sides felt that they could appeal to Hilberg as an authority. This conflict intensified in the subsequent 'Historians debate' in Germany in the course of which Hans-Ulrich Wehler was able to expose the lack of theoretical underpinning among contemporary historians (Hans-Ulrich Wehler, 2007). The resulting paradox is that the preface to *The Destruction of the European Jews* suggests an intentionalist approach, while the body of the work points to a structuralist one. The genocide of the Jews cannot be explained simply with reference to Hitler's intentions and the monstrous propensities of the Nazi perpetrators. Thus far Hilberg's study points in the direction of Hannah Arendt's thesis of the 'banality of evil,' but is by no means identical with it. Hilberg's further research, together with his two revised editions of his magnum opus, can also be interpreted as a response to the controversies triggered by his book in the United States, Germany, and Israel.

While the book he wrote in 1976 on *The Role of the German Railroads in the Destruction of the Jews* appears to be based purely on facts, it depends on his trust in empirical reconstruction for its intelligibility. In *Perpetrators, Victims, Bystanders: The Jewish Catastrophe 1933–1945*, a book that appeared in 1992, we can see how the historical record is distorted by the perpetrator/victim dichotomy that also dominates the public discussions of this question. The importance of the indifference of the bystanders as a crucial concept for explaining the behavior of the majority of people remains a closed book to him. His next publication *Sources of Holocaust Research: An Analysis* (2001) reflects this contradictory treatment of his sources. Historical

sources do not speak for themselves; one has to be able to read them. His memoir *The Politics of Memory*: *The Journey of a Holocaust Historian*, which appeared simultaneously in English and German, corrects a number of erroneous accounts. He is right to castigate the intellectual plagiarisms of competitors such as Lucy Dawidowicz and Nora Levin. Nora Levin (1916–89) published her book *The Destruction of European Jewry 1933–1945* in 1968 and in it she mercilessly plunders both Gerald Reitlinger and Hilberg. Lucy Dawidowicz's success (1975) hurt Hilberg even more deeply. Her book was tailored to the tastes of the American public and it enabled her to make the leap in status from scholarly outsider to officially acknowledged expert, a status Hilberg never achieved in the United States. His autobiography *The Politics of Memory: The Journey of a Holocaust Historian* comes very close to an idea formulated by Theodor W. Adorno that was very familiar to Hilberg's mentor Franz Neumann. "When Theodor Adorno condemned those who would write poetry after Auschwitz as barbaric, I had a problem. I am, of course, no poet, but have been a footnote writer. Are footnotes less barbaric?" (Hilberg, *The Politics of Memory*: 138.)

See also: Bureaucratization and Bureaucracy, History of; Cold War: Impact on Social and Behavioral Sciences; Empirical Social Research, History of; Genocide and War; Genocide, Historical Aspects of; Holocaust, The; National Socialism and Fascism; Public Sphere: Nineteenth and Twentieth-Century History; War Propaganda; War and Nationalism.

Bibliography

Browning, Christopher, 1986. The Revised Hilberg. In: Simon Wiesenthal Center Annual, vol. III, pp. 289–300.

Hilberg, Raul, 1996. The Politics of Memory: The Journey of a Holocaust Historian. Ivan R. Dee, Chicago, p. 156.

Kohler, Lotte, Sauer, Hans, 1992. Hannah Arendt/Karl Jaspers, Correspondence 1926–1969 (Robert, Rita Kimber, Ed. and Trans.). Harcourt Brace Jovanovich, New York, pp. 549–551.

Mills C. Wright, 1942[1964]. Now in Power, Politics and People. The Collected Essays of C. Wright Mills. New York, p. 170

Rabinbach, Anson G., 2000. Hannah Arendt und die New Yorker Intellektuellen. Anthony Grafton, 'Arendt und Eichmann am Esstisch,'. In: Smith, Gary (Ed.), Hannah Arendt Revisited: "Eichmann in Jerusalem" und die Folgen. Frankfurt am Main, pp. 33–56.

Wehler, Hans-Ulrich, 2007. Intentionalisten, Strukturalisten und das Theoriedefizit der Zeitgeschichte. In: Frei, Norbert, Broszat, Martin (Eds.), Der 'Staat' Hitlers und die Historisierung des Nationalsozialismus. Göttingen, p. 72.

Hintze, Otto (1861–1940)

Witold Neugebauer, University of Wurzburg, Wurzburg, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 6708–6711, © 2001, Elsevier Ltd.

Abstract

The Berlin historian Otto Hintze continues to fascinate those engaged in international historical research. In Germany, there was a particularly noticeable increase of interest in Hintze in the 1960s, in conjunction with an enduring resurgence of social historical schools. The rare combination of detailed investigation through decades of sources on the one hand and the elaborated approach of history's systematic contiguous disciplines on the other brought to light results which are still of great interest in modern research. This is especially true for the comparative approach to political structural history, particularly that of the European states from the height of the Middle Ages to the nineteenth century. From an early stage, Hintze was at pains to incorporate the categories of political science and economics, constitutional law, and sociology into historiographical analysis, thus broadening the scope of the discipline. Despite modernization theory and more recent concepts of the comparative analysis of nation building, Hintze's advocacy of the comparative approach to constitutional and governmental history and the methodological initiatives taken in his works are still of great significance. Indeed, Hintze's approach remains indispensable to the investigation of the processes of state making.

The Berlin historian Otto Hintze continues to fascinate those engaged in international historical research. In Germany, there was a particularly noticeable increase of interest in Hintze in the 1960s, in conjunction with an enduring resurgence of social historical schools. The rare combination of detailed investigation through decades of sources on the one hand and the elaborated approach of history's systematic contiguous disciplines on the other brought to light results which are still of great interest in modern research.

This is especially true for the comparative approach to political structural history, particularly that of the European states from the height of the Middle Ages to the nineteenth century. From an early stage, Hintze was at pains to incorporate the categories of political science and economics, constitutional law, and sociology into historiographical analysis, thus broadening the scope of the discipline. Despite modernization theory and more recent concepts of the comparative analysis of nation building, Hintze's advocacy of the comparative approach to constitutional and governmental history and the methodological initiatives taken in his works are still of great significance. Indeed, Hintze's approach remains indispensable to the investigation of the processes of state making.

Otto Hintze was born in the small Pomeranian town of Pyritz on August 27, 1861, as the son of a middle-ranking civil servant. After studying for two years at the University of Greifswald, Hintze transferred to the Friedrich-Wilhelms-Universität of Berlin in 1880–81. There he studied, for example, under Johann Gustav Droysen, who was still in his final years of life. Hintze completed his doctoral studies in the spring of 1884 with a dissertation on medieval history (*König Wilhelm der Holländer. Ein Beitrag zur Geschichte des deutschen Interregnums*), supervised by the historian Julius von Weizsäcker. Gustav Schmoller, the economic and political scientist who was soon to have a considerable impact on Hintze, was not yet named as one of his academic teachers. Hintze followed his first rather narrow course of study in history with a second in legal and constitutional history (sciences of state). According to the records, this was a formative stage for Hintze, during which Schmoller introduced him to archival research on the Prussian administrative files. Even this, his very first research on unpublished source materials, was carried out in connection with early comparative projects. At the same time, it was Hintze's first practical preparation for his work on the *Acta Borussica*, Gustav Schmoller's editing and research project on Prussian history in the eighteenth century in particular.

Hintze was engaged in this exhaustive project, which was based at the Prussian Academy of Sciences, in various functions over a period of decades. He was initially contracted in 1888 as a research fellow. In this capacity, he acquired an unparalleled command of the published and unpublished source material on the Brandenburg-Prussian state, its regions and provinces, classes and estates, before expanding the scope of his research to include the sixteenth and nineteenth centuries. He completed his postdoctoral thesis, a monograph on the history of the Prussian state economy (*The Prussian silk industry in the 18th century*), in 1895. In 1899 he was appointed professor at the University of Berlin, where – turning down several offers of chairs at other universities – he advanced to the status of full professor from 1900 to 1902–07, and continued to teach until prematurely given emeritus status in 1920. Besides his broad research, which had long since extended beyond the boundaries of Prussia, he owed this distinguished academic career to Gustav Schmoller's excellent contacts in political and administrative circles which is of vital importance in the view of the various animosities among contemporary historians.

The initial focus of Hintze's extensive writings was on constitutional and economic issues of Prussian history, culminating in 1901 in a lengthy monograph on the organization of the Prussian state administrative authorities around 1740. In this volume Hintze analyzed the making of the Prussian state, its development from a territorial state to a 'grand state.' He then reluctantly accepted an official commission to compile a comprehensive survey of the history of Brandenburg-Prussia. Several editions of this work were published in 1915–16, leading to heated (interdenominational) controversies during World War I.

In the course of archival research spanning two decades, Hintze acquired an excellent grasp of the source material, establishing an empirical basis that enabled him, even in further-reaching investigations, to take concrete historical life into consideration, particularly that of more recent centuries. For Hintze, the exploration of Prussian history was never a limited end in itself, but a paradigm for the historical process *per se*. This does not mean to say that he took this historical example as a yardstick by which the development of other states was to be measured, however. By his own account, it was at a early stage of his career – during his studies of legal and constitutional history under Schmoller – that Hintze had identified a much broader area as his future field of research: general constitutional and administrative history, from the comparative perspective. Furthermore, in choosing his Prussian topics of inquiry, he had always considered their usefulness for the 'entirety of my academic plans and studies' (1892).

Long before the World War I, which marked a turning point in Hintze's life, he became engaged in other, broader fields of research. In the course of his dispute with the protagonist of the older German historical school of economics, Wilhelm Roscher, Hintze had presented his conception of a comparative approach to the analysis of world history as early as 1897. In contrast to Roscher, whom he accused of neglecting the fundamental social aspects of state structures, Hintze devised a procedure that integrated both social conditions and the effects of foreign policy and power politics into the analysis of the formation of political structures. Hintze's main concern was to identify the recurring patterns in the development of states by means of a typological and comparative procedure. 'In the social region, there is a tendency towards regulated organic development: the making of states follows this developmental trend to a certain extent, simultaneously regulating it, but is by no means dependent on it alone.' Hintze maintained that the implications of 'power interests' were not to be underestimated, that the establishment of a state was 'at least as dependent on its external position of power as on the preconditions of the particular social structure for domestic governmental activity,' and that the 'external conditions cannot fail to influence social development.' It was this 'political side' of the formation of political structures which Hintze felt was neglected in any typological theory of states which, following the Aristotelian example, merely tried to identify analogous sequences of state systems. Hintze considered the process of state making to be largely dependent on the 'interrelations of the states, on the forces prevailing amongst them, on the rise and fall of the neighbouring states,' i.e., on what Leopold von Ranke had called the 'great world relationships.' (*see* Ranke, Leopold von (1795–1886)) Hintze considered power as a given category, which – like the phenomenon of state rivalries – he did not attempt to deduce in any further detail. For him, the central aim was to define and register 'types conditioned by world history' phenomena observed in groups of states that had maintained close political relations. In this sense, he explained afterwards, he understood state making to be a 'general, regulative principle,' thus implying that he did not intend strict rules to be postulated. It seems that in his early writings, Hintze still assumed a cautious primacy of foreign policy factors in the process of state making, but that – remarkable in the history of science – he turned away from this explanatory model over the course of his further research and analysis.

In a whole series of individual studies, whose scope was gradually to broaden, reaching the dimensions of world history in about 1930, Hintze applied and tested his typologically-based comparative approach. He developed what could be termed a spatial-typological procedure, which makes (European) historical development appear to be that of a bonded unit of regions. Above all, Hintze was keenly interested in the differentiation of a 'core Europe' (with a close genetic and spatial relationship to the Carolingian Empire) and a ring of 'peripheral' countries. He tested this spatial typological and structural historical differentiation at various junctures in his comparative studies of constitutional history – in his works on the history of the European estates, for example. It also becomes apparent in the comparison of continental western and central European absolutism on the one hand and the insular parliamentarianism of England on the other, a differentiation which has admittedly been called into question in recent Anglo-Saxon research. The 'peripheral' states of England, Sweden, Poland, and Hungary were of great interest to Hintze on account of their numerous structural parallels, especially where the 'parliamentarian' tradition was concerned.

Hintze set himself the aim of writing a comprehensive 'general comparative constitutional and administrative history of the modern states, namely the Latin and Germanic peoples,' but the manuscript was no longer published from around 1930 onwards. Central parts of this work were lost after Hintze's death; the fragments that survived are currently being edited by an international team. New sources reveal that Hintze remained keenly interested in the 'individuality' of states and political processes. Of course, the historical legacy of Hintze's work should not be allowed to conceal his innovative potential.

Otto Hintze sought to establish contact with the systematic contiguous disciplines at an early stage of his career, and not merely in reaction to the traumatic upheaval of political cultures subsequent to the 1914/1918 war. On the contrary, he had already begun to lose interest in the Prussian paradigm in the years preceding the war. His openness to the more modern approaches taken by social history, psychology, and sociology became evident as early as 1900, in the mediating statement he made during the so-called Lamprecht dispute about the relationship of 'the individualistic and the collectivist conception of history' (1897). Hintze pointed to the significant contribution made to historiography by Comte and Spencer (*see* Comte, Auguste (1798–1857); Spencer, Herbert (1820–1903)), both of whom he referred to in his lectures. For Hintze, individuality and collectivity constituted the two poles of historical insight, and he maintained that the methods of social psychology afforded great opportunities to go beyond Ranke. Admittedly, Hintze's growing interest in the concepts of sociology characterizes him as an outsider among the historians of his time. Hintze's great 1920s discussions on the theory and philosophy of history, not only with Weber (*see* Weber, Max (1864–1920)), but also with Oppenheimer, Sombart, and Troeltsch, as well as with the jurists Kelsen and Smend, merely marked the culmination of a continuous

development in his work. This is demonstrated, for example, by Hintze's discussion of Sombart's *Modern Capitalism*, in which the Berlin historian also referred to Marx. The extent to which Hintze started from the assumption of a primary, autonomous concept of society is a matter of debate. While he regarded the state as the dominant factor, his later studies and the surviving fragments of manuscripts written prior to 1933 reveal marked changes and a broadening of approach – he now tended towards a concept of society which was not derived from the state alone. For Hintze, social and socio-economic factors began to gain in causal-analytic significance; the same applies to the concept of class, though Hintze never approached Marxist conceptions with anything but a distant interest. It is thus also a matter of contention whether Hintze's typological procedure, along with the principle of 'clear abstraction' (*anschauliche* Abstraction), *ipso facto* neglected the element of dynamics. Hintze's procedure was related (or rather contiguous) to Weber's conception of the ideal type (*Idealtypus*), though always referring to the central question of the individualities of historical developments. It should be noted that Georg Jellinek had made pertinent suggestions for a typological methodology as early as 1900 in his *Allgemeine Staatslehre* (*General Theory of State*), and that Hintze was able to develop these ideas in his search for phenomena of 'relative comparability' (Kocka). In any case, Hintze saw his work as a necessary supplement to both the legacy of Ranke and modern sociology.

Hintze's reception of Weber's writings also predates 1914. Both the strength of his systematic interests and the positioning of methodological boundaries are evident in Hintze's second unpublished magnum opus, which can only be reconstructed to some extent from lecture notes. After 1920, Hintze had intended to publish this work as a 'general theory of state and society on a historical basis,' a continuation of and supplement to his general constitutional history. He had intended to paint a 'typical picture of the modern state in terms of its shared properties and deveopmental trends, its various historical stages and individual forms.' In this 'policy' of Hintze, which he evolved in his middle period, the primacy of the state and of the law was (still) very apparent. 'The state must be considered from the legal, the philosophical, and the historical perspective' – in this order. Social factors were suited and subordinated to this conception. In contrast to his classical studies of comparative constitutional history, in which industry and workers (or the labor movement) played at most a marginal role, Hintze integrated the social developments of the nineteenth century into his 'policy' by means of critical discussion of Marx and Lassalle. In so doing, Hintze was able to identify developments in the middle classes and nonmanual occupations that were to be examined in greater detail by later social historians.

All this makes Hintze still of interest to present-day historians. Of course, not all of his works are relevant today, as a brief glance at his political writings will show. However, his openness to the contributions of the social and political sciences – highly unusual among his contemporaries in the field – and the scope of his view, which finally extended beyond euro-centrism, mean that he remains an important, innovative figure in modern political structural history. Admittedly, the reconstruction of his *oeuvre* is essential to apply his approach in practical research also for further developing his theories, yet this is forever hindered by the fact that central works were never published and can only be substituted and inferred. After 1933 Otto Hintze, who had married the Jewish historian Hedwig Guggenhämer, became a *persona non grata* in Germany. He ordered that his personal papers should be destroyed after his death, which for the most part they were. Hintze died in his Berlin apartment on April 25, 1940, long since isolated from academic life.

See also: Comparative History; Constitutionalism, Comparative; Constitutions; Historical Thought and Historiography: Christian Tradition; Middle Ages, The; Nation, Sociological Aspects of; Nations and Nation-States in History; Political History: History of Politics; State Formation, Theory of; State and Society: History of the Concept.

Bibliography

Büsch, O., Erbe, M. (Eds.), 1983. Otto Hintze und die moderne Geschichtsschaft [Otto Hintze and Modern Historiography]. Colloquium Verlag, Berlin.

Hintze, O., 1908. Historische und politische Aufsätze [Historical and Political Essays], vol. 4. Verlag Deutsche Bucherei, Berlin.

Hintze, O., 1962–1967. In: Oestreich, G. (Ed.), Gesammelte Abhandlungen [selected works], second and third ed, vol. 3. Vandenhoeck & Ruprecht, Göttingen.

Hintze, O., 1998. In: Di Costanzo, G., Erbe, M., Neugebauer, W. (Eds.), Allgemeine Verfassungs- und Verwaltungsgeschichte der neueren Staaten. Fragmente [General Constitutional and Administrative History of the Newer States. Fragments], vol. 1. Palmar, Naples, Italy.

Kocka, J., 1972. Otto Hintze. In: Wehler, H.-U. (Ed.), Deutsche Historiker, vol. 3. Vandenhoeck & Ruprecht, Göttingen, pp. 41–64.

Kocka, J., 1981. Otto Hintze, Max Weber und das Problem der Burokratie [Otto Hintze, Max Weber and the problem of bureaucracy]. Historische Zeitschrift 233, 65–105.

Neugebauer, W., 1993. Otto Hintze und seine Konzeption der 'Allgemeinen Verfassungsgeschichte der Neueren Staaten' [Otto Hintze and his conception of the 'General consitutional history of the newer states']. Zeitschrift für historische Forschung 20, 65–96.

Neugebauer, W., 1998. Die wissenschaftlichen Anfänge Otto Hintzes [The academic origins of Otto Hintze]. Zeitschrift der Savigny-Stiftung für Rechtsgeschichte, Germanistische Abteilung 115, 540–551.

Neugebauer, W., 1999. Zum schwierigen Verhältnis von Geschichts-, Staats- und Wirtschaftswissenschaften [On the difficult relationship between history, political science and economics]. In: Kocka, J. (Ed.), Die Königlich Preußische Akademie der Wissenschaften im Kaiserreich. Akademie Verlag, Berlin, pp. 235–275.

Neugebauer, W., 2000. Zur Quellenlage der Hintze-Forschung [On source material in research on Hintze]. Jahrbuch fur die Geschichte Mittel- und Ostdeutschlands 45, 323–338.

Oestreich, B., 1989. Otto Hintze. In: Erbe, M. (Ed.), Berlinische Lebensbilder. Geisteswissenschaftler [Berlin Lives. Arts Scholars]. Colloquim Verlag, Berlin, pp. 287–309.

Smith, L.S., 1967. Otto Hintze's comparative constitutional history of the West. Doctoral dissertation, Saint Louis, MI.

Hippocampal Amnesia

Andrew Mayes, School of Psychological Sciences, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In global organic amnesia, anterograde amnesia in which recall and recognition of recently encountered facts and personal experiences is impaired, co-occurs with retrograde amnesia in which the same kind of memory disturbance is found but for memories that were formed prior to the causative brain lesion. Variants of this syndrome are triggered by lesions to any one of several interconnected brain regions, including the medial temporal lobes (MTL), midline diencephalon, basal forebrain, and retrosplenial cortex. Most amnesia-related research has been done on the MTL, which include the archicortical hippocampus and interconnected neocortical structures, such as the perirhinal cortex (PRC). What these structures do depends on how they process their inputs, what these are, and to where the processed products are sent. The hippocampus receives and mnemonically integrates processed object-related and spatiotemporal context information and has a different structure from the perirhinal and parahippocampal cortices, which receive mainly object and spatial context information respectively. There is evidence that selective hippocampal amnesia disrupts free and cued recall of recently encountered facts and events, but leaves familiarity for recently encountered stimuli intact. However, other hippocampal amnesics have equally impaired recall and familiarity memory and it has been argued that patients with selectively impaired recall have milder amnesia, smaller hippocampal lesions, and would show deficits for *strong* familiarity that diagnoses recognition very accurately. There are several reasons for not accepting this argument. There is no direct evidence for these three claims. Also, PRC lesions disrupt familiarity memory, but leave recall relatively intact, which cannot be explained by the strength confound claim. It seems more likely that global patients have subtle extrahippocampal damage that disrupts functioning in an extended PRC system that mediates stimulus familiarity. Other modeling, lesion and imaging evidence supports the selective recall-disruptive effect of hippocampal lesions and helps interpret it. This work suggests that hippocampus rapidly creates sparse, 'pattern-separated' memory representations of object-spatiotemporal context associations that stress *differences* between its often similar inputs whereas the medial temporal neocortices create pattern-generalized representations that stress *similarities* across their inputs. The pattern-separated representations are good at supporting recall through readily 'pattern-completed' partial cues. They do not support familiarity, which requires pattern-generalized representations, and is achieved through global matching. Whether hippocampal lesions disrupt recall, but not familiarity for premorbid episodes but not facts is underexplored. It is also still disputed to what extent hippocampal amnesics show impaired *perception* and immediate memory for spatiotemporal information.

Introduction

Organic amnesia typically involves two kinds of memory disorder: anterograde amnesia (AA) and retrograde amnesia (RA). AA comprises an impaired ability to recall and recognize recent experiences (episodic memories) and recently encountered facts (semantic memories). In severe amnesia, this deficit becomes apparent as soon as a patient is distracted following a new experience or encounter with a new fact. RA is similar in that it also involves an impaired ability to recall and recognize episodic and semantic information, but this information was put normally into memory prior to the brain damage that caused the memory syndrome. It is often claimed that RA shows a temporal gradient in which recall and recognition of memories formed longer before the triggering brain trauma are less affected and perhaps even normal. However, this claim remains controversial. Also, although AA and RA usually occur together, it remains unresolved to what extent the severity of AA correlates with that of RA (Mayes et al., 1997) and it has even been argued that selective AA or RA is found in some patients.

In one form or another, this syndrome can be caused by damage to any of a variety of brain structures. Thus, it can be caused by damage to different parts of the medial temporal lobes (MTL), the fornix, parts of the midline diencephalon that include the mammillary bodies, the mammillothalamic tract, the anterior thalamic nucleus, parts of the dorsomedial thalamic nucleus as well as other less well specified thalamic nuclei, the retrosplenial cortex, the ventral midbrain nucleus of Guddens, and structures in the basal forebrain, such as the septum and the diagonal band of Meynert. It has also been suggested that damage to some prefrontal regions and even parts of the parietal cortex may cause amnesic-like symptoms. Consistent with the similarity of the memory deficits that damage to each of these structures produces, it has long been acknowledged that they are highly interconnected so as to form, for example, a limbic memory circuit (Delay and Brion, 1969). The structures clearly form one or more functionally connected memory systems that mediate one or more kinds of episodic and semantic memory. Functional magnetic resonance imaging (fMRI) has shown that amnesia-causing lesions to non-MTL structures that leave the MTL structurally intact still disturb the activity and presumably the functional efficiency of the MTL (Snaphaan et al., 2009).

Although causing AA and RA, selective damage to many of these structures has been claimed to leave other cognitive functions and kinds of memory unaffected. Thus, patients with 'pure' amnesia often show good, perhaps unaffected, levels of intelligence and, by inference, attention. Strikingly, semantic memories that were heavily overlearnt prior to brain damage, such as applies to knowledge of word meanings, appear intact. Immediate memory on tests, such as digit span and Corsi

 http://dx.doi.org/10.1016/B978-0-08-097086-8.51039-3

blocks, may be normal as are kinds of longer-term memory that, in contrast to recall and recognition of experiences and facts, do not involve an aware feeling of memory. These preserved procedural/nondeclarative/implicit kinds of memory include skill/habit memory, simple classical conditioning, simple category memory, simple kinds of perceptual memory, and priming (see Spiers et al., 2001). Pure amnesics in whom this pattern of functioning is found are, however, rare and it is much more common to find patients in whom brain damage extends beyond the functional systems that mediate various aspects of aware episodic and semantic memory. In these patients, the partial disruption of brain regions critical for immediate memory, and various aspects of non-aware memory, intelligence and attention, leads to additional memory impairments as well as other cognitive disturbances. Such impure amnesias are often seen in those with closed head injuries and early Alzheimer's disease.

Severity of amnesia is indicated by how far below normal a patient's recall and recognition are in comparison to the mean level of healthy people with the same intelligence. This comparison is necessary because intelligence influences how good memory should be in healthy people by determining richness of memory encoding and flexibility of retrieval, and these processes will vary in pure amnesics because their intelligence also does. It is accepted that severity of amnesia varies from a level hard to distinguish from normal down to AAs that make life feel as if one is constantly waking from a dream, forgetting faster than one can recall what has just happened, and RAs where the past is virtually a blank slate so that one is left with little sense of self. Although severity must reflect the extent to which memory systems have been rendered dysfunctional, its precise determinants remain unresolved. This is partly because severity is confounded with the functional heterogeneity of amnesia, i.e., some lesions disrupt memory differently from other lesions. Nearly all researchers now agree this heterogeneity exists, although they disagree about its nature (e.g., Montaldi and Mayes, 2010; Wixted and Squire, 2011). Disagreement has been especially long-lasting and strong about the kind of functional heterogeneity revealed by damage to certain MTL structures. In particular, it is disputed in what way hippocampal damage (hippocampal amnesia) disrupts memory differently from damage to the perirhinal cortex (PRC) and, to a lesser degree, damage to the parahippocampal cortex (PHC).

This article focuses on the memory and non-memory effects of hippocampal damage (hippocampal amnesia) and how these effects differ from those of PRC and/or PHC damage. A framework for discussion will initially be provided. This framework will consider the inputs of MTL structures, their interconnections and their individual connections with extra-MTL brain structures and how knowledge of this constrains the likely functions of MTL structures and the functional systems of which they form part. As part of this initial discussion, the kinds of differential functions that have been considered for the hippocampus and some other MTL regions will be briefly outlined. The lesion evidence will then be reviewed after which evidence from neuroimaging studies, including fMRI, that are relevant to the interpretation of lesion studies and the possible functions of the hippocampus and other MTL regions will be considered.

Hippocampal and Other MTL Structure Functions: What Are the Constraints and What Are the Functions that Have Been Proposed?

The originating sites of inputs to different MTL structures should constrain *what* information each structure receives and processes whereas each structure's cytoarchitectural structure should constrain *how* it will *process* its inputs. The brain sites to which each MTL structure outputs the products of its processing should also constrain the overall *function* that is performed by the functionally interlinked system of structures of which it forms a part. These constraints are further tightened by the physiological and biochemical characteristics of MTL structures, knowledge of which is being steadily expanded by appropriate experimental and modeling work.

A hierarchical view of the connections to and within the MTL as is shown in **Figure 1**a, has changed little in 20 years (Zola-Morgan et al., 1994; Wixted and Squire, 2011) as Murray and Wise (2012) have pointed out. According to this view, many associative neocortex regions send their inputs to the PRC and PHC and from these structures to entorhinal cortex (ERC) and thence the hippocampus. The hippocampus sends backprojections to many of these associative neocortical sites. Inputs to the PRC and PHC come from distinct neocortical regions. PRC inputs primarily come from temporal cortex (notably from inferior temporal neocortex areas (TE/TEO)), insula and orbitofrontal cortex, which has led to the suggestion that PRC receives and processes nonspatial-object information (objects, faces, and words). In contrast, PHC inputs primarily come from parietal cortex, retrosplenial cortex and different frontal cortex regions, which has led to the suggestion that PHC receives and processes contextual information perhaps with a strong emphasis on spatial information. The strength of these PRC/PHC input differences should not be exaggerated, however, because these structures are directly connected, particularly strongly from PHC to PRC. The hippocampus receives both object-related and contextual information, which it is generally believed to bind/associate together into associative memory representations.

As is shown in **Figure 1**b, the extent to which PRC and PHC form the base of an MTL hierarchy with the hippocampus at its top has recently been criticized. Murray and Wise (2012) have pointed out that the hippocampus and PRC have separate evolutionary origins and are now juxtaposed because the hippocampus, which was originally a medial cortical structure, has been caudally and laterally displaced into the temporal lobe during mammalian evolution. Also, contrary to the strictly hierarchical MTL view, there are direct hippocampal-prefrontal projections and two way hippocampal-prefrontal connections via the retrosplenial cortex as well as other hippocampal-association cortex connections, none of which involve the PRC and PHC (see Aggleton, 2012). The functional implications of these connections remain to be fully identified as do its inputs from subcortical regions, such as the amygdala, basal forebrain (e.g., septum and diagonal band of Broca), ventral tegmentum, raphe nuclei, locus coeruleus, thalamic nuclei (e.g., paratenial nuclei, nucleus reuniens) and supra- and retromammillary regions. The hierarchical conception also

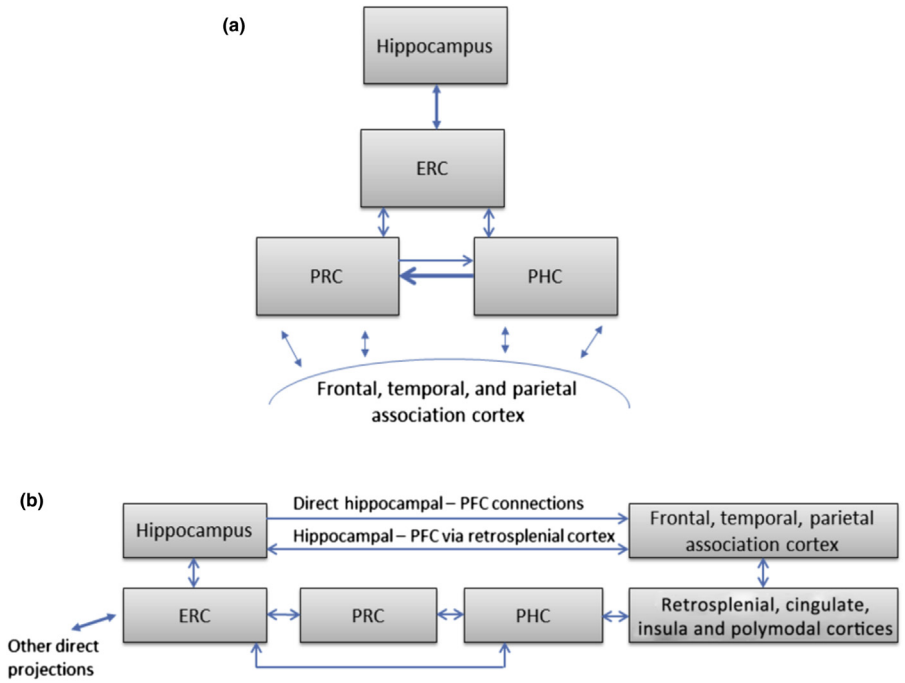

Figure 1 (a) Traditional view of hippocampus as the top structure in an information integration hierarchy with perirhinal and parahippocampal cortices at its base. Entorhinal cortex (ERC), perirhinal cortex (PRC), and parahippocampal cortex (PHC). (b) Modification of traditional hierarchical view, which stresses the greater number of cortical regions that feed into the hippocampus and the importance of prefrontal cortex (PFC) connections.

tends to underemphasize the subcortical projections of the hippocampus via the fornix to medial diencephalic nuclei, such as the mammillary bodies and anterior thalamus, and thence to the retrosplenial cortex, which connects strongly to the hippocampus (Figure 1).

The different evolutionary histories of the hippocampus and PRC/PHC are consistent with the cytoarchitectural differences between these two groups of structures. Whereas PRC and PHC are neocortical, the hippocampal complex (including dentate gyrus, cornu ammonis (CA) fields and subicular, presubicular, and parasubicular cortices) is a very different, somewhat heterogeneous archicortical region with a less layered structure. There are, of course, detailed differences of cytoarchitecture, but, despite this, it seems very likely that the broad kinds of processing engaged in by the PRC and PHC should be very similar to each other, albeit largely applied to different kinds of input. In contrast, the hippocampal processing of its distinct inputs should be very different from that of the PRC and PHC, although hippocampal subregions would also be expected to process somewhat distinct inputs perhaps slightly differently from each other.

The great number of hippocampal, PRC and PHC interconnections that are still being identified (see van Strien et al., 2009) does not preclude that these structures can work separately from each other when engaged in specific kinds of functioning that will involve separate systems of non-MTL structures. Sometimes the structures work together. In rats, evidence indicates that, following even unilateral lesions to the CA fields and dentate gyrus, there was reduced immediate early gene activity in PRC, PHC, and subiculum (Jenkins et al., 2006), although performance on a spatial working memory task was unaffected. But they can also work independently of each other (as again indicated by immediate early gene activity as a marker of neuronal responsivity). Thus, the hippocampus identifies spatial and temporal associative novelty (Jenkins et al., 2004; Amin et al., 2006) without affecting PRC and PHC activity whereas PRC identifies novelty of visually presented objects (see Aggleton and Brown, 2005). Although there may be some differences in MTL organization between rodents and nonhuman primates, let alone humans, if one can extrapolate, this work suggests that hippocampal and PRC lesions may sometimes disrupt the same memory functions and sometimes *selectively* disrupt different memory functions.

There are, therefore, general anatomical reasons for thinking that the hippocampus, PRC, and PHC process somewhat different inputs and that the neocortical MTL structures will transform/process their inputs in broadly different ways from the hippocampal complex, but the nature of these processing differences is unconstrained. However, extensive animal research with long-term potentiation (LTP) and long-term depression (LTD) and single unit recording suggests that the processing should involve LTP- and LTD-like storage of transformed spatiotemporal context inputs in the hippocampus and primarily LTD-like storage of transformed object inputs in the PRC (see Montaldi and Mayes, 2010, for a brief discussion).

Hippocampal Amnesia: The Lesion Evidence

The most famous case of hippocampal amnesia was HM, who developed a dense AA and what appeared to be RA with

a temporal gradient, following the surgical removal in 1953 of a large part of his MTL bilaterally to treat his intractable epilepsy at the age of 27. His case is famous because over the next 50 years he was the most studied amnesic and was the first to incontrovertibly show that MTL damage causes amnesia in humans (Corkin, 2013). However, his brain damage was not selective to the hippocampus because his amygdala and ERC were also extensively damaged. Damage extended into the parahippocampal gyrus, including parts of the PRC, and into the lateral temporal cortex (Corkin et al., 1997). Damage to the hippocampus is also caused by hypoxia that can occur following cardiac arrest, near drowning, attempted hanging, and carbon monoxide poisoning. However, hypoxic damage often and perhaps always extends beyond the hippocampus. Toxins, such as domoic acid, have also been shown occasionally to damage the hippocampus and amygdala relatively selectively and trigger amnesia (Teitelbaum et al., 1990). Autoimmune disorders, associated with several different autoantibodies, including those that act on voltage-gated potassium channel complexes and *N*-methyl-D-aspartate (NMDA) receptor sites, can disrupt the hippocampus and cause amnesia in what is sometimes referred to as limbic encephalitis (Vincent et al., 2010; Zandi et al., 2009). Such encephalitides appear to be more common than was once believed, but once again, they usually tend to affect more than the hippocampus.

Completely selective hippocampal damage (including the CA fields, dentate gyrus and subicular complex), therefore, probably does not exist in humans and even relatively selective damage is rare so that only a few cases have been reported (see Montaldi and Mayes, 2010). With these cases, the extent of confirmatory magnetic resonance imaging (MRI) has been variable so a question always remains about how much undetected dysfunction/damage there is outside the hippocampus and possibly the MTL. This may be the main reason why a dispute has existed for over a decade about how closely hippocampal amnesia resembles global organic amnesia in showing equally impaired recall and recognition of recently encountered stimuli. On the one hand, patients have been reported with AAs in which recall and recognition are equally impaired following hippocampal lesions (for example, Manns and Squire, 1999; see Wixted and Squire, 2011, for a review). On the other hand, patients have been reported in whom recall is more seriously disrupted than item recognition, which sometimes appears to be unimpaired (for example, Vargha-Khadem et al., 1997; Mayes et al., 2002; see Montaldi and Mayes, 2010; Brown et al., 2010, for reviews).

There is a well-established dual process view (for example, Mandler, 1980) that recognition is driven by two kinds of memory: recollection and familiarity. Recollection is a form of recall in which a previously studied stimulus cues the recall of details experienced when it was encountered on a previous occasion (such as what thoughts about the stimulus came to mind). It is, therefore, a form of associative memory in which contextual details are retrieved and help confirm that the stimulus was encountered earlier in a specific situation. In contrast, familiarity involves no recall, but the feeling that the stimulus has been encountered before (whether nonspecifically or in a specific context is still disputed). Both these kinds of memory vary in strength as indicated by confidence in decisions, speed of decision, and accuracy, but recollection is typically stronger. They depend on several processes working together and the relative contribution of recollection and familiarity to performance varies markedly across recognition tasks. Familiarity and recollection can be measured by several methods, each of which depends on hard-to-prove assumptions so there is some dispute about measurement effectiveness, although there are methods for dealing with this that include getting convergent results from more than one method (see Migo et al., 2012). In hippocampal patients with impaired item recognition, when measured, familiarity has been found to be as impaired as recollection (see Wixted and Squire, 2011, for a review). In patients with relatively spared recognition, when measured, familiarity has been found to be completely intact in the face of impaired recollection (see Montaldi and Mayes, 2010, for a review).

There is currently no agreed explanation why these two patterns of memory deficit have been found in patients with apparently fairly selective hippocampal damage. The differences certainly exist and are not an artifact of one pattern being found only with particular tests because sometimes the same tests have been used with both kinds of patients. For example, some 'global' patients are equally impaired on both recall and recognition performance with the Doors and People Battery (Baddeley et al., 1994) whereas 'selective' patients are more impaired on the recall tests from the Battery even though the recognition tests are difficulty matched (see Montaldi and Mayes, 2010 for a discussion). Nor is it likely, given the numbers involved, that 'selective' patients were premorbidly differentially good at familiarity memory or that 'global' patients were premorbidly differentially bad at familiarity memory. Nevertheless, there is a need for researchers to select appropriate tests and measures of familiarity and recollection.

The Doors and People Battery is special because its recognition tests are performed no better by healthy people than are its recall tests. This is not typical because recall performance is usually worse than recognition performance, which is explained in terms of the greater demands often made by recall tests. If this kind of difficulty is honored by the hippocampus and it mediates familiarity as well as recollection, one would expect that more of the structure would be needed to mediate recall efficiently so that small lesions should disrupt recall more than recognition as is found in 'selective' patients. Such an explanation would require that these patients should have smaller lesions and milder recall/recollection impairments. At present, however, there is no evidence that 'global' patients have more impaired recall/recollection or larger hippocampal lesions than 'selective' patients (see Montaldi and Mayes, 2010; Brown et al., 2010). One exception to this is perhaps patients who have 'recovered' from transient global amnesia. It has been shown that patients who are past the acute phase of this disorder often have acute punctate lesions of the hippocampus and a selective impairment of recollection (Jager et al., 2009). These patients may have milder recollection impairments and smaller lesions than typical 'global' patients.

If 'selective' hippocampal amnesia does not result from a simple artifact, it may still not reveal the critical memory role of the hippocampus in healthy people. This is because many

patients show variable degrees of memory recovery in the months following hippocampal damage (e.g., Henke et al., 1999; Holdstock et al., 2008). If future work shows that some patients recover familiarity memory alone whereas others do not, this would suggest the 'selective' patients do not refute the view that the hippocampus is vital for mediating familiarity memory as well as recollection/recall in healthy people. This might occur because 'selective' patients somehow exploit the redundant (degenerate) mediation of familiarity, but are much less able to do this with recollection. In other words, in addition to the primary role of the hippocampus and structures functionally linked to it in mediating familiarity, there are other normally inhibited, structures also capable of mediating familiarity (albeit rather less efficiently) after a period of reorganization following destruction of the hippocampus (which can be total) (see Fanselow, 2010, for a version of this idea). Evidence that 'selective' patients show completely preserved familiarity and the inability to explain why no equivalent redundant mediation of recollection exists make this interesting possibility implausible. An alternative possibility is that compared to 'global' patients, the hippocampus is less damaged in 'selective' patients so that they have more residual hippocampus, which is capable of supporting normal levels of familiarity, but not of recollection, following some recovery time. Both these possibilities imply that the hippocampus of healthy people plays a vital role in familiarity memory so there should be hippocampal activity changes when healthy people experience just familiarity, but this is strongly disputed (see Montaldi and Mayes, 2010; Wixted and Squire, 2011). The lack of evidence that 'selective' patients have less impaired recollection/recall and more hippocampal damage than 'global' patients makes the 'more residual hippocampus' explanation also seem very implausible (see Brown et al., 2010).

However, what needs to be known is how much residual hippocampus is left in individual patients. This means we need to know how much residual *functional* hippocampus is left but evidence about this is currently conflicting. On the one hand, Maguire et al. (2001, 2005) used fMRI to show that successfully recalling factual information failed to activate the remaining 50% of hippocampal gray matter volume of the 'global' patient, VC, but did do so in the 'selective' patient, Jon, who also showed a remaining hippocampal gray matter volume around 50%. These volume estimates were made using structural MRI. On the other hand, Jeneson et al. (2010) have argued that, there is likely to be nearly complete loss of hippocampal neurons when structural MRI measures show that hippocampal volume has reduced a third or more. If correct, this would imply that neither VC nor Jon should have had any residual functional hippocampus. This argument was based on two MRI-scanned patients who subsequently underwent postmortem histological analysis (Rempel-Clower et al., 1996). The conclusion is quite plausible because the residual gray matter, visible on MRI in hippocampal patients, does not necessarily comprise neuronal cell bodies. Nevertheless, it is too early to reject the possibility that the correspondence between structural MRI measures and postmortem histological measures is quite poor and much more research needs to be done in order to improve this correspondence. But, even if Maguire and her colleagues were detecting activations in residual hippocampus with patient Jon, they did this with factual recollection and not with familiarity on its own. Proper support for the residual hippocampus explanation requires that familiarity on its own activates the residual hippocampus in 'selective' but not 'global' patients, but hippocampal activation with familiarity on its own is still disputed even in healthy people with larger hippocampal structures.

Another explanation for the conflicting findings is that 'global' patients have more damage/dysfunction than 'selective' patients in nonhippocampal MTL and/or extra-MTL regions that mediate familiarity memory in the healthy brain. This explanation is plausible because using MRI to identify what may be relatively subtle brain damage and dysfunction is difficult and so MRI studies of hippocampal patients have rarely taken sufficient measures. The explanation also fits better with the observation that inconsistent results have been much less prominent in the animal literature where *selective* hippocampal lesions have usually been found to have little effect on recognition and not to disrupt familiarity at all when it has been measured (e.g., Sauvage et al., 2008; see Brown et al., 2010, for a review). The explanation needs to be systematically investigated by identifying the familiarity-mediating regions in healthy people with fMRI analysis of familiarity and connectivity analyses. Once this is done, differences in gray matter, white matter connections and/or signs of abnormal activity in the familiarity-system can be examined in a large group of hippocampal patients, who have been administered appropriate tests and measures of recollection and familiarity. If the 'global' patients show more damage and/or dysfunction extending into the familiarity-system than the 'selective' patients, especially if this disruption correlates with their familiarity impairments, the validity of the 'selective' effect would be supported.

Although no such study has yet been completed in patients with relatively selective hippocampal lesions, one has been completed with patients who suffered fornix damage following colloid cyst surgery (Tsivilis et al., 2008; Vann et al., 2009). Familiarity and recollection were assessed in a large group of patients using three different measures that gave convergent results. Patients with largely intact fornices and mammillary bodies (the volume of which was measured as a proxy for the fornix because it could be measured more accurately) were compared with very damaged/atrophied fornices/mammillary bodies, carefully matching for any additional brain damage and other neuropsychological measures. The patients with more extensive fornix/mammillary body damage were more impaired on recall/recollection measures, but not on the familiarity measures or on their Doors and People recognition performance. As a single case study of Carlesimo et al. (2007) found a similar pattern of performance in a patient with bilateral stroke damage primarily focused on the mammillothalamic tract, it looks as though this 'extended hippocampal system' may form part of the recollection/recall network.

Wixted and Squire (2011) have argued that both lesion and fMRI studies of recollection and familiarity are subject to a confound. They would not be convinced even by the kind

of large group study described above unless it found that selective hippocampal lesions left familiarity unimpaired when it was strength matched to recollection, which *was* impaired. They believe that lesion studies confound the recollection/familiarity comparison with the strong/weak recognition memory comparison so one cannot know whether a lesion is selectively disrupting recollection or strong recognition memory regardless of whether this is supported by recollection or familiarity. This is because recollection typically gives rise to more accurate (strong) recognition than familiarity, although highly confident and rapid familiarity judgments can lead to very accurate recognition that is matched to recollection (for example, see Montaldi et al., 2006; Montaldi and Mayes, 2010; Kafkas and Montaldi, 2012). Song et al. (2011a) found that their hippocampal patients showed an impaired level of highly confident and accurate familiarity. This, however, indicates nothing surprising because it was already known that these patients' familiarity was impaired, i.e., they were 'global' hippocampal amnesics. What *needs* to be examined is whether 'selective' hippocampal patients show sparing not just for familiarity in general, but also for very accurate familiarity.

The confound argument would, however, have difficulty explaining why hippocampal damage sometimes seems to disrupt recollection selectively whereas PRC (or other MTL neocortex) damage disrupts familiarity selectively. It is hard to explain how fairly weak/inaccurate familiarity needs far less of the resources of the hippocampus than stronger/more accurate recollection but that weak/inaccurate familiarity needs *more* of the resources of the PRC unless these structures have radically different roles in the processing of familiarity and recollection. Double dissociations of this kind have been reported in which impaired recollection was always associated with hippocampal damage and impaired familiarity was associated with PRC damage (e.g., Bowles et al., 2007, 2010), combined MTL neocortical damage (Wolk et al., 2011) or with ERC damage (Yonelinas et al., 2007). In each of these studies, PRC, etc. damage would not have been extensive so it is unclear whether extensive or complete damage would disrupt recollection. However, it is interesting to note that a study by Barbeau et al. (2011) found intact visual recall and impaired visual recognition in a patient with extensive right PRC damage whereas another patient with extensive right hippocampal damage and spared PRC showed the reverse pattern of impairment. Also, research with rats, using the receiver operating characteristics method of calculating recollection and familiarity, has found that, whereas hippocampal lesions only disrupt recollection (Sauvage et al., 2008), substantial and focal lesions of the amygdala (which is strongly connected to PRC) selectively disrupt familiarity (Farovik et al., 2011).

'Global' deficit patients are taken to support the view that the hippocampus supports familiarity as well as recollection for complex combinations of attributes that include spatio-temporal as well as object attributes (see Wixted and Squire, 2011). This view allows that, provided a stimulus involves only one or two attributes, familiarity may be adequately supported by the PRC without the hippocampus. The view seems to imply that functional differences between MTL structures mainly relate to the different attributes that they integrate into memory representations. The selective double dissociation pattern supports a different view, but this is not just about retrieval. Rather, it supports a view according to which PRC (and PHC) receives object attribute inputs, which it rapidly integrates into object memory representations that stress the similarities of the input to other inputs (it pattern 'generalizes'). These representations support familiarity well through a global match procedure in which success depends on the summed similarity of the representation to other representations. In contrast, the hippocampus receives object and spatiotemporal context attribute inputs, which it rapidly integrates into memory representations in which the differences from other, even similar, inputs is emphasized (pattern separation). This is achieved by the sparse encoding of inputs so that relatively few hippocampal neurons code for each input combination so there is relatively little overlap. Such memory representations support pattern completion (recollection/recall) from representation components that act as cues, but they are very poor at supporting global match. Unlike familiarity-related representations, these representations are also flexible as many cues can trigger recall, include all kinds of associative relationships and so have been described as relational (Konkel and Cohen, 2009). So, this view is that, not only do the MTL structures receive different inputs, but the hippocampus encodes and retrieves its inputs in a different way from PRC and PHC. The view is supported by neural network modeling (Norman and O'Reilly, 2003; Norman, 2010) as well as fMRI and other evidence about hippocampal pattern separation (see Yassa and Stark, 2011, for a review).

There is little evidence about whether the pattern of RA differs from that of AA in hippocampal amnesia. It is disputed whether systems consolidation slowly occurs so that memory for events and facts gradually becomes independent of the hippocampus and is unaffected by hippocampal lesions so that RA is temporally graded (see Squire and Bayley, 2007) or whether this is true of factual memories, but not of personal event memories unless they have been transformed into semanticized or gistlike representations (see Winocur and Moscovitch, 2011). According to the latter view, the RA gradient for detailed, untransformed episodic memories is flat. Only one study has so far investigated whether relatively selective hippocampal damage disrupts only recollection for premorbidly acquired fact and personal experience memories. Gilboa et al. (2006) used the receiver operating characteristics procedure for estimating familiarity and recollection in a 51-year old colloid cyst patient with moderately selective hippocampal atrophy. This patient showed preserved fact memory, which probably depended only on his normal familiarity because the procedure suggested that the controls only used familiarity. In contrast, the study showed very impaired recollection for detailed episodic memories and a flat RA gradient. The study leaves unresolved how hippocampal RA affects factual recollection, whether familiarity for episodic details exists and whether it is impaired in hippocampal RA.

The view that hippocampal damage just affects a particular kind of aware associative long-term memory has been

challenged in several ways in the last decade. First, some lesion and fMRI evidence has suggested that the hippocampus is necessary for discriminating between complex spatial scenes, particularly when these are pictured from different perspectives, even when long-term and working memory seem not to be required (see Lee et al., 2012, for a review). This suggests that the hippocampus is necessary for integrating object representations within an allocentric spatial framework, i.e., for perception as well as memory of such scenes. Without such high-level perceptual representations, recall memory may be lost abnormally fast in the minutes following learning as has been reported (see Isaac and Mayes, 1999a,b) and may explain why memory remains normal in amnesics for as long as interfering inputs are minimized (Della Sala, 2010; Dewar et al., 2010). Fast forgetting would be expected because, without distinctive kinds of hippocampally-dependent coding, such as spatial-object coding, the memory representation remaining will be very similar to the representations of many other inputs and suffer excessive interference (see Lee et al., 2012 for the same argument applied to PRC and object memory).

Second and relatedly, hippocampal amnesia disrupts imagination of future or fictitious scenes and events (e.g., Hassabis et al., 2007). Although this may be explicable in terms of a failure of scene *recall memory* and the central role of using such recall constructively in imagination, Mullally et al.'s (2012) finding that hippocampal lesions largely abolish the boundary extension effect suggests that both recall and imagination may depend on the ability to construct spatial scene representations. During encoding, healthy people construct a spatial scene representation that encloses objects/event details. This scene representation extends beyond its real boundaries and produces illusory boundary extension distortions in which the scene is expanded relative to the objects it encloses. The effect applies to haptic as well as visual presentations and suggests that hippocampal amnesics are unable to create the spatial aspects of scenes, something very similar to a perceptual process. The hippocampus may also be critical for temporal sequence memory and perhaps processing (Mayes et al., 2001; Devito and Eichenbaum, 2011), so it may be critical for creating a temporospatial framework for objects. Although it is controversial whether this subserves a high-level integrated form of perception, it is likely that this kind of memory representation underlies associative recall of facts and events. Selective hippocampal damage disrupts recognition and presumably recollection of certain kinds of association (e.g., face–voice associations) where spatiotemporal memory information should not help performance (see Mayes et al., 2007). Nonspatiotemporal components, like faces and voices, may, therefore, be mnemonically linked only in the hippocampus, but this may be automatically done within a spatiotemporal framework even though this is not needed to support recollection performance.

Third, it has been claimed that hippocampal damage disrupts working (immediate) memory for object-location associations (e.g., Hannula et al., 2006; Olson et al., 2006; Pertzov et al., 2013). Although this might be expected if the hippocampus is critical for perception of such associative information, both claims have been disputed by Jeneson and Squire (2012). They do not dispute that deficits can be found at very short delays, but argue that if working memory capacity is exceeded, if rehearsal is hard, or if attention is diverted, then impairments may result because the key distinction is between sub- and supraspan memory. This raises the interesting question of whether associative memory at these short delays depends on the same or different storage mechanisms from longer-term memory.

Fourth, even more controversially, it has been claimed that the hippocampus is critical for associative memory not just with but also without memory awareness (i.e., associative priming) (Henke, 2010; but see Verfaellie et al., 2012). If correct, this would suggest that the hippocampus does the processing to enable certain kinds of associative memories to be stored and later retrieved whether these are linked to conscious feelings of memory (as recollection is) or not, which leaves open whether other connected structures generate conscious memory feelings.

Fifth, Sharon et al. (2011) found that a hippocampal amnesic was able to learn and retain conscious memories of object-word associations normally through an incidental and inferential 'fast mapping' procedure, like that used by infants to learn words rapidly. The authors argued that these memories were accompanied by feelings of memory and lasted for at least a week and were probably mediated by temporal polar regions. Clearly this requires replication with a larger group of hippocampal patients, but, if true, would, unlike Henke's claim, narrow the conditions under which the hippocampus is necessary for forming associative memories.

There is growing evidence that the hippocampus may be functionally heterogeneous. One important distinction is between the anterior and posterior hippocampus (ventral and dorsal in rodents). Whereas posterior hippocampus is more strongly connected to PHC and retrosplenial cortex, anterior hippocampus is more strongly connected to the amygdala and to the PRC and seems to be functionally related more strongly to stress, emotions, such as anxiety and frustration, and social cognition and less to spatial memory than is posterior hippocampus (Fanselow and Dong, 2010; Maguire et al., 2000; Ranganath and Ritchey, 2012). The extent to which these differences are reflected in anterior and posterior hippocampal lesions in humans will be hard to investigate with selectively damaged patients.

Imaging Work that Is Relevant for Interpreting Hippocampal Amnesia

Research with fMRI has investigated whether encoding that leads to recollection or familiarity alone, or the occurrence of recollection or familiarity alone at test, has generally found that recollection activates the hippocampus whereas familiarity does not whereas familiarity correlates with reduced PRC activity relative to novelty (Ranganath et al., 2004; Kafkas and Montaldi, 2012; Montaldi et al., 2006; see Montaldi and Mayes, 2010, for a review). However, Wixted and Squire (2011) have argued that the blood oxygen level dependent (BOLD) response/recognition memory strength/accuracy function may be different for the hippocampus and PRC and that familiarity of similar accuracy to recollection would activate the hippocampus. In two studies, rating

recognition confidence and using the remember/know procedure for measuring word recollection and familiarity, they showed that encoding that produced strong familiarity, and strong familiarity at test, activated the hippocampus as much as recollection (Song et al., 2011b; Smith et al., 2011). However, Montaldi et al. (2006) and Kafkas and Montaldi (2012) have also matched strong familiarity with recollection for accuracy and found that recollection activated the hippocampus more and that familiarity did not activate the hippocampus as a function of confidence, reaction time, and accuracy. Their memory stimuli were everyday scene and object pictures respectively so two interpretations of these conflicting findings are possible.

The hippocampus may mediate familiarity as well as recollection for concrete word stimuli, but not for everyday scenes (comprising objects in a spatial context) and objects. Alternatively, the remember/know procedure may be almost impossible to follow for words because, unlike with scene and object pictures, when people reencode studied words at test they fail to identify recalled word–meaning related information as recollected and wrongly report familiarity. Furthermore, familiarity reports get more confident because unidentified recollections get richer and more detailed. Other evidence supports this interpretation. Thus, Qin et al. (2011) found that hippocampal activity increased as more information/detail was reported, but it did not increase as subjective confidence (and probably accuracy) increased provided amount recollected remained constant. In fMRI studies that examine word familiarity effects, the amount of information found familiar should always be the same so, if familiarity is like recollection of fixed amounts of information, one would expect no increase in hippocampal activity as confidence/accuracy increases. It seems more likely that the hippocampus is not involved in studied word familiarity, but activates more as amount recalled in unidentified recollection increases. However, this remains to be shown.

This interpretation is also supported by two sets of imaging studies. The first set of studies examined five pharmacoresistant epileptic patients who had intracranial EEG (iEEG) electrodes placed in their hippocampus and PRC. In the first study (Staresina et al., 2013), which used recordings from the hemisphere with no seizure-onset zone, the patients encoded words paired with one of two sources (colors or scenes) and were then given a recognition and source memory test for each studied word. The size of the peak iEEG response for correct rejections of unstudied words, recognition without source memory of studied words (arguably mainly familiarity), and recognition with source memory (recollection) was measured. The *relative* ordering of this peak response differed between hippocampus and PRC. In the hippocampus, recollection > familiarity = correct rejection whereas in PRC, correct rejection > recollection > familiarity. Even if the two regions show different memory strength/BOLD response functions, this indicates the presence of two memory signals according to the very strict nonmonotonic requirements of state-trace theory (Newell and Dunn, 2008). Given all available evidence, these two signals are probably familiarity and recollection. In the second study (Staresina et al., 2012), both recording methods used (fMRI and iEEG) showed that, in the hippocampus, an early recollection effect was followed by a late *postdecision* item novelty (correct rejection) effect whereas, in PRC, an early novelty effect was followed by a prolonged recollection effect. This temporal pattern suggests that a rapid PRC item familiarity signal triggers the hippocampus to recollect associated source information, which then perhaps activates PRC representations. Although this interpretation does not fit with the rat immediate early gene data described earlier, it is compatible with the increased functional connectivity of hippocampus and PRC that was found. No support is given for the idea that the hippocampus mediates word familiarity.

In the second set of studies, fMRI data were analyzed with both univariate and multivariate pattern analyses. In the first study (Liang et al., 2013) PRC novelty responses were maximal for faces whereas PHC novelty responses were maximal for scenes. Although these results fit with other findings indicating that object stimuli (like faces) are processed more anteriorly (PRC) whereas visuospatial scene stimuli are processed more posteriorly (PHC), multivariate analyses found that both PHC and PRC showed distinctive coding activity patterns for 'nonpreferred' stimuli that comprised visual and spoken words and sounds as well as faces and scenes. So, *both* these MTL neocortical areas code distinctively for object-related and more visuospatial scene-related information, which the authors argued probably depended on meaning-related differences between stimulus classes. In contrast, multivariate analysis showed that anterior hippocampus did not discriminate between stimulus classes whereas posterior hippocampus discriminated scenes more distinctively than other stimulus classes: these findings fit with functional anterior/posterior hippocampal differences and suggest that anterior hippocampus codes its inputs in a more domain general or 'abstract' way. In the second study (LaRocque et al., 2013), PHC, PRC, and hippocampal patterns of encoding activity that led to subsequent recognition of multiple stimulus categories, including faces, bodies, objects, and scenes, were examined using multivariate analyses. In PHC and PRC, the neural encoding pattern *similarity* of a stimulus to that of all other encoded stimuli was found to predict subsequent recognition whereas, in the hippocampus, subsequent recognition was better for stimuli that showed more pattern distinctiveness (i.e., dissimilarity of the hippocampal neural coding pattern) from other studied stimuli (i.e., better pattern-separated stimulus representations). These studies, therefore, suggest that the inputs of these MTL structures are not as distinct as is often assumed, but that PHC and PRC *code* their inputs very differently from the way the hippocampus codes its. Furthermore, whereas the hippocampus processes its inputs to produce pattern-separated representations that support recollection (pattern completion), the PHC and PRC code their inputs to create similarity-enhancing representations that support recognition through global similarity matching (familiarity).

Conclusion

Although some of the lesion and imaging data appear conflicting and cannot yet be fully explained, a description of what

happens in hippocampal amnesia can be tentatively made. Following selective damage, the hippocampus is no longer able to create flexible, integrated representations of its object-related and scene inputs within a spatiotemporal framework. In particular, it can no longer rapidly create sparse, pattern-separated associative memory representations of all kinds within this framework and later reactivate these from partial cues through pattern completion (recollection). There is no convincing lesion evidence that it plays any role in familiarity. Furthermore, the lesion evidence is supported by fMRI evidence, which indicates that it *does not create* memory representations that enhance similarity with other inputs so it cannot support global matching and probably familiarity. Instead, this familiarity-related processing is done by PRC and PHC. It remains unclear whether hippocampal damage disrupts high-level spatial scene perception, subthreshold immediate object-location memory (although it disrupts such associations at very short delays), and novel associative priming. It is also unclear whether, under certain conditions, arbitrary associations can be rapidly encoded and normally retained as aware, flexible recallable memories. Hippocampal damage impairs rapidly acquired semantic as well episodic memories whenever these involve the recall of associations. Finally, hippocampal damage probably causes RA for facts and events that were not overlearnt premorbidly. However, it remains unclear whether recall deficits for these memories reduce with memory age or whether impairment of detailed episodic recall remains constant as memories age.

See also: Episodic Memory; Future Thinking; Hippocampus and Related Structures; Implicit Memory; Learning and Memory, Neural Basis of; Recognition; Retroactive Interference; Working Memory, Neural Basis of.

Bibliography

Aggleton, J.P., 2012. Multiple anatomical systems embedded within the primate medial temporal lobe: implications for hippocampal function. Neuroscience Biobehavioral Reviews 36, 1579–1596.

Aggleton, J.P., Brown, M.W., 2005. Contrasting hippocampal and perirhinal cortex activity using immediate early gene activity. Quarterly Journal of Experimental Psychology 58B, 218–233.

Amin, E., Pearce, J.M., Brown, M.W., Aggleton, J.P., 2006. Novel temporal configurations of stimuli produce discrete changes in immediate early gene expression in the rat hippocampus. European Journal of Neuroscience 24, 2611–2621.

Baddeley, A.D., Emslie, H., Nimmo-Smith, I., 1994. Doors and People: A Test of Visual and Verbal Recall and Recognition. Thames Valley Test Company, Flempton, Bury St. Edmunds.

Barbeau, E.J., Pariente, J., Felician, O., Puel, M., 2011. Visual recognition memory: a double anatomo-functional dissociation. Hippocampus 21, 929–934.

Bowles, B., Crupi, C., Mirsattari, S.M., Pigott, S.E., Parrent, A.G., Pruessner, J.C., Yonelinas, A.P., Kohler, S., 2007. Impaired familiarity with preserved recollection after anterior temporal-lobe resection that spares the hippocampus. Proceedings of the National Academy of Sciences of the United States of America 104, 16382–16387.

Bowles, B., Crupi, C., Pigott, S.E., Parrent, A.G., Wieber, S., Janzen, L., Kohler, S., 2010. Double dissociation of selective recollection and familiarity impairments following two different surgical treatments for temporal-lobe epilepsy. Neuropsychologia 48, 2640–2647.

Brown, M.W., Warburton, E.C., Aggleton, J.P., 2010. Recognition memory: material, processes, and substrates. Hippocampus 20, 1228–1244.

Carlesimo, G.A., Serra, L., Fadda, L., Cherubini, A., Bozzali, M., Caltagirone, C., 2007. Bilateral damage to the mammillo-thalamic tract impairs recollection but not familiarity in the recognition process: a single case investigation. Neuropsychologia 45, 2467–2479.

Corkin, S., 2013. Permanent Present Tense: The Man with No Memory, and What He Taught the World. Allen Lane, London.

Corkin, S., Amaral, D.G., Gonzalez, R.G., Johnson, K.A., Hyman, B.T., 1997. H.M.'s medial temporal lobe lesion: findings from magnetic resonance imaging. Journal of Neuroscience 17, 3964–3979.

Delay, J., Brion, S., 1969. Le syndrome de Korsakoff. Masson, Paris.

Della Sala, S., 2010. Forgetting due to retroactive interference in amnesia: findings and implications. In: Della Sala, S. (Ed.), Forgetting. Psychology Press, Hove, East Sussex, pp. 185–209.

Devito, L.M., Eichenbaum, H., 2011. Memory for the order of events in specific sequences: contributions of the hippocampus and medial prefrontal cortex. Journal of Neuroscience 31, 3169–3175.

Dewar, M., Cowan, N., Della Sala, S., 2010. Forgetting due to retroactive interference in amnesia: findings and implications. In: Della Sala, S. (Ed.), Forgetting. Psychology Press, Hove, East Sussex, UK.

Fanselow, M.S., 2010. From contextual fear to a dynamic view of memory systems. Trends in Cognitive Sciences 14, 7–15.

Fanselow, M.S., Dong, H.-W., 2010. Are the dorsal and ventral hippocampus functionally distinct structures? Neuron 65, 7–19.

Farovik, A., Place, R.J., Miller, D.R., Eichenbaum, H., 2011. Amygdala lesions selectively impair familiarity in recognition memory. Nature Neuroscience 14, 1416–1417.

Gilboa, A., Winocur, G., Rosenbaum, R.S., Poreh, A., Gao, F., Black, S.E., Westmacott, R., Moscovitch, M., 2006. Hippocampus 16, 966–980.

Hannula, D.E., Tranel, J., Cohen, N.J., 2006. The long and the short of it: relational memory impairments in amnesia, even at short lags. Journal of Neuroscience 26, 8352–8359.

Hassabis, D., Kumaran, D., Maguire, E.A., 2007. Using imagination to understand the neural basis of episodic memory. Journal of Neuroscience 27, 14365–14374.

Henke, K., 2010. A model of memory systems based on processing modes rather than consciousness. Nature Reviews Neuroscience 11, 523–532.

Henke, K., Kroll, N.E.A., Behniea, H., Amaral, D.G., Miller, M.B., Rafal, R., Gazzaniga, M.S., 1999. Memory lost and memory regained following bilateral hippocampal damage. Journal of Cognitive Neuroscience 11, 682–697.

Holdstock, J.S., Parslow, D.M., Morris, R.G., Fleminger, S., Abrahams, S., Denby, C., Montaldi, D., Mayes, A.R., 2008. Two case studies illustrating how relatively selective hippocampal lesions in humans can have quite different effects on memory. Hippocampus 18, 679–691.

Isaac, C.L., Mayes, A.R., 1999a. Rate of forgetting in amnesia: I. Recall and recognition of prose. Journal of Experimental Psychology: Learning, Memory and Cognition 25, 942–962.

Isaac, C.L., Mayes, A.R., 1999b. Rate of forgetting in amnesia. II. Recall and recognition of word lists at different levels of organization. Journal of Experimental Psychology: Learning, Memory and Cognition 25, 963–977.

Jager, T., Szabo, K., Griebe, M., Bazner, H., Moller, J., Hennerici, M.G., 2009. Selective disruption of hippocampus-mediated recognition memory processes after episodes of transient global amnesia. Neuropsychologia 47, 70–76.

Jeneson, A., Squire, L.R., 2012. Working memory, long-term memory, and medial temporal lobe function. Learning & Memory 19, 15–25.

Jeneson, A., Kirwan, C.B., Hopkins, R.O., Wixted, J.T., Squire, L.R., 2010. Recognition memory and the hippocampus: A test of the hippocampus: a test of the hippocampal contribution to recollection and familiarity. Learning & Memory 17, 63–70.

Jenkins, T.A., Amin, E., Pearce, J.M., Brown, M.W., Aggleton, J.P., 2004. Novel spatial arrangements of familiar visual stimuli promote activity in the rat's hippocampal formation but not the para hippocampal cortices: a c-Fos expression study. Neuroscience 124, 43–52.

Jenkins, T.A., Amin, E., Brown, M.W., Aggleton, J.P., 2006. Changes in immediate early gene expression in the rat brain after unilateral lesions of the hippocampus. Neuroscience 137, 747–759.

Kafkas, A., Montaldi, D., 2012. Familiarity and recollection produce distinct eye movement, pupil and medial temporal lobe responses when memory strength is matched. Neuropsychologia 50, 3080–3093.

Konkel, A., Cohen, N.J., 2009. Relational memory and the hippocampus: representations and methods. Frontiers of Neuroscience 3, 166–174.

LaRocque, K., Smith, M., Carr, V., Witthoft, N., Grill-Spector, K., Wagner, A., 2013. Global similarity and pattern separation in the medial temporal lobe predict subsequent memory. Journal of Neuroscience 33, 5466–5499.

Lee, A.C.H., Yeung, L.-K., Barense, M.D., 2012. The hippocampus and visual perception. Frontiers in Human Neuroscience 6, 1–17.
Liang, J.C., Wagner, A.D., Preston, A.R., 2013. Content representation in the medial temporal lobe. Cerebral Cortex 23, 80–96.
Maguire, E.A., Frith, C.D., Rudge, P., Cipolotti, L., 2005. The effect of adult-acquired hippocampal damage on memory retrieval: an fMRI study. Neuroimage 27, 146–152.
Maguire, E.A., Gadian, D.G., Johnsrude, I.S., Good, C.D., Ashburner, J., Frackowiak, R.S., Frith, C.D., 2000. Navigation-related structural change in the hippocampi of taxi drivers. Proceedings of the National Academy of Sciences of the United States of America 97, 4398–4403.
Maguire, E.A., Vargha-Khadem, F., Mishkin, M., 2001. The effects of bilateral hippocampal damage on fMRI regional activations and interactions during memory retrieval. Brain 124, 1156–1170.
Mandler, G., 1980. Recognizing-the judgment of previous occurrence. Psychological Review 87, 252–271.
Manns, J.R., Squire, L.R., 1999. Impaired recognition memory on the Doors and People Test after damage limited to the hippocampal region. Hippocampus 9, 495–499.
Mayes, A.R., Daum, I., Markowitsch, H., Sauter, W., 1997. The relationship between retrograde and anterograde amnesia in patients with typical global amnesia. Cortex 33, 197–217.
Mayes, A.R., Holdstock, J.S., Isaac, C.L., Hunkin, N.M., Roberts, N., 2002. Relative sparing of item recognition memory in a patient with adult-onset damage limited to the hippocampus. Hippocampus 12, 325–340.
Mayes, A.R., Isaac, C.I., Holdstock, J.S., Hunkin, N.M., Montaldi, D., Downes, J.J., MacDonald, C., Cezayirli, E., Roberts, J.N., 2001. Memory for single items, word pairs, and temporal order of different kinds in a patient with selective hippocampal lesions. Cognitive Neuropsychiatry 18, 97–123.
Mayes, A.R., Montaldi, D., Migo, E., 2007. Associative memory and the medial temporal lobe. Trends in Cognitive Sciences 11, 126–135.
Migo, E.M., Mayes, A.R., Montaldi, D., 2012. Measuring recollection and familiarity: improving the remember/know procedure. Consciousness and Cognition 21, 1435–1455.
Montaldi, D., Mayes, A.R., 2010. The role of recollection and familiarity in the functional differentiation of the medial temporal lobes. Hippocampus 20, 1291–1314.
Montaldi, D., Spencer, T.J., Roberts, N., Mayes, A.R., 2006. The neural system that mediates familiarity memory. Hippocampus 16, 504–520.
Mullally, S.L., Intraub, H., Maguire, E.A., 2012. Attenuated boundary extension produces a paradoxical memory advantage in amnesic patients. Current Biology 22, 261–268.
Murray, E.A., Wise, S.P., 2012. Why is there a special issue on perirhinal cortex in a journal called hippocampus? The perirhinal cortex in historical perspective. Hippocampus 22, 1941–1951.
Newell, B.R., Dunn, J.C., 2008. Dimensions in data: testing psychological models using state-trace analysis. Trends in Cognitive Neurosciences 12, 285–290.
Norman, K.A., 2010. How hippocampus and cortex contribute to recognition memory: revisiting the complementary learning systems model. Hippocampus 20, 1217–1227.
Norman, K.A., O'Reilly, R.C., 2003. Modelling hippocampal and neocortical contributions to recognition memory: a complementary learning systems account. Psychological Review 110, 611–646.
Olson, I.R., Page, K., Moore, K.S., Chatterjee, A., Verfaellie, M., 2006. Working memory for conjunctions relies on the medial temporal lobes. Journal of Neuroscience 26, 4596–4601.
Pertzov, Y., Miller, T.D., Gorgoraptis, N., Caine, D., Schott, J.M., Butler, C., Husain, M., 2013. Binding deficits in memory following medial temporal lobe damage in patients with voltage-gated potassium channel complex antibody-associated limbic encephalitis. Brain 136, 2474–2485.
Qin, S., van Marle, H.J.F., Hermans, E.J., Fernandez, G., 2011. Subjective sense of memory strength and the objective amount of information accurately remembered are related to distinct neural correlates at encoding. Journal of Neuroscience 31, 8920–8927.
Ranganath, C., Ritchey, M., 2012. Two cortical systems in memory-guided behaviour. Nature Reviews Neuroscience 13, 713–726.
Ranganath, C., Yonelinas, A.P., Cohen, M.X., Dy, C.J., Tom, S.M., D'Esposito, M., 2004. Dissociable correlates of recollection and familiarity within the medial temporal lobes. Neuropsychologia 42, 2–13.
Rempel-Clower, N.L., Zola, S.M., Squire, L.R., Amaral, D.G., 1996. Three cases of enduring memory impairment after bilateral damage limited to the hippocampal formation. Journal of Neuroscience 16, 5233–5255.
Sauvage, M., Fortin, N., Owens, C., Yonelinas, A., Eichenbaum, H., 2008. Recognition memory: opposite effects of hippocampal damage on recollection and familiarity. Nature Neuroscience 1, 16–18.
Sharon, T., Moscovitch, M., Gilboa, A., 2011. Rapid neocortical acquisition of long-term arbitrary associations independent of the hippocampus. Proceedings of the National Academy of Sciences 108, 1146–1151.
Smith, C.N., Wixted, J.T., Squire, L.R., 2011. The hippocampus supports both recollection and familiarity when memories are strong. Journal of Neuroscience 31, 15693–15702.
Snaphaan, L., Rijpkema, M., van Uden, I., Fernandez, G., de Leeuw, F.-E., 2009. Reduced medial temporal lobe functionality in stroke patients: a functional magnetic resonance imaging study. Brain 132, 1882–1888.
Song, Z., Wixted, J.T., Hopkins, R.O., Squire, L.R., 2011a. Impaired capacity for familiarity after hippocampal damage. Proceedings of the National Academy of Sciences of the United States of America 108, 9655–9660.
Song, Z., Wixted, J.T., Smith, C.N., Squire, L.R., 2011b. Different nonlinear functions in hippocampus and perirhinal cortex relating functional MRI activity to memory strength. Proceedings of the National Academy of Sciences of the United States of America 108, 5783–5788.
Spiers, H.J., Maguire, E.A., Burgess, N., 2001. Hippocampal amnesia. Neurocase 7, 357–382.
Squire, L.R., Bayley, P.J., 2007. The neuroscience of remote memory. Current Opinion in Neurobiology 17, 185–196.
Staresina, B.P., Fell, J., Dunn, J.C., Axmacher, N., Henson, R.N., 2013. Using state-trace analysis to dissociate the functions of the human hippocampus and perirhinal cortex in recognition memory. Proceedings of the National Academy of Sciences of the United States of America 110, 3119–3124.
Staresina, B.P., Fell, J., Do Lam, A.T.A., Axmacher, N., Henson, R.N., 2012. Memory signals are temporally dissociated in and across human hippocampus and perirhinal cortex. Nature Neuroscience 15, 1167–1173.
Teitelbaum, J.S., Zatorre, R.J., Carpenter, S., Gendron, D., Evans, A.C., Gjedde, A., Cashman, N.R., 1990. Neurologic sequelae of domoic acid intoxication due to the ingestion of contaminated mussels. New England Journal of Medicine 322, 1781–1787.
Tsivilis, D., Vann, S.D., Denby, C., Roberts, N., Mayes, A.R., Montaldi, D., Aggleton, J.P., 2008. A disproportionate role for the fornix and mammillary bodies in recall versus recognition memory. Nature Neuroscience 11, 834–842.
Vann, S.D., Tsivilis, D., Denby, C.E., Quamme, J.R., Yonelinas, A.P., Aggleton, J.P., Montaldi, D., Mayes, A.R., 2009. Impaired recollection but spared familiarity in patients with extended hippocampal system damage revealed by three convergent methods. Proceedings of the National Academy of Sciences of the United States of America 106, 5442–5447.
Vargha-Khadem, F., Gadian, D.G., Watkins, K.E., Connelly, A., Van Paesschen, W., Mishkin, M., 1997. Developmental amnesia: effect of early hippocampal pathology on episodic and semantic memory. Science 277, 376–380.
van Strien, N.M., Cappaert, N.L.M., Witter, M.P., 2009. The anatomy of memory: an interactive overview of the parahippocampal–hippocampal network. Nature Reviews Neuroscience 10, 272–282.
Verfaellie, M., LaRocque, K.F., Keane, M.M., 2012. Intact implicit verbal relational memory in medial temporal lobe amnesia. Neuropsychologia 50, 2100–2106.
Vincent, A., Irani, S.R., Lang, B., 2010. The growing recognition of immunotherapy-responsive seizure disorders with autoantibodies to specific neuronal proteins. Current Opinion in Neurology 23, 144–150.
Winocur, G., Moscovitch, M., 2011. Memory transformation and systems consolidation. Journal of the International Neuropsychology Society 17, 766–780.
Wixted, J.T., Squire, L.R., 2011. The medial temporal lobe and the attributes of memory. Trends in Cognitive Sciences 15, 210–217.
Wolk, D.A., Dunfee, K.L., Dickerson, B.C., Aizenstein, H.J., DeKosky, S.T., 2011. A medial temporal lobe division of labor: insights from memory in aging and early Alzheimer disease. Hippocampus 21, 461–466.
Yassa, M.A., Stark, C.E.L., 2011. Pattern separation in the hippocampus. Trends in Cognitive Sciences 34, 515–525.
Yonelinas, A.P., Widaman, K., Mungas, D., Reed, B., Weiner, M.W., Chui, H.C., 2007. Memory in the aging brain: doubly dissociating the contribution of the hippocampus and entorhinal cortex. Hippocampus 17, 1134–1140.
Zandi, M.S., Irani, S.R., Follows, G., Moody, A.M., Molyneux, P., Vincent, A., 2009. Limbic encephalitis associated with antibodies to the NMDA receptor in Hodgkin lymphoma. Neurology 73, 2039–2040.
Zola-Morgan, S., Squire, L.R., Ramus, S.J., 1994. Severity of memory impairment in monkeys as a function of locus and extent of damage within the medial temporal lobe memory system. Hippocampus 4, 483–495.

Hippocampus and Related Structures

Menno P Witter, Norwegian University of Science and Technology NTNU, Trondheim, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The hippocampus is part of the cortex of the forebrain in mammals and is closely linked to a restricted number of related cortical areas that are here collectively referred to as the parahippocampal region. This collection of cortical regions, which in humans is located in the medial portion of the temporal lobe, is critically involved in learning and memory processes, in particular in episodic memory. This article aims to summarize the connectional organization of the hippocampal and parahippocampal domains, largely within the context of their relevance for episodic memory.

What Is the Hippocampus and What Are Related Structures?

The hippocampus is a macroscopically defined cortical structure that is present in the brains of all mammals. In all mammalian species the appearance of the hippocampus is alike in that it is an enrolled continuation of the most medial rim of the cortical surface of the hemisphere. The cortices directly adjacent to the hippocampus, which collectively are referred to as the para-(next-to) hippocampal region, are considered as the related structures. The particular arrangement of these different cortical structures can be easily appreciated in sections taken perpendicular to the long axis of the hippocampus (Figure 1).

The hippocampus and the parahippocampal region of different species are differently positioned in the brain, such that in species with a clearly developed temporal lobe, the position of the hippocampus is more ventrally and anteriorly (humans and monkeys) compared to the situation in for example the rat, where the hippocampus looks more like a c-shaped structure positioned in the posterior third of the hemisphere, such that its dorsal one-third has a marked spatial relation with the posterior half of the corpus callosum. However, these differences in position do not alter the major characteristics and the topological relations between the hippocampus and the parahippocampal structures (see below). Nor does it influence the fact that the most anterior/ventral portion of the hippocampus has a close spatial relationship with the amygdala in all mammals.

General Features of the Hippocampus–Parahippocampal Region and Species Differences

The hippocampus is a three-layered cortical structure comprising three major subdivisions, which can be easily recognized in all species (Figure 1). The first field, the dentate gyrus, is a c- or v-shaped structure, which is characterized by a densely packed cell layer, mainly consisting of granular cells; i.e., cells with a round soma, and a prominent apical dendritic tuft and less developed or almost absent basal dendrites. Enclosed in this c-shaped cortical blade is a polymorphic cell group generally referred to as the hilus of the dentate gyrus. The second subfield is the Ammon's Horn or Cornu Ammonis, generally abbreviated as CA-field. The CA-field has been subdivided into CA1, CA2, and CA3. Some authors have opted to differentiate a CA4 area, which is spatially close to the dentate gyrus. Since this is an ill-defined area, this practice will not be followed here. Likewise, no further subdivisions in CA3 and CA1 will be made. The third subdivision of the hippocampus is the subiculum. Note that according to some authors a prosubiculum can be differentiated from the subiculum proper. Also, readers may find texts in which reference is made to the subicular complex, consisting of (pro)subiculum, presubiculum, and parasubiculum. Based on connectional arguments, it is preferred to consider the prosubiculum as part of the subiculum that partially overlaps spatially with the neighboring CA1, while the presubiculum and parasubiculum are functionally different from the subiculum and in this account they are considered as part of the parahippocampal region (Witter, 2010; Witter and Amaral, 2004).

The parahippocampal region, in contrast to the hippocampus, comprises cortical regions that have more than three layers, generally five or six. In all species, the parahippocampal region comprises five main domains, the pre- and parasubiculum, the perirhinal and entorhinal cortices, and a fifth area which in primates is commonly referred to as the parahippocampal cortex. In nonprimates, the most likely homolog for the latter area is the so-called postrhinal cortex (Burwell et al., 1995; Witter et al., 1989). The respective names of some of the areas that make up the parahippocampal region refer to the rhinal sulcus, which can be distinguished in all mammals. However, its particular features differ quite a bit between species. Whereas in rodents it is a distinguishable fissure that runs along the anteroposterior extent of the hemisphere, in humans it is only present in about half of the population, as a shallow and short indentation in the most medial-anterior temporal lobe.

A remark is appropriate concerning the question whether nonmammalian species have a hippocampus and related structures. In reptiles, cortical areas have been described which resemble, both in terms of cell types as well as connectivity, the mammalian dentate gyrus and Ammon's Horn. Part of the olfactory related cortex of reptiles may be considered homologous to the mammalian parahippocampal region (ten Donkelaar, 1997). There is evidence that lesions of these structures in reptiles interfere with learning. With respect to birds, there is convincing evidence that a particular region in the forebrain, homolog to the reptilian cortex, is functionally important for spatial memory. This region is generally referred to as the hippocampus, and although structurally it looks quite different from the mammalian hippocampus, connectionally it is quite similar. Currently, there is no agreement about the

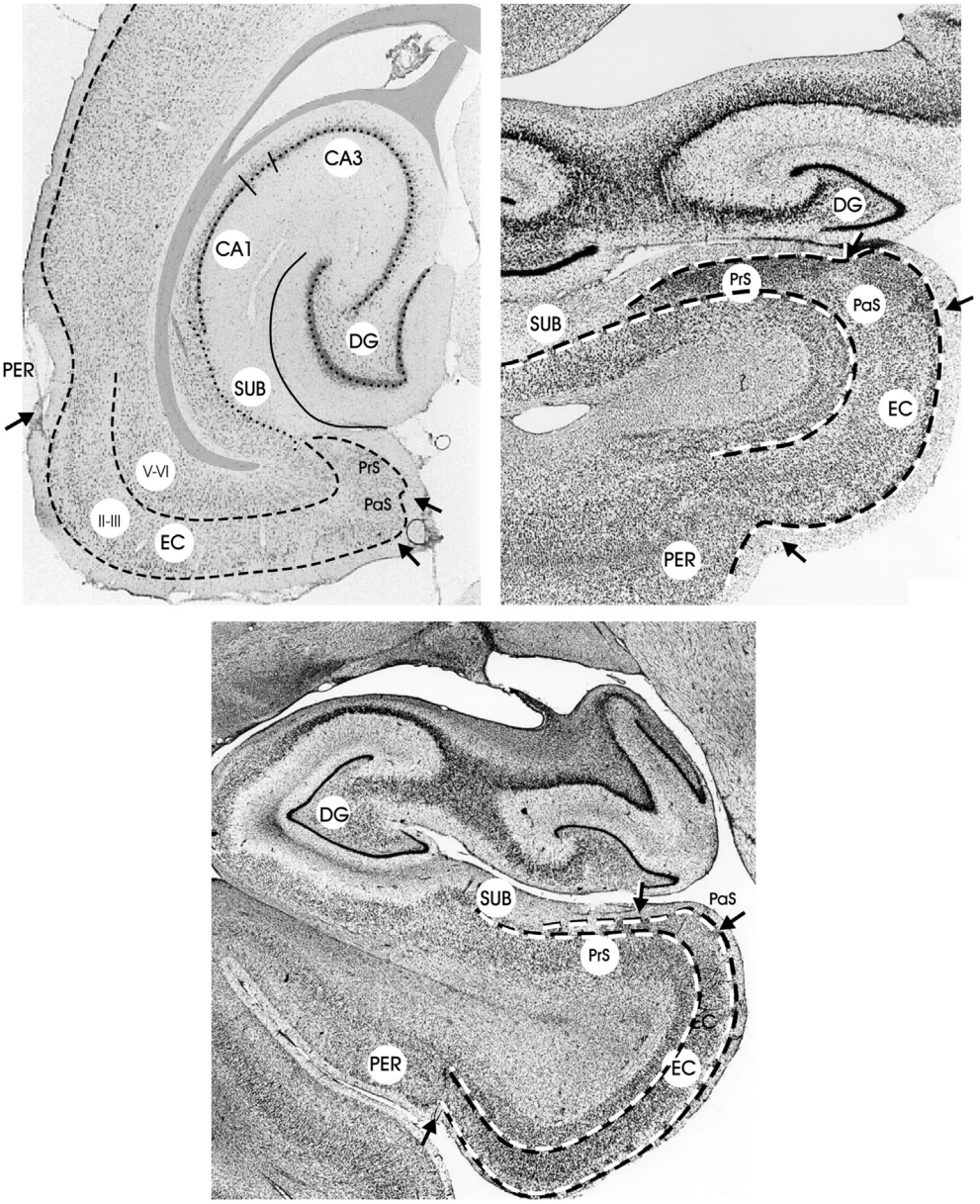

Figure 1 Three sections of the hippocampus and related parahippocampal structures cut perpendicular to the long axis of the hippocampus of the rat (upper left), the monkey (upper right), and the human (lower panel; modified after Scharfman et al., 2000; with permission of The New York Academy of Sciences).

detailed homology with the mammalian hippocampus and the subdivisions therein (Clayton, 1998; Treves et al., 2008). Finally, even in bony fish, a functionally and structurally comparable area has been described located in the most lateral wall of the telencephalic cortical sheet (Rodriguez et al., 2002).

Wiring of the Hippocampus and Parahippocampal Region

Local Circuits

The subdivisions of the hippocampus and the parahippocampus were originally described based on differences in overall features of cell types, the size and staining properties of the neuronal somata and other striking architectural differences. By now it is well established that the most commonly used divisions, as described above, also differ remarkably with respect to local wiring between their excitatory principal cells. There are additional differences in the organization of inhibitory networks that will not be described in detail here. The dentate gyrus (DG) is characterized by a network of granular cells that project massively to CA3 by way of the so-called mossy fiber pathway. These efferent axons also provide a strong collateral local input to the mossy cells in the polymorph layer or hilus, which in turn project back to the granular cells. The network in CA3 features a strong auto-associative network

through which CA3 pyramidal cells provide a major input onto themselves, which is a unique feature among hippocampal and parahippocampal areas. Pyramidal cells further project to CA2 and CA1 through the Schaffer collaterals and send an output to the septal complex (Witter, 2010). The local network in CA2 is yet poorly explored but recent findings consistently point out that the properties of the individual neurons and the network are strikingly different from what has been reported for CA3 and CA1. CA2 neurons may inhibit CA3 and CA1 in a feedback and feedforward manner, respectively (Jones and McHugh, 2011). The principal cells in CA1 are monosynaptically connected and electrically coupled to each other. They are further connected through a complex series of interneurons. The latter feature is not unique to CA1 but the CA1 has been studied in most detail (Chamberland and Topolnik, 2012). The CA1 neurons project to the subiculum, as well as to a number of cortical and subcortical domains. The organization of the subiculum is yet poorly understood but it may share neuronal properties with CA1, and no elaborate associative network has been described (Kim and Spruston, 2012). Together with CA1 it originates most of the hippocampal output (Witter and Amaral, 2004).

The parahippocampal domains are all multilayered and for most of them our knowledge in the intrinsic organization is still rather sparse. While in entorhinal cortex differences between layers with respect to local networks have been reported, published data on intralaminar connectivity are restricted to known connections from layer V to layers II and III and vice versa. Such interconnectivity between deep and superficial layers seem to be less obvious on pre- and parasubiculum. The entorhinal cortex is subdivided into the lateral and medial entorhinal cortex that differ with respect to types of inputs (Figure 2; Canto et al., 2008). The lateral subdivision receives a characteristic input from the perirhinal cortex, which is mediated through an overall inhibitory network of connections (de Curtis and Pare, 2004). In contrast, the medial entorhinal cortex is characterized by inputs from the preparasubiculum as well as from postrhinal and retrosplenial cortices, both of which are largely mediated by way of excitatory networks, although the inputs from the presubiculum do include a relatively well-developed inhibitory component. There is recent evidence that one of the main principal cell types in layer II, the stellate cells in respectively the lateral and medial entorhinal cortex, are different form each other with respect to electrophysiological properties (Canto and Witter, 2012).

Cortico-parahippocampal–hippocampal Circuits

Most detailed information about the connectivity of the system comes from anatomical tracing studies in the rat, although for an ever-increasing number of pathways relevant detailed information is available for the monkey (Suzuki and Amaral, 2004), the mouse (Witter, 2010), as well as a host of other species, including cats, guinea pigs, and bats (Witter et al., 1989). The picture that emerges from all these studies is that this connectivity is rather well conserved, such that a general, nonspecies specific, description may suffice. The hippocampus has two major pathways through which it is connected to the rest of the brain. The first pathway makes

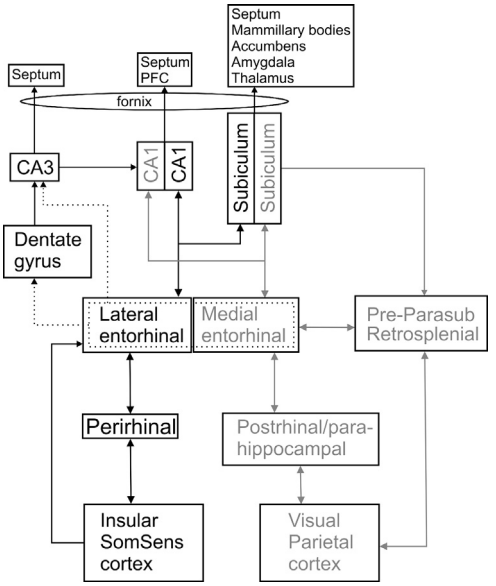

Figure 2 Summary diagram of the connectivity of the hippocampus and parahippocampus. Indicated are the two major communication pathways of the hippocampus by way of the parahippocampus (bottom part of the figure) and the fornix (top part of the figure). Within the parahippocampal–hippocampal network the lateral (black) and medial (gray) parallel pathways are indicated, as well as how these pathways distribute to different components of the hippocampus (converging onto DG and CA3 – dashed lines; targetting different neurons in CA1 and subiculum, which maintain this input selectivity in their interconnectivity – solid black and gray lines).

use of the so-called cortico-parahippocampal–hippocampal circuit, and the second is formed by the fornix (Figure 2).

A striking feature of the cortico-parahippocampal–hippocampal circuit is that the hippocampus does not have direct connections with the majority of the functionally different cortical regions. In contrast, this communication is mediated by way of the parahippocampal region. Of this region, the perirhinal and postrhinal/parahippocampal cortices maintain the highest number of corticocortical connections, including connections with all major sensory realms, such as visual, auditory, somatosensory, and gustatory ones. Also, massive connections exist with the other two major higher-order association domains, the parietal and prefrontal cortices, but connections with the motor cortex are less prevalent. The perirhinal and postrhinal/parahippocampal cortices, in turn, provide a major input to the superficial layers II and III of the entorhinal cortex, which are further reached by inputs from olfactory structures, the amygdala, and from the remaining two components of the parahippocampal region; i.e., the pre- and parasubiculum. Neurons in entorhinal layers II and III, which are the recipients of this wide variety of inputs, give rise to the major cortical input to the hippocampus, the so-called perforant pathway. This name is derived from the traditional descriptions by Ramon Y Cajal (1911) who noted that fibers from the entorhinal cortex perforate the underlying white matter and the cortical lamina of the subiculum, to enter the hippocampus. The perforant pathway projection harbors two different systems. Entorhinal layer II cells distribute their axons to the dentate

gyrus and CA3, whereas layer III cells send their axons to CA1 and the subiculum. In addition to this intriguing difference in target specificity, the terminal distributions of these two components of the perforant pathway are entirely different. Layer II cells give rise to a widespread projection to distributing throughout the transverse axis of the field, likely targeting most of not all of the principal neurons in DG and CA3. In contrast, axons of layer III cells only target restricted groups of the available neurons in CA1 and the subiculum (see below for functional relevance; see also Figure 2). Hippocampal output from CA1 and subiculum is distributed to a large extent by way of projections back to the entorhinal cortex (van Strien et al., 2009), from where ongoing output projections originate to a host of cortical and subcortical structures, including projections to perirhinal and parahippocampal/postrhinal cortices, prefrontal areas, amygdala, ventral striatum, and midline thalamus. Additional outputs reach other cortical areas either by way of the white matter tract underlying the entorhinal cortex and subiculum (angular bundle) or by way of the fornix (Figure 2; see Section The Fornix)

Within the cortico-parahippocampal–hippocampal system at least two parallel pathways have been described, mediated by way of the two main entorhinal subdivisions. One of the pathways carries inputs from the perirhinal cortex and lateral entorhinal cortex into and away from the hippocampus, whereas the second mediates transfer of information from and to the parahippocampal/postrhinal cortex and medial entorhinal cortex (Ranganath and Ritchey, 2012). In view of the anatomical organization of the perforant pathway outlined above, it is likely that in the hippocampus these two pathways converge at the level of the dentate gyrus and CA3, whereas they are kept more or less separate at the level of CA1 and the subiculum (van Strien et al., 2009). The position of the pre- and parasubiculum in the network is still poorly understood. They are almost exclusively reciprocally connected to the medial entorhinal cortex and the anterior nucleus of the thalamus, likely mediating the main connectivity with the vestibular-based head directional input system (Taube, 2007).

The Fornix

The fornix is a major fiber bundle, which connects the hippocampus to the hypothalamus, in particular to the mammillary bodies (Witter and Amaral, 2004). On its way, the fornix also issues fibers to the septal complex, the ventral striatum, and the amygdala. The fornix is not a pure output pathway since projections from the septal complex to the hippocampus (see section Functions of Hippocampus–Fornix Connectivity) and in part to the entorhinal cortex travel by way of the fornix. In addition, the commissural connections between the left and right hippocampi, and CA1/subicular projections to prefrontal cortex also partially travel by way of the fornix. The hippocampal projection to the mammillary bodies is part of the traditionally described limbic or Papez circuit, which includes the mammillo-thalamic tract connecting the mammillary bodies with the anterior complex of the thalamus, which in turn projects to large portions of the limbic cortex, including anterior- and posterior-cingulate cortex, and pre- and parasubiculum. All these structures, in turn, provide input to the hippocampus, either directly or indirectly by way of the entorhinal cortex (Sugar et al., 2011; Vann et al., 2009).

The Hippocampus Is Not a Homogeneous Structure

When one looks at the overall anatomical organization of the parahippocampal–hippocampal system, it appears that the hippocampus is rather homogeneous along its long axis. However, differences in numbers of neurons between the dorsal/posterior and ventral/anterior hippocampus have been reported, indicative for differences in the overall neuronal network (Witter, 2010). Moreover, in the rat and in the monkey, it has been established that the dorsal/posterior portion of the hippocampus receives its major cortical input from the more sensory-related or exteroceptive lateral and caudal portions of the entorhinal cortex, whereas the ventral/anterior hippocampus receives inputs from more medial parts of the entorhinal cortex, which most likely convey information that reflects the interoceptive status of the individual. In addition, direct reciprocal connections with the amygdala are limited to the ventral/anterior and intermediate levels of the hippocampus and are absent for the dorsal/posterior portion of the hippocampus (Witter et al., 1989). Behavioral data in support of this notion are the findings that in rats the dorsal hippocampus is critical for the initial learning and long-term memory of spatial information, whereas the ventral hippocampus is not essential (Moser and Moser, 1998). Recently, brain imaging studies have revealed that in humans also functional differences are present between anterior and posterior halves of the hippocampus; however the precise nature of these differences is still an issue of debate (Ranganath and Ritchey, 2012; Small, 2002).

Functions of the Hippocampus and Parahippocampal Region

Functions of the Cortico-Hippocampal System

Ample clinical and experimental evidence supports the view that the hippocampus and parahippocampal structures are essential for memory processes, although the unique contributions of each of the different subfields are as yet incompletely understood. The results of behavioral studies, in particular, have provided insights into how each of the components of the hippocampal–parahippocampal system contributes to memory processes. Instead of contributing in a unitary way to memory, it appears evident that individual structures make selective contributions. Such studies indicated that the hippocampal formation itself is relevant for processes like pattern separation and pattern completion, both in terms of spatial and nonspatial components and possibly even the timing of the event (Eichenbaum et al., 2012). The hippocampus does not critically contribute to object recognition memory. Instead, it is the laterally positioned component of the parahippocampal region, the perirhinal cortex, which appears to be critically involved in this task (Brown et al., 2010). In contrast, the more posterior portions of the parahippocampal region, the parahippocampal cortex in primates or the postrhinal cortex in nonprimates, may be critically involved in spatial or

topographical memory (Eichenbaum et al., 2012; Ranganath and Ritchey, 2012). Insight in the functional relevance of the entorhinal cortex has increased dramatically over the last decade since the discovery of spatially modulated grid cells in the medial entorhinal cortex (Hafting et al., 2005). Together with place cells in the hippocampus and head direction cells in presubiculum, they form important networks likely serving path integration; i.e., the internal generation of a representation of space (Moser et al., 2008). In the lateral entorhinal cortex, grid cells have not been found, but in contrast cells apparently code for the position and qualitative aspects of objects; i.e., they likely represent space based on external sensory information (Deshmukh and Knierim, 2011).

Functions of Hippocampus–Fornix Connectivity

As indicated above, the fornix provides the hippocampus with another pathway, beside the cortical route, to communicate with a host of subcortical brain structures, including the mammillary bodies (Aggleton et al., 2010). Although the relevance of the fornix with respect to learning and memory has been highly debated, recent evidence clearly indicate that fornix damage does impair certain aspects of learning and memory. It has been proposed that projections of the hippocampus to the mammillary body and to the anterior thalamus are parallel pathways contributing differentially to these processes (Aggleton et al., 2010). This is in line with convincing data from patients with thalamic infarctions pointing to the importance of the connection from the mammillary complex to the anterior complex of the thalamus (Van der Werf et al., 2000).

The fornix also mediates connectivity from the hippocampal–parahippocampal structures to the ventral striatum and the amygdala. In the ventral striatum, interactions may take place between hippocampal, amygdaloid, and prefrontal outputs, leading to the selection of appropriate behavioral strategies; e.g., in learning paradigms (Pennartz et al., 2011). The functional relevance of the reciprocal connections between the amygdala and hippocampus, and for that matter also those between the amygdala and the parahippocampal region is still poorly understood (Pitkanen et al., 2000). It is assumed that amygdala inputs may provide the hippocampal system with a marker for relevance to a particular memory episode.

The connections to and from the septal complex of both the hippocampus and parahippocampal structures use the fornix as well as a number of other white matter pathways. Although much about these interactions has yet to be resolved, it is clear that the normal functioning of the hippocampal–parahippocampal–cortical interactions depend strongly on a functional septal complex. It is in particular the fact that the septal complex supplies the hippocampus and parahippocampal region with most of its cholinergic input, as well as the role of the septal complex in the synchronization of electrical activity, which relates to its functional relevance (Hasselmo, 2006; Pignatelli et al., 2012).

Clinical Relevance

The hippocampal–parahippocampal structures play an established role in learning and memory processes, which is supported by reports concerning pathological changes seen in diseases that are associated with learning and memory disorders such as Alzheimer's disease. However, other neurological–psychiatric disorders are associated with pathological and/or functional alterations in these regions as well (Small et al., 2011). One striking observation is the laminar specificity of the pathology in entorhinal cortex for each of these diseases (Braak and Braak, 1992). In case of Alzheimer's disease, the initial pathology is seen specifically in layer II, in case of epilepsy, the most striking initial pathology is observed in layer III of the medial entorhinal cortex, whereas in case of schizophrenia, the pathology appears to involve a neuronal displacement in layer II.

See also: Alzheimer's Disease, Neural Basis of; Amnesia: General; Autobiographical Memory; Declarative Memory, Neural Basis of; Emotion, Neural Basis of; Episodic Memory; Hippocampal Amnesia; Human Spatial Orientation, Neural Basis of; Learning and Memory, Neural Basis of; Limbic System; Neural Representations of Objects; Retrieval; Spatial Navigation; Temporal Lobe; Vestibular System.

Bibliography

Aggleton, J.P., O'Mara, S.M., Vann, S.D., Wright, N.F., Tsanov, M., Erichsen, J.T., 2010. Hippocampal-anterior thalamic pathways for memory: uncovering a network of direct and indirect actions. European Journal of Neuroscience 31 (12), 2292–2307.

Braak, H., Braak, E., 1992. The human entorhinal cortex: normal morphology and lamina-specific pathology in various diseases. Journal of Neuroscience Research 15 (1–2), 6–31.

Brown, M.W., Warburton, E.C., Aggleton, J.P., 2010. Recognition memory: material, processes, and substrates. Hippocampus 20 (11), 1228–1244.

Burwell, R.D., Witter, M.P., Amaral, D.G., 1995. Perirhinal and postrhinal cortices of the rat: a review of the neuroanatomical literature and comparison with findings from the monkey brain. Hippocampus 5 (5), 390–408.

Canto, C.B., Witter, M.P., 2012. Cellular properties of principal neurons in the rat entorhinal cortex. II. The medial entorhinal cortex. Hippocampus 22 (6), 1277–1299.

Canto, C.B., Wouterlood, F.G., Witter, M.P., 2008. What does the anatomical organization of the entorhinal cortex tell us? Neural Plasma 2008, 381243.

Chamberland, S., Topolnik, L., 2012. Inhibitory control of hippocampal inhibitory neurons. Frontiers in Neuroscience 6, 165.

Clayton, N.S., 1998. Memory and the hippocampus in food-storing birds: a comparative approach. Neuropharmacology 37 (4–5), 441–452.

de Curtis, M., Pare, D., 2004. The rhinal cortices: a wall of inhibition between the neocortex and the hippocampus. Progress in Neurobiology 74 (2), 101–110.

Deshmukh, S.S., Knierim, J.J., 2011. Representation of non-spatial and spatial information in the lateral entorhinal cortex. Frontiers in Behavioral Neuroscience. 5, 69.

Eichenbaum, H., Sauvage, M., Fortin, N., Komorowski, R., Lipton, P., 2012. Towards a functional organization of episodic memory in the medial temporal lobe. Neuroscience & Biobehavioral Reviews 36 (7), 1597–1608.

Hafting, T., Fyhn, M., Molden, S., Moser, M.B., Moser, E.I., 2005. Microstructure of a spatial map in the entorhinal cortex. Nature 436 (7052), 801–806.

Hasselmo, M.E., 2006. The role of acetylcholine in learning and memory. Current Opinion in Neurobiology 16 (6), 710–715.

Jones, M.W., McHugh, T.J., 2011. Updating hippocampal representations: CA2 joins the circuit. Trends in Neuroscience 34 (10), 526–535.

Kim, Y., Spruston, N., 2012. Target-specific output patterns are predicted by the distribution of regular-spiking and bursting pyramidal neurons in the subiculum. Hippocampus 22 (4), 693–706.

Moser, E.I., Kropff, E., Moser, M.B., 2008. Place cells, grid cells, and the brain's spatial representation system. Annual Review of Neuroscience 31, 69–89.

Moser, M.B., Moser, E.I., 1998. Functional differentiation in the hippocampus. Hippocampus 8 (6), 608–619.

Pennartz, C.M., Ito, R., Verschure, P.F., Battaglia, F.P., Robbins, T.W., 2011. The hippocampal-striatal axis in learning, prediction and goal-directed behavior. Trends in Neuroscience 34 (10), 548–559.
Pignatelli, M., Beyeler, A., Leinekugel, X., 2012. Neural circuits underlying the generation of theta oscillations. Journal of Physiology – Paris 106 (3–4), 81–92.
Pitkanen, A., Pikkarainen, M., Nurminen, N., Ylinen, A., 2000. Reciprocal connections between the amygdala and the hippocampal formation, perirhinal cortex, and postrhinal cortex in rat. A review. Annals of the New York Academy of Sciences 911, 369–391.
Ramon Y Cajal, S., 1911. Histologie du Systeme Nerveux de l'Homme et des Vertebres. Maloine, Paris.
Ranganath, C., Ritchey, M., 2012. Two cortical systems for memory-guided behaviour. Nature Reviews in Neuroscience 13 (10), 713–726.
Rodriguez, F., Lopez, J.C., Vargas, J.P., Broglio, C., Gomez, Y., Salas, C., 2002. Spatial memory and hippocampal pallium through vertebrate evolution: insights from reptiles and teleost fish. Brain Research Bulletin 57 (3–4), 499–503.
Scharfman, H.E., Witter, M.P., Schwarcz, R., 2000. The parahippocampal region. Implications for neurological and psychiatric diseases. Introduction. Annals of the New York Academy of Sciences 911, ix-xiii.
Small, S.A., 2002. The longitudinal axis of the hippocampal formation: its anatomy, circuitry, and role in cognitive function. Reviews in Neuroscience 13 (2), 183–194.
Small, S.A., Schobel, S.A., Buxton, R.B., Witter, M.P., Barnes, C.A., 2011. A pathophysiological framework of hippocampal dysfunction in ageing and disease. Nature Reviews in Neuroscience.
Sugar, J., Witter, M.P., van Strien, N.M., Cappaert, N.L., 2011. The retrosplenial cortex: intrinsic connectivity and connections with the (para)hippocampal region in the rat. An interactive connectome. Frontiers in Neuroinformatics 5, 7.
Suzuki, W.A., Amaral, D.G., 2004. Functional neuroanatomy of the medial temporal lobe memory system. Cortex 40 (1), 220–222.
Taube, J.S., 2007. The head direction signal: origins and sensory-motor integration. Annual Reviews in Neuroscience 30, 181–207.
ten Donkelaar, H.J., 1997. Reptiles. In: Nieuwenhuys, R., ten Donkelaar, H.J., Nicholson, C. (Eds.), The Central Nervous System of Vertebrates. Springer, Berlin, pp. 1315–1524.
Treves, A., Tashiro, A., Witter, M.E., Moser, E.I., 2008. What is the mammalian dentate gyrus good for? Neuroscience 154 (4), 1155–1172.
Van der Werf, Y.D., Witter, M.P., Uylings, H.B., Jolles, J., 2000. Neuropsychology of infarctions in the thalamus: a review. Neuropsychologia 38 (5), 613–627.
van Strien, N.M., Cappaert, N.L., Witter, M.P., 2009. The anatomy of memory: an interactive overview of the parahippocampal–hippocampal network. Nature Reviews in Neuroscience 10 (4), 272–282.
Vann, S.D., Aggleton, J.P., Maguire, E.A., 2009. What does the retrosplenial cortex do? Nature Reviews in Neuroscience 10 (11), 792–802.
Witter, M.P., 2010. The hippocampus. In: Puelles, L., Paxinos, G., Watson, C. (Eds.), The Mouse Nervous System. Academic Press, San Diego, CA.
Witter, M.P., Amaral, D.G., 2004. Hippocampal formation. In: Paxinos, G. (Ed.), The Rat Nervous System. Elsevier Academic Press, San Diego, CA, pp. 635–704.
Witter, M.P., Groenewegen, H.J., Lopes da Silva, F.H., Lohman, A.H., 1989. Functional organization of the extrinsic and intrinsic circuitry of the parahippocampal region. Progress in Neurobiology 33 (3), 161–253.

Relevant Websites

www.temporal-lobe.com – Connectivity Database.
www.rbwb.org – Architectural 3D and 3D Database.

Hirschman, Albert O. (1915–2012)

Jeremy Adelman, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article surveys the life and work of Albert O. Hirschman. It chronicles the major intellectual influences and his publications from the 1940s to the 1990s. He contributed to the fields of economic development, history of economic thought, methods in social science theory, and democratic theory. His range reflected a varied and complex life history.

Albert O. Hirschman was one of the many Mitteleuropean intellectuals thrust out of the continent by the spread of fascism and the crisis of the interwar years to contribute to the transformation of the global landscape of the social sciences. It has been a convention to stress the ways in which figures as disparate as Karl Polanyi, Hannah Arendt, and Friedrich Hayek exported continental intellectual traditions to new grounds and expanded on the core they had developed in Budapest, Heidelberg, and Vienna.

Hirschman's career and output were different from many of his peers. Through his multistaged life, he kept evolving, folding in elements and traditions he encountered into the ones he carried with him; each step yielded a new way of thinking and a different set of research questions. Hirschman's style of social science, therefore, represented a compound of influences built up over the course of the twentieth century and shaped by European, American, and Latin American debates. His work also spanned from concerns about balance payments and interstate conflicts in Europe, to the microfoundations of economic development in the Third World, to concerns about collective action, to the intellectual history of capitalism, and meditations on language and democracy. It would be hard to imagine an intellectual who could have intervened so originally in so many big debates. That he was able to do so in large measure reflects Hirschman's ever-evolving intellectual style.

Born in Berlin in 1915, Hirschman grew up in the cultural and political heart of the Weimar Republic. Such a cosmopolitan setting, and the spirit of experimentation and tolerance, left their imprints on the young Hirschman. His parents were members of the flourishing assimilated Jewish bourgeoisie; his father Carl was a surgeon, his mother's (Hedwig) parents belonged to the influential Marcuse clan. They sent him to the French Gymnasium, which emphasized classical learning in math and humanities, especially the tradition of German idealism, as well as languages. As a teenager, Hirschman got drawn into the political maelstrom of the waning Weimar years. His sympathies were with the left wing of the Social Democratic Party, which was locked in a battle with Communists and fascists. But it was more than a political contest for Hirschman; it was also an intellectual cause, as Hirschman was captivated by Marxist political economy. His readings convinced him to study economics at the University of Berlin for one fitful semester in the fall of 1932. Given the electoral agitation at the time, there was hardly a moment to study, however.

Hitler's rise to power interrupted Hirschman's fledgling career. At the same time, his father died of cancer in early 1933. This double blow severed Hirschman's ties to Berlin and he fled in April of that year to Paris. One might say that the end of adolescence closed the door on a boy's dream we might associate with the Weimar Republic, but not entirely. A love of Goethe and a dedication to fight for and understand the cosmopolitan values of his deceased republic accompanied Hirschman through his life.

The flight from Berlin was the first of many; intolerance hounded him from country to country. Someone more vulnerable to embitterment might have considered this trademark of life in the modern age, a sign of decline from the exalted ideas of the Enlightenment. Certainly many other Mitteleuropa exiles spread a shadow of doubt and pessimism over the modern condition. The Frankfurt School is perhaps the most renown. While Hirschman would be buffeted about to become a more broadly European intellectual, schooled in Paris (École des Hautes Études Commerciales), London (the London School of Economics), and Trieste (the University of Trieste), he did not give up hope that the specter of fascism spreading across the continent would eventually roll back. Along the way, Hirschman amalgamated Anglo-American empirical political economy, French interests in geography and statistics, and an important strain of Italian political philosophy associated with liberal socialism. He became increasingly doubtful of the old rigidities of his Marxist ancestry without ever fully disavowing it. These were also years of heady political involvement, which saw Hirschman fight in the Spanish Civil War, participate in the Italian underground, join the French army to thwart the onslaught of the Wehrmacht, and operate a subrosa scheme to help refugees escape from Marseilles in late 1940. An important trait for his social science was the strong connection between action and learning; Hirschman did not believe that scholarship was a separate endeavor from social change. But instead of seeking solutions to perfect the world, Hirschman became more and more aware that his cause was to improve it.

Eventually, the Vichy police came close to chasing him down, and so Hirschman fled once more – this time to the United States. As luck would have it, the Rockefeller Foundation offered him a scholarship to study at Berkeley. He lived there for almost 2 years, got married to Sarah Chapiro, and authored his first book, *National Power and the Structure of Foreign Trade* (University of California Press, 1945). That work distilled many of his economic thoughts that had been incubating especially in Italy and France, to understand the economic dimensions of Germany's expansionist urge. In

examining the role that asymmetrical trade relations played in shaping a foreign policy outlook, Hirschman argued that increasing interdependence between countries of unequal resources and endowments in an age of heightened national security predisposed hegemons to expand and resort to coercive measures. The solution to this structural problem was a change of norms and governing systems that drained sovereignty from national states to supranational cooperative arrangements. *National Power* was innovative. It was a precursor to the field of international political economy and would be picked up again in the 1970s as a foundational text; but at the time it appeared, the world's economists were already fixed on solving postwar problems and not concerned with prewar causes. It was soon forgotten. One aspect of his thinking is worth noting: Hirschman was uncompelled by the debates over Keynesian economics even though he was at ground zero when *The General Theory* appeared on the bookstands in London in 1936. He could not get too excited about rival grand theoretical claims about the self-correcting market. His quarry lay elsewhere: how to fathom the underlying roots of Europe's economic turmoil, concerns that would lead him eventually to the instabilities and disequilibrium of the development process more generally.

After serving on the Italian campaign as a member of the Office of Strategic Services, Hirschman returned to the US and settled with his family in Washington DC. He worked for the Federal Reserve Board in the brain trust that advised the policy makers behind the Marshall Plan. These were not uncreative years, but Hirschman grew increasingly frustrated by the combination of bureaucratic infighting and the polarizing climate that had hard line free marketeers sniping away at US lending to European countries that did not make price stability the primary goal. But it was especially the McCarthy purges of the civil service that sealed Hirschman's fate and sent him packing once more. He did not know it, but the FBI had been trailing Hirschman from the day he arrived at Berkeley. In late 1951, facing unemployment and with a wife and two daughters to support, Hirschman desperately sought help from an old friend who was working at the World Bank (known then as the International Bank for Reconstruction and Development). He learned of a venture the bank was sponsoring to pilot one of the first efforts to invest in development in what would soon be called 'the Third World.' The country was Colombia.

Within months, the Hirschman family and their bags arrived in Bogotá. If Hirschman's first flight led to his Europeanization, and his flight from Europe to his Americanization, this flight led to Hirschman's Latin Americanization. With this phase came his reinvention. Yet, some basic traits of his style were becoming clear. He was no orthodox thinker. He defied categorization. And when times were bleak, it was all the more important to think differently about the source of the problem with an eye to potential remedies. But it was the encounter with the challenges of capitalist development and democracy in Latin America that brought his imagination into relief. In Colombia, he worked not in an ivory tower but as a consultant helping tackle everyday problems of investment in irrigation schemes and housing projects. The World Bank had sponsored a development plan, to be coordinated out the Planning Department of the Central Bank of Colombia, to target investments into sectors and industries that were deemed most likely to pull the country out of its cycle of violence and what was generally regarded as 'backwardness.' Hirschman's job was to go around the country assessing candidates for loans according to a master plan. The day-to-day work was exciting, but the plan was not.

Hirschman chafed. The team of foreign 'experts' worked off a scheme that Hirschman saw as increasingly unhelpful. The urge to follow the script of an abstract model left out a couple of key issues, as far as Hirschman was concerned. The technocrats overlooked much of the dynamism of the private sector on the ground. They put too much stock on their aggregate numbers as a compass for investment decisions. And they scorned local learning, feeling that their expertise and being from the outside made their knowledge untainted and 'objective.' This was absurd, to Hirschman. After 2 years, he was tired of the wrangling in the Central Bank and struck out on his own to be an independent economic consultant, doing similar evaluative work for investors, especially Colombia's various credit agencies.

This was Hirschman's first real exposure to field work, and he became an addict. Indeed, his approach to development economics remained deeply empirical as a result, and not just any empirical style worked for him; he preferred the power of the case study, a method increasingly used and criticized in the social sciences. By evaluating housing projects, pulp and paper plants, road works, Hirschman felt that he could close the gap between the analytical tools and the subject of analysis; he preferred smaller scales and partial insights to large, sweeping causal claims. And he preferred to learn from peoples' actions and behavior.

But there was more than a research style at stake. Hirschman also resisted the kind of certainties that sprung from abstract theories, especially in the emerging field of development economics. Whereas early development economics put the accent on Third World peoples as essentially caught in time, backward, simple folk, who needed outside forces to shake them from their traditions and traps, Hirschman was one of the first scholars to apply a style he borrowed from an old associate from his Berkeley and Washington years, Alexander Gerschenkron. This was the belief that not all societies had to follow the same road toward progress, stages, or phases. Development was not of a piece. There was a second feature to his style, and that was the view that on the ground things were going much better than economists tended to believe. This contrasted with the prevailing view that only external forces were autonomous enough from the lockjaw of tradition to provoke change, as well as the conviction that only concerted all-encompassing plans would work because the forces of tradition would undermine any partial effort and spin the country into disequilibrium. Hirschman thought this was not just wrongheaded; it did not square with empirical findings from the field – and in fact had to override them in order to stay plausible. He found the circular reasoning at odds with the real world of development.

These ideas took final shape when he was offered a two-year fellowship at Yale, thanks in part to the interjection of another old associate, Thomas Schelling. At Yale, he was able to convert his insights into his second book, *The Strategy of Economic Development* (Yale University Press, 1958) one of the masterworks in the history of economic thinking. The book would

catapult Hirschman onto a new career path. But behind its prose was more than learning from the development experience on the ground. Hirschman had been reading a generation of Latin American scholarly political economy that influenced him deeply. He was exposed to a group of economists, in particular, who emphasized the role of deep structures that patterned the nature of Latin American capitalism differently than in so-called developed countries; Latin America was not backwards or *un*developed. It was *under*developed because of its capitalism. No scheme contrived from lessons of the 'developed' world would work very well. Moreover, they emphasized the strains and tensions associated with the development process, and not the goal of seamless, frictionless planning. Development was not just about 'catching up' – a basic precept of standard theory from Europe and North America.

The combination of these Latin American influences shaped *The Strategy of Economic Development*, a book that explicitly offered a counternarrative to elaborate stage theories and the preference for balance and integration. Instead, Hirschman accented the role of disequilibria and tension points in the economic process that compelled people to confront problems and make decisions to solve them. Indeed, a key to the book was that what was scarce in the development process was not capital (which had to be imported) but decision-making capacities (which had to be homegrown). Tensions were not to be avoided; they were not the results of bad policy; they were what good policy should be inducing. In this sense, Hirschman advocated a model of unbalanced growth, not balanced growth. A second aspect of this approach was a concept of linkages, specifically the 'backwards' and 'forwards' activities generated around a specific industry, like mining or fishing. The development process should focus not on overarching concerted plans, but spotlighting sectors and industries with greater linkage effects.

The appearance of *The Strategy of Economic Development*, followed by the Cuban Revolution a year later, put Hirschman in big demand. At middle age, he catapulted into the citadels of American higher education, from Yale to Columbia, Harvard, and finally to the Institute for Advanced Study in Princeton. In the meantime, *Strategy* was the first of what would be four volumes about development. Its sequel turned more explicitly to the study of three case studies of development policy. In emphasizing the centrality of policy over planning, Hirschman was trying to draw attention to the role of states in development, states that need not be treated as just as dragging anchors. With the support of the Twentieth Century Fund, he embarked on a study of agrarian reform policy in Colombia, regional development in Brazil, and inflation and stabilization in Chile. Whereas *Strategy* was based on his Colombia experience without detailing it, this work was explicitly empirical and shaped in concert with a pioneering scholar of policy decision-making processes, Charles Lindblom. Each case chronicled the role of social scientists and 'reform-mongering' policy makers willing to take risks and imagine a different social order without collapsing into defeatism. *Journeys Toward Progress: Studies of Economic Policy-Making in Latin America* (The Twentieth Century Fund, 1963) aimed to show that development was happening on the ground in Latin America, that Latin Americans were not bereft of their own abilities, and highlighted the role that ideology played in mobilizing resources. It was a clarion call at a time in which more and more scholars were impressed by the revolutionary alternative posed by the Cuban Revolution.

The third volume was also empirical, but this time the case studies were many. In 1964–65, Hirschman embarked on a worldwide tour at the behest of the Brookings Institution to evaluate World Bank projects in Asia, Africa, southern Europe, and Latin America. Behind the scenes some of the top brass at the World Bank were keen to have an assessment of a decade of Bank lending. *Development Projects Observed* (The Brookings Institution, 1967) was not exactly what they were hoping for. Instead of concrete evaluations of each project and specific lessons that would amount to best practice recommendations, Hirschman's conclusions were almost lyrical. He sought to 'sing the epic' of development by recounting the ways in which projects – as opposed to strategies and policies of the earlier volumes – could yield to a great adventure. Indeed, he wanted economists to be open to the element of surprise and unintended consequences (which, in his recounting, more often tended to be beneficial than bad), or what he called 'side effects' that were not originally baked into the design and implementation of the project. It was hardly an easy lesson to implement, but it was a plea for open-mindedness in decision making and evaluation. The book did sire one of Hirschman's most famous aphorisms – and Hirschman would soon become renowned for them – 'the hiding hand.' Playing on Adam Smith's notion of invisible, anonymous forces guiding the fortunes of producers and consumers, Hirschman gave it a dialectical twist, to argue that the hiding hand is one that deliberately obscures itself so that people may act; that sometimes knowing the full challenge ahead operates to dissuade people from action; that the anticipated costs of a project had to be deliberately hidden in order to summon its creative energies and unleash its potential. In this fashion, even failures could be good news because they might yield positive side effects, and one can learn from them. By this point, Hirschman's attention to the role of psychological factors and the role of behavior in social change was coming to full fruition.

These books capped an empirical trilogy. Along the way, Hirschman penned a series of increasingly conceptual papers that raised basic questions about the methodological and epistemic features of development thinking, which had him addressing ever broader-gauged issues in the social sciences. These were gathered together in an anthology called *Bias for Hope: Essays on Development and Latin America* (Yale University Press, 1971). The title itself bespoke his concern to resist the rising sense of unease even as the development decade drew to a close and offer alternative ways to think about the connections between politics and economics, while he waded into controversies like foreign aid, industrial policy, and tax schemes. One of the consistent messages was the Latin America's economic difficulties did not represent the end of a model of growth, the futility of reformist policy making, or the pointlessness of favoring certain industrial strategies. More important were the methodological essays. One had him coining the term 'possibilism' (which was really borrowed from Kierkegaard) to denote a specific analytic temperament, one that casts light on what appear to be immutable, stubborn, and impervious obstacles to change, to find options ('Introduction:

Political Economics and Possibilism'). With a little imagination, some lateral thinking and daring, alternatives were possible. "Aren't we interested in what is (barely) possible, rather than what is probable?," he asked himself in his notes as he wrote one of these essays. The ethical compass for the possibilist was a notion of freedom defined, as Hirschman put it, as 'the right to a nonprojected future,' the liberty to explore fates that were not predicted or projected by iron laws of social science. A second injunction was for social scientists to stop obsessing over certainty and prediction – which reminded him of Flaubert's injunction against *la rage de vouloir conclure* as sure to lead us to dead ends and foreclosed outcomes. Maybe we should be more humble and more hopeful, he mused. Instead, social scientists should understand complexity of human action and the unpredictability of its effects. Third, he staked his case for the idea of reform in more fulsome terms. In a blistering attack on how economists harped on about the entrenched obstacles to change and the need for new 'paradigms,' Hirschman applauded the effort to change; what was important about Sisyphus's seemingly endless labors was that he tackle them with a smile on his face ('Obstacles to Development: A Classification and a Quasi-Vanishing Act'). This was always a more appealing approach than the mindless overconfidence in the solvability of all problems or its twin, the fatalism that nothing can be changed willfully at all.

Still, it was hard to deny that development was running into disasters. Hirschman had to confront them first hand. When the Biafran Civil War erupted in Nigeria, Hirschman lacerated himself for not having anticipated this side effect in his evaluation of a World Bank-funded railway project a few years earlier. He had sensed ethnic and regional tensions, but he never imagined things could go so awry; here was a failure of imagination. In response to this war, and to the rising agitation against the US escalation of the Vietnam War, the rise of the civil rights movement, the consumer right movement, he wrote a very quick little book that would make him one of the most celebrated social scientists on the planet. When *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States* (Harvard University Press, 1970) appeared, it was clear that he had struck a nerve. No book so effectively explained to its readers how and why people responded to the deterioration of the private and public institutions. He mixed together social psychology, consumer behavior theory, and political science to illuminate how people wrestled with rival internal propensities at any given time. Just as the development process was a complex one, Hirschman was now arguing for a complex notion of individual and social behavior – an approach that would poise him in a conflict with a subsequent drive to assert the behavioral primacy of 'the rational actor.' At the time, many people read the book as charting human responses as alternatives to each other, as contradictions. But this was not Hirschman's argument. Rather, what fascinated him was the way in which voice, the 'political' response, and exit, the 'economic' reaction, might even supplement or complement each other. Wielding the exit option, say threatening to use school vouchers, could give more power to the voice of school reformers. Other times, exit had deleterious effects on voice, as in the case of Nigerian railway managers, who's response to shippers leaving price-gouging railway rates was to hire truckers, and along the way to aggravate internecine ethnic rivalries. By the end of the book, Hirschman himself was warning that the practitioners of voice had to keep their craft alive lest people succumb to too much exit. Either way, the very wording exit, voice, and loyalty captivated readers more often than some of the subtleties of the book's content. This was perhaps an instance of the cover inviting premature judgment.

There was a way in which Hirschman's fame in the 1960s was a function of the decade's underlying hopefulness. As the 1970s unfolded, and the mood turned decidedly bleaker, Hirschman found it increasingly difficult to hang on to his optimism. But he did try to understand the turn; indeed disenchantment and disappointment became subjects of his analysis. One of his most famous essays took aim at the way in which economists and political scientists invoked social disparities to account for the discontent. Called 'The Changing Tolerance for Income Inequality in the Course of Economic Development,' *Quarterly Journal of Economics*, 87 (November 1973: pp. 544–565), it drew attention to subjective and perceptive forces in determining whether people deferred to or protested widening gaps. Along the way, he coined another famous term, 'the tunnel effect,' to denote the ways in which peoples' moods could shift from elation to outrage without being objectively worse off – or vice versa.

But he was not blind to the shift around him. It had an important effect on his work. Hirschman made several decisive choices. One was to leave Harvard to become a permanent faculty member at the Institute for Advanced Study in Princeton, where he would join with the anthropologist Clifford Geertz in trying to chart a course of social science that would differ from the mainstream drift of the academic disciplines.

Hirschman's subjects also made two turns. The first was, as he put it, 'to retreat to history.' In the face of a rising 'neoconservative' tide, especially in economic thinking associated with free market ideology and precepts about the futility of collective action, Hirschman went back in time to examine the origins of thinking about capitalism. What he aimed to reveal was a general appreciation that, on the eve of capitalism's triumph, of a more complex sense of the self, one that time narrowed into the 1970s version of *homo economicus*. There was also a more nuanced understanding of the mutual dependence of the state and market – not an intrinsic hostility. What came of this was a luminescent book about the history of thinking about capitalism, *The Passions and the Interests: Political Arguments for Capitalism before Its Triumph* (Princeton University Press, 1977), which insisted that there were alternative ways of thinking about markets and politics, ways that were more humane, more creative – and ultimately more liberating than the schema advanced by critics and defenders of capitalism in the 1970s. This book is generally considered a milestone in the history of economic thought. It also set the stage for a series of essays he would later write in the 1980s that would draw upon the historical arguments to intervene in contemporary debates. These essays argued for a more complex notion of the economy itself, one that would admit multiple rationalities, a sharper connection between politics and economics, and an eloquent defense of the moral responsibilities of social scientists. These essays were gathered together in two important anthologies, *Essays in Trespassing: Economics to Politics and Beyond* (Cambridge University Press, 1981) and *Rival Views of Market Society and Other Recent Essays* (Viking, 1986).

The second response to the malaise of the 1970s was to return to the field of development in Latin America. This was, in some senses, a very odd decision. After all, matters had gotten decidedly worse. Salvador Allende was overthrown in Chile; the Brazilian military turned the screws on civilians; the Mexican miracle was going sour after the massacre of 1968, and the Argentine coup of 1976 gave new meaning to martial brutality. It seemed, moreover, that the old model of industrialization was indeed finally spent. Yet, it was precisely these mounting challenges that compelled him to write essays that questioned whether Latin American economies were 'spent' and whether the authoritarian turn was so 'determined' by the economic crises. He was resolved to shake off heavily structuralist thinking and underlying pessimism in Latin America. What was unique about many of these essays was how they emerged from extended conversation and collaboration with a younger cohort of Latin American social scientists keen to throw off the strong Marxisant leanings but without abandoning the interest in class relations or political economy. These included the Argentine Guillermo O'Donnell and Brazilian Fernando Henrique Cardoso. But just as the retreat to history admitted a historic shift, Hirschman also had to confess that his original field, development economics, had also run aground. His 'Decline and Fall of Development Economics,' a masterpiece of living intellectual history, was perhaps his most pessimistic publication. These essays would also reappear anthologized in *Essays in Trespassing*. The return to Latin America culminated in his last field research project conducted in 1982–83, when Hirschman evaluated several dozen grassroots projects funded by the Inter-American Foundation. Based on extensive interviews across the region, he showed that the poor did organize collectively to pursue common goals, and sometimes transformed their lives as a result – this often in a context of dire economic circumstances in the midst of a debt crisis and dictatorship. His least well-known book, *Getting Ahead Collectively: Grassroots Experiences in Latin America* (Pergamon Press, 1984) was bereft of much theory. But it did anticipate a turn to the study of social movements, bottom-up development, and microfinance.

One might argue that Hirschman then had two tracks – one about development and its macro causes and consequences, and one a concern with microfoundations of behavior in the marketplace and politics. Ambitiously, Hirschman sought to bring these two sides together in a series of lectures delivered at Princeton University in 1979 and later published as *Shifting Involvements: Private Interest and Public Action* (Princeton University Press, 1982). That book took elements of *Exit, Voice and Loyalty*, especially the observation that peoples' responses to circumstances oscillated according to diminishing returns to their consumption, civic, and political habits. It also borrowed from his development thinking about democracy and capitalism. What it argued was that people were both pursuing private and collective aims and would swing in pendular fashion between them depending on their relative rewards, and that political regimes, dictatorships as well as democracies, were shaped by and shaped these strategies. *Shifting Involvements* was possibly Hirschman's most ambitious book even though it was very short 134 pages. It was also most severely reviewed, very often because Hirschman's essayistic style had to gloss over big thorny issues to which social scientists had devoted a great deal of attention. But it did enjoy a better reception in Latin America and in Europe, where social scientists were more anxious to draw out similar connections, even if some of Hirschman's causal connections were not so clear.

There was a seam in Hirschman's narrative style: the delicate equipoise between impassioned observation and critical involvement. Here was knowledge aimed at changing understandings of the world. The readerly experience of a Hirschman book or essay was intended to be one that destabilized the commonsensical and the orthodox. Whether it was the gurus of 'balanced growth' in the 1960s or the Milton Friedmanite zealots of the 1980s, Hirschman's purpose was to challenge closed certainties. Nor did he spare his colleagues on the Left. He was a sympathetic skeptic of the hardened 'dependency' theories or 'structuralist' explanations for the troubles in Latin America. It is often forgotten that in his famous book exposing the wordplay of what we now call 'neoliberal' apostles in *Rhetoric of Reaction: Perversity, Futility, Jeopardy* (Harvard University Press, 1991), one chapter was devoted to progressive forms of intransigence.

That book also brought to fruition many years of thinking about the role of discourse in politics. In a sense, Hirschman was always aware of the power of words and their ties to ideology and perceptions. But as neoconservatives grew increasingly triumphal in the 1980s, Hirschman grew more concerned that they were also closing themselves off from the prospects of being wrong or from learning anew. He was as ever always concern with the prospects for reform and improvement. Hence the laser beam on 'reaction,' which was, he charted in the book, a semantic hostility to the very idea of reform and deliberate improvability. He had seen some of the consequences unfold in Germany. It was this concern about the democratic fabric that compelled him first to write an essay for *The Atlantic Monthly* and then a complete book on the subject. In it, he coined three word games that intransigents play – futility (that all reforms were doomed to fail anyway), jeopardy (that deliberate reforms might undermine improvements), and perversity (that deliberate reforms might make things worse).

As his career drew to a close, Hirschman returned more and more often to Europe. His work was being translated into French and German, his essays circulating widely around the continent, where he was also being showered with honors and awards. Poignantly, he began to return with more frequency to Berlin, and was there through much of the transition in 1989–90, to watch his natal city reunited and democracy restored. In a way, he felt his hopefulness redeemed. The events spawned a series of reflections on the importance of humility and revisiting one's concepts, a principle he called 'self-subversion.' His last major published essay was called 'Social Conflicts as Pillars of Democratic Market Society,' *Political Theory*, 22 (May 1994), pp. 203–218. It warned that, in a reunified post-1989 Germany, citizens should not rush to serve democracy with shared, community values or norms to ease the frictions and tensions of pluralism. Nothing would be more pernicious – and one can almost hear the voice of a century's watching Germans wrestle with the challenge – than to deal with conflict by reaching for national or collective ideals. Just as he had once told development economists

to treat tension and conflict as the requisites of economic change, he was now urging Europeans to see social conflict as a premise for democratic life.

Hirschman was a master at turning familiar phrases and common coinages on their heads precisely to lure his reader into discovering that realities were not as fixed as they seemed. This was one of the reasons why Hirschman was an early trespasser into psychology and was always eager to loosen the constraints of predictive thinking (in mainstream American social science) or heavy structuralism in development thought. Belatedly, he is now being recognized as one of the founders of a more expanded behavioral social science and for this reason the subject of a revival after many years of disdain from disciplinary purists. Many people wondered why Hirschman did not win a Nobel Prize for economics, which *The Economist* noted, in its obituary to him, he richly deserved. The fact is, Hirschman defied the familiar ways in which the academy and its disciplines organized its intellectual fiefdoms. What made him so original was that he emerged from the margins of the university and so was never truly of the university. This freed him to cross boundaries with great abandon. But it also made him very hard to categorize, even though he always considered himself at heart an economist.

See also: Capitalism; Culture and Economic Development; Economic Growth; Economics and Ethics; Ethical Turn in International Development; Global Development: Markets and Wellbeing; Industrialization; Political Economy, History of; South America: Sociocultural Aspects; Sustainable Development: An Economic Perspective.

Bibliography

A Bias for Hope: Essays on Development and Latin America, 1971. Yale University Press, New Haven, CT. Reprinted by Westview Press (1985).

A Propensity to Self-Subversion, 1995. Harvard University Press, Cambridge, MA.

Crossing Boundaries: Selected Writings and an Interview, 1998. Zone Books, New York.

Development Projects Observed, 1967. Brookings Institution, Washington, DC. Reedited, with a new preface by the author, by the Brookings Institution (1995).

Essays in Trespassing: Economics to Politics and Beyond, 1981. Cambridge University Press, New York.

Exit, Voice, and Loyalty: Responses to Decline in Firms. Organizations and States, 1970. Harvard University Press, Cambridge, MA.

Getting Ahead Collectively: Grassroots Experiences in Latin America, 1984. Pergamon Press, New York.

Journeys toward Progress: Studies of Economic Policy-Making in Latin America, 1963. Twentieth Century Fund, New York. Reprinted by W.W. Norton with a new preface (1973).

National Power and the Structure of Foreign Trade, 1945. University of California Press, Berkeley and Los Angeles. Reprinted 1969; paperback edition with new introduction, 1980.

Rival Views of Market Society and Other Recent Essays, 1986. Viking/Penguin, New York. Paperback edition, with new preface, Harvard University Press (1992).

Shifting Involvements: Private Interest and Public Action, 1982. Princeton University Press, Princeton, NJ.

The Passions and the Interests: Political Arguments for Capitalism Before Its Triumph, 1977. Princeton University Press, Princeton, NJ. Princeton Classics Edition, 2013, with a foreword by Amartya Sen and a second preface by the author and Afterword by Jeremy Adelman.

The Rhetoric of Reaction: Perversity, Futility, Jeopardy, 1991. Belknap Press of Harvard University Press, Cambridge, MA.

The Strategy of Economic Development, 1958. Yale University Press, New Haven, CT. Reprinted 1978 by W.W. Norton; in 1988 by Westview Press.

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 11

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser—Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 11

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 11 *xliii*

VOLUME 11

Historical Archaeology 1
Sarah Croucher

Historical Demography 5
Osamu Saito

Historical Explanation, Theories of: Philosophical Aspects 10
C Behan McCullagh

Historical Geography 17
Gordon M Winder

Historical Linguistics: Numerical Methods 23
Sheila M Embleton

Historical Sociology 27
Nicholas Hoover Wilson and Julia Adams

Historical Thought and Historiography: Christian Tradition 31
Christoph Markschies

Historical Thought and Historiography: Classical Period (Especially Greece and Rome) 35
Wilfried Nippel

Historical Thought and Historiography: Current Trends 39
Georg G Iggers, Supriya Mukherjee, and Qingjia E Wang

Historical Thought and Historiography: East Asia 48
Masayuki Sato

Historical Thought and Historiography: Indigenous Cultures in the Americas 54
Berthold Riese

Historical Thought and Historiography: Islamic Tradition 61
Richard S Humphreys

Historical Thought and Historiography: Jewish Tradition 67
M Brenner

Historical Thought and Historiography: Modern History (Since the 18th Century) — the West 72
Georg Gerson Iggers

Historical Thought and Historiography: South Asia 78
T Raychaudhuri

Historical Thought and Historiography: Southeast Asia 82
Anthony Reid

Historical Thought and Historiography: Sub-Saharan Africa 89
M Diawara

Historicism 93
Donald R Kelley

History and Literature 97
Paul M Lützeler

History and the Social Sciences 102
W Spohn

History from Below, the History of Everyday Life, and Microhistory 108
Andrew I Port

History: Overview 114
Jörn Rüsen

History, Professionalization of 120
Rolf Torstendahl

History: Representation Within and Functions of 125
Chris Lorenz

History: Theories and Methods 131
Chris Lorenz

HIV Risk Interventions 138
JaNelle M Ricks, Ralph J DiClemente, and Puja Seth

Hochschild, Arlie (1940—) 145
Anne Cleary

Holocaust, The 151
Saul Friedländer

Home, Anthropology of 155
Thomas Hauschild

Homelessness in the United States 159
Martha R Burt

Homicide 163
Wendy C Regoeczi and Terance D Miethe

Honesty and Humility 169
Ingo Zettler and Benjamin E Hilbig

Honneth, Axel (1948–) 175
Simon Thompson

Honor and Shame 181
Charles Stewart

Honor Killing: Multidimensional and Multilevel Perspectives 185
Tenzin Dorjee and Stella Ting-Toomey

Hooking up and Casual Sex 192
Paula England and Shelly Ronen

Hope and Hopelessness 197
Eddie MW Tong

Horticultural Societies 202
Paul Sillitoe

Hospitals and Health: Psychosocial Aspects 208
Jeffrey A Alexander

Hotelling, Harold (1895–1973) 215
Ingram Olkin and Allan R Sampson

Household Archaeology 219
Ruth Tringham

Household Economics 224
Shoshana Grossbard

Households in Anthropology 228
Sylvia J Yanagisako

Housing and Cities 233
Ray Forrest

Housing Economics 239
Richard Arnott

Human Behavioral Ecology 246
Raymond Hames

Human Cognition, Evolution of 252
Harry Heft

Human Cooperation, Evolution of 259
Sarah Mathew

Human Development and Health 267
Linda K George

Human Development, Successful: Psychological Conceptions 271
Werner Greve

Human Development, Theories of 276
Richard M Lerner, Rachel M Hershberg, Lacey J Hilliard, and Sara K Johnson

Human Ecology 283
Dudley L Poston, Jr

Human Evolutionary Genetics 289
Pascale Gerbault and Mark G Thomas

Human Factors and Ergonomics 297
Maureen Hassall, Tania Xiao, Penelope Sanderson, and Andrew Neal

Human Genome Diversity Project: History 306
Joanna Radin

Human Genome Project: History and Assessment 311
Hub Zwart

Human Genomics in Asia 318
Haidan Chen and Tikki Pang

Human Language, Evolution of 323
Josef P Rauschecker

Human Language: Evolutionary Precursors 329
Yen Ying Lim and Peter J Snyder

Human Mate Choice 335
Yasuo Ihara

Human Mate Choice, Evolution of 340
Judith A Easton, Cari D Goetz, and David M Buss

Human Resource Management, Psychology of 348
Wayne F Cascio

Human Rights and Social Work 353
Elisabeth Reichert

Human Rights, Anthropology of 360
Mark Goodale

Human Rights, History of 367
Antoon De Baets

Human Rights: Legal Aspects 375
Jonathan Klaaren

Human Sciences, History of 380
Stephen P Turner

Human Spatial Orientation, Neural Basis of 386
Aiden E Arnold, Irene Liu, Kara Murias, Edward Slone, Ford Burles, Veronica Guadagni, and Giuseppe Iaria

Human Terrain System (United States): Critique 392
Maximilian C Forte

Human–Computer Interfaces 400
J May

Human–Environment Interactions: Case Studies 405
CG Knight

Hume, David (1711–76) 410
John P Wright

Humor, Anthropology of 416
Henk Driessen

Hunger and Eating, Neural Basis of 420
Neil E Rowland

Hunter-Gatherer Societies, Archaeology of 423
Vivian Scheinsohn

Hunting and Gathering Societies: Anthropology 428
Nurit Bird-David

Huntington Samuel P. (1927–2008) 432
Andrew Tsz Wan Hung

Hurst, James Willard (1910–97) 438
Aviam Soifer

Husserl, Edmund (1859–1938) 441
Christian Thiel

Hutterite Demography 445
Douglas L Anderton

Hybridity 448
Nestor G Canclini

Hypertension: Psychosocial Aspects 453
Dmitry M Davydov and David Shapiro

Hypnosis 458
Burkhard Peter

Hypnosis, Neural Basis of 465
Victoria Holec and Ian Q Whishaw

Hypnosis, Psychiatry of 472
Nicholas A Covino

Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of 476
Yvonne Kuepper

Hypothalamus 482
Phillip Grant

Hypothesis Testing in Statistics 491
Roger L Berger

Hypothesis Testing: Methodology and Limitations 494
TAB Snijders

Hypothesis Tests, Multiplicity of 500
Lyle V Jones, Charles Lewis, and John W Tukey

Hysteria: History and Critiques 506
Santosh K Chaturvedi and Soumya Parameshwaran

Iconicity in Linguistics 512
John Haiman

Idealization, Abstraction, and Ideal Types 516
Steve Clarke

Identity and Identification 521
Akeel Bilgrami

Identity and Identification, Social Psychology of 526
Nick Hopkins

Identity in Anthropology 532
TM Luhrmann

Identity in Childhood and Adolescence 537
Jane Kroger

Identity Movements 543
Simon Langlois

Identity Offline and Online 547
Rob Cover and Steven Doak

Ideological Constraint: History and Current Status of the Concept 554
John P Connolly

Ideologies, Institutions, and the New Institutionalism 561
Dirk Sauerland

Ideology, Sociology of 573
Michel Dubois

Ignorance, History of Concept 579
Barbara Hoenig

Illich, Ivan (1926–2002) 584
Chen Hon-Fai

Illiteracy, Sociology of 589
Narayana Jayaram

Illness and Dyadic Coping 593
Guy Bodenmann

Illness Behavior and Care-Seeking 596
Howard Leventhal, Jessica Yu, and Elaine A Leventhal

Imaging Methods in Clinical Psychology 603
Sara V Tupak and Thomas Straube

Imaging Methods in Psychotherapy 612
Sara V Tupak and Thomas Straube

Imitation during Infancy and Early Childhood 621
Rebecca A Williamson

Immigration and Crime in the United States 628
Richard Stansfield

Immigration, Applied Research on 635
Maria E Enchautegui

Immigration, Citizenship, and Integration: Social Work Connections 641
Ransford Danso

Immigration, Emigration, and Citizenship 648
David Cook-Martín

Immigration: Health and Mortality 654
Robert A Hummer, Jennifer E Melvin, and Monica He

Immigration Policy 662
Sofia A Perez

Immigration: Political Aspects 669
Christian Joppke

Immigration: Social Psychological Aspects 675
Eva GT Green, Nicole Fasel, and Oriane Sarrasin

Imperialism 682
David A Lake

Imperialism, History of 685
Simon C Smith

Implementation: Political 692
Eugene Bardach

Implementation Science 695
Dean Fixsen, Karen Blase, Allison Metz, and Melissa Van Dyke

Implicit Association Test 703
Sarah Teige-Mocigemba and Karl Christoph Klauer

Implicit Memory 709
Chad J Marsolek

Implicit Social Cognition 714
Adam Hahn and Bertram Gawronski

Imprisonment: Sociological Aspects 721
Tanya N Whittle

Incest, Inbreeding, and Their Consequences 725
Alan H Bittles

Incest Prohibition, Origin and Evolution of 730
Arthur P Wolf

Income Distribution and Inequality: Measurement Issues 732
David S Johnson and Timothy M Smeeding

Income Inequality 739
Dennis Gilbert

Income Taxes 744
Agnar Sandmo

Indicator: Methodology 750
Kenneth A Bollen and Shawn Bauldry

Indigeneity 755
Jeremy Beckett

Indigenous Child Welfare Practice: The Canadian Case 762
Naadli T Ormiston and Kundoqk J Green

Indigenous Knowledge and Intellectual Property Rights 769
Jane E Anderson

Indigenous Knowledge: Science and Technology Studies 779
Kavita S Philip

Indigenous Management Styles 784
Charmine EJ Härtel

Indigenous Psychology 788
Rogelia Pe-Pua

Indigenous Rights 795
Jacquelyn A Jampolsky and Kristen A Carpenter

Indigenous Social Work 804
Michael A Hart

Individualism versus Collectivism: Philosophical Aspects 811
Philip Pettit

Individuals in Society: History of the Concept 816
Ron Harré

Indo-European Languages 820
Philip Baldi and Pierluigi Cuzzolin

Industrial Cluster Dynamics 828
Max-Peter Menzel

Industrial Dynamics: Patenting Perspectives 834
Mei-Chih Hu

Industrial Ecology 843
Reid Lifset and Thomas E Graedel

Industrial Geography 854
Amy K Glasmeier and Christa Lee-Chuvala

Industrial Relations, History of 860
Patrick Fridenson

Industrial Societies 863
Alf Hornborg

Industrialization 868
Richard Biernacki

Industrialization, Typologies and History of 872
PK O'Brien

Industrial–Organizational Psychology: Science and Practice 879
Wayne F Cascio

Inequality Analysis: Overview 885
Guillermina Jasso

Inequality: Comparative Aspects 894
Gerald D Berreman

Inequality Measurement 899
Frank A Cowell

Inequality, Social 908
Douglas S Massey

CONTRIBUTORS TO VOLUME 11

Julia Adams
Yale University, CT, USA

Jeffrey A. Alexander
School of Public Health, University of Michigan, Ann Arbor, MI, USA

Jane E. Anderson
Department of Anthropology and Museum Studies, New York University, NY, USA; and Institute for Intellectual Property and Social Justice, Howard University of Law, Washington DC, USA

Douglas L. Anderton
University of South Carolina, Columbia, SC, USA

Aiden E. Arnold
University of Calgary, Calgary, AB, Canada

Richard Arnott
University of California Riverside, Riverside, CA, USA

Philip Baldi
Penn State University, University Park, PA, USA

Eugene Bardach
Goldman School of Public Policy, University of California at Berkeley, Berkeley, CA, USA

Shawn Bauldry
University of Alabama, Birmingham, AL, USA

Jeremy Beckett
University of Sydney, Pyrmont, NSW, Australia

Roger L. Berger
School of Mathematical and Natural Sciences, Arizona State University, Phoenix, AZ, USA

Gerald D. Berreman[†]
University of California, Berkeley, CA, USA

Richard Biernacki
University of California, San Diego, CA, USA

Akeel Bilgrami
Columbia University, New York, NY, USA

Nurit Bird-David
University of Haifa, Haifa, Israel

Alan H. Bittles
Duke University, Durham, NC, USA; Murdoch University, Perth, WA, Australia; and Edith Cowan University, Perth, WA, Australia

Karen Blase
FPG Child Development Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Guy Bodenmann
University of Zurich, Zürich, Switzerland

Kenneth A. Bollen
University of North Carolina, Chapel Hill, NC, USA

M. Brenner
University of Munich, Munich, Germany

Ford Burles
University of Calgary, Calgary, AB, Canada

Martha R. Burt
Urban Institute, Washington, DC, USA

David M. Buss
The University of Texas at Austin, Austin, TX, USA

Nestor G. Canclini
Autonomous Metropolitan University, Mexico and National System of Researchers, Mexico

Kristen A. Carpenter
University of Colorado Law School, Boulder, CO, USA

Wayne F. Cascio
University of Colorado Denver, Denver, CO, USA

Santosh K. Chaturvedi
National Institute of Mental Health and Neurosciences, Bangalore, India

[†] Deceased.

Haidan Chen
College of Humanities and Development Studies, China Agricultural University, Beijing, China

Steve Clarke
Charles Sturt University, Sydney, NSW, Australia; and The University of Oxford, Oxford, UK

Anne Cleary
School of Sociology, University College Dublin, Dublin, Ireland

John P. Connolly
The University of Texas at Arlington, Arlington, TX, USA

David Cook-Martín
Grinnell College, Grinnell, IA, USA

Rob Cover
School of Social Sciences M257, The University of Western Australia, Crawley, WA, Australia

Nicholas A. Covino
Massachusetts School of Professional Psychology, Boston, MA, USA

Frank A. Cowell
London School of Economics and Political Science, London, UK

Sarah Croucher
Wesleyan University, Middletown, CT, USA

Pierluigi Cuzzolin
Università degli Studi di Bergamo, Bergamo, Italy

Ransford Danso
School of Social Work, Ryerson University, Toronto, ON, Canada

Dmitry M. Davydov
Russian Institute for Advanced Study, Sholokhov Moscow State University for the Humanities, Moscow, Russia; and Institute of General Pathology and Pathophysiology, Russian Academy of Medical Sciences (RAMS), Moscow, Russia

Antoon De Baets
University of Groningen, Groningen, The Netherlands

M. Diawara
Center for Research on Local Knowledge, Bamako, Mali

Ralph J. DiClemente
Rollins School of Public Health, Emory University, Atlanta, GA, USA

Steven Doak
School of Social Sciences M257, The University of Western Australia, Crawley, WA, Australia

Tenzin Dorjee
California State University Fullerton, Fullerton, CA, USA

Henk Driessen
Radboud University Nijmegen, Nijmegen, The Netherlands

Michel Dubois
GEMASS — CNRS, Paris, France

Judith A. Easton
Texas State University, San Marcos, TX, USA

Sheila M. Embleton
York University, Toronto, ON, Canada

Maria E. Enchautegui
Urban Institute, Washington, DC, USA

Paula England
New York University, New York City, NY, USA

Nicole Fasel
University of Lausanne, Lausanne, Switzerland

Dean Fixsen
FPG Child Development Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Ray Forrest
City University of Hong Kong, Kowloon Tong, Hong Kong

Maximilian C. Forte
Concordia University, Montreal, QC, Canada

Patrick Fridenson
EHESS, Paris, France

Saul Friedländer
University of California, Los Angeles, CA, USA

Bertram Gawronski
University of Texas at Austin, Austin, TX, USA

Linda K. George
Duke University, Durham, NC, USA

Pascale Gerbault
Research Department of Genetics, Evolution and Environment, University College London, London, UK

Dennis Gilbert
Hamilton College, Clinton, NY, USA

Amy K. Glasmeier
Massachusetts Institute of Technology, Cambridge, MA, USA

Cari D. Goetz
California State University, San Bernadino, CA, USA

Mark Goodale
George Mason University, Burke, VA, USA

Thomas E. Graedel
School of Forestry & Environmental Studies, New Haven, CT, USA

Phillip Grant
Justus-Liebig-University Giessen, Giessen, Germany

Eva G.T. Green
University of Lausanne, Lausanne, Switzerland

Kundoqk J. Green
School of Social Work, University of Victoria, Victoria, BC, Canada

Werner Greve
University of Hildesheim, Hildesheim, Germany

Shoshana Grossbard
San Diego State University, San Diego, CA, USA

Veronica Guadagni
University of Calgary, Calgary, AB, Canada

Adam Hahn
University of Cologne (Universität zu Köln), Köln, Germany

John Haiman
Macalester College, St Paul, MN, USA

Raymond Hames
University of Nebraska, Lincoln, NE, USA

Ron Harré
University of Oxford, Oxford, UK; and Georgetown University, Washington, DC, USA

Michael A. Hart
University of Manitoba, Winnipeg, MB, Canada

Charmine E.J. Härtel
UQ Business School, The University of Queensland, Brisbane, QLD, Australia

Maureen Hassall
The University of Queensland, St Lucia, QLD, Australia

Thomas Hauschild
Eberhard-Karls University, Tübingen, Germany

Monica He
University of Pennsylvania, Philadelphia, PA, USA

Harry Heft
Denison University, Granville, OH, USA

Rachel M. Hershberg
Institute for Applied Research in Youth Development, Tufts University, Medford, MA, USA

Benjamin E. Hilbig
School of Social Sciences, University of Mannheim, Mannheim, Germany; and Max Planck Institute for Research on Collective Goods, Bonn, Germany

Lacey J. Hilliard
Institute for Applied Research in Youth Development, Tufts University, Medford, MA, USA

Barbara Hoenig
Innsbruck University, Innsbruck, Austria

Victoria Holec
Canadian Centre for Behavioural Neuroscience, University of Lethbridge, Lethbridge, AB, Canada

Chen Hon-Fai
Lingnan University, Hong Kong

Nick Hopkins
School of Psychology, University of Dundee, Dundee, UK

Alf Hornborg
Lund University, Lund, Sweden

Mei-Chih Hu
National Tsing Hua University, Hsinchu, Taiwan

Robert A. Hummer
University of Texas at Austin, Austin, TX, USA

Richard S. Humphreys
University of California at Santa Barbara, Santa Barbara, CA, USA

Andrew Tsz Wan Hung
The Hong Kong Polytechnic University – Hong Kong Community College, Kowloon, Hong Kong

Giuseppe Iaria
University of Calgary, Calgary, AB, Canada

Georg Gerson Iggers
University at Buffalo (State University of New York), Buffalo, NY, USA

Georg G. Iggers
State University of New York, Buffalo, NY, USA

Yasuo Ihara
The University of Tokyo, Bunkyoku, Tokyo, Japan

Jacquelyn A. Jampolsky
University of Colorado, Boulder, CO, USA

Guillermina Jasso
New York University, New York, NY, USA

Narayana Jayaram
Tata Institute of Social Sciences, Mumbai, India

David S. Johnson
U.S. Bureau of Economic Analysis, Washington, DC, USA

Sara K. Johnson
Institute for Applied Research in Youth Development, Tufts University, Medford, MA, USA

Lyle V. Jones
University of North Carolina at Chapel Hill, NC, USA

Christian Joppke
University of Bern, Bern, Switzerland

Donald R. Kelley
Rutgers University, New Brunswick, NJ, USA

Jonathan Klaaren
University of the Witwatersrand, Johannesburg, South Africa

Karl Christoph Klauer
University of Freiburg, Freiburg, Germany

C.G. Knight
Pennsylvania State University, University Park, PA, USA

Jane Kroger
University of Tromsø, Tromsø, Norway; and Western Washington University, Bellingham, WA, USA

Yvonne Kuepper
Justus-Liebig-University Giessen, Giessen, Germany

David A. Lake
University of California, San Diego, La Jolla, CA, USA

Simon Langlois
Université Laval, Québec City, QC, Canada

Christa Lee-Chuvala
Massachusetts Institute of Technology, Cambridge, MA, USA

Richard M. Lerner
Institute for Applied Research in Youth Development, Tufts University, Medford, MA, USA

Elaine A. Leventhal
Rutgers University, New Brunswick, NJ, USA

Howard Leventhal
Rutgers University, New Brunswick, NJ, USA

Charles Lewis
Fordham University, Bronx, NY, USA

Reid Lifset
School of Forestry & Environmental Studies, New Haven, CT, USA

Yen Ying Lim
Warren Alpert School of Medicine of Brown University, Lifespan Hospital System, Providence, RI, USA

Irene Liu
University of Calgary, Calgary, AB, Canada

Chris Lorenz
VU University Amsterdam, Amsterdam, The Netherlands; and Ruhr-University Bochum, Bochum, Germany

T.M. Luhrmann
Stanford University, Stanford, CA, USA

Paul M. Lützeler
Washington University, St. Louis, MO, USA

Christoph Markschies
Theologische Fakultät der Humboldt-Universität zu Berlin, Berlin, Germany

Chad J. Marsolek
University of Minnesota, Minneapolis, MN, USA

Douglas S. Massey
Princeton University, Princeton, NJ, USA

Sarah Mathew
Arizona State University, Tempe, AZ, USA

J. May
University of Sheffield, Sheffield, UK

C. Behan McCullagh
La Trobe University, Bundoora, VIC, Australia

Jennifer E. Melvin
University of Texas at Austin, Austin, TX, USA

Max-Peter Menzel
Institute of Geography, Universität Hamburg, Hamburg, Germany

Allison Metz
FPG Child Development Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Terance D. Miethe
University of Nevada, Las Vegas, NV, USA

Supriya Mukherjee
University of Manitoba, Winnipeg, MB, Canada

Kara Murias
University of Calgary, Calgary, AB, Canada

Andrew Neal
The University of Queensland, St Lucia, QLD, Australia

Wilfried Nippel
Humboldt-Universität zu Berlin, Philosophische Fakultät I, Institut für Geschichtswissenschaften, Berlin, Germany

P.K. O'Brien
University of London, London, UK

Ingram Olkin
Stanford University, Stanford, CA, USA

Naadli T. Ormiston
Indigenous Studies Program, Camosun College, Victoria, BC, Canada; and School of Social Work, University of Victoria, Victoria, BC, Canada

Tikki Pang
Lee Kuan Yew School of Public Policy, National University of Singapore, Singapore, Singapore

Soumya Parameshwaran
Kasturba Medical College, Mangalore, India

Rogelia Pe-Pua
University of New South Wales, Sydney, NSW, Australia

Sofia A. Perez
Boston University, Boston, MA, USA

Burkhard Peter
MEG-Stiftung, München, Germany

Philip Pettit
Princeton University, Princeton, NJ, USA and Australian National University, Canberra, ACT, Australia

Kavita S. Philip
University of California, Irvine, CA, USA

Andrew I. Port
Wayne State University, Detroit, MI, USA

Dudley L. Poston
Texas A&M University, College Station, TX, USA

Joanna Radin
Yale University, New Haven, CT, USA

Josef P. Rauschecker
Georgetown University, Washington, DC, USA; and Institute for Advanced Studies, Technische Universität München, Garching, Germany

T. Raychaudhuri
Oxford, UK

Wendy C. Regoeczi
Cleveland State University, Cleveland, OH, USA

Elisabeth Reichert
Southern Illinois University at Carbondale, Carbondale, IL, USA

Anthony Reid
Australian National University, Canberra, ACT, Australia

JaNelle M. Ricks
Rollins School of Public Health, Emory University, Atlanta, GA, USA

Berthold Riese
Sun Yatsen University, Canton, People's Republic of China

Shelly Ronen
New York University, New York City, NY, USA

Neil E. Rowland
University of Florida, Gainesville, FL, USA

Jörn Rüsen
Institute for Advanced Study in the Humanities, Essen, Germany; and University of Witten/Herdecke, Bochum, Germany

Osamu Saito
Hitotsubashi University, Tokyo, Japan

Allan R. Sampson
University of Pittsburgh, Pittsburgh, PA, USA

Penelope Sanderson
The University of Queensland, St Lucia, QLD, Australia

Agnar Sandmo
Norwegian School of Economics, Bergen, Norway

Oriane Sarrasin
University of Sussex, Brighton, UK

Masayuki Sato
University of Yamanashi, Kōfu, Yamanashi, Japan

Dirk Sauerland
Witten/Herdecke University, Witten, Germany

Vivian Scheinsohn
Instituto Nacional de Antropología y Pensamiento Latinoamericano (INAPL)/Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Universidad de Buenos Aires (UBA), Buenos Aires, Argentina

Puja Seth
Rollins School of Public Health, Emory University, Atlanta, GA, USA

David Shapiro
University of California, Los Angeles (UCLA), Los Angeles, CA, USA

Paul Sillitoe
Durham University, Durham, UK

Edward Slone
University of Calgary, Calgary, AB, Canada

Timothy M. Smeeding
University of Wisconsin—Madison, Madison, WI, USA

Simon C. Smith
University of Hull, Hull, UK

T.A.B. Snijders
University of Groningen, Groningen, The Netherlands

Peter J. Snyder
Warren Alpert School of Medicine of Brown University, Lifespan Hospital System, Providence, RI, USA; and Rhode Island School of Design, Providence, RI, USA

Aviam Soifer
William S. Richardson School of Law, University of Hawai'i, Honolulu, HI, USA

W. Spohn
Free University Berlin, Berlin, Germany

Richard Stansfield
Rutgers University, Camden, NJ, USA

Charles Stewart
University College London, London, UK

Thomas Straube
Institute of Medical Psychology and Systems Neuroscience, University of Münster, Münster, Germany

Thomas Straube
University of Münster, Münster, Germany

Sarah Teige-Mocigemba
University of Freiburg, Freiburg, Germany

Christian Thiel
Institut fuer Philosophie, Universität Erlangen-Nürnberg, Erlangen, Germany

Mark G. Thomas
Research Department of Genetics, Evolution and Environment, University College London, London, UK

Simon Thompson
University of the West of England, Bristol, UK

Stella Ting-Toomey
California State University Fullerton, Fullerton, CA, USA

Eddie M.W. Tong
National University of Singapore, Singapore, Singapore

Rolf Torstendahl
Uppsala University, Uppsala, Sweden

Ruth Tringham
University of California, Berkeley, CA, USA

John W. Tukey[†]
Princeton University, Princeton, NJ, USA

Sara V. Tupak
Institute of Medical Psychology and Systems Neuroscience, University of Münster, Münster, Germany

Stephen P. Turner
University of South Florida, Tampa, FL, USA

Melissa Van Dyke
FPG Child Development Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Qingjia E. Wang
Rowan University, Glassboro, NY, USA

Ian Q. Whishaw
Canadian Centre for Behavioural Neuroscience, University of Lethbridge, Lethbridge, AB, Canada

Tanya N. Whittle
University of Delaware, Newark, DE, USA

Rebecca A. Williamson
Georgia State University, Atlanta, GA, USA

Nicholas Hoover Wilson
Yale University, CT, USA

Gordon M. Winder
Ludwig-Maximilians University Munich, Munich, Germany

Arthur P. Wolf
Stanford University, Stanford, CA, USA

John P. Wright
Central Michigan University, Mt. Pleasant, MI, USA

Tania Xiao
The University of Queensland, St Lucia, QLD, Australia

Sylvia J. Yanagisako
Stanford University, Stanford, CA, USA

Jessica Yu
Rutgers University, New Brunswick, NJ, USA

Ingo Zettler
University of Copenhagen, Copenhagen, Denmark; and University of Tübingen, Tübingen, Germany

Hub Zwart
Radboud University of Nijmegen, Nijmegen, The Netherlands

[†] Deceased.

Historical Archaeology

Sarah Croucher, Wesleyan University, Middletown, CT, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M. Hall, volume 10, pp. 6718–6720, © 2001, Elsevier Ltd.

Abstract

Historical archaeology refers to the study of archaeological remains dating from 'modern' periods, usually within the last 500 years. This article outlines the major themes of historical archaeology, with particular attention to capitalism and colonialism. It discusses new theoretical developments, in particular the expansion of the field beyond North America and the Atlantic World, and greater engagement with social theory.

The term historical archaeology is somewhat of a misnomer for this field. Technically, historical archaeology could apply to any period for which historical documentation is available to supplement the material record studied by archaeologists. We might think of historical archaeology as more specifically the archaeology of the modern world (although modernity itself is a contested term). This period coincides roughly with the beginning of capitalism and European colonialism, with most historical archaeologists studying material remains, and the social relations to which they relate, from the late 1400s onward. Historical archaeology thus encounters some of the material with which we ourselves have some familiarity. The field itself was first broadly constituted in North America in the 1960s, and by the 1990s was firmly established with a focus on the major themes discussed here (Little, 1994).

Several key social, cultural, and economic issues are most commonly investigated through historical archaeology's broad framing of modernity (cf Funari et al., 1999). Capitalism has been, and continues to be, one of the central topics for historical archaeology. Our field studies the emergence of capitalism in Europe, the Americas, and elsewhere. This is often thought of in linear terms, with capitalist ways of life first being thought to emerge in England. Studies of household space and landscapes show that a process of enclosure began to commodify agrarian space, households increasingly began to differentiate work and domestic space, and food preparation and consumption began to separate (Johnson, 1996). Such changes have been argued to form a genealogy of later capitalist shifts in the United States, often framed as an eighteenth-century 'Georgian Order' in housing and other material culture (Deetz, 1996; Leone, 1984).

Historical archaeological studies of capitalism go far beyond economic shifts; fundamentally these studies are interested in the social changes associated with new ways of organizing labor and bodies, often concomitant with the rise of the individual (Shackel, 1993). Beyond British genealogies and the eighteenth-century Georgian Order, capitalism has been addressed in a range of different contexts. Plantations, for instance, have been central to wider debates as to the development of capitalist production (Mintz, 1985; Wolf, 1982). Studying the landscape of plantations, archaeologists have identified ways in which space was used to help create new means of viewing labor, in line with the development of particular technologies of labor that are seen to be tied up in capitalism (Delle, 1998). Other archaeologists have addressed capitalism through the world systems theory outlined by Immanuel Wallerstein (2000). Thus, the archaeology of Russian America can be studied as a part of a global system of capitalism, linking changes in material culture and oppression of native communities to wider worldwide economics (Crowell, 1997). Industrial archaeology – the study of factories and domestic spaces in industrialized locations – has also been a vital part of the study of capitalism. This involves the study of factories themselves, including their development as workplaces separated from the home with clear class hierarchies, and the concomitant growth of new domestic spaces for working class laborers and their families (Casella and Croucher, 2010; McGuire and Reckner, 2005; Mrozowski, 2000; Palmer, 2005; Symonds, 2005). This turn to industrialization is studied hand in hand with the general shift toward urbanization, and extensive studies of working class 'slums' have taken place in cities such as New York (Mayne and Murray, 2001).

The second central research topic of historical archaeology is the study of colonialism. This is often a corollary of developing capitalism, but is regularly separated out in analyses (cf Croucher and Weiss, 2011; Orser, 1996). The archaeology of Spanish colonialism forms one broad subfield, in locations such as Mexico (Rodríguez-Alegría, 2005), the Caribbean and Florida (Deagan, 1998), and California (Voss, 2008). The archaeology of British colonialism forms another significant subfield (Lawrence, 2003) and this includes locations as diverse as North America (Deetz, 1996), Australia (Paterson and Wilson, 2000), and South Africa (Klose and Malan, 2000). Beyond this, increasing attention is being paid to the archaeology of French colonialism, particularly in the Caribbean (Dawdy, 2008; Kelly and Hardy, 2011). Smaller pockets of study have addressed Dutch colonialism in South America (Hall, 2000; Schrire, 1995), Russian colonialism in the Americas (Lightfoot et al., 1998), and Omani colonialism in East Africa. Recent directions in this field, drawing on postcolonial theory, have turned from the study of colonizer to that of the colonized, placing increasing importance on the way in which indigenous cultures resisted colonial rule and continued to thrive, even as forms of material culture may seem to change. Such studies emphasize the ever-changing nature of indigenous societies, and the need to understand these as occupying the same temporal space – i.e., *not* that of the 'primitive' – as colonizers (Liebmann and Murphy, 2011; Silliman, 2009).

The study of these issues is addressed through varied forms of material culture. Historical archaeologists excavate, just as in any other field, and many studies are based on quantitative analysis of artifact patterns at different sites (e.g., Shackel, 1993). Other studies address landscape-based approaches, studying buildings or larger-scale landscapes to assess changing patterns through time (e.g., Delle, 1998). Alongside these sources, historical archaeologists always incorporate, to varying degrees, historical data. This field is by no means a 'handmaiden to history' (Noël Hume, 1964); instead, we are able to bring new insights to historical data through our integration of material perspectives. Oral historical interviews may also be undertaken in conjunction with archaeological excavation, and these may yield an interesting interplay between recollection and material facts (Whitehead and Casella, 2011).

The beginnings of historical archaeology were rooted in a search for a distinctly *American* past. Plantations in the Chesapeake have long, for instance, been a mainstay of historical archaeology. However, there has also always been diversity in the field, if nothing else, through the study of Spanish colonialism. But in the last few decades there has been significant growth in the idea of historical archaeology as a global field. At first this included the recognition that colonial South Africa had interesting comparisons with the U.S. (Falk, 1991), and the study of European colonialism in Australia, as mentioned above, has also figured into historical archaeology writ large. By the 1990s 'postmedieval' archaeologists in the United Kingdom (those archaeologists studying roughly the same period as US-based historical archaeologists) were beginning to engage in serious debate with U.S. scholars about the Atlantic World. But this period also witnessed the push for something that was more recognizably postcolonial in theoretical orientation, that is, some way in which historical archaeology outside of the Atlantic World, and areas which did not experience white settler colonialism, might also be addressed through the broader rubric of historical archaeology. One of the areas in which this development has been most notable has been sub-Saharan Africa (Croucher, 2014; Reid and Lane, 2004; Richard, 2013). This move has come with a call to incorporate postcolonial theory into historical archaeology, so that the dominant narrative is no longer that of a teleological expansion of European colonialism and capitalism throughout the world, and in which local voices come to be privileged in the determination of heritage (Shepherd, 2002; Stahl, 2009).

One of the central subfields of historical archaeology is the study of the African Diaspora. This began with the excavation of plantations in the U.S., and many studies have been written about the materiality of enslaved life in the Americas, demonstrating the rich social and cultural lives of these groups. For instance, on Jamaica, Mark Hauser (2008) has used locally produced ceramics as a means to trace participation in autonomous economic transactions by African Jamaicans. The deliberate deposition of artifacts, and artifacts associated with burials often shows the way in which enslaved and free black communities participated in ritual practices linked to African traditions (Armstrong and Fleishman, 2003; Davidson, 2010). Archaeological studies in the Americas, and now in East Africa, have also shown how self-emancipated communities built autonomous lives outside of bondage (Marshall, 2013). Studying free communities allows archaeologists to understand the way in which agency was expressed by African descendant individuals and groups, and shifts our present-day focus on the African Diaspora away from slavery. As such there has been a significant turn toward the study of free black communities in the U.S. in particular, with growing attention focused on the effects of systematic racism through the nineteenth and twentieth centuries (Leone et al., 2005; Mullins, 1999). Themes of social justice, racialization, and the study of those groups who may lack widespread historical recognition are also present in studies of wider immigrant groups in the U.S. This includes the study of Irish Americans in the industrial northeast and overseas Chinese in the West (Voss, 2005).

Feminist and gendered approaches to archaeology have been taken up widely within historical archaeology. Since much of the structure of gendered life today has been actively generated in the last few hundred years, historical archaeologists are able to trace the way in which material culture has been used to create and reinforce gender norms and identities. For instance, Diana Wall (1994) has shown that ceramics for various forms of dining in nineteenth-century New York City were related to the growing development of middle- and upper-class values in relation to the separation of the domestic sphere. Masculinities have also been studied in historical archaeology; Rodney Harrison (2002) has shown how impressive flaked glass tools, known as Kimberley spear points, were central to new ways of defining masculinity for Aboriginal men in late nineteenth- and early twentieth-century Australia. Sexuality has also begun to be integrated into historical archaeology, with a range of case studies showing the value of analyzing former brothel contexts to understand more about the lives of the women and men who formerly inhabited these sites (Seifert, 1991). Sexuality has been recognized as one of the ways through which colonial subjects were controlled. In eighteenth-century California, controlling indigenous and colonist women's sexuality was achieved through various spatial means, including the redesign of the Presidio in San Francisco (Voss, 2008). Critiques of mainstream feminist approaches have also been raised through drawing historical archaeology into conversation with black feminist theory (Battle-Baptiste, 2011; Franklin, 2001). These demonstrate the general lack of diversity within the field and have forced archaeologists to think about intersectionality, particularly as it pertains to African Diaspora archaeology.

The lack of diversity in historical archaeology is being addressed in part through a revisioning of some projects in relation to local communities. Following on from public pressure in cases with egregious mismanagement such as the African Burial Ground in New York City, historical archaeologists are beginning to recognize the importance of being led by communities as we engage in our research (La Roche et al., 1997). The nature of our field means that there are often communities who have some form of direct ties to the material that we work with. Listening to such groups and having an open and collaborative research process may mean that we come to produce work that holds much greater value to a public beyond our discipline, and that in the process also pays greater attention to the wealth of knowledge held in communities outside of traditional scholarly venues (Agbe-Davies, 2010; McDavid, 2002).

Historical archaeologists also have their eyes turned to the future, present, and very recent past. Forensic archaeology overlaps with historical archaeology in chronological terms, and historical archaeologists engage with the practice and theory of locating and identifying remains associated with war, mass killings, and other human rights violations of the last century (Crossland, 2000; González-Ruibal, 2008; *see* Conflict and War, Archaeology of: Behavior and Social Organization). Studies of contemporary material culture, practiced using the methodologies of historical archaeology, are also allowing archaeologists to analyze the materiality of contemporary phenomena. For example, Jason De León (2012) has studied the material traces of, and objects carried by, undocumented migrants crossing the U.S. border from Mexico. This shows the way in which migrants struggle to carry useful objects purchased with limited means. Such archaeological projects fuse the methodology of historical archaeology with ethnographic study to bring forward new insights that emerge when material culture is taken seriously. Interdisciplinarity – drawing on varied historical and ethnographic data – is the hallmark of historical archaeology. This field is by no means a handmaiden to history; historical archaeologists are focused on material culture, and thus are able to bring new insights into periods which may seem much more familiar than the traditional areas of archaeological study.

See also: Anthropology and History; Apprenticeship, Anthropological Aspects; Archaeology, Politics of; Archaeology, Theory in; Capitalism; Colonization and Colonialism, History of; Ethnic Identity and Ethnicity in Archaeology; Gender, Feminism, and Sexuality in Archaeological Studies; Marxist Archaeology; Materiality and Culture; Migrations, Colonizations, and Diasporas in Archaeology; Modernization and Modernity in History.

Bibliography

Agbe-Davies, Anna S., 2010. Concepts of community in the pursuit of an inclusive archaeology. International Journal of Heritage Studies 16 (6), 373–389.

Armstrong, D., Fleischman, M., 2003. House-yard burials of enslaved laborers in eighteenth-century Jamaica. International Journal of Historical Archaeology 7 (1), 33–65.

Battle-Baptiste, Whitney, 2011. Black Feminist Archaeology. Left Coast Press, Walnut Creek, CA.

Casella, Eleanor Conlin, Croucher, Sarah K., 2010. The Alderley Sandhills Project: An Archaeology of Community Life in (Post)-Industrial England. Manchester University Press, Manchester, UK.

Crossland, Zoë, 2000. Buried lives. Archaeological Dialogues 7 (02), 146–159.

Croucher, Sarah K., Weiss, Lindsay, 2011. The archaeology of capitalism in colonial contexts, an introduction: provincializing historical archaeology. In: Croucher, S.K., Weiss, L.M. (Eds.), The Archaeology of Capitalism in Colonial Contexts: Postcolonial Historical Archaeologies. Springer, New York, pp. 1–38.

Croucher, S.K., 2014. Capitalism and Cloves: An Archaeology of Plantation Life on Nineteenth-Century Zanzibar. Springer/Society for Historical Archaeology, New York.

Crowell, Aron, 1997. Archaeology and the Capitalist World System: A Study from Russian America. Plenum Press, New York.

Davidson, J.M., 2010. Keeping the Devil at Bay: The Shoe on the Coffin Lid and Other Grave Charms in Nineteenth- and Early Twentieth-century America. International Journal of Historical Archaeology 14 (4), 614–649.

Dawdy, Shannon Lee, 2008. Building the Devil's Empire: French Colonial New Orleans. The University of Chicago Press, Chicago.

De León, Jason, 2012. "Better to be hot than caught": excavating the conflicting roles of migrant material culture. American Anthropologist 114 (3), 477–495.

Deagan, Kathleen, 1998. Transculturation and spanish american ethnogenesis: the archaeological legacy of the quincentenary. In: Cusick, J.G. (Ed.), Studies in Culture Contact: Interaction, Culture Change and Archaeology. Visiting Scholar Conference Volumes. Southern Illinois University, Center for Archaeological Investigations, Carbondale, pp. 23–43. Occasional Paper No. 25.

Deetz, James, 1996. In Small Things Forgotten: An Archaeology of Early American Life, second ed. Anchor Books, New York.

Delle, James A., 1998. An Archaeology of Social Space: Analyzing Coffee Plantations in Jamaica's Blue Mountains. Plenum Press, New York.

Falk, Lisa (Ed.), 1991. Historical Archaeology in Global Perspective. Smithsonian Institution Press, Washington, DC.

Franklin, Maria, 2001. A black feminist-inspired archaeology? Journal of Social Archaeology 1 (1), 108–125.

Funari, Pedro Paulo A., Jones, Siân, Hall, Martin, 1999. Introduction: archaeology in history. In: Funari, P.P.A., Hall, M., Jones, S. (Eds.), Historical Archaeology: Back from the Edge. Routledge, London, pp. 1–20.

González-Ruibal, Alfredo, 2008. Time to destroy: an archaeology of supermodernity. Current Anthropology 49 (2), 247–279.

Hall, Martin, 2000. Archaeology and the Modern World: Colonial Transcripts in South Africa and the Chesapeake. Routledge, New York.

Harrison, Rodney, 2002. Archaeology and the colonial encounter: kimberley spear points, cultural identity and masculinity in the North of Australia. Journal of Social Archaeology 2 (3), 352–377.

Hauser, Mark W., 2008. An Archaeology of Black Markets: Local Ceramics and Economies in Eighteenth-Century Jamaica. University Press of Florida, Gainesville.

Johnson, Matthew, 1996. An Archaeology of Capitalism. Blackwell, Cambridge, MA.

Kelly, K.G., Hardy, M.D., 2011. French Colonial Archaeology in the Southeast and Caribbean. University Press of Florida, Gainesville.

Klose, Jane, Malan, Antonia, 2000. The ceramic signature of the cape in the nineteenth century, with particular reference to the Tennant street site, cape town. The South African Archaeological Bulletin 55 (171), 49–59.

La Roche, Cheryl J., Blakey, Michael L., 1997. Seizing intellectual power: the dialogue at the New York african burial ground. Historical Archaeology 31 (3), 84–106.

Lawrence, S. (Ed.), 2003. Archaeologies of the British: Explorations of Identity in Great Britain and Its Colonies. Routledge, London, pp. 1600–1945.

Leone, Mark P., 1984. Interpreting ideology in historical archaeology: using the rules of perspective in the William Paca Garden in Annapolis, Maryland. In: Miller, D., Tilley, C. (Eds.), Ideology, Power and Prehistory. Cambridge University Press, New York, pp. 25–35.

Leone, Mark P., Janifer LaRoche, Cheryl, Babiarz, Jennifer J., 2005. The archaeology of black americans in recent times. Annual Review of Anthropology 34, 575–598.

Liebmann, Matthew, Melissa S. Murphy, 2011. Rethinking the archaeology of "Rebels, Backsliders, and Idolaters". In: Liebmann, M., Murphy, M.S. (Eds.), Enduring Conquests: Rethinking the Archaeology of Resistance to Spanish Colonialism in the Americas. School for Advanced Research, Advanced Seminar Series, J.F. Brooks, general editor. School for Advanced Research Press, Santa Fe, NM, pp. 3–18.

Lightfoot, Kent G., Martinez, Antionette, Schiff, Ann M., 1998. Daily practice and material culture in pluralistic social settings: an archaeological study of culture change and persistence from Fort Ross, California. American Antiquity 63 (2), 199–222.

Little, Barbara J., 1994. People with history: an update on historical archaeology in the United States. Journal of Archaeological Method and Theory 1 (1), 5–40.

Marshall, Lydia Wilson, 2013. Fugitive slaves and community creation in nineteenth-century Kenya: an archaeological and historical investigation of Watoro villages. Azania: Archaeological Research in Africa 48 (1), 148–149.

Mayne, Alan, Murray, Tim (Eds.), 2001. The Archaeology of Urban Landscapes: Explorations in Slumland. Cambridge University Press, New York.

McDavid, Carol, 2002. Archaeologies that hurt; descendants that matter: a pragmatic approach to collaboration in the public interpretation of African-American archaeology. World Archaeology 34 (2), 303–314.

McGuire, RandallH., Reckner, Paul, 2005. Building a working class archaeology. In: Casella, E., Symonds, J. (Eds.), Industrial Archaeology: Future Directions. Springer, New York, pp. 217–241. Contributions to Global Historical Archaeology.

Mintz, Sidney W., 1985. Sweetness and Power: The Place of Sugar in Modern History. Penguin Books, New York.

Mrozowski, Stephen, 2000. The growth of managerial capitalism and the subtleties of class analysis in historical archaeology. In: Delle, J.A., Mrozowski, S., Paynter, R. (Eds.), Lines that Divide: Historical Archaeologies of Race, Class, and Gender. University of Tennessee Press, Knoxville, pp. 276–305.

Mullins, Paul R., 1999. Race and Affluence: An Archaeology of African America and Consumer Culture. Contributions to Global Historical Archaeology. Kluwer Academic/Plenum Publishers, New York.

Noël Hume, I., 1964. Archaeology: handmaiden to history. The North Carolina Historical Review 41 (2), 214–225.

Orser Jr., Charles E., 1996. A Historical Archaeology of the Modern World. Plenum, New York.

Palmer, Marilyn, 2005. Understanding the workplace: a research framework for industrial archaeology in Britain. Industrial Archaeology Review 27 (1), 9–17.

Paterson, Alistair, Wilson, Andrew, 2000. Australian historical archaeology: retrospects and prospects. Australian Archaeology 50, 81–89.

Reid, Andrew, Lane, Paul, 2004. African Historical Archaeologies. Contributions to Global Historical Archaeology. Kluwer Academic/Plenum Publishers, New York.

Richard, François G., 2013. Hesitant geographies of power: the materiality of colonial rule in the Siin (Senegal), 1850–1960. Journal of Social Archaeology 13 (1), 54–79.

Rodríguesz-Alegría, Enrique, 2005. Eating like an Indian: negotiating social relations in the spanish colonies. Current Anthropology 46 (4), 551–573.

Schrire, Carmel, 1995. Digging through Darkness: Chronicles of an Archaeologist. University Press of Virginia, Charlottesville.

Seifert, D.J., 1991. Within site of the White House: the archaeology of working women. Historical Archaeology 25 (4), 82–108.

Shackel, Paul A., 1993. Personal Discipline and Material Culture: An Archaeology of Annapolis, Maryland. University of Tennessee Press, Knoxville, TN.

Shepherd, Nick, 2002. The politics of archaeology in Africa. Annual Review of Anthropology 31, 189–209.

Silliman, Stephen W., 2009. Change and continuity, practice and memory: native american persistence in colonial New England. American Antiquity 74 (2), 211–230.

Stahl, Ann B., 2009. The archaeology of African history. International Journal of African Historical Studies 42 (2), 241–255.

Symonds, J., 2005. Experiencing industry. In: Casella, E.C., Symonds, J. (Eds.), Industrial Archaeology, Contributions to Global Historical Archaeology. Springer, New York, pp. 33–57.

Voss, Barbara L., 2005. The archaeology of overseas chinese communities. World Archaeology 37 (3), 424–439.

Voss, Barbara L., 2008. The Archaeology of Ethnogenesis: Race and Sexuality in Colonial San Francisco. University of California Press, Berkeley, LA.

Wall, Diana diZerega, 1994. The Archaeology of Gender: Separating the Spheres in Urban America. Plenum Press, New York.

Wallerstein, Immanuel, 2000. The Essential Wallerstein. The New Press, New York.

Whitehead, Sarah, Casella, Eleanor Conlin, 2011. Mrs. Perrin's "Tranklements": community life and class distinction in (Post)Industrial-Era Cheshire. In: Beaudry, M.C., Symonds, J. (Eds.), Interpreting the Early Modern World. Springer, New York, pp. 177–195. Contributions to Global Historical Archaeology.

Wolf, Eric R., 1982. Europe and the People without History. Univeristy of California Press, Berkeley, Los Angeles, California.

Historical Demography

Osamu Saito, Hitotsubashi University, Tokyo, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Historical demography was born as a separate discipline in historical science in the 1950s, when the concepts and techniques of modern demography were applied by French scholars to historical material. Its progress and achievements since then have changed our perceptions on, first, the Malthusian interpretations of premodern economy and society and, second, the onset of demographic transition that followed. The achievements of this discipline are assessed by conducting a survey of sources and methods, as well as of microbehavioral and macroaggregative research results on both topics. Then, attention turns to cross-Eurasian comparative studies of demographic and social structure.

Some of the central questions that historical demographers are concerned with may be traced back to the writings of the classical economist Thomas Robert Malthus. His influential thesis postulated that human beings are able to reproduce themselves 'in a geometrical ratio' while the capacity of the earth to produce subsistence increases 'only in an arithmetical ratio.' In the long run, therefore, population increase of the past must have been kept under control either by the 'positive check' of higher mortality or by the 'preventive check' of abstinence or late marriage. This theory exerted profound influence on our understanding of the nature of early-modern demography and the transition to its modern regime. However, it was not until the post-World War II period that historical demography was born as a separate discipline in historical science: in the 1950s, concepts and techniques of modern demography were applied by French scholars to historical material. Its progress and achievements since then have changed our perceptions on Malthusian and related questions.

A Social Science History

Historical demography is a social science history. It is distinguished from other branches of history by its explicit use of concepts and tools derived from a discipline in social sciences (in this case, demography). Its goal is to establish historical facts by applying such conceptual frameworks to historical sources, and its methodological novelty lies in quantification. The research results of historical demography as such have often questioned the conventional pictures of past populations, thereby raising doubts about the assumptions upon which many theories rest. The position of historical demography in the social and behavioral sciences thus resembles that of New Economic History, but it should be realized that there are a few notable differences in emphasis between these two social science history fields. While the application of theoretically constructed econometric models is given prime importance in New Economic History, emphasis in historical demography has always been on measurement. Attempts to formulate models to explain historical facts have never been put aside, but more effort has been directed to the development of finer measures and the identification of patterns, temporal, and cross-cultural, by using those quantitative measures.

Sources, Methods, and Impacts

Innovation in methodology often goes with the discovery of new data sources. In historical demography, too, its birth was closely tied with French scholars' application of the family reconstitution method, a form of nominal linkage, to registers of baptism, burial, and marriage kept in Christian parish churches. Parish register demography, as one may call this type of research, is an attempt to reconstruct a demographic profile of the past based on a set of family histories reconstituted according to the rules set out by the demographer Louis Henry, who turned to historical material to identify what he called 'natural fertility,' that is, fertility before the deliberate effort to reduce family size came to be practiced.

The kind of information that historical demography has supplied since then is the sort that it had always been thought would be available only from modern vital records or sample surveys. Thanks to laborious work by historical demographers, demographers are now able to describe past populations in much the same vocabulary as those of the present. However, there are a few drawbacks. In this kind of research, analysis ought to be restricted to persons whose burial dates are known, or who are known to have survived until their reproductive period was over. Since more people tended to leave the community with advancing age, it is very difficult to compute death probabilities across all the age groups, although a large enough sample, such as the one constructed by E. A. Wrigley and his associates at the Cambridge Group for the History of Population and Social Structure, does enable researchers to estimate age-specific patterns of mortality even from family reconstitution (Wrigley et al., 1997).

Furthermore, demographic research based on historical microdata has enabled researchers to work with 'cohort' measures. Most demographic concepts, such as life table and marital fertility, are designed to portray the life events of a group of people who were born or married in the same period (termed 'cohort'), but usually it is not possible for present-day demographers to single out people born in a particular period and to trace them to the moment they die. Instead of using

cohort measures, they have to rely on 'period' rates to account for demographic events in a specified period, in which the behavior of one cohort is bound to be mixed with that of other cohorts. In this sense, historical microdata have given demographers an opportunity to work with methodologically ideal populations.

For historians, on the other hand, historical demography has offered them a different set of merits. First, population becomes tractable as 'population change' and can now be decomposed into demographic components such as births, deaths, and migrations. This in turn enables historians to relate those demographic phenomena to changes in economic circumstances. For example, one of the results of the earliest family reconstitution studies in France showed us how changes in grain prices were correlated with those in births and deaths. Similarly, one can examine interrelationships between time series of demographic rates and real wages as exemplified in Wrigley and Schofield (1981).

Another impact of historical demography is being felt in social history. Since parish register demography reconstructs families and their life events that occurred in familial contexts, it has offered social historians invariable insights into the demographic domain of family life. Marriage, for example, is no longer a subject tied only with legal and ecclesiastical concerns. Who married whom, at what age, and with what age gap between spouses – all these questions can now be answered. In addition, information about prenuptial pregnancy, gender differentials in infant and child mortality, and other related topics can shed light on the intimacy of family life in the past.

Other related but somewhat different approaches have been adopted by another group of social historians. Statistical studies of the family and household, launched by Peter Laslett at the Cambridge Group in the 1960s, have been conducted in close association with research in parish register demography (Laslett and Wall, 1972). The research along this line strongly suggests that types of family system, whether or not a new household was formed on marriage, whether the presence of live-in servants in the household was frequent or not, the mean age at marriage, and the frequency of celibacy were all closely interrelated.

These links with social history have led to the burgeoning of family and household studies in history. Topics studied thus far include servanthood; gender inequality within the household; care for the elderly, widows, and orphans; and the succession of headship and property. There are also issues that have enabled historians to get further into the inner domain of the individual. Abortion, infanticide, child neglect, and other means of family limitation embedded in traditional reproductive culture are such issues.

However, the source material being used in historical demography and its related research fields is not confined to the parish registers. In fact, there is a serious drawback in the use of parish registers as a source. Since only people who registered baptism, marriage, and burial in a single community were recorded, it is impossible to reconstruct families who eventually left the community. It is also impossible to know how many people existed in that community at a particular point in time. All this implies that aggregative studies are extremely difficult to conduct, implying that the size of a settlement population, its birth and death rates, and in- and out-migration ratios can hardly be computed. Also difficult to conduct, for the reasons already mentioned, are mortality studies.

It is not surprising, therefore, that in many Western European countries other microdata are being used. Especially useful are listings of households and population, most of which were one-off 'censuses' taken for various purposes. Also, there are places where census-type records were regularly taken. In traditional Japan, for example, registers of village populations were compiled by officials every year, while in some parts of imperial China similar records were taken triennially. Such repetitive census-type data are useful not only for household analysis but also for demographic studies, since probabilities of entry or disappearance can be calculated at a regular interval. Since the early 1990s, these East Asian sources, together with similar data from three European populations in Belgium, Italy, and Sweden, have been intensively explored by an international team of scholars for the studies of mortality, fertility, and nuptiality in the past. This Eurasian collaborative work has so far produced volumes on mortality and fertility, reexamining the Malthusian paradigms with those individual-level data in different economic, institutional, and cultural contexts (Bengtsson et al., 2004; Tsuya et al., 2010).

Two Dimensions: Micro and Macro

Two broad areas of issues have been occupying a central place in the historiography of historical demography. One is the Malthusian models of positive and preventive checks and the other is the thesis of demographic transition. The former concerns premodern settings, while the latter is about the decline in fertility and mortality from high premodern rates to lower modern rates (for stages and patterns of the transition, see Chesmais, 1992). In both fields of research, however, it is useful to distinguish microbehavioral from macroaggregative dimensions.

Microdemography

As noted in this article, parish register demography was born as a fertility-oriented research field. Malthus's notion of the positive check posits the idea that overpopulation tended to be checked by a rise in mortality, which postulates that the reproductive behavior of couples was governed by the invariable 'passion between the sexes.' The classic theory of demographic transition assumes that pretransition populations placed 'strong pressure on their members to reproduce,' and hence that marital fertility was uncontrolled and its level close to the physiological maximum. Both Louis Henry, who coined the term 'natural fertility,' and Ansley Coale, who developed a technique to measure a deviation from the natural, uncontrolled fertility pattern in the course of his own project on the European fertility decline, assumed that if the couple has no target size of family and was not concerned by the existing number of children (termed 'parity'), marital fertility was not controlled and hence its level must be invariably high. In other words, the decision made by a couple to end childbearing (termed 'stopping') is assumed to mark the onset of fertility limitation. Much effort, therefore, has been made to measure

fertility levels and identify patterns by examining age, parity, birth interval, and other factors associated with individual couples' reproductive behavior (see Coale and Watkins, 1986).

However, according to a large body of empirical studies so far conducted, the original assumptions about reproductive behavior prior to the demographic transition are no longer tenable. First, pretransition of marital fertility was far more variable than previously imagined. Few populations exhibited levels comparable to that exemplified by the Hutterites, a small group of Anabaptist settlers in North America whose total marital fertility rate (TMFR, defined as the number of children a married woman would have within her reproductive period) stood as high as 11. In most cases, it was kept at moderate levels with a TMFR ranging from 6 to 9 and, as in East Asia, may well have been as low as 4–5.5 even if the 15–19-year-old age group is included in the reproductive period.

Second, it is confirmed that stopping behavior was never common among premodern populations. It should be noted, however, that some elite subpopulations, such as Catholic aristocrats in France and Italy and bourgeois families in Calvinist Geneva, exhibited a distinct stopping behavior, so that they can be regarded as forerunners of modern family limitation. Third, moreover, the couple's behavior to 'space' births at particular intervals, as contrasted with stopping, seems to have been more important in accounting for the moderate levels of marital fertility in traditional times. Unlike stopping, birth spacing could have been achieved either through deliberate means such as contraception or through unconscious actions such as breastfeeding, and many research results suggest that the role played by differences in feeding practices of the infant was unmistakable.

Fourth, it is now widely recognized that in many instances marital fertility was rising, rather than falling, in the period immediately prior to the onset of modern fertility decline. This was the case in some of the historic Eurasian populations as well as in twentieth-century developing countries (Dyson and Murphy, 1985; Knodel, 1988; Tsuya et al., 2010; Wrigley, 1998). One of the factors suggested as having affected this rising tendency is a reduction in the level of breastfeeding, which was often associated with early phases of urbanization. Another is an improvement in maternal health, which can be inferred by various measures such as the declining incidence of stillbirths. The latter is a promising area of microdemographic study, because even from family reconstitution, it is possible to investigate the interplay between fertility and mortality measures at very early phases of life.

The conclusion that pretransition fertility was not uncontrolled, whether consciously or unconsciously, demolishes the conventional divide between the modern and premodern regimes. Furthermore, findings from microdemographic data analysis have made it difficult to see the demographic transition in a dichotomous framework. To what extent the transition was a behavioral revolution still remains debatable.

Macrodemography

The concept of 'reproduction' can also be applied to a population as a whole. It means the way in which the population replaces itself through natural processes of birth and death. The overall level of such replacement may be indicated by the total number of children that an average woman will have during her childbearing period. This is called the total fertility rate (TFR). Since it does not take mortality into account, the population's natural process should be measured in terms of both fertility and mortality such as life expectancy at birth (e_0). Thus, one may describe historical populations by referring to the two, TFR and e_0. According to Livi-Bacci (2007), for example, Paleolithic populations and the Russians in 1879 happened to roughly exhibit the same level of TFR, a little fewer than 6.5 children per woman. But mortality was different. The life expectancy is supposed to have been below 20 years for the Paleolithic, whereas that for Russia was close to 30. As a result, the difference in the rate of natural increase was considerable. It was only marginally above zero in the Paleolithic case, whereas, more than 10 000 years later, the Russian population was growing at somewhat over 1% per annum.

This, however, should not be taken to imply that the driving force in population history was always mortality decline, with fertility remaining high throughout the long period before the modern age. Indeed, over time both fertility and mortality may have exhibited substantial volatility. Neolithic settlers are believed to have been characterized by fertility and mortality higher than those of Paleolithic hunters and gatherers, reflecting a combination of higher population density and sedentism among the farming populations (Bocquet-Appel, 2011). A much more recent case in point is England from 1550 to 1800. Wrigley and Schofield (1981) have documented that from the late sixteenth century, when England's rate of natural increase stood at below 1% per annum, the country's TFR and e_0 simultaneously declined toward the late seventeenth century. From that period to the second half of the eighteenth century, the two measures came back to a position of higher fertility and higher life expectancy again. This time, however, the speed of rise in fertility exceeded that in life expectancy, so the country's population grew faster than 1% per annum.

Since information from family reconstitution studies does not add up to macro, aggregate measures, complex and sophisticated estimation techniques, such as back projection, are needed to reconstruct the history of a past population at large. Judging from a few well-grounded back projection works such as Wrigley and Schofield (1981), first, there is now agreement that the role of nuptiality should be stressed in accounting for how overall fertility changed in pretransition Western Europe. The age at marriage was instrumental in adjusting the actual level of fertility in relation to changing economic circumstances. This is what the classic Malthusian model of the preventive check predicted, thus suggesting that such a homeostatic mechanism was at work in the long- to medium-run time frame of the Western European past.

Second, this does not imply that mortality crises were not important. Short-term fluctuations in mortality were frequent in various regions in Europe, most of which were subsistence crises occasioned by crop failure. The demographic effects of famine are well known. The incidence not only takes a heavy death toll but also reduces fertility (Dyson and Ó Gráda, 2002). The frequency of famines, therefore, can account for much of the fluctuation in human population. However, the famine mortality was particularly high when accompanied by an outbreak of infectious disease, and it is the changing levels of 'exposure' to viruses and other microorganisms that acted as

a major determinant to mortality over the long run. And such exposure was in many cases not directly related to humans' economic conditions but had more to do with cross-cultural encounters of populations, each of which tended to function as a distinct disease pool. The supposed link between mortality and the general standard of material life, therefore, was less direct than often imagined.

This suggests that the overall relationship between economy and population growth was rather weak. It is true that during the period of fertility decline, neither nuptiality nor mortality could offset the declining trend of marital fertility and, hence, the corresponding rise in population. This, however, should not be taken to imply that economic growth, which accompanied this demographic transition in most cases, changed the whole relationship between the demographic variables a 'demographic regime' (Wilson and Airey, 1999). Neither industrialization nor urbanization was the crucial determinant of the onset of fertility decline. Even the modern decline of mortality was not in close association with an increase in the general standard of living.

Finally, mention must be made to the role of migration. The basic demographic equation tells us that a population will change according to the addition of births and immigrants, and the subtraction of deaths and emigrants. In other words, without information about in- and out-migration, it is impossible to account for any macrochange. Migration, therefore, is as important a topic as nuptiality, fertility, and mortality in historical demography.

Unfortunately, however, as historical sources rarely recorded actual population movements, not much has been done on this important topic. The historiography reveals that the volumes of in- and out-migration, distance, and pull-and-push mechanisms are issues relatively better covered thus far. Much less known is the demographic impact of such migration flows. Were death probabilities in areas that migrants headed for different from those in their native communities? Was their marriage and reproductive behavior altered by their geographical mobility? Or was even their parents' demographic decision making affected by the prospect that their children would eventually leave for a place of, presumably, better opportunities? It is only in cases where no effect of these was substantial that we do not need to know anything other than the rate of net migration (immigrants minus emigrants), which tended to be rather small. More specifically, any interpretation of results from family reconstitution studies rests upon the assumption that 'leavers' would behave just as 'stayers' did. All the demographic measures derived from this method relate only to the population of stayers (i.e., those born locally who continued to live in their native community). The mean age at marriage, for example, could have differed between these two populations. Although it turns out that the difference was minimal in the case of English family reconstitution, the possible effects of migration on marriage as well as the other demographic variables can be determined only empirically (Wrigley et al., 1997, Ch. 5).

One of the few findings we have in this regard is about rural-to-urban migration. In a traditional setting, the urban population was distinguished by a greater exposure to diseases and, hence, higher mortality than in the countryside. Obviously, even in such a mortality-dominant urban demography, the marriage market may also have been affected by the size of in-migration as well as that of labor demand. Moreover, the rural sector should not be considered demographically homogeneous. There are European case studies suggesting that 'proto-industrial' villages exhibited an earlier age at marriage and higher fertility rates than farming villages. To what extent this was universal is debatable. What is certain, however, is that the destination and volume of each migration stream between the rural and urban sectors could affect the overall rate of population change in a preindustrial setting.

The picture becomes less clear for the period of modern fertility transition, which in many cases coincided with extremely large volumes of migration streams, both long- and short-distance movements. It is probably safe to suppose that rural-to-urban migration in this period also meant deterioration in the disease environment, while transoceanic migration to land-abundant settlements led to a lower age at marriage and higher fertility. What is less clear is its influence on the population sending those migrants in the course of change. In the literature on demographic transition, migration, especially townward migration, tends to be regarded as having facilitated the decline of fertility. On the other hand, it is often noted that rural populations did not readily respond to changing economic and social circumstances by lowering their fertility. One suggested reason why it was still beneficial for peasant families to keep their fertility relatively high is the expanding prospects for the younger generation: when it was felt that they could emigrate and form their own households elsewhere, stayers' immediate need for the limitation of fertility may have been reduced.

The Demographic Landscape in Cross-Cultural Perspective

At a more generalized level, historical demographers have developed a notion of a homeostatic demographic regime. Many societies in the past had an ability to maintain equilibrium between their resources and the demographic characteristics; it is the ability to restore the original state of equilibrium, even if an external shock forces the system away from that position. When the ways in which this homeostasis is achieved are similar for a group of populations, it is probably appropriate to call that particular mode of interrelationships between the demographic variables a 'demographic regime.' Given empirical evidence so far accumulated, therefore, we may see the traditional Western European populations as characterized by a particular mode of demographic regime. In this regime, both fertility and mortality levels were relatively low, and the key to its adjusting mechanism was nuptiality acting to ensure the system would restore equilibrium.

Turning to the nuptiality patterns, John Hajnal in his influential articles has shown that in traditional Europe, west of the line from St. Petersburg to Trieste, the mean age at marriage was relatively high and celibacy not negligible, and that all this was closely tied with the nuclear family system and its associated principle of neolocalism. In contrast, non-European countries exhibited comparatively low ages at marriage and very low proportions never married, suggesting that different demographic regimes were operating in other parts of the world (Hajnal, 1965, 1983).

Take China and Japan. According to in-depth regional studies in traditional periods, the level of life expectancy at birth (e_0) in northern China was probably in the early 1930s, while it stood at 39 in central Japan (for China, see Bengtsson et al., 2004, Ch. 10; for Japan, see Liu et al., 2001, Ch. 7). Compared with the Western European average, China's mortality was on the high side and Japan's was on the low side. On the other hand, as noted in this article, TMFRs in both countries were low by Western European standards.

The Chinese and Japanese results raise some interesting questions. First, it is clear that the much lower level of marital fertility did not reflect stopping behavior, but spacing, thus confirming much of what the work on Europe has suggested. What is not certain at this stage is whether the observed long birth interval was a mere reflection of low fecundity, or a product of deliberate actions. While the factors accounting for the former vary from breastfeeding to malnutrition, the latter raises further questions: to what extent abortion, infanticide, and child neglect were practiced; which of the three was most responsible; and whether it was motivated by despair or guided by the notion of 'planning,' however rudimentary it would have been.

Second, the East Asian nuptiality pattern was undoubtedly different from that of Western Europe. There was no neolocal principle working. Nor was the age at marriage functioning so as to adjust overall fertility to economic conditions. Within East Asia, while China and Japan seem to have shared the characteristics of universal marriage, the female mean age at marriage was unmistakably higher in Japan than in China.

Third, however, if attention is turned to the status of females and family systems, we realize that there were some fundamental differences between China and Japan. For example, while no particular anomalies are observed from Japanese sex-specific life tables, the Chinese studies point out that parents in northern China are likely to have used infanticide on girls (Bengtsson et al., 2004), and according to an analysis of the relationship between changes in remarriage and death rates, females' remarriage among Chinese was more restricted than among Japanese. Clearly the status of girls and women was very different in traditional Chinese society. More specifically, it is the family system that differed. Predominant in Japan was the stem family system, under which the family household was vertically structured, and customs such as adoption, marriage, and remarriage were all for the family line to perpetuate. In China, on the other hand, what operated were the rules of joint family. Under that system, the household was laterally structured and the composition tended to be perennially complex.

Finally, all this suggests that Hajnal's dichotomous framework for marriage and family formation systems should be expanded to accommodate a third type (i.e., the stem family system), which will go some way toward a better understanding of cross-cultural differences in marriage and fertility behavior and in social structure of the past as well as the present (Saito, 2005).

To conclude, historical demography as it now stands is both social science oriented and interdisciplinary. It has questioned the Malthusian interpretations of early-modern economy and society and the processes of the demographic transition. Recently, moreover, it has shed new light on cross-Eurasian comparisons in demographic and social structure.

See also: Economic History; Family and Kinship, History of; History and the Social Sciences; Quantification in History; Social History.

Bibliography

Bengtsson, T., Campbell, C., Lee, J.Z., et al., 2004. Life under Pressure: Mortality and Living Standards in Europe and Asia, 1700–1900. MIT Press, Cambridge, MA.

Bocquet-Appel, J.-P., 2011. When the world's population took off: the springboard of the Neolithic Demographic Transition. Science 333, 560–561.

Chesnais, J.-C., 1992. The Demographic Transition: Stages, Patterns, and Economic Implications. Oxford University Press, Oxford.

Coale, A.J., Watkins, S.C. (Eds.), 1986. The Decline of Fertility in Europe. Princeton University Press, Princeton.

Dyson, T., Murphy, M., 1985. The onset of fertility transition. Population and Development Review 11, 399–440.

Dyson, T., Ó Gráda, C. (Eds.), 2002. Famine Demography: Perspectives from the Past and Present. Oxford University Press, Oxford.

Hajnal, J., 1965. European marriage patterns in perspective. In: Glass, D.V., Eversley, D.E.C. (Eds.), Population in History. Edward Arnold, London, pp. 101–143.

Hajnal, J., 1983. Two kinds of pre-industrial household formation system. In: Wall, R., Robin, J., Laslett, P. (Eds.), Family Forms in Historic Europe. Cambridge University Press, Cambridge.

Knodel, J.E., 1988. Demographic Behavior in the Past: A Study of Fourteen Village Populations in the Eighteenth and Nineteenth Centuries. Cambridge University Press, Cambridge.

Laslett, P., Wall, R. (Eds.), 1972. Household and Family in Past Time. Cambridge University Press, Cambridge.

Liu, T.-J., Lee, J., Reher, D., Saito, O., Wang, F. (Eds.), 2001. Asian Population History. Oxford University Press, Oxford.

Livi-Bacci, M., 2007. A Concise History of World Population, fourth ed. Blackwell, Oxford.

Saito, O., 2005. The third pattern of marriage and remarriage: Japan in Eurasian comparative perspectives. In: Engelen, T., Wolf, A.P. (Eds.), Marriage and the Family in Eurasia: Perspectives on the Hajnal Hypothesis. Aksant, Amsterdam, pp. 165–212.

Tsuya, N.O., Feng, W., Alter, G., Lee, J.Z., et al., 2010. Prudence and Pressure: Reproduction and Human Agency in Europe and Asia, 1700–1900. MIT Press, Cambridge, MA.

Wilson, C., Airey, P., 1999. How can a homeostatic enhance demographic transition theory? Population Studies 53, 117–128.

Wrigley, E.A., 1998. Explaining the rise in marital fertility in England in the 'long' eighteenth century. Economic History Review 51, 435–464.

Wrigley, E.A., Schofield, R.S., 1981. The Population History of England, 1541–1871: A Reconstruction. Edward Arnold, London.

Wrigley, E.A., Davies, R.S., Oeppen, J.E., Schofield, R.S., 1997. English Population History from Family Reconstitution 1580–1837. Cambridge University Press, Cambridge.

Historical Explanation, Theories of: Philosophical Aspects

C Behan McCullagh, La Trobe University, Bundoora, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

There have been three distinct ways of approaching the explanation of historical events: metahistorical, prescriptive, and descriptive. Metahistorians explain events as being driven by some grand overarching pattern of historical change. The best known are Enlightenment theories of progress toward greater freedom and rationality, and classical theories of the cyclical growth and decay of civilizations. Prescriptive theories state how historical events should be explained: causal, hermeneutic, structural, and narrative models have been proposed. Descriptive theories, on the other hand, analyze forms of explanations commonly found in history books. Arguments to the best explanation are commonly used to infer what happened in the past. Causal and rational explanations commonly explain why significant events and actions occurred. Finally, postmodern theories of language, culture, and narrative have driven some writers to skepticism about descriptions and explanations of the past. I suggest that such skepticism is not warranted.

There are three fairly distinct kinds of theory about the nature of historical explanations. First, there are reflections upon accounts people have given of basic processes of historical change that determine the general direction of history. These accounts are what J.-F. Lyotard calls 'the grand narratives' or 'metanarratives' of history. In Western culture, these have generally been theories of historical progress and theories of cyclical change. Such theories have been proposed by theologians, political theorists, and philosophers, but seldom by professional historians, who are understandably wary of generalizations of such immense scope. In recent years, both the truth and the point of such explanatory theories have been queried, and as a result such theories of world history are now seldom produced. Some historians still offer quite general explanatory interpretations of national and international history, but these too are often held to be partial, in both senses. For convenience, historical explanations of this first kind will be referred to as 'general interpretations of historical change.' The closest thing to a theory about such explanations is the discussion by J.-F. Lyotard, so his writing will be used as a platform for some remarks about them.

Second, there are four carefully articulated theories of how historians should explain events. These are written by philosophers rather than historians. Some think historical explanations should be scientific, employing general laws. Some declare that history is not a scientific but a hermeneutic discipline looking for the meaning of events rather than their causes. A few favor comprehensive, many-layered structural history, noting how basic geographic and social structures influence historical change. Finally, there are those who, in a postmodern vein, see history as no more than a form of literature, with historical events acquiring meaning from their place in a discourse. The philosophers who promote these theories of historical explanation would say they are describing historical practice, but in fact they are much less interested in analyzing historical practice than in promoting a form of explanation that they think is very important in understanding history. For that reason these will be called 'prescriptive theories of historical explanation.'

The third kind of theory of historical explanation is a theory that claims to be descriptive and analytic, focusing on historical practice to discover the variety of historical explanations historians actually produce and their criteria of acceptability. The task of this theory is to provide clear analysis of typical historical explanations, pointing out the logical standards that it seems they aim to satisfy. The two most common kinds of explanation produced by historians are explanations of the origins of physical artifacts, such as documents, paintings, coins, and buildings, to discover what happened in the past, and explanations of past actions and events to discover why they occurred. To discover what happened in the past, historians often employ a form of argument known as an argument to the best explanation. On the other hand, explanations of why people acted as they did, and why major events occurred, frequently identify reasons and other causes of actions and events that made their occurrence probable, or more probable than an envisaged alternative.

The article will briefly describe each of these theories of historical explanation, and make one or two comments upon them. It will conclude by drawing attention to the skepticism of some recent philosophers of history about the truth of historical descriptions and the correctness of historical explanations.

General Interpretations of Historical Change

In the culture of Western Europe, there have been several grand theories of historical change. Christianity, more particularly the Bible, brought with it a theory of progress in history, of God establishing his kingdom on earth, first through his prophets, then through the Israelites, through Jesus, and finally through the Church, until Christ returns in triumph. St. Augustine presented an immensely influential version of this story in *The City of God*. With the progress in science achieved by the application of reason in the seventeenth and eighteenth centuries, a secular theory of historical progress was created by Kant, Condorcet, and others, who saw reason and the methods of science driving human progress toward ever-increasing, this-worldly happiness. Hegel and Marx described the process of human progress as dialectical, contradictions being overcome by new stages of development, either in the advancement of reason, freedom, and self-consciousness in the case of Hegel or in autonomy and prosperity, in the case of Marx. The optimism implicit in all these theories of progress was largely shattered by

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63087-8

the experiences of the twentieth century, of hideous world wars, of vast economic depression, of rampant infectious diseases, and of tyranny, greed, and madness defeating enlightenment in so many countries. As Lyotard puts it, Auschwitz invalidates the Hegelian assumption that the real is rational; Budapest 1956 invalidates the Marxist claim that what is communist is proletarian (Lyotard, 1987, p. 162).

A few writers have argued that progress has indeed continued throughout the twentieth century. With the defeat of fascism and the disintegration of Soviet communism, F. Fukuyama sees the emergence of liberal democratic capitalism as the final stage of historical development, *The End of History*. He says that it best satisfies the two driving forces in history, the desire for material prosperity and the desire for personal recognition. Tyrannies had to go, as tyrants showed no respect for the common people. In democracies everyone has a say, or at least a vote. It seems obvious, however, that if the vote is not enough to ensure the recognition and respect people desire, then history has not ended after all.

Lyotard has three main objections to the grand theories of historical progress mentioned above. The first has already been stated: all are contradicted by the horrors of the twentieth century. Second, Lyotard objects that these theories have constrained historical writing in an unacceptable way. Under the influence of such theories, historians have written particular histories as instances of the theory of progress they most favored. For example, they depicted political revolutions as steps toward political freedom, and scientific revolutions as evidence of the growth of reason. In this way, Lyotard finds metanarratives oppressive, and prefers that people should write their narrative stories as they like (Lyotard, 1984).

Finally, Lyotard says that "all great historical upheavals" such as the French Revolution "are formless and figureless in historical human nature" (Lyotard, 1987, p. 174). The enthusiasm such an upheaval generates is sublime, he says, and incapable of true representation, although we like to suppose that all who participated were motivated by a desire for greater freedom. All that historians can truly describe, according to Lyotard, are individual human actions and responses, 'little narratives' of particular events.

A few comments on Lyotard's ideas seem called for. First, he is right to deny inevitable progress toward a more rational, free, self-conscious, prosperous, and autonomous society. Theories of progress all underestimate the desire and ability of powerful people to exploit others for their own gain. More realistic, it seems, are some cyclical theories of historical change. Aristotle and Polybius produced such theories, which were revived during the Renaissance in the writing of Louis Le Roy and Jean Bodin in France, and of Machiavelli and later Vico in Italy. In Aristotle, they found a doctrine of the natural succession of political structures from monarchy to aristocracy to democracy and back to monarchy again, each stage established to resolve injustices that had arisen in the previous one. From Polybius they came to see cycles in the growth, flourishing, and decay of states, some attributing decay to states overextending themselves, others seeing it as a result of moral failure among its leaders. However, as Bodin said, the causes of a state's collapse can be varied, including military conquest by its enemies (1965, p. 59). There is no general theory of historical change that is true without exception. Cyclical theories were popularized in the twentieth century by Manuel Spengler's (1926–1928) *The Decline of the West* and Toynbee's (1951–1961) *A Study of History*.

Second, while metahistories may have been used to legitimate certain accounts of particular historical events, to the exclusion of others, some generalizations about the motivation for political revolutions and other events seem in order. The leaders of the French Revolution and the pamphlet writers who supported them did legitimate their attack on the monarchy by an appeal to reason, liberty, and human rights. Of course there were other, more self-interested motives at work as well, but that is not to deny the legitimation given to the revolutionary movement by these ideas. Rather than dismiss generalizations about the motives for revolutions and other large-scale events as all mistaken, historians should take care to qualify them so that they are fairly accurate. Generalizations are useful ways of summing up individual stories and describing social movements.

Finally, one of the chief objections to metahistorical explanations of historical change is something that Lyotard does not mention. They frequently assume that a whole series of events were all justified by the same ideals, when in fact this is false. For example, many political events and revolutions have been justified in the name of liberty and human rights, but the ideas of liberty and human rights were not always the same in every case. The rights and liberties sought in the English Revolution, the French Revolution, and the Russian Revolution, for example, differed significantly. Therefore, it is misleading to suggest that they were all legitimated by the same ideals.

This typically historicist objection to metahistory is important, but not entirely decisive. While different particular rights and liberties were sought at different times and places, all can still be called rights and liberties. Indeed, all drew upon traditions of thought about human reason and dignity found in classical and biblical writing, which were applied to the particular situations. It is wrong to suggest that these events all had precisely the same legitimation, but it is also wrong to deny that the values they appealed to were quite unrelated.

Prescriptive Theories of Historical Explanation

In the last section, we considered some philosophical responses to theories of historical change, which is to theories that identified certain more or less constant sources of historical developments. These theories provided substantive explanations of history. We now turn to theories of historical explanation of quite a different kind. These are theories about the logical structure of explanations of historical events. They are not theories about what produces historical change but theories about how historians should explain what has happened in the past. These theories are prescriptive, describing how their authors think historical events should be explained. There are in fact many forms of explanation commonly used by historians, but these theories focus upon just those that their authors judge, for different reasons, to be the most important.

Theories of Scientific Explanation

In the nineteenth century, both John Stuart Mill and the historian Henry Thomas Buckle argued that historians should try to explain historical events as scientists do, by discovering

the laws of society that governed their occurrence. With these laws, the occurrence of historical events could be deduced, given certain antecedent conditions, just as the occurrence of natural events can be deduced from the laws of nature.

When K.R. Popper came to propose a theory of historical explanation, he quoted Mill's theory with approval, but thought it could be made more precise. "I suggest," he wrote, "that to give a causal explanation of a certain *specific event* means deducing a statement describing this event from two kinds of premises: from some *universal laws*, and from some singular or specific statements which we may call the *specific initial conditions*" (Popper, 1961, p. 122). Popper went on to point out that the laws used by historians are familiar, everyday ones of no particular interest. He opposed those historicists who believe there are general laws of historical development, and he opposed those who claim that historical events are unique, and so have nothing general about them whatever. Those who present general laws of history, such as the Marxist theory that all history is of class struggle, are really offering interpretations, Popper said, describing those changes in history that interest them.

The philosopher who most clearly and forcefully presented a deductive model of historical explanation was C.G. Hempel, in his article 'The Function of General Laws in History.' Drawing upon the tradition of Mill, Popper, and others, he produced the following analysis: The explanation of an event of some specific kind *E*

consists of

1. *a set of statements asserting the occurrence of certain events* $C_1 \ldots C_n$ *at certain times and places,*
2. *a set of universal hypotheses, such that*
   a. *the statements of both groups are reasonably well confirmed by empirical evidence,*
   b. *from the two groups of statements the sentence asserting the occurrence of event E can be logically deduced.*
   (Hempel, 1965, p. 232)

The letters *C* and *E* are intended to suggest cause and effect. The aim of historical explanations, he wrote, is that of "showing that the event in question was not 'a matter of chance,' but was to be expected in view of certain antecedent or simultaneous conditions" (Hempel, 1965, p. 235). He admitted that the relevant laws are often difficult to formulate precisely, and that often the consequence of the initial conditions occurs only with a degree of probability (Hempel, 1965, p. 237). Consequently, he called the explanations offered in history incomplete 'explanation sketches' (Hempel, 1965, p. 238).

Hempel's formulation of the deductive theory of explanation was subject to much discussion and criticism. The main objections were as follows:

1. The model does not state conditions sufficient to produce a good explanation,
   a. because sometimes events that enable one to predict an outcome are not the events that caused it. For instance, from a sudden fall in a barometer one can predict a storm, but that event does not explain the occurrence of the storm, because the fall in the barometer was not necessary for the storm's occurrence.
   b. Furthermore, for an explanation to be satisfactory, it must describe the kinds of causes that are of interest to the reader. If a man with a gastric ulcer is suffering from stomach pain after eating parsnips, a doctor would attribute the pain to his gastric ulcer, whereas his wife would attribute it to his eating parsnips. For the doctor, eating parsnips is normal, but having an ulcer is not, whereas the wife is familiar with the man's ulcer and picks out his eating parsnips as the event that made the difference. This objection assumes that the function of causal explanations is to explain a contrast between a given event and another expected, normal event, rather than to show why an event was predictable.
2. The model does not state conditions necessary for a good explanation,
   a. because often all that is required to explain an event is to show that the event was more probable than a contrasting state, which might not be very probable at all. Consider the question 'Why was the monarchy replaced by a republic in France in 1792?' This question could be asking, why was there a change in constitution and not just a change of government, why was the republic set up in 1792 and not earlier, and why was the establishment of a republic in France more likely than any other kind of constitution, such as a limited monarchy of the sort that resulted from the English Revolution?
   b. Furthermore, many explanations in history are not intended to show that an event was very probable, or that it was more probable than a contrasting state. For instance, historians often want to know the reasons why people behaved as they did, reasons that may or may not have made their actions very probable.

There is some merit in Hempel's theory that historical explanations are meant to show events to have been very probable. Admittedly, some explanations have different aims. Certainly, causal explanations must refer to events that were necessary for the occurrence of the event being explained. Nevertheless, some causal explanations are not contrastive but are meant to explain what brought about a certain event by pointing out events and circumstances that made its occurrence very probable. However, this can very seldom be done by citing a single law relating cause and effect. Normally, historical events are the product of several tendencies at work in a situation and are explained as an outcome of those. Historians try to make the outcome appear as highly probable as possible, but often that is not very probable, just more probable than anything else in the circumstances.

Theories of Hermeneutic Explanation: Explaining the Meaning of Actions

While G. Vico and J.G. von Herder had written of the importance of understanding human cultures as complex, almost organic units, the theory of hermeneutic understanding, in opposition to scientific understanding, was first fully applied to history by W. Dilthey. The theory had its origin in the lectures on hermeneutics by F. Schleiermacher, whose biography Dilthey published early in his career (in 1871). Following Schleiermacher, Dilthey said that the way to understand human actions is by using empathy to imagine the experiences and thoughts that moved people to act as they did. We interpret

what people do and say, he said, by drawing upon our own experiences of life to imagine what was going on in their minds to make them produce the expression that is being interpreted. After he had developed this psychological theory of understanding, Dilthey became interested in the meaning human actions have in furthering the aims of the social or cultural structures to which they contribute. He wrote:

> *To start with, the parts of a life have a meaning according to their relation to that life, its values and purposes, and according to the place they occupy in it. Furthermore, historical events become significant through being links in a system of interactions in which they co-operate with other parts to bring about values and purposes in the whole.* (Dilthey, 1961, p. 148)

The structures he mentions as being "the simplest and most homogeneous" are "education, economic life, law, politics, religions, social life, art, philosophy, science" (Dilthey, 1961, p. 146). These different structures often express similar values and beliefs and so are interconnected. These common ideas can create a common tendency in a particular culture for an age. People are constantly improving the structure and effectiveness of these organizations, and their actions can be interpreted as having such historical significance.

The main objection to Dilthey's theory of historical understanding was clearly stated, and addressed, by Gadamer in *Truth and Method* (1986). Dilthey had said that we understand other people's actions on the basis of our general knowledge of life in the world. Gadamer pointed out that often the historian's world is remote from the world of the people being studied. Dilthey, he said, "took no account whatsoever of the historical nature of experience" (Gadamer, 1986, p. 212). Heidegger explained that understanding things is not something we do after we have apprehended them, but rather it is what we do in apprehending them, in reading a text, or in recognizing a machine. Our interpretation of things is a function of our own historical nature (Gadamer, 1986, pp. 230–2): "that we study history only insofar as we are ourselves 'historical' means that the historicalness of human There-being [*Dasein*] in its expectancy and its forgetting is the condition of our being able to represent the past" (Gadamer, 1986, p. 232). Our understanding of what things mean is first of all provided by the traditions in which we have been raised (Gadamer, 1986, p. 250).

Once the historicity of historical understanding is fully understood, the question of the objectivity of history becomes quite urgent. Is there any sense in which one's understanding of the past could be called 'objective'? Gadamer has a couple of things to say about this. First, he insists upon coherence as a condition of correct understanding: "The movement of understanding is constantly from the whole to the part and back to the whole. Our task is to extend in concentric circles the unity of the understood meaning. The harmony of all the details with the whole is the criterion of correct understanding" (Gadamer, 1986, p. 259). He remarks that when it is difficult to find a coherent interpretation, we suspect that we are dealing with a world whose patterns of meaning we do not really understand. This leads to his second important point. Gadamer employs the concept of a 'horizon,' which he describes as "the range of vision that includes everything that can be seen from a particular vantage point" (Gadamer, 1986, p. 269). Historians occupy a particular vantage point in the history of the tradition that formed them. To understand a past event, they must do so from the point of view of those involved. They can find that historical horizon because they have learned of it from within their own tradition. Then when historians come to report their understanding of the past, they do so in terms of their present worldview. Gadamer explains, "In the process of understanding there takes place a real fusing of horizons, which means that as the historical horizon is projected, it is simultaneously removed" (Gadamer, 1986, p. 273). It is but a short step from Gadamer's account of historical understanding to postmodern theories, which describe history as a cultural construct, usually in narrative form, having no particular relation to the real events of the past whatever.

At the same time as German philosophers were reacting to attempts to make historical explanations scientific by developing theories of meaning, British philosophers were pointing out that historical explanations are typically accounts of the reasons for which people acted as they did. R.G. Collingwood wrote as follows:

> *For science, the event is discovered by perceiving it, and the further search for its cause is conducted by assigning it to its class and determining the relation between that class and others. For history, the object to be discovered is not the mere event, but the thought expressed in it. To discover that thought is already to understand it. After the historian has ascertained the facts, there is no further process of inquiring into their causes. When he knows what happened, he already knows why it happened.* (Collingwood, 1961, p. 214)

Collingwood said that knowledge of past minds is arrived at by drawing inferences from observable facts, more particularly by forming and testing hypotheses about what people had in mind when acting as they did.

Collingwood's attack on scientific models of explanation and espousal of rational explanations was given fresh expression by W.H. Dray in his first book *Laws and Explanation in History* (1957). He criticized the 'covering-law models' of Popper and Hempel and developed an account of rational explanation that was less idealistic and more logical than Collingwood's. He said that in deciding reasons for actions, historians relied on general knowledge of the form "When in a situation of type $C_1 \ldots C_n$ the thing to do is x" (Dray, 1957, p. 132). Such statements, he said, were 'principles of action,' which are universally known and respected. They explain actions by showing that the action was the rational thing to have done in the circumstances.

Today, we would give more respect to Collingwood's account of explanation than to this early statement of Dray's, because people's reasons for acting are not always universally appropriate. Indeed, even from the agent's point of view, they are sometimes not very rational at all. All that matters when explaining the reasons for an action is that they were the reasons for which the agent acted.

Hempel responded to Dray by arguing that reasons only explain actions if we know that people holding such reasons were likely to act on them (Hempel, 1963). According to Dray, historical explanations should show that actions were reasonable responses to a situation. Hempel said they must

always show that the action was to be expected in the circumstances. I think reasons can cause actions by inclining people to perform them, but they do not always make the actions appear either very rational or predictable.

Theories of Structural Explanation in History

There are two distinct kinds of structural explanations in history, both closely associated with the work of the French Annales School in the twentieth century.

Braudel's Structural Theory

The first kind of structural explanation is most famously exemplified by F. Braudel's large history, *The Mediterranean and the Mediterranean World in the Age of Philip II* (1972b). In it, Braudel explains the geographic and social structures in which observable actions occurred, and how those structures limited or promoted certain possibilities. He wrote:

> *when I think of the individual, I am always inclined to see him imprisoned within a destiny in which he himself has little hand, fixed in a landscape in which the infinite perspectives of the long term stretch into the distance both behind him and before. In historical analysis as I see it, rightly or wrongly, the long run always wins in the end … it indubitably limits both the freedom of the individual and even the role of chance.* (Braudel, 1972b, p. 1244)

The constraints that Braudel had in mind were largely geographical and technological, but he included frameworks of belief as well (see Braudel, 1972a, p. 18).

Braudel showed how geography and technology limited agricultural production and trade, and so the growth of wealth in the sixteenth century. The size of populations, and their vulnerability to famine and disease, also played a part. In addition to these slowly changing structures of geography and technology, Braudel also described cycles of economic and social change that occurred more frequently. These constituted processes of change within economic, social, and institutional structures that historians could model and describe in a theoretical way. He provided a careful analysis of the cycles of mercantile capitalism between the fourteenth and eighteenth centuries, which formed the background for the history of Philip II in the sixteenth century. As Braudel showed, these structures limited the possibilities of human action, as well as making the effects of certain events very probable. However, Braudel gave no systematic account of the relation between these two kinds of structures and the actions of individuals, which constituted the third level of historical description he identified.

Structures of Meaning

The second kind of structural explanation favored by historians of the Annales School, such as Lucien Febvre, Marc Bloch, and more recently by Emmanuel Le Roy Ladurie and many others, is one that explains the meaning of an utterance, a gesture, or a symbol by referring to a structure of meanings, a language, of which it is a part. Dilthey had anticipated this theory of meaning in some of his later writings, although he is best known for relating meaning to intention. This French structuralist theory of explanation emerged from work by Saussure in linguistics and Lévi-Strauss in anthropology, which explained the meaning of one part of a language by referring to other parts to which it was systematically related. Historians found such patterns of meaning in social events. As Jonathan Culler has put it

> *The cultural meaning of any particular act or object is determined by a whole system of constitutive rules [such as rules of a marriage ceremony or of football]—rules which do not so much regulate behaviour as create the possibility of particular forms of behaviour make it possible to marry, to score a goal, to write a poem, to be impolite. It is in this sense that a culture is composed of a set of symbolic systems.* (Culler, 1973, p. 22)

People are seldom aware of the rules governing the symbols they use, any more than they are aware of the rules of language, but the rules can be discovered by historians and anthropologists studying their behavior. The rules can be implicit or explicit. Historians prefer to study the explicit rules governing behavior, usually found in the discourse by which it is described, because then the meaning given to the behavior is truly historical, and not something merely inferred by the historians.

Theories of Narrative Explanation

Once historical explanations were thought to display the meaning of historical events, rather than just describe their causes, the way was open to view historical narratives as forms of explanation. This is precisely what Hayden White did in his seminal book *Metahistory* (1973). In the theoretical introduction to that book, White describes 'explanation by emplotment,' 'explanation by formal argument,' and 'explanation by ideological implication.' An historical event that is merely listed in a chronicle, he said, has no particular meaning or significance. However, once it is located in a story, it is given both meaning and significance, by being related to what preceded and followed it. Following Northrop Frye, White identified four forms of emplotment: romance, tragedy, comedy, and satire, each with its characteristic structure. Thus, for example, "Romance is fundamentally a drama of self-identification symbolised by the hero's transcendence of the world of experience, his victory over it, and his final liberation from it" (White, 1973, p. 8). As well as telling stories of different kinds, he said, historians also describe the relations between events in the past in several different ways. Following Stephen C. Pepper, he identified four kinds of 'argument' about the nature of historical events. The formist depicts them as unique, the organicist sees particular events as parts of a whole, the mechanist presents events as instances of general laws at work in history, and the contextualist describes events as products of tendencies at work in their environment. Finally, narratives can represent certain ideological views of historical change and give events meaning by locating them within one of those. Following Karl Mannheim, White picks out four such ideologies: that of the conservatives, who see history as the evolution of existing structures; the liberals, who think that reason and goodwill will eventually see that certain ideals are attained; the radicals, who think utopian conditions should be achieved by revolutionary means; and

finally the Anarchists, who believe that an ideal state will be achieved by destroying existing structures and leaving people free to act on their own humanity. White found 'affinities' between these three groups of four. For example, he said romantic plots often employed a formist presentation of events, and embodied an anarchist ideology.

White correctly points out that explanations of these kinds can be viewed as interpretations. Interpretations are one of several more or less equally well-justified accounts of a subject, which give meaning and significance to its parts. White did not consider whether one interpretation could be better than another, but F.R. Ankersmit did in his book *Narrative Logic* (1983, Chap. 8), in which he declared that narrative interpretations should aim for maximum scope, not only in the number of events they describe but also in the number they illuminate. This seems to be on the right track.

Descriptive Theories of Historical Explanation

Hayden White denied that his analysis of historical explanation was exhaustive, but it was far less prescriptive than previous theories, indicating a large range of ways in which historians could account for events in the past. An avowedly descriptive, analytic approach to historical explanations is that provided by McCullagh in *The Truth of History* (1998) and *The Logic of History* (2004). McCullagh analyzes forms of explanation common in historical writing to discover both their structure and their conditions of acceptability.

Historical inquiry usually begins with the collection of documents considered relevant to a topic of interest about people's behavior in the past. One way of drawing inferences about the past from present evidence is by judging the best explanation of that evidence. To choose between the possible explanations, scholars refer to the several criteria for judging the best explanation (see McCullagh, 1984, Ch. 2). For example, the four gospels in the New Testament of the Bible describe appearances of Jesus to his disciples as one who was alive, even though he had died. That is a topic of interest to Christians, and some have studied the significance of the gospels' accounts of the resurrection appearances very carefully (e.g., Craig (2008), Ch. 8 and Licona (2010)). Historians consider several different explanations of the gospels' reports of Jesus' appearances to his disciples after his crucifixion, trying to decide what really happened. One possible explanation is that Jesus' body was stolen from the tomb and the disciples experienced hallucinations of his risen body; another is that the reports are really myths designed to support the faith of the disciples and others in Jesus' divinity; and the third is that God raised Jesus from the dead and he appeared to the disciples, which is what Christians believe. The Christian hypothesis neatly explains the experiences of the disciples reported in the gospels, and their subsequent confident declarations that Jesus had risen from the dead in the book of Acts, so it has impressive explanatory scope and power. However, it is implausible, because it is a law of nature that after two days dead bodies do not live again, and the Christian hypothesis is ad hoc, because without any supporting evidence it declares that God raised Jesus, to avoid the problem of its natural implausibility. Having said that, however, the alternative explanations of reports of Jesus' appearances are not very convincing either: mass hallucinations are very improbable, and if the resurrection reports were a myth, it is unlikely that the disciples would have proclaimed them so vigorously, at some cost to themselves (for instance, Stephen was killed for saying that Jesus lived, and was the promised Messiah). (For a more detailed consideration of the arguments, see McCullagh (2012)).

Arguments to the best explanation are designed to consider the credibility of hypotheses about unknown causes of present data. Once historians have amassed information about a topic in history, they go on to offer explanations of events they judge to have been significant. This requires them to identify circumstances and events that were contingently necessary and jointly sufficient to have made the occurrence of their chosen events probable. They might want to know all the important causes of the events, or just the more important, or the conditions that made the events that occurred more probable than others that might have been expected in the circumstances. The events they explain are often actions. To explain these, historians often refer to the beliefs, values, and intentions of the agents; their interests, instincts, and concerns; and the rules of their culture and the expectations of their society. (See McCullagh, 2004, Ch. 4.) Sometimes, historians explain events that are described, not as actions, but as changes in a social structure: for instance, the defeat of an army, the collapse of a bank, and an increase in tribal hostility. Historians employ generalizations about common causes of such events to identify the particular causes at work in particular cases. Often these generalizations refer to tendencies or dispositions for such events to occur when triggered by events of a certain kind. For instance, tribal conflict within a state is likely to increase after the removal of a central government that previously suppressed it. Sometimes an event is caused, not by another event so much as by prevailing conditions. For instance, if one army is much weaker than another, then in a conflict between them, the first army is likely to be defeated. All these generalizations have been learned from historical experience.

When philosophers of history write about explanations, they usually have causal explanations or narrative explanations in mind. Historians can provide other kinds of explanation as well. They can explain what happened in the course of a certain event, how a certain change from one state to another occurred, what the structure of a certain institution or society was, as well as why an event occurred. The adequacy of such explanations is judged by deciding how adequately they answer the questions they are addressing. The questions provide both the logical and conceptual parameters of the required explanation.

The Challenge of Postmodern Theory

Historians have traditionally assumed that the conclusions they draw about the past from present evidence are very often true, even though it is logically possible for the conclusions of the inductive arguments they use to be false. They have also assumed that their causal explanations of past events are generally correct.

Over the last few decades, however, these assumptions about the truth of historical descriptions and the correctness of causal explanations have been challenged. This skepticism can be attributed to three streams of philosophical discussion. The first, attributed in the first place to F. de Saussure's analysis of language, is often called 'the linguistic turn' or 'the semiotic turn' (Spiegel, 2005, Introduction). Saussure said that the meaning of words is explained, not by their relation to things in the world but by their relation to one another. Some concluded that the world we believe in is really nothing but a linguistic construction, a product of our particular language, having no particular relationship to the world at all. The second source of postmodern skepticism is from the work of anthropologists like C. Geertz, who noted how the world seems different to people of different cultures and concluded that the world we believe in is simply a construction of our particular culture. In fact our understanding of the causes of events is not always the same as that of people in other cultures. Even within our own culture, our understanding of the causes of natural and social events changes from time to time with changes in scientific and social theories. The third discussion to promote skepticism about the truth of history has been among philosophers of history themselves, starting with H. White's *Metahistory*, which drew attention to the way historical narratives are structured (see Section Theories of Narrative Explanation above), and implied that their structure is determined, not by past events, but by the historians' preferences. The most strenuous argument for skepticism about the truth of history, drawing upon these and other sources, is A. Munslow's *The Future of History*. He is convinced that we cannot prove history true, and says that we should judge histories by their emotional, moral, or aesthetic value.

A thorough reply to the skeptics would take us too far afield. Suffice it to say that our descriptions of the world have truth conditions, and we are justified in believing them to be true when we know the kinds of experiences we would expect to have if they were true. We may believe that descriptions of the world are true if the kind of experiences they imply have indeed been experienced. Languages and cultures differ, but sentences describing the world in each language and culture have truth conditions, and can be judged true or false depending upon whether the experiences they imply have been experienced. To overcome skepticism about our knowledge of the world and of the past, we must investigate the relations between language and experience more fully than the skeptics I have mentioned. In addition, with particular regard to history, we must study the arguments by which historians defend their descriptions and explanations, to see whether they warrant belief in their conclusions. (See McCullagh, 1984, 1998, and 2004, which attempt to do just that.)

See also: Explanation: Conceptions in the Social Sciences; Functional Explanation: Philosophical Aspects; History: Overview; History: Representation Within and Functions of; History: Theories and Methods; Rational Choice Explanation: Philosophical Aspects.

Bibliography

Ankersmit, F.R., 1983. Narrative Logic. A Semantic Analysis of the Historian's Language. Martinus Nijhoff, The Hague, The Netherlands.
Braudel, F., 1966. La Méditerranée et le Monde Méditerranéen l'Epoque de Philippe II, second ed. Librairie Armand Colin, Paris.
Braudel, F., 1972a. History and the social sciences. Annales: economies, sociétés, civilisations. In: Burke, P. (Ed.), Economy and Society in Early Modern Europe. Essays from Annales. Routledge and Kegan Paul, London.
Braudel, F., 1972b. The Mediterranean and the Mediterranean World in the Age of Philip II. Collins, London.
Collingwood, R.G., 1961. The Idea of History. Oxford University Press, Oxford, UK.
Craig, W.L., 2008. Reasonable Faith. Christian Truth and Apologetics, third ed. Crossway Books, Wheaton, IL.
Culler, J., 1973. The linguistic basis of structuralism. In: Robey, D. (Ed.), Structuralism: An Introduction. Clarendon Press, Oxford, UK.
Dilthey, W., 1961. Meaning in history. In: Rickman, H.P. (Ed.), W. Dilthey's Thoughts on History and Society. George Allen and Unwin, London.
Dosse, F., 1987. L'Histoire en miettes: Des Annales a la nouvelle histoire. Éditions La Découverte, Paris.
Dosse, F., 1994 (trans. Conroy, P.V., Jr.). New History in France. The Triumph of the Annales. University of Illinois Press, Urbana, IL.
Dray, W., 1957. Laws and Explanation in History. Oxford University Press, Oxford, UK.
Dray, W., 1974. The historical explanation of actions reconsidered. In: Gardiner, P. (Ed.), The Philosophy of History. Oxford University Press, Oxford, UK.
Gadamer, H.-G., 1960. Wahrheit und Methode. J C B Mohr, Tübingen, Germany.
Gadamer, H.-G., 1986. Truth and Method. Crossroad, New York.
Hempel, C.G., 1963. Reasons and covering laws in historical explanation. In: Hook, S. (Ed.), The Philosophy of History: A Symposium. New York University Press, New York, pp. 143–163.
Hempel, C.G., 1965. The function of general laws in history. In: Hempel, C.G. (Ed.), Aspects of Scientific Explanation and Other Essays in the Philosophy of Science. Free Press, New York.
Licona, M.R., 2010. The Resurrection of Jesus: A New Historiographical Approach. IVP Academic, Downers Grove, IL.
Lyotard, J.-F., 1979. La Condition postmoderne: rapport sur le savoir. Editions de Minuit, Paris.
Lyotard, J.-F., 1984 (trans. Bennington, G., Massumi, B.). The Postmodern Condition: A Report on Knowledge. Manchester University Press, Manchester, UK.
Lyotard, J.-F., 1987. The sign of history. In: Attridge, D., Bennington, G., Young, R. (Eds.), Post-structuralism and the Question of History. Cambridge University Press, Cambridge, UK.
Manuel, F.E., 1965. Shapes of Philosophical History. Stanford University Press, Stanford, CA.
McCullagh, C.B., 1984 (reprinted 2009). Justifying Historical Descriptions. Cambridge University Press, Cambridge, UK.
McCullagh, C.B., 1998. The Truth of History. Routledge, London.
McCullagh, C.B., 2004. The Logic of History. Routledge, London.
McCullagh, C.B., 2012. The Resurrection of Jesus: Explanation or Interpretation? Southeastern Theological Review 3.1. Southeastern Baptist Theological Seminary, Wake Forest, NC, pp. 41–53.
Munslow, A., 2010. The Future of History. Palgrave Macmillan, Basingstoke, UK.
Owensby, J., 1994. Dilthey and the Narrative of History. Cornell University Press, Ithaca, NY.
Popper, K.R., 1961. The Poverty of Historicism. Routledge and Kegan Paul, London.
Spengler, O., 1926–1928 (trans. Atkinson, C.F.). The Decline of the West, vols 2. Allen and Unwin, London.
Spiegel, G.M., 2005. Practicing History. Routledge, New York.
Toynbee, A.J., 1951–1961. A Study of History, vols 10. Oxford University Press, London.
White, H., 1973. Metahistory: The Historical Imagination in Nineteenth-Century Europe. The Johns Hopkins University Press, Baltimore, MD.
White, H., 1978. Interpretation in history. In: White, H. (Ed.), Tropics of Discourse. Essays in Cultural Criticism. The Johns Hopkins University Press, Baltimore, MD.

Historical Geography

Gordon M Winder, Ludwig-Maximilians University Munich, Munich, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by B. Graham, volume 10, pp. 6737–6743, © 2001, Elsevier Ltd.

Abstract

Historical geography stems from efforts to incorporate historical research at the regional scale on landscape change and past geographies into geography. The scope of this early work is outlined, but while such matters continue to be important in historical geography research, the field of study now embraces many other research foci. Studies of capitalism and its transformative effects; research on environmental change and human–environment relations; and work interpreting the meanings of historical landscapes and landscape change are three other main constitutive areas of study. However, the field is rapidly expanding around new themes, such as postcolonial reinterpretations of the history of empires, historical geographies of the colonized world, the connections between identity formation and the spaces and places of nationalism, ethnicity, and gender. These and other new themes are outlined, and the point is made that historical geography is now a broad interdisciplinary field of inquiry and not a narrowly defined subdiscipline of geography. This is a cause of some adjustment within geography, and the concerns of historical geographers for their endeavor are discussed.

Historical geographers aim to incorporate historical research into geography by examining past geographies, landscape change and the meanings of landscapes, capitalism and its transformative effects on economies and societies, urbanization and demographic change, as well as environmental change and human–environment relations. Consequently the field of historical geography is dynamic with many new themes, such as the history of mobility, travel writing, postcolonial reinterpretations of the history of empires, historical geographies of the colonized world, the connections between identity formation and the spaces and places of nationalism, ethnicity, and gender, emerging in recent years. These and other new themes are outlined below. Historical geography is now a broad interdisciplinary field of inquiry and not a narrowly defined subdiscipline of geography. This is a cause of some adjustment within geography.

Setting a Place for History within Geography

Despite earlier antecedents, modern historical geography really stems from the late nineteenth century, and particularly the work of French geographers, notably Paul Vidal de la Blache, on regional landscape formation, and of the German school of *Anthropogeographie*. Offshoots from these beginnings were established in Britain and the USA, and by the 1920s historical geography seemed poised to play the same preeminent role in human geography as geomorphology had in physical geography. During the mid-twentieth century, however, geography tended to emphasize spatial relationships within a regional geography approach and, in this context, historical geographers struggled to assert historical methodologies and research questions within the discipline. Historical geographers like Derwent Whittlesey in the USA and H.C. Darby in the UK focused on landscape change and process, and each cultivated the core practice of researching past geographies using forms of cross-sectional analysis. In parallel with the pattern of explanation in geomorphology current before 1950, Derwent Whittlesey organized his research around a series of reconstructions of past geographies. Detailed reconstructions of past geographies became associated with the work of many historical geographers, including H.C. Such analysis paralleled patterns of explanation in geomorphology at the time, but focused on cultural landscapes rather than physical ones. Their reconstructions of past geographies highlighted change over time but the explanatory power of this 'sequent occupance' or reconstruction approach remained at best weak. Nevertheless, detailed reconstructions of past geographies remain an important tool in historical geography, as is illustrated in many historical atlases published toward the end of the twentieth century. In Los Angeles, Carl Sauer researched the diffusion of cultural objects and practices, and the study of cultural landscapes. Borrowing from German geography as well as biology and anthropology, he supervised many PhD students and, although a pioneer of cultural geography, did much to pave the way for historical explanation in geography.

The generation of historical geographers emerging after World War II confronted new challenges. As the quantitative revolution swept through geography, they wrestled with whether history was needed in geography if geography is spatial science? As the Cold War shaped universities and societies, they were confronted with political issues: can historical explanation be informed by Marxism? Some new centers of historical geography research emerged, and notably at the University of Wisconsin, Madison, where the Canadian Andrew Hill Clark fostered research on regions, economies, and cities within North America. In the UK, Alan Baker at Cambridge University and Robin Butlin at Leeds were leading figures in a growing subdiscipline. Some of the postwar generation of historical geographers harnessed diffusion theory for research into disease and epidemiology. In Britain, E.A. Wrigley pioneered demographic history. Others drew their explanatory models from settlement, agriculture, economic, and urbanization studies, but generally they advanced explanation through careful reconstructions using archival data. Growing numbers of historical geographers met in both regional and national

associations and in an internationalizing but loosely organized conference, now known as the International Conference of Historical Geographers (ICHG). The establishment of the *Journal of Historical Geography* in 1975 signaled the consolidation of the subdiscipline.

Human geography in general began to embrace and discuss historical explanation in new ways in the 1980s and 1990s. Spurred on by debates in the social sciences and history over historical materialism and structuralism, and by feminist theory and phenomenology, many human geographers looked for alternatives to positivist social science. In 1990, David Harvey confidently argued that analysis of the production of space, the geographical imagination, and the historical geography of concepts of space and time were central to any critical enquiry into the contemporary world. He pushed for both a critical historical geography of capitalist transformations of the world and a critical inquiry into the ways in which geographers framed time, change, and development. Simultaneously, Brian Harley had challenged conventional histories of cartography, and mapping practices within historical geography, by drawing out the ways in which cartography had been structured to serve various state purposes. Cultural geographer Dennis Cosgrove interrogated the concept of landscape in its historical European uses and found similar issues with geographers' studies of landscape change. In this context, the Canadian historical geographer Cole Harris argued for geography as a project of regional-historical synthesis, and against the growth of spatial science at the expense of humanistic geographies. As these and other geographers called for critical inquiry into the subjectivities of geography and their histories, it seemed that historical geography would be central to a human geography that was moving beyond positivist framings of knowledge.

Historical geographers drew inspiration from a wide range of social theory, and humanistic and feminist approaches. They researched diverse archival sources and wrestled with how these could be interpreted and what it meant to work in the archives. They brought new concepts, frames of reference, subjectivities, questions and topics as well as new quantitative methods, cartographic technologies, and forms of textual and visual analysis to their desks, as they pursued alternative forms of historical explanation. However, the distinctive approaches that emerged cut across earlier framings of historical geography as a subdiscipline of geography organized around inquiries into regional and landscape change and the analysis of past geographies, even as such research reached a high point of sophistication. The result has been prolific growth of historical geography but around many divergent and entangled research themes.

Regional Historical Geography

Regional historical geographies remain vital within the subdiscipline. In North America, Donald Meinig and the many students of Andrew Clark were very active over several decades in research and commentary in and around this approach. Their projects emphasized the dilemmas of (re)settlement and human–environment relations, and both drew on and contributed to theory and debates in economic, agricultural, and urban history. Perhaps their efforts reached an apogee with the publication of Donald Meinig's four volume work *The Shaping of America* (1998–2004), Carville Earle's *Geographical Inquiry and American Historical Problems* (1992), Cole Harris' *The Resettlement of British Columbia* (1998), and the *Historical Atlas of Canada* (1987–93) published in three volumes. As the reference to the historical atlas of Canada should make clear, thematic mapping has an important role in this research. North American regional historical geographers have contributed two key organizing ideas: the geographical personality of particular regions, and the formation of culture areas. This regional or area studies approach was not confined to rural landscape change, but included important work on the appropriation of 'new' lands through colonialism, on urban systems, and on industry. Nor was it confined to North America or to 'national' projects: historical geographers have written on an array of regions and areas. Generally theirs are synthetic works that draw together research on diverse phenomena and at many scales. Such research is now of considerable interest to those historians who are rescaling their research either to global, world, empire, regional, or even microscales.

Capitalist Transformation

The historical geography of capitalism and its transformative effects through history are a second main area of research, one which now constitutes a more substantial body of work than the regional studies stream but intersects with it. How capital works and acts to construct and destroy specific landscapes, work rules, economies, and social identities is a central area of inquiry. But capital is not conceived as the only actor, and attention is directed to the everyday life worlds, responses, resistances, and life chances of those living in a transforming, modernizing world. As a result, historical geographers working in this stream research diverse topics. In the UK, Derek Gregory and John Langton published monographs on regional transformations during the English industrial revolution, while M. Overton remapped the agrarian transformation in England. In Montreal, Sheri Olson and Pat Thorndon led research into a wide range of urban experiences of the industrial city, and identified the adaptive strategies and experiences of individuals in demographic and household economy terms and by class, ethnicity, gender, age, and family status. Alan Pred contributed complex situated montages of the spaces and experiences of world's fairs and dock work. Robert Lewis analyses the remaking of North America's metropolitan industrial spaces while I have applied network theory from economic geography to reinterpret the geographies of agricultural machinery production. Richard Harris unmasked North American suburban development processes and experiences, in the process challenging the Chicago School models on factual grounds. Richard Dennis surveyed English industrial cities, and researches the new urbanism emerging in the nineteenth century as a result of new apartment buildings and new forms of mobility. The construction of new spaces of consumption and segregated ethnic residential neighborhoods, and the contestation of public space have been other important areas for historical geographic research, but there is a fantastic array of related studies. For example, Mona

Domosh has interrogated understandings of New York's department stores by considering them in the contexts of the emergence of consumption districts and from the perspective of a gendered streetscape. In *The Ride to Modernity* (2001), economic geographer Glen Norcliffe explored Canadians' peculiar encounter with the bicycle and its spatial ramifications and social meanings as one of the first mass-produced consumer luxuries. At Berkeley, Richard Walker has scrutinized capitalist framings and constructions of nature. Historical geographers working in this broad research area have developed synergies and alliances with economic geographers, social and urban historians, demographers, labor, economic and business historians, and cultural geographers. Their research has not only broken with long-held historical narratives summarizing spatial processes under industrial capitalism, but has also challenged representations of the significance of present-day reconfigurations of urban and economic space.

Environmental Transformations

Environmental transformations and human–environment relations comprise a third area of research within historical geography. There are ancient scholarly works on these subjects and modern scholars included Alexander von Humboldt, George Perkins Marsh, Vidal de la Blache, and Élisée Reclus, so this is a research area with an illustrious pedigree in geography. In addition, as Alan Baker declared, these roots of a historical geography of human–environment relations are entangled with those of historical ecology as practiced in anthropology and biogeography. So Pierre Gourou's ground-breaking research into laterization in tropical environments is also part of this legacy. Midway through the twentieth century, *Man's Role in Changing the Face of the Earth* (1956), an encyclopedic survey on the subject edited by William Thomas, drew together an interdisciplinary who is who of the field. This feat was echoed in 1990 with another edited volume *The Earth as Transformed by Human Action*. These themes remain important foci for research within historical geography which continues to feature work reconstructing past natural environments and researching human impact on natural environments. Here historical geographers like Donald Meinig, who worked out the historical environmental limits of the American Great Plains and the South Australian wheat belt, J.M. Powell, who evaluated water management in Victoria, Australia, Eric Pawson, writing on New Zealand's environmental history, Craig Colten, whose *An Unnatural Metropolis* (2005), studied hazard and vulnerability in New Orleans, and Michael Williams, Graeme Wynn, and Michael Roche, who each wrote on forest (mis) management, have made substantive contributions. While some physical geographers, like Stanley Trimble and Andrew Goudie, who penned an introduction to the subject that has served the purpose for a long time now, have contributed to the research area, and one of the most read special issues of *The Journal of Historical Geography* focused on the history of climate change discourse, there is considerable scope for further involvement.

These efforts of historical geographers are now entangled with those of environmental historians such as Donald Worster, Richard White, Alfred Crosby, William Cronon, Caroline Merchant, Richard Grove, Christian Pfister, Christof Mauch, Frank Uekotter, John McNeil, David Nye, and Theodore Steinberg. Environmental history sprouted in the USA but has now been set to work in many lands. Its practitioners have tended to embrace environmentalist values and to write critically of past human interactions with environments. Their research projects are pitched at diverse scales from global patterns of species diffusion to local studies of dams and wetlands. They write histories of global, urban and marine environments, animals, green politics and environmental movements, perceptions of the environment, and science networks related to environmental knowledge. There is considerable traffic in the ideas and practices between them and the historical geographers researching environmental transformations and human–environment relations, and there are multiple entangled strands of research in this thriving field of interdisciplinary research.

Landscape Change

Landscape interpretation and change is a further key theme for historical geographers, but the foundational work in this theme came from cultural and, to a more limited extent, physical geographers rather than historical geographers. The physical geographer W.G. Hoskins (who published *The Making of the English Landscape* in 1955), and the cultural geographers J.B. Jackson (founding editor of the American journal *Landscape*) and Carl Sauer, were major mid-twentieth century figures writing to understand the making of national and regional landscapes. Their interests have been taken up inside cultural geography in the subsequent works of Wilbur Zelinsky, Pierce Lewis, Fred Kniffen, Yi-Fu Tuan, and by historical geographers Michael Conzen, Derek Holdsworth, and many others in North America, and of Della Hooke, Hugh Prince, and others in the UK. This tradition emphasizes landscape design, environmental awareness, and their manifestation in, especially, vernacular landscapes. The subject matter ranges from urban and suburban streetscapes, through public and religious buildings and monuments, factories, office towers and banks, to backyards, public gardens, cemeteries, and national parks. Particular attention has been paid to non-Western landscape design as, for example, in Hong-key Yoon's publications on geomancy. Historical geographers share their interests in these topics with professionals in cultural geography, anthropology, archeology, architectural history, folk studies, and museum studies.

Cultural–Historical Geography

The emergence in the 1980s of a new cultural geography transformed writing in historical geography on the meanings of landscape. By asserting that landscapes are material objects that manifest many discourses the new cultural geography fostered a burgeoning, critical literature on the meanings of landscape. This literature, heralded by Denis Cosgrove and Stephen Daniels' edited collection *The Iconography of Landscape* (1988), emphasizes symbolic landscape representation in the past, the

power relations manifest in such representations, and their work in forming identity, memory, and heritage. Attention has turned to many forms of landscape representation that were not the main foci for earlier research including landscape painting, travel writing, literature, money, and exhibitions. Cultural geographers like Denis Cosgrove, Stephen Daniels, Brian Osborne, James Duncan, Jeanne Kay Guelke, Pyrs Gruffudd, Emily Gilbert, Mona Domosh, Veronica della Dora, and Suzanne Seymour have studied past representational systems. There are new historical geographies of the present which analyze the use of aids to memory and nostalgia as active forces in today's landscapes. Among others, Nuala Johnson and Brian Graham have paid particular attention to landscapes of memory and heritage landscapes. As is increasingly the case, there are connections between scholars working on this theme and those working on other themes: here Don Mitchell's *The Lie of the Land: Migrant Workers and the California Landscape* (1996) can serve as an illustrative example.

Indeed, there is now common usage of the term 'cultural-historical geography' to capture a style of critical geographical research that uses historical studies and cultural theory to research not only landscape but also imperialism, colonialism, the environmental consequences of colonialism, and the history of geographical and environmental thought. Whether this term can capture the thrust of the emerging research enterprise is less certain. The roles of geographers in past imperialism, cartography, land survey, and orientalism are properly the subjects of critical enquiry. Historical geographers are interrogating imperialism as part of postcolonial reinterpretations and critiques of the history of empires and are writing historical geographies of the colonized world. Thus, the spaces, networks, exhibitions, maps, and landscapes of European empires and their settler colonists are being researched, for example, in the works of Felix Driver, David Gilbert, Stephen Legg, Alan Lester, Cole Harris, Brenda Yeoh, Mike Heffernan, and Kay Anderson. Tracing their entry point to Edward Said's *Orientalism* and the works of Michel Foucault, critical geographers like Derek Gregory and Joanne Sharp are investigating the representational systems and framings of geopolitical discourses, including the impacts of the Cold War on North American social science, which Matthew Farish explored in *The Contours of America's Cold War* (2010), and the geographies of the War on Terror, which Derek Gregory maps in *The Colonial Present* (2004). In their work, American imperialism is in focus.

National Styles

Different contexts have inflected specific 'national' foci for historical geography research, and this means that it is problematic to read off the content of historical geography from trends in the UK and USA alone. In New Zealand and Australia, historical geography has long been identified with research into environmental management, land use change, colonial land alienation and resettlement, and now with interpretation of colonial cultural landscapes. By contrast, in Canada, urban, industrial, and cultural landscape foci are each equally as important as environmental history or colonialism. In Japan, pre-nineteenth century geographies have been a main focus of interest but there is new research emerging on more recent periods, especially Japan's engagement with modernity and imperialism. In both Singapore and Israel, historical geographers have been directly confronting the colonial and post-colonial policies that have shaped their countries. There is a long-established and growing community of historical geographers in China. In Germany and Eastern Europe, environmental and land use change has emerged as the mainspring of historical geography, but in each case, it is problematic to make more than these cursory comments, since the field is in flux and 'national' styles of historical geography are inherently problematic. There is considerable networking across national borders as witnessed, among other things, by the history of the ICHG. The ICHG began life in 1975 with a meeting of British and Canadian historical geographers in Kingston, Canada, expanded to become an Australian, British, Canadian, New Zealand, and US conference, and, in 1986, at the Baton Rouge meeting, changed its name to the ICHG to reflect its international character. The most recent meetings were held in Kyoto in 2009 and Prague in 2012, and the next conference is scheduled for London in 2015.

Fluid Historical Geographies

There are many other themes that could be profitably included in this review but a full account of historical geography's tangled web is really beyond the scope of this paper. Historical geographies of identity probe the connections between nationalism, trans-nationalism, ethnicity, identity, globalization, and place. Geography, as a scientific and educational practice, has a historiography that is being critically reviewed. There are collaborative projects in GIS and remote sensing to recreate and reinterpret past geographies. Capturing historical cadastral, topographical, and other map and aerial photo data in a GIS has enormous potential to enhance analysis of landscape and environmental change. While Japanese historical geographers use cadastral maps of ancient Japan to assess recent earthquake damage, Anne Knowles has been remapping aspects of the holocaust, and British collaborations locate William Wordsworth as he penned his poems in the Lake District. Brian Graham and Catherine Nash identified interest in the study of modernity, but even the title of their reader, *Modern Historical Geographies* (2000), asserts an array of approaches, concepts, and perspectives. Attention continues to be directed to many different eras. Historical geographers write from many perspectives, pursue an array of approaches, and connect with different themes during their individual careers.

Fluidity is also true of the institutional frameworks for historical geography. As governments transform their research funding opportunities, geographical research is increasingly organized in research groups and teams, and the institutional contexts for historical geography morph. As individual geography departments refocus their programs, there is often no room for a dedicated 'historical geographer' and few such positions are advertised but it is also expected that geographers of all kinds will be historically informed and inquiring. This means that many departments are homes to excellent historical geographic research but that this is seldom organized under the framework of a department of geography with a chair in

the historical geography of a particular region. Colleagues appointed in GIS, physical, cultural, economic, social, or urban geography turn their attentions to historical problems and research. Some historians are housed in geography departments and there is also movement the other way. In these changed circumstances, knowledge of regional geographies and histories, the historical geography of capitalism, environmental transformations, human–environment relations, past landscapes, archival resources, and historical research methods remain crucial to much research.

Historical Geography in Review

The ways in which historical geographers conceive of their field of study have changed considerably over the last 50 years. Contrasting three reviews of historical geography illustrates the point. In his 1962 classification, H.C. Darby framed historical geography as a subdiscipline of geography which comprised research into geographies of the past, changing landscapes, the role of the past in the present, and the role of geographical factors and explanations in historical research, or geographical history. This latter term has been used in different ways within debates over the scope and content of historical geography. It was used either as a synonym for historical geography or as a descriptor of the special case of the use of geographic factors to explain historical phenomenon. US historical geographer Carville Earle argued for a revised historical geography, one that focused on geographical explanation of historical problems. Darby's seemingly programmatic statement reflected the struggles within mid-twentieth century geography and has been criticized from many quarters in subsequent years. Nevertheless, his review was meant to open a general field of inquiry rather than to constrain research.

When defining historical geography for *The Dictionary of Human Geography* (2000), Daniel Clayton declared that "geographical work on the past has become eclectic and no new orthodoxy for historical geography is likely to emerge." Such work was self-consciously interdisciplinary or without disciplinary label. Clayton left the impression that 'historical geography' was a label attached to a modernist project of academic disciplines and one with diminishing purchase as critiques of the foundational narratives of modernity intensified. For Clayton, the enduring legacy of the subdiscipline was its flagship *Journal of Historical Geography*, which he saw as a valuable arena for research and debate.

By contrast, in his *Geography and History* (2003), Alan Baker acknowledged no *old* orthodoxy for historical geography and instead identified three general approaches within a broad interdisciplinary field of historical geography – locational geographies and histories, environmental histories and geographies, landscape studies – and a fourth, integrative approach that emphasized region and area. He argued that regional geographies and histories have a claim to be at the center of historical geographic research because they are integrative in an explicitly geographic way, and might draw from any of the other three approaches. Baker was at pains to highlight "the diversity, the coherence and the relevance of historical geography" and "to reconcile some of the oppositions perceived within it and also between historical geography and history." While he unabashedly sought to discern order in the development of historical geography, he welcomed experiment, dialog, and encounter and denied that there ever was a narrowly defined historical geography concerned only with the reconstruction of past geographies and their changes over time.

Historical geography is, as ever, fluid. There are discernible trajectories but just how these are evaluated depends a great deal upon your point of view. Many historical geographers express concern over the prospects for their subdiscipline and discipline, but their concern seldom arises from any sense of seeing their subject area 'invaded' by other interest groups. Instead, they often express the belief that critical geographers represent historical geography as stuck in positivist, cross-sectional, and regional practices while ignoring most of the substantive research in the field, and contend that, because other interesting and productive foci have emerged and present challenging and inspirational findings, there is now too little attention to region, urban system, ecosystem, or water shed in (historical) geographic research. They perceive a danger that the sweep of regional and systems change will be missed and historical explanation eschewed.

At the same time many historical geographers interpret the same developments by celebrating the continued growth and expansion of a dynamic field of interdisciplinary research with many new and interesting concepts, theories, and frameworks for research. Historical geographers are in dialog with historians of all kinds, economic, social, and cultural geographers, geographical information scientists, even physical geographers. Theorizing space, place, landscape, maps, populations, cities, and environment is increasingly important in history. This is seen in the many collaborative projects and teams of researchers that have emerged in environmental history, urban history, social science history, history of cartography, landscape and cultural studies, and historical demography. For example, some recent works in New Zealand history, such as Eric Pawson and Tom Brooking's *Environmental Histories of New Zealand* (2002) and *Seeds of Empire* (2011), or Barbara Brookes, Anne Cooper, and Robin Law's *Sites of Gender* (2003), feature teams of geographers, historians, and social scientists. Indeed, the authors publishing in the pages of *The Journal of Historical Geography* have increasingly diverse backgrounds. Within geography time, memory, and the past have proved to be of continuing interest as geographers seek to research uneven development in space just as the world seems more intimately wired and networked than ever before. In each and every thematic subdiscipline, geographers need to develop relevant historical geographies.

Historical geography is currently expanding as critical geographical work on the past proliferates. The subdiscipline's borders are open and bridges across the 'divide' with history are multiplying. Historical geography, more eclectic and more experimental than ever before, is increasingly a site for interdisciplinary research as social scientists and historians try to establish which aspects of the twenty-first century world are new and how they are significant. Simultaneously, institutional support for the subdiscipline has narrowed as geography departments favor thematic specializations, albeit with historical research interests. This situation could be read as a sign of the long-term success of

historical geography: it has helped to establish a permanent place for history within geography.

See also: Cultural Geography; Gender and Feminist Studies in Geography; Landscape; Nature and Society in Geography; Postcolonial Geography; Regional Geography; Time-Space in Geography; War and Terrorism, Geography of.

Bibliography

Baker, A.R.H., 1999. Reflections on the relations of historical geography and the annales school of history. In: Clark, S. (Ed.). The Annales School: Critical Assessments, vol. II. Routledge, London, pp. 96–129.

Baker, A.R.H., 2003. Geography and History: Bridging the Divide. Cambridge University Press, Cambridge.

Baker, A.R.H., 2007. Classifying geographical history. The Professional Geographer 59 (3), 344–356.

Butlin, R.A. (Ed.), 1993. Historical Geography: Through the Gates of Space and Time. Edward Arnold, London.

Butlin, R.A., 2009. Geographies of Empire: European Imperialism and Colonialism C. 1880–1960. Cambridge University Press, Cambridge.

Colten, C., 2005. An Unnatural Metropolis: Wresting New Orleans from Nature. LSU Press, Baton Rouge.

Cosgrove, D., Daniels, S. (Eds.), 1988. The Iconography of Landscape: Essays on the Symbolic Representation, Design and Use of Past Environments Cambridge.

Driver, F., 2001. Geography Militant: Cultures of Exploration and Empire. Blackwell, Oxford.

Goudie, A., 1981. The Human Impact on the Natural Environment: Past, Present and Future. Blackwell, Oxford (Seventh ed., 2013).

Graham, B., Nash, C., 2000. Modern Historical Geographies. Pearson Education, Harlow.

Harris, R.C., 1971. Theory and synthesis in historical geography. Canadian Geographer 15, 157–172.

Harris, R.C., 1997. The Resettlement of British Columbia: Essays on Colonialism and Geographical Change. UBC Press, Vancouver.

Harris, R.C., 2004. How did colonialism dispossess? Comments from an edge of empire. Annals of the Association of American Geographers 94 (1), 165–182.

Harvey, D., 1990. Between space and time: reflections on the geographical imagination. Annals of the Association of American Geographers 80, 418–434.

Lewis, R., 2008. Chicago Made: Factory Networksin the Industrial Metropolis. Chicago University Press, Chicago.

Meinig, D., 1988–2004. The Shaping of America: A Geographical Perspective on 500 Years of History. Yale University Press, New Haven.

Olsson, S., Thornton, P., 2011. Peopling the North American City: Montreal 1840–1900. McGill-Queen's University Press, Montreal and Kingston.

Philo, C., 1994. History, geography and the 'still greater mystery' of historical geography. In: Gregory, D., Martin, R., Smith, G. (Eds.), Human Geography: Society, Space and Social Science. MacMillan, Basingstoke, pp. 252–281.

Smith, N., Godlewska, A. (Eds.), 1994. Geography and Empire. Blackwell, Oxford.

Williams, M., 1994. The relations of environmental history and historical geography. Journal of Historical Geography 20, 3–21.

Historical Linguistics: Numerical Methods

Sheila M Embleton, York University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Numerical methods in historical linguistics strive to determine how closely two languages are related. Pairwise, this can be used to reconstruct a tree for a language family (lexicostatistics). More ambitious methods assign dates to splits within trees (glottochronology). The methods have been extended to distant relationships, often to try to answer whether the languages involved are related at all. All methods, no matter how sophisticated, rest on the simplest observation: since languages change over time, originally related languages gradually become less similar. This enables inferences about diachrony from synchronic information only, attractive when little is known about history. Methodological difficulties come in quantifying rates of change, measuring difference, and accounting for factors interfering with normal change processes (e.g., mutual influence between languages). The attractions of numerical methods are the same as in any humanities discipline: speed, objectivity, replicability, and ability to handle large volumes of data efficiently.

Introduction

The main goal of numerical methods in historical linguistics has always been to answer the question of how closely two languages are related. Applied pairwise to languages, this can be used to reconstruct a tree for a whole language family (lexicostatistics). More ambitious methods have additionally assigned dates to the splits within those trees (glottochronology). The questions and methods have been extended to involve distant (hence controversial) relationships between languages, so the basic question becomes not just how closely two languages are related, but whether they are related at all. All methods in this area, no matter how sophisticated, rest on the simplest of observations: since languages change over time, two originally related languages gradually become less similar over time. This enables the linguist to infer something about diachrony from synchronic information only, particularly attractive in situations where little is known about the history. The methodological difficulties come in quantifying rates of change, measuring difference, and accounting for the myriad of factors that can interfere with the so-called normal change processes (e.g., mutual influence between languages). The attractions of numerical methods in historical linguistics are the same as those attributed to quantitative methods in almost any discipline in the humanities or social sciences: speed, objectivity, replicability, and ability to handle large volumes of data efficiently.

Early Methods

The earliest precursor to a numerical method in historical linguistics is a statement made by Robert Gordon Latham (1850: p. 565, see also Hewes in Hymes, 1960: p. 338): "the average rate at which languages change is capable of being approximated" and "the maximum difference, at a given period, between two or more languages is also capable of being approximated". He never pursued this, and the next milestone in the development of numerical methods in historical linguistics was a statement by Sapir (1916: p. 76): "The greater the degree of linguistic differentiation within a stock the greater is the period of time that must be assumed for the development of such differentiation". Sapir's influence on later linguists was great, so even though he himself never followed this statement up by taking the basic steps of quantifying linguistic differentiation or rate of change, this can be seen as the genesis of the most widely known numerical method in historical linguistics, that of Swadesh (1950, and see below).

The first explicit numerical treatment of genetic proximity of languages was that of Kroeber and Chrétien (1937), following a method created by two anthropologists, Jan Czekanowski (1928) and Stanislaw Klimek (1935), for use in physical anthropology and ethnography. For each pair of languages, on a list of N features, calculate a (number of features shared by both languages), b (number of features in the first language but not in the second), c (number of features in the second language but not in the first), and d (number of features in neither language). This produces a 2×2 contingency table, a well-studied mathematical entity enabling various forms of analysis. In particular, the quantity Nr^2, where $r = (ad - bc)/\sqrt{(a+b)(c+d)(a+c)(b+d)}$, is χ^2 (one degree of freedom), enabling one to test the statistical significance of the relationship between the two languages. The values of r for all pairs of languages under consideration can become the input to various methods (e.g., hierarchical cluster analysis) for family tree reconstruction. The method is simple but not unproblematic (e.g., construction of the feature list, difficulties with nonindependence of features, interpretation of correlations especially if negative) (for complete treatment and commentary, see Embleton, 1986).

The second numerical method in historical linguistics is Ross's root-retention model (1950). For N features and m languages, construct an $N \times m$ table in which a cell has a cross if that language exhibits that feature. The question of relationship of languages can then be recast as whether or not the crosses are distributed randomly in the table. For each pair of languages, calculate n_1 (number of features found in the first language, the number of crosses in the first column), n_2 (number of features found in the second language, the number of crosses in the second column), and r (number of features shared by both

languages, the number of rows with a cross in both columns). Then the probability for that pair of languages of obtaining the shared number of crosses or more is

$$\sum_{R=r}^{N} = n_1 \frac{\left\{ {n_1 \atop R} \right\} \left\{ {N-n_1 \atop n_2 - R} \right\}}{\left\{ {N \atop n_2} \right\}}$$

for which a normal approximation exists. These probabilities can be tested for significance, and could become the input for various family tree reconstruction methods. Interestingly, Ross's method can be shown to be mathematically equivalent under certain conditions to that of Kroeber and Chrétien, and thus subject to the same problems (for complete treatment and commentary on this and the method in general, see Embleton, 1986). Although certain elements of these models recur today, both have largely been abandoned, although the problems encountered and lessons learned remain relevant.

Historically, the next method to evolve was Swadesh's, but given its centrality to the field and ongoing relevance, the author deals with that in a separate section. A final numerical method to consider is that of counterindications originated by Gleason (1959), later rediscovered by others. All topologically possible family trees for the languages in question are constructed, each is evaluated for how well it fits against a fixed list of features (every time the tree does not fit the distribution of that feature, a counterindication is recorded), and the tree with the best score (even if not perfect) is chosen. This method is difficult to implement for two reasons. Even with modern high-speed computing and sophisticated algorithms, the number of topologically possible trees for even a moderately sized language family quickly becomes unworkably large. The method is also highly vulnerable to the effect of borrowing across languages. A variant of Gleason's method, relying on hierarchical clustering and attempting to estimate branch lengths (i.e., used glottochronologically), is shown in Embleton (1986), along with a complete discussion of the method itself. Taylor et al. (2000), using recent advances and sophisticated techniques from computer science and biological cladistics, are developing a character-based (rather than distance-based) method for family tree reconstruction with some similarities to Gleason's method. Despite difficulties with dealing with borrowing in realistic amounts, such methods are promising partly because they closely parallel the cognitive processes of skilled practitioners of traditional comparative historical reconstruction techniques, while being more objective and simultaneously handling vast amounts of data.

Swadesh

By far the best known numerical method in historical linguistics is that introduced by Swadesh (e.g., 1950), actualizing Sapir's statement by regarding lexical replacement as analogous to radioactive decay. As with carbon-14 dating, dates could then be provided for branching points in the tree (glottochronology). Take a list of N meanings of supposed basic core vocabulary, found universally in all cultures, expressing basic noncultural concepts, expected to be more resistant to change and borrowing than other vocabulary items. Such a list, now known as a Swadesh list, contains such items as lower numerals, simple kinship terms, body parts, topographical terms, widespread flora and fauna, pronouns (personal, demonstrative, and interrogative), verbs denoting basic actions, and naturally occurring noncultural phenomena. Assuming a constant rate of replacement, R, for words representing meanings on this list, one can calculate the time elapsed (time depth) since any pair of languages split using the formula

$$\frac{-1}{2R} = In\left\{\frac{c}{N}\right\}$$

where c is the number of cognates on the Swadesh list between the two languages compared. By calculating time depths for all pairs of languages under consideration, we can construct a family tree with dates.

Swadesh's method was greeted with widespread enthusiasm and widely applied in the late 1950s, the 1960s, and the early 1970s, almost with a sense of euphoria about finally having an exact tool for obtaining insights into hitherto unknowable linguistic prehistory. The debate over Swadesh's method quickly became emotional, but regardless of the researcher's subjectivity, a common core of problems can be identified, sometimes with potential solutions.

Difficulties with particular meanings on the Swadesh list include the following:

1. Some meanings cannot be translated into the target language. This is fairly rare, and best treated as missing data, reducing N accordingly.
2. Sometimes there are multiple possible translations in the target language for a meaning on the list. Different researchers might make different but equally valid choices from the range of possibilities ('multiple synonymy'). Solutions include always counting only the most frequent possible translation, choosing at random, always choosing to maximize the cognate count, always choosing to minimize the cognate count, or various fractional scoring systems.

Difficulties with cognacy judgments (remember that one is not necessarily dealing with languages where the history and hence cognacy are well known) are as follows:

1. Due to accumulated sound change, cognates may not be recognizable, causing us to overestimate time depth.
2. Chance similarities may be wrongly interpreted as cognates, causing us to underestimate time depth. In practice, one might hope that this and the preceding cancel out.
3. Sometimes only one part of a word is cognate to a word in another language ('partial cognation'); how should one score this? Solutions include always counting this as positive, always counting this as negative, or various fractional scoring systems.

Problems with the universality of R, which was shown to vary over time, between languages, and according to meaning. R had been estimated from languages with written historical records as about 14% per millennium for the standard 100-word Swadesh list and about 19% per millennium for the

standard 200-word Swadesh list. Not only might the very fact that these languages are written bias the rate (one might expect less replacement in a language with a lengthy written history), but 11 of the 13 pairs used to establish the rate were Indo-European (representing only three of its branches).

Problems with the family tree (Stammbaum) model, generally centered on the splits not being clean (thus giving any precise date is artificial and meaningless). Languages are often in close physical proximity even after they split, mutually influencing one another; tendencies in the two languages predating the split may come to fruition only afterward. Both these interfere with the model, but are more properly problems with the family tree model, and Swadesh's adoption of it, rather than problems with Swadesh's method itself.

Borrowings between languages (related or unrelated) can alter replacement rates quite substantially. For some languages, even the Swadesh list contains a high amount of borrowed vocabulary.

For various reasons (see Embleton, 1986, and elsewhere), the use of Swadesh-style lexicostatistics and glottochronology waned by the mid-1970s, although applications lingered on in several language areas, and the methods continued to be refined by applied mathematicians.

Sankoff

Most refinements introduced by applied mathematicians involved adding further parameters to the basic Swadesh model. For example, parameters were added to account for drift and for 'chance recurrent cognation' (i.e., when a word was replaced twice, such that the original word was restored), replacement rates were allowed to vary from meaning to meaning, a very crude allowance was made for borrowing, and there was talk of allowing the rates to vary from language to language and from time to time (further details, along with some models, can be found in Embleton, 1986). Some of these are useful refinements (e.g., allowing replacement and borrowing rates to vary over time, between languages, and between meanings), but these are very much a mathematician's way of solving the problem, producing quite elegant solutions. But in the real world in which the linguist works, all those parameters have to be estimated – and this is surprisingly difficult even for well-studied languages for which we actually have both long and extensive written records. It is also the case that an extraordinary amount of data is required to estimate so many parameters, and this is most often simply impossible. So these elegant theoretical solutions have very little practical import.

However, the single most important development was the incorporation of a geographical dimension into the family tree, thus allowing for borrowing to be largely dealt with. The basic insight here is due to Sankoff (1972), with correction and further refinement by Embleton (see 1986 for the most comprehensive treatment). The resulting set of differential equations required to reconstruct the tree and estimate the branch lengths is complicated, reflecting the increased power of the method (equations are given in Embleton, 1986, and elsewhere). The parameters required all have some reasonable chance of being estimated, even for lesser studied families; these are a similarity measure for each language pair, replacement rates for each language, borrowing rates between each language pair (not necessarily symmetric), and knowledge of which languages are in contact with one another.

Starostin

Largely independently of the work described above, research on numerical methods in historical linguistics was proceeding in the Soviet Union (later Russia or 'in diaspora'). Unlike the impetus for Swadesh's work, which was largely to aid in providing histories and trees for relatively little studied language families (e.g., Amerindian), much of this work gained its impetus from questions of more distant linguistic relationship. This had its effect on the development of the methods, because it was crucial that the methods be able to produce useful results even at much greater time depths, where Swadesh's method typically has difficulty distinguishing the signal (cognacy, true genetic relationship) from the noise (chance resemblance, contact relationship).

Historically, the first advance here was due to Dolgopolsky (1986), who capitalized on the fact that different meanings change at different rates. He compiled a list of the 15 most stable meanings, based on examination of 140 languages. One can then calculate the probability of fortuitous coincidence of the words representing those 15 most stable meanings for any language pair; if the probability is significantly less than that expected by mere chance, this can be taken as evidence of relationship. This method is most striking in its extreme simplicity, but can also be related to some more complex and ongoing work, largely carried out in Eastern Europe, founded on the observation that several characteristics of a word (e.g., length, number of meanings, age) are intercorrelated with its frequency. Words of more recent origin tend to be less frequent, longer, and have fewer meanings, whereas older words tend to be more frequent, shorter, and have more meanings. The age–frequency relationship in particular is being extensively researched, seeking a small hard core of ancient short vocabulary that may reveal distant relationships, and may even be able to be harnessed into a lexicostatistical or even glottochronological method.

Another advance has been made by Sergei Starostin, with work first published in the 1980s and continuing until his death in 2005, known as root glottochronology or etymostatistics (see Starostin, 1999). Compared to standard Swadesh-style glottochronology, the main innovation is the observation that while glottochronology regards any change (however small) in a word's meaning as a loss of that word from that meaning, root glottochronology does not. Thus, while standard glottochronology sees English *dog* and German *Hund* in the meaning 'dog' as noncognate, root glottochronology notices the existence of the English word *hound*; this more conservative approach to the maintenance of a language's root stock enables one to work better at greater time depths. Starostin's method also gives more accurate results for known cases where the time depths are more shallow (standard glottochronology tends to give dates that are too recent). Root glottochronology takes a text, containing a given set of words in the parent language, and then measures the number of these words still extant (in any derivative, with any semantic shift) in the daughter language. It proceeds only with detailed comparative historical and etymological analysis

(eliminating look-alikes and loans), and can also be repeated for more than one text (hence a different set of words), giving greater precision and potentially some estimate of error (see Starostin, 1999, for formulas and more detailed explanation).

Ringe

Various linguists have used probabilistic methods to assess similarity between vocabularies of two different languages, usually focusing on phonological similarities (an early example is Ross, 1950). The basic question is whether the observed degree of similarity between the two vocabularies is greater than what could be expected based on chance. Although the problem is conceptually simple, it quickly becomes mathematically intractable even between just two languages, as one has to wrestle with different phoneme distributions, different phoneme frequencies, how to measure phonetic distance between different phone(me)s, different phonotactic structures, and lengths of roots, quite apart from the questions (familiar from glottochronology) of how much semantic latitude to allow between the words compared. The most recent and sophisticated treatment of this problem is by Ringe (e.g., 1998 and references therein; McMahon and McMahon, 2005: pp. 50–68), who attempts to account for many of the above issues, albeit with simplifying assumptions. This research remains active, yet without consensus as to its utility, partly because of the degree of the simplifying assumptions, and also partly because it has mostly been applied in the realm of long-distance relationships, where the emotions over the results can sometimes cloud the objectivity with which the method is viewed.

Current Work and Future Directions

Research in numerical methods in historical linguistics has reentered the mainstream after several decades on the sidelines. Several recent collections, published in mainstream venues, demonstrate the diversity and debates within the field; see in particular two special issues of journals (McMahon, 2005; Wichmann and Grant, 2010), as well as Holman et al. (2011), Eska and Ringe (2004), and Levinson and Gray (2012). The most active areas of current research are improvements to the basic Swadesh model; exploration of the age–frequency relationship; root glottochronology; computer-implemented tree reconstruction using methods borrowed from biological cladistics, due to the conceptual parallels between language change and biological evolution (computational cladistics, computational phylogenetics); and Ringe's probabilistic methods. The degree of mathematical, statistical, and computational expertise required is extremely high, which prevents the majority of historical linguists from ever mastering the methods, and consequently renders them unable to do more than use results produced by others or participate in research teams where others control the methods. There is a growing realization, however, that numerical methods while not providing any magic solutions to any long-standing problems, make a useful contribution as one of the tools that we use to explore our linguistic history and prehistory.

See also: Language Families, Archaeology and History of; Lexicology and Lexicography.

Bibliography

Czekanowski, Jan, 1928. Na marginesie recenzji P.K. Moszyń-skiego o Książce: Wstęp do Historji Slowian. Lud 7 (Series II). Reprint Lwow.

Dolgopolsky, Aaron, 1986. A probabilistic hypothesis concerning the oldest relationships among the language families in northern Eurasia. In: Shevoroshkin, Vitaliy, Markey, Thomas (Eds.), Typology, Relationship and Time: A Collection of Papers on Language Change and Relationship. Karoma, Ann Arbor, MI, pp. 28–50.

Embleton, Sheila M., 1986. Statistics in Historical Linguistics. In: Quantitative Linguistics, vol. 30. Brockmeyer, Bochum, Germany.

Eska, Joseph F., Ringe, Donald, 2004. Recent work in computational linguistic phylogeny. Language 80, 569–582.

Gleason Jr., Henry A., 1959. Counting and calculating for historical reconstruction. Anthropological Linguistics 1, 22–32.

Holman, Eric W., Brown, Cecil H., Wichmann, Søren, et al., 2011. Automated dating of the world's language families based on lexical similarity. Current Anthropology 52 (6), 841–875.

Hymes, Dell H., 1960. Lexicostatistics so far. Current Anthropology 1, 3–44, 338–345.

Klimek, Stanislaw, 1935. The structure of California Indian culture, culture element distributions. University of California Publications in American Archaeology and Ethnology 37 (1), 1–70.

Kroeber, Alfred L., Chrétien, Charles, D., 1937. Quantitative classification of Indo-European languages. Language 13, 83–103.

Latham, Robert G., 1850. The Natural History of the Varieties of Man. Van Voorst, London.

Levinson, Stephen C., Gray, Russell D., 2012. Tools from evolutionary biology shed new light on the diversification of languages. Trends in Cognitive Sciences 16, 167–173.

McMahon, April, 2005. Introduction. Transactions of the Philological Society 103 (2), 113–119.

McMahon, April, McMahon, Robert, 2005. Language Classification by Numbers. Oxford University Press, Oxford.

Ringe Jr., Donald R., 1998. Probabilistic evidence for Indo-Uralic. In: Salmons, Joseph C., Joseph, Brian D. (Eds.), Nostratic: Sifting the Evidence. Benjamins, Amsterdam, pp. 153–197.

Ross, Alan S.C., 1950. Philological probability problems. Journal of the Royal Statistical Society B 12, 19–59.

Sankoff, David, 1972. Reconstructing the history and geography of an evolutionary tree. American Mathematical Monthly 79, 596–603.

Sapir, Edward, 1916. Time Perspective in Aboriginal American Culture. In: Memoir 90, Anthropological Series, vol. 13. Government Printing Bureau, Ottawa, Canada, pp. 389–462.

Starostin, Sergei, 1999. Comparative-historical linguistics and lexicostatistics. In: Renfrew, Colin, McMahon, April, Trask, Larry (Eds.), Time Depth in Historical Linguistics, vol. 2. McDonald Institute for Archaeological Research, Cambridge, UK, pp. 377–413.

Swadesh, Morris, 1950. Salish internal relationships. International Journal of American Linguistics 16, 157–167.

Taylor, Ann, Warnow, Tandy, Ringe, Donald, 2000. Character-based reconstruction of a linguistic cladogram. In: Smith, John Charles, Bentley, Delia (Eds.), Historical Linguistics 1995. Selected Papers from the 12th International Conference on Historical Linguistics, Manchester, August 1995, General Issues and Non-Germanic Languages, vol. 1. Benjamins, Amsterdam, pp. 393–408.

Wichmann, Søren, Grant, Anthony, P. (Eds.), 2010. Quantitative approaches to Linguistic Diversity. Diachronica 27(2), 168 pp. (pp. 341–358).

Historical Sociology

Nicholas Hoover Wilson and Julia Adams, Yale University, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C. Tilly, volume 10, pp. 6753–6757, © 2001, Elsevier Ltd.

Abstract

'Historical sociology' is the systematic examination of the past to inform and reconstruct empirical findings and theories about the world. Its roots lie with sociology's classical founders, but it experienced vibrant growth and reconstruction after World War II. Today it is characterized by methodological pluralism, as well as a wide engagement with social-scientific questions of large-scale social change, the emergence and institutionalization of modern societies, and the epistemological foundations of social science history.

Historical sociology is the systematic examination of the past to inform and reconstruct empirical findings and theories about the world, and it is as old as sociology itself. Virtually all of the classical founders of the discipline – including Auguste Comte, Harriet Martineau, Karl Marx, Max Weber, Emile Durkheim, Georg Simmel, and W.E.B. Du Bois – invoked and analyzed history in their arguments about the nature and fate of modern social relations. Today, this engagement with history remains the vibrant core of historical sociology, even as it is complemented by more formal and abstract methodological reflections on the roles of time and sequence in processes of social change.

To emphasize the continuity of historical sociology between the age of the classics and now, however, is not to deny the dynamism of the field. Indeed, in one sense historical sociology has only recently returned to its roots. The field's classical representatives were keen to understand the genesis and vicissitudes of modernity in all aspects, and they situated their investigations in the broadest global and indeed imperial context (Steinmetz, 2013; but cf Magubane, 2004; Connell, 1997). Yet, as sociology as a whole – and historical sociology along with it – entered a new phase after World War II, many earlier and even contemporaneous insights were cast aside. This shift had many causes, but it was partly due to the analytical pull of Marxian analysis and responses to it, which narrowed privileged substantive objects of inquiry to structural political economy and the rise and dynamics of the modern nation-state.

Narrowing historical sociology to these overarching concerns and questions provided some real intellectual benefits. For example, research in historical sociology galvanized around a focused research program, which led to a series of paradigmatic works and generated insights into long-standing questions of social change. Barrington Moore, Jr's examination of the *Social Origins of Dictatorship and Democracy* (1966), a key early work in this vein, examined the interaction of class and commercializing agriculture in the shape modernizing European regimes took; Theda Skocpol's *States and Social Revolutions* (1979), another canonical text, compared several modern-world revolutions in the hopes of uncovering their common causes; Michael Mann's *Sources of Social Power* (1986) expanded a multicausal view of social change throughout human history; and the influential volumes edited by Charles Tilly (1975) and Evans et al. (1985) revolutionized the study of state formation and the administrative dynamics of governance. Immanuel Wallerstein's *The Modern World System 1* (2011[1974]), the most global of the major second-wave works, also insisted on the politico-economic foundation of social change and the post-1500 supersession of empires by states.

While each of these works in historical sociology's 'second wave' emphasized different aspects of social change, taken together they explored how modernizing states were 'potentially autonomous' from the societies they governed, and indeed how interstate competition and states' efforts to extract resources from their populations explained the forms regimes throughout the world took. Moreover, they also emphasized the other face of this process – how increasingly sophisticated political and state development intertwined with social and protest movements among the populations governed by emerging states, and how these populations sought to be recognized and to transform social relations through collective social action and protest.

The remarkable insights of the second wave of historical sociology were purchased at some cost. To facilitate comparison, and to more easily generalize their findings beyond the historical contexts they examined, second-wave authors privileged seemingly objective factors like political economy, international relations, and interest-based action. This shift away from contextual understanding of historical actors' outlooks and motives for their actions and toward structural constraints had substantive and methodological consequences. Substantively, it meant that the self-understanding of historical actors, as well as the categories usually applied to those understandings in sociology – including forms of identity – received relatively short shrift. For example, second-wave analyses of violence tended to concentrate on the size and organization of armies, the frequency and structure of urban disturbance, and the ways states financed military ventures. Meanwhile, however, key questions of how explicit violence and more subtle forms of coercion signaled the rise of modern self-understandings and forms of discipline were left aside, and it has only been recently that such foundational work on the subject as Norbert Elias' multivolume *The Civilizing Process* (2000) and Michel Foucault's *Discipline and Punish* (1977) have been reincorporated into the field of historical sociology.

Methodologically, moreover, authors in the second wave usually stressed the independence of their historical cases, in order to more clearly analyze the causes of the phenomenon at hand. Yet many world-historical events and processes are *not* independent of one another, and indeed, a crucial aspect of historical change lies in their interrelationship. Consider Theda Skocpol's groundbreaking study of great revolutions (1979). In order to explain the causes of social revolutions *in general*, Skocpol treated the Russian, French, and Chinese revolutions as independent cases and compared their causes with one another using John Stuart Mill's famous method of causal inference. Jack Goldstone's (1986) innovative work on demography and the institutional basis of revolutions extended the comparative ambit even more broadly throughout the early modern world. However, as William Sewell, Jr (1996) and Michael Burawoy (1989) argued, revolutionaries understood themselves to be mimicking their predecessors (or at least trying to avoid the mistakes of earlier revolutions), so that later revolutions descended from earlier ones rather than simply appearing alongside them in the historical record. As do historians, historical sociologists continue to grapple with the importance of sequence and causality in history (see, e.g., Abbott, 1995; Haydu, 1998).

Partly to address these kinds of substantive and methodological gaps in the second wave, historical sociologists in a nascent 'third wave' have looked to 'repressed' aspects of the second wave while pushing historical sociology in fresh directions (Adams et al., 2005). This ongoing movement has yet to coalesce around central findings in the same way as the second wave ultimately did, but so far it has begun a fruitful dialogue between the second wave's emphasis on state–society relations and larger questions of culture, religion, race and ethnicity, gender and sexuality, agency, and the generative properties of social interaction. Moreover, this dialogue has already yielded some provocative new directions of study. For example, Jack Goldstone recently suggested (2001), in contrast to his earlier (1986) work on demography and the causal basis of early modern revolutions, that revolutions be treated as emergent phenomena. Rogers Brubaker (1992), Margaret Somers (2008), and Ron Aminzade (2013), in quite different ways, have brought culture to the forefront in analyzing the development of varied patterns of national citizenship. Eiko Ikegami (1995) and Philip Gorski (2003) both demonstrated how culturally rooted modes of social discipline transformed over time in relationship to the foundation of modern states. Adams (2005), Charrad (2001), and Clemens (1997: esp. Ch. 6) have addressed how gender (historically specific forms of masculinities and femininities) intertwined with political and social regimes in the early modern and modern eras in ways intrinsic to institutional outcomes like state formation and political upheaval. And George Steinmetz (2007) and Julian Go (2011) revitalized interest in the sociological studies of empire, while Miguel Centeno (1997), James Mahoney (2010), Yang Su (2011), Xiaohong Xu (2013), and others have pushed the analysis of regime transformation, development, and social violence well beyond Europe's boundaries.

This third-wave dialogue with earlier findings in historical sociology was accompanied by revitalized interest in the field among a broad class of neoinstitutional and economic social scientists. Led by Viviana Zelizer (1985), Frank Dobbin (1997), Neil Fligstein (1993), Monica Prasad (2006), Bruce Carruthers (1999), Marion Fourcade (2009), and Lyn Spillman (2012), these works used historical evidence and analysis of change over time to trace the divergent institutional structures that shaped contemporary political economy. These range from nationally rooted conceptualizations of appropriate market boundaries (in the case of Dobbin) to how the economic evaluation of children shows the changing configurations of market, state, and family (in Zelizer's case) to how economic crises filter through firms' 'conceptions of control' (Fligstein) to how interactions of the political style of organized labor and state regulation produce alternative 'politics of free markets' in advanced economies (Prasad). Today, because of the ongoing globalization of these institutions as well as the underlying documentary sources, a clear new frontier of economic sociology will be systematically evaluating developments in emerging economies in the same light.

Beyond its new substantive directions for research, the third wave is also awash in exciting methodological discussion and experimentation. Some have borrowed modes of analysis and methods from other parts of the social sciences to make innovative historical arguments. For example, Edgar Kiser (e.g., Kiser and Linton, 2002) and Ivan Ermakoff (2008) used game-theoretic models of social interaction to explain macrolevel outcomes in regime behavior, and Andreas Wimmer and Yuval Feinstein (2010) employed statistical modeling of territorial change among systems of political regimes to rigorously test second-wave theories. In terms of broader discussion, the map of explicit methodological formulations in historical sociology today has at least three noteworthy locations. Some (e.g., Ragin, 1989; Mahoney and Rueschemeyer, 2003) have sought to buttress and update Skocpol's invocation of Mill's comparative method to accommodate large-N studies and the complex realities of less-bounded entities and muddy social causation. Others have taken a very different tack, emphasizing the deep interpretative grounding of historical events in the meanings and worldviews of the participants (Bonnell and Hunt, 1999; Reed, 2011). Still others have sought to steer a course between these alternatives, emphasizing the possibility of 'eventful' social change based on the emergence, invocation, and transformation of local social meanings (Clemens, 2007; Padgett and Powell, 2012; Sewell, 2005). Together, these approaches have addressed (and questioned) how to sustain persuasive social-scientific arguments engaging historical events when the independence of the events being analyzed cannot be taken for granted and how people's shared capacities and collective imaginations develop and come to obtain explanatory force.

Since these deeper methodological debates swirl around specificity versus generalizability and the roles of context and culture in the explanation of social change – areas of discussion that are quite familiar to academic historians – the relationship between historical sociology and history itself is changing as well. In an earlier edition of this very encyclopedia article, for example, Charles Tilly described the difference between the practices of history and historical sociology in these terms: "Professional historians…specialize in particular times and places, for example, modern Latin America of ancient China…meanwhile, sociologists divided up their own specialties mostly by structures and processes – families, religions, industrialization, crime, and so on – and drew their evidence chiefly

from observation of the countries they currently lived in" (2001, p. 6753). This line is significantly more blurred today. In part, this is because connections among the disciplines have strengthened as universities reconfigured following the global financial crisis of 2008. Thus, while some continue to insist on a sharp distinction between the practices of both disciplines, many other historical sociologists have begun employing archival materials, and historians have increasingly explicitly evaluated social-scientific theories (Mayrl and Wilson, n.d.).

More broadly, however, the blurring and productive tension among history, historical sociology, and other social sciences reflects what Abbott has called the "Chaos of Disciplines" (2001: esp. Ch. 4) – that is, the complicated struggle to construct a productive dialogue between history and social science while also maintaining a legitimate position as a scientific enterprise (for recent surveys of this terrain, see Lachmann (2013) and Lang (2012)). This tense process is characterized by the emergence of 'historical' branches of inquiry in social science disciplines adjacent to sociology. For example, scholars housed in political science or in crossover spaces produce work largely indistinguishable from third-wave historical sociology: for instance, *The Search for American Political Development* (Orren and Skowronek, 2004) and the body of excellent historically informed investigations of welfare policy processes (compare, for example, Weir (1992) and Kahl (2005) with Fox (2012)) as well as work by historical sociologists of interest in comparative politics and the study of American political processes (e.g., DeLeon et al., 2009; Riley, 2005; Slez and Martin, 2007). Likewise, neoinstitutional economics has long emphasized the dependence of economic development on preceding historical events, in both historical-descriptive (North and Weingast, 1989) and more formalized modes (Greif, 2006). Yet, as each of these subdisciplines turns to history, its practitioners also continually are made to justify themselves as producing something 'more than history,' and are frequently asked to defend the validity, methodology, and generalizability of their findings to their own disciplinary mainstreams.

Beyond these constraints of mainstream justification, this give-and-take of disciplinary cross-fertilization and justification produces enormous opportunity for the future of historical sociology. The improved dialogue among historically oriented sociologists, historians, political scientists, and economists has created a novel and important opportunity: to use the deep emphasis of history on the embeddedness of action within the tectonics of social structures to unwind the decades-long 'Parsonian Bargain' that has cordoned off the challenges of social science history from the discipline of economics in particular. Call it, if you like, another aspect of the 'Piketty moment' (Piketty, 2014). In addition, the range of claims made about the scope conditions of social-scientific theories across historically oriented inquiries presents a chance for a thoroughgoing reconstruction of the sociology of knowledge (e.g., Camic et al., 2011; Kennedy, 2014; Reed, 2011), probing the making of seemingly objective knowledge from socially embedded positions by both historical actors and social scientists and historians today. Finally, the dramatic resurgence of proto-, semi-, and nonstate forms of governance across the world – comprising part of contemporary globalization – impels analysis by sociologists, and social science historians more generally. If comparative historical sociology did not exist, one would certainly have to invent it.

See also: Action, Theories of Social; Bureaucracy, Sociology of; Capitalism: Global; Collective Behavior, Sociology of; Community Sociology; Differentiation: Social; Economic Sociology; Institutions; Knowledge, Sociology of; Legitimacy, Sociology of; Marxism in Contemporary Sociology; Modernity; Political Sociology; Power in Society; Social Changes: Models; Social Identity in Sociology; Sociology, Epistemology of; Tradition: Social.

Bibliography

Abbott, Andrew, 2001. Chaos of Disciplines. University of Chicago Press, Chicago, IL.
Abbott, Andrew, 1995. Sequence analysis: new methods for old ideas. Annual Review of Sociology 21, 93–113.
Adams, Julia, 2005. The Familial State. Ruling Families and Merchant Capitalism in Early Modern Europe. Cornell University Press, Ithaca, NY.
Adams, Julia, Clemens, Elisabeth S., Orloff, Ann Shola (Eds.), 2005. Remaking Modernity: Politics, History, and Sociology. Duke University Press.
Aminzade, Ronald, 2013. Race, Nation, and Citizenship in Post-Colonial Africa: The Case of Tanzania. Cambridge University Press, Cambridge.
Bonnell, Victoria E., Hunt, Lynn Avery, 1999. Beyond the Cultural Turn: New Directions in the Study of Society and Culture. University of California Press.
Brubaker, Rogers, 1992. Citizenship and Nationhood in France and Germany. Harvard University Press, Boston, MA.
Burawoy, Michael, 1989. Two methods in search of science. Theory and Society 18 (6), 759–805.
Camic, Charles, Gross, Neil, Lamont, Michele, 2011. Social Knowledge in the Making. The University of Chicago Press, Chicago, IL.
Carruthers, Bruce G., 1999. City of Capital: Politics and Markets in the English Financial Revolution. Princeton University Press.
Centeno, Miguel Angel, 1997. Blood and debt: war and taxation in Nineteenth-Century Latin America. American Journal of Sociology 102 (6), 1565–1605.
Charrad, Mounira, 2001. States and Women's Rights: The Making of Postcolonial Tunisia, Algeria, and Morocco. University of California Press, Berkeley, LA.
Clemens, Elisabeth S., 1997. The People's Lobby: Organizational Innovation and the Rise of Interest Group Politics in the United States, 1890–1925. University of Chicago Press.
Clemens, Elisabeth S., 2007. Toward a historicized sociology: theorizing events, processes, and emergence. Annual Review of Sociology 33, 527–549.
Connell, R.W., 1997. Why is classical theory classical? American Journal of Sociology 102 (6), 1511–1557.
de Leon, Cedric., Desai, Manali., Tugal, Cihan, 2009. Political articulation: parties and the constitution of cleavages in the U.S., India, and Turkey. Sociological Theory 27 (3), 193–219.
Dobbin, Frank, 1997. Forging Industrial Policy: The United States, Britain, and France in the Railway Age. Cambridge University Press.
Elias, Norbert, 2000. The Civilizing Process: Sociogenetic and Psychogenetic Investigations, rev. ed. Blackwell Publishers, Oxford.
Ermakoff, Ivan, 2008. Ruling Oneself Out: A Theory of Collective Abdications. Duke University Press.
Evans, Peter B., Dietrich, Rueschemeyer, Theda, Skocpol (Eds.), 1985. Bringing the State Back In. Cambridge University Press, Cambridge.
Fligstein, Neil, 1993. The Transformation of Corporate Control. Harvard University Press.
Foucault, Michel, 1977. Discipline and Punish: The Birth of the Prison. Random House LLC.
Fourcade, Marion, 2009. Economists and Societies. Princeton University Press, Princeton.
Fox, Cybelle, 2012. Three Worlds of Relief: Race, Immigration, and the American Welfare State from the Progressive Era to the New Deal. Princeton University Press, Princeton, NJ.
Go, Julian, 2011. Patterns of Empire: The British and American Empires, 1688 to the Present. Cambridge University Press, New York.
Goldstone, Jack, 1986. Revolution and Rebellion in the Early Modern World. University of California Press, Berkeley/Los Angeles, CA.
Goldstone, Jack, 2001. Toward a fourth generation of revolutionary theory. Annual Review of Political Science 4, 139–187.

Gorski, Philip, 2003. The Disciplinary Revolution: Calvinism and the Rise of the State in Early Modern Europe. University of Chicago Press, Chicago.
Greif, Avner, 2006. Institutions and the Path to the Modern Economy. Lessons from Medieval Trade. Cambridge University Press, Cambridge.
Haydu, Jeff, 1998. Making use of the past: time periods as cases to compare and as sequences of problem solving. American Journal of Sociology 104 (2), 339–371.
Ikegami, Eiko, 1995. The Taming of the Samurai: Honorific Individualism and the Making of Modern Japan. Harvard University Press.
Kahl, Sigrun, 2005. The religious roots of modern poverty policy: catholic, lutheran, and reformed protestant traditions compared. European Journal of Sociology 46 (1), 91–126.
Kennedy, Michael, 2014. Globalizing Knowledge: Intellectuals, Universities, and Publics in Transformation. Stanford University Press, Stanford, CA.
Kiser, Edgar, Linton, April, December 2002. The hinges of history: statemaking and revolt in early modern France. American Sociological Review 67, 889–910.
Lachmann, Richard, 2013. What Is Historical Sociology. John Wiley & Sons.
Lang, Matthew, 2012. Comparative-historical Methods. SAGE Publications.
Magubane, Zine, 2004. Bringing the Empire Home: Race, Class, and Gender in Britain and Colonial South Africa. University of Chicago Press.
Mahoney, James, 2010. Colonialism and Postcolonial Development. Cambridge University Press, Cambridge.
Mahoney, James, Rueschemeyer, Dietrich, 2003. Comparative Historical Analysis in the Social Sciences. Cambridge University Press.
Mann, Michael, 1986. The Sources of Social Power: A History of Power from the Beginning to A.D. 1760. Cambridge University Press, Cambridge.
Mayrl, Damon, Nicholas Hoover Wilson. n.d.What do historical sociologists do all day? Historical methods in theory and practice. Manuscript.
Moore, Barrington, 1966. Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World. Beacon Press, Boston.
North, Douglass C., Weingast, Barry R., December 1989. Constitutions and commitment: the evolution of institutional governing public choice in Seventeenth-Century England. The Journal of Economic History 49 (4), 803–832.
Orren, Karen, Skowronek, Stephen, 2004. The Search for American Political Development. Cambridge University Press, Cambridge.
Padgett, John F., Powell, Walter W., 2012. The Emergence of Organizations and Markets. Princeton University Press.
Piketty, Thomas, 2014. Capital in the Twenty-first Century (Arthur Goldhammer, Trans.). Harvard University Press, Boston, MA.
Prasad, Monica, 2006. The Politics of Free Markets: The Rise of Neoliberal Economic Policies in Britain, France, Germany, and the United States. University of Chicago Press.
Ragin, Charles C., 1989. The Comparative Method: Moving beyond Qualitative and Quantitative Strategies. University of California Press.
Reed, Isaac Ariail, 2011. Interpretation and Social Knowledge. On the Use of Theory in the Human Sciences. The University of Chicago Press, Chicago.
Riley, Dylan, 2005. Civic associations and authoritarian regimes in inter-war Europe: Italy and Spain in comparative perspective. American Sociological Review 70, 288–310.
Sewell Jr., William H., 1996. Three temporalities: towards and eventful sociology. In: The Historic Turn in the Human Sciences. University of Michigan Press, Ann Arbor, pp. 245–280.
Sewell Jr., William H., 2005. A theory of structure: duality, agency, and transformation. In: Logics of History: Social Theory and Social Transformation. University of Chicago Press, Chicago.
Skocpol, Theda, 1979. States and Social Revolutions: A Comparative Analysis of France, Russia, and China. Cambridge University Press, Cambridge.
Slez, Adam, Martin, John Levi, 2007. Political action and party formation in the United States constitutional convention. American Sociological Review 72, 42–67.
Somers, Margaret, 2008. Genealogies of Citizenship. Markets, Statelessness, and the Right to Have Rights. Cambridge University Press, Cambridge.
Spillman, Lyn, 2012. Solidarity in Strategy: Making Business Meaningful in American Trade Associations. University of Chicago Press.
Steinmetz, George, 2007. The Devil's Handwriting. University of Chicago Press, Chicago.
Steinmetz, George, 2013. Sociology and Empire: The Imperial Entanglements of a Discipline. Duke University Press.
Su, Yang, 2011. Collective Killings in Rural China during the Cultural Revolution. Cambridge University Press.
Tilly, Charles, 2001. Historical sociology. In: International Encyclopedia of the Behavioral and Social Sciences, pp. 6753–6757.
Tilly, Charles (Ed.), 1975. The Formation of States in Western Europe. Princeton University Press, Princeton.
Wallerstein, Immanuel, 2011. The Modern World-system I: Capitalist Agriculture and the Origins of the European World-economy in the Sixteenth Century, with a New Prologue. University of California Press.
Weir, Margaret, 1992. Politics and Jobs: The Boundaries of Employment Policy in the United States. Princeton University Press, Princeton, NJ.
Wimmer, Andreas, Feinstein, Yuval, 2010. "The rise of the nation-state across the world, 1816 to 2001. American Sociological Review 75 (5), 764–790.
Xu, Xiaohong, 2013. Belonging before believing: group ethos and bloc recruitment in the making of Chinese communism. American Sociological Review 78 (5), 773–796.
Zelizer, Viviana A. Rotman, 1985. Pricing the Priceless Child: The Changing Social Value of Children. Princeton University Press.

Historical Thought and Historiography: Christian Tradition

Christoph Markschies, Theologische Fakultät der Humboldt-Universität zu Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ever since the beginnings of orderly reflections on the topic of history in Christianity and the first literary attempts to write history, four basic problems have been characteristic of all such attempts. All Christian historical thought and also all Christian historiography can be characterized by basic decisions concerning the following four problem areas: (1) the classification as a specific inside or outside perspective on Christianity; (2) the respective understanding of the position of one's interpretation on a range between historical and theological standards; (3) specific mechanisms of selection; and (4) a small number of models by the help of which the past is reconstructed. These four basic problems of Christian interpretations of history cannot be solved by means of an isolated special theological hermeneutics. They call for an original contribution to the general discussion of history and of the interpretation of history.

Ever since the beginnings of orderly reflections on the topic of history in Christianity and the first literary attempts to write history (by the author of *The Gospel According to St. Luke*, AD *c*.81–96, cf Luke 1, 1–4 and Eusebius of Caesarea, AD *c*.264–340, *Ecclesiastical History*), four basic problems have been characteristic of all such attempts. Their uninterrupted significance from the ancient world up to the present allows development of a history of problem areas of Christian historical thought. All Christian historical thought and also all Christian historiography can be characterized by basic decisions concerning the following four problem areas: (1) the classification as a specific inside or outside perspective on Christianity (Section The Problem of Inside Perspective and Outside Perspective); (2) the respective understanding of the position of one's interpretation on a range between historical and theological standards (Section Christian Interpretation of History – Historical and Theological Standards); (3) specific mechanisms of selection (Section The Mechanisms of Selection of Christian Interpretation of History); and (4) a small number of models by which the past is reconstructed (Section The Basic Models of a Christian Interpretation of History).

The Problem of Inside Perspective and Outside Perspective

A Christian understanding of history describes Christianity from an inside perspective; but in many cases such descriptions can, to some extent be regarded as outside perspectives. This is because, together with an outside perspective on Christianity, for example, by non-Christians, there are also outside perspectives of Christian minorities on Christian majorities, and vice versa. Such perspectives have a special forming influence on the understanding of history: The *Ecclesiastical History* by Theodoret of Kyrrhos (AD *c*.393–466), one of the most definitive works of Christian historians of the second generation after Eusebius, regards historical developments as formed by the debates between orthodox church ('orthodoxy') and apostate heresy. The *Ecclesiastical History* by his contemporary Philostorgius (AD *c*.368–425) has the same perspective, however, with the opposite judgment: the theological tradition of the Alexandrine presbyter Arius (AD *c*.260–336), which has influenced theological thought mainly in Great Britain up to the present, is regarded as true orthodoxy by Philostorgius but as apostate heresy by Theodoret. Besides these, there have always existed more or less critical descriptions from an outside perspective by non-Christians (e.g., by the Roman historian Ammianus Marcellinus, AD *c*.330–395, in the heyday of Christian historiography in late antiquity) in which Christian historians were tackled implicitly or explicitly.

It was only in modern Europe that Christian historians have made determined attempts toward a reduction of inside-perspective elements in Christian historiography. This is, of course, mainly with respect to the narrow horizon of thought formed according to the self-awareness of the three Christian denominations that arose in Central Europe after the religious wars of the sixteenth and seventeenth centuries (the Religious Peace of Augsburg 1555 and the Peace of Westphalia 1648). At the beginning of the modern era, reflecting denominational antagonism, the Lutheran, Reformed, and Roman Catholic interpretations of history entered into competition. The protagonist of the overcoming of this denominational approach to the history of Christianity was Gottfried Arnold (1666–1707; Professor of History in Giessen 1697–98). He tried to write an '*unpartheiische*', which is a nondenominational church history. In the twentieth century, following the institutionalization of the ecumenical movement, which has tried to overcome the denominational splittings of the sixteenth century, the call for a general 'ecumenical church history' (Ernst Benz) arose. In a widespread late twentieth-century attempt to conceptualize the alternatives, 'church history' as opposed to 'history of Christianity' has been used. 'Church history' describes a perspective orientated toward the Christian denominations, and 'history of Christianity' describes a trans-denominational perspective.

Christian Interpretation of History – Historical and Theological Standards

A Christian interpretation of history experiences a tension between historiography (or history) on the one hand and

Christian theology on the other. This tension has existed from the very beginning insofar as theology (which in this context means the theory of the Christian faith) naturally favors certain theoretical options when reconstructing a history, for example, its assumption of a god acting in history, which other interpretations of history may hold but are by no means bound to hold. Such specific theoretical options always have methodical consequences; this is what the religious philosopher Ernst Troeltsch (1865–1923) showed in his clear analysis and what has been called 'historical or dogmatic method' (Troeltsch, 1922: pp. 729–753). But one must not overestimate the 'theological' elements of Christian interpretations of history before the modern era: from the fourth century they had as a matter of course been orientated toward the literal and methodical standards of the interpretation of history of their respective times. This theoretical background was broken only very partially by specific religious or theological concepts. The first author of an ecclesiastical history, Eusebius of Caesarea, already followed the forms and standards of the historiography of his time in his historical writing, as can be seen in the prologue (Eusebius, historia ecclesiastica 1, 1, 3f). Consequently, his account leaves out central questions of a theological interpretation of history. Because pagan ancient historiography very rarely explicitly treated the difficult relation between divine acting in history through fate (τυχη/*tyche* or *fatum*) and human freedom, Eusebius, too, never picked out the exact relation between man's freedom and God's acting in history as a central theme.

The fundamental problem of the relation between an interpretation of history according to modern standards on the one hand and specific theological premises on the other hand intensified in modern Europe. In the course of the modern secularization of historical thought, under the premise of *etsi Deus non daretur* (God is not necessary), new determinants of history (such as fate, chance, reasons of state, and ethics) were established in place of the Christian approach to a history regarded as being under the law of God. In consequence, the problem of how – under these conditions – a theological view of history was possible increased. Most Christian interpretations of history after the Enlightenment have, however, tried to take up the problem, often through a religious or theological interpretation of the philosophical framework of the historiographic theory. The Berlin theologian Friedrich Schleiermacher (1768–1834), for example, made the following point in his notes of the years 1796–99: "*Die Historie ist immer religiös und die Religion muß ihrer Natur nach historisch sein*" ("history is always religious and religion must be historical by nature", Schleiermacher, 1984: p. 25). At the same time the possibility of a Christian interpretation of the world, which was lost during the period of Enlightenment, was meant to be restored by means of such a religious charging up of history.

The Mechanisms of Selection of Christian Interpretation of History

All interpretation of history and, furthermore, all historical writing, is characterized by acts of selection from a huge amount of material. Within the framework of a Christian interpretation of history, first the basic choice of subject must be made: to date, it has been heavily disputed whether a Christian interpretation of history merely ought to deal with certain fields of general history such as Christianity, the history of certain denominations or ecclesiastical groups, or a particular history formed in a particular way by a transcendental power ('*Heilsgeschichte*', 'salvation history'), or whether a Christian interpretation of history ought rather to come up with a distinctive Christian perspective on history as a whole called 'world history' since the end of the eighteenth century. The controversies over whether the chosen ideological perspective necessarily leads to a concentration on an ecclesiastical history, or on a history of Christianity, or alternatively on the wider framework of a universal history, has always been an issue in Christian historical writing.

The first Christian author of a historical monograph, Eusebius, chose to concentrate on an ecclesiastical history without any detailed consideration of the political history of his time. A universal history, now lost, was presented by the Greek presbyter Philippus of Side less than 100 years later, before AD 439. The Christian interpretation of history that has had the strongest influence up to the modern era, and which was fundamental in the Middle Ages, was written by the North-African theologian Augustine (AD 354–430). It takes the perspective of a universal history: after the city of Rome had been captured by the Goths in AD 410, both pagan and Christian interpretations of history, which until then had their focus in the Empire, broke down because of the obvious collapse of the Roman Empire. As a consequence, the bishop was bound to develop his own distinctive concept of a universal history which would be able to convey a new sense of history. Augustine presented it with his work *De Civitate Dei*, '*The City of God*' (AD 413–426; for details see Section The Basic Models of a Christian Interpretation of History).

While the Middle Ages mainly expressed concepts of universal history based on Augustine's, for example, by Otto of Freising (*c*.1112–58), the traditional Christian universal–historical interpretation broke down during the Enlightenment. Theological categories, such as 'God's revelation' or 'God's acting', which had been systematic premises of the Christian interpretation of a universal history, and had consequently formed a unity of transcendental and earthly reality, became problematic. After the attempts of the nineteenth century to recreate such a concept of unity on a new basis, and thus to once more present a Christian interpretation of a universal history (in very different ways by Schleiermacher, Hegel, and Ranke), the twentieth century progressively withdrew from such theological interpretations of history. Christian interpretation and presentation of history in the twentieth century were more or less confined to the specific context of Christianity and church, even if the theoretical basis was sometimes broader. That was the systematic reason for the modern criticism of concepts which attempted to create a close correlation between the history of human development and God's acts of salvation, by way of 'salvation history', the theological interpretation of certain incidents.

Along with the fundamental question of whether a Christian interpretation of history must be understood universally, or be restricted to a history of Christian institutions, there are other mechanisms of selection taken from general historiography which are of importance to a Christian interpretation of

history. Historians of the ancient world and the Middle Ages used, as important mechanisms of selection, the questions of the meaning of an incident, of its benefit for future generations, or of its documentation in the respective privileged sources, and so forth. The bias thus created is obvious. However, in general, the critical sense with which historians before the modern era treated their own modes of selection is underestimated. Such reflections were usually made in the prologue of historiographical works.

The vehement calling into question of traditional mechanisms of selection in the general history of the twentieth century (see, e.g., the criticism of any historical interest with a biased concept of histories of ideas or politics, the criticism of the concentration on Europe, and of the neglect of feminist research) reached the historiography of Christian denominations with a certain delay; its consequences on a general Christian historical awareness only gradually become apparent.

The Basic Models of a Christian Interpretation of History

Christian conceptions of history up to the modern era usually have reconstructed it on the basis of a small number of comparatively simple models. Eusebius describes the history of the church according to a model of a process of constant 'advances', something which is in strict contrast to the historiography of his times with its stress on cyclic processes and the decadence of traditional Roman values. The history of his own religion teleologically leads to the Christian empire of Constantine. Augustine breaks with Eusebius' model of a progress with his concept of two cities (*civitates*) wandering through the ages. He differentiates between the *civitas Dei*, the 'citizenry of God' as an earthly community of the heavenly citizens, and the *civitas diaboli*, 'the citizenry of the devil'. Of course, the Roman state and the community of those who do not live under God and toward him are not identical in the sense of the state as such being the *civitas diaboli*. But Augustine takes up the traditions of radical Christian criticism of the state, and with his anthropologically based historiographic model gives a clear analysis of the social deficiencies of his time. The European Reformation breaks with this conception. It mainly follows the models of decadence developed by the medieval church reformers, which can also be found with humanistic authors: the promising beginnings of the first Christians had decayed in a middle period which is only now brought to light again. This conception was monumental, and had a strong anti-Catholic impetus in the so-called Magdeburger Zenturien – a team effort by protestant theologians (Basel 1559–74) led by the strict Lutheran Matthias Flacius Illyricus (1520–75), which has been known under that name since the eighteenth century, and which has remained a fragment. Along with a comparatively simple modeling, the work is characterized by the use of an enormous number of sources. Cardinal Caesar Baronius (1538–1607) reacted to this challenge with his *Annales Ecclesiastici*. But this catholic reply, in terms of method only, mirrored the protestant work, and it tried to answer the reproach of having broken with the truth with a proof of uninterrupted continuity from the age of the early Christian church.

The model of decadence gained special influence (e.g., in Goethe's work); Arnold applied as a basis – despite his programmatic orientation toward an impartial standpoint. Arnold explained the history of Christianity according to a model of progressive decadence, beginning after the pure earliest church. Despite his persistent interest in counterevidence to official protestant ecclesiastical historiography, he held to his conception of decadence, and to a structural principle of the denominational history of Christianity. Such simple basic models for the reconstruction of the history of Christianity were by no means, or first, only questioned by the European Enlightenment. Augustine's criticism of Eusebius' conception of progress in *De Civitate Dei* is an example, because it was motivated by an empirical diagnosis of society and by critical work on biblical texts. Humanist historians of the Reformation must also be mentioned, particularly Philipp Melanchthon (1497–1560) and Johannes Sleidanus (1505–56). In his short world history and in a work on contemporary history entitled *Denkwürdigkeiten, betreffend Staat und Religion unter Kaiser Karl V*, strict models of decadence are missing. Sleidanus always endeavored to describe religious changes with regard to imperial power and state power; he can be regarded as one of the progenitors of a strong orientation toward political history.

Along with the orientation toward simple conceptions of progress or decadence, the history of Christian historiography, and the Christian awareness of history, has called into question oversimplified models up to the present. In Ranke's formula, for example, each epoch is immediate to God; divisions into periods are good indicators for the respective underlying models. In this respect, denominational Christian interpretation of history tends to differ from the standards of general history; protestant conceptions, for example, speak of a 'period of Reformation' as a threshold between the Middle Ages and the modern era.

The four basic problems outlined above have characterized Christian concepts of history from its beginnings, and they show that the crucial changes in Christian interpretation of history caused by the increasing influence of the methodical standards of modern historiography and finally by the victory of historicism, must not be overlooked in theology, either. The victory of historicism has had a significant consequence on academic theology as a whole, because the historization of all theological disciplines was demanded (among others by Troeltsch), and was partially carried out. It is equally clear and almost undisputed that the general 'crisis of historicism' also had an impact on theological form of Christian interpretation of history, something which had already been diagnosed by contemporary church historians like Karl Heussi (1877–1961). Mainly in continental Protestant theology, this general crisis was accompanied by a theological criticism of general historicism. This development is now called an 'antihistorical revolution' (Graf, 1988: pp. 377–405), a term which points to its consequences: the removal of church history as a leading subject of academic theology.

In the twentieth century, there was an interesting revitalization of apocalyptic thought on a large scale, the relation of which to the 'crisis of historicism' is not immediately clear. This form of Christian theology, which goes back to a movement in Hellenistic Judaism, and which had already left deep marks on

certain biblical texts, is characterized by a distinctive awareness of history. The present time is interpreted as the 'last age', and forms part of a certain fixed scheme of the course of history, while there is no historical interest in the identification of certain 'signs of the times'. In the ancient world, the formation of a Christian awareness of history was accompanied by a drastic loss of meaning of apocalyptic elements for Christianity. The eruptive revival of apocalyptic trends in the Middle Ages, the early modern era, and the present has in most cases been accompanied by a protest against Christian interpretations of history orientated toward the standards of the respective times.

Current Problems and Questions of Christian Interpretation of History

The systematization of current schools of Christian interpretation of history meets with the difficulty of the actual presence of pluralism (Nowak, 1997: pp. 3–12), perceived a 'new confusion'. Such concepts are present in the institutional fields of denominational church history as well as independent of these. The Christian inside perspective, considered to embody prejudice in some accounts, presents a challenge to other approaches to the problems of Christian historicism; they have not, however, discarded it. This shows the limited significance of the presently popular alternatives of a 'church history' and a 'history of Christianity'. In both cases, attempts to take an outside perspective are, more or less admittedly, wrapped in an inside perspective.

Within the horizon of the standards of history at the beginning of the third millennium on the one hand, and of Christian theology on the other, almost all schools have tried to achieve careful detheologization. This, however, is rarely picked out as a central theme. It is only very occasionally that a decidedly theological contribution to the general discussion on history is demanded (Herms, 1997: pp. 305–330); a 'resort to pragmatism' can be noted (Nowak, 1997: p. 12). Even if some historians still demand a concept of universal history in Christian terms as regards the selection of subject matter, most publications actually present contributions to a discipline of general history. That such presentations are in most cases characterized by the leading assumptions of universal history (and theology) is obvious to any sensitive examination; but it is only rarely admitted or made explicit. The presently vigorous call for an approach to church history, or history of Christianity, as historical–cultural studies also implies certain universal–historical structures in the shape of an '*histoire totale*'.

The critical calling into question of historiographical mechanisms of selection and of modeling, which mainly arose in the twentieth century, has increasingly been accepted by the institutionalized Christian interpretation of history. But it is very rare that it is understood as one of the consequences of fundamental decisions made in the course of the European Reformation movement which have created a high level of ideology-critical potential.

These four basic problems of Christian interpretations of history cannot be solved by means of isolated special theological hermeneutics. They call for an original contribution to the general discussion of history and of the interpretation of history. This is especially true in view of the latest developments in 'writing history in the postmodern era': can the concept of a literal sense as the regulatory concept of the process of interpretation really be given up in the course of the deconstruction of its own narrative? Does not a Christian interpretation of history which does away with the 'grand narratives', and the 'master processes' dissolve itself? Is the notion of 'God's acting' dispensable in a Christian interpretation of history? Here, a new and interesting challenge is put to the long-neglected field of a theory of Christian interpretation of history. It is possible that a theoretically renewed, clear-cut, Christian interpretation of history helps to preserve in a positive sense two functions of the Christian interpretation of history under the conditions of the postmodern era: the function of forming society, and the function of critically analyzing society. But with all this, the traditional objections to a simplistically designed history of salvation remain important.

See also: Historical Thought and Historiography: Classical Period (Especially Greece and Rome); Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; History: Theories and Methods.

Bibliography

Bebbington, D.W., 1990. Patterns in History: A Christian Perspective on Historical Thought. Baker Book House, Grand Rapids.

Graf, F.W., 1988. Die "antihistorische Revolution" in der protestantischen Theologie der zwanziger Jahre. In: Rohls, J., Wenz, G. (Eds.), Vernunft des Glaubens. Festschrift W. Pannenberg. Vandenhoeck & Ruprecht, Göttingen, Germany, pp. 377–405.

Herms, E., 1997. Theologische Geschichtsschreibung. Kirchliche Zeitgeschichte 10, 305–330.

Markschies, C., 1995. Arbeitsbuch Kirchengeschichte. JCB Mohr, Tübingen.

Markschies, C., 2001. Kirchengeschichte. In: Goertz, H.-J. (Ed.), Geschichte. Ein Grundkurs, second ed. Rowohlt, Reinbek bei Hamburg, Germany, pp. 408–422.

Meinhold, P., 1967. Geschichte der kirchlichen Historiographie. Karl Alber, Freiburg, Germany.

Nowak, K., 1997. Wie theologisch ist die Kirchengeschichte? Theologische Literaturzeitung 122, 3–12.

Schleiermacher, F., 1984. Schriften aus der Berliner Zeit 1796–1799. In: Meckenstock, G. (Ed.), Kritische Gesamtausgabe KGA I/2. W. de Gruyter, Berlin.

Tillich, P., 1968. In: Braaten, C. (Ed.), A History of Christian Thought from Its Judaic and Hellenistic Origins to Existentialism. SCM Press, London, pp. 33–34.

Troeltsch, E., 1922. Ueber historische und dogmatische Methode in der Theologie. In: Zur religiösen Lage, Religionsphilosophie und Ethik. JCB Mohr, Tübingen, pp. 729–753.

Historical Thought and Historiography: Classical Period (Especially Greece and Rome)

Wilfried Nippel, Humboldt-Universität zu Berlin, Philosophische Fakultät I, Institut für Geschichtswissenschaften, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since the subsequent fifth century BC, Greek historians developed a critical attitude to the past, invented methods that made it possible to discern between facts and fiction. The adoption of historiography by the Romans provided the model for constructing the past in the shape of national history. The Christian writers invented ecclesiastical historiography as a subject of its own and a model for universal chronicles and histories starting from the creation.

Historiography, the rational reconstruction of the past in the medium of a prose that claims to literary quality was developed in Greece during the fifth century BC. Its emergence reflects the peculiar intellectual and political developments of this age. The high cultures of Egypt and Mesopotamia had known lists of rulers, their foundations, festivals, and wars that represented the official version according to which the monarch inevitably would crush enemies. There was no need and no chance to develop a conception of the past in which the course of events would be ascribed to the contending efforts of multiple actors, the result of which was unpredictable and thus needed to be reconstructed and narrated in each case. The historical books of the Old Testament, in a certain sense, took account of the contingency of human affairs but at the same time attributed the outcome of events to God's intervention.

The Origins of Greek Historiography

After the development of literacy during the eighth century BC, Greek commemoration of a distant past had materialized in the epic tradition (Homer, Hesiod). Its combination of 'historical' and 'mythical' elements became subject to criticism in an age when sixth century-Ionian philosophy (as represented by the Milesians Thales, and Anaximander) had attempted at a 'scientific' explanation of the physical world and of human society. About 500 BC, Hecataeus of Miletus tried to reconstruct the genealogical facts out of the mythological traditions. He wrote a pioneering work of geography that reflected the expansion of the geographical and historical horizon of the Greeks as a result of their colonization (since *c.*750 BC). This included their expansion on the shores of the Mediterranean and Black Seas, their voyages of discoveries, and the information that the Eastern Greeks (as subjects of the vast Persian Empire) had acquired about parts of Asia. Historiography, in a narrower sense, developed out of the experience of the Persian Wars. However, the realization that the great victories of the Greeks in 490 and 480–479 BC had been achieved, so to speak, against all probability apparently developed only with a certain distance to the events themselves.

Herodotus (*c.*485–after 425 BC; a widely traveled native of Halicarnassus in Asia Minor) opened his account of the Persian Wars with the statement that he wanted to present the results of his research (*historie*) in order to commemorate the deeds of both Greeks and Barbarians and to establish the cause of their conflict. *Historie* is a general term for a rational inquiry. It was only subsequently (especially with Aristotle's *Poetics*) that it became the label for the literary genre that Herodotus had created. Herodotus pointed out that his study was the work of an individual and that he felt obliged to an impartial approach. Greek historiography was developed by free intellectuals and did not start as officially committed accounts. Apparently, especially exiled individuals like Herodotus – and subsequently inter alia Thucydides, Xenophon, and Polybius – were able to take on the position of an (at least, subjective) neutral observer. Herodotus did not only provide the first work of an *histoire événementielle* based on stories that might have been generated two generations before his own time but also continued the tradition of geographical and ethnographical accounts by incorporating digressions on the various peoples (from the Egyptians to the Scythians) under Persian control. Here he relied on local traditions and autopsy and sometimes on reported stories, the authenticity of which he himself doubted.

Herodotus' younger contemporary Thucydides of Athens (*c.*460–400 BC) created a quite different model of historiography. He deliberately distanced himself from Herodotus (though without naming him) as he thought Herodotus to have shown too much confidence in his sources. Thucydides' statements on his methodological principles are (with the notable exception of Polybius) almost without parallel in ancient historiography. He chose as his subject the Peloponnesian War between Athens and Sparta and their respective allies (431–404 BC; Thucydides' account, however, is incomplete and breaks off at 411 BC). According to his own words he had started to take notes at the outbreak of the war as he had anticipated the unprecedented dimensions of this war which affected the whole Greek world. Later, he used his time in exile (the consequence of his defeat as an Athenian general in 424) to collect all available facts by interrogating protagonists and eyewitnesses of both sides. Being confident that he was able to reconstruct the objective course of events he did not name his sources or quote varying, or even conflicting, versions by different witnesses but digested all information into an authoritative account of 'how it had really happened' (an aside of his which was subsequently to be taken up by Ranke). He wanted to differentiate the long-term causes, which led to a climate of mutual distrust between the two powers, from the immediate occasions, which made them take military action. War and civil war were seen by Thucydides as revealing the condition humaine, love of power, ambition, greed, and fear as constituents of people's invariable nature; thus the thoroughgoing analysis of human action should provide insights of everlasting value.

Already Herodotus had incorporated some speeches into his narrative in order to explain the motives behind the decisions of protagonists but Thucydides went much further by systematically using this stylistic device based on the techniques of rhetoric that were taught by contemporary Sophists. As Thucydides explained, his methodological principle was 'to make the speakers say, what, in my opinion, was called for by each situation while keeping as closely as possible to the general sense of the words that were actually used.' Those speeches served to present opposing arguments (in pairs of speeches ascribed to antagonists) and to discuss the political, moral, and juridical implications of certain decisions or to express the self-image of a society (especially in the 'Funeral Speech' attributed to Pericles). The unity of style, the high level of abstraction, the implicit cross-references, and the shortness of the speeches point out that they were primarily a medium to carry the historian's interpretation. Since Thucydides, the inclusion of more or less fictitious speeches was a standard feature of ancient historiography. Thucydides made two exceptions to his principle of writing a strictly contemporary history. He gave an account of the development of early Greek history in order to prove that due to the level of its material culture, a war of such dimensions like the Peloponnesian War would have been impossible. That was a sort of conjectural history based on theories of societal development and on the interpretation of cultural survivals. He also gave a selective description of the 50-year period between the end of the Persian Wars and the outbreak of the Peloponnesian War in which he stressed the rise of Athenian power and the distrust it fostered with the Spartans. This account also served to bridge the gap between Herodotus' narrative and his own.

Thucydides' model of historiography differed from the Herodotean by eliminating the features of cultural history and ethnographical discourse. But both corresponded with respect to concentrating on large-scale wars, which involved great parts of the Greek world, and with regard to treating matters of constitutional history and internal politics of the states concerned only insofar as they were relevant for the course of events. Their different attitudes to the period of time that could be reconstructed by drawing on oral traditions notwithstanding, they both primarily relied on oral sources. As a rule, down to the Christian historians of Late Antiquity ancient historiographers did not consult documentary evidence systematically like inscriptions or state papers in archives and did not engage seriously in 'antiquarian' research on constitutional, legal, and religious institutions. That is why they could not think of replacing the narratives of former historians by a new account based on other better sources. Thus Thucydides was a continuator of Herodotus, Xenophon, and other fourth century historians who started their Greek histories at the point where Thucydides' narrative had broken off. Historians who wanted to treat former times again, cannibalized (often in combination with harsh criticism) the accounts of their predecessors.

The Variety of Greek Historiographical Production

Thucydides had shaped the trend-setting model of military and political history that was followed by a great number of Greek writers including the historians of Sicily. However, his successors embellished the rhetorical elements or amplified scenes with emotive appeals to readers. Throughout antiquity, Herodotus enjoyed a dubious reputation as being not only the father of history but also, at the same time, a narrator of fabulous stories (in Cicero's famous phrase). The broader Herodotean approach, however, was not totally outdone; it was taken up again by the historians of Alexander the Great and survived in subsequent works with strong geographical and ethnographical interests like that of Posidonius (*c.*135–50 BC) and in the specialized geographical works from Eratosthenes (third century BC) to Strabo (Augustan age).

Besides the mainstream of military and political history, a number of specialized historical genres developed since the late fifth and during the fourth centuries. These included: universal history (Ephoros of Cyme, fourth century BC, whose – later lost – work was used by Diodorus Siculus, first century BC); foundation histories, local chronicles, and constitutional histories (e.g., on Athens); cultural histories (Dicaearchus, pupil of Aristotle, late fourth century BC); biographies (late fifth century BC sketches subsequently to be used by Plutarch, AD *c.*45–120; and Xenophon's life of the Spartan king Agesilaus II, fourth century BC).

During the Hellenistic Age, an enormous number of historical works of various kinds were produced by nearly a 1000 historians known to us by name. This tremendous increase of historical production reflected the political reorganization of the Mediterranean world by the establishment of the vast Hellenistic empires (and subsequently the rise of Rome to a world power). It was fostered by the patronage of Hellenistic monarchs and the establishment of libraries and scholarly institutions, especially in Alexandria and Pergamum. Almost all of this extremely rich production has, however, got lost and is only known from quotations and references in subsequent works. For example, the various reports on Alexander's campaigns by court historiographers, officials, and generals only emerge with the compilation of Diodorus; comprehensive accounts of Alexander's history based on the works of his companions dating (?) from the second and third centuries AD (Arrian and Curtius Rufus).

Polybius (*c.*200–120 BC) is the only Hellenistic historian of whose historical work a substantial amount has survived. Polybius, a military fighter of the Achaean Confederacy (on the Peloponnese) was one of 1000 leading Achaeans who were deported to Italy after 167 BC. He came into close contact with prominent figures of the Roman nobility, especially Scipio Aemilianus (the destroyer of Carthage). The subject of his historical work was Rome's rise to world dominion in the time from the Punic Wars to the destruction of the Macedonian state (167 BC). Polybius calls it a 'pragmatic history'; it focuses on political and military actions and is especially based on eyewitness reports and first-hand knowledge of the theaters of war. His intentions and explicit methodological reflections (combined with polemic against several former historians) imply a conscious return to the Thucydidean model of historiography. But Polybius understood Rome's military success as having depended on its uniquely stable, and at the same time flexible, political system. In order to explain this point to his Greek readers, he incorporated a description of the Roman constitution and military system and thus gave a systematical

treatment of institutions otherwise not to be found in ancient historiography. (His analysis of Rome's 'mixed constitution' would become one of the most important subjects in the reception of ancient political theory since the Renaissance.)

Roman Republican Historiography

Roman history was first made the subject of historiographical works by Greek authors like the Sicilian Timaeus of Tauromenium (*c.*350–260 BC) and other Greek writers who were especially interested in the stories about the foundation of Rome. Latin literature only developed from the middle of third century BC as a product of the reception of Greek culture. Certain records of the past, however, had been preserved since early times: public documents like laws, decrees of the senate and international treatises, and family records of the aristocracy. And there were the *annales* kept by the pontiffs in which the names of consuls (the eponymous magistrates) were listed, triumphs and religious prodigies were noted, and some miscellaneous information was collected. About 200 BC, the senator Fabius Pictor provided the first history of Rome written by a Roman, yet in Greek. Drawing on the models of Hellenistic historiography Fabius Pictor covered the time from the founding of the city until 200 BC. He probably gave elaborate accounts of the early, and of the contemporary periods and only a sketchy one of the time in between, a proportion that would become typical of many subsequent writers. He was followed by other historiographers writing in Greek until the middle of second century. The older Cato (234–149 BC) was presumably the first to write a historical work in Latin. His *Origines* applied the interest in foundation history not only to the earliest history of Rome but also to the emergence of the cities in Italy and paid particular attention to the unification of Italy. His treatment of political history focused on the justification of his own policy and thus provided the model for the *ex parte* tendency of Roman contemporary historiography. Roman historiography was ethnocentric and it was the product of the leisure-time activity of senators. Their intentions were to demonstrate Rome's cultural equality with the Greeks, to justify her domination over the Mediterranean world, to praise the wisdom and success of the ruling aristocracy, and to take sides with respect to contemporary conflicts. From the late second century onward, a number of Latin historiographical works were composed which followed in their structure the pontifical *annales* (which were published in a full version in 120 BC) and gave a year-by-year account from the origins of the city. They apparently embellished their descriptions of the early republic and especially the 'Struggle of the Orders' between patricians and plebeians by projecting back features of the internal conflicts that had become almost endemic since the tribunate of Tiberius Gracchus (133 BC). Some historians seem to have taken into account certain results of serious antiquarian research, others had no scruples and they invented 'documents.' (It was the Greek author Dionysius of Halicarnussus who, living in Augustan Rome, showed the greatest interest in evidence of an antiquarian nature.) All these Latin 'annalistic' works (as well as the earlier historiographical attempts previously mentioned) are only indirectly known. They were used by Livy (*c.*59 BC–AD 17) who was the first Roman historian without a record of political and military experience of his own. His history covered the time from the foundation of the city to 9 BC. Out of his 142 books, only 35, which reach down to 167 BC (with *lacunae*) have survived. Livy often struggled with the contradictory accounts of his predecessors and the quality of his own descriptions obviously depended on the quality of the sources that was available to him. He did not invent his own versions of the course of events but his work shows the problematic nature of a history totally made on the basis of other, often overtly partisan, historical works that were mostly accepted at face value.

The Thucydidean model of contemporary history was imitated by Sallust (86–35 BC) who after the failure of his political career wrote monographs on the conspiracy of Catiline (63 BC) and on the Roman Wars against the Numidian king Jugurtha (111–106 BC). He interpreted the political crisis of the Late Republic as the consequence of a moral decline following from the abandonment of traditional values. (No other contemporary work on the Late Republic has survived; a comprehensive account of this period is only available in works by subsequent authors such as Appian, second century, and Cassius Dio, third century, who both wrote in Greek.)

Roman Imperial Historiography

Writing contemporary history could become a dangerous job under the Emperors as the fate of the historian Cremutius Cordus demonstrates. Under the reign of Tiberius (in AD 25) he was indicted for *lese majesty* because of the republican sympathies he had shown in his work on the Civil War period (he committed suicide; his work was burnt). The short sketch by his contemporary Velleius Paterculus, a former soldier, covered Roman history from the very beginnings to his own days in only two books and culminated in the panegyric eulogy of the emperor Tiberius.

The historiography of Tacitus (AD *c.*55–120) reflects a new climate of intellectual freedom during the reigns of Nerva and Trajan in comparison to the experience with the tyrannical régime of Domitian (under whom Tacitus had started his public career). His (only partly preserved) works, the *Histories* and the *Annals*, cover the time AD 69–96 and AD 14–68, respectively. Tacitus kept the annalistic form and he paid special attention to the proceedings of the senate. But the form was often at odds with the contents of the works, as the policies of the rulers could not easily be fitted to a chronological scheme that had been adequate for a senatorial régime with yearly changing consuls at the head. Tacitus showed certain nostalgia for the republican liberty without indulging into illusions about returning to a republican constitution. Though claiming to write *sine ira et studio* his personal judgments on rulers (especially the gloomy picture of Tiberius) would serve to uphold the mirror of an honest ruler and his comments on senatorial proceedings reflected his standards for dignified behavior under a monarchy. Tacitus' work with its stress on the psychology of emperors and senators marks the final point of Latin senatorial historiography. (However, subsequently, there was an annalistic history of Rome from the beginnings to AD 229 written in Greek by the senator Cassius Dio.)

Apart from short summaries of Roman history, during the second century serious historiography gave way to imperial biographies. Suetonius (AD *c.*70–140, secretary for the library and subsequently for the imperial correspondence under

Hadrian) gave the history of lives and times of 12 rulers from Caesar to Domitian. They were arranged in a more topical than strictly chronological manner and drew on archival materials as well as traditions stemming from court gossip. Modeled on Suetonius but of a highly fictitious character (with numerous references to imaginary sources) was the *Historia Augusta*, a series of biographies of emperors for the time from Hadrian (AD 117–138) to the accession of Diocletian (AD 285) which was probably produced in the late fourth century by one or more anonymous writers who took on the names of bogus authors.

The Impact of Christianity

The religious tensions of Late Antiquity due to the rise of Christianity only had a limited impact on pagan historiography, which continued to concentrate on short summaries of Roman history. The general restraint from religious polemic also holds true for the work of Ammianus Marcellinus (*c.*330–395), a Greek from Syrian Antioch and former soldier. His history, written in Latin, is otherwise a great exception as he embarked on a continuation of Tacitus that treated the time from Nerva's accession to the throne (96) to the emperor Valens' disastrous battle of Adrianople against the Goths in 378 (only the exhaustive parts on the years from 353 onward have survived).

Christian authors did not enter into a competition with pagan ones on the traditional fields of political and military historiography but they developed models of historical writing that would become the most important innovations until Renaissance times. From the third century they compiled chronicles of universal history, which harmonized the data of biblical, Greek, and Roman history. The chronicle – in Greek – by Eusebius (*c.*260–339, bishop of Caesarea, adviser, and later biographer of Constantine the Great), its Latin translation and continuation (until 378) by Hieronymus and several successor works provided the model for a view of world history from the times of creation. Eusebius, again, created with his *Church History* a new historiographical genre not only with concern to the subject, the history of the Christian 'nation,' but also in his approach: dispensing with speeches, giving bibliographical surveys, and quoting verbatim from letters, official documents, and other sources. The awareness that their theological approach might make their accounts challengeable led Eusebius and subsequently inter alia Socrates and Sozomenus (both early fifth century) to prove thus the accuracy of their factual reports.

Concluding Remarks

Given the complexity and variety of ancient historiography it is almost impossible to summarize its impact on the subsequent Western tradition, especially since the Renaissance, in a few sentences. The most important legacy of Greek historians was the development of a critical attitude to the past, the invention of methods that made it possible to discern between facts and fiction. The shortcomings of their attempts in view of modern eyes should not obscure this unique contribution to the development of the Western mind. The adoption of historiography by the Romans provided the model for constructing the past in the shape of national history. The Christian writers invented ecclesiastical historiography as a subject of its own and a model for universal chronicles and histories starting from the creation. From the Renaissance, Thucydides and Tacitus were taken as the masters of political history, Polybius as the model military historian. With the growing interest in Oriental cultures Herodotus gained a new reputation as a cultural historian. Ancient historiography, which had covered republican city-states as well as world monarchies, was as a whole considered as teaching political and moral lessons of everlasting value (and thus was also of great importance for political theory). The tradition that historiography and antiquarian studies were two quite different subjects survived until the late eighteenth century. There were a vast number of commentaries on ancient texts, collections of inscriptions and coins, learned dissertations on the constitutional, legal, military, and religious institutions of the ancient world but nobody thought of replacing the ancient historiographical works by writing anew narrative histories of antiquity. Gibbon was one of the first to do this and to integrate the results of antiquarian research but he still avoided the period covered comprehensively by ancient historiography and thus started his own narrative only with the second century AD. Doubts about the reliability of ancient historiography, especially of the Roman annalistic tradition, developed in learned dissertations from the seventeenth century. However, it was only in the early nineteenth century, starting with Niebuhr, that a modern historiography of antiquity based on source criticism and systematic researches on the monumental remains of the ancient world began to be composed.

See also: Ancient Greece and Rome, History of; Christianity, Origins of; Historical Thought and Historiography: Christian Tradition; History: Overview; History: Representation Within and Functions of; History: Theories and Methods; Lithics and Archaeology.

Bibliography

Dorey, T.A. (Ed.), 1966. Latin Historians. Routlegde and Kegan Paul, London.
Fornara, C.W., 1983. The Nature of History in Greece and Rome. University of California Press, Berkeley, CA.
Hornblower, S. (Ed.), 1994. Greek Historiography. Clarendon Press, Oxford, UK.
Marincola, J. (Ed.), 2007. A Comparison to Greek and Roman Historiography. Blackwell, Oxford.
Meier, C., 1975. Geschichte (II. Antike). In: Geschichtliche Grundbegriffe, 2, pp. 595–610.
Meister, K., 1990. Die Griechische Geschichtsschreibung. Kohlhammer, Stuttgart, Germany.
Mellor, R., 1988. Roman historiography and biography. In: Grant, M., Kitzinger, R. (Eds.), Civilization of the Ancient Mediterranean, vol. III. Charles Scribner's, New York, pp. 1541–1562.
Momigliano, A.D., 1955. Contributi alla storia degli studi classici e del mondo antico, vol. 9. Edizioni di storia e letteratura, Rome.
Momigliano, A.D., 1978. Greek historiography. History & Theory 27, 1–28.
Momigliano, A.D., 1990. The Classical Foundations of Modern Historiography. University of California Press, Berkeley, CA.
Strasburger, H., 1977. Umblick im Trmmerfeld der griechischen Geschichtsschreibung. In: Historiographia Antiqua (Ed.), Commentationes Lovanienses in Honorem W. Peremans Septuagenarii Editae. Leuven University Press, Leuven, Belgium, pp. 3–52.
Usher, S., 1988. Greek historiography and biography. In: Grant, M., Kitzinger, R. (Eds.), Civilization of the Ancient Mediterranean, vol. III. Charles Scribner's, New York, pp. 1525–1540.
Winkelmann, F., 1991. Historiographie. Reallexikon für Antike und Christentum 15, 724–766.
Woodman, A.J., 1988. Rhetoric in Classical Historiography. Croom Helm, London.

Historical Thought and Historiography: Current Trends

Georg G Iggers, State University of New York, Buffalo, NY, USA
Supriya Mukherjee, University of Manitoba, Winnipeg, MB, Canada
Qingjia E Wang, Rowan University, Glassboro, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G.G. Iggers, volume 10, pp. 6771–6776, © 2001, Elsevier Ltd.

Abstract

This article is a survey of major trends of historical thought and writing since 1970 from a global perspective. The 1970s marked a turning away from analytical social science–oriented historical studies to interpretative cultural approaches. They also saw the inclusion of women, ethnic minorities, and formerly colonial peoples into historical accounts. The end of the Cold War marked a further reorientation in the 1990s and since. For the most part, the focus on culture and on previously neglected segments of the population continued but there was an increased turn away from Eurocentrism to global perspectives in historical writing.

Introduction: Dominance of Social Science Models of History around 1970

The years since 1970 and particularly since 1990 have seen important reorientations in historical thought and historical writing in the West and elsewhere. These changes are closely related to fundamental transformations in the conditions of modern life.

Briefly comparing the intellectual and historiographical situation in the year 1970 with that in 2013, one can distinguish certain dominant trends then and now. In neither case, however, is it possible to reduce historical thought to simple common denominators. Nevertheless in 1970, many historians, at least in the West and increasingly in non-Western countries, were convinced that history was a science, albeit a social science, engaged in recreating a real past as it had actually occurred. Since then, this confidence in the objectivity of science, including historical inquiry, has come under severe attack, as have the notions still dominant in the early 1970s of a coherent process of history. This view of history as a social science replaced an earlier orientation, sometimes labeled 'historism' or the Rankean paradigm, which had been dominant in much of the nineteenth century and well into the twentieth century, which saw history as one of the humanities, a *Geistes-* or *Kulturwissenschaft*, which required a logic of inquiry different from that of the hard sciences, namely, one that takes into account the elements of meaning and intention in human life that cannot be reduced to abstract generalizations. This conception went back to the German historian Leopold von Ranke (1795–1886), who laid the basis for history as a professional academic discipline that viewed history as a special sort of science in the sense that it founded the historical narrative on the critical examination of primary sources. Until that point, historiography in the West from the Ancient Greeks to the Enlightenment, while committed to the truth, basically took the form of rhetorical presentation. Now a sharp distinction was made between the scholarly work of professional historians, committed to the strict examination of primary sources, and 'amateur' historians who were not trained in these rigorous methods.

The new social science orientations went in several different directions. However, they shared certain assumptions. They criticized the older historiography for narrating a history of events, primarily political, focusing on leading personalities in positions of power, and called for a history that proceeded analytically, seeking to examine the social context in which events occurred. While the older historiography relied heavily on philology as a tool for the examination and understanding of texts to fathom human intentions, many professional historians now turned to the neighboring social sciences, sociology, economics, and psychology, and in France also to geography, which analyzed social structures and processes and stressed the interdisciplinary nature of historical studies. It strove for a history that was more comprehensive than that of the conventional narrative political historiography, now including broader segments of the population.

At this point in the mid-twentieth century, there existed a marked difference between the ways historical studies were conducted generally in the West and in Japan, to a lesser extent also in India and China, on the one hand, and in much of Latin America, the Moslem world, and Sub-Saharan Africa on the other. In the former, historical studies had been professionalized since the late nineteenth century and were largely concentrated in academic institutions, universities, and research centers; in the latter, they were generally, particularly in Latin America, the work of gentlemen scholars interested in local or regional history, or of political activists struggling against colonialism in India and Sub-Saharan Africa. In the period between 1945 and 1970, professionalization had set in the latter countries as well. In the Arab world and in Spanish America, universities had a long tradition, but it was only in the mid-twentieth century that research universities in the modern sense were established. In Brazil, the first modern style university was founded in São Paolo in the 1930s. In 1948, the establishment by the University of London of the School of Oriental and African Studies was followed beginning in the 1950s with the founding of research-oriented universities in Sub-Saharan Africa, most notably the University of Ibadan in Nigeria. Although historical studies in the developing countries generally followed an agenda focused on the search for

national identity from a postcolonial perspective, it increasingly took over from the West not only critical research methods but also shared in the general transformation of historical thought in the West.

Social science–oriented history in the period after 1945 ranged from the studies of the material and biological bases of historical life of the French *Annales* in the 1960s to the new historical demography in France and Great Britain, to quantitative urban history and electoral behavior in the United States, to the formulation of quantifiable models of economic cycles in the United States, Great Britain, and France, to a 'historical social science' in West Germany strongly influenced by Max Weber. Nevertheless, despite their differences, they all agreed that history had more kinship with the social sciences than with the humanities. This assumed, as the older historiography had also done, that there was a real historical past that constituted the object of historical studies. It sought to make historical method conform to the methodologies of other social sciences. Hence, it generally preferred different kinds of sources than did the older historiography, for example, hard economic and demographic data. This led Geoffrey Barraclough to comment in a survey written in the 1970s for United Nations Educational, Scientific and Cultural Organization on trends in recent historical studies that

> the search for quantity is beyond all doubt the most powerful of the new trends in history, the factor above all others which distinguishes historical attitudes in the 1970s from historical attitudes in the 1930s. (p. 39)

As we shall see, Barraclough's assertion about the centrality of empirical, quantitative history was, in many ways, already outdated when he made the observation in 1979, and it was not fully correct even then. Quantitative methods were not as dominant as he suggested. They played an important role in the United States and to an extent in France, but very little of one in Germany or yet in India or Latin America.

The Impact of the Annales and Marxism on the Writing of History

Two historiographical orientations that did not fit Barraclough's paradigm of social science history but that had a great influence internationally, including outside the West, should be mentioned here. Possibly no journal had as important an impact in France and internationally in the second half of the twentieth century as the *Annales*, although in French it later contained English abstracts to reach an international public. Peter Burke spoke of a 'French Historical Revolution' initiated by the Annales School. The revolutionary aspect consisted in the school's radical rejection of the manner in which historiography had been traditionally practiced at the Sorbonne with the focus on the narrative of primarily political events. Instead, the Annales sought to abolish the boundaries between the traditional disciplines in order to integrate them into the 'sciences of man' (*sciences de l'homme)* in which historians would work closely not only with sociologists and economists but also with 'scientists' in the fields of linguistics, semiotics, psychoanalysis, literature, and the arts. But it is not correct to speak of an 'Annales Paradigm' as Troian Stoianovich does in his book. The strength of the Annales lay in its openness, its commitment to exploring a large number of methodological approaches. The journal was founded in 1929 by Lucien Febvre and Marc Bloch as *Annales d'histoire économique et sociale*. It built on the ideas embodied in Henri Berr's *Revue de synthèse historique* (1900) which, as the title suggested, advocated a broad historical synthesis of the humanities and the social sciences. This was also the aim of Febvre and Bloch, but they went in very different directions. Bloch's work in *Feudal Society* and in his book on the character of French rural history sought impersonal structures, histories essentially without people, while Febvre in his book on Luther's and Rabelais' religious views focused on individuals. But in both cases, such as in Bloch's *Royal Touch* about the belief in medieval England and France that the king with his touch could heal scrofula, they were concerned less with ideas than with collective mentalities. Thus, Febvre in *The Problem of Unbelief in the Sixteenth Century: The Religion of Rabelais* concluded that Rabelais could not have been an atheist because the structure of the language that set the context for the collective mentalities of his age constrained him to be a believer. The journal, which had been suspended during the war, was reborn in 1946, this time with the subtitle *Economies, Sociétés, Civilisations* to emphasize its interdisciplinary character more strongly. The group now acquired a firm institutional basis with the creation of the *Maison des Sciences de l'Homme* in 1968 in which the key social science and humanistic disciplines were housed under one roof, and soon afterward the *Ecole des Hautes Etudes en Sciences Sociales*, which replaced the Sorbonne as the most important center of historical studies. Fernand Braudel in his book on the Mediterranean (1949, 1966) indicated the direction in which much Annales historiography went. In the place of one historical time, he saw a plurality of times, in this book the stationary time of the Mediterranean as a geographic space (*longue durée)*, the slow changes in social and economic structures (*conjonctures*), and finally the fast time of political events (*événements*), which interested him the least.

The significance of a nondogmatic Marxism on historical thought in Japan, India, Sub-Saharan Africa, and Latin America, and to a lesser extent in Western Europe, should not be underestimated. Marxism–Leninism represented a dogmatic form of Marxist orthodoxy, which dominated in the Soviet Union and in the countries dominated by it and in a somewhat different form in Maoist China, in both cases imposed by authoritarian regimes. By the time the Soviet Union collapsed, Marxism–Leninism had lost much of its credibility. Even in the Soviet Union, a historian such as Aaron Gurevich in *Categories of Medieval Culture* (1971) moved away from historical materialism to a cultural approach reminiscent of that of the Annales. The Polish historian Witold Kula in *Measures and Men* (1970) explored the symbolic meaning of weights throughout Western history. He and other Polish historians contributed regularly to the Annales.

In Western Europe, a thorough revision of classical Marxist theory had already taken place with the Hungarian Georg Lukács' *History and Class Consciousness* (1923), which was withdrawn from publication under pressure from the

Communist International but which continued to circulate in unauthorized editions. Lukács strove to free Marx's theory from economic determinism and stressed the cultural side of Marx's work. Basic ideas of Lukács' critique were taken up by Max Horkheimer, the founder of the Frankfurt School of Critical Theory, indebted to Marx, but freed of the speculative and authoritarian aspects of orthodox Marxist doctrine and extending the critique of capitalism to the realm of culture. Thus, the foundations were laid for what came to be known as Western Marxism. The immediate effect of Western Marxist ideas on historical writing in the West was limited. Marxism increasingly became an academic pursuit located in universities devoid of the firm support it had earlier in the industrial working class movement.

Marxism occupied a much more significant role in the economically less developed areas of the world, India, Latin America, Sub-Saharan Africa, and at the time China, in forms that were inspired by Western ideas but took on different forms that reflected local conditions. Here too, Marxist theorists were mostly located in academic institutions, but there was still a revolutionary tradition that built on Marxist ideas and for some had inspiration in the one successful revolution in Latin America, namely, in Cuba. Important ideas for revolutionary movements in these countries came from the revision of orthodox Marxism by the Italian communist Antonio Gramsci, whose writings in Mussolini's prisons in the 1920s and 1930s only became known in the 1960s. In them he explained the defeat of the communist workers movement in Italy in terms that involved factors insufficiently developed in traditional Marxist social theory. A critic of the orthodox Marxist reliance on an economic analysis of society, he maintained that the strength of capitalism rested ultimately on its cultural 'hegemony' by which control was maintained not solely by political and economic coercion over the working population but through a culture in which the values of the dominant classes were imposed and accepted by the masses of the population. According to him, revolutionary action required a counterculture challenging the dominant values of the capitalist classes. But Gramsci also criticized the orthodox Marxist focus on an industrial proletariat and expanded the idea of the working classes to include broad segments of the working population and the poor, to which he referred as the 'subaltern' classes, in an unevenly economically developed society as the Italy of his time. His analysis fitted much more closely the condition of unevenly developed countries, especially Latin America, than did orthodox Marxist doctrines.

Challenges to Social Science Approaches to History: Postmodernism

The widely held confidence in social science and modernization was radically challenged in the 1970s and 1980s, most vocally by postmodernist philosophers and literary critics, but more quietly also by historians, particularly cultural historians. While it is difficult to reduce these critics to a common denominator, they agreed in repudiating one assumption that they considered to be central to the modern historical outlook, namely, the belief in the coherence and continuity of world history, the invention of a 'master,' or 'metanarrative' (Lyotard) that sees history as an onward development to the modern Western world with its promise of enlightenment, emancipation, and well-being for the world as a whole. Practitioners of the established historiography, critics charged, had overlooked the close relationship between Western progress and imperialism abroad and the consolidation of power relations based on race, gender, and class at home, taken for granted in their view of history. But history, they argued, had no coherence, thus eliminating the possibility of a grand narrative and opening the way to innumerable separate histories.

Closely related to this critical view was the rejection of the possibility of objective historical inquiry into the past. According to Peter Novick, the idea and ideal of objectivity that lay at the very center of the profession ensured "the separation of history from fiction" (pp. 3–4). For Hayden White, however, history was just another 'linguistic mode' based on the aesthetic and moral preferences of the historian, and no more objective than fiction. Postmodernist philosophers further noted that the assumption that behind appearances lies a solid, foundational real world is erroneous. Thus, historical consciousness, according to Keith Jenkins, is "just one more foundationless, positioned expression in a world of foundationless, positioned expressions" (p. 6).

The Turn to Cultural History and the Return to Narrative

Such extreme epistemological relativism was largely restricted to literary theorists such as Hayden White, Roland Barthes, and Jacques Derrida. Few practicing historians, however, would agree that there are no criteria of objectivity. Historians continue to hold that historical works, in contrast to historical novels, must proceed from research that requires standards of objectivity shared by the scholarly community. Postmodernism had reflected a general mood that was shared by many historians. Part of this mood was a general skepticism regarding the nature and quality of modern civilization and its science, the recognition that the historical science of the West, whether in its historicist or its social science form, had legitimized relationships of power in which a middle and upper class elite had been privileged at the expense of the broad masses of the population, including women and the non-Western world. Thus, by the 1970s, the old scientific ethos of the social science historians was shaken. In l979, in a now famous article, Lawrence Stone noted that in the 1970s a change had occurred in the way history was viewed and written. The belief, central to social science history, that "a coherent scientific explanation of change in the past" (p. l9) was possible, was widely rejected. In its place there emerged an interest in the most varied aspects of human existence, accompanied by the conviction that "the culture of the group, and even the will of the individual, are at least as important causal agents of change as the impersonal forces of material output and demographic growth" (p. 9). Yet, although Stone emphatically rejected the illusion of 'coherent scientific explanation,' he nowhere suggested that historical narrative should surrender its claim to rational inquiry and realistic reconstruction.

The Impact of the Linguistic and the Cultural Turn on the Writing of History

The new directions that historical studies took must be understood within the context of the political climate of the 1960s, which saw the emergence of a women's movement, the rise of ethnic consciousness, the struggle for racial equality, anticolonialism, and environmentalism, which all pointed at the inadequacies of the social and economic order of an industrialized, modernizing Western world and the social science that served it. This marked a significant reorientation in historical writing, a new focus on the experiences of women and of marginal parts of the population, homosexuals, ethnic minorities, and colonial people, as well as of other nonconformist groups. This reorientation has to be understood within the context of the institutional changes in higher education, the rapid expansion of the universities, the changes in the student body with the entrance of segments of the population that had been underrepresented before, and the increasing presence of women in the profession. These changes were most pronounced in the United States but also occurred elsewhere. This deeply affected historical thought and writing. It also included a widespread critique of the established forms of social history, and incidentally also of traditional Marxism, which were criticized for concentrating on anonymous social processes, and producing a history without people. They sought a turn away from what they termed macrohistory to microhistories, which dealt with the life experiences of common human beings.

While few social historians could be identified as postmodernists or Marxists, certain central ideas of postmodernism and of cultural Marxism impacted a great deal of writing in social history, which increasingly moved in the direction of cultural history. The idea of the unity or direction of history gave way to the concentration on small units on the local level. The subjects of the new culturally oriented social history were most often "the lives and feelings and behavior of the poor and obscure rather than the great and powerful" (Stone, 1979: p. 19). Anthropology and semiotics replaced sociology and economics as the key auxiliary sciences of social history. But anthropology underwent a sea of change. While classical anthropology, e.g., that of Bronislaw Malinowski in the l920s and 1930s, had tended to a static analysis of 'primitive' and 'timeless' societies, anthropologists had now become increasingly historical in outlook, recognizing like Pierre Bourdieu that all, even the most 'primitive' cultures, evolve and that modern societies also are proper subjects of anthropological study. In Germany, this anthropological turn in the l980s took the form of *Alltagsgeschichte* (history of everyday life) (Hans Medick, Alf Lüdtke, and Lutz Niethammer), in Italy of the microstoria movement (Carlo Ginzburg and Giovanni Levi) around the journal *Quaderni Storici*.

With the turn to cultural history and cultural anthropology, the scope of history was substantially broadened. The turn to microstoria also carried with it a new emphasis on the history of emotions. The Annales had from the beginning of the school explored what they considered to be collective mentalities. They now examined changing attitudes on death, childhood, and sexuality, first with a quantitative collective approach, then with a focus on individual emotions and attitudes, as in Emmanuel Le Roy Ladurie's story of the heretic village Montaillou. The Annales launched a major study of book production and reading habits; microhistorians like Carlo Ginzburg and Hans Medick did this on an individual and local level.

The gap created between social and cultural histories began to close markedly in the 1980s and 1990s. In the 1980s, feminist historians, dissatisfied by the secondary status of women's history ('herstory') in a profession dominated by seemingly masculinist prescriptions of historical practice, turned to gender as a "useful category of historical analysis" (Scott). By emphasizing the culturally and socially 'constructed' nature of sexual identity, gender, a term borrowed from linguistics and literary criticism, revealed hierarchies and relations of power in societies mediated through language and discourse. Thus, postmodernist textual analyses of gendered meanings were used to challenge not only established conventions of sex, bringing to light their contingent nature, but also social institutions and even scholarly conventions of historical practice, which revealed a masculine bias. Natalie Davis is a prime example of this merger of social, cultural, and gender histories. Beginning with a focus on the experience of women in an early modern French social and cultural context, she expanded her scope to include the interplay of culture and religion between the European and the North African shores of the Mediterranean through the lens of a sixteenth-century Moslem. More recently, she turned to slave cultures with roots in Africa in eighteenth-century colonial Suriname.

The Impact of the Cultural Turn on Marxist Historiography

Gramsci's writings deeply influenced the Subaltern Studies group in India formed in the 1980s. Like Gramsci, whose term subaltern they had taken over, they called for a history from below which they saw the subaltern classes as potentially active revolutionary forces as in *The Elementary Aspects of Peasant Insurgency*, the classical monograph by Ranajit Guha, the founder of the group. The Studies laid foundations for a postcolonial historiography, which, however, distanced itself from the political leaders of the independence movement in India who came from a ruling elite, which in their opinion merely replaced the former British ruling elite. They called for a historiography in which categories of class, ethnicity, and gender would be integrated, but held that a Marxist conception of class conflict misses the analysis of power to which Foucault had pointed as an infinitely complex network of micropowers that permeate every aspect of social life and culture. At this point, they went further than E.P. Thompson, to whom they also felt indebted, in the direction of culturalism; the concept of subaltern replaced that of class, while subaltern agency replaced class consciousness, and because they were sympathetic to the postmodern critiques of power, subalternity became a condition embedded in and resulting from discourse.

Perhaps no book had as great an impact on the reorientation of Marxist social history, not only in the English-speaking world but also beyond, as E.P. Thompson's *The Making of the English Working Class* (1963). Thompson maintained the concept of class and the notion that "the class experience is largely determined by the productive relations

into which men are born – or enter involuntarily" (p. 9). But he stressed that class must be seen not as a 'structure,' or even a 'category' but as something that in fact happens in human relationships (p. 9). The English working class was thus not primarily the product of economic forces as it had been in the 'Communist Manifesto' or in structuralist, functionalist sociology, but an active agent, coming from a specific cultural tradition, which made itself as much as it was made. Hence his reliance on qualitative elements that made up a culture such as literature, art, folklore, religion, and symbolism. Moreover, he emphatically rejected the orthodox Marxist notion that sees the past as a step to the future.

Yet Thompson's historiographical approach soon came under attack from cultural Marxists who argued that he had not broken sufficiently with classical Marxist notions of class. Even if he saw class in terms of culture, they charged that his approach to culture still focused on an industrial working class in which those not directly linked to the industrial work process, including women, played little of a role. They began to ask "why did he (Marx) give so little attention to the experience of women and colonial peoples in favor of the experience of the European, white, male, working class?" (Wickham, 2007: p. 3), which became the concern of much of social history.

Reorientation after 1990: Toward a Global Academic Community

The main currents that have been discussed with the turn to culture and the opening of history to women and portions of the population that had been previously neglected continued to shape historical writing after 1990. Nevertheless, just as the political currents of the 1960s deeply influenced historical thought and writing worldwide, so the changed political constellation of the 1990s after the dissolution of the Soviet empire between 1989 and 1991 and the end of the Cold War presented historians with new challenges. The universal peace that the American political scientist Francis Fukuyama predicted in his essay, 'The End of History?,' in 1989, in which he foretold a worldwide, even if gradual, acceptance of American style free enterprise and democratic institutions after the collapse of Soviet communism, did not materialize in the way he had foreseen. Samuel Huntington who in *The Clash of Civilizations* had written of an irreconcilable conflict between Islamic and Western – and incidentally also Chinese culture – operated with very simplistic notions of a timeless Islam as a unified culture, overlooking the divisions in the Islamic world, its history, the effects of modernization, the role of economic factors, and finally the interrelation between Islamic societies and the modern West. Huntington's analysis would leave little space for the Arab Spring of 2011.

On one level, Fukuyama's prediction proved to be at least partly true, namely, as to the extension of Western style capitalism to large areas of the world as part of a process of globalization. This period was marked by what appeared to be the inevitable victory of corporate capitalism, not only in the West where it had begun well before 1989, but also in countries like China and somewhat later in Vietnam, which proclaimed themselves to be Communist. If the first decades after 1945 had been marked by a general prosperity in the West and in Japan accompanied by a higher living standard for the bulk of the working population, since the 1970s much of this was reversed with the stagnation of wages and dismantlement of basic parts of the social network, without any effective challenge from the general population. Marxism as an economic or political challenge was effectively dead, at least in the West. The financial crisis of 2008 pointed at the dysfunctions of capitalism without resulting in any substantial reforms; the Occupy Wall Street movement of 2011 contributed to an awareness of the growing social injustices created by corporate capitalism, but was supported by an ineffective small minority. Nor, as Fukuyama had wrongly foreseen, did the emergence of capitalist economies usher in a process of democratization. In China, for example, and also elsewhere, the introduction of capitalism was accompanied by a stringent dictatorship, and by a semi-dictatorship in Russia.

What did take place was a process of globalization that not only involved the worldwide transformation of the economy, intensified by the new information technology, but also was accompanied by the increasing homogenization in everyday life, in consumer behavior, patterns of urbanization, metropolitan architecture, popular culture in film and music, and the relations of the sexes and of the generations. Nevertheless, globalization took on heterogeneous forms on the social and cultural level, reflecting local circumstances and heritages. And these generated resistance against the effects of globalization on traditional ways of life, which, not seldom, involved violence.

More than in the period before 1990, but beginning already then, an international scholarly community came into existence, as non-Western scholars, foremost Indian, Middle Eastern, Latin American, and increasingly historians from Sub-Saharan Africa, occupied important academic positions at American, British, and Australian universities and to some extent at the *Ecole des Hautes Etudes en Sciences Sociales* in Paris. It is an important fact that the language of international communication was increasingly English. Scholars in the non-Western world had now increased input into these discussions, including those inspired by the Indian Subaltern Studies. In return, Western, particularly North American and British scholars cooperated closely, as to an extent they already had before 1990, with scholars in Latin America, Africa, India, and East Asia, so that studies in many fields have been increasingly internationalized. But there are still very definite limits to international exchanges. English language literature is widely read in the original or in translation. Many of the important works in history or related social science and humanistic disciplines are translated from English into non-Western languages, as are important French and German books and articles. But very few Chinese, Japanese, Korean, Farsi, Turkish, or Arabic writings have been translated into English. Thus, for example, in recent years 13 times more books were translated from English into Chinese than vice versa (Sachsenmaier, 2011: p. 42). The result is that international communication is still dominated by Anglo-America, with little awareness of the theoretical discussions outside the West, with the exception of Anglophone scholarship in India. And much history continues to be written in national languages for national audiences. To be sure, in recent years the international communication

among historians across the world has been facilitated considerably by the advance of computer technology and the wide use of the Internet.

Major Historiographical Trends after 1990

We can generally identify several major trends in historical writings across the world after 1990, which tend often to be interrelated: (1) the move away from a focus on national history to comparative transnational and also global approaches, (2) a shift away from a Eurocentric perspective on history to one that recognizes the significance of non-Western cultures and the interaction between the West and these cultures, (3) greater interdisciplinarity as part of comparative studies, and (4) an expansion of the scope of history to include women and sections of the population who previously had been marginalized. Beyond this, women and ethnic minorities increasingly entered the profession, particularly but not only in the United States, giving impetus to women's, gender, and ethnic studies. At the same time, the extreme cognitive relativism identified with the postmodern linguistic and cultural turns lost in significance (Spiegel) as its insufficiency in understanding the complexities of the modern, global world became clear. As seen, the forms of relativism that reduced history to a form of literary fiction had only a limited effect on how history was actually written and its advocates were found largely among academic literary scholars, not among historians.

However, while these are important trends in the ways history is written today, there is no new paradigm that defines historical studies. In fact, there is not only considerable diversity but there are also continuities with traditional ways of writing history. Although much history turned away from the concentration on the nation state, the nation state survives in historical writing, even if the definition of what constitutes a nation underwent changes. The turn toward transnational and global approaches was perhaps most developed in two countries with very different historiographical traditions, China and the United States. American historians had for a long time been interested in the world outside of America, but for a long time from a Western orientation – this meant Europe, particularly Western and Central Europe – now they increasingly dealt with the rest of the world, including former colonized areas such as Latin America. This involved a move away from the older conceptions of the cultural and even racial superiority of the West. Chinese historiography in the modern period had never been unaware of the histories of the Western world, as had been widely but incorrectly presupposed. In fact, world history had centered on the West, although now it also paid greater attention to the so-called developing countries (Sachsenmaier, 2011: pp. 171–231). It is striking that this tendency accompanied the economic modernization of China. Not withstanding the Marxist ideology professed by the state, much of historical writing in China since the 1990s was far removed from Marxist assumptions. At a time when conceptions of modernization came increasingly under scrutiny in the West, and both Western scholars like Shmuel Eisenstadt and Indian historians such as Dipesh Chakrabarty argued that there were multiple modernities, differing in the non-Western world from the European model, Chinese historians often tended to identify modernity with its Western form.

The Transformation of the Concept of the Nation

Notwithstanding the continuing concentration on the nation in much historiography throughout the world, the concept of the nation underwent changes. Reflecting the increasing number of ethnic minorities in post-1990 Western and Central Europe and in Anglo-America, as well as Australia and New Zealand, many historians no longer viewed nations as homogeneous, but as heterogeneous, multicultural societies. Similarly, it was now argued that the widely held view of Europe's coherence overlooked the various social and cultural strains transcending the borders that compose it. The same, of course, also applies to Latin America, Africa, and major Asian regions. Latin American historians became increasingly aware of the diversity between and within its regions and sought to reconstruct what they called the "peculiarities of Latin America," adapting a formulation from E.P. Thompson. There was an increasing tendency not only to write but also to teach national history from a multicultural perspective. In the United States, the *National Standards for United States History* and the *National Standards for World History* stressed the pluralism of the ethnic cultures that make up the United States and the role of minorities and women in history, an approach that permitted global intercultural comparisons. In India, the legacy of the Subaltern Studies meant the recognition of the need to go beyond unitary interpretative frameworks, such as that of the nation state. It was acknowledged that nationalist histories had evolved as anticolonial histories, yet from the perspective of the Subaltern Studies group much of this anticolonial historiography had merely replaced British elites with Indian elites ignoring the large subaltern sections of the population as active agents in the anticolonial struggle. This did not mean a decline in nation-oriented histories. Historians in non-Western regions have continued to delve into the distinct cultural traditions of a region's past in order to deepen their understanding of the interplay between tradition and modernity and past and present. Yet more recently, historians have sought to avoid dealing with national units altogether, to insist on relationships, transfers, and interactions independent of national lines in what French historians have recently called *histoire croisée* or 'entangled histories' (Werner, Zimmermann). Here, one may also mention Gerhard Haupt's and Jürgen Kocka's recent reader on comparative transnational approaches.

The Reorientation of Historical Writing and the Turn from National to Global History

It was in 1990 that as a consequence of the process of globalization the term 'global history' first became widely used. To begin with, it was interchangeable with a much older term, 'world history.' Attempts to write world history had a long tradition in Chinese, Islamic, and Western historiography. Ranke still undertook to write a world history that concentrated exclusively on the Western world, as did other nineteenth- and twentieth-century historians who dealt with the non-European

world in an age of imperialism and colonialism mainly in its relation to the West. The last major 'universal history' that sought to encompass the whole world was the English *Universal History* launched in 1736, thereafter largely replaced by Eurocentric or nation-oriented histories. This changed markedly after 1990 when the study of world or global history first received a firm professional basis with the foundation of the World History Association and the launching of its *Journal of World History* in 1990. In 1991, the bilingual English and German *Comparativ: Zeitschrift für Globalgeschichte* followed and in 2006, the *Journal of Global History*. Again, the United States and China, and to an extent Japan, became the key centers for the pursuit of world or global history, directed not only to research but also to instruction. In the United States, undergraduate courses in Western Civilization, which had been widely introduced shortly after World War I, now gave way to courses in world history. In China, institutes of world history were established. For the first time after 2006, a number of histories of historiography appeared, which encompass historical writing throughout the globe from its beginnings. Going beyond this, a small circle of American historians pursued what they called Big History (Christian, Spiers). While historians traditionally had considered the past prior to writing as prehistorical, they now wanted to go back to the very beginnings of the universe, the Big Bang, and trace it through the creation of the Solar System, the first appearance of human life and its evolution to the modern period. Such an approach had to break with the established methods of historical inquiry that focused on humanity. It wished to link history more closely to the natural sciences and give it a material basis, but inevitably involved large-scale speculation. A more solid attempt to link history to a natural setting was contained in William H. McNeill's early contribution in *Plagues and Peoples* (1976), which examined the impact of disease on human history in an environmental context.

In fact, since the 1970s, increasing attention has been given to environmental history reflecting the growing popular awareness of the dangers posed to the environment. First centered in the United States, environmental history was soon pursued throughout the world, particularly in India and Latin America. Scholars writing from an Indian perspective focused on the politics of environmental ideas, manifested in the colonial context allegedly as a movement to protect native wildlife over native people and in the postcolonial period a growing polarization between the environmentalism of the affluent and the environmentalism of the poor (Guha). In the United States, J. Reill in his environmental history of the world in the twentieth century stressed that traditional historiography has failed to understand the central role that the unparalleled human manipulation and neglect of the environment with its catastrophic consequences has played in the past century.

Marxist Historiography, the Annales, and the Historiography of Politics after 1990

A brief comment on the role of Marxist ideas in historical writing after 1990: The collapse of the Soviet Union and the communist regimes in its Eastern European subject states carried with it the definite demise of Marxism–Leninism as a historical doctrine. It became increasingly evident that the Marxist model of historical development had become increasingly irrelevant in the contemporary world. But, as seen above, Marxism–Leninism had long before lost its intellectual credibility and Marxist theory in Western Europe had moved away from Marxist orthodoxy to stressing the role of culture and recognizing that the development of a postindustrial capitalism had created social realities that no longer corresponded to Marx's conception of industrial economies with its clearly delineated bourgeois and proletarian classes. Marxism, moreover, had lost its foundation in a revolutionary proletariat, which no longer existed in the form Marx had conceived it. In the economically less evenly developed societies with large preindustrial agrarian populations, the classical Marxist analysis of classes fitted even less. After 1990, Marxism as a revolutionary movement was restricted to unevenly developed areas of Latin America and India. Elsewhere, but also in Latin America and India, the advocates of Marxist social theories were largely intellectuals generally found at the universities; Marxism became one philosophical orientation among others. But while Marxist ideas continued to occupy a role in the historiography of non-Western countries, it did so more in its Gramscian than in its orthodox form. In the West, historians who had long identified themselves as Marxists such as Perry Anderson, the editor of the *New Left Review*, and Eric Hobsbawm concluded that Marxism in the form in which they had once understood it was certainly dead, even if it still had an important critical function. The *History Workshop*, which in 1976 as a Marxist-inspired journal had carried the subtitle *Journal of Socialist Historians*, changed in 1982 to *Journal of Socialist and Feminist Historians*, in 1995 dropped the subtitle altogether with the explanation that "the political conditions in which we work have changed almost out of recognition … since we last amended our mast head."

This shift away from older positions occurred not only in Marxist but also in non-Marxist historiography, as in the case of the Annales School. The Annales had pioneered an openness to many areas of social existence that had long been ignored. Yet there was something provincial about a great deal of historiography carried on its name. Its turn away from political history went so far that it failed to deal with the decisive role that politics has played in its social and cultural contexts. Annales studies focused on the European, for the most part the French Middle Ages and the French Ancien Regime, largely giving little attention to political aspects. Bloch, Braudel, Pierre Chaunu, and Denys Lombard to be sure pursued a broader transnational perspective and Maurice Agulhon and Marc Ferro turned to the political world of the nineteenth and twentieth centuries, respectively. Only in the 1930s did the Annales confront problems facing the world at the time and extending to the non-Western world. In the 1990s, they recognized its limitations when they dropped the subtitle *Economies, Sociétés, Civilisations*, which they had used since immediately after World War II, which had consciously eliminated politics as a prime concern of history and downgraded the role of narratives. The journal's new subtitle *Histoire Sciences Sociales* reflected the openness to Western as well as non-Western societies and cultures that marked a great deal of contemporary historiography. Similarly, *Geschichte und Gesellschaft*, which in

the 1980s had given increased attention to cultural aspects of history, since the 1990s turned to transnational, global topics.

And the Annales gave important impulses to the role that memory occupied in the reconstruction of the historical past. Of course, memory, false or true, had always served the formation of collective identities. Already Maurice Halbwachs, a student of Emile Durkheim, had sought to establish the connection between a social group and its collective memory, which also played a significant role in the work of Lucien Febvre and Marc Bloch. But it was only since the 1980s, and particularly since the 1990s, that memory began to occupy a very central place in historical thought. A key work was the seven-volume *Lieux de Mémoire* (Sites of Memory), published between 1986 and 1993, which proceeded on the assumption that the history of the French nation could not be adequately represented by a narrative based on documentary evidence, but involved the images that the French had of their past. Thus, sacred sites, festivals, myths, songs, literature and art all played a role in fashioning national consciousness as it was remembered. But from a very different perspective, memory played a crucial role in the reconstruction of the horrors of the twentieth century, particularly the Holocaust, as they were reflected in the memories of victims. In Russia, the Memorial group sought to reconstruct what happened under Stalin on the basis of the memories of survivors.

This leads us to a consideration of the role of political history since the shift of historiography to cultural studies. Political history was by no means dead and a good deal of it continued to pursue traditional topics with traditional methods of relying on archival documents. But much historiography of politics was affected by the cultural trend. Labor, workers, slaves, natives, and women became the focus of a growing body of historical writing by academic historians. What occurred now was not the abrogation of political history but the cross-fertilization of political, social, and cultural history. Yet it is striking that the fifth volume of *The Oxford History of Historical Writing* (2011), the most comprehensive and up-to-date survey of historical writing since 1945, reflecting the disdain of many cultural historians for politics, does not include a chapter on the writing of political history and largely ignores the topic.

Conclusion

A very fundamental reorientation of historical studies accompanied by a reorientation of historical consciousness took place in the course of the 1960s, leading to an increased openness, reflecting the increasing attention paid to women and to other segments of the population that had previously been marginalized. As the central role of Europe, and more gradually also that of the United States, declined in the postcolonial age, history also assumed an increasingly global perspective. The collapse of the Soviet empire did not fundamentally change this in the period after 1990. Although these years saw the expansion of Western capitalism and a growing Western influence on social and cultural aspects, it also witnessed an increasing consciousness on the part of the non-Western world of their indigenous traditions resulting not in homogeneity on a global level but on diversity in an increasingly modernized world. This, as seen, was also reflected in approaches to history. There were two shifts in emphasis in historical writings, even if they did not constitute a fundamental reorientation. One was a growing global perspective accompanied by a decline, but not an abandonment of national history; another an awareness of the role of politics, which had been played down and too often totally ignored by historians committed to the cultural turn. Moreover, with the turn to transnational and global histories, which inevitably dealt with structures and processes, some historians tended to turn again to more analytical social science approaches that had been rejected by many cultural historians since the 1970s.

See also: Cultural History; Environmental History; Gender History; Global History: Universal and World; Historical Thought and Historiography: East Asia; Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; Historical Thought and Historiography: South Asia; Historical Thought and Historiography: Southeast Asia; Historical Thought and Historiography: Sub-Saharan Africa; History and the Social Sciences; History, Professionalization of; History: Overview; History: Representation Within and Functions of; History: Theories and Methods; Linguistic Turn and Discourse Analysis in History; Political History: History of Politics; Social History; Subaltern History.

Bibliography

Abromeit, J., 2011. Max Horkheimer and the Foundations of the Frankfurt School. Cambridge University Press, Cambridge.
Appleby, J., Hunt, L., Jacob, M., 1994. Telling the Truth about History. Norton, New York.
Barraclough, G., 1978. Main Trends in History. Holmes & Meier, New York.
Braudel, F., 1972–1974. The Mediterranean and the Mediterransean World in the Age of Philip II. Harper and Row, New York.
Burguière, A., 2006. The Annales School: An Intellectual History. Cornell University Press, Ithaca, New York.
Burke, P., 1990. The French Historical Revolution: The Annales School, 1929–1989. Polity Press, Cambridge.
Chakrabarty, D., 2000. Provincializing Europe: Postcolonial Thought and Historical Difference. Princeton University Press, Princeton.
Christian, D., 1993. Maps of Time: An Introduction to Big History. Longman, Essex, England.
Eley, G., 2005. A Crooked Line: From Cultural History to the History of Society. University of Michigan Press, Ann Arbor.
Evans, R., 1999. In Defence of History. W.W. Norton, New York.
Falola, T. (Ed.), 1993. African Historiography: Essays in Honor of Jacob Ade Ajayi. University of Amsterdam Press, Amsterdam.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Guha, R., 1983. Elementary Aspects of Peasant Insurgency in Colonial India. Oxford University Press, Delhi.
Guha, R., 2000. Environmentalism: A Global History. Longman, New York.
Haupt, H.G., Kocka, J. (Eds.), 2009. Comparative History and the Quest for Internationality: Central European Approaches and New Perspectives. Berghahn Books, New York.
Hobsbawm, E., 2011. How to Change the World: Reflections on Marx and Marxism. Yale University Press, New Haven.
Hunt, L. (Ed.), 1989. New Cultural History. University of California Press, Berkeley, CA.
Iggers, G.G., 2005. Historiography in the Twentieth Century: From Scientific Objectivity to the Postmodern Challenge. Wesleyan University Press, Hannover, New Hampshire.
Iggers, G.G., Mukherjee, S., Wang, E., 2008. A Global History of Modern Historiography. Longman/Pearson, Harlow, England.
Jenkins, K. (Ed.), 1997. The Postmodern History Reader. Routledge, London.

Kocka, J., Mitchell, A. (Eds.), 1993. Bourgeois Society in Nineteenth-Century Europe. University of Minnesota Press, Minneapolis.
Lyotard, J.F., 1984. The Postmodern Condition: A Report on Knowledge. University of Minnesota Press, Minneapolis.
Malerba, J., 2009. A história na América Latina: ensaio de critica historiográfica. Editora FGV, Rio de Janeiro.
McNeill, W.H., 1976. Plagues and Peoples. Anchor Press, Garden City, NY.
Moya, J.C. (Ed.), 2011. The Oxford Handbook of Latin American History. Oxford University Press, Oxford.
Novick, P., 1988. That Noble Dream: The Objectivity Question and the American Historical Profession. Harvard University Press, Cambridge, Massachusetts.
Robinson, C., 2003. Islamic Historiography. Cambridge University Press, Cambridge.
Sachsenmaier, D., 2011. Global Perspectives on Global History. Theories and Approaches in a Connected World. Cambridge University Press, Cambridge.
Schneider, A., Woolf, D. (Eds.), 2011. The Oxford History of Historical Writing, vol. 5. Oxford University Press, Oxford.
Scott, Joan Wallach, 1988. Gender and the Politics of History. Columbia University Press, New York.
Spiegel, G.M., 2005. Practicing History: New Directions Historical Writing after the Linguistic Turn. Routledge, London.
Spiers, F., 1996. The Structure of Big History: From the Big Bang until Today. University of California, Berkeley.
Stoianovich, T., 1976. French Historical Method: The Annales Paradigm. Cornell University Press, Ithaca, NY.
Stone, L., 1979. The revival of narrative: reflections on a new old history. Past and Present 85, 3–24.
Thompson, E.P., 1964. The Making of the English Working Class. Victor Gollancs Ltd, London.
(Special issue of Storia della Storiografia). In: Wang, Q.E. (Ed.), The Ebb and Flow of Marxist Historiography.
Wang, Q.E., Iggers, G.G. (Eds.), 2002. Turning Points in Historiography: A Cross-cultural Perspective. University of Rochester Press, Rochester.
Werner, W., Zimmermann, B., 2006. Beyond comparison; histoire croisée and the challenge of reflexivity. History and Theory 45, 30–50.
White, H., 1973. Metahistory: The Historical Imagination in Nineteenth-Century Europe. Johns Hopkins University Press, Baltimore.
White, H., 1987. The Content of the Form. Johns Hopkins University Press, Baltimore.
Wickham, C., 2007. Marxist History-Writing for the Twenty-First Century. Oxford University Press, Oxford.
Woolf, D., 2011. A Global History of History. Cambridge University Press, Cambridge.

Historical Thought and Historiography: East Asia

Masayuki Sato, University of Yamanashi, Kōfu, Yamanashi, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Historiography has been one of the primary cultural undertakings in East Asia. This is in contrast to the cultures of Europe, India, and Islam, where the concentration of cultural power has not been fixed in history. The purpose of writing history in East Asia was based on the Chinese philosophical premise that historical facts constituted the only certain and immutable reality, because people could not alter that which had already happened. This idea took history to be the axis of East Asian civilization. History set itself apart from political affairs, stood above administration, and deemed itself the foundation for all human judgment.

The art of recalling the past has developed very differently in Europe and East Asia. Since the time of Herodotus and Thucydides, Western historiography has centered on histories written by individuals for individuals, whereas East Asian historiography has since ancient times pivoted on public historiography. This distinctive characteristic of East Asian historiography could be described as 'publicly sponsored history', and constitutes a tradition that continues today as the *basso continuo* of culture.

'East Asia' as a historical and cultural concept, comprises present-day China, Japan, Korea, and Vietnam. In these areas, history developed around China, which overwhelmed the peripheral countries with its physical size, and political and cultural ascendancy. This historic East Asia has been termed 'the Chinese world order' or a system of international relations with China at the center. The system reached Korea and Vietnam in the early Han period (202 BC–AD 9), and Japan in the AD first century.

This sinocentric system also formed the common cultural heritage of East Asia: the classical Chinese language and Chinese writing system, the Confucian social and familial order, Buddhism through Chinese translations, the legal and administrative systems, and the art of historiography. The peripheral states of the East Asian world might also be called sinographic cultures, in the sense that they used classical Chinese as a common language. This is, in a sense, comparable with the use of Latin in medieval and early modern Europe (Fairbank et al., 1989).

Role and Purpose of Historiography

For the most part, the task of historiographical compilation in East Asia was traditionally a state-run project. The 'official history' produced by that compilation, along with materials collected for the purpose, constituted the core of East Asian historiography. Historiography may be said to be the primary cultural undertaking in East Asia. This is in contrast to the cultures of Europe, India, and Islam, where the concentration of cultural power has not been fixed in history.

For 2000 years Chinese historiography centered on the 'official history' compiled by each successive dynasty as a state enterprise. Later generations positioned the *Shiji* (*The Records of the Grand Historian*) by Sima Qian (145–*c*.86 BC) as the first official history. Since then 24 other official histories have been compiled. A characteristic of these official histories is that they have an encyclopedic tinge; that is, the body of the work originated by Sima Qian endowed the entire culture, its politics, economics, society, culture, technology, etc., with a unified structure. History was written as a means of comprehensively describing such a world system (Watson, 1958). As such, historiography in East Asia is perhaps equivalent to 'primary cultural undertakings' such as Biblical commentary and the *Corpus Iuris Iustanianus* in the West, the *Laws of Manu* in India, and the *Koran* in the Muslim world.

The purpose of writing history in East Asia was based on the Chinese philosophical premise that historical facts were the only certain and immutable reality. Chinese metaphysics was not premised on a revealed religion based on the existence of a unique, almighty God, like Judaism, Christianity, and Islam. Rather, it took the world as an ever-changing phenomenon, as presented in the *Yijing* (*The Book of Changes*). Therefore, it sought immutable reality in history, because human beings could not alter that which had already happened. This idea took history to be the axis of East Asian civilization.

This culture of history later spread throughout East Asia in tandem with Confucianism, creating a common historical culture throughout East Asia. The proclamation of this philosophy may be found in Confucius' statement that "All the empty words I want to write down are neither so clear nor so startling as seeing their meaning in action" (*Shiji*).

The Chinese had developed their own way of constructing an immutable past. In China and Korea, it was standard practice that, once the state's historical compilation bureau had completed compiling the official history of the previous dynasty, the bureau destroyed all the sources it had collected. This was to prevent the revision or rewriting of the official history, for once it was published by the government, the history itself took on the character of a sacred text. The most certain way to endow the official history with the imprimatur of authority was to destroy the sources on which it had been based. During the Korean Chosŏn dynasty (1392–1910), for example, the source materials were burned after use. When that was done, the account embodied in the official histories became the facts of history.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62029-9

In this fashion, East Asian cultures preserved the ideal that history was the sole immutable basis for human judgment. The biographies that comprise over half the material in the official histories maintained this tradition of objective narration in their own way. In the biographies, as in other sections of the official history, the texts first set forth what they believed to be 'fact'; this was followed by additions by historians who added their own evaluation. This vast corpus of biographies forces us to consider why the historians believed biography to be a necessary part of a history. The answer is that in a culture lacking a unitary supreme being, the records of the lives of eminent individuals become the only true sacred texts (Yoshikawa, 1963).

Historical Objectivity

The East Asian tradition was maintained by seeking objectivity in historical narration in the lives of the historians who wrote it even more than the narration itself. The following text presents an instance of a fifteenth-century Korean historian as a distinguished figure.

It is recorded that in 1431, as the compilation of the *Taejong sillok* (*the Veritable Records of King Taejong*) was nearing completion, his successor King Sejong (1418–50) asked the compilers to show him their work in advance. "In the previous dynasty, every monarch personally reviewed the veritable records of his predecessor; but King Taejong did not review the Veritable Records of King Taejo." Sejong's senior ministers replied that, "If Your Majesty were to review [the work in progress], later monarchs would surely revise [the historians' work]. Then, [future] historians would suspect that the monarch might look at the draft, and they would inevitably fail to record the facts completely. Then, how would we transmit [facts] faithfully for the future?"

'History' in modern East Asia is generally parsed with the two-character compound Chinese word *lishi* (Japanese, *rekishi*; Korean, *yoksa*), but until the middle of the nineteenth century, it was expressed in the single-character term *shi* (Japanese, *shi*; Korean, *sa*). This word *shi* itself originally denoted the historian himself, and only later, by analogy, did it come to refer to the product of the historian's craft, the written work of history itself. This is particularly important when we note that in the languages of Europe, it was precisely the reverse: the word 'historian' was formed to refer to the maker of the written 'history' (Sato, 2007). It is thus not surprising that historical consciousness in East Asia tends to focus far more on the attitude of the historian than on the actual works of history themselves.

Development of East Asian Historiography

China

In China, Sima Qian substantially established the tradition of Chinese historiography by inventing his own format for compilation. He succeeded his father as the grand historian of the Han emperor, and with his access to court archives he was able to complete the *Shiji* (*The Record of the Grand Historian*) in 130 volumes. It covers a history of 3000 years from the (legendary) Huangdi dynasty to his own period, and covers China and Central and South Asia. The format consists of five sections: 'Benji', or imperial annals (12 volumes); 'Biao', a chronology of the successive dynasties (10 volumes); 'Shu', treatises on economy, technology, astronomy, etc., (8 volumes); 'Shi-jia', annals of feudal states and nobles (30 volumes); and 'Liezhuan', biographies of famous persons and foreign peoples (70 volumes). Sima's format is encyclopedic in nature, systematizing the world in a single, unified structure; his historiographical format dominated East Asian historiography until the mid-nineteenth century.

Ban Gu (32–92 BC) followed Sima Qian's work, but limited himself to the history of a single dynasty, which he completed as the *Hanshu* (*History of the [former] Han dynasty*). His practice, in which each Chinese dynasty compiled the history of its predecessor, became standard for all succeeding official histories. During the Tang dynasty (618–906) this historical culture truly flourished, as history became an independent subject in the Chinese classification of learning (McMullen, 1988). Historians were closely linked with governmental bureaucracy and their team method of compilation started in this period and continued until the twentieth century.

A new format for compiling historiography was proposed by Sima Guang (1019–86) in his 294-volume *Zizhi tongjian* (*The Comprehensive Mirror for Aid in Government*) in 1084. This was an extensive historiography that employed the method of chronological order to help readers find certain historical events through dynasties. It covers 16 dynasties, for a total of 1362 years, and includes information from dynastic histories, chronicles, biographies, novels, etc. It is necessary to refer here also to the Hanxue (the Han School), the development of which established historical study based on textual criticism of historical documents in the Qing dynasty. Qian Daxin (1728–1804) was one of the great figures among them, whose *Nianershi kaoyi* (*Verification of the text of the 22 histories*) was an examination of characters and language, dates and places, genealogies, and institutions in the text of the 22 official histories.

Japan

The idea of official history spread from China to the other East Asian countries. These nations compiled their own official histories, modeled on the Chinese format and written in classical Chinese, not in their native languages.

The first Japanese state historiography was the *Nihon shoki* (*The Chronicle of Japan*, 720) compiled by Prince Toneri, and modeled after the Chinese format. It covers the period from the Age of Gods to the end of the reign of Empress Jitô (686–697), arranged in chronological order, in annalistic form. Toneri's historians used history as a forum to establish Japanese national identity and consciousness. This is in marked contrast to the *Kojiki* (*Record of Ancient Matters*, 712) commissioned by Ôno Yasumaro, which was based on a transcription of oral tradition (Sakamoto, 1991).

This tradition continued until the late nineteenth century. The *Honchô tsugan* (*Comprehensive Mirror of Japan*, 1644–70), compiled by Hayashi Razan and his son Hayashi Gahô,

followed the tradition of official histories sponsored by the Tokugawa shogunate and the *Dai Nihonshi* (*the History of Great Japan*, 1657–1906) compiled under the support of the Mito clan.

Apart from these traditions, Japanese historiographical writings included auxiliary sciences as bibliography, classics, textual criticism, and philology, which had developed in Japan to the same levels as in the West by the nineteenth century. Such undertakings in Japan originated when variant readings or conflicting readings of the *Chronicle of Japan* (720) were handed down intact to posterity. In this important respect, China differed markedly from Japan. In China, history was transmitted as an authentic, authoritative text, which admitted no variants. This difference is richly illustrated by the work of Hanawa Hokiichi (1746–1821). His *Gunsho ruijû* (*Classified collection of Japanese classics and documents*), begun in 1779 and still in progress, first appeared in 1819, with 1270 titles in 530 volumes; another 2103 titles in 1150 volumes appeared in 1822. The objective was not to establish the authoritative text and obliterate the others, but to make available all the variant texts, which later generations might judge. This collection compares favorably with such European endeavors as the *Monumenta Germaniae Historica*, the *Rerum Italicarum Scriptores*, etc. (Maruyama, 1974).

The Meiji government, established in 1868, which was responsible for the restoration of the emperor as the supreme sovereign and which led the subsequent modernization of Japan, continued the tradition of 'official histories' when it founded the Office for the Compilation of National History (Shushikyoku). At first, the aim was to create an official history of Japan, but for various reasons, the objective was amended to that of compiling historical documents. The original idea of compiling a national history was abandoned. The work of the Office continues today in its successor, the Historiographical Institute of Tokyo University. Moreover, virtually every prefecture, district, and locality in Japan, even down to the level of individual villages, continues the tradition of official history at the local level today, compiling local historical sources and publishing local histories. Similarly, many individual government bureaus, as well as corporations, universities, and other corporate entities, regularly commission their own histories, extending the model of 'official history' to the private realm.

Korea

Around the fourth century, Korea founded its historiographical tradition, although epoch-making historiography did not appear until the twelfth century. Kim Pu-sik (1075–1151) compiled the *Samguk sagi* (*History of the Three Kingdoms*) in 1145, the oldest extant history of Korea, covering times from antiquity to the decline of unified Silla in 935. It follows the format of Sima Quan's official history, as is shown by the name *sagi* (the Korean pronunciation of *Shiji*). Kim Pu-sik intended to provide moral lessons for posterity. Iryŏn (1206–89), a high-rank Korean Buddhist monk, compiled the *Samguk yusa* (*Memorabilia of the Three Kingdoms*) in 1281, which is the only extant history of Korea before the Chosŏn dynasty (1392–1910), together with the *Samguk sagi*. This work is unique in the sense that he wrote it privately, including myths, tales, and social issues, and even refers to the sources he used.

Sŏ Kŏ-jŏng (1420–88) compiled the *Tongguk t'onggam* (*Comprehensive Mirror of Korean History*) as a Korean official history in 1485, modeled on Sima Guang's *Zizhi tongjian*. It covers the period from ancient times to the end of the Koryŏ dynasty in 1392. This work was compiled under the influence of neo-Confucianism, and reflects the moralistic purpose of praising good and condemning evil. An Chŏng-bok (1712–91) likewise published his *Tongsa kangmok* (*Annotated Account of Korean History*), 20 volumes, in 1778, a chronological treatment of Korean history from Tan'gun, the legendary founder of Korea, to the decline of Koryŏ in 1392 (Baker, 2012).

Vietnam

In Vietnam, just as in Japan and Korea, some official histories were published including *Dai Viet Su Ky Toan Thu* (*Complete Book of the Records of Vietnam*) in 1479, which is based on the style of the *Zizhi tongjian* of Sima Guang (Wade, 2012).

Historiology in East Asia

The study of historical theory in East Asia began with Liu Zhiji's *Shitong* (*Comprehensive Historiography*), written in China in the eighth century (708). It was the first thorough treatise in Chinese – or for that matter in any language – on historical criticism. It was also a history of Chinese historians and a theory of historiography, and deals with such questions as selection and criticism of historical materials, historical objectivity, judgment in history, and historical causality. The work remained a constant reference for Chinese historians until the nineteenth century, and gave rise to the metahistorical tradition in Chinese historiography.

Liu's theory of history was further developed by Zheng Qiao (1104–1162) in his *Tongzhi* (*Comprehensive Monographs*, 200 volumes). In the introduction to the *Tongzhi*, he proposed his core idea of 'synthesis' and 'interrelatedness and causality'. By use of these concepts, he argued for the importance of the continuity of historical development. The voluminous *Tongzhi* also includes '20 Summaries', noteworthy encyclopedic chapters.

Zhang Xuecheng (1738–1801) wrote his *Wenshi tongyi* (*General Survey of Literature and History*), a work on the theory and history of historiography (only published in 1920). The first chapter starts with 'The Six Classics are all history'. Here he expressed his philosophy of history: that the very idea of the 'Way' or the 'Sacred' (*tao*) can only be known through men's actions (*shi*), which are recorded in the *Six Classics* (Nivison, 1966).

In comparison to the voluminous amounts of historical works written in China, very few works of historical criticism were produced. This is quite thought provoking, for it signifies that traditional East Asian historiography was normative in nature. Little discussion regarding the nature of history was needed. Moreover, relatively few works of metaphysical interpretation of the course of history were penned, the exception being Wang Fuzhi's (1619–92) posthumously

published discussion of historical logic and historical tendencies. East Asians valued detailed precision in history, rather than the sort of philosophical interpretations developed in Europe. Some exceptions, informed by Buddhism, were found in medieval Japan. Jien (1155–1225), a prominent Japanese Buddhist cleric, wrote his self-deprecatingly titled *Gukanshô* (*Notes of the Views of a Fool*, 1220), in which he proposed his philosophy of history with the term *dôri*, an idea of historical inevitability. Through this he explained historical processes in terms of the state of development and proposed his own periodization. Similarly, a century later, Kitabatake Chikafusa proposed in his *Jinnô Shôtôki* (*Chronicle of the True Descent of the Divine Sovereigns*, 1339) that Japan is a land founded by the gods and maintained by the blessings of an unbroken lineage of divinely descended emperors (Brownlee, 1991).

The Rise of Modern Historiography in East Asia

Western concepts of 'civilization and enlightenment' were introduced systematically to East Asia beginning in the mid-nineteenth century. They had a transformative effect on notions of history and historiography as well. The movement toward 'modern' and 'scientific' history in East Asia began with a variety of intellectual experiments among Japanese scholars in the 1870s and the 1880s. As Korean and Chinese intellectuals ventured to Japan toward the end of the nineteenth century, 'modern' Western notions and methods of historical inquiry and explanation gradually filtered across the entire East Asian region.

Western historiography was introduced through translations of European and American works into Japanese. Japanese intellectuals created their own neologisms to translate Western terms by combining two or three Chinese characters. Newly invented terms written with Chinese characters were later exported to China and Korea. In consequence, East Asian scholars now share most academic terms in the liberal arts and sciences.

The fact that such translations were feasible and successfully received, demonstrates the stage that Japanese historiography had achieved. By introducing Western historiography through translation, rather than keeping it as the preserve of a few language specialists, Western history and historiography were transformed into common intellectual property for all East Asian historians. This was now made accessible to others who worked on their own national histories.

The successful importation of Western historiography relied on the strength of extant traditional Japanese forms of historiography and auxiliary sciences. Native pioneers in the introduction of Western historiography, historians like Shigeno Yasutsugu (1827–1910) and Kume Kunitake (1839–1931), were initially trained in *Kôshôgaku* tradition (Chinese, *kaozhengxue*: textual criticism originating from the so-called Han School in the Qing period). They were thus well prepared to accept the empirical methods of Leopold von Ranke (1795–1886) and were particularly receptive to his notions of a rigorously critical approach to sources and the avoidance of value judgments. Ranke's ideas had been imported to Japan, when his student Ludwig Riess (1861–1928), who was invited to Tokyo Imperial University as a professor of history in 1887.

The acquisition and acceptance of European historical theory began with translations of leading contemporary European works on historical method and theory, such as those by J.G. Droysen, E. Bernheim and C.V. Langlois, and C. Seignobos. Under the strong influence of Bernheim, Tsuboi Kumezo (1859–1936) published his *Shigaku kenkyûhô* (*Methods of Historical Research*, 1903) in Japan, and Liang Qichao (1873–1929) published the *Zhongguo lishi yanjiufa* (*Methods of Chinese Historical Research*, 1922) in China. Until then to study history in traditional East Asia had meant becoming conversant with the official Chinese histories. The introduction of Western ideas of historiography opened new horizons for students of history. Historians now began to inquire into the 'historical facts'. They went beyond the historical narrative to examine the underlying facts themselves.

A typical example of this new consciousness is found in Shigeno Yasutsugu, nicknamed 'Doctor Obliteration'. Applying the European historical method to Japanese history, he reexamined the *Taiheiki* (*Chronicle of the Great Peace*, 1370–71), and proved that the fourteenth-century military commander Kojima Takanori never existed (Sato, 2008).

A similar consciousness was found in China in the case of Gu Jiegang (1893–1980), whose *Gushibian* (*Critiques of Ancient History*, 7 volumes, 1926–41) contributed greatly to the modernization of Chinese historiography. Gu took a finely honed critical scalpel to the world of ancient Chinese legend, which had long been accepted as fact. His school was nicknamed 'ancient doubters'. His method was rooted in the Qing dynasty (1644–1911) school of ancient doubters; however, it was at the same time backed by the new European historical method (Ku, 1931).

Similarly, in Korea, Yi Pyŏng-do (1896–1989) had been trained in modern historical methods. He stressed the necessity of critical and objective methods of treating historical sources in his *Han'guksa taegwan* (*Grand Survey of Korean History*, 1963).

With the introduction of Western culture to Japan, history became reinstitutionalized, beginning with the governmental section and the newly established Imperial University. However, it should not be forgotten that the introduction of Western historiography reduced East Asian historiography from its former position as a state enterprise to the status of a single academic field. Individual historians started to have their own 'specialty' in order to survive in the new university setting. In order to distinguish themselves from eminent amateur historians, their research was published as monographs in newly established historical journals.

In the West, the 'professionalization' of historical research resulting in the establishment of history as an independent university academic field was a great achievement for a discipline that had been neglected institutionally. By contrast, in East Asia, the introduction of modern Western historical research heralded the end of the prestigious tradition of a form of historiography that had aimed at a comprehensive description of the entire world. Since then, the subsequent evolution of East Asian historiography has developed largely along the lines of 'modern' historiography that prevail in other areas of the developed world.

Development of Modern Historiography

Since modern historiography was introduced to East Asia, its development varied according to the situation in each country.

In China, newly discovered archaeological results urged a new perspective of ancient history on Chinese historians. In particular, the discovery of oracle bone inscriptions laid the foundations for the new Chinese ancient history led by such historians as Luo Zhenyu (1866–1940) and Wang Guowei (1877–1927). These new works were later synthesized with Marxist materialism by Guo Moruo (1892–1978). His *Zhongguo gudai shehui yanjiu* (*Study of Ancient Chinese Society*, 1930) is the first Marxist work on ancient Chinese society. Thereafter, especially with the establishment of the People's Republic of China in 1949, Marxism became dominant among Chinese historians. Traditional historiographical achievements were mostly replaced by reverse interpretations, a trend that was highlighted in the period of the Cultural Revolution (1966–76). After the 1990s, the worldwide information revolution and the introduction of the latest historical methods of social history began to change Chinese historiography, opening it to the rest of the world (Ng and Wang, 2005).

In Japan, soon after the introduction of modern European historiography, the emperor-centered state promulgated a new political orthodoxy, and began to restrict intellectual freedom, especially in the study of ancient Japanese history. Since the 1930s, social and economic history began to apply the Marxist framework to Japanese history, mostly to the study of Japanese capitalism. After World War II, Marxism emerged as the dominant historiographical method, and it was only in the 1970s that Japanese historians began to accept the idea of social history imported from France, Germany, and England. Since the late 1980s, many Marxist historians quietly shifted from socioeconomic history to social history.

Metamorphoses of Public History

Currently, in East Asia, the aura of official history still colors a collective historiographical consciousness. The past is still seen as the most trustworthy 'mirror on mankind'. What is sought from official historiography is not 'a' history, but 'the' history, the reliability of which, people tend to believe, is guaranteed by public editorship.

In China, 24 official dynastic histories have been published and the tradition of public history continues to the present day. The *Qingshigao* (*Draft of the History of the Qing Dynasty*) was completed in 1927, and was finally published in Beijing in 1977. This applies to local history too; the publication of prefectural histories sponsored by local governments began in the Ming dynasty (1368–1644), and continues even in contemporary China.

In Japan, this idea is seen in school history textbooks and local histories. The Japanese Ministry of Education censors all school textbooks, thereby restoring the notion of 'official history' in the realm of history textbooks. Japanese schools teach students 'what' rather than 'why' and fill students' heads with information regarding a fixed course of Japanese history or world history without suggesting that different viewpoints exists. Out of 47 Japanese prefectures 43 have compiled their own prefectural histories and some have been compiled two or three over the past 120 years. Today, one finds hardly a city, town, or village anywhere in Japan that has not done likewise. These local histories are premised on the traditional historiographical notion that there is a historical truth that is embodied in authoritative narratives.

The tradition of official history is also alive in present-day South Korea. The 25-volume *History of the Republic of Korea*, whose publication was completed in 1979 by the National History Compilation Committee (established in 1948), is a state-produced history of the homeland.

Since the late nineteenth century, the practice of history has undergone professionalization and specialization on a worldwide scale, and East Asia too has been drawn into this whirlpool. Since then the style of historiographical practice in East Asia has, in most respects, resembled that of Western countries. Historians, as individuals, are expected to author numerous research papers and books. However, many historians are also involved in publicly sponsored historiography. Countless university-based Japanese historians, for example, serve on the editorial boards of local and prefectural history projects. We can thus observe the continuing coexistence of the two differing traditions of historical study. To put it differently, East Asian historians simultaneously inhabit two worlds of historiography: one a world of Rankean (or traditional East Asian) factual objectivity and another of post-Rankean interpretations of varied and opposing voices (Sato, 2007).

See also: Asia, Sociocultural Aspects: China; East Asian Studies: Politics; Japan: Sociocultural Aspects; Needham, Joseph (1900–95); Religions of East Asia; Time, Chronology, and Periodization in History.

Bibliography

Baker, Don, 2012. Writing history in pre-modern Korea. In: Rabasa, J., Sato, M., Tortarolo, E., Woolf, D. (Eds.), The Oxford History of Historical Writing Vol. 3: 1400–1800. Oxford University Press, London, pp. 103–118.

Brownlee, John S., 1991. Political Thought in Japanese Historical Writing: From Kojiki (712) to Tokushi Yoron (1712). Wilfrid Laurier University Press, Waterloo, ON.

Fairbank, John K., Reischauer, Edwin O., Craig, Albert M., 1989. East Asia: Tradition & Transformation, rev. ed. Wadsworth Publishing Company, Belmont.

Ku, Chieh-kang [Gu, Jiegang], 1931. The Autobiography of a Chinese Historian (A.W. Hummel, Trans.). E J Brill ltd, Leyden.

Maruyama, Masao, 1974. Studies in the Intellectual History in Tokugawa Japan. Tokyo University Press, Tokyo.

McMullen, David, 1988. State and Scholars in T'ang China. Cambridge University Press, Cambridge.

Ng, On-Cho, Wang, Q. Edward, 2005. Mirroring the Past: The Writing and Use of, History in Imperial China. University of Hawaii Press, Honolulu.

Nivison, David S., 1966. The Life and Thought of Chang Hsueh-ch'eng. Stanford University Press, Stanford.

Prusek, Jaroslav, 1963. History and epics in China and the West. Diogenes 42, 20–43.

Sakamoto, Taro, 1991. The Six National Histories of Japan (J.S. Brownlee, Trans.). University of British Columbia Press, Vancouver.

Sato, Masayuki, 2007. The archetype of history in the Confucian ecumene. History and Theory 46, 217–231.
Sato, Masayuki, 2008. The search for scholarly identity. In: Rüsen, Jörn (Ed.), Time and History. Berghahn Books, pp. 200–211.
Wade, Geoff, 2012. Southeast Asian historical writing. In: Rabasa, J., Sato, M., Tortarolo, E., Woolf, D. (Eds.), The Oxford History of Historical Writing Vol. 3: 1400–1800. Oxford University Press, Oxford, pp. 119–147.
Watson, Burton, 1958. Ssu-ma Ch'ien: Grand Historian of China. Columbia University Press, New York.
Yoshikawa, Kōjiro, 1963. Man and the concept of history in the East. Diogenes 11, 14–19.

Relevant Websites

http://www.princeton.edu/~classbib/ – Classical Historiography for Chinese History.
http://www.fordham.edu/halsall/eastasia/eastasiasbook.asp – East Asian History Sourcebook.
http://archive.org/details/chengyutung – University of Toronto Cheng Yu Tung East Asian Library.

Historical Thought and Historiography: Indigenous Cultures in the Americas

Berthold Riese, Sun Yatsen University, Canton, People's Republic of China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Historiography and historical thinking among indigenous people of the Americas have been mostly expressed in objects of art (sculpture, murals, and pictorial manuscripts) and in oral traditions. A concept of unified history – beginning with the origin of the world and eras of deities, continuing without break to the creation of humans and human history, and concluding with the ethnocentric tribal history continuously updated to the present – characterizes all groups. Inherent problems of interpretation of sources are the foreign languages and distortions that indigenous traditions have suffered in colonial times under massive European influence by conquerors, missionaries, and administrators. The best coverage of indigenous history and historical thinking is available for the Aztec (Mexico), Mixtec (southwestern Mexico), Highland Maya (Mexico and Guatemala), Lowland Maya (Mexico), Chimu (north coast of Peru) and Inka (highlands of Peru), and, to a lesser degree, North American Creek (Southeastern United States) and Plains Indians (midwestern Canada and United States).

Concepts of History and Their Formation

History is here conceived as information transferred by organized human groups and handed down from generation to generation. Since information can be contained in natural objects as well as in artifacts, a basic distinction needs to be drawn between historical information gained from either of these sources. The way people develop consciousness of their history is by taking information from these sources; storing it in their memory or otherwise; conceptualizing it by means of selection, rearrangement, and interpretation (cf. Markowitsch, 2002); and agreeing collectively on its meaning and its canonical forms. Since the trajectory from information gathering to its societal formation and use is a cultural process, each society develops and cultivates particular ways to do so. All cultures, however, make use of similar media for preservation and proliferation of history, especially oral performances (stories and myths) and figural visualization (monuments and artifacts). In addition to this, more advanced civilizations often use writing as the preferred means for disseminating and handing down historical concepts to future generations.

Mesoamerica

The Mesoamerican culture area existed from *c.* 2000 BC until the Spanish conquest in 1521. It was the home of important civilizations, among whom the Olmec, Epi-Olmec, Zapotec, Mixtec, Toltec, Maya (Highland and Lowland), and Aztec are particularly noteworthy. The only people in Mesoamerica (and in all of precolonial America) to have developed a fully functioning logo-syllabic writing system were the Epi-Olmec and the Lowland Maya (*c.* AD 100–1540). Rather incomplete writing systems (partial writing) were characteristic of their northern and western neighbors – that is, the Zapotec, the Mixtec, and the Aztec. Taken together, these cultures form the core of the cultural area Mesoamerica, which is characterized by similar, and in many respects uniform, conceptions of history. For the purposes of this article, they can therefore be treated as a unit.

Components of Mesoamerican Historiography

Persons (human and imagined), places, and actions are the major elements of Indian historiography in Mesoamerica. Additionally, natural phenomena (animals, plants, geomorphology, weather and astronomical phenomena, etc.) are noted. Judgments, expression of emotions, plans concerning the future, and afterthoughts, as reflections on history, are not expressed. The Mesoamerican *Weltbild* operates with three basic components. Cycles of time express repetitions that are believed to operate in history. Connections and explanations of meaning beyond the plainly visible historical facts are achieved by way of analogy (i.e., interpreting something that is identical or similar in form to be also identical in substance or related in a causal way) and by way of the causal-principle "post hoc ergo propter hoc." Another method is the reification of originally mental constructs and their material (written) expression. This is particularly significant when information is handed down over a long time, since history is largely conceived as the history of living entities and actions.

Analogy, reification, and cyclic rearrangements are the means by which complex historical constructs are created. Theogonical myths of creation and of successive world ages are, chronologically speaking, the first topic. Deities are the protagonists of these remote ages; tribes and their leaders, places of origin, and stations of tribal migrations enter the scene only later during semihistorical and fully historical times.

Reflections on history are not expressed explicitly, but can be deduced in part from historical writing: the human fate is overwhelmingly a world of suffering and danger, predetermined by deities and only to a very small degree influenced by individual behavior. Ancient cultures are studied by their successors not only by way of oral tradition and picture writing but also through artifactual, architectural, and artistic remains. These ancient cultures are conceived as more or less mythical according to their historical remoteness, and are often glorified; for example, the Toltec are considered the inventors of crafts and arts in the eyes of the Aztec and, to a lesser degree, the late postclassic Highland Maya (Kiche and Cakchiquel),

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62030-5

whereas the even remoter culture of Teotihuacan is believed to have been populated by giants and deities.

Historical Chronology

Mesoamerican cultures have developed methods of chronological dating (Riese, 1997). A cyclical and public calendar based on the solar year of 365 days provides the public year count and structures public events. In addition, a ritual 260-day almanac for private and personal soothsaying, which is likewise cyclical, is a second important chronological frame. Both combined and/or extended lead to supplementary large-scale dating devices, such as the Year Bundle (Azt. *xiuhmolpilli*) or Calendar Round and the Long Count. The basic chronological entity of all these variant calendars is the day. The Lowland Maya included in their historical writing the calculation of temporary celestial phenomena, especially lunar, Venus, and Mars cycles, and observed the seeming positions of planets in constellations of fixed stars, comparable to the Zodiac of Ancient Mediterranean and recent European cultures. These celestial phenomena are related to human affairs, and thus especially the sun and the planet Venus are considered factors in their history and forces that exert their influence in the near future (warfare).

By means of such chronological techniques, Mesoamerican people created histories of anthropomorphic and theriomorphic deities and conceived overarching world eras, manifest in the Aztec 'Legend of the Suns' (Riese, 2007) and the famous 'Calendar' or 'Sun Stone' in the Mexican National Museum. These demonstrate that they believed in the existence of five successive eras ('suns'), each with a different duration of a few hundred to several thousand years. Each era was quite distinct with regard to the creatures inhabiting it and their food, and each was finally destroyed in cataclysms by wild animals (jaguars), air (wind), fire, and water (rain). The reader may note the analogy to ancient Greek and Chinese theories of the elements. A new creation marked the beginning of each world age. Even the contemporary world age of the Aztecs around AD 1500 was considered to be of limited duration. It would be destroyed by earthquakes. Within this framework, humans appeared relatively late, at the beginning of the fifth era only. The Lowland Maya had a different and noncatastrophic notion of world eras with apparently smooth transition periods managed by councils of deities presided by God L. Occasionally, they connected themselves as human beings in a speculative way with the world of their deities anteceding the current human epoch. This holds especially true for the ruling dynasty of Palenque, which considered itself to be descended from the gods of the preceding world age. Every Mayan world age lasted exactly 1 872 000 days, which is approximately 5000 Gregorian years.

Factual Content

Origins of Humans

Mesoamerican tribes believed that most political groups of later times originated from a distinct place, which could be the tree of Apoala among the Mixtec, or the Seven Caves (Azt. *Chicomoztoc*), the island of Aztlan in the center of a lake, or a mythical city of Great Culhuacan (Azt. *Huei Culhuacan*) for the central Mexicans. These origins are often relatively recent. According to the Aztec main tradition, they emerged together with six or seven other tribes in AD 1064 from the Seven Caves. In some areas, the origin is placed in a more historical setting, such as the city of Tollan. This tradition was especially popular in Highland Guatemala among the Kiche and Cakchiquel. Tollan was the capital of the famous Toltec, who were actually a historical people admired for their skills, accomplishments, and political power and flourishing about AD 1000. Tollan is today a ruined city in the modern state of Hidalgo, Mexico. Migration myths, apart from being descriptions of historical events, which are to a certain extent factual – as far as can be concluded from language, history, and geography – seem to have functioned as a means of ex-post-facto political legitimization. A successfully migrating and thus surviving tribe legitimizes its right to rule over the traversed territories and the cities within, or at least to exert indirect political control in cases where the areas in question are too distant for factual dominance.

Polities and Tutelary Deities

The interrelation of the people with their pantheon becomes apparent in the system of tutelary deities of tribal groups and city-states: Quetzalcoatl for the Toltec (capital: Tollan) around AD 900–1100 and later also for the Chololtec (capital: Cholullan in the modern state of Puebla, Mexico) around AD 1000–1500, Tohil for the Kiche (capital: Utatlán) around AD 1000–1500, Huitzilopochtli for the Aztecs (capital: Tenochtitlan or Mexihco) around AD 1350–1500, and Camaxtli for the Tlaxcaltec (capital: Tlaxcallan in the modern state of Tlaxcala, Mexico) of the same period. In this manner, origin-related, transhistorical interrelations between humans and deities take form and are kept alive in a similar way as they are known for Old World civilizations (Sumer, Ancient Egypt, Babylon, and Jewish culture of the Old Testament) with deities intervening actively in the fate of their tribe. Tribes with their leaders, ruling dynasties arising from them during the process of becoming sedentary (Breuer, 1990), and their tutelary deities are the main actors in Mesoamerican history. Rulers are described with their most important *rites de passage* (birth, inauguration, marriage, procreation of offspring, and death), their martial triumphs, and – to a lesser extent – their civilizing deeds (introduction of cultivated plants or technical skills, monument erection, temple building, and water management), but not with any personal characteristics.

Geographical Horizon

After settling down, topics such as wars, marriage alliances, conspiracies, insidious martial acts, and self-sacrifice appear as historical episodes in tribal history, and the delimination of national territory is documented in special maps or cadastral collections. It is in these later times that major wars between Mesoamerican polities become a subject of history. In central Mexico, a confederacy of three Mexican cities – Tenochtitlan, Tlacopan, and Tetzcuhco – threw off the yoke of Tepanec rule of Azcapotzalco and gained complete independence, prestige, and property of lands (Riese, 2011). Later Aztec historians were well aware of the political and social importance of this war of liberation.

The Mayan area was never a unified political entity. It stayed separated into three regions, with little interaction and

territorially small city-states. In the southern Maya Lowlands, the city-states of Calakmul and Tikal fought a hegemonial war for the central Petén region for some 200 years (Martin and Grube, 2000) and during this time formed changing alliances with second-level cities. In northern Yucatan, repeated foreign intrusions by Mixtec and Toltec conquerors and later small-scale internal wars among a dozen petty states were decisive in the cultural memory, giving it a distinct fatalistic touch (Morley & Barrera Vásquez, 1949). The Guatemalan Highlands were the scene of frequent warfare between Mam, Kiche, and Cakchiquel (Carmack, 1973), events that are amply documented in their chronicles (Brinton, 1885; Schultze Jena, 1944; Recinos, 1957) and dramas (Rabinal, 1994).

Since the range of historical accounts rarely transgresses the polity and its immediate neighbors and never exceeds the limits of the Mesoamerican core area, 'world history' from a precolonial Indian perspective is very limited. History is primarily tribal or city-state history; regional history occasionally appears during selected short periods of single-power dominance and in a weaker form through ceremonial relations, especially the consecration of rulers in the holy city of Cholollan, and pilgrimage to important shrines (e.g., the Sacred Cenote in Chichén Itzá). History was never world history, not even pan-American history, neither in abstract thinking nor in political action. This compares, albeit on a narrower basis, with ancient and medieval concepts of the world in Middle Eastern and European cultures. No statements revealing knowledge or speculation about countries and peoples beyond this horizon can be found, even though archaeological investigations indicate external contacts in the form of trade relations with peoples far outside Mesoamerica, namely, with the North American Southwest, the Northeastern Woodlands (Mississippian cultures), Panama, and Ecuador (Riese, 2010). This seemingly odd and willful limitation of the historical horizon is characteristic of many early states and of most chiefdoms worldwide.

Social History

It is men who dominate the historical records as well as the actual political scene, and these are in general members of the nobility, since all Mesoamerican societies were ranked. The historical records, starting in about AD 300, give special emphasis to particularly successful individuals, such as the Toltec rulers Quetzalcoatl and Huemac, '8 Deer Tiger Claw' among the Mixtec, and the rulers 'Pakal' of Palenque, 'Rising Sun' of Copán, and 'Kakupakal' of Chichén Itzá among the Classic Lowland Maya. Many more are known from postclassic times, especially in central Mexico as witnessed in the Annals of Quauhtitlan (Lehmann, 1974), Highland Guatemala, and northern Yucatán.

In addition to politically active members of the ruling nobility, historical records of the Maya and the Aztec convey a few names of artists (painters, sculptors, poets, and performing artists), thus demonstrating to some extent interest in cultural and art history on a personal basis beyond the dynastic record.

Taking into account that history is predominantly the subject of males, the reader's attention must be drawn to the occasional significance of women among Mesoamerican people, especially in the Mayan city-states of Palenque and Yaxchilán around AD 600–700 and in early Aztec history beginning in 1400 with 'Ilancueitl,' mother of the first official Aztec ruler, Acamapichtli, and with several other prominent women. These women presumably never equaled men in rank, but gained prominence in their patrilineal societies as substitutes for a male heir, or as royal consorts upholding dynastic continuity. It remains unclear whether this role involved real implementation of power or whether it was confined to representational duties, although – in keeping with ethnological theory of gender studies – a rather representative and symbolic function is suggested. Women are also conceptualized as instruments of political provocation, especially when abducted as prisoners of war, raped or forced into marriage, and sometimes even put to sacrificial death. Respectful treatment of women is rarely reported, albeit in Aztec and Mixtec historiography a few cases are described.

Reflections on History

It is impossible to give a definite description of historical thinking among Mesoamerican Indians, since no explicit statements related to this can be found in the sources, and indirect allusions are vague. It is, however, manifest that humans are controlled and directed by deities to whom they are compelled to pay tribute in the form of prayers, offerings, and cult practice (there are 18 major feasts per year) in order to maintain and ensure their and the world's existence and well-being. This is most evident in the Aztec's dogmatic obligation of invigorating the sun on its daily course over the firmament with sacrificed human hearts. This compulsion to human sacrifice led to sacrificial excesses later witnessed by the Spanish conquerors. Prior to the conquest, several Indian philosophers had already criticized the practice of human sacrifice as incorrect by expressing their desire for the idealized Toltec times (around AD 1000) to return, when people sacrificed flowers and animals instead of human beings. Human sacrifice included self-sacrifice, especially drawing blood from various body parts (tongue, penis, and thigh).

The general Mesoamerican outlook on human life is pessimistic and fatalistic. This manifests itself especially in the divinatory rituals of the Maya, Zapotec, and Aztec, where indifferent or positive prophecies are outnumbered by negative ones. The collective fatalistic behavior of the colonial Yucatec Maya and – two centuries later – of the Itzá, a late enclave in colonial New Spain that succumbed only in 1697 to Spanish dominance, concerning their own political fate might be an extreme example of this attitude toward the future, gained from analysis of the past. However, it should be taken into account that the Yucatec case could well be the result of traumatic experiences during the Spanish conquest, since only colonial sources give information on this topic. Thus, it is difficult to evaluate how strong fatalism was traditionally. The same holds good for the often-cited pessimistic and seemingly inactive attitude of the Aztec ruler Moteuczuma confronting the Spanish incursion into his territory in 1519.

Human life is considered to be hard and fateful (like in old Jewish faith). Consolatory compensation is offered by the Aztec concept of the otherworld, where sweet paradise awaits those who endure dangerous labor and hardship on earth, favoring warriors dying in battle (cp. Islamic belief) and

childbearing women. Yet we lack a theologically differentiated comprehension of the Aztec idea of an afterlife, since this is not further described in the historical accounts (Eschmann, 1976). The Lowland Maya left more written and pictorial documents regarding this topic, especially in their rich funeral ceramic paintings (Riese, 2004).

Reflections on antecedent cultures and polities can rarely be found among the descriptive historical traditions of the central Mexican Aztecs, other than general admiration for cultural accomplishments of the Toltec. The major city-ruin of Teotihuacan, which had been destroyed and abandoned 700 years prior to the arrival of the Aztecs in the valley of Mexico, is taken by myth as the place of origin of the gods. Even earlier cultures are remembered indirectly by conserving their artifacts and imitating their products, as has been amply shown in the excavations of the Major Aztec Temple of Tenochtitlan (nowadays the 'Centro Histórico' of Mexico City).

Sources

Numerous monuments executed in lasting materials (stone) are an important source for the study of historical thinking of Mesoamerican people. They date back as far as 1000 BC, beginning with sculptures portraying Olmec rulers, and stone slabs illustrating Zapotec conquests (so-called *danzantes*). The first real inscriptions date from the first two centuries AD, called the Epi-Olmec period. Especially noteworthy is the stela from La Mojarra, although its long inscription presumably representing the Proto-Zoquean language of the Isthmus region has not yet been deciphered, contrary to much publicized claims. It is presumably historiographic in content and presents a royal portrait with a long dated text. Similarly, the Zapotec produced narrative sculptures, albeit with rather short accompanying texts (Marcus, 1992).

Murals from Teotihuacan (Miller, 1973) and ceramic paintings of the Classic Lowland Maya (Kerr, 1989–2000, and Internet database) with their pictorial and partially narrative depictions and hieroglyphic explanations are major sources for reconstructing the Mesoamerican Classic from AD 300–800, since several thousands of them have been secured (mostly illegally by grave robbers) and are accessible to scholarship. These paintings show scenes from royal palaces, royal visits, banquets, payment of tribute, hunting of deer and birds, ball games, market scenes, processions, and postwar ceremonies. Most illustrations on Maya vases probably allude to the otherworld. Since daily life can be assumed to have served as the calcque for the concept of this otherworld, it is possible to reverse this relation and deduce everyday life of the Maya from ceramic paintings. In many instances, concrete persons are named as protagonists of the scenes or recipients of the vessel, and sometimes the artists of the painting are also named.

Further sources of historical records are pictorial and written manuscripts with historical contents, which – due to their fragility – generally do not date back too far. Only about a dozen precolonial manuscripts exist, four from the Maya Lowlands and the rest Mixtec in origin. Perhaps another hundred copies and extracts compiled during early colonial times have also been preserved. Most of those are of central Mexican and Mixtec origin, with very few from the Tarascans and Zapotec regions. Yucatecan (Barrera Vásquez and Rendón, 1948) and Highland Maya (Schultze Jena, 1944; Recinos, 1957) sources written in Indian languages mostly lack pictorial elements.

Christian missionaries introduced writing with the Latin alphabet in early colonial times (AD 1520–1650). Written documents of this type are especially numerous in central Mexico. Basically European in form, and to some extent even in content, the object of these written sources is almost completely related to the colonial situation (lawsuits dealing with inheritances, properties, and privileges, and history of the Spanish conquest). Many colonial writings obviously connect pre-Spanish traditions of partial writing and chronology with oral traditions and form a new integrated genre. In the Guatemalan highlands ('Popol Vuh' and 'Annals of the Cakchiquels,' Brinton, 1885), as well as with the Yucatec Maya ('Books of Chilam Balam,' which means books of a prophet named 'Jaguar'). These sources not only are less numerous than those of European confection but also show less European influence in content. Topicalizing other Mesoamerican regions (Tarascans, Zapotec, and Nicarao), only a few historical documents have been preserved and/or included in Spanish historiographic treatises. They reflect study of now-lost Indian documents and interviews with knowledgeable Indian informants. Thus, Gonzalo Fernández de Oviedo y Valdés (1959) reports on the Nicarao and their tribal history. The Nicarao lived at the southern periphery of Mesoamerica and are name givers to the modern state of Nicaragua. The Franciscan missionary Juan de Torquemada (1975–1983) conveyed comparable traditions of the Pipil living in modern Guatemala and El Salvador. The accounts of the Mixtecs by Francisco de Burgoa (1674) are comprehensive but not very detailed, and the historian Francisco Antonio de Fuentes y Guzmán (1972) includes in his writings summaries of many now-lost Indian documents from Highland Guatemala. A major European element in sources, which have been written in the indigenous languages but with Latin letters, is the effort to produce some sort of synthesis, often named 'world history' and 'natural history.' This effort can be noted, for instance, in a general chronology drawn from uncoordinated lists recording royal successions of individual Indian city-states as presented by the anonymous *Annals of Quauhtitlan* (Lehmann, 1974) and in the writings of Domingo Chimalpahin (1963–1965). As suggested by its subdivision into 12 books, the topics, and the illustrations, the *Historia General* of the Franciscan friar Bernardino de Sahagún (1950–1981) obviously resembles medieval European natural history and its antique origins (Pliny the Elder), in this case containing authentic information of pre-European Aztec culture provided by his Indian informants.

Other Regions of the Americas

On their first contacts with indigenous peoples of the Americas, European conquerors came across various forms of 'picture writing' (thus the then current nonscholarly terminology) that was used to write down historical events. Besides Mesoamerica, treated in the foregoing paragraphs, this was observed in several areas of North America, predominantly in the eastern woodlands (Voegelin, 1954) and on the plains and prairies west of the Mississippi (Mooney, 2011). They were in use much

later than in Latin America, since large parts of North America were conquered and settled by Europeans only during the eighteenth and nineteenth centuries. South America was in this respect quite the reverse, even though pictorial representations in painting (Moche and Nasca), weaving (Paracas and Nasca), rock engraving (Nasca), and string notations (Quechua: *khipu*) were developed early in its cultural history and highly elaborated. No such records have been reported from the vast lowlands of South America (Venezuela, Brazil, eastern Bolivia, etc.).

North America

Conceptually and structurally similar to Mesoamerican historical tradition are the eighteenth- and nineteenth-century historical traditions of the Lenape (Algonkin language family) and of the Indians of the Great Plains in North America (Kiowa, Blackfoot, and Teton), both recorded in non-book-like fashion since books did not exist in Native North America prior to the advent of Europeans. The Lenape (formerly called Delaware) and other related groups of the East (e.g., the Creek) (Gatschet, 1884) cultivated an oral and possibly pictorially written tradition of the tribe's migrations led by certain chiefs, while the Plains Indians, being acephalic societies, depicted outstanding events from the group's daily life following the year's seasons. These are called 'Winter Counts' (Handbook, 2001) and conspicuously stress intertribal warfare, but not, however, the then already more decisive wars against Anglo-American intruders.

The South American Andes

The study of historical concepts in the Central Andes, which were lastly dominated by the Inka, is rendered more difficult due to a lack of precolonial written records. All records of Indian history from this culture area originate in early colonial times, and are based on oral traditions and the system of knotted cords (*khipu*) only. Thus, factual history for the Chimu kingdom on the north coast, and the Inka empire in the highlands around the capital city of Cuzco, does not go back beyond AD 1400, when, retrospectively, mythical legend begins.

Andean history was put into writing almost exclusively by Spaniards (de Betanzos, 1987) or mestizos (Garcilaso de la Vega, 1609) and thus makes use of categories and forms of a European conception of history. This holds true also for writings that are sometimes considered as 'ancient Indian' such as the preciously illustrated '*Nueva Corónica y Buen Gobierno*' of Guaman Poma de Ayala (1936), which is European in its format and illustrations, even though the author is Indian, writes a mixture of Spanish and Quechua, and represents an Indian perspective in his writing. The anonymous historical drama 'Ollantay' is, according to its form and subject, rather more similar to a baroque theater play from Europe than an expression for Indian thinking and acting, albeit written in a stylistically pure form of Quechua, the language of the Inka empire, which makes it appear very Indian in conjunction with its theme, a heartbreaking episode of the late Inka empire (Middendorf, 1890).

The general assertion that something like an 'Andean historical concept' existed and is reflected in the sources, expressed by prestigious scholars, is not a helpful concept, since it remains vague and originates in projections by western scholars and their historical thinking based mainly on sources written by European authors.

Recent Achievements and Problems Ahead

Brain research and historical case studies of the last decades have brought to light the extreme extent of human molding, transforming, and distorting of memorized information. This has led to the consequence of neoskepticism toward so-called historical truth and the equation of historical memories with factual events as conveyed in oral history and written records based on these (Markowitsch, 2002). These new insights should be kept in mind, and all reconstructions of history gained by using humanly produced sources are to be reevaluated.

Without explanations from the creators themselves, it also remains difficult to understand the content, intent, and symbolism of the great range of historical objects and sources, even though researchers have presented seemingly comprehensive and highly sophisticated interpretations, in particular Seler (1960–1967), Eschmann (1976), and Jansen and Pérez Jiménez (2007), to mention only the well-conceived efforts.

The deciphering of the Epi-Olmec script is a pressing research desideratum since it would give us direct access to historical writing antedating hitherto useable records by 200 years. Hieroglyphic captions of Mayan ceramic paintings are well understood but not yet fully recontextualized to their place of origin, due to the fact that most of them originate in clandestine excavations. Most other Mayan inscriptions associated with distinct persons that designate singular objects like 'earring' or '(sack full of) beans' are short; also, they neither contribute to a differentiated understanding of the depicted scene, since they only name the object, nor do they help much in a broader comprehension of local history, even when they mention the owner or receiver of said object. Analogy from early colonial or even ethnographic sources of the present to foster our understanding of ancient Mayan sources is fraught with pitfalls, but nevertheless it is advanced from time to time by scholars who seem to wish to transcend our current knowledge using dubious research methods.

Among non-Mayan and non-Aztec Indian documents, most progress has been made in understanding Mixtec pictorial historiography, thanks to the lifelong research of two scholars, the Mexican Alfonso Caso (1977–1979) and the Dutch Maarten Jansen.

For Andean history, the most valuable recent research has been the use of colonial administrative sources as well as many contributions of field archaeology using 'sophisticated analytical techniques enabling us to gain a more precise environmental and chronological framework for historical projections.

Texts from early colonial times (until about 1650) written in the indigenous languages by means of the Latin alphabet (Crónica mexicayotl, 2004) are easier to approach, although they are still not very well understood either. Most of them originated with the Aztecs, some of them with the Yucatecan and Highland Maya, while there are only very few from other cultures and regions. Whereas the more technical problems of

translation can be met by increasing language competence, the main problem of ethnohistory (the technical term for the discipline of history that analyzes sources in the writing systems and languages of the people studied, Ethnohistory, 1972–1975) is to grasp concepts, metaphors, and so on to understand the *Weltbild* of the Indians, to separate it from European projections and intrusions, and finally to evaluate their emic intent and assess their factual content.

Devoting further efforts to research of chronological problems is most likely rewarding, especially for Mesoamerica. Mayan, Aztec, and Mixtec time counts are not only of high accuracy but also omnipresent in historical records (Prem, 2008a,b), and they are partly interrelated. For the time of the Maya Classic from AD 300 to 900, there is not a single year without several exactly datable historical accounts. Central and western Mexican historiography covers a time span from approximately AD 600 until the Spanish conquest and is structured with about the same amount of dates. However, due to their calendrical system, central Mexican dates are chronologically ambiguous, since they recur every 52 years, while Mayan dates are still floating in (western) time, for lack of a concluding correlation of western (Gregorian or Julian) chronology and the Mayan Long Count. The interrelations of dates in the different historical calendar systems (of Mixtec, Aztec, and Lowland and Highland Maya), the comparison of dates with chronological reconstructions in archaeology, and absolute dating by way of historical astronomy will some future day render it possible to compile a historical chronological chart for the Mesoamerican core area covering the period from AD 100 to 1521.

For North America and the Andean Culture Area, chronology is purely archaeological, and thus less precise and not directly correlated with historical events. But advances in chemical and physical dating techniques of archaeological artifacts will help greatly under these conditions, to gain more precise insights into the chronological development of cultures.

Because of Mesoamerica's high civilizing and political complexity as well as its dynamics, the analysis of overarching political and cultural systems and their changes in time is a very important field of research. Points in time of such systems are distinct and often short periods during the formative Olmec epoch, the middle classic Teotihuacan horizon, the Toltec dominance and expansion during early postclassic times, and the Aztec empire in late postclassic times, respectively. A great amount of effort has already gone into acquiring knowledge on the postclassic times in central Mexico, Oaxaca, northern Yucatan, and the Lowland Maya Classic. Yet the abundant archaeological findings of this period, including monumental sculpture and buildings, can be considered only the backdrop for political and economic history, worldview, and symbolic language.

General anthropological theories have made use of the respective Indian cultures, mostly Aztec and Inka, since Bandelier (1880) classified the Aztecs within the scope of an evolutionary scheme developed by Lewis Henry Morgan. It was followed over 100 years later by Breuer's (1990) more sophisticated analysis and has been amplified toward cultural aspects by art historical and architectural studies (Riese, 2011). American cultures are particularly relevant for this kind of research because it is assumed that they developed independently of Old World civilizations, and thus serve as models for testing general anthropological theories.

In spite of the importance of indigenous American historical documents, artifactual and written, we should be cautious to accept newfound documents and artifacts at first glance as authentic, since forging has been a constant and prolific industry for over 150 years, and too many scholars are prone to accept such forgeries as genuine documents, since they seem to offer useful information and often fit well into the expectations of current research. The most disturbing forgeries in recent decades were a series of written 'histories' on the Yucatan Maya, probably produced in Mexico, and a vast number of 'colonial documents' purportedly containing detailed information of the late Inka culture, produced in southern Italy. A more ancient problem in this respect is, of course, the authenticity of the North American Walam Olum (Voegelin, 1954) attributed to the Lenape Indians.

See also: Americas, Sociocultural Overviews: Central America; Historical Thought and Historiography: Current Trends; Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; Imperialism, History of; Time, Chronology, and Periodization in History.

Bibliography

Bandelier, Adolph Francis Alphonse, 1880. On the social organization and mode of government of the ancient Mexicans. In: Annual Report of the Peabody Museum, vol. 12. Harvard University, Cambridge, Mass.: Peabody Museum, pp. 557–699.
Barrera Vásquez, Alfredo, Rendón, Silvia, 1948. El Libro de los Libros de Chilam Balam. Fondo de Cultura Económica, México.
Betanzos, Juan de., 1987. Suma y Narración de los Incas. Madrid.
Breuer, Stefan, 1990. Der archaische Staat: Zur Soziologie Charismatischer Herrschaft. Dietrich Reimer Verlag, Berlin.
Burgoa, Francisco de, 1674. Geográfica Descripción de la Parte Septentrional del Polo Artico de la América. Juan Ruiz, México.
Brinton Daniel Garrison (Ed.) 1885. The Annals of the Cakchiquels. Brinton's Library of Aboriginal American Literature, 6. Philadelphia.
Carmack, Robert M., 1973. Quichean Civilization: The Ethnohistoric, Ethnographic, and Archaeological Sources. The University of California Press, Berkeley & Los Angeles.
Caso, Alfonso, 1977–1979. Reyes y Reinos de la Mixteca, vol. 2. Fondo de Cultura Económica, México.
Chimalpahin, Domingo, 1963–1965. Relationen, vol. 2. Cram de Gruyter & Co, Hamburg.
Crónica mexicayotl, 2004. In: Berthold, Riese (Ed.), Collectanea Instituti Anthropos, vol. 44. Academia Verlag, Sankt Augustin.
Eschmann, Anncharlott, 1976. Das Religiöse Geschichtsbild der Azteken (Indiana, Beiheft 4). Gebrüder Mann Verlag, Berlin.
Ethnohistory. Handbook of Middle American Indians, vols. 12–15, 1972–1975. University of Texas Press, Austin.
Fernández de Oviedo y Valdés, Gonzalo, 1959. Historia General y Natural de las Índias, vol. 5. Ediciones Atlas, Madrid.
Fuentes y Guzmán, Francisco Antonio de, 1972. Obras Históricas, vol. 3. Ediciones Atlas, Madrid.
Garcilaso de la Vega, el Ynga., 1609. Primera Parte de los Comentarios Reales. Pedro Crasbeeck, Lisboa.
Gatschet, Albert Samuel. (1884). A Migration Legend of the Creek Indians. Philadelphia
Handbook of the North American Indians, Vol. Ed. Raymond J. DeMallie, Vol.13: Plains, 2001, Smithsonian Institution, Washington, DC.
Jansen, Maarten, Pérez Jiménez, Gabina Aurora, 2007. Encounter with the Plumed Serpent: Drama and Power in the Heart of Mesoamerica. University of Colorado Press, Boulder.
Kerr, Justin, 1989–2000. The Maya Vase Book, vol. 6. Kerr Associates, New York [Illustrations also available on the Internet under ‚Kerr Portfolio'.].

Lehmann, Walter (Ed.), 1974. Die Geschichte der Königreiche von Colhuacan und Mexico. Quellenwerke zur Alten Geschichte Amerikas, vol. 1. W. Kohlhammer Verlag, Stuttgart.

Marcus, Joyce, 1992. Mesoamerican Writing Systems: Propaganda, Myth, and History in Four Ancient Civilizations. Princeton University Press, Princeton, NJ.

Markowitsch, Hans-Joachim, 2002. Dem Gedächtnis auf der Spur. Vom Erinnern und Vergessen. Wissenschaftliche Buchgesellschaft, Darmstadt.

Martin, Simon & Nikolai Grube, (2000). Chronicle of the Maya Kings and Queens. Thames and Hudson, London.

Middendorf, Ernst Wilhelm, 1890. Ollanta. Ein Drama der Keshuasprache. Verlag F.A. Brockhaus, Leipzig.

Miller, Arthur G., 1973. The Murals of Teotihuacán. Dumbarton Oaks, Washington, DC.

Mooney, James, 2011. Calendar History of the Kiowa Indians. Kessinger Publishing Company.

Morley, Sylvanus Griswold, Barrera Vásquez, Alfredo, 1949. The Maya Chronicles (= Carnegie Institution of Washington, Publication 585.). Carnegie Institution of Washington, Washington, DC.

Poma de Ayala, Felipe Guaman, 1936. Nueva Corónica y Buen Gobiemo (Codex péruvien illustré). In: Travaux et Mémoirs de L'Institut d'Ethnologie, vol. 23. Institut d'Ethnologie, Paris.

Prem Hanns J., 2008a. Geschichte Alt-Amerikas. In: Oldenbourg Grundriss der Geschichte, vol. 23. R. Oldenbourg Verlag, München [also available as e-book].

Prem Hanns J., 2008b. Manual de la Antigua Cronología Mexicana. Miguel Angel Porrúa, México.

Rabinal Achi, 1994. Un drame dynastique maya du quinzième siècle. In: Breton, Alain (Ed.), Recherches Américaines, vol. 1. Societé des Américanistes et Société d'Ethnologie, Paris.

Recinos, Adrián, 1957. Crónicas indígenas de Guatemala. Editorial Universitario, Guatemala.

Riese, Berthold, 1997. Zeitstrukturen in Mesoamerika. In: Müller, Klaus E., Rüsen, J. (Eds.), Historische Sinnbildung. Rowohlt, Reinbeck bei Hamburg, pp. 240–250.

Riese, Berthold, 2004. Die Maya und der Tod. In: Archiv für Religionsgeschichte, vol. 6. Saur, München, pp. 135–158.

Riese, Berthold, 2007. Aztekische Schöpfungs- und Stammesgeschichte. In: Ethnologische Studien, vol. 38. Lit Verlag, Münster.

Riese, Berthold, 2010. Der Untergang der Sonnengötter. Die Hochkulturen des Alten Amerika. Herder Verlag, Freiburg im Breisgau.

Riese, Berthold, 2011. Das Reich der Azteken. Geschichte und Kultur. C.H. Beck Verlag, München [also available as e-book].

Sahagún, Bernardino de, 1950–1981. General History of the Things of New Spain. School of American Research, Santa Fe.

Schultze Jena, Leonhard (Ed.), 1944. Popol Vuh. Quellenwerke zur Alten Geschichte Amerikas, vol. 2. W. Kohlhammer Verlag, Stuttgart & Berlin.

Seler, Eduard, 1960–1967. 5 vols. & index vol. Gesammelte Abhandlungen zur Amerikanischen Altertums- und Sprachenkunde. Akademische Druck- und Verlagsanstalt, Graz.

Torquemada, Juan de, 1975–1983. Monarquía Indiana. 7 vols. Universidad Nacional Autónoma de México [originally published in 1615], Mexico.

Voegelin, C.F. (Ed.), 1954. Walam Olum or Red Score. Indiana Historical Society, Indianapolis.

Historical Thought and Historiography: Islamic Tradition

Richard S Humphreys, University of California at Santa Barbara, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Islamic historiographic tradition developed across Muslim societies between the early eighth century and modern times, with a distinctive body of practices and forms owing little to those used in neighboring or preceding societies. History never became a formal discipline but was composed by self-taught amateurs, chiefly bureaucrats and religious scholars. Historical writing began with the biography of Muhammad and the controversial events of Islam's first decades, but eventually grew to encompass both history since the Creation and the events and notable persons of particular regions. In the twentieth century, Muslim historians have increasingly adopted the historiographic models of post-Enlightenment Europe.

'Islamic historiography' is a useful but somewhat misleading term commonly applied to the historical writing produced in the predominantly Muslim societies of Asia and Africa between the rise of Islam (post-CE 622) and the late nineteenth century. In contrast, history written after that point may have a distinctive regional or cultural orientation, but it increasingly falls within the post–Enlightenment historiographic discourse established and dominated by Western Europe. The term is useful because it allows us to discern the degree to which the enormous historical literature produced under the aegis of Muslim-dominated societies and polities embodies an unbroken (albeit complex and constantly changing) tradition of debate over the meaning and interpretation of the past. It is misleading insofar as it beguiles us into thinking that this literature was produced entirely by Muslims, or that it is exclusively grounded in Islamic doctrine and focused on religious issues. In fact, a number of important works within the 'Islamic' historical tradition (admittedly a small proportion of the total) were produced by non-Muslims (Hoyland, 1997). More importantly, much historical writing by Muslims focuses on decidedly nonsacred topics and presents these topics in a manner that owes little to Islamic religious concepts and concerns.

History as a Form of Knowledge and a Profession

Quite apart from the ambiguity of the term, Islamic historiography represents an intriguing paradox. Historical works were produced in great profusion by all sorts of authors and compilers from at least the mid-eighth century on, and it is evident from the number and distribution of manuscripts that these works reached a wide readership. Widely recognized technical terms for historical writing emerged very early and remained in constant use thereafter: *ta'rikh* (literally, 'dating', hence 'chronicle'), *akhbar* ('recitals of events'), and, less commonly, *sira* ('manner of life', hence 'biography'). One must not think of these terms as referring to sharply defined categories. For example, *ta'rikh* (the most widely used term) applied not only to chronologically organized works but also to historical writing of every kind, including alphabetically arranged biographical dictionaries. *Sira* most often referred to biographies of the Prophet Muhammad, but was sometimes also attached to the lives of sultans and other secular figures (Rosenthal, 1968; *Encyclopaedia of Islam* sub verbo *ta'rikh*; Robinson, 2003).

Historical writing in all genres was produced in great quantities. However, although it permeated medieval Islamic culture and was universally seen as a distinct form of literary practice, it never achieved the status of a 'science' (Arabic, *ilm*) – i.e., it never became a formal subject of study and training in the scholarly curriculum. Within the normative classification of knowledge constructed by medieval Islamic culture, history remained always on the margins of scholarship and literature. For religious specialists, it was ancillary to the core disciplines of legal studies (*fiqh*), prophetic tradition (*hadith*), and Qur'anic exegesis (*tafsir*). It was useful for identifying and evaluating the scholars who had produced and maintained religious knowledge generation after generation, for throwing some light on the circumstances in which revelation had come, for demonstrating how Islam had expanded throughout the world through divine providence, and even for clarifying the origins and meaning of the sectarian conflicts that continued to afflict the community of believers. In contrast, secular historians (who began to emerge on the scene in the tenth century) tended to focus on recent and contemporary events among the political elite. These were often bureaucrats and courtiers, as one might expect. However, many were religious scholars by training and profession, albeit men with close ties (as judges or official preachers) to the regime. They might well regard history as *magister vitae*, a trove of practical wisdom, but only rarely did they see it as the foundation for a systematic investigation of politics and society (Robinson, 2003).

Even in the humble roles of handmaid and schoolmaster, however, historical writing still generated many works of great literary and methodological sophistication. The political (and occasionally social and economic) analyses of the best Muslim historians yield nothing to their counterparts in the ancient, medieval, and Renaissance West. All this suggests that history played a much more fundamental role in defining Islamic self-understanding than that culture's spokesmen were able to articulate. On the broadest level, history – even the history produced by secular writers – showed Muslims how God was working his purposes out in the deeds and sufferings of the community of believers. In particular, it

measured the fidelity (or lack thereof) of the believers to the revelation brought by Muhammad. History challenged Muslims to measure their progress on the path to salvation or perdition as the inevitable Day of Judgment approached (Khalidi, 1994).

Historians were in effect self-taught amateurs, men who made their living as bureaucrats or religious scholars and wrote history in their spare time. They learned how to do it not through formal study with established masters, but simply by studying the works of their predecessors and contemporaries and adapting the forms and approaches of these writings to their own purposes. History, in short, was a literary practice, not a discipline. Some histories were written at the behest of rulers anxious that their deeds be properly memorialized for future generations, but even these do not represent a true professionalization of history. In late medieval and early modern times, we do find quasiofficial court historiographers under the Ottomans, Safavids, and Mughals between the sixteenth and mid-nineteenth centuries, but there was no permanent office devoted to this purpose. For the most part, the job was handed over on an ad hoc basis to any member of the ruler's entourage who was thought to possess sufficient political acumen and literary skill (Quinn, 2000; Robinson, 2003).

Two outstanding scholars did attempt to make history into a generally recognized formal science, but in spite of their great prestige and influence, neither achieved his goal. The Qur'anic exegete, jurist, and historian Abu Ja'far al-Tabari (838–923) tried to demonstrate in his vast *Chronicle of Prophets and Kings* that historical reports could be sifted and presented according to the rigorous criteria applied to *hadith* (the dicta ascribed to the Prophet, which embodied his teaching and example) (*Encyclopaedia of Islam* sub verbo 'Tabari', *ta'rikh*; Robinson, 2003). However, very few followed him down that path, even though his reconstruction both of pre-Islamic history and of Islam's first two centuries provided an authoritative framework for almost all later chroniclers. His method simply was not workable for contemporary events from the ninth century on. As for early Islamic history, most of Tabari's successors were content simply to condense and paraphrase his work (however, there was a significant, and perhaps unsurprising, revival of interest in the early sources for the Arab-Muslim conquests, *c*.630–651, among both Andalusian scholars contending with the Reconquista and Syro-Egyptian historians in the era of the Crusades). Five centuries later, the North African Ibn Khaldun (1332–1406) – a gifted legal scholar and logician as well as a historian – argued in his famous *Muqaddimah* that history was the science of society, and as such ought to be treated as a branch of formal philosophy. Again, his argument did not suffice to move historiography up to that exalted status, in spite of the substantial circulation that his work enjoyed (Mahdi, 1957; *Encyclopaedia of Islam* sub verbo Ibn Khaldun; Al-Azmeh, 1990). On the other hand, he was not without influence. The historians of fifteenth-century Egypt (some of them his direct students) paid far closer attention to social and economic life than had earlier historians. Likewise, certain officials in the seventeenth- and eighteenth-century Ottoman Empire were quite interested in his theories on the causes for the rise and fall of dynasties.

The Producers of Historical Knowledge

As we have seen above, history was written (and read) by highly literate men with some time on their hands. Given the social and educational structures of medieval Muslim societies, that means that it was written chiefly by chancery secretaries (Arabic, *kuttab*; singular, *katib*) and religious scholars (Arabic, *ulama*; singular, *alim*). It is tempting to describe these two groups in terms of well-defined ideal types, and thus to posit opposing historical perspectives of palace and mosque. In reality, these perspectives overlapped in complex and constantly changing ways (Rosenthal, 1968; Fleischer, 1986; Humphreys, 1991; Quinn, 2000; Robinson, 2003).

Throughout Islamic times, the *ulama* were given state appointments as judges or mosque officials, although it is true that a great many of the most prestigious scholars prized their independence and refused to accept such positions. But after the mid-eleventh century, even the independent-minded scholars found it difficult to refuse professorships in the burgeoning religious colleges (madrasahs) that were endowed and to some degree controlled by the rulers and their senior military commanders. However scrupulous they might be, scholars could hardly avoid some entanglement with the princely courts. Quite apart from questions of material interest, most also had a profound fear of social disorder – as a popular dictum had it, one night of anarchy was worse than 60 years of tyranny. Thus, they seldom challenged or even questioned the underlying political order of their societies, however severely they might criticize particular rulers.

On the other hand, their relations with the princes were always ambivalent. They knew that rulers came and went like the clouds. They might collaborate with the regime in power but rarely identified themselves closely with it. Their job was to embody Islamic beliefs and values in their own lives and to propagate these within society at large. Religious scholars came from a wide range of social backgrounds and interacted constantly with the urban population. They indeed derived much of their prestige (the only solid basis of their social influence) from their accessibility to all.

As historians, then, the religious scholars were the spokesmen for the community and its concerns. As such, they stood not so much against the princely courts as apart from them. They naturally placed the rulers and their doings at center stage in their historical writing, since the rulers were the heads of the Muslim community within the territories they ruled – responsible not only for defense, public order, and justice but also for ensuring that Islamic religious practices, moral norms, and legal procedures were maintained. Without a Muslim ruler, however incompetent or corrupt, an authentic Islamic way of life was impossible. However, rulers were not the sole guardians of these things; scholar–historians were acutely aware of the indispensable role of their own class. They asserted this role in every word and gesture, and by recording their own deeds in countless biographical dictionaries of eminent religious scholars.

In contrast to the religious scholars, chancery secretaries were men of the regime, who shared and indeed articulated its perspectives on rulership and society. The middle and upper ranks especially tended to live in a world apart from the people they governed. Their lives and fortunes were held at the whim

of an autocrat, whose purposes and moods were of far greater importance to them than the problems of the larger society outside the palace walls. In spite of their narrow focus, the secretaries often took a surprisingly critical view of the regimes they served. Usually, they judged them in accordance with the political traditions of pre-Islamic Iran, although these were often blended with specifically Islamic values derived from the Qur'an and teachings ascribed to the Prophet (*hadith*). These traditions centered on the unique role of the ruler in upholding justice within his realms. In practice, that meant respect for established social hierarchies, equitable taxation, enforcing right religion, and control of the army. The chancery secretaries might well be sincere Muslims, but for them, the ruler and his senior officials were the sole actors who really counted. Hence, their historiography is strongly focused on the court and on the struggle for power. The best historians provide searching analyses of this struggle, to be sure, but with a few notable exceptions, they portray it as the affair of a few score individuals, not of broad social groups and forces.

The Problem of Origins, CE 622–825

Islamic historiography was born in obscurity. The earliest historical texts that we now have only acquired something close to their present form in the late eighth century. For earlier stages of development, we are reduced to inference or, more accurately, speculation. We can at least be sure that Islamic historical writing emerged within Arabophone Muslim circles, for the most primitive elements in this literature are couched in archaic Arabic and reflect almost exclusively the internal concerns of the Muslim community. Islamic historiography emerged out of the triumph and trauma of the seventh century: The sudden rise of a new religion among the peoples of Arabia, the vast conquests that they made under the banner of that religion, the profound transformation that religion and empire wreaked on their traditional culture and sociopolitical organization, and the bitter civil strife that rent a unified community of believers into hostile sects and factions. To make sense of these tremendous events, early Muslims had to devise some way of recalling, ordering, and interpreting them (Khalidi, 1994; Donner, 1998; Robinson, 2003).

By the early eighth century, they discovered an overarching pattern of meaning and value in the Qur'an. The Qur'an (the collection of the revelations received by Muhammad between his call to prophethood *c*.610 and his death in 632) is far from a historical record of the earliest Muslim community. It provides no narratives of Muhammad's life and struggles; it only alludes to a few of the events that formed that community. Rather, it places Islam within the long sequence of prophetic messages through which God has repeatedly called his creatures back to obedience. In the Qur'an, Muslims could see how a prophet had been sent to every person, bearing the promise of salvation for those who believed and the threat of damnation for those who refused. Most people did reject their prophets and suffered destruction. In contrast, a very few (most notably the Jews and Christians) accepted their prophets and prospered so long as they adhered to the messages these prophets had brought them. Over time, however, even these believing people tended to fall away, and God called them to repentance by periodic chastisement and humiliation. The Qur'an thus laid down for Muslims a paradigm that explained the course of historical events. This paradigm contained three elements: the covenant offered by God to his people, the ever-present threat that this covenant might be betrayed by its recipients, and the desperate hope for redemption even after betrayal. In the circumstances of the seventh and eighth centuries, the Qur'anic paradigm of Covenant–Betrayal–Redemption was very compelling indeed (Humphreys, 1991; Donner, 1998).

Paradigms do not create literary genres. To recount the things that had befallen them, the early Muslims developed short prose narratives (Arabic, *akhbar*; singular, *khabar*) focused on discrete events. The earliest historical narratives about the seventh-century conquests and the internal struggles within the community of believers sprang up in a spontaneous, uncontrolled manner. They represented the contradictory memories and claims of innumerable tribal spokesmen and sectarian advocates, not a master narrative propounded and controlled by official spokesmen of the caliphal court. They were originally (and perhaps remained for some decades) oral accounts, and changed constantly as they were told and retold. Moreover, highly partisan narratives inevitably provoked counter-narratives. With a very few exceptions, it is no longer possible to recover the primitive form of any of these early narratives. That was no doubt the case even for the scholars of the mid-700s (Wellhausen, 1899; Noth, 1994; Donner, 1998; Robinson, 2003).

The earliest methodical effort to identify, collect, and reduce to writing the most reliable (i.e., religiously and politically acceptable) of these floating, infinitely plastic narratives took place around 700 under the aegis of the Umayyad caliphs. It is important to note that these first written versions were not books meant for public circulation, but rather notes and drafts for personal use. The information in them was for the most part transmitted orally, not via written copies. Only by the 730s can we think of more or less organized collections of such narratives (Schoeler and Toorawa, 2009). In any case, the Umayyads could impose no monopoly on the process of collection and redaction; indeed, their version of Islam's first century was nearly obliterated when they were overthrown by the Abbasids in 749–50 and subjected to a systematic *damnatio memoriae*. Only between 750 and 830 did a group of scholars, some with ties to the Abbasid court and some quite independent of it, began to generate a stable corpus of historical narratives. This corpus represents what we now regard as the 'mainstream' version of early Islamic history. The Abbasid court used its patronage to encourage the articulation of an overarching narrative that would support its claims to political and religious authority. Its capacity to control this process was very limited, however, and numerous competing versions produced by rivals and opponents of the Abbasids survive, albeit in fragmentary or masked form. However obscure the 'realities' of early Islamic history may be, the historical debate of the late eighth and early ninth centuries is quite visible (Lassner, 1986; Robinson, 2003; Borrut, 2010).

The narratives collected, redacted, or (sometimes) created in the early Abbasid period were usually gathered in 'monographs' devoted to a single event (typically a battle) or category of information (e.g., lists of provincial governors). However, this period also witnessed the first major syntheses. Ibn Ishaq

(d.767) produced his famous biography of the Prophet Muhammad (the oldest complete work of Islamic historiography to reach us, albeit in redacted form), and the following decades saw a host of works on the early conquests, the first caliphs, etc. Structurally, these syntheses are compilations of discrete reports (*akhbar*) dealing with specific episodes, not continuous integrated narratives. On the other hand, the editors shaped their compilations in accordance with a conscious (although usually unstated) design and agenda. Midrange syntheses of this kind dominated Islamic historiography down to the late ninth century, and they never ceased to be produced altogether.

The early Abbasid period also witnessed a plethora of efforts to develop a reliable absolute chronology, although the first essays in this direction seem to have been made under the later Umayyads. Chroniclers naturally used the *hijri* calendar, traditionally instituted by the caliph Umar I in 642. But of course, events that had occurred before 642 had to be backdated, while the precise dates for a host of later events could only be guessed at. Nevertheless, Muslim historians quickly came to feel that precise dating was an essential element of their craft; the effort to pin things down and put them in the correct chronological order became a permanent characteristic of Islamic historical practice. By the 850s at latest, we begin to see annalistic chronicles of the kind long familiar in Byzantine historiography – viz., a year-by-year calendar of significant events (usually recounted in very summary fashion), beginning at a fixed point and coming down to the present. In the Byzantine tradition that fixed starting point was typically the Creation, while early Muslim historians favored the Hijra (CE 622). The annalistic structure of the chronicle was a rigid but very useful way of ordering events, and by the eleventh century, it had become by far the most widely used format in Arabic language historiography (Rosenthal, 1968).

The early ninth century witnessed the appearance of one last genre of synthetic historiography, the prosopography. The first major extant work of this class was the *Large Book of Generations* (*Kitab al-tabaqat al-kabir*) of Ibn Sa'd (d.845). His goal was to compile concise biographies of all those persons who had transmitted Islamic religious knowledge from generation to generation, beginning with the Prophet himself. The biographical entries in Ibn Sa'd's compilation seldom provided a detailed account of the events of a person's life, let alone any effort to probe the person's personality. Rather, they aimed at evaluating his (or her) religious standing and reliability as a scholar. Although his own book was little read until the eleventh century, Ibn Sa'd had clearly addressed a deeply felt need, for succeeding centuries (including the twentieth) saw a profusion of such works. In addition to countless biographical compilations dedicated exclusively to religious scholars, there were many others dealing with poets, men of letters, mystics, physicians, bureaucrats, and ultimately all those (including a number of women) who had earned a place within the cultural memory of Islam (Robinson, 2003).

By the last quarter of the ninth century, Muslim historians began to produce a series of grand syntheses, which claimed as their subject matter all of Islamic history, or even the history of the world since the Creation. Among many important works, pride of place undoubtedly belongs to the *Chronicle of Prophets and Kings* (*Ta'rikh al-rusul wa'l-muluk*) of al-Tabari (838–923). Apart from its methodological rigor, al-Tabari's compilation is distinguished by its vast scale, its meticulous attention to Biblical and Iranian history – Rome is conspicuous by its absence – and its inclusion of conflicting perspectives on the critical events of Islam's first 250 years. al-Tabari was a somewhat controversial figure in his own time, but his authority among later generations was immense. Most of his successors were content simply to abridge and condense his account of early Islamic history, but from him they did learn to place Islam within God's plan of creation and salvation by linking it to the religions and empires that had preceded it (Humphreys, 1991; Khalidi, 1994; Donner, 1998).

The Mature Tradition, 950–1800

By the time al-Tabari died, the Covenant–Betrayal–Redemption paradigm that had shaped historical discourse throughout the eighth and ninth centuries seemed only remotely connected to the issues now confronting Muslims throughout the world. Historians now confronted the collapse of the unitary caliphate and the rise of a host of warlord-dominated regimes throughout the Islamic lands – in short, the political disintegration of the community of believers. Moreover, the religious conflicts that had arisen in Islam's first two centuries could no longer be healed; by the mid-tenth century what had been fluid if bitterly contested allegiances were transmuted into rigidly defined sectarian divisions. Political and religious fragmentations were unwelcome facts, but facts they incontrovertibly were, and they deeply shaped the historical writing of the following centuries.

First, the new environment provoked a change in subject matter. In striking contrast to al-Tabari and his predecessors, historians now began to focus on recent and contemporary events rather than the community's early history. Likewise, historians tended to frame their work in local and regional rather than universal terms; their terrain was more often the history of Isfahan or Damascus than of the Islamic world as a whole. This emphasis on recent and regional or local events encouraged historians to adopt dynastic history as the framework for their narratives, for dynasties (and the warlords who created them) were the most visible political reality in their world. The turn to dynastic history was linked to new interpretive concerns. For historians observing the rise and fall of the warlord states, the central problem was not a ruler's role in the grand drama of salvation, but rather the reasons for the success or failure of his regime (Humphreys, 1991; *Encyclopaedia of Islam* sub verbo *ta'rikh*; Robinson, 2003; Hirschler, 2006).

Within this new historiographic culture, universal history – i.e., history that claimed to include the whole Islamic world since the time of the Prophet, or even the Creation – continued to be written, and in fact enjoyed a significant revival in the thirteenth and fourteenth centuries (Radtke, 1992). But even the finest work of this genre, the *Complete History* (*al-Kamil fi al-ta'rikh*) of Ibn al-Athir (1160–1234), is strongly inflected by

the distinctive concerns of its era. Within its strict annalistic framework is imbedded a quite visible sequence of dynastic histories. It does cover all the lands of Islam with remarkable balance and detail, but it does so by examining each region seriatim in its survey of the events of each year. Finally, its primary focus is the mundane political process and (implicitly) the pragmatic lessons that could be inferred from that process.

Two other universal histories of great originality are equally products of their time and place. The *Comprehensive Chronicles* (*Jami' al-tawarikh*) composed under the supervision of Rashid al-Din (d.1310), vizier of the Mongol empire in Iran and Iraq, is a work that embodies the aspirations to universal rule of the rulers for whom it was written, for it deals with the history of every person about whom Rashid al-Din could obtain any information. Even the usually despised and neglected Franks are included within his vast tableau. It is an irreplaceable encyclopedia of the historical knowledge available to Iranian scholars in the early fourteenth century. Finally, the *Book of Examples* (*Kitab al-'Ibar*) of Ibn Khaldun (1332–1406) is distinguished not only by its famous introductory volume (the *Muqaddimah*) but by a well-constructed, comprehensive, and analytic survey of peoples and cultures – a survey that nicely displays the socioeconomic acumen of its author.

Apart from these shifts in subject matter and emphasis, the mid-tenth century witnessed the bifurcation of historical writing into two linguistic branches. Arabic (the traditional language of Islamic scholarship) was now joined by Persian, the language spoken at the emerging princely courts of Iran and Muslim Central Asia (Meisami, 1999; Quinn, 2000). Arabic continued to be used throughout the Iranian cultural zone (which came to include Central Asia, Anatolia, and much of South Asia) for the religious sciences, but Persian rapidly supplanted it in other arenas: secular and religious poetry, government administration, didactic prose, and history. The use of Persian was more than a linguistic phenomenon, for Persian historiography quickly developed distinctive characteristics of its own – e.g., a preference for lengthy continuous narratives, an explicit appeal to the political precepts and institutions of ancient Iran, and a chronology that tends to focus on princely reigns rather than the annual calendar. By the eleventh century, Arabic and Persian historiography clearly embody two distinct traditions, each developing in its own way. This picture was further complicated by the emergence of Turkish as a major vehicle of literary expression in the late fifteenth century. From this point down to the mid-nineteenth century, Turkish language works were composed throughout the Middle East – no surprise, since the political and military elites from the Nile to northwest India were overwhelmingly of Turkish origin during that period. The main centers of Turkish literature were located in predominantly Turkish-speaking areas – Central Asia and the core Ottoman lands of Anatolia and the eastern Balkans. In spite of some specific characteristics of its own and many works of outstanding quality, however, Turkish historiography tended to model itself upon the Persian tradition, and in some regions, the two languages coexisted almost on an equal basis as vehicles of literary expression (Fleischer, 1986; Faroqhi, 1999).

The Transition from Islamic to European Models, 1800–1950

By 1800, the Muslim peoples of Asia and North Africa faced severe military and economic challenges from Europe. However, most Muslim polities (with the important exceptions of the Mughal Empire in India and the sultanates of Java) still retained their political independence, and the deep-rooted societal and cultural structures of the Islamic lands seemed intact. By 1900, however, almost every Muslim state had been swallowed up or made a client regime, and Muslim societies were enmeshed in massive, uncontrollable transformation on every level of life.

Change on such a scale inevitably compelled the creation of a radically new historiography, although the process was slow and complex. Works of great merit continued to be produced within the existing traditions throughout the first half of the nineteenth century. Indeed, excellent historical scholarship of a decidedly traditional cast continued to be produced well into the twentieth century. The best historians were fully cognizant of the new world they inhabited, but for many decades, traditional historiographic forms and approaches seemed adequate to the task. By the 1870s, this was no longer the case for many writers. First, existing models of historical writing simply could not address the cultural and ideological dimensions of the European challenge. No less important, by this time, many Muslim intellectuals had studied in Europe or with European teachers, and they felt the powerful allure of Enlightenment and post-Enlightenment thought. To them it was self-evident that European military and economic power was grounded in the power of European ideas (Hourani, 1962).

European ideas and literary models were easier to perceive than to internalize, and the late nineteenth century witnessed a difficult struggle by Muslim intellectuals and scholars to articulate new approaches to historical knowledge. Most late-nineteenth-century historical writing is in some sense a hybrid product. The point is illustrated by two major works, both marked by broad scope, critical use of documents, and awareness of European scholarship. Ahmet Cevdet Pasa's history of the Ottoman Empire (12 vols.: 1855–84) adheres to the traditional chronicle form, while Ali Mubarak's historical geography of Egypt (20 vols.: 1886–88) is broadly modeled on a classic work by the fifteenth-century Cairene scholar al-Maqrizi (d.1442). A model for historiography in a modern (i.e., European) style was provided by the *History of Islamic Civilization* (*Ta'rikh al-tamaddun al-islami*, 5 vols.: 1902–06) of Jurji Zaydan, a Syrian Christian novelist and man of letters. For both political and cultural reasons, Zaydan had no imitators, and only after World War I did Middle Eastern historiography – particularly in the new Turkish Republic and Egypt – begin to adhere systematically to contemporary European modes of research and presentation. This shift was more visible and consistent in works dealing with modern history (Delanoue, 1982; Crabbs, 1984; Di-Capua, 2009).

In both Turkey and Egypt, the new approach was strongly encouraged and to some degree controlled by the regime. Atatürk established an official historical society, which generated a nationalist interpretation of Ottoman, and modern Turkish history, which was almost uncontested down at least to

the 1970s. In Egypt, King Fu'ad (1917–35) oversaw a similar enterprise, but did so by hiring a number of European scholars to create a narrative of Egypt's rapid path to modernization under the strong leadership of its nineteenth-century viceroys (who also happened to be his own ancestors). Alongside this official initiative, however, emerged a cluster of Egyptian academic historians who strove to pursue an independent path. They were trained at the new King Fu'ad (now Cairo) University under the aegis of Shafiq Ghurbal (d.1961), himself trained at the University of London. By the end of the 1930s, even studies of early and medieval Islamic times were brought within the new framework. The 'Europeanization' of historiography affected not only research methods and literary structure but also the kinds of questions that Muslim scholars asked about their past. Even where (for religious or political reasons) they rejected the frames of reference and interpretations put forward by European historians, they could not avoid wrestling with the problems that these scholars had defined (Di-Capua, 2009).

This transformation in approach clearly reflects the professionalization of history after World War I – a sharp departure from traditional practice in the Islamic world. This process began with the establishment of European-style universities in Cairo and Istanbul just before the war, along with the creation of national societies for historical research. Simultaneously, history was enlisted in the struggles for national independence and cultural autonomy of the interwar and postwar periods. Moreover, by World War I, a small but growing number of Muslim scholars were taking advanced degrees from the leading universities in Europe; the prestige and authority that they enjoyed when they returned home in effect imposed the new model of historical thought and writing on their students.

Cross-cultural influence in historiography has not flowed in a single direction, however. Even in the 1920 and 1930s, a few Muslim scholars published their work in Western languages, and since the 1950s, this trend has grown enormously. A growing number of scholars from Islamic Asia and Africa have not only taken their advanced degrees in Europe and North America but also been appointed to many of the most prestigious teaching posts there. This situation has begun to create a real dialogue among a plurality of theoretical and cultural perspectives. Even though the methods and practices of traditional Islamic historical writing have all but disappeared, the history of the Islamic world has increasingly fallen again into the hands of its own scholars – the natural state of things, perhaps, but one that was hardly the case in the nineteenth and twentieth centuries.

See also: Historical Thought and Historiography: South Asia; Historical Thought and Historiography: Southeast Asia; Historical Thought and Historiography: Sub-Saharan Africa; Nationalism, Historical Aspects of: Arab World; Religion, History of; Religious Nationalism and the Secular State; Time, Chronology, and Periodization in History.

Bibliography

Al-Azmeh, Aziz, 1990. Ibn Khaldun: An Essay in Reinterpretation. Routledge, London and New York.
Borrut, Antoine, 2010. Entre mémoire et pouvoir: l'espace syrien sous les derniers Omeyyades et les premiers Abbassides. Brill, Leiden.
Crabbs, Jack A., 1984. The Writing of History in Nineteenth-Century Egypt: A Study in National Transformation. Wayne State University Press, Detroit, MI.
Delanoue, Gilbert, 1982. Moralistes et politiques musulmans dans l'Égypte du XIXe siècle. Institut Français d'Archéologie Orientale, Cairo.
Di-Capua, Yoav, 2009. Gatekeepers of the Arab Past: Historians and History Writing in Twentieth-Century Egypt. University of California Press, Berkeley, CA.
Donner, Fred M., 1998. Narratives of Islamic Origins: The Beginnings of Islamic Historical Writing. Darwin Press, Princeton, NJ.
Ende, Werner, 1977. Arabische Nation und islamische Geschichte: die Umayyaden im Urteil arabischen Autoren des 20. Jahrhundert. Orient-Institut der Deutschen Morgenländischen Gesellschaft, Beirut.
Encyclopaedia of Islam, second ed., 1960–2005 See entries for ta'rikh and individual historians.
Faroqhi, Suraiya, 1999. Approaching Ottoman History: An Introduction to the Sources. Cambridge University Press, Cambridge.
Fleischer, Cornell, 1986. Bureaucrat and Intellectual in the Ottoman Empire: A Biographical Study of the Historian Mustafa Ali, 1541–1600. Princeton University Press, Princeton, NJ.
Hirschler, Konrad, 2006. Medieval Arabic Historiography: Authors as Actors. Routledge, London.
Hourani, Albert H., 1962. Arabic Thought in the Liberal Age, 1797–1939. Oxford University Press, London.
Hoyland, Robert G., 1997. Seeing Islam as Others Saw It: A Survey and Evaluation of Christian, Jewish, and Zoroastrian Writings on Early Islam. Darwin Press, Princeton, NJ.
Humphreys, R. Stephen, 1991. Islamic History: A Framework for Inquiry, rev. ed. Princeton University Press, Princeton, NJ.
Keaney, Heather N., 2013. Medieval Islamic Historiography: Remembering Rebellion. Routledge, London.
Khalidi, Tarif, 1994. Arabic Historical Thought in the Classical Period. Cambridge University Press, Cambridge, UK.
Lassner, Jacob, 1986. Islamic Revolution and Historical Memory: An Inquiry into the Art of Abbasid Apologetics. American Oriental Society, New Haven, CN.
Lewis, Bernard, Holt, Peter Malcolm (Eds.), 1964. Historians of the Middle East. Oxford University Press, London.
Mahdi, Muhsin, 1957. Ibn Khaldun's Philosophy of History. Allen and Unwin, London.
Meisami, Julie Scott, 1999. Persian Historiography to the End of the Twelfth Century. Edinburgh University Press, Edinburgh, UK.
Mustafa, Shakir, 1983–1993. Al-Ta'rikh al-'arabi wa'l-mu'arrikhun. In: Arabic Historiography and Historians, third ed., 4 vols. Dar al-'Ilm li'l-Malayin, Beirut.
Noth, Albrecht, 1994. The Early Arabic Historical Tradition: A Source-Critical Study (L.I. Conrad, Trans.), second ed. Darwin Press, Princeton, NJ.
Quinn, Sholeh A., 2000. Historical Writing during the Reign of Shah 'Abbas: Ideology, Imitation, and Legitimacy in Safavid Chronicles. University of Utah Press, Salt Lake City, UT.
Radtke, Berndt, 1992. Weltgeschichte und Weltbeschreibung im mittelalterlichen Islam. Orient Institut der Deutschen Morgenländischen Gesellschaft, Beirut.
Robinson, Chase F., 2003. Islamic Historiography. Cambridge University Press, Cambridge, UK.
Rosenthal, Franz, 1968. A History of Muslim Historiography, second ed. Brill, Leiden.
Schoeler, Gregor, Toorawa, Shawkat M., 2009. The Genesis of Literature in Islam; From the Aural to the Read. Edinburgh University Press, Edinburgh, UK.
Schoeler, Gregor, Vagelpohl, Uwe, Montgomery, James E., 2011. The Biography of Muhammad: Nature and Authenticity. Routledge, London.
Wellhausen, Julius, 1899. Prolegomena zur ältesten Geschichte des Islams. In: Wellhausen, J. (Ed.), Skizzen und Vorarbeiten, vol. 6. George Reimer, Berlin.

Historical Thought and Historiography: Jewish Tradition

M Brenner, University of Munich, Munich, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 6793–6798, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article combines a brief analysis of premodern Jewish historical traditions from the Hebrew Bible and Rabbinic sources to sixteenth-century chronicles with a stronger emphasis on modern Jewish historiography as it emerged in the early nineteenth century. It points to the fact that modern Jewish historiography did not grow out of earlier Jewish historical traditions but rather out of the critical approaches that characterized the early nineteenth-century German academic environment. During the nineteenth and twentieth centuries several schools of Jewish history writing developed, usually influenced by their immediate environment.

Biblical Traditions of History Writing

Together with the historical traditions of ancient Greece the Hebrew Bible contains the oldest corpus of historical writings. While not all of the various books, which entered the Jewish biblical canon reveal a strong sense of history, quite a few – such as Samuel, Kings, and Chronicles – are primarily historical accounts. In contrast to the Greek historians, the authors and editors of the Bible were interested not in recording history but in giving a theological sense to history. The numerous biblical passages which command the Jews to remember (*zakhor*) do not imply a special interest in history itself but express a definition of the Jews as a historical people held together first and foremost by its will to observe its historically revealed laws and to remember its common history. God is described in terms of His historical deeds. He is the 'God of the fathers' and 'Your God who brought you out of the Land of Egypt.'

With the canonization of the Hebrew Bible in the second-century Christian Era (CE) Jewish historical writing undertakes a clear shift. The first-century CE produces the last major Jewish historical writings for about 15 centuries. In a historical work, of which only two out of five parts survived, Philo of Alexandria describes the persecution of Egyptian Jews by the Roman governor Flaccus and the delegation of Alexandrian Jews (including himself) to the emperor Caligula. The most important historian in this period is Flavius Josephus, whose account of the Jewish War against Rome in CE 66–73 was written a few years later in Roman exile and – together with his *Antiquities* – stands in the tradition of Roman rather than Jewish historiography.

Historical Thought in Rabbinic Literature and Medieval Sources

There is hardly a more telling symptom for the decline of Jewish historical writing than the fact that Josephus' histories fell into oblivion among the Jews during the Talmudic period. The same holds true for many historical accounts, which did not enter the Jewish Biblical canon, such as the Books of the Maccabees, which describe the rise of the Hasmoneans against Hellenism. Preserved by non-Jews, they were rediscovered by Jews only in the sixteenth century. During the Middle Ages the history of the Second Temple period – beyond the Biblical accounts themselves – was known among Jews in rather distorted versions of the original accounts, such as the *Book of Antiochus* and the tenth-century *Book of Yossipon* (believed to be Josephus' original account), both produced in southern Italy.

The Talmud, which was to become the single most important source for Rabbinic Judaism over the centuries, is unhistorical in its nature. Although this vast corpus contains quite a few passages of historical value, its editors reveal no special interest in the chronology of events. Thus, rabbis of different centuries are made to engage in conversations, and biblical events are often presented out of their chronological order.

Medieval Jewish historical thought is framed by the memory of the biblical past on the one hand and by the longing for a messianic future on the other hand. The time span between a golden past and an ideal future is usually interpreted with clear reference to the biblical course of events. Biblical place names like Ashkenaz and Sefarad designate European countries, Christianity is Edom, and Islam is Yishmael. Vicious rulers are called Haman, like the archetypical biblical villain from the Book of Esther, and saviors are termed Mordechai after the hero of the same story. Messianic expectations led to concrete calculations about the End of the Days, and contemporary tragedies were often seen as the birth pangs of the Messianic Age.

Altogether, traditional Jewish historical thought as expressed in Rabbinic literature and medieval sources reveals at least a partly cyclical understanding of time. Every year when celebrating Passover observant Jews remember the exodus from Egypt, and during three annual days of fasting they mourn the destruction of Jerusalem. Each rescue from persecution is regarded a new Purim (in commemoration of the Book of Esther). Every Sabbath morning a weekly portion of the Pentateuch is read in the synagogue, thus reviving sequentially each year the historical accounts of the Five Books of Moses.

The most prominent genre of historical writing during the Middle Ages was the chronologies, which often tried to establish a reliable 'chain of tradition' (*shalshelet hakabbala*) from Moses through the leading Rabbis of the Mishna and Talmud and up to the authors' own time and place. They constituted a refutation of the anti-Talmudic challenge made by the Karaite sect and at the same time underlined the importance of the local Jewish community in which they were produced.

Examples of this genre are the *Letter of R. Sherira b. Hanina* from the academy at Pumbedita in Babylonia, who in the late tenth century responded to scholars from Kairouan, and the twelfth-century *Sefer Hakabbala* (*Book of Tradition*) of the Spanish scholar Abraham ibn Daud. A stronger emphasis on contemporary events can be seen in the detailed accounts of the Crusade massacres concerning the cruelties, which befell the Jewish communities of the Rhineland in the eleventh and twelfth centuries.

Among the most popular sources reflecting medieval and early modern Jewish historical thought in Central Europe were Memor Books, which contain collections of prayers together with a necrology of distinguished persons, and recorded local martyrs and persecutions. The oldest known Memor Book is from Mainz and dates back to 1296. A few hundred Memor Books are extant, mainly from seventeenth-century rural communities in southern Germany.

Sixteenth-Century Jewish Historiography as a Response to Catastrophe

The sixteenth century saw a sudden upsurge of Jewish historical writing with the publication of 10 important historical works, more than all of those known to us from the Middle Ages. Quantitatively as well as qualitatively the development was significant, with a new interest in history itself, as Yosef Hayim Yerushalmi demonstrated in his seminal study *Zakhor* (1989). Writers like Yosef ha-Kohen, Abraham Zacuto, and Samuel Usque wrote in the wake of the expulsion of Jews from Spain and Portugal (1492 and 1497), which affected not only their own lives but also completed the wave of expulsions of Jews from Western Europe. The new search for explanation is apparent in their writings, perhaps most pointedly in Solomon ibn Verga's *Shevet Yehuda* (The Staff of Judah, 1554), but also in the historical oeuvre of the only Ashkenazic author of this group, David Gans from Prague. His *Tsemah David* (Offspring of David, 1592) contains the most precise account of Jewish and world history undertaken by a Jewish author until that time.

The Italian Jew, Azaria de' Rossi, who completed his *Me'or Enayim* (A Light to the Eyes, 1573–5) at about the same time, has the merit of recovering ancient sources that had been lost to Jewish tradition because they did not enter the biblical canon. Among them are works by Philo of Alexandria and the apocryphal *Letter of Aristeas*. For the first time the historical interest of Jewish writers did not concentrate on the biblical past or speculations about the messianic future but dealt with the concrete history of the Jews during recent centuries and in their own time. These authors also showed a genuine interest in the histories of other nations. The impact of their work on later Jewish historiography was, however, minimal. It was only as a result of the encounter with academic scholarship in the early nineteenth century that modern Jewish history writing developed.

The Beginnings of Modern Jewish Scholarship and Critical Historiography in the Nineteenth Century

The beginnings of modern Jewish scholarship can be dated to the small circle of Jewish university students from Berlin who founded the *Verein für Cultur und Wissenschaft der Juden* in 1819. Leopold Zunz, the leading spirit of this group and the dominant figure in the field, which would become known as *Wissenschaft des Judentums*, published in 1818 a systematic agenda for modern Jewish research under the title *Etwas ueber die rabbinische Literatur* (*On Rabbinic Literature*). The study of history, which had been marginal in the works of eighteenth-century Jewish enlighteners (*maskilim*), now became a central pillar of Jewish scholarly enterprise. From the beginning, a certain apologetic character was inherent in many attempts at reclaiming the Jewish past. As Immanuel Wolf, a founding member of the *Verein* expressed it: "Scientific knowledge of Judaism must decide on the merits and demerits of the Jews, their fitness or unfitness to be given the same status and respect as other citizens."

In a more direct way, the scholarly studies of Leopold Zunz served as instruments to achieve emancipation for German Jews. Like Wolf, Zunz believed that 'the equality of the Jews in customs and life will follow from the equality of *Wissenschaft des Judentums*.' In his *Gottesdienstliche Vorträge der Juden, historisch entwickelt* (1832), he documented the long tradition of Jewish sermonic literature, thus rebutting the Prussian government's dismissal of German synagogue sermons as innovations in Jewish liturgy. Five years later, in his study on Jewish surnames, *Die Namen der Juden*, Zunz proved the ancient origin of many 'non-Jewish' names among Jews and thus reverted to a decree that forbade the use of Christian first names among Jews.

It would, however, not do justice to the enormous enterprise of nineteenth-century *Wissenschaft des Judentums* to view it predominantly under the aspect of apologetics, as did later critics, especially Zionists, most prominently Gershom Scholem. Another important motif for the pioneers of Jewish historiography was to retrieve the lost treasures of the past for a rapidly assimilating Jewish community as a means of strengthening their particular identity as modern Jews. As Yosef Hayim Yerushalmi writes, "History becomes what it had never been before—the faith of fallen Jews" (1989: p. 86). Finally, the purely scholarly interest in the Jewish past at a time in which history writing and reading became one of the most popular intellectual endeavors should not be disregarded.

There was no bridge between those first scholarly accounts of Jewish history and earlier Jewish writing. It was through the non-Jewish world that historiography entered Jewish writing. Consequently, it was not a Jew, but a Protestant French author in Dutch exile, Jacques Basnage, who wrote the first multivolume postbiblical Jewish history. He characterized his *Histoire du peuple Juif depuis Jésus Christ jusqu'à présent* … (1706–11, rev. ed., 1716–21) in his subtitle as a continuation of Josephus, and there was indeed some truth to this statement. To be sure, his account would in no way measure up to modern scholarly standards, and was still full of missionary intentions, but it was nevertheless the first serious attempt to write a comprehensive postbiblical Jewish history.

Political Agendas and the Writing of Jewish History

The first Jewish writer to embark on such an endeavor was Isaak Markus Jost, a member of the *Verein für Cultur und Wissenschaft*

der Juden and one of the pioneers of *Wissenschaft des Judentums*. The modern secular education he received as a student at Berlin and Goettingen is visible in his attempt to present a sober and objective view of Jewish history in his nine-volume *Geschichte der Israeliten seit der Zeit der Maccabäer bis auf unsere Tage* (1820–28). Jost's work is clearly centered on the German–Jewish experience and draws a picture of Judaism as a religion, which has lost all its original national characteristics. Depicting modern Jews as a purely religious community was part of the larger German–Jewish struggle for Emancipation. The title of his later work, the three-volume *Geschichte des Judentums und seiner Sekten* reveals the urge to present Judaism as a religion analogous to Christianity.

When Heinrich Graetz wrote his 11-volume *Geschichte der Juden von den ältesten Zeiten bis auf die Gegenwart* – the most popular such history ever written – one generation later (1853–76), he deviated from the early works of *Wissenschaft des Judentum* representatives in his stronger emphasis on the existence of a Jewish people and by a more passionate presentation. Although he was later claimed as a proto-Zionist writer, it is important to note that his definition of the Jewish nation is that of a community held together by an essentially religious idea. Like Jost, for whom modern Jewish history begins with the enlightened Prussian monarch Frederick the Great, Graetz located the beginnings of Jewish modernity in Germany, with the appearance of the philosopher Moses Mendelssohn and the Berlin Jewish Enlightenment.

It was not until the late nineteenth century that a clear break from the earlier Germanocentric view of Jewish history and from the strong emphasis upon a 'history of suffering and scholarship' (Graetz) appeared in the work of the Russian-Jewish historian Simon Dubnow. His 10-volume *World History of the Jewish People* was written in Russian but published first in German translation (1925–29). Motivated partly by his own political agenda – he was the founder of the autonomist Jewish movement which represented Jewish diaspora nationalism in Eastern Europe – his version of Jewish history centered on institutions of Jewish life in the diaspora, most notably the *kehilla*, the semiautonomous Jewish community.

Dubnow was also among the founders of the YIVO Institute for Jewish Research – an institution established in Berlin and Vilnius in 1925 to systematically research the Jewish past and present in Eastern Europe and other Ashkenazic communities. In contrast to the traditional German dominance of nineteenth-century *Wissenschaft des Judentums* and to Hebrew-language Zionist scholarship, the YIVO deliberately presented its research in Yiddish, the language spoken by the Jewish masses of Eastern Europe.

From the end of the nineteenth century, Jewish historiography was increasingly influenced by new subdisciplines of *Wissenschaft des Judentums*, such as ethnography, sociology, and demography. The *Gesellschaft fuer Juedische Volkskunde*, a society of Jewish folkloristics founded in 1898 and led by the Hamburg rabbi Max Grunwald, played a pioneering role in the scholarly research of Jewish folk traditions around the globe. Its journal, the *Mitteilungen fuer Juedische Volkskunde*, published the most important findings in the field for almost three decades (1898–1929). In Russia, an expedition to the traditional communities of the Ukraine led by the playwright and folklorist An-Ski from 1912 to 1915 was the culmination of a long search for the remnants of rural Jewish life.

Jewish demographers created their own institutional framework when they established an Office for Statistics among the Jews in Berlin, which from 1904 published its own journal and was closely related to the burgeoning interest in Jewish sociology best expressed in the pioneering works of Arthur Ruppin. Most of these endeavors were clearly related to the 'Jewish Renaissance' propounded by the emerging Zionist movement. Zionist scholars reacted against what they alleged to be the orientation of nineteenth-century Jewish scholars exclusively to the Jewish past and broadened their interest to include contemporary issues within their research.

After the establishment of the Hebrew University in 1925, scholars with a Zionist outlook formed a vaguely connected group of Jewish historians, sometimes referred to as the 'Jerusalem School,' emphasizing the centrality of Palestine in the course of Jewish history. Their most outspoken representative was the later Israeli minister of education Benzion Dinur (Dünaburg), who in 1935 helped to establish the most important Hebrew-language historical journal, *Zion* (an earlier series of the journal had been aborted), together with the first professor of Jewish history at the Hebrew University in Jerusalem, German-born Yitshak (Fritz) Baer.

From the very early works of Zionist historiography, usually written by Zionist activists who were not professional historians (such as Nahum Sokolov and Adolf Böhm), a teleological view of history leading to the return to the Jewish homeland is clearly visible. At the same time, the Jewish past in the Land of Israel is placed in the center of interest at the expense of diaspora history. In contrast to most historians, Benzion Dinur thus commences the period of exile not with the destruction of the Second Temple in the first-century CE but much later, with the Muslim conquest of Jerusalem in the seventh-century CE. In a similar vein, he concludes the period of diaspora domination as early as 1700, when a group of pious Jews from Eastern Europe made their way to Palestine. In his *History of Zionism, 1600–1918* (1919) Nahum Sokolov anachronistically dates the Zionist movement back to the seventeenth century.

While Dubnow's conception of Jewish history represented a corporate view of Jewish diaspora life and the Jerusalem School emphasized the centrality of Palestine, the last historian who attempted a single-handed multivolume universal Jewish history, Salo Wittmayer Baron, stood for a more positive approach to individuals' diaspora experience. His 16-volume *A Social and Religious History of the Jews* (1952–76) reaches only the year 1650. His more affirmative view of the diaspora was certainly influenced by his own experience of spending most of his life in the United States, where he was appointed in 1930 as the first professor of Jewish history at a Western university, Columbia University in New York.

Jewish historiography has thus to a large extent been determined by the political context, the geographical situation, and the ideological and religious outlook of the respective historians. Since all these factors differed considerably in the Jewish world, the outcome was a highly diverse array of Jewish historical writing – even though, at least up until the second half of the twentieth century, most of those who undertook to write Jewish history were themselves Jews. The destruction of European Jewry, the establishment of the State of Israel, and the flourishing of Jewish studies in the United States led to an entirely new situation for Jewish historiography after World War II.

Twentieth-Century Trends

With the exception of Baron no single historian embarked on a monumental Jewish history project in the tradition of Jost, Graetz, and Dubnow in the course of the twentieth century. There were a few attempts to publish Jewish world histories as collective enterprises, as early as the ambitious plan developed by Ismar Elbogen of the Berlin *Hochschule für die Wissenschaft des Judentums* in the late 1920s; these, however, fell victim to the economic and political crisis of the early 1930s. More successful, but for financial reasons also never completed, was a multi-volume project initiated by scholars of the Hebrew University.

After World War II Jewish history became more and more a field for specialists who would concentrate on limited areas and periods. This tendency toward specialization was also the result of an increasing professionalization and acceptance of the subject as a distinct field to be taught at university level, with new centers emerging in Israel and the United States. Until the 1960s, the teaching of Jewish history in the US had been restricted to a handful of universities and a few institutions with a strong emphasis on Jewish education, such as rabbinical seminaries and Jewish teachers' colleges or the Jewish-sponsored Brandeis University. With the establishment of new Israeli universities the singular role of the Hebrew University in Israel diminished, while a true explosion in the field of Jewish studies enabled a similar diversification at American universities beginning in the 1970s. European institutes of higher learning followed to a more modest degree in the 1980s and 1990s. By 2000, Jewish history had become a highly specialized and diversified field represented at almost every major institution of higher learning in the United States and many European countries.

This development led to a more complex picture of Jewish historiography. The personalities and political programs of individual historians and the formations of more or less clearly defined schools no longer dominate the professional scene. Zionist historians who had fought an ideological battle over Jewish history, mostly against their non-Zionist Jewish colleagues, up to the 1930s were essentially left without opponents as a result of the Holocaust and the destruction of alternative Jewish options, such as diaspora nationalism and Bundist socialism. In the State of Israel, apologetical historiography (now mainly directed against Arab and British versions of the history of Palestine) did not entirely disappear but was more and more balanced by the emergence of a younger generation of historians, by whom the existence of the Jewish state was already taken for granted. This resulted in a gradual break with long-accepted notions rather than an intellectual revolution as more and more historians and sociologists of a new postwar generation attacked the older schools of Zionist historiography in the 1980s and 1990s.

While mainly questioning issues of Israeli history or prehistory, such as the role of the Zionist leadership in Palestine with respect to attempts at the rescue of European Jewry in World War II, and the accepted notions surrounding the first Arab–Israeli War in 1948–49, the views of the so-called 'New Historians' also challenge more general Zionist notions of Jewish history. In a highly politicized context the two sides of the 'Israeli Historians' Debate' stand not only for different views on Jewish or Israeli history but mark also different paths for Israel's political future. In contrast to Israeli historiography, which has been dominated by a classical left-liberal Zionist intellectual establishment, the 'new' historians and sociologists often represent a post-Zionist stance and as such constitute a major challenge to the master narrative of Jewish and Zionist history.

In the last decade of the twentieth century, Jewish historiography moved from its early Germanocentric and later Eurocentric emphases to a stronger integration of widely neglected communities. In particular, the histories of Jews in the Muslim world have slowly been integrated into the general picture. This is partly the result of the emergence, mainly in Israel, of historians of oriental background, but also of the establishment of a few academic positions concentrating on Sephardic Jewry at American universities. It was in the United States, where the history of Jewish women emerged as a subfield of gender studies, with significant publications in the 1990s about Jewish women in Imperial Germany and in the United States, and a historical encyclopedia of Jewish women. New trends among general historians, including postmodern ideologies, have also left their imprint on Jewish historiography.

Europe, too, witnessed a revival of Jewish historical scholarship in the 1990s. The larger Jewish communities of Western Europe, France and Britain, emerged as new centers of Jewish scholarship, although most of the historical studies there deal with their respective national Jewish histories. Central and East European institutions relating to the Jewish past were destroyed by the Nazi onslaught, together with those who participated in them. Ironically, the Nazis entered into much more systematic Jewish historical research than non-Jewish scholars had done in the Imperial and Weimar periods; they established research institutes for Jewish history in Munich, Frankfurt, and Berlin. Studies undertaken there, however, were preoccupied not with Jewish history per se, but with the so-called 'Jewish Question' and sought to justify Nazi discrimination and persecution against the Jews from a pseudo-scholarly perspective.

It is interesting to note, though, that some of the research begun there – especially on the Court Jews and on Sephardic Jews in Hamburg – was published with only slight alterations in the 1950s. More modern Jewish historical research in Germany developed as a result of the establishment of the first Jewish Studies institutes and of the emergence, starting in the 1960s, of a new interest in German–Jewish history among modern German historians. While Jewish history is mainly written by Jewish historians and for a Jewish public both in prewar Europe and in today's Israel and America, the postwar European situation is different. With only small Jewish communities, research is mainly undertaken by non-Jewish scholars, and mainly read by non-Jews.

Major Issues in Modern Jewish Historiography

Perhaps even more than scholars of other particular histories Jewish historians have been divided by a set of fundamental questions. First of all, there is the question of whether Jewish history exists at all or whether the history of Jews merely constitutes a chapter within the histories of the nations among whom they lived. If Jewish history exists in a broader sense, the question arises whether it is the history of a people or a purely

religious history – or whether at some point the national history turned into a religious history. Finally, an additional problem of Jewish history is its theological interpretation by some modern Jewish and non-Jewish thinkers, which makes Jewish history appear as though it escapes the regular laws of history. This last argument is strengthened by the conspicuous role of anti-Semitism throughout the centuries and the claim of the uniqueness of the Holocaust, which brought about the near destruction of European Jewry.

The doubt whether Jewish history exists per se seems rather superfluous in an age of expanding programs in Jewish history and within the framework of an article discussing the rich tradition of Jewish history writing. Nevertheless, one can still find this opinion among some modern historians. It is based partly on the Christian theological notion that active Jewish history came to an end with the early stages of Christianity, and partly on the secular claim that Jews lack a clearly defined political history based on a common territory. In the nineteenth century such a belief was inherent in the theological characterization of ancient Judaism as 'late Judaism,' while on the other side of the spectrum of historical outlook Marxist historians denied a separate course of Jewish history on the basis that Jews not only lacked a territory of their own but also a division into classes of society in the modern period. Nowadays, the questioning of the existence of a Jewish history throughout the ages is less explicit. More indirectly, however, such a view can be detected in the almost exclusive interest of many historians in the histories of the Jews in their respective societies. Thus, German Jewry or French Jewry is often seen as isolated entities, as part of German or French history, but without any connection to the developments of other Jewish communities. On the other hand, there has been a contrary tendency among some Zionist historians to neglect the respective contexts of Jewish histories and view them only through a national Jewish perspective.

If we accept the notion that Jewish history exists as such and is stretched over numerous centuries and over five different continents, the question remains whether it is primarily a national or a religious history. The pioneers of Jewish historiography in nineteenth-century Germany often preferred to depict Jewish history in terms of the development of a religious idea (the purest form of monotheism) and its bearers. This concept was in line with the demands of Emancipation that Jews integrate into European societies as citizens who only differed in their religious creed. Thus, the notion of a Jewish people was characterized as an ancient idea long left behind. With diaspora nationalists like Simon Dubnow and the emerging Zionist school the national element was restored to Jewish historiography. Both schools emphasized that Jews never ceased to constitute a nation even if they lacked their own territory. Dubnow went so far to stress that a nation that had overcome the need for its own territory constituted the highest and most sophisticated form of nationalism. Zionists, on the other hand, emphasized the unbroken longing for return to the Land of Israel and the use of the Hebrew language over the centuries.

A final peculiarity of Jewish history is the theological weight it enjoyed especially in a Christian context. In medieval and early modern Europe Jews often constituted the only non-Christian minority. The traditional position Jews enjoyed in Christian Europe was a mixture of protection and discrimination. Jews served as the classical 'other,' being protected because they bore testimony to Jesus Christ while at the same time discriminated against as Christ-killers and as unbelievers. On the basis of this peculiar role, many Christian authors of postbiblical Jewish history regarded this particular history as outside the course of 'regular' history. They claimed accordingly that it should be judged by theological rather than historical criteria. On the Jewish side, the most prominent proponent of a Jewish '*Sonderweg*' theory, albeit with a different background, was the philosopher Franz Rosenzweig, a student of Friedrich Meinecke, who stressed in his main work *The Star of Redemption* (1921) that general rules of history do not apply to Jewish history. More recently, the discussion of the uniqueness of the Holocaust has brought the question of a peculiar course of Jewish history to a different level. In this view, it is not Jewish history that is incomparable to other histories, but rather the persecution of the Jews, which culminated in the attempt of their total annihilation.

In his programmatic and provocative essay, 'Ghetto and emancipation' (1928), the historian Salo Baron fiercely objected to what he called a 'lachrymose version of Jewish history.' At the beginning of the twenty-first century, and half a century after the greatest tragedy of the Jewish people, numerous voices once again express their reservations against a one-sided picture of Jewish history reduced to a history of persecution. While they stress the importance of research on anti-Semitism and the Holocaust, they also warn not to read Jewish history backwards starting from its most tragic chapter.

See also: Historical Thought and Historiography: Christian Tradition; Historical Thought and Historiography: Current Trends; Historical Thought and Historiography: Islamic Tradition; Jewish Diaspora; Judaism and Gender; Judaism; Religion, History of.

Bibliography

Aschheim, S.E., 2012. At the Edges of Liberalism: Junctions of European, German, and Jewish History.

Baskin, J.R., Seeskin, K., 2010. The Cambridge Guide to Jewish History, Religion, and Culture. Cambridge University Press, Cambridge.

Botticini, M., Eckstein, Z., 2012. The Chosen Few: How Education Shaped Jewish History, 70–1492. Princeton University Press, Princeton.

Frank, D.H., Leaman, O., 1997. History of Jewish Philosophy. Routledge, London.

Funkenstein, A., 1993. Perceptions of Jewish History. Berkeley University Press of California, Berkeley, CA.

Gurock, J.S., American Jewish Historical Society, 1998. American Jewish History: A Eight-volume Series. Routledge, New York.

Halevi-Wise, Y. (Ed.), 2012. Sephardism: Spanish Jewish History and the Modern Literary Imagination. Stanford University Press, Stanford, California.

Liberles, R., 1995. Salo Wittmayer Baron: Architect of Jewish History. New York University Press, New York.

Meyer, M.A., 1974. Ideas of Jewish History. Behrman House, New York.

Myers, D., 1995. Reinventing the Jewish Past: European Jewish Intellectuals and the Zionist Return to History. Oxford University Press, New York.

Myers, D.N., Ruderman, D.B. (Eds.), 1998. The Jewish Past Revisited, Reflections on Modern Jewish Historians. Yale University Press, New Haven, CT.

Robinson, I., 2008. Translating a Tradition: Studies in American Jewish History. Academic Studies Press, Boston.

Schorsch, I., 1994. From Text to Context: The Turn to History in Modern Judaism. University Press of New England, Hanover, NH.

Yerushalmi, Y.H., 1989. Zakhor: Jewish History and Jewish Memory, second ed. Schocken, New York.

Historical Thought and Historiography: Modern History (Since the 18th Century) – the West

Georg Gerson Iggers, University at Buffalo (State University of New York), Buffalo, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article offers a brief survey of the main currents of historical thought and writing in the West since the eighteenth century. It begins with the emergence of a secular outlook in Enlightenment approaches to history. It then proceeds to the transformation of historical studies in the nineteenth century marked by professionalization. Strongly influenced by the work of Leopold von Ranke, historians saw their discipline as a science (Wissenschaft), scientific in the sense of reliance on the critical examination of the sources, distancing itself from nonprofessional amateur historians. Its approach to historical presentation in the form of narrative focusing largely on political events and political leaders differed sharply from broader generalizing approaches as represented in the philosophies of history in very different forms by Georg Friedrich Hegel, Karl Marx, and Auguste Comte. Although Ranke maintained a European perspective, his followers largely saw history in the service of the national state and nationalism. The Rankean paradigm, which became the basis of history as an academic discipline worldwide, was challenged in the first part of the twentieth century by new social science-oriented approaches that moved from narrative to the analysis of social structures and processes of change. They shared with the older paradigm the conviction that there was an objective historical past that could be known and that history had direction and coherence culminating in the modern world. These two interrelated conceptions were challenged in the last third of the twentieth century and the beginning of the twenty-first century by social scientists and particularly humanists who questioned the possibility of scientific objectivity and called for the replacement of the Eurocentrism of much of professional historiography up to that point by a global perspective that gives proper attention to gender and to ethnic and social diversities.

Reorientation of Historical Thought in the Eighteenth Century: The Emergence of a Secular Outlook

There is good reason for beginning this article on Western historical thought and historical writing in the modern age with the eighteenth century. The Western tradition is much older, going back at least to the Hellenic Age. There is much continuity between the older traditions, particularly those of the humanistic historiography of the Renaissance and historical writing in the eighteenth century. However, there are also profound changes in the conceptions of history and of historical knowledge and in the institutional context in which history is written that mark a sharp break with previous periods.

Henri Berr and Lucien Febvre in the article 'History' in the first edition of the *Encyclopedia of the Social Sciences* in 1932 identified two characteristics of the modern conception of history as basic to the modern historical outlook and practice: the commitment to establishing the 'truth' and to placing historical accounts within 'universal history.' A third is the transformation of historical studies into a professional discipline. The older ancient and humanistic historiography, they argued, saw history as a form of rhetoric in the service of politics and morality and as an art form. 'History as it is conceived today,' they noted, 'may blossom into art, may be crowned with philosophy; but it is primarily and necessarily the solid establishment of facts' (Berr, p. 358), however difficult this may be.

This formulation presupposes a sharp distinction between history as literature and as a scientific discipline, even if its approach to science takes into account the elements of contingency and particularity that distinguish history from the generalizing natural sciences, but it assumes that historical accounts based on a solid factual basis narrate a coherent development that reflects historical reality. There is a good deal to be said for this description of the transformation that historical studies underwent in the eighteenth century and particularly in the nineteenth century, but especially from the perspective of today it seems one-sided and incomplete.

On two levels, historical thought and historiographical practice in the eighteenth century moved toward greater historical realism. On the one hand, an increasingly secular outlook insisted on the exclusion of the supernatural and held that historical explanation must not violate a common understanding of natural reality. This secular approach was relatively new in the eighteenth century. On the other hand, it coexisted often with a residual religiosity, as it still did even in the nineteenth century with a conception of a hidden divine plan, as in Leopold von Ranke's trust in the 'hand of God' (2011, p. 87), or in a less hidden form in G. W. F. Hegel's dialectic. Thus Enlightenment historians such as J. C. Gatterer and A. L. Schloezer did not challenge the biblical chronology or the biblical account, although for the postbiblical period, or for the nonHebraic ancient peoples, they insisted on an approach in accord with natural reality. Apart from the great works of synthesis in the eighteenth century, David Hume, Catharine Macaulay, Montesquieu, Voltaire, Edward Gibbon, G. T. Raynal, and Adam Ferguson, there exists a tradition of erudition that introduced standards of source criticism, which went back to the humanist scholarship of the Renaissance and Reformation and to the erudition of seventeenth-century historical scholarship. The belief that great philosophic works of synthesis are not enough, and that scholarly history cannot be written until the authenticity and veracity of the sources on which it rests have been established, expressed itself in the great

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62033-0

methodical collection of sources begun in the seventeenth century with the Jesuit *Acta Sanctorum*, published after 1643, to provide a firm basis for ecclesiastical history, and with Ludovico Muratori's efforts in the *Scriptores rerum italicarum* (1723–51) to lay the foundation for an Italian national history. The German scholars at the universities, first J. L. Mosheim and J. S. Semler at Halle and then J. A. Ernesti and J. O. Michaelis at Göttingen, established a tradition of historical exegesis that laid the foundations for historical hermeneutics and the new discipline of historical philology that emerged in the latter part of the eighteenth century and became the key elements of the critical method of B. G. Niebuhr and Ranke in the nineteenth century.

A key concept of eighteenth century thought was the linear character of history. History was no longer understood as a cumulation of events but as a continuous story possessing inner coherence. Reinhart Koselleck has suggested that between 1750 and 1850 the concept 'history' (*Geschichte*) replaced that of 'histories' (*Geschichten*). Each of the great historical works of the eighteenth century followed a continuous, coherent story whether in Gibbon's *Decline and Fall of the Roman Empire*, Hume's *History of England*, or Voltaire's *Essai sur les moeurs*.

The conception of world history that arose with the Enlightenment constituted not only an extension of the historical outlook but at the same time also a narrowing. This conception dominated almost all of Western historical thought and historical writing until well into the second half of the twentieth century. The history of the world was identified with the history of civilization and the latter rather narrowly with that of the West. Two variations emerged in the conception of what constituted world history, one identified with Condorcet, centering on an idea of progress on a world scale, and the other identified with Herder, which stressed the diversity of peoples and cultures. On closer examination, these two conceptions shared many characteristics, specifically in their stress on development.

The Enlightenment has too many aspects to permit one to reduce it to a simple formula. Two ideas were central to it, but neither was free of contradictions. The first assumed the essential equality of all human beings in all cultures; the second maintained that only human beings who had attained a high degree of culture and education deserved to be free and equal citizens. The latter view distinguished between societies that were in the vanguard of civilization, understood as Western civilization, and less developed peoples. If G. W. Leibniz and Voltaire had still regarded China as a civilization equal and possibly superior to that of the West, Christoph Meiners considered the yellow peoples as inferior to Europeans, and the Blacks as barely different from animals. The notion emerged of a distinctive Western civilization with roots in classical Greece. J. J. Winckelmann's rediscovery of Greek art in 1764 contributed to an idealization of Greek standards of beauty and personal culture, which were adopted as models for a modern culture distinguishing the West from other cultures and giving it its superior standing. Meiners' scientific racism represented an extreme position, but the idea of the unique position of the West and the relative inferiority of all non-Western cultures and peoples became a central assumption in Western historical consciousness.

Nineteenth-Century Conceptions of History and Historical Science: between Romantic Historism and Scientistic Positivism

This outlook was dominant well into the twentieth century. It reflected the changes that had taken place in the world in political and military strength, economic power, and scientific and technological development. The two divergent historical outlooks in the first half of the nineteenth century, one represented by Hegel's dialectical philosophy of history and the other by Ranke's historism, shared this outlook despite their philosophic differences. Hegel's perception of the world as an ongoing dialectical process by which, through progressive contradictions, the absolute mind, reason, fulfills itself and takes on concrete forms in the institutional life of the modern European world, and specifically in the Prussian political and social order after the French Revolution and Napoleon, reflects its origins in specifically German political and philosophical traditions. However, it has parallels in less involuted intellectual currents of thought outside of Germany, from Turgot, Condorcet, and Thomas Paine in the late eighteenth century to Henri de Saint Simon, Auguste Comte, J. S. Mill, Thomas H. Buckle, Herbert Spencer, and Benedetto Croce in the nineteenth century. For all of them world history is the history of the European world, a world that for some may include the United States but for all excludes the Slavic peoples. For Hegel only the West has a history. China, India, and Persia may have been the earliest civilizations, but they have remained stagnant. According to Hegel, Black Africans are incapable of cultural creativity. Karl Marx with his conception of static Asiatic modes of production took over much of Hegel's conception of the inferiority of non-Western societies, and he and Friedrich Engels used this to justify Western imperialism as the only way to bring the non-Western world into the realm of modern civilization.

Theoretically, Ranke's conception of the historical world represented the antithesis of Hegel's. Ranke rejected the idea of progress as well as the schematization of historical development as it occurred in Hegel's philosophy of history. History for him deals with the particular and the contingent and avoids generalizations and determinism. In fact, however, Ranke saw history as a linear process at the culmination of which stood the postNapoleonic Western and Central European world. Ranke proclaimed that "every age is equally immediate to God" (2011, p. 21) but at the same time denied that China or India had a true history. Except for his early work on the Serbs, his writings dealt with the countries he defined early on as the Latin and Germanic. His incomplete *World History* is a history of the West.

Whatever the philosophical and political differences, nineteenth-century thinkers as well as the educated public generally turned to history as the key to understanding things human that replaced traditional philosophy. The charges, often repeated, that the eighteenth century, particularly the Enlightenment, ignored history, generalized, and moralized, and did not recognize individuality and uniqueness of historical contexts are disproved when we look at the great historical works of the eighteenth century, which we have already mentioned. Despite the impact of romanticism in the first part of the nineteenth century, with its stress on imagination and

poetry, a cult of science dominated historical study and writing, which was not necessarily in conflict with a romantic outlook. This search for a scientific approach to history took two different directions: one that has frequently been called positivism and the other historism, neither of which can easily be reduced to a common denominator. The first of these orientations took the natural sciences as its model. Building on Turgot and Condorcet, Auguste Comte sought to formulate laws of development by which societies moved from simpler to more complex forms of organization. The driving force was the transformation of knowledge from theological and then metaphysical stages to positive science that would provide the means for the scientific organization of society. However, positivism remained largely grand theory with little empirical application to historical inquiry. More significant was Karl Marx's approach to history. Marx saw the driving force of history in the lawful development of economic forces, which constituted the basis of historical transformation, and the revolt of this basis against the social, political, and intellectual superstructure. In the course of this conflict the contradictions within the social order, which involved inequality, would be overcome in a classless society. Marx thus intended to replace Hegel's idealist dialectic with a materialistic one. If positivism both in its Comtean and Marxist forms in its search for a scientific history envisaged a utopia at the end of history, the main currents of historical scholarship sought to avoid a speculative approach to history but nevertheless did not free themselves from metaphysical and even theological presuppositions that they refused to acknowledge. Ranke may perhaps serve as the most influential representative of this orientation. He too was committed to a strictly scientific approach to history. For him, however, this approach must recognize that general laws have no application to history that deals with the particular and the contingent. For him the task of the historian was to recapture the past as 'it actually was' (*wie es eigentlich gewesen*), free of any bias, requiring the application of philological methods to the critical analysis of the sources as F. A. Wolf and Niebuhr had applied them. However, for him history was more than the reconstruction of the past. It involved interpretation and synthesis, as it did for J. G. Droysen. Ranke was convinced that the particular event contained the general, and that the general became apparent to the researcher who immersed himself in the sources, because – and here we return from the realm of strict science to that of metaphysics – the historian can observe the general tendencies that manifest themselves in history, tendencies that do not represent God's direct intervention in the world, but nevertheless reflected divine purpose.

Professionalization of Historical Studies

The stress on the scientific character of historical studies led to the formation of a professional discipline of history. Professionalization was not restricted to history but was part of a process in which scholarly studies in all fields, the liberal arts, as well as the natural sciences, were institutionalized at centers of higher learning, primarily the universities and research institutes. The German model was imitated elsewhere, relatively early in France with the introduction of the German seminar method at the Ecole Pratique des Hautes Etudes in Paris in 1868 and in the United States with the introduction of a research-oriented graduate program at The Johns Hopkins University in Baltimore in the 1870s leading to a PhD. With some delay the German model was adopted in Japan, Great Britain, the Netherlands, Italy, Spain, Russia, Scandinavia, Greece, Romania, Latin America, and elsewhere, each time, it must be stressed, with major modifications that reflected different scholarly traditions. Thus, the sharp distinction between the cultural and the natural sciences that dominated German thought played less of a role elsewhere. A basic tenet of professional historians was the total objectivity or what Ranke called the 'impartiality' of the historian. In fact, however, the new discipline was highly politicized everywhere. This intensified interest in archival research went hand in hand with the wave of nationalism that gripped all Western countries and also modernizing non-Western countries such as Japan. The broad cosmopolitan view of many historians in the eighteenth century gave way to a new focus on national history. Ranke still wrote the histories of a variety of European states, Serbia, England, France, Prussia, and also the Reformation in Germany and the Papacy. Later historians, who took him as a model for their own scholarship, in most cases concentrated on their own nation.

However, the development of modern historical scholarship was not a German monopoly. The opening of archives and the collection of documents, primarily those of the Middle Ages, became a prerequisite for scholarly national history. In Paris the Ecole des Chartes was founded in 1821 to train scholars in the critical use of documents. Muratori's collection of Italian documents in the eighteenth century has already been mentioned; this was followed by the *Monumenta Germaniae historica* launched in Germany in 1819 and by similar undertakings in France, Great Britain, Spain, Poland, Bohemia, Greece, and elsewhere. The Rankean belief that immersion into the sources guarantees an impartial view of the interrelation among historical events turned out to be an illusion, something that was openly recognized by the scholars of the Prussian school such as Heinrich von Sybel, J. G. Droysen, and Heinrich von Treitschke who frankly saw historical scholarship as a tool for the legitimation and propagation of the German national cause under the Hohenzollern rule. Moreover, the process by which scholars were recruited guaranteed a high degree of homogeneity in social composition and political outlook.

Professionalization was not only to guarantee objectivity but also to provide a scientific or scholarly discourse distinct from the literature. Ranke had stressed the scientific character of history, but at the same time he also recognized that history combines science with art. The great prestige that history enjoyed everywhere in the nineteenth century – a number of the leading French statesmen from François Guizot to Alexis de Tocqueville, Alphonse de Lamartine, Adolphe Thiers, and Jean Jaurès were important historians – rested on the literary quality of their histories. History could not escape being literature, and imaginative literature at that. In fact there existed a close parallel between the great histories written by professional historians and the great realistic novels of the nineteenth century – Stendhal, Honoré de Balzac, Charles Dickens, Gustave Flaubert, Gottfried Keller, Gustav Freytag, Ivan S. Turgenev,

Fyodor Dostoyevski, and even Thomas Mann's *Buddenbrooks* (1902). They all constructed a plot that followed a temporal line and involved individual subjects as agents. The great historians of the century, Jules Michelet, Denis Fustel de Coulanges, Thomas B. Macaulay, and Theodor Mommsen, who in 1902 received the Nobel Prize for literature, were read as great literature, as was Ranke himself. Michelet is a striking example of the crossing of borders between critical scholarship and imaginative literature, the exhaustive mining of the archives for the purpose of creating a national memory and identity.

However, the role professionalization played in nineteenth-century historical writing should not be overestimated as it often has been. A great deal of history did not follow the dominant paradigm of narrative national history such as three of the most important and lasting works of the midcentury, Fustel de Coulanges's *La Cité antique* (*The Ancient City*), de Tocqueville's *L'Ancien régime et la Révolution Française* (*The Old Regime and the French Revolution*), and Jacob Burckhardt's *Die Kultur der Renaissance in Italien* (*The Civilization of the Renaissance in Italy*). Moreover, many leading writers of history such as Thomas Carlyle, Macaulay, John Green, George Bancroft, Lamartine, Louis Blanc, and, of course Tocqueville, were not professional historians.

The great literary histories of the eighteenth century, for example, those of Hume, Gibbon, and Voltaire, were intended to entertain, educate, and edify a broad educated public. Professional historians wrote often for specialists. In fact the division between literary and professional history and their audiences was not as deep as it may appear. Certainly Hume, Gibbon, and also William Robertson were serious scholars, and the great professional historians of the nineteenth century, such as Ranke, Fustel de Coulanges, Theodor Mommsen, and Lord Acton were read avidly by a broad public for whom history had become the key science replacing philosophy as a guide to understanding the human world.

Twentieth-Century Reorientations of Historical Thought and Scholarship: The Turn to the Social Sciences

Outside the discipline proper, critical methods were applied to many areas: the history of religion (F. D. Strauss, Ernest Renan), the history of law, the history of literature and art, and other fields. While the historians proper neglected the study of society and economics, an important group of economic historians, the so-called Historical School of National Economy in Germany (Bruno Hildebrand, Karl Knies, Wilhelm Roscher, and Gustav Schmoller) pursued a middle course between an empirical approach, seeking generalizations of economic behavior and development, reminiscent of positivism, and an intense regard for the particular historical setting of culture and society in which economic behavior occurred, which reflected historicist principles. The late nineteenth century and the turn to the twentieth century witnessed profound reexaminations of the historical outlook that had been dominant until then. On the one hand, there was a continued, even a heightened faith in science, technology, and the progress of Western civilization that largely survived the trauma of World War I. On the other hand, there was an increasing sense of discomfort with the quality of modern civilization. Darwinism, or rather Social Darwinism, played a fundamental role in both these outlooks. The world was seen as a struggle for the survival of the fittest that could forecast the progressive victory of the white races over the rest of the world, which for Oswald Spengler was a desperate struggle for survival. At the turn to the twentieth century, some Europeans and Americans foresaw progress in terms of growing tolerance and democracy, while others preached racism and ultranationalism. As far as historical thought was concerned, both Burckhardt and Friedrich Nietzsche voiced their discomfort with a modern industrial society with its concomitant social and cultural aspects, a mass civilization moving to commercialism and democracy, and extended this dismay to a critique of the modern faith in science. Nietzsche, Wilhelm Dilthey, Henri Bergson, and Georges Sorel in differing ways called for a restoration of the vital against the mechanical aspects of life and culture. Established patterns of historical scholarship, as practiced by the historical profession, were called into question simultaneously at the end of the nineteenth century from Russia and Poland to North America. Generally, however, the critics of the historical profession did not criticize the latter not because it was too scientific but rather because it was insufficiently so and had not taken into consideration the tremendous changes in modern society and civilization in an industrial and democratic age.

The most spectacular confrontation with the established profession occurred in Germany with the so-called *Methodenstreit*, occasioned by the publication of the first volume of Karl Lamprecht's *Deutsche Geschichte* in 1891. Lamprecht criticized his colleagues for their narrow focus on the state and their neglect of broader cultural and social aspects of history. He also accused them of working with an outdated concept of science and historical science that restricted itself to events and personalities rather than taking into consideration the structural context of historical change. The German professoriate, reflecting the peculiar semiautocratic character of the Prussian German state, vociferously rejected Lamprecht's initiative and saw in it a threat to the established social and political order. Lamprecht's attempt to formulate laws of historical development was rejected internationally, but his call for a broadening of historical perspective to encompass society, economics, and culture was generally welcomed outside of Germany. Historians in Russia (Paul Vinogradoff, Vasiliev Kliuchevskii, and Paul Miljukov), Romania (Nikolae Iorga), Poland (Jan Bujak and Franticzek Rutkowski), Norway (Halvdan Koht), Belgium (Henri Pirenne), as well as the circle around the journal *Revue de synthèse historique* founded by the philosopher Henri Berr in 1900 as a forum for international discussion, and the American 'New Historians' (F. J. Turner, J. H. Robinson, Charles Beard, Carl Becker, H. E. Barnes) called for an expansion of the scope of history and a closer cooperation with the empirical social sciences. Two variations of a social science approach to history need to be mentioned. The first is Max Weber's historical sociology that moves between the historistic emphasis on the recognition of the uniqueness of cultures and a historical social science working with clearly defined concepts and a rigorous logic of inquiry. Weber's conception of social causation and historical development was in some ways indebted to Marx, but unlike Marx emphasized the

interrelation between culture, including religion, and economics, and placed much greater weight on contingency. Weber recognized that societies constituted meaningful relationships that required an interpretive sociology but avoided the sharp distinction that idealistic philosophers such as Wilhelm Windelband, Heinrich Rickert, and Wilhelm Dilthey made between the generalizing natural sciences and the cultural sciences (*Geisteswissenschaften*) that favored empathetic understanding (*Verstehen*).

A second approach to a history taking into account the social sciences was offered by the so-called *Annales* School. It is difficult to reduce the *Annales* to a common denominator. The journal by this name, launched in France in 1929 by Lucien Febvre and Marc Bloch with the support of Henri Pirenne, continued initiatives begun early in the century in Henri Berr's already mentioned *Revue de synthèse historique*. A similar project was initiated in Poland in the journal *Roczniki Dziejow Spolecznych i Gospodarczych*, the editors of which, Rutkowski and Bujak, were in close contact with the *Annales*. In the place of the concentration of the established historiography on events and great personalities, the *Annales* explored impersonal forces and structures.

The 'human geography' of Paul Vidal de la Blache played an important role in the early works of Lucien Febvre and Marc Bloch. Human history had to be understood in the concrete context of space and time. Unlike the German Friedrich Ratzel, who saw soil and race (*Volk*) as determining factors in history, Vidal de la Blache, like Febvre, Bloch, and later Fernand Braudel, recognized the role that human agency and culture played in the formation of the geographic setting. The historians of the *Annales* attempted to pursue a 'total history,' generally limited to a restricted region, in which all aspects, geographic, economic, political, religious, and demographic, were to be included. Quite consciously the dividing line between the established disciplines were overcome, and in the *Maison des Sciences de l'Homme*, opened in Paris in 1968, economists, sociologists, anthropologists, linguists, psychoanalysts, and literary critics were housed with history as the integrating science.

In other ways too, the *Annales* represented a radical break with older historical conceptions. The nation no longer represented the basic unit of history; it was replaced by the region or by a broad comparative approach that extended beyond Europe (Braudel and Pierre Vilar). The concept of a linear time, so basic to all forms of nineteenth century historical thought, and with it the idea of progress, were abandoned, and a historical epoch viewed in its own terms, not as part of a historical process. Burckhardt had already done something similar in 1860 in *Die Kultur der Renaissance in Italien* and had worked with a very broad concept of culture, as had also Jan Huizinga in *Herfsttij der Middeleuwen* (1919), but unlike the *Annales* historians they focused heavily on elites. In contrast the *Annales* dealt with the population at large, but ran the danger that at times their history became impersonal and void of concrete human beings. In retrospect, the racist historiography of German National Socialism of the period 1933–45, although it built on widespread Social Darwinist ideas and ultranationalism, remained an interlude, which even in Germany had limited influence. The Marxism–Leninism of the Soviet Union and its dependent states in its dogmatic forms, in many ways a perversion of Marxism, similarly represents an interlude. As early as 1956, historians in Poland and Hungary, and until the freeze after the Soviet occupation in Czechoslovakia, and even in the Soviet Union (Aaron Gurevich), participated in the broader historical discussions of the international community.

New Orientations after 1945: The Turn to a Cultural Dimension

A subsequent article, 'Historiography and Historical Thought: Current Trends,' will deal with trends in historical thought and study in the second half of the twentieth century. The victory of the Allies in World War II and the dominance of the United States in the postwar world revived older beliefs in progress and science. The 1950s and 1960s witnessed a turn to the quantitative social sciences in historical studies. The years since the late 1960s have seen a reorientation in historical outlook in the Western world generally, including to a lesser extent also Eastern Europe and Russia. The conflicts of the 1960s, the racial confrontation in the United States, the Vietnam War, the student revolts throughout Europe and North America, and the suppression of the Prague Spring all contributed to a change in outlook as did the fundamental technological changes that resulted in the transformation of an industrial society concentrated on production into an economy based on information technology, in which consumption and services play a key role and the old socioeconomic classes gave way to new ones. The confidence in science and economic growth was severely tested as a new generation became increasingly conscious of ecological concerns and called for the emancipation of subordinated segments of the population such as women and ethnic minorities at home and the peoples of the former colonies abroad. The very quality of the modern Western world, including its science, came under attack. All this had consequences for historical studies after the 1960s.

See also: Historicism; History and the Social Sciences; Positivism, Sociological.

Bibliography

Bann, S., 1984. The Clothing of Clio: A Study of the Representation of History in Nineteenth-Century Britain and France. Cambridge University Press, New York.
Barnes, H., 1962. A History of Historical Writing. Dover, New York.
Barraclough, G., 1978. Main Trends in History. Holmes & Meier, New York.
Bentley, M., 1999. Modern Historiography. Routledge, London.
Berr, H., Febvre, L., 1932. History. In: Encyclopedia of the Social Sciences. Macmillan, New York. vol. 7357–368.
Borguière, A., 2006. The Annales School: An Intellectual History. Cornell University Press, Itaca, NY.
Breisach, E., 1983. Historiography: Ancient, Medieval, and Modern. University of Chicago Press, Chicago, IL.
Budd, A. (Ed.), 2009. The Modern Historiography Reader: Western Sources. Routledge, London.
Burke, P., 1990. The French Historical Revolution: The Annales School, 1929–1989. Polity Press, Cambridge, UK.
Butterfield, E., 1955. Man on His Past: The Study of the History of Historical Scholarship. Cambridge University Press, Cambridge, UK.
Carbonell, C., 1981. L'Historiographie. Presse Universitaire de France, Paris.
Chickering, R., 1993. Karl Lamprecht: A German Academic Life (1856–1915). Humanities Press, Atlantic Highlands, NJ.

Den Boer, P., 1998. History as a Profession: The Study of History in France, 1815–1914. Princeton University Press, Princeton, NJ.
Evans, R., 2009. Cosmopolitan Islanders: British Historians and the European Continent. Cambridge University Press, Cambridge, UK.
Fueter, E., 1911. Geschichte der neueren Historiographie (History of Modern Historiography). Oldenbourg, Munich, Germany.
Gazi, E., 2000. Scientific History: The Greek Case in Comparative Perspective (1850–1920). Peter Lang, New York.
Gooch, G., 1913. History and Historians in the Nineteenth Century. Macmillan, London.
Higham, J., Gilbert, F., Krieger, L., 1965. History. Prentice Hall, Englewood Cliffs, NJ.
Iggers, G., 1968. The German Conception of History: The National Tradition of Historical Thought from Herder to the Present. Wesleyan University Press, Middletown, CT.
Iggers, G., 2005. Historiography in the Twentieth Century: From Scientific Objectivity to the Postmodern Challenge, second ed. Wesleyan University Press, Middletown, CT.
Iggers, G. (Ed.), 2011. Leopold von Ranke: The Theory and Practice of History. Routledge, London.
Iggers, G., Wang, E., Mukherjee, S., 2008. A Global History of Modern Historiography. Pearson/Longman, Harlow.
Kelley, D., 2006. Frontiers of History: Historical Inquiry in the Twentieth Century. Yale University Press, New Haven, CT.
Kelley, D., 2003. Fortunes of History: Historical Inquiry from Herder to Huizinga. Yale University Press, New Haven, CT.
Keylor, W., 1975. Academy and Community: The Foundation of the French Historical Profession. Harvard University Press, Cambridge, MA.
Koselleck, R., 1975. Geschichte, Historie. In: Brunner, O., Conze, W., Koselleck, R. (Eds.), Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland, vol. 2. Klett, Stuttgart, Germany, pp. 593–717.
Kramer, L., Maza, S. (Eds.), 2002. A Companion to Western Historical Thought. Oxford University Press, Oxford.
Novick, P., 1988. That Noble Dream: The Objectivity Question and the American Historical Profession. Cambridge University Press, Cambridge, UK.
Offenstadt, N., 2011. L'Historiographie. Presse Universitaire de France, Paris.
Macintyre, S., Maiguashca, J., Pók, A. (Eds.), 2011. The Oxford History of Historical Writing vol. 4. Oxford University Press, pp. 1800–1945.
Reill, P.H., 1975. The German Enlightenment and the Rise of Historicism. University of California Press, Berkeley, CA.
Ringer, F., 1969. The Decline of the German Mandarins: The German Academic Community 1890–1933. Harvard University Press, Cambridge, MA.
White, H., 1973. Metahistory: The Historical Imagination in Nineteenth-Century Europe. The Johns Hopkins University Press, Baltimore, MD.
Woolf, D. (Ed.), 1998, A Global Encyclopedia of Historical Writing, vol. 2. Garland Publishing, New York.
Woolf, D., 2011. A Global History of History. Cambridge University Press, Cambridge, UK.

Historical Thought and Historiography: South Asia

T Raychaudhuri, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 6804–6808, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article traces the main features of the South Asian historiographical tradition from ancient times to recent developments. It identifies the central concerns of ancient and medieval Indian historiography – their emphasis on dynastic and political history, and moves on to discuss in some detail the trends in historical writings in more recent times, especially since the World War II. It takes into account both the work of South Asian historians and the international scholarship on the history of South Asia. The impact of decolonization and the social sciences on the historiography of the region is emphasized.

This article addresses two interlinked subject matters, the history of South Asia to which scholars from all over the world have contributed, and historiographical developments in the region of South Asia itself. It briefly covers the historical tradition in premodern times and primarily focuses on developments since World War II.

Historiography of Ancient India

Historical Traditions in Ancient India

Compared to classical Europe, ancient India was rather deficient in historical traditions. The Hindu mythological texts called the *Puranas* (circa first to fourteenth century) contained genealogical lists of Indian dynasties. These lists, by way of narrating the achievements of the kings, also recorded rudimentary information about wars, conquests, and the rise and fall of empires. A very small number of works chronicle the history of individual kings and dynasties. Of these, the most famous is Kalhana's *Rajatarangini* (Stein, 1900), a history of the Kashmir princes. Incidentally, the author recognized one historical principle, namely, that the historian must try to be unbiased in his/her judgment, 'like a judge.'

The Buddhists and Jains developed a tradition of sacred history, especially hagiographies centered on the life stories of their founders. Some of the stories concerning the Buddha, like the account of his death (*Mahaparinivvana-sutta*, Bongard-Levin, 1986) bear the stamp of authenticity.

European Scholarship on Ancient India

In the nineteenth century, European scholars like Pargiter tried to reconstruct the dynastic history of ancient India using the Puranic lists, inscriptions, coins, archaeological remains, and foreign accounts of India like those of Alexander's historians and the Chinese Buddhist pilgrims. European Orientalists also tried to reconstruct the religious and social history of India, using religious texts and the *smritis*, normative writings laying down the rules by which the Brahminical Hindu was expected to live in matters, both ritualistic and secular (like inheritance and punishment for crimes). They also undertook the massive task of publishing critical editions of religious texts that laid the foundation for serious work on religious, cultural, and social history. A large number of German, French, and East European scholars have contributed to this enterprise.

Historiography of Ancient India: the Indian Contribution

A large body of Indian scholars started research on ancient Indian history following the lead of the European scholars. Their work was heavily focused on archaeological evidence, especially the study of inscriptions. The archaeological work, by Europeans and Indians, helped move back the known beginning of civilization in South Asia to about 3000 BC through the discovery of the Indus Valley Civilization. The work of the Indian Indologists also reflected the newfound nationalistic pride in the glories of ancient India. Some of it focused on traditions of democratic/oligarchic politics and developed statecraft to challenge the notion that these were European imports into India. However, the central concern of the earlier work was with the meticulous reconstruction of dynastic history, especially the correct chronology. Some scholars like R.K. Mukherji and R.C. Majumdar did research on the history of Indian cultural influence in Southeast Asia and invented the now discarded notion of Greater India.

'Orientalism': A Controversy

Edward Said's *Orientalism* (1995), which projects the view that Western scholarship on Asia projected the notion of an inferior 'other' of Europe and thus reflected the new relationships of power generated by imperialism, has had a seminal influence on the historical work on South Asia (since 1980) even though Said's primary point of reference was the Middle East. The controversy generated by this work has drawn attention to the many strands of Orientalist studies in the West and to the fact that French and German studies of South Asia do not project any unquestioned superiority. For ancient India, the controversy has led to more research on Western historiography than to any reorientation in the exploration of substantive themes.

Indology after World War II

The post-World War II historiography of ancient India has moved away from these earlier preoccupations and, under the influence of Marxism as well as anthropology, concerned itself

with discovering structures like feudalism and patterns of large-scale social change over time. The writings of Kasambi (1965), Sharma (1965), Romila Thapar et al. (1990) are prime examples of this initiative. The multivolume *History and Culture of the Indian People* edited by Majumdar (1951–1969) reflects a more traditional approach, while Louis Dumont's *Homo Hierarchicus* (1972) applies Levi-Strauss' structural analysis to the study of the caste system.

Historiography of the Indo-Islamic Period

The thirteenth century saw Turkish and Afghan dynasties and tribes, Muslims by faith, firmly established as rulers in many parts of India. The sultanate based in Delhi, where there was a succession of dynasties, became the seat of a pan-Indian empire with fluctuating boundaries. These dynasties were succeeded by the Mughals or Timurids in the sixteenth century and their effective rule lasted until the late eighteenth century. This long period, marked by the encounter between the older Indian civilization and the religion and culture of Central Asia's Islamic peoples, has been the subject matter of a rich historical literature.

The Medieval Persian Chronicles

India's Central Asian invaders brought with them a rich tradition of historiography. This literature, written mostly in Persian, is known as *tarikhs*, a term, translated as annals. The *tarikhs* are narratives of dynastic successions, wars, and conquests. Some of the writers had leanings toward Islamic fundamentalism and hence were inclined to present the wars of the Muslim invaders and Hindu dynasts as *jihad* or 'holy' wars. These accounts were later used by some British writers to project the Indo-Islamic era as an age of Muslim tyranny. This perception informed the ideology of India's proto-nationalism, which saw the struggle of the Rajput clans and other Hindu chiefs to retain their territories as the Hindu struggle for independence. However, chronicles provide ample evidence of a pragmatic bent in the policies of the Muslim kings, anxious not to alienate their Hindu subjects. The historians at the court of the third Mughal emperor Akbar, project his policy of peace with all, which was not to the liking of the more orthodox elements at his court. One component of the historical literature of the period is the memoirs written by the emperors, Babar and Jahangir, and by Babar's daughter, Gulbadan. These sensitively written accounts offer a portrait of the age and the family life of the Mughals.

Regional Histories

The Indo-Islamic era also produced a rich and varied body of historical writings written in the regional languages. Best known among these are the Marathi *bakhars*, modeled on the Persian *tarikhs*, the chronicles of the Ahom kings of Assam, known as *buranjis*, and the *Madla Panji* retained in Orissan temples which recorded the annals of the local dynasties. Besides these, the Vaishnavas in Bengal produced a body of hagiographical literature celebrating the lives of their founder, Chaitanya and other pioneers of the movement. All these products of the period have been later used to reconstruct the history of the period.

Colonial and Postcolonial Historiography of the Indo-Islamic Era

The earliest 'modern' historical writings on the Indo-Islamic era go back to the nineteenth century. Works such as Orme's *History of Indostan* are amateur efforts focused on the political history of the period. The famous collection of selected source material, translations from the Persian chronicles, Eliot and Dawson's *History of India as Told by Her Own Historians* (1981), contain introductions that are essentially historical statements. One of the compiler's stated objects was to prove how very inferior 'Muslim rule' was compared to its British counterpart. In general histories of India, written by administrators–scholars like Vincent Smith (1981) the same concern to prove the superiority of British rule is less obviously present. It only appears as a view of Indian history that projects a long-term proneness toward political unification that finally succeeds only under the British. Such works contain an implicit evaluation of moral and material welfare of the people in pre-British times. The twentieth-century economic historian W.H. Moreland (1920) undertook a comparison between India at the death of emperor Akbar with conditions in the nineteenth century. Serious academic research on the Indo-Islamic era, based on the study of the chronicles as well as documentary evidence like the *akhbars* or newsletters of the Mughal period, belongs to the twentieth century. It produced magnificent works of scholarship such as Sir Jadunath Sarkar's multivolume histories of Aurangzeb and the later Mughals focused on political narratives. There were other studies dealing with the religious policies of the Mughals and art and architecture of the period. The burgeoning nationalist consciousness sought to emphasize the syncretist traditions of the Indo-Islamic era, implying that the conflict between communities was a product of British rule.

Since World War II, a major development in this field of study resulted from the work of a group of scholars based at Aligarh. Irfan Habib's *Agrarian System of Mughal India* (2000), drawing upon a wide range of source material, was hailed as a pathbreaking contribution to the study of agrarian history. The research of the Aligarh 'school' uses the methods of modern historiography and has drawn upon a vast quantity of new material. While some of the scholars subscribe to Marxist theories of history, the impact of Marxism is not very obvious in their work.

In recent years some Western historians have projected the view that the Mughal Empire, like all its predecessors in India, was a 'segmentary state' under a thin veneer of centralization. This has led to a controversy with some of the Aligarh historians.

Historiography of the Colonial Era

Colonial Historiography under Colonial Rule

British historian's writing in the colonial era produced a large body of works that recorded the history of the conquest, the

wars with indigenous powers, and histories of the local powers like the Marathas, Sikhs, and Mysore with whom they had to contend for power. These histories are in the purely narrative mode, supplemented by ethnographic accounts of various elements in Indian society. They are informed by a simple assumption regarding the legitimacy and naturalness of the British conquest. One of the pioneers in the field, James Mill (Mill and Wilson, 1997), writing in the early decades of the nineteenth century, first projected a clear notion of cultural and racial superiority, significantly absent from earlier British writings on India.

British writings in the latter half of the nineteenth and early decades of the twentieth century often assumed a defensive posture in response to the growing nationalist criticism of their rule, especially the charge that the colonial nexus had induced a drainage of resources from India, thus steadily impoverishing the country. British historians challenged this view and instead projected a picture of economic transformation and development as a result of colonial rule. This controversy has not yet died out. The British historiography of colonial rule before World War II almost entirely focused on the activities of the government. A number of studies dealt with individual viceroys. Some twentieth-century works briefly refer to the burgeoning nationalist movement (one historian described it under the heading 'Political Sentiment') but with one exception they dismissed it as a phenomenon not to be taken seriously until the very eve of the transfer of power. However, the British colonial historiography increasingly adopted the methods of modern research with its emphasis on 'objective' analysis of records. But since the records in question were for the most part products of the colonial government, the resulting research inevitably reflected the built-in bias of the documents, a bias that the historian generally shared. Indian historians, excepting those writing on economic history, rarely ventured into comments on the history of British rule. Exceptions to this pattern were the publicists writing on the rising of 1857 (one of whom described it as India's First War of Independence) and authors who considered British rule to be beneficial for India. The strong ideological and political undertone manifest in the historiography of the colonial era has assumed more subtle forms in the postcolonial period.

Historiography of Modern India after Decolonization

Since the transfer of power in 1947, attention has shifted from the working of the British administration to the structure and dynamics of indigenous politics, though a number of monographs on the political history of the Raj, as well as its administrative and constitutional initiatives, were written during this period. Indian historians now came to be deeply concerned with the history of the Indian nationalist movement and its social–cultural origins. A group of historians approached the subject from the overt perspective of nationalism; others were influenced by Marxist class analysis and saw Indian nationalism as an expression of colonial bourgeois aspirations. A group of British historians centered at Cambridge explored the close relationship between British administrative and colonial initiatives on the one hand, and the structure and dynamics of indigenous politics on the other. The competition between those who were admitted to a share in power by the colonial ruler and those who were not was seen to be the central feature of indigenous politics. Nationalism itself was perceived to be a strategy reflecting the aspirations of the 'have-nots.' This view was questioned by Indian historians who saw the thesis as a continuation of the old colonialist perception dismissing nationalism as a conspiracy of the self-seeking few. However, the thesis was directly, or indirectly, questioned outside India as well. A group of historians under the leadership of Professor Low of the Australian National University recognized the reality of Indian nationalism and proceeded to unravel its local and class roots. American scholars brought to the subject the insights of social sciences, especially sociology and anthropology. There were considerable developments in economic history, using economic theory as well as statistical methods. The new techniques were used to tackle the old questions regarding the negative or positive impact of colonialism. The debate, as noted, remains inconclusive. The inadequacy of statistical evidence is one reason for this state. It should be noted here that Indian historical research has now become an international enterprise to which the Japanese, Russian, and continental scholars have made significant contributions.

One major new development in the subject is centered on the series of publications entitled *Subaltern Studies* (Guha et al., 1988). Borrowing the concept of the Subaltern from the Italian Marxist, Gramsci, the contributors to this series have focused on the history of the underprivileged elements in South Asian society and used the postmodernist techniques of discourse analysis to unravel the autonomous consciousness of the dispossessed and the record of their resistance to exploitation and domination by the elite. This movement, too, has become an international enterprise with historians and social scientists from all parts of the Anglophone world contributing to the *Subaltern Studies*.

Pakistani historians and their preindependence predecessors who focused on the development of Muslim consciousness and aspirations in the subcontinent have been deeply influenced by the theory of two nations, the ideological basis of the demand for Pakistan. Their perception is in sharp contrast to the nationalist perception of coexistence and mutual accommodation between the major ethnic cultures of the region. The history of Sri Lanka, despite its many links with the civilization and society of the mainland, has been unjustifiably peripheral to the concerns of the historians of India. But there was a strong tradition of historiography in the region drawing, inter alia, on the Buddhist religious texts for the ancient period. Some of the products of this historiography are influenced by Marxist theories of history.

One marked feature of developments in the historiography of modern India is the diversification of its subject matter. History of art and its links with nationalist sentiment, women's studies, and ecological history are some of the areas explored in recent years. This has helped bring the subject within the sphere of contemporary intellectual interest and brings to an end its intellectual isolation.

See also: Colonization and Colonialism, History of; Historical Sociology; Historical Thought and Historiography: Current Trends; History and the Social Sciences; History: Overview; Imperialism, History of; Orientalism; Subaltern History; Time, Chronology, and Periodization in History.

Bibliography

Anand, S., 1991. Modern Indian Historiography: From Pillai to Azad. MG Publishers, Agra, India.

Bayly, C.A., 01 January, 1997. Modern Indian historiography. In: Bentley, Michael (Ed.), Companion to Historiography.

Bongard-Levin, G.M. (Ed.), 1986. Tripitaka, SutraØitaka, Mahaparinivvanasutra. International Institute of Buddhist Studies, Tokyo.

Devahuti, D., Indian History and Culture Society, 1979. Problems of Indian Historiography. D.K. Publications, Delhi.

Dumont, L., 1972. Homo Hierarchius: The Caste System and Its Implication. Paladin, London.

Elist, H.M., Dowson, J., 1981. In: Nizami, K.A. (Ed.), The History of India as Told by its Own Historians: The Muhammadan Period. Delhi, India.

Gallagher, J., Seal, A., Johnson, G., 1973. Locality, Province and Nation Essays on Indian Politics 1870 to 1940. Cambridge University Press, Cambridge, UK.

Guha, R., Chakravarti, S., Gayatri, 1988. Selected Subaltern Studies. Oxford University Press, New York.

Habib, I., 2000. Agrarian System of Mughal India 1556–1707, second ed. Oxford University Press, New Delhi, India.

Kasambi, D.D., 1965. The Culture and Civilisation of Ancient India in Historical Outline. Routledge, London.

Konkle, M., 2004. Writing Indian Nations: Native Intellectuals and the Politics of Historiography, 1827–1863. University of North Carolina Press, Chapel Hill.

Majumdar, R.C., 1951–1969. History and Culture of the Indian People, vol. 10. Bharatiya Vidya Bhavan, Bombay, India.

Mill, J., Wilson, J.H., 1997. The History of British India. Routledge, London.

Moreland, W.H., 1920. India at the Death of Akbar. Macmillan, London.

Nicholson, A.J., 2010. Unifying Hinduism: Philosophy and Identity in Indian Intellectual History. Columbia University Press, New York.

Palit, C., Institute of Historical Studies (Kolkata, India), 2008. Indian Historiography. Institute of Historical Studies, Kolkata.

Philips, C.H. (Ed.), 1961. Historians of India, Pakistan and Ceylon. Oxford University Press, London.

Said, E.W., 1995. Orientalism. Penguin, Harmondsworth, UK.

Sarkar, S., 1985. Modern India: 1885–1947, second ed. Oxford University Press, Delhi. Macmillan, New Delhi. 1997. Writing Social History.

Sharma, R.S., 1965. Indian Feudalism, 300–1200. Calcutta University Press, Calcutta.

Singh, G.P., 2003. Ancient Indian Historiography: Sources and Interpretations. D.K. Printworld, New Delhi.

Smith, V., 1981. In: Spear, T.G. (Ed.), The Oxford History of India. Oxford University Press, Delhi, India.

Stein, M.A., 1900. Kalhana's Rajatarangini: A Chronicle of the Kings of Kashmir. Westminster, UK.

Tejani, S., 2008. Indian Secularism: A Social and Intellectual History, 1890–1950. Indiana University Press, Bloomington, Ind.

Thapar, R., 1990. A History of India. Penguin, Harmondsworth, UK.

Frykenberg, R.E., India to 1858; Raychaudhari, T., India, 1858 to the 1930s; Robin J. Moore, India in the 1940s Winks, R.W., 1999. The Oxford History of the British Empire, Historiography, vol. 5. Oxford University Press, Oxford.

Historical Thought and Historiography: Southeast Asia

Anthony Reid, Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Southeast Asian literary traditions, extant from the fifteenth and sixteenth centuries, included Chinese-style dynastic histories, narratives of Buddhist sacred sites, and royal chronicles influenced by both Buddhist and Islamic patterns. More modern types of religious and secular narratives developed by the late eighteenth and early nineteenth centuries. For some of these, we can trace the external stimuli, but others appear to have developed autonomously. These traditions were overwhelmed in the late nineteenth and early twentieth centuries by Western narratives taught in colonial schools, often of a triumphalist imperial kind. These were adapted for their own purposes by Marxist and nationalist writers intent on change. Initially ignorant, embarrassed, or romantic about the older indigenous writing, Southeast Asia's new national historians rediscovered the great diversity of these traditions as they became professionalized with the development of university systems. International historians writing in English remain influential in these professional circles, particularly in the growing fields of social, intellectual, economic, and environmental history and transnational trends across the region. Controversies continue to swirl around the definition of the nation, violent traumas such as 1965–66 in Indonesia and 1975–78 in Cambodia, and the monarchy in Thailand.

Embracing the southeastern peninsula of Eurasia and the Indonesian and Philippine archipelagoes, Southeast Asia has recorded its history in a dozen written languages and countless oral traditions. Indic models of writing were influential in the millennium before CE 1200, and Chinese and Islamic ones in the subsequent six centuries, while European enlightenment interpretations of history became increasingly dominant in the nineteenth and twentieth centuries. Diversity has always been a hallmark of this region. The Vietnamese emulated the Chinese literary tradition with their dynastic chronicles; cyclical Indic ideas had a longer career in the Theravada Buddhist societies of the mainland, while linear Muslim and Christian ideas became influential in the islands. Nevertheless, synchronicities are apparent in the developments within seemingly autonomous literary and historiographic traditions.

Oral poetry and traditions were important for most peoples, but particularly for those in the highlands and eastern islands (including the Philippines), which did not absorb Indic or Chinese literary forms. As late as the 1960s, at least one people in Roti (near Timor) retained oral historical narratives that accurately recorded the names of chiefs for three or four centuries. Most of the written chronicles that took shape in the sixteenth and seventeenth centuries incorporated older traditions with the repetitive features of oral narrative.

Chronicle Traditions

The dynastic chronicles of Dai Viet were expressed in Chinese and followed the model of didactic Chinese chronicles heavily informed by the moral perspective of Tang and Sung Confucianism. Some such work survives from the fourteenth century, but the most important is the strictly Confucian chronicle composed by Ngo Si Lien in 1479. This in turn incorporates the *Dai Viet su ky* of Le Van Huu (1272), which included a historical justification for the independent imperial status of Dai Viet – then occupying little more than the Red River delta.

The remaining written traditions grew out of Indic models in Sanskrit and Pali, or in the Malay case from Islamic writing in Arabic and Persian. Sacred histories, in Mon (the then language of what is today southern Burma), Burmese, Thai, and Khmer, as well as in Pali, traced the establishment of Buddhist law and practice from the time of the Buddha to the founding of local temples in Southeast Asia. Among the earliest examples are the northern Thai *Mulasasana* (around 1420) and the Pali *Jinakalamalipakaranam* (1516), both written by monks in Chiang Mai.

Royal and dynastic chronicles are almost old. The *Pararaton*, a unique chronicle in old Javanese of the Javanese kingdoms of Singosari and Majapahit down to 1481, was written after the fall of Majapahit to Muslim forces around 1527. The earliest extant *yazawin* (royal genealogy or chronicle) in Burmese was written by the monk Shin Thiawuntha in 1520. Dynastic chronicles (*phongsawadan*) were being compiled in Thai from at least the sixteenth century, although the Luang Prasoet text conventionally dated 1680 is the oldest of the extant versions that have come down to posterity. Written by an astrologer at the cosmopolitan court of King Narai (1657–88), this chronicle marks a trend to more secular dynastic history well aware of Siam's position among its neighbors as well as European, Indian, and Chinese visitors.

The advent of Islam created more discontinuities in the south than Theravada Buddhism did in the north. As Southeast Asian Islam shifted from an Arabic to indigenous idiom in the fifteenth and sixteenth centuries, the Malay language written in Arabic script became its favored vehicle. Only in Java and South Sulawesi, and of course in non-Muslim Bali, did the older literary idiom and alphabet continue alongside the Arabic/Malay one. Probably the oldest extant Malay chronicle is that of the Kings of Pasai, in north Sumatra, which became the first substantial Islamic state in Southeast Asia in the 1290s. This *Hikayat Raja-raja Pasai* may have been first compiled by Pasai Malay captives in Java in the late fourteenth century and represents a collection of semimythical stories about the Pasai kings, much more reflective of local oral traditions than of mainstream Islamic historiography. Other Malay royally

focused chronicles of the sixteenth or seventeenth centuries take this storytelling mode to an art form that can be illuminating for the historian even when devoid of chronology. The *Hikayat Banjar*, for example, tells origin stories about the leading Islamic state in southern Borneo through a conservative ethic extolling the virtues of a Javanese cultural model.

Most reprinted today is the Genealog of (Melaka) Kings (*Sulalat'us-Salatin*), a version of which Thomas Stamford Raffles printed and popularized in 1821 under the name The Malay Annals (*Sejarah Melayu*), in a conscious attempt to provide a national history for Malays. It weaves through a host of lively anecdotes its political message that the great metropolis of Melaka fell to the Portuguese in 1511 because the fundamental contract between the ruler and the ruled was violated by the last Melaka ruler. Originating in some form a generation after this Malay disaster, the text appears to have been essentially created in the early seventeenth century by the Johor writer Tun Seri Lanang, although subsequently much modified. Not much later, but a great deal more Islamic in spirit, was the encyclopedic Garden of Kings (*Bustan as-Salatin*), written by the Gujarat (India)-born theologian Nuruddin Ar-Raniri around 1640. Written in Malay at the behest of the Sultan of Aceh, in northern Sumatra, this includes a substantial chapter on the history of Aceh, and another on Melaka, and represents the most skillful use in Malay of Islamic traditions of historiography.

Java's conversion to Islam was a turbulent and drawn-out process centered in the fifteenth and sixteenth centuries. The syncretic Islamo-Javanese courts that emerged from it produced a chronicle tradition in Javanese verse known as *babad*, part mythical, part prophetic and instructional, and part genealogy. The central narratives designed to show the legitimacy of the Mataram kings who followed are collectively known as the *Babad Tanah Jawi*. All modern historians, beginning with Raffles (see below), have drawn on this set of traditions, although the task of editing a canonical text has so far defeated modern scholarship. The most chronologically accurate of the early Javanese histories, the Kartasura verse chronicle *Babad Ing Sangkala*, was written in 1738 but thought to incorporate a relatively accurate text originally compiled before 1670.

An Autonomous Modernity?

There are interesting harbingers of a more modern linear concept of historical writing before the encounter with European models in the nineteenth century. In South Sulawesi (Celebes), an Indic-derived alphabet was only introduced for the Makasar and Bugis languages (probably from Sumatra, where it had not been used for chronicles) in the fifteenth century. It appears to have been adopted especially for the recording of aristocratic genealogies, essential for asserting supernatural origins and establishing appropriate marriage partners. From around 1600, such genealogies were transformed into state chronicles at Makasar, then a major maritime power influenced by both European and Islamic literary models. Court diaries also began to be written in the 1620s to record royal births, deaths, and marriages, as well as notable foreign visitors, wars, and innovations. These state histories adopt a critical tone toward their sources, a distinctly linear notion of the progress of the kingdom through various innovations, and a desire to learn from failures as well as successes. The Goa (Makasar) chronicle began with an explanation of why it was important that future generations remember the past, 'to avoid the twin dangers, either of our thinking that we were all great kings, or of others alleging that we were worthless.'

The late eighteenth century witnessed more abundant historical writing in a self-conscious, secular, and linear pattern in several Southeast Asian traditions. While the better survival of manuscripts written after the wars of the eighteenth century may be one explanation, there also appears to have been a creative sense of crisis and new beginning that prompted reflection on the past. An early sign of this trend was U Kala's *Great Chronicle of Kings* (*Mahazawingyi*), the first comprehensive and chronologically reliable Burmese chronicle, written in the 1720s. In the late eighteenth century this was incorporated into the first history that might be called national, the "New Chronicle of Myanma" (*Myanma Yazawinthit*), also the first to use its sources in a critical sense.

Among Vietnamese, the new trend was embodied by the scholar-official Le Quy Don (*c.*1726–84), who wrote both a Le Dynasty history in biographical form – *Dai Viet Thong Su* The Complete History of Dai Viet – and the encyclopedic *Phu Bien Tap Luc* (*Miscellaneous Records of Pacification in the Border Area*). The latter addressed technological, economic, cultural, and political developments in the southern Vietnamese kingdom (known to foreigners as Cochinchina), which was a rival of his own. The Thai-language kingdoms endured a mauling from Burma in the late eighteenth century, and virtually all responded in the period 1790–1830 by carefully reconstructing their histories to draw moral lessons from disaster and to establish the legitimacy of new lineages. In Java for a similar reason, there was a spate of chronicle writing after a long period of warfare that ended with the peace of Gianti in 1755.

European Enlightenment Influences

The Philippines was the area best endowed with early European writings because of the prominence in the colonization of the islands of Spanish religious orders, each of which produced chronicles. Among the best of these, containing much material about the pre-Hispanic Philippines, were those of the Franciscan Marcelo de Ribadneyra (1601), the Jesuit Pedro Quirino (1604), and the Dominican Diego Aduarte (1640). An outstanding later example of the genre was the 14-volume chronicle of Spanish and missionary activity in the islands by the Recollect Juan de la Concepcion (1788–92). As nineteenth century Spanish accounts became ever more negative about the achievements of the Filipinos, the first generation of nationalists turned for inspiration to an outstanding Spanish secular writer, Antonio de Morga, whose *Sucesos de las Islas Filipinas* (1609) received its first modern edition (1868) in English rather than Spanish. The intellectual pioneer of Philippine nationalism, Jose Rizal (1861–96), published in Paris a new Spanish edition of the work in 1890, with the professed intention "to awaken your [Filipino] consciousness of our past, already effaced from your memory, and to rectify what has been falsified and slandered."

European writer-officials such as de Morga sometimes incorporated translations of local chronicles into their texts, even if increasingly skeptical about their mythic and didactic purposes. The earliest version of the Thai royal chronicle was that acquired and translated by the Dutch factor Jeremias van Vliet in 1640, with unacknowledged help from his well-connected local wife. More influential on European enlightenment ideas, however, was the elegant description of Siam in the 1680s by an envoy of King Louis XIV, Simon de la Loubère. A number of other Dutch, French, and English accounts of seventeenth-century Southeast Asia were influential for the expanding European understanding of the world, and many remain useful sources today. Empirically, the most ambitious such work was the encyclopedic *Oud en Nieuw Oost-Indiën* (8 volumes, 1724–26) of Francois Valentijn, who compiled it in retirement in Dordrecht after serving for 16 years as a Dutch Reformed Pastor in Ambon and Java.

The nineteenth century began with a number of scholar-officials participating self-consciously both in enlightenment ideas about history and progress, and in Britain's rediscovery of Southeast Asia. William Marsden (1754–1836), later Vice-President of the Royal Society, was a pioneer with his *History of Sumatra* (1783). As governor of Java during the British wartime interregnum (1811–16), Stamford Raffles (1781–1826) was able to command both more resources and more influence with his *History of Java* (2 volumes, 1817). He popularized the romantic ruins of vanished kingdoms in Java and elsewhere and the notion that modern Southeast Asians were in a state of decline from their 'classic' age, which could only be reversed by enlightened British rule. The Edinburgh-trained doctor John Crawfurd (1783–1868) deserves particular attention as an historian and encyclopedist of Southeast Asia as a whole. His *History of the Indian Archipelago* (3 volumes, 1820), followed by the *Descriptive Dictionary of the Indian Islands and Adjacent Countries* (1856), covered almost the whole region, based on first-hand knowledge of Java and Singapore as an administrator, and the official missions he led to Burma, Siam, and Viet Nam in the 1820s, each of which yielded published reports that began the rendering into English of the chronicle traditions of these countries.

Writers in Malay were the first to respond to this new European scholarship, notably those who cooperated in the quest for manuscripts by English and Dutch officials. Munshi Abdullah bin Abdul-Kadir (1797–1854) was a Malay writer for Stamford Raffles and his successors in Singapore, and used his understanding of British mores to launch a critique of the Malay society of the sultanates. In his vivid autobiographical works, *Kisah Pelayaran Abdullah ke Kelantan* (*The Story of Abdullah's Voyage to Kelantan*, 1838) and *Hikayat Abdullah* (1849), he introduced such concepts as progress, race, freedom, and the rule of law into Malay historical writing. South of Singapore, the Riau court writer Raja Ali Haji (*c.*1809–70) was in regular contact with the Dutch Malay scholars of his day and learned something of their perspective on history. The spirit of his two outstanding histories, *Silsilah Melayu dan Bugis dan sekalian raja-raja-nya* (*Genealogy of Malay and Bugis Princes*, 1865) and *Tuhfat al-Nafis* (*The Precious Gift*, around 1866), is nevertheless heavily didactic in the best Islamic tradition.

As Southeast Asia became partitioned among the colonial powers in the second half of the nineteenth century, the more talented colonial officials pioneered modern-style 'national' histories in English, French, and Dutch. These usually combined some of the chronicle record of ancient kingdoms, the interventions of foreigners, and the rise of colonial power within the borders later marked out by European treaties. Sir Arthur Phayre, an army officer who spent the years 1835–67 in key posts in Burma, was the first of these with his *History of Burma* (1883), followed in 1926 by a book of the same name by G. E. Harvey. Adhémard Leclère wrote a *Histoire du Cambodge* in 1914, sometime before the more demanding equivalent for Viet Nam, *Histoire Moderne du Pays d'Annam (1592–1820)* (1920), by Charles Maybon. A history of Siam was first published in French on the basis of missionary accounts by Turpin in 1771, but the first of the modern European histories was that of the British Consul in Chiang Mai, W. A. R. Wood (1926), relying much on the chronicles. The Netherlands-Indian school system, like those of colonial Viet Nam and Burma, required 'national' textbooks in the colonial language and idiom, of which the most influential was the Dutch-triumphalist *Geschiedenis van Nederlandsch Indië* (1930) of F. W. Stapel. In Malaya, however, there was a deliberate distinction between a colonially centered history of the place in English and a dynastic chronicle-centered story of the Malay people in Malay, the latter based on Sir Richard Winstedt's *Tawarikh Melayu* ('Malay History,' 1918).

Although such writings were influential in creating the sense of a national identity (albeit colonially dominated), it was European archeology that excited the imagination of both Southeast Asians and outsiders. Raffles had brought the great monuments of Borobudur and Prambanan to world attention, and later in the century, Dutch scholars began systematic work to document them. The ruins of Angkor became known to Europeans in the 1860s, which began a romantic quest for ancient civilizations. The modern restoration of the ruins of Angkor began under French auspices in 1907, while the Dutch and English were not far behind in restoring the monuments of Java and Burma, respectively. The inscriptions of these and other temples allowed the histories of the temple-building Hindu–Buddhist kingdoms to be reconstructed, through the work of Louis Finot and George Coedès in Cambodia, N. J. Krom and W. F. Stutterheim in Java, and the 'Greater India' school led by R. C. Majumdar in Calcutta. In Java and Cambodia in particular, the fascination with these grand but mysterious pre-1400 kingdoms was a more powerful spur to nationalism than were the later kingdoms that seemingly either collaborated or failed in the collision with European power.

Nationalist and Marxist History

All of Southeast Asia except Siam/Thailand was under Western colonial control from 1900 to 1941 and was ruled by the Japanese for nearly four turbulent years (1942–45) thereafter. The sudden Japanese surrender in August 1945 inaugurated a period of revolutionary struggles in the successor states, with relative peace imposed on Indonesia only around 1966, Viet Nam after 1975, and Cambodia in the 1990s. The first half of the twentieth century was a period of extraordinary debate and rethinking of the meaning of history, as a Western-educated generation increasingly detached from cultural connections to

the immediate past developed new ways of reinterpreting the new Western narratives for modernizing purposes. First Social Darwinism and then Marxism were seized upon as means to explain what had to be done to save the nation from colonial humiliation. By the 1950s, a nationalist, anticolonial historiography was dominant everywhere except Malaya/Malaysia and Thailand, turning militant opponents of the colonial order into the most exemplary heroes. Marxist categories were selectively adopted but were influentially used everywhere to devalue the precolonial past as generically 'feudal.' Debates remained intense between communist, religious, and ethnonationalist readings, and often between local and national perspectives.

In Indonesia, nationalist writers suddenly occupied center stage when the Japanese occupation removed all Dutch textbooks. Muhammad Yamin (1903–62), a Sumatran convert to a Java-centric view of national history, became the most influential writer and patron of a nationalist history that glorified Sriwijaya and Majapahit as precursors of modern Indonesia. His works on the fourteenth-century Javanese kingdom of Majapahit, *Gajah Mada* (1945) and *Tatanegara Majapahit* (3 volumes, 1962), have long-term significance.

In Thailand, a similarly influential figure through the most nationalist phase was Wichit Wathakan (1898–1962), a commoner who became an immensely popular writer of historical novels and plays, as well as a 12-volume world history (1931). After the antimonarchy revolution of 1932, he replaced Prince Damrong Rajanubhab (1862–1943) as the government's arbiter of culture and history. Where Damrong had chronicled a Thai identity developed from the monarchy's gradual adoption of liberal modern values, Wichit glorified the purity and heroism of the Thai race.

In Viet Nam, the debate between traditionalists, nationalists, and Marxists centered on the evaluation of the Nguyen dynasty, which unified the country with some French assistance in 1802, and of the Tay Son rebellion, which preceded it in the late eighteenth century. Tran Trong Kim (1883–1945), author of the first modern Vietnamese history (1928, constantly reprinted until the 1960s) and briefly Prime Minister under the Japanese, broke with the dynastic chroniclers in accepting the Tay Son leader Quang Trung as a national hero. Nevertheless he accepted the first Nguyen emperor Gia Long, who defeated Quang Trung's forces, as a dynastic founder and unifier. Dao Duy Anh's *Short History of Vietnamese Culture* (1938) was the first and perhaps the best of the Marxist-influenced studies that saw Vietnamese culture as constantly evolving through social forces. For later Marxists, the Tay Son were heroic peasant rebels, while the last dynasty was a disaster because it failed to adapt to historical conditions.

The nationalist format remained dominant within Southeast Asia into the 1970s, and was given international currency by the Indo-China conflicts. The Viet Nam war stimulated some fine research and writing in English and French seeking the key to Viet Nam's resistance to US intervention in a long tradition of anti-Chinese and anti-French patriotism. John McAlister's *Viet Nam: The Origins of Revolution* (1969), David Marr's *Vietnamese Anti-Colonialism 1885–1925* (1971), and Frances Fitzgerald's *Fire in the Lake: The Vietnamese and the Americans in Vietnam* (1972) were all influential at the height of the conflict. These helped to stimulate a strong generation of US students entering the field of Southeast Asian history in the 1970s, even though few could find academic positions in their own country in the 1980s.

Traditional monarchies were forcibly rejected by the nationalist revolutions in both Indo-China and Indonesia, both because they were deemed to have been artificially sustained by colonialism, and because they did not fit with revolutionary aspirations for the new egalitarian man. The historians also took a very critical view of nineteenth century monarchs who had failed to resist the Europeans. Some Marxists shared Tan Malaka's view that Indonesia had no history worth writing about "except one of slavery." Nationalists, however, typically reached further back into a mistier past, when conquering kings were uncritically celebrated as a boost to national self-confidence. Only the Thai (very narrowly) and Malay monarchies survived the revolutionary era of the 1940s. The establishment version of Thai nationalist history constructed by King Vajiravudh and Prince Damrong in the first two decades of the past century was shaken by the military revolution of 1932 and almost forgotten amid the strident nationalism of the 1940s, but reestablished itself thereafter in the long reign of King Bhumibol (1950–present). It stressed the unique virtues of Thai monarchy in preserving Thailand's independence and fashioning a flexible modernity. The contestation of this increasingly rigid format by professional and international historians (led by Thongchai Winichakul) became since the 1990s the dominant controversy of Thai history.

Postnationalist, Regional, and Thematic Histories and Debates

The 1970s and 1980s witnessed a number of Cold War authoritarian regimes, which imposed their narrow versions of the nationalist (or in Indochina Marxist-nationalist) historiography in schools and universities. This provoked a reaction toward more critical studies, distrustful of meta-narratives especially when they served state interests. Professional historians who had served the dictators, such as Professor Nugroho Notosusanto who became education minister under General Suharto's long presidency (1966–98) of Indonesia, tended to be discredited when their patrons fell from power. The many Filipino historians who were generously funded to write the multivolume *Tadhana: The History of the Filipino People* (1976–80) were fortunate that only President Ferdinand Marcos' name appeared on the books, which were scarcely mentioned after the Dictator's overthrow in 1986.

The heaviest national imprint on historiography was, however, that of Viet Nam, because questioning that narrative was long equated with supporting southern separatism or American intervention. Historians of the South Vietnam regime (1954–75), often themselves refugees from the north, did not challenge the grand narrative for suppressing southern identity but rather for leading naturally to communism. It was left to a new Vietnamese-American generation of the children of refugees to begin since 2000 to write sympathetically about losers and victims of the *Nam Tien* (southward movement) of the Vietnamese state – Cham, Khmer, various once-autonomous communities, and local regionalisms. Keith Taylor at

Cornell University was the most prominent Western historian to break the mold, seeming to reject his own pioneering *The Birth of Vietnam* (1983) in a series of dramatic concessions to Indochinese pluralities. The nationalist narrative had downplayed or decried the division of Viet Nam into two warring polities throughout the period 1600–1777, whereas Li Tana's *Nguyen Cochinchina: Southern Vietnam in the Seventeenth and Eighteenth Centuries* (1998) celebrated the southern kingdom's diversities. Among the volumes signaling the new trend were Keith Taylor and John Whitmore, *Essays into Vietnamese Pasts* (1995), and Nhung Tuyet Tran and Anthony Reid, *Viet Nam: Borderless Histories* (2006).

Anglo-Irish Benedict Anderson (*b*.1936) made his career at Cornell University, where his brand of culturally informed political and intellectual history had even more influence on historians than on his own discipline of political science. His pathbreaking *Imagined Communities: Reflections on the Origins and Spread of Nationalism* (1983) was preceded by a series of more specific studies on Javanese and Indonesian political culture through Japanese occupation and revolution. In the 1970s and 1980s, deeply alienated from Suharto's Indonesia, he addressed the political identity of Thailand and the Philippines in a revisionist manner, undercutting the idea of Siam's uncolonized uniqueness. Among the able historians trained at Cornell, Michigan, and elsewhere were a number of Southeast Asians who had an impact in their own countries as well as internationally. Reynaldo Ileto (*b*.1946) pioneered a new kind of history from below by beautifully demonstrating how popular Christianity provided the idiom of resistance in his *Pasyon and Revolution: Popular Movements in the Philippines* (1979).

Nidhi Eoseewong (or Aeusrivongse, *b*.1940) wrote almost entirely in Thai after returning to teach in Chiang Mai, but his admirers eventually ensured a translation in *Pen & Sail: Literature and History in Early Bangkok* (2005). Shifting agency in Thai history away from kings and Westerners to broader socioeconomic changes, he argued that the emergence of a Sino-Thai market economy in the early nineteenth century lay behind a new modern spirit in literature. Thai historiography was also enriched by a generation of (largely Sino-Thai) student leftists in the 1970s, who were either arrested during the military coup of 1976 or escaped to briefly join the communist insurgency, but were eventually able to study overseas. Kasian Tejapira at Cornell sensitively explored the relations between Chineseness, Thainess, communism, and democracy, while Thongchai Winichakul's Sydney thesis, published as *Siam Mapped: A History of the Geo-body of a Nation* (1994), injected into the debate on nationalism the importance of mapping and spatial representation.

Taking the critique of nationalist histories to the furthest extreme was the work of political scientist James Scott (*b*.1936), who wrote in an increasingly historical vein about the everyday resistance of peasant societies. His *The Art of Not Being Governed: An Anarchist History of Upland Southeast Asia* (2009) argued powerfully that the statelessness of many Southeast Asian peoples was a rational preference for freedom rather than a sign of backwardness. The interdisciplinarity of Southeast Asian history was indeed its comparative advantage. Besides Scott and Anderson, the cultural anthropologist Clifford Geertz (1926–2006) was perhaps the most influential Southeast Asianist among historians. History is seldom absent from his studies of Java, Bali, and Islam, but most salient in *Agricultural Involution: The Process of Agricultural Change in Indonesia* (1963) and *Negara: The Theatre State in Nineteenth Century Bali* (1980). The latter influentially argued that elaborate royal rituals were the defining acts of Bali states, not simply their decoration or arcane cosmology.

The construction of a broader Southeast Asian historical identity was primarily the work of Western scholars, although a few early nationalists such as Jose Rizal initially sought a broader regional identity (in his case Malay) as a means to escape from colonial entrapments. Another Asian pioneer of this idea was Nguyen Van Huyen (1908–75), later Education Minister in the Hanoi government. As a student and Vietnamese teacher in Paris he discovered Dutch scholarship on Indonesia, and used it to good effect in his study of common patterns of house building in the region, probably the first academic study of Southeast Asia by a Southeast Asian – *Introduction à l'Etude de l'Habitation sur Pilotis dans l'Asie du Sud-est* (1934).

The term Southeast Asia (*Südostasien*) was first found useful by German-speaking prehistorians like F. Heger and Robert Heine-Geldern who used it from the early 1900s to describe the dispersion of bronze-age Dongson drums throughout the region, and similarities in the Hindu–Buddhist temple art and architecture. Primarily French scholarship on the inscriptions uncovered relations between the ancient kingdoms of Angkor, Champa, Dai Viet, Pagan, and Java. George Coedès (1886–1969) first brought this material together in a pathbreaking book he entitled *Histoire Ancienne des États Hindouisés d'Extrême-Orient* (Hanoi, 1944), although better known in its English translation (1968). Coedès broke with the 'Greater India' school that regarded the Southeast Asian kingdoms as Indian colonies, but continued to see Indian agency through visiting traders and scholars. Later historians such as Paul Mus, Bernard Groslier, Oliver Wolters, and Claude Jacques were inclined to see Southeast Asians borrowing creatively what they found useful in Indian civilization. Sheldon Pollock's *The Language of the Gods in the World of Men: Sanskrit, Culture and Power in Premodern India* (2006) usefully proposed Sanskrit civilization as a common process throughout South and Southeast Asia, rather than a transfer from the former to the latter.

Histories of the whole region began to appear in English in the 1950s, typically from the pen of those who, like Coedès, had begun their teaching and researching within the region. The most influential such pioneer was D. G. E. Hall (1891–1979), who established the first British chair of Southeast Asian History at the School of Oriental and African Studies in 1949 and wrote an often-reprinted *History of South-East Asia* (1955). Hall had learned his Southeast Asia as history professor at the University of Rangoon (1921–34), partly from more experienced colleagues such as Burma prehistorian G. H. Luce and economic historian J. S. Furnivall (1878–1960). The latter's theoretical reach was broader through his comparative work to explicate Southeast Asian colonial societies, and his still useful *Netherlands India: A Study of Plural Economy* (1939) popularized the argument that plural societies resist the homogenizing effects of capitalism. The other colonial seedbed of historians of the region was the University of Malaya,

opened in Singapore in 1949 and eventually spawning a range of institutions in independent Malaysia and Singapore. Regional historians such as Northcote Parkinson, C. D. Cowan, John Bastin, Wang Gungwu, Jeyamalar Kathirithamby-Wells, David Bassett, Jan Pluvier, Anthony Reid, Leonard Andaya, and Barbara Andaya learned much of their Southeast Asian perspective by teaching in these institutions. By contrast, the strong generation of late colonial-era Dutch and French historians had difficulty finding successors amidst postcolonial bitterness and disillusion.

The initiative passed after World War II largely to the United States, Australia, and Japan. In the United States, there was significant work on comparative colonial history by Rupert Emerson and others before World War II, and a new wartime focus on the region stimulated by the Institute of Pacific Relations. Southeast Asian area studies made a vigorous new beginning, however, in the 1950s, notably at Cornell and Yale, focusing initially on the dynamic of the diverse new nationalisms. Pre-twentieth century history developed more slowly, initially reliant on colonially experienced British imports, with first D. G. E. Hall and then Oliver Wolters (working particularly on early Chinese materials) pioneering the teaching of the subject at Cornell University. Many of Wolters' students, and later those of Thai historian David Wyatt, pioneered the creative work of combining the chronology of colonial sources with the cultural and religious meanings found in often opaque indigenous chronicles. Special mention should be made here of Merle Ricklefs' (*b*.1943) pathbreaking volumes on Javanese history from seventeenth to twentieth centuries, starting with the Cornell thesis published as *Jogjakarta under Sultan Mangkubumi, 1749–92: A History of the Division of Java* (1974).

In the 1970s, Southeast Asian studies were in a slump in the United States, but booming in Australia, where many of the new graduates moved and graduate training began. Particularly, at the Australian National University in Canberra (Anthony Reid, David Marr, Ann Kumar, A. C. Milner, and later Craig Reynolds and Reynaldo Ileto) and Monash University in Melbourne (John Legge, David Chandler, and M. C. Ricklefs) important schools developed for the training of Southeast Asian historians in multidisciplinary area studies contexts.

Japanese historians also saw Southeast Asia as a kind of frontier for better understanding regional and global dynamics, free of the dominant obsessions with the West and China. Although some pioneers like Tatsuro Yamamoto (prehistory) and Seiichi Iwao (colonial and maritime) began doing seminal work in the 1930s, Japanese historians began to make a bigger mark internationally from the 1970s. For modern Indonesia some trained at Cornell, like Akira Nagazumi (1929–87) at Tokyo University and Takeshi Shiraishi, who also taught at Cornell (1987–96) before moving to Kyoto and in 2007 to head the regionally focused Institute for Developing Economies (IDE) in Tokyo. The strongest Japanese contribution to understanding Southeast Asian history was, however, that of Japanese-trained scholars working on the societies of Mainland Southeast Asia. Their work was much better informed by Chinese sources than that of Western or local historians, and it showed a rigorous empirical spirit in economic history. Japanese writing on the history of rice agriculture in Southeast Asia was particularly strong, as demonstrated for English readers in the volume edited by Yoneo Ishii, *Thailand: A Rice-Growing Society* (1978), or some of Yumio Sakurai's work on Vietnam and Junko Koizumi's on Thailand. Much of this work was concentrated in specialist organizations set up in the 1960s, such as the IDE (1958), the Center for Southeast Asian Studies at Kyoto University (1965), and the Tokyo University of Foreign Studies (1949, on older roots).

The initial focus of the new nationalisms made the first generation of American-trained historians little inclined to perceive the region as a whole. The most important US history of the region, *In Search of Southeast Asia* (Steinberg, 1971), required the collaboration of six country specialists and is basically divided by country. The teaching of Southeast Asia outside the region itself requires a regional approach, however, and as new approaches were sought many of them found it helpful to choose a broader (or narrower) focus than the nation states. The *Cambridge History of Southeast Asia* (ed. Nicholas Tarling, 1992), dominated by historians based in Australia, New Zealand, and Southeast Asia, took the heroic line that each chapter should cover the whole region, but be chronologically and thematically demarcated.

The approach of Fernand Braudel and the *Annales* School was brought to Southeast Asia in the 1980s in relation to the Early Modern, a period that had until then been largely abandoned to the 'European expansion' school. Denys Lombard (1936–98), founder of the interdisciplinary journal *Archipel*, developed in *Le Carrefour Javanais: Essai d'Histoire Globale* (3 volumes, 1990) the kind of 'total history' favored by the *Annales*, seeking to peel off progressively the Western, Islamic, and Hindu–Buddhist layers of his subject. Anthony Reid (*b*.1939) acknowledged a debt to Braudel in his *Southeast Asia in the Age of Commerce, c.1450–1680* (2 volumes, 1988–93), notably his interdisciplinarity, his 'mediterranean world' focus, and his elevation of structures, *mentalités*, and the everyday above political *événements*.

Reid's proposed pan-Southeast Asian coherence and a mid-seventeenth century general crisis, shifting the Southeast Asian trajectory away from engagement with world commerce and ideas, were robustly countered by Victor Lieberman. He perceived striking continuities in the way state consolidation proceeded in Mainland Southeast Asia, from 23 polities in 1340 to 3 by 1820. His insight that this consolidating process was common also to parts of Europe led eventually to the most ambitious exercise in world history by a Southeast Asianist, Lieberman's *Strange Parallels: Southeast Asia in Global Context, c.800–1830* (2 volumes, 2003–09). This rejected European exceptionalism by explaining the parallels between what he called the 'protected zones' of Eurasia – western Europe, Japan, and Mainland Southeast Asia – because they were environmentally protected as Russia and China were not from the great continental movements of nomadic horsemen. Lieberman's work also extended Reid's use of climatic factors to explain some of these parallels, identifying a number of other crisis moments such as the fourteenth and late eighteenth centuries.

Most of the controversies in the field occurred on a narrower stage. The unusual prominence of women, and of entrepreneurial Chinese minorities, in Southeast Asia has received

increased attention by the newer historians. The rise of China has again problematized interstate relations in the precolonial era and the nature of the 'tribute system' by which Southeast Asian polities had conducted commercial and political relations with the north. Here as elsewhere environment, sexuality, body, and health are ever more central to younger historians. Perhaps the most sensitive issue is that of state violence since independence. Although the Pol Pot regime (1975–78) fell long ago, and President Suharto in 1998, the large-scale massacres they perpetrated in Cambodia and Indonesia (1965–66), respectively, remain inadequately explained and bitterly controversial. In Burma/Myanmar, Vietnam, and Laos, the regimes that perpetrated violence against opponents and minorities are still in a position to restrict discussion of these matters internally. Contrary to earlier expectations, the prestige of academic writing in English has strengthened even within the region, driven by the usual global competitive factors, although only Singapore and the Philippines participate significantly in it. Elsewhere, the most vigorous vernacular historical debates are not in academic forums but in newspapers, films, and popular novels such as the remarkable four volumes written in internment by Pramoedya Ananta Toer (1925–2006), describing the birth of Indonesian national consciousness in the early twentieth century.

See also: Colonization and Colonialism, History of; Historical Thought and Historiography: Christian Tradition; Historical Thought and Historiography: Current Trends; Historical Thought and Historiography: East Asia; Historical Thought and Historiography: Islamic Tradition; Historical Thought and Historiography: South Asia; Nationalism, Historical Aspects of: East Asia; Vietnamese Revolution, The.

Bibliography

Ahmad, Abu Talib, Tan, Liok Ee (Eds.), 2002. New Terrains in Southeast Asian History. Ohio University Press, Athens.

Coedès, G., 1968. The Indianized States of Southeast Asia. University of Hawaii Press, Honolulu, HI.

Cowan, C.D., Wolters, O.W. (Eds.), 1976. Southeast Asian History and Historiography: Essays Presented to D. G. E. Hall. Cornell University Press, Ithaca, NY.

Grabowski, V. (Ed.), 2011. Southeast Asian Historiography: Unravelling the Myths. River Books, Bangkok.

Hall, D.G.E. (Ed.), 1961. Historians of South East Asia. Oxford University Press, London.

Reid, A., Marr, D. (Eds.), 1979. Perceptions of the Past in Southeast Asia. Asian Studies Association of Australia/Heinemann, Singapore.

Tarling, N. (Ed.), 1992, The Cambridge History of Southeast Asia, vol. 2. Cambridge University Press, Cambridge, UK.

Historical Thought and Historiography: Sub-Saharan Africa

M Diawara, Center for Research on Local Knowledge, Bamako, Mali

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 10, pp. 6813–6817, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article covers the recent origins of written history in sub-Saharan Africa, the 'discovery' and scrutiny of its oral history traditions, the interaction of anthropology and history, the Anglo-American debate about the nature of sub-Saharan African historiography, and the role of the diaspora both in Africa and more broadly. Despite continuing economic dependence, indigenous perspectives upon sub-Saharan Africa and its history are growing, fostered in part by the diaspora itself across Europe, North America, and Asia.

This article concentrates on sub-Saharan Africa. It covers recent origins of written history, the 'discovery' and valuation of its oral traditions, the interaction of anthropology and history, the Anglo-American debate, and the role of the diaspora. Intellectual dependence resembling the economic impasse of the continent is worsening. A glimpse at what is on the horizon suggests the tide is turning, the groundswell stemming from the unprecedented creativity of local bards, both griots and other producers of historical material. In fact, a wider influence of African studies on intellectual life can be detected. This resurgence is amplified by the diaspora, which actively propagates its share of the African experience. Hope is raised by the fact that disciplines applied to Africa are inevitably linked to academic trends in Europe and the US. No longer cramped by a narrow view of a particular field of study, Africanists have demonstrated the benefits of multidisciplinary collaboration. In addition, Africanists have been instrumental in fostering a dialogue and creating an interface between traditional academic approaches in Anglo-Saxon and French-speaking countries.

The discovery of Africa as an object of scholarly research is part of a wider expansion of academic pursuit. It coincided with the emergence of women's studies. The focus of investigation shifted to the underdogs, including slaves, colonized nations, and other silent majorities of our day as well as the Middle Ages, in Africa, and elsewhere (Wachtel, 1971; Clancy, 1979; Ginzburg, 1980; Feiermann, 1993).

This article concentrates on sub-Saharan Africa. It covers recent origins of written history, the 'discovery' and valuation of its oral traditions, the interaction of anthropology and history, the Anglo-American debate, and the role of the diaspora.

Origins of History

The first Sahel states emerged on the northern fringe of the savanna: Ghana (eighth to eleventh century), Mali (thirteenth to fourteenth century), and Songhay (twelfth to sixteenth century).

Arab authors were the most knowledgeable on Africa's past. Those who made the largest contributions to reconstructing the history of Africa, and Western and Central Africa in particular, include al-Bakri (1029–94), al-Idrisi (1154), al'Umari (1301–49), and Hassan Ibn Mohammed al-Wuzza'n, known as Leo the African (1494–1552). Ibn Khaldun (1332–1406) is unquestionably the 'Father of History,' even though Ibn Battuta (1304–69) remains the best thanks to his personal accounts and the detail in his descriptions (Levtzion and Hopkins, 1981).

Written European sources followed, beginning in the fourteenth century. Their authors were Portuguese navigators. These documents preceded the substantial addition made by codification or oral records in Arabic by Mahmoûd Kâti and Es-Sa'di, both Timbuktu scholars.

In the course of the nineteenth and twentieth centuries, accounts written by Africans in indigenous languages have abounded. A large share of these histories was in Kiswahili, Hausa, Fulfulde, Kanembu, Diula, and Malagasy.

The first professional historians, caught in the Hegelian trap, discounted the history of Africa (Neale, 1986; Mudimbe and Jewsiewicki, 1993: p. 1). However, the continent and its 'people with no history' were deemed worthy of a present. Ethnography was called upon (Mudimbe and Jewsiewicki, 1993: p. 8) to justify the backwardness of the region. Underdeveloped Africa was expected to make do with a 'prehistory' (Moniot, 1991: p. 197).

The International African Institute was established in London in 1928. The School of Oriental and African Studies began in 1948 to teach a new subject called history of Africa (Vansina, 1992: p. 77). An overdue offspring of academic research was born. Curiously, African universities in French colonies, the local intelligentsia, and their backers in France did not see any need for the discipline (Vansina, 1992; Moniot, 1995: p. 647).

History and Anthropology

Almost 8 years after the publication of *African Historiographies* (Jewsiewicki and Newbury, 1986), Jewsiewicki and Mudimbe published *History Making in Africa* (1993). The majority of the authors who contributed to the former work were historians. Anthropologists dominated the latter. This shift in the weight accorded to the two disciplines is a sign of the times. When focused on Africa, anthropology and history do not fit neatly into the university compartmentalization scheme. Rebellious grassroots reality has gradually overcome the rigidity that had governed the allocation of professorships.

As time passed most of the convenient cubbyholes have been banned to the realm of memory. It was once the fashion to contrast societies with no history, acephalous, precapitalist societies, with societies blessed with a state and a history culminating in capitalism. The first group was viewed as inherently handicapped by communalism, polytheism, oralism, and tribalism. The second group of nations was endowed with the virtues of individualism, monotheism, literacy, and polity, to name just a few. Criticism of the concepts of tribe and ethnicity took two different approaches, a French school and an Anglo-Saxon school (Coquerel, 1983; Young, 1986; Chrétien and Prunier, 1989; Barber et al., 1990; Coplan, 1991, 1993; van Jaaresveld, 1992; Harries, 1993; Jenkins, 1997).

The adoption of precolonial Africa as a concept for a catchall analysis of the continent marked a milestone. This approach created the illusion of a different, idealized world. Postcolonial Africa was viewed as the child of a past to which the umbilical cord had been severed. This is the country cherished by the nationalists during their struggle for independence and the image invoked by politicians. Cheikh Anta Diop was fond of climbing the continent's family tree back to its glorious Egyptian ancestors, while Joseph Ki-Zerbo (1972) and J.F. Ade Ajayi convey us to a not so distant past, labeled the precolonial era (Mudimbe and Jewsiewicki, 1993: pp. 1–2). Historians responsible for this breakdown into periods have changed their minds and now support the anthropologists' view of the existence of a continuum between the colonial societies and their worthy successor, sovereign Africa (Bayart, 1989; Piault, 1987).

The focus of historians on the past and the concentration of ethnologists on the present explain the farsightedness of the former and the shortsightedness of the latter. The outlook of these two isolated disciplines changes radically as soon as historians discover humanity as the center of their field of investigation and anthropologists take account of time, the framework for human behavior, and activity.

The Silent Make Themselves Heard

Africa is known for its chanted, sung, or simply recounted oral history. The *griots*, professional chroniclers of West Africa, the *biiru*, specialists in Rwandan royal ritual, and ordinary storytellers in the cities and the country have provided a wealth of material to researchers accustomed to studying written material.

In 1962, Jan Vansina published his authoritative work, *De la tradition orale*. His name is invariably associated with the acceptance of oral accounts as a source of African history and with the elaboration of a rigorous scientific method. It took historians, or at least part of their fraternity, a long time to learn to appreciate the 'auriture' of oral genres (Coplan, 1991, 1993: p. 8).

Zealous pursuit of sources engenders such excesses as equating the death of an old timer with the gutting of a library. This attitude refers us to a kind of precolonial library, to paraphrase Mudimbe. It is also called the memory of the griots and 'oral archives' (Bazin, 1979: p. 450). Writing history Western style entails the fault of reducing documents recorded by interviewers to simple sources. This is denounced by Jewsiewicki (1986: p. 16) and Shoenbrun (1993: p. 55).

An increasing hue and cry (Farias, 1992) has been raised in opposition to this tendency. The critics are anxious to upgrade the role of reputedly passive traditionists (Ba, 1980). While David W. Cohen (1989: p. 12) insists on the contrast between material provided by the Africanist community and that produced by Africans who collect data on a daily basis, Jan Vansina was an early advocate of another reading of these documents, as revealed in a text of Jewsiewicki (1979: p. 81).

As soon as we escape the pull of the two fetishes constituted by the roving reporter and the Western-style historiographer, we are attracted by other types of history, and vice versa. We can then pay heed to nonformalized oral texts, unofficial samples taken from lower rungs of the social ladder. Their particular form and their nonverbal character require another method of analysis, yielding a more fertile approach to Africanist research (Olivier de Sardan, 1976; Diawara, 1990; Barber, 1991). The field is further enriched by the involvement of anthropologists who tap the upstream end of oral production and characterize its interface with society (*surface sociale*). Their tools bridge a major gap neglected by historical studies, which concentrate entirely on the oral text and lose sight of the life, environment, and experiences of participants in society (Moniot, 1986: p. 58).

The history of the continent has become a professional endeavor. It involves "interrogating and revealing the historian's operation" (Moniot, 1995: p. 647). From now on, the Africanists can forget the "automatic legitimation of being specialists in a small exotic field" (Cooper, 1995: p. 236).

The Anglo-American Debate

Vansina (1992: p. 91) takes advantage of the lessons of historiography consisting in recognizing the past in the present and vice versa, and aims his harshest criticism at the postmodern drift of David W. Cohen. The latter (1994: p. XV) stresses the fact that historical knowledge is not the monopoly of poetry or formalized narratives but that it is constantly being proclaimed and cited. Henige (1995: p. 314) vehemently attacks what he calls Cohen's "predilection […] for subordinating the substance of sources to their interpretation."

The Diaspora Includes Africa; In Fact, It Is Africa

What position do African specialists take in this controversy, which almost boils down to an 'American–American' debate? Jewsiewicki (1986: p. 9) makes a statement that has lost none of its pertinence for historians and other Africanists, "The Africanist historian and the African historian share neither the same responsibilities nor the same existential constraints" (cf Vansina, 1994: p. 221). This is certainly true, except that another schism has split the African scholars' camp, divided between a diaspora and a home community, and occasionally bridged by a coming and going between the two shores. Africans abroad still focus more on debates that are often kindled overseas, in America or Europe.

Nevertheless, practically everything written is in the language of academics, which is not the idiom of indigenous populations. A UNESCO policy to remedy this situation, adopted in the

1970s, remains a pious vow even though issues of *l'Histoire Générale de l'Afrique* have appeared in Kiswahili and Hausa. Anyway, who is able to read them? Where can you get a copy and what would it cost? Does the body of literature constitute history for the average person? The question posed by the *African Historiographies* subtitle, 'What history for which Africa?' is as relevant as ever. The role of Africans in this international forum appears to be rather marginal. The empire of Western historiography is bigger than ever (Vansina, 1994: p. 220). The hopes raised by studies in the English-speaking world, which benefits from the North American alternative to the intellectual heirs of the British Empire and from a more substantial organization (Jewsiewicki, 1986: p. 15), continue to fade. In the meantime, the New World option has become an attractive outlet for the brain drain from tropical dictatorships. Zaire has forfeited its chances to develop a school of thought liberated from the domination of Brussels. The ex-colonies reproduce the models elaborated in the former centers of colonial power (Chrétien, 1997) despite the fact that English-speaking and French-speaking African authors have begun to read one another's works (Coquery-Vidrovitch, 1997). Their marginality becomes even more pronounced as the tendency to consume (È Nzièm, 1986: p. 25) concepts developed outside Africa grows. It makes little difference whether they are the products of non-Africans or of Africans of the diaspora, which never stops expanding (Jewsiewicki, 1979: p. 81).

Intellectual dependence resembling the economic impasse of the continent is worsening. While additions to the circle of recognized Africanists are becoming rare, a heavy storm of productivity is brewing in Africa, as far as reading about and interpreting the past are concerned. Student theses, cassettes recorded by individuals in different villages, radio stations, and other relevant institutions are mushrooming. The harvest has never been so bountiful (Cohen, 1994: p. XV; Diawara, 1997).

This analysis is not intended to deny the essential, unanticipated impact of African studies on social sciences in general. The 'dissolution of world history' by the history of Africa is a reality (Feiermann, 1993). In evaluating African studies, one is inclined to dismiss the crucial role played by the diaspora, which actively propagates its share of the African experience (Macamo, 1999). Expatriates are deprived of their birthright by attributing any success they achieve to their severance of ties to their dismal native lands in the tropics. Mudimbe and Appiah (1993: p. 113), referring to the impact of African studies on American disciplines, cite the necessity "to locate ourselves as much in the current history of those disciplines as in the current state of African Studies […]."

In fact, a wider influence of African studies on intellectual life is detected – in Western Europe and elsewhere as well as in America (Feiermann, 1993: p. 200, note 13). The impact of Africanist historiography on that of France, the former métropole, is undeniable (Coquery-Vidrovitch and Jewsiewicki, 1986: pp. 143–144).

A glimpse at what is on the horizon suggests the tide is turning, the groundswell stemming from the unprecedented creativity of local bards, both griots and other producers of historical material. This resurgence is amplified by a foreign legion, which gives a boost to those who have remained on their native soil. Hope is raised by the fact that disciplines applied to Africa are inevitably linked to academic trends in Europe and the US. No longer cramped by a narrow view of a particular field of study, Africanists have demonstrated the benefits of multidisciplinary collaboration. In addition, Africanists have been instrumental in fostering a dialogue and creating an interface between traditional academic approaches in Anglo-Saxon and French-speaking countries.

See also: African Studies: History; Anthropology, History of; Historical Thought and Historiography: Christian Tradition; Near Middle East/North Africa Studies: Culture; Near Middle East/North Africa Studies: Society and History.

Bibliography

Atieno, O.E.S., Ogot, B.A., 2001. African Historians and African Voices: Essays Presented to Professor Bethwell Allan Ogot on His Seventieth Birthday. P. Schlettwein, Basel, Switzerland.

Ba, A.H., 1980. La tradition vivante. In: Histoire Générale de l'Afrique, vol. 1. UNESCO, Paris, pp. 191–230.

Barber, K., 1991. I Could Speak until Tomorrow. Oriki Women and the Past in a Yoruba Town. Edinburgh University Press, Edinburgh, UK.

Barber, K., Farias, M., Fernando de, P. (Eds.), 1990. Self-Assertion and Brokerage. Early Cultural Nationalism in West Africa. Birmingham University African Studies, Series 2, Birmingham.

Bayart, J.-F., 1989. L'État en Afrique. La Politique du Ventre. Fayard, Paris.

Bazin, J., 1979. La production d'un récit historique. Cahiers d'Études Africaines 73–76 19 (1–4), 435–483.

Chrétien, J.-P., 1997. Une crise de l'histoire de l'Afrique en langue française. Politique Africaine 68, 141–149.

Chrétien, J.-P., Prunier, G. (Eds.), 1989. Les Ethnies ont une Histoire. Karthala-ACCT, Paris.

Cissé, Y.T., Kamissoko, W., 1988. La Grande Geste du Mali des Origines à la Fondation de l'Empire. Karthala-Arsan, Paris.

Cissé, Y.T., Kamissoko, W., 1991. Soundjata, la gloire du Mali. Arsan Karthala, Paris.

Clancy, M.T., 1979. From Memory to Written Record. England 1066–1307. Arnold, London.

Cohen, D.W., 1989. The undefining of oral tradition. Ethnohistory 36, 9–17.

Cohen, D.W., 1994. The Combing of History. University of Chicago Press, Chicago.

Cooper, F., 1995. Work, class and empire: an African historian's retrospective on E.P. Thompson. In: Social History, pp. 235–241.

Coplan, D.B., 1991. Fictions that save: migrants' performance and Basotho national culture. Cultural Anthropology 6, 164–191.

Coplan, D.B., 1993. History in eaten whole: consuming trapes in Sesotho auriture. In: Mudimbe, V.Y., Jewsiewicki, B. (Eds.), History and Theory. Wesleyan University Beiheft 32, Middletown, CT, pp. 80–104.

Coquerel, P., 1988. Le verouillage ethnique en Afrique du Sud. UNESCO-OUA.

Coquerel, P., 1989. Les Zulu dans l'Afrique du Sud contemporaine. In: Chrétien, J.-P., Prunier, G. (Eds.), Les Ethnies ont une Histoire. Karthala-ACCT, Paris, pp. 417–425.

Coquery-Vidrovitch, C., 1997. Réflexion comparée sur l'historiographie africaniste de langue française et anglaise. Politique Africaine 66, 91–100.

Coquery-Vidrovitch, C., Jewsiewicki, B., 1986. Africanist historiography in France and Belgium. Traditions and trends. In: Jewsiewicki, B., Newbury, D. (Eds.), African Historiographies. What History for Which Africa? Sage, London, pp. 139–150.

Cuoq, J., 1975. Recueil des sources arabes concernant l'Afrique occidentale du VIIe au XVIe siècle (Bilâd al-Sûdân). Editions du CNRS, Paris.

Diawara, M., 1989. Women, servitude and history: the oral historical traditions of women of servile condition in the kingdom of Jaara (Mali) from the fifteenth to the mid-nineteenth century. In: Barber, K., Farias Moraes de, P.F. (Eds.), Discourse and Its Disguises. The Interpretation of African Oral Texts. Centre of West African Studies, Birmingham, pp. 109–137.

Diawara, M., 1990. La Graine de la Parole. Franz Steiner Verlag, Stuttgart, Germany.

Diawara, M., 1994. Der Blick vom anderen Ufer. Oder: Die Entdeckung des Weissen. In: Schlumberger, J.A., Segl, P. (Eds.), Europa—aber was ist es? Aspekte seiner Identität in interdisziplinärer Sicht. Böhlan, Cologne, Germany, pp. 255–283.

Diawara, M., 1997. Über die Bedeutung der Arbeit in der Welt der Mande. In: Spittler, G., Beck, K. (Eds.), Arbeit in Afrika. LIT, Munster, pp. 49–68.
Diop, C.A., 1960. L'Afrique Noire Précoloniale. Présence Africaine, Paris.
Diop, C.A., 1967. Antériorité des Civilisations Nègres: Mythe ou Vérité Historique? Présence Africaine, Paris.
Diouf, M., SEPHIS (Programme: Netherlands), & Centre for Studies in Social Sciences, 2003. Historians and Histories, What for?: African Historiography between the State and the Communities. SEPHIS, Amsterdam, Netherlands.
Duby, G., Lardreau, G., 1980. Dialogues. Flammarion, Paris.
È Nzièm, N., 1986. African historians and Africanist historians. In: Jewsiewicki, B., Newbury, D. (Eds.), African Historiographies. What History for Which Africa? Sage, London, pp. 20–27.
Falola, T., Aderinto, S., 2010. Nigeria, Nationalism, and Writing History. University of Rochester Press, Rochester, NY.
Farias, P.F.M., 1992. The oral traditionist as critic and intellectual producer: an example from contemporary Mali. In: Falola, T. (Ed.), African Historiography—Essays in Honour of Jacob Ade Ajayi. Longman, London, pp. 20–44.
Feiermann, S., 1993. African histories and the dissolution of world history. In: Bates, R.H., Mudimbe, V.Y., O'Barr, J. (Eds.), Africa and the Disciplines. The University of Chicago Press, Chicago, pp. 167–212.
Ginzburg, C., 1980. Cheese and the Worms. The Cosmos of a Sixteenth-Century Miller. Routledge & Kegan Paul, London.
Harries, P., 1993. Imagery, symbolism and tradition in a South African Bantustan: Mangusuthu Buthelezi, Inkatha, and Zulu history. In: Mudimbe, V.Y., Jewsiewicki, B. (Eds.), History Making in Africa, History and Theory. Wesleyan University Beiheft 32, Middletown, CT, pp. 105–125.
Henige, D., 1995. Homphaloskepsis and infantilizing of history. Journal of African History 36, 311–318.
Jaaresveld, F.A., 1992. Recent Afrikaner historiography. Itinerario 16 (1), 93–106.
Jenkins, R., 1997. Rethinking Ethnicity: Arguments and Explorations. Sage, London.
Jewsiewicki, B., 1979. L'histoire en Afrique ou le commerce des idées usagées. Revue Canadiénne des Etudes africaines 13 (1–2), 69–87.
Jewsiewicki, B., 1986. Introduction. One historiography or several. A requiem for Africanism. In: Jewsiewicki, B., Newbury, D. (Eds.), African Historiographies. What History for Which Africa? Sage, London, pp. 9–17.
Jewsiewicki, B., 1991. Le primitivisme, le postcolonialisme, les antiquités 'nègres' et la question nationale. Cahiers d'Études Africaines 121–122 31 (1–2), 191–213.
Jewsiewicki, B., Newbury, D. (Eds.), 1986. African Historiographies. What History for Which Africa?. Sage, London.
Ki-Zerbo, J., 1972. Histoire de l'Afrique Noire: d'Hier à demain. Hatier, Paris.
Levtzion, N., 1973. Ancient Ghana and Mali. Methuen, London.
Levtzion, N., Hopkins, J.F.P., 1981. Corpus of Early Arabic Sources for West African History. Cambridge University Press, Cambridge, UK.
Macamo, E.S., 1999. Was ist Afrika? Zur Geschichte und Kultursoziologie eines modernen Konstrukts. Duncker & Humblot, Berlin.
McIntosh, R.J., McIntosh, S.K., 1979. Terra-cotta statuettes from Mali. African Art 12 (2), 52–53.
Miller, J., 1980a. Listening for the African past. In: Miller, J. (Ed.), The African Past Speaks. Dawson-Archon, Folkstone, UK, pp. 1–59.
Miller, J. (Ed.), 1980b. The African Past Speaks. Dawson-Archon, Folkstone, UK.
Miller, S.A., Riding, I.J., 2011. Native Historians Write Back: Decolonizing American Indian History. Texas Tech University Press, Lubbock, Tex.
Mkandawire, P.T., 2005. African Intellectuals: Rethinking Politics, Language, Gender, and Development. Zed Books, New York.
Moniot, H., 1970. Les sources de l'histoire africaine. In: Deschamps, H. (Ed.), Histoire Générale de l'Afrique. T I, PUF, Paris, pp. 123–147.
Moniot, H., 1986. Profile of a historiography. Oral tradition and historical research in Africa. In: Jewsiewicki, B., Newbury, D. (Eds.), African Historiographies. What History for Which Africa? Sage, London, pp. 50–58.
Moniot, H., 1991. L'histoire africaine, un observatoire de la periodisation. Périodes: la construction du temps historique. Actes du Ve colloque d'histoire du temps présent. Éditions de l'École des Hautes Études en Sciences Sociales et Histoire au Présent, Paris.
Moniot, H., 1995. L'histoire à l'épreuve de l'Afrique. Cahiers d'Études Africaines 138–139 35 (2–3), 647–656.
Mudimbe, V.Y., Appiah, K.A., 1993. The impact of African studies on philosophy. In: Bates, R.H., Mudimbe, V.Y., O'Barr, J. (Eds.), Africa and the Disciplines. The University of Chicago Press, Chicago, pp. 113–135.
Mudimbe, V.Y., Jewsiewicki, B., 1993. Africans' memories and contemporary history of Africa. In: Mudimbe, V.Y., Jewsiewicki, B. (Eds.), History Making in Africa, History and Theory. Wesleyan University Beiheft 32, Middletown, CT, pp. 1–11.
Neale, C., 1986. The idea of progress in the revision of African history, 1960–1970. In: Jewsiewicki, B., Newbury, D. (Eds.), African Historiographies. What History for Which Africa? Sage, London, pp. 112–122.
Olivier de Sardan, J.-P., 1976. Quand nos Pères Etaient Captifs Paris.
Piault, M. (Ed.), 1987. La Colonisation, Rupture ou Parenthèse Paris.
Shoenbrun, D.L., 1993. A past whose time has come: historical context and history in Eastern Africa's Great Lakes. In: Mudimbe, V.Y., Jewsiewicki, B. (Eds.), History Making in Africa, History and Theory. Wesleyan University Beiheft 32, Middletown, CT, pp. 32–56.
Stock, B., 1983. The Implication of Literacy. Princeton University Press, Princeton, NJ.
Vansina, J., 1962. De la Tradition Orale. Musée Royal de l'Afrique Centrale, Annales, Sciences Humaines No. 36, Tervuren.
Vansina, J., 1985. Oral Tradition as History. University of Wisconsin Press, Madison, WI.
Vansina, J., 1992. Some perceptions on the writing of African history: 1948–1992. Itinerario 16 (1), 77–91.
Vansina, J., 1994. Living with Africa. University of Wisconsin Press, Madison, WI.
Wachtel, N., 1971. La Vision des Vaincus. Gallimard, Paris.
White, D.G., 2008. Telling Histories: Black Women Historians in the Ivory Tower. University of North Carolina Press, Chapel Hill.
Young, C.M., 1986. Nationalism, ethnicity and class in Africa: a retrospective. Cahiers d'Études Africaines 103, 421–495.

Historicism

Donald R Kelley, Rutgers University, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by P. Rossi, volume 10, pp. 6757–6762, © 2001, Elsevier Ltd.

Abstract

Historicism was a coinage of the later eighteenth century, when Novalis contrasted it with other methods of knowing (chemical, mathematical, artistic, etc.) and Schlegel associated it not with Kantian philosophy but with the modern science of philology. The term was occasionally used pejoratively in philosophical polemic but it blossomed in the work of the Historical Schools of language, law, and religion and in interpretation of the work of Giambattista Vico. It figured prominently in the early twentieth-century 'crisis of historicism' discussed by Ernst Troeltsch, Karl Mannhein, and others, and it was surveyed in philosophical terms by Friedrich Meinecke as a union of the principles of individuality and development. It continues today in scholarship, in both old and so-called new forms.

Historicism (German *historismus*, French *historisme*, Italian *storicismo*) refers to historical interpretation as opposed to abstract analysis of the past but it is also a term of constantly changing meanings and associations. The term is of Romantic origins associated first with the nineteenth-century German Historical Schools and then more generally with historical method as applied to all the arts and sciences and to human life as a whole. 'Historicism' (*historismus*) appeared first in a fragment of Novalis, who contrasted it with other methods (chemical, mathematical, artistic, etc.) and associated it with 'the system of confusion'. Contemporaneously, Friedrich Schlegel related historicism not to Kantian philosophy but to the modern science of philology. The word was occasionally used in philosophical polemics, and again the usage might seem pejorative since it was opposed to academic philosophy and especially unhistorical Kantian idealism. Felix Dahn argued, "historicism is above all a methodological moment, not a speculative principle…; its goal is [not philosophy but] life". Christlieb Julius Braniss opposed it to the reductionist and deterministic 'philosophy of naturalism', and in 1848 he distinguished between naturalism and historicism, which attempted to interpret all levels of existence in historical terms. In 1879, Karl Werner applied the phrase 'philosophical historicism' to the work of Giambattista Vico. This connection was later endorsed by Benedetto Croce, Friedrich Meinecke, Erich Auerbach, and others.

In 1895 in his inaugural lecture on the study of history, Lord Acton pointed to "that influence for which the depressing names historicism and historical-mindedness have been devised" – 'all things', for him, including law, theology, science, and philosophy itself. Yet he also wrote, "History is not only a particular branch of study, but a particular mode and method of knowledge in other branches", and "The great point is the history of history". Yet by the late nineteenth-century historicism had acquired a largely pejorative meaning because of its associations with relativism and the threats posed to the assumptions and values of philosophy, theology, and economics, which represented three of the absolutes of Western culture, nineteenth-century field of economics, especially in 1883 by Karl Menger's liberal assault on 'the errors of historicism', that is, the irrational methods of Gustav Schmoller and others members of the so-called younger historical school of economics. In a sense, this war of methods (*Methodenstreit*) recapitulated the struggles between the historical and philosophical schools in the nineteenth century. Now, however, such struggle deployed the weaponry of modern positivism and quantitative techniques. Liberal, market-based rejection of historicism, indeed of historical method in any sense, has persisted in many areas of social sciences, as well as in the humanities, as in Russian formalism, structural linguistics, and the New Criticism.

In theology the errors of historicism seemed equally offensive. In the pre–World War I years, religious controversies raged around the various forms of 'modernism', which had been denounced in Pope Pius IX's 'Syllabus of Errors' of 1864, and were accompanied by a whole list of other secularizing '-isms'. Positivism, psychologism, and historicism all posed threats not only to traditional moral values but also to the validity of both reason and revelation. The historicizing of Christianity was the avowed aim of the historical school of religion (*religionsgeschichtliche Schule*) headed by Max Weber's friend Ernst Troeltsch, whose classic work, *Die Soziallehren der christlichen Kitchen und Gruppen* (1912), marked the intersection of the socioeconomic and theological problematics of historicism; and this discussion too, was carried on with heavy anxiety about questions of traditional norms and the elimination of metaphysical and metahistorical absolutes.

Philosophers had worried about the threat of history since the late eighteenth century, and their anxieties resurfaced in the early twentieth century as a 'crisis of historicism'. In 1910, Edmund Husserl denounced historicism as the enemy of 'philosophy as a rigorous science'. Martin Heidegger, following Friedrich Nietzsche's famous critique of 'the use and abuse of history', contrasted it with true historicity (*Geschichtlichkeit*) grounded in contemporary existence. Like Braniss, Troeltsch opposed 'historicism' to 'naturalism' and traced this war of methods back to the seventeenth century with special reference to Vico's attack on Cartesianism; and he celebrated historicism for removing the 'dead hand' of dogma while at the same time fearing the threat it posed to philosophical and moral tradition. Following Troeltsch, Karl Heussi distinguished between 'history for the sake of history' (*l'histoire pour l'histoire*),

relativism, radical evolutionism, and speculative philosophy of history and analyzed them all under the rubric, 'the crisis of historicism'.

Historicism represented a problem for all the human sciences and, for Karl Mannheim, it constituted a complete *Weltanschauung* beyond the level of conscious reflection or ideology, so that history itself was caught in its net. It served the modern dynamic world as 'timeless reason' had served a more static world. "Historicism has developed into an intellectual force of extraordinary significance", he says; "it epitomizes our *Weltanschauung*… The historicist principle not only organizes, like an invisible hand, the work of the cultural sciences (*Geisteswissenschaften*), but also permeates everyday thinking." For Mannheim, historicism was a product not only of 'mere' historiography but of 'problems involved in the new ways of facing life'. Yet like Meinecke, Mannheim's purpose is to situate historicism within the confines of formal philosophy. As a counterpart to the older faith in reason, now itself historicized, historicism also needed to be the object of critical theory. Yet historicism was not absolute, but rather bound to assumptions and spatial, temporal, and material conditions that had undercut universalist conceptions. The virtue of historicism for Mannheim was that it set dynamism at the center of its conceptualizations instead of relativizing it as in 'the old static system'. In this way, historicism became 'the Archimedean lever' for the modern worldview and life experience, that is, made it in effect the condition of human, historical, and indeed philosophical understanding.

Croce, following his predecessors Vico and Hegel, thought that the Germans had not pushed historicism far enough. He rejected the claims of modern social science, such as those of Weber and Emile Durkheim, to universal status, regarding them all as open to contingency and subject to historical conditions. Human values had the same continent status, and Croce had no fear that relativism was a major threat, since for him historicism was "a logical principle…, the very category of logic." In Germany, historicism was the reigning condition of contemporary thought and life – and according to his famous aphorism, "every true history is contemporary history". Indeed philosophy itself was 'absolute historicism', so that historicism was not the source of the intellectual crisis of the twentieth century but rather its potential solution. Croce's views and prejudices were carried over into the Anglophone world by R.G. Collingwood, who translated his work and followed his interpretation of historiographical tradition and who turned back to the classical view of history as a form of 'inquiry' into human behavior, proceeding through the interpretation of evidence and aiming ultimately at 'human self-knowledge'.

The best known interpretation of historicism was Friedrich Meinecke's *Die Entstehung des Historismus* (1936; trans. *Historicism*: *The Rise of a New Historical Outlook*), which offered a comprehensive map of European historical thought since the seventeenth century and gave systematic and philosophical form to the Rankean principles of individuality and development. As 'forerunners' he recognized the Lord Shaftesbury, Gottfried Wilhelm Leibniz, Gottfried Arnold, Vico, and Joseph-François Lafitau, who all reflected the rejection of natural law, abstract reason, and universal laws as the primary way to understand life over time. Historicism was the product of 'the loosening up of dogmatic and ecclesiastical thought after the wars of religion' and the rise to prominence of the key principles of individuality and development. Vico was important but also a 'solitary and unrecognized figure' and continued to assume that God ruled the world, though he realized that nature was a process of development, which had its own laws. Lafitau contributed to the historicist outlook by his comparative study of Native American Indians with savages of antiquity.

According to Meinecke, Voltaire was the first great cultural historian, and he published his 'essay' on universal history of civilization and the 'human spirit' in 1769. The restriction of his point of view was shown in his study of the age of Louis XIV, which he regarded as superior to any previous age in history. Another candidate for primacy was Montesquieu, who applied 'empirical realism' to the science of politics, though his notions about geography and climate reintroduced a kind of determinism that Meinecke found retrograde. Yet Montesquieu had significant views about the ideas of individuality and development so crucial to Meinecke's idea of historicism. As Voltaire spoke of the 'spirit' of peoples, so Montesquieu spoke of the 'spirit' of the laws, though Voltaire had more of a historical perspective. Other significant French historians of that age included A.R. Turgot, Antoine Condorcet, Antoine Goguet, and Nicolas Boulanger. These were exceptional figures, in Meinecke's view, since unfortunately the abstract notions associated with the French Revolution "froze history and historical thought into a fixed and static state".

For Meinecke, British contributors to historicism started with David Hume, who was a philosopher before he was a historian and had already begun passing beyond Enlightenment rationalism to an appreciation for the variety of human nature over space and time. Hume also rejected religious belief for a considered Deism, and he treated English history as a development. Edward Gibbon showed a 'world historical grasp of his subject', and he had a sense of 'the community of peoples closely bound together by their intellectual outlook and their destinies'. The third great historian of the English Enlightenment was William Robertson, whose work 'illustrates a certain progress in the handling of universal history'. Robertson also had a good idea of the development of peoples from a primitive stage. In the great intellectual transformation leading to the Romantic movement ideas of the two great classics of Western tradition, Homer and the Bible, also marked important changes. In 1766, Adam Ferguson, friend of Adam Smith, published a book on the history of civil society, says Meinecke, "along sociological and positive lines" from a primitive level to that of the high culture of the Enlightenment, and some of his remarks constituted 'a sign of a new age of historical thought'. A greater figure contemporary to Ferguson was Edmund Burke, who (like Justus Moser in Germany) tried 'to infuse new life into the nature of the state'. He saw the dangers of the ideology of natural law and revolution and turned to traditionalism as the key to history. According to Meinecke, Moser tried to read the effaced writing of the past in the palimpsest of the present, though for Meinecke, this was not yet quite full-blown historicism.

For Meinecke, the first large-scale expression of historicism came in Germany, and again he attends to familiar names – four main figures – of German thought to illustrate his argument. G.E. Lessing had a great appreciation of the principles of individuality and development, shown especially in his work

on the education of the human race (1780). However, the developing character of historicism was suggested still better in J.J. Winckelmann's *History of Ancient Art* (1764), which maintained contact with Herodotean *historia* and which referred to the history of the history of art. Moser was concerned with local and antiquarian history and thus to the individualizing aspect of historicism, and in his theory of 'total impression' he showed an awareness of individual psychology and social patterns, or cultural synthesis and continuity, in history. Karl von Irwing (not mentioned by Meinecke) was among the first to adopt (in 1779) the terminology and conceptualization of 'culture' (derived from the Latin *cultura*, meaning the cultivation or mastery of material or spiritual activities), and he posited a series of evolutionary stages from the state of nature to mature civilization expressed most fully in the nation-state and the arts and sciences necessary for this institution. In the late eighteenth-century Germany 'culture' (*Kultur*) came to replace 'spirit' (Geist, esprit) as a way of expressing the character of society, and it was adopted by a number of scholars who restored the juncture of learning and thought, although Meinecke discusses only J.G. Herder.

According to Meinecke, the greatest of the pioneers of historicism in the early years were Moser, Herder, and Goethe. Herder began as a literary critic and ended as a philosopher of history. His early work was on primitive poetry, myth, and language, but his efforts then expanded into larger studies of national life, especially in Germany, and as Meinecke remarked, he "portrayed himself in the character of Petrarch". In 1774, there appeared his preliminary study of the philosophy of history, and then his major work on the topic 10 years later, which extended the ideas of national individuality and development not merely to Germany but to a whole range of nations. In his organic view of history, shared with other Enlightenment scholars, he proposed a conception of several progressive stages of development. He simultaneously shifted from describing the 'spirit' of a people to describing a people by its 'culture'. At the end of his life, and under the influence of Hamann, he reinforced his distrust of rationalism with a 'metacritique' of Kantian abstract metaphysics by insisting that 'One thinks with words' and philosophy loses its human base when it tries to go beyond language – and so history. The proper question was not 'what is reason?' but 'what is humanity?' To his mind, of course, humanity had to be understood in terms of its origins, languages, and development. Meinecke does not mention this work, nor is it much appreciated in histories of philosophy.

"Goethe was only marginally a historian," Meinecke admits, but "Without Goethe we should not be where we are." For Goethe too, as 'Herder's pupil', history was part of nature and also had 'laws of development'. For him, 'the past and the present are one', and Meinecke examines Goethe's views on art, literature, life, travels, and world events at the time of Napoleon to show his adherence to the principles of individuality and development. For Goethe, 'history must always be rewritten'. This appreciation of Goethe concludes the survey of early historicism offered by Meinecke, who overlooks the linguistic and literary – and indeed in many ways the antiphilosophical orientation of historicism as it emerged in the age of Romanticism. Meinecke also neglected the contributions of the Historical School, Barthold Georg Niebuhr, J.G. Eichhorn, Karl Friedrich von Savigny, Jakob Grimm, Wilhelm Roscher et al., and the scholarly tradition as a whole. By reducing historicism in effect to a philosophical doctrine Meinecke violates the original impulse, for in general historicism was not a theory but a scholarly practice and a mode of interpretation. It was part of an effort not to make history into a philosophical doctrine but rather to transform it into a foundational discipline to which philosophy itself could be subject.

In 1796, J.G. Eichhorn joined cultural and literary history in a way that showed the connection between literature or language and history in a prophetic statement: "The history of arts and sciences, their origins, progress, and various transformations can never be separated from the history of the social conditions, for culture and literature are twin sisters, children of the same father, who continually provide support for each other. Culture, the first-born, prepares the birth of her younger sister, and thereafter they live and work together, inseparably and unseparated, and also die together. Without the history of one the life of the other is incomplete and incomprehensible." Eichhorn was one of those scholars who inaugurated the 'higher criticism' of the Bible, which was followed by the mythological criticism of David Friedrich Strauss. Biblical criticism, too, became an adjunct of the Historical Schools. In general, as G.P. Gooch put it, "The scope of history has gradually widened until it has come to include every aspect of the life of humanity."

One of the main products of historicism in full flower was the Prussian School of historical writing, of which Heinrich Treitschke, son of an ennobled Saxon general, was perhaps the leading and most passionate figure. Treitschke was a supporter of Bismarck and in 1874 he moved to the University of Berlin. After Treitschke, came Leopold von Ranke, who took a more neutralist attitude toward history, and turned the agenda of historicism toward political history. "The chief creation of more recent history," remarked Burckhardt, "is the Great Power, the life form of the most significant peoples." In the nineteenth century German historians sought continuity in their national tradition and the character of their people and institutions, and they were frequently followed in this by other national historians, including France, England, and the United States toward political history, and Burckhardt, whose work helped turn historical studies to cultural and artistic history. The older 'new history' in Germany, France, and the United States (though not England) continued this culturalist turn in the work of Karl Lamprecht, James Harvey Robinson, and later the Annales school. Meinecke and Troeltsch came to the University of Berlin in 1914 and 1915 (in history and philosophy), respectively, and their views were central to the discussion of the 'crisis of historicism' in the 1920s.

The later semantic history of historicism has become increasingly muddled because of misappropriations of the word for philosophical or ideological but quite unhistorical or even antihistorical purposes, beginning with Karl Popper's Poverty of Historicism (1957), which identifies historicism with naive biological determinism (precisely the opposite of the view of Braniss) and a total innocence of history, experience, and even common sense. Another instance is the book of Maurice Mandelbaum, who writes in *History, Man, and Reason*, "Historicism is a belief that an adequate understanding of the

nature of any phenomenon and an adequate assessment of its value are to be gained through considering it in terms of the place which it occupied and the role which it played within a process of development." Michael Gillespie, along similar lines, thinks historians go 'critically astray' in tracing the source of historicism to the attitudes associated with historical scholarship, especially that of early modern Europe. On the contrary, historicism was and is an alternative to conventional philosophy and scientific naturalism, and it flourished in the rich soil of literary and antiquarian learning and the teachings of the Historical Schools. In fact historicism is positioned in opposition to the scientisms that evade questions of point of view, perspective, cultural context. It emphasizes the necessity of interpretation in posing questions about a past that is not only a 'foreign country' but also largely inaccessible except through traces and testimonies which happen to have survived and which must be expressed in the language and conditions of the present – a present that is itself on its way soon to becoming a past.

The so-called New Historicism, popularized since 1980, goes back at least to 1942, when Croce wrote, "The New Historicism accepts, extends, and applies Vico's principle that men know only what they do." The newer New Historicism, though avoiding philosophical claims, likewise turns, as Vico had done, to cultural forms, especially literature and art, as a way of revealing underlying interests and power structures and penetrating to inarticulate levels and marginal groups of society. In general, this New Historicism was part of the linguistic and textualist turn taken by history and the human sciences, including philosophy, in the past generation. Included too were the closely allied innovationist movements that call themselves 'the new cultural history' and 'a new philosophy of history'. Despite attempts to deprecate the Old Historicism on ideological as well as methodological and terminological grounds, the New Historicism shared or inherited some of the same assumptions and goals and in a long perspective forms part of the same story.

See also: Cultural History; Darwin, Charles Robert (1809–82); Enlightenment; Hegel, Georg Wilhelm Friedrich (1770–1831); Hermeneutics; History, Professionalization of; History: Theories and Methods; Intellectual History; Kant, Immanuel (1724–1804); Mannheim, Karl (1893–1947); Wundt, Wilhelm Maximilian (1832–1920); von Ranke, Leopold (1795–1886).

Bibliography

Acton, John Emerich Edward, 1985. In: Fears, J. Rufus (Ed.), Essays in the Study and Writing off History. Indianapolis.
Babbach, Charles R., 1995. Heidegger, Dilthey, and the Crisis of Historicism. Ithaca.
Collingwood, Robin G., 1993. The Idea of History. Clarendon Press, Oxford.
Croce, Benedetto, 1941. History as the Story of Liberty (S. Sprigge, Trans.). Norton, New York.
Croce, Benedetto, 1960. History, Its Theory and Practice (D. Ainslie, Trans.). Russell, New York.
Croce, Benedetto, 1966. Philosophy, Poetry, History (C. Sprigge, Trans.). Oxford University Press, Oxford.
Gillespie, Michael A., 1984. Hegel, Heidegger, and the Ground of History. University of Chicago Press, Chicago.
Heussi, Karl, 1932. Die Krisis des Historismus. Tübingen.
Iggers, Georg G., 1983. The German Conception of History: The National Tradition of Historical Thought from Herder to the Present. Wesleyan University Press, Middletown, CT.
Mandelbaum, Maurice, 1971. A History, Man, and Reason Study in Nineteenth-Century Thought. Johns Hopkins University Press, Baltimore.
Mannheim, Karl, 1952. Historicism. In: Kecskemeti, P. (Ed. and Trans.), Essays on the Sociology of Knowledge. Routledge and Kegan Paul, London.
Meinecke, Friedrich, 1972. Historism: The Rise of a New Historical Outlook (J.E. Anderson, Trans.) Routledge, London.
Page, Carl, 1995. Philosophical Historicism and the Betrayal of First Philosophy. Pennsylvania State University Press, University Park.
Popper, Karl, 1957. The Poverty of Historicism. Boston.
Roberts, David, 1987. Benedetto Croce and the Uses of Historicism. Berkeley.
Thaden, Edward C., 1999. The Rise of Historicism in Russia. Peter Lang, New York.
Thomas, Brook, 1991. The New Historicism and Other Old-Fashioned Topics. Princeton University Press, Princeton.
Troeltsch, Ernst, 1922. Historismus und seine Probleme. Tübingen.
Troeltsch, Ernst, 1960. The Social Teaching of the Christian Churches (O. Wyon, Trans.). Harper, New York.

History and Literature

Paul M Lützeler, Washington University, St. Louis, MO, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 6817–6822, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

Ever since Aristotle, theoreticians have analyzed the differences and commonalities of narration in historiography and literature. Both in the modern novel as well as in modern historiography one has tried to get rid of narration as a basic principle. Narration survived all new experiments in the two areas. The interrelation of factual and fictional narration is on the mind of many theoreticians these days. Hayden White's thesis of the poetic structure of historiography has been discussed in a controversial manner. In order to point out the differences in the two modes of narration one has to understand the varying truth concepts behind historiography and the novel. As Käte Hamburger has pointed out, it is the as-if structure of fiction that makes the novel different from historiography. What both forms of narration have in common are the creation of meaning, the function of memory, and the formation of identity.

Literature and history have many things in common and one of them is the aesthetic form of narration. The many contemporary experiments notwithstanding, modern historiography and the modern novel still share the common feature of narration. In narratology, some theoreticians have differentiated between the narrative modes in the two fields; others have shown how they conflate and that they can hardly be distinguished. This article wants to show what narration in literature and history has in common and in what respect the two modes differ from each other.

For Aristotle (see Chapter 9 of his *Poetics*) the matter of the division of labor between historiographer and poet was easily decided: the historiographer relates 'what has happened' and the poet 'what might happen.' In contrast to Plato's devastating criticism of literature, Aristotle considered poetry to be of high value. In his eyes, literature dealt with problems of a more general and thus more significant nature, while the historian had to report about specific, and thus, less important matters.

Until approximately 1750, historiography was considered a branch of literature. Since the eighteenth century (when Aristotle's distinctions were rediscovered), the differences between the two disciplines have been stressed, and it was only then that the term 'literature' became synonymous with 'poetry' and 'fiction.' Gottsched and Lessing still shared Aristotle's prejudice against history. For them the field of history was merely a quarry from which they mined some of the raw material they used to build their preconceived literary plots. Of real importance to them was a moralistic story, and examples were taken from the repertoires in the areas of history, legend, and mythology. However, the late eighteenth century is an important turning point. As Koselleck has demonstrated, it is the time of the development of a genuinely new philosophical awareness about history (Koselleck, 1976). The concept of a continuously progressing history as such replaces both the religious concept of a Christian teleology as well as the idea of history as a chaotic conglomerate of countless individual stories of an exemplary nature. Through this philosophical development, literature is seen as part of history and thus it loses its Aristotelian superiority.

Today, the relationship between fiction and historiography poses a number of hermeneutic questions, and there is no modern Aristotle who might offer a simple and acceptable solution fitting for our time. Narration can be viewed as one of the primary tools of knowledge and it determines the structure and aesthetic form in the presentation of a real or a fictional event. Narration discloses the dimension of historical reality as well as the sphere of poetic fiction. The logic of narration provides access to the logic of historiography and fiction; it makes the structure of both historical consciousness and fantasy tangible. In every narrative, real or fictional events are supported by a structure through which happenings are accentuated, selected, or eliminated according to the law of relevance. Neither in historiographical nor in fictional narration are we dealing with a reproduction or duplication of events, but with a certain organization of actual or imaginary occurrences or experiences.

Narration in Historiography

Perhaps because it is so difficult to liberate it from its established forms, narration is constantly under attack. In the discussion of modern historiography from 1970 to 2000, it looked as if 'narration' would be replaced by 'description' and 'explanation.' 'Historiography,' Szondi argued, must "divest itself increasingly of its narrative character"; narrative historiography must "turn into description if a new concept of history is to do justice to our knowledge of history as an anonymous process, as a sequence of events and changing systems" (Szondi, 1973). The arguments of the antinarrativists can be summarized in the following way: the narrative recording of events in historicism was based on the belief that history was made by individuals. However, this categorization did not adequately take into account the development of social processes. Processes and structures can only be described not narrated. Thus the old *histoire événementielle* has to be replaced by the new *histoire structurale*. 'The nouvelle histoire' of French historians in particular, with their descriptions of structures underlying long periods of time, and their quantifying methods, forced the old paradigm of writing history by narrating events into a defensive position (Furet, 1982; LeGoff and Nora, 1974).

However, according to some historiographers, history as a narration of events and history as a description of structures are not mutually exclusive. These historiographers concede that history as a narration of events is no longer the center of the historiographer's work and that it no longer provides the obvious support of historical writing. Nevertheless, they also concede that narration is allowed to stand as a possibility alongside the writing of history as a structural analysis of processes (Borst, 1976). Other scholars will have nothing to do with a strict separation of the two methods and point out that narration and description are intertwined. The argument is that the fact that modern history is a series of processes, it can only be understood through the alternate explanation of events through structures and vice versa (Koselleck, 1976; Kocka, 1984). Although the structures become more accessible through description and events through narration, nevertheless in practice it would be impossible to maintain a demarcation between the two methods. The complementary characters of description and narration are increasingly emphasized (Stone, 1979).

The defenders of the descriptive method like to confront the narrators among the historiographers with their lack of theory. They assert that the historiographer must not only narrate but also explain, and that this is only possible by means of a theoretical concept of the realities, as it has long been practiced by social scientists. The proponents of the narrative method in writing history defend themselves vehemently against the charge of being antitheoretical. From their initially defensive position, they have turned to attack. Danto has asserted that narration is vital for description and explanation; that the form of narration is at the same time a form of explanation. The narrative form always surpasses the given and categorically uses all-encompassing concepts. Narrative structures influence our historical thinking in the same way theories do in the case of the sciences (Danto, 1965). Ricoeur expresses the same idea in his phrase "to tell a story means to explain it" (Ricoeur, 1983). Rüsen agrees that narration is a part of the logical structure of historiography and that consequently "description and explanation have a function subordinate to the narrative scheme" (Baumgartner and Rüsen, 1982).

Narration in Literature

In connection with Gottfried Benn's poetics, Grimm described a type of novel from which the narrator had disappeared. It was called "the novel of the phenotype" or "the inward directed novel," labels indicating a tendency toward lack of action and lack of characters (Grimm, 1963). The world and its events are banned here. The reduction of the 'I,' the isolated subject, is significant for this novel: the 'I' speaks only of itself. Consequently, one reads in Rilke's *Aufzeichnungen des Malte Laurids Brigge* from 1910: "It must have been before my time that stories were narrated, really narrated. I have never heard a person narrate a story." Benjamin's discussion of the decline of storytelling in his essay *Der Erzähler* about the Russian writer Nikolai Lesskow echoes Rilke's statement (Benjamin, 1972).

The aesthetic designation of the phenotype novel – exemplified in the novels of Virginia Woolf, Benn, and Rilke – is also appropriate for the writing of some contemporary authors; the authors of the French *nouveau roman* (like Alain Robbe-Grillet) as well as postmodern American writers (like William Gass) and contemporary German authors (like Peter Handke) come to mind. In the novel of the avant-garde (Joyce, Dos Passos, Gide, Döblin, Musil, Broch), narration was either complemented or largely replaced by montage techniques and cognitive means, by reflection or essayistic writing. Here it was not so much a matter of narration of events, action, and mimesis as one of the presentation of ideas and the provision of knowledge (Hinck, 1995).

However, the best-known postwar novelists (like Garcia-Marquez, Vargas Llosa, Philip Roth, Bellow, or Grass) uphold traditional established narrative models. One can write purely narrative novels. However, nonnarrative epic fiction is nonexistent. Like historiography, the novel obviously cannot dispense entirely with narration and thus it makes sense, now as ever, to adhere to the science of narrative technique (Lützeler, 1986). In the field of narrative technique, literary scholars are far ahead of the narratologists in the camp of the historiographers. Historians could learn from reader theories as developed by such literary theorists as Jonathan Culler ('the competent reader'), Umberto Eco ('the model reader'), Stanley Fish ('the informed reader'), Wolfgang Iser ('the implied reader'), Michel Riffaterre ('the superreader'), and Gerald Prince ('the narratee') (Culler, 1975; Eco, 1979; Fish, 1970; Iser, 1974; Riffaterre, 1971; Prince, 1982).

The Conflation of Historiography and Fiction

In *Metahistory*, Hayden White draws interesting parallels between historiography and fiction. He strives for a poetics of historiography and attempts to describe the forms of historical presentation in order to compare them with literary forms (White, 1973). White starts with the premise that no sequence of events – by itself – is a story; it is the task of the historiographer to transfer the events, the facts, to the narrative framework of a story. The connections between happenings do not occur spontaneously; they are the result of the historiographer's reflection. In each historiographical text, White distinguishes between the 'what' (the narrated events) and the 'how' (the type of narration, the 'emplotment'). He was not the only scholar to see that historiographical and fictional narration are related, but he was the first to make a systematic attempt to study the structure of historical consciousness – to show the fictional elements and their formal manifestations in historiography.

The situation in nineteenth-century historiography and philosophy of history serves as an example. White could have given his book the title *The Rhetoric of Historiography* as a parallel to Wayne C. Booth's 1961 study *The Rhetoric of Fiction*. In *Metahistory* he concludes that the writing of history has little to do with such qualities as true or false, correct or incorrect, and that ultimately historical writing cannot be distinguished from fiction. The historian's style, White argues, depends on the combination of the following elements: first, there are three different strategies; he calls them 'explanation by formal argument,' 'explanation by ideological implication,' and 'explanation by emplotment.' Specific modes of articulation correspond to each strategy. In this context we are only interested in the

strategy of 'emplotment' (narration) with its four corresponding 'poetic' modes of 'romance,' 'comedy,' 'tragedy,' and 'satire.'

White's references to certain structural similarities between the unfolding of tragedies or comedies and that of historical representations are interesting, but these similarities should by no means be construed as an actual fictionalization of history. Demandt has shown how thought structures were mirrored in the poetic metaphors used by historiographers, and how, in turn, the ideologies of the historiographers influenced their use of rhetorical figures (Demandt, 1978). However, the use of metaphors does not signify a fictionalization of history. There exist basic narrative concepts that are equally valid for historical and for fictional narration. Without action, images, and metaphors, narration per se is unimaginable.

Returning to White's example of the different depictions and interpretations of the French Revolution, it is evident that this is not a case of fictionalization because the historical process is not depicted as an 'as if,' but as something that actually did occur. No one will deny that the depiction of history is a creative process. However, its qualitative difference from fictional narration remains, despite the use of literary interpretation patterns and flowery metaphors. While fiction has as its basis the so-called 'as-if' structure, and thus proves to be pretense, illusion, and unreality, historiography always relates to actual events. Hamburger clarified the difference through the example of a historical report about Napoleon and a novel about the French emperor. The historiographer cannot see Napoleon 'in the subjectivity, the ego-originality of his inner thoughts, of his very "existence".' If he does so, however, we "find ourselves in a novel about Napoleon, a piece of fiction" (Hamburger, 1968).

The narrative techniques with which novelists depict the psyches of their characters have been analyzed by Dorrit Cohn. She differentiates between psychonarration, quoted monologue, and narrated monologue (Cohn, 1978). These are narrative techniques reserved for fictional narration, which cannot occur in historical writing. Just as the use of a literary (tragic or comic, for example) interpretive category cannot turn a historical narrative into a fictional one, a historical novel does not become a piece of historiography simply by including historical reality in the fictional field. Käte Hamburger validly defines the line between fiction and reality, 'where there is no crossing over from one category to the other,' when she discusses the different perceptions of mimesis in novel and historiography. According to her, there is a definite relationship between the narrating process and the narrated material in historiography, while characters created by novelists are simply narrated persons. In fiction, there is no relationship, only functional context, between the narrating process and the narrated material. In other words, the fictionalization of the persons described consists in the fact that the characters are not presented as objects but as subjects, as "ego-origines" (Hamburger, 1968).

This difference between historiography and fiction also becomes evident when we look at the results of more recent studies of the theory of functionality. Assmann has shown that scholars in the fields of analytic language philosophy and of pragmatic communication theory insist on this distinction (Assmann, 1980). In the analytic philosophy of language the point is made that fiction has a very peculiar status as far as its logic of statements is concerned: the authors of fiction are not obliged to come up with any proof; their judgments are quasi, pseudo, or as-if statements. In communication theory, fictionality is studied with respect to its reception effects: it involves the elimination of the normal connection that exists between the participants of an act of communication. Dealing with fiction leads to a temporary liberation from thinking in terms of fixed institutions; it means the absence of social sanctions and a lack of verification.

The Concepts of Truth

Historia vero testis temporum, lux veritas, vita memoriae, magistra vitae, nuntia vetustatis: Cicero's classic definition of the function of historical writing is still valid today – with a few deviations and reservations. The modern formulations are somewhat less pretentious, but given such terms as 'demand for truth,' 'imparting of meaning,' 'provision of identity,' and 'memory,' the function of historiography is not defined so differently from Cicero's wording. Only the postulate *magistra vitae* is viewed more skeptically. Authors of novels are equally concerned with the pursuit of truth, the imparting of meaning, the provision of identity, and memory. Friedrich Schlegel wrote in his introduction to the *Wissenschaft der europäischen Literatur*: "As history aims at recognition and truth it approaches science, but as it is also representation (…) it is related to art." Here the double-faced aspect of historiography has been precisely formulated.

No historiographer would be so arrogant as to assume that his or her report could encompass the whole truth of a historical event. Historiography can claim with justification that its statements are 'truthful' insofar as they have been gained 'scientifically,' that is in so far as they can be proven. And they can be proven because they can be proven wrong: the truth capacity of the narrative statements of historical writing is based on the fact that they can be falsified. In this respect, one can imagine no greater difference between historical writing and fiction.

Since the statements of a novel are fictional to begin with, they cannot be falsified. While writers of fiction may embellish an event from the past with fantasy, historiographers must check the sources, which dictate what they cannot say. The truth of a poetic statement is not achieved by fiction imitating the principles of historiography. The propensities for truth and truthfulness itself are presented differently in the two narrative modes. The intent to provide the truth does exist in both literature and historical writing, but it is aimed in different directions. Historiographers attempt to achieve the most exact comprehension and representation of actual historical events; novelists have no such ambition. On the contrary, their strength lies not in the realm of facts but in that of possibilities.

The aesthetics of the modern as well as of the postmodern novel are focused on the opening up of possibilities. Modernist literature aims to distance itself from the mimetic realistic principle that dominated the novel of the second half of the nineteenth century. Musil, for example, in *Der Mann ohne Eigenschaften*, plays with several types of utopias and arrives at the category of 'possibility,' in opposition to confidence in the power of history. The same is true in the postmodern novels of Thomas Pynchon.

The Concept of Totality

In examining the truth postulate in literature and historiography, one must consider the premises on which historical and novelistic narration are based. In traditional narration the novelist, like the historiographer, is guided by ideas of totality as we know them, on the one hand from the aesthetics of Hegel to Lukács, and on the other hand from Hegel's and Marx's substantialist philosophy of history. In both cases, the attempt is made to see events in such a way that they make sense within the context of a totality. Here the historical and artistic totalities are structurally identical. Danto has criticized as outdated the substantialist philosophy of history, with its concept of a teleological, continuous development. He states that this philosophy acts as if it knew of a 'divine plan' of history, as if one could determine past and present from the perspective of the future. Danto wants to replace the speculative and dogmatic substantialist philosophy of history with an analytical (descriptive and coordinating) theory of history. In this theory, the vantage point would not be from a postulated future goal back to the present and past, but rather from the past to the present, with no certainties about the future (Danto, 1965). The ancient trust, reaffirmed by Nietzsche, in the continuity of history is questioned, as are the Christian and Marxist tenets of the eschatological or utopian direction of history.

Such beliefs once formed the foundation for the concept of continuity in history. Many contemporary historians shun these labels and are consequently faced with two choices: giving up the concept of continuity altogether, or adopting it from the tradition of historiography and perhaps developing it in their own fashion. Historians at present are more aware than ever of the heuristic character of the concept of continuity; they see it merely as a 'regulative idea' in historical narration (Rüsen, 1982). Today, the focus is more on discontinuity. In this respect the impact of Foucault as an outright anti-Hegelian thinker (especially with his 1969 study *L'archéology du savoir*) is not to be underrated.

The dismissal of the concepts of continuity and totality has also had an impact on the writing of the novel. The idea of the 'self-contained' work of art that provides totality by way of typical characters is dead. Its place has been taken by the 'open' novel with its open-endedness, as first envisioned by the early German Romantics. In his *Theorie des Romans* (1916), Georg Lukács declared that the modern novelist could outdo the historiographer if, after the loss of totality of meaning and being, he or she would counter the partial insights into the course of history with concern for a new, homogeneous cosmology. In contrast, however, the modern novel plays an important role as an instrument of cognition because such concepts of systems are now distrusted. Not a new cosmology of being, but a view of the disparity of what is to be, of the diversity of possibilities, is what this medium can provide.

In twentieth-century philosophy, a similar discussion to the one between Plato and Aristotle about the truth value of fiction took place. Adorno was convinced that, due to its autonomy, art 'as a creative, nonalienated work' was the last vestige of freedom. According to him, art and literature alone can unveil what is hidden in reality and can destroy ideology and lies (Adorno, 1970). What Adorno criticized as a Web of ideological deceptions is close to what Roland Barthes calls the 'mythicizing' of everyday life. Like Adorno, Barthes affirms that poetic language maintains direct contact with reality and, unlike historiography, escapes mystification (Barthes, 1957). In contrast to Adorno and Barthes, Eagleton and Jameson argue that the writing of literature is just another form of ideological production (Eagleton, 1976). Jameson maintains the following position: "Ideology is not something which informs or invests symbolic production; rather the aesthetic act is itself ideological, and the production of aesthetic or narrative form is to be seen as an ideological act in its own right" (Jameson, 1981).

All four authors make extremely one-sided statements. The common ground of historiography and literature is narration. And when narration, through the eyes of Jameson, is seen as an ideological act per se, we cannot expect much truth in either historiography or fiction. But if, on the other hand, literature is able to destroy ideology and deception – to cite Adorno – this virtue should also be attributed to historiography. Narration can lead in both directions: it can help to approach truth or it can produce ideological veils. Neither literature nor historiography is immune to the virus of error or the distortion of truth, but both are capable of making reality be seen more clearly.

Meaning, Memory, and Identity

Among the functions of historiography and the novel are, along with the search for truth, the imparting of meaning and the provision of identity. The imparting of meaning (*Sinnbildung*) is far from the creation of meaning (*Sinnstiftung*) (Rüsen, 1982). Surely there were, and still are, historiographers and novelists who have declared themselves ideological creators of meaning, but neither modern novelists nor critical historiographers are interested in such ambitious goals. In historical narration, the imparting of meaning is achieved through memory alone. And although novelistic narration cannot be reduced to memory, it plays an important role here as well. This is especially evident when historians fail to bring to mind certain aspects of historical experience through memory; then it is often the novelist who must assume this task. The 1950s in West Germany provide a clear example of this phenomenon. With regard to the National Socialist past, repression rather than grief was prevalent. It was a novel that broke this taboo: *Die Blechtrommel*, by Günter Grass (1959). This literary breakthrough was fallowed by the works of the historiographers. George Steiner listed a number of examples of this function of the "writer as remembrancer" (Steiner, 1973). It is particularly in cases of the conscious falsification of history by historians that the writer steps in.

Writers and historiographers can help define the identity of a generation, a nation, or a culture. Since the Enlightenment and secularization, historiography and fiction have been burdened increasingly with the function of providing an identity. By dealing with memories, the experiences of the present attain a coherent perspective that provides the reader with an orientation for the future. Repeatedly correlating experiences of the present with those of the past can bring about the dissolution of stagnant patterns of identity, with a critical view of ideological prejudices (Lützeler, 1998). The factual orientation of the historiographer as well as the novelist's interest in what is possible are part of the provision of

identity. By means of memory, a potential of historical and literary meaning is activated that, along with the demonstration of actual limits and the allusion to yet untested possibilities, makes the provision of identity possible. Historical and literary consciousness is a complex construct in which memory is the prerequisite for interpretation and expectation. To structure an identity means, first, to identify with stories. During an identity crisis, society may show greater interest in historical remembering (as was the case in Europe in the nineteenth century), or demand a fictional provision of identity (as is presently the case in Latin America). The weaker the link with the historical past and the more suspect it appears, the stronger will be the desire for fictional patterns of identity.

See also: History: Theories and Methods.

Bibliography

Adorno, T.W., 1970. Ästhetische Theorie. Suhrkamp, Frankfurt, Germany.

Assmann, A., 1980. Die Legitimität der Fiktion. Fink, Munich, Germany.

Barthes, R., 1957. Mythologies. Editions du Seuil, Paris.

Baumgartner, H.M., Rüsen, J., 1982. Erträge der Diskussion. In: Lämmert, E. (Ed.), Erzählforschung. Ein Symposium. Metzler, Stuttgart, Germany.

Benjamin, W., 1972. Gesammelte Schriften 2. Suhrkamp, Frankfurt, Germany.

Borst, A., 1976. Das historische Ereignis. In: Baumgartner, H.M., Rüsen, J. (Eds.), Seminar: Geschichte und Theorie: Umrisse Historik. Suhrkamp, Frankfurt, Germany.

Cohn, D., 1978. Transparent Minds: Narrative Modes for Presenting Consciousness in Fiction. Princeton University Press, Princeton, NJ.

Culler, J., 1975. Structuralist Poetics. Structuralism, Linguistics and the Study of Literature. Routledge and Kegan Paul, London.

Chandler, J., 2009. The Cambridge History of English Romantic Literature. Cambridge University Press, Cambridge, UK.

Carter, R., McRae, J., 2002. The Routledge History of Literature in English: Britain and Ireland. Routledge, London.

Danto, A.C., 1965. Analytical Philosophy of History. Cambridge University Press, Cambridge, UK.

Demandt, A., 1978. Metaphern für Geschichte. Beck, Munich, Germany.

Eagleton, T., 1976. Criticism and Ideology. A Study in Marxist Literary Theory. Humanities Press, Atlantic Highlands, NJ.

Eco, U., 1979. The Role of the Reader. Indiana University Press, Bloomington, IN.

Fish, S., 1970. Literature in the reader: affective stylistics. New Literary History 2, 123–162.

Furet, F., 1982. L'atelier de l'histoire. Flammarion, Paris.

Grimm, R., 1963. Strukturen; Essays zur Deutschen Literatur. Sachse und Pohl, Göttingen, Germany.

Gies, D.T., 2004. The Cambridge History of Spanish Literature. Cambridge University Press, Cambridge, UK.

Gray, R.J., 2004. A History of American Literature. Blackwell Pub., Malden, MA.

Hamburger, K., 1968. Die Logik der Dichtung. E Klett, Stuttgart, Germany.

Hinck, W., 1995. Geschichtsdichtung. Vandenhoeck & Ruprecht, Göttingen, Germany.

Hart, J.L., 2011. Literature, Theory, History. Palgrave Macmillan, New York.

Iser, W., 1974. The Implied Reader; Patterns of Communication in Prose Fiction from Bunyan to Beckett. Johns Hopkins University Press, Baltimore, MD.

Jameson, F., 1981. The Political Unconscious: Narrative as a Socially Symbolic Act. Cornell University Press, Ithaca, NY.

Jorgensen, A., University of York, 2010. Reading the Anglo-saxon Chronicle: Language, Literature, History. Brepols, Turnhout.

Kocka, J., 1984. Zurück zur Erzählung? Plädoyer für historische Argumentation. Geschichte und Gesellschaft 10, 395–408.

Koselleck, R., 1976. Wozu noch Historie? In: Baumgartner, H.M., Rüsen, J. (Eds.), Seminar: Geschichte und Theorie. Suhrkamp, Frankfurt, Germany.

Kay, S., Cave, T., Bowie, M., 2003. A Short History of French Literature. Oxford University Press, Oxford.

Kelleher, M., O'Leary, P., 2006. The Cambridge History of Irish Literature. Cambridge University Press, Cambridge.

LeGoff, J., Nora, P., 1974. Faire de l'histoire. Gallimard, Paris.

Lützeler, P.M., 1986. Zeitgeschichte in Geschichten der Zeit. Bouvier, Bonn, Germany.

Lützeler, P.M., 1998. Nomadentum und Arbeitslosigkeit. Identität in der Postmoderne. Merkur 52, 908–918.

Loewenstein, D., Mueller, J.M., 2002. The Cambridge History of Early Modern English Literature. Cambridge University Press, Cambridge, UK.

Lee, P.H., 2003. A History of Korean Literature. Cambridge University Press, Cambridge.

Mair, V.H., 2001. The Columbia History of Chinese Literature. Columbia University Press, New York.

Marsh, R.J., 2007. Literature, History and Identity in Post-Soviet Russia, 1991–2006. Peter Lang, Oxford.

Marcus, G., Sollors, W., 2009. A New Literary History of America. Belknap Press of Harvard University Press, Cambridge, MA.

Norton, D., Norton, D., 2000. A History of the English Bible as Literature. Cambridge University Press, Cambridge, UK.

New, W.H., 2003. A History of Canadian Literature. McGill-Queen's University Press, Montreal, QC.

Prince, G., 1982. Narratology: The Form and Functioning of Narrative. Mouton, Berlin, Germany.

Platt, K.M.F., Brandenberger, D., 2006. Epic Revisionism: Russian History and Literature as Stalinist Propaganda. University of Wisconsin Press, Madison, WI.

Ricoeur, P., 1983. Temps et récit. Seuil, Paris.

Riffaterre, M., 1971. Essais de stylistique structurale. Seix Barral, Spanish, Barcelona.

Rüsen, J., 1982. Die vier Typen des historischen Erzählens. In: Koselleck, R. (Ed.), Formen der Geschichtsschreibung. Deutscher Taschenbuch Verlag, Munich, Germany.

Roskies, D.G., Diamant, N., 2012. Holocaust Literature: A History and Guide. Brandeis University Press, Hanover.

Steiner, G., 1973. The writer as remembrancer: a note on poetics 9. Yearbook of Comparative and General Literature 22, 51–57.

Stone, L., 1979. The revival of narrative reflections on a new old history. Past and Present 85, 3–24.

Szondi, P., 1973. Für eine nicht mehr narrative Historie. In: Koselleck, R., Stempel, W.D. (Eds.), Geschichte – Ereignis und Erzählung. Fink, Munich, Germany.

Schmidt, J., 2002. The Joys of Philology: Studies in Ottoman Literature, History and Orientalism, 1500–1923. Isis Press, Istanbul.

White, H., 1973. Metahistory: The Historical Imagination in Nineteenth-century Europe. The Johns Hopkin University Press, Baltimore, MD.

Widdowson, P., Barker, S., Gill, J., 2010. Literature as History: Essays in Honour of Peter Widdowson. Continuum, London.

Wellbery, D.E., Ryan, J., Gumbrecht, H.U., 2004. A New History of German Literature. Belknap Press of Harvard University Press, Cambridge, MA.

Wells, R.H., Burgess, G., Wymer, R., 2000. Neo-historicism: Studies in Renaissance Literature, History, and Politics. D.S. Brewer, Woodbridge, Suffolk, UK.

Watkins, J., 2002. Representing Elizabeth in Stuart England: Literature, History, Sovereignty. Cambridge University Press, Cambridge, UK.

History and the Social Sciences

W Spohn, Free University Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 6829–6835, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

History and the social sciences have developed and maintained complex relationships made up of commonalities and differences as well as divergences and convergences. In principle, both disciplinary areas are concerned with the scientific study of the human life-world and the whole range of human behavior, individual and collective, in time and space as well as in the past and the present. Both disciplinary areas originated, with the rise of the modern world, as a common enterprise to develop systematic, secular, and empirically validated knowledge about the historical-social reality. In this fundamental sense, history can be and should be seen as part of the social sciences or vice versa the social sciences as part of the historical sciences – both constituted by substantially interrelated modes of sociohistorical research (Gulbenkian Commission, 1996; Hall, 1999).

History and the social sciences have developed and maintained complex relationships made up of commonalities and differences as well as divergences and convergences. In principle, both disciplinary areas are concerned with the scientific study of the human life-world and the whole range of human behavior, individual and collective, in time and space as well as in the past and the present. Both disciplinary areas originated with the rise of the modern world as a common enterprise to develop systematic, secular, and empirically validated knowledge about the historical-social reality. In this fundamental sense, history can be and should be seen as part of the social sciences or vice versa the social sciences as part of the historical sciences – both constituted by substantially interrelated modes of sociohistorical research (Gulbenkian Commission, 1996; Hall, 1999).

Despite their fundamental commonality, however, history and the social sciences have followed quite different and diverging epistemological and methodological tracks during their academic institutionalization in Western Europe and North America in the nineteenth and early twentieth centuries. On the one hand, the social sciences (economics, sociology, and political science as well as anthropology) came to be oriented toward the successful epistemological model of the natural sciences and defined themselves primarily as nomothetic and generalizing disciplines. On the other, history, predominantly influenced by the humanities and critical against speculative social laws, developed traits of an idiographic and particularistic discipline. Against these traditional disciplinary definitions, a converging interdisciplinary cooperation between history and the social sciences has developed only in the second half of the twentieth century. The social-scientific orientation of history was programmatically formulated in approaches such as the 'new social history,' 'la nouvelle histoire,' 'Social Science History,' or 'Historische Sozialwissenschaft,' developing in leading journals such as *Annales Economies, Sociétés, Civilisations* (1946), *Past and Present* (1952), *Journal of Social History* (1967), *Journal of Interdisciplinary History* (1970), *Social Science History* (1976), and *Geschichte und Gesellschaft* (1975). In a parallel movement, the historical orientation of the social sciences crystallized in new areas like historical sociology, historical economics, historical political science, and historical anthropology. Sometimes even the epistemological identity of history and the social sciences has been stated in a fundamental sense.

More recently, however, this *rapprochement* between history and the social sciences was halted and partly reversed. Within the study and presentation of history a 'revival of the narrative,' a 'new cultural history,' and concepts of the 'historical cultural sciences' have emerged, emphasizing what distinguishes them from the social sciences proper. But it should be stressed that, in a parallel movement, the social sciences have been similarly challenged from within by alternative cultural, interpretative-hermeneutic, and deconstructive-postmodernist epistemologies. Often the so-called 'cultural turn' in history and the social sciences is seen as a replay of the traditional juxtaposition of idiographic and nomothetic positions and conservatively used to reconfirm the traditionally separate institutionalization of both disciplinary areas. But properly understood, this cultural turn can and should be rather viewed as a challenge to the still prevailing principled opposition of generalizing or particularizing disciplinary epistemologies and methodologies. The future task, then, would be to open the historical and social sciences toward a common and multivocal, intra- and interdisciplinary discourse of sociohistorical inquiry.

Disciplinary Divergence

In its originally institutionalized – but nowadays increasingly challenged – meaning, history is primarily concerned with the study, description, interpretation, and presentation of historical facts or relics from the past, focusing on: idiographic methods, the analysis of historical contexts and temporal change, and the conceptualization of historical phenomena as particular individualities, their inductive interpretation, and narrative presentation (Burke, 1992; Zunz, 1985). In the opposite – but also increasingly questioned – direction, the social sciences such as sociology, economics, political science, and anthropology have been traditionally understood as concerned with the study, generation, and explanation of social facts in the present emphasizing nomothetic methods,

systematic analysis, and theoretical explanation by reference to general laws or social regularities (Monkkonen, 1994; Skocpol, 1984).

This problematically polarizing division between history and the social sciences developed with the institutionalization of the university systems during the nineteenth century under the impact of the natural sciences. Science, since then, came to be equated primarily with the natural sciences and particularly with its conspicuous inorganic and organic branches of physics and biology. Taking the natural sciences as epistemological models, the social sciences emerged as an attempt to discover the general laws and evolutionary mechanisms of modern society (Lepenies, 1988; Gulbenkian Commission, 1996). Thus, political or classical economy as developed by Adam Smith or David Ricardo tried to find the general laws of modern capitalist market society or in a critical direction as formulated by Karl Marx to uncover the anatomy of emerging modern capitalism. In the same steps, the early founding fathers of sociology such as Claude-Henri Saint-Simon, Auguste Comte, or Herbert Spencer intended to find the structural and evolutionary laws of modern society. At the same time, with the advent of modern mass politics in the American and French revolutions, the emerging political science as in Alexis de Tocqueville or John Stuart Mill attempted to formulate the general features and bases of modern democratic polities. Finally, with the growing awareness of non-Western societies under the impact of European colonialism and imperialism, anthropology tried to analyze the archaic or 'primitive' societies as structural counterparts and evolutionary prestages of modern society.

By contrast, history in view of the contingent flow of historical events and human actions as well as the role of cultural and spiritual forces came to be skeptical of the search for quasi-natural laws of the historical-social world. Instead, history followed the humanities, art, and literature as culturally oriented disciplines that seemed to be closer to the spiritual, creative, and poetic aspects of historical development. Particularly under the impact of German 'historicism,' the epistemological and methodological principles of historical knowledge developed into a radical opposition to the positivist search for historical-social facts. The principle of Leopold von Ranke to narrate history 'wie sie eigentlich gewesen ist' formulated the conviction that the historical method has to grasp the historical reality in its manifold particularities. In epistemological terms as developed by Johann Gustav Droysen and Wilhelm Dilthey, history cannot be subsumed under deterministic-mechanistic laws, but has to be understood in its concrete, experience-near, and life-full quality. On this basis, not only were the historical subdisciplines such as political, legal, cultural, social, and economic history principally viewed as hermeneutic-interpretative disciplines, but also economics, sociology, and political science as promoted among others by Gustav Schmoller, Werner Sombart, or Max Weber were primarily conceptualized as cultural sciences rather than as social sciences in order to grasp the individual and collective meanings structuring socioeconomic and political actions and processes. These epistemological and methodological differences and divergences were characteristic for the relationships of history and the social sciences not only in Germany, but also in all Western societies. However, through the particular influence of German 'historicism' refuting on ontological grounds the applicability of explanatory social-scientific methods to historical events and cultural individualities, these divergences often developed into a principled opposition (Iggers, 1993; Lepenies, 1988).

The interdisciplinary bridging of this epistemological, analytical, and methodological opposition only became possible with the increasingly historical orientation of the social sciences as well as the growing social-scientific and sociological orientation of history (Abrams, 1982; Lipset and Hofstadter, 1970; Skocpol, 1984). In the first half of the twentieth century, this converging movement was seldom although of crucial significance in preparing the ground for the later developing convergences between the historical and social sciences. Of lasting influence was, first, Emile Durkheim and the continuation of his approach by Marcel Mauss and Henri Berr, attempting to combine the sociology of social change with anthropological comparisons and historical developments (Raphael, 1994; Stoianovich, 1976). Of fundamental importance was also, second, the German historical-comparative sociology, particularly in the hands of Max Weber, Otto Hintze, Werner Sombart, and in their continuity also of Norbert Elias, who tried to explain the specificities of the Western civilization, its unique economic, political, and cultural order by historical and universal comparisons (Lepenies, 1988). Third, strongly influenced by European sociological traditions, modernization theory developed in the United States as a universal social-scientific approach closely connected with historical-comparative modernization research on the historical prerequisites and developments of modern societies on a world scale.

Interdisciplinary Convergence

Influenced by these precursors, the historical and social sciences combined programmatically, first in the Annales school (Raphael, 1994; Stoianovich, 1976) and then later in the American Social Science History approach (Tilly, 1981), some proponents of English Marxist historiography (Kaye, 1984) and in their emulation in the West German 'Historische Sozialwissenschaft' (Kocka, 1977; Wehler, 1980). This interdisciplinary convergence led to different forms of 'social-scientific history' or 'historical social science' – despite many national particularities. They materialized in the 1960s and 1970s in a parallel movement in the Western societies (Iggers, 1993). In critical opposition to the received tradition of narrative and exclusively interpretive history, the major common methodological aims of social-scientific history or historical social science can be summarized in the following principles. First, social-scientific historical approaches should be oriented to a structural and processual perspective on the various (economic, sociopolitical, and cultural) dimensions of macrohistorical change. Second, they should shift historical analysis from a 'top-down' elite perspective to a 'bottom-up' popular or people's history. Third, social-scientific historical approaches should give up the traditional emphasis on political and intellectual history and turn toward economic and social history. Fourth, they should make systematic use of the available social-scientific, theoretical, comparative, and

quantitative methods in history and favor systematic explanation instead of hermeneutic interpretation. Finally, this movement to social-scientific history or historical social science was often motivated by a critique of traditional elitist-conservative historiography and a commitment to emancipatory, democratic, or enlightened values.

The common methodological aims of social-scientific history or historical social science, at the same time, were colored and refracted by the respective national intellectual and cultural environments. In France, the 'nouvelle histoire' promoted within the long established interdisciplinary tradition of the Annales school (Raphael, 1994; Stoianovich, 1976) was characterized by a rather balanced multilevel (*longue durée, conjonctures, évènements*) and multidimensional (economy, social structure, and civilizational mentalities) mode of historical analysis. In England, a strong neo-Marxist influence shaped the rise of economic and social history with a particular emphasis on social inequality and class history (Kaye, 1984). In the United States the interdisciplinary social science history was influenced by the development of historical-comparative multidimensional modernization research and the special attraction of econometric, statistical, and social research methods for the quantification of historical processes (Tilly, 1981). In Germany, Marxian and Weberian traditions of scholarship combined with elements of Western modernization theories and led to 'Historische Sozialwissenschaft' and 'Gesellschaftsgeschichte,' in which the analysis of social inequality and political power with a particular emphasis on the German 'special path' were treated (Kocka, 1977; Wehler, 1980). At the same time, these national variations reflect not only differing intellectual and theoretical traditions, but also a different emphasis on the use of qualitative and quantitative methods in historical-social inquiry. This is also indicated by the paradigmatic oscillations between the concepts of (1) historical social science, (2) structural history, and (3) social-scientific history.

'Historical social science' ('Historische Sozialwissenschaft') serves as a loosely defined summary category. Still, the historical-comparative analysis of large-scale and multidimensional processes of social change from traditional life-worlds to modern society has been at the center. In particular, the concepts of historical-comparative modernization research, historical sociology, societal history, and synthetic structural history, with different theoretical and methodological orientations, belong to this category. In the French Annales school in a mixture of Durkheimian and Marxian traditions, the comparative analysis of feudalism by Marc Bloch and the social history of capitalism by Fernand Braudel (to be followed by Immanuel Wallerstein's history of the modern world system) are important examples. Among English Marxist historians, the debate on the transition from feudalism to capitalism in a line from Maurice Dobb to Perry Anderson and the social history of modern classes, developed in different directions by Eric Hobsbawm and Edward Thompson, carried the field. In the United States, modernization research guided by a mixture of Max Weber's comparative sociology and Talcott Parsons' structural functionalist theory of social change was of major importance, resulting in path-breaking studies like Neil Smelser's analysis of social change in the English industrial revolution; Shmuel Eisenstadt's civilizational comparison of empires; Barrington Moore's historical explanation of the development of twentieth-century dictatorship and democracy; Tilly's edited volume on the formation of national states in Western Europe; and Reinhard Bendix's study on corresponding historical changes in the legitimacy of political domination (see as an overview Skocpol, 1984). In Germany, the modernization paradigm in a mixture of Weber and Marx directed Ralf Dahrendorf's sociological analysis of Germany's troublesome process of democratization or Hans-Ulrich Wehler's (1987–95) and Jürgen Kocka's (1993) comparative social-historical investigations of the German 'special path' and the role of the social classes within it.

'Structural history,' again, has had different meanings with different outlooks. It materialized in innovative areas of a variety of 'new histories' along the lines between the main subsystems of modern society and the related social-scientific disciplines. In the 'New Economic History' of the 1960s, the application of economic concepts and theories has given rise to the history of economic growth, technology, capitalist markets, and institutions; here, David Landes' (1969) *Prometheus* or Sidney Pollard's (1981) *Peaceful Conquest* are examples of a qualitative type, whereas Gianni Tonioli's (1991) comparative history of European industrialization – debating Alexander Gerschenkron's theory of economic backwardness – represents a quantitative type of economic history. In the 'New Social History' in the 1960s and 1970s, sociological concepts were used in the historical analysis of demographic development, social structure, and social mobility; social groups, occupational structures, and social classes; social institutions, associations, and interest groups; as well as social movements, social protest, and political mobilization (Kocka, 1977). This was combined with the application of political-scientific approaches to the history of political modernization, state formation, nation building, the development of democratic regimes and party systems, as well as political power and contestation (Tilly, 1984). Further, social-scientific approaches influenced by social anthropology have also been applied to cultural phenomena and developments such as family and socialization, religion and secularization, as well as value systems and mental structures (see different essays in Rürup, 1977).

The concept of 'social-scientific history' has been used primarily for the application of rigorous quantitative research methods in structural or social history (Landes and Tilly, 1971; Schröder, 1994). Not always, but often it has been equated with historical social research or the systematic quantitative operationalization and measurement of historical-social variables. In addition, this often combined with a deductive application of theoretical models to historical data and their explanation through statistical-mathematical and increasingly computer-assisted methods. Also this type of social-scientific historical research followed, either in close cooperation or in separate organization, the main areas of the 'new histories.' Particularly, the 'New Economic History' has been suited to cliometric, statistical, and econometric research methods as for instance in Paul Bairoch's and Maurice Lévy-Leboyer's (1981) analysis of economic growth disparities since the industrial revolution. But also in the new social history, social-demographic family, mobility, and migration studies such as Richard Wall et al. (1983) as well as time series of social protest

such as Charles Tilly et al. (1975) are prominent examples. In the new political history, the historical change of voting behavior, in particular, became an important field of historical social research (e.g., Best and Thome, 1991). But also in the cultural history of mentalities as carried through particularly in the Annales tradition, the application of quantitative methods were fruitfully developed (e.g., Shapiro and Markoff, 1998).

Interdisciplinary Challenges

Successful as these different new historical subdisciplines in their qualitative and quantitative orientations have been, the guiding vision of social-scientific approaches to history has been the combination of the various structural histories in a unifying societal history, 'histoire totale' or 'Gesellschaftsgeschichte' (Iggers, 1993; Wehler, 1980). This vision follows the epistemological ideal of merging the historical sciences and the social sciences in a common paradigm – as essentially expressed in the paradigmatic concepts of social science history, historical sociology, or 'historische Sozialwissenschaft' (Abrams, 1982; Smith, 1991; Wehler, 1980). With this unifying vision, however, the disciplinary traditions and methodological differences between historicized conceptions of the social sciences and social-scientific approaches to history become easily blurred. On the one hand, historical approaches in the social sciences basically proceed within a generalizing frame of reference (Monkkonen, 1994). For instance, historical economics follows the basic idea of uncovering time/space bound economic laws or regularities (Kindleberger, 1990; Monkkonen, 1994; Wallerstein, 2000). Also historical sociology is primarily oriented to the general features of time/space bound social processes, patterns, and figurations (as exemplary studies see Eisenstadt, 1996; Mann, 1986/1993, Smelser, 1991). In parallel, historical approaches in the political sciences concentrate on general features of political development (Monkkonen, 1994). Also, historicized versions of anthropology are interested in the general evolutionary patterns of human behavior (Monkkonen, 1994). On the other hand, social-scientific approaches to history as multidimensional structural or combined societal history usually remain in the confines of a narrative frame of reference – even when applying rigorously social-scientific theoretical concepts and research methods. Thus, in spite of all interdisciplinary convergence, there have been limits to a full merger between historically oriented social science and social science oriented history.

A second *problematique*, expressed in the oscillating meanings of historical social science, structural history, and social-scientific history, concerns the relationships between theory and method. Structural approaches to history often use social-scientific concepts in order to systematize, interpret, and explain long-term historical processes in one or another societal dimension. Social-scientific approaches to history are often oriented to theoretical models in order to explain deductively quantified historical-developmental data. Approaches in historical social science and related synthetic orientations to history often use multidimensional theories of social change in order to explain causal connections between the various societal dimensions. Many varieties of historical social science, structural history, and social-scientific history use theories and concepts, deductive and inductive modes of explanation or interpretation, as well as qualitative, quantitative, and comparative methods of analysis. In an effort to systematize these methodological strategies in historical-social inquiry, Theda Skocpol proposed to distinguish model-theoretical, causal-analytical, and interpretive modes of explanation (Skocpol, 1984: pp. 356–391). Charles Tilly, concentrating on the types of comparison and the number of cases and variables involved, differentiated between individualizing, generalizing, encompassing, and variation-finding forms of comparison (Tilly, 1984: p. 84ff). Charles Ragin (2000), attempting to bridge qualitative and quantitative modes of analysis, distinguished between case-analytical and variation-oriented methodological strategies. Fruitful as these methodological distinctions are, particularly the methodological relationships between explanatory and interpretive research strategies, the theoretical sources and analytical consequences of these methodological distinctions for sociohistorical inquiry still remain unexplored (Hall, 1999: p. 173).

A third increasingly salient but divisive *problematique* concerns the role of culture in the differing modes of historical social science, structural history, and social-scientific history. Not only in practice, but also often programmatically, social-scientific approaches to sociohistorical inquiry tend to favor the analysis of the economic, social, and political dimensions and to limit the analysis of culture to value systems or cultural institutions. The contents and forms of culture in their constitutive or causal role in history are often neglected and the corresponding hermeneutic or deconstructive methods to interpret their meanings, symbols, values, and identities are often seen as secondary. Against these prevailing tendencies of social-scientific approaches to culture, there emerged an increasingly influential challenge by the 'new cultural history' or cultural-scientific versions of history (Hunt, 1989; Dirks et al., 1994; Mergel and Welskopp, 1997). This 'cultural turn' consists of different countercurrents against the social-scientific vision of a systematic, generalizing, and explanatory structural and societal history. First, the revival of the narrative insists on the interpretive, literary, and narrative tasks of history (Stone, 1979). Second, the development of microhistory or 'Alltagsgeschichte' emphasizes the complexity of life-worlds as such or as microfoundations of historical macrostructures (Luedtke, 1989). Third, the 'linguistic turn' presupposes the all-pervasiveness of language in the social world, thus stressing the analytical task of deconstructing discourses and decoding linguistic-symbolic structures (Hunt, 1989). Fourth, the 'cultural turn' is consciously oriented to those social groups such as women, ethnic minorities, or the underclass, seen as neglected by the predominating social-scientific approaches to history (e.g., *History Workshop Journal* 1976 or *Radical History* 1975). Finally, some authors are influenced by a postmodernist perspective emphasizing the fragmentation and decentering against the structuring and ordering of the social world (Iggers, 1993). The result is not only the generation of a whole new set of historical subdisciplines such as women's and gender history, ethnic and minority history, regional and local history as well as the history of identities and memory, but also an influential shift from social-scientific history to cultural-scientific history (see however Bonnell and Hunt, 1999).

In a sense, the 'cultural turn' signifies a revitalization of historicism and cultural-scientific approaches to history – reemphasizing qualitative, interpretative, ethnographical, and hermeneutic methods against systematic-explanatory orientations and, with it, reversing the causal or structural determination of history from 'society' to 'culture.' At the same time, this theoretical and analytical movement reflects a parallel shift in the social sciences from social-structural to institutional and cultural paradigms (Skinner, 1985; Matthes, 1992; McDonald, 1996). For instance, in economics and economic sociology the new institutionalism has gained wide recognition, but also cultural neo-Weberian approaches reappear (North, 1981; Monkkonen, 1994; Swedberg, 1998). In various domains of sociology as well, the 'new institutionalism' and various forms of a 'new cultural sociology' become influential (Monkkonen, 1994; Crane, 1994). Further, in political science the traditional institutional approaches become challenged by cultural paradigms (Monkkonen, 1994). All this is influenced by a crucial shift in anthropology from social and structural anthropology to various forms of interpretative, ethnographic, historical, and cultural anthropology (Monkkonen, 1994). As a result, social-scientific history is not merely confronted with systematic, generalizing, and explanatory forms of the social sciences, but also with individualizing, interpretative, and reflexive varieties of the cultural sciences. In other words, new opportunities of interdisciplinary cooperation emerge.

Interdisciplinary Perspectives

The interdisciplinary cooperation and convergence of history and the social sciences have opened up novel and dynamical fields of sociohistorical research. At the same time, the *idées directrices* of that historical-social-scientific convergence have often privileged quasi-natural-scientific epistemologies and methodologies. As a consequence, generalizing modes of sociohistorical inquiry stand in tension with particularizing modes; structural and macroanalytical theories used in historical-social analysis conflict with action and microanalytical theories; causal explanation is often limited to the application of theoretical models and causal-analytic methodologies in opposition to interpretation and hermeneutic strategies; and, accordingly, there are tensions between qualitative and quantitative methods, including generalizing and particularizing methods of comparison. These tensions are revealed, on the one hand, within the social-scientific convergence of history and the social sciences in the paradigmatic oscillations between the guiding concepts of historical social science, structural history, and social-scientific history. These tensions develop, on the other hand, within the wider historical sciences, in the often polarized opposition to renewed historicist and culturalist, narrative and hermeneutic or deconstructivist modes of historical-social inquiry. At the same time, these theoretical and methodological oppositions and polarizations are not solely characteristic for the historical sciences. Parallel disciplinary developments can also be observed in the social sciences and the widespread and often sharp oppositions to cultural-scientific approaches in the various social-scientific disciplines.

These epistemological, theoretical, and methodological oppositions and conflicts have often led into a stalemate in both disciplinary areas of the historical and social sciences between renewed forms of narrative-interpretative and anthropological-cultural history in contradistinction to the nomothetic social sciences. However, these oppositions in the historical as well as the social sciences can and should be used to reenergize a wider intra- and interdisciplinary cooperation between the historical, social, and cultural sciences. These renewed cooperative efforts should concentrate on the crucial epistemological, theoretical, methodological, and analytical dimensions in sociohistorical inquiry within and across the established academic disciplines: on the epistemological level, combining social-scientific and cultural-scientific as well as generalizing and particularizing modes of knowledge; on the theoretical level confronting systematically social-structural and cultural approaches to historical-social reality; on the analytical level, mediating between the social-structural, institutional and cultural dimensions as well as between the global, national, regional, and local macro-/microdimensions in sociohistorical inquiry; and on the methodological level, clarifying the relationships between qualitative and qualitative as well as historical, social-scientific, and cultural-scientific comparative research strategies (Bourdieu, 1995; Haupt and Kocka, 1996; Spohn, 1998). All these renewed epistemological, theoretical, methodological, and analytical efforts of an inter- and intradisciplinary cooperation should not be done in segmented isolation, but in a multidisciplinary and multivoiced context of practiced sociohistorical inquiry. Such an opening of the historical and social sciences, however, is not only a matter of an intellectual intra- and interdisciplinary cooperation, but also demands an institutional cross-cutting between and across the segmented academic disciplines of the historical and social sciences (Gulbenkian Commission, 1996).

See also: Anthropology and History; Cultural History; Demography, Early History of; Economic History; Economics, History of; Family and Kinship, History of; Field Theory; Historical Sociology; History: Theories and Methods; Social History; Weberian Social Thought, History of.

Bibliography

Abrams, P., 1982. Historical Sociology. Open Books, Shepton Mallet, UK.
Bairoch, P., Lévy-Leboyer, M. (Eds.), 1981. Disparities in Economic Development since the Industrial Revolution. St. Martin's Press, New York.
Backhouse, R., Fontaine, P., 2010. The History of the Social Sciences since 1945. Cambridge University Press, Cambridge.
Best, H., Thome, H. (Eds.), 1991. Neue Methoden der Analyse historischer Daten. Scripta Mercaturae Verlag, St. Katharinen.
Bhattacharya, S., 2007. Centre for studies in civilizations (Delhi, India), & project of history of Indian science, philosophy, and culture. In: Development of Modern Indian Thought and the Social Sciences. Centre for Studies in Civilizations, New Delhi.
Bonnell, V.E., Hunt, L. (Eds.), 1999. Beyond the Cultural Turn: New Directions in the Study of Society and Culture. University of California Press, Berkeley, CA.
Bourdieu, P. (Ed.), 1995. Histoire sociale des sciences sociales. Actes de la Recherche en sciences sociales. 106/107.
Burke, P., 1992. History and Social Theory. Polity Press, Cambridge, UK.
Crane, D. (Ed.), 1994. The Sociology of Culture. Emerging Theoretical Perspectives. Blackwell, Oxford, UK.

Dirks, N.B., Eley, G., Ortner, S.B. (Eds.), 1994. Culture/Power/History: A Reader in Contemporary Social Theory. Princeton University Press, Princeton, NJ.
Eisenstadt, S.N., 1996. Japanese Civilization. A Comparative View. The University of Chicago Press, Chicago.
Fleck, C., Beister, H., 2011. A Transatlantic History of the Social Sciences: Robber Barons, the Third Reich and the Invention of Empirical Social Research. Bloomsbury Academic, London.
Gulbenkian Commission on the Restructuring of the Social Sciences, 1996. Open the Social Sciences: Report of the University Press. Stanford University Press, Stanford, CA.
Hall, J.R., 1999. Cultures of Inquiry. From Epistemology to Discourse in Sociohistorical Research. Cambridge University Press, Cambridge, UK.
Haupt, H.-G., Kocka, J. (Eds.), 1996. Geschichte und Vergleich. Campus Frankfurt.
Hunt, L. (Ed.), 1989. The New Cultural History: Essays. University of California Press, Berkeley, CA.
Iggers, G.G., 1993. Geschichtswissenschaft im 20. Jahrhundert. Ein kritischer Überblick im internationalen Zusammenhang. Vandenhoek und Ruprecht, Göttingen.
Kaye, H.J., 1984. The British Marxist Historians an Introductory Analysis. Cambridge University Press, Cambridge, UK.
Kindleberger, C.P., 1990. Historical Economics. Art of Science? University of California Press, Berkeley, CA.
Kocka, J., 1977. Sozialgeschichte; Begriff, Entwicklung, Probleme. Vandenhoek und Ruprecht, Göttingen.
Kocka, J. (Ed.), 1993. Bürgertum und bürgerliche Gesellschaft im 19. Jahrhundert. Beck, München.
Landes, D.S., 1969. The Unbound Prometheus: Techological Change and Industrial Development in Western Europe from 1750 to the Present. Cambridge University Press, Cambridge, UK.
Landes, D.S., Tilly, C., 1971. History as Social Science. Prentice-Hall, Englewood Cliffs, NJ.
Lipset, S.M., Hofstadter, R. (Eds.), 1970. Sociology and History: Methods. Basic Books, New York.
Lepenies, W., 1988. Between Literature and Science: The Rise of Sociology. Cambridge University Press, Cambridge, UK.
Luedtke, A. (Ed.), 1989. Alltagsgeschichte: Zur Rekonstruktion Historischer Erfahrungen und Lebernweisen. Campus, Frankfurt.
Mann, M., 1986/1993. The Sources of Social Power, 2 vols. Cambridge University Press, Cambridge, UK.
Matthes, J. (Ed.), 1992. Zwischen den Kulturen? Die Sozialwissenschaften vor dem Problem des Kulturvergleichs. Vandenhoek und Ruprecht, Göttingen.
McDonald, T.J. (Ed.), 1996. The Historic Turn in the Human Sciences. Michigan University Press, Ann Arbor, MI.
Mergel, T., Welskopp, T. (Eds.), 1997. Geschichte zwischen Kultur und Gesellschaft. Beiträge zur Theoriedebatte. Beck, München.
Monkkonen, E.H. (Ed.), 1994. Engaging the Past. The Uses of History. Across the Social Sciences. Duke University Press, Durham, NC.
North, D.C., 1981. Structure and Change in Economic History. Norton, New York.
Oakley, A., 2000. Experiments in Knowing: Gender and Method in the Social Sciences. New Press, New York.
O'Connor, A., 2001. Poverty Knowledge: Social Science, Social Policy, and the Poor in Twentieth-Century U.S. History. Princeton University Press, Princeton, NJ.
Pollard, S., 1981. Peaceful Conquest: The Industrialization of Europe 1760–1970. Oxford University Press, Oxford, UK.
Ragin, C.C., 2000. Fuzzy-set Social Science. University of Chicago Press, Chicago.
Raphael, L., 1994. Die Erben von Bloch und Febvre. Klett und Cotta, Stuttgart.
Ross, D., Porter, T.M., 2003. The Modern Social Sciences. Cambridge University Press, Cambridge.
Rürup, R. (Ed.), 1977. Historische Sozialwissenschaft: Beitr. zur Einf in d. Forschungsprais. Vandenhoek and Ruprecht, Göttingen.
Schröder, C., 1994. Historische Sozialforschung. Minerva, St. Katharinen.
Shapiro, G., Markoff, J., 1998. Revolutionary Demands: A Content Analysis of the Cahiers De Doléances of 1789. Stanford University Press, Stanford, CA.
Skinner, Q. (Ed.), 1985. The Return of Grand Theory in the Human Sciences. Canto, Cambridge, UK.
Skocpol, T. (Ed.), 1984. Vision and Method in Historical Sociology. Cambridge University Press, Cambridge, UK.
Smelser, N.J., 1991. Social Paralysis and Social Change: British Working-Class Education in the Nineteenth Century. University of California Press, Berkeley, CA.
Smith, D., 1991. The Rise of Historical Sociology. Temple University Press, Philadelphia, PA.
Spohn, W. (Ed.), 1998. Kulturanalyse und vergleichende Forschung Comparativ. Leizpiger Beiträge zur Universalgeschichte und vergleichenden Gesellschaftsforschung 8 (1).
Stoianovich, T., 1976. French Historical Method: The Annales Paradigm. Cornell University Press, Ithaca, NY.
Stone, L., 1979. The revival of the narrative: reflections on a new old history. Past and Present 85, 3–24.
Swedberg, R., 1998. Max Weber and the Idea of Economic Sociology. Princeton University Press, Princeton, NJ.
Staum, M.S., 2011. Nature and Nurture in French Social Sciences, 1859–1914 and beyond. McGill-Queen's University Press, Montréal.
Tilly, C., 1981. As Sociology Meets History. Academic Press, New York.
Tilly, C., 1984. Big Structures, Large Processes, Huge Comparisons. Russel Sage Foundation, New York.
Tilly, C., Tilly, L., Tilly, R., 1975. The Rebellious Century 1830–1930. Harvard University Press, Cambridge, MA.
Toniolo, G., Sylla, R. (Eds.), 1991. Patterns of European Industrialization in the 19th Century. Routledge, London.
Wall, R., Robin, J., Laslett, P. (Eds.), 1983. Family Forms in Historic Europe. Cambridge University Press, Cambridge, UK.
Wallerstein, I.M., 2000. The Essential Wallerstein. New Press, New York.
Wagner, P., 2001. A History and Theory of the Social Sciences: Not All that Is Solid Melts into Air. SAGE, London, England.
Wehler, H.-U., 1980. Historische Sozialwissenschaft und Geschichtsschreibung. Vandenhoek and Ruprecht, Göttingen.
Wehler, H.U., 1987–95. Gesellschaftsgeschichte Deutschlands, 3 vols. Beck, München.
Zunz, O. (Ed.), 1985. Reliving the Past: The World of Social History. University of North Carolina Press, Chapel Hill, NC.

History from Below, the History of Everyday Life, and Microhistory

Andrew I Port, Wayne State University, Detroit, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'People's history' focuses on the lives of ordinary people, with an eye to their struggles, everyday practices, beliefs, values, and mentalities. Influenced by the *Annales* School and cultural anthropology, but reacting against traditional social history's emphasis on social structures and serial trends, its practitioners emphasize the importance of individual agency while trying to demonstrate the complexity of lived experience, the fluidity of identity, and the subjective nature of meaning. Important types of 'people's history' include 'history from below,' *Alltagsgeschichte* (the 'history of everyday life'), and 'microhistory,' all of which involve a dramatic reduction of historical scale, focusing on a single individual, community, or spectacular event.

The Origins, Characteristics, and Goals of 'People's History'

'People's history' marks a radical departure from traditional mainstream historiography, which long tended to concentrate on high-level politics and diplomacy, warfare, and the lives of great statesmen. The move away from this focus on the high and mighty began most systematically during the period between the two world wars. Trying to understand social dynamics and the ways in which societies change, the members of the French *Annales* School embraced an interdisciplinary, materialist approach that looked to the social sciences for inspiration. At the same time, they shifted their attention away from major events and powerful individuals to underlying structural forces, mentalities, and the masses. Similar concerns and methods inspired a group of Marxist historians active in the United Kingdom after World War II, as well as a number of historians working later in the United States. Turning their attention to the lives and struggles of ordinary people, they focused on social relations at the grass roots, popular forms of protest, everyday activities such as work and leisure, as well as attitudes, beliefs, practices, and behavior. This became known in the 1960s as 'history from below.'

It was out of this tradition, which significantly broadened the array of 'legitimate' topics considered worthy of historiographical investigation, that new variations of 'people's history' – *Alltagsgeschichte* ('the history of everyday life') in Germany, 'microhistory' in Italy and France – emerged in the 1970s and 1980s. Despite some differences in their method and use of sources, the two share a number of essential characteristics. In the first place, their practitioners focus on the qualitative, quotidian, 'lived' experiences of ordinary people, i.e., on the actions, practices, habits, values, beliefs, mentalities, and feelings of the oppressed, excluded, pauperized, and marginalized: those who have traditionally been excluded from historical accounts and remained "largely anonymous in history – the 'nameless' multitudes in their workaday trials and tribulations" (Lüdtke, 1995: 4).

To that end, they dramatically reduce the scale of their historical investigation, confining it to a single individual, small community, or seemingly obscure event which is then subject to painstaking microscopic analysis involving an intensive study of the available documentary material. Immersing themselves like detectives in seemingly mundane sources related to themes such as everyday family life, gender relations, leisure activities, and popular culture, they search for and try to tease out the meaning of 'various clues, signs, and symptoms' – all in the belief that 'microscopic observation' will uncover and explain the 'previously unobserved,' as well as other revealing phenomena lost to the type of 'conventional analysis' that takes a more comprehensive bird's-eye view (Levi, 1991: 97, 106; Brewer, 2010: 97; Muir, 1991: xvi–xvii). The focus on spectacular stories and obscure people are, in short, 'devices' to get at larger issues (Lepore, 2001: 144) – while the 'lingering' over mere details represents an innovative attempt to achieve "the old dream of a total history, but this time reconstructed from the bottom up" (Steege et al., 2008: 375; Revel, 1995: 497).

Such histories usually fall into one of two categories: the 'episodic' and the 'systematic' (Gregory, 1999: 102). The first type, which tends to take a narrative approach and rely heavily on 'thick description,' focuses on a single, spectacular episode or event usually involving one person or a small group of individuals – such as the investigation of a heretical sixteenth-century Italian miller by Inquisition officials (Ginzburg, 1980), the elaborately staged murder of dozens of cats by disgruntled apprentice printers in Paris in the 1730s (Darnton, 1984), or an antisemitic riot incited by accusations of blood libel in a small Prussian town in the early twentieth century (Smith, 2002). The other type assiduously reconstructs the complex web of familial and extrafamilial social relations in a small community. Prominent examples include Giovanni Levi's study of social interaction in a village in the Piedmont in the 1690s – "a banal place and an undistinguished story," in the words of the author (Levi, 1988) – and David Sabean's dense studies of property, production, and kinship in the southern German village Neckarhausen from 1700 to 1870 (Sabean, 1990, 1998).

The goals and value of such studies are manifold. In the first place, they are at pains to demonstrate the ways in which ordinary people had agency and were not merely the victims of large, amorphous, impersonal forces, i.e., that they were the active subjects of their own lives – not just the passive objects of 'history.' A great deal of attention is thus paid to their ability to withstand hegemonic forms of dominance and control, i.e., to "an individual's constant negotiation, manipulation, choices and decisions in the face of a normative reality which,

though pervasive, nevertheless offers many possibilities for personal interpretations and freedoms" (Levi, 1991: 94). This focus on obscure individuals and the 'business of coping' (Bucur et al., 2009: 206) is also important because it vividly reveals the multiplicity and extreme complexity of lived experience, especially its more subjective dimensions in 'concrete life situations.' Such close-up investigations – which combine the "warmth of the narrator's intimate glance and the coldness of the scientist's detached observation" (Ginzburg, 1993: 16) – lay bare the 'social production and construction of meaning' by individuals themselves in ways that can 'hold up against the ruling discourses of a particular society,' the 'ambiguities and contradictions' of perceptions and behavior, as well as the way in which 'power relations are reproduced and challenged in everyday exchanges' (Eley, 1989: 315, 322–323; Bucur et al., 2009: 195–196).

The Role of Agency and the Reaction against Structures, Serials, and Teleologies

'People's history' in all its forms traces its intellectual heritage to *Annales*, but it is at the same time a conscious reaction against the work of that school, as well as against the type of interdisciplinary 'social science' history it inspired in France, the United States, and Germany beginning in the 1950s and 1960s. The latter, as practiced by leading historians such as Fernand Braudel in France and later Hans-Ulrich Wehler in Germany, relied heavily on sociological theories and techniques, focusing on aggregate trends and numerical series (e.g., births, marriages, deaths, prices, occupations) in an attempt to uncover long-term, large-scale, measurable regularities, formulate historical 'laws,' and arrive at blanket generalizations about human behavior (Revel, 1995; Eley, 2005).

Those who became disillusioned with this quantitative approach launched a multipronged critique. They objected, in the first place, to the way it seemingly crushed "all individuals to insignificance under the weight of vast impersonal structures and forces" (Muir, 1991: xxi) and thus failed to take into account the agency of ordinary people. Instead of consigning the so-called masses to historical impotence, history 'from below' tries to show how they themselves contribute to, appropriate, and shape broad supraindividual forces and structures. They see historical change and continuity, in other words, as the result of actions by groups and individuals who are '*simultaneously* both objects of history and its subjects' (Lüdtke, 1995: 6–8).

In a sense, the way in which they problematize the relationship between agency and the iron cage of structure is reminiscent of studies by so-called new cultural historians, who question Marxist claims about the primacy of the 'material' and the 'social' in the complicated relationship between base and superstructure (Hunt, 1989). That said, historians of everyday life and mirohistorians have been just as critical of postmodernism, which the new cultural historians have strongly embraced, as they have been of structuralism. Whereas the latter seemingly erased the role of individuals, the former has tried to 'kill off the subject completely' (Steege et al., 2008: 377) and thus call into question all claims about human agency.

Those who practice various types of 'people's history' also question the seemingly uniform, homogeneous way in which social science history has portrayed social life across different societies, as well as within any single society. They point instead to the great variety and complexity of social relations, experiences, behavior, and identities at the grass roots. This, they claim, is something that a focus on aggregate trends, large-scale processes, as well as overarching categories and groups (such as 'workers') is, by its very nature, unable to grasp. They draw attention to other supposed weaknesses of social science history as well: its reliance, for example, on documents – usually produced by those in positions of power – containing material that can be serialized. What about documents and documented practices, they ask, that do not readily lend themselves to quantification, or those that are 'documentarily unique' – what François Furet has suggestively referred to as the *hapax legomenon* (Ginzburg, 1993: 21). The claim here is that seemingly unique events and actions, as well as single cases or episodes, that violate norms can, in fact, illustrate underlying social structures, behaviors, and practices that would otherwise remain undetected.

Besides criticizing the methodological shortcomings of quantitative social science history and its 'macro' approach, which is seen as too deterministic and mechanistic, the practitioners of 'people's history' have also strenuously objected to the way it embraces grand narratives of progress – above all those associated with 'modernization theory' and its "single, linear progressive model…against which all societies are measured." The gold standard of development has been Western affluence and the triumph of modern consumer society (Brewer, 2010: 93–94). By turning their attention to the ways in which the 'losers' of history and the 'casualties of progress' experienced modernity in all its facets, as well as to 'fragmentation, contradictions, and plurality of viewpoints,' microhistorians and historians of the everyday look at social strategies and 'concrete experiences' that buck general trends – and thus fail to accord with the 'grand contours' of history identified by those who take a more capacious view of historical developments. This allows them to emphasize the importance of contingency and uncover, furthermore, the "social context in which an apparently anomalous or insignificant fact assumes meaning," thus revealing "the hidden incoherences of an apparently unified social system" (Levi, 1991: 107; Lüdtke, 1995: ix, 7).

Political Sensibilities and Intellectual Influences

The foregoing suggests the political agenda of everyday history and microhistory, whose practitioners – while highly critical of Marxist structural history – share Marxist critiques of "liberal conceptions of modernity" and the steamroller effects of capitalism, industrialization, and bureaucracy on the downtrodden (Gregory, 1999: 101). Coming from a leftist background, their political sensibilities were (re)formed in the crucible of the 1970s, a period of political inertia and economic stagnation that not only marked the end of the *trentes glorieuses* in the West, but also called into question – following the turbulence of the late 1960s – Marxism's 'scientific' premises and promises about the progressive nature

and emancipatory power of the industrial proletariat (Bell, 2002). It was in this atmosphere, characterized by a 'new reticence' about progress and modernity, that a certain nostalgia for the preindustrial past flourished. In fact, this backward-looking 'romantic opposition to capitalism' and late modernity helps explain why so many pioneering microhistorical studies have looked at communities and mentalities during the early modern period (Ginzburg, 1993: 20).

The critique of Western ethnocentrism that accompanied this rejection of modernist teleology reflects the strong influence that cultural anthropology has had on the field of everyday history. In fact, its practitioners self-consciously adopt a variety of ethnographic methods and insights, such as Franz Boas and Claude Lévi-Strauss's ideas about 'cultural relativism,' as well as concepts such as Clifford Geertz's "thick description," which emphasizes the importance of social context, signs, and symbolism in understanding the cultural meaning of human behavior and practices. Like ethnographers, they try, by way of written documents and other sources, to immerse themselves in the daily lives of those they study, with an eye to uncovering routine and repetitive acts that, they believe, reveal the dominant and underlying forms of a given culture. These are then 'decoded' or read like any 'text,' be it written or spoken, gestured or performed (Highmore, 2002b; Rosenhaft, 1987).

As the practitioners of microhistory and the history of everyday life have pointed out themselves, there are a number of other pan-European influences on their work – from the literary practices of Leo Tolstoy in *War and Peace* to modernist authors such as Marcel Proust and Italo Calvino, from Siegfried Kracauer's ideas about cinema and the supposedly discontinuous nature of reality to 'neorealist' Italian film. Other discernible influences include sociologist Erving Goffman's theories about the performative aspects of everyday interactions, Pierre Bourdieu's 'theory of practical action,' Henri Lefebvre's evocative work on 'everyday life in the modern world,' as well as Michel de Certeau's ruminations on 'social practices' and the 'reappropriation' of cultural objects in resistance to the rhythms and demands of consumerist modernity. Michel Foucault's work on persecution, madness, and the oppressive nature of modern institutions, as well as Sigmund Freud's attention to language and the way in which 'trifles,' 'slips,' and other 'insignificant details' can alert us to hidden meanings, have been important as well (Levi, 1991; Brewer, 2010; Muir, 1991; Highmore, 2002b).

Among historians the greatest influence has undoubtedly been E.P. Thompson, the doyen of self-critical Marxist historians associated with the British New Left. His pioneering work *The Making of the English Working Class* effectively launched 'history from below' in the 1960s, with its emphasis on consciousness, culture, and the everyday conditions and practices of those obscure individuals he hoped "to rescue…from the enormous condescension of posterity." Equally important were Thompson's ideas about crowd rationality and the central role of experience in identity formation (Thompson, 1963, 1971; Eley, 1989: 313). 'History from below' subsequently flourished in the United Kingdom and the United States, but it was in Italy and Germany that it assumed radical new forms. The main practitioners in Italy were active in the northern part of the country, which, as a whole, did not have a strong social history tradition. Carlo Ginzburg, Giovanno Levi, and other Italian 'microhistorians' coalesced around the Bolognese journal *Quaderni storici*, which began publishing specimens of the new genre in the late 1970s. Along with Lutz Niethammer, Göttingen historians Alf Lüdtke and Hans Medick established the practice of *Alltagsgeschichte* north of the Alps in Germany, where they published much of their own innovative work, as well as that of their colleagues, in the journal *Historische Anthropologie*. Though they share comparable concerns and reflect similar influences, *Alltagsgeschichte* and microhistory are not entirely the same undertaking. Whereas the latter gravitates toward the spectacular and 'unique' in order to get at lived experience itself, the former focuses on largely unconscious, routine, and repetitive acts in an attempt to reconstruct social relationships and daily transactions – which, they believe, are themselves the key to understanding historical development (Gregory, 1999: 103–104).

There are no American or French 'schools' strictly speaking, but Natalie Zemon Davis, Robert Darnton, and a number of other scholars in the United States produced a series of path-breaking studies that belong to this field – as did the French historian Emmanuel Le Roy Ladurie, whose reconstruction of social life in a medieval village became a model of the genre (Le Roy Ladurie, 1978).

Critiques of Microhistory and *Alltagsgeschichte*

In their classic studies *Montaillou* and *The Cheese and the Worms*, Le Roy Ladurie and Carlo Ginzburg both relied heavily on detailed interrogations of alleged heretics conducted by members of the Inquisition in France and Italy. Reading these sources against the grain, they picked out banal details and documentary fragments to get at larger historical issues as well as at more general socioeconomic and cultural patterns – such as the complex relationship between elite and popular culture. Since the focus is on those who, as a rule, left behind few written documents themselves, other important sources have included reports by police and church officials, teachers, physicians, and factory inspectors; personal correspondence and travelogues; parish registers, wills, notarial records, and protocols.

The often ingenious use of seemingly mundane source material has been widely praised by other scholars. But this reliance on documents left by the rich and powerful to get at the lives of the poor and oppressed has also been a source of hefty criticism. For instance, many of these documents, such as legal ones used in court cases, were designed to deceive and distort reality in ways that some historians have supposedly failed to grasp (Kuehn, 1989). The need for careful – and the possibility of sloppy – source critique is not peculiar to microhistory and the history of everyday life, of course. Still, its practitioners – especially microhistorians – have been especially sensitive to the potential dangers of using documents originally intended to convey information that is much different from the focus of their own scholarly pursuits. This is why many have consciously integrated their

research procedures and source analysis into the narrative itself. As Ginzburg and Levi explain, the 'obstacles interfering with the research,' such as 'lacunae or misrepresentations in the sources,' were 'constituent elements of the documentation' and thus became an 'intrinsic part of the account.' Similarly, the 'hypotheses, the doubts, the uncertainties,' as well as the researcher's very procedure and point of view are all included in the narration (Ginzburg, 1993: 22–23, 28; Levi, 1991: 106). By 'revealing their bag of tricks,' they create a sort of Brechtian 'estrangement' effect that 'defamiliarizes' or 'denaturalizes' the seemingly familiar and thus offers an especially fitting form for depicting the complexity of the everyday (Highmore, 2002b: 21–23).

Another weighty critique concerns the possibility that these historians will interpret sources, persons, and events in an anachronistic or ahistorical manner. The use of modern categories and discourses to describe and explain earlier experiences (instead of examining them 'on their own terms'), as well as the imposition of present sensibilities on the past, are all potential hazards of the enterprise – but again, not ones strictly limited to the various forms of 'people's history.' Still, the difficulty of interpreting past behavior and understanding the meaning(s) that historical actors ascribed to their actions is a staggeringly difficult enterprise (Bell, 2002: 271). Microhistorians and historians of the everyday also run the risk, and have been faulted for, romanticizing the past, focusing on 'picturesque detail,' and sentimentally celebrating ordinary people as heroes, proto-feminists, or freedom fighters; others have accused them of 'trivializing' the crimes committed by oppressive regimes such as the Third Reich by overstating the possibilities for resistance and nonconformity at the grass roots (Lüdtke, 1995: 10–12).

All of this draws attention to one of the great philosophical debates that microhistory and the history of everyday life bring front and center: the relationship between structure and agency, between free will and determinism. In fact, one of the greatest challenges of the genre is navigating between "the Scylla of blind historical forces that determine individual behaviors, and the Charybdis of a romanticized self-determination by radically free historical actors" (Gregory, 1999: 105) – i.e., not exaggerating or distorting the possibilities of agency in highly repressive societies (or in ones that used more refined methods of repression and surveillance in an effort to ensure conformity), or of downplaying the extent to which larger structures and institutional forces shape the lives of ordinary individuals. As Karl Marx famously put it in *The Eighteenth Brumaire of Louis Bonaparte* (1852), "Men make their own history, but they do not make it just as they please; they do not make it under circumstances chosen by themselves, but under circumstances already existing, given, and transmitted from the past." But historians of everyday life try to show the ways in which 'collective phenomena' are themselves the very result of individual agency and the actions of ordinary people: "The systems that presumably govern the world are not alien to everyday life but immanent in it; the living of everyday life itself is itself part and parcel of the process by which structure and ideology are reproduced and transformed." In other words, ordinary people themselves supposedly "hold the lion's share of responsibility for making that collective world" (Steege et al., 2008: 367–368, 372).

Scholars working on everyday life in Germany during the Third Reich have, in fact, been especially adept at describing the 'ambivalences and ambiguities' (Eley, 1989: 325) – the 'dark side' (Port, 2013) – of behavior at the grass roots, pointing out the ways in which ordinary Germans supported and profited from the regime's discriminatory policies at the same time that they bucked the system and pursued their own interests. Alf Lüdtke has aptly described this practice involving both complicity and resistance as *Eigensinn*, a notoriously difficult term to translate or even define precisely, but one that draws attention to the 'sliver of autonomy' (Bucur et al., 2009: 190) enjoyed by the seemingly powerless and weak – while, at the same time, rightly emphasizing the ambiguous and multivalent nature of most behavior (Lüdtke, 1995). This draws attention, in turn, to the way in which the everyday provides both a 'training ground' of sorts for conformity, as well as a site where conformity can be 'evaded' and acts of nonconformity can take place (Highmore, 2002b: 5).

Historians of everyday life and microhistorians have also had to contend with another criticism frequently leveled since local or 'case' studies first became popular in the 1960s: this involves the extent to which the individual, community, or event at the center of a given study are truly 'representative' of larger trends. All persons, communities, and historical events are in some way 'unique,' of course. But since the adoption of social science methods and insights in historical research, as well as E.H. Carr's calls in the early 1960s for a greater emphasis on causation and sophisticated historical analysis (as opposed to 'mere' narration), historians have been at pains to make larger analytical claims about the importance of their own often narrowly focused studies and relate the particular to the general, i.e., to larger historical trends and developments.

Microhistorians have been especially cognizant of this thorny challenge, given the radically reduced scale of their investigations, as well as their resolute insistence on context. Some concede the need for a comparative dimension, i.e., the need to compare their findings and conclusions to other similarly circumscribed studies. But they rightly insist that some issues can best be understood at the micro level – and others, for that matter, at the macro level. Institutions and institutional power, for example, would fall into the latter category, intimate relationships within the family and the very fluidity of identity in the other. A focus on formal power relations, institutions, and high-level policy often fails to comprehend or explain how and why things developed as they did on the ground – often in ways not intended by authorities and elites, as recent research on everyday life in the Soviet bloc has revealed (Port, 2007). In other words, the 'microphysics of power' at the grass roots can effectively serve to undermine and mitigate against the power of larger structural forces and processes (Steege et al., 2008: 361). Still, integrating and relating those two levels in a meaningful manner – balancing intimacy and distance, establishing causal connections – is one of the greatest challenges faced by microhistorians and historians of the everyday. The following cartographic analogy is useful for making this point: "it is a mistake to hand someone an architectural blueprint if he or she wants to drive to another city. The key is knowing which map we need in a particular instance, and

how to combine harmoniously maps of different scales" (Gregory, 1999: 109). Siegfried Kracauer believed that Marc Bloch, one of the founders of the *Annales* School, offered a venerable model in his two-volume *Feudal Society* (1939–40), which provides "a constant back and forth between micro- and macrohistory, between close-ups and extreme long shots" (Levi, 1991: 27).

In response to the issue of 'representativeness,' microhistorians in particular have embraced the concept of the 'normal exception,' a term originally coined by Edoardo Grendi. These are forms of seemingly unusual behavior that defy prevailing norms, which is the very reason they leave traces in the archives. But they do so on a regular basis, and – though contrary to what those in power and other elites consider to be 'normal' behavior – are completely normal for ordinary individuals, especially those on the socioeconomic margins of society. Such 'transgressions' are, in fact, 'perfectly representative of their own social milieu' (Muir, 1991: xiv). Of course, the ability to recognize something as a 'normal exception' in the first place posits an a priori grasp of larger patterns only afforded by a macro view. The focus on that which fails to fit into grand frameworks or narratives, that throws a wrench in the relentless search for regularity and predictability, highlights, in short, the complexity of reality, of the 'really real' – which is, after all, what everyday life at the grass roots truly claims to represent, or at least one significant aspect of it.

The Future of 'People's History'

The history of everyday life and microhistory enjoyed tremendous popularity in the 1980s. *Alltagsgeschichte* itself was hailed at the time as "the most important new departure" in West German historiography (Eley, 1989: 297). But by the end of the following decade, it, along with social history as a whole, had been largely eclipsed by the 'new cultural history' popular in the United States, which looked to the insights of postmodernism and the so-called linguistic turn for inspiration. Despite important differences in terms of method, assumptions, and intellectual influences, all of these new approaches shared a certain affinity in terms of what they criticized about the older historiography's focus on major events, powerful individuals, and dominant institutions – and then, with the rise of social science history, on the gothic structures and intangible historical forces associated with a steadfast materialist interpretation of 'how history works.'

Even though 'history from below,' microhistory, and *Alltagsgeschichte* are no longer as popular as they once were, reports of their death have indeed been 'greatly exaggerated' (Steege et al., 2008: 358). A number of historians in Europe and the United States continue to produce important specimens of this genre and continue to discuss its theoretical underpinnings and potential contributions – for example, in writing about the states that made up the now-defunct Soviet bloc. By examining the complex relationship between 'state' and 'society,' their post-1945 history suggests the fruitful ways in which an understanding of life at the grass roots can help explain developments and trajectories that an investigation of institutions, parties, and high-level politics cannot alone. It shows, in short, that even under regimes where the power relationship between the 'rulers' and 'ruled' was clearly asymmetrical, developments at the micro level could and did have a major influence on those at the macro.

It is a truism that contemporary politics can have a significant effect on the types of history pursued by a given generation, as well as on the research topics that interest its members. The sudden popularity of 'history from below' in the 1960s – especially the interest in ordinary people, social protest, and the behavior of 'crowds' – clearly reflected the political turmoil, social upheaval, and critical atmosphere of that decade. Changing political sensibilities and disappointments have since then marked a dramatic shift away from an interest in the plight of the working classes to that of other groups, including women, ethnic minorities, and colonial subjects. As Wendy Goldman has noted, "Today, historians seem more interested in identity, nationality, ethnicity, globalisation, and transnational histories, which reflect in turn preoccupations of our contemporary world" (Bucur et al., 2009: 199). That is undoubtedly true, yet recent political developments in a world increasingly dominated by global financial institutions and structures seemingly beyond the control of ordinary individuals – from the 'Arab Spring' in the Middle East to the emergence of the 'Occupy Movement' in the West – might very well spark a renewed interest in 'people's history' and the emphasis it places on the power and agency of ordinary individuals.

See also: Annales School; Anthropology and History; Boas, Franz (1858–1942); Cultural History; Ethnography; Everyday Life, Anthropology of; Family and Kinship, History of; Foucault, Michel (1926–84); Geertz, Clifford (1926–2006); Goffman, Erving (1922–1982); Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Levi-Strauss, Claude (1908–2009); Life-style, History of the Concept; Materiality and Culture; Thick Description: Methodology.

Bibliography

Bell, D.A., 2002. Total history and microhistory: the French and Italian paradigms. In: Kramer, L., Maza, S. (Eds.), A Companion to Western Historical Thought. Blackwell Publishers, Malden, MA and Oxford, pp. 262–276.

Brewer, J., 2010. Microhistory and the histories of everyday life. Cultural and Social History 7 (1), 87–109.

Bucur, M., Gavrilova, R., Goldman, W., Healy, M., Lebow, K., Pittaway, M., 2009. Six historians in search of Alltagsgeschichte. Aspasia 3, 189–212.

Darnton, R., 1984. The Great Cat Massacre and Other Episodes in French Cultural History. Basic Books, New York.

Eley, G., 1989. Labor history, social history, *Alltagsgeschichte*: experience, culture, and the politics of the everyday – a new direction for German social history. The Journal of Modern History 61 (2), 297–343.

Eley, G., 2005. A Crooked Line: From Cultural History to the History of Society. The University of Michigan Press, Ann Arbor, MI.

Ginzburg, C., 1980. The Cheese and the Worms: The Cosmos of a Sixteenth-Century Miller (Anne Tedeschi, Trans.). The Johns Hopkins University Press, Baltimore, MD.

Ginzburg, C., 1993. Microhistory: two or three things that I know about it. Critical Inquiry 20 (1), 10–35.

Gregory, B., 1999. Is small beautiful? Microhistory and the history of everyday life. History and Theory 38 (1), 100–110.

Highmore, B., 2002a. Everyday Life and Cultural Theory: An Introduction. Routledge, London and New York.

Highmore, B., 2002b. Introduction: questioning everyday life. In: Highmore, B. (Ed.), The Everyday Life Reader. Routledge, London and New York, pp. 1–34.

Hunt, L., 1989. Introduction: history, culture, and text. In: Hunt, L. (Ed.), The New Cultural History, pp. 1–22.
Kuehn, T., 1989. Reading microhistory: the example of Giovanni and Lusanna. The Journal of Modern History 61 (3), 512–534.
Le Roy Ladurie, E., 1978. Montaillou: Cathars and Catholics in an French Village (Barbara Bray, Trans.). Scolar, London, pp. 1294–1324.
Lepore, J., 2001. Historians who love too much: reflections on microhistory and biography. The Journal of American History 88 (1), 129–144.
Levi, G., 1988. Inheriting Power: The Story of an Exorcist (L.G. Cochrane, Trans.). The University of Chicago Press, Chicago, IL.
Levi, G., 1991. On microhistory. In: Burke, P. (Ed.), New Perspectives on Historical Writing. Polity Press, Cambridge, pp. 93–113.
Lüdtke, A., 1993. Eigen-Sinn. Fabrikalltag, Arbeitererfahrungen und Politik vom Kaiserreich bis in den Fascismus. Ergebnisse Verlag, Hamburg.
Lüdtke, A., 1995. Introduction (W. Templer, Trans.). In: Lüdtke, A. (Ed.), The History of Everyday Life: Reconstructing Historical Experiences and Ways of Life. Princeton University Press, Princeton, NJ, pp. 3–40.
Muir, E., 1991. Introduction: observing trifles. In: Muir, E., Ruggiero, G. (Eds.), Microhistory and the Lost Peoples of Europe. The Johns Hopkins University Press, Baltimore and London, pp. vii–xxviii.
Port, A.I., 2007. Conflict and Stability in the German Democratic Republic. Cambridge University Press, Cambridge and New York.
Port, A.I., 2013. East German workers and the "dark side" of Eigensinn: divisive shop-floor practices and the failed revolution of 17 June 1953. In: Berghoff, H., Balbier, U. (Eds.), The East German Economy, 1945–2010: Falling Behind or Catching Up? Cambridge University Press, Cambridge and New York, pp. 111–128.
Revel, J., 1995. Microanalysis and the construction of the social. In: Revel, J., Hunt, L. (Eds.), Histories: French Constructions of the Past. The New Press, New York, pp. 492–502.
Rosenhaft, E., 1987. History, anthropology, and the study of everyday life: a review article. Comparative Studies in State and History 29 (1), 99–105.
Sabean, D.W., 1990. Production, Property, and Family in Neckarhausen, 1700–1870. Cambridge University Press, Cambridge and New York.
Sabean, D.W., 1998. Kinship in Neckarhausen, 1700–1870. Cambridge University Press, Cambridge and New York.
Smith, H.W., 2002. The Butcher's Tale: Murder and Anti-Semitism in a German Town. W.W. Norton & Co., New York.
Steege, P., Bergerson, A.S., Healy, M., Swett, P.E., 2008. The history of everyday life: a second chapter. The Journal of Modern History 80 (2), 358–378.
Thompson, E.P., 1963. The Making of the English Working Class. Vintage Books, New York.
Thompson, E.P., 1971. The moral economy of the English crowd in the eighteenth century. Past & Present 50, 76–136.

Relevant Websites

www.microhistory.org – Microhistory.org.
www.microhistory.eu/bibliography.html – Microhistory.org Bibliography.
www.mulino.it/edizioni/riviste/issn/0301-6307 – il Mulino.
www.werkstattgeschichte.de – History Workshop.

History: Overview

Jörn Rüsen, Institute for Advanced Study in the Humanities, Essen, Germany; and University of Witten/Herdecke, Bochum, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents history as a fundamental factor of cultural life. It addresses its anthropological fundaments in the human endeavor to make sense out of the experience of time. It tackles the narrative form of historical knowledge and the typology of historical representation. Special reference is made to the problem of intercultural communication and the power of ethnocentrism, its danger and chances to overcome it.

Anthropological Universals

On a general and fundamental anthropological level, history is the totality of the cultural practices through which human beings interpret their past in order to understand the present and anticipate the future. In other words, history is the generation of sense out of the experience of time.

But not every interpretation of time is historical. What determines the historical quality of an interpretation is the fact that time confronts human beings in two ways: *as real change within the world*, in which contingent events occur and are experienced, and as an *interior, temporal extension of consciousness* (and of the unconscious). Here, memories, imaginations, knowledge, intentions, and objectives, guided by the elemental mental forces of fear, anxiety, desire, hope, and longing, play a central role. We interpret the 'objective' time of real change (the present) through the recollection of another time (the past); thus the 'subjective' time into which we project a future is bound to our approach to past experiences. This is human culture's specifically historical procedure for experience and interpretation. What defines the historical interpretation of time is its implicitly meaningful and thus orientative conception of the course of time. The latter comprises the past, the present, and the future and at the same time articulates experiences and intentions; it is thus empirical and normative at the same time. The philosophy of history explicates this conception of the course of time as a particular mental structure. Even without philosophical reflection or justification, such a conception of the course of time organizes historical thought, and shapes its content. History interprets the contingent events of the past and defines their significance for the present. It organizes historical narrative as a particular kind of explanation (Danto, 1965). It is this 'explanation through narration' that defines the logic of historical thought in its academic form as well.

Conceptions of Time

The different conceptions of the course of time important for historical interpretation employ various temporal categories in various constellations (e.g., cyclical time, linear time, the concept of the *kairos*), each encouraging particular expectations: decline, progress, permanence, recurrence, revolution, destruction, and so on. They relate real temporal processes to the active intentions and self-conceptions of the subjects that make use of them. In other words, history mediates the subjectivity of acting and suffering human beings with the objectivity of the temporal process of their world (Rüsen, 2007). In early societies, this is accomplished by an appeal to divine forces, which impel the transformation of the world and the human spirit at the same time; in other societies, these divine forces are replaced by humankind as a spiritual and moral being, its ability to create symbolic worlds or to plan the conquest of the natural and social world through knowledge. Finally, nature too can be invoked as a sort of metaphysical instance for the continuity between the two temporal dimensions of human historicity. Nature can thus be understood as that which compels the human race, by means of social antagonisms, to create culture (as in Kant, for example), or as struggle of the races in Social Darwinist terms. Psychoanalysis, on the other hand, posits an interior, psychological nature, in order to explain the historical movement of the human world by means of the dynamics of the unconscious.

The idea of one single 'history' comprising all temporal change within the human world is of modern origin (Koselleck, 1985). However, it has a premodern equivalent in older cosmogonies, theogonies, and anthropogonies, as well as in the concept of redemptive history.

History as Event and Narration

In every culture and epoch 'history' has a *double meaning*. On the one hand, it designates *temporally organized events* in the past. At the same time, and inseparably from the first meaning, 'history' also means *the account of these events* that articulates their significance for our understanding of ourselves and of our world. The first meaning has to do with temporal succession, the second with its narrative representation.

Not everything that has happened can be called history, but rather only that which is somehow meaningful in the present: as for example the foundation of a city or a kingdom is significant to its inhabitants, as a person's birthday is significant to him or her, and so on. Usually this event is described in a sequence of occurrences. These may often be the actions of human beings, heroes, or divine beings, as in myth or epic; however, they may also involve more complicated phenomena, such as the development of a manufacturing process, the evolution in the meaning of a concept, the epochal transformation of a mentality, an 'axial time,' and so on.

In its second meaning, history is the representation of the past in the culture of the present. Here 'history' is the product of

historical consciousness: By the work of memory the past has come to life and meaning in a narrative form. In this way the past 'in itself,' that is, everything that happened, and has persisted as trace, relic, memorial, tradition, ethic, and so on becomes history 'for us.' I suggest to distinguish and organize the basic mental procedures involved in the following way:

- *Perception* of a different time that is owed to the fascination or the burdening awareness of the archaic, the obsolete, the mysterious trace, the persistent memorial, and so on;
- *Interpretation* of this time as temporal movement in the human world, according to some meaningful aspects (e.g., as evidence of the permanence of certain values, as examples of a general rule, as progress, etc.);
- *Orientation* of human practice by interpretation that can be generated in two ways – both by an external and an internal perspective. In the external perspective, human activity gets its temporal dimension relevant for shaping its intentions. Examples for this perspective are an increase of political legitimacy through political participation, or the restitution of the world before its destruction, or the renewal of acceptable life forms against the decline of morals. In the internal perspective human subjectivity gets its temporal dimension, where identity as a strong belief in belonging to and be different from others is at stake (e.g., "We are the children of the sun," or "We as a nation stand for the universality and achievements of human rights," or "We belong to the communion of saints," or "We represent true spirituality in contrast to the others' materialism"); and finally
- *Motivation for action* provided by this orientation (e.g., willingness to sacrifice oneself, even die for a certain ideal etc., or to kill others for the sake of historical ideas of national greatness, the missionary spirit, etc.)

Memory, History, and Master Narratives

There is an ongoing debate about the relationship between *memory* and *historical consciousness* (Megill, 1998; Seixas, 2004). Memory in the everyday sense refers to past experiences of individuals in their own lives, whereas historical consciousness primarily thematizes a past that exceeds an individual life span. Both, however – the individual memory on which individuality and social affiliation are built as well as the expansion of 'history' beyond the limits of a single life span back into the past – are inseparable like two sides of a coin: human beings tend to transform their own identities into temporally more comprising (including) structures, such as the community of saints, the nation, the people, the culture, and so on. The basis for the individuals' self-esteem and their practical temporal orientation tends to exceed their life span or to outlive them. It thus usually reaches far back into the past (Assmann, 1995).

'History' generally performs its orientative function by a 'master narrative' or 'metanarrative.' This describes the conception and evolution of the addressee's world in order to legitimate its normative framework. At the same time, it provides an experiential inventory that is available whenever problems of orientation require a solution by consensus. It depicts, confirms, and legitimates not only systems of politics, economics, and environment, but also the collective and individual subjectivity and identity of the people. Master narratives represent social affiliation and differentiation, internally and externally.

Master narratives generally organize social inclusion and exclusion *ethnocentrically*, ascribe positive normative characteristics and weight to one group (mostly one's own) and negative ones to other different groups. One's own lifeworld appears as cultured or civilized, whereas other worlds are looked upon as savage and barbaric. This difference can be articulated in different ways: as utterly opposed to one's own, as deviation from a norm, as retarded in development, etc. Just as history presents past experiences as a 'mirror' in which the present perceives itself, the image of others in the same historical mirror has a contrastive function: it defines a self-image. Under special conditions of an orientation crisis, however, the 'other' may be ascribed admirable features, as we can find in the image of the 'noble savage.'

Historical Culture

The interpretative work of historical consciousness and its product called 'history' becomes manifest in a society's *historical culture* (Karlsson, 2011). Historical culture is multidimensional, like every other part of human culture. It has religious, ethical, pedagogical, political, and rhetorical expressions; its cognitive substance always aims at the knowledge of 'wie es eigentlich gewesen' (how it really was). In an artificial way, we can distinguish five basic dimensions of historical culture as ideal types, each of which is quite different in its logic and thus accountable to different criteria of meaning:

- The *political* dimension is concerned with the legitimation of a political order in the context of historical thinking. Historical consciousness inscribes this order into the identity concepts of the people. Doing so the decisive principle of making sense of the past is *legitimacy* of power and domination.
- The *aesthetic* dimension is concerned with the emotional effectiveness of historical interpretations, or the domain which appeals to and affects the human senses. A strong historical orientation must always engage the senses. Pictures, sculptures, public monuments, and museums serve this purpose. Masquerades, dances, and music can all have historical content. Many older master narratives are composed in a poetic form and are ritually performed. Historical knowledge must employ literary models to become discursive. In doing so historical culture is guided by the sense criterion of formal coherence in a narrative representation and (traditionally called 'beauty').
- The *cognitive* dimension is concerned with the knowledge of past events significant for the present and the future. Without the element of knowledge, the recollection of the past cannot effectively be introduced into discourses concerned with the interpretation of current temporal experience. Mythical master narratives, too, have a cognitive status, though scientific method and practice may deprive them of it. They lose their orientative power when confronted with a knowledge of the past that possesses a more

elaborated relation to experience. In this dimension historical thinking refers to the sense criterion of truth.

- The *ethical* dimension is constituted by norms and values which stem from the social context of the historians. Here the past is subordinated to the judgments of the present. This relationship can be intensified to a concept of historical responsibility (Bédarida, 1995; Carr et al., 2004) that is based on an intergenerational connection between forefathers and offspring. The sense criterion in this dimension is the distinction between good and evil.
- The *religious* dimension refers to the temporal extension of sacred elements of human life. Here history is shaped by the sense criterion of redemption. Traditionally it opens up a perspective of transcendence in the representation of the past, but in modern times it may achieve a highly ideological form of innerworldly or secular religion (as developed in Marxism or fascism). Today, historical culture is characterized by an unreconciled relationship between secularism (most effectively manifested in historical studies as an academic discipline) and still powerful religious approaches to human life.

The interrelationship of these dimensions is complex, flexible, changeable, and full of tensions and conflicts. The different sense criteria try to dominate and instrumentalize each other. They need a discursive mode of interrelationship within which they may complement each other. A unifying principle of this discourse may be gained out of a cultural anthropology of historical thinking. It has to emphasize the potentials of overcoming ethnocentric factors in historical culture by referring to fundamental and universal values inherent in the cultural nature of every human being (Antweiler, 2012; Rüsen, 2012).

Typologies of Historical Representation

Synchronic and diachronic *typologies of the historical* can be generated out of the particulars of historical thought, and inquire into its carriers, functions, class and gender specificity, and other factors within the context of the social framework of determinations (Rüsen, 1996). Yet they only become historically specific, when they are capable to differentiate the specific interpretative work of historical consciousness and of its product – 'history.' Here history appears as the cultural orientation system of temporal consciousness – constructed according to the different potentials of historical meaning. Such a differentiation gained by a metatheoretical perspective has yet been little tackled. Previous efforts have usually focused on functional and formal aspects, and have less thematized the total domain of human temporal interpretation but simply the domain of historical writing. If at all only the Western tradition was considered. Nonetheless, such typologies can be used as analytical tools for the research in the wide areas of historical culture, especially for intercultural comparisons.

Nietzsche, for example, distinguished the monumental, the antiquarian, and the critical uses of history in terms of their advantages and disadvantages for life (Nietzsche, 1985).

Hayden White (1973) presented a very influential model for the analysis of historical writing, and applied it to the most significant historians and philosophers of history in the nineteenth century. It consists of a complex strategy of creating historical meaning by *emplotment* (with its variations of romance, tragedy, comedy, and satire), *formal argument* (with its variations of formalist, organicist, mechanistic, and contextual thinking), *ideological implication* (with its variations of anarchism, conservatism, radicalism, and liberalism), and *tropes* of historical discourse (with their variations of metaphorical, metonymic, synecdochic, or ironic uses of language).

White developed this typology for the analysis of historiographical texts; nevertheless, it can be applied to other manifestations of history, such as historical museums (cf Bann, 1984).

Rüsen (2005: pp. 9–19) also developed a universally applicable typology of historical sense generation according to its most basic principles: its particular references, its conception of temporal movement, its communicative form, and its identity concept. All these factors are communicated and constituted by the principle of historical sense:

- The *traditional type* recalls the origins grounding present relations. Traditional history represents the course of events in terms of the persistence of these origins and their normative power, and represents identity by affirming previous interpretive models of human subjectivity (e.g., social roles).
- The *exemplary type* recalls past facts that concretize the rules of human conduct. Exemplary history represents a course of events in terms of the transtemporal applicability of these rules, and constructs identity as rule-based competence. This narrative type is represented in Europe by the classical motto *historia vitae magistra*, in China sometimes by the metaphor of the mirror. This type is predominant in the historical thought of great civilizations up to modern times.
- The *genetic type* recalls qualitative changes in the past by which other and strange conditions become familiar ones in the present. The genetic construction of meaning presents a course of events as evolution, in which conditions change in order to become dynamically permanent. It constructs identity as the synthesis of permanence and change, in other words as a dynamic process. This type predominates in modern societies.
- The *critical type* recalls past events in order to use them as counterfacts against established patterns of historical orientation. Critical history constructs identity by negating pregiven identity-forming interpretive models, in other words, as the power to say 'no.'

Media

The diverse manifestation of history, in its dual significance as fact and as interpretation, crucially depends on the *medium* through which temporal experience and interpretation is communicated. *Language* is always predominant, although there are other creative areas of historical perception and representation, such as the visual and musical arts, architecture, and dance. The modes by which language puts human communication in order are decisive for the expression of historical consciousness.

History was originally communicated orally, as it is done now – in the private or informal sphere. In preliterate cultures, history becomes *oral history* or *oral tradition* (Vansina, 1985). Such a history is usually politically conservative: the traditional world order becomes obligatory; new experiences have to be integrated into it so as to confirm the collective's basic norms and identity. Such representations are sensually appealing to the senses and concrete; they are communicated face to face and correspondingly have immediate effects. The criteria of reference are of a mythical nature. Its validity claims are raised and made plausible through ritual and repetition. In modern historical research, *oral history* is the method for gaining historical knowledge from the recollections of people as actors and witnesses. Thus bit provides an important source for research about the mechanisms of biographical memory as an element of historical culture in the intersection of the public sphere and individual biography. The memories of witnesses about historically significant events, communicated orally, have obtained an important role in public historical culture – especially by the testimonies of survivors of great catastrophes such as the Holocaust (Langer, 1991).

The medium of *writing* has dominated for thousands of years and continues to dominate contemporary historical culture (in which, of course, oral elements continue as important and effective). Written history frees from the immediacy of the communicative situation and creates a distance between history as content and history as a form of communication (Ong, 1982). This distance increases the experiential horizon of historical consciousness to a large degree; it facilitates new methods for accumulating as well as managing experience. It provides new opportunities for objectivation in the approach to historical experience and it discloses new opportunities for subjectivation through historical interpretation as well: the narrator can become an author, and readers can obtain an expanded scope for critical readings. Validity claims are raised either through canonization, or become an opportunity for discursive justification. In either case, interpretation becomes a discrete practice of historical meaning construction; here lies its potential of specialization and the corresponding strategies of mediation and popularization. The original relationship between poetry and truth is dissolved; mythology has now become the potential subject for fundamental critique. Explicit conceptualization has received an essential cognitive element in the historical interpretation, so that history can, at least in the long run, become an object of study and theory.

Entirely *new media* have recently transformed the historical mode of thought. We cannot yet distinguish distinct lines of development and stable structures, but we can certainly indicate new developments from which we may expect transformations of the most basic kind.

The Internet has provided historical culture with an enormous input and enlargement of information and communication. It has opened up new possibilities of constructing individual historical perspectives according to the specific life situation of different people. It offers unlimited access to accumulating knowledge of the past and challenges the recipients to give it a peculiar meaning for their lives. By using this possibility everyone has become his or her own historian (Foner, 2002). This could lead to an anarchic fragmentation of historical culture and at the same time to an intensified communication with new chances of participation. It weakens pregiven (e.g., national) borders and limits of historical identity thus opening up new horizons of awareness: to be part of humanity as the broadest scope of placing one's own life into the temporal change of the human world.

This enlargement of the realm of historical culture goes along with a fundamental change in making sense of the past. In the historical culture of the public sphere, collective memory is flooded by a torrent of historical images. The forms of consciousness elaborated by scholars and literates (literacy) – and above all by the distancing work of rationality – are in danger of decreasing in significance, and especially in political efficiency. The grammar of history is progressively becoming imagology of presentations in which every era is contemporaneous, and the fundamental idea of a fundamental linear movement of time is disappearing. The constitutive difference of temporality thus might be suspended into a universal contemporaneity that can no longer be narratively put into order. Whether there can then be a specific 'historical order' within the orientative temporal continuity between the past, the present, and the future has at least become a problem. The very term of *posthistoire*, and the related discussion of a mode of life without genuinely historical interpretation (Niethammer, 1992), suggests that these questions are now open. At the same time, there has been an immense increase in access to data. New storage media allow new modes of historical experience, and radically call into question previous criteria of significance. At the same time, new media of communication – such as the Internet – permit no isolated, politically sanctioned decision about the meaning of history The abundance of possibilities and the diversity of new voices require new strategies, new forms, and new contents of historically grounded participation and exclusion. In each case, fixed ideas of the permanence and substance of individual and collective identity are being outstripped by the diversity of global communication, in favor of more dynamic and more open differentiations. This process then provokes reactions, which stubbornly insist on ethnocentric distinctions, often expressed through these new media as well.

Developments

The transformation in the media of historical thought as sketched above suggests a *transcultural*, *macrohistorical transformation* that will ultimately leave no culture unaffected. This may provide us with a cognitive mean for a diachronic classification of historiographical variety. Such a classification could be organized in terms of epochal distinctions, with still a typological quality. For example, White's or Rüsen's typologies could both be used to characterize historical processes. White's sequence of tropes can be understood as cyclical, and Rüsen's distinction between traditional, exemplary, and genetic historical thought can describe the great transcultural epochs where the critical type of historical thought mediates the transition between the other types.

One of the most important distinctions concerns the relationship between myth and 'history,' where the latter treats worldly affairs. Myth on the contrary, relates to a divine

prehistory as the source of meaning (Eliade, 1991). In contrast, 'historical thought' in the stricter sense discerns the meaning of history in the immanent dimension of a chain of events (*res gestae*). The Judeo-Christian sacred history mediates between these two modes of thought, and allows events in the world to appear as the manifestation of God's power and God's will. In other cultures, the divine and the mundane are mediated differently: mundane but somehow significant events have to be integrated into a comprehensive cosmic order. History becomes meaningful through nature, which is conceived as divinely ordered. With increasing complexity in human praxis, more and more mundane incidents are integrated into this order until it is finally 'historicized' and 'denaturalized.' Yet this process is often opposed by attempts to renaturalize it as biological racism.

Comprehensive large-scale developments of historical thought have very often been conceptualized as leading to the current status of historical studies. This has often been criticized as too one-sided, so the diversity of the new media and the different dimensions of the historical culture are emphasized instead. Nevertheless, plausible accounts of a developmental process can be articulated (in the form of typological hypotheses). These accounts can be used to discern and analyze transformations of historical thought in the diversity of different cultures.

Max Weber's conception of a universal process of *rationalization and disenchantment* appears as relevant as ever (Weber, 1988). The classical philosophy of history of Voltaire, Kant, and Hegel presented these criteria in the form of the actual course of events. Since historical studies acquired increased competence in investigating and representing real events in the past, the question of meaning was investigated within the framework of a logic of historical thought (Dilthey, Max Weber, Rickert). Presently this question is perceived primarily as the concern of the central verbal procedure of historical thought, as historical narrative (Danto, 1965). Issues of scholarly standards and truth claims (Evans, 1999) might be obscured by only focusing on the rhetorical and poetic strategies of historical writing (Rigney, 1990).

Despite this formalization of historical understanding the structures of real historical processes remain important points of reference. Here the global and often conflictual transformations brought about by modern industrial societies are in the foreground. Their developmental dynamics can be understood only by a concept of time in which the horizon of experience and expectation structurally diverge, and which can only be reconnected by a fundamental temporalization of the human world (Koselleck, 1985).

Narratives of progressive rationalization and disenchantment must, of course, be complemented by an account of a critical and transcendent re-enchantment resistant to methodical rationality, such as the endlessly renewed rationality of sense (value rationality); it refuses to dissolve in the methodical rationality of the control of experience. Historical studies, the development of which in Europe from the eighteenth century until now has been the most effective factor in the modernization of the historical culture, must employ sense criteria that methodical rationality alone cannot generate (Iggers et al., 2008).

The rationalization of modern historical culture is constituted by a universalization of the norms and values that render the past significant for the present. In the realm of these values, the ethnocentric logic of identity formation plays an enormously important role. It has always had a universalizing tendency, insofar as it describes its own world as both unique and normatively obligatory. In order to overcome the conflict-bearing power of ethnocentrism, one should emphasize the concept of 'humanity.' Humanity is a common frame of reference for historical identity. In the focus of historical experience, humanity is extended to the limits of the known world. In the logical framework of ethnocentrism this universality creates a fundamental, structural conflict of cultures in historical thought that has repeatedly led to violence, to the legitimation of domination, and even to genocide.

Problems of Intercultural Communication

This leads us to a fundamental problem in contemporary historical thought. It is shaped by the rationalism that culminates in historical scholarship and related human sciences (Bentley, 1997). This rationalization of history as scholarship corresponds to the modern rationalization of all knowledge, without which the organization of modern societies would be inconceivable. Is historical thought inseparable from the value system of Western culture in its current form along with this methodological rationality? If so, how can it conceive history in such a way that it accounts for the diversity in different cultures' articulations of 'history' as the orientative framework of human praxis? Does this diversity, once it has been researched and analyzed by scholarship, not always appear in the light of a particular interpretation – namely the interpretation to which methodological reason essentially belongs? (Seth, 2004)

Current tendencies in historical thought (postmodernism, postcolonialism, the 'cultural turn') are determined to overcome modernity in this sense. They want to create a pluralism of 'histories' that can break the spell of a uniforming rationality and a generalizing normativity. They thereby destroy the orientative power of modern historical thought with its notion of a unified direction of development, or 'progress.' At the same time they problematize the claims of scholarship that describe the laws of such development along with the desire for mastering them, e.g., as we can find in Marxism.

In the age of globalization, in spite of the critique of the claim for hegemony inherent in Western historical thought, we cannot do without either methodological rationality or the universal regulative of communication. We need this regulative to understand the present by the interpretation of the past, to define inclusion and exclusion, and to act in anticipation of the future. If the ethnocentric logic of such a universalization is to be broken or overcome, the truly universal regulative of historical thought and of historical culture can only *be a mutual and thereby universal recognition of difference and particularity*. The practice of this recognition is based upon and guided by a fundamental and universal idea of human dignity, and this idea serves as a generally acceptable principle of criticism. It is and will remain counterfactual, even utopian; nevertheless, it may disclose new dimensions of historical experience and lead to new and politically significant orientations of human life in time (Kozlarek et al., 2012).

Ethnocentrism is still effective in historical thinking. It is mainly criticized as an established form in the tradition of

historical studies as they have emerged in the West since the end of the eighteenth century. Here universalistic concepts of historical interpretation can be deciphered as generalized cultural peculiarities of the West. Postcolonialism therefore contradicts these ideas, but by doing so it uses the logic of ethnocentrism by negation. Other attempts of refusing Western traditions in historical thinking replace them by non-Western ideas, and by doing so, they create new ethnocentric attitudes that are becoming visible, e.g., as Sinocentrism (Huang, 2007; Sato, 2007).

There is no neutral standpoint beyond different cultural orientations. Postmodernism therefore proposes a relativistic idea of history, where every human is his or her own historian, and where the general validity of historical cognition is only a 'noble dream' (Novick, 1988). But this relativism destroys any possibility of critical argumentation in intercultural communication. It prevents a cognitive achievement, which may meet the challenge of globalization, which has to bring cultural difference into a new mode of thinking beyond the destructive power of ethnocentrism. In this new mode, difference and unity have to be recognized and synthesized at the same time.

This is possible by referring to anthropological universals in historical thinking. They frame the diversity of human life forms in time and space by a new idea of humankind (Antweiler, 2012; Rüsen and Laass, 2009). In difference to the tradition of philosophical thinking about history, which has unified it as one encompassing development of human life forms, this idea makes the unity visible in the diversity of life forms and their changes. The unity of humankind can only be understood by referring to the diversity and changeability of civilizations. By historical thinking and its hermeneutical approach to cultural difference this synthesis of unity and diversity can be brought about and achieved.

This new way of historical thinking, by which it meets the challenge of globalization, has to answer another challenge as well. The terrors and catastrophes of the twentieth century, such as the Holocaust, demand for new categories of historical thinking and new forms of historical narration. Without the categories of inhumanity and suffering historical thinking will miss this challenge. And without applying the possibility of integrating senselessness into its meaning – construction concerning the past, historical thinking will miss the dark side of human history, that cannot certainly be overlooked any longer (Friedländer, 1998).

See also: Comparative History; Cultural History; Global History: Universal and World; Historical Thought and Historiography: Current Trends; Historicism; History, Professionalization of; History: Representation Within and Functions of; History: Theories and Methods; Identity and Identification; Integration and Multiculturalism: Focus on Western Europe; Intellectual History; Media and History; Modernization and Modernity in History; Multiculturalism; Time, Chronology, and Periodization in History.

Bibliography

Antweiler, C., 2012. Anthropological Basics towards an Inclusive Humanism – On a Realistic Cosmopolitanism in the Age of Globalization. V&R unipress, Göttingen.
Assmann, J., 1995. Collective memory and cultural identity. New German Critique 65, 125–133.
Bann, S., 1984. The Clothing of Clio. Cambridge UP, Cambridge.
Bédarida, F. (Ed.), 1995. The Social Responsibility of the Historian. Berghahn Books, Oxford.
Bentley, M. (Ed.), 1997. Companion to Historiography. Routledge, London.
Carr, D., 1986. Time, narrative and history. In: Edie, J.M. (Ed.), Studies in Phenomenolgy and Existential Philosophy, second ed. Indiana UP, Bloomington, IN, 1991.
Carr, D., Flynn, T., Makkreel, R. (Eds.), 2004. The Ethics of History. Northwestern UP, Evanston, IL.
Danto, A., 1965. Analytical Philosophy of History. Cambridge UP, Cambridge.
Eliade, M., 1991. Myth of the Eternal Return, or Cosmos and History. Princeton UP, Princeton, NJ.
Evans, R., 1999. In Defense of History. W.W. Norton, London.
Foner, E., 2002. Who Owns History? Rethinking the Past in a Changing World. Hill and Wang, New York.
Friedländer, S., 1998. Writing the history of the Shoa: some major dilemmas. In: Blanke, Horst-Walter, Jaeger, Friedrich, Sandkühler, Thomas (Eds.), Dimensionen der Historik: Geschichtstheorie, Wissenschaftsgeschichte und Geschichtskultur heute. Geburtstag, Cologne, pp. 407–414.
Huang, C., 2007. The defining character of Chinese historical thinking. History and Theory 46, 180–188.
Iggers, G.G., Wang, Q.E., Mukherjee, S., 2008. A Global History of Modern Historiography. Pearson Longman, Harlow.
Karlsson, K.-G., 2011. Processing time – on the manifestation and activation of historical consciousness. In: Bjerk, Helle, Lenz, Claudia, Thorstensen, Erik (Eds.), Historizing the Using of the Past. Scadinavian Perspectives on History Culture, Historical Consciousness and Didactics of History Related to World War II. Transcript, Bielefeld, pp. 129–144.
Koselleck, R., 1985. Futures Past: On the Semantics of Historical Time. MIT Press, Cambridge, MA.
Kozlarek, O., Rüsen, J., Wolff, E. (Eds.), 2012. Shaping a Humane World. Civilizations, Axial Times, Modernities, Humanisms. Transcript, Bielefeld, Germany.
Langer, L., 1991. Holocaust Testimonies. The Ruins of Memory. Yale UP, New Haven.
Megill, A., 1998. History, memory, identity. History of the Human Sciences 11 (3), 37–62.
Müller, K., Rüsen, J. (Eds.), 1997. Historische Sinnbildung—Problemstellungen, Zeitkonzepte, Wahrnehmungshorizonte, Darstellungsstrategien. Rowohlts Enzyklopädie, Reinbek.
Niethammer, L., 1992. Posthistoire: Has History Come to an End? Reinbek, London.
Nietzsche, F., 1985. The Use and Abuse of History (A. Collins, Trans.). Macmillan, New York.
Nora, P., 1989. Between memory and history: Les Lieux de Mémoire. Representations 26, 7–25.
Novick, P., 1988. That Noble Dream. The "Objectivity-Question" and the American Historical Profession. Cambridge UP, New York, Cambridge.
Ong, W., 1982. Orality and Literacy: The Technologizing of the Word. Routledge, London.
Ricoeur, P., 1984–1985. Time and Narrative, vols. 3. Chicago UP, Chicago.
Ricoeur, P., 1998. Das Rätsel der Vergangenheit. Erinnern—Vergessen—Verzeihen. Essener Kulturwissenschaftliche Vorträge, vol. 2. Wallstein, Göttingen.
Rigney, A., 1990. The Rhetoric of Historical Representation: Three Narrative Histories of the French Revolution. Cambridge UP, New York.
Rüsen, J., 1996. Some theoretical approaches to intercultural comparison of historiography. History and Theory, Theme Issue: Chinese Historiography in Comparative Perspective 35, 5–22.
Rüsen, J., 2005. History. Narration – Interpretation – Orientation. Berghahn Books, New York.
Rüsen, J. (Ed.), 2007. Time and History: The Variety of Cultures. Berghahn Books, New York.
Rüsen, J. (Ed.), 2012. Approaching Humankind – Towards an Intercultural Humanism. V&R unipress, Göttingen.
Rüsen, J., Laass, H. (Eds.), 2009. Humanism in Intercultural Perspective: Experiences and Expectations. Transcript, Bielefeld.
Sato, M., 2007. The archetype of history in the confucian ecumene. History and Theory 46, 218–232.
Seixas, P. (Ed.), 2004. Theorising Historical Consciousness. University of Toronto Press, Toronto.
Seth, S., 2004. Reason or reasoning? Clio or Siva? Social Text 78, 22 (1), 85–101.
Vansina, J., 1985. Oral Tradition as History. James Currey and Heine-mann, London.
Weber, M., 1988. In: Lassman, Peter, et al. (Eds.), Science as a Vocation. Hammersmith.
White, H., 1973. Metahistory. The Historical Imagination in Nineteenth Century Europe. Johns Hopkins UP, Baltimore.

History, Professionalization of

Rolf Torstendahl, Uppsala University, Uppsala, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A distinction is made between a historical professional who is able to live from writing or else producing history, and a historical professional who is regarded as a fully competent historian by the academic community. The history of professionals in the first sense is followed from Antiquity (especially old China) to the present. The second sense of the term is newer than the phenomenon that started to develop in the eighteenth century and found a full dissemination in the nineteenth and twentieth centuries. Different stages of historical professionalism are discerned, starting with Ranke's followers from around 1830, whose state-centered professionalism was followed by a methodological phase from the 1880s. After 1950, other criteria than method were given priority, and these have come to compete with each other.

It is evident that the term 'professional' in this connection has two different meanings, which are related to each other but are not equivalent. The first (a) is professional in the sense of being able to live by the trade (in the meaning of an occupation, such as the French word 'profession'). The second (b) is professional as a qualified person who is recognized by society (not least by his or her fellows in academia) as competent to practice the trade with a closure directed against nonprofessionals, that is, amateurs and those writing on history without being 'real' researchers.

The Historian's Profession as Breadwinning

In the first sense of professionalism (see Professions and Professionalization, History of), there have been professional historians for a long time but there is no distinct collective process of professionalization. The writing of history is age-old. Some famous history writers were active in European antiquity, and especially Herodotus and Thucydides of Greece have won lasting fame. These and other early European historians were not writing history for their living. In China from at least the second century BC, and later in Korea and Japan, historians were engaged to describe the history of their country under a dynasty. This usage involved a real employment in China, and it even entailed a personal responsibility of the historian for what he wrote. Sima Qian (born *c.* 145 BC) was employed as Grand Scribe (or Historian) by Emperor Wu of the Han dynasty in 105 BC, which made it possible for him to live by his professional work. He was able to do researches in the imperial library and archives and thus composed his major history, the *Shiji*. Besides that, he worked on an improved calendar system and a sort of theoretical discourse on history (Wang, 1995; Nienhauser 2011). He and other Chinese dynastic historians were obviously fully occupied and cannot even normally have worked alone but must have had a small office with clerks at their disposal (there were subordinate scribes). The Chinese professionalism among its historians had no known counterparts. In Korea, the first well-known historians (eighth century AD) were politicians who wrote on their own behalf. In India, there were no real historians until the thirteenth century because of the dominant cyclical time perception in Hindu culture. The late Roman and medieval European historiographers had little profit from their writing and could not live from it. "A book belonged to its author for as long as it stayed in his hands. After that it was free as air" (Smalley, 1974). Some late medieval chroniclers in Europe, however, managed to live by their work as historians. Such chroniclers were employed by princes or cities in order to relate the important events that took place within the principality or city.

In Europe, the chroniclers employed by princes and cities (those active in monasteries were not employed for the purpose and so were not formally living by their historical writings) were followed later by official historiographers hired by kings and states for the 'proper' recording of the wars in which their masters were engaged. Titles like Historiographer of the Realm (in variations all over Europe) make clear the official mission of the historian. The French Historiographe du Roi was a title used from 1554 to 1824, and 104 persons bore the title. It was not so much a function as a title with a reward, and the historiographers were not highly regarded at court (Boer, 1998). For most such official historiographers, the revenue from the title was only a side income and not enough to live on. Quite another type of professional historian in this sense arose with the great wars in Europe in the seventeenth century: those who lived from their reports from the war scenes. They established a form of market situation for themselves, and especially those behind the Theatrum Europaeum were successful in their branch, a sort of intermediary between history and journalism.

University professors of history appeared from the seventeenth century but became a group of importance only from the eighteenth century and then primarily in German states. Until the late nineteenth century, only full professors normally were able to live, and sometimes live well, by their salaries, while their subordinate teachers were poorly paid if at all (Lingelbach, 2011).

Already from the late eighteenth century, another type of professional historian, in sense (a) of the word, appeared in Europe: literary men writing history and more or less living by their literary accomplishments. Men like Voltaire, Gibbon, Robertson, and, to an extent, Schlözer and Herder made themselves known as historians (though primarily as learned and literary men), and their works were sold in editions of considerable numbers. Their ambition was to a large extent to give to the reader an impression of historical personalities and their struggle with moral problems, a sort of applied political science or applied moral philosophy. These historians did not

strive for new factual knowledge at first hand but rather a better insight into the moral problems of politics pursued by states and kings. Such literary professionals who worked with historical subjects were more successful than the contemporary professors of the universities of Europe, who also earned their living from history and who were mostly not well known.

Still, in the (a) sense of the word 'professional,' the authors of the 'broad' literature on history had precursors in the eighteenth century in the form of school textbooks and, in England, orientations for magistrates written by employees in the legal system (Steedman, 2009). Popular narratives of history became common in Europe and the United States from the early nineteenth century. Many narrators of history wanted to give to their countrymen a better understanding of the past of their nation. Publishing narratives on national history for the broad layers of the population sometimes also gave a living to their authors already in the first half of the nineteenth century (Burke, 2011). It is easy to see the continuation of such popular narrative into the twentieth century. The number of persons who have had the possibility to earn their living by telling history in different media and not only in printed books has increased enormously. The demand for history as entertainment, originating in the eighteenth century among the upper classes, has grown and spread in society. In the twentieth century, publishers have been interested in different types of books, such as specialized historical works, textbooks for schools and universities, historical journals, historical fiction, and pure historical entertainment. Adaptations of history in various forms for radio and television have also gradually increased their share in the total turnout, and art film has been used for telling stories with a historical setting. The comprehensive number of persons who make their living from these forms of writing and processing history has never been calculated.

In the twentieth century, though starting around 1850, there has also been a great growth in the number of professional historians in sense (a) of the word, because of the growth of institutions where historians got employment, including the publishing establishments mentioned above (Lingelbach, 2003, 2011). Most numerous are, of course, those employed in teaching institutions. Here must be included all teachers specializing in history, both in secondary schools and at university level. Historians working in archives and museums also belong to this category. With the great expansion of secondary and tertiary education, the sector has been steadily growing in the twentieth century, especially after World War II. For France in 1910, Boer has estimated the number of historians in these categories to 1045 (Boer, 1998). On the eve of the 1950s, the number of professional historians in France was approximated as less than 3000, but then this number does not include archivists and historians employed in museums. In 1967, the number of the same limited category of professional historians had grown to close to 8000 (Charle, 1995). German and Australian calculations go in the same direction (Macintyre, 2000; Raphael, 2000).

Establishing Professionalism among Historians as a System of Norms

The second sense in which the term professional is used, "professional as a qualified person who is recognized by society (not least by his or her fellows in academia) as competent to practice the trade," has a more complicated historical reference. Historians who were recognized as competent historians by their contemporaries have existed for a long time. But in regard to the recognition by fellow historians, especially those with university posts, this is something much newer. If recognition is taken to mean an inclusion into a network, group, or formal association, only the history of the period from 1750 provided any means for the realization of such professionalism. If, further, a sort of common program of the network or group is required, the field is narrowed even more.

In the eighteenth century, German universities were flourishing and universities in other parts of Europe and its colonies were not. Because of the low level of state power in Germany, not only free authors could express their views but so could civil servants, such as professors. This made German universities, especially those in Halle and Göttingen, serve as nurseries for new notions in the humanities. Linguistics and history prospered. In the nineteenth century, 130 Protestant historians (plus an unknown number of Catholics) became full professors of history at German universities. For the first time, an occupational group of historians developed and a discourse on history emerged (Blanke, 1991). Mostly, however, these were individual accomplishments and the level of coordination between the learned men was low. Some historians (e.g., Johann Martin Chladenius and Johann Christoph Gatterer) lectured on rules and wrote on the theory of history (Reill, 1975). The process of standardization of methodological rules was, however, regarded with skepticism by some of the most important historians, such as August Ludwig Schlözer, who talked about imprisonment in scholarly rules, as Blanke has noted (Blanke, 1991). As to rules for the critical use of sources and general methodological emphasis, the introductory book by Friedrich Rühs from 1811 is noteworthy.

There were hardly any 'schools' of historians before well into the nineteenth century. Academic teaching of history and research in history were well established in many countries in the eighteenth century, though without a canon of methodology or a professed normative system, which applied to history as a discipline. History was rather regarded as a part of a wider discipline of morality, the French *sciences morales et politiques*. Past events could be used to show the consequences of different types of conduct, and new Enlightenment notions made it possible to analyze historical processes without referring to God or providence (Heilbron, 1990). For centuries, history had been used for this purpose, but the morality concerned changed its content and its stability in the eighteenth century.

The change, brought about primarily by Leopold von Ranke around 1830, was to create the conditions for a network of historians under the guidance of the master. He had meetings with his disciples, later known as his seminar, where problems of history were discussed. Ranke himself was a spokesman for three main principles of historical interpretation: the state as a central entity, the state system as the polity of international relations, and the necessity for the historian to evaluate each single entity for its role in history and not from its relation to contemporary politics. The first of these principles was a common ground for many thinkers of the period, but combined with the other two it formed a specific analytic system (Torstendahl, 2009). The third principle has been partly

traced back to forerunners in the eighteenth century (Meinecke, 1936; Blanke and Rüsen, 1984; Blanke, 1991) and it created the basis of the thought system, which became known as *Historismus*. Ranke's influential teaching made his version of this *Historismus* the basis of a conscious professional identity, which spread in Europe and had made inroads in many other parts of the world, especially the United States and East Asia (Berg, 1968; Eskildsen, 2008, 2009; Torstendahl, 2010).

In the German academic community of historians, the combination of *Historismus* and teaching through seminars became an instrument for the professionalization of historians. A common evaluative system was created in the latter half of the nineteenth century. A methodological part was often combined with individualization of persons and collective entities of the past, especially among the historians of the Middle Ages who were concerned with the editing of the *Monumenta Germaniae historica* (initiated in 1819; first volume published in 1826). This work elaborated further methodological principles for editing that were developed in France by Jean Mabillon and others from the seventeenth century onward, and that gave rise to the Ecole des Chartes (founded in 1821) at Paris. Methodology became important and the principles for the evaluation of sources became more and more systematized. A methodological direction of historians within the *Historismus* has therefore been discerned, with a philosophical direction as its counterpart. The former was characterized by its devotion to empirical studies with a sincere application of methodological principles in order to obtain objective knowledge, while the philosophical direction generally rejected the possibility of reaching any form of certainty and proclaimed the relativity of knowledge. In fact, it will seem that the differences within *Historismus* were great. On the one hand were those who saw historical relativism as fundamental and regarded history primarily as a means of orientation in the present. On the other hand were those who wanted to overcome as much as possible the constraints of the sources for understanding the past, though still with a considerable relativism in regard to values (Oexle, 1996; Scholtz, 1997).

A new turn of historical professionalism grew out of the empirical *Historismus*. It was important that new methodological manuals were published and widely accepted in the European community of historians. J. G. Droysen had lectured repeatedly on methodology and the principles of *Historismus* but the published version, his *Grundriss der Historik* (1868), was very concise. Ernst Bernheim's *Lehrbuch der Historischen Methode* (1889) and Charles-Victor Langlois' and Charles Seignobos' *Introduction des Etudes Historiques* (1898) were, in this regard, epoch making. These books were primarily methodological, but in both their different epistemological foundations played a great role. The books received a very wide circulation. They taught how to find secure knowledge in history. Their different philosophical bases occasionally gave rise to debates but their differences did not prevent their rule system from becoming widely appreciated among historians (Torstendahl, 2003).

An important result of the methodological turn of professionalism was that historians turned their attention to more and more limited problems. Around 1900, it was no longer usual for professional historians to make the whole history of a country the subject of their study. Rather they focused on a certain political event and its prehistory, for which they now were expected to gather all relevant material. The proper methods to evaluate sources were often regarded as a dividing line between professionals and amateurs.

Several factors helped to make methodology a central element in the professionalism of historians. One was the rise of national historical associations. This was a process that took place all over Europe (and also beyond Europe in some countries) in the latter half of the nineteenth century. Philosophical and political issues divided all such associations and, thus, only methodology was really a common basis. This was even more the case with the international congresses of historians, from 1926 organized by the Comité International des Sciences Historiques. The congresses can be traced back to 1898 but got their form later. In the programs of the international congresses of 1900 and 1903, methodology played an important role and the object of the discussion was which were the conclusions for the philosophical foundations that must be drawn from the established methods, rather than the other way around. After the congress of 1908, the French historian Simiand complained in a review article that methods had not been sufficiently emphasized in the congress contributions (Erdmann, 2005).

In the last decades of the nineteenth century and the early part of the twentieth, the historical associations of the European countries also in most cases started journals (Stieg, 1986). The influential German *Historische Zeitschrift* was, however, founded in 1859 as a private initiative by Heinrich von Sybel. The earliest still existing seems to be a Danish journal from 1838. During the twentieth century, the creation of journals of history became usual in all parts of the world, and some of these are much more than national journals but quite general (such as the *American Historical Review*, founded in 1895) or thematic for studies of the history of a major part of the world (such as the *Journal of Asian Studies*, founded in 1957). Other journals and reviews with a thematic or rather broad methodological focus, such as the *Revue de Synthese Historigue* (1900–31), the *Annales d'Histoire Economique et Sociale* (founded in 1929, and from 1946 called *Annales: Economies, Sociétés, Civilisations*), *Past and Present* (founded in 1952), the *Journal of Modern History* (founded in 1929), and so on, were later creations. In most of the journals, articles on themes of history were combined with critical reviews of books. An important element in such reviews was to examine the methods used and the shortcomings of authors in respect to methodological rules. Thus, in the form of critical reviews of the works on history, the methodological norms were spread. These norms were also commonly accepted. In the case that a historian wanted to contradict criticism, he or she had to show that the principles had been observed rather than that these principles were not valid. This is something that characterizes methodology but is not valid for the kind of substantive norms that formed the major part of the professionalism that Ranke had established, based on his kind of *Historismus*.

For the formation of the profession of historians, the debates on the application of historical methods were very important. A series of such debates took place in different countries in the first decades of the twentieth century. There was a big debate in Germany, called the *Methodenstreit*, on the correct use of methods in regard to any scientific activity. Generalization and interpretation were advanced as

alternatives. Karl Lamprecht, with his concept of social and cultural history, was rather lonely in Germany and is often regarded as defeated. In fact, part of the issue was the theories of *Historismus* and the alternative professed by Lamprecht. Another part dealt with methods and their application in the strict sense, where individualization (W. Windelband, H. Rickert, M. Weber, F. Meinecke, and others) was opposed to views based on structures and collectivities (W. Sombart, H. Pirenne, F. Simiand, and others) (Iggers, 1984). The struggle was reflected in other countries such as Finland, where Lamprecht's influence was strong. In other countries (e.g., the Scandinavian countries and China), there were other straightforward struggles about the use of methodological principles and their consequences for national history. The historians who advanced new and radical consequences of methodologies, which were already well known, had a program. They wanted to establish firm knowledge in history, a knowledge that was contrasted with guesses and conjectures, not historical relativism. In the United States, the focus of the debate was rather on the possibility of establishing objectivity (Novick, 1988).

Historical professionalism also increased through the ongoing specialization of all historical fields. In the beginning of the twentieth century, it was no longer the normal thing for a historian to take as the theme of an investigation the history of a country during a certain period. This had been quite usual in the middle of the nineteenth century, but in the first decades of the twentieth century this had changed. History had begun its separation into specific subfields, and economic, social, and cultural history was regarded as major examples of this. Specialized research institutes came into existence, and specialized journals were published: the *Economic History Review* from 1927, the *Vierteljahrschrift für Sozial- und Wirtschaftsgeschichte* from 1903, the *Journal of Contemporary History* from 1966, and the *Vierteljahrshefte für Zeitgeschichte* from 1953. Similar journals were published in many countries with specific reference to the specialty in the national context. Specialization went further, chronologically and thematically. Investigations became very specialized in regard to field and object of study, and period of investigation. There was a dramatic change in the themes regarded fit for doctoral dissertations from the middle of the nineteenth century to 1920. Dissertations then were also outcomes of extensive archival research, where the author could and should show his or her (some women were achieving the doctoral degree at the time) acquaintance with sources and with the appropriate methods. Academic professionalism aiming at firm knowledge had found its way of expressing itself in such specialized studies.

At the same time as the second phase of historical professionalism in the sense of recognition among equals was triumphing in the interwar years, a distrust in the validity of its central notion, the idea of secure knowledge, was winning ground among historians. The philosophical direction of the *Historismus* had long since brought about doubts as to the possibility of obtaining security in historical research. Its proponents were influenced by the old tradition of hermeneutics, grown out from Bible studies, for acquiring a system of understanding. It contained two main requirements: preknowledge as the conceptual apparatus and direction of interest, and empathy as the means of acquiring an understanding of human thoughts and actions. Hermeneutics was regarded as a means to acquire more or less probable statements on the past, and the most general formulations included an ambition to maximize the probability through research. Collingwood's *The Idea of History* (1946) and Hans-Georg Gadamer's *Truth and Method* (1960) were influential in Britain and Germany respectively, and their works were complemented by the Frenchman Paul Ricoeur's *Time and Narrative* (1983–85).

Common to these works was that many historians accepted them as important foundations for the work of the historian. This meant that a new professional identity arose among some historians, conflicting with the earlier established group identity based on firm knowledge. As long as a basic norm in the hermeneutical camp was to increase the probability of historical statements through empirical research, the conflict could be kept within close bounds. After the 1980s, this was no longer possible. Various theoretical directions have taken the relativistic claim further. The debate in the 1980s and 1990s has gone out from Hayden White and his rhetorical analysis of historiography (White, 1978). The analyses of his standpoint have been numerous, above all in the United States, and have led to a variety of relativistic positions as regards the theory of history based on the assumption that history is basically a rhetorical and esthetic discipline. At the same time, in Europe, postmodernism has influenced other philosophers of history. A far-going relativism in different forms thus has more or less superseded the earlier basis of professional identity among historians.

At the same time as the idea of firm knowledge became questioned, the emphasis of professionalism was moved from the basis of methodology to the field of what was historically important. Social theories in the broad sense of theories of the social sciences had for a long time occasionally been used by historians. Braudel's book on the Mediterranean world from 1949 was breaking new ground in this respect. More regularly, historians have made models of important or interesting history the foundation of professionalism (or professional schools) from the 1960s onward. In economic history, it was already earlier the rule to rely on and use theories of economics, but theories of management and organizations and sociological theories were used to a large extent by historians in historical studies only in the 1970s and 1980s and more specifically Marxism in the 1970s. The professionalism of economic historians has come to rely on the ability to apply such theoretical instruments as well as specific economic (statistical) methodology. The analysis of social structures and social dynamics has also developed through the use of social theory. This approach had one of its origins in the Annales circle, then most often directed toward demography and geography and their social consequences. More explicitly, social science theory in general, including sociological and organizational theory, was proposed as the proper instrument for historical research by the so-called Bielefeld school, consisting of a group of historians who had some relation to the University of Bielefeld in the 1970s.

What they proposed was in line with what had been done by several American historians earlier, such as Merle Curti and Stephan Thernstrom, but their efforts had not been coordinated into a program. The Bielefeld historians advanced the concept of a 'historical social science' (Kocka, 1977; Wehler, 1980). This type of history found much appeal in the 1970s

and 1980s. Although parts of this enthusiasm for social sciences were criticized later, it is rather the theories used than the close link to social science (including social anthropology, which found growing appeal among historians) that have met opposition. The use of social science theories has thus been one source of professional identity among historians in recent decades.

All through the twentieth century, historical professionalism has also been challenged from political quarters. Autocratic regimes all over the world have tried to turn historians into docile instruments for the propaganda of the ruling ideology. Many have suffered for their professional ideals; others have accepted the norms of the state instead of the norms of the profession (Baets, 2009).

By the early twenty-first century, when historians have tried to base their professionalism on interpretative theories, methodological rules, and social science theories, all without general acceptance, the history of historiography seems to have met a new interest. Different standpoints to historical knowledge may unite in the interest in investigating the history of historiography from different angles, and it is equally important for all different standpoints. The role of this disciplinary self-understanding for a renewal of professionalization should not be overrated, however. The history of the discipline cannot meet the requirement of providing the norms through which a closure can be upheld against nonprofessionals, and as regards norms there is no unanimity.

Historical professionalism has developed in stages from Rankean *Historismus*, to methodological claims to find a firm historical knowledge, then further to fields of importance. No single normative basis is currently prevailing, and the frail identity of the profession seems to rely on the quasi-theoretical field of its own history.

See also: Historicism; History: Overview; History: Representation Within and Functions of; History: Theories and Methods; Professions and Professionalization, History of; Professions, Sociology of.

Bibliography

Baets de, A., 2009. Responsible History. Berghahn Books, New York.
Berg, Gunter, 1968. Leopold von Ranke als akademischer Lehrer. Studien zu seinen Vorlesungen und seinem Geschichtsdenken. Vandenhoeck & Ruprecht, Göttingen, Germany.
Blanke, H.W., 1991. Historiographiegeschichte als Historik. Fromann-Holzboog, Stuttgart- Bad Cannstadt, Germany.
Blanke, H.W., Rüsen, J. (Eds.), 1984. Von der Aufklärung zum Historismus. Zum Strukturwandel Historischen Denkens. Schöningh, Paderborn, Germany.
Boer, P. den, 1998. History as a Profession. The Study of History in France, 1818–1914. Princeton University Press, Princeton, NJ.
Burke, P., 2011. Lay history: official and unofficial representations, 1800–1914. The Oxford History of Historical Writing 4, 115–132.
Charle, C., 1995. Être historien en France: une nouvelle profession? In: Bédarida, F. (Ed.), L'Histoire et le Métier d'Historien en France 1945–1995. La Maison des Sciences de 1'Homme, Paris, pp. 21–44.
Erdmann, K.D., 2005. Toward a Global Community of Historians. Begrhahn, New York & Oxford.
Eskildsen, K.R., 2008. Leopold Ranke's archival turn: location and evidence in modern historiography. Modern Intellectual History 5, 425–453.
Eskildsen, K.R., 2009. Leopold von Ranke, la passion de la critique et le séminaire d'histoire. In: Jacob, C. (Ed.), Lieux de savoir. Espaces et communautés, Paris, pp. 462–482.
Heilbron, J., 1990. Ontstaan van de Sociologie [The Rise of Social Theory]. Prometheus, Amsterdam.
Iggers, G., 1984. The 'Methodenstreit' in international perspective. The reorientation of historical studies at the turn from the nineteenth to the twentieth century. Storia della Storiografia [History of Historiography] 6, 21–32.
Kocka, J., 1977. Sozialgeschichte: Begriff, Entwicklung, Probleme. Vandenhoeck und Ruprecht, Göttingen, Germany.
Lingelbach, G., 2003. Klio macht Karriere. Die Institutionalisierung der Geschichtswissenschaft in Frankreich und den USA in der zweiten Hälfte des 19. Jahrhunderts. Vandenhoeck & Ruprecht, Göttingen, Germany.
Lingelbach, G., 2011. Institutionalization and professionalization. The Oxford History of Historical Writing 4, 78–96.
Macintyre, S., 2000. History in a new country. Australians debate their past. In: Torstendahl, R. (Ed.), An Assessment of Twentieth Century Historiography. Royal Academy of Letters, History and Antiquities. Stockholm, Sweden, pp. 71–86.
Meinecke, F., 1936. Die Entstehung des Historismus. R. Oldenburg, Munich, Germany.
Nienhauser, W.H., 2011. Sima Qian and the Shiji. The Oxford History of Historical Writing 1, 463–484.
Novick, P., 1988. That Noble Dream: The 'Objectivity Question' and the American Historical Profession. Cambridge University Press, Cambridge, UK.
Oexle, O.G., 1996. Geschichtswissenschaft im Zeichen des Historismus. Vandenhoeck und Ruprecht, Göttingen, Germany.
Raphael, L., 2000. Flexible response? Strategies of academic historians towards larger markets for national historiographies and growing scientific standards. In: Torstendahl, R. (Ed.), An Assessment of Twentieth Century Historiography. Royal Academy of Letters, History and Antiquities. Stockholm, Sweden, pp. 127–147.
Reill, P.H., 1975. The German Enlightenment and the Rise of Historicism. University of California Press, Berkeley, LA.
Scholtz, G. (Ed.), 1997. Historismus am Ende des 20. Jahrhunderts. Eine Internationale Diskussion. Akademie-Verlag, Berlin.
Smalley, B., 1974. Historians in the Middle Ages. Thames and Hudson, London.
Steedman, C., 2009. Labours Lost. Domestic Service and the Making of Modern England. Cambridge University Press, Cambridge, UK.
Stieg, M.F., 1986. The Origin and Development of Scholarly Historical Periodicals. University of Alabama Press, Tuscaloosa, AL.
Torstendahl, R., 2003. Fact truth and text. The quest for a firm basis for historical knowledge around 1900. History and Theory 42, 305–331.
Torstendahl, R., 2009. Historical professionalism. A changing product of communities within the discipline. Storia della Storiografia 56, 3–26.
Torstendahl, R., 2010. From all-round to professional education. How young historians became members of an academic community in the nineteenth century. Leidschrift 25, 17–31.
Wang, Q.E., 1995. Time perception in ancient Chinese historiography. Storia della Storiografia 28, 69–86.
Wehler, H.-U., 1980. Historische Sozialwissenschaft und Geschichtsschreibung: Studien zu Aufgaben und Traditionen deutscher Geschichtswissenschaft. Vandenhoeck and Ruprecht, Göttingen, Germany.
White, H., 1978. Tropics of Discourse. Essays in Cultural Criticism. Johns Hopkins University Press, Baltimore, MD.

History: Representation Within and Functions of

Chris Lorenz, VU University Amsterdam, Amsterdam, The Netherlands; and Ruhr-University Bochum, Bochum, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

History as a discipline is a continuation of extradisciplinary forms of dealing with the past and has therefore never enjoyed a monopoly on its representation. Moreover, within the discipline there has always been a plurality of representations competing with one another. This article traces the modern problem of the narrative representation of history to its origins in the Aristotelian and the rhetorical traditions. The same goes for the modern question whether historical narratives can be true.

The modern doubts as to the practical functions of history are also traced to its origins. Although traditionally history's function was conceived of as pragmatical (*historia magistra vitae*), this idea presupposed that the past was basically similar to the present and the future. Since the rise of modernity in the eighteenth century this presupposition increasingly lost its plausibility. Because change became permanent, there was a rupture between the present and the past. Although this rupture was the very condition for the possibility of the 'disciplinization' of history, it rendered traditional pragmatical conceptions of history obsolete. Nietzsche's distinction between the monumental, the antiquarian, and the critical functions of history – taken over by, among others, Michel Foucault – is presented as a solution to the practical problem that still has validity.

Introduction

History as an institutionalized, academic discipline is a continuation of extradisciplinary forms of dealing with the past and, therefore, never enjoyed a monopoly on its representation. Narrative representations of origins – be it in the form of myths, legends, religions, or ideologies – are found in most cultures. These different forms of representing the past offer collectives an *orientation in time*: they give a *meaning* to the present by locating the present on a time axis between a future and a past.

Next to the variety of disciplinary forms of historical representation (Stern, 1970; Kelley, 1991; Lorenz, 1997), these extradisciplinary representations of history still exist, often competing with another and with disciplinary history. The problem how to demarcate the extradisciplinary from the disciplinary ways of dealing with the past – including their functions – has therefore haunted history as a discipline from its very beginning. Traditionally, the discipline's identity has been based on its *truth claim, based on its critical methods* and its *archival foundation* . This article will only deal with the disciplinary representations of the past in the West. It will *only* take extradisciplinary representations into account in order to cast light on the origins and the official borderlines of the discipline (Fain, 1970; McNeill, 1986).

Historical Representation and Its Function as a Historical Problem

Although the concept of representation has a clear-cut meaning in philosophy – that of a picture, in which the *structure* of the pictured phenomenon is *maintained* (Mormann, 1997) – in history the term is usually used as an equivalent of concepts like *image* or *construction*. In titles like 'The representation of Hitler in the US press, 1933–1945' or 'The representation of the Black Woman in the literature of the Antebellum South,' historians emphasize the constructed and interpretative character of historical reconstructions: history cannot directly be presented, but is *re*presented by the historian in language. This emphasis is rooted in an awareness that historical representations are *no mirror images* of history itself and is usually accompanied by an awareness of the fundamental *plurality* of historical reconstructions. At the same time, it is sometimes accompanied by an awareness of the *practical* functions (or the *politics*) of representation.

For the consciousness of this 'representational' character of history, two developments were crucial. The first development was the conception of history as an *object of knowledge*, independent of its individual representations. According to Koselleck, this development originated in Europe, in the mid-eighteenth century, when history as a process – *historia* – was first distinguished from the traditional *historiae*, i.e., the *stories* told about the past.

History could only be conceived of as an independent object of knowledge when it was no longer regarded as a binding *tradition* or as a practical *craft* belonging to literature or rhetoric, i.e., as the art of telling (his)stories in order to persuade a public, as had been the case in Europe from classical antiquity until the mid-eighteenth century. And history's functions only developed into a problem after the traditional idea, that past experiences were relevant for present purposes and that histories could teach by example what to do and what not to do (*historia magistra vitae*), began to loose its plausibility. 'History' thus is the direct product of the idea of 'progress' and of 'modernity.'

The second development, that promoted the awareness of the representational character of history, was the 'discovery' of its *perspectival character* during the Enlightenment. Although the possibility of multiple stories about the same events was known already – and a central concern of rhetorics – Enlightenment historians were the first to recognize explicitly that the historical representations *necessarily* vary with the *perspectives* of the historians and with the *development of time*. The Enlightenment also stimulated the tendency to identify cognitive, scholarly activity with *empirical* disciplines, which was a consequence of the competition of the *Geisteswissenschaften* with the modern natural sciences (Scholtz, 1991; Momigliano, 1990).

Koselleck dates the fundamental change in the relationship between Europeans and their sense of the past halfway through the eighteenth century, according to Pocock these changes had already started in the seventeenth century (Koselleck, 1979; Pocock, 1962). The traditional idea of a continuity between the past, the present, and the future was gradually replaced by the experience of *permanent and 'progressive' change*, so typical of modernity (Scholtz, 1991; Hardtwig, 1990). Because everything was changing all of the time, the past turned into a 'foreign country' (Plumb, 1973; Lowenthal, 1985; Rüsen, 1983–89).

Early in the nineteenth century, this experience of *difference* between the past and the present would lead to the association of history with the notions of *non*recurrence and *non*-generalizability, later expressed in the concepts of *uniqueness* and *individuality*. *History*, thus, came to be conceived of not just as random change through time – which could be cyclical and thus repeatable – but as *change as a process of directional and irreversible development*.

With the emergence of this specific new sense of the past, the stage was set for the development of history and the social sciences into separate disciplines. The object of history was increasingly identified with *non*recurring phenomena in the human past, while the social sciences were identified with recurrent phenomena in the human present. Because recurrence is a precondition for generalization and thus for (structural and processual) theories, history and theory gradually turned into opposites, excluding one another. Thus the idea was born that the forms of representation in history differed fundamentally from those in the social sciences. While historical representation was usually characterized as *time oriented* and as *narrative*, social scientific representational forms were characterized as *analytical* and *theoretical* (Meran, 1985; Burke, 1992; Lorenz, 1997). The modern question of narrative in history was thus born in the eighteenth century as a concomitant of the rise of modernity.

The Classical Roots of the Modern Debate (1): The Aristotelian Heritage

Both the Enlightenment conception of history and its critique in Herder and in later Romantic thinkers – the two conceptions of history that still define the terms of the modern debate – must be viewed in the context of Aristotelian philosophy, because the Aristotelian definitions and distinctions of disciplines were still very influential in the eighteenth century (Aristotle, 1963).

In Aristotle's (384–322 BC) philosophy, there is a basic distinction between empirical knowledge of singular facts and universal knowledge of general facts. Concrete, empirical knowledge just formulates the *contingent* and *particular* characteristics of its object, while theoretical knowledge formulates its *necessary* and *universal* characteristics. These latter characteristics are viewed as *recurrent* and *unchanging* and were therefore limited to the domain of *nature*. Theoretical knowledge was therefore limited to the domain of nature too. Theoretical knowledge, according to Aristotle, does not only inform us *what* reality looks like, but also *why* this is the case: theoretical knowledge is always *explanatory* knowledge, because it informs us about the *causes* of what happens. Therefore explanations in science are causal explanations and are ideally based on theoretical knowledge. This idea would dominate later philosophy of science in the form of positivism from Auguste Comte to Carl Hempel (Meran, 1985).

Within Aristotelian philosophy the domain of human activity or *culture* (praxis) held a fundamentally different status from nature, because contingency, particularity, and change characterize the first domain. This domain, therefore, was *not* open to theoretical knowledge and to scientific explanation *in principle* ('*de singularium non est scientiae*,' as Duns Scotus wrote in the thirteenth century). History, according to the Aristotelian system, belonging to the domain of culture, could never be regarded as a science and was located in the domain of *practical* knowledge, more akin to poetry (including literature).

Although both poets and historians dealt with the same phenomena, i.e., *human actions* and *events* (albeit respectively in verse and in prose), history only dealt with what was *real* and poetry with what was *imagined*. The historian, according to Aristotle, "describes what happened" while the poet "describes what can happen" (Aristotle, 1963: p. 419). This distinction was later codified in the distinction between the *res factae* (Latin: res = affairs; factae = done) or 'things done,' which was the domain of history, and *res fictae* (Latin: fictae = imagined), which only concerned poets.

The duty of the historian to stick to the facts and to tell the *truth*, the *whole* truth and *nothing but* the truth, derived mainly from the model of Thucydides. It received its canonical formulations in Cicero (106–43 BC) and Lucian (*c*.AD 125–200), who also linked truth in history directly to *impartiality* by avoiding moral blame and envy (*sine ira et studio*). The German historian Leopold von Ranke (1795–1886) would only repeat these classical ideals in the nineteenth century, with great force and effect (Iggers, 1997).

According to Aristotle, the poet enjoyed a 'constructive liberty' not shared by the historian. While the unit of the historian is defined by what had happened to specific subjects at a specific *time* at a specific *place*, the unit of the poet was *not* defined in such a way. In poetry, actions and events are put together by a unit of *meaning*, that is by constructing a whole with a beginning, middle, and an end, held together by a *plot*. The parts of the poetic unit all fulfill a function in – i.e., derive their meaning from – this whole. Because of its constructive quality, Aristotle held that poetry was more philosophical than history. This characterization was also based on the idea that poetry – for instance, tragedy – was concerned with *general* problems – types – and not only with particulars, like history.

The Aristotelian definition of history would define the parameters of the debates on historical representation up to the present. The *opposition of nature and culture* and the identification of nature with necessity and repetition, and culture with contingency and change, have defined the methodological battlefields in the philosophy of history. This opposition linked nature to (universal) theory while excluding (particular) history from theory a priori. As influential has been the *opposition of poetry and history* and the identification of history with the domain of the factual in opposition to the domain of fictionality.

Last but not least, the Aristotelian *opposition of historical units and poetical units* has put the *relationship between chronological*

and narrative ordering of events on the agenda of the philosophy of history i.e., the relationship between chronology and narrative. This opposition raised the question what narrative explanations are and how they relate to causal explanations.

The Classical Roots of the Modern Debate (2): The Rhetorical Tradition

Another fundamental root of the modern debate on historical representation can be traced back to classical *rhetoric* (Koselleck et al., 1982). Accounts of the past were dealt with by classical authors like Aristotle, Cicero, Lucianus, and Quintilianus (*c*.AD 35–96) primarily in the context of rhetoric, especially *forensic rhetoric*, i.e., rhetoric used to persuade a judicial court. So, the problem how to compose and write an account of the past was primarily dealt with as a problem in a *communicative* context and not as a *cognitive* problem, as was the case since the nineteenth century. Therefore the question how to produce a *convincing* history loomed larger than how to write a *truthful* history – which in turn explains why rhetoric later acquired a doubtful reputation until it was officially expelled from the scholarly domain by Descartes, only to return in the late twentieth century in the train of postempiricism and post-positivism (Nelson et al., 1987). This origin in forensic rhetoric also explains why the domain of history was identified with the domain of *human action*.

Now to be convincing, according to rhetoric, truth may help, because Plato argued that true arguments are the most convincing. Quintilianus, however, later pointed out that truth as such was neither a necessary nor a sufficient condition for being convincing. Being *convincing* is directly related to being perceived as *probable* rather than as being factually *true*, because some truths are hard to believe. In order to be convincing, according to rhetoric, past action primarily must be explained in (motivational) terms *understandable* and *plausible* for the *audience* addressed. Therefore, the reconstruction of past actions was basically conceived of in the rhetorical tradition as an act of *translation* of factual information by a speaker, who has a specific *aim* and who addresses a specific *audience*.

The rhetorical emphasis on probability instead of on truth was a consequence of its focus on *communication* and the awareness that factual truth is not a communicative category. It did *not* imply any plea for blurring the distinction between facts and fiction in history, nor the idea that truth in history fulfilled no practical function. To the contrary, the truth of the past was sought after, because only *true* accounts of past actions were *useful* as *exempla* for later generations (Momigliano, 1990; Hay, 1977). History was thus usually regarded as *inherently pragmatic* for some 20 centuries until the presuppositions of this pragmatic conception lost their plausibility with the rise of 'modernity' in the second half of the eighteenth century.

The Modern Debate on Historical Representation (1): The Problem of Narrative

Basically, the theories of historical representation developed since the eighteenth century can be classified on basis of their relationship with the Aristotelian and the rhetorical characterizations of history. There is a genus of theories, that can be seen as a continuation of Aristotelian theory, and there is a genus of theories, that opposes Aristotelian theory. Both genera contain several species.

The first genus consists of theories that have transformed the Aristotelian analysis of *poetical or literary narrative* into theories of *historical narrative*. According to these theories, historical narratives are structured by plots (P. Ricoeur; P. Veyne; H. White; D. Carr), colligatory concepts (W. Walsh), configurational ideas (J-.G. Droysen; L. Mink), or metaphors (F. Ankersmit). This view implies that historical narratives are *not* structured by a chronological order; instead they are structured by linguistic instruments that synthesize chronologically disparate elements into a *synchronic unity* (as, e.g., is done by concepts like 'The Enlightenment,' 'feudalism,' or 'imperialism'). Narratives, thus, are fundamentally different from chronologically ordered historical representations, such as chronicles and annals.

The *fundamental* difference between literary and historical narrative, according to most of these narrative theories, is the difference between fiction and fact: while the construction of literary narratives is only restricted by what is *plausible*, historical narratives are restricted by what is factually *true*. Therefore, history could and did claim a scholarly status in being empirical in the broad sense of a *Wissenschaft*.

This position was first developed by the Historical School from Niebuhr and Ranke onward and is philosophically legitimized by historicism that posits a fundamental connection between the narrative form of representation and the notions of time and of individuality. The basic idea is that narrative is the linguistic medium, which shows how a phenomenon (a person, nation, idea, etc.) develops its defining characteristics – its *identity* or *individuality* – in time, just as a biography shows how an individual has developed its identity in time (Lübbe, 1977; Veyne, 1978; Ankersmit, 2012). And although most narratives appear to follow the natural flow of time, it is, in fact, the other way around: the central idea of a narrative (e.g., the Enlightenment, the Age of Revolutions, etc.) determines which temporal units and orders (periodization, etc.) are the relevant ones for the narrative in case. Narrative time is thus always *constructed* time. Droysen formulated this insight some 140 years ago (Droysen, 1857).

Within this first genus of narrative theories, there are several species that show important differences among each other. The first species, exemplified by Louis Mink, Herman Lübbe, and Frank Ankersmit, has developed the argument that explanation by narrative is *sui generis* and is logically independent of causal explanation. All members of this species hold that the *narrative mode of representation* is *explanatory in itself*.

The second species, exemplified by Arthur Danto and Jörn Rüsen, holds that explanation by narrative, although different from lawlike explanation, is based on causal explanation nevertheless. The narrative mode of representation is analyzed here as a set of *causal chains*.

A third species of narrative theories, exemplified by Paul Ricoeur, David Carr, and Paul Veyne, has developed the argument that historical explanation has a narrative form and a plot structure. Ricoeur and Carr, moreover, argue that human experience itself is narratively structured and that there is an *a priori fit between human experience and narrative explanation*.

Both human individuals and communities can be said to have the same narrative form of self-constitution, because their identities can only be represented in the form of narratives.

Whatever their differences, all of these three species of narrative theories regard narratives primarily as explanatory schemes. All three try to answer the Aristotelian question how narratives explain, if not straightforward causally.

A fourth species of narrative theories is not primarily focussed on issues of *explanatory logic*, but on the *rhetorical problem* why narratives are regarded as *plausible* and *convincing*. Ann Rigney, and Hayden White in his *Metahistory*, exemplify this type of narrative theory. First, White argues that all historical narratives fulfill an ideological function, aimed at specific audiences. Second, he argues that historical narratives contain one of four literary plot structures: comedy, tragedy, satire, or romance. Third, he argues that all history can be emplotted freely in these four manners. Narratives are accepted by an audience as probable and as plausible, according to White, when the narrative plot structure and its ideological implication are *recognized* and *accepted as such*.

The acceptability of historical narratives, according to White, is thus *relative to audiences* and *relative to ideological aims* and *not* only dependent on the 'objective' truth of the factual statements in the narrative. In this respect, White's theory has a rhetorical ring, although his fixation on plot structures is at the same time Aristotelian. However, on the issue of narrative truth, White deviates fundamentally from both Aristotelian and rhetorical presuppositions. Contrary to Aristotle, he blurs the borderline between fact and fiction in history by stating that historical narratives are 'fictions of factual representation.' Consequently, according to White, there are no epistemological grounds for preferring one historical representation over another, only political and aesthetic ones. So, along with rhetoric, he holds that plausibility bears no necessary relation to truth. However, contrary to both Aristotelian and rhetorical theory, White severs the crucial link between research and composition of historical narrative: the composition of narratives is not treated as following from – and restricted by – research, but as a fully independent process (Paul, 2011).

The second genus of theories of historical representation somehow opposes Aristotelian theory. Historically this genus derives from eighteenth-century Enlightenment *conceptions of history*. Contrary to Aristotle, Enlightenment historians like Voltaire, Condorcet, and Chladenius, held that history could be both *theoretical* – in the sense of containing knowledge of a general character – and *scientific* – in the sense of being empirical. 'Theoretical' or 'hypothetical' history was thus claimed as a legitimate 'scientific' project, in which 'theory' and empirical knowledge could be integrated. Theoretical knowledge was no longer exclusively identified with the domain of nature and no longer exclusively interpreted as knowledge of what is necessary and universal. All knowledge of a general character was regarded as theoretical, including empirical generalizations and theories of the middle range. In the twentieth century, this type of position has been defended by the proponents of the *conception of history as a social science*, who conceived of history as a *mix* of particular and general characteristics. As the general characteristics could be theorized, they emphasize the importance of *theories* and *models* in historical reconstructions (Wehler, 1973; Kocka, 1977; Tilly, 1981; Burke, 1992; Iggers, 1997). This type of representations, being explicit about the concepts and arguments used, is often called *analysis* instead of narrative, emphasizing its affinity to the social sciences. This happened especially when 'traditional' history and its function were seriously questioned between 1960 and 1980.

The Nineteenth-Century Origins of the Modern Debate on the Functions of History

From classical antiquity until the eighteenth century the pragmatic presupposition that past experience was meaningful for present purposes was not questioned seriously, nor that the (empirical) truth coincides with the (ethical) good. From the end of the eighteenth century onward, however, historians and philosophers began to doubt the salutary nature of history and therefore the function of history: when history is no longer experienced as a meaningful and salutary process, what good is the study of history then? Nietzsche's all too nasty question in *The use and disadvantage of history for life* (1874), whether the study of history had *any use at all*, was the logical endpoint of this vanishing faith in history.

European historians basically countered these doubts by two strategies of demarcation. Firstly, by attempts to enhance the scholarly status and 'objective' credentials of history by increasingly expelling the rhetorical and artistic dimensions from its disciplinary domain. An increasing tendency to conceive of history as a *Wissenschaft*, opposed to *Kunst* or art, by identifying the discipline with its empirical research methods (*Forschung*) at the cost of the literary and artistic methods of its presentation (*Darstellung*), was the result (see Hardtwig, 1990; Scholtz, 1991).

Secondly, these doubts were countered by demarcating disciplinary history sharply from any *substantial philosophy of history*, especially from its Enlightenment and Hegelian variants (including Marxism), and from any *practical function*. In the famous words of Ranke, historians were not to pass any judgment on the meaning of the past – as the Enlightenment had done – but only 'to show what actually had happened.' By this demarcation strategy, the *reflection on the practical functions of history* was excluded from the disciplinary discourse and was expelled to the *philosophy* of history.

This epistemological demarcation had become possible, while history was professionalizing and attaching itself to the (usually: national) *states* both institutionally and financially (Berger and Lorenz, 2010). Because both the state elite's and the majority of the professional historians presupposed that education in (national) history was essential for 'nation building,' the practical functions of history ceased to be a problem. The practical function of history was now conceived of as *indirect* – through individual and collective identity formation alias *Bildung* – rather than as direct through *exempla*. This view was also held by proponents of a 'contemplative' function of history, like Ranke.

The ideal of objectivity was silently identified in historical discourse with promoting the *practical* cause of the 'impartial' nation state. This identification was not seen as a problem, because in the dominant conceptions of history, the historical process *itself* was identified with the genesis of nations and

states. It would take two World Wars, before this identification turned into a serious problem due to an increasing lack of plausibility. Nietzsche's nasty question, until then comfortably put aside by many professional historians as 'philosophical,' had to be faced at last: what good is the study of history anyway?

Nietzsche's analysis and classification of the functions of history in *The uses and abuses of history*, although framed as a critique of the nineteenth-century discipline of history in Germany, has proved to be very fruitful up to the present day (Rüsen, 1983–89; Kocka, 1989; Hardtwig, 1990; Tollebeek, 1990) Basically, Nietzsche distinguishes three ways to deal with the past and three types of history that correspond with three different attitudes and basic needs of human beings. Therefore, history has three basic functions that serve as mutual correctives and cannot be reduced to each other.

The first way to deal with the past is pragmatic, as an active *acting* being. Actors may address the past in search of examples, which, according to Nietzsche, deserve imitation. The past addressed in this way is treated as 'a collection of monuments.' That's why Nietzsche calls this pragmatic type of history *monumentalistic history*.

The second way to deal with the past is full of piety, as a *contemplative* being, when one is looking for one's roots in the past in search of one's identity. The past addressed in this way is treated as ' a collection of antiquities' or as 'a museum'; therefore, this 'pious' type of history, i.e., history just for the sake of the past, is labeled *antiquarian history* by Nietzsche.

Because man is not only directed toward the past, but also toward the future, it is impossible always to treat the past as a 'museum.' In order to create something *new* it is necessary to break with the past to some extent and to replace it. Therefore, it is also necessary to address the past in a critical way, as something to be criticized as fit for replacement. This type of handling the past, which Nietzsche compares to 'cutting one's roots loose with a knife,' is called *critical history*.

All three types of history have their own type of dangers and deficiencies that can only be remedied by history of the other types. Monumental history, for instance, tends to equalize the past and the present and to be very selective, obliterating differences and rendering huge parts of the past, unfit for the monumental function. This tendency can only be counteracted by the antiquarian impulse, which in turn suffers from 'historical myopia' and from a 'blind drive to collect.' This tendency, in turn, can only be counteracted by the critical impulse, which in turn must be held in check by the other two in order to prevent critical history from degenerating into a war against one's past.

Historiography can be seen as the combined product of these three impulses, in varying mixes. Disciplinary history since the nineteenth century, however, has usually – although not exclusively – embraced the antiquarian self-image and legitimization, because the recognition of both the monumentalistic and the critical approaches to the past would raise the problem of its practical functions and its claim to 'objectivity.' In accordance with this contemplative self-image of most historians, the antiquarian presentation *of identity* – i.e., the narrative reconstruction of the genesis of individuality in time – is usually presented as the discipline's only practical function. As in Nietzsche's time, the monumental and the critical functions of history are usually not discussed within the disciplinary boundaries (Angehrn, 1985). At times, however, for example in the 1970s, the situation has been different, because of an influential number of 'critical' historians in the profession (Kocka, 1989).

Symptomatic of this state of affairs is that recently the monumental function of history has been delegated to a separate domain, i.e., the domain of *collective memory* and of *lieux de mémoire*. The same holds for the critical function of history that is central to projects like gender history, postcolonial studies, and Foucault's critical history (which replaces the Enlightenment self-image of the West by an image of a disciplined 'carceral society'). These forms of critical history were often *not* regarded as belonging to the domain of history proper or are at best tolerated in its margins (Foucault, 1980; Scott, 1988).

Mainly in one historiographical domain – that of the *Holocaust* – reflection on the functions of history has included both its critical and monumental functions, because the very possibility of the antiquarian approach is fundamentally questioned here. Not accidentally, increasing skepticism concerning future 'progress' – and thus of 'modernity' – is causing increasing doubts as to the object-character of 'the past.' Historical representation *as such* has turned into a problem, because it is doubted whether the traditional categories and procedures of historical representation in order to 'objectify' the past, can function in this unimaginable domain where the very temporal distinction between past and present is problematic. Some hold that psychoanalytic categories like trauma, repression, transfer, splitting off, and working through are more apt to characterize the relationship between the historical representations and represented historical reality. Others have suggested that the 'overwhelming' dimensions of the Holocaust can better be characterized through the notion of the *sublime* (Friedländer, 1992; LaCapra, 1994; Frei and Kansteiner, 2013).

Mainly in this historiographical domain, the discussion about the functions of history is directly connected to the discussion on historical representation and its limits. From there, the discussion is being continued concerning some of the discipline's fundamental presuppositions, such as the presupposition that the past is 'at a distance' from the present and the future – and thus no longer 'present.' Along this route the debate is nowadays reflecting on the assumptions that gave birth to history as a discipline more than 200 years ago (Lorenz and Bevernage, 2013).

See also: Historical Thought and Historiography: Christian Tradition; Historical Thought and Historiography: Current Trends; Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; Historicism; History and the Social Sciences; History: Theories and Methods; Holocaust, The.

Bibliography

Angehrn, E., 1985. Geschichte und Identität. Berlin, De Gruyter, New York.
Ankersmit, F.R., 2012. Meaning, Truth and Reference in Historical Representation. Ithaca, Cornell University Press, New York.

Aristotle, 1963. In: Bambrough, R. (Ed.), The Philosophy of Aristotle. New York.
Berger, S., Lorenz, C. (Eds.), 2010. Nationalizing the Past. Historians as Nation Builders in Modern Europe. Palgrave Macmillan, Houndsmills.
Burke, P., 1992. History and Social Theory. Polity press, Cambridge.
Carr, D., 1986. Time, Narrative and History. University of Indiana Press, Bloomington.
Danto, A.C., 1965. Analytical Philosophy of History. Cambridge University Press, Cambridge, New York.
Droysen, J.-G., 1857–82. Historik. Stuttgart/Bad Cannstatt 1977.
Fain, H., 1970. Between Philosophy and History: The Resurrection of Speculative Philosophy of History within the Analytic Tradition. Princeton University Press, Princeton.
Fay, B., Pomper, Ph, Vann, R. (Eds.), 1998. History and Theory. Contemporary Readings. Cambridge.
Foucault, M., 1980. Power/Knowledge. Harvester Wheatsheaf, Brighton.
Frei, N., Kansteiner, W. (Eds.), 2013. Den Holocaust erzählen. Historiographie zwischen wissenschaftlicher Empirie und narrativer Kreativität. Jena.
Friedländer, S. (Ed.), 1992. Probing the Limits of Representation. Nazism and the "Final Solution". Cambridge, Mass.
Gardiner, P. (Ed.), 1959. Theories of History; Readings from Classical and Contemporary Sources. New York.
Gossman, L., 1990. Between History and Literature. Harvard University Press, Harvard.
Hardtwig, W., 1990. Geschichtskultur und Wissenschaft. Munich.
Hay, D., 1977. Annalists and Historians. Western Historiography from the 8th to the 18th Century. London.
Iggers, G., 1997. Historiography in the Twentieth Century. From Scientific Objectivity to the Postmodern Challenge. Wesleyan University Press, Hanover and London.
Kelley, D. (Ed.), 1991. Versions of History from Antiquity to the Enlightenment. Yale University Press, New Haven and London.
Kocka, J., 1977. Sozialgeschichte. Begriff – Entwicklung – Probleme. Göttingen.
Kocka, J., 1989. Geschichte und Aufklärung. Aufsätze, Göttingen.
Koselleck, R., 1979. Vergangene Zukunft. Zur Semantik Geschichtlicher Zeiten. Suhrkamp, Frankfurt a.M.
Koselleck, R., Lutz, H., Rüsen, J. (Eds.), 1982. Formen der Geschichtsschreibung. Munich.
LaCapra, D., 1994. Representing the Holocaust. History, Theory, Trauma. Ithaca and London.
Lorenz, Chr, 1997. Konstruktion der Vergangenheit. Eine Einführung in die Geschichtstheorie. Cologne.
Lorenz, Chr, Bevernage, B. (Eds.), 2013. Breaking up Time. Negotiating the Borders between the Present, Past and Future. Göttingen.
Lowenthal, D., 1985. The Past is a Foreign Country. Cambridge University Press, Cambridge.
Lübbe, H., 1977. Geschichtsbegriff und Geschichtsinteresse. Analytik und Pragmatik der Historie. Schwabe and Co, Basel and Stuttgart.
McClosky, D. (Ed.), 1987. The Rhetoric of the Human Sciences: Language and Arguments in Scholarship and Public Affairs. University of Wisconsin Press, Madison.
McNeill, W., 1986. Mythistory and Other Essays. Chicago University Press, Chicago and London.
Megill, A. (Ed.), 1994. Rethinking Objectivity. Duke University Press, Durham.
Meran, J., 1985. Theorien in der Geschichtswissenschaft. Göttingen.
Mink, L., 1987. Historical Understanding. Cornell University Press, Ithaca, NY.
Momigliano, A., 1990. The Classical Foundations of Modern Historiography. University of California Press, Berkeley.
Mormann, T., 1997. Ist der Begriff der Repräsentation obsolet? Zeitschrift für philosophische Forschung 51, 349–366.
Nelson, J., Megill, A., Mc Closkey, D. (Eds.), 1987. The Rhetoric of the Human Sciences. Language and Argument in Scholarship and Public Affairs. University of Wisconsin Press, Madison.
Nietzsche, F., 1981(1874). Vom Nutzen und Nachteil der Historie für das Leben. In: Nietzsche, F. (Ed.), Unzeitgemäße Betrachtungen. Frankfurt a.M., pp. 95–185.
Novick, P., 1988. "That Noble Dream". The 'Objectivity Question' and the American Historical Profession. Cambridge University Press, Cambridge.
Paul, H., 2011. Hayden White. The Historical Imagination. Polity Press, London.
Plumb, J., 1973. The Death of the Past. Harmondsworth, Penguin, England.
Pocock, J., 1962. The origins of the study of the past: a comparative approach. Comparative Studies in Society and History 4, 209–246.
Ricoeur, P., 1983–85. Temps et récit, 3 vols. Seuil, Paris.
Rigney, A., 1990. The Rhetoric of Historical Representation: Three Narrative Histories of the French Revolution. Cambridge University Press, Cambridge.
Rüsen, J., 1983–89. Historische Vernunft. Grundzüge einer Historik, 3 vols. Göttingen.
Scholtz, G., 1991. Zwischen Wissenschaftsanspruch und Orientierungsbedürfnis. Zu Grundlage und Wandel der Geisteswissenschaften. Suhrkamp, Frankfurt a.M.
Scott, J., 1988. Gender and the Politics of History. Columbia University Press, New York.
Stern, F. (Ed.), 1970. Varieties of History. From Voltaire to the Present. Macmillan, London.
Tilly, C., 1981. As Sociology Meets History. Academic Press, New York.
Tollebeek, J., 1990. De Toga Van Fruin. Denken over Geschiedenis in Nederland 1860–1960. Amsterdam.
Veyne, P., 1978. Comment on Écrit L'histoire? Seuil, Paris.
Walsh, W., 1967. Philosophy of History: An Introduction. Harper & Row, New York.
Wehler, H.-U., 1973. Geschichte Als Historische Sozialwissenschaft. Suhrkampf, Frankfurt a.M.
White, H., 1973. Metahistory. The Historical Imagination in Nineteenth Century Europe. Johns Hopkins University Press, Baltimore.

History: Theories and Methods

Chris Lorenz, VU University Amsterdam, The Netherlands; and Ruhr-University Bochum, Bochum, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since the institutionalization of history as a distinct academic discipline in the nineteenth century, the truth claim of history has basically been founded on some idea of the historical method. What was considered to constitute this 'method' of history, however, has varied considerably over competing conceptions of history and over time. The various conceptions have been connected to different theories about what constituted history *itself* (as an object) and what was considered as reliable *knowledge* of this object (if possible at all). Therefore, the debates about the methods of history have always been embedded in debates about theories of history – and vice versa. As natural science was often considered to be the methodological opposite of history, images of natural science have also influenced the debate about historical method and theory. This article traces the main course of these debates since the nineteenth century and offers some ways to classify them.

Introduction

Since history turned into an academic discipline in the nineteenth century and since professional historians claimed a privileged epistemological status for their approach vis-à-vis nonprofessional approaches to history, debates about the role of methods and theories in history have followed a clear pattern. In periods of relative uncertainty and of crisis about the status of history as a distinctive academic discipline – such as 1880–1920 and 1965–85 – many historians participate in these debates. In those periods, *Methodenstreit* is more or less intense and visible, while in periods of relative certainty and of academic recognition, this type of debate usually recedes into the background. So, seen from an *historical point of view*, reflection on the role of methods and theories in history is in a sense a symptom of crisis – at least outside the small group of historians and philosophers who *permanently* reflect on history in their capacity as philosophers of history. This direct connection between a 'reflective mood' among professional historians and their sense of professional uncertainty makes sense because, as will be elucidated below, method and/or theory have usually been presented as the discipline's *cognitive foundations*, especially its claim to produce 'objective' knowledge of the past. Therefore doubts and debates about method and theory in history are always doubts and debates about the discipline's cognitive status, its academic credentials and its claim to objectivity.

Seen from the *point of view of the sociology of science*, debates about the role of methods and theories in history have usually been directly linked to battles waged by the proponents of different conceptions of history for academic legitimacy and supremacy, that is, for academic recognition and reputation. In periods of serious competition for 'intellectual capital' (Bourdieu) between several competing factions, operating within the same 'academic field,' debates about methods and theories in history have been intense. This not only holds true for history, but also for other social sciences such as sociology, economics, psychology and, therefore, is a characteristic of pluralistic or polyparadigmatic disciplines.

Characteristic for the periods of increased interest in theory and method is also an increased interest in the history of the discipline from both an intellectual angle and an institutional angle. This interest in disciplinary history is usually motivated by a desire to locate the origins of the preferred conception of historical theory and method in the discipline and to generate academic credentials by locating prestigious founding fathers or precursors of the preferred conception in the discipline's pedigree – the closer to 'the origins,' the better. Therefore, in history, an increased interest in the history of history writing, alias in *historiography*, can usually be observed alongside an increased interest in theory and method: theoretical and methodical reflections usually are accompanied by reflections on the origins, nature, and variety of historiography.

Seen from a *philosophical point of view*, debates about the role of theories and methods in history often have a blurred character due to the vague philosophical status of the central concepts of theory and method. Therefore, some elementary conceptual clarifications are called for in order to answer the question why theory and method in history have turned into a heavily contested area at times since history turned into an academic discipline.

Theory and Method in History: Some Preliminary Distinctions

In discussing the role of theory and method in history it is necessary to introduce some basic distinctions. Due to the double meaning of history as *res gestae* and as *historia res gestae*, theory in history can refer both to *history as a process* and the *knowledge of that process*. Therefore the *first basic distinction* to make is the distinction between (1) *theories, that deal with characteristics of history as a process* (such as, for instance, Marx's theory of history as class struggle or Treitschke's theory of history as nation formation), and (2) *theories that deal with the characteristics of knowledge of history* (such as, for instance, that scientific knowledge of history has the form of lawlike knowledge or the form of concrete particulars). Theories of type (1) can be called *substantial* or *material* theories of history, while theories of type (2) can be called *cognitive* theories of history, i.e., theories of historical knowledge.

Cognitive theories of history can further be subdivided in (1) *epistemological theories* and (2) *methodological theories*, while substantial theories of history can also be subdivided in finer subsets. Substantial and cognitive theories usually have been interrelated, because presuppositions about *what* can be known

about the past are linked to presuppositions of *how* the past can be known. Substantial theories, for instance, that hold that the historical process is characterized by underlying mechanisms and regularities, analogous to theories of classical mechanics, have been interrelated with cognitive theories that posit that real knowledge of history implies knowledge of general laws. And the substantial theory that the historical process is characterized by a fundamental contingency and *not* by hidden regularities alias a 'motor' has been interrelated with cognitive theories that posit that real knowledge of history is knowledge of particular facts instead of general laws.

Substantial theories, in this way, not only have consequences for methodological positions, but also for epistemological positions. Substantial theories, that posit some kind of affinity or even identity of the knowing subject and the object of history in contrast with nature (as was the case with Hegelian idealism and its offshoots in historicism, such as Droysen's and Dilthey's versions of historicism), have usually been interrelated with methodological theories that posit a distinct kind of method for the human sciences – such as *understanding* or *Verstehen* or *interpreting* – in contrast with the methods of the natural sciences – usually defined as some form of causal or lawlike *explanation*. And these distinct methodological theories have, in their turn, usually been interrelated with distinct epistemological theories, linking methods of understanding to hermeneutic or phenomenological theories of knowledge and methods of explanation to empirical theories of knowledge. Before taking a better look at this terrain, however, it is first necessary to clarify the central *concepts* of method and theory as well as their interrelationships in history. When analyzed more closely, both method and theory in history appear to have a broad variety of meanings.

Historical Method(s): Concept and History

Analyzing the meaning of 'method' in history, it is useful to start with a clarification of the term 'method' in other disciplines first (J. Meran, 1985). In the mathematical sciences, it is common to refer to axiomatic-deductive, to infinitesimal and to inductive methods, meaning *a set of formal, technical procedures* in order to solve problems within a specified range. Outside the mathematical sciences, the meaning of method is usually far less defined. In philosophy, for instance, there is a long tradition referring to methods of all sorts, ranging from Descartes' *Discourse on method* to Kant's *transcendental method*, from Marx's *dialectical method* to Husserl's *phenomenological method* and to de Saussure's *structuralist method*, to name just a handful. These 'methods' in philosophy, however, do *not* constitute defined formal rules for specified ranges of problem solving, but constitute complex sets of ideas, concepts, and rules of thumb for analyzing specified domains such as scientific knowledge, the human consciousness, the history of mankind, the meaning of concepts, or the meaning of language in use or of language as a system of signs. 'Method' in philosophy, thus, is definitely *not* identical with a set of formal techniques, but a loosely circumscribed research program. This diffuse and nonformal meaning of method is also characteristic of the human sciences.

In history, the term historical method was first introduced in a systematic way in the sixteenth century by Jean Bodin in his treatise of source criticism, *Methodus ad facilem historiarium cognitionem* (1566). Characteristically, Bodin's treatise intended to establish the ways by which reliable *knowledge of the past* could be established by checking sources against each other and by so assessing the reliability of the information conveyed by them, relating them to the interests involved. Historical method was thus at the same time intended to defeat skeptical doubts with regard to the possibility of reliable knowledge of the past. Skepticism (alias historical Pyrrhonism) had been fed by a large quantity of scriptural forgeries parading as original historical documents, primarily produced by competing churches and nobles to back up their claims to rights and property. Therefore, the traditional claims to impartiality, truth and trustworthiness of historians, dating back to classical Antiquity, were no longer believed, and were now replaced by a claim to impersonal and intersubjective method.

This direct *relationship between history's claim to truth and method* would characterize all later treatises on historical method, although what was meant by method did change over time. In the eighteenth century, historians of the Enlightenment, such as Schlözer, included not only the source critical techniques as *the* historical method (*Forschung*) but also the way the results of historical research were synthesized into a textual whole (*Darstellung*) and presented to an audience. This broader conception of method had its roots in the classical rhetorical tradition and would only be rediscovered in historical theory from the 1970s onward. The German historian and historical theorist Johann Gustav Droysen (Droysen, 1857–1882) has become one of the pivotal historians behind this 'rediscovery' of the textual and rhetorical dimensions of history, because his treatment of historical method in the earlier and the later versions of his *Historik* embodies the transition from the broader to the narrower conception of historical method (Meier and Rusen, 1986; Nelson et al., 1987).

The larger part of the nineteenth and twentieth century would be dominated by the research-oriented conception of historical method of the so-called *Historical School* in Germany, led by historians as Berthold Niebuhr and Leopold Ranke. Their conception of history has long been regarded as the beginning of modern, scientific history, and harked back to the narrow conception of historical method, limiting the methodical character of history to source criticism (Grafton, 1997). The historical composition and presentation – the *Darstellung* – was no longer regarded as methodical and was now seen as the literary, artistic, or aesthetic component of history. Therefore, the question of whether history belongs to the arts or to the sciences could and did arise from the nineteenth century onward, a question that became part of the continuing *Methodenstreit* (F. Stern, 1970; Fay et al., 1998).

In the nineteenth century, during the process of rapidly increasing disciplinary differentiation and academic institutionalization, it became customary in the human sciences to identify disciplines with *one* specific method that was presented as the guarantee of its 'truth' and 'objectivity.' In this process, sociology was identified with *the* sociological method and history with *the* historical method. Canonical formulations of this conception of historical method can be found in E. Bernheims *Lehrbuch der historischen Methode* (Leipzig, 1889) or in Ch. Langlois and Ch. Seignobos *Introduction aux études historiques* (Paris, 1897).

Historical method in the nineteenth century usually referred to the techniques for conducting historical *research* on written materials, the so-called *sources*. These techniques – partially taken over from philology – were traditionally subdivided into three parts, corresponding with three subsequent steps in research: *heuristics*, that is the techniques for locating the relevant sources; *source criticism*, the techniques by which the temporal and spatial origins of the sources are established as well as their authenticity; and *interpretation*, by which the 'cleansed' information derived from the sources is put together by means of interpretation in order to infer 'what really happened.'

Usually a distinction was made between *primary* and *secondary* sources. Primary sources consist of the written material produced more or less contemporaneous with the object of research, ideally recorded by eyewitnesses (model: the official records of parliamentary debates). The time span between primary sources and what they represent is thus as small as possible, because distance in time is regarded as a major source of unreliability of the information conveyed. Secondary sources consist of records written a considerable time *after the fact* and not by eyewitnesses (model: 'hear say'). As the information in secondary sources is based on *other* sources, it is, at best, 'second hand.' Therefore, it is regarded as less reliable than primary source information. According to the dominant, Western conception of history since the nineteenth century, serious academic history can only be based on primary sources, because those are the only reliable source of information on the past. Therefore, the archives were usually considered to be the workshop of the professional historian.

According to the same nineteenth-century conception, nonscriptural or 'low' cultures were primitive *by definition* and simply had no history comparable with 'high cultures.' In this way the scriptural, source-based conception of historical method defined – and limited – the object of academic history to the history of 'high cultures,' especially the history of the West. It would last until the second half of the twentieth century before these biases in the very conception of historical method would be seriously criticized, especially by 'history from below' and by *postcolonial theory*. It dawned on an increasing number of historians that entangled with the history of 'the West' there was and is a history of 'the rest.' As a consequence, *global* and *world history* have been on the rise since the 1990s, using comparative and transnational approaches (Haupt and Kocka, 2010).

Another traditional distinction was that between *internal* and *external* source criticism. *Internal* source criticism is actually *textual* criticism, i.e., the hermeneutic procedures by which a text is reconstructed (if it is passed down in an incomplete form or in several variants) as a whole and the textual procedures by which authorship and textual authenticity are checked and its meaning established. *External* source criticism consists of those procedures by which a source is checked by information derived from *other* sources – written and material. The so-called *auxiliary sciences*, such as paleography, chronology, sillography, toponomy, numismatics, and diplomactics, usually play an important role in this phase when historians are dealing with older periods. And when modern historians are dealing with nonscriptural, oral cultures or with cultures, that left few or no scriptural traces, archaeology and anthropology often function as auxiliary science. Only in the course of the twentieth century did reference to *multiple methods* within *one* discipline became the rule, thereby reopening the problem of truth and objectivity and paving the way for a new wave of skepticism. For if *the* method can no longer be regarded as *the* procedural guarantee for the truth and objectivity of the discipline, what could? So an influential argument ran (R. Bernstein, 1983; Megill, 1994). Recognition of plural methods within one discipline, therefore, has often been regarded as a threat to the discipline's academic credentials.

From the late 1960s onward, the process of methodological differentiation within all of the human sciences took an unprecedented pace, leading to a temporary fragmentation of most disciplinary fields and a questioning of most traditional disciplinary boundaries and self-definitions. Significantly, the new, 'polyparadigmatic' self-image was at the same time regarded by many as a *crisis* of their discipline. This also held for history in the 1960s and 1970s regarded by many as one of the social sciences (Novick, 1988; Appleby et al., 1994; G. Iggers, 1997). Later, the changes in styles of 'doing history' were often characterized in terms of an ever increasing number of 'turns' (linguistic, cultural, spatial etcetera) (American Historical Review, 2012).

Historians identified the quantitative method, the biographical method, the comparative method, the ideal-typical method, understanding as a method ('Verstehen'), the causal method ('Erklärung'), discourse analysis as a method, the micro-historical method, and to a psychoanalytical method in history. Other historians claimed that the concept of method had become severely overstretched, holding that most advertised *methods* in fact were *theories* about historical knowledge, that could only be checked by *the* historical method of source criticism. Therefore, a closer look at the relationship between historical method and theories is necessary.

Theories of Historical Knowledge (1): Empiricism and Antiempiricism

The connections between historical method and historical theory can be located both on the level of *epistemology* and on the level of *methodology*. As epistemology is the theory of knowledge, historical epistemology deals with the theory of historical knowledge, i.e., the conditions of its possibility, its range, sources, and foundations. Basically two epistemological positions have been defended in relation to historical knowledge from the nineteenth century onward, empiricism and antiempiricist idealism.

The problem whether and how it is possible to acquire knowledge of other minds, i.e., the *mental states of others*, such as thoughts and intentions, separated from the knowing subject in both time and space – the typical problem of the historian – has been a fundamental one since the mid-nineteenth century. According to *empiricism*, mental states could only be known as far as they correlated with empirically observable epiphenomena. The empirical dispositions or expressions *of* other minds are seen as the only access *to* other minds (Hempel, 1966).

The proponents of phenomenological and hermeneutic views (*hermeneutics* is the theory of interpretation, originally of

text interpretation), to the contrary, have always affirmed this possibility of a nonempirical access to mental phenomena. Phenomenology claimed *introspective methods* as the access to the mental universe; hermeneutics, far more influential in the theory of history than phenomenology, claimed that *'reading' other minds* could be conceived of as *analogous to reading texts*.

The fundamental cognitive activity in all hermeneutic theories is the *interpretation* of meaningful symbols and symbolically structured phenomena (like actions and cultural objects), whereas all empiricist theories regard the *observation* of objects as the fundamental cognitive activity. Just as it is possible to know the ideas expressed in a text by reading and interpreting, it is possible to know the ideas expressed in actions and in all artificial objects by *inferring* the unobservable ideas from their observable expressions through *reenactment* and *interpretation* by the historian (Droysen, 1977; Dilthey, 1981; R. Collingwood, 1993). This possibility was guaranteed, according to the hermeneutic views, by the basic unity of mankind, implying the *basic unity* of subject and object of knowledge in the human sciences. Therefore each *Alter* was an *Alter Ego*, accessible through a process of intuitive identification and transfer in thought or idea (therefore idealism). So in historical epistemology, just as in the other human sciences, the basic divide between 1850 and 1970 has been the one between empiricist and antiempiricist, idealist positions. It would last until the 1980s, with the rise of postmodernism and its offshoots in the discipline of history, before the unitary presuppositions of hermeneutic theories would be seriously criticized and questioned. Western (White) Male was no longer accepted by all as the model of all mankind. This has led to a renewed emphasis on the basic diversity of knowing subjects and their epistemological and methodological relevance for disciplines like history (Scott, 1991).

Theories of Historical Knowledge (2): Unity vs Duality of Method

Between 1850 and 1970, this basic divide in the domain of epistemology roughly corresponded with the basic divide in the domain of historical *methodology*, i.e., the divide between the proponents of the theory of the *unity of scientific method* and the proponents of the theory of a fundamental *methodological dualism of the human and the natural sciences*.

Methodology as such aims at the *rational reconstruction* of the normative rules that are constitutive for a specific discipline and its development. Rational reconstruction, in contrast with historical reconstruction, presupposes that disciplinary practices and their development can be reconstructed in terms of *logical* and *empirical* arguments; this presupposition is, of course, an *idealization* of the rationality of real disciplinary practice, because apart from rationality, power and privilege also play their role in reality. Methodology abstracts from those nonrational factors in disciplinary life and aims to reconstruct the cognitive criteria and ideals, by which the practitioners of a discipline judge each other's products and evaluate their relative quality. Methodology, therefore, *presupposes* the existence of a basic cognitive consensus and rationality among the practitioners of a discipline. This presupposition, of course, has been criticized both for the natural as for the human sciences (Goodman, 1978; Rorty, 1986; Novick, 1988).

Next to the basic ideal of truth, other cognitive ideals of a discipline are reconstructed by its methodology, such as the *explanatory ideals* of the discipline and its *criteria of progress*. At this level the basic methodological divide between the proponents of the *unity* of scientific method and the proponents of the *dualism* of scientific methods must be located, because the divide revolves around the question of whether there is *one* explanatory ideal for *all* disciplines or not. This issue has dominated the methodological debates for more than a century since 1850 and is known under the labels of the *explanation vs understanding controversy* (Gardiner, 1959; Hausmann, 1991). According to the *unity of science view*, also often labeled as *positivism*, there was just *one* model or logic of explanation and of development for *all* empirical disciplines. Classical mechanics in the nineteenth century was usually considered to be the paradigm for all other disciplines, including human sciences like history. The only difference between the various disciplines was their *phase of development*. Underdeveloped disciplines (like history) showed little similarity to Newtonian mechanics while developed ones did. Immature disciplines only produced 'explanation sketches' (C. Hempel) that had to be developed into real explanations. The only difference between disciplines was thus a matter of scientific *maturity* and not a matter of explanatory *logic*.

Real *scientific explanation* boiled down to *subsuming an event* to be *explained under a general law*. Therefore, the event to be explained can be *logically deduced* from the general law in combination with the given circumstances (the so-called initial conditions) and is at the same time, given the law and the initial conditions, predictable. According to this view on the logic of scientific explanation, that is also known under the headings of the *subsumtion model*, the *deductive-nomological model*, or the *covering law model of explanation*, explanation and prediction in science are basically the same; explanation is really prediction in the past alias retrodiction. Scientific explanation, in this view, is showing why what happened, was – given the known laws and circumstances – expected to happen because it was bound to happen: explanation boiled down to show what happened was *logically necessary*. Given the fact that the general laws were usually (though not exclusively) interpreted as *causal* laws, formulating universal and invariant connections between causes and effects, scientific explanation was usually seen as causal explanation and *explanatory necessity* also as *causal necessity*: scientific explanation specifies the causes of what happened by formulating its causal determinations.

Now although this unitary theory of scientific explanation was amended and modified in time, in order to account for statistical and probabilistic forms of explanation in the natural sciences, its status was based on the *assumption* that 'positivist' theory at least provided an adequate reconstruction of *natural* science and its development from the seventeenth century onward. This assumption has been increasingly undermined by *postempiricist* and *postpositivist* philosophy of science roughly since the 1960s and 1970s. Both the empiricist underpinnings of the covering law model and its logical claims were undermined by historical examples, arguing that the supposedly paradigmatic cases in the history of science did not fit the basic logical models of philosophy of science. Therefore this 'turn' in

philosophy of science is often called the *historical turn*, because history was mobilized against explanatory logic (Lakatos and Musgrave, 1970). However, before examining the implications of postempiricism and postpositivism in philosophy of science for the opposition between classical unitary and dualistic methodological theories, an overview of the dualistic positions and arguments is called for.

Theories of Historical Knowledge (3): The Autonomy of History

The philosophical defense of the *autonomy of historical explanation* vis-a-vis the explanatory models of natural science usually has started off with a critique of the unitary claim. From Droysen and Dilthey, through Windelband, Rickert, Weber, Collingwood, and Gadamer, to White and Ricoeur, the argument has been developed that historical explanation is not only *empirically* different, but *logically* different from the covering law model. This claim of autonomy has traditionally been defended in three variants: an ontological version, an epistemological version, and a methodological version. The *ontological argument* for methodological dualism was first systematically developed by J.G. Droysen and later by R.G. Collingwood and his school. It usually harks back to Vico and Hegel, because it is based on an idea of a *subject–object identity* in history: in history, mankind confronts itself, and history, therefore, is basically the autobiography of mankind. History is fundamentally the process in which mankind *expresses* and *develops* itself in the form of diverse cultures; and because mankind has *made* history, it is capable of *knowing* history in a direct way, according to this argument, just as the writer of an autobiography knows the history in case in a direct way. Historians basically try to reconstruct cultures and the particular ideas expressed in its cultural practices (such as religions and rituals) and its material artifacts (such as its temples and sacred objects). When historians offer *explanations* for the past, they construct *intelligible and plausible connections* between the culture and the ideas expressed in it on basis of their research of the sources under consideration. When they succeed in this task, they offer an *understanding* of the phenomena under consideration (*Verstehen*). *Understanding history* is conceived of as basically similar to *understanding a text*; the major difference is that the texts of history have more or less disappeared and therefore must be reconstructed first on the basis of their traces by research. This type of intelligibility is possible, because in history man confronts man and therefore intuitive understanding is possible in contrast to the situation in the sciences of nature, where the knowing subject confronts an external object.

The intelligibility sought for by historians – *historical understanding* – is thus centered on *connecting ideas to their expressions in particular spatiotemporal contexts*, and proceeds by going back and forth between parts and whole in the so-called *hermeutic circle* or *spiral*. Historical understanding is therefore logically different from scientific explanation: the type of connections, which historians present, are not causally determined and cannot be deduced from general laws. *Historical* connections are *not necessary but contingent* connections, which always could have been different from their actual form.

Neither are they based on empirical observation, but on plausible *inferences*, based on critical research of the sources and on analogical reasoning. Historical understanding at best produced *plausible* and *probable* knowledge, and never *certain* or *founded* knowledge, as scientific explanation, based on sensory experience and logic presumably did. And being dependent on intellectual, interpretative *activities* by the historian, which could be trained but not formalized, historical understanding, thus, remained 'subjective,' i.e., embedded in the culture, time, and place of the historian. On the basis of this argument, Gadamer in his *Truth and Method* (1960) would radically historicize the idea of historical understanding.

The second type of argument in defense of methodological dualism is not ontological, but basically *epistemological*. It was first systematically developed by Wilhelm Dilthey and bears a resemblance to Droysen's argument, because it also harks back to Hegel's idea of a subject–object identity in history with its implied 'history as the autobiography of mankind model.' However, Dilthey does not ground historical method directly in the subject–object structure of the historical process *itself*, but in the *inner experience* ('Erlebnis') of historical facts as *mental facts*. The 'mental facts' of history and of other *Geisteswissenschaften* consist of feelings, thoughts, and strivings of human beings living in the past that are only accessible through the mind and the inner experience of the historian. Like Droysen, Dilthey posits that the historian makes the mental facts of the past intelligible as expressions of ideas, by putting them in a particular spatiotemporal context alias in a *Lebenszusammenhang*. And like Droysen, he contrasts historical understanding with scientific explanation by rejecting causal, lawlike reasoning for history, but unlike Droysen, Dilthey emphasizes the contrast of the 'inner experience' of the historian with the 'outer experience' of the scientist as the foundation of methodological dualism.

The third type of argument in defense of methodological dualism is of a *methodological* character. It is *not* based on an ontological claim that history deals with an object fundamentally different from nature, and *neither* is it based on an epistemological claim that the historical facts are different from natural facts, being accessible to an 'inner' way of knowing. According to the methodological argument, as developed by Windelband, Rickert, and Weber, *two fundamental ways of knowing or methodological approaches* can be distinguished that are connected by two fundamentally different aims or knowledge interests. The first way to approach an object is a *generalizing* method; in this case, one abstracts from the particularity of the objects of knowledge and tries to detect and explain its general characteristics. According to Windelband, this generalizing method was specific for the *nomothetic* sciences (Greek: nomos = law; tithemi = to posit), such as the natural sciences. The second fundamental approach of an object is the *particularizing* method; in this case, one abstracts from the general characteristics of an object of knowledge and tries to describe its specific characteristics. According to Windelband, this particularizing method was specific for the *idiographic* sciences (Greek: idios = particular; graphein = to describe), such as history. However, both the generalizing and the particularizing method are *independent* of the nature of their objects

of knowledge; nature can be studied in an idiographic way (for instance, in geology) and history can be studied in a nomothetic way (for instance, in developmental psychology). It all depends on one's methodological *interest*.

Rickert and Weber would develop Windelband's basic ideas by introducing the dichotomy between sciences of nature (*Naturwissenschaften*) and sciences of culture (*Kulturwissenschaften*). Rickert introduced the idea that knowledge interests behind disciplines were guided by *values* (*Wertbezogenheit*); Weber pointed out that values usually change through time. According to him the change of the guiding values behind the sciences of culture could explain why those disciplines did not accumulate knowledge in the systematic form like the natural sciences, and why they seemed to enjoy an 'eternal youth' (Weber, 1922). At best the sciences of culture could accumulate knowledge at the intermediate level *between* individualizing descriptions and universal explanations in the form of *ideal types*, according to Weber. He suggested comparison as the method at this intermediate level. Lübbe, Veyne, and Mink would develop the Weberian methodological position later, identifying history with the particularizing methodological interest (Lübbe, 1977; Veyne, 1978; Mink, 1987).

All in all, *different conceptions of historical method* are connected to *different theories of historical knowledge*. Since the characteristics of historical knowledge are usually defined in contrast with scientific knowledge as the major contrast class, theories of natural science are also presupposed in theories of historical knowledge. Fundamental changes in theories of natural science, therefore, affect theories of historical knowledge by *detour*. The rise of postempiricism, postpositivism, and the historical turn in philosophy of natural science, thus has undermined most traditional, classical defenses of methodological dualism in an indirect way by undermining the classical unitary model of science. A few remarks must suffice.

Postempiricism has effectively undermined the presupposition that scientific knowledge can ever be grounded on *certain* foundations in sensory experience. The opposition between grounded and certain knowledge on one side – traditionally attributed to the natural sciences – and inferential and probable knowledge on the other – traditionally attributed to the human sciences – has thus vanished. This opposition has been replaced by the insight that *all* sensory experience is guided and framed by concepts and theories. All empirical knowledge, therefore, is based on interpretative *activities* of the knowing subject because it is *theory laden*. Therefore, *all* knowledge is *fallible* and open to later *re*interpretations. Postempiricism, thus, has dissolved the presupposed absolute contrast between a 'subjective' kind of knowledge, based on interpretation – characteristically for the *Geisteswissenschaften*, the *Kulturwissenschaften*, or the historical sciences – and an 'objective' kind of knowledge, based on empirical observation and experiment, not in need of interpretation. Postempiricism has thus relativized both the distinction between theoretical knowledge and empirical knowledge, on one side, and the distinction between interpretative knowledge and empirical knowledge, on the other. It has, thus, brought *conceptualization* and *interpretation* to center stage in *all* disciplines (Skinner, 1985). Doing so, postempiricism has undermined traditional underpinnings of the idea of fundamentally distinct scientific and humanistic methods and the idea of a fundamental contrast between the 'objectivity' of science and the 'subjectivity' of the humanities. Both scientific domains aim at *empirical* argument and *intersubjectivity* (Megill, 1994; Oexle, 1998; Daston and Gallison, 2007).

Postpositivism, based on postempiricism, has in turn effectively undermined the presupposition that all scientific explanations can be reduced to one basic model, i.e., the subsumtion model of explanation. By discrediting this presupposition for *natural* science (Salmon, 1990), postpositivism a fortiori has discredited the subsumtion model for the humanities. Moreover, the fact that the explanation of mental states by nonmental states has not succeeded up to the present, has also discredited unitary views (Bohman, 1991).

In the theory of history, postempiricism and postpositivism have led to a renewed reflection on the nature of historical knowledge in *postmodernism* and in a renewed interest in approaches inspired by *literary theory*. After the classical science model was discredited, the roots of history in rhetoric and literature were rediscovered, leading to a renewed interest in the linguistic forms in which history is represented, i.e., in theories of historical *representation*. The postmodern vanguard, often identified with the *linguistic turn*, pushed both theory and method in history into the background, locating historical representation *beyond* epistemology and methodology, on the terrain of aesthetics, rhetoric, and politics (White, 1973; Megill, 1985; Ankersmit and Kellner, 1995; Fay et al., 1998).

This move, in fact, harks back to the *pre*methodical and *pre*disciplinary roots of history, cutting history loose from its claim to constitute a truth-seeking discipline and to produce 'objective' knowledge (Rüsen, 1983–1989; Lorenz, 1997). Nevertheless, the pendulum is already swinging back, leading to broader approaches of history that keep up the cognitive claims of the discipline based on *both* research (Forschung) and composition (Darstellung), while integrating the literary and the practical dimensions of history. Moreover, recently unity of science claims has returned in some forms under the banner of the so-called *digital humanities* that use tools provided by computing (such as data visualization, information retrieval, data mining, text mining and statistics). Therefore the present landscape of theory and methods in history shows even more variety than 10 years ago (Berry, 2012).

See also: Biography: Historical; Comparative History; Data Bases and Statistical Systems: Archives and Historical Databases; Discourse Studies; Hermeneutics, History of; Hermeneutics; Historical Thought and Historiography: Christian Tradition; Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; Historicism; History and Literature; History and the Social Sciences; History, Professionalization of; History: Representation Within and Functions of; Positivism, History of; Psychohistory; Public Sphere: Nineteenth and Twentieth-Century History; Quantification in History; Subaltern History; Time, Chronology, and Periodization in History.

Bibliography

AHR Forum: historiographic "turns" in critical perspective. American Historical Review 117 (3), June 2012, 671–813.

Ankersmit, F.R., Kellner, H. (Eds.), 1995. A New Philosophy of History. University of Chicago Press, Chicago, London.

Appleby, J., Hunt, L., Jacob, M., 1994. Telling the Truth about History. Norton, New York.

Bernheim, E., 1889. Lehrbuch der historischen Methode. Duncker, Leipzig.

Bernstein, R., 1983. Beyond Objectivism and Relativism: Science, Hermeneutics and Practice. Oxford.

Berry, D.M. (Ed.), 2012. Understanding Digital Humanities. Palgrave, Basingstoke.

Bodin, J., 1595. Methodus Ad Facilem Historiarium Cognitionem. Apud Iacobum Stoer, Geneva.

Bohman, J., 1991. New Philosophy of Social Sciences: Problems of Indeterminacy. Cambridge University Press, Cambridge, UK.

Collingwood, R.G., 1993(1946). The Idea of History. Knox, Oxford.

Daston, L., Galison, P., 2007. Objectivity. Zone Books, New York.

Dilthey, W., 1981(1905–1911). Der Aufbau der geschichtlichen Welt in den Geisteswissenschaften. Suhrkamp, Frankfurt a.M.

Droysen, J.-G., 1857–1882. Historik, Stuttgart/Bad Cannstatt 1977.

Fay, B., Pomper, P., Vann, R. (Eds.), 1998. History and Theory. Contemporary Readings. Blackwell, MA.

Gadamer, H.-G., 1972(1960). Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik. Tübingen.

Gardiner, P. (Ed.), 1959. Theories of History; Readings from Classical and Contemporary Sources. The Free Press, New York.

Ghandi, L., 1998. Postcolonial Theory: A Critical Introduction. Columbia University Press, New York.

Goodman, N., 1978. Ways of Worldmaking. Indianapolis, IN.

Grafton, A., 1997. The Footnote: A Curious History. Harvard University Press, Cambridge, MA.

Haupt, H.-G., Kocka, J. (Eds.), 2010. Comparative and Transnational History. New York/Oxford.

Hausmann, T., 1991. Erklären und Verstehen. Zur Theorie und Pragmatik der Geschichtswissenschaft. Suhrkamp, Frankfurt a.M.

Hempel, C., 1966. Explanation in science and history. In: Dray, W. (Ed.), Philosophical Analysis and History. Harper & Row, New York, pp. 95–127.

Iggers, G., 1997. Historiography in the Twentieth Century. From Scientific Objectivity to the Postmodern Challenge. Wesleyan University Press, Hanover and London.

Koselleck, R., 1979. Vergangene Zukunft. Zur Semantik Geschichtlicher Zeiten. Suhrkamp, Frankfurt a.M.

Kuhn, T., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.

Lakatos, I., Musgrave, A. (Eds.), 1970. Criticism and the Growth of Knowledge. Cambridge University Press, Cambridge.

Lorenz, C., 1997. Konstruktion der Vergangenheit. Eine Einführung in die Geschichtstheorie. Cologne-Weimar-Vienna.

Lorenz, C., 2011. History and theory. In: Woolf, Daniel, Schneider, Axel (Eds.), The Oxford History of Historical Writing, vol. 5. Oxford UP, Oxford, pp. 13–35.

Lübbe, H., 1977. Geschichtsbegriff und Geschichtsinteresse. Analytik und Pragmatik der Historie. Schwabe and Co, Stuttgart, Basel.

Megill, A., 1985. Prophets of Extremity. Nietzsche, Heidegger, Foucault, Derrida. University of California Press, Berkeley.

Megill, A. (Ed.), 1994. Rethinking Objectivity. Duke University Press, Durham.

Meier, C., Rüsen, J. (Eds.), 1986. Historische Methode. Deutscher Taschenbuch Verlag, Munich.

Meran, J., 1985. Theorien in der Geschichtswissenschaft. Cambridge University Press, Göttingen.

Mink, L., 1987. Historical Understanding. Cornell University Press, Ithaca, NY.

Nagl-Docekal, H., 1982. Die Objektivität der Geschichtswissenschaft. Oldenburg, Vienna and Munich.

Nelson, J., McClosky, D., Megill, A. (Eds.), 1987. The Rhetoric of the Human Sciences: Language and Arguments in Scholarship and Public Affairs. University of Wisconsin Press, Madison.

Novick, P., 1988. "That Noble Dream". The 'Objectivity Question' and the American Historical Profession. Cambridge University Press, Cambridge.

Oexle, O.-G. (Ed.), 1998. Naturwissenschaft, Geisteswissenschaft, Kulturwissenschaft; Einheit – Gegensatz – Komplementarität?. Göttingen.

Rickert, H., 1896–1902. Die Grenzen der naturwissenschaftlichen Begriffsbildung. Eine logische Einleitung in die historischen Wissenschaften, 2 vols. Freiburg.

Ricoeur, P., 1983–1985. Temps et récit, 3 vols. Paris.

Rorty, R., 1986. Philosophy and the Mirror of Nature. Blackwell, Oxford.

Rüsen, J., 1983–1989. Historische Vernunft. Grundzüge einer Historik, 3 vols. Göttingen.

Salmon, W., 1990. Four Decades of Scientific Explanation. University of Minnesota press, Minneapolis, MN.

Scott, J., 1991. Women's history. In: Burke, P. (Ed.), New Perspectives in Historical Writing, pp. 42–66. Polity Press, Cambridge.

Seignobos, Ch, 1897. Introduction Aux Études Historiques. Hachette, Paris.

Skinner, Q. (Ed.), 1985. The Return of Grand Theory in the Human Sciences. Cambridge University Press, Cambridge.

Stern, F. (Ed.), 1970. Varieties of History. From Voltaire to the Present. Macmillan, London.

Veyne, P., 1978. Comment on Écrit L'histoire? Seuil, Paris.

Weber, M., 1973(1922). Gesammelte Aufsätze zur Wissenschaftslehre. Mohr Seibeek, Tübingen.

White, H., 1973. Metahistory. The Historical Imagination in Nineteenth Century Europe. Johns Hopkins University Press, Baltimore.

Windelband, W., 1915. Geschichte und Naturwissenshaft. In: Präludien. Aufsätze und Reden zur Einleitung in die Philosophie und ihrer Geschichte, vol. 2. Mohr, Tübingen, pp. 136–161.

HIV Risk Interventions

JaNelle M Ricks, Ralph J DiClemente, and Puja Seth, Rollins School of Public Health, Emory University, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S.C. Kalichman, volume 10, pp. 6880–6882, © 2001, Elsevier Ltd.

Abstract

Human Immunodeficiency Virus (HIV) infection, the cause of Acquired Immune Deficiency Syndrome (AIDS), has become a significant threat to global public health, faster than any previous epidemic. Recent decades have seen significant progress in the global HIV/AIDS response, largely due to increased availability of treatment and targeted efforts to decrease transmission and incidence. This article highlights HIV risk prevention strategies that have been developed, implemented, and demonstrated efficacy in reducing the risk of HIV infection among adolescents, women, men who have sex with men, and heterosexually active men.

Significant progress has been made in the global human immunodeficiency virus/acquired immune deficiency syndrome (HIV/AIDS) response during recent decades. Since 2001, new cases of HIV have been reduced by more than 50% in 25 of the worst-affected countries (UNAIDS, 2013). In 2012, there were approximately 2.3 million new cases of HIV reported globally, showing a 33% decline in the number of new infections from 3.4 million in 2001. The greatest declines have been observed in sub-Saharan Africa (25% between 2001 and 2011) and the Caribbean (42% between 2001 and 2011). During the same time period, the number of AIDS deaths also declined from 2.3 million in 2005 to 1.6 million in 2012, an accomplishment largely attributable to wider availability of antiretroviral therapy and the continuing decrease in HIV transmission and incidence (UNAIDS, 2012b).

Despite the progress made in the last decade, the burden of this epidemic continues to present a challenge to the global community in expanding and sustaining commitments to combating HIV. The 2013 UNAIDS report estimated that there were approximately 35.3 million people living with an HIV (PLHIV) infection in 2012 (UNAIDS, 2013). Sub-Saharan Africa remains disproportionately affected, and in 2011, 23.5 million PLHIV resided in this region. Southern Africa is the worst affected subregion in the world, accounting for 34% of all PLHIV in 2009 (UNAIDS, 2011a). Women of various global regions comprise half of all adults (15 years and older) living with HIV. This proportion has remained constant over the last 15 years (UNAIDS, 2013).

Across the globe, there is growing recognition that a comprehensive HIV prevention strategy is an optimal approach to combat the HIV/AIDS epidemic, as it integrates diverse biomedical, behavioral, and structural intervention components to prevent HIV transmission. Current strategies for HIV prevention incorporate the following: (1) biomedical interventions, such as treatment as prevention, preexposure prophylaxis (PrEP), and voluntary medical male circumcision (VMMC); (2) behavioral interventions that create demand, improve adherence to risk-reduction mechanisms, and provide psychosocial services to improve acceptance and utilization of essential HIV prevention and treatment services; and (3) structural interventions, such as policy-related changes, sterile syringe, and condom distribution.

This article highlights HIV risk prevention strategies that have been developed, implemented, and demonstrated efficacy in reducing the risk of HIV infection among adolescents, women, men who have sex with men (MSM), and heterosexually active men.

Adolescents and Youth

Impact and Epidemiology

An estimated 5.4 million adolescents and youth in the age-group of 15–24 years were living with HIV in 2012. Ninety-seven percent of new infections occurred in low- and middle-income countries (UNAIDS, 2011b). In 2011, sub-Saharan Africa accounted for approximately 85% (1.8 million) of all adolescents living with HIV (UNAIDS, 2011a). However, in a concentrated epidemic such as the one in the United States, adolescents are responsible for nearly half of all new HIV infections, despite representing only 25% of the total population (Guttmacher Institute, 2006a,b).

Between 2001 and 2011, new adolescent HIV infections decreased globally, although the difference was not statistically significant (UNAIDS, 2012a). The most notable reduction was seen in countries in sub-Saharan Africa and the Caribbean (UNAIDS, 2012c). The decline of new HIV infections among young people worldwide has been occurring simultaneously with behavioral changes such as waiting longer to become sexually active, having fewer partners, and an increased use of condoms with multiple partners (UNAIDS, 2012a). However, many adolescents still do not receive the support that they need to prevent transmission. For example, from 2001 to 2012, areas of the Middle East and North Africa experienced an estimated 50% increase in the number of new HIV infections among young people (UNAIDS, 2012c). Many young people in the age-group of 10–14 years are unaware of their HIV status. Additionally, there is a gender divide, as rates among young women in the age-group of 15–24 years are twice as high as those for young men, accounting for 22% of all new HIV infections and 31% of new infections in sub-Saharan Africa (UNAIDS, 2011b).

Selected Predictors of Risk

Adolescent HIV risk has been associated with a broad range of psychosocial predictors on multiple ecological levels, including individual, relationship, social, and structural. Many of the predictors in this constellation of risk are amenable to modification, an important consideration from a preventive perspective.

At the individual level, variability in perception of risk susceptibility impacts willingness to engage in sexual risk behavior. Similarly, evidence indicates that individual characteristics such as personality traits, psychological states (e.g., impulsivity, low self-esteem, psychological stress, depression), and self-efficacy (e.g., condom negotiation, condom use) play a role. Adolescents' HIV risk behavior also seems to be associated with other risk behaviors such as substance use, delinquency, and pregnancy.

Relationship characteristics also can have a significant influence on adolescents' risk behavior and subsequently their odds of contracting HIV. Although some commonalities in predictors of risk transcend gender, the role of social and intimate relationships creates important differences between male and female adolescents. Females are affected by issues such as partner's age, relationship control, fear of condom negotiation, and intimate partner violence – all factors that have been associated with increased HIV risk. For males, relationship factors more often provide protection against HIV transmission. For example, belief in contraception responsibility, partner's sexual inexperience, and early stages of the relationship have been protective against risky sexual behavior and associated with increased condom use.

On a social level, societal factors have an extensive impact on HIV risk, given their role in shaping cultural norms and influencing behavior. In a supportive social environment, young people can access HIV prevention, treatment, care, and support services. For example, faith-based organizations can play an important role in helping adolescents and youth access services. In Malaysia, the Department of Islamic Development partnered with the Ministry of Health to implement training programs to sensitize religious leaders on HIV-related issues and to reduce HIV-related stigma and discrimination within the communities (Malaysia AIDS Council, 2011). Similarly, among African-American youth, religion is one of the most ubiquitous influences. Involvement in and influence by religion has been shown to be significantly greater for African-American youth than their Caucasian and Latino counterparts. Understandably, religion, particularly religious persecution, has been shown to increase HIV risk among adolescent and young African-American MSM.

Structurally, gender inequalities, poverty, limited access to education, and gender-based discrimination have been shown to also significantly exacerbate risk of HIV infection among adolescent and young women. Widespread gender norms and violence contribute to heightened vulnerability in this population. For example, poverty and aspirations for a better life often motivate young women to engage in transactional sexual relationships, which increase risk of HIV acquisition. Regarding violence, findings of a South African study showed a relationship between intimate partner violence and HIV transmission among young women (Jewkes et al., 2010). Additionally, young women have lower comprehensive knowledge about HIV than young men. In sub-Saharan Africa, only 28% of young women demonstrated comprehensive knowledge compared with 36% for young men (UNAIDS, 2013). Eliminating new HIV infections among children requires ensuring that HIV prevention services reach this population and keep adolescent girls HIV-negative.

Intervention Efficacy

A new analysis featured in a 2013 UNICEF report shows that by increasing investment in high-impact interventions, two million adolescents, particularly girls, could avoid becoming infected by 2020 (UNAIDS, 2013). High-impact interventions include condoms, communications for behavior change (e.g., mass media), and targeted approaches for at-risk and marginalized populations. This prevention approach is also characterized by use of combined scientifically proven, cost-effective, and scalable interventions targeted to the right populations in the right geographic areas. This is in addition to investments in other sectors such as education, social protection and welfare, and strengthening health systems (UNICEF, 2013).

High-impact HIV prevention programs for adolescents have been implemented and evaluated in a multitude of settings, including community settings, schools, clinical sites, detention centers, and substance abuse treatment centers. Sufficient evidence indicates that regardless of the intervention venue, participation in an intervention yields some degree of positive risk reduction (Mullen and Ramirez, 2002; Ross et al., 2006). The most frequently reported outcomes are reduction of unprotected sexual intercourse or increased condom use and decreased frequency of sex. Other observed modifications include delay of sexual initiation, increased sexual knowledge, reduced sexually transmitted infection (STI) incidence, and fewer sexual partners. Although many interventions have effectively reduced sexual risk behaviors, these programs often result in only small to moderate effects and lack long-term sustainability.

Review of youth-based interventions in all countries reveals that tailored interventions for particular groups (e.g., African-American female adolescents, MSM, intrauterine device users) are generally more effective than broad-based interventions, whether implemented in high- or low-prevalence countries, or in generalized or concentrated epidemics. Recognition and responsiveness to the heterogeneity of adolescents allows prevention efforts to target specific behavioral risk characteristics, developmental levels, sexual preferences, and gender differences. Globally, strategic information on adolescents and youth within the HIV response, however, is lacking, particularly from key subpopulations. Therefore, global organizations such as UNICEF have called for a focused effort on collecting strategic information on adolescents who inject drugs or engage in transactional sex, young MSM, street-connected and institutionalized adolescents, and those from disadvantaged socioeconomic or ethnic backgrounds.

Implications for Practice: Improving Future HIV Prevention Efforts for Adolescents

Multiple efficacious HIV prevention interventions exist for adolescents. The 2013 UNAIDS report on HIV, adolescents, and youth calls for an expansion of successful approaches to HIV prevention for adolescents and youth. Preventing HIV in adolescents will depend not only on the development and evaluation of innovative intervention approaches but also on effectively translating and integrating efficacious programs into self-sustained components of clinic practice, school curricula, or community programs. Increasing evidence shows that it is possible to adopt or adapt existing HIV prevention interventions for adolescents in new contexts while also maintaining the core elements, theory, internal logic, and positive outcomes of the original program (McKleroy et al., 2006). However, implementing existing programs in a new community requires attention to cultural and contextual factors as well as specific HIV-related risks of the new target population. Failing to make these considerations may result in interventions that lack relevance, acceptability, and sustainability in the target population (Castro et al., 2004). Therefore, a critical step is implementing tailored interventions that target risk factors for specific groups of adolescents. This will help ensure effectiveness in reducing high-risk sexual behavior and STIs.

Investments should focus on interventions that can be efficiently scaled up and reach those most vulnerable. Such an effort would help extend, reinforce, and amplify intervention strategies and have a greater impact on HIV risk among adolescents.

Women

Impact and Epidemiology

When the HIV/AIDS epidemic was first recognized in the early 1980s, it was gender-structured to primarily involve networks of men. In the 1990s, HIV infection rates among women greatly increased and have remained stable, affecting a growing number of women. Women make up a greater proportion of new diagnoses than men; therefore, the number of HIV infections among this group is increasing in many countries. In 2012, women accounted for half of the 35.3 million global total cases of infection, primarily as a result of heterosexual transmission (UNAIDS, 2013).

HIV/AIDS is the leading cause of death for women in their reproductive years (15–49). Among young people in the age-group of 15–24 years, the HIV prevalence rate for young women is twice that of young men. Additionally, HIV-positive women are at increased risk for STIs, including *Trichomonas vaginalis*, *Chlamydia trachomatis*, and human papillomavirus; cervical dysplasia; and genital cancer. The high rate of HIV infection among women has dramatic consequences for both their own health and public health.

Selected Predictors of Risk

Much of the existing HIV prevention literature portrays women as especially vulnerable to HIV infection because of biological susceptibility. Male–female HIV transmission is twice as efficient as female–male, and multiple female-specific biological characteristics (e.g., sex during menstruation, oral contraceptives use) may enhance the efficiency of heterosexual transmission. HIV risk for women is also deeply intertwined with socially and culturally propelled imbalances of gender and power. For example, violence against women has been associated with HIV infection. Globally, 10–60% of women in the age-group of 15–49 years experience abuse (WHO, 2013). Previous research has indicated that fear of violence undermines the ability of women and girls to negotiate safer sex. It also has been well documented that abuse history is associated with HIV risk behavior such as lower rates of condom use, higher lifetime number of sex partners, earlier initiation of sexual activity, higher rates of drug and alcohol use, and a history of STIs. Moreover, abuse history has been associated with lack of responsiveness to STI and HIV risk-reduction interventions.

Intervention Efficacy

Both male and female condoms have been shown to be effective in preventing HIV transmission. In contrast to male condoms, which require men's willingness to use them, female condoms allow women to exercise power in their sexual decision-making without having to engage in condom negotiation with their male partners. Findings on the acceptability of the female condom, however, have been mixed and they are not used as widely due to several barriers. Therefore, behavioral interventions that focus on delaying sexual debut, decreasing the number of sexual partners and concurrent partners, and increasing safe sex behaviors, including condom use, are often components of several HIV prevention programs.

Reviews of HIV prevention programs in the United States for women have suggested that the most efficacious interventions (1) are guided by social psychological theories; (2) include only women; (3) emphasize gender-related influences, such as gender-based power imbalances, and sexual assertiveness; and (4) are led by peers. Multiple HIV behavioral interventions for women in the United States have been shown to be effective in reducing risky sexual behavior and increasing condom use. However, their impact on HIV transmission and incidence seems to be limited. Additionally, when examining developing countries, behavioral interventions have been shown to be less effective in reducing HIV transmission and risky sexual behavior among women.

Implications for Practice: Improving Future HIV Prevention Efforts for Women

Recent findings on HIV risk interventions for women emphasize the need to focus on both biological and behavioral interventions in order to reduce HIV transmission and incidence. Behavioral interventions are essential to help increase women's HIV knowledge, develop skills, and reduce risky sexual behavior. However, they also can enhance the

effectiveness of biomedical interventions by increasing condom use, HIV testing, and the uptake of other biomedical interventions.

Evidence also suggests that future research needs to more adequately address the contextual conditions impeding women's ability to protect themselves against HIV. Therefore, structural and policy-related interventions are also important. Globally, there is increasing recognition that gender should be addressed more systematically in all HIV/AIDS prevention programs, though progress is still slow. In 2012, 81% of countries reported that they included specific measures for women in national HIV strategies but only 41% allocated specific budget lines for such activities (UNAIDS, 2012b). It is important to realize that the most evolved set of HIV prevention intervention programs are structural interventions that address women's social risk for disease through myriad public health approaches from education, behavioral strategies, and biomedical interventions to policy. Ultimately, interventions for women that ignore the social, economic, cultural, and political realities of their environment will not be optimally effective.

Men Who Have Sex with Men

Impact and Epidemiology

MSM constitute the largest proportion of new infections in most of the developed world. Global prevalence rates, however, may be significantly underestimated, given only 87 countries currently report HIV prevalence among MSM. Data are most sparse for the Middle East and African regions, where criminal sanctions against same-sex sexual behavior can make epidemiological assessment challenging. Among MSM in Asia, the Caribbean, and Latin America, data on newly identified HIV epidemics are beginning to surface (UNAIDS, 2012a,c).

Surveillance data from low- and middle-income countries indicate that MSM are 13–19.3 times more likely to be HIV-infected than the general population. HIV prevalence among MSM ranges from 0 to 40%. Countries that have an HIV prevalence rate above 20% include Bolivia, Jamaica, Mexico, Thailand, Trinidad, and Zambia. In Pattern 1 countries (i.e., industrialized countries with large numbers of reported HIV/AIDS case such as North America, Western Europe, Australia), MSM continue to make up the largest proportion of new AIDS cases each year (UNAIDS, 2012c). In the United States, young MSM accounted for 72% of new infections among 13–24 year olds, and MSM in this age-group represented 30% of new infections among all MSM in 2010 (Centers for Disease Control and Prevention, 2012).

Selected Predictors of Risk

Currently, knowledge of HIV status is low among MSM. Patterns of sexual risk behavior may vary substantially depending on one's knowledge of his HIV status and that of his sexual partner. Evidence indicates that MSM aware both of their own and of their partner's HIV status are less likely to engage in discordant unprotected anal or oral intercourse. In several MSM subgroups, individual HIV testing behavior is impacted by social norms, low perceived risk of HIV infection, fear of a positive test result, and not knowing available testing locations. Additionally, for many MSM, stigma, discrimination, and laws criminalizing sex between men cause reluctance to disclose their sexual activities to health-care workers, thus limiting their access to voluntary HIV testing.

It has been suggested that the rise of the Internet as a facilitator of sexual networking has increasingly contributed to the reemergence of HIV among MSM (Centers for Disease Control and Prevention, 2003). However, evidence of the impact of advances in electronic communication on sexual risk behavior is inconsistent. Although HIV status disclosure is higher among men meeting partners online than those who meet partners in public locations, e-dating increases the ability to meet potential sexual partners, including those who prefer unprotected sex.

Intervention Efficacy

To date, HIV prevention approaches have been insufficient to mitigate the HIV epidemics in MSM. A number of behavioral interventions have been somewhat efficacious in reducing risk behavior among MSM but do not effectively decrease incidence of new HIV cases. Results of a systematic review concluded that studies with strong research designs (e.g., theory-based, measured biological outcomes, included control condition) indicated that individual-level, small group, and community-level behavioral interventions for MSM (e.g., individual counseling, peer education, group counseling, community building) reduced unprotected anal sex by 27% compared with minimal or no intervention. Effectiveness was greatest among interventions that included efforts to promote personal skills such as keeping condoms available and behavioral self-management, indicating merit of including such content in the development and delivery of new interventions for MSM. These interventions were most effective for majority white populations, illuminating the gap in outreach for some high-risk populations that have been critically underserved in intervention research, particularly MSM in countries where English is not the primary language (Jones et al., 2008).

Behavioral interventions are necessary to address HIV in MSM but are insufficient alone. New biomedical modalities such as daily oral PrEP and rectal microbicides have shown promise. For example, PrEP is a method in which individuals who are not HIV-positive take medication to reduce their risk of acquiring HIV. Robust data suggest PrEP's efficacy among MSM. A study by Grant and Lama (2010) demonstrated a 44% reduction in risk of HIV acquisition in this population. Recommendations from the Centers for Disease Control and Prevention indicate that PrEP should be offered to high-risk MSM "at substantial, ongoing, high risk for acquiring HIV infection" (Centers for Disease Control and Prevention, 2011). Little is known, however, about how to measure the duration of risk among MSM. Thus, it remains unclear how to identify MSM at ongoing high risk and how to determine the necessary duration of their PrEP use.

HIV testing programs for MSM are of special concern, given difficulty engaging these men in testing behavior that has been observed in many settings. Studies from Brazil, China,

Ukraine, and Uganda suggest that in some areas, offering community-based and mobile testing at convenient times and in culturally appropriate venues can increase voluntary testing rates. Additionally, 'test-and-treat' programs may be effective for MSM. Data show decreases in HIV incidence among individuals who know their status, as these individuals are less likely to engage in riskier sexual behavior. Early HIV identification through self and partner testing may enhance prevention efforts. Yet, questions remain on how to address the deleterious effect of social stigma on delaying individuals from accessing testing and treatment services.

Use of technology-enabled HIV prevention efforts has presented a unique opportunity to access MSM populations. Evidence indicates that scalability of HIV prevention programs can be enhanced, as online capacities offer the opportunity to increase fidelity of the delivery of program content with minimal incremental costs. Randomized control trials have demonstrated the feasibility of effectively enrolling, surveying, and administering online HIV prevention interventions to subpopulations that are often more difficult to reach through conventional means (e.g., rural) (Bowen et al., 2007; Preston and D'Augelli, 2004). Additionally, online research allows for enrollment of participants from broad geographic areas, increasing the extent to which research results may be generalizable. Modern computing platforms and survey utilities can also enhance measurement of complex behavioral patterns. For example, an Internet-administered questionnaire to collect information about sexual networks reported a more precise breakdown of the timing of sexual relationships compared with interviewer-administered methods (Rosenberg and Sullivan, 2010). Identified limitations of use of new communication technologies and online facilities include concerns with selection bias, ability to consent, and survey completion. Additionally, misclassification issues such as fraudulent data collection (e.g., survey submission by ineligible individuals) may arise where monetary incentives are offered.

Implications for Practice: Improving Future HIV Prevention Efforts for MSM

Reducing HIV incidence among MSM will require a diverse set of interventions, based on the understanding of patterns of spread and local norms. Coordinated behavioral and biomedical approaches could significantly decrease HIV incidence in MSM if delivered to scale. For example, efforts to increase HIV testing and treatment and also optimize PrEP uptake may be enhanced with technology-enabled strategies to achieve scale and strategic use of personally delivered package components. These prevention tools could substantially reduce new HIV infections in MSM; however, their effectiveness will continue to be undercut if structural considerations are not also properly addressed. Stigma, discrimination, and social-level homophobia continue to limit access and uptake of necessary services such as testing, condom availability, and PrEP. Expanding coverage and reaching these men requires structural changes and reform of policy that enable culturally tailored care. Safe and affirming spaces for prevention, treatment, and care must be a focus of future HIV prevention efforts for MSM.

Heterosexual-Active Men

Impact and Epidemiology

Heterosexual behavior constitutes a growing mode of HIV transmission in the United States and is a primary mode of transmission worldwide (UNAIDS, 2013). Although heterosexual men play a key role in HIV transmission to women, as a risk group, they have been and remain virtually invisible in most HIV/AIDS prevention theory, research, and interventions (Exner et al., 1999; Raj and Bowleg, 2012). In many countries, the focus of the epidemic has historically been on women and youth. For example, women are considered to be particularly vulnerable to HIV infection because of sociocultural and biological factors. Models of gender and the influence of male power and privilege also undergird important theoretical frameworks for understanding women's HIV risk. This has directed the focus to women in HIV public health prevention campaigns, leaving heterosexual men to receive considerably less attention in the epidemic.

Selected Predictors of Risk

Societies worldwide frequently position male sexual power as a prime driver of the HIV epidemic. In many contexts, the relational nature of dominant male norms exacerbates the risk of HIV transmission for women. This routine depiction of masculinity paints men as invulnerable and may play a role in preventing men from engaging in safer sexual behaviors and delaying them from seeking HIV preventative care and treatment. For example, masculinity is sometimes defined by having sex with multiple or concurrent partners. 'Traditional' male gender roles are also associated with less consistent condom use, less belief in male responsibility for contraception, negative attitudes toward condoms, abusive responses to condom negotiation, and intimate partner violence – a known correlate of transmission risk among women.

Intervention Efficacy

Behavioral interventions that focus on strategies to decrease risk behaviors among heterosexual men are typically created to support biomedical interventions. For example, there is compelling evidence that VMMC reduces the risk of heterosexually acquired HIV infection in men by approximately 60% (Bailey et al., 2007; Gray et al., 2007). By the end of 2011, VMMC had been performed on more than 1.3 million men in African countries, where this intervention approach is emphasized, nearly doubling the number completed from 2010 to 2011. Randomized controlled trials have shown that when performed by well-trained health professionals in properly equipped settings, male circumcision is safe. Additionally, a systematic review identified an estimated 20% reduction of HIV incidence among women whose male sexual partners had been circumcised (Weiss et al., 2009).

PrEP is an HIV prevention method in which HIV-negative individuals take a daily pill to reduce their risk of becoming infected. Findings from randomized trials have shown that PrEP is effective in protecting heterosexual men against HIV acquisition in diverse geographical settings. Challenges to PrEP have included the cost, potential drug toxicity,

and viral resistance. Additionally, there is evidence of behavioral disinhibition, leading to increased behavioral risk compensation (i.e., increased sexual risk-taking among individuals using known effective HIV prevention methods) among heterosexually active men. For example, a 2013 study discovered a significant increase in unprotected sex with outside partners among heterosexual male PrEP users (Mugwanya et al., 2013). PrEP, however, has been shown to reduce the risk of HIV infection among adult men and is particularly effective when combined and integrated with other HIV prevention methods.

Implications for Practice: Improving Future HIV Prevention Efforts for Heterosexual Men

WHO/UNAIDS recommendations emphasize that VMMC should be considered an efficacious intervention for HIV prevention in countries and regions with heterosexual epidemics, high HIV, and low male circumcision prevalence. Because VMMC provides only partial protection, it should only be one element of a comprehensive HIV prevention package that includes HIV testing and counseling services, promotion of safer sex practices, provision of condoms, and proper sexual health education to support skills and motivation to engage in these strategies. Including a prioritized focus on PrEP for high-risk heterosexually active men can also maximize prevention impact of an integrated approach.

Changing male cultural and socialization patterns will certainly not be an easy task but should be targeted. Focusing HIV prevention efforts on heterosexual male sexual behavior can make a difference in the epidemic among men and their female partners. Future programming that is invested in the intersections between masculinities, gender equality, and health can bolster a thriving culture of human rights and social justice.

Conclusion

Three decades after the first reported cases of AIDS, the HIV/AIDS pandemic continues to cast a tremendous burden. Considering the global scope of human suffering, negative social and economic impact, loss of productivity, and medical care costs related to HIV/AIDS morbidity and mortality, increasing the development, testing, availability, and implementation of efficacious risk-reduction interventions for various populations remains paramount. This article described global HIV risk prevention strategies that have been developed, deployed, and demonstrated efficacy in reducing HIV risk behaviors and infection rates among adolescents, women, MSM, and heterosexually active men.

Sufficient evidence suggests that the necessary tools to substantially alter the HIV epidemic regardless of mode of transmission are available. A growing number of countries are demonstrating increased control of the epidemic through effective prevention and treatment efforts. Additionally, meeting the scientific challenge of a functional cure is promising as the possibility of a vaccine draws closer. Much, however, remains to be done scientifically, programmatically, and socially. Key barriers to knowledge of HIV status and access to prevention, care, and treatment services continue to thrive in the forms of human rights violations, sexual orientation legislation, maltreatment of women and girls, marginalization of children, and stigma and discrimination. To address these issues, there is a continued need for budgetary support, political leadership, and commitment in partnership with affected communities.

In the absence of a highly effective vaccine or cure and with the current state of growing fiscal constraints and infrastructure and limited resources worldwide, maximizing prevention efforts will require utilization of evidence-based approaches for optimal combination of biomedical, behavioral, and structural interventions. When implemented in prevention efforts tailored to both HIV-negative people and PLHIV, combination prevention is a comprehensive strategy to ensure reduction in HIV incidence and strong control of the HIV/AIDS pandemic.

See also: Adolescent Health and Health Behaviors; Globalization and Health; Health Behavior, Psychosocial Theories of; Health Education and Health Promotion; Health Promotion in Schools; Health Risk Perception; Infectious Diseases: Psychosocial Aspects; Medicine and Health for Sexual Minorities; Men's Health; Mortality and the HIV/AIDS Epidemic; Sexual Risk Behaviors; Sexually Transmitted Infections: Social Network Analysis.

Bibliography

Bailey, R.C., Moses, S., Parker, C.B., Agot, K., Maclean, I., Krieger, J.N., Ndinya-Achola, J.O., 2007. Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. Lancet 369 (9562), 643–656. http://dx.doi.org/10.1016/S0140-6736(07)60312-2.

Bowen, A.M., Horvath, K., Williams, M.L., 2007. A randomized control trial of Internet-delivered HIV prevention targeting rural MSM. Health Education Research 22 (1), 120–127. http://dx.doi.org/10.1093/her/cyl057.

Castro, F., Barrera Jr., M., Martinez Jr., C., 2004. The cultural adaptation of prevention interventions: resolving tensions between fidelity and fit. Prevention Science. Retrieved from: http://link.springer.com/article/10.1023/B: PREV.0000013980.12412.cd.

Centers for Disease Control and Prevention, 2003. Internet use and early syphilis infection among men who have sex with men – San Francisco, California, 1999–2003. Morbidity and Mortality Weekly Report 52, 1229–1232.

Centers for Disease Control and Prevention, 2011. Interim guidance: preexposure prophylaxis for the prevention of HIV infection in men who have sex with men. Morbidity and Mortality Weekly Report 60, 65–68.

Centers for Disease Control and Prevention, 2012. Estimated HIV incidence among adults and adolescents in the United States, 2007–2010. HIV Surveillance Supplemental Report 17 (4).

Exner, T., Gardos, P., Seal, D., Ehrhardt, A., 1999. HIV sexual risk reduction interventions with heterosexual men: the forgotten group. AIDS and Behavior. Retrieved from: http://link.springer.com/article/10.1023/A:1025493503255.

Grant, R., Lama, J., 2010. Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. New England Journal of Medicine. Retrieved from: http://www.nejm.org/doi/full/10.1056/Nejmoa1011205.

Gray, R.H., Kigozi, G., Serwadda, D., Makumbi, F., Watya, S., Nalugoda, F., Wawer, M.J., 2007. Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial. Lancet 369 (9562), 657–666. http://dx.doi.org/10.1016/S0140-6736(07)60313-4.

Guttmacher Institute, 2006a. Facts on American Teens' Sexual and Reproductive Health. New York.

Guttmacher Institute, 2006b. Facts on Sexually Transmitted Infections in the United States. New York.

Jewkes, R., Dunkle, K., Nduna, M., Shai, N., 2010. Intimate partner violence, relationship power inequity, and incidence of HIV infection in young women in South Africa: a cohort study. The Lancet. Retrieved from: http://www.sciencedirect.com/science/article/pii/S014067361060548X.

Jones, K.T., Gray, P., Whiteside, Y.O., Wang, T., Bost, D., Dunbar, E., Johnson, W.D., 2008. Evaluation of an HIV prevention intervention adapted for black men who have sex with men. American Journal of Public Health 98 (6), 1043–1050. http://dx.doi.org/10.2105/AJPH.2007.120337.

Malaysia AIDS Council, 2011. HIV & Islam: Responsible Religious Response to HIV and AIDS in Malaysia. Retrieved from: http://www.aidsalliance.org/includes/document/hivislambrief2010_1.pdf.

McKleroy, V.S., Galbraith, J.S., Cummings, B., Jones, P., Harshbarger, C., Collins, C., Deborah Gelaude, J.W.C., 2006. Adapting evidence-based behavioral interventions for new settings and target populations. AIDS Education and Prevention 18 (Suppl.), S59–S73. Retrieved from: http://guilfordjournals.com/doi/abs/10.1521/aeap.2006.18.supp.59.

Mugwanya, K.K., Donnell, D., Celum, C., Thomas, K.K., Ndase, P., Mugo, N., Baeten, J.M., 2013. Sexual behaviour of heterosexual men and women receiving antiretroviral pre-exposure prophylaxis for HIV prevention: a longitudinal analysis. The Lancet Infectious Diseases 13 (12), 1021–1028. http://dx.doi.org/10.1016/S1473-3099(13)70226-3.

Mullen, P., Ramirez, G., 2002. Meta-analysis of the effects of behavioral HIV prevention interventions on the sexual risk behavior of sexually experienced adolescents in controlled studies in the United. Journal of Acquired Immune Deficiency Syndromes. Retrieved from: http://europepmc.org/abstract/MED/12107363.

Preston, D., D'Augelli, A., 2004. The influence of stigma on the sexual risk behavior of rural men who have sex with men. AIDS Education and Prevention. Retrieved from: http://guilfordjournals.com/doi/abs/10.1521/aeap.16.4.291.40401.

Raj, A., Bowleg, L., 2012. Heterosexual risk for HIV among black men in the United States: a call to action against a neglected crisis in black communities. American Journal of Men's Health 6 (3), 178–181. http://dx.doi.org/10.1177/1557988311416496.

Rosenberg, E., Sullivan, P., 2010. A New Online Survey Module to Measure Sexual Partnership Timing with Results from a Focus Group of MSM. Sex::Tech, San Francisco. Retrieved from: http://scholar.google.com/scholar?q=A+new+online+survey+module+to+measure+sexual+partnership+timing%2C+with+results+from+a+focus+group+of+MSM.&btnG=&hl=en&as_sdt=0%2C48#0.

Ross, D., Dick, B., Ferguson, J., 2006. Preventing HIV/AIDS in young people: a systematic review of the evidence from developing countries. UNAIDS Interagency Task Team on Young People. Retrieved from: http://www.popline.org/node/176536.

UNAIDS, 2011a. Global HIV/AIDS Response: Epidemic Update and Health Sector Progress towards Universal Access. Geneva.

UNAIDS, 2011b. World AIDS Day Report 2011. Geneva.

UNAIDS, 2012a. HIV and Outreach Programmes with Men Who Have Sex with Men in the Middle East and North Africa. Geneva.

UNAIDS, 2012b. UNAIDS Report on the Global AIDS Epidemic. Geneva.

UNAIDS, 2012c. World AIDS Day Report 2012. Geneva.

UNAIDS, 2013. Global Report: UNAIDS Report on the Global AIDS Epidemic 2013. Geneva.

UNICEF, 2013. Towards an AIDS-Free Generation Children and AIDS.

Weiss, H., Hankins, C., Dickson, K., 2009. Male circumcision and risk of HIV infection in women: a systematic review and meta-analysis. The Lancet Infectious Diseases. Retrieved from: http://www.sciencedirect.com/science/article/pii/S147330990970235X.

WHO, 2013. Fact Sheet: Violence against Women: Intimate Partner and Sexual Violence against Women. Geneva.

Hochschild, Arlie (1940–)

Anne Cleary, School of Sociology, University College Dublin, Dublin, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Arlie Russell Hochschild is frequently cited as one of the most influential sociologists of the twentieth and twenty-first centuries. This is in part due to her contribution to the sociology of emotions, and she remains a central figure in this field of study. Within this general development lie concepts such as emotional labor and emotion work which have entered the lexicon of sociology and other disciplines and have extended her influence over a wide field. Her work moves thematically through the family, work, relationships, love and caring, and on each of these topics she has contributed unique insights. Underpinning this analysis is a desire to demonstrate the centrality of emotions in social life and how emotions explicate linkages between macro and micro domains. Hochschild's orientation is increasingly that of a public sociologist, and in this way her ideas are accessible to an audience beyond academia. This analysis of her work traces the origins and key elements of her theoretical ideas and considers her contribution to contemporary sociology and society.

Introduction

Arlie Russell Hochschild was born in Boston, Massachusetts, in 1940. Her keen observational powers were apparent from an early age as is evident from the many references to her personal life in her writing. At age 12 years she was transported into a very different cultural environment when her diplomat father was assigned to the American Embassy in Tel Aviv. The family subsequently lived in New Zealand where Hochschild began her college education and – she has noted – her experience of diverse cultures, and of being an outsider within these cultures, was instrumental in moving her toward her future discipline. The diplomatic world of symbolism and performance also provided material for her future study, and she has described being aware from a young age of the distinction between inner and outwardly expressed emotions. By the time she returned to the United States to enter Swarthmore College as an undergraduate, her cultural experiences were far wider than the average American student. The setting for her postgraduate study – the University of California in Berkeley – had a similarly transformative effect on her outlook. She entered Berkeley in the early 1960s and the campus soon became, as she describes herself, "the epicenter of national ferment" (Hochschild, 2013a: 2). The often violent campaign for social and political change, along with the feminist movement, influenced Hochschild's emerging ideas:

> With these movements swirling about me, I set about the quiet, steady work of gathering intellectual tools with which to clearly understand what was going on … and my search began to link the public trends astir at that time to the more private world of the family.
>
> 2013a: 2, 3

In this way Hochschild was attempting to decipher how the family, more generally intimate relationships, would emerge out of this era of change. The survival of family and relationships has been a major seam of interest in her work and is a topic to which she has returned throughout her career – most recently in 2013 with her collection of essays 'So How's the Family?'

Academic Career, Honors, and Awards

Arlie Hochschild graduated from Swathmore College with a degree in International Relations in 1962 and was awarded an MA (1965) and PhD (1969) in sociology from the University of California, Berkeley. She began her academic career in the University of California, Santa Cruz in 1969 and in 1971 she returned to the Department of Sociology in Berkeley as an Assistant Professor. She was appointed a Full Professor there in 1983 and has remained in Berkeley throughout her academic career except for a year spent teaching and researching in India as part of a Fulbright Scholarship and a semester as Lang Visiting Professor in Swarthmore College in 1992.

Hochschild has been the recipient of a Guggenheim Memorial Fellowship (1976–77), a Fulbright Scholarship (1997–98), the Andrew Mellon Foundation Emeritus Award (2007), and the Arensberg Award for lifetime achievement for Research on Work (2006); four awards from the American Sociological Association (ASA) – The Charles Cooley Award (1983) for her book The Managed Heart; The Jessie Bernard Award (2008) for The Second Shift, The Time Bind, and Global Women; the Award for Public Understanding of Sociology (2000), and a Lifetime Achievement Award (2001) from the Sociology of Emotions Section of the ASA. She has received research funding from the Ford Foundation (1990–91) and the Alfred P. Sloan Foundation (1997) – the latter to establish a Center for Working Families in the University of California, Berkeley. Hochschild has been conferred with honorary doctorates by Swarthmore College (1993), The University of Oslo (2000), Aalborg University (2004), The University of Lapland (2012), and Mount St Vincent University in Canada (2013). Her books appear in more than 15 languages, and her works have been the subject of a number of international conferences and seminars – most recently in 2012 when Zeppelin University in Germany hosted the

International Conference in Honour of Arlie Russell Hochschild.

Theoretical Influences

Arlie Hochschild began her sociological studies in a theoretical environment in which the study of emotions was primarily sited within the discipline of psychology. Within sociology emotional life was either ignored, subsumed within other theoretical categories, or left theoretically underdeveloped as in the work of Simmel and Mead. Durkheim, who studied an area (suicide) closely linked to emotions downplayed its personal, emotional, dimension. Parsons, working within the dominant theoretical perspective of the mid-twentieth century, namely structuralism and functionalism, did include emotions but as peripheral rather than as core to his theory of social action. Yet, within the Department of Sociology at Berkeley she found theoretical influences – notably those of Erving Goffman – which were more in harmony with her emerging ideas. Goffman had incorporated emotions into his work – notably in his dramaturgical perspective – but while acknowledging his influence Hochschild has argued that Goffman's conceptualization remains at surface, rather than at deep, level and are situation specific ("His characters had feelings … but we couldn't know, from Goffman, where these feelings came from"(2003: 7)). C. Wright Mills was another early influence, in particular his idea that the personality can be sold in the market, but Hochschild adds the proviso that people 'actively manage feelings' in this transaction (2012a: 229). She also drew on concepts developed by Marx, notably his ideas about alienation, as well as elements from Freud and Darwin, to develop 'A new social theory of emotion' (2012a: 228):

> Drawing on Dewey, Gerth and Mills, and Goffman within the interactional position, I explore what gets 'done to' emotions and how feelings are permeable to what gets done to them. From Darwin, in the organismic tradition, I posit a sense of what is there, impermeable, to be 'done to,' namely, a biologically given sense related to an orientation to action. Finally through Freud, I circle back from the organismic to the interactional tradition, tracing trough an analysis of the signal function of feeling how social factors influence what we expect to and thus what feelings 'signal.'
>
> Hochschild (2012a): 232

Hochschild's starting point is that human emotion and feeling are intrinsically social and that emotions are key to social life ("… it is through emotion that we know the world" (2013a: 4)) and to understanding links between micro and macro domains. As with Goffman, emotions are managed performances by individuals but Hochschild argues that these performances are conducted within cultural constraints and guided by rules (Feeling Rules) that vary according to social class, gender, etc. What people feel and express therefore depends on cultural norms and one's social category or position and these norms are absorbed from childhood within the family and elsewhere. The expression and management of emotion involves conscious effort and the way we do this 'emotion work' can be at the surface or deep level. When internal feelings are manipulated to produce a desired performance this involves emotional labor. If this process moves into the public sphere, and emotions and feelings are regulated by the workplace or market, then emotional labor becomes a commodity and this can have alienating consequences for the individual – a concept Hochschild based on Marx's theory.

Major Works and Publications

The basis of Hochschild's ideas were laid out in a 1979 American Journal of Sociology paper (Hochschild, 1979) but more specifically in The Managed Heart, published in 1983, and subsequent publications indicate the thematic development of her work. The next two books (The Second Shift (1989) and The Time Bind (1997)), focus on another major interest – family relationships and the gender division of labor in households. Her 2002 collection with Barbara Ehrenreich (Global Women) explores how modern family life is connecting to the market for its needs, and two more recent works (The Commercialization of Intimate Life (2003) and The Outsourced Self (2012b)) develop this outsourcing theme. In Hochschild's most recent book (So How's the Family and Other Essays), published in 2013, she spans and reviews her major thematic interests and emphasizes the linkages between them.

In the Managed Heart Hochschild develops the idea that institutions can manipulate emotions based on a study of airline flight attendants and bill collectors. Friendly, empathetic, service is an important selling point for the airline industry (as in most service industries), and this is ensured by selection of suitable personnel and training to fit the person to the job. In this way, Hochschild explains, feeling rules are no longer at the discretion of the worker and emotion work "… comes to belong more to the organisation and less to the self" (Hochschild, 2012a: 198). Some element of this is generally part of working in the service sector but the practice, she argues, becomes more insidious as organizations attempt to control emotions at a deeper level ("they suggest how to imagine and thus how to feel" (2012a: 49). This can result in emotional dissonance or alienation when strains develop between what the employee actually feels and what she/he is expected to present to the public. This aspect of Hochschild's theory has been criticized as overly deterministic, as creating "an illustration of emotionally crippled actors" (Bolton and Boyd, 2003: 289). Bolton and Boyd (2003) suggest that such interactions are two way, that workers have more control over their emotions than Hochschild implies, and that airline attendants can and do resist the demands of management. They argue that forms of emotional self-management vary considerably in the contemporary workplace, and this is supported by research findings indicating more nuanced feeling rules and emotions in present-day work settings (see King, 2012). Similarly, Wouters (1989) rejects the idea that having one's emotions commodified is inherently alienating. Yet, Brook (2009: 9) maintains that the commodification of emotions is acutely alienating and threatening to self-identity and that Hochschild's concept of emotional labor, although narrower in scope, is a valid interpretation of Marx's theory of alienation.

Hochschild has referred to families "as our most precious and emotionally powerful form of mutual commitment" (2013: 4) and the family has been a consistent seam in her work. Her concept of the family is broad; that is, constellations of people in close relationship to each other and the underlying theme is how people are managing love and care in contemporary society. In The Second Shift and The Time Bind she explores how the family is absorbing "… the revolution in the role of women" (Hochschild, 2003: 7) that has resulted from the large-scale entry of women into the labor market. In The Second Shift she describes how couples divide up their family and work responsibilities with women expected to undertake additional physical labor as well as a third 'shift' of emotional work in the home. Men do disproportionately less emotional work in families than women – the result, Hochschild suggests, of their higher status within these relationships. In The Time Bind she again examines the interaction between family and work life but this time she attempts to control for the work environment. The impetus for this study came from a finding in The Second Shift that the sharing of family roles was associated with marital happiness that led her to examine work and home balance in family-friendly companies. In these environments, however, where gender equality was facilitated, family/work tensions were still apparent, and there was evidence of some subjects seeking refuge from the family in the world of work. Although she presented these findings at that time as evidence of the negative impact of the market on family and personal life, it was clear that broader issues were also at play, which relate to the nature and status of caring. According to Hochschild caring work is viewed as low status and "doesn't necessarily bring joy" (2003: 5) – a view she admits holding herself from an early age – and this will remain an issue even if the market accommodates family and work lives. Hochschild has since remarked that combining family and work responsibilities relates to something deeper and is fundamentally about a balance of social and moral worlds (2013a: 276). These deeper themes are developed in her next group of books.

The Commercialization of Intimate Life and The Outsourced Self (and Global Women which she coedited with Barbara Ehrenreich) continue to explore how people are 'doing' love and care in a society in which the majority of parents work outside the home, and marriage and informal systems of care are becoming increasingly precarious. The Commercialization of Intimate Life describes the impact of this confluence of factors on families including the children who are positioned as eavesdroppers in this ongoing saga. However, as Robert Bellah notes on the cover of this book, this is not a celebration of the traditional family but rather an exposure of its shortcomings and inability to cope with the crisis in love and care. Hochschild implicates "the stalled revolution for gender equality" (Hochschild, 2003: 170) in this impasse. Her view is that caring and nurturance have been demoted and marginalized, contrary to the original tenets of feminism, which sought to engage both parents equally in childrearing. Present-day families, unable, and sometimes unwilling, to cope with the demands of both work and care roles, are outsourcing the latter. As society increasingly consumes rather than produces care, she argues, this outsourcing of care has become inevitable.

In The Outsourced Self Hochschild follows the consequences of the commodification of carework and growing pressure from the market to commercialize the self (2012a: 14). Experts are now retained for activities once regarded as ordinary and instinctive and the intimate realm is increasingly influenced by the market. It is a far cry, she suggests, from the 'just do' spirit of spontaneous community cooperation of former times (2012a: 14). Another theme here is the gendered nature of carework in that women are disproportionately employed in areas which require care and emotional labor and how emotional labor is now crossing borders and has a transglobal dimension. Intense emotional labor is required when women cross continents to care for children leaving their own children in the care of others. In the Outsourced Self she describes an extreme expression of the encroachment of the market into intimate life referencing surrogacy clinics in India. In these clinics surrogates produce babies for couples in more affluent regions and Hochschild refers to these centers as a point "on an ever-widening two-lane global highway" that connects poor to rich nations (Hochschild, 2011: 262). Those who run the clinics view surrogacy as a straightforward, mutually beneficial, commercial transaction and appear to be immune to the emotional labor involved as the women suppress feelings in order to detach from the babies. Hochschild notes that "this ideal of the de-personalized pregnancy is eerily reminiscent" of Huxley's novel Brave New World (2011: 265).

These examples of twenty-first century emotional labor connect to Hochschild's theory of emotions and her analysis of caring work, and it also links to broader conceptual themes concerning the colonization of the lifeworld as developed by Habermas (1987). But while Habermas confined this incursion mainly to the public sphere, Hochschild, as Musial (2013: 44) suggests, has demonstrated the extent of its invasiveness into people's intimate lives. In her most recent publications, Hochschild revisits this issue, pointing to the dangers of a market culture that is profoundly changing the nature of emotional attachment, where we feature as consumers rather than creators of love, and asks "What, if anything, is too emotional to sell?" (Hochschild, 2011: 263).

Impact of Hochschild's Contribution

Arlie Hochschild's contribution is multifaceted but within academia her primary influence lies in her theoretical and empirical development of emotional life and more specifically in marking out the Sociology of Emotions as a field of study. In doing so she shifted attention to the social dimensions of emotions demonstrating, in contrast to biological approaches, that emotions are fundamentally social and that social factors infiltrate the entire process and context. Her viewpoint is determinedly sociological although amenable to cross-disciplinary approaches, and while this work is more generally sited within microsociology it has strong links to macro concerns via her studies of institutions. Emotion, as she has said, 'is the heart of what sociology is' (D'Oliveira-Martins, 2014) and provides an explanatory bridge between the self and society which makes this study area central to sociological inquiry. Hochschild's approach can be traced from Simmel and Goffman but, as the introduction to her theoretical development above implies, Hochschild did not merely build on the

nascent ideas of Goffman and others but reached deeper and broader in producing a theoretical framework that made emotional life theoretically and empirically visible. She was not the first sociologist to identify emotions as important but it was Hochschild who delineated this area providing a paradigmatic framework and workable concepts – an achievement Goffman doubted was possible (see D'Oliveira-Martins, 2014). In doing so she initiated a seam of analysis within sociology which has increased in importance over the past three decades, and the influence of her theoretical ideas span many disciplines and arenas. As Bolton and Boyd (2003) note, little has been written in the area of emotion over the past 30 years which does not cite The Managed Heart.

Her most widely disseminated contribution has been her linking of emotions and work, describing how workers carry out emotional labor that can be incorporated by the capitalist labor process. The concept of emotional labor has been mentioned in over half a million publications and cited in 10 000 academic papers and the number of citations has grown exponentially since 2000 (Hochschild, 2013b). The number and variety of these empirical studies attest to the validity and efficacy of these concepts and to the vibrancy of the subject area – as the editors of a recent collection of papers on this theme suggest (Grandey et al., 2013). Anat Rafaeli's (2013) analysis of emotional labor as a social meme of particular societies is an example of the theoretical outreach of this work and implies an analytical program which is simultaneously moving deep and wide. There have been calls for more attention to be paid to core issues (see Garey and Hansen, 2012; Wharton, 2013), but it is clear that these concepts have proven sociological and cross-disciplinary appeal. Moreover, the concept of emotional labor has been adopted by organizations and incorporated into the intellectual arsenal of business gurus. As Hochschild notes, "… capitalism has found a use for emotion management, and so it has organised it more efficiently and pushed it further" (2012a: 186).

Hochschild's theoretical influence is, however, broader than just the analysis of work and labor. At a basic level her framework allows us to understand emotions in social life and she provides diverse empirical evidence to demonstrate the importance of emotions and emotion patterns in society. Wharton, a frequent commentator on her work, has remarked that the main concerns of "Hochschildian sociology" – relationships and families and the encroachment by the market into these intimate realms – signify the importance of her work (Wharton, 2011: 463/464). Ways of belonging and relationships have transformed in contemporary society but the need for attachment and meaning remain key to social and psychic survival. Her ideas, developed in The Second Shift and The Time Bind and now extended to families in the global economy, provide a basis from which to tease out the complexities of present-day emotional attachments and the caring responsibilities that go with them.

Some of Hochschild's views on the family may need re-evaluation to take account of contemporary relationship trends. Similarly, structural issues, especially class, may have a greater impact on emotional expression than her work implies (Simon and Nath, 2004). According to Hochschild, emotional scripts and agendas for men and women differ with women taking on more emotionally expressive duties and men more likely to stay affectively neutral. This scenario does not take sufficient account of changes in family dynamics and gendered emotion work which is emerging from current empirical work. Women are increasingly taking on traditionally male attributes and becoming more economically independent while some men's lives have been transformed by adverse economic conditions (Lively, 2013: 242). Noelle Chesley's (2011) work illustrates how job instability among lower socioeconomic male groupings is driving more equitable family arrangements as men are pushed into more active parenting and household roles. Similarly, analysis of male emotions within the sphere of masculinities theory is deconstructing binary notions of emotional lives and traditional concepts of male emotions. The pressure on some groups of men to suppress their feelings and produce culturally acceptable emotional performances is emerging as a consistent trend and high male suicide rates may be a manifestation of this, representing the externalized cost to society of this repression of emotions (Cleary, 2012). Alongside this is a broader movement assessing work–life balance that is reevaluating conventional, including gendered, ideas of living and working and in which traditional female emotion skills have a high currency (Lively, 2013: 242).

These themes relate to Hochschild's discussion of dominant emotional narratives such as the extolling of the rational, emotion-free voice in western thought and the implications of this for social and political life. Yet, according to Hochschild, we are beginning to value authentic, spontaneous, feeling to prize the 'unmanaged heart,' a trend, she suggests, is "emerging in response to the rise of corporate guile" (2012a: 190/192). At the same time, as intimate life moves into the market language and feeling rules are changing as we become separated from the acts by which we used to say we care. Here Hochschild is connecting to a body of work mapping emotional discourses over time (see, e.g., Elias, 2000; Wouters, 2007). These ideas also reflect those of Giddens (although she rejects a close connection – see Wilson and Lande, 2005) as well as Bourdieu's notion of emotional capital. In a shifting, fragmented world where people struggle with issues of self-identity, her focus on emotion work helps to illuminate the activities of individuals attempting to create meaning. As Hochschild has remarked this puts the study of emotions at the very heart of twenty-first century sociology in tracing links between private troubles and public issues (2013a: 7).

Hochschild's role as a public sociologist is an increasingly important dimension of her activities although this facet has always been present in her work, and she speaks of "a powerful split between two parts of myself – one geared to 'doing' and the other to 'figuring things out'" (2013a: 1). Her narratives of public concern move through the contradictions that continue to exist in work and home, the impact of the care void on children and older people, the extent to which caring work is being outsourced to the market, the downsizing of public services, the growing gap between rich and poor, and the challenges of globalization. The realities of the global care industry are vividly described in her profiles of mothers doing the emotional labor of detaching themselves from their own children to care for others in more affluent parts of the world. Her public role and the communication of her ideas to an audience beyond academia is facilitated by the accessibility of

her published work. Hochschild is the recipient of three New York Times Notable Books of the Year, and Judith Shulevitz described The Outsourced Self as in the genre of popular or public sociology and Hochschild as "eliciting thoughtful reflections from ordinary people" (New York Times/Shulevitz, 2012). Her ability to convey her ideas in sparse and easily understood ways was also the basis of an ASA citation which referred to her "creative genius for framing questions and lines of insight often condensed into paradigm shifting words and phrases" (American Sociological Association, 2013). Another, related, feature is her ability to insert her own life into her work which resonates with feminist thinking. Although some colleagues cautioned against linking personal and intellectual stories, Hochschild believes this subjectivity is key to her observation. Very few theorists place themselves to such an extent within the story and while, at times, this may appear sentimental, even atavistic, in tone (as in her description of summers spent in rural Maine) it hits a more political, and realistic, target in her account of working and parenting as a young academic in the 1970s (Hochschild, 2003). The fact that Hochschild journeyed through an inequitable gender system is clear from her narrative (see Hochschild, 2003: 227–254, Inside the clockwork of male careers). Her struggles were/are not the same as the majority of women grappling with the complexities of combining employment and motherhood but her narratives reach over this divide due to her willingness to go out there and engage with people using her very considerable ability to empathize with her respondents. These personal inserts, along with her engagement with those she writes about, bring a particular explicitness as well as integrity to her work. She has shown herself to be an honest appraiser of feminism and its appropriation by capitalism which "incorporated the idea of equal pay for equal work and diversity but dispensed with any challenge to the priorities of the system women wanted in on" (2003: 254). These themes signal another significant achievement – becoming a senior academic in a sparse female terrain and going further to enter an even rarer sphere – the annals of encyclopedic collections as a major sociological theorist.

Barbara Ehrenreich (2011) has referred to Hochschild as an iconic thinker but she has, as demonstrated above, underpinned the originality of her thinking with a strong theoretical base and workable concepts. She has produced a body of work, including a subdiscipline, which is firmly established within sociology and widely known across other disciplines and beyond academia. This work engages with many of the key issues of our times – primarily how we love and care for each other and is underpinned by data that she continues to collect and mine herself. As a public intellectual she has delivered her message from both small and illustrious platforms using the discipline of sociology as the channel and in this way "she exemplifies the ideals of engaged and informed sociology" (American Sociological Association, 2013). The communication of this message is also aided by another factor that Wharton (2011) has referred to – her personal interest in the work of those who seek to follow in her footsteps. Few academics of her level give such time to mentor interested scholars, doing emotion work herself as she weaves issues from across the world into the emotions paradigm. She once told an interviewer that she likes to take her readers on an intellectual journey, guiding them to explore the important issues of our time (see Wilson and Lande, 2005). Assessing four decades of her very considerable body of work one can conclude that Arlie Russell Hochschild has accomplished this.

See also: Alienation, Sociology of; Culture and Emotion; Dehumanization; Emotion and Expression; Emotional Regulation; Emotions, Sociology of; Family as Institution; Love and Intimacy, Psychology of; Macrosociology-Microsociology; Marriage and the Dual-Career Family; Work and Industry, Sociology of.

Bibliography

American Association of Sociology, 2013. Citation for Jessie Bernard Award. Website: asanet.org/about/awards/Bernard/hochschild.cfm (accessed 14.11.13.).

Bolton, S.C., Boyd, C., 2003. Trolley dolly or skilled emotion manager? Moving on from Hochschild's managed heart. Work, Employment and Society 17 (2), 289–308.

Brook, P., 2009. The alienated heart: Hochschild's 'emotional labour' thesis and the anticapitalist politics of alienation. Capital & Class 33 (2), 7–31.

Chesley, N., 2011. Stay-at home fathers and breadwinning mothers: gender, couple dynamics, and social change. Gender & Society 25 (4), 642–664.

Cleary, A., 2012. Suicidal action, emotional expression, and the performance of masculinities. Social Science & Medicine 74 (4), 498–505.

D'Oliveira-Martins, M., 2014. Emotional Labor around the World: An Interview with Arlie Hochschild. Global Dialogues, Journal of the International Sociological Association 4 (3). http://isa-global-dialogue.net/emotional-labor-around-the-world-an-interview-with-arlie-hochschild/.

Ehrenreich, B., 2011. Foreword. In: Grandey, A., Diefendorff, J.A., Rupp, D. (Eds.), Emotional Labor in the 21st Century: Diverse Perspectives on Emotional Regulation at Work. Routledge, New York & Sussex, UK, pp. xi–xii.

Ehrenreich, B., Hochschild, A.R. (Eds.), 2002. Global Women: Nannies, Maids, and Sex Workers in the New Economy. Metropolitan Press, New York.

Elias, N., 2000. The Civilising Process (E. Dunning, J. Goudsblom, S. Mennell, Trans.). Blackwell Publishing Company, Oxford (First published 1996).

Garey, A.I., Hansen, K.V. (Eds.), 2012. Introduction: an eye on emotion in the study of families and work. At the Heart of Work and Family: Engaging the Ideas of Arlie Hochschild. Rutgers University Press, New Brunswick, NJ, pp. 1–14.

Grandey, A., Diefendorff, J.A., Rupp, D., 2013. Bringing emotional labor into focus: a review of three research lenses. In: Grandey, A., Diefendorff, J.A., Rupp, D. (Eds.), Emotional Labor in the 21st Century: Diverse Perspectives on Emotional Regulation at Work. Routledge, New York & Sussex, UK, pp. 3–27.

Habermas, J., 1987 (T. McCarthy, Trans.). The Theory of Communicative Action, vol. II. Beacon Press, Boston.

Hochschild, A.R., 2013a. So How's the Family? And Other Essays. The University of California Press, Berkeley, CA.

Hochschild, A.R., 2013b. Foreword. In: Grandey, A., Diefendorff, J.A., Rupp, D. (Eds.), Emotional Labor in the 21st Century: Diverse Perspectives on Emotional Regulation at Work. Routledge, New York & Sussex, UK, pp. xiii–xixv.

Hochschild, A.R., 2012a. The Managed Heart: Commercialization of Human Feeling. The University of California Press, Berkeley, CA (First published 1983, reissued with new preface 2012).

Hochschild, A.R., 2012b. The Outsourced Self: Intimate Life in Market Times. Picador/Metropolitan Books, New York.

Hochschild, A.R., 2011. Childhood at the global crossroads. In: Garey, A.I., Hansen, K.V. (Eds.), At the Heart of Work and Family: Engaging the Ideas of Arlie Hochschild. Rutgers University Press, New Brunswick, NJ, pp. 262–268.

Hochschild, A.R., 2003. The Commercialization of Intimate Life. University of California Press, Berkeley, CA.

Hochschild, A.R., 1997. The Time Bind: When Work Becomes Home and Home Becomes Work. Metropolitan, New York.

Hochschild, A.R., 1989. The Second Shift: Working Parents and the Revolution at Home. Viking, New York.

Hochschild, A.R., 1979. Emotion work, feeling rules and social structure. American Journal of Sociology 85 (3), 551–575.

King, D., 2012. It's frustrating! Managing emotional dissonance in aged care work. Australian Journal of Social Issues (Special Issue: Emotions in Social Life and Social Policy) 47 (1), 51–70.

Lively, K.J., 2013. Social and cultural influencers: gender effects on emotional labor at work and home. In: Grandey, A., Diefendorff, J.A., Rupp, D. (Eds.), Emotional Labor in the 21st Century: Diverse Perspectives on Emotional Regulation at Work. Routledge, New York & Sussex, UK, pp. 223–249.

Lloyd Smith, S., Spring 1999. Arlie Hochschild: soft-spoken conversationist of emotions. Soundings 11, 120–127.

Musial, M., 2013. The outsourced self: intimate life and market times by Arlie Russell Hochschild. Theory Culture Society 30 (3), 40–45.

Rafaeli, A., 2013. Reflecting on emotional labor as a social meme. In: Grandey, A., Diefendorff, J.A., Rupp, D. (Eds.), Emotional Labor in the 21st Century: Diverse Perspectives on Emotional Regulation at Work. Routledge, New York & Sussex, UK, pp. 294–299.

Simon, R.W., Nath, L.E., 2004. Gender and emotion in the United States: do men and women differ in self-reports of feelings and expressive behaviour? American Journal Sociology 109 (5), 1137–1176.

Shulevitz, J., May 25, 2012. Paying for Labors of Love (The Outsourced Self, by Arlie Russell Hochschild). New York Times.

Wharton, A.S., 2011. The sociology of arlie hochschild. Work and Occupations 38 (4), 459–464.

Wharton, A.S., 2013. Back to the future. In: Grandey, A., Diefendorff, J.A., Rupp, D. (Eds.), Emotional Labor in the 21st Century: Diverse Perspectives on Emotional Regulation at Work. Routledge, New York & Sussex, UK, pp. 300–305.

Wilson, N.H., Lande, B.J., 2005. Feeling capitalism: a conversation with Arlie Hochschild. Journal of Consumer Culture 5 (3), 275–288.

Wouters, C., 1989. The sociology of emotions and flight attendants: Hochschild's managed heart. Theory, Culture & Society 6 (1), 95–123.

Wouters, C., 2007. Informalisation: Manners and Emotions since 1890. Sage, London.

Holocaust, The

Saul Friedländer, University of California, Los Angeles, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 6887–6891, © 2001, Elsevier Ltd.

Abstract

This article considers the systematic extermination of the Jews of Europe by the Nazis during World War II. It will also allude to the historical background of the extermination and to related contemporary circumstances. A first section of the article deals with some aspects of the historical context, a second one will briefly present the immediate setting of the 'Final Solution to the Jewish Question,' while the third and central part will address the extermination *per se* and the reactions to it.

'The Holocaust' is the term most widely used in reference to the systematic extermination of the Jews of Europe by the Nazis during World War II. Usually, the term also alludes to the historical background of the extermination and to related contemporary circumstances. The first section of this entry will deal with some aspects of the historical background, the second will briefly present the immediate context of the 'Final Solution to the Jewish Question,' while the third and central part will address the extermination per se and the reactions to it. This chronological presentation does not imply an unavoidable passage from one phase to the next. However, while the link between the evolution of German history during the first two decades of the twentieth century and the onset of the Nazi regime was in no way evident, the causal relation between the anti-Jewish policies of the early war period and the extermination was much closer.

The Background

The historical origins of the Holocaust remain an object of ongoing debate. Whereas some scholars consider the extermination of the Jews as the consequence of a specific evolution of German history, at least from the end of the nineteenth century onwards, others interpret the Holocaust as an outgrowth of modernity, the implementation of a pitiless instrumental rationality at the service of a totalitarian racial utopia. Both approaches are not incompatible.

The rise of racial anti-Semitism, the spreading of social Darwinism, and the publication of the first theories of racial anthropology and eugenics turned the late nineteenth century into a breeding ground of ideological trends that were to become central tenets of the Nazi worldview. And, although political anti-Semitism declined in the Wilhelminian Reich at the beginning of the 1900s, an increasingly vocal anti-Jewish hostility spread throughout wide sectors of German society. In this evolution, the outbreak of World War I became a major turning point.

After a brief period of national unity, the unforeseen prolongation of the conflict and the growing political polarization which it induced triggered a radicalization of anti-Jewish attitudes, mainly in the nationalist-conservative camp. The defeat of Germany, the revolutionary upheavals that followed, and the birth of a republic hated and rejected by an important sector of the population, as well as the vast majority of the traditional elites, fanned the already virulent anti-Semitism even further. The Jews were accused of bearing the responsibility for the national humiliation (the 'stab in the back' myth), of controlling the new republic, of instigating bolshevism and, above all, of conspiring to dominate the world, as described in the most widely read anti-Semitic forgery of the century, *The Protocols of the Elders of Zion* (1919).

It was in this postwar atmosphere that Adolf Hitler began his political career by joining the tiny German Workers' Party (DAP) in Munich, in late 1919. In 1920, the DAP became the National Socialist German Workers' Party (NSDAP) with anti-Semitism figuring prominently in its 25 point program. Notwithstanding the failed putsch of 1923 and Hitler's brief incarceration, the resumption of his political agitation in 1925 along with systematic organization, unrelenting propaganda, and, more than anything else, the impact of the Great Depression propelled the party into prominence on the national scene.

Anti-Semitism lay at the center of Hitler's worldview from the outset of his political career and remained so after his accession to power on 30 January 1933. Reduced to its core, this anti-Semitism viewed historical evolution as determined by a merciless racial struggle dominated by the confrontation between Aryan and Jew. The outcome of this struggle was to decide the fate of mankind and particularly of the Aryan race. Hitler's faith had an apocalyptic dimension, presented in terms of 'scientific' theories. But how far this worldview concretely led to the systematic planning of the anti-Semitic policies of the regime, or, obversely, to what extent these measures were the result of uncoordinated initiatives stemming from a growing radicalization of the party and its agencies, remains an open question. One could argue that the ideological guidelines indicated a general framework for action, but that the 'twisted road' followed by the implementation of the anti-Jewish policies was dictated by changing circumstances. In any case, during the 1930s, no plan for a physical extermination of the Jews existed; the ever-more extreme anti-Jewish policies of the prewar years had two successive and identifiable aims: the exclusion of the Jews from the national racial community (the *Volksgemeinschaft*) and, thereafter, their expulsion from the Reich.

From the outset of the new regime the Jews were dismissed from the civil service, from a growing number of professions, and from all major fields of German cultural life. Moreover, all that belonged to or had been influenced by the 'Jewish spirit' had to be cleansed from the German cultural sphere: Freud's or

Einstein's theories, Heine's poetry, Mendelssohn's music, books, works of art, memorial plaques, films, libretti – anything. A stream of anti-Jewish incitement was daily spewed out by all the propaganda channels of party and state, under the guidance of Propaganda Minister Joseph Goebbels and chief party ideologist, Alfred Rosenberg; it found its most rabid expression in *Der Stürmer*, the 'specialized' anti-Jewish organ of the Gauleiter of Franconia, Julius Streicher. The entire education system soon absorbed the anti-Semitic message, as did the Hitler Youth (HJ) and the League of German Girls (BdM). The indoctrinated youngsters of the 1930s were to become the soldiers of the 1940s.

In September 1935, the Nuremberg Laws added strict biological separation to social and cultural segregation by forbidding marriage and sexual relations between Jews and Germans (the Law for the Protection of German Blood and Honor). The Reich Citizenship Law reduced the Jews to second-class citizens. Finally, the first 'Supplementary Decree' to the Reich Citizenship Law defined who was a Jew or, alternatively, a 'mixed breed' of the first or second degree (November 1935). Any transgression of the 'Blood Law' was considered as 'race defilement' (*Rassenschande*), and the offenders (usually the Jewish male offenders) were sent to jail or a concentration camp.

In 1936, the *Reichsführer SS*, Heinrich Himmler, was put in command of all German police forces. This reorganization of the control and terror apparatus led to the appointment of SS General Reinhard Heydrich, previously chief of the security service of the SS (*Sicherheitsdienst* or SD) to the command of both the SD and the newly established Security Police (*Sicherheitspolizei*), which included the Gestapo and the Criminal Police. After the outbreak of the war, these agencies were further centralized within the Reich Security Main Office (*Reichssicherheitshauptamt* or RSHA). While the planning of direct anti-Jewish terror was mainly Himmler's and Heydrich's preserve, the plundering of Jewish assets and enterprises was orchestrated by Hermann Goering's Four Year Plan Administration, and, after the dismissal of the conservative minister of the economy, Hjalmar Schacht (who had managed to protect some measure of Jewish economic life), by his Nazi successor, Walther Funk.

After the annexation of Austria, in March 1938, the drive to compel the Jews to leave the 'Greater German Reich' ('Old Reich' and Austria) became overwhelming. During that year, 'compulsory Aryanization' and, ultimately, the prohibition of any Jewish economic activity destroyed the material basis of Jewish existence in Germany. Notwithstanding the widespread reticence of most countries to take in Jewish refugees (easily identifiable since, at the demand of the Swiss government, their passports were marked with a 'J'), forced emigration and even expulsion over the borders became standard Nazi practice. Finally, on 9 and 10 November, following the killing of a German diplomat by a Jewish youth in Paris, a nationwide eruption of state-organized violence, the so-called 'Night of the Broken Glass,' destroyed the remnants of Jewish life in Germany. The Jews of the Reich now desperately attempted to flee. The panic soon extended to the Jews of the 'Protectorate of Bohemia and Moravia,' after the German occupation of Prague in March 1939. Yet, it was the German attack on Poland, on 1 September 1939 and the outbreak of World War II that entirely changed the fate of European Jewry.

The Immediate Context of the 'Final Solution' (1939–41)

During the first 2 years of the war, the goal of Nazi anti-Jewish policies still appeared uncertain. Jewish emigration from the Reich and occupied Europe remained a possibility, although wartime conditions soon reduced it to a trickle. The goal frequently alluded to was a territorial concentration of Europe's Jews, first in the easternmost part of Poland, in the Lublin area of the newly established 'General Government' (the name used for the nonannexed area of Poland which was put under the authority of the Nazi 'Governor General,' Hans Frank), then, after the victory over France, in the French island colony of Madagascar, and, finally, after the attack on the USSR, in Arctic Russia or Siberia. None of these plans ever materialized, except for some short-lived attempts to deport Jews from Vienna and the Protectorate to the Lublin area, in the fall of 1939 and in early 1941.

A few months after the occupation of Poland, the Nazis started to deport hundreds of thousands of Jews and Poles from territories annexed to the Reich, such as the Warthegau (a part of western Poland), into the General Government; they drove the Jews into ghettos in larger towns and even dumped Jews from the Reich into Vichy France, in the fall of 1940. The deportations within ex-Poland may have been linked in part to the regime's policy of ingathering ethnic Germans from the Baltic countries, the Russian occupied territory of Poland, and the Balkans in the 'Great Germanic Reich.' Thus, the deportations of Jews and of Poles from the Warthegau were meant to make space for the incoming *Volksdeutsche*. The policy was pursued under the orders of Himmler, as newly appointed 'Reich Commissar for the Strengthening of Germandom' (RKF). It was meant to be the first stage of what became known as 'General Plan East' (*General Plan Ost*), aimed at establishing German domination in the East for centuries to come. The enslavement of entire nations and the death of tens of millions of local inhabitants was taken into account. This policy was probably not directly linked to the genesis of the 'Final Solution' since it cannot explain the deportations from the west (to be described later on).

Simultaneously, from October 1939, an entirely separate but not unrelated operation had started in the Reich: the systematic mass murder of the mentally ill, the so-called euthanasia or 'Aktion T-4': personnel were trained and killing methods were developed that were to be used for the mass extermination of Jews after the scaling down of T-4, in the summer of 1941, to a less visible 'euthanasia' operation (Aktion 14f13).

In Western Europe, the much smaller Jewish populations were not secluded (except for the establishment of a semi-ghetto in Amsterdam) or as brutally treated as the Jews of Poland. But their exclusion from public service and the liberal professions, growing curtailment of their economic activity, the registration of persons and property, limitations on travel, and police surveillance turned them into pariahs. Their persecutors were not only the Germans, but also satellite regimes such as Vichy France.

After the German occupation of Yugoslavia and Greece in April 1941, Serbia remained under direct military control and Croatia became a satellite; in both countries, harsh anti-Jewish measures were introduced. Slovakia's anti-Jewish policy

followed strict German orders. Romania implemented fierce anti-Semitic measures of its own, and only Hungary and Bulgaria maintained semi-independent policies in this domain.

In order to facilitate their control over the Jews – beyond territorial concentration – as well as to degrade them, the Nazis compelled all those over the age of 6 to display a 'Jewish star' on their clothing, first in Poland, then, from September 1941 on, in the Reich and the Protectorate, and, in the spring of 1942, in all countries of occupied Western Europe. But the principal instrument of Nazi domination was the 'Jewish Councils.' These councils, composed of representatives of the Jewish communities, received German orders, conveyed them to the Jewish populations, and were responsible, upon pain of death, for their implementation. Moreover, in Eastern Europe the Jewish Councils became agencies of a Jewish self-administration of sorts, struggling against impossible odds; justly or unjustly, they often became targets of sarcasm, resentment, and even hatred in their own communities. Some historians have adopted a highly critical stance toward the councils, considering them as instruments in the hands of the Nazis for the destruction of their own people. Others have questioned this overall characterization as the differences in the behavior of the councils were considerable. All in all, the councils' members were ordinary people confronted with entirely unprecedented circumstances: their own reasoning was incompatible with the murderous strategy of the Nazis.

The Extermination

Systematic mass extermination of Jewish populations began with the German attack on the USSR, on 22 June 1941. Whereas in most newly occupied Soviet territory mainly Jewish men were executed until August 1941, in the Baltic countries and in Transnistria (where the Romanian army was the principal ally of Germany), entire Jewish communities were butchered from the outset. The killing of Jews in party or state positions had been ordered by Heydrich on the eve of the campaign; the perpetration of pogroms by local inhabitants was also explicitly fostered by the Germans. Whether the killing that engulfed the entire Jewish population from August 1941 onwards was independent of any instrumental goals, or whether, as with the starving to death of two million Russian POWs before the spring of 1942 (around 3.3 million died before the end of the war), it derived from difficulties experienced by the Wehrmacht in securing sufficient food supplies for the army is not clear. In any case, the mass murder of Jews on Soviet territory (around 500 000 victims by the end of 1941) was not perpetrated solely by the 'intervention units' of the Security Police and the SD (*Einsatzgruppen*), by local auxiliaries, and by the Romanian army in the south, but also by tens of Reserve Police Battalions and regular Wehrmacht units. In other words, not only the 'troops of the ideological war' (the SS) but ordinary men, possibly ten of thousands of them, were involved, often willingly so, in the mass murder of the Jews.

Few historians would now argue that the decision to exterminate all the Jews of Europe predated the attack on the USSR. Some contend that this decision was taken by Hitler soon after the German onslaught, during the period of 'euphoria,' due to the apparent success of the invasion, and that it found its expression in Goering's letter to Heydrich of 31 July 1941, ordering him 'to take all necessary measures to prepare the Final Solution of the Jewish Question in Europe.' Others argue that such an order was issued in the early fall, when doubts may have risen in Hitler's mind about the prospects of a quick victory in the East. Others still link the order to the USA's entry into the war, in other words to the transformation of the European war into the 'world war,' which, as Hitler had 'prophesied' in his Reichstag speech of 30 January 1939, would see 'the extermination of the Jewish race in Europe.' This last interpretation seems bolstered by documentary evidence. On 18 December, according to Himmler's appointments diary, the SS chief met with Hitler and wrote down: 'Jewish question/ eradicate as partisans.' The reference to 'partisans' is clearly a way of describing an enemy lurking behind the lines, an internal foe, who, just as in 1917/18, would stab Germany in the back if given the chance. Finally, some historians doubt the existence of any clear order given at the highest level and consider total extermination as resulting from the cumulative effects of local initiatives.

By October 1941, the killings had extended to Jews from the Greater Reich and the Protectorate who were deported mainly to Lodz, Riga, Kovno, and Minsk. In Riga and Kovno, thousands were shot on arrival. The others were exterminated later with the inhabitants of their ghettos. They had also spread to Serbia, where the Wehrmacht executed some 6000 Jews in reprisal for partisan attacks. During the same month, Jews were forbidden to leave the continent, 'in view of the coming final solution of the Jewish Question' and the construction of the Belzec extermination camp (possibly that of Sobibor too) began in the General Government.

In early December, the first extermination site became operational at Chelmno, near Lodz, principally for the murder of Jews from the Warthegau. On 20 January 1942, in a meeting in Berlin (*am Grossen Wannsee*), Heydrich informed an array of state secretaries, senior SS officers, and other selected officials of the general extermination program that, according to his computations, was to apply to 11 million European Jews. From then on total extermination became a problem of logistics and of some measure of haggling between the Germans and various governments or local authorities about handing over their Jews. In almost all cases, foreign Jews were arrested and delivered first, local Jews followed (as will be indicated further on, in some rare cases Jews were not handed over at all).

The arrests of Jews, mainly in Western Europe, and the logistics of the deportations were coordinated between the Higher SS and Police Leaders (HSSPF), Himmler's delegates to each occupied country or major area, local authorities and police forces, and the Jewish Affairs section at the RSHA (IVB4) under Adolf Eichmann. Furthermore, the deportations demanded joint planning with the Reich Railways Authority (*Reichsbahn*) and its territorial subdivisions (*Ostbahn*, etc.). Once the Jews reached the camps, they passed under the command of the SS Economic and Administrative Main Office (*SS Wirtschafts- und Verwaltungshauptamt*) under SS General Oswald Pohl (except for the *Aktion Reinhard* camps). In the camps, the Jews were destined for extermination; however, mainly in Auschwitz, selections separated the healthiest and the strongest from the mass of the inmates, submitting them to slave labor before sending them to their death.

The first stage of the extermination chiefly included the Jews of the Warthegau and of the General Government who, through 1942, were killed in the *Aktion Reinhard* camps (Belzec, Sobibor, Treblinka, and Majdanek), as well as in Chelmno. The deportation of the Jews of Warsaw to Treblinka began in July 1942. Simultaneously, deportations from Western Europe (France, Belgium, and Holland) started. By then, the principal Auschwitz extermination camp (Auschwitz II or Birkenau) had become fully operational. The killing method shifted increasingly from the use of carbon monoxide gas to that of Zyklon B, a powerful pesticide utilized in the fumigation of army barracks and ships' holds. Before they were killed, the victims' valuables and clothes were taken away and their hair was shorn; after the gasings, gold teeth, and artificial limbs were removed from the corpses. The German economy recycled the 'booty': gold was smelted and delivered to the *Reichsbank;* hair was used for mattresses, slippers, etc.; clothes were distributed by the *Winterhilfe* (the German winter-relief organization). Homes, furniture, and other belongings left behind by the deportees were distributed to deserving *Volksgenossen* or acquired, usually at dirt-cheap prices. This allocation of spoils spread to occupied countries in the West. Particularly valuable objects, mainly art, looted all over Europe by the *Einsatzstab Reichsleiter Rosenberg* or ERR, were offered to Nazi leaders (Hitler and Goering headed the list), to German museums, and various other German institutions.

In 1943, the last batch of German Jews was deported to the extermination camps; the deportations from Western Europe continued, and, simultaneously, the Jews of the Balkans and, by the end of the year, of Italy (following the German occupation of the greater part of the peninsula) were included in the killing process. From March 1944 onward, after the occupation of Hungary by German troops, this country's Jews became the last large national group of victims (over 400 000) to be sent to Auschwitz-Birkenau. On 2 November 1944, Himmler ordered the cessation of the mass exterminations; his order was frequently disregarded, even after the liberation of a growing number of camps by the Allies. 'Death marches,' organized by camp commanders or SS personnel, went on throughout the winter of 1944–45 and practically until the German surrender in May 1945. By the time the war in Europe drew to a close, approximately 5.5 million Jews had been exterminated.

The exterminations on Soviet territory had become widely known by the fall of 1941. Information about the mass killing of European Jewry beyond the Soviet border spread throughout the continent, and in Allied and neutral countries from the second half of 1942 onward, although the technical details of the 'Final Solution' may not have been widely known until 1944. The overall reaction was one of indifference and inaction, sometimes of outright collaboration. While Vichy authorities and police collaborated with the Germans in the arrests and deportations, Switzerland hermetically closed its borders to the Jews who attempted to flee. In general terms, the Christian churches intervened feebly, mainly on behalf of converted Jews. Thousands of Jewish children were hidden in Christian institutions; in most cases, their lot was conversion. The silence of Pope Pius XII, notwithstanding some help from the Vatican in hiding Jews in the fall of 1943, has, to this day, remained a matter of heated debate.

In some cases, the reactions were different. Spain and Portugal allowed the transit of Jews through their territory. Sweden attempted to take in the Jews of Norway when the deportations started and it took in almost all the Jews of Denmark (most of whom were transported to safety by the Danes themselves). Protests against the anti-Jewish measures were staged in Holland, in early 1941; the Italians tried to protect the Jews in their own country as well as those under their control. Bulgaria refused to deliver its own Jewish citizens, although it did deliver the foreign Jews. Finland rejected all German demands regarding its small Jewish population.

'The abandonment of the Jews' by the Western Allies is largely undisputed nowadays. Rejected or fatally delayed rescue plans involved transfers of ransom money to the Germans or their allies, or the bombing, in 1944, of the Auschwitz gas chambers and of the railway line connecting the camp with Hungary. Such refusal to help, whether known or merely sensed, could but contribute to the feeling of total isolation of the Jews of Europe in the face of the German extermination process.

The concrete possibilities of Jewish resistance were almost nil. Moreover, by the time most European resistance movements became active, the bulk of European Jewry had been exterminated. Hence Jewish armed revolt in the ghettos (the best-known instance being the Warsaw ghetto revolt of April 1943) or even in the camps, although hopeless, carried the utmost symbolic significance. As for the immense majority of European Jews, they had practically no options even when impending murder was obvious: flight and hiding were limited options; despair and numbness, overwhelming.

See also: Antisemitism; Collective Memory, Psychology of; Contemporary History; Dictatorship in History; Ethnic Cleansing, History of; Genocide, Historical Aspects of; Historical Thought and Historiography: Jewish Tradition; Judaism and Gender; National Socialism and Fascism; Racism, History of; Second World War, The; Violence, History of; War Crimes Tribunals, Crimes against Humanity and Transitional Justice.

Bibliography

Aly, G., 1995. Endlösung. S. Fisher, Frankfurt am Main, Germany.
Bankier, D., 1992. The Germans and the Final Solution. Oxford University Press, Oxford, UK.
Bauer, Y., 1994. Jews for Sale? Yale University Press, New Haven, CT.
Browning, C.R., 1992. Ordinary Men, first ed. HarperCollins, New York.
Burrin, P., 1994. Hitler and the Jews. Edward Arnold, New York.
Friedländer, S., 1997. Nazi Germany and the Jews, first ed. HarperCollins, New York.
Gerlach, C., 1997. Kalkulierte Morde. Hamburg, Germany.
Gutman, Y., 1982. The Jews of Warsaw 1939–1943. Indiana University Press, Indiana, IN.
Herbert, U. (Ed.), 1998. Nationalsozialistische Vernichtungspolitik. Frankfurt Mainam, Germany.
Heer, H., Naumann, K. (Eds.), 1995. Vernichtungskrieg. Hamburg, Germany.
Hilberg, R., 1993. Die Vernichtung der Europäischen Juden, vols 3. Fischer. Frankfurt, Germany.
Longerich, P., 1998. Politik der Vernichtung. Piper, Munich, Germany.
Marrus, M.R., Paxton, R.O., 1981. Vichy France and the Jews. Basic Books, New York.
Mommsen, H., 1991. The realization of the unthinkable. In: Mommsen (Ed.), From Weimar to Auschwitz. Princeton University Press, Princeton, NJ.
Trunk, I., 1972. Judenrat. Macmillan, New York.
Wyman, D.S., 1984. The Abandonment of the Jews, first ed. Pantheon Books, New York.

Home, Anthropology of

Thomas Hauschild, Eberhard-Karls University, Tübingen, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 6891–6895, © 2001, Elsevier Ltd.

Abstract

Discussions on the human 'home' and related categories always oscillates between emotionally charged categories of belonging and more practical reasonings about the location of persons and groups in the world – one argument today being represented by the 'invention of tradition' approach, the other recently tending towards sociobiological and ethological claims about the existence of a human instinct which orientates us towards total 'environs' or natural habitats. This article enquires into very different ways of constructing material and spiritual homes, trying to show that the striving for a 'home' might be a human universal, but that, on the other side, the ideal is never totally realized, as human fragility and dispersal drives us always and again away from our homes. In the era of 'moving targets,' anthropological fieldwork has partly lost its central referent: the small group presumably inhabiting a restricted area since immemorial times. However, we should acknowledge that humans all over the world are struggling to recreate homes under the impact of growing mobility and globalization. The anthropological field therefore re-emerges over again, sometimes 'under fire,' sometimes through more peaceful visits of anthropologists to other people's homes.

Definition and Meaning of 'Home'

The term 'home,' being of Anglo-Saxon and Germanic origins, often carries rational and irrational connotations at the same time: house, landscape, habitat versus place of belonging, object of nostalgia, place of origin. The German equivalents, *Heim* and *Heimat,* originally meant 'house' and 'owned land,' but later on denoted ideal and material spaces offering homeliness or German *Gemütlichkeit.* Not coincidentally, the term *Heimat* was re-introduced into the German written language through the reception of British gothic ideas and romanticism. Moreover, *Heimat* was adapted by many other languages as a lean word signifying a mixture of connotations all tending towards an emotionalized interpretation of the place a person feels adapted to: 'my home is my castle.' While loaded with emotional charge, the term 'home' usually also carries allusions to very profane and unromantic connotations, such as in the euphemistic names given to institutions catering for the poor, old, and disabled, or in the meanings expressed by the untranslatable German term *unheimlich* (the 'uncanny' of Freudian philosophy).

Anthropologically relevant references to 'home' initially appeared during the nineteenth century in Anglo-Saxon and German romantic and gothic styles of writing in which it denoted *Ur-Heimaten* or certain groups of humankind, such as the 'Aryan' origins in Asia or people's origins in other 'cradles' (Kuper 1988). Later on the term 'home' acquired a more specific political meaning as in the case of the 'homelands' assigned by the South African apartheid regime to ethnicized groups of 'black' citizens, which had the purpose of containing them in reservations or similar constructions of 'home rule.' More recently, the term has gained some significance again in anthropological studies of migrant, diaspora, and fugitive populations (Malkki 1995; Vertovec and Cohen 1999).

The related discussions move between two poles, one pole being represented by the 'invention of tradition' approach (Hobsbawm and Ranger 1982), and the other pole tending towards sociobiological and ethological claims about the existence of a human instinct which orientates us towards distinct 'environs' or natural habitats (Wilson 1975). From an anthropological perspective, the often rather unreflected and untheorized use of the term 'home' in scientific texts offers a good opportunity for checking the process of translation between European and non-European cultures. For the discussion of universalistic assumptions about a human drive for being 'at home' involves very heterogeneous forms of belonging, alienation, or indifference towards a 'home.' Often the 'drive' to be at home seems quite easily contented with illusions of a spiritual home; yet in some cases historical processes lead to violent conflicts about presumed 'home countries' between certain groups.

Mary Douglas (1966) introduced anthropologists to a world of symbolic attributions, in which the body can signify everything, and everything can mean the body. The same applies to homes, when looked at as a rational construction of territorial belonging on the one hand, and as an image of embeddedness in a particular social and geographical landscape, or nostalgia about such a situation, on the other.

Ways of Being 'At Home'

Obviously, the portable shelter constructed by a group of Kung hunters and gatherers is not the same as the proverbial British or German 'castle.' Anthropologists observed, however, that so-called bush people in Africa, traditional Australian aborigines, as well as groups inhabiting the rainforests of South Asia and South America, have developed elaborate systems of navigation in their regions, based for example on landmarks or tactile qualities of the ground (Widlok 1997). In the nineteenth century, Australian aborigines imagined their ancestors to wander around and with their actions invent the current topographical shape of their habitat. In so doing they related to the 'ground' in a way not at all alien to European conceptions of identity defined by geographical, social, genealogical, and individual physical characteristics of a given group in a given area (Spencer and

Gillen 1938). Attracted by some of the benefits of provincial town life or driven away by colonial powers from their places, their twentieth-century descendants may insist on returning to their promised lands, if their means of living allow such a step (Berndt 1977). Welfare programs can thus paradoxically lead to a partial return to traditional ways of living and especially of nurturing their ancestors' grounds. However, often we are no longer able to discern whether indigenous nativists adhere to old projections of a social and spiritual life onto certain 'hunting grounds' and places of survival through gathering activities, or whether this just follows Western concepts of belonging, grounded and capitalizing on geographical features of a given area.

There is varied ethnographic and mythological evidence of responses to loss of old nurturing grounds, including the wish to re-enchant the relationship between humans, heaven, and nature, although so far there is not enough comparative research available on the symbolism of land and house. Yet there is some evidence that human groups that succeed in storing greater amounts of food develop more closely-knit local mythologies bound up with the topography of landmarks, human dwellings, places for growing and harvesting food, as well as holy places. Horticulturists and cattle-breeding populations living in segmentary societies or low-grade forms of centralization tend to organize the household, the surrounding area, and faraway places that are of interest to the group by socially constructing ancestors and territorial spirits or divinities (Lienhardt 1961, Levi-Strauss 1955). The 'juridic domain' (Fortes 1969) concerning the simple facts of life – dwelling together, dividing up food, working a territory, keeping a herd, and maintaining property rights – is thus organized by an explicit ideology of belonging, descent, holy grounds, and the profane (Kramer and Kramer-Marx 1993). A supernatural interpretation of shrines and natural features, even of artificially organized biotopes, can be the territorial expression of a temporal cult of vegetation which unites social groups in a harmonious view of the ego, the household, the family, and other related persons with its ground, including all the plants and animals living or to be raised therein (Turner 1969).

A 'system of total exchange' (Mauss 1960), of generalized reciprocity, seems to pervade and bind human groups and their natural surroundings together. Paradoxically, this can privilege those groups which have reasons to claim to have been the first humans in a particular region and can give them a superior status, heavily defended in conflict, or lost and exchanged with the status of a pariah. The conflict between Hutu and Tutsi in Central Africa has had such a component (Maquet 1961). In a time of international arms traffic, and armed interventions into local cultures, tensions of this kind can lead to severe cases of mass violence.

There is a definite difference between such closely knit groups or networks of groups and the nation states of world society, in so far as the 'total social fact' (Mauss 1960) of traditional societies is organized into a tight overlapping of the domains of landscape, social structure, cosmology, etc. (Fortes 1969). The father or uncle can be the earthmaster, celebrating a holy shrine, and the next moment he can be the leader of a group of young warriors. This does not mean, though, that the social reality of these groups is more homogeneous than ours. Whilst subscribing, like their fellows, to a certain degree to the governing harmonious ideology of space, time, and home, some outsiders, skeptics, and dissidents always disagree and hold rational or disenchanted views of their group's attempts to maintain a homeland (Radin 1957).

By exchanging 'stigma' for 'charisma,' such outsiders can strongly affect or change their society. Thus, the classical 'big men' of New Guinea who constructed the 'pig festival' through the vegetational and reproductive cycles draw on a whole system of credits and cliental relations countering the heroic culture of the 'great men' who search for the spiritual gratification of the warrior (Godelier 1982). Early imperialist states seem to have been based on a similar mixture of exploiting nature, fellow, and other groups, on the one hand, and heroic fantasies about extending a holy landscape to other regions, on the other (Wolf 1999).

The study of nomadism might be particularly instructive with respect to human ideas about finding or having a home (Mauss 1960). For the layperson, nomads just seem to wander around without clearly relating to certain features of the landscapes. The experience of wandering and fugitive groups in Europe, such as the Sinti, Roma, and many others, may support this conception of a 'wandering' part of humanity that does not feel bound to certain cycles of exchange. Wanderers are always suspect to sedentary populations, and this resentment can remain through long periods of sedentarization, if it is supported by political powers. One of the starting points of German anthropology in the nineteenth century was a series of heated discussions about whether or not Jews felt any obligation of fairness and civil behavior towards non-Hebrew populations (Gilman 1990).

In the anthropological data, nomadic constructions of a unified horizon of landscapes abound, as does information on nomads treating their tents, carriages, or cars as 'mobile homes' (Rao 1987). Motifs in Oriental carpets, manufactured by nomads, often appear to carry the signum of descent and territorial rights as well as some kind of report of the group's history and historical claims for grounds and herds (Schuster and Carpenter 1996).

Not only nomadic populations but many other groups convey an ideal of homogeneity, tradition, and *mise en scene* of an environment of 'being at home' through their material culture, such as dress and costumes. This continues right up to European medieval, Ottoman, and today's international regulations of clothing and corporeal signs. Also the often painfully achieved imprints on the skins of members of more marginal populations seem to have the same rational and spiritual aim of demonstrating a form of belonging together (Clastres 1973). It is the invention of an 'environment,' in the sense that this term is used in modern art. Human environmental creativity ascribes to a certain space the property of being a home. These autopoietic processes of representing and creating an order of time, space, and body are perhaps most obvious in rural societies with their regional and class-bound costumes, their festivals, shrines, and clear-cut regional polities of power.

But even in the case of feudal societies, which often enough did not allow their underclass members to leave their 'homes' – even in this continuum of slavery on the one hand and enthusiasm for one's own 'ground' on the other – we can find distinct groups who, through historical experience or the

subtleties of property rights, seem to have acquired a very distanciated attitude towards their houses, fields, and cattle (Lopreato 1967). In times of famine or collective violence, these groups were probably among the first to leave their grounds and wander to the cities or the promised lands, over the mountains and across seas, as soon as the opportunity was offered by the modernization of the world economy into the global capitalistic project.

Modern and Postmodern Homes

In cities and other modernized landscapes, anthropologists have observed distinct techniques deployed in designing the 'home' of migrated groups or former peasant inhabitants of the area (Hannerz 1980). Conscious city design, the building of ghettos for reasons of controlling the 'bio-power,' and the process of identity formation in subgroups of a given population all go hand in hand here. Homeland associations as well as informal networks amongst industrial workers and employees ascribe to their private and sometimes also their professional life characteristics of tribal life. The elite of industrialized areas tend to form closely-knit networks, sometimes secret, sometimes formally recognizable, which transform their actors into something reminiscent of the earthmasters and segmentary war chiefs as encountered in traditional societies.

In the third millennium, two-thirds of the world population are expected to live such a metropolitan way of life, and from what we know about the older metropolitan areas in the world, we can deduce that the postmodern metropolis will be partly made by neighborhoods defining and defending themselves like traditional 'villages in the city,' like those in Naples, Cairo, Rio de Janeiro, or New York. These 'lieux' and 'milieux' of trust and similarity will be defined by criteria of class and taste as well as ethnicity and religious confession. Homelessness will abound as well as those of the better off desperate to (re)create a home as an anchor to a very mobile style of life (Dehavenon 1999). The inhabitants of big cities try to overcome the anomic atmosphere of 'non-lieux' by staging and topographically creating city life through festivals and customs, charity and sport events, as well as a home life through decoration, invitations of friends, meals, and so on (Augé 1992; Cieraad 1999).

In addition to the withdrawals and the small refuges of identity, however, an extension of the word 'home' has become important. With the advent of the nation states, millions and even billions of humans can unite in very large conglomerations of landscapes, and by means of national symbolism they try to find not only a common political but also symbolic denominator. An organicist idea of the state – 'a sound family makes a sound state' (Norman 1991) – may help to transfer feelings which were originally connected with house, family, regional polity, and local religion to the higher spheres of a fatherland united in the sign of a mother language, a home country (Gellner 1998).

This is where ethnicizing or nationalizing politics have often overlapped with anthropological science, a science often intended historically as a kind of school for learning how to invent nations. Not only under Hitler and Stalin did the 'nations of anthropology' (Borneman 1995) stand in close conjunction with their host nations, so did American cultural anthropology and other well-respected disciplines.

Today's 'United Nations' comprise groups that are divided by very different models of inventing traditions as well as interior constructions. Some of them are founded on the basis of the one and only territory alone, others claim that a common language, cultural heritage, race, or historical traumatization provide the fundament of their existence. However, the whole system of nations nowadays demands that no single person, house, or territory can be without a host nation. This is supported by older ethnographic constructions, which always attributed a group of people to a certain territory. Today, such origins of ethnic topographies frequently turn into a topography of terror, while at the same time international movements of refugees and migration make such attributions obsolete. Another form of homebuilding, however, will apparently continue to exist even in the most technologically advanced systems on Earth. Against all the prognostics of the social sciences, the 'home' seems to survive even its most important carrier, the family. Statistical evidence indicates that single households abound in all metropolitan areas of the world, yet this 'yuppie' culture maintains that people ought to belong to some house, some household (often lovingly decorated and furnished), a certain area, a group of friends and colleagues. Big festivals like Christmas – which is, indeed, the only feast of a truly international character celebrated today, not only in the West but in most parts of the world including China and Africa – provide this 'home' with an anchor in time and space. Groups of families and friends reunite during these days to have traditional forms of communality, exchange, and mythopoesis about a bearded old man leaving his occult home in the last area on the world that still proves unreachable for globalization, the polar circle, carrying with him gifts and goods that will help to make people feel more at home (Miller 1993).

Final Thoughts

In the era of 'moving targets' (Marcus 1999), anthropological fieldwork has partly lost its central referent: the small group inhabiting a restricted area since immemorial times – the field. However, it seems that humans all over the world are struggling to recreate homes under the circumstances created by today's form of globalization. The anthropological 'field' therefore re-emerges over and over again, sometimes 'under fire' (Nordstrom 1995) and sometimes in a peaceful dialogue with the anthropologist, that 'professional stranger' who leaves his home to study others' homes.

See also: Collective Beliefs, Sociological Explanation of; Collective Memory, Anthropology of; Community Studies: Anthropological; Community and Society: History of the Concepts; Community, Social Context of: The US Case; Everyday Life, Anthropology of; Family as Institution; Family: The Anthropology of the Concept and Its History; Folklore; Households in Anthropology; Kinship in Anthropology; Nationalism, Sociology of; Place in Geography; Tradition, Anthropology of.

Bibliography

Augé, M., 1992. Non-Lieux. Introduction à une anthropologie de la surmodernité. Éditions du Seuil, Paris.

Berndt, R., 1977. Aborigines and Change: Australia in the 70s. Australian Institute of Aboriginal Studies. Canberra, ACT.

Borneman, J., 1995. American anthropology as foreign policy. American Anthropology 97, 663–672.

Cieraad, I. (Ed.), 1999. At Home: An Anthropology of Domestic Space. Syracuse University Press, New York.

Clastres, P., 1973. De la torture dans les sociétés primitives. L'Homme 13, 114–120.

Dehavenon, A.L., 1999. There's No Place Like Home. Anthropological Perspectives on Housing and Homelessness in the United States. Westport, CT, Bergin and Garvey.

Douglas, M., 1966. Purity and Danger: An Analysis of Concepts of Pollution and Taboo. Routledge and Kegan Paul, London.

Fortes, M., 1969. Kinship and the Social Order. University of Chicago Press, Chicago.

Gellner, E., 1998. Nationalism. New York University Press, New York.

Gilman, S., 1990. Jewish Self-Hatred: Anti-Semitism and the Hidden Language of Jews. Johns Hopkins University Press, Baltimore, MD.

Godelier, M., 1982. La Production des Grands Hommes. Fayard, Paris.

Hannerz, U., 1980. Exploring the City. Columbia University Press, New York.

Hobsbawm, E., Ranger, T. (Eds.), 1992. The Invention of Tradition. Cambridge University Press, Cambridge, UK.

Kramer, F.W., Kramer-Marx, G., 1993. Zeitmarken. Die Feste von Dimodonko. Trickster, Munich, Germany.

Kuper, A., 1988. The Invention of Primitive Society: Transformations of an Illusion. Routledge, London.

Lévi-Strauss, C., 1955. Tristes Tropiques. Plon, Paris.

Lienhardt, G., 1961. Divinity and Experience. The Religion of the Dinka. Oxford University Press, Oxford, UK.

Lopreato, J., 1967. Peasants No More: Social Class and Social Change in an Underdeveloped Society. Chandler, Scranton, PA.

Maquet, J., 1961. The Premise of Inequality in Ruanda. Routledge, London.

Marcus, G. (Ed.), 1983. Elites. Ethnographic Issues. University of New Mexico Press, Albuquerque, NM.

Marcus, G., 1999. Ethnography Through Thick and Thin. Princeton University Press, Princeton, NJ.

Malkki, L., 1995. Purity and Exile. Violence, Memory and National Cosmology among Hutu Refugees in Tanzania. University of Chicago Press, Chicago.

Mauss, M., 1960. Essai sur le Don. In: Mauss, M. (Ed.), Sociologie et Anthropologie. Presses Universitaires de France, Paris.

Miller, D. (Ed.), 1993. Unwrapping Christmas. Clarendon Press, Oxford, UK.

Nordstrom, C. (Ed.), 1995. Fieldwork Under Fire. Comparative Studies of Violence and Survival. University of California Press, Berkeley, CA.

Norman, K., 1991. A sound family makes a sound state. Ideology and upbringing in a German village. Stockholm Studies in Social Anthropology, 24. Almqvist and Wiksell International, Stockholm.

Radin, P., 1957. Primitive Man as Philosopher. Dover, New York.

Rao, A., 1987. The Other Nomads: Peripatetic Minorities in Cross-Cultural Perspective. Böhlau, Vienna.

Schuster, C., Carpenter, E., 1996. Patterns that Connect. Social Symbolism in Ancient and Tribal Art. Abrams, New York.

Spencer, B., Gillen, F.J., 1938. The Native Tribes of Central Australia. Routledge, London.

Turner, V., 1969. The Ritual Process. Structure and Antistructure. Routledge and Kegan Paul, London.

Vertovec, S., Cohen, R. (Eds.), 1999. Migration, Diaspora and Transnationalism. Elgar, Cheltenham, UK.

Widlok, T., 1997. The shared cognition of Hai//om bush people. Journal of the Royal Anthropological Institute 3, 317–332. Orientation in the wild.

Wilson, E., 1975. Sociobiology, abr, ed. Belknap, Cambridge, MA.

Wolf, E., 1999. Envisioning Power. Ideologies of Dominance and Crisis. University of California Press, Berkeley, CA.

Homelessness in the United States

Martha R Burt, Urban Institute, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Homelessness has now been a fixture on the political scene in the United States for three decades. The scope of the problem has not been debated for more than a decade, thanks to regular nationwide counts and prevalence estimates. Progress is being made on reducing and ending homelessness, using several approaches that have been evolving since 2000, some more recently than others, amid continuing debate. This article examines definitions of homelessness in the United States and how useful estimates of population size and dynamics have been established since 2007. It then provides basic descriptive information about homelessness in the United States, looks at its patterns and antecedents, describes the growth of homeless-specific service programs, and reviews research on the costs of homelessness. It ends with probable future directions for homeless research and policy.

Defining Homelessness

Homelessness did not exist as a public policy issue from the end of the Great Depression until the early 1980s. Even with increasing public monies flowing to homeless service agencies as the 1980s progressed, no formal definition of homelessness was available until Congress passed the Stewart B. McKinney Homeless Assistance Act of 1987, and even then the definition varied, and still does vary in different federal agencies, depending on the goals of their homeless assistance programs. The Department of Housing and Urban Development (HUD) is the only federal agency given responsibility for reducing the number of homeless people in the country, and thus the one whose definition has had the greatest reach and impact. Since 1987, HUD has defined people as literally homeless if they sleep in a shelter for homeless people or in places not meant for human habitation such as vehicles, abandoned building, outdoor locations, or transportation facilities or stations. Over the years this definition has expanded to include people at imminent risk of homelessness because they cannot stay where they are and have no housing alternatives, and other adjustments have been made to accommodate the needs of particular federally funded programs. The US Department of Education has a different definition of homelessness, pertinent to its mission to assure that children affected by parental homelessness can continue their education, with a priority of receiving it in their original school.

For programmatic purposes and to target resources, HUD also defines a condition known as chronic homelessness, for which a person must have at least one disability that impairs functioning and have been homeless for at least a year or for at least 12 months out of the previous 3 years. This definition applied initially only to single adults, but was expanded with 2009 legislation to include a parent in a homeless family.

Homeless Numbers

Due to the previously described changes in the definition of homelessness, there are no good estimates of the size of the literally homeless population for the 1960s and 1970s. The best guess is that in the late 1970s formal shelters (i.e., missions) for people experiencing homelessness might have been able to accommodate about 50 000 people at any one time in the entire country, and that the number of literally homeless people probably was no more than 100 000 and could have been considerably less. Even through much of the 1980s, before HUD established its definition of literal homelessness, studies of people experiencing homelessness in the United States used study-specific definitions of homelessness. By 1984 the point-in-time (PIT) estimate was up to 250 000–300 000, based largely on the perceptions of local service providers. The first study to use a statistically valid, nationwide, sampling methodology (Burt and Cohen, 1989) occurred in 1987 and produced a point-in-time estimate of 500 000–600 000 people.

After HUD established its definition, to be credible studies had to use it as the basis for at least their core population estimate, even if they also described people facing a housing crisis who might not yet meet the full definition of literal homelessness. This was the case for the next nationally representative study performed, the National Survey of Homeless Assistance Providers and Clients (Burt et al., 2001), whose data were collected in 1996. It estimated that between 444 000 (for October) and 842 000 (for February) persons *used homeless assistance programs* (*not all* were homeless people) on any given day. All of these numbers include any children accompanying a homeless adult, but not runaway and homeless youth.

HUD began collecting information about homeless populations and services for the United States as a whole in 2005, and believed it to be sufficiently reliable by 2007 to start reporting it to Congress in Annual Homeless Assessment Reports, which have been improving in accuracy and completeness since that time (HUD 2008–12). Included in its statistics are (1) data that HUD requires most communities in the country to collect at a specific point in time (PIT), during a 24-h period usually done in the last week of January, for unsheltered as well as sheltered people (this count is required every other year, but many communities do it annually) and (2) continually for users of emergency shelters and transitional housing programs to produce an annual prevalence figure. Both are lower bound estimates, but at least they are consistent from year to year.

PIT counts have dropped slightly between 2007, when almost 672 000 people were counted, and 2011, when about 636 000 people were counted. The drop was almost entirely in the unsheltered count (down 13%), which largely includes single adults, and reflects policies emphasizing supportive housing for people who have been homeless for years and have significant levels of disability. One-year estimates of sheltered homeless people are also down between 2007 and 2011, from about 1.6 million to 1.5 million people, or about 5%. But the trend for sheltered families was in the opposite direction. Sheltered family households increased 28% between 2007 and 2010, due to the Great Recession that began toward the end of 2007 and combined massive unemployment with housing loss due to foreclosures. Families using shelter began to decline by 2011, but still remained almost 14% higher than was the case in 2007.

Major Characteristics of Homeless People in the United States

Demographic Characteristics

Characteristics of people experiencing homelessness come from computerized data collected continuously in communities throughout the country to reflect *shelter and transitional housing users* over the course of a year (USDHUD, 2012 – no similar information is available for homeless people who never used a shelter throughout the year). In 2011, 63% of persons in shelter were single adults, a decline from 70% in 2007 due to the increase in sheltered homeless families. Of single adults in shelter, 72% were male; 45% were white non-Hispanics, 35% were African American non-Hispanics, and 8% were Hispanic; 70% were between 31 and 61 years old; and 43% had at least one disability – characteristics that fluctuated by one or two percentage points from year to year. Slightly more than one in three stayed in shelter for a week or less out of the reporting year, and about three in five stayed less than a month.

Of persons in homeless families using shelter in 2011, 60% were children. Among the adults, 79% were female; 28% were white non-Hispanics, 44% were African American non-Hispanics, and 21% were Hispanic; 56% were between 18 and 30 years old and another 41% were ages 31–50. Only 16% of adults in homeless families had a disability. About one in four stayed in shelter for a week or less out of the reporting year, and about half stayed less than a month.

("Other Characteristics," reported in the original, are not available from the Annual Homeless Assessment Report, and the 1996 data are too old to use.)

Patterns of Homelessness

Research on patterns of homelessness in the early 1990s (Culhane and Kuhn, 1998) differentiated three patterns of homelessness among single adults in New York and Philadelphia – first-time/crisis/transitional, episodic, and chronic – that subsequent research has been confirmed in other cities and later years. The first and largest group includes people who are homeless only once or twice and usually for a relatively short period of time (e.g., less than 3, 6, or 12 months). Often they have been precariously housed and find themselves homeless after an unexpected crisis such as unemployment, divorce/separation, or eviction. They usually secure some type of housing relatively quickly and often do not become homeless again. Episodically homeless people are defined by their pattern of cycling in and out of homelessness over significant periods of time. Chronically homeless adults, the smallest group, have relied on shelters or lived on the streets for many months or years, and many also spend time in various institutions such as inpatient treatment centers, detoxification programs, and jails. PIT studies, which overestimate the proportion of homeless people with long spells, usually find that about one-quarter of homeless adults are in the first group, and that half or more have extensive histories of either episodic or continuous homelessness. Studies that are able to observe flows into and out of the homeless condition over time give a different picture, showing only 10–20% with long-term homelessness and up to half with quite short-term or crisis patterns (Culhane and Hornburg, 1997: 101–134). More recent research confirms this picture (USDHUD, 2012).

The chronic, long-term homeless may have only a few distinct spells of homelessness, but each one lasts for a very long time; they have a high prevalence of disabilities that pose barriers to securing stable employment and housing. They absorb a highly disproportional share, often 40–50%, of homeless assistance and other resources and, for this reason, have been the focus of public policy initiatives to create supportive housing that they will accept and that ends their homelessness.

More recent research on homelessness patterns among families (Culhane et al., 2011) finds the same three patterns but the characteristics of the episodic and chronic/long-term groups are reversed. Among families, those using shelter and transitional housing programs for more than a year tend to be the *most* stable and have the most human capital, but are waiting in homeless assistance programs until they receive a rent subsidy. The episodic group of families is the smallest (about 8%), and more closely resembles the long-time homeless single adults in spending more time in other institutions when not in the shelter system.

Antecedents of or Factors Associated with Homelessness

Both societal and personal factors affect homelessness. The societal factors predict *when* and *where* homeless populations will increase or decrease; the personal factors help to predict *who* will be vulnerable to these societal pressures. The two sides of the equation determining the ability to afford housing are household income and housing costs. In the United States, housing costs for the lowest-cost units have increased significantly since the mid-1980s, while the earning ability of most people has continued to decline. Homelessness most recently became a public issue in the United States in the early 1980s, during a time of record-high unemployment (the recession of 1981–82) that reduced household income among both low-wage workers and workers in manufacturing, many of whom lost their jobs. The record prosperity of the United States during most of the 1990s into the 2000s and beyond has been accompanied by the development of a two-tiered labor market

that has relegated many people without a college or technical education to low-paying service jobs and produced the highest levels of income inequality for more than 50 years. At the same time, wealthier households have pushed housing costs to record-high levels. The differential between earning power and housing costs sets the societal stage for homelessness. A public safety net of housing and other programs could assure that households do not founder under such circumstances, but US public policy has moved in the other direction.

That being said, certain people are more likely than others to become homeless. Poverty is the *sine qua non* of homelessness. The explanatory value of demographic characteristics such as race/ethnicity diminish or disappear in many analyses that simultaneously include income variables plus characteristics such as physical health, childhood experiences, and variables relating to alcohol and drug use and mental health problems. Regression results from an increasing number of studies on both families and singles confirm that sexual or physical abuse by a household member, foster care, or other out-of-home placement in childhood differentiate both homeless families and homeless singles from their housed counterparts. Physical and mental health conditions in adulthood, and early and current alcohol or drug use are also usually found to increase the probability that a person will experience homelessness when other things are held constant.

Growth in Shelter/Housing Capacity

Residential components of US homeless assistance have grown tremendously since the early 1980s, but different segments have seen growth at different times. Growth first occurred in emergency shelter, which went from a rough estimate of 100 000 beds in 1984 to 275 000 in 1988, according to USDHUD estimates. But there growth mostly stopped; the estimate of emergency shelter capacity for 1996 was about 239 000 beds, for 2007 it was about 211 000 beds, and for 2011 it was about 226 000 beds.

In contrast, system capacity for transitional housing and permanent supportive housing (PSH), which barely existed before 1990, has expanded greatly. The major growth in transitional housing occurred in the 1990s, with about 160 000 beds by 1996 and 211 000 beds by 2007. It has recently been in decline due to policy changes, and totaled about 202 000 in 2011. The Expansion of PSH began even later and is still increasing, again due to policy emphasis. The 1996 estimates were for about 114 000 PSH beds, which rose to about 189 000 beds in 2007 and reached about 267 000 beds by 2011, with more on the way. This growth coincided with the initial availability and eventual expansion of federal funding for new types of shelter/housing stimulated by homeless legislation, to which many local initiatives have added resources, mostly for PSH.

Policy Changes and Their Relationship to Evidence of Effectiveness

In 2000 and thereafter, major efforts in the United States have focused on ending long-term homelessness for persons with complex health needs and disabling conditions by providing PSH. Motivations for this strategy are twofold: avoiding the costs incurred when this population uses expensive crisis public services, and improving health outcomes. The findings of numerous studies amply attest to reductions in crisis public service use and cost when chronically homeless people move into PSH (summarized in Culhane et al., 2007; Burt et al., 2011). A very extensive and complex study in Los Angeles (Flaming et al., 2009) examined service use across nine public agencies for over 10 000 recipients of General Relief, many of them homeless. The study found cost offsets for housing placement while considering several levels of service need and use of services. Along with numerous other studies, the Los Angeles research showed that public costs go down when individuals are no longer homeless. It also showed the benefits of the more costly services accompanying PSH for the sickest and most disabled persons – the first time that the effect of services per se, distinct from just the housing, can be clearly detected in both health outcomes and cost offsets.

Advocates used the evidence for cost avoidance, along with results showing that agencies providing PSH could move long-term homeless people with significant levels of disability off the streets and keep them housed, and documentation that the number of disabled chronically homeless individuals was small enough to be manageable, to encourage federal decision makers to adopt a change in national policy – toward a commitment to ending chronic homelessness. Many communities across the country followed suit, leading to more than doubling the availability of PSH units in the 2000s, as noted above. The result has been documented declines in chronic homelessness (USDHUD, 2012).

Other policy emphases in recent years include a major federal investment ($1.5 billion) in homelessness prevention and rapid rehousing through the Homelessness Prevention and Rapid Rehousing Program, which ran from 2009 through early 2012 as a response to the severe economic downturn then prevailing. Compared to PSH, however, much less evidence exists to document the effectiveness of these strategies.

Future Directions for Policy and Research

In the United States the argument has come full circle after almost three decades. The notion that homelessness was a problem of 'housing, housing, and housing' was made forcefully in the early 1980s, and has now proved to be quite accurate. Service programs have helped many homeless people in many ways. But to help people leave homelessness for good, or to divert people from entering into homelessness, the resources required are housing resources. These include help with landlords, back rent, evictions, and move-in expenses as diversion outlays, and actual housing as a means to end homelessness.

Future research could profitably take two directions. The first direction would focus on what supports or services made available *just before* a homeless episode or at the moment of its precipitating event could avert the transition to literal homelessness by stabilizing the situation or simply redefining

the issue. For instance, some community support services for persons with chronic mental illness offer 'respite' facilities where a person can stay for several days during a crisis in their illness. Respite stays avert complete dissolution of the person's life in the community, including loss of housing. It would be important to learn how these types of interventions can work best, for whom, and when. The second direction for research would focus on what services *in addition to housing* would make the most difference for people who are already homeless. Some communities in the United States have already shifted resources from emergency shelter and transitional housing programs to affordable and supported housing, actually reducing the need for emergency shelter beds. More attention needs to be focused on developing a range of such approaches, because it should be clear that continuing to increase the capacity of the homeless service network will only serve to bring more and more people into the category of homeless.

See also: Addictions: General Considerations; Drug Use and Abuse: Psychosocial Aspects; Drugs: Illicit Use and Prevention; Housing Economics; Housing and Cities; Mentally Ill, Public Attitudes towards; Poverty, Sociology of; Social Stratification; Street Children: Cultural Concerns; Unemployment: Structural; Violence, Criminalization, and the Homeless; Welfare Policy Comparisons; Welfare.

Bibliography

Burt, M.R., Aron, L.Y., Lee, E., 2001. Helping America's Homeless: Emergency Shelter or Affordable Housing? Urban Institute Press, Washington, DC.

Burt, M.R., Cohen, B.E., 1989. America's Homeless: Numbers, Characteristics, and the Programs that Serve Them. Urban Institute Press, Washington, DC.

Burt, M.R., Wilkins, C., Mauch, D., 2011. Medicaid and Permanent Supportive Housing for Chronically Homeless Individuals: Literature Synthesis and Environmental Scan. http://aspe.hhs.gov/daltcp/reports/2011/ChrHomlr.pdf.

Culhane, D.P., Hornburg, S., 1997. Understanding Homelessness: New Policy and Research Perspectives. Fannie Mae Foundation, Washington, DC.

Culhane, D.P., Parker, W.D., Poppe, B., Gross, K.S., Sykes, E., 2007. Accountability, cost-effectiveness, and program performance: progress since 1998. Chapter 12. In: Dennis, Deborah, Locke, Gretchen, Khadduri, Jill (Eds.), Toward Understanding Homelessness: The 2007 National Symposium on Homelessness Research. Department of Health and Human Services and Department of Housing and Urban Development, Washington, DC.

Culhane, Dennis P., Kuhn, Randall, 1998. Patterns and determinants of public shelter utilization among homeless adults in New York City and Philadelphia. Journal of Policy Analysis and Management 17 (1), 23–43.

Culhane, Dennis P., Min Park, Jung, Metraux, Stephen, 2011. The patterns and costs of services use among homeless families. Journal of Community Psychology 39 (7), 815–825.

Flaming, D., Burns, P., Matsunaga, M., Sumner, G., Moreno, M.H., Toros, H., Doan, D., 2009. Where We Sleep: Costs When Homeless and Housed in Los Angeles. Economic Roundtable, Los Angeles.

US Department of Housing and Urban Development, 2012. The 2011 Annual Homeless Assessment Report to Congress. https://www.onecpd.info/resources/documents/2012AHAR_PITestimates.pdf.

Homicide

Wendy C Regoeczi, Cleveland State University, Cleveland, OH, USA
Terance D Miethe, University of Nevada, Las Vegas, NV, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article synthesizes the large volume of literature on structural and individual correlates of homicide offending. The cross-national research on poverty, inequality, family structure, segregation, racial and ethnic composition, and population structure is reviewed. Individual level studies are summarized with respect to their findings on gender, race and ethnicity, age, relationship between victim and offender, circumstances, alcohol and drug use, and prior criminal record. The article concludes with a discussion of homicide prevention strategies.

A vast literature exists on the subject of homicide. The research in this area can be grouped broadly into aggregate-level studies that explain variation in homicide rates across cities, counties, states, regions, and countries, and individual-level studies that examine the victim, offender, incident, and situational contexts of homicide events.

Aggregate-Level Studies

Wide geographic variation exists in homicide rates. As will be described in more detail, the major theoretical perspectives that have been applied to explaining this variation include strain, social disorganization, opportunity, and social control perspectives.

Concentrated Poverty

Disparities in concentrations of poverty have been a focal point of a number of aggregate-level studies on homicide. Macro-level strain theories view these concentrations as an indication of the extent to which individuals are blocked from accessing legitimate means to attain culturally defined goals related to material success. Concentrations of poverty are also central to the arguments of social disorganization theorists, who emphasize the barriers it creates to the development of collective efficacy within neighborhoods and communities. In areas with high concentrations of poverty, unemployment, and female-headed households, neighbors are unlikely to develop the kinds of feelings of mutual trust and cohesion that provide the basis for informal social control of residents' behavior.

Studies conducted using geographic units within the United States consistently find that poverty is related to homicide rates. Recent cross-national research also supports the view that poverty is a significant structural correlate of homicide rates internationally.

Inequality

Considerable debate exists concerning the importance of relative versus absolute poverty in the production of high homicide rates in particular geographic locations. Those who argue for the importance of relative poverty believe that the disparity between one's own socioeconomic position and those around them is more likely to produce strain and motivation to use illegal means to achieve a similar standard than living in areas in which essentially no one has a high standard of living.

Various types of inequality have been examined for their relationship to homicide, including economic, racial, and gender inequality. Economic inequality has received the most attention of these inequalities. Beginning with the prominent study conducted by Blau and Blau (1982), measures of economic inequality have been assessed as a major structural correlate of homicide rates on a variety of geographic levels. Economic inequality can be expected to produce high homicide rates for several reasons. In line with the arguments of strain theorists, economic inequality may reflect blocked aspirations and opportunities for the least fortunate. In addition, economic inequality generates relative deprivation. As Jacobs and Richardson (2008: p. 31) argued, "In contrast to absolute poverty, visible economic disparities and the resulting unfavorable comparisons magnify the natural suspicions held by the least fortunate that the social procedures used to distribute rewards are not just."

The most widely used measure of inequality in this body of research is the Gini coefficient. The Gini coefficient ranges from a score of 0 (complete equality) to 1 (complete inequality) and is a reflection of how much area exists between the Lorenz curve and the line of equality (*see* Inequality Measurement).

Economic inequality has been found to explain variation in homicide rates across U.S. counties. Within the cross-national literature on homicide, inequality has emerged as an important predictor of variations in homicide rates across nations. Pridemore (2011), however, has argued that cross-national studies that find inequality to have a significant, positive relationship with homicide fail to control for poverty, and that once poverty is included in the model, any significant effect of inequality tends to disappear. Several studies support this position.

Racial Inequality

Research examining the association between racial inequality and homicide addresses the question of whether structural disparities between racial groups are a greater catalyst for lethal violence than economic inequality in general.

The findings with respect to the impact of racial inequality on homicide rates are mixed and appear to depend on the type of homicide examined. For example, racial inequality has been shown to be related to intraracial homicide offending; however, racial inequality does not appear to predict homicide offending between racial groups (i.e., interracial homicides).

Family Structure

Family structure measures commonly are included in aggregate-level studies of homicide rates. They typically serve as proxies for the level of informal social control in a given area. Locations with high proportions of female-headed households, for example, are theorized to exercise less supervision over residents. Furthermore, single-parent families do not have the time, energy, or financial resources to develop significant relationships with neighbors or to mobilize their community to address common problems. Thus, unchecked, violent crime in these areas will flourish. The prevalence of female-headed households is consistently a strong and significant predictor of homicide rates.

The divorce rate and the percent of males that are divorced also have been widely used in aggregate studies of homicide rates. The divorce rate is another means of tapping into the prevalence of one-parent families and is expected to be associated with homicide rates for the same theoretical reasons as the share of female-headed households. In other words, high levels of family disruption resulting from divorce weaken both formal and informal social controls, consistent with a structural–control perspective. Furthermore, divorce can generate additional interpersonal conflict between exspouses, directly contributing to increased rates of intimate partner homicides. From a normative conflict perspective, high divorce rates are symptomatic of a widespread collapse of traditional institutions beyond the family and may produce disorganization in which multiple family forms seek to coexist. Studies of the United States find that when examined outside of a disadvantage index, the divorce rate has a significant positive effect on homicide rates.

The percent of divorced males, on the other hand, is interpreted more commonly from a routine activities and lifestyle perspective. More specifically, a higher prevalence of divorced males in the population may be indicative of more men frequenting high-risk locations (e.g., bars, nightclubs) and engaging in high-risk behaviors that increase their attractiveness as targets and increase their contact with motivated offenders.

Racial Residential Segregation

Residential segregation often is operationalized using the index of dissimilarity, D, which measures the maximum vertical line between the Lorenz curve and the diagonal line of evenness. In essence, it provides an indication of what percent of a group's population would need to move to realize an even distribution. Other measures of residential segregation include Lieberson's (1981) P^* index.

Within the United States, this area of research has focused largely on the residential segregation experienced by African Americans, but recently it has expanded to include an examination of residential segregation among Hispanics. The extent to which racial residential segregation affects homicide victimization varies depending on the measure of segregation used.

Disadvantage Indexes

Strong interrelationships among structural correlates of violence cause significant difficulties in trying to estimate their unique impacts on homicide rates. Researchers typically have handled these problems with multicolinearity among the explanatory variables by using principal components analysis to identify common factors. Although the specific combination of variables included in the factor analysis varies across studies, the most common types of factors generated are a disadvantage–resource deprivation index (typically composed of measures of poverty, income, unemployment, and family structure but also may include inequality and racial residential segregation), residential instability (i.e., proportion of renters, vacancies, residents who have moved within past 5 years), and population structure (population size and density).

The disadvantage index is most consistently associated with homicide rates across units of analysis. The association between concentrated disadvantage and homicide may be more complex than a straightforward unconditional linear relationship. There is some indication, for example, that its influence is conditional upon the type of homicide examined. Furthermore, its impact among certain subgroups may be nonlinear such that the effect tops out after a certain point.

Racial and Ethnic Composition

The inclusion of measures of the relative size of racial or ethnic minority groups within an area is a relatively common practice in homicide research. Among studies conducted in the United States, it has become virtually routine to include a measure of the percent of the population that is black in multivariate analyses of homicide rates. Often, percent black is relegated to the role of a control variable. It is viewed by many as an indirect measure of Wolfgang and Ferracuti's subculture of violence in which subcultural values supporting the use of violence are argued to develop among groups of urban minority males. A positive relationship between percent black and homicide rates has emerged in a number of studies. Negative or nonsignificant relationships also are common and often are argued to support a benign neglect hypothesis.

Measures of the relative size of other minority groups, such as Hispanics and Native Americans, have become more common in the homicide literature in recent years. Their impact on homicide rates is not consistently significant.

Population Size and Density

Aspects of population structure, such as size and density, generally serve as indicators of criminal opportunity. Routine activities and opportunity theories argue that variation in offenses across time and space are driven in part by differences in opportunity structure. Greater access by motivated offenders to attractive targets in the absence of guardianship will generate higher levels of violent crime. All else being equal, areas with

larger numbers of residents will include more motivated offenders and suitable targets (*see* Routine Activities, Criminal Opportunities, Crime and Crime Prevention).

Analyses of population size and density measures have produced mixed findings. When significant, the effect is positive, supporting arguments that areas with larger populations will have increased homicide rates.

Individual-Level Studies

Research examining collections of homicide cases has produced a body of knowledge on the characteristics of homicide victims, perpetrators, and incidents. These characteristics of homicide identified in individual-level studies are summarized in the following sections.

Gender

A universal feature of homicide studies is the overrepresentation of males as perpetrators. Males are also disproportionately victims of homicide. The higher prevalence of homicide by male offenders has been explained by a variety of biological, psychological, and sociological factors. Biosocial theories have emphasized an interaction between biological and environmental factors. Although several biological traits have been shown to be associated with increased aggression in males, such as high testosterone, impulsivity, and low empathy, environmental factors can both facilitate and constrain their influences. Sociological theories stress the role of social learning processes, differential opportunities, subcultural influences, lifestyle differences, and gender role socialization practices in producing the well-established gender disparities in homicide offending and victimization.

One of the few exceptions to the general pattern of men's disproportionate involvement in lethal violence is the significant share of female victims killed by intimate partners. Research conducted across a number of countries has demonstrated the central role of male sexual jealousy and proprietariness in domestic violence, generally, and women killed by their abusers, specifically. The two main perspectives regarding the source of male sexual jealousy and their desire to control women are feminist theories and evolutionary psychology. For feminist theorists, the foundation of the jealousy and control lie in patriarchal social structures and cultural processes. For evolutionary psychologists, the universal nature of the overrepresentation of males in lethal violence and the common situational context and motives driving men to kill women worldwide is not consistent with the notion that their basis is entirely social or cultural. In their view, sexual proprietariness has developed among males in reaction to reproductive competition among men. The threat of losing sole access to a female partner due to perceived infidelity or her departure from the relationship would set off such a response among males.

The other general exception to this pattern is the involvement of women as perpetrators of infanticide and homicides of young children. Research on infanticide indicates that these often are committed after a hidden pregnancy to prevent parents from discovering the child's existence or when women (typically young girls) otherwise feel unable to care for the child. Infanticides often involve strangulation or suffocation of the victim where the victim's body subsequently is dumped or otherwise hidden from discovery. Cross-nationally, the most significant predictors of infant homicide victimization rates are female labor force participation and income inequality.

Race and Ethnicity

Racial and ethnic minorities often are involved disproportionately as victims and offenders of homicide. In the United States, the elevated risk of homicide among African Americans in particular is well documented. In Canada, the United States, and Australia, the native populations often have the highest rates of homicide offending and victimization. Ethnic minorities in many European countries also have elevated risks of homicide victimization. For example, in Finland, individuals of Roma descent have higher risks for homicide victimization; in Sweden, it is individuals of Yugoslavian descent; and, in the Netherlands, it is those of Antillean descent.

A number of theoretical perspectives have been applied to explain the overrepresentation of racial and ethnic minorities among homicide offenders and victims. Central to many of these arguments are the crime-producing effects of structural inequality and residential instability experienced by many minority group members. These sources of economic disadvantage in minority communities generate feelings of alienation, despair, and frustration, resulting in a greater prevalence of economically motivated homicides (i.e., those committed during a robbery or burglary) and also generating greater interpersonal conflict and hence dispute-related homicides.

Regarding the overrepresentation of African Americans as homicide offenders and victims within the United States specifically, explanations have emphasized the profound impact of the historical legacy of slavery and racism that has limited the opportunity for African Americans to achieve justice and resolve conflicts in everyday life through legitimate avenues. Within this context, violence often is viewed as self-help or as other manifestations of an alternative avenue of social control. This is particularly likely to occur in cases in which minority populations lack trust in the police and the criminal justice system more generally to provide just responses in dispute resolutions.

Age

As with other types of criminal offending, individuals in their late teens and early twenties are the most likely to be homicide offenders and victims. A number of sociological theories have been applied as explanations of the age distribution of homicide offending. In particular, the high concentration of homicide arrestees among young adults has been attributed to this age-group's low self-control, high impulsivity, weakened bonds and attachments to social institutions (e.g., family, schools, religion), and routine activities and lifestyles that expose them to risky and dangerous situations. In addition to maturational reform, changes occurring with respect to lifestyles and routine activities, increased stakes in conformity, and the development

of attachments to significant others as individuals enter into adulthood explain the lower levels of homicide offending in older age groups.

Relationship between Victim and Offender

In spite of the fear they generate, homicides committed by strangers constitute a minority of cases. People are most likely to be killed by someone known to them, particularly friends and acquaintances. Research indicates that the situational context underlying intimate partner, family, friend or acquaintance, and stranger homicides differs. Intimate partner homicides typically involve individuals of the same race and approximate age but the opposite sex. They are virtually the only type of homicide in which female victims outnumber male victims. Although the immediate precipitating event in these acts of lethal violence is often a dispute, intimate partner relationships that end in homicide almost always are characterized by a history of domestic violence against the female partner. Additional risk factors for intimate partner homicides, especially of female victims, include estrangement, forced sex, the presence of stepchildren in the home, violence during pregnancy, access to guns, prior incidents of stalking, threats to kill, and earlier episodes of choking or strangulation.

Family homicides are largely expressive acts of violence and often involve men killing other men. Within the United States, they are primarily committed with firearms. When they involve older participants (i.e., not children), they commonly involve alcohol use and fit the image of confrontational homicides, thus sharing situational characteristics with homicides involving friend and acquaintances.

Homicides occurring between friends and acquaintance homicides are largely the domain of males. Participants tend to be of similar age and racial background. Their situational contexts are typically dispute related. They often fit the image of character contests and begin as verbal confrontations that escalate into physical attacks. They occur disproportionately in the evenings, on weekends, and in public locations.

Stranger homicides involve a diverse array of situational contexts. They are more likely than other types of homicides to be instrumentally motivated and to be committed by males. Although some fit the typical conception of an innocent bystander or random victim killed by someone with whom they had no prior relationship, it is also common for stranger homicides to occur between individuals involved in criminal lifestyles (e.g., drug dealing, prostitution).

Circumstances

The literature frequently classifies homicides as expressive or instrumental. Expressive homicides are spontaneous, emotionally driven acts and are the most common situational context for lethal violence. Instrumental homicides are those that are motivated by prospective goals, including the acquisition of money or status. Homicides most commonly classified as instrumental are those occurring during the commission of other crimes, such as robberies and rapes. These types of homicides constitute a minority of homicides.

Expressive and instrumental homicides differ in terms of their situational context. Instrumental homicides are more likely to involve strangers, multiple offenders, young offenders, and men killing other men. In contrast, expressive homicides tend to involve older offenders, interracial attacks, and nonstrangers.

Expressive homicides often take the form of confrontational disputes involving character contests. Character contests involve defenses of one's social self in response to status threats, often in public in the presence of others. Prior disputes between the victim and offender are common. In some, the victim plays a contributory role by initiating the verbal confrontation or being the first to resort to physical violence, leading to their depiction as victim-precipitated homicides.

Homicides occurring within the context of illegal drug transactions are frequent circumstances for robbery homicides in many urban areas. Dealers are targeted as easy prey for ripping off both money and drugs given their lack of access to legal recourse. Disputes over quantity and quality of drugs, drug-selling territory, and drug debts also produce lethal confrontations in which robberies may occur as well.

High-profile types of homicides – such as mass murders, serial killers, and homicides committed by law enforcement – are rare and yet receive the most media attention.

Alcohol and Drug Use

It is common for both the victim and offender to have been consuming alcohol at the time of the homicide incident. Cross-nationally, approximately half of homicide victims are under the influence of alcohol when they are killed; about one-third would meet the threshold for intoxication. Alcohol use among homicide victims tends to increase with age. Inconsistent conclusions have emerged with respect to whether it is more common among female or male victims. A history of alcohol dependence is also relatively common among homicide offenders. Examinations of cross-cultural variations in the influence of alcohol on homicide have produced mixed findings. Some research reports that levels of alcohol use in homicides are relatively consistent across locations, but others conclude alcohol has a greater impact on homicide in specific countries (e.g., Russia).

Most theories of alcohol use and homicide focus on the impact of alcohol on an individual's cognitive functioning and reasoning. Alcohol reduces inhibitions and removes normative rules prohibiting the use of violence. This is particularly likely to be the case in situations in which active constraint is required in the sense that violence is perceived as an effective dispute-resolution mechanism given the dynamics of the specific situation.

Although not as prevalent, a quarter or more homicide victims are under the influence of illicit substances at the time of their death. As with alcohol, this has been explained by the influence of illicit substances on brain functioning, perception, and the creation of delusions and hallucinations. The overlap between individuals who use and sell drugs may be an equally important factor. The risk of violence to those involved in drug markets is well documented and remains one of the leading explanations for the tremendous upsurge in lethal violence in the late 1980s and early 1990s in the United States.

Prior Criminal Record

Prior criminal records are common among both homicide victims and offenders. The significant involvement of homicide

victims in criminal activity frequently has been cited as evidence in support of an overlap in populations of homicide offenders and victims. That a significant portion of homicide victims and offenders are drawn from the same population of individuals involved in deviant lifestyles is consistent with the arguments of a number of criminological theories. Chief among these are routine activities theory and the subculture of violence. The 'principle of homogamy' derived from lifestyle and routine activities theories asserts that a person's risk of victimization is correlated with the frequency with which they interact or have contact with members of demographic groups that account for a disproportionate share of offenders. To the extent that the subculture of violence thesis argues that subgroups of the population are geared normatively toward violence, shared perceptions and misperceptions regarding engagement in violence among both victims and offenders mean the two populations will not be distinct, instead interchanging roles in a system of subcultural relationships.

Prevention

Deterrence has been heavily relied on as a strategy for reducing crime generally and violent crime specifically. Although the exact deterrence strategies that have been used vary across countries, these approaches include the implementation of mandatory minimum sentences, the development of three-strikes laws, making it easier to process juvenile offenders as adults, and the use of the death penalty. In general, efforts to increase the severity of punishment have shown only limited success in reducing serious violent crime. More targeted strategies, however, such as those that take aim at the illegal carrying of firearms in high–violent crime areas and directed police patrols in those localities, have been shown to be effective.

As cynicism has mounted over the ability of the criminal justice system to reduce violence in general and lethal violence in particular, homicide prevention efforts increasingly have turned toward multiagency collaborations and public health approaches. Within the United States, successful programs such as Boston's Operation Ceasefire follow this model by having developed partnerships between a number of criminal justice agencies and local activist groups. The program relies on a strategy of 'pulling levers,' in which attention was focused on a small group of individuals involved in gang violence and trafficking illegal firearms. These chronic gang members are given a message of no tolerance for violence and every available legal 'lever' is used to respond to violence when it occurs to create a swift and intense response.

The implementation of death review committees composed of participants from a broad spectrum of social service and criminal justice agencies to conduct in-depth examinations of child fatalities, deaths from domestic violence, and – in a few instances – all homicides within a given jurisdiction have helped to identify system gaps and other policy changes that can prevent future homicides.

The preventive aspects of restricting the availability of alcohol establishments, increasing alcohol taxes, and making other efforts to reduce alcohol consumption are supported by a body of cross-national research on alcohol and homicide. Given the prevalence of alcohol and drug use among homicide offenders and victims, providing substance abuse treatment to those at risk for serious violence likely will yield long-term benefits in terms of homicide reduction. Increased access to mental health services, particularly among children exposed to violence, also has been shown to be a promising avenue.

Increased resources for domestic violence victims in the form of hotlines, shelters, and legal advocacy services; widespread use of lethality assessments in cases of nonlethal intimate partner violence; and the development and strict enforcement of policies prohibiting possession of firearms among individuals arrested for domestic violence are valuable approaches for preventing intimate partner homicides more specifically.

See also: Age and Crime; Collective Efficacy and Crime; Control Theories and Crime; Crime and Gender; Crime, Race and Ethnicity; Domestic Violence: Sociological Perspectives; Inequality Measurement; International and Comparative Criminology; Routine Activities, Criminal Opportunities, Crime and Crime Prevention; Strain Theories and Crime.

Bibliography

Anderson, E., 1999. Code of the Street: Decency, Violence, and the Moral Life of the Inner City. W.W. Norton, New York.

Beaulieu, M., Messner, S.F., 2010. Assessing changes in the effect of divorce rates on homicide rates across large U.S. cities, 1960–2000. Revisiting the Chicago school. Homicide Studies 14, 24–51.

Blau, J., Blau, P., 1982. Metropolitan structure and violent crime. American Sociological Review 47, 114–128.

Block, C.R., Devitt, C.O., Fonda, D., Fugate, M., Marting, C., McFarlane, J., 2000. The Chicago Women's Health Study: Risk of Serious Injury or Death in Intimate Partner Violence: a Collaborative Research Project. U.S. Department of Justice, National Institute of Justice, Washington, DC.

Braga, A.A., Kennedy, D.M., Waring, E.J., Piehl, A.M., 2001. Problem-oriented policing, deterrence, and youth violence: an evaluation of Boston's Operation Ceasefire. Journal of Research in Crime and Delinquency 38, 195–226.

Bye, E.K., 2008. Alcohol and homicide in Eastern Europe: a time series analysis of six countries. Homicide Studies 12, 7–27.

Campbell, J.C., Glass, N., Sharps, P.W., Laughon, K., Bloom, T., 2007. Intimate partner homicide: review and implications of research and policy. Trauma, Violence, & Abuse 8, 246–269.

Cohen, L.E., Felson, M., 1979. Social change and crime rate trends: a routine activity approach. American Sociological Review 44, 588–608.

Daly, M., Wilson, M., 1988. Homicide. Aldine de Gruyter, New York.

Dobash, R.E., Dobash, R.P., Cavanagh, K., Lewis, R., 2004. Not an ordinary killer—just an ordinary guy. When men murder an intimate partner. Violence against Women 10, 577–605.

Dugan, L., Nagin, D.S., Rosenfeld, R., 1999. Explaining the decline in intimate partner homicide: the effects of changing domesticity, women's status, and domestic violence resources. Homicide Studies 3, 187–214.

Farrington, D.P., Loeber, R., Stallings, R., Homish, D.L., 2012. Early risk factors for young homicide offenders and victims. In: DeLisi, M., Conis, P.J. (Eds.), Violent Offenders: Theory, Research, Policy, and Practice, second ed. Jones & Bartlett, Burlington, MA, pp. 21–39.

Fiala, B., LaFree, G., 1988. Cross-national determinants of child homicide. American Sociological Review 53, 432–445.

Gartner, R., 1990. The victims of homicide: a temporal and cross-national comparison. American Sociological Review 55, 92–106.

Jacobs, D., Richardson, A.M., 2008. Economic inequality and homicide in the developed nations from 1975 to 1995. Homicide Studies 12, 28–45.

Johnson, H., 2012. When feminism meets evolutionary psychology: the enduring legacy of Margo Wilson. Homicide Studies 16, 332–345.

Krivo, L., Peterson, R.D., 2000. The structural context of homicide: accounting for racial differences in process. American Sociological Review 65, 547–559.

Kuhns, J.B., Wilson, D.B., Clodfelter, T.A., Maguire, E.R., Ainsworth, S.A., 2011. A meta-analysis of alcohol toxicology study findings among homicide victims. Addiction 106, 62–72.

LaFree, G., 1999. A summary and review of cross-national comparative studies of homicide. In: Smith, M.D., Zahn, M.A. (Eds.), Homicide: A Sourcebook of Social Research. Sage Publications, Thousand Oaks, CA, pp. 125–145.

Land, K.C., McCall, P.L., Cohen, L.E., 1990. Structural covariates of homicide rates: are there any invariances across time and social space? American Journal of Sociology 95, 922–963.

Landberg, J., Norstrom, T., 2011. Alcohol and homicide in Russia and the United States: a comparative analysis. Journal of Studies on Alcohol & Drugs 72, 723–730.

Lieberson, S., 1981. An asymmetrical approach to segregation. In: Peach, C., Robinson, V., Smith, S. (Eds.), Ethnic Segregation in Cities. Croom Helm, London, pp. 61–82.

Liem, M., Ganpat, S., Granath, S., Hagstedt, J., Kivivuori, J., Lehti, M., Nieuwbeerta, P., 2013. Homicide in Finland, the Netherlands, and Sweden: first findings from the European Homicide Monitor. Homicide Studies 17, 75–95.

McCall, P.L., Land, K.C., Parker, K.F., 2010. An empirical assessment of what we know about structural covariates of homicide: a return to a classic 20 years later. Homicide Studies 14, 219–243.

Merton, R.K., 1938. Social structure and anomie. American Sociological Review 3, 672–682.

Messner, S.F., Golden, R., 1992. Racial inequality and racially disaggregated homicide rates: an assessment of alternative theoretical explanations. Criminology 30, 421–448.

Messner, S.F., Rosenfeld, R., 1997. Political restraint of the market and levels of criminal homicide: a cross-national application of institutional-anomie theory. Social Forces 75, 1393–1416.

Miethe, T.D., Regoeczi, W.C., 2004. Rethinking Homicide: Exploring the Structure and Process Underlying Deadly Situations. Cambridge University Press, Cambridge, UK.

Parker, K.F., 2004. Industrial shift, polarized labor markets and urban violence: modeling the dynamics between the economic transformation and disaggregated homicide. Criminology 42, 619–645.

Parker, K.F., McCall, P.L., 1999. Structural conditions and racial homicide patterns: a look at the multiple disadvantages in urban areas. Criminology 37, 447–478.

Parker, R.N., 1995. Bringing "booze" back in: the relationship between alcohol and homicide. Journal of Research in Crime and Delinquency 32, 3–38.

Polk, K., 1994. When Men Kill: Scenarios of Masculine Violence. Cambridge University Press, New York.

Pratt, T.C., Cullen, F.T., 2005. Assessing macro-level predictors and theories of crime: a meta-analysis. Crime and Justice: A Review of Research 32, 373–450.

Pridemore, W.A., 2008. A first test of the poverty-homicide hypothesis at the cross-national level. Criminology 46, 133–154.

Pridemore, W.A., 2011. Poverty matters: a reassessment of the inequality-homicide relationship in cross-national studies. British Journal of Criminology 51, 739–772.

Roth, R., 2009. American Homicide. Harvard University Press, Cambridge, MA.

Sampson, R.J., 1987. Urban black violence: the effect of male joblessness and family disruption. American Journal of Sociology 93, 348–382.

Sampson, R.J., Raudenbush, S.W., Earls, F., 1997. Neighborhoods and violent crime: a multilevel study of collective efficacy. Science 227, 916–924.

Shaw, C.R., McKay, H.D., 1942. Juvenile Delinquency and Urban Areas: A Study of Rates of Delinquents in Relation to Differential Characteristics of Local Communities in American Cities. University of Chicago Press, Chicago, IL.

Shihadeh, E.S., Flynn, N., 1996. Segregation and crime: the effect of Black social isolation on the rates of Black urban violence. Social Forces 74, 1325–1352.

Wadsworth, T., Kubrin, C.E., 2004. Structural factors and black interracial homicide: a new examination of the causal processes. Criminology 42, 647–672.

Wilson, J.M., MacDonald, J.M., Tita, G.E., 2010. Localized homicide patterns and prevention strategies: a comparison of five project safe neighborhood sites. Victim and Offenders 5, 45–63.

Wolfgang, M., Ferracuti, F., 1967. The Subculture of Violence: Toward an Integrated Theory in Criminology. Tavistock, London.

Honesty and Humility

Ingo Zettler, University of Copenhagen, Copenhagen, Denmark; and University of Tübingen, Tübingen, Germany
Benjamin E Hilbig, School of Social Sciences, University of Mannheim, Mannheim, Germany; and Max Planck Institute for Research on Collective Goods, Bonn, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Based on recent lexical studies across various languages, Honesty-Humility has been suggested as a sixth basic factor of personality. Specifically, according to the HEXACO Model of Personality, Honesty-Humility represents individual differences in active cooperativeness – operationalized via the facets sincerity, fairness, greed avoidance, and modesty. Herein, research corroborating Honesty-Humility as a basic personality factor is summarized and supplemented with an overview of the broad array of criteria this factor has been shown to influence. In addition, we briefly touch on aspects that are still understudied concerning Honesty-Humility, especially in terms of its biological basis, development, and broader societal consequences.

The personality factor named Honesty-Humility is defined as "the tendency to be fair and genuine in dealing with others, in the sense of cooperating with others even when one might exploit them without suffering retaliation" (Ashton and Lee, 2007: p. 156). Correspondingly, Honesty-Humility encompasses attributes such as being fair-minded, honest, modest, and sincere as opposed to being boastful, deceitful, greedy, and hypocritical (Ashton and Lee, 2008a). According to the HEXACO Model of Personality (Ashton and Lee, 2007), Honesty-Humility represents one of the (six) basic dimensions of personality, subsuming aspects of reciprocal altruism in terms of active cooperation. As such, Honesty-Humility has not only been considered an important construct for personality psychology, but also, more generally, for many socially important criteria (e.g., Ashton and Lee, 2008b; Dunlop et al., 2012; Lee et al., 2013). Herein, we describe the recovery of Honesty-Humility as a core personality factor, its theoretical and empirical background, various aspects it has been linked to, potential future directions, and (un)solved methodological issues.

Theoretical and Empirical Background

The recovery of Honesty-Humility as a basic personality factor can be traced to several recent lexical studies performed across different languages and cultures. The underlying reasoning behind lexical studies is that aspects of individual differences are communicated about, thereby becoming part of everyday language, and are thus ultimately included in dictionaries. In lexical studies, researchers present lists of personality-descriptive words (typically adjectives) drawn from dictionaries to participants who rate the extent to which they (or close acquaintances) possess these attributes. Data from large samples are then (factor) analyzed, resulting in a relatively small set of mostly unrelated factors that summarize a wide range of individual differences.

In recent years, findings from lexical studies across more than 10 languages – including Croatian, Dutch, English, Filipino, French, German, Greek, Hungarian, Italian, Korean, Polish, and Turkish – have suggested that personality might best be described by six factors (e.g., Lee and Ashton, 2008), resulting in the HEXACO Model of Personality. Honesty-Humility is one of these six factors alongside Emotionality, eXtraversion, Agreeableness, Conscientiousness, and Openness to experience. Note that findings from earlier lexical studies have often been considered support for the Five-Factor Model of Personality, and the remaining five factors of the HEXACO model are similar to these in content. At closer inspection, however, some of these earlier studies described a factor that can be interpreted as Honesty-Humility, though it was typically not considered a fully-fledged factor in its own right (for an overview, see Ashton and Lee, 2001). Also, early (English) lexical studies, which served as the primary basis for the original Five-Factor Model, were typically limited in terms of how many items (adjectives) could be considered, thus providing less precise factor solutions (cf e.g., Lee and Ashton, 2008).

Importantly, lexical studies merely provide a condensed descriptive simplification and grouping of personality-descriptive adjectives. As such, they do not inherently yield a theoretical account of basic personality. However, beyond the mere lexical findings, the HEXACO model (Ashton and Lee, 2007; Ashton et al., 2014) does encompass theoretical reasoning for why and how its six dimensions make up the core dimensions of personality. More precisely, drawing connections to evolutionary biology, Honesty-Humility, Agreeableness, and Emotionality are suggested to represent aspects of altruism, namely active reciprocal altruism (Honesty-Humility), reactive reciprocal altruism (Agreeableness), and kin altruism (Emotionality). The three other factors, in turn, are taken to represent engagement or endeavor with regard to socially related (Extraversion), task-related (Conscientiousness), and idea-related (Openness to experience) areas.

Correspondingly, Honesty-Humility reflects a blend of characteristics, which contribute to (long-term) mutual cooperation as opposed to actively exploiting others. In this vein, Lee and Ashton (2004) operationalized Honesty-Humility as comprising the facets of Sincerity ("a tendency to be genuine in interpersonal relations"), Fairness ("a tendency to avoid fraud and corruption"), Greed Avoidance ("a tendency to be

uninterested in possessing lavish wealth, luxury goods, and signs of high social status"), and Modesty ("a tendency to be modest and unassuming," all p. 334). Unlike Agreeableness (comprising the facets Forgivingness, Gentleness, Flexibility, and Patience), Honesty-Humility refers to how people typically behave when they can actively shape cooperation (nonexploitation). Agreeableness, on the other hand, refers to how people react to (perceived) exploitation, that is, their tendency to forgive or retaliate.

Summarizing thus far, Honesty-Humility has been recovered as a basic dimension in several lexical studies across different languages, linked theoretically to established concepts from evolutionary biology, and proposed to play an important role for various social criteria – especially in the realm of behaviors reflecting active cooperation. Next, we provide an overview of research testing these conjectures, i.e., (1) that Honesty-Humility represents a core personality factor, and (2) that it accounts for cooperation in terms of active reciprocal altruism.

A Basic Personality Dimension Explaining Active Cooperation?

The claim that Honesty-Humility represents a core personality dimension has been examined in several investigations. In particular, its relation to established personality attributes – especially those typically thought to lie outside of the range captured by the Big Five – has been examined. For instance, Honesty-Humility has been consistently associated with constructs of the 'Dark Triad,' i.e., Machiavellianism (a personality construct capturing attributes such as being deceitful, ruthless, and self-centered), narcissism (a personality construct capturing attributes such as grandiosity, entitlement, and feeling superior to others), and psychopathy (a personality construct capturing attributes such as having a lack of empathy, a parasitic lifestyle, or high impulsivity). Quite irrespective of whether self- and/or observer-reports are relied on, correlations between Honesty-Humility and the Dark Triad constructs are typically between $r \approx -0.40$ and $r \approx -0.70$, and Honesty-Humility has been found to explain unique variance in the prediction of the Dark Triad constructs beyond the Big Five (Lee and Ashton, 2014; Lee et al., 2013). A similar pattern has emerged for relations between Honesty-Humility and other traits that are not fully accounted for within the Big Five space, for instance attributes of Paunonen's Supernumerary Personality Inventory (Lee et al., 2005) or egoism (De Vries et al., 2009). So, Honesty-Humility has been found to broadly and uniquely cover several aspects of the personality space, thus confirming the implication of lexical findings that it represents a basic dimension of personality.

Concerning the core theoretical notion that Honesty-Humility essentially relates to active reciprocal altruism, a series of critical tests can be highlighted. Specifically, Honesty-Humility was shown to predict dictator game giving (e.g., Hilbig and Zettler, 2009), contributions to public goods (Hilbig et al., 2012), choice of the cooperative option in prisoner dilemmas (Zettler et al., 2013), and the willingness to maximize joint outcomes (at some personal cost) in the framework of social value orientations (e.g., Hilbig et al., 2014). As such, Honesty-Humility predicted both self-reported and actual active cooperation, typically yielding small- to medium-sized effects. Also, Honesty-Humility repeatedly accounted for incremental variance beyond the other basic factors (either in terms of the Five Factor or the HEXACO model) and was by far the strongest – and in several cases the only – predictor. In addition, recent investigations (e.g., Hilbig et al., 2013a) tested the theoretical distinction between Honesty-Humility (active cooperation) and Agreeableness (reactive cooperation) according to the HEXACO model, and found empirical support for the dissociation of these factors.

Honesty-Humility and Life Outcomes

Although Honesty-Humility was introduced relatively recently, it has been investigated with increasing frequency – spanning broad areas of interest. Herein, we will roughly sketch the most dominant of these research streams, namely sex-, money-, and power-related variables, deviant behavior and crime, as well as sociopolitical attitudes.

Based directly on links between Honesty-Humility and Machiavellianism, narcissism, and exploitativeness, Honesty-Humility has also been found to be negatively associated with selfish tendencies in a broader sense. For instance, a study by Lee et al. (2013; for similar results, see e.g., Ashton and Lee, 2008b) found that people low in Honesty-Humility showed more selfish tendencies concerning sex, money, and power. More precisely, across self- and peer-reports, Honesty-Humility was negatively related to a sex-related measure encompassing short-term mating orientation and behavior, a ludus love style, and the tendency for sexual quid pro quo. Also, Honesty-Humility accounted for a money-related measure capturing conspicuous consumption and materialism as well as for a power-related measure comprising a desire for power and social dominance orientation (i.e., the tendency to endorse social hierarchy and inequality). In the analyses, Honesty-Humility was typically the strongest predictor as compared to the other five factors of the HEXACO model, to the Big Five measured via corresponding inventories, or to composite measures of the Dark Triad constructs. Thus, extending the findings outlined above, Honesty-Humility appears to impact a broader set of selfish, egoistic, or self-centered tendencies and behaviors that are of societal interest.

Another fruitful research stream has linked Honesty-Humility to counterproductive activities, deviant behavior, and crime. Since low Honesty-Humility entails one's pull toward prestige, high social status, or luxury goods, a link between this trait and individuals' willingness to engage in criminal or deviant behavior (in order to reach these aims) is typically expected. Correspondingly, across different countries and across a variety of samples including students, employees, and community samples, individuals high in Honesty-Humility have been found to show less deviant or crime-related behavior subsuming fraud, theft, vandalism, alcohol, drug consumption etc. Results typically indicate medium to strong effects of Honesty-Humility, outperforming all remaining core dimensions of personality (e.g., Ashton and Lee, 2008a; Dunlop et al., 2012; Zettler and Hilbig, 2010). In a similar vein, Honesty-Humility often added incremental variance beyond variables such as self-control or integrity test scores, which have been

considered among the most important predictors of deviance-related criteria. It must be noted, however, that this line of research is consistently limited by a lack of actual behavior as the criterion and a strong reliance on *reported* behavior. Overcoming this limitation seems necessary before a final judgment on the influence of Honesty-Humility for crime-related behavior can be made, although the extant evidence is entirely theory-consistent.

Furthermore, a growing line of research has linked Honesty-Humility to social, ecological, economic, or political attitudes and behavior. For instance, Honesty-Humility has been found to be positively associated with proenvironmental attitudes and ecologically responsible behavior (Hilbig et al., 2013b). This is compatible with the notion that solving ecological issues typically requires cooperative action from an individual or group (e.g., the 2°C climate target, i.e., reducing carbon dioxide emissions to a level limiting global warming to 2°C) and/or forgoing personal luxury (e.g., refraining from heavy air conditioning). Concerning politics, Honesty-Humility has been repeatedly associated with low social dominance orientation (see above) and (in some countries) a more left-wing political orientation and corresponding voting behavior (e.g., Chirumbolo and Leone, 2010). The conceptual link at this point is arguably an unconditional cooperative orientation, also implying a preference for social equality. Relatedly, Pozzebon and Ashton (2009) found that Honesty-Humility was linked to values and corresponding behavior concerning benevolence (positively), hedonism (negatively), and power (negatively), and Lee et al. (2009) demonstrated that Honesty-Humility is the basic personality factor most strongly related to one of the two central dimensions of personal values, namely self-transcendence versus self-enhancement (whereas the other core personal value, Openness to change vs Conservation, is most strongly associated with Openness to experience). Besides such direct relations, the notion that Honesty-Humility relates to a central dimension of personal values is further supported by the finding that people overestimate the similarity between themselves and close acquaintances particularly in terms of Honesty-Humility and Openness to experience, but not on other basic factors. This could be taken to imply that people emphasize similarity in characteristics associated with Honesty-Humility and Openness to experience as these most strongly correspond to important values for one's identity (Lee et al., 2009). Overall, alongside Openness to experience, Honesty-Humility has been established as the basic personality factor associated most strongly with ecological, political, and morality-related values, attitudes, and behavior.

Beyond the research summarized herein, other investigations have successfully linked Honesty-Humility to a broad range of variables in the realm of personality disorder traits (e.g., Antagonism), risk taking and sensation seeking (e.g., ethical risk taking), religiousness, creativity, and other domains – a continuously updated list of published articles concerning the HEXACO model in general and Honesty-Humility in particular can be found at www.hexaco.org.

Interaction with Situations

Beyond such 'main effects' of Honesty-Humility on the Dark Triad constructs, crime, sociopolitical attitudes etc., several studies have also attempted to place Honesty-Humility into the broader framework of person–situation interaction. Indeed, apart from general notions contending that behavior will always result from an interplay between person and situation factors, the definition of Honesty-Humility itself explicitly entails a corresponding prediction: Ashton and Lee (2007) described Honesty-Humility as the tendency for fairness and cooperation "*even when* one might exploit [others] without suffering retaliation" (p. 156, emphasis added). In other words, people high in Honesty-Humility are expected to show cooperative behavior irrespective of whether or not the situation would allow them to exploit others. Those low in Honesty-Humility, by contrast, should, by implication, exploit others if and only if this is possible without running the risk of retaliation or other negative consequences. As a core methodological strength, Honesty-Humility is thus defined in a way that allows for experimental tests.

In one early straightforward test of the hypothesized person–situation interaction, Hilbig and Zettler (2009) asked participants to allocate goods (100 points) between themselves and another person in both the dictator and the ultimatum game. Importantly, in the dictator game the other person has no power over the offer, that is, she or he has to accept it (even if the 'offer' is zero, i.e., the dictator keeps everything). In the ultimatum game, on the other hand, the other person may reject the offer in which case neither of the participants receives any payoff at all. So, in the latter case, attempted exploitation (by the allocator) may be met with retaliation (by the recipient). Hilbig and Zettler (2009) found an interaction between Honesty-Humility and this situational manipulation (dictator vs ultimatum game) in line with the definition of Honesty-Humility: Those high in Honesty-Humility allocated a fair amount of goods to the recipient in both conditions, whereas those low in Honesty-Humility were selfish in the dictator game (in which they did not need to fear retaliation), but shifted toward a fair offer in the ultimatum game.

A similar pattern of interaction has been found with regard to two social dilemmas, namely, public goods and prisoner's dilemma. Concerning the public goods dilemma, Hilbig et al. (2012) manipulated the anonymity and presence versus absence of potential punishment for free-riding, and found that those high in Honesty-Humility, once again, contributed substantially to the pubic goods quite irrespective of this manipulation. Those low in Honesty-Humility, in turn, increased their contributions strategically only if potential punishment for free-riding was introduced. In the prisoner's dilemma, Zettler et al. (2013) manipulated the potential benefits (and risks) of choosing the noncooperative option. Again, it was found that those high in Honesty-Humility chose the cooperative option relatively irrespective of how attractive (and nonrisky) defection was, whereas those low in Honesty-Humility conditioned their decision on such incentives for defection (cooperating if and only if defection was not very tempting and rather risky).

Importantly, this interaction pattern has also been found in an applied setting. In a field study with employees, Zettler and Hilbig (2010) investigated whether Honesty-Humility interacted with the situational factor of organizational politics (in simple terms, whether the work environment is characterized by self-serving activities) in bringing about counterproductive work behavior. Indeed, the results suggested a similar pattern: Those high in Honesty-Humility showed relatively little

counterproductive work behavior, again quite irrespective of the level of organizational politics in their workplace. Those low in Honesty-Humility also refrained from counterproductive work behavior whenever the level of organizational politics was low, and, in turn, when the (subjective) probability for being sanctioned for self-serving behavior was rather high; however, under high levels of organizational politics they showed substantially more counterproductive work behavior.

Methodological Issues

The most widely used measures to assess Honesty-Humility are the corresponding scales from the different versions of the HEXACO Personality Inventory-Revised (HEXACO-PI-R, Ashton and Lee, 2009; Lee and Ashton, 2004, 2006). More precisely, three versions of the HEXACO-PI-R exist, namely one version (HEXACO-60) assessing Honesty-Humility with 10 items, one version (HEXACO-100) assessing Honesty-Humility with 16 items, and one version (HEXACO-200) assessing Honesty-Humility with 32 items. The HEXACO-PI-R is available in different languages as a self- and, in several cases, an observer-report form – including Chinese, Croatian, Dutch, English, French, German, Hungarian, Italian, Japanese, Korean, Lithuanian, Polish, Romanian, Serbian, Spanish, and Turkish. Sample items for Honesty-Humility are "I wouldn't pretend to like someone just to get that person to do favors for me" (belonging to the sincerity facet), "I would never accept a bribe, even if it were very large" (belonging to the fairness facet), "Having a lot of money is not especially important to me" (belonging to the greed avoidance facet), or "I am an ordinary person who is no better than others" (belonging to the modesty facet).

Although all three versions have been found to yield good psychometric properties in general, they entail the usual trade-off between precision on the facet level and brevity. Specifically, in the HEXACO-60, each facet is represented by two or three items only. Therefore, facet scores will typically not yield satisfactory reliability. In the HEXACO-100, each facet is represented by four items, resulting in the possibility to obtain facet scores with acceptable reliability (the internal consistency estimates typically are between $\alpha \approx 0.60$ and $\alpha \approx 0.85$). In the HEXACO-200, finally, each facet is represented by eight items providing relatively robust facet-level information (internal consistency estimates typically are between $\alpha \approx 0.75$ and $\alpha \approx 0.90$).

Besides the HEXACO-PI-R, other measures assessing Honesty-Humility (or essentially identical constructs) have also been introduced, including the Questionnaire Big Six Scales (QB6; Thalmayer et al., 2011), scales derived from the International Personality Item Pool (IPIP; Ashton et al., 2007), or a relatively brief measure (consisting of 24 items) of the HEXACO factors (Brief HEXACO Inventory, BHI, De Vries, 2013). These notwithstanding, most research concerning Honesty-Humility has been conducted with versions of the HEXACO-PI-R. On the one hand, this fact facilitates comparisons between research findings and fosters general conclusions about the validity of Honesty-Humility as represented by the HEXACO-PI-R (which arguably is most directly tied with the HEXACO model). On the other hand, it could also be considered a drawback because findings might be limited to some unknown specifics of the HEXACO-PI-R. So, future research might profit from an increased use of alternative questionnaires assessing Honesty-Humility.

Also, the currently available measures are constructed for adults in particular, both in terms of wording and/or the context referred to (e.g., bribes, taxes, or the work life). As such, none allow for assessing the six HEXACO factors of (younger) children. Thus, results on the importance of Honesty-Humility are limited to adults, so far, and it could be a fruitful future direction to extend this knowledge to children and adolescents. In addition, there is a continued need for evidence based on methods other than questionnaires, e.g., direct observation of behavior, implicit attitude measures, or physiological variables to provide a broader understanding of Honesty-Humility and its nature and consequences.

A crucial aspect in the assessment of Honesty-Humility is its relation to response biases such as socially desirable responding. Indeed, given that Honesty-Humility encompasses attributes such as being sincere, honest, and loyal versus deceitful, boastful, and hypocritical (Ashton and Lee, 2008a; Lee and Ashton, 2004, 2006), Honesty-Humility can be considered a socially desirable characteristic in itself. So, one might argue that individuals scoring high in Honesty-Humility are not actually more sincere, honest, or the like but rather exhibit a stronger tendency to claim possessing such attributes – at least, when self-reports of Honesty-Humility are used. Although research on the link between Honesty-Humility and socially desirable responding is limited, so far, some conclusions can be pointed out: First off, an increasing number of findings has suggested that self- and observer-reports of Honesty-Humility are typically correlated just as strongly as self- and observer-reports of other traits, namely between $r \approx 0.50$ and $r \approx 0.60$ (e.g., Dunlop et al., 2012; Lee and Ashton, 2006). This hints that Honesty-Humility is, at best, as much affected by socially desirable responding as the other basic factors. Perhaps more importantly, a positive link between Honesty-Humility and truthful responding has been found based on observation of actual behavior (Hershfield et al., 2012). This could not be the case if (many) individuals actually low in Honesty-Humility strongly biased their answers upward and thus scored high on the scale. Likewise, this could not be the case if (many) individuals actually high in Honesty-Humility biased their answers downward due to their modesty. Similarly, those low in self-reported Honesty-Humility have been shown to be more likely to commit moral transgressions but concurrently less likely to be upfront about their wrongdoings as compared to those high in Honesty-Humility (Hilbig et al., in press). So, overall, it seems justified to conclude that (self-)reports of Honesty-Humility are valid and no more strongly affected by socially desirable responding than other factors. Nonetheless, still more fine-grained research aiming to tease apart Honesty-Humility and socially desirable responding is needed to ultimately test concerns about biases when assessing Honesty-Humility.

Future Directions

In light of the fact that Honesty-Humility has been established rather recently as a core personality dimension, it is unsurprising

that several issues will require additional attention in the next years. Here, we will suggest some potential future issues for the Honesty-Humility factor – clearly inspired by already existing preliminary hints or by corresponding research concerning other personality traits.

From an overarching perspective, research on Honesty-Humility could play a more important part in other disciplines, both within psychology and, more generally, in the social sciences. So far, most studies have been conducted through the lens of personality psychology and, though to lesser degree, with regard to ethical decisions and behavior, sociopolitical attitudes and behavior, and social-dilemma decision-making. By comparison, the Big Five are still the prime model of basic personality structure in other fields such as educational psychology, sport psychology, and clinical psychology. Nonetheless, the definition of Honesty-Humility and some of the extant empirical toe-holds indicate that this construct might also be crucial for important aspects within these fields. For instance, Honesty-Humility could predict bullying, long-term team success, or the success of interventions aiming to reduce aggressive or exploitative behavior. Thus, it seems a fruitful agenda to consider Honesty-Humility more thoroughly in other fields of psychology and beyond.

Other important aspects which are, as yet, clearly under-investigated concern the biological basis and development of Honesty-Humility. That is, the role of genetics (heredity) versus environmental influences (e.g., upbringing) for Honesty-Humility has not yet been investigated. On a related matter, the question of change over time and, if observable, its direction remains to be addressed. One might speculate, for instance, whether Honesty-Humility increases with age or whether particular contexts (e.g., one's social and/or work environment) actually influence individuals' levels of Honesty-Humility. Furthermore, research on (zero-acquaintance) personality judgments concerning Honesty-Humility is largely missing, that is, whether an individual's level in Honesty-Humility can, to some extent, be correctly inferred by strangers (as is the case for Extraversion, to name an example). Findings concerning these issues (i.e., genetics, development, and personality judgment) might further complement the existing notions on the evolutionary basis of the HEXACO model. From a methodological point of view, corresponding research would also imply a stronger reliance on different paradigms, study designs, and/or assessment methods (e.g., longitudinal investigations, twin and adoption studies).

Conclusions

Recently, Honesty-Humility has been recovered as a basic dimension of personality and research on this construct has been impressively fruitful. There is consistent support that Honesty-Humility reflects a core dimension of personality, that it uniquely predicts a broad range of behaviors which represent active cooperation, that it is related to criteria somewhat outside of the Big Five space, to deviant behavior, and/or to sociopolitical attitudes, and that it entails a specific pattern of person–situation interactions as implied by the HEXACO Model of Personality. These extant findings notwithstanding, further investigations will be needed to establish more comprehensive knowledge on Honesty-Humility. Given the indisputable importance of Honesty-Humility for many aspects of broad societal relevance, a still deeper understanding of its determinants and development, its interplay with contextual factors, and its consequences may contribute to facing future challenges inherent in human interaction between individuals, dyads, groups, societies, and even cultures.

Acknowledgment

Preparation of this manuscript was supported by grants from the German Research Foundation (DFG) to the first (ZE 948/1-1) and last (HI 1600/1-1) author. We thank Michael C. Ashton and Kibeom Lee for helpful suggestions and comments.

See also: Agreeableness; Big Five Factor Model, Theory and Structure; Conscientiousness; Cooperation and Competition; Dark Triad: The "Dark Side" of Human Personality; Factor Analysis and Latent Variable Models in Personality Psychology; Neuroticism; Openness to Experience; Personality Assessment, Faking and; Personality Assessment: Overview; Personality Assessment; Personality and Economics; Personality, Evolutionary Models of; Personality, Trait Models of; Political Psychology; Social Dilemmas, Psychology of; Social Dominance Orientation; Workplace Deviance.

Bibliography

Ashton, M.C., Lee, K., 2001. A theoretical basis for the major dimensions of personality. European Journal of Personality 15 (5), 327–353.

Ashton, M.C., Lee, K., 2007. Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. Personality and Social Psychology Review 11 (2), 150–166.

Ashton, M.C., Lee, K., 2008a. The HEXACO model of personality structure and the importance of the H factor. Social and Personality Psychology Compass 2 (5), 1952–1962.

Ashton, M.C., Lee, K., 2008b. The prediction of honesty-humility-related criteria by the HEXACO and Five-Factor models of personality. Journal of Research in Personality 42 (5), 1216–1228.

Ashton, M.C., Lee, K., 2009. The HEXACO-60: a short measure of the major dimensions of personality. Journal of Personality Assessment 91 (4), 340–345.

Ashton, M.C., Lee, K., De Vries, R.E., 2014. The HEXACO honesty-humility, agreeableness, and emotionality factors: a review of research and theory. Personality and Social Psychology Review 18 (2), 139–152.

Ashton, M.C., Lee, K., Goldberg, L.R., 2007. The IPIP-HEXACO scales: an alternative, public-domain measure of the personality constructs in the HEXACO model. Personality and Individual Differences 42 (8), 1515–1526.

Chirumbolo, A., Leone, L., 2010. Personality and politics: the role of the HEXACO model of personality in predicting ideology and voting. Personality and Individual Differences 49 (1), 43–48.

De Vries, R., 2013. The 24-item brief HEXACO inventory (BHI). Journal of Research in Personality 47 (6), 871–880.

De Vries, R.E., De Vries, A., De Hoogh, A., Feij, J., 2009. More than the Big Five: egoism and the HEXACO model of personality. European Journal of Personality 23 (8), 635–654.

Dunlop, P.D., Morrison, D.L., Koenig, J., Silcox, B., 2012. Comparing the Eysenck and HEXACO models of personality in the prediction of adult delinquency. European Journal of Personality 26 (3), 194–202.

Hershfield, H.E., Cohen, T.R., Thompson, L., 2012. Short horizons and tempting situations: lack of continuity to our future selves leads to unethical decision making and behavior. Organizational Behavior and Human Decision Processes 117 (2), 298–310.

Hilbig, B.E., Glöckner, A., Zettler, I., 2014. Personality and prosocial behavior: linking basic traits and social value orientations. Journal of Personality and Social Psychology 107 (3), 529–539.

Hilbig, B.E., Moshagen, M., Zettler, I. Truth will out: linking personality, morality, and honesty through indirect questioning. Social Psychological and Personality Science, in press.

Hilbig, B.E., Zettler, I., 2009. Pillars of cooperation: honesty-humility, social value orientations, and economic behavior. Journal of Research in Personality 43 (3), 516–519.

Hilbig, B.E., Zettler, I., Heydasch, T., 2012. Personality, punishment and public goods: strategic shifts towards cooperation as a matter of dispositional honesty-humility. European Journal of Personality 26 (3), 245–254.

Hilbig, B.E., Zettler, I., Leist, F., Heydasch, T., 2013a. It takes two: honesty-humility and agreeableness differentially predict active versus reactive cooperation. Personality and Individual Differences 54 (5), 598–603.

Hilbig, B.E., Zettler, I., Moshagen, M., Heydasch, T., 2013b. Tracing the path from personality – via cooperativeness – to conservation. European Journal of Personality 27 (4), 319–327.

Lee, K., Ashton, M.C., 2004. Psychometric properties of the HEXACO personality inventory. Multivariate Behavioral Research 39 (2), 329–358.

Lee, K., Ashton, M.C., 2006. Further assessment of the HEXACO personality inventory: two new facet scales and an observer report form. Psychological Assessment 18 (2), 181–191.

Lee, K., Ashton, M.C., 2008. The HEXACO personality factors in the indigenous personality lexicons of English and 11 other languages. Journal of Personality 76 (5), 1001–1053.

Lee, K., Ashton, M.C., 2014. The Dark Triad, the Big Five, and the HEXACO model. Personality and Individual Differences 67, 2–5.

Lee, K., Ashton, M.C., Pozzebon, J.A., Visser, B.A., Bourdage, J.S., Ogunfowora, B., 2009. Similarity and assumed similarity in personality reports of well-acquainted persons. Journal of Personality and Social Psychology 96 (2), 460–472.

Lee, K., Ashton, M.C., Wiltshire, J., Bourdage, J.S., Visser, B.A., Gallucci, A., 2013. Sex, power, and money: prediction from the Dark Triad and honesty-humility. European Journal of Personality 27 (2), 169–184.

Lee, K., Ogunfowora, B., Ashton, M.C., 2005. Personality traits beyond the Big Five: are they within the HEXACO space? Journal of Personality 73 (5), 1437–1463.

Pozzebon, J.A., Ashton, M.C., 2009. Personality and values as predictors of self- and peer-reported behavior. Journal of Individual Differences 30 (3), 122–129.

Thalmayer, A.G., Saucier, G., Eigenhuis, A., 2011. Comparative validity of brief to medium-length Big Five and Big Six personality questionnaires. Psychological Assessment 23 (4), 995–1009.

Zettler, I., Hilbig, B.E., 2010. Honesty-humility and a person-situation interaction at work. European Journal of Personality 24 (7), 569–582.

Zettler, I., Hilbig, B.E., Heydasch, T., 2013. Two sides of one coin: honesty-humility and situational factors mutually shape social dilemma decision making. Journal of Research in Personality 47 (4), 286–295.

Honneth, Axel (1948–)

Simon Thompson, University of the West of England, Bristol, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Axel Honneth is the leading light of the current, third generation of critical theorists. He has sought to refound a critical theory of society on a revivified version of Hegel's idealist philosophy, and, in particular, on the idea of recognition that he believes lies at its heart. According to Honneth's theory, there are three types of institutions – the family, law, and the market – in which three distinctive forms of recognition – love, respect, and esteem – may be experienced. For him, social justice is achieved to the extent to which all individuals have an opportunity to receive these various types of recognition in practice.

Honneth's Output, Academic, and Intellectual Achievements

Overview of Honneth's Output

Honneth is the author of a number of monographs, cowritten volumes, and collections of essays in a number of areas of critical theory, social philosophy, and the theory of recognition. Of his extensive oeuvre, the two most important works are the monographs *The Struggle for Recognition* (1995b) and *Freedom's Right* (2014). He has also written a number of shorter books, including *Reification* (2008) and *The Pathologies of Individual Freedom* (2010). In addition, Honneth has coauthored two volumes: *Social Action and Human Nature* (1988) with Han Joas, and *Redistribution or Recognition?* (2003) with Nancy Fraser. Several collections of his articles and essays have also been published, including *Disrespect* (2007) and *The I in We* (2012).

Honneth's first published book, cowritten with Hans Joas, was *Social Action and Human Nature* (1988; first published in German in 1980). Departing from Marxism, and discussing a range of contemporary thinkers including Nobert Elias, Michel Foucault, and Jürgen Habermas, this volume presents an argument for a 'philosophical anthropology' and for what Honneth and Joas call a 'historical anthropology.' These are regarded as two theoretical standpoints which seek to understand human nature as intersubjectivist in nature, displaying historical variations, but retaining a fixed core.

The Critique of Power (1991) and *The Fragmented World of the Social* (1995a) gather together essays, most of which were first published between 1982 and 1992, just prior to the publication of *The Struggle for Recognition*. In the first of these volumes, Honneth investigates the idea of a critical theory, and considers what form it might take in the contemporary world. He starts from Max Horkheimer's version of critical theory, but ends up examining the work of a number of contemporary theorists, including Foucault but above all Habermas. In the second of these volumes, Honneth again reflects on the nature of critical theory. In this book, he looks with increasing interest at potential French resources for such a theory. This time, his attention is directed not just to Foucault, but to Jean-Paul Sartre, Pierre Bourdieu, and others too. In these essays, it is possible to see the emergence of a number of themes that are of great importance to Honneth in his later work, including an attempt to refound critical theory in light of contemporary intellectual and social developments, and also the further development of an intersubjective account of human identity. Having said this, the theory of recognition with which Honneth is now strongly associated did not emerge clearly until the publication of *The Struggle for Recognition*.

First published in German as *Kampf um Anerkennung* in 1992, Honneth's *The Struggle for Recognition* (1995b) is the first key work in his oeuvre. Most of his writings over the 20 or so years since it was first published elaborate, defend, revise, and extend the theoretical framework articulated there. While it is true that the ideas which feed into Honneth's account of recognition can be seen in some of his earliest essays, the idea of a struggle for recognition as the leitmotif of his work seems to have emerged around 1990. In that year, he was awarded his Habilitation entitled '*Kampf um Anerkennung*,' and he published an important article 'Integrity and Respect: Principles of a Conception of Morality Based on the Theory of Recognition' (1992), which was initially given as a lecture in 1990. The theoretical framework which Honneth articulates in *The Struggle for Recognition* is discussed in more detail below.

In *Redistribution or Recognition? A Political-Philosophical Exchange* (2003), Honneth engages in an important debate with Nancy Fraser about the politics of recognition. In this exchange, both restate their own accounts of recognition, and both seek to refute each other's objections to their accounts. This book is important for understanding the scale of Honneth's ambition in claiming that all forms of injustice can be understood as types of misrecognition. He also seeks to show that his recognition-based theory of justice is not a form of ethical perfectionism, and nor it is one of deontological proceduralism. This part of his argument can be read as an extension of the idea of a 'formal conception of ethical life,' which he introduced in his *The Struggle for Recognition*.

Disrespect (2007) and *The I in We* (2012) are two collections of essays in which Honneth continues to finesse his theory of recognition. Most of the essays in the first of these collections were originally published from 1994 to 2000. The second collection brings together essays written between about 2004 and 2010. This latter book is divided into four sections, which, it could be argued, correspond to the major building blocks of Honneth's work: an interpretation of G.W.F. Hegel's philosophy of intersubjectivity, a defense of the idea of recognition as the foundation of a theory of social justice, an account of the 'paradoxical' developments of capitalism, and a distinctive

interpretation of object-relations psychoanalysis. *Pathologies of Reason* (2009) brings together essays first published between 2003 and 2007 on the German tradition of critical theory, including discussions and analyses of Theodor Adorno, Walter Benjamin, Sigmund Freud, Franz Neumann, and others.

Reification: A New Look at an Old Idea (2008) and *The Pathologies of Individual Freedom* (2010) are two relatively short books marking further developments of important aspects of Honneth's thought. The first of these volumes reprints the Tanner Lectures on Human Values, which Honneth delivered at the University of California, Berkeley in 2005. In these lectures, he seeks to revivify György Lukács' idea of reification by placing it in the framework of his theory of recognition. As a result, Honneth stretches the idea of recognition downward, as it were, below the level of reflexive ethical life to that of humans' most elementary and practical relationship to the world around them. *The Pathologies of Individual Freedom* is a short book – or a long essay – on the account of ethical life which Hegel presented in *The Philosophy of Right*. Against his earlier view, defended in *The Struggle for Recognition*, Honneth acknowledges here that an intersubjective account of recognition remains a key element in Hegel's thought from the early Jena writings of 1802–06 to the mature *Philosophy of Right* of 1821.

It is likely that *Freedom's Right* (2014) will mark a major development in Honneth's intellectual evolution. In this lengthy monograph, he presents a careful and detailed analysis of the ways in which freedom is grounded in the institutions of modern society. Continuing his earlier concerns, *Freedom's Right* draws on and reinterprets Hegel's idea of ethical life, especially as this is presented in *The Philosophy of Right*. This monograph also sees Honneth thinking more systematically about the democratic public sphere and the democratic state than he has done before. Since it was only published in English in 2014, it is impossible to know as yet what impact *Freedom's Right* will have on the various fields – including critical theory, social philosophy, theories of recognition, and so on – in which Honneth is influential. However, given his current importance, and the weightiness of this volume, it would seem safe to predict that its impact will be wide-ranging and profound.

Honneth's Academic and Intellectual Achievements

Critical Theory

Honneth regards himself first and foremost as a critical theorist in the tradition that began with the founding of the Institute for Social Research – or Frankfurt School – in 1923. He has considered the idea of critical theory on a number of occasions, seeking to determine what such a theory could and should look like today (see especially Honneth, 2009). Committed to the cause of human liberation or emancipation, critical theorists seek to show how particular social formations can act as obstacles to such liberation, and they do so by formulating their critiques in terms that are drawn from an account of those social formations themselves. To put this another way, critical theorists believe that there are resources in the present, identified by their critiques, out of which a better, more emancipated future can be created. It is this particular concatenation of beliefs, which gives critical theory what Fraser describes as "its distinctive dialectic of immanence and transcendence" (Fraser and Honneth, 2003: 202).

The dominant figures in the first generation of critical theorists were Adorno and Horkheimer. In their early work, these thinkers remained to some degree loyal to the Marxian belief that the proletarian class could and would act as the historical agent of human liberation. Later on, following a growing awareness about the nature of actually existing communism, Adorno in particular offered a much more pessimistic account of the prospects for human emancipation. In the second generation of critical theorists, Habermas is the figure who stands above all others. He sought to find a new way of justifying at least a cautious optimism about the possibility of a more emancipated future. He did so by developing a theory of 'communicative rationality,' which seeks to root universally valid moral claims in structures of human communication. Honneth, the leading light of the current, third generation of critical theorists, has sought to refound a critical theory of society on a revivified version of Hegel's idealist philosophy, and, in particular, on the idea of recognition which he believes lies at its heart.

Hegelian Roots

It follows from this that it is of vital importance to understand Honneth's relationship to Hegel. The closeness and intensity of this relationship started to become clear as Honneth's theory of recognition began to emerge around 1990. In *The Struggle for Recognition*, Honneth explicitly acknowledged his debate to Hegel (Honneth, 1995b: part I). At this stage, however, he concentrated on Hegel's Jena writings of 1802–06 rather than his later, better known work. Honneth did so because he was convinced that, with the publication of the *Phenomenology of Spirit* of 1806, Hegel had abandoned his insight into the central role of recognition in the intersubjective formation of human identity and social relations, and had replaced it with a transcendental philosophy at the heart of which lay the supra-individual entity of *Geist* (Honneth, 2007: 132–133). More recently, however, Honneth has changed his mind. Partly as a result of recent Hegel scholarship (e.g., Neuhouser, 2000), he now sees recognition as a central motif both of Hegel's *Phenomenology* of 1806 and of his *Philosophy of Right* of 1821 (Honneth, 2012: part I).

This reevaluation sees Honneth underline the significance of his debt to Hegel. In recent work, he has sought on several occasions to rework and resuscitate in particular Hegel's idea of ethical life (*Sittlichkeit*), and the idea of actualized freedom which stands at its core (see especially Honneth, 2010). For Hegel, ethical life is the system of customs and mores in which all individuals are embedded, and that shapes the ways in which they relate to one another. Ethical life has three principal parts. The family is the most elementary form of human sociality, governed by sentiment and feeling. Civil society is the realm in which individuals relate to one another as bearers of rights with whom they may chose to enter contractual relations. In the state, individuals regard themselves as part of a community (as they did unreflexively in the family), but they do so without losing the subjective freedom which they enjoy in civil society. Overall, the aim of Hegel's account of ethical life is to identify and specify the several sorts of material conditions which are jointly necessary to enable the

development of self-consciousness or the actualization of freedom. It will now be shown how, in Honneth's hands, these three aspects of ethical life are transmogrified into the three types of institutions in modern society – the family, law, and the market – in which three distinctive forms of recognition – love, respect, and esteem – may be experienced.

Idea of Recognition

To begin with, it is necessary to look more closely at Honneth's idea of recognition, and in particular to examine the various roles that it plays in his theoretical system. To identify several of the more important of these roles, recognition explains formation of individual identities, underlies the dynamics of social and political struggles, serves as the measure of social and ethical progress, and operates as a metric for social justice. In order to understand how it can do all this, it is important to understand that Honneth regards recognition as a relationship between two (or more) parties in which one positively acknowledges or affirms another (Honneth, 2012: 80). The first relationships of recognition are those between infants and their caretakers, but then they expand into ever wider circles of significant and then generalized others. In practice, of course, such positive acknowledgment cannot be guaranteed. From this perspective, Honneth suggests, struggles for recognition can be seen as collective attempts to move a society closer to one in which the right sort of relationships of recognition obtain, in which all individuals have roughly equal opportunities to achieve self-realization, and in which therefore social justice is achieved.

So far, reference has been made to the idea of recognition, and to relationships of recognition, *simpliciter*. But it is important to understand that for Honneth, recognition is a single genus with three distinct species (Ikäheimo, 2002), which he describes as love, respect, and esteem. These are all forms of positive acknowledgment, but each has its own internal logic and its own institutional location, which Honneth refers to as the family, the legal system, and the market, respectively. It is now necessary to consider each of these three modes of recognition in turn.

In his account of love, Honneth updates and reworks Hegel's account of the family. He does so by drawing on a form of object-relations psychoanalysis, that associated above all with the English psychoanalyst D.W. Winnicott (Honneth, 1995b: 95–107; 2012: part IV). From Winnicott and others, Honneth gets an account of the way in which the individual's identity is forged in the intense relationship between newborn infant and primary caretaker. Drawing on this psychoanalytical theory, Honneth seeks to 'de-transcendentalize' Hegel's account of the family, and to place it instead on a solid empirical footing. Seen from this psychoanalytical perspective, Honneth regards love as the "affectionate attention of concrete others" (1995b: 87). To be more specific, it is a relationship in which one person attunes themselves to another, and is in particular responsive to that other's needs. From the recipient's perspective, the experience of love enables them to acquire the confidence to articulate their needs with self-confidence and without shame (Honneth, 2003: 139). For Honneth, this mode of recognition is particularly significant since it is developmentally prior to respect and esteem. In other words, an individual must first experience love before they can experience either of the two other modes of recognition (Honneth, 1995b: 107, 176).

On Honneth's account, as we have just seen, love is initially and primarily experienced in the family. The individual's first experience of recognition is the love they receive (all being well) from their primary caretakers. Such recognition is then experienced in gradually widening circles of significant others, from siblings and other family members, to friends and other carers, and so on. Having said this, it is important to emphasize that, for Honneth, this expansion has strict limits. Love, in this specific sense at least, is limited to circles of significant others, and cannot be shown, for instance, to all of humanity or all of creation: it is "constituted by strong emotional attachments among a small number of people" (Honneth, 1995b: 95). It is worth noting, finally, that in more recent work Honneth concedes that the family is not a sphere purified of all forms of relationship except that of love. Instead he describes a 'tension' between affection and justice in which family members are united by relationships of love, but must at the same time seek to determine the moral obligations which they have to one another (Honneth, 2007: Chapter 7). In this way, Honneth acknowledges that there is not an exclusive one-to-one relationship between particular modes of recognition and particular social institutions.

In addition to love, recognition also takes the form of respect or what Honneth sometimes calls "legal recognition" (e.g., Honneth, 2003: 170). This mode of recognition is not precisely correlated with a single aspect of Hegel's ethical life. In one sense, respect is the sort of attitude that individuals in civil society take to one another when they regard each other as possessors of subjective rights. This mode of recognition is also related to Hegel's account of the state, in which individuals come to understand themselves as equal citizens subject to the authority of the same legal order. In Honneth's case, respect is regarded as a relationship in which one person acknowledges another's capacity for rational autonomy, and thus acknowledges their right to be in charge of their own life. In more recent work, Honneth has fleshed out this conception of respect a little further. Following Kant, he suggests that the person who respects another 'thwarts' their own 'self-love' in order to give due weight to the interests of that other (Honneth, 2012: 85; quoting Kant, 1785: 17–18). To respect another, that is to say, is to negate one's desires or to limit one's egoistic actions in order to take appropriate cognizance of or to take up an appropriate attitude to that other.

In practice, Honneth suggests, people treat each other respectfully by regarding each other as bearers of rights guaranteed to them by their state. For instance, when a person does not try to prevent another from speaking freely, they acknowledge the other's right to freedom of expression, and thus they respect them as a person who is capable of the responsible exercise of their ability to express their own opinions and to hear the opinions of others. Thus the institutional framework in which this mode of recognition is located is the legal system, and in particular in the system of subjective rights, to be found in the modern state. This having been said, Honneth is of course well aware that contemporary doctrines of human rights suggest that, rather than being restricted to the legal systems of particular states, the scope of such rights is universal. According to these doctrines, he writes, "it is assumed

that human beings themselves are capable of granting the universal rights that guarantee a human life with dignity for all" (Honneth, 2007: 209). Again this suggests that there is not a wholly exclusive relationship between this mode of recognition and a particular type of social institution.

The third and final mode of recognition which Honneth identifies is esteem. Once more, the relationship between this type of recognition and Hegel's account of ethical life is not straightforward. Esteem, for Honneth, is a mode of recognition experienced by participants in the market. Thus it may appear to be closest to the sort of attitude which individuals take to one another in Hegel's civil society. For Hegel, however, individuals in civil society act selfishly in pursuit of their own interests, and, while they may regard each other as bearers of subjective rights, they only contract with one another in order to advance their interests. For Honneth, by contrast, esteem is a relationship in which people do not merely receive their contractual dues, but are rather valued for their achievements or the contributions which they make to shared societal goals (Honneth, 1995b: 122). As he puts it, "in the system of market-mediated exchange, subjects mutually recognize each another as private autonomous beings who act for each other and thereby sustain their livelihood through the contribution their labour makes to society" (Honneth, 2012: 64).

According to Honneth, if these relationships of esteem are arranged as they should be, then esteem will be symmetrical, in the sense that "every individual is free from being collectively denigrated, so that one is given the chance to experience oneself to be recognized, in light of one's own accomplishments and abilities, as valuable to society" (Honneth, 1995b: 130). However, things rarely work out this well in practice. In 'developing capitalist' society, there was a 'one-sided, ideological valuing of certain achievements,' so that other sort of achievements – including those connected to 'household work' – were systematically undervalued (Honneth, 2003: 141). And, in the most recent 'neoliberal' phase of capitalist development, workers have been forced to become 'entreployees' who are expected to be self-motivated, flexible, and responsible for setting and achieving the work goals which they have set themselves (Honneth, 2012: 173–175). However, since the societal goals by which individuals' contributions are judged are subject to constant interpretation and reinterpretation, then, Honneth suggests, it is always possible for groups to struggle to be esteemed for achievements which were previously undervalued. The struggles of second-wave feminists to make women's housework and extradomestic labor suitably acknowledged and rewarded would be a good example of this phenomenon in practice.

Struggles for Recognition

These final remarks suggest that in practice relationships of recognition may be distorted, incomplete, or entirely absent. According to Honneth, the various sorts and degrees of misrecognition and nonrecognition thus experienced are regarded as profound injustices. From this perspective, struggles for recognition can be seen as attempts by groups of individuals to repair, restore, or create *de novo* relationships in which they receive appropriate acknowledgment. It is for this reason, to cite the subtitle of Honneth's *The Struggle for Recognition*, that the language of recognition provides "the moral grammar of social conflicts" (Honneth, 1995b). In other words, a wide range of social conflicts can be understood as struggles in which groups strive to obtain the acknowledgment from others which they believe is their due. In this way, Honneth naturalizes Hegel's famous account of the struggle for recognition between lord and bondsman in Chapter 4 of the *Phenomenology*. For Honneth, rather than regarding such struggles as necessary for the evolution of consciousness, they are instead seen as the motors of moral and social progress.

It is important to understand that, for Honneth, struggles over each mode of recognition unfold in their own way, since each contains its own distinctive developmental potential. This is least so in the case of love. Honneth has generally denied that this first mode of recognition can develop in the same way that both respect and esteem can. Having said this, in his contributions to Redistribution or Recognition?, he does hint that it may be possible for individuals to come to receive a more appropriate kind of love from their significant others. As he says, a "different or expanded kind of care" may be possible in which "newly developed or previously unconsidered needs" can be met (Honneth, 2003: 144). To a much greater degree, struggles for recognition can both widen the scope and deepen the content of respect and esteem. In both cases, recognition may develop along two distinct axes. So far as respect goes, struggles may result in more people being considered as eligible subjects for rights, and at the same time the content of those rights may become more extensive. T.H. Marshall's account of the evolution from civil through political to social rights is often cited in this regard (Marshall, 1950). In the case of esteem, there are once again two dimensions along which such recognition can develop: in the process of 'individualization,' unique individuals rather than social groups become the potential objects of esteem (Honneth, 1995b: 125), and, in the process of 'pluralization,' a stable hierarchy of esteem is gradually replaced by a 'horizontal competition' between values (Honneth, 1995b: 122).

It is important finally to understand that for Honneth the point of this analysis is not just to identify the characteristic lines along which struggles for respect and esteem may develop. It is also to show how such struggles lead to social and ethical progress, as society moves closer to a condition in which every one of its component parts receives appropriate acknowledgment. If the right sort of relationships of recognition are in place, then all individuals will have roughly equal opportunities to enjoy the acknowledgment which they need and deserve. The receipt of such recognition is a necessary condition of their having an opportunity to achieve self-realization, and it is for this reason that the justice of a particular regime of recognition is to be judged by its ability to put those conditions of recognition in place. As Honneth says: "The justice or well-being of a society is proportionate to its ability to secure conditions of mutual recognition under which personal identity-formation, hence individual self-realization, can proceed adequately" (Honneth, 2003: 174). Here we see clearly how the descriptive and prescriptive elements of Honneth's theory come together: recognition explains the formation of individual identity, and at the same time provides the metric for social justice.

How the Work Has Been Received and What Long-Term Impact It Will Have

How the Work Has Been Received

It has been suggested that Honneth is widely acknowledged to be the leading figure in the current, third generation of critical theorists. As Director of the Institute of Social Research since 2001, he has overseen a range of important and innovative research programmes, gathered together under the overall banner of 'Paradoxes of capitalist modernization' (http://www.ifs.uni-frankfurt.de/institut/). Honneth's influence is, however, not confined to Frankfurt narrowly or German-speaking world more widely, and his appointment to a chair at Columbia University in 2011 is one clear indication that his influence is now being felt worldwide. To put it in slightly passé terms, there is no doubt that Honneth's work transcends what used to be widely thought of as a divide between 'continental' and 'Anglo-American' philosophy. In a phrase, he is read not only by scholars of Hegel, but also by those of John Rawls, and there is no doubt it is his own profound understanding of both of these traditions which makes this possible. (See, for example, his essays on social justice (Honneth, 2012: Chapters 3, 7).)

It has also been suggested that Honneth's theory of recognition lies at the heart of his philosophical system. Hence it is not surprising that it is here that his influence has been at its strongest. Honneth is of course not the only important contemporary exponent of the theory of recognition. The continuing importance of Charles Taylor's seminal essay on 'The Politics of Recognition' cannot be overrated (Taylor, 1995). In addition, Fraser's strongly anti-Hegelian account of recognition (Fraser and Honneth, 2003) has served as an important challenge to Honneth's own Hegelian account. In addition, a number of other monographs on recognition have made significant contributions to the contemporary debate. Here Paul Ricoeur's *The Course of Recognition* (Ricoeur, 2005), in which he distinguishes between identification, self-recognition, and mutual recognition, deserves special mention. Finally, although it is not the principal focus of his work, it is arguable that James Tully's agonic conception of recognition (Tully, 2008) is another important rival to Honneth's Hegelian conception. All this having been said, there is no doubt that Honneth has set the agenda for contemporary debates about the philosophy and politics of recognition, and that all those wishing to write about this subject must write about him, if only to disagree.

There is, in addition, an extensive range of ways in which Honneth's theory of recognition has been applied in order to cast light on a wide variety of theoretical issues and practical problems. To give just one small example, in a volume on *The Politics of Misrecognition* (Thompson and Yar, 2011), contributors draw on Honneth's theoretical framework in order to discuss matters as diverse as racism, slavery, the unequal division of labor, disability, criminal harm, and leprosy. Beyond this one book, the list of topics to which Honneth's ideas have been applied could be extended much further to include the politics of identity and difference, multiculturalism, social movements, global justice, political economy, and so on.

What Long-Term Impact the Work Will Have

It is never easy to gauge the long-term impact of a particular thinker's work. Some drop suddenly out of intellectual fashion, while others enjoy unexpected revivals. This having been said, the steady and accelerating growth of interest in Honneth's work shows no sign of abating. Since his theoretical framework is constantly evolving, its continuing vitality would suggest that Honneth's existing readers will not stop reading him, and that many new ones will start. His theory of recognition, in particular, provides a highly flexible framework within which a wide variety of theoretical issues and practical problems can be usefully examined. The fact that it combines, among other things, an intersubjective account of the formation of individual identity, a framework for understanding the political struggles of social movements, an account of the nature of social justice, and a perspective from which capitalist labor relations can be evaluated, helps to explain why Honneth's influence is now widespread and profound, and why it will continue to be so for the foreseeable future.

See also: Critical Theory; Frankfurt School: Institute for Social Research; Habermas, Jürgen (1929–); Hegel, Georg Wilhelm Friedrich (1770–1831); Justice: Philosophical Aspects; Multiculturalism; Self: Philosophical Aspects; Social Movements, History of: General; Taylor, Charles (1931–).

Bibliography

Works by Honneth

Fraser, N., Honneth, A., 2003. Redistribution or Recognition? A Political-Philosophical Exchange. Verso Press, London.
Honneth, A., Joas, Hans, 1988. Social Action and Human Nature. Cambridge University Press, Cambridge.
Honneth, A., 1991. The Critique of Power: Reflective Stages in a Critical Social Theory. MIT Press, Boston.
Honneth, A., 1992. Integrity and respect: principles of a conception of morality based on the theory of recognition. Political Theory 20, 187–201.
Honneth, A., 1995a. The Fragmented World of the Social: Essays in Social and Political Philosophy. State University of New York Press, Albany.
Honneth, A., 1995b. The Struggle for Recognition: The Moral Grammar of Social Conflicts. Polity Press, Cambridge.
Honneth, A., 2007. Disrespect: The Normative Foundations of Critical Theory. Polity Press, Cambridge.
Honneth, A., 2008. Reification: A New Look at an Old Idea. Oxford University Press, Oxford.
Honneth, A., 2009. Pathologies of Reason: On the Legacy of Critical Theory. Columbia University Press, New York.
Honneth, A., 2010. The Pathologies of Individual Freedom: Hegel's Social Theory. Princeton University Press, Princeton.
Honneth, A., 2012. The I in We: Studies in the Theory of Recognition. Polity Press, Cambridge.
Honneth, A., 2013. Freedom's Right: The Social Foundations of Democratic Life. Polity Press, Cambridge.

Works by Others

Bankovsky, M., Le Goff, A. (Eds.), 2012. Recognition Theory and Contemporary French Moral and Political Philosophy. Manchester University Press, Manchester.
Burns, T., Thompson, S. (Eds.), 2013. Global Justice and the Politics of Recognition. Palgrave Macmillan, London.

Deranty, J.-P., 2009. Beyond Communication: A Critical Study of Axel Honneth's Social Philosophy. Brill, Leiden.
Hegel, G.W.F., 1806[1977]. The Phenomenology of Spirit. Oxford University Press, Oxford.
Hegel, G.W.F., 1821[1991]. Elements of the Philosophy of Right. Cambridge University Press, Cambridge.
Ikäheimo, H., 2002. On the genus and species of recognition. Inquiry 45, 447–462.
Kant, I., 1785[1959]. Fundamental Principles of the Metaphysic of Morals. Bobbs-Merrill, Indianapolis.
Marshall, T.H., 1950[1997]. Citizenship and Social Class. Pluto, London.
McBride, C., 2013. Recognition. Polity Press, Cambridge.
Neuhouser, F., 2000. Actualizing Freedom: The Foundations of Hegel's Social Theory. Harvard University Press, Harvard.
Petherbridge, D. (Ed.), 2011. Axel Honneth: Critical Essays. Brill, Leiden.
Petherbridge, D., 2013. The Critical Theory of Axel Honneth. Lexington Books, Lanham.
Ricoeur, P., 2005. The Course of Recognition. Harvard University Press, Harvard.
Schmidt-am-Busch, H.-C., Zurn, C. (Eds.), 2009. The Philosophy of Recognition. Historical and Contemporary Perspectives. Lexington Books, Lanham.
Taylor, C., 1995. The Politics of Recognition, in: Idem, Philosophical Arguments. Harvard University Press, Harvard.
Thompson, S., 2006. The Political Theory of Recognition: A Critical Introduction. Polity Press, Cambridge.
Thompson, S., Yar, M., 2011. The Politics of Misrecognition. Ashgate, Aldershot.
Tully, J., 2008. Public Philosophy in a New Key, Two Volumes. Cambridge University Press, Cambridge.

Honor and Shame

Charles Stewart, University College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Honor once defined northern Europeans, then it defined others on the margins of Europe, and now it is a problematic feature of others within Euro-American societies. This article traces honor and shame as paired concepts in social scientific thought beginning with their introduction by anthropologists of the Mediterranean in the 1960s. Men claimed honor through acts of bravado or violence that commanded social respect. Transgressions of a man's property, including his women, caused shame and required retaliation. Honor had been a familiar idea in medieval and early modern Europe, but it disappeared with the emergence of strong centralized states, and it became a fascinating anachronism in Mediterranean societies. Eventually it disappeared from the northern Mediterranean littoral only to resurface in the heart of modern Western societies among urban street gangs or in 'honor killings,' which have been on the rise in Europe and North America since the turn of the twenty-first century.

An historical perspective on honor and shame reveals this conceptual pair to be an othering label – a substandard ethos that Euro-American societies left behind with the advent of state judicial control. A strong whiff of what Fabian (1983) termed 'allochronism' hovers and makes contemporary honor and shame societies noncoeval with modern Western societies. The analysis of this situation in the first part of this article allows readers to understand what social scientists have understood by 'honor and shame' and where such social and cultural configurations could be found. This leads to a paradoxical observation. Whereas honor and shame had largely become extinct in Western societies by 1900, at the dawn of the twenty-first century they re-surfaced at the root of social panics over 'honor killings' in countries such as Sweden, Germany, and the USA. These acts of violence are interpreted as evidence of the unassimilated character of recent migrants from the Middle East and South Asia. 'Honor killings' drew upon a complex of ideas so foreign (or so anachronistic) that police and judicial systems in the West took quite some time to comprehend what was happening. Far from dying the death that seemed destined for them, globalization has returned honor and shame with new force. Social scientists must now consider how to understand and respond to honor and shame in a transnational situation where distant homelands can nourish these values and incite action in diasporas around the world.

Honor in Anthropology

Social anthropologists working in Mediterranean societies introduced the gender-linked concepts of honor and shame in the 1960s. Men claimed honor by staking out a reputation for aggressive machismo, while women submitted to their control with quiet modesty. Mediterraneanists treated honor and shame as a system of social values – very much in the Durkheimian tradition, as 'collective representations.' Honor comprised both an individual's sense of self-worth and this person's reputation in the surrounding community. Shame, by contrast, arose from the failure to act according to social values and it entailed public disgrace.

Anthropologists assumed that the honor and shame configuration was distinctively Mediterranean but its features can be found in many societies, especially in the Middle and Near East; from Morocco to India. The recognition of this wide dispersion, along with the demise of area studies, caused researchers to become more interested in local vocabularies and practices. Ethnographers ceased attempting to build a generalizable analytic model of honor and shame by the late 1980s. Historians of early modern Europe and America then became the social scientists most interested in investigating honor as an important social concept underpinning the social institution of dueling, for example. The concepts of honor and shame opened new insights for social and cultural historians wishing to explore the intimate topics of gender, sexuality, aggressive pride, self-regard, and public reputation.

By the middle of the twentieth century the word 'honor' was well on its way to obsolescence in ordinary English language usage. The practice of dueling to redress injuries to one's honor had ended and consequently honor had ceased to be a topic of public concern. The term was mainly frozen into stock expressions such as 'word of honor' or fixed in the aspic of formalized conventions, such as the Queen's Honors List, or the school honor roll. Honor was still a familiar word, yet people were not absolutely certain what it meant; it was rich in connotations, but poor in denotation.

The Oxford-trained anthropologist Julian Pitt-Rivers, scion of a venerable English family whose forebears may well have participated in the culture of honor in the nineteenth century, capitalized on the evocativeness of 'honor.' He reintroduced the word in a wide-ranging essay that drew principally on data from early modern European history. In that period honor drifted away from its medieval sense of a monarch's gift of fiefs in return for military services. By the seventeenth century it had come to mean a general social reputation and then in the early nineteenth century it took on a more interiorized sense of integrity or moral dignity. If publicly insulted – called a liar or cuckolded, for example – a man was within his rights to demand the satisfaction of a duel to restore his honor. Under

Napoleonic law a man could kill his wife with impunity if he caught her *in flagrante delicto*.

Pitt-Rivers used an amalgamation of the historical senses of honor to interrogate his own contemporary ethnographic data from a Spanish village. His essay appeared alongside studies of other Mediterranean societies in a book entitled *Honour and Shame: The Values of Mediterranean Society* (Peristiany, 1965) that launched honor and shame in anthropology. The premise of the volume was that "[M]editerranean peoples … are constantly called upon to use the concepts of honor and shame in order to assess their own conduct and that of their fellows" (Peristiany, 1965: p. 10).

For the most part honor and shame were conceived to be opposites and furthermore to orient different gender-linked ideals. Men were most likely to claim honor aggressively through acts of violence or cunning that earned them a fearsome reputation. This reputation was maintained by an ongoing performance of swagger and gruffness that staked out an inviolable space around a man's property, livestock, and women. As Peristiany noted (1965: p. 9), if provoked, "a Greek Cypriot, a Bedouin and a Berber may answer 'I also have a moustache'."

For their part, women protected this honor by comporting themselves with modesty. Among the Sarakatsani shepherds of northern Greece described by Campbell (1964: p. 287), women did not run for fear that they might fall over backwards and expose their underwear, thereby shaming themselves irretrievably. Ideally a young woman should stay at home working on her dowry, well away from the public gaze so that her name would not be on people's lips.

Comparative Considerations

The Mediterranean focus of the *Honour and Shame* volume spurred further ethnographic research within this 'culture area' and also attempts to theorize why such a similar set of values had been arrived at in this region. Virtually all anthropologists assumed that honor and shame were relatively consistent in form throughout the Mediterranean until Herzfeld (1980) pointed out, on the basis of research in two very different Greek communities, that honor amounted to 'socially appropriate behavior.' In a quiet, law-abiding village, macho swagger and aggressivity do not form the basis of honor.

The appreciation of the relativity of honor forced specialists into a theoretical stock taking (Gilmore, 1987). Two definable tendencies emerged: (1) honor and shame vary in form in different communities, and indeed may be found in completely non-Mediterranean societies, such as Japan where there is a great emphasis on the avoidance of shame; (2) although not exclusive to the Mediterranean, Mediterranean societies do, nonetheless, seem to converge on some strikingly similar cultural expressions of honor and shame. The gestural sign of the billy goat's horns made with extended index and baby finger, for example, signifies that a man has been cuckolded in a number of Mediterranean societies. The northern European resort to the image of the cuckoo bird to represent marital betrayal (hence 'cuckold') indicates the imagery of a different cultural area. In Mediterranean societies goats furnish the relevant vocabulary. A man who has been betrayed by his wife wears horns in Greek (*keratás*) and in Italian (*cornuto*) while in Spain he is just a big goat (*cabrón*). Unlike rams, billy goats allow their females to be mounted by other males without protest.

Performance: Class and Gender

In early modern Europe honor was mainly an affair of the aristocracy; slights from social inferiors could be ignored without consequences. The majority of early anthropologists of the Mediterranean worked in small rural communities where everyone was putatively a social equal. All men began with an equal portion of honor that they could increase or lose. A single system of honor and shame thus permeated the whole community. In larger, stratified towns or cities, however, there could be several parallel systems of honor by class. Honor itself was not economically determined, but the circles within which it was contested, and the modes of contestation, were.

There can be few societies lacking altogether in notions of precedence and social failure. The particular expressive performance of prestige in Mediterranean societies was surely what attracted the attention of anthropologists. These performances might involve elaborate verbal displays, such as the spontaneous exchange of insulting rhyming couplets, or the skilled telling of tales of sexual prowess or daring animal theft. The anthropologists recording these stories usually came from the middle classes of northern European or North American societies. Their fascination with these spectacularly different social practices precluded the recognition of their own participation in systems of honor (Pina-Cabral, 1989: p. 402). If transgressed a middle-class person might conceivably report the incident to the police and perhaps bring charges. If defamed in public, this person might register an accusation of libel and fight the case through the legal system rather than by direct, violent retribution. These practices protect personal reputation and arguably involve honor and shame, but it is not *the* system of honor and shame that Mediterraneanists had in mind.

Mediterranean honor assumed its particular form as a mode of direct action in the absence of moderating state institutions. Even where the law might be present, honor ignores its offices. The place to find similar configurations of honor and shame in many Western societies is often among the working classes or in criminal subcultures. American inner-city gangs have developed an honor culture that is perhaps even more 'Mediterranean' than that of Mediterranean societies as evidenced in the following statement by a Los Angeles Chicano gang member:

> Honor is important to me and every homey [gang member] because if you ain't got honor, you got nothing, man. I mean, no amount of money can buy you honor, so you do everything you can to protect it. I mean, without honor, nobody will respect you, not your friends or your family.
>
> (Jankowski, 1991: p. 143)

That middle-class anthropologists were apparently oblivious to these alternative systems of honor in their own societies exemplifies the earlier contention that parallel, discrete

codes for honorable comportment may operate in stratified societies.

Feminist studies of honor and shame have rejected the idea that honor is exclusive to men, and contended that women are not to be construed primarily as the passive defenders of male honor. Although their silence has been taken as a sign of submissiveness it might actually be a form of subversive irony through which women mock the supposed gender hierarchy. Furthermore, women's performances of suffering (their own pain, or on behalf of family members) can be understood as claims to respect in the community of fellow women sufferers. Dubisch (1995) has observed that Greek women engage in competitive suffering as a means of gaining prestige in the community. A gender studies model, then, where performances and social constructions of masculinity and femininity are treated independently, is increasingly replacing the honor and shame model, which linked them together in a version of normal patriarchy. Abu Lughod's (1986) study of honor and poetry among Egyptian Bedouins opened another dimension by contrasting honor as public performance of strength with the intimate vulnerability expressed in songs, thus situating honor and shame in the anthropology of emotions.

The Migration of Honor

Just as the initial optimism in the revelatory power of honor and shame began to wear out in anthropology, scholars in other fields discovered the formulation and adopted it with excitement. Where only a few books about honor and dueling in European history had been written before 1970, by the 1990s many titles had been published analyzing these themes in several European countries. In his *Southern Honor* (1982), Wyatt-Brown acknowledged the influence of Pitt-Rivers and speculated that the similarities between the American South and the Mediterranean arose because "primitive conditions of life often produce elemental social values sometimes roughly parallel from one society to another."

Like *mana* (to which Pitt-Rivers compared it), taboo, or shaman, honor belongs to middle-range anthropological theory – terms discovered ethnographically in one region and then applied to guide investigation in other societies. The Azande distinction between witchcraft (innate) and sorcery (learned), for example, has been used in the identification of witchcraft in numerous other societies. Honor and shame appear to have been discovered in Mediterranean anthropology and to have traveled out to other societies and other disciplines. Yet if we look more closely honor and shame had already traveled from early modern European history into Mediterranean anthropology via Pitt-Rivers' essay. That these concepts now seem more attractive to Europeanist historians than they do to Mediterraneanist anthropologists may well be because they actually arise directly from these historical societies. Honor and shame were indigenous concepts in early modern European society.

The future of honor and shame in Mediterranean studies is murky precisely because they do not emerge from local vocabularies. It is striking how few Mediterranean societies actually have words for 'honor' in the sense advanced by Pitt-Rivers and how infrequently equivalent terms are employed in daily life. In Greece the words *timí* 'honor' and *philótimo* 'sense of honor' are rarely heardwhereas in Italy the operative concept is neither shame nor honor, even though the latter would seem to be readily accessible in the Italian word *onore*. Instead, exemplary moral fiber is displayed by following 'the code of silence' (*omertá*) and not disclosing details of criminal or other in-group activities to inquisitive authorities. In his early monograph describing an Andalusian village, Pitt-Rivers (1954) focused attention on 'shame' (*vergüenza*) and made no mention of 'honor.'

The conception of honor brought into anthropology by Pitt-Rivers was always polysemous and this was its theoretical attraction. Honor could involve criminal aggression or Christian virtue, its forms extended into gender relations and equally into the politics and economics of patronage and mafia protection. By the 1990s anthropologists found that the vagueness of the terms honor and shame, and their mismatch with local concepts blunted their usefulness for ethnographic research and for comparison. Honor and shame looked set to be dissolved into the local terminologies uncovered by particularist ethnography, or absorbed into other frames of analysis altogether, such as psychological anthropological concepts of self and emotion.

Honor, Murder, and Multiculturalism

Honor did not, however, disappear. On the contrary, it acquired sharper definition than ever as the phenomenon of 'honor killings' began to command public attention in Western countries (Ewing, 2008). By 2010 an estimated 12 honor killings per year were occurring in each Germany and Britain with comparable statistics found in other countries with immigrant populations from areas such as Pakistan, Afghanistan, and Turkey. Police seminars, courtroom hearings, and even the European Parliament became venues where society attempted to understand the meaning of 'honor' in order to avert the damage it was causing (Wikan, 2008). Typically in these cases, kinsmen kill a young woman for pursuing a romantic relationship, or even just a Western lifestyle, which they oppose. Men may also be victims. The woman's alleged lover may be targeted, and overt homosexuality might also offer a pretext for murder. The meaning of honor has thus narrowed to refer to a kin group's protection of its reputation against taint from the sexual behavior of group members that is perceived as shameful.

In Britain, 17-year-old Shafilea Ahmed was murdered by her first-generation Pakistani migrant parents, who were sentenced to 25 years in prison in 2012. The motive for the murder might have been Shafilea's refusal of an arranged marriage during a family visit to Pakistan. In understanding them as honor murders authorities have come to grasp such crimes as corporate family decisions, condoned by a number of kin (including women). The implications for criminal charges of complicity or accessory to murder are enormous. Honor killings indicate a fatal clash between Western modernity with its recognition of individual equality and autonomy and non-Western traditions grounded in gender inequality and patriarchy. Migration to Europe has created a rift in values between generations. First-generation migrants commit the majority of honor killings

against second- and third-generation migrants who have become Westernized.

A Swedish woman of Kurdish descent, Pela Atroshi, was murdered in Iraqi Kurdistan and her father and uncle were sentenced to 6 months suspended sentences by an Iraqi court for the deed (Wikan, 2008: 39ff.). Upon return to Sweden, Pela's sister presented evidence that two of her uncles (themselves Swedish citizens) had been involved in the crime and a Swedish court sentenced them to life imprisonment. This case shows the mismatch between European laws and those of societies that recognize honor as a mitigating circumstance. Wikan raises the question that now hangs over many Western countries: "Where should a liberal state stand in dealing with illiberal immigrant groups?" (2004: p. 197). The matter has become a diagnostic test of the limit of cultural relativist, and anticolonial positioning.

Throughout its development as a concept, honor has been out of time or out of place, or both, and this has fueled its fascination for social scientists. At first it was a medieval idea surprisingly found in the contemporary Mediterranean. Then it was a Mediterranean system of values that illuminated historical European societies. Later it turned up in the inner cities of America rather than out there in the Mediterranean or Near East where it belonged. With the phenomenon of honor killings it has come to roost in the West, yet it is anomalous, perhaps symptomatic of historical transition. Honor is only unstably in place; brought into prosperous Western countries by immigrants yet many of these immigrants have become citizens. It is a temporal paradox because it has not been superseded and relegated to the European past. Honor has become, rather, a growing problem for the future of liberal plural societies, which are caught in the dilemma of how to respect cultural difference while maintaining fundamental values of individual freedom and self-determination.

See also: Culture and Emotion; Emotions, Psychological Structure of; Emotions, Sociology of; Guilt; Mafia, Anthropology of; Shame and the Social Bond; Tribe; Violence in Anthropology; Violence, History of.

Bibliography

Abu-Lughod, L., 1986. Veiled Sentiments: Honor and Poetry in a Bedouin Society. University of California Press, Berkeley, CA.

Campbell, J.K., 1964. Honour, Family and Patronage: A Study of Institutions and Moral Values in a Greek Mountain Community. Clarendon Press, Oxford.

Dubisch, J., 1995. In a Different Place: Pilgrimage, Gender and Politics at a Greek Island Shrine. Princeton University Press, Princeton, NJ.

Ewing, K., 2008. Stolen Honor: Stigmatizing Muslim Men in Berlin. Stanford University Press, Stanford, CA.

Fabian, J., 1983. Time and the Other: How Anthropology Makes Its Object. Columbia University Press, New York.

Gilmore, D.D. (Ed.), 1987. Honor and Shame and the Unity of the Mediterranean. American Anthropological Association Special Publication 22, Washington, DC.

Herzfeld, M., 1980. Honor and shame: problems in the comparative analysis of moral systems. Man 15, 339–351.

Jankowski, M.S., 1991. Islands in the Street: Gangs and American Urban Society. University of California Press, Berkeley, CA.

Peristiany, J.G. (Ed.), 1965. Honour and Shame: The Values of Mediterranean Society. Weidenfeld and Nicolson, London.

Pina-Cabral, J., 1989. The Mediterranean as a category of regional comparison: a critical view. Current Anthropology 30, 399–406.

Pitt-Rivers, J.A., 1954. The People of the Sierra. Weidenfeld and Nicolson, London.

Pitt-Rivers, J.A., 1965. Honour and social status. In: Peristiany, J.G. (Ed.), Honour and Shame: The Values of Mediterranean Society. Weidenfeld and Nicolson, London.

Wikan, U., 2004. Deadly distrust: honor killings and Swedish multiculturalism. In: Hardin, R. (Ed.), Distrust. Russell Sage, New York, pp. 192–204.

Wikan, U., 2008. Honor of Fadime: Murder and Shame. University of Chicago Press, Chicago, IL.

Wyatt-Brown, B., 1982. Southern Honor: Ethics and Behavior in the Old South. Oxford University Press, New York.

Honor Killing: Multidimensional and Multilevel Perspectives

Tenzin Dorjee and Stella Ting-Toomey, California State University Fullerton, Fullerton, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Honor killing is a crime and a social and intercultural moral dilemma in terms of understanding its motivations and possible prevention. This article presents multiple perspectives to illuminate the multifaceted dimensions of honor killing. Various contemporary ethical and theoretical perspectives on honor killing are addressed. The article is organized into four sections: honor killing as contextualized and defined; honor killing as a complex social, moral, and intercultural issue; social ecological and intercultural facework perspectives on honor killing, with an illustrative application; and finally, conclusions and future directions.

Defining and Contextualizing Honor Killing

'Honor killing or honor-related violence' is a multifaceted phenomenon that reflects a heinous social crime, a moral dilemma within societies, and an intercultural challenge. In general, while killing can be accidental, premeditated killing with malice is murder. From this perspective, honor killing is murder, typically, the murder of a female family member by her own family and/or friends to avenge shame or disgrace brought by infidelity or any other culturally unacceptable behaviors (Nasrullah et al., 2009). Women are the majority of victims of honor killing but men can suffer a similar fate (Idriss and Abbas, 2011; Welchman and Hossain, 2005). According to U.N. reports, every year as many as 5000 women are killed in the name of honor (Chesler, 2010; Smartt, 2006), which most observers think is an underestimate. The world-wide average age of female honor killing victims is 23 years. Men are also victims of honor killing (Chesler, 2010; Nesheiwat, 2005).

While honor killings can be found around the world, this article focuses on their occurrence within the cultural context of South Asia and the Middle East and addresses different aspects of honor killing from interdisciplinary perspectives. The key English terms to understand the motivations of this form of killing are 'honor' and 'perception.' 'Honor' is the English translation of the term '*izzat*' in Hindi–Urdu languages and cultures. '*Izzat*' is a broad term that includes honor, reputation, dignity, respect, social standing, and justice. In many cultures 'family *izzat*' guides social relationships and interactions (Encarta, 2009). All family members are expected to preserve and enhance family *izzat*. Social mobility from a low-status position to a high-status position via academic and professional achievement or matrimonial alliances, for example, a poor person marrying into a rich family, can enhance family *izzat*. Given the history of the caste system in India, social position is intimately related to one's *izzat* (Mandelbaum, 1988). While promoting family *izzat* would usually be commended, disgracing family *izzat* entails serious social consequences including social ostracism and even murder in the name of honor. The term 'honor killing' thus illustrates the dark side of *izzat*.

Honor killing is believed to restore family *izzat* (i.e., honor, dignity, reputation, respect, social standing, and justice). It is primarily a family *izzat* restoration strategy. Honor killing is based on a perceived disgrace or threat to the family *izzat*. For example, if a family member or friend believed that another family member had done something to disgrace her or his family (e.g., interreligious dating, disobeying the parents' wishes), the offending person would be punished accordingly in an authoritarian family system, including by death if the offense was sufficiently serious. Thus, the perception of bringing disgrace to the family is central to understanding honor killing or *izzat*-based killing (Carline, 2011; Perlmutte, 2011).

Dimensional Complexity of Honor Killing

Honor killing is a multidimensional phenomenon that requires interdisciplinary and social–ecological perspectives for breadth and depth of understanding. Below, we discuss the two major perspectives taken on the issue: conflict face-negotiation theory (FNT) (Ting-Toomey, 2005; Ting-Toomey and Kurogi, 1998) and the social–ecological framework (SEF) (Bronfenbrenner, 1977; Stokols, 1996; Ting-Toomey, 2011).

Honor Killing: The Sociocriminal Approach

Honor killing is often primarily approached as a social issue or crime and is addressed according to the legal constitution of a country (Idriss andAbbas, 2011). The main goals of this approach are social justice for the victims of honor killing and their families and prevention of honor killing in the first place (Idriss and Abbas, 2011; Welchman and Hossain, 2005).

Coomaraswamy (2005) offers a sociocriminal approach to understanding and preventing honor killing based on women's human rights and criminal law: "First, we need to ensure that we draw upon existing international standards and laws in upholding state responsibility. As Special Rapporteur, I called upon states not to invoke custom, tradition or religious considerations to avoid their obligations to eradicate violence against women and girls. …" (Coomaraswamy, 2005: pp. xii–xiii). She recommended developing national penal codes to provide protection to women irrespective of cultural practice. In this regard, scholars have investigated laws pertaining to honor killing in various national contexts such as Bangladesh, Pakistan, India, Lebanon, Jordan, Iraqi Kurdistan, Palestine, Egypt, Israel, Latin America, Italy, and United Kingdom

(see Welchman and Hossain, 2005) and other countries (see Idriss and Abbas, 2011). The laws of these different countries vary greatly in their provisions for social justice for the victims of honor killing or honor crimes. Women's rights proponents continue to advocate for the protection of women's rights, including the prevention of honor killing and full justice to its victims. A major challenge, of course, is to create uniform national penal codes prohibiting or criminalizing honor killing in nations and societies with incompatible moral values.

Honor Killing: A Moral Dimension

A moral order is the "theory by which a group understands its experience and makes judgments about proper and improper actions" (Pearce and Littlejohn, 1997: p. 51). From an intergroup perspective, moral conflicts that arise from issues such as honor killing are embedded in "incommensurate moral orders" (Pearce and Littlejohn, 1997: p. 54). These moral orders represent the most difficult forms of conflict to manage due to the multiple ways in which group members view being, knowledge, and moral values (Pearce and Littlejohn, 1997). In short, protagonists of honor killing strongly believe it to be the right course of action and represents moral justice (as a means of preserving traditional moral order), while antagonists strongly believe it is a morally reprehensible action and should be condemned and eradicated. Thus, multiple and competing ethical perspectives are brought to bear on the practice, and that creates a moral dilemma (Dorjee et al., 2013; Ting-Toomey, 1999, 2011).

Ethical absolutists believe in universal standards, notwithstanding cultural and contextual differences, and they strongly condemn honor killing as immoral action (Dorjee et al., 2013). In contrast, ethical relativists strongly believe that cultural contexts are vital to understanding different moral standards and they may suggest caution before condemning honor killing. While relativism promotes respect for all cultural standards and practices, it can perpetuate negative values and standards such as honor killing, *sati* tradition, and genital mutilation. Ethical universalists would condemn honor killing (and like practices) because they strongly believe in universal standards systematically derived from cross-cultural comparative standards and humanistic concerns. But it is extremely challenging to develop cross-cultural universal standards. Possibly, a metaethics position (Ting-Toomey, 1999, 2011) that transcends all other ethical positions can be used to understand honor killing. Such a perspective would look at honor killings case by case, always from a multilayered, contextual framework, which does not prejudge the morality of such actions. Two cases of honor killing may differ greatly due to multiple social, cultural, and contextual factors.

Honor killing posed a direct challenge to key metaethical concepts such as global social justice, identity equality and respect, human rights and freedom from pain and suffering (Dorjee et al., 2013). Importantly, how can we best ensure these metaethical issues are justly and globally respected and honored? One promising metaethical standpoint is found in the Dalai Lama's *Secular Ethics* approach. His understanding of secularism is informed by the constitution of India (Smith, 2011) wherein secularism is defined as *respecting all people regardless of faith or even no faith*. In his two classic books, the Dalai Lama (2001, 2011) persuasively argues that secular ethics is rooted in compassion – a sense of caring for others or warm heartedness – that we have inherited biologically from our mothers before we come to know of any religion or ideology. Secular ethics and the so-called 'middle way' approach can be useful in addressing many human conflict issues and cultural dilemmas (see Dorjee, 2013).

Honor Killing: Intercultural Conflict Dimension

Honor killing can also be viewed as an example of intercultural conflict, defined as the implicit or explicit emotional struggle or frustration between individuals of different cultures over perceived incompatible beliefs, values, norms, face-saving concerns, goals, scarce resources, processes, and/or outcomes in a communication situation (Ting-Toomey, 2005, 2011). From an intercultural or relational perspective, honor killing is directly related to face-saving concerns both within and between group community members. Face is an important aspect present in all social interactions. Its meanings, however, differ in specific cultures. The English term *face* (as used, e.g., in the term 'saving face') refers to a claimed sense of desired social image in a relational setting (Ting-Toomey, 1988). On the deeper level, face is a concept people use to gauge their own social self-worth and the social worth of others (Baig et al., 2014). Ting-Toomey's (2005, 2009) update of FNT explains the influence of cultural norms and values on people's behavior in different cultural communities and how face is managed in various conflict situations.

FNT recognizes three types of face concerns: self-face concern, other-face concern, and mutual-face concern. From cross-cultural perspectives, individualists are likely to be more concerned with self-face issues, but less with other and mutual-face issues. On the contrary, collectivists are likely to be more concerned with other and mutual-face issues, but less with self-face concerns. Since Hu's (1944) initial conceptualization about the meaning of Chinese face, several researchers have examined other Asian culture-specific meanings pertaining to the concept of face. For example, the Japanese term for face is *kao* (Haugh and Watanabe, 2009); and the Thai term *nâa* concerns face (Ukosakul, 2009). In Indian social–cultural contexts, *izzat* largely connotes communal face concerns. The Tibetan term *la* gya also refers to group-based face concerns such as community honor, ethnic honor, and national honor. These culturally specific explorations of face provide insight into how face concerns emerge through interaction as well as the various expected communicated rules in these cultures. Nayar (2004) explored the idea of *izzat* in intergenerational contexts with Punjabi Sikhs in Canada, and found that the value of *izzat* is "extremely strong and is expressed through the notion of behaving properly to save face, be it the family's or community" (p. 166).

The above cultural terms highlight the importance of maintaining honor or face in interpersonal and intergroup relationships and interactions. In conflict situations, individuals need facework management strategies and skills to effectively negotiate conflict goals including identity goals and relational goals. According to FNT, facework strategy consists of

two types: face giving strategy and face saving strategy (Lee and Park, 2011; Ting-Toomey, 2005). As social beings, all people want approval, respect, and acceptance. In this regard, it is vital to give others 'face,' i.e., by not humiliating others, acknowledging and respecting their identity, and communicating supportive messages in social interactions. However, when face is threatened or undermined, people employ face saving or restoration strategy. In minor face loss situations (such as not getting the expected jobs), individuals can engage in situational scapegoating attributions such as blaming all the foreigners coming in to take away the jobs. However, in major face loss situations, individuals may use sabotage strategies or even undertake drastic measures such as committing suicide or engaging in honor killing to restore face following deep face shame and humiliation (Perlmutte, 2011; Leung and Cohen, 2011; Scheff, 2011). In the case of the honor killing of Banaz Mahmod (see below) to restore family honor or izzat, the father authorized the killing of his own daughter. Societal, acculturational, situational, relational, and personal factors can profoundly shape *izzat* outlook and perspective (see Dorjee et al., 2013).

Conflict FNT in Multilayered Social–Ecological Contexts

We believe the conjoint framework of FNT and the SEF can provide the most meaningful explanation of honor killing.

Conflict FNT

Research on face and facework can be found in a wide range of disciplines such as anthropology, psychology, sociology, linguistics, management, international diplomacy, and human communication studies, among others. The concept of face has been used to explain linguistic politeness rituals, apology acts, embarrassment situations, requesting behaviors, and conflict interactions.

Ting-Toomey's (1988, 2005) conflict version of FNT explains the culture-based, individual-based, and situational-based factors that shape communicators' tendencies in approaching and dealing with conflicts. Based on the results of several large cross-cultural conflict data sets, a third version of the FNT appeared in 2005, and contained 7 premises or assumptions and 24 theoretical propositions.

Of the seven assumptions, the four relevant to this discussion are as follows: (1) people in all cultures try to maintain and negotiate face in all communication situations; (2) the concept of face is especially problematic in emotionally threatening or identity vulnerable situations when the situated identities of the communicators are called into question; (3) the value dimensions, in conjunction with individual, relational, and situational factors, influence the use of particular facework behaviors in particular cultural scenes; and (4) intercultural facework competence refers to the optimal integration of knowledge, mindfulness, and communication skills in managing vulnerable identity-based conflict situations appropriately, effectively, and adaptively.

When an individual's face image or a community's face image is being threatened in a conflict situation, she or he would likely experience identity-based frustration, emotional vulnerability, anger, hurt, and even a desire for vengeance (Zhang et al., 2012). Collectivists tend to emphasize the importance of communal face or *izzat* more highly than individualists. Individuals can experience face threats at both individual and group membership levels. In the 2005 FNT version, the following conditions were posited concerning the valence direction of an intercultural or intergroup *face threatening process (FTP)*. First, the more important, the culturally appropriate facework rule that has been violated, the more severe the perceived FTP. Second, the larger the cultural distance between the conflict parties, the more mistrust or misunderstanding accumulates in the FTP. Third, the more important the perceived conflict topic or imposition of the conflict demand, as interpreted from distinctive cultural angles, the more severe the perceived FTP. Fourth, the more power the conflict initiator has over the conflict recipient, the more severe the perceived FTP by the recipient and also the initiator. Fifth, the more harm the FTP produces, the more time and effort are needed to repair the FTP. Sixth, self-face or in-group community-face concern becomes incrementally more salient if several of these conditions are present in a FTP. For example, individuals are likely to move toward self-face saving and in-group face-saving emphasis if they perceive the escalation of the various face-threatening conditions directed at them or their salient in-groups. Cultural worldview perspectives, individual personality tendencies, and situational pressures frame the underlying interpretations of what count as a severe intercultural 'face-threatening' interaction episode. All five of these face-threatening conditions can aptly explain the intensity and perceived face-threatening level in multiple honor killing cases and in conjunction with the twisted, dark side of in-group face or *izzat*.

Multilayered SEF

In conjunction with the conflict FNT, the SEF provides us with a set of conceptual tools to sharpen our analysis from the macrolevel viewfinder to the microlevel cultural clash in an honor killing case.

Honor killing is not a random act of taking someone's life. It involves both intercultural/intergroup factors such as incommensurate moral orders, role membership expectations, institutional procedural breakdowns, and communication malfunctions to interpersonal attitudinal rigidity factors. Therefore, research that focuses on any one contextual level underestimates the effects of other contexts. For example, looking at honor killing as a mere violation of criminal law or strictly as a cultural issue would exclude insights gained from examining the interplay of macro- and microlevel factors in honor killing.

SEF is the study of people in embedded environments and the reciprocal influence between human behaviors and multiple environmental contexts (Bronfenbrenner, 1977; Stokols, 1996). Contemporary interdisciplinary research suggests the importance of a multilevel-theorizing approach to understand the complexity of human behavior (such as honor killing) in embedded contexts (Oetzel et al., 2006). A multilevel perspective enables us to understand (and possibly

challenge) the deeply held assumptions of a particular conflict approach, family/community reactions, institutional bureaucratic procedures, role expectancy violations, and cultural/religious worldview clashes.

According to Ting-Toomey and Oetzel (2013), SEP consists of five contextual levels of analysis: macrosystem, exosystem, mesosystem, microsystem, and chronosystem. The 'macrolevel' analysis focuses on the history, values, beliefs, and ideologies of a culture that influence interactants' conflict perspectives and behaviors. In comparison, the 'exolevel' analysis looks at the influence of established institutions and their formal procedures or policies on the community and individuals' reactions and actions. Meanwhile, the 'mesolevel' analysis examines the influences of immediate groups (e.g., family) and organizations (e.g., the workplace or religious groups) on the conflict parties and conflict episodes. The 'microlevel' analysis looks at both intrapersonal (e.g., identity issues and attributions) and interpersonal (e.g., the actual face to face or social media interactions) levels of interpretation and communication manifested in the actual conflict settings. The final level, the 'chronosystem level' refers to the developmental time span of the case history. Together, these multicontextual levels of analyses can provide a comprehensive understanding of diverse conflict situations – from everyday intercultural-intergroup conflicts to abusive family conflict events.

Furthermore, social ecological analytical concepts include three types of cross-level effects: (1) top-down effects, (2) bottom-up effects, and (3) interactive effects (Klein et al., 1999; Ting-Toomey and Oetzel, 2013). *Top-down effects* refer to how the larger cultural and institutional forces shape the intercultural conflict stance or practice – from the macro, the exo, the meso, to the microlevel. From this vantage point, individuals' conflict ideologies and practices are shaped by the various layered structures in which people are nested hierarchically (e.g., does the larger cultural worldview or institutional level emphasize communicating apology or forgiveness and does the effect spill downward to the lower levels?). *Bottom-up effects* focus on how lower levels (e.g., individuals and interpersonal relationships) affect higher levels (e.g., workplace, media news to institutional policies, and then cultural change). These types of effects are not as prevalent in the literature as are top-down effects, but they are no less important. *Interactive effects* involve simultaneous and mutual effects at more than one level. In some cases, the effects of one level (e.g., culture) moderate the outcomes at another level (e.g., family policies). Unlike top-down or bottom-up effects, interactive effects assume simultaneous process impacts at multiple contextual levels.

Conflict Face Negotiation and Social–Ecological Integrative Lens: A Theory–Practice Application Case

SEF enables the multicontextual level analysis of richly dense conflict cases such as the honor killing story presented below. In applying the analytical concepts of SEP in a systematic manner, we can explore the interplay of various complex-layered factors that frame honor killing, lead to deeper insights and suggest more compassionate ways to circumvent such heinous crimes in the future. Due to space limitation, we will present an excerpted honor killing story from the work of Dorjee et al. (2013; for a complete story rendition and detailed analysis, please consult Dorjee et al., 2013). The case amplifies intercultural conflict among the key characters: the victim – Banaz Mahmod; her father – Mahmod Mahmod; and her uncle – Ari Mahmod, The case also exemplifies the cultural clash of the mainstream U.K. community, the Kurdish community living in the UK, and the Mahmod family. An abridged story of her brutal murder is described below.

Banaz Mahmod was a 20-year-old, Iraqi-Kurdish female, living in Surrey, UK. She was murdered in her home in January 2006. Strangled with a bootlace, her body was stuffed into a suitcase. Several months later, the suitcase was found buried more than 100 miles away, under a house in a Birmingham suburb. Banaz's father (Mahmod Mahmod) and uncle (Ari Mahmod), along with three other accomplices, were arrested as murder suspects in the honor killing.

Following cultural practice, Banaz had been forced to marry an older man at the age of 17. She returned to her parent's home when her 'arranged' husband proved to be abusive and violent. Later she met Rahmat Sulemani, a young Iranian Kurd, and fell in love with him. A cousin saw Banaz kissing Sulemani outside the tube (train) station informed the father and uncle, who then decided to kill her for this offense.

Banaz learned from her mother about the death penalty and went to the police with the information, but she was evidently not persuasive and her situation was ignored. On 23 January 2006, Banaz also informed local police about the threats to Mr Sulemani's life and stated again the threat to her's. Although the police offered her a safe place to stay, she returned home believing that her mother would protect her from harm. The very next day, her father and uncle recruited three young Kurdish men to kill Banaz in the family home. Investigators found her body 3 months later on 28 April 2006.

Her uncle and father were charged with murder, while the three other men were charged with perverting the course of justice. The father and uncle were found guilty and sentenced to at least 20 and 23 years of life imprisonment, respectively. During the court hearings, it became obvious that Banaz was murdered to redeem the family honor. The prosecuting attorney told the court that her family decided that Banaz had shamed the family and should pay the ultimate price. The father was without remorse and clearly believed that the perceived loss of reputation was more important than his daughter's life.

Social Ecological Macrolevel and Exolevel Analysis

Macrolevel analysis provides insights into the historical, ideological, and cultural contexts related to the honor killing of Banaz Mahmod. Iraq, Iran, and Kurdistan have been involved in conflicts for decades (Hardy, 2005). The Kurds were caught in the middle of the Iran–Iraq war that lasted for 8 years; 100 000 Kurds who fled to Iran were killed. Against this backdrop, the Mahmod family escaped the Iraqi Hussein's regime and sought political asylum in the United Kingdom. They settled in UK and, as with most immigrants, tried to retain their native culture and traditions.

In particular, they adhered closely to the most repressive form of patriarchal ideology according to which women are not entitled to make any decisions about their own lives

(Idriss, 2011). In the case of Banaz and her sisters, patriarchal ideology (a macrolevel factor) strongly influenced the male family members' decision to kill her. Both ethical absolutism and ethical relativism can constrain our understanding of this murder case because the former involves ethnocentric culture-bound evaluation and the latter perpetuates honor killing in the name of an ethnorelative cultural viewpoint.

A metaethics perspective, however, elucidates the story. From this point of view, Banaz Mahmod had universal and civil rights (e.g., the right to life and justice, and the right to form voluntary social relationships) regardless of her cultural group membership. These rights were violated in her killing. Other factors involved her death are historical conflicts and religious differences. Banaz was an Iraqi and her suitor Sulemani was an Iranian (although they both were Kurds).

From the point of view of FTP, the following features of the killing are salient: First, this expectation of arranged marriages and fidelity within these arranged relationships is extremely strong, as is the patriarchical *motif* within Kurdish families. Maintaining *izzat* (honor, respect, and face concern) is what one does within a collectivistic Muslim cultural community. As Condition 1 in the conflict FTP states: First, the more important the culturally appropriate facework rule is violated, the more severe the perceived FTP. Condition 3 posits: The more important the perceived conflict topic (i.e., Banaz's running from her marriage and engaged on her own intergroup dating practice) or imposition of the conflict demand, as interpreted from distinctive cultural angles, the more severe the perceived FTP. Finally, Condition 6 in FTP states: Self-face or in-group community-face concern becomes incrementally more salient if several of these conditions are present in FTP. The perception of her father, uncle, and the *cultural* in-group was that Banaz had clearly violated important (even sacred) cultural norms; she defied patriarchy and fell in love with someone from the out-group. Death was the only appropriate punishment. Indeed, it can probably be said that none other was even contemplated. The topic of face-violation shame was perceived as so severe that her life had to be snuffed out prematurely. Concurrently, on the social–ecological exolevel analysis, the conflict FTP also posits in Condition 2: The larger the cultural distance between the conflict parties, the more mistrust or misunderstanding cumulate in the FTP. How the British police engaged in standard procedures to deal with the Banaz Mahmod's family conflict case events showed a wide gap of cultural distance and cultural insensitivity and blundering practices. At this level, established institutions, formal procedures, and their accompanying policies also contributed to the tragic murder of young Banaz. This highlights the importance of using the SEF to train all police officers to understand the cross-level effect between macrolevel/exolevel and microlevel conflict factors in exacerbating a life threatening case to the eventuality of a heinous murder outcome case.

Social Ecological Mesolevel and Microlevel Analysis

At this level, individual's behavior is influenced by immediate groups, community support or lack of it, and certain organizations. For example, the father was continuously harassed and rejected by others in his in-group community for his dereliction of duty as a man and a father when Banaz's elder sister Bekhal had absconded the parental home following the uncle, Ari Mahmod's threat to her life for talking to an unrelated man. Thus, to restore and also protect the honor, status, moral face of the family, the father decided for Banaz's murder. Indeed, the Daily Mail (2007) reported that Banaz's family and the traditional members of the Iraqi-Kurdish community in the UK believe that her family's actions were justified. This critical point also highlights the cross-level interactive effect between the meso- (i.e., the in-group community pressure and reaction) and microlevel (i.e., the actual honor killing conflict story) factors.

The microsystem social–ecological level of analysis comprised of both intrapersonal (i.e., personal and social identity issues) interpretations and interpersonal conflict communication processes as they played out over time. Three incidents stand out at this level in Banaz's story. The first is her forced marriage (at the age of 17) and her subsequent divorce from an abusive husband. This alone may have angered her family as they believed she 'brought shame' to the family via divorce. Second, her secret romantic relationship with Sulemani angered her family for multiple reasons, chief among them the apparent promiscuity that this suggested (their kiss at the exit of the train station was photographed by a cousin and presented to the father as incontrovertible proof of her insult to izzat). Third, due to concerns about community izzat and family honor, members of the Kurdish community were enlisted to help the family slaughter the young woman.

The dynamic of face concerns and pressurized negative reactions (both real and potential ones) from Mr Mahmod's in-group community seemed to have ruled out any option for constructive conflict negotiation or dialogue. Concurrently, police blunders and cultural insensitivity also contributed to the tragic death. The conjoint perspectives elucidated in this discussion contribute to a broader and deeper understanding of honor killing.

Conclusions and Future Directions

Honor killing, we have argued, is a complex, multifaceted cultural behavior that should be looked at from a conjoint ecological and social, psychological perspective. The SEF provides analytical tools to examine various social, cultural, and moral contexts and their interplay; FNT and the motif of *izzat* offer deeper explanations of the clash of intergroup 'facework violation' situations.

Honor is an element of face that involves the emotions of acceptance and rejection, pride and shame, punishment and vengeance, and restorative justice and forgiveness. Honor killing is therefore an ultimate, drastic face restoration move that cleanses izzat and restores in-group approval.

Six directions for future research on honor killing should be considered. First, what is the relationship between the various face concerns (i.e., self-face, other-face, mutual-face, and community-face) and the various dark facets of face emotions (i.e., shame, guilt, pain, hate, revenge, retribution, and redemption)? Second, how can we restore face imbalance when cultural or community face has gone beyond the tipping point? Third, how can we educate or train individuals in various public institutions (e.g., police, counsellors, social

workers) to be culturally face-sensitive and simultaneously be humanely astute in their situational appraisal of a cultural dilemma? Fourth, what would forgiveness look like in the context of such a heinous crime – how does face concern relate to intercultural-intergroup faith, hope, compassion, and restorative justice? Fifth, culture-sensitive face attunement, dialogue, and forgiveness are interdependent and intertwined. What culturally intelligent strategic facework actions at each chronosystem turning point could have changed the outcome? Finally, to whom should other family members (or perhaps, other in-group community resources, sacred symbols, or languages) appeal in times of cultural crisis that would defuse murderous rage?

Honor killing is an inverse mirror image of key metaethical concepts such as global social justice, interpersonal equality and respect, cultural sensitivity, individual and human rights, and the triumph of freedom from the conditions of human suffering. From a humanistic perspective, if we were to actively seek knowledge and understanding of 'unfamiliar others,' they would presumably feel accepted and thus, less likely to need to save face through honor killing. From SEF, institutional host members and immigrant individuals also need to come together to learn a community peace-building dialogue and language – in order to live with dreams, hopes, forgiveness, and ultimately, humanitarian renewal and faith.

See also: Children, Rights of: Cultural Concerns; Cultural Rights and Culture Defense: Cultural Concerns; Diversity and Disagreement in Ethics: Philosophical Implications; Domestic Violence: Sociological Perspectives; Ethics and Values; Global Violence Against Women; Identity and Identification; Individualism versus Collectivism: Philosophical Aspects; Integration and Multiculturalism: Focus on Western Europe; Justice, Conflicts, and the Justice of Conflict Resolution; Justice: Philosophical Aspects; Multiculturalism; Relativism: Philosophical Aspects; Shared Belief: Philosophical Aspects; Social Dominance Orientation; Value Pluralism.

Bibliography

Baig, N., Ting-Toomey, S., Dorjee, T., 2014. Intergenerational narratives on face: a south asian Indian American perspective. Journal of International and Intercultural Communication 7 (2), 127–147.

Barton, F., 12 June 2007. The Tragic Story of Banaz Mahmod: She Fell in Love at 19, so her Family Killed Her. Retrieved from: http://www.dailymail.co.uk/news/article-461378/The-tragic-story-Banaz-Mahmod-fell-love-19-family-killed-her.html.

Bronfenbrenner, U., 1977. Toward an experimental ecology of human development. Retrieved from: American Psychologist 32, 513–531. http://jmsw.org/hdf/facultystaff/Tudge/Bronfenbrenner%201977.pdf.

Carline, A., 2011. Honour and shame in domestic homicide. In: Idriss, M.M., Abbas, T. (Eds.), Honour, Violence, Women and Islam. A GlassHouse Book, Routledge, Abingdon, Oxon, pp. 80–95.

Chesler, P., 2010. Worldwide trends in honor killings. Middle East Quarterly 17, 3–11.

Coomaraswamy, R., 2005. Preface: violence against women and "crimes of honour,". In: Welchman, L., Hossain, S. (Eds.), 'Honour': Crimes, Paradigms and Violence against Women. Zed Books, London, pp. xii–xiii.

Court Sees 'Honour Killing' Video, 13 March 2007. BBC News. Retrieved from: http://news.bbc.co.uk/2/hi/uk_news/england/6446977.stm.

Daily Mail, 20 July 2007. Two-hour Rape and Torture of Honour Killing Girl Murdered by Her Father. Retrieved from: http://www.dailymail.co.uk./news/article-469553/Two-hour-rape-torture-honour-killing-girl-murdered-family.html.

Dorjee, T., 2013. Intercultural and intergroup conflict resolution: nonviolence and middle way approaches. In: Oetzel, J.G., Ting-Toomey, S. (Eds.), The Sage Handbook of Conflict Communication, second ed. Sage, Los Angeles, CA, pp. 687–711.

Dorjee, T., Baig, N., Ting-Toomey, S., 2013. A social ecological perspective in understanding 'honor killing': an intercultural moral dilemma. Journal of Intercultural Communication Research 41, 1–21.

Encarta, 2009. World English Dictionary. Izzat. Retrieved from: http://encarta.msn.com/dictionary_561536116/izzat.html.

Giles, H., Willemyns, M., Gallois, C., Anderson, M.C., 2006. Accommodating a new frontier: the context of law enforcement. In: Fielder, K. (Ed.), Social Communication. Psychology Press, New York, pp. 129–162.

Hardy, R., 22 September 2005. The Iran-Iraq War: 25 Years on. Retrieved from BBC News: http://news.bbc.co.uk/2/hi/middle_east/4260420.stm.

Haugh, M., Watanabe, Y., 2009. Analysing 'Face-in-interaction' in Japanese: insights from intercultural business meetings,'. In: Bargiela-Chiappini, F., Haugh, M. (Eds.), Face, Communication, and Social Interaction. Equinox Publishing, London, U.K, pp. 289–304.

Hu, H.C., 1944. The Chinese concepts of "face." American Anthropologist 46, 45–64.

Idriss, M.M., 2011. Honour, violence, women, and Islam – an introduction. In: Idriss, M.M., Abbas, T. (Eds.), Honour, Violence, Women and Islam. A GlassHouse Book, Routledge, Abingdon, Oxon, pp. 1–15.

Idriss, M.M., Abbas, T. (Eds.), 2011. Honour, Violence, Women and Islam. A GlassHouse Book, Routledge, Abingdon, Oxon.

Klein, K.J., Tosi, H., Cannella, A.A., 1999. Multilevel theory building: benefits, barriers, and new developments. Academy of Management Review 24, 243–248.

Lama, D., 2001. Ethics for the New Millennium. Riverhead Trade, USA.

Lama, D., 2011. Beyond religions: ethics for a whole world. Houghton Mifflin Harcourt, USA.

Lee, E.L., Park, H.S., 2011. Why Koreans are more likely to favor "apology," while Americans are more likely to factor "thank you." Human Communication Research 37, 125–146.

Leung, A., Cohen, D., 2011. Within- and between-culture variation: Individual differences and the cultural logics of honor, face, and dignity culture. Journal of Personality & Social Psychology 100, 507–526.

Mandelbaum, D.G., 1988. Women's Seclusion and Men's Honor: Sex Roles in North India, Bangladesh, and Pakistan. University of Arizona Press, Tucson.

Nasrullah, M., Haqqi, S., Cummings, K.J., 2009. The epidemiological patterns of honour killing of women in Pakistan. European Journal of Public Health 19 (2), 193–197.

Nayar, K.E., 2004. The Sikh Diaspora in Vancouver: Three Generations amid Tradition, Modernity, and Multiculturalism. University of Toronto Press, Toronto, Canada.

Nesheiwat, F., 2005. Honor crimes in Jordan: their treatment under Islamic and Jordanian criminal laws. Penn State International Law Review 23, 251–281.

Oetzel, J.G., Ting-Toomey, S., Rinderle, S., 2006. Conflict communication in contexts: a social ecological perspective. In: Oetzel, J.G., Ting-Toomey, S. (Eds.), The Sage Handbook of Conflict Communication: Integrating Theory, Research, and Practice. Sage, Thousand Oaks, CA, pp. 727–740.

Pearce, W.B., Littlejohn, S.W., 1997. Moral Conflict: When Social Worlds Collide. Sage, Thousand Oaks, CA.

Perlmutte, D., 2011. The semiotics of honor killing and ritual murder. Anthropoetics: The Journal of Generative Anthropology 17. http://www.anthropoetics.ucla.edu/ap1701/.

Scheff, T., 2011. Social-emotional world: mapping a continent. Current Sociology 59, 347–360.

Smartt, U., 2006. Honour killings. Justice of the Peace 170, 4–7.

Smith, D.E., 2011. India as a Secular State. Princeton University Press, New Jersey, USA.

Stokols, D., 1996. Translating social ecological theory into guidelines for community health promotion. American Journal of Health Promotion 10, 282–298.

Ting-Toomey, S., 1988. Intercultural conflicts: a face-negotiation theory. In: Kim, Y.Y., Gudykunst, W.B. (Eds.), Theories in Intercultural Communication. Sage Publications, Newbury Park, CA, pp. 213–235.

Ting-Toomey, S., 1999. Communicating across Cultures. Guilford, New York.

Ting-Toomey, S., 2005. The matrix of face: an updated face-negotiation theory. In: Gudykunst, W.B. (Ed.), Theorizing about Intercultural Communication. Sage Publications, Thousand Oaks, CA, pp. 71–92.

Ting-Toomey, S., 2009. A mindful approach to managing conflicts in intercultural-intimate Couples. In: Karis, T.A., Killian, K. (Eds.), Intercultural Couples: Exploring Diversity in Intimate Relationships. Sage, Thousand Oaks, CA, pp. 455–458.

Ting-Toomey, S., 2011. Intercultural communication ethics: multiple layered issues. In: Cheney, G., May, S., Munshi, D. (Eds.), The Handbook of Communication Ethics. Lawrence Erlbaum Publishers, Mahwah, NJ, pp. 335–352.

Ting-Toomey, S., Kurogi, A., 1998. Facework competence in intercultural conflict: an updated face negotiation theory. International Journal of Intercultural Relations 22, 187–225.

Ting-Toomey, S., Oetzel, J.G., 2013. Culture-based situational conflict model: an update and expansion. In: Oetzel, J.G., Ting-Toomey, S. (Eds.), The Sage Handbook of Conflict Communication, second ed. Sage, Los Angeles, CA, pp. 763–789.

Ukosakul, M., 2009. Significance of 'face' and politeness in social interaction as revealed through Thai 'face' idioms. In: Bargiela-Chiappini, F., Haugh, M. (Eds.), Face, Communication, and Social Interaction. Equinox Publishing, London, UK, pp. 289–304.

Welchman, L., Hossain, S. (Eds.), 2005. 'Honour': Crimes, Paradigms and Violence against Women. Zed Books, London.

Yousafzai, M., Lamb, C., 2013. I Am Malala: The Girl Who Stood up for Education and Was Shot by the Taliban. Little, Brown, and Company, USA.

Zhang, R., Ting-Toomey, S., Dorjee, T., Lee, P., 2012. Culture and self-construal as predictors of relational responses to emotional infidelity: China and the United States. Chinese Journal of Communication 5, 137–159.

Relevant Websites

www.bbc.co.uk/ethics/honourcrimes/crimesofhonour_1.shtml – BBC – Ethics – Honour Crimes.

www.opendemocracy.net/…/use-and-abuse-of-honour-based-violence-in – The Use and Abuse of Honour Based Violence in the UK.

www.amnestyusa.org – Violence Against Women Information – Amnesty International USA – 'Women's Rights' Violence against Women.

www.globalissues.org – Women's Rights – Global Issues – Issues, Articles.

www.who.int/mediacentre/factsheets/fs239 – WHO – Violence against women – World Health Organization.

Hooking up and Casual Sex

Paula England and Shelly Ronen, New York University, New York City, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Hookups' refer to casual sexual encounters that do not necessarily involve intercourse. While premarital sex has been common for decades, the hookup culture is distinctive in that sexual behavior is accepted in casual liaisons. Several forms of gender inequality in the hookup culture are identified: less attention is given to women's than men's sexual pleasure, women are judged more harshly than men for engaging in casual sex, and sexual coercion and assault are not uncommon. Sexual behavior does not differ appreciably between youth in different social classes, but more disadvantaged youth contracept less consistently and are more likely to have unplanned pregnancies. Media attention on hookup culture mirrors past feminist debates about sexuality, since the central question is whether hooking up is good or bad for women.

'Hookup' is the term that youth use to refer to casual sexual encounters. Hookups often begin in a social setting, such as a party, during which two students meet and something sexual happens between them that night. Those 'hooking up' are usually not strangers, and they do not necessarily have an interest in further contact or a romantic relationship when they have these mostly consensual encounters. Hookups can involve a range of sexual behaviors: from kissing or 'making out' on a party dance floor to touching, oral sex, or intercourse in private settings. The term has become widespread among students across public and private universities in the United States (Glenn and Marquardt, 2001; England et al., 2008). Sex outside of romantic relationships is also common among less privileged students who do not attend universities. Here, we focus largely on the growing literature on heterosexual hooking up.

Trends in Premarital and Casual Sex

Hookups involve premarital sexual behavior of some sort, not always intercourse. Formerly a taboo, intercourse before marriage has become largely accepted in many affluent societies. Highly religious subgroups teach that it is morally wrong, and their members often delay sex longer than others, but often end up having sex before marriage despite their belief against it (Regnerus, 2007). Approximately 90% of Americans born after 1960 have had premarital sex (Finer, 2007). Similarly, among French men and women born after 1967, only 11 and 7%, respectively, reported that their first intercourse was after marriage (Hubert et al., 1998: p. 55).

While premarital sex is noncontroversial for most, in the United States the higher rate of teen births than in other nations has led to substantial policy concern about teen sex. Defining 'teen' as those under 20, Hofferth et al. (1987: p. 49, Tables 3 and 4) use the National Survey of Family Growth, and report no clear trend in teenage intercourse between those born in 1938 and 1950, but for those cohorts born between 1950 and 1967, they find a clear upward trend for both black and white women in the proportion of unmarried teens who had experienced intercourse. Similarly, Laumann et al. (1994: p. 325), using data from the 1992 National Health and Social Life Survey, find no trend in mean age at first intercourse between black or white women born in 1933–42 and those born 1943–52, but find declines in mean age between the 1943–52 and 1953–62 birth cohorts. However, since the late 1980s, involving cohorts born since 1970, there has been a decrease in teen sex in the United States, perhaps in response to HIV. For example, in 1988, 51% of teens aged 15–19 years had had intercourse at least once, but among those at these ages in 2008, only 43% had had intercourse (Martinez et al., 2011).

The trend toward a higher proportion of youth having premarital sex does not tell us whether the bulk of premarital sex is experienced by couples planning to marry each other, couples in monogamous relationships without a plan to marry, or in more casual liaisons. Hookups, it is claimed, are a new phenomenon in that they are entirely outside even the modicum of commitment entailed in a boyfriend–girlfriend romantic relationship. Indeed, qualitative research on hookups claims that sometimes the sex comes first in a hookup and later some small proportion of the couples who hook up decide to be in a relationship (England et al., 2008). Of course, casual sexual behavior is not entirely new, as the decades-old use of the phrase 'one night stands' reveals. In fact, however, we lack systematic data on change over time in whether nonmarital sex is in casual or relational contexts. Yet, many authors suggest that the prevalence, visibility, and acceptance of casual sex have increased in recent decades (e.g., Garcia et al., 2012), leading to a hookup 'culture,' which implies that casual sexual behaviors are normatively accepted, at least within parts of the youth culture. It is perhaps more accurate to call it a subculture, since, for example, highly religious youth, especially those in conservative denominations, are less likely to hook up (Burdette et al., 2009), and many students have few if any hookups (England et al., 2008).

Behaviors Occurring in Hookups and Contexts in Which Hookups Occur

While hookups, as students use the term, always include some sexual behavior, they do not always involve intercourse.

England et al. (2008) report on data from the Online College Survey of Social Life, a survey of students from a number of colleges and U.S. universities. The survey asked students if they had hooked up in college, and instructed them to "use whatever definition of hookup you and your friends use." Of those responding students who were seniors, 76% had hooked up in college; the median number of hookups (retaining those with none in the analysis) was 5 (England et al., 2008: p. 533). Those who had hooked up were asked what happened sexually on their most recent hookup; among heterosexual students 31% 'made out' and/or did some nongenital touching, but had no genital contact, 16% had some hand/genital contact but did not have oral sex or intercourse, 15% had oral sex but not intercourse, and the remaining 38% had intercourse (England et al., 2008: p. 533). Thus, while the term 'hookup,' as used by students, always implies some physical contact, it does not necessarily entail intercourse, and hookups are quite diverse. Sometimes, students hook up two or more times with the same person. Of all the recent hookups students reported on, about half were the first time hooking up with this person, 11% were the second time, 8% were the third time, and so on with fewer for each successive time. Students were more likely to have intercourse on a hookup if they had hooked up with the person before (England et al., 2008: p. 534).

The term 'hookup' is in use among some high school students. Manning et al. (2006) asked secondary school students who had had intercourse whether they had done so with someone who was not a boyfriend or girlfriend; 61% of sexually active teens had done so. Most of these were not 'one night stands' and were not with strangers; rather they were typically with friends or with ex-boyfriends or ex-girlfriends.

Among university students, hookups can begin at social events such as parties. For those hookups in which only kissing or 'making out' occurs, they sometimes take place entirely at the social event, rather than in a private place. Armstrong and Hamilton's (2013) qualitative study of women's experiences at a US university identifies a 'party pathway' that some students follow. Women who follow this pathway most frequently are from relatively high socioeconomic status backgrounds; many of them join sororities, and few are very serious students. The university party scene is time-consuming, and sometimes comes into conflict with the academic portion of the university experience (Armstrong and Hamilton, 2013).

Parties are sites in which highly gendered norms prevail in initiations of sexualized dancing known as 'grinding' (Ronen, 2010), pairings that can lead to hookups. Sexualized dancing and making out on the dance floor may be a prelude to leaving to go somewhere private to pursue sexual activity. In addition, parties are sometimes the sites of sexual assault or rape. Parties are often hosted by fraternities in their residential houses; public spaces are dark, and music is loud, making conversations difficult. Partygoers seeking an alternative may be invited to fraternity members' bedrooms. Women are encouraged, if not compelled, to dress provocatively to go to parties: perpetuating gendered standards of appearance by which women end up more often cold, exposed, or uncomfortable. Gendered norms lead to expectations that women will be 'nice.' The combination of these factors at party settings sometimes leads to sexual assault (Armstrong et al., 2006).

Hookup Culture and Gender Inequality

Researchers have identified a gendered double standard in hookup culture, under which women are judged more harshly than men for casual sex or for having what is perceived as too many partners. Men do not face the same risks to their reputations as women (Sprecher et al., 1987; Crawford and Popp, 2003; Bogle, 2008; Kreager and Staff, 2009). They may even earn the distinction of 'player' or 'stud' for hooking up, especially within male peer groups. A woman who has many hookups is more likely to be known as a 'slut' or 'whore,' pejorative terms that may render her less attractive as a longer term partner, or girlfriend. Of course, the double standard is not a recent innovation, but has been present for centuries, and always entails applying some standard of what sexual behavior is unacceptable more stringently to women than men. However, the rise of casual sex or hookup behavior has shifted *what* sexual standard is applied differentially to men than women. When any sex before marriage was less accepted than it is now, condemnations of women focused on whether they had had premarital sex, with more leeway given to men. The rise of the hookup as a social form has changed the focus to number of sexual partners, and how readily one has intercourse in casual contexts. Some qualitative data suggest that this is now the focus of the sexual condemnation of women (Armstrong et al., 2012).

This cultural double standard is observed beyond students' attitudes about hooking up. It seems rooted in a cultural belief that men are more entitled to sexual agency and pleasure than women. This is one way to interpret the fact that, both in public and private, sexual behaviors prioritize men's pleasure over women's. The common form of dancing at college parties involves woman's buttocks 'grinding' against men's groins, a configuration best suited to produce men's genital stimulation and pleasure (Ronen, 2010). Moreover, reports of orgasm in hookups occurring in private show a large gender gap. Men have orgasms much more often in hookups than women; however, in sex in relationships, the gap is much narrower. Armstrong et al. (2012) show that women report orgasms in 42% of hookups with intercourse compared to 74% of relationship events involving intercourse. The authors draw on qualitative interviews with men and women to suggest that, while women concern themselves with men's pleasure in both relationships and hookups, many men readily admit that they focus only on their own pleasure, and devote considerably less effort to women's pleasure in hookups. This suggests that the notion of women's entitlement to pleasure is less accepted in the hookup than the relational context.

A final type of gender inequality in the hookup culture is sexual coercion or assault. Of course, coercive sexuality can and does happen in relationships as well as in hookups.

In the party scene, expectations that party attendants drink heavily and trust their partymates enhance the risks of sexual assault, as does women's frequent dependence on male party hosts for rides back to their own residences (Armstrong et al., 2006). Sexual assault is even more common in poor communities among youth not attending university (Miller, 2008; Jones, 2010).

Gender Differences in Preferences for Hookups versus Relationships

Conventional wisdom suggests that men are more interested in sex, especially casual sex with varying partners, while women prefer sex within monogamous relationships. Research backs some of this up, showing that women connect sex and love more than men, and women desire fewer sexual partners (Baumeister et al., 2001; McHugh, 2006). These observations would suggest that men would embrace the hookup culture more than women. Consistent with this, one qualitative study of university students showed that women talked more than men about preferring relationships to hookups (Bogle, 2008). Of course, like all gender differences, these preferences regarding the casual sex in hookups involve overlapping distributions, so some women enthusiastically embrace hooking up. In trying to explain the existing gender difference, some authors prefer an evolutionary argument (Baumeister and Vohs, 2004; Regnerus and Uecker, 2011). Regnerus and Uecker (2011) found that casual sex with a number of partners was associated with depression for female but not male youth and interpret this in terms of hard-wired sex differences in psychology. Others suggest that women, faced with the double bind of potential stigma, and seeing that hookups often do not work in their favor, prefer to get into relationships to be able to have sexual pleasure and a good reputation (Bogle, 2008; Backstrom et al., 2011). In other words, women's preference for relational sex may be, in part, a result of the double standard in which men do not respect casual partners and thus do not care about their pleasure (Armstrong et al., 2012).

Although women have orgasms more often in relational sex than in hookups, not all research supports the idea that relationships are better for young women. Based on their ethnographic and interview study of women students at a Midwestern 4-year public university, Hamilton and Armstrong (2009) conclude that women are more often harmed by bad relationships than by casual hookups gone awry. Women complained about boyfriends who tried to control the time they spent with friends. For those from upper middle-class backgrounds, relationships interfere with what the young women see as 'self-development projects,' or the main goals of college – schoolwork and meeting people. Relationships create a risk of getting drawn toward marriage too early, and for women from working class backgrounds, relationships with men back home sometimes pull them out of college. This suggests that hooking up may have emerged in part in response to the lengthening of the typical time in the life cycle between first sex and marriage.

Social Class Differences in Sexual Behavior and Its Consequences

In the United States, youth from more advantaged homes, as assessed by the education of their parents, have their first intercourse slightly later; the difference in timing is about a half year between a youth whose mother did not complete high school and one who is a college graduate (England et al., 2012). However, Wu and Martin (2013) show that neither racial nor socioeconomic differences in age at first intercourse explain much of racial and socioeconomic differences in the probability of having a nonmarital birth. The class difference that is more consequential is that advantaged youth and young adults use birth control more consistently. As a result, they are much less likely to become parents early, to have nonmarital births, or to have unintended births at any age (England et al., 2012; Musick et al., 2009). A number of explanations have been offered for the class differences in consistency of contraceptive use. Economists argue that young women who know that their educational and job trajectory will lead to good paying jobs have higher opportunity costs of having a birth that will derail this trajectory (Hotz et al., 1997). Another possibility is that class differences in parenting styles contribute to differences in having sufficient planfulness and self-regulation to achieve efficacy at contraception (England et al., 2012).

In the low-income context where birth control use is often inconsistent, casual sex is likely to lead to births that are not only outside of marriage, but are outside of any durable relationship. Qualitative work on those living in impoverished inner-city neighborhoods show that pregnancies are common in both casual and serious relationships. In both cases, couples often rally to try to raise the child together, although most relationships do not last longer than a few years (Edin et al., 2011; Edin and Nelson, 2013). A nationally representative sample of nonmarital urban births in the United States shows that approximately 80% of the parents are romantically involved at the birth, and half are cohabiting; however, fewer than a third remain together 5 years after the birth (McLanahan, 2011). The child is then generally raised by the mother, and often later co-resides with one or more of the mother's subsequent male partners (McLanahan, 2011).

Are there class differences in the propensity to have sex in casual versus relational contexts? Manning et al. (2006) examined the determinants of engaging in heterosexual intercourse within 'nonromantic' (by which they meant outside of girlfriend–boyfriend relationships) versus romantic dyads. They found that among secondary school students who had had intercourse, controlling for other variables, experience with nonrelational sex was more common for African Americans than those of other races, but was also more common for students whose mothers had higher education. Another possible indicator of whether sex is fairly casual is how early within a partnership it occurs. England et al. (2011) found no effect of a youth's mother's education on whether youth had sexual intercourse within 1 week of meeting a partner. Thus, it appears that the most important sex-related difference by social class is how much birth control is practiced rather than sexual practices themselves.

Attention to Hooking up by Journalists and Other Cultural Commentators

Perhaps galvanized by the new term, hookup culture has received attention in the popular press, which features debate over whether the phenomenon is good or bad, especially for women. Journalist Laura Sessions Stepp (2007) argues that women seem to have unwisely accepted hooking up even though it does them emotional harm. Feminist writer Jessica Valenti (2009) offers a critique of this perspective as promoting repressive understandings of feminine purity. What's harmful according to Valenti is not casual sex but a 'moral panic' about casual sex. The term 'moral panic' refers to frenzied public concern about social phenomena, which turns to volatile outcry that is disproportionate to the problem identified. Feminist writer Ariel Levy (2005) suggests a different critique of the hookup culture than that offered by Stepp. She provocatively asks whether the gendered and sexualized culture that characterizes parties where students hook up can really be empowering for women, who seem too concerned with status, or whether men will find them 'hot,' to consider their own sexual pleasure. In Levy's view, it is this 'raunch culture' of constant, excessive, and internalized sexualization that leads to women's lack of access to pleasure. These journalists and cultural commentators, while not exhaustive or entirely representative of commentators' views, are suggestive of the diversity of opinion in contemporary debates about hooking up in the media. Furthermore, they show that contributors to the debate consistently return to questions about gender inequality and whether and how hooking up might be detrimental to women (Armstrong et al., 2010).

A defining part of hookup culture is a separation of sex from romantic relationships, just as the sexual revolution of the 1960s followed by the early women's liberation and gay liberation movements separated sex from marriage. While hookup culture is not political like those social movements of the 1960s to 1980s, it yields some similar cultural consequences. Both the hookup culture and preceding sexual and feminist 'revolutions' liberalized sexual norms and lead to the acceptability of sex for pleasure. The controversies surrounding the hookup culture echo aspects of the feminist 'Sex Wars' of the 1980s, when feminists debated the legitimacy of sexual practices such as viewing pornography, lesbian sadomasochism, and commercial sex work (Samois, 1982; Lederer, 1980; Duggan and Hunter, 2006). The fundamental controversy centered on whether women's sexual agency led to empowered experiences of pleasure, or rather, whether certain types of women's sexuality spelled danger because they were shaped by male power and employed in the service of patriarchy (Vance, 1984). For example, in an unusual alliance with conservatives, some feminists argued that pornography is inherently patriarchal and harmful to women (Dworkin, 1989; MacKinnon, 1985, 1991; Russell, 1993). Against this alliance, self-described 'pro-sex' feminists argued for women's equal right to define and enjoy sexual pleasure, including pornography (Snitow et al., 1983; Rubin, 1993; Califia, 1994). Political and legal battles ensued around antipornography legislation, championed by feminists Andrea Dworkin and Catherine Mackinnon. According to their perspective, pornography enacted and promoted male violence against women, and so banning pornography would be a step in the direction of gender equality. Feminists opposed to the antipornography legislation argued that this logic misunderstood the sources of female sexual subordination and promoted traditional restrictions on women's sexual agency (Hunter and Law, 1987). Contemporary debates about hooking up suggest a similar divide between those who think casual sex can be pleasurable and those who see danger in hookup culture. Like the Sex Wars before them, contemporary debate is mired in an irresolvable quandary, which suggests that hookup culture offers some combination of both liberated access to pleasure and increased risk of danger.

See also: Adolescent Sexual Risk; Femininities; Heterosexual Expression at Older Ages; Heterosexuality in Africa; Heterosexuality in China; Heterosexuality; Intimate Partner Violence; Marital Quality and Health; Masculinity, Scripts of; Methodological Issues in the Study of Sexuality; Parallel Relationships; Popular Sexual Cultures in Western Europe (1930s to the Present); Satisfaction and Commitment in Long-Term Heterosexual Relationships; Sex Life Challenges: The Finnish Case; Sexual Behavior and Social Networks; Sexual Debut; Sexual Dysfunction, Epidemiology of; Sexual Motivation; Sexuality Education; Sexuality Over the Life Course; Sexuality and the Internet; Sexuality in Australia; Sexuality, Theories of; Sexually Transmitted Infections: Social Network Analysis.

Bibliography

Armstrong, Elizabeth A., Hamilton, Laura, 2013. Paying for the Party: How College Maintains Inequality. Harvard University Press.

Armstrong, Elizabeth A., Hamilton, Laura, Sweeney, Brian, 2006. Sexual assault on campus: a multi-level, integrative approach to party rape. Social Problems 53, 483–499.

Armstrong, Elizabeth A., Hamilton, Laura, England, Paula, 2010. Is hooking up bad for young women? Contexts 9 (3), 22–27.

Armstrong, Elizabeth A., England, Paula, Alison, C. K., Fogarty, 2012. Accounting for women's orgasm and sexual enjoyment in college hookups and relationships. American Sociological Review 77 (3), 435–462.

Backstrom, Laura, Armstrong, Elizabeth A., Puentes, Jennifer, 2011. Women's negotiation of cunnilingus in college hookups and relationships. Journal of Sex Research 49 (1), 1–12.

Baumeister, Roy F., Vohs, Kathleen D., 2004. Sexual economics: sex as female resource for social exchange in heterosexual interactions. Personality and Social Psychology Review 8, 339–363.

Baumeister, Roy F., Kathleen, R. Catanese, Vohs, Kathleen D., 2001. Is there a gender difference in strength of sex drive? Theoretical views, conceptual distinctions, and a review of relevant evidence. Personality and Social Psychology Review 5, 242–273.

Bogle, Kathleen A., 2008. Hooking up: Sex, Dating and Relationships on Campus. New York University Press, New York.

Burdette, Amy M., Ellison, Christopher G., Hill, Terrence D., Glenn, Norval D., 2009. 'Hooking up' at college: does religion make a difference? Journal for the Scientific of Religion 48, 535–551.

Califia, Pat, 1994. Public Sex: The Culture of Radical Sex. Cleis Press Pittsburgh, Pittsburgh, Pennsylvania.

Crawford, Mary, Popp, Danielle, February 2003. Sexual double standards: a review and methodological critique of two decades of research. The Journal of Sex Research 1, 13–26.

Duggan, Lisa, Hunter, Nana D., 2006. Sex Wars: Sexual Dissent and Political Culture. Taylor & Francis, New York.

Dworkin, Andrea, 1989. Pornography: Men Possessing Women. Plume, New York.
Edin, Kathryn, Nelson, Timothy, 2013. Doing the Best I Can: Fatherhood in the Inner City. University of California Press, Berkeley, CA.
Edin, Kathryn, Nelson, Timothy, Reed, Joanna Miranda, 2011. Daddy, baby: momma, maybe: low-income urban fathers and the 'package deal' of family life. In: Carlson, Marcia, England, Paula (Eds.), Social Class and Changing Families in an Unequal America. Stanford University Press, Stanford, pp. 83–107.
England, Paula, Shafer, Emily Fitzgibbons, Fogarty, Alison C.K., 2008. Hooking up and forming romantic relationships on today's college campuses. In: Kimmel, Michael S., Aronson, Amy (Eds.), The Gendered Society Reader, third ed., pp. 531–547.
England, Paula, McClintock, Elizabeth, Shafer, Emily Fitzgibbons, 2011. Birth control use and early, unintended births: evidence for a class gradient. In: Carlson, Marcia, England, Paula (Eds.), Social Class and Changing Families in an Unequal America. Stanford University Press, Stanford, pp. 21–49.
Finer, Lawrence B., 2007. Trends in premarital sex in the United States, 1954–2003. Public Health Reports 122, 7–78.
Garcia, Justin R., Reiber, Chris, Massey, Sean G., Merriwether, Ann M., 2012. Sexual hookup culture: a review. Review of General Psychology 16 (2), 161–176.
Glenn, Norval, Marquardt, Elizabeth, 2001. Hooking up, Hanging Out, and Hoping for Mr. Right: College Women on Mating and Dating Today. Institute for American Values, New York.
Hamilton, Lisa, Armstrong, Elizabeth A., 2009. Gendered sexuality in young adulthood: double binds and flawed options. Gender & Society 23 (5), 589–616.
Hofferth, Sandra L., Kahn, Joan R., Baldwin, Wendy, 1987. Premarital sexual activity among U.S. Teenage women over the past three decades. Family Planning Perspectives 19 (2), 46–53.
Hotz, V. Joseph, Alex Klerman, Jacob, Willis, Robert J., 1997. The economics of fertility in developed countries. In: Rosenzweig, Mark R., Stark, Oded (Eds.), Handbook of Population and Family Economics. North-Holland Elsevier Press, New York, pp. 275–342.
Hubert, Michel, Bajos, Nathalie, Sandfort, Theo, 1998. Sexual Behavior and HIV/AIDs in Europe: Comparisons of National Surveys. UCL Press, London.
Hunter, Nan D., Law, Sylvia A., 1987. Brief Amici Curiae of Feminist Anti-Censorship Taskforce, et al., in American Booksellers Association v. Hudnut. University of Michigan Journal of Law Reform 21, 69.
Jones, N., 2010. Between Good and Ghetto. Rutgers University Press, New Brunswick, NJ.
Kreager, Derek A., Staff, Jeremy, 2009. The sexual double standard and adolescent peer acceptance. Social Psychology Quarterly 72 (2), 143–164.
Laumann, Edward O., John, Gagnon, Robert, Michael, Stuart Michaels, 1994. The Social Organization of Sexuality. University of Chicago Press, Chicago.
Lederer, Laura, 1980. Take Back the Night: Women on Pornography. William Morrow & Company, New York.
Levy, Ariel, 2005. Female Chauvinist Pigs: Women and the Rise of Raunch Culture. Free Press, New York.
MacKinnon, Catharine A., 1985. Pornography, civil rights, and speech. Harvard Civil Rights/Civil Liberties Law Review 20, 10–68.
MacKinnon, Catharine A., 1991. Pornography as defamation and discrimination. Boston University Law Review 71, 793.
Manning, Wendy D., Giordano, Peggy C., Longmore, Monica A., 2006. Hooking up: the relationship contexts of 'nonrelationship' sex. Journal of Adolescent Research 21, 459–483.
Martinez, Gladys, Copen, Casey E., Abma, Joyce C., 2011. Teenagers in the United States: sexual activity, contraceptive use, and childbearing, 2006–2010 National Survey of Family Growth. National Center for Health Statistics. Vital Health Statistics 23 (31).
McHugh, Maureen C., 2006. What do women want? A new view of women's sexual problems. Sex Roles 54, 361–369.
McLanahan, Sara, 2011. Family instability and complexity after a nonmarital birth: outcomes for children in fragile families. In: Carlson, Marcia, England, Paula (Eds.), Social Class and Changing Families in an Unequal America. Stanford University Press, Stanford, pp. 108–133.
Miller, J., 2008. Getting Played: African American Girls, Urban Inequality, and Gendered Violence. New York University Press, New York.
Musick, Kelly, England, Paula, Edgington, Sarah, Kangas, Nicole, 2009. Education differences in intended and unintended fertility. Social Forces 88 (2), 543–572.
Regnerus, Mark, Uecker, Jeremy, 2011. Premarital Sex in America: How Young Americans Meet, Mate, and Think about Marrying. Oxford University Press, USA.
Regnerus, Mark, 2007. Forbidden Fruit: Sex and Religion in the Lives of American Teenagers. Oxford University Press, Oxford.
Ronen, Shelly, 2010. Grinding on the dance floor: gendered scripts and sexualized dancing at college parties. Gender & Society 24 (3), 355–377. http://dx.doi.org/10.1177/0891243210369894.
Rubin, Gayle, 1993. Thinking sex. In: Abelove, Henry, Barale, Michele Aina, Halperin, David M. (Eds.), Lesbian-Gay Studies Reader. Routledge, New York.
Russell, Diana E.H., 1993. Against Pornography: The Evidence of Harm. Russell Pub.
Samois, 1982. Coming to Power: Writings and Graphics on Lesbian S/M; S/M, a Form of Eroticism Based on a Consensual Exchange of Power. Alyson Publications.
Snitow, Ann Barr, Stansell, Christine, Thompson, Sharon (Eds.), 1983. Powers of Desire: The Politics of Sexuality. Monthly Review Pr, New York.
Sprecher, Susan, McKinney, Kathleen, Orbuch, Terri L., 1987. Has the double standard disappeared?: an experimental test. Social Psychology Quarterly 50 (1), 24–31.
Stepp, Laura Sessions, 2007. Unhooked: How Young Women Pursue Sex, Delay Love and Lose at Both. Riverhead Books (Hardcover).
Valenti, Jessica, 2009. The Purity Myth. Seal Press, Berkeley, CA.
Vance, Carole S. (Ed.), 1984. Pleasure and Danger: Exploring Female Sexuality. Routledge & Kegan Paul Books, London and New York.
Wu, Lawrence L., Martin, Steven, 2013. Sexual Onset and Premarital First Births: Insights from a Sequential Hazard Model. Unpublished paper, Department of Sociology, New York University, New York.

Hope and Hopelessness

Eddie MW Tong, National University of Singapore, Singapore, Singapore

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C.R. Snyder, volume 10, pp. 6907–6909, © 2001, Elsevier Ltd.

Abstract

This article provides a brief review of the literature on hope and hopelessness. The major approaches and measurements of hope and hopelessness are discussed. Differences between hope, hopelessness, and related states such as optimism and anxiety are briefly described. Approaches to increase hope and decrease hopelessness are outlined. The article also raises several questions for future research.

Hope is a virtue highly regarded by ancient writers and contemporary authors. It is featured in the famous myth of Pandora, who, upon opening a sacred jar, unleashed all forms of evils into the world, but found Elpis, the Spirit of Hope, remaining in the jar. Many religions, including Buddhism, Christianity, Hinduism, Islam, and Judaism, propounded hope-related beliefs. For instance, the Apostle Paul considered hope to be among the three greatest Christian virtues, alongside faith and love (1 Corinthians, 13: 13). Hope was also endorsed by many philosophers, including Plato, Sophocles, and Nietzsche, as a virtue or even a necessity. Even Napoleon is said to have once proclaimed: "A leader is a dealer in hope." Hope is also a regular theme in modern popular culture. For example, hope empowered the fight against tyranny in Star Wars and sustained the arduous journey to Mordor in the Lord of the Rings.

Many psychologists see hope as critical to psychological and physical well-being. For example, Snyder (2002) wrote that hope enhances physical and psychological health and plays beneficial roles in psychotherapy. Vaillant (2003) sees hope as important in assisting the elderly to adapt to aging. Hope has been found to positively predict interest and effort in studying, and learning (Pekrun et al., 2002). Hope compels one to remain committed to difficult goals, often in spite of insurmountable obstacles, instead of abandoning them. Hopelessness, however, not only causes one to abandon goals but also can lead to negative, sometimes devastating, consequences including depression and suicide. This article aims to give a brief review of the literature on hope and hopelessness, touching on the different theoretical approaches, measurement methods, and empirical evidence.

Approaches and Measures of Hope

One well-known approach posits hope as a cognitive expectancy with respect to specific goals. This approach has its roots in the 1950s and 1960s during which psychologists such as Stotland (1969) and psychiatrists such as Menninger (1959) defined hope as a positive expectation for goal attainment. More recently, Snyder (2002) significantly extended the cognitive approach through his theory of hope, which postulates hope as a cognitive, goal-oriented mindset. In his model, hope comprises two expectancies. The first concerns the belief that one is capable of executing the means to attain desired goals (agency thinking) and the second refers to the belief that one is capable of generating those means (pathways thinking). In this model, the combined experience of agency thinking plus pathways thinking gives the experience of hope.

Snyder and his colleagues developed the Hope Scale (with both trait and state versions) to assess agency and pathways thinking and, in a solid program of research, demonstrated that the measure reliably predicted numerous desirable outcomes, such as academic achievement, athleticism, physical well-being, and psychological adjustment (Snyder, 2002). Although agency thinking and pathways thinking are distinct constructs, they are represented by a single higher-order hope factor and thus can be combined as a unitary measure (Snyder et al., 1991).

In positing that hope is the combination of agency thinking and pathways thinking, Snyder (2002) argues that hope would not occur if either one of these expectancies is absent. His view has sparked considerable debate. Some researchers have pointed out that his model cannot explain why some people can remain hopeful even when they feel that there is nothing they can do to get what they want (Aspinwall and Leaf, 2002; Bruininks and Malle, 2005). Also, people often feel hope for others whose problems they cannot do anything to help (Averill et al., 1990; Bruininks and Malle, 2005). These findings suggest that hope can be felt without pathways thinking, which directly contradicted Snyder's view. Indeed, it was found that only agency thinking correlated positively with self-report ratings of hopefulness (Tong et al., 2010). Also, while acknowledging that the Hope Scale can predict important outcome variables, researchers have pointed out that the scale assesses goals, not necessarily hope. Finally, hope is widely seen as an emotion. Even if it does comprise agency thinking and pathways thinking, hope cannot be reduced to just these cognitive expectancies (Tennen et al., 2002).

An alternative approach conceptualizes hope as an emotion. For instance, Plutchik (1980) states that hope is a positive emotion characterized by a sense of anticipation. Fredrickson (2009) sees hope as a positive emotional reaction when things are going well and the future is hard to predict, and theorizes that hope broadens resources by enabling us to see the broader scheme of things. Pekrun (e.g., Pekrun and Stephens, 2010) posits that hope is an activating positive emotion, one that promotes task focus, intrinsic motivation, and flexible cognitive strategies.

Researchers in this approach also examine the adaptive functions and components associated with hope. Hope, like other emotions, is an adaptive response. It is associated with an adaptation function that has been evolved to facilitate effective navigation of specific environmental demands and comprises several components (e.g., subjective feeling, motivational urges, cognitive processes, physiological responses) that serve this adaptation function. This approach does not discount cognitive expectancies as an integral part of the hope experience, but conceptualizes them as among the several components that support its adaptation significance.

The adaptation function of hope is to keep a person focused an obstructed goal, instead of abandoning it (Lazarus, 1991). What could be the components of hope? Subjectively, hope appears to be, at best, moderately pleasant to experience (Smith and Ellsworth, 1985). This may not come as a surprise because the type of situations that typically evoke hope are those in which a valued goal has yet to be attained and formidable obstacles exist to prevent one from attaining it. Its motivational urge is to sustain focus on the unattained goal and fuel the expectation that the goal can be attained despite the uncertain prospect. Hope is also associated with a specific manner of perceiving events. A hopeful person tends to be highly cognizant of whether the current situation is favorable to achieving a goal, view the chances of attaining the goal as uncertain but not impossible, and appraise the situation as controlled by forces that are hard or impossible to control (Roseman et al., 1996; Smith and Ellsworth, 1985). A common thread in these findings is that, in contrast to Snyder's model, people often remain hopeful even when they know that there is nothing they can do to get what they want. In such cases, their hope appears to be sustained only by the sense that their goals can somehow be attained, even if they believe that they are incapable of attaining them (Aspinwall and Leaf, 2002; Averill et al., 1990; Bruininks and Malle, 2005; Tong et al., 2010).

These findings are typically based by methods that directly assess hope experiences. Some studies used the recall method in which participants recollected personal hope experiences and rated measures that assessed specific components (e.g., Bruininks and Malle, 2005; Roseman et al., 1996; Smith and Ellsworth, 1985). Other studies instructed participants to rate their current experience on simple but face-valid hope items such as "How hopeful do you feel now?" (e.g., Smith and Ellsworth, 1987; Tong et al., 2010). Such methods examine how hope is understood by laypersons, instead of predefining hope, which is the core of Snyder's approach. However, despite its apparent advantage in ecological validity, these methods have their share of problems. Recall data can be questioned because of memory concerns. The linguistic labels 'hope' and 'hopeful' can be understood differently by different participants (e.g., participants can interpret these labels to mean optimism, wishing, wanting, or other related constructs). This problem is exacerbated whenever there exists varying cultural interpretations of the labels. Further, there remains much research to be done on other components of the hope experience. It is not clear, for instance, whether hope has a specific and identifiable bodily expression. Does it carry a unique facial expression? Does it have an identifiable bodily posture and movement? Is it associated with a distinctive set of autonomic and neurological activation?

The third, new but promising, approach contains elements of both Snyder's cognitive approach and the emotion-componential approach. Scioli et al. (2011) argue that existing conceptualizations of hope focus too strongly on its goal-oriented function, and ignore the wider cultural and spiritual systems in which hope could be situated. They propose a multifaceted, contextualized view of hope that incorporates perspectives from nursing, philosophy, psychology, psychiatry, and theology. They postulate that hope is driven by four domains. First, in the domain of attachment, hope can be instilled by a supportive social environment (e.g., family, community). Accordingly, the hopeful individual can trust in the availability of a social other who endorses his/her endeavors and can assist him/her (Herth, 1991). Second, in the domain of mastery, hope is inspired by the sense that goals can be achieved and challenges can be overcome. This domain most resembles the goal-oriented function in Snyder's goal-setting framework. Third, the domain of survival captures hope in extreme situations – crises that involve loss, pain, and even death. In these circumstances, personal resilience is thoroughly tested and hope is often dependent not on personal strengths, but on social networks and religious beliefs. This domain addresses the point that some researchers have made, as outlined above, about hope being not related to just personal control. Finally, in the domain of spirituality, hope is inspired and sustained by spiritual or religious beliefs. It is served by religious activities (e.g., prayer), perceived spiritual presence, and a sense of union with something that is beyond us and yet part of us.

An important implication of this approach is that hope does not exist in a social or spiritual vacuum. While acknowledging that hope can be based on one's sense of control, this approach also proposes that hope can be evoked by an external system that includes the family, culture, and religious institutions. Hence, although one can find hope through personal strengths, hope is often enabled because of emotional or instrumental support from others and the sense of assurance or guidance from a higher spiritual power (see also Herth, 1991). This approach shows promise. State and trait measures of the four hope domains have been developed and preliminary data show acceptable psychometric properties (Scioli et al., 2011). But this approach is at the infancy stage of development, and many questions await investigation. For instance, the exact mechanisms by which social environments and religious institutions inspire hope remain unclear.

Definitions and Measures of Hopelessness

Hopelessness is widely defined as a negative cognition about the future. It is associated with a lack of enthusiasm, a motivational tendency to give up, and a dearth of positive expectancies (Beck et al., 1974). Based on this definition, the Beck Hopelessness Scale was developed. Other hopelessness measures have been developed for specific populations. For instance, based on the Beck Hopelessness Scale, Kazdin et al. (1983) developed and validated the Hopelessness Scale for Children (ages seven and older). Fry (1984) developed the Geriatric Hopelessness Scale to measure diminished abilities, lost interpersonal worth, and difficulty in recovering spiritual faith in the elderly.

The relationship between hopelessness, and depression and suicidal behavior has been the subject of a large body of research. The influential hopelessness theory of depression (Abramson et al., 1989) situates hopelessness as a proximal and sufficient cause of depression. The theory argues that the combination of cognitive vulnerability and negative life events causes depression. Cognitive vulnerability is the tendency to attribute negative events to stable and global causes, view them as likely to further produce negative effects, and construe the self is worthless. This cognitive bias is theorized to produce hopelessness, which, in turn, increases the risk of depression. There is strong empirical support that cognitive vulnerability interacts with negative events to predict depression and the process is mediated by hopelessness (e.g., Hankin et al., 2005; Metalsky and Joiner, 1992).

Research has found hopelessness to be a strong predictor of suicidal cognition and behavior (e.g., Beck et al., 1990; Beck et al., 1989). Some researchers have argued that hopelessness is a stronger predictor of suicidal cognition and behavior than depression is (e.g., Wetzel et al., 1980). There is also evidence that hopelessness mediates the relationship between depression and suicidal tendencies (Beck et al., 1990; Petrie and Chamberlain, 1983). Understandably, clinicians have advocated reducing hopelessness as a key intervention strategy to reduce suicide risk (Beck et al., 1990; Brown et al., 2005).

Therefore, there is striking evidence that hopelessness is a key precipitating factor of depression and suicidal tendency. More could be done to examine the conditions under which hopelessness leads not to suicidal responses and depression but to other negative outcomes. For instance, not everyone who feels hopeless about the future makes suicidal attempts; some might resort to violence to get what they want. Not everyone who feels hopeless experiences depression; some might feel unfair that their wishes have not been fulfilled and hence anger.

Similarities and Differences between Hope and Hopelessness, and Other Constructs

Hope is frequently compared to optimism. Research has indicated that hope and optimism can be subsumed under a single higher-order factor that reflects constructive future-focused thinking (Bryant and Cvengros, 2004). These two constructs do overlap to some extent. Optimism is considered to be a generalized expectation that the future will be positive (Scheier and Carver, 1985). This definition coincides considerably with Snyder's hope, which involves positive, goal-focused cognitive expectancy.

However, the two constructs are very different. In Snyder's (2002) view, optimism is the anticipation of attaining the desired outcome (outcome expectancy), whereas hope is the interplay between agency and pathways. Research has also differentiated these constructs. In one study, participants were asked to describe hope and optimism and their descriptions were coded (Bruininks and Malle, 2005). The results showed that participants perceived the optimistic person as having greater control over events than the hopeful person. They saw optimism-related events as more in one's favor than hope-related events. Hope was also perceived as lower than optimism in terms of the subjective likelihood of goal attainment. The results also suggest that optimism is more strongly associated with positive affect than hope is.

Hope is also different from challenge/determination. To be accurate, there are some similarities between them. In both constructs, attention is focused on an important but unrealized goal. Both emotions motivate perseverance. Also, measures of hope and challenge/determination are positively correlated (e.g., Smith and Kirby, 2009).

However, challenge/determination appears to be a stronger motivator than hope in compelling a person to learn and master a tough situation or skill (Lazarus et al., 1980; Smith and Kirby, 2009). As such, it is thought that a prime benefit of challenge/determination is personal growth and mastery. In contrast to hope, which, as mentioned above, has been found to be associated with a low sense of control over the goal situation, challenge/determination has been found to be associated with appraisals of high control and high coping potential (Smith and Kirby, 2011). Compared to hope, challenge/determination compels a person to invest more effort to attain the targeted goal.

Hope is also not wishing and fantasizing. Bruininks and Malle (2005) found that in wishing and fantasizing, one perceives even less control over the goal situation compared to hope. Also, the desired outcome one fantasizes about tends to be perceived as unlikely or impossible to materialize. Hence, a person who simply wishes or fantasizes is likely to abandon the goal. In contrast, hope induces a stronger commitment to the goal, even if the person feels that he or she has little control over the situation.

Wanting and desiring are similar to hope in terms of the strong focus on important goals. However, Bruininks and Malle (2005) found that wanting and desiring, more than hope, occur when the outcome is appraised as highly possible to occur. In contrast, situations that elicit hope tend to be those in which the goal is appraised as difficult (but not impossible) to obtain and in which the circumstances are deemed as difficult to control (Roseman et al., 1990).

Hope bears some similarity to self-efficacy. Snyder's pathways and agency thinking seem similar to Bandura's (1997) outcome expectancy and efficacy expectancy, respectively. However, according to Snyder (2002), pathways thinking reflects the belief of whether the self has the overall capability to generate the routes to goals, but outcome expectancy refers to whether one can generate solutions specifically for the current problem, and agency thinking reflects the intention that one *will* perform the goal-approaching actions but efficacy expectancy refers to the belief about whether one *can* perform these actions. Also, self-efficacy thoughts are accorded the pivotal driving role in Bandura's theory, whereas agency and pathways are given equal weight in Snyder's model.

Hopelessness and anxiety share some features. Both are highly distressful and are associated with negative evaluation of the future. Both are associated with a low sense of control over the situation. However, according to Pekrun and Stephens (2010), the main difference between them lies in anxiety being an activating state (i.e., physiologically activating, high arousal) and hopelessness being the opposite, deactivating state. More research is needed to clarify the differences between these two constructs.

Increasing Hope and Decreasing Hopelessness

A point shared by the different approaches outlined above is that hope occurs in response to goals. Hence, what goal is being pursued can influence the amount of hope one can feel. To this end, it is important to set realistic goals. It is also important to set goals that one wants rather than those that other people want. Other strategies include producing an array of goals, setting goals that stretch one's talents, and setting concrete and achievable goal markers (Nezu and D'Zurilla, 1981).

According to Snyder (2002), hope can be increased by two channels – agency and pathways. Strategies that enhance the sense that one is able to do things that bring about the desired outcome should increase agency. These include recalling previous incidents of successful goal attainment, focusing on attainable goals, positive self-talk, anticipating and viewing impediments as challenges, and even engaging in energy-boosting activities such as exercise. Strategies that help a person to improve his or her ability to generate multiple routes to goal attainment can increase pathways. These include breaking a goal into smaller attainable steps, developing mental maps to goals, generating different solutions to problems, learning and mastering particular skills, and developing supportive social networks.

Mirroring the hopelessness theory of depression, Needles and Abramson (1990) propose the recovery model of depression to fight hopelessness. The recovery model proposes that an enhancing inferential style over positive events should increase hopefulness and also reduce depressive symptomatology. This inferential style involves attributing positive events to global and stable causes, inferring positive consequences, and believing in one's self-worth. Numerous studies have found support for this model (e.g., Johnson et al., 1998; Needles and Abramson, 1990).

Conclusion and Caveats

Among the many ways that make humans unique as a species is the ability to imagine. This ability underlies hope because hope is about envisioning a better future. However, like other positive emotions, there are downsides to hope. People can hope for outcomes that are unlikely to occur (Averill et al., 1990). Hope can degenerate into wishing thinking, in which one fails to see that the odds are impossible to overcome and that the goal is impossible to attain, and yet continues to pursue it or waits for it to materialize. Also, it is one thing to feel hopeful for the moment, but another to remain hopeful over a long period of adversity. Techniques that enable a person to remain hopeful over a period and difficult period are in need. Nevertheless, hope is a critical emotion because it sustains people when they feel helpless, enabling them to focus on their dreams despite what others perceive to be crippling obstacles.

Acknowledgment

This article is based in part on the previous contribution by Dr. Charles R. Snyder. It is an honor to be given the opportunity to build on his work. This article is written in memory of him and a tribute to his excellent contribution to research on hope.

See also: Affect-Regulation Motivation; Coping, Personality and; Emotion, Perception and Expression of; Emotional Regulation; Emotions: Methods of Assessment; Optimism and Pessimism; Self-Determination Theory; Self-Regulated Learning; Suicide.

Bibliography

Abramson, L.Y., Metalsky, F.L., Alloy, L.B., 1989. Hopelessness depression: a theory based subtype of depression. Psychological Review 96, 358–372.

Aspinwall, L.G., Leaf, S.L., 2002. In search of the unique aspects of hope: pinning our hopes on positive emotions, future-oriented thinking, hard times, and other people. Psychological Inquiry 13, 276–288.

Averill, J.R., Catlin, G., Chon, K.K., 1990. Rules of Hope. Springer-Verlag, New York.

Bandura, A., 1997. Self-efficacy: The Exercise of Control. Freeman, New York.

Beck, A.T., Brown, G., Berchick, R.J., Stewart, B.L., Steer, R.A., 1990. Relationship between hopelessness and ultimate suicide: a replication with psychiatric outpatients. American Journal of Psychiatry 147, 190–195.

Beck, A.T., Brown, G., Steer, R.A., 1989. Prediction of eventual suicide in psychiatric inpatients by clinical ratings of hopelessness. Journal of Consulting and Clinical Psychology 57, 309–310.

Beck, A.T., Steer, R.A., Kovacs, M., Garrison, B., 1985. Hopelessness and eventual suicide: a 10-year prospective study of patients hospitalized with suicidal ideation. American Journal of Psychiatry 142, 559–563.

Beck, A.T., Weissman, A., Lester, D., Trexler, L., 1974. The measurement of pessimism: the hopelessness scale. Journal of Consulting and Clinical Psychology 42, 861–865.

Brown, G.K., Ten Have, T., Henriques, G.R., Xie, S.X., Hollander, J.E., Beck, A.T., 2005. Cognitive therapy for the prevention of suicide attempts: a randomized controlled trial. Journal of the American Medical Association 294, 563–570.

Bryant, F.B., Cvengros, J.A., 2004. Distinguishing hope and optimism: two sides of a coin, or two separate coins? Journal of Social and Clinical Psychology 23, 273–302.

Bruininks, P., Malle, B.F., 2005. Distinguishing hope from optimism and related affective states. Motivation and Emotion 29, 327–355.

Fredrickson, B.L., 2009. Why choose hope. Psychology Today.

Fry, P.S., 1984. Development of a geriatric scale of hopelessness: implications for counseling and intervention with the depressed elderly. Journal of Counseling Psychology 31, 322–331.

Hankin, B.L., Fraley, R.C., Abela, J.R.Z., 2005. Daily depression and cognitions about stress: evidence for a trait-like depressogenic cognitive style and the prediction of depressive symptoms in a prospective daily diary study. Journal of Personality and Social Psychology 88, 673–685.

Herth, K., 1991. Development and refinement of an instrument to measure hope. Scholarly Inquiry for Nursing Practice 5, 39–51.

Johnson, J.G., Han, Y.S., Douglas, C.J., Johannet, C.M., Ruseel, T., 1998. Attributions for positive life events predict recovery from depression among psychiatric inpatients: an investigation of the needles and Abramson model of recovery from depression. Journal of Consulting and Clinical Psychology 66, 369–376.

Kazdin, A.E., French, N.H., Unis, A.S., Esveldt-Dawson, K., Sherick, R.B., 1983. Hopelessness, depression, and suicidal intent among psychiatrically disturbed inpatient children. Journal of Consulting and Clinical Psychology 51, 504–510.

Lazarus, R.S., 1991. Emotion and Adaptation. Oxford University Press, New York.

Lazarus, R.S., Kanner, A.D., Folkman, S., 1980. Emotions: a cognitive phenomenological analysis. In: Plutchik, R., Kellerman, H. (Eds.), Theories of Emotion. Academic Press, New York, pp. 189–217.

Metalsky, G.I., Joiner, T.E., 1992. Vulnerability to depressive symptomatology: a prospective test of the diathesis-stress and causal mediation components of the hopelessness theory of depression. Journal of Personality and Social Psychology 63, 667–675.

Menninger, K., 1959. The academic lecture of hope. American Journal of Psychiatry 116, 481–491.

Needles, D.J., Abramson, L.Y., 1990. Positive life events, attributional style, and hopefulness: testing a model of recovery from depression. Journal of Abnormal Psychology 99, 156–165.

Nezu, A., D'Zurilla, T.J., 1981. Effects of problem definition and formulation on decision making in the social problem-solving process. Behavior Therapy 12, 100–106.

Pekrun, R., Goetz, T., Titz, W., Perry, R.P., 2002. Academic emotions in students' self-regulated learning and achievement: a program of quantitative and qualitative research. Educational Psychology 37, 91–106.

Pekrun, R., Stephens, E., 2010. Achievement emotions: a control-value approach. Social and Personality Psychology Compass 4, 227–294.
Petrie, K., Chamberlain, K., 1983. Hopelessness and social desirability as a moderator variables in predicting suicidal behavior. Journal of Consulting and Clinical Psychology 51, 485–487.
Plutchik, R., 1980. Emotion: A Psychoevolutionary Synthesis. Harper & Row, New York.
Roseman, I.J., Spindel, M.S., Jose, P.E., 1990. Appraisal of emotion-eliciting events: testing a theory of discrete emotions. Journal of Personality and Social Psychology 59, 899–915.
Roseman, I.J., Antoniou, A.A., Jose, P.E., 1996. Appraisal determinants of emotions: constructing a more accurate and comprehensive theory. Cognition and Emotion 10, 241–277.
Scheier, M.F., Carver, C.S., 1985. Optimism, coping, and health: assessment and implications of generalized outcome expectancies. Health Psychology 4, 219–247.
Scioli, A., Ricci, M., Nyugen, T., Scioli, E.R., 2011. Hope: its nature and measurement. Psychology of Religion and Spirituality 3, 78–97.
Smith, C.A., Ellsworth, P.C., 1985. Patterns of cognitive appraisal in emotion. Journal of Personality and Social Psychology 48, 813–838.
Smith, C.A., Ellsworth, P.C., 1987. Patterns of appraisal and emotion related to taking an exam. Journal of Personality & Social Psychology 52, 475–488.
Smith, C.A., Kirby, L.D., 2009. Relational antecedents of appraised problem-focused coping potential and its associated emotions. Cognition and Emotion 23, 481–503.
Smith, C.A., Kirby, L.D., 2011. The role of appraisal and emotion in coping and adaptation. In: Contrada, R.J., Baum, A. (Eds.), Handbook of Stress Science: Biology, Psychology, and Health. Springer, New York, pp. 195–208.
Snyder, C.R., 2002. Hope theory: rainbows in the mind. Psychological Inquiry 13, 249–275.
Snyder, C.R., Harris, C., Anderson, J.R., Holleran, S.A., Irving, L.M., Sigmon, S.T., Yoshinobu, L., Gibb, J., Langelle, C., Harney, P., 1991. The will and the ways: development and validation of an individual-differences measure of hope. Journal of Personality and Social Psychology 60, 570–585.
Stotland, E., 1969. The Psychology of Hope. Jossey-Bass, San Francisco.
Tennen, H., Affleck, G., Tennen, R., 2002. Clipped feathers: the theory and measurement of hope. Psychological Inquiry 13, 311–317.
Tong, E.M.W., Fredrickson, B.L., Chang, W., Lim, Z.X., 2010. Re-examining hope: the roles of agency thinking and pathways thinking. Cognition and Emotion 24, 1207–1215.
Vaillant, G., 2003. Aging Well: Surprising Guideposts to a Happier Life from the Landmark Harvard Study of Adult Development. Little Brown, Boston.
Wetzel, R.D., Margulies, T., Davis, R., Karam, E., 1980. Hopelessness, depression, and suicidal intent. Journal of Clinical Psychiatry 41, 159–160.

Horticultural Societies

Paul Sillitoe, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

After distinguishing horticulture from agriculture, this article goes on to introduce various horticultural societies according to agronomic regimes – shifting cultivation, bush fallow, ley, and agricultural systems – and sociocultural orders – egalitarian/tribal societies, chiefdoms, peasantries, and industrial-urban – to give a typology of horticultural societies. Following this, it discusses the extensive changes that have occurred in horticultural societies, often initiated in the name of economic development – such as the introduction of new crop varieties – and mounting concerns for the sustainability of some horticultural regimes with pressure on natural resources and accelerating environmental degradation, notably through global climate change.

A horticultural society is one in which people's livelihoods depend on the cultivation of small plots, commonly called gardens. These frequently support a wide range of crops and meet staple subsistence needs. The proportion of produce sold on the market varies, to finance the purchase of commodities manufactured elsewhere (e.g., clothing, processed food, hardware, etc.). The scale of society and its organization may vary widely too, from small tribal groups to large peasant populations.

The Term 'Horticulture'

Horticulture is distinguished from agriculture by scale of production, specialization, and commercialization. Agriculture features in industrialized societies supplying food to large urban populations. It is, in some regards, the reverse of horticulture. It involves the farming of large fields, usually under a monocrop (all wheat, potatoes, beans, etc.), whereas the gardens of horticulturalists usually feature intercropping, the mixing up of several different crops, as opposed to their segregation in separate rows or plots. There are sound ecological reasons for this practice, including maximizing on nutrient and water availability, controlling pests and diseases, increasing photosynthetic efficiency, exploiting microenvironmental variations to best effect, and so on. Agriculture features highly mechanized production techniques, whereas horticultural societies rely more on human labor and animal drought power. Farmers in agricultural systems meet few of their own subsistence requirements, specializing in the large-scale production of a few foodstuffs for sale on the market. They engage in agribusiness, intensive production often making use of inputs to boost yields, frequently obtained from off-farm and industrially manufactured (inorganic fertilizers, biocides, etc.). The term horticulture has a quite different meaning in such urban-industrial contexts, which can confuse the unwary. It refers to the intensive production, often under artificial glasshouse conditions, of high value crops, such as early vegetables, salads, and flowers by commercial growers.

Other food production regimes from which horticultural societies are distinguished include gatherer-hunters and pastoralists (Sutton and Anderson, 2004: pp. 125–289), all of which support lower population densities than industrial-urban systems. These categories were bequeathed to us by Victorian scholars who thought of them comprising an evolutionary hierarchy, putting horticultural societies somewhere in the middle of the range between the gatherer-hunter and industrial-urban poles. This scheme continues to inform thinking to the current day (as evidenced by articles in this encyclopedia), and underpins notions of development conceived as facilitating the passage of communities to the industrial-urban pole. Anthropologists continue to use it although uneasy with the continued evolutionary implications, arguing that we should not judge urban-industrial life, notably liberal market democracy, as the superior socioeconomic order to which all human beings should necessarily aspire. It is an approximate classification scheme. We know, for instance, that Australian Aborigines (taken as a classic example of hunter-gatherer life) managed natural resources using fire to promote new vegetative growth, following what is dubbed a 'firestick farming' regime, and harvested in a way that ensured plants such as yams reproduced optimally next season (Gammage, 2011). At some hazy point, populations are thought to manage their natural resources sufficiently closely to be practising farming; this is accompanied by a more sedentary lifestyle. They qualify as horticultural societies when they emphasize crop cultivation over animal herding, although mixed arable and livestock farming is common. The label horticultural society covers a wide range of different subsistence regimes and sociocultural orders. The term is more commonly used in North America than Europe, where instead of lumping these regimes together, it is more usual to distinguish the different socioagronomic orders discussed below.

Agronomic Regimes

We find varying degrees of sedentariness among people who depend largely on cultivation to supply their food. The variation informs a preliminary distinction between different horticultural strategies. There is also some correlation with increasing population densities.

The places where we find the least permanent settlement and physically most direct technologies have cultivation systems that fallow the land not by rotating crops but by

rotating land use. We call this type of horticulture **shifting cultivation**, because after people have grown crops and fertility declines and yields start to fall (these cultivators not using any imported fertilizer to supplement the soil's nutrients), they abandon the plot to the forest and move elsewhere to clear a new garden on fresh land. Secondary regrowth meanwhile recolonizes the abandoned area, which may be left fallow for tens of years to allow nature to restore its fertility; following which it may be cleared and planted again in the next generation. Shifting cultivation is labor intensive, featuring hand tools, such as axes, machetes, hoes, and pointed digging sticks. We find a sexual division of labor, orchestrated into a joint effort. Men usually do the heavy axe-related work. They fell the trees and clear the forest. They also enclose the area to be gardened with rough fencing if this is necessary to keep out animals. The subsequent burning off of the dry and dead vegetation cleared from the plot is often done by both men and women working together. This burning is an important step, reducing standing vegetation to ash, which is readily mixed with the soil, liberating locked up nutrients in a form more available to planted crops. The release of nutrients (such as nitrogen, potash, phosphorus, etc.) through burning has led to the coining of the term 'swidden' to describe this type of cultivation. (Swidden is an old English word meaning to burn.) Following clearing and burning off, women usually take over and are responsible for breaking up the soil, planting crops, weeding, and finally, harvesting the plot. People living around the world in Asia, Africa, the Pacific, and Latin America grow a large variety of crops using this cultivation technique: for example, grains such as corn, millet, and sorghum; roots such as manioc, yams, and sweet potato; and fruits such as bananas, pawpaws, and melons.

Many Amazonian Indian populations practice shifting cultivation, horticultural crops supplying the bulk of their food regardless of the time spent gathering and hunting in the surrounding forest (Watters, 1971). Among the Yanomamo, plantains and bananas are the staple crop. They select well-drained sites, relatively clear of thorny underbrush. They do not attempt to burn large tree trunks, left strewn across plots in a jumble, serving as boundaries between patches of crops that belong to those who plant them. Newly established gardens yield crops in cycles, the rapidly growing ones such as maize coming first, followed by others including manioc, taro, sweet potato, and mapuey. The Yanomamo also grow chewing tobacco, cotton for hammock yarn, drugs for shaministic rituals, and many fruits including papaya, avocados, gourds, and hot peppers. A garden produces constantly as it matures, but a regular supply of food demands careful management. Gardens have direction, moving little by little as people gradually abandon older areas as thorny underbrush encroaches. They constantly clear adjacent land and transplant secondary plantain suckers as they appear. Sometimes, villages are obliged to establish new gardens elsewhere. This is usually a political decision, communities moving when threatened by raids from enemy villages. It requires a large effort to establish a new garden some way off, to transport needed cuttings and seeds. New gardens also produce sporadically at first, and people may face hunger, whereas established 'rolling gardens' offer predictable subsistence.

At some ill-defined point, agronomists distinguish **bush fallow** systems. These are broadly similar to shifting systems regarding technology and cropping strategies but with a reduced fallow interval and longer periods under cultivation. Instead of one or two years under cultivation followed by two or more decades under fallow forest, these horticulturalists may farmland for 5 to 10 years with a fallow interval of similar duration under bush or savanna vegetation. Stock such as cattle, goats, sheep may graze on the fallow. Settlement is increasingly stationary, and land holdings are usually more clearly defined with agreed field boundaries. These fallow systems sometimes evolve from shifting regimes with increasing population densities. They may demand management of soil fertility through manuring or fertilizer application, otherwise yields will decline over time. Degradation of the natural resource base may occur with increasing population pressure and demand to grow cash crops.

Again, at some ill-defined point, we distinguish **ley** systems. These are ones in which farmers systematically use the land to pasture livestock during the nonarable phase. At one end are unregulated leys where people graze their animals on unmanipulated natural regrowth of grasses and herbs. They have communal grazing rights, and practice little pasture management. At the other end are regulated leys where farmers manage the grazing vegetation, frequently putting down grasses and legumes as fodder. They maintain individual pastures, which they may fence off. They may graze meadows directly or use them to produce hay or silage. At this point, we usually think of them as agricultural regimes.

Bush fallows and unregulated leys are common in interlacustrine East Africa (Allan, 1967). The Lugbara, for example, have fields and homesteads scattered across open upland country. Homesteads comprise compounds, clusters of family huts and granaries, under which are ghost shrines and magical plants. Fields surround them. Each homestead cultivates three types of fields. Men do the heavy work of opening the land, women weed and harvest. Both sow, depending on the crop, men more handling cash crops. They intercrop in all. Postharvest care and use of crops are the responsibilities of women. The 'riverine fields,' irrigated by channels from streams, are highly fertile and need only short fallows; they support sweet potato, maize, sugarcane, and bananas. The 'home fields' fertilized with ashes and manure support nutrient-demanding crops such as beer brewing white sorghums. The 'outside fields' support staple eleusine, together with sorghum, simsim, and pigeon pea, cultivated for 2 or 3 years, after which they are left under bush fallow to restore fertility. People allow unregulated grazing of cattle and goats on fallow. They recognize different fallows, from recent and regaining fertility; to areas a few year old, the presence of certain plants showing restored fertility; through to cleared land left for a few months before hoeing; and finally uncultivated land. Each woman in a homestead should have areas in all fields, of equal fertility. If not, quarreling is likely. The fields of a homestead are likely to be scattered, due to differences in soil and water. This sets a limit on the maximum size of farming groups. A household short of certain farmland may move and take up residence with kinsmen elsewhere, usually a wife's or mother's brother.

It is generally agreed that with the invention of the plow and particularly associated agricultural machinery we enter the realm of **agriculture**. The plow allows people to harness either the strength and energy of draught animals or engine-powered machines to assist in their preparation of land for the planting of crops. The result is that one person, together with his team of oxen or whatever, can accomplish considerably more work alone, releasing others to follow other occupations. We have the emergence of increasingly specialized labor, a move from so-called organic to mechanical solidarity. We find the practice of crop rotation and occasional fallowing to maintain soil fertility, together with the increasing use of amendments such as manure, fertilizer, biocides, and so on. In regions that suffer water deficits, people may have irrigation technologies. In some places there are plantations of perennial commercial tree crops. The upshot is permanent farming, with clearly demarcated fields. While it is inappropriate to go into the growth of the first civilizations among the early cultivators of the Middle East, Mesoamerica, and the Far East, we should note some of the consequences of these developments, founded on a settled agricultural way of life (Cowan and Watson, 1992). They allowed for greater population densities, which, combined with the relative permanence of settlement that results with agriculture, led to the emergence of wider-scale political groups than based on the family or tribe. These developments opened the way over millennia to the evolution of industrial-urban societies, sometimes with associated peasantries.

Before turning from the agronomic aspects of horticultural societies to consider the social orders associated with them, it needs to be stressed that while the foregoing regimes have been presented with an evolutionary dimension it would be wrong to consider them as discrete steps leading from shifting cultivation to mechanized agriculture. This typology is no more than a low-level generalization by which we can usefully classify cultures. When we look closely we find that many societies do not fit comfortably into any one category but span two or more, as the East Uganda example illustrates. In the Baliem valley of New Guinea too, where sweet potato is the staple and people herd pigs, we find three sorts of cultivations. The largest are the permanent plots on the valley floor, some of which the Dani recultivate immediately upon harvest and others they leave fallow for months, even years. These gardens feature elaborate ditch systems that drain away/store water, delineate and protect cultivations from pigs and serve as mulching basins supplying organic rich mud for slopping over beds to ensure continued fertility. Other gardens men clear in a classic shifting mode on the mountain slopes, burning off plots and planting for a year or two. The third and smallest are kitchen gardens maintained adjacent to houses supporting a wide range of valuable crops such as tobacco, sugarcane, and bananas, continually fertilized with household waste. Households commonly have several different widely separated fields at different stages.

Sociocultural Orders

We now turn briefly to look at the gross nature of the social orders found associated with horticulturalists. It is difficult to distinguish between above horticultural regimes according to political economy (how people distribute resources). They occur in one of three broad categories: egalitarian with balanced reciprocity, chiefdoms with redistribution systems, or peasantaries in nations with markets. There is some correlation between permanence of settlement, cultivation regime and polity, with shifting cultivators frequently living in egalitarian orders and ley farmers in chiefdoms.

The **egalitarian societies** featuring balanced reciprocity are characterized by groupings commonly called tribes. These are relatively large groups of people (i.e., in comparison to hunter-gatherer bands), who occupy specific territories, where they lay specific claim and rights to resources. Characteristically these societies recognize all individuals as equal and place no persons above others, with authority and power to rule over them. They are what we call stateless or acephalous, and have evolved various strategies to maintain social order without formal government, judiciary, and so on; employing, for instance, descent and age in Africa, and sociopolitical exchange in Melanesia. The productive group is not the tribe, but usually the family, with some help from extra kin in occasional large cooperative ventures. Under this household mode of production, farmers consume most of their produce, although they distribute some to others. The formal distribution of food and goods is likely to be significant, taking place predominantly in reciprocal transactions. Relatives and friends who give things to one another keep a careful account. They give, not with the expectation of receiving an immediate return gift (or payment), but with the expectation of an equal return sometime later. There is often a competitive element in these transactions, those doing best winning higher status. They sometimes constitute grand institutions and are attended with some ceremony. Exchange is a facet of social and political relations and may exist independently of material wants and needs. This takes us away from subsistence and horticulture toward sociopolitical issues.

In Melanesia, sociopolitical exchanges featuring balanced reciprocity characterize the social order. Among New Guinea Highlanders, for example, each Dani family supplies its own food, and provides its own shelter. It occupies a compound with men's and women's houses, and pig sty, often comprising several related families and friends. Clan affiliation determines access to land and all have equal rights. These are rights of usufruct, not land ownership. Men do the heavy, spasmodic garden and construction work, and are more involved in public activities such as politics, disputes, and rituals. Women undertake routine cultivation work and domestic responsibilities, herding pigs and tending family needs. There is little call for distribution of goods with this minimal sexual division of labor. There is some trade with people elsewhere for a few imported goods. Sometimes, friends and neighbors make up a cooperative work party to garden or build a house. But there are no leaders. The Dani combine for conviviality, each person doing the same thing, the output comprising the sum total of their efforts (it is more fun mudding gardens together for example, even though this is a one-man job). Hostilities are common, fighting with bows and arrows and spears. Exchange promotes social order, primarily featuring pigs, which are usually killed on ceremonial occasions. The giving-and-receiving of animals constantly refurbishes the social network; people are bound to one

another through their lifetimes by the exchange of hundreds, even thousands, of pigs.

The **chiefdoms** with redistribution systems are hierarchically organized societies, where persons occupy different social ranks. In the top rank are the rulers, commonly called chiefs. These persons have authority vested in them, often sanctioned by some supernatural connection, and govern in a manner those living in their chiefdom, kingdom, tribe or whatever. Everyday rule is usually delegated locally to middle rank persons, who are responsible for day to day government, administration of justice, and so on. These chiefdoms are often feudal in nature, with these local leaders owing allegiance to the chief and supplying a military force if required by him. Societies organized in this way exploit raw materials on a larger scale. They do this, for example, by territorial conquests (as epitomized by the Zulu under their king Shaka) and through extensive trading networks (e.g., the Aztecs controlled mining activities as far away as northern Mexico). These societies can also organize larger productive groups, through the authority of their rulers, who by so doing, not only facilitate cooperation but also provide some sort of rationale for the existence of their power (e.g., the farming terraces, irrigation works, storehouses for produce, etc. constructed by the Inca). Redistributive systems predominate in these societies, where those of lower rank give food and goods to those higher in the form of tribute. The chief will rearrange everything he receives and add to it any surplus goods, which his probably superior productive position supply (i.e., from his large, communally worked gardens), and then redistributes it all to his subjects (usually via local-level officers in his government).

The chiefdoms of South Africa are good examples of this political order. The horticulture practised is broadly similar to East Africa, the main crops being maize and millet; together with groundnuts, gourds, sugarcane, and pumpkins. Farming is seasonal, dependent on the summer rains. One source of rulers' political authority is the belief that they can control the advent of the rains. Many have a precarious existence; food supplies fluctuating annually between plenty and scarcity bordering on famine, the Swazi for example talk of lean summer months as the time 'to swallow the pickings of the teeth.' Horticulture is not a prestigious occupation and is left primarily to women, whose main tool is the iron hoe. Except that is for plowing, which involves highly valued oxen handled by men, and the cultivation of cash crops, notably tobacco and cotton. Households cultivate large fields using ox-drawn plows. The power of rulers also relates to their right to allocate land to their subjects, who depend on it for their livelihood. The Swazi say that their political overlords 'serve' them the land, securing their rights by allegiance and usage. Women are allocated garden areas by their husbands at marriage. Homestead plot size depends primarily on supply of resident labor. People use grazing lands communally. They value pastoralism highly, having the so-called African 'cattle complex,' these animals have great symbolic value being exchanged to ratify marriage, sacrificed to propitiate ancestors, and paid in tribute to rulers. They call the king the 'Bull' of the nation, and the royal herd far exceeds any other, supplying animals redistributed to subjects at national events and sacrifices. In addition to sex and age, pedigree also structures the division of labor. Men and women cooperate in agriculture, man's work is heavier and more spasmodic, woman's routine and more continuous. Aristocrats do not engage in manual labor to the same extent as their subjects, they are responsible for successfully organizing their efforts, supervising work parties, and arranging for specialists to doctor land and seed. The rulers have several 'gardens of kingship' in different localities, which local headmen maintain by arranging their communal cultivation. Homesteads rely on kinsmen and neighbors to assist at bottlenecks; customarily organized into two competitive work parties they spur each other on in songs, rhythmically coordinating their work through music. The size of communal work parties reflects status; they enhance prestige, being rewarded with beer and food at the host's home.

The **peasantaries** associated with markets are also incorporated into hierarchically organized societies. They are usually incorporated within nation states where they occupy marginal and low positions, relatively powerless, and commonly exploited. These horticultural communities are in peripheral, underprivileged situations, frequently taken advantage of, even abused by, the dominant society. They usually feature poverty. Peasants depend in some regards on economic and political institutions beyond their local control. These bodies, which rule significant aspects of their lives, may supply them with certain services. Persons may participate in informal reciprocal exchange networks within their communities, but outside they are obliged to engage with impersonal commodity markets. While they supply their own subsistence needs, like tribespeople – from whom many of the Third World's peasants have descended in recent times – they are obliged also to cultivate a surplus to sell on outside markets to earn cash. In addition to using this money to purchase goods manufactured elsewhere, they also need some cash to pay the taxes levied on them by the colonial/national governments to which they are subject. They may also need to purchase outside inputs, such as inorganic fertilizers and biocides, to maintain their yields, especially if their farming system is under increasing demographic pressure, perhaps resulting in increasing environmental degradation, or subject to enforced 'development' (such as having to cultivate 'high-yielding varieties'). This may contribute to growing poverty. What peasants receive financially in return for their surplus horticultural produce is on the whole unfair, the élite making a disproportionate profit to support their privileged position and lifestyle by controlling access to imperfect markets and the arrangements for the distribution of produce.

The Mayan Indians of Mexico have occupied a peasant status increasingly since the Spanish conquest (Vogt, 1990). They sell crops surplus to their subsistence needs, largely maize and beans, and trade evaporated salt at fiesta day markets. In turn they purchase goods such as coffee, sugar, fish, liquor, and bread. They have been involved in marketing for centuries, historical evidence showing that they traded with Aztecs in pre-Hispanic times. Maize is their staple, believed to have an 'inner soul' in the ear, together with beans, squashes, chiles, chayotes, and fruits such as peaches. They keep chickens and pigs, and sheep for wool woven into *chamarras*, blankets and shawls, and mules for carrying heavy loads. The Mayans cultivate three types of land: small individually owned garden plots located near Highland homes, their ancestral patrimony; larger *ejido* plots acquired during the 1940s agrarian reform program, allocated by local

committees to households; and plots in the lowlands rented from wealthy *Ladino* landowners by patrilineal groups, which they divide equally among themselves. They use centuries' old swidden techniques, working with axes, machetes, bill hooks, and hoes. After cutting and burning vegetation, they plant using pointed digging sticks. In old fields they cut and rake maize stubble into piles for burning, and then hoe sheep manure into the turned topsoil. They prepare fields for planting at the onset of the rainy season. Ritual attends planting, a critical task in which a shaman offers candles, incense, liquor, and prayers to the Earth owner, who controls wind and clouds. People cultivate fields for varying periods of time, depending on elevation and soil quality, before allowing brush to grow again. They store threshed maize in granaries or indoor grain bins prior to sale.

Again this sociocultural typology, like the previous agronomic one, is only gross, the elements constituting approximate cultural sets, which have proved useful in organizing our ideas over many years. It continues to structure our thoughts in a preliminary fashion. There is some broad correlation between peoples' horticultural arrangements that is, how they produce food and other items necessary for life, and how they are socially organized, both to effect production and to achieve distribution of its fruits. But this is not to imply, as some writers argue, that mode of production determines the nature of social arrangements. Furthermore, many of these societies have undergone substantial social change under colonialism and subsequently globalization.

Future Directions

Several questions remain to be answered about horticultural societies, particularly with the occurrence of extensive social change. Many aspects are still poorly understood, insight sometimes blinkered by orthodoxy. The Baliem Valley for example supports one of the highest population densities in New Guinea ($\sim$160 persons km^{-2}), with no evidence that it is pushing against the environmental limits. How can the region's intensive horticulture system support such a high population density without any outside fertilizer amendments? The accepted shifting cultivation model, which pertains to this part of the world, predicts evidence of environmental strain, of the kind commonly associated with overpopulation. Yet the Highland region's sophisticated horticultural regime maintains soil fertility sufficient for the repeated cultivation of some crops (Sillitoe, 2010).

A major contributor to our understanding of horticultural societies beyond anthropology and geography has been 'farming systems research' conducted under the aegis of international development agencies (Shaner et al., 1982). This involves multidisciplinary teams, using survey techniques largely, to detail the ecological, technical, and socioeconomic aspects of horticultural regimes and identify constraints on production amenable to scientific research. It was initially associated with 'modernization,' seeking to advance the development of horticultural societies into commercial agricultural ones. It was thought that this would alleviate poverty, leading to increased production of a food surplus necessary to support increasing urban populations. To this end, agricultural research institutes worldwide embarked on programs to breed new 'high yielding varieties' of crops and develop associated technologies to increase production. The scientifically successful 'green revolution' resulted, diminishing the threat of famine. Nonetheless while it benefited wealthy landowners, it largely passed poor farmers by. It also worryingly increased dependency on outside manufactured inputs (inorganic fertilizers, fossil-fueled machinery, chemical pesticides, and herbicides) and gave rise to concern about long-term sustainability with evidence of increasing environmental pollution, susceptibility to mutated pathogens, insecure supply of needed inputs, and so on. There is a danger that the current push for genetically modified crops may exacerbate these problems. The move toward cash cropping also put farmers at the mercy of market fluctuations, sometimes causing their families considerable hardship, further increasing resistance to modernization's inducements to embrace commercial production.

There are growing concerns about the sustainability of various horticultural regimes, and other livelihood strategies, with changes that are occurring to the natural environment globally, largely resulting it is thought from human activities. Currently popular, and regularly featured in the media, is global climate change, which if meteorologists' modeling predictions are accurate, will have far reaching ecological consequences. Others are pollution of soil, water, and air with industrial and other waste, land degradation due to inappropriate farming practices, and extinction of life forms and threat to the planet's biodiversity with unknown environmental consequences. Many of those living in horticultural societies, who comprise the rural poor in many developing countries, are particularly vulnerable because of their limited capacity to cope with climate variability against a background of increasing global political-economic pressures, which include calls for adoption of clean energy, and often declining sustainability of livelihoods.

If changes of the magnitude predicted occur, the knock-on effects for human activities, including horticulture, are likely to be considerable, and research into these and ways to cope are seen increasingly as a priority. The 'sustainable livelihoods' approach has emerged in development contexts, which aims to approach livelihood strategies holistically, as structured around livelihood assets – comprising five types of capital (social, physical, natural, financial, and human) that are assessed according to vulnerability to shocks – which set within various socioeconomic structures and processes (government, laws, civil institutions, companies, etc.) determine livelihood outcomes (Carney, 2002). It is a flexible, propoor, people-centered approach that seeks systematically to understand livelihood strategies with a view to informing development interventions to improve the security and living standards of communities.

Key future questions are not only to what extent are variability and climatic extremes impacting on horticultural communities but also what adaptations are people currently making to cope with these and how effective are they in reducing vulnerability? It is evident that a significant part of the burden of coping with predicted environmental change will depend on the resourcefulness of communities and local means, and we need to understand the implications from their

perspective. In other words, we not only have to further understand horticultural regimes but also investigate the feasibility of people innovating on them to manage the impacts of forecast change and their possible social implications. For instance, the scope for scaling up the *baira* 'floating gardens' a half-forgotten local tradition recently seen again on the Bangladesh floodplain where climatic extremes, including floods, cyclones, heat waves, and variable rainfall, challenge millions of poor vulnerable people, trapping them in a cycle of poverty (Irfanullah et al., 2011).

The growing concern over the implications of predicted global environmental change has reinforced the rethink of priorities that occurred in the 1980s when it was realized that some small-scale subsistence focused horticultural societies were strongly resisting change as envisioned by modernization and communities were being increasingly exposed to serious perturbations beyond their control in trying to accommodate to development interventions. The 'participation' approach to development emerged as it was increasingly acknowledged that poor people need to be involved in the planning and implementation of any interventions intended to assist them. This promises to advance understanding and the interests of horticultural societies in the future, particularly with the concomitant emergence of 'indigenous knowledge' research, which represents the coming together of applied farming systems research and academic disciplines such as anthropology, geography, and ecology, which have long esteemed understanding of horticultural societies (Sillitoe et al., 2005). The assumption is that any development assistance needs to be in tune with local sociocultural and environmental contexts. It also furthers meaningful appreciation of the complex relations that inform natural resources management, as understood by local farmers, those who contend daily with them – the collapse of many interventions has been attributed to a failure to comprehend the complexity of their horticultural systems. These changes in research and development agendas bode well for the future in furthering understanding and promoting the interests of horticultural societies *along their own lines*.

See also: Hunting and Gathering Societies: Anthropology; Irrigation Societies; Pastoralism in Anthropology; Peasants and Rural Societies in History (Agricultural History).

Bibliography

Allan, W., 1967. The African Husbandman. Oliver and Boyd, London.
Altieri, M., 1995. Agroecology: The Science of Sustainable Agriculture. Westview Press, Boulder.
Bayliss-Smith, T.P., 1982. The Ecology of Agricultural Systems. Cambridge University Press, Cambridge, UK.
Carlstein, T., 1982. Time Resources, Society and Ecology. George Allen and Unwin, London.
Carney, D., 2002. Sustainable Livelihoods Approaches: Progress and Possibilities for Change. Department for International Development, London.
Cernea, M.M., Kassam, A.H. (Eds.), 2006. Researching the Culture in Agri-Culture. CABI Publishing, Wallingford.
Cowan, C.W., Watson, P.J., 1992. The Origins of Agriculture. Smithsonian Institution Press, Washington.
Gammage, B., 2011. The Biggest Estate on Earth: How Aborigines Made Australia. Allen & Unwin, Sydney.
Gliessman, S.R., 1990. Agroecology: Researching the Ecological Basis for Sustainable Agriculture. Springer Verlag, New York.
Grigg, D.B., 1974. The Agricultural Systems of the World. Cambridge University Press, Cambridge, UK.
Irfanullah, H.M., Azad, M.A.K., Kamruzzaman, M., Wahed, M.A., 2011. Floating gardening in Bangladesh: a means to rebuild lives after devastating flood. Indian Journal of Traditional Knowledge 10 (1), 31–38.
Ruthenberg, H., 1976. Farming Systems in the Tropics. Clarendon Press, Oxford.
Shaner, G.W., Phillipp, P.F., Schmehl, W.R., 1982. Farming Systems Research and Development. Westview Press, Boulder.
Sillitoe, P., 2010. From Land to Mouth: The Agricultural 'Economy' of the Wola of the New Guinea Highlands. Yale University Press, New Haven.
Sillitoe, P., Dixon, P., Barr, J., 2005. Indigenous Knowledge Inquiries: A Methodologies Manual For Development. Intermediate Technology Publications, London.
Sutton, M.O., Anderson, E.N., 2004. Introduction to Cultural Ecology. Berg, Oxford.
Vogt, E.Z., 1990. The Zinacantecos of Mexico: A Modern Maya Way of Life. Holt, Rinehart and Winston, New York.
Watters, R.F., 1971. Shifting Cultivation in Latin America: FAO Forestry Development Paper No. 17, Rome.

Hospitals and Health: Psychosocial Aspects

Jeffrey A Alexander, School of Public Health, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the United States, increasingly, strong market and financial pressures are interacting with traditionally dominant institutional forces to shape the behavior of hospitals, particularly those that operate as nonprofit organizations. This situation contrasts with that of many other developed nations where hospitals and health care organizations are operated largely under public auspices. Escalating demands for greater efficiency, cost containment, and higher quality, together with counterpressures to maintain local autonomy and professional control, has resulted in both new organizational practices and organizational forms. This article describes the social forces affecting hospitals and other health care organizations, and the changes that these organizations have experienced as they attempt to accommodate potentially conflicting environmental demands.

In the United States, increasingly, strong market and financial pressures are interacting with traditionally dominant institutional forces to shape the behavior of hospitals, particularly those that operate as nonprofit organizations. This situation contrasts with that of many other developed nations where hospitals and health care organizations are operated largely under public auspices. Escalating demands for greater efficiency, cost containment, and higher quality, together with counterpressures to maintain local autonomy and professional control, has resulted in both new organizational practices and organizational forms (Guterman et al., 2010). Many of these new practices and structures are borrowed in modified form from the corporate sector. These effects are widely expressed in hospital ownership, organizational structures, and governance, as well as in work practices and controls. The most notable examples include the consolidation of hospitals into complex health systems, the consequent shift in the locus of governance away from local communities, and the increasing diversity and complexity of compensation arrangements between health care delivery organizations and physicians. Operational effects have included greater emphasis on both efficiencies in the use of human resources, increasing value in health care, and improved quality of care. Though specific environmental pressures vary in different countries, changes in the broader market and institutional contexts that hospitals face are key in understanding their evolving structure and behavior.

During the second half of the twentieth century, the dominant mode of health care delivery in the United States has evolved from individual, patient–physician relationships to care provided by organizations (Starr, 1982). At the same time, social scientists have moved from examining hospitals as closed systems, which operate independently of their environment, to seeing them as open systems, both affected by and responding to external forces such as regulation, professional norms, shifting reimbursement policies, and significant efforts to 'reforming' and aligning incentives (Scott, 1998). This perspective assumes that structural forms and organizational behavior in the health care sector are influenced by factors such as regulation, change in professional power, competition among providers, and mutualistic behavior among different types of providers. The context in which hospitals operate has been described as having two essential components: the technical environment and the institutional environment. The technical aspects of an organization's task environment involve the production of goods and services, and the market forces that reward their efficient and effective production. Institutional environments encompass the rules and requirements to which organizations must conform, if they are to receive the social and political support and legitimacy needed to survive. Increasingly, the two types of forces are seen as competing.

Organizations in the health care sector have traditionally been subject to both strong institutional and strong technical pressures. These forces have operated in separate but parallel fashion to shape the practices and structures of hospitals (Alexander and Scott, 1984). For example, hospitals traditionally maintained very loose control over physician staff as a reflection of strong professional requirements for autonomy. At the same time, they were accountable for the 'technical' quality of care through legal and quality assurance controls. Recently, market forces have played a more visible role in shaping national-policy debate concerning health care access, quality, and costs. For example, hospitals now compete actively with each other for patients and for government and health plan reimbursement that is increasingly tied to outcomes such as quality, patient satisfaction, and cost reduction. In view of this, some have argued that the influence of traditional institutional structures and practices, such as regulation, professional norms, and local control, have weakened and are now subordinated to the technical demands of the market. However, there is evidence that historical institutional forces have not been eliminated, but, instead, are interacting with the new 'market' ethos to constrain and shape the structures and practices of hospitals, including those originating in the corporate sector (Thomas and Helmreich, 2002). These interactions have given rise to a unique set of changes and practices that differ from both those common to the corporate sector and those common to hospitals in more institutionally dominant eras.

Institutional Environment of Hospitals

Institutional environments feature elaborate rules and requirements to which individual organizations must conform, if they are to receive support and legitimacy. Attention is

directed away from control and coordination of technical processes and toward conformity to externally defined requirements or regulations. The more institutionalized the environment of an organization, the more its structures and procedures are governed by external controls. This control requires specific arrangements in the organization, which are independent of contributions to technical performance (Scott, 1995). For much of the twentieth century and up until today, hospitals and other health care organizations have been subject to strong institutional pressures, including a broad array of governmental regulations and requirements and a strong ethos of local, community control. Further, numerous professional norms and prerogatives have governed what types of personnel might be hired, how tasks are distributed among them, and what procedures must be followed in performing these tasks (Nembhard et al., 2009).

The health care field in the United States has long been characterized by institutional pressures that are both strong and heterogeneous (Stevens, 1989). All branches of the federal government make decisions centrally, which affect the entire health care sector. Local and state governments also exercise authority to regulate health care. This fragmented structure of authority makes it possible for various actors to pass laws to meet particular needs, and regulations are therefore uncoordinated, at best, and inconsistent or conflicting, at worst. Further, a long tradition of volunteerism in American hospitals has complemented the inculcation of professional norms into the institutional structure of these organizations. The notions of local control, community accountability, and philanthropic support of hospitals have served to keep decision-making from becoming overly centralized. At the same time, this has discouraged the formation of a cohesive, integrated health care system.

Technical Environment of Hospitals

Notwithstanding the institutional demands on hospitals, the technical requirements of modern medicine are considerable and often require tight internal controls and careful coordination, if performance is to be effective. Hospitals have also increasingly been pressured to improve efficiency in the use of resources. Retrospective cost reimbursement was singled out as the primary explanation for both rising costs and inappropriate utilization. Critics have also identified professional specialization and hospital-based care as major contributors to rising costs. Growing specialization and fragmentation in delivery rendered services increasingly impersonal, while the predominance of inpatient services threatened accessibility and provision of comprehensive, family-oriented care.

Historically, political attention had focused on expanding access to health care, as exemplified by the passage of legislation in the 1960s extending health insurance coverage to many of the elderly (Medicare) and the poor (Medicaid). The 1980s, however, marked the beginning of a new era that shifted the focus to controlling costs (Scott et al., 2000). Several years of double-digit increases in prices threatened the economic interests of actors who paid for health care; in response, they initiated policies and practices to control costs. In 1984, the federal government's Medicare program became the first major payer to attempt to control hospital costs by paying a fixed fee for each service, rather than the 'usual and reasonable' fees that hospitals routinely charged. This key change in payment policy had significant financial and market implications for hospitals. The resulting reductions in payments to hospitals and other organizational providers cut their budgets and contributed to cash-flow problems. Hospitals and other providers tried to compensate by increasing their patient and service volume. In turn, this created more competition for patients.

A series of remedies have been proposed to deal with the cost escalation crisis. The solutions imposed have often clashed with the historical institutional functions of hospitals, namely, providing care to the community, running on a voluntary basis, and engaging in philanthropic enterprise. Direct regulatory approaches to containing costs in the 1970s have shifted to greater emphasis on legal, budgetary, and market responses to addressing these issues. Taken as a whole, these remedies fall under the general rubric of a competitive approach to controlling costs, even though, all involved substantial government oversight and occasionally intervention. Such approaches are consistent with a general societal shift toward privatization and market solutions to larger social problems. While recently court rulings have relaxed some regulations stifling collaboration among doctors and organizational providers, new regulations have reduced price-fixing and other forms of anticompetitive behavior. Together these rulings have provided both inducements and constraints on organizational behavior, such as severe limitations on hospital acquisitions and mergers.

Budgetary remedies to these conflicts have been applied at both the federal and state levels. At the federal level, the U.S. government initially budgeted a fixed reimbursement for each inpatient admission based on diagnosis. States also enacted rate-setting programs that prospectively set reimbursement levels for hospitals and other providers. These budgetary interventions have forced health care providers to attend to efficiency and have raised the level of competition among these providers. Finally, market remedies have included the introduction of the so-called 'managed care' strategies, designed to reduce utilization and thereby costs (*see* Health Insurance: Economic and Risk Aspects). Examples of these strategies have included competitive bidding and selective contracting in government programs and the use of preferred providers and risk-based financial incentives by private insurance firms. More recently, however, both state and federal governments have begun to publish information on the performance of health care providers (on both cost and quality), aimed at redressing the traditional asymmetry of information between providers and consumers of health care and have introduced new reimbursement approaches that provide incentives for hospitals and other providers to better coordinate care and reduce duplication of services (Conrad and Perry, 2009).

As expected, these interventions have lowered health care costs, decreased utilization, and, perhaps to a lesser extent, increased quality of care (Robinson, 1996; Fung et al., 2008). But there have also been other, unanticipated effects. Since the 1990s, there have been an increasing number of exits of traditional organizations from the health care system. Although most hospitals and other organizational providers that have closed have tended to be primarily inner-city and rural institutions and have done so for primarily

technical/market reasons, many others have remained operating despite financial losses and technical inefficiencies. The survival of so many marginal institutions in the face of increasing pressures from a market-based system reflects the importance of local, community support for these institutions, and the salience of legitimacy in organizational survival (Rohrer, 1989). This illustrates well the complex combination of technical and institutional forces that shape organizational life in the health care sector.

Hospital Responses to Hybrid Environments

A fundamental question becomes, how do health care organizations operating within such hybrid environments accommodate different environmental demands that are often diametrically opposed to each other? The convergence of traditional institutional and new technical forces impinging upon hospitals has created new organizational, strategic, and survival challenges for health care providers. Neo-institutional theory suggests that organizations functioning in sectors that are highly developed both institutionally and technically will develop more complex and elaborate administrative systems and will experience higher levels of internal conflict (Scott, 1995). Historically, we know this to be true for many organizations in the health care sector. Hospitals, for example, have typically operated within a dual hierarchy of control, represented on the one hand by hospital administration and on the other by hospital medical staff and physicians. These dual lines of control operate in parallel with each other, but rarely have intersected in such a way as to disrupt the relative spheres of influence over issues related to control, resource allocation, and professional prerogatives.

Conflict has been expressed in hospitals largely through increasing parity between managerial and professional logic and ideologies. The emphasis of managerial ideology on rational systems and cost-effective care has increasingly conflicted with the professional orientation toward individual doctor–patient relations and the microdelivery of specific medical services (Goodrick et al., 1997). Conflicting claims to the right way to deliver health care, the presence of alternative rationales for organizing work (professional versus managerial), and the inherent conflicts between technical and institutional pressures in the health care sector have eroded the legitimacy of hospitals. This has opened the way for a number of new organizational forms, practices, and strategies, particularly those that have symbolically represented the new ethos of privatization and corporate rationality. For example, health organizations are increasingly lobbying government and engaging in marketing campaigns in order to influence legislation and public opinion. Rather than being passive recipients of institutional changes in the broader environment, they seek actively to shape societal beliefs toward health care and the way health care is organized.

Organizational Heterogeneity

Health care delivery has been dominated traditionally by hierarchical organizations such as hospitals and health care systems. However, as competitive pressures and the demands of managing care under limited resources have increased over the past decade, a wide variety of new organizational forms have emerged. These range from managed care organizations, physician–hospital arrangements, accountable-care organizations, and physician practice management companies, to diagnosis-specific service contracts with specialized providers, such as home health care and rehabilitation firms (Iglehart, 2011).

As traditional institutional norms and practices have eroded and have been replaced by a stronger corporate rationality, we have observed greater heterogeneity among organizations in the health care sector. This new ethos suggests there is no best way to organize in order to achieve technical efficiency and positive performance in the marketplace. More variation in organizational forms is exhibited, as hospitals become less concerned with mapping the characteristics of the institutional environment onto their own structures and more concerned with achieving technical efficiencies and buffering their technical cores from environmental disturbances.

A clear response to the new market forces in health care has been the adoption of structures and practices from the corporate sector (Burns, 1990). These structures include multidivisional organizational designs, parent holding-company arrangements, contract management of facilities, and vertical integration across a number of different service lines. While these practices clearly have gained popularity during the 1990s, there is limited evidence that they actually improve the quality or lower the cost of care. Neo-institutional theorists have argued that while early adopters of these corporate structures may have done so for technical–rational reasons, later adopters may have viewed them as the prevailing institutional practices under the new market-based health care environment.

Hospital Governance

Historically, most nonprofit hospitals were overseen by autonomous governing boards composed largely of representatives of the communities the hospitals served. This type of governance arrangement reflected a longstanding institutional belief that the control of nonprofit hospitals should be vested in local communities. However, during the 1980s and 1990s, two developments have altered this traditional governance model of nonprofit hospitals, both of which are part of a larger trend toward the so-called 'corporatization' of health care. One development is the consolidation of over 40% of US hospitals into complex health care systems. In such systems, the locus of control for member hospitals is shifted to varying degrees outside local communities to system-level governing boards and corporate offices (Toomey and Toomey, 1993). Another development is corporate restructuring of hospitals, where hospital assets and decision-making authority are transferred to multiple corporate entities that may include a parent holding company and various subsidiaries. Such restructuring often entails the creation of multiple governing bodies that can attenuate the decision-making authority of the provider's community-based board.

Under a traditional model of governance, nonprofit hospitals resemble consumer cooperatives in that community representatives are in a position to determine hospital policies.

As such, a traditional form of governance is arguably a constraint on hospital price inflation, since community representatives have an incentive to resist price increases that they ultimately will incur in the form of higher insurance premiums. Also, form of governance may impact the accessibility of hospital services since a community-based board may be inclined to invest resources to expand the accessibility of hospital services to its own community (e.g., through the provision of uncompensated care). Policymakers have expressed concern that system membership and corporate restructuring are diminishing the role of communities in the control of nonprofit hospitals, and government agencies that have oversight responsibilities for nonprofit hospitals are currently examining alternatives.

Indeed, hospitals have come under increased public scrutiny for practices that appear to be inconsistent with their not-for-profit (NFP) ethos, which traditionally focuses on serving their communities in a charitable manner (Seay and Sigmond, 1989). Recent investigations into the activities of NFP hospitals have pointed to weak or lax management and governance and claim a variety of misdeeds such as excessive compensation of executives, inadequate provision of charity care, and deceptive pricing of services. These and related concerns have given rise to questions as to whether there is, in fact, sufficient difference between nonprofit and for-profit hospitals (Kane, 2007). This is another example of how hospitals are often caught between their traditional philanthropic missions and new market forces.

Physician–Organization Structures

As payment reform and competitive pressures have compelled hospitals to consolidate and provide services along a broad continuum of care, the need to integrate formerly loosely coupled elements of these organizations has increased. Perhaps, the most problematic aspect of such integration is the alignment of incentives and strategic goals between delivery organizations and physicians. Hospitals seek to engage in strategic relations with physicians in order to: (1) protect against the risks of agency from physicians who have traditionally borne little risk for the care of patients, (2) reduce transaction costs by developing economies of scale and integrating physicians into the structure of the delivery organization, (3) achieve market power both up- and downstream over suppliers and customers respectively, and (4) develop capabilities that differentiate systems from competitors in the areas of innovation, technology, and quality. While these market-driven reasons are clearly critical to hospitals and health care systems, the alignment of economic and strategic interests between organizations and physicians is difficult given the differences in cultural context, socialization, orientation to patient care, and economic incentives (Dukerich et al., 2002).

As managed care organizations and health care regulators have recognized the need to align the economic incentives of providers with the goals of cost containment and high quality, they have changed methods of provider compensation from simple discounting to risk sharing (e.g., capitation, *per diem*, or per case reimbursement for hospitals). Providers under these systems assume actuarial and utilization risk and reward. In principle, risk sharing promotes appropriate behavior, since high utilization and costs will result in deficits for either the physician group or its physicians. In practice, the physician group's form of compensation, incentives, and governance moderate these effects. For example, the incentives inherent in risk sharing can be lost if individual physician compensation is not strongly linked to holding costs below the group's fixed budget. Alternatively, if physicians feel they have no say in the formulation of incentives in compensation systems, they may be less willing to abide by them.

These examples suggest that the incentives created by risk sharing are effective only under certain conditions. Several authors have recognized that the effects of such payment mechanisms may partially depend on organizational factors such as type of governance, physician leadership, information dissemination, sanctions, and organizational culture (Kralewski et al., 2000; Alexander et al., 2000). Aside from such fundamental differences in orientation to quality and related mistrust over motives and goals, many of the problems of alignment can be traced to medical staff governance itself. In contrast to hospitals, which are organized and governed along bureaucratic lines with clearly established lines of authority and accountability, the typical hospital medical staff in the United States is organized as an association of physicians with ties to the hospital, but with little else to bind them together as a collective entity. They are essentially membership organizations that lack formal authority to impose accountability standards on their members, or to direct integrated activity with hospitals. Because of its loose structure and emphasis on individual physician interests, the association form of medical staff governance is not well suited to promote collective responsibility for patient-care quality, operational efficiency, or other areas of performance. Many states, in fact, have long had in place policies that prohibit corporations (i.e., hospitals) from employing physicians directly – the so-called 'corporate practice of medicine doctrine' – which serves to reinforce the legitimacy of the voluntary medical staff with all its previously noted difficulties (Alexander and Young, 2010).

Physicians and organizational health care providers now recognize the need to work together to manage risk and coordinate care under new reimbursement schemes. To realize these objectives, to avoid institutionalized constraints inherent in hospital–medical staff relations, and to develop the administrative infrastructure necessary for managing service delivery, many health care systems have established market-based organizational arrangements that link the delivery organization with one or more physician groups. Market-based arrangements are different from internal hospital structures in that they tend to be formally structured, contractual in character, corporate in organization, and include physicians outside the formal membership of the medical staff.

Given fundamental differences between methods of organizing by physicians and delivery organizations, however, it is highly questionable whether structural solutions to achieving greater integration between physicians and delivery organizations can be completely effective. There is only limited evidence to suggest that structural solutions work and considerably more evidence that self-selection by physicians into closely aligned structural arrangements may account for more of the explained variance in alignment than structure itself. Hospitals' attempts

to integrate and align with physicians have not always rationalized the delivery of health care, but instead in many cases have made it more complex. Physicians and their interests vary widely, and hospitals have discovered that no one structural vehicle best accommodates the interests of all physician subgroups. Consequently, it is not uncommon for multiple market-based arrangements to exist within a single organization. This makes the coordination of care and alignment of incentives at least as difficult, if not more so, as it has been under the traditional medical staff model.

Organization of Work in Hospitals

The high degree of specialization in health care means that each professional brings only a fraction of the knowledge needed to care for patients. A rich literature shows that the actions of these professionals must be well coordinated to be effective. Despite the imperative for such collaboration, effective care coordination in hospitals is often missing from professional interactions because of the hierarchical, individualistic culture of medicine (IOM, 2000, 2001), which is deeply rooted in the socialization process for health professionals and its institutionalized value system (Leape and Berwick, 2005).

Largely as a result of increasing competitive pressures and external emphases on cost containment and quality improvement (QI), the organization of work within hospitals has shifted. The reciprocal interdependence inherent in treating patients whose clinical conditions are evolving requires that care providers in the hospital setting participate in ongoing knowledge-sharing to arrive at a complete picture of the patient's situation and an effective treatment plan. Two trends are particularly in evidence: the use of teams to perform managerial and clinical work, as opposed to performing such work individually; and the use of contracts to obtain labor externally rather than to hire employees.

Because of demands by managed care firms, revised payment policies, and integrated health care systems to provide comprehensive care to patients and/or enrollees in health plans, there is growing recognition that patients need comprehensive care in order to address their multiple problems effectively. Given the broad spectrum of knowledge necessary for comprehensive care, there is growing belief that teams of practitioners with complementary roles and skills are best suited to providing high quality, cost-effective patient care (Fried and Rundall, 1994). Teamwork in the clinical round is also perceived as an antidote to fragmentation in services, a traditional problem in health care delivery. For example, repeated patient assessments by various health care providers add unnecessary costs to patient care, when one combined assessment would suffice. Advocates of team-oriented care cite studies that consistently show that collaboration among physicians, nurses, and other health care personnel enhances the quality of care and patient satisfaction (Nikolaus et al., 1999; Le et al., 1998).

Team-based care stands in marked contrast to the traditional hierarchical work relations in most hospitals, which accord the physician a role of absolute autonomy and control over patient-care processes. The growing emphasis on prevention, wellness, and cost-effectiveness in care delivery make this hierarchical structure outmoded. Flatter structures and more collegial relations among providers appear to be the developing norm. For example, many nurses advocate a collaborative relationship with physicians where nurses have primary responsibility for health maintenance, prevention, and diagnosis and treatment of routine cases, while physicians retain responsibility for the diagnosis and treatment of complex cases and the prescription of medications.

Contracting for labor services, as an alternative to employment, stems from the need for more flexibility in an increasingly competitive environment. Hospitals need to be able to make rapid changes in their workforce in response to demands from a volatile market. Rather than staffing to accommodate periods of peak demand with attendant downtime for slack periods, hospitals have increasingly adopted a contingent approach to staffing, responding to the particular needs of the organization at a given time (Katzman, 1999; Rizzuto et al., 1999). Contracting typically occurs in housekeeping, business, and clinical services. Proponents argue that contracting frees top managers to concentrate on core services, brings specialized expertise to the organization, improves flexibility through the ability to terminate contracts, reduces supply costs through superior purchasing power, and promotes the recruitment of specialists who may be in short supply. Critics argue, however, that the contract-management firm and the contracting organization may fail to clarify performance goals or that contracted staff may have difficulty working with other departments and full-time personnel. Perhaps most importantly, some believe that contracting for a wide variety of services without a clear strategy risks damaging organizational coherence and control (D'Aunno et al., 1996).

Quality of Care in Hospitals

Another illustration of the organizational response to technical and institutional pressures lies in the area of health care quality (McGlynn et al., 2003). On the one hand, decline in public support and trust for hospitals has reduced their impetus to assure quality and discipline members of their organizations for lapses in quality. Responsibility for quality has begun to shift to the payers of health care and their agents (e.g., managed care organizations), as well as to the courts, which adjudicate an increasing number of malpractice suits against hospitals and physicians. These trends reflect a growing tendency to think of health care as a commodity rather than an entrusted service provided by professionals and professional organizations. On the other hand, there is greater emphasis placed on QI systems within organizations as a counterbalance to external controls over quality and as a response to market pressures to increase the efficiency and effectiveness with which care is delivered (Kimberly and Minvielle, 2000). Again, this duality of external versus internal control is a juxtaposition based in the competing demands placed on hospitals by institutional and technical forces.

Hospitals, in particular, are being challenged by increasingly competitive markets and regulatory changes to produce high quality care without raising costs. Given the multiple pressures and problems associated with health care quality, hospitals face the additional challenge of developing policies, procedures, and

incentives to improve quality within a domain that has traditionally been the province of highly autonomous professionals. QI practices have emerged in recent years partially as an alternative to quality assurance and as a reaction to the shortcomings of outcome-based control (McLaughlin and Kaluzny, 2006). With mounting pressure to maintain or improve quality in the face of reduced resources, hospitals have begun to look to process improvements as a mechanism to provide high quality care, while containing health care costs. While the adoption of QI appears to be an internally generated response to changing market conditions, for the most part, accrediting bodies such as the Joint Commission on Accreditation of Health care Organizations have also encouraged hospitals to adopt QI.

Given the traditional emphasis on provider–patient relations, individualized approaches to improving quality have received the vast majority of attention to date. Although the knowledge and practices of individual physicians are important for high quality care, increasingly clinicians practice within groups or systems of care. As such, the quality of care delivered depends critically on the ability of these groups or systems to prevent errors, coordinate care among settings and practitioners, and ensure that relevant, accurate information is available when needed.

One of the most promising recent developments is the effort to institutionalize QI in hospitals as a coordinated, organization-wide approach to care strategies (Berwick et al., 1990; Gaucher and Coffey, 1993; Carman et al., 1996). These practices, collectively, are designed to systematically examine and improve processes of care and care support in organizational settings. Typically, these QI efforts are strongly rooted in evidence-based procedures and rely extensively on data collected from hospitals about care processes and patient outcomes.

However, while few would dispute the importance of providing high quality care in hospitals, many hospital managers feel that they cannot afford to invest heavily in QI programs because such investments would erode their bottom line and make them less competitive in markets where cost containment receives primary emphasis. By contrast, proponents of investing in QI efforts maintain that enhancing quality makes the hospital more attractive to patients, physicians, and payers, thereby contributing positively to the hospital's financial health. Although this issue has been widely debated in terms of 'the business case for quality,' the absence of empirical research in the area has opened the debate to competing technical and institutional beliefs rooted in the competing values of professionalism, market competition, and community control (Hendricks and Singhal, 2001; Ittner and Larcker, 1996; Lederer and Rhee, 1995).

Hospital Behavior in Other Nations

The focus of this article has been to associate change in hospital structure and behavior to change in the broader market and institutional contexts in which these organizations operate. The United States has been prominently featured in this discussion because of the somewhat unusual combination of strong technical and institutional pressures to which hospitals must respond. Other countries, however, may diverge markedly in the degree to which either institutional or market forces play a role in shaping hospital behavior. For example, in countries where health care delivery is highly centralized and funded exclusively by the government, one would expect to find much more structural homogeneity among hospitals, in marked contrast to the increasing diversity and differentiation among US hospitals. In other nations, where privatization and multiple sources of health care funding predominate, hospitals are more likely to resemble in their structure and behavior of other commercial vendors of services to the market.

Recent trends in OECD countries would seem to indicate that they are facing technical pressures for increased efficiency similar to those in the United States (Normand, 1993). Significant reductions both in the number of hospital beds per 1000 and in average length of hospital stay have resulted, particularly in Scandinavian countries. In the countries of the former Soviet Union, on the other hand, average length of stay has not dropped, despite mounting financial strain due to economic instability. One potential explanation for this is lingering institutional norms regarding appropriate clinical practice (Hensher et al., 1999). This may illustrate how specific institutional beliefs about health and health care in a given nation may shape hospital behavior in ways that differ substantively from the ways in which institutional forces shape hospital behavior in the United States. For example, in Taiwan, the deeply seated belief that family is the primary care giver is reflected in the hospital practices of incorporating family members in patient education, extended visiting hours, and even boarding with the patient during the hospital stay. Cross-national differences in hospital structure and behavior are thus likely to be a reflection of both the structure and content of the broader environment from which hospitals obtain patients, critical resources, and legitimacy.

See also: Comparative Health Systems; Doctor–Patient Interaction in the West: Psychosocial Aspects; Gender and Health Care; Globalization and Health; Health Care Markets: Theory and Empirical Evidence; Health Care, Rationing of; Health Insurance: Economic and Risk Aspects; Medical Profession; Public Health as Social Science; Socioeconomic Status and Health Care.

Bibliography

Alexander, J.A., Scott, W.R., 1984. The impact of regulation on the administrative structure of hospitals. Hospital and Health Service Administration 29, 71–85.

Alexander, J.A., Vaughn, T., Burns, L., 2000. Organizational, strategic and marketing determinants of operating policies of physician organization arrangements. Health Services Management Research 13, 1–15.

Alexander, J.A., Young, G., 2010. Overcoming barriers to improved hospital-physician collaboration and alignment: governance issues. In: Crossan, J., Tollen, L. (Eds.), Partners in Health: How Physicians and Hospitals Can Be Accountable Together. Jossey Bass, San Francisco, CA.

Berwick, D.M., Godfrey, A.B., Reossner, J., 1990. Curing Health Care: New Strategies for Quality Improvement. Jossey-Bass, San Francisco, CA.

Burns, L.R., 1990. The transformation of the American hospital: from community institution toward business enterprise. Comparative Social Research 12, 77–112.

Carman, J.M., Shortell, S.M., Foster, R.W., Hughes, E.F., Boerstler, H., O'Brien, J.L., O'Conner, E.J., 1996. Keys for successful implementation of total quality management in hospitals. Health Care Management Review 21 (1), 48–60.

Conrad, D.A., Perry, L., 2009. Quality-based financial incentives in health care: can we improve quality by paying for it? Annual Review of Public Health 30, 357–371.

D'Aunno, T., Alexander, J.A., Laughlin, C., 1996. Business as usual? Changes in health care's workforce and organization of work. Hospital and Health Service Administration 41, 3–18.

Dukerich, J.M., Golden, B.R., Shortell, S.M., 2002. Beauty is in the eye of the beholder: the impact of organizational identification, identity, and image on the cooperative behaviors of physicians. Administrative Science Quarterly 47, 507–533.

Fennell, M., Alexander, J., 1993. Perspectives on organizational change in the US medical care sector. Annual Review of Sociology 19, 89–112.

Fried, B., Rundall, T., 1994. Managing groups and teams. In: Shortell, S.M., Kaluzny, A.D. (Eds.), Health Care Management: Organization Design and Behavior, third ed. Delmar Publishers, Albany, NY.

Fung, C.H., Lim, Y.-W., Mattke, S., Damberg, C.L., Shekelle, P.G., 2008. Systematic review: the evidence that publishing patient care performance data improves quality of care. Annals of Internal Medicine 148 (2), 111–123.

Gamm, L.D., 1992. Health care markets as interorganizational fields: a conceptual perspective. Health Services Management Research 5 (1), 44–53.

Gaucher, E.J., Coffey, R.J., 1993. Total Quality in Healthcare: From Theory to Practice. Jossey-Bass, San Francisco, CA.

Goodrick, E., Meindl, J.R., Flood, A.B., 1997. Business as usual: the adoption of managerial ideology by U.S. hospitals. Research in the Sociology of Health Care 14, 27–50.

Guterman, S., Davis, K., Stremikis, K., Drake, H., 2010. Innovation in medicare and medicaid will be central to health reform's success. Health Affairs 29 (6), 1188–1193.

Hendricks, K.B., Singhal, V.R., 2001. Firm characteristics, total quality management and financial performance. Journal of Operations Management 19, 269–285.

Hensher, M., Edwards, N., Stokes, R., 1999. International trends in the provision and utilization of hospital care. British Medical Journal 319, 845–848.

Iglehart, J.K., 2011. Assessing an ACO prototype: medicare's physician group practice demonstration. New England Journal of Medicine 364 (3), 198–200.

Institute of Medicine, 2001. Crossing the Quality Chasm: A New Health System for the 21st Century. National Academy Press, Washington, DC.

Institute of Medicine, 2000. To Err Is Human: Building a Safer Health System. National Academy Press, Washington, DC.

Ittner, C., Larcker, D., 1996. Measuring the impact of quality initiatives on firm financial performance. In: Ghosh, S., Fedor, D. (Eds.), Advances in the Management of Organizational Quality, vol. 1. JAI Press, Greenwich, CT, pp. 1–37.

Kane, N.M., 2007. Tax-exempt hospitals: what is their charitable responsibility and how should it be defined and reported? St Louis University Law Journal 51 (2), 459–473.

Katzman, C.N., 1999. Outsourcing keeps growing. Contract management survey finds times are good for firms serving healthcare industry. Modern Healthcare 29 (35), 42, 44, 46–50.

Kimberly, J.R., Minvielle, E., 2000. The Quality Imperative: Measurement and Management of Quality in Health Care. Imperial College Press, London.

Kralewski, J.E., Rich, E.C., Feldman, R., Dowd, B.E., Bernhardt, T., Johnson, C., Gold, W., 2000. The effects of medical group practice and physician payment methods on costs of care. Health Services Research 35 (3), 591–613.

Le, C.T., Winter, T.D., Boyd, K.J., Ackerson, L., Hurley, L.B., 1998. Experience with a managed care approach to HIV infection: effectiveness of an interdisciplinary team. American Journal of Managed Care 4 (5), 647–657.

Leape, L., Berwick, D., 2005. Five years after to Err is human what have we learned? JAMA 293 (19), 2384–2390.

Lederer, P.J., Rhee, S., June 1995. Economics of total quality management. Journal of Operations Management 12, 353–367.

McGlynn, E.A., Asch, S.M., Adams, J., Keesey, J., Hicks, J., DeCristofaro, A., Kerr, E.A., 2003. The quality of health care delivered to adults in the United States. New England Journal of Medicine 348 (26), 2635–2645.

McLaughlin, C.P., Kaluzny, A.D., 2006. Continuous Quality Improvement in Health Care, third ed. Jones and Bartlett Publishers, Sudbury, MA.

Nembhard, I.M., Alexander, J.A., Hoff, T., Ramanujam, R., 2009. Understanding implementation failure in health care delivery: a role for organizational research and theory. Academy of Management Perspectives 23 (1), 1–27.

Nikolaus, T., Specht-Leible, N., Bach, M., Oster, P., Schlierf, G., 1999. A randomized trial of comprehensive geriatric assessment and home intervention in the care of hospitalized patients. Age and Ageing 28 (6), 543–550.

Normand, C., 1993. Comparisons among national health care systems in the European marketplace. Hospital Formulary 28 (Suppl. 1), 6–11.

Rizzuto, G.T., Yallapragada, R.R., Roe, C.W., Authement, J.P., 1999. Contract labor use in healthcare. Hospital Topics 77 (4), 15–19.

Robinson, J.C., 1996. Decline in hospital utilization and cost inflation under managed care in California. Journal of the American Medical Association 276 (13), 1060–1064.

Rohrer, J.E., 1989. Closing rural hospitals: reducing 'institutional bias' or denial of access? Journal of Public Health Policy 10 (3), 353–358.

Scott, W.R., 1995. Institutions and Organizations. Sage, Thousand Oaks, CA.

Scott, W.R., 1998. Organizations: Rational, Natural and Open Systems, fourth ed. Prentice-Hall, Upper Saddle River, NJ.

Scott, W.R., Ruef, M., Mendel, P.J., Caronna, C.A. (Eds.), 2000. Institutional Change and Healthcare Organizations: From Professional Dominance to Managed Care. University of Chicago Press, Chicago.

Seay, J.D., Sigmond, R.M., Spring 1989. Community benefit standards for hospitals: perceptions and performance, the future of tax-exempt status for hospitals. Frontiers of Health Services Management 5 (3), 3–39.

Starr, P., 1982. The Social Transformation of American Medicine. Basic Books, New York.

Stevens, R., 1989. In Sickness and in Wealth: American Hospitals in the Twentieth Century. Basic Books, New York.

Thomas, E.J., Helmreich, R.L., 2002. Will airline safety models work in medicine? In: Rosenthal, M.M., Sutcliffe, K.M. (Eds.), Medical Error: What Do We Know? What Do We Do? Jossey-Bass, San Francisco, CA.

Toomey, R.K., Toomey, R.E., 1993. The role of governing boards in multihospital systems. Health Care Management Review 18 (1), 21–30.

Hotelling, Harold (1895–1973)

Ingram Olkin, Stanford University, Stanford, CA, USA
Allan R Sampson, University of Pittsburgh, Pittsburgh, PA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 6921–6925, © 2001, Elsevier Ltd.

Abstract

Harold Hotelling (1895–1973) was an internationally renowned American leader in developing the intellectual foundations of statistics, and in furthering areas of mathematical economics. His best known statistical research concerns multivariate analysis, where he introduced Hotelling's T^2, principal components analysis and canonical correlations analysis. Hotelling's T^2 is a multivariate generalization of Student's *t*-test, principal components are uncorrelated linear combinations of the underlying data capturing the data's variability, and canonical variables and correlations help to simplify the dependence relationships between two sets of multivariate measurements. Hotelling's major contributions to mathematical economics concern natural resource economics, spatial economic competition, and welfare economics. Hotelling additionally played a prominent role in the administration of statistics in universities and national societies.

Harold Hotelling (1895–1973) was one of a handful of statistical researchers who in the first half of the twentieth century developed the solid intellectual foundations of the then newly developing discipline of statistics. In addition, he was also an important contributor to several areas of mathematical economics. His academic career includes: professorships at Stanford University, Columbia University, and the University of North Carolina at Chapel Hill, and his professional leadership involved major statistical and economic societies. He is perhaps best known for his statistical work in multivariate analysis, introducing the concepts of principal components, canonical correlations and Hotelling's-T^2. In mathematical economics, he contributed to natural resource economics, spatial economic competition, and introduced the welfare equilibrium proposition.

Early Life

Born on 29 September 1895 in Fulda, Minnesota, Harold Hotelling, whose parents were Clair Hotelling and Lucy Rawson, was the oldest of six children. Having moved with his family to Seattle at age nine, Hotelling studied journalism at the University of Washington, taking time from his studies to work on a weekly paper. In 1919, he received his BA in Journalism, graduating Phi Beta Kappa. Encouraged, in part, by the mathematician and well-known historian, Eric Temple Bell, he went on to graduate study in mathematics at the University of Washington, receiving an MA in 1921. It was during this time that he married Floy Tracy in 1920. He received his Ph.D. in Mathematics in 1924 from Princeton University, doing dissertation research on dynamic topology with Oswald Veblen, although his original intent was to study mathematical economics. His interest in economics took root while growing up and in his belief in 'the importance ... of a study of economics [in] that [it] might possibly impress some people to the point of making some changes in the institutions with which they had become familiar [and] which I [Hotelling] was already prepared to condemn vigorously.' Alternatively, he 'couldn't escape the lure of mathematics' and notes that he 'hoped and [it] appeared that the mathematics would be extremely prolific of new economic theorems and ideas if followed out' (Hotelling, 1963).

Beginning at Stanford University

Invited to join the Food Research Institute at Stanford University, Hotelling was a research associate there from 1924 to 1927, following which he was an associate professor in Stanford's Department of Mathematics (1927–31). An early indication of the breadth of his intellectual interests is illustrated by the fact that in 1927 he taught courses on mathematical statistics, differential geometry, and analysis situs (a forerunner of topology), and in that year published papers on the application of the latter to statistics, and on differential equations subject to error. While working in the Food Research Institute, he was introduced to the work of R. A. Fisher (five years his elder), and through their mutual correspondence they became friends. In fact, Hotelling wrote the book review in the *Journal of the American Statistical Association* (Hotelling, 1927) of the first edition of Fisher's *Statistical Methods for Research Workers* (and also did subsequent reviews there of the Second through Seventh Edns.). After the planned visit of Fisher to Stanford in the summer of 1928 was abandoned, Hotelling visited Fisher at the Rothamsted Experimental Station during the second half of 1929. Their connections continued and at the outbreak of World War II, Hotelling, along with three others, wrote to Fisher volunteering to care for Fishers' six daughters in the United States during the war. However, a last minute cancellation kept the Fisher children in Britain (Box, 1978: p. 377).

While at Stanford, Hotelling wrote a number of path-breaking papers in statistics and in mathematical economics. The most notable among these statistical results was his generalization of Student's *t*-ratio to handle multivariate response data (Hotelling, 1931). This generalization remains known as Hotelling T^2. His paper with Holbrook Working (Working and Hotelling, 1929) on modeling and linear prediction first demonstrated that the standard error of the estimate of a linear regression depended on the independent variable's value, thereby giving rise to the well-known

Working–Hotelling curved simultaneous confidence regions for a linear regression. His research contributions to mathematical economics were equally notable. His 1929 paper provided a game-theoretic solution given by two differential equations to a problem of price and location competition between two entities in a spatial setting. His 1931 paper on the economics of exhaustible resources is recognized as the seminal paper in the field of natural resource economics. In that paper, Hotelling demonstrates that in an equilibrium setting, prices of exhaustible natural resources will tend to rise over time at a percentage rate equaling the interest rate.

Columbia University and the Statistical Research Group

Hotelling moved to Columbia University in 1931 as a Professor of Economics, having been recruited to further the development of mathematical economics, and also to initiate mathematical statistics. Throughout his career at Columbia and later at the University of North Carolina at Chapel Hill, Hotelling was superlative at both attracting outstanding researchers and mentoring young researchers destined for stardom. The Nobel-prize winning economist, Milton Friedman, was among the earliest mentored by Hotelling, spending 1933–34 at Columbia learning mathematical economics from him. Later mentoring Kenneth Arrow, another Nobel-prize winning economist, Hotelling was also instrumental in attracting Abraham Wald, W. Allen Wallis, and Jacob Wolfowitz to the statistics facility at Columbia. As a teacher, Hotelling was perceived to be organized and comprehensive, interweaving theory with application. Importantly, an appreciation of research was imparted to his students, and he was able at an interpersonal level to instill in his students their own confidence to be successful at research (Madow, 1960).

Hotelling was one of three original voting fellows of the Institute of Mathematical Statistics on 12 September 1935, and was President of that Institute in 1941. He began his association with the Cowles Commission, participating in its first Summer conference in 1935, and served as President of the Econometric Society during 1936–37. In 1939–40, he was a Visiting Lecturer at the Indian Statistical Institute and was President of the Indian Statistical Congress (1940).

Along with Wallis and Wolfowitz, Hotelling was a charter member of the Statistical Research Group (SRG), at Columbia, for which Wallis was appointed Director of Research. The SRG, founded on 1 July 1942, and in existence until September 30, 1945, was an extraordinary research group of statisticians whose goals were to provide research support and analyses for the armed forces in furthering the quality and accuracy of their war efforts. The origins of sequential analysis stem from Wald's work for the SRG, and among Hotelling's contributions were techniques for control charts for multivariate data (Kruskal, 1980). (A detailed history of the SRG is provided by Wallis, 1980).

Early in Hotelling's career at Columbia, his wife, Floy, died in 1932, leaving behind two small children. In 1934, Hotelling married Susanna Edmundson and together they had six children, five sons and a daughter (who died in her infancy).

Hotelling, during his prewar years at Columbia, continued to produce ground-breaking research in statistics and economics. His 1933 and 1936 papers introduced and lay the groundwork for principal component analysis as a multivariate statistical technique for simplifying the structure of large numbers of correlated measures (Hotelling, 1933, 1936a). In 1936, he extended notions of multiple correlation to canonical correlations in order to measure and study the relationships between two dependent sets of correlated responses (Hotelling, 1936b). In the field of economics, his 1935 paper provided a consistent theory uniting demand and utility by means of a constrained optimization problem and his presidential address to the Econometric Society in 1938 (Hotelling, 1938) introduced a welfare equilibrium proposition which leads in certain societal settings to the argument that goods be priced at their marginal costs. In 1940, his innovative and insightful paper on the teaching of statistics in universities was published (Hotelling, 1940). In this prescient paper, Hotelling observes the inevitability of the increased demand for statisticians, and argues that the teaching of statistics in universities would be best done solely by statisticians in separate departments of statistics (a quite novel idea in 1940!). With regard to this article, Jerzy Neyman (1960) later wrote 'in my opinion the establishment of the excellent departments of statistics in such universities as Chicago, Columbia, Harvard, North Carolina, and Stanford is due to a considerable extent to the publication of Hotelling's article.'

Career at University of North Carolina

After the war in 1946, Hotelling was recruited by Frank Graham, President of the University of North Carolina at Chapel Hill to start the Department of Mathematical Statistics there upon his arrival. In addition to being appointed professor of mathematical statistics and professor of economics, he was named Associate Director of the Institute of Statistics (directed by Gertrude Cox), which was constituted with Chapel Hill's department and the Department of Experimental Statistics at North Carolina State, as well as related groups in psychometrics, social sciences and biostatistics. It is notable that when Hotelling was planning to go to Chapel Hill, he proposed that Wald join him there. In order to entice Wald to remain at Columbia, a Department of Mathematical Statistics was promptly proposed and officially approved in December 1946 (Hoxie, 1955). As at Columbia, Hotelling attracted outstanding statisticians to the Chapel Hill faculty including R. C. Bose, Wassily Hoeffding, P. L. Hsu (who had been at Columbia with Hotelling), William Madow, Herbert Robbins, S. N. Roy, and Walter Smith. Hotelling relinquished the chairmanship to George Nicholson in 1952, and devoted the remainder of his career there to his research and his students.

His research at North Carolina involved, on the whole, elaborations and in-depth further study of some of his earlier research. Noteworthy among these is his work in 1947 and 1951 on hypothesis tests related to multivariate analysis of variance, which lead to the testing criterion referred to as the Lawley–Hotelling trace. He formally retired in 1966, but continued to be active in the department. He was elected to the

National Academy of Sciences in 1970, and received a number of other academic honors during this period. Subsequent to a serious stroke in mid-1972, Harold Hotelling died in Chapel Hill, North Carolina on 26 December 1973.

Hotelling's Contribution to Multivariate Statistical Analysis

One of the origins of statistical hypothesis testing is a test that the mean μ of a normal distribution has a specified value μ_0. The resultant Student's t-test, obtained in 1908, remains as one of the most used statistical procedures. The t-statistic is

$$t = \sqrt{N(\bar{x} - \mu_0)/s}$$

where $\bar{x}$ and s, respectively, represent the mean and standard deviation of a sample of size N. Hotelling recognized that many experiments will have correlated measures $x_1, \ldots, x_{,p}$ which arise from a multivariate normal distribution, and he provided the Hotelling's T^2 test to handle the means of multiple measurements (Hotelling, 1931). For a sample of size N let $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_p)$ denote the sample mean vector and $S = (s_{ij})$ represent the sample covariance matrix.

The T^2-statistic is

$$T^2 = N(\bar{x} - \mu_0)S^{-1}(\bar{x} - \mu_0)'$$

where $\mu_0 = (\mu_{10}, \ldots, \mu_{p0})$ is the specified mean vector in the hypothesis. When the hypothesis is true, $(N-p)T^2/p(N-1)$ has an F-distribution with p and $N-p$ degrees of freedom.

Analogous to Student's t-test, the T^2-statistic can also be used to test for the equality of the mean vectors of two normal distributions that have the same unknown covariance matrix but unequal means, μ and ν. Now the statistic becomes

$$T^2 = \frac{N_1 N_2}{N_1 + N_2}(\bar{x} - \bar{y})S^{-1}(\bar{x} - \bar{y})'$$

where $\bar{x}$ and $\bar{y}$ are mean vectors from samples of size N_1 and N_2, respectively, and S represents the pooled sample covariance matrix. Although there are many parameters in this model, under the null hypothesis the distribution of $(N_1 + N_2 - p - 1)$ T^2/p has an F-distribution with p and $N1 + N_2 - p - 1$ d.f.

The T^2-statistic additionally arises as a linear discriminant, as noted by Fisher, and is the sample analog of the Mahalanobis squared distance.

Quite remarkably, Hotelling noted that for the two-sample problem, the T^2 statistic is invariant under the transformation $\bar{x} \to \bar{x}A$, $\bar{y} \to \bar{y}A$, where A is any nonsingular matrix. This anticipated a general theory of statistical invariance developed much later. In particular, the T^2-test has a number of optimal properties among invariant tests: (a) the T^2-statistic is the only invariant function of the sufficient statistics $\bar{x}$ and S; (b) as a test statistic it is the uniformly most powerful test among all invariant tests; and (c) the noncentral distribution depends on the scalar parameter $(\mu - \nu)\sum^{-1}(\mu - \nu)'$, which is the only invariant of the parameters.

Large data sets either in exploratory data analysis, data mining, or model building are accompanied by a large number of variables. The reduction of numbers of variables is often a first step in trying to uncover patterns in the data. Because multivariate measures $x_1, \ldots, x_p$ are related, there is the need to disentangle the correlational structure. To accomplish this, Hotelling (1933, 1936a) introduced the p principal components $z_1, \ldots, z_p$, which are linear combinations of the underlying measures, namely $z_k = \sum_{j=1}^{p} a_{ki}x_i$, $k = 1, \ldots, p$,where the coefficients are normalized by $\sum_{i=1}^{p} a_{ki}^2 = 1$. The first set of coefficients $a_{11}, \ldots, a_{1p}$ is chosen so as to maximize the variance of z_1. The second set of coefficients $a_{21} \ldots, a_{2p}$ is chosen to maximize the variance of z_2 subject to the constraint that z_1 and z_2 are uncorrelated. At each step the variance is maximized, but with the additional constraints that each z_k be uncorrelated with $z_1, \ldots, z_{k-1}$. Their ordered variances Var $(z_1) \geq \ldots \geq$ Var (z_p) turn out to be the characteristic roots of the population covariance matrix.

In effect, principal components provide transformed coordinates that are uncorrelated, thereby disentangling the measures. Because the principal components are ordered according to their variances, this permits the researcher to focus on a subset of principal components that contributes substantially to the total variance. In multiple regression, the predictor variables often are closely related, and the problem of multicollinearity results. Principal components is one important procedure that introduces orthogonality and reduces dimensionality, thus removing multi-collinearity.

To build a hierarchy of correlational measures one can start with the ordinary Pearson product moment correlation between x and y. With multiple measures $x_1, \ldots, x_p$ and y one generates the multiple correlation of y anclthe linear combination $y^* \equiv \sum a_i x_i$, which is the maximum correlation between y and y^*. Suppose now there are multiple measures $x_1, \ldots, x_p$ and $y_1, \ldots, y_q$. To analyze the relation between these two sets of variates, Hotelling (1936b) defined canonical correlations between u_k and v_k, which are the correlations of linear combinations $u_k = \sum a_{kj} x_j$ and $v_k = \sum b_{kj} y_j$ where $k = 1, 2, \ldots, K \equiv \min(p, q)$. The coefficients are normalized so that the variances of u_k and v_k are 1, and are chosen so as to maximize the correlations, and such that $u_1, \ldots, u_k$ are mutually uncorrelated and $v_1, \ldots, v_k$ are mutually uncorrelated $k = 1, \ldots, K$. The new variables u_k and v_k, $k = 1, \ldots, k$ are called the canonical variables; furthermore, it turns out that u_i and $v_j (i \neq j)$ are uncorrelated. The discovery of the corresponding matrix factorization is quite remarkable. A special case of this factorization leads to the singular value decomposition, which is a very important representation in numerical linear algebra.

As a result of this decomposition, one can transform the original variables to new coordinates $(u_1, v_1), \ldots, (u_K, v_K)$, that exhibit a simple structure: (1) the pairs are uncorrelated; (2) the pairs are ordered decreasingly in accordance to their correlations (which are non-negative); that is, (u_1, v_1) are maximally correlated; and (3) the variances of the u_k's and v_k's are normalized to be one. In a 1935 paper, Hotelling had earlier focused on the first pair (u_1, v_1), and he called u_1 the most predictable criterion and v_1 the best predictor.

Principal components and canonical correlations can be viewed in terms of roots of a determinantal equation and estimated from their sample analogues. It was this context that opened an entire area of statistical research on the distribution of the roots of random determinantal equations. In 1939, a number of papers appeared simultaneously in the statistical literature in which the central distribution of the roots were

obtained. From then until the present, there is a continuing flow of theory concerning the noncentral distributions of roots of sample determinantal equations.

Hotelling's training as a mathematician provided him with a warehouse of tools that he applied to statistical problems. Among his first papers was a presentation in 1926 to the American Mathematical Society on the use of topological principles in the derivation of distributions. Many of his later results were obtained with geometrical arguments. He was able to exploit asymptotic theory to obtain normal approximations and limiting results. Hotelling, in 1930, was the first to attempt to state precise conditions for the consistency and asymptotic normality of the maximum likelihood estimator. With Lester Frankel in 1938, he obtained expansions for the distribution of Student's *t*-statistic, thereby leading to a useful normal approximation.

With his interest in relations between variables, Hotelling recognized that the required normality assumptions may be unrealistic. This led him to study the nonparametric version of the product moment correlation. Together with Margaret Pabst, Hotelling studied the Spearman rank correlation coefficient and they showed that for large samples it approximately follows a normal distribution. In 1953, Hotelling continued his analysis of correlations and obtained expansions of Fisher's *z*-transformation of Pearson's product moment correlation. This work remains among the definitive results in this area.

Hotelling wrote on matrix calculations and developed methods that still stand. His work on weighing designs led to a number of combinatorial generalizations. Although Hotelling had a mathematical focus for his research, he was not averse to working on applied problems. The range of his applied papers include results concerning the physical state of protoplasm, duration of pregnancy, causes of birth rate fluctuations, anthropomorphic measures of Southwest Indians, and effects of pregnancy on dental caries.

Harold Hotelling's originality and imagination is seen in both the genius of his research that produced innovative methods that stimulated whole areas of modern statistics and in his prescient leadership in the discipline of statistics. Further details concerning Hotelling, including his complete bibliography, can be found in Olkin *et al.* (1960), in Pfouts (1960), and in three articles appearing as 'Three papers in honor of Harold Hotelling at 65' in *The American Statistician* **(14),** published in 1960 (pp. 15–25).

See also: Conflict: Anthropological Aspects; Conflict: Organizational; Cross-Cultural Research Methods in Psychology.

Bibliography

Box, J.F., 1978. R. A. Fisher: The Life of a Scientist. Wiley, New York.

Hotelling, H., 1927. Review of 'Statistical methods for research workers. Journal of the American Statistical Association 22, 411–412.

Hotelling, H., 1931. The generalization of Student's ratio. Annals of Mathematical Statistics 2, 360–378.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24, 417–441 and 498–520.

Hotelling, H., 1936a. Simplified calculation of principal components. Psychometrika 1, 27–35.

Hotelling, H., 1936b. Relations between two sets of variates. Biometrika 27, 321–377.

Hotelling, H., 1938. The general welfare in relation to problems of taxation and of railway and utility rates. (Presidential address to the Econometric Society). Econometrica 6, 242–269.

Hotelling, H., 1940. The teaching of statistics. Annals of Mathematical Statistics 11, 457–470.

Hotelling, H., 1963. 'Remarks by Harold Hotelling at a Seminar in which Honorary Degree Recipients were asked to give brief intellectual biographies.' As transcribed from a tape recording — not corrected or edited. University of Rochester, Private Communication.

Hoxie, R.G., 1955. The Department of Mathematical Statistics. In: A History of the Faculty of Political Science. Columbia University. Columbia University Press, New York, pp. 250–255.

Kruskal, W.H., 1980. Comment: First interaction with Harold Hotelling; testing the Norden bombsight. Journal of the American Statistical Association 75, 331–333.

Madow, W.G., 1960. Harold Hotelling as a teacher. American Statistician 14, 15–17.

Neyman, J., 1960. Harold Hotelling: A leader in mathematical statistics. In: Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G., Mann, H.B. (Eds.), Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford University Press, Stanford, CA, pp. 6–10.

Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G., Mann, H.B., 1960. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford University Press, Stanford, CA.

Pfouts, R.W., 1960. Essays in Economics and Econometrics: A Volume in Honor of Harold Hotelling. University of North Carolina Press, Chapel Hill, NC.

Wallis, W.A., 1980. The Statistical Research Group, 1942–1945. Journal of the American Statistical Association 75, 320–330.

Working, H., Hotelling, H., 1929. Applications of the theory of error to the interpretation of trends. Journal of the American Statistical Association 24 (March suppl.), 73–85.

Household Archaeology

Ruth Tringham, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 6925–6929, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Household archaeology developed in the 1980s as a reaction to the dominant regional scale of traditional ecological–evolutionary models in archaeology by focusing on the study of intrasettlement relations. The first household archaeology studies were designed to provide detailed microscale analysis of demographic, social and economic processes of evolutionary change. These studies focused on what households do, rather than what their social form is. The initial impetus for household archaeology came from archaeologists who were excavating well-preserved architectural remains in Mesoamerica, the southwest US and in Europe. Household archaeology studies were enriched by models (mostly Marxian) that emphasized the role of dominance structures and social inequality within and between households in larger socioeconomic transformations. Archaeological evidence of the spatial differentiation of domestic production and its architectural context became the heart of these household archaeology studies. In the late 1990s household archaeology transformed into a number of strands that embraced the post-processual critique in archaeology, recognizing the historical contingency of the social practice of individuals in the household and their role as the foundation of larger social transformations, and the active role of materials use spaced en rules throughout including houses – in creating and maintaining the world of these household actors. This theme has continued into the twenty-first century.

Introduction: The Emergence of Household Archaeology

Household archaeology emerged in the late 1970s and early 1980s as a late development of social archaeology on both sides of the Atlantic. The aim of household archaeology is to create a context in which a humanized reconstruction of the past may be nurtured, through the study of intrasettlement relations. Such an investigation at the microscale involves the intensive recording of archaeological observations and detailed excavation of an area broad enough to include architectural and other features of the built environment. The criticism of such an effort is that it produces an isolated picture of one part of a social group – perhaps only one household – that does not provide a statistically valid extrapolation to the rest of a settlement, or from one settlement to the regional scale.

Household archaeology developed both as an extension of and a reaction, by both processual ('new') and Marxist-oriented archaeologists to the social evolutionary features of social archaeology (Robin, 2013). Social archaeology provided data and inferences for explanations of change and variation of past cultural systems, focusing on intersettlement relations, especially in terms of the interaction of political leaders on a regional basis within the parameters of ecological features such as distance and resource distribution. Marxist-oriented studies added the dimensions of ideology and power relations in production and distribution.

A criticism of the macroscale models of social change has been that, in identifying the longer-term evolutionary trajectory of, for example, resource control, hierarchical social relations, and population growth, and in establishing the general social form of a group, the context of shorter term changes and variability is given a lower priority in explanation. In this way a dehumanization of the past has been created in order to 'write history (or prehistory) large.'

Household archaeology developed in the 1980s and 1990s as the result of a growing interest in intrasettlement relations, but it is a sub-discipline that is neither unified, organized, nor recognized within archaeology. The aims and agendas of the archaeologists who carry out studies that are termed 'household archaeology' are as variable as are the theoretical strands of archaeology itself.

Household Archaeology: Archaeology at a Small Scale

When it was initially developed in the late 1970s the study of households aimed at filling in the most detailed level of settlement pattern and activity analysis, which could then be extrapolated to make more general statements about demographic trends, specialized production, class structures and so on (Flannery, 1976). Later, in the important volumes edited by Wilk and Rathje and Wilk and Ashmore, the importance of households as 'fundamental elements of human society' was emphasized (Ashmore and Wilk, 1988, p. 1; Wilk and Rathje, 1982). At this time it was suggested that, since households are the level at which social groups articulate directly with economic and ecological processes, their study would offer a chance for archaeologists to examine social adaptation (in a functionalist sense) with direct reference to the empirical details of the archaeological record. It was assumed this would be done within the methodological framework of scientific logical positivism. In other words, household analysis would allow archaeologists to 'bridge the existing 'mid-level theory gap' in archaeology' (Wilk and Rathje, 1982, p. 617). Thus household archaeology developed alongside the empirical theory-building in archaeology, known also as middle range research, that lies at the heart of the history of processual archaeology at the end of the twentieth century (Smith, 2011).

It is clear that household archaeology in the 1980s implied examining and analyzing social change at a microscale as a complement to analyses of social and economic change at a broader scale of regions and continents. Household archaeology served to provide a richness of detail, but was not an end in itself. The ultimate target was to understand social evolutionary change.

Sources of Inspiration for Household Archaeology

The explicit inspiration for household archaeology comprised the studies of households and multiscalar histories written by historians (including historians of the so-called Annales school) (Braudel, 1972), and by anthropologists, especially those interested in historical demography and historical anthropology (Arnould and Netting, 1982; Laslett and Wall, 1972; Netting et al., 1984), (Hammel and Laslett, 1974, p. 76). It is the latter who coined the term 'co-resident domestic group,' as a way to bring together the enormous diversity of the minimal social group – family, household, housefuland to use a term that was not subject to 'cultural bias' and would have value for 'objective scientific analysis.' The '"co-resident domestic group" consists of those who share the same physical space for the purpose of eating, sleeping and taking rest and leisure, growing up, childrearing and procreating' (Hammel and Laslett, 1974, p. 76).

Archaeologists envied the historians and social anthropologists their knowledge about the kinship relations and the demographic composition of the members of a domestic group that allowed them to explore the 'co-resident domestic group' in terms of variability in their size and kinship composition, and the cooperative activities of their members (Smith, 2012). Traditional scientific archaeological methodology puts limitations on the investigation of the social composition of or relations within the domestic group. In view of the dangers inherent in using architectural features or any other archaeological data to identify a 'household' on a prehistoric site, Wilk and Rathje pointed out that it was more productive for archaeologists to investigate questions such as what co-residents of a domestic group (households) do, rather than what their social form is (who lives there and how they are related) (Wilk and Rathje, 1982). Thus, for example, it was argued that to understand transformations in the social relations of production in Neolithic Southeast Europe, it was not essential to be able to identify 'households' in the historical and anthropological sense of the term (Tringham, 1990, p. 603). It was sufficient to be able to investigate changes in co-residence and the cooperative activities of kin-based co-resident domestic groups.

It is not surprising that the impetus for household archaeology came from archaeologists who had focused on the excavation and interpretation of well-preserved architectural remains. Architecture – more broadly, the built environment – provides the essential context of co-resident domestic groups. Archaeologists working in prehistoric Mesoamerica led the field (Ashmore and Wilk, 1988; Wilk and Rathje, 1982; Yaeger, 2013). Simultaneous and parallel developments were occurring in the Southwest US with archaeologists working with spectacular Pueblo architecture, such as Grasshopper Pueblo (Ciolek-Torrello, 1986; Riggs, 2007), and in Europe and the Near East with archaeologists working in Neolithic architectural contexts (Bogucki, 1988; Tringham, 1990, 1991). The Chacmool conference on Household and Community provided a welcome synthesis of the diverse routes to household archaeology in place by 1989 (MacEachern et al., 1989).

Modeling Household Activities in Archaeology

In modeling the economic and social role of co-resident domestic groups, participants of 'household archaeology' have followed social anthropologists in their models of the domestic mode of production (Chayanov, 1966 (reprinted 1986), Netting et al., 1984; Sahlins, 1972). These models characterized self-sustaining agriculturists and horticulturists who live in permanent villages and carry out relatively small-scale production (Isendahl and Smith, 2013). Such studies have emphasized the importance of the function of households as the organizing units of cooperative labor, production, consumption and sharing of resources, the transmission of property and rights from one generation to another, and the creation and maintenance of ties and alliances with other units through marriage and other forms of exchange. These same essential features of social reproduction became the objects of investigation of household archaeology (Ashmore and Wilk, 1988; Wilk and Rathje, 1982).

The study of 'what households do' has been enriched by the Marxist focus on ideology, power and social inequality. In the domestic mode of production the pattern of domination and inequality between and within different households of a settlement and between different settlements, although temporary and flexible, is thought to comprise a crucial variable in stimulating and enabling larger-scale transformations such as the intensification of production, the increase of the labor supply, and the increase in cultural complexity (Meillassoux, 1978).

Inequalities certainly exist, along with innumerable variations, between households and are the result of differences in the cycles of demographic and economic evolution through which households pass – cycles in which they expand, accumulating wealth, power, property, and members, and then contract (Cuellar, 2013). Thus, at any one time the households of a settlement will demonstrate differences in membership, the composition of the labor force, and the structure of dominance within the co-resident group and within the settlement as a whole. They will demonstrate differences in relations with groups outside the village, and, most importantly, differences in access to the products and processes of production. This concept of small-scale social inequality is the basis of a number of archaeological models that use the emerging importance of households as the unit of social reproduction to explain larger social and economic transformations (Rathje and McGuire, 1982; Tringham, 1990).

These differential patterns of social and economic relations reflecting social inequalities, however, have such a low archaeological visibility that these societies have traditionally been regarded as 'egalitarian.' The difficulty in recognizing inequality in 'egalitarian' societies is due to the fact that it is ephemeral and not cumulative. An additional significant factor, however,

is that inequality in such societies is not based on such criteria as the ownership of land and other material property, which are easily definable or easily recognizable in the archaeological data (Price and Feinman, 1995). Control of labor – not land, capital or property ownership – is the crucial variable on which inequality is based in precapitalist self-sustaining societies. Thus, those who control the circulation of labor and its products, control the social reproduction of society. At the scale of the domestic unit or household, this may be an individual within the household, or it may be 'senior' households who control the labor of 'junior' households. It has been argued that inequality at this scale creates tensions in numerous spheres and is a crucial variable in the process of socioeconomic transformation that can be observed in the archaeological record (Tringham, 1990, p. 608).

Archaeological Evidence of Household Inequality

The primary requirement for the examination of the social context of production and economic life at an archaeological site (in other words, of household archaeology) is that a number of domestic architectural units be exposed by excavation. This will enable a comparative view of co-resident domestic groups across the settlement at any one time. This information is enhanced by the provision of such a picture through time. Only when a large number of domestic structures is exposed is it possible to tell whether, for example, all the structures in a given excavated occupation horizon represent residences or parts of residences; whether the differentiated units of co-residence can also be differentiated as units of cooperative production; whether all the domestic structures were occupied at the same time; whether they all became part of the archaeological record at the same time or whether they were destroyed or abandoned at different times, and, finally whether they appear to be in a parallel stage of their use-lives at the time of destruction or abandonment or whether some seem to have been more recently established than others (Paynter, 1989). The study of archaeological architecture in terms of the use-life of the buildings has greatly enhanced these efforts for the purposes of household archaeology (Tringham, 1994). In this way the context and nature of the construction, maintenance, occupation, modification, and abandonment/destruction, and replacement of a residence all become objects of knowledge. Traditionally, they would have comprised elements that were subsumed under the formal features of the building. The investigation of a building through different stages of its use-life, brings to light many important aspects that would otherwise have been ignored; for example, modifications such as the addition or removal of rooms during the cycle of growth and decline of a domestic unit; efforts to prolong the use-life of a building by repairs and renewals; symbolic emphasis placed on particular use-life stages such as re-plastering, the destruction of old houses (for example by fire), or founding of new houses (Tringham, 1994).

Inequalities within and between domestic units are manifested archaeologically by the spatial differentiation of the production process (that is, in the production, consumption, and distribution of food and nonfood resources) (Wright, 2014). In each case the pattern of economic activity (and inequality) should have a distinctive appearance, demonstrating differential access to the materials, means, process, and products of production. The expectation is that different households will show differential access not only to the products of exchange with other groups but also to the raw materials, techniques, and equipment for making use of resources, such as food items, minerals, and ceramics. Such a differentiation of economic activity is frequently, but not necessarily, accompanied by specific styles of material elaboration, such as ceramic decoration.

Trends in Household Archaeology

For many years household archaeology remained marginal to the mainstream of processual (new) archaeology. In his review of Ashmore and Wilk's edited volume on Household Archaeology in Mesoamerica, David Freidel noted that for the interest in and impact of 'household archaeology' to be sustained, it would need to be incorporated into a more ambitious theoretical program in archaeology (Freidel, 1989). Its incorporation into Marxist models of power and social inequality certainly helped in this respect. The theoretical basis that has contributed most to sustaining interest in household archaeology, however, has been that of feminist anthropology and theories of social practice (Bourdieu, 1977; Moore, 1988). A common theme of these diverse poststructural theories has been their development beyond neo-Marxian models of power and labor control to models of how ideologies of gender, class, labor, production, identity, etc., are constructed, maintained and transformed. They focus on domestic arenas and the contexts of daily practice as being the most significant places where these negotiations between social agents take place. During the later 1990s these models came to characterize the many faces of post-processual archaeology, one of which is an archaeology of households that has transformed into an archaeology of gendered places (Tringham, 1994).

In the archaeology of gendered places, the archaeological record is not only analyzed and observed at a microscale, but also interpreted at a microscale (Casella and Fowler, 2005). It challenges the search for the universal category of 'kin-based co-residential domestic group' and embraces the richness of the diversity of the social context of domestic action that has been forefronted in social anthropology for many years (Moore, 1988). It lifts the household domain out of its previous position of 'assumed general knowledge' and as marginal to the 'great events' of history. It provides a richly-based, multiscalar theoretical framework to consider the role of the labor of players in prehistory: women, men and children (Tringham, 1991).

The theoretical basis of feminist anthropology and social practice theory obviously creates some serious epistemological challenges to the scientific logical positivist methodology that characterizes mainstream processual (new) archaeology. It is as difficult to identify 'gender' in archaeological architecture as it is to find 'households' and 'families.' Responses to this challenge have been articulated in many publications (Hodder, 1999; Wylie, 1992). They are based in a celebration of the ambiguity of archaeological data, a critical awareness of the scientific process of inference, and a richly researched

utilization of sources for interpretation – ethnography, history, art and literature. Above all in the post-processual epistemological framework, archaeology is identified as an interpretive discipline rather than a hypothetico-deductive search for true (pre)history (Baugher-Perlin and Spencer-Wood, 2010). Thus in order to think about the prehistoric construction and practice of gender ideologies and dominance relations within or between households, a researcher may interpret the archaeological data and their meaning in many ways. It is not required to identify gender or 'domestic unit' in the archaeological record.

The database for an archaeology of gendered places remains especially but not exclusively in the record of the built environment. Along with the nature of household archaeology itself, however, the excavation and interpretation of archaeological architecture has changed. According to post-processual concepts, the built environment can be thought of as both a passive container that has a use-life and reflects social action and as an 'active' arena or medium for and symbolic expression of social actions. Thus social actions, practices, negotiations, and dominance ideologies of whoever passes through a building are constrained and channeled by the materiality of that building. (Tringham, 1994). In practice this has led to the necessity of retrieving a highly detailed microstratigraphic record of the traces of human activity, in which the sequence of every depositional event is recorded and interpreted. It is time-consuming, expensive, but immensely rewarding (Hodder, 1999; Tringham, 1994).

Along similar lines, more recent household archaeology has embraced the ambiguities recognized by practice theory anthropologists about kinship categories, including 'family' and 'household.' In their view, it is the meaning and cultural value ascribed to a relationship that are significant rather than some set of biological ties (Gillespie, 2000). The enactment of kinship often involves the negotiation of identity and the identification of relationships through particular places on the landscape – buildings and sets of material culture (Joyce, 2000).

At the beginning of the twenty-first century, 'what households do' is still a focus of household archaeology, but the 'doing' is far more complex than physical action. Moreover, a 'co-resident domestic group' is no longer a faceless economic category, but comprises the complex life-histories of many people as well as buildings and places. Finally, household analysis is no longer a step along the ultimate goal of understanding social evolutionary change. It plays an essential role in the creation of a multiscalar history that considers such questions as how – in practice – was a sense of place created; how was social memory about this group maintained; what were the myths about garbage; what would the walls of a house say if only they could speak (Tringham, 2000).

See also: Ancestors, Anthropology of; Anthropology, History of; Ethnoarchaeology; Family and Kinship, History of; Gender, Feminism, and Sexuality in Archaeological Studies; Historical Demography; Horticultural Societies; Kinship in Anthropology; Marxist Archaeology; Mesoamerica, Archaeology of; North America, Archaeology of.

Bibliography

Allison, P., 2002. The Archaeology of Household Activities. Routledge, London.
Arnould, E.J., Netting, R.M., 1982. Households: changing form and function. Current Anthropology 23, 571–576.
Ashmore, W., Wilk, R.R., 1988. House and household in the Mesoamerican past: an introduction. In: Wilk, R.R., Ashmore, W. (Eds.), Household and Community in the Mesoamerican Past. University of New Mexico Press, Albuquerque, NM, pp. 1–28.
Baugher-Perlin, S., Spencer-Wood, S.M., 2010. In: Baugher, Sherene, Spencer-Wood, Suzanne M. (Eds.), Archaeology and Preservation of Gendered Landscapes [Electronic Resource]. Springer, c2010, New York, London.
Bogucki, P., 1988. Forest Farmers and Stockherders: Early Agriculture and its Consequences in North-Central Europe. New Studies in Archaeology. Cambridge University Press, Cambridge, UK.
Bourdieu, P., 1977. Outline of a Theory of Practice. Cambridge University Press, Cambridge, UK [trans. Nice R].
Braudel, F., 1972. The Mediterranean and the Mediterranean World in the Age of Philip II. Harper and Row, New York [trans. Reynolds S].
Casella, E., Fowler, C., 2005. The Archaeology of Plural and Changing Identities : Beyond Identification. Kluwer Academic/Plenum, New York.
Chayanov, A.V., 1966. The Theory of Peasant Economy. University of Wisconsin Press, IL (reprinted 1986).
Ciolek-Torrello, R., 1986. Room function and household at Grasshopper Pueblo. In: Benson, C., Upham, S. (Eds.), Mogollon Variability. New Mexico State University, The University Museum, Las Cruces, NM, pp. 107–119. Occasional Papers 15.
Cuellar, A.M., 2013. The Archaeology of Food and Social Inequality in the Andes. Journal of Archaeological Research (2), 123.
Flannery, K.V., Winter, M.C., 1976. Analyzing household activities. In: Flannery, K.V. (Ed.), The Early Mesoamerican Village. Academic Press, New York.
Freidel, D., 19 May 1989. Household and community in the Mesoamerican past. Science 863–864.
Gillespie, S., 2000. Beyond kinship: an introduction. In: Gillespie, S.D., Joyce, R.A. (Eds.), Beyond Kinship: Social and Material Reproduction in House Societies. University of Pennsylvania Press, Philadelphia, PA, pp. 1–21.
Hammel, E.A., Laslett, P., 1974. Comparing household structure over time and between cultures. Comparative Studies in Society and History 16, 73–109.
Hodder, I., 1999. The Archaeological Process: An Introduction. Blackwell, Oxford, UK.
Isendahl, C., Smith, M.E., 2013. Sustainable agrarian urbanism: the low-density cities of the Mayas and Aztecs. Cities 132.
Joyce, R., 2000. Heirlooms and houses: materiality and social memory. In: Gillespie, S., Joyce, R. (Eds.), Beyond Kinship: Social and Material Reproduction in House Societies. University of Pennsylvania Press, Philadelphia, PA, pp. 189–212.
Laslett, P., Wall, R. (Eds.), 1972. Household and Family in Past Time. Cambridge University Press, Cambridge, UK.
MacEachern, S., Archer, D.J.W., Garvin, R.D. (Eds.), 1989. Households and Communities. The 21st Annual Conference. Chacmool Calgary, AB.
Meillassoux, C., 1978. The economy' in agricultural self-sustaining societies: a preliminary analysis. In: Seddon, D. (Ed.), The Relations of Production: Marxist Approaches in Economic Anthropology. Frank Cass, London, pp. 127–157 [trans. Lackner H].
Moore, H.L., 1988. Feminism and Anthropology. University of Minnesota Press, Minneapolis, MN.
Netting, R.McC., Wilk, R.R., Arnould, E.J. (Eds.), 1984. Households: Comparative and Historical Studies of the Domestic Group. University of California Press, Berkeley, CA.
Paynter, R., 1989. The archaeology of equality and inequality. Annual Review of Anthropology 18369–18399.
Price, T., Feinman, G.M., 1995. In: Price, T. Douglas, Feinman, Gary M. (Eds.), Foundations of Social Inequality. Plenum Press, New York.
Rathje, W.L., McGuire, R.H., 1982. Rich men … poor men. American Behavioral Scientist 25 (6), 705–716.
Riggs, C., 2007. Architecture and identity at Grasshopper Pueblo, Arizona. Journal of Anthropological Research 63 (4), 489–513.
Robin, C., 2013. Archaeology and Everyday Life: Understanding Everyday Life. University Press of Florida.
Robin, C., 2013. Social Theory and Everyday Life. University Press of Florida.
Sahlins, M., 1972. Stone Age Economics. Aldine, Chicago.
Smith, M., 2011. Empirical urban theory for archaeologists. Journal of Archaeological Method & Theory 18 (3), 167–192.
Smith, M. (Ed.), 2012. The Comparative Archaeology of Complex Societies. Cambridge University Press, New York.

Tringham, R., 1990. Conclusion: Selevac in the wider context of European prehistory. In: Tringham, R., Krstic, D. (Eds.), Selevac: a Neolithic Village in Yugoslavia, Monumenta Archaeologica No. 15. UCLA Institute of Archaeology Press, Los Angeles, pp. 567–616.

Tringham, R., 1991. Households with faces: the challenge of gender in prehistoric architectural remains. In: Gero, J.M., Conkey, M.W. (Eds.), Engendering Archaeology: Women and Prehistory. Basil Blackwell, Oxford, UK, pp. 93–131.

Tringham, R., 1994. Engendered places in prehistory. Gender, Place, and Culture 1, 169–203.

Tringham, R., 2000. The continuous house: A view from the deep past. In: Gillespie, S., Joyce, R. (Eds.), Beyond Kinship: Social and Material Reproduction in House Societies. University of Pennsylvania Press, Philadelphia, PA.

Wilk, R.R., Rathje, W.L., 1982. Household archaeology. American Behavioral Scientist 25 (6), 617–640.

Wright, K., 2014. Domestication and inequality? Households, corporate groups and food processing tools at Neolithic Catalhoyuk. Journal Of Anthropological Archaeology 331–333.

Wylie, A., 1992. The interplay of evidential constraints and political interests: recent archaeological research on gender. American Antiquity 57, 15–35.

Yaeger, J., 2013. Maya household archaeology and settlement survey, then and now. Antiquity (337), 921.

Household Economics

Shoshana Grossbard, San Diego State University, San Diego, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Household economics analyzes all decisions made by households at both the microeconomic and macroeconomic level. One influential tradition based on the New Home Economics emphasizes the role of household production and takes a microeconomic perspective. It includes unitary models conceiving of the household as a unified decision-making unit and models recognizing that individual members of multiperson households keep their own preferences and constraints. Applications include all decisions possibly made by households, such as labor supply, consumption, savings, demand for health, time uses other than paid work, household financial arrangements, migration, marriage and divorce, other sexual behavior, fertility, demand for education, intergenerational transfers, and home-leaving by adult children. Household economists are interested in the effects of a wide range of government policies on household decision making. Many of these applications of household economics have been pursued independently, but recently there is a tendency for more integration and cross-fertilization between the various models that can be classified under the umbrella of 'household economics.'

Household economics analyzes all decisions made by households, including those regarding consumption, savings, labor supply, leisure, home production, health, education, fertility, marriage, divorce, environmental protection, emotional well-being, and participation in institutions.

There are many kinds of households. Most – but not all – households are families living together, and therefore household economics overlaps with economics of the family. It also overlaps with demographic economics.

History

Even though Malthus and some other classical economists wrote on fertility and other household decisions, at the end of the nineteenth century the marginalist revolution moved the focus of economics to market transactions and pushed part of household economics into a black box. The parts of household economics that continued to be studied by mainstream economists include consumption and labor supply, two areas closely tied to the monetary economy. During this period, some agricultural and home economists studied household production (e.g., Reid, 1934). The New Home Economics (NHE), a movement toward the reintegration of the household into economic analysis, was initiated by Nobel laureate Gary Becker and Jacob Mincer at Columbia University in the 1960s (Grossbard-Shechtman, 2001; Grossbard, 2011).

Some Key Modeling Assumptions

Five elements of theoretical models of the household are considered here: number of household members, subject matter of the decision, type of decision making, consideration of household production, and number of time periods considered.

The most common household models assume one-person households and analyze decisions regarding consumption, savings, and labor supply. Regarding type of decision making, if multiperson households act as if they were one unit the models are called 'unitary,' a term introduced by Browning et al. (1994). Some unitary models assume that the household has one altruistic head, which leads everyone in the household to act in the household's best interest, even a 'Rotten Kid' (Becker, 1981). A problem with unitary models is what Apps and Rees (2009) call the anonymity implication: unitary models assume that it does not matter who within the household earned a particular income. This implication has been refuted by many (a partial summary is provided by Haddad et al., 1997).

Alternative models have assumed that individual members of multiperson households keep their own preferences and constraints (Donni and Chiappori, 2011). These alternatives include consensual models, models of independent individual decision making, and bargaining models. Consensual models follow Samuelson (1956) in assuming that households have a social welfare function reflecting a household consensus. Independent individual models – such as Becker's (1973) competitive marriage models, Keeley's (1977) marital search model, and Grossbard-Shechtman's (1984) model of marriage and labor markets – assume that marriage markets are in equilibrium and that individuals with separate preferences and constraints agree on exchanges based on equilibrium prices established in markets. Bargaining models assume that household members play games of a cooperative (e.g., McElroy and Horney, 1981; Chiappori, 1988) or a noncooperative nature (e.g., Konrad and Lommerud, 1995). All three types of nonunitary model analyze how outcomes vary according to factors such as age, sex ratios in marriage markets, and divorce laws on intrahousehold distribution of goods and individual access to leisure (and thus labor supply). Becker (1973) coined such factors market factors, McElroy (1990) called them extrahousehold environmental parameters (EEP), and Browning et al. (1994) 'distribution factors.'

Household production plays a key role in many household decision-making models, starting with the original NHE models of consumption (Mincer, 1963; Becker, 1965; Lancaster, 1966) and labor supply (Mincer, 1962). Models also differ in terms of the number of periods that they consider.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.94008-X

Microeconomists tend to develop one-period models, an exception being Ghez and Becker's (1975) early dynamic model of consumption and labor supply. Macroeconomists – who often analyze time series – are more likely to consider changes over time. A recent survey of macrolevel time use analyses can be found in Aguiar et al. (2012). Furthermore, related economic analyses of home production in farming households have continued within the discipline of agricultural economics (e.g., Taylor and Adelman, 2003).

A Wide Range of Applications

Household economics continues to cover the study of labor supply, consumption, and savings, as was the case before the arrival of the NHE. However, in the 50 years since the birth of the NHE, many more areas of application have sprung up.

Grossman (1972), also part of the NHE, pioneered the use of household production models in analyses of household demand for health (see also Anderson and Grossman, 2009). This has become a large field, including the study of obesity (e.g., Fertig et al., 2009) and other harmful behaviors. Among the early applications of household production models to the study of time uses other than paid work are those by Biddle and Hamermesh (1990) and the analyses reviewed by Juster and Stafford (1991). Time allocation models examine a variety of time uses, from sleep (Biddle and Hamermesh, 1990) to social interactions (Saffer and Lamiraud, 2012) and household activities conserving the environment (Matsumoto, 2013).

Household economics includes the study of household financial arrangements (e.g., Woolley, 2003; Negrusa and Oreffice, 2011) and migration. Also included are the study of marriage and divorce, applications that originated with Becker (1973), and of other sexual behavior, including timing of first intercourse (e g. Cannonier, 2012) and marital fidelity (Fair, 1978; Smith, 2012; Bishai and Grossbard, 2010). Unmarried cohabitation among heterosexual couples (Grossbard-Shechtman, 1982) and the division of chores work inside the household are also part of the subject matter (Becker, 1981; Gimenez-Nadal and Sevilla, 2012). The formation of same-sex couples is a recent and growing application of household economics (e.g., Jepsen and Jepsen, 2002; Badgett et al., 2008; Grossbard and Jepsen, 2008; Negrusa and Oreffice, 2011).

Children are a major focus of household economics. Microeconomic analyses of fertility were pioneered by Leibenstein (1957) and Becker (1960) and have generally assumed that couples made fertility decisions as one unit. Browning (1992) summarized a large portion of the economic literature dealing with children and published prior to 1980. More recent studies examining women's decision to have a child either alone or as a couple include that by Ekert-Jaffe and Grossbard (2008). Analyses of children based on a collective model also introduce individual differences between the spouses (e.g., Cherchye et al., 2012). Parents invest in their children in different ways, including caring for them (Folbre and Yoon, 2007), paying for their education, choosing between sending them to school and to work (Kis-Katos, 2012), and giving them monetary transfers (Becker, 1981). Economists study such investments in the context of a wide range of children outcomes (e.g., Currie, 2011; Conway and Li, 2012; and Cunha et al., 2010). Gifts of time and money from children to older parents have also been analyzed by household economists (e.g., Becker, 1981; Pezzin et al., 2009) and so have unmarried adult children's decisions to live away from their parents (e.g., Le Blanc and Wolff, 2006; Del Boca and Chiuri, 2010).

Policy, Laws, and Technology

Household economics is interested in the effects of a wide range of government policies on household decision making. This encompasses inter alia the effects of taxes on labor supply (e.g., Beninger et al., 2006), of drug policies on drug consumption (e.g., Becker et al., 2006), and of paid parental leaves on fertility and labor supply (e.g., Datta Gupta et al., 2008).

Household economics also analyzes how laws affect outcomes of household decision making. This includes studies of the effects of divorce laws on divorce (e.g., Becker, 1981; Wolfers, 2006) and labor force participation (e.g., Peters, 1986; Chiappori et al., 2002), marriage laws on economic growth (e.g., Schoellman and Tertilt, 2006) and fertility (e.g., Guttierez and Suarez Beccera, 2012), property laws on maintenance of physical capital (e.g., Ellickson, 2008), and bride price refundability on sexual fidelity (Bishai and Grossbard, 2010).

A relatively new area of research deals with the effects of technological changes in household production on economic growth (e.g., Greenwood et al., 2005) and fertility (Bailey and Collins, 2011). How changes in contraceptive technology have affected fertility and wages (Bailey et al., 2012) as well as labor supply (Goldin and Katz, 2002; Oreffice, 2007) is also a relatively new direction for research on household economics. With the increased availability of merged household and firm data it is hoped that economists will also produce analyses of how firms' policies toward their workers affect various household outcomes such as labor supply and fertility (Grossbard-Shechtman, 2003).

Toward More Integration

Since the start of the NHE in the 1960s, household economics has experienced big leaps forward. However, these leaps have been uneven, with some applications and methodologies leaping more than others. For a long time microhousehold economics dominated, but in recent years the growth in macrohousehold economics is noticeable. In part, it has been propelled by the increased availability of time use data both at the national and international levels. However, the expansion of household economics into the study of outcomes not previously considered part of economics characterizes the microapplications of household economics more than its macrolevel applications, and the more empirically oriented researchers more than the scholars specialized in theory.

More integration and cross-fertilization between the various branches of household economics would be desirable. Given that similar theoretical and empirical models can often be applied to many different kinds of outcomes being researched, there is much to gain from models developed separately for the

study of a restricted set of household outcomes being exported to researchers dealing with different household outcomes. This holds for both theoretical and empirical models and micro- and macroeconomic models of household decisions.

Journals such as the *Review of Economics of the Household* and the *Journal of Population Economics,* as well as edited books such as *Household Economic Behaviors* (Molina, 2011), contribute to such cross-fertilization by publishing a variety of models dealing with household economics.

See also: Families and Households, Formal Demography of; Family Theory: Economics of Childbearing; Family Theory: Economics of Intergenerational Relations; Family Theory: Economics of Marriage and Divorce; Gender, Economics of; Labor Supply; Time Use Research: Recent Developments; Time Use and Gender.

Bibliography

Aguiar, M., Hurst, E., Karabarbounis, L., 2012. Recent developments in the economics of time use. Annual Review of Economics 4, 373–397.

Anderson, R., Grossman, M., 2009. Health and the household. Review of Economics of the Household 7, 219–226.

Apps, P.F., Rees, R., 2009. Public Economics and the Household, Cambridge. Cambridge University Press.

Badgett, M.V.L., Gates, G., Maisel, N., 2008. Registered domestic partnerships among gay men and lesbians: the role of economic factors. Review of Economics of the Household 6, 327–346.

Bailey, M.J., Collins, W.J., 2011. Did improvements in household technology cause the baby boom? Evidence from electrification, appliance diffusion, and the Amish. American Economic Journal: Macroeconomics 3, 189–217.

Bailey, M., Hershbein, B.J., Miller, A., 2012. The opt-in revolution? Contraception, fertility timing and the gender gap in wages. American Economic Journal: Applied Economic 4, 225–254.

Becker, G.S., 1960. An economic analysis of fertility. In: National Bureau Committee for Economic Research, Demographic and Economic Change in Developed Countries, a Conference of the Universities. Princeton University Press, Princeton, NJ.

Becker, G.S., 1965. A theory of allocation of time. Economic Journal 75, 493–515.

Becker, G.S., 1973. A theory of marriage: part I. Journal of Political Economy 81, 813–846.

Becker, G.S., 1981. A Treatise on the Family, first ed. Harvard University Press, Cambridge.

Becker, G.S., Murphy, K.M., Grossman, M., 2006. The market for illegal goods: the case of drugs. Journal of Political Economy 114, 38–60.

Beninger, D., et al., 2006. Evaluating the move to a linear tax system in Germany and other European countries. Review of Economics of the Household 4, 159–180.

Biddle, J., Hamermesh, D., 1990. Sleep and the allocation of time. Journal of Political Economy 98, 922–943.

Bishai, D., Grossbard, S., 2010. Far above rubies: the association between bride price and extramarital sexual relations in Uganda. Journal of Population Economics 23, 1177–1188.

Browning, M., 1992. Children and household economic behavior. Journal of Economic Literature 30, 1434–1475.

Browning, M., Bourguignon, F., Chiappori, P.-A., Lechene, V., 1994. Incomes and outcomes: a structural model of intra-household allocation. Journal of Political Economy 102, 1067–1096.

Cannonier, C., 2012. State abstinence education programs and teen birth rates in the US. Review of Economics of the Household 10, 53–76.

Cherchye, L., De Rock, B., Vermeulen, F., 2012. Married with children: a collective labor supply model with detailed time use and intrahousehold expenditure information. American Economic Review 102, 3377–3405.

Chiappori, P.-A., 1988. Rational household labor supply. Econometrica 56, 63–90.

Chiappori, P.-A., 2011. Collective labor supply with many consumption goods. Review of Economics of the Household 9, 207–220.

Chiappori, P.-A., Fortin, B., Lacroix, G., 2002. Marriage market, divorce legislation, and household labor supply. Journal of Political Economy 110, 37–71.

Conway, Karen S., Li, M., September 2012. Family structure and child outcomes: a high definition, wide angle snapshot. Review of Economics of the Household 10, 345–374.

Cunha, F., Heckman, J.J., Schennach, S.M., 2010. Estimating the technology of cognitive and noncognitive skill formation. Econometrica 78, 883–931.

Currie, J., 2011. Inequality at birth: some causes and consequences. American Economic Review 101, 1–22.

Datta Gupta, N., Smith, N., Verner, M., 2008. The impact of Nordic countries' family friendly policies on employment, wages, and children. Review of Economics of the Household 6, 65–89.

Del Boca, D., Chiuri, M.C., 2010. Home-leaving decisions of daughters and sons. Review of Economics of the Household 8, 393–408.

Donni, O., Chiappori, P.-A., 2011. Non-unitary models of household behavior: a survey of the literature. In: Molina, J.A. (Ed.), Household Economic Behaviors. Springer, New York.

Ekert-Jaffe, O., Grossbard, S., 2008. Does community property discourage unpartnered births? European Journal of Political Economy 24, 25–40.

Ellickson, R.C., 2008. The Household: Informal Order Around the Hearth. Princeton University Press, Princeton.

Fair, R., 1978. A theory of extramarital affairs. Journal of Political Economy 86, 45–61.

Fertig, A., Glomm, G., Tchernis, R., 2009. The connection between maternal employment and childhood obesity: inspecting the mechanism. Review of Economics of the Household 7, 227–256.

Folbre, N., Yoon, J., 2007. What is childcare? Lessons from time use surveys from major English-speaking countries. Review of Economics of the Household 5, 223–248.

Ghez, G., Becker, G.S., 1975. The Allocation of Time and Goods over the Life Cycle. Columbia University Press, New York.

Gimenez-Nadal, J.I., Sevilla, A., 2012. Trends in time allocation: a cross-country analysis. European Economic Review 56, 1338–1359.

Goldin, C., Katz, L., 2002. The power of the pill: oral contraceptives and women's career and marriage decisions. Journal Political Economy 110, 730–770.

Greenwood, J., Seshadri, A., Yorukoglu, M., 2005. Engines of liberation. Review of Economic Studies 72, 109–133.

Grossbard, S., 2011. Independent individual decision-makers in household models and the new home economics. In: Molina, J.A. (Ed.), Household Economic Behaviors. Springer, New York.

Grossbard-Shechtman, A., 1982. A theory of marriage formality: the case of Guatemala. Economic Development and Cultural Change 30, 813–830.

Grossbard-Shechtman, A., 1984. A theory of allocation of time in markets for labor and marriage. Economic Journal 94, 863–882.

Grossbard-Shechtman, A., 2001. The New Home Economics at Columbia and Chicago. Feminist Economics 7, 103–130.

Grossbard-Shechtman, A., 2003. Introduction. In: Grossbard-Shechtman, S. (Ed.), Marriage and the Economy. Cambridge University Press, Cambridge, NY.

Grossbard, S., Jepsen, L., 2008. The economics of gay and lesbian couples: introduction to a special issue on gay and lesbian households. Review of Economics of the Household 6 (4), 311–325.

Grossman, M., 1972. On the concept of health capital and the demand for health. Journal Political Economy 80, 223–255.

Gutierrez, E., Suarez Becerra, P., 2012. The relationship between civil unions and fertility in France: preliminary evidence. Review of Economics of the Household 10, 115–132.

Haddad, L., Hoddinott, J., Alderman, H., 1997. Intrahousehold Resource Allocation in Developing Countries: Models, Methods and Policy. John Hopkins University Press, Baltimore and London.

Jepsen, L., Jepsen, C., 2002. An empirical analysis of the matching patterns of same-sex and opposite-sex couples. Demography 39, 435–453.

Juster, F.T., Stafford, F.P., 1991. The allocation of time: empirical findings, behavioral models, and problems of measurement. Journal of Economic Literature 29, 471–522.

Keeley, M.C., 1977. The economics of family formation. Economic Inquiry 25, 238–250.

Kis-Katos, K., 2012. Gender differences in work-schooling decisions in rural North India. Review of Economics of the Household 10, 491–519.

Konrad, C., Lommerud, K.E., 1995. Family policies with non-cooperative families. Scandinavian Journal of Economics 97, 581–601.

Lancaster, K.J., 1966. A new approach to consumer theory. Journal of Political Economics 74, 132–157.

Le Blanc, D., Wolff, F.-C., 2006. Leaving home in Europe: the role of parents' and children's incomes. Review of Economics of the Household 4, 53–73.

Leibenstein, H., 1957. Economic Backwardness and Economic Growth. John Wiley, New York.
Matsumoto, S., 2013. Spouses' time allocation to pro-environmental activities: who is saving the environment at home? Review of Economics of the Household. Online first. http://link.springer.com/article/10.1007/s11150-013-9177-x.
McElroy, M.B., 1990. The empirical content of Nash-bargained household behavior. Journal of Human Resources 25, 559–583.
McElroy, M.B., Horney, M.J., 1981. Nash-bargained household decisions: toward a generalization of the theory of demand. International Economic Review 22, 333–349.
Mincer, J., 1962. Labor force participation of married women: a study of labor supply. In: Gregg, L.H. (Ed.), Aspects of Labor Economics. Princeton University Press, Princeton, NJ.
Mincer, J., 1963. Market prices, opportunity costs, and income effects. In: Christ, C. (Ed.), Measurement in Economics. Stanford University Press, Stanford, CA.
Molina, J.A. (Ed.), 2011. Household Economic Behaviors. Springer, New York.
Negrusa, B., Oreffice, S., 2011. Sexual orientation and household financial decisions: evidence from couples in the United States. Review of Economics of the Household 9, 445–463.
Oreffice, S., 2007. Legalization of abortion, intra-household bargaining power, and labor supply. Review of Economics of the Household 5, 181–207.
Peters, H.E., 1986. Marriage and divorce: informational constraints and private contracting. American Economic Review 76, 437–454.
Pezzin, L.E., Pollak, R., Schone, B., 2009. Long-term care of the disabled elderly: do children increase caregiving by spouses? Review of Economics of the Household 7, 323–339.
Reid, M., 1934. The Economics of Household Production. John Wiley and Sons, New York.
Saffer, H., Lamiraud, K., 2012. The effect of hours of work on social interaction. Review of Economics of the Household 10, 237–258.
Samuelson, P., 1956. Social indifference curves. Quarterly Journal of Business and Economics 70, 1–22.
Schoellman, T., Tertilt, M., 2006. Marriage laws and growth in Sub-Saharan Africa. American Econonomic Review: Papers and Proceedings 96, 295–298.
Smith, I., 2012. Reinterpreting the economics of extramarital affairs. Review of Economics of the Household 10, 319–344.
Taylor, J., Adelman, I., 2003. Agricultural household models: genesis, evolution, and extensions. Review of Economics of the Household 1, 33–58.
Wolfers, J., 2006. Did unilateral divorce laws raise divorce rates? A reconciliation and new results. American Economic Review 96, 1802–1820.
Woolley, F., 2003. Control over money in marriage. In: Grossbard-Shechtman, S. (Ed.), Marriage and the Economy. Cambridge University Press, New York.

Households in Anthropology

Sylvia J Yanagisako, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Anthropologists use the term household loosely to refer to a residential unit, whose members cooperate in some activities of production, consumption, or child rearing. They differentiate the household from the family by defining the former as a residential unit and the latter as a kinship unit. Close scrutiny of this distinction reveals key assumptions that have shaped the way anthropologists and other social scientists have conceptualized and studied not only households, but also families, domestic groups, and kinship.

Problems in Defining the Household

The grounds for analytically separating families and households lie in the broader claim that kinship and locality are two distinct principles of organization (Keesing, 1958). This claim, in turn, derives from the discipline's focus in the nineteenth and early twentieth centuries on the principles organizing social groups and relations in nonstate societies. Kinship, which was assumed to be universally defined on the basis of genealogical linkages rooted in biological reproduction, was considered the most important among these principles. As a key mode of recruiting members to (and limiting membership in) politically important groups such as lineages and clans, kinship (genealogical propinquity) was distinguished from locality (spatial propinquity), which defined residential units such as hamlets and villages. As anthropologists expanded their range of ethnographic research to include peasant communities, they drew on the distinction between kinship and locality to differentiate family from household. The family as a corporate kin group was distinguished from the household as a collection of kin and sometimes nonkin who share a common residence. The corporate character of the peasant family was viewed as rooted in the jural rights to property, usually land, which its members held in common. Family members supposedly retained these rights regardless of whether they were members of the same household. Hence, family and household, it was argued, are both 'logically distinct' and 'empirically different' (Bender, 1967: 493).

Actual usage of the term household in the anthropological literature, however, calls into question both the logical distinction and the empirical difference between families and households. For one, residential units composed of individuals who are not genealogically linked, such as orphanages, college dormitories, men's houses (in societies where gender segregation among adults is practiced), convents, and army barracks, usually are not considered households. At the same time, however, families whose members move between residential units that are located on different continents are referred to as 'transnational households' (Basch et al., 1994; Jones and de la Torre, 2011). As a general practice, anthropologists attach the term household only to groups of people that share both a common living space and activities that are identified as domestic. Domestic activities include those activities related to food production and consumption or to reproduction, including childbearing and child rearing. As these are the very same activities that are considered the functional basis of the 'family,' in practice, households are identified using both 'kinship' and 'residence' criteria. Thus, the claim for the logical distinction and empirical difference between family and household is not affirmed by scholarly practice.

Because 'domestic' activities are sometimes carried out by groups of people who do not reside together, some anthropologists (Bender, 1967) have suggested that it would do better to employ alternate terms in our ethnographic descriptions. Among the alternatives that have proposed are 'coresidential group,' 'budget unit,' and 'domestic group.' Less important than the terms employed, however, is the need for ethnographers to describe precisely the composition, organization, activities, and functions of these units. In addition, the cultural meanings people attach to them must be explicated (Yanagisako, 1984). In the absence of this knowledge, anthropologists are likely to project their own cultural meanings and expectations on those units.

The difficulties of identifying households are illustrated by innumerable cases (see Segalen, 1984), which demonstrate that neither the boundaries nor the functions of household are easily discerned. Questions have been raised as to how to treat residential groups that move through a seasonal cycle of dispersal and concentration, how to handle the movement of people between dwelling units, whether to define as a single household the huts or houses that share a common yard or those that share a common stove, and whether to include servants, apprentice, boarders, lodgers, and nonkin in the household. More recently, studies of transnational migration (Rouse, 1991; Basch et al., 1994; Parreñas, 2005; Boehm, 2012) have raised the question as to whether residential units located on different continents constitute a single household. As the speed and ease of long-distance communication have been revolutionized by new media, including instant messaging, e-mail, and video calling, some scholars (Madianou and Miller, 2011) have claimed that the nature of family relationships is changing. Although the emerging scholarship on digital media and migration has focused on families rather than households – sometimes slipping between these terms – it remains to be seen whether new forms of communication will lead to the reconceptualization of what constitutes a household. Since studies of gay and lesbian kinship (Weston, 1991) effectively challenged the assumption that heterosexual marriage and parenthood are the universal,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12088-4

natural core of families, studies of gay and lesbian households have increased.

Problems in the Classification of Households Types

Much of the literature on households has been concerned with explaining variation among households that are distributed over time and space. In describing this variation, most scholars have focused on the character of the genealogical links between the members of the household as its defining feature. Terms, such as 'stem-family household,' 'nuclear family household,' and 'patrilateral extended household' are based on the configuration of genealogically defined kin types included in the household – in other words, its household composition. In addition, the most genealogically 'close' members of the household are assumed to form its structural core. If a household contains two adult brothers, their wives, and children, it is labeled a fraternal-joint household. The assumption here is that the relationship between the brothers constitutes the structural core that shapes the relations of all the members of the household. Accordingly, studies of household change over time track changes in the genealogical composition of the household.

The classification of households on the basis of their genealogical composition rests on the shaky assumption that people who are similarly linked genealogically will have the same or greatly similar social relations. In other words, it is assumed that the labeling of a household as a 'stem-family household' captures its organizational structure and function. But, a stem-family household found at one time and place does not necessarily share the same organizational dynamics as a stem-family household at another time and place. A household, or any social unit, is much more than the sum of its genealogical links. It is the negotiated outcome of the culturally meaningful actions and relations among its members. The actions and relations of households' members with identical genealogically defined links can differ radically. Even households with the same genealogical composition in the same community may have significantly different structures of authority and divisions of labor, and they may engage in a range of different activities (Yanagisako, 1984).

What Explains Variation among Households?

A key question that has fueled comparative studies of households by anthropologists, sociologists, historians, demographers, and other social scientists is what explains variations among them. Two related questions have been asked of variations observed in household composition. First, why do a significant proportion of the households in a society differ from what people report to be the ideal type of household? Second, irrespective of these ideals, what explains the distribution of household types within and between societies? Three types of variables have been offered to answer both these questions: demographic variables, economic variables, and social stratification variables.

Demographic variables affect the size and composition of households, thus acting as constraints on the attainment of the culturally ideal household type. Age at marriage, life expectancy, and fertility shape the composition of households at any given time in a population. The economy of the household is also shaped by demographic processes. Among peasant for whom the household is the unit of production, changes in the demographic structure of the household as it moves through the developmental cycle entail changes in the ratio of consumers to workers (Chayanov, 1966).

Fortes's (1958) model of the developmental cycle of domestic groups draws on these demographic processes to explain the varieties of domestic groups found in a society. Fortes hypothesized that the variety of household types found in a community or society at any point in time are merely different phases in the developmental cycle of a 'single general form.' As all households pass through different stages of the developmental cycle at different times, a census taken at a particular point in time will catch only a few in the ideal, complete phase. Hence, an ideal household type, such as the patrilineal stem-family household of Japan (which includes a married couple, their adult son and his wife and children) is better regarded as "a transitory state in the development of the household, rather than as a form" (Hammel, 1975: 142).

Fortes's model has been useful in helping us understand the impact of events, such as birth, marriage, and death on the composition of households. It cannot account for all the observed variation in households, however, because other cultural and social forces (see below) are also involved. Fortes himself later (1978: 18) conceded that a model of a uniform developmental cycle could not explain why the actual histories of different families entail different developmental sequences. The model may also blind us to historical shifts wherein more recently formed households embark on a different trajectory than older ones. In addition, while demographic factors undoubtedly shape households, it is a mistake to treat them as exogenous, biological constraints on the composition and economy of households. To do so confuses social replacement with biological reproduction and overlooks the strategies people actively employ to control household size and composition, including practices of marriage, fosterage, adoption, timing and spacing of births, and timing of household division. The collection of longitudinal data on households, whether by drawing on historical sources or through retrospective oral histories, is necessary to understanding both continuity and change. Cohort analysis is an effective tool for helping ascertain similarities and differences in the household histories of members of successive cohorts, defined on the basis of year of birth, marriage, or other historically significant events experienced by the members of a community.

Economic variables are also adduced to explain cross-societal variation in household composition and organization. In some cases a mix of 'economic' and 'ecological' factors, including labor needs of production, defensive needs, care of children, taxation, and conscription practices of the state, are identified as the functional basis for the prevalence of certain household types. In others, the labor requirements of a particular productive activity are singled out as the primary determinant of household type. In their cross-societal analysis of 60 randomly selected societies, Pasternak et al. (1976) conclude that labor requirements are the most powerful determinant of the formation of extended family households.

They hypothesize that in societies where work outside the home makes it difficult for a mother to tend to children and other domestic tasks or where the outside activities of the father make it difficult for him to perform his subsistence work, extended family households will prevail. Although they are able to demonstrate a correlation between activity requirements and presence of extended family households, they lack diachronic data to support the hypothesized causal relation between their independent variable (activity requirements) and their dependent variable (presence of extended family households). Consequently, they must resort to arguing that it is more plausible that activity requirements cause extended family households rather than vice versa. It seems as plausible, however, that household composition and productive activities do not lie in a simple, unidirectional causal relation but instead shape each other.

Attempts like that of Pasternak et al. to identify the determinants of household composition and structure through statistical analysis of cross-societal comparisons have declined significantly over the past 40 years as the inherent problems and limitations of such studies have been increasingly recognized. Among these problems is that of discerning whether appropriate units have been selected for comparison. As all societies contain different frequencies of a range of household types, it is misleading to classify entire societies into those with 'extended family households' versus those with 'independent family households.' A better understanding of the relation between household composition and 'economic' factors and processes may be gained by investigating the ways in which diverse types of households in a society articulate with one another rather than by reducing the entire society to a single household type.

Labor requirements are frequently used to explain the frequency with which a particular household type appears in a society or section of society. Netting (1974) cites several ecological studies as demonstrating that household composition among horticulturalists and agriculturalists varies with the type and amount of labor required for effective crop production. His own study attributes the differences between the large, extended families of shifting cultivators, and the nuclear family households of intensive agriculturalists in the same Nigerian plateau environment to different labor requirements (Netting, 1965).

Both cross-societal and intrasocietal studies that seek the determinants of household variation in productive labor requirements assume that once having identified the primary productive activity of the household, researchers can ascertain its labor requirements and proceed to show how the household efficiently meets them. In doing so, however, they tend to overlook the fact that it is not simply technological requirements that necessitate a household type, but the entire manner in which production is socially organized. In other words, a productive activity can be accomplished by a variety of household types depending on how that production is socially organized, including the division of labor by sex and age, the use of hired labor, cooperation between households in productive tasks, and the exchange of labor between households. In assuming that the household is a self-contained labor unit, researchers fail to sufficiently investigate the existence of other forms of labor organization. The social units engaged in production may not be the same at different phases of the productive cycle. For example, different social units may work together in the planting, weeding, and harvesting of crops. To attribute household composition to labor requirements begs the question, because it assumes the prior existence of an entire social organization of production, including age and gender divisions of labor. To overcome this narrow focus on the technical requirements of production, the interplay of a wider range of cultural, political, and economic forces shaping productive relations within and between households must be considered.

Social stratification: differential access to resources, in particular differences in ownership and control of the means of production, also shape household variation in both peasant communities and urban, industrial populations. A past tendency to neglect social stratification in studies of households and other domestic groups can be attributed to the past assumption that the communities studied by ethnographers were basically homogeneous. When anthropologists studied peasant communities, for example, the stark contrast with the landowners – who were generally not studied – led them to overlook significant differences among the peasants. Consequently, there has been a tendency to overlook stratification as a significant factor in the analysis of household variation among peasants. Yet, peasant communities are rarely homogenous; instead they are usually internally differentiated. Differences in wealth among peasants are also often obscured by kinship ideals and commitments that connect the landed to the landless and the land-rich to the land-poor. Significant wealth differences may exist, however, between the smaller coresidential units of a single kin-based compound or between brothers, who head the main and branch households of a family. In other words, ideologies of family and kinship may obscure stratification among kin.

Changes in Households over Time

Anthropological studies of household and family change over time have generally focused on the effects of industrialization, urbanization, and the increasing integration of communities into capitalist market relations. Some of these challenge or argue for the refinement of the dominant social theory that family, household, and kinship decline in importance as a society 'modernizes' and becomes increasingly urbanized and industrialized. They demonstrate that, in some cases, urbanization and industrialization have strengthened the extended family rather than weakened it and that new functions are assumed by extended families and extended family households. For example, urban migration may contribute to the maintenance of extended family households. While these studies provide an important correction to the theory of modernization and kinship decline, they are marred by the inconsistent use of ambiguous terms, including 'extended family household.' The argument over whether extended families have declined with urban-industrialization is in part the result of differences in the way in which researchers define and conceptualize household type. Quite often it is assumed that people who live in extended family households pursue common strategies and cooperative actions under the direction

of a central authority – the head of the household. This not only gives undue weight to centralized authority as the defining feature of the extended family, but it also overlooks the heterogeneity in the goals and strategies of the members of a household.

Critical Approaches to Household

The adaptationist, utilitarian approach to households that prevailed in much of the anthropological literature up until the 1980s has been challenged by theoretical shifts in the discipline that have had an important impact on the discipline in general. Among these perhaps the most significant has been the emergence of feminist theories of gender.

In the 1970s, one of the first questions feminist anthropologists raised in rethinking anthropological analysis of women's status and roles was whether ethnographers had overlooked women's goals, strategies, and forms of action, especially in the 'domestic sphere.' M. Wolf's (160) study of Taiwanese joint family households, for example, revealed that family and household division was brought about as much by women's attempts to advance their own goals and those of their 'uterine family' as it was by conflicts among brothers. Collier (1974) argued that this process occurs commonly in societies where men gain power by holding together large bodies of coresident kin, but where women (especially younger women) gain power by breaking up such units. By attending to the informal power women often wield in societies where men are granted formal authority, feminist anthropologists alerted us to the tensions and conflicts and, therefore, the complex processes shaping households and other domestic groups. Likewise, investigations of women's relations with kin and others outside their household demonstrated that it is not just men who link mother–child units to larger institutional structures. The involvement of women in informal exchange networks between families and households had been overlooked by androcentric studies that had equated male relations with the 'public sphere.' Once studies showed that the female and male members of households – as well as the young and the old, those born in the household and those who had married into it – often do not share the same goals and strategies, it was no longer possible to retain the naive concept of a unified 'household strategy.' Finally, by calling into question the presumed universality and immutability of gender identities and the gender division of labor, feminist anthropologists denaturalized gender systems and called for the study of its historical and cultural production. They revealed the gender assumptions embedded in concepts of the 'public' and 'private' spheres (Reiter, 1975), in definitions of the 'domestic sphere' and 'domestic groups,' including family and household (Yanagisako, 1979) and in theories of kinship in general (Yanagisako and Collier, 1987).

Rethinking Households

Functionalist analyses of 'household adaptation' reified the household, treating it as a social actor endowed with human consciousness and motives. Households were described as having goals, fashioning strategies, and adapting to the changing pressures of a modernizing society. Such language demonstrates the ease with which an analytic construct can be reified into a social actor. The imputation of consciousness and will to households is rooted in the assumption that those who live together share a collective motive and will – in other words, a unified household strategy. It obscures the complex social processes through which people negotiate the relations that produce households, despite different and sometimes conflicting goals, strategies, and ideas about what it means to 'live together.'

Neither the composition nor the dynamic structure of households can be understood without a thorough understanding of the cultural meanings and sentiments that shape the social actions and relations of the people who live in them. This does not mean that households cannot be compared across societies, but rather that any comparative study of households must explicate the culturally specific and heterogeneous meanings and motives shaping people's residential preferences and relations. Such an expanded strategy of research and analysis will enable anthropologists to move beyond crude comparisons of distributions of genealogically defined household types within and across societies to understanding households as the negotiated outcomes of culturally meaningful action.

See also: Family: The Anthropology of the Concept and Its History; Genealogy in Anthropology; Home, Anthropology of; Kinship in Anthropology; Marriage; Matrifocality.

Bibliography

Basch, L., Schiller, N.G., Blanc, C.S., 1994. Nations Unbound: Transnational Projects, Postcolonial Predicaments, and Deterritorialized Nation-States. Gordon and Breach, Langhorne, PA.

Bender, D.R., 1967. A refinement of the concept of household: families, co-residence, and domestic functions. American Anthropologist 69, 493–504.

Boehm, D., 2012. Intimate Migrations: Gender, Family and Illegality among Transnational Mexicans. New York University Press, New York.

Chayanov, A.V., 1966. The Theory of Peasant Economy. R.D. Irwin, Homewood, IL.

Collier, J.F., 1974. Women in politics. In: Rosaldo, M.Z., Lamphere, L. (Eds.), Women, Culture, and Society. Stanford University Press, Stanford, CA.

Fortes, M., 1958. Introduction. In: Goody, J.R. (Ed.), The Developmental Cycle of Domestic Groups. Cambridge University Press, Cambridge, UK.

Fortes, M., 1978. An anthropologist's apprenticeship. Annual Review of Anthropology 7, 1–30.

Hammel, E.A., 1975. Reflections on the zadruga. Ethnologica Slavica 7, 141–151.

Jones, R., de la Torre, L., 2011. Diminished tradition of return? Transnational migration in Bolivia's Valle Alto. Global Networks 11 (2), 180–202.

Keesing, F.M., 1958. Cultural Anthropology: The Science of Custom. Rinehart, New York.

Madianou, M., Miller, D., 2011. Migration and New Media: Transnational Families and Polymedia. Routledge, New York.

Netting, R.M., 1965. Household organization and intensive agriculture: the Kofyar case. Africa 35, 422–429.

Netting, R.M., 1974. Agrarian ecology. Annual Review of Anthropology 3, 21–56.

Parreñas, R.S., 2005. Children of Global Migration: Transnational Families and Gendered Woes. Stanford University Press, Stanford, CA.

Pasternak, B., Ember, C.R., Ember, M., 1976. On the conditions favoring extended family households. Journal of Anthropological Research 32, 109–123.

Reiter, R.R., 1975. Men and women in the south of France: public and private domains. In: Reiter, R.R. (Ed.), Towards an Anthropology of Women. Monthly Review Press, New York.

Rouse, R., 1991. Mexican migration and the social space of postmodernity. Diaspora 1, 8–23.

Segalen, M., 1984. Nuclear is not independent: organization of the household in the Pays Bigouden Sud in the nineteenth and twentieth centuries. In: Netting, R.M., Wilk, R.R., Arnold, E.J. (Eds.), Households: Comparative and Historical Studies of the Domestic Group. University of California Press, Berkeley and Los Angeles, CA.

Weston, K., 1991. Families We Choose: Lesbians, Gays, Kinship. Columbia University Press, New York.

Wolf, M., 1972. Women and the Family in Rural Taiwan. Stanford University Press, Stanford, CA.

Yanagisako, S.J., 1979. Family and household: the analysis of domestic groups. Annual Review of Anthropology 8, 161–205.

Yanagisako, S.J., 1984. Explicating residence: a cultural analysis of changing households among Japanese-Americans. In: Netting, R.M., Wilk, R.R., Arnold, E.J. (Eds.), Households: Comparative and Historical Studies of the Domestic Group. University of California Press, Berkeley and Los Angeles, CA.

Yanagisako, S.J., Collier, J.F., 1987. Toward a unified analysis of gender and kinship. In: Collier, J.F., Yanagisako, S.J. (Eds.), Gender and Kinship: Essays toward a Unified Analysis. Stanford University Press, Stanford, CA.

Housing and Cities

Ray Forrest, City University of Hong Kong, Kowloon Tong, Hong Kong

© 2015 Elsevier Ltd. All rights reserved.

Abstract

We live in an urban age. Most people now live in cities and cities are mostly housing. This article emphasizes the centrality of housing in everyday life and as a focal point of social and political conflict. It highlights the connections between housing and the life course, health, education, and social status. It describes the key terms that social scientists have employed to explore housing and cities and the different theoretical perspectives that have been used to explain the distribution of urban housing.

Introduction

In the middle of the nineteenth century, Britain became the first country in the world where more people lived in cities than in rural areas. Some 150 years later, around 2010, the rest of the world had reached that point. Whichever country we are living in, we are now likely to be city dwellers. The major cities are still places of power and government. Some are still places of manufacturing. But more than ever they are places of residence. A process of proletarianization and rural–urban migration that began in England in the eighteenth century continues today on an even greater scale. From the vantage point of Western Europe or North America, industrialization and mass migrations from the countryside to rapidly expanding cities are seen as historical phenomena. However, from an African, Latin American, or Asian perspective these processes are very current and are generating many of the same kinds of urban problems that were evident in previous centuries in Northern Europe and North America. People flock to the cities in search of a better life, pulled by the prospects of better wages and pushed by social disadvantage and the land dispossessions of encroaching urbanization and corporate landgrabs (Araghi, 2009). When they get to the city, however, they often struggle to find decent housing amidst a straining infrastructure that cannot cope with the accelerating population influx. The result can be overcrowding, declining sanitary conditions, and disease.

In Engels' nineteenth century Manchester or Sims' nineteenth century London, housing conditions could be appalling. Keating's (1976) selection for the social explorers of Victorian cities offers this description:

> The room when we entered it was in a condition beyond description. The lady was washing the baby, and she made that an excuse for the dirt of everything else. Two ragged boys were sitting on the filthy floor, a dirty little girl was in a corner pulling a dirty kitten's tail, and the smoke from the untidiest grate I ever saw in my life was making the half-washed baby sneeze its little head nearly off. The family, all told, that slept in this room was seven. There was a bed and there was a sofa-so I concluded the floor must have been the resting-place of some of them. (p. 76)

Closely comparable descriptions could be found of housing conditions in contemporary cities but probably in other parts of the world. The geography of urbanization and industrialization may have changed but housing inequalities remain stark. Indeed, from a global perspective, it could be argued that housing problems in cities are getting worse rather than better. Slum housing is endemic and homelessness, overt and hidden, accounts for millions of urban residents (Tipple and Speak, 2009).

The condition and location of housing connects to all the key aspects of everyday life in the city, including schooling, employment opportunities, social status, and personal wealth. It is precisely this pivotal position of urban housing to the shaping of life chances that has made it such a politically charged and contentious field. Conflicts in cities are, and have so often been, around housing issues of residential segregation, the redevelopment of residential neighborhoods, evictions and displacements, and housing costs. The Glasgow Rent Strike in 1915 is the best-known historical example, notable because it was organized mainly by working class women and during wartime. Manuel Castells' *The City and the Grassroots* (1983) highlights the importance of housing issues in mobilizing urban protest or 'social' movements in a wide range of social and cultural contexts. Such bottom-up demands for better housing, or resistance to redevelopment, have strong class dimensions but may also be rooted in shared identities of gender, ethnicity, or sexual orientation.

Naming the Parts

At the simplest level, housing in cities is usually described and classified in terms of 'tenure' or whether people own or rent. Renters either rent from private landlords, public landlords, or from a housing association or housing society. These latter kinds of landlords may have charitable or quasicharitable status and may have religious or trade union foundations. In the past, large employers provided rental accommodation but this has generally declined in importance in most countries.

Urban home ownership, as distinct from more traditional rural-based, semifeudal ownership, can be divided between mortgaged and outright ownership. If a person is an urban home owner, the person will probably have borrowed money from a bank, building society, or some other kind of financial institution. The person only actually owns the property when that loan is paid off, typically close to or at retirement. In some societies, the family also remains an important source of financial assistance for housing, particularly for young people seeking to buy their first home.

Much of the housing literature focuses on movement or nonmovement between different types of tenures as a key element of residential differentiation in cities. The classic progression in more affluent societies is for children to leave the parental home and rent privately for a while before buying their first home. For lower income households, purchase may be too expensive and social renting may be a short-term or long-term housing destination.

Tenure patterns vary across societies as do definitions of tenures. Historically, private landlords were dominant providers in expanding cities but declined in importance as forms of home ownership and public or social landlordism became established. The private landlord class often comprised a myriad of small-scale property owners, perhaps with only one or two properties to rent. High-speed urbanization generates housing demands that necessitate a more coordinated, institutional response requiring large, private financial institutions and/or state intervention. Thus, private landlordism tended to survive best in countries where it had a strong institutional foundation, as in the former West Germany.

It is often believed that there is a close relationship between growing affluence and levels of individual home ownership – a linear model with rising incomes and rising individual property ownership growing in tandem. However, some of the richest countries in the world, such as Switzerland, Japan, or Germany, have relatively low levels of home ownership. To understand and explain the cases of Switzerland and Germany or any other country's housing patterns, and how these patterns have evolved in specific cities, thus requires an understanding of distinct social and political histories and of institutional structures and relationships. Harloe's analysis of the development of different forms of social housing after the Second World War is an excellent example of this approach. He shows the particular and contingent relationships and circumstances involved in the development of social or public housing in different societies. Kurz and Blossfeld (2004) offer an equally nuanced analysis of home ownership and social structures.

By way of summary, two dimensions of urban housing provision should be highlighted. The first relates closely to the preceding discussion of tenures but is also aimed at complicating these categories. Housing analysts have often conceptualized urban housing provision in terms of 'commodification' and 'decommodification.' In other words, what proportion of a city's housing is provided via market mechanisms or allocated on some other basis such as bureaucratic assessment of need? Second, how important are informal processes? How far are property rights formally recognized and formally regulated? Here, the author refers most typically to unauthorized or illegal, self-build types of housing associated most closely with the cities of the developing world.

Millions of people find their own housing solutions in the absence of formal and/or affordable forms of provision. It has been estimated that over a third of the urban labor force in developing countries are housed in the informal sector. When we think therefore of housing in cities, we need to think beyond the regulated and formal forms of provision of the Western city and consider the critical role of informal self-provision. Moreover, without the initiative and resilience shown by poor migrants to cities, many aspects of the so-called formal sector would be unsustainable. For example, it is often the poor and marginal and those in the informal housing and labor market that do the critical recycling work of cities – foraging among the urban detritus as a means of survival.

Thus, when we consider housing and cities, the conventional starting point is usually tenure – who owns and who rents? However, such categories contain such variation of circumstances and conditions within and between societies that they lack explanatory power. It is more appropriate to think about housing and cities in terms of formal/informal or commodified/decommodified and to see such categories as continuous rather than polar opposites. These concepts can then be applied to different stages of the housing process in cities – through its production, distribution, and exchange. These different stages involve a range of institutions and actors, including government, quasi-government, private and voluntary sectors, households, and families. Ball and Harloe (1992) referred to such configurations as "structures of housing provision" – "combinations of social agents involved in housing provision that relate to each other in empirically observable ways" (p. 3).

Connections

David Donnison, an early and influential housing specialist, neatly captured the centrality of housing to so many aspects of the everyday life of urban populations. In his introduction to *The Government of Housing* (1967) he observed that "House and home stand at the center of people's lives, providing a shelter for sleep and for half their waking activities, a shield against the elements and the world-which yet admits both in a selective and controlled fashion and a storage place and showcase for most of their possessions. To study housing is to explore a cross section of a whole society and its affairs" (p. 9). The dwelling is thus a key nodal point, symbolically and practically. Where you live and what you live in are important signifiers of your social status. You are judged consciously and unconsciously, purposively and inadvertently, by your address. High-status addresses cost money. The rich live in big houses. Wander around any city and see social status imprinted on the residential structure. It is one of the most visible expressions of social division. Moreover, as information technology has become increasingly sophisticated in its capture and deployment of household data, addresses have become much more than mere signifiers of social status. Zip codes or postcodes now convey risk and danger, safety and desirability, and investment and disinvestment. They show whether certain neighborhoods contain 'people like us' or 'the others,' whether they are places to avoid or places to embrace. And residential reputation may hinder or enhance employment opportunities.

Various urban analysts have observed that the contemporary city is a city of walls and enclaves (e.g., Marcuse, 1995). These walls may be virtual and metaphorical. People living in the same neighborhood or even the same street may occupy different worlds. They may share the same space but have little in common. In recent decades, these barriers and distinctions have taken on an increasingly physical reality. There are real walls that distinguish and separate different kinds of housing. Gated residential communities have

multiplied in rich and poor countries. They may be demographic communities (for the retired, for the young family) or defensive communities protected by electronic gates, security cameras, and private guards. This is not a new phenomenon. Historically, many cities had private and usually prestigious residential areas, often with imposing gates and bylaws specifying what was and was not allowed within the perimeter. But gating is now pervasive and is without doubt a defining feature of housing patterns in the contemporary city (Webster et al., 2002). It is also a notable feature of social status and social security, particularly in the affluent residential enclaves of the expanding cities of Asia, Latin America, and Africa.

Demography is also a major factor to consider in relation to urban housing. The stage in the family life course affects not only the choice of dwelling type, dwelling size, and the affordable price but also the critical connections. The classic life cycle model progresses from leaving the parental home to getting married, having children, the departure of children, retirement, and eventual mortality. The first stage after leaving the parental home represents the need for a new dwelling unit and the last stage releases an empty unit for reuse. As individuals progress across the life cycle, households typically grow in size (full nester household with children) and then shrink when children leave (empty nester). Of course, this has always been an approximate and idealized version of real life. Housing histories are messier and some argue that the modernized version is less predictable than was the case for previous generations. People stay single for longer, relationships break down, and adult children return home for a while. There is more divorce and more remarriage, women have children later in life, there is greater longevity, and so on.

When a family breaks up, two housing units are required rather than one. One or both partners may lack the resources to buy or rent in the market and need social housing of some kind. Changing social norms and behavior generate new types of demands for housing. City housing built for one generation has to accommodate the very different demands of the next. This results in formal and informal adaptations. Large houses become flats. Rooms are subdivided. Once fashionable areas become the reception areas for poorer migrants to the city. In many European cities, much of the housing is well over a century old. For example, a fifth of the dwellings in the United Kingdom, France, or Italy were built before 1918. During building booms and even in cities where there are few institutional or other barriers to redevelopment, it would typically take half a century to replace the building stock. Even if that were achieved, building forms could never keep pace with changing social norms.

However we conceive of the family life course in relation to urban housing, different groups will have different needs and preoccupations. This will affect locational and other residential preferences. Younger single people and childless couples spend less time at home and will probably want to be more immersed in the urban milieu than families with young children. Young families are likely to seek bigger gardens on the suburban or rural fringes; as children reach their teens, the pressure on parents will be to return to a more urban environment. One of the most critical issues for families with children concerns schooling. Here the connection between residential location and school choice can be direct. Geographical catchment areas can determine the majority of allocated places to schools. Schools acquire good and bad reputations. In cities where market processes dominate, strong reputations translate into higher house prices. Real estate agents further reinforce this process and neighborhoods with the most desirable schools become expensive places to buy. Even if there is a substantial social housing sector, school catchment areas are likely to affect the pattern of allocations. Those able to wait, the least desperate, can hang on until a dwelling in the right location becomes available. This interface between housing and education has become a major battleground in many cities as families vie for position in increasingly competitive education markets. In knowledge-based societies where the returns for high-quality education seem to be increasing (as do the penalties for low levels of education) and where in many societies one child is the norm, this competition has become more intense. For aspirant parents with resources, the choice may be between paying for expensive private education or moving to a more expensive neighborhood. For poorer households, the problem is that the best schools may become colonized by those able to pay higher property prices. This process can thus reinforce existing patterns of social and spatial disadvantage.

A full appreciation of the dynamics of neighborhood change is beyond the scope of this article. The key point to emphasize is that housing cannot be separated from its neighborhood context or from broader issues of residential design, planning, and architecture. Here, there are links to current debates about the future shape of cities and to environmental concerns about urban sprawl and the need for higher density living. Residential planning can also reflect and reinforce class and ethnic segregation.

More fundamental challenges to the design of dwellings and neighborhoods have often come from feminist writers arguing that cities are essentially designed by men for men. The suburban ideal in particular has been critiqued as predicated on the male breadwinner model and dependent on the unpaid domestic work of women. "Despite very real differences in the structure and support for suburban living, women everywhere were taken for granted as its main custodians and beneficiaries and separated, appropriately enough it was assumed, from male-dominated economic and political activity in the city" (Strong-Boag et al., 1999; 171). The world may have moved on somewhat, at least in some societies, but even today few would challenge Dolores Hayden's assertion in 1984 that "no country has yet created an urban fabric and an urban culture to support men and women on equal terms as citizens and workers" (p. 232).

Understanding the Distribution of Housing in Cities

Social scientists have approached the question of who gets what housing, and why, from different theoretical perspectives. Housing is both physical form and social construction. What is built, when, by whom, and for whom are ultimately social and political processes involving the exercise of power, which is unequally distributed. Although most people can exercise some choice about where they live, such choice is exercised within highly variable constraints.

Addressing these issues conceptually is most easily conceived of in terms of approaches that focus on individual behavior, those emphasizing institutional actors and those adopting a more structural perspective. These approaches overlap and are not necessarily mutually exclusive.

By far the most significant influence on studies of housing and cities has been the work of the Chicago School of the 1920s and 1930s. Chicago, at that time, was one of the most rapidly expanding cities in the world and Park, Burgess, McKenzie, and associates used the city as a laboratory to explore the dramatic social and economic transformations that were taking place. The key features of the ecological approach, and its shortcomings, are well-trodden ground and those seeking detailed elaboration should consult the suggestions for further reading. But essentially, the city was conceived of as a built environment in which individuals and groups competed for survival. As in the natural world, this competition involved processes of invasion, displacement, and succession. The evolving structure of cities could be understood therefore as a biological model, mediated but not primarily determined by, cultural factors. The research agenda and conceptual schema were understandably a reflection of the forces shaping Chicago at that time. Here was a city expanding rapidly through in migration with the central business district (CBD) encroaching on the residential areas in the city center. This process pushed the population outwards into the suburbs leaving behind 'zones of transition' – areas of residential decay and falling property values awaiting commercial and industrial redevelopment. It was these cheap areas of dilapidated housing in the inner city that provided accommodation for poorer migrants to the city.

While the work of the Chicago School has been critiqued for being strong on description of the changing residential patterns but weaker on explanation, it represents the first systematic, social science research on urban residential processes. It was the foundation stone for sociological analysis of contemporary urban society – indeed, for sociology as a discipline. Moreover, although there have since been considerable theoretical advances in understanding the processes that sort and sift residential areas, its basic approach to analyzing urban social structures has influenced generations of urban geographers and sociologists. It is also often forgotten that the Chicago research encompassed ethnographic work on processes of adaptation and assimilation into this new way of urban life, and on social groups on the margins of urban society. It was thus the forerunner to various studies of residential communities, both mainstream and marginal, which followed.

In research on housing and cities, mainstream economics kept itself somewhat distant from the other social science disciplines. This can be partly attributed to the greater importance attached to disciplinary purity among economists compared to say sociology or geography – the disciplines that have tended to dominate urban studies as a subdiscipline. However, neoclassical economics has made an important and distinctive contribution in this field. Individual competition has always been the hallmark of neoclassical economic models. In the context of household residential preferences and demands, one thus has utility maximizing subjects trading housing costs against transport costs. Stated simply, housing costs decrease, and transport costs increase, as we move further from the CBD. These models became increasingly sophisticated and included additional trade-offs and locational influences including leisure, densities, neighborhood characteristics, environmental quality, and ethnicity. Economists were particularly influential in the development of ideas of 'filtering' as a way of understanding the operation of housing policies, housing markets, and housing submarkets. One application of filtering theory was in relation to vacancy chains – the chain of moves set in motion when a dwelling is vacated. For example, building a new dwelling at the top end of a housing market may generate a long series of moves ultimately releasing a cheap dwelling at the bottom end of the market. Unsurprisingly, such models have been controversial in housing policy debates.

Neoclassical economists' concerns with the aggregate effects of individual residential locational decisions can be linked to another strand of research that emerged in the 1960s. This drew particularly on psychology and the sociology of the family and the household to explore perception and decision making. Residential mobility within cities came under closer scrutiny. Why did households move? How did individuals, households, and groups respond to changes in the external environment? This generated increasingly complex models of the process of residential location and of the interaction between internal (e.g., family size) and external (e.g., changing neighborhood characteristics) drivers of mobility decisions. One of the best known studies of residential mobility actually preceded much of this literature and had a strong social policy orientation. Rossi's *Why Families Move?*, originally published in 1955, was concerned with a perceived dysfunctional pathology of American society. Postwar United States was characterized by excessive mobility that potentially threatened social cohesion and ideas of community. The research reassuringly found that residential mobility was, however, primarily driven by family growth factors rather than by some innate, pervasive, and unrelenting desire to move among contemporary Americans.

In the 1960s, the field remained dominated by studies that gave preeminence to the role and experience of the individual in housing systems and that tended to neglect or downplay the wider socioeconomic/political context in which housing choices were made. The latter part of this decade, however, saw significant changes in approaches to housing and cities. As with the Chicago research, issues of immigration and segregation were also the main driving forces for the classic study undertaken in Birmingham, UK, by John Rex and Robert Moore. Published in 1967, *Race, Community and Conflict* set out to develop a sociology of race relations through an exploration of urban segregation in area of Birmingham, which had become a key reception point for new black immigrants to the city. In doing so, they explicitly drew on the ecological ideas of the Chicago School combined with a Weberian conception of classes as status groups – in this case, housing classes. Rex and Moore described a process of city growth involving migration from the central areas to the suburbs. Two forces were at work here – the expansion of the CBD and a common pursuit of a middle class, suburban lifestyle. The professional class moved to the detached houses in the inner suburbs; the lower middle class to semidetached suburbia, and a skilled working class to public housing suburbia. This process of migration involved the progressive abandonment of the nineteenth century housing. The less prestigious Victorian housing was bought

up by landlords and sublet as lodging houses to new migrants creating a twilight area of short life properties, physical decay, overcrowding, and a poorer immigrant class. In this competition for housing resources, there was a status hierarchy. At the top were the outright owners of whole houses followed by owners of mortgaged dwellings. At the bottom were tenants of rooms in lodging houses. The achievement of suburban home ownership was a 'scarce resource' in which there were winners and losers. The pattern of housing distribution could thus be the basis for potential conflict between different groups.

With regard to race relations, the concern was the perceived cleavages between a white middle class in suburban home ownership, a white working class in public rental housing, and the black immigrant population. Because of income barriers and eligibility rules for public rental housing, new immigrants were effectively excluded from the two main forms of urban housing provision. Previously, the literature on urban housing had focused on residential choice and mobility, on atomized individuals competing in the market. The significance of Rex and Moore's study was that it emphasized conflict and constraint, was explicitly rooted in sociological theory, and linked spatial form and social process.

Inevitably, their analysis was subject to considerable critique. For example, the assumption that all groups were competing for the same scarce resource, suburban home ownership, was questioned by some as was the conception of housing classes as the basis for social conflict. Could the struggle over housing really be analyzed as a class struggle over the distribution of life chances in the city – a domain of class conflict distinct from the world of work? Was it not the case that the distribution of housing reflected, rather than caused, social stratification? Both Weberians and Marxists had more to say on this issue.

From a Weberian perspective, the British sociologist Ray Pahl, focused on the role of urban gatekeepers as a key element of the wider context in which individual choices were exercised (Pahl, 1967). Housing choices were not made in a vacuum but within conditions created by the various institutions and actors that controlled the production and allocation of housing and other urban resources. Pahl was still concerned with spatial patterns and saw residential location as an inherent source of inequality and conflict since every space is unique. The issue was how a particular spatial distribution occurred. For Pahl, the key to understanding the spatial distribution of housing lay in the way power was exercised, and decisions made, by urban managers. This *urban managerialist* perspective spawned a whole series of research studies on banks, building societies, real estate agents, local governments, and landowners. The decisions of mortgage lenders or those who allocated public housing were argued to be crucial to explaining sociospatial distributional patterns in the residential city. Particular controversial issues related to the redlining (or whitepainting) of neighborhoods by lending institutions, thereby starving them of mainstream lending and thus, it was argued, reinforcing a downward spiral of decay.

Again, race and ethnicity were central to these discussions. In these institutional studies, the social and spatial consequences of lending or allocation decisions were often attributed to the attitudes and prejudices of particular institutional actors. However, the unjust distribution of housing resources was more likely to be the product of more complicated intra-institutional dynamics. Sometimes decisions taken at different levels of an organization, and for unconnected reasons, combined to produce discriminatory outcomes.

It should be noted that urban managerialism, as a distinct approach, had more academic and policy traction in Europe, and notably the United Kingdom, than in North America. Some commentators have suggested that this is because of the more interventionist character of housing provision in Europe compared to the more market-driven housing markets in the United States and Canada. Housing resource allocators, especially with regard to social housing, were more visible in places like Britain.

A more fundamental break with previous conceptions of the production, distribution, and consumption of urban housing came from within the Marxist paradigm. A fundamental conceptual problem with the managerialist/institutionalist perspective (which Pahl himself recognized) was that 'the managers' of housing systems could not be blamed for creating scarcity – they merely had to make decisions on how to allocate scarce housing resources. The Marxist literature seemed to offer an answer. Urban housing was produced, consumed, and distributed according to the logic of contemporary capitalism. Cities, and thus housing provision, were shaped according to the needs of capital accumulation. Housing tenures (ownership, private renting, and public renting) were economic devices to speed up the rotation of capital in cities. Cities were conceived of as essentially 'transaction maximizing systems' with residential real estate as a key sector for capital investment and realization. Conflicts over housing were class struggles in a system in which inequalities were endemic. The early Marxist contributions powerfully shifted the focus of analysis from individuals and institutions to social and economic structures. In doing so, they were often criticized for being overly economistic. People often seemed to have disappeared from view under the weighty analyses of the economic forces shaping housing provision in the capitalist city. The problems of methodological individualism in work on urban housing had been theoretically addressed but in doing so, much of the richness and detail – the understanding of everyday life in residential areas – seemed to have been lost. Housing might be capitalist but there were different forms of capitalism and different capitalist cities with distinct cultural, economic, political, and policy histories. It was thus recognized that there was a need to work back from the high level of abstraction offered by Marxist writers on housing toward studies that allowed more room for temporal and spatial contexts.

Concluding Observations

Reflecting on housing and cities at the end of the first decade of the twenty-first century, four aspects stand out. First, under postmodern influences, ideas of diversity and difference in housing have seen renewed interest in detailed ethnographies and community-type studies of neighborhoods. In other words, there has been recognition of the need for more empirical research on the experience of contemporary city life – a return to some of the original concerns of the Chicago School. Second, and conversely, urban housing is more than ever

conceived of economically rather than socially. Residential real estate has been deeply implicated in various financial crises and economists have become more interested in the impact of housing markets on wider economies (*Critical Review*, 2009). Third, as indicated earlier, in China the pace of urbanization and of new residential construction is unprecedented. Social science research on housing and cities has been predominantly focused on Western developed societies and cities. New conceptual tools are now required to capture the changed social and cultural geography of housing and urban development. Fourth, the last three decades or so has been the era of neoliberal policy dominance. Urban housing has generally been in the forefront of market-driven policy developments whether in the United Kingdom, China, or Russia. The longer term consequences of these major political and policy shifts in relation to housing provision are only now emerging and are likely to become a focal point of future research on housing and cities.

See also: Gated Communities; Housing Economics; Just City; Marxist Geography; Place in Geography; Social Geography; Spatial Equity; Urban Geography; Urbanization; Wealth Distribution.

Bibliography

Araghi, F., 2009. The invisible hand and the visible foot: peasants, dispossession and globalization. In: Akram-Lodhi, A. Haroon, Kay, Cristóbal (Eds.), Peasants and Globalization: Political Economy, Rural Transformation and the Agrarian Question. Routledge, New York.

Ball, M., Harloe, M., 1992. Rhetorical barriers to understanding housing provision: what the 'provision thesis' is and is not. Housing Studies 7 (1), 3–15.

Burgess, E., Bogue, D.J. (Eds.), 1964. Contributions to Urban Sociology. University of Chicago Press, Chicago.

Castells, M., 1983. The City and the Grassroots. Edward Arnold, London.

Clapham, D., Clark, W., Gibb., K., 2012. The Sage Handbook of Housing Studies. Sage, London.

Critical Review, 2009. Special Issue: Causes of the Financial Crisis, 21 (2–3).

Donnison, D., 1967. The Government of Housing. Penguin, Harmondsworth.

Galster, G., 1996. William Grigsby and the analysis of housing submarkets and filtering. Urban Studies 33 (10), 1797–1806.

Harloe, M., 1995. The People's Home?: Social Rented Housing in Europe & America. Blackwell, Oxford.

Harvey, D., 2011. The Enigma of Capital. Profile Books, London (see particularly Chapters 6 and 7).

Hayden, D., 1986. Redesigning the American Dream. Norton and Company, New York.

Henderson, J., Karn, V., 1984. Race, class and the allocation of public housing in Britain. Urban Studies 21, 115–128.

Keating, P., 1976. Into Unknown England, 1866–1913. Selections from the Social Explorers. Fontana Collins, London.

Keating, P. (Ed.), 1976. Into Unknown England 1866–1913: Selections from the Social Explorer. Manchester University Press, Manchester.

Kurz, K., Blossfeld, P., 2004. Home Ownership and Social Inequality in Comparative Perspective. Stanford University Press, Stanford.

Lamarche, F., 1976. Property development and the economic foundations of the urban question. In: Pickvance, C. (Ed.), Urban Sociology: Critical Essays. Tavistock, London.

Marcuse, P., 1993. What's new about divided cities? International Journal of Urban and Regional Research 17 (3), 355.

Marcuse, P., 1995. Not chaos but walls: postmodernism and the partitioned city. In: Watson, S., Gibson, K. (Eds.), Postmodern Cities and Spaces. Blackwell, Oxford.

Muth, R.F., 1969. Cities and Housing. University of Chicago Press, Chicago.

Pahl, R., 1967. Whose City? Penguin, Harmondsworth.

Rex, J., Moore, R., 1967. Race, Community and Conflict. Oxford University Press, Oxford.

Rossi, P., 1955. Why Families Move: A Study in the Social Psychology of Urban Residential Mobility. Free Press, Glencoe.

Smith, N., 1996. The New Urban Frontier: Gentrification and the Revanchist City. Routledge, New York.

Strong-Boag, V., Dyck, I., England, K., Johnson, L., 1999. What women's spaces? Women in Australian, British, Canadian and US suburbs. In: Harris, R., Larkham, P.J. (Eds.), Changing Suburbs-Foundation, Form and Function. E&F Spon, London.

Tipple, G., Speak, S., 2009. The Hidden Millions: Homelessness in Developing Countries. Routledge, London.

United Nations Human Settlements Programme, 2003. The Challenge of Slums. Earthscan Publications, London.

Webster, C., Glasze, G., Frantz, K., 2002. The global spread of gated communities. Environment and Planning B: Planning and Design 29 (3), 315–320.

Housing Economics

Richard Arnott, University of California Riverside, Riverside, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the economics of housing. The organization is conventional. The first three sections treat the microeconomics of housing. The first section reviews the demand for housing, the second the supply of housing, and the third section how demand, supply, and the allocation mechanism (principally markets) interact to determine the allocation of housing. The fourth section treats two subjects that build on housing microeconomics, real estate economics (housing from an investment perspective) and real estate macroeconomics (principally housing market cycles and fluctuations). The final section reviews housing policy.

Housing Economics

Housing has a set of characteristics that together make it different from all other commodities, which are as follows:

1. Necessity: Housing satisfies a basic human need, shelter.
2. Importance: Most households spend more on housing than on any other class of commodity.
3. Spatial fixity: Most housing units are rooted to a particular location. Living in a particular housing unit, a household, therefore, consumes not only the characteristics of the housing unit per se, but also the characteristics of its location – its neighbors and neighborhood, its accessibility to other locations, its taxes and public services, and its locational amenities and disamenities.
4. Durability: Housing is the most durable of major commodities. If built soundly and maintained well, a housing unit will last virtually indefinitely.
5. Heterogeneity and complexity: To describe a housing unit reasonably accurately requires enumerating very many characteristics (complexity). Also, housing units exhibit considerable heterogeneity in characteristics, and households exhibit considerable heterogeneity in tastes.
6. Housing as an asset: Because housing is durable and the expenditure share on housing high, the housing stock constitutes a substantial proportion of a nation's wealth and of most households' portfolios. The demand for housing, therefore, combines consumption and asset motives.
7. Housing and the macroeconomy: Since housing is durable, modest fluctuations in housing demand translate into large fluctuations in housing investment. Due to the importance and volatility of housing investment and the magnitude of the wealth effects generated by movements in housing values, the performance of the housing market can have a major impact on the macroeconomy.

The interactions of these characteristics cause the economics of housing to be significantly different from the economics of other commodities. Section Housing Supply will examine the supply of housing and Section Housing Demand the demand. Section Housing Allocation and Markets will discuss how supply and demand interact to determine equilibrium in the housing market. Section Real Estate Economics will look at the housing market from the perspective of real estate economics. Section Housing Policy will consider the role of government with respect to housing. The focus throughout will be on housing economic theory, in contrast to empirical work.

Housing Supply

Before discussing how housing is supplied, a quantifiable definition of 'housing' is needed. The appropriate definition depends on context. In some contexts, highly aggregated definitions that ignore housing heterogeneity – the number of units or total floor area – are most useful. In other contexts, each housing unit is usefully viewed as a bundle of many characteristics. In yet other contexts, intermediate levels of aggregation are most helpful. The definition employed strongly influences the way the market is conceptualized and housing supply and demand described.

Most housing economics is conducted at a high level of aggregation. A nation's monetary authority is interested in the value of the nation's housing stock and the level of housing investment, not in where housing units are located within cities or how many units have two bedrooms. Similarly, the government's housing ministry is principally concerned with housing affordability, crowding, sanitary facilities, and structural soundness, and not in how many units have fireplaces.

At the highest level of aggregation, a housing unit is described simply by the flow of housing services it provides, which can be measured by its rent. Holding its characteristics fixed, an increase in the unit's rent corresponds to a rise in the price of housing services. A housing supply curve may then be defined relating the units of housing services supplied to the price of housing services. Because housing is durable, the elasticity of the housing services supply curve depends on the time horizon. Within, say, a month, the housing supply curve is almost vertical; little can be done beyond maintenance to increase the quantity of housing services supplied. As the time horizon increases, the housing services supply curve becomes increasingly elastic; within 6 months, a rise in the price of housing services will induce some rehabilitation; within a year, some new housing will be constructed; and so on.

In many situations, the analyst is concerned with the interaction between particular housing submarkets. For example, the government might be interested in the effects of

a subsidy on owner-occupied housing. To analyze this policy, the housing market would be divided into owner-occupied and rental submarkets. The two submarkets would interact on both the supply (the subsidy would encourage the conversion of housing from rental to owner occupancy) and demand (the subsidy-induced fall in the consumer price of owner-occupied housing services would cause a decrease in rental housing demand) sides. Other divisions of the housing market into submarkets are appropriate in other contexts. If the government is concerned with the living standards of the poor, the housing market is usefully divided into quality submarkets; if it is concerned with downtown revitalization, a division into downtown and suburban submarkets is suitable; and so on.

The discussion thus far has treated the supply for housing services, which is a flow. Lurking in the background is the supply of housing stock. The two are related through the user cost of capital (which is the ratio of the annualized cost of employing a unit of capital to its supply price, and equals the interest rate plus the rate of depreciation plus the maintenance cost rate less the rate of capital gains) and the opportunity cost of land, with appropriate adjustment for the various taxes applied to housing. Housing economists are sometimes careless in failing to distinguish two 'prices' of housing – rent, the price per unit of housing services per unit time, and value, the asset price of a unit of housing stock.

In the short run, rent adjusts to clear the market for housing services. This translates, via the user cost of capital and the opportunity cost of land, into the equilibrium asset price or market value of housing. If the market value of housing exceeds the market value of land plus construction cost, new construction is triggered, which increases the supply of housing stock and of housing services, causing rent to fall. This process continues until a long-run equilibrium is attained where the market-clearing rent equals construction cost times the user cost of capital plus the opportunity cost of land.

The above conceptualization of the housing market is especially well suited for aggregate analysis. The remainder of the section reviews several alternative conceptualizations, each of which has its particular domain of application.

The focus of the Mills–Muth model (see Mills (1972) and Muth (1969), and Brueckner (1986) for an integrative treatment) is on urban spatial structure – how housing density (the ratio of floor to land area), housing rent, and land rent vary with location. Ignore the spatial fixity and durability of structures, so that static analysis can be applied. A landowner adds capital to his land until marginal factor cost equals marginal revenue product. Under competition, land rent equals the corresponding producer's surplus. This analysis relates land rent and housing density to housing rent at a particular location. At the level of the market, housing rents adjust so that in equilibrium at each location the quantity of housing supplied equals the quantity demanded, with more attractive and accessible locations commanding higher rents. Then from the relationship between housing rent and housing density, the equilibrium urban spatial structure is obtained.

The Mills–Muth model ignores the durability of structures. A related class of models, reviewed in Brueckner (2000), examines the spatial structure of a growing city, accounting for durability. A developer who owns vacant land then faces the decision of when to build as well as at what density. The relationship between predevelopment land value and post-development housing value and density can be straightforwardly derived. Normally vacant land is developed from the center outwards. Subsequently, there are successive waves of redevelopment. Each wave entails construction at higher density with the time between waves depending on the city's growth rate. This process is evident in some cities but in others has been curtailed by redevelopment restrictions.

The Mills–Muth model and dynamic durable housing models treat location, housing density, and unit size. Housing units also differ in quality. If two housing units of the same size at the same location command different rents, the higher rent unit is presumably of higher quality. Since the poor typically live in low-quality housing, quality differentiation is particularly important to consider when evaluating policies aimed at improving low-income housing. Models of housing quality differentiation examine the filtering process (Grigsby, 1963). An urban location goes through a life cycle. Housing is of high quality when constructed but depreciates with age, and as it does so is occupied by successively poorer households. Eventually, rehabilitation/renovation or redevelopment of the site at higher density becomes profitable, and the cycle repeats. Since the typical neighborhood was initially developed in a short time span, the filtering process gives rise to neighborhood cycles too. Filtering models suggest that it is efficient for the poor to live in hand-me-down housing, which provides an argument against public housing – housing constructed and operated by the government specifically for occupancy by poor households.

More recent filtering models (Sweeney, 1974; Arnott et al., 1983; Rothenberg et al., 1991) view the housing market as a hierarchy or continuum of quality-differentiated submarkets. At any instant, the housing market is described by the housing stock at the various qualities. Rent as a function of quality adjusts to clear the market. Construction occurs at those qualities where housing value equals or exceeds construction costs plus the value of land. Maintenance is carried to the point where the last dollar spent increases housing value by one dollar, and a housing unit's value is the present discounted value of rent less maintenance expenditures. These models can be enriched to treat housing differentiation by unit size, location, and density (Anas and Arnott, 1991).

Filtering models provide a strong conceptual basis for policy analysis and are intuitively appealing, accommodating all the standard housing conversion processes – construction, maintenance, rehabilitation, upward and downward conversion (combining and subdividing housing units), abandonment, and demolition and reconstruction. Despite their popularity among academic housing economists, filtering models have not yet been widely adopted by the housing policy community.

Working with detailed descriptions of housing units is normally distracting, but in measurement is essential. The minister of housing may be uninterested in whether the poor have fireplaces (as long as they receive heat), but is interested in whether the average quality of low-income housing has increased. This requires obtaining a quality index, an aggregate of many housing characteristics – including fireplaces. Investors, housing policy analysts, and macroeconomists are all concerned with the time trend of housing prices, which

requires developing price indices. And for tax assessment purposes, housing values need to be estimated accurately, which necessitates detailed description. Hedonic analysis utilizes such detailed description. The market values (or rents) of housing units are regressed against a long list of their characteristics (including public services, neighborhood and neighbor characteristics, accessibility measures, and environmental variables). The regression coefficients are interpreted as implicit market prices for characteristics (Rosen, 1974). Hedonic analysis has proved very useful not only in housing valuation and housing price and quality index construction, but also in inferring the market's valuation of various neighborhood characteristics, such as parks, trees, and improved air quality. But the implicit conceptualization it employs – separate markets for each of the hundreds of housing characteristics – is unwieldy for most housing policy analysis.

Industry studies focus on market organization – industry concentration and the form of strategic interaction between firms. The conventional wisdom that the housing market is competitive or monopolistically competitive, while inadequately researched, appears sound.

Housing Demand

The previous section indicated that how housing supply is described depends on how housing is conceptualized. The same is true of housing demand. Most housing demand studies employ the housing services conceptualization. Accordingly, there is a very large empirical literature estimating the income and price elasticities of demand for housing services. Excellent reviews are provided in Goodman (1996) and Olsen (1986). Muth, whose classic *Cities and Housing* (1969) provides the foundation of modern housing economic analysis, has argued that since the housing expenditure share has shown remarkable stability over time and location – about 25% of after-tax income – the long-run price and income elasticities of housing must be close to unity. Most empirical studies, however, estimate significantly lower price and income elasticities. How can the results of these empirical studies be reconciled with Muth's observation?

The literature has concentrated on four classes of explanations as follows:

1. With high moving costs (including psychic costs), a household does not instantaneously change its housing when its income or the price of housing changes; rather, housing decisions are long term. This explains why experiments investigating how households alter their housing consumption under temporary rent subsidies have found a generally price-inelastic response. Relatedly, the demand for housing should depend more on permanent than on transitory income.
2. That many moves are triggered by demographic change (e.g., marriage, births, retirement, and divorce) makes it more difficult to identify the economic determinants of housing choice.
3. Tenure choice with borrowing constraints. The typical household starts its life cycle by renting, often underconsuming housing to save up for the down payment on an owner-occupied housing unit. Having saved up enough, it moves into a starter home or condominium, aiming to trade up when it can afford to. The timing of these moves depends on the realization of the household's uncertain income prospects and demographic changes, and on the evolution of housing prices, interest rates, and down payment requirements.
4. The asset or portfolio demand for housing. When deciding how expensive a home to buy, a household considers how much housing it would like to include in its wealth portfolio. A household's demand for owner-occupied housing achieves a balance between its consumption and portfolio demands. Households adjust the discrepancy between these demands through tenure choice, customization, and furnishing (Henderson and Ioannides, 1983). A household whose asset demand exceeds its consumption demand may purchase a large home but spend little on decoration, or may purchase additional housing units and rent them out. A household whose asset demand falls short of its consumption demand may choose either to own a cheaper house and decorate it nicely or to rent.

Estimating price and income elasticities accounting for all these complications, as well as the tax treatment of housing, is indeed a daunting task. Despite these difficulties, considerable progress has been made in the empirical analysis of housing demand, reflecting advances in data collection, computing power, applied econometrics, and housing economic theory, and in the use of controlled experimentation. Regrettably, however, little work has been done on estimating filtering models (Rothenberg et al., 1991 is a notable exception), which has impeded their use in policy analysis.

Housing Allocation and Markets

Throughout history, most consumer goods have been allocated through markets. The twentieth century witnessed a sharp rise in the size and scope of government. The communist countries adopted central planning, which entailed government construction, operation, and direct allocation of housing. Elsewhere too, government intervention in the housing sector increased. For a quarter century the pendulum has been swinging in the opposite direction, but by historical standards government intervention in the housing sector remains high. Even in the United States, with the most free market of Western economies, much low-income housing continues to be allocated outside the market, and the allocation of market housing to be deliberately influenced by a myriad of government policies.

Why have governments everywhere doubted the ability of markets to allocate housing? Textbooks in microeconomic theory give a standard list of (classic) market failures or inefficiencies – public goods, natural monopoly, and technological externalities. None is important in the context of housing. The primary motive for extensive government intervention vis-à-vis housing has rather been to improve equity. Some believe that the income distribution generated by the free market is just but most do not. How then is equity to be achieved within a generally market-oriented economy? There are two broad approaches. One is income redistribution, general

egalitarianism; the other is effective redistribution through the taxation/subsidization or nonmarket allocation of specific commodities, specific egalitarianism. Most economists favor general egalitarianism because it respects consumer sovereignty. But most governments evidently favor specific egalitarianism. They – and the person on the street – feel that the consumption of some goods (basic housing, clothing, food, education, and medical care) is more meritorious than others (alcohol, cigarettes, drugs, and luxuries). Many governments have favored the nonmarket allocation of housing for the poor over subsidization, because it permits more refined targeting and may be administratively simpler. There has been remarkably little modern economic analysis of nonmarket housing allocation (Anas and Cho, 1988, is a notable exception). The rest of the section will, therefore, consider how the market allocates housing.

The central contribution is Alonso (1964). He asked: How does the market determine the pattern of household location within a city? Suppose that there are two household groups, rich and poor, that each household derives utility from only other goods and housing floor area, and that locations differ only in accessibility. In equilibrium, all rich households achieve the same utility level (otherwise migration would occur from low-utility to high-utility locations). Associated with this utility level is the equilibrium housing bid-rent curve for the rich, which indicates the housing rent (per unit floor area) payable by the rich at different locations consistent with this utility level. The equilibrium bid-rent curve for the poor is derived in an analogous manner.

Two basic principles apply: (1) A housing unit goes to whomever bids the most for it and (2) rent equals the bid rent of whomever bids the most for it. The top panel of **Figure 1** plots the equilibrium bid-rent curves for the two groups as a function of accessibility, a. Applying the two principles implies that housing between accessibility levels a′ and a″ goes to the rich and that between a″ and a‴ to the poor, and that housing rent as a function of accessibility is given by the 'upper envelope' of the equilibrium bid-rent curves. The lower panel of the figure shows the stock of housing available at different

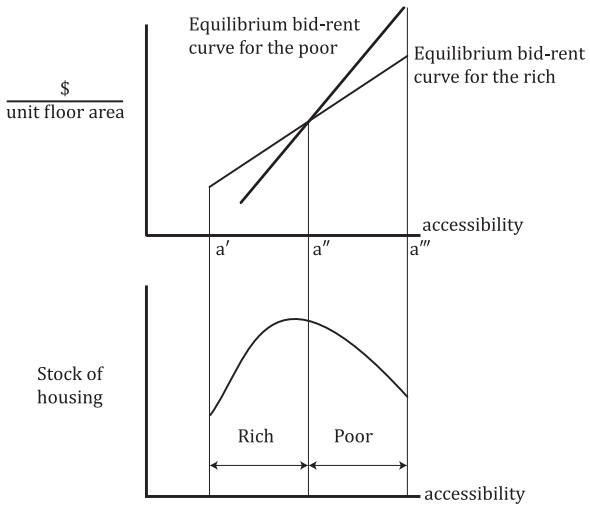

Figure 1 Equilibrium bid-rent curves.

accessibility levels. The equilibrium utility levels are determined by the conditions that, for each group, the quantity of housing demanded equals the quantity of housing supplied. The slope of a group's housing bid-rent curve is the premium that group is willing to pay for improved accessibility. As drawn, the poor outbid the rich at more accessible locations because they are willing to pay a higher premium per unit floor area for improved accessibility. This result has been used in explaining the post-World War II suburbanization of wealthier US households.

The accessibility-rent tradeoff is central in determining the pattern of household location within a metropolitan area. The above analysis can be extended to treat other factors, such as the relative value different groups attach to amenities, disamenities, public services, environmental quality, and neighborhood quality and composition. Each group has an equilibrium bid-rent surface over all locations within the metropolitan area, and lives where it outbids all other groups. Groups may differ not only in income but also in demographic status, race, job location, etc. Extending the analysis to account for taste heterogeneity within groups yields the probabilities that a particular location is occupied by a household in a particular group. The same competitive allocation mechanism operates with respect to other housing attributes.

The asset value of a housing unit is the discounted present value of expected future rents minus expected costs, adjusted for uncertainty. The volume of construction depends on profitability, as measured by asset price minus supply price; similarly, the volume of rehabilitation depends on the gain in asset price through rehabilitation minus the cost; and maintenance expenditure depends on how steeply housing value rises with quality.

The above discussion indicates how a perfectly competitive housing market works. In actual housing markets, housing units and households exhibit multidimensional heterogeneity. To find a unit that suits its tastes, a household undertakes costly search over vacant units. The friction of search confers market power on landlords, who are consequently price-setters rather than price-takers. With entry and exit close to free, actual housing markets are, therefore, more accurately described as monopolistically competitive, with excess capacity taking the form of vacant units. Furthermore, housing markets are matching markets, with market tightness (the ratio of searching households to vacant units) playing an important role in the market adjustment process. Down payment requirements on mortgages and credit restrictions on developers also significantly affect the market's operation. These are typically treated as exogenous but should be derived as a feature of equilibrium.

Whether the gain in accuracy from treating the housing market as monopolistically rather than perfectly competitive justifies the increased analytical complexity, is an empirical issue and depends on context. Normally little is sacrificed in using perfectly competitive models to analyze the positive effects of specific housing policies. But imperfectly competitive models should be used to provide microfoundations for housing macroeconomics since nonprice adjustment is important, and to evaluate welfare changes from housing policies.

In the frictionless world of perfect competition, and without taxes, a household would be indifferent between renting and

owning. That tenure choice is in fact important indicates that it should be considered from the perspective of imperfect competition. Though often treated in the context of housing demand, tenure choice is an aspect of market equilibrium since it concerns the contractual form through which housing services are exchanged. Renting and single-family owner occupancy are the most common contractual forms; others include condominium ownership (under which common areas are jointly owned) and cooperative ownership (under which a resident is essentially a shareholder in the building). With ownership comes extended property rights, including claim to the property's capital gains, stronger legal protection from search, more privacy, and the ability to customize, but also increased risks and responsibilities.

Real Estate Economics

How does residential real estate economics differ from traditional housing economics? Both are concerned with the housing sector, but real estate economics regards it from an investment perspective, housing economics from a policy perspective, especially public policy with respect to housing the poor. The state of real estate economics 20 years ago was excellently surveyed in DiPasquale and Wheaton (1996). In the intervening 20 years, reflecting the shift from public to private provision of housing, much of the best research on housing economics has been done under the rubric of real estate economics.

One important development has been the extension of real estate investment theory to treat uncertainty satisfactorily. Because real estate investment decisions are made under uncertainty, which gradually resolves itself, and are largely irreversible, they are usefully viewed as the exercise of real options (Titman, 1985; Capozza and Li, 1994), to which the theories of options valuation and irreversible investment can be applied.

Traditionally, housing macroeconomics was concerned with the relationship between housing macroeconomic variables (stocks, starts, completions, house values and rents, the mortgage rate, and down payment constraints) and how they are affected by the state of the macroeconomy and monetary policy, as well as with the internal dynamics of real estate cycles (Stein, 1995 is an important early paper). The housing sector in turn was viewed as affecting the broader macroeconomy through investment and employment in the construction industry. Over the last quarter century, two major changes occurred in financial markets that have increased the interdependence between the real estate sector and the rest of the economy. The first was the liberalization of mortgage markets, which allowed financial institutions considerably more flexibility than before in their design of mortgage instruments. The second was the creation of secondary mortgage markets (mortgage-backed securities) that allowed banks to diversify their mortgage risk. After a mortgage crisis in the US in 2008 triggered the Great Recession, study of the real and financial links between the real estate sector, especially the housing sector, and the rest of the macroeconomy, has intensified.

Drawing on the theory of asymmetric information, the microeconomic side of this literature has focused on the optimal design and regulation of contracts between banks and borrowers, developers and contractors, landlords and tenants, and homesellers, homebuyers, and real estate agents. Private contracting should be regulated so that private contracts allocate risk efficiently and provide appropriate incentives. Of special interest has been the appropriate regulation of mortgages and mortgage-backed securities, and of their markets.

Macroeconomists have come to recognize that the real estate sector affects the rest of economy in ways other than just through investment and employment in the construction industry. In the US, between 1983 and 2008 real home prices almost doubled, resulting in very large positive wealth effects that stimulated consumption and investment throughout the economy; the subsequent fall in housing prices during the Great Recession had a corresponding dampening effect. Excessively liberal mortgage financing and weak regulation of secondary mortgage markets led to widespread mortgage default, a banking crisis and a more general financial crisis, as well as undermining consumer confidence. Macroeconomists (e.g., Iacoviello, 2005) have responded by adding housing and housing finance modules (though taking the form of the mortgage borrowing constraint as given) to their models, which are linked to the rest of the economy, through not only the employment market but also the banking sector. To satisfactorily link the microeconomic and macroeconomic literature will require relating the equilibrium set of mortgage-finance related contracts to the state of the macroeconomy and of financial regulation.

Housing Policy

Housing policy debates tend to pit economists against planners and other social scientists. The divisions between the two groups are social philosophical. Since economists value individual freedom to choose, they hesitate more than planners to impose their vision of the good life on others. Also, economists generally emphasize efficiency and believe it is best achieved via markets, with government enforcing property rights and intervening minimally to mitigate market failure, while planners tend to be suspicious of markets, objecting to the inequities and apparent disorder they generate. For all these reasons, planners almost invariably favor more government intervention than do economists.

The planner's view dominated housing policy in western Europe from the 1920s through to the 1980s, especially in the Scandinavian countries where housing markets were largely supplanted by government allocation. The objective of decent and affordable housing for all was largely achieved, but at the cost of considerable inefficiency. During the last decade, these countries have been dismantling the Byzantine housing allocative and regulatory structures that had evolved, and moving to increasingly free housing markets. Thus, the economist's view is currently in the ascendancy.

Economists' belief in the efficiency of markets is at least well reasoned. Two central results in microeconomic theory are the First Theorem of Welfare Economics, which states that under perfect competition a market equilibrium is efficient, and the Second Theorem, which states that (under convexity conditions) any efficient allocation can be achieved as a competitive

equilibrium with appropriate lump-sum redistribution. If, therefore, an economy were to satisfy the Theorems' conditions, efficiency and equity could be simultaneously achieved if the government only enforces property rights and redistributes income appropriately, leaving everything else for markets to do, and almost all government intervention in the housing sector would be harmful. Whether real world economies come close to satisfying the Theorems' conditions is clearly crucial in determining the appropriate role of government in housing provision, and therefore needs to be considered carefully.

Through the 1980s, most neoclassical economists believed that, since buyers and sellers are many and entry and exit free, housing markets are almost perfectly competitive. Nowadays, most housing economists would disagree. Housing markets have many characteristics that violate the assumptions of perfect competition: housing is a heterogeneous good; transactions costs are important; asymmetric information is endemic; and insurance and futures markets in housing are highly incomplete. As well, because the government cannot observe an individual's 'need' but must instead infer it from a signal (e.g., income) that the government can manipulate, all redistributive policies are distortionary. Equity is then most efficiently achieved through a portfolio of policies, including perhaps housing assistance for the poor. Since the optimal portfolio is sensitive to details of the economic environment, housing economists now tend to base their policy advice as much on the results of controlled experiments and empirical studies as on a priori reasoning. While consideration of market imperfections expands the scope of potentially ameliorative government intervention, the political economy literature warns that government, too, is highly imperfect.

Government withdrawal from the housing sector has been especially pronounced in direct housing provision. The current view is that low-income housing projects were a mistake. They led to the alienation and ghettoization of the poor, and were excessively costly (since it is efficient for the poor to live in filtered-down housing). Low-income housing policy in the United States today encourages household mixing by race, ethnicity, and income, through subsidizing the nonpoor to live in the projects and the poor to move into middle-income neighborhoods. Other countries are experimenting with selling project units for owner occupancy.

A classic debate in housing economic policy is whether intervention is more effective on the demand or supply side of the market. Generally speaking, housing demand policies can achieve better targeting, but may drive up rents in the short run until the lagged supply response kicks in. Unlike most other safety-net programs, households eligible for housing subsidy programs are typically not entitled to benefits, which are instead rationed.

Virtually all housing policies apply to either rental or owner-occupied housing but not to both. Most policy makers encourage homeownership, believing it to foster social behavior, civic responsibility, and commitment to neighborhood. The empirical evidence is, however, weak since controlling for the effect that responsible and well-functioning households are more likely to choose homeownership is difficult. Furthermore, homeownership is especially risky for low-income households. Homeownership is encouraged through subsidized and guaranteed mortgages, mortgage interest deductibility under the income tax, tax inducements for first-time homeownership, and subsidies for home improvements.

There are several other elements common to the housing policy debate in most countries.

1. Rent control (reviewed in Arnott, 1995): Economists have been virtually unanimous in their opposition to traditional (now termed 'first-generation') rent controls, which impose ceilings on rents below market-clearing levels. Almost all jurisdictions that have had first-generation rent controls have either dismantled them or replaced them with second-generation programs that stress tenant security rather than rent reduction.
2. Tax policy: Most countries use tax policies to achieve housing policy goals.
3. Eviction rights versus security of tenure: The nature of the landlord–tenant relationship bears the seeds of conflict. Policy concerning protection against eviction for nonpayment of rent entails a classic tradeoff between risk sharing and incentives. Predictably, relative to the United States, European jurisdictions favor the tenant.
4. Zoning, building codes, and density and height restrictions: These significantly affect the spatial pattern of rents and hence household location by income-demographic group. In many cities, height and redevelopment restrictions have forced up downtown property values, driving many poor households previously resident there to surrounding locations. Also, affluent North American suburbs have used minimum lot size zoning and other regulations to exclude the poor.
5. Antidiscrimination: Housing audits have been used in the United States to identify discrimination. A pair of individuals, one of whom belongs to a potentially discriminated-against group, are paid to pose as prospective tenants or prospective homebuyers, and are given the same resume; differences in their treatment are then attributed to discrimination. Also, recent empirical research has attempted to identify discrimination by banks in mortgage lending.
6. Homelessness: In most Western countries, homelessness is now a very visible housing problem. While many causes are sociological – the weakened bonds of family and community, higher mobility, deinstitutionalization of the mentally ill, and increased drug use – economic causes are important too (O'Flaherty, 1996). The withdrawal of governments from the housing sector has caused low-quality housing rents in many cities to rise so much that homelessness is a rational response for some very low-income individuals and households.

Due to rapid population growth and urbanization, by 2050 most of the world's largest cities will be in developing countries. Unlike European and North American cities, their cities tend to have dual housing markets, a formal and an informal housing market. Informal housing ranges from squatter housing to housing in violation of zoning and building regulations, and has dubious legal status. Furthermore, many of these cities are in countries with large informal economies, resulting in household income being difficult to measure. Informality raises a special set of issues with respect to housing policy (Arnott, 2009).

See also: Censuses: Current Approaches and Methods; Cities: Internal Structure; Data Bases and Statistical Systems: Economics (General); Data Bases and Statistical Systems: Economics (Statistical Systems); Economic Geography; Economics, History of; Encyclopedias, Handbooks, and Dictionaries; Families and Households, Formal Demography of; Gated Communities; Gentrification; Home, Anthropology of; Homelessness in the United States; Household Archaeology; Household Economics; Housing and Cities; Infrastructure Investment; Location Theory; Mathematical Models in Geography; Microeconomics, History of; Political Economy, History of; Property: Legal Aspects; Regional Science; Residential Segregation in the United States; Social Simulation: Computational Models; Spatial Mismatch; Spatial Optimization Models; Suburbanization and Suburbanism; Urban History; Urban Internal Spatial Structure; Urban Regeneration.

Bibliography

Alonso, W., 1964. Location and Land Use: Towards a General Theory of Rent. Harvard University Press, Cambridge, MA.

Anas, A., Arnott, R., 1991. Dynamic housing market equilibrium with taste heterogeneity, idiosyncratic perfect foresight and stock conversions. Journal of Housing Economics 1, 2–32.

Anas, A., Cho, J.R., 1988. A dynamic, policy-oriented model of the regulated housing market: the Swedish proto-type. Regional Science and Urban Economics 18, 201–231.

Arnott, R., 1995. Time for revisionism on rent control? Journal of Economic Perspectives 9, 99–120.

Arnott, R., 2009. Housing policy in developing countries: the importance of the informal economy. In: Spence, M., Annez, P., Buckley, R. (Eds.), Urbanization and Growth. The World Bank for the Commission on Growth and Development.

Arnott, R., Davidson, R., Pines, D., 1983. Housing quality, maintenance, and rehabilitation. The Review of Economic Studies 50, 467–494.

Brueckner, J.K., 1986. The structure of urban equilibria: a unified treatment of the Muth–Mills model. In: Mills, E.S. (Ed.), Handbook of Regional and Urban Economics, vol. 2. North-Holland, Amsterdam, The Netherlands, (Chapter 20).

Brueckner, J.K., 2000. Urban growth models with durable housing: an overview. In: Huriot, J.M., Thisse, J.F. (Eds.), Economics of Cities: Theoretical Perspectives. Cambridge University Press, Cambridge, pp. 263–289.

Capozza, D., Li, Y., 1994. The intensity and timing of investment: the case of land. American Economic Review 84, 889–940.

DiPasquale, D., Wheaton, W.C., 1996. Urban Economics and Real Estate Markets. Prentice Hall, Englewood Cliffs, NJ.

Goodman, A.C., 1996. Topics in empirical urban housing research. In: Arnott, R. (Ed.), Encyclopedia of Economics, vol. 1. Harwood, Amsterdam, The Netherlands.

Grigsby, W.G., 1963. Housing Markets and Public Policy. University of Pennsylvania Press, Philadelphia, PA.

Henderson, J.V., Ioannides, Y., 1983. A model of housing tenure choice. American Economic Review 73, 98–113.

Iacoviello, M., 2005. Housing prices, borrowing constraints, and monetary policy in the business cycle. American Economic Review 95, 739–764.

Mills, E.S., 1972. Studies in the Structure of the Urban Economy. The Johns Hopkins University Press, Baltimore, MD.

Muth, R.F., 1969. Cities and Housing. University of Chicago Press, Chicago.

O'Flaherty, B., 1996. Making Room. Harvard University Press, Cambridge, MA.

Olsen, E.O., 1986. The demand and supply of housing service: a critical survey of the empirical literature. In: Handbook of Regional and Urban Economics, vol. 2, Chapter 25.

Rosen, S., 1974. Hedonic prices and implicit markets. Journal of Political Economy 82, 35–55.

Rothenberg, J., Galster, G.C., Butler, R.V., Pitkin, J., 1991. The Maze of Urban Housing Markets. University of Chicago Press, Chicago.

Stein, J.C., 1995. Prices and trading volume in the housing market: a model with down-payment effects. Quarterly Journal of Economics 110, 379–406.

Sweeney, J.L., 1974. A commodity hierarchy model of the rental housing market. Journal of Urban Economics 1, 288–323.

Titman, S., 1985. Urban land prices under uncertainty. American Economic Review 75, 505–514.

Human Behavioral Ecology

Raymond Hames, University of Nebraska, Lincoln, NE, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Behavioral ecology is an explicitly Darwinian approach to the study of behavior. It emphasizes the adaptive value of behavior patterns of individuals as they attempt to solve social and environmental problems that affect their chances of survival and reproduction. The goal of human behavioral ecology is to explain behavioral diversity as a consequence of environmentally contingent responses made by individuals. While some studies show differential reproductive outcomes, most studies rely on proxy measures of fitness. As such, it is the application of a series of behavioral models from evolutionary biology such as optimal foraging theory, parental investment, life history theory, kin selection, reciprocal altruism, and mating strategies. Much of the research to date has focused on behavioral strategies of producing food, resource transfers, marriage, investing in offspring, and managing social interactions between offspring, mates, kin, and unrelated individuals. A large fraction of this work has been accomplished by anthropologists working among traditional peoples.

Aside from behavioral ecology, evolutionary approaches to the study of human behavior fall under a variety of names such as sociobiology, biocultural or biosocial science, human ethology, socioecology, evolutionary anthropology, evolutionary ecology, gene-culture coevolution (or dual inheritance theory), and evolutionary psychology (Cronk, 1991: p. 25). While some of these names are synonyms, others represent different emphases described below. Human behavioral ecology attempts to explain behavioral diversity as a consequence of environmentally contingent responses made by individuals in their attempts to maximize their inclusive fitness. Critical to this goal is the specification of environmental factors that constrain an individual's attempt to maximize fitness. As a consequence, individuals develop behavioral strategies designed to solve a variety of recurrent adaptive problems set by nature.

Origins and Early History of the Field

Behavioral ecology is a relatively new discipline and, for the most part, it is practiced by anthropologists. The publication of E.O. Wilson's *Sociobiology* (1975) along with the earlier work of biosocial anthropologists such as Robin Fox and Lionel Tiger (Cronk, 1991) and, in particular, evolutionary biologist Richard Alexander (1974) were critical in stimulating interest in applying Darwinian approaches to the study of human behavioral diversity. Very quickly Wilson's sociobiological approach was found to be limiting because it was largely based on older ethological approaches which posited relatively simple, general, and rigid models of the relationship between genes and human behavior (Winterhalder and Smith, 1992: p. 9) (*see* Sociobiology: Overview; Sociobiology: History of). This was especially apparent in his final chapter of *Sociobiology* (1975: pp. 547–575) where he considered humans. For example, in that work Wilson suggested that certain behavioral traits such as territoriality among foraging peoples were universal (1975: pp. 564–565). This claim and the approach that generated it was shown to be limited in a paper by Dyson-Hudson and Smith (1978), a kind of approach commonly used today in behavioral ecology to understand human behavioral variation. Using a strategic cost-benefit model developed by the avian behavioral ecologist J. Brown, they showed that territoriality was variable among tribal populations and was determined by specific environmental factors consistent with evolutionary theory.

As will be more fully described later, environmental cost–benefit approaches in the context optimization and game theoretic models are at the core of much of behavioral ecology. It is vital to understand that behavioral ecology is a branch of evolutionary biology and the core of evolutionary biology is about changes in the frequencies of genes through time. However, behavioral ecology is agnostic about the causative role of genes in the study of human behavioral variation. This position was clarified by Alan Grafen (1984) with his notion of the phenotypic gambit. Phenotypic variation is shaped by an individual's genotype interacting with a variety of cultural, developmental, environmental, and other nongenetic factors. The degree to which variation in any trait is a consequence of genetic variation is not directly relevant to the formulation or evaluation of behavioral ecological models. This approach is consistent with analyses in nonhuman behavioral ecology (Alcock, 2005).

Relationship to Other Fields Using Evolutionary Approaches to the Study of Human Behavioral Variation

The field of evolutionary psychology (Buss, 2005) is often inaccurately equated with behavioral ecology. It is therefore critical to distinguish between these complementary Darwinian approaches to the study of human behavioral variation. Behavioral ecology attempts to develop hypotheses regarding variation in behavioral strategies that individuals employ to maximize their inclusive fitness. In some studies, the central focus is on the reproductive consequences of behavior. In contrast, evolutionary psychologists are interested in elucidating the cognitive mechanisms (or mental modules or cognitive architecture) that evolved in humans in the 'environment of evolutionary adaptedness'

(*see* Evolution of Cognition: An Adaptationist Perspective; Human Mate Choice, Evolution of). This approach stresses the fact that all adaptations are to historic environments and it is problematic to assume that traits that were adaptive in the past are adaptive in current environments. As a consequence, evolutionary psychologists seek to uncover mental modules that were designed by natural selection to solve problems such as: What criteria do people use to select mates? How does mate selection criteria differ between the sexes? Who are my kin? How are exchange relationships managed? From this perspective, the actual reproductive consequences of behavior are often of no immediate concern. Other differences distinguishing these fields include behavioral ecology's reliance on naturalistic fieldwork among tribal peoples in contrast to evolutionary psychology's questionnaire surveys and laboratory experiments typically with Western college students, and behavioral ecology's focus on relationship behavior and reproduction in contrast to evolutionary psychology's emphasis on evolved psychological mechanisms that induce behavior. However, behavioral ecologists are beginning to test evolutionary psychological models of mate selection models in nonwestern populations. For example, Marlowe (2004) examined male and female mate preferences among the hunting and gathering Hadza, discovering similarities but also some significance discordances between this population and modern people typically studied by evolutionary psychologists. Recent research on operational sex ratios and marriage markets (Pollet and Nettle, 2008) suggests that evolutionary psychologists' reliance on Trivers' ideas on sex differences in mate choice may require modification. A general critique of too great a reliance on westerners in evolutionary and general psychology research on basic psychological traits is the one by Henrich et al. (2010) emphasizing Marlowe's point about the limitations of research on a 'thin slice of humanity' otherwise known as undergraduate college students.

Multilevel selection (Wilson and Wilson, 2007) and dual inheritance theory (Boyd and Richerson, 1985; *see* Gene-Culture Coevolution) are of interest to some behavioral ecologists because both approaches rely on adaptationist theory. In multilevel selection, it is argued that selection can occur at a variety of biological levels (gene, individual, or group). This approach is of interest to some behavioral ecologists because of its potential to solve first- and second-order collective action problems in the study of cooperation, exchange, and coalitional strategies in warfare.

What is sometimes known as dual inheritance theory or gene-culture coevolution (Boyd and Richerson, 2005) is founded on the realization that cultural traits or memes are similar to genes because both are replicators that use vehicles (bodies or minds) to transmit themselves across generations through social learning 'biases' and are subject to selective forces. Cultural traits differ from genetic traits in their mode of transmission. Certain cultural traits may dominate a cultural pool because they directly enhance the fitness of an individual (e.g., use of spices for meat, see Billing and Sherman, 1998). On the other hand, cultural practices such as the use of fire for cooking (Wrangham et al., 1999) had an enormous impact on our entire digestive system leading to numerous genetic changes. Boyd and Richerson (2005) further propose there are psychological mechanisms such as the tendency to emulate high status people, or copy what the majority of individuals do that are designed by natural selection and crucial for the transmission of culture. In addition, cultural traits may be transmitted laterally within generations instead of lineally through generations allowing them to spread more rapidly. Compared to behavioral ecology and evolutionary psychology, dual inheritance models have only recently begun to be tested empirically, both in the lab using experimental methods similar to evolutionary psychology and in the field in similar populations as to those studied by human behavioral ecologists (*see* Cultural Evolution: Overview).

Relationship to Traditional Social and Behavioral Sciences

Behavioral ecology is distinguished from traditional anthropological and social science theory in its emphasis on the individual or gene as the unit of selection and the idea that much of the content of culture is a consequence of decisions by individuals to enhance their inclusive fitness. Ever since its founding as a discipline, anthropologists have argued that individuals sacrifice themselves to enhance the integration or survival of the social systems in which they participate or are much like puppets or robots manipulated by culture (Cronk, 1991). Furthermore, any attempt to reduce or link cultural behavior to lower levels of analysis such as psychology was specifically condemned. Human behavior was believed to have emergent properties irreducible to the facts of psychology and biology. While it is clear that some dimensions of human behavior are not usefully reduced to underlying psychological or biological processes, the ever-growing body of evidence discussed later tells us that humans are biological and psychological entities and many aspects of human behavioral diversity can be usefully linked to underlying levels of analysis.

Theoretical Armamentarium

Behavioral ecology is strongly informed by a series of theoretical breakthroughs developed in the early 1960s and 1970s in the evolutionary analysis of social and foraging behavior. These breakthroughs occurred in the context of optimization and game theoretic models that predict behavioral strategies consistent with fitness maximization under a variety of constraints. These include inclusive fitness (or kin selection, Hamilton, 1964), senescence (Williams, 1957), reciprocal altruism (Trivers, 1971), sexual selection and parental investment (Trivers, 1972), sex biasing in parental investment (Trivers and Willard, 1973), optimal foraging theory (OFT) (Winterhalder and Smith, 2000), and, most recently, life history theory (Stearns, 1992). These works clearly demonstrated that behavioral variation was an apt subject for investigation from an evolutionary perspective consistent with the modern synthesis and an individual centered view of evolution. As noted, some behavioral ecologists are closely examining the utility of multilevel and cultural selection, especially in the area of cooperation. Early on in the development of behavioral ecology there was an expectation that kin selection would inspire

considerable research especially in the areas of food exchange and other forms of cooperative behavior. As will be evident later on, few studies have made kin selection a central focus of investigation.

Areas of Application

Foraging Theory

The foraging strategies or the collection of wild resources has been a fundamental area of interest because of its implication for the structuring of socioeconomic relations between the sexes, parental investment, group size, and mobility patterns. One of the earliest applications of behavioral ecological research came with the application of OFT (Stephens and Krebs, 1986) in the study of hunting, fishing, and gathering (or foraging behavior) among modern hunter–gatherers and foraging horticulturalists. OFT allows researchers to develop fundamental hypotheses that predict which food resources foragers will pursue when encountered during a search (diet breadth), where foragers will travel to search for resources (patch choice models), and how long they will stay in patches before moving to other areas (marginal value theorem). These models assume that foragers have a goal to maximize their net rate of return while foraging. By doing so they are ultimately able to maximize their fitness. The rate of return is indexed by a currency such as calories or another attribute of food value (e.g., Hill, 1988). This assumption is reasonable because food resources are fundamental to survival and reproduction for one of two reasons. When the environment is poor, foragers are likely to be 'resource limited.' Therefore, maximizing one's net rate of return allows foragers to acquire the greatest amount of food while foraging. On the other hand, when the environment is relatively rich, foragers may be 'time limited' such that efficient acquisition of food resources leaves foragers with more time to engage in alternative fitness-enhancing activities such as childcare (Winterhalder and Smith, 2000).

A number of qualitative and quantitative tests of optimal foraging models have been applied to the problem of diet breadth (Hill and Kaplan, 1992). Qualitative tests predict directional tendencies in prey choice by showing that diet breadth (the number of different species that will be pursued upon encounter), for example, will expand or contract depending on the encounter rates of highly profitable species. Quantitative tests predict the precise number of species a forager will pursue under a variety of environmental conditions (Hill and Kaplan, 1992), which species will enter or drop out of the optimal diet breadth, and the exact rank-ordering of species in the diet breadth. To complicate matters, most studies of diet breadth show a sophisticated division of labor between men and women, and between children and adults (Hill and Kaplan, 1992), which seem to be based on game versus vegetable products for the former and skill and strength levels for the latter. Another constraint influencing diet breadth is pregnancy and dependent childcare for women. In addition, the acquisition of large game or difficult-to-acquire resources may be a means to signal mate value (Smith and Bliege Bird, 2000) as well as to provision the hunter's family.

Although OFT was designed to account for foraging behavior in animals and then later extended to human foraging, recent applications have been made with considerable success in the areas of the origins of agricultural and pastoral food production (Winterhalder and Smith, 2000) and in archaeology (Bird and O'Connell, 2006). It has also rekindled anthropological interest in the problem of conservation among native peoples. Thus far, research indicates that true conservation is rare among native peoples but even more importantly researchers are attempting to specify the conditions under which conservation is likely to emerge as an adaptive strategy (Hames, 2007).

Resource Transfers

The movement of goods, and, to some extent, services, between families and individuals has received considerable attention by behavioral ecologists. Traditionally, such research is subsumed under the terms of exchange or reciprocity (for an extensive review see Gurven, 2004). Behavioral ecologists term such activities transfers because it more neutrally characterizes resource movements between groups and/or individuals compared to traditional terms such as exchange or reciprocity (which assume that the transfer of goods or services between individuals will be returned in some material form at some point in time). In the analysis of transfers, one assumes that giving or yielding a resource that one has acquired will be done such that the owner of the resource gains something in return or avoids a cost in defending that resource. As a consequence, models of transfer assume that the benefits of giving are greater than the costs of monopolizing the resource. The models of transfer developed thus far include (1) scrounging (or tolerated theft); (2) risk minimization; (3) trade or exchange; and (4) costly signaling.

At one extreme, coercive models of scrounging or tolerated theft assume that the resource acquired cannot be easily or effectively defended against those who lack resources, and that the benefit of giving a portion of the resource presents a higher payoff than the cost of defending it against another who is in greater need and thus more willing to contest the resource. At the other extreme are models of sharing, which assume that individuals give resources to others contingent upon an expectation that something of value will be returned to the giver in the future. For example, when resources cannot be predictably acquired (especially true in hunting), risk minimization or pooling strategies are employed through the mechanism of reciprocal altruism. Individuals who acquire resources share with those who have been unlucky in the hunt with the expectation that in the future, should they return home without game, others who have acquired game will share with them. Kin selection models have been used to explain why resource transfers tend to be biased toward close kin; costly signaling (and sexual selection models) argue that actors attempt to acquire resources that require considerable skill to demonstrate their desirability as a mate or ally; and trade models suggest that transfer occurs through specialization via the mechanism of reciprocal altruism.

In collaboration with gene-culture coevolutionists and experimental economists, described earlier, a number of

behavioral ecologists have used experimental economics methods (e.g., dictator, ultimatum, and public goods games) to illuminate how culture and market integration affects patterns of cooperation and exchange (Henrich et al., 2005) across 15 different societies (*see* Human Cooperation, Evolution of). The comparative results indicate that the canonical model (maximizing selfishness) is incorrect in all societies. Instead there is considerable cross-cultural variation and much of the variation is explained by the degree of market integration and the importance of cooperation in daily life.

Marital Strategies

Most marital models are generally founded on Trivers' (1972) observations about mate selection whereby females seek good genes and/or male investment while males seek numerous mates with high reproductive value. Approaches are further informed by a variety of models ranging from the polygyny threshold model (Borgerhoff Mulder, 1990) and pathogen stress (Low, 1990) to male coercion (Marlowe, 2000) and socially imposed monogamy (Alexander, 1974). The polygyny threshold model predicts that females will select mates according to the quantity of resources they control. In short, wealthy males are much more likely to mate polygynously than poor males. In most cases, ownership of material resources (e.g., agricultural land or livestock) is what differentiates poor from wealthy men.

The polygyny threshold model assumes female choice in marital affairs. In many cases, this assumption is clearly suspect and male coercion or parental or kin control may determine a woman's marital fate. Equally troubling is the fact that in many cases female fertility declines under polygyny (especially true for junior wives) and monogamous women oftentimes have higher fertility. At this point, it is clear that polygyny will nearly always enhance a man's fertility but it frequently impairs that of a woman. This latter issue is leading researchers to seek factors that could compensate for lowered fertility. For example, it may be the case that although a woman may bear fewer offspring under polygyny she may bear an equal number of grandchildren as in the case of Mormon polygyny (Josephson, 1993). This would be true if her sons were polygynous, married earlier than average, or married women with high reproductive value.

Although polygyny is permitted in nearly three-quarters of all ethnographic populations, monogamy and the rare case of polyandry (a woman having more than one husband simultaneously) are being investigated through modified versions of the polygyny threshold model. In the case of monogamy it is becoming clear that extramarital mating, or special matings (e.g., concubinage in historic societies) and remarriage following divorce (serial monogamy) is more common among wealthy males. Although such systems are officially monogamous, from a reproductive point of view they produce the high male reproductive variance characteristic of polygynous systems. However, monogamy in many cases is ecologically imposed such that a single male may not have the resources to support more than one wife and in extreme cases difficult environments may lead to polyandry (Starkweather and Hames, 2012). Finally, the advent of the state and legal imposition of socially imposed monogamy is currently poorly understood because these systems often have extremes in male wealth that should lead to polygyny (Geary and Flinn, 2001).

However, in some societies where economic stratification is absent and men are unable to differentially control resources, some men are polygynous because they are exceptionally productive (e.g., excellent hunters) or have achieved political power or status (e.g., distinguished warriors, leaders, or shamans).

Reproduction: Parenting and Life History Strategies

Studies of parental effort have focused on the issues of birth spacing, differential investment in offspring, and the interaction between mating and parenting. The ornithologist Lack was the first to recognize the trade-off between offspring number and offspring quality (or the ability of offspring to thrive and eventually reproduce). Since parental resources are limited, more offspring, through a shortening of the birth interval, will mean less investment per offspring and the potential for reducing each offspring's fitness. This trade-off suggests that there will be an optimal number of offspring that maximizes a parent's fitness and this number will vary with environmental circumstances. In general, one would predict that where parental resources are abundant birth intervals should be short and where parental resources are scarce birth intervals should increase. The cross-cultural evidence from natural fertility traditional societies strongly supports the relationship between parental resources and reproductive success: wealthy parents or culturally successful parents have higher fertility than poor parents (Smith, 2004).

In a seminal article, Irons (1979) suggested that there ought to be a correlation between cultural and reproductive success. Cultural systems are structured so as to encourage individuals to achieve certain kinds of culturally valued statuses. These achievements vary culturally depending on local environmental constraints. Both in traditional and in a number of recent societies, which have not entered the most recent demographic transition, culturally valued statuses are achieved through hunting prowess, religious rank, herd ownership, political leadership, occupational status, or land holdings. At this point, there are more than 30 studies (Hopcroft, 2006) showing in nearly all cases a positive correlation between cultural achievement and fitness (either fertility or survivorship of offspring) or factors strongly associated with fitness (number of wives, frequency of extrapair copulations). However, it is very important to note that this relationship is demonstrated for male but not female reproductive success. Significantly, attributes that lead to high male fitness such as polygyny decrease female fitness at least over the short term but perhaps not the long term. However, the relationship between cultural and reproductive success does not appear to hold in the modern context which is addressed in the section 'Life history.'

Parental investment

In general the amount of investment an offspring receives from parents depends on (1) the relatedness between parent and

offspring; (2) the effect of investment on the offspring; and (3) the effect of investment on parents' future reproduction and survival. These hypotheses have been positively evaluated using cross-cultural and ethnographic cases studies (Daly and Wilson, 1984). For example, the probability of infanticide, abuse, and neglect has been shown to be strongly sensitive to these factors. Infants who are unrelated to a woman's current husband or who have congenital defect or who are born in times of economic privation or whose continued investment would harm the survivorship of older offspring receive little or no investment. The thinking explicit in this approach is that parents adjust their investment in children according to their perceptions of a child's quality, paternity, current environmental conditions, and the needs of other offspring. Further investigations into differential parental investment have been inspired by the Trivers and Willard (1973) model where sex of offspring may influence investment. This model predicts that mothers in good condition will invest more in the sex that has greatest variance in reproductive success (usually males) while mothers in poor condition will invest more in offspring with least reproductive variance (usually females). Research on societies as diverse as hunter–gatherers, pastoralists to eighteenth-century Germany and nineteenth-century India has demonstrated the predicted differential investment by sex and distortions in sex ratios. However, research in modern industrialized nations is not always consistent.

Parenting or mating effort?

Male economic activities, especially hunting, are conventionally seen as kinds of parental effort by males. However, we have long known that males frequently transfer resources to households other than their own. For some theorists, such behavior may be fruitfully viewed as costly signaling or attempts to create new mating relationships or acquire social allies (Smith and Bliege Bird, 2000). Essentially, males are attempting to display honest but costly signals of fitness to others.

Life history

Life history strategies deal with age-specific schedules of mortality and fecundity and the traits that are directly a result of these schedules (Hill and Kaplan, 1999: p. 397; Kaplan et al., 2000). It deals with fundamental developmental events such as age at first and last reproduction, growth rates, birth intervals, and senescence. The theory is based on the assumption that somatic effort (growth, development, and maintenance) and reproductive effort (e.g., mating and parenting) are negatively correlated. Formal research in this area is relatively new but rapidly developing and has focused on the peculiar human trait of menopause, timing of reproduction, and demographic transition theory.

The grandmother hypothesis attempts to explain the evolution of menopause, a life history trait that is nearly absent among mammals but well documented in humans. The model predicts that women will cease reproduction prior to death in order to ensure required investment in current dependent offspring and to assist their children in the rearing of grand offspring. Although attractive, the model does not, as yet, show that continued reproduction until death is not a superior alternative and does not account for longevity of reproductive potential in males (Peccei, 2001). However, it has generated considerable research on the role that grandmothers play in the survivorship and fertility of their kin (Sear and Ruth, 2008) and alloparenting or cooperative breeding (Kramer, 2010).

Throughout history, humans have passed through a number of demographic transitions beginning with a high mortality and high fertility regime to today's worldwide transition to a low mortality and low fertility regime. A central issue behavioral ecologists are confronting is the apparent paradox between fertility and wealth. In tribal and in many state-level historical populations, there is a consistent positive correlation between fertility and wealth. This means that the traditional human pattern was to convert wealth into reproduction. However, at the end of the twentieth century, there is a global tendency for parents to reduce reproduction when wealth increases. Several explanations are offered for this trend. The 'human capital' perspective notes that in the current environment there is an increased requirement of parental investment needed to make children competitive in a modern labor market where earning capacity is determined by parental investment in such things as education (Kaplan and Lancaster, 2010). Others would argue that low reproduction in the context of high wealth is a consequence of a changed environment: for the first time in human history, we possess safe and effective means of contraception, an innovation that humans did not previously possess. This has led to research demonstrating that there is a kind of correlation between wealth and reproductive success, at least for men (Hopcroft, 2006). This research shows that wealthy men have greater numbers of premarital and extramarital unions than poor men (Perusse, 1993). This lends some credence to the perspective that some of the adaptations we currently possess cannot always be understood by looking at the reproductive consequences of behavior.

Summary

Human behavioral ecology is a young and explicitly Darwinian approach to the study of human behavioral variation. It redresses an absence of evolutionary considerations in the analysis of human behavior and social organization in the social and behavioral sciences. Since its inception in the late 1970s it has spawned the development of new journals and academic specialties explicitly devoted to training students to query the evolutionary bases of behavior. Its penetration into academic life has been uneven. It is most prominent in those disciplines such as anthropology and psychology where human evolution is considered relevant for a holistic understanding of human behavior or in economics where self-interest, methodological individualism, and maximization are key elements in modeling behavior. It has been resisted in disciplines such as sociology and political science where findings on human evolution are either ignored or, at best, regarded as irrelevant.

See also: Adaptation, Fitness, and Evolution; Cognition, Evolution of; Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Evolution, History of; Evolutionary Demography: A Darwinian Renaissance in Demography; Evolutionary Social Psychology; Fertility Theory: Embodied-Capital Theory of Life History Evolution; Gene–Culture Coevolution; Human Cooperation, Evolution of; Human Mate Choice, Evolution of; Kin Selection; Kinship, Evolution of; Marriage Systems, Evolution of; Sociobiology, History of; Sociobiology: Overview.

Bibliography

Alcock, J., 2005. Animal Behavior: An Evolutionary Approach. Sinauer, Sunderland, MA.

Alexander, R., 1974. The evolution of social behavior. Annual Review of Ecology and Systematics 5, 325–383.

Billing, J., Sherman, P., 1998. Antimicrobial functions of spices: why some like it hot. Quarterly Review of Biology 73, 3–49.

Bird, D., O'Connell, J., 2006. Behavioral ecology and archaeology. Journal of Archaeological Research 14 (2), 143–188.

Borgerhoff Mulder, M., 1990. Kipsigis women's preferences for wealth men: evidence for female choice in mammals? Behavioral Ecology and Sociobiology 27, 255–264.

Boyd, R., Richerson, P., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago.

Boyd, R., Richerson, P., 2005. Not by Genes Alone: How Culture Transformed Human Evolution. University of Chicago Press, Chicago.

Buss, D.M., 2005. The Handbook of Evolutionary Psychology. Wiley, NY.

Cronk, L., 1991. Human behavioral ecology. Annual Review of Anthropology 20, 25–53.

Daly, M., Wilson, M., 1984. Sociobiological analysis of human infanticide. In: Hausfater, Glenn, Hrdy, Sarah Blaffer (Eds.), Infanticide. Aldine, New York, pp. 487–502.

Dyson-Hudson, R., Smith, E.A., 1978. Human territoriality: an ecological reassessment. American Anthropologist 80, 21–41.

Geary, D.C., Flinn, M., 2001. Evolution of human parental behavior and the human family. Parenting: Science and Practice 1, 5–61.

Gurven, M., 2004. To give and to give not: the behavioral ecology of human food transfers. Behavioral and Brain Sciences 27, 543–583.

Grafen, A., 1984. Natural selection, kin selection, and group selection. In: Krebs, J.R., Davies, N.B. (Eds.), Behavioural Ecology: An Evolutionary Approach. Sinauer, Sunderland, MA, pp. 62–84.

Hames, R., 2007. The ecologically noble savage debate. Annual Review of Anthropology 36, 177–190.

Hamilton, W., 1964. The genetical evolution of social behavior. Journal of Theoretical Biology 7, 1–52.

Henrich, J., Heine, S., Norenzayan, A., 2010. The weirdest people in the world. Behavioral and Brain Sciences 33, 61–135.

Hill, K., 1988. Macronutrient modifications of optimal foraging theory: an approach using indifference curves applied to some modern foragers. Human Ecology 16, 157–197.

Hill, K., Kaplan, H., 1992. The evolutionary ecology of food acquisition. In: Smith, E., Winterhalder, B. (Eds.), Evolutionary Ecology and Human Behavior. Aldine de Gruyter, New York, pp. 167–202.

Hill, K., Kaplan, H., 1999. Life history traits in humans: theory and empirical studies. Annual Review of Anthropology 28, 397–430.

Hopcroft, R., 2006. Sex, status, and reproductive success in the contemporary United States. Evolution and Human Behavior 27 (2), 104–120.

Irons, W., 1979. Cultural and biological success. In: Chagnon, N., Irons, W. (Eds.), Evolutionary Biology and Human Social Behavior: An Anthropological Perspective. Duxbury Press, North Scituate, MA, pp. 257–272.

Josephson, S.C., 1993. Status, reproductive success, and marrying polygynously. Ethology and Sociobiology 14, 391–396.

Kaplan, H., Hill, K., Lancaster, J., Hurtado, A.M., 2000. A theory of human life history evolution: diet, intelligence, and longevity. Evolutionary Anthropology 9, 156–185.

Kaplan, H., Lancaster, J., 2010. Embodied capital and extra-somatic wealth in human evolution and human history. In: Muehlenbein, M. (Ed.), Human Evolutionary Biology. Cambridge University Press, New York, pp. 439–455.

Kramer, C., 2010. Cooperative breeding and its significance to the demographic success of humans. Annual Review of Anthropology 39, 417–436.

Low, B., 1990. Marriage systems and pathogen stress in human societies. American Zoologist 30, 325–339.

Marlowe, F., 2000. Paternal investment and the human mating system. Behavioural Processes 51, 45–61.

Marlowe, F.W., 2004. Mate preferences among Hadza hunter-gatherers. Human Nature 15, 365–376.

Peccei, J.S., 2001. A critique of the grandmother hypotheses: old and new. American Journal of Human Biology 13 (4), 434–452.

Perusse, D., 1993. Cultural and reproductive success in industrial societies: testing the relationship at the proximate and ultimate levels. Behavioral and Brain Sciences 16, 267–322.

Pollet, T.V., Nettle, D., 2008. Driving a hard bargain: sex ratio and male marriage success in a historical US population. Biology Letters 4 (1), 31–33.

Sear, R., Ruth, M., 2008. Who keeps children alive? A review of the effects of kin on child survival. Evolution and Human Behavior 29 (1), 1–18.

Smith, E.A., 2004. Why do good hunters have higher reproductive success? Human Nature 15 (4), 343–364.

Smith, E.A., Bliege Bird, R.L., 2000. Costly signaling and turtle hunting. Evolution and Human Behavior 21, 110–122.

Starkweather, K., Hames, R., 2012. A survey of non-classical polyandry. Human Nature 23, 149–172.

Stearns, S., 1992. The Evolution of Life Histories. Oxford University Press, Oxford, UK.

Stephens, D., Krebs, J., 1986. Foraging Theory. Princeton University Press, Princeton.

Trivers, R., 1971. The evolution of reciprocal altruism. Quarterly Review of Biology 46, 35–57.

Trivers, R., 1972. Parental investment and sexual selection. In: Campbell, B. (Ed.), Sexual Selection and the Descent of Man: 1871–1971. Aldine, Chicago, pp. 136–179.

Trivers, R., Willard, D., 1973. Natural selection of parental ability to vary the sex ratio of offspring. Science 179, 90–92.

Wilson, E.O., 1975. Sociobiology: The New Synthesis. Harvard University Press, Cambridge, MA.

Williams, G.C., 1957. Pleiotrophy, natural selection, and the evolution of senescence. Evolution 11, 32–39.

Wilson, D.S., Wilson, E.O., 2007. Rethinking the theoretical foundation of sociobiology. The Quarterly Review of Biology 82, 327–348.

Winterhalder, B., Smith, E., 1992. Evolutionary ecology and the social sciences. In: Winterhalder, B., Smith, E. (Eds.), Evolutionary Ecology and Human Behavior. Aldine de Gruyter, Hawthorne, NY, pp. 3–23.

Winterhalder, B., Smith, E., 2000. Analyzing adaptive strategies: human behavioral ecology at twenty-five. Evolutionary Anthropology 7, 51–72.

Wrangham, R.W., Jones, J.H., Laden, G., Pilbeam, D., Conklin-Brittain, N., 1999. The raw and the stolen: cooking and the ecology of human origins. Current Anthropology 40, 567–594.

Human Cognition, Evolution of

Harry Heft, Denison University, Granville, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Investigations from diverse domains of study – from developmental and comparative psychology to developmental biology, evolutionary theory, and paleoarchaeology – have been converging to offer a fresh understanding of the evolution of modern human cognition. This emerging picture suggests that our hominid ancestors engaged in complex forms of cooperative social processes and protocultural practices and that *Homo sapiens* evolved in relation to the functional opportunities afforded by these background ecological conditions. Particularly significant is the emergence of processes supporting the cross-generational transmission of information during ontogeny and the pedagogical and technical means to preserve the knowledge gains of prior generations.

Introduction

Scientific inquiry into the evolution of modern human cognition continues to be entangled in centuries' long nativism–empiricism debates from epistemology. In the wake of the nineteenth- and twentieth-century developments in science and technology, these alternatives have taken on distinctive forms. Nativism has been naturalized by being grounded in an individual's genetic endowment, which reflects outcomes of natural selection processes once operating on the species. Empiricism has shed any allusion to a blank slate, if it ever did hold such a view (even John Locke assumed some characteristic ways in which the mind operates). Today, all camps agree that biological and experiential factors play crucial, contributory roles in the ontogenesis of human cognition. However, this appearance of partial unanimity should not mask the striking, underlying legacy of differences that persists.

Over the past half century, nativism in psychology – which previously had been mired in imprecise and largely descriptive instinct theories – received a fresh boost from Chomsky's treatment of language as an innate biological endowment, operating independently of other cognitive functions, and requiring minimal environmental input to set in motion. This nativist approach to language prompted others to propose a wide-ranging set of innate, modularized cognitive functions that account for the varied, reasoning-like cognitive operations of the human mind. These modules are viewed as being grounded in encapsulated, brain processes that are products of selection over phylogenesis. Although these nascent functions are 'triggered' by environmental circumstances, their specific character is not acquired (learned). In some circles, these functions are conceptualized as computational processes operating on symbolic, mental representations of the environment and of thought. As a shorthand, the author will refer to this group of approaches, as some adherents refer to themselves, as *neo-nativist*. Neo-nativist proposals are located at the convergence of several twentieth-century developments: psycholinguistics, neo-Darwinian population biology (the synthesis of Darwinian selection and genetics), computer and cognitive science, as well as certain strains of comparative psychology and ethology.

In contrast, an alternative line of thought has emerged over this same time period, although its roots can be traced back to earlier efforts in developmental biology and embryology (Gottlieb, 1992). This alternative shares the empiricist emphasis on the environmental context for developing psychological processes, but is otherwise quite different from that lineage. For convenience, this approach will be referred to here as *transactionist*. The principal differences between standard empiricist approaches and transactionist approaches rest on their conceptualization of the organism. Empiricist approaches have tended historically to assume that the organism is a relatively passive recipient of the press of environmental circumstances, whereas transactionist views emphasize the active character of organismic functioning and the reciprocal, mutually constituting nature of environment–organism relations over time. From the standard empiricist approach, the individual may be sensitive, and even responsive, to contingent relations among, for example, patterns of stimulation; however, psychological processes do not contribute in any substantive way to the character of those relations. Transactionism emphasizes that, with few exceptions, organisms are active *participants* in animal–environment systems (Gottlieb, 2003; Oyama, 1985; Thelen and Smith, 1996). A transactionist treatment of the evolution of modern human cognition grows, in part, from developmental biology and the confluence of ecological psychology, embodied cognition, dynamical systems thinking, and evidence from paleoarcheology.

Undeniably, neo-nativist approaches give considerable attention to the environment, but most substantively with regard to its role in phylogenesis. Grounded as these approaches are in neo-Darwinian thinking, it is ancestral environments rather than contemporary conditions and their influence on ontogenesis that are accorded central importance. From a neo-nativist perspective, experience over ontogeny serves as an overlay to the expression of innate structures. There is a body of research that would seem to be consistent with neo-nativist claims of innate, modularized cognitive functions appearing as early as 3 months (e.g., Carey, 2009; Spelke and Kinzler, 2009). However, this line of research is faced with the seemingly irresolvable challenge of how to assess infant knowledge prior to any postnatal experience. Also related to

this challenge is the problematic nature of the very concept of the innate or instinctive, which has been the target of considerable criticism by behavioral biologists for over half a century (e.g., Hailman, 1969; Lehrman, 1953). In addition, evidence has shown that some of the neo-nativist laboratory findings are highly dependent on the specific testing conditions employed (e.g., Learmonth et al., 2002; Thelen and Smith, 1996). From a transactionist perspective, the neo-nativist stance is judged to overlook the constitutive role that current environmental structures and sociocultural–historical processes play in ontogenetic development. Related to these points, nativist approaches are criticized for failing to recognize the ongoing role of the individual and the group in shaping environmental influences. Obviously, these are contentious issues between the two camps. For brevity, this essay primarily focuses on features of a transactionist approach.

Hominidae

Perhaps the major challenge facing any attempt to understand the evolution of modern human cognition is the absence of other extant species within the genus *Homo* besides us. For this reason, one must turn to two sources of comparative evidence: (1) other living genera of the Hominidae family: *Pongo* (orangutans), *Gorilla*, and *Pan* (chimpanzee) and (2) paleontological and archaeological remains of ancestral species within the *Homo* lineage. There are obstacles to be faced in pursuing either line of investigation. On the one hand, comparisons with other living primates must be tempered by the recognition that the common ancestor to these lineages is estimated to have existed roughly 6 My. The very long stretches of time over which these lineages have been evolving separately should be borne in mind. In the case of the paleontological and archaeological record, on the other hand, evidence can be difficult to come by. Only the most durable traces remain in the face of the assaults of time. Moreover, the identification and sequencing of distinct prior species is quite fluid. The number of identified species of *Homo* has been growing with each decade – there are minimally 20 at present – and the evidence continues to accumulate and prior findings reevaluated in light of them. Still, it is clear that a view that human evolution was a linear succession of species can no longer be sustained.

When modern humans are compared with chimps and bonobos, striking similarities are found and also differences. The differences are particularly intriguing because of the apparently high degree of genotypic similarity between these species. Among the most obvious psychological differences between humans as compared to apes are several qualities: greater complexity and variation in patterns of social interaction and cooperation among humans and the extended period of postnatal neurological immaturity and plasticity in humans, particularly in the first years of life. As it will be seen, these two qualities are related. Impressive indications of cooperation have been observed among nonhuman primates, both in the field and in the laboratory, especially with regard to hunting, defense, and problem solving generally. However, the patterns of cooperation that arise within the functioning of human social groups, including participation in sociocultural structures, symbolic means of engaging the world, and the greater potential to plan prospectively, give rise to possibilities that far outstrip what is seen among nonhuman primates.

Before further considering the functional possibilities that arise out of human social interaction, one must first face up to the challenge that cooperation generally poses for standard accounts of evolution. The challenge is that cooperative actions typically appear to incur costs to the actor, and for this reason, they would appear to run counter to the Darwinian assumption that individuals function to maximize their reproductive success. As such, an individual's actions that do not contribute to his or her own reproductive success, and the innate structures in which those actions are grounded will not be sustained over time in the population as a matter of course. Part of the conceptual challenge here is the neo-Darwinian assumption, which has dominated evolutionary thinking in biology and psychology for decades, that selection processes operate only at the level of the individual, or reducing the focus of selection to even a more microlevel, to the genes. Attempts to reconcile cooperation and self-interest at the level of individual selection have taken two forms: inclusive reproductive fitness or kin selection (promoting the reproductive success of individuals with similar genotypes) and reciprocity (actions that may produce short-term losses but with the expectation of long-term positive gains from others).

Of late, continuing debates over the limitations of these preceding claims have led to reconsideration of the possibility that selection operating simultaneously at multiple levels, including that of the group (posited earlier by Darwin in *The Descent of Man*), may be more robust grounds for explaining cooperative actions than neo-Darwinian explanations (Sober and Wilson, 1998). If individuals within a group on average benefit from cooperative processes, and if cooperation contributes to the longevity of the group and as a consequence, the reproductive success of sufficient numbers of individuals, then cooperation among individuals could be selected for and sustained over time at the group level. Notable has been the recent embrace of this viewpoint by one of the most prominent early advocates of the inclusive fitness view (Wilson, 2012). This emphasis on group selection poses significant challenges to the Neo-Darwinian point of view that has long held sway in psychology, principally in the hands of the so-called evolutionary psychologists. Importantly, as it will be seen, inclusion of group selection as a consideration regarding cooperative behavior is consistent with shifting views on the nature of infant development in psychology.

Intersubjectivity

Western conceptions of child development have long adopted the view that development is largely an individual journey and that the social aspects of development only assume a significant place beyond the first year once certain cognitive foundations are established by the individual and, in turn, symbolic thought begins to develop. Neither is this view no longer tenable empirically, nor is the tacit assumption – a remnant of outmoded 'faculty psychologies' – that social and cognitive development is separable. Beginning with research in the 1960s, it has become clear that human infants are remarkably

responsive toward other persons. Attention is directed more toward humanlike facial patterns than toward other features in the surround, particularly if those patterns are in motion. Further, cases of imitation by neonates have been reported in recent decades, although not without some challenge. Even by the most conservative interpretation, however, it is clear that to the extent that neonates can control their direction of gaze, they attend especially to social others, or at least humanlike visual patterns. Moreover, recent evidence indicates that throughout the first months, there is a growing attunement to the actions and even intentions of others that are directed toward the self (Reddy et al., 2013).

Often, findings such as these have been taken to indicate a preparedness for facial patterns per se, without viewing this sensitivity in a wider context that cuts across time-worn separate psychological domains. Attention (cognition) to faces is inseparable from attention to information of a social nature and ultimately an interpersonal nature. Specifically, these actions have been viewed by some researchers as evidence for the beginnings of 'intersubjectivity,' which refers to behavior that suggests the establishing of mutual awareness or attention between individuals, at least for brief intervals. Importantly, beginning around 6 months of age, mutual attention broadens such that infants begin to demonstrate sensitivity to the direction of another's gaze and to pointing, and by 12 months, such joint attention is commonplace among human infants. Notably, joint attention appears to be unique among humans. Although a few marginal cases of joint attention have been reported with laboratory-reared primates, it has not yet been observed in the wild. Whereas a wide range of animals, including birds and nonhuman primates, clearly learn about features of the environment through observing the actions of conspecifics, and in addition, adults of some species have been found to leave materials about for the young to practice manipulating (e.g., in the support of nest building), joint attention and intersubjectivity seem to be nearly unique human phenomena.

This capacity for establishing mutual awareness is a momentous step in human cognitive evolution because with its emergence, the infant's attention can be guided by the actions of more experienced individuals to particular environmental features. The significance of such human intersubjective processes for learning and development cannot be overstated; it underwrites all forms of pedagogy – the intentional transmission of information between generations. Broadly considered, it is the foundation for a great deal that transpires cognitively in ontogeny as well as for cultural evolution and cross-generational teaching. This claim is consistent with the findings that infants and children who have difficulties establishing or maintaining intersubjective processes face enculturation difficulties.

Hominin Evolution

The findings in support of the early social nature of infant functioning converge with the picture of hominin evolution that has been emerging over the past several decades. Cultural processes among hominin did not begin with the appearance of *Homo sapiens*. Paleoarchaeological evidence has revealed the establishment of semipermanent settlements beginning most likely with *Homo habilis* (2.3–1.6 My), and clearly with the appearance of *Homo erectus* (1.8 My). Such evidence indicates that at this point in their evolution, hominins were intensely social in ways not seen among extant nonhuman primates. (Cole and Hatano, 2007; Megarry, 1995).

The construction and maintenance of even the most basic settlement arrangements require a combination of foresight and interpersonal awareness as well as a rich degree of communication that distinguishes the hominid line at this juncture of evolution (Sterelny, 2003). The evidence for extensive migration by *H. erectus* and *H. sapiens* further suggests advanced patterns of cooperative behavior, communication, and planning. Wide dispersal through migrations by *H. sapiens* and *Homo neanderthalensis* is not seen among the other genera of the Hominidae family.

Evidence for semipermanent settlements and migration coincides with the findings of positive correlations between the size of the neocortex and group size among primates. Extrapolating this relationship to humans, Dunbar (1995) estimated that human social groups would not exceed 150 individuals and that prediction seems to hold up in comparative assessments of human societies. This is not to say that recognition of others would be limited to that number; indeed, human memory for faces has been found to be prodigious. Rather, this number points to a threshold for possessing sufficient knowledge of others and for self-monitoring to maintain interpersonal, cooperative relationships at the scale of communities. It seems likely that increased neocortical size and community size may have coevolved. As social groups become larger, more possibilities for cooperation and collaboration driven by enhanced prefrontal executive processes can result. Group success relative to other groups would favor those individuals who could function cooperatively in communities. As noted, human social groups historically were rather dispersed geographically, and for this reason, the 'ways of the group' were likely to vary due to local conditions and social history. Individuals would benefit from prolonged early periods of neurological plasticity in order to learn the ways of the specific group into which they were born.

Toolmaking

The archeological record of changes in toolmaking also points to the critical relationship between social and cognitive processes and to the role that pedagogy likely played in the appearance of early cultural processes. Taken broadly, a tool is any material feature of the environment used by an animal to mediate its engagement with other features of the environment. Working from this definition, many organisms are tool users. For example, chimps have often been observed utilizing a branch as a probe in a termite hill or a leaf as a vessel for transporting water to the mouth, elephants use branches as switches to swat insects, dolphins make use of sponges to harvest crustaceans from the sea floor, and some birds employ pine needles to probe bark for grubs. However, the modification of material features for the purpose of tool use is far less common. Tool construction requires considerable foresight (executive control) because the individual is operating to

realize some affordance that does not presently exist. In addition to tool construction, another distinguishing feature of advances in tool use is innovation. The animal must recognize that some feature has the potential for serving as a tool with the appropriate modification. In such cases, the animal needs to be attentive to the feature's dual affordance character: that once modified it can be wielded or otherwise manipulated by the individual and that it is suitable for the task in question (van Leeuwen et al., 1994).

Critical for the present discussion is the manner in which knowledge of tool use and tool construction is transmitted between generations. Research among primates has shown that the young of a species must have the opportunity to observe tool use by others and also have the resources for tool use available to them. With humans, the possibilities for transmitting information are greatly expanded by means of joint attention and guided participation, and later language, allowing for demonstration and instruction. In short, master–apprentice relationships that require intentional instruction are only possible on the back of these uniquely human functions.

When examining the evidence for early toolmaking by hominins, researchers are limited to those materials that can withstand the effects of time. For this reason, our knowledge of early hominin tool use and manufacture is limited primarily to stone implements and shells. Hides, sinews, and later, even fabric may well have been used extensively, but only in exceptional cases of preservation will such materials be found. Nonhuman primates have most likely been using selected rocks to, for example, crack nuts for millennia, and there is little reason to assume that these actions have changed over time. Evidence of stone toolmaking has been found only in the proximity of hominin remains or settlements. The pattern that is revealed in the record of stone tools over time among hominids is marked by very long periods of design stability, followed by small innovations that are preserved for considerable durations and then followed by a period of rapid innovation and change that continues to the present day.

The earliest manufactured stone tools have single chipped edges and have been dated at 2.7 My. The design was remarkably stable, spanning roughly 1 My. These artifacts (Oldowan tools) have been attributed to *Australopithecus garhi*, and were later produced by *H. habilis* (Mithen, 1996). In spite of the primitive appearance of these tools, their production requires some technical skill and foresight. The technique is commonly referred to as stone knapping, and it involves utilizing one rock as a percussor to chip away at another, less dense rock, calving off stone flakes in order to leave behind a sharp edge (Roux and Bril, 2005). Single-edged stone tools can be used for cutting, scraping, or gouging. These percussive actions are not random, but rather must be controlled as to force, angle, and location in order to produce a conchoidal fracturing of the implement. The production of these artifacts suggests a degree of executive control over motor functions yet to be found in nonhuman primates. It has been proposed that these skills point to cortical reorganization involving extensive neural connections between the prefrontal and cerebellar regions (Donald, 2001).

Further, production of this type of artifact necessitates an understanding of materials including the probable patterns of fracture of particular rocks. It is most unlikely that the skills associated with these techniques could be acquired through mere imitation. The complexity of the process and, significantly, its stability of production point to intentional instruction from artisan to learner along the lines of joint attention and guided instruction as described above. In short, these rather simple appearing implements index a major cognitive shift in hominid evolution: a new mode of deliberative information transmission across generations (Tomasello, 2001). At the same time, the recurrence of these patterns of production over vast stretches of time points to limitations in innovation and invention, or alternatively, the difficulty of preserving innovation across generations. In this regard, the cumulative results of innovation crucially depend on the possibilities of preserving through social transmission the gains of prior generations (Tomasello's 'ratchet effect'). Finally, the materials utilized in toolmaking have sometimes been traced to locations at a distance from where they have been found, suggesting an impressive level of deliberate planning rather than the fortuitous use of found materials. All these characteristics of even such relatively simple tools point to notable cognitive advances associated with increased cortical size in the hominin lineage.

The production and utilization of double-edged, quasi-symmetrical, chipped rocks with sharp points (Acheulean tools), which are attributed to *H. erectus*, begin to appear between 1.7 and 0.25 My (Abramiuk, 2012). These bifacial tools can be utilized for a far wider range of purposes than their unifacial predecessors. Designs varied in width and length, adapted for various tasks. Clearly, such tools require a greater degree of skill and planning yet seen, pointing to advances in planning with particular ends and purposes in mind, an understanding of materials, and the operation of pedagogy that permits the social transmission of the considerable know-how needed to produce these artifacts. This evidence points to the collaborative synergy of cognitive, motor, and social functions.

Symbolic Cognition

Communication, for present purposes, refers to the intentional sharing of information between individuals. (Obviously, communication from a cross-species perspective has a much broader and varied definition.) Fundamentally, it is a means by which one individual directs the attention of another to a feature of the environment or, later in development, to an abstraction such as an idea. Pointing is perhaps the most basic gestural form of interpersonal communication, and, as discussed above, it requires establishing an intersubjective relation (mutual awareness) among individuals.

Symbolic forms of intentional communication typically involve the utilization of auditory and visual patterns whose meaning is based on a conventional relationship between the symbol and its referent. In short, there must be a common understanding among interlocutors – from merely two individuals to a society to a global network – as to what an auditory or visual pattern refers. In turn, referential meanings reside in a shared semantic network which, as a matter of course, will require considerable exposure to discern. The capacity to

engage in symbolic communication in the form of an oral language or gesture appears to be universal, stemming at least in part (see below) from our phylogenetic history. Other forms of communication, such as pictures and written language, also depend on a symbolic capacity, and yet these forms derive from particular sociocultural histories and, hence, may not necessarily be universal. In general, symbolic thought comes to mediate most activities of a sociocultural nature. Theoretical traditions that emphasize the central role of activity in knowing (e.g., Vygotskian approaches) claim that what begins as interpersonal symbolic communication eventually becomes internalized and serves as the basis for intrapersonal symbolic thought. The emergence of symbolic communication broadly is tied to phylogenetic changes in the neocortex, as well as enhanced possibilities for early neuroplasticity (see, Deacon, 1997; Donald, 2001).

Much uncertainty shrouds the evolutionary transition from gestural communication along the lines of pointing to symbolic communication. The gap between these two forms of communication seems considerable. In the absence of other extant species of *Homo*, the transition from gesture to symbol may not be as apparent as it otherwise might be (Donald, 2001). However, vestiges of an intermediate form of communication can still be seen among modern humans, and occasionally isolated in some neurological cases, in the form of nonverbal, mimetic (meaningful) expression. Gesture is an integral part of human oral communication as a complement to speech (Goldin-Meadow, 2003), and it can also function alone as a symbolic form of communication, often of an iconic character, by representing, for example, some action or some feeling. Mimetic gesture can substitute for speech when interlocutors do not share a common language, and when circumstances are not conducive for communicating verbally. Notably, mimetic gesture can readily be seen in most cultures where traditional ritualized actions (e.g., religious practices) continue to function even in the midst of oral and written language. Donald speculates that mimetic (nonverbal) expression is a phase of communication that separately operated prior to oral language, and has subsequently been enfolded phylogenetically into communication functions generally. The continuing presence of mimetic expression offers the tantalizing hypothesis that the capacity to think symbolically evolved gradually, with mimesis being a crucial early step. Recent research on embodied meaning (e.g., Glenberg and Kaschak, 2002) and on the relationship between gesture and thinking (e.g., Goldin-Meadow and Beilock, 2010) aligns with this hypothesis.

Still, the question remains as to the ontogenesis of symbolic communication and of language more widely. Research by Goldin-Meadow et al. (2007) demonstrates the invention of idiosyncratic gestural communication (mimetic expression) among a sample of infants who are deaf siblings. These findings suggest that the rudiments of expressive actions at least do not require much in the way of environmental priming. Can the same be said of spoken language? The identification of common syntactic structures seemingly at work even across very diverse languages pointed to the possibility of innate brain structures (a 'language acquisition device') that have arisen through phylogenesis. These claims, which were accompanied by a barrage of arguments against the alternative possibility that language development could possibly proceed through learning and conditioning, left few cogent alternative accounts of language development for decades. But new possibilities have started to emerge that not so much undermine the evolutionary roots of language, but rather shift its emphasis somewhat by not losing sight of the fact that *H. sapiens* are fundamentally sociocultural, biological beings.

The Human Eco-niche

The expression 'ecological niche' (or eco-niche) is typically employed to refer to the subset of features and resources in a given locale in relation to which an animal carries out its daily activities. The standard use of eco-niche carries with it a presupposition that the organisms must fit adaptively 'into' their eco-niche. From this vantage point, adaptation is viewed as the organism's efforts to 'answer' a functional problem posed by the environment. In an influential paper, Lewontin (1978) argued that this conceptualization of adaptation, and the unidirectional view of environment–organism relations it conveys, is exceedingly narrow in its failure to recognize – as Darwin (1881) demonstrated in his final book – that organisms commonly alter their surrounds in order to better function in them. In recent years, this notion of 'niche construction' has received considerable attention in large measure due to the work of Odling-Smee et al. (2003), who have documented the ubiquitous character of niche construction across living organisms, from microorganisms and plants to animals. Living organisms alter their surrounds to varying degrees, and these actions play a *constituent* role in the fashioning and maintenance of their eco-niche. Moreover, such actions can, in some cases, alter environmental selection pressures and over time, resulting in shifts in the genotypic makeup of the population. In this way, actions of organisms can at times be at the leading edge of phylogenetic change.

Because niche construction is so commonplace in nature, it is no longer viable to claim, as once was the case, that what distinguishes hominins from other living things is the fact that they alter their environment (*Homo faber*). Nor are hominins to be distinguished by the *extent* to which they do so. Insects surely are the winners there. Instead, what is distinctive is the manner in which the fashioning of the environment through daily actions, and cumulatively over historical time, has come to play an essential role in supporting and shaping the ontogenetic development of humans as social beings.

Although ontogenetic development occurs in relation to some enduring properties of environments (e.g., a gravitational field), it also does so in relation to repeating patterns of activity among the individual's group. Particular patterns of group function typically stem from a cumulative history of addressing local challenges and, to a lesser extent, from the infusion of influences from external sources. An extended period of neurological and behavioral immaturity among infants and children affords opportunities for extensive discovery and learning from others, particularly with the scaffolding from intentional, guided teaching. Significantly, given this extended period of immaturity in humans, the actions of others in

structuring the developmental environment (niche construction) become especially crucial. What is found among newborns is considerable flexibility in adopting the ways of living that are specific to their social group, and in turn the development of skills though pedagogy (e.g., literacy) for retaining the knowledge gains of previous generations. Further, the learning of rituals, traditions, and social norms promotes group cohesiveness.

Such an emphasis on the essential role of the human eco-niche in ontogeny has reopened the debate on language origins as well as the development of other modes of cognition. Claims of innateness in cognitive abilities can be reevaluated by recognizing that biological processes at all points of development and to varying degrees are context sensitive. Contextual influences are among the *constitutive factors* in ontogenetic development, and human development in particular (Blumberg, 2005; Oyama, 1985). For this reason, ongoing investigative attention needs to be directed at contexts for development.

We are beginning to learn that infants are exquisitely sensitive to regularities in visual events and in the auditory stream (Aslin et al., 1999). Growing evidence, for example, is pointing to infants' sensitivity to the statistical regularities present in infant-directed speech (Saffran, 2003). It is an intriguing possibility that forms of infant-directed speech, indeed, even the structure of human languages themselves, might have emerged pragmatically from sociocultural practices over millennia as human societies discovered effective ways to make linguistic information available to be discovered by each generation (Deacon, 1997). Such practices surely would be shaped by constraints in human ways of functioning, reflecting an organism–environment mutuality that characterizes ecosystem processes generally. The possibility for acquiring language is species specific, no doubt. But the structure of language that seems to work most effectively, as well as the means of teaching it, may be the result of a slow process of sociocultural discovery. Variation in paths of development, in contrast, may reflect practices that have arisen in relation to particular local circumstances and that have proven to be effective for enculturation to the ways of the group. The view that emerges is an on-going, synergistic, coevolution of ecological structures and cognitive possibilities over multiple time scales.

Likewise, synergistic processes could be operating in the case of less canalized forms of cognition in tandem with the development of techniques and tools for thinking. Literacy emerged as a sociocultural practice, and we are still in the process of discovering ways to facilitate the development of its requisite skills. Similarly, techniques for representing spatial relationships (e.g., maps) are cultural innovations, and efforts to enhance the utilization of these representations are ongoing (Uttal et al., 2013). The latter two examples also point to a significant shift in cognitive possibilities among humans, namely, the development of tools that have the effect of distributing task demands across the individual–artifact relation (Clark, 2008). Specifically, the creation of symbolic notation for the means of keeping records has allowed individuals and societies (e.g., Connerton, 1989) to expand memory in the form of written or graphic records. Moreover, problem-solving capacities can be enhanced by using a display of information, such as calculations on paper, as a 'work space' for cognitive tasks (Wilson, 2004).

The pace of cultural change has been accelerating, and with it the cognitive changes needed for living in an increasingly complex world in a sustainable manner. The recent development of electronic global communication networks is yet another reflection of the enduring tendency for all animals to transform their eco-niche. However, the far-reaching nature of these changes has few parallels in human history, comparable perhaps only to the invention of writing, mathematics, and time. Its influence on the continuing evolution of human cognitions is barely visible on the horizon.

See also: Biological Preconditions for Language Development; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Cultural Evolution: Overview; Cultural Evolution: Phylogeny versus Reticulation; Cultural Evolution: Theory and Models; Gesture in Linguistics; Human Language, Evolution of; Human Language: Evolutionary Precursors; Language Acquisition; Language Development: Evolutionary Theories; Williams Syndrome; Writing Systems, Psychology of.

Bibliography

Abramiuk, M., 2012. The Foundations of Cognitive Archeology. MIT Press, Cambridge.
Aslin, R., Saffran, J., Newport, Elissa L., 1999. Statistical learning in linguistic and nonlinguistic domains. In: MacWinney, B. (Ed.), The Emergence of Language. Lawrence Erlbaum Associates, Mahwah, NJ.
Blumberg, M., 2005. Basic Instinct: The Genesis of Behavior. Thunder's Mouth Press, New York.
Carey, S., 2009. The Origin of Concepts. Oxford University Press, New York.
Clark, A., 2008. Supersizing the Mind: Embodiment, Action, and Cognitive Extension. Oxford University Press, New York.
Cole, M., Hatano, G., 2007. Cultural-historical activity theory: integrating phylogeny, cultural history, and ontogenesis in cultural psychology. In: Kitayama, S., Cohen, D. (Eds.), Handbook of Cultural Psychology. Guilford Press, New York, pp. 109–135.
Connerton, P., 1989. How Societies Remember. Cambridge University Press, Cambridge.
Darwin, C., 1881. The Formation of Vegetable Mould, through the Action of Worms, with Observations on Their Habits. John Murray, London.
Deacon, T., 1997. The Symbolic Species: The Co-evolution of Language and the Brain. W.W. Norton, New York.
Donald, M., 2001. A Mind So Rare: The Evolution of Human Consciousness. W.W. Norton, New York.
Dunbar, R., 1995. Neocortex size and group size in primates: a test of a hypothesis. Journal of Human Evolution 28, 287–296.
Glenberg, A., Kaschak, M., 2002. Grounding language in action. Psychonomic Bulletin & Review 9, 558–565.
Goldin-Meadow, S., 2003. The Resilience of Language. Psychology Press, New York.
Goldin-Meadow, S., Beilock, S.L., 2010. Action's influence on thought: the case of gesture. Perspectives on Psychological Science 5, 664.
Goldin-Meadow, S., Mylander, C., Franklin, A., 2007. How children make language out of gesture: morphological structure in gesture systems developed by American and Chinese deaf children. Cognitive Psychology 55, 85–135.
Gottlieb, G., 1992. Individual Development & Evolution: The Genesis of Novel Behavior. Oxford University Press, New York.
Gottlieb, G., 2003. On making behavioral genetics truly developmental. Human Development 46, 337–355.
Hailman, J., 1969. How an instinct is learned. Scientific American 221 (6), 98–106.
Learmonth, A., Nadel, L., Newcombe, N., 2002. Children's use of landmarks: implications for modularity theory. Psychological Science 13, 337–341.

Lehrman, D., 1953. A critique of Konrad Lorenz's theory of instinctive behavior. The Quarterly Review of Biology 28, 337–363.
Lewontin, R., 1978. Adaptation. Scientific American 239, 212–228.
Megarry, T., 1995. Society in Prehistory: The Origins of Human Culture. New York University Press, New York.
Mithen, S., 1996. The Prehistory of the Mind: The Cognitive Origins of Art and Science. Thamnes & Hudson, London.
Odling-Smee, F.J., Laland, K., Feldman, M., 2003. Niche Construction: The Neglected Process in Evolution. Princeton University Press, Princeton, NJ.
Oyama, S., 1985. The Ontogeny of Information: Developmental Systems and Evolution. Cambridge University Press, New York.
Reddy, V., Markova, G., Wallot, S., 2013. Anticipatory adjustments to being picked up in infancy. PLoS One 8, e65289. http://dx.doi.org/10.1371/journal.pone.0065289.
Roux, V., Bril, B. (Eds.), 2005. Stone Knapping: The Necessary Conditions for a Uniquely Hominin Behavior. McDonald Institute for Archaeological Research, University of Cambridge, Cambridge, UK.
Saffran, J., 2003. Statistical language learning: mechanisms and constraints. Current Directions in Psychological Science 12, 110–114.
Sober, E., Wilson, D., 1998. Unto Others: The Evolution and Psychology of Unselfish Behavior. Harvard University Press, Cambridge, MA.
Spelke, E., Kinzler, K., 2009. Core knowledge. Developmental Science 10, 89–96.
Sterelny, K., 2003. Thought in a Hostile World: The Evolution of Human Cognition. Blackwell Publishing, Oxford, UK.
Thelen, E., Smith, L., 1996. A Dynamic Systems Approach to the Development of Cognition and Action. MIT Press, Cambridge, MA.
Tomasello, 2001. The Cultural Origins of Human Cognition. Harvard University Press, Cambridge, MA.
Uttal, D., Miller, D.I., Newcombe, N., 2013. Exploring and enhancing spatial thinking: links to achievement in science, technology, engineering, and mathematics? Current Directions in Psychological Science 22, 367–373.
van Leeuwen, L., Smitsman, A., van Leeuwen, C., 1994. Affordances, perceptual complexity, and the development of tool use. Journal of Experimental Psychology: Human Perception and Performance 20, 174–191.
Wilson, E.O., 2012. The Social Conquest of the Earth. Liveright Publishing, New York.
Wilson, R., 2004. Boundaries of the Mind. Cambridge University Press, New York.

Human Cooperation, Evolution of

Sarah Mathew, Arizona State University, Tempe, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Human cooperation is a novel evolutionary puzzle because we cooperate with genetically unrelated individuals in groups that comprise millions of people. Direct reciprocity, especially when considering errors in behavior, has shed light on pairwise cooperation among well-known but genetically unrelated people. Three mechanisms have been identified to explain large-scale cooperation in humans. Cooperation can evolve by indirect reciprocity in groups as long as people can observe each other's behavior, or can garner honest information via gossip on who cooperates and defects in their interactions. Humans also have a disposition to cooperate and to punish those who do not, even at a cost to oneself, and such costly punishment can sustain cooperation in even larger groups of people who do not know much about each other. Cultural group selection explains the scale of human cooperation, why it is variable, and why norms enforced by sanctions are group-beneficial. Support for these theories has come from laboratory experiments using a variety of behavioral economic games, and from field studies in small-scale societies. Key open questions include understanding what characterizes goodness in indirect reciprocity, why gossip is sufficiently accurate, and why people are motivated to engage in costly punishment.

Introduction

Humans are much more cooperative than any other vertebrate species. Division of labor and delayed exchange of valuable commodities occurs in virtually every human society. The sick and injured are fed and cared for. Public capital facilities like communal shelters, ceremonial grounds, and defenses are commonplace in human societies. Such cooperation is virtually absent in other vertebrates. Humans even help strangers – donate blood, give directions, offer their seat, engage in philanthropy – more than most other vertebrates help individuals they have known their entire life. Language too is a low-cost signaling system, made possible because cooperation ensures that communication is sufficiently honest. Vertebrates that do not cooperate are limited to costly, self-verification of signals.

This high degree of cooperation makes humans an ultra-social species. Ultrasociality is rare in nature because of the detrimental effect of free riders – individuals who take the benefits of cooperation without paying the cost. Nonetheless when ultrasociality does occur, it has spectacular ecological consequences. Two eusocial taxa – ants and termites – make up one-third of the terrestrial animal biomass of the planet. Humans have higher biomass than any terrestrial mammal aside from the domestic cows we use in farming. Thus, if the free rider problem is solved, there are tremendous evolutionary gains to be had from cooperation.

Human ultrasociality is a novel evolutionary puzzle because it is not explained by genetic relatedness. Most ultrasociality in nature occurs among genetically related individuals (Smith and Szathmary, 2000). In multicellular organisms, cells are clones. In eusocial insects, workers tend to be siblings. But humans cooperate in groups comprised of hundreds, thousands, and even millions of genetically unrelated individuals. A growing body of research suggests that our extreme reliance on culture – socially transmitted knowledge, beliefs, and norms – might be the key ingredient that fueled this transition (Richerson and Boyd, 2005; Henrich and Henrich, 2007; Bowles and Gintis, 2011). Piecing together this puzzle has illuminated a diverse range of human phenomena such as our proclivity to punish wrong doers, our attention to fairness, laws and norms, morality and justice, and our penchant for gossip.

How Do We Know That Humans Are Cooperative?

Key insights on human cooperation have come from economic experiments, which have demonstrated that people all over the world violate the assumption that humans have purely self-interested preferences (J. Henrich et al., 2004; Camerer, 2003). In the Ultimatum Game, two anonymous players get to divide a sum of money. The proposer offers a division. If the responder accepts the offer, the two players get their proposed shares. If the responder rejects, they both receive nothing. If people are money-maximizers the lowest possible offer will be made and any offer will be accepted. Another game commonly used is the dictator game. Here the experimenter gives one of the players – the dictator – a sum of money. The dictator can either keep all of it for herself, or give some to the second player – the recipient. If players are money-maximizers, dictators will keep their full endowment. A third game is the public goods game. Here a group of participants are given a sum of money. They may contribute a portion of this endowment to a common pool, which is multiplied by the experimenter and divided equally among players regardless of their personal contribution. The multiplication factor is set so that a player can maximize her income by not contributing anything. In a seminal study Henrich et al. (Henrich et al., 2005) conducted these three experiments across 15 nonindustrialized societies including mobile hunter-gatherers, small-scale horticulturalists, acephalous pastoralists, and agriculturalists. People behaved quite differently in different societies, but in no society was their behavior consistent with money-maximization. Proposers offered more than the minimum possible amount, responders in many societies rejected low offers, dictators transferred money to the recipient, and participants contributed to the common pool.

Research on contemporary small-scale foragers also paints a picture of societies that are remarkably cooperative. Among the Shiwar, a forager-horticulturalist society in South America, more than 50% of adult men had disabilities that kept them from foraging for more than a month (Sugiyama, 2004). In any other vertebrate society, such a lengthy disability would mean death. Human foragers instead regularly survive these periods because others provide for them during such times. An analysis of the Hiwi and Ache, two forager societies in South America suggest that there is extensive food sharing beyond the household (Gurven, 2004). Among the Hiwi, in more than 75% of the foraging events, some amount of food was given to another household. In about 15% of foraging events, 90–100% of the food acquired is given to another household.

The human life history itself would be impossible without intensive cooperation, so much so that humans are considered to be cooperative breeders (Hrdy, 2009; Hill and Hurtado, 2009). Compared to the typical primate, humans invest in higher quality offspring. Human offspring take longer to reach sexual maturity and become self-sufficient foragers only by their late teens, but they have larger brains and longer lifespan. One would expect that such prolonged care of offspring would require longer inter-birth intervals – the usual pattern observed in mammals whereby quality comes at the expense of quantity. But surprisingly humans have higher fertility and shorter inter-birth intervals than chimpanzees, even though chimpanzee juveniles need no care. In a synthetic review of human life history traits, Kaplan et al. (2000) posit that food transfers are at the nexus of our escape from this trade-off. Post-reproductive women and adult males produce more food than they consume. Reproductive women and children produce less than they consume. Food transfers from these net-producers to these net-consumers allow reproductive aged women to support multiple dependent offspring at a time, and allow children to stay dependent for long.

Property rights are omnipresent in human societies of all types, from hunter-gatherers to nation states. There are tremendous benefits to be had if community members do not steal each other's unguarded property. People would not have to constantly guard or fight for their property and can utilize this time in production. Cooperation of this form must have been present among Paleolithic foragers. People would be disinclined to create complex tools and technology that require time and effort now but would pay off later, if they could lose these items to predatory band members. Human males spend far less effort staying in close proximity to their mates for mate-guarding than do males of most other animals. Free of the need to always mate-guard, men can spend their day foraging separately from women, specializing in different foods than women do.

Why Is Cooperation an Evolutionary Puzzle?

If cooperation is beneficial, and organisms that cooperate are immensely ecologically successful, why does evolution seldom favor cooperation? To see why, consider the payoff structures that organisms may encounter in nature that creates opportunities to benefit from cooperation. The payoff structure is the cost and benefit for each of the actors, given their behavior and the behavior of their partner(s). In some forms of cooperation, individuals have no incentive to defect (i.e., to not cooperate) as long as their partner cooperates. Table 1 illustrates such a payoff structure, also referred to as a mutualism or coordination game. Here two foragers have the option of either finding small game on their own, or coordinating with each other to hunt larger game. Assume that if they attempt to hunt the larger prey on their own they will fail. As long as the two foragers can communicate that they are willing to coordinate their foraging attempt, they have no reason to hunt alone. Realizing the benefits of cooperation in this case is easy: mutually agree to cooperate.

Table 1 Mutualistic interaction

		Forager 1	
		Hunt alone	*Hunt together*
Forager 2	Hunt alone	7, 7	7, 0
	Hunt together	0, 7	10, 10

An illustration of mutualistic payoffs where as long as one player cooperates, it is in the best interest of the other player to cooperate.

Table 2 Prisoner's dilemma game

		Player 2 (acquired fish)	
Prisoner's dilemma payoff structure		*Keep fish*	*Share fish*
Player 1 (acquired tubers)	Keep tubers	7, 7	12, 5
	Share tubers	5, 12	10, 10

An illustration of payoffs in a prisoner's dilemma where it is in a player's best interest to defect, regardless of whether the other player cooperates or defects.

Problematically, in many of the opportunities for cooperation in nature, individuals have an incentive to defect even when their partner cooperates. Such payoff structures are best illustrated by the prisoner's dilemma game (Table 2). Here one forager has spent the day fishing and the other acquiring tubers. At the end of the day, they can eat what they have personally acquired or they can share a portion of their day's catch with each other. Assuming that their optimal daily diet is a mix of tuber and fish, their joint payoff is maximized if they both share. But, forager 1 is better off keeping her fish if her partner keeps the tubers, but also if her partner shares the tubers. Likewise for forager 1. Thus both foragers will keep their catch forgoing the big gains they as a pair could have had from cooperating. The *n*-person public goods game illustrates this tension for a group of people: the group maximizes its payoff if everyone contributes to the public good, but each person has a higher payoff if they take the benefit without paying the cost of producing the public good.

Cooperation in a prisoner's dilemma or public goods game presents an evolutionary puzzle. To obtain the benefits of cooperation individuals have to behave in a way that maximizes the pair's or group's payoff, not their own. Evolutionary pathways to cooperation identify mechanisms by which cooperators can end up with higher payoffs (and consequently higher fitness) than defectors despite this disadvantage.

The Evolution of Small-Scale Cooperation

Kin Selection

Kin selection (Hamilton, 1964) explains a wide range of cooperation that occurs in nature, and plays an important role

in human cooperation too. Altruistic behavior can evolve if individuals preferentially direct their behavior toward kin. This is because if a person carries the cooperative allele, her relatives are more likely than a random individual in the population to also possess the cooperative allele. This means that if individuals cooperate only with kin, cooperators will be more likely than defectors to also receive the benefits of cooperation. Such positive assortment allows cooperation to evolve.

Although nepotism is one aspect of human cooperation, it does not fully explain human cooperation. Nepotistic sentiments do exist. The meaning of the word – handing favors unfairly to relatives – illustrates the struggle to suppress this form of cooperation in order to organize cooperation at a larger scale. Food sharing in hunter-gatherers suggests that humans have evolved other pathways to cooperation that may often compete with our nepotistic motivations. A study among reservation-dwelling Ache of Paraguay (Gurven et al., 2001) shows that offspring, siblings and parents get a larger percentage of an individual's share of a resource item. However, close kin outside of the nuclear family like nieces, nephews, aunts, uncles and cousins receive less than what an unrelated band member receives. Another study conducted in the same population (Allen-Arave et al., 2008) show that foragers do not preferentially share with households that have their kin. Instead, foragers give away food to households that need food, and to households that have given them food in the past. Moreover, when there are disadvantages to cooperating with kin, humans are able to manage even high-stakes cooperation with nonkin. Among the Turkana, a nomadic pastoral society in East Africa, warriors venture on stealth missions with two to five other men, in which they plan to take cattle of neighboring ethnic groups. These missions can be risky, and the team needs to engage in a fire-fighting if detected. Although a high degree of cooperation and trust in one's companions is required, such teams are often comprised of age-mates and friends, not brothers. When I have enquired about this to Turkana participants I interviewed, they explain that one of the brothers needs to look after the herds, and that if the raiding party happen to be killed on this mission, a loss of two sons would be catastrophic for a family.

Misfiring of the nepotistic psychology in novel environments also does not satisfactorily explain human cooperation. Based on the high degree of cooperation in foraging societies, evolutionary social scientists have often assumed that humans evolved in foraging bands comprised of kin (e.g., Maner and Kenrick, 2010). If humans had evolved in kin-based bands, then the misfiring (Barkow et al., 1992) of the nepotistic psychology in present-day environments can account for human ultrasociality. However, this is an unlikely scenario. First, there is no evidence that Paleolithic foragers lived in kin-based bands. Data from two contemporary hunter-gather societies, the Ju/Hoansi and Ache, show that only up to a quarter of the band is comprised of a person's kin (Hill et al., 2011). The other half of the band is affines, and the remaining quarter is neither kin nor affines. Second, in acephalous nomadic pastoral societies like the Turkana, multiple settlements cooperate in warfare (Mathew and Boyd, 2011). Turkana settlements disperse and regroup seasonally like hunter-gather bands. With a band-level cooperative psychology, we should expect the Turkana to cooperate on the scale of settlements. Yet, an average Turkana raid has warriors from four different settlements. Lastly, people in societies with higher market integration and larger community size are more generous in the dictator game and more willing to punish unfair offers in the ultimatum game when playing with anonymous community members (Henrich et al., 2010). The misfiring hypothesis would suggest that people living in small-scale societies that are more similar to ancestral social environments should cooperate in anonymous interactions with their peers at least as much as people in large market-based societies.

Direct Reciprocity

Direct reciprocity (Trivers, 1971) can sustain cooperation among pairs of individuals who have the opportunity to interact repeatedly. Direct reciprocity is best exemplified by the *Tit for Tat* (*TFT*) strategy, where an individual cooperates in the first round in a repeated prisoner's dilemma game, and afterward acts just as her partner acted in the previous round (Axelrod and Hamilton, 1981). Reciprocal strategies like *TFT* can outcompete defectors because its practitioners are more likely to accrue the long-run benefits of repeated cooperation than defectors who attempt to exploit them in the first round but lose out afterward. Although there are limited examples of reciprocity in other animals, some forms of human cooperation fit with reciprocity. For instance, among the Hiwi and Ache, the amount of one family's food production given to another family correlates with the amount of the second family's food production given to the first family, controlling for biological kinship, geographic proximity, and the overall amount of this resource the family gives away (Gurven, 2004). Consistent with reciprocity models, humans have 'cheater detection' systems for social exchange – specialized cognitive systems to recognize when a partner has violated a social contract (Cosmides et al., 2010).

Although *TFT* was shown to be the winning strategy against a host of other strategies (Axelrod and Hamilton, 1981), reciprocity in nature does not follow its prediction too closely. The condition for *TFT* to be an evolutionarily stable strategy against a pure defector strategy is $wb > c$, where b is the benefit the recipient gets, c is the cost for the giver, and w is the probability the pair will interact again. According to this condition if a pair will interact twice on average, they could be as cooperative as siblings. If so, there should be rampant high-stakes pairwise cooperation in humans and other animals. Yet, reciprocity has rarely been documented in other animals, although repeated interactions occur in many animal societies (Clutton-Brock, 2009). Humans display conditional cooperation, whereby we are more likely to cooperate with partners who cooperate, but the strategy we utilize pays attention to much more than what occurred in the previous interaction. For instance, people pay attention to their partner's intent, not their behavior (Falk et al., 2008). Friendship is an appropriate domain for reciprocity – it is a long-term relation between two unrelated individuals characterized by mutual aid and companionship. Yet, a study conducted in USA, Japan, and China indicate that a characteristic of friendship is that people do not track each other carefully and tolerate imbalances in help received and given (Xue and Silk, 2012).

Why is there a mismatch between reciprocity theory and real-world pairwise cooperation? One reason is that in reciprocity models which strategy prevails depends on what mix of strategies

are allowed to compete. This ambiguity prompted a recent study to examine what occurs when all possible strategies are allowed (van Veelen et al., 2012). The analysis by van Veelen et al. reveals that even when there are considerable gains to cooperation, there is only moderate cooperation in the population. The surprising pattern is caused by strategies that arise via mutation that can cooperate and thrive just as well as the common strategy, but once common, pave the way for less cooperative strategies. The population thus cycles between states that are dominated by more and less cooperative strategies. Interestingly, the authors find that if there is a small degree of assortment, repeated interactions yield much higher levels of cooperation. In humans, such assortment can arise if interactions are directed toward kin. Importantly, humans have yet another pathway for assorting. Because we acquire information through social learning, and learn what the local norms are about how to behave from our local group, interacting with cultural group members can also lead to assortment (Boyd and Richerson, 2010).

Another reason why real-world pairwise cooperation is unlike *Tit for Tat* is because *TFT* is a poor strategy when players make mistakes (Fudenberg and Maskin, 1990), fueling a quest for a variety of alternative strategies (reviewed in McElreath and Boyd, 2008; Nowak, 2006). Real-world interactions are plagued by genuine mistakes. People intend to cooperate, but may defect instead. For example, Sally agrees to pick Tom from the airport, but wrote the date incorrectly in her calendar and does not show up when Tom arrives. Such errors are called implementation errors. In another kind of error, people cooperate, but their partner may perceive it as a defection. For example, Sally arrives at the airport half an hour late because of traffic. Tom waited for 20 min and then took a taxi home. These are labeled perception errors. Resolving these errors have led to a diverse range of strategies that are not just retributive like *Tit for Tat*, but are additionally contrite – sorry for mistakes made and willing to make amends (Boyd, 1989), forgiving – occasionally cooperating with those who defect (Nowak and Sigmund, 1992), and flexible – shifting from cooperation to defection and vice versa based on remuneration under these tactics (Nowak and Sigmund, 1993). More investigation on this topic could illuminate further the emotions and behavioral strategies that underpin pairwise social relations in humans.

The Evolution of Large-Scale Cooperation

Human cooperation cannot be explained by nepotism and direct reciprocity alone. Humans cooperate with strangers, and cooperate in groups comprising millions of genetically unrelated individuals. In state societies, people regularly trade with anonymous strangers, and through taxes, contribute toward public goods like defense and capital facilities, and provide for individuals who are needy. Even without coercion from the state, people voluntarily cooperate: they donate blood, contribute to charity, assist in disaster relief, organize community events, protest unjust laws, give up their seat on a bus for an elderly person, return items to lost-and-found locations, give directions to travellers, etc.

Even politically uncentralized societies can organize cooperation at an impressively large scale. Among the Turkana, an acephalous tribal-scale pastoral society in East Africa, in addition to the clandestine stealth raids described earlier, men also mobilize for large-scale raids in which they engage in a firefight with neighboring ethnic groups (Mathew and Boyd, 2011). Such raids are a high-stakes form of cooperation – 1% of participants get killed in these raids, and 20% of all males born in the study population die in warfare. Although individuals take great risks in combat to overcome the opponent's resistance, all combatants can acquire the loot. Some gains like territory and deterrence are obtained even by noncombatants. Yet, the scale of Turkana raids is large. An average raiding party comprised of 315 warriors, who came from five different age groups, four different settlements, and three different subterritories of Turkana society. These men do not all know each other. Moreover, a warrior has on average only four close kin with him in such a raiding party.

Three mechanisms have been identified to explain the evolution of large-scale cooperation in humans: indirect reciprocity, punishment, and cultural group selection.

Indirect Reciprocity

Humans not only help those who helped them, but also help those who helped *others*. Alexander (Alexander, 1987) posited that such indirect reciprocity could explain society-wide cooperation enforced by moral systems that are based on how people treat others. Nowak and Sigmund (Nowak and Sigmund, 1998) theoretically demonstrated that indirect reciprocity can be evolutionarily stable. In this model, a pair of individuals is sampled at random from the social group. One of them is assigned to be the donor, and the other the recipient. The donor can cooperate by conferring a benefit b at cost c to the recipient, or the donor can defect. In the next round, another pair is sampled at random, and the game continues. Each act of cooperation increases a person's 'image score' while defection lowers this score. It turns out that a 'discriminator' strategy, which cooperates if its partner has a sufficiently high image score, can be evolutionarily stable against defectors if $qb > c$. Here q is the probability that the donor knows the image score of the recipient. Image-scoring was found to perform well only under a narrow set of conditions, and so has been replaced by 'standing' strategies which attend also to information on the donor's own score and to whether the recipient's past defections were justified (Panchanathan and Boyd, 2003; Leimar and Hammerstein, 2001). Nonetheless, the general consensus from the models is that cooperation can evolve by indirect reciprocity in communities where people can observe each other's behavior, or can garner honest information via gossip on who cooperates and defects in their interactions.

Key features of human cooperation are consistent with indirect reciprocity. Laboratory experiments have demonstrated that participants are motivated to help a novel partner who has helped other players. For instance, in one study (Ule et al., 2009) participants were paired at random with another participant each round. In one of the treatments the player in the donor role could help, do nothing, or harm one's partner. The majority of participants (57%) were other-regarding discriminators who helped if their partner had a reputation for helping others, and passed up on the opportunity to help if their partner did not have a history of helping others. A related observation is that humans are more cooperative when there

are cues that they are being observed, suggesting that reputation matters. In an anonymous one-shot experimental study, participants were more generous to a stranger merely when stylized eye spots were drawn on the computer screen in which they took the study (Haley and Fessler, 2005). Nonverbal cues like mutual eye-gaze and touching increased contributions in a public-goods game (Kurzban, 2001). These subtle cues that we are being watched increase cooperation beyond the laboratory too (Bateson et al., 2006). Concern for reputation fits indirect reciprocity. However, individuals who strategically cooperate only when their own reputation is at stake resemble the self-interested discriminator strategy in the study by Ule et al. where 10% of players helped their partners when their own reputation was poor rather than when their partner had a good reputation. More theoretical work is required to understand the prevalence of self-interested discriminators and why they do not undermine cooperation sustained by other-regarding discriminators.

The above models of indirect reciprocity only explain pairwise helping among members of a social group. Can indirect reciprocity sustain large-scale cooperation? To answer this, one study examined players' behavior when rounds of the indirect reciprocity game alternated with rounds in which the group played a public goods game (Milinski et al., 2002). Players were much more cooperative in this game than if they were playing consecutive rounds of the public-goods games. Thus, the reputation consequence in the pairwise game was motivating players to behave well in the *n*-person game too. Panchanathan and Boyd (2004) theoretically demonstrated that indirect reciprocity can indeed support large-scale cooperation. If individuals in a social group engage in both *n*-person and pairwise cooperation, and if those who fail to contribute in the public-goods stage are denied help in pairwise exchange, then contributing to public goods is evolutionarily stable. Attitudes toward Turkana warriors who shirk on the battlefield are consistent with indirect reciprocity's role in sustaining *n*-person cooperation (Mathew and Boyd, 2014). In vignette studies, Turkana subjects were asked either about a fictitious coward who lagged behind others on a raid, or about an unskilled warrior who put in the effort but did not perform well. Participants were more likely to say that they would not help the coward if he needed to borrow an animal from them at a time of need. Being denied such help is consequential for a Turkana herder. Herdsmen lose their stock in raids, epidemics and droughts, and without aid from their relatives and community they cannot rebuild their herds. The threat of losing such assistance must motivate warriors to not shirk on the battlefield.

Indirect reciprocity has illuminated why humans mind other's business, and how this enables cooperation, but there remain three open questions. One question is why indirect reciprocity is even needed to sustain pairwise cooperation, when direct reciprocity could suffice. Second, it is clear that cooperation should be somehow contingent on the recipient having a good standing, but what is the best system of assigning 'goodness'? For instance, if I am in good standing and I am kind to a defector, should it lower, increase, or have no effect on my standing? Should my act of kindness be viewed differently if I myself was in bad standing? To tackle this question Ohtsuki and Iwasa (2004) examined a vast range of assignment systems and found eight leading strategies. All eight of these strategies share the property that cooperating with good folks is good and defecting on good folks is bad. Moreover, in terms of all leading eight strategies it is good to defect on bad folks. Empirical studies on how humans actually adjust their impressions of others can help narrow the search further. A third issue is that indirect reciprocity models assume that information quality is sufficiently high because mechanisms like gossip allow reputational information to be broadcast. But why are people motivated to tattle tale when they personally do not stand to gain from it? And why do people not manipulate information for their own gains enough to undermine the quality of reputational information and thereby unravel cooperation? A better grasp of these issues will provide a more accurate picture of how indirect reciprocity evolved, and will get us closer to describing the structure and nuances of our moral psychology.

Punishment

Cooperation can be maintained if individuals cooperate, and at some cost to themselves, punish those who defect. Such altruistic punishment can be evolutionarily stable against defectors as long as punishment imposes a larger cost on defectors than the benefit defectors acquire by free riding (Boyd et al., 2003). An advantage of punishment is that it can maintain cooperation in much larger groups than indirect reciprocity because people do not need to know the previous behavior of everyone else in the group. For instance, indirect reciprocity alone is unlikely to maintain cooperation on Turkana raids, as raiding parties comprise a few hundred men who are drawn from different segments of society and do not know each other well. This may be why the Turkana occasionally also dispense corporal punishment to cowards and deserters in raids. A problem with punishment is that it can be undermined by second-order free riders. In indirect reciprocity, the discriminator strategy benefits from refusing to help someone in bad standing, and so are evolutionarily stable against unconditional cooperators (Panchanathan and Boyd, 2004). In contrast, because punishment is costly, unconditional cooperators or second-order free riders fare better than punishers by gaining the benefit of punishment without paying the cost (Yamagishi, 1986).

Empirically, humans are motivated to cooperate and to punish those who do not, even at a cost to themselves (Gintis et al., 2005). In a classic experiment Fehr and Gachter (2002) compared players' contribution in multiple rounds of a public goods game when players could either punish or not punish shirkers in their group. When punishment was allowed, players used it even though it cost them – 84% of players punished at least once – and consequently contribution levels increased each round. In contrast, contributions decreased each round in a typical public goods experiment without punishment because conditional cooperators who start out contributing stop doing so when they notice free riders in their group. Remarkably, experiments also reveal that the motivation to engage in costly punishment extends to uninvolved third-party observers. In the third-party punishment game, an allocator receives an endowment and can allocate any amount of it to the recipient. After the allocation, a third person – the observer – can reduce her own endowment to subtract money from the allocator. Henrich et al. conducted this experiment across 15

small-scale societies (Henrich et al., 2006). In every society, there were observers who would punish unfair allocators, even though such punishment is costly to the observer and provides the observer no personal benefit.

There is evidence that punishment supports cooperation outside of laboratory experiments too. Since their inception, state societies have utilized punishment to curb norm violators and criminals, and enforce rule of law. However, given that punishment is costly to implement, it is important to know if punishment occurs in politically uncentralized societies where there are no police and courts responsible for implementing punishment. Here the Turkana provide a useful case study. Turkana warriors report that they do indeed punish cowards and deserters in raiding parties (Mathew and Boyd, 2011). A group of age-mates find the unsuspecting violator, restrain him by holding or tying him, and beat or whip him. All the while, they reprimand him for what he did, teach him how he should have behaved, and urge him to not repeat the offense in future raids. After this, the violator may be required to kill an animal from his herd for the others to eat. The Turkana view such punishment as an intervention to correct men who have erred, rather than as an opportunity to injure a person or dispense violence (Mathew and Boyd, 2014). Age-mates and other community members discuss the matter beforehand during which differing opinions may be heard about whether punishment is warranted. Some violators are only criticized verbally and warned that if they repeat it, they will face more serious sanctions.

Punishment is generally accepted to play some role in large-scale cooperation, but there are ongoing debates about how important its role is. One point of contention is how punishment compares to rewards and other forms of positive reinforcement. Punishment may not be efficient from the group's perspective (Ohtsuki et al., 2009). Because costly punishment lowers the payoff of both the punisher and punished, in some experiments it yields lower average payoffs for the group (Dreber et al., 2008). When both the option for rewarding cooperators and punishing defectors is allowed, rewarding increased contribution levels more than punishment in the repeated public goods game (Rand et al., 2009). Other experiments contradict these findings. When the time horizon is longer than in typical laboratory experiments, punishment increased average group payoffs (Gächter et al., 2008). In the study by Ule et al. (2009) where individuals had the opportunity to either help, not help, or punish a partner, players used indirect sanctions much more often than direct punishment. However, the actions of the 10% of players who were other-regarding costly punishers were necessary to make defectors have a lower payoff than other-regarding discriminators.

Another issue is whether there is sufficient empirical evidence for costly punishment outside of laboratory experiments (Guala, 2012). Some scholars contend that in the real world punishment is typically implemented through low-cost means. For instance, the Turkana first criticize cowards and deserters, and utilize corporal sanctions when such verbal sanctions do not suffice. Moreover, a group of age-mates convene to punish a violator. Such coordinated punishment is much cheaper to implement and can evolve under a less restrictive range of conditions (Boyd et al., 2010). While a solo punisher may face physical harm and retaliation, age-mates who deliver punishment in a group escape these costs. Nonetheless, in addition to the chance of potentially losing a valuable social tie, organizing corporal punishment takes time and effort. Age mates discuss the violation, gather local age-mates, plan the punishment, find the unsuspecting violator, and discipline him. These costs, albeit lower than what is faced by a solo punisher, matter when considering the evolution of a trait. So, the question of why people are motivated to bear costs to punish free riders remains a crucial open problem.

Although some theoretical solutions to the second-order free rider problem have been proposed (e.g., Boyd et al., 2003; Henrich and Boyd, 2001), more work is needed to clarify the issue. There are virtually no empirical studies of second-order free riding in small-scale societies, which raises many questions. Do societies have cultural norms against second-order free riding? Are people motivated to sanction second-order free riders, or are we motivated to reward punishers? Is third- or fourth-order free-riding a concept in any society? Thus it remains to be seen how humans solved the critical problem of second-order free riding and evolved the capacity for punishment.

Cultural Group Selection

Indirect reciprocity and punishment have explained important aspects of human cooperation and moral motivations, but they are not complete explanations. Sanctions have the unfortunate property that they can sustain a vast range of outcomes (Boyd and Richerson, 1992). Sanctions can maintain both cooperative and noncooperative outcomes. For instance mutual enforcement can maintain the norm 'You may steal your neighbor's unguarded cows to feed your family' or 'You may not steal your neighbor's unguarded cows to feed your family'. Sanctions can also maintain norms on a wide range of scales. For example, 'Do not steal your clansmen's cattle, but the cattle of the rest are for brave men to steal' or 'Do not steal your tribesmen's cattle, but the cattle of other tribes are for brave men to steal' are both group-beneficial norms, but one benefits clans, the other benefits tribes. In the domain of property rights, cooperation occurs on the scale of clans in the former society, and on the much larger scale of tribes in the latter society.

Experimental findings too confirm what casual observations tell: human societies vary widely in the norms they enforce, and these norms support cooperation at vastly different scales. Herrmann et al. (Herrmann et al., 2008) conducted a public goods game with punishment in 16 countries, and found that participants of some countries punished individuals who contributed to the public goods, while in other countries people punished noncontributors. Antisocial punishment was more common in countries with weak rule of law and low civic cooperation. Another study comparing behavior across 15 small-scale societies (Henrich et al., 2010) showed that people in societies with higher market integration and large community size are more generous in the dictator game and more willing to punish unfair offers in the ultimatum game. This suggests that large, market-based societies have norms of fairness in transient interactions that support cooperation with anonymous strangers, whereas small-scale societies lack these norms.

A complete account of human cooperation would specify the processes that select among the many possible norms that

sanctions can sustain to yield the patterns of cooperation we observe across societies.

Cultural group selection (Boyd and Richerson, 1985) can explain why sanctions and norms support cooperation on the scale of cultural groups – symbolically marked groups of people with similar norms and beliefs. Social learning can maintain between-group variation in behavior despite substantial migration, making cultural group selection much more plausible than genetic group selection (Henrich and Boyd, 1998; Boyd and Richerson, 2010). Bell et al. (2009) tested this claim by estimating between-group variation in cultural and genetic traits among contemporary societies. The analysis shows that cultural FST values are orders of magnitude larger than genetic FST values, supporting the view that cultural group selection is far more plausible than genetic group selection. How do social learning mechanisms maintain sufficient between-group behavioral variation? Conformism – imitating the behavior of the majority – is one type of social learning mechanism that can evolve in a wide range of conditions that favor social learning (Boyd and Richerson, 1985; Henrich and Boyd, 1998). If people have conformist learning biases, then migrants into a group will adopt the behavioral variant of the new group, because they adopt the common variant in the new environment. A second way by which social learning biases can maintain between-group variation is by punishment (Richerson et al., 2003). If norm violations are punished, then new immigrants into a group will quickly learn to adopt the customs of the new social environment. As long as there is sufficient between-group variation, there are ample means by which cultural groups can compete: warfare, differential population growth, and greater success at being imitated by other groups. Consequently norms that benefit the cultural group can evolve.

There is some evidence that the scale of human cooperation is consistent with cultural group selection. Consider the following study among the Turkana (Mathew and Boyd, 2011). The Turkana are subdivided into two-dozen territorial sections each comprising 20 000–50 000 people. Participants from one territorial section – the Kwatela – were narrated one of two scenarios. In one scenario, two Kwatela warriors go to the neighboring Turkana territorial section – the Lukumong. They steal cows, and drive them back to Kwatela land. In the other scenario, two Kwatela warriors go to the neighboring tribe – the Toposa – and steal cows, which they drive back to Kwatela land. Participants, who were all Kwatela, were then asked a series of questions to elicit how they judged the warriors. Strikingly, although in both scenarios the Kwatela benefit, participants reacted extremely differently. They felt that the warriors who raided Lukumong cattle did something very wrong, deserved sanctions, and were undeserving of their help. In contrast, the warriors who raided Toposa cattle were praiseworthy and deserving of help. This means that Turkana norms governing warfare create benefits exactly at the scale of the cultural group – not at the smaller scale of territorial sections, and not at the larger scale that includes other tribes. The pattern of male mortality in Turkana suggests that the norms really matter. About half of adult male mortality is due to interethnic warfare, and only 1% is due to internal interpersonal violence. Thus, despite the ubiquity of firearms within Turkana settlements, the Turkana successfully suppress interpersonal violent conflict, and most of the conflict is between the Turkana and other tribes. If this pattern is general, there is a compelling case to be made that the ultrasocial unit in human societies is the cultural group. As cultural groups get bigger, so then will the scale at which humans cooperate.

Conclusion

For a long time, evolutionary theory did not explain human behavior as adequately as it explained the behavior of other animals. But recent models of cooperation have begun to integrate distinctive features of humans such as culture and moral sentiments into the evolutionary framework. Direct reciprocity, and the range of reciprocal strategies that deal with errors, have provided a useful framework for understanding pairwise relations in humans. Indirect reciprocity and punishment models have illuminated how humans evolved to cooperate in large groups of unrelated individuals. Finally, cultural group selection explains the scale of human cooperation, and why it is variable. There remain many unresolved questions: What reciprocal strategy underpins friendships? What is goodness? Why is gossip honest enough? Why are people motivated to punish? More theoretical work and empirical studies across a wide range of societies can help fill these gaps, and provide us a more detailed picture of how we became ultrasocial.

See also: Altruism and Prosocial Behavior, Sociology of; Behavioral Economics; Charity and Philanthropy, Economics of; Collective Behavior, Sociology of; Common Law; Conflict and War: Anthropological Aspects; Conventions and Norms: Philosophical Aspects; Cooperative Game Theory; Cross-Cultural Research Methods in Sociology; Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Decision and Choice: Bounded Rationality; Deterrence Theory: Crime; Deterrence: Sanction Perceptions; Deterrent Effect of Police and Prisons; Economic Anthropology; Egalitarianism; Emergent Properties; Empathy: A Social Neuroscience Perspective; Equity in Law; Ethnic Conflicts; Ethnocentrism; Evolutionary Approaches to Human Behavior: Philosophical Aspects; Evolutionary Economics; Experimental Economics; Fertility Theory: Embodied-Capital Theory of Life History Evolution; Fertility Theory: Theory of Life History Evolution; Field Experiments; Field Observational Research in Anthropology and Sociology; Friendship, Anthropology of; Game Theory: Noncooperative Games; Game Theory; Gene–Culture Coevolution; Human Behavioral Ecology; Human Cognition, Evolution of; Hunting and Gathering Societies: Anthropology; Institutions; Intentionality in Language and Communication, Emergence of; Justice, Conflicts, and the Justice of Conflict Resolution; Kin Selection; Kinship, Evolution of; Kinship, Formal Models of; Law and Economics; Law and Morality: An Analytical Perspective; Law: Anthropological Aspects; Market and Nonmarket Allocation; Mind, Theories of; Morality 'East' and 'West': Cultural Concerns; Morality: Evolution of; Natural Law; Neural Foundation of Morality; Norms; Philosophy of Law; Property Rights, History of; Public Goods: International; Punishment: Comparative Politics of; Punishment: Social and Legal Aspects; Punishment; Quantitative Cross-national Research Methods; Religion:

Morality and Social Control; Rule of Law (and Rechtsstaat); Sanctions in Political Science; Social Networks; Sociality In Anthropology; Sustainability and Sustainability Science; Tradition: Social; Trust: Philosophical Aspects; Welfare: Philosophical Aspects.

Bibliography

Alexander, R.D., 1987. The Biology of Moral Systems. Aldine de Gruyter.

Allen-Arave, W., Gurven, M., Hill, K., 2008. Reciprocal altruism, rather than kin selection, maintains nepotistic food transfers on an Ache reservation. Evolution and Human Behavior 29, 305–318.

Axelrod, R., Hamilton, W.D., 1981. The evolution of cooperation. Science 211 (4489), 1390–1396.

Barkow, J.H., Tooby, J., Cosmides, L. (Eds.), 1992. The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, Oxford.

Bateson, M., Nettle, D., Roberts, G., 2006. Cues of being watched enhance cooperation in a real-world setting. Biology Letters 2 (3), 412–414.

Bell, A.V., Richerson, P.J., McElreath, R., 2009. Culture rather than genes provides greater scope for the evolution of large-scale human prosociality. Proceedings of the National Academy of Sciences 106 (42), 17671–17674.

Bowles, S., Gintis, H., 2011. A Cooperative Species: Human Reciprocity and Its Evolution. Princeton University Press, Princeton, NJ.

Boyd, R., 1989. Mistakes allow evolutionary stability in the repeated prisoners-dilemma game. Journal of Theoretical Biology 136 (1), 47–56.

Boyd, R., et al., 2003. The evolution of altruistic punishment. Proceedings of the National Academy of Sciences 100 (6), 3531–3535.

Boyd, R., Gintis, H., Bowles, S., 2010. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. Science 328, 617–620.

Boyd, R., Richerson, P., 2010. Transmission coupling mechanisms: cultural group selection. Philosophical Transactions of the Royal Society (B) 365 (1559), 3787–3795.

Boyd, R., Richerson, P.J., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago, IL.

Boyd, R., Richerson, P.J., 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethology and Sociobiology 13 (3), 171–195.

Camerer, C.F., 2003. Behavioral Game Theory: Experiments on Strategic Interaction. Princeton University Press, Princeton, NJ.

Clutton-Brock, T., 2009. Cooperation between non-kin in animal societies. Nature 462, 51–57.

Cosmides, L., Barrett, H.C., Tooby, J., 2010. Adaptive specializations, social exchange, and the evolution of human intelligence. Proceedings of the National Academy of Sciences 107, 9007–9014.

Dreber, A., et al., 2008. Winners don't punish. Nature 452, 348–351.

Ein-Dor, T., et al., 2010. The Attachment Paradox: How can so many of us (the insecure ones) have no adaptive advantages? Perspectives on Psychological Science 5 (2), 123–141.

Falk, A., Fehr, E., Fischbacher, U., 2008. Testing theories of fairness – Intentions matter. Games and Economic Behavior 62, 287–303.

Fehr, E., Gachter, S., 2002. Altruistic punishment in humans. Nature 415, 137–140.

Fudenberg, D., Maskin, E., 1990. Evolution and cooperation in noisy repeated games. The American Economic Review 80 (2), 274–279.

Gächter, S., Renner, E., Sefton, M., 2008. The long-run benefits of punishment science. Science 322 (5907), 1510.

Gintis, H., et al. (Eds.), 2005. Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life. MIT Press, Cambridge, MA.

Guala, F., 2012. Reciprocity: Weak or Strong? What punishment experiments do (and do not) demonstrate. Behavioral and Brain Sciences 35 (1), 1–15.

Gurven, M., 2004. Reciprocal altruism and food sharing decisions among Hiwi and Ache hunter–gatherers. Behavioral Ecology and Sociobiology 56, 366–380.

Gurven, M., et al., 2001. Reservation food sharing among the Ache of Paraguay. Human Nature 12 (4), 273–297.

Haley, H., Fessler, D., 2005. Nobody's watching? Subtle cues affect generosity in an anonymous economic game. Evolution and Human Behavior 26, 245–256.

Hamilton, W.D., 1964. The genetical evolution of social behaviour. Journal of Theoretical Biology 7 (1), 17–52.

Henrich, J., et al., 2005. "Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. Behavioral and Brain Sciences 28, 795–815.

Henrich, J., et al., 2006. Costly punishment across human societies. Science 312 (5781), 1767–1770.

Henrich, J., et al. (Eds.), 2004. Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-scale Societies. Oxford University Press, Oxford.

Henrich, J., et al., 2010. Markets, religion, community size, and the evolution of fairness and punishment. Science 327, 1480–1484.

Henrich, J., Boyd, R., 1998. The evolution of conformist transmission and the emergence of between-group differences. Evolution and Human Behavior 19, 215–241.

Henrich, J., Boyd, R., 2001. Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. Journal of Theoretical Biology 208, 79–89.

Henrich, J., Henrich, N., 2007. Why Humans Cooperate: A Cultural and Evolutionary Explanation. Oxford University Press, Oxford.

Herrmann, B., Thoni, C., Gachter, S., 2008. Antisocial punishment across societies. Science 319, 1362–1367.

Hill, K., et al., 2011. Co-residence patterns in hunter-gatherer societies show unique human social structure. Science 331 (6022), 1286–1289.

Hill, K., Hurtado, M., 2009. Cooperative breeding in South American hunter-gatherers. Proceedings of the Royal Society of London B 276, 3863–3870.

Hrdy, S., 2009. Mothers and Others: The Evolutionary Origins of Mutual Understanding. Harvard University Press, Cambridge, MA.

Kaplan, H., et al., 2000. A theory of human life history evolution: diet, intelligence, and longevity. Evolutionary Anthropology 9 (4), 156–185.

Kurzban, R., 2001. The social psychophysics of cooperation: nonverbal communication in a public-goods game. Journal of Nonverbal Behavior 25, 241–259.

Leimar, O., Hammerstein, P., 2001. The evolution of cooperation through indirect reciprocity. Proceedings of the Royal Society of London B 268, 745–753.

Maner, J.K., Kenrick, D.T., 2010. When Adaptations Go Awry: Functional and Dysfunctional aspects of social Anxiety. Social Issues and Policy Review 4 (1), 111–142.

Mathew, S., Boyd, R., 2011. Punishment sustains large-scale cooperation in prestate warfare. Proceedings of the National Academy of Sciences 108 (28), 1091–6490.

Mathew, S., Boyd, R., 2014. The cost of cowardice: punitive sentiments towards free riders in Turkana raids. Evolution and Human Behavior 35 (1), 58–64.

McElreath, R., Boyd, R., 2008. Mathematical Models of Social Evolution: A Guide for the Perplexed. University of Chicago Press, Chicago, IL.

Milinski, M., Semmann, D., Krambeck, H.J., 2002. Reputation helps solve the 'tragedy of the commons'. Nature 415, 424–426.

Nowak, M., 2006. Five rules for the evolution of cooperation. Science 314, 1560–1563.

Nowak, M., Sigmund, K., 1992. Tit for tat in heterogeneous populations. Nature 355 (6357), 250–253.

Nowak, M.A., Sigmund, K., 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in Prisoner's Dilemma. Nature 364 (6432), 56–58.

Nowak, M.A., Sigmund, K., 1998. Evolution of indirect reciprocity by image scoring. Nature 393, 673–676.

Ohtsuki, H., Iwasa, Y., 2004. How should we define goodness?—reputation dynamics in indirect reciprocity. Journal of Theoretical Biology 231 (1), 107–120.

Ohtsuki, H., Iwasa, Y., Nowak, M.A., 2009. Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. Nature 457 (7225), 79–82.

Panchanathan, K., Boyd, R., 2003. A tale of two defectors: the importance of standing for evolution of indirect reciprocity. Journal of Theoretical Biology 224 (1), 115–126.

Panchanathan, K., Boyd, R., 2004. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. Nature 432 (7016), 499–502.

Rand, D.G., et al., 2009. Positive interactions promote public cooperation. Science 325, 1272–1275.

Richerson, P.J., Boyd, R., 2005. Not by Genes Alone: How Culture Transformed Human Evolution. The University of Chicago Press, Chicago, IL.

Richerson, P.J., Boyd, R., Henrich, J., 2003. Cultural evolution of human cooperation. In: Hammerstein, P. (Ed.), Genetic and Cultural Evolution of Cooperation. MIT Press, Cambridge, MA.

Smith, J., Szathmary, E., 2000. The Origins of Life: From the Birth of Life to the Origin of Language. Oxford University Press, Oxford.

Sugiyama, L.S., 2004. Illness, injury, and disability among Shiwiar forager-horticulturalists: implications of health-risk buffering for the evolution of human life history. American Journal of Physical Anthropology 123 (4), 371–389.

Trivers, R.L., 1971. The evolution of reciprocal altruism. The Quarterly Review of Biology 46 (1), 35–57.

Ule, A., et al., 2009. Indirect punishment and generosity toward strangers. Science 326 (5960), 1701–1704.

van Veelen, M.J., et al., 2012. Direct reciprocity in structured populations. Proceedings of the National Academy of Sciences 109 (25), 9929–9934.

Xue, M., Silk, J., 2012. The role of tracking and tolerance in relationship among friends. Evolution and Human Behavior 33 (1), 17–25.

Yamagishi, T., 1986. The provisioning of sanctioning as a public good. Journal of Personality and Social Psychology 51, 100–116.

Human Development and Health

Linda K George, Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

There are strong reciprocal relationships between health and human development. Developmental disabilities and delays *in utero* and in childhood are associated with compromised health at all later stages of life, including middle and old age. Conversely, physical or mental illness early in the life course stunts normal development during childhood/adolescence and is associated with disadvantaged socioeconomic status throughout adulthood.

There are strong reciprocal links between health and human development. Throughout the life course, good health provides opportunities for cognitive, emotional, and social development and poor health constrains development. Conversely, successful human development promotes health and developmental difficulties pose risks to optimal health. Although these conclusions are – and can only be – based on observational data, accumulated research findings are sufficiently consistent and strong to permit the inference that these reciprocal effects are causal.

Despite abundant evidence that health and human development operate reciprocally on each other, methodological challenges to research on this topic are substantial. Current findings typically are based on short-term longitudinal data rather than the decades of monitoring required for a fine-grained understanding of the dynamics of health and human development. In addition, research on the links between health and human development during middle-age and late life is threatened by selective mortality – i.e., by the fact that poor health and developmental failures or delays result in early death, leaving an unrepresentatively healthy and developmentally advanced population available for study.

Conceptualization and Definitions

Both 'health' and 'human development' are broad terms, covering a lot of conceptual and empirical ground. Both are defined quite broadly here. Health encompasses the broad range of indicators of physical and mental illness, as well as disability and mortality. It also is important to note that health is strongly age related. Chronic physical illnesses, disability, biological dysregulation (e.g., increased inflammation and decreased immune function), and mortality all increase with age. Mental illness has a different relationship with age, peaking in adolescence, and early adulthood.

Human development is conceptually very broad, although most research on human development can be usefully described as falling into two traditions. One tradition focuses on infancy and childhood and broadly traces the unfolding of universal characteristics and capacities as individuals traverse the early part of the life course. It is generally assumed that developmental processes and outcomes are characteristic of the species. Thus, variability is expected to be modest, although both personal experiences and broader social and cultural contexts can affect the nature and timing of developmental phenomena to a limited degree. This tradition also focuses on developmental milestones and posits that development is cumulative – i.e., that failure to attain one developmental milestone precludes the achievement of later developmental challenges.

The second tradition focuses on adulthood and aging rather than early life and was developed by scholars working from life span (the term preferred by psychologists) and life course (the term preferred by sociologists) perspectives. In this tradition, 'development' is viewed much differently. In particular, there is no expectation of universal developmental tasks. Instead, the focus is on patterns of skills, attitudes, and behaviors that emerge over the life course. These patterns both have implications for subsequent health and are, in part, a function of health. A major theme of research in this tradition is that both developmental pathways are strongly affected by social context – that social and cultural forces are the primary determinants of human behavior and that these forces will exert strong influences on the nature and timing of even those behaviors that are widely characteristic of the species. Because of dramatic differences in social systems and cultural traditions, wide individual and group variations in patterns of behavior over the course of adulthood are expected.

The Link between Health and Human Development

Scholars studying developmental processes first became interested in illness as a threat to normal development. Serious illnesses and/or disabilities can constrain an individual's ability to master developmental tasks and/or delay his or her achievement (e.g., Newacheck and Taylor, 1992; Verhoof et al., 2012). Moreover, failure to master early developmental tasks can disrupt the achievement of later ones, resulting in a snowballing pattern of developmental disadvantage. Although early research focused primarily on disruptions to highly specific developmental tasks (e.g., cognitive development, motor skills), recent research casts a wider net, including, for example, studies of how severe childhood illness can delay or disrupt the development of social skills, social relationships with family members and peers, self-esteem, and academic achievement (e.g., Verhoof et al., 2012).

Another body of research examines the relationships between health and human development from the opposite

point of view: how patterns of development influence subsequent health. From this perspective, successful accomplishment of developmental tasks (be they biologically driven or socially/culturally prescribed) is viewed as a form of resource accumulation that affects subsequent health. Failure to achieve developmental skills can affect health in both the short and long term. During childhood and adolescence, developmental failure is strongly linked to mental illness and psychological distress. But developmental failures or delays also affect health much later in the life course. A broad body of research now documents the links between early achievements – including intelligence, educational achievement, self-esteem, feelings of self-efficacy, and the social skills required for successful relationships – and health in middle and later life (e.g., Goodman et al., 2011).

Although these are distinct bodies of research, both are part of a broader and deeper intellectual transformation of how health and illness are understood by both scientists and the larger public. From the middle of the nineteenth century until the middle of the twentieth century, the *medical model* dominated our understanding of the nature and causes of illness. The medical model views illness as biological dysfunction caused by the failure of the organism's biological defenses to prevent damage or disturbance. It is a complex model, recognizing that threats to the body's integrity can be generated by either external forces (injuries, viral, and bacterial antagonists) or internal dysfunctions (congenital deficits, cellular dysregulation). But it is an incomplete model that ignores health-relevant social, psychological, and cultural factors. Since the 1950s, a multidisciplinary model, often called the *biopsychosocial model of illness*, has become the dominant theory of illness among both scientists and the general public. The major tenet of this model is that illness is a product of the complex and interacting effects of biological, psychological, social, and cultural factors. An important corollary of this model is that biological intervention alone typically cannot prevent, cure, or manage illness. Both the causes of and appropriate interventions for illness and premature mortality occur at multiple levels. Research on the reciprocal and complex links between human development and health has been an important demonstration of the utility of the biopsychosocial model of illness.

The Effects of Health on Human Development: State of the Evidence

A broad body of research documents the strong and robust effects of health at one part of the life course on subsequent developmental outcomes. The strongest evidence is from studies of the effects of childhood physical or mental illness on subsequent development during both childhood and adulthood.

Childhood physical illness is associated with several types of developmental problems and/or delays. The primary developmental outcomes examined are cognitive development and academic performance, emotional control and maturity, successful relationships with peers, and behavioral problems at home or at school (e.g., aggression, withdrawal). Numerous studies document strong and robust relationships between childhood chronic physical illness and developmental problems, especially poor cognitive/academic performance, problems developing relationships with peers, and low self-esteem (Forrest et al., 2011; Martinez et al., 2011; Pinquart, 2013; Pinquart and Shen, 2011). These problems tend to cumulate over time, leading to increasing developmental problems in adolescence (Verhoof et al., 2012). These findings apply to many specific illnesses, including cancer, cystic fibrosis, Type 1 diabetes, and epilepsy.

Childhood mental illnesses also exact a severe developmental toll (Forrest et al., 2011; Martinez et al., 2011). Indeed, children with mental disorders appear to be at greater risk of stigma, bullying, and social isolation than those with physical illnesses (Walker et al., 2008).

The developmental deficits resulting from childhood physical and mental illness continue into adulthood. Childhood physical illness is associated with lower levels of educational achievement and income during adulthood (Contoyannis and Dooley, 2010; Currie, 2009; Goodman et al., 2011). Childhood mental illness is especially damaging, exacting more severe costs on adult socioeconomic achievement than childhood physical illness (Goodman et al., 2011; Smith and Smith, 2010). This pattern may reflect the fact that childhood physical illnesses vary substantially in their effects on adult socioeconomic status (SES). Children suffering from illnesses that can be managed effectively typically attain levels of education, occupation, and income similar to those of their healthy peers (Newacheck and Taylor, 1992). In contrast, illnesses that result in frequent incapacity or threats to survival (such as cystic fibrosis) and those that involve cognitive functioning (such as epilepsy) are associated with poorer adult socioeconomic outcomes.

There is also an intergenerational component to the effects of health on development during childhood and adolescence. A broad body of research consistently reports that children who live with a parent who suffers from chronic physical illness, chronic or intermittent mental illness, and/or disability are at increased risk of developmental problems or delays (e.g., Pakenham and Cox, 2012; Sieh et al., 2013). Again, a variety of developmental problems are observed, including cognitive and learning problems, behavior problems, general social competence, and the ability to forge successful relationships with peers.

Given the strong evidence that chronic illness (of self or parent) places children and adolescents at risk of developmental problems, recent research appropriately focuses on the processes and mechanisms that account for these relationships. This is a complex issue because there are multiple pathways by which illness affects development. In broad brush, however, these pathways can be placed in two categories: characteristics of the illness and family dynamics.

For chronically ill children, characteristics of the illness, its sequellae, and the nature and efficacy of medical management all play a role in decreasing or increasing the risk of developmental problems. Some illnesses have direct effects on developmental outcomes (e.g., epilepsy can impair memory). In other cases, treatments to manage chronic illnesses create obstacles to normal development. Some medications have side effects that place the child at risk of developmental problems (e.g., inability to concentrate); other treatments (e.g.,

chemotherapy) can impair function in multiple domains. Illness severity affects development beyond the symptoms and impairments resulting from the disease. For example, the more absences from school as a result of the illness, the more severe are the developmental problems (McCubbin, 1988). Finally, the stigma associated with chronic illness can lead to rejection – or even bullying – from peers and isolation of the child (Walker et al., 2008).

Family dynamics also play a strong role in increasing or reducing the ill child's risk of developmental problems. Quality of parenting and the parents' methods of coping with a chronically ill child are especially important (Gannoni and Shute, 2010; Tartakovsky and Hamama, 2011). The entire dynamic of family life is changed by the presence of a chronically ill child. Parenting requires more time, special skills, and maintaining a balance between 'normalizing' the child's life as much as possible, while also providing high levels of supervision and medical management. Demands change over time as the illness changes and as the child ages. Providing care to a chronically ill child over long periods of time is draining and affects all areas of parents' lives (Avison et al., 1991).

Family dynamics also help explain why children often experience developmental problems when a parent is chronically ill. Two processes seem to be especially important in this regard. First, if parental illness is accompanied by high levels of marital discord, the risk of developmental problems in children is increased (Pakenham and Cox, 2012). Second, the risk of developmental problems in children is dramatically increased if the ill parent is 'emotionally unavailable' to the child (Sieh et al., 2013).

The Effects of Human Development on Health: State of the Evidence

Research on the ways that human development affects health is somewhat more difficult to identify than research on the effects of health on development. The difficulty rests, in part, on the large number of indicators that researchers view as fitting under the conceptual umbrella of human development. Physical, cognitive, emotional, and social dynamics are all viewed as legitimate developmental phenomena and each of those categories includes a large number of more specific indicators. The focus here is on developmental tasks that emerge during early life, when both normal development and its timing have been well documented.

David J.P. Barker (1997) initiated the large and increasing body of research on what is now variously known as the fetal origins hypothesis, the fetal programing hypothesis, and the thrifty phenotype hypothesis. Barker argued that human development begins in the womb and demonstrated that problems of fetal growth can affect health decades later. Although initially received with skepticism by the scientific community because of the confounding effects of environmental factors, strong support for the theory now exists. A number of fetal characteristics have been linked to health later in life, including birth weight, length of gestation, and maternal gestational diabetes. Low birth weight has received most attention and can lead to what is now known as the thrifty phenotype. In brief, when a fetus receives insufficient nutrients for optimal development, it will make metabolic adaptations to survive and develop as normally as possible. Those adaptations, however, are permanently programed into the fetus and have negative implications for later health (Victoria et al., 2008). All three fetal characteristics are associated with both childhood illnesses (Wehby et al., 2013) and increased risk of coronary heart disease, hypertension, stroke, diabetes, and some cancers in middle and later life (Barker et al., 2013; Durnwald and Landon, 2013). This body of research is arguably the strongest evidence that developmental problems cast a long arm, affecting health during middle and late life. From a more practical perspective, this research provides compelling evidence of the importance of optimal prenatal care.

Developmental trajectories after birth also affect subsequent health. Developmental problems in childhood are associated with health problems that emerge in adolescence or young adulthood. Strauss et al. (2012), for example, report that poor cognitive and social skills in childhood are risk factors for the onset of schizophrenia in adolescence and early adulthood. A wide range of developmental difficulties are associated with poor self-rated health, depressive and anxiety disorders, more school absences for short-term illnesses, and more hospital days (Blanchard et al., 2006; Boyle et al., 2011).

Developmental problems in childhood and adolescence also predict poor health in middle and later life. Deficits in cognitive and social skills during adolescence are associated with increased prevalence and earlier onset of chronic illnesses during middle and old age (Kaestner and Callison, 2011). Childhood behavior problems, especially aggression, predict the experience of sudden 'health shocks' during adulthood (Temcheff et al., 2011). An insecure attachment style, which is established by age two, is associated with adult illnesses and health risks including greater reactivity to stress, use of psychotropic and illicit drugs to control affect, cardiovascular disease, and metabolic syndrome (Maunder and Hunter, 2001).

In perhaps the most comprehensive examination to date, Nobel Prize winner James Heckman (2012) recently documented the wide ranging ways in which and ages at which developmental disabilities exact a heavy cost on health for decades.

Methodological Challenges

Several methodological issues complicate efforts to better understand the links between development and health; two are especially important. First, accurate depiction of the complex, reciprocal, and time-dependent relationships between development and health requires longitudinal data collected over most of the life course, i.e., from birth until late life. This long time span is needed because it is not until middle age that a significant proportion of the population experiences significant health problems, although the antecedents of those problems often occur decades earlier. Although the number of studies monitoring individuals over long segments of the life course has increased in the past two decades, none of them cover the optimal age range. As a result, scientists rely on cross-sectional comparisons and short-term longitudinal data sources. Until data covering the entire life course are available,

however, there is always the threat that what appear to be age changes are actually cohort differences.

Second, selective mortality poses a serious threat to our understanding of the links between development and health because both are related strongly to mortality. Selective mortality bias is a result of the fact that survivors of a given cohort are healthier and more socially and economically advantaged than cohort members who die. Consequently, as cohorts age, survivors become increasingly 'elite.' As a result of selective mortality, research that examines the links between development and health – especially studies restricted to adults – almost certainly *underestimate* the effects of development on health and, vice-versa, the effects of health on development. Selective mortality also may affect patterns observed among survivors. For example, the narrowing of SES differences in health during late life may result from earlier mortality of low SES persons rather than a 'biological leveling' at the end of life (see Zajacova and Burgard, 2013, for an extensive examination of compositional changes in aging cohorts due to selective mortality).

Future Directions

Future research on the links between health and human development is likely to focus on two major topics. First, efforts will continue to identify the specific pathways and mechanisms by which development and health affect each other. Although extant research has made important contributions to this issue, many mechanisms remain unknown. In addition, replication of findings to date is especially important because of reliance on cross-sectional and short-term longitudinal research designs. Second, future research will address the extent to which and ways in which developmental and health problems early in life can be compensated for or reversed. There will be formidable obstacles to achieving these research goals, but the benefits to human health and well-being will be worth the efforts.

See also: Adolescent Health and Health Behaviors; Chronic Illness, Psychosocial Coping With; Health Psychology in Childhood; Life Course Epidemiology; Stress, Coping and Health.

Bibliography

Avison, W.R., Noh, S., Speechley, K.N., 1991. Parents as caregivers: caring for children with health problems. In: Albrecht, G.A., Levy, J.L. (Eds.), Advances in Medical Sociology, vol. 2. JAI Press, Greenwich, CT, pp. 65–94.

Barker, D.J.P., 1997. Maternal nutrition, fetal nutrition, and disease in later life. Nutrition 13, 807–813.

Barker, D.J.P., Osmond, C., Thornburg, K.L., Kajantie, E., Eriksson, J.G., 2013. The intrauterine origins of Hodgkin's lymphoma. Cancer Epidemiology 37, 321–323.

Blanchard, L.T., Gurka, M.J., Blackman, J.A., 2006. Emotional, developmental, and behavioral health of American children and their families: a report from the 2003 national survey of children's health. Pediatrics 117, e1202–e1212.

Boyle, C.A., Boulet, S., Schieve, L.A., Cohen, R.A., Blumberg, S.J., Yeargin-Allsopp, M., Kogan, M.D., 2011. Trends in the prevalence of developmental disabilities in U.S. children, 1997–2008. Pediatrics 127, 1034–1042.

Contoyannis, P., Dooley, M., 2010. The role of child health and economic status in educational, health, and labour market outcomes in young adulthood. Canadian Journal of Economics 43, 323–346.

Currie, J., 2009. Healthy, wealthy, and wise: socioeconomic status, poor health in childhood, and human capital development. Journal of Economic Literature 47, 87–122.

Durnwald, C., Landon, M., 2013. Fetal links to chronic disease: the role of gestational diabetes mellitus. American Journal of Perinatology 30, 343–346.

Forrest, C.B., Bevans, K.B., Riley, A.W., Crespo, R., Louis, T.A., 2011. School outcomes of children with special health care needs. Pediatrics 128, 303–312.

Gannoni, A.F., Shute, R.H., 2010. Parental and child perspectives on adaptation to childhood chronic illness: a qualitative study. Clinical Child Psychology and Psychiatry 15, 39–53.

Goodman, A., Joyce, R., Smith, J.P., 2011. The long shadow of childhood physical and mental problems on adult life. PNAS Proceedings of the National Academy of Sciences of the United States of America 108, 6032–6037.

Heckman, J.J., 2012. The developmental origins of health. Health Economics 21, 24–29.

Kaestner, R., Callison, K., 2011. Adolescent cognitive and noncognitive correlates of adult health. Journal of Human Capital 5, 29–69.

Martinez, W.C., Legato, J.S., Lauren, J., 2011. Social competence in children with chronic illness: a meta-analytic review. Journal of Pediatric Psychology 36, 878–899.

Maunder, R.G., Hunter, J.J., 2001. Attachment and psychosomatic medicine: developmental contributions to stress and disease. Psychosomatic Medicine 63, 556–567.

McCubbin, M.A., 1988. Family stress, resources, and family types: chronic illness in children. Family Relations 37, 203–209.

Newacheck, P., Taylor, W.R., 1992. Childhood chronic illness: prevalence, severity, and impact. American Journal of Public Health 82, 364–371.

Pakenham, K.I., Cox, S., 2012. Test of a model of the effects of parental illness on youth and family functioning. Health Psychology 31, 580–590.

Pinquart, M., 2013. Self-esteem of children and adolescents with chronic illness: a meta-analysis. Child Care, Health, and Development 39, 151–161.

Pinquart, M., Shen, Y., 2011. Behavior problems in children and adolescents with chronic physical illness: a meta-analysis. Journal of Pediatric Psychology 36, 1003–1016.

Sieh, D.S., Oort, F.J., Visser-Meily, J.M.A., Meijer, A.M., 2013. Mediators for internalizing problems in adolescents of parents with chronic medical conditions. Journal of Developmental and Physical Disabilities published online March 29, 2013 (10.1007/s10882-013-9345-1).

Smith, J.P., Smith, G.C., 2010. Long-term economic costs of psychological problems during childhood. Social Science & Medicine 71, 110–115.

Strauss, G.P., Allen, D.N., Miski, P., Buchanan, R.W., Kirkpatrick, B., Carpenter Jr., W.T., 2012. Differential patterns of premorbid social and academic deterioration in deficit and nondeficit schizophrenia. Schizophrenia Research 135, 134–138.

Tartakovsky, E., Hamama, L., 2011. Mothers' acceptance-rejection of their children infected with HI: the role of the mothers' social axioms, psychological distress, and relationships with the partner. Journal of Pediatric Psychology 36, 1030–1042.

Temcheff, C.E., Serbin, L.A., Martin-Storey, A., Stack, D.M., Ledingham, J., Schwartzman, A.E., 2011. Predicting adult physical health outcomes from childhood aggression, social withdrawal, and likeability: a 30-year prospective longitudinal study. International Journal of Behavioral Medicine 18, 5–12.

Verhoof, E., Maurice-Stam, H., Heymans, H., Grootenhuis, M., 2012. Growing into disability benefits? Psychosocial course of life of young adults with chronic somatic disease or disability. Acta Paediatrica 10, e19–e26.

Victoria, C.G., Adair, L., Fell, C., Hallal, P.C., Martorell, R., Richter, L., Sachdev, H.S., 2008. Maternal and child undernutrition: consequences for adult health and human capital. The Lancet 371, 340–357.

Walker, J.S., Coleman, D., Lee, J., Squire, P.N., Friesen, B.J., 2008. Children's stigmatization of childhood depression and ADHD: magnitude and demographic variation in a national sample. Journal of the American Academy of Child & Adolescent Psychiatry 47, 912–920.

Wehby, G.L., Nyarko, K.A., Lopez-Camelo, J.S., 2013. Fetal health shocks and early inequalities in health capital accumulation. Health Economics. http://dx.doi.org/10.1002/hec.2901.

Zajacova, A., Burgard, S.A., 2013. Healthier, wealthier, and wiser: a demonstration of compositional changes in aging cohorts due to selective mortality. Population Research and Policy Review 32, 311–324.

Human Development, Successful: Psychological Conceptions

Werner Greve, University of Hildesheim, Hildesheim, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by W. Greve, volume 10, pp. 6970–6974, © 2001, Elsevier Ltd.

Abstract

Evaluating personal development as 'successful' presupposes the existence of normative criteria. Although psychological theories and empirical facts by themselves do not form a sufficient basis for legitimate evaluations and norms, they are nevertheless a necessary component. Accordingly, this article discusses psychological concepts that empirically describe and theoretically explain the courses and outcomes of human development. In particular, the concepts of plasticity, resilience, individual capacity to act, and social integration are examined. These concepts and the resulting psychological model of successful development will be explained by focusing on recent models of developmental regulation.

From its outset, human development is dependent on the presence of essential and beneficial conditions in the physical and social environment. Some of these conditions are fundamental and the same for all human beings (oxygen, water, food), while others are more complex and necessary to varying degrees (e.g., support of caregivers, language-education, culturally specific competencies). Human beings also differ from one another in their psychological and physical needs: congenital and acquired vulnerability and deficiencies need special support and encouragement, some of them may even require specialist intervention or preventive treatment (e.g., specialized diets in the case of inherited metabolic disturbances such as phenylketonuria) (*see* Human Development, Theories of). At any event, there needs to be a minimum degree of *fit* between the individual and his or her environment before development can actually take place. Within these bounds, however, a very wide range of human development can unfold: An individual's predispositions and qualities encounter an immeasurably vast range of diverse environmental influences (socialization, critical and nonnormative life events, historical influences, etc.). These influential factors operate throughout an individual's ontogenesis (Baltes et al., 2006). Human development never reaches completion, is multidirectional and seldom irreversible, while at the same time featuring a high degree of interindividual *variability* and intraindividual *plasticity* (Baltes, 1987, 1997; *see* Neural Plasticity). Moreover, there is now a broad consensus that human development does not take place uniformly in an irreversible sequence of different stages and phases, but is a process that individuals themselves actively influence in substantial ways (Brandtstädter, 2006; *see* Brain Maturation and Plasticity during Infancy and Adulthood).

In view of this broad range of possibilities, the question of how and under what circumstances human development can *succeed* proves to be complex in several respects. The purpose of this article is to deal with two questions. The more important, of course, is the question of what *makes* development successful (from a psychological point of view); Which are the prerequisites of an 'art of life' (Staudinger, 1999)? To answer this question, however, the term 'successful' needs to be clarified.

What Is 'Successful' Development? The Problem of Criteria

Since the work of David Hume (1739–40/1978) many have argued that while empirical sciences like psychology may, for reasons of principle, apply criteria that are value judgments, they cannot justify or legitimize the use of such criteria. According to this view, no matter how good science may be, it can only describe and explain the world, but it cannot judge it. Is the question of what makes human development 'successful' hence unanswerable from the start, at least from the viewpoint of an empirical science?

Is and Ought – Bridging a Presumed Gap

Actually, the relationship between empirical description and explanation on the one hand and value judgments and normative standards on the other is in fact more complex than that. Even if empirical facts, to begin with, do not constitute *sufficient* conditions for making evaluations, as a rule they are in many cases *necessary* ones (Brandtstädter and Schneewind, 1977). Any decision or action is predicated not only upon preferences and value judgments but also upon expectations which take account of empirical facts. For instance, the demand for a juvenile who commits crimes to be 'educated' by punishment assumes that punishment actually serves to influence his or her future actions. This is an assumption that can, and indeed must, be empirically tested. In this respect, science plays a key role in answering the question of successful development, even if the criterion for measuring 'success' has had to be determined on a prescientific basis.

Yet it is open to question whether assessments of human action and biographies really are as completely arbitrary as is suggested by a Humean perspective. Without wishing to get involved in the finer points of the philosophical debates surrounding the 'deducibility' of normative statements from descriptive ones, it can nevertheless be said that all positions purporting to offer proof of the inherent distinction between 'is' and 'ought' in fact tend to presuppose this 'gap' of principle rather than to prove its existence (Mavrodes, 1968). In search for firm foundations for value judgments as well as for empirical descriptions, the last line of defense is the practice

of simply not calling certain statements into question because the grounds for doubting them appear more dubious than the statements themselves.

So, although we can forget any prospect of finding 'the' best biography per se in the foreseeable future, indisputably negative examples can easily be identified within the variations realized. For instance, there is no doubt that a person who persistently and repeatedly violates others for no reason, leads a less 'productive' life than one who shows reasonable consideration for others' interests and needs when taking his or her actions and decisions. Likewise, one can hardly describe an old age marred by illness and loneliness, causing distress particularly to the aged person but also to his or her relatives and friends, as 'successful' without arousing the suspicion that this is an artificial, academic discussion or simply cynicism. The details of examples such as these will be controversial, and here too particular constellations will lead to changes in our assessments. Nevertheless, such examples show that in many cases value judgments ('better' – 'worse') are scarcely more doubtful than the 'brute' facts. This does not entail, however, that value judgments and normative statements can be derived from scientifically proved descriptive statements. Empirical sciences such as psychology do need additional (normative) premises if they are to progress from describing and explaining the world toward altering it. Yet, some of these premises are hardly controversial but rather undisputable.

Nature and Culture: Flexible Intersections

Human development is more than just a biological process. Since we are "unfinished animals … who complete themselves through culture" (Geertz, 1973: p. 44), human ontogenesis is in need of compensatory support by cultural and individual scaffolding and self-regulatory efforts. The 'incomplete architecture' of human development (Baltes, 1997) requires and, at the same time, allows the individual shaping of the life course (*see* Life Course: Sociological Aspects). This, of course, is the very reason why we are in need for (evaluative) criteria of what 'successful' development actually means. As a consequence of the individual's possibility to influence his or her life course, the search for universal criteria of successful development which can be applied to all people irrespective of their individual, cultural and historical circumstances, is likely to be fruitless. The actional perspective of human development entails that people's 'success' depends to no small degree on what they themselves consider to be successful. Within the development-regulating normative rules and developmental tasks of a culture, which incidentally are themselves subject to cultural development processes, the individual him- or herself organizes the ways of shaping his or her own life, looking for and construing orientations that give meaning to his or her personal development (Brandtstädter, 2006).

And yet, the individual perspective has to be complemented with a general one. Even if we are the coproducers of our development (Lerner and Busch-Rossnagel, 1981), it is constrained by the laws of nature, in particular by biological laws. Moreover, human development obviously follows a universal pathway. Since the course of human development is, like any other aspect of human beings, a result of evolutionary processes (Bjorklund and Pellegrini, 2002), its functional necessities delineate the individual's margins for self-regulation, let alone self-formation (*see* Lifespan Development: Evolutionary Perspectives). Beyond securing (or at least supporting) the individual's survival at every stage, development serves the ultimate criterion of evolution, i.e., reproduction. Beyond the preparing function of childhood and youth ('deferred adaptations'; Bjorklund and Pellegrini, 2002), the criterion of reproductive functionality (i.e., 'fitness') includes the entire life span, even up to its final stages in old age (Greve and Bjorklund, 2009; *see* Fertility Theory: Theory of Life History Evolution). The sufficient fulfillment of the respective functional necessities of every developmental stage, both with respect to survival of this particular stage and to the respective contribution to reproduction (i.e., the survival of the next generation(s)), represents another necessary condition for 'successful' development (at every stage and hence in general).

Individual development is the result of the complex interactions of both biological and cultural (social) influences. These influences, in turn, are subject to permanent changes: Beyond the biological evolution, cultural factors evolve as well (Jablonka and Lamb, 2005), and so does the mutual influence of biological and cultural influences (and their interactive interplay) (*see* Culture and the Self: Implications for Psychological Theory). Within the boundaries set by biological and cultural preconditions, the individual's actions play a pivotal role in this multilayered complexity. Hence, particularly those developmental processes which enable in the individual to act and react in his or her environment provide a basis for a further group of necessary conditions for developmental 'success.'

Taken together, these considerations point toward a solution to the question of criteria. Beyond the fulfillment of the fundamental conditions essential for life, the prerequisites for successful development are clearly those meta-competencies and meta-resources which ensure that individuals can make use of opportunities and encouragement, can compensate for deficiencies and vulnerabilities, can choose and realize their own developmental goals, and thus can actively shape their development within the framework defined by immediate and potential possibilities (Brandtstädter and Schneewind, 1977). This meta-perspective implies a multicriteria approach on the one hand, and the integration of subjective and objective criteria of 'success' on the other. In particular, the adaptivity of the developing individual, allowing and facilitating it to adapt to the biocultural niche it is confronted with, should be of particular importance.

What Makes Development Successful?

Hence, in order to answer the question of what makes development a successful one, we have to look out for those conditions that enable (or support) the individual to act and react in a way that his or her survival can be secured, capacity to reproduce (or to contribute to the reproduction of sufficiently close relatives) is maintained (or enhanced), capacity

to pursue individual goals is maintained, and capacity to compensate for losses and obstacles is preserved.

Efficacy: The Capacity to Act

If human development, though canalized by biological constraints and regulated by cultural requirements, is essentially shaped by individual action, a necessary prerequisite for successful development is the individual's capacity to act, i.e., to choose developmental goals, to make decisions, to form intentions and to put them into practice, and also to overcome resistance and to revise strategies where necessary (Brandtstädter, 2006). Acting successfully entails first the precise competencies needed to carry out a particular action in practice. Additionally, an adequate representation of these competencies and potentials in the person's self-concept is a prerequisite for an appropriate and autonomous choice to be made from among the possible developmental options.

Although difficulties of coping with developmental tasks occur throughout the life course, juvenile development serves as a good example with which to illustrate (*see* Childhood and Adolescence: Developmental Assets). Challenges such as physical changes and sexual maturation, the need of developing interpersonal skills (e.g., for relationships with the opposite sex), acquiring education, and, in particular, forming a personal and social identity are not just the storms to be weathered during adolescence; these challenges are vitally bound up with laying the groundwork for successfully solving the tasks and meeting the demands of adult life (Crockett and Crouter, 1995). In particular, social developmental tasks play a particularly prominent role during the years of youth (Petersen and Leffert, 1995). Especially at a juvenile age, the dualism involved in social integration is patently evident: at one and the same time, social integration is an aim in itself and a prerequisite for successfully fulfilling the developmental task of preparing adulthood. However, both the necessity and the opportunity to change and challenge social contexts are characteristic features of adolescence too. Since the dividing line between creatively transgressing the social rules and a deviant or pathological disregard for social limits can be a fuzzy one, these dilemmata of individuation versus integration and of autonomy versus adaptation may serve as a background for the explanation of juvenile delinquency (Coie and Dodge, 1998). Delinquent juveniles, in particular those whose deviance is 'adolescence-limited' (Moffitt, 1993), have evidently not managed to cope successfully with these tasks. Thus, personal resources of coping with discrepancies between the social demands on the one hand and the personal goals, desires, and competencies on the other gain a key function in explaining successful juvenile development. In particular, the goodness of fit between the individual and his or her social environment thus becomes an important precondition and, at the same time, a central criterion for developmental success of adolescence.

Coping: Compensate for Losses, Regulating Disappointments

Throughout our life span, but in later stages of life more than in youth and early adolescence, individuals need to compensate for losses and blocked goals: In order to regain one's capacity to solve life-problems actively (Heckhausen et al., 2010), we have to overcome disappointments to readjust our aspirations and find new goals.

Actually, human development in general can be regarded as a hierarchical, interactive sequence of micro- and macroscopic coping processes (*see* Human Development, Theories of). While current coping processes and resources are shaped by developmental conditions, coping reactions can also act as forces in creating future development. That leads on to the investigation of developmental constellations of individual and social resources that make an individual either resilient against or vulnerable to new strains as they arise. The discussion of the conditions under which people cope successfully with aging illuminates this aspect very well. Numerous studies have shown that people's ability to function and their sense of well-being are not greatly impaired as they grow old, if at all. Indicators of overall psychological status such as self-esteem, general sense of well-being, or being convinced of having control over one's own life are variables which, taken together, do not show any substantial covariance with age (e.g., Brandtstädter et al., 1993; Staudinger et al., 1995). As a consequence, the stereotype of aging being a process strongly characterized by deficiencies and loss has gradually gone out of favor as being too one-sided and incomplete, to be superseded by a more multifaceted view of 'productive' and 'successful' aging (Baltes and Baltes, 1990; *see* Retirement and Encore Adulthood: The New Later Life Course). These findings are most surprising when set against the well-documented fact that, whether in terms of cognitive development, changing sensory functions or morbidity, aging does entail a real increase in the losses and impairment people suffer (Bond et al., 1995; *see* Human Development, Theories of).

Resilience: The Integration of Regulational Layers

To this end, a plethora of processes of self-regulation performs an integrating, 'orchestrating' system (Staudinger et al., 1995); this ties in with the 'developmental system's' view on human development (proposed both from psychological and biological perspectives; e.g., Ford and Lerner, 1992; Oyama et al., 2001). Self-regulating processes can essentially be classified into two modes (Brandtstädter, 2006; Brandtstädter and Greve, 1994). In the 'assimilative' mode, goals are maintained even against resistance, and strategies are pursued tenaciously until they are achieved. Obviously development can only succeed if people are sufficiently willing and capable to pursue his or her developmental goals and to solve developmental tasks. Anyone who always gives up immediately will achieve neither essential tasks nor those of his or her choosing. Yet, sometimes a goal will be – or become – unreachable (e.g., once a woman has passed the menopause, it is just impossible to bear a child). Accordingly, in the 'accommodative' mode, when the individual is confronted with goals that are truly or are perceived to be unachievable, or with irreversible losses and deterioration, preferences and orientations are adapted in a way that avoids meaninglessness and depressive tendencies, and preserves subjective well-being, but at the same time brings new,

achievable goals into view, which revive the capacity and motivation to act.

Thus, the apparent contradiction between the growing strains and losses on the one hand and the stability of the self and well-being on the other is resolved once the personal constellations of 'resilience' are made the focus of the theoretical perspective (Baltes and Baltes, 1990; Brandtstädter, 2006; Greve and Staudinger, 2006; Leipold and Greve, 2009; Staudinger et al., 1995). In addition to the model of developmental self-regulation already mentioned (Brandtstädter, 2006), the several models of developmental regulation provide important insights that help to explain resilience in old age. For instance, the model of 'selective optimization with compensation' (Baltes and Baltes, 1990) claims that losses can be partly cushioned or alleviated by way of selective improvement (training, expertise), and compensated for as people reorient their lives or change their strategies. The life span theory of motivation (Heckhausen et al., 2010) adds that processes of 'secondary control' support the individual's strive to regain goals if formerly pursued aspirations are blocked (e.g., the goal of having a child after passing the menopause).

Unfortunately, we do not know much about the developmental conditions of the individual's ability to regulate one's development both in the assimilative and the accommodative mode (Thomsen and Greve, 2013). If the developing individual's plasticity and adaptivity are the key condition to fulfill biological, cultural, and individual goals, the conditions for the development of adaptivity are of particular importance for developmental 'success' (under all conceivable circumstances and conditions).

Conclusion: Courses and Conditions of Successful Development

In terms of cognitive functions, social integration, and self-regulation, human development features a considerable amount of plasticity (Baltes et al., 2006), which not only makes up the core prerequisite for the dynamics of resilience and the coping processes discussed here, but indeed is also a general precondition for the very possibility of life span development. Self-regulation processes are a central criterion for adaptation processes on the one hand and they substantially control them on the other.

Personal well-being and psychological health, and social integration can only be maintained and improved if the individual can be protected against unfavorable conditions by resilient resource constellations, if personal and social coping resources can help to cushion and compensate for the crises and losses that will inevitably occur, and if the individual capacity to act is largely maintained, whether by stabilizing the required conditions or by finding a new orientation toward achievable objectives. Actually, all of these aspects can be seen both as the products and the producers of successful development.

Development essentially consists in upholding and implementing the individual's capacity to adapt to the discrepancies in new situations and developmental tasks as he or she seeks (via assimilation or accommodation) to attain a solution to, or dissolution of, the discrepancies that are generating a crisis or creating a burden, if these are unavoidable in the long run. Thus, successful development implies a progressive adaptation which simultaneously maintains or indeed expands this plasticity and adaptivity. As far as we know today, this is possible – and usual – at least up to the latest years of very old age (Baltes, 1997). That leads to the somewhat tautological-sounding statement that successful development essentially means ensuring that development, i.e., progressive adaptation, will always be possible and really will occur.

The emerging perspective of adaptivity and activity as the roots of successful human development suggests essential new research perspectives focusing on sources of resiliency and vulnerability at any stage of human development. Moreover, even if the preconditions and processes are known, which provide successful development, the ways to improve the individual's capacity, to acquire, or to use them are still to be investigated. Thus, studies in applied developmental science are needed in order to discover techniques and strategies of individual and social improvement of both of these core conditions, with respect to special interventions, e.g., in the case of juvenile delinquency, as well as with respect to primary prevention of developmental crises, e.g., in the case of successful aging in very old age.

See also: Aging Mind: Facets and Levels of Analysis; Aging, Theories of; Childhood and Adolescence: Developmental Assets; Culture and the Self: Implications for Psychological Theory; Fertility Theory: Theory of Life History Evolution; Human Development, Theories of; Lifespan Development: Evolutionary Perspectives; Motivation: Life Course and Sociological Perspectives; Neural Plasticity; Retirement and Encore Adulthood: The New Later Life Course.

Bibliography

Baltes, P.B., 1987. Theoretical propositions of life-span developmental psychology: on the dynamics between growth and decline. Developmental Psychology 23, 611–626.

Baltes, P.B., 1997. On the incomplete architecture of human ontogeny. American Psychologist 52, 366–380.

Baltes, P.B., Baltes, M.M., 1990. Psychological perspectives on successful aging: the model of selective optimization with compensation. In: Baltes, P.B., Baltes, M.M. (Eds.), Successful Aging. Perspectives from the Behavioral Sciences. Cambridge University Press, New York.

Baltes, P.B., Lindenberger, U., Staudinger, U.M., 2006. Lifespan theory in developmental psychology. In: Lerner, R.M. (Ed.), Handbook of Child Psychology, pp. 569–664.

Bjorklund, D.F., Pellegrini, A.D., 2002. The Origins of Human Nature. Evolutionary Developmental Psychology. APA, Washington.

Bond, L.A., Cutler, S.J., Grams, A. (Eds.), 1995. Promoting Successful and Productive Aging. Sage, Thousand Oaks, CA.

Brandtstädter, J., 2006. Action perspectives on human development. In: Lerner, R.M. (Ed.), Theoretical Models of Human Development, Handbook of child psychology, sixth ed., vol. 1. Wiley, New York, pp. 516–568.

Brandtstädter, J., Greve, W., 1994. The aging self: stabilizing and protective processes. Developmental Review 14, 52–80.

Brandtstädter, J., Schneewind, K.A., 1977. Optimal human development: some implications for psychology. Human Development 20, 1–30.

Brandtstädter, J., Wentura, D., Greve, W., 1993. Adaptive ressources of the aging self: outlines of an emergent perspective. International Journal of Behavioral Development 16, 323–349.

Coie, J.D., Dodge, K.A., 1998. Aggression and antisocial behavior. In: Eisenberg N. (Ed.) Social, Emotional, and Personality Development. In: Damon, W. (Ed.-in-Chief), Handbook of child psychology, vol. 3. Wiley, New York, pp. 779–862.

Crockett, L.J., Crouter, A.C., 1995. Pathways through adolescence: an overview. In: Crockett, L.J., Crouter, A.C. (Eds.), Pathways through Adolescence. Lawrence Erlbaum, Mahwah, NJ.

Ford, D.H., Lerner, R.M., 1992. Developmental Systems Theory: An Integrative Approach. Sage Publications, Thousand Oaks, CA.

Geertz, C., 1973. The Interpretation of Cultures. Selected Essays. Basic Books, New York.

Greve, W., Bjorklund, D., 2009. The Nestor-effect. Extending evolutionary developmental psychology to a lifespan perspective. Developmental Review 29, 163–179.

Greve, W., Staudinger, U.M., 2006. Resilience in later adulthood and old age: resources and potentials for successful aging. In: Cichetti, D., Cohen, D. (Eds.), Developmental psychopathology, second ed., vol. 3. Wiley, New York, pp. 796–840.

Heckhausen, J., Wrosch, C., Schulz, R., 2010. A motivational theory of lifespan development. Psychological Review 117, 32–60.

Hume, D., 1739–40/1978. A treatise of human nature. In: Selby-Bigge, L.A. (Ed.). Clarendon, Oxford, UK.

Jablonka, E., Lamb, M.J., 2005. Evolution in Four Dimensions. MIT Press, Cambridge, MA.

Leipold, B., Greve, W., 2009. Resilience – a conceptual bridge between coping and development. European Psychologist 14, 40–50.

Lerner, R.M., Busch-Rossnagel, N.A. (Eds.), 1981. Individuals as Producers of Their Development. Academic Press, New York.

Mavrodes, G.I., 1968. On deriving the normative from the nonnormative. Papers of the Michigan Academy of Science, Arts, and Letters 53, 353–365.

Moffitt, T.E., 1993. Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy. Psychological Review 100, 674–701.

Oyama, S., Griffiths, P.E., Gray, R.D. (Eds.), 2001. Cycles of Contingency: Developmental Systems and Evolution. MIT Press, Cambridge, MA.

Petersen, A.C., Leffert, N., 1995. What is special about adolescence? In: Rutter, M. (Ed.), Psychosocial Disturbances in Young People. Cambridge University Press, Cambridge (UK), pp. 3–36.

Staudinger, U.M., 1999. Social cognition and a psychological approach to an art of life. In: Blanchard-Fields, F., Hess, T. (Eds.), Social Cognition and Aging. Academic Press, New York, pp. 343–375.

Staudinger, U.M., Marsiske, M., Baltes, P.B., 1995. Resilience and reserve capacity in later adulthood: potentials and limits of development across the life span. In: Cicchetti, D., Cohen, D. (Eds.), Developmental Psychopathology, vol. 2. Wiley, New York, pp. 801–847.

Thomsen, T., Greve, W., 2013. Accommodative coping in early adolescence: an investigation of possible developmental components. Journal of Adolescence 36, 971–981.

Human Development, Theories of

Richard M Lerner, Rachel M Hershberg, Lacey J Hilliard, and Sara K Johnson, Institute for Applied Research in Youth Development, Tufts University, Medford, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Relational developmental systems (RDS) models frame contemporary scholarship in human development – whether in infancy, childhood, adolescence, adulthood, or late adulthood – and encourage investigations of variation both within time and across people in their trajectories of *individual↔context relations* using person-centered and change-sensitive methods. Accordingly, we contrast RDS theory-framed research with prior theoretical approaches, specifying implications of RDS theories for description, explanation, and optimization of development across the life span. We note implications for programs and policies and for promoting social justice among diverse individuals, families, and communities.

Developmental science seeks to describe, to explain, and to optimize within-individual changes and differences among individuals in these within-individual changes across the life span (Baltes et al., 1977; Lerner, 2012). Although the goals of description, explanation, and optimization existed in prior instantiations of the field – in child psychology and then in developmental psychology (Lerner, 2012) – contemporary developmental scientists approach these three objectives differently than in the past. Whether studying infancy, childhood, adolescence, or the adulthood and aging portions of the life span, contemporary scholarship in human development attempts to explain how mutually influential relations between individuals and their contexts provide the basis for individual behavior and development (i.e., bidirectional, reciprocal, synergistic, or fused relations; e.g., Tobach and Greenberg, 1984). Today, then, developmental scientists focus on systematic and successive alterations in the course of these relations and on the integration of multiple attributes of the individual (e.g., physiological, cognitive, emotional, motivational, and behavioral characteristics) and multiple levels of the ecology of human development. These levels range from the biological level through the sociocultural and historical levels, including the designed and natural environments (Bronfenbrenner and Morris, 2006; Lerner, 2002, 2006).

This focus of contemporary developmental science is framed within relational developmental systems (RDS) theoretical models (e.g., Overton, 2010b, 2012; Overton and Müller, 2013; see Lerner et al., 2013; for an example of such framing within the study of adolescent development). These models integrate the actions of the individual on the context and the actions of the context on the individual, represented as *individual↔context relations* (Brandtstädter, 1998, 2006; Lerner, 2002, 2006). The emphasis on RDS models results in the view that developmental science is a nonergodic field (Nesselroade and Molenaar, 2010). The ergotic theorem holds that data sets are marked by (1) homogeneity across individuals in a three-dimensional matrix involving persons, variables, and time; and (2) stationarity of individuals' scores on variables across time (Molenaar, 2007). Developmental scientists, however, believe that there is variation both within time and across people in their trajectories of *individual↔context relations* (i.e., *across time differences*). In other words, people differ in their paths across the life span. Accordingly, the assumptions of homogeneity and stationarity of the ergodic theorem are rejected in contemporary developmental science (Molenaar, 2007, 2010) when framed by RDS theories. As a consequence of nonergodicity, developmental scientists stress the importance of person-centered and change-sensitive methods in their descriptive and explanatory efforts.

To explain these contemporary features of theory in the study of human development we will first contrast today's RDS theory-framed field with prior approaches to theory. Accordingly, we review the evolution of ideas about developmental theory from the end of the nineteenth century and across the twentieth century. We describe the bases of the emergence of interest in RDS theories and then present an overview of key features of these conceptions. This discussion will enable us to specify the implications of RDS theories for description, explanation, and optimization of development across the life span. Our discussion of optimization enables us to formulate some concluding comments about the links between RDS-predicated programs and policies for promotion of social justice among diverse individuals, families, and communities.

Developmental Science: Past and Present

During the latter years of the nineteenth century and for much of the twentieth century, and perhaps especially in the United States and Western Europe, the study of human development became a visible subfield of psychology (see Cairns and Cairns, 2006; for a review). In this literature, and its antecedents in philosophy (see Baltes, 1983 for a review), development was envisioned to be a life-span phenomenon (e.g., Erikson, 1959; Hall, 1904, 1922). The majority of the scholarship about human development in the United States and Western Europe, however, was focused on infancy and childhood (e.g., Binet and Simon, 1905a,b; Gesell, 1929; Piaget, 1923; Preyer, 1882; Terman, 1925).

As a consequence, across this historical period, child psychology emerged as a specific subarea of psychology, spurred on by the research of scientists studying this age period, by the founding of several university centers and institutes devoted to the study of children (e.g., in Iowa, involving

scholars such as Boyd R. McCandless; Minnesota, involving scholars such as Dale B. Harris; and Yale, involving scholars such as Arnold Gesell), and by the work occurring within land-grant universities in the United States in the field of home economics that focused on children (and families) (Cairns and Cairns, 2006; Lerner and Simon, 1998). At the same time, many of these bases of child psychology also created a purportedly multidisciplinary instantiation of scholarship devoted to the study of children, that is, child *development*. In 1933, the Society for Research in Child Development (SRCD) was founded to promote such a multidisciplinary approach to the study of children (and to the application of child development research). In actuality, however, SRCD was, from its outset, dominated by scholars trained in psychology, and this is still the case today. It is not surprising, then, that whether labeled child psychology or child development, the study of the early portion of the life span was framed in very similar ways by scholars studying children (*see* Childhood and Adolescence: Developmental Assets).

At its inception, the child development (or child psychology) field was framed by Cartesian split conceptions of change across ontogeny and by reductionist accounts of the bases of human development (Overton, 2006, 2010b). The core conceptual issues of child development were the nature–nurture, the continuity–discontinuity, and the stability–instability controversies (Lerner, 2002). These debates and proposed 'solutions' involved, for instance, reducing development to being a phenomenon explained by either nature variables (genes or maturation; e.g., Hamburger, 1957) or operant or respondent stimulus-response connections (e.g., Bijou and Baer, 1961). This split, reductionist ontology about development meant that the epistemological route to learning about the basis of development was to identify the essential (nature *or* nurture) explanatory variable(s). Accordingly, the study of development was also marked by variable-centered analyses, as exemplified by the tables of contents of the editions of the *Handbook of Child Psychology* (e.g., Carmichael, 1946, 1954; Mussen, 1970, 1983) and other major compendiums (e.g., Reese and Lipsitt, 1970; Stevenson, 1963) published during this period.

However, as early as 1970, Mussen, the editor of the third edition of the *Handbook of Child Psychology*, pointed to the potential implications of the growing interest among some scientists to move away from a reductionist approach, which involved descriptions of the variables purportedly accounting for ontogenetic structure and function, and toward an approach that viewed development as involving interrelations among variables (from multiple levels of organization). Mussen (1970) said that "the major contemporary empirical and theoretical emphases in the field of developmental psychology … seem to be on *explanations* of the psychological changes that occur, the mechanisms and processes accounting for growth and development" (p. vii). By pointing to the interest in change processes, Mussen was implying that we needed something more to explain the process of development, unless we believed that nature or nurture variables explained themselves in structure or function.

That 'something more' was already emerging within the study of development – at a series of conferences held at West Virginia University in the late 1960s and early 1970s about the nature and implications of a life-span view of human development (e.g., Baltes and Schaie, 1974; Nesselroade and Reese, 1973; Schaie, 1970) (*see* Lifespan Development, Theory of). These West Virginia University conferences, and the edited books that derived from them, laid the foundation for contemporary RDS theories. First, the conferences and associated books discussed the philosophical, theoretical, and methodological problems associated with split/reductionist accounts of development. Second, they introduced ideas about the potential for plasticity (i.e., the potential for systematic change) in development across life, and they pointed to the role of potentially mutually influential relations between individuals and their normative age- and history-graded contexts, and their nonnormative experiences, in instantiating this plasticity. Finally, they underscored the fundamental necessity of studying intraindividual changes (and interindividual differences in intraindividual changes) involved in these individual↔context relations in order to describe, explain, and optimize the course of human development. These ideas would act synergistically with growing scholarship in Europe that provided theory and data fostering a 'reversal' of focus for developmental inquiry – from variable-centered to person-centered approaches to human development (e.g., Magnusson, 1999). These ideas were also synergistic with work in sociology at the time demonstrating that the course of life was shaped by historical events that one encountered at particular times and in particular places (Elder, 1974) (*see* Motivation: Life Course and Sociological Perspectives).

When taken together, these ideas pointed to the vacuity of split/reductionist models (and their attendant methods). In turn, this work underscored the importance of time and place, person↔context relations, plasticity, and the need for longitudinal (change-sensitive) methods to study intraindividual change across life, as well as the diverse life paths of these intraindividual changes. These ideas, when considered together, presented a major challenge to the then-dominant metatheoretical and theoretical ideas in the field. Indeed, the new ideas about human development formed at the West Virginia University conferences were akin to a Kuhn-like (Kuhn, 1962) paradigm shift (Lerner et al., 2013; Overton and Lerner, 2012), from the then still dominant models, methods, and ontogenetic foci associated with child development.

In addition to being primarily descriptive and normative (Mussen, 1970), with the norms usually generated by studying only a small portion of humanity (i.e., European American middle-class children in the main; Hagen et al., 1990), the extant 'paradigm' was as likely (if not more likely) to use cross-sectional research to study development than it was to employ longitudinal methods. The use of cross-sectional designs (and data analysis methods, e.g., R-technique analyses; e.g., see Cattell, 1966, and for more current versions of these ideas see Nesselroade and Molenaar, 2010) was predicated on the assumption of the applicability of the ergodic theorem (e.g., Molenaar, 2007), that is, as noted, that one could assume interindividual homogeneity and across-time stationarity of the modeled parameters (but not necessarily of an individual's observed values). In addition, because of the presumption that the descriptions being generated from

research were explained by the split, reductionist conception framing the research, studies during this period were also largely atheoretical in character (Bronfenbrenner, 1963; McCandless, 1967; Mussen, 1970). As such, selections of methods (e.g., the use of experimental designs, as in Reese and Lipsitt, 1970) or data analytic techniques (e.g., analysis of variance, factor analyses) were given priority over theory-based questions in the framing of research.

The work being generated through the advent of the life-span (Baltes et al., 2006) and life-course (Elder and Shanahan, 2006) perspectives, and the emergence of what are now termed RDS theoretical models (Overton, 2011; Overton and Müller, 2013) were linked – at least initially through the West Virginia University conferences – to these perspectives. For example, Riegel's (1976a,b) dialectic model, Lerner's (1978; Lerner and Kauffman, 1985) developmental contextual model, Magnusson's (1999) individual-context model, and Overton's (1984, 2003) organismic contextual model coalesced to shift the 'paradigm.' These strands merged in the 1970s, 1980s, and 1990s to create a focus on person↔context relations and *person↔context models* (Cairns and Cairns, 2006; Lerner, 2002, 2006). As we have noted, these models involved the belief that time and place matter in regard to shaping the course of life (Bronfenbrenner, 2005; Elder, 1998, 2013) and emphasized that developmental science needed to study both the person and the diversity of people in order to understand human development.

In short, the outcome of these theoretical ideas is that the study of development is the study of nonergodic phenomena. Accordingly, across the last decades of the twentieth century there were innovations in methods of designing research about, and analyzing data from, the study of within-person trajectories and differences between people in these trajectories. Developmental psychology per se, however, had neither sufficient conceptual tools nor methodological means to gather and interrelate variables from the multiple levels of analysis needed to describe the *individual↔context relations* involved in human development, to explain these relations, or to test these explanations through optimization efforts. As such, ideas from several disciplines were drawn on to create integrative, nonreductionist, and systems models of human development and, as well, methods from numerous fields (e.g., computer science, economics, neuroscience, molecular biology, sociology, and statistics, as well as psychology) were used to study development. By the beginning of the twenty-first century, true developmental science had been created (Magnusson and Cairns, 1996).

In sum, then, in much of the twentieth century the study of human development was embedded in a Cartesian world view: It gave great credence to a split conception of the world and thus perpetuated interest in the nature–nurture issue (Overton, 1973, 2010b). Similar core conceptual issues within the field included the continuity–discontinuity issue and the stability–instability issue, both framed within this split conception (see Lerner, 2002; Overton, 2006, 2010b, 2012; for reviews). These (false) conceptual commitments led researchers to readily accept genetic reductionism and embrace several theories (or approaches) that claimed to show how genes provided the fundamental material bases of human behavior and development (e.g., behavior genetics, sociobiology, or evolutionary psychology). The apparent truth of these assumptions was maintained, in part, by avoiding consideration of the ecology or context of human development (e.g., see Greenberg, 2011; Gottlieb, 1998; Overton, 2012; for critiques).

Moreover, scientists working within this context focused on what might be called the 'generic human being' (Emmerich, 1968) and on the earliest years of life, progressing in the best cases to the years surrounding puberty (Brim and Kagan, 1980). Today, however, in the middle of the second decade of the twenty-first century, developmental science conceptualizes the entire span of human life as involving change. It also rejects prior conceptions of a nature–nurture divide and its associated problems, embracing instead the importance of the *individual↔context relations* noted earlier.

RDS Theories

RDS theoretical models are, today, at the cutting edge of theory and research within developmental science (e.g., see the two volumes of the *Handbook of Life-Span Development*; Lamb and Freund, 2010; Overton, 2010a). Such theory focuses on the processes that govern mutually influential relations between individuals and their contexts. Brandtstädter (1998) terms these relations 'developmental regulations' and notes that where developmental regulations involve *mutually beneficial individual↔context relations*, they constitute *adaptive* developmental regulations.

With RDS models, all levels of organization within the ecology of human life are involved in mutually influential relationships across the breadth of the entire life course (Bronfenbrenner and Morris, 2006; Riegel, 1975, 1976a,b). The broadest level of the context is history (temporality) and, as a consequence of all levels of person and context being embedded in history, there is always a potential for systematic change – for plasticity (Lerner, 1984) – in *individual↔context relations*. Indeed, given that variation exists across time within contexts and, as well, across contexts within time, differences in time and place constitute vital contributors to plasticity across the life span – even into the 10th and 11th decades of life (Baltes et al., 2006; Elder, 1980; Elder and Shanahan, 2006; Elder et al., 1993). As a consequence, human life is variegated and changing. The processes of human development are neither uniform nor permanent across individuals, settings, or history. As such, the impact of temporality in human development involves the individual life span as well as generational and historical time (Elder, 1998; Elder et al., 1993) (*see* Motivation: Life Course and Sociological Perspectives; Ecology of Aging).

In addition, the array of individual and contextual variables involved in these potentially plastic *individual↔context relations* is virtually infinite. Estimates are that the odds of two genetically identical genotypes arising in the human population is about one in 6.3 billion, and each of these potential human genotypes may be coupled across life with an even larger number of life course trajectories of social experiences (Hirsch, 2004). Thus, the number of human phenotypes that can exist is fundamentally infinite, and the diversity of development becomes a prime, substantive focus

for developmental science. To emphasize the point, it is useful to note that recent studies have produced ample empirical information about the relational developmental character of genetic processes (e.g., see Charney, 2012; for a review).

Current genetic research (1) underscores the embeddedness of genetic processes within processes from other levels of organization of the fused, relational developmental system; (2) documents the idea that *gene↔context relations* are part of the plasticity of within-individual change; and (3) provides a conception of the genome that contradicts the concept of genes found in behavioral genetics, sociobiology, and evolutionary psychology. For instance, Charney (2012) points to how the contemporary scientific study of genetics is signaling not only a nonsplit approach to developmental science but, as well, is tolling a death knell for these genetic reductionist approaches. He says that,

> Rather than being an unchanging template, DNA appears subject to a good deal of environmentally induced change. Instead of identical DNA in all the cells of the body, somatic mosaicism appears to be the normal human condition. And DNA can no longer be considered the sole agent of inheritance. We now know that the epigenome, which regulates gene expressivity, can be inherited via the germline. These developments are particularly significant for behavior genetics for at least three reasons: First, epigenetic regulation, DNA variability, and somatic mosaicism appear to be particularly prevalent in the human brain and probably are involved in much of human behavior; second, they have important implications for the validity of heritability and gene association studies, the methodologies that largely define the discipline of behavior genetics; and third, they appear to play a critical role in development during the perinatal period and, in particular, in enabling phenotypic plasticity in offspring (p. 331).

This evidence, along with that supplied by others (e.g., Ho, 2010; Ho and Saunders, 1984; Greenberg, 2011; Gissis and Jablonka, 2011; Hood et al., 2010; Jablonka and Lamb, 2005; Molenaar, 2010), creates the basis for a true Kuhnian paradigmatic revolution in the study of development (Kuhn, 1962). The findings presented by these scholars constitute anomalies (in effect, falsifications) of the 'old' genetic reductionist paradigm. These anomalies result in a crisis for the reductionist paradigm and, critically, a basis for science (and for working scientists) to turn toward an available, alternative paradigm. This new paradigm is RDS theories and, consistent with Kuhn's (1962) discussion of scientific revolutions, the very findings that are anomalies in (falsifications of) genetic reductionist models (and methods) are integrated within the now-dominant paradigm (Overton and Lerner, 2012).

Simply, to the extent that developmental scientists have successfully validated their theoretical models of the intra-individual change trajectories linked to specific instantiations of *individual↔context relations*, their work should be able to be applied to enhance the likelihood that diverse individuals will be on better (i.e., healthier, more positive) trajectories. Although developmental science will only move asymptotically toward this knowledge base, current emphases in developmental science (Fisher et al., 2012) stress that scholarship should be directed to identifying the answers to questions about how to use integrated explanatory/optimization work to create a more socially just world for diversity of individuals in our global community. This contemporary interest in the connection between basic and applied developmental science that serves social justice may portend a continued focus on an applied developmental science that contributes to social justice in future decades (Lerner, 2012).

Conclusions about RDS Theories

From the perspective of RDS models, to adequately describe and explain human development across the life span, scientists need to focus on the substantive significance of diversity. They must attend to the diversity of *individual↔context relations* that exist within people across time (this diversity makes implausible the ergodic theorem postulation of stationarity across the life span) and they must attend the variation across people in these varying trajectories of within-individual change (this diversity makes implausible the ergodic theorem postulation of homogeneity across people) (Molenaar, 2007). In addition, an RDS theoretical approach to understanding the life course has implications as well for optimization, the third goal of developmental science.

Optimization within RDS Theoretical Approaches to Developmental Science

The focus on diversity in regard to describing and explaining developmental change that is emphasized within RDS theories also involves the expectation that, as a consequence of health-supportive alignments between people and settings, positive changes can be promoted across all instances of variation in *individual↔context relations*. With this stance, diversity becomes the necessary subject of inquiry in developmental science. That is, to understand the bases of and, in turn, to promote *individual↔context relations* that may be characterized as healthy, positive, adaptive, or resilient – which are relations reflecting the maintenance or enhancement of links that are mutually beneficial to individuals and context – scholars must ask a complex, multipart question (Lerner et al., 2012, 2013). Specifically, researchers must ascertain,

- *what fundamental attributes of individuals* (e.g., what features of biology and physiology, cognition, motivation, emotion, ability, physiology, or temperament); among individuals of
- *what status attributes* (e.g., people at what portions of the life span, and of what sex, race, ethnic, religious, geographic location, etc. characteristics); in relation to
- *what characteristics of the context* (e.g., under what conditions of the family, the neighborhood, social policy, the economy, or history); are likely to be associated with
- *what facets of adaptive functioning* (e.g., maintenance of health and of active, positive contributions to family, community, and civil society)?

Moreover, proponents of RDS models believe that the relative plasticity of the *individual↔context relations* creates a synthesis between the explanatory and optimization goals of

developmental science (Baltes et al., 1977; Lerner, 2002). To test explanations of developmental change, scholars need to institute or evaluate programs that are aimed at altering the bidirectional relations expected to produce changes in behavior and development. These actions must necessarily be embedded in the ecologically valid settings of human development in order to have generalizability to the lived experiences of individuals and, therefore, must constitute intervention (applied) research and, at the same time, research testing basic explanatory processes of human development. As such, in contemporary developmental science, the commonly regarded split between basic and applied research is also regarded as a flawed holdover from earlier eras (Fisher et al., 2012; Lerner and Overton, 2008). In short, the application of developmental science (optimization) is a coequal partner with description and explanation within contemporary developmental science.

The emerging approach to developmental science also brings to the fore a concern for promoting social justice (Fisher et al., 2012; Lerner and Overton, 2008). Arguably, the most arduous test of the integrated explanatory/optimization scholarship conducted by developmental scientists involves efforts to bring to scale changes in the developmental system for diverse individuals. If our explanatory models can fully account for the system of *individual↔context relations* that alter the course of development for all individuals, then we should be able to promote more positive development among all individuals. Developmental scientists should be able to specify *what* characteristics, of *what* individuals, should be integrated with *what* features of the ecology of human development, at *what* points across ontogeny, to produce *what* instances of (more optimal) changes in behavior and development (*see* Ecology of Aging). Longitudinal projects, such as the Berlin Aging Study (Baltes et al., 2006) or the 4-H Study of Positive Youth Development (J. Lerner et al., 2013), involve such specifications: in those studies, findings illustrate the importance of research framed by RDS models in explaining the course of development and generating rich ideas for optimization strategies.

Accordingly, rather than optimize their theories to promote success of a 'generic' or 'ideal' person, within-individual developmental systems work should be able to be applied to enhance the likelihood that diverse individuals will be on better (i.e., healthier, more positive) trajectories. Answering the above noted 'what' questions is essential if scholars are to use their integrated explanatory/optimization work to enhance policy and programming, and thereby, create a more socially just world for a diversity of individuals in our global community (Fisher et al., 2012). RDS theories suggest that, through the plasticity of individual↔context relations, means may be found to move all people toward living on a more 'level playing field.' Evidence from RDS-predicated research may be used, then, in policy and program innovations aimed at leveling the playing fields of life for diverse individuals, families, and communities.

Conclusions

The future of RDS theories in developmental science, as a superordinate frame for scholars in psychology, sociology, economics, biology, medicine, education, and other fields interested in describing, explaining, and optimizing the course of human life, seems assured. One bit of evidence in support of this prediction is that the forthcoming, seventh edition of the *Handbook of Child Psychology* will be published in 2015 by John Wiley and Sons with the expanded title of the *Handbook of Child Psychology and Developmental Science* (Lerner, 2015). Across the four volumes of this publication, volume editors (Volume 1, on theory and methodology; Volume 2, on processes of cognitive development; Volume 3, on processes of socioemotional development; and Volume 4, on ecological processes) emphasize the use of RDS perspectives and, as well, point to the implications of this theoretical frame for both understanding and enhancing *individual↔context relations*.

In the future, developmental science will emphasize rigorous, theory-predicated research about the mutually influential relations among individual and ecological processes, about the embodiment of human development within the rich and complex ecology of human life. Methodological innovations about how best to study the relational developmental system in manners maximally sensitive to time, place, and human diversity will therefore continue to be areas of active scholarship. Here we expect greater attention will be paid to tools for multimethod research, to creative ways to partition the x-axis in manners reflecting the diversity of developmental processes, to systems science methods (Urban et al., 2011), and to the triangulation between micro and macro analyses of within-individual change (Werner, 1948, 1957).

In turn, we believe that the application of developmental science will continue to be a core and integrated concern of developmental scientists. The theoretical orientations and interests of new cohorts of developmental scientists, the requirements imposed by funders for producing scholarship that matters in the real world, and the needs for evidence-based means to address the challenges of the twenty-first century will coalesce to make Kurt Lewin's (1952: p. 169) statement "There is nothing so practical as a good theory," an oft-proven empirical reality. Indeed, we believe that the scientific and societal value on which the developmental science of the future will be judged will be whether its theoretical and methodological tools are productive at promoting positive human development across the life span for the diverse people of the world. Therefore, developmental science as a means for promoting social justice may be the most significant lens through which the future contributions of developmental science will be viewed.

Acknowledgment

This work was supported in part by grants from the John Templeton Foundation.

See also: Bruner's Theory of Cognitive Development; Childhood and Adolescence: Developmental Assets; Children, Rights of: Cultural Concerns; Children, Value of; Cognitive Development, Educational Theories of; Cognitive Development: Child Education; Contextual Studies:

Methodology; Control Variables in Research; Cross-Cultural Research Methods in Sociology; Data Collection: Interviewing; Developmental Research across Cultures and Nations: Challenges, Biases, and Cautions; Education and Learning: Lifespan Perspectives; Equivalence and Transfer Problems in Cross-Cultural Research; Families as Educational Settings; Family and Schooling; Gibson's Theory of Perceptual Learning; Human Development, Successful: Psychological Conceptions; Infancy and Human Development; Intelligence, Prior Knowledge, and Learning; Learning Theories and Educational Paradigms; Learning and Instruction: Social-Cognitive Perspectives; Lifelong Learning and Its Support with New Technologies; Lifespan Development, Theory of; Longitudinal Analyses of Sexual Development through Early Adulthood; Longitudinal Data: Event–History Analysis in Discrete Time; Media Effects on Children; Neo-Piagetian Theories of Cognitive Development; Neo-Vygotskian Developmental Theory; Personality Development: Systems Theories; Piaget's Theory of Cognitive Development; Prenatal and Infant Development: Overview, Current Trends, Future Directions; Qualitative Methodology in Developmental Psychology; Scientific Concepts: Development in Children; Self and Identity Development During Adolescence across Cultures; Self-Concepts: Educational Aspects; Self-Development in Childhood and Adolescence; Self-Efficacy: Education Aspects; Vygotsky's Theory of Cognitive Development; Vygotsky's Theory of Human Development and New Approaches to Education.

Bibliography

Baltes, P.B., 1983. Life-span developmental psychology: observations on history and theory revisited. In: Lerner, R.M. (Ed.), Developmental Psychology: Historical and Philosophical Perspectives. Erlbaum, Hillsdale, NJ, pp. 79–112.

Baltes, P.B., Lindenberger, U., Staudinger, U.M., 2006. Life span theory in developmental psychology. In: Lerner, R.M. (Ed.), Handbook of Child Psychology: Theoretical Models of Human Development, sixth ed., vol. 1. Wiley, New York, pp. 569–664.

Baltes, P.B., Reese, H.W., Nesselroade, J.R., 1977. Life – Span Developmental Psychology: Introduction to Research Methods. Brooks/Cole, Monterey, CA.

Baltes, P.B., Schaie, K.W., 1974. Aging and IQ: the myth of the twilight years. Psychology Today 7, 35–40.

Bijou, S.W., Baer, D.M., 1961. Child Development: A Systemic and Empirical Theory, vol. 1. Appleton-Century-Crofts, New York.

Binet, A., Simon, T., 1905a. Sur la necessite d'etablir un diagnostic scientific des etats inferieurs de l'intelligence. L'Annee Psychologique 11, 162–190.

Binet, A., Simon, T., 1905b. Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. L'Annee Psychologique 11, 191–244.

Brandtstädter, J., 1998. Action perspectives on human development. In: Lerner, R.M. (Ed.), Theoretical Models of Human Development. In: Damon, W. (Editor-in-Chief), The Handbook of Child Psychology, vol. 1, fifth ed. Wiley, New York, pp. 807–863.

Brandtstädter, J., 2006. Action perspectives on human development. In: Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology. In: Lerner, R.M. (Ed.), Theoretical Models of Human Development, vol. 1, sixth ed. John Wiley, Hoboken, NJ, pp. 516–568.

Brim Jr., O.G., Kagan, J., 1980. Constancy and change: a view of the issues. In: Brimm Jr., O.G., Kagan, J. (Eds.), Constancy and Change in Human Development. Harvard University Press, Cambridge, MA, pp. 1–25.

Bronfenbrenner, U., 1963. Developmental theory in transition. In: Stevenson, H.W. (Ed.), Child Psychology. Sixty-Second Yearbook of the National Society for the Study of Education, Part 1. University of Chicago Press, Chicago, pp. 517–542.

Bronfenbrenner, U., 2005. Making Human Beings Human: Bioecological Perspectives on Human Development. Sage, Thousand Oaks, CA.

Bronfenbrenner, U., Morris, P.A., 2006. The bioecological model of human development. In: Lerner, R.M. (Ed.), Handbook of Child Psychology. In: Damon, W., Lerner, R.M. (Editors-in-Chief), Theoretical Models of Human Development, vol. 1, sixth ed. John Wiley & Sons, Hoboken, NJ, pp. 793–828.

Cairns, R.B., Cairns, B., 2006. The making of developmental psychology. In: Lerner, R.M. (Ed.), Handbook of Child Psychology, Theoretical Models of Human Development, sixth ed., vol. 1. Wiley, Hoboken, pp. 89–165.

Carmichael, L. (Ed.), 1946. Manual of Child Psychology. Wiley, New York.

Carmichael, L. (Ed.), 1954. Manual of Child Psychology, second ed. Wiley, New York.

Cattell, R.B., 1966. Psychological theory and scientific method. In: Cattell, R.B. (Ed.), Handbook of Multivariate Experimental Psychology. Rand McNally, Chicago, IL, pp. 1–18.

Charney, E., 2012. Behavior genetics and post genomics. Behavioral and Brain Sciences 35 (5), 331–358.

Elder Jr., G.H., 1974. Children of the Great Depression: Social Change in Life Experiences. University of Chicago Press, Chicago.

Elder Jr., G.H., 1980. Adolescence in historical perspective. In: Adel son, J. (Ed.), Handbooks of Adolescent Psychology. Wiley, New York, pp. 3–46.

Elder Jr., G.H., 1998. The life course and human development. In: Damon, W. (Ed.), Handbook of Child Psychology. In: Lerner, R.M. (Ed.), Theoretical Models of Human Development, vol. 1, fifth ed. John Wiley, New York, pp. 939–991.

Elder, G.H., 2013. Studying lives in changing times: a life course journey. In: Lerner, R.M., Petersen, A.C., Silbereisen, R.K., Brooks-Gunn, J. (Eds.), The Developmental Science of Adolescence: History through Autobiography. Psychology Press, New York, NY.

Elder Jr., G.H., Modell, J., Parke, R.D. (Eds.), 1993. Children in Time and Place: Developmental and Historical Insights. Cambridge University Press, New York.

Elder Jr., G.H., Shanahan, M.J., 2006. The life course and human development. In: Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology. In: Lerner, R.M. (Ed.), Theoretical Models of Human Development, vol. 1, sixth ed. John Wiley, Hoboken, NJ, pp. 665–715.

Emmerich, W., 1968. Personality development and concepts of structure. Child Development 39, 671–690.

Erikson, E.H., 1959. Identity and the life cycle. Psychological Issues 1, 50–100.

Fisher, C.B., Busch-Rossnagel, N.A., Jopp, D.S., Brown, J.L., 2012. Applied developmental science, social justice, and socio-political well-being. Applied Developmental Science 16 (1), 54–64.

Gesell, A.L., 1929. Maturation and infant behavior pattern. Psychological Review 36, 307–319.

Gissis, S.B., Jablonka, E. (Eds.), 2011. Transformations of Lamarckism: From Subtle Fluids to Molecular Biology. The MIT Press, Cambridge, MA.

Gottlieb, G., 1998. Normally occurring environmental and behavioral influences on gene activity: from central dogma to probabilistic epigenesis. Psychological Review 105, 792–802.

Greenberg, F., 2011. The failure of biogenetic analysis in psychology: why psychology is not a biological science. Research in Human Development 8 (3–4), 173–191.

Hagen, J.W., Paul, B., Gibb, S., Wolters, C., March 1990. Trends in research as reflected by publications in Child Development: 1930–1989. In: Paper Presented at the Biennial Meeting of the Society for Research on Adolescence, Atlanta, GA.

Hall, G.S., 1904. Adolescence: Its Psychology and Its Relations to Psychology, Anthropology, Sociology, Sex, Crime, Religion, and Education. Appleton, New York.

Hall, G.S., 1922. Senescence: The Last Half of Life. Appleton, New York.

Hamburger, V., 1957. The concept of development in biology. In: Harris, D.B. (Ed.), The Concept of Development. University of Minnesota Press, Minneapolis, MN, pp. 49–58.

Hirsch, J., 2004. Uniqueness, diversity, similarity, repeatability, and heritability. In: Garcia Cole, C., Bearer, E., Lerner, R.M. (Eds.), Nature and Nurture: The Complex Interplay of Genetic and Environmental Influences on Human Behavior and Development. Erlbaum, Mahwah, NJ, pp. 127–138.

Ho, M.W., Saunders, P.T. (Eds.), 1984. Beyond Neo-Darwinism: Introduction to the New Evolutionary Paradigm. Academic Press, London.

Ho, M.W., 2010. Development and evolution revisited. In: Hood, K.E., Halpern, C.T., Greenberg, G., Lerner, R.M. (Eds.), Handbook of Developmental Systems, Behavior and Genetics. Wiley Blackwell, Malden, MA, pp. 61–109.

Hood, K.E., Halpern, C.T., Greenberg, G., Lerner, R.M. (Eds.), 2010. The Handbook of Developmental Science, Behavior and Genetics. Wiley Blackwell, Malden, MA.

Jablonka, E., Lamb, M.J., 2005. Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life. MIT Press, Cambridge, MA.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.

Lamb, M.E., Freund, A.M. (Eds.), 2010. Handbook of Life-Span Development. In: Lerner, R.M. (Editor-in-Chief), Social and Emotional Development, vol. 2. Wiley, Hoboken, NJ.

Lerner, J.V., Bowers, E.P., Minor, K., Boyd, M.J., Mueller, M.K., Schmid, K.L., Napolitano, C.M., Lewin-Bizan, S., Lerner, R.M., 2013. Positive youth development: processes, philosophies, and programs. In: Lerner, R.M., Easterbrooks, M.A., Mistry, J. (Eds.), Handbook of Psychology. In: Weiner, I.B. (Editor-in-Chief), Developmental Psychology, vol. 6, second ed. Wiley, Hoboken, NJ, pp. 365–392.

Lerner, R.M., 1978. Nature, nurture, and dynamic interactionism. Human Development 21, 1–20.

Lerner, R.M., 1984. On the Nature of Human Plasticity. Cambridge University Press, New York.

Lerner, R.M., 2002. Concepts and Theories of Human Development, third ed. Lawrence Erlbaum Associates, Mahwah, NJ.

Lerner, R.M., 2006. Developmental science, developmental systems, and contemporary theories of human development. In: Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology. In: Lerner, R.M. (Ed.), Theoretical Models of Human Development, vol. 1, sixth ed. John Wiley, Hoboken, NJ, pp. 1–17.

Lerner, R.M., 2012. Developmental science and the role of genes in development. GeneWatch 25 (1–2). http://www.councilforresponsiblegenetics.org/genewatch/GeneWatchPage.aspx?pageId=413.

Lerner, R.M. (Editor-in-Chief), 2015. Handbook of Child Psychology and Developmental Science, seventh ed. Wiley, Hoboken, NJ.

Lerner, R.M., Agans, J.P., Arbeit, M.R., Chase, P.A., Weiner, M.B., Schmid, K.L., Warren, A.E.A., 2012. Resilience and positive youth development: a relational developmental systems model. In: Handbook of Resilience in Children, pp. 293–308.

Lerner, R.M., Easterbrooks, M.A., Mistry, J., 2013. Foundations of development across the life span. In: Lerner, R.M., Easterbrooks, M.A., Mistry, J. (Eds.), Handbook of Psychology. In: Weiner, I.B. (Editor-in-Chief), Developmental Psychology, vol. 6, second ed. Wiley, Hoboken, NJ, pp. 3–17.

Lerner, R.M., Kauffman, M.B., 1985. The concept of development in contextualism. Developmental Review 5, 309–333.

Lerner, R.M., Overton, W.F., 2008. Exemplifying the integrations of the relational developmental system: synthesizing theory, research, and application to promote positive development and social justice. Journal of Adolescent Research 23, 245–255.

Lerner, R.M., Simon, L.A.K., 1998. The new American outreach university: challenges and options. In: Lerner, R.M., Simon, L.A.K. (Eds.), University-Community Collaborations for the Twenty-First Century: Outreach Scholarship for Youth and Families. Garland Publishing, New York, pp. 3–23.

Lerner, R.M., Weiner, M.B., Arbeit, M.R., Chase, P.A., Agans, J.P., Schmid, K.L., Warren, A.E., 2012d. Resilience across the life span. Annual Review of Gerontology and Geriatrics 32 (1), 275–299.

Lewin, K., 1952. Field Theory in Social Science: Selected Theoretical Papers. Tavistock, London.

Magnusson, D., 1999. Individual development: toward a developmental science. In: Proceedings of the American Philosophical Society, vol. 143, pp. 86–96.

Magnusson, D., Cairns, R.B., 1996. Developmental science: towards a unified framework. In: Cairns, R.B., Elder Jr., G.H., Costello, E.J. (Eds.), Developmental Science. Cambridge University Press, New York, pp. 7–30.

McCandless, B.R., 1967. Children. Holt, Rinehart & Winston, New York.

Molenaar, P.C.M., 2007. On the implications of the classical ergodic theorems: analysis of developmental processes has to focus on intra-individual variation. Developmental Psychobiology 50, 60–69.

Molenaar, P.C.M., 2010. On the limits of standard quantitative genetic modeling of inter-individual variation: extensions, ergodic conditions and a new genetic factor model of intra-individual variation. In: Hood, K.E., Halpern, C.T., Greenberg, G., Lerner, R.M. (Eds.), Handbook of Developmental Systems, Behavior and Genetics. Wiley Blackwell, Malden, MA, pp. 626–648.

Mussen, P.H. (Ed.), 1970. Carmichael's Manual of Child Psychology, third ed. Wiley, New York.

Mussen, P.H. (Ed.), 1983. Handbook of Child Psychology. Wiley, New York.

Nesselroade, J.R., Molenaar, P.C.M., 2010. Emphasizing intra-individual variability in the study of development over the lifespan. In: Overton, W.R. (Ed.), Cognition, Biology, and Methods across the Life Span. In: Lerner, R.M. (Editor-in-Chief), Handbook of Life-Span Development, vol. 1. Wiley, Hoboken, NJ, pp. 30–54.

Nesselroade, J.R., Reese, H.W. (Eds.), 1973. Life-Span Developmental Psychology: Methodological Issues. Academic Press, New York.

Overton, W.F., 1973. On the assumptive base of the nature-nurture controversy: additive versus interactive conceptions. Human Development 16, 74–89.

Overton, W.F., 1984. World views and their influence on psychological theory and research: Kuhn–Lakatos–Laudan. In: Reese, H.W. (Ed.), Advances in Child Development and Behavior, 18. Academic Press, New York, pp. 191–226.

Overton, W.F., 2003. Development across the life span: philosophy, concepts, theory. In: Lerner, R.M., Easterbrooks, M.A., Mistry, J. (Eds.), Weiner, Irving B. (Editor-in-Chief), Handbook of Psychology: Developmental Psychology, vol. 6. Wiley, New York, pp. 13–42.

Overton, W.F., 2006. Developmental psychology: philosophy, concepts, methodology. In: Lerner, R.M. (Ed.), Handbook of Child Psychology. In: Damon, W., Lerner, R.M. (Editors-in-Chief), Theoretical Models of Human Development, vol. 1, sixth ed. John Wiley & Sons, Hoboken, NJ, pp. 18–88.

Overton, W.F. (Ed.), 2010a. Cognition, Biology, Methods. In: Lerner, R.M. (Editor-in-Chief), The Handbook of Life-Span Development, vol. 1. Wiley, Hoboken, NJ, pp. 1–29.

Overton, W.F., 2010b. Life-span development: concepts and issues. In: Overton, W.R. (Ed.), Cognition, Biology, and Methods across the Life Span. In: Lerner, R.M. (Editor-in-Chief), Handbook of Life-Span Development, vol. 1. Wiley, Hoboken, NJ.

Overton, W.F., 2011. Relational developmental systems and quantitative behavior genetics: alternative of parallel methodologies. Research in Human Development 8, 258–263.

Overton, W.F., 2012. Evolving scientific paradigms: retrospective and prospective. In: L'Abate, L. (Ed.), The Role of Paradigms in Theory Construction. Springer, New York, pp. 31–66.

Overton, W.F., Lerner, R.M., 2012. Relational developmental systems: paradigm for developmental science in the post-genomic era. Behavioral and Brain Sciences 35 (5), 375–376.

Overton, W.F., Müller, U., 2013. Meta-theories, theories, and concepts in the study of development. In: Lerner, R.M., Easterbrooks, M.A., Mistry, J. (Eds.), Weiner, Irving B. (Editor-in-Chief), Comprehensive Handbook of Psychology: Developmental Psychology, vol. 6. Wiley, New York, pp. 19–58.

Piaget, J., 1923. La pensee l'enfant. Archives of Psychology, Geneva 18, 273–304.

Preyer, W.T., 1882. Die Seele des Kindes: Beobachtungen über die geistige Entwicklung des Menschen in den ersten Lebensjahren. Grieben, Leipzig.

Reese, H.W., Lipsitt, L.P. (Eds.), 1970. Experimental Child Psychology. Academic Press, New York.

Riegel, K.F., 1975. Toward a dialectical theory of human development. Human Development 18, 50–64.

Riegel, K.F., 1976a. The dialectics of human development. American Psychologist 31, 689–700.

Riegel, K.F., 1976b. From traits and equilibrium toward developmental dialectics. In: Arnold, W.J., Cole, J.K. (Eds.), Life-Span Developmental Psychology: Dialectical Perspectives on Experimental Research. Academic Press, New York, pp. 1–45.

Schaie, K.W., 1970. A reinterpretation of age-related changes in cognitive structure and functioning. In: Goulet, L.R., Baltes, P.B. (Eds.), Life-Span Developmental Psychology: Research and Theory. Academic Press, New York, pp. 485–507.

Stevenson, H.W. (Ed.), 1963. Child Psychology. Sixty-Second Yearbook of the National Society for the Study of Education, Part 1. University of Chicago Press, Chicago, IL.

Terman, L.M. (Ed.), 1925. Genetic Studies of Genius, I: Mental and Physical Traits of a Thousand Gifted Children. Stanford University Press, Stanford, CA.

Tobach, E., Greenberg, G., 1984. The significance of T.C. Schneirla's contribution to the concept of levels of integration. In: Greenberg, G., Tobach, E. (Eds.), Behavioral Evolution and Integrative Levels. Lawrence Erlbaum, Hillsdale, NJ, pp. 1–7.

Urban, J.B., Osgood, N., Mabry, P., 2011. Developmental systems science: exploring the application of non-linear methods to developmental science questions. Research in Human Development 8, 1–25. http://dx.doi.org/10.1080/15427609.2011.549686.

Werner, H., 1948. Comparative Psychology of Mental Development. International Universities Press, New York.

Werner, H., 1957. The concept of development from a comparative and organismic point of view. In: Harris, D.B. (Ed.), The Concept of Development. University of Minnesota, Minneapolis, pp. 125–148.

Human Ecology

Dudley L Poston, Jr., Texas A&M University, College Station, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The author distinguishes human ecology from ecology and focuses on sociological human ecology as elaborated by Amos Hawley. The author outlines the ecological orientation and introduces several examples of applications of human ecological theory to empirical investigations of the demographic processes. Human ecology is concerned with the organizational aspects of human populations that arise from their sustenance-producing activities. The author also outlines the theoretical and conceptual bases of the four referential constructs of human ecology, namely, organization, population, technology, and environment, reviews some of the empirical and theoretical literature spanning more than five decades, and shows in particular that demographic models benefit from using the ecological perspective.

Introduction

Otis Dudley Duncan once wrote about the difficulties defining human ecology and the perspective of human ecology as follows: "Even a provisional statement of [its] concerns will doubtless encounter strong objections from one or another group of scientists and thinkers who regard their studies of man as exemplifying the ecological viewpoint" (1959: p. 679). In this article, the author first presents various definitions and approaches of human ecology to illustrate its variant aspects and features. In the major part of this article, the author focuses on what he refers to as 'sociological human ecology,' discusses its principal dimensions and characteristics, and outlines its conceptual rubrics.

What Is Human Ecology?

The word ecology is from the Greek *oikos* (οἶκος) meaning 'household,' or 'place to live'; ecology may be defined as the "study of the interrelationships of organisms with their environment and each other" (Smith and Pimm, 2013). One of the earliest statements was that of the Greek philosopher Theophrastus, an associate of Aristotle, who focused on the "interrelationships between organisms and between organisms and their nonliving environment" (Smith and Pimm, 2013). Ernest Haeckel (1868 [1876]) was the first to actually use the term 'ecology' in his study of plants, and the term made its way into the English language with the translation of his book in 1876. The term 'human ecology' was first used by Robert Park and Ernest Burgess in their *Introduction to the Science of Sociology* (1924). For the first several decades after the term was first introduced, there was little agreement among sociologists about its meaning and focus (Alihan, 1938; Gettys, 1940; Firey, 1945). Even to this day, there are an assortment of approaches and orientations.

Freese (2001) has noted that the approaches to human ecology mainly differ according to the preferred vantage points of the scholars. He has discussed three orienting perspectives: (1) the biological perspective began with Haeckel who, following Darwin, defined ecology within the context of evolution and the "adaptations of organisms to the biotic and abiotic features of their environments" (Freese, 2001: p. 6974); since humans are living organisms, human ecology then emerged, although biologists per se have taken only a passing interest in the ecology of humans; (2) the metaphorical approach began with theories of community succession that biologists used in studying natural communities; Freese (2001) has written that some sociologists have used 'nature' as a metaphor to understand social organization; and (3) the ideological approach started in the 1960s with the genesis of the modern environmental movement worldwide; here the "biological and metaphorical interpretations of ecology are freely interwoven and supplemented with value judgments" (Freese, 2001: p. 6974) more or less in line with ecological science.

The major treatment of human ecology in sociology is Amos Hawley's book, *Human Ecology: A Theory of Community Structure* (1950), which to this day is the definitive exposition of the field. Hawley's book sets out the subject matter of human ecology and its approach. According to Hawley, human ecology deals with "how growing, multiplying beings maintain themselves in a constantly changing but ever restricted environment" (1950: p. 66). For human populations, this requires examining the ways in which individuals act collectively to achieve more effective use of their habitat (Poston and Frisbie, 2005).

Despite this clear and unambiguous sociological statement, some sociologists have ascribed to human ecology perspectives that are inconsistent with Hawley's thinking and that of McKenzie (1924, 1934, 1968), his predecessor and teacher. Here are three examples.

First, the sociobiologist Pierre van den Berghe has noted that "sociologists who claim to be ecologists … have reduced this specialty to a pedestrian kind of social geography (where) they largely plot social characteristics of people on maps" (1990: p. 174). Second, sociologists John Logan and Harvey Molotch have written that "in human ecology, spatial relations are the analytical basis for understanding urban systems" (1987: p. 4). And, third, the social theorist Manuel Castells has discussed the parallels between Marxian and ecological thinking and has observed that the results obtained by ecology have no more value for establishing

a theory of space than a mass of sociocultural correlations (1979: pp. 122–123). Not only is Castells' comment misguided, it does not capture the important materialistic and organizational similarities and differences between Marxist and ecological theory (see Hawley, 1984).

The above representations of human ecology as focusing on spatial relations are due in part to the unfortunate statement of McKenzie (1924) when he defined human ecology as the "study of the spatial and temporal relations of human beings as effected by the selective, distributive and accommodative forces of the environment." Hawley noted that although this simple, lucid statement inspired a great amount of empirical investigation, it caused human ecology to be regarded as nothing more than the descriptive study of spatial distributions, an outcome that McKenzie later noted was a misplacement of emphasis. Attention to spatial patterns, McKenzie recorded in his notes, should be subordinate and incidental to the analysis of sustenance relations (see Hawley's remarks in McKenzie (1968: pp. xiii–xiv); Poston and Frisbie, 1998, 2005).

There are other examples of the misuse or misunderstanding by social scientists of human ecology. Some refer to human ecology as studies using spatial rather than individual units of analysis (Robinson, 1950), as analyses of the physical features of geographic and built-up areas (Zorbaugh, 1929; Suttles, 1972), or as the factor analyses of the characteristics of aggregate units, i.e., factorial ecology (Berry and Rees, 1969).

These illustrations all exemplify Duncan's observation that "the term ecology is sometimes applied rather casually – even irresponsibly. [Frequently] studies adopting the label bear only a tenuous relationship to any systematic, scientific conception of the field" (Duncan, 1959: p. 680).

Sociological Human Ecology

Human ecology is a field of study grounded in the four referential constructs of population, technology, organization, and environment. The unit of analysis is the human population, circumscribed more or less in a territorial fashion. Its major assumptions are that populations have unit character and integrity, and that properties and attributes of these populations are more than the summation of their component parts (Poston and Frisbie, 1998, 2005).

Human ecology is concerned with the organizational aspects of human populations that arise from their sustenance-producing activities. These activities are necessary for the collective existence of the populations and must be adapted to the changing conditions confronting them. Included are an ever-changing and mediating environment, their technological repertoires, and the size, composition, and distribution of the populations themselves (Duncan, 1959; Frisbie and Poston, 1975, 1978a, 1978b; Poston, 1980, 1981; Poston and Frisbie, 1998, 2005).

Human ecologists address questions such as, what are the structural arrangements that characterize a population's sustenance-related endeavors? Under what conditions does one form of sustenance structure appear rather than another? What are the consequences for populations of varying configurations of sustenance-producing activities?

The answers lie in the fact that populations survive by virtue of collective organization. Human ecology is concerned with the determinants and consequences of sustenance organization, a consideration, by the way, that addresses the interplay between human ecology and demography. Thus, it is not surprising that much of the empirical literature of human ecology focuses on demographic applications, a point that will be made more evident below.

Human ecology offers demography an aggregate perspective for the analysis of the demographic processes. A fundamental tenet of human ecology is that a population redistributes itself through the vital processes and migration to achieve a balance or equilibrium between its size and life chances (Hawley, 1968: p. 331; also see Davis (1963)). Duncan (1959: p. 708) also emphasized the important ecological connections between organization and population size. Hawley (1950) noted that human populations will adjust their size through any of the demographic processes to maintain an equilibrium with their sustenance organization. Stated in another way, "demographic structure contains the possibilities and sets the limits of organized group life" (Hawley, 1950: p. 78; see also Poston, 1983; Poston and Frisbie, 1998, 2005).

Human Ecology's Four Rubrics

The author noted above that human ecology is grounded in the four referential constructs of population, technology, organization, and environment and now discusses each of these rubrics separately.

Organization

It is not an overstatement that organization is the fundamental element of the subject matter of human ecology. This is so because it is a social organization that mediates the balance among population size, growth, and distribution and the natural environment upon which it depends (Micklin, 1973). Human ecology is concerned with the organizational aspects of human populations arising from their sustenance-producing activities (Frisbie and Poston, 1978b: p. 14). In fact, the two broad goals of human ecology are to establish (1) the causes and (2) the consequences of particular characteristics of sustenance organization in human populations (Gibbs and Martin, 1959: p. 33).

There is major agreement regarding the centrality of organization within human ecology (Duncan, 1959; Hawley, 1950; Gibbs and Martin, 1959; Micklin, 1973; Poston et al., 1984; Namboodiri, 1988, 1994; Poston and Frisbie, 1998, 2005). However, despite its central position in human ecology and in the ecological theory of migration, the idea of sustenance organization was for decades in a primitive state of development both conceptually and empirically. Indeed, most of the research on sustenance organization that ecologists conducted in the 1950s and 1960s treated the concept as if it referred solely to the division of labor. This occurred even though there was little in the ecological literature to warrant such a limitation.

The notion of organization in human ecology is multifaceted. A major dimension of sustenance organization involves what Hawley referred to as the "arrangement of differentiated parts suited to the performance of a given function or set of functions" (1950: p. 178). This is sustenance differentiation, i.e., the extent to which the population is differentiated in its sustenance activity.

Sustenance differentiation consists of two elements: (1) the number of activities and (2) the degree of uniformity in the distribution of the population across the activities. A high degree of sustenance differentiation obtains when there are a relatively large number of activities characterizing the population and when the population members are evenly distributed across these activities (Gibbs and Poston, 1975).

Scholars since Durkheim (1893 [1960]) have included this dimension as a major component of the division of labor. There are many measures of sustenance differentiation, six of which have been elaborated by Gibbs and Poston (1975).

Another dimension of sustenance organization is functional interdependence; it can be combined with sustenance differentiation to form the other side of the division of labor (Gibbs and Poston, 1975). The degree of functional interdependence in a population depends on (1) the number of exchange linkages, (2) the variety of products involved, and (3) the volume of exchange flows (Eberstein and Frisbie, 1982). Empirical indicators of functional interdependence are often based on commodity-flow data.

A third dimension of sustenance organization is the volume of sustenance produced by the population, that is, the degree of productivity of the particular configuration of sustenance activities.

A fourth dimension of sustenance organization is the degree of efficiency of the sustenance organization. Given the level of sustenance produced, how efficiently does this occur? How much effort is required to produce the sustenance, whatever its volume?

A final structural characteristic or dimension of sustenance organization is the degree to which population members are engaged in sustenance-related pursuits (Poston and Johnson, 1971; Martin and Poston, 1972, 1976). What patterns of utilization of population members characterize the organization of one ecological unit versus another, especially with regard to ascribed statuses? How fully realized are the potential contributions of population members? To what extent do inequalities exist in the population by ascribed statuses? The degree to which populations differentiate by ascribed statuses in allocating sustenance roles to their members is an important dimension of sustenance organization, especially if the analyst is interested in sustenance productivity and other input-related functions (Poston and Frisbie, 1998, 2005).

Population

It goes without saying that of the four ecological concepts, population is the most advanced in terms of conceptual and operational detail. This is easily understood since an entire specialization, demography, is devoted to the study of population characteristics and dynamics. However, with few exceptions (Poston and White, 1978; Frisbie and Poston, 1978b; Namboodiri, 1994; Poston and Bouvier, 2010), ecologists seldom examine dimensions of the population as influences on population redistribution. Yet it is well known from demographic research that such population variables as age, race, and sex composition have predictable effects on demographic outcomes, especially on migration. Given the very advanced nature of the population rubric, additional attention is not necessary. The details about this rubric may be viewed in any introductory demography text (Poston and Bouvier, 2010).

Technology

Of the four basic ecological categories, technology is the most critical for the adaptation of human populations. Lenski has noted that technology is the 'prime mover' in the process of social change and adaptation for at least three reasons: (1) it sets the boundaries for feasible social and economic options; (2) technological change appears to be more easily accepted by the population than change in organization or ideology; (3) it is "easier to compare the effects of alternative tools or techniques than it is to compare the effects of alternative systems of social organization or alternative ideologies" (Lenski, 1970: p. 102; Poston and Frisbie, 1998, 2005).

The concept of technology is prominent in ecological and other macrolevel sociological theories. And there is a consensus with respect to its definitions. Three dimensions figure prominently in the definitions: material features (tools, capital equipment, and machines), information (knowledge, techniques, and scientific discovery), and energy. These are the same three ecosystem 'commodity' flows that Duncan (1964) identified as basic to the survival of populations.

Scholars have given only minimal attention to applying the technology component of the ecological complex to the study of populations below the societal level. As a consequence, there are few guidelines to suggest points of departure in specifying particular technological applications as foci.

One approximation toward conceptualization is the belief that one of the long-recognized technological keys to the establishment and growth of population aggregates is the presence and development of adequate transportation facilities. More than 100 years ago, Cooley (1894 [1930]: pp. 75–83) observed that population and wealth will tend to come together wherever there is a break or an interruption in routes of transportation. The development of transportation facilities partially determines industrial concentration and influences the expansion of local populations (Hawley, 1981). Since the availability of transportation is a major determinant of the ease of access of a population to its environment, a population's ability to compete with other populations, and the efficiency of sustenance extraction, an important dimension of technology, should involve mobility facilitating technology (Poston and Frisbie, 1998, 2005).

Two empirical indicators of this dimension of technology are the presence of an interstate highway crossing a county (or state or province) and the intersection in the area of two or more interstate highways. While these measures may be 'obvious,' the obviousness of their influence does not imply either triviality or simplicity. In fact, there is a large literature dating back to the 1960s that testifies both to the importance

and complexity of the impact of interstate arteries on subarea population change in general and change due to migration in particular (Dickinson, 1964; Wheat, 1969; Gauthier, 1970; Fuguitt and Beale, 1976; Briggs, 1980; Lichter and Fuguitt, 1980).

A second kind of technological application deals with the acquisition of sustenance. At a minimum, ecologists need to develop indicators of this dimension that reflect technological inputs affecting both primary and transformative sustenance activities. Many demographers, for instance, have noted that areas for which agricultural enterprise constitutes a major economic base are apt to experience demographic, especially migration, losses as agricultural production becomes increasingly mechanized and productive and capital intensive. However, previous research has found that (1) where production is highly land intensive or (2) where large volume and capital intensive production of food and fiber predominate, positive demographic changes predominate (Frisbie and Poston, 1978b).

Regarding the transformative component of sustenance acquisition, a useful indicator of the employment of available technology is new capital expenditures. These will index at least the hardware and capital-equipment dimension of technology in the manufacturing sector, i.e., the capital goods, equipment, and machines that figure prominently in the definitions of technology cited above. Of course, it is clear that capital may be substituted for labor, so that high levels of new capital expended might well mean a leveling off, if not an outright reduction in, local employment opportunities (Poston and Frisbie, 1998, 2005).

Environment

In human ecological terms, the environment is defined as "whatever is external to and potentially or actually influential on a phenomenon under investigation" (Hawley, 1968: p. 330). The concept of environment occupies a central position in the general theoretical framework of human ecology mainly because the environment is the ultimate source of sustenance for a population (Hawley, 1968: p. 330). However, little empirical research in sociological human ecology takes the environment directly into account, perhaps because of its breadth. That is, by definition, the environment "has no fixed content and must be defined anew for each different object of investigation" (Hawley, 1968: p. 330). In fact, some hold that the environment is the "least well conceptualized of the variables constituting the ecological complex" (Berry and Kasarda, 1977: p. 14).

However, close scrutiny of the ecological treatment of the environment reveals an implicit specificity not apparent in the above general definition. The environment comprises not everything external to the phenomenon of interest, but only those externalities that, by virtue of the limits they set on the acquisition of sustenance, affect the life chances of an organized population with a given technological repertoire. As Leo Schnore noted over 50 years ago, "the environment is viewed as a set of limiting conditions, which may be narrow or broad, depending upon the technological devices and modes of organization that prevail in a given population" (Schnore, 1958: p. 628; see also Michelson, 1970). Therefore, the human ecologist should narrow the arena of inquiry to those factors that, in light of existing technology, serve as limiting (or enabling) resources for the adaptation and growth of populations.

The author notes here that it is not useful to consider social and economic activities (or aberrations) of local populations to be part of their environment. Certain of these activities, for example, employment in given industries, are best viewed as aspects of ecological organization. Others, such as crime and deviance, rates of mortality and morbidity, unemployment, education, and income levels, are best conceived as indicators of different aspects of life chances that emerge from a population's organized efforts to adapt to the environment. In a real sense, the latter variables tend to indicate the degree of success or failure of the adaptive process. In short, they may reveal a disequilibrium between population and life chances. As such, they should be useful in helping to account for variation in a demographic process such as migration and thus should be included in models designed to explain migration. But they should not be conceptualized as aspects of the environment.

Despite the difficulties that arise in attempts to give conceptual and operational substance to the concept, it is clear that the ecological environment has two broad and distinct dimensions: the physical and the social. Hawley has written that the environment "includes not only the physical and biotic elements of an occupied area but also the influences that emanate from other organized populations in the same and in other areas. In certain circumstances the latter acquire a more critical importance than the former" (1981: p. 9).

Specifically, Hawley distinguished two dimensions, the biophysical and the ecumenic. The "former includes physiographic features, climate, soil characteristics, plant and animal life, mineral and other materials," and so forth. In contrast, the ecumenic refers to the "ecosystems or cultures possessed by peoples in adjacent areas and beyond" (Hawley, 1986: p. 14).

Poston et al. (2009) have undertaken an analysis among the states of the United States of the effects of physical climate on net migration. They gathered data on 11 different climate variables and used factor analysis to reduce them to the three dimensions of temperature, humidity, and wind. They showed that the temperature and humidity dimensions were significantly associated with migration, even after controlling for the effects on migration of factors dealing with ecological organization, the social environment, and population (see also Poston and Bouvier, 2010).

Conclusion

In this article, the author attempted to distinguish human ecology from ecology, and then to focus on sociological human ecology as mainly elaborated by Hawley (1950, 1998). The author outlined the ecological orientation, and also introduced several examples of applications of human ecological theory to empirical investigations of the demographic processes.

It was necessary to first set out the general orientation of sociological human ecology, mainly because of the fact that even today, despite the immense number of publications providing evidence to the contrary, the field is still

misunderstood by many sociologists and social scientists to be either a descriptive exercise or any kind of aggregate analysis. The author showed that some still hold that human ecology represents spatial or aggregate investigations of human phenomena. This representation minimizes considerably the rich sociological context of human ecology and indicates misunderstanding, perhaps even ignorance, of its orientation and subject matter.

The broad theoretical purview of human ecology may be distinguished from a narrower focus of ecological theory on demography. Human ecology is concerned with the organizational aspects of human populations that arise from their sustenance-producing activities. For the purposes of this selection, the author noted that human ecology offers demography a specific aggregate perspective for the analysis of the demographic processes (Poston and Bouvier, 2010).

The last part of this section outlined and articulated the theoretical and conceptual bases of the four referential constructs of human ecology, namely, organization, population, technology, and environment. The author reviewed some of the empirical and theoretical literature spanning more than five decades and showed in particular demographic models of migration benefit from use of the ecological perspective. Accentuated were the explicitly sociological features of the ecological perspective in a demonstration of its fruitful employment in demographic investigations. The strictly spatial studies that so many have thought to be ecological not only are not ecological, but also are usually not sociological. Moreover, they are theoretically lacking and are of little utility for demographic investigations (Poston and Frisbie, 1998, 2005).

See also: Community Sociology; Community, Social Context of: The US Case; Ecology of Aging; Environmental Sociology; Labor, Division of; Migration and Development; Organizational Ecology; Residential Segregation in the United States; Urban Sociology.

Bibliography

Alihan, M.A., 1938. Social Ecology: A Critical Analysis. Columbia University Press, New York.

Berry, B.J.L., Kasarda, J.D., 1977. Contemporary Urban Ecology. Macmillan, New York.

Berry, B.J.L., Rees, P.H., 1969. The factorial ecology of Calcutta. American Journal of Sociology 74, 445–491.

Briggs, R., 1980. The Impact of Interstate Highway System on Nonmetropolitan Growth. Final Report. U.S. Department of Transportation, Washington, D.C.

Castells, M., 1979. The Urban Question: A Marxist Approach. MIT Press, Cambridge.

Cooley, C.H., 1894 [1930]. The theory of transportation. In: Angell., R.C. (Ed.), Sociological Theory and Social Research. Holt, Rinehart, and Winston, New York, pp. 75–83.

Davis, K., 1963. The theory of change and response in modern demographic history. Population Index 29, 345–366.

Dickinson, R.E., 1964. City and Region. Routledge & Kegan Paul, London.

Duncan, O.D., 1959. Human ecology and population studies. In: Hauser, P.M., Duncan., O.D. (Eds.), The Study of Population. University of Chicago Press, Chicago, pp. 678–716.

Duncan, O.D., 1964. Social organization and the ecosystem. In: Faris., R.E.L. (Ed.), Handbook of Modern Sociology. Rand-McNally, Chicago, pp. 37–82.

Durkheim, Emile, 1893 [1960]. The Division of Labor in Society. The Free Press, New York.

Eberstein, I.W., Frisbie, W.P., 1982. Metropolitan function and interdependence in the U.S. urban system. Social Forces 60, 676–700.

Firey, W., 1945. Sentiment and symbolism as ecological variables. American Sociological Review 10, 140–148.

Freese, L., 2001. Human ecology. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Elsevier, New York, NY, pp. 6974–6978.

Frisbie, W.P., Poston Jr., D.L., 1975. Components of sustenance organization and nonmetropolitan population change: a human ecological investigation. American Sociological Review 40, 773–784.

Frisbie, W.P., Poston Jr., D.L., 1978a. Sustenance differentiation and population redistribution. Social Forces 57, 42–56.

Frisbie, W.P., Poston Jr., D.L., 1978b. Sustenance Organization and Migration in Nonmetropolitan America. Iowa Urban Community Research Center, University of Iowa, Iowa City.

Fuguitt, G.V., Beale, C.L., 1976. Population Change in Nonmetropolitan Cities and Towns. USDA Economic Research Service. AER-323.

Gauthier, H.L., 1970. Geography, transportation, and regional development. Economic Geography 46, 612–619.

Gettys, W.E., 1940. Human ecology and social theory. Social Forces 18, 469–476.

Gibbs, J.P., Martin, W.T., 1959. Toward a theoretical system of human ecology. Pacific Sociological Review 2, 29–36.

Gibbs, J.P., Poston Jr., D.L., 1975. The division of labor: conceptualization and related measures. Social Forces 53, 468–476.

Haeckel, Ernst, 1868 [1876]. The History of Creation, or, the Development of the Earth and Its Inhabitants by the Action of Natural Causes: A Popular Exposition of the Doctrine of Evolution in General, and of that of Darwin, Goethe and Lamarck in Particular (Translated from the German by E.R. Lankester). Henry S. King & Co, London.

Hawley, A.H., 1950. Human Ecology: A Theory of Community Structure. Ronald Press, New York.

Hawley, A.H., 1968. Human ecology. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences. Crowell, Collier and Macmillan, New York, pp. 328–337.

Hawley, A.H., 1981. Human ecology: persistence and change. In: Short., J.F. (Ed.), The State of Sociology: Problems and Prospects. Sage, Beverly Hills, pp. 119–140.

Hawley, A.H., 1984. Human ecological and Marxian theories. American Journal of Sociology 89, 904–917.

Hawley, A.H., 1986. Human Ecology: A Theoretical Essay. University of Chicago Press, Chicago.

Hawley, A.H., 1998. Human ecology, population, and development. In: Micklin, M., Poston Jr., D.L. (Eds.), Continuities in Sociological Human Ecology. Plenum Press, New York, pp. 11–25.

Lenski, G.E., 1970. Human Societies. McGraw-Hill, New York.

Lichter, D.T., Fuguitt, G.V., 1980. Demographic response to transportation innovation: the case of the interstate highway. Social Forces 59, 492–512.

Logan, J.R., Molotch, H.L., 1987. Urban Fortunes: The Political Economy of Place. University of California Press, Berkeley.

Martin, W.T., Poston Jr., D.L., 1972. The occupational composition of White females: sexism, racism and occupational differentiation. Social Forces 50, 349–355.

Martin, W.T., Poston Jr., D.L., 1976. Industrialization and occupational differentiation: an ecological analysis. Pacific Sociological Review 19, 82–97.

McKenzie, R.D., 1924. The ecological approach to the study of the human community. American Journal of Sociology 30, 287–301.

McKenzie, R.D., 1934. The field and problems of demography, human geography and human ecology. In: Bernard, L.L. (Ed.), The Fields and Methods of Sociology. R. Long & R. R. Smith, New York, pp. 52–66.

McKenzie, R.D., 1968. In: Hawley., A.H. (Ed.), Roderick D. McKenzie on Human Ecology: Selected Writings. University of Chicago Press, Chicago.

Michelson, W., 1970. Man and His Environment. Addison-Wesley, Reading, MA.

Micklin, M., 1973. Introduction: a framework for the study of human ecology. In: Micklin, M., Hinsdale, I.L. (Eds.), Population, Environment and Social Organization: Current Issues in Human Ecology. Dryden Press, pp. 2–19.

Namboodiri, K., 1988. Ecological demography: its place in sociology. American Sociological Review 53, 619–633.

Namboodiri, K., 1994. The human ecological approach to the study of population dynamics. Population Index 60, 517–539.

Park, R.E., Burgess, E.W., 1924. Introduction to the Science of Sociology. University of Chicago Press, Chicago.

Poston Jr., D.L., 1980. An ecological analysis of migration in metropolitan America, 1970–75. Social Science Quarterly 61, 418–433.

Poston Jr., D.L., 1981. An ecological examination of southern population redistribution, 1970–75. In: Poston Jr., D.L., Weller, R.H. (Eds.), The Population of the

South: Structure and Change in Social Demographic Context. The University of Texas Press, Austin, pp. 137–154.

Poston Jr., D.L., 1983. Demographic change in nonmetropolitan America in the 1960s and 1970s: population change versus net migration change. The Rural Sociologist 3, 28–33.

Poston Jr., D.L., Bouvier, L.F., 2010. Population and Society: An Introduction to Demography. Oxford University Press, New York.

Poston Jr., D.L., Frisbie, W.P., 1998. Human ecology, sociology and demography. In: Micklin, M., Poston Jr., D.L. (Eds.), Continuities in Sociological Human Ecology. Plenum Press, New York, pp. 27–50.

Poston Jr., D.L., Frisbie, W.P., 2005. Ecological demography. In: Poston Jr., D.L., Micklin, M. (Eds.), Handbook of Population. Springer Publishers, New York, pp. 601–623.

Poston Jr., D.L., Frisbie, W.P., Micklin, M., 1984. Sociological human ecology: theoretical and conceptual perspectives. In: Micklin, M., Choldin., H.M. (Eds.), Sociological Human Ecology: Contemporary Issues and Applications. Westview, Boulder, pp. 91–123.

Poston Jr., D.L., Gotcher, D.J., Gu, Y., 2009. The effect of climate on migration, United States, 1995–2000. Social Science Research 38, 743–753.

Poston, D.L., Johnson, G.C., 1971. Industrialization and professional differentiation by sex in the metropolitan southwest. Social Science Quarterly 52, 331–348.

Poston Jr., D.L., White, R., 1978. Indigenous labor supply, sustenance organization and population redistribution in nonmetropolitan America: an extension of the ecological theory of migration. Demography 15, 637–641.

Robinson, W.S., 1950. Ecological correlations and the behavior of individuals. American Sociological Review 15, 351–357.

Schnore, L.F., 1958. Social morphology and human ecology. American Journal of Sociology 63, 620–634.

Smith, R.L., Pimm, S.L., 2013. "Ecology." Encyclopædia Britannica Online Academic Edition. Encyclopædia Britannica Inc., Chicago accessed 24.05.2013. http://www.britannica.com/EBchecked/topic/178273/ecology.

Suttles, G., 1972. The Social Construction of Communities. University of Chicago Press, Chicago.

van den Berghe, P.L., 1990. Why most sociologists don't (and won't) think evolutionarily. Sociological Forum 5, 173–185.

Wheat, L.F., 1969. Highway Research Record 277, 9–24.

Zorbaugh, H., 1929. The Gold Coast and the Slum. University of Chicago Press, Chicago.

Human Evolutionary Genetics

Pascale Gerbault and Mark G Thomas, Research Department of Genetics, Evolution and Environment, University College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.L. Mountain, volume 10, pp. 6984–6991, © 2001, Elsevier Ltd.

Abstract

Traditionally, our knowledge of human evolution has come from the fossil and material culture records, and has been studied by paleontologists, anthropologists, archaeologists, anatomists, and – to some extent – linguists. In the past 25 years genetics has made substantial contributions to this field. While much of this has been driven by advances in molecular biology techniques (i.e., the methods used to obtain genetic data), the principles underlying how genetic data can be used to make inferences about our evolutionary past come from the field of population genetics. In this article we discuss how population genetics has been used to address two sets of questions regarding human evolution: (1) when and where did human populations originate (demographic history) and (2) to what extent and in what ways have humans adapted to changes in our ecology by natural selection (adaptation history).

Introduction to Population Genetics

Genetic information is carried by the sequence of bases in deoxyribonucleic acid (DNA). One of the most important features of the DNA molecule is its ability to be replicated, and so be passed on from a cell to its daughter cells, and from one generation to the next, mostly unchanged. Much of human DNA (referred to as the human genome) appears not to have a direct function, and is sometimes referred to as junk DNA. In the past it was generally thought that the only functionally important parts of our genome were genes – regions of DNA that provide information on how to make specific proteins – and regions near those genes that control their expression. However, in recent years other genomic regions have been shown to be important, such as those coding for functionally active ribonucleic acid (RNA) molecules.

During the replication process, some changes can occur in the DNA sequence; any such change is termed a mutation. Mutations can occur in reproductive and nonreproductive cells (i.e., germ and somatic cells, respectively). However, only mutations occurring in germ cells are heritable and therefore a substrate for evolutionary processes. Mutations give rise to new genetic variants, called alleles (Kimura, 1971), and the overall constellation of alleles in an individual is known as its genotype. It should be noted that mutations are relatively rare – even on the evolutionary timescale – so most of the DNA sequences in our genome are identical between individuals, and indeed between humans and other primates. However, the human genome is very large (around 3 billion base pairs) so while comparatively rare, sites in the genome (loci) that are variable between individuals are still numerous. In addition to mutation, a process called recombination can shuffle the distribution of alleles along DNA sequences into new combinations (see later).

The field of population genetics is chiefly concerned with describing and understanding the distribution and fate of genetic variation. Ultimately, an allele only has two fates, loss (from the population) or fixation (i.e., loss of variation at that locus). However, in the intervening time between mutation and loss or fixation a locus will be variable or polymorphic. The distribution and fate of genetic variation in a population is shaped by three processes: mutation (including recombination), genetic drift, and natural selection. Mutation, as explained above, gives rise to variation, natural selection favors (and so increases/decreases in frequency) particular alleles, and genetic drift is the random sampling of alleles from one generation to the next through reproduction, and leads to random changes in allele frequencies through time.

When a new mutation occurs in the protein-coding region of a gene, or in other functionally important parts of the genome, it may change the biological characteristics (phenotype) of an individual. While phenotypes are only partly determined by genotype, they can affect the overall fitness of individuals, and hence lead to differential survival and reproduction. This process is known as natural selection, and it can act on phenotypes related to early development and survival to reproductive age (viability), on success in attracting a mate (sexual selection), on the consequent ability to fertilize (gamete selection), or on the number of progeny produced (fecundity). The sum of these various selection stages constitutes the fitness of a phenotype, which is partly dependent on environmental variables. The selection coefficient is the relative fitness of a genotype or genotypes in relation to others.

A new mutation might not affect the underlying phenotype. Such mutations give rise to alleles that are neither advantageous nor disadvantageous, and are said to be neutral. The fate of neutral mutations is determined by random genetic drift – the chance passing on alleles to future generations. Genetic drift reduces genetic diversity through time since it ultimately leads to the loss or fixation of alleles, but its effect depends strongly on the effective size of the population considered.

The effective population size (N_e) is the size of a randomly mating population that would show the same degree of genetic drift as the actual population (Wright, 1931). The relationship between effective and census population sizes (N_e and N, respectively) varies depending on certain properties of the population considered (Rice, 2004). For example, if we assume

that individuals mate randomly and generations are not overlapping, then when N changes through time it can be shown that N_e is the harmonic mean of N over a given period of time. In such conditions, N_e is disproportionately affected by small values of the fluctuating population size, and N_e will be smaller than the average value of N across generations. However, these assumptions do not always hold (Rice, 2004).

The effect of genetic drift can be illustrated by looking at the chance an allele has of being fixed in a population. For autosomes (chromosomes other than the sex chromosomes: X or Y) the probability of fixation of a new allele equals its relative frequency in the population and is therefore $1/(2N_e)$; $2N_e$ because every individual carries two copies of each autosome. For the Y chromosome (one copy, only carried by males) or the mitochondrial DNA (mtDNA; only transmitted by females) this probability is approximately $1/(0.5N_e)$, and for the X chromosome (2 copies in females, 1 copy in males) this probability becomes approximately $1/(1.5N_e)$. This highlights that the larger the population the less is the effect of drift since the smaller the chance a new allele has of being fixed in this population.

An appreciation of the importance of genetic drift emerged from a major branch of population genetics called Neutral Theory (Kimura, 1971). Prior to this it was thought that natural selection was the most important force shaping the fates of alleles, and so evolution. Neutral Theory itself arose out of the realization that there was far more genetic variation in natural populations (including humans) than could be maintained by natural selection alone. It postulates that most mutations give rise to alleles that are either disadvantageous – so lost rapidly from the population – or are selectively neutral, or sufficiently nearly neutral that drift rather than selection is the main force governing their fate. Neutral Theory does admit the possibility of selectively advantageous alleles arising by mutation, but assumes that such events are sufficiently rare that they contribute little to overall patterns of genetic variation (Kimura, 1968).

One interesting prediction of Neutral Theory relates to the rate of change of DNA sequences through time. If selection plays a major role in the determination of survival of new mutations, then a constant mutation rate should not lead to a constant rate of evolution (substitution rate, or the rate of fixation of new alleles). Instead the substitution rate would be expected to vary episodically through time, as selection intensity acting on different traits varies. Alternatively, if the Neutral Theory holds then it is expected that the substitution rate is set by the underlying mutation rate and the proportion of new alleles that are selectively disadvantageous (the selective constraint). Given that the mutation rate (μ) and selective constraint for any given gene is approximately constant through time, this predicts that the rate of genetic change is also constant over all evolutionary lineages. This is known as the molecular clock hypothesis (Kimura, 1968; Zuckerkandl and Pauling, 1965) and it permits the estimation of lineage/species divergence times using genetic data.

Neutral Theory has had a profound effect on population genetics; so much so that while there are undoubtedly examples where natural selection has shaped the fate of alleles, neutrality is now widely accepted as the null hypothesis – the default assumed status of alleles at polymorphic loci.

Given that patterns of genetic variation in populations are shaped by mutation, drift (itself shaped by demographic history) and natural selection, it stands to reason that those patterns of variation in populations contain information about mutation, demographic history, and natural selection. However, extracting that information to make inferences about our evolutionary past is not trivial, primarily because any particular pattern of variation in one or a set of populations can be the result of a very wide range of different evolutionary scenarios (equifinality).

Population Genetics Inferences: Demographic History and Natural Selection

One of the advantages of the neutral model of evolution is that it makes predictions about the relationship between patterns of genetic variation, the mutation rate, and demographic parameters, such as the effective population size (Kimura, 1968; Zuckerkandl and Pauling, 1965). These predictions include: (1) the expected level of diversity within a species at equilibrium is a function of the mutation rate and the effective population size; (2) as two species diverge, they accumulate differences at the same rate (i.e., substitution rate) at which neutral mutations arise; and (3) the expected allele frequencies in a sample of DNA sequences is a function of the effective population size and of the sample size.

These predictions can then be compared to empirical data (Zuckerkandl and Pauling, 1965). From there, any significant deviations from neutral expectations would suggest the locus or loci under study did not evolve neutrally but instead were subjected to other forces, including natural selection. In other words, the study of natural selection is mainly based upon tests of significant deviation from the null hypothesis of neutral evolution, rather than any direct measurements of selection. However, an important complicating factor is that real populations typically have complicated demographic histories that cannot be collapsed to a single population size parameter, and some demographic histories can lead to patterns in genetic variation that mimic those formed by natural selection (e.g., Currat et al., 2006).

A wide-range of methods have been developed to detect genomic signature of natural selection. These methods differ primarily in the nature of the genetic data they consider and the evolutionary time frame they are sensitive to. Allele frequencies and other measures of genetic variation can be used to estimate the extent and nature of demographic changes, and population structuring, or departures from neutrality.

The extent of Hardy–Weinberg disequilibrium can be used as a crude test of neutral evolution. The Hardy–Weinberg theorem assumes an idealized population with an infinite number of randomly mating individuals, no selection, no mutation, and no migration, and predicts that for a single locus with two alleles, the three genotypes *AA*, *Aa*, and *aa* follow the proportions p^2, $2pq$, and q^2, where p and q are the initial relative frequencies of the two alleles *A* and *a*, respectively. Different types of natural selection will alter these proportions in distinct ways (Nielsen, 2005). However, any deviation from Hardy–Weinberg equilibrium can be due to various factors, such as population structure, and should thus be treated with caution

as evidence for selection. Importantly, natural selection affects only those selected alleles, while demographic factors and genetic drift act on the whole genome.

Classical methods for detecting selection are based on the distribution of allele frequencies in single nucleotide polymorphisms (SNPs), also called the allele- or site-frequency spectrum (SFS)-based methods. The allele-frequency spectrum can be 'unfolded' or 'folded' depending on whether the derived and the ancestral (i.e., fixed in other Apes) allele can be distinguished or not, respectively. This basically involves identifying the number of allele-frequency classes observed, and counting the number of loci falling in each frequency class. When considering the SFS negative selection tends to increase the proportion of low-frequency variants compared to neutral expectations (Nielsen, 2005).

Alternatively, a new mutation arising in a population may be advantageous, in which case it will increase in frequency in the population by *positive selection*. Much interest has focused on positive selection due to its association with adaptation and the evolution of new forms and functions. Over a prolonged period, positive selection increases the fixation rate of beneficial function-altering alleles. It can be detected by comparison of changes between species. The first test of selection based on detecting regions where patterns of variation depart from those expected under neutrality is the Hudson–Kreitman–Aguade (HKA) test (Hudson et al., 1987). It compares levels of diversity in different genes or genomic regions, calibrated by between-species divergence rates for the same regions, to test whether these levels are significantly increased or reduced in the region of interest, compared to theoretical expectation or empirical data for presumed neutral parts of the genome (Nielsen, 2005).

Balancing selection is a case of positive selection where the new variant is advantageous in combination of other alleles. Balancing selection thereby increases the proportion of intermediate-frequency variants with respect to neutral expectations, whereas positive selection increases the proportion of high-frequency variants.

A *selective sweep* is an example of positive selection where the new favored variant reaches a high frequency in the population, resulting in an overall decrease of the genetic variation at the selected locus, as well as in the genomic region surrounding it. A selective sweep tends to increase the proportion of low-frequency variants in respect to neutral expectations.

A widely-used type of a selection test based on the SFS is Tajima's *D* test (see Tajima, 1989). In cases of a selective sweep or purifying selection, this test detects an excess of low-frequency variants (indicated by negative values of *the* Tajima's *D* statistic) compared to what is expected under neutrality, or to empirical data for presumed neutral parts of the genome. Alternatively, in cases of balancing selection this test detects an excess of alleles of intermediate frequency (indicated by positive values of *D*). Other extensions of this type of tests are based on similar principles (e.g., Fay and Wu, 2000).

Selection can also differ between populations. This may happen due to adaptations to local environment, in which cases the level of population differentiation may increase for loci under different selection in different populations. The Lewontin–Krakauer (1973) test was one of the first neutrality tests to use the level of genetic differentiation between populations. This test rejects neutral evolution when differentiation between populations at specific locus is larger than that expected under a neutral model, or is outside the empirical range observed in presumed neutral regions of the genome. Other tests have been developed under the same principle (reviewed in Nielsen (2005)). Notably, Akey et al. (2002) used genome-wide data to look at the variation in F_{ST} (a traditional measure of allele-frequency difference between populations) in humans. Various statistical methods can use genome-wide scans to detect selective sweeps without prior hypotheses on candidate genes or regions under selection (Nielsen, 2005).

Another feature of selected genomic regions is an increase in the level of linkage disequilibrium (LD). When a new allele arises it is physically linked to alleles at other loci nearby on the same chromosome. This generates a nonrandom association of alleles known as LD, and initially, any new allele will be in complete LD with other nearby alleles. These combinations of alleles in a region of a chromosome are sometimes referred to as haplotypes. Over time, recombination breaks down the association between alleles at nearby loci, and the allele of interest will become associated with an increasing number of different haplotypes. LD also decreases with increasing genetic distance from the selected site as recombination shuffles allele combinations in proportion to distance away from it on the chromosome. Thus, the number and lengths of haplotypes associated with an allele of interest can act as a proxy for the age of that allele. If an allele has been positively selected then it will rise to high-frequency quicker than expected under genetic drift alone. It therefore follows that high-frequency alleles that are still in high LD with other nearby alleles (i.e., are recent in origin) are good candidates for selection. These principles have been used to develop a number of statistics for detecting recent and strong natural selection. For example, the extended haplotype test identifies tracts of homozygosity (identity in randomly drawn pairs of sequences) associated with a core haplotype using the 'extended haplotype homozygosity' (EHH) statistic (e.g., Voight et al., 2006). A haplotype containing an allele that has been positively selected is expected to display high EHH values and high frequencies. In contrast, haplotypes that reach high frequencies due to genetic drift are likely to have taken longer to reach those high frequencies and are therefore likely to have been subjected to more recombination and mutation events, and would consequently show lower values of EHH.

A fundamental problem with some methods of detecting genomic signatures of natural selection is the bias generated when genetic data are obtained not through direct sequencing, but by genotyping SNPs that have been discovered in other samples. This is called *ascertainment bias*. Patterns observed in the data, such as allele frequencies, population differentiation, and LD, all depend on the procedure used to discover these SNPs. For example, ascertained SNPs will almost invariably be biased toward more common variants, thus skewing statistics based in the SFS. Additionally, many SNPs were identified in only a subset of populations, particularly Europeans. When the SNP-discovery protocol is known, statistical methods can be used to correct for ascertainment biases to some extent (e.g., Voight et al., 2006).

More importantly, most methods for detecting selection are challenged by the confounding effects of demographic history

(e.g., Przeworski, 2002; Currat et al., 2006). For example, Tajima's *D* test can reject a neutral model of evolution if the population has undergone an expansion or was strongly structured (e.g., Przeworski, 2002) since an expansion can lead to an increase in the proportion of low-frequency variants, mirroring the effect of a selective sweep, and the population structure can lead to an increase in the proportion of intermediate-frequency variants, mirroring balancing selection. Allele surfing, a process whereby a rare allele is driven to high frequencies at the wave front of an expanding population, can also mirror a selective sweep (e.g., Currat et al., 2006). Alternatively, a preferential loss of low-frequency variants is expected to occur during a population bottleneck, thereby leading to an excess of intermediate-frequency variants, mirroring balancing selection.

In consequence, failure to reject a neutral model of evolution for a locus might be the result of a particular demographic history. Methods to account for the confounding effects of demography usually involve comparisons of distributions of genomic patterns of diversity at a locus of interest – such as selection–detection statistics – to those derived from the rest of the genome (which is presumed to be mostly neutral), or generation of distributions of statistics under specific demographic scenarios by simulation, and detection of outliers (e.g., Xue et al., 2009). These methods involve modeling specific demographic histories by computer simulation. Alternatively, a recently developed method for genome-wide data sets accounts for shared population history and gene flow by generating an empirical pattern of covariance in allele frequencies between populations from a set of markers (Coop et al., 2010). This is then used as a null model for identifying loci involved in local adaptation by looking at unusual correlations between allele frequencies and ecological variables (Coop et al., 2010).

An alternative approach to detecting selection makes use of advances in ancient DNA analysis to directly assess the rate of allele-frequency change through time (Wilde et al., 2014; Sverrisdottir et al., 2014). While this approach currently requires the assumption of population continuity between temporally distinct DNA samples (Sverrisdottir et al., 2014), it does have some key advantages over other methods. First, it is a direct method to test for selection, unlike others based only on contemporary DNA samples. Secondly, given sufficient data it has the potential to detect episodic changes in selection intensity through time. Thirdly, unlike some approaches it is not dependent on estimates of mutation or recombination rates (Wilde et al., 2014).

The Origins of Modern Humans: The Genetic Record

Population genetics is grounded in mathematics and probability theory (e.g., Kimura, 1968). The field has however become fundamentally data driven since the discovery of the ABO blood group system. Cavalli-Sforza and Edwards (1964) estimated genetic distances between five populations based on five different blood groups and inferred a tree in which Europeans were separated from an Afro-Asian lineage. A larger study analyzing 35 proteins linked Europeans and Asians to the exclusion of Africans, estimating a Europeans/Asians split ~55 000 years ago and an older divergence from Africans ~120 000 years ago (Nei and Roychoudry, 1974).

As genetic data increased, multivariate methods, such as principal component analysis (PCA), started to be used to summarize patterns of variation for multiple genetic markers (including HLA and blood group protein markers) over space (Cavalli-Sforza et al., 1994). Spatial gradients of allele frequencies were often interpreted as corresponding to hypothesized linguistic or cultural expansions. It has however been shown that spatial patterns of genetic variation can arise even when no major demographic events have occurred (Novembre and Stephens, 2008) but simply because similarity between locations tends to decay with geographic distance, as predicted under an isolation by distance (IBD) model.

One of the earliest influential genetic studies of human origins examined mtDNA variation from worldwide population sample of 147 individuals (Cann et al., 1987). The inferred maximum parsimony tree indicated that the deepest branches were in Africa and a molecular clock estimate placed the deepest split, that is, the time to the most recent common ancestor, at just less than 200 000 years ago. Despite the caveats of this study, notably the 'African sample' was actually a sample of African Americans, and the maximum parsimony analysis used mid-point rooting, instead of the more reliable out-group rooting, these results were corroborated some years later by a more robust study (Ingman et al., 2000). Since then, estimates of the time to the most recent common ancestor (TMRCA) for mtDNA and the Y chromosome have been highly influential in studies of human evolution. Estimates for mtDNA TMRCA currently range around 140–240 kya (e.g., Behar et al., 2012) while, until recently, estimates for the Y chromosome TMRCA ranged between 60 and 140 kya (e.g., Wei et al., 2013).

These TMRCA estimates for mtDNA and the Y chromosome differ considerably and might initially seem contradictory. However, large differences are expected for two major reasons. First, the variance in expected TMRCAs is large compared to its mean, as predicted by coalescent theory (Kingman, 1982). Second, the mode of inheritance of mtDNA and the Y chromosome (maternally and paternally inherited, respectively) mean they may have different effective population sizes.

Coalescent theory is a retrospective mathematical model of gene genealogies for a sample of DNA sequences under a defined population history (Kingman, 1982). Demographic processes shape the pattern in which those lineages (i.e., branches on a genealogical tree) connect to one another. We say those lineages converge or 'coalesce' when this process is looked at backward through time. Gene coalescence can be thought of as drift run backwards through the genealogical tree. However, while drift and coalescence are conceptually related, simulating coalescence backward through time is considerably more computationally efficient than simulating drift forward through time, because it only considers the sample, not the whole population. Coalescent theory shows that the expected TMRCA is $2N_e$ in generations, and the final time interval, when the remaining two lineages coalesce (i.e., join) into the MRCA, represents more than half of the variance in the TMRCA. This is because coalescences take longer when there are fewer lineages and those times will tend to strongly affect the shape of the genealogy.

These high-expected variances in TMRCAs have been highlighted by the recent discovery of a Y chromosome lineage branching at the basal portion of the Y chromosome genealogical tree (Mendez et al., 2013). This Y chromosome lineage was identified in individuals of West Africa origin and recent West African descent, and was labeled A00 (Mendez et al., 2013). The updated Y chromosome genealogical tree provided a TMRCA estimate of 338 kya (95% confidence interval (237–581) kya) (Mendez et al., 2013). This estimate does not only predate current mtDNA TMRCA but also the earliest anatomically modern human fossils. This demonstrates that interpretation of TMRCA date estimates for a single non-recombining region of the genome should be treated with caution – they do not necessarily represent the founding dates of a population or species. Furthermore, because of the ubiquity of past migration and the stochastic process inherent to gene genealogies, it is difficult to infer the geographic origin of a genealogical lineage from its current distribution (Beaumont et al., 2010).

Regardless of the TMRCA estimates, studies of mtDNA variation (Cann et al., 1987; Ingman et al., 2000) and of nonrecombining regions of the Y chromosome variation (e.g., Ke et al., 2001) in modern populations are consistent with, and have been interpreted as strongly supporting a recent African origin of our species. This relies on the observation that Africa is the source of the deepest lineages and harbors the greatest diversity. The genetic diversity of both systems is characteristic of a rapidly expanding population or one that was subject to positive selection, that is, long terminal branches and excess of low-frequency polymorphisms. A similar signal was also identified in various autosomal regions (e.g., Voight et al., 2005). A demographic process where modern humans expanded out of Africa between 50 000 and 100 000 years ago and replaced other archaic humans has been shown to better explain the general pattern of genetic diversity observed (e.g., Fagundes et al., 2007).

Although classification systems vary, archaic humans refer to any *Homo*-related fossil remain that is not anatomically modern humans (*Homo sapiens sapiens*). Traditionally, and prior to the development of molecular genetics techniques, information on the morphology and behavior of our ancestors and the related species at different times came from fossil and archaeological evidence. One of the most widely known archaic human groups is Neanderthal (*Homo neanderthalensis*). It represents a morphologically distinct group, with robust morphology, large brains, and prominent brow ridges. These fossils date to between 250 000 and 39 000 years ago (Higham et al., 2014) and have been found in Europe and western and central Asia. In contrast, the earliest widely accepted fully modern human skull is Omo I, from Ethiopia, and dates to 195 000 years ago (McDougall et al., 2005).

Recently, the genomes of two archaic human forms, Neanderthals (Green et al., 2010) and Denisovans (Reich et al., 2010), have been partially sequenced, thereby providing some estimates of the relationships of ancient and modern humans. This was performed by first focusing on biallelic SNPs where two present-day humans carry different alleles and the archaic human genome carried the derived allele, that is, not matching chimpanzee (Green et al., 2010; Reich et al., 2010). Then a measure called *D(H1,H2, A, chimpanzee)* was computed to assess the difference in proportion of matching when the derived allele in the archaic human (*A*) genome matched the modern human genome *H1* more often than the modern human genome *H2*. *D* is positive if the archaic genome matches *H1* more often and negative if it matches *H2* more often. This measure led to an estimate of the proportion of Neanderthal ancestry in genomes as 1–4% (Green et al., 2010) and 4–6% of Denisovan ancestry in Melanesian genomes (Reich et al., 2010). It thus appears that Neanderthals and Denisovans did contribute some ancestry to non-African modern humans.

Adaptations and Detection of Positive Selection

We now inhabit radically different ecological (e.g., hot/cold and tundra/forest), cultural (e.g., variety of food resources and of their uses and social and mating systems) and demographic (high population densities and rates of gene flow) environments worldwide. Environmental changes have triggered human adaptation, a process that involves physiological, biochemical, and behavioral adjustments. Genetic adaptations are additional responses of humans to changing environment.

Genome-wide scans have the potential to generate lists of putatively selected genes that can be further studied from a functional perspective. The most striking examples of signals of selection include those identified in genomic regions involved in human pigmentation (e.g., Wilde et al., 2014) and for genetic polymorphisms conferring some resistance to malaria (Kwiatkowski, 2005) and other infectious diseases (e.g., Fumagalli et al., 2012). Both emphasize that the various environments we have evolved in, either in the form of climate factors (e.g., Hancock et al., 2010) – such as ultraviolet exposure – or of pathogens (e.g., Fumagalli et al., 2012) have driven components of our evolution. In this respect, evidence for correlations between patterns of genetic variation and environmental variables (e.g., Hancock et al., 2010), such as temperature, precipitation, solar radiation, latitude, or elevation, have been identified. However, identifying the precise selective pressure and assessing the functional advantage of candidate genetic variants remains a challenge.

EDAR

This challenge is well exemplified by genes involved in hair follicle development, which have shown higher levels of population differentiation than expected under neutrality. In particular, the ectodysplasin-A receptor gene (*EDAR*; MIM 604095), located on chromosome 2q12.3, plays a central role in generation of the primary hair follicle pattern. In humans a derived allele in the *EDAR* gene, called *EDAR370A*, results in a valine-to-alanine substitution at position 370 of the protein sequence. This allele is associated with thicker hair in East Asian populations (Kamberov et al., 2013) has shown a strong signal of positive selection using tests based on the SFS (e.g., Kelley et al., 2006), haplotype structure (Voight et al., 2006) and population differentiation (e.g., Xue et al., 2009). Using haplotype analysis, the *EDAR370A* allele has been dated to between 1133 and 73 996 years ago (Bryk et al., 2008), but the reasons for strong selection are not yet known.

The only selection hypothesis explicitly formulated so far is that this allele may have been beneficial under Asian climate conditions some 25 000 to 10 000 years ago; a time when the climate was significantly colder and dryer than present in eastern and northern Asia. This hypothesis proposes that the *EDAR370A* allele contributed to increase lubrication and humidification of exposed surfaces (face and scalp) during the dryer and colder Ice Age in East Asia (Chang et al., 2009).

This has however not been supported by a recent study reporting an *EDAR370A* mouse model (Kamberov et al., 2013). In this study *EDAR370A* mice showed (1) increased hair thickness; (2) a higher mammary branch density and smaller mammary fat pad area; and (3) an increase in eccrine sweat gland density (Kamberov et al., 2013). This study also showed that *EDAR370A* was associated with (4) tooth variation (including single and double shoveling of the upper incisors); and (5) a higher active eccrine sweat gland density in Han Chinese (Kamberov et al., 2013). A spatially explicit forward in time simulation model, integrating the origins and spread of farming populations and fitting to observed modern allele frequencies, suggested that *EDAR370A* arose in East Asia around 30 000 years ago (95% credible interval 13 175–39 575 years ago) with selection a coefficient among the highest estimated for the human genome (0.122 with a 95% credible interval 0.030–0.186).

This nonetheless does not explicitly tell us which of *EDAR370A* pleiotropic effect/s was/were the actual target of natural selection. For example, as *EDAR* affects the development of sweat glands and hair morphology, it may be hypothesized that *EDAR370A* alters thermoregulation, and possibly mate preference. Alternatively, mammary gland branching and/or fat pad size may have been the adaptive phenotype, either because of associated benefits on lactation (not-yet assessed) or because of increased mate preferences (Dixson et al., 2011). Identifying the selective pressure on *EDAR370A* and the target phenotype of this selection are made difficult by the various developmental pathways in which *EDAR* is involved.

LCT

Pathways involved in nutrient intake, digestion, and metabolism are essential to energy production and growth, and consequently to development and survival. Since the range expansions out of Africa, our species has undergone several dietary transitions, including a shift from a diet based on animal hunting and a broad range of food gathering to a diet based on farming domesticated plants and animals (Luca et al., 2010). This Neolithic transition had profound effects on our diet, particularly dietary breadth and carbohydrate content. The Neolithic transition dates back to about 11 000 years ago in its core regions and marks the transition from foraging (i.e., hunter-gathering) to food-producing (i.e., farming) societies. It is associated with changes that include an increasingly sedentary lifestyle; the development of alternative economies that focus on animal and/or plant domesticates; and technical innovations that include polished stone tools and pottery.

The dietary breadth of farming populations is known to be narrower than that of hunter-gatherers (Luca et al., 2010). Patterns of genetic variation in the gene coding for lactase (*LCT*) provide what is probably the strongest evidence of genetic adaptation to dietary specialization (Holden and Mace, 1997; Gerbault et al., 2011). Lactase is the enzyme that digests the milk sugar lactose. It is expressed in young mammals but its expression usually decreases after the weaning period is over; a trait known as lactase nonpersistence. However, in certain human populations many adults continue to express lactase throughout adulthood, a trait termed lactase persistence.

Large differences in the frequency of lactase persistence in different populations have been known for some time (Holden and Mace, 1997). More recently, at least four independent alleles associated with lactase persistence have been identified indicating convergent evolution in different geographic locations (Enattah et al., 2007; Gerbault et al., 2011). In Europe and southern and central Asia, a single allele (–13 910*T*) predominates, whereas in the Middle East and Africa three additional alleles (13 915*G*, 13 907*G*, and 14 010*C*) are commonly found in lactase-persistent individuals. Furthermore, the correlation between the distribution of lactase persistence and the distribution of pastoralism/dairying (Holden and Mace, 1997) provide strong support for a gene-culture coevolutionary process.

Selection coefficients on the -13 910*T* allele have been estimated to be between 1.4 and 19% using extended haplotype homozygosity (Bersaglieri et al., 2004), between 5.2 and 15.9% using spatially explicit simulation modeling (Itan et al., 2009), and around 2.4% using ancient DNA (Sverrisdottir et al., 2014), among the highest of the human genome over the past 30 000 years. Furthermore, the age estimates for the –13 910*T* allele are relatively recent: 2188–20 650 years ago using extended haplotype variation (Bersaglieri et al., 2004), 7475–10 250 years ago using closely linked microsatellite variation (Mulcare, 2006) and 6256–8683 years ago using simulation modeling (Itan et al., 2009). These age estimates bracket dates for the domestication of milkable animals and dairying.

Our knowledge of the evolution of lactase persistence in Europe has benefitted considerably from archaeological data, notably from the age of dairy animals at death (e.g., Vigne, 2008) and lipid residue analyses from potsherds (Evershed et al., 2008). Both have shown that dairying was an early feature of European Neolithic economies. By considering ancient DNA data on the occurrence of the –13 910*T* allele in combination with archaeological data on milk use, it appears that dairying was practiced before lactase persistence arose or became common(Burger et al., 2007).

Conclusion

The propensity of our species to colonize a wide range of environments illustrates our species plasticity and adaptability. While much of that plasticity and adaptability is underwritten by our capacity for cumulative culture (Powell et al., 2009), it now seems certain that some biological adaptation has occurred. Our ability to detect the genomic signatures of natural selection has improved dramatically in the last 10 years. This has been simultaneously driven by advances in the molecular techniques used to generate genetic data, by improvements in ancient DNA technologies, and critically, by

innovations in statistical inference and computer simulation modeling. However, challenges remain in disentangling the relative effects of mutation, recombination, migration, population structure, genetic drift, and natural selection on patterns of genetic diversity. Strong evidence for the adaptive status of a trait entails (1) evidence of differential fertility or mortality dependent on particular genetic variation; (2) evidence from *in vitro* and/or *in vivo* studies of functional differences between genotypes that affect the reproductive success of their carriers; and (3) evidence of concordance between the distribution of a genetic traits and the environmental factors that drive selective pressures. Because providing these three types of evidence remains a challenging task, few strong examples of natural selection exist. We have presented two of them, those involving *EDAR* and *LCT* genes.

As molecular techniques improve and become easier and cheaper to perform, more genetic information accumulates, bringing new challenges in handling and analyzing so much data (Pool et al., 2010). With improvements in our understanding of metabolic networks and biological pathways, it becomes increasingly evident that mutations leading to changes in a gene product can affect phenotypes in multiple and subtle ways. This makes the targeting of selective pressure and their potential adaptive consequences difficult. Integrative computational modeling conditioned on multiple data types offers one solution to this problem, and provides a robust means of testing evolutionary hypotheses.

See also: Adaptation, Fitness, and Evolution; Darwinism; Evolution, History of; Genetics and Anthropology; Human Behavioral Ecology; Microevolution.

Bibliography

Akey, J.M., Zhang, G., Zhang, K., Jin, L., Shriver, M.D., 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Research 12, 1805–1814.

Beaumont, M.A., Nielsen, R., Robert, C., Hey, J., Gaggiotti, O.E., Knowles, L., Estoup, A., Panchal, M., Corander, J., Hickerson, M., Sisson, S.A., Fagundes, N.J., Chikhi, L., Beerli, P., Vitalis, R., Cornuet, J.M., Huelsenbeck, J.P., Foll, M., Yang, Z., Rousset, F., Balding, D., Excoffier, L., 2010. In defense of model-based inference in phylogeography. Molecular Ecology 19 (3), 436–446.

Behar, D.M., Van Oven, M., Rosset, S., Metspalu, M., Loogvali, E.L., Silva, N.M., Kivisild, T., Torroni, A., Villems, R., 2012. A "Copernican" reassessment of the human mitochondrial DNA tree from its root. American Journal of Human Genetics 90, 675–684.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., Hirschhorn, J.N., 2004. Genetic signatures of strong recent positive selection at the lactase gene. American Journal of Human Genetics 74, 1111–1120.

Burger, J., Kirchner, M., Bramanti, B., Haak, W., Thomas, M.G., 2007. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. Proceedings of National Academy Science USA 104, 3736–3741.

Bryk, J., Hardouin, E., Pugach, I., Hughes, D., Strotmann, R., Stoneking, M., Myles, S., 2008. Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation. PLoS ONE 3, e2209.

Cann, R.L., Stoneking, M., Wilson, A.C., 1987. Mitochondrial DNA and human evolution. Nature 325, 31–36.

Cavalli-Sforza, L., Edwards, A.W., 1964. Analysis of human evolution. Genetics Today 3, 923–933.

Cavalli-Sforza, L., Menozzi, P., Piazza, A., 1994. The History and Geographic of Human Genes. Princeton University Press, Princeton, New Jersey, USA.

Chang, S.H., Jobling, S., Brennan, K., Headon, D.J., 2009. Enhanced EDAR signalling has pleiotropic effects on craniofacial and cutaneous glands. PLoS ONE 4, e7591.

Coop, G., Witonsky, D., Di Rienzo, A., Pritchard, J.K., 2010. Using environmental correlations to identify loci underlying local adaptation. Genetics 185, 1411–1423.

Currat, M., Excoffier, L., Maddison, W., Otto, S.P., Ray, N., Whitlock, M.C., Yeaman, S., 2006. Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in *Homo sapiens*" and "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans". Science 313 (5784), 172; author reply 172.

Dixson, B.J., Vasey, P.L., Sagata, K., Sibanda, N., Linklater, W.L., Dixson, A.F., 2011. Men's preferences for women's breast morphology in New Zealand, Samoa, and Papua New Guinea. Archives of Sexual Behavior 40, 1271–1279.

Enattah, N.S., Trudeau, A., Pimenoff, V., Maiuri, L., Auricchio, S., Greco, L., Rossi, M., Lentze, M., Seo, J.K., Rahgozar, S., Khalil, I., Alifrangis, M., Natah, S., Groop, L., Shaat, N., Kozlov, A., Verschubskaya, G., Comas, D., Bulayeva, K., Mehdi, S.Q., Terwilliger, J.D., Sahi, T., Savilahti, E., Perola, M., Sajantila, A., Jarvela, I., Peltonen, L., 2007. Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans. American Journal of Human Genetics 81, 615–625.

Evershed, R.P., Payne, S., Sherratt, A.G., Copley, M.S., Coolidge, J., Urem-Kotsu, D., Kotsakis, K., Ozdogan, M., Ozdogan, A.E., Nieuwenhuyse, O., Akkermans, P.M., Bailey, D., Andeescu, R.R., Campbell, S., Farid, S., Hodder, I., Yalman, N., Ozbasaran, M., Bicakci, E., Garfinkel, Y., Levy, T., Burton, M.M., 2008. Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding. Nature 455, 528–531.

Fagundes, N.J., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F.M., Bonatto, S.L., Excoffier, L., 2007. Statistical evaluation of alternative models of human evolution. Proceedings of the National Academy of Sciences of the United States of America 104, 17614–17619.

Fay, J.C., Wu, C.I., 2000. Hitchhiking under positive Darwinian selection. Genetics 155, 1405–1413.

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admetlla, A., Pattini, L., Nielsen, R., 2012. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genetics 7, e1002355.

Gerbault, P., Liebert, A., Itan, Y., Powell, A., Currat, M., Burger, J., Swallow, D.M., Thomas, M.G., 2011. Evolution of lactase persistence: an example of human niche construction. Philosophical Transactions of the Royal Society London B: Biological Sciences 366, 863–877.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H., Hansen, N.F., Durand, E.Y., Malaspinas, A.S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prufer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Hober, B., Hoffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., De La Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., Paabo, S., 2010. A draft sequence of the Neanderthal genome. Science 328, 710–722.

Hancock, A.M., Alkorta-Aranburu, G., Witonsky, D.B., Di Rienzo, A., 2010. Adaptations to new environments in humans: the role of subtle allele frequency shifts. Philosophical Transactions of the Royal Society London B: Biological Sciences 365, 2459–2468.

Higham, T., Douka, K., Wood, R., Ramsey, C.B., Brock, F., Basell, L., Camps, M., Arrizabalaga, A., Baena, J., Barroso-Ruiz, C., Bergman, C., Boitard, C., Boscato, P., Caparros, M., Conard, N.J., Draily, C., Froment, A., Galvan, B., Gambassini, P., Garcia-Moreno, A., Grimaldi, S., Haesaerts, P., Holt, B., Iriarte-Chiapusso, M.J., Jelinek, A., Jorda Pardo, J.F., Maillo-Fernandez, J.M., Marom, A., Maroto, J., Menendez, M., Metz, L., Morin, E., Moroni, A., Negrino, F., Panagopoulou, E., Peresani, M., Pirson, S., De La Rasilla, M., Riel-Salvatore, J., Ronchitelli, A., Santamaria, D., Semal, P., Slimak, L., Soler, J., Soler, N., Villaluenga, A., Pinhasi, R., Jacobi, R., 2014. The timing and spatiotemporal patterning of Neanderthal disappearance. Nature 512, 306–309.

Holden, C., Mace, R., 1997. Phylogenetic analysis of the evolution of lactose digestion in adults. Human Biology 69, 605–628.

Hudson, R.R., Kreitman, M., Aguade, M., 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116, 153–159.

Ingman, M., Kaessmann, H., Paabo, S., Gyllensten, U., 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408, 708–713.

Itan, Y., Powell, A., Beaumont, M.A., Burger, J., Thomas, M.G., 2009. The origins of lactase persistence in Europe. PLoS Computational Biology 5, e1000491.

Kamberov, Y.G., Wang, S., Tan, J., Gerbault, P., Wark, A., Tan, L., Yang, Y., Li, S., Tang, K., Chen, H., Powell, A., Itan, Y., Fuller, D., Lohmueller, J., Mao, J., Schachar, A., Paymer, M., Hostetter, E., Byrne, E., Burnett, M., Mcmahon, A.P., Thomas, M.G., Lieberman, D.E., Jin, L., Tabin, C.J., Morgan, B.A., Sabeti, P.C., 2013. Modeling recent human evolution in mice by expression of a selected EDAR variant. Cell 152, 691–702.

Ke, Y., Su, B., Song, X., Lu, D., Chen, L., Li, H., Qi, C., Marzuki, S., Deka, R., Underhill, P., Xiao, C., Shriver, M., Lell, J., Wallace, D., Wells, R.S., Seielstad, M., Oefner, P., Zhu, D., Jin, J., Huang, W., Chakraborty, R., Chen, Z., Jin, L., 2001. African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes. Science 292, 1151–1153.

Kelley, J.L., Madeoy, J., Calhoun, J.C., Swanson, W., Akey, J.M., 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Research 16, 980–989.

Kimura, M., 1968. Evolutionary rate at the molecular level. Nature 217, 624–626.

Kimura, M., 1971. Theoretical foundation of population genetics at the molecular level. Theoretical Population Biology 2, 174–208.

Kingman, J.F.C., 1982. The coalescent. Stochastic Processes and Their Applications 13, 235–248.

Kwiatkowski, D.P., 2005. How malaria has affected the human genome and what human genetics can teach us about malaria. American Journal of Human Genetics 77, 171–192.

Lewontin, R.C., Krakauer, J., 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74, 175–195.

Luca, F., Perry, G.H., Di Rienzo, A., 2010. Evolutionary adaptations to dietary changes. Annual Review of Nutrition 30, 291–314.

McDougall, I., Brown, F.H., Fleagle, J.G., 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature 433, 733–736.

Mendez, F.L., Krahn, T., Schrack, B., Krahn, A.M., Veeramah, K.R., Woerner, A.E., Fomine, F.L., Bradman, N., Thomas, M.G., Karafet, T.M., Hammer, M.F., 2013. An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. American Journal of Human Genetics 92, 454–459.

Mulcare, C.A., 2006. The Evolution of the Lactase Persistence Phenotype. Genetic, Evolution and Environment. University of London, London.

Nei, M., Roychoudhury, A.K., 1974. Genic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. American Journal of Human Genetics 26 (4), 421.

Nielsen, R., 2005. Molecular signatures of natural selection. Annual Review of Genetics 39, 197–218.

Novembre, J., Stephens, M., 2008. Interpreting principal component analyses of spatial population genetic variation. Nature Genetics 40, 646–649.

Pool, J.E., Hellmann, I., Jensen, J.D., Nielsen, R., 2010. Population genetic inference from genomic sequence variation. Genome Research 20, 291–300.

Powell, A., Shennan, S., Thomas, M.G., 2009. Late Pleistocene demography and the appearance of modern human behavior. Science 324, 1298–1301.

Przeworski, M., 2002. The signature of positive selection at randomly chosen loci. Genetics 160, 1179–1189.

Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L., Maricic, T., Good, J.M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E.E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M.V., Derevianko, A.P., Hublin, J.J., Kelso, J., Slatkin, M., Paabo, S., 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468, 1053–1060.

Rice, S.H., 2004. Effective Population Size. Evolutionary Theory: Mathematical and Conceptual Foundations. Sinauer Associates, Inc., Sunderland, MA, USA.

Sverrisdottir, O.O., Timpson, A., Toombs, J., Lecoeur, C., Froguel, P., Carretero, J.M., Ferreras, A., Arsuaga Gotherstrom, J.L., Thomas, M.G., 2014. Direct estimates of natural selection in iberia indicate calcium absorption was not the only driver of lactase persistence in Europe. Molecular Biology and Evolution 31, 975–983.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595.

Vigne, J.D., 2008. Zooarchaeological aspects of the neolithic diet transition in the Near East and Europe, and their putative relationships with the neolithic demographic transition. In: Bocquet-Appel, J.P., Bar-Yosef, O. (Eds.), The Neolithic Demographic Transition and Its Consequences. Springer, New York.

Voight, B.F., Adams, A.M., Frisse, L.A., Qian, Y., Hudson, R.R., Di Rienzo, A., 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proceedings of the National Academy of Sciences of the United States of America 102, 18508–18513.

Voight, B.F., Kudaravalli, S., Wen, X., Pritchard, J.K., 2006. A map of recent positive selection in the human genome. PLoS Biology 4, e72.

Wei, W., Ayub, Q., Chen, Y., Mccarthy, S., Hou, Y., Carbone, I., Xue, Y., Tyler-Smith, C., 2013. A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Research 23, 388–395.

Wilde, S., Timpson, A., Kirsanow, K., Kaiser, E., Kayser, M., Unterlander, M., Hollfelder, N., Potekhina, I.D., Schier, W., Thomas, M.G., Burger, J., 2014. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proceedings of the National Academy of Sciences of the United States of America 111, 4832–4837.

Wright, S., 1931. Evolution in Mendelian populations. Genetics 16, 97–159.

Xue, Y., Zhang, X., Huang, N., Daly, A., Gillson, C.J., Macarthur, D.G., Yngvadottir, B., Nica, A.C., Woodwark, C., Chen, Y., Conrad, D.F., Ayub, Q., Mehdi, S.Q., Li, P., Tyler-Smith, C., 2009. Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. Genetics 183, 1065–1077.

Zuckerkandl, E., Pauling, L., 1965. Molecules as documents of evolutionary history. Journal of Theoretical Biology 8, 357–366.

Human Factors and Ergonomics

Maureen Hassall and Tania Xiao, The University of Queensland, St Lucia, QLD, Australia
Penelope Sanderson and Andrew Neal, The University of Queensland, St Lucia, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Human factors and ergonomics applies scientific information about human cognition and behavior to support the design of products and systems that enhance human well-being and performance. Human factors and ergonomics professionals investigate human–system interaction at the individual, group, and organizational levels. The aim is to design products that augment human physical, cognitive, and motivational capabilities and improve the social, technical, and organizational aspects of a system. Human factors and ergonomics has been used to improve safety and productivity particularly in complex sociotechnical systems.

Introduction

"Everything we invent or make is ultimately designed for human use and benefit" (Chapanis, 1996). As technology and work systems have evolved they have become larger, more complex, and more interconnected. The rate of change in such systems and their increasing complexity have increased the likelihood that unintended and adverse consequences will occur. Human factors and ergonomics has emerged as a profession whose practitioners endeavor to ensure that tools, products, and systems are designed to be safe, usable, and useful for humans. Specifically, human factors and ergonomics is "the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data and methods to design in order to optimize human well-being and overall system performance" (www.iea.cc; www.hfes.org). According to Dul et al. (2012), human factors and ergonomics is unique because "(1) it takes a systems approach (2) it is design driven and (3) it focuses on two closely related outcomes: performance and well-being" (p. 377).

The scope of human factors and ergonomics work has evolved to be very broad, as shown in **Figure 1**. Human factors and ergonomics work is typically performed at one of three levels of analysis – the individual, group, or organizational level. The individual level of analysis focuses on interactions that an individual human might have with an engineered system, technology, or tool. The group level of analysis focuses on how groups or teams of humans interact with technology and how these interactions influence performance. The organizational level of analysis investigates the influence that organizational design factors (e.g., organizational structure, organizational processes, and work design) and the use of technology have on human well-being and overall productivity. For each of these levels of analysis, human factors and ergonomics professionals draw on knowledge from different disciplines (shown in left column of **Figure 1**) to investigate a range of topics (center column of **Figure 1**) in order to formulate design interventions intended to human–system interactions (right column of **Figure 1**).

In this article, we review the background and scope of human factors and ergonomics work. In the background section, we discuss the origins and history of human factors and ergonomics. We then describe the scope of human factors and ergonomics work in three sections that cover the individual, group, and organization levels of analysis. We conclude the article by outlining some of the future directions and challenges that the human factors and ergonomics discipline faces.

Background

The history of human factors and ergonomics has been traced back to 1857 when Polish scientist Jastrezebowski proposed the term 'ergonomics' to represent the science of human work (Jastrzebowski, 1857[2001]). After the introduction of mass production, the scientific study of human physical labor was reinvigorated by Scientific Management. During this era, scientific studies of labor adopted techniques such as time and motion studies to decompose work into standardized functions that represent the 'best' way for workers to perform a task efficiently and with minimal stress.

The next major change in labor studies occurred during the two World Wars of the twentieth century. In World War I, intense requirements were placed on factory workers which resulted in unacceptable levels of fatigue and injuries (Zionchenko and Munipov, 2005). The responsibility of improving human well-being at work while ensuring high levels of productivity was allocated to specially developed health boards and committees staffed by physiologists, psychologists, engineers, and doctors (Zionchenko and Munipov, 2005).

In World War II, rapid technical advancements changed the nature of human work in a number of ways. For example, newly developed fighter planes required pilots to manage complex cockpits and deal with low oxygen and with high g-forces that they would not survive without supplementary air and special flying suits (Chapanis, 1996). New technology also led to the development of screen-based work including radar and sonar monitoring jobs (Chapanis, 1996). During this time, psychologists, physiologists, and physicians were recruited to work cooperatively with military engineers to create designs that would handle safety and functionality issues associated

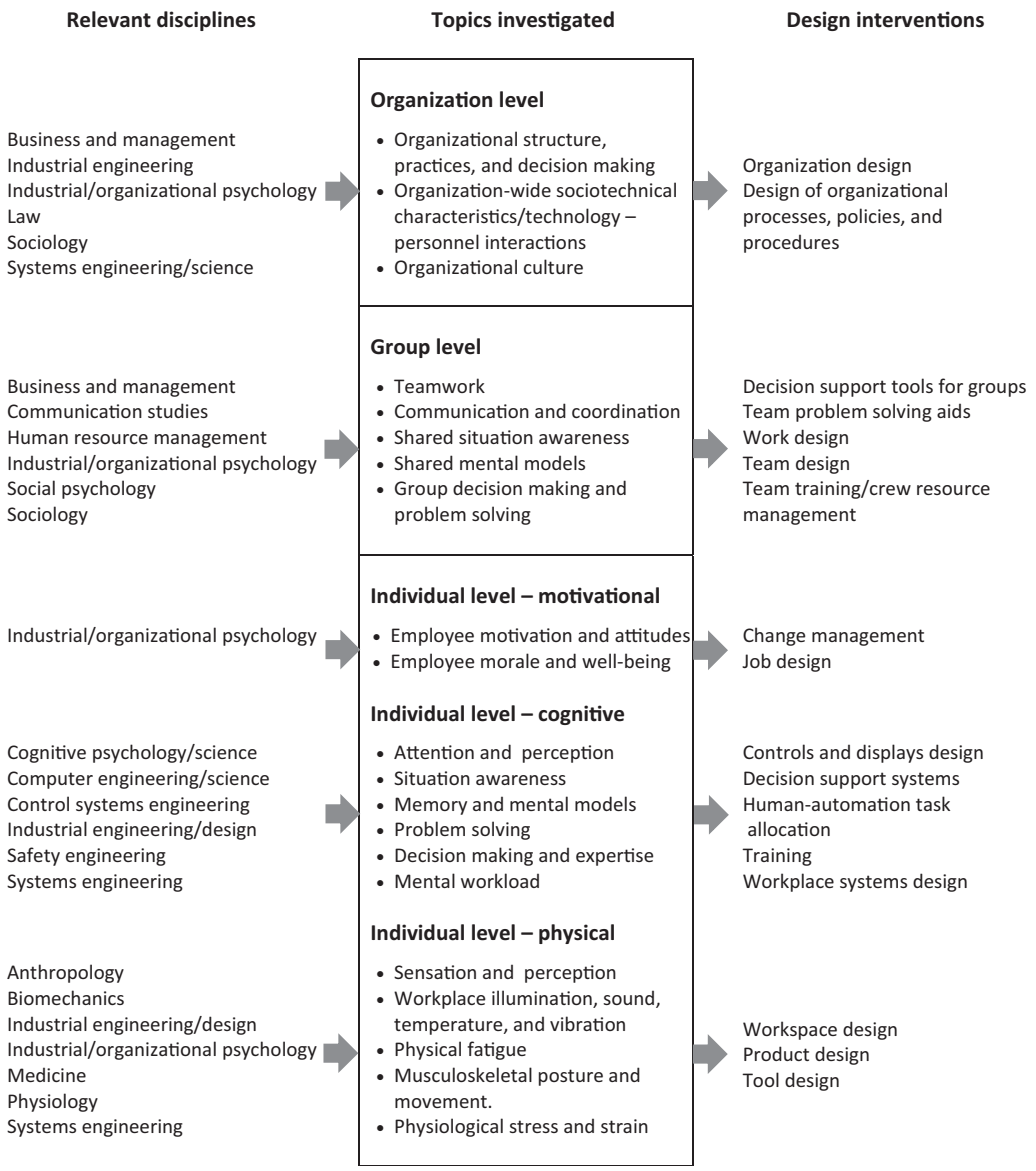

Figure 1 Scope of human factors and ergonomics.

with human–technology interactions (Zionchenko and Munipov, 2005).

After World War II, the size and complexity of work systems continued to grow. For example, nuclear weapons technology was converted for use in electricity generation which lead the development of complex, safety critical nuclear power plants. The Three Mile Island nuclear power plant accident in 1979 illustrated the importance of ensuring that industrial systems were built to suit the human operators and controllers. The accident was partly caused by the fact that the system design provided operators with confusing and misleading information that led them to make incorrect decisions (Vicente, 1999). The lessons from Three Mile Island, combined with the proliferation of computerized interfaces in industry, led to the emergence of cognitive systems engineering (Hollnagel and Woods, 1983). Cognitive systems engineering focuses on designing complex sociotechnical systems to facilitate effective and adaptive human diagnosis and control during normal and abnormal operating situations.

As the human factors and ergonomics discipline evolved, its scientists and practitioners began to form into professional societies. Two of the more prominent societies today are the International Ergonomics Association (IEA) and the USA-based Human Factors and Ergonomics Society (HFES). Within these two societies are subgroups that focus on specific aspects and application areas of human factors and ergonomics, such as anthropometry, cognitive engineering and decision making, ergonomics in design, and individual differences in performance. Further subgroups focus on specific applications such as aerospace, aging, healthcare, safety, and transportation. Today, human factors professionals draw on knowledge from many disciplines to take an

evidence-based systems approach to improving human interaction within systems.

Individual Level of Analysis

Much human factors and ergonomics work focuses on the interaction between individual humans and the technical and social systems within which they work. Human factors professionals seek to understand the people who will work in/with the system, the functions people and technology need to perform, and how people might go about performing their required functions. With such an understanding, human factors and ergonomics professionals can help design systems that accommodate the natural variability associated with human performance in ways that help humans achieve the overall system purpose safely and productively while minimizing workplace stress. Studies that investigate interactions between individual humans and sociotechnical systems usually concentrate on *physical ergonomic*s or on *cognitive ergonomics* which will be discussed later. There is also a body of research examining motivational factors.

Physical Ergonomics

Physical ergonomics focuses on designing jobs, tools, equipment, and workspaces to fit the physical attributes of workers. As IEA notes, "physical ergonomics is concerned with human anatomical, anthropometric, physiological and biomechanical characteristics as they relate to physical activity" (www.IEA.cc). Studies involve understanding the capabilities, constraints, and variability associated with the human body in order to design products, processes, and work environments that are productive, safe, and satisfying for people to use (e.g., Ahram and Karwowski, 2013). Physical ergonomics professionals draw on human anthropometrical, biomechanical, and physiological information to design physical environments that accommodate and enhance human sensation and perception, human physical postures, and human musculoskeletal performance (Kroemer and Grandjean, 2009; Marras and Karwowski, 2006).

Sensation and perception are important because they are the processes by which humans receive and process information from their environment (Wickens et al., 2004). Human sensory systems include vision, audition, olfaction, gustation, somasthesis, and the vestibular system. Physical ergonomics professionals aim to design work environments that enhance human sensation and perception and exploit its strengths, rather than harm it. This requires an understanding of how environmental factors such as light, sound, vibration, and temperature can be sources of comfort, discomfort, or even harm to humans. For example, light can help or distract humans depending on the interaction between properties of the light (e.g., the size, luminosity, differentiation, and quality of light), the capabilities, limitations, and variability inherent in the human visual system, and characteristics of the work that the human is required to do. Similarly, human performance can be affected by mismatches among the levels of sound, temperature, and vibration in a system, tolerability limits associated with human senses, and the requirements of the task. Successful physical ergonomics interventions reduce the occurrence of chronic illnesses and impairments (e.g., hearing loss); reduce stress; and improve safety, well-being, and productivity.

Physical ergonomics professionals also investigate the relationships between work space and tool design on the one hand, and human anthropometry, physiological, and biomechanical data on the other. They study how these relationships might affect human safety, health, comfort, and performance (Bridger, 2003). Topics of interest include posture, manual handling, repetitive movements, musculoskeletal workload, and physiological stress and strain. Physical ergonomists aim to produce designs that minimize fatigue, discomfort, musculoskeletal damage, and repetitive strain injuries (Wickens et al., 2004). They perform scientific studies in order to make recommendations on how the design of portable devices, computer input tools, workplace furniture, work station layout, and jobs could be changed to improve user health, safety, comfort, and performance. Examples of some physical ergonomics projects include designing cell phones to suit the attributes of human hands in ways that permits the phone to be operated with one hand only, designing office chairs with the adjustability needed to accommodate the range of human body shapes and sizes, and designing torqueless tools for astronauts working in zero gravity while enclosed in cumbersome space suits.

Cognitive Ergonomics

Cognitive ergonomics focuses on human mental processing abilities and limitations (www.iea.cc). Following the introduction of computerized technology, work has become less physically demanding, and more cognitively demanding. Many tasks place heavy demands on attention, perception, memory, and reasoning. Understanding the strengths and vulnerabilities of human cognitive processes is required to design products, interfaces, and systems that can be used safely and effectively by humans (Hollnagel and Woods, 1983). Cognitive ergonomics draws on knowledge from cognitive psychology, cognitive science, computer science, and control engineering. Cognitive ergonomists seek to design tools and systems based on a scientific understanding of attention, perception, memory, mental models, expertise, and mental workload in order to enhance human situation awareness, problem solving, and decision-making capabilities (Harris, 2013).

Designing systems that enhance human attention and perception is important, because the latter are the processes that collect, filter, and interpret information from the environment. The failure to attend to appropriate cues was a significant factor in Eastern Air Lines Flight 401 crash in 1972 where the entire flight deck crew were so focused on dealing with a faulty landing gear light, they failed to notice that the plane was losing altitude and failed to hear an altitude warning alarm (National Transport Safety Board, 1973). Attention and perception are important components of situation awareness.

Situation awareness has been variously defined. One commonly used definition comes from Endsley (1988) who defines situation awareness as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their

status in the near future" (p. 97). Another definition is that situation awareness is the "continuous extraction of environmental information, integration of this information with previous knowledge to form a coherent mental picture, and the use of that picture in directing further perception and anticipating future events" (Dominguez, 1994). Inaccurate situation awareness has been shown to be a contributing factor to industrial accidents including aviation accidents (Jones and Endsley, 1996) and process industry accidents (Kaber and Endsley, 1998). Hence, creating designs that facilitate accurate situation awareness is a critical prerequisite for ensuring humans correctly respond to situations.

Accurate situation awareness relies, in part, on human memory and so-called mental models. Memory is the human capacity to retain information for future use. Memory is often categorized into working memory and long-term memory (Wickens et al., 2004). Working memory stores limited information on a short-term, temporary basis so that it can be referenced and manipulated. Long-term memory stores information and its relevant associations on a more permanent basis. The term 'mental models' (Gentner and Stevens, 1983) refers to a person's mental representation of a system or object that includes its purpose and descriptions of how it functions and behaves. Mental models are useful constructs to help explain how people comprehend the current status, predict potential future states, and to troubleshoot any issues with a system or product. Inaccurate mental models contributed to the Three Mile Island nuclear reactor incident. Operators believed that an illuminated control panel light for a pressure relief valve indicated that the valve was closed, whereas the light illuminated only when the instruction was given to close the valve and did not indicate the actual physical position of the valve (Sanders, 1998). During the accident, the light was illuminated so operators thought the valve was closed, but it was actually stuck open. This confusion contributed to the accident. Human factors and ergonomics professionals aim to design systems that shape and reinforce accurate mental models so that humans can make correct diagnoses and take appropriate actions in normal and emergency situations.

Human interactions within engineered systems and products usually require them to make decisions and often to solve problems. Decision making is the process that humans use to determine a forward path of action. Problem solving is the mental process used to analyze and resolve problems. To make decisions, people can use different strategies and can be influenced by different biases (Klein, 2009; Kahneman, 2011). For example, people can make decisions using slow and deliberative processes or rapid heuristic processes. These processes can be influenced by biases that can lead to the selection of choices and solutions that are not normatively optimal. Such biases include confirmation bias, hindsight bias, anchoring, and framing effects. Bias in decision making can initiate and escalate situations that lead to accidents. For example, the decision making associated with not redesigning o-rings and then launching space shuttle *Challenger* when the temperature was below 53 °F was found to be a significant factor in the loss of the shuttle and all its crew (Rogers et al., 1986). Confirmation bias was found to be a contributing factor in the misinterpretation of the well integrity test results which, in turn, contributed to the Deepwater Horizon disaster (Hopkins, 2012). Conversely, correct decision making was identified as the first factor that contributed to the successful landing of US Airways flight 1549 on the Hudson River in 2009 (National Transport Safety Board, 2010). Studies that investigate human decision making in real-world contexts are often called naturalistic decision-making studies (Klein et al., 1993). Human factors and ergonomics professionals perform such studies to better understand how to design systems and tools that support human decision-making processes and minimize biases.

Another area of interest to human factors and ergonomics professionals working in cognitive ergonomics is the study of expertise. Human factors and ergonomics professionals seek to understand the cognitive mechanisms, task factors, and environmental situations that contribute to the superior performance that we describe as expertise. One model that is used to understand expert decision making is the Recognition Primed Decision model which suggests that experienced people use their recognition of a situation and prior experience with similar situations to make quick decisions on how to act (Klein et al., 1993). Understanding the factors that can influence the acquisition and execution of expertise can help human factors and ergonomics professionals design tasks and training intended to facilitate the delivery of expert performance. For example, by studying expert land mine detection personnel and using the information to train others, Staszewski (2004) was able to show an improvement in overall detection rates for low metal mines from less than 30% to greater than 80% detection.

Another topic of interest in cognitive ergonomics is that of mental workload. Mental workload is a function of the cognitive demands associated with task performance and of the mental processing ability of the person performing the task (Parasuraman et al., 2008). Performance and safety issues can arise when the cognitive demands of a task exceed the cognitive capabilities of the people performing the task. For example, one of the key factors that contributed to the 1994 explosion at the Texaco Milford Haven refinery was the overloading of control room operators who had to act on 275 poorly prioritized alarms in 11 min and who had to manage a situation using control room displays that did not provide an overview of the process they were controlling (Health and Safety Executive, 1994). Mental workload problems can also arise when the task allocation between humans and automation does not adequately engage human mental processing abilities. Giving automation the responsibility to perform most control functions can leave the human operator with a poor understanding of system functionality and with a low level of control skills that hinder effective human control of events that the automation is not designed to manage (Moray, 1986). For example, the loss of Air France flight 447 with 228 people on board in 2009 occurred when the pilots were unable to effectively manage the plane after the autopilot disconnected at high altitude because the system, including the technology and training provided, was not designed to help the pilots correctly diagnose and manually respond to the unexpected situation (BEA, 2012).

Cognitive ergonomics has also contributed in a number of ways to design safer, more productive systems by improving the design of tools, interfaces, and automated processes that augment the capabilities of individual users. For example, cognitive ergonomics studies were used during the TADMUS

program – a program that was motivated by the 1988 shooting down an Iran commercial passenger airline by the *USS Vincennes*. TADMUS generated a number of design and training solutions to help Navy decision makers correctly identify objects picked up by radar, thereby avoiding such accidents (Cooke and Durso, 2008). To improve interfaces and automated processes, cognitive systems engineers have developed analytic frameworks such as Cognitive Work Analysis (Rasmussen et al., 1994; Vicente, 1999). Cognitive Work Analysis underlies the Ecological Interface Design (EID) approach to designing interfaces and it also guides the allocation of function between humans and automated systems. EID interfaces are intended to support the range of cognition that operators might use to manage anticipated and unanticipated scenarios while controlling complex systems (Vicente and Rasmussen, 1992). EID interfaces have been shown to improve operator problem solving, diagnosis, and control, especially for unanticipated events (e.g., Jamieson, 2007). They have also been shown to be easier to learn and use and more effective in military command and control (Hall et al., 2012).

Motivational Factors

Motivation of the individual worker is not often covered in treatments of human factors and ergonomics because it is conventionally more strongly associated with organizational psychology. Nonetheless, individual motivation is an important factor that has received more attention in, for example, the francophone work psychology tradition, which has contributed ideas to cognitive ergonomics (e.g., Pinder, 2007). Motivational research has examined the role of employee motivation, attitudes, and well-being within sociotechnical systems (Rousseau, 1977). Work and job design are major factors that influence motivation (Campion and Thayer, 1985, 1989). For example, job simplification is a widespread practice, being driven by technological, regulatory, and societal changes. Job simplification involves the proceduralization of work and the removal of autonomy and discretion on the part of the employee. Research has consistently demonstrated that job simplification adversely affects employee motivation and well-being. People need enriched jobs that provide autonomy, variety, and meaning (Pan and Werblow, 2012). In contrast, job enrichment leads to improvements in job satisfaction, productivity, and other employee and business benefits (Humphrey et al., 2007). How worker motivation is ensured during organizational design and preserved during organization change will be covered in the section on the group level of analysis in human factors and ergonomics.

Employee motivation and attitudes are also important during organizational change and when improving occupational safety. Employee motivation and attitudes play a critical role in organizational change. Organizational change initiatives frequently fail due to employee resistance. Similarly, in the area of occupational safety, employee motivation and attitudes play a critical role in determining whether employees will comply with safe work practices and take a proactive approach to safety (e.g., by helping co-workers).

The work done by the human factors and ergonomics community to improve human well-being and system performance at an individual level has more recently been extended to the group and organizational interactions. We discuss the work done at group and organization levels in the next two sections of the article.

Group Level of Analysis

Humans often perform work and interact with systems in groups or teams. For example, in the cockpits of commercial aircraft the captain and co-pilot work together to successfully manage flight operations. Similarly, in medical operating rooms, surgeons, anesthetists, and nurses need to work together to ensure the surgical procedure is successful. As a result, human factors and ergonomics studies can focus on how groups or teams of humans interact to perform tasks in engineered systems. Human factors and ergonomics professionals draw on knowledge from industrial and organizational psychology, social psychology, sociology, industrial engineering, and other disciplines to investigate how system design influences team effectiveness. Team effectiveness is affected by both individual and group attributes. Individual attributes include the physical, cognitive, and motivational attributes discussed in the previous section. Group attributes include teamwork, shared situation awareness, shared mental models, and group decision making. These will be discussed next.

Human factors and ergonomics professionals are interested in how to design products and systems that enhance effective teamwork. A first important prerequisite of effective teamwork is that team members should understand the objectives of the team, where the team fits with other teams, and the overall organizational objectives. A second important prerequisite is that team members should understand their own and other team members' roles, responsibilities, and preferences. When team members understand the objectives and task demands required of the team as well as the knowledge, tendencies, and attitudes of team members, they are better able to coordinate team activities (Cannon-Bowers and Salas, 2006). A third important prerequisite of effective team work is good communication and coordination. The link between effective teamwork and improved performance has been made in the number of contexts including healthcare (e.g., Wheelan et al., 2003) and aviation (e.g., Leedom and Simon, 1995; Stout et al., 1997). The *Herald of Free Enterprise* sinking which resulted when the ferry left port with its bow doors open is an example of the consequences that can occur as a result having systems that tolerate poor teamwork.

A fourth important aspect of team effectiveness is 'shared' situation awareness. In a study of the process industry, it was found that 50% of operations failures associated with reported incidents were due to failures in shared situation awareness (Bullemer and Reising, 2013). Shared situation awareness has two components: individual team members' situation awareness (described in previous section) and the shared understanding of the situation that is present in the team at a given point in time. The shared understanding of the situation can be affected by the context (e.g., objectives to achieve, resources available), environmental conditions (e.g., climatic or workplace conditions), and temporal elements (e.g., time pressure,

risk levels) (Salas et al., 2006). The quality of a team's shared situational awareness can affect team communication and coordination activities as well as the attitudes of, and demands placed on, team members. Shared situation awareness also depends, in part, on shared mental models. Understanding shared mental models helps reveals (1) the commonality between the mental representations held by team members and (2) how the mental representation of the system is distributed or shared among different team members. As Cooke and Gorman (2006) have noted, by considering both perspectives "it is possible to (1) assess the degree to which each member has positional knowledge or knowledge relevant to his or her own role, and (2) the degree to which each team member has inter-positional knowledge or knowledge pertinent to other team roles" (p. 272). With such an understanding, human factors and ergonomics professionals can advise on ways that system designs can promote effective shared situation awareness and shared mental models within groups and teams.

Group decision making is also a key topic for human factors and ergonomics at the group level. Group decision making differs from individual decision making in that it can be influenced by group biases such as group think, internal conflict, group processes, group ethics, social norms of the group, and the combination of individual member preferences and needs. Understanding group decision-making processes is important because it has been identified as a significant contributing factor in major accidents including the loss of the space shuttle *Challenger* (Rogers et al., 1986) and in the Deepwater Horizon disaster (Hopkins, 2012). By understanding group decision-making processes and influences, human factors professionals can design decision support tools and problem-solving aids that promote more effective group decision making.

Other design outcomes that might result from the group level of analysis include changes to team training initiatives, job design, and team design. For example, crew resource management (CRM) is a human factors-inspired training program that focuses on improving crew performance, coordination, and activity management within a particular context. CRM was initially designed to help pilots work together in cockpits to successfully control airplanes through a variety of flight conditions and contingencies. The success of CRM was highlighted in the investigation report of US Airways Flight 1549 crash landing in the Hudson River which stated that "contributing to the survivability of the accident was (1) the decision-making of the flight crewmembers and their crew resource management during the accident sequence" (National Transport Safety Board, 2010). The principles underlying CRM have been extended to other activities such as maintenance work and to other domains such as healthcare. An example of a human factors and ergonomics-inspired job and team redesign initiative that lead to significant improvements was the redesign of an emergency response organization within a nuclear power plant (Klinger and Klein, 1999). The redesign recommendations included suggested changes to the team composition and emergency response operations room layout which led to fewer people, improved situation awareness, and better response outcomes.

Organization Level of Analysis

Human factors and ergonomics professionals recognize that human well-being and overall system performance can be affected by organizational factors. The study of such factors is referred to as organizational ergonomics or macroergonomics. Organizational ergonomics draws on information from organizational science, management science, complexity science, and sociotechnical systems theory to understand how organizational structures, policies, practices, and processes affect human performance and productivity. The aim of organizational ergonomics is to ensure the organization as a whole is designed to facilitate safe and effective interactions between technological subsystems, personnel subsystems, and the external environment. As Hendrick (2005) has noted, "work systems that are fully harmonized and compatible with their sociotechnical characteristics…can result in synergistic improvements in various organizational effectiveness criteria, including health, safety, comfort, and productivity" (p. 43). Topics of interest include organizational design, work design, and the design of organization-wide processes.

Organizational design is the design of the whole organization and it can be described using three dimensions: complexity, formalization, and centralization (Hendrick and Kleiner, 2002). Complexity is the degree to which organizational subunits are differentiated in function and interlinked by coordinating processes. Formalization is the degree to which organizational processes are standardized. Centralization is the extent to which decision making and authority are concentrated within the organization. Organization complexity, formalization, and centralization influence how work is managed at the team level. These three organizational attributes can also influence the culture and behavioral norms within subunits and teams. The importance of organizational factors on human well-being and overall system performance was highlighted as a significant contributor to the 2000 Walkerton water contamination incident that resulted in 2400 people becoming ill and 7 dying (Leveson et al., 2003) and the 2003 space shuttle *Columbia* disaster which resulted in the loss of the shuttle and its crew (Columbia Accident Investigation Board, 2003). Human factors and ergonomics professionals can investigate and recommend organizational design interventions that promote effective interactions between people and technology at an individual, team, subsystem, and organizational level.

Work design involves determining how to arrange work to enhance employee satisfaction and motivation while delivering organizational quality and productivity requirements. Work design initiatives consider task variability factors, work scheduling factors (e.g., shift rosters, flexible working arrangements), physical location of work, productivity requirements, regulatory requirements, employee safety, and social and motivational needs as well as the needs of other stakeholders. Understanding the influence that work design has on human–system interactions is important when introducing organizational level change. For example, between 50 and 75% of technology implementation projects experience some degree of failure (Umble and Umble, 2002). Some of the reasons for these failures can be summarized as a failure to consider system-wide effects of how the technology impacts

user needs, the current work design arrangements, and the organization's current use of technology. Human factors and ergonomics professionals can use macroergonomic techniques to help ensure that organization-wide change projects enhance effective interactions between work designs, workers, and technology at an organization level (Moro, 2009).

The term 'organizational culture' refers to the shared values and assumptions of the members of that organization, and is sometimes described as 'the way things are done around here'. Culture plays an important role in shaping employee behavior and employee response to new systems and processes. For example, safety management systems are now mandatory in many jurisdictions; however, a safety management system is unlikely to be effective unless there is a culture in which safety is valued and employees feel able to raise safety issues without fear of reprisal (Reason, 1997).

Another area where human factors and ergonomics professionals provide advice is on the design of organizational processes. An example of such an organizational process that is frequently recommended by macroergonomics professionals is participatory ergonomics. Participatory ergonomics, an extension of participatory management, is the process by which employees who are affected by a problem or initiative are involved in identifying the interventions that best suit them and that also help to achieve the organizational objectives. As Hendrick (2005) comments, "the participatory approach has proven to be particularly effective in establishing an ergonomics and safety culture, which sustains performance and safety improvements that initially result from the macroergonomic intervention" (p. 55).

Organizational ergonomics can deliver significant improvements to human well-being and to overall system performance. Hendrick (2005) reported that organizational ergonomics interventions have achieved improvements of 50–90% in productivity or quality. The improvements were obtained by adopting the following macroergonomic criteria for effective work system design (Hendrick, 2005):

1. Joint design. The approach used should be human-centered, where the personnel subsystem is designed jointly with the technological subsystem and the design process involves extensive employee participation.
2. Humanized task approach. The allocation of functions and tasks to humans or machines should first consider human professional, cultural, and psychosocial needs before allocating the function or task to a human or a machine.
3. Sociotechnical systems approach. Consideration and evaluation of an organization's sociotechnical characteristics should be integrated into the process of designing any work system.

Conclusion

The nature of human–systems interaction at the individual, group, and organization level will continue to change as technology, people, and systems themselves continue to change. We conclude this article with a brief discussion on the future challenges and opportunities facing the human factors and ergonomics profession.

The development and expansion of digital technology will continue to create paradigm shifts in people's work and lifestyles (e.g., voice recognition, robotics, and augmented cognition). Such paradigm shifts offer opportunities for new interactive experiences, new social challenges, new types of work designs, and new forms of organization. For example, the research currently being undertaken in 'nano' technologies – molecular scale technologies – could lead to the emergence of nanoergonomics studies that seek to improve human interactions with extremely small (nano-sized) devices and machines (Karwowski, 2006). Human factors and ergonomics professionals will need to stay abreast of technology advances in order to continue to propose design improvements that enhance human well-being and overall system performance.

However, the human factors and ergonomics discipline has traditionally been one that reacts to technology developments and safety issues. The discipline needs to become more proactive in anticipating and driving technology advances. Human factors and ergonomics needs to "try to anticipate how the nature of work will change so that it can both foresee what work will be and propose what work should be" (Hollnagel, 2014). As a field of endeavor, human factors and ergonomics also needs to drive technology advancements by integrating itself into every step in the system lifecycle, from operational needs analysis, to concept development, engineering design, implementation and commissioning, operations and maintenance, and even decommissioning (Chapanis, 1996). To be more proactive, human factors and ergonomics professionals need to concentrate on the initial steps of the system development lifecycle. With early involvement, human factors and ergonomics professionals can help to identify, specify, and help create new technology that aids human well-being and overall system performance, rather than just trying to improve existing technologies. Examples of technologies created from human factors and ergonomics initiatives have begun to emerge especially from the healthcare sector. For example, human factors and ergonomics researchers are developing implantable early warning devices for people who are at risk of having a heart attack (Day and Young, 2012). This work needs to be extended across other industries and to help improve other types of human–system interactions.

Another challenge that human factors and ergonomics faces is the need to continue to work on the positioning of the discipline in relation to science and engineering. As a scientific discipline, human factors and ergonomics has accumulated, and must continue to accumulate, a substantial body of knowledge from systematic studies that relate to human–system integration (Dul et al., 2012). As an engineering discipline, human factors and ergonomics has applied, and needs to continue to apply, the principles and knowledge gained from scientific studies to the specification and design of technology, products, and systems. By continuing to advance on both the scientific and engineering fronts, the human factors and ergonomics discipline will increase its contribution to the design of systems that enhance effective interactions between humans and their environment.

In this article, we have provided a high level overview of human factors and ergonomics. For further reading we recommend the following texts:

- Karwowski, W., 2006. International encyclopedia of ergonomics and human factors. CRC/Taylor & Francis, Boca Raton, FL.
- Salvendy, G., 2006. Handbook of human factors and ergonomics. John Wiley & Sons, Hoboken, NJ.
- Wickens, C.D., Lee, J.D., Liu, Y., Becker, S.E., 2004. An introduction to human factors engineering. Pearson Prentice Hall, Upper Saddle River, NJ.

See also: Decision Making, Psychology of; Decision Making: Nonrational Theories; Group Processes, Social Psychology of; Industrial–Organizational Psychology: Science and Practice; Organizational Climate in the Work Setting; Technological Change and the Sociotechnical System, Applied Psychology of; Work Motivation; Workplace Stress.

Bibliography

Ahram, T.Z., Karwowski, W. (Eds.), 2013. Advances in Physical Ergonomics and Safety. CRC Press, Boca Raton, FL.

BEA, 2012. Final report: On the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro - Paris. Ministry of Ecology, Sustainable Development, Transport and Housing, France.

Bridger, R.S., 2003. Introduction to Ergonomics. Taylor & Francis, New York.

Bullemer, P.T., Reising, D.V., 2013. Improve the operations team situation awareness: lessons learned from major process industry incidents. Paper presented to the American Fuel & Petrochemical Manufacturers (AFPM) Annual Meeting San Antonio, TX, 17–19 March.

Campion, M.A., Thayer, P.W., 1985. Development and field evaluation of an interdisciplinary measure of job design. Journal of Applied Psychology 70, 29–43.

Campion, M.A., Thayer, P.W., 1989. How do you design a job? Personnel Journal 68, 43–46.

Cannon-Bowers, J.A., Salas, E., 2006. Team effectiveness and competencies. In: Karwowski, W. (Ed.), International Encyclopedia of Ergonomics and Human Factors, second ed. CRC Press, Boca Raton, FL, pp. 2379–2383.

Chapanis, A., 1996. Human Factors in Systems Engineering. Wiley, New York.

Columbia Accident Investigation Board, 2003. Columbia Accident Investigation Board Report, vol. 1. National Aeronautics and Space Administration, Washington, DC.

Cooke, N.J., Durso, F.T., 2008. Stories of Modern Technology Failures and Cognitive Engineering Successes. CRC Press, Boca Raton, FL.

Cooke, N.J., Gorman, J.C., 2006. Assessment of team cognition. In: Karwowski, W. (Ed.), International Encyclopedia of Ergonomics and Human Factors, second ed. CRC Press, Boca Raton, FL, pp. 271–276.

Day, M.C., Young, C., 2012. This is your heart speaking. Call 911. Ergonomics in Design: The Quarterly of Human Factors Applications 20, 4–12.

Dominguez, C., 1994. Can SA be defined? In: Vidulich, M., Dominguez, C., Vogel, E., McMillan, G. (Eds.), Situation Awareness: Papers and Annotated Bibliography. Air Force Systems Command, Wright-Patterson Air Force Base, OH.

Dul, J., Bruder, R., Buckle, P., Carayon, P., Falzon, P., Marras, W.S., Wilson, J.R., Van Der Doelen, B., 2012. A strategy for human factors/ergonomics: developing the discipline and profession. Ergonomics 55, 377–395.

Endsley, M.R., 1988. Design and evaluation for situation awareness enhancement. Proceedings of the 23rd Annual Meeting of the Human Factors and Ergonomics Society 32, 97–101.

Gentner, D., Stevens, A.L., 1983. Mental Models. L Erlbaum Associates, Hillsdale, NJ.

Hall, D.S., Shattuck, L.G., Bennett, K.B., 2012. Evaluation of an ecological interface design for military command and control. Journal of Cognitive Engineering and Decision Making 6, 165–193.

Harris, D. (Ed.), 2013. Engineering Psychology and Cognitive Ergonomics: Understanding Human Cognition. 10th International Conference, EPCE 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, 21–26 July 2013, Proceedings Part. Springer, New York.

Health and Safety Executive, 1994. The explosion and fires at the texaco refinery, Milford Haven, 24 July 1994: A report of the investigation by the Health and Safety Executive into the explosion and fires on the Pembroke Cracking Company Plant at the Texaco Refinery, Milford Haven on 24 July 1994. HSE, UK.

Hendrick, H.W., 2005. An overview of macroergonomics. In: Moray, N. (Ed.), Ergonomics: Major Writings. Taylor & Francis, New York, pp. 38–58.

Hendrick, H.W., Kleiner, B.M., 2002. Macroergonomics: Theory, Methods, and Applications. Lawrence Erlbaum Associates, Mahwah, NJ.

Hollnagel, E., 2014. Human factors/ergonomics as a systems discipline? "The human use of human beings" revisited. Applied Ergonomics 45, 40–44.

Hollnagel, E., Woods, D.D., 1983. Cognitive systems engineering: new wine in new bottles. International Journal of Man-machine Studies 18, 583–600.

Hopkins, A., 2012. Disastrous Decisions: The Human and Organisational Causes of the Gulf of Mexico Blowout. CCH Australia Limited, Sydney, Australia.

Humphrey, S.E., Nahrgang, J.D., Morgeson, F.P., 2007. Integrating motivational, social, and contextual work design features: a meta-analytic summary and theoretical extension of the work design literature. Journal of Applied Psychology 92, 1332–1356.

Jamieson, G.A., 2007. Ecological interface design for petrochemical process control: an empirical assessment. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans 37, 906–920.

Jastrzebowski, W.B., 1857[2001]. An outline of ergonomics, or the science of work based upon the truths drawn from the science of nature (Originally published in four parts in Nature and Industry, 29, 227–231; 30, 236–244; 31, 244–251; 32, 253–258). In: Karwowski, W. (Ed.), International Encyclopedia of Ergonomics and Human Factors. Taylor & Francis, New York, pp. 161–175.

Jones, D.G., Endsley, M.R., 1996. Sources of situation awareness errors in aviation. Aviation, Space, and Environmental Medicine 67, 507–512.

Kaber, D.B., Endsley, M.R., 1998. Team situation awareness for process control safety and performance. Process Safety Progress 17, 43–48.

Kahneman, D., 2011. Thinking, Fast and Slow. Penguin Group, London, England.

Karwowski, W., 2006. The discipline of human factors and ergonomics. In: Sakvendy, G. (Ed.), Handbook of Human Factors and Ergonomics. John Wiley & Sons, Hoboken, NJ, pp. 1–37.

Klein, G.A., 2009. Streetlights and Shadows: Searching for the Keys to Adaptive Decision Making. MIT Press, Cambridge, MA.

Klein, G.A., Orasanu, J., Calderwood, R., Zsambok, C.E., 1993. Decision Making in Action: Models and Methods. Ablex Publishing, Norwood, NJ.

Klinger, D., Klein, G., 1999. An accident waiting to happen. Ergonomics in Design: The Quarterly of Human Factors Applications 7, 20–25.

Kroemer, K.H.E., Grandjean, E., 2009. Fitting the Task to the Human: A Textbook of Occupational Ergonomics. Taylor & Francis, Bristol, PA.

Leedom, D.K., Simon, R., 1995. Improving team coordination: a case for behavior-based training. Military Psychology 7, 109–122.

Leveson, N.G., Daouk, M., Dulac, N., Marais, K., 2003. Applying STAMP in Accident Analysis. Proceedings of the second workshop on the Investigation and Reporting of Accidents (IRIA03), 177–198.

Marras, W.S., Karwowski, W., 2006. The Occupational Ergonomics Handbook. CRC/Taylor & Francis, Boca Raton, FL.

Moray, N., 1986. Monitoring behavior and supervisory control. In: Boff, K.R., Kaufman, L., Thomas, J.P. (Eds.), Handbook of Perception and Human Performance. Wiley, New York.

Moro, F.B.P., 2009. Macroergonomics and information systems development. International Journal of Human-Computer Interaction 25, 414–429.

National Transport Safety Board, 1973. Eastern Air Lines Inc. L-1011, N310EA, Miami, Florida, 29 December 1972. Aircraft Accident Report. National Transport Safety Board, Washington, DC.

National Transport Safety Board, 2010. Loss of Thrust in Both Engines After Encountering a Flock of Birds and Subsequent Ditching on the Hudson River US Airways Flight 1549 Airbus A320-214, N106US Weehawken, New Jersey, 15 January 2009. Aircraft Accident Report. National Transport Safety Board, Washington, DC.

Pan, W.S.Y., Werblow, J., 2012. Does good job enrichment policy and practices impact employee's job satisfaction? Journal of Global Business Issues 6, 1–5.

Parasuraman, R., Sheridan, T.B., Wickens, C.D., 2008. Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. Journal of Cognitive Engineering and Decision Making 2, 140–160.

Pinder, C.C., 2007. Work motivation. In: Rogelberg, S. (Ed.), Encyclopedia of Industrial and Organizational Psychology. SAGE Publications Inc., Thousand Oaks, CA.

Rasmussen, J., Pejtersen, A.M., Goodstein, L.P., 1994. Cognitive Systems Engineering. Wiley, New York.

Reason, J.T., 1997. Managing the Risks of Organizational Accidents. Ashgate, Aldershot, UK.

Rogers, W.P., Armstrong, N.A., Acheson, D.C., Covert, E.E., Feynman, R.P., Hotz, R.B., Kutyna, D.J., Ride, S.K., Rummel, R.W., Sutter, J.F., Walker Jr, A.B.C., Wheelon, A.D., Yeager, C.E., 1986. Report of the Presidential Commission on the Space Shuttle Challenger Accident. Washington, DC.

Rousseau, D.M., 1977. Technological differences in job characteristics, employee satisfaction, and motivation: a synthesis of job design research and sociotechnical systems theory. Organizational Behavior and Human Performance 19, 18–42.

Salas, E., Muniz, E.J., Prince, C., 2006. Situation awareness in teams. In: Karwowski, W. (Ed.), International Encyclopedia of Ergonomics and Human Factors, second ed. CRC Press, Boca Raton, FL, pp. 903–906.

Sanders, A.F., 1998. Elements of Human Performance: Reaction Processes and Attention in Human Skill. Lawrence Erlbaum Associates, Mahwah, NJ.

Staszewski, J.J., 2004. Models of human expertise as blueprints for cognitive engineering: applications to landmine detection. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 48, 458–462.

Stout, R.J., Salas, E., Fowlkes, J.E., 1997. Enhancing teamwork in complex environments through team training. Group Dynamics: Theory, Research, and Practice 1, 169–182.

Umble, E.J., Umble, M.M., 2002. Avoiding ERP implementation failure. Industrial Management 44, 25–33.

Vicente, K.J., 1999. Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-based Work. Lawrence Erlbaum Associates, Mahwah, NJ.

Vicente, K.J., Rasmussen, J., 1992. Ecological interface design: theoretical foundations. IEEE Transactions on Systems, Man, and Cybernetics 22, 589–606.

Wheelan, S.A., Burchill, C.N., Tilin, F., 2003. The link between teamwork and patients' outcomes in intensive care units. American Journal of Critical Care 12, 527–534.

Wickens, C.D., Lee, J.D., Liu, Y., Becker, S.E.G., 2004. An Introduction to Human Factors Engineering. Pearson Prentice Hall, Upper Saddle River, NJ.

Zionchenko, V., Munipov, V., 2005. Fundamentals of ergonomics. In: Moray, N. (Ed.), Ergonomics: Major Writing. Taylor & Francis, New York, NY, pp. 17–37.

Human Genome Diversity Project: History

Joanna Radin, Yale University, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Human Genome Diversity Project was intended to be a large-scale effort to sample and archive human genetic variation. It was initially proposed by geneticists and evolutionary biologists in the early 1990s and was soon met by strong criticism. These criticisms came from multiple fronts, including biological and cultural anthropologists as well as indigenous rights activists and community members. Project organizers, who identified as antiracist and uninterested in reaping commercial benefits, were surprised to learn that many of these critics perceived the project as racist and neocolonial. These concerns led the initiative to falter, particularly in the United States. While the Human Genome Diversity Project did not achieve its original aims, it demonstrated the need to attend to the intertwined ethical and epistemological challenges of investigating human difference in a genomic age.

The Human Genome Diversity Project

The Human Genome Diversity Project (hereafter referred to as the Diversity Project) was intended as a large-scale effort to sample and archive human genetic variation. It was publicly proposed by geneticists and evolutionary biologists in the summer of 1991, as part of an effort to overcome a perceived Eurocentric bias in the Human Genome Project (Cavalli-Sforza et al., 1991).

The Diversity Project was also positioned as urgent. Organizers highlighted the importance of collecting genetic material from indigenous populations who were perceived to be vanishing (Roberts, 1991). Although the Diversity Project was not conceived of as a biomedical endeavor, justifications for the focus on such communities situated them both as normal controls and natural genetic laboratories for the detection of inherited diseases.

Those who supported the Diversity Project were self-described antiracist scientists. They were also disinterested in reaping commercial profits from the enterprise (Reardon, 2005). So, it was an unpleasant surprise for them to be accused of racism, colonialism, and biopiracy.

It is not an overstatement to say that the Diversity Project has been central to the formulation of many social scientists' engagement with past, present, and future horizons of the life sciences. At stake in social scientific interest in the Diversity Project are several interrelated questions: Who gets to speak authoritatively about what it means to be biologically human? Who are the bodies from which such biological understandings are built? And, who can benefit from that knowledge?

Antecedents

In accounting for why Diversity Project organizers were caught off guard by criticism of their initiative, it is important to recognize that their proposal was not without precedent. Attempts to map 'human biodiversity' preoccupied nineteenth-century anthropologists and eugenicists (Marks, 1995a). Those projects focused primarily on morphology; that is, what could be seen from looking at or measuring a body or a bone. This emphasis on what could be seen with the naked eye began to shift in the twentieth century with the rise of serology, which depended on the detection of blood groups and immunological markers (Silverman, 2000). An increasingly molecularized view of human difference gained momentum after World War II as these serological techniques were applied in view of the new science of human population genetics.

Population genetics was the interwar product of what has been called the 'evolutionary synthesis' in biology, which reconciled Darwinian ideas about evolution with Mendelian concepts of inheritance (Provine, 1971). Much of the initial mathematical theory of this approach was worked out through the study of observable traits in small interbreeding groups of nonhuman organisms, like fruit flies. Soon, however, interest grew in attempting to extend the techniques of population genetics to humans (Smocovitis, 2012). In the wake of genocide legitimated by Nazi racial science, the quantitative approach of population genetics appeared to provide a more politically neutral and objective approach to the study of human difference.

In 1950, Theodosious Dobzhansky, a leading fruit fly population geneticist, and Sherwood Washburn, a prominent physical anthropologist, brought together colleagues in both fields for a meeting to outline strategies for implementing a research program in human population genetics (1950). The meeting was held at Cold Spring Harbor Laboratories on Long Island. It was there, at the site of the Eugenics Record Office from the 1910s to the 1940s, that the value of a human population genetics focused primarily on geographically isolated and so-called primitive populations was emphasized.

Many of the ideas presented at the 1950 Cold Spring Harbor Symposium were put into practice by attendees who were subsequently affiliated with the ecologically oriented International Biological Program – an ambitious effort to take stock of the biosphere – which ran from 1964 to 1974. Under the auspices of the International Biological Program's Human Adaptability arm, population geneticists and anthropologists began to collect genetic material, including blood, from various populations, especially those they identified as 'primitive' and on the verge of disappearing (Santos, 2002). Although they also gathered a range of measurements and

materials, many believed that the study of blood samples using old serological and new biochemical techniques would provide more objective evidence for the characterization of genetic variation in humans. Instead of looking for observable traits, they could use serological and other biochemical techniques to locate variability in the blood and then use mathematical analyses to extrapolate frequency and patterns of inheritance. These scientists also believed that it was important to preserve the blood they salvaged, such that it could continue to be studied as new molecular techniques were developed. Unprecedented access to technologies of cold storage and air travel greatly facilitated the stockpiling of hundreds of thousands of blood samples (Radin, 2013).

Although the Human Adaptability arm of the International Biological Program was relatively uncontroversial, at least compared to the Diversity Project, it would be inaccurate to say that it did not receive any criticism at all. At the time of its planning, certain human biologists criticized the International Biological Program's focus on 'primitive' groups, pointing to the growing sense that such classifications were illusory and even racist. Other anthropologists, while not opposing the study of such communities per se, were disdainful of geneticists' encroachment into the sites of their fieldwork. Geneticists, they often complained, relied on anthropologists to do the important work of learning family histories and gaining trust, but did not repay those contributions with coauthorship credits.

Luca Cavalli-Sforza, who would become a leader in the Diversity Project, was a participant in the blood collection efforts of the Human Adaptability arm of the International Biological Program, at times relying on the expertise of a cultural anthropologist named Colin Turnbull. The criticism Cavalli-Sforza faced in the 1990s for having proposed a project that was so similar to the stocktaking work of the International Biological Program presented challenges that he was unprepared to effectively negotiate (Santos, 2002). While the seeds of this pushback had already been sown in the 1960s, it was amidst a tighter coupling of science, medical markets, and global communication that they began to bear bitter fruits (Lewin, 1993).

Initial Controversy

The planning for the Diversity Project was structured by a series of three meetings funded by the National Science Foundation and the Department of Energy. (The latter grew out of the Atomic Energy Commission, which had long been interested in studying the impact of radiation on the human genome, including during the International Biological Program.) During this planning stage, the Diversity Project was criticized by a number of groups.

Among those who spoke out against the proposed Diversity Project were biological anthropologists who had training in population genetics. They were mistrustful of an effort that sought to understand human diversity but that failed to explicitly include anthropologists, especially those whose careers had been devoted to the genetic bases of human evolution and variation (Marks, 1995b). Cultural anthropologists expressed mistrust as well. Some of their concerns pertained to the categorization of indigenous peoples as 'primitive' and therefore closer to nature. They viewed this as a form of essentialism that perpetuated rather than overturned a long history of scientific racism. Still other critiques expressed concern about the potential for indigenous peoples to be subjected to a new extractive form of colonialism linked to an emerging biotechnology industry (Lock, 1994; Hayden, 1996; Haraway, 1997).

The Diversity Project also received criticism from outside academia. The first such criticisms were leveraged by an activist group called the Rural Advancement Foundation International (RAFI). RAFI had been vocal on issues of intellectual property and biodiversity for a number of years, although they had previously focused on agricultural issues. RAFI leadership contacted a number of indigenous rights organizations, alerting them to the Diversity Project. Drawing connections between plant bioprospecting and genetic sampling of human genetic material, in June of 1993 RAFI charged that Diversity Project organizers were focused on preservation of potentially valuable genetic material at the expense of conservation of indigenous life ways (RAFI, 1993).

RAFI also attempted to establish a link between the Diversity Project and the patenting of indigenous DNA, which Diversity Project organizers vigorously denied as a goal of their effort. At the root of that particular accusation was the recognition of US National Institutes of Health policies that encouraged patenting of innovations resulting from federally funded biomedical research. Although individual participants in the Diversity Project opposed patenting, they were unsuccessful in diffusing suspicion about their relation to commercial interests. A few months later, the Third World Network, an important source of information on the needs and rights of people in the global south, echoed RAFI's call for an immediate halt to the Diversity Project. It was at the end of 1993 that the World Council of Indigenous Peoples branded the initiative with the label 'Vampire Project,' in which Diversity Project organizers were accused of extracting the lifeblood of indigenous peoples for their own benefit (Reardon, 2005).

Old Questions, New Contexts, Unexpected Outcomes

While the Diversity Project's emphasis on collecting blood from vanishing communities was very similar to goals of the International Biological Program, many things had changed in the approximately 20 years between the two initiatives. Among the most significant were the emergence of an indigenous rights movement and the development of new technologies, namely, the ability to amplify and sequence DNA.

Although scientists had long known about the molecular structure of DNA, making sense of it remained an arduous and highly complex task. This changed in the late 1980s with the advent of a technique known as polymerase chain reaction (PCR). It became easier to work with small amounts DNA, which could in turn provide huge amounts of genetic data. Scientists began to thaw blood samples collected during the International Biological Program in an attempt to recover bits of DNA that had been preserved (Merriwether, 1999).

One of the other significant biomedical breakthroughs of that period was a reliance on preserved cell lines that would allow for the continual production of high-quality DNA.

A white blood cell that was 'transformed' into a cell line by being cultured with a virus would continue to produce DNA. Such transformed cell lines came to be referred to as 'immortal,' because they were now able to outlive the bodies from which they were extracted and to be used for unanticipated purposes (Lock, 2001; Landecker, 2007). Improvements in the practice of tissue culture motivated Diversity Project scientists to collect new blood with the intention of extracting, manipulating, and thus maintaining the cells contained within.

New genomic practices and routines such as the reliance on cell lines and use of PCR not only provided a justification for revisiting the study of biological diversity but also contributed to the construction of race (M'charek, 2005; Reardon, 2005).

For example, one of the earliest debates about the Diversity Project actually took place between two of its organizers, Cavalli-Sforza and geneticist Allan Wilson. At the crux of their disagreement was the most appropriate way to determine what would count as a population. Cavalli-Sforza advocated for an approach that had been used during the International Biological Program: the selection of populations based on language spoken, which he believed correlated with genes. Wilson favored an approach that used a grid to sample populations at regular intervals around the globe. Ultimately, concerns about the cost and time involved in implementing Wilson's grid proposal led organizers to remain committed to a mode of identifying populations in terms of language. Any residual resistance was mooted by Wilson's untimely death from cancer in June of 1991.

The idea that the Diversity Project was a 'Vampire Project' or a new form of biologically based colonialism drew on widespread discomfort with the idea that body parts could be alienated from persons yet kept alive (Lock, 2001). It was a concern that was felt deeply and rippled widely, riding the waves of a coalescing indigenous rights movement. The indigenous rights movement built on several decades of human rights work that had contributed to situating identity as a strategic resource in political struggles.

Some of these struggles were articulated in a special 1996 issue of the journal *Cultural Survival Quarterly*, commissioned to respond to the charges RAFI and others had made against the Diversity Project. Jonathan Friedlaender, the then Program Officer for Physical Anthropology at the US National Science Foundation, was asked to serve as a guest editor (Friedlaender, 1996). Friedlaender, who had conducted his first fieldwork in Melanesia during the International Biological Program and who had for many years used blood samples from members of communities in that region to trace patterns of human diversity found himself at the center of controversy surrounding the use of indigenous DNA when he was wrongly accused by RAFI of attempting to patent Melanesian genes.

The special journal issue featured short essays from a range of individuals, including scientists affiliated with the Diversity Project as well as indigenous rights activists and a representative of RAFI. Solomon Island activist Ray Liloqula suggested that there was a lack of alignment in the interests of scientists and the people they sought to study. He stated that "Indigenous people know who they are, where they come from, what genetic traits are in their families, and have always known where they are going. The problem at the present time is having enough knowledge to cope with the newly introduced ways of life, sickness, and social problems, problems associated with new way of economic and political development, and from being exposed to the wider world" (Liloqula, 1996). Aroha Te Pareake Mead, a Maori activist, argued that, "Western science goes to great lengths to de-humanize the humanness or life-force of human genes; hence, terms such as 'specimens', 'materials', 'properties', and 'collections' are adopted as a means to ignore the essence of life contained within … the same perspective is carried over to issues of replication, immortalization, transgenic engineering, and cloning" (Mead, 1996).

Diversity Project organizers working in North America responded to these critiques by attempting to develop a set of ethical practices intended to ensure that people understood the goals of the project and had enrolled of their own volition. This culminated in the drafting of a 'model ethical protocol,' which yielded the idea of group consent (Greely, 1998). From the perspective of these Diversity Project organizers, group consent was an ethically innovative extension of the biomedical practice of obtaining informed consent to participate in research from individuals. However, the assumptions embedded in the well-intentioned ideas of group consent and research partnership sparked new concerns that directed criticisms back to the epistemological foundations of the project and its focus on populations (Reardon, 2005).

The model ethical protocol also attempted to recast indigenous peoples as research partners. This intervention was received differently both across and within various US communities. Some African-American biological anthropologists felt that any effort to enhance the diversity and participation in the Human Genome Project should also include attention to the genomes of those of African ancestry. In response, Diversity Project organizers sought to expand the list of populations to be studied beyond those seen as rare and endangered to include also African-Americans and Mexican-Americans. At the same time, other African-American biological anthropologists resisted the call for participation, invoking a dark legacy of experimentation, as in the notorious Tuskegee syphilis experiments. Hopi geneticist Frank Dukepoo felt that while some American Indians would willingly participate, many others were likely to perceive an invitation to participate in genetic research as an extension of several centuries of colonization and assimilation (Dukepoo, 1998).

What united all these responses was the concern that despite efforts to make the Diversity Project more inclusive, organizers would retain the authority to define the key questions guiding the research. Those questions might not be those of greatest interest to various participants and, perhaps more insidiously, their answers might produce new kinds of racially tinged harms.

Ultimately, these criticisms caused the Diversity Project to falter in the United States. Yet, in the European Union, laboratories were created to support the study of European genetic diversity. Eventually, the US National Research Council recommended that a small number of research projects be funded. Consortia for the creation and storage of cell lines were also established.

In April 2002, under the rubric of the Human Genome Diversity Project, 1064 cell lines made from the human tissues of individuals and population from around the globe were made available via the Fondation Jean Dausset Centre pour

l'Etude du Polymorphisme Humaine (CEPH) in Paris, France (Cavalli-Sforza, 2004; Cann et al., 2002). The cell lines were collected from various laboratories by the Diversity Project and CEPH and frozen on liquid nitrogen "to provide unlimited supplies of DNA for studies of sequence diversity and history of modern human populations" (Cann et al., 2002). DNA extracted from these samples has been circulated to over 100 researchers leading to more than 200 scientific papers (CEPH, 2000).

A significant portion of this research has been done under the auspices of the International Haplotype Map (also known as the HapMap), begun in 2002 as a public–private effort to link population-based variability to studies of complex diseases like diabetes, cancer, or heart disease (Gibbs et al., 2003). HapMap researchers used cell lines stored at CEPH and also created new ones from populations classified in terms of 'geographical ancestry' as Asian, European, and African. Although HapMap's organizers attempted to define the undertaking in opposition to the Diversity Project, this mode of delimiting populations inspired similar critiques, including concern about the reification or reinscription of race (Fujimura and Rajagopalan, 2011; Duster, 2006). Geographical ancestry appeared to be merely a new way of tracking older and potentially stigmatizing ideas about race (Hamilton, 2008).

Legacies

It is perhaps more meaningful to conclude that the Diversity Project achieved goals that were different than those proposed by its organizers than it is to conclude it was a failure. As one important commentator has observed, the debates sparked by the Diversity Project raised "questions about how to order and classify an aspect of nature to advance human understanding [which] proved inseparable from an allied set of questions about how to organize human differences for the purposes of creating credible and legitimate systems of governance" (Reardon, 2005: p. 3). The political dimensions of these debates were not only enacted in meetings and publications but in the laboratory as well. In the words of another influential commentator, "genetic diversity is not an object that lies waiting for scientists to discover, nor can it be treated as a construct of scientists. Genetic diversity is enacted in a complex scientific practice. It is not only dependent on the scientist and the DNA but also on the various technologies applied to produce it" (M'charek, 2005: p. 4).

These social scientific critiques of the Diversity Project were a result of bringing the practices of cultural anthropology – namely, ethnography – to a new kind of field site: the anthropological genomic laboratory and its allied sites of knowledge making. This reconfiguration of expertise has had the result of making cultural anthropologists, who had been largely excluded from the planning stages of the Diversity Project, newly relevant. It has also resulted in calls for novel biocultural approaches to the study of what it means to be human in the field of Science and Technology Studies as well as in Anthropology (Fullwiley, 2011; Palsson, 2008).

Despite the forces that derailed the Diversity Project in the 1990s, indigenous people – who have not, in fact, disappeared – continue to be the focus of human genetic diversity mapping projects. In addition to the HapMap, in 2005, National Geographic in partnership with the Waitt Family Foundation and IBM launched the Genographic Project. Led by population geneticist Spencer Wells, a student of Cavalli-Sforza, this effort aimed to continue the project of collecting genetic material from groups who were once again cast as vanishing. Like the Diversity Project, and the International Biological Project before it, the Genographic Project trades on notions of endangered diversity to encourage the salvage of genetic material from members of supposedly vulnerable indigenous communities (Tallbear, 2007). That such projects continue to be formulated and funded make clear that lessons of the Diversity Project are still unfolding, generating unexpected challenges for the natural and social sciences, alike (Kowal et al., 2013).

See also: Biobanking: Ethical Issues; Bioethics: Genetics and Genomics; Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute; Eugenics, History of; Expertise, Anthropological Approaches To; Genetics and Anthropology; Genetics and Indigenous Communities: Ethical Issues; Genetics and Social Justice; Genetics and Society; Genetics: The New Genetics; Genographic Project; Genomics, Ethical Issues In; Human Evolutionary Genetics; Human Genome Project: History and Assessment; Indigeneity; New Genetics and Race; Race: Genetic Aspects.

Bibliography

Cann, H., de Toma, C., Cazes, L., Legrand, M., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W., et al., 2002. A human genome diversity cell line panel. Science 296, 261–262.

Cavalli-Sforza, L.L., 2004. Diversity project takes time but reaps rewards. Nature 428, 467.

Cavalli-Sforza, L.L., Wilson, A., Cantor, C.R., Cook-Deegan, R., King, M.-C., 1991. Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the human genome project. Genomics 11, 490–491.

Cold Spring Harbor Symposia on Quantitative Biology. Cold Spring Harbor Labs, 1950. The Science Press, Cold Spring Harbor, NY.

Dukepoo, F.C., 1998. Commentary on "scientific limitations and ethical ramifications of a non-representative Human Genome Project: African American responses" (F. Jackson). Science and Engineering Ethics 4 (2), 171–180.

Duster, T., 2006. Molecular reinscription of race: unanticipated issues in biotechnology and forensic science. Patterns of Prejudice 40, 427–441.

Friedlaender, J.S., 1996. Genes, people, and property: furor erupts over genetic research on indigenous groups. Cultural Survival Quarterly 20, 22.

Fujimura, J.H., Rajagopalan, R., 2011. Different differences: the use of 'genetic ancestry' versus race in biomedical human genetic research. Social Studies of Science 41, 5–30.

Fullwiley, D., 2011. The Enculturated Gene: Sickle Cell Health Politics and Biological Difference in West Africa. Princeton University Press, Princeton.

Gibbs, R., et al., 2003. The international HapMap project. Nature 426, 789–796.

Greely, H.T., 1998. Informed consent, stored tissue samples, and the human genome diversity project: protecting the rights of research participants. In: Weir, R. (Ed.), Stored Tissue Samples: Ethical, Legal, and Policy Implications. University of Iowa Press, Iowa City.

Hamilton, J., 2008. Revitalizing difference in the HapMap: race and contemporary human genetic variation research. Race, Pharmaceuticals, and Medical Technology, Fall, 471–477.

Haraway, D.J., 1997. Modest_Witness@Second_Millennium. FemaleMan_Meets_OncoMouse: Feminism and Technoscience. Routledge, New York; London.

Hayden, C., 1996. A biodiversity sampler for the millennium. In: Franklin, S., Ragone, H. (Eds.), Reproducing Reproduction: Kinship, Power and Technological Innovation. University of Pennsylvania Press, Philadelphia.
Kowal, E., Radin, J., Reardon, J., 2013. Indigenous body parts, mutating temporalities, and the half-lives of postcolonial technoscience. Social Studies of Science 43, 465–483.
Landecker, H., 2007. Culturing Life: How Cells Became Technologies. Harvard University Press, Cambridge, MA.
Lewin, R., 29 May 1993. Genes from a disappearing world. New Scientist, 25–29.
Liloqula, R., 1996. Value of life: saving genes versus saving indigenous peoples. Cultural Survival Quarterly 20, 42. http://www.culturalsurvival.org/ourpublications/csq/article/value-life-saving-genes-versus-saving-indigenous-peoples (accessed 9.10.13.).
Lock, M.M., 1994. Interrogating the human genome diversity project. Social Science and Medicine 39, 603–606.
Lock, M.M., 2001. The alienation of body tissue and the biopolitics of immortalized cell lines. Body and Society 7, 63–91.
M'charek, A., 2005. The Human Genome Diversity Project: An Ethnography of Scientific Practice. Cambridge University Press, Cambridge, UK; New York.
Marks, J., 1995a. Human Biodiversity: Genes, Race, and History. Alaine de Gruyter, New York.
Marks, J., April 1995b. The human genome divesity project: good for if not good as anthropology? Anthropology Newsletter 36, 72.
Mead, A.T.P., 1996. Genealogy, sacredness, and the commodities market. Cultural Survival Quarterly 20, 46. http://www.culturalsurvival.org/ourpublications/csq/article/genealogy-sacredness-and-commodities-market (accessed 9.10.13.).
Merriwether, D.A., 1999. Freezer anthropology: new uses for old blood. Philosophical Transactions: Biological Sciences 354, 121–129.
Palsson, G., 2008. Genomic anthropology: coming in from the cold? Current Anthropology 49, 545–567.
Provine, W., 1971. The Origins of Theoretical Population Genetics. University of Chicago Press, Chicago, IL.
Radin, J., 2013. Latent life: concepts and practices of human tissue preservation in the international biological program. Social Studies of Science 43, 483–508.
RAFI, May 1993. Patents, indigenous peoples, and human genetic diversity. RAFI Communique.
Reardon, J., 2005. Race to the Finish: Identity and Governance in an Age of Genomics. Princeton University Press, Princeton, NJ.
Roberts, L., 1991. A genetic survey of vanishing peoples. Science 252, 1614–1617.
Santos, R., 2002. Indigenous peoples, postcolonial contexts, and genomic research in the late twentieth century: a view from Amazonia (1960-2000). Critique of Anthropology 22, 81–104.
Silverman, R., 2000. The blood group 'fad' in post-war racial anthropology. In: Marks, J. (Ed.), Kroeber Anthropological Society Papers. University of California Press, Berkeley.
Smocovitis, V.B., 2012. Humanizing evolution: anthropology, the evolutionary synthesis and the prehistory of biological anthropology, 1927-1962. Current Anthropology 53 (S5), 46–56.
Tallbear, K., 2007. Narratives of race and indigeneity in the genographic project. The Journal of Law, Medicine & Ethics 35, 412–424.

Relevant Websites

http://www.cephb.fr/en/hgdp/diversity.php/ – CEPH, 2000.
http://www.councilforresponsiblegenetics.org/.
http://www.etcgroup.org/.
https://genographic.nationalgeographic.com/.
http://hagsc.org/hgdp/index.html.
http://hapmap.ncbi.nlm.nih.gov/.
http://indigenousgenomicsgovernance.org/.

Human Genome Project: History and Assessment

Hub Zwart, Radboud University of Nijmegen, Nijmegen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The *Human Genome Project* (HGP) has been hailed as an important milestone in the history of science, in the history of humanity even, and as a project whose completion would not only transform the practice of medicine, but change forever the course of human history as well. By setting up a flanking program devoted to anticipating and addressing the ethical, legal, and social implications of genomics, the HGP has had a substantial influence on the social sciences and humanities fields involved in studying contemporary science as well. In this article, the HGP is first of all placed in a historical perspective (as a key chapter in the history of the life sciences as such). As to the history of the HGP proper (1990–2003), three stages are distinguished. Finally, the author assesses what the impact of the HGP has been, both for the life sciences as such and for the society in a broader sense. Although biomedical benefits (in the form of new treatments, etc.) have been sparse (in comparison to the stellar expectations of the early years), the HGP did change the way we think about ourselves and our own history: Its 'cultural' relevance has been quite significant.

Introduction

The *Human Genome Project* (HGP) has been hailed as an important milestone in the history of science, in the history of humanity even, and as a project whose completion would not only 'transform the practice of medicine' but 'change forever the course of human history' (Davies, 2001/2002, p. xiii). It was argued that, due to the HGP, the 'childhood of the human race is about to come to an end' (Davies, 2001/2002, p. 10). It is biology's version of the Manhattan Project, or, as its Director Francis Collins phrased it in 1993, the HGP is 'more important than putting a man on the moon or splitting the atom' (Davies, 2001/2002, p. 69). From a history of biology perspective, the HGP (whose imminent completion was announced at a Press Conference in June 2000) has been described as the completion of the 'century of the gene' (Fox-Keller, 2000) that began with the rediscovery of the work of Gregor Mendel in the spring of 1900, while the complete sequence became available in 2003, exactly 50 years after the discovery of the structure of DNA by James Watson and Francis Crick. Many have argued that the HGP represents a 'paradigm shift' in biology (Gilbert, 1991), dramatically changing the manner in which biomedical research is done (Collins, 1999). 'Gene hunting' gave way to a much more systematic and data-driven approach. ('Data driven' describes a form of research that, rather than relying on the conscious designing of experiments to test hypotheses, relies on amassing large amounts of data that can be systematically analysed, browsed, and trawled with the help of computer programs looking for patterns and associations.) Enormous amounts of sequencing data were assembled in large-scale databases, allowing for a more comprehensive analysis of living systems.

Besides a series of high-impact papers in top journals such as *Nature* and *Science* (notably International Human Genome Sequencing Consortium IHGSC, 2001, 2004; Venter et al., 2001), the project has unleashed a cascade of publications of breath-taking proportions, covering various genres (research articles, books, comments, interviews, etc.), not only in the life sciences, but also in the social sciences and the humanities. In this article, this quickly expanding library of HGP discourse can only be analysed in brief outline. It will consist of three parts. The central part entails a concise history of the HGP in which three stages will be distinguished. Yet, as the HGP can only be properly understood against the backdrop of previous developments that evolved in the course of the twentieth century, the history of the HGP proper will be preceded by a brief 'prehistory' in which the main events that lead to the launching of the HGP are outlined. The third and final part will be devoted to assessing the meaning and significance of the HGP in retrospect. How do we assess its importance, both for the life sciences as such and for society at large? Seen from this perspective, the HGP no longer functions as an end point (i.e., the completion of century of the gene), but rather as a point of departure (for next generation genomics and other postgenomics developments such as systems biology and synthetic biology).

Prehistory: The Century of the Gene

Michel Foucault's inaugural lecture as Professor of the History of Systems of Thought at the Collège de France (Foucault, 1971) contains a brief but interesting reflection on Gregor Mendel. How could it be, Foucault asks his audience, that Mendel's work, his 'truth,' was more or less completely ignored by the biologists of his time? For Foucault, Mendel was a kind of 'monster' (p. 37), someone who, in 1865, spoke of objects and relied on methods that were completely alien to the biology of his era. In short, Foucault casts him as someone who lived and worked outside established networks, outside official academic discourse: A lost voice who spoke too soon.

A key aspect of Mendel's untimely approach (of his 'epistemic mutation') consisted in the use of a small alphabet of symbols (Aa, Bb, Cc, etc.) to refer to pairs of (dominant or recessive) 'factors,' as he called them (1866/1913), that could be either present or absent in his model organism, the garden pea. What was so rigorously new, according to Foucault, was that Mendel spoke about these factors as autonomous units,

independent of biological categories such as species or sex, as objects of study in their own right. All the rest, everything pertaining to the garden pea as a living organism, was filtered out as it were. For Mendel, *Pisum sativum* was simply a handy research tool for studying what later would come to be called 'genes.' From 1900 onward, this way of looking at the phenomena of life (studying genes with the help of a select number of model organisms that basically functioned as research contrivances, as items in the scientific tool box) came into vogue, and genetics became a core discipline of biology.

In recent decades, this picture of Mendel as a garden monk living in extreme isolation has been revised. There were kindred minds to be found among his contemporaries, who were working and thinking along similar lines. One of them was Francis Galton who, like Mendel, used statistics to study life (Müller-Wille and Rheinberger, 2012). In order to articulate his views on heredity, Galton employed an interesting metaphor:

> Ova and their contents are, to biologists looking at them through microscopes, much what mail-bags and the heaps of letters poured out of them are to those who gaze through the glass windows of a post office … they cannot read a single word of what the letters contain (1875, p. 82).

What Mendel and Galton had in common was the intuition that the basic units of life could be represented as letters of an alphabet. Whenever biologists peered into the nucleus of an ovum through their microscope, they witnessed a process that was comparable to the processing of piles of documents containing characters, but due to the low resolution power of the instruments at their disposal, these documents were as yet unreadable for the scientists of life. Nonetheless, they were written in an alphabet (of a Mendel-like nature) and transferred as it were by ancestors to their descendants (Galton's ovum as a kind of transgenerational communication establishment). In fact, in ancient Greek, the term 'elements' (*stoicheia*) means building blocks and may therefore refer not only to atoms as the building blocks of matter, or to Euclid's building blocks of elementary geometrical knowledge, but also to letters of the alphabet.

One could argue that genomics (the sequencing of genomes, the entirety of an organism's hereditary information, with the help of automated high-throughput sequencing machines) from the very beginning has built on this metaphor. It constitutes an effort to dive deeper into the nuclei of cells, in order to decipher, scan and process these formerly unreadable letters, with the help of the high resolution provided by sophisticated automated sequencing technologies. Thus, the 'century of the gene' (Fox-Keller, 2000), which began with the rediscovery of Mendel's effort to talk about life in terms of a small alphabet of symbols, came to a conclusion with the famous Press Conference of 26 June 2000 when Bill Clinton, Francis Collins, and Craig Venter proudly announced that the HGP was galloping toward completion, producing spates of letters: the human genome sequence, written in the A, C, G, T alphabets the 'code of life.' (ACGT stands for the four nucleic acid bases that make up an organism's DNA. The 'A' stands for adenine, the 'T' for thymine, the 'C' for cytosine, and the 'G' for guanine. These four nucleic acids are the 'letters' of the genetic code.)

The story of what happened between these two poles of the human genome narrative (i.e., the rediscovery of Mendel in the spring of 1900 and the public announcement of the upcoming completion of the human genome sequence in June 2000) is a complex one. Only a few of the major events can be highlighted here.

In the 1930s and 1940s, to begin with, an important transformation took place in the landscape of science. The center of gravity shifted from Central Europe (Vienna, Berlin, and Zürich) to the United Kingdom and the United States, while the academic lingua franca shifted from German to English, and the frontier of the scientific revolution shifted from quantum physics to the molecular life sciences. This transformation was exemplified by two physicists who migrated not only from the German- to the English-speaking world, but also from quantum physics to biology. One of them was Nobel laureate Erwin Schrödinger who, in 1944, gave a series of lectures in Dublin entitled *What is life?* In these lectures, he voiced the idea that the 'genom' (as he spelled it) was basically an 'aperiodic crystal' containing a code comparable to the Morse code. Now that physicists had discovered the elementary building blocks of energy and matter, time had come to use the tested methods and powerful tools of physics (such as crystallography) to unravel the elementary building blocks of life.

At that time, however, Max Delbrück, although younger than Schrödinger, had already put to practice what the latter advocated. He went to Cambridge and later to Caltech in California where he opted for a virus (the bacteriophage) as his model organism. For Delbrück, genes were for biology what atoms had been for physics. And therefore, he began his new career as a biologist by looking for a 'minimal' organism that could somehow play the role that the hydrogen atom had played in quantum physics. This search for the 'hydrogen atom of biology' led to the bacteriophage (literally, the bacteria eater), a virus that attacks and destroys bacteria. He settled upon the phage as his model organism because he wanted to study a living entity that came as close as possible to being 'the gene in itself' ('Das Gen an sich;' Fischer, 1985, p. 98). Thus, Schrödinger and Delbrück played a pivotal role in the emergence of a new field: molecular biology. In fact, the three scientists who in 1962 were awarded the Nobel Prize for their discovery of the structure of DNA, namely, Francis Crick (a physicist), James Watson (a biologist), and Maurice Wilkins (a physicist), had all read (and confessed to be very much inspired by) Schrödinger's lectures, which had been published as a book in 1946, while James Watson took Delbück's course on bacteriophage research at Cold Spring Harbor. Thus, the discovery of the structure of DNA was more or less a direct result of the broader transformation described above: the massive migration of physicists to the new biology (cf Kay, 2000).

Now that the structure of DNA had been revealed, the next big objective was to read and decipher the 'code' embedded in it. After DNA (the double helix), the next big challenge was sequencing the whole genome (the collection of all the genes as well as the noncoding segments present in human DNA). In 1985, Robert Sinsheimer, President of the

University of California, Santa Cruz, and Charles DeLisi from the Department of Energy, organized meetings to discuss the feasibility of sequencing the human genome. One of the first questions to be addressed was whether this would call for a large-scale, international, top–down programmatic effort (Big Science), or rather for a bottom–up, researcher-driven approach. Most participants in the debate tended toward the former and in the course of the 1980s, one of the most ambitious and controversial projects in the history of biology was designed. The final preparations coincided with the collapse of the Berlin Wall and the subsequent reunification of Germany and, to some extent, of Europe as a whole. Euphoria was in the air. On 9 November 1989, as citizens from East Berlin flooded into the Western parts of the divided city, *Nature* published an article on HGP scientists developing a joint database where they could deposit their sequencing materials, flanked by an article about societal issues to be addressed. The project was formally launched on 1 October 1990, with James Watson as its first Director. In 1992, Watson stepped down and was succeeded by Francis Collins while Craig Venter, partly because of his frustrations with the bureaucracy and the power structure of the official HGP, launched a nonprofit, privately funded human genome sequencing project of his own devise. He would establish The Institute of Genomics Research and later Celera Corporation to achieve this goal.

The Human Genome Project: A Concise History

The history of the HGP can roughly be divided into three stages (Zwart, 2008). The first stage (1988–92) was a period of development and implementation, beginning with the appointment of James Watson at the National Institutes of Health (NIH) in 1988 and ending with the latter's resignation (on 10 April 1992). The highlight event was the official launch of the program on 1 October 1990. Simultaneously (also in 1990) a flanking project was launched to anticipate and address the ethical, legal, and social implications (ELSIs) of the sequencing effort, also funded by NIH. In fact, 3–5% of the budget was devoted to this type of research. Thus, NIH became the biggest sponsor of bioethical and other forms of ELSI research in the United States.

The second stage (1993–98) set in with the appointment of Francis Collins (1993) as Watson's successor. The HGP now really gained momentum. For the next few years, work on the HGP came to involve a relatively large number of research groups of varying sizes and sequence capacities. The human genome was divided into 23 natural subunits (the chromosomes) and research labs competed for funding in the context of the NIH peer-review system. By 1998, only 4% of the human genome had been sequenced in this manner. That same year, however, was a turning point in many ways: the onset of the third and final stage (1998–2001).

First of all, former NIH scientist Craig Venter announced that his privately funded Celera Corporation, relying on the latest automated sequencing machines, in combination with his controversial whole-genome shotgun method, would sequence the human genome much faster than the public consortium envisioned (Shreeve, 2004; Venter, 2007). This announcement entailed a direct challenge to the public effort headed by Collins, which was now under the threat of becoming redundant (Cook-Deegan, 1994/1995). The latter responded by formulating a more focused strategy of action. From now on, the bulk of the work would be done by a limited number of big laboratories with massive sequence capacity, such as the Sanger Centre lead by John Sulston (Cambridge, UK) and the Whitehead Institute lead by Eric Lander (Cambridge, Massachusetts): a select number of key players that came to be known as the G-5. Moreover, it was decided that publication of a finished sequence would be preceded by a working draft version, again a departure from the original plan (Davies, 2001/2002, p. 162). Due to this reorientation, the HGP was suddenly transformed into a 'no-holds-barred race to claim the human blueprint.' A dramatic competition ensued, eagerly covered by the media, until on 26 June 2000 the 'race' or 'war' was suddenly declared over, as both sides agreed they had passed the line more or less simultaneously. Collins and Venter cordially made their appearance at the famous press conference at the White House in order to announce, in the presence of President Clinton, that the endeavor to sequence the human genome was nearing its completion . Both teams sealed a truce by publishing working draft versions in a coordinated manner in February 2001, in *Science* and *Nature*, respectively (IHGSC, 2001; Venter et al., 2001), although the fighting between representatives of both 'armies' over who had actually won the race, or war, would continue in various settings until today. Notably, John Sulston continued to question the role of Celera in sequencing the human genome (Sulston and Ferry, 2002/2003).

After 2001, there has been a considerable aftermath. As indicated, the 2001 publication contained a draft version, and the HGP was officially declared complete on 14 April 2003, although a 'final' final version of the finished sequence (with 99.99% accuracy) was published in 2004 (IHGSC, 2004), while further analyses and papers on the HGP continue to occur. The human sequence is now seen as the starting point for future research. This is also visible in the images used to elucidate its meaning (Zwart, 2010). During the Press Conference of June 2000, Collins compared the HGP to the expedition of Lewis and Clark, commissioned by President Thomas Jefferson in the beginning of the nineteenth century to map the western part of North America from the Mississippi to the Pacific. This resulted in a famous map (below, Figure 1) depicting a more or less empty landscape, with a focus on its physical features, such as rivers and mountains. A few years later, when presenting a vision for the future of genomics research (Collins et al., 2003), a different picture emerged (below, Figure 2). Rather than with an open landscape, one is now faced with buildings and emerging infrastructures. The genomics future is envisioned as a three-floor building designed in the style of Frank Lloyd Wright. The basic topology, as it were, has shifted. The landscape has become inhabited. "All the initial objectives of the HGP have now been achieved," Collins et al. argue, "and a revolution in biological research has begun" (p. 835). It will now quickly spread from the basement to biology, health, and society.

Figure 1 The Lewis and Clark map. Source: http://memory.loc.gov/cgi-bin/query/r?ammem/gmd:@field(NUMBER+@band(g4126s+ct000763))

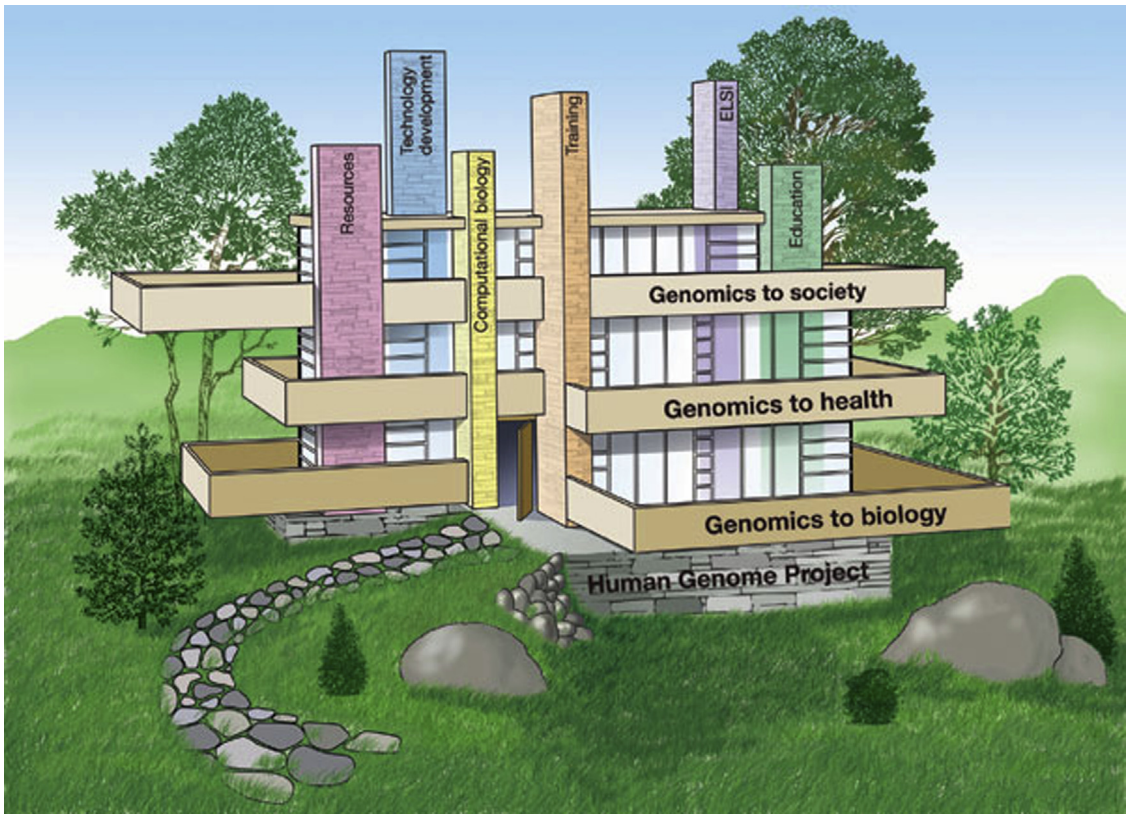

Figure 2 A vision for genomics. Source: Collins, Francis, Green, Eric D. Green, Guttmacher, Alan, Guyer, Mark S., [NHGRI], 2003. A vision for the future of genomics research, a blueprint for the genomics era. Nature 422, 835–847.

Figure 3 The 26 June 2000 Press Conference announcing the near completion of the HGP.

Meaning and Significance: Assessing the HGP

One of the questions raised with regard to the HGP has been to what extent it can really be regarded as the advent of 'Big Science' in biology (Vermeulen, 2009). Both Watson and Collins on various occasions placed the HGP on an equal footing with other Big Science endeavors such as the Manhattan Project, the Apollo Program, and the European Organization for Nuclear Research (CERN). The HGP was regarded as 'Big' by many, first of all because of the size of the funding involved (three billion dollars as core funding) in combination with the international scale of the collaboration as well as the fact that the project was managed in a more or less centralized manner by Watson and subsequently Collins as its Director. Yet, Big Science in physics (where the concept first emerged) is typically organized around a centralized large-scale research facility, attracting hundreds of researchers (such as the Large Hadron Collider of CERN in Geneva), whereas HGP's infrastructure remained much more distributed, while its basic technologies were not available from the outset, but were rather developed as the project evolved. Or, as Müller-Wille and Rheinberger phrase it in slightly Marxist terms, the HGP created not only a plethora of scientific products (data, papers, and the like), but also the means for their production (2012, p. 201).

This may be seen as a purely 'semantic' question, but in the context of ongoing controversies and reflections as to what the HGP has been able to achieve, the issue proves quite relevant. Building on the widespread impression that in the end (in terms of benefits for society) the HGP did not quite live up to the stellar ambitions that were voiced in the 1990s, some may now question the viability of a Big Science approach in the life sciences domain as such, whereas others may argue that, in view of the bewildering complexity of living systems, the HGP simply was not Big enough.

This brings us to the second issue over which controversy is still raging: what has been the societal impact of the human sequence? Although it is beyond doubt that genomics has irreversibly changed the life sciences as a global research field, so far, concrete clinical benefits, for instance, in terms of treating cancer, have been sparse. During the 26 June 2000 Press Conference, great expectations were voiced in this respect. Clinton claimed that, because of the HGP, our children's children will probably know the word cancer only as a constellation of stars. Yet, in many ways, the outcomes of the HGP were something of a disappointment. Let me go into this experience of disillusionment in somewhat more detail.

The first major disappointment had to do with the number of genes on the human genome, namely, 22 500, which was an astonishingly low number compared to previous estimates that ranged from 100 000 up to 350 000 (Zwart, 2007). The human genome, moreover, appeared to be quite similar to the genomes of other species, such as the chimpanzee and the laboratory mouse. There is nothing special about our sequence. This in itself constitutes a nice example of what Sigmund Freud (1917/1947) has termed a narcissistic offense. According to Freud, all major scientific breakthroughs (and he mentions the discoveries of Copernicus and Darwin as paradigmatic examples) entail serious narcissistic offenses that challenge our basic views about ourselves. We tend to see ourselves as highly unique and central, but the truth as laid bare by science is that we are not. Yet, we are the only species on earth able and willing to think about our own evolution or about our place in the universe, and we are certainly the only species who has sequenced its own genome (or may come up with such an idea at all). An interesting paradox: food for philosophy indeed. Where does our uniqueness come from?

As we cannot attribute it solely to our DNA, it must come from elsewhere. It must be closely interwoven with our cultural history, over and above our biological evolution. We must be, to a certain extent, manmade. We are not only the producers, but also the products of our technologies, cultures, and societies. It is precisely here that the HGP did have an impact. It has provided us with additional and valuable sources of information concerning early human history, notably by considering not only the human genome itself, but also the genome of others species (such as dogs, horses, rice, cereals, maize, etc.) as historical archives for bioarchaeological research (Zwart and Penders, 2011). In other words, the 'cultural impact' of genomics, by broadening our historical horizon, has been quite significant.

Thus, the experience of disillusionment notably pertained to the biomedical results. While it is widely acknowledged the HGP did have an enormous impact on the way in which biological research is done, its medical applications so far have been quite limited. Although the HGP did succeed in transforming laboratory life, so far the consequences for health care, medicine, and human daily existence in general have been quite modest. This is also reflected in the societal debate triggered by the HGP. Initially, serious concerns were voiced, for instance, by Francis Fukuyama (2002) who believed that the essence of what we are is endangered by biotechnology now that the human sequence is about to be revealed. He postulated the existence of a set of typically 'human' genes, which he termed the Factor X, a genetic endowment that distinguished humankind in essence from other species and should be safeguarded from biotechnological interference. In France, novelist Michel Houellebecq devoted his successful novel *Elementary Particles* (1998) to a scientist who managed to develop an algorithm that allowed humankind to drastically ameliorate its genetic constitution, thereby altering the course of future history (an event on the scale of a 'metaphysical mutation'). Yet, it soon became apparent that such scenarios are fairly unrealistic and that the Factor X is nowhere to be found. A related concern raised in the early days of the HGP was that of genetic discrimination: The idea that human genomics would lead to the emergence of a social underclass of carriers of 'bad genes,' as dramatized in the movie *Gattaca* (released in 1997), for example. Again, these apprehensions, grounded in rather deterministic understanding of the relationships between genes and traits, proved unwarranted. Life is too complex to be understood in such a linear and straightforward fashion. So far, the societal relevance of the HGP has been quite limited indeed. This is also reflected in ELSI research, where big concerns gradually gave way to much more everyday discussions.

In 2010, the journal *Nature* published a retrospective. The basic question was has the proclaimed genomics revolution arrived? Collins (2010a) and Venter (2010) both contributed to the issue. Their answer and that of other contributors was, essentially, not yet. The most important outcome of the HGP resides in the acknowledgment that life is much more 'complicated' (Hayden, 2010) than was envisioned back in 1989, when Francis Collins wrote his very first e-mail and the HGP was about to be launched.

This issue was taken up more elaborately by Francis Collins (2010b) (now Director of NIH) in his latest book *The Language of Life*, written 10 years after the jubilant Press Conference of June 2000. Although he still sees the HGP as 'one of the boldest scientific efforts that humankind has ever mounted' (p. 299), he at the same time acknowledges that the enthusiasm that had peaked during the days of the June 2000 Press Conference has more or less evaporated:

> Nearly a decade has passed since that moment of celebration. Virtually all biomedical researchers would agree that their approach to understanding how life works has been profoundly and irreversibly affected by access to the complete DNA sequence of the human genome … But the effect on the public of all the hoopla in 2000 has been mixed. Most people know that the genome has been spelled out, but they have lost track of what has happened since then. They remember the ascent of the mountain, but they are unaware of the rewards that are starting to appear in the valley. (p. 3)

The revolution is still raging, that is the basic message of his latest book; it is ongoing and societies better prepare themselves for a massive disruptive change, but this will not come through the sequencing of the human genome as such, but rather though its subsequent stage: The era of next-generation sequencing and the $1000 genome (Davies, 2010) and, as a consequence of that, the advent of personal genomics and personalized medicine:

> Healthy individuals are increasingly able to discover some of their body's inner secrets and take appropriate action. The potential for individual prediction is beginning to spill out to the general public, offering the opportunity to take more control of your fate … None of this is happening overnight, [but] without question, man's knowledge of man is undergoing the greatest revolution since Leonardo. (p. 5)

But still, up to this point at least, the transformations brought about by next-generation sequencing are limited. In comparison to the dramatic debates of previous decades, current deliberations focus much more on everyday and perhaps even trivial examples, such as the credibility of providers of personalized genome tests such as 23andMe.

One way of seeing this is to argue that, apparently, supporters of the HGP consciously created a hype in the 1990s, by willfully exaggerating its societal potential. Subsequently, after publication of the human sequence in 2001, they shifted the argument saying that life proved to be much more complex than was expected at the beginning, so that much more research (i.e., much more funding) will be needed to harvest the potential benefits. Another, less cynical, but in my view more credible, way of looking at it is by arguing that a project of the size of HGP is likely to go through a series of stages. From this perspective, Watson and Crick represent the era of childhood and playfulness, building models with toylike components without much regard for formalities, such as research projects or management structures. Subsequently, during the period of latency (the 1970s and 1980s), the focus shifted toward the development of basic tools and skills (such as the Sanger sequencing method) in preparation for a distant and still fairly abstract future. The HGP itself, then, represents the adolescence stage: the era of great expectations and fierce collisions. Finally, at present, human genome sequencing has entered the stage of adulthood and normal science. Genomics

has become embedded in the life sciences and as such, has acquired a certain level of normalcy, so that its implications for human existence are less dramatic, more everyday, perhaps even trivial, and its promises more realistic.

As was already indicated above, a significant impact of the HGP resides in its cultural relevance: The way we see ourselves and think about ourselves as humankind. First of all, genomics provides us with new sources of information, allowing us to revisit early human history in a detailed fashion, from the days of human migration out of Africa up to the emergence and spread of agriculture in the context of the so-called Neolithic revolution. Subsequently, it is clear that the HGP has prepared the ground for, and will gradually be eclipsed by, so-called postgenomic developments, such as personalized genomics (or next-generation sequencing) already mentioned above.

In the early days of the HGP, the human sequence was often presented as the 'blueprint' of human existence or even as 'the near-complete instructions for how to build and run a human body,' downloadable from the Internet (Ridley, 1999, p. 1). Paradoxically, the project that started off as an exemplification of a genetic deterministic mind-set eventually undermined the credibility of precisely the genetic reductionist view of life by revealing the astonishing complexity of living systems. For example, it paved the way for the emergence of epigenetics: the awareness that the vicissitudes of an organism (nurture) may result in modifications of its genome (notably on the level of gene expression), which can be passed on to the future generations. This goes against the long-standing dogma that acquired features cannot be inherited. Another development is the shift from 'reading' (i.e., sequencing) DNA to 'rewriting' DNA on the molecular level, that is synthetic biology: the intermediary zone between genomics and nanoscience. Similar realignments can be discerned elsewhere, for instance, between genomics and brain research (cognomics). Thus, the HGP has definitely strengthened the trend, already characteristic of contemporary technoscience as such, toward the emergence of a plethora of transdisciplinary research conglomerates (molecular biology, bioinformatics, genomics, synthetic biology, epigenetics, cognomics, etc.) by which genomics itself is bound to be engulfed sooner or later.

See also: Biobanking: Ethical Issues; Bioethics in the Post-genomic Era; Bioethics: Genetics and Genomics; Direct-to-Consumer Personal Genetic Testing; Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute; Genetics and Society; Genetics and the Media; Genetics: The New Genetics; Genographic Project; Genomics, Ethical Issues In; Human Evolutionary Genetics; Human Genome Diversity Project: History.

Bibliography

Collins, F., 1999. Medical and societal consequences of the Human Genome Project. New England Journal of Medicine 341, 28–37.

Collins, Francis, Green, Green, Eric D, Guttmacher, Alan, Guyer, Mark S, [NHGRI], 2003. A vision for the future of genomics research, a blueprint for the genomics era. Nature 422, 835–847.

Collins, Francis, 1 April 2010a. Has the revolution arrived? Nature 464, 674–675.

Collins, F., 2010b. The Language of Life. DNA and the Revolution in Personalised Medicine. Harper, New York.

Cook-Deegan, Robert, 1994/1995. The Gene Wars. Science, Politics and the Human Genome. Norton, New York/London.

Davies, Kevin, 2001/2002. Cracking the Genome. Inside the Race to Unlock Human DNA. John Hopkins University Press, Baltimore and London.

Davies, Kevin, 2010. The $1,000 Genome: The Revolution in DNA Sequencing and the New Era of Personalized Medicine. Free Press.

Fischer, Peter, 1985. Licht und Leben. Ein Bericht über Max Delbrück, den Wegbereiter der Molekularbiologie. Universitätsverlag Konstanz, Konstanz.

Foucault, Michel, 1971. L'ordre du discours. Gallimard, Paris.

Fox-Keller, Evelyn, 2000. The Century of the Gene. Harvard University Press, Cambridge, Mass.

Freud, Sigmund, 1917/1947. Eine Schwierigkeit der Psychoanalyse. In: Gesammelte Werke, XII. Imago, London. 3–12.

Fukuyama, Francis, 2002. Our Posthuman Future. Consequences of the Biotechnology Revolution. Farrar, Straus & Giroux.

Galton, Francis, 1875. A theory of heredity. Contemporary Review 27, 80–95.

Gilbert, Walter, 1991. Towards a paradigm shift in biology. Nature 349, p. 99.

Hayden, Erika, 1 April 2010. Human genome at ten: life is complicated. Nature 464, 664–667.

IHGSC, 2004. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945.

International Human Genome Sequencing Consortium [IHGSC], 15 February 2001. Initial sequencing and analysis of the human genome. Nature 405, 860–921.

Kay, Lily, 2000. Who Wrote the Book of Life? A History of the Genetic Code. Stanford University Press, Stanford.

Mendel, Gregor, 1866/1913. Versuche über Pflanzenhybriden. In: Mendel, G. (Ed.), Versuche über Pflanzenhybriden: zwei Abhandlungen Hrsg. E. von Tschermak. Ostwald's Klassiker der exakten Wissenschaften 121. Engelmann, Leipzig, pp. 3–46.

Müller-Wille, Staffan, Rheinberger, Hans-Jörg, 2012. A Cultural History of Heredity. The University of Chicago Press, Chicago and London.

Ridley, Matt, 1999. Genome. The Autobiography of a Species in 23 Chapters. Harper, New York, etc.

Schrödinger, Erwin, 1944/1967. What Is Life? The Physical Aspect of the Living Cell/ Mind and Matter. Cambridge University Press, London.

Shreeve, James, 2004. The Genome War. How Craig Venter Tries to Capture the Code of Life and Save the World. Knopf, New York.

Sulston, John, Ferry, Georgina, 2002/2003. The Common Thread. Science, Politics, Ethics and the Human Genome. Bantam/Corgi.

Venter, J. Craig, et al., 16 February 2001. The sequence of the human genome. Science 291 (5507), 1304–1351.

Venter, J. Craig, 2007. A Life Decoded. My Genome: My Life. Viking/The Penguin Group, New York.

Venter, J. Craig, 1 April 2010. Multiple personal genomes await. Nature 464, 676–677.

Vermeulen, Niki, 2009. Supersizing Science; On Building Large-scale Research Projects in Biology. Maastricht University Press, Maastricht.

Zwart, Hub, 2007. Genomics and self-knowledge. new genetics and society. New Genetics and Society 26 (2), 181–202.

Zwart, Hub, 2008. Understanding the Human Genome Project: a biographical approach. New Genetics and Society 27 (4), 353–376.

Zwart, Hub, 2010. The adoration of a map. Reflections on a genome metaphor. Genomics, Society and Policy 5 (3), 29–43.

Zwart, Hub, Penders, Bart, 2011. Genomics and the Ark. An ecocentric perspective on human history. Perspectives in Biology and Medicine 54 (2), 217–231.

Relevant Websites

http://en.wikipedia.org/wiki/Human_Genome_Project – HGP.
http://www.nih.gov/ – NIH.
http://www.genome.gov/ELSI/ – ELSI.
http://www.genome.gov/ – NHGRI.
http://www.sanger.ac.uk/about/history/hgp/ – Sanger Institute.
http://en.wikipedia.org/wiki/J._Craig_Venter_Institute – J. Craig Venter Institute.

Human Genomics in Asia

Haidan Chen, College of Humanities and Development Studies, China Agricultural University, Beijing, China
Tikki Pang, Lee Kuan Yew School of Public Policy, National University of Singapore, Singapore, Singapore

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the past decade, Asia has been actively engaged in human genomic studies and has made great contributions to the field. There is an increase in the number of genomics institutes, consortiums, and initiatives across the continent to study the association between genetic variation and disease. Despite these laudable efforts, Asia faces tremendous challenges in terms of funding, regulation, collaboration, and ethical, legal, and social issues related to genomics. These need to be addressed in the near future to promote the development of genomic medicine.

Introduction

The past decade has seen the advancement of human genetics and genomics worldwide. Various countries have invested seriously in this field, with the expectation of gaining a better understanding of human health and disease. Biobanks and genetic databases have been built to facilitate interdisciplinary clinical research. Regional and international efforts have also been made to promote the success of genomic medicine. For example, in Europe, the Biobanking and Biomolecular Resources Research Infrastructure has been constructed to secure global competitiveness of research and industry of the European Union, and benefit the health of European citizens. International genomic research collaborations, such as the Human Genome Project (HGP), the International HapMap Project, and the International Cancer Genome Consortium, have produced open-access genetic databases for researchers around the globe (Kaye, 2011; Vaught et al., 2009). The completion of the sequencing of the human genome, the mathematical analysis of DNA variants called single nucleotide polymorphisms (SNPs), and the progress in genome-wide association studies have contributed dramatically to biomedical research. Yet the majority of medical genomics research has mainly focused on people of European descent, the results of which will not be beneficial to the greatest segment of the world's population. Given that the lack of diversity in these studies will bias our understanding of the association between genetic variants and diseases, researchers have recognized the importance of inclusion of populational diversity in pharmacogenomic and genome-wide disease studies (Bustamante et al., 2011).

The geopolitical continent of Asia contains almost two-thirds of the world population, with rich cultural, ethnic, linguistic, and genetic diversity. Asian nations, particularly those still developing, are often depicted as the epicenter of pandemics, e.g., severe acute respiratory syndrome (SARS) and avian influenza outbreaks (Tan et al., 2012). At the same time, these countries also face growing chronic disease burdens, but the causes of diseases in Asian populations remain unknown (Jemal et al., 2011). Thus, genomic research on Asian populations is essential to further enhance our knowledge of the genetic basis of complex chronic and infectious diseases, and to cope with Asian and global public health challenges. In this article, we first give a brief introduction to the contributions of Asian nations to human genomics, placing greater focus on the genomics consortiums and initiatives in the region: Asian Cohort Consortium (ACC), Human Genome Organization (HUGO) Pan-Asia SNP consortium (PASNP 1.0) and HUGO Pan-Asia Population Genomics Initiative (PASNP 2.0). We then discuss the challenges that Asia faces in bringing genomic science and technology to bear on applications of genomic medicine. Finally, we outline several future perspectives for the advancement of genomics in Asia.

Asia's Contribution

The world's largest geopolitical continent, Asia, exhibits heterogeneous geographical, cultural, and socioeconomic characteristics. The investment and progress of genomic science and technology of each country across the continent are uneven. Wealthy countries, such as Singapore and Japan, have made tremendous investments in genomics research and development (R&D) toward efficient diagnosis tools and personalized medicine. The Genome Institute of Singapore (GIS), located at Singapore's biomedical hub, Biopolis, is one of the centers of excellence in genomics in Asia and indeed the world. GIS gained world recognition through its rapid response to SARS and identification of five strains of SARS coronavirus in 2003 to effectively contain a pandemic. The strength of GIS's research lies in three milestone arenas: "infectious diseases, cross-national science diplomacy, and regenerative medicine" (Fischer, 2013). Japan has established several research institutions, e.g., RIKEN Center for Genomic Medicine, and research infrastructures, e.g., Biobank Japan, which are steering the nation ahead on the path toward genomic applications in medicine and health (Yoshizawa et al., 2014).

Emerging economies, such as China and India, are the rising powers in the global biomedicine field and are now challenging the developed economies (Salter and Faulkner, 2011). China was the sole developing country to participate in the HGP, contributing toward sequencing of 1% of the human genome. The Beijing Genomics Institute (BGI) has become a global sequencing powerhouse that sequenced the first Asian human genome, and a number of plant and animal genomes, such as rice, silkworm, chicken, and panda (Normile, 2012). India is endeavoring to foster economic development and address local health needs through investment in genomics-based innovation. The Indian Genome Variation Consortium, a government-funded collaborative network of

several local institutions, is one of the examples that reflect these efforts (Hardy et al., 2012). Still other Asian countries, such as Indonesia and Philippines, have emerged in recent years as active actors in the field of genomics, partly due to their ethnic and linguistic diversity (Liu, 2008). Such efforts by nations across the continent are crucial for Asia's genomic research collaboration networks, such as ACC, APSNP 1.0 and 2.0. We introduce these networks in detail below.

Asia Cohort Consortium

First proposed in November 2004, in Seoul, the ACC is a collaborative cancer cohort research project, involving more than one million healthy people across Asia who will be followed over time until various disease endpoints are reached. The ACC seeks to understand the relationship between genetics, environmental exposures, and the etiology of disease, and to discover early detection biomarkers. The ACC has two missions: "(i) to serve as a platform for cross-cohort collaborative projects and combined analysis and (ii) to act as an incubator for new cohorts" (Song et al., 2012). The ACC has approximately 50 active members from Bangladesh, China, India, Iran, Japan, South Korea, Malaysia, Singapore, Taiwan, and the United States, among others. It is cochaired by John Potter of the Fred Hutchinson Cancer Research Center, Seattle, Washington, USA, and Daehee Kang of Seoul National University College of Medicine, South Korea. Its coordinating center is located at the Fred Hutchinson Cancer Research Center to provide support for scientific collaboration, coordination and communication, data operations, and statistical consultation (Rolland et al., 2011).

As most studies on the associations between body mass index (BMI) and the risk of death have been conducted on North American and European populations, and less is known about these associations in Asian populations, the focus of ACC's projects is on BMI in Asian populations. To date, the cross-cohort collaborative projects of ACC have yielded about seven articles, showing the association between BMI and risk of death; BMI and diabetes; BMI, tobacco smoking, alcohol consumption, and risk of cancer of the small intestine; BMI and risk of death from pancreatic cancer; meat consumption and cause-specific mortality; and BMI and cardiovascular disease mortality in Asian populations (https://www.asiacohort.org/index.html (accessed 03.07.14.)).

PASNP 1.0 and 2.0

The SNP consortium was established in early 1999 through a collaborative effort among major pharmaceutical companies, the Wellcome Trust, and five leading academic centers to provide a resource pool for clinical genomic research and for the discovery of novel diagnosis and personalized therapies (Holden, 2002). This consortium largely focused on Caucasian populations, while genomic variation among Asian peoples remained unexplored. In 2007, HUGO Pan-Asia SNP Consortium (PASNP 1.0) was set up with 93 researchers from 40 institutions in 11 Asian countries to map human genetic diversity in Asia. Coordinated from the GIS, the PASNP took the first steps toward collaboration among Asian scientists.

Based on strong bioinformatics teams from China, India, Japan, Singapore, etc., diverse ethnicities from Indonesia, Malaysia, Philippines, and Taiwan, and the scientists' common interests, the consortium used samples from more than 1900 individuals representing 73 populations to conduct migration studies (Normile, 2004, 2009). The PASNP consortium's first report, published in *Science* in December 2009, provided a physical map of human variation in Southeast Asian, East Asian (EA), and Central South Asian populations, and showed strong evidence for the southern migration route to Asia, though it did not completely rule out a two-wave model of migration (Abdulla et al., 2009).

In order to advance the work of PASNP, the HUGO Pan-Asia Population Genomics Initiative was launched in 2011 as a version 2.0 of PASNP 1.0. It established a larger network with hundreds of researchers, and constructed a gene pool with diverse people, data, and cultures, including those from mainland Central Asia and the Pacific Islands. The goal of PASNP 2.0 is to further explore Asian migration patterns, Asian genetic diversity, and local adaptation, and eventually to translate genomics knowledge to the practice of genomic medicine. The data collected by PASNP 2.0 will be open to the worldwide scientific community for further genomic studies (http://papgi.org/index.php/About_Us (accessed 03.07.14.)).

Despite the achievements that Asian countries and consortiums have made, they still face formidable challenges in terms of funding, regulation, collaboration, and the ethical, legal, and social issues (ELSI) of genomics in Asia. As we show in the next section, the development of human genomics in Asia faces not only similar issues as do other national and transnational endeavors, such as funding, standardization of data and samples, harmonization of ELSI practices and regulations, and gaining public trust, but also some Asia-specific issues and local concerns in the practice of Asian science and collaboration.

Current Challenges

Funding

Funding for genomic science is becoming more competitive and more difficult to obtain, both from the government and from industry sources. For this reason, ensuring sustainable funding is one of the major challenges for genomic studies (Vaught et al., 2009). Even in Singapore, where the government had attracted many top scientists from the West with substantial funds, after two 5-year periods of multibillion dollar investments, with little or no warning the core budgets for bioscience institutes under the Agency for Science, Technology and Research (A*STAR) was cut in the third-year period beginning in 2011. A*STAR bioscientists, including those from GIS, had to compete for grants from the new Industry Alignment Fund, and collaborate with clinicians on industrial and clinical applications (Fischer, 2013; Normile, 2011).

Elsewhere, BGI took an entrepreneurial approach to fuel its growth of sequencing and bioinformatics capabilities. It borrowed 1.5 billion U.S. dollars from a government-owned bank and must begin to pay back the principal within 10 years. Since sequences do not make money, BGI established subsidiary

companies to expand the application of genomic science and technology in order to pay off the debt and generate more income (Normile, 2012). For example, BGI Diagnostics offers health services, such as noninvasive prenatal test and tests for various genetic disorders; BGI Ark Biotechnology Co. LTD has set up a transgenic platform, a cloning platform, an experimental farm, and an animal model department (see BGI's website). In addition to the challenges of funding, some Asian countries also find regulation a problem.

Regulation

When appointed as the first director of the HGP, James Watson resolved that the U.S. National Institutes of Health should allocate part of its funds to study the ELSI of genomics in the face of unprecedented challenges posed by genomics and its future application. Research into the ELSI of genomics was initiated first in the USA, thence spreading to Canada and Europe (Zwart and Nelis, 2009). Awareness of the issues in the East happened much later, but such research was supposed to study similar issues as those in the West for the purpose of some outcomes leading to better regulation of genomics. More recently, Yoshizawa et al. identified heterogeneous regulatory frameworks and ELSI practice among EA countries. They found that EA countries' regulation of genomics had a relatively inconsistent and mixed character. Consider the aspect of privacy as an example: Japan, Singapore, South Korea, and Taiwan have regulations and oversights to protect personal information, while China and Indonesia exert little or no control over privacy issues. A similar situation is seen with ethics review. Even though there are ethics review committees in nearly all of these countries, discrepancies in ethics review capacities and the implementation of oversight between different countries are noticeable (Yoshizawa et al., 2014).

What is perhaps of greater concern is that, in overall terms, there are limited or even absent regulations in many developing countries. Some of these countries may have no or limited capacity to regulate innovative genomic medicine. This could be a major barrier to the application of emerging genomic medicine products to improve global health. Therefore, it is significant to include the developing countries in the International Conference on Harmonization, as well as other international consortia, so that developed countries can help improve the regulatory capacity of developing countries (Hardy et al., 2008). Yet international health research collaboration between developed and developing countries in the field of human genetics is often a sensitive issue.

Biopiracy

There are concerns about possible exploitation of the populations in the developing countries in the commercialization of human tissues, and unfairness in benefit sharing and ownership between developed and developing countries (Schulz-Baldes et al., 2007). Biodiversity-rich countries like China, India, and Indonesia are concerned about the exploitation of genetic resources by more developed countries. They have been uneasy about equitable benefit sharing, ownership, and intellectual property rights in international research collaboration. As a consequence, China promulgated the Interim Measures for the Administration of Human Genetic Resources, issued by its Ministry of Science and Technology and the Ministry of Health in 1998. This directive stipulates that only through collaboration with Chinese parties can foreign researchers get access to Chinese genetic resources (Chen, 2013).

Fearing that Indonesia may not benefit from vaccines developed by international scientists and multinational pharmaceutical companies, the Indonesian government stopped sending H5N1 virus specimens to the World Health Organization in 2006. This stance was supported by Malaysia, Thailand, and other developing countries to communicate their desire for mutual trust, transparency, and equity between the developed and developing nations in the virus sharing mechanism. Moreover, Indonesia revised its health law and enacted the Ministry of Health Regulation on Material Transfer Agreement in 2009 in order to safeguard national sovereignty over its biological materials (Sedyaningsih et al., 2008). Although India has ethical guidelines and regulations in place to prevent biopiracy, there is a lack of adherence to regulations and an absence of strict implementation of measures to monitor the misuse of genetic samples (Kumar, 2009). China faces similar problems. To protect the use of Chinese genetic resources in research collaboration, the Office of Legislative Affairs of the State Council, People's Republic of China, published the exposure draft of Measures for the Administration of Human Genetic Resources online in 2012 to solicit advice from the public. As of this writing, the official measures have still not been issued (http://www.gov.cn/gzdt/2012-10/31/content_2254379.htm (accessed 09.07.14.)). The biopiracy issue reflects complexities in cross-border collaboration as well as the importance of mutual trust and equity among scientists in using human genetics.

Collaboration

Although there are potential dangers of biopiracy and unfair benefit sharing in collaboration between developing countries and more developed countries (north–south collaborations), collaboration should nevertheless be promoted as a strategy to help build developing nations' research and translational capacities (Schulz-Baldes et al., 2007). Increasingly, there is also a trend toward collaboration among developing countries (south–south collaborations) in building human resource capacity (Ivers et al., 2010). In the case of the PASNP consortium, Edison Liu, one of the key organizers of the consortium, found it challenging to coordinate science collaboration among Asian scientists. As he pointed out:

> Scientists in Asia have a tendency to look past each other and focus on collaborations with the United States or Europe, partly because these collaborations get them more credit from their school administrations. Also, in Asia, most countries see each other as competitors. Just getting people together is an accomplishment.
>
> Liu, 2008

The science culture in Asia is different from that among scientists in the West and tends to be more hierarchical and bureaucratic. Another challenge rests with the disparity in research capacities and research infrastructures between Asian

countries (Liu, 2008). Besides these challenges, Asian genomic research collaborations face similar challenges as those of the international science communities more generally. As large-scale genomic research collaborations require data comparisons and validations across different populations, the lack of harmonized ELSI and lack of a regulatory infrastructure for genomic research and application are major hurdles. It is also difficult but crucial to ensure the quality of samples and the interoperability of data for collaboration (Song et al., 2012; Vaught et al., 2009). Finally, the translation of genomics knowledge to genomic medicine and innovative medicine products necessitates the interaction and collaboration of all stakeholders. They are not only scientists, clinicians, policy-makers, and industry people, but also patients, consumers, and healthy individuals.

Publics

Engaging the public in biobank and genomic studies is of great importance because researchers need access to patients' and to the public's biological samples. There are a good number of qualitative and quantitative studies about public perception of biobanks and genetic databases in Western countries to explore the public's reasons for participating and not participating in biobank studies; their views of issues such as informed consent, benefit sharing, commercialization, and internationalization; and their concerns over privacy, discrimination, ownership, and the return of results. However, except for a few limited studies in Asian countries, among them China, Japan, and Singapore, little is known about the Asian public's perception of biobanks (Gottweis et al., 2011).

Investigation of public perceptions of biobanks reveals several factors that affect the public's willingness to donate their samples for research and to participate in the studies. These factors include the public's understanding of biobanks, its trust in research institutions and the scientists running those institutions, and its consideration of privacy, discrimination, and commercialization (Gottweis et al., 2011). In Asia, there is tendency for the public to distrust science. For example, Singapore's growing emphasis on commercialization of science discourages the public's participation in research. In Japan, public trust in expertise was eroded following the earthquake, tsunami, and nuclear accident in 2011 (Arimoto and Sato, 2012). There are cases reporting public distrust of experts, and public concerns over the misuse of genetic information and genetic discrimination after genetic testing in China and South Korea. In Taiwan, tensions and distrust arose between the public and authorities of the Taiwan Biobank due to a lack of open science communication (Yoshizawa et al., 2014). Well-known cases of scientific fraud in South Korea (Bonetta, 2006), Japan (Tsurimoto et al., 2009), and China (Lin, 2013) have also contributed to the lack of trust between the public and scientists. All of these signal a call for an interface of science and society to actively interact with the public.

Future Perspectives

As a result of rich and diverse genetic resources, together with political will of various countries, and the great efforts that scientists have made, Asia is rising in the field of human genomics in the postgenomic era. Building regional collaborative genomics networks, such as ACC and PASNP 1.0 and 2.0, is the first and vital step toward better science, medicine, and health in Asia. The research conducted by these networks is competitive in the global arena and bridges the gaps between Asia and Western countries. However, compared with research collaboration in Europe and North America, Asian genomics networks face more challenges than their counterparts in the West. The emerging economies and the developing countries in Asia have not established consistent and robust regulatory frameworks to govern the conduct of genomic research. Furthermore, their ELSI practice and regulation is often influenced by local sociopolitical and cultural concerns and tends to be heterogeneous. This has been a disincentive to regional collaboration on genomic science and technology, which requires the harmonization of ELSI practice and regulation across Asia. Therefore, in the future, it will be helpful to place more emphasis on research into ELSI of genomics and develop regulatory infrastructures, which could become a tool to secure the public's trust in science. It is also important to encourage both north–south collaborations and south–south collaborations. Experience gained from PASNP could contribute to further collaboration on the basis of previous networks and research collaboration.

Editor's Note

Relevant aspects of genomics research in other countries on the Asian continent are covered in other articles of this Encyclopedia (see cross references). See also e.g. (Najmabadi, et al., 2003; Kumar, 2012; Prainsack, 2007).

See also: Anthropology, Genomics, and Human Variation: National Roots; Biobanking: Ethical Issues; Bioethics in the Post-genomic Era; Bioethics: Genetics and Genomics; Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute; Gene–Environment Interactions in Health and Well-Being; Genetics and Indigenous Communities: Ethical Issues; Genetics and Society; Genetics: Legal Aspects; Genetics: The New Genetics; Genomics, Ethical Issues in; Human Genome Diversity Project: History; Human Genome Project: History and Assessment.

Bibliography

Abdulla, M.A., et al., 2009. Mapping human genetic diversity in Asia. Science 326 (5959), 1541–1545.

Arimoto, T., Sato, Y., 2012. Rebuilding public trust in science for policy-making. Science 337 (6099), 1176–1177.

Bonetta, L., 2006. The aftermath of scientific fraud. Cell 12 (5), 873–875.

Bustamante, C.D., Burchard, E.G., De La Vega, F.M., 2011. Genomics for the world. Nature 475 (7355), 163–165.

Chen, H., 2013. Governing international biobank collaboration: a case study of China Kadoorie biobank. Science Technology and Society 18 (3), 321–338.

Fischer, M.M.J., 2013. Biopolis: Asian science in the global circuitry. Science Technology and Society 18 (3), 379–404.

Gottweis, H., Chen, H., Starkbaum, J., 2011. Biobanks and the phantom public. Human Genetics 130 (3), 433–440.

Gottweis, H., Gaskell, G., Starkbaum, J., 2011. Connecting the public with biobank research: reciprocity matters. Nature Reviews Genetics 12 (11), 738–739.
Hardy, B.-J., et al., 2008. Science and society – the next steps for genomic medicine: challenges and opportunities for the developing world. Nature Reviews Genetics 9, S23–S27.
Hardy, B.-J., et al., 2012. Genomic applications in medicine and health in India. In: Kumar, D. (Ed.), Genomics and Health in the Developing World. Oxford University Press, Oxford, pp. 916–921.
Holden, A.L., 2002. The SNP Consortium: summary of a private consortium effort to develop an applied map of the human genome. Biotechniques S22–S26.
Ivers, L.C., et al., 2010. South-south collaboration in scale-up of HIV care: building human capacity for care. Aids 24, S73–S78.
Jemal, A., et al., 2011. Global cancer statistics. CA: A Cancer Journal for Clinicians 61 (2), 69–90.
Kaye, J., 2011. From single biobanks to international networks: developing e-governance. Human Genetics 130 (3), 377–382.
Kumar, N.K., 2009. India's preparedness in tackling biopiracy and biobanking: still miles to go. In: Sleeboom-Faulkner, M. (Ed.), Human Genetic Biobanks in Asia: Politics of Trust and Scientific Advancement. Routledge, London and New York, pp. 142–150.
Kumar, D. (Ed.), 2012. Genomics and Health in the Developing World. Oxford University Press, Oxford.
Lin, S., 2013. Why serious academic fraud occurs in China. Learned Publishing 26 (1), 24–27.
Liu, E., 2008. Genomics takes hold in Asia. Interview by David Cyranoski. Nature 456 (7218), 12.
Normile, D., 2004. Genetic diversity – consortium hopes to map human history in Asia. Science 306 (5702), 1667.
Normile, D., 2009. SNP study supports southern migration route to Asia. Science 326 (5959), 1470.
Normile, D., 2011. Winds of change leave bioscientists scrambling. Science 332 (6026), 165.
Normile, D., 2012. China's sequencing powerhouse comes of age. Science 335 (6068), 516–519.
Najmabadi, H., et al., 2003. The Iranian Human Mutation Gene Bank: a data and sample resource for worldwide collaborative genetics research. Human Mutation 21 (2), 146–150.
Prainsack, B., 2007. Natural forces: the regulation and discourse of genomics and advanced medical technologies in Israel. In: Glasner, P., Atkinson, P., Greenslade, H. (Eds.), New Genetics, New Social Formations. Routledge, London, pp. 231–252.
Rolland, B., Smith, B.R., Potter, J.D., 2011. Coordinating centers in cancer epidemiology research: the Asia cohort consortium coordinating center. Cancer Epidemiology Biomarkers & Prevention 20 (10), 2115–2119.
Salter, B., Faulkner, A., 2011. State strategies of governance in biomedical innovation: aligning conceptual approaches for understanding 'rising powers' in the global context. Globalization and Health 7. http://dx.doi.org/10.1186/1744-8603-7-3.
Schulz-Baldes, A., Vayena, E., Biller-Andorno, N., 2007. Sharing benefits in international health research. Research-capacity building as an example of an indirect collective benefit. EMBO Report 8 (1), 8–13.
Sedyaningsih, E.R., et al., 2008. Towards mutual trust, transparency and equity in virus sharing mechanism: the avian influenza case of Indonesia. Annals Academy of Medicine Singapore 37 (6), 482–488.
Song, M., et al., 2012. Asia cohort consortium: challenges for collaborative research. Journal of Epidemiology 22 (4), 287–290.
Tan, Y., Lee, K., Pang, T., 2012. Global health governance and the rise of Asia. Global Policy 3 (3), 324–335.
Tsurimoto, T., et al., 2009. Report from the working group of the molecular biology society of Japan for the investigation of fraud in research papers. Genes to Cells 14 (8), 903–908.
Vaught, J., Kelly, A., Hewitt, R., 2009. A review of international biobanks and networks: success factors and key benchmarks. Biopreservation and Biobanking 7 (3), 143–150.
Yoshizawa, G., et al., 2014. ELSI practices in genomic research in East Asia: implications for research collaboration and public participation. Genome Medicine 6. http://dx.doi.org/10.1186/gm556.
Zwart, H., Nelis, A., 2009. What is ELSA genomics? science & society series on convergence research. EMBO Reports 10 (6), 540–544.

Relevant Websites

https://www.asiacohort.org/ – Asia Cohort Consortium.
http://www.genomics.cn/en/index – Beijing Genome Institute.
http://biobankjp.org/index.html – Biobank Japan.
http://www.gis.a-star.edu.sg/internet/site/ – Genome Institute of Singapore.
http://www.igvdb.res.in – Indian Genome Variation Consortium.
http://papgi.org/index.php/Main_Page – Pan-Asian Population Genomics Initiative.
http://www4a.biotec.or.th/PASNP – Pan-Asian SNP Consortium.

Human Language, Evolution of

Josef P Rauschecker, Georgetown University, Washington, DC, USA; and Institute for Advanced Studies, Technische Universität München, Garching, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Understanding the evolution of language, like that of any other cognitive brain system, benefits from a direct comparison of humans with other living species. Generally speaking, the brains of humans and monkeys show a great deal of anatomical similarity. The auditory cortical system in both species, for instance, is organized into a ventral and a dorsal pathway that originate, respectively, in areas anterior and posterior to a primary field and project to distinct areas of prefrontal cortex. In both species, the ventral pathway performs hierarchical decoding of spectrally complex sounds ('auditory object recognition'), which includes decoding of speech sounds ('speech perception') and their ultimate linking to meaning ('semantics') in humans. The dorsal auditory pathway connects posterior auditory regions in both species with parietal, premotor, and prefrontal cortex and subserves a general function in audiomotor integration and control. This includes processing of auditory space and motion as well as online control of speech. The dorsal stream also concatenates sounds into temporal sequences and has thus developed an important role in sentence processing. Besides these similarities in brain organization, future comparative studies on the evolution of language will have to test whether humans and monkeys show principal differences in brain connectivity (e.g., new pathways appearing as a result of specific mutations) or whether species differences are merely of a quantitative nature. The evidence so far does not seem strong enough to rule out the latter possibility.

Introduction

Great thinkers like Francis Crick have lamented on occasions that "the trouble with language is that there are no animal models." Yet the evolutionary origins of language must be latent in animals living today. Language as complex as in humans may not exist in other animals, but its basic components for decoding sounds within the auditory system, for production of sounds within the motor system, and the sensorimotor coupling between the two should be traceable as precursors of human language in animal models. In addition to these basic elements, working memory and the processing and storage of sequences should be crucial for language systems. The ability to learn the perception and production of communication sounds ('vocal learning') has been considered a critical facet of human language as well and has triggered an important comparison with birdsong (Doupe and Kuhl, 1999; Fitch, 2000; Petkov and Jarvis, 2012). Nonhuman primates, on the other hand, seem to have only a limited capacity for vocal learning, but the neuroanatomy of their cerebral cortex is so intriguingly similar to that of humans that it invites direct comparison as well. The challenge of using the monkey as an animal model for human language consists, therefore, mostly in looking for elements that are underdeveloped compared to humans. The core of this article is built on the assumption that, despite its shortcomings, the brain of nonhuman primates contains sufficiently many similarities with the human brain, so it makes the hypothesis worth considering that something as intricate and unique as the human language system has indeed evolved from simpler versions. Their elements, after some improvement and in combination with each other, might have enabled the evolution of language. Other animal models such as songbirds, considerations of fossile data, and research on nonauditory forms of communication remain of great importance for a complete understanding of language evolution, but are not part of this account.

Species-Specific Communication Sounds in Nonhuman Primates

Spoken language starts with the processing of complex auditory signals. Physiological recordings in nonhuman primates suggest that neurons already at the secondary stage of processing along the auditory cortical pathway (the lateral belt areas; Pandya and Sanides, 1972; Morel et al., 1993) can show a preference for species-specific communication calls (Rauschecker et al., 1995). These findings in rhesus monkeys extend earlier accounts from squirrel monkeys that have also demonstrated vivid responses to species-specific calls at the behavioral and neural level (Winter and Funkenstein, 1973; Wollberg and Newman, 1972). Neuroethologically, it is important to demonstrate such responses to sounds that are specific to communication within the same species; only in this way a system can be established that works for conspecifics and does not respond to signals from other species, as similar as they might be acoustically. It would not make sense, for instance, to expect selective responses to human speech in the brains of monkeys and vice versa. On the other hand, sounds produced in certain behavioral situations, e.g., at the sight of a threat, as 'eagle' or 'snake' alarm calls, might carry distinct semantic meaning triggering specific forms of behavior in members of the same species (Cheney and Seyfarth, 1990).

The response tuning of single neurons (and neuronal assemblies) in nonprimary auditory cortex to species-specific communication sounds is generated by convergence of input from lower order neurons that themselves respond to simple sounds like tones, frequency-modulated sweeps, or band-passed noise bursts. Neurons are sensitive to highly specific combinations of such inputs, and combining signals in a nonlinear conjunctive AND-logic leads to the existence of neurons that respond specifically to certain types of calls. There is every reason to believe that the human auditory cortex contains similar neurons with 'combination sensitivity'

(Suga, 1988; Margoliash and Fortune, 1992), which respond specifically to certain acoustic-phonetic features, syllables, or words. As in the monkey, such a system should be organized into a similar hierarchy from rather simple to more complex neurons, whose incidence increases from primary auditory cortex to more anterior and ventral regions of the superior temporal lobe (Rauschecker, 1998; Rauschecker and Tian, 2000; Romanski and Averbeck, 2009; Kikuchi et al., 2010, 2014; Fukushima et al., 2014) and into ventrolateral prefrontal cortex (Romanski et al., 2005; Cohen et al., 2009).

Human Auditory and Language Cortex

Early studies of human auditory cortex with functional magnetic resonance imaging (MRI) have shown that primary auditory cortex responds best to tones, while at the next stage, the equivalent of the lateral belt in the monkey, band-passed noise bursts are more effective stimuli (Wessinger et al., 2001). Further along the anteroventral pathway, cortical regions are selectively activated by words and intelligible speech sounds (Binder et al., 2000; Scott et al., 2000). This hierarchical organization of the auditory ventral stream with regard to speech sound processing was recently corroborated with more refined imaging techniques (Chevillet et al., 2011). Furthermore, a meta-analysis of more than 100 neuroimaging studies of human speech processing has demonstrated that cortical regions lateral to primary core areas are sensitive to phonemes. Farther afield in the anterior superior temporal gyrus (STG), words are processed; finally, in the most anterior locations of the superior temporal sulcus, short phrases lead to selective activation (DeWitt and Rauschecker, 2012, 2013).

In an exciting series of recent studies (Steinschneider, 2011; Chang et al., 2010; Mesgarani et al., 2014), direct recordings from human auditory cortex have become more feasible. This approach, begun by neurosurgeons in epileptic patients (Penfield and Roberts, 1959; Ojemann et al., 1988), is now utilized with semichronically implanted electrode arrays in electrocorticogram recordings from the cortical surface. Results show neurons in nonprimary human auditory cortex very similar to those described above in nonhuman primates. Most specifically, neurons have been shown to respond to phonetic elements and formant transitions, as one would expect.

Invariance and Categorization of Speech Sounds

Invariant representation of sounds is another important step toward establishing a usable system for auditory communication, such as speech. A particular speech sound has to transmit the same meaning regardless of pitch (fundamental frequency) or loudness and irrespective of idiosyncrasies in the speaker's voice or accent. There is evidence that some invariances are formed along the anteroventral stream (DeWitt and Rauschecker, 2012). Evidence for this comes also from studying the representation of other 'auditory objects' (Leaver and Rauschecker, 2010; Zatorre et al., 2004), i.e., sounds that carry meaning in a wider sense and can form categories of sounds. Speech sounds are only one such category; musical instrument sounds or animal calls would be other examples of categories. It turns out from these studies that the 'speech sound' category is indeed processed in clusters within the same cortical region (anterolateral region of STG in the left hemisphere), whereas musical instrument sounds are processed in a more anterior region of STG in the right hemisphere (Leaver and Rauschecker, 2010). All of this suggests that neurons whose selectivity is generated by combining certain acoustic elements in specific ways cluster together forming invariances and categories. Presumably, similar clustering mechanisms lead to the development of functionally specialized cortical regions that process certain configurations, like faces in the visual domain (Kanwisher et al., 1997) or voices in audition (Belin et al., 2000). Both of these regions exist in humans and nonhuman primates (Petkov et al., 2006), which is a strong argument in favor of a direct evolutionary lineage.

Other reports have found that premotor regions may be involved in the formation of speech sound categories as well (Chevillet et al., 2013), which suggests that the way a sound is produced will also influence the way it is processed and stored. Thus, it appears possible that speech invariances are formed in different ways: First, on the basis of spectrotemporal information, which is pooled along the frequency domain in the sense of an OR-logic within the auditory ventral stream, and second, in the domain of articulatory motor gestures, which are formed originally for speech production, but are invoked during the processing of speech as well. The latter route is reminiscent of the 'motor theory of speech perception' (Liberman and Mattingly, 1985), in that it requires the matching of motor information with spectrotemporal information on the acoustic–phonetic side, but there are two important differences here: neither is the process special for speech nor are the motor gestures alone sufficient for speech perception.

As noted above, the same dual way of encoding information holds for other doable sounds that can be classified into discrete categories (Warren et al., 2005). Such auditory objects are represented in anterior regions of the STG (Leaver and Rauschecker, 2010), but premotor cortex participates in their encoding as long as they can be produced (Scott, 2005). Monkeys are naturally handicapped by their less sophisticated vocal apparatus, which limits their vocal repertoire and their capacity for mimicking sounds (Fitch, 2000). The involvement of the dorsal pathway (including premotor regions) in the processing and categorization of self-produced sounds will therefore have to be tested by other means, e.g., by the monkeys' ability to synchronize with other individuals' cage rattling ('drumming') (Remedios et al., 2009) or by teaching them to produce sound sequences by pressing buttons or leavers rather than their vocal apparatus (Artchakov et al., 2012).

Dorsal Stream for Sensorimotor Integration and Control: Sequences, Sentences, and Beyond

The involvement of the dorsal auditory pathway, including premotor and inferior parietal regions, becomes especially evident, when the encoding and representation of sound sequences is investigated (Leaver et al., 2009). During the learning of such sequences in humans, the basal ganglia are

actively engaged, whereas after the sequences become highly familiar, the same sequences activate increasingly more frontal cortical areas. Analogous findings have been reported in rhesus monkeys (Pasupathy and Miller, 2005). It appears, therefore, that the basal ganglia are responsible for the concatenation or 'chunking' of sequential auditory information. The chunks represent information about conditional probabilities for one sound being followed by another (Rauschecker, 2014). Once the chunks have been formed by basal ganglia action, they are once again stored in prefrontal regions. A similar chunking process occurs with cued sequences of learned finger movements (Koechlin and Jubault, 2006). This process involves prefrontal cortex near Broca's area and has therefore been compared with models of language (Hagoort, 2005), redefining Broca's area in terms of chunking ('unification') of semantic, syntactic, and phonological information. Other accounts have compared Broca's area with a readiness platform, where sequences are put together before they are launched (Wise, 2003). The existence of a homologue of Broca's area in nonhuman primates has been postulated on anatomical grounds (Petrides and Pandya, 2009) and involves similarities in cytoarchitectonics and connectivity (see below). Whether the homology can be backed up in terms of functional and computational mechanisms remains to be seen.

Thus, the role of the dorsal stream, as defined above, can be conceptualized into one of sensorimotor integration and control and applies to all kinds of sequential stimuli, even beyond the auditory domain. Specifically for speech, the dorsal processing stream plays a role in speech production as well as categorization of phonemes during online processing of speech (Rauschecker and Scott, 2009; Rauschecker, 2011; **Figure 1**). The former role conforms to the classical idea of an 'efference copy' or feed-forward internal model (Wolpert et al., 1995) and allows for fast and efficient online control of speech production. By contrast, the latter function can be formalized as an inverse model during real-time speech processing, creating the affordances of the speech signal in a Gibsonian sense (Gibson, 1966; Rauschecker, 2005). Both functions require a direct or indirect connection between sensory and motor cortical structures of the brain, whereby subcortical structures (e.g., the basal ganglia) provide an additional link setting up transitional probabilities during associative learning of sound sequences.

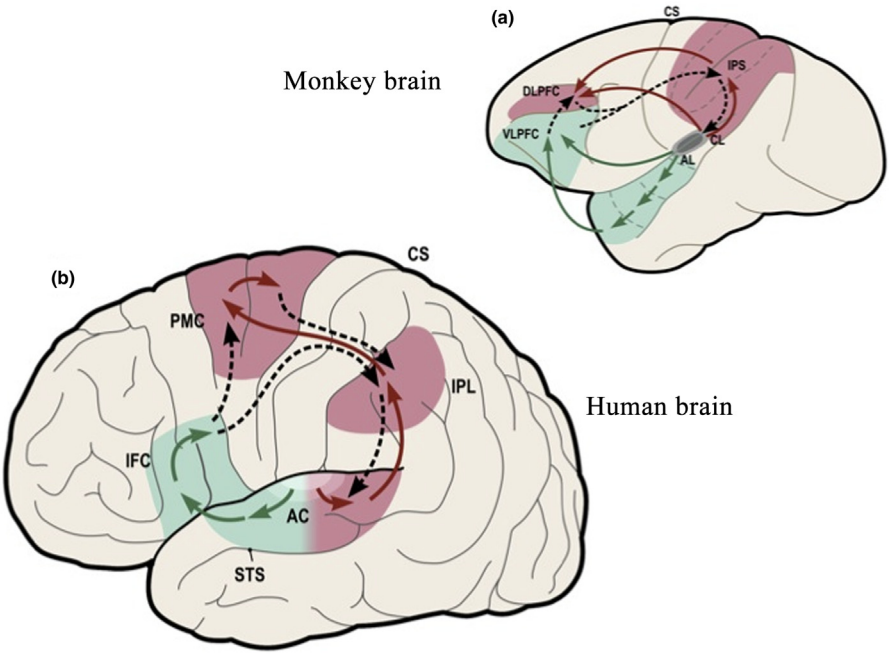

Figure 1 Ventral and dorsal streams for the processing of complex sounds in the primate brain: (a) in the rhesus monkey (Modified from Rauschecker, J.P., Tian, B., 2000. Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proceedings of the National Academy of Sciences of the United States of America 97, 11800–11806.); (b) in the human (Simplified from Rauschecker, J.P., Scott, S.K., 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature Neuroscience 12, 718–724.). The ventral stream (in green) plays a general role in auditory object recognition, including perception of vocalizations and speech. The dorsal stream (in red) pivots around inferior/posterior parietal cortex, where a quick sketch of sensory event information is compared with an efference copy of motor plans (dashed lines). Thus, the dorsal stream plays a general role in sensorimotor integration and control. In clockwise fashion, starting out from auditory cortex, the processing loop performs as a forward model: Object information, such as vocalizations and speech, is decoded in the anteroventral stream all the way to category-invariant inferior frontal cortex (IFC, or VLPFC in monkeys) and transformed into articulatory representations (DLPFC or ventral PMC). Frontal activations are transmitted to the IPL and posterior superior temporal (pST) regions, where they are compared with auditory and other sensory information. AC: auditory cortex; AL: anterolateral area; CL: caudolateral area; STS: superior temporal sulcus; IFC: inferior frontal cortex; DLPFC, VLPFC: dorsolateral and ventrolateral prefrontal cortex; PMC: premotor cortex; IPL: inferior parietal lobule; IPS: inferior parietal sulcus; CS: central sulcus. Composite figure adapted from Rauschecker, J.P., 2011. An expanded role for the dorsal auditory pathway in sensorimotor control and integration. Hearing Research 271, 16–25.

In summary, a system like human language (or its predecessors in nonhuman primates or in a common ancestor of today's humans and nonhuman primates) requires the interplay of ventral- and dorsal-stream systems. Comparing human and monkey brain connectivity, there may be quantitative differences in the strengths of these connections within the dorsal stream, but not a difference in principle (Frey et al., 2008, 2014). Similarly, in the ventral stream, the fine-grain organization of cortical areas and the fine-tuning of its neuronal elements may be richer in humans than in monkeys, providing humans with a perceptual network for the detection of more subtle differences in the acoustic signal. The decisive distinction between humans and monkeys may, therefore, lie in a third component: the prefrontal network, where ventral and dorsal information converges and gets integrated. With its own hierarchical organization, the prefrontal network provides the substrate for recursive processing of nested sequences, as they are typical for human grammatical language structures (Friederici, 2004, 2012). Again, however, this emergent new ability of humans may be based on a quantitative rather than principal difference in human and monkey brain organization, which ties in the existing strengths of both ventral and dorsal processing streams with frontoparietal networks underlying working memory (Bornkessel-Schlesewsky and Schlesewsky, 2013; Bornkessel-Schlesewsky et al., in preparation).

Future Studies

To test the evolutionary similarity of human and monkey ventral and dorsal streams in even greater detail, two things have to be done:

1. Connectivity studies in both species will have to investigate in great detail which areas are indeed connected. This will establish a greater amount of homology than other approaches, especially when the same techniques of structural and functional imaging are utilized. While anatomical tracer studies in monkeys will remain the gold standard (Romanski et al., 1999; Petrides and Pandya, 2009; Hackett, 2011), noninvasive fiber tractography using MRI-based technology will gain increasing importance as its resolution improves, because the exact same approach can be used in both species. Early attempts using diffusion tensor imaging have had insufficient power to resolve crossing fibers within a single voxel or disentangle fibers with crossing trajectories (Anwander et al., 2007; Catani et al., 2005; Croxson et al., 2005; Rilling et al., 2008). Such studies have, therefore, remained inconclusive with regard to monkey–human homologies in language evolution. High-angular-resolution techniques, such as diffusion spectrum imaging (DSI), have been utilized successfully in humans (e.g., Frey et al., 2008) as well as in monkeys (Schmahmann et al., 2007; Wedeen et al., 2008). Cross-validation studies of autoradiographic tract tracing and DSI in monkeys have shown a remarkable concordance of results between tracer studies and DSI (Schmahmann et al., 2007), but further improvements in resolution and reductions in scan time are certainly possible, before such studies can become routine. Functional studies based on blood oxygenation level-dependent responses are feasible in both species as well (Petkov et al., 2006) and can elucidate connectivity to a certain extent. However, microstimulation techniques as another approach to analyze functional connectivity (Petkov et al., in preparation) are limited to animal studies.
2. Behavioral monkey studies have to be designed that test the above concepts and go beyond traditional models of 'what' and 'where' processing. Both of these modes can certainly still be considered as characteristic for the two streams, but as generalized models are developed (Rauschecker, 2011), more appropriate monkey studies will have to follow. These studies will have to focus on the computational transformations that occur between the various processing stages or cortical modules rather than merely the connectivity describing different anatomical pathways. This is how processing streams are really characterized.

Conclusions

In this article, I have argued that the anatomical similarities between the human and nonhuman primate brains are sufficiently close that a quantitative rather than qualitative difference might have resulted in the ability of humans to develop a species-specific communication system into 'real' language. The similarities carry over from audition to speech and form the substrate for language, including those aspects of sentence processing and beyond that are often claimed to require uniquely human computational mechanisms. The ability of monkeys to communicate more effectively might be hampered not only by their less flexible vocal apparatus but by a number of other factors: on the receptive side, denser and more fine-tuned organization of ventral-stream structures for the decoding of complex sounds, more differentiation in higher superior temporal regions that allow better invariance formation, and categorization of auditory objects into fine-grained representations. On the productive side, a similar refinement may have taken place: the ability to produce more intricate, coordinated movements, and, thus, more finely distinguishable sounds, which matched the expansion of perceptual space.

What might have ultimately enabled the evolution of an effective language system, however, might have been the development of a more effective network connecting the ventral and dorsal streams by virtue of a denser, more highly developed prefrontal network. This network, in conjunction with a more effective working memory system resulting from it, would have enabled the combinatorial explosion of vocabulary as well as semantic interpretations, and this would have elevated language to an instrument for communicating abstract thought that was much more limited in nonhumans. This, even more than the ability to learn and memorize patterns of vocal production, might have launched a new era of evolution that ultimately culminated in human language.

Acknowledgments

The collection of the underlying data and the writing of this article were supported by grants from the National Science Foundation

(BCS-0519127, OISE-0730255) and from the National Institutes of Health (R01NS052494, RC1DC010720). This article has been expanded from a previous communication by the author (Rauschecker, 2012).

See also: Biological Preconditions for Language Development; First Language Acquisition, Developmental Psychology of; Gibson's Theory of Perceptual Learning; Language Development, Theories of; Language Development: Emergentist Theories; Language Development: Evolutionary Theories; Mirror Neurons, Theory of; Neurophysiological Markers of Language Development; Tactile Processing.

Bibliography

Anwander, A., Tittgemeyer, M., von Cramon, D.Y., Friederici, A.D., Knosche, T.R., 2007. Connectivity-based parcellation of Broca's area. Cerebral Cortex 17, 816–825.

Artchakov, D., Ortiz, M., Kusmierek, P., Cui, D., VanMeter, J., Jääskeläinen, I., Sams, M., Rauschecker, J.P., 2012. Representation of sound sequences in the auditory dorsal stream after sensorimotor learning in the rhesus monkey. Society for Neuroscience – Abstract, 368.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309–312.

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S., Springer, J.A., Kaufman, J.N., Possing, E.T., 2000. Human temporal lobe activation by speech and nonspeech sounds. Cerebral Cortex 10, 512–528.

Bornkessel-Schlesewsky, I., Schlesewsky, M., 2013. Reconciling time, space and function: a new dorsal-ventral stream model of sentence comprehension. Brain and Language 125, 60–76.

Bornkessel-Schlesewsky, I., Schlesewsky, M., Small, S.L., Rauschecker, J.P. The neurobiological roots of language in primate audition: common computational properties, in preparation.

Catani, M., Jones, D.K., ffytche, D.H., 2005. Perisylvian language networks of the human brain. Annals of Neurology 57, 8–16.

Chang, E.F., Rieger, J.W., Johnson, K., Berger, M.S., Barbaro, N.M., Knight, R.T., 2010. Categorical speech representation in human superior temporal gyrus. Nature Neuroscience 13, 1428–1432.

Cheney, D.L., Seyfarth, R.M., 1990. How Monkeys See the World. University of Chicago Press, Chicago.

Chevillet, M.A., Riesenhuber, M., Rauschecker, J.P., 2011. Functional localization of the ventral auditory "what" stream hierarchy. Journal of Neuroscience 31, 9345–9352.

Chevillet, M.A., Jiang, X., Rauschecker, J.P., Riesenhuber, M., 2013. Automatic phoneme category selectivity in the dorsal auditory stream. Journal of Neuroscience 33, 5208–5215.

Cohen, Y.E., Russ, B.E., Davis, S.J., Baker, A.E., Ackelson, A.L., Nitecki, R., 2009. A functional role for the ventrolateral prefrontal cortex in non-spatial auditory cognition. Proceedings of the National Academy of Sciences of the United States of America 106, 20045–20050.

Croxson, P.L., Johansen-Berg, H., Behrens, T.E., Robson, M.D., Pinsk, M.A., Gross, C.G., Richter, W., Richter, M.C., Kastner, S., Rushworth, M.F., 2005. Quantitative investigation of connections of the prefrontal cortex in the human and macaque using probabilistic diffusion tractography. Journal of Neuroscience 25, 8854–8866.

DeWitt, I., Rauschecker, J.P., 2012. Phoneme and word recognition in the auditory ventral stream. Proceedings of the National Academy of Sciences of the United States of America 109, E505–E514.

DeWitt, I., Rauschecker, J.P., 2013. Wernicke's area revisited: parallel streams and word processing. Brain and Languages 127, 181–191.

Doupe, A.J., Kuhl, P.K., 1999. Birdsong and human speech: common themes and mechanisms. Annual Review of Neuroscience 22, 567–631.

Fitch, W.T., 2000. The evolution of speech: a comparative review. Trends in Cognitive Sciences 4, 258–267.

Frey, S., Campbell, J.S., Pike, G.B., Petrides, M., 2008. Dissociating the human language pathways with high angular resolution diffusion fiber tractography. Journal of Neuroscience 28, 11435–11444.

Frey, S., Mackey, S., Petrides, M., 2014. Cortico-cortical connections of areas 44 and 45B in the macaque monkey. Brain and Language 131, 36–55.

Friederici, A.D., 2004. Processing local transitions versus long-distance syntactic hierarchies. Trends in Cognitive Sciences 8, 245–247.

Friederici, A.D., 2012. The cortical language circuit: from auditory perception to sentence comprehension. Trends in Cognitive Sciences 16 (5), 262–268.

Fukushima, M., Saunders, R.C., Leopold, D.A., Mishkin, M., Averbeck, B.B., 2014. Differential coding of conspecific vocalizations in the ventral auditory cortical stream. Journal of Neuroscience 34, 4665–4676.

Gibson, J.J., 1966. The Senses Considered as Perceptual Systems. Allen & Unwin, London.

Hackett, T.A., 2011. Information flow in the auditory cortical network. Hearing Research 271, 133–146.

Hagoort, P., 2005. On Broca, brain, and binding: a new framework. Trends in Cognitive Sciences 9, 416–423.

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. Journal of Neuroscience 17, 4302–4311.

Kikuchi, Y., Horwitz, B., Mishkin, M., 2010. Hierarchical auditory processing directed rostrally along the monkey's supratemporal plane. Journal of Neuroscience 30, 13021–13030.

Kikuchi, Y., Horwitz, B., Mishkin, M., Rauschecker, J.P., 2014. Processing of harmonics in the lateral belt of macaque auditory cortex. Frontiers in Neuroscience 8, 204.

Koechlin, E., Jubault, T., 2006. Broca's area and the hierarchical organization of human behavior. Neuron 50, 963–974.

Leaver, A.M., Rauschecker, J.P., 2010. Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. Journal of Neuroscience 30, 7604–7612.

Leaver, A.M., Van Lare, J., Zielinski, B., Halpern, A.R., Rauschecker, J.P., 2009. Brain activation during anticipation of sound sequences. Journal of Neuroscience 29, 2477–2485.

Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech perception revised. Cognition 21, 1–36.

Margoliash, D., Fortune, E.S., 1992. Temporal and harmonic combination-sensitive neurons in the zebra finch's HVc. Journal of Neuroscience 12, 4309–4326.

Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F., 2014. Phonetic feature encoding in human superior temporal gyrus. Science 343, 1006–1010.

Morel, A., Garraghty, P.E., Kaas, J.H., 1993. Tonotopic organization, architectonic fields, and connections of auditory cortex in macaque monkeys. Journal of Comparative Neurology 335, 437–459.

Ojemann, G.A., Creutzfeldt, O., Lettich, E., Haglund, M.M., 1988. Neuronal activity in human lateral temporal cortex related to short-term verbal memory, naming and reading. Brain 111 (Pt 6), 1383–1403.

Pandya, D.N., Sanides, F., 1972. Architectonic parcellation of the temporal operculum in rhesus monkey and its projection pattern. Zeitschrift für Anatomie und Entwicklungsgeschichte 139, 127–161.

Pasupathy, A., Miller, E.K., 2005. Different time courses of learning-related activity in the prefrontal cortex and striatum. Nature 433, 873–876.

Penfield, W., Roberts, L., 1959. Speech and Brain Mechanisms. Princeton University Press, Princeton, NJ.

Petkov, C.I., Jarvis, E.D., 2012. Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. Frontiers in Evolutionary Neuroscience 4, 12.

Petkov, C.I., Kayser, C., Augath, M., Logothetis, N.K., 2006. Functional imaging reveals numerous fields in the monkey auditory cortex. PLoS Biology 4, e215.

Petkov, C.I., Kikuchi, Y., Milne, A., Mishkin, M., Rauschecker, J.P. & Logothetis, N.K. Effective connectivity in primate fronto-temporal pathways assessed with combined microstimulation and functional imaging, in preparation.

Petrides, M., Pandya, D.N., 2009. Distinct parietal and temporal pathways to the homologues of Broca's area in the monkey. PLoS Biology 7, e1000170.

Rauschecker, J.P., 1998. Cortical processing of complex sounds. Current Opinion in Neurobiology 8, 516–521.

Rauschecker, J.P., 2005. Vocal gestures and auditory objects. Behavioral and Brain Sciences 28, 143–144.

Rauschecker, J.P., 2011. An expanded role for the dorsal auditory pathway in sensorimotor control and integration. Hearing Research 271, 16–25.

Rauschecker, J.P., 2012. Ventral and dorsal streams in the evolution of speech and language. Frontiers in Evolutionary Neuroscience 4, 7.

Rauschecker, J.P., 2014. Is there a tape recorder in your head? How the brain stores and retrieves musical melodies. Frontiers in Systems Neuroscience 8, 149.

Rauschecker, J.P., Scott, S.K., 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature Neuroscience 12, 718–724.

Rauschecker, J.P., Tian, B., 2000. Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proceedings of the National Academy of Sciences of the United States of America 97, 11800–11806.

Rauschecker, J.P., Tian, B., Hauser, M., 1995. Processing of complex sounds in the macaque nonprimary auditory cortex. Science 268, 111–114.

Remedios, R., Logothetis, N.K., Kayser, C., 2009. Monkey drumming reveals common networks for perceiving vocal and nonvocal communication sounds. Proceedings of the National Academy of Sciences of the United States of America 106, 18010–18015.

Rilling, J.K., Glasser, M.F., Preuss, T.M., Ma, X., Zhao, T., Hu, X., Behrens, T.E., 2008. The evolution of the arcuate fasciculus revealed with comparative DTI. Nature Neuroscience 11, 426–428.

Romanski, L.M., Averbeck, B.B., 2009. The primate cortical auditory system and neural representation of conspecific vocalizations. Annual Review of Neuroscience 32, 315–346.

Romanski, L.M., Averbeck, B.B., Diltz, M., 2005. Neural representation of vocalizations in the primate ventrolateral prefrontal cortex. Journal of Neurophysiology 93, 734–747.

Romanski, L.M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P.S., Rauschecker, J.P., 1999. Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nature Neuroscience 2, 1131–1136.

Schmahmann, J.D., Pandya, D.N., Wang, R., Dai, G., D'Arceuil, H.E., de Crespigny, A.J., Wedeen, V.J., 2007. Association fibre pathways of the brain: parallel observations from diffusion spectrum imaging and autoradiography. Brain 130, 630–653.

Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123, 2400–2406.

Scott, S.K., 2005. Auditory processing–speech, space and auditory objects. Current Opinion in Neurobiology 15, 197–201.

Steinschneider, M., 2011. Unlocking the role of the superior temporal gyrus for speech sound categorization. Journal of Neurophysiology 105 (6), 2631–2633.

Suga, N., 1988. Auditory neuroethology and speech processing: complex-sound processing by combination-sensitive neurons. In: Edelman, G.M., Gall, W.E., Cowan, W.M. (Eds.), Auditory Function. Wiley, New York, pp. 679–720.

Warren, J.E., Wise, R.J., Warren, J.D., 2005. Sounds do-able: auditory-motor transformations and the posterior temporal plane. Trends in Neurosciences 28, 636–643.

Wedeen, V.J., Wang, R.P., Schmahmann, J.D., Benner, T., Tseng, W.Y., Dai, G., Pandya, D.N., Hagmann, P., D'Arceuil, H., de Crespigny, A.J., 2008. Diffusion spectrum magnetic resonance imaging (DSI) tractography of crossing fibers. NeuroImage 41, 1267–1277.

Wessinger, C.M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., Rauschecker, J.P., 2001. Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. Journal of Cognitive Neuroscience 13, 1–7.

Winter, P., Funkenstein, H.H., 1973. The effects of species-specific vocalization on the discharge of auditory cortical cells in the awake squirrel monkey (*Saimiri sciureus*). Experimental Brain Research 18, 489–504.

Wise, R.J., 2003. Language systems in normal and aphasic human subjects: functional imaging studies and inferences from animal studies. British Medical Bulletin 65, 95–119.

Wollberg, Z., Newman, J.D., 1972. Auditory cortex of squirrel monkey: response patterns of single cells to species-specific vocalizations. Science 175, 212–214.

Wolpert, D.M., Ghahramani, Z., Jordan, M.I., 1995. An internal model for sensorimotor integration. Science 269 (5232), 1880–1882.

Zatorre, R.J., Bouffard, M., Belin, P., 2004. Sensitivity to auditory object features in human temporal neocortex. Journal of Neuroscience 24, 3637–3642.

Human Language: Evolutionary Precursors

Yen Ying Lim, Warren Alpert School of Medicine of Brown University, Lifespan Hospital System, Providence, RI, USA
Peter J Snyder, Warren Alpert School of Medicine of Brown University, Lifespan Hospital System, Providence, RI, USA; and Rhode Island School of Design, Providence, RI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Language is a uniquely human trait, and yet the need to communicate transcends all species as it facilitates, at the very least, basic survival needs. While human language is unique and without direct comparators, there is evidence for proto-language in animals from birds to primates. Although the characteristics of these forms of communication may not be equivalent to language per se, in this article, we discuss several evolutionary antecedents, or precursors to human language, such as the organization of nervous systems with regional specialization, a mirror neuron system, and the ability to perform manual gestures.

All species on this earth must communicate with each other, whether through vocalizations or gestures, at the very least to facilitate basic survival needs such as sexual selection or alerting conspecifics about the presence of predators. Such intraspecies social communication is a core mediator of survival and natural selection and, as a result, innumerable researchers across a range of disciplines – from sociobiology to cognitive psychology, from linguistics to ecology – have attempted to study, define, and understand social communication. Language, however, is the human capacity to acquire and use complex systems of communication, such as vocal sounds, gestures, or written symbols, to express everything from basic needs to abstract thoughts and feelings. Language then, is a highly complex, rule-based, symbolic means of communication that has allowed humans to maintain complex social structures, to communicate abstract concepts, to share and to anticipate future events, and to convey subtle differences in emotions and intent. It is different from that of other species as our nervous systems have evolved and adapted such that we are not only capable of reasoning with abstraction, but we have also developed a system of communication that enables us to communicate such complex thoughts to each other. However, with the growth of evolutionary psychology, there has been increased evidence for proto-language in animals from birds to gorillas, and while these characteristics may not be considered to be equivalent to human language, they may be precursors to human language.

The complexity of human language has undoubtedly increased over the course of hominid evolution, and although human language is markedly different even from the intricate communication systems exhibited by our closest relatives, the other four extant genera of great apes, there remain elements of speech or communication necessary for human language to exist that are shared by our hominid ancestors, and even by other nonhominid species; elements that are in fact relied on across the entire vertebrate line in order to facilitate communication within species. For example, a key requirement for language is an ability to relate specific vocalizations to individual events or objects (classical conditioning) (Snyder and Harris, 1997). A second precondition for the support of higher cognitive functions, including language, appears to be regional (including inter-hemispheric) specialization of the neocortex. Neuropsychologists, neurologists, neurosurgeons, and practitioners of related disciplines have long studied the causal relations between structural and/or functional cerebral asymmetries and variations in cognitive, motor, and sensory functions. One only has to recall Broca's (1865) recognition of the localization of the motor–speech region for perhaps the best historical example tying speech–motor control to (left-) hemispheric cortical specialization. We now recognize that both handedness and language functions usually co-vary with respect to hemispheric dominance and, hence, anything related to the lateralization of the central nervous system (CNS) (e.g., handedness) also relates to what may be the quintessentially lateralized human brain function – language (Snyder, 2000).

An Asymmetrical Brain

Cerebral lateralization involves the division of information processing across the two cerebral hemispheres and although hemispheric specialization was once considered – even as recently as the 1970s – to be a uniquely human trait, an asymmetrical brain is ubiquitously shared among all vertebrates. One reason for the existence of an asymmetrical brain may be that the advantages associated with possessing such cerebral asymmetry far outweigh the ecological disadvantages associated with lateral biases. It is possible that lateralization may offer increased neural capacity because the specialization of one hemisphere for a particular function ultimately leaves the other hemisphere free to perform other tasks or functions, thus negating the need for duplicate cognitive functions within the two hemispheres (Denenberg, 1981). Another advantage of lateralization is that it may be an efficient way of increasing the brain's ability to simultaneously process multiple sources of information. A classic example of the advantages of lateralization in assisting the simultaneous processing of information is a study conducted in lateralized domestic chicks. A group of domestic chicks were strongly lateralized by exposing the eggs to light before hatching, a condition that can align and strengthen visual lateralization in chicks, and another group was weakly lateralized (i.e., eggs were incubated in the dark).

The chicks were then tested on a dual task that involved the left-hemispheric control of pecking responses and discriminating grain from pebbles, and the right-hemispheric control of monitoring overhead to detect a model predator (Rogers et al., 2004). Strongly lateralized chicks (light-incubated) did not only learn to discriminate between grain (food) and pebble faster than weakly lateralized (dark-incubated) chicks, but they were also able to detect the model predator sooner and were more responsive to the model predator than weakly lateralized chicks (Rogers et al., 2004). Although weakly lateralized chicks often failed to detect the model predator when the chicks were required to also discriminate between grain and pebbles, they had no difficulty in performing this foraging task when no model predator was present, suggesting that strong lateralization is required for chicks to be able to perform two separate tasks simultaneously (Rogers et al., 2004; Rogers, 2000). This advantage of lateralization has similarly been observed in pigeons (Güntürkün et al., 2000), and also in Australian parrots (Magat and Brown, 2009).

Lateralization may also have been naturally selected because it is a convenient way of preventing the simultaneous processing of incompatible responses in organisms with laterally placed eyes, such as the teleost fish (*Girardinus falcatus*). In one of the first experiments to demonstrate asymmetry in lower vertebrates, Cantalupo and colleagues demonstrated that both immature and mature teleost fish showed a bias to turn right when first presented with a simulated predator, but that there was a bias to turn left on subsequent presentations of the predator. The authors speculated that if there were an inborn tendency to monitor for potential predators using the left eye, presumably because the right cerebral hemisphere is selectively involved in emotional expression and control, it is possible that extensive exposure in a particular environment could reduce the strength of that asymmetry (Cantalupo et al., 1995).

Another function that is needed for language is the capacity to build large "vocabularies" of vocal sounds, which is based on the classical conditioning of associations between objects or actions, and vocalizations or speech sounds. This capacity appears to be related to lateral asymmetries for motor control, and also laterally placed nuclei. In a study of 70 African Grey parrots (*Psittacus erithacus*) that are known to exhibit possibly the greatest capacity, in captivity, to develop large lexicons of human speech sounds, right-footed African Greys had significantly larger lexicons than left-footed African Greys, and this difference in lexical size was not accounted for by any differences in training efforts by owners or by socialization with conspecifics (Snyder and Harris, 1997). Although there is a dearth of direct brain lesion studies in parrots – for obvious reasons – this strong relationship between footedness and lexical size in the African Grey is supportive of lesion studies in other bird species. For example, lesion studies in chickens have shown that the intermediate medial hyperstriatum ventrale (IMHV) region of the left hemisphere is involved in both short-term and long-term memory formation and storage, while the right IMHV is more preferentially involved in the initial encoding of memory for future recall (Rogers and Anson, 1979; Cipolla-Neto et al., 1982). Further evidence for hemispheric lateralization of memory formation in chicks also appears in the form of clear changes in synaptic density within the left, but not the right, IMHV (Bradley et al., 1981). Lesion and autoradiography studies of regional differences in CNS metabolic activity have also demonstrated a critical role of the left IMHV in the formation of long-term memory in a variety of learning tasks (Rose and Csillag, 1985; Patterson et al., 1990).

Clearly, the neural substrates of memory formation in the avian brain are complex, and involve regions beyond the IMHV; however, the results of these studies do suggest that the left and the right hemispheres of the chicken brain are specialized for different aspects of stimulus analysis and encoding of information for later recall. Given this, it is possible that one hemisphere of the parrot brain (left hemisphere) is also more involved in the encoding of new words into a parrot's lexicon. Although lexical size is more dependent on the parrot's ability to learn, rehearse, and memorize a novel set of vocal calls that are nonspecies specific, than on motor control, right-footed African Grey parrots are able to better engage in intentional or unintentional classical conditioning to form better associations between their owners' vocal calls and particular events, thus gaining a larger lexicon (Snyder and Harris, 1997). Thus, right-footedness in African Grey parrots may represent a behavioral index of contralateral hemispheric specialization for both motor control and for other 'higher-level' cognitive function, i.e., categorization and long-term mnestic processing of human speech sounds (Snyder and Harris, 1997).

Taken together, there are evolutionary advantages for CNS regional specialization of complex higher cognitive functions, and it is likely that the language evolved as a result of more basic functions that have been conserved across the vertebrate line, such as the ability to engage in classical conditioning and categorical learning. As the mammalian and avian lines diverged about 350 million years ago, it is possible that we are either looking at homologous brain organization as a result of common early ancestor(s). Alternatively, this common pattern of CNS organization between mammals and birds may represent a classic example of convergent evolution. Importantly though, regardless of whether it is homology or convergent evolution, the asymmetric organization of the CNS in mediating higher cognitive functions is likely to be evolutionarily advantageous and suggests that precursors to human language would most likely require a lateralized brain.

Mirror Neuron System

One of the most important cognitive capacities that serves as a prerequisite for the evolution of language, specifically the ability to understand the actions of conspecifics and to both imitate and learn from them, is governed by what is now known as the 'mirror neuron system.' Mirror neurons were first discovered in area F5 of the monkey ventral premotor cortex, and represent a distinctive class of neurons whose activity is modulated when a monkey performs a particular action and when it observes another monkey or human performing a similar action (Di Pellegrino et al., 1992; Rizzolatti et al., 1996). However, mirror neurons only respond when a monkey observes an object-directed action, the mere sight of an object alone is insufficient to trigger activation of the mirror neuron system (Rizzolatti and Luppino, 2001; Rizzolatti and Craighero, 2004). In particular, mirror neurons require some

form of interaction between an effector (either hand or mouth) and an object. The function of mirror neurons has been contested, and they have been implicated in a range of functions from imitation and action understanding, to empathy development. It is likely though, that rather than possessing a unique functional role, mirror neurons subserve a range of neural mechanisms that map visual representations of actions onto their motor counterparts. An important example of this is gesturing as a form of communication in apes, which, as we discuss below, may be a manual precursor to oral language.

As mirror neurons are characterized by their response to both observation and execution of the same action, the human ability to imitate has been proposed to have evolved from the mirror neuron system in monkeys (Arbib, 2005). Indeed, the mirror neuron system has been studied in humans, and while there is no direct evidence of any mirror neurons in humans, functional neuroimaging and magnetic stimulation studies suggest that a similar mirror system exists in humans. One example is a functional magnetic resonance imaging (fMRI) study conducted in normal human volunteers where participants were subjected to two conditions – observation only and observation-execution (Iacoboni et al., 1999). In an 'observation only' situation, participants were only required to observe the stimuli. In the 'observation-execution' situation, the same stimuli were presented but participants were also required to lift their right finger as quickly as possible in response to the stimuli. Interestingly, fMRI showed strong activity in the left *pars opercularis* of the inferior frontal gyrus (IFG), the right parietal region, the right parietal operculum, and the right STS region – but only in the 'observation-execution' situation. Similarly, in another study where human volunteers were required to imitate finger movements that were either goal oriented or actions with no explicit goal, the greatest blood oxygen level-dependent fMRI signal was observed in the *pars opercularis* of the IFG when goal-oriented movements were performed (Koski et al., 2002). This particular activation in the IFG follows a pattern that is consistent with that observed in the area F5 mirror neurons in monkeys, and anatomical studies have suggested that area F5 in the ventral premotor cortex of monkeys is homologous to Brodmann's area 44 in humans (Rizzolatti and Arbib, 1998), which is activated by nonlinguistic motor functions such as gestures and sensorimotor learning (Binkofski and Buccino, 2004) (Brodmann's area 44 differs from area 45 in that area 45 is activated by spoken and signed language (Horwitz et al., 2003)).

An important feature of the neural areas implicated in human language is that they are represented primarily in the left cerebral hemisphere (Hickok and Poeppel, 2004, 2007; Poeppel and Hickok, 2004). This left-hemispheric bias has been proposed to have arisen from the incorporation of vocalization (Corballis, 2010), as vocalization appears to be left hemispheric in nonhominid species such as frogs (Bauer, 1993) and mice (Ehret, 1987). It is important to consider though, that while the primate mirror neuron system can be activated by the sounds of actions, primate vocalizations themselves do not appear to activate this system (Kohler et al., 2002). However, some studies have found that the mirror neuron system in humans is lateralized to the left hemisphere for manual operations unrelated to language. When participants were required to perform a task in which they had to identify pictures of a rotated hand as either a left or a right hand, a task which required participants to simultaneously imagine their own hand rotated to match the pictured hand, activation of the inferior parietal cortex was restricted only to the left hemisphere regardless of whether the pictured hand was a left or a right hand (Sekiyama et al., 2000). Observation of a split-brained patient also showed that only the left hemisphere was activated when the patient engaged in an action observation task, suggesting that any right-hemisphere activation in intact humans (such as that observed by Aziz-Zadeh et al. (2005)) may be a result of data transmission across the corpus callosum from the left hemisphere. Taken together, it appears that in the majority of individuals, the mirror neuron system is predominantly left hemispheric particularly when it pertains to language functions. It is possible then that Broca's area may have evolved on the top of the primate mirror neuron system for grasping (ventral premotor area F5), with its capacity to recognize, imitate, and generate a set of actions, and may indeed even be a reason for the higher prevalence of right handedness in humans (Arbib, 2005; Corballis, 2010).

Gestural Theory of Language Evolution

Given the homology between Broca's area in humans and the ventral premotor area F5 in monkeys, several authors have proposed that language may have evolved from manual gestures rather than directly from vocalizations (Arbib, 2005; Corballis, 2010; Fogassi and Ferrari, 2007; Rizzolatti and Craighero, 2004; Goldin-Meadow, 2005; Rizzolatti and Arbib, 1998). In fact, one does not need to look far to see the intimate relationship between manual gestures and vocal speech. Manual hand gestures are used extensively in people with normal hearing, even in individuals who are blind (Willems and Hagoort, 2007; Goldin-Meadow, 2005), and hand movements are often seen in the early developmental years of young children. For example, babbling in infants is often accompanied by rhythmic hand movements (Masataka, 2001), and word productions in children aged between 11 and 13 months are often accompanied by finger pointing, and gestures that indicate object recognition (Bates and Snyder, 1987; Iverson and Goldin-Meadow, 2005). Finally, language does not necessarily require vocalization. The signed languages of the deaf also have the linguistic complexities and sophistication of spoken languages (Emmorey et al., 2002), and Broca's area has been found to be activated by both speaking and signing (Horwitz et al., 2003).

Neurological research in nonhuman primates have shown that when primates generate vocal calls, the neural circuits activated involve the cingulate cortex, diencephalon, and brainstem structures – circuits that are radically different to those involved when humans speak (Jürgens, 2002; Rizzolatti and Sinigaglia, 2008), and circuits that are more primitive in nature. Further, while vocalizations occur frequently in all animals, the ability to learn vocalizations appear to be restricted to very few nonhuman species, including elephants, seals, killer whales, and some birds, and of the primates, only humans are capable of learning vocalizations (Jarvis, 2006). Attempts to teach human language to nonhuman primates have quickly revealed that primates are unable to produce

anything resembling human language or speech. In perhaps the most famous example, a baby chimpanzee that was reared in a human family was able to articulate only three or four words, and its verbal prowess was very quickly eclipsed by those of the human children in the family (Hayes and Hayes, 1952). However, despite the inability of primates to learn vocalizations, gorillas and bonobo chimpanzees have been found to be capable of socially learning and attending to the communicative bodily gestures of the recipient, both of which are critical to spoken language (Liebal et al., 2004a, 2004b; Pika et al., 2003, 2005). This has led some to argue that the inability to engage in vocal learning and to produce speech may be a result of the limitations of the primate vocal apparatus rather than an inability to produce language itself. Indeed, attempts to teach language to nonhuman primates based on manual gestures have been slightly more successful. For example, researchers at the University of Nevada have taught the chimpanzee Washoe how to communicate using American Sign Language (ASL), and after developing an impressive vocabulary, she was able to combine signs into two- or three-word sequences to perform simple requests (Gardner and Gardner, 1969), making Washoe the first nonhuman to be able to communicate using ASL.

Examinations of food-begging manual gestures and intraspecies communicative exchanges in chimpanzees have revealed a right-hand bias (Meguerditchian et al., 2010; Hopkins et al., 2005). Interestingly, the right-hand preference for species-typical gestures was similar to the right-hand bias directed at humans when requesting for food, but hand preferences for a noncommunicative action (e.g., nose wipe) was not correlated with any communicative hand preference (Meguerditchian et al., 2010). It was thus proposed that as communicative gestures consistently show a right-hand preference, it is possible that chimpanzees, our closest primate relative, may possess a specific left-lateralized cerebral system for communication, and this system may be involved in the production of gestures.

Taken together, while the ability to learn vocalizations is also critical to spoken language, it is possible that this ability occurred much later in the evolutionary path toward human language. Rather, given our primate heritage, our early ancestors might be more likely to have been preadapted to generate spontaneous and intentional communication based on manual and other bodily gestures rather than on vocalizations, and thus, gestural actions in nonhuman primates may constitute an important evolutionary precursor to the emergence of human language.

In considering these aspects, Michael Arbib has proposed seven stages in the evolution of human language that would provide a useful framework of the multitude of evolutionary precursors to human language (Arbib, 2005, 2009).

Stage 1: Grasping and manual gestures or actions that are similarly shared with nonhuman primates.

Stage 2: A mirror neuron system for grasping that can recognize and encode.

Stage 3: An ability to perform simple imitations of object-directed grasping through repeated exposure to stimuli.

Stage 4: An ability to perform more complex imitations for grasping, especially an ability to also recognize conspecifics' gestures as a set of familiar actions and then repeat them, or that the gestures performed may be a combination of novel actions that are approximate to or variants of already familiar actions.

Stage 5: Protosign, which is a gestural-based communication system that formalizes and disambiguates imitations, evolves (e.g., being able to distinguish 'bird' from 'flying,' or 'food' and 'eating').

Stage 6: Protosign would provide the early scaffolding of control mechanisms required for the development of proto-speech, which is the ability to control the vocal apparatus with increasing flexibility until proto-speech becomes dominant.

Stage 7: Language. With a language-ready brain and the capacity to engage in manual–vocal communication, humans gradually shifted from action-object frames to verb-argument structures to syntax and semantics. The increased need for communication in order to survive in a hostile world, coupled with possible intermarriages between groups and tribes who developed separate languages may have thus spurred on the rapid development of human language. It is this final stage that has possibly the least to do with any biological evolution, but may have arisen primarily from cultural evolution and historical change instead.

Genetic and Physical Considerations

It is without any doubt that the emergence of human language has been a result of several quantum leaps in our physiological and genetic structure. For example, the human version of the Forkhead-box P2 (*FOXP2*) gene is functionally identical in all normal human populations worldwide, but differs significantly from our nearest primate relative, the chimpanzee. *FOXP2* in humans and chimpanzees differ by two amino acids – threonine to asparagine substitution at position 303 and asparagine to serine substitution at position 325. This seemingly minor difference of two amino acids has been enough to result in major behavioral and cognitive functions in the lineage of humans. The *FOXP2* gene has been dubbed the 'language gene' as a mutation of this gene in an English family known as the KE family has been found to result in severe deficits in vocal articulation (Belton et al., 2003; Lai et al., 2001). Neuroimaging studies have reported that unlike their unaffected relatives, affected members of the KE family show no activation in Broca's area when they were required to covertly (silently) generate verbs (Liégeois et al., 2003), and significantly reduced gray matter density bilaterally in the caudate nucleus, the cerebellum, and the left and right IFG (Belton et al., 2003) – the cerebral regions that are implicated in both motor and language functions.

Another important factor to consider in the evolutionary path toward human language is the unusual vocal tract of humans. Our descended larynx, the large and rounded shape of our tongue, and the angle between our pharyngeal and oral cavities are fundamentally different to the vocal tracts of other nonhuman species (Lieberman, 1984; Negus, 1949). Lieberman argued that it is this difference that enables humans to produce a greater range of vocalizations and have greater

diversity of formant frequencies. However, cineradiographic observations of nonhuman mammal vocalizations have shown that some mammals lower their larynx during vocalization (e.g., dogs barking) (Fitch, 2005). There are also a number of nonhuman species that have a permanently descended larynx, such as koalas, deer, and lions. These observations suggest then that lowering the larynx to generate more impressive vocalizations (e.g., lion roaring) is a primitive trait, and natural selective forces that are independent of speech may have driven laryngeal descent at some stage in our evolutionary path toward language, and this lowering of the larynx in turn may have acted as a preadaptation for our much later phonetic capabilities (Fitch, 2005).

Conclusions

The purpose of this article was *not* to argue that language runs on a continuum from fish to humans. As discussed above, there have indeed been some dramatic changes in our physiological and genetic structure to allow for human language to emerge. However, despite these important differences between humans and other nonhuman species, there are still some basic requirements that evolutionarily, may serve as precursors to human language, such as an asymmetrical brain, a mirror neuron system, and the ability to perform manual gestures. By themselves, each factor would not be sufficient to produce language as we know it; however, as Charles Darwin would put it, "I cannot doubt that language owes its origin to the imitation and modification, aided by signs and gestures, of various natural sounds, the voices of other animals, and man's own distinctive cries. When we treat of sexual selection we shall see that primeval man, or rather some early progenitor of man, probably used his voice largely, as does one of the gibbon-apes at the present day, in producing true musical cadences, that is in singing; we may conclude from a widely-spread analogy that this power would have been especially exerted during the courtship of the sexes, serving to express various emotions, as love, jealousy, triumph, and serving as a challenge to their rivals. The imitation by articulate sounds of musical cries might have given rise to words expressive of various complex emotions." (Darwin, 1871: p. 56).

Bibliography

Arbib, M.A., 2005. From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. The Behavioural and Brain Sciences 28, 105–124.

Arbib, M.A., 2009. Evolving the language-ready brain and the social mechanisms that support language. Journal of Communication Disorders 42, 263–271.

Aziz-Zadeh, L., Koski, L., Zaidel, E., Mazziotta, J., Iacoboni, M., 2005. Lateralization of the human mirror neuron system. Journal of Neuroscience 26, 2964–2970.

Bates, S., Snyder, L.S., 1987. The Cognitive Hypothesis in Language Development. University of Illinois Press, Champaign, IL, US.

Bauer, R.H., 1993. Lateralization of neural control for vocalization by the frog (*Rana pipiens*). Psychobiology 21, 243–248.

Belton, E., Salmond, C.H., Watkins, K.E., Vargha-Khadem, F., Gadian, D.G., 2003. Bilateral brain abnormalities associated with dominantly inherited verbal and orofacial dyspraxia. Human Brain Mapping 18, 194–200.

Binkofski, F., Buccino, G., 2004. Motor functions in the Broca's region. Brain and Language 89, 362–369.

Bradley, P., Horn, G., Bateson, P., 1981. Imprinting. An electron microscopic study of chick hyperstriatum ventrale. Experimental Brain Research 41, 115–120.

Broca, P., 1865. Sur le siege de la faculte du langage articule. Bulletin de la Societe d'Anthropologie 6, 337–393.

Cantalupo, C., Bisazza, A., Vallortigara, G., 1995. Lateralization of predator-evasion response in teleost fish (*Girardinus falcatus*). Neuropsychologia 33, 1637–1646.

Cipolla-Neto, J., Horn, G., Mccabe, B.J., 1982. Hemispheric asymmetry and imprinting: the effect of sequential lesions of the hyperstriatum ventrale. Experimental Brain Research 48, 22–27.

Corballis, M.C., 2010. Mirror-neurons and the evolution of language. Brain and Language 112, 25–35.

Darwin, C., 1871. The Descent of Man, and Selection in Relation to Sex. Murray, London.

Denenberg, V.H., 1981. Hemispheric laterality in animals and the effects of early experience. Behavioral and Brain Sciences 4, 35–36.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., Rizzolatti, G., 1992. Understanding motor events: a neurophysiological study. Experimental Brain Research 91, 176–180.

Ehret, G., 1987. Left hemisphere advantage in the mouse brain for recognizing ultrasonic communication calls. Nature 325, 249–251.

Emmorey, K., Damasio, H., Mccullough, S., Grabowski, T., Ponto, L.L., Hichwa, R.D., Bellugi, U., 2002. Neural systems underlying spatial language in American sign language. NeuroImage 17, 812–824.

Fitch, W.T., 2005. The evolution of language: a comparative review. Biology and Philosophy 20, 193–230.

Fogassi, L., Ferrari, P.F., 2007. Mirror neurons and the evolution of embodied language. Current Directions in Psychological Science 16, 136–141.

Gardner, R.A., Gardner, B.T., 1969. Teaching sign language to a chimpanzee. Science 165, 664–672.

Goldin-Meadow, S., 2005. Hearing Gestures: How Our Hands Help Us Think. Harvard University Press, USA.

Güntürkün, O., Diekamp, B., Manns, M., Nottelmann, F., Prior, H., Schwarz, A., Skiba, M., 2000. Asymmetry pays: visual lateralization improves discrimination success in pigeons. Current Biology 10, 1079–1081.

Hayes, K.J., Hayes, C., 1952. Imitation in a home-raised chimpanzee. Journal of Comparative and Physiological Psychology 45, 450–459.

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92, 67–99.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nature Reviews, Neuroscience 8, 393–402.

Hopkins, W.D., Russell, J., Freeman, H., Buehler, N., Reynolds, E., Schapiro, S.J., 2005. The distribution and development of handedness for manual gestures in captive chimpanzees (*Pan troglodytes*). Psychological Science 16, 487–493.

Horwitz, B., Amunts, K., Bhattacharyya, R., Patkin, D., Jeffries, K., Zilles, K., Braun, A.R., 2003. Activation of Broca's area during the production of spoken and signed language: a combined cytoarchitectonic mapping and PET analysis. Neuropsychologia 41, 1868–1876.

Iacoboni, M., Woods, R.P., Brass, M., Bekkering, H., Mazziotta, J.C., Rizzolatti, G., 1999. Cortical mechanisms of human imitation. Science 286, 2526–2528.

Iverson, J.M., Goldin-Meadow, S., 2005. Gesture paves the way for language development. Psychological Science 16, 367–371.

Jarvis, E.D., 2006. Selection for and against vocal learning in birds and mammals. Ornithological Science 5, 5–14.

Jürgens, U., 2002. Neural pathways underlying vocal control. Neuroscience and Biobehavioural Reviews 26, 235–258.

Kohler, E., Keysers, C., Umiltà, M.A., Fogassi, L., Gallese, V., Rizzolatti, G., 2002. Hearing sounds, understanding actions: action representation in mirror neurons. Science 297 (5582), 846–848.

Koski, L., Wohlschläger, A., Bekkering, H., Woods, R.P., Dubeau, M.C., Mazziotta, J.C., Iacoboni, M., 2002. Modulation of motor and premotor activity during imitation of target-directed actions. Cerebral Cortex 12, 847–855.

Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., Monaco, A.P., 2001. A forkhead-domain gene is mutated in a severe speech and language disorder. Nature 413, 519–523.

Liebal, K., Call, J., Tomasello, M., 2004a. Use of gesture sequences in chimpanzees. American Journal of Primatology 64, 377–396.

Liebal, K., Pika, S., Tomasello, M., 2004b. Social communication in siamangs (*Symphalangus syndactylus*): use of gestures and facial expressions. Primates 45, 41–57.

Liégeois, F., Baldeweg, T., Connelly, A., Gadian, D.G., Mishkin, M., Vargha-Khadem, F., 2003. Language fMRI abnormalities associated with *FOXP2* gene mutation. Nature Neuroscience 6, 1230–1237.

Lieberman, P., 1984. The biology and evolution of language. Harvard University Press, Cambridge, MA.

Magat, M., Brown, C., 2009. Laterality enhances cognition in Australian parrots. Proceedings. Biological Sciences 276, 4155–4162.

Masataka, N., 2001. Why early linguistic milestones are delayed in children with Williams syndrome: late onset of hand banging as a possible rate–limiting constraint on the emergence of canonical babbling. Developmental Science 4, 158–164.

Meguerditchian, A., Vauclair, J., Hopkins, W.D., 2010. Captive chimpanzees use their right hand to communicate with each other: implications for the origin of the cerebral substrate for language. Cortex 46, 40–48.

Negus, V.E., 1949. The comparative anatomy and physiology of the larynx. Hafner Publishing Company, New York.

Patterson, T.A., Gilbert, D.B., Rose, S.P.R., 1990. Pre- and post training lesions of the intermediate medial hyperstriatum ventrale and passive avoidance learning in the chick. Experimental Brain Research 80, 189–195.

Pika, S., Liebal, K., Tomasello, M., 2003. Gestural communication in young gorillas (*Gorilla gorilla*): gestural repertoire, learning, and use. American Journal of Primatology 60, 95–111.

Pika, S., Liebal, K., Tomasello, M., 2005. Gestural communication in subadult bonobos (*Pan paniscus*): repertoire and use. American Journal of Primatology 65, 39–61.

Poeppel, D., Hickok, G., 2004. Towards a new functional anatomy of language. Cognition 92, 1–12.

Rizzolatti, G., Arbib, M.A., 1998. Language within our grasp. Trends in Neurosciences 21, 188–194.

Rizzolatti, G., Craighero, L., 2004. The mirror-neuron system. Annual Review of Neuroscience 27, 169–192.

Rizzolatti, G., Fadiga, L., Gallese, V., Fogassi, L., 1996. Premotor cortex and the recognition of motor actions. Brain Research. Cognitive Brain Research 3, 131–141.

Rizzolatti, G., Luppino, G., 2001. The cortical motor system. Neuron 31, 889–901.

Rizzolatti, G., Sinigaglia, G., 2008. Further reflections on how we interpret the actions of others. Nature 455, 589.

Rogers, L.J., 2000. Evolution of hemispheric specialization: advantages and disadvantages. Brain and Language 73, 236–253.

Rogers, L.J., Anson, L.M., 1979. Lateralization of function in the chicken fore-brain. Pharmacology, Biochemistry, and Behavior 10, 679–686.

Rogers, L.J., Zucca, P., Vallortigara, G., 2004. Advantages of having a lateralized brain. Proceedings. Biological Sciences 271, S420–S422.

Rose, S.P.R., Csillag, A., 1985. Passive avoidance training results in lasting changes in deoxyglucose metabolism in left hemisphere regions of chick brain. Behavioral and Neural Biology 44, 315–324.

Sekiyama, K., Miyauchi, S., Imaruoka, T., Egusa, H., Tashiro, T., 2000. Body image as a visuomotor transformation device revealed in adaptation to reversed vision. Nature 407, 374–377.

Snyder, P.J., 2000. Evolutionary bases for cerebral localization of higher cognitive functions. Brain and Language 73, 127–131.

Snyder, P.J., Harris, L.J., 1997. Lexicon size and its relation to foot preference in the African Grey parrot (*Psittacus erithacus*). Neuropsychologia 35, 919–926.

Willems, R.M., Hagoort, P., 2007. Neural evidence for the interplay between language, gesture, and action: a review. Brain and Language 101, 278–289.

Human Mate Choice

Yasuo Ihara, The University of Tokyo, Bunkyoku, Tokyo, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Human mate choice is partially affected by individual preference for physical or mental characteristics of potential partners. This article examines genetic influences on human mate choice and possible consequences of mate choice on the genetic composition of populations. It considers two examples of physical traits on which human mate choice may depend, namely the human major histocompatibility complex and facial characteristics.

Introduction

Human mate choice could be affected by a myriad of factors. Social norms often encourage or discourage certain types of unions on the basis of kin relations, while individuals may have preferences for certain physical or mental characteristics possessed by potential partners. This article focuses on those aspects of human mate choice driven by individual preferences and briefly reviews some of the recent developments in the field. Firstly, the article discusses the interaction of genes and mate choice in humans. Individual preferences affecting human mate choice may have a genetic basis. More precisely, differences among individuals in mating preferences may be partially accounted for by differences in genes among those individuals. At the same time, mate choice may also have an impact on genes. This is because if individuals mate selectively with respect to traits that are at least partially heritable, that could give rise to nonrandom association of different genes or genetic diversification between populations. Secondly, as an example of a highly heritable physical trait that is relevant to human mate choice, the article considers the major histocompatibility complex (MHC). The MHC plays an important role in the immune system of vertebrates, including humans, and is characterized by its extreme polymorphism so that virtually every individual, except for the case of monozygotic twins, has a different MHC type. Recent studies have suggested that the MHC plays a role in mate choice in humans as well as in nonhuman animals. Thirdly, the article considers a more manifest and partially heritable target of human mate choice, namely facial characteristics. In particular, recent studies have made close examinations to confirm the anecdotal observation that spouses of heterosexual unions resemble each other. Facial similarity between such spouses raises a possibility that individuals may tend to find self-resembling opposite-sex faces attractive and that choice of marriage partners may be affected by this preference.

Genes and Human Mate Choice

One way to explore genetic influences on human behavior is to compare similarities in behaviors of interest within monozygotic (MZ) twin pairs and within dizygotic (DZ) twin pairs, so-called twin studies. MZ twins have virtually identical genotypes, while DZ twins share on average 50% of their genes that are identical by descent. Hence, assuming that the shared environmental effect does not differ between MZ and DZ twins, higher correlation for MZ than DZ twins indicates that behavioral variation among individuals is partially due to genetic variation.

Zietsch et al. (2011) investigated the genetic and environmental influences on mate choice in a large sample of twins and their partners and parents ($N > 20\,000$ individuals), by analyzing key traits potentially relevant to mate choice. There were positive correlations between partners of twins for education, income, religiosity, social attitudes, and age, suggesting that twins may tend to choose similar partners with respect to these traits. However, the correlations were not significantly higher for MZ than DZ twins, indicating no genetic influence on mate choice. These results are consistent with an earlier study by Lykken and Tellegen (1993) based on a smaller sample of twins (152 and 117 pairs of MZ and DZ twins, respectively). In contrast, Rushton and Bons (2005) found that spouses of MZ twins (174 pairs) were more similar to each other than were those of DZ twins (148 pairs) in their responses to personality, attitude, and demographic questions, which was indicative of a genetic influence.

As mentioned in this article, human mate choice is affected by various factors, thus making it difficult to detect genetic influences on 'realized' mate choice. Nevertheless, genetic influences on individual preferences, whether or not they are fulfilled in actual mate choice, may be revealed by asking people what they like. Phillips et al. (2010) asked twin pairs to complete a psychometric scale measuring mate preference toward altruistic traits in a potential mate (MPAT scale; Phillips et al., 2008). Analyzing the sample of 70 MZ and 87 DZ twin pairs (all females), they found a significant positive correlation in the responses to the MPAT scale for MZ twin pairs but not for DZ twin pairs. Hence, the variance in female preference toward altruistic traits is partially explained by the genetic variance. More recently, Zietsch et al. (2012) investigated genetic influences on long-term mate preferences by asking more than 4000 twins to evaluate the relative importance of 13 traits that might be present in a potential mate (Buss and Barnes, 1986): kind and understanding, easygoing, intelligent, healthy, physically attractive, exciting personality, wants children, good earning capacity, creative and artistic, good housekeeper, good heredity, university graduate, and religious. They compared correlations between MZ twins and between DZ twins for the reported relative importance of each trait and found that

correlations were higher for MZ than DZ twins for almost all the traits for both sexes, suggesting widespread genetic influences on mating preferences.

Laurent et al. (2012) analyzed the genome-wide data on single nucleotide polymorphisms (SNPs) in three populations of African, European-American, and Mexican origins from the HapMap 3 database (Altshuler et al., 2010) to detect extreme patterns of genetic similarity or dissimilarity between spouses. The average kinship coefficient, a measure of relative genetic similarity, between spouses was computed for each of 19 979 genes and tested whether it deviated from zero in a statistically significant manner by comparing it against null distributions obtained by randomization procedures. They provided a list of genes that showed significant signatures of nonrandom mating (although most of them did not reach statistical significance under a more stringent criterion). The list included some genes that might influence physical appearance and thus were potentially interesting in the context of mate choice. For example, assortative mating (i.e., a tendency to mate with someone similar to self for a given gene) was detected in the European-American population for MLPH, a gene involved in pigmentation, and PLOD1, a gene associated with epidermis development. Also, in the Mexican population, disassortative mating (i.e., tendency to mate with someone dissimilar to self) was detected for KRT32 and KRT35, two keratin genes involved in epidermis development.

Genes involved in skin appearance exhibited signatures of nonrandom mating in the European-American and Mexican populations but not in the African population: in fact, most of the genes on the list were population specific. As indicated by Laurent et al. (2012), the limited overlap between populations might be due to population-specific mating preferences culturally maintained over a large number of generations. If that is the case, this can be regarded as an example of gene–culture coevolution, where culture contributes to shape the genetic compositions of populations (see Laland et al., 2010).

Human Mate Choice Based on the Major Histocompatibility Complex

The MHC plays a fundamental role in the immune system of vertebrate species. MHC molecules are cell surface glycoproteins whose principal function is to present foreign antigens to T cells to initiate immune response against invading parasites. The extreme polymorphism of the genetic loci encoding MHC molecules is thought to be maintained by such mechanisms as heterozygote advantage, frequency-dependent selection, maternal–fetal interactions, and nonrandom mating (Apanius et al., 1997).

In 1976, Yamazaki and colleagues reported the results of their experiments indicating that house mice apparently have a mating preference dependent on MHC similarity between individuals (Yamazaki et al., 1976). Succeeding studies have confirmed that male and female mice are more likely to mate with individuals having MHC genotypes dissimilar to theirs than with MHC-similar individuals (Egid and Brown, 1989; Potts et al., 1991; reviewed in Penn and Potts, 1999) and that the recognition of MHC types of others is mediated by olfaction (Yamazaki et al., 1979; reviewed in Penn and Potts, 1998). Disassortative mating for the MHC tends to result in MHC-heterozygous and thus immunocompetent offspring as well as inbreeding avoidance. Possible influences of the MHC on mate choice have also been explored in other species of mammals and in birds, reptiles, and fish (reviewed in Piertney and Oliver, 2006).

The possibility that the MHC may be also relevant to human mate choice was first tested experimentally by Wedekind et al. (1995). In their experiment, 44 male university students were asked to wear a T-shirt for two consecutive nights and live as much as possible 'odor-neutral' during these two days. Then 49 female university students evaluated the odors of the T-shirts worn by males who were similar or dissimilar to the females at MHC loci. It turned out that females who were not taking oral contraceptives tended to rate the odors of MHC-dissimilar males more pleasant and sexier than those of MHC-similar males. Furthermore, the odors of MHC-dissimilar males were more likely to remind the females of their current or previous mates than were those of MHC-similar males. In a subsequent experiment by Wedekind and Füri (1997), 121 male and female participants evaluated the odors of T-shirts worn by four males and two females. Consistent with the earlier study, pleasantness ratings given to the odors were negatively correlated with the degree of MHC similarity between smellers (males and females not using oral contraceptives pooled) and T-shirt wearers (males and females pooled). These results seem to support the hypothesis that humans have an MHC-dependent odor preference that may consciously or unconsciously influence mate choice decisions, as in some species of nonhuman vertebrates.

However, further experiments by other research groups have produced mixed results (reviewed in Havlicek and Roberts, 2009). First, Jacob et al. (2002) asked 49 female raters to evaluate the odors of T-shirts worn by six male donors and found that the donor of a female's most preferred odor shared significantly more MHC alleles with her than did the donor of her least preferred odor. This result seems to contradict the previous studies by Wedekind and colleagues, which found a negative impact of MHC similarity on perceived odor pleasantness, although Jacob et al. (2002) argued that the results are consistent if odor pleasantness is maximized for an optimal MHC similarity. Second, in Thornhill et al. (2003), 97 male and 100 female participants provided body odor by wearing T-shirts and also evaluated the odors of opposite-sex participants. For the analysis, data on the odors of 56 males and 48 females rated by 77 males and 65 females not using oral contraceptives were used. Attractiveness ratings given to a female's odor tended to negatively correlate with MHC similarity between the female and male raters, suggesting that males prefer the odor of MHC-dissimilar females. In contrast, no evidence of MHC-dependent preference was found in female raters. Third, Santos et al. (2005) collected the body odor of 29 male and 29 female university students not by T-shirts but by necklaces with a cotton sachet. The male and female participants subsequently evaluated the odors of opposite-sex participants and classified them into three categories: pleasant, indifferent, or unpleasant. MHC similarity was statistically associated with perceived odor pleasantness only in the case when female raters evaluated male odor: more specifically, females were more likely to classify a male's odor

as indifferent when they shared fewer MHC alleles with the odor donor. Fourth, following the original experiment by Wedekind et al. (1995), Roberts et al. (2008) asked each female participant to evaluate the odors of six males who were similar or dissimilar to her at MHC loci. In total, ratings given to the odors of 97 males by 110 females not using oral contraceptives were analyzed. In spite of the nearly identical experimental design, the results of Wedekind et al. (1995) were not replicated successfully: Roberts et al. (2008) found no association between perceived odor pleasantness and MHC similarity.

If humans do have such MHC-dependent odor preference as claimed by Wedekind et al. (1995), it might have an effect on their actual choice of marriage partners. This raises an interesting hypothesis to be tested: married couples may be more dissimilar to each other at MHC loci than expected by chance. Indeed, Ober et al. (1997) provided evidence supporting the hypothesis. They studied 411 married couples in the Hutterites, a North American reproductive isolate known to have limited MHC variation, and found that the number of couples matching for a five-locus MHC haplotype, 44, was significantly smaller than expected by chance. However, Hedrick and Black (1997), who examined the numbers of shared MHC alleles among 194 couples from 11 South Amerindian tribes, did not find any significant deviation from those expected by random mating. Ihara et al. (2000) also compared the observed and expected MHC similarity between wives and husbands for 154 couples from six prefectures in the Tohoku region of Japan and 291 couples from 16 prefectures all over Japan. They too did not find any evidence of disassortative mating for MHC loci (for related studies, see also Havlicek and Roberts, 2009).

MHC similarity between spouses has also been investigated using SNP data. Chaix et al. (2008) analyzed similarity between spouses for 9010 SNPs distributed across the MHC region for 28 couples of European-American origin and 27 couples of African origin from the HapMap 2 data set (Frazer et al., 2007). It turned out that European-American spouses were significantly less similar to each other than were random pairs of individuals, while the similarity in African spouses was not different from that in random pairs. That is, MHC-disassortative mating was detected in the European-American sample but not in the African sample. In addition, when the spousal similarity was examined for more than 3 million SNPs distributed genome-wide, African spouses were significantly more similar to each other than were random pairs, whereas any significant tendency was not observed in European-American spouses. The genome-wide similarity in African spouses – or, more specifically, couples from the Yoruba ethnic group from Nigeria – was interpreted as a result of socio-demographic processes specific to this population rather than individuals' mating preference. More recently, Detri et al. (2010) performed a similar analysis using the HapMap 3 data set (Altshuler et al., 2010), which was not available at the time of the publication of Chaix et al. (2008). Based on their analysis, Detri et al. (2010) argued that the observation by Chaix et al. (2008) of MHC dissimilarity among European-American spouses from the HapMap 2 data set was not robust, and a similar analysis on the HapMap 3 data set did not find such a tendency (for further discussion, see Detri and Roth, 2012; Laurent and Chaix, 2012).

Human Mate Choice Based on Facial Characteristics

Casual observers have frequently noted facial resemblance between spouses, and researchers have shown that this impression is apparently true. In an early experiment by Griffiths and Kunz (1973), participants matched facial photographs of married couples more correctly than expected by chance. Since the result was the same even when photographs were limited to those of couples married less than one year, it was suggested that people tend to marry someone facially similar to them (see also Hinsz, 1987), even if they also grow physically similar after marriage (Zajonc et al., 1987).

More recently, Bereczkei and colleagues have devised a simple procedure to detect facial similarity between individuals (Bereczkei et al., 2002, 2004). Each of their group photos, or 'tableaus,' contained five photographs of different faces, one on the left-hand side and four on the right-hand side. In one kind of tableau, a husband was placed on the left and his wife as well as three control females of similar age on the right. A set of such tableaus was presented to judges, who were instructed to choose one of the four female faces that they thought was most similar to the male face. For their sample of 32 couples, Bereczkei et al. (2002) found that the proportion of judges who chose the photograph of the wife as the most similar to the husband was significantly greater than expected by chance, confirming the findings of earlier studies that spouses tend to be facially similar to each other.

Furthermore, using another kind of tableau, Bereczkei et al. (2002) also examined facial resemblance between a wife and her mother-in-law. This second set of tableaus was motivated by the hypothesis that human mate choice may be influenced by parental phenotypes, learned during the early stage of development through parent–offspring interaction, analogous to the phenomenon known as sexual imprinting in nonhuman animals (ten Cate and Vos, 1999). In a nutshell, they tested whether or not a man tends to be attracted by women who resemble his mother when she was younger. The results of the experiment were as follows: a wife tended to be judged as similar to her husband's mother at the age when the son was between two and eight years old, and the degree of resemblance was even higher between wives and husbands' mothers than between wives and husbands. In addition, according to the husbands' responses to the short form of the EMBU (Swedish acronym for Egna Minnen Beträffande Uppfostran ["My memories of upbringing"]). (Arindell et al., 1999), a self-report scale to quantify one's relationship with the parents during childhood, a wife was more likely to resemble her husband's mother if the husband had been less frequently rejected by the mother during childhood. In sum, these results indicated the existence of a "sexual-imprinting-like" mechanism, by which a man tends to prefer a woman who resembles his mother.

There were two major caveats to the experiment of Bereczkei and colleagues. First, Bereczkei et al. (2002) considered a putative sexual-imprinting-like effect only on male mate choice. Whether or not a woman also prefers a man resembling her father was not investigated. Second, more importantly, the observed resemblance between wives and mothers-in-law might have been created jointly by the wife–husband resemblance owing to a preference for self-resembling partners and the son–mother resemblance due to shared genes, rather than

by maternal influence on the son's mate choice. To deal with these issues, Bereczkei et al. (2004) investigated mate choice of 26 women from adoptive families, in which daughters were not genetically related to their adoptive parents. Using a similar technique as the previous study, they examined facial similarity between husbands and wives, between husbands and wives' adoptive fathers and between husbands and wives' adoptive mothers. As it turned out, husbands were judged as resembling their wives, confirming the previous study. Husbands were also judged as similar to the wives' adoptive fathers (at the age when the daughters were between two and eight years old), while statistically significant resemblance was not found between husbands and the wives' adoptive mothers (when the daughters were between two and eight years old). Recalling that the women were genetically unrelated to the adoptive parents, the observed resemblance between husbands and wives' adoptive fathers could not be a joint result of wife–husband resemblance and daughter–father resemblance. Hence, this result supports the existence of a sexual-imprinting-like phenomenon in humans (see also Little et al., 2003; Wiszewska et al., 2007; but see Nojo et al., 2012; Rantala and Marcinkowska, 2011).

Instead of searching for similarity in the faces of established couples, another line of studies has directly investigated people's preference for self-resembling faces. For this purpose, researchers have used computer graphic techniques to generate same- or opposite-sex facial images resembling experimental subjects in a subtle way so that they will not be conscious of the nature of the images (Penton-Voak et al., 1999). In particular, DeBruine (2004) showed facial images thus generated to 53 male and 55 female participants and asked them to evaluate the attractiveness of the faces. Participants were divided into same-sex, same-phenotypic-category (East Asian, European, or West Asian) groups, and each participant performed pairwise comparisons of faces resembling the members of the same group, including her or himself. When evaluating images of the same sex as the participants, both males and females found self-resembling faces more attractive than other group members judged them. In contrast, when evaluating opposite-sex images, neither males nor females found self-resembling faces significantly more attractive than other group members judged them. In other words, self-resemblance enhanced the attractiveness of same-sex faces but not that of opposite-sex faces.

However, an obvious question is: what does the word 'attractiveness' exactly mean when it is applied to same-sex or opposite-sex faces? More to the point, did DeBruine's participants use the same criteria when evaluating same-sex and opposite-sex faces? Besides, the finding that self-resemblance is irrelevant to the attractiveness of opposite-sex faces does not seem congruent with the observed facial resemblance between spouses. In a follow-up study, DeBruine (2005) approached these issues. In an experimental framework nearly identical to that of the previous study, 66 male and 78 female participants made pairwise comparisons of opposite-sex facial images resembling either themselves or other participants. They were asked to evaluate the faces based on three criteria: trustworthiness, attractiveness for a long-term relationship (e.g., marriage), and attractiveness for a short-term relationship (e.g., a 'one-night stand'). The results of the experiment suggested that self-resemblance enhances trustworthiness but represses attractiveness as a short-term partner of opposite-sex faces. Regarding attractiveness for a long-term relationship, facial resemblance did not have any significant effect.

A possible interpretation of these results is that a preference for self-resemblance is context dependent: people may tend to find, perhaps unconsciously, someone who is facially similar to themselves more desirable in social contexts, whereas they tend to find an opposite-sex individual resembling themselves less desirable in sexual contexts. It has been argued that a preference for self-resemblance in social contexts accords with inclusive fitness theory (Hamilton, 1964), which predicts that animals behave favorably toward close kin (DeBruine, 2002; Krupp et al., 2008). It has also been suggested that a preference against self-resemblance in sexual contexts may be an adaptation to avoid inbreeding (DeBruine et al., 2011). Since 'attractiveness' of an opposite-sex face is likely to involve both social and sexual contexts, it may be perceived on the basis of a balance between the two opposing effects of self-resemblance. This balance is unlikely to be constant across situations, which might partially explain the observed variation in the results of recent experiments (Lass-Hannemann et al., 2010; Kocsor et al., 2011; Nojo et al., 2011; Watkins et al., 2011).

Conclusion

We have reviewed some of the recent research on human mate choice driven by individuals' preference. We first looked at the interaction between genes and mate choice in humans and then explored two examples of physical traits on which human mate choice may depend: the MHC and facial characteristics. A number of researchers have contributed to this field to improve our understanding of the nature of human mate choice. Nevertheless, our current knowledge of the phenomenon is far from sufficient and further investigation is clearly needed.

See also: Behavioral Genetics: Psychological Perspectives; Gene–Culture Coevolution; Genetics of Personality; Human Cooperation, Evolution of; Human Mate Choice, Evolution of; Incest Prohibition, Origin and Evolution of; Incest, Inbreeding, and Their Consequences; Kin Selection; Kinship, Evolution of; Marriage Systems, Evolution of.

Bibliography

Altshuler, D.M., Gibbs, R.A., Peltonen, L., et al., 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467, 52–58.

Apanius, V., Penn, D., Slev, P.R., Ruff, L.R., Potts, W.K., 1997. The nature of selection on the major histocompatibility complex. Critical Reviews in Immunology 17, 179–224.

Arindell, W.A., Sanavio, R., Aguilar, G., et al., 1999. The development of a short form of the EMBU: its appraisal with students in Greece, Guatemala, Hungary and Italy. Personality and Individual Differences 27, 613–628.

Bereczkei, T., Gyuris, P., Koves, P., Bernath, L., 2002. Homogamy, genetic similarity, and imprinting; parental influence on mate choice preferences. Personality and Individual Differences 33, 677–690.

Bereczkei, T., Gyuris, P., Weisfeld, G.E., 2004. Sexual imprinting in human mate choice. Proceedings of the Royal Society of London B 271, 1129–1134.

Buss, D.M., Barnes, M., 1986. Preferences in human mate selection. Journal of Personality and Social Psychology 50, 559–570.

Chaix, R., Cao, C., Donnelly, P., 2008. Is mate choice in humans MHC-dependent? PLoS Genetics 4, e1000184.
DeBruine, L.M., 2002. Facial resemblance enhances trust. Proceedings of the Royal Society of London B 269, 1307–1312.
DeBruine, L.M., 2004. Facial resemblance increases the attractiveness of same-sex faces more than other-sex faces. Proceedings of the Royal Society of London B 271, 2085–2090.
DeBruine, L.M., 2005. Trustworthy but not lust-worthy: context-specific effects of facial resemblance. Proceedings of the Royal Society B 272, 919–922.
DeBruine, L.M., Jones, B.C., Watkins, C.D., et al., 2011. Opposite-sex siblings decrease attraction, but not prosocial attributions, to self-resembling opposite-sex faces. Proceedings of the National Academy of Science USA 108, 11710–11714.
Detri, A., Cenik, C., Kraft, P., Roth, F.P., 2010. Absence of evidence for MHC-dependent mate selection within HapMap populations. PLoS Genetics 6, e1000925.
Detri, A., Roth, F.P., 2012. Response to "MHC-dependent mate choice in humans: why genomic patterns from the HapMap European American data set support the hypothesis". Bioassays 34, 576–577.
Egid, K., Brown, J.L., 1989. The major histocompatibility complex and female mating preferences in mice. Animal Behaviour 38, 548–550.
Frazer, K.A., Ballinger, D.G., Cox, D.R., et al., 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851–861.
Griffiths, R.W., Kunz, P.R., 1973. Assortative mating: a study of physiognomic homogamy. Social Biology 20, 448–453.
Hamilton, W.D., 1964. The genetical evolution of social behaviour. I. Journal of Theoretical Biology 7, 1–16.
Havlicek, J., Roberts, S.C., 2009. MHC-correlated mate choice in humans: a review. Psychoneuroendocrinology 34, 497–512.
Hedrick, P.W., Black, F.L., 1997. HLA and mate selection: no evidence in South Amerindians. American Journal of Human Genetics 61, 505–511.
Hinsz, V.B., 1987. Facial resemblance in engaged and married couples. Journal of Social and Personal Relationships 6, 223–229.
Ihara, Y., Aoki, K., Tokunaga, K., Takahashi, K., Juji, T., 2000. HLA and human mate choice: tests on Japanese couples. Anthropological Science 108, 199–214.
Jacob, S., McClintock, M.K., Zelano, B., Ober, C., 2002. Paternally inherited HLA alleles are associated with women's choice of male odor. Nature Genetics 30, 175–179.
Kocsor, F., Rezneki, R., Juhász, S., Bereczkei, T., 2011. Preference for facial self-resemblance and attractiveness in human mate choice. Archives of Sexual Behavior 40, 1263–1270.
Krupp, D.B., Debruine, L.M., Barclay, P., 2008. A cue of kinship promotes cooperation for the public good. Evolution and Human Behavior 29, 49–55.
Laland, K.N., Odling-Smee, J., Myles, S., 2010. How culture shaped the human genome: bringing genetics and the human sciences together. Nature Reviews Genetics 11, 137–148.
Lass-Hannemann, J., Deuter, C.E., Kuehl, L.K., et al., 2010. Effects of stress on human mating preferences: stressed individuals prefer dissimilar mates. Proceedings of the Royal Society B 277, 2175–2183.
Laurent, R., Chaix, R., 2012. MHC-dependent mate choice in humans: why genomic patterns from the HapMap European American dataset support the hypothesis. Bioassays 34, 267–271.
Laurent, R., Toupance, B., Chaix, R., 2012. Non-random mate choice in humans: insights from a genome scan. Molecular Ecology 21, 587–596.
Little, A.C., Penton-Voak, I.S., Burt, D.M., Perrett, D.I., 2003. Investigating an imprinting-like phenomenon in humans: partners and opposite-sex parents have similar hair and eye colour. Evolution and Human Behavior 24, 43–51.
Lykken, D.T., Tellegen, A., 1993. Is human mating adventitious or the result of lawful choice? A twin study of mate selection. Journal of Personality and Social Psychology 65, 56–68.
Nojo, S., Ihara, Y., Furusawa, H., Akamatsu, S., Ishida, T., 2011. Facial resemblance and attractiveness: an experimental study in rural Indonesia. Letters on Evolutionary Behavioral Science 2, 9–12.
Nojo, S., Tamura, S., Ihara, Y., 2012. Human homogamy in facial characteristics: does a sexual-imprinting-like mechanism play a role? Human Nature 23, 323–340.
Ober, C., Weitkamp, L.R., Cox, N., et al., 1997. HLA and mate choice in humans. American Journal of Human Genetics 61, 497–504.
Penn, D., Potts, W., 1998. How do major histocompatibility complex genes influence odor and mating preferences? Advances in Immunology 69, 411–436.
Penn, D.J., Potts, W.K., 1999. The evolution of mating preferences and major histocompatibility complex genes. American Naturalist 153, 145–164.
Penton-Voak, I.S., Perrett, D.I., Peirce, J.W., 1999. Computer graphic studies of the role of facial similarity in judgements of attractiveness. Current Psychology 18, 104–117.
Phillips, T., Barnard, C., Ferguson, E., Reader, T., 2008. Do humans prefer altruistic mates? Testing a link between sexual selection and altruism towards non-relatives. British Journal of Psychology 99, 555–572.
Phillips, T., Ferguson, E., Rijsdijk, F., 2010. A link between altruism and sexual selection: genetic influence on altruistic behaviour and mate preference towards it. British Journal of Psychology 101, 809–819.
Piertney, S.B., Oliver, M.K., 2006. The evolutionary ecology of the major histocompatibility complex. Heredity 96, 7–21.
Potts, W.K., Manning, C.J., Wakeland, E.K., 1991. Mating patterns in seminatural populations of mice influenced by MHC genotype. Nature 352, 619–621.
Rantala, M.J., Marcinkowska, U.M., 2011. The role of sexual imprinting and the Westermarck effect in mate choice in humans. Behavioral Ecology and Sociobiology 65, 859–873.
Roberts, S.C., Gosling, L.M., Carter, V., Petrie, M., 2008. MHC-correlated odour preferences in humans and the use of oral contraceptives. Proceedings of the Royal Society B 275, 2715–2722.
Rushton, J.P., Bons, T.A., 2005. Mate choice and friendship in twins: evidence for genetic similarity. Psychological Science 16, 555–559.
Santos, P.S.C., Schinemann, J.A., Gabardo, J., Bicalho, M.G., 2005. New evidence that the MHC influences odor perception in humans: a study with 58 Southern Brazilian students. Hormones and Behavior 47, 384–388.
ten Cate, C., Vos, D.R., 1999. Sexual imprinting and evolutionary processes in birds: a reassessment. Advances in the Study of Behavior 28, 1–31.
Thornhill, R., Gangestad, S.W., Miller, R., et al., 2003. Major histocompatibility complex genes, symmetry, and body scent attractiveness in men and women. Behavioral Ecology 14, 668–678.
Watkins, C.D., DeBruine, L.M., Smith, F.G., et al., 2011. Like father, like self: emotional closeness to father predicts women's preferences for self-resemblance in opposite-sex faces. Evolution and Human Behavior 32, 70–75.
Wedekind, C., Füri, S., 1997. Body odour preferences in men and women: do they aim for specific MHC combinations or simply heterozygosity? Proceedings of the Royal Society of London B 264, 1471–1479.
Wedekind, C., Seebeck, T., Bettens, F., Paepke, A.J., 1995. MHC-dependent mate preferences in humans. Proceedings of the Royal Society of London B 260, 245–249.
Wiszewska, A., Pawlowski, B., Boothroyd, L.G., 2007. Father-daughter relationship as a moderator of sexual imprinting: a facialmetric study. Evolution and Human Behavior 28, 248–252.
Yamazaki, K., Boyse, E.A., Mike, V., et al., 1976. Control of mating preferences in mice by genes in the major histocompatibility complex. Journal of Experimental Medicine 144, 1324–1335.
Yamazaki, K., Yamaguchi, M., Baranoski, L., et al., 1979. Recognition among mice: evidence from the use of a Y-maze differently scented by congenic mice of different major histocompatibility types. Journal of Experimental Medicine 150, 755–760.
Zajonc, R.B., Adelmann, P.K., Murphy, S.T., Niedenthal, P.M., 1987. Convergence in the physical appearance of spouses. Motivation and Emotion 11, 335–346.
Zietsch, B.P., Verweij, K.J.H., Burri, A.V., 2012. Heritability of preferences for multiple cues of mate quality in humans. Evolution 66, 1762–1772.
Zietsch, B.P., Verweij, K.J.H., Heath, A.C., Martin, N.G., 2011. Variation in human mate choice: simultaneously investigating heritability, parental influence, sexual imprinting, and assortative mating. American Naturalist 177, 606–616.

Human Mate Choice, Evolution of

Judith A Easton, Texas State University, San Marcos, TX, USA
Cari D Goetz, California State University, San Bernadino, CA, USA
David M Buss, The University of Texas at Austin, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Evolutionary perspectives on human mating have generated a substantial corpus of work that reveals important sex differences in mate choice and preferences. Both men and women engage in multiple mating strategies, from long-term committed relationships to short-term sexual liaisons. This article summarizes how sex-specific adaptive problems faced by our ancestors shaped men's mate choices, including preferences for cues to fertility, sexual availability, faithfulness, and numerous partners; and women's mate choices, including preferences for cues to ambition, dependability, and investment potential. We also examine how these preferences differ in long-term compared to short-term mating contexts. Understanding the evolved nature of these differences is crucial not only in expanding understanding of human mating behaviors, but also in understanding the sexual conflict that results from sex differences in mate choice.

There are robust sex differences in men's and women's preferences for a romantic partner as a result of differences in mating-related adaptive problems recurrently faced by men and women over human evolutionary history. Mate choice also differs as a function of the type of relationship men and women are seeking. Using the lens of sexual strategies theory (Buss and Schmitt, 1993), this article highlights how men's and women's mate choice preferences help them solve sex-specific and temporal-specific mating-related adaptive problems.

Mating Strategies

Humans use a complex set of mating strategies. One of these strategies is long-term mating, which typically involves heavy commitment, pair-bonding, investment by both parents in offspring, and a reported feeling of being in love (e.g., Buss, 2003; Symons, 1979; Trivers, 1972). At the other end of the mating spectrum is short-term mating, which typically involves much less commitment and investment and places a greater emphasis on sexual behavior than long-term mating (e.g., Buss and Schmitt, 1993; Buss, 2003). Men and women can both employ a combination of mating strategies; that is, they can alternate between long-term and short-term mating or use some intermediate mating strategy. They may also use a dual mating strategy of maintaining a long-term relationship while seeking extra-pair copulations (Buss and Schmitt, 1993; Gangestad and Simpson, 2000; Schmitt et al., 2001a; Schmitt, 2005).

The frequency at which men and women employ each strategy differs partly as a function of their minimum obligatory parental investment. At a minimum women must bear the costs of gestation and lactation, whereas at a minimum men need only to contribute sperm and a few moments of their time (Trivers, 1972). The greater minimum obligatory parental investment women bear should cause them to be selective in their mating habits. The high metabolic costs of gestation and lactation and the decreased survival odds for children raised by a single mother (Daly and Wilson, 1988; Geary, 2007; Hewlett, 1988; Hill and Hurtado, 1996; Hrdy, 2007) further suggest women benefit by allocating more effort to a long-term relationship with an investing partner than to many casual relationships. Additionally, once a woman is pregnant she cannot become more pregnant by having sex with additional men, further suggesting the direct reproductive benefits to women of mating with multiple men are minimal. Alternatively, men's minimum parental obligation indicates they can be less selective in their mating habits, at least in short-term or polygynous mating contexts. Inseminating one woman does not necessarily prevent a man from inseminating additional women, suggesting the reproductive benefit to men of mating with multiple women is large (Trivers, 1972). Stated differently, for men a major constraint on their reproductive success historically has been sexual access to fertile women. Men and women historically have faced different mating-related adaptive problems as a result of these and other key differences in their reproductive biology, including internal female fertilization, lactation, and sharper age-graded fertility for women compared to men. These adaptive problems and how men and women solve them will be discussed below.

Men's Mating Strategies and Mate Choice

Short-Term Mating

As a result of men's minimum parental obligation, one effective method for men to increase their reproductive success is to increase the number of their sexual partners (Buss and Schmitt, 1993; Symons, 1979), suggesting they may benefit more from adopting a short-term mating strategy than a long-term strategy. For example, a man who already had one child could double his reproductive success by engaging in an additional short-term mateship resulting in pregnancy and successful childbirth, all else equal. Previous research suggests historically it was not the amount of times a man impregnated one woman that increased his reproductive success but instead it was the number of women he impregnated that mattered (see Betzig, 1988 for review).

Therefore one adaptive problem ancestral men would have to solve in order to increase their reproductive success would be the problem of partner number (Symons, 1979). One way to solve this would be to have sexual desire for a large number of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.81049-1

sexual partners. Over ancestral history those men with this desire would theoretically have had more sexual partners, increasing their reproductive success more than men who did not possess this desire, suggesting men may have evolved a preference for sexual variety (Symons, 1979). In support of this, men report wanting to have a greater number of sexual partners than women across various time periods from the next month to the rest of their lifetime (Buss and Schmitt, 1993; Schmitt et al., 2001b; 2003; Simpson and Gangestad, 1991). Men also report having sexual fantasies about a greater number of sexual partners than do women (Ellis and Symons, 1990). Additionally men can solve the adaptive problem of acquiring multiple sexual partners by relaxing their standards – lower standards increase the chances they will find a partner (Buss and Schmitt, 1993). Kenrick et al. (1990) asked men to report their minimum acceptable threshold for a sexual partner, a date, and a marriage partner. Men reported lower standards for sexual partners than for dates or marriage partners on qualities, such as intelligence (Kenrick et al., 1990). Men also report lower standards for sexual partners or one-night stands than do women (Buss and Schmitt, 1993; Greitemeyer, 2005; Kenrick et al., 1990, 1993; Li and Kenrick, 2006; Regan, 1998a,b; Schieb, 2001: Schützwohl et al., 2009). Along with lowering their standards for a sexual partner, men can increase the numbers of partners they have by decreasing the amount of time they wait to engage in sexual intercourse with a new partner (Buss and Schmitt, 1993). When asked how long they would have to know someone before having sex, men were more willing to have sex after less time elapsed than were women (Buss and Schmitt, 1993; Schmitt et al., 2001c, 2003). Sexual attraction to a partner more quickly develops in men than in women (Clark and Hatfield, 1989; Clark et al., 1999; Hald and Høgh-Olesen, 2010; Rubin et al., 1981) and men frequently misperceive women's friendliness as a signal of sexual interest (Abbey, 1982; Haselton, 2003; Haselton and Buss, 2000; Perilloux et al., in press).

Not only did ancestral men have to solve the problem of partner number, but they also needed to determine which women were available to mate with and which of those women were fertile. Those men who spent time trying to mate with women who were not sexually accessible would produce less offspring than men who selected mates who were sexually accessible. Similarly, men trying to mate with women who were unable to conceive, such as those already pregnant, lactating, prepubescent, or menopausal would have left fewer offspring than men who selected fertile partners (Symons, 1979). This suggests men have evolved mate preference mechanisms that have two primary functions: select women who are available and displaying cues to fertility (Buss and Schmitt, 1993; Singh, 1993; see Buss, 2003 for review). Men value promiscuity and devalue prudishness and sexual inexperience in potential short-term partners, suggesting a preference for sexually accessible women (Buss and Schmitt, 1993; Oliver and Sedikides, 1992; Schmitt et al., 2001a). Additionally, evidence indicates men prefer women who are younger than they are and the age gap in preferences and actual mate selections tends to increase with men's age (Buss, 1989; Buunk et al., 2001; Greenlees and McGrew, 1994; Kenrick and Keefe, 1992). Youth in women is an indication of future reproductive potential, that is the potential for having a larger number of children in the future (Buss, 2003; Symons, 1979).

Men's psychological mechanisms recurrently used estrogen-dependent physiological cues to successfully estimate the age and hence reproductive value of prospective partners. Age-related cues include feminine facial and vocal qualities, smooth and clear skin, full lips, thick full hair, firm and symmetrical breasts, and small waists relative to the size of hips (Feinberg et al., 2006; Fink et al., 2001; Marlowe, 1998; Singh, 1993, 1995, 2002; Thornhill and Gangestad, 2008; Thornhill and Grammar, 1999). Women possessing these traits are perceived to be physically attractive, are highly desired as potential mates by men, and have high levels of estrogen and progesterone, which both increase conception likelihood (Jasienska et al., 2004; Singh, 2002; see Sugiyama, 2005 for review).

To be successful at mating with multiple partners, men also need to avoid getting entangled in a committed relationship and having to provide investment (Buss, 2003; Buss and Schmitt, 1993; Symons, 1979). Mating with a woman who required a large amount of investment or time from a man would decrease his opportunities to mate with other women. One way men can solve this problem is to prefer women who will not require a relationship or investment. Men perceive a commitment with a short-term partner as undesirable in comparison to a commitment with a long-term partner (Buss and Schmitt, 1993). Another way men can solve the problem of avoiding commitment is to experience a negative affective shift after sexual intercourse (Haselton and Buss, 2001). Men who pursue a short-term mating strategy rate first sexual encounters as better than later ones, lose interest in partners after a few days, weeks, or months, and decrease their perceptions of a partner's physical attractiveness after having sexual intercourse with her, and this perception continues to decrease further with time (Haselton and Buss, 2001). Ancestrally men who experienced a negative affective shift after sexual intercourse could more quickly detach from a potential long-term relationship than men who did not experience such a shift, providing those men with more time to acquire multiple mates.

Long-Term Mating

Although these findings suggest ancestrally men could best increase their reproductive success by employing a short-term mating strategy, men could also benefit by employing a long-term mating strategy. Seeking out multiple mates can take time and energy away from actual mating. Men can save that time and energy by staying with one long-term partner (Buss, 2003). Men who engage in short-term mating can never be certain that any offspring produced are theirs because they cannot be certain they have sole reproductive access to any of their partners. To solve this problem of paternity uncertainty, men can maintain a long-term relationship that provides them exclusive reproductive access to a woman (Alexander and Noonan, 1979; Buss and Schmitt, 1993; Thornhill and Gangestad, 2008), particularly if they employ mate guarding tactics to prevent insemination from other men (Buss, 1988a; Buss et al., 1992; Buss and Shackelford, 1997; Daly et al.,

1982; Shackelford et al., 2005a). Additionally, because high-quality women are highly desired by men, they can be choosier and demand more time and investment from men before they engage in sexual intercourse (Buss, 2003; Buss and Shackelford, 2008; Symons, 1979). By being willing to enter a long-term relationship, men may be able to obtain women of higher quality than if they attempt multiple short-term matings (Buss and Schmitt, 1993; Gangestad and Simpson, 1990; Symons, 1979; Thornhill and Gangestad, 2008). Although men might produce fewer offspring by not having multiple partners, they could still increase their reproductive success by producing high-quality offspring who would not only be likely to survive but who eventually would be highly sought after as mates. Additionally, regardless of offspring quality, the survival rate of offspring is increased when investment is provided by both parents (Daly and Wilson, 1988; Geary, 2007; Hewlett, 1988; Hill and Hurtado, 1996; Hrdy, 2007).

Evidence shows that men do engage in long-term mating. When asked how much they are currently seeking a long-term mate, men do not differ from women in their desire to find a long-term mate (Buss and Schmitt, 1993; but see Schmitt et al., 2001b for sex difference). Men place a greater emphasis on faithfulness and chastity in a long-term mate than in a short-term mate (Buss, 1989; Buss and Barnes, 1986; Buss and Schmitt, 1993; Buss et al., 2001). Furthermore, men view promiscuous behavior in potential long-term mates as undesirable (Buss and Schmitt, 1993). A preference for long-term mates possessing these traits helped ancestral men solve the adaptive problem of paternity uncertainty because faithful and chaste women would be less likely to be promiscuous (see Buss, 2003 for review).

Ancestrally, attempting to solely pursue short-term mating would have been a high-risk strategy for men. Given that highly desirable women can successfully demand large investment (Buss and Shackelford, 2008), and that women in general prefer long-term relationships (see below), men who attempted to only engage in short-term mating risked being shut out of the mating game entirely (Buss and Schmitt, 1993; Buss, 2003; Schmitt, 2005). It is more likely ancestral men pursued some combination of serial long-term relationships with occasional opportunistic short-term mating or extra-pair copulations, especially those men high enough in status or with enough resources to cross the polygyny threshold (Andersson, 1994; Daly and Wilson, 1983; Gangestad, 1993; Marlowe, 2000; Schmitt, 2005). This has changed drastically in the modern first-world environment – the anonymity of modern society makes it easier for men to deceive women into brief casual sexual liaisons at low costs to the men (Carroll et al., 1985; Garcia and Reiber, 2008; Paul and Hayes, 2002; Paul et al., 2000; Reiber and Garcia, 2010; Uecker and Regnerus, 2010). This is because the modern environment parasitizes men's evolved mating mechanisms, which gives the impression adopting a short-term mating strategy is universally the most beneficial mating strategy for men (see Jonason et al., 2009 as an example of this fallacy). However, men are not equally successful at short-term mating. Men possessing characteristics deemed highly valuable to women seeking short-term mates (see below) are considerably more successful than men who do not possess those characteristics (Jackson and Kirkpatrick, 2007; Landolt et al., 1995; Li and Kenrick, 2006). These findings indicate that men can rely on a suite of mating strategies to increase their reproductive success, and that men's mate preferences shift as a function of the type of strategy currently being executed.

Of course, these mating strategies are all enacted within different cultural contexts that sometimes constrain and sometimes enable one mating strategy over another (e.g., Gangestad et al., 2006). In cultures that cloister, veil, or guard women, for example, opportunities for short-term mating are severely restricted. In more sexually permissive cultures and in large cities that offer greater anonymity, short-term mating strategies are easier to implement.

Women's Mating Strategies and Mate Choice

Long-Term Mating

Women have a greater minimum obligatory parental investment than men, making them more selective in their mating decisions (Trivers, 1972). In traditional societies, which best resemble those of our ancestors, women spend much of their adult life either pregnant or lactating, and therefore are limited in the number of offspring they can produce (Sugiyama, 2005). Ancestrally, the most effective way for most women to increase their reproductive success was to heavily invest in each offspring produced to insure offspring lived until reproductive age. Paternal investment would have increased the likelihood of offspring survival and future success on the mating market (Geary, 2007; Hewlett, 1988; Hill and Hurtado, 1996; Hrdy, 2007). Therefore, ancestral women who were better at selecting and obtaining mates who were willing and able to invest in their offspring were more likely to leave surviving offspring than women who did not select, or could not obtain, investing partners (Buss, 1989, 2003; Buss and Schmitt, 1993; Schmitt, 2005; Symons, 1979).

Women's mating mechanisms, therefore, must solve the adaptive problem of determining investment ability in potential partners. Ancestrally, women could have ascertained this by a man's potential for accruing resources, which in the modern environment equates with monetary resources (Symons, 1979; Buss, 2003). Ancestral men with the ability to accrue resources were better able to provide tangible investment, such as shelter and food, than could men who lacked this ability. Women who selected men who could accrue resources as partners would have produced more offspring who survived than women who chose other men. In modern environments, women consistently rate qualities synonymous with economic resources, such as good earning capacity, wealth, good financial prospects, and a reliable future career, as more important for a potential long-term mate to possess than do men (e.g., Bereczkei et al., 1997; Buss, 1989; Buss and Barnes, 1986; Buss and Schmitt, 1993; Buss et al., 2001; Greenlees and McGrew, 1994; Kenrick et al., 1993, 1990; Pawlowski and Dunbar, 1999; Sprecher et al., 1994; Waynforth and Dunbar, 1995). Ambition is correlated with the drive and ability to acquire resources, and women prefer long-term partners who are ambitious (e.g., Buss, 1989; Buss and Barnes, 1986; Buss et al., 2001; Kenrick et al., 1993, 1990) and will even prefer

ambitious yet physically unattractive men to nonambitious yet physically attractive men (Sritharan et al., 2010; Townsend and Wasserman, 1998; Wilbur and Campbell, 2010). Another way women can determine a man's potential to acquire economic resources is to evaluate his social status. Men with higher social status tend to accrue greater resources than do men with lower social status. Women prefer long-term mates with high social status (e.g., Buss, 1989; Kenrick et al., 1993, 1990; Li et al., 2002; Rajecki et al., 1991; Townsend and Levy, 1990 a,b). Additionally, women tend to rate men low in these qualities as undesirable as a long-term mate (Buss and Schmitt, 1993).

Just determining a man's investment ability would not have been enough for ancestral women to insure offspring survivability. They also had to solve the adaptive problem of determining if the man would be willing to invest his resources and time specifically in her and any potential offspring. It would not help offspring survival if a man had plentiful resources but spread them among many women or diverted them to another woman's offspring (Buss et al., 1992; Symons, 1979). One marker of this willingness is a man's commitment to the relationship, which can be determined by cues of fidelity and love (Buss, 2003). In one study, women rated acts of resource display, such as 'surprised her with a gift' and 'bought a special present for her' as indicative of a man being in love. Also rated highly were acts indicative of fidelity, such as 'remained faithful when they were separated for more than a month' and 'gave up going out with other women for her' (Buss, 1988b). Similarly men interested in keeping a mate frequently use resource display and commitment-increase retention tactics (Buss, 1988a; Buss and Shackelford, 1997; Shackelford et al., 2005). These behaviors indicate a man is willing to commit his resources and time to one woman and their joint offspring, increasing the chances offspring will receive necessary investment to survive.

The odds that offspring would have survived ancestrally also could have been increased by selecting long-term mates who provided not just resources, but who also displayed characteristics indicative of being a good parent and being willing to help raise offspring. Offspring raised by both biological parents have an increased likelihood of surviving than do offspring raised by one parent (Daly and Wilson, 1988; Geary, 2007; Hewlett, 1988; Hill and Hurtado, 1996; Hrdy, 2007). Cues indicating potential good parenting skills include dependability and emotional stability, kindness and sincerity, and similar personalities, interests and backgrounds (see Buss, 2003 for review). Women consistently rate these characteristics as important for potential long-term mates to possess. Kindness, dependability, and emotional stability often rank at the top of women's list of required preferences for a long-term partner (Buss, 1989; Buss and Barnes, 1986; Buss et al., 2001; Evans and Brase, 2007; Fletcher et al., 2004; Kenrick et al., 1993, 1990; Li et al., 2002; Regan, 1988a; Scheib, 2001). Women can assess a man's ability to be a good parent by observing his interactions with children. Interacting with children increases women's perceptions of a man's attractiveness in comparison to perceptions of his attractiveness when he is alone (Brase, 2006; La Cerra, 1995). Additionally, women appear to be able to detect men's interest in infants simply by viewing images of men's faces. Women rate men expressing greater interest in infants as more attractive potential long-term mates than men who do not show such interest (Roney et al., 2006). Women whose psychological mechanisms led them to select men possessing these traits and with an interest in children would have had more help raising offspring, increasing offspring survivability, and ultimately the women's own reproductive success, than would women whose psychological mechanisms led them to select other men as partners. Although long-term mating is extremely beneficial to women, there are contexts in which women can also benefit from engaging in short-term mating.

Short-Term Mating

Pregnancy and lactation require large amounts of investment and resources but parental investment can have diminishing returns (Cashdan, 1993; Low, 1990). Therefore if women can obtain their own resources, they do not necessarily stand to benefit from obtaining *more* resources from a potential mate. Ancestrally women may have obtained their own resources through their family and could therefore afford to trade-off requiring investment from potential partners for other noninvestment qualities (Gangestad, 1993). In these cases, women would actually benefit by shifting toward a more short-term mating strategy (Gangestad and Simpson, 2000) because men of high presumptive genetic quality often choose to focus more effort on short-term mateships rather than long-term mateships (Apicella and Marlowe, 2007; Buss, 2003; Marlowe, 1999; in birds: Møller and Thornhill, 1998; Rhodes et al., 2005; Schmitt, 2005; Thornhill and Gangestad, 2008). Ancestrally women could have also obtained resources by having a long-term committed partner. When this occurred, women may have chosen to risk the certain investment from their long-term partner by engaging in extra-pair copulations with men of high presumptive genetic quality (Buss, 2003; Gangestad and Simpson, 2000; Greiling and Buss, 2000; Thornhill and Gangestad, 2008). What this suggests is women may have evolved adaptations for employing either a short-term mating strategy or a dual mating strategy, but these adaptations evolved to solve different mating-related adaptive problems than did men's (Buss, 2003; Buss and Schmitt, 1993; Gangestad and Simpson, 2000; Greiling and Buss, 2000; Schmitt et al., 2001a; Schmitt, 2005; Thornhill and Gangestad, 2008).

Unlike for men, there is no direct or linear increase to a woman's reproductive success by having sexual intercourse with multiple partners. Instead, women can increase their reproductive success, in principle, by having sexual intercourse with men possessing 'good genes' (e.g., Gangestad and Simpson, 2000; Greiling and Buss, 2000; Simpson and Gangestad, 1991; Thornhill and Gangestad, 2008; Trivers, 1972). Good genes are currently thought to be those indicative of good health and increased chances of survivability (see Andersson, 1994 for review). Mating with a man possessing good genes therefore increases chances for offspring survival. Additionally, because women who employ a short-term mating strategy tend to prefer men who advertise their good genes, mating with a man who possesses good genes should increase the chances a woman's male offspring will be able to find a mate and produce offspring, thereby increasing her

reproductive success ('sexy sons' hypothesis; Buss, 2003; Fisher, 1930; Kokko et al., 2002; Smith, 1984; Thornhill and Gangestad, 2008).

If women recurrently engaged in short-term mating to obtain partners with 'good genes,' then their psychological mechanisms should have evolved to solve the adaptive problem of detecting genetic quality of potential partners (Thornhill and Gangestad, 2008). Ancestral women who selected partners with good genetic quality would have been more likely to produce offspring with increased survival odds than women who selected partners with decreased genetic quality.

Women obviously cannot detect underlying genetic quality in a potential mate, but instead can use statistically correlated phenotypic markers of good genes. One such set of phenotypic markers are those indicating masculinity (Frederick and Haselton, 2007; Penton-Voak et al., 1999; Thornhill and Gangestad, 1993, 1999). Masculinity is positively correlated with body symmetry, and a high degree of body symmetry indicates low amounts of developmental instability. Developmental instability is a measure of ability to overcome exposure to prenatal pathogens and mutations. Therefore, men with low levels of developmental instability can be considered to have higher fitness and better genes than men with greater amounts of developmental instability (Gangestad and Simpson, 2000; Gangestad and Thornhill, 1997; Gangestad et al., 1994; Thornhill and Gangestad, 1999, 2008). Women interested in short-term mating rate men displaying characteristics perceived as masculine as more attractive than those appearing more feminine (Gangestad, 1993; Johnston et al., 2001; Penton-Voak et al., 2003, 1999) and prefer to short-term mate with men considered attractive (Buss and Schmitt, 1993; Greiling and Buss, 2000; Kenrick et al., 1993; Li and Kenrick, 2006; Rhodes et al., 2005; Simpson and Gangestad, 1992; Wiederman and Dubois, 1998; Wilbur and Campbell, 2010). These findings suggest women have evolved psychological mechanisms functioning to select short-term partners with high genetic quality.

Women may not engage in short-term mating solely to obtain good genes for their offspring. Instead, given the great importance of long-term relationships to women, they may engage in short-term mating to assess a man as a potential long-term partner (Buss and Schmitt, 1993; Buss, 2003; Greiling and Buss, 2000; Shackelford et al., 2004). Choosing a poor long-term partner would have been extremely costly to ancestral women, especially in exogamous societies, which would have left women without physically close familial support (Buss, 2003; Buss and Schmitt, 1993). Women who used short-term mating to assess potential partners for long-term relationships, and were able to correctly determine which ones would make a good long-term partner, would have had greater reproductive success than those women who could not make such determinations (Buss, 2003; Buss and Schmitt, 1993). Women should, therefore, have evolved psychological mechanisms to solve the adaptive problem of partner assessment.

Women can solve the problem of partner assessment by selecting short-term partners with similar characteristics to highly preferred long-term partners (Buss and Schmitt, 1993). When considering potential short-term partners, women consider a man's financial resources and generosity important (Wiederman and Dubois, 1998) – two traits women seeking long-term partners also consider important (Buss, 1989; Buss and Barnes, 1986; Buss et al., 2001; Kenrick et al., 1993, 1990; Li et al., 2002; Scheib, 2001). In general, women's preferences for potential short-term partners are as stringent as their preferences for a long-term mate, suggesting they may use short-term mating as an assessment strategy (Buss and Schmitt, 1993; Li and Kenrick, 2006; Kenrick et al., 1993, 1990; Regan, 1998a,b; Shackelford et al., 2004; Sprecher and Regan, 2002; Surbey and Conohan, 2000). The assessment problem can also be solved by specifically decreasing preferences for traits in short-term partners that are detrimental to quality if present in a long-term partner. For example, women should not want a promiscuous long-term partner due to the risk of costly investment loss, and if using short-term mating as an assessment tool, should also express a decreased interested in promiscuous men as short-term partners (Buss, 2003; Buss and Schmitt, 1993; Schmitt et al., 2001b). Women seeking a short-term partner rate promiscuity as undesirable in their potential mates (Buss and Schmitt, 1993; Schmitt et al., 2001b; Wiederman and Dubois, 1998). Additionally if women use short-term mating as an assessment strategy for a long-term relationship, they should experience an affective shift after having sexual intercourse with men who would be good long-term partners to promote the formation of a long-term pair bond (Haselton and Buss, 2001). These findings suggest women do engage in short-term mating to assess men as potential long-term partners, although men too might use short-term mating to gauge long-term quality.

Women may also engage in short-term mating to acquire immediate resources instead of the promise of future resource investment they seek in long-term partners. Ancestral women may have needed immediate resources, such as scarce game meat to help provision offspring, and could have acquired such highly valued resources by exchanging them for sex (Buss, 2003). For example, better hunters in the Aché (a traditional hunter-gatherer society) have more mistresses and those mistresses report having more illegitimate children by the more skilled hunters (Hill and Hurtado, 1996; Hill and Kaplan, 1988). If acquiring resources sooner rather than later helped women survive or produce more viable offspring, then women should have evolved psychological mechanisms for preferring as short-term partners men who are willing to provide immediate resources (Buss, 2003; Buss and Schmitt, 1993; Symons, 1979). Buss and Schmitt (1993) asked women to rate four behaviors related to early resource provisioning: 'spends a lot of money on me early on,' 'gives me gifts early on,' 'has an extravagant lifestyle,' and 'stingy early on.' Women valued the first three behaviors for potential short-term partners more than for long-term partners, but devalued the last behavior for potential short-term partners (Buss and Schmitt, 1993). Other research has corroborated these findings – when seeking a short-term partner women place a premium on immediate resource investment (Bereczkei et al., 1997; Fletcher et al., 2004; Kenrick et al., 1990, 1993; Regan, 1998 a,b; Townsend and Levy, 1990a,b). The strong demonstrated preference women seeking short-term partners have for immediate investment

suggests women do use short-term mating as an opportunity to obtain immediate material benefits for themselves or any current offspring.

Although long-term mating is often a more beneficial strategy to women's reproductive success than short-term mating, evidence suggests women do at times engage in short-term mating. Statistically speaking, every time a man has sex with a new partner, a woman is also having sex with a new partner. This suggests the mean number of sex partners for men and women must be identical, assuming an equal sex ratio in the mating pool (Buss, 2003; Greiling and Buss, 2000). Similarly, men's strong desire for sexual variety could not have evolved if ancestral men never obtained access to multiple sexual partners (Schmitt et al., 2001b; Schmitt et al., 2001c; Schmitt et al., 2003; Smith, 1984), suggesting that at least some ancestral women were willing to engage in brief sexual encounters with multiple men. In recent studies on hook-up behavior, between 50 and 75% of college women report engaging in at least one casual sexual encounter within the last year, and these numbers closely parallel college men's reports of their own casual sex behavior (Garcia and Reiber, 2008; Grello et al., 2006; Paul and Hayes, 2002; Paul et al., 2000; Reiber and Garcia, 2010; Uecker and Regnerus, 2010).

Additional support for women having an evolved short-term mating psychology comes from men's physiology. As a function of the risk of sperm competition, species' testes size is correlated with the species' specific mating strategies. Sperm competition occurs when the sperm of two or more males occupy a female's reproductive tract simultaneously (Baker and Bellis, 1993 a,b; Parker, 1970; Smith, 1984). More monogamous primate species face less risk of sperm competition and therefore have smaller testes relative to body size than do polygamous primate species. For example, male gorillas have sole reproductive access to a harem of female gorillas and their testes size is less than 0.1% of their total body weight, whereas male chimpanzees compete for access to fertile females and their testes are about 0.3% of their weight. Human male testes size falls in the middle of these two species, suggesting humans are not entirely monogamous such as the gorillas, or entirely promiscuous such as the chimpanzees (see Pound et al., 2006; Shackelford et al., 2005b, c for review). If human females did not engage in short-term mating, there would be no selection pressure on males' physiology to cope with the threat of sperm competition, and human testes would be proportionally similar to the testes of gorillas. However, because human testes are not proportionally similar to gorilla testes, it is reasonable to infer that across evolutionary history, human males did have to solve the adaptive problem of sperm competition some of the time, providing evidence for females occasionally employing some form of a short-term mating strategy. As with men, it is clear women rely on a suite of strategies and their mate preferences shift accordingly to match the current strategy being executed.

Conclusion

As a result of differences in minimum obligatory parental investment, men and women differ in how they can best increase their reproductive success. When possible, men's reproductive success historically was best be increased through low-cost opportunistic mating or polygynous mating, whereas women's reproductive success is best increased through high-quality long-term investment mating. Although men and women both employ a variety of mating strategies, their preferences when pursuing these strategies fundamentally differ. One consequence of these sex differences is sexual conflict, jealousy, sexual aggression, sexual harassment, and rape are often products of differences in men's and women's mate preferences and desires. Understanding the evolved nature of men's and women's mate choice contributes to our understanding of these darker elements of human mating in addition to our broader understanding of mating psychology.

See also: Fertility Theory: Embodied-Capital Theory of Life History Evolution; Marriage Systems, Evolution of; Sociobiology: Overview.

Bibliography

Abbey, A., 1982. Sex differences in attributions for friendly behavior: do males misperceive females' friendliness? Journal of Personality and Social Psychology 32, 830–838.

Alexander, R.D., Noonan, K.M., 1979. Concealment of ovulation, parental care, and human social evolution. In: Chagnon, N.A., Irons, W. (Eds.), Evolutionary Biology and Human Social Behavior: An Anthropological Perspective. Duxbury Press, North Scituate, MA, pp. 402–435.

Andersson, M., 1994. Sexual Selection. Princeton University Press, Princeton, NJ.

Apicella, C.L., Marlowe, F.W., 2007. Men's reproductive investment decisions: mating, parenting, and self-perceived mate value. Human Nature 18, 22–34.

Baker, R.R., Bellis, M.A., 1993a. Human sperm competition: ejaculate adjustment by males and the function of masturbation. Animal Behaviour 46, 861–885.

Baker, R.R., Bellis, M.A., 1993b. Human sperm competition: ejaculate manipulation by females and a function for the female orgasm. Animal Behaviour 46, 887–909.

Bereczkei, T., Voros, S., Gal, A., Bernath, L., 1997. Resources, attractiveness, family commitment: reproductive decisions in human mate choice. Ethology 103, 681–699.

Betzig, L., 1988. Mating and parenting in Darwinian perspective. In: Betzig, L., Borgerhoff-Mulder, M., Turke, P. (Eds.), Human Reproductive Behavior: A Darwinian Perspective. Cambridge University Press, Cambridge, pp. 3–20.

Brase, G.L., 2006. Cues of parental investment as a factor in attractiveness. Evolution and Human Behavior 27, 145–157.

Buss, D.M., 1988a. From vigilance to violence: tactics of mate retention in American undergraduates. Ethology & Sociobiology 9, 291–317.

Buss, D.M., 1988b. Love acts: the evolutionary biology of love. In: Sternberg, R.J., Barnes, M.L. (Eds.), The Psychology of Love. Yale University Press, New Haven, CT, pp. 100–118.

Buss, D.M., 1989. Sex differences in human mate preferences: evolutionary hypotheses testing in 37 cultures. Behavioral and Brain Sciences 12, 1–49.

Buss, D.M., 2003. The Evolution of Desire: Strategies of Human Mating, rev. ed. Free Press, New York.

Buss, D.M., Barnes, M.F., 1986. Preferences in human mate selection. Journal of Personality and Social Psychology 50, 559–570.

Buss, D.M., Larsen, R.J., Westen, D., Semmelroth, J., 1992. Sex differences in jealousy: evolution, physiology, and psychology. Psychological Science 3, 251–255.

Buss, D.M., Schmitt, D.P., 1993. Sexual strategies theory: an evolutionary perspective on human mating. Psychological Review 100, 204–232.

Buss, D.M., Shackelford, T.K., 1997. From vigilance to violence: mate retention tactics in married couples. Journal of Personality and Social Psychology 72, 346–361.

Buss, D.M., Shackelford, T.K., 2008. Attractive women want it all: good genes, economic investment, parenting proclivities, and emotional commitment. Evolutionary Psychology 6, 134–146.

Buss, D.M., Shackelford, T.K., Kirkpatrick, L.A., Larsen, R.J., 2001. A half century of American mate preferences: the cultural evolution of values. Journal of Marriage and the Family 63, 491–503.

Buunk, B.P., Dijkstra, P., Kenrick, D.T., Warntjes, A., 2001. Age preferences for mates related to gender, own age, and involvement level. Evolution and Human Behavior 22, 241–250.

Carroll, J.L., Volk, K.D., Hyde, J.S., 1985. Differences between males and females in motives for engaging in sexual intercourse. Archives of Sexual Behavior 14, 131–139.

Cashdan, E., 1993. Attracting males: effects of parental investment on mate attraction strategies. Ethology and Sociobiology 14, 1–24.

Clark, R.D., Hatfield, E., 1989. Gender differences in receptivity to sexual offers. Journal of Psychology and Human Sexuality 2, 39–55.

Clark, C.L., Shaver, P.R., Abrahams, M.F., 1999. Strategic behaviors in romantic relationship initiation. Personality and Social Psychology Bulletin 25, 707–720.

Daly, M., Wilson, M., 1983. Sex, Evolution, and Behavior, second ed. Willard Grant Press, Boston, MA.

Daly, M., Wilson, M., 1988. Homicide. Aldine de Gruyter, Hawthorne, NY.

Daly, M., Wilson, M., Weghorst, S.J., 1982. Male sexual jealousy. Ethology and Sociobiology 3, 11–27.

Ellis, B.J., Symons, D., 1990. Sex differences in sexual fantasy: an evolutionary psychological approach. The Journal of Sex Research 27, 527–555.

Evans, K., Brase, G.L., 2007. Assessing sex differences and similarities in mate preferences: above and beyond demand characteristics. Journal of Social and Personal Relationships 24, 781–791.

Feinberg, D.R., Jones, B.C., Law Smith, M.J., Moore, F.R., DeBruine, L.M., Cornwell, R.E., et al., 2006. Menstrual cycle, trait estrogen level, and masculinity preferences in the human voice. Hormones and Behavior 49, 215–222.

Fletcher, G.J.O., Tither, J.M., O'Loughlin, C., Friesen, M., Overall, N., 2004. Warm and homely or cold and beautiful? Sex differences in trading off traits in mate selection. Personality and Social Psychology Bulletin 30, 659–672.

Fink, B., Grammar, K., Thornhill, R., 2001. Human (*Homo sapiens*) facial attractiveness in relation to skin texture and color. Journal of Comparative Psychology 115, 92–99.

Fisher, R.A., 1930. The Genetical Theory of Natural Selection, third ed. Oxford University Press, USA, New York.

Frederick, D.A., Haselton, M.G., 2007. Why is muscularity sexy? Tests of the fitness indicator hypothesis. Personality and Social Psychology Bulletin 33, 1167–1183.

Gangestad, S.W., 1993. Sexual selection and physical attractiveness: Implication for mating dynamics. Human Nature 4, 205–236.

Gangestad, S.W., Haselton, M.G., Buss, D.M., 2006. Evolutionary foundations of cultural variation: evoked culture and mate preferences. Psychological Inquiry 17, 75–95.

Gangestad, S.W., Simpson, J.A., 1990. Toward an evolutionary history of female sociosexual variation. Journal of Personality 58, 69–96.

Gangestad, S.W., Simpson, J.A., 2000. The evolution of human mating: the role of trade- offs and strategic pluralism. Behavioral and Brain Sciences 23, 675–687.

Gangestad, S.W., Thornhill, R., 1997. The evolutionary psychology of extra-pair sex: the role of fluctuating asymmetry. Evolution and Human Behavior 18, 69–88.

Gangestad, S.W., Thornhill, R., Yeo, R., 1994. Facial attractiveness, developmental stability, and fluctuating asymmetry. Ethology and Sociobiology 15, 73–85.

Garcia, J.R., Reiber, C., 2008. Hook-up behavior: a biopsychosocial perspective. Journal of Social, Evolutionary, and Cultural Psychology 2, 192–208.

Geary, D.C., 2007. Evolution of fatherhood. In: Salmon, C.A., Shackelford, T.K. (Eds.), Family Relationships: An Evolutionary Perspective. Oxford University Press, USA, New York, pp. 115–144.

Greenlees, I.A., McGrew, W.C., 1994. Sex and age differences in preferences and tactics of mate attraction: analysis of published advertisements. Ethology and Sociobiology 15, 59–72.

Greiling, H., Buss, D.M., 2000. Women's sexual strategies: the hidden dimension of short-term mating. Personality and Individual Differences 28, 929–963.

Grello, C.M., Welsh, D.P., Harper, M.S., 2006. No strings attached: the nature of casual sex in college students. The Journal of Sex Research 43, 255–267.

Greitemeyer, T., 2005. Receptivity to sexual offers as a function of sex, socioeconomic status, physical attractiveness, and intimacy of the offer. Personal Relationships 12, 373–386.

Hald, G.M., Høgh-Olesen, H., 2010. Receptivity to sexual invitations from strangers of the opposite sex. Evolution and Human Behavior 31, 453–458.

Haselton, M.G., 2003. The sexual overperception bias: evidence of a systematic bias in men from a survey of naturally occurring events. Journal of Research in Personality 37, 34–47.

Haselton, M.G., Buss, D.M., 2000. Error management theory: a new perspective on biases in cross-sex mind reading. Journal of Personality and Social Psychology 78, 81–91.

Haselton, M.G., Buss, D.M., 2001. The affective shift hypothesis: the functions of emotional changes following sexual intercourse. Personal Relationships 8, 357–369.

Hewlett, B.S., 1988. Sexual selection and paternal investment among Aka pygmies. In: Betzig, L., Borgerhoff-Mulder, M., Turke, P. (Eds.), Human Reproductive Behavior: A Darwinian Perspective. Cambridge University Press, Cambridge, pp. 263–276.

Hill, K., Hurtado, A., 1996. Ache Life History: The Ecology and Demography of a Foraging People. Aldine de Gruyter, Hawthorne, NY.

Hill, K., Kaplan, H., 1988. Tradeoffs in male and female reproductive strategies among the ache: Part 1. In: Betzig, L., Borgerhoff-Mulder, M., Turke, P. (Eds.), Human Reproductive Behavior: A Darwinian Perspective. Cambridge University Press, Cambridge, pp. 277–289.

Hrdy, S.B., 2007. Evolutionary context of human development: the cooperative breeding model. In: Salmon, C.A., Shackelford, T.K. (Eds.), Family Relationships: An Evolutionary Perspective. Oxford University Press, USA, New York, pp. 39–68.

Jackson, J.J., Kirkpatrick, L.A., 2007. The structure and measurement of human mating strategies: toward a multidimensional model of sociosexuality. Evolution and Human Behavior 28, 382–391.

Jasienska, G., Ziomkiewicz, A., Ellison, P.T., Lipson, S.F., Thune, I., 2004. Large breasts and narrow waists indicate high reproductive potential in women. Proceedings of the Royal Society of London B 271, 1213–1217.

Johnston, V.S., Hagel, R., Franklin, M., Fink, B., Grammer, K., 2001. Male facial attractiveness: evidence for hormone-mediated adaptive design. Evolution and Human Behavior 22, 251–267.

Jonason, P.K., Li, N.P., Cason, M.J., 2009. The "booty call": a compromise between men's and women's ideal mating strategies. Journal of Sex Research 46, 1–11.

Kenrick, D.T., Groth, G.E., Trost, M.R., Sadalla, E.K., 1993. Integrating evolutionary and social exchange perspectives on relationships: effects of gender, self-appraisal, and involvement level on mate selection criteria. Journal of Personality and Social Psychology 64, 951–969.

Kenrick, D.T., Keefe, R.C., 1992. Age preferences in mates reflect sex differences in reproductive strategies. Behavioral and Brain Sciences 15, 75–133.

Kenrick, D.T., Sadalla, E.K., Groth, G.E., Trost, M.R., 1990. Evolution, traits, and the stages of human courtship: qualifying the parental investment model. Journal of Personality 58, 97–116.

Kokko, H., Brooks, R., McNamara, J.M., Houston, A.I., 2002. The sexual selection continuum. Proceedings of the Royal Society of London B 269, 1331–1340.

La Cerra, M.M. (1995). Evolved mate preferences in women: Psychological adaptations for assessing a man's willingness to invest in offspring (Doctoral dissertation). University of California, Santa Barbara (Dissertation Abstracts International: Section B: the Sciences & Engineering March 1995, vol. 55 (9-B), p. 4149).

Landolt, M., Lalumière, M., Quinsey, V., 1995. Sex differences in intra-sex variations in human mating tactics: an evolutionary approach. Ethology and Sociobiology 16, 3–23.

Li, N.P., Bailey, J.M., Kenrick, D.T., Linsenmeier, J.A.W., 2002. The necessities and luxuries of mate preferences: testing the tradeoffs. Journal of Personality and Social Psychology 82, 947–955.

Li, N.P., Kenrick, D.T., 2006. Sex similarities and differences in preferences for short-term mates: what, whether, and why. Journal of Personality and Social Psychology 90, 468–489.

Low, B.S., 1990. Sex, power and resources: male and female strategies of resource acquisition. International Journal of Contemporary Sociology 27, 49–73.

Marlowe, F., 1998. The nubility hypothesis: the human breast as an honest signal of residual reproductive value. Human Nature 9, 263–271.

Marlowe, F., 1999. Male care and mating effort among Hadza foragers. Behavioral Ecology and Sociobiology 46, 57–64.

Marlowe, F., 2000. Paternal investment and the human mating system. Behavioural Processes 51, 45–61.

Møller, A.P., Thornhill, R., 1998. Male parental care, differential parental investment by females, and sexual selection. Animal Behaviour 55, 1507–1515.

Oliver, M.B., Sedikides, C., 1992. Effects of sexual permissiveness on desirability of partner as a function of low and high commitment to relationship. Social Psychology Quarterly 55, 321–333.

Parker, G.A., 1970. Sperm competition and its evolutionary consequences in the insects. Biological Reviews 45, 525–567.

Paul, E.L., Hayes, K.A., 2002. The casualties of 'casual' sex: a qualitative exploration of the phenomenology of college students' hookups. Journal of Social and Personal Relationships 19, 639–661.

Paul, E.L., McManus, B., Hayes, K.A., 2000. "Hookups": characteristics and correlates of college students' spontaneous and anonymous sexual experiences. The Journal of Sex Research 37, 76–88.

Pawlowski, B., Dunbar, R.I., 1999. Impact of market value on human mate choice decisions. Proceedings of the Royal Society of London B 266, 281–285.

Penton-Voak, I.S., Little, A.C., Jones, B.C., Burt, D.M., Tiddeman, B.P., Perrett, D.I., 2003. Female condition influences preferences for sexual dimorphism in faces of male humans (*Homo sapiens*). Journal of Comparative Psychology 117, 264–271.
Penton-Voak, I.S., Perrett, D.I., Castles, D.L., Kobayashi, T., Burt, D.M., Murray, L., et al., 1999. Menstrual cycle alters face preference. Nature 399, 741–742.
Perilloux, C., Easton, J.A., Buss, D.M. The misperception of sexual interest: Who misperceives and who is misperceived? Psychological Science, in press.
Pound, N., Shackelford, T.K., Goetz, A.T., 2006. Sperm competition in humans. In: Shackelford, T.K., Pound, N. (Eds.), Sperm Competition in Humans: Classic and Contemporary Readings. Springer, New York, pp. 3–31.
Rajecki, D.W., Bledsoe, S.B., Rasmussen, J.L., 1991. Successful personal ads: gender differences and similarities in offers, stipulations, and outcomes. Basic and Applied Social Psychology 12, 457–469.
Regan, P.C., 1998a. Minimum mate selection standards as a function of perceived mate value, relationship context, and gender. Journal of Psychology and Human Sexuality 10, 53–73.
Regan, P.C., 1998b. What if you can't get what you want? Willingness to compromise ideal mate selection standards as a function of sex, mate value, and relationship context. Personality and Social Psychology Bulletin 24, 1294–1303.
Reiber, C., Garcia, J.R., 2010. Hooking up: gender differences, evolution, and pluralistic ignorance. Evolutionary Psychology 8, 390–404.
Rhodes, G., Simmons, L.W., Peters, M., 2005. Attractiveness and sexual behavior: does attractiveness enhance mating success? Evolution and Human Behavior 26, 186–201.
Roney, J.R., Hanson, K.N., Durante, K.M., Maestripieri, D., 2006. Reading men's faces: women's mate attractiveness judgments track men's testosterone and interest in infants. Proceedings of the Royal Society B 273, 2169–2175.
Rubin, Z., Peplau, L.A., Hill, C.T., 1981. Loving and leaving: sex differences in romantic attachments. Sex Roles 1, 821–836.
Scheib, J.E., 2001. Context-specific mate choice criteria: women's trade-offs in the contexts of long-term and extra-pair mateships. Personal Relationships 8, 371–389.
Schmitt, D.P., 2005. Fundamentals of human mating strategies. In: Buss, D.M. (Ed.), The Handbook of Evolutionary Psychology. Wiley, New York, pp. 258–291.
Schmitt, D.P., 118 Members of International Sexuality Description Project, 2003. Universal sex differences in the desire for sexual variety: tests from 52 nations, 6 continents, and 13 islands. Journal of Personality and Social Psychology 85, 85–104.
Schmitt, D.P., Couden, A., Baker, M., 2001a. The effects of sex and temporal context on feelings of romantic desire: an experimental evaluation of Sexual Strategies Theory. Personality and Social Psychology Bulletin 27, 833–847.
Schmitt, D.P., Shackelford, T.K., Buss, D.M., 2001b. Are men really more 'oriented' toward short-term mating than women? A critical review of theory and research. Psychology, Evolution & Gender 3, 211–239.
Schmitt, D.P., Shackelford, T.K., Duntley, J., Tooke, W., Buss, D.M., 2001c. The desire for sexual variety as a tool for understanding basic human mating strategies. Personal Relationships 8, 425–455.
Schützwohl, A., Fuchs, A., McKibbin, W.F., Shackelford, T.K., 2009. How willing are you to accept sexual requests from slightly unattractive to exceptionally attractive imagined requestors? Human Nature 20, 282–293.
Shackelford, T.K., Goetz, A.T., Buss, D.M., 2005a. Mate retention in marriage: further evidence of the reliability of the mate retention inventory. Personality and Individual Differences 39, 415–425.
Shackelford, T.K., Goetz, A.T., LaMunyon, C.W., Quintus, B.J., Weekes-Shackelford, V.A., 2004. Sex differences in sexual psychology produce sex similar preferences for a short-term mate. Archives of Sexual Behavior 33, 405–412.
Shackelford, T.K., Pound, N., Goetz, A.T., 2005b. Psychological and physiological adaptations to sperm competition in humans. Review of General Psychology 9, 228–248.
Shackelford, T.K., Pound, N., Goetz, A.T., LaMunyon, C.W., 2005c. Female infidelity and sperm competition. In: Buss, D.M. (Ed.), The Handbook of Evolutionary Psychology. Wiley, New York, pp. 372–393.
Simpson, J.A., Gangestad, S.W., 1991. Individual differences in sociosexuality: evidence for convergent and discriminant validity. Journal of Personality and Social Psychology 60, 870–883.
Simpson, J.A., Gangestad, S.W., 1992. Sociosexuality and romantic partner choice. Journal of Personality 60, 31–51.
Singh, D., 1993. Adaptive significance of waist-to-hip ratio and female physical attractiveness. Journal of Personality and Social Psychology 65, 293–307.
Singh, D., 1995. Female health, attractiveness and desirability for relationships: role of breast asymmetry and waist-to-hip ratio. Ethology and Sociobiology 16, 465–481.
Singh, D., 2002. Female mate value at a glance: relationship of waist-to-hip ratio to health, fecundity and attractiveness. Neuroendocrinology Letters 23, 81–91.
Smith, R.L., 1984. Human sperm competition. In: Smith, R.L. (Ed.), Sperm Competition and the Evolution of Mating Systems. Academic Press, New York, pp. 601–659.
Sprecher, S., Regan, P.C., 2002. Liking some things (in some people) more than others: Partner preferences in romantic relationships and friendships. Journal of Social and Personal Relationships 19, 463–481.
Sprecher, S., Sullivan, Q., Hatfield, E., 1994. Mate selection preferences: gender differences examined in a national sample. Journal of Personality and Social Psychology 66, 1074–1080.
Sritharan, R., Heilpern, K., Wilbur, C.J., Gawronski, B., 2010. I think I like you: spontaneous and deliberate evaluations of potential romantic partners in an online dating context. European Journal of Social Psychology 40, 1062–1077.
Sugiyama, L.S., 2005. Physical attractiveness in adaptationist perspective. In: Buss, D.M. (Ed.), The Handbook of Evolutionary Psychology. Wiley, New York, pp. 292–343.
Surbey, M.K., Conohan, C.D., 2000. Willingness to engage in casual sex: the role of parental qualities and perceived risk of aggression. Human Nature 11, 367–386.
Symons, D., 1979. The Evolution of Human Sexuality. Oxford University Press, New York.
Thornhill, R., Gangestad, S.W., 1993. Human facial beauty: averageness, symmetry and parasite resistance. Human Nature 4, 237–269.
Thornhill, R., Gangestad, S.W., 1999. Facial attractiveness. Trends in Cognitive Science 3, 452–460.
Thornhill, R., Gangestad, S.W., 2008. The Evolutionary Biology of Human Female Sexuality. Oxford University Press, New York.
Thornhill, R., Grammar, K., 1999. The body and face of a woman: one ornament that signals quality? Evolution and Human Behavior 20, 105–120.
Townsend, J.M., Levy, G.D., 1990a. Effects of potential partners' physical attractiveness and socioeconomic status on sexuality and partner selection. Archives of Sexual Behavior 19, 149–164.
Townsend, J.M., Levy, G.D., 1990b. Effect of potential partners' costume and physical attractiveness on sexuality and partner selection. Journal of Psychology: Interdisciplinary & Applied 124, 371–389.
Townsend, J.M., Wasserman, T., 1998. Sexual attractiveness: sex differences in assessment and criteria. Evolution and Human Behavior 19, 171–191.
Trivers, R., 1972. Parental investment and sexual selection. In: Campbell, B. (Ed.), Sexual Selection and the Descent of Man. Aldine de Gruyter, New York, pp. 136–179.
Uecker, J.E., Regnerus, M.D., 2010. Bare market: campus sex ratios, romantic relationships, and sexual behavior. The Sociological Quarterly 51, 408–435.
Waynforth, D., Dunbar, R.I.M., 1995. Conditional mate choice strategies in humans: evidence from 'lonely hearts' advertisements. Behaviour 132, 755–779.
Wiederman, M.W., Dubois, S.L., 1998. Evolution and sex differences in preferences for short-term mates: results from a policy capturing study. Evolution and Human Behavior 19, 153–170.
Wilbur, C.J., Campbell, L., 2010. What do women want? an interactionist account of women's mate preferences. Personality and Individual Differences 49, 749–754.

Human Resource Management, Psychology of

Wayne F Cascio, University of Colorado, Denver, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Human resource management (HRM) includes five broad activities that together constitute the HRM system: staffing, retention, development, adjustment, and managing change. Its objectives are to optimize the performance and productivity of all workers in an organization, and to help line managers manage those workers more effectively. Modern HRM traces its roots to a variety of historical influences, beginning with the industrial revolution. New challenges face today's HR managers based on the new paradigm of work: anytime, anywhere, in real space or in cyberspace.

At a broad level, human resource management (HRM) is concerned with choices – choices that organizations make from a wide variety of possible policies, practices, and structures for managing employees. Managing employees effectively has important consequences for individuals themselves as well as for the organizations that employ them. Thus there exists a substantial and growing body of research evidence showing a strong connection between how firms manage their people and the economic results they achieve (Crook et al., 2011).

As practiced in work organizations, HRM includes five broad activities that together constitute the HRM system: staffing, retention, development, adjustment, and managing change (Cascio, 2013). These activities can be carried out at the individual, work team, or larger organizational unit (e.g., department) level. They are the responsibility of every line manager, as well as of the HRM department.

Staffing comprises the activities of (1) identifying work requirements within an organization, (2) determining the numbers of people and the skills mix necessary to do the work, and (3) recruiting, selecting, and promoting qualified candidates. *Retention* comprises the activities of (1) rewarding employees for performing their jobs effectively, (2) ensuring harmonious working relations between employees and managers, and (3) maintaining a safe, healthy work environment. *Development* is a function whose objective is to preserve and enhance employees' competence in their jobs by improving their knowledge, skills, abilities, and other characteristics; HR specialists use the term 'competencies' to refer to these items. *Adjustment* comprises activities intended to maintain compliance with the organization's HR policies (e.g., compensation and discipline) and business strategies (e.g., innovation and cost leadership). *Managing change* is an ongoing process whose objective is to enhance an organization's ability to anticipate and respond to developments in its external and internal environments, and to enable employees at all levels to cope with the changes. As Jack Welch, legendary former CEO of General Electric, noted, "Look, HR should be every company's 'killer app.' What could possibly be more important than who gets hired, developed, promoted, or moved out the door? Business is a game, and as with all games, the team that puts the best people on the field and gets them playing together wins. It's that simple" (Welch and Welch, 2006, p. 92).

Responsibilities and Objectives

HRM activities can be carried out at the individual, work team, or larger organizational unit (e.g., department) level. Sometimes they are initiated by the organization (e.g., recruitment efforts or management development programs), and sometimes they are initiated by the individual or work team (e.g., voluntary retirement or safety improvements). It might appear that the activities of staffing, retention, development, adjustment, and managing change are the special responsibilities of the HR department. However, these responsibilities also lie within the core of every manager's job throughout any organization – and because line managers have the authority (the organizationally granted right to influence the actions and behavior of the workers they manage), they have considerable impact on the ways that workers actually behave. This implies two things: (1) a broad objective of HRM is to optimize the performance and productivity of all workers in an organization and (2) a special objective of the HR department is to help line managers manage those workers more effectively. The HR department accomplishes this special objective through policy initiation and formulation, advice, service, and control in close communication and coordination with line managers.

Generally speaking, HR specialists provide the technical expertise in each area, while line managers (or, in some cases, self-directed work teams) use this expertise in order to manage people effectively. In a small business, however, line managers are responsible for both the technical and managerial aspects of HRM.

For example, to enhance retention, line managers are responsible for treating employees fairly, resolving conflicts, promoting teamwork, and providing pay increases based on merit. To help line managers do these things effectively, the HR department must devise a compensation and benefits system that employees will perceive as attractive and fair, establish merit increase guidelines that will apply across departments, and provide training and consultation to line managers on all employee relations issues, such as conflict resolution and team building.

The Evolution of Human Resource Management

To appreciate where the field is today, it is instructive to examine its historical roots. As it is practiced in most developed

countries around the world, modern HRM has emerged from nine interrelated sources.

1. Rapid technological change that increased the specialization of labor associated with the industrial revolution
2. The emergence of free collective bargaining, with constraints established for unions and for employers
3. The scientific management movement
4. Early industrial psychology
5. Governmental personnel practices growing out of the establishment of the United States Civil Service Commission
6. The emergence of personnel specialists and the grouping of these specialists into personnel departments
7. The human relations movement
8. The behavioral sciences
9. Social legislation and court decisions

The following sections describe each of these sources in more detail.

The Industrial Revolution

Three characteristics of the industrial revolution were the development of machinery, the linking of human power to the machines, and the establishment of factories in which a large number of people were employed. The result was a tremendous increase in job specialization as well as in the amount of goods that workers could produce. 'Division of labor' became a rallying cry of this revolution.

Clearly, the industrial revolution greatly accelerated the development of business and commerce. Owners and entrepreneurs generally did quite well for themselves, but the average citizen fared poorly in comparison to today's workers in terms of purchasing power and working conditions. Labor was considered a commodity to be bought and sold, and the prevailing political philosophy of *laissez-faire* resulted in little action by governments to protect the lot of workers. Twelve-hour workdays were common, tens of thousands of workers were killed on the job, wages were often below the level needed to support a family, job insecurity and unemployment were constant threats, and blatant discrimination and harsh treatment were everyday experiences. Against this backdrop, it was inevitable that workers would organize collectively to improve their wages and working conditions. Thus, an increasingly assertive and sometimes radical trade union movement grew (Kaufman, 2008).

The Emergence of Free Collective Bargaining

Because of the way they were abused, workers organized collectively to protect themselves and to improve their lot in life. From their perspective, the industrial revolution fostered specialization, along with the need for workers within each specialization to organize themselves against its abuses. Trade unions, also called 'labor unions,' spread rapidly, and so did the incidence of strikes (see also Labor Unions). The first overt union activity in the United States took place in 1794, when the Philadelphia shoemakers attempted to raise their wages in reaction to a wage cut by their employers. The employers sued the union, arguing that workers who jointly attempted to raise their wages constituted an illegal conspiracy in restraint of trade. In 1806, a federal court ruled in favor of the employers and fined the employees involved. Until this 'conspiracy doctrine' was overturned in 1842 in the case of *Commonwealth of Massachusetts v. Hunt*, workers were discouraged from forming unions. In the *Hunt* case, the court ruled that labor unions were not criminal per se, for they could have honorable as well as destructive objectives. In short, a union's conduct will determine whether it is legal or illegal (Commons and Filmore, 1958). Nevertheless, in the United States, courts generally adopted an antiunion stance until 1935. In that year, the National Labor Relations Act was passed. It focused largely on labor's right to organize. It became an unfair labor practice for an employer to coerce or restrain employees in the exercise of their rights, to dominate or interfere with a labor organization, or to refuse to bargain collectively with a legal representative of the employees. Subsequent legislation refined, broadened, and set legal limits on the scope of management and union activities. From the perspective of labor, the development of free collective bargaining and the USA labor movement created the need for what is now recognized as effective HRM.

Scientific Management

From the perspective of management, the scientific management movement also created the need for effective HRM in order to bring order and rationality to the practice of management. Frederick Winslow Taylor was the prophet of scientific management, and the stopwatch was his 'bible' (Bell, 1972). Taylor began his experiments in the steel industry in 1885. His objective was the systematic analysis of work, which he broke down into its smallest mechanical elements, and then rearranged into the most efficient combination. The objective was to identify the 'one best way' to perform work. In addition to the scientific study of the task itself ('time and motion' study), Taylor argued that the individuals selected to do the work should be as perfectly matched physically and mentally to the demands of the job as possible, and that overqualified individuals should be excluded.

Employees should be trained carefully by supervisors (whose own work was also divided into specialties) to ensure that they perform the work exactly as specified by prior scientific analysis. In no case, however, should employees ever be called upon to work at a pace that would be detrimental to their health.

Finally, to provide an incentive for the employee to follow the detailed procedures specified (and closely supervised by line supervisors on the shop floor), Taylor felt that employees should receive an addition of between 30 and 100% of their ordinary wages when the task was done right and within the time limits specified. The first piece-rate incentive system had arrived.

Taylor also was interested in the social aspects of work, although he saw little good emerging from the social interaction within work groups. He felt that work groups fostered a level of individual efficiency equal to that of the least productive worker in the group. In other words, he believed that the efficiency of the group would not be any greater than that of its least productive member. Thus Taylor tried to institute a kind of social physics into the workplace.

Nevertheless, jobs designed only to maximize efficiency lead to predictable psychological consequences, and these have been well documented (Blauner, 1964; Kornhauser, 1965). Such jobs often lead to job dissatisfaction, surface attention to work, depersonalization, feelings of alienation (i.e., powerlessness and meaninglessness), and frustration due to lack of personal growth and development. In the context of automobile assembly line jobs in the early 1900s, Henry Ford lamented, "Why is it that I always get a whole person when what I really want is a willing pair of hands?" (quoted in Labich, 1994). On balance, however, while many of Taylor's ideas were controversial, there is no doubt that application of scientific management principles resulted in much higher productivity than would otherwise have been possible.

Early Industrial Psychology

The objective of early industrial psychology was to improve organizational goals (productivity and efficiency) primarily by applying psychology with an emphasis on individual differences through employee selection and training (Koppes, 2007). Hugo Munsterberg's 1913 book, *Psychology and Industrial Efficiency*, described experiments in selecting streetcar operators, ship's officers, and telephone switchboard operators. His contributions to human resource (or personnel) management are noteworthy for his emphasis on job analysis in terms of (1) the abilities required to do the job and (2) the development of testing devices – aptitude as well as work sample tests – that were able to forecast subsequent job performance (Munsterberg, 1913).

Munsterberg's 1917 book, *Business Psychology*, included the results of questionnaires sent to hundreds of laborers. The questionnaires requested anonymous data on attitudes toward their work, their involvement in the labor movement, and their leisure activities.

Both Munsterberg and Taylor were concerned with how workers were selected for jobs, as well as with the individual worker's feelings and aspirations. Although they differed on several issues (notably, on how social structure affected productivity), Munsterberg's writings clearly suggest that he felt that the two emerging fields of personnel selection and social psychology (the study of how people affect and are affected by one another) were complementary (Landy, 1992; Moskowitz, 1977).

When the United States entered World War I in 1917, psychologists Walter Dill Scott, Walter VanDyke Bingham, and Robert Yerkes collaborated in developing intelligence tests that could be used to select and classify recruits. The Army Alpha was administered to individuals with English-language skills, while the Army Beta was administered to recruits who were illiterate or did not have strong English-language skills. These tests paved the way for large-scale intelligence testing and for later expansion of psychological testing into government, industry, and education (Vinchur and Koppes, 2011).

Other developments at about the same time included the use of rating sheets for interviewers, advances in checking references given by workers, and advances in statistical methods for estimating validity (the extent to which selection devices accurately forecast job performance). Other kinds of psychological measures also appeared during and after World War I, such as measures of aptitude, interest, and personality. The measurement of attitudes was made possible by the scaling techniques of L. L. Thurstone in 1927, and improved further by Rensis Likert in 1932. Likert-type rating scales (e.g., scales with possible answers like *strongly agree*, *agree*, *neither agree nor disagree*, *disagree*, and *strongly disagree*) are still popular today.

The US Civil Service Commission

Employment in the federal government had reached 130 000 by the 1880s, and by 1920 it had surpassed 600 000, making it the nation's largest employer. Prior to civil service reform in 1883, federal employment was subject to the patronage, or 'spoils,' system. Government jobs were rewards that elected officeholders used to raise money and to reward their political supporters, based on the philosophy that "to the victor goes the spoils" (Kaufman, 2008). That system ended after a disappointed office seeker shot and killed President James Garfield in 1881, an event that galvanized public opinion against the spoils system. In its place, the Pendleton Act of 1883 established the US Civil Service Commission (today known as the US Office of Personnel Management). A forerunner in progressive personnel policies, the act provided that fair, open, and competitive examinations be administered to implement merit-based employment in the public service. It also provided a measure of employment security for those selected, prohibited discharge for refusing to engage in political activity, encouraged a nonpartisan approach to appointments, and mandated that a commissioner administers the act (Van Riper, 1958). Perhaps, the major impact of this act was to protect employment and promotion policies in the federal government from political influence. Over the years, these policies have influenced personnel practices in state and local governments as well as in private industry.

Private Industry's Approach to Personnel Management

Historians consider 1912 as the approximate date of the emergence of the modern personnel department (Eilbert, 1959). The term 'personnel,' with its connotation of managing people in organizations, began to appear around 1909. Between 1900 and 1920, while advances were being made in scientific management, industrial psychology, and the federal civil service, a number of personnel specialists or employment managers emerged in companies such as B. F. Goodrich, National Cash Register, and Standard Oil of California. The specialists managed such areas as employment, employee welfare (financial, housing, medical, and educational), wage setting, safety, training, and health. The employment management movement, born in Boston with the founding of the Employment Managers' Association of Boston in December 1912, was an early forerunner of the modern human resources department. Meyer Bloomfield, Louis Brandeis, and the Filene Brothers were key players in the development of this movement. It represented a newfound appreciation of the role of labor – the human factor – as a source of competitive advantage if it was managed wisely and humanely (Kaufman, 2008).

The Human Relations Movement

In Chicago in 1923, the Hawthorne Works of the Western Electric Company provided the setting for one of the most famous behavioral research efforts of all time. The purpose of the research was to study the effects of illumination (lighting) on worker productivity. In one experiment, production went up when the illumination increased. In another experiment, production went up when the illumination decreased. After three years of continually inconsistent results, the researchers concluded that in a situation involving people, it was impossible to change one condition (the lighting) without affecting other variables at the same time. Employee motivation and work group morale seemed to be the really crucial factors.

The results of subsequent experiments at the Hawthorne Works by Professors Elton Mayo, Fritz Roethlisberger, and T. North Whitehead of the Harvard Business School indicated that productivity was directly related to the degree of group teamwork and cooperation. The level of teamwork and cooperation, in turn, seemed to be related to the interest of the supervisor and the researchers in the work group, the lack of coercive approaches to productivity improvement, and the participation afforded the workers in changes affecting them (Pennock, 1930; Roethlisberger and Dickson, 1939). In short, the researchers came to view the organization of workers as a *social system*, in contrast to Taylor's view of that organization as a *technical–economic system*. Interest and research in topics such as leadership, motivation, group processes, and job satisfaction have direct antecedents in the Hawthorne research (Vinchur and Koppes, 2011).

The conviction that group behavior and workers' feelings were associated with morale and productivity characterized much of the research and theorizing in the human relations movement in the 1940s and the 1950s. Influential thinkers such as Douglas McGregor (Theory X and Theory Y managers), Chris Argyris (modern organizations are in conflict with the personalities of mature adults), and Rensis Likert (the 'linking pin' model of integrating small groups into the organization) focused on the fit between organizations and individuals. Their ideas can be summed up concisely: organizations and people need each other, and a good fit between the two benefits both (Shafritz and Ott, 1996).

In practice, unfortunately, many managers seemed to use human relations for the short-term purpose of manipulating workers to increase output, rather than for the long-term goal of satisfying worker needs while meeting organizational needs. Failure to find evidence that these programs made a difference in workers' satisfaction, or that happy workers were productive workers, helped kill this approach to HRM (Brayfield and Crockett, 1955).

The Behavioral Sciences

The behavioral-science approach to managing people is an outgrowth of human relations studies, although it embraces a wider base of academic and applied disciplines and is concerned with a wider range of problems. *Behavioral sciences* refer to the social and biological sciences concerned with the study of human behavior. Much of the knowledge of HRM and many of its practical applications have come from such behavioral-science disciplines as the following:

1. *Industrial/organizational psychology*: the study of the behavior of people at work (see Industrial-Organizational Psychology: Science and Practice)
2. *Social psychology*: the study of how people affect and are affected by one another
3. *Organization theory*: basic philosophies about why organizations exist, how they function, how they should be designed, and how they can be effective
4. *Organizational behavior*: the study of the causes of individual and group behavior, and of how this knowledge can be used to make people more productive and satisfied in organizational settings
5. *Sociology*: the study of society, social institutions, and social relationships

Another area that has exerted a considerable influence on HRM theory and practice is *general systems theory*. Systems theory has developed from such fields as biology, physics, and cybernetics (the study of mechanical–electrical communications systems and devices). It has been applied and elaborated in every behavioral science field.

We now know that the ways people behave in organizations cannot be explained simply by human relationships. The organization itself, through its unique 'culture,' molds, constrains, and modifies human performance. *Organizational cultue* has been described as the DNA of an organization – invisible to the naked eye, but critical in shaping the character of the workplace (Tetenbaum, 1999). *Culture* is the pattern of basic assumptions a given group has invented, discovered, or developed in learning to adapt to both its external and internal environments. This pattern of assumptions has worked well enough to be considered valid and, therefore, to be taught to new members as the correct way to perceive, think, and feel in relation to those problems (Martin, 2002; Schein, 2004). The way the organization is structured, the authority attached to different positions, and job and technology requirements clearly affect behavior, as do stories, legends, and myths about key people and events; what leaders pay attention to, measure, and control; and any implicit and possibly unconscious criteria that leaders use to determine who fits key slots in the organization (Schneider et al., 2011). Even though the present understanding of the determinants and effects of behavior is incomplete, there is a better sense of the ways in which separate influences interact with each other to affect individuals.

Social Legislation and Court Decisions

Governments in most developed countries around the world have passed social legislation, and courts have issued decisions on unfair discrimination in employment. While some countries are more aggressive in enforcing these issues than others, the laws and court decisions frame national policy on equal employment opportunity (EEO). In the United States, discrimination in employment is prohibited on the basis of race, religion, national origin, gender, age, and disability. Aggrieved individuals have the right to pursue their claims in court. At the organizational level, HRM professionals work with top

managers to develop and implement EEO policies to ensure that individuals are treated fairly in all aspects of the employment relationship.

Learning from the Past, Looking toward the Future

As noted earlier, the field of human resource management system is composed of five broad activities: staffing, retention, development, adjustment, and managing change. The field as it is known today has emerged from nine interrelated sources: the industrial revolution, the emergence of free collective bargaining, scientific management, early industrial psychology, governmental personnel practices growing out of the establishment of the US Civil Service Commission, the emergence of personnel specialists and the grouping of these specialists into personnel departments, the human relations movement, the behavioral sciences, and social legislation and court decisions.

Today's HR professionals need to be proactive, not reactive. They must anticipate changes in the social, legal, political, and economic environments, and then facilitate effective organizational responses to them. Some of these issues include finding and keeping top talent in tight labor markets, and capitalizing on the information revolution in ways that will benefit organizations, such as social media and other Internet-based recruitment sources. HR professionals must take the lead in establishing sensible privacy-protection policies for employees, and developing interactive, company-based intranets to facilitate employee inquiries and inputs. Other challenges include managing the performance of geographically dispersed employees and teams (virtual teams), and providing training and development opportunities on demand.

Many of the ways of working that we have taken for granted are disappearing – working nine to five, spending an entire career with only one company, taking weekends off, working with people we have known well, and working in offices we go to every day. Similar changes are happening in organizations. For example, the idea that hierarchy is the best way to manage information flows, that most people will work with team members in the same office, or that the majority of talent will be held within the boundaries of an organization. All of this is shifting, and what is coming in its place is less knowable and less understandable – accelerating technology, accelerating globalization, accelerating demographic changes, and accelerating social trends (Gratton, 2011).

Indeed, this is today's paradigm of work: anytime, anywhere, in real space or in cyberspace. This presents ongoing challenges in areas such as work-life fit (Galinsky and Matos, 2011); developing a total rewards package that will attract, retain, and motivate employees at all levels (Lawler, 2011); measuring social networks to identify trusted connections (Stephenson, 2011); promoting collective leadership (Boudreau and Ziskin, 2011); and promoting and enabling ambidextrous leadership that will identify new growth businesses, while remaining focused on the execution of existing ones (Probst et al., 2011). HRM will contribute to the bottom-line success of organizations and to the betterment of employee welfare by developing proactive responses to the workplace challenges wrought by this new paradigm.

See also: Industrial–Organizational Psychology: Science and Practice; Job Analysis and Work Roles, Psychology of; Job Design and Evaluation: Organizational Aspects; Labor Unions; Organizational Climate in the Work Setting; Wellbeing and Burnout in the Workplace, Psychology of.

Bibliography

Bell, D., 1972. Three technologies: size, measurement, hierarchy. In: Davis, L.E., Taylor, J.C. (Eds.), Design of Jobs. Penguin, London.
Blauner, R., 1964. Alienation and Freedom. University of Chicago Press, Chicago, IL.
Boudreau, J.W., Ziskin, I., 2011. The future of HR and effective organizations. Organizational Dynamics 40, 255–266.
Brayfield, A.H., Crockett, W.H., 1955. Employee attitudes and employee performance. Psychological Bulletin 52, 396–424.
Cascio, W.F., 2013. Managing Human Resources: Productivity, Quality of Work Life, Profits, ninth ed. McGraw-Hill/Irwin, New York, NY.
Commons, J.R., Filmore, E.A. (Eds.), 1958. Labor Conspiracy Cases, 1806 to 1842. A Documentary History of American Industrial Society, vol. 3. Russell & Russell, New York, NY.
Crook, T.R., Todd, S.Y., Combs, J.G., Woehr, D.J., Ketchen Jr., D.J., 2011. Does human capital matter? A meta-analysis of the relationship between human capital and firm performance. Journal of Applied Psychology 96, 443–456.
Eilbert, H., 1959. The development of personnel management in the United States. Business History Review 33, 345–364.
Galinsky, E., Matos, K., 2011. The future of work-life fit. Organizational Dynamics 40, 267–280.
Gratton, L., 2011. Workplace 2025 – what will it look like? Organizational Dynamics 40, 246–254.
Kaufman, B.E., 2008. Managing the Human Factor: The Early Years of Human Resource Management in American Industry. Cornell University, ILR Press, Ithaca, NY.
Koppes, L.L., 2007. History of industrial/organizational psychology in North America. In: Rogelberg, S.G. (Ed.), Encyclopedia of Industrial and Organizational Psychology, vol. I. Sage, Thousand Oaks, CA, pp. 312–317.
Kornhauser, A., 1965. Mental Health of the Industrial Worker. Wiley, New York, NY.
Labich, K., 1994, Nov. 14. Why companies fail. Fortune, p. 64.
Landy, F.J., 1992. Hugo Munsterberg: victim or visionary? Journal of Applied Psychology 77, 787–802.
Lawler III, E.E., 2011. Creating a new employment deal: total rewards and the new workforce. Organizational Dynamics 40, 302–309.
Martin, J., 2002. Organizational Culture: Mapping the Terrain. Sage, Thousand Oaks, CA.
Moskowitz, M.J., 1977. Hugo Munsterberg: a study in the history of applied psychology. American Psychologist 32, 824–842.
Munsterberg, H., 1913. Psychology and Industrial Efficiency. Houghton Mifflin, Boston, MA.
Pennock, G.A., 1930. Industrial research at Hawthorne and experimental investigation of rest periods, working conditions, and other influences. Personnel Journal 8, 296–309.
Probst, G., Raisch, S., Tushman, M.L., 2011. Ambidextrous leadership: emerging challenges for business and HR leaders. Organizational Dynamics 40, 326–334.
Roethlisberger, F.J., Dickson, W.J., 1939. Management and the Worker. Harvard University Press, Cambridge, MA.
Schein, E.H., 2004. Organizational Culture and Leadership. Jossey-Bass, San Francisco, CA.
Schneider, B., Ehrhart, M.G., Macey, W.H., 2011. Perspectives on organizational climate and culture. In: Zedeck, S. (Ed.), APA Handbook of Industrial and Organizational Psychology, vol. 1. American Psychological Association, Washington, DC, pp. 373–414.
Shafritz, J.M., Ott, J.S. (Eds.), 1996. Classics of Organization Theory, fourth ed. Wadsworth, Belmont, CA.
Stephenson, K., 2011. From Tiananmen to Tahrir: knowing one's place in the 21st century. Organizational Dynamics 40, 281–291.
Tetenbaum, T., 1999, Autumn. Beating the Odds of Merger & Acquisition Failure. Organizational Dynamics, pp. 22–36.
Van Riper, P.P., 1958. History of the United States Civil Service. Row-Peterson, Evanston, IL.
Vinchur, A.J., Koppes, L.L., 2011. A historical survey of research and practice in industrial and organizational psychology. In: Zedeck, S. (Ed.), Handbook of Industrial and Organizational Psychology, vol. 1. American Psychological Association, Washington, DC, pp. 3–36.
Welch, J., Welch, S., 2006, July 17. So Many CEOs Get this Wrong. BusinessWeek, p. 92.

Human Rights and Social Work

Elisabeth Reichert, Southern Illinois University at Carbondale, Carbondale, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Human rights have provided a contentious issue in the social work profession, and have a patchy implementation globally. This is changing as social workers and educators throughout the world embrace concepts of human rights. The International Federation of Social Workers and International Association of Schools of Social Work have incorporated human rights throughout their mission statements. However, while human rights have now become more accepted within social work education and practice, social workers still have to understand what human rights actually represent. Application of human rights to social work policies and practices remains a work in progress.

Introduction

Human rights have recently been prioritized for inclusion in social work education and practice. In this article, I describe human rights and argue for their further development in the profession.

Overview of Human Rights and Social Work

The connection between human rights and social work has become increasingly important within the last decade. In 2004, at a meeting in Adelaide, Australia, the International Federation of Social Workers (IFSW) and International Association of Schools of Social Work (IASSW) jointly made a major change within the international code of ethics for social workers throughout the world by agreeing an ethics document that covered practice, education, and research. This agreement changed IFSW's 1994 ethics document that did not mention human rights as a social work principle. The 2004 document incorporated human rights as a dominant guidepost for social work policies and practice and turned social work into a human rights profession. These two international umbrella organizations for the social work profession now expect practitioners in their member countries to adhere to human rights principles. Thus, for example, the US-based Council of Social Work Education has required human rights to become part of the social work curriculum that US schools of social work must teach for accreditation (Council of Social Work Education, 2008). Outside the United States, social work educators have also integrated human rights into their works (e.g., Ife, 2008; Wetzel, 1993; Wronka, 2008; Dominelli, 2009).

Why have human rights been recently integrated into the wider social work profession? This integration of human rights with social work may have always been there, without being specifically stated as currently. In the past, social work principles often related to human rights principles, but little formal effort that specifically linked human rights with social work. What is now happening is formal recognition of the strong connection between human rights and social work. For instance, the American social work code of ethics today still does not mention human rights within its ethical statements (National Association of Social Workers, n.d.). Yet, a reading of that code indicates a strong connection to human rights principles. Values stated in the code include social justice, dignity and worth of the individual, and integrity, all of which require reference to human rights. Although human rights have always been explicitly part of contemporary social work, the profession's international organizations have made this connection formal and explicit in the *Global Agenda for Social Work and Social Development* (IASSW, IFSW, ICSW, 2004).

What Are Human Rights?

One of the major challenges within the social work profession is that of defining human rights. Understanding the meaning of human rights is crucial to social workers using these within their practice. Unfortunately, in today's world, politicians and government officials often leave out the term human rights when criticizing others for all types of sins, including torture, suppression of political speech, religious discrimination, and gender bias. These acts violate human rights, but social workers' concerns go far beyond a litany of violations. Often overlooked, human rights include positive rights, like a right to water, food, education, adequate health care, housing, work, clean environment, leisure time, and other important quality of life matters (United Nations, 1948). Rarely do government leaders discuss these equally important human needs within the context of human rights (George, 2003). By failing to connect social, cultural, and economic needs to the overall concept of human rights, governments leave huge gaps regarding the realization of human rights.

Universality and Indivisibility

Human rights cannot exist within a vacuum. They must be linked to an overall goal of how best to manage and enforce these rights. Within this goal lies the notion of universality: Every individual, wherever she or he resides, has a claim to specific human rights principles, which should be accepted as valid globally. Governments should not have the authority to deny an individual their human rights. Rather, governments have an obligation to establish frameworks within which to promote human rights. Problems with universality arise when a government does not accept a particular need as a human

right, for example, the United States does not acknowledge medical care as a human right even though the Universal Declaration of Human Rights (UDHR) specifically classifies medical care as one of these – Article 25, UDHR. American laws may require public hospitals to treat individuals for emergency care, but there is no other legal requirement to provide medical care. In contrast, other Western countries view medical care as a right that governments must provide. Some Asian countries reject UDHR definitions of human rights for being embedded in Western concepts that focus on individual, not group rights (Yip, 2005).

Other commonly accepted human rights include free speech, right of assembly, and elections (Art. 19–21). While more and more countries seem accepting of these universal values (note recent developments in the Middle East), differences exist in their interpretation of these. For instance, China has restricted some opinions from being communicated over the Internet; other countries accept greater latitude in allowable content.

Universality does not mean that everyone is alike or should ignore differences. Universality allows all types of diversity and differences among individuals and groups. Its main purpose is to ensure that individuals and peoples everywhere have basic rights within their lives, but not that these will be the same everywhere (Reichert, 2011). For example, European countries fund medical care in diverse ways (Esping-Andersen, 1990).

Within a human rights framework, indivisibility goes alongside universality. Indivisibility refers to the necessity for governments and individuals to recognize each human right and not selectively promote some while ignoring others. International agreement may exist on the definition of a human right like medical care, but a country may refuse to enforce that right. Without adequate health care, an individual may ignore treatment of an illness. Impaired health can reduce an individual's enjoyment of other human rights, like employment or participation in community elections. The denial of one human right impacts on the enjoyment of other rights, making indivisibility crucial to any human rights framework.

How is it possible to promote human rights when a country can pick and choose what human rights it accepts or ignores? This is a major dilemma within the structure of human rights. Although every country belonging to the United Nations agrees to accept the UDHR as a guiding document, many countries fail to embrace fully all the rights covered by that document. A consequence of the parochial treatment of human rights is the belittling of what human rights mean (MacNaughton and Frey, 2011). Social workers may become frustrated when confronted with entrenched government structures that diminish the enjoyment of human rights. For human rights to have meaning to the profession, social workers have to develop interventions that can overcome obstacles to their realization. Possible approaches for integrating human rights into social work practice are discussed in Section Human Rights within Social Work Context.

Universal Declaration of Human Rights

Any understanding of contemporary human rights, whether by students, social workers, government officials, lawyers, or other interested persons, has to begin with a thorough analysis of the UDHR (United Nations, 1948). The United Nations (UN) adopted the UDHR in 1948 and requires all member nations to adhere to it. Social workers were involved through Katherine Kendall as IASSW's representative. The UDHR lists fundamental freedoms and rights that each member nation is expected to promote among citizens and residents.

The drafting of the UDHR occurred with the horrors perpetrated by Nazi Germany in mind (Morsink, 1999). Nazi Germany had labeled numerous individuals and groups as non-Germans, depriving them of rights and benefits granted to ethnic German citizens, with the eventual goal of exterminating non-Germans (Arendt, 1956). While Nazi Germany's genocidal policies provided the most immediate rationale for universal rules that would protect vulnerable populations, other factors also played a role (Reichert, 2011). Poverty and unemployment had afflicted many individuals throughout the world during the Great Depression of the 1930s. The colonization of African and Asian countries by Europeans also provided fuel for a universal rule against the exploitation of peoples by more powerful ones. The list of issues concerning the human condition seemed endless after World War II. Their discussion produced a broad-based document defining the essential underpinnings for human existence. Given vast cultural and other differences among individuals, finding common needs and freedoms for everyone to share was crucial. The outcome of the ensuing intensive soul-searching was the UDHR. While not perfect, its principles provide a worthy foundation for advancing political and social rights that everyone should enjoy, regardless of where she or he resides.

The UDHR does not contain language that legally binds a nation to enforce human rights. Although many countries have integrated human rights into their own laws, this integration remains a work in progress (Petersmann, 2002). Reference to the UDHR as a standard for acceptable conduct regarding human rights highlights its significance. However, the conflict between the principle of national sovereignty and a state's duty to care for each citizen results in international reluctance to enforce adherence to the UDHR (Dominelli, 2009).

To understand the fundamental concepts of human rights, social workers should read the UDHR and analyze the document as a whole. It begins by recognizing the inherent dignity and equal and inalienable rights of all members of the human family as the foundation of freedom, justice, and peace in the world (Preamble). Human rights should be protected by laws and everyone should enjoy those rights. The UDHR often refers to freedom as an important goal without defining this term. When different cultures and political systems are involved, defining freedom becomes difficult. I argue that freedom in the UDHR does not refer to a Western or American type of freedom, but one that allows for general improvement in social and economic conditions. A common understanding of rights and freedoms plays a key role in realizing universal human rights. Although what constitutes human rights and freedoms can be contested, nations can agree some similarities while acknowledging that differences in interpretations will occur owing to cultural, economic, political, and social diversity among them.

Political and Civil Human Rights

The UDHR lists a number of political and civil rights that all member countries of the UN have accepted as a core foundation for their political and civil rights frameworks (Art. 1–21), and upon which they are expected to construct their political structures. Those human rights include prohibitions against slavery, torture, arbitrary arrest and detention, gender and ethnic discrimination, and arbitrary interference with an individual or family's privacy. Other political and civil rights listed provide the right to asylum, fair legal systems, presumption of innocence against criminal charges, freedom of movement, freedom of expression, fair elections, and peaceful assembly.

However, some countries may interpret these rights in ways that violate the clear intent of the UDHR. Such responses illicit questions such as: Should a country be allowed to detain individuals suspected of terrorist acts for unlimited periods of time without trial or a hearing to examine the evidence of suspected terrorist acts? Should a country be allowed to torture suspects accused of terrorism to extract information that might prevent future terrorist acts? Should previously accepted guidelines defining torture be revised to allow treatment that was previously accepted as torture? Should a country be allowed to execute suspected terrorists without trial or even arrest, as occurs through the use of unmanned drones to kill suspected terrorists? These acts appear to violate political human rights contained in the UDHR. Yet, what enforcement mechanism can prevent such human rights violations, particularly if the accused is a major power like the United States? Various courts may consider individual human rights cases to provide a remedy. But when a country adopts general policies that violate established human rights, or refuses to acknowledge the legitimacy of the International Criminal Court at The Hague, human rights violations will continue. This lack of enforcement poses a significant challenge in promoting human rights.

Cultural, Economic, and Social Human Rights

A second category or type of human right contained in the UDHR involves cultural, economic, and social needs (Art. 22–27). It provides that everyone has the right to social security, education, housing, medical coverage, employment, and leisure time. Everyone also has the right to participate in the cultural life of a community and enjoy the protection of interests resulting from scientific, literary, or artistic creations.

Cultural, economic, and social human rights differ from political and civil human rights in that a community or government will have to be more proactive in insuring their realization. Political and civil human rights aim to prevent governments from discriminating against particular ethnic groups. As long as the government does not discriminate, no human right is violated. However, cultural, economic, and social rights often require a government to provide every citizen or resident with services like health care or education. This can be costly and some governments may lack resources for funding these. Some people contend that economic and social human rights dilute the traditional human rights focus on political and civil rights (*The Economist*, 2007). This view maintains that food, jobs, and housing are necessities, but not rights that can be violated, unlike a government detaining individuals without a fair trial. This clarity on political rights seldom holds for economic and social rights.

Social workers can question the opinion that economic and social human rights are not true human rights because they will most likely encounter situations where economic and social human rights are at issue. In contemporary human rights, the lack of economic and social rights can be linked to dictatorships in several countries, including Germany under the Nazis. In others, for example, during the Great Depression of the 1930s, the United States had to respond to its residents' economic and social welfare by creating work projects and other economic assistance programs (Sunstein, 2004). Economic and social human rights are difficult to separate from political and civil rights. Neither type of human right should be subordinated to the other, because the deprivation of one can easily lead to the deprivation of other. Additionally, where countries fail to address economic and social inequality, economic growth may suffer (Lowrey, 2012; *The Economist*, 2012).

Other Provisions of the UDHR

Aside from promoting political, civil, economic, and social human rights, the UDHR encourages working toward a social and international order in which human rights can be realized (Art. 28). Without UN member countries cooperating on international concerns and disputes, human rights principles can easily be ignored. An example of this concept is the failure of the international order to uphold a multinational environmental treaty to reduce greenhouse emissions that affect more than one country. Environmental rights were not part of the UDHR, but social workers are demanding their inclusion (Dominelli, 2012).

Another provision of the UDHR states that everyone has a duty to respect the human rights of others and the community in which the individual resides (Art. 29). Human rights impose duties upon governments and individuals regarding recognition of their importance to the community as a whole.

Challenges to the UDHR

Major challenges to the UDHR concern its enforcement and interpretation. Whose voices are being heard when interpreting human rights? An individual with income in the top 10% of income earners might object to redistributing that income among lower income individuals through taxation. Higher income earners might view those earning less as lazy or undeserving, but human rights are not designed to function in that way. All voices should be heard when conceptualizing human rights. Human rights belong to everyone, not a select group. If someone does not have adequate medical care or food with which to survive, then government and other members of the community should provide that medical care and food. Every individual is responsible for satisfying those needs, but circumstances may prevent the individual from doing so, for example, lacking resources for buying medical care or food. Labeling a person as lazy or undeserving is unhelpful in addressing human rights issues because the structural causes of an individual's predicament can be ignored. Governments and

community organizations should address structural obstacles to satisfying human rights by asking: Why does a person not have adequate medical care? Is it the individual's fault or are there cultural and social reasons for it? Why does a person not have sufficient food? Is the individual making unwise choices in purchasing items, perhaps preferring lottery tickets or tobacco products over food? Does the individual simply have no income? The human rights mission of professional social work is to examine circumstances behind failure to meet human rights goals and try to remedy these (IFSW and IASSW, 2004).

Even where countries generally agree the interpretation of a human right, enforcement can be difficult. For example, countries may agree that women have equal status to men, but underlying social and cultural norms may prevent its realization. Women may not enjoy the same level of respect as men in holding public office, managing large companies, occupying upper administration positions within education, and any number of significant roles within society. Contesting such human rights violations can be detrimental to women's career or lives, not to mention the emotional stress of raising objections to entrenched norms. Despite the Convention to Eliminate Discrimination Against Women and others linked to ethnicity, disability, and age, an egalitarian human rights framework can be difficult to create because this would require some people to relinquish current privileges.

No simple answers can confront human rights challenges. However, social work occupies a unique position in educating about and confronting human rights violations. Ethical statements require social workers to educate themselves about human rights and address their violation. This task is not easy, but one that ethical social workers are duty bound to pursue.

Covenants and Other Documents on Human Rights

The UN has drafted numerous documents following the UDHR to provide wider acceptance and guidance on realizing human rights. While there are many, I have space to focus on two key covenants or international treaties that are binding on nations adopting them: the International Covenant on Civil and Political Human Rights and the International Covenant on Economic, Social, and Cultural Rights (United Nations, 1976, www.un.org). Many nations have endorsed these covenants and, at least on paper, have integrated these human rights into their legal systems. The European Court of Human Rights requires member countries to adhere to provisions within the UDHR and these covenants (European Court of Human Rights, www.echr.coe.int). The UN does not require member nations to adopt these or any other covenants, and some have not. One notable exception to the adoption of the covenant on economic and social rights is the United States, which refuses to accept various UN covenants or treaties on human rights. Its failure to promote UN documents on human rights or endorse their wider acceptance lies partially within the American legal framework for approving treaties and partly within its political structure where national sovereignty must prevail over UN treaties. Nations that sign international treaties, like the covenants on human rights, must relinquish some control over their legal systems, since they are obliged to incorporate human rights into their laws. American politicians have not been willing to do this and have failed to approve the vast majority of UN human rights treaties, including the Convention on the Rights of the Child.

International Covenant on Civil and Political Rights

The UDHR laid the foundation for the modern era of human rights. However, the UN did not draft the UDHR in legally binding language. To legally require countries to acknowledge and enforce human rights, the UN has proposed various covenants or conventions, which are to have the force of law once a member country had given approval. One of the first major covenants that the UN drafted after the UDHR was the International Covenant on Civil and Political Rights. Its counterpart was the International Covenant on Economic, Social, and Cultural Rights.

The International Covenant on Civil and Political Rights duplicates UDHR provisions that address what might be classified as freedom from wrongful treatment by a government. No country may

- torture someone (Art. 7);
- practice slavery or forced labor (Art. 8);
- arbitrarily arrest or detain an accused (Art. 9);
- discriminate against persons on the basis of a status, including gender (Art. 26); or
- interfere with an individual's right or privacy (Art. 17), freedom of expression (Art. 19), or use of its legal system (Art. 16).

Other sections of this covenant include the right of self-determination among peoples (Art. 1) and the right of minority groups within a state to enjoy their own culture, religion, and language (Art. 27). Countries that agree to follow this covenant must submit periodic reports to a Human Rights Committee as to progress made in satisfying its human rights provision (Art. 40, para. 1).

Social workers can embrace this covenant's political and civil rights that guarantee that everyone can voice opinions, preserve cultural backgrounds, enjoy freedom of speech, freedom of movement, and freedom from arbitrary arrest and torture. They can also integrate these into social work policies and practices.

International Covenant on Economic, Social, and Cultural Rights

The counterpart to the International Covenant on Civil and Political Rights is the International Covenant on Economic, Social, and Cultural Rights. This latter covenant also derives its provisions from the UDHR but focuses on needs or quality of life issues. For this reason, some countries, most notably the United States, do not consider this one as pressing as that on civil and political rights. Unlike many other countries, the United States does not recognize the Covenant on Economic, Social, and Cultural Rights as a valid expression of human rights, and so refuses to ratify it.

The Covenant on Economic, Social, and Cultural Rights requires that countries provide technical and vocational guidance, training programs, and policies to achieve economic,

social, and cultural development, as well as full and productive employment (Art. 6). Women are guaranteed conditions of work equal to those of men, and workers should have a decent living (Art. 7). This emphasis on work is a key part of the covenant, which acknowledges that societal harm can occur through unemployment or unfair working conditions. Some countries may not have the resources to fulfill the work requirements, but, at a minimum, governments should assist residents in obtaining suitable employment opportunities. Other parts of the covenant require states to provide education (Art. 13), food and housing (Art. 11), health care (Art. 12), and social security (Art. 9) to every resident, although many Western governments are pursing austerity measures despite the risks in socioeconomic rights.

Like the Covenant on Civil and Political Human Rights, the Covenant on Economic and Social Rights provides for reports on steps taken to fulfill these rights, but allows states more leeway in satisfying them. This can be an implicit acknowledgment that many countries might have difficulties meeting the covenant's requirements.

Other Important Human Rights Documents

While the UDHR and covenants on political and socioeconomic human rights form the core of contemporary human rights, the UN has proposed numerous other documents for adoption and inclusion into member states' legal system. Briefly, these include the:

- *Convention on the Elimination of All Forms of Discrimination against Women* (CEDAW, 1979). This aims to eliminate distinctions, exclusions, or restrictions made on the basis of sex, which aim to impair women's human rights (Art. 1).
- *Convention on the Rights of the Child* (CRC, 1989). Member nations approving this Convention are obliged to recognize the special needs of children and respect their human rights. This convention specifically recognizes the responsibilities, rights, and duties of a child's parents or guardians in raising the child. No provision of this Convention aims to deprive parents of their accustomed role as the child's caregivers.
- *Convention on Elimination of All Forms of Racial Discrimination* (CEAFRD, 1965). This seeks to prevent ethnic discrimination.
- *Convention on Rights of Persons with Disabilities* (CRPD, 2009). This promotes the rights of disabled people.
- *Principles for the Older Person* (POP, 1999). These have less force than conventions, but can assist in reducing age discrimination.
- *Declaration on the Right to Development* (DRD, 1986). Developmental rights are contentious, but this declaration helps countries to industrialize.
- *Millennium Declaration* (MD, 2000). The *Millennium Declaration* stresses broad goals (Millennium Development Goals) to be achieved by the year 2015, including eradication of poverty and hunger, achieving universal primary education, ensuring environmental sustainability, reducing child mortality, empowering women, and developing of a global partnership for development.

The adoption of UN documents on human rights coupled with difficulty in their enforcement has led to the formation of organizations like Human Rights Watch. This regularly investigates human rights abuses around the world and provides annual surveys on compliance with human rights principles (Human Rights Watch, www.hrw.org). Surveys issued by Human Rights Watch highlight human rights violations in individual countries, and provide annual updates. Human Rights Watch covers practically every country, including Western countries like the United States that seem to place themselves above others in the human rights arena.

The social work profession has much to gain from integrating human rights into its policies and practices. Its review of UN documents on human rights can clarify the issues that directly impact upon a social worker's education and career.

Human Rights within Social Work Context

The connection between human rights principles and social work policies and practices requires more than a study of UN documents. Social workers must understand the meaning of human rights and develop the tools for integrating human rights into the profession. The IFSW and IASSW have developed and stated the core purposes of social work throughout the world, *The Global Standards Document* (Sewpaul and Jones, 2004). Those purposes include:

- Facilitating the inclusion of socially excluded, dispossessed, and vulnerable groups of people;
- Addressing and challenging barriers, inequalities, and injustices in society;
- Forming working relationships with individuals, families, groups, organizations, and communities to enhance their well-being and problem-solving capacities;
- Assisting and educating people to obtain services and resources in their communities;
- Formulating and implementing policies and programs that enhance people's well-being, and promoting development and human rights, and collective social harmony;
- Encouraging people to advocate on pertinent local, national, regional, and international concerns;
- Engaging in social and political action to influence social policy and economic development and effect change by critiquing and eliminating inequalities; and
- Promoting respect for traditions, cultures, and religions among different ethnic groups and societies insofar as these do not conflict with human rights.

These core activities place social workers at the forefront of understanding clients, their needs, and cultures within which they exist. Social workers should not be passive in this pursuit but should take a leading role rather than a secondary role to lawyers or government officials when addressing human rights issues affecting clients. The IFSW and IASSW documents encourage social workers to adopt more active roles in addressing the many political, economic, and social issues that afflict countries throughout the world. One example of this is community-based human rights advocacy practice, an approach to community organizing that combines theoretical and practice frameworks from law and community organizing in social work practice. Human rights, such as the right to housing, health, and education, are at the forefront of such

practice. Individual and community constituents are trained to advocate for entitlements and maneuver within systems to access rights (Grodofsy, 2012). A social worker applying human rights to their practice should first analyze a particular set of circumstances in relation to human rights principles. Questions to ask include: What human rights issues are relevant to the situation? What are the specific provisions in human rights documents that apply? What controversy or conflict might arise from the human rights relevant to this situation? Does cultural relativism play a role in the social worker's perception of what constitutes human rights?

Cultural Relativism

Another key concept in analyzing human rights within a social work situation is that of cultural relativism. In cultural relativism, all points of view are equally valid, and any truth is relative, with truth belonging to the individual or her or his culture (Reichert, 2011). All ethical, religious, and political beliefs are truths related to the cultural identity of the individual or society. Cultural relativism is appropriate for certain human rights; for instance, language, food. However, others like clothing, art, and architecture differ from one culture to another and maintaining these differences remains important (Pasamonik, 2004). Also, cultural relativism suggests an irreducible diversity among cultures because each culture is unique as a whole, but with parts so intertwined with others that none can be understood or evaluated without reference to the others and the cultural whole, the so-called pattern of culture (Lawson, 1998).

The phrase 'cultural relativism' creates confusion when discussing human rights because it refers to a view that all cultures are equal and that universal values are secondary when examining cultural norms. No outside value is superior to that of the local culture. So, for example, if local cultures allow female genital mutilation, the human right prohibiting cruel or degrading treatment cannot prevent it (Reichert, 2011). Such conundrums indicate that like universality, cultural relativism has limits. A social worker should examine the historical background of a cultural practice that appears to violate contemporary human rights principles and ask Which group has determined the cultural practice? Has the decision-making process been democratic, with voices from many groups heard in determining it? This analysis will often find that cultural relativism rests on outdated or oppressive foundations.

Conclusion

The social work profession is well suited to integrating human rights principles into its policies and practices. This integration presents many challenges, including lack of government support for their fulfillment. Moreover, enforcement is weak because nations are reluctant to intervene in sovereign nations' internal affairs. And, cultural relativism can confuse practitioners wishing to act. However, by understanding what human rights mean to particular individuals in specific contexts, and acting to apply human rights to such situations, social workers can confront human rights challenges within the profession. The IFSW–IASSW guidelines are useful in developing this dimension of social work theory and practice.

See also: Children, Rights of: Cultural Concerns; Civil Liberties and Human Rights; Cultural Rights and Culture Defense: Cultural Concerns; History of Human Rights; Human Rights, Anthropology of; Human Rights: Legal and Sociolegal Perspectives; Indigenous Rights; International Association of Schools of Social Work (IASSW); International Council on Social Welfare; Social Justice and Social Inequalities; Social Work Education; United Nations and Its Agencies.

Bibliography

Arendt, A., 1956. On Totalitarianism. Andre Deutsch, London.
CEAFRD, 1965. Convention on Elimination of All Forms of Racial Discrimination. www.un.org (accessed 12.12.13.).
CEDAW, 1979. Convention on Elimination of All Forms of Discrimination against Women. www.un.org (accessed 12.12.13.).
Council of Social Work Education, 2008. Educational Policy and Accreditation Standards. CSWE, Alexandria: Virginia.
CRC, 1989. Convention on the Rights of the Child. www.un.org (accessed 12.12.13.).
CRPD, 2009. Convention on Rights of Persons with Disabilities. www.un.org (accessed 12.12.13.).
Dominelli, L., 2009. Social Work in a Globalizing World. Polity Press, Cambridge.
Dominelli, L., 2012. Green Social Work. Polity Press, Cambridge.
DRD, 1986. Declaration on Right to Development. www.un.org (accessed 12.12.13.).
The Economist, 24 March 2007. Stand up for Your Rights, p. 12.
The Economist, 13 October 2012. Special Report: For Richer, For Poorer, pp. 1–24.
Esping-Andersen, G., 1990. The Three Worlds of Welfare. Routledge, London.
European Court of Human Rights. www.echr.coe.int (accessed 20.12.13.).
George, S., 2003. Globalizing rights? In: Gibney, M.J. (Ed.), Globalizing Rights. Oxford University Press, Oxford.
Grodofsy, M., 2012. Community-based human rights advocacy practice and peace education. International Social Work 55 (5), 740–753.
Human Rights Watch. www.hrw.org (accessed 20.12.13.).
International Association of Schools of Social Work (IASSW), International Federation of Social Workers (IFSW), International Council on Social Welfare (ICSW), 2004. The Global Agenda for Social Work and Social Development. www.iassw-aiets.org (accessed 20.12.13.).
Ife, J., 2008. Human Rights and Social Work: Towards Rights Based Practice, second ed. Cambridge University Press, Cambridge.
IFSW, IASSW, 2004. Core Purposes. The Global Standards. www.ifsw.org (accessed 20.12.13.).
IFSW, IASSW, n.d. Statement of ethical principles and guidelines. The Ethical Documents. www.ifsw.org (accessed 20.12.13.).
Lawson, S., 1998. Democracy and the problem of cultural relativism: normative issues for international politics. Global Society: Journal of Interdisciplinary International Relations 12 (2), 251–271.
Lowrey, A., 16 October 2012. Income inequality may take toll on economic growth. New York Times. www.nytimes.com (accessed 20.12.13.).
MacNaughton, G., Frey, D.F., 2011. Decent work for all: a holistic human rights approach. American University International Law Review 26 (2), 441–483.
McDowell, T., Libal, K., Brown, A.L., 2012. Human rights in the practice of family therapy: domestic violence, a case in point. Journal of Feminist Family Therapy 24 (1), 1–23.
MD, 2000. Millennium Declaration. www.un.org (accessed 20.12.13.).
Morsink, J., 1999. The Universal Declaration of Human Rights: Origins, Drafting, and Intent. University of Pennsylvania Press, Philadelphia, PA.
National Association of Social Workers, n.d. Code of Ethics. www.naswdc.org (accessed 12.12.13.).
Pasamonik, B., 2004. The paradoxes of tolerance. Social Studies 95 (5), 206–211.
Petersmann, E., 2002. Time for a United Nations "global compact" for integrating human rights into the law of worldwide organizations: lessons from European integration. European Journal of International Law 13 (3), 621–650.
POP, 1999. Principles for the Older Person. www.un.org (accessed 12.12.13.).

Reichert, E., 2004. Understanding Human Rights: An Exercise Book. Sage Publications, Thousand Oaks, CA.
Reichert, E. (Ed.), 2007. Challenges in Human Rights: A Social Work Perspective. Columbia, New York.
Reichert, E., 2011. Social Work and Human Rights, second ed. Columbia University Press, New York.
Reinbold, J., 2011. Political myth and the sacred center of human rights: the universal declaration and the narrative of "inherent human dignity". Human Rights Review 12, 147–171.
Sewpaul, V., Jones, D., 2004. Global standards for social work education and training. Social Work Education 23 (5), 493–513.
Steen, J., 2012. The human rights philosophy as a values framework for the human behavior course: integration of human rights concepts in the person-in-environment perspective. Journal of Human Behavior in the Social Environment 22 (7), 853–862.
Sunstein, C., 2004. The Second Bill of Rights: FDR's Unfinished Revolution and Why We Need it More than Ever. Basic Books, New York.
United Nations, 1948. Universal Declaration of Human Rights. United Nations Document a/810. UN, New York.
United Nations, 1976. International covenants on political and civil rights and Economic, Social, and Cultural Rights. www.un.org (accessed 12.12.13.).
Wetzel, J., 1993. The World of Women: In Pursuit of Human Rights. Macmillan, London.
Wronka, J., 2008. Human Rights and Social Justice: Social Action and Service for the Helping and Health Professions. Sage, Los Angeles, CA.
Yip, K., 2005. A dynamic Asian response to globalisation in cross-cultural social work. International Social Work 48 (5), 593–607.

Relevant Websites

www.echr.coe.int.
www.hrw.org.
www.iassw-aiets.org.
www.icsw.org.
www.ifsw.org.
www.naswdc.org.
www.un.org.

Human Rights, Anthropology of

Mark Goodale, George Mason University, Burke, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by K. Hastrup, volume 10, pp. 7007–7012, © 2001, Elsevier Ltd.

Abstract

Anthropology has had a curious and contested disciplinary relationship with the postwar human rights project. This article analyzes the history of this relationship from the period of the drafting of the University Declaration of Human Rights to the present. After describing both political and critical approaches to human rights by anthropologists, the article discusses contemporary anthropology of human rights that emerged after the end of the Cold War. This article reviews the major topics and areas of theoretical contribution in the anthropology of human rights and concludes with a discussion of future directions.

Overview

Although anthropologists took an interest in human rights from the creation of the modern human rights system after the Second World War, only recently has it been possible to speak of the anthropology of human rights. Unlike the anthropology of religion, kinship, politics, or law, for example, the anthropology of human rights developed almost accidentally as a response to changes in historical and geopolitical incidences that took place largely after the end of the Cold War. Anthropologists in the field, in the midst of ongoing research on problems such as social change, legal identity, transnational networks, and indigenous mobilization, found themselves confronting emergent forms of political, legal, and cultural practices that were being shaped by a suddenly urgent discourse of human rights. Throughout the 1990s and 2000s, anthropologists tracked this rapid expansion of human rights from a number of different methodological and theoretical perspectives. As the liminal period of the post-Cold War arguably gives way to a historical phase of re-differentiated hierarchy and constraint in which human rights play an important – although not globally hegemonic – role in politics, social ethics, and legal regulation, the anthropology of human rights will continue to develop as an important disciplinary subfield and mode of critical analysis.

This article on the anthropology of human rights is structured as follows. In the next section, the history of anthropology's often fraught relationship with human rights is surveyed and a timeline of key turning points in this history is described. As a matter of historiography, the emphasis on a particular chronology is meant to provide an interpretative framework rather than establish a set of unassailable historical facts. The historical narrative of anthropology and human rights has itself been a matter of some dispute. Some of this contention will be explained and an effort will be made to achieve some measure of balance among the most notable differing perspectives. After examining the history of anthropology's various engagements with human rights, the article will then identify and describe key themes that emerge from this dynamic subfield and which mark off the anthropology of human rights as a distinctive source of ideas for human rights scholars, practitioners, and policy-makers. Much of the early anthropological skepticism toward certain aspects of the postwar human rights project has been carried forward in the form of a growing body of critical anthropological theory. Nevertheless, this has not created a divide between the findings that have emerged from the ethnography of human rights practices and the demands of human rights activists and policy-makers for human rights theory with wider application. Instead, somewhat surprisingly, recent work in the anthropology of human rights has been taken up in policy documents and commissioned reports among a range of international institutions, including the Office of the United Nations High Commissioner for Human Rights, the World Bank, and the Committee on the Elimination of Discrimination against Women. Even the President of Ireland, Michael D. Higgins, a noted human rights activist and author, has made several studies in the anthropology of human rights central to policy speeches and annual addresses. Finally, the article considers future directions in the anthropology of human rights with a focus on promising avenues for new research and public scholarship.

History

Outside of anthropology, a debate is currently raging between scholars who view the development of ideas and practices of human rights over the *longue durée* and those who see human rights as a much more recent and unprecedented phenomenon. To the former, the history of human rights must be analyzed as part of an incremental, centuries-long process in which conceptions of universal rights were extracted from an even longer intellectual history with roots in classical antiquity and then ecclesiastical Europe; these ideas were eventually made the basis for political ideologies with practical consequences for politics and modes of governance; and, finally, these ancient ideas of human rights found expression in the 1948 Universal Declaration of Human Rights (UDHR), which must be understood as both a culmination and an apotheosis. For the latter, the so-called revisionists, the story of human rights is much more contemporary. As a rhetoric of political activism and moral outrage, the history of human rights really begins in the 1970s. On this view, the UDHR is merely a relic with little relevance for understanding the set of historical and political factors that made the emergence of

human rights both possible and, in a sense, inevitable. (For examples of these debates, see Alston, 2013; Bass, 2010; Glendon, 2001; Hunt, 2007; Moyn, 2010.)

The history of the relationship between anthropology and human rights is marked by a similar debate, not over origins, but over the role anthropologists played during the drafting of the UDHR, the extent to which anthropology as a discipline had a distinct perspective at all on human rights in the aftermath of the Second World War and throughout the early years of the Cold War, and the form and implications of more recent orientations to human rights by individual scholars and professional associations.

The Creation of the UDHR

The devastation of the Second World War formed the backdrop for the effort to create an international system that would once and for all prevent global war, military imperialism, and genocide. Much like the earlier and short-lived League of Nations, the new United Nations system was predicated on a set of universal moral propositions, the most important being the proposition that all human beings are born equal in dignity. As a consequence of this proposition, leading proponents of the postwar settlement thought it necessary to go a step further and draw up a declaration of universal rights that would serve as the framework through which human dignity would be enshrined, promoted, and eventually protected.

An international drafting committee was assembled, chaired by Eleanor Roosevelt. During the contentious process of forging consensus around a working draft of what became the UDHR, the United Nations Educational, Scientific, and Cultural Organization (UNESCO) was asked to help the drafting committee by conducting a comprehensive survey of world opinion on the question of the types of universal norms that should be represented in a declaration of human rights (see, e.g., Morsink, 1999). UNESCO convened a panel of leading philosophers, theologians, writers, and diplomats and this group – which has come to be known as the 'Philosophers' Committee' – designed a lengthy questionnaire, which was distributed to individuals around the world. The responses to this questionnaire were collected and analyzed during a meeting of the panel at UNESCO headquarters in Paris, in 1947. After this meeting, a report was sent to the drafting committee to assist it as it neared completion of a first draft of the UDHR (UNESCO, 1949).

One of the individuals who responded to the UNESCO questionnaire was the American cultural anthropologist Melville Herskovits. (The questionnaires were sent to individuals based on their cultural backgrounds, scientific expertise, and religious affiliations. They were not sent to institutions or representatives of nation-states.) Herskovits participated in both the American Anthropological Association (AAA) and the US National Research Council and was a leading international anthropologist of the time who was strongly influenced by his former professor Franz Boas's theory of cultural particularism. He sent his formal response to the UNESCO questionnaire to the Executive Committee of the AAA, and it then asked the editor of the AAA's scientific journal, *American Anthropologist*, to publish Herskovits's 'Statement on Human Rights' in the journal's 1947 December edition, which he did (see Goodale, 2009). In publishing the Statement, the Executive Committee adopted it on behalf of the AAA without consultation or a vote of the general membership (the AAA was a much different organization in the mid-twentieth century).

The Statement on Human Rights rejected the legitimacy of a UDHR on three distinct grounds: the epistemological, the empirical, and the ethical. First, it argued that the claims of human rights involved abstract moral assertions that could not be scientifically demonstrated by anthropology. It was, therefore, impossible to know whether or not the universal propositions of human rights were true through the techniques of scientific investigation. Second, the Statement suggested that the weight of anthropological research could not be used even for indirect support for the claims of universal human rights. Anthropologists had conducted a wide range of studies of legal and political systems by the middle of the twentieth century, and this research revealed a panorama of cultural diversity. It was simply not possible, according to the Statement, to read this body of anthropological research as evidence of preexisting normative universality. And finally, the Statement dwelled at some length on the broader political economy of human rights and the role of Western powers in promoting particular legal and political approaches to social problems. It argued that regardless of good intentions, an overarching declaration of human rights from the new United Nations would eventually work against the vitality of the local value systems that held together small-scale societies (AAA, 1947).

Regardless of these concerns from at least one prominent member of the discipline that was widely regarded during the mid-twentieth century as the primary source of scientific information about the world's cultures, the Drafting Committee finalized the UDHR, and it was presented to and adopted by the General Assembly on 10 December 1948 (for a general history of the drafting process, see Morsink, 1999).

1948–1980s

The publication of the Statement on Human Rights in *American Anthropologist* led to three brief comments in 1948 and 1949 (Barnett, 1948; Steward, 1948; Bennett, 1949). But from then on, there is scant evidence that anthropologists from any national tradition took an active research interest in human rights for at least several decades. Indeed, it is true enough to say that anthropologists played almost no role in the development of human rights in the initial decades after 1948 and that the discipline had no discernable position toward or interest in human rights, the Herskovits-authored Statement on Human Rights notwithstanding. Nevertheless, it has been common for scholars in more recent years to assert that the Statement on Human Rights represented a widespread anthropological position on human rights at the time and that this early orientation was the source of decades of shame and disciplinary embarrassment (see, e.g., Engle, 2001; Messer, 1993). But archival and historical research revealed that despite active academic and political involvement in the debates over race in the 1950s, the civil rights and antiwar movements in the 1960s, and the antinuclear and environmental movements of the 1970s, anthropologists were largely

silent on the question of human rights and its relation to the discipline until the 1980s (Goodale, 2009).

1980s–1999

The period 1980s was a time of epistemological ferment for anthropology, particularly in the United States, where critical scholars built on earlier interests in political economy and broader Marxist approaches to social theory to scrutinize the institutional orientation of the discipline toward the often marginalized peoples that were its traditional objects of knowledge. Although this crisis of representation in anthropology was more often than not articulated through the language of then-cutting edge social and critical theory, typically continental German or French, the debates at the time had an indirect implication for the anthropology of human rights. The auto-critique of anthropological positivism during the 1980s created cover for the emergence of much more formally political forms of anthropological knowledge, even more so than earlier movements toward Marxist anthropology, which, at least in its well-defined British version, remained largely an innovation confined to anthropological theory.

At the same time, by the 1980s, human rights had become a much more visible framework through which grievances, social struggles, and projects for justice could be articulated. From the global antiapartheid movement to the antinuclear green movements of Western Europe, 'human rights' entered the political and social vocabulary at the same time that the ratification of major international human rights documents accelerated and new human rights laws were drafted. For the history of anthropology and human rights, the most important of these was International Labor Organization Convention No. 169, the so-called indigenous people's 'Bill of Rights.'

As early as the 1970s, anthropologists like David Maybury-Lewis had used their anthropological expertise to advocate for the 'cultural survival' of small-scale populations, many of whom lived in isolated, resource-rich areas that made them vulnerable to exploitation and displacement. Throughout the 1980s, the expanding indigenous rights movement within the broader human rights coalition provided a further opening for anthropologists to become involved politically as advocates and as cultural experts. Reflecting on this shift in the discipline during this period, scholars like Terry Turner argued that human rights became a means through which anthropologists could participate in an 'emancipatory cultural politics' on behalf of indigenous populations (Turner, 1997). This heightened awareness of human rights paved the way for a series of institutional shifts, again, primarily within American anthropology, which included the creation of a standing Committee for Human Rights within AAA (1995) and, most importantly, the drafting of the 1999 'Declaration on Anthropology and Human Rights.'

This official statement of the Association's (if not the discipline's) orientation toward human rights was – unlike the 1947 Statement on Human Rights – approved by a majority of Association members through a general vote. In addition, at a substantive level, the 1999 Declaration was an explicit and dramatic repudiation of Herskovits's 1947 Statement. Not only does the Declaration recognize the cross-cultural validity of human rights, but it also maintains that anthropologists have a disciplinary obligation to use their research expertise to expand the definition and meanings of human rights even beyond those established in current international law (AAA, 1999). Following the promulgation of the Declaration, the Association became increasingly active in human rights advocacy on behalf of vulnerable populations and victims of human rights abuses (including academics) around the world.

In the years leading up to the 1999 Declaration, some of the same activist-anthropologists who played key roles in the drafting of the Declaration and its promotion within the discipline prepared a special 1997 issue for the *Journal of Anthropological Research* that laid the intellectual and disciplinary groundwork for this new orientation and understanding of anthropology's responsibilities toward human rights advocacy (see the articles by Hatch, Messer, Nagengast, Turner, and Zechenter). Among other things, these essays reflected a certain defensiveness by some anthropologists about the way in which concepts such as cultural relativism had come to be associated with justifications for human rights abuses. Despite the fact that the danger that human rights advocacy could devolve into cultural imperialism "chill[ed] the anthropological soul," as Nagengast argued in her contribution (1997: 349), the path had been cleared – both institutionally and intellectually – for a new phase in anthropology's engagement with human rights.

1999–2007

At the same time when anthropologists were developing institutional and epistemological frameworks that made anthropology newly relevant to human rights advocacy, and vice versa, another, quite different process of historical importance was taking place. The end of the Cold War was a geopolitical watershed moment for the broader international human rights movement. The dissolution of the Realpolitik logics of the Cold War conflict created a vacuum in international relations in which new (or recreated) transnational forms of moral practice and political struggle could emerge. In just over a decade, human rights discourse saturated everything from constitutional reform to international development projects. By the mid-1990s, the rhetoric of human rights had already become the "only legitimate language of democratic transition," as Richard A. Wilson put it (2001: 1); many countries, particularly in Latin America, had implemented international human rights norms through wide-ranging domestic reforms in women's health, bilingual education, and rural development; and a movement to create a permanent international criminal court to prosecute gross human rights violations culminated in the adoption of the Rome Statute by the UN General Assembly.

Even as some anthropologists participated in this proliferation of human rights as advocates, cultural experts, and witnesses, other scholars responded by developing methodologies of critical analysis that coalesced into a proper anthropology of human rights. Here, ethnographers studied the close details of human rights advocacy and political reform across an ever-widening range of regional and social scale. The critical ethnography of human rights was largely agnostic about the social, political, and moral claims of the broader human

rights project that it documented across a rich cross-cultural panorama. Important examples of this first extended period in the anthropology of human rights would include Richard A. Wilson's study of human rights in the creation of a postapartheid South Africa (2001); the edited volume on culture and rights that collected early essays that were the fruits of the first sustained ethnographic study of human rights in the post-Cold War period (Cowan et al., 2001); Sally Engle Merry's (2006b) innovative multisited and transnational ethnography of the system used to monitor compliance with the Convention on the Elimination on All Forms of Discrimination Against Women (CEDAW); Harri Englund's provocative study of the ways in which the hegemony of human rights discourse made Malawians 'prisoners of freedom' (2006); and Winifred Tate's ethnographic account of the 'culture and politics' of human rights advocacy in conditions of civil war, postconflict negotiation, and transnational criminality (2007).

As a matter of disciplinary recognition and development, the emergence of the anthropology of human rights as a mode of critical reflection on the rise and an increasing hegemony of human rights around the world (which included the individual and institutional participation of anthropologists as activists) culminated in both a major 2006 retrospective in *American Anthropologist* entitled 'Anthropology and Human Rights in a New Key' (with articles by Cowan, Goodale, Merry, Riles, Speed, and Wilson) and a volume of ethnographic essays on the 'practice of human rights' (Goodale and Merry, 2007) that would come to shape debates on human rights outside of anthropology and in the field of international human rights itself.

2007–Present

The anthropology of human rights has continued to mature and play a part in wider debates in the discipline from the late 2000s to the present. Evidence for this can be found in the number of doctoral dissertations and junior research projects that are self-consciously located within an anthropology of human rights; the growth in conferences, association panels, and workshops in this new subfield; and in the move by anthropology departments and interdisciplinary fields to develop specialisms that emphasize the anthropological study of human rights.

The next-generation ethnographies of human rights practices build on the theoretical and disciplinary foundation established during the 1990s and 2000s. They are marked by increasing levels of methodological and epistemological innovation, as human rights discourse becomes further naturalized within what James Ferguson (2006) called the 'neoliberal world order.' Among the more notable of these recent works are Harri Englund's ethnography of human rights and popular media in Malawi (2011); Galit Sarfaty's study of the fight by human rights lawyers to change the internal culture of the World Bank (2012); Kamari Maxine Clarke's ethnographic critique of the 'fictions of justice' that reinforces the legitimacy of the International Criminal Court (2012); Karen Faulk's ethnography of tensions between neoliberal citizenship and human rights advocacy in Argentina; and Lori Allen's account of the way a culture of cynicism shapes human rights activism in occupied Palestine (2013).

Major Topics

As an emergent and increasingly influential subfield of social and cultural anthropology, the anthropology of human rights has made contributions to theory and methodology in a number of different areas. It has also made its mark outside of anthropology and the academy as international lawyers, policy-makers, and government officials turn to the anthropology of human rights as a source of information and ideas about human rights in policy documents, reports to governing agencies, and in major public speeches (see, e.g., Higgins, 2012, 2013).

Culture

As a category for organizing difference, culture was problematic for human rights from the very beginning. In his 1947 Statement, Herskovits worried at length that existing cultural diversity and richness were not compatible with countervailing universalisms like human rights, which asserted transcultural patterns of normative belief and practice that both flew in the face of anthropological evidence and threatened to homogenize culture within the broader political economy of the postwar settlement.

But despite these concerns, cultural diversity was not affected by the promulgation of the UDHR in 1948, to a large extent, because the broader influence of human rights, not to mention the kind of global transformation imagined by Herskovits, stalled out almost immediately within the Cold War. Even Eleanor Roosevelt herself recognized that the project of human rights would be years, even decades, in the making, and that cultural patterns would not begin to shift until what she called a 'curious grapevine' could take root and spread to places where "governments are not so anxious for it" (quoted in Korey, 1998). By the time this kind of large-scale change did begin, anthropologists of human rights observed that the impact on culture defied stark dichotomies between particularity and universality, and culture itself facilitated the transnationalization of human rights norms.

In the groundbreaking volume *Culture and Rights* (2001), anthropologists reflecting on a decade of ethnographic research after the end of the Cold War noticed two key phenomena at work. In the first, existing cultural practices and modes of normative understanding served as mechanisms for what Sally Engle Merry (2006a) would later describe as 'vernacularization': the translation of increasingly hegemonic international human rights norms into forms of ethical and political practices rooted in the particular. And in the second, scholars in *Culture and Rights* described and theorized the emergence of a transnational culture *of* human rights. Here, an anthropological account of culture proved useful in understanding the formation of new categories of collective action within international agencies, transnational NGOs, and in political and social movements shaped by the logics of human rights. In both, the use of the culture concept by anthropologists of human rights to understand human rights praxis moved against the grain of wider trends in anthropological and social theory, which had become skeptical of the analytical utility of culture in the face of the supposed hybridization of globalization.

Power

Human rights, as Burke might have said, acting through people, is power; and "considerate people, before they declare themselves, will observe the use which is made of *power*; and particularly of so trying a thing as *new* power in *new* persons, of whose principles, tempers, and dispositions they have little or no experience" (Burke, 1910[1790]: 7, emphasis in original). Anthropologists of human rights have been keenly attuned to another problem originally identified by Herskovits: the ways in which a global project of human rights creates heightened tensions at points of existing conflict by overwhelming alternative forms of social change and dispute resolution. Of course, from the perspective of human rights advocacy, this is power exercised for the good; indeed, as expressed most forcefully in international treaties like CEDAW, the postwar human rights project demands change, replacement, even the suppression of modes of cultural practice that remain unaccountable and, therefore, illegitimate. As CEDAW's oft-cited Article 5 puts it, states parties to the international women's bill of rights shall "modify the social and cultural patterns of conduct of men and women, with a view to achieving the elimination of prejudices and customary and all other practices" that are based on local theories of gender inequality.

But as anthropologists have shown, there are often costs to the ways in which human rights tends to bracket and then marginalize other cultural logics of social justice, development, conflict transformation, and public ethics. And this extension of power can have unintended consequences. One of the more far-reaching examples of the way anthropologists have explored the implications of the unpredictable power of human rights is Harri Englund's (2006) ethnography of development in Malawi. As he explains, the concept of human rights was officially translated in the local language with a phrase that meant "the freedom one is born with" (2006: 51). Over time, people took this emphasis on human rights-as-freedom to mean the freedom to challenge existing cultural norms around fashion, obedience in public schools, and sexual behavior, rather than the structural, economic, and political conditions that reinforced a legacy of inequality and public corruption. The result, according to Englund, was that Malawians were eventually "disempower[ed] through translation." Human rights discourse saturated every aspect of public life in Malawi, which is what government officials and transnational aid workers sought. But as human rights were (mis-)rendered in a local moral vernacular, it was transformed beyond recognition in ways that foreclosed its use as a language of much needed social change.

Epistemology

When Herskovits argued that anthropology was not able to make definitive claims about universal human rights because it was a "science of mankind" and, therefore, concerned only with empirical questions of human behavior as expressed through "patterns of culture," he could not have foreseen the epistemological innovations in the discipline that would expand its objects of knowledge and transform its domains of inquiry. This is not to say, however, that in later decades anthropologists brushed aside Herskovits's early arguments in order to confront the basic ontological and philosophical problems that remained essential to human rights. Much of the intellectual labor devoted to human rights remained within spheres like critical legal studies, political theory, and moral philosophy. Rather, anthropologists used ethnographic research to make a subversive case for broadening the grounds on which the fundamental moral and theoretical questions of human rights could be asked and answered. This, in turn, carried important implications for the epistemology of human rights, especially in the post-Cold War, when human rights discourse become increasingly interwoven into everyday forms of legal, political, and social practices.

Anthropologists of human rights observed early on that ideas about human rights were fundamental to human rights praxis. Ordinary social actors, often caught up in moments of crisis or dislocation, were never able to simply harness human rights instrumentally or co-opt its imaginaries of justice as if it were a normative toolkit waiting to be opened when needed. Instead, the logics of human rights advocacy demanded self-consideration as much as social change; people were invited, encouraged, compelled to remake themselves as citizens of a different moral universe. Ethical theorizing in terms of this often radically alternative moral universe became a distinct form of social practice, and the anthropologist became both a witness to and co-participant in this transformation as part of the ethnographic encounter (see Goodale, 2006).

What resulted was an ethnographic record of innovative, and potentially transformative, folk models of human rights that were deeply embedded in the circumstances of their creation. Perhaps, the best example we have of a fully articulated local account of human rights is Shannon Speed's ethnography of the aftermath of the Zapatista rebellion in Chiapas (2007). Both during and after the violence, international and transnational human rights organizations flooded into the Chiapas region. Those promoting indigenous rights-as-human rights were particularly influential in shaping the way Zapatista's resistance was expressed. Indigenous political leaders formed 'good governance councils' in which the ideas of human rights were debated at length, remade, and then used to represent Zapatista moral values as these grounded Zapatista political action.

Transnational Advocacy

The transnational human rights networks that emerged after the end of the Cold War did precisely what Eleanor Roosevelt had expected them to do: they challenged the sovereignty of the state and made it possible to create new public spheres that were both translocal and firmly rooted in sites of intimate contestation. Scholars like Annelise Riles (2000) studied these networks from the 'inside out' and helped shape the broader understanding of the global assemblages that were changing the ontology of social relations during a time of geopolitical and global economic transformation.

But anthropologists of human rights also revealed the extent to which networks of human rights advocacy are shaped by the political economies of local conflict in ways that can change their normative valence and render them, in certain circumstances, incapable of fulfilling their broader mandate for social change and moral transformation. For example, Winifred Tate's (2007)

long-term ethnography of the ongoing conflict between the Colombian state and the Revolutionary Armed Forces of Colombia (FARC) shows how human rights activists struggle to translate the language and moral logics of universal human rights into a category of instrumental action, one that can respond to the challenges of historical trauma, multiple and ambiguous accounts of culpability for atrocity, the legacy of structural violence, and enduring patterns of economic inequality with roots in the colonial period.

Also Sally Engle Merry's (2006b) study of the institutions that monitor national compliance with CEDAW illustrates in rich detail the ways in which human rights activists themselves must navigate multiple and not always overlapping cultures of advocacy and resistance. Representatives of national human rights ministries often find themselves in the position of having to advocate at the same time for both compliance with an international human rights treaty and the integrity and legitimacy of cultural practices that would seem to violate the treaty. Nevertheless, these kinds of dichotomies do not negate the reach of international human rights law into national and local conflicts. Instead, as Merry argues, they reflect the ways in which the practice of human rights creates its own categories of contested identity and power with uncertain implications for both transnational human rights advocacy and the promotion of national(-ist) heritage.

Critique and Engagement

Finally, the anthropology of human rights, perhaps more than other academic orientations toward human rights, grapples uneasily with the dilemma of how to develop a rigorous and ethnographic account of human rights that is both critical and ethically attuned to the conditions of vulnerability that lead to abuse and exploitation. This tension has expressed itself in different ways for different anthropologists, some of whom (like Winifred Tate and Shannon Speed, for example) began their careers as human rights activists before making the social and political processes they were committed to the objects of ethnographic research and reflection.

However, the tension between critique and engagement, skepticism and advocacy, and resistance and commitment is not only a challenge for anthropologists of human rights. As ethnographic research has demonstrated, it is a fundamental social and moral fact for the practice of human rights itself. This is, in part, because as a working theory of both social practice and political change, the idea of human rights demands a form of self-reflection and self-constitution that is meant to sow doubt – about existing cultural practices, about folk theories of the person, about hierarchies of power. Yet, the transition from the old to the new – at once illegitimate authentic – is fraught with moral slippage and unanticipated consequences.

An important recent example from the ethnography of human rights practice is Lori Allen's (2013) study of the role of human rights discourse within the politics of Palestinian resistance to Israeli occupation of the West Bank. Although the language of human rights was used as early as the late 1970s in Palestine as a grassroots rhetorical strategy to advocate for victims of the occupation to an international audience, a professional cadre of activists and NGOs eventually came to restrict the use of human rights within tightly controlled social and political spaces. At the same time, the set of Palestinian grievances went unmet for decades, including the continued violation of human rights, the failure to achieve political independence, and an inability to shift favorably political opinion within Israel.

The result was that ordinary Palestinians came to view the claims of human rights with cynicism and even suspicion. But rather than rejecting human rights entirely, locals within the struggle made an organic critique of human rights part of a broader critical and emancipatory discourse in support of Palestinian autonomy, anti-imperialism, and grassroots (as against interventionist) advocacy. With decades of engagement with human rights part of the history of the Palestinian struggle against occupation, activists were able to appropriate, or push against, the logics and expectations of human rights with a high degree of contextual awareness and political realism.

Future Directions

The anthropology of human rights is now well established as a distinct area of research and source of anthropological theory. Institutionally, scholars and graduate students who work in the anthropology of human rights commonly, but not exclusively, come from the ranks of legal and political anthropology. Because human rights has become an increasingly pervasive mode of contemporary world-making, anthropologists encounter traces of this influence across a wide range of cultural practices, political movements, and moral projects.

This is not to say, however, that the status of human rights is uncontested – quite the contrary. As the liminal period of the post-Cold War gives way to cultural re-differentiation, the establishment of new hierarchies, and the narrowing of spaces of political and social experimentation, human rights will continue to jostle with alternative forms of moral praxis and collective- and self-constitution. While the postwar project of human rights matures through its translation into mundane, even banal, processes of constitutional reform, good governance, and neoliberal economic restructuring, its potential for catalyzing radical transformation and moral upheaval will likely diminish. The anthropology of human rights will become less the study of a political and moral discourse during a time of often giddy transition and seemingly limitless possibility, and more the study of an established contemporary secular universalism among a host of competitors.

See also: Advocacy in Anthropology; Borders, Anthropology of; Citizenship, Anthropology of; Colonialism, Anthropology of; Common Sense, Anthropology of; Conflict: Anthropological Aspects; Cosmopolitanism; Cultural Relativism, Anthropology of; Nongovernmental Organizations: Anthropological and Historical Aspects; Refugees in Anthropology; Values, Anthropology of; Violence in Anthropology; War: Anthropological Aspects, Historical Development of.

Bibliography

Allen, Lori, 2013. The Rise and Fall of Human Rights: Cynicism and Politics in Occupied. Stanford University Press, Stanford (Stanford Studies in Human Rights).

Alston, Philip, 2013. Does the past matter? On the origins of human rights. Harvard Law Review 126, 2043–2081.

American Anthropological Association, 1947. Statement on human rights. American Anthropologist 49 (4), 539–543.

American Anthropological Association, Committee on Human Rights, 1999. Declaration on Anthropology and Human Rights. http://aaanet.org/about/Policies/statements/Declaration-on-Anthropology-and-Human-Rights.cfm (accessed 14.10.14.).

Barnett, H.G., 1948. On science and human rights. American Anthropologist 50, 352–355.

Bass, Gary J., 11 November 2010. The old new thing. review of Samuel Moyn's the last Utopia. The New Republic 35–39.

Bennett, John W., 1949. Science and human rights: reason and action. American Anthropologist 50, 329–336.

Burke, Edmund, 1910[1790]. Reflections on the French Revolution. J.M. Dent & Sons, London.

Clarke, Kamari Maxine, 2009. Fictions of Justice: The International Criminal Court and the Challenge of Legal Pluralism in Sub-Saharan Africa. Cambridge University Press, Cambridge.

Cowan, Jane, Dembour, Marie-Bénédicte, Wilson, Richard A. (Eds.), 2001. Culture and Rights: Anthropological Perspectives. Cambridge University Press, Cambridge.

Engle, Karen, 2001. From skepticism to embrace: human rights and the American Anthropological Association from 1947–1999. Human Rights Quarterly 23, 536–559.

Englund, Harri, 2006. Prisoners of Freedom: Human Rights and the African Poor. University of California Press, Berkeley.

Englund, Harri, 2011. Human Rights and African Airwaves: Mediating Equality on the Chichewa Radio. Indiana University Press, Bloomington.

Ferguson, James, 2006. Global Shadows: Africa in the Neoliberal World Order. Duke University Press, Durham.

Glendon, Mary Ann, 2001. A World Made New: Eleanor Roosevelt and the University Declaration of Human Rights. Random House, New York.

Goodale, Mark, 2009. Surrendering to Utopia: An Anthropology of Human Rights. Stanford University Press, Stanford.

Goodale, Mark, 2006. Ethical theory as social practice. American Anthropologist 108 (1), 25–37.

Goodale, Mark, Merry, Sally Engle (Eds.), 2007. The Practice of Human Rights: Tracking Law between the Global and the Local. Cambridge University Press, Cambridge.

Higgins, President Michael D., 27 September 2013. The human rights discourse: some issues of source and prospects for achievement. In: The Annual Human Rights Lecture to the Human Rights Committee of the Law Society of Ireland. http://www.president.ie/speeches/6664-2/ (accessed 20.11.13.).

Higgins, President Michael D., 10 December 2012. The human rights discourse: its importance and its challenges. In: The Human Rights Commission's Annual Lecture, International Human Rights Day. http://www.president.ie/speeches/the-human-rights-discourse-its-importance-and-its-challenges-by-president-michael-d-higgins-the-human-rights-commissions-annual-lecture-international-human-rights-day-monday/ (accessed 20.11.13.).

Hunt, Lynn, 2007. Inventing Human Rights: A History. Norton, New York.

Korey, William, 1998. NGOs and the Universal Declaration on Human Rights: "A Curious Grapevine". St. Martin's Press, New York.

Merry, Sally Engle, 2006a. Transnational human rights and local activism: mapping the middle. American Anthropologist 108 (1), 38–51.

Merry, Sally Engle, 2006b. Human Rights and Gender Violence: Translating International Law into Local Justice. University of Chicago Press, Chicago.

Messer, Ellen, 1993. Anthropology and human rights. Annual Review of Anthropology 22, 221–249.

Morsink, Johannes, 1999. The Universal Declaration of Human Rights: Origins, Drafting, Intent. University of Pennsylvania Press, Philadelphia.

Moyn, Samuel, 2010. The Last Utopia: Human Rights in History. Belknap Press of Harvard University Press, Cambridge.

Nagengast, Carole, 1997. Women, minorities, and indigenous peoples: universalism and cultural relativity. Journal of Anthropological Research 53 (3), 349–369.

Riles, Annelise, 2000. The Network Inside Out. University of Michigan Press, Ann Arbor.

Sarfaty, Galit, 2012. Values in Translation: Human Rights and the Culture of the World Bank. Stanford University Press, Stanford (Stanford Studies in Human Rights).

Speed, Shannon, 2008. Rights in Rebellion: Indigenous Struggle and Human Rights in Chiapas. Stanford University Press, Stanford.

Steward, Julian, 1948. Comment on the statement on human rights. American Anthropologist 50, 351–352.

Tate, Winifred, 2007. Counting the Dead: The Culture and Politics of Human Rights Activism in Colombia. University of California Press, Berkeley.

Turner, Terence, 1997. Human rights, human difference: anthropology's contribution to an emancipatory cultural politics. Journal of Anthropological Research 53, 273–291.

UNESCO, 1949. Human Rights: Comments and Interpretations. Columbia University Press, New York.

Wilson, Richard A., 2001. The Politics of Truth and Reconciliation in South Africa. Cambridge University Press, Cambridge.

Human Rights, History of

Antoon De Baets, University of Groningen, Groningen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, six basic debates about human rights are clarified from a historical perspective: the origin of human rights as moral rights connected to the natural law doctrine and opposed to positive rights; the wave of criticism of their abstract and absolute character by nineteenth-century liberal, conservative, and socialist thinkers; their extension from the rights of man to the rights of all human beings and from individual rights to individual and group rights; the tension between national and international control over the protection of human rights and between domestic and international jurisdiction; the controversy over the indivisibility of the three generations of rights; and, finally, the problem of the universal or relative character of human rights as viewed from historical and anthropological angles. The article also tries to answer the questions of whether there has been more progress than failure in the field of human rights and of whether the pressure of inhumanity around us is necessary to trigger human rights progress.

The expression 'history of human rights' has two different meanings. It can be understood (as it usually is) as the history of the concept of human rights, or (less frequently) as the history of the enjoyment of human rights (and their abuse) in practice. Therefore, the basic debates about the concept of human rights are first clarified here from a historical perspective. Then, an attempt is made to answer the questions of whether the history of human rights has been characterized by success or failure and of whether failure triggered progress.

Moral versus Positive Rights

Developing the human rights concept took two millennia in Western history and was so full of unforeseen links, ruptures, reversals, and coincidences, that, until only a few decades ago, it convincingly could be called a unique product of that history. One symbolic pivotal point in this development was 1689–1690, just after the Glorious Revolution of 1688. In these years, the English philosopher John Locke (1632–1704) published his *Two Treatises on Government*, which contained the first classical formulation of human rights (then still called 'natural rights'). In retrospect, this work stood at the juncture of a set of amazingly diverse ideas, some of which went back to antiquity and some of which will survive through the twenty-first century. Locke's point of departure was the doctrine of natural law as it had developed since antiquity and had been transformed, in certain respects almost beyond recognition, by medieval Christianity and early-modern philosophers. According to Locke, the doctrine maintained that human nature endowed all humans with certain basic moral entitlements (such as the rights to life, liberty, and property), to which the manmade laws of society, both positive and customary, had to conform. These entitlements were called natural rights and assumed to possess five characteristics: they were discoverable by reason, itself a faculty of human nature; they existed before humans entered society; they applied to all humans by virtue of their being human; they were inalienable (i.e., could not be taken away); and they were restricted by the recognition of the rights of others only. As such, these rights limited the power of the society and its political and legal manifestation, the state. Respect for them was the foremost test for the legitimacy of political power.

The natural law doctrine had antecedents. Ancient ideas on rationalism, equality, and the restrained exercise of power formed its substratum. Despite its strong claim of immutability, however, many of its components, such as the exact catalog of moral principles or the source of authority for these principles, turned out to be rather flexible. Although temporarily fallen into eclipse because of the ascendancy of the sixteenth-century absolutist state, the natural law doctrine would be fostered by three developments initiated in that very period: with the Renaissance came an emphasis on individualism and secularism; the outcome of the religious struggles following the Reformation taught the necessity of tolerance; and the seventeenth-century scientific revolution marked the triumph of rationalism. These three developments reversed the philosophical priorities of the hierarchical medieval societies. Natural law gradually became associated with rights rather than duties and with individuals rather than groups.

The perception of the source from which these natural rights emanated also changed. Whereas Christianity had maintained that God was the ultimate source and political absolutism had invested the monarch with divine rights, the consent of the governed came to be seen as the final authority now. Individuals gave this consent in a social contract, which recognized that they possessed natural rights before entering society. Governments owed their origin to that contract, by which free and equal citizens agreed to entrust to a neutral sovereign body some of their nonnatural rights, to accept its authority and rules as long as it protected them, and to rebel against it if it did not. This became the basis for the rule of law and the separation of, and balance between, the legislative, executive, and judicial functions of government.

Although previous political theories paved the way and later ones expressed similar ideas, it was Locke's version that became iconic, especially outside England. In the age of Enlightenment, it exerted profound influence on the politicians behind the great democratic revolutions at the end of the eighteenth century. In the United States and France, natural rights, mainly of a political and civic character, were incorporated into the declarations of the revolutionaries. In 1789, natural rights were

transformed into the 'rights of man' (*droits de l'homme*). This marked the beginning of a slow process of codification of rights within the constitutions of most states and in international law that still has not ended.

This brief overview shows that, although many aspects of the human rights conception have an ancient pedigree, other crucial aspects have not. The gradual appearance of the concept was the result of a unique historic combination of particular series of events and philosophical thought reflecting upon them. The natural law doctrine itself eventually knew a curious fate: with unprovable tenets and with the possibility of abusive claims 'in the name of human nature,' it became overshadowed by the powerful state-centered thinking of the nineteenth century (as it had been in the sixteenth), but it survived. It was rehabilitated in a modest form during World War II. But the present approach sees human rights as an operationalization of the basic concept of human dignity rather than of human nature.

Absolute versus Relative Rights

The debate on moral versus positive rights was continued by liberal, conservative, and socialist critics of human rights, typically in such works as those of Jeremy Bentham (1748–1832), Edmund Burke (1729–1797), and Karl Marx (1818–1883). These critics argued against the abstract and absolute character of human rights and maintained that rights had to be connected to the society in which they historically were to be exercised. The liberal utilitarian Bentham believed only in the force of positive legislation. For him, natural rights were imaginary, 'nonsense upon stilts.' He feared that they were powerful rhetoric in the hands of rulers and a substitute for effective legislation. Unlike Bentham, the conservative Burke did not deny the existence of natural rights. He feared these rights, however, for reasons contrary to Bentham's: in the hands of the common people, natural rights would stimulate revolutionary sentiments and cause social upheaval leading to terror and to the destruction of the social fabric. Rights (and the duties linked to them) ought to be derived from the principles and values emanating from the particular traditions of a society. For the socialist Marx, the human rights doctrine was not radical enough. It left the inequalities generated by capitalism untouched (particularly private property) and therefore served the interests of the ruling bourgeois class. Because of these inequalities, the workers were unable to enjoy their rights fully, which therefore remained largely formal. As a class, the workers had a duty to create the conditions for social change by claiming collective social and economic rights with which to overturn the exploitative economic system. The new, egalitarian society to be established after the struggle had been successful, would be one without a state, laws, or rights.

These three critics did not stand alone, but their arguments gained a classic character and inspired powerful currents. Bentham's ideas gave impetus to the tradition of legal positivism. Ideas similar to Burke's would feed the romanticism and historicism of the nineteenth century to the extent that they also emphasized cultural identity and diversity. Marx's concerns activated the thinking about social and economic rights, heightened awareness of the need for social legislation, and influenced the 1917 Russian Revolution.

Men versus All Humans and Individuals versus Groups as Rights-Holders

Clearly, until long after its introduction, the concept of 'rights of man' covered only restricted categories of humans, usually male tax-paying citizens. The large majority of women, slaves, and foreigners were excluded. The struggle to include these categories into the ranks of rights-holders would lead to the final transformation of the concept into 'human rights.' As early as 1791, Olympe de Gouges campaigned for women's rights. With the movement against slavery, another group followed. Categories of individuals deemed particularly vulnerable (next to women and slaves, also children, aliens, refugees, prisoners, the disabled) eventually were protected by special regulations. Likewise, the difficult problem of how to realize group rights was introduced at an early stage in the human rights debate. Groups deemed vulnerable (minorities, colonized peoples, indigenous peoples) increasingly received special attention. Specific treaties drafted for national, ethnic, or religious minorities came into effect long before general treaties. The League of Nations, founded in 1919, did not formally recognize human rights but added pioneering treaties for minorities, refugees, and mandate territories.

From the movement to enlarge the concept of rights from selected human beings to all human beings, two further questions can be derived. The first of these is what the consequences are of the anthropocentric character of human rights, as it excludes other sentient beings. The other problem is whether past and future generations have human rights. As to past generations, it is difficult to ascribe rights to the dead, who are not human beings, but *past* human beings. Nevertheless, in this context, one could think of the duties of the living toward the dead concerning the dead body, the funeral, and the grave; concerning the will, the identity, and the heritage of the deceased; and concerning the images and texts about them. As to future generations, the preservation of the cultural heritage of humanity and of the natural environment is a chief concern as is the relationship between human rights and the human genome. As to future generations in relation to past generations, the accountability of successor governments for human rights violations committed by their precursors and their duty to investigate and prosecute these violations – with its corollary, the right to the truth – so as to combat impunity, became the focus of a new field of study: transitional justice (the full range of processes and mechanisms associated with a society's attempts to come to terms with the legacy of large-scale past abuses to ensure accountability, serve justice, and achieve reconciliation).

Domestic versus International Jurisdiction over Human Rights

Since 1800, two tendencies, one international, the other national, partly opposite, partly complementary, have dominated the debate over who was to supervise the protection of

human rights. A process of nationalization started when human rights were incorporated into the constitutions of growing numbers of countries. At the same time, human rights became part of the agenda of intergovernmental meetings, leading to the humanitarian interventions of the nineteenth century (continuing a tradition dating from the seventeenth) and the gradual incorporation of human rights into international law. The decisive historical moment for internationalization, however, came during World War II, when the link between human rights violations and war (hence, *a contrario*, the link between human rights protection and international peace and security) became clearly visible and, shortly after, when the Nazi and Japanese atrocities led to worldwide condemnation embodied in the international war crimes tribunals of Nuremberg and Tokyo. Although the United Nations only paid secondary attention to human rights in their 1945 Charter, this did not prevent work on an International Bill of Human Rights, consisting of a morally compelling but not legally binding Universal Declaration of Human Rights, adopted in 1948, and two International Covenants, one on civil and political rights and the other on economic, social, and cultural rights, both approved in 1966 and entered into force in 1976.

The idea of self-determination appealed to peoples under colonial rule or comparable alien subjugation and proved to be a crucial concept. It was a dangerous concept also because narrowing self-determination to only one of its possible interpretations – political independence, hence secession – was potentially explosive and because consensus was lacking as to who the rights-bearer ('a people') exactly was. Self-determination was promoted to a group right only in the 1966 covenants.

In the efforts to transform human rights into positive rights by making them enforceable in positive law, usually three stages are distinguished: standard-setting (codification of standards), ratification (making standards binding), and implementation (enforcing standards via sanctions). Seminal steps for enforcing standards were the elaboration of complaint mechanisms for state and nonstate actors (mainly since the 1970s), the appointment of a High Commissioner for Human Rights (1993), and the adoption of the Statute of an International Criminal Court with universal jurisdiction to prosecute alleged perpetrators (including heads of state and government) of genocide, crimes against humanity, war crimes committed in international and internal conflicts, and crimes of aggression (1998). Important preparatory work for these new devices was done by a coalition of nongovernmental organizations, the representatives of civil society. The elaboration of human rights institutions and mechanisms on a world scale was paralleled, and often preceded, by similar developments on a regional scale in Europe, America, Africa, and the Arab world.

In the midst of this evolution, the state was the key player. The cherished principles of the Westphalian world order, established in 1648, were national sovereignty and interstate equality. Their corollary, domestic jurisdiction, excluded interference from abroad in the internal affairs of the state. The vigorous doctrine of nationalism and successive waves of decolonization since the late-eighteenth century strengthened these principles. Paradoxically, recognition of national sovereignty made possible progress in international law with its core principle of interference in the internal affairs of states, for two reasons. First, states slowly realized that when they adopted human rights as the constitutional basis of the rule of law, they implicitly created a duty to protect them everywhere, including beyond their borders. Second, it became painfully obvious that the states charged with the protection of human rights very often were the main perpetrators of human rights violations. As a consequence, states eventually began to accept, although reluctantly, an international monitoring element that restricted their sovereign domestic power. The number of human rights campaigns and humanitarian interventions increased drastically in the 1990s, but their enabling conditions, forms (diplomatic, political, economic, or military), potential abuse, and effect were hotly disputed.

First- versus Second- versus Third-Generation Human Rights

As the list of human rights gradually expanded, analysts grouped them into 'generations.' First-generation rights, tied to the idea of liberty, were traditional or individual rights of a political and civic nature. They were characterized as negative rights ('freedoms from') because they favored government abstention. Second-generation rights, tied to the idea of equality, were modern or collective rights of an economic, social, or cultural nature. They were characterized as positive rights ('rights to') because they favored government intervention. Third-generation rights, tied to the idea of fraternity, were modern or collective rights of a global nature. They were characterized as solidarity rights and usually included the established right of self-determination and the newer rights to development, human security, peace, and democracy. The rights of indigenous peoples and the right to a life-sustaining environment frequently are added, but these also are listed as fourth-generation rights by some.

Each generation of rights emerged in a particular historical context. The first generation was the result of the so-called democratic revolutions of the late-eighteenth century. Increasing political emancipation and democratic participation in the next two centuries gave them further impetus. The second generation developed in the nineteenth century as the socialist tradition with its reformist and revolutionary strategies introduced ideas of welfare and social legislation to counter the exploitation of working classes. The third generation usually is associated with demands from colonial peoples, and after their independence, from the South for a worldwide redistribution of power and wealth, and with the delicate balance between global and local trends.

At the 1993 World Conference on Human Rights in Vienna, the second such conference ever and the first after the Cold War, the indivisibility of human rights was reaffirmed strongly. Even so, seven rights acquired a special status as they were the only ones stipulated by the 1966 covenants to be nonderogable, even in times of public emergency. They are the rights to life; to freedom of thought, conscience, and religion; to be recognized as a person before the law; and the prohibitions against torture and slavery, against imprisonment for nonfulfillment of contractual obligations, and against retroactive application of criminal laws.

The bewildering diversity of human rights nevertheless triggered much controversy on their compatibility and rank. Because all human rights were conceived as inalienable, they sometimes came into mutual conflict. A classic example of a conflict within the same generations of rights is the tension between free expression and privacy: both have to be balanced according to fair rules. But most conflicts occurred between, rather than within, generations of rights. Whereas first-generation rights assumed the character of universally enforceable standards, second- and third-generation rights were rather aspirations of the fulfillment which was utterly dependent on the creation and application of expensive state programs. Faced with all these demands, some states, mainly in the West, gave priority to first-generation rights, whereas others, mainly in the East and the South, campaigned, if not for the primacy, then at least for the equivalence of second- and third-generation rights to those of the first generation.

The strongest debates were held on whether political freedom should take priority over economic development or the reverse. During the Cold War (and even beyond), those defending the priority of political freedom (free minds) were accused of perpetuating inequality, those defending the priority of economic development (full bellies), of perpetuating dictatorial government. The hypothesis that political freedom or democracy has to be suspended until satisfactory levels of economic development are attained is contradicted by experience: in such cases, freedom usually is trampled underfoot, with no economic development attained after all. Unaccountable governments generally exacerbate economic and environmental crises. Moreover, how can a present suspension of freedom ever be conceived as a suitable preparation for its future protection? The reverse hypothesis, that economic development and political participation are interdependent, seems to have firmer ground: the exercise of civil and political rights is dependent to a certain degree on the fulfillment of economic and social rights (such as the rights to food and to education) and, in its turn, it is a requirement for citizens to participate actively in the economic development of the society. In short, there is no evidence for a conflict between political rights and economic performance. Research by the United Nations Development Program (UNDP) demonstrated that democracies are better than dictatorships in at least three crucial areas: managing conflicts; avoiding catastrophes and managing sudden downturns that threaten human survival; and helping spread the word about critical health issues.

This brings us to the major question of the relationship between human rights and political systems. Beginning with the Universal Declaration of Human Rights, all international human rights instruments advocate a democratic society as the best political system to protect human rights. In addition, the preamble of the Universal Declaration strongly condemns the dictatorships of the past. From this, the United Nations infer a simple definition of a democratic society: it is a society that recognizes and respects the human rights set forth in the Universal Declaration of Human Rights. This definition, simple as it seems, is strong because it combines the minimalist conception of democracy as the rule of the majority with a maximalist human rights-oriented application of the rule of law. Although, strictly speaking, no state in the world lives up to such an ideal definition, it points to the fact that a democracy is the best framework to facilitate proper conditions for the protection of human rights. In any case, human rights and democracies both emerged historically as the results of the 'democratic revolutions' of the late-eighteenth century.

Universality versus Relativity of Human Rights

The history of human rights expounded earlier, pointed to the Western roots of the human rights concept. Elements of the concept, such as some natural law principles, were available outside the West, but the combination of factors that led to its formulation was unique. As the West conquered the rest of the world, it introduced imperial dominance and violence. At the same time, however, its notions of nationalism, freedom, and human rights – developed initially for home consumption – fell on fertile ground around the globe: they proved able to transcend their particular roots and context. In the hands of colonized peoples, these notions echoed similar indigenous values and increasingly served to justly criticize Western exploitation in the language of the West itself. Thus, the claim of the universality of human rights has an anthropological basis, not a historical one: it is based on the worldwide appeal of a specific system of thought. In an unambiguous sense, this is logical because human dignity, as the foundational idea behind human rights, is a universal value with venerable traditions everywhere.

The critics of the universality of human rights remained skeptical. During the preparatory work on the 1948 Universal Declaration of Human Rights, leading U.S. anthropologists emphasized the cultural relativity of values and denounced the universalist pretensions of the intended declaration as cultural imperialism. Moreover, the historic critics of the universality of human rights, presented earlier, were succeeded by neo-Marxists, communitarians, and postmodernists. But the most influential group of critics was the vociferous ruling elites of some Asian and African countries. They challenged specific formulations of human rights, such as religion or marriage, and advocated, instead of universality, the principle of, for example, 'Asian values' or of domestic jurisdiction or of the primacy of second- and third-generation rights over first-generation rights. Remarkably, the victims of imprisonment, torture, and other violations in non-Western countries (among them many future elites of these countries), if they survived the prisons of dictators, found such criticism of the ruling elites unconvincing. They argued that denying universal human rights to citizens outside the West was itself a sign of cultural arrogance. One cannot deny progress in this regard: whereas the 1948 Universal Declaration of Human Rights was approved by a vote of 48 to 0 with 8 abstentions, at the 1993 Vienna Conference, 171 states unanimously reaffirmed the universality claim laid down in this declaration. This made the universality claim more representative and hence more solid.

Progress and Failure in the History of Human Rights

Recently, historian Bruce Mazlish defended the thesis that inhumanity breeds humanity. He argued that World War II and the Holocaust shocked the world to such a degree that, in

reaction, a human rights revolution was unleashed. This opinion can be called the *inhumanity thesis of human rights*. It is a strong thesis but not an unproblematic one.

If we want to test it, the first question is whether we can know anything at all about the moral nature of human beings. Are human beings good or evil? Many people, often inspired by religion, strongly believe in the intrinsic goodness of humans. In contrast, others, like Immanuel Kant, hold that human beings are antagonistic and bellicose. Still others, like the anthropologist Marshall Sahlins, think that goodness and badness are intertwined features and that the question of the good or evil nature of humans is a false question. If human beings cannot be called unambiguously either good or bad, how can we know whether inhumanity has increased or decreased in the course of history? In absolute terms, the case for an overall increase of inhumanity is quite defensible. Population growth on earth and the increasing efficiency of repressive technologies make this probable. Our perception, however, may be misled: we may overestimate the present absolute level of inhumanity because of our lack of knowledge of inhumanity in the past. Alternatively, we may underestimate the present level because we are too impressed by the recently expanding human rights awareness. In relative terms, the question is different: given that an equal amount of inhumanity probably leads to greater absolute destruction in modern times, were past societies more repressive or more humane toward their citizens than those of the twenty-first century? When we look at the number of victims of human rights catastrophes as a percentage of the total population, it is not certain at all, as many think, that present-day rates are higher than those of the past. Peter Singer has argued that in many tribal societies, despite the absence of machine guns and high explosives, the percentage of the population killed annually in warfare far exceeded that of any modern society. A probable absolute increase of inhumanity in history, then, does not imply that humankind is more barbarous now than it was in the past, nor that it is less. If anything, it seems likely that, in relative terms, levels of inhumanity are fluctuating or even possibly in decline.

If the level of inhumanity is not necessarily on the rise, why did large-scale inhumanity not trigger a human rights regime before the eighteenth century? We can only speculate that, in the further past, massive violence was perceived to belong to the *condition humaine*. At the same time, attention to the fate of victims of crime has an impressive pedigree: everywhere, religious and ethical systems have preached principles of humanity throughout history (although these systems often defined humanity in a limiting sense). We are forced to conclude, however, that the accumulative force of humanitarian attempts in the further past apparently did not pass a critical threshold.

Another question concerns the watershed in human rights thought as it first appeared in the United States and France in the late-eighteenth century. According to Lynn Hunt, one of the factors explaining this human rights takeoff was the appearance around 1760 of a new genre of novels – epistolary novels – which made their readers sensitive to new experiences and enhanced in them the empathy and sympathy necessary to imagine new political conceptions, such as equality and human rights. Another equally strong factor was the development of new views of individual autonomy, which brought more respect for the body, making the practice of torture suddenly look unacceptable. The first factor, the influence of epistolary novels, seems implausible. It cannot be doubted that reading these hugely successful novels enhanced sympathy. But that is not enough to foster human rights. The sympathy aroused when reading epistolary novels is particularizing – directed at the protagonists of the novel – and not universalizing: it often is accompanied by a rejection of all those who are their adversaries. Thus, the transfer of sympathy is limited, and it stops shortly after the novel is read. No inclusive sympathy is reached in this way. There is additional support for this objection: nearly all of the tremendously popular protagonists of these novels were women, but women's rights were about the last category of rights to be realized. In addition, one of the foremost human rights campaigners, Mary Wollstonecraft, was skeptical about this novel reading.

In contrast, the second factor, the general outburst of revulsion against torture, is plausible. The intolerable character of ill-treatment led to the formulation of rights to protect physical integrity. It begs the question, however, of why torture generally was not seen as repulsive *before* 1760. It therefore is clear that, in addition to the torture factor, other conditions were necessary to spread the new insights. Hunt stressed that philosophers and legal scholars writing about human rights were read widely on both sides of the ocean. She also emphasized the practical advantages of thinking in universal terms: in America, the English colonies did not aim at reform but at independence, which made reference to a universal set of values attractive and necessary; in France, the new principles were cast in the language of universality and so opened vast horizons of applicability hitherto unthought-of. This is not to say that the newly formulated rights were enforced immediately. Trials of major perpetrators of human rights violations, for example, had not yet taken place. In short, the inhumanity thesis holds only partially for the human rights takeoff of 1760–1800. The question of why some relatively successful concept of humanity arose in France and the United States in the eighteenth century, but not elsewhere, merits more attention.

The next question is intriguing. After so promising a start around 1760–1800, why did the human rights idea fall into eclipse between 1800 and 1940? It implies another question: did the idea *really* become eclipsed after 1800? Some developments testify to the continuation of the élan during these 140 years. One could point to the strength of the movement for the abolition of slavery, the Geneva Conventions, the gradual development of customary rules of international humanitarian law, and the Hague Conventions. These developments of the nineteenth century were followed by myriad important if relatively modest human rights initiatives in the interwar period, including those of the League of Nations. Although research may further reveal that the eclipse was not as total as it might look at first, it seems unlikely that this period will prove an uninterrupted triumph of human rights. The human rights idea was attacked by conservatives, liberals, and socialists. Powerful currents of defensive nationalism, imperialism, and racism in the nineteenth century, and scores of dictatorships appearing in the

twentieth century, discouraged thoughts about universal human rights before 1940.

For our purposes, it is noteworthy that an initiative such as the establishment of the Red Cross was propelled by a war: the 1859 Battle of Solferino. The Red Cross would become the official custodian of the Geneva Conventions. As this still is the most important instrument for humanitarian action in times of war, the Solferino connection adds considerable credibility to the inhumanity thesis. In contrast, World War I yields an ambiguous picture. On the one hand, although it saw a genocide (the Armenian genocide), an attempt to try its perpetrators failed. On the other hand, one of its results was the League of Nations, an initiative meant to stimulate international cooperation. It seems that none of the steps taken in the period between 1800 and 1940 was decisive.

The next question is whether the inhumanity of war and genocide really triggered the human rights revival of 1940–1948, as was postulated by Mazlish. The factors 'war' and 'genocide' played different roles. A total war of aggression was being visibly prepared by Nazi Germany from at least 1938. H.G. Wells and others mounted a human rights campaign immediately after the start of the war in September 1939. Some of Wells's ideas were picked up by Franklin Roosevelt in his Four Freedoms speech in 1941. It is still not clear, however, whether the main document of the new epoch, the United Nations Charter (1945), played a catalyst role in the campaign. Some emphasize that the mere mention of human rights in the Charter was sufficient to establish a Human Rights Commission, whose first assignment was to draft the Universal Declaration of Human Rights.

The other factor, genocide, was more ambiguous. From the summer of 1942, it gradually dawned upon the Allies that the Nazis were performing a large-scale extermination campaign, but the truth in its full brutality became generally visible only in the last months of the war. This was sufficient to unchain the process to draft the 1948 Genocide Convention, which represented a serious breach of the sacred principle of nonintervention in domestic affairs of states. Strangely enough, however, there was little connection between the debates about the Genocide Convention and those about the Universal Declaration, although they were held simultaneously, and it therefore is not clear whether the Holocaust was incessantly on the minds of the drafters of the Universal Declaration. We must realize how uncertain the semantic impact of the term 'genocide' still was between 1945 and 1948: this new concept was to conquer the minds of the masses only slowly. The Nazi genocide, when it was on the minds of the drafters of the Universal Declaration, was immersed in a more general aversion to Nazi barbarity. Even if such awareness of Nazi barbarity was omnipresent, the idea that the worst part of that barbarity constituted a unique crime called 'genocide' only slowly emerged in wider circles.

Recently, the historian Samuel Moyn has challenged the view of the deep roots of human rights. For him, human rights have been really successful only since 1977 when they finally left their United Nations quarantine and became embedded in a lasting mass movement. That mass movement was the result of the emergence of Helsinki groups and other human rights organizations in Eastern Europe and the Union of Soviet Socialist Republics, the incorporation of human rights into the politics of President Carter in 1977, and the conferment of the Nobel Peace Prize on the grassroots organization Amnesty International later that year. These factors coincided with structural developments, such as the collapse of the credibility of the geopolitical rhetoric of the Cold War and the postcolonial disillusionment with revolutionary political utopias. According to Moyn, human rights emerged as a minimalist and practical utopia, with moral rather than political overtones. Moyn also argued that the belated success of human rights was accompanied by the invention of a longstanding tradition starting in 1760, whereas, in reality, the survival of the idea of human rights was but a story of contingencies.

Moyn's analysis is convincing with two important caveats. First, his debunking of human rights antecedents before 1977 depends on a particular appreciation of historical rhythms. For example, one of the most notable developments, the incorporation of human rights notions into most of the world's constitutions well before the 1970s, is called 'a fashion' by Moyn. There was not only coincidence after coincidence, in myriad ways there was also an accumulated pressure not to let the human rights idea die. From 1945 to 1948, attention to human rights was huge in comparison to anything that had happened on that front before 1945, but it was modest in comparison to anything that was to happen after 1977. In short, the breakthrough of 1977 was unthinkable without this preparatory work. Second, like Mazlish, Moyn underestimated the strength of humanitarianism, largely because he equated it with philanthropy (*social* humanitarianism). Despite some overlap, *legal* humanitarianism is very different from philanthropy. The Red Cross not only crafted the subtle edifice of the Geneva Conventions but also played a substantial role in the development of several human rights (such as the right to the truth) and of court statutes (such as the 1998 statute of the International Criminal Court).

It is clear from these factors that inhumanity did not play a major role in the breakthrough of 1977. The wars and genocides of the 1960s and 1970s surely provoked indignation, but they had no lasting resonance among human rights activists. Still, one inhumanity-related factor did play a role. Outraged by the repression of Latin American dictatorships in the 1970s, Amnesty International launched its antitorture campaign in 1973. This campaign met with overwhelming approval and perhaps became the major key to Amnesty's success. Here, as in the 1760s, the general repugnance against torture was a powerful stimulus.

Juridical globalization from 1998 was an important next step. In contrast to the boost of 1977, this recent wave of global justice *does* have a direct connection with war and genocide: the wars in the former Yugoslavia and the ethnic cleansing and genocide there and in Rwanda led to the establishment of international tribunals, which formed the laboratory for the International Criminal Court and accelerated the process of approval of the latter's statute in 1998. This indicates at least modest progress.

Some, however, insist that all these results are meager. Where some see progress, these critics see regression or stagnation. This skepticism is not unreasonable. Indeed, those defending the judgment of recent progress in human rights may be misled: *any* progress in this area is impressive when the start is recent and the level of departure low. The judgment of recent failure of human rights, in its turn, has to meet several empirical objections also. First, the imperfect application of the concept of human rights in everyday reality does not

necessarily discredit the concept itself. Second, among the growing number of planetary *lieux de mémoire,* which stimulate global awareness, some of the strongest belong to the class of inhumanities: the Holocaust, Hiroshima, and 11 September are examples. Third, even skeptics have to admit that a world without a Universal Declaration is worse off than one with such a declaration. The Geneva Conventions are ratified by all countries in the world. There is also the idea of genocide. It was given shape in a convention in 1948, but the first sentence ever for genocide was handed down in 1998: half a century later. Since then, convictions for genocide have multiplied rapidly. Last but not least, the political regime best suited to protect human rights – democracy – has made considerable progress over the past decades. The UNDP recently showed that the share of democracies increased from less than a third of countries in the early 1970s to three-fifths in 2008. At the same time, however, the reported levels of human rights violations have remained virtually unchanged globally between roughly the same decades. It thus seems that the rise of democracy, casting doubt on the judgment of recent failure, was matched by stagnation rather than a rise in human rights protection, casting doubt on the judgment of recent success. Both judgments can be the product of ahistorical myopia. It further shows that even if democracy is the regime most favorable to human rights, progress is not to be taken for granted.

It is now possible to evaluate the inhumanity thesis. The relationship between inhumanity and humanity is not a necessary one. There were many instances in which inhumanity only gave birth to more inhumanity and to new cycles of violence. Conversely, progress in human rights thought often was unrelated to any act of inhumanity. Proof for the causal role of the inhumanity factor was weakest during the most important breakthrough, the one of 1977. At several moments in history, however, inhumanity *did* propel humanism. And it was not always the worst inhumanities that provoked a response in terms of human rights. Even the awareness of the risks of inhumanity often was sufficient. The unbearability of repetition probably played a role as well: it may explain why the human rights response was stronger after 1945 than after 1918. This is hopeful. For if the inhumanity thesis were necessarily true, we would need more human rights catastrophes to book more human rights progress. And that would be a self-defeating paradox. But progress in human rights is not given, it is fought for. As Kant wrote, "Nothing entirely straight can be fashioned from the crooked wood of which humankind is made."

See also: Courts and Adjudication; Dehumanization; Democracy, History of; Genocide, Historical Aspects of; Human Rights, Anthropology of; Natural Law; War Crimes Tribunals, Crimes against Humanity and Transitional Justice.

Bibliography

American Anthropological Association Executive Board, 1947. Statement on human rights. American Anthropologist 49, 539–543.

Amnesty International, 1961–2013. Report. Amnesty International, London.

Andreopoulos, George (Ed.), 1994. Genocide: Conceptual and Historical Dimensions. University of Pennsylvania Press, Philadelphia, PA.

Burgers, Jan Herman, 1992. The road to San Francisco: the revival of the human rights idea in the twentieth century. Human Rights Quarterly 14, 447–477.

Cassese, Antonio, Gaeta, Paola, Jones, John (Eds.), 2002. The Rome Statute of the International Criminal Court: A Commentary. Oxford University Press, Oxford.

Castberg, Frede, 1968. Natural law and human rights: an idea-historical survey. In: Eide, Asbjörn, Schou, August (Eds.), International Protection of Human Rights. Almqvist & Wiksell, Stockholm, pp. 13–34.

Chalk, Frank, Jonassohn, Kurt, 1990. The History and Sociology of Genocide: Analyses and Case Studies. Yale University Press, New Haven–London.

De Baets, Antoon, 2004. A declaration on the responsibilities of present generations toward past generations. History and Theory: Studies in the Philosophy of History 43, 130–164.

De Baets, Antoon, 2007. A successful utopia: the doctrine of human dignity. Historein: A Review of the Past and Other Stories 7, 71–85.

De Baets, Antoon, 2009a. The impact of the Universal Declaration of Human Rights on the study of history. History and Theory: Studies in the Philosophy of History 48, 20–43.

De Baets, Antoon, 2009b. Responsible History. Berghahn, New York and Oxford.

De Baets, Antoon, 2011a. Historical Imprescriptibility. Storia della Storiografia/History of Historiography no. 59–60, 128–149.

De Baets, Antoon (Ed.), 2011b. History and Human Rights: Special Issue, vol. 59–60. Storia della Storiografia/History of Historiography, 43–149.

De Baets, Antoon, 2012. Does inhumanity breed humanity? investigation of a paradox. History and Theory: Studies in the Philosophy of History 51, 451–465.

Donnelly, Jack, 1982. Human rights and human dignity: an analytic critique of non-western conceptions of human rights. The American Political Science Review 76, 303–316.

Forsythe, David (Ed.), 2009. Encyclopedia of Human Rights. Oxford University Press, Oxford.

Henckaerts, Jean-Marie, Doswald-Beck, Louise (Eds.), 2005. Customary International Humanitarian Law. Cambridge University Press, Cambridge.

Henkin, Louis (Ed.), 1981. The International Bill of Rights: The Covenant on Civil and Political Rights. Columbia University Press, New York.

Human Rights Watch, 1989–2013. World Report. Human Rights Watch, Washington.

Hunt, Lynn, 2007. Inventing Human Rights: A History. Norton, New York and London.

Jones, Peter, 1994. Rights. Macmillan, Basingstoke.

Jongman, Albert (Ed.), 1996. Contemporary Genocides: Causes, Cases, Consequences. PIOOM, Leiden.

Kant (Immanuel), 1785/2002. In: Hill, Thomas, Zweig, Arnulf (Eds.), Groundwork for the Metaphysics of Morals. Oxford University Press, Oxford.

Kant, 2006. In: Kleingeld, Pauline (Ed.), Toward Perpetual Peace and Other Writings on Politics, Peace, and History. Yale University Press, New Haven and London.

Laqueur, Walter, Rubin, Barry (Eds.), 1979. The Human Rights Reader. New American Library, New York, NY.

Mazlish, Bruce, 2009. The Idea of Humanity in a Global Era. Palgrave Macmillan, New York.

Morsink, Johannes, 1999. The Universal Declaration of Human Rights: Origin, Drafting, and Intent. University of Pennsylvania Press, Philadelphia.

Moyn, Samuel, 2010. The Last Utopia: Human Rights in History. Harvard University Press, Cambridge, MA, and London.

Nowak, Manfred, 2005. U.N. Covenant on Civil and Political Rights: CCPR Commentary, second ed. Engel, Kehl am Rhein, Strasbourg, and Arlington, VA.

Preliminary note: Complete versions of the Human Rights Instruments Mentioned. Available at 'Network of Concerned Historians' www.concernedhistorians.org/content/to.html.

Rosenbaum, Alan (Ed.), 1980. The Philosophy of Human Rights: International Perspectives. Aldwych Press, London.

Rummel, Rudolph, 1994. Death by Government. Transaction Publishers, New Brunswick and London.

Sen, Amartya, 1999. Democracy as a universal value. Journal of Democracy 10, 1–17.

Simma, Bruno (Ed.), 2002. The Charter of the United Nations: A Commentary. Oxford University Press, Oxford.

Singer, Peter, 2002. One World: The Ethics of Globalization. Yale University Press, New Haven, CT.

Triffterer, Otto (Ed.), 1999. Commentary on the Rome Statute of the International Criminal Court. Nomos, Baden-Baden.

UNESCO (Ed.), 1949. Human Rights: Comments and Interpretations. Wingate, London and New York.

United Nations Commission on Human Rights, 2003. Continuing Dialogue on Measures to Promote and Consolidate Democracy. Office High Commissioner Human Rights. E/CN.4/2003/59 www.ohchr.org.

United Nations Commission on Human Rights, 2005. Updated Set of Principles for the Protection and Promotion of Human Rights through Action to Combat Impunity. Office High Commissioner Human Rights. E/CN.4/2005/102/add.1 www.ohchr.org.

United Nations Development Programme, 2000. Human Development Report. Theme: Human Rights and Human Development. Oxford University Press, Oxford.

United Nations Development Programme, 2002. Human Development Report. Theme: Deepening Democracy in a Fragmented World. Oxford University Press, Oxford.

United Nations Development Programme, 2010. Human Development Report. Theme: The Real Wealth of Nations: Pathways to Human Development. Oxford University Press, Oxford.

United Nations Economic and Social Council, 1985. Siracusa Principles on the Limitation and Derogation Provisions in the International Covenant on Civil and Political Rights. E/CN.4/1985/4, annex.

Human Rights: Legal Aspects

Jonathan Klaaren, University of the Witwatersrand, Johannesburg, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Human rights can be viewed from legal and sociolegal perspectives. From a legal perspective, human rights are the rights derived from the Universal Declaration of Human Rights, the International Covenant on Economic, Social and Cultural Rights, and the International Covenant on Civil and Political Rights and other human rights instruments; they are enforced on both international and domestic levels. Sociolegal perspectives, embedded within the disciplines of history, sociology, anthropology, and international relations, instruct that human rights are not constrained only by law; they have their own history, sociology, social life, and transnational activist networks. The issue may be posed whether the field of human rights has moved beyond a topic of interdisciplinary research to become a distinct discipline.

Introduction

An article entitled 'Human Rights: Legal and Sociolegal Perspectives' must designate its subject matter. For this purpose, the term 'human rights' is used here to identify primarily universal or international human rights rather than rights within nation-states. Thus, by human rights, this article understands the international human rights regime rather than national ones. The two remits overlap but are distinct (Henkin, 1979). International human rights are often identified and given legal weightage by a series of international human rights treaties and declarations adopted by the United Nations (UN) (or other international organizations) in the two decades following the end of the Second World War beginning with the Universal Declaration of Human Rights (Sieghart, 1983).

It is also necessary to be clear about the perspectives from which this subject will be approached. The article is written from both the legal and the sociolegal perspective. It thus surveys a range of perspectives including but not limited to those within the discipline of law. Other articles in this encyclopedia address matters of legal doctrine related to human rights (see cross-references above).

Human rights almost need no introduction in an encyclopedic work, as their significance is widely understood. As the pioneering American law professor Louis Henkin wrote, we live in 'the age of rights' (Henkin, 1990): Many if not most of the significant issues of our day are framed in and resolved within a human rights frame of reference (Sarat and Kearns, 2001: p. 2). With significant variations across the diverse societies of the globe, this has been the case, broadly speaking, since the end of the Second World War (Minow, 2002). Further, much of the modernizing and progressive energy of social-change advocates intending to make the world a better place is often expressed in human rights terms. Indeed, this has been the case with the drive to ensure equality within and among nations, as well as with the drive to end poverty, as we have begun to recognize a human right to development (Alston, 1988). Indeed, as the world continues to globalize yet does so in new ways, the doctrine of human rights as distinct from national or constitutional rights takes on new significance. New constitutions and innovative doctrines in international law draw their social significance and legitimacy partially but increasingly from human rights.

A work of survey must start and end its mapping somewhere. In its main body, this article will first detail a legal perspective on human rights. This presentation of significant human rights instruments, institutions, and legal issues is largely internal to the human rights field and to the legal discipline. Second, the article turns to a survey of four sociolegal (interdisciplinary) perspectives on human rights, addressing works from history, sociology, anthropology, and international relations. The article concludes with a brief reflection on the development of human rights as a perspective (an intellectual discipline) in its own right.

Human Rights: A Legal Perspective

Within the discipline of law, there are a myriad of approaches to human rights. The aim here is to identify the key conceptual distinctions made, and institutions resourced, in order to provide a legal perspective on human rights.

One should start with the legal delineation and source of universal human rights. Popular usage conflates as human rights several fields of law that are often kept conceptually distinct within the discipline of law: international human rights law, international criminal law (Slye and Van Schaak, 2009; Van Schaak and Slye, 2010), international humanitarian law (Sassioli et al., 1999), and international refugee law (Goodwin Gill and McAdam, 1996; Hathaway, 2005). In practice, however, specific human rights issues may be addressed from legal perspectives based in more than one of these fields. Moreover, emerging issues (e.g., the development of international criminal tribunals and the field of transitional justice) often arise at the intersection of one or more of these fields.

International human rights law scholars tend to distinguish between international instruments under the auspices of the UN, on the one hand, and regional instruments of international law on the other. The Convention on the Elimination of All Forms of Discrimination Against Women, which was adopted by the UN General Assembly in 1979, is an example of the former, as is the African Charter on Human and Peoples' Rights, agreed to by the Organization of African Unity in 1986. Other regions adopting regional instruments include Europe (the European Convention for the Protection of Human Rights

and Fundamental Freedoms) and the Americas (the American Convention on Human Rights).

A further distinction within international human rights law is that between civil and political rights on the one hand and socioeconomic rights on the other. This distinction builds on the existence of three separate UN human rights instruments: the Universal Declaration of Human Rights (adopted in 1948), the International Covenant on Economic, Social and Cultural Rights (1966), and the International Covenant on Civil and Political Rights (1966). Recent scholarship contests this distinction and stresses the historical and conceptual indivisibility and interrelatedness of all forms of human rights (Kabasakal Arat, 2006; Whelan and Donnelly, 2007).

From a legal perspective, human rights occasion what one might call the ideal of legal compliance with human rights. In this ideal, human rights are identified and agreed to by the UN (or a regional body of states) in the form of human rights treaties. These legally binding treaties are then enforced in three possible ways: through actions among the states themselves, through actions of agencies of the international community of states, or through domestic actions of the states ratifying and incorporating those treaties. In this ideal, the specific mode of enforcement does not assume any great significance but the fact of compliance – of adherence to human rights – does matter.

Given this ideal of legal compliance with human rights, one might identify two broad levels of formal enforcement of human rights: the international level, understood to include the community of states as well as institutions such as international courts and tribunals, and the domestic level, understood to include constitutions as well as domestic institutions such as courts and human rights commissions. Some human rights legal scholars are of the opinion that the impact of human rights treaties has been more significant through their use at the national level than through the supervisory measures they provide for, such as state reporting and individual complaints, at international levels (Heyns and Viljoen, 2001). Indeed, international human rights legal scholars have paid careful attention to whether international human rights treaties make a difference (Hathaway, 2002; Goodman and Jinks, 2003, 2004; Halliday and Schmidt, 2004).

At the international level, states can and do pressure each other regarding compliance with international human rights law. For instance, the signing of the Helsinki Final Act of 1975 established human rights as a norm ostensibly binding on all states of Europe. Following 1975, states could legitimately call each other to account on human rights issues, citing the Final Act (Thomas, 1999). More recently, over the past two decades, states have established a number of independent international tribunals with powers and procedures akin to those of domestic courts (see e.g., http://jura.ku.dk/icourts/icourtfinder/). States may have established these institutions in order to increase their own credibility in a number of multilateral settings (Helfer and Slaughter, 2005, 1997).

At the domestic level, the primary institutions seen to be implementing human rights law are domestic constitutions, courts, and human rights commissions. Thus some human rights scholars and organizations have examined the impact of the evolution of the international human rights regime on national bills of rights, arguing that the development of the regime has transformed many of the debates around domestic bills of rights (Alston, 1999). Others have argued for the articulation of a common law basis for international human rights law in order to bypass the question of the status of human rights law as 'incorporated' or 'unincorporated' (see below) in English courts (Hunt, 1997), while still others have argued that the initial performance of most African human rights commissions has been disappointing (Human Rights Watch, 2001).

Along with attention to these two levels of formal enforcement, human rights legal scholars also pay attention to questions of culture (including the issue of cultural relativism) (An'Naim, 1995) and of civil society organizing and advocating for human rights concerns (Welch, 1995).

Deeply intertwined – at least from a legal perspective – with the enforcement process are a number of issues and questions of doctrinal law. For example: (1) what rules govern the interpretation of international human rights treaties? (2) is a particular human rights instrument a treaty (and thus binding at international law) or not? (3) has the treaty been ratified by sufficient numbers of states to have entered into force? (4) has a particular state ratified the treaty and, if so, when? (5) did a particular state make any reservations (excluding or modifying its effect) to a human rights treaty? (6) what is the domestic legal effect of the human rights instrument? (7) has the state temporarily derogated from or suspended the human rights instrument? and (8) has the potential enforcement of a human rights matter exhausted domestic remedies? (Martin et al., 1997). Much of human rights law, from the point of view of its practitioners is taken up in addressing and resolving these doctrinal legal questions.

Human Rights: Four Sociolegal Perspectives

Consistent with their claimed universal nature, human rights have overlapped with nearly every discipline and significant community of practice. For instance, political scientists have long been interested in the dynamics and power of human rights, and scholars of human rights have often been its powerful practitioners and politicians. An example of the latter is Michael Ignatieff, serving as the Leader of the Official Opposition in Canada for several years. Broadly speaking, disciplines can offer both internal and external perspectives on human rights. The four disciplines surveyed below – history, sociology, anthropology, and international relations/politics – arguably go a step beyond and offer a perspective that may be constitutive of the discipline of human rights itself.

History and Human Rights

As a relatively recent development, historians have begun to explore the histories of human rights. For instance, a timely and recent cultural and intellectual history of human rights traces the roots of the concept to the rejection of the practice of torture as a means for finding the truth (Hunt, 2007). In a different vein, Moyn has argued that in the 1970s "the moral world of Westerners shifted, opening a space for the sort of utopianism that coalesced in an international human rights movement that had never existed before" (2010: p. 1). First, other histories narrow their focus and probe the origins of human

rights within a particular national jurisdiction, such as in South Africa (Dubow, 2012). The study of human rights takes such scholars into a particular specialist area of law – human rights – but human rights arguably presents no greater methodological challenge than the one faced by historians of business studying the development of the railroad industry in the nineteenth century. For historians, the entire field of social activity is available as data and evidence. Still, these histories of human rights have been influenced by two hallmarks of a sociolegal perspective: first the attention to the 'gap' between the law 'on the books' and the law 'on the ground' and, second, the attention to the variety of ways in which law and the variety of interpretations of law interrelate and influence each other.

Sociology and Human Rights

While the recent attention to human rights within sociology has picked up from the mid-2000s, there has been earlier attention. One strand of such attention draws upon the key normative concerns of the discipline. Turner, for instance, made an argument for a sociology of human rights that would serve as a substitute for national citizenships, which he saw as inherently limited (Turner, 1993). Sociologists have continued to work within this universalizing moral and normative vein (Soysal, 1994; Turner, 2010; Friesen, 2011a). Friesen has argued that "A human rights approach offers an opportunity to unite apparently disparate concentrations of the discipline – such as race, class, gender, sexual orientation – into a larger framework which analyzes equality and fairness on a *human* scale" (Friesen, 2011b). The sociological perspective can also bring two important correctives to practitioners of all types working in the field of human rights: a certain degree of reflexivity about the use of human rights tools as well as a way to exhibit "a reliance on legal institutions that remains aware of their limits and their own direction" (Hagan and Levi, 2007: p. 373).

Anthropology and Human Rights

Anthropologists have recently realized that they have much to offer to the field of human rights. Anthropologists emphasize what Richard Wilson terms 'fully documenting the social life of rights,' by which he means "the social forms that coalesce in and around formal rights practices and formulations, and which are usually hidden in the penumbra of the official political process" (Wilson, 2006: p. 78). For instance, one anthropological strand of productive research has been concerned with human rights and the police (Hornberger, 2011, 2010). Anthropologists have also offered the insight that human rights thinking itself originates within a specific culture that has often resided within and been symbiotic with American institutions, including notably American law schools (Riles, 2006).

One concept key to human rights that can be emphasized from the discipline of anthropology is what Mark Goodale, working with an ethnographic and critical approach to human rights, has termed 'betweenness.' Goodale and Merry describe "the locations where human rights discourse emerges in practice as 'between' the global and the local" (Goodale, 2007: p. 22; Goodale and Merry, 2007). For them, "betweenness is meant to express the ways in which human rights discourse unfolds ambiguously, without a clear spatial referent, in part through transnational networks, but also, equally important, through the projection of the moral and legal imagination by social actors whose precise locations … within these networks are (for them) practically irrelevant" (Goodale and Merry, 2007).

International Relations/Politics and Human Rights

International relations/politics has taken on the study of human rights as one of its core topics of research. Perhaps the classic theme in this field is to ask the question whether the norms of international human rights have had any effect on domestic governments (e.g., Risse et al., 1999). The answer, by and large, is often positive although usually not determinative or conclusive. An additional theme is to examine the working of the human rights movement. Here, one work that has achieved great prominence and influence is Keck and Sikkink (1998). Keck and Sikkink introduce network analysis to the study of human rights activists and work with the concept of transnational activist networks. In their view, human rights networks are "vehicles for communicative and political exchange, with the potential for mutual transformation of participants" (1998: p. 214).

Conclusion: A Reflection on Human Rights as a Distinct Discipline

This section asks the question whether the field of human rights is, or is becoming its own discipline, in the sense of an intellectual school of thought (which often bear a parallel to institutional developments in centers of knowledge production such as universities).

There is clearly a tradition of human rights scholarship that spans both a significant time period and a range of perspectives. For instance, one can track the development from the formative and field building work of Jack Donnelly (2013) to the more critical approach of Makau wa Mutua (1996, 2001, 2002). The former was deeply committed to the universal impulse of the human rights field and the latter, while departing from and using that assumption as a starting point, is nonetheless fundamentally critical of earlier work.

Moreover, approaches based within the human rights field have shown their power to reveal insights. For instance, one approach that might be understood to be located within a distinct human rights field is that of Martha Minow. She studied a range of responses to human rights violations, genocide, and other instances of mass violence. In her view, law can provide a repertoire of such responses (1998). This focus has led Minow to identify an institutional characteristic of responses to human rights violations that can be understood to run parallel to the betweenness concept identified at the heart of human rights practice by Goodale and Merry. In Minow's view (2002: p. 59), "[t]he advocates of human rights have generated innovative institutions and practices, yet the successive and multiple nature of these innovations can contribute to the sense that human rights are merely ideal aspirations and, in practice,

insecure. Yet this tradition of institutional innovation may be the most significant legacy of the human rights movement."

Furthermore, analyses based within the human rights field have had significant impact in the world of the academe as well as in the policy world. For instance, as influential a thinker as the philosopher and economist (not to mention his being a winner of the Nobel Prize) Amartya Sen has worked extensively within the field of human rights (1997, 2004). This impact has also been in the form of new knowledge creation. The field of transitional justice saw great growth as a specialized and potentially distinct field of practice and knowledge in the 1990s and 2000s. However, transitional justice seems not to have achieved status on its own as a discipline. Instead, the growth of the transitional justice field (which may best be understood to be included within that of human rights) seems rather to indicate the depth and diversity of topics within the human rights discipline.

An alternative to seeing human rights as a distinct discipline is to see it as an interdisciplinary field of study. Here a parallel to sociolegal studies itself may be illustrative. Sociolegal scholars have also asked whether their corner of the academy is a distinct discipline. By and large, the answer has been in the negative. It would seem that human rights rather than sociolegal studies has a better claim to (and chance for) distinct academic discipline status. Indeed, if the study of human rights is not a distinct field, then human rights is at least a topic for interdisciplinary work (Freeman, 2011). This article thus far might be seen as an example of an interdisciplinary approach toward a new and significant research topic. Work such as the recent history of human rights (Moyn, 2010) demonstrates a step toward such a development.

See also: Civil Liberties and Human Rights; Human Rights, Anthropology of; Human Rights, History of; International Law and Treaties; Torture: Legal Perspectives.

Bibliography

Alston, Philip, 1988. Making space for new human rights: the case of the right to development. Harvard Human Rights Yearbook 1, 3.

Alston, Philip, 1999. Promoting Human Rights through Bills of Rights: Comparative Perspectives. Oxford University Press, Oxford.

An-Na'im, Abdullahi, 1995. Problems of universal cultural legitimacy for human rights. In: Naim, Abdullahi Ahmed An (Ed.), Human Rights in Cross-Cultural Perspectives: A Quest for Consensus. University of Pennsylvania Press, Philadelphia, pp. 331–367.

Donnelly, Jack, 2013. Universal Human Rights in Theory and Practice. Cornell University Press, Ithaca.

Dubow, Saul, 2012. South Africa's Struggle for Human Rights. Jacana, Auckland Park, South Africa.

Freeman, Michael, 2011. Human Rights: An Interdisciplinary Approach. Polity, Cambridge.

Friesen, B., 2011a. Globalizing the human rights perspective. In: Blau, J., Frezzo, M. (Eds.), Sociology and Human Rights. Pine Forge, Thousand Oaks, CA, pp. 79–102.

Friesen, B., 2011b. Sociological Spring: Human Rights and the Discipline. Available at: http://www.asanet.org/footnotes/nov11/spring_1111.html.

Goodale, Mark, 2007. Introduction: locating rights, envisioning law between the global and the local. In: Goodale, Merry (Eds.), pp. 1–38.

Goodale, Mark, Merry, Sally Engle (Eds.), 2007. The Practice of Human Rights: Tracking Law between the Global and the Local. Cambridge University Press.

Goodman, Ryan, Jinks, Derek, 2003. Measuring the effects of human rights treaties. European Journal of International Law 14 (1), 171–183.

Goodman, Ryan, Jinks, Derek, 2004. How to influence states: socialization and international human rights law. Duke Law Journal, 621–703.

Goodwin-Gill, Guy S., McAdam, Jane, 1996. The Refugee in International Law. Clarendon Press, Oxford.

Hagan, John, Levi, Ron, September 2007. Justiciability as field effect: when sociology meets human rights. Sociological Forum 22 (3), 372–380.

Halliday, Simon, Schmidt, Patrick Delbert (Eds.), 2004. Human Rights brought Home: Socio-Legal Perspectives on Human Rights in the National Context, vol. 3. Hart Publishing, Oxford.

Hathaway, James C., 2005. The Rights of Refugees under International Law. Cambridge University Press, New York.

Hathaway, Oona A., 2002. Do human rights treaties make a difference? The Yale Law Journal 111 (8), 1935–2042.

Helfer, Laurence R., Slaughter, Anne-Marie, 1997. Toward a theory of effective supranational adjudication. Yale Law Journal 107, 273.

Helfer, Laurence R., Slaughter, Anne-Marie, 2005. Why states create international tribunals: a response to professors Posner and Yoo. California Law Review, 899–956.

Henkin, Louis, 1979. Rights: American and human. Columbia Law Review 79 (3), 405–425.

Henkin, Louis, 1990. The Age of Rights. Columbia University Press, New York.

Heyns, C., Viljoen, F., 2001. The impact of the United Nations human rights treaties on the domestic level. Human Rights Quarterly 23, 483.

Hornberger, Julia, 2010. Human rights and policing: exigency or incongruence? Annual Review of Law and Social Science 6, 259–283.

Hornberger, Julia, 2011. Policing and Human Rights: The Meaning of Violence and Justice in the Everyday Policing of Johannesburg. Routledge.

Hunt, Lynn, 2007. Inventing Human Rights: A History. W.W. Norton & Co., New York.

Hunt, Murray, 1997. Using Human Rights Law in English Courts. Hart Publishing, Oxford.

Ignatieff, Michael, 2001. Human Rights as Politics and Idolatry. Princeton University Press, Princeton, N.J.

Kabasakal Arat, Zehra F., 2006. Forging a global culture of human rights: origins and prospects of the international bill of rights. Human Rights Quarterly 28 (2), 416–437.

Keck, Margaret E., Sikkink, Kathryn, 1998. Activists beyond Borders: Advocacy Networks in International Politics. Cornell University Press, Ithaca, NY.

Klug, Heinz, 2000. Constituting Democracy: Law, Globalism, and South Africa's Political Reconstruction. Cambridge University Press, Cambridge, UK.

Martin, Francisco F., et al., 1997. International Human Rights Law and Practice: Cases, Treaties and Materials. Kluwer Law International, The Hague, Boston.

Minow, Martha, 1998. Between Vengeance and Forgiveness: Facing History after Genocide and Mass Violence. Beacon, Boston, MA.

Minow, Martha, 2002. Instituting Universal Human Rights Law: the invention of tradition in the twentieth century. In: Sarat, A., Garth, B., Kagan, R. (Eds.), Looking Back at Law's Century. Cornell University Press, Ithaca, NY, pp. 58–77.

Moyn, Samuel, 2010. The Last Utopia: Human Rights in History. Harvard University Press, Cambridge, MA.

Mutua, Makau, 2002. Human Rights: A Political and Cultural Critique. University of Pennsylvania Press, Philadelphia, PA.

Mutua, Makau, 2001. Savages, victims, and saviors: the metaphor of human rights. Harvard International Law Journal 42 (1), 201–245.

Mutua, Makau, 1996. The ideology of human rights. Virginia Journal of International Law 36.

Nowrojee, Binaifer, 2001. Protectors or Pretenders? Government Human Rights Commissions in Africa, Human Rights Watch, New York.

Riles, Annelise, 2006. Anthropology, human rights, and legal knowledge: culture in the iron cage. American Anthropologist 108, 52–65.

Risse, Thomas, Ropp, Stephen, Sikkink (Eds.), 1999. The Power of Human Rights: International Norms and Domestic Change. Cambridge University Press, New York.

Sarat, Austin, Kearns, T., 2001. The unsettled status of human rights: an introduction. In: Sarat, A., Kearns, T. (Eds.), Human Rights: Concepts, Contests, Contingencies, pp. 1–24.

Sassòli, Marco, Bouvier, Antoine A., Olson, Laura M., Dupic, Nicolas A., Milner, Lina, 1999. How Does Law Protect in War?: Cases, Documents, and Teaching Materials on Contemporary Practice in International Humanitarian Law. International Committee of the Red Cross.

Sen, Amartya, 2004. Elements of a theory of human rights. Philosophy & Public Affairs 32 (4), 315–356.

Sen, Amartya, 1997. Human Rights and Asian Values. Carnegie Council on Ethics and International Affairs, New York.

Sieghart, Paul, 1983. The International Law of Human Rights. Clarendon Press, Oxford.

Slye, Ronald, Van Schaack, Beth, 2009. International Criminal Law: Essentials. Wolters Kluwer Law & Business, Austin.

Soysal, Yasemin Nuhoglu, 1994. Limits of Citizenship: Migrants and Postnational Membership in Europe. University of Chicago Press, Chicago.

Thomas, Daniel, 1999. The Helsinki accords and political change in Eastern Europe. In: Risse, T., Ropp, S., Sikkink, K. (Eds.), The Power of Human Rights: International Norms and Domestic Change. Cambridge University Press, New York, pp. 205–233.

Turner, Bryan S., 1993. Outline of a theory of human rights. Sociology 27 (3), 489–512.

Turner, Bryan S., 2010. Vulnerability and Human Rights. Pennsylvania State University Press, University Park, PA.

Van Schaack, Beth, Slye, Ronald, 2010. International Criminal Law and Its Enforcement: Cases and Materials. Foundation Press, New York.

Welch, Claude E., 1995. Protecting Human Rights in Africa: Roles and Strategies of Nongovernmental Organizations. University of Pennsylvania Press, Philadelphia, PA.

Whelan, Daniel J., Donnelly, Jack, 2007. The West, economic and social rights, and the global human rights regime: setting the record straight. Human Rights Quarterly 29 (4), 908–949.

Wilson, Richard A., 2006. Afterword to anthropology and human rights in a new key: the social life of rights. American Anthropologist 108, 77–83.

Human Sciences, History of

Stephen P Turner, University of South Florida, Tampa, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term Human Sciences is primarily a French usage, but it refers back to a much deeper tradition in the literature claiming that works of the spirit and human experience cannot be reduced to the realm of causal science, and require different methods. Following Kant, much of this discussion has focused on the problem of the conceptual formation of human experience. Methodologically, discussion has shifted back and forth between an emphasis on concepts, on experience, and external facts. Foucault and Bourdieu extended the critical possibilities of the notion of misrecognition to ally the human sciences with critique.

The social sciences have been known by many names, including the moral sciences, the *Geisteswissenschaften*, the behavioral sciences, and the 'sciences of man.' Each of these names conveys a particular philosophy or methodological ideology, and typically includes or excludes some other discipline, such as psychology or law, and emphasizes some disciplines, such as linguistics, as especially exemplary. But the boundary lines and differences are not very precise, and they are evident more in the history of methodological disputes than in actual usage and administrative practice.

The *Sciences Humaine*, or Human Sciences, is a term of French origin. It appears in Francophone settings in numerous official and academic contexts, for example, in the name of the *L'Institut des Sciences Humaines et Sociales*, part of the *Center National de la Recherche Scientifique*. Part of the mission of the former institute is the promotion of interdisciplinary work, and the term 'human sciences' carries a suggestion of interdisciplinarity. The name also appears in the designation of faculties, in such combinations as Philosophy and the Human Sciences, Letters and the Human Sciences, Law and the Human Sciences, and Human and Social Sciences. The term *Sciences de l'Homme* overlaps this usage, but implies both a narrower definition, limited to the traditional social sciences, and a broader one, in that it implies nothing about methodology, and can be stretched to include human biology and the neurosciences. Often the term is used to encompass all of 'sciences' of the human, including history, geography, economics, law, political science, administrative science, educational science, linguistics, psychology, and the traditional social sciences. It is occasionally extended to include the humanities in general, as well as theology.

As the label 'Human and Social Sciences' suggests, however, it is often used as a means of distinguishing the approach from at least some of the social sciences, or some aspects of the social sciences. The American National Endowment for the Humanities uses the phrase "those aspects of social sciences which have humanistic content and employ humanistic methods," which captures this distinction. But in English usage, the term science is used as a way of distinguishing its methods from those of the humanities. In Continental and especially German usage, the term 'science' (*Wissenschaft*) is applied to any form of organized knowledge including those using the methods of the humanities. The journal *History of the Human Sciences* includes "traditional humanistic studies as well as the social sciences," and "publishes wide-ranging material in the fields of sociology, psychology, political sciences, the history and philosophy of science, anthropology, classical studies and literary theory."

Despite this amorphousness, the term designates something coherent: a more or less common approach opposed to the 'moral sciences' of J.S. Mill and the positivism of Auguste Comte and to their scientistic heirs. The central philosophical argument for this opposition was the irreducibility of something (meaning, intention, subjective experience, language, culture, historical individuals, etc.) to science, or to any law-like or causal analysis. In addition, it was claimed that there was knowledge, or access, to these irreducible things through other means, such as understanding, introspection, or empathy, reflexivity, or through hermeneutical or phenomenological analysis, or through a scientific analysis of a different, structural, kind, of language, artifacts, or cultural objects, practices, and patterns of human conduct.

The Philosophical Background

In a late interview, Michel Foucault, who was not only an influential figure in, but also an historian and architect of the human sciences, commented that "we are all Neo-Kantians" (Foucault, 1994: vol. 1, 546; quoted in Han, 2002: 3). This is true, in very complex but important ways, and applies more generally to much of what goes under the label 'Human Sciences.' To understand his comment it is necessary to explain Kantianism and neo-Kantianism and their persistent importance for the Human Sciences.

The Kantian heritage runs through the history of the human sciences in two basic ways. The first is the obvious one: for Kant the human world was the world of free will and spirit, guided by the moral law, in contrast to the physical world, governed by deterministic law. The second is related to the Kantian dictum that we can know and understand only what we make, and not the world as it is: "we have complete insight only into that which we can make ourselves and according to our own concepts" (*Critique of Judgment*, §68; quoted in Beiser, 2006: 13). But the path from Kant is anything but simple. The German idealists radicalized and with Hegel historicized these ideas, and they were taken up in misunderstood but

fertile ways by other thinkers, especially in France (Schmaus, 2004), and blended, not to say entangled, with other ideas.

Among the other issues that early nineteenth century discussions were entangled with were theological ones, especially issues over Biblical interpretation. These led to the notion of hermeneutics as a distinct method, not only for the interpretation of texts but for human understanding. The idea of understanding as a distinct mode or kind of knowledge specific to the study of humans led to a large discussion of the distinction between explanation and understanding, and of their possible relations about the nature of man.

The theological entanglements were connected with the traditional idea that the Christian message could not be understood except through faith: that only an act of faith opened one to the true meaning of Christianity and experience of Christ. The sharpness of this discussion was a result of the growing conflict between science and theology: theologians worked hard to find a preserve of modes of thought, belief, and knowledge that science could be excluded from. Inner knowledge, inner experience, was made central to the experience of religion, and thus to one's humanity.

The German Idealists who followed Kant and radicalized him developed various forms of what Arthur Schopenhauer was to call, in the title of his major work, *The World as Will and Representation* (2010), by giving a large role to Will or construction in their accounts of and extending these notions to the social and political world, especially to the nation as an object of thought. Schopenhauer denied the subject-object distinction, claiming that our only knowledge was of ideas, or representations. Johann Gottlieb Fichte (2005), similarly, rejected the Kantian idea of the unknowable thing itself, and instead focused on subjectivity as the source of our representations and therefore all that we can know. With Fichte we get the idea that consciousness is a social phenomenon, and that consciousness of other rational individuals precedes the 'call' which brings out individual consciousness. This is the beginning of the idea of intersubjectivity. But with Fichte we also see the development of some more sinister notions of national mentalities that are not unrelated to his philosophy,

The idea of a *Volksgeist* is one of the many collectivist notions of mental life that figures in the early history of the human sciences. This concept had a special historical importance. It was associated with the aristocrat and professor of law Friedrich Carl von Savigny, who argued that the specific form of law of a state arose from the historically rooted spirit of the people. Von Savigny, by taking up a career as a professor, raised the status of the German universities, contributing to the process that led them in the late nineteenth century to be the world standard. Ideas with collectivist character, such as the idea of a shared *Weltanschauung* (world view), mystical ideas about community and about the nation as a community of fate, as well as medievalist nostalgia for organically unified societies all went together and merged in an anti-individualist, anti-liberal, anti-economistic, and anti-English mélange, which was associated with German national identity.

These ideas were not originally associated with Kantianism. Indeed Johann Gottfried von Herder, Kant's student, developed his own ideas of *Volk*-thought in opposition to Kantianism. But the affinities between German Idealism in the period between Kant and Hegel and collectivist ideas were strong. Nevertheless, the picture is not simple: there was German liberalism and individualism that Herder, Wilhelm von Humboldt, Rudolph Gneist, and others also promoted, which typically connected the idea of individual development and freedom with the idea of national development and freedom. Reconciling the two kinds of freedom was a characteristic concern of these thinkers, and of Hegelians, one often resolved by notions of the mystical harmony between the two.

The relationship between German nationalism and ideas about the sciences of the human was to persist as the methodological discussions developed. The earliest and most intense of these methodological discussions preceded the idea of a human 'science.' They were concerned with historical knowledge, which raised the question of the individual in acute form. German historicism is usually thought to begin with the notion of the distinct individual developing according to its own being, a doctrine that applied to *Kollektivgebilde* or collective beings such as nations or even shared ideas as well as to individuals (Bos, 2010: 209). If each individual had his or her own principle of development, it was necessary for the historian to respect this: to exhibit the specific and unique character of the individual in time.

This had methodological implications that immediately distinguished history from generalizing sciences. Humboldt developed these thoughts by arguing that history itself did not have an aim or *telos*. This argument was taken up in the middle of the century by Johann Gustav Droysen, specifically against Hegelian accounts of history. For Droysen, the father of historical hermenuetics, objectivity in the sense of Leopold von Ranke's history as it actually happened was overrated, and Hegelian history in terms of the absolute idea was delusional. In a famous statement, he said:

> I want to appear to have nothing more – but also nothing less – than the relative truth of the standpoint bequeathed to me by my fatherland, and political and religious convictions, my earnest study. (Quoted in Maclean, 1982: 361)

This self-confessed reflexive relativism contained, among other things, a sense of the relativity of historical standpoints, the relativity imposed by the need to speak to present audiences. For Droysen this went with a higher valuation of understanding: he is the father of *Verstehen* as a goal of historical inquiry. Interpersonal understanding is both the basis of social existence and of historical inquiry (Maclean, 1982: 359).

The Neo-Kantians

With this we have the fundamental elements of the human sciences, which cohere with Wilhelm Dilthey and the neo-Kantians. The relations between Dilthey and the neo-Kantians is complex, but the bottom line is this: neo-Kantianism was a failure on its own terms, but bequeathed problematics that are still with and part of the human sciences; Dilthey recognized the tensions between the elements of his own thought,

which included Kantian elements, and never fully resolved them, but did develop his ideas in a direction that points directly to later human science.

For the Kantians, the characteristic problem is objectivity. The central question is how, given this dependence of knowledge on our construction of the world, is objectivity possible. For Kant, the answer was that there was only one correct and necessary construction of the world. Neo-Kantianism abandoned this, but only in part. Neo-Kantianism is often dated to the 1872 dissertation of Hermann Cohen, which purported to explicate Kant, but in fact turned him on his head. Cohen's reasoning was this:

> If … I take cognition not as a form and manner of consciousness, but as a *Faktum* that has established itself in science and that continues to establish itself on given foundations, then the enquiry is no longer directed to a subjective fact; it is directed instead to a fact that, to whatever extent self-propagating, is nevertheless objectively given, a fact grounded in principles. In other words, the enquiry is directed not to the process and apparatus of cognition, but to its result, to science itself. Then the unequivocal question arises: from which presuppositions does this fact of science derive its certainty? (Paulson, 1992: 330)

Rather than grounding science against the skeptic, as Kant meant to do, Cohen took the existence of an ordered body of knowledge as itself the objectively given fact, a fact that needs to be understood by identifying the presuppositions that make this fact possible.

This reasoning was soon generalized from physics to other domains of knowledge – historical knowledge, law, theology, and eventually to sociology, first by Georg Simmel and then by many other conceptual system builders. Each of these domains could be understood as conceptually self-constituted, distinct, and made distinct by the different grounding or organizing concepts that the everyday factual assertions within these domains presupposed.

By the end of the nineteenth century, when this strategy was at the peak of its influence, the entire neo-Kantian project of understanding domains as conceptually preconstituted, organizable into a scheme of intelligible categories (which could be discovered by conceptual analysis), was beginning to break up. The problem was relativism: the prospect of multiple presuppositional 'constitutings' of the subject matter in a domain, or defining a domain, fatally undermined the circular notion of a *Faktum der Wissenschaft*, which depended on the idea that one was not merely imposing a frame or 'lens' but identifying the facts in terms of the uniquely necessary conceptual structure of the science itself, and validating this structure by reference to a uniquely valid description of the facts themselves.

Dilthey and the Human Sciences

Dilthey was the thinker who put these issues together, in the form of a full consideration of the prospects of the human sciences as an alternative to the conception found in Mill and Comte. Although his basic starting point was Kantian, and concerned with systematic knowledge, he added to this the distinctive problem of human experience. We can think of the problematic of the human sciences in terms of a Kantian triangle: experience, the concepts forming the experiences, and the objects, texts, overt practices, and externals that are given meaning by humans. The neo-Kantians de-emphasized experience in favor of the concepts forming it and the evidence of concepts and their order. This led to a model of particular orders of life that was itself rigid, while life itself was not: human experience flowed and changed. Dilthey insisted on reminding the neo-Kantians that their constructions of experience were abstractions. But at the same time he argued that we have access to experience.

> The human sciences have the advantage over the cognition of nature that their object is not mere appearance given in external sensation as the mere phenomenal reflex of something real but an immediate reality itself. And this reality, moreover, is given in the form of a coherence experienced from within. Yet the very manner in which reality is given in inner experience gives rise to great difficulties in apprehending it objectively. (Dilthey, 1957: (*Gesammelte Schriften* vol. 5), 317–318); cited in Ermarth, 1978, 99)

Experience is thus a problem and an opportunity: an opportunity because we have a means other than reason and inference to understand others and the experiences we share with them, namely *Verstehen* or understanding as distinct from *Verstand*: "We explain through purely intellectual processes, but we understand through the cooperation of all the powers of the mind activated by apprehension" (Dilthey, 1957: (*Gesammelte Schriften* vol. 5), 172). And our capacity to understand, which depends on all our capacities as experiencing beings, allows us access to the experiences of others, such as the figures of history.

He also stressed another difference between the human and the natural sciences: relexivity. As he put it, "in the human sciences there prevails a special fundamental relation (*Grundverhaltnis*) such that we are living in society at once as an active agent (*wirkendes Element*) and also as aware of our effects (*unserer Wirkungen uns bewusst*)" (Dilthey, 1977: (*Gesammelte Schriften* vol. 18), 18; cited in Ermarth, 1978: 107). But also that we are both the products of our society and culture and their maker: "conditioned and unconditioned" (*bestimmt-unbestimmt*) ("Die Breslauer Ausarbeitung" (unpublished), 99; cited in Ermarth, 1978, 121).

The possibility and necessity of reflection on this situation is distinctive to the human sciences. And it leads to the conclusion that "we belong integrally to a total mental realm of which we are a part" (Dilthey, 1957: (*Gesammelte Schriften* vol. 5), 35; cited in Ermarth, 1978, 121). "The foremost trait of the individual is that it lives in something which is not itself" (Dilthey, 1959: (*Gesammelte Schriften* vol.1), 30; cited in Ermarth, 1978: 121). This led Dilthey to an insistence on the importance of culture as the total mental realm of which we are a part, but also to an insistence on the non-fixity of the mental realm, the vitality of the course of life, and the impossibility of the kind of neat reductions of a realm of thought to a hierarchy of conceptual relations central to neo-Kantianism. With Dilthey, as we have seen, there are differences in emphasis between different approaches to the human sciences with respect to which of these are prioritized. Nevertheless, the

central problem of the human sciences is that concepts and the subjective experiences of others need to be inferred from externals of some kind, including traditional sources, such as historical documents, but also more generally the externals of human action and expression.

Later History

The subsequent history of the human sciences does not follow a direct path. Two major influences, Durkheim and Weber, bear on the development of these ideas, but neither falls neatly within this development. Works in philosophy, such as Gaston Bachelard's *Poetics of Space* (1994), the *Annales* School of historians and its focus on mentalities, the rise of humanistic or phenomenological psychology and sociology, and the focus on everyday life of Walter Benjamin and French Marxism, all contribute more directly.

Durkheim's 'scientific' ideas, notably his ontology of the collective consciousness, were being abandoned or de-emphasized by the time of his death, but other 'Collective' ideas, especially those associated with cultural differences, collective memory, and techniques of the body, were promoted by the followers of Durkheim. One of the surviving ideas was the notion that the fundamental categories of thought were 'social.' With the emphasis on mentalities of the *Annales* School of historians, this produced a kind of working consensus about the pervasiveness of cultural influence and the possibility of studying it, especially through artifacts. An influential example of this kind of thinking was the work of Philippe Ariès on the history of childhood (1962). By looking at paintings and other expressions of culture, including books of advice, he showed that the concept of childhood had changed, and that childhood in the modern sense was an invented category.

Weber's influence was on the opposite pole: his argument that direct observational understanding of meaning was possible, and that explanations of action required both adequacy at the level of meaning and adequacy at the level of cause, together with the example of his own influential imaginative reconstruction of the mental and motivational work of Protestant early capitalists, had an important role. It validated the 'scientific' status of *Verstehen*, and also provided a model of social science hermeneutics.

The idea of culture, however, in a sense similar to Durkheim's, became a staple of French thought, a central element of the human sciences, and evolved in a way that merged with the concept of ideology, and provided the human sciences with a means of critique. This was a merger foreordained in the experience of neo-Kantianism. The experience of neo-Kantianism was that the frames which could be constructed to account for and provide conceptual order to a domain by identifying its presuppositions produced a serious problem of underdetermination. A domain like the law could be made sense of in terms of a variety of conflicting concepts – justice, equity, equality, fairness, legal validity, legal certainty, and so forth. Some of these accounts reflected political or valuative positions. Each of them could be interpreted as favoring one side or another in the social and political conflicts of the period.

Methodological Problems

The human sciences faced this diversity of analytic perspectives as a second-order problem. At the first level, the problem of the human sciences was to identify the frames, presuppositions, and categories that the participants themselves used to order their existence. The second-order problem was to select the account of these frames that provided a vocabulary for talking about them. When Foucault (1990: 257) discusses problematization, he characterizes it in these terms: "the totality of discursive or non-discursive practices that introduces something into the play of true and false and constitutes it as an object for thought (whether in the form of moral reflection, scientific knowledge, political analysis, etc.)." This captures some key ideas: the notion that there is a totality made up of mental or discursive elements that organize and present the world to us as meaningful and truth-relevant.

At the second level, however, there were many variations of these ideas. They range from the 'Structuralism' of thinkers like Claude Levi-Strauss to the more phenomenological reconstruction of lived experience through narrative. The difference is this: for the structuralist, the ordering principles need not be at all close to the conscious awareness of those whose conduct is being explained. They are mental, and shared, but they are revealed from the organization of such things as artifacts, the physical layout of villages, or the inner logic of myths exhibiting binary oppositions. In contrast, for the analyst concerned with lived experience, it is important to closely match and analyze the conscious life and thought of the subject, in spite of the fact that the researcher must rely on external evidence. This is of course another form of Dilthey's problem of the impossibility of capturing 'life' in thought and the necessity of relying on external facts to make inferences about the organizing mental frame on which our individual thought depends.

Karl Mannheim creatively extended the basic model of neo-Kantian constitution of objects of thought to the world of political ideology, and what he called 'utopias.' His thought was that utopias were partly true but idealized characterizations of real states of affairs that gave the appearance of reality. He quotes Droysen's idea of history as the realization of ideas:

> All movement in the historical world goes on in this way: Thought, which is the ideal counterpart of things as they really exist, develops itself as things- ought to be … (Mannheim, 1936: 199, §77)

as well as Droysen's comment on the implications of this for critique:

> Thoughts constitute the criticism of that which is and yet is not as it should be. Inasmuch as they may bring conditions to their level, then broaden out and harden themselves into accord with custom, conservatism, and obstinacy, new criticism is demanded, and thus on and on. (1936: 199, §78)

Reality is not yet in accord with the idea: thus it is open to critique from the standpoint of the idea. But the normal situation, Mannheim argues, is one in which the utopian thinker

transforms immediate experience in accordance with the organizing principle of the utopian image and wish:

> the utopian element—i.e. the nature of the dominant wish—which determines the sequence, order, and evaluation of single experiences. This wish is the organizing principle which even moulds the way in which we experience time. (1936: 209)

In short, experience itself is distorted or molded in a way that conceals reality and makes it appear in experience in accordance with the utopian wish. A similar idea of misrecognition is central to Foucault and Pierre Bourdieu. The thought is that it is intrinsic to the manner of thinking or ideology that it misrepresents or conceals aspects of reality, and that these aspect can be revealed, non-ideologically, by the analyst-critic.

The special value of these approaches has been in the way they have expanded the range of subject matter. The more 'experience' oriented approaches have revealed a great deal about the subjective experiences of social actors, and explored the diversity of subjective experiences, particularly with respect to gender and the experience of social exclusion. The kinds of analysis more oriented to external objects has brought into view a wide range of cultural objects and provided a second order account of their meaning, especially their role in misrecognition. These analyses have provided new accounts of exclusion and domination through practices that are nominally neutral or directed toward other purposes.

Nevertheless, there are issues with these techniques of analysis. The problem of circularity is central: the 'externals' used for the analyses are given meaning and selectively appropriated to provide a subject matter for analysis; the constitutive frame ascribed to the participants is inferred from the meanings of the external facts that are analyzed. This is a direct inheritance from neo-Kantianism, with the same result: a vast diversity of analyses. The inference to the frame is itself problematic. Not only is there a high degree of underdetermination in reasoning from an assemblage of meaningful things to the supposed presuppositions or constitutors of meaning, the argument is essentially functionalist.

The functionalist character of the reasoning (the end being the meaning of the objects or external facts, the frame functioning to give it a specific meaning) is especially problematic in connection with claims about misrecognition. Misrecognition implies that there is a purpose to the frame, namely to exclude or oppress, and the analyst infers this purpose or function from the facts of exclusion. But there is no feedback mechanism that assures that this purpose will continue to be pursued. So there is no way to assure that the oppression and exclusion are not merely accidental by-products of a given arrangement, and not its purpose.

These very abstract considerations can be brought to earth by considering their bearing on claims about racism and sexism: Are ostensibly neutral practices racist or sexist? The mere fact that their application leads to exclusion or oppression does not make them so. There needs to be a kind of purposive force. And inferences to this kind of force require evidence of feedback, or of some sort of supra-individual agent, which is not available. So these arguments collapse into functionalism (Elster, 2010).

The Sociology of the Human Sciences

The relation of the human sciences to ideological and political concerns has often been close. The fact that some core human science concepts lead directly to the possibility of critique has made them attractive to particular social and political movements. Foucault and Bourdieu, for example, provided analyses that were particularly influential on the Left and for feminism. The notion of misrecognition could be used as means of unmasking liberal ideological constructions of bureaucratic practices. Bourdieu and Jean-Claude Passeron's *Les Heriters* (1964), for instance, became an influential book among the teachers who worked in the educational system it described, which interpreted the 'neutral' liberal standards of the educational system as means of reproducing class distinction. A large part of the popularity of these thinkers was a result of these political affinities. But the history of the human sciences contains some quite different affinities. Werner Sombart's nationalistic methodological work, *Die Drei Nationaloekonomie* (1930) associated the need to understand experience through *Verstehen* with the German character, and derived this methodological impulse from Goethe and Nietzsche, and opposed the alien use of abstractions in economics. The concept of *Volksgeist*, notoriously, served nationalistic purposes.

Yet the most important 'sociological' aspect of the notion of the human sciences has been its role in opposing scientistic forms of social inquiry, and indeed the image of the social sciences as scientific in the sense of the natural sciences. By implicitly pointing to the specifically human aspect of these sciences – the fact of intelligible human experience – the term grounds a rejection of the reduction of this experience to quantification or experiment.

See also: Collective Memory, Anthropology of; Durkheim, Emile (1858–1917); Foucault, Michel (1926–84); Hermeneutics; Historicism; Kantian Ethics and Politics; Mannheim, Karl (1893–1947); Mauss, Marcel (1872–1950); Sociology, History of; Structuralism; Subaltern History; Weber, Max (1864–1920).

Bibliography

Ariès, P., 1962. Centuries of Childhood: A Social History of Family Life (R. Baldick, Trans.). Random House, New York (first published 1960).

Bachelard, G., 1994. The Poetics of Space: The Classic Look at How We Experience Intimate Spaces (M. Jolas, Trans.). Beacon Press, Boston (first published 1958; first English translation 1964).

Beiser, F., 2006. Kant and *Naturphilosophie*. In: Friedman, M., Nordmann, A. (Eds.), The Kantian Legacy in Nineteenth-Century Science. MIT Press, Cambridge, MA, pp. 7–26.

Bos, J., 2010. Individuality and interpretation in nineteenth-century German historicism. In: Feest, U. (Ed.), Historical Perspectives on Erklären and Verstehen. Springer, Dordrecht, pp. 207–220.

Bourdieu, P., Passeron, J.-C., 1964. Les Heriters: Les étudiants et la culture. Minuit, Paris.

Dilthey, W., unpublished ca. 1880. "Die Breslauer Ausarbeitung." Göttingen Nachlass. Cod. Ms. W. Dilthey, no.2. Niedersächsische Staats- und Universitätbibliothek, Göttingen.

Dilthey, W., 1957. In: Misch, G. (Ed.), Die geistige Welt: Einleitung in die Philosophie des Lebens. Erste Hälfte: Abhanlungen zur Grundelgung der Geistswissenschaften, second ed. BG Teubner, Stuttgart. Vandenhoeck & Ruprecht, Göttingen. (first published 1924; English edition Poetry and Experience 1985).

Dilthey, W., 1959. In: Groethuysen, B. (Ed.), Einleitung in die Geisteswissenschaften: Versuch einer Grundlegung für das Studium der Gesellschaft und der Geschicte, fourth ed. BG Teubner, Stuttgart. Vandenhoeck & Ruprecht, Göttingen. (English edition Introduction to the Human Sciences, 1989).
Dilthey, W., 1977. In: Johach, H., Rodi, F. (Eds.), Die Wissenschaften vom Menschen, der Gesellschaft und der Geschichte: Vorarbeiten zur Einleitung in die Geisteswissenschaften. BG Teubner, Stuttgart. Vandenhoeck & Ruprecht, Göttingen. (first published 1865–1880).
Elster, J., 2010. Obscurantisme dur et obscurantisme mou dans les sciences humaines et sociales. Diogène 1, 231–247.
Ermarth, M., 1978. Wilhelm Dilthey: The Critique of Historical Reason. The University of Chicago Press, Chicago, IL.
Fichte, J.G., 2005. The Science of Knowing: Fichte's 1804 Lectures on the Wissenschaftslehre (W. Wright, Trans.). State University of New York Press, Albany (first published 1804).
Foucault, M., 1990. In: Kritzman, L. (Ed.), Politics, Philosophy, Culture: Interviews and Other Writings 1977–1984. Routledge, New York (first published 1988).
Foucault, M., 1994. Dits et écrits: 1954–1988, vols. 1–4. Gallimard, Paris.
Han, B., 2002. Foucault's Critical Project: Between the Transcendental and the Historical (E. Pile, Trans.). Stanford University Press, Stanford, CA.
Maclean, M.J., 1982. Johann Gustav Droysen and the development of historical hermeneutics. History and Theory 21 (3), 347–365.
Mannheim, K., 1936. Ideology and Utopia: An Introduction to the Sociology of Knowledge (L. Wirth, E. Shils, Trans.). Harcourt, Brace & World, New York (first published 1929).
Paulson, S.L., 1992. The Neo-Kantian dimension of Kelsen's pure theory of law. Oxford Journal of Legal Studies 12 (3), 311–332.
Schmaus, W., 2004. Rethinking Durkheim and His Tradition. Cambridge University Press, Cambridge.
Schopenhauer, A., 2010. The World as Will and Representation. Cambridge University Press, Cambridge (first published 1818).
Sombart, W., 1930. Die drei Nationalökonomien. Duncker & Humblot, München/Leipzig.

Relevant Websites

http://www.fhhs.org/ – Forum for History of Human Science.
http://hhs.sagepub.com/ – History of the Human Sciences.
http://www.artfulsoftware.com/humanscienceresearch.html – Human Science Research Studies.
http://www.templeton.org/what-we-fund/core-funding-areas/science-and-the-big-questions/human-sciences – John Templeton Foundation: Human Sciences.
http://www.theoryofknowledge.net/areas-of-knowledge/the-human-sciences/human-sciences-starting-points – Theory of Knowledge. Net: human sciences starting points.
http://en.wikipedia.org/wiki/Category:Human_sciences – Wikipedia Category: Human sciences.
http://www.marxists.org/reference/subject/philosophy/works/ge/dilthey.htm – William Dilthey Introduction to the Human Sciences.

Human Spatial Orientation, Neural Basis of

Aiden E Arnold, Irene Liu, Kara Murias, Edward Slone, Ford Burles, Veronica Guadagni, and Giuseppe Iaria,
University of Calgary, Calgary, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by P.E. Sharp, volume 15, pp. 10413–10417, © 2001, Elsevier Ltd.

Abstract

The ability to spatially orient is one of the most complex cognitive skills. It relies on the processing of different spatial information available in the environment, leading to the use of different cognitive strategies useful for orientation. The neural mechanisms underlying such a complex phenomenon include networks that require the proper functioning of a variety of cortical and subcortical regions. Here, we review the basic behavioral and neural mechanism responsible for the ability to orient in spatial surroundings, as well as its development (and lack of) in children.

Humans Navigate Using Egocentric and Allocentric Spatial Information

The ability to navigate in spatial environments is one of the most fundamental behaviors that animals and humans display. It allowed our ancestors to hunt and forage, explore new environmental niches, and settle communities. It remains a critical ability in the modern world by allowing humans to wayfind through increasingly complex and distributed urban surroundings.

Humans navigate through environments by relying on two forms of topological memory that produce spatial representations differing in frames of reference, learning strategies, and neuroanatomical structures. The first memory system generates egocentric spatial representations that are used to guide wayfinding to locations in a person's immediate environment by relating spatial information to the body of the individual (Mou et al., 2008). The second memory system generates allocentric spatial representations that are used to guide wayfinding to locations outside of a person's immediate environment by using long-term memories about the arrangement of environmental landmarks independent of a person's actual location.

Egocentric Spatial Representations Are Generated in the Parietal Cortex

Past research in animals and humans suggests that Brodmann area 7a in the posterior parietal cortex (PPC) is the primary cortical area involved in generating and maintaining the egocentric spatial representations used for navigation (Burgess et al., 2001; Byrne et al., 2007). While sensory, information from the environment is processed in modality specific areas of the brain, area 7a and other posterior and medial regions in the parietal cortex integrate inputs from these sensory areas into sensorimotor maps. These sensorimotor maps code body-to-object relationships that can be used to guide wayfinding by outputting signals to the motor cortex. Neurons in area 7a exhibit 'gain field' responses that modulate their firing rate to visual stimuli depending on (1) the orientation of gaze direction relative to the head (Andersen et al., 1987), (2) the orientation of the head relative to the body, and (3) the orientation of the body relative to the environment (Snyder et al., 1998). These gain field responses are believed to allow for effective and coordinated translations between the various body-centered egocentric reference frames necessary for movement through spatial environments (Burgess, 2008; Deneve et al., 2001; Zipser and Andersen, 1988).

Sequence-based circuit dynamics in the PPC have been shown to provide stable firing characteristics in relation to egocentric navigation, suggesting that parietal-based spatial representations rely on a form of working memory that encode environmental information relevant to a current navigation task. Using two-photon microscopy to image layer two-thirds of PPC neurons while mice ran through a virtual T-maze, Harvey et al. (2012) found that temporal sequences of parietal neurons in mice predicted whether a left or right turn was made at the T intersection, a decision based on visual cues located early in the maze that coded the location of a reward. The researchers found that left and right, turns produced dissociable temporal sequences between the visual cue and the decision point at which a turn was required. High variability in the temporal sequence was found to be associated with a response error on that trial. This suggests the firing dynamics of certain PPC circuits specify choice-related information and may provide the basis for a spatial working memory system that is essential for movement planning and navigation. Additionally, the researchers found that the active cell in the sequence encoded information about the spatial and temporal progression through the maze independent of actual location. These findings imply that the PPC is able to use circuit dynamics to encode choice-specific information related to accurate navigation, as well as progression through an environment by self-motion, independent of location in the maze.

Allocentric Spatial Representations Are Processed in the Medial Temporal Lobe

The ability to represent the spatial layout of environmental features is contingent upon the effective integration of various types of neurons in the medial temporal lobes that exhibit distinct firing properties in relation to certain aspects of space.

These cell types are known as place, head direction, grid, and view cells.

Place Cells

Through an elegant set of studies in the 1970s, O'Keefe et al. provided the first evidence that a subset of neurons within the rat hippocampus fired when the animal entered a specific area of an environment (O'Keefe and Dostrovsky, 1971; O'Keefe, 1976; O'Keefe and Nadel, 1978). Termed 'place cells,' these neurons display place fields – receptive fields within an environment – where there is an increase in firing rate as the rat approaches a specific location in the environment. Although one place cell represents a limited portion of an environment, taken as a population the hippocampal place cells are able to represent any location through a unique combination of firing rates from different place fields. Based on these findings, O'Keefe and Nadel (1978) hypothesized that place cells serve as the neural basis of a cognitive map, an internal representation of the spatial layout of an environment and spatial relationship between landmarks within it, that is used by animals and humans for flexible and efficient navigation.

Subsequent studies on the signaling properties of place cells suggest possible mechanisms through which place fields are generated. Looking at the influence of environmental landmarks, Muller and Kubie (1987) found that a white cue card placed on the boarder of a gray cylinder produced a set of distinct place fields. When the experimenters rotated the white cue card along the edge of the cylinder, they found that the associated place fields rotated to a similar degree. However, the place fields randomly changed their angular location when the cue card was removed, although they remained, firing in only a single location. This suggests that salient landmarks are able to modulate place fields but are not necessary for their ongoing activity. These findings provide converging evidence toward one of O'Keefe's (1976) original findings that place cells persist even in dark environments when landmarks are not viewable.

If place fields are not fixed to environmental landmarks, how else might they be generated? A study by Sharp et al. (1995) suggests that vestibular signaling plays an equally important role in stabilizing place fields. The researchers recorded place fields in a black and white striped cylinder that was constructed so that the walls and the floor could be rotated independently. They found that when the walls and floor were slowly rotated together in the dark, the place fields rotated as well. However, when either the floor or the wall was rotated quickly with the lights on, the place fields remained stable. It was concluded that the rats could use vestibular cues about self-motion to detect incongruences between self-motion and visual inputs. Another prominent model that has been proposed to modulate place fields is the 'boundary vector model' (Hartley et al., 2000). Under this model, boundary vector cells (BVCs) – neurons with receptive fields that increase firing when moving toward an environmental boundary – are input to hippocampal place cells, which sum the BVC inputs and generate place fields at a given location relative to the surrounding boundaries.

Place cells have also been directly recorded in the human hippocampus and parahippocampal region. By implanting electrodes in the medial temporal lobes of patients with epilepsy, Ekstrom et al. (2003) recorded from 317 neurons while patients played a navigation game in a simple virtual environment. The researchers observed that 26% of the cells responded to the patient's location in the environment, 21% responded to the desired destination of the game, and 12% responded to views of specific landmarks. Of the cells in the hippocampus, 24% were bona fide place responsive cells. Studies in humans using functional magnetic resonance imaging (fMRI) have supported the role of the hippocampus in encoding and using cognitive maps of virtual environments (Hartley et al., 2003; Iaria et al., 2007; Iaria et al., 2003; Maguire et al., 1998).

Head Direction Cells

Just as place cells are able to represent an individual's location independent of the orientation of their body, scientists have discovered a cell class that represents the orientation of an individual's head independent of their actual location. Termed 'head direction' (HD) cells, these neurons fire over a rotational range of 90° with maximal firing occurring near the middle of the range (Taube et al., 1990a) and are contingent upon the orientation of the head within an environment. Different HD cells fire at different orientations so that the entire 360° range can be represented by the combination of relative HD cell firing rates (Taube et al., 1990b). These cells are largely found along an anatomical circuit that resembles Papez's circuit (Papez, 1937), stretching from the dorsal tegmental nucleus to the lateral mammillary nucleus and the anterior thalamus into the postsubiculum. HD cells have also been found in the retrosplenial (Chen et al., 1994) and medial entorhinal cortex (Sargolini et al., 2006).

The HD cell signal is believed to arise from angular head velocity signals produced by the vestibular system. Lesion studies in rats provide support for this by showing that disruption of the vestibular system removes the directional firing of HD cells along the anterior thalamic nucleus (Stackman and Taube, 1997). Salient environmental landmarks, as well as the geometry of environmental boundaries, have also been found to modulate the directional orientation of HD cell firing fields (Taube et al., 1990b), suggesting that HD signals may be modulated in part by visual input.

Grid Cells

First described in 2004 by Fyhn et al. (2004), 'grid' cells represent one of the most important discoveries in understanding the cellular basis of how the brain maps spatial environments. These cells are located primarily in the medial entorhinal cortex and display multiple place fields arranged in a triangular grid. Like place cells, the relative firing rate of grid cells at a population level is able to represent the entire spatial layout of an environment. As such, grid cells have been proposed as an alternative means through which location in an environment can be represented and their firing properties have been shown to influence place cell activity (McNaughton et al., 2006; Fyhn et al., 2007) through projections from the dorsal medial entorhinal cortex to the dorsal hippocampus.

Systems neuroscience models of grid cell dynamics hypothesize that speed-modulated signals from HD cells input to velocity controlled oscillators that form the basis of grid cell activity (Burgess et al., 2007; Jeewajee et al., 2008). Supporting these models, the deep layers of the medial entorhinal cortex have been found to contain not only grid cells, but also HD cells and conjunctive cells that represent an individual's location and head orientation (McNaughton et al., 2006; Sargolini et al., 2006). This integration of speed, velocity, and orientation information in the medial entorhinal cortex has led some researchers to suggest that grid cells form the basis of a path integrator that allows for navigation through environments using path integration (i.e., updating of location relative to a fixed point by using self-motion cues about spatial displacement) (Fyhn et al., 2007; Hafting et al., 2005; McNaughton et al., 2006). McNaughton et al.'s (2006) model builds on the idea of a path integrator and suggests that the outputs of grid cell firing into the hippocampus are used to set the spatial scale upon which the size of place fields are generated.

View Cells

Spatial view cells are neurons in the human parahippocampal region that have been found to display selective firing when an individual looks at a certain object or person within a spatial environment. In their study using electrodes implanted in epileptic patients, Ekstrom et al. (2003) found that 88% of view cells fired preferentially to a single object during navigation through a virtual town. These cells were found to be either location dependent (30%) or location independent (70%), the latter responding when an object was viewed from any position within the environment. The prevalence of location-independent view cells led the researchers to hypothesize that the parahippocampal region is extracting allocentric information based on salient landmarks, and inputs a coarse representation of space to the hippocampus. The hippocampus then integrates various inputs into a flexible cognitive map of the environment used for navigation. The role of the parahippocampal region in processing allocentric representations of space is supported by the research by Epstein and Kanwisher (1998) on a 'parahippocampal place area' identified using fMRI in humans, which is believed to encode the geometry of local spatial environments.

Hippocampal Structure and Function Are Associated with Allocentric Spatial Representations in Humans

Converging with Ekstrom et al.'s (2003) findings on place and view cells in the human hippocampus, numerous brain imaging studies have further implicated the hippocampus as a critical structure underlying the formation and use of allocentric spatial representations during navigation. Wolbers and Büchel (2005) found that the hippocampus was necessary for learning the spatial layout of a virtual environment, and that activity increased in the left hippocampus on trials where the greatest spatial learning was required. Similarly, Iaria et al. (2007) found that the left anterior hippocampus was preferentially engaged while participants learned the spatial layout of six landmarks in a virtual city. However, the researchers found that the right posterior hippocampus displayed increased activity while participants used their knowledge of the arrangement of landmarks to navigate between them in the shortest route possible. This suggests a possible dichotomy between the functional role of the left and right hippocampus in relation to encoding and using a cognitive map within a virtual environment.

The functional role of the hippocampus was comprehensively evaluated in a study by Spiers and Maguire (2006). The researchers scanned the brains of London taxi drivers while they navigated through an interactive virtual simulation of London. The task required that the participants engage in both directed (i.e., computer generated) and spontaneous route planning. After the scan session, the researchers asked that the taxi drivers watch a video that had been recorded of them performing the task and give a verbal description of what they were thinking during the experiment. This allowed them to pair neural activity with in-task demands, as well as subjective cognitive processes. The researchers observed that the left hippocampus, but not the right, showed increased activity while planning directed routes. However, neither the left or right hippocampus showed increased activity during spontaneous route planning. These results suggest that the role of the hippocampus during human navigation may be somewhat transient, showing increases in neural firing during periods of attentive route planning.

Variability in the structural characteristics of the hippocampus has also been found to be significantly related to navigation skills in humans. A study by Bohbot et al. (2007) found that participants who used a spatial learning strategy during a radial arm maze task had significantly more gray matter in their hippocampus than nonspatial learners. Conversely, the nonspatial learners were found to have significantly more gray matter in their caudate nucleus, an area associated with response learning during wayfinding. The influence of hippocampal microstructure was further evaluated by Iaria et al. (2008) in a diffusion tensor imaging study. Participants with high fractional anisotropy – a, measure of the structural integrity of brain structures – in their right hippocampus were found to encode allocentric spatial representations quicker and were able to use them more efficiently to orient and navigate within a virtual environment.

Egocentric and Allocentric Spatial Representations Are Integrated across a Distributed Brain Network into Mental Representations of an Environment

A direct implication from the decades of research on single and multicell recording in animal models is that spatial navigation occurs through a distributed brain network that integrates various features of space into a flexible mental representation. Positron emission tomography (PET), fMRI, and electroencephalography studies in humans have begun to illuminate the intricacies involved in the neural systems underlying spatial navigation and given insight into the neural basis of cognitive maps.

One of the pioneering studies on the human navigation network was carried out by Maguire et al. (1998) using PET imaging while participants navigated a virtual town. The researchers found widespread activity throughout the brain in structures that were homologous to spatial areas identified in animal models. Of particular interest was regional activity in the right hippocampus, which reflected accurate navigation through the town, and in the right caudate nucleus, which correlated with speed of navigation. These two areas were supported by activities in inferior and medial parietal regions, suggesting that allocentric spatial representations processed within the medial temporal lobes were being integrated with egocentric representations in the parietal cortex in order to successfully navigate through the town.

The differential engagement of hippocampal and caudate structures during navigation were further delineated through a series of studies in the 2000s. Iaria et al. (2003) asked individuals to solve a virtual-arm radial maze while they recorded brain activity. The authors found that participants were able to solve the task by spontaneously using one of two different strategies: a spatial strategy in which individuals made use of environmental landmarks in order to get to different locations, or a nonspatial strategy in which they navigated by counting number of turns without relying on environmental landmarks. The researchers found that the spontaneous use of spatial vs nonspatial strategies in humans correlated with increased activity in the hippocampus and caudate nucleus, respectively. Similar findings were reported by Hartley et al. (2003) who showed that good navigators used a hippocampal dependent system, while wayfinding and a caudate dependent system while following an explicit route. This hippocampus/caudate nucleus dichotomy was taken further in a study by Marchette et al. (2011). The researchers showed that variability in place vs response strategies during navigation could be explained by the proportional engagement of the hippocampus and caudate nucleus relative to one another. Taken together, these studies suggest that there is interindividual and task specific modulation of a core network that supports spatial navigation, and that variability in this modulation can influence spatial behavior.

Systems neuroscience models put forth by Burgess et al. (Burgess, 2008; Burgess et al., 2001; Byrne et al., 2007) describe a computational perspective on how medial temporal-based allocentric representations and parietal-based egocentric representations integrate within the human brain. These models suggest that hublike structures in the cortex – the precuneus, retrosplenial cortex, and hippocampus, primarily – integrate allocentric spatial representations that code landmark and location-dependent information with egocentric spatial representations that code spatial information in the immediate environment. The result is a flexible mental representation – a cognitive map – that can be used to simulate movement through space and plan routes to guide wayfinding. Arnold et al. (2013) complement this model from a cognitive perspective by hypothesizing that the navigation strategies used by humans exist on a continuum defined by the degree of integration, where highly efficient modes of navigation require coordinated processing of both egocentric and allocentric spatial representations.

Variability in Large-Scale Neural Network Dynamics Influences the Capacity to Navigate Using Mental Representations

While systems neuroscience models provide a comprehensive account of the computational processes that underlie spatial navigation, they provide limited insight into how variability in network dynamics influences spatial behavior. To address this, Arnold et al. (2013) investigated whether individual variability in using egocentric and allocentric spatial representations to accurately orient could be explained through variability in certain network dynamics. The researchers had participants perform an fMRI study where they (1) encoded the spatial layout of a virtual city and (2) used landmarks within, the city to make orientation decisions. Analysis of the whole group data revealed engagement of a canonical navigation network during the encoding task and engagement of a retrosplenial-based system during the orientation task. However, when accuracy of orientation performance was included in the analysis, good performers were found to engage an additional network involving the left hippocampal region and the right supramarginal gyrus over and above the network expressed by the whole group. This suggests that variability in the configuration of task-related functional networks influences the ability to integrate allocentric and egocentric spatial representations for the purpose of orientation.

The researchers then used the functional networks identified during the orientation task in a resting-state fMRI analysis to further investigate whether variability in network properties influenced spatial behavior. Two specific topological characteristics indexing network integration were evaluated using graph theory – global efficiency across the entire functional network and node centrality of the additional network regions expressed by accurate participants. Both whole network global efficiency and node centrality in the left hippocampal region were found to be highly predictive of accuracy, during the orientation task. These results suggest that intrinsic connectivity properties of functional networks measured during rest influence the degree to which individuals are able to generate mental representations in order to effectively orient. Similar results have also been documented in the neural encoding of spatially relevant landmarks, where connectivity measured through resting-state fMRI signal between the hippocampus and parahippocampal gyrus relates to an individual's navigation ability (Wegman and Janzen, 2011).

Children's Ability to Navigate Is Related to the Development of Specific Functional Brain Networks

Development of spatial cognition in children is supported by prominent functional and structural brain changes that occur in a nonlinear fashion (Johnson, 2001). A longitudinal study of children and adolescents between the ages of 4 and 21 years showed that gray matter first matures at the primary sensorimotor cortex, and the frontal and occipital pole. The remaining higher order cortical areas develop in a parietal to frontal direction, and the superior temporal cortex is last to mature (Gogtay et al., 2004). There are also significant

large-scale structural network changes in terms of topological properties – brains of children older than 8 years of age begin to exchange local efficiency and modularity for increased global efficiency. In addition, the number and distribution of connector hubs show age-related variations that eventually localize to multimodal association areas. Finally, interregional structural connectivity increases in areas that are the last to mature (Khundrakpam et al., 2012).

These developmental patterns underlie the acquisition of spatial orientation skills, which progresses from using simple environmental cues to forming complex mental representations of large-scale environments. Children as young as 8-month old show primitive beacon use by associating a cue object with a goal location during visual searches (Crowther et al., 2000). With mobility, infants begin to use beacons as distant clues to orient toward. Higher level encoding of body movement after the first year of life forms the basis of path integration and allow children to continuously update their spatial location from vestibular and proprioceptive systems (Loomis et al., 2001). Parietal and frontal lobe maturity is accompanied by increased attention, executive function, and memory skills, and as a result, children are better able to encode distances and direction between landmarks and goal locations (Bushnell et al., 1995). Language (Shusterman et al., 2011) and body orientation (Wang and Spelke, 2002) further support children in heading orientation and learning routes based on sequential turns. Finally, a protracted development of the temporal lobe result in complex cognitive maps as one of the last spatial orientation strategies to mature (Piper et al., 2010).

There has been limited study into developmental disorders of orientation skills in children. To date there are published examples of congenital malformations (Griffiths and Hunt, 1984; Iaria et al., 2005), perinatal injury (Brunsdon et al., 2007; Pavlova et al., 2007), and early traumatic brain injury (Lehnung et al., 2003) resulting in topographical disorientation, but full descriptions of the deficits are restricted to isolated case studies. Specific visuospatial deficits associated with impaired orientation skills have been demonstrated in William's syndrome (Nardini et al., 2008) indicating a possible model to investigate how genetic differences may contribute to navigation capacity. The careful study of individual variability or deficits in orientation skills as they emerge during childhood may provide better understanding of brain development and the neurobiological underpinnings of these skills (Anderson et al., 2011). The interaction of different lesions through periods of development has potential for illuminating the brain's strategies for adaption to insult.

Recently, a set of case studies describing a newly identified condition termed developmental topographical disorientation (DTD) (Iaria and Barton, 2010; Iaria et al., 2009) provide evidence that the lack of development in the ability of children to mentally represent an environment can occur in the absence of brain lesions. In the first documented case study, Iaria et al. (2009) describe an individual who was unable to generate cognitive maps of novel environments despite having no other cognitive deficits or structural abnormalities in the brain. Using fMRI, the researchers measured brain activity while this individual explored a virtual environment with several distinct landmarks. Compared to a control group, this individual displayed significantly lower brain response in the hippocampus and retrosplenial cortex while attempting to encode the spatial layout of the virtual city. Though based on a single case, these results suggest that within DTD there is selective disengagement of brain regions critical for encoding the arrangements of landmarks, which may translate into a selective impairment in generating cognitive maps. A follow-up study on 120 new cases of DTD (Iaria and Barton, 2010) supports the findings described in the original first case, where the ability to encode cognitive maps, was the primary navigation skill impaired on a test battery assessing a variety of orientation skills. The neural mechanisms underlying this unique developmental condition are under investigation. However, given the knowledge gained in the last few years, it is reasonable to speculate that a selective orientation defect as described in individuals with DTD may be related to the lack of proper functional connectivity between the various brain regions critically involved in the complex phenomenon of spatial orientation.

See also: Aging and Memory; Hippocampus and Related Structures; Learning and Memory, Neural Basis of; Spatial Navigation.

Bibliography

Andersen, R.A., Essick, G.K., Siegel, R.M., 1987. Neurons of area 7 activated by both visual stimuli and oculomotor behavior. Experimental Brain Research 67, 316–322.

Anderson, V., Spencer-Smith, M., Wood, A., 2011. Do children really recover better? Neurobehavioural plasticity after early brain insult. Brain 134, 2197–2221.

Arnold, A.E.G.F., Burles, F., Krivoruchko, T., Liu, I., Rey, C.D., Levy, R.M., Iaria, G., 2013. Cognitive mapping in humans and its relationship to other orientation skills. Experimental Brain Research 224, 359–372.

Arnold, A.E.G.F., Protzner, A., Bray, S., Levy, R.M., Iaria, G., 2013. Neural network configuration and efficiency underlies individual differences in spatial orientation ability. Journal of Cognitive Neuroscience. http://dx.doi.org/10.1162/jocn_a_00491 (online ahead of print).

Bohbot, V.D., Lerch, J., Thorndycraft, B., et al., 2007. Gray matter differences correlate with spontaneous strategies in a human virtual navigation task. Journal of Neuroscience 27, 10078–10083.

Brunsdon, R., Nickels, L., Coltheart, M., Joy, P., 2007. Assessment and treatment of childhood topographical disorientation: a case study. Neuropsychological Rehabilitation 17, 53–94.

Burgess, N., 2008. Spatial cognition and the brain. Annals of the New York Academy of Sciences 1124, 77–97.

Burgess, N., Barry, C., O'Keefe, J., 2007. An oscillatory interference model of grid cell firing. Hippocampus 17, 801–812.

Burgess, N., Becker, S., King, J.A., O'Keefe, J., 2001. Memory for events and their spatial context: models and experiments. Philosophical Transaction of the Royal Society of London. Series B, Biological Sciences 356, 1493–1503.

Bushnell, E.W., McKenzie, B.E., Lawrence, D.A., Connell, S., 1995. The spatial coding strategies of one-year-old infants in a locomotor search task. Child Development 66, 937–958.

Byrne, P., Becker, S., Burgess, N., 2007. Remembering the past and imagining the future: a neural model of spatial memory and imagery. Psychological Review 114, 340–375.

Chen, L.L., Lin, L.H., Green, E.J., Barnes, C.A., McNaughton, B.L., 1994. Head-direction cells in the rat posterior cortex. I. Anatomical distribution and behavioral modulation. Experimental Brain Research 101, 8–23.

Crowther, H.L., Lew, A.R., Whitaker, C.J., 2000. The development of beacon use for spatial orientation in 6–8.5-month-old infants. Infant Behavior and Development 23, 41–59.

Deneve, S., Latham, P.E., Pouget, A., 2001. Efficient computation and cue integration with noisy population codes. Nature Neuroscience 4, 826–831.
Ekstrom, A.D., Kahana, M.J., Caplan, J.B., et al., 2003. Cellular networks underlying human spatial navigation. Nature 425, 184–188.
Epstein, R., Kanwisher, N., 1998. A cortical representation of the local visual environment. Nature 392, 598–601.
Fyhn, M., Hafting, T., Treves, A., et al., 2007. Hippocampal remapping and grid realignment in entorhinal cortex. Nature 446, 190–194.
Fyhn, M., Molden, S., Witter, M.P., et al., 2004. Spatial representation in the entorhinal cortex. Science 305, 1258–1264.
Gogtay, N., Giedd, J.N., Lusk, L., et al., 2004. Dynamic mapping of human cortical development during childhood through early adulthood. Proceedings of the National Academy of Sciences of the United States of America 101, 8174–8179.
Griffiths, P., Hunt, S., 1984. Specific spatial defect in a child with septo-optic dysplasia. Developmental Medicine and Child Neurology 26, 395–400.
Hafting, T., Fyhn, M., Molden, S., et al., 2005. Microstructure of a spatial map in the entorhinal cortex. Nature 436, 801–806.
Hartley, T., Burgess, N., Lever, C., et al., 2000. Modeling place fields in terms of the cortical inputs to the hippocampus. Hippocampus 10, 369–379.
Hartley, T., Maguire, E.A., Spiers, H.J., Burgess, N., 2003. The well-worn route and the path less traveled: distinct neural bases of route following and wayfinding in humans. Neuron 37, 877–888.
Harvey, C.D., Coen, P., Tank, D.W., 2012. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. Nature 484, 62–68.
Iaria, G., Barton, J.J.S., 2010. Developmental topographical disorientation: a newly discovered cognitive disorder. Experimental Brain Research 206, 189–196.
Iaria, G., Bogod, N., Fox, C., 2009. Developmental topographical disorientation: case one. Neuropsychologia 47, 30–40.
Iaria, G., Chen, J., Guariglia, C., et al., 2007. Retrosplenial and hippocampal brain regions in human navigation: complementary functional contributions to the formation and use of cognitive maps. European Journal of Neuroscience 25, 890–899.
Iaria, G., Incoccia, C., Piccardi, L., et al., 2005. Lack of orientation due to a congenital brain malformation: a case study. Neurocase 11, 463–474.
Iaria, G., Lanyon, L., Fox, C., Giaschi, D., 2008. Navigational skills correlate with hippocampal fractional anisotropy in humans. Hippocampus 18, 335–339.
Iaria, G., Petrides, M., Dagher, A., Pike, B., 2003. Cognitive strategies dependent on the hippocampus and caudate nucleus in human navigation: variability and change with practice. Journal of Neuroscience 23, 5945–5952.
Jeewajee, A., Barry, C., O'Keefe, J., Burgess, N., 2008. Grid cells and theta as oscillatory interference: electrophysiological data from freely moving rats. Hippocampus 18, 1175–1185.
Johnson, M.H., 2001. Functional brain development in humans. Nature Reviews Neuroscience 2, 475–483.
Khundrakpam, B.S., Reid, A., Brauer, J., et al., 2012. Developmental changes in organization of structural brain networks. Cerebral Cortex. http://dx.doi.org/10.1093/cercor/bhs187 (online ahead of print).
Lehnung, M., Leplow, B., Ekroll, V., et al., 2003. Recovery of spatial memory and persistence of spatial orientation deficits after traumatic brain injury during childhood. Brain Injury 17, 855–869.
Loomis, J.M., Klatzky, R.L., Golledge, R.G., 2001. Navigating without vision: basic and applied research. Optometry and Vision Science 78, 282–289.
Maguire, E., Burgess, N., Donnett, J., et al., 1998. Knowing where and getting there: a human navigation network. Science 280, 921–924.
Marchette, S.A., Bakker, A., Shelton, A.L., 2011. Cognitive mappers to creatures of habit: differential engagement of place and response learning mechanisms predicts human navigational behavior. Journal of Neuroscience 31, 15264–15268.
McNaughton, B.L., Battaglia, F.P., Jensen, O., et al., 2006. Path integration and the neural basis of the 'cognitive map'. Nature Reviews Neuroscience 7, 663–678.
Mou, W., Xiao, C., McNamara, T.P., 2008. Reference directions and reference objects in spatial memory of a briefly viewed layout. Cognition 108, 136–154.
Muller, R.U., Kubie, J.L., 1987. The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. Journal of Neuroscience 7, 1951–1968.
Nardini, M., Atkinson, J., Braddick, O., Burgess, N., 2008. Developmental trajectories for spatial frames of reference in Williams syndrome. Developmental Sciences 11, 583–595.
O'Keefe, 1976. Place units in the hippocampus of the freely moving rat. Experimental Neurology 51, 78–109.
O'Keefe, J., Dostrovsky, J., 1971. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain Research 34, 171–175.
O'Keefe, J., Nadel, L., 1978. The Hippocampus as a Cognitive Map. Clarendon, Oxford.
Papez, J., 1937. A proposed mechanism of emotion. Archives of Neurology and Psychiatry 38, 725–743.
Pavlova, M., Sokolov, A., Krägeloh-Mann, I., 2007. Visual navigation in adolescents with early periventricular lesions: knowing where, but not getting there. Cerebral Cortex 17, 363–369.
Piper, B.J., Acevedo, S.F., Craytor, M.J., et al., 2010. The use and validation of the spatial navigation Memory Island test in primary school children. Behavioural Brain Research 210, 257–262.
Sargolini, F., Fyhn, M., Hafting, T., et al., 2006. Conjunctive representation of position, direction, and velocity in entorhinal cortex. Science 312, 758–762.
Sharp, P.E., Blair, H.T., Etkin, D., Tzanetos, D.B., 1995. Influences of vestibular and visual motion information on the spatial firing patterns of hippocampal place cells. Journal of Neuroscience 15, 173–189.
Shusterman, A., Ah Lee, S., Spelke, E.S., 2011. Cognitive effects of language on human navigation. Cognition 120, 186–201.
Snyder, L.H., Grieve, K.L., Brotchie, P., Andersen, R.A., 1998. Separate body- and world-referenced representations of visual space in parietal cortex. Nature 394, 887–891.
Spiers, H.J., Maguire, E.A., 2006. Thoughts, behaviour, and brain dynamics during navigation in the real world. Neuroimage 31, 1826–1840.
Stackman, R.W., Taube, J.S., 1997. Firing properties of head direction cells in the rat anterior thalamic nucleus: dependence on vestibular input. Journal of Neuroscience 17, 4349–4358.
Taube, J.S., Muller, R.U., Ranck, J.B., 1990a. Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. Journal of Neuroscience 10, 420–435.
Taube, J.S., Muller, R.U., Ranck, J.B., 1990b. Head-direction cells recorded from the postsubiculum in freely moving rats. II. Effects of environmental manipulations. Journal of Neuroscience 10, 436–447.
Wang, R., Spelke, E., 2002. Human spatial representation: insights from animals. Trends in Cognitive Sciences 6, 376–382.
Wegman, J., Janzen, G., 2011. Neural encoding of objects relevant for navigation and resting state correlations with navigational ability. Journal of Cognitive Neuroscience 23, 3841–3854.
Wolbers, T., Büchel, C., 2005. Dissociable retrosplenial and hippocampal contributions to successful formation of survey representations. Journal of Neuroscience 25, 3333–3340.
Zipser, D., Andersen, R.A., 1988. A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. Nature 331, 679–684.

Relevant Website

www.gettinglost.ca.

Human Terrain System (United States): Critique

Maximilian C Forte, Concordia University, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The U.S. Army's launch of the Human Terrain System (HTS) in 2007, initially attempting to incorporate anthropologists into a program designed to gather 'ethnographic intelligence' in support of counterinsurgency requirements, arguably met with limited success and spawned intense debate through 2011. While the HTS involved a range of social scientists and regional experts, anthropologists in US universities almost exclusively dominated in articulating critiques of HTS. Fundamental ethical, methodological, and certain theoretical tools of the discipline were put to the test, and sharpened. Yet some critical political positions were marginalized, while key epistemological and political stances on other arenas of collaboration with the state, and the nature and scope of critique itself, were at best addressed ambiguously.

Unfolding with particular intensity from late 2007 through 2011, debates abounded about the role of academics, and specifically anthropologists, in the U.S. Army's Human Terrain System (HTS) program. Especially challenging for an understanding of what was argued and why is the fact that there were several very different critiques that often contrasted with one another. On one side, criticisms of HTS were to be found among those employed in the program, whether military or civilian; among disillusioned former proponents of the program, in academia; and within the military that housed HTS. HTS suffered from numerous defectors, dissenters, and internal whistleblowers.

The production of multiple variants of human terrain analysis at various levels of the U.S. military and intelligence was also sometimes accompanied by either implicit or explicit attempts to distance these variants from HTS, creating a sense that HTS was viewed by even those closest to it – in institutional and political terms – as a potential embarrassment. One might also argue that HTS itself emerged as a debate, at least in part, in the form of a critique of now mainstream anthropology. On the other side, diverse criticisms emerged from the American Anthropological Association (AAA), from the independent Network of Concerned Anthropologists (NCA), and from disparate activists, independent journalists, and bloggers in anthropology, both within the United States and abroad. These different, though related, actors tended to bring to the fore qualitatively divergent critiques, using different reasoning, and to different ends. Yet, on some key issues, most of the dominant voices in the debate sometimes tended to agree. The resulting picture is complex enough that it is only with considerable difficulty that one could argue that any one side of this intellectual and political struggle had somehow succeeded in achieving all of its key aims by the end of this period.

Theory and Practice of the HTS

With the war in Iraq beginning to turn against US-led forces by the end of 2003, marked by increased casualties inflicted by Iraqi resistance fighters and the emergence of increasingly sophisticated and lethal 'improvised explosive devices' (IEDs), some in the U.S. military felt that while technologically superior, their forces could not win the war to pacify Iraq without some drastic change in thinking and practice on the ground. The abundance of 'situational awareness' (e.g., where Iraqi tank units were located) was cast as inadequate when what was really needed was 'cultural awareness' to understand the implacable resistance in order to defeat it. This new approach was varyingly called 'culture-centric' or 'population-centric,' and as Scales (2004) explained, "wars are won as much by creating alliances, leveraging nonmilitary advantages, reading intentions, building trust, converting opinions, and managing perceptions – all tasks that demand an exceptional ability to understand people, their culture, and their motivation.

These imperatives place an increased premium on the ability of the U.S. military to understand the nature and character of war, as well as the cultural proclivities of the enemy" (see also McFarland, 2005). Declaring current US occupations to be part of 'World War IV,' Scales also asserted that 'culture awareness' and the ability to "build ties of trust [that] will offer protection to our troops more effectively than body armor." He concluded: "To win World War IV, the military must be culturally knowledgeable enough to thrive in an alien environment. Victory will be defined more in terms of capturing the psycho-cultural rather than the geographical high ground. Understanding and empathy will be important weapons of war" (Scales, 2006). General David Petraeus, once widely proclaimed to be the intellectual grandmaster of contemporary counterinsurgency theory, standing out for his academic qualifications, similarly argued that the decisive terrain of warfare would the 'cultural terrain': "cultural awareness is a force multiplier, reflects [*sic*] our recognition that knowledge of the cultural 'terrain' can be as important as, and sometimes even more important than, knowledge of the geographic terrain.

This observation acknowledges that the people are, in many respects, the decisive terrain, and that we must study that terrain in the same way that we have always studied the geographic terrain" (Petraeus, 2006: p. 51). In advocating what some misinterpreted as a 'gentler' approach to war, "emphasizing cultural knowledge and ethnographic intelligence as major components of its counterinsurgency doctrine," what was formulated was an allegedly new 'cultural turn,' one that

required the expertise of anthropologists in particular in order to understand 'adversary societies' (Jager, 2007: p. v). An early attempt to define the 'human terrain' concept as "the human population and society in the operational environment as defined and characterized by socio-cultural, anthropologic, and ethnographic data and other non-geophysical information about the human population and society" was, however, both overly broad and internally repetitive (Kipp et al., 2006: p. 14, fn 2). These, nonetheless, were some of the main founding lines of thinking on which HTS was instituted.

Putting this thinking into practice, some of the first advocates and directors of HTS came together specifically to counter the impact of IEDs in Iraq, in what was a founding platform for HTS as a practice, identifying the bomb makers as the intelligence priority rather than focusing on the bomb technology alone (Hedges, 2004). Montgomery McFate, the first anthropologist to serve as the social science director at HTS, linked knowledge of the 'human terrain' with work to counter-IEDs (McFate, 2005b). With the thinking behind HTS put into place, and the right network of military and political elites backing such a project, HTS came into being in the same year as the release of the *Counterinsurgency Field Manual* that was directed by Petraeus (U.S. Army, 2006).

After justifying the HTS concept to the military hierarchy, employees of the U.S. military set about trying to convince anthropologists to support the effort. First, the lead anthropologist advocating HTS, Montgomery McFate, began by asserting that anthropologists could help to remedy the lack of knowledge about 'adversary culture' on the part of soldiers, diplomats, and intelligence agents and thus help prevent any abuses of the people occupied by U.S. forces. This lack of cultural knowledge was the product of "the almost total absence of anthropology within the national-security establishment" (McFate, 2005a: p. 24; 2005c?). (However, according to the Center for Naval Analyses, the U.S. armed forces employed a total of 532 persons with degrees in anthropology (Clinton et al., 2010).) Two of the more basic rationales that were used simultaneously in order to appeal to anthropologists were (1) anthropological knowledge could save lives and (2) by contributing to the military and intelligence apparatuses, anthropologists could exercise influence and help change policy while making their discipline more relevant. However, according to McFate's colleagues in the military, simply providing 'cultural awareness' would not be sufficient: what needed was 'ethnographic intelligence' to penetrate opposing networks on a global scale (see Renzi, 2006: p. 17).

HTS directors and publicists initially claimed that saving lives was the intended outcome and benefit of the program. HTS would reduce the need for 'kinetic' (violent) operations that arose from cultural misunderstandings, while simultaneously providing aid to local communities in an effort to win 'hearts and minds' and thus again reducing the need for violence. The latter were some of the dominant working assumptions utilized by advocates for the program. HTS officials repeatedly emphasized that HTS was not an intelligence program and that HTS personnel did not take part in targeting decisions. HTS deployed human terrain teams consisting of persons who were to have expertise in the Middle East and Afghanistan as well as in ethnographic research, in order to map tribal networks and social structures. HTS officials claimed that there was no substantive conflict between its methods and the AAA's code of ethics: HTTs would seek informed consent from Iraqis and Afghans; those interviewed would not be coerced; and, according to the defenders of the program, identities would be kept confidential.

By 2010, HTS grew in size to include 550 employees (Clinton et al., 2010: p. 76). HTS fielded 10 HTTs in Iraq and 17 more in Afghanistan (Clinton et al., 2010: p. 19). According to recent revelations, HTS received at least $726 million in government funding from 2007 through 2013 (Vanden Brook, 2014). On the battlefield, HTT members could initially earn up to $300 000 annually. In late 2008, however, the program ceased to be run by private military contractors, forcing salaries down to the low $200 000 range. The salaries paid proved to be a significant attraction, especially during a severe recession, as attested to by officials in charge of recruitment (Clinton et al., 2010: p. 89). McFate herself had previously noted that the "majority of newly minted anthropologists brutally compete for a limited number of underpaid university faculty appointments" (2005a: p. 28), thus holding out an alternative. By the middle of 2009, however, only six HTS employees possessed a PhD in anthropology, and four held MAs in anthropology, while most of the HTS staff had degrees in fields unrelated to the stated objectives of the program (Commission on the Engagement of Anthropology with U.S. Security and Intelligence Communities CEAUSSIC, 2009: pp. 13, 63).

The phrase, "controversy plagued the program," was probably one of the media's most frequently repeated statements about HTS (Forte, 2011: p. 150). The program had sprung hundreds of leaks, adding to considerable criticism to the program and attracting more, including from members of Congress that had previously celebrated the program. HTS suffered three fatalities; one HTT member was tried and convicted for the murder of a detainee; and another was arrested on espionage charges. Internal U.S. Army investigations have substantiated reports of waste and fraud in addition to racism and sexual harassment. HTS came under review by the House Armed Services Committee (HASC). During 2013, a rash of U.S. Army documents obtained by *USA Today* were released to the public, along with reports that demonstrated the extreme degree of corruption, mismanagement, and abuses within the program (see Vanden Brook, 2013; Forte, 2013).

Possibly one of the most essential working assumptions of HTS, bordering on a myth, was that US occupation forces, when possessing cultural knowledge, would avoid provoking the resentment of the dominated population and thus increase the chance that those dominated would instead turn against resistance forces and support the US occupation and whatever local government it helped to install. According to Gusterson, this assumption reduced the fight against occupation to a mere cultural misunderstanding (quoted in Udris et al., 2010). Logically, it would also imply that the US invasions and occupations of Afghanistan and Iraq were also the product of cultural misunderstanding on the part of the United States. HTS advocates also claimed that the program would save lives, make war more humane, serve to defeat 'insurgents,' and help to 'anthropologize' the military. On these counts and more, such advocates were not able to stand up to the many, varied criticisms, most of which passed without any official reply from HTS directors who in many

cases refused to answer even basic questions about the program (see Price, 2009a).

Contexts, Conjunctures, and Recursion

While it is logical that both HTS and its critiques first arose within a given context that played a fundamental role in motivating, shaping, and informing their articulation, that context itself consists of multiple historical foundations and more recent conjunctures. First, the immediate context that took shape as a result of the attacks of 11 September 2001 allowed the US political directorate to convert the meaning of the attacks into an opportunity for expanding and reinforcing the national security state and launching ambitious military adventures abroad, both of which (Afghanistan and Iraq) initially drew popular support at home, until the reality of determined resistance set in. This ushered in an "applied reactionary agenda to legitimize and facilitate the active involvement of anthropologists with the US military in the 'war on terrorism' in Iraq, Afghanistan, and beyond" (Sluka, 2010: p. 100). Second, in accordance with the dominant neoliberal program that rose to power in the early 1980s, national security would no longer be the exclusive domain of the state, but would also be outsourced to a growing range of private contractors for whom war was literally business. Third, in order to sustain popular support, and given the history of domestic social strife that arose as a result of the draft for the war against Vietnam, the U.S. military switched to an all-volunteer force, offering numerous incentives and benefits that would not otherwise be available to the 'average' citizen. Fourth, with what are remembered as 'blunders' in Vietnam, Lebanon, and Somalia, the current militarized mass-mediated culture would focus attention on: technological advancements in ordnance and logistics ('smart' bombs and 'precision' airstrikes); better 'knowing the enemy' (during the 1983 invasion of Grenada, US troops lacked local maps, and did not know in advance that English was the local language); and exit strategies and perhaps timelines. In other words, the ethos would be that of winnable wars, and winnable because they were founded on smart thinking. However, other, longer-term foundations had to first be in place for any of the above to take hold so quickly and so firmly.

In terms of the long-term continuities that continue to produce militarization and securitization agendas that focus on foreign targets, this would involve a comprehensive understanding of US history, and the world-systemic history from which it emerged, beginning with the foundation of colonies in the sixteenth century, the conquest of Indigenous populations, the development of nationalist agendas, independence from Britain, the usurpation of territory both Indigenous and Mexican, attempts to annex parts of the Caribbean, and then a wider search for foreign markets – all with corresponding ideological or even theological rationalizations. Ranging from ideas of divine providence and manifest destiny, the idea of an 'American exceptionalism' arose, one that would persist in defying both questions and objections. To veer off the track of foreign intervention would mean being 'isolationist' and it implies a sense of irresponsibility, a formulation that makes sense only in an imperial political culture. Even in the midst of the most brutal wars fought abroad, missionary zeal could readily furnish the justificatory code: it was the responsibility of Americans to teach others how to improve themselves, and the duty of others to obey.

A canonical faith in America as a force for global good became fixed in place and continues to exert some force. Ideas of 'liberating' others from themselves, that is, from their own bad habits of savagery, and from their domestic orders of oppression, became authorized truths of US expansionism and interventionism. In light of both the conjunctures and continuities that are merely sketched here, the content of debates surrounding HTS appear to be reformulating and reinscribing prior histories, practices, and hegemonic ideas of care, protection, progress, the legitimacy of the violence of the mighty, the enlightened and even blessed nature of the dominant, the right to rule and the duty to serve. Had these contexts not been in place, it is difficult to see how the propositions and debates could have arisen, using the intellectual resources that were deployed, and achieving a sense of purposes and outcomes that came with very high stakes.

Institutionalized anthropology in the United States, never wholly insulated from the contexts above, also furnished its own critical parameters and resources that underpinned several critiques, emerging from prior confrontations with the militarization of anthropology in the performance of counterinsurgency. Indeed, the very articulation of the code of ethics officially adopted by the AAA in 1971, which would become central to a large number of the critiques of HTS, came about as a result of the first time that anthropology was used for counterinsurgency, during the Vietnam War years (Fluehr-Lobban, 2003). This first AAA code of ethics strongly condemned secret research and affirmed that the paramount responsibility of anthropologists was to those they study, not their own governments and not their sponsors. Anthropologists were called upon to do "everything in their power to *protect* the physical, social, and psychological welfare" of those they studied (AAA, 1971; emphasis added). That code of ethics "emerged as a direct result of the protest against counterinsurgency research in Vietnam, Project Camelot in Latin America, Project Agile in Thailand, and the Himalayan Border Countries Project in India" (Sluka, 2010: p. 102). In the case of Project Camelot, anthropologists reported that 'for years afterward' the 'aftershocks' of the Project continued to harm their research, "as potential academic collaborators and other interlocutors wondered if they were really working for the CIA" (Gusterson, 2009: p. 48). Thus both the protection of others, and professional self-interest, became prominent features of the earliest ethical critiques of collaboration with the U.S. military in counterinsurgency.

Recursion thus played a role in helping to both substantiate and actualize the main lines of debate the second time that counterinsurgency would be raised to beckon anthropological support, when HTS was launched. If active critics within the AAA had the war experiences of the late 1960s and the 1971 code of ethics to resort to, proponents of HTS made it easier to draw such historical linkages. In the Vietnam War, the Civil Operations and Revolutionary Development Support (CORDS) program involved both propaganda and extreme violence as part of the intended US 'pacification.' Early advocates of HTS hailed it as 'a CORDS for the twenty-first century'

(see Kipp et al., 2006). The authors not only refer to the program in 19 parts of their article, they have significantly revised the history of CORDS to argue that it was actually successful, or potentially so (Kipp et al., 2006: p. 11), without considering its lethal dimension. CORDS gave rise to the Phoenix program in 1967 which, as one of the leading anthropological critics of HTS noted, "led to the capture, torture, and assassination (euphemistically called 'neutralization' by the Pentagon) of tens of thousands of civilians" in Vietnam (González, 2009: pp. 5, 59). Thus, thanks in large part to the advocates of HTS, anthropologists' concerns about the use of ethnographic data for the purposes of intelligence-gathering to support targeted killings once more became paramount with suspicions about the true functions of HTS, masked by its preferred euphemisms. One more instance of recursion resurfaced with González' investigation into the roots of the 'human terrain' concept, which he found dated back to urban policing of inner cities of the United States in the 1960s, specifically directed at quelling the perceived threat of militants of the Black Panther Party, an experience also tied to the FBI's Counter-Intelligence Program (COINTELPRO) aimed at spying on, infiltrating, and destabilizing a wide range of activist groups (González, 2009: pp. 29–29).

This recursion would become salient given two separate events: first, the revelations in the press that the anthropologist who headed the social science directorate of HTS, Montgomery McFate, had worked as corporate spy in infiltrating and undermining various activist groups, including the antigun rights lobby (Corn et al., 2008); and, second, the use of 'human terrain mapping' in a project coordinated between the CIA and the New York City Police Department, and adopted more recently by the Los Angeles Police Department (Apuzzo and Goldman, 2011a,b,c; DGI, 2013; Ruderman, 2012). These revelations would also spur the critique that counterinsurgency is also conducted domestically, even in the 'affective labor' involved in spreading ideas that construct the new 'military normal' in everyday American life (Lutz, 2009: p. 34). Having acknowledged the degree of recursion in these critiques it would be, however, a significant error to assume that the dividing lines in the debates around HTS were instantly formulated with little development and deliberation, or that there was any spontaneous rush to condemn the program (see Forte, 2010).

The Diversity of Critiques and Critics

While many different camps were involved in criticizing HTS, these tend to stand out more in hindsight as few were ever explicit and self-conscious in denominating themselves or precisely articulating their differences. In some cases, selective actors would themselves communicate multiple messages that were sometimes contradictory in terms of their root assumptions or intended outcomes. Critiques of HTS fit into nine areas:

- Organizational: focusing on labor relations and (mis) management within HTS' own ranks (U.S. Army, 2010).
- Financial: focusing primarily on issues of the cost effectiveness of HTS, and criticizing the high salaries paid and instances of fraud. Also included here was recognition of the hard realities of a shortage of full time, if any, academic employment for those with doctorates in anthropology (McFate, 2005a: p. 28), and the resulting heightened anxiety within anthropological ranks.
- Institutional: whether within organized sections of the U.S. military or the AAA, specific actors had their own institutions' interests to defend against HTS. The AAA concluded that HTS was not a valid practice of anthropology. Elements in the military criticized the duplication of functions at higher cost, represented by HTS, as well as criticism of counterinsurgency doctrine, and the skewed budgets that resulted from what military critics viewed as misplaced priorities.
- Professional: these criticisms emphasized the actual or potential harm to anthropology's reputation posed by HTS, the potential jeopardy to anthropological fieldworkers who might be mistakenly associated as U.S. military or intelligence agents, as well as criticizing the perceived low level of professionalism within HTS. In terms of reputation, there was also some criticism of the mainstream media's attempt at creating celebrity status for Montgomery McFate.
- Military strategic: doubts cast about the efficacy of counterinsurgency, its dubious likelihood of success, and the deficiencies of using civilian nonexperts.
- Methodological: critique of the research methods and theoretical models adopted as HTS standards. In particular, these critiques tended to dispute whether Human Terrain Teams actually did anything that could be called ethnography, as well as finding fault with the structural-functionalist models of society and culture present in both the HTS training handbook (HTS, 2008) and in the Army's counterinsurgency Field Manual (US Army, 2006).
- Scholarly: the quality of both the research output and the qualifications of the researchers were the focus of these critiques.
- Ethical: in general terms this involved the passive failure to protect one's informants by neglecting or downplaying a wide variety of basic research procedures regarding informed consent, confidentiality, and 'do no harm.' Indeed, the concern voiced by some critics was that HTS involved the active contribution to doing real harm, charging that HTS served as a 'refined targeting' instrument (CEAUSSIC, 2009: p. 32), while also stressing its other, covert function as a means of gathering intelligence on 'enemies' (Vanden Brook, 2013).
- Political: the broadest and arguably most profound of the critiques tended to place the HTS phenomenon within wider contexts of the militarization of the social sciences, and foreign intervention and occupation. There also tended to be a critical view of 'humane war' and 'humanitarian intervention,' and of avoiding the cul-de-sac of liberal critiques of the liberalism found within justifications for HTS. Political critiques were mindful of the history of U.S. imperialism and the shifting roles of anthropology in that history. This camp also tended to take issue with both the dominance of discussions of professional ethics in critiques of HTS, as well as the conflict of interest presented by military anthropologists in the AAA serving on commissions that passed judgment on both HTS and military anthropology itself.

Despite the wide range of criticism, the debates tended to render prominent only in selective schools of criticism. Critiques were thus dominated primarily by ethical and professional concerns, with some assumptions shared with HTS proponents themselves (see Zehfuss, 2012). There was thus a concomitant concern with the 'tone' of debate, with calls for collegial respect, viewing the stronger expressions of criticism with some distaste as if mere ideological polemics. In that vein, both sides of the debate – HTS advocates and mainstream academic critics, especially military anthropologists – tended to betray shared affinities, a mutual sense of belonging, and common concerns in terms of 'care' and 'protection' of those under the US occupation. There was, in some cases, a shared concern in finding practical solutions (that took occupation as a given, and thus out of bounds as a target of critique) and providing useful advice to military and political authorities. The mainstream media also routinely privileged these voices over others, even though all criticisms were, for the most part, rendered marginal at best (see Price, 2009a).

As suggested above, there were almost as many distinct actors with interests vested in the critique of HTS, as there were distinctive foci of the critiques. In terms of American anthropology, the AAA the Executive Board and the Commission on the Engagement of Anthropology with the U.S. Security and Intelligence Communities (CEAUSSIC) spoke out in terms of professional ethics, while the independent NCA propelled and echoed many of the same ethical concerns, with a few added critiques of military anthropology, counterinsurgency, and occupation. Within the U.S. military, certain agencies and individuals spoke out both privately and publicly, even if only to advance competing forms of human terrain analysis and thus competing for budgetary support for their alternative programs. Within the HTS itself, there were numerous dissatisfied employees, critical former employees, and outspoken defectors who openly exposed the inner workings of HTS (Allison, 2010a,b; Price, 2010). Some of the politically most critical voices, however, were attached to none of the above groupings, opting instead for independent and individual expression of their critiques. Interestingly, anarchist anthropologists generally did not participate in any of the debates surrounding HTS or militarized anthropology more broadly. Those with stakes in the outcomes of the debates were considerably more numerous and varied, ranging from a large array of private defense contractors, to U.S. military officials, the White House, the HASC, and a seemingly endless number of nationalistic military and security think tanks and commentators in the media.

In terms of professional concerns, dealing with the reputation of the discipline and its fieldworkers, the Executive Board of the AAA disapproved of HTS in October 2007, as an "unacceptable application of anthropological expertise," specifically criticizing the program for "placing anthropologists in positions in which their work will be in violation of the AAA Code of Ethics" and threatening the well-being of "both other anthropologists and persons other anthropologists study" (American Anthropological Association Executive Board AAA-EB, 2007). Following a lengthy investigation of HTS, the AAA's Commission (CEAUSSIC) condemned the program in the following terms: "When ethnographic investigation is determined by military missions, not subject to external review, where data collection occurs in the context of war, integrated into the goals of counterinsurgency, and in a potentially coercive environment – all characteristic factors of the HTS concept and its application – it can no longer be considered a legitimate professional exercise of anthropology" (CEAUSSIC, 2009: p. 4). CEAUSSIC emphasized not only "the incompatibility of HTS with disciplinary ethics," but also "the problem of allowing HTS to define the meaning of 'anthropology' within DoD" (CEAUSSIC, 2009: p. 3). The AAA also revised its code of ethics, and in 2009 it reinstated the provisions against secret research.

The shadow cast over American anthropology by the mass-mediated public persona of HTS' lead anthropologist, Montgomery McFate, also spurred further criticisms of the program from the standpoint of professional ethics and professional reputation. In 2008, *Mother Jones* revealed that, prior to joining HTS, McFate had for years worked as a corporate spy, penetrating the antigun lobby and environmental and animal rights activist groups (Corn et al., 2008). McFate was also found to have plagiarized passages in the Army's new *Counterinsurgency Field Manual* (Price, 2007). Price revealed how McFate's doctoral thesis may have made her Irish Republican Army informants vulnerable to cooption and defeat. The thesis includes statements such as "knowledge of the enemy leads to a refinement in knowledge of how best to kill the enemy" (Price, 2009c). Though some anthropologists were concerned about how McFate propagated espionage and unethical practices through her own work, there were particularly strong critiques of McFate's writings on the history of anthropology (see, for example, Lawless, 2009; Sluka, 2010). According to McFate, the discipline had crashed into 'the abyss of postmodernism' since the end of the Vietnam War, which resulted in the "curious and conspicuous lack of anthropology in the national security arena" when the discipline had in the past had a long and 'fruitful' relationship with colonial authority and administration (McFate, 2005a: p. 24).

From Ethics to Politics

Much of the institutionalized defense of the reputation of anthropology revolved around the question of the ethics of research involving human subjects. Ethics for the most part were considered in relation to specific research procedures and immediate relationships, with an emphasis on protection: do no harm, seek informed consent, protect informants' confidentiality, and respects one's obligations to those studied first and foremost. Criticisms of HTS on ethical grounds came from a wide array of anthropologists. The more politically informed approaches argued that it is very difficult to avoid harming other human beings when one is serving one of the primary belligerents in an armed conflict. Others questioned whether informed consent is truly sought and obtained, especially when HTT members travel with armed soldiers and are often themselves armed. HTS members who have sworn that they obtained consent, privately explained that the consent was in written form and that copies were deposited in digital format with HTS, thus creating a trail that could expose their

informants' identities. The U.S. Army itself did much the same, when in posting numerous photos of HTTs in action in Afghanistan, they clearly showed the faces of individuals serving as informants. As summarized in the words of the NCA, "Human Terrain Mapping … does not provide for meaningful informed consent; it endangers those studied; it generates knowledge that will disappear into Pentagon computers; and it increases the likelihood that other anthropologists will be perceived as spies and military operatives, thereby endangering the research, perhaps even the physical safety, of professional colleagues" (NCA, 2007: p. 4).

Anthropologists also directed criticisms at the scholarly and methodological aspects of human terrain research. Many correctly observed that HTS generally failed to deploy people with prior experience in or knowledge about Iraq and Afghanistan, and that most of its personnel lacked any ability to speak local languages. In some cases, HTTs were forced to rely on interpreters in Afghanistan who themselves lacked knowledge of the local language in a given area of operations. The argument that social science can be carried out while forming part of an armed squad was also doubted, even by one Marine commander who observed that, "when you're in uniform you have all the coercive force of the U.S. government" (CEAUSSIC, 2009: p. 50). Indeed, forming part of a patrol, and working while under the watchful eyes of a military escort, dictated the length of the average interview that could be conducted with villagers without coming into the sights of a sniper: 7 min (Allison, 2010a). What counterinsurgency theorist David Kilcullen called 'armed social science' (Packer, 2006) also involved what in the military was known as 'windshield ethnography' (Edwards, 2013: p. 548) that is, high-speed assessments based on immediate impressions and brief, formal interviews, under adverse conditions. Few anthropologists would class this as ethnography. Also questioned was whether academic inquiry can produce the mission-relevant objective facts that were sought, given the simplified structural-functionalist and instrumentalist notion of 'culture' employed by HTS (Zehfuss, 2012: p. 177).

What were arguably the most incisive, far-reaching, and most fertile of the bodies of criticism were those focusing on the political. If furthered, these could have expanded into wider critiques of 'humanitarianism' and 'humane warfare,' as well as critiques of where institutional anthropology's roles stand with regard to the state and the military. Given the primacy of professional reputation, ethics, and scholarship, it was the political critique that still remains to be fully explored. To a significant extent, the political critique emerged out of a contextualized and historicized understanding of how U.S. anthropology came to adopt the ethical statements that it had. Reflecting on the ethics emerging from the Vietnam War experience, for example, on the need to consider the history and interests of those sponsoring research, could lead to a rejection of all research in support of the U.S. military: "On that basis alone – the appalling record of the US military establishment with respect for example to human rights and civilian casualties – any anthropologist who chooses to work for them now can hardly be said to be taking this precaution seriously" (Sluka, 2010: p. 102). Critical anthropologists were particularly mindful then of the '9/11 context' and the 'war on terror' that threatened to quell dissent in US universities (once again) while marshalling teaching and research toward the demands of the national security state.

In addition, the political critique understood the limits sets by an exclusive focus on research ethics and how the latter could dovetail with HTS' self-representation in one key regard: a shared focus on the protection of others. Thus by setting certain common standards beyond which critique was meant to cease, the ethics focus missed the point, such as the broader mission to which HTS contributed. Expressing this argument, cultural anthropologist Price (2009b) explained: "no matter how anthropological contributions ease and make gentle this conquest and occupation, it will not change the larger neocolonial nature of the larger mission; and most anthropologists are troubled to see their discipline embrace such a politically corrupt cause." Moreover, the issue at focus is war as a tool of U.S. foreign policy, rather than finding ways of supposedly 'improving' war (Lutz in Udris et al., 2010).

By 2011, the political critique became as much a critique of the placement of professional ethics at the center of debates with HTS. These critics understood that, "the convenience of a categorical separation of 'the ethical' from 'the political' has left large issues unaddressed" (Price, 2009b) and that the context of the debate is '100% political' (Jackson, 2009). While more ambiguous in the conclusions he drew from this, George Marcus (2009) also observed that, "without the discussion of the political contexts of anthropological research, the discussion of ethics can be formulaic, moralistic, however virtuous, and frankly naïve investing only in terms of established professional imaginaries of practice." Noting the limits of criticisms that suggested that only the involvement of anthropologists in HTS was problematic (thus 'quarantining' the discipline while implying that members of other disciplines should feel unconstrained to participate in HTS), others observed that,

> the separation between politics and ethics is both problematic and futile. It is difficult to see how 'political' questions regarding 'justifications' could be separate from ethics. Indeed, some of the critique framed in terms of professional ethics seems implicitly reliant on a prior judgment about the permissibility of using military force. That is, a wider (or simply different) understanding of ethics influences how the problematic is set up, and whether a war is seen as legitimate. Yet, if anthropologists grant that they might make a contribution to military operations in different circumstances, then their refusal seems not really based on the claimed professional ethics.
>
> Zehfuss, 2012: p. 186

Impacts and Consequences

In promoting itself as the smarter way of doing war, HTS advocates initially failed to realize that 'winning hearts and minds' at home (let alone in very different societies) would prove to be a mission never accomplished. This is due to several factors, including the vocal opposition of directors and members of the AAA; the formation of the NCA, whose advocacy rallied attention and support for anthropological opposition to participation in counterinsurgency; declining public support for the wars in Iraq and Afghanistan; and rising discontent within the ranks of HTS itself, finding expression in

numerous damaging leaks previously noted. In personal communications, some individuals who were part of HTS have admitted that the combined efforts of the AAA, NCA, and media had an impact in severely limiting the number of anthropologists willing to join HTS (see also Sluka, 2010: pp. 108–109). The rise to prominence of these debates within anthropology has already begun to make a mark in terms of teaching materials in the discipline, which either devote new sections in revised editions to the HTS debate, or incorporate the works of those directing HTS or who sympathized with it (see Robben and Skuka, 2012; Ferraro, 2009; Erickson and Murphy, 2013).

On the other hand, as outlined in the previous section, there was serious difficulty among critics in moving the debate beyond a matter of disciplinary reputation and professional ethics. All too often discussion was narrowed down to the minutiae of research procedures, or slightly expanded to consider whether HTS was involved in espionage and lethal targeting. The attempt by the AAA's own CEAUSSIC to separate critique of HTS from political concerns was in essence an attempt to depoliticize discussion, which is itself a political act. Also political was the AAA's decision to include a number of military anthropologists in CEAUSSIC, in a situation where their own interests were directly at stake.

One important question was never raised in the debates: whether anthropological knowledge can lend itself to harmful uses. A second was why were anthropologists gathering and publishing such potentially harmful knowledge to begin with? With what appears to be a rapid recession in funding for HTS since 2013, a downsizing in its numbers, and tougher questions from the mass media and former congressional supporters, it might be that such questions will remain unaddressed until the next time something like HTS emerges.

See also: Conflict and War: Anthropological Aspects; Geographic Information Systems; Internet and Privacy; Military Geography; Military Sociology; Surveillance Studies; War, Sociology of.

Bibliography

Allison, J., 5 December 2010a. The Leavenworth Diary: Double Agent Anthropologist Inside the Human Terrain System. Zero Anthropology. Available at: http://zeroanthropology.net/all-posts/the-leavenworth-diary-double-agent-anthropologist-inside-the-human-terrain-system/.

Allison, J., 2010b. Whose side are you on boys? The problem for embedded anthropologists in the US human terrain system teams. Forum Archaeologiae – Zeitschrift für klassische Archäologie 55 (6). Available at: http://homepage.univie.ac.at/elisabeth.trinkl/forum/forum0610/55price.htm.

American Anthropological Association (AAA), 1971. Principles of Professional Responsibility. American Anthropological Association, Washington, DC. Available at: http://www.aaanet.org/cmtes/ethics/AAA-Statements-on-Ethics.cfm.

American Anthropological Association, Executive Board (AAA-EB), 31 October 2007. American Anthropological Association Executive Board Statement on the Human Terrain System Project. American Anthropological Association. Available at: http://www.aaanet.org/issues/policy-advocacy/statement-on-HTS.cfm.

Apuzzo, M., Goldman, A., 31 August 2011a. Inside the Demographics Unit, the CIA Spy Team the NYPD Says Doesn't Exist. Associated Press. Available at: http://www.huffingtonpost.com/2011/08/31/inside-the-demographics-u_n_943139.html.

Apuzzo, M., Goldman, A., 22 September 2011b. NYPD Eyed US Citizens in Intel Effort. Associated Press. Available at: http://www.huffingtonpost.com/2011/09/22/nypd-eyed-us-citizens-in-_n_975731.html.

Apuzzo, M., Goldman, A., 17 October 2011c. CIA Role within NYPD Raises Eyebrows. Associated Press. Available at: http://www.nbcnewyork.com/news/local/CIA-Spy-Muslim-NYPD-Ray-Kellty-New-York-City-Police-Investigation-131970938.html.

Clinton, Y., Foran-Cain, V., McQuaid, J., Norman, C.E., Sims, W.H., 2010. Congressionally Directed Assessment of the Human Terrain System. Center for Naval Analyses, Alexandria, VA.

Commission on the Engagement of Anthropology with US Security and Intelligence Communities (CEAUSSIC), 2009. Final Report on the Army's Human Terrain System Proof of Concept Program. Available at: http://www.aaanet.org/cmtes/commissions/ceaussic/upload/ceaussic_hts_final_report.pdf.

Corn, D., Ridgeway, J., Schulman, D., 29 July 2008. There's Something about Mary: Unmasking a Gun Lobby Mole. Mother Jones. Available at: http://www.motherjones.com/politics/2008/07/theres-something-about-mary-unmasking-gun-lobby-mole.

DGI, 2013. LAPD's Implementation of Geospatial Intelligence in Human Terrain Mapping. DGI. Available at: http://www.wbresearch.com/dgieurope/human-terrain-mapping.aspx.

Edwards, D.B., 2013. Counterinsurgency as a cultural system. In: Erickson, P.A., Murphy, L.D. (Eds.), Readings for a History of Anthropological Theory, fourth ed. University of Toronto Press, Toronto, ON, pp. 545–560.

Erickson, P.A., Murphy, L.D. (Eds.), 2013. Readings for a History of Anthropological Theory, fourth ed. University of Toronto Press, Toronto, ON.

Ferraro, G. (Ed.), 2009. Classic Readings in Cultural Anthropology, second ed. Wadsworth Cengage, Stamford, CT.

Fluehr-Lobban, C. (Ed.), 2003. Ethics and the Profession of Anthropology: A Review of Issues and Principles, second ed. University of Pennsylvania Press, Philadelphia, PA.

Forte, M.C., 28 May 2010. Time Line and FAQ for the Human Terrain System and Responses by the Network of Concerned Anthropologists and the American Anthropological Association. Zero Anthropology. Available at: http://zeroanthropology.net/2010/05/28/time-line-and-faq-for-the-human-terrain-system-and-responses-by-the-network-of-concerned-anthropologists-and-the-american-anthropological-association/.

Forte, M.C., 2011. The human terrain system and anthropology: a review of ongoing public debates. American Anthropologist 113 (1), 149–153.

Forte, M.C., 19 February 2013. Documents: Investigations into the U.S. Army's Human Terrain System. Zero Anthropology. Available at: http://zeroanthropology.net/2013/02/19/documents-investigations-into-the-u-s-armys-human-terrain-system/.

González, R.J., 2009. American Counterinsurgency: Human Science and the Human Terrain. Prickly Paradigm Press, Chicago, IL.

Gusterson, H., 2009. Militarizing knowledge. In: Network of Concerned Anthropologists, Eds. The Counter-Counterinsurgency Manual: or, Notes on Demilitarizing American Society. Prickly Paradigm Press, Chicago, IL, pp. 39–55.

Hedges, S.J., 23 October 2004. U.S. Battles Low-Tech Threat: Military Shifts Tactics to Deal with Insurgents' Lethal, Hidden Bombs. Chicago Tribune. Available at: http://articles.chicagotribune.com/2004-10-23/news/0410230230_1_humvees-armored-roadside-bombs.

Human Terrain System (HTS), 2008. Human Terrain Team Handbook. Human Terrain System, Fort Leavenworth, KS.

Jackson, J., 30 November 2009. Reflections on the Soon-to-be Released CEAUSSIC Report. American Anthropological Association. Available at: http://blog.aaanet.org/2009/11/30/reflections-on-the-soon-to-be-released-ceaussic-report/.

Jager, S.M., November 2007. On the Uses of Cultural Knowledge. Strategic Studies Institute. Available at: http://www.strategicstudiesinstitute.army.mil/pdffiles/pub817.pdf.

Kipp, J., Grau, L., Prinslow, K., Smith, D., September–October 2006. The Human Terrain System: A CORDS for the 21st Century. Military Review. Available at: http://usacac.army.mil/CAC2/MilitaryReview/Archives/English/MilitaryReview_20061031_art005.pdf, 8–15.

Lawless, R., 6–8 November 2009. Empires and the Sullying of Anthropology: From Malinowski to Human Terrain Systems. CounterPunch. Available at: http://www.counterpunch.org/2009/11/06/empires-and-the-sullying-of-anthropology/.

Lutz, C., 2009. The military normal. In: Network of Concerned Anthropologists, Eds. The Counter-Counterinsurgency Manual: or, Notes on Demilitarizing American Society. Prickly Paradigm Press, Chicago, IL, pp. 23–37.

Marcus, G., 7 December 2009. Origin Story and Grande Finale. American Anthropological Association. Available at: http://blog.aaanet.org/2009/12/07/ceaussic-origin-story-and-grand-finale/.

McFarland, M., March–April 2005. Military cultural education. Military Review. Available at: http://www.au.af.mil/au/awc/awcgate/milreview/mcfarland.pdf.

McFate, M., March–April 2005a. Anthropology and counterinsurgency: the strange story of their curious relationship. Military Review 24–38. Available at: http://www.au.af.mil/au/awc/awcgate/milreview/mcfate.pdf.

McFate, M., May–June 2005b. Iraq: the social context of IEDs. Military Review 37–40. Available at: http://usacac.army.mil/cac2/AIWFC/COIN/repository/Social_Context_of_IEDs_Mil_Review-McFate(May-Jun05).pdf.

McFate, M., July 2005c. The military utility of understanding adversary culture. Joint Forces Quarterly 38, 42–48. Available at: http://www.dtic.mil/doctrine/jfq/jfq-38.pdf.

Network of Concerned Anthropologists (NCA), 2007. The network of concerned anthropologists pledges to Boycott counterinsurgency. Anthropology News 48 (9), 4–5.

Packer, G., 18 December 2006. Knowing the Enemy: Can Social Scientists Redefine the "War on Terror"? The New Yorker. Available at: http://www.newyorker.com/archive/2006/12/18/061218fa_fact2?currentPage=all.

Petraeus, D.H., January–February 2006. Learning counterinsurgency: observations from soldiering in Iraq. Military Review. Available at: http://usacac.army.mil/CAC2/MilitaryReview/Archives/English/MilitaryReview_2006CR1031_art010.pdf.

Price, D.H., 30 October 2007. Pilfered Scholarship Devastates General Petraeus's Counterinsurgency Manual. CounterPunch. Available at: http://web.archive.org/web/20080101185509/http://www.counterpunch.org/price10302007.html.

Price, D.H., 7 April 2009a. Counterinsurgency's Free Ride: The Press and Human Terrain Systems. CounterPunch. Available at: http://www.counterpunch.org/2009/04/07/counterinsurgency-s-free-ride/.

Price, D.H., 9 September 2009b. Anthropological Engagements with Military and Intelligence Agencies: Ethics, Politics, and Ongoing Discourse. American Anthropological Association. Available at: http://blog.aaanet.org/2009/09/11/ceaussic-anthropological-engagements-with-military-and-intelligence-agencies/.

Price, D.H., 1 December 2009c. A BetterWay to Kill? Human Terrain Systems, Anthropologists and the War in Afghanistan. CounterPunch. Available at: http://www.counterpunch.org/2009/12/01/human-terrain-systems-anthropologists-and-the-war-in-afghanistan/.

Price, D.H., 15 February 2010. Human Terrain Systems Dissenter Resigns, Tells Inside Story of Training's Heart of Darkness. CounterPunch. Available at: http://www.counterpunch.org/2009/04/07/counterinsurgency-s-free-ride/.

Renzi, F., September–October 2006. Networks: terra incognita and the case for ethnographic intelligence. Military Review 16–22. Available at: http://www.au.af.mil/au/awc/awcgate/milreview/renzi.pdf.

Robben, A.C.G.M., Skuka, J.A. (Eds.), 2012. Ethnographic Fieldwork: An Anthropological Reader. Wiley-Blackwell, Malden, MA.

Ruderman, W., 21 August 2012. Secretive Police Unit Sought to Map Terrorist Havens in City, Official Testifies. The New York Times. Available at: http://www.nytimes.com/2012/08/22/nyregion/nypd-unit-sought-havens-for-terrorists.html?_r=0.

Scales Jr., R.H., 2004. Culture-centric warfare. U.S. Naval Institute Proceedings 130 (10), 32–36. Available through Academic Search Complete, EBSCOhost.

Scales Jr., R.H., July 2006. Clausewitz and world war IV. Armed Forces Journal. Available at: http://web.archive.org/web/20061226205524/http://www.armedforcesjournal.com/2006/07/1866019.

Sluka, J.A., 2010. Curiouser and curiouser: Montgomery McFate's strange interpretation of the relationship between anthropology and counterinsurgency. PoLAR: Political and Legal Anthropology Review 33 (S1), 99–115.

Udris, D., Der Derian, J., Udris, M., 2010. Human Terrain (Documentary). Udris Film, Providence, RI. Oxyopia Productions & Global Media Project.

U.S. Army, 2006. FM 3-24, Counterinsurgency. Headquarters, Department of the Army, Washington, DC.

U.S. Army, 12 May 2010. Findings and Recommendations, AR 15–6 Investigation Concerning Human Terrain System (HTS) Project Inspector General Complaints. United States Army Training and Doctrine Command, Fort Monroe, VA. Available at: https://app.box.com/s/2mv0g54xsr41aegwbw9i.

Vanden Brook, T., 28 February 2013. Army Plows Ahead with Troubled War-Zone Program. USA Today. Available at: http://www.usatoday.com/story/news/world/2013/02/17/human-terrain-system-iraq-afghanistan/1923789/.

Vanden Brook, T., 2 January 2014. Human Terrain System Tab up to $726M since '07. USA Today. Available at: http://www.militarytimes.com/article/20140102/NEWS05/301020029/Human-Terrain-System-tab-up-726M-since-07.

Zehfuss, M., 2012. Culturally sensitive war? The human terrain system and the seduction of ethics. Security Dialogue 43 (2), 175–190.

Relevant Websites

American Anthropological Association (AAA). Ethics Resources. Available at: http://www.aaanet.org/cmtes/ethics/Ethics-Resources.cfm (accessed 12.09.13.).

Forte, M.C., 2010. Bibliography and Archive: The Military, Intelligence Agencies, and the Academy (with Special Reference to Anthropology). Zero Anthropology. Available at: http://zeroanthropology.net/all-posts/bibliography-and-archive-the-military-intelligence-agencies-and-the-academy-with-special-reference-to-anthropology-documents-news-reports/ (accessed 02.01.14.).

Human Terrain System. Available at: http://humanterrainsystem.army.mil/ (accessed 07.01.14.).

Network of Concerned Anthropologists. Available at: https://sites.google.com/site/concernedanthropologists/ (accessed 07.01.14.).

Human–Computer Interfaces

J May, University of Sheffield, Sheffield, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 7031–7035, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Human–computer interaction (HCI) is the study of how people use technological artifacts, and their design. Unified cognitive architectures such as GOMS and Soar, derived from artificial intelligence, have proven useful theoretically, but too detailed for general application in design. The recognition that design flaws had to be identified early in the design process has lead to the idea of user-centered design, with techniques such as QOC supporting the design process, and the development of evaluation methods, such as cognitive walkthroughs and codifications of practical advice in the form of guidelines. To help designers understand why usability problems might have arisen, cognitive psychology theories are being couched as supportive evaluation methods. One such approach, ICS, is an example of a trend away from the dyadic, turn-taking relationship between a single user and a single computer, towards a more complex view of computers as part of our everyday environment. This trend is increasing the involvement of social and organizational theorists. From another perspective, computer scientists see systems as built up from 'interactors' and argue that the user's cognitive processes can be modeled as just another set of interactors. Syndetic models examine the overall system of human and computer interactors. In general, HCI is focusing less on the individual, has a great deal to gain from other social sciences, and has the potential to provide insight into human behavior in return.

The increasing contribution of other disciplines is making human–computer interaction focus less on the individual, with the consequence that a great deal can be gained from other fields within the social sciences, and perhaps, provide more insight into human behavior in return.

Human–computer interaction, more widely known by its acronym HCI, is primarily the study of how people use the complex technological artifacts that now pervade their lives. A secondary, but arguably more important focus, is on how these artifacts can be designed to facilitate this use. As an applied discipline, HCI encompasses theoretical and academic researchers from psychological and computer science backgrounds, who may be interested in the area as a domain within which to apply and evaluate their theories of human or system behavior, as well as designers and engineers from industrial and commercial backgrounds, for whom the pay-off is a saleable product, built faster and more economically than their competitors'. In recent years, with the rapid technological development of computing devices and their increasing ubiquity within telecommunications, the field has expanded to include social and organizational researchers, who examine the impact that the technologies have upon human–human interaction. This itself feeds back to the identification of novel applications and the development of further devices. HCI is now an active, multifaceted, and methodologically diverse component of the domain that it studies, both following and directing technical innovation.

Origins in Human Factors and Cognitive Psychology

The roots of HCI predate the invention of computers, and the term itself has only been used since the 1970s. At their simplest, computers are just machines, built to help people carry out tasks, and the dominant paradigm for studying HCI has been that of ergonomics or human factors, with study of the *interaction* between the human and computer subjugated to that of the human's *use* of the computer. An earlier name for the field was man–machine interaction, and within the field, the words person and human are still generally replaced by the term 'user.' In the same way that early time and motion researchers would analyze the tasks that a factory or office worker performed, breaking their job down into its constituent elements, measuring the time each part took and assessing their relationships to each other to identify any 'wasted,' redundant, or unnecessary actions, so too have HCI researchers attempted to decompose the user's task, and to identify task steps that are difficult to perform, hard to learn, or confusing.

This emphasis on task analytic decomposition reflects the early predominance within HCI of cognitive psychologists, who brought with them information processing models of human thought. These models were themselves based on a metaphor of the mind as a computer: one that received sensory input, processed it with access to memory and inferential resources, and produced motor output to act upon the world. By far the most influential of these within HCI has been the GOMS model (Card et al., 1983), which stands for Goals, Operators, Methods, and Selection. Based upon an intentionally simplistic conception of a model human processor (MHP), a GOMS analysis would identify the goals that the user has to reach, and the operators or mental actions available to the user to carry out the task of reaching the goal. Frequently used series of operators become chunked as easily accessible 'methods' that are linked to a subgoal or specific part of the task, and selection routines enable the user to 'decide' which alternative methods or operators to apply if there is more than one possible route to the goal. The MHP provides estimates of the length of time that executing each basic sensory, cognitive, and motor operator takes (these values being empirically derived from cognitive psychology experiments), and so once an accurate GOMS analysis of the task has been produced, it is possible to sum the durations of each component to produce a prediction of the time taken for the whole task to be

performed. Subsequent elaboration of the approach to allow the perceptual, cognitive, and motor parts of the architecture to operate in parallel, and to model the processing of concurrent activities using critical path analysis, has had notable success in modeling the efficiency of telephone toll and assistance workstations (Gray et al., 1993).

Despite successes such as this, a limitation on the application of GOMS in practice is that it is intended to model only the optimal performance of a skilled user in ideal circumstances, and so it is best suited to industrial-style settings where a particular task may be executed hundreds or thousands of times a day. An interface design that is a second or two quicker to use in these situations will thus result in a large saving in time, and presumably money, for the organization that is paying for it. GOMS is unable to say anything about another crucial aspect of interface use, that is, how long it might take someone to become an expert, how many errors a nonexpert might make, or how pleasant or unpleasant the experience of executing the operations might be. Indeed, the critical assumption within GOMS that operators can be chunked as methods is not actually addressed by the model itself. It remains the job of the analyst to recognize which sequences of operators can be represented as methods, and how finely these methods need to be related hierarchically. Recognizing that this development of knowledge about the task is the key to learning about an interface, Newell (1990) proposed what he termed a unified cognitive architecture specifically to allow problem solving and learning to be modeled as a process of goal decomposition, such that goals are iteratively broken down into subgoals. By specifying the operation of this architecture within a 'programming language' called Soar, Newell provided a technique for predicting the way that users would integrate their postulated initial knowledge with information derived from the computer interface through their interaction with it. In effect, a Soar analyst writes a program in the form of 'If,…Then…' rules, and this is then 'run' to simulate a user's learning behavior.

The GOMS and Soar research communities have had a great deal of influence within the cognitive psychology side of HCI, but as simulation models, their attention to detailed aspects of the interface is both their strength and weakness. While it allows them to identify small but critical aspects of a design as crucial problems for a user, they both require complete specifications of an interface, and for Soar, the task context and information available to a user. They therefore come into their own for the contrastive evaluation of extant alternatives, but are hard to apply systematically at earlier stages of design, when the question is 'how?' rather than 'which?' The same is arguably true of newer, more detailed techniques such as LICAI, which is based on Kintch's construction–integration model of goal formation in text comprehension (Kitajima and Polson, 1997), and EPIC (Kieras and Meyer, 1997). EPIC builds on the model human processor by incorporating many theoretical concepts such as the distinction between declarative and procedural knowledge drawn from Anderson's ACT class of models, and limited capacity visual and auditory working memories coordinated by executive processes. However, as Kieras and Meyer themselves note, 'EPIC was not developed primarily for addressing HCI problems but is a larger scientific endeavor to represent important theoretical concepts of human intelligence or abilities' (Kieras and Meyer, 1997: p. 394).

Development of Usability Evaluation Methods

The apparent concentration of cognitive psychologists on the development of cognitive architectures that would ideally explain all of human behavior, as well as the human use of computers, can justifiably be seen as a problem from the point of view of their colleagues who are involved in the day-to-day design of computer interfaces. It was accepted fairly quickly that full GOMS models were too laborious to be practical, and a simpler (but inevitably less reliable) 'keystroke level model' was developed that avoided the need for a full task analysis. Soar models also require a programming and analysis commitment that can seem disproportionate in the context of design, especially when an apparently small modification in the design requires the whole model to be reprogrammed and rerun. It can also be difficult to identify reasons for the simulated behavior that results from an analysis. These cognitive architectures, then, may prove to be more useful in informing their core disciplines than in practical application.

In reaction to these perceptions of too much cost for too little benefit, a number of 'cut down' usability evaluation methods have been developed. All are to some extent derived from well-founded and principled psychological theories, but by encapsulating key components of the modeling techniques within a process of some form, seek to minimize the overhead costs of learning the theory itself, and to provide a positive cost–benefit trade-off by reducing the need for user trials to be conducted with early prototypes or mock-ups of interfaces (Nielsen and Mack, 1994). Cognitive walkthroughs, for example (Polson et al., 1992), provide a checklist by which aspects of the interface can be compared with the user's notional task, step by step, and screen by screen. The steps of the checklist are in fact derived from cognitive complexity theory, and the method focuses on the semantic appropriateness of the verbal labels or commands used throughout the interface, although the practitioners applying it do not need to know this.

Other approaches seek to avoid any role for theory at all, and base themselves in the codification and reuse of practical experience. From the earliest days of HCI, sets of guidelines have been compiled, both within design teams to provide a 'house style' and in a wider sense as a shared set of 'best practices,' so that designers encountering a problem for the first time have access to the solutions found or attempted by others. Some, such as those presented by Gardiner and Christie (1987), have an empirical basis, and specify specific design features such as acceptable screen color combinations and choices for typefaces. Others are less formal, and direct the designer towards desirable attributes or practices such as aiming for consistency, compatibility and coherence. The rapidly evolving nature of the domain conspires against such stored knowledge, however, for anything designed before the mid-1990s is almost certain to be based on expired technology, whose constraints no longer apply. Concrete prescriptions about design solutions become outdated too quickly for their

promulgation, while the more abstract (such as 'be consistent') have proven too abstract to be of much real use.

HCI and the Design Process

This focus upon evaluation reveals a fundamental problem in early notions of the relationship between design and modeling. An implicit assumption of many HCI approaches was that design could be thought of as a sequential or 'waterfall' process, with discrete stages progressing from a requirements analysis, through specification layers, to implementation, with a final evaluation of the finished product. By locating evaluation at this final stage, HCI staff found themselves limited to what became known as a 'fire-fighting' role, being asked to repair problematic designs at a point when any redesign would result in a product falling behind schedule, further worsening the cost implications of their involvement. Organizationally, it is very difficult to convince people that the design they have been committed to for months has some basic usability flaws.

As Landauer (1995) has pointed out in his econometric analysis of the computer industry, design flaws have to be identified much earlier in design if their identification is to be helpful and cost effective to repair. Fortunately, few designs actually proceed in such a strictly formal manner, with a more serendipitous, parallel, and iterative process being more typical. Such cyclical processes allow HCI contributions to be made at more appropriate junctures, and have lead to an emphasis upon user-centered design (e.g., Norman and Draper, 1986), in which user testing becomes a central aspect of design throughout the process, design itself is directed towards prototyping, and designers are encouraged to 'plan to throw one away,' following Brooks' seminal advice (Brooks, 1975).

A very different approach to providing support to designers in these less structured situations comes in the form of a notational decision support tool called QOC, or questions, options, and criteria (MacLean et al., 1991). Aimed at recording the 'design rationale,' QOC represents the design space through which a product has traveled, listing the questions that the design team were trying to answer at various points, the options that were considered, and the criteria that they used to choose one option over another. By making decisions and trade-offs explicit, and recording the reasons that were cited for making them, QOC provides a mechanism both for managing the decision process and for recording it for later review. It has been found to be particularly helpful in providing background information for people who were not part of the decision at the time, either because they only become involved at a later stage or because they were working on some other part of a large project. QOC is important in that it represents a wider concern with design as a process, rather than simply with evaluating its final products.

Involvement in design at a point before the final product has been created means, inevitably, that HCI methods need to provide advice to designers rather than evaluating finished designs. While 'cut-down' usability inspection methods and heuristic evaluations can be conducted on partial designs and mock-ups, it has been argued that their pragmatic separation from theory weakens their value. While they may well identify problems in a design, the fact that they are not associated with an explanatory theoretical underpinning means that the causes of the problems can remain unexplained. The benefit to the designer of a rapid identification of problems is accompanied by the cost of working out why the problem has occurred and what is to be done about it. May and Barnard (1995) recommended that HCI researchers should work towards providing 'supportive evaluation methods' that assist the designers by accounting for usability problems in terms of some theoretical concepts that are helpful in suggesting redesigns.

They referred to a distinction made by Long (1987) between the 'discovery representation' of a scientific theory and its 'application representations.' The discovery representation allows academics and researchers to explore phenomena, such as typical error patterns in HCI, or the way that users learn about interfaces and solve problems. The application representations, on the other hand, are encapsulations of the theory specifically targeted towards practical application. These may take several different forms, depending upon the intended practitioners and the expected purpose. Unlike cut-down methods, the application representations retain a direct link to the parent theory and, in order to use them, information has to be expressed in theoretically important ways. The advantage for practitioners is that the results of their use is also couched in theoretically relevant forms, and so is linked to an explanatory body of ideas that can be recruited to assist in redesign.

May and Barnard (1995) present the example of a cognitive architecture called interacting cognitive subsystems, or ICS. As a discovery representation, it is unlike GOMS, Soar or EPIC in that it does not seek to simulate cognitive processing, but provides a framework for identifying bottlenecks in the flow of information through a series of cognitive subsystems, each of which is specialized for processing a different level of mental representation. Like the simulation models, these levels include sensory, cognitive, and motor representations. The modeling process involves describing the likely pattern of exchanges between the various levels, the degree of skill-based behavior of relevant aspects of a task, the involvement of the user's recent and longer-term memories, and, crucially, the structure of the user's conceptual task or of perceptual representations of the interface. In this discovery form, the modeling is too complex to be carried out by anyone other than an expert in the theory itself, and the results are couched in terms of the processes and their configuration, rather than specific design consequences.

To support the use of ICS within design, two alternative application representations have been developed: one is a production rule-based 'expert system,' which partly replicates the role of the ICS theorist in constructing a principled model of an interaction; the other is a pencil-and-paper technique for reasoning about the changes over time in the structure of the mental representations encoded at the various ICS subsystems (Barnard and May, 1999). Both of these application representations require some familiarity with the core concepts of ICS, but neither requires practitioners to construct a full ICS model themselves. Their value lies in providing problem-solving tools for designers to apply to partially specified or speculative ideas about interfaces, before any commitment has been made.

Influx of Novel Approaches

In moving away from a simulation of a user interacting with a computer, ICS is part of a trend away from the concentration upon a dyadic, turn-taking relationship between a single user and a single computer. Classical HCI models can be caricatured as having a human issuing commands to a computer, waiting for it to carry them out, and then inspecting the screen while the computer patiently waits for them to form some new goals. While this might have been an appropriate model for HCI in the early 1980s, the advent of distributed computing architectures and the massive interconnection of computing systems over the Internet means that it is inadequate. Even modern personal computers will be carrying out multiple tasks, doing background processing almost in-between each keystroke. The humans are not always isolated users pursuing solitary goals, either, but are using computers as part of their everyday activities, often in social or work settings that involve interacting with other humans with the support of their computers. An analysis of interaction that does not allow for such factors has inevitable flaws.

These developments have led to a new involvement within HCI of social and organizational psychologists, as well as researchers from other social sciences such as ethnomethodology, anthropology, and sociology (Dourish and Button, 1998). This has led to an influx of novel theoretical ideas, in particular Vygotsky's activity theory (e.g., Nardi, 1995) and Suchman's situated action (Suchman, 1987). These ideas re-emphasize the relevance of the context within which human actions are formed, and as Carroll (1997) noted, have brought a continental European perspective into what had previously been a largely North American and British endeavor. Instead of the user and their system, activity theory deals with an 'activity system' composed of social aspects such as attitudes, beliefs, and experiences as well as technological factors.

Another development which moves away from the traditional HCI dyad has come from a computer science perspective, where the idea of a system as a single entity has long been outdated. With the introduction of 'object oriented' ideas, computer scientists have moved towards seeing complex systems as built up from a large number of quasi-independent 'interactors' that exchange information and tasks, with no constraint for these interactors to be part of a single machine. A large body of theory has evolved to model such distributed and concurrent activity, and a logical step has been to see a user's cognitive processes as just another set of interactors. The unit of analysis for such 'syndetic models' (Duke et al., 1999) is the overall system composed of human and computer interactors, and their conjoint activity can be modeled using mathematical formalisms to prove literally that certain interactions can or cannot occur, or that certain assumptions must be met for them to be possible.

Prospects for HCI as a Discipline

Both the social and the computational contributions to HCI currently fall into the discovery mode of analysis, in Dowell and Long's terms, rather than the application mode. It is not yet clear whether these novel directions will crystallize into methods or representations that are directly applicable by design practitioners, or whether they will serve more to direct the commercial development of products away from the classical PC in an office towards other niches. They may also represent the evolution of HCI as a discipline in its own right, with its own methods and theories, rather than as an interdisciplinary area colonized by theories developed in cognitive psychology and artificial intelligence. The tension between academic theorists and practical developers will undoubtedly continue to be a major dynamic upon the field, with its predominance of conference papers enabling nonempirical presentations of novel technologies or design tools to share the limelight with hard-nosed experimental evaluations. In recent years, the increasing theoretical quality of the few specialized HCI journals may provide justification for those who have argued for universities to establish HCI departments in their own right, with their own undergraduate courses. On the other hand, the progressive involvement of computing artifacts within our everyday lives is making their use an ordinary part of human behavior. The increasing contribution of other disciplines is making HCI focus less on the individual, with the consequence that a great deal can be gained from other fields within the social sciences and, perhaps, provide more insight into human behavior in return.

Editor's Note

Research on the human-computer interface has exploded since Turkle's original entry and nearly all of it serves to confirm her prescient conclusion, that "With the computer-human interfaces that researchers project in the near-future, the user will experience the interface as a creature (as in increasingly advanced robots) as prosthetics (a pair of glasses that also serve as a computer screen for mobile net browsing), a piece of clothing (with embedded processors that for example use the temperature of the human body as the input to warm or cool the skin), or as a virtual world that people can fully immerse in and navigate by moving their bodies as one would move in the real physical world." All of these projected future technologies now exist, although some are more developed (and commercially available) than others. Recent sources on human-computer interactions that readers will find useful include: Grudin (2013) and Graimann et al. (2010). The reader's attention is also directed to the International Journal of Human-Computer Interaction, a Taylor and Francis journal where a great deal of new research on the topic appears monthly.

See also: Science and Technology, Social Study of: Computers and Information Technology.

Bibliography

Barnard, P.J., May, J., 1999. Representing cognitive activity in complex tasks. HCI 14, 93–158.

Brooks, F.P., 1975. The Mythical Man Month; Essays on Software Engineering. Addison-Wesley, Reading, MA.

Card, S.K., Moran, T.P., Newell, A., 1983. The Psychology of Human–Computer Interaction. Erlbaum, Hillsdale, NJ.

Carroll, J.M., 1997. Human–computer interaction: psychology as a science of design. International Journal of Human Computer Studies 46, 501–522.

Dourish, P., Button, G., 1998. On 'technomethodology': foundational relationships between ethnomethodology and system design. HCI 13, 395–432.
Duke, D.J., Barnard, P.J., Duce, D.A., May, J., 1999. Syndetic modeling. HCI 13, 337–393.
Gardiner, M.M., Christie, B. (Eds.), 1987. Applying Cognitive Psychology to User–Interface Designs. Wiley, Chichester.
Gray, W.D., John, B.E., Atwood, M.E., 1993. Project Ernestine: validating a GOMS analysis for predicting and explaining real-world task performance. HCI 8, 237–309.
Graimann, B., Allison, B., Pfurtscheller, G., 2010. Brain-Computer Interfaces: Revolutionizing Human-Computer Interaction. Springer, London and New York.
Grudin, J., 2013. Human-computer interaction. Annual Review of Information Science and Technology 45 (1), 367–430.
Kieras, D.E., Meyer, D.E., 1997. An overview of the EPIC architecture for cognition and performance with application to human–computer interaction. HCI 12, 391–438.
Kitajima, M., Polson, P.G., 1997. A comprehension-based model of exploration. HCI 12, 345–389.
Landauer, T., 1995. The Trouble with Computers. MIT Press, Cambridge, MA.
Long, J.B., 1987. Cognitive engineering and HCI. In: Warr, P. (Ed.), Psychology at Work. Pelican, Harmondsworth, UK, pp. 73–93.
MacLean, A., Young, R., Bellotti, V., Moran, T., 1991. Questions, options and criteria: elements of design space analysis. HCI 6, 201–250.
May, J., Barnard, P.J., 1995. Towards supportive evaluation during design. Interacting with Computers 7, 115–143.
Nardi, B.A. (Ed.), 1995. Context and Consciousness: Activity Theory and Human–Computer Interaction. MIT Press, Cambridge, MA.
Newell, A., 1990. Unified Theories of Cognition. Harvard University Press, Cambridge, MA.
Nielsen, J., Mack, R.L. (Eds.), 1994. Usability Inspection Methods. Wiley, New York.
Norman, D.A., Draper, S.W., 1986. User Centered System Design. Erlbaum, Hillsdale, NJ.
Polson, P.G., Lewis, C.H., Rieman, J., Wharton, C., 1992. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. International Journal of Man Machine Studies 36, 741–773.
Suchman, L.A., 1987. Plans and Situated Actions: The Problem of Human–Machine Communication. Cambridge University Press, New York.

Human–Environment Interactions: Case Studies

CG Knight, Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A comparative case study is a research approach to formulate or assess generalizations that extend across multiple cases. The nature of comparative case studies may be explored from the intersection of comparative and case study approaches. Case studies and qualitative comparative analyses share a common explanatory limitation with quantitative models – they suggest reasonable explanations without being able to prove analytic assertions with scientific certainty. Strategies in comparative case studies include research on the basis of community, place, perspective, structure, system, scale, transects, gradients, time, events, analogs, or theory as organizing principles. However, the major organizing principle of a comparative case study often includes others that are subsidiary but inherent in the characteristics of the cases selected. There are no formal rules for evaluating comparative case studies, but reasonable criteria against which studies can be assessed do exist. The intriguing variety of strategies for organizing comparison means that this approach to understanding will continue to be a vital part of human–environment scholarship.

A comparative case study is a research approach to formulate or assess generalizations that extend across multiple cases. The fundamental intent of a comparative case study – development of knowledge – applies equally in the natural and social sciences, with the latter incorporating qualitative analysis in a discursive mode as well as the quantitative and positivist modes of the sciences. Comparative case studies are particularly important to questions of human–environment interaction, combining case study and comparative traditions and thus extending the depth of the individual case study across several instances in time and/or space.

Case Study and Comparative Method

A case study is a rigorous, detailed, spatially bounded, temporally limited description and analysis of an event, process, place, or community intended by the researcher to exemplify a more general insight, principle, or theory. It "is an exploration of a 'bounded system' or case (or multiple cases) over time through detailed, in-depth data collection involving multiple sources of information rich in context" (Creswell, 1997: p. 61). Case studies of particular communities or places are a common and classical mode of analysis in anthropology, human geography, and sociology.

Comparative analysis, in contrast, explores similarities revealed in different situations or cases sharing some common element(s), while differing in others. Comparative analyses are also common among social sciences, varying in the depth to which components are explicated, from statistical abstractions through vignettes to a level of detail common in individual case studies. Often, a case study has unique elements whose effects on the validity of generalizations may be ambiguous or uncertain. Thus, case studies addressed through a conjunction of insights from alternative perspectives will reveal commonalties of particular interest. Comparison between cases can help to delineate such relationships and potentially to clarify ambiguity in a singular case.

Qualitative and Quantitative Aspects

Comparative analyses may proceed via quantitative analysis of numerical abstractions from cases, through detailed narrative description and qualitative analysis. There is no necessary relationship between qualitative and quantitative approaches in case studies and comparative analyses. Qualitative case studies are frequent in anthropology, some schools of sociology, and contemporary cultural geography. Case studies can also be internally as well as comparatively quantitative, particularly in human and economic geography. While comparative studies are often qualitative, they may be quantitative when there are a large number of cases with a small number of variables, as in social survey research.

Explanatory Status of Comparative Case Studies

Both case studies and comparative analyses can be *inductive* or *deductive* (Hempel, 1965; George and Bennett, 2005). However, case studies and qualitative comparative analyses share a common explanatory limitation with models – they are fundamentally *heuristic* rather than *definitive* in that they cannot prove the validity of generalizations beyond providing persuasive evidence in particular contexts (Oreskes et al., 1994). Conversely, a case study illuminating and supporting a general principle does not preclude another case casting doubt. As Leach (1968: p. 344) stated about the comparative method in anthropology, "…comparison…is not, and cannot be, a disguised form of scientific experiment leading to an explanation." Moreover, neither misinterpretation of the case nor failure of the case to illustrate a generalization necessarily means that a hypothesis is not valid. However, a comparative case study approach strengthens the heuristic value of individual cases by bringing in-depth understanding in different circumstances, with greater confidence in the conclusions drawn from them. "The generality of the case study…is relative to the methodological value of the test process involved in the experiment itself and 'holds until proved otherwise'" (Hamel, 1993: p. 40). Multiple cases strengthen the confidence of the researcher that the generality is sound.

Comparative Case Study Strategies

Researchers investigating human–environment issues can avail themselves of many alternative strategies for developing comparative case studies. Strategies included here are certainly not exhaustive or mutually exclusive, but they suggest ways in which comparative case studies can be structured creatively for investigation, presentation, and analysis. Often, the power of the comparative analysis is strongly related to the selected strategy. These strategies include research on the basis of community, place, perspective, structure, system, scale, transects, gradients, time, events, analogs, or theory as organizing principles. In practicality, the major organizing principle of a comparative case study often includes subsidiary dimensions inherent in the nature of the cases selected. For example, selection of case communities along a transect incorporates both community and place perspectives as well, even though the transect is the organizing theme. Examples of each strategy are included in Table 1.

Community

Comparative case studies in disciplines such as sociology and anthropology often focus on studying different social communities to ascertain, for example, generalities surrounding a particular cultural pattern (relations between matrilineal systems and matrilocal residence); social experience (immigrants in the new world city); or community structure (comparative political systems). In human–environment studies, comparison between communities is also important. Examples of such comparative case studies include investigation of traditional or vernacular architecture; studies of pastoral lifestyles and range management strategies; cross-cultural studies of local environmental relations; analysis of ethnic farming populations in the same locale; and studies of traditional navigation technologies and knowledge systems used by island inhabitants.

Table 1 Comparative case studies: a literature guide

Community	Bennett, 1969
	Mortimore, 1999
	Greenberg, 2011
Place	Moran, 1993
	Turner et al., 1993
	Yarnal et al., 2009
Perspective	Knight, 1974
	Meinig, 1979
Structure	Goldschmidt, 1966
	Moran, 1995
	Turner and Brush, 1987
System–subsystem	DeWalt and Pelto, 1985
	Mitchell, 1974
Scale	Wilbanks and Kates, 1999
	Easterling et al., 1998
Transects	Kromm and White, 1992
	Moran and Brondízio, 1998
Gradients	Allan, 1965
	Porter, 1965
Gradients and time	Porter, 2006
	Singh et al., 2013
Time	Broek, 1932
	Carlstein, 1982
Events	Blaikie, 1985
	Burton et al., 1993
	Bankoff et al., 2004
	Piguet et al., 2011
	Landesman and Weisfuse, 2014
Analogs	Glantz, 1988
	Wood and Knight, 1974
Theory	Boserup, 1966
	Worster, 1985

Place

Places that provide examples of a shared geographical setting or similar human–environmental process may be comparatively studied. Examples include studies of human–environment relations in specific kinds of environment, such as mountains, flood plains, and island or coastal areas; research on societal responses to extreme population density; and investigation of places affected by technological transformation, such as impacts of transportation change (superhighways, dieselization of railroads), agricultural mechanization leading to rural depopulation, management of water resources, and transitions in household energy use in developing countries.

Perspective

An alternative to selection of different cases analyzed from a similar perspective is to bring multiple perspectives to a single case. Such an approach explicitly invokes a multioperationism desirable in all case studies – the convergence of different kinds of evidence toward a more complete and rich understanding. One common method for an approach based on perspective is the *emic* versus *etic* familiar from anthropology – the understanding held by the individuals involved, in contrast to the analytical viewpoint of the researcher. This distinction is now an important element of postmodern qualitative inquiry across the social sciences. The use of perspective in comparative study includes landscape analysis, analysis of environmental controversy, and multiperspective community or place studies.

Structure

Comparative case studies may be developed on the proposition that selected cases reflect the functioning of a common underlying structure, somewhat analogous to comparative linguistics. For cross-cultural translation of color terms, a formal Munsell color chart can provide a common structure on which to map various language terms for color. Less formally, comparative analysis may be based on typologies that provide the framework for comparison. In the human–environment realm, structural frameworks used as the basis for comparative case studies include Linnaean plant classification compared to indigenous ethnobotanical plant classification, fundamental requisites of a society, comparative farming systems, and human response systems in the face of natural hazards.

System–Subsystem

A comparative case study may be organized to look at various levels of a complex system to establish relationships,

mechanisms, and emergent properties that vary by the level under investigation. Examples include studies of the public role in environmental management at local, state, and national levels; studies of agricultural change viewed at the family, village, society, and regional levels; and studies of the driving forces of deforestation, from family- and farm-level decisions, through regional and national forces to the global economy.

Scale Analysis

Case studies of the occurrence of phenomena at different spatial scales may be undertaken to ascertain variance or invariance of form or process. Scale analysis has become an important element in the study of the natural environment, including hydrological, climatological, and ecological systems. In human–environment investigations, scale analysis is undertaken in studies of the causes and potential impact of climate change to suggest local and larger scale determinants of greenhouse gas emissions; appropriate levels of generalization of impacts, such as climate change on farming; and initiatives for mitigation.

Transects

Geographical transects provide one kind of spatial sampling that researchers can use to develop information and insight that can be generalized across a society or region. A transect provides one means of minimizing sample bias in an attempt to develop generalizations and analysis of differences within a society or region that might reasonably be assumed to be homogeneous. Such studies include investigation of market towns along a transect in an agricultural region; research on the dynamics of colonist settlement along new roads in rain forest areas; or exploratory archeological investigation of settlement remains in a region. In transect studies, the area being investigated may be found not to be homogeneous, suggesting comparative study of social differences (such as ethnic enclaves in an agricultural region) or environmental differences (such as local availability of ground water in the High Plains of the United States). One special kind of transect selected in recognition of difference rather than homogeneity is the gradient.

Gradients

Cultural ecologists in geography and anthropology often undertake comparative case study analyses using sample cases selected across environmental, spatial, or social gradients rather than in areas representative of a homogeneous area. One principle underlying such studies is that society adapts to environmental differences; thus the purpose of study is to understand the nature of adaptations and how they function in the face of environmental change. Examples include studies of agricultural adaptation along an altitudinal or climatic gradient; investigation of the impact of proximity on market involvement in a developing country; or research on the human–environment interactions in different societies occupying similar environments, either across a social gradient within a similar environmental region, such as farmers in Africa, or living interspersed in the same location, such as pastoralist and agriculturalist land use management.

Time

A temporal approach to comparative case study examines the same place at different times, different places at the same time that reflect temporal differences, or different places at different times. In the first approach, a single place is examined at different times. Such a study might show the development of successive human–environment relations as population or other variables change through time. When a geographical location was occupied by different societies through time, such studies in geography were offered as evidence of 'sequent occupance (sic),' one of the strategies used to refute environmental determinism. Investigation of different places at the same time could be seen as a quasitemporal process when diffusion of an innovation or process of change is spreading geographically and some areas represent an older core location with earlier adaptation, while others represent preadaptation or emerging stages. An example might be the diffusion of a cash-cropping complex in a traditional rural society where various locations along a space–time gradient represent stages of adaptation of a new technology.

Events

In the evolution of comparative case study in the human–environment realm, one salient mode of analysis focused on human response to natural hazards. Natural hazards research sought commonalties, first, in human response to floods, followed by attention to other natural hazards, to the hazardousness of places, and eventually to technological hazards. In all of these realms, comparative case studies provided the core strategy for the development of general frameworks for natural hazard analysis. Seen more generally, selection of hazard case studies for comparative research may be organized around extreme geophysical events, such as tornadoes, hurricanes, and floods, or long-term environmental threats such as drought, deforestation, or soil depletion and erosion. Comparison was also based on spatial dimensionality and temporal occurrence of hazards, as well as on hazard geometry. The event-based approach can be extended to other kinds of events: human and ecological impacts of dams; local resistance to imposed political or technological change; or suburban sprawl.

Analogs

An analog is based on the premise that objects or events having some similarity may also have other similarities. Thinking by analogy, in fact, underlies the comparative case study approach, in that the similarity among cases makes them, in fact, analogous in some respect. As a formal organizing principle, analogs inherently involve two or more cases that share sufficient commonalty that insight can be gained by comparison. There is no formal procedure for developing analogs; rather, good analogies depend on "pragmatic considerations regarding contexts, interests and purposes" (Jamieson, 1988: p. 81) and represent a priori insight before the full development of the comparative cases. In comparative case study using analogs,

there is no necessary juxtaposition of the cases in time or space; rather, comparison originated in a conceptual linkage that appeared promising to the researcher. Analogous case studies are strengthened when they are selected with explicit similarity criteria.

Comparative case studies developed within an analog framework include studies of the impact of modern technology in developing countries; society's response to acid rain, ozone depletion, and future greenhouse gas stabilization; responses to past climate change; and adaptation to drought. Often, comparative case studies built upon analogs are oriented to the future, asking whether insights gained from the analogs can be applied to new decisions.

Theory

Comparative case studies can address theory, wherein the cases are selected to test hypotheses from the theory. Examples include comparative case studies of the process of agricultural change with respect to population growth, such as the Boserup hypothesis; modeling society's progress through economic, demographic, disease, or mobility transitions; testing Wittfogel's concept of the hydraulic civilization; or explicating a framework for understanding the processes of environmental adaptation. In such studies, individual cases are not only compared with other empirical cases, but compared also with the expectations from a priori theory.

The Comparative Case Study Method

The comparative case study approach draws upon the same kinds of method that are common to the case study and comparative methods. The researcher links a hypothesis to certain cases using one or more of the strategies described previously, gathers information about the cases, develops evidence from the cases, presents the cases, and constructs explanation or generalization from the evidence. Insight, creativity, intuition, and even serendipity cannot be dismissed as important elements that the researcher uses to formulate hypotheses and to develop the strategy of research.

The depth of analysis of the constituent cases in a comparative study selected for *instrumental* purposes may not be equivalent to that of an individual case examined for its *intrinsic* interest (Creswell, 1997). Focus on an issue or hypothesis draws attention to specific relevant issues or variables in the cases. Yin (1989) referred to this difference as *holistic* versus *embedded* analysis, the latter connoting the selection of cases within a wider research question.

Tools for Analysis

Researchers developing new case studies for comparative analysis have a wide variety of available tools and data sources, including interviews, surveys, participant observation, ethnographic research, field observation, officially collected information, archival research, environmental analysis, remote sensing, qualitative analysis, quantitative analysis, systems analysis, structural analysis, and various modes of modeling. In contrast, researchers undertaking meta-analysis of previous studies may be confined to information developed through other researchers' selection of tools and information; the researcher may undertake further study to amplify information on the case. As with individual case studies, the results of comparative analyses are usually narrative, although normative, probabilistic, statistical, simulation, or graphic models are possible modes of conceptualizing and communicating research outcomes.

Evaluation of Comparative Case Studies

Although there are no formal rules for evaluating comparative case studies, there exist reasonable criteria against which studies can be assessed. These include establishing the relevance of the cases; thorough identification and consultation of information sources; the adequacy and accuracy of reflecting information from sources; correct interpretation of sources; objectivity in reporting and interpretation; and internal consistency as evidenced by triangulation or multioperationism. If the cases purport to be a 'natural' experiment, controls provide an important strengthening of the argument. Additional criteria include the possibility of replication; reporting of contrary evidence and/or interpretation; specification of exceptions and special circumstances; recognition of and explanation for contrary cases; rigor versus superficiality; and value-added insight and understanding. Primary case studies must also meet the canons of field or archival work in the relevant discipline; comparative case studies built from meta-analysis of previously published work have potential for review by original researchers.

Conclusion

Comparative case studies offer a richness of insight that may be lost in the abstractions of statistical analysis, while maintaining the tradition of the case study as a means to explore generalities in the complexity of the real world. The intriguing variety of strategies for organizing comparison means that this approach to understanding human relationships to the environment will continue to be a vital part of human–environment scholarship.

See also: Comparative History; Comparative Method in Evolutionary Studies; Comparative Studies: Method and Design; Human Ecology.

Bibliography

Allan, W., 1965. The African Husbandman. Barnes and Noble, New York.

Bankoff, G., Frerks, G., Hilhorst, T., 2004. Mapping Vulnerability: Disasters, Development, and People. Earthscan, London.

Bennett, J.W., 1969. Northern Plainsmen: Adaptive Strategy and Agrarian Life. Aldine, Chicago.

Blaikie, P.M., 1985. The Political Economy of Soil Erosion in Developing Countries. Longman, London.

Boserup, E., 1966. The Conditions of Agricultural Growth: The Economics of Agrarian Change under Population Pressure. Aldine, Chicago.

Broek, J.O.M., 1932. The Santa Clara Valley, California: A Study in Landscape Changes. Oosthoek, Utrecht, The Netherlands.

Burton, I., Kates, R.W., White, G.F., 1993. The Environment as Hazard. Guilford Press, New York.
Carlstein, T., 1982. Time Resources, Ecology and Society. George Allen and Unwin, London.
Creswell, J.W., 1997. Qualitative Inquiry and Research Design, Choosing among Five Traditions. Sage, Thousand Oaks, CA.
DeWalt, B.R., Pelto, P.J. (Eds.), 1985. Micro and Macro Levels of Analysis in Anthropology: Issues in Theory and Research. Westview Press, Boulder, CO.
Easterling, W.E., et al., 1998. Spatial scales of climate information for simulating wheat and maize productivity: the case of the US great plains. Agricultural and Forest Meteorology 90, 51–63.
Eisenstadt, S.N., 1968. Comparative study. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences, vol. 14. Macmillan, New York, pp. 421–429.
George, A.L., Bennett, A., 2005. Case Studies and Theory Development in the Social Sciences. MIT Press, Cambridge MA.
Glantz, M.H. (Ed.), 1988. Societal Responses to Regional Climatic Change: Forecasting by Analogy. Westview Press, Boulder, CO.
Goldschmidt, W.R., 1966. Comparative Functionalism: An Essay in Anthropological Theory. University of California Press, Berkeley, CA.
Greenberg, M., 2011. The Environmental Impact Statement after Two Generations: Managing Environmental Power. Routledge, New York.
Hamel, J., 1993. Case Study Methods. Sage, Newbury Park, CA.
Hempel, C.G., 1965. Aspects of Scientific Explanation: and Other Essays in the Philosophy of Science. Free Press, New York.
Jamieson, D., 1988. Grappling for a glimpse of the future. In: Glantz, M.H. (Ed.), Societal Responses to Regional Climatic Change. Westview Press, Boulder, CO, pp. 73–93.
Knight, C.G., 1974. Ecology and Change: Rural Modernization in an African Community. Academic Press, New York.
Kromm, D.E., White, S.E. (Eds.), 1992. Groundwater Exploitation in the High Plains. University Press of Kansas, Lawrence, KS.
Case studies. In: Landesman, L.Y., Weisfuse, I.B. (Eds.), 2014. Public Health Preparedness and Response to Disasters. Jones & Bartlett Learning, Burlington, MA.
Leach, E., 1968. The comparative method in anthropology. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences, vol. 1. Macmillan, New York, pp. 339–344.
Meinig, D.W., 1979. The Interpretation of Ordinary Landscapes: Geographical Essays. Oxford University Press, New York.
Mitchell, J.K., 1974. Community Response to Coastal Erosion: Individual and Collective Adjustments to Hazard on the Atlantic Shore. University of Chicago, Department of Geography. Research Series No. 156.
Moran, E., Brondízio, E., 1998. Land use change after deforestation in Amazonia. In: Liverman, D., et al. (Eds.), People and Pixels: Linking Remote Sensing and Social Science. National Academy Press, Washington, DC, pp. 94–120.
Moran, E.F., 1993. Through Amazonian Eyes: The Human Ecology of Amazonian Populations. University of Iowa Press, Iowa City, IA.
Moran, E.F., 1995. The Comparative Analysis of Human Societies: Toward Common Standards for Data Collection and Reporting. L. Rienner, Boulder, CO.
Mortimore, M., 1999. Working the Sahel: Environment and Society in Northern Nigeria. Routledge, London.
Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation, and confirmation of numerical models in the earth sciences. Science 263, 641–646.
Piguet, E., Pécoud, A., de Guchteneire, P. (Eds.), 2011. Migration and Climate Change. Cambridge University Press, New York.
Porter, P.W., 1965. Environmental potentials and economic opportunities – a background for cultural adaptation. American Anthropologist 67, 409–420.
Porter, P.W., 2006. Challenging Nature, Local Knowledge, Agroscience, and Food Security in Tanga Region, Tanzania. University of Chicago Press, Chicago.
Ragin, C.C., Becker, H.S., 1992. What Is a Case? Exploring the Foundations of Social Inquiry. Cambridge University Press, Cambridge, UK.
Ragin, C.C., 1987. The Comparative Method: Moving beyond Qualitative and Quantitative Analysis. University of California Press, Berkeley, CA.
Singh, S.J., Haberl, H., Chertow, M., Mirtl, M., Schmid, M., 2013. Long Term Socio-ecological Research: Studies in Society-nature Interactions across Spatial and Temporal Scales (Human–Environment Interactions). Springer, Dordrecht, The Netherlands.
Stake, R., 1995. The Art of Case Study Research. Sage, Thousand Oaks, CA.
Turner, B.L., Brush, S.B. (Eds.), 1987. Comparative Farming Systems. Guilford, New York.
Turner, B.L., Hyden, G., Kates, R.W., 1993. Population Growth and Agricultural Change in Africa. University of Florida Press, Gainesville, FL.
Wilbanks, T.J., Kates, R.W., 1999. Global change in local places: how scale matters. Climatic Change 43, 601–628.
Wood, J.S., Knight, C.G., 1974. Cross-cultural perspectives on human adjustment to arid margins. In: Newman, J.L. (Ed.), Drought, Famine and Population Movements in Africa. Syracuse University Foreign and Comparative Studies/Eastern Africa Series No. 17, Syracuse, NY, pp. 75–126.
Worster, D., 1985. Rivers of Empire: Water, Aridity, and the Growth of the American West. Pantheon, New York.
Yarnal, B., Polsky, C., O'Brien, J., 2009. Sustainable Communities on a Sustainable Planet: The Human-Environment Regional Observatory Project. Cambridge University Press, Cambridge, England.
Yin, R.K., 1989. Case Study Research: Design and Methods. Sage, Newbury Park, CA.

Hume, David (1711–76)

John P Wright, Central Michigan University, Mt. Pleasant, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

David Hume played a major role in the founding of the social and behavioral sciences in the Scottish eighteenth-century Enlightenment. The article discusses his general psychological and philosophical principles, his skepticism, his contributions to moral and political theory, to economics and politics, to history and historiography, and to the philosophical, psychological, and sociological study of religion.

David Hume, a leading figure in the Scottish Enlightenment, contributed significantly to the disciplines of psychology, history, political science, and economics, as well as to more theoretical discussions in ethics, aesthetics, epistemology, and philosophy of religion. He was a controversial figure in his own day because of his criticisms of both natural and revealed religion, and because of the political opinions he expressed in his popular political essays and his very successful *History of England*. His skeptical arguments concerning induction are still discussed today by philosophers interested in probability theory, and his account of the foundations of morality enters into current disputes about the nature of moral judgment.

Life and Writings

Hume was born in Edinburgh on 26 April 1711. The family was of genteel but modest means: His father, Joseph Home, an advocate and member of the minor gentry of Scotland, died when he was 2 years old; and his mother Katherine, who never remarried, raised him, his sister, and older brother on their small estate of Ninewells on the Scottish Borders. Hume probably began taking classes at Edinburgh University in 1721 where he studied Greek and Latin, scholastic logic and metaphysics, natural philosophy, and took an extramural class in mathematics. There is no record of his ever having taken a degree (which was not unusual at this period) and, in a short autobiography written just before his death, he noted that, while his family intended that he study law, his own interests were in "Philosophy and general Learning." He stated that he planned his first philosophical work, *A Treatise of Human Nature*, before he left University, probably in 1725 or 1726, and from his correspondence we learn that he began struggling to write about psychological phenomena in 1727.

Suffering from complaints that we would now consider to be largely psychosomatic, Hume decided in 1734 that he needed to pursue a more active life; he left Scotland for Bristol where he worked briefly for a sugar merchant. After a few months, he went to France where he spent the next 3 years writing and studying. He first went to Reims, but soon settled at La Flèche in Anjou, in the vicinity of the famous Jesuit college where more than a 100 years earlier René Descartes had been a student. The first two books of his *Treatise of Human Nature* ("Of the Understanding" and "Of the Passions") were published in February of 1739 over a year after he returned to England, and the third ("Of Morals") was published 21 months later, after he returned to Scotland. In his autobiography Hume claims that the *Treatise*, from which he hoped to gain literary fame, "fell dead-born from press"; however, there is evidence that it did make a stir among the Edinburgh literati, and it has been taken by scholars since that time as Hume's most important and seminal work. It was followed in 1741–42 by his very successful *Essays, Moral and Political*, written to reach a more popular audience.

Hume unsuccessfully applied for university chairs at Edinburgh University in 1744–45 and at the University of Glasgow in 1751–52, where he had strong support from his friends (including Adam Smith and physician William Cullen) but was opposed by many of the clergy who accused him of irreligion. While his *Treatise* contained no explicit discussions of religion, this was changed in his rewriting of its first book, which was published under the name of *Philosophical Essays Concerning Human Understanding* in 1748. It contained his famous argument against rational belief in testimony for miracles, as well as the basis for his later attack on the design argument for the existence of God. Continuing to search for a career, Hume spent some time during this period as Secretary to General St Clair and participated in a rather futile attack in 1746 on the coast of Brittany during the War of Austrian Succession. In 1747, he traveled with the General as his aide-de-camp on a Military embassy to the capitals of Britain's allies Austria and Sardinia.

In 1749, Hume returned once more to Scotland and began an immensely productive decade in which he became one of the leading British writers of the day. In 1751, he moved from Ninewells to Edinburgh where he was to live for much of the rest of his life as a member of what has become known as the Scottish Enlightenment. In 1751, he was the elected secretary of the Philosophical Society of Edinburgh, the forerunner of the Royal Society of Edinburgh, a group devoted to the study of natural science, medicine, and technology. (He was later, in 1762, to present Benjamin Franklin's paper describing the use of lightning rods to the Society.) In 1751, he also published his *Enquiry Concerning the Principles of Morals*, a reworking of the third book of his *Treatise of Human Nature*. His correspondence at this time also indicates that he had already written substantial portions of his posthumously published *Dialogues Concerning Natural Religion*. In 1752, Hume published his popular *Political Discourses*, which included important essays on economics, and began work on what was to become his most successful piece of writing, namely, his multivolume *History of*

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61055-3

England. He became Keeper of the Advocates Library (later incorporated into the National Library of Scotland) in January, which gave him direct access to a large collection of historical sources. The first volume of the *History*, concerned with the reign of James I and Charles I, was published in 1754 and the remaining volumes between 1756 and 1761. In 1758, Hume published *Four Dissertations*, which included his "Natural History of Religion," "Of the Passions" (a reworking of the second book of the *Treatise*), "Of Tragedy," and "Of the Standard of Taste."

In 1763, after the end of the Seven Years' War, Hume was invited to go to France as Secretary to the Ambassador, Lord Hertford. In France, where his writings were admired, he was lionized both by the mighty and by the intellectuals. He developed close relations with many of the *philosophes*, especially d'Alembert. Hume was *Chargé d'affaires* for about 4 months in 1765 after Hertford left Paris. He returned to England at the beginning of 1766 bringing with him Jean Jacques Rousseau, on the advice of his Parisian friends; Rousseau sought refuge in England from persecution on the Continent. Hume and other British friends procured a pension for him from the King. However, this whole affair turned into a disaster as a result of Rousseau's paranoia. Rousseau denounced Hume as his enemy. Hume's defense of his own conduct was published by d'Alembert and Suard in Paris in October of 1766. His last public appointment was as an Undersecretary of State from February 1767 to January 1768. He retired to Edinburgh where he continued to publish corrected editions of his writings – that is, all but the *Treatise* which he publicly disowned in 1772 in favor of his later philosophical writings including the two *Enquiries*, *A Dissertation on the Passions*, and his *Natural History of Religion*. Hume died on 25 August 1776 as one of Britain's foremost men of letters.

General Philosophical and Psychological Principles

In his *Treatise of Human Nature*, Hume set out to provide a foundation for all the sciences – natural as well as social – on a science of human nature. The central principle of that science was the association of ideas which, Hume wrote, 'enters into all his reasonings.' He held that ideas are associated through resemblance, contiguity in space and time, and causality. He wrote that these three modes of association "are the only links that bind the parts of the universe together, or connect us with any person or object exterior to ourselves." Mental energy – what he called 'force and vivacity' – is transferred via the associational links from one perception to the related ones. This theory played a central role in his account of the source of our inductive reasoning processes and our natural common sense belief in reality in his theory in Book 1, "Of the Understanding," in his theory of the indirect passions in Book 2, "Of the Passions," and in the account of sympathy which forms the basis of the theory "Of Morals" which he developed in Book 3. Hume attributed the association of ideas to the faculty of imagination, and contrasted its conclusions with those based on reason and the senses.

Hume saw himself as following the lead of recent British philosophers including Locke and Hutcheson who had begun to apply to the human sciences the experimental method that had proved so successful in the physical sciences. However, he argued that, unlike the physical sciences, it is impossible to carry out controlled experiments in the science of human nature because introspection interferes with the natural principles it seeks to discover. Rather, he argued that the human sciences must base their principles on the observation of human behavior. Many of his inferences were drawn from his vast knowledge of history, though in Book 2 of the *Treatise* he considered the way different psychological effects result from different social situations.

In his discussion "Of the Understanding" in Book 1 of the *Treatise*, Hume followed Locke in arguing that all our ideas are derived from experience, that is, from the *impressions* of sensation and reflection. In Hume's hands, this principle led to a deep skepticism; he argued that our most natural beliefs concerning the nature of space and time, the existence of a world independent of our senses, of a single self which lasts throughout our lives, and of objective causal power cannot be derived through reason. Rather, these beliefs have their source in 'fictions' or 'illusions' generated by the imagination and the association of ideas.

Scholars still disagree about the nature of Hume's own basic philosophical commitments. Some, following the logical positivists, hold that Hume rejected all metaphysical claims, including the claim that there are unknown powers in the natural world, which are independent of human understanding. For these interpreters, his principle that all ideas are derived from impressions, trumps the fictions produced by the imagination. Others regard this 'copy principle' only as a source of the extreme or Pyrrhonian skepticism that he described so vividly in the conclusion to Book 1 of his *Treatise*. They note that he himself maintains what he called 'a mitigated skepticism, or Academical philosophy' in response to his own and other's Pyrrhonian arguments. According to the mitigated skepticism that he adopted at the end of his *Enquiry Concerning Human Understanding*, "philosophical decisions are nothing but the reflections of common life methodized and corrected." On this reading, natural instinctive beliefs, guided by the principles of the association of ideas, provide Hume with what he called 'a legitimate Ground of Assent.' These natural beliefs gain legitimacy from their universality and the fact that they are necessary for human survival. The overall aim of Book 1 of his *Treatise* and his first *Enquiry* was to show the narrow limits of our human understanding, and the way that the natural processes of the imagination are required to give us a purchase on reality.

While he used induction in his own science of human nature, Hume argued that induction is not itself founded in reason, but has its source in custom and habit. From our past experience of the constant conjunction of one kind of event followed by another we infer like events in the future. But this inference itself presupposes the principle of the uniformity of nature. Hume argues that any claim to justify *this* principle by past experience would be circular, and that it is impossible to demonstrate that the laws of nature will continue in the future as they have in the past. He concludes that the inference from past regularities to future events, even in experimental science, is not founded on reason but on 'an instinct or mechanical tendency,' which we share with other animals.

Hume extended his claim that nothing but constant experience makes us ascribe causal necessity to human actions.

He argued against those who claim that while physical events are necessitated by their causes, the human will is free. He noted that while we have a false sense of liberty when we reflect on our own actions, we clearly accept necessity when we consider the actions of others. We recognize that their actions are regular, and can generally be inferred from their motives, character, and circumstances. Thus, he argued, there is just as much evidence of necessity in human actions as in physical events, and in both cases a scientifically-minded-person will assume necessity even when the regularity is not apparent. Hume argued further that we assume necessity in holding persons responsible for their actions. Punishment and reward are only justified if actions flow from established character traits, not from random actions. It is only if intentional actions are determined by something fixed and durable in a person's character that they are the proper subjects of praise and blame, and of moral responsibility.

Moral and Political Theory

In Book 3 of his *Treatise of Human Nature ("Of Morals")*, Hume followed Francis Hutcheson in arguing that our moral judgments are at bottom based on sentiment rather than reason. But he differed from Hutcheson in arguing that it is our *sympathy* with the pleasure and pains of others, rather than an innate moral sense, which forms the basis of our moral judgments. While he argued that there are no objective moral facts, Hume stressed in his *Enquiry Concerning the Principles of Morals* that before we can arrive at a correct moral judgment we need to employ our reason in making careful distinctions, comparing different situations, and determining general facts.

He distinguished between *natural* virtues such as benevolence, love of children, and pity for the unfortunate, and *artificial* virtues such as adherence to rules of ownership of property, promise keeping, and allegiance. While we have a natural motivation to be benevolent, the artificial virtues depend on the development of social conventions. Hume argued that these virtues are valued merely because of their utility for the maintenance of society – a claim he subsequently extended to the most important natural virtues.

Hume begins his account of the origin of justice in the *Treatise* by focusing on the development of the conventions of private property. Such conventions become necessary because of the selfishness and limited generosity inherent in human nature and the instability and scarcity of external goods. Hume argues that men have an unquenchable desire to gain more and more goods for themselves and their families, and that it is this desire more than any other passion which is directly destructive of society. Like Hobbes before him, he postulates a state of nature that is 'full of war, violence, and injustice.' While, like Hobbes, he thinks that a solution must come from 'judgment and understanding' rather than the passions, he argued that these faculties operate in an unexceptional way according to the principles he laid down in Book 1 of the *Treatise*. A convention or agreement to refrain from the goods of others "arises gradually, and acquires force by a slow progress, and our repeated experience of the inconveniences of transgressing it." It is through custom and habit that we develop a convention for the orderly transference of goods through mutual consent. Thus, while promoting their own self-interest men develop "a general sense of common interest … which induces them to regulate their conduct by certain rules" for the stability of possessions. It is this common interest or *utility* which forms the basis of the moral value we ascribe to those who follow these rules in an established society.

But while *utility* accounts for the need for general rules for the possession and transfer of property and performance of contracts, Hume argues that specific rules determining the right of property must be based on nonrational principles of the imagination, or the association of ideas. Thus, for example, the ascription of property to the first possessor, while not more useful than giving it to the person who will improve it, is adopted because the mind moves naturally from the object to what is contiguous. Again, the imagination favors the rule of inheritance from parent to child, given the causal relation between them. While he rejects any natural right to what one has mixed one's labor with, Hume still argues that the imagination will favor that relation because of the association of the cause (the work of the laborer) with the effects (the change in the object on which the labor is applied). But this can only be one of a variety of principles which determine the positive law for the possession and transfer of property, and the balance between these various rules can only be established by custom and opinion.

Hume presented a parallel account of the origin of government, and the virtue of allegiance to it once it is established. The necessity of rulers would have become apparent as soon as human beings began to live in large impersonal societies where they were naturally inclined to seek short-term gain without considering their long-range interest in maintaining the rules for the stability of property. Reason suggests the solution of accepting magistrates who enforce the rules of justice and punish offenders impartially. But again, on Hume's account this solution developed gradually: men first gave their allegiance to leaders in war when threatened by a common enemy, and those leaders and successors became their rulers.

Hume opposed the Lockean or Whig theory that allegiance was based on a promise made in an original contract. He pointed out that there was no historical evidence for such a contract, that promises themselves require a convention as much as government, and that popular opinion gives no credence to such an opinion, except in Britain where it was inculcated by Whig party politics. But while he opposed the Whig theory, he agreed that government can only persist with the acquiescence of the governed – and not simply by force. For Hume, this acquiescence is maintained by custom and opinion, and he wrote that governments whether monarchical or democratic 'are equally legal, if established by custom or authority.' Hume was equally opposed to the Tory theory of "Passive Obedience" which was based on a theory of Divine right and that active resistance to the established government was never justified. He recognized that when government no longer serves the interests for which it was formed 'the obligation to obedience must cease.' What then results is what he called an 'appeal to heaven, that is, by war and violence.' However, violent revolution can only be 'the last refuge in desperate cases.'

While the general necessity of allegiance is based on its utility for the public good and that of each subject, Hume held

that this necessity did not solve the problem of determining who had the right to govern – a problem which was still an issue in Britain in his own day. He held that there was no rational solution, but that the imagination suggested several which had been adopted historically when succession was in question – including long possession, present possession, conquest, family succession, and positive law. He pointed out that while the last of these proves the best solution, even it depends on accepted custom and opinion. He argued that given the heavy costs of civil war and that a constant threat of rebellion would encourage authoritarian government, habits of obedience are necessary in any political society.

Economics and Politics

Hume was an important forerunner of classical liberal economic theory. His friend Adam Smith noted that Hume was the first to argue that economic forces determine political structures rather than vice versa and that the development of modern liberty and government by law was due to the development of trade and industry. Hume saw in the rise of the merchant and manufacturing classes a check to the despotism that existed in premodern society. He argued that it was the interest of these classes to live by equal laws and to limit the power of both the aristocracy and monarchy. His analysis was based on comparisons of the governments of modern Europe with those of the ancient and medieval worlds, as well as the rise of parliament in the early seventeenth century that led to the English Civil War. While he accepted James Harrington's social determinism, his originality lay in the postulation of commercialism as the key factor limiting the absolute power of the monarch.

Developing the ideas of Bernard Mandeville, Hume opposed those in his day who held that the creation and consumption of luxury goods enfeeble both the individual and the state as a whole. Hume argued that the increase and consumption of commodities 'which serve the ornament and pleasure of life, are advantageous to society,' and that a consumer society encourages the development of hard-working persons whose labor can easily be turned to the service of the state in times of national need.

Hume regarded economic behavior as deeply rooted in human psychology. He argued that human beings have an insatiable desire for activity, and that there is nothing that satisfies this desire more than lucrative employment. There is, he argued, no pleasure which feeds on itself so much as that of the merchant who sees his fortune grow. The habit of industry stimulated by desire for this pleasure creates a frugality which does not naturally exist in an agrarian society based on inherited wealth. The love of gain, which marks the activity of the merchant, results in large sums of money that can be lent at low interest to stimulate further economic growth.

In his essay "That Politics may be reduced to a Science," Hume stated that while individual human actions are often unpredictable, politics, which studies the effects of laws and constitutions can arrive at generalizations through experience which approach the certainty of mathematics. He argued, for example, that nonrepresentative democracy inevitably degenerates into anarchy and eventually to despotism – as happened in ancient Athens and at the end of the Roman Republic. However, he also saw the danger of generalizing from limited historical experience, as he thought Machiavelli had done in drawing conclusions about the instability of monarchical forms of government from 'the furious and tyrannical governments of ancient times' and the small principalities of Renaissance Italy. Hume noted that Machiavelli's principles had been proved false by the experience of the large European hereditary monarchies of the seventeenth and eighteenth centuries. While he himself put forward generalizations concerning different forms of government, his skepticism led him to recognize that these would need to be modified by future circumstances that had not yet existed in human history.

Hume was a critic of the civic humanist tradition that held that good government must be based on the virtue of politicians. He argued that good constitutions must recognize that human beings are predominantly self-interested. He recognized three forms of legitimate government in the eighteenth-century world in which he lived – republics such as that of the Netherlands, civilized monarchies such as France, and the mixed form of government of Britain which, he thought, tended more toward the republican form because of the independence of parliament and the separation of powers. Each form of government promoted rule by law, as opposed to rule by the arbitrary discretion of the monarch, and in each there was an inevitable struggle between authority and liberty. The danger of civilized monarchy such as existed in France lay in the heavy burden of taxes on the land which impoverished the peasants, as well as creating a dangerous inequality which weakened the overall power of the state. But Hume also saw a potential instability in the republican and mixed forms of government of the Netherlands and Britain that relied on public debt to finance the interests of the state. Nevertheless, he favored the latter forms of government since they encouraged commerce and equal distribution of wealth. Hume wrote that equality "diminishes much less from the *happiness* of the rich than it adds to that of the poor" while at the same time it "augments *the power of the state*."

While he was an early supporter of American independence, the "Wilkes and Liberty" riots of the late 1760s and early 1770s convinced him that the balance of power in Britain had shifted too much in favor of liberty. This and other statements in his later life have led some scholars to consider him a supporter of counter-revolution and a forerunner of the conservatism of Edmund Burke.

Historiography and History

Hume was both a philosopher of history and a philosophical historian. He both articulated and applied principles for the assessment of historical sources. These principles are carefully laid out in his essay "Of Miracles" where he discusses the criteria we ought to use to judge the reliability of witnesses, as well as the need to balance those results against the intrinsic likelihood of the occurrence of the event. Following discussions of 'probability' in John Locke and other early modern writers, he formulated criteria for the reliability of written and oral testimony: these included considerations such as consistency of different reports of the event, the character and expertise of the witnesses, their number, and their self-interest

in convincing others of their accounts of actions and events. Hume held that we need to balance these considerations of the reliability of testimony against the absolute likelihood of the occurrence of a miracle. Given that a miracle is defined as a violation of a law of nature and the existence of such a law is determined on the basis of an invariable experience, there is extremely strong evidence against the existence of any miracle. He argues that no historical testimony of a miracle has succeeded in overtopping the negative evidence that arises from the very nature of the claim that a miracle exists.

More generally, Hume argues that we must balance the testimony given by historical witnesses against the inherent probability of the occurrence of any event. In his essay "Of the Populousness of Ancient Nations," a major pre-Malthusian demographic study, he challenged the commonly held view, based on the writings of ancient historians, that the ancient world was more heavily populated than modern Europe. He supported his argument on a number of grounds including the lack of opportunity of ancient slave populations to reproduce, on the destructiveness of ancient wars, and on the unlikelihood that ancient economies could support vast populations. In his *History of England*, Hume also frequently appealed to general principles of human nature, those which he believed to be established by the newly emerging human sciences, to evaluate the credibility of his historical sources.

Hume sought to write an impartial *History of England* which favored neither Whigs nor Tories, though, in his own day, it was criticized by both. He also broke with those who wrote histories from a religious perspective, seeing either the hand of providence in the affairs of men or the world as degenerating from a golden age. His aim was to write a history which recognized advances in civilization where they occurred, but also recognized the tenuous nature of human progress which was constantly under threat of relapsing into barbarism.

In the *History*, Hume charted the vast changes in the English constitution from the *Magna Carta* to the Revolution of 1688, and the overthrow of James II. His *History* was controversial in his own day because he argued that there was 'no regular plan of liberty' before the reign of the first two Stuart kings, when Parliament began to act independently of the monarch. On Hume's account, the reign of the Tutor monarchs including Elizabeth I was despotic, and that the country did not make any significant economic or political progress until the first half of the seventeenth century.

Hume held that there are a variety of causes that contribute to historical change. As already mentioned, he considered the rise of a middle class as playing a major role in the development of English liberty. He also took account of the unintended consequences of historical actors. While the religious fanaticism of the Puritans led to the execution of Charles I in 1649 and the dictatorship of Cromwell during the Commonwealth period, its ultimate result was the 'mixed form of government' on which English liberty was established. His moral praise was always for those who after 1642, whether supporters of the king or parliament, favored the balance of power. This led Hume to 'shed a generous tear for Charles I' who he believed had honestly supported limiting the power of the monarchy after 1642, but also led Hume to condemn the actions of James II in the 1680s because he violated what had by then become the established constitution.

Religion

Hume was a major Enlightenment critic of religion – particularly of monotheism, which he considered inherently intolerant and subversive of morality. In his *Natural History of Religion*, he argued that religion has its primary roots in our fear of the unknown forces that affect life and death, as well as our natural tendency to think of those forces in humanlike terms. Since these forces affect us in arbitrary ways, we seek to placate them as we would human tyrants, through praise and debasement of ourselves. This, he argued, is particularly true of monotheistic religions, where the single deity is thought of as infinitely more powerful than humans. Debasing actions such as fasting or self-punishment are performed to please such a deity – not moral actions which, according to Hume's own philosophy, have their basis in our sympathy with the pain and pleasure of our fellow humans. Hume explained the intolerance of Christian sects to one another as arising from their reasoning that a single God would not allow different rituals and beliefs to His devotees. Polytheists, on the other hand, can easily accept their neighbors' worship of different gods, are willing to incorporate them into their own pantheon and adopt their practices. He attributed intolerance to "almost all religions, which have maintained the unity of God."

Like many of his Enlightenment contemporaries, Hume was particularly critical of the clergy. He argued that the character that must be maintained by members of this profession inclines to hypocrisy, since no one can maintain genuine religious sentiments for any sustained period. Also, it was in their self-interest to maintain the superstitious beliefs and practices of their own followers, as well as their opposition to rival sects. To avoid the public disorder caused by religious factions Hume favored the existence of an established church.

He criticized the natural or philosophical religion embraced by many of his enlightened eighteenth-century contemporaries. In his posthumously published *Dialogues Concerning Natural Religion*, he presented the most comprehensive critique of the intelligent design argument before Darwin. Hume argued that there is no reason before experience to think that order and apparent design must proceed from intelligence, and that it is only our parochial experience of human intelligence that suggests such a connection. Through the *persona* of his skeptic *Philo*, he argued that the analogy between the order found in nature, and the order and purpose found in the works of man is weak; that we experience order arising from other causes than intelligence; that our experience of the universe itself is limited in space and time and cannot form the basis for principles which operated at the beginning of the universe; and that intelligence in the only case that we experience it – namely in human beings – itself arises from natural causes. However, Hume rejected the possibility that there was chance or indeterminacy in nature, and was unable to put forward an empirically based alternative explanation for the order and design apparent in nature.

Hume also argued that the problem of evil was insoluble by any rational means. In his essay "Of a particular providence and a future state," he pointed out that one cannot legitimately infer more goodness in the creator than is actually found in the creation which is observable to us. While one can speculate about an afterlife that compensates for the injustices and

suffering of this life, one cannot establish it on any rational basis. Indeed, in his *Dialogues*, his skeptic *Philo* argues that the presence of natural and moral evil in this world should lead us to the conclusion that its creator, if powerful, is entirely indifferent to human conceptions of good and evil.

Conclusion

While he was still completing his *Treatise of Human Nature* in 1740, Hume wrote to Francis Hutcheson that in a cool moment he believed that his ideas would be more likely to furnish hints to the future rather than add to the sum total of human knowledge. This reflection was quite prescient. In 1776, the year of Hume's death, Jeremy Bentham, the founder of utilitarianism, wrote that the scales were lifted from his eyes when he read Book 3 of the *Treatise*, where he found "demonstrated, after a few exceptions made" that "the foundations of all *virtue* are laid in *utility*." Bentham added that the key difference between him and Hume was that while Hume used the idea of *utility* to account for the way things are, he (Bentham) used it to account for the way things *ought to be*. The most influential contemporary critic of his epistemology, Thomas Reid, wrote to Hume that he would always avow himself to be his "Disciple in Metaphysics." However, while Reid presented a naturalistic account of our cognitive faculties, he and other writers of the Scottish School of Common Sense rejected Hume's empiricism, his stress on the association of ideas in human thinking, and his skepticism. For them, unlike Hume, there is no conflict between our natural common sense beliefs and the foundations of scientific knowledge. While the most influential of his writings – his *History of England* – was the standard for its subject until well into the nineteenth century, Hume's rejection of the standard Whig account of development of the English constitution was vehemently opposed by republicans such as Thomas Jefferson. In general, while Hume's writings were seminal, his successors have challenged many of his arguments, and their central role has been to contribute to further empirical study of the many subjects he dealt with.

See also: American Revolution, The; Balance of Power, History of; Bourgeoisie and Middle Classes, History of; Causation: Physical, Mental, and Social; Causes and Laws: Philosophical Aspects; Civil Society, Concept and History of; Conservatism: Historical Aspects; Contractarianism; Conventions and Norms: Philosophical Aspects; Economic History; English Revolution, The; Enlightenment; Ethics and Values; Free Will and Action; Freedom: Political; Historical Thought and Historiography: Christian Tradition; History and the Social Sciences; Intellectual History; Justice: Philosophical Aspects; Liberalism: Historical Aspects; Logical Positivism and Logical Empiricism; Mental Causation; Parliaments, History of; Political History: History of Politics; Political Parties, History of; Religion, History of.

Bibliography

Baier, A., 1991. A Progress of Sentiments: Reflections on Hume's Treatise. Harvard University Press, Cambridge, MA.
Bonge, L., 2000. David Hume: Prophet of the Counter-Revolution, second ed. Liberty Fund, Indianapolis.
Buckle, S., 2001. Hume's Enlightenment Tract: The Unity and Purpose of Hume's Enlightenment Tract. Oxford University Press, Oxford, UK.
Garrett, D., 1997. Cognition and Commitment in Hume's Philosophy. Oxford University Press, UK.
Harris, J.A., 2015. Hume: An Intellectual Biography. Cambridge University Press, Cambridge, UK.
Hume, D., 1983. In: Todd, W.B. (Ed.), The History of England from the Invasion of Julius Caesar to the Revolution in 1688. Liberty Classics, Indianapolis, IN.
Hume, D., 1987. In: Miller, E.F. (Ed.), Essays Moral, Political and Literary. Liberty Classics, Indianapolis, IN.
Hume, D., 1998. In: Beauchamp, T.L. (Ed.), An Enquiry Concerning the Principles of Morals. Clarendon Press, Oxford, UK.
Hume, D., 2001. In: Beauchamp, T.L. (Ed.), An Enquiry Concerning Human Understanding. Clarendon Press, Oxford, UK.
Hume, D., 2007a. In: Beauchamp, T.L. (Ed.), A Dissertation on the Passions, the Natural History of Religion: A Critical Edition. Clarendon Press, Oxford, UK.
Hume, D., 2007b. In: Norton, D.F., Norton, M.J. (Eds.), A Treatise of Human Nature, vol. 2. Clarendon Press, Oxford, UK.
Hume, D., 2007c. In: Coleman, D. (Ed.), Dialogues Concerning Natural Religion. Cambridge University Press, Cambridge, UK.
Kail, P., 2007. Projection and Realism in Hume's Philosophy. Oxford University Press, Oxford.
Mossner, E.C., 1980. The Life of David Hume, second ed. Clarendon Press, Oxford, UK.
Norton, D.F., Taylor, J. (Eds.), 2009. The Cambridge Companion to Hume, second ed. Cambridge University Press, Cambridge, UK.
Read, R., Richman, R.K. (Eds.), 2007. The New Hume Debate, revised ed. Routledge, London.
Smith, N.K., 1941. The Philosophy of David Hume: A Study of Its Origins and Its Central Doctrines. Macmillan, London.
Stewart, J., 1992. Opinion and Reform in Hume's Political Philosophy. Princeton University Press, Princeton, NJ.
Stewart, M.A., Wright, J.P., 1994. Hume and Hume's Connexions. Edinburgh University Press, Edinburgh, Scotland.
Wennerlind, C., Schabas, M. (Eds.), 2008. David Hume's Political Economy. Routledge, Abingdon, Oxford.
Wright, J.P., 1983. The Sceptical Realism of David Hume. Minnesota University Press, Minneapolis, MN.

Relevant Websites

http://www.humesociety.org/.
http://www.humesociety.org/hs/index.html.

Humor, Anthropology of

Henk Driessen, Radboud University Nijmegen, Nijmegen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The capacity to joke and perceive something as humorous is universal. However, what is held to be funny is relative. In most societies, there is a distinction between the diffuse humor of the common people and the institutionalized humor by experts at specific times and places. Humor ranges from mild joking to biting satire. Worldwide topics relate to the boundaries between nature and culture, self and other. Many jokes deal with commonly experienced ambiguities, dilemmas, and paradoxes. Functions of humor include entertainment, release of tension, cultural critique, and the marking of group boundaries. This article reviews some forms, domains, and functions of humor.

The Nature of Humor

The terms humor, joking, and laughter are often used interchangeably and are part of the semantic field of the comic that in turn belongs to the domain of expressive culture. There is no single acceptable definition of humor but scholars generally agree that it revolves around a sense of the incongruous or ludicrous. Humor is omnipresent yet elusive, nonsensical yet serious, friendly yet aggressive or hostile, universal yet specific. Humor and joking are part of the human condition yet may differ from time to time, place to place, and even from person to person. A joke can make sense across two millennia, as this 'stupid peasant' joke from ancient Greece testifies: "A peasant, having heard that parrots live for a 100 years, bought one to see if it was true." A joke may also change or lose its point as it moves between cultures. The 'dirty jokes' that provoke laughter in Tamil popular theater would hardly draw a smile from Dutch observers. Humor is notoriously hard to translate from one language to another.

A joke is something said or done in a playful manner to excite mirth. Jokes are relativizers, creating in a playful manner distance from turmoil and toil, misery and misfortune. Humor permeates all domains of public and private life. Worldwide topics include sex, food, health, death, the stranger, and lunatic. Understanding humor often requires intellectual effort, yet scholarly attention partly spoils the punch line. The comic is part of everyday life yet stands apart. Humor creates a momentary world of make-believe. It is closely related to laughter, yet one can appreciate a joke without laughing, and not all events and actions that trigger laughter are humorous.

The paradoxical nature of humor is further complicated by the fact that it is perception and performance, cognition and action, intellectual and emotional, institutionalized and spontaneous. Cognition is implied in grasping a double entendre or perceiving an incongruence. On the other hand, much humor touches the emotional stratum of cultural expectations and involves mimic, gestures, pantomime, and horseplay. An apt one-line summary of humor is "a play upon form" (Douglas, 1968) or "a play upon meanings" (Zijderveld, 1983).

The Study of Humor

The paradoxical nature of humor helps to explain why philosophers have been interested in it since ancient times as well as social scientists have long neglected it. Philosophers including Plato, who detested humor; Erasmus and Rabelais, who celebrated it; and Bergson and Plessner, who have devoted learned treatises to the mechanisms of joking and laughter. In their perspective, there is a bias toward intellectual and linguistic mechanisms, especially the awkward incongruity, perceived in surprise, which most thinkers define as the essence of humor.

Freud promoted the psychological study of joking and considered it a regression to infantile modes of thinking, an expression of anxiety, and a substitute for aggression and libido. Freud concentrated on *der Witz*, which is the quintessence of Yiddish humor: a short, pointed, witty story or pun. During the 1980s and 1990s, humor research has expanded both in terms of disciplines involved and number of publications. In 1988, the International Society for Humor Studies, an interdisciplinary professional association, launched the journal *Humor: International Journal of Humor Research*. Linguists and psychologists dominate the inquiry into humor but more recently it has become increasingly popular among scholars of cultural and communication studies, anthropologists, sociologists, political scientists, and clinical researchers (cf Nilsen, 1993). Over the past 15 years, publications have grown exponentially.

Anthropology

It is said that Bronislaw Kaspar Malinowski, one of the founding fathers of social anthropology, defined his discipline as the science of the sense of humor. Indeed, the anthropologist shares with the humorist a strategy – to use a Geertzian phrase – of making the familiar strange and the strange familiar. There are also family resemblances between anthropologists and tricksters (cf. Van Meijl, 2005). Both compensate physical weakness or lack of power with cunning and subversive humor. Common sense is disrupted, the unexpected evoked in order to enhance the awareness that reality is socially and culturally constituted. This relativist perspective may be essential for the cross-cultural study of humor, but the deep contextualization needed to understand humor gets in the way of anthropology's comparative perspective (Driessen, 1997).

Despite Malinowski's claim and anthropology's enormous archive of cultural variation, until the mid-1980s the vast majority of anthropologists neglected humor as a research

topic. So did sociologists (Zijderveld, 1983, 2011; Kuipers, 2006). Some early anthropologists studied humor from a folklorist and antiquarian perspective, but rarely as a resource for cultural analysis. In some fieldwork guides, humor is not even mentioned in spite of the fact that it is a finely calibrated barometer of the preoccupations shared by members of a society or group. Without understanding humor and laughter, one may never get to the core of a culture (Brandes, 1980). Radcliffe-Brown's article (1940) was a breakthrough, although his interest was in the structural implications and functions of joking relationships rather than their symbolic content. Mary Douglas found the essence of the joke in its attack on control. In the 1970s and early 1980s, anthropologists used semiotic and narrative approaches to study joking per se, not just its functions (Handelman and Kapferer, 1972). The first comprehensive, comparative anthropological study of humor and laughter appeared in the mid-1980s (Apte, 1985), and this event marked a considerable increase in case studies on humor in articles and books. Yet, humor remains for the most part a spin-off of other research topics, despite a growing awareness among anthropologists that it provides clues to what really matters in groups and societies. A good example is the intensification of political humor just before and throughout the so-called Arab Spring in 2011. The widespread use of humor in cartoons, comedy, graffiti, and raps expressed subversion and acted as a powerful vehicle for political activism. While the Arab world, in particular Egypt, has a long history of lampooning leaders, relatively new humorous genres such as rapping, stand-up comedy, and joking in blogs proliferated thanks to the new social media. Another striking example is the Italian comedian and agitator Beppe Grillo. This political buffoon has become a serious political force with his protest movement 'Five Stars.' It caused a landslide in the municipal elections of early 2012.

Anthropologists and comics share a relativist perspective. It is no surprise that the anthropological essay (Miner, 1956) and the ethnographic documentary have been parodied both within and without anthropology. A recent example on YouTube is 'Fest des Huhnes' (Festival of the Chicken) in which an African anthropologist documents the 'strange' customs of a mountain 'tribe' in Austria. Horace Miner's famous essay on body ritual among the Nacirema is a classic in teaching anthropology. Grant McCall's (1971) parody of British society and structural analysis is another example. In a brief article, taking 'Decimal Day' when the United Kingdom decimalized its currency as a starting point, McCall uses Lévi-Strauss, Dumont, and Srinivas in his burlesque of the British estate system, beverages, and other 'primitive classifications based on the four-part scheme.' A widely known case of self-mockery is the bestseller *The Innocent Anthropologist* by museum anthropologist Nigel Barley (1983).

The Many Faces of Humor

Genres

In most societies and cultures, there is a distinction between the diffuse joking of people as part of daily sociability and the institutionalized joking by virtuosi at specific times and places, such as the clown in wedding ceremonies in Polynesia, the television comedian on New Year's Eve, the trickster in a Javanese *wayang* play, or the regular stand-up comedy acts in New York, Amsterdam, or Sydney cafés. Jokes and pranks may provide the main fabric of which men are bound to each other on a daily basis as in Andalusian agrotowns (Brandes, 1980; Driessen, 1983).

Humor ranges from mild jokes to biting satire. A first type is benign humor that is simple, informal, and harmless, intended for diversion and repose. It does not unsettle the course of daily life. In metropolitan Egypt, for instance, during the regimes of Nasser, Sadat, and Mubarak, joking was the lubricant of social life. Persons who lacked a sense of humor 'had thick blood' and were scorned. A second type of joking is wit, banter, persiflage, or punning. This is a game of language and intellect, using irony to tackle discrepancies between appearance and reality. A third type is black or gallows humor, which challenges and mocks tragedy and misery, for instance, the 'Darwin Awards' on the Internet. A fourth form is satire, militant, aggressive, or hateful joking, meant to criticize authority, common sense, and conventions, as in Carnival around the world. The trickster of the Winnebagoo, Pueblo, and other Native Americans is a quintessential satirist, as were the Eurasian court jokers. The political cartoonist is a modern example of a satirist (Jansen, 2009) as is the stand-up comedian in metropolitan society. Finally, obscene and sick humor are categories found in many different societies. A recent example from the Netherlands in 2012 are so-called April Fool's jokes in which a suspicious parcel was placed at a synagogue in a rural town and a 'sea mine' dropped into a city canal. After massive interventions by the fire brigade and security forces it turned out that both objects were fake and part of a First of April 'joke.'

Domains of Joking

Worldwide, topics of humor deal with the boundaries between nature and culture, in particular with sex, food, health, and death as well as with social boundaries (us and them, strangers and lunatics). It has often been noted that it is hard to translate a joke from one language to another, especially in the case of puns and witticisms. Language is both a topic and vehicle of humor. Linguistic play and rhetorical devices, such as incongruity, paradox, and irony, are central to puns, tall tales, and riddles. Anthropologist Napoleon Chagnon, struggling to learn Yanomamö, was the victim of verbal humor. Informants made a sport of inventing obscene and ridiculous names for people such as 'dirty rectum' or 'feces of the harpy eagle' which Chagnon wrote down and used without knowing their meaning (Driessen, 1997). Apaches stage joking imitations of the 'whiteman' with exaggerated, loud, and fast speaking (Basso, 1979). In many societies, for instance, the Inuit, verbal virtuosity is expressed in duels in which humorous and obscene insults are recited to ridicule opponents. In Egypt, these verbal mock fights are ritualized in male coffee houses, for instance, "Your father puts on his underpants with a shoehorn"; "when your father sees a garbage truck coming, he lies down like a heap."

There is also a strong interdependence between gender and joking (Herzog, 1999; Crawford, 2003). Sexuality is a core subject of joking everywhere and humor is an important

element in attraction among the sexes. In most societies, there is gender inequality in joking. While men tend to use more aggressive humor and laugh louder in public spaces, women tend to remain more passive and restrained. Men play more prominent roles in ritual public humor. There are exceptions. Older women may publicly engage in humor and even compete with men. In Pacific island communities, female clowns play important ceremonial roles (Hereniko, 1995). In many Mediterranean societies, women perform for women mock imitations of men, their sexuality included. In Western society, professional female comedians are recently gaining prominence, an obvious indication of female emancipation.

Age is important in joking. Psychologists have noted the crucial role of humor in personality development and language learning. It serves as a general educational device. Children and adolescents have a special sense of humor with which they mock adults. Adults often find amusement in encouraging children to imitate adult etiquette. There are striking gender differences in the nature, devices, and approval of children's jocular behavior.

A widespread kin-related form of institutionalized humor is the joking relationship. Some categories of kin are licensed, sometimes even obliged, to tease, ridicule, and insult persons and groups belonging to another category, who must accept it without taking offense. In many societies, kin of the opposite sex who are potential sexual partners are involved in joking relationships. In Western society, mild joking is allowed with grandparents. Outside the kinship domain, reciprocal joking relations may occur, for instance, on the work floor or in hospitals and nursing homes. Over the past few years, the author observed many cases of reciprocal joking between nurses and elderly clients in a Dutch rest home. One particular nurse was walking the thin line between joking and humiliation when she asked a lady of 87 who regularly received brief social visits by a male neighbor of 79 whether she had thought of putting a condom on her bedside table.

Where governments fear freedom of expression and rigidly demand conformity, political humor is repressed yet flourishes, as is the case in most dictatorships. 'Joking up' is a weapon of the subordinated, whereas 'joking down' may act as a means of the powerful to express their superiority. Among near-equals, humor may play a role in what Freud called the narcissism of minor differences.

One of the main domains of joking in the globalizing world is ethnicity. Ethnic jokes make fun of supposed and farcically exaggerated effects in certain ethnic groups and nations. The deficiencies appear in pairs, e.g., the stupid and canny, the stupid and cowardly, and militarism and cowardice. Such dualities relate to commonly experienced ambiguous and contradictory situations (Davies, 1990). The telling of jokes may be, but not necessarily, an expression of racism and discrimination. Ethnic and many other jokes have a complex relationship with everyday experience.

Religion is replete with ironies and paradoxes. Consequently, joking, laughing Gods, and holy fools are to be found in the beliefs and rituals of many religions, local and global. There are counter-virtuosi who disturb and suspend the rigid ritual order by burlesque, inversion of roles, and sexual and scatological humor. Berger (1997) even argued for a 'theology of the comic' in which the experience of the comic is a promise of redemption.

Joking is not only a feature of face-to-face communication but also a major element in newspapers and magazines, movies, radio, television, and the Internet. Comic books have become a popular art form. There is musical humor and humor in novels and poetry, increasingly taken seriously by scholars of the humanities. Humor is also an effective device in advertising. In sum, humor permeates all domains of (semi-) public and (semi-) private life.

Functions and Meanings

Given the Protean nature of humor and its widely diverse manifestations, it is not surprising that there is no universal structure of meaning to humor. Its content largely relates to commonly experienced ambiguities, dilemmas, and paradoxes that are part of the human condition. Humor is one of the ways to deal with such experiences, and this is the universal meaning of humor. Much humor involves entertainment. It produces pleasure and smoothes social interaction. One of the most widespread popular psychological and functionalist theories has it that humor plays an important role in conflict resolution by releasing tension. Related theories hold that it is a tool of cultural critique and social control by highlighting taboos and key values. Another social function of humor and joking is that they mark group boundaries and help to promote a sense of belonging both in face-to-face and digital communities, sometimes by testing and challenging social cohesion. Humor and its different genres can also effect greater commitment, regulate embarrassing behavior, and sustain serious interaction. And they are relativizers, creating in a playful manner distance from turmoil, misfortune, and misery.

All these functions are present in the humor and laughter of young marginalized men in a city of Northern Morocco (Juntunen, 2002). In this context, humor and laughter serve as a relief-providing vehicle for young men with limited access to economic and political resources. Besides providing relief, joking among these young men also has the function of identity and solidarity building. The following joke popular among these men shows how the backwardness of the peasant turns into passive resistance. A man from the Ministry of Agriculture went on a tour of inspection in the countryside. He came across a peasant with a flock of sheep and asked him: Where do the sheep get their water? The black ones or the white ones? The black ones. There is a well over there. And the white ones then? The white ones as well. And where do they graze? The black ones or the white ones? The black ones. Up over there. And the white ones? The white ones as well. And how much do they pay for the wool here? Of the black ones or the white ones? The black ones. Twenty dirhams a kilo. And the white ones? The white ones as well. The civil servant has been writing everything in his official file but now had started to lose his temper. He asked sharply: Why do you keep on asking 'the black ones or white ones?' Because the black ones are mine. And the white ones? The white ones as well.

Trends in humor research show an increased interest in the healing power of humor and laughter in cross-cultural communication, psychotherapy (laughing sessions), and

clinical research. Cliniclowns in child hospitals seem to have relaxing and comforting effects on seriously ill children. Ethnic humor is a topic of growing relevance as are differences in joking styles and tastes. Recent issues of the journal *Humor* include articles on the fear of being laughed at, Chinese humor, and humor in poetry and politics.

See also: Comparative Method in Anthropology; Cultural Critique, Anthropological; Cultural Relativism, Anthropology of; Ethnicity: Anthropological Aspects; Ethnography in Applied Social Research; Gender: Anthropological Aspects; Kinship in Anthropology; Play, Anthropology of; Popular Culture; Trickster Ethnography.

Bibliography

Apte, M.L., 1985. Humor and Laughter: An Anthropological Approach. Cornell University Press, Ithaca.

Barley, N., 1983. The Innocent Anthropologist: Notes from a Mud Hut. British Museum Publications, London.

Basso, K.H., 1979. Portraits of 'the Whiteman'. Linguistic Play and Cultural Symbols among the Western Apache. Cambridge University Press, Cambridge.

Berger, P.L., 1997. Redeeming Laughter. The Comic Dimension of Human Experience. Walter de Gruyter, New York.

Brandes, S., 1980. Metaphors of Masculinity: Sex and Status in Andalusian Folklore. University of Pennsylvania Press, Philadelphia.

Crawford, M., 2003. Gender and humor in social context. Journal of Pragmatics 35, 1413–1430.

Davies, C., 1990. Ethnic Humor Around the World. A Comparative Analysis. Indiana University Press, Bloomington.

Douglas, M., 1968. The social control of cognition. Some factors in joke perception. Man (N.S.) 3, 361–376.

Driessen, H., 1983. Male sociability and rituals of masculinity in rural Andalusia. Anthropological Quarterly 56, 125–133.

Driessen, H., 1997. Humor, laughter and the field: reflections from anthropology. In: Bremmer, J., Roodenburg, H. (Eds.), A Cultural History of Humor. Polity Press, Oxford.

Handelman, D., Kapferer, B., 1972. Forms of joking activity: a comparative approach. The American Anthropologist 74, 484–518.

Hereniko, V., 1995. Woven Gods, Female Clowns and Power in Rotuma. University of Hawaii Press, Honolulu.

Herzog, T., 1999. Gender differences in humor appreciation revisited. Humor 12, 411–423.

Jansen, W., 2009. What Happened to Zina Diali? Gender in Algerian Cartoons. The Politics of Dissent in North Africa. Yale University, New Haven.

Juntunen, M., 2002. Between Morocco and Spain. Men, Migrant Smuggling and a Dispersed Moroccan Community. Institute for Asian and African Studies, Helsinki.

Kuipers, G., 2006. Good Humor, Bad Taste: A Sociology of the Joke. Mouton de Gruyter, Berlin.

McCall, G., 1971. An anthropology of decimalisation. New Society 437, 224–226.

Meijl, T. van, 2005. The critical ethnographer as a trickster. Anthropological Forum 15, 235–245.

Miner, H., 1956. Body ritual among the Nacirema. American Anthropologist 58, 503–513.

Nilsen, D.L.F., 1993. Humor Scholarship. A Research Bibliography. Greenwood Press, Westport.

Radcliffe-Brown, A.R., 1940. On joking relationships. Africa 13, 195–210.

Zijderveld, A.C., 1983. The sociology of humor and laughter. Current Sociology 31, 1–103.

Zijderveld, A.C., 2011. Waarom wij lachen: over de grap, de spot en de oorsprong van humor. Cossee, Amsterdam.

Hunger and Eating, Neural Basis of

Neil E Rowland, University of Florida, Gainesville, FL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E.M. Stricker, volume 10, pp. 7074–7078, © 2001, Elsevier Ltd.

Abstract

This article summarizes our current understanding of the theoretical bases of eating behavior, and the associated neural structures and signal molecules. In particular, the interaction between brain stem and forebrain regions subserving energy regulation is outlined, as well as the essential integration of these systems with those involved in rewarding aspects of food intake.

Energy transfer is the essential foundation of life. All biological processes require energy to proceed, and they do so by a series of biochemical reactions or transformations. Animals continually lose energy in the form of heat loss to the environment and to produce muscular activity for the heart, diaphragm, and skeletal motion. To fuel these processes requires that we have a continuous supply of suitable energy. That supply is food, which we, both, store and use as biological energy. This energy is most commonly expressed in terms of kilocalories (kcal) or kilojoules (kJ), although the kilo prefix is sometimes omitted.

History, Homeostasis, and Allostasis

Early theorists realized that survival is a question of making energy consumed equal to energy expended. Homeostasis is a concept, advanced by Walter Cannon (1929), which is consistent with this premise. Basically, it expanded on an earlier concept by Claude Bernard that relative constancy of the internal milieu – in this case energy content or flow – was essential for life. Cannon argued that dedicated physiological regulatory mechanisms were responsible and necessary for homeostasis.

Dual Center Theory

Recognizing that energy input via eating was essential, Eliot Stellar (1954) proposed a dual center model for feeding in which one region of the brain (the ventromedial hypothalamus) was considered a satiety center and an adjacent region (the lateral hypothalamus), a hunger center. These two regions each receive neural and hormonal inputs relevant to a variety of aspects of food, but ultimately the balance of neural activity between these two structures was thought to determine whether eating would occur (for example, due to hunger) or would not occur (for example, due to satiety). Today, some 60 years later, our knowledge of the brain structures, cells, and genes involved has proliferated greatly, yet the theoretical underpinning of much research is still fundamentally based on these opposing or reciprocal concepts of excitation of feeding (hunger) and inhibition of feeding (satiety). Another type of proliferation has taken place during these 60 years – our national waistlines as an obesity epidemic – because at a population and individual-level food intake has systematically exceeded expenditures. There is nothing constant about that fact, so the original concepts of balance or homeostasis are either inadequate or incomplete.

Allostasis

One of the problems with the simplest models of homeostasis is that they are reactive: when something drifts out of range, it is corrected. Yet many examples suggested that organisms were able to predict events, and several embellishments were made to the basic model. Most recently, the concept of allostasis has been promoted (Sterling, 2012). To simplify greatly, the concept of allostasis means that organisms do not regulate for constancy, but rather change the levels of regulation depending upon the prevailing environmental conditions. Another problem with classical homeostasis is that it requires some internal representation of the ideal: a set point, and it has proven difficult to demonstrate such an entity. In contrast, allostasis does not require a set point, simply an adaptive system with multiple inputs hard wired in such a way that the net output will be predictable for a given set of inputs. Applied to body weight which, as will be discussed is the output of multiple systems, allostasis is closely related to the idea of settling point (Wirtshafter and Davis, 1977).

Environmental Influences on Eating

Perhaps the conscious sensations that most affect our eating behavior are the chemical senses of smell and taste, and their synthesis as flavor. In an evolutionary context, these senses were important to locate food and evaluate its safety. Today, finding safe food in a modern society is not usually a problem: indeed, our brain is poorly prepared for the cornucopia of alimentary sensation with which it is bombarded. I will mention four aspects of taste/flavor that affect consumption: quality, intensity, variety, and economics.

Taste Quality

Foods or fluids that are sweet, via a family of so-called sweet receptors on the tongue and palate, are most readily accepted and ingested. Substances that are bitter generally are rejected. Fat-containing foods are also highly palatable, perhaps because

of their texture although recent work suggests there may also be fatty acid receptors on the tongue. Substances that combine sweet and fat are almost irresistible, like many desserts! Salty tastes also enhance the palatability of some commodities.

Taste Intensity

Intensity is related to the relative concentration of the tastant molecules in a sample, and their access to the relevant sensory receptors. Many elementary tastants show a preference-aversion function. When the concentration is very low, the taste is not or barely detectable (bland) and intake is low. As concentration increases, intake also increases up to a maximum beyond which it falls again – very high concentrations become less palatable. Sugar has a particularly long and pronounced preference side to this function.

Food Variety

Not only do real foods contain more than one tastant or odorant, we generally eat several different foods in a short time, or at least have a choice. The phenomenon of sensory-specific satiety (more accurately, satiation, see later) demonstrates that when people are given a succession of choices of flavors of even the same food (for example, yogurt) the amount consumed is considerably more than when only one flavor is presented. In its simplest form, this resembles habituation (to a given flavor) and dishabituation to the new flavor. The phenomenon of alliesthesia is related – the perceived palatability of a commodity declines as more of it is consumed.

Economics

Ecological foraging theories emphasize that the energetic effort expended to obtain food cannot, in the long term, exceed the energy yield from that food. In modern societies, effort is most usually in the form of time and monetary exchange. In particular, food that is immediately available at modest cost promotes 'opportunistic' eating – for example, vending machines or fast food (Rowland, 2012). Consumer economics describes the relationship between consumption and cost as a demand curve: not surprisingly, demand declines as price increases with the caveats that not all commodities show the same shape of demand curve and that price embodies aspects of time and money, as well as perceived social benefits. This cost–benefit aspect of the environment impacts the allostatic mechanism(s) mentioned previously.

Hunger, Satiation, and Satiety

Hunger

Hunger is the unpleasant sensation that motivates us to seek out and consume food. Homeostatic theories emphasized that hunger should be derived from internal deficiencies, such as an empty stomach of low availability of circulating metabolic fuels to critical sensors, presumably in the brain. However, hunger is not unconditionally correlated with such internal signals and people report hunger even when they are well-nourished. A concept of opportunistic feeding emphasizes the key role that the mere presence or prospect of food has on consumption. Our distant ancestors lived under continuous food insecurity – they did not know when the next meal would arrive and so the best survival strategy was to eat when food is available. Today, it may be that lower income individuals have the same perceived food insecurity and that may contribute to their overeating and relatively higher incidence of obesity. That does not deny the fact that a period of either acute or chronic lack of food (deprivation) will not increase hunger ratings and the motivation to seek food – they do – but deprivation is not the only trigger of hunger. Ghrelin is a hormone released from cells in the stomach when food is not present, and is thought to contribute to the sensation of hunger related to deprivation (Mason et al., 2014).

Satiation and Satiety

Satiation and satiety refer, respectively, to the early and late phases when hunger is absent after eating. Thus, satiation refers to processes that terminate an ongoing feeding episode while satiety refers to processes that suppress hunger for a more extended period of time. Once food has passed the oral cavity and reached the stomach, it initiates a temporally overlapping cascade of events that contribute to satiation and satiety. Food has bulk, and gastric distention is one such factor. The chemicals in the food, either before or after breakdown by digestive enzymes, stimulate specific receptors on enteroendocrine cells lining the gastrointestinal tract that lead among other things to release of hormones that either enter the blood stream and/or act in a paracrine fashion to stimulate local sensory nerve endings (Mayer, 2011). Either way, via neural and vascular pathways, these signals reach the brain. Some of these signals emanating from the upper parts of the tract, such as cholecystokinin, start signaling while eating is still going on and are candidate satiation hormones. Others, such as peptide YY are released from cells further down the intestine and so are not stimulated until food had been partly digested and propelled to that location by peristalsis, and are instead candidate satiety hormones. These neural and hormonal signals are thought to impinge on a common region of the hindbrain, the nucleus of the tractus solitarius (NTS) and/or the adjacent area postrema (Grill, 2010).

Neurobiology of Hunger and Satiety

Several key neural systems have been identified that participate, at a minimum, in the ebb and flow of hunger and satiety (Berthoud, 2011; Grill and Hayes, 2012). The most studied are systems that derive from cell bodies in the arcuate nucleus, which is situated medially at the base of the hypothalamus. Within the arcuate nucleus, there are two principal phenotypes of neurons. One type coexpresses genes for neuropeptide Y (NPY) and agouti-related peptide (AgRP). Administration of either NPY or AgRP into the brain of rodents causes them to hoard and eat food, and so these neurons are believed to be endogenously activated during states of hunger. A second type expresses the proopiomelanocortin (POMC) gene and, of several products from this gene transcript, the most relevant is α-melanocyte stimulating hormone (MSH). Administration of

α-MSH into the brain of rodents causes them to eat less, and so these neurons are believed to be normally activated during states of satiation or satiety. Alpha-MSH exerts its anorectic action primarily via type 3 and/or 4 melanocortin (MC3, MC4) receptors found in high abundance in the paraventricular nucleus (PVN) of the hypothalamus. Additionally, AgRP acts as an antagonist or blocker at MC4 receptors, so assuring a reciprocal action between these neuronal types. Further, both of these classes of neurons have receptors for blood-borne insulin and leptin, and so are influenced by these circulating hormones related to energy status. Recent evidence suggests that there are several functional subtypes within these two classes of neurons, including different parts of the brain to which they project. Thus, the 'hypothalamocentric' theories of food intake focus around the excitatory and inhibitory controls of these neurons, and whether they could eventually be manipulated for therapeutic purposes.

An additional consideration is the role of the hindbrain and in particular NTS. This is essentially the only region of the brain beside the PVN that expresses high levels of MC4 receptors, suggesting a prominent role for α-MSH in this region of the brain as well. There are bidirectional axonal connections between the NTS and the hypothalamus (and other forebrain regions) suggesting that descending projections from the hypothalamus can modulate the excitability of primary afferent information (taste, enteroendocrine) information ascending from the NTS to the rostral brain (Grill and Hayes, 2012).

Reward Systems

The systems described in the previous section are well-suited to homeostatic-like control of food intake, by controlling and modulating the flow of primary sensory information including endogenously secreted molecules. However, this does not address the issue of how or why taste and availability factors so profoundly influence total food intake and eventually body weight in an allostasis-like manner. Food is well known to act as a primary reinforcer of behavior, and a common reward system may be responsible for reinforcing a number of classes of behavior. This common substrate is a small population of neurons in the ventral tegmental area of the midbrain that express the transmitter dopamine and send axonal projections to the ventral striatum or nucleus accumbens and the frontal cortex. These neurons are potently activated by food and by cues related to the anticipation of food. There is some evidence that, rather than signal reward per se, these neurons encode differences between expected reward and that received (a prediction error). Additional signal molecules related to reward are endogenous opioids and endocannabinoids, both of which are known to affect food intake. Thus, activity within the reward system is thought to affect food intake, to some extent independent of the homeostasis-like systems described previously, although the systems must interact (Berthoud, 2011).

See also: Food Preferences, Psychology and Physiology of; Hypothalamus; Neurotransmitters; Reinforcement: Neurochemical Substrates; Taste and Smell, Psychology of.

Bibliography

Berthoud, H.-R., 2011. Metabolic and hedonic drives in the neural control of appetite: who is the boss? Current Opinion in Neurobiology 21, 888096.

Cannon, W.B., 1929. Organization for physiological homeostasis. Physiological Reviews 9, 399–431.

Grill, H.J., 2010. Leptin and the systems neuroscience of meal size control. Frontiers in Neuroendocriniology 31, 61–78.

Grill, H.J., Hayes, M.R., 2012. Hindbrain neurons as an essential hub in the neuroanatomically distributed control of energy balance. Cell Metabolism 16, 296–309.

Mason, B.L., Wang, Q., Zigman, J.M., 2014. The central nervous system sites mediating the orexigenic actions of ghrelin. Annual Review of Physiology 76, 519–533.

Mayer, E.A., 2011. Gut feelings: the emerging biology of gut-brain communication. Nature Reviews Neuroscience 12, 453–466.

Rowland, N.E., 2012. Order and disorder: temporal organization of eating. Behavioural Brain Research 231, 272–278.

Sterling, P., 2012. Allostasis: a model of predictive regulation. Physiology & Behavior 106, 5–15.

Stellar, E., 1954. The physiology of motivation. Psychological Review 61, 5–22.

Wirtshafter, D., Davis, J.D., 1977. Set points, settling points, and the control of body weight. Physiology & Behavior 19, 75–78.

Hunter-Gatherer Societies, Archaeology of

Vivian Scheinsohn, Instituto Nacional de Antropología y Pensamiento Latinoamericano (INAPL)/Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Universidad de Buenos Aires (UBA), Buenos Aires, Argentina

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H. Lourandos, volume 10, pp. 7078–7082, © 2001, Elsevier Ltd.

Abstract

Hunter-gatherer archaeology is the archaeological field of study that deals with hunter-gatherers, which are societies that procure most of their food from hunting, gathering, and fishing. Humans have been hunter-gatherers for most of their existence. Until the first food-producing societies developed, hunter-gatherers lived almost everywhere and this subsistence regime constituted the only way of living that humans had. So temporally, hunter-gatherer archaeology extends from the very first hunters documented in the prehistoric past until today. This article reviews the main characteristics of hunter-gatherer archaeology, the history of research, and current developments.

Introduction

At the heart of many anthropological theoretical frameworks is the assumption that we as humans are fundamentally hunter-gatherers. By the 1950s, paleoanthropologist Raymond Dart considered that hunting and a lust for meat "… drew the man-apes out of the forest and was a main force in human evolution" (Sussman, 1998: 121). Dart defined the basis of what was called the "hunting hypothesis" (Cartmill, 1997: 511), and later the writer Robert Adrey (1976) popularized it. Although the hunting hypothesis has since been disproved, when it was established that our hominin ancestors were mostly scavengers (Binford, 1981; Shipman, 1983), this view of the pivotal role of hunting and gathering lifeways in human evolution has had considerable traction in popular literature (e.g., Stewart, 1969). The hunter hypothesis is still evident in social scientific thought as well. For instance, a textbook on Evolutionary Psychology proposes that "… the mind is a set of information-processing machines that were designed by natural selection to solve adaptive problems faced by our hunter-gatherer ancestors" (Cosmides and Tooby, 1997: 1). Therefore, since the role of hunting and gathering in our history and our present continues to be debated, it should come as no surprise that hunter-gatherers have been a highly visible subject in anthropology and archaeology since the 1960s.

Hunter-Gatherer Archaeology

Hunter-gatherer archaeology is the archaeological field of study that deals with hunter-gatherers or foragers (often used interchangeably, i.e., Kelly, 1995: xiv, although not without controversy). Hunter-gatherers are "those who do or did procure most or all of their food from hunting, gathering and fishing" (Kelly, 1995: 3). But, as always in human matters, nothing is that simple. Kelly himself states that hunter-gatherers "do grow some of their own food, trade with agriculturalist for produce or participate in cash economies" (Kelly, 1995: 3).

Humans have been hunter-gatherers for most of their existence. Up to 12 000 years ago, when the first food-producing societies developed (Lee and Daly, 1999), hunter-gatherers inhabited nearly all environments globally and constituted the only lifeway available to humans. Thus, spatially hunter-gatherers are "archaeologically documented in environments ranging from tropical rain forests to the Arctic – and nearly everywhere archaeologists have looked in between" (Johnson, 2014: 18). Temporally, hunter-gatherer archaeology could be extended to the present day because foragers still persist in certain parts of the world. Even in the modern cities, the homeless could be considered as foragers acting in an urban environment (from an activist perspective, see Zimmerman et al., 2010). As Zimmerman et al. (2010: 444) observe, "There can be archaeology of ten minutes ago as easily as there can be archaeology of 10 000 years ago," which is a valid statement for hunter-gatherer archaeology more generally.

Main Characteristics

The archaeology of hunter-gatherers is often very different from the kind of archaeology seen in the media. Far away from castles, marble monuments, and splendorous graves, included in the field of classical or urban archaeology, it can frequently be very basic and straightforward.

The main features of hunter-gatherer archaeology are:

1. An interest in the recovery of any minute item related to human behavior: Because hunter-gatherers are mostly nomadic (although sedentary lifestyle is recorded in some hunters-gatherers, mostly exploiting aquatic resources, as in the Northwest Coast of North America, see Ames, 1985), they have fewer material possessions than, for instance, sedentary agriculturists. Their objects are more transportable, namely, more compact, smaller, and less fragile than the possessions of sedentary people. Also, in contrast to the villages and cities of sedentary groups, hunter-gatherers have sparser population distributions and correlatively they create more archaeological sites although with lower artifact densities. Finally, one must consider the durability of the materials used by hunter-gatherers. Many materials used by hunter-gatherers were perishable and their preservation depends on the particular conditions of deposition

or their antiquity, with the noteworthy exception of lithic materials (see below). In this sense, hunter-gatherer archaeologists are focused even in the smallest details in their research, in a context of relatively artifact scarcity (when compared to a village or urban site).

2. A strong focus on technological issues, especially those related with lithic materials: The poor preservation of many materials and the excellent preservation of stone is the reason that much hunter-gatherer research deals mainly with lithic technology. When dealing with the archaeology of the earliest humans or human ancestors, usually only lithic artifacts are recovered. In more recent times, new technologies emerged and more materials survive to be recovered by archaeologists (as worked bone, ceramics, metals, etc.). However, all hominins and humans must have used other materials that simply were not preserved. This fact poses a major challenge to hunter-gatherer archaeology since, as the motto goes, the absence of evidence it is not evidence of the absence. For instance, the lack of wooden artifacts does not mean that wood was not exploited: It could simply not have been preserved. Therefore, although stone tools are used in other contexts, such as urban sites, lithic materials are the bulk of what archaeologists analyze in hunter-gatherer sites.
3. An interest in human mobility patterns and circulation: The fact that hunter-gatherers are mostly nomadic or mobile has naturally led to considerable research in this area (see below).
4. An emphasis on understanding subsistence patterns: Many of the archaeological remains found at hunter-gatherer sites were associated with food procurement.
5. Less interest in the symbolic aspects and belief systems arguably due to problems of interpretation, given the lack of context or continuity to interpret symbols present in rock art or decorated artifacts. But also, as the theoretical framework mainly used in hunter-gatherer has been processualism (see below), this lack of interest in symbolic issues was linked to a preference for materialist perspectives (see Archaeology, Theory in). However, although less often employed than in other areas of archaeological focus, interpretive theoretical frameworks have recently developed in hunter-gatherer studies (see below).

Anthropology, Archaeology, and Hunter-Gatherers

As Bettinger (1987: 121) has noted "there is little that separates archaeologist from cultural anthropologist where hunter-gatherers are concerned and many scholars move back and forth between the study of living and extinct hunter-gatherers without perceptible difficulty." This is because archaeologists have utilized ethnographic information to develop analogies (i.e., Cummings, 2013) or provide a frame of reference (Binford, 2001) for research in different times and from many theoretical standpoints (see below). Hunter-gatherer studies are today at the main intersection between anthropology and archaeology. In that intersection, anthropology contributes the theoretical standpoint and the linking between living context and archaeological context whereas archaeology has contributed 'deep' time and a variety of behaviors that anthropology cannot record. But, hunter-gatherer anthropological knowledge was created in a colonial or postcolonial context (Cummings, 2013) whereas hunter-gatherer archaeology has mostly focused in hunter-gatherers before they occupied marginalized geographical and social spaces. Archaeology is thus the only discipline that allows for the study of hunter-gatherers before colonialism. One of the great strengths of archaeology is its ability to provide examples of human behavior that are not recorded in the present or the historical record, but that nonetheless characterized the human experience at some point in our species' history. The present therefore can be useful in studying the past (as it is in Middle Range Theory, for instance; see below), but it should not limit our interpretations of past hunter-gatherer societies.

Bettinger identifies two perspectives on hunter-gatherers studies in the early and mid-twentieth century. The first characterized hunter-gatherers as primitives and remnants of the Stone Age (i.e., Murdock, 1934 and its suggestive title) which "emphasized the things (hunter-gatherers) lacked that more advanced cultures would later have" (Bettinger, 1987: 122). This model endured until the advent of the second in 1966, when the "Man the Hunter" conference was held in Chicago. Organized by Lee and DeVore (1968), this symposium brought together researchers from all over the world to develop a new model of hunter-gatherer societies. Dubbed the "generalized foraging model" (Isaac, 1990), this new model relied heavily on anthropological theories derived from the New Ecology and Neofunctionalism (Bettinger, 1987).

The generalized foraging model considered five characteristics: (1) egalitarianism, (2) low population density, (3) lack of territoriality, (4) a minimum of food storage, and (5) band as a preferred form of social organization (Kelly, 1995). As Kelly (1995) has pointed out, the most critical part of this model was Marshall Sahlins' (1968) formulation of the "original affluent society," which generated a picture of hunter-gatherers as "spending relatively little time working, (having) all the food they needed and (spending) leisure hours sleeping or socializing" (Kelly, 1995: 16). Thus, from this perspective agriculture was not a breakthrough but rather "a lifeway adopted only under dire circumstances" (Kelly, 1995: 18). As an answer to *Man the Hunter*, the volume *Women the Gatherer* (Dahlberg, 1981) challenged the gender bias implied by both the title and content of the former by stressing the role of women in hunter-gatherer societies and their importance as providing critical resources to the diet. Although this book was important in terms of establish gender bias operating in hunter-gatherer studies, it suffered from the same fault that *Man the Hunter*: hid hunter-gatherer variability (Wobst, 2011) in order to emphasize similarities (Kent, 1996).

The movement to critique this homogenizing perspective of hunter-gatherers began almost simultaneously with the work of Wilmsen (1989) and Headland and Reid (1989) in different parts of the world. They suggested a different view, that of hunter-gatherers as dependent on trade with food-producing neighbors as well as their own part-time cultivation or pastoralism. This 'revisionist critique' has also generated reactions (see Shott, 1992 for a review) but slowly it began to gain adherents leading to a wide diversification in hunter-gatherer studies.

An Archaeological Perspective: Hunter-Gatherers and Processualism

In the 1960s neofunctionalism took the lead in anthropology. Neofunctionalism was derived from Systems Theory, and its core concepts were incorporated into the emergent New Archaeology or Processualism (see Archaeology, Theory in). Starting as a critique of the so-called normative theory (which asserted that artifacts reflected cultural norms), which was advocated by archaeologists and anthropologists employing a culture history theoretical perspective, Processual Archaeologists advocated for a scientific archaeology following the Hempelian deductive-nomological model of explanation and the Hempelian hypothesis formation/testing model (O'Brien et al., 2005: 110). Processualism also criticized culture history's focus on homogeneity and instead began to consider variability within ancient societies. With the advent of Processualism, the scene was set for a new perspective on hunters-gatherers. In fact, most of the research by one of the main advocates of Processualism, Lewis Binford, was devoted to hunter-gatherers.

As Binford stated, processualist archaeologists needed to construct Middle Range Theory as a mean by which the static archaeological context was connected with the dynamic and systemic context (behavior) that created it. In order to construct Middle-Range theories, Processualism explored actualistic studies (including experimental archaeology, ethnoarchaeology, and taphonomy). Among them, ethnoarchaeology gave a new impulse to the archaeological study of modern hunter-gatherers (see Ethnoarchaeology). A good example of an archaeological processualist perspective on hunter-gatherers is the Forager-Collector continuum model posited by Lewis Binford (1978, 1980). Departing from his own ethnoarchaeological work, Binford identified two extremes on a spectrum of mobility patterns among hunter-gatherers: from foragers on one end to collectors on the other. Binford attributed residential mobility, or movement of an entire group from one camp to another, to the former and logistic mobility, or movement of a small group from a residential base camp to a place to perform a specific task, to the latter, with a continuum of possibilities in between. Binford argued that the kind of mobility was dependent on the kind of environment under study: Where access to resources is temporally or spatially limited, a variety of measures will be taken to extend the utility of those resources (Bettinger, 1987). In environments that are spatially and temporally homogeneous (without seasons of shortage), there is no need to store food in anticipation of future scarcity. Under these conditions, hunter-gatherers exploit the resources near their residential base and then move to an unexploited area. Thus, foragers move consumers to food resources and map onto resources (Bettinger, 1987). In environments where resources are spatially and temporally heterogeneous (with periods of shortage), incongruities between distribution of population and resources require some form of compensation. Minimally, hunter-gatherers have to implement storage (a practice associated with resources obtained in bulk over a short period). Thus collectors move residentially to key locations that are not necessarily related to food supply, and then use logistical travels by small groups to bring resources to the residential base. "The forager responds to temporal and spatial resource shortage by residential movement; the collector responds to temporal shortage by storage and to spatial mismatches between population and resources by logistical mobilization" (Bettinger, 1991: 70). After its proposal, this model was tested in many archaeological cases around the world and found to be an accurate predictor of hunter-gatherer mobility.

Alternative models signaled differences among hunter-gatherers in different terms. For instance, Woodburn (1980, 1982) proposed a distinction between delayed and immediate return hunter-gatherers where delayed return groups focused on long-term storage, which eventually leads to inequality, and immediate return groups focused on producing food for rapid consumption. Bettinger and Baumhoff (1982) developed a traveler-processor model which posited a continuum of hunter-gatherer strategies by specifying relationships between population and resources and settlement and subsistence patterns. They posited that as population increased from low (traveler situation) to high densities (the processor situation), diet breadth and processing cost would increase.

Finally, Processualism and later Evolutionary Ecology or Behavioral Ecology (Winterhalder and Smith, 2000; Kelly, 1995) favored optimal foraging models starting in the 1980s. Initially taken from biology, these studies provided "a yardstick of objective economic rationality as a basis for the comparative study of human behavior" without which "it would be impossible to know whether human behavior was governed by cultural norms and ideas or by rational, self-interested individuals" (Bettinger, 1987: 135). In anthropology, optimal foraging theory assumes that human decisions are made to maximize net rate of energy gain. Some of the models deal with mobility although others include different ways to evaluate energy cost and gains. In the case of diet or diet breadth models, optimal foraging theory is employed to evaluate how many different species hunter-gatherers should exploit while patch choice models examine how hunter-gatherers will select an environmental patch to exploit.

Hunter-Gatherer Archaeology Today

In his review paper in the late twentieth century, Shott (1992: 864) argued for a synthesis of the general and the particular, one that would combine "a legitimate concern for case-specific historical developments with a commitment to general theory and explication" and asserted that archaeology would have an important role to play in forging this synthesis. As Shott expected, the 1990s was indeed a stimulating decade in hunter-gatherer studies. New research capacities provided by computers and the Internet allowed for the elaboration of a series of 'global' studies on modern hunter-gatherers aimed at elucidating general patterns (Cashdan, 2001; Henrich et al., 2006; Hill et al., 2011 among others).

Processualist archaeologists also have been elaborating studies at a global scale (Binford, 2001; Johnson, 2014; Kelly, 1995; Marlowe, 2005). For instance, using a Behavioral Ecology perspective, Kelly (1995) reviewed a series of hunter-gatherer groups focusing in subsistence, mobility, exchange, group size, and other characteristics to analyze hunter-gatherer

variability. Kelly's has been a preferred textbook in many archaeological courses around the world and it was highly influential. Nevertheless his database pales in comparison to Binford's (2001), in which by using environmental and ethnographical data he constructs generalizations and propositions "developed from ethnographic patterning among hunter-gatherers as a tool for archaeological research" (Johnson, 2014: 7).

From Binford's and Kelly's works, certain general patterns on modern hunter-gatherers were established. For instance, that hunter-gatherers' dependence on plants for subsistence decreases as latitude increases (at low latitudes hunter-gatherers are often primarily dependent on plant foods whereas at high latitudes, low population density hunter-gatherers are primarily dependent on animal foods) and that variation in subsistence is conditioned by effective temperature (a measure of growing season), population density, and availability of aquatic resources (Binford, 2001), allowing higher population densities along productive coasts and large rivers (Binford, 2001; Kelly, 1995). Also it was proposed that foraging groups in productive environments travel shorter distances and have smaller home ranges and higher population densities than societies in less productive environments (Kelly, 1995; Binford, 2001) as long as groups that rely more on hunting tend to have lower population densities, occupy larger areas, and move more frequently over greater distances (Kelly, 1995).

Postprocessualist archaeologists also have dealt with hunter-gatherers. As an example of this theoretical framework, Classen's (2012) work on Eastern US caves is noteworthy. Claasen draws on a body of evidence of 'cave rituals and ritual caves' established for Mesoamerica to elaborate a list of expectations about characteristics of ritual caves and offerings for hunter-gatherer groups. Her list for defining sacred caves includes: presence of flowing water, items of personal adornment left as offerings, items frequently burned, and/or broken, among others. Claasen then looked for these characteristics in a series of Eastern US caves occupied during the Archaic period. She reviewed the characteristics of nonhuman bone, mollusks, hematite, rock features, and stone tools recorded in those caves to finally conclude that they were utilized for ritual activities. She actually proposes that "Caves are so important in historic accounts of origins and homes of deities and culture heroes that rather than defaulting to camping to explain the material culture found in caves, the default explanation should become ritual" (Classen, 2012: 259–260). In terms of frameworks, today, most hunter-gatherer archaeologists are engaging a personal theoretical mix which includes Processualist and Postprocessualist approaches. The paradigmatic case in anthropology should be Tim Ingold's works (discussed below).

Biases

While promising, this state of the art still presents certain biases. One of the most pervasive is that most studies of modern hunter-gatherers are devised to model our ancestor's behavior (i.e., Marlowe, 2005; Ströhle et al., 2010; King, 2013). In addition, a great deal of hunter-gatherer archaeology is directed to early hominins or the 'peopling' of certain regions. In the Americas, this is manifest in the large body of work focusing on the Pleistocene–Holocene transition. Late Holocene hunter-gatherers, those who lived after the Ice Age, receive far less attention in international publishing despite an abundance of examples and the good preservation of sites from these times. For instance, and just as an example, on examination of all issues of the journal *Before Farming* (from the first issue in 2003 to the last in 2012), which specializes in hunter-gatherers, almost 30% of the papers published pertained to Paleolithic Europe and the Levant, followed by modern and prehistoric South Africa (15%), and modern and prehistoric Australia (11%). There are also spatial biases: The case of South American hunter-gatherers is illuminating. Kelly's (1995) database contains only 8 South Americans groups (5 of them from Amazonia) versus 50 North American groups (Kelly, 1995; Figure 1.1, 4–5). And in a recent book specifically devoted to South American archaeology (Silverman and Isbell, 2008), only 9 out of 51 articles referred to hunter-gatherers.

Conclusion

As Johnson (2014: 2) notes "Perhaps the most significant change in hunter-gatherer studies over the last several decades has been the shift from expecting hunter-gatherers to have similar properties (subsistence, settlement pattern, technology, social organization) wherever they are found to recognizing that hunter-gatherer adaptations should vary (often widely) along many different dimensions in different environments and under different levels of population density and regional organization diversity." After many years of "evolutionary paradigms dominating the discourse on hunter-gatherer variability and change" (Sassaman and Holly, 2011), the variability we expect to see among hunter-gatherers has begun to permeate the ways we construct our knowledge about them. Ingold's (2000) works were very influential in archaeology for expanding our view of hunter-gatherer diversity. His main critique is directed at attempts to apply models designed for the study of nonhuman foraging behavior to the analysis of human hunting and gathering. "This application results from a conflation of rational choice theory, drawn from classical microeconomics, with the theory of natural selection, drawn from evolutionary biology… A properly ecological account of hunting and gathering requires however that we dissolve this opposition, showing how people develop their skills and sensitivities through histories of continuing involvement with human and non-human constituents of their environments" (Ingold, 2000: 10). Also postprocessualists and interpretive archaeologists have begun to tackle hunter-gatherers (i.e., see above and Cannon, 2011). That diversity is reflected in a recent book which presents a wide array of theoretical stances (Cummings et al., 2014). As the middle ground claimed by Shott is still in construction, creative decades are coming for hunter-gatherer archaeology. Probably, the territory located at the great divide between Processualism and Postprocessualism will be the most fruitful.

See also: Archaeology, Theory in; Arctic Archaeology and Prehistory; Australia and New Guinea, Archaeology of; Ethnoarchaeology; Europe, Archaeology of; Fishing Societies; Food Production, Origins of; Hunting and Gathering Societies: Anthropology; Lithics and Archaeology; Migrations, Colonizations, and Diasporas in Archaeology; North America, Archaeology of; Pacific Islands, Archaeology of; South America, Archaeology of.

Bibliography

Adrey, R., 1976. The Hunting Hypothesis. Bantam Books, New York.

Ames, K.M., 1985. Hierarchies, stress, and logistical strategies among hunter–gatherers in northwestern North America. In: Price, T.D., Brown, J.A. (Eds.), Prehistoric Hunter–Gatherers: The Emergence of Cultural Complexity. Academic Press, Orlando, pp. 155–180.

Bettinger, R., 1987. Archaeological approaches to hunter-gatherers. Annual Review of Anthropology 16 (1), 121–142.

Bettinger, R., 1991. Hunter-Gatherers: Archaeological and Evolutionary Theory. Springer, New York.

Bettinger, R., Baumhoff, M., 1982. The Numic spread: Great Basin cultures in competition. American Antiquity 47 (3), 485–503.

Binford, L., 1978. Nunamiut Ethnoarchaeology. Academic Press, New York.

Binford, L., 1980. Willow smoke and dogs' tails: hunter-gatherer settlement systems and archaeological site formation. American Antiquity 45, 4–20.

Binford, L., 1981. Bones: Ancient Men and Modern Myths. Academic Press, New York.

Binford, L., 2001. Constructing Frames of Reference. An Analytical Method for Archaeological Theory Building Using Hunter-Gatherer and Environmental Data Sets. University of California Press, Los Angeles.

Cannon, A., 2011. Structured Worlds: The Archaeology of Hunter Gatherer Thought and Action. Equinox, London.

Cartmill, M., 1997. Hunting hypothesis of human origins. In: Spencer, F. (Ed.), History of Physical Anthropology: An Encyclopedia. Garland, New York, pp. 508–512.

Cashdan, E., 2001. Ethnic diversity and its environmental determinants: effects of climate, pathogens, and habitat diversity. American Anthropologist 103 (4), 968–991.

Classen, C., 2012. Cave rituals and ritual caves in the Eastern United States. In: Sundstrom, L., DeBoer, W. (Eds.), Enduring Motives: The Archaeology of Tradition and Religion in Native America. University of Alabama Press, Tuscaloosa, pp. 253–263.

Cosmides, L., Tooby, J., 1997. Evolutionary Psychology: A Primer. http://www.cep.ucsb.edu/primer.html (accessed 01.12.14.).

Cummings, V., 2013. Can the study of modern hunter-gatherers helps us to understand the past? In: The Anthropology of Hunter-Gatherers: Key Themes for Archaeologists. Bloomsbury, London, pp. 1–12.

Cummings, V., Jordan, P., Zvelebil, N., 2014. The Oxford Handbook of the Archaeology and Anthropology of Hunter-Gatherers. Oxford University Press, Oxford.

Dahlberg, F., 1981. Woman the Gatherer. Yale University Press, Yale.

Headland, T., Reid, L., 1989. Hunter-gatherers and their neighbors from prehistory to the present. Current Anthropology 30, 43–66.

Henrich, J., McElreath, R., Barr, A., Ensimger, J., Barrett, C., Bolyanatz, A., Cardenas, J.C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., Ziker, J., 2006. Costly punishment across human societies. Science 312, 1767–1770.

Hill, K.R., Walker, R.S., Božičević, M., Eder, J., Headland, T., Hewlett, Hurtado, A., Marlowe, F., Wiessner, P., Wood, B., 2011. Co-residence patterns in hunter-gatherer societies show unique human social structure. Science 331 (6022), 1286–1289.

Ingold, T., 2000. The Perception of the Environment: Essays in Livelihood, Dwelling and Skill. Routledge, London.

Isaac, B., 1990. Economy, ecology, and analogy: the !Kung San and the generalized foraging model. In: Isaac, B. (Ed.), Early Paleoindian Economies of Eastern North America, Research in economic anthropology, vol. 5. JAI Press, Greenwich, pp. 323–335.

Johnson, A., 2014. Exploring adaptive variation among hunter-gatherers with Binford's frames of reference. Journal of Archaeological Research 22 (1), 1–42.

Kelly, R., 1995. The Foraging Spectrum: Diversity in Hunter-Gatherer Lifeways. Smithsonian Institution Press, Washington, DC.

Kent, S., 1996. Cultural Diversity among Twentieth-Century Foragers: An African Perspective. Cambridge University Press, Cambridge.

King, B., 2013. The modern obesity epidemic, ancestral hunter-gatherers, and the sensory/reward control of food intake. American Psychologist 68, 88–96.

Lee, R., DeVore, I., 1968. Man the Hunter. Aldine Publishing Co., Chicago.

Lee, R., Daly, R., 1999. The Cambridge Encyclopedia of Hunters and Gatherers. Cambridge University Press, Cambridge.

Marlowe, F., 2005. Hunter-gatherers and human evolution. Evolutionary Anthropology. Issues, News and Reviews 14 (2), 54–67.

Murdock, G., 1934. Our Primitive Contemporaries. Macmillan, Oxford.

O'Brien, M., Lyman, R., Schiffer, M., 2005. Archaeology as a Process: Processualism and Its Progeny. University of Utah Press, Salt Lake City.

Sahlins, M., 1968. Notes on the original affluent society. In: Lee, R., DeVore, I. (Eds.), Man the Hunter. Aldine Publishing Co., New York, pp. 85–89.

Sassaman, K., Holly Jr., D., 2011. Hunter-Gatherer Archaeology as Historical Process Amerind Studies in Archaeology. University of Arizona Press, Tucson.

Shipman, P., 1983. Early hominid lifestyle: hunting and gathering or foraging and scavenging. In: Cluton-Brock, J., Grigson, C. (Eds.), Animals and Archaeology: Hunters and Their Prey, vol. 1, pp. 31–49.

Silverman, H., Isbell, W., 2008. Handbook of South American Archaeology. Springer.

Shott, M., 1992. On recent trends in the anthropology of foragers: Kalahari revisionism and its archaeological implications. Man 843–871.

Stewart, G.R., 1969. Earth Abides. Houghton Mifflin, Boston.

Ströhle, A., Hahn, A., Sebastian, A., 2010. Latitude, local ecology, and hunter-gatherer dietary acid load: implications from evolutionary ecology. American Journal of Clinical Nutrition 92 (4), 940–945.

Sussman, R., 1998. The myth of man the hunter/man the killer and the evolution of human morality. In: Sussman, R. (Ed.), The Biological Basis of Human Behavior: A Critical Review. Prentice Hall, Upper Saddle River, pp. 121–129.

Wilmsen, E., 1989. Land Filled with Flies: A Political Economy of the Kalahari. The University of Chicago Press, Chicago.

Winterhalder, B., Smith, E.A., 2000. Analyzing adaptive strategies: human behavioral ecology at twenty-five. Evolutionary Anthropology Issues News and Reviews 9 (2), 51–72.

Wobst, M., 2011. Epilogue: foragers in hindsight or theory and method meet history. In: Sassaman, K., Holly, D. (Eds.), Hunter-Gatherer Archaeology as Historical Process Amerind Studies in Archaeology. University of Arizona Press, Tucson, pp. 248–258.

Woodburn, J., 1980. Hunters and gatherers today and reconstruction of the past. In: Gellner, A. (Ed.), Soviet and Western Anthropology. Duckworth, London, pp. 95–117.

Woodburn, J., 1982. Egalitarian societies. Man 17, 431–451.

Zimmerman, L., Singleton, C., Welch, J., 2010. Activism and creating a translational archaeology of homelessness. World Archaeology 42 (3), 443–454.

Hunting and Gathering Societies: Anthropology

Nurit Bird-David, University of Haifa, Haifa, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The perspectives taken on the contemporary peoples referred to by this term have changed considerably since its early uses in the eighteenth century. I describe its evolving meanings, and then several clusters of features partially shared by these peoples, including sharing, perceptions of the environment, animism, relational knowledge, and relational ontologies. Though their cases are rare, they remain pivotal in modern thought as our ultimate 'Other' and our ultimate 'Us.'

'Hunting and gathering' is a conventional term for subsisting on the 'natural produce' of the earth by hunting and trapping wild animals; gathering wild roots, fruits, and berries; fishing; and collecting honey of wild bees, as opposed to producing food by agriculture and herding. By 'hunting and gathering societies' (HGS) ethnographers mean contemporary indigenous people who fully or partly engage in these activities, now or formerly, or who see themselves as hunters. (For brevity, the terms 'hunter-gatherers,' 'hunters,' and 'foragers' are sometimes used synonymously.) The term is also used in general reference to prehistoric populations, living before the agricultural revolution (around 10 000 years BP). In common usage, it also stands for alien Other, diversely constructed by modernity in order to define itself. In this article, the primary focus is placed on the first meaning, while note is taken of its brush with the other two. The term is first examined historically, with an outline of its origins and changes of usage over time. Then, several clusters of features that are partially shared by these societies are presented in model terms.

Contested Meanings

In order to understand the expanding meanings of HGS it is essential to go back to the origin of the field and follow its historical development. In the seventeenth century, writers invoked 'primitive peoples,' who hunt and gather for their subsistence, in discussing the 'natural state' of humankind. By the eighteenth century, these were typed 'hunter-gatherers,' in distinction from 'herders,' within a broad 'savage' class, in writings on the origins of property relations. In the nineteenth century, evolutionist work continued to draw on specific peoples, often without referring to the class as a whole, in studies of the origins of a variety of other social institutions, from kinship through politics to religion.

The notion consolidated as a central term in early twentieth-century cultural ecology theory, most influentially in Julian Steward's work (1936). Concerned with the relation between society and environment, Steward argued that for any given mode of subsistence, successful adaptation determines a certain level of sociocultural integration. In the case of hunting-gathering, he argued that this is a 'band,' which universally is patrilineal, exogamous, and territorial. HGS were associated with bands, and hunting peoples were comparatively studied as band societies (see, for example, Damas, 1969), until the 1966 symposium 'Man the Hunter' (Lee and DeVore, 1968). This interdisciplinary symposium commonly is considered the beginning of modern hunter-gatherer studies.

The symposium was inspired by the assertion that human beings had subsisted by hunting and gathering for over 99% of the time of human life on earth. Linking cultural-ecological and (neo)behavioral-evolutionary perspectives, its major objective was to learn about human origins and nature by comparing 'living hunters.' Uniform patterns found among them, it was posited, can be regarded as hunter-gatherers' successful adaptations, in present and past.

Serious reservations were raised about its agenda. One objection, for example, was that no contemporary group could be seen as 'living relics,' or as direct descendants, of the Pleistocene. Even if a group had been found to subsist purely on hunting, its people would have had a history as long as any other, and they would not have lived any longer in a world populated by hunters as in the Pleistocene. Another objection was that over the span of 2 000 000 years humans had changed considerably, biologically, and psychologically, enough to render homogenizing them inappropriate.

The ethnographic comparison, at the same time, profoundly overturned accepted wisdom. Contra Steward's persistent thesis, research revealed that most bands actually are socially fluid and nomadic. In many cases, gathering makes no less important contribution to subsistence than hunting. Contra contemporary stereotypes and bourgeois intuition, the comparison showed that these peoples meet their needs with surprisingly little labor. (Marshall Sahlins provocatively dubbed them 'the original affluent society.')

Outweighing the reservations, this successful outcome entangled HGS with 'Man the Hunter's' project. HGS came to mean a type of people, rather than diverse peoples engaged in a type of action, encompassing populations over a vast evolutionary span of time. The term now embodied the hypothesis that, at their core, these populations have similar sociocultural forms, determined by adaptation to the environment. From a fuzzy term of reference to 'very simple' peoples, and then to 'band societies,' it connoted a firm, homogenous and exclusive social category, a category of humanity defined techno-economically.

Subsequent ethnographic research, which burgeoned in the wake of 'Man the Hunter,' grew diverse in focus and approach, as seen, for example, in the series of International Conferences on Hunting and Gathering Societies (1978 to date, e.g., Leacock

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12090-2

and Lee, 1980; Ingold et al., 1988[1986]: Vols. I, II; Burch and Ellana, 1994; Wenzel et al., 2000; Barnard, 2004). Issues arising from late twentieth-century indigenous circumstances and cultures were addressed, including political struggles and self-administration in modern states, and indigenous peoples' own perceptions of their environment, economy, and self. Historical, interpretative, regional, and applied perspectives commanded attention; evolutionary-ecological studies of foraging behavior increasingly took a separate course. Indigenous people started contributing to academic research, as anthropologists and interested parties. This growing ethnography showed a diversity of social forms among hunting-gathering peoples, from egalitarian to hierarchical and even slave-owning societies. The 1970s meaning of HGS as a firm, homogenous and exclusive social category was increasingly challenged. In an attempt to save it, the category was diversely divided into subtypes (e.g., see Woodburn, 1982). Or, it was recharacterized, for example, by a collective appropriation of nature (see Ingold, 1986), and by distinct cultural perceptions of the environment (see Bird-David, 1990, 1999; Ingold, 2000).

The 1980s and the early 1990s saw many controversies, sparked partially by inflexible readings of the term HGS. Scholars disputed the category as pertaining to a class of societies that are both singularized by hunting-gathering and distinguished from agricultural and pastoral peoples, because the ethnography revealed continuities between hunter-gatherers and other societies, as well as a great diversity among the hunter-gatherers themselves. Other scholars questioned the categorization of societies by mode of subsistence, as research showed heterogeneous and changing economies. Hunting-gathering was found to be articulated with other pursuits, going back in some cases centuries and even millennia, and, at any rate, under speeding changes since the mid-twentieth century. Students debated whether any living group sufficiently complies with the subsistence criterion and whether this criterion generates a meaningful category, one whose members share more than the selecting mode of subsistence. Even the field at large was critiqued as spurious, by researchers seeing it in terms of the 1960s objective to learn about prehistory and evolution. Their criticism drew on ethnohistorical and archeological work suggesting arguably that these societies have not escaped the world and regional forces.

The new millennium has seen a surge of renewed interest in hunter-gatherers' studies, and once again their discussions in interdisciplinary forums bringing together the four fields of anthropology. The diversity of HGS is addressed, accommodating within the category the assemblage of indigenous peoples discussed by practitioners within the field of hunter-gatherer studies (see Kelly, 1995). Attention is paid not only to their diversity but also – through such issues as their relations with others, the changes they undergo, and their socialities and ontologies – to both their resilience and vulnerability.

HGS, then, is a historically situated term, which has acquired multiple layers of meanings over the years. It is a circumlocution for a historically situated field of research that provides a stimulating comparative context and a scholarly communal framework within which to explore a broad array of distinctive issues. Ethnographers construct 'partial connections' (in Strathern's sense of the term, 1991) between these societies by studying their respective groups while keeping the others 'in mind.' They call to attention partially shared features, which determine common issues in the field, and contribute distinctive perspectives to anthropology, generally. Hunting and gathering activities are a focus of attention, rather than a criterion of demarcation, for scholars and for the indigenous peoples themselves.

Issues

This section presents, in model terms, some of the partially shared features that came to light in the course of ethnographers' comparative discussions of their field experiences and material.

Hunting and Gathering

In many of these societies, people combine these activities with other economic pursuits (e.g., trade in minor forest produce, horticulture, work for neighbors, wage labor, social benefits, etc.). However, hunting and gathering themselves carry special significance; they are a total phenomenon encompassing economic, social, and symbolic aspects. Hunting-gathering, moreover, often influences the way in which people carry out such other activities. They often return to hunt and gather at least periodically, when not prevented from doing so. They regard what they collect as 'real food,' and while they do pursue other activities, in some cases, hunting-gathering still provides a major part of their diet. Some postforagers return to their traditional lands in attempts to creatively combine traditional and modern lifeways.

Band Society

In many of these societies people live in small groups (described in early work as 'bands') consisting of several families. The groups are flexible and fluid. Making visits, and even shifting residence, from place to place, are common, including in new modern settlements, and highly valued. Men and women often know each other personally; they all are (or at least are classified as) family and relatives. Their sense of themselves as family often survives growing integration into nation-states.

Sharing

The land and its resources in many of these societies are often collectively appropriated (Ingold, 1986). Their sharing cannot be understood as a form of exchange (Woodburn, 1998) but as a core part of their modes of sociality (Bird-David, 1994). Game is publicly shared; other resources are shared in response to direct demand for sharing (Peterson, 1993), especially foraged resources. Sharing and the expectation of it is part of broader nonaggrandizing economic reasoning and conduct. While hard to express in modern urbanite language, sharing extends not only to distributive but also to social experience. People in the group share each other: they share company, dwelling spaces and work, stories, memories, and

personal histories. They experience their selves as coalescent, rather than as a group of separate individuals, with cultural priority given to immediate relations (Bird-David, 1994). Sharing, which is normatively promoted, fosters an egalitarian society in many cases (see Woodburn, 1982). While closely connected, people usually enjoy considerable personal autonomy.

Giving Environment, Animism, Relational Epistemology, and Conservation

Nature is locally experienced neither as a mechanical system nor as objects for human use and ownership, but as an assembly of sentient beings related to and shared with (Bird-David, 1990). These beings (animals, forest, winds, hills, etc.) are frequently approached as relatives, often as grandparents. People care for them, and this helps sustain their collective identity. People engage with these beings through shamanistic practices and by pursuing hunting-gathering. To hunt and to gather are perceived not just as subsistence pursuits but also as means of communication with the 'natural' relatives. Even when forced to reside away, and to adopt modern lifestyles, people commonly practice their rituals, and return to their places to hunt-gather and 'keep in touch' with their relatives. This animistic perception is integrated with deep ecological understanding of the neighborhood. It often propagates responsiveness to and responsibility for constituents of the environment, along with its utilization (Bird-David, 1999), however, with conservationist effect limited to certain situations (Bird-David and Naveh, 2008).

Relations with Others

With the benefit of growing understandings of HGS themselves, attention has expanded to their relations with others, including traditional agricultural neighbors, ancient civilizations, and modern states. In many parts of the world, HGS have been living for centuries if not millennia not just in contact with others but in symbiotic relations with their neighbors, arguably as equal to or dependent on them (see Ikeya et al., 2009). Tropical forest hunter-gatherers of Central Africa (alias Pygmies) are known for their close ties with cultivators, and for links of fictive kinship and patron–client relations with them. Southeast Asian groups are known for close trade links between foragers and farmers, and historical shifts back and forth between foraging and shifting cultivation. South Asian foragers, as now newly appreciated, have maintained their foraging lifeways for millennia amidst Hindu civilizations (Gardner, 2013). HGS's conflicts, resistance, and self-determination in both developed and developing nation-societies, such as Canada, Australia, and Botswana, are extensively studied (Biesele et al., 1999).

Hunter-Gatherers' Ontologies

The 2000s surge of interest in indigenous ontologies, responding to the general 'ontological turn' in anthropology, has rejuvenated interest in hunter-gatherers' relational ontologies. Their ways of being human, and their relations with humans, nonhumans, and ancestors – as conceived, lived, experienced, and acted upon – are given serious attention as ontological potentialities. In the booming study of the indigenous ontologies, HGS are sometimes discussed together with other small-scale indigenous peoples such as, for example, Amazonian peoples. The pendulum may have swung too sharply to the opposite end, overlooking some possibly distinctive hunter-gatherers' ontological senses amidst general commonalities. HGS's worlds are often constitutive of relations of sharing and living together with humans and nonhumans, present and ancestral, and with their lands. Experienced relations suggest the ontological status of an interlocutor as subject or, at times, object premising that everything has ontological potentiality for communication. The ontological boundary between subject and object, human and nonhuman, is permeable, variable, and negotiable (see Poirier, 2013).

Conclusions

Once the world's sole inhabitants, HGS constitute now only a tiny fraction of the modern world population. Yet, they remain resilient as well as vulnerable, with surprisingly many groups still unstudied, especially in Central Africa and South Asia (see Fortier, 2009 on the latter). Their lives have been changing over the past 150 years during which they have been the subject of anthropological interest. But the perspectives that anthropologists have been taking on them during these years have changed more profoundly. HGS have found ways of partially continuing and renewing their hunting-gathering traditions, and their modes of sociality and ontologies, despite extraordinary pressures and violence, during these 150 years. Anthropologists have dramatically changed their perspectives on HGS: from regarding them in the nineteenth century as savages and creatures of nature – through approaching their lifeways in the early twentieth century as cultural-ecological adaptations to nature, and in the mid-twentieth century as cases from which to learn about human evolution and prehistory – now, to seriously addressing and learning from them alternative senses of nature and ontological potentialities.

Though their cases are rare, HGS continue to creatively help us think about a whole range of key issues, including socio-environmental ethics and praxis, child care and gender, conceptions of society and nature, human relationships and sharing economies, showing us markedly different ways from familiar models. In broadening scholars' imagination, that cannot but be biased by urban life experiences and intuitions, HGS can help scholars think about other forms of life: human life in the past as well as possible futures. HGS remain pivotal in modern thought as our ultimate 'Other' and our ultimate 'Us.'

See also: Australian Aboriginal Society and Culture: An Overview; Comparative Method in Anthropology; Economic Anthropology; Indigeneity; Modes of Production; Patron–Client Relations, Social and Anthropological Study of; Primitive Society.

Bibliography

Bird-David, N., 1990. The giving environment: another perspective on the economic system of gatherer-hunters. Current Anthropology 31, 183–196.

Bird-David, N., 1994. Sociality and immediacy: or past and present conversations on bands. Man 29 (3), 583–603.

Bird-David, N., 1999. Animism revisited: personhood environment and relational epistemology. Current Anthropology 40 (Suppl. 1), S67–S91.

Bird-David, N., Naveh, D., 2008. Relational epistemology, immediacy, and conservation: or, what do the Nayaka try to conserve? Journal for the Study of Religion, Nature and Culture 2 (1), 55–73.

Barnard, A. (Ed.), 2004. Hunter-gatherers in History, Archeology and Anthropology. Berg, Oxford.

Biesele, M., Hitchcock, R., et al. (Eds.), 1999. Hunter-Gatherers in the Modern World: Conflict, Resistance, and Self-determination. Berghahn Books, Providence, RI.

Burch Jr., E.S., Ellana, L.J. (Eds.), 1994. Key Issues in Hunter-Gatherer Research. Berg, Oxford.

Damas, D. (Ed.), 1969. Contribution to Anthropology: Band Societies. National Museum of Canada, Ottawa (Bulletin 228).

Fortier, J., 2009. The ethnography of South Asian foragers. Annual Review of Anthropology 38 (1), 99–114.

Gardner, M.P., 2013. Understanding anomalous distribution of hunter-gatherers: the Indian case. Current Anthropology 54 (4), 510–513.

Ingold, T., Riches, D., Woodburn, J. (Eds.), 1988. Hunters and Gatherers, 2 vols. Berg, Oxford.

Ingold, T., 1986. The Appropriation of Nature: Essays on Human Ecology and Social Relations. University of Iowa Press, Iowa City.

Ingold, T., 2000. The Perception of the Environment: Essays in Livelihood, Dwelling and Skill. Routledge, London and New York.

Ikeya, K., Ogawa, H., et al. (Eds.), 2009. Interactions between Hunter-Gatherers and Farmers: From Prehistory to Present. National Museum of Ethnology, Osaka.

Kelly, R.L., 1995. The Foraging Spectrum: Diversity in Hunter-Gatherer Lifeways. Smithonian Institution Press, Washington.

Leacock, E., Lee, R.B. (Eds.), 1980. Politics and History in Band Societies. Cambridge University Press, Cambridge.

Lee, R.B., DeVore, I. (Eds.), 1968. Man the Hunter. Aldine, Chicago.

Peterson, N., 1993. Demand-sharing: reciprocity and the pressure for generosity among foragers. American Anthropologist 95, 860–874.

Poirier, S., 2013. The dynamic reproduction of hunter-gatherers' ontologies and values. In: Boddy, J., Lambek, M. (Eds.), A Companion to the Anthropology of Religion. John Wiley and Sons Inc.

Steward, J.H., 1936. The economic and social basis of primitive bands. In: Lowie, R.H. (Ed.), Essays in Anthropology in Honor of A.L. Kroeber. University of California Press, Berkeley.

Strathern, M., 1991. Partial Connections. Roman and Littlefield, Sabage, MD.

Wenzel, G.W., Hovelstrud-Broda, G., et al. (Eds.), 2000. The Social Economy of Sharing: Resource Allocation and Modern Hunter-gatherers. National Museum of Ethnology, Osaka.

Woodburn, J., 1982. Egalitarian societies. Man (New Series) 17, 431–451.

Woodburn, J., 1998. Sharing is not a form of exchange: an analysis of property-sharing in immediate-return hunter-gatherer societies. In: Hann, C.M. (Ed.), Property Relations: Renewing the Anthropological Tradition. Cambridge University Press, Cambridge, pp. 48–63.

Huntington Samuel P. (1927–2008)

Andrew Tsz Wan Hung, The Hong Kong Polytechnic University – Hong Kong Community College, Kowloon, Hong Kong

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Samuel Phillips Huntington was an influential American political scientist. He was also a consultant to various American government agencies. He upholds the idea of conservative realism in politics. His research covers several areas of political science, such as civil–military relations, modernization and political development, comparative politics, and international relations. Regarding the role of military, he argues for autonomous military professionalism. In discussing about modernization of developing countries, he emphasizes the priority of political order over democracy. In the case of America, as well as other democratizing countries, he worries that the excess of political participation would diminish the authority and effectiveness of state governance, enhancing political instability, and, at worst, leading to the reverse of democratization. After the end of the Cold War, Huntington characterizes the international order as multicivilizational, in which the clash of civilizations will direct the development of global politics. In such a multicultural situation, Huntington urges Americans to affirm their unique national identity by revitalizing Anglo-Protestant culture and its religiosity, which has defined America since its founding.

Biography

Samuel Phillips Huntington was a prominent American political scientist. He was born in 1927 and brought up in a middle-class family in New York City. His father, Richard Thomas Huntington, was a publisher of a tourist magazine, and his mother, Dorothy Sanborn (née Phillips) was a short-story novelist. Huntington was the only son. He studied at Peter Stuyvesant High School. He wrote poetry when he was still a youth. At 16 years of age, he entered the Yale University, and has graduated with 'exceptional distinction' at 18 years of age; and then served in the US Army. In 1948, he acquired master's degree in political science from the University of Chicago.

In 1951, he was awarded a doctorate by Harvard University, where he finished his PhD dissertation in 4 months. He believed that his rapid completion of dissertation precipitated his diabetes, by which he need to administer himself three daily insulin injections ever since. During his study in Harvard University, he began teaching at 23 years of age in 1950. In 1957, his first important book, *The Soldier and the State: The Theory and Politics of Civil–Military Relations*, was published. This book brought to him widespread vehement attacks. Huntington's argument for the military professionalism in this book was considered by some senior professors in Harvard as a laudation to militarism, and Huntington was denied tenure in 1959. He then went to Columbia University as an associate professor of government and a deputy director of its Institute of War and Peace Studies. His two other books on military politics, *The Common Defense: Strategic Programs in National Politics* and *Changing Patterns of Military Politics* were then published respectively in 1961 and 1962.

In 1963, Huntington was asked to return to Harvard as a tenured professor of government. He then, coauthored with Zbigniew Brzezinski, his old Harvard colleague, and published *Political Power: USA–USSR* (1964), which was to compare the political systems in the US and the Soviet Union. In 1965, he became a fellow of the American Academy of Arts and Sciences.

In 1967, Huntington worked in Johnson Administration and wrote a 100-page report on the Vietnam War, *The Bases for Accommodation*, published in 1968b *Foreign Affairs*. In the article, while Huntington supported America to defeat North Vietnam, he criticized America for endorsing wrong strategy. To win the Vietnam War, Huntington argued for a strategy of 'forced draft urbanization,' a strategy of carpet bombing the rural countryside and to force the rural population moving into city; thus, weakening the support of the Vietcong by the ethnic and communal network in rural areas. However, the article caused students' tremendous protest. His office, the Center for International Affairs, was firebombed. The words "War Criminal Lives Here" were painted on his house front door.

At the same year, another important book, *Political Order in Changing Societies* (1968a), was published. In 1970, he published an edited volume, *Authoritarian Politics in Modern Society: The Dynamics of Established One-Party Systems*, which argued for durability of authoritarian politics. And *No Easy Choice: Political Participation in Developing Countries*, coauthored with *Joan M. Nelson*, was also published in 1976. These three books are considered as Huntington's important works on political development in developing countries, demonstrating his persistent thesis: If political participation outstrips institutionalization, political decay instead of development will result. His writings on political development and the Vietnam War built his reputation as a conservative hawk.

In 1970, Huntington cofounded the prominent *Foreign Policy* magazine with Warren Demian Manshel. Huntington remained as coeditor until 1977. In this time, he published two books about the instability of American democracy due to excessive political participation: *The Crisis of Democracy: Report on the Governability of Democracies to the Trilateral Commission* (1975), coauthored with Michel Crozier and Joji Watanuki, and *American Politics: The Promise of Disharmony* (1981).

During 1977 and 1978, Huntington was invited by Zbigniew Brzezinski, at that time President Carter's National Security Adviser, to work at the Carter White House as

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61219-9

coordinator of security planning for the National Security Council. He authored Presidential Directive 18 (PD-18), 'US National Strategy,' an analysis of US–Soviet strategic competition. He predicted that the ideological and military threat posed by the Soviet Union was just temporarily, the US would finally win. He suggested the US to expand the military power and to dispose the Rapid Deployment Force in the Persian Gulf to protect the region from Soviet intrusion. Huntington's analysis helped to inform the foreign policy from Carter administration to Reagan government.

In the 1980s, he became a member of the Presidential Commission on Long-Term Integrated Strategy. He also served as director of Harvard's Center for International Affairs from 1978 to 1989, and as president of the American Political Science Association from 1986 to 1987. In 1989, he started the John M. Olin Institute for Strategic Studies, and was its director until 1999. In this period, he has published *American Military Strategy* (1986).

The end of the Cold War further shaped Huntington's thought on the new international order and the US role in it. In *The Third Wave: Democratization in the Late Twentieth Century* (1991), he explored the features of third-wave democracy movement and demonstrates factors that advance or reverse the trend of democratization in the developing countries. He also listed the political, cultural, and economic obstacles to democratization. In particular, he discussed Confucianism and Islamic faith, and their relation with democratization in different countries in East Asia and the Middle East. The exploration of cultural obstacle of democratization seemed to pave the way to his most famous thesis 'The Clash of Civilizations and the Remaking of World Order,' which appeared in a 1993 *Foreign Affairs* article and in the 1996 book of the same title. These two books, as well as his earlier writing on political development, critically refute the over optimism about the universality of democracy assumed by many American scholars, especially associated with the thesis of 'End of History' posed by Francis Fukuyama in 1989. After the publication of the article, he served as chairman of the Harvard Academy for International and Area Studies from 1996 to 2004. In this period, he had published *Who Are We?: The Challenges to America's National Identity* (2004). He also involved in editing several books with other scholars on topics about military, political development, globalization, and culture.

Huntington died on 24 December 2008, Martha's Vineyard, 81 years of age. He married with Nancy Arkelyan Huntington for 51 years and they have got two sons. He was a self-declared conservative and a lifelong democrat, but an old-fashioned one. He claimed himself as a follower of Reinhold Niebuhr, the leading Protestant theologian of twentieth-century America.

Main Writings

The Soldier and the State: The Theory and Politics of Civil–Military Relations

This book was partly provoked by the firing of General Douglas MacArthur by President Harry Truman in the Korean War. The event inspired him to explore the civil–military relations in a liberal democratic society. Before, as civil–military relations are expected to be compatible with American liberalism, people used to endorse liberal idea of subjective civilian control that requires the military to involve in institutional politics, thus military's autonomy is legally and institutionally restricted. However, Huntington, in this book, argues for the theory of objective civilian control, which stresses "autonomous military professionalism" (Huntington, 1957: 83), and endorses conservatives, anti-individualistic ethos. While civilian officials make decision on policies, we should let military to have autonomy in all technical, professional matters, so that the military can work out the means to execute the policies. Huntington thinks that professionalizing military can best serve the security of American nation, instead of democratic force.

Huntington first provides a historical and theoretical analysis of officership as a profession, and argues that the expertise of the military officer is "the management of violence" (Huntington, 1957: 11). Liberalism believes in individualism, have faith in human capacity for advancement, and distrusts in power and military. For Huntington, such a view is too optimistic. Conservative realism rather believes "the permanence, irrationality, weakness, and evil in human nature" (Huntington, 1957: 79). The origin of violence lies in selfishness and weakness of human nature that can only be controlled by "organization, discipline, and leadership" (Huntington, 1957: 63). Military ethics stresses the importance of order and power in sustaining state security. It emphasizes the priority of community over individual, as individuals can only realize themselves by participating in the great continuing organic body. Thus, "[t]he military ethic is basically corporative in spirit" (Huntington, 1957: 64). Then, Huntington argues that the practice of soldier is a professional matter. An efficient society should leave room to soldiers to cultivate their professional abilities and have their own professional ethic. He reviews the prevalence of liberalism and the arduous development of the professionalism in the history of America from 1789 to World War II (WWII). He shows that the constant suspicion of professionalism by liberalism that had repeatedly attempted to 'enlighten' the soldiers, and thus reducing the efficiency of the military, which indeed threatens the security of the nation.

Finally, Huntington explores the civil–military relations during and after the WWII. In his analysis of the political roles of the Joint Chiefs and the Department of Defense, Huntington points out that the shortcomings of these institutions lie in the liberal ideology, which makes people to be unrealistic in thinking about war. Huntington, therefore, urges to abandon liberalism for military professionalism that is required in the coming age. The book is ended by adoring the elite military college at West Point.

Political Order in Changing Societies

In this book, Huntington argues for the priority of political order and stability over individual liberty in political development. In his studies of international politics during the 1950s and 1960s, Huntington notices a dramatically growth of incidence of political violence and a decline of political order in many modernizing countries. By using historical

comparative analysis of these different political orders, he shows that the political instabilities are not because of impoverished economy or backward politics and culture; rather they are largely due to the fast social change and mass political participation that come with the slow progress of political institution. Americans are blind by their happy history. They usually assume that all the goodness, such as democracy, economic development, and social justice, would automatically come together in modernization. However, it is not the case in many modernizing countries because of their different histories, geographies, and social conditions. Thus, he argues that for these developing countries, the primary concern is not about liberty; rather it is the question of how to establish a legitimate political order. He states, "Men may, of course, have order without liberty, but they cannot have liberty without order" (Huntington, 1968a: 7–8).

In these modernizing countries, the rise of economic growth and the living standard, and the literacy of people, coupled with the spread of Western values, led to the aspiration of political participation. Through political participation, socially mobilized individuals expressed their demands to their government. However, as these governments were low-level institutionalized, and the new values were resisted by traditional groups with deep-rooted privilege, the demands upon government were hard to be expressed in legitimate way and to be gathered within the political system. Thus, the rise of political participation engenders political instability.

For Huntington, the key to social stability is a strong political institution, which can respond political participation by social forces from different groups. Political institution is about political organization and procedure of administration, which aims at "maintaining order, resolving disputes, selecting authoritative leaders, and thus promoting community among two or more social forces" (Huntington, 1968a: 8–9). It can be measured by its adaptability, complexity, autonomy, and coherence. The more the adaptable, complicated, independent from other social forces, and coherent an organization is, the more highly institutionalized it is.

Huntington's discussion about military intervention in politics is especially interesting to many scholars. Instead of assuming the fondness of military to interrupt, the political intervention of these specialize social groups usually due to "the absence of effective political institutions capable of mediating, refining, and moderating group political action" (Huntington, 1968a: 196). Huntington calls this as "the praetorian society" (Huntington, 1968a: 195). As the society is lacking agreements among groups, public recognized political leaders or legitimate intermediaries to moderate conflicts between social forces, there would be different kinds of direct political participation. "The wealthy bribe; students riot; workers strike; mobs demonstrate; and the military coup" (Huntington, 1968a: 196). The method of military intervention is simply more dramatic and more effective to arbitrate the conflicts than others. Generally speaking, the military prefer to take the role of guardian, avoiding dangerous components brought into power of governors. If the military must govern, their leaders tend to be conservative. They usually resist to reform and to assimilate the newly politicized groups that the society requires. It makes those groups become the resentful enemies of the country.

Thus, for Huntington, military regimes are good at concentrating power and executing political reform, they are also strong in extending power to the urban middle class; but they are impotent in dealing with the complicated society, in coordinating and aggregating interests of different groups. Thus, Huntington suggests to "convert military intervention in politics into military participation in politics" (Huntington, 1968a: 243), which assimilates urban elites into politics while limits their political participation, so that the social and economic reforms are protected from traditional forces in the society.

Furthermore, the political reform must be slow and gradual. Historically speaking, the attempts of radical political reforms usually failed, because it would cause too much resistance from different parties. Instead, Huntington suggests the strategy of the "proper mix of Fabianism and blitzkrieg" (Huntington, 1968a: 346). It means that the reform cannot be fast; it should be step by step, drafted in relative secrecy and working on only one issue at a time. Once the time is right, the issue must be rapidly resolved, and then to erase it from the political agenda before the protest has been lodged by the opponents.

The Crisis of Democracy

In 1960s, there were many different kinds of social movements in America such as movement for civil rights, antiwar, feminism and sex liberation, etc. Popular participation has greatly increased, and authority of existing systems was challenged. The predominance of equality as a social, political, and economic goal was reasserted. For Huntington, the 1960s was "a decade of democratic surge and of the reassertion of democratic egalitarianism" (Crozier et al., 1975: 60). The immediate influence on government of democratic surge is contradictory: "The vitality of democracy in the United States in the 1960s produced a substantial increase in governmental activity and a substantial decrease in governmental authority" (Crozier et al., 1975: 64). Decline in governmental authority with high social expectation and mass political participation causes, what Huntington calls, "democratic distemper" (Crozier et al., 1975: 102), a threat to political stability. The distemper, for Huntington, stems from an "excess of democracy" (Crozier et al., 1975: 113). The remedy is not to have more democracy, but to restore balance between governmental activity and governmental authority, that is also "to restore balance between vitality and governability in the democratic system" (Crozier et al., 1975: 113). To achieve this moderation, we have to acknowledge two points. First, democracy is only one method of constituting authority; it is neither the only method nor the universal one. Sometimes, "the claims of expertise, seniority, experience, and special talents may override the claims of democracy as a way of constituting authority" (Crozier et al., 1975: 113). Second, the successful operation of democracy requires certain intervention from the state, and indifference and noninterference of political groups. If the political requirements have exceeded what the political institution can offer, the political authority may be damaged. Huntington, therefore, concludes that

although democracy is generally recognized as a good, it also has a limit; to maximize it is not always a good thing to do.

American Politics: The Promise of Disharmony

In this book, Huntington questions, why contemporary Americans are so rebellious, while most people in history usually conform to convention and follow rules. Huntington thinks that Americans, in the 1960s and 1970s, have experienced a period of creedal passion, in which people were aware of a gap between American ideal and its institutions with the atmosphere of discontent, and hence the political participation increases. He has also demonstrated three other creedal passion periods such as the American Revolution, the Jacksonian Era, and the Progressive Era. Different from other continental countries, Americans perceive their national identity as not defined by ethnics, heritage, or certain organic characteristic, but by political ideals. Huntington calls these ideals the 'American Creed,' which stresses democracy, equality, liberty, constitutionalism, and individualism. However, in reality, political institution inevitably attempts to maintain the political order involving a set of interests and certain restrains on freedom. This induces a permanent gap between ideals and institutions; Huntington calls it the 'IvI gap,' which leads to tension and disharmony in American politics. Americans' responses to the gap have shown a 'consciousness cycle' from moralism, cynicism, complacency to hypocrisy. At times, people have a highly intensity of belief in the ideals with a clear perception of the gap; they would respond in a moralistic way that attempts to eliminate the gap by changing the political system and reducing the power of government. When people find that the reform is impossible to achieve, they turn into cynicism and endure the gap. Cynicism then becomes complacency by denying the gap to deal with the cognitive dissonance, and finally develops into hypocrisy and ignores the gap. After a period, moralism will be reinvigorated. The cycle of creedal passion periods seems to happen about every 60 years. Huntington expects that the cycle would probably be continued. Nevertheless, because of continuing political development and America's increasing engagement in the international politics, a large complicated political institution is required. Huntington, therefore, argues that Americans should make a balance in the search of ideal and the political development. Although many critics find that America is a lie since its reality is far from its ideals, Huntington disagrees. He states, "America is not a lie; it is a disappointment. But it can be a disappointment only because it is also a hope" (Huntington, 1981: 262).

The Third Wave: Democratization in the Late Twentieth Century

In this book, Huntington identifies three waves of global democratization. The first started from Jacksonian democracy in the US in 1820s, and then extended up to 29 democracies in 1926. However, the rise of Mussolini in Italy started the first reverse wave that reduced democratic states down to 12 in the world. The second began with the triumph of the Allies in the WWII that reached 36 democratic countries in 1962, but back down to 30 in the second reverse wave during 1960–75. The third wave started from 1974 up to the time Huntington writing the book, meanwhile there are more than 30 countries making transition to democracy. This book mainly focuses on the study of the third-wave democracy by comparative approach.

Huntington has listed five common factors that have given rise to the occurrence of the third-wave democratization: (1) the aggravating legitimate crisis of authoritarian regimes by widespread of democratic values and the economic failure, (2) the fast global economic growth of the 1960s, which increases people's living standards and the growth of urban middle class, (3) the expansion of Christianity in the third-wave democracy and the alternation of attitude toward democracy by the Catholic Church, shown in the Second Vatican Council and the National Catholic churches' opposition toward authoritarianism, (4) the influence from the new policy of the external actors, such as the policy shift of the Soviet Union, and the promotion of human rights and democracy by European Community and the United States, and (5) the snowballing effect by successful democratization of other countries, especially to those countries that are geographically close and culturally similar.

Based on the experience of the last two reverse waves, Huntington also anticipates three factors that are likely to cause a third reverse wave: (1) Democratic states' failure of effective operation could undermine their own legitimacy, especially in the face of international economic crisis. (2) The reverse development from democracy to authoritarianism and dictatorship by nearby states, or the rise of power of nondemocratic states can also have reverse snowballing effect. (3) Different old and new forms of authoritarianism may emerge that meet the need of the time.

Finally, Huntington has discussed political, cultural, and economic obstacles to democratization. Economically, poverty is probably the main obstacle to democratization. The "transitions to democracy should occur primarily in countries at the middle levels of economic development" (Huntington, 1991: 60). Politically, the lack of experience with democracy and the existence of long-term authoritarian ruler would hamper the democratic development. Culturally, Huntington states, "[a] profoundly antidemocratic culture would impede the spread of democratic norms in the society, deny legitimacy to democratic institutions, and thus greatly complicate if not prevent the emergence and effective functioning of those institutions" (Huntington, 1991: 298). In particular, he discusses Confucianism in the East Asia countries and Islamic faith in the Middle East, elaborating how these traditions have various influences to the development of democracy in different countries. Basically, Huntington believes that the interaction between the cultural tradition and the social development is complicated. Furthermore, "cultures historically are dynamic rather than passive. The dominant beliefs and attitudes in a society change" (Huntington, 1991: 311). Therefore, these cultural obstacles are not necessarily preventing democratic development.

The Clash of Civilizations and the Remaking of World Order

With the end of the Cold War, Huntington claims, the international order has changed. For the first time in history, "global

politics has become multipolar and multicivilizational" (Huntington, 1996: 21). The primary distinctions among people in this new world are no longer ideological or economic, but cultural. The main thesis of this book is to show that cultural identity, or civilizational identities at the broadest level, are shaping the global political order in the post-Cold War world (Huntington, 1996: 20). By 'civilization,' Huntington means "a culture writ large" (Huntington, 1996: 41), "the highest cultural grouping of people and the broadest level of cultural identity" (Huntington, 1996: 43). He has identified eight existing civilizations: Sinic, Japanese, Islamic, Hindu, Slavic-Orthodox, Western, Latin America, and African. In the new international order, the interaction among states will be multidirectional, shaped by power and wealth, as well as their cultural preference and religion. The clash of civilizations will direct the global politics in the future. He also argues that modernization is not necessarily Westernization. The attempt to establish a universal civilization by the West is a kind of Western cultural hegemony. It is also the source of "the resentment and hostility of non-Western peoples against the West" (Huntington, 1996: 59).

As the influence of the West is declining, the balance of global power has shifted. The West has to face the increasing challenge from other civilizations that want to shape the world in non-Western style. Asian civilizations are enlarging their political, military, and economic strength; Islamic countries have experienced rapid growth in population and expanded religious ideology. The revitalization of Islam "embodies the acceptance of modernity, rejection of Western culture, and the recommitment to Islam as the guide to life in the modern world" (Huntington, 1996: 110). These non-Western civilizations generally maintain the values of their own religion and culture, and posed the threat to the Western civilization. It also shows that the world is becoming modernized without being Westernized.

Huntington predicts that the idea of Universalism by the West will cause growing clashes with other civilizations. He anticipates that the clashes with China and Muslim countries would be the most serious. An alignment between Muslim and Sinic cultures may be formed to fight against their common enemy, the West. Huntington has identified three cases by the West that would be perceived by non-Western countries as the attempt to enforce the Western cultural hegemony: (1) the proliferation of weapons to maintain military superiority, (2) the promotion of human rights and democracy, and (3) the restriction of non-Western immigration.

Furthermore, the conflicts at the local fault line, mainly between Muslims and non-Muslims, would create a "kin-country syndrome" (Huntington, 1996: 254) whereas states of a civilization would mobilize support from states of their civilization in the war against another civilization, their common foe. It would escalate the threat broader, and thus the core states would give effort to stop these wars.

The world is now inevitably multicultural. The United States and the West are facing both internal and external challenges. The internal is the declining of the Western values. The external is the rising of the non-Western cultural identities. On the one hand, Huntington suggests the West to recognize the fact of multicivilizational international order, to enhance the understanding of the religious and philosophical assumptions underlying other civilizations, and to identify commonalities among Western and other civilizations. On the other hand, Huntington believes that the endurance of the West depends on Americans' reaffirming their recognition of the West. So Americans have to perceive the Western civilization as a unique, not universal; to consolidate and to renew their own culture. This idea has also motivated him to write the following book, to explore the nature of American identity.

Who Are We?: The Challenges to America's National Identity

For Huntington, Americans are facing the identity crisis; the United States is under the threat of disintegration. In this book, he argues for the revitalization of America by reembracing "the cultural core" (Huntington, 2004: 59): the Anglo-Protestant culture, values, and religious commitment, without the elements of race and ethnicity.

In the book, Huntington refutes two widely accepted theses: (1) America is an immigrant nation and (2) American identity is solely defined by the American Creed. Huntington examines the history and argues that the America's founders in the eighteenth century should be considered as settlers rather than immigrants. They are not immigrants because they have founded a new polity with a distinct culture rather than arriving in an existing one. Thus, the origin of America was not "a nation of immigrants" (Huntington, 2004: 37). Rather it was a nation established by the Anglo-Saxons with core national culture: English and Anglo-Protestantism. In the nineteenth century, the substance of American national identity includes four elements: race (white), ethnic (Anglo-Saxon), religious culture (Anglo-Protestantism and English), and ideology (American Creed). For Huntington, America would not be the America today if it was not founded "by British Protestant, but by French, Spanish, or Portuguese Catholics … it would be Quebec, Mexico, or Brazil" (Huntington, 2004: 59).

After the United States has been established by the Anglo-Saxons, wave after wave of immigrants have attended, joined the settlement, and assimilated by the institutions, language, and culture of the core nation. This has also weakened the elements of ethnic and race in defining American identity. For Huntington, the main challenge was the mass immigration from Mexican with the Hispanic 'culture of Catholicism' in the Southwestern America in 1970s (Huntington, 2004: 254). They are different from previous immigrants because they want to maintain their own culture and are unwilling to be assimilated. At the same time, the universities and intellectuals started to embrace multiculturalism and criticized against the Anglo-Protestantism. Huntington worries that the reluctance to be assimilated by Hispanics could lead to "a bifurcated America, with two languages, Spanish and English, and two cultures, Anglo-Protestant and Hispanic" (Huntington, 2004: xvi). Indeed, apart from the rise of multiculturalism, the American national identity is also threatened by the globalization of business, professional and academic elites, the effect of diasporas, and the collapse of the Soviet Union, which left America lost the enemy against which to define itself.

Given the fact of multicultural situation, many liberal elites suggest to sustain the solidarity of the United States by embracing the American Creed. However, for Huntington, simply relying on the American Creed or any civic ideology is

too thin to maintain a nation for long. In the past, few nations have ever been successfully defined solely by a set of political principles. Some communist states, such as the Soviet Union, Yugoslavia, and Czechoslovakia, that attempted to use ideology alone to unite people of different cultures, have finally collapsed. On the contrary, the declining of communist ideology in China, with a core Han culture, has posed no threat to the solidarity of a country. Apart from ideology, identity must contain other substances such as shared emotion, language, memory, and culture. In particular, the 'core' Anglo-Protestant culture, which has been upheld by American of all races, ethnicities, and religions, is indispensable for American national identity. Indeed, the values of American Creed are not abstract and universal; they are also the product of the Anglo-Protestantism. Once the Anglo-Protestantism is rejected, and thus the American Creed is separated from its origin, the emotional and psychological force of the Creed would also be diminished. In the face of powers from different civilizations and countries, Huntington rejects the idea of cosmopolitanism, which opens America to be reshaped by other people and cultures. He also refutes the idea of imperialism, which attempts to use American power to reshape the world according to American values. Rather, Huntington suggests the nationalist approach, which recognizes and accepts the distinctiveness of America from other countries. The difference between America and other countries is largely defined by its Anglo-Protestant culture and its religiosity. Americans should strengthen the American identity, preserve and revitalize their culture that has defined America since its founding.

His Influence

Huntington is described as "[o]ne of the most influential political scientists of the last 50 years" by Corydon Ireland in Harvard News Office. He has established a new model of analyses of international and domestic conflicts. He also served as a mentor to dozens of famous policy makers and intellectuals such as Francis Fukuyama, the author of *The End of History and the Last Man* (1992), and Fareed Zakaria, the former managing editor of *Foreign Affairs* and the current editor of *Newsweek International*. His writings have shaped American views on civilian–military relations, democracy, modernization, political development, comparative government, and international relationship. His *The Soldier and the State*, in its 15th printing, is still considered a standard title about how military affairs interact with the political realm. The book *Political Order in Changing Societies* is also frequently reprinted, most recently in 2006. It remains the requisite reading for political science students today. In particular, after 9/11 terrorist attack, his *The Clash of Civilizations* is applauded as prescient to the global conflict of cultures, and has shaped recurrent debates on international relations. The book was translated into 33 languages; there also exists a pirated translation in Iran. The book shows a critique to those who thought the end of the Cold War meant the universal triumph of Western values.

See also: Civilizations; Conservatism: Theory and Contemporary Political Ideology; Democratic Transitions; Development and the State; Globalization and World Culture; Integration and Multiculturalism: Focus on Western Europe; Migration and Development; Militaries and Militarization, Anthropology of; Military History; Military, War, and Politics; Modernization, Sociological Theories of; Multiculturalism, Anthropology of.

Bibliography

Art, R.J., Davis, V., Huntington, S.P. (Eds.), 1985. Reorganizing America's Defense: Leadership in War and Peace. Pergamon-Brassey's, Washington.

Berger, P.L., Huntington, S.P. (Eds.), 2002. Many Globalizations: Cultural Diversity in the Contemporary World. Oxford University Press, Oxford.

Besteman, C., Gusterson, H. (Eds.), 2005. Why America's Top Pundits Are Wrong: Anthropologists Talk Back. University of California Press, Berkeley, pp. 43–82.

Biancardi, F., 2003. Democracy and the Global System: A Contribution to the Critique of Liberal Internationalism. Palgrave Macmillan, New York, pp. 95–135.

Bonney, R., 2008. False Prophets: The 'Clash of Civilizations' and the Global War on Terror. Peter Lang International Academic Publishers.

Crozier, M.J., Huntington, S.P., Watanuki, J., 1975. The Crisis of Democracy: Report on the Governability of Democracies to the Trilateral Commission. New York University Press, New York.

Harrison, L.E., Huntington, S.P. (Eds.), 2000. Culture Matters: How Values Shape Human Progress. Basic Books, New York.

Huntington, S.P., 1957. The Soldier and the State: The Theory and Politics of Civil-Military Relations. Belknap Press of Harvard University Press, MA. Reprinted in 1981.

Huntington, S.P., 1961. The Common Defense: Strategic Programs in National Politics. Columbia University Press, New York.

Huntington, S.P., 1962. Changing Patterns of Military Politics. Free Press of Glencoe, New York.

Huntington, S.P., Brzezinski, W., 1964. Political Power: USA–USSR. Viking Press, New York.

Huntington, S.P., 1968a. Political Order in Changing Societies. Yale University Press, New Haven. Reprinted in 2006.

Huntington, S.P., 1968b. The bases of accommodation. Foreign Affairs 46 (4), 642–656.

Huntington, S.P., 1970. Authoritarian Politics in Modern Society: The Dynamics of Established One-party Systems. Basic Books, New York.

Huntington, S.P., 1976. No Easy Choice: Political Participation in Developing Countries. Harvard University Press, MA.

Huntington, S.P., 1981. American Politics: The Promise of Disharmony. Belknap Press, MA.

Huntington, S.P., Nye, J.S. (Eds.), 1985. Global Dilemmas. Center for International Affairs, Harvard University, MA.

Huntington, S.P., 1986. American Military Strategy. Institute of International Studies, University of California, Berkeley.

Huntington, S.P., 1991. The Third Wave: Democratization in the Late Twentieth Century. University of Oklahoma Press, Norman.

Huntington, S.P., 1996. The Clash of Civilizations and the Remaking of World Order. Simon & Schuster, New York.

Huntington, S.P., et al., 1996. The Clash of Civilizations? The Debate. Foreign Affairs, New York. Twentieth Anniversary Edition, 2013.

Huntington, S.P., 2004. Who Are We?: The Challenges to America's National Identity. Simon & Schuster, New York.

Weiner, M., Huntington, S.P. (Eds.), 1987. Understanding Political Development: An Analytic Study. Little, Brown, Boston.

Hurst, James Willard (1910–97)

Aviam Soifer, William S. Richardson School of Law, University of Hawai'i, Honolulu, HI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

James Willard Hurst was the most influential American legal history scholar. This article examines his professional and personal life and addresses his lasting impact. Rather than focusing on the opinions of appellate courts, Hurst looked to the practical, everyday influences of law and lawyers and pioneered a new approach to American legal history.

J. Willard Hurst of the University of Wisconsin was the most influential scholar within the realm of American legal history. By emphasizing factors beyond legal doctrine and by altering the way people think about time, place, and change itself, Hurst enlivened and transformed an entire academic field. Doing legal history from the outside – from the perspective of sociological jurisprudence – led Hurst to turn his back on the 'great man' tradition and to reject the previously dominant focus on formal judicial opinions. He firmly opposed teleological accounts that tended to celebrate the gradual improvement of law over time. Instead, Hurst focused on economic matters, as shaped by the general culture, and on the ordinary processes of statutes, administrators, unremarkable cases, and the everyday work of lawyers. In particular, Hurst concentrated on tracking what he termed 'the expansion of middle-class energy,' which he called 'our prime inheritance.'

Throughout his decades of scholarship, teaching, and mentoring, Hurst combined careful and unremitting attention to massive low-level details with sweeping descriptions of general patterns. Hurst not only redefined an entire academic field, but he also did a great deal to open the borders of legal history to other disciplines. As leading legal historian Lawrence Friedman said: "Willard Hurst, as a scholar, did what few scholars have ever succeeded in doing: he created a whole new field of study, and virtually out of nothing" (Friedman, 1965).

Biographical Background

James Willard Hurst was born on 6 October 1910 in Rockford, Illinois, where his father ran a movie theater and had other business interests. When Hurst attended Williams College from 1928 to 1932, he majored in economics, and became editor in chief of the student newspaper and president of the International Affairs Club. He graduated Phi Beta Kappa and wrote a prize-winning student essay on international implications of the English Gold Clause crisis of 1931, published in the *Journal of Political Economy* (Hurst, 1932). Hurst then attended Harvard Law School, and reported that he was disgruntled about the emphasis on abstractions and the failure to pay attention to the depression and to the ideas of the legal realists. A Note Editor of the *Harvard Law Review*, Hurst graduated in 1934 and spent the 1935–36 academic year as a research fellow for Felix Frankfurter, working primarily on the lecture series Frankfurter published as *The Commerce Clause under Marshall, Taney, and Waite* (Frankfurter, 1937).

Hurst next served as law clerk for 80-year-old United States Supreme Court Justice Louis D. Brandeis during the 1936 term, one of that Court's most contentious ever. Brandeis thought the states to be invaluable laboratories for progressive regulation and lawmaking, and he urged young men who worked for him to return to their home regions. Hurst became one of the very few to heed Brandeis's advice. In 1937, Hurst declined an invitation to teach at Yale, instead choosing to teach at the University of Wisconsin, not far from his hometown. The fact that the law school was connected to the rest of the University appealed to Hurst – it was easy to work together with people in economics, sociology, and ultimately, the history department (Garth and Sterling, 1998). In the course of rejecting the offer from Yale, Hurst wrote to Frankfurter that Wisconsin afforded 'just about an ideal 'laboratory' situation.' Despite Yale's offer of higher rank and double the salary, Hurst advised Frankfurter that he preferred Madison, Wisconsin, 'particularly because of this extraordinarily interesting state government machinery here on the doorstep' (Soifer, 1992).

By 1941, Hurst had compiled several books of teaching materials and coauthored *Law in Society* (Hurst and Garrison, 1941) with his dean, Lloyd Garrison. This pioneering casebook focused on the myriad ways that the plight of a widow – a woman whose husband, a carnival worker, was crushed under a company tractor – intersected with legal institutions beyond the courts. This book is often said to be the first to examine norm creation through the historical process of public lawmaking (Tomlins, 2000b [34 Law Soc'y Rev.]).

After Pearl Harbor, Hurst and his wife, Frances Wilson Hurst, moved to Washington, D.C. Hurst first worked for the general counsel of the Board of Economic Warfare and then joined the Navy, where his assignments included work on a highly publicized treason case, leading eventually to *The Law of Treason in the United States* (Hurst, 1971b).

Even before returning to the University of Wisconsin after World War II, Hurst began to write about the crucial link he perceived between legal history and 'opportunities to apply informed, humane reason to influence the course of events' (Soifer, 1992). The 'Wisconsin Idea' of enlightened political leadership had lasting appeal for Hurst, and he repeatedly rejected job offers to teach at other prestigious institutions. Although Hurst was extremely reluctant to travel from his Wisconsin base, he nonetheless was directly involved in launching the international Law and Society movement, as well as in pioneering a new approach to legal history that has had worldwide impact (Christopher Tomlins, 2000a). He taught

full time at Wisconsin until 1980 and continued to publish after he retired. Hurst died in Madison on 18 June 1997.

Hurst's Scholarship

Hurst's scholarly career repeatedly underscored a core idea. In 1971, he summarized the enormous project he had undertaken: "A realistic history of law in the United States will be a social history of law, taking law as man-made, and as the product of both deliberation and drift, but not of any immanent superior order of reality" (Hurst, 1971a).

Hurst sought to describe and analyze how most people responded to a central problem. That basic issue, according to Hurst, requires consideration of 'whether men will be able to use the power of the politically organized community to secure the basic conditions of a decent life for the individual, without thereby destroying the very values they are trying to realize' (Soifer, 1992).

Throughout his decades of scholarship and teaching, Hurst sought to develop a new integration of the social science ideas of his era with careful historical consideration of the largely unexamined mass of law as it functioned on the ground, far removed from the doctrines promulgated by selected appellate judges that dominated – and still dominate – legal education in the United States. His ultimate goal was to address the needs of society by using law in all its forms. Lawyers thus were to be trained not only to be 'officers of the court,' but also to be 'officers of the state.' In particular, Hurst concentrated on tracking what he termed 'the expansion of middle-class energy.' Legendary for his diligence and prodigious scholarly output, Hurst also spent a great deal of time encouraging younger generations of legal historians. The correspondence that Hurst wrote to those who sent him their drafts and articles was noteworthy for the insightful and detailed feedback he provided – which he typed himself, replete with typos – as well as for his quiet, heartfelt encouragement.

Robert Gordon described the great reach of Hurst's innovative academic project as follows:

> After expanding the relevant record beyond prior appellate cases and their facts to all prior legal acts and the social influences on them, and the pending proceedings to the overall task of designing workable legal-regulatory arrangements for the present age, Hurst wrote his books as background briefs for present-day lawmakers.
>
> Gordon, 1975

To Hurst, social patterns mattered greatly. But he perceived that generalized thoughtlessness and a strong tendency to accept inertia and drift often served to counterbalance, or even to disable, other strong, competing commitments.

Through his prodigious output of scholarship on markets, legislation, monetary policy, and administrative actions large and small, Hurst pointedly emphasized the frequent unthinking clash between core values in the United States. He seemed to delight in clearly describing basic sobering paradoxes. For example, Hurst observed: "Individuals realize their humanity only in society, but they realize their individuality only in self-awareness, which consists in some sense of separateness." Moreover, he noted that such tensions afford protective coloration for 'shrewd men' who 'warp individualistic symbols of the constitutional ideal to sanction unchecked organized power' (Soifer, 1992).

As Hurst challenged traditional legal distinctions between public and private activity, he pushed himself and those who followed him to develop both extensive local data and expansive theories to capture and explain overarching patterns. In *The Growth of American Law* (Hurst, 1950), Hurst argued that 'law has been more the creature than the creator of events.' This famous book emphasized the impact of physical, technological, and social facts on nineteenth-century law and paid particular attention to the often inglorious role of the legal profession. Some have criticized Hurst for being too much a functionalist, and for inadequately attending to issues of power. Yet Hurst's 946-page *Law and Economic Growth: The Legal History of the Lumber Industry in Wisconsin, 1836–1915* (Hurst, 1964) highlighted the failures of laissez-faire. It also extensively documented Wisconsin's broad and detailed state regulation of a single industry during a period generally thought to have been at or near the apex of free-market ideology.

Hurst's Impact

William Novak asserted in 2000 that Hurst must surely rank as 'a broad-gauged socio-legal thinker of the first order, writing in the tradition of Tocqueville, Weber, Holmes, and Pound' (Novak, 2000). Hurst was original in his emphasis on place and time, particularly through his focus on his native Midwest. He was also innovative through his unusual illustration of the difficulty in discerning any clear patterns of progress in legal developments in the United States. Indeed, Hurst concentrated on exploring the intricate network of constraints – some perceived, some never understood, some formally legal, some not legal at all – that repeatedly confined and constrained the freedom of everyone.

In Hurst's view, Americans competed fervently, if often blindly, to manipulate whatever system they found in place in their particular context. He generally found it necessary to bemoan the power of deep habits that often produced irrational development and substantial depredations over time. Yet, Hurst also provided crisp accounts of the constitutional impulse among Pike Creek squatters, for example, as well as broad studies of such issues as the fierce, largely irrational conflict over gold and silver currency in the late nineteenth century.

Hurst sometimes has been inaccurately caricatured as part of the celebratory framework of some forms of 'consensus history.' Yet, the very broad-gauged quality of his work (an ongoing quest, in Novak's words, "to try to twist the tail of the cosmos") emphasized the mainstream at the cost of excluding, or at least understating, the importance of minorities and dissenting viewpoints throughout American legal history (Novak, 2000). Hurst's faith in fact-gathering led him to produce not merely a new way of looking at law but also an extended historical critique of what he called 'bastard pragmatism' – the propensity of people throughout United States history to

allow themselves to be dominated by the short term and the bottom line (Hartog, 1994).

Through his career, Hurst emphasized the difficult, if not impossible, goal of increasing both public and private accountability. In *Law and the Conditions of Freedom* (Hurst, 1956), he noted a clear historic preference among Americans for 'the promise of the future over commitments to the past, but also for the ambitions of the present over the claims of the future.' To Hurst, facts were crucial. Only through facts could a doubting optimist, such as Hurst himself, imagine enlightened leaders actually enhancing the general pursuit of 'lives of good conscience and fulfillment.' He maintained a broad view of the appropriate role for government.

If Hurst sometimes seemed unabashed in adopting what Daniel Ernst called 'a presentist approach to the past,' he remained first and foremost an innovative, exceptionally influential historian (Ernst, 2000).

There continues to be a considerable amount of scholarship published to discuss and honor Hurst, and there have been numerous symposia considering his work. After Hurst's death in 1997, the Law and Society Association established the James Willard Hurst Prize in legal history and, in 2001, the American Society for Legal History and the Institute for Legal Studies at the University of Wisconsin Law School created the Hurst Summer Institute in Legal History to assist and encourage legal historians early in their careers. The unusual number and breadth of symposia dedicated to Hurst only begin to suggest how significantly Hurst's emphasis on the impact of markets, legislation, monetary policy, social habits, and administrative actions transcended earlier boundaries. Through his enormous output of scholarship, his legendary generosity in giving scholarly assistance to others, and his noteworthy teaching, Willard Hurst changed the way the legal history of the United States is written and taught. By now, as Lawrence Friedman put it, "You're either a Hurstian or a reviser of Hurst" (Margolick, 1990).

Although Hurst has often been criticized for not considering groups that have been subordinated within the legal system, this retrospective critique does not negate the major importance and continuing relevance of Hurst's work. Hurst not only redefined an entire academic field, but he did much to expand the borders of legal history to other disciplines. The framework he applied in creating the first comprehensive overview of law in modern America has continued to open doors and to refresh legal history. Hurst's career challenges his successors to consider carefully the intricate, crucial, often paradoxical intertwining of legal ideas and social realities within the many realms in which people actually live.

See also: Law and Society; Law as an Instrument of Social Change; Law, Sociology of; Law: History of Its Relation to the Social Sciences; Legal Education; Legal Realism; Supreme Courts.

Bibliography

Abrahamson, Shirley S., Frank, John P., Luce, Charles F., Friedman, Lawrence M., Foster, Bill, Skilton, John, Mermin, Samuel, Macaulay, Stewart, Eskridge Jr., William N., McEvoy, Arthur F., November–December 1997. Tributes to James Willard Hurst. Wisconsin Law Review 6, 1123–1210.

Diamond, Stephen, 1979. Law and social order in the United States – Hurst, JW. Michigan Law Review 77, 784–794.

Erlanger, Howard, 1985. Review symposium: the work of J. Willard Hurst. American Bar Foundation Research Journal, 113–144.

Ernst, Daniel R., Spring 2000. Willard Hurst and the administrative state: from Williams to Wisconsin. Law and History Review 18 (1), 1–36.

Frankfurter, Felix, 1937. The Commerce Clause under Marshall, Taney, and Waite. University of North Carolina Press, Chapel Hill.

Friedman, Lawrence M., 1965. Contract Law in America: A Social and Economic Case Study. University of Wisconsin Press, Madison.

Garth, Bryant G., Sterling, Joyce, 1998. From legal realism to law and society: reshaping law for the last stages of the social activist state. Law and Society Review 32 (2), 409–471.

Gordon, Robert W., 1975. Introduction: J. Willard Hurst and the common law tradition in American legal historiography. Law and Society Review 10 (1), 9–55.

Hartog, Hendrik, Autumn 1994. Snakes in Ireland: a conversation with Willard Hurst. Law and History Review 12 (2), 370–390.

Hurst, James W., Garrison, Lloyd K., 1941. Law in Society: A Course Designed for Undergraduates and Beginning Law Students; Cases and Other Materials. College Typing Co, Madison, WI.

Hurst, James W., 1932. Holland, Switzerland, and Belgium and the english gold crisis of 1931. The Journal of Political Economy 40, 638–660.

Hurst, James Willard, 1950. The Growth of American Law. Little, Brown and Company, Boston.

Hurst, James Willard, 1956. Law and the Conditions of Freedom in the Nineteenth-Century United States. University of Wisconsin Press, Madison.

Hurst, James Willard, 1964. Law and Economic Growth: The Legal History of the Lumber Industry in Wisconsin, 1836–1915. Harvard University Press, Cambridge.

Hurst, James W., 1971a. Legal elements in United States history. 5 Perspectives in American History 28, 91.

Hurst, James W., 1971b. The Law of Treason in the United States. Collected Essays. Greenwood Pub. Corp, Westport, CT.

Konefsky, Alfred S., Spring 2000. The voice of Willard Hurst. Law and History Review 18 (1), 147–165.

Landauer, Carl, Spring 2000. Social science on a lawyer's bookshelf: Willard Hurst's law and the conditions of freedom in the nineteenth-century United States. Law and History Review 18 (1), 59–96.

Margolick, David, 23 March 1990. Law: at the bar. The New York Times. http://www.nytimes.com/1990/03/23/us/law-at-the-bar.html?src=pm.

Novak, William J., Spring 2000. Law, capitalism, and the liberal state: the historical sociology of James Willard Hurst. Law and History Review 18 (1), 97–146.

O'Neil, Robert M., Garrison, Lloyd K., Fleming, Robben W., Luce, Charles F., Cohen, Milton H., Murphy, Earl F., Friedman, Lawrence M., Gordon, Robert W., November–December 1980. Dedications to J. Willard Hurst. Wisconsin Law Review 6, 1093–1130.

Scheiber, Harry N., February 1970. At the borderland of law and economic history: the contributions of Willard Hurst. American Historical Review 75 (3), 744–756.

Soifer, Aviam, March 1992. In retrospect: Willard Hurst, consensus history, and the growth of american law. Reviews in American History 20 (1), 124–144.

Tomlins, Christopher, 2000a. Engaging Willard Hurst: a symposium. Law and History Review 18, 7–221.

Tomlins, Christopher, 2000b. Framing the field of law's disciplinary encounters: a historical narrative. Law and Society Review 34, 911–972.

Tushnet, Mark, 1972. "Lumber and the legal process. Wisconsin Law Review 1, 114–132.

Husserl, Edmund (1859–1938)

Christian Thiel, Institut fuer Philosophie, Universität Erlangen-Nürnberg, Erlangen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Edmund Husserl (1859–1938) was one of the most influential philosophers of the twentieth century. Inventor of the 'phenomenological method' for the analysis of physical and mental phenomena, with 'intentionality' at its center, and initiator of the 'phenomenological' movement of contemporary human sciences, he became the great critic of naive 'naturalism' in philosophy and psychology ('psychologism'), and of what he regarded as the loss of orientation, certainty, and thereby of credibility of modern science. His appeal for 'going back to the things themselves,' and for overcoming the crisis of modern science by a foundation of all science and philosophy in a clearly understood 'life-world' (*Lebenswelt*) inspired the work of a great number of philosophers, including Gadamer, Heidegger, Ingarden, Levinas, Merleau-Ponty, Ricoeur, and Sartre, as well as social scientists such as Buytendijk, Kurt Lewin, Luckmann, and Alfred Schütz. Husserl's impact on the social sciences was mainly indirect, through (1) revealing the deficiencies of a merely naturalistic approach in science, (2) supplying very detailed methods for the description of phenomena, relations and situations, and (3) stimulating applications of the phenomenological method in many areas of social science, e.g., in sociology as brought into effect by Schütz.

Life and Academic Career

Edmund Gustav Albrecht Husserl was born in Prossnitz, Moravia, on 8 April 1859. He studied mathematics, physics, astronomy, and philosophy at the universities of Leipzig (1876–78), Berlin (1878–81), and Vienna (1881–82), where he obtained his PhD degree in mathematics with Leo Königsberger in 1883. After a short period as assistant to the mathematician Karl Weierstrass in Berlin, and a year of military service, Husserl continued his philosophical studies with Franz Brentano in Vienna from 1884 to 1886, and with Carl Stumpf in Halle between 1886 and 1887, where he qualified as lecturer in 1887 with a postdoctoral thesis entitled 'Ueber den Begriff der Zahl. Psychologische Analysen' ('On the Concept of Number. Psychological Analyses').

Husserl became a professor in 1894, and moved to Göttingen in 1901. In 1916, he accepted a call to Freiburg to succeed Heinrich Rickert in the chair of philosophy there. Husserl retired in 1929. After the seizure of power in Germany by the National Socialists in 1933, his status of emeritus professor was suspended because of his Jewish descent, and, after a short reinstatement, definitively revoked in 1936 after the Nuremberg race laws. Husserl died in Freiburg on 27 April 1938.

Writings, Subjects, Approaches, and Contexts

Husserl's significance for philosophy, which is certainly greater than that for the social and behavioral sciences, rests on his establishment of a new way of dealing with philosophical problems, which he called 'phenomenology,' turning this expression into a 'technical' term by giving it a clearer delimitation and a weightier role than it had enjoyed with earlier philosophers such as Lambert or Hegel. The methods and content of Husserl's phenomenology, however, changed considerably during their development, which may be summarized in terms of three 'turns' of direction and aims: (1) the turn from early foundational work to antipsychologism, (2) the shift from descriptive to 'constitutive' phenomenology, and (3) coping with the 'crisis' of modern science by a new focus on the life-world.

From Early Foundational Work to Antipsychologism

In *Philosophie der Arithmetik* (*Philosophy of Arithmetic*, 1891, Husserl, 1950: Vol. XII, pp. 1–283), Husserl made an attempt at a descriptive-psychological foundation of arithmetic, presumably having in mind an extension of Weierstrass's program of basing the vast number of mathematical concepts on a small number of fundamental ones, by providing an even deeper foundation for the latter by grounding them in the psychical acts underlying the mathematician's work. In the course of these investigations, Husserl introduced (independently of C. von Ehrenfels, who is normally credited with this discovery) 'gestalt qualities,' which he treated under the name *figurale Momente*.

Less successful (with hindsight) was Husserl's criticism of G. Frege's definition of number. It invoked an objective but sharp countercritique by Frege, which seems to have persuaded Husserl of the impossibility of taking the special science of psychology (which proceeds inductively, falling short of universally valid statements), as a foundation for mathematics or even for all special sciences ('psychologism'). Therefore, in *Logische Untersuchungen* (1901), aiming at a foundation of logic, which at that time was usually set up on a psychological basis, Husserl advocated a consequent antipsychologism, which influenced contemporary philosophy profoundly. This concept was extended in his work *Philosophie als strenge Wissenschaft* (*Philosophy as a Rigorous Science*) (Husserl, 1950: Vol. XXV, pp. 3–62) to a critique of historicism and (under the label of 'naturalism') scientism (1910–11).

The positive foundation of logic was to be laid by a new foundational discipline to which Husserl, oddly enough, gave the name 'descriptive psychology,' despite his claim that it differed from empirical psychology as well as from Brentano's descriptive classification of psychical (mental) phenomena; it was to be 'descriptive' in the sense of presenting the

immediately given mental facts without presuppositions (*voraussetzungslos*). According to this earliest form of phenomenology, that which is intended in a mental act (perception, judgment, idea, etc.) is transcendent to the act itself and given only in 'adumbrated' (*abgeschattet*) manner, through changing phenomena, and therefore inadequately. By contrast, everything immanent to the act is free of adumbration and can therefore be regarded as given in an absolutely adequate manner. Moreover, intuition of what is immanently given appears as free of structural complications like those attributed to judgment in contradistinction to sensation. The contents of the sphere of immediate experience (*Erlebnisse*) are beyond every doubt; with 'intentionality' at its center, it is the secure ground of a foundational discipline, which is capable of justifying statements about essences (*Wesen*) and their relations, by a method, which Husserl first called 'eidetics,' but later called 'intuition of essences' (*Wesensschau*).

Particular relevance to a philosophy of the sciences may be claimed for Husserl's remarks on 'pure logic.' Pure logic is an a priori theory of science dealing with two kinds of fundamental categories: categories of meaning (concept, proposition, truth, inference, etc.) on the one hand, and categories of objects (object, state of affairs, connection, number, etc.) on the other. Husserl takes the unity of the propositions of a science (or a discipline) to consist in the unity of its explanatory principles, and finds his next task in the uncovering of the conditions of the very possibility of theories. The route is to proceed from preliminary studies, e.g., a 'purely logical grammar' investigating the components and connections of propositions, via the statement of laws obtaining between such categories, to an a priori 'theory of possible theory forms.' These investigations, continued only in *Formale und transzendentale Logik* (*Formal and Transcendental Logic*, 1929) and in *Erfahrung und Urteil* (Husserl, 1939), share a methodical aim: the "fixation of unambiguous, sharply distinct verbal meanings" (Husserl, 1901: Vol. I, p. 244; Husserl, 1939 (English translation 1973): Vol. I, p. 238).

This search for 'verbal meanings' is also one of the foremost goals of contemporary methodical philosophy and philosophy of science. Husserl, however, did not approach the joint methodical aim by actually constructing a language of science, but recommended an 'intuitive presentation (*Vergegenwärtigung*) of essence in adequate ideation,' obviously conceived as independent of and antecedent to any linguistic construction. The ideation to be performed is not an idealization (as in geometry), but pure description, aiming at the elevation of that which is inexact and essentially vague to something typical, a 'type.' By setting out in this way, phenomenology increasingly moved away from the other methodical currents of modern philosophy and came closer to the philosophy of life (*Lebensphilosophie*), from which, however, Husserl differs sharply methodically by ascribing a foundational value to experience, after its precise rendering as 'pure experience' (*reines Erlebnis*) in which meaning, anterior to everything real, is believed to constitute itself (philosophical intuitionism).

The Step from Descriptive to 'Constitutive' Phenomenology

Object and method of phenomenology were extended and partly revised in *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie* (Husserl, 1913), and pure logic was also characterized in a new way. In Husserl (1901), the predicate 'ontological' had been restricted (diverging from tradition) to sciences such as geography, astronomy, and anatomy, which deal with concrete objects, and not with abstract objects (like the theoretical or nomological sciences). Now, the old term 'ontology' is taken up again in a new sense. Husserl takes being to be subdivided into highest kinds or 'regions,' so the sciences, which deal with them yield 'regional ontologies.' Above these (material) ontologies we find the formal sciences, mathematics and logic, which are aimed at 'form essences' (*Formwesen*) and give rise to a *mathesis universalis*. Treating the 'empty forms of region' as such, they comprise the forms of all possible material ontologies, and are nothing else but a 'formal ontology.'

In *Formale und transzendentale Logik*, formal logic was subdivided into (1) a pure doctrine of forms of meaning, or logical grammar, (2) the logic of consequence, or logic of noncontradiction, and (3) the logic of truth. The purely formal ontology was characterized more clearly as the logic of objects by severing it from the 'apophantic' logic of judgment, while the nomological sciences are deductive, formal–analytic sciences describing 'definite manifolds.' Fundamentally, new is Husserl's assessment of ontology as a naive–dogmatic discipline that is turned into a science only by its founding on a 'transcendental logic.' The term 'transcendental,' as explained in 1936, refers to "the motif of inquiring back into the ultimate source of all the formations of knowledge, the motif of the knower's reflecting upon himself and his knowing life" (Husserl, 1954: p. 100). Husserl's reflective 'pure consciousness' of 1913 is now determined further by its function as the 'transcendental ego,' having everything as its object, while never becoming an object itself, and granting access only through 'eidetic intuition.'

The absolute sphere, which is the foundation of ontology and may be distinguished from the ontological sphere as 'semasiological' (R. Zocher), is the object of a 'critical' transcendental philosophy, which Husserl can only conceive of as a fully developed phenomenology. It makes use of the grasping of essences as a tool for analyses, which approach the problem of foundation not as a methodological task, but as a problem of constitution in the classical sense: in this sphere, according to Husserl, valid statements may be obtained only by a sequence of increasingly radical reduction steps, each of which is to consist in a certain 'abstention' (*epoché*). The first two steps, i.e., 'phenomenological reduction,' leading to the phenomena, and 'eidetic reduction,' leading to the essential structure of consciousness, carry us on to the transcendental sphere. Access to it, however, is only given by 'transcendental reduction,' which is the 'absolutely universal reduction,' opening up the transcendental subjectivity, which constitutes the meaning of the world and its contents, and thereby all objecthood.

The Crisis of European Science and the Life-World

The period between World War I and the seizure of power by National Socialism in Germany abounds with diagnoses of crises, from the foundational crises in most sciences (claimed for psychology, sociology, economics, etc., following the

foundational crisis of classical mathematics so dramatically exposed by H. Weyl in the 1920s) to a general crisis of the times, of intellectual and moral orientation, and of European ('rational') culture generally.

Husserl gave much thought to this phenomenon and made out a growing loss of confidence in the sciences hitherto regarded as the firm foundation of Europe's rational culture. He analyzed this situation comprehensively in an article *Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie* published in 1936 in a new and little-known Serbian journal and later forming the first two parts of the monograph bearing the same title (Husserl, 1954). The sciences, it seemed to him, while 'functioning' as well as ever technically, had lost sight of their origin as well as of their aims, and thereby lost their former role and status as a guide and a standard of reliable knowledge and moral responsibility. Husserl hoped for a renewal by recognition of the fact that all scientific knowledge and activity is grounded in and grows out of everyday life. This is a realm usually not reflected upon but capable of thorough analysis as our 'life-world,' the domain not only of the physically 'given,' but primarily of our social life. Our immediate experience is not that of objects given to our consciousness, but that of situations in which we find ourselves with other people, in social groups and institutions, participating in common meaningful action, handling (linguistic and other) symbols, and approaching with understanding even "the great symbolic systems such as language, myth, religion, art, etc.; all of which are essential elements of the *Lebenswelt* and therewith of highest interest to the social sciences" (Schütz, 1959: p. 96). The phenomenological analysis of these largely 'prepredicative' constituents of the life-world was envisaged by Husserl in the *Crisis* treatise and begun in his late- and posthumously published work *Erfahrung und Urteil* (1939), where he has recourse to the life-world and to the primacy of praxis over mere contemplation. But in front of them, he places the 'prepredicative' evidence of perception and of judgment based on perception, and thereby once more a prelinguistic foundational domain. The new approach remained to be developed by others.

Contexts of Discussion and of Influence

Notwithstanding the fact that "Husserl literally 'thought through his pen'" (Embree et al., 1997: p. 327), entrusting an incredibly large part of his ideas to shorthand memos that remained unpublished during his lifetime (and part of them until now), his thinking did not develop monologically, without context. During Husserl's student years, the careful study of Descartes, Leibniz, the British empiricists and Brentano was recommended to him by T. Masaryk (today less famous for his *Versuch einer concreten Logik* of 1887 than for his later career as the first president of Czechoslovakia), and Husserl himself testifies to the influence of William James's *Principles of Psychology* of 1890 on the shaping of his phenomenological thought. Having acquired some fame as a creative philosopher, Husserl corresponded with many experts in science and philosophy, and discussed his ideas at invited lectures at other German universities, and abroad in Vienna, Prague, Paris, Amsterdam, London, and elsewhere – events certainly prepared by the translation of the *Logische Untersuchungen* and the *Ideen* into several foreign languages.

The beginning of the 'phenomenological movement' may be seen in the exchange of ideas with students and collaborators of T. Lipps in Munich, who learned about Husserl's investigations through their fellow student J. Daubert after he met Husserl in Göttingen in 1902. In spite of having nothing in print, Daubert had considerable influence on the 'Munich phenomenologists' (among them M. Geiger, A. Pfänder, and A. Reinach), some of whom (like H. Conrad-Martius, A. Koyré, and E. Stein, later one of Husserl's private assistants) came to Göttingen to work with Husserl. They became the leading members of the 'Göttinger philosophische Gesellschaft' and main authors of the *Jahrbuch für Philosophie und phänomenologische Forschung*, founded in 1913, which published some of the great contributions (phenomenological and other) to contemporary philosophy, such as Husserl's *Ideen* in 1913, M. Scheler's *Der Formalismus in der Ethik und die materiale Wertethik* (*Formalism in Ethics and the Material Ethics of Value*) in 1913–16, and M. Heidegger's *Sein und Zeit* and O. Becker's *Mathematische Existenz* (*Mathematical Existence*), both in 1927. After its discontinuation in 1930, the publications' goals and role were taken up by the American journal *Philosophy and Phenomenological Research*, founded in 1940.

In Husserl's most active period, private assistants helped organize his literary production, defended his positions (e.g., E. Fink against neo-Kantian critics), and prepared lectures and notes for publication (e.g., the *Zur Phänomenologie des inneren Zeitbewußtseins*, prepared by E. Stein and edited by M. Heidegger in 1928, and, posthumously in 1939, *Erfahrung und Urteil* by L. Landgrebe). In 1940, H.L. Van Breda saved most of Husserl's *Nachlass* (including 40 000 pages of manuscripts, Husserl's private library and much of his correspondence), which today forms the core of the Husserl Archives at the Catholic University at Louvain, Belgium. Philosophical and related work in the phenomenological spirit is, beyond this, upheld by 'phenomenological societies' in several countries.

Impact on the Social and Behavioral Sciences

Husserl's own writings did not take up any topics directly relevant to the social and behavioral sciences and therefore had little influence on them during his lifetime. Yet, they greatly stimulated interest in the phenomenological method in other authors, who explored its applicability and fruitfulness within the social sciences, broadening and strengthening the phenomenological movement. Whereas the investigations of A. Reinach into the a priori foundations of civil law, and of M. Scheler into the origins of cultural development, are contributions to the philosophy of the social sciences rather than to social sciences proper, the potentiality of a phenomenological approach was recognized early in the psychology of personality (especially in the psychology of the subconscious and in psychiatry), and in experimental psychology, where it opened up new vistas for social psychology and the psychology of perception (D. Katz, A. Metzger, M. Wertheimer): the development of gestalt psychology (W. Köhler, K. Koffka) and of K. Lewin's field theory are unthinkable without the impact of Husserl's phenomenology (see also MacLeod, 1947). In the

field of sociology, the *formal sociology* (or sociological formalism) of A. Vierkandt and A. Gurwitsch have drawn explicitly on phenomenological ideas, and more recently, sociological systems theory, social constructionism and ethnomethodology have been clearly influenced by phenomenological thought (P. Berger, H. Garfinkel, E. Goffman, T. Luckmann, Niklas Luhmann, I. Srubar).

According to A. Schütz, "Husserl's signal contribution to the social sciences consists … in the wealth of his analyses pertinent to problems of the *Lebenswelt* and designed to be developed into a philosophical anthropology" (Schütz, 1959: p. 97). This development was brought about by Schütz himself, who met Husserl personally only in 1932, after publication of his own pioneering work *Der sinnhafte Aufbau der sozialen Welt* (*The Phenomenology of the Social World*, 1972), but established close intellectual contacts for the six remaining years of Husserl's life. Schütz became the towering figure of phenomenological sociology, and it is due to the considerable impact of his work that Husserl's phenomenology, beyond its historical merit of exhibiting the limits of naturalistic and atomistic thinking, has gained a lasting influence on the social and behavioral sciences of today.

See also: Hermeneutics; Phenomenology in Sociology; Phenomenology: Philosophical Aspects; Philosophy, Psychiatry, and Psychology; Psychologism, History of; Religion, Phenomenology of; Schütz, Alfred (1899–1959).

Bibliography

Bell, D., 1990. Husserl. Routledge, London.

Embree, L., et al., 1997. Encyclopedia of Phenomenology. Kluwer, Dordrecht, The Netherlands.

Herzog, M., Graumann, C.F. (Eds.), 1991. Sinn und Erfahrung. Phänomenologische Methoden in den Humanwissenschaften. Asanger, Heidelberg, Germany.

Husserl, E., 1901. Logische Untersuchungen (Trans. into English, 1970 Logical Investigations. Routledge & Kegan Paul, London, and Humanities Press, New York). Niemeyer, Halle, Germany.

Husserl, E., (rev 1976), 1950, 1952. Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie, vol. 1. Niemeyer, Halle, Germany; Biemel, W. (Ed.), vol. 2. Nijhoff, The Hague, The Netherlands; Biemel, M. (Ed.), vol. 3. Nijhoff, The Hague, The Netherlands (Trans. into English 1980–89. Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy, vol. 1. Nijhoff, The Hague, The Netherlands; vols. 2 and 3. Kluwer, Dordrecht, The Netherlands).

Husserl, E., 1939. Erfahrung und Urteil. Untersuchungen zur Genealogie der Logik. Academia, Prague (Trans. into English 1973. Experience and Judgment. Investigations in a Genealogy of Logic. Northwestern University Press, Evanston, IL, and Routledge & Kegan Paul, London).

Husserl, E., 1954. Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie: Eine Einleitung in die phänomenologische Philosophie. Nijhoff, The Hague, The Netherlands (Trans. into English 1970. The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy. Northwestern University Press, Evanston, IL) (Pts. 1 and 2 had already been published in the journal Philosophia (Belgrade) 1 (1936), pp. 77–176).

Husserl, E., 1950. Ff. Husserliana: Edmund Husserl, Gesammelte Werke. Nijhoff, The Hague, The Netherlands, and Kluwer, Dordrecht, The Netherlands.

MacLeod, R.B., 1947. The phenomenological approach to social psychology. Psychological Review 54, 193–210.

Schütz, A., 1959. Husserl's importance for the social sciences. In: Van Breda, H.L., Taminiaux, J. (Eds.), Edmund Husserl 1859–1959: Recueil Commémoratif publié à l'Occasion du Centenaire de la Naissance du Philosophe. Nijhoff, La Haye, The Netherlands, pp. 86–98.

Schütz, A., 1962–97. Collected Papers. Nijhoff, The Hague, The Netherlands (Natanson, M. (Ed.), 1962. The Problem of Social Reality, vol. I; Brodersen, A. (Ed.), 1967. Studies in Social Theory, vol. II; Schütz, I. (Ed.). 1966. Studies in Phenomenological Theory, vol. III).

Smith, B., Smith, W.D. (Eds.), 1995. The Cambridge Companion to Husserl. Cambridge University Press, Cambridge, UK.

Wolf, K.H., 1979. Phenomenology and sociology. In: Bottomore, T., Nisbet, R. (Eds.), A History of Sociological Analysis. Heinemann, London, pp. 499–556.

Hutterite Demography

Douglas L Anderton, University of South Carolina, Columbia, SC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Hutterites, a pronatalist religious sect that showed no evidence of intentional fertility limitation prior to 1950, have been widely used as a standard for high, or even 'natural,' fertility. A common and communal lifestyle among Hutterites gives further reason to expect less sociodemographic variation in fertility than many high-fertility populations and enhances their value as a historical fertility standard. The early Hutterite population has been the most commonly used age-specific fertility pattern, as a standard representing natural fertility, in a wide variety of demographic indices, simulations, and models. In turn, methods incorporating the Hutterite fertility standard have facilitated the study of departures from uncontrolled fertility, infecundability, and other biological aspects of fertility.

The unique demography of the Hutterites – a contemporary Mennonite sect of Germanic origin who lives in the Great Plains of the United States and Canada – has played an important role in the development of demographic theory and methods. Prior to 1950, they appear to have practiced no birth control for religious reasons and their high fertility levels give reason to believe they did not intentionally control their fertility (Eaton and Mayer, 1953). Their communally shared lifestyle, including the sharing of material resources, is also presumed to limit social or cultural variations in Hutterite fertility (Larsen and Vaupel, 1993). Because of their uniformly high and apparently uncontrolled fertility, studies of the patterns and determinants of fertility among the Hutterites have helped to shape our theoretical expectations for populations that do not make any conscious efforts to limit childbearing. And, as a result, the Hutterites have also been widely used in demographic methods as a fertility, and biological infertility, standard to which other populations are compared.

Hutterite Fertility

Many different estimates of fertility and fertility determinants among the Hutterites have been published in different studies (see Larsen and Vaupel, 1993). All of these studies agree, however, that Hutterite fertility is exceptionally high. Eaton and Mayer (1953) reported a median completed family size of 10.4 children among married women over the age of 45 in 1950. Recalculations using the data published by Eaton and Mayer report a lower completed family size with a median of 9.4 children, and a mean of 8.97 among the same women (Lang and Göhlen, 1985). These estimates are consistent with the modal completed family size of 10 children, and the maximum of 15 children, reported for the Hutterite sample analyzed by Larsen and Vaupel (1993). Even in this very high-fertility population, however, fecundability and the probability of conception decline substantially with age. The monthly probability of conception is estimated to decline by about two-thirds between ages 20 and 35 (Larsen and Vaupel, 1993), likely reflecting age-related influences such as higher probabilities of fetal loss, lower coital frequency, longer postpartum amenorrhea, etc., among older age women. Nonetheless, the age-specific influences on Hutterite fertility resulting in a decline of fecundability across ages appear to be largely biological in nature. Compared with other populations, the Hutterite population has exceptionally high fertility levels and the age-related decline in fecundability offers no apparent evidence of intentional fertility limitation.

The Hutterites as a Natural Fertility Standard

Because of their doctrines and behavior, the Hutterites are presumed to represent a *natural fertility* population. That is, as Henri (1961) argued, their childbearing patterns reflect no efforts to limit fertility bound to the number of children already born, i.e., parity-*specific* fertility limitation. In fact, the Hutterites were one of the populations Henri (1961) used to define and illustrate the concept of natural fertility. He also ensured a special standing for the Hutterites as a high-fertility standard when he found that they had the highest age-specific fertility levels of 13 historical populations he studied, all of which appeared to be natural fertility populations.

The Hutterites as a 'Standard Natural Fertility Population' – Coale's 'I' Indices

Hutterite fertility levels and age-specific fertility patterns have been used as standards in some of the most widely used fertility measures of recent years. The Princeton European Fertility Project (Coale and Watkins, 1986) popularized the use of Coale's (1969) 'I' indices, which compare fertility in a population with that of the Hutterites through indirect standardization. Coale's index of overall fertility (I_f) compares the number of births to women in a population with the number of births these women would have had if they had experienced the same high age-specific fertility rates as married Hutterite women (i.e., 0.550 births per year among women aged 20–24, 0.502 to those aged 25–39, 0.447 to those aged 30–34, 0.406 to those aged 35–39, 0.222 to those aged 40–44, and 0.061 to those aged 45–59, see Henri, 1961). Coale's index of marital fertility (I_g) similarly compares births

to married women with the fertility they would have had if they had experienced the Hutterite age-specific marital fertility rates. And, Coale's index of illegitimate fertility (I_h) draws the same comparison for nonmarried women, again using Hutterite age-specific marital fertility rates as a standard. A final fourth index is intended to capture effects of the proportion married upon fertility (I_m) through comparing the fertility married women would have with that all women would have, if both experienced Hutterite age-specific marital fertility rates.

Age Patterns of Hutterite Fertility as a Standard – Coale and Trussell's '*M*' and '*m*' Indices

Coale and Trussell's (1978) more formal mathematical modeling of age-specific marital fertility control also incorporates the fertility pattern of the Hutterites and Henri's (1961) other natural fertility populations as a standard for comparison. Coale and Trussell used 10 of Henri's natural fertility populations, including the Hutterites, to specify a standard age-specific pattern for natural fertility (*see* Demographic Models). The Coale–Trussell model attempts to parameterize the differences in overall levels and age patterns of fertility from these natural fertility expectations in populations with varying levels of fertility control. The level of fertility is then captured by two parameters of the model, Coale and Trussell's '*M*' index of fertility levels and their '*m*' index indicating departures from a natural fertility age pattern (1978).

Hutterite Fertility Determinants in Micro- and Macrosimulations

Not surprisingly, demographic simulation models have also made use of Hutterite fertility standards. Ridley and Sheps' (1966) initial (i.e., Cohort A) microsimulation of fertility, for example, used values of fecundability and postpartum non-susceptibility, which were derived from the Hutterite population. Bongaarts' more recent (1976) macrosimulation of fertility also incorporated Henri's (1961) natural fertility estimates as standards for age-specific patterns of fecundity and infecundity associated with sterility. In these simulations, the Hutterite fertility standards are used to represent biological fertility parameters in the absence of known parity-specific fertility control.

Limitations of the Hutterite Fertility Standard

As the theoretical definition of natural fertility has been challenged, so have the measures of fertility control which have utilized the Hutterite fertility standard. Comparisons with a Hutterite standard population, such as Coale's I_g index, may reflect differences in fertility, which do not arise from parity-specific control (Santow, 1995). And, as in the case of Coale's '*M*' and '*m*' indices, such comparisons may also fail to capture the heterogeneity of fertility behavior in populations being compared with the natural fertility standards (Okun, 1994), including the possibility of fertility control, which is not closely tied to parity such as birth spacing. And, in the last half of the twentieth century, Hutterite fertility began to show evidence of fertility control (Peter, 1987; White, 2002). Thus, one limitation of the Hutterite standard is the fact that it represents a time and culture, which now belong to history and may diverge over time from a hypothetical contemporary natural fertility population.

Although comparative indices and methods face many of the same limitations as any indirect standardization, incorporating the Hutterites as a constant standard population for empirically based comparisons has, nonetheless, been of considerable advantage to demographers. Use of the Hutterite standard has allowed meaningful comparisons of fertility levels and patterns across widely different historical populations. The Hutterite standard has also provided indices with a clearly interpretable and unambiguous reference to an empirically observed high-fertility standard.

The Hutterites and Biodemography

Underlying their use as a high-fertility standard in many different methods, the Hutterites in and of themselves offer a unique opportunity for demographers to explore biological aspects of fertility and fecundity. An absence of any evident parity-specific fertility control, along with very high fertility levels, makes the Hutterite population as the one that may be used to study biological determinants of fertility without the confounding influence of fertility limiting behaviors. However, a troubling concern in the study of historical populations that lack widespread evidence of fertility control is the extent to which there may be many subgroups within the population with different fertility levels and patterns arising from different behavior and sociocultural circumstances. The Hutterites' communal lifestyle and uniformly high fertility give no reason to suspect wide differences in fecundability and fertility from social or cultural forces. And the uniformity of the population in many demographic behaviors (e.g., universal marriage and early ages at marriage, short durations of breast-feeding and early weaning, short birth intervals, absence of induced abortion, low levels of premarital conception, etc.) is likely to minimize many sources of variation in fertility compared with other populations. This relative homogeneity, or absence of subgroup variation, is especially important to the recent use (and assumptions) of event history methods to estimate the underlying biological patterns of fecundability among the Hutterites. This allows researchers estimating fecundability with Hutterite data to assume the underlying patterns and variations they discover arise from biological variations (see Larsen and Vaupel, 1993, for a review and example of such studies).

As further studies provide more refined estimates of Hutterite fertility and fecundity, these will likely continue to serve as empirically grounded standards for our demographic expectations in the absence of intentional fertility limitation. The theoretical importance of these expectations to demography is evident in the significant role, which studies of the Hutterites have had in shaping the concept of natural fertility and debates over the nature of fertility limitation. Their methodological import is clear in the wide variety of methods already grounded in the Hutterite fertility standard.

See also: Demographic Measurement: General Issues and Measures of Fertility; Demography, Early History of; Demography: History Since 1900; Fertility Theory; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Fertility and Culture: Anthropological Insights.

Bibliography

Bongaarts, J., 1976. A dynamic model of the reproduction process. Population Studies 30 (1), 59–73.

Coale, A.J., 1969. The decline of fertility in Europe from the French Revolution to World War II. In: Behrman, S.J., et al. (Eds.), Fertility and Family Planning: A World View. University of Michigan Press, Ann Arbor, MI.

Coale, A.J., Trussell, T.J., 1978. Technical note: finding the two parameters that specify a model schedule of marital fertility. Population Index 44 (2), 203–213.

Coale, A.J., Watkins, S.C. (Eds.), 1986. The Decline of Fertility in Europe. Princeton University Press, Princeton, NJ.

Eaton, J.W., Mayer, A.J., 1953. The social biology of very high fertility among the Hutterites: the demography of a unique population. Human Biology 25, 206–264.

Henri, L., 1961. Some data on natural fertility. Eugenics Quarterly 8, 81–91.

Lang, H., Göhlen, R., 1985. Completed fertility of the Hutterites: a revision. Current Anthropology 26 (3), 395.

Larsen, U., Vaupel, J.W., 1993. Hutterite fecundability by age and parity: strategies for frailty modeling of event histories. Demography 30 (1), 81–102.

Okun, B., 1994. Evaluating methods for detecting fertility control: Coale and Trussell's model and cohort parity analysis. Population Studies 48 (1), 193–222.

Peter, K.A., 1987. The Dynamics of Hutterite Society: An Analytical Approach. University of Alberta Press, Altona.

Ridley, J.C., Sheps, M.C., 1966. An analytic simulation model of human reproduction with demographic and biological components. Population Studies 19 (3), 297–310.

Santow, G., 1995. Coitus interruptus and the control of natural fertility. Population Studies 49 (1), 19–43.

White, K.J.C., 2002. Declining fertility among north American Hutterites: the use of birth control within a Dariusleut colony. Social Biology 49 (1–2), 58–73.

Hybridity

Nestor G Canclini, Autonomous Metropolitan University, Mexico and National System of Researchers, Mexico

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term hybrid has been employed since ancient times to name processes of syncretism in migrations and colonization, but it has historically been more significant in biology. Since the end of the twentieth century, it has been widely used in social sciences and technology. The intensification of globalized interculturality in the arts and music, of intercontinental distribution of communication industries and of population displacements has led to an increase in studies on hybridity in the fields of anthropology, cultural, aesthetic and religious studies, and cultural politics. At times, this term takes priority over such notions as mestizaje, syncretism, creolization, and transculturation. Studies on hybridity open the way to gaining a deeper critical insight into the diversified development of societies, their mixing patterns, and the intercultural movements of hegemony and resistance.

'Hybridity,' originally a biological term, has been used by scholars in the social sciences, literary, artistic, and cultural studies to designate processes in which discrete social practices or structures, previously existing in separate form, are combined to generate new structures, objects, and practices.

Even if the study of hybridity processes has only been systematized during the last few decades, cultural mixing – or hybridization – has existed since the first exchanges among peoples, the first migrations, and wars of conquest. The Tower of Babel is a story of *mestizaje*. Plutarch used the term syncretism for political alliances. Pliny the Elder mentioned hybridization when referring to the migrants to Rome in his day. The historians of the Christian evangelization of Africa, Asia, and the Americas (G. Prius, S. Gruzinski) have recorded that converts did not abandon their traditional beliefs but reelaborated and fused them with those of the western missionaries.

The research into hybridity processes has become more extensive during the 1980s and 1990s, particularly in English and Spanish, whereas in other languages, such as French, the term coexists with similar concepts like *mestizaje* and syncretism, favored by French scholars. Some authors place these terms under 'hybridization,' thereby granting it the flexibility to include processes of intercultural confluence.

A study of various databases in July 2012 in order to compare references to hybridity in social sciences publications from 1990 to 2000 with those from 2000 to 2012 reveals a notable increase during the second period. Jstor in English brought up 1366 references for the first period and 2934 for the second. According to Scopus, articles on hybridity have risen from 94 to 730 during the respective periods. Whereas Google Scholar identifies 9290 references in English during the former period and 21 700 during the latter, its Spanish results show a decline from 1650 during the first period to 1400 during the second.

The Term Hybridization in Biology, Social Sciences, and Technologies

The use of hybridity in the humanities and the social sciences has been questioned due to the biological origin of the concept. Some point to the dangers of transferring the sterility commonly associated with this term to society and culture. Whereas this criticism is still found in recent studies, it is an extension of a nineteenth-century belief, when hybridization was looked upon with suspicion because it was believed to harm social development. However, this opinion has changed since 1870 with Mendel's theory of the enrichment produced by genetic crossbreeding and its examination during the twentieth century (Olby, 1984). In strict terms, biology, particularly botany, teems with fertile hybridizations that take advantage of cell characteristics of different plants to improve the growth, resistance, quality, and economic and nutritional value of food stuffs derived from these combinations (Callender, 1988). Hybridization of human DNA with bacteria to produce protein is also common; this is, in fact, the main means to generate insulin.

The biological background of the term does not explain its acceptance in the social sciences. Like other concepts imported from biology, such as *reproduction*, which had to be redeveloped in order to speak about social, economic, and cultural reproduction, the theoretical solidity and explanatory power of hybridity in social sciences depend upon the manner in which the term is relocated within the networks of specific social and cultural concepts.

On the one hand, hybridity is linked with more traditional notions of social thought that allude to the separation and mix among cultures and strata: race, ethnicity, *mestizaje*, miscegenation, syncretism, identity, multiculturalism, heterogeneity, creolization, transculturation, bricolage. On the other, the understanding of this term, its treatment and epistemological consistency depend upon the manner in which different authors link the concept with terms and theories, such as tradition, modernity, contradiction, globalization, postmodernity, and border studies.

Several disciplines have played a role in the sociocultural construction of hybridity. As a linguistic process, it was developed by Mikhail Bakhtin, among others, who used it to describe the coexistence of elite and popular languages in the same group, or even in a single speaker, and by Homi K. Bhabha, who defines hybridity as a metonym of presence and places it amidst power relations, noting differences between colonial and resistance hybridization, depending

on whether the linguistic and literary discourses serve the purposes of domination or subversion (Bhabha, 1994). Some historians of the Americas speak of hybridity to explain the interweaving of pre-Columbian imagery with the imagery of the Spanish colonizers and with that of later modernity cultures (Gruzinski, 1999).

From the end of the twentieth century, various disciplines and schools of cultural thought have adopted the concept of hybridity for creative processes in the visual arts and literature. It has become a descriptor of heterogeneous esthetic repertoires arising from the mixing of resources, such as tradition and modernity, artisanry and technology, local cultures and transnational iconographies (García Canclini, 1995; Papastergiadis, 2005). It is also useful for recognizing the increasingly important hybrid musical forms and genres, such as jazz, reggae and fusions of rock with Celtic music, salsa, tango, etc. Its reach even destabilizes biological orders and classifications, such as transexuality. Likewise, hybridity in education, so some suggest, makes children aware of multicultural creativity in order to steer them away from racist attitudes (Gallager-Geurtsen, 2009). This term is used, moreover, for automobiles that combine electric energy with internal combustion, gastronomy that fuses ethnic and national traditions and eclectic architectures, such as those that mix styles from different epochs within the same building, following the logic of postmodernism. Is there a danger that the elasticity of this concept so undermines its rigor that it eventually becomes banal? For this reason, it is worth bearing in mind the debates of those who reject the notion of hybridization and of those who compare it with other similar concepts.

Mestizaje, Syncretism, Transculturation, Creolization, and Hybridization

A number of historians and anthropologists have shown the decisive role of *mestizaje* in certain areas, such as the Mediterranean, since the classical period of Greece (Laplantine and Nouss, 1997). But the more intense – or researched – periods of interethnic and intercultural mixing occurred at the time of European expansion across the Americas.

Mestizaje became the prevailing process within the cultures of the so-called New World, because of the mixing of colonizers from Spain and Portugal, and, later, from Britain and France with the indigenous peoples and, subsequently, with African slaves. Nowadays, less than 10% of the population of Latin America is descended from indigenous peoples alone (this percentage decreases in the USA and Canada). Communities of European descent that have not mixed with indigenous peoples are also a minority. Thus, the makeup of all of the Americas requires the notion of *mestizaje*, both biologically, in terms of the production of phenotypes from genetic crossbreeding, and culturally, in terms of the fusion of European and indigenous American customs, beliefs, and modes of thought.

For a very long time, studies foregrounded the physiognomic and chromatic aspects of *mestizaje*. Skin color and physical features are still critical to the ideological construction of subordination and discrimination; however, in the social sciences and in more democratic political thought, *mestizaje* currently centers on the cultural dimension of the combination of identities. In anthropology, cultural studies, and cultural policy making, the issue is presented as devising ways to share multicultural life styles (Kymlicka, 1995).

Of the diverse crossings among cultures, fusions of religions have received the most attention. Syncretism is the specialists' term for this issue. European expansionism in Africa, Latin America, and Asia gave rise to multiple mixed forms: heterogeneous beliefs fused and new divinities emerged (the Virgin of Guadalupe in Mexico, the combination of African ancestors, indigenous figures, and Catholic saints in Brazilian Umbanda). Practically, every religion has fundamentalist tendencies that seek to defend pure forms, but the magnitude of migration and the transcontinental circulation of rituals and beliefs during the eighteenth and nineteenth centuries stressed these hybridizations and increased tolerance toward them to such a point that in countries such as Brazil, Cuba, Haiti, and the USA, it was frequent to find a double or triple religious affiliation. For example, a Catholic might participate in an African-American ritual or in a New Age ceremony. If we broaden the concept of syncretism to include affiliation to several belief systems as well as religion, the phenomenon grows remarkably, particularly among the many people who turn to indigenous or Asian medicine for certain ailments, to allopathic medicine for others, and even to Catholic or Pentecostal rituals. These syncretic practices are often combined with musical fusion and multicultural social organization systems, as is the case in Cuban *Santería*, Haitian Voodoo, and Brazilian *Candomblé*.

It was some of these phenomena that concerned the Cuban historian and anthropologist, Fernando Ortiz, when he proposed the term *transculturation* to designate "processes in which features are not mechanically aggregated, but give way to a new phenomenon, independent and original." Bronislav Malinowski's endorsement of that utterance in an introduction to Ortiz's book, and its future development by literary critics Angel Rama and Mary Louise Pratt in their studies of travel writing, are still appealing; particularly in literature and in Anglo-Saxon and Latin American cultural studies.

The word creolization has also been used to refer to intercultural combinations, especially in anthropology and linguistics. It designates, *sensu stricto*, the language and culture created by variations on a dominant language with other languages in the context of the slave trade. It includes the variations of French in the Americas and the Caribbean (Louisiana, Haiti, Guadeloupe, Martinique) and in the Indian Ocean (Reunion, Mauritius), or of Portuguese in Africa (Guinea, Cape Verde), in the Caribbean (Curacao) and Asia (India, Sri Lanka). But whereas it presents paradigmatic tensions between orality and writing, elite and popular sectors, center and periphery, in a diversity *continuum*, Hannerz suggests its usage could be extended to the transnational field to name "processes of cultural confluence" characterized "by power, prestige, and material resources inequality" (Hannerz, 1997). He is not the only author to draw attention to the inequality and discontinuity of hybridity processes, but by emphasizing the significance of examining the increasing movement between center and periphery and the asymmetries among markets, states, and educational levels he challenges

those who consider *mestizaje* to be, simply, intercultural homogenization and reconciliation (Hannerz, 1996).

Hybridization, Modernity, and Globalization

Hybridization

Whereas scholars often use the terms mentioned above almost as synonyms, *mestizaje* takes precedence in studies on the mixing of 'races' or ethnic groups, and syncretism refers to religious fusions or traditional symbolic movements. Besides the fact that *mestizo* has no English equivalent (some Anglophone scholars include the Spanish term in their studies), research carried out during the 1980s in Spanish, English, and occasionally in French and Portuguese adopted 'hybridity' to include diverse intercultural mixes, such as ethnic *mestizaje*, syncretism, in relation to beliefs, and other modern fusions of arts and crafts with industry, of elite culture with popular and traditional cultures and with the mass media (Bhabha, 1994; García Canclini, 1995; Bernd and de Grandis, 1995).

Modernity

Modernity, of which nation-states are an expression, has coherently organized – at times in reality, at others in appearance – the diversity of cultures. There are fewer than 200 nation-states which comprise more than 600 languages and some 5000 ethnic groups. These states have had, at once, to organize multiculturality by means of diverse policies and to facilitate communication among the ethnic groups and the different languages in each territory. Even in the cases in which a single language is imposed upon the 20, 50, or 100 communities in the nation, this common language adopts diverse modulations in the English- or Spanish-speaking country when it is combined or hybridized with the words, syntactic structures, and semantic references of the speakers of other tongues.

In the meantime, two other characteristics of modern development, transnational expansion of capitalism and of mass communication technologies, advance relationships and exchanges between local traditions and cultures of modernity. The crafts and celebrations of peasant groups travel, together with their producers, to cities and often to other countries, or they enter into contact with foreign tourists. Ethnic music, linked to specific rituals of a certain group, is not only recorded on CDs, movies, and videos to be distributed worldwide; it is also mixed with classical and contemporary music, with jazz, rock, salsa, and other local and regional kinds of music. This is how the *chicha*, a hybrid of Andean and Caribbean rhythms, was produced. Other examples include the jazzy reinterpretation of Mozart by the Afro-Cuban group Irakere, the recreation of English, Indian, and other music by the Beatles, Peter Gabriel, and many other musicians.

The strict borders established by nation-states have become blurred. Few cultures today are stable units with precise boundaries defining the settled area of the territory. But the multiple opportunities to hybridize imply neither indetermination nor unrestricted freedom. Hybridization occurs under specific historical and social conditions, in the midst of production and consumption systems that sometimes operate as constraints, as can be witnessed by the lives of many migrants. The city is an exemplary social entity that supports and conditions hybridity. The multilingual, multicultural megalopolises, for example, Paris in the nineteenth and the twentieth centuries, Vienna at the beginning of the twentieth century, London, Berlin, New York, Los Angeles, Buenos Aires, Sao Paulo, Mexico, and Hong Kong are studied as centers where hybridization kindles greater conflict and cultural creativity (Appadurai, 1996; Hannerz, 1996).

The wide range of processes of exchange and mixing involved in hybridization has led to the construction of typologies. Peter Burke's study is one of the most meticulous accounts of the diversity of meanings and inequalities. Supporting his proposal with cultural encounters from every continent, this historian distinguishes, firstly, intercultural *imitation* from *appropriation* and *reuse*; secondly, he mentions *borrowing* and, finally, his alternative: *bilateral transculturation*, *accommodation* and *negotiation*. It is, as he observes, the processes of accommodation and appropriation that provide the most diverse creativity and that require a reconsideration of *translation* in order to introduce cultural concepts to other linguistic communities, which may not have heard of them (Burke, 2009).

Globalization

Globalization stresses these trends of modernity by creating world markets of material goods and money, messages, and migrants (workers, tourists, exiles). These constant flows and interactions undermine borders, frontier posts, and the autonomy of local traditions. At the same time, they cultivate many more forms of hybridization than they did in the past, such as commercial, communicational, and consumer practices. What must be added to the modalities embodying hybridization – migration, travel, and educational integration policies introduced by nation-states – are the mixtures created by cultural industries. Industrialization encompasses not only the mass media (radio, movies, television, video), but also the industrial production for massive audiences of international exhibitions, touring folkloric groups, and globalized theme parks such as Disney's (Harvey, 1996).

However, the leading globalization analysts note that it not only integrates and generates *mestizajes* but also segregates, produces new inequalities, and stimulates controversial responses. Identity alliances emerge (Latins or Arabs in the USA or in Europe) as do regional organizations (the European Union, NAFTA, Mercosur) that aim, in part, to give different groups the strength to counteract the indiscriminate interaction of globalism (Beck, 2000). Some group members find the means in these alliances to resist both globalization and hybridization.

Homogeneity vs Differences?

Although most regional agreements between countries only regulate free trade, economic and financial globalization increases interdependence and reciprocal influences between cultural traditions and means of expression, formerly so distant from each other. The global movement of goods and messages stimulates creative practices and consumer habits in

western countries to include African or Asian objects, forms, and esthetics in their repertoire. Conversely, African and Asian countries are no longer the destination for a small number of missionaries, traders, anthropologists and adventurers, but for large numbers of material and symbolic objects produced in the West. This hemispheric interaction blurs the profiles that had, in former centuries, established sociocultural autonomy in each region. Contributing factors include not only the global circulation of capitals, material, and media goods but also the geographic expansion of migrations. Two examples illustrate this: many more Arabs and Koreans are now settling in western societies and Latin Americans are moving their point of reference from the United States of America and Europe to seek employment in Arab and Asian nations as well.

These concepts have also extended their horizon of application. The notion of diaspora, utilized historically to describe the dispersion of Jewish, Greek, and Armenian peoples, is now used, as Ien Ang observes, to analyze almost any group living outside its country of origin. The migrants' increased possibilities of communication with their original homeland through improved transport, the mass media, and digital networks means that they no longer see themselves as minorities within the host nations but as members of global diasporas; the 'overseas Chinese,' or the dispersal of the Chinese throughout the world, provides an eloquent example (Ang, 2003). Whereas these groups engage in processes of hybridization with their new surroundings, they reaffirm their customs, at the same time, as they are often in touch with their country of origin. Hybridization destabilizes identities, as it blurs the boundaries between the Chinese and the non-Chinese, Islamists, and westerners, but the new institutions that organize diaspora, such as the transboundary trade networks and web pages, create a form of transnational nationalism.

Recent globalization produces a more complex combination than in the past between homogenizing tendencies and affirmation of differences. Technological advances and the potential to have customers all over the world have engendered movements of homogenization, such as equalization, to use a musical term: electronic devices alter the timbre variations and local melodic features for the purpose of stylizing them and eliminating whatever might sound discordant to listeners from other cultures (de Carvalho, 1995). But readings that attempt to reconcile diversities reveal the covert inequalities in the production, circulation, and consumption of culture. An analysis of the strategies of commercial hybridization should locate them in what Ulf Hannerz calls the "political economy of culture inherent to the mestizo *continuum*"; that is, the uneven distribution of centers and peripheries, and even "the coexistence of various *continuums* instead of a single and exclusive *continuum*" (Hannerz, 1997: pp. 115–116). That the countries of the center pay more attention to underdeveloped countries, and that the migrants and Latin American messages are hybridized in the center (ethnic musical discs recorded in New York, Miami, Mexico, and Buenos Aires, and these cities converted, at the same time, by MTV into the capitals of Latin American rock) does not signify the abolition of center and periphery.

Unlike the conventional debates about the issue of homogeneity versus differences, the contemporary scene inspires a consideration of the zones of exchange, instead of choosing one or the other. The question is how to find a common denominator and how members of different cultures coexist, if they are only partly hybridized. Which settings foreground this reformulation of the problem most readily? Metropolises and cosmopolitan cities, ports and borders have been throughout history typical cases of intersections among hybrid cultures; they also sustain their differences and reinvent traditions, languages, gastronomy, and the arts. At the present time, the Internet has strengthened multicultural coexistence and created an increase in worldwide conflicts. As many authors write, we may experience intercultural tensions anywhere; any place can become a border.

Burke suggests that throughout history there have been four ways of responding to this situation: (1) *the imitation of the foreign models*, based upon admiration, tolerance, and in many other cases upon an all too harmonious vision of the exchanges between differences; (2) *resistance* to any external force perceived as an invasion, which has antecedents in the distant past, such as the prohibition of foreign books, studying abroad, when Philip II was on the throne in Spain (1559), or the Muslims' banning of the printing press until 1800, because it threatened the oral transmission of knowledge; (3) *cultural purification*, or a reaction against the acceptance of the foreign and a deconstruction of what has been borrowed and assimilated, that is, a dehybridization such as that which occurred during Nazism, the French response to the *défi americain* and the more recent measures taken by European countries requiring that immigrants observe the traditions and language of their adopted country; and (4) *cultural segregation*, which applies, for example, to some Asian groups who, finding it difficult to keep their entire territory uncontaminated, select areas to remain free from westernization or defend the indigenes as a source from which to regenerate communities within capitalism; it also applies to those such as the Japanese who lead a double life, by adhering to a foreign culture at work and turning to their own traditions during their spare time (Burke, 2009).

In sum, the increase in exchanges at this stage of globalization is not a one-way street. There are homogenizing tendencies that include diverse labor and consumer practices as compatible parts of the same mode of production. Contrary to the fear of the Americanization of the planet or the exclusive domination of the English language, several other languages, such as Chinese, Spanish, and Arabic are enjoying a period of expansion. Moreover, antiglobalization and deglobalization movements are strengthening local traditions, often acknowledging, nonetheless, that they are making productive use of appropriations from external sources.

Studies on '*structural hybridism*' (Nederveen Pieterse, 2000) of the contemporary world thus contribute to a social theory of nonunidirectional globalization, understood not as a homogenization process but as diversified development through the mixings and coexistence of cultural traditions that never quite fuse. It is worth recalling the unequal distribution of development and access to the processes of hybridization, because it depends on the position of each culture or group in power hierarchies. The combinatory faculties of producers and consumers are different in the movements of *endogenous hybridization*, or the local and

self-managing circuits, from those of coercive *heteronomous hybridization* that concentrate on appropriating characteristics from many cultures and making them circulate among a few transnational enterprises that produce messages and goods, administration, and editing of multicultural meaning.

Just as this occurs with the *spatial* distribution of hybridization strategies, so Brian Stross has described the differences of the '*hybridization cycles*.' He states that we pass, in historical terms, from more heterogenic to more homogeneous forms, and then again to other more heterogenic forms, depending on our ability to construct hegemony and establish a given state of intercultural combinations.

The Future

Beyond those naive conceptions of hybridization as processes of universal reconciliation, it is clear by now that the greatest interdependence between cultures creates, at once, multiple conflicts and rapprochement between symbolic and material practices. By no means is this the case only outside the homeland: we have frontier experiences, even in our place of birth.

Having passed from societies that impose restraints on multiculturality (each minority in its own quarter) to a more open and diffused interculturality, anthropologists and cultural analysts argue that we should focus on the zones of exchange, rather than on identity studies (ethnic and national). Do the concepts of hybridization, *mestizaje* and syncretism help us to organize conceptually those more complex processes of interaction? They do, so long as we do not expect a fixed meaning for these concepts and recognize the range of situations that might occur on the zones of exchange. Concepts are, as Mieke Bal writes, the sites of debate, where consciousness of differences and mixings is raised, and, subsequently, attempts to reach a reciprocal understanding are proposed. Agreeing doesn't mean agreeing on content, but agreeing on the basic rules of the game: if you use a concept at all, you use it in a particular way, so that you can meaningfully disagree on content (Bal, 2002).

Specialists in the fields of anthropology and cultural studies maintain that the evident relevance of hybridization requires changes to the methods normally used in the social sciences. It is no longer possible to analyze large units, that is, nations, states, transnational enterprises, without noticing their internal diversities and modulations, as often happens in disciplines that analyze at macro level, such as sociology and economics. Nor is it acceptable that anthropology, which has traditionally studied small communities, should continue solely to identify their homogeneity, thereby ignoring the heterogeneity and hybridization caused by the transnational goods and messages that intersect these communities. A commitment to a political negotiation of diversity and hybridizations requires a '*mestizo* way of thinking' (Laplantine and Nouss, 1997; Gruzinski, 1999).

Fundamentalisms are not the only opposition to religious syncretism and intercultural *mestizaje*. Many reject hybridization because it generates insecurities within cultures and conspires against their ethnocentric self-esteem. It also challenges trends in modern analytical thought accustomed to conceptual pairings: civilized–wild, national–foreign. This schematism ignores many ways in which culture is shared nowadays, such as someone who is Brazilian by nationality, Portuguese by tongue, Russian or Japanese by origin, and Catholic or a believer in an African-American religion.

In the future, hybrid thought will not work so much with radically different homogenous compact units as it will with intersections, transitions, and transactions; it will express less concern for identity than for heterogeneity and multiple senses of belonging. In order to gain a greater insight into the contemporary world, it is essential to study how cultures mix, what are the rhythms and modalities of this mixing, what they lose and gain in the process, which modes of coexistence they have chosen, and how they conduct their negotiations to sustain their differences. Moreover, we must understand hybridity and learn to be cosmopolitan 'multicultural gourmets,' who travel through diverse cultures and savor the differences, if we are to comprehend globalization and the ethnic and national crises in societies. In sum, if we are to comprehend society.

See also: Creolization: Sociocultural Aspects; Globalization and World Culture; Multiculturalism, Anthropology of; Specialization and Recombination of Specialties in the Social Sciences; Syncretism.

Bibliography

Ang, I., June 2003. Together-in-difference: beyond diaspora, into hybridity. Asian Studies, Review 27 (2), 141–154.
Appadurai, A., 1996. Modernity at Large: Cultural Dimensions of Globalization. University of Minnesota Press, Minneapolis, MI.
Bal, M., 2002. Travelling Concepts in the Humanities: A Rough Guide. University of Toronto Press, Canada.
Barbero, J.M., 1987. De Los Medios a Las Mediaciones. Ediciones G., Gil, México.
Beck, U., 2000. What Is Globalization? Polity Press, Oxford, UK.
Bernand, C., 1993. Altérités et métissages hispano-américains. In: Descamps, C. (Ed.), Amériques Latines: Une Altérité. Editions du Centre Pompidou, Paris.
Bernd, S., de Grandis, R., 1995. Imprivesiveis Americas: Questóes de Híbrida (ao Cultural, nas Americas). Sagra-Luzzatto, Porto Alegre, Brazil.
Bhabha, H.K., 1994. The Location of Culture. Routledge, London and New York.
Burke, P., 2009. Cultural Hybridity. Polity Press, Cambridge.
Callender, L.A., 1988. Gregor Mendel: an opponent of descent with modification. History of Science XVI.
Canclini, N.G., 1995. Hybrid Cultures. University of Minnesota Press, Minneapolis, MI.
Carvalho, José Jorge de, 1995. Hacia una etnografía de la sensibilidad musical contemporánea. UnB, Departamento de antropología, Brasilia. (Serie Antropología).
Gallager-Geurtsen, T., 2009. Inspiring hybridity: a call to engage with (in) global flows of the multicultural classroom. October–December. Multicultural Perspectives 11 (4), 200–203.
Gruzinski, S., 1999. La pensée métisse. Fayard, Paris.
Hannerz, U., 1996. Transnational Connections. Routledge, London.
Hannerz, U., 1997. Fluxos, fronteiras, híbridos: Palavras-chave da antropología, transnacional. Manet, Rio de Janeiro 3 (1), 7–39.
Harvey, P., 1996. Hybrids of Modernity. Routledge, London.
Kymlicka, W., 1995. Multicultural Citizenship. Clarendon Press, Oxford, UK.
Laplantine, F., Nouss, A., 1997. Le Méttissage. Dominos, Paris.
Nederveen Pieterse, J. (Ed.), 2000. Global Futures: Shaping Globalization. Zed Books, New York.
Olby, R.C., 1984. Historiographical problems in the history of genetics. Rivista di Stork della Scienza 1 (1), 25–38.
Papastergiadis, Nikos, August 2005. Hybridity and ambivalence: places and flows in contemporary, art an culture. Theory, Culture & Society 22 (4), 39–64.
Polar, A.C., 1997. Mestizaje e hibridez: los riesgos de las metáforas. Apuntes (Miscegenation and Hybridity: The Dangers of Metaphor. Notes). Revista Iberoamericana LXIII (180), 341–344.

Hypertension: Psychosocial Aspects

Dmitry M Davydov, Russian Institute for Advanced Study, Sholokhov Moscow State University for the Humanities, Moscow, Russia; and Institute of General Pathology and Pathophysiology, Russian Academy of Medical Sciences (RAMS), Moscow, Russia
David Shapiro, University of California, Los Angeles (UCLA), Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

From the standpoint of the social and behavioral sciences, hypertension has attracted a great deal of attention. It forms the backbone of the growing fields of health psychology and behavioral medicine. The evidence is supportive of the role and significance of psychosocial factors in the etiology, course, and treatment of hypertension. This knowledge has the potential for deepening our understanding of the mechanisms involved in hypertension, improving the effectiveness of both drug and non-drug treatment, and facilitating efforts for prevention. There is great need for further systematic research and especially for theoretical and empirical integration of psychosocial and biological knowledge of the disorder.

Hypertension is a complex disorder of physiological regulation determining increases in blood pressure that stay high over time. Leaving aside the less common cases of so-called secondary hypertension, the causes of which are well known (e.g., endocrine diseases, kidney diseases, drugs, and tumors), about 95% of cases of hypertension are likely to be the consequence of an interaction between many environmental and individual factors. US surveys indicate an age-adjusted prevalence of hypertension (>140/90 mmHg) of between 28 and 30% in adults aged 18 and over. Interest in the role of psychosocial factors in hypertension has a long history beginning with theories about social stress, affective processes, and individual behavioral dispositions in its etiology. Noteworthy are the early psychosomatic writings about the significance of anger expression and suppression for high blood pressure.

Direct and Indirect Evidence

Numerous studies provide evidence of a definite role for psychosocial factors in hypertension. Direct evidence is derived from cross-sectional studies of individuals with hypertension compared to control samples and longitudinal studies to determine what factors are associated with the development of hypertension. Indirect evidence comes primarily from studies of psychosocial factors affecting blood pressure in healthy individuals with or without various predisposing characteristics. Psychosocial factors are also relevant to the degree of adherence to medications, engagement in health-related behaviors affecting (buffering or increasing) risk for hypertension, and effectiveness of antihypertensive medications and of nondrug methods of managing hypertension based on psychosocial concepts and methods.

Blood Pressure Reactivity and Hypertension

Elevations in blood pressure can be elicited by many different stimuli and circumstances involving psychological and social factors. Examples are exposure to noxious, painful, threatening, or otherwise stressful stimuli and events, participation in demanding tasks requiring repeated behavioral and/or mental adjustments, continuous striving and active coping, frustration in achieving goals, mental effort especially in solving difficult problems, momentary moods and behaviors involving fear, anger, and other emotional states, inhibition or suppression of emotional states and reactions, conflicts about the expression of angry feelings or other negative emotions, and interpersonal and group conflict. Moreover, even blood pressure responses to simple physical efforts like postural change may be affected by psychosocial factors (Davydov et al., 2012). Social-cultural norms define what are appropriate and desired reactions and behaviors. The extent to which blood pressure is affected is determined by the balance between demands placed on the individual and on one's resources and ability to cope with the demands.

A common theme in the literature is that individual differences in cardiovascular reactivity (CVR) to various stimuli and events serve as a risk factor for hypertension and cardiovascular disorders in general (Carroll et al., 2012, 2011). Increased emotional reactions, sympathetic activity and associated hormonal and humoral changes, and the sensitivity of peripheral tissue are considered mechanisms of excessive CVR and pathways to established hypertension, where activation of the sympathetic nervous system seems to play a pivotal role (Gamboa et al., 2012; Lovallo, 2003; Singh et al., 2010). Most antihypertensive medications appear to regulate blood pressure by interfering with these mechanisms. A defect in baroreflex and vagus regulations may also contribute to hypertension (Davydov et al., 2010a, 2007; Lohmeier et al., 2009; Paton et al., 2005; Sartori et al., 2005; Thrasher, 2006).

Psychosocial factors are associated with higher casual or ambulatory levels of blood pressure, suggesting that the transformation from excessive reactivity to a chronic elevation may take place with repeated instances of blood pressure increases. Thus, in some samples of subjects, levels of blood pressure, especially during sleep, have been found to be positively associated with negative moods (e.g., angry, stressed) during the day and such personality traits as hostility and anxiety (Davydov et al., 2010a; Shapiro et al., 2001, 1997b). Physiological models have been developed to describe the cardiovascular dynamics underlying the transformation from

early heightened reactivity to later established high levels of blood pressure.

Hypertension may be related to a physiological mechanism of protection of positive mood with a resulting negative somatic outcome (high blood pressure) for the maintenance of positive subjective well-being in the face of aversive, threatening or unfriendly surroundings (Berendes et al., 2013; Davydov et al., 2012). For example, chronic pain and chronic negative mood may increase hypertension risk as a physiological compensation mechanism regulating arousal systems in persistent coping with negative challenges (Davydov et al., 2012; Drouin and McGrath, 2013; Olsen et al., 2013). This may explain associations found between negative mood and subsequent elevation in blood pressure and a proposal to use blood pressure reactivity to some challenges as an objective metric of individual capacity for mental resilience against chronic or occasional adversity (Davydov et al., 2012, 2010b).

Methodological and Conceptual Issues in Research on CVR

Debates continue about the evidence supporting the CVR hypothesis based on responses to laboratory stressors (Zanstra and Johnston, 2011). Most laboratory studies of CVR are limited in the range of cognitive, social, emotional, physical, and complex tasks studied. Inter-task CVR consistency is not high, indicating that no single task may be able to represent a CVR trait of the individual (Ottaviani et al., 2008). Individuals vary widely in their blood pressure responses to various circumstances. Variations in the kinds of individuals and methods chosen for study may result in different findings and conclusions. Laboratory settings may impose constraints on the participants so that they may not respond to challenges and threats with their preferred method of coping. And individuals vary in the degree to which their responses persist over time for any one provocation or generalize across provocations.

One alternative to laboratory procedures is to assess CVR in everyday settings using diaries and ambulatory blood pressure measures (Shapiro et al., 1993; Zanstra and Johnston, 2011). However, a problem of naturalistic studies is that each participant in a specific sample may be differentially exposed to daily life stressors during a selected sampling period. Averaging CVR evoked by different stressors has been proposed as a way to improve reliability of measures obtained under controlled laboratory conditions for evaluating CVR as a trait of the individual (Kamarck et al., 2000; Zanstra and Johnston, 2011).

The association between perceptions of stress and the blood pressure response to such stress may not be consistent because some individuals tend to intentionally or non-intentionally over- or under-report moods and emotional states while at the same time showing decreased or increased physiological reactivity to provocations or diminished or elevated blood pressure levels in general. These individuals either generally do not feel the reported emotions or wish to make a favorable impression on other people (Davydov et al., 2013). These characteristics have been labeled as 'defensiveness,' 'alexithymia,' or 'response style' (Shapiro et al., 1996). Individuals may learn to cope effectively with stress or keep their emotional responses under control. For example, experienced air traffic controllers or other professionals dealing with emergencies do not necessarily show elevated levels of blood pressure (Sega et al., 1998). By taking these various factors into account, further sense may be made out of the inconclusive findings and the soundness of the reactivity model.

The ability to predict the development of sustained high levels of blood pressure, established hypertension, and cardiac events may be improved by considering other psychological factors such as various personality traits and coping styles (anger/hostility, defensiveness, emotional suppression, denial, repression, anxiety, lack of control and helplessness, Type A and D personality, depression, pessimism) (Brydon et al., 2010; Einvik et al., 2011; Everson-Rose and Lewis, 2005; Gleiberman, 2007; Lévesque et al., 2010; Rutledge and Linden, 2003; Smith and MacKenzie, 2006). In the social domain, contributing factors are low socioeconomic status, social disintegration, and family disruption (poverty, high crime rates, high divorce rates), everyday hassles, occupational stress (job strain, job dissatisfaction, high job demands, little independence in one's work duties or latitude in the ability to make decisions), low levels of family and social support, social isolation, taxing responsibilities such as providing care to disabled family members, marital conflict, minority social status, ethnicity, and gender (Everson-Rose and Lewis, 2005; Kamarck et al., 2005; Lin et al., 2011). For example, in a large prospective study, after controlling for multiple confounding factors, greater reactivity to a mental stress task was associated with higher resting systolic blood pressure, more upward drift in resting systolic blood pressure, and hypertension 12-years later, providing some support for the reactivity hypothesis (Carroll et al., 2011).

Exploiting mediation and moderation analyses may help explain associations between psychosocial factors, CVR, and hypertension. For example, defensiveness was found to determine hypertension based on CVR (Rutledge and Linden, 2003). Anger and hostility do not necessarily have the same implications for men and women. An important missing link between CVR and subsequent established hypertension has been suggested by research on the degree to which individuals recover from reactions to stress and on rumination (Davydov et al., 2011; Ottaviani et al., 2011). In a laboratory study, it was shown that recovery after stress (recalling an experience 'that made you very angry') was much greater if the participant was distracted. Moreover, in a 24-h ambulatory recording, systolic blood pressure at times that the person was thinking about the anger experience was 20 mmHg higher compared to times without such rumination (Ottaviani et al., 2011). See Gerin et al. (2012) for an overview of this framework.

Biopsychosocial Models

The ability to predict whether repeated elicitations of increases in blood pressure will lead eventually to chronically elevated blood pressure and to cardiovascular disease is probably best improved when certain biological factors are included in the test models, such as obesity, physical activity, physical

conditioning, salt sensitivity, excitable autonomic nervous system, deficiencies in certain vitamins and minerals, intake of sodium and potassium, presence of other disorders such as diabetes, high glucose, insulin, or lipid levels, certain racial characteristics, and other genetic or constitutional predispositions such as family history of hypertension and cardiovascular disease (Lovallo, 2003; Singh et al., 2010). In a study of occupational stress in nurses, those nurses for whom both parents had hypertension and who were hostile, defensive, or anxious had elevated ambulatory blood pressure, which was highest during work days (Shapiro et al., 1996). A similar finding is reported by Vögele and Steptoe (1993) who found the highest CVR in adolescent boys who had a tendency to suppress anger expression and whose parents had high blood pressure, indicating a raised risk of future cardiovascular disease. Other psychosocial, life style, and habitual behavior patterns need to be considered as they are relevant to one's diet, activity level, and intake of caffeine, alcohol, and other substances. For example, certain personality traits like a high level of trait hostility may be associated with unhealthy health habits such as smoking, low levels of exercise, poor physical conditioning, poor nutrition, and increased number of social, family, and occupational stresses and conflicts. These associations are explained by a bidirectional model of person–environment interrelationships when stressful and other risk factors are considered as a combination of random and non-random events with the latter influenced by an individual's own reactivity phenotype (Davydov et al., 2012).

Diagnostic Issues

Psychological factors are relevant to the diagnosis of hypertension. Some patients may show elevations in blood pressure during a doctor's visit and evaluation ('white coat effect') but not at other times. This effect has been related to an alerting response to the examination or anxiety about one's health. In one study, trait hostility was found relevant to this effect. Those individuals who were highly hostile were consistently high in blood pressure during the examination and over the course of the day, whereas low hostile individuals were more likely to show elevated pressure only during the evaluation and not during the day (Jamner et al., 1993). Another study showed that if blood pressure was measured in alerting (standing) posture or at the start of doctor's visit higher blood pressure was related to lower level of depression, but if blood pressure was measured in the same subjects in a relaxing (lying) position or several minutes later during the same doctor's visit higher blood pressure was already related to higher level of depression (Davydov et al., 2012). By having the patient record blood pressure, concurrent behavioral activity, and self-reports at home and in everyday settings, useful information may be obtained on blood pressure responses and variability at different times and occasions that may guide decisions about treatment and increase the patient's adherence to medications or otherwise facilitate pressure reduction. Reliable wireless devices and 'cloud' communication technologies for building real-time and continuous ambulatory health tele-monitoring systems are now available for this purpose (Chan et al., 2012).

The Problem of Blood Pressure Variability

Blood pressure can be quite variable both in the short-term (minutes to hours) with a range from a small change to 60 mmHg over the course of meetings, public speaking, conflict, pain, waking from sleep, driving, posture, eating, smoking, etc. (usually of 10–15 mmHg for daytime and 5–10 mHg at night), and in the long-term (days to months) with a usual range of 10–20 mmHg over the course of chronic stress (e.g., work-related), visits to doctor, season, infection, shift work, menstrual cycle, depression episodes, stopping addictive habits, etc. (Rothwell, 2010). Short-term (e.g., within 24-h) and long-term (e.g., clinical and 24-h visit-to-visit) blood pressure variability were found to be only weakly correlated ($r < 0.2$) that challenges their equal prognostic value for cardiac events and cardiovascular mortality (Mancia et al., 2012). Moreover, in some cases enhanced blood pressure variability (e.g., high circadian difference of systolic blood pressure) was of a protective nature against follow-up cardiovascular mortality, but in others (e.g., high standard deviation of systolic blood pressure during the night) it was an independent predictor of cardiac events (Mancia, 2012).

These facts provided the impetus for the investigation of blood pressure in different settings with the control for their effects in various individual psychosocial and physical conditions for the correct diagnosis and the evaluation of drug effects (Davydov et al., 2012, 2010a). Thus, ambulatory blood pressure recording methods buttressed by information obtained from diaries of self-reports of moods and the situations encountered in everyday life offer many advantages, including a large number of blood pressure measurements under different conditions (work vs non-work or off days, sleep vs waking periods) and in a variety of repeatedly occurring situations, locations, and activities. Valuable data have been obtained from studies of special populations: paramedics, traffic control agents, secretaries, laboratory technicians, emergency dispatchers, nurses, bus drivers, and college students (Shapiro et al., 2002). The research is challenging because of the complexities and uncontrolled nature of the observations, but novel analytical tools show great promise (Davydov, 2013; Davydov et al., 2010a; Ottaviani et al., 2011).

Psychosocial Factors and Treatment of Hypertension

Psychosocial factors play a role in the treatment and management of hypertension. The response to antihypertensive drugs and to various behavioral methods may partly depend on psychosocial factors, although research is sparse on this topic and these factors are not usually considered in the choice of medications and methods used to treat patients. In one example, individuals with high levels of hostility were found less responsive to diuretics than non-hostile individuals, suggesting that the former might be more responsive to agents directly affecting the sympathetic nervous system (Pasic et al., 1994). Another study found that in emotionally stable people, training for avoiding negative stimuli perception was associated with elevated post-intervention blood pressure and higher blood pressure reactivity to stress (Higgins and Hughes, 2012). Ethnicity can affect the response to treatment

of hypertension (Kramer et al., 2004). Psychological factors influence degree of adherence to medications and may affect the degree to which side effects of medications occur and affect quality of life (Farrell et al., 2011).

A great deal of research has been devoted to the evaluation of nondrug treatments of hypertension derived from psychological and behavioral principles (Brannon and Feist, 2009). The methods include progressive relaxation, simple relaxation, biofeedback-guided relaxation, biofeedback training involving blood pressure, heart rate, and other cardiovascular measures, biofeedback training using measures associated with sympathetic nervous system functioning, cognitive-behavioral methods of facilitating coping with stress and dealing with anger ('stress management training'), hypnosis, and autogenic training. These procedures seem most suitable as alternatives in patients with mild hypertension, especially those who are free of significant cardiovascular disease risk factors (family history, high cholesterol level, smoking) and in other patients as a means of reducing the amount of medication needed for blood pressure control or completely eliminating the need for such medication (Guo et al., 2008). Reducing drug requirements may be of special value in dealing with undesired side effects of medications in some patients (Shapiro et al., 1997a). As yet there is no consensus about the effectiveness of these procedures, and major trials remain to be done (Dickinson et al., 2008). Other traditional methods such as yoga and meditation may be useful as a means of increasing awareness and control of mental and bodily functioning. Regular physical exercise can facilitate blood pressure reduction and improve the coronary circulation. Whether or not people exercise, avoid or take medications, or use herbal, homeopathic, or various nondrug methods is determined by many different psychological, social, and cultural factors (Glynn et al., 2010).

See also: Behavioral Medicine; Cardiovascular Reactivity and Recovery; Coronary Heart Disease: Psychosocial Aspects; Depression, Pessimism, and Health; Gender and Physical Health; Personality and Health; Race: Ethnicity and Health; Stress and Cardiac Response.

Bibliography

Berendes, A., Meyer, T., Hulpke-Wette, M., Herrmann-Lingen, C., 2013. Association of elevated blood pressure with low distress and good quality of life: results from the nationwide representative German health interview and examination survey for children and adolescents. Psychosomatic Medicine 75, 422–428.

Brannon, L., Feist, J., 2009. Health Psychology: An Introduction to Behavior and Health, seventh ed. Cengage Learning, Belmont, CA.

Brydon, L., Strike, P.C., Bhattacharyya, M.R., Whitehead, D.L., McEwan, J., Zachary, I., Steptoe, A., 2010. Hostility and physiological responses to laboratory stress in acute coronary syndrome patients. Journal of Psychosomatic Research 68, 109–116.

Carroll, D., Ginty, A.T., Painter, R.C., Roseboom, T.J., Phillips, A.C., de Rooij, S.R., 2012. Systolic blood pressure reactions to acute stress are associated with future hypertension status in the Dutch Famine Birth Cohort Study. International Journal of Psychophysiology 85, 270–273.

Carroll, D., Phillips, A.C., Der, G., Hunt, K., Benzeval, M., 2011. Blood pressure reactions to acute mental stress and future blood pressure status: data from the 12-year follow-up of the West of Scotland study. Psychosomatic Medicine 73, 737–742.

Chan, M., Estève, D., Fourniols, J.-Y., Escriba, C., Campo, E., 2012. Smart wearable systems: current status and future challenges. Artificial Intelligence in Medicine 56, 137–156.

Davydov, D.M., 2013. Asymmetry in activity of cardiovascular system: its origin and role in the maintenance of human resilience and adaptation to stressors. Pathogenesis 11, 33–38.

Davydov, D.M., Luminet, O., Zech, E., 2013. An externally oriented style of thinking as a moderator of responses to affective films in women. International Journal of Psychophysiology 87, 152–164.

Davydov, D.M., Shapiro, D., Cook, I.A., Goldstein, I., 2007. Baroreflex mechanisms in major depression. Progress in Neuro-Psychopharmacology and Biological Psychiatry 31, 164–177.

Davydov, D.M., Shapiro, D., Goldstein, I.B., 2010a. Relationship of resting baroreflex activity to 24-hour blood pressure and mood in healthy people. Journal of Psychophysiology 24, 149–160.

Davydov, D.M., Stewart, R., Ritchie, K., Chaudieu, I., 2010b. Resilience and mental health. Clinical Psychology Review 30, 479–495.

Davydov, D.M., Stewart, R., Ritchie, K., Chaudieu, I., 2012. Depressed mood and blood pressure: the moderating effect of situation-specific arousal levels. International Journal of Psychophysiology 85, 212–223.

Davydov, D.M., Zech, E., Luminet, O., 2011. Affective context of sadness and physiological response patterns. Journal of Psychophysiology 25, 67–80.

Dickinson, H., Campbell, F., Beyer, F., Nicolson, D., Cook, J., Ford, G., Mason, J., 2008. Relaxation therapies for the management of primary hypertension in adults: a cochrane review. Journal of Human Hypertension 22, 809–820.

Drouin, S., McGrath, J.J., 2013. Blood pressure and pain sensitivity in children and adolescents. Psychophysiology 50, 513–520.

Einvik, G., Dammen, T., Hrubos-Strøm, H., Namtvedt, S.K., Randby, A., Kristiansen, H.A., Somers, V.K., Nordhus, I.H., Omland, T., 2011. Prevalence of cardiovascular risk factors and concentration of c-reactive protein in type D personality persons without cardiovascular disease. European Journal of Cardiovascular Prevention and Rehabilitation 18, 504–509.

Everson-Rose, S.A., Lewis, T.T., 2005. Psychosocial factors and cardiovascular diseases. Annual Review of Public Health 26, 469–500.

Farrell, K., Shen, B.-J., Mallon, S., Penedo, F.J., Antoni, M.H., 2011. Utility of the millon behavioral medicine diagnostic to predict medication adherence in patients diagnosed with heart failure. Journal of Clinical Psychology in Medical Settings 18, 1–12.

Gamboa, A., Okamoto, L.E., Diedrich, A., Choi, L., Robertson, D., Farley, G., Paranjape, S., Biaggioni, I., 2012. Sympathetic activation and nitric oxide function in early hypertension. American Journal of Physiology. Heart and Circulatory Physiology 302, H1438–H1443.

Gerin, W., Zawadzki, M.J., Brosschot, J.F., Thayer, J.F., Christenfeld, N.J.S., Campbell, T.S., Smyth, J.M., 2012. Rumination as a mediator of chronic stress effects on hypertension: a causal model. International Journal of Hypertension 2012, 9.

Gleiberman, L., 2007. Repressive/defensive coping, blood pressure, and cardiovascular rehabilitation. Current Hypertension Reports 9, 7–12.

Glynn, L.G., Murphy, A.W., Smith, S.M., Schroeder, K., Fahey, T., 2010. Interventions used to improve control of blood pressure in patients with hypertension. Cochrane Database of Systematic Reviews, CD005182.

Guo, X., Zhou, B., Nishimura, T., Teramukai, S., Fukushima, M., 2008. Clinical effect of qigong practice on essential hypertension: a meta-analysis of randomized controlled trials. Journal of Alternative Complementary Medicine 14, 27–37.

Higgins, N.M., Hughes, B.M., 2012. Individual differences in the impact of attentional bias training on cardiovascular responses to stress in women. Anxiety Stress and Coping 25, 381–395.

Jamner, L.D., Shapiro, D., Hui, K.K., Oakley, M.E., Lovett, M., 1993. Hostility and differences between clinic, self-determined, and ambulatory blood pressure. Psychosomatic Medicine 55, 203–211.

Kamarck, T.W., Debski, T.T., Manuck, S.B., 2000. Enhancing the laboratory-to-life generalizability of cardiovascular reactivity using multiple occasions of measurement. Psychophysiology 37, 533–542.

Kamarck, T.W., Schwartz, J.E., Shiffman, S., Muldoon, M.F., Sutton-Tyrrell, K., Janicki, D.L., 2005. Psychosocial stress and cardiovascular risk: what is the role of daily experience? Journal of Personality 73, 1749–1774.

Kramer, H., Han, C., Post, W., Goff, D., Diez-Roux, A., Cooper, R., Jinagouda, S., Shea, S., 2004. Racial/ethnic differences in hypertension and hypertension treatment and control in the multi-ethnic study of atherosclerosis (MESA). American Journal of Hypertension 17, 963–970.

Lévesque, K., Moskowitz, D.S., Tardif, J.-C., Dupuis, G., D'antono, B., 2010. Physiological stress responses in defensive individuals: age and sex matter. Psychophysiology 47, 332–341.

Lin, S.X., Carnethon, M., Szklo, M., Bertoni, A., 2011. Racial/ethnic differences in the association of triglycerides with other metabolic syndrome components: the multi-ethnic study of atherosclerosis. Metabolic Syndrome and Related Disorders 9, 35–40.

Lohmeier, T.E., Hildebrandt, D.A., Dwyer, T.M., Iliescu, R., Irwin, E.D., Cates, A.W., Rossing, M.A., 2009. Prolonged activation of the baroreflex decreases arterial pressure even during chronic adrenergic blockade. Hypertension 53, 833–838.

Lovallo, W.R., 2003. Psychophysiological reactivity: mechanisms and pathways to cardiovascular disease. Psychosomatic Medicine 65, 36–45.

Mancia, G., 2012. Short- and long-term blood pressure variability: present and future. Hypertension 60, 512–517.

Mancia, G., Facchetti, R., Parati, G., Zanchetti, A., 2012. Visit-to-visit blood pressure variability in the European Lacidipine Study on Atherosclerosis: methodological aspects and effects of antihypertensive treatment. Journal of Hypertension 30, 1241–1251.

Olsen, R.B., Bruehl, S., Nielsen, C.S., Rosseland, L.A., Eggen, A.E., Stubhaug, A., 2013. Hypertension prevalence and diminished blood pressure-related hypoalgesia in individuals reporting chronic pain in a general population: the tromsø study. Pain 154, 257–262.

Ottaviani, C., Shapiro, D., Davydov, D.M., Goldstein, I.B., 2008. Autonomic stress response modes and ambulatory heart rate level and variability. Journal of Psychophysiology 22, 28–40.

Ottaviani, C., Shapiro, D., Fitzgerald, L., 2011. Rumination in the laboratory: what happens when you go back to everyday life? Psychophysiology 48, 453–461.

Pasic, J., Shapiro, D., Jamner, L.D., Hui, K.K., 1994. Hostility and the response to diuretic in mild-to-moderate hypertension. American Journal of Hypertension 7, 503–508.

Paton, J.F.R., Boscan, P., Pickering, A.E., Nalivaiko, E., 2005. The yin and yang of cardiac autonomic control: vago-sympathetic interactions revisited. Brain Research Reviews 49, 555–565.

Rothwell, P.M., 2010. Limitations of the usual blood-pressure hypothesis and importance of variability, instability, and episodic hypertension. Lancet 375, 938–948.

Rutledge, T., Linden, W., 2003. Defensiveness and 3-year blood pressure levels among young adults: the mediating effect of stress-reactivity. Annals of Behavioral Medicine 25, 34–40.

Sartori, C., Lepori, M., Scherrer, U., 2005. Interaction between nitric oxide and the cholinergic and sympathetic nervous system in cardiovascular control in humans. Pharmacology & Therapeutics 106, 209–220.

Sega, R., Cesana, G., Costa, G., Ferrario, M., Bombelli, M., Mancia, G., 1998. Ambulatory blood pressure in air traffic controllers. American Journal of Hypertension 11, 208–212.

Shapiro, D., Goldstein, I.B., Jamner, L.D., 1996. Effects of cynical hostility, anger out, anxiety, and defensiveness on ambulatory blood pressure in black and white college students. Psychosomatic Medicine 58, 354–364.

Shapiro, D., Hui, K.K., Oakley, M.E., Pasic, J., Jamner, L.D., 1997a. Reduction in drug requirements for hypertension by means of a cognitive-behavioral intervention. American Journal of Hypertension 10, 9–17.

Shapiro, D., Jamner, L.D., Davydov, D.M., James, P., 2002. Situations and moods associated with smoking in everyday life. Psychology of Addictive Behaviors 16, 342–345.

Shapiro, D., Jamner, L.D., Goldstein, I.B., 1993. Ambulatory stress psychophysiology: the study of "compensatory and defensive counterforces" and conflict in a natural setting. Psychosomatic Medicine 55, 309–323.

Shapiro, D., Jamner, L.D., Goldstein, I.B., 1997b. Daily mood states and ambulatory blood pressure. Psychophysiology 34, 399–405.

Shapiro, D., Jamner, L.D., Goldstein, I.B., Delfino, R.J., 2001. Striking a chord: moods, blood pressure, and heart rate in everyday life. Psychophysiology 38, 197–204.

Singh, M., Mensah, G.A., Bakris, G., 2010. Pathogenesis and clinical physiology of hypertension. Cardiology Clinics 28, 545–559.

Smith, T.W., MacKenzie, J., 2006. Personality and risk of physical illness. Annual Review of Clinical Psychology 2, 435–467.

Thrasher, T.N., 2006. Arterial baroreceptor input contributes to long-term control of blood pressure. Current Hypertension Reports 8, 249–254.

Vögele, C., Steptoe, A., 1993. Anger inhibition and family history as modulators of cardiovascular responses to mental stress in adolescent boys. Journal of Psychosomatic Research 37, 503–514.

Zanstra, Y.J., Johnston, D.W., 2011. Cardiovascular reactivity in real life settings: measurement, mechanisms and meaning. Biological Psychology 86, 98–105.

Hypnosis

Burkhard Peter, MEG-Stiftung, München, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Hypnosis is defined as the art of creating an alternative reality through imagination. The experience of evidence (illusion) and involuntariness (shift in agency) seems helpful, if not crucial, to the feeling of being hypnotized. Several variables are important to the quality of the hypnotic response, such as hypnotizability, hypnotic or trance state, hypnotic phenomena, hypnotic suggestions, context variables, motivation, and expectancy as well as relationship (hypnotic rapport). These factors are described, illustrated in its meaning and practical use, and substantiated by research results. Applications in medicine, dentistry, and psychotherapy are described, and relevant data supporting the efficacy of its use in these fields are provided. The role of Franz Anton Mesmer in historical reference as founder of modern hypnotism is challenged.

Hypnosis can be defined as the art of creating an alternative reality by imagination, which, ideally, should be experienced like a hallucination or illusion. In order to accomplish this, a person, with the help of hypnotic induction rituals, is 'put into' a hypnotic or 'trance' state, which is assumed to be a distinctive state different from normal experience. It is thought that being in a hypnotic trance makes it easier for the hypnotized person to use his/her hypnotizability to construe this alternative reality. Two experiences are helpful, if not crucial, for the feeling to be hypnotized: evidence (i.e., the subjective conviction of experiencing it 'as real'), and a shift in agency, from voluntariness to involuntariness (i.e., allowing for a kind of an 'alien control').

The more intense and real (evident) this alternative reality feels and the more it is experienced in form of hypnotic phenomena (i.e., hallucinations, illusions, and involuntary responses), the more likely normal reality is dismissed or dissociated, partially or as a whole, during the time of the trance state; and the more likely therapeutic relevant features of this alternative reality will be eventually implemented in everyday life. In that way, hypnotized patients can better tolerate aversive medical procedures, or, in the course of a hypnotherapy, can change their feelings, cognitions, and behavior.

Usually, the hypnotic trance is induced, and hypnotic phenomena are suggested by another person (heterohypnosis). However, they can also be self-induced (self-hypnosis). Several variables, which are more or less interconnected, play a role in the quality of the hypnotic response: the amount of hypnotic suggestibility or hypnotizability, the presence of a trance state and its depth, the kind of hypnotic phenomenon that is intended, the way in which suggestions aiming at a specific hypnotic phenomenon are presented, and the variables related to the context in which suggestions are presented. Motivation and expectancies are important too, as well as variables pertaining to the relationship between the person who gets hypnotized and the particular hypnotist or hypnotherapist in addition to his or her hypnotic skills.

Hypnotizability

The expressions hypnotizability, hypnotic susceptibility, hypnotic responsiveness, or hypnotic suggestibility are used interchangeably when the measurement of suggestibility is preceded by a formal hypnotic induction procedure. If there is no preceding hypnotic induction, terms like nonhypnotic, imaginative, or waking suggestibility are used. Hypnotic and nonhypnotic suggestibility are highly correlated. To assess hypnotizability, subjects or patients are presented with various suggestions that are intended to elicit certain hypnotic phenomena. The amount of hypnotizability is measured by the number of suggestions that are passed successfully. Similar to other personality traits, hypnotizability is normally distributed, with approximately 20% being high hypnotizables, 60% medium hypnotizables, and 20% low hypnotizables – depending on the choice of cutoff values (e.g., Peter et al., 2014).

Because correlations are unstable, the amount of an individual's hypnotizability cannot be reliably predicted from other personality traits; it can only be measured with hypnotizability scales. In recent years, a number of studies support the assumption that there may be two distinctive groups of high hypnotizables: individuals who like to be involved in hypnosis ('hypnosis prone'), that is, the classical 'fantasizers' with high imaginative abilities ('fantasy proneness'), and individuals who are highly dissociative and probably vulnerable to psychopathology (Peter et al., 2014). Twin studies and longitudinal studies point to the heritability of hypnotizability, as well as recent results from genetics: hypnotizability has been associated with genetic polymorphism in catechol-*O*-methyltransferase, an enzyme that is involved in dopaminergic metabolism. This led to the conclusion that high hypnotizables supposedly show greater frontal lobe dopamine-mediated flexibility than low hypnotizables (Lichtenberg et al., 2004). This might be the reason why 'hypnotic prone' individuals obviously showed markedly higher scores in charming ('charming/histrionic') and optimistic ('optimistic/rhapsodic') personality traits. This special attitude of social approachability of high hypnotizables (in case that they are securely attached; see Peter et al., 2014) can obviously also be artificially created by oxytocin administration (Bryant and Hung, 2013) or by disrupting the dorsolateral prefrontal cortex with low frequency repetitive transcranial magnetic stimulation (Dienes and Hutton, 2013) – i.e., a temporal enhancement of (state) hypnotizability. Furthermore, according to recent data (Geiger et al., 2014), hypnotizability seems to be positively correlated with intelligence after all, at least in women; the negative correlation found in men might explain why this association was not found in earlier studies.

Hypnotic Trance

Suggestions presented in hypnotizability scales are normally preceded by a so-called hypnotic or trance induction, during which the person is invited – after some introductory and explanatory words – to relax and enter a hypnotic trance. The benefit of using trance induction procedures has been discussed controversially. According to Braffman and Kirsch (1999), hypnotizability test scores were improved by preceding trance induction only in a small part of subjects, and only marginally so. The hypnotic induction showed no effect for about half of the subjects' and in fact a negative effect for a quarter of the subjects. Does this mean hypnosis induction rituals are expendable or even detrimental in some individuals? From a socio-cognitive perspective, trance induction works by creating expectancies of how one will response to the subsequent suggestions, and these expectancies make it easier for subjects or patients to use their suggestibility for achieving the aims of hypnosis. Hence, in clinical practice, hypnosis is preceded by a shorter or longer, direct or indirect induction of a hypnotic trance, before actual suggestions are implemented during hypnotherapeutic work.

Both the wording and grade of elaboration in presenting the induction formula are of particular importance for the effect of the subsequent hypnotic suggestions. For example, responses to suggestions in hypnotizability scales are less pronounced if introducing words convey the expectation that relaxation is induced rather than hypnosis (Gandhi and Oakley, 2005). Obviously, highly hypnotizable individuals need to hear precisely what they should be doing, that is, whether they are expected to relax or to enter a hypnotic trance. Culturally shaped meanings imply that these two are not exactly the same. High hypnotizables seem to respond easily to simple direct suggestions while medium hypnotizables need longer and more elaborated suggestions (Szabó, 1996), like the so-called indirect approaches in Ericksonian hypnosis, such as conversational inductions, embedded commands, or therapeutic metaphors.

Neuroimaging studies show that, in highly hypnotizable individuals, the induction of a hypnotic state is accompanied by changes in brain activity. These changes consist mainly of decreased activation of certain prefrontal regions (Dietrich, 2003; Egner et al., 2005) and relate to the so-called default mode network (DMN; McGeown et al., 2009). With this, the old controversy in hypnosis research between state theorists and nonstate theorists (e.g., socio-cognitive theorists) seems to be resolved, at least insofar as the latter denied the existence of a specific hypnotic state. However, the significance of the effect of hypnotic induction in relation to hypnotizability is still unclear (e.g., Mazzoni et al., 2012).

Hypnotic Phenomena: Practical Uses and Scientific Results

Hypnotic phenomena are characterized by changes in perception, experience, cognition, and behavior, which are different from an individual's everyday experience, and different from deliberately produced imagination. Accompanied by psychophysiological and neurophysiological concomitants, they exhibit a high degree of evidence, so that they can be experienced as actual hallucinations or illusions. This can be called the ratification of a hypnotic trance.

Hypnotically suggested behaviors are perceived to have a high degree of involuntariness or automatism. The experience of agency and the perceived control of an action should be significantly altered, sometimes right up to the experience of an alien control. In addition to the feeling of evidence of the imaginations, the perceived involuntariness of the suggested action is an essential element to the experience of being hypnotized. By involuntariness, the hypnotized person indicates that she/he willingly transfers the agency of her/his actions to another agent and allows a kind of alien control. Without voluntary help, a suggested action can occur (such as an arm lifting without voluntary effort following a suggestion for arm levitation) or an action may not be executed in spite of deliberate effort to do so (a cataleptic arm cannot be bent, or one's eyes, when closed, cannot be opened willfully). These are classical dissociative phenomena, which are typically perceived as unusual or unfamiliar and therefore attributed to the state of being hypnotized. The connection of two in everyday life's continuous mental processes is interrupted, such as that between intention ('I want to open my eyes') and execution ('I can open them').

The important factor is the exact attribution of this source of alien control because it conveys, among other things, the preconceived notion that hypnosis is operated by a 'lack of will,' in the case of the 'hypnotic power' of a stage hypnotist. A neutral attribution in scientific studies, for example, usually refers to the hypnotizability of the subject. In the psychotherapeutic context, an internal locus of reattribution is exclusively appropriate, such as one's unconscious resources, the power and capabilities of the unconscious mind, or unconscious creativity, in order to expand a patient's sense of self-efficacy.

The classic hypnotic responses (hypnotic phenomena), which are also produced through hypnotizability assessments, can be categorized according to phenomenological aspects into the following groups.

Motor-Kinesthetic Phenomena

Motor-kinesthetic phenomena consist of so-called ideomotor behaviors, which are experienced as involuntary (e.g., hand levitation, eyelid closure), catalepsy, or paretic inhibition of behaviors. Most people can experience motor-kinesthetic phenomena, which do not typically require a high degree of suggestibility (Peter et al., 2013). Direct suggestions for ideomotor movements, such as for eye closure or arm levitation, are more easily and often followed than suggestions for the lack of bodily mobility, which challenge one's personal agency, such as eye catalepsy (the eyes cannot voluntarily be opened) or entire bodily paralysis (the muscles cannot be moved willingly).

The psychotherapeutic object for a full body paralysis through motor inhibition accompanied by sensory deprivation through eyelid closure can be seen in facilitating the construction of an alternative reality by increasing the capacity for imagination and deepening the absorption. Thus, the prerequisites for classical (Pavlovian) conditioning are created, i.e., for psychophysiological and emotional learning. If one's

own cause of an action cannot be experienced because of a catalepsy or paresis, the question arises as to whom or to what agency and control has been transferred. In the history of hypnosis, this is traditionally a 'therapeutic Tertium (third),' whose special form is shaped by socio-cultural beliefs. The contemporary gestalt of this therapeutic Tertium are the 'unconscious' resources of a person, which are often summarily referred to, especially in the context of Milton Erickson, as 'the unconscious' or, as is the actual traditional term, the 'subconscious' (Dessoir, 1896). The ritual of arm levitation is well suited for making this therapeutic construct of the unconscious available and to get in contact with it ("The more you come in contact with your unconscious, the higher goes your arm…"), such as through ideomotor signaling ("If your unconscious mind knows the answer to the question XY, then the right hand goes up; if not, the left hand…"). Through this or a similar therapeutic Tertium (unconscious knowledge, inner wisdom, hidden resources, etc.; see Peter, 2009a), a patient's sense of self-efficacy can be restored and strengthened. In some severely disturbed patients, when suitable internal resources are lacking, a targeted therapeutic input is necessary. But even in this case, the form of a 'representative' figure (Bongartz and Bongartz, 2009) can be introduced and integrated into the patient's reservoir of abilities.

A series of research studies have shown that motor-kinesthetic phenomena are not mere compliance reactions. Peter et al. (2012a) were able to demonstrate that electromyographic (EMG) muscle activity of the deltoid muscle during hypnotically induced arm levitation was 27% less than deliberately lifting and holding up the arm without hypnosis. Neurological studies have shown that the perception of involuntariness and alien control activates different parts of the brain than deliberate initiation and control of the movement. Blakemore et al. (2003) found activation of the parietal operculum only when an ideomotor arm levitation following a hypnotic induction was externally attributed. Pyka et al. (2011) found that a change in the authorship of an action (i.e., perceived control or agency) was associated with a change in the region of the neighboring angular gyrus. In a recent exploratory study, Deeley et al. (2013) showed brain images of ideomotor movements while using a joystick: Loss of perceived control of movements was associated with reduced connectivity between supplementary motor area and motor regions. In addition, reduced awareness of involuntary movements were associated with less activation of parietal cortices (BA 7), which are thought to be responsible for awareness of the bodily self and for somatosensory awareness (Insula).

Sensory-Affective Phenomena

Sensory-affective phenomena include positive and negative hallucinations (or illusions) among all sensory modalities. These phenomena can be experienced by fewer people, which, among other things, depend on the sensory modality that is suggested to be hallucinatory or illusory, and on whether the experienced hallucination is challenged by external evidence. Positive visual phenomena (e.g., hallucinating a nonexistent person), for example, are easier to experience for more people than negative visual phenomena (e.g., a real existing person cannot be seen) or a combination of negative and positive olfactory hallucinations (e.g., to enjoy a strong ammonia smell as a sweet scent of violets). Taste hallucinations (a sweet or sour taste in the mouth) are, in turn, easier to hallucinate than positive auditory hallucinations (e.g., listening to music). The latter is, in turn, easier than hallucinated auditory deafness (e.g., one can no longer hear an actual speaking voice).

These phenomena find their traditional use in medical and dental applications, especially for hypnotically induced pain control; an efficient hypnotic anesthesia is by definition a negative-kinesthetic hallucination. The hypnotic amelioration of pain is, however, dependent on a patient's (trait) suggestibility and of the special features of a given pain. The hypnotic control of somatosensoric aspects of pain is, for example, more dependent on a patient's existing suggestibility, in contrast to the change of the affective parts of psychosomatic pain.

In psychotherapeutic practice, these sensory-affective phenomena depend highly on the personal meaning or the affective significance of the patient and, therefore, contribute to his/her (state) suggestibility. In trauma therapy, for example, one can regain control over intrusions through numerous uses of positive and negative hallucinations and illusions: Using all modalities and qualities of the senses and through spatial orientation, the imagination of a 'safe place' as a hallucinated alternative reality should first be constructed. The next part consists of a planned deconstruction of the traumatic experience. Those aspects of hallucinated perceptions and exuberant emotions are dissociated with the help of negative hallucinations so that they can be better processed cognitively and emotionally. Only then does a gradual confrontation with the traumatic event become useful in order to establish a new interpretation and subsequent reintegration into one's biography. Generally speaking, each hypnotic age regression entails positive and negative illusions and hallucinations, i.e., many different sensory-affective phenomena, which are based on real episodic and procedural memories from one's personal history, or are entirely constructed in therapy ('created memory'). This is especially useful when maladaptive representations of a patient's biography must be reconstructed in order to view certain aspects of the past from a more adaptive perspective. This possibility, with the help of hypnosis or even by virtue of a given high suggestibility, of creating paramnesic phenomena has led to major controversies and law trials in the 1990s on the advantages of hypnotically uncovering 'repressed memories' versus the danger of the implantation of profound 'false memories' (cf Spiegel and Scheflin, 1994).

There are now numerous studies that show that the perceived quality of sensory phenomena is verifiable through neurophysiological studies (e.g., cf Peter, 2009b). Such studies prove that sensory phenomena are caused not simply by a mechanism of distracting one's perception but are the product of profound dissociative processes (cf Friederich et al., 2001). A basic neural mechanism for dissociative hypnotic phenomena, which play a role in negative hallucinations or illusions, seems to be that established normal functional connectivity networks are suspended by appropriate hypnotic suggestions, so that there will be a disruption in the synchronization of relevant areas of the brain that effect some action or

perception (for pain, see Miltner and Weiss, 2007; Rainville et al., 1997). In contrast, those functional connectivity networks, which constitute a perception in normal circumstances, are activated when suggestions for positive phenomena are given (e.g., see Derbyshire et al., 2004 with respect to pain; Kosslyn et al., 2000 for visual illusions).

Cognitive Phenomena

Most textbooks consider the sensory phenomena as cognitive phenomena. In the phenomenological classifications established here, cognitive phenomena are limited to explicit cognitive deficits, such as partial amnesia (e.g., following a suggestion to 'forget' a number, one cannot perform an arithmetic problem in which that number appears) or following, with seeming compulsivity, a given posthypnotic suggestions (e.g., upon a previously established signal, for which source amnesia is suggested, one reproduces a suggested hypnotic phenomenon following the termination of hypnosis). More complex phenomena that resemble neurological or psychiatric disorders and include sensory phenomena are self-referenced illusions, such as being unable to recognize one's own reflection in the mirror or believing it to be that of an opposite sex fraternal twin (Cox and Barnier, 2013).

In the so-called golden era of hypnosis in Europe at the end of the nineteenth century, one thought that psychotherapeutic change could occur simply with phenomena such as amnesia and posthypnotic suggestion. However, one came to find that both were rare and relatively ineffective for many patients. Today they do not play a dominate role in hypnotherapy, at least not in the form of direct suggestions ("You will forget that for decades you have been a heavy smoker, and you will become nauseated whenever see a cigarette or smell cigarette smoke" – given the variety of smoking-related stimuli in an everyday environment, one would likely become desensitized to such a suggestion). More conceivable for psychotherapy are suggestions that are aimed at more complex changes in the self-concept of a person ("You are leaving your old lifestyle as a smoker behind you and develop a new self that can easily refrain from smoking").

That last example exemplifies that hypnosis in psychotherapy is not something in which the patient will follow well-intentioned suggestions without criticism or resistance. Rather, he should be encouraged to change more or less extensively and even profoundly his/her self-concept and conviction about his/her capacities to think, experience, and act along with his social role. This, actually, is not a new concept as the classical theory of suggestion already stated that any heterosuggestion must be accepted autosuggestively in order to be effective (Langen, 1972: p. 7).

Various findings suggest that hypnosis shifts the balance between anterior effector and posterior sensory processing (Spiegel and Kosslyn, 2004). This can also be substantiated, based on recent data, as to the so-called DMN (e.g., see McGeown et al., 2009). The DMN is usually in a deactivated state when the brain is occupied with external tasks. It is activated when there are no external tasks to which to tend and a person has the time and inclination to let one's mind wander and lead in internal self-talk. In particular, the posterior portions of the DMN, the precuneus, and the posterior cingulate are regarded as the cortical correlates of self-referential processes, which are substantially involved when one constructs a mental image of oneself (Raichle et al., 2001). Following a hypnotic induction, the anterior portions of the DMN in the prefrontal cortex specifically seem to be deactivated. To that effect, McGeown et al. (2009) believe that the hypnotic situation followed by further suggested tasks will be inferred by subjects as an external task. In psychotherapeutic terms, the trance induction seems to contribute to the ceasing of the patient's negative self-talk and prepares him/her to attend to another's voice that offers alternatives for cognitive restructuring (Kraiker, 1985). Results of Cojan et al. (2009) and Pyka et al. (2011) show a significant activation of the posterior portions of the DMN when special hypnotic tasks are desired. These studies have shown, for example, that hypnotic paralysis is not caused by a direct inhibition of motor centers by the orbitofrontal cortex, as Ward et al. (2003) have shown. Rather, a hypnotically induced immobility is apparently mediated by the precuneus and thus possibly due to an extensive change in the self-representation of the hypnotized person. The suggestion for paralysis would not be interpreted strictly in the sense of "My arm is paralyzed" – the results of Ward et al. (2003) might stipulate that – rather it is much more complex like "I am hypnotized and have been told that my arm cannot move, thus I cannot move it." One can hope that future clinical studies will show that more comprehensive hypnotherapeutic changes in self-concept, particularly in depressive or anxious patients, will correspond to analogue changes in the brain.

Ability to Hypnotize

One could assume that it takes some skill or ability to use hypnosis that might go beyond the mastery of techniques. This is still a largely unexplored area. Our own results indicate that successful users of hypnosis possess the previously described hypnotic prone personality style and have good personal experiences with hypnosis (Peter et al., 2012b).

Context of Hypnosis

In contrast, the numerous contextual factors that may influence the degree of a subject's or patient's hypnotizability have been well investigated by a variety of social psychological studies. It is obvious that task related motivation and expectations (see Lynn et al., 2008) are important factors, not only for highs but also for medium suggestible people, in order to experience hypnotic phenomena. The so-called indirect hypnotic techniques of Milton H. Erickson, particularly the elaborate instructions used to create and use hypnotic rapport, the therapeutic relationship in the hypnotic context, serve this purpose (see Rossi, 1980; for further details).

Techniques and Rituals

The techniques and rituals of the hypnotic induction as well as the subsequent utilization of the hypnotic trance state for

certain purposes, such as in medicine or psychotherapy, are very simple in its direct form. They are easy to learn and require no prerequisites. The techniques are described in detail in many manuals. In a simple, direct form, they are obviously on the hypnotizability of the subject or patient. Depending on the hypnotic phenomenon, they may only be suitable for high hypnotizables and only with limited effectiveness. The large number of medium hypnotizables requires techniques that present the necessary phenomenon in an elaborate form (Szabó, 1996). This function can be fulfilled by the indirect techniques created by Milton H. Erickson and his followers (Revenstorf and Peter, 2009).

Areas of Application

When hypnosis is used in the field of medicine or dentistry, it is referred to as clinical or medical/dental hypnosis, while it is referred to as hypnosis therapy or hypnotherapy, and sometimes therapy in trance, when used in psychotherapy. The use of hypnosis has increased since the end of the last century in medicine, dentistry, and especially in psychotherapy in German speaking countries as well as in other European and in English speaking countries. Since the 1990s, a special focus has been placed on the use of the so-called Ericksonian hypnosis or hypnotherapy. The scientific evidence regarding the efficacy of medical hypnosis is good, but unsatisfactory in the field of hypnotherapy (for more details see Hagl, 2013, 2014).

The still valid meta-analysis of the efficacy of hypnosis conducted by Flammer and Bongartz (2003) incorporated 57 randomized control trials (RCTs), in which the clinical use of hypnosis was compared with untreated or treatment-as-usual control groups, and yielded a weighted effect size of $d = 0.56$. Approximately three quarters of the 57 studies (from the years 1973–2002) investigated the application of hypnosis for somatic complaints or as an adjunct to medical treatments. Studies on affective and obsessive–compulsive disorder, however, have been lacking since the 1970s. There were some studies of anxiety disorders, which focused mainly on test anxiety with an effect size of $d = 0.69$, and seven studies on smoking cessation, which had an effect size of $d = 0.59$. A further meta-analysis by Flammer and Alladin (2007) of 21 RCTs considered only trials for the treatment of psychosomatic disorders that had a weighted effect size of $d = 0.61$. In general, the use of hypnosis in medicine is considered to be rather effective. The effectiveness of hypnosis for the reduction of experimental and clinical pain is considered good (Montgomery et al., 2000). The effectiveness of reducing anxiety and stress prior to surgical and other medical invasive procedures is well documented. Tefikow et al. (2013) calculated of 34 studies, small to medium effect sizes: $g = 0.53$ for the reduction of psychological stress, $g = 0.36$ for pain intensity, $g = 0.56$ for impairment caused by pain, and $g = 0.36$ for drug consumption. These results prompted the *Wissenschaftliche Beirat Psychotherapie* (WBP; Scientific Advisory Board) of the Federal Republic of Germany to pronounce the recognition of hypnotherapy as a treatment for the ICD diagnosis group F54 (mental and social factors in somatic disease) in 2006; this implies that the use of hypnosis is financed by legal and private health insurance.

There is a particularly diverse and extensive collection of individual reports and case studies for the use of hypnosis with dissociative, posttraumatic stress, and anxiety disorders. There are only a few reports for the treatment of obsessive–compulsive disorder and even fewer for depression alone. For anxiety disorders, the evidence of effectiveness is focused mainly on hypnotherapy for patients with test anxiety. In a meta-analysis of eight studies (Flammer, 2006) in which hypnosis was compared with a placebo of simply paying attention to patients, there was a medium effect size of $d = 0.66$. Another eight RCTs compared with a waiting list control group had an effect size of $d = 1.02$. There are only two controlled studies for posttraumatic stress and acute stress disorders, which showed a slight superiority of hypnosis compared to psychodynamic (Brom et al., 1989) or cognitive behavioral therapy (CBT; Bryant et al., 2005). For affective disorders, only one study thus far has met RCT standards: Alladin and Alibhai (2007) compared the combination of hypnosis with CBT and CBT alone, and found the former was slightly superior. In an older meta-analysis, Kirsch (1996) found a substantial effect of hypnotherapy for eating disorders when hypnosis was used in addition to behavioral weight loss programs. More recent data are not available.

There is some controversy over the results with substance dependence, particularly in regard to the value of hypnosis for smoking cessation. A Cochran study (Barnes et al., 2010) was judged to be ineffective. The WBP (2006) in Germany, however, has scientifically recognized hypnotherapy for smoking cessation (as well as methadone withdrawal). Still, the overall effectiveness of hypnosis for psychic disorders is largely not well proven by RCTs.

Another application of hypnosis is within the area of basic scientific research, referred to as scientific hypnosis. Researchers from other psychological disciplines, such as bio- and neuropsychology, make use of hypnotic techniques to investigate cognitive and emotional phenomena, often in the last two decades with the aid of brain imaging.

Stage hypnosis, in which hypnotic techniques are used for entertainment purposes, is still being used to the chagrin of almost all professional users of hypnosis. With the exception of a few countries such as Israel, Sweden, and Denmark, it is not illegal in most countries. Lay applications of hypnosis for show as well as healing purposes can be seen throughout the entire history of hypnosis as an accompanying phenomenon.

History

Hypnosis has a long history and is sometimes seen as the mold for all forms of psychotherapy, and it certainly applies for all psychodynamic therapies. Behavioral therapies date back to other sources, particularly with respect to experimental and learning psychology. The physician Franz Anton Mesmer, with his theory of animal magnetism, is often considered to be the great grandfather of modern hypnosis. This, however, has to be doubted for several reasons. His

theory postulates a universal fluid, animal magnetism, of which the flow is disrupted in ill people and had to be harmonized though certain physical applications such as the so-called passes or through a baquet. This theory has nothing to do with the modern understanding of psychotherapy. During the Romantic Era of the first third of the nineteenth century, the understanding of the original physical magnetism as conceived by Mesmer changed basically. The still practiced passes were no longer thought to transfer a physical force but to create a special somnambulistic state, which empowers the sick with special abilities, such as self-diagnosing and self-therapy. From a modern view, one could see this as a rudimentary concept of orientation toward the resources of the patient. This new and even more psychological understanding of the 'magnetic rapport' dates back to Puységur, who was one of Mesmer's students and first formulated the approach that was 100 years later elaborated by Janet and Sigmund Freud at the end of the nineteenth century. It is referred today as the therapeutic relationship and is thought to play an important role in psychotherapy and hypnotherapy (Peter and Iost-Peter, 2014). Mesmer's original ideas continue to live today in many lay and healing practitioners as fluidists and magnetists. Lay magnetists and lay hypnotists were an inseparable part of the history of hypnosis; they continued the techniques when they were forgotten in professional circles and even banned from them. James Braid, for example, developed his idea of monoideism – induced nervous sleep (Neurypnology, from which the term hypnosis is derived) – from a lay magnetist. Freud was impressed by the stage hypnotist Hansen in Vienna, then later learned eventually at the end of the nineteenth century from Jean Martin Charcot in Paris and Hippolyte Bernheim in Nancy. The theory of the latter about the effect of hypnotic suggestion would dominate the basis on scientific research of hypnosis starting in the 1920s by Clark Hull (cf 1933) in Wisconsin and later at Yale and Ernest Hilgard (cf 1979) in the 1960s at Stanford. Milton H. Erickson was a student of Clark Hull and developed a very divergent opinion about hypnosis: instead of standardizing the suggestions, as Hull used and demanded particularly for the purposes of research, the hypnotic inductions and suggestions should be adjusted to the needs and individual possibilities of each hypnotized person. From this, he developed, with great creativity, a variety of elaborate verbalizations for suggestions and suggestive strategies for therapeutic change, which are known today as the so-called indirect or Ericksonian hypnosis, and are indispensible to modern hypnotherapy. As mentioned previously, the major advantage of these techniques is that hypnosis can be used with more patients, namely those who are of medium suggestibility. Thus Erickson becomes the father of modern hypnosis and hypnotherapy. If one wants to find the great grandfather of today's psychotherapy and hypnosis, then the Exorcist Johann Joseph Gassner fulfills this role better than Franz Anton Mesmer because his special form of exorcism involves many elements of hypnotic rituals and also comprises the important principles of behavior-therapeutic self-control (Peter, 2005).

See also: Behavior Therapy: Background, Basic Principles, and Early History; Classical Conditioning Methods in Psychotherapy; Psychological Treatment, Effectiveness of; Therapist–Patient Relationship; Virtual Reality in Psychotherapy.

Bibliography

Alladin, A., Alibhai, A., 2007. Cognitive hypnotherapy for depression: an empirical investigation. International Journal of Clinical and Experimental Hypnosis 55 (2), 147–166.

Barnes, J., Dong, C.Y., McRobbie, H., Walker, N., Mehta, M., Stead, L.F., 2010. Hypnotherapy for smoking cessation. Cochrane Database of Systematic Reviews 10, CD001008.

Blakemore, S.J., Oakley, D.A., Frith, C.D., 2003. Delusions of alien control in the normal brain. Neuropsychologia 41 (8), 1058–1067.

Bongartz, B., Bongartz, W., 2009. Stellvertretertechnik. In: Revenstorf, D., Peter, B. (Eds.), Hypnose in Psychotherapie, Psychosomatik und Medizin. Ein Manual für die Praxis, second ed. Springer, Heidelberg, pp. 268–276.

Braffman, W., Kirsch, I., 1999. Imaginative suggestibility and hypnotizability: an empirical analysis. Journal of Personality and Social Psychology 77, 578–587.

Brom, D., Kleber, R.J., Defares, P.B., 1989. Brief psychotherapy for posttraumatic stress disorders. Journal of Consulting and Clinical Psychology 57, 607–612.

Bryant, R.A., Hung, L., 2013. Oxytocin enhances social persuasion during hypnosis. PLoS One 8. http://dx.doi.org/10.1371/journal.pone.0060711.

Bryant, R.A., Moulds, M.L., Guthrie, R.M., Nixon, R.V., 2005. The additive benefit of hypnosis and cognitive-behavioral therapy in treating acute stress disorder. Journal of Consulting and Clinical Psychology 73, 334–340.

Cojan, Y., Waber, L., Schwarz, S., Rossier, L., Forster, A., Vulleumier, P., 2009. The brain under self-control: modulation of inhibitory and monitoring cortical networks during hypnotic paralysis. Neuron 62, 862–875.

Cox, R.E., Barnier, A.J., 2013. Shifting self, shifting memory: testing the self-memory system model with hypnotic identity delusions. International Journal of Clinical and Experimental Hypnosis 61 (4), 416–462.

Derbyshire, S.W., Whalley, M.G., Stenger, V.A., Oakley, D.A., 2004. Cerebral activation during hypnotically induced and imagined pain. NeuroImage 23, 392–401.

Deeley, Q., Walsh, E., Oakley, D.A., Bell, V., Koppel, C., Mehta, M.A., et al., 2013. Using hypnotic suggestion to model loss of control and awareness of movements: an exploratory fMRI Study. PLoS ONE 8 (10), e78324. http://dx.doi.org/10.1371/journal.pone.0078324.

Dessoir, M., 1896. Das Doppel-Ich. Günther, Leipzig.

Dienes, Z., Hutton, S., 2013. Understanding hypnosis metacognitively: rTMS application to left DLPFC increases hypnotic suggestibility. Cortex 49, 386–392.

Dietrich, A., 2003. Functional neuroanatomy of altered states of consciousness: the transient hypofrontality hypothesis. Conscious and Cognition 12 (2), 231–256.

Egner, T., Jamieson, G., Gruzelier, J., 2005. Hypnosis decouples cognitive control from conflict monitoring processes of the frontal lobe. NeuroImage 27, 969–978.

Flammer, E., 2006. Die Wirksamkeit von Hypnotherapie bei Angststörungen. Hypnose-ZHH 1 (1 and 2), 173–198.

Flammer, E., Alladin, A., 2007. The efficacy of hypnotherapy in the treatment of psychosomatic disorders: meta-analytical evidence. International Journal of Clinical and Experimental Hypnosis 55, 251–274.

Flammer, E., Bongartz, W., 2003. On the efficacy of hypnosis: a meta-analytic study. Contemporary Hypnosis 20 (4), 179–197.

Friederich, M., Trippe, R.H., Oezcan, M., Weiss, T., Hecht, H., Miltner, W.H.R., 2001. Laser-evoked potentials to noxious stimulation during hypnotic analgesia and distraction of attention suggest different brain mechanisms of pain control. Psychophysiology 38 (5), 768–776.

Gandhi, B., Oakley, D.A., 2005. Does "hypnosis" by any other name smell as sweet? the efficacy of "hypnotic" inductions depends on the label "hypnosis". Conscious and Cognition 14, 304–315.

Geiger, E., Peter, B., Prade, T., Piesbergen, C., 2014. Intelligence and hypnotic susceptibility: is there a connection? International Journal of Clinical and Experimental Hypnosis 62 (3), 310–329.

Hagl, M., 2013. Zur Wirksamkeit von Hypnose und Hypnotherapie. Ein Forschungsbericht für die Jahre 2010 bis 2012. Hypnose-ZHH 8 (1 and 2), 145–181.

Hagl, M., 2014. Studien zur Wirksamkeit von klinischer Hypnose und Hypnotherapie im Jahre 2013. Hypnose-ZHH 9 (1 and 2), 147–167.
Hilgard, E.R., 1979. A Saga of Hypnosis: Two Decades of the Stanford Laboratory of Hypnosis Research 1957–1979. Dept. of Psychology, Stanford University, Stanford, CA.
Hull, C.L., 1933. Hypnosis and Suggestibility: An Experimental Approach. D. Appleton-Century Company, New York.
Kirsch, I., 1996. Hypnotic enhancement of cognitive-behavioral weight loss treatments – another meta-reanalysis. Journal of Consulting and Clinical Psychology 64, 517–519.
Kosslyn, S.M., Thompson, W.L., Constantini-Ferrando, M.F., Alpert, N.M., Spiegel, D., 2000. Hypnotic visual illusion alters color processing in the brain. American Journal of Psychiatry 157, 1270–1284.
Kraiker, C., 1985. Kognitive Modelle hypnotischer Phänomene. In: Peter, B. (Ed.), Hypnose und Hypnotherapie nach Milton H. Erickson: Grundlagen und Anwendungsfelder. Pfeiffer, München, pp. 20–30.
Langen, D., 1972. Kompendium der medizinischen Hypnose: Einführung in die ärztliche Praxis. Karger, Basel.
Lichtenberg, P., Bachner-Melman, R., Epstein, R.P., Crawford, H.J., 2004. Hypnotic susceptibility: multidimensional relationships with Cloniger's tridimensional personality questionnaire, COMT polymorphisms, absorption, and attentional characteristics. International Journal of Clinical and Experimental Hypnosis 52 (1), 47–72.
Lynn, S.J., Kirsch, I., Hallquist, M., 2008. Social cognitive theories of hypnosis. In: Nash, M.R., Barnier, A.J. (Eds.), The Oxford Handbook of Hypnosis. Oxford University Press, Oxford, pp. 111–140.
Mazzoni, G., Venneri, A., McGeown, W.J., Kirsch, I., 2012. Neuroimaging resolution of the altered state hypothesis. Cortex 49, 400–410.
McGeown, W.J., Mazzoni, G., Venneri, A., Kirsch, I., 2009. Hypnotic induction decreases anterior default mode activity. Consciousness and Cognition 18, 848–855.
Miltner, W.H., Weiss, T., 2007. Cortical mechanisms of hypnotic pain control. In: Jamieson, G.A. (Ed.), Hypnosis and Conscious States. The Cognitive Neuroscience Perspective. Oxford University Press, Oxford, pp. 51–66.
Montgomery, G.H., DuHamel, K.N., Redd, W.H., 2000. A meta-analysis of hypnotically induced analgesia: how effective is hypnosis? International Journal of Clinical and Experimental Hypnosis 48 (2), 138–153.
Peter, B., 2005. Gassner's exorcism – not Mesmer's magnetism – is the real predecessor of modern hypnosis. International Journal of Clinical and Experimental Hypnosis 53 (1), 1–12.
Peter, B., 2009a. Therapeutisches Tertium und hypnotische Rituale. In: Revenstorf, D., Peter, B. (Eds.), Hypnose in Psychotherapie, Psychosomatik und Medizin. Ein Manual für die Praxis, second ed. Springer, Heidelberg, pp. 69–77.
Peter, B., 2009b. Is it useful to induce a hypnotic trance? A hypnotherapist's view on recent neuroimaging results. Contemporary Hypnosis 26 (3), 132–145.
Peter, B., Bose, C., Piesbergen, C., Hagl, M., Revenstorf, D., 2012b. Persönlichkeitsprofile deutschsprachiger Anwender von Hypnose und Hypnotherapie. Hypnose-ZHH 7 (1 and 2), 31–59.
Peter, B., Geiger, E., Prade, T., Vogel, S.E., Piesbergen, C., 2014a. Norms of German adolescents for the Harvard group scale of hypnotic susceptibility, form A. International Journal of Clinical and Experimental Hypnosis.
Peter, B., Iost-Peter, A., 2014. Der "Fall Wolfart" oder das Problem mit dem magnetischen Rapport. Zur Entwicklung der therapeutischen Beziehung in den Anfängen der Psychotherapie. Hypnose-ZHH 9 (1 and 2), 169–207.
Peter, B., Piesbergen, C., Lucic, T., Staudacher, M., Hagl, M., 2013. The role of tactile support in arm levitation. American Journal of Clinical Hypnosis 56 (2), 115–142.
Peter, B., Schiebler, P., Piesbergen, C., Hagl, M., 2012a. Electromyographic investigation of hypnotic arm levitation: Differences between voluntary arm elevation and involuntary arm levitation. International Journal of Clinical and Experimental Hypnosis 60 (1), 88–110.
Peter, B., Vogel, S.E., Prade, T., Geiger, E., Mohl, J., Piesbergen, C., 2014. Hypnotisability, personality style and attachment. An exploratory study. Part 1: general results. American Journal of Clinical Hypnosis 57.
Pyka, M., Burgmer, M., Lenzen, T., Pioch, R., Dannlowski, U., Pfleiderer, B., et al., 2011. Brain correlates of hypnotic paralysis – a resting state fMRI study. NeuroImage 56 (4), 2173–2182.
Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., Shulman, G.L., 2001. A default mode of brain function. Proceedings of the National Academy of Sciences of the United States of America 98 (2), 676–682.
Rainville, P., Duncan, G.H., Price, D.D., Carrier, B., Bushnell, M.C., 1997. Pain affect encoded in human anterior cingulate but not somatosensory cortex. Science 277, 968–971.
Revenstorf, D., Peter, B. (Eds.), 2009. Hypnose in Psychotherapie, Psychosomatik und Medizin. Manual für die Praxis, second ed. Springer, Heidelberg.
Rossi, E.L. (Ed.), 1980. The Collected Papers of Milton H. Erickson, vols. 1–4. Irvington Publishers, New York.
Spiegel, D., Kosslyn, S.M., 2004. Glauben ist Sehen: Die Neurophysiologie der Hypnose. Hypnose und Kognition 21 (1 and 2), 119–137.
Spiegel, D., Scheflin, A.W., 1994. Dissociated or fabricated? Psychiatric aspects of repressed memory in criminal and civil cases. International Journal of Clinical and Experimental Hypnosis 42 (4), 411–432.
Szabó, C., 1996. Differences between direct and indirect hypnotic techniques. In: Peter, B., Trenkle, B., Kinzel, F.C., Duffner, C., Iost-Peter, A. (Eds.), Munich Lectures on Hypnosis and Psychotherapy. M.E.G.-Stiftung, München, pp. 171–175.
Tefikow, S., Barth, J., Maichrowitz, S., Beelmann, A., Strauss, B., Rosendahl, J., 2013. Efficacy of hypnosis in adults undergoing surgery or medical procedures: a meta-analysis of randomized controlled trials. Clinical Psychology Review 33, 623–636.
Ward, N.S., Oakley, D.A., Frackowiak, R.S.J., Halligan, P.W., 2003. Differential brain activations during intentionally simulated and subjectively experienced paralysis. Cognitive Neuropsychiatry 8, 295–312.
WBP – Wissenschaftlicher Beirat Psychotherapie nach § 11 PsychThG, 2006. Gutachten zur wissenschaftlichen Anerkennung der Hypnotherapie. Hypnose-ZHH 1 (1 and 2), 165–172.

Relevant Websites

http://www.asch.net.
http://www.ishhypnosis.org.
http://www.meg-hypnose.de.
http://www.meg-stiftung.de.
http://www.sceh.us.

Hypnosis, Neural Basis of

Victoria Holec and Ian Q Whishaw, Canadian Centre for Behavioural Neuroscience, University of Lethbridge, Lethbridge, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.A. Karlin, volume 10, pp. 7101–7105, © 2001, Elsevier Ltd.

Abstract

Hypnosis has long been used as a healing practice and is still used for therapeutic purposes. Initially, hypnosis was studied in nonhuman animals and humans using electroencephalograms. New brain imaging techniques such as functional magnetic resonance imaging allow for a better understanding of neural activity patterns during hypnosis. These studies have revealed that attention and default network activity are altered in hypnosis. Specifically, activity patterns point toward a systems view of hypnosis that discounts a unitary state hypothesis. Future imaging research might be directed toward investigating these neural correlates within the framework of a two-system approach.

Introduction

The following article will give a brief overview of the origins of hypnosis. After the historical account, the therapeutic uses will be discussed along with an account of the current state of the literature on its neural correlates. Subsequently, animal hypnosis will be briefly described, followed by an account of the role of cognitive control and attention in hypnosis. The final sections will discuss some limitations of current methodologies for testing and a proposal for future research.

A Brief History of Hypnosis

The term hypnotism was coined by the English surgeon James Braid, the so-called *father of hypnosis*, in 1842 (Charcot, 1889). However, hypnosis had been used as a healing practice long before that time. Early accounts of its practice date back to 200 BC, when yogis in ancient India described transcendental meditation (Jaltare and Gopinath, 2012). Hypnosis was publicized during the time of the German physician Franz Anton Mesmer (1734–1815); it was then known as *animal magnetism*. This term coined by Mesmer described the reciprocal transference of energy that occurred between animate beings, which Mesmer believed to house a certain *life energy* or *ethereal fluid* (Crabtree, 2012; Waterfield, 2003). It was through Mesmer's highly controversial, but highly publicized treatment of diseases with magnetism, that hypnosis as a practice became popularized. Mesmer's work in the mid-eighteenth century can be considered the basis of Western hypnosis as it is contemporarily understood, and the colloquial term *mesmerized* originated directly from his name (Waterfield, 2003). Mesmer's healing technique involved sweeping motions of the hands over the body of a patient in order to apply what he thought was the power of the magnetic fluid (Crabtree, 2012).

Mesmer's student Puységur noticed that a magnetized person typically displayed six characteristics: (1) an altered state of consciousness that resembled sleepwalking; (2) a special connection with the person performing the magnetism; (3) suggestibility (i.e., the performance of suggestion and the susceptibility to such suggestion); (4) amnesia of the events occurring during magnetism; (5) ability to read the thoughts of the magnetizer; and (6) personality change during magnetism (Crabtree, 2012). Puységur further described this altered or divided state of consciousness as *magnetic sleep* or *somnambulism*, as he believed that this state was the same as natural sleepwalking. He did not himself believe in the magnetic fluid that was central to Mesmer's theory, and instead focused on the psychological aspect of rapport between the magnetizer and the patient. Around the time Puységur started his own practice in 1784, Mesmer's healing techniques were deemed nonscientific by a French scientific commission who argued that Mesmer's practice dealt with imagination, not physical entities (White, 1941).

In the nineteenth century, when the English surgeon Braid had learned about magnetism, the terminology shifted from magnetism to hypnotism (Crabtree, 2012). In fact, Braid initially used the term neurohypnotism, of which hypnotism is an abbreviation. He believed that the phenomena produced by hypnosis were purely physiological, and that the mind slipped out of gear when fatigued by any lasting sensation. It was in that state that the patient was rendered susceptible to hypnotic direction. As Crabtree (2012) points out, hypnotism was not popular in England during Braid's lifetime, but was readily adopted in France and in the United States.

At the turn of the twentieth century, hypnosis became popularized through the Nancy School of hypnosis initially through Hippolyte Bernheim, whose focus was the role of suggestion in hypnosis. Other European neurologists and psychologists, including the French neurologist Jean-Martin Charcot at the Salpêtrière hospital and his student Sigmund Freud, soon opposed Bernheim's focus on suggestion. Conversely, Charcot considered hypnosis a neurosis, the states of which were naturally derived, and not brought about by suggestion. Charcot delineated three different states: (1) cataleptic state, (2) lethargic state, and (3) somnambulistic state. At this time, the terminology of neurosis was heavily used in the new psychoanalytic theory, of which Freud and Charcot were avid proponents. Indeed, Charcot

considered hypnosis an artificially created neurosis, which was identical with hysteria (Crabtree, 2012). Both the Nancy and Salpêtrière schools of thought coexisted at the time, but it was Bernheim's theory that resulted in the dominant views about hypnosis and suggestion.

Since the mid-1900s, increased efforts have been directed at the scientific study of hypnosis. The American psychologist Robert White (1941) saw hypnosis as an experimental technique by which to reorganize the behavior of a person, which in turn allowed for the study of artificially reproduced behaviors and a variety of psychological problems. White also highlighted the fact that hypnosis allows for studying individual differences in behavior, as the technique generally did not elicit the same phenomena among individuals. He proposed that hypnosis be part of social psychology, and asked scientific questions about the phenomena that it encompasses. In short, White (1941) considered hypnosis as a meaningful, goal-directed striving. In the 1990s, a sociocognitive perspective on hypnosis purported by the American psychologist Nicholas Spanos emerged (Crabtree, 2012). His explanations of hypnosis held that the patient or participant enacts a role. The experimenter and participant subsequently interact during the state of hypnosis via subtle cues.

Today, researchers continue to attempt to understand the underlying mechanisms of hypnosis. One controversy that has remained is whether hypnosis presents an altered state of consciousness, or whether it is an ordinary psychological process related to others including role enactment, imagination, relaxation, and attention (Jaltare and Gopinath, 2012; Karlin and Orne, 2001; Kluft, 2012). There have been two camps of hypnosis researchers. The first camp proposes that hypnosis is a discrete, unique state of consciousness (Raz, 2011). The second camp ascertains that hypnosis is a form of common wakefulness rather than a discrete state. This debate will be touched upon again toward the end of the article. First, however, some contemporary uses of hypnosis as a therapeutic means will be discussed.

Therapeutic Hypnosis

Hypnosis is currently used for various therapeutic treatments, including pain management and cessation of addictive behaviors. The following section will elucidate the role of hypnosis in such treatments in greater detail, as it is widely used and known to laypeople through its therapeutic applications. In Mesmer's days, both mineral magnetism, which was performed using magnets, and animal magnetism were used as therapy for ailments such as convulsions, blindness, asthma, edema, and other nervous complaints (Oon, 2008). Mesmer himself acknowledged that there were some limitations to which illnesses or symptoms could be cured via hypnosis. These were largely restricted to nervous disorders, whereas *organic* or permanent conditions could benefit indirectly, such as via improvement of the patient's mood (Oon, 2008).

Pain management is one area in which hypnosis is most often used. A review by Askay et al. (2009) highlighted the use of virtual reality hypnosis in pain management. The authors explained the advantages of virtual reality hypnosis, including: (1) diminished dependency on a trained, skilled hypnotist; (2) cost-effectiveness; and (3) superiority to audio hypnosis. In virtual reality hypnosis for pain treatment, a patient enters a virtual, three-dimensional world via virtual reality technology. Askay et al. (2009) argued that the patient thereby enters a *presence*, attention becomes focused toward that presence and away from the pain. For instance, research also found that virtual reality hypnosis was successful for treatment of acute pain after burns (Patterson et al., 2004) and of chronic neuropathic pain (Oneal et al., 2008). Further, in participants with low hypnotizability (i.e., the trait measure assessing a person's ability to be hypnotized; Hilgard and Hilgard, 1975), hypnotizability was enhanced through virtual reality hypnosis (Askay et al., 2009). The authors proposed that this feature of attention capture combined with hypnotic suggestion may be an ideal way to approach pain management. Indeed, in healthy participants undergoing pain stimulation, virtual reality hypnosis was associated with a reduced pain-related functional magnetic resonance imaging (fMRI) signal in the anterior cingulate cortex, insula, thalamus, and primary and secondary somatosensory cortices (Hoffman et al., 2004).

Another way hypnosis can be used is by administering it oneself. That is, rather than having a dedicated hypnotist perform the suggestion, patients are given the tools to hypnotize themselves. Self-hypnosis as a form of treatment for labor pain in women giving birth was recently investigated by Werner et al. (2012). While successes with this method had been previously reported (Martin et al., 2001; Mehl-Madrona, 2004), in this randomized, controlled trial, women using self-hypnosis as a pain management strategy were still found to use as many conventional pain management methods (e.g., epidural analgesia) as women who did not use self-hypnosis. Self-hypnotized women also reported a similar experience of labor pain compared to those who did not use self-hypnosis. The authors concluded that in previous studies, participants were self-selecting hypnosis for treatment of labor pain, and therefore may have had systematically different characteristics prior to the intervention. The use of a randomized, controlled, large-size trial in the study by Werner et al. (2012), however, did not allow for a self-selection bias, and thus may not have found any treatment effects.

A recent randomized, controlled study investigated the underlying mechanisms of hypnotic intervention in pain management (Goodin et al., 2012). The authors examined pain ratings, as well as pain-provoked changes of cortisol and the soluble tumor necrosis factor-α receptor II (sTNFaRII) after intervention with hypnosis. Highly susceptible participants were randomly assigned to an experimental group who received intervention with hypnosis, or to a control group. The authors found that pain ratings were significantly reduced in those participants undergoing hypnosis, but neither cortisol nor sTNFaRII levels were reduced in reaction to acute pain. Because effect sizes were moderate in the pre- to post-intervention assessment, the authors did not rule out a role of the hypothalamic-pituitary-adrenal (HPA) axis altogether and concluded that future research was needed to further investigate how hypnosis might affect chronic pain as evidenced by a dysfunctional HPA axis.

Jensen et al. (2012) reviewed the use of functional neuroimaging to evaluate pain treatments including hypnosis. The authors stated that psychological pain treatments have gained

greater popularity either as a standalone method or in combination with pharmacological treatments. Among these treatments are distraction techniques, clinical hypnosis, and mental imagery. In typical hypnotic suggestion for pain relief, the patient initially receives suggestions for: (1) relaxation, followed by subsequent suggestions for one or more of (2) distracted attention, (3) positive affect, and (4) heightened self-efficacy to install a feeling of control over the pain (Green et al., 2005; Jensen et al., 2012). As this article reviewed some of the neuroimaging techniques including fMRI and positron emission tomography (PET), it allowed for an indication of underlying neural correlates. From this review, the following correlates of psychological pain modulation became evident in healthy participants undergoing pain stimulation: dorsolateral and ventrolateral prefrontal cortex (dlPFC, vlPFC), anterior cingulate cortex (ACC), and periaqueductal gray (PAG). In particular, ACC activity was shown to correlate with modifying pain unpleasantness, and activity in primary somatosensory cortex (SI) has been shown to correspond with modifying pain intensity (Rainville et al., 2002). However, in patients with chronic pain such as fibromyalgia or chronic back pain, when pain was modulated via hypnosis, no corresponding activity specifically in ACC, but rather a unique change in overall activity patterns occurred (Derbyshire et al., 2009; Nusbaum et al., 2011).

Lastly, Hill (2010) reported several studies by Ernest Rossi, who unprecedentedly investigated the epigenetic changes associated with hypnotic intervention, which he initially called *psychosocial genomics*. Genes related to well-being and health were reportedly expressed in the hours and days after hypnotic intervention. Using deoxyribonucleic acid (DNA) microarrays, Rossi and colleagues (Atkinson et al., 2010; Rossi et al., 2008) were able to assess epigenetic changes brought about by experience-dependent gene expression. Blood samples were obtained from participants to construct DNA microarrays prior to hypnosis, 1 h following hypnosis, and 24 h following hypnosis. After participants underwent hypnosis, epigenetic changes were assessed, as were subjective ratings of hypnosis using the Stanford Hypnotic Susceptibility Scale (SHSS; Hilgard, 1965). In a pilot study (Rossi et al., 2008), upregulation of 15 genes was found. Some of these genes were characteristic of stem cell growth and proliferation, plasticity associated with memory, learning, and novelty, and stress reduction (Hill, 2010).

In summary, while hypnosis seems to be effective for treatment of pain in healthy participants who undergo pain stimulation, the story tends to be more complicated for those with chronic pain (Jensen et al., 2012; Nusbaum et al., 2011) or acute labor pain (Werner et al., 2012). Additionally, physical markers of pain modulation such as cytokines and cortisol have not shown modulation through hypnosis (Goodin et al., 2012). However, pain management via hypnosis can possibly elucidate some of the underlying neural correlates of hypnosis. In particular, in healthy participants undergoing pain stimulation, distinct neural activity markers were found during pain treatment (Rainville et al., 2002). Specifically, many areas involved in the so-called *default network* tend to show activity during hypnotic pain modulation. In the following section, the role of cognitive control and attention in hypnosis will be discussed.

Hypnosis, Consciousness, and Cognitive Control

The State Debate

As mentioned earlier, the contemporary debate about hypnosis pertains to whether it is a discrete state or a continuum of an existing state. That is, are people who are hypnotized in a distinct state, which is different from other states of consciousness, or can the performances of those hypnotized be equally well achieved by others not hypnotized? Proponents of the state theory purport that hypnosis is a distinct state of awareness, which is: (1) characterized by heightened suggestibility, (2) enhanced imagination, (3) reduced cognitive control, (4) involuntary behaviors, and (5) that such a state will one day be demonstrated on the neurophysiological level (Waterfield, 2003). Opponents of the state theory propose that: (1) responsiveness to suggestion is a normal psychological phenomenon, (2) individual differences in responsiveness are due to a person's expectations and attitudes toward hypnosis, (3) hypnotic phenomena are within human abilities, (4) involuntary behaviors have an alternative explanation, and (5) no neurophysiological proof for a hypnotic state will be found. Even though the state theory has faced much skepticism in the last four decades, as Waterfield (2003) pointed out, the antistate theorists thus far have not been able to *disprove* the state theory.

An early view of the state theory of hypnosis was motivated by the fact that hypnotic procedures result in bizarre behaviors (such as often seen in stage hypnosis shows) that some perceived as exceeding the capabilities of nonhypnotized individuals (Crabtree, 2012). However, as Crabtree went on to explain, such instances are indeed not unique to hypnosis. He cited two examples. The first was of a patient at the dentist's office who under hypnotic suggestion did not report any pain while six of his teeth were pulled. The second was of a railway worker whose toes were severed when his foot was run over by a railway car, while he was hurrying to push a coworker out of the way of an oncoming train. He also reported no pain, and had no awareness that his toes were run over until several minutes after the rescue of his coworker. As can be readily seen from these accounts, bizarre behaviors alone do not qualify as a classification criterion for a hypnotic state.

One prominent view is that of Ernest Hilgard (Hilgard, 1992; Hilgard and Hilgard, 1975). He described a theory called neodissociationism, which stipulates that there is a hierarchy of control systems that operate at any given time in any given individual. Hypnosis serves the function of modifying this hierarchical arrangement so that some control systems become separated or *dissociated* from others. Abilities for different tasks of varying orders of magnitude, such as writing a letter or raising a family, are named *subsystems* (Waterfield, 2003). At any point in time, different subsystems may be active, whereas others may not be. A central control system monitors these subsystems. Then, dissociation occurs through separation of subsystems from each other or from the central control system (Waterfield, 2003). Hilgard (1992) also stated that the concept of consciousness is an illusion, as people engage in multiple behaviors all the time, and a conscious representation of this engagement is never complete. He saw awareness as a tool that shifted among aspects of these behaviors or of the external environment.

Especially interesting to Hilgard was the idea that some attentive processes can occur without awareness. Those were the processes that he described as dissociated from the conscious experience of the person.

In summary, the debate about whether hypnosis is a discrete state or lies on a continuum of normal psychological phenomena continues. The view of neodissociationism proposed by Hilgard and Hilgard (1975) provided a basis for studying neural correlates, which has been done since. The next section will discuss how initial electroencephalogram (EEG) studies in animals attempted to answer the question of whether hypnosis occurs as a dissociated neural correlate without its equivalent behavioral output. Further, the dissociationist view is present again in contemporary fMRI studies that focus on attentional processes and their dissociation from external events.

Toward a Neural Basis

Recent literature points toward the fact that the state debate has largely been a theoretical one, and that little attention has been paid to a neural basis (Crabtree, 2012; Deeley et al., 2012). To begin, one might ask the question about how a distinct state of hypnosis is separable from other states of consciousness. One pathway for answering this question may come from the field of animal hypnosis, which has been prominent since the time of Pavlov (Whishaw, 1984). The term animal hypnosis must not be confused with Mesmer's animal magnetism, which was derived from the Latin *animus*, spirit or mind. In animal hypnosis, animal actually refers to the living creatures. Specifically, animal hypnosis describes the immobilization of an animal, such as a rabbit, by holding it turned onto its back, until its' breathing becomes regular and it relaxes its limbs. The animal will stay immobile for a period of up to 5 min (Whishaw, 1984). The brain activity accompanying such a hypnotic state in an animal is similar to the waking state activity of an animal that is sitting still. One theory of what happens in the hypnotic state is that the initial restraint of the animal activates a neural system in the brain that inhibits other neural systems that produce movement responses, which renders the animal immobile (Klemm, 1971).

Whishaw et al. (1982) set out to test which of two different EEG types that are present in normal rabbits can be observed during hypnosis. The authors explained that in the 1960s with the increased use of EEG, a new *state hypothesis* of animal hypnosis became popular. This hypothesis postulated that the animal is in a dissociated state in which EEG activity patterns are disconnected from the usual motor activity that accompanies them. In the example of the rabbit, two types of brain activity typically correlate with the animal's movements such as hopping, turning, and head movements. These activity patterns are classified as: (1) low voltage fast activity (LVA) and (2) hippocampal rhythmical slow activity (RSA, now better known as theta). These patterns can occur during movements such as those described above, and during resting state, or immobility. Immobility activity is typically of lower frequency. Whishaw et al. (1982) hypothesized that if higher frequency EEG activity was present in the hypnotized rabbit, then one could speak of a state dissociation brought on by the uncoupling of movement and movement-dependent EEG. If, however, the slower activity was present, then one could speak of hypnosis as a nondistinct state, as this activity can spontaneously occur during resting activity. Results indicated that only slower EEG activity found spontaneously in immobilized rabbits was present in hypnotized rabbits. Higher frequency activity could not be elicited through brain stimulation without accompanying movements, so a dissociation could not be achieved.

These initial accounts of animal hypnosis suggest that the state dissociation hypothesis is not supported by evidence derived from EEG in animals. The next section will discuss some of the correlates seen in fMRI studies in humans during hypnosis. These studies will focus largely on the default network, which in humans is an activity pattern that correlates with resting state activity. Because hypnosis tries to achieve a type of resting state, this framework has received particular attention. In fact, overlapping neural correlates between hypnosis and the default network have been found.

The Default Network and Attention

The default network, or default mode network (DMN), describes a network of brain structures that are more active during low-demand tasks or the brain's resting state, rather than being associated with specific tasks (Deeley et al., 2012). It can be described as an internal system, which is linked to self-related processes, and stands in opposition to the extrinsic system, which is linked to external stimuli (Demertzi et al., 2011). Deeley et al. (2012) pointed out that there are actually several networks involved in resting state activity, including the dorsal attention network, the sensorimotor network, and the DMN. The core regions of DMN consist of medial PFC, superior frontal cortex, ACC, posterior cingulate cortex, precuneus, and retrosplenial cortex, lateral parietal cortex, inferior parietal lobule, lateral temporal cortex, and parahippocampal gyri (Deeley et al., 2012; Lipari et al., 2012; Stawarczyk et al., 2011). These structures comprise networks that pertain to executive function and attention.

Activity in the DMN has been shown to be reduced during the resting state. Here, participants were instructed to lie quietly in an alert state or view a passive stimulus such as a fixation cross (Greicius et al., 2003). The amount of activity decrease has been shown to directly relate to the amount of cognitive demand required by the cognitive task. That is, DMN activity seemed to persist during a low-demand task, but it was reduced during a high-demand task. Similarly, DMN activity has also been shown to be reduced under hypnotic suggestion (Deeley et al., 2012). Demertzi et al. (2011) explained that reduction in connectivity during hypnosis among DMN areas including posterior retrosplenial, cingulate, and precuneal areas might be due to a decreased demand for interaction with the external and a focus toward the internal world. This explanation was supported by other findings (Greicius et al., 2003). This reduced connectivity along the frontoparietal axis might thus function as a way to eliminate sensory awareness in hypnosis. Another recent study showed evidence that this altered connectivity may underlie hypnotizability, as participants with high hypnotizability showed greater activity in the DMN (Hoeft et al., 2012).

However, Deeley et al. (2012) conceded that the cognitive and subjective correlates of the DMN are generally difficult to assess, as experimental interference may elicit unwanted goal-directed activity in this network. Thus, the authors proposed that hypnosis lends itself well to measure task-unrelated activity and its behavioral or cognitive correlates. Because participants can report self-ratings of hypnotic depth and further describe their experience in the alert state, these subjective accounts may be correlated to DMN activity.

Lifshitz et al. (2012) described that there are similarities in DMN activity between meditation and hypnosis, as both processes seem to rely on focused attention. Focused attention is the sustained narrowing of attention toward an internal aspect, such as breathing. Meditation, like hypnosis, has been shown to alter the neural expression of pain and also the connectivity in the default network. In a previous study by the same group, it was found that posthypnotic suggestion to perceive printed words as meaningless led to an increase in performance on a Stroop word-color task. This was correlated with a decrease in information processing in the visual cortex as well as a reduction in ACC activity (Raz, 2005). On this task, participants have to voice out the ink color of a written word, and not let the actual word, which spells out a different color to the ink, distract them. The posthypnotic suggestion by Raz (2005) thus improved participants' ability to pay attention to the color, rather than the word, which corresponded to a reduction in DMN activity. Long-term meditators too show this reduction in activity. This finding serves as evidence that focused attention is one common denominator of modulation in DMN processes. Attention is the common factor in both hypnosis and DMN processes, and it may be the case that this commonality lies in focusing attention inward to disregard external stimuli.

In an unprecedented study, Lipari et al. (2012) investigated DMN activity in an individual described as a *hypnotic virtuoso*, a very highly hypnotizable person who typically displays all phenomena associated with hypnosis. The authors deemed that this was an optimal participant to investigate pure hypnosis (i.e., hypnosis in the absence of possible distractions elicited by suggestion or mental tasks used to enter hypnosis). This study was motivated by the debate of whether hypnosis is a distinct brain state or not. Because wide individual differences even in DMN activity are seen during hypnosis, the neural correlates of pure or unconditioned hypnosis might give an insight into whether there is a unitary state that can be classified as hypnosis. Changes in the DMN were recorded using EEG and fMRI during task-free mentation of the hypnotic virtuoso. During hypnosis, there was increased activity in posterior regions of the DMN, including precuneus, posterior cingulate cortex, retrosplenial cortex, and inferior parietal lobule. Activity in the prefrontal regions of the DMN was also enhanced. This was in contrast to many studies, including those cited above, that typically found an activity reduction in anterior DMN regions, such as ACC and medial PFC. The authors explained that this could be due to the general mild hypnotizability of participants used in previous studies. In addition, this finding could be due to focusing attentional resources. This is in line with the internal focus of attention as it is steered away from external stimuli. The gathering of attentional resources may elicit activity in attentional networks, which are said to include PFC. In that sense, this case study of pure hypnosis may have provided a clearer picture of DMN activity during hypnosis.

As mentioned earlier, individual differences across participants in hypnosis have made it difficult to outline common features. Hoeft et al. (2012) investigated the functional basis for hypnotizability and found that many of these individual differences were due to the degree of participants' hypnotizability. Individuals who were highly hypnotizable showed greater activity in the dorsolateral PFC, which is a main hub in executive control. Further, there was increased connectivity between ACC and dorsolateral PFC in highly hypnotizable participants during the resting state. Compared with the resting state, highly hypnotizable individuals also showed decreased activity in ACC and dorsolateral PFC in hypnosis.

Priftis et al. (2011) tested the neuropsychological phenomenon of neglect in a posthypnotic suggestion experiment. Neglect typically follows damage to the parietal-frontal network in the right hemisphere, resulting in a complete unawareness of the contralateral (i.e., left) side of space and self. The authors explained that visuospatial attention can be deployed in two ways: (1) through a first subsystem, which is localized in the left hemisphere, and orients visuospatial attention from the left to the right, and (2) through a second subsystem, which is localized in the right hemisphere and orients visuospatial attention from the right to the left. The left-hemisphere subsystem is said to be more powerful. Neglect could thus be explained by citing these two subsystems. According to Priftis et al. (2011), neglect would be a result of ipsilesional hyperattention, rather than contralesional hypo-attention. To test this idea, posthypnotic suggestion was used in half the participants engaging in an attentional task. Indeed, neglectlike behavior was induced in participants under post-hypnotic suggestion. They further observed an asymmetrical allocation of attention to the right side of space, indicated by the fact that attention to the right visual field resulted in slower reaction times for stimuli shown in the left field compared to attention to the left visual field for stimuli shown in the right field. This finding supported the fact that neglect might be due to an ipsilesional hyperattention. It further corroborated the role of attention to hypnosis. During hypnosis, attention to an entire visual field was induced.

In a resting-state fMRI study (Pyka et al., 2011), the experimenter induced a left-hand paralysis in participants via hypnotic suggestion. Areas of the DMN were of particular interest, and fMRI analysis was based on these regions. Results showed that there was increased connectivity of the precuneus with the right dorsolateral PFC and angular gyrus. However, connectivity of the medial PFC and primary motor cortex was unchanged. The authors concluded that the precuneus was important for maintaining an altered state of consciousness.

Processes related to the DMN seem to be changed in hypnosis. This may indicate that hypnosis renders cognitive control processes such as attention to external stimuli unnecessary. Rather, cognitive control seems to be directed inward, which can explain some higher activation in the DMN toward hypnosis as observed by Lipari et al. (2012). Instead of surrendering to hypnosis, individuals indeed show enhanced control over sensorimotor function, but as Hoeft et al. (2012) point out, this is done with a sense of involuntariness and somewhat dissociated from the deliberation enabled through

cognitive control. This is consistent with the idea that through hypnotic suggestion, sensory experiences can be modulated, such as to decrease pain perception or reduce anxiety. Through this inward focusing of attention, it might be the case that awareness of external stimuli is diminished. Further, some attentional phenomena, such as visuospatial hemispheric neglect, can be induced via hypnosis (Priftis et al., 2011).

Even though many differences exist across studies in terms of methodology and in terms of which structures within or outside of the DMN show activity changes during hypnosis, there are some generalizations that can be derived from these accounts. First, activity changes in the DMN occur. In some cases, these changes present as an activity increase, and in other cases, as a decrease. There is no unequivocal consensus on which structures exactly are involved, but they typically present a subset of DMN structures. Second, attentional processes play a role in hypnosis. Attention is not a unified construct. Instead, many attentional processes can occur simultaneously. Attention can be focused on one stimulus, while others might go completely unnoticed (Crabtree, 2012; Hilgard, 1992). In hypnosis, it seems to be the case that some attentional processes are dissociated from external events and directed inward, as they occur in meditation. Third, cognitive control seems uncoupled from other systems in the brain, such as sensorimotor, visual, or auditory systems. Hypnosis can produce involuntary behaviors that are feasible during the normal state, but the executive control systems may be dissociated from them. That is, control over activation in subsystems in the brain is surrendered, or simply not attended to. Combined, these results point toward a systems view of hypnosis that discounts a distinct or unitary state hypothesis. Activity as found in hypnosis can be found in normal waking states as well. It seems to be the case, however, that during hypnosis, the executive regions that are implicated in the DMN and said to control attention, planning, volition, and behavioral flexibility, do not consolidate such subsystem activation.

A Note on Neuroimaging and EEG

With the advance of fMRI, hypnosis became one of the behaviors heavily investigated through neuroimaging. In fact, a majority of recent experiments use fMRI to examine neural correlates of hypnosis. Sometimes, fMRI is used in conjunction with PET or EEG. Karlin and Orne (2001) described EEG in detail, as this method was the prevailing one prior to the dawn of fMRI to study hypnosis. One method in EEG is to present stimuli to a participant, which result in acute alterations in activity. These changes are termed *event-related evoked potentials*, as they correlate directly to the event that is occurring in the environment, such as an auditory or visual stimulus. To distinguish event-related potentials from noise, stimulus-related activity is subtracted from a baseline of spontaneous signal or artifacts from the EEG recording. It has been shown that highly hypnotizable participants show brain activity consistent with self-reports of hallucinations during hypnosis (Crawford, 2001). Specifically, these participants showed increased or faster positive evoked potentials and decreased or slower negative evoked potentials.

Some caveats of neuroimaging studies must be kept in mind. As Raz (2011) pointed out, there is a danger of taking neuroimaging studies on hypnosis at face value without realizing some of the limitations this technique poses. He explained that the analysis of fMRI images takes multiple steps and a lot of human judgment to arrive at conclusions. Further, fMRI images are noisy, actual signals are weak, and false or coincidental patterns are readily discovered. Careful consideration must also be applied to not commit the tempting mistake of reverse inferences. Those are inferences that extrapolate a certain mental state from the activation of a certain brain region. However, this is a dangerous undertaking, as one region can be involved in multiple tasks, and activity alone does not predict its responsibility for a certain behavior. Temporal resolution in fMRI is poor, as accumulation of the fMRI signal is incremental and takes time. However, on the upside, new techniques such as diffusion tensor imaging and transcranial stimulation that can be used in conjunction with fMRI allow for the opportunity of corroborating results through converging evidence. Further, novel imaging analysis software allows for a fine-grained examination of brain activity and considers activation patterns across individual voxels without averaging voxels together (Raz, 2011). This averaging can detect activity more easily, which may lead an experimenter to commit a type I error (i.e., concluding there is a difference between experimental conditions when there is not).

Proposal for Future Research

The psychologist Daniel Kahneman, recipient of Nobel Memorial Prize in Economic Sciences in 2002, has proposed that there are two modes of thinking, which he terms System 1 and System 2 (Kahneman, 2011). System 1 can be described as automatic, intuitive, and effortless. It is fast and outside of voluntary control, and constantly integrates and updates incoming information. System 2, on the other hand, can be described as controlled and willful. It follows rules, demands attention, requires mental effort, and is involved in complex computations. The two systems are competitive and complementing each other in a dynamic and interactive way.

Even though Kahneman does not directly relate these systems to brain correlates, it is likely that System 1 involves parietal-frontal networks that mediate implicit memory, whereas System 2 involves temporal-frontal circuits and explicit memory. Thus, areas involved in System 1 might be those that maintain short-term memory and automatic, unintentional memory processes such as primed recall and procedural memories. Areas involved in System 2 might be those that pertain to executive and attentional functions and planned decisions, as well as conscious, thought-out memory processes and semantic and episodic memories.

In hypnosis, it is possible that System 2 is initially required to direct attentional focus inward. This idea is supported by psychophysiological studies that show a requirement for increased mental effort in focused attention. Specifically, Kahneman and Beatty (1966) reported that pupillary dilation proportionally increased with the amount of effort required on a cognitive task. Such pupillary dilation is also evident during hypnotic suggestion, when the focusing of attention occurs (Rossi et al., 2008). While System 2 then maintains this state of focused attention, only System 1 is left to deal with external

stimuli, such as instructions for behavioral activity. Procedural processes thus would be easily elicited and carried out, thus exemplifying the state of suggestibility that often characterizes hypnosis. Future brain imaging research might direct efforts toward identifying the neural correlates of System 1 and System 2 and how activity changes within these systems during hypnotic suggestion.

See also: Attention, Neural Basis of; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Consciousness, Neural Basis of; Hemispatial Neglect, Neural Basis of; Hypnosis; Imaging Methods in Psychotherapy; Virtual Reality in Psychotherapy.

Bibliography

Askay, S.W., Patterson, D.R., Sharar, S.R., 2009. Virtual reality hypnosis. Contemporary Hypnosis 26 (1), 40–47.

Atkinson, D., Iannotti, S., Cozzolino, M., Castiglione, S., Cicatelli, A., Vyas, B., Rossi, E., 2010. A new bioinformatics paradigm for the theory, research, and practice of therapeutic hypnosis. American Journal of Clinical Hyponsis 53 (1), 27–46.

Charcot, J.M., 1889. Magnetism and hypnotism. Forum of New York 8, 566–577.

Crabtree, A., 2012. Hypnosis reconsidered, resituated, and redefined. Journal of Scientific Exploration 26 (2), 297–327.

Crawford, H., 2001. Neuropsychophysiology of hypnosis: toward an understanding of how hypnotic interventions work. In: Burrows, G., Stanley, R., Bloom, P. (Eds.), Handbook of Clinical Hypnosis. Wiley, New York, pp. 61–84.

Deeley, Q., Oakley, D.A., Toone, B., Giampietro, V., Brammer, M.J., Halligan, P.W., 2012. Modulating the default mode network using hypnosis. International Journal of Clinical and Experimental Hypnosis 60 (2), 206–228.

Demertzi, A., Soddu, A., Faymonville, M.E., Bhari, M.A., Gosseries, O., Vanhaudenhuyse, A., Laureys, S., 2011. Hypnotic modulation of resting state fMRI default mode and extrinsic network connectivity. Progress in Brain Research 193, 309–322.

Derbyshire, S.W., Whalley, M.G., Oakley, D.A., 2009. Fibromyalgia pain and its modulation by hypnotic and nonhypnotic suggestion: an fMRI analysis. European Journal of Pain 13 (5), 542–550.

Goodin, B.R., Quinn, N.B., Kronfli, R., King, C.D., Page, G.G., Haythornthwaite, J.A., McGuire, L., 2012. Experimental pain ratings and reactivity of cortisol and soluble tumor necrosis factor-α receptor II following a trial of hypnosis: results of a randomized controlled pilot study. Pain Medicine 13, 29–44.

Green, J.P., Barabasz, A.F., Barrett, D., Montgomery, G.H., 2005. Forging ahead: the 2003 APA division 30 definition of hypnosis. International Journal of Clinical and Experimental Hypnosis 53, 259–264.

Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V., 2003. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. Proceedings of the National Academy of Sciences 100 (1), 253–258.

Hilgard, E.R., 1965. Hypnosis. Annual Review of Psychology 16, 157–180.

Hilgard, E.R., 1992. Divided consciousness and dissociation. Consciousness and Cognition 1, 16–31.

Hilgard, E.R., Hilgard, J.R., 1975. Hypnosis in the Relief of Pain. William Kaufmann, Inc, Los Altos, CA.

Hill, R., 2010. An introduction to psychosocial genomics: how the body speaks to us about the effects of noninvasive processes such as therapeutic hypnosis. Australian Journal of Clinical Hypnotherapy and Hypnosis 30 (2), 5–16.

Hoeft, F., Gabrieli, J.D.E., Whitfield-Gabrieli, S., Haas, B.W., Bammer, R., Memon, V., Spiegel, D., 2012. Functional brain basis of hypnotizability. Archives of General Psychiatry 60 (10), 1064–1072.

Hoffman, H.G., Richards, T.L., Coda, B., Bills, A.R., Blough, D., Richards, A.L., Sharar, S.R., 2004. Modulation of thermal pain-related brain activity with virtual reality: evidence from fMRI. Neuroreport 15 (8), 1245–1248.

Jaltare, K., Gopinath, J.K., 2012. An exploratory study on the effects of hypnosis on the subjective well being of post graduate students in Bangalore. Indian Streams Research Journal 2 (7), 1–6.

Jensen, K.B., Berna, C., Loggia, M.L., Wasan, A.D., Edwards, R.R., Gollub, R.L., 2012. The use of functional neuroimaging to evaluate psychological and other non-pharmacological treatments for clinical pain. Neuroscience Letters 20, 156–164.

Kahneman, D., 2011. Thinking, Fast and Slow. Doubleday, Canada.

Kahneman, D., Beatty, J., 1966. Pupil diameter and load on memory. Science 154 (3756), 1583–1585.

Karlin, R.A., Orne, E., 2001. Neural basis of hypnosis. International Encyclopedia of the Social and Behavioral Sciences, 7101–7105.

Klemm, W.R., 1971. Evoked responses in brain motor areas during the immobility reflex ("animal hypnosis"). Physiology & Behavior 6, 137–144.

Kluft, R.P., 2012. Hypnosis in the treatment of dissociative identity disorder and allied states: an overview and case study. South African Journal of Psychology 42 (2), 146–155.

Lifshitz, M., Campell, N.K.J., Raz, A., 2012. Varieties of attention in hypnosis and meditation. Consciousness and Cognition 21, 1582–1585.

Lipari, S., Baglio, F., Griffanti, L., Mendozzi, L., Garegnani, M., Motta, A., Pugnetti, L., 2012. Altered and asymmetric default mode network activity in a "hypnotic virtuoso": an fMRI and EEG study. Consciousness and Cognition 21, 393–400.

Martin, A.A., Schauble, P.G., Rai, S.H., Curry Jr., R.W., 2001. The effects of hypnosis on the labor processes and birth outcomes of pregnant adolescents. Journal of Family Practice 50, 441–443.

Mehl-Madrona, L.E., 2004. Hypnosis to facilitate uncomplicated birth. American Journal of Clinical Hypnosis 46, 299–312.

Nusbaum, F., Redoute, J., Le Bars, D., Volckmann, P., Simon, F., Hannoun, S., Sappey-Marinier, D., 2011. Chronic low-back pain modulation is enhanced by hypnotic analgesic suggestion by recruiting an emotional network: a PET imaging study. International Journal of Clinical and Experimental Hypnosis 59 (1), 27–44.

Oneal, B.J., Patterson, D.R., Soltani, M., Teeley, A., Jensen, M.P., 2008. Virtual reality hypnosis in the treatment of chronic neuropathic pain: a case report. International Journal of Clinical and Experimental Hypnosis 56 (4), 451–462.

Oon, Z., 2008. A critical presentation of the life and work of Frank Anton Mesmer MD and its influence on the development of hypnosis. European Journal of Clinical Hypnosis 8 (1), 32–40.

Patterson, D.R., Tininenko, J.R., Schmidt, A.E., Sharar, S.R., 2004. Virtual reality hypnosis: a case report. International Journal of Clinical and Experimental Hypnosis 52 (1), 27–38.

Priftis, K., Schiff, S., Tikhonoff, V., Giordano, N., Amodio, P., Umilta, C., Casiglia, E., 2011. Hypnosis meets neuropsychology: simulating visuospatial neglect in healthy participants. Neuropsychologia 49, 3346–3350.

Pyka, M., Burgmer, M., Lenzen, T., Pioch, R., Dannlowski, U., Pfleidere, B., Konrad, C., 2011. Brain correlates of hypnotic paralysis – a resting-state fMRI study. Neuroimage 56, 2173–2182.

Rainville, P., Hofbauer, R.K., Bushnell, M.C., Duncan, G.H., Price, D.D., 2002. Hypnosis modulates activity in brain structures involved in the regulation of consciousness. Journal of Cognitive Neuroscience 14 (6), 887–901.

Raz, A., 2005. Attention and hypnosis: neural substrates and genetic associations of two converging processes. Journal of Clinical and Experimental Hypnosis 53 (3), 237–258.

Raz, A., 2011. Does neuroimaging of suggestion elucidate hypnotic trance? International Journal of Clinical and Experimental Hypnosis 59 (3), 363–377.

Rossi, E., Iannotti, S., Cozzolino, M., Castiglione, S., Cicatelli, A., Rossi, K., 2008. A pilot study of positive expectations and focused attention via a new protocol for optimizing therapeutic hypnosis and psychotherapy assessed with DNA mircoarrays: the creative psychosocial genomic healing experience. Sleep and Hypnosis 10 (2), 39–44.

Stawarczyk, D., Majerus, S., Maquet, P., D'Argembeau, A., 2011. Neural correlates of ongoing conscious experience: both task-unrelatedness and stimulus-independence are related to default network activity. PLoS ONE 6 (2), 1–14.

Waterfield, R., 2003. Hidden Depths: The Story of Hypnosis. Brunner-Routledge, New York.

Werner, A., Uldbjerg, N., Zachariae, R., Rosen, G., Nohr, E.A., 2012. Self-hypnosis for coping with labor pain: a randomised controlled trial. British Journal of Obstetrics and Gynaecology 120, 346–353.

Whishaw, I.Q., 1984. Animal hypnosis. In: Corsini, R.J. (Ed.), Wiley Encyclopedia of Psychology. John Wiley and Sons, New York (Reprinted in the Concise Encyclopedia of Psychology, 1987).

Whishaw, I.Q., Flannigan, K.P., Schallert, T., 1982. An assessment of the state hypothesis of animal hypnosis through an analysis of neocortical and hippocampal EEG in spontaneously immobile and hypnotized rabbits. Electroencephalography and Clinical Neurophysiology 54, 365–374.

White, R.W., 1941. A preface to the theory of hypnotism. Journal of Abnormal and Social Psychology 36 (4), 477–505.

Hypnosis, Psychiatry of

Nicholas A Covino, Massachusetts School of Professional Psychology, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In its various incarnations, hypnosis is among the oldest of the clinical techniques in medicine and it has been the subject of research and theory by most major figures in psychiatry and psychology from Sigmund Freud to the present. History reveals a series of cycles of ascent and decline in popularity for hypnosis along with outlandish claims of cures balanced by serious scholarly investigation. Individuals with a variety of medical and psychological ailments have been assisted by hypnotic interventions. However, despite its seniority as a treatment modality, numerous methodological and definitional issues plague those who study this area. This article will define the topic, review some of the history and evolution of the technique, outline some of the areas in which hypnosis is thought to be an effective adjunct to clinical care, and present some of the methodological complexities and limitations in the general application of hypnosis.

Definition

This treatment technique is thought to have received its name from the English surgeon James Braid (1795–1860) who used the Greek term 'ὕπνOς' (sleep) to describe his clinical observations of the physiological state of those whom he treated with this technique. Hypnosis is the direct descendant of the seventeenth- and eighteenth-century practices then called 'Magnetism,' 'Somnambulism,' and after a charismatic healer 'Mesmerism' although its roots can be traced to the religious and healing arts of antiquity. Today practitioners of hypnosis adopt the American Psychological Association's (1993) definition of hypnosis as a "procedure wherein changes in sensations, perceptions, thoughts, feelings or behaviors are suggested."

Relevant History

Chroniclers of hypnosis (see Ellenberger, 1970; Gauld, 1992) point to the work of Franz Anton Mesmer (1734–1815) as the cornerstone for the establishment of this technique in modern medicine. Mesmer's belief that an imbalance in magnetic fluid led to disease states in humans gave rise to the ritualized application of magnets to individuals with a variety of medical and psychological problems. Not only the metal apparatus, but also certain people were thought to possess the 'animal magnetism' to restore the magnetic equilibrium and health. French government and scientific investigators who tested his work were able to dispute Mesmer's disease theory, but not the results. People continued to recover from a variety of ailments by means of suggestions (i.e., ideas) provided by charismatic individuals, and the technique soon spread throughout Europe and the USA.

Sigmund Freud (1856–1939) made his living as a young psychiatrist by treating patients, almost exclusively, with hypnosis. As a student of the famous neurologist Jean Martin Charcot (1825–93) and the physician Hippolyte Bernheim (1840–1919), he learned about the efficacy of an idea that could facilitate an action or lead to symptom relief. Freud's early technique employed suggestions to deny his patients' complaints energetically and to provide assurance and motivation for the performance of more adaptive behaviors. Freud also observed hypnotized subjects seeming to possess knowledge without realization, carrying out posthypnotic suggestions without awareness of the source, and blindly following the commands of the hypnotist. As a translator of both Charcot's and Bernheim's writings on suggestion, hysteria and hypnosis and his opportunity to witness the work of both scholars, Freud was highly familiar with hypnosis and he became convinced that there were powerful, unconscious mental forces extant in humans that influenced their behavior.

In several months of intensive study at Charcot's Salpêtrière, Freud learned that traumatic events create trance states in vulnerable patients that appeared similar to those created by hypnosis during clinical interview. His teachers believed that individuals vulnerable to autosuggestion trapped painful emotion in the nervous system that gave rise to psychiatric symptoms. Patients with hysteria, especially, seemed to suffer painful reminiscences of the trauma that were expressed cognitively, affectively, somatically, and behaviorally. Since they appeared to possess similar characteristics, hypnotic trance was employed to assist patients with hysteria to revisit the scene of the original trauma, reexperience the original affects for cathartic discharge and recover from their symptoms through catharsis and insight. This technique was the basis for Freud's uncovering treatment in the *Studies on Hysteria* (Breuer and Freud, 1893–1895) and hypnosis became one of the underpinnings of the theory and technique of psychoanalysis.

The power of suggestion, the influence of unconscious mental processes in humans, and the use of hypnosis to access memories were not the only contributions of hypnosis to psychoanalysis. Freud was fascinated by the question of 'influence' and what led people to follow the directions of others. He reasoned that group activity and hypnotic treatments encourage people to minimize individual interests, personal control, discernment, personality style, and so on, and it led people to develop intense emotional relationships with others for reasons that are difficult to account for objectively. Freud wrote that the rapport created between doctor and patient in hypnosis was comparable to that of tribal members for the chief and he used it as a paradigm for his

understanding of therapeutic influence and transference throughout his career.

Where Mesmer might be credited with taking hypnosis away from the exorcists and 'healers' and placing it within traditional medicine, Freud explored hypnosis for what it could offer as a psychotherapeutic technique and for what it might contribute to his model of the mind. However, his several reports that psychoanalysis began when he 'rejected' hypnosis relegated this technique to an inferior place among methods for mental health treatment. Very few of his followers would risk alloying the 'copper of suggestion with the pure gold of psychoanalysis.' Freud's efforts to distance himself and psychoanalysis from hypnosis had something of the same dampening effect as Mesmer's failure to validate magnetic theory with scientific study.

The birth of psychology and social science research techniques during the twentieth century brought new methods of inquiry into the essence and effects of hypnosis. The rise of interest in mind–body medicine also fueled clinical interest and investment in this area. Both of these developments, along with the creation of several professional journals and international societies for the study and practice of hypnosis, contributed to the continued development of this field.

Therapeutic Applications

Medicine

A technique that can alter sensations, perceptions, thoughts, feelings, and/or behavior can be especially helpful in today's medical setting (Covino and Frankel, 1999). Patients and practitioners have long experienced anecdotal reports of symptom relief from something that a physician said, did, or proffered. Researchers in the laboratory and clinical setting have been laboring to establish empirical support for the use of hypnosis with a variety of symptomatic complaints and to manage common emotions in the medical setting to facilitate treatment.

Most clinicians have witnessed the influence of a patient's perception on the subjective experience of pain. Where an objective physiological event almost always serves as the cause of an injury, the meaning that the person makes of the event determines the amount of suffering.

In experimental laboratories, hypnotized subjects have demonstrated an ability to tolerate electric shock, ischemic, and cold pressor pain better than their nonhypnotized peers or when they employed other pain-management strategies. A substantial body of research finds that hypnotic suggestions for analgesia, distraction, and dissociation are superior to placebo, relaxation, and cognitive interventions in the management of a variety of painful conditions. Clinical research finds success for hypnosis in the management of pain associated with surgery, cancer, reflex sympathetic dystrophy, burns, interventional radiology, angioplasty, bone marrow transplant, and colonoscopy (Patterson, 2010).

A number of experimenters have been successful at producing raucous breathing and decreased pulmonary function among a subset of asthma patients by means of suggestion alone. Whether these patients were invited to believe erroneously that they were encountering an allergen or breathing a bronchoconstricting drug, the idea alone was enough to affect their breathing adversely. Several well-designed studies have found that hypnosis training has a better impact on asthma symptoms, 'as needed' medication use, and pulmonary function tests than education, relaxation, guided imagery, and placebo inhalers. Furthermore, asthma patients who were trained to use hypnosis to manage their breathing have demonstrated a superior ability to tolerate methacholine challenge testing.

Allergists, dermatologists, and others have reported successful treatment of warts, eczema, psoriasis, and other skin conditions with hypnosis. In clinical studies, suggestions to imagine bathing in salty water, increasing warm and tingling feelings on the skin, reducing blood flow to infected areas, and general healing have assisted patients to relieve itchiness, reduce inflammation, and decrease the number and size of warts. Likewise, patients with severe burns have been able to refrain from scratching, tolerate dressing changes, and reduce their use of pain medications with the help of hypnosis.

Patients with irritable bowel syndrome experience emotional distress in the form of anxiety and anger that adversely impacts gastric secretion, motility, and pain. A considerable number of these patients find relief from placebo interventions and several studies report alterations in gut physiology with hypnosis. Gastric acid secretion, gut motility, bowel habits, abdominal distention, and abdominal pain have been reported to improve with very brief hypnotic interventions aimed at increasing relaxation, controlling smooth muscle and gut activity, and bowel function.

Hypnotic training has been found to diminish the severity and frequency of bleeding episodes among patients with hemophilia. Training a small sample of patients to use hypnosis for relaxation and anxiety reduction has controlled factor usage, spontaneous bleeding and the need for transfusion. In addition, preoperative training in self-hypnosis for the control of anxiety and bleeding has been found to reduce the number of blood transfusions required by surgical patients.

One of the most common causes of noncompliance with chemotherapy among cancer patients is the significant distress that they experience from the associated nausea and vomiting. Antiemetic drugs often have their own side effects and are typically ineffective with the anticipatory emesis that is often a conditioned response. General anesthesia, also, commonly causes postoperative nausea and vomiting among surgical patients. Well-designed clinical interventions have found support for hypnosis training over relaxation techniques, distraction or individual counseling in the management of nausea and the frequency and intensity of emesis. This has been the case for pediatric as well as adult cancer patients.

Numerous methodological difficulties limit confidence and widespread application of hypnotic techniques in medicine, however, there is a substantive body of research to support its efficacy in pain relief (Montgomery et al., 2000; Patterson, 2010). As with most clinical research, the heterogeneity of the patients in terms of illness onset, type, course, previous treatments, comorbidity and so on, limits what can be said with confidence regarding the use of hypnosis in medicine. Also, many experimenters fail to include an adequate description of hypnosis, or might include hypnotic interventions as a component of more complex psychosocial interventions, or

fail to ascertain whether the subject was indeed hypnotized during the test. Nonetheless, qualified empirical evidence exists to support the efficacy of hypnotic techniques in the treatment of pain, asthma, dermatological disorders, irritable bowel syndrome, hemophilia, and hyperemesis (Pinnell and Covino, 2000). Furthermore, the use of hypnosis for certain surgical patients shortens operating time and increases speed of recovery (Lang and Rosen, 2002).

Psychotherapy

The addition of hypnosis to cognitive–behavioral treatments for anxiety disorders improves the effect of traditional psychotherapeutic interventions. The ability to alter perception, adjust expectations, increase motivation, and imagine more adaptive behaviors facilitates change, especially among those suffering from anxiety disorders. Where it was once thought that autonomic nervous system hyperarousal was the result of so-called stressful life events, recent research highlights the importance of cognitive processes such as perception. One person's strain might well be another's relief, depending upon the meaning that is made of the situation. An invitation from the hypnotist to perceive an event in a more benign fashion creates the opportunity for emotional relief and more adaptive behavior. Imaginative individuals can, also, use their capacity to picture frightening situations more vividly to expose themselves to feared stimuli successfully. These patients are able to achieve some of the same mastery as '*in vivo* exposure' without the time and travel. Behaviors can be mentally rehearsed, problems anticipated, and resolved with the help of suggestion, focused concentration and imagery. Likewise, unique motivational ideas can be the object of attention without distraction or criticism, and used to provide significant force for behavioral change.

Most efforts to use hypnosis for smoking cessation and weight control have seen modest results at best. Patients often report achieving relaxation and satisfaction with the technique, but pounds lost have been few and smoking abstinence is rarely achieved by more than a few program participants. This might well be due to the manner in which suggestions are employed to motivate patients and reduce appetite and desire. Much more is known these days about the complex manner in which patients cycle through processes of habit change as well as the benefits of medication support. Further work in this area would do well to integrate differing hypnotic approaches into the diverse phases and processes that people use in making habit change happen.

A number of clinicians have advocated the use of hypnosis for patients in psychotherapy. The hope in this area is to assist people to recover memories of difficult experiences as well as to teach a coping mechanism for managing extreme anxiety and dissociative symptoms. These therapists propose that the absorption, dissociation, and intensity of emotions that often accompanies hypnotic trance are similar to the experiences of many victims of trauma. In a manner similar to state-dependent learning, and in much the same way that Charcot and Freud employed this technique, patients are invited to experience hypnosis, revisit the initial trauma, and have an opportunity to express, integrate, and control their associated emotions. In addition, those who spontaneously experience dissociative states are taught to enter into a hypnotic trance to reduce anxiety and to achieve some ability to terminate or control hyperarousal.

In a fashion that mirrors Freud's later work, the use of hypnosis as a psychotherapeutic technique is the focus of considerable controversy. In general, as was the case with Freud, the main idea of memory recovery sits in the context of credibility and veridicality. For some patients and practitioners, it is painful and angering to raise the issue of truth in memory, since so many of those who were physically and sexually abused in their youth were frightened into secrecy. Yet, Freud seemed to see the inevitable intertwining of suggestion with dissociation in hypnosis as limiting the utility of this technique in psychotherapy.

Laboratory research finds that some modest gains in memory can be achieved from efforts to recall, especially, when these are made proximate to the experience. Cues, reviews, relaxation, and other motivational techniques also seem to aid the recovery of memory. Nonetheless, hypnotic techniques have generally failed to improve recall of word lists and eyewitness reports. On the other hand, mounting evidence suggests that people in clinical and experimental situations can be lead to the erroneous belief that an event or an experience did indeed occur. The fact that these circumstances are more likely to take place in situations of heightened emotionality, uncertainty, and unfamiliarity makes many clinicians cautious about using hypnosis as an 'uncovering' technique. Furthermore, the research evidence that exposure to hypnosis can both distort memory and increase subjects' confidence in erroneous memories has led courts of law to bar testimony obtained through the use of hypnosis.

Essential Elements and Putative Mechanisms

Despite its long history as a psychological intervention, it is not yet possible to define the mechanism of action of hypnosis. With increased access to sophisticated neuroimaging tools, there is enthusiasm for research that might define the brain mechanisms that facilitate hypnotic phenomena (Raz and Shapiro, 2002). However these studies have, thus far, failed to identify a specific neurological marker or neurophysiological substrate for hypnosis. There are, however, several ways that the beneficial effects of hypnosis have been understood.

Suggestions are ideas presented to a subject for uncritical acceptance. In hypnosis, the subject is asked to concentrate on these ideas in a manner that diminishes distraction, lessens logical analysis and maximizes acceptance and acquiesce. Suggestions are not commands, but invitations that can be rejected or accepted. However, like the response of patients to placebo medications, there is often a sense of involuntariness to hypnotic responses and an attribution of causation to the hypnotist. The physical context, psychological expectations, or the pressure to fit in well with a group can also provide direction for a person's thoughts, feelings, or behaviors. Some theorists maintain that these psychosocial factors alone or in combination result in successful hypnotic response.

The ability to become absorbed in an idea as if it is real and to use fantasy and imagery to alter cognition and/or behavior is

also seen as a determinant of successful hypnotic response. Such special concentration is thought to invite a state of dissociation where consciousness is divided. The ability to separate awareness of sensation, memory, time, behavior, and so forth is not only part of daily experience (e.g., shifting the gears of a car without awareness, daydreaming), but it can be created by hypnotists and imaginative patients. Such ability is thought to facilitate the use of hypnosis in management of painful conditions and anxiety. Since the higher cognitive faculties are blocked from elaborating pain or anxiety signals, emotional or behavioral responses are muted or stopped.

There are a number of clinical, and perhaps nonspecific, factors that facilitate or hinder the response of patients to hypnosis. Patients must be motivated to follow the suggestions of the hypnotist. A fearful, overly defensive, unwilling person will rarely succeed with hypnosis and one who feels safe, valued, and cared for by an 'expert' most likely will. Research indicates that hypnotizability appears to be a rather normally distributed, stable, trait. Although there are some that claim that anyone can achieve some benefit from hypnotic interventions, many hypnosis experts find that those with heightened ability are the most helped. Certain conditions favor suggestion such as ambiguity, uncertainty, novelty, and hyperemotionality. Good hypnotic technique endeavors to recruit or develop some of these elements.

See also: Freud, Sigmund (1856–1939); Hypnosis, Neural Basis of; Hysteria: History and Critiques; Psychoanalysis, History of; Psychological Treatment, Effectiveness of.

Bibliography

American Psychological, Division of Psychological Hypnosis, 1993. Cost analysis of adjunct hypnosis with sedation during outpatient interventional radiologic procedures. Hypnosis, Psychological Hypnosis 2 (3).

Breuer, J., Freud, S., 1893–1895. Studies on hysteria. In: Strachey, J. (Ed.), The Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 2. Hogarth Press, London, pp. 1–312.

Covino, N.A., Frankel, F.H., 1999. Clinical hypnosis. In: Nicoli Jr., A.M. (Ed.), The Harvard Guide to Psychiatry. The Harvard University Press, Cambridge, MA, pp. 552–556.

Ellenberger, H.F., 1970. The Discovery of the Unconscious: The History and Evolution of Dynamic Psychiatry. Basic Books, New York.

Gauld, A., 1992. A History of Hypnotism. Cambridge University Press, New York.

Lang, E.V., Rosen, M.P., 2002. Cost analysis of adjunct hypnosis with sedation during outpatient interventional radiologic procedures. Radiology 222, 375–382.

Montgomery, G.H., DuHamel, K.N., Redd, W.H., 2000. A meta-analysis of hypnotically induced analgesia: how effective is hypnosis? International Journal of Clinical and Experimental Hypnosis 48, 138–153.

Patterson, D.R., Jensen, M.P., 2003. Hypnosis and clinical pain. Psychological Bulletin 129, 495–521.

Patterson, D.R., 2010. Clinical Hypnosis for Pain Control. APA, Washington, DC.

Pinnell, C.M., Covino, N.A., 2000. Empirical findings on the use of hypnosis in medicine: a critical review. The International Journal of Clinical and Experimental Hypnosis 48, 170–194.

Raz, A., Shapiro, T., 2002. Hypnosis and neuroscience. A cross talk between clinical and cognitive research. Archives of General Psychiatry 59 (1), 85–90.

Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of

Yvonne Kuepper, Justus-Liebig-University Giessen, Giessen, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S. Levine, volume 10, pp. 7109–7113, © 2001, Elsevier Ltd.

Abstract

The Hypothalamic-pituitary-adrenal axis (HPA axis) has been shown to play a major role in the regulation of a variety of physiological as well as psychological processes. This article will give an overview over its structure and its diverse functions, such as basic bodily functioning, behavior, stress reactions or psychopathology. Furthermore, the underlying mechanisms as well as the diverse internal and external parameters which modulate basal HPA axis activity and reactivity as well those processes associated with HPA axis activity will be discussed.

The Hypothalamic-Pituitary-Adrenal Axis: Basic Concepts and Regulation

The hypothalamic-pituitary-adrenal axis (HPA axis) has been shown to play a major role in the regulation of a variety of physiological as well as psychological processes, most prominent among them being the stress reaction. Its primary effector-hormone in humans, cortisol, released from the adrenal cortex, is considered one of the major stress hormones. The HPA axis constitutes a closely regulated hormonal axis, which plays a critical role in maintaining homeostases and shows a high reactivity to internal as well as external stimuli.

On activation through a variety of stimuli, the paraventricular nucleus (PVN) within the hypothalamus secretes corticotropin-releasing hormone (CRH) and arginine-vasopressin (AVP) into the hypophyseal portal vein. CRH constitutes the major stimulus for the anterior pituitary to produce and secrete adrenocorticotropic hormone (ACTH) into the general circulation. AVP works in synergy with CRH in stimulating ACTH release. ACTH then stimulates glucocorticoid (GC) secretion, with cortisol being the primary GC in humans, and to a lesser degree mineralocorticoid (mainly aldosterone) secretion from the adrenal cortices (see Figure 1). The HPA axis is closely regulated through negative feedback loops; i.e., circulating GCs participate in their own regulation by inhibiting ACTH production in the pituitary as well as CRH production in the hypothalamus (see Figure 1).

Under normal resting conditions – that is in the absence of stressful stimuli – the release of CRH and AVP occurs in a pulsatile fashion and follows a clear circadian rhythm, which also is reflected in the rhythmic release of cortisol. This circadian rhythm is ultimately under the control of the suprachiasmatic nucleus (SCN, the primary endogenous pacemaker), via its connections to the PVN. In humans, cortisol levels throughout the day are characterized by showing the lowest levels around midnight, followed by a steady rise until awakening and reaching a peak within the first 30–60 min after awakening (cortisol awakening response). Levels continuously decline over the course of the day and finally reach the nadir again during sleep. This circadian rhythmicity may, however, be modulated by changes in internal or external conditions, such as time of food intake, light conditions and timing, shift work or activity status (e.g., physical activity like sports will increase cortisol levels); it is disrupted by stress. During acute stress reactions, CRH and AVP release is massively increased, resulting in augmented production of ACTH as well as GCs in turn, which are pivotal in allowing an adequate adaption process to stress. Cortisol increase in response to a stressor is comparably slow in healthy humans, starting approximately 10 min after onset of stress and being downregulated to baseline levels within the first hour after cessation of the stressor. GC release magnitude in healthy individuals, however, is limited by hormonal as well as neural mechanisms, as hyperproduction of GCs is involved in a variety of severely negative health-outcomes (see below for details). Animal models as well as studies in humans indicate that resilience to stress is associated with a rather rapid activation of the stress response plus its efficient deactivation; i.e., a close and effective control of HPA axis activity in the form of termination of GC release via negative feedback loops. A further important point to keep in

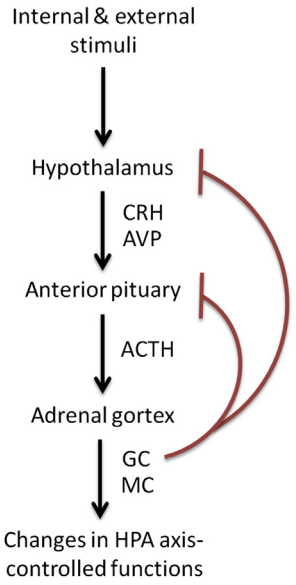

Figure 1 Simplified, schematic view of HPA axis. For details, please refer to the text; CRH: corticotropin-releasing hormone, AVP: vasopressin, ACTH: adrenocorticotropic hormone, GC: glucocorticoids, MC: mineralocorticoids.

mind when analyzing GCs and their psychobiological relevance is that a high proportion of serum GCs is bound to proteins and therefore biologically inactive. In humans, approximately 90% of plasma cortisol is bound to corticosteroid-binding globulin (CBG) and a further 7% to serum albumin, leaving only *c*. 3% free and therefore biologically active.

The effects of GCs are facilitated via the glucocorticoid receptor (GR), type II, and the mineralocorticoid receptor (MR), type I. Both receptors are part of the steroid receptor family; i.e., they constitute ligand-dependent transcription factors. Both are localized within the cytoplasm. Upon hormone binding, the receptor–hormone complex enters the nucleus and acts as a transcription factor on specific target sites in the genome – i.e., it is able to modulate gene transcription. Beyond these genomic effects, more rapid nongenomic effects have been described. These are thought to play an important role in coordinating a rapid adaptive response to stress. The mechanisms, however, by which this rapid nongenomic signaling is achieved are not fully understood (for review, see Groeneweg et al., 2012). GRs are found to be distributed widely throughout the mammalian central nervous system, most notably in the hypothalamus and the pituitary gland but also in parts of the limbic system such as the amygdala and hippocampus. MRs seem to be somewhat more restricted in their anatomical localization, having been shown in limbic structures; i.e., predominantly in the hippocampus but also in the amygdala.

HPA axis activity is controlled by a multitude of afferents that are thought to coordinate HPA axis activity status with internal and external stimuli. Of note, there appear to be different neural pathways that respond to different stimuli activating the PVN neurons and triggering HPA axis activation. From the brainstem pathways project directly into neurons of the PVN, with catecholamines promoting HPA axis activation in response to hemorrhage, respiratory disease, hypotension and potentially immune challenge as well. As mentioned above, the SCN takes part in controlling circadian rhythmicity via connections to the PVN. Further activational input might derive from connections to the amygdala, which is known to elicit stress responses and also plays a pivotal role in emotional regulation. Another limbic structure might also excite HPA axis activity: the bed nucleus of stria terminalis , which is involved in controlling basic homeostatic processes. Beyond direct limbic pathways, aminergic inputs from both the locus caeruleus, the main CNS site producing noradrenaline , as well as the raphe nuclei, which produce the monoamine serotonin (5-HT), may activate the HPA axis as well. Since direct neural pathways to the HPA from both systems are comparably limited, however, it is thought that more indirect pathways – e.g., through the more numerous innervations of limbic areas – are at work as well. (see Herman and Cullinan, 1997 for details).

HPA Axis: Diverse Functions and Clinical Correlates – A Short Overview

The effector hormones of the HPA axis, GCs and mineralocorticoids, play a role in the regulation of a variety of basic physiological functions and processes, such as carbohydrate metabolism, protein metabolism, fat metabolism, sleep, connective tissue, the hematopoietic system, the cardiovascular system, speed of signal transduction in the central nervous system, water and electrolyte balance and immune system activity. The proper functionality of the HPA axis is pivotal for maintaining bodily functions under normal conditions, keeping up homeostases and, thereby, health of the individual.

These diverse functions are also reflected in the variety of symptoms which might be caused in case of dysregulation of HPA axis activity. For example, hypercortisolism (also referred to as Cushing syndrome), which might be caused by exogenous factors, such as systemic administration of ACTH or corticoids during treatment of autoimmune diseases or allergies. Alternatively, it can also be caused by an increased production of ACTH in the pituitary, which results in increased corticoid production in turn (central Cushing syndrome). Adrenal Cushing's disease is characterized by an increased production of corticoids from the adrenal glands caused by neoplasia (adenomas or carcinomas), which is accompanied by an inhibition of ACTH release due to increased negative feedback-loop activation. Symptoms of hypercortisolism are diverse and often severe, underlining the relevance of HPA axis functioning, especially the need to closely regulate levels of fluctuating GC in the body (see Table 1). However, not only increased levels of cortisol are detrimental to health, this is also true for hypocortisolism which also has severe negative effects on one's health (see Table 1).

Furthermore, dysregulation of HPA axis activity has been shown to play a role in the etiology of several psychopathological conditions such as depression, burn-out syndrome, anxiety disorders, posttraumatic stress disorder (PTSD), borderline personality disorder, anorexia nervosa, schizophrenia or bipolar disorder. Of the variety of HPA axis-associated psychopathological conditions, depression probably is the most widely studied. Patients have been shown to exhibit marked changes in HPA axis activity as well as its regulation; Circadian rhythm of cortisol is often disturbed, hyperfunctionality of HPA axis and accompanying increased cortisol levels have been shown and the ability to down-regulate HPA-axis activity to baseline levels after exposition to a stressor additionally appears to be disrupted. In about 50–60% of patients suffering from acute clinical depression, such changes in HPA axis regulation might be found. Furthermore, successful antidepressant treatment is able to normalize these dysregulations. Interestingly, however, HPA axis activity is not fully normalized in all patients. This lack of normalization is a clinical indicator of the risk of relapse in patients. Today it is believed that the abnormal functionality of the HPA axis is one of the key pathological mechanisms in depression. In this context, cognitive impairments, which often accompany depression, are of interest as well. For one thing, they can be used as predictors of outcome of antidepressant therapy as well as risk of relapse. Furthermore, they often persist even when psychopathological symptoms are relieved. This lead to the assumption that these cognitive impairments are caused by changes in functional networks involved in the etiology of depression, which is further

Table 1 Hyper- and hypo-cortisolism: symptom overview

Hypercortisolism (Cushing's disease)
Obesity of the upper body
Thin arms and legs
Moon-mace (round, red, full face)
Acne or skin infections
Purple marks (1/2 inch or more wide), called striae, on the skin of the abdomen, thighs and breasts
Thin skin with easy bruising
Backache which occurs with routine activities
Bone pain or tenderness
Collection of fat between the shoulders (buffalo hump)
Thinning of the bones, which leads to rib and spine fractures
Weak muscles
Fatigue
Headache
Increased thirst and urination
Reduced sexual desire
Diabetes mellitus, type II
Mental changes, such as depression, anxiety or changes in behavior
In women: Excess hair-growth on the face, neck, chest, abdomen and thighs, and menstrual cycle that becomes irregular or stops
In men: erectile dysfunction
In children: reduced growth rate
Hypocortisolism (Addison's disease)
Muscle weakness
Muscle pain
Joint pain
Fatigue/weakness
Loss of appetite
Weight loss
Cravings for salt
Low blood pressure
Low blood sugar
Darkening skin
Changes in mood, depression and irritability, being emotionally hypersensitive
Diarrhea
Shakiness
Body aches
Sensitive skin
Nausea
Vomiting
Hypoglycemia

underlined by findings that show hypercortisolism to be associated with impaired cognitive functioning, especially memory deficits. One underlying mechanism appears to be hippocampal volume reduction, which is thought to be caused by chronically increased circulating GCs. These do not only cause atrophy of hippocampal neurons, but also inhibit long-term potentiation. Several lines of evidence suggest that these effects are mostly mediated by GRs.

Given the wide variety of health and disease related processes the HPA axis is involved in, it is not surprising that a wide body of research was concerned with finding those factors which determine HPA axis basal activity as well as reactivity. A major focus of research regarding the HPA axis and its psychobiological correlates has therefore been into individual differences in stress reactivity and associated psychopathological states. By now, several sources of individual differences in HPA functioning have been identified and will be discussed in the following paragraphs.

HPA Axis: Determinants of Individual Differences in Functioning

Interactions with Other Hormone- and Neurotransmitter Systems

When trying to understand the determinants of individual differences in HPA axis functionality, one has to keep in mind that the HPA does not act independently from other neuroendocrine systems. Indeed, it has been shown that HPA axis basal activity as well as reactivity is modulated by and itself modulates the activity of a variety of other hormone- and neurotransmitter systems, such as 5-HT, dopamine, noradrenalin, sex steroids, oxytocin and vasopressin or growth hormone. Therefore, a certain percentage of individual differences in HPA axis functioning must to attributable to the complex interactions with other systems.

A well-studied example of such complex interplay between systems is the bidirectional functional connectivity between the HPA axis and the sex steroids, the neuroendocrine basis of the reproductive system. GCs might inhibit sex steroid activity in multiple levels: by inhibiting the release of (1) gonadotropin-releasing hormone in the hypothalamus, (2) the *pituitary gonadotropins*, or (3) the gonads. Furthermore, GCs might cause a resistance to sex steroids in their target tissues. Within subjects exhibiting chronic HPA axis activation (e.g., patients with anorexia nervosa or competitive athletes) suppression of gonad activity has been shown, i.e., GCs do not only modulate sex steroid activity, but the sex steroids might modulate HPA axis activity as well. Estrogen response elements, for example, have been found in the promoter region of the CRH gene, indicating that estradiol might modulate CRH expression via its genomic effects. This is further supported by the fact that a direct stimulation of CRH gene expression has been shown. These and other results are thought to play a role in the underlying mechanisms of the well described sex-differences in HPA axis activity.

Growth hormone release has been demonstrated to be inhibited by prolonged HPA axis activation from stress. Interestingly, however, with the onset of a stressor growth hormone levels actually might be temporarily increased. This increase is thought to be mediated via the genomic effects of GCs, which bind at GC response elements localized in the promoter region of the growth hormone gene.

The peptide hormone oxytocin, which has been implicated in the regulation of social behaviors as well as emotion regulation and is thought, inter alia, to possess anxiolytic and stress-reducing properties has been shown to interact with HPA axis functionality. Oxytocin might reduce HPA axis responsivity to stress, and is thought that the stress buffering effect of social support is mediated via oxytocin and its ability to attenuate HPA axis stress reactivity. In humans, for example, intranasal application of oxytocin reduces cortisol responses during couple conflict. Interestingly, social support during a psychosocial stress situation as such is able to reduce not only subjective negative affective responses to stress but also cortisol

release; an effect which is even more pronounced after intranasal application of oxytocin.

Effects of Adverse Environmental Factors

A further source of variance in basal HPA axis activity as well as reactivity appears to be adverse environmental stimuli, especially experiences during early life.

The adverse effects of wearing circumstances of life or stressful minor and major life events on indicators of psychological and physical health have been shown repeatedly. As summarized in several review articles, a clear association has been shown between adverse environmental conditions and the occurrence as well as progression of several somatic (e.g., delayed wound healing, inflammatory disease, tinnitus, multiple sclerosis, coronary heart disease and others), somatoform as well as psychological illnesses (e.g., affective disorders, PTSD and others) (McEwen, 2006; Nater et al., 2006).

One underlying mechanism thought to play a role in explaining the detrimental effects of stress on physiological as well as psychological health, especially severe stressors experienced during early phases of development, is stress-induced abnormal functioning of the CNS; inter alia dysregulation of HPA axis basal activity as well as reactivity. Indeed, the acute stress burden of any given persons – e.g., problems related to work or family, chronic illness etc. – appears to modulate acute activity of the HPA axis. It has been shown that chronic stress might change the normal circadian rhythm of cortisol; i.e., is characterized by elevated basal cortisol levels. Furthermore, the neuroendocrine acute stress response appears to be reduced, and recovery times are augmented; i.e., downregulation of cortisol levels to baseline levels.

The effects of early stressful or traumatic environmental conditions, however, appear to be far more important, as they are often found to be associated with lifelong maladaptive changes in HPA axis regulation. A number of studies show that stressful experiences, especially during early life, might have long-lasting effects on HPA axis activity. For example, children who were subject to long-term separation from parents or lost their parents to death have been shown to exhibit reduced basal cortisol levels, nonresponsivity in the dexamethasone suppression test (indicating a dysfunctional feedback regulation of the HPA axis) as well as a hyperactive HPA axis. Further support for the assumption that stressful environmental stimuli might cause changes in HPA axis function stem from numerous studies in patients suffering from different forms of anxiety disorders (inter alia, posttraumatic stress disorder , panic disorder or social phobia) and with a history of a variety of early life stressors, who exhibit marked alterations in HPA axis activity; e.g., alterations in cortisol response to stress or changes in basal activity of CRH or abnormal reactions to the dexamethasone suppression test (for review, see Faravelli et al., 2012). Similar results have been shown in numerous animal studies, further underlying the strong adverse effects early life stress might have on HPA functionality. For example, rhesus macaques which have been reared in separation from their mothers, show substantially increased neuroendocrine stress responses when compared with mother-reared animals. These results are further supported by separation experiments in rodents. Particular noteworthy in the context of animal studies are those showing how HPA axis activity and its downstream effects on behavior as well as personality might be modulated by epigenetic changes caused by early environmental stimuli. In this context, a study on rats was able to demonstrate how maternal care might have long-lasting effects on the offspring's personality, behavior as well as HPA axis functioning, which were mediated by epigenetic changes in response to a specific environmental stimulus. The offspring of naturally high-nurturing rat mothers were compared to those of low-nurturing mothers. The adult offspring of low-nurturing mothers were characterized by increased anxiety and stress sensitivity, also reflected in heightened corticosterone secretion under stressful conditions as well as increased GR expression rates in the hippocampus. Subsequent analyses showed this increase in GR expression to be mediated through an increase in methylation markers in the promoter region of the GR gene, which appear to prevent binding of the transcription factor nerve growth factor-inducible protein A and those reducing transcription rate of the GR gene.

Taken together, the data available today show early life stress to consistently be associated with marked changes in HPA axis regulation on several levels (CRH levels, feedback regulation of HPA axis as well as stress reactivity), thereby disrupting an adaptive response to stress – i.e., a rapid activation followed by an efficient termination.

Effects of Genetic Factors

It is important to keep in mind that HPA axis functionality is not only modulated by environmental stimuli, but is also partially determined by genetic factors. This means that naturally occurring variance within those genes involved in HPA axis regulation – e.g., encoding for GC and MRs, enzymes involved in GC synthesis, or regulator hormones like CRH – might be associated with changes in activity of their respective protein products and thus also with changes in HPA axis functionality.

An example of such effects is a single nucleotide polymorphism (SNP) in the GR gene (NR3C1), the Asn363Ser (rs6195, N363S, A to G transition, Codon 363 (Exon 2); the major allele encodes for asparagine, while the minor G-allele encodes for serine). Homozygote carriers of the G-allele have been found to show increased corticosteroid sensitivity as well as an augmented HPA axis reactivity to acute psychosocial stress. Furthermore, carriership of the minor allele has been associated with higher waist-to-hip ratio, higher body-mass index (BMI) as well as increased insulin secretion after dexamethasone intake. These associations reflect changes in metabolic activity that are at least partially under HPA axis control. The critical minor allele appears to increase cortisol reactions to stress, this in turn is thought to cause a bodywide overexposition with cortisol, which itself is associated with several adverse metabolic effects. Another SNP within the GR gene which appears to modulate GR receptor sensitivity is the Bcl-1 (rs41423247, C > G, 647 base pairs downstream of exon 2). Again, the SNP has been shown to be associated with GC hypersensitivity as well as associated metabolic changes. Regarding cortisol reactions to an acute stressor, results are less clear, whereby some data indicate that homozygosity for the minor allele is associated with reduced cortisol reactions.

However, this association appears to be further modulated by gender, with women showing increased cortisol reactions and men reduced levels of cortisol.

The MR is another important protein involved in regulating HPA axis activity. Within its gene (NR3C2) a functional SNP, I180V (rs5522, a missense SNP, isoleucine–valine), has been described, which is associated with increased physiological stress reactivity; i.e., cortisol as well as heart rate and also increased emotional reactivity.

A further well-studied example for the relevance of genetic variance in determining HPA axis activity is an SNP within the FKBP5 gene, which encodes a protein taking part in modulating GR sensitivity as a cochaperone. When FKBP5 binds the GR it attenuates GR affinity for cortisol, which is also accompanied by a reduction in the amount of active GR translocated to the nucleus. Several SNPs within this gene have been identified and studied with respect to their associations with stress sensitivity as well as HPA functionality. The rs1360780 (C/T) has given rise to the most consistent findings. The T-allele is associated with increased FKBP5 expression, which in turn is associated with relative GR resistances. Furthermore, TT-homozygote subjects show a substantially slower endocrine stress-recovery; i.e., downregulation of cortisol after exposure to a psychosocial stress is significantly slower as when compared with C-allele carriers. These results were further corroborated in infants exposed to a stressor, with T-allele carriers showing increased cortisol responses. This SNP might also modulate antidepressant-response in patients, with TT-homozygotes exhibiting more depressive episodes, but also showing a better response to treatment.

In line with the complex interactions with other hormone and transmitter systems described above, genetic variance within genes encoding regulatory proteins of those systems have been shown to be associated with changes in HPA axis functionality as well. Widely studied examples include genetic indicators of serotonergic (e.g., the 5-HTTLPR) or dopaminergic activity (e.g., the COMT val^{158}met polymorphism) or brain-derived neurotropic factor (BDNF); e.g., the BDNF val^{66}met polymorphism. A more detailed summary of genetic markers implemented in regulation of stress reactivity and HPA axis functioning might be found in suitable reviews (see references, e.g., Stephens and Wand, 2012; DeRijk and de Kloet, 2005; Derijk and de Kloet, 2008; Feder et al., 2009).

Importantly, the associations regarding the different genetic variants described above by the way of examples were not replicate in all studies, and further research is needed to understand the exact mechanisms involved in order to explain these discrepancies (e.g., DeRijk and de Kloet, 2005; Luijk et al., 2010). One particularly promising approach to explain the often inconsistent results of molecular–genetic association studies, such as described above, are gene × environment interactions – i.e., persons with different genotypes reacting differently to the same environmental stressor. The following paragraph will give a short summary of chosen aspects regarding gene × environment interactions and their relevance for HPA axis activity.

Gene × Environment Interactions

By now, several studies have shown the high relevance of gene × environment interactions in explaining HPA axis basal activity as well as stress reactivity. Furthermore, there are several examples which show how individual differences in facets of trait personality (e.g., neuroticism or anxiety) which have been shown to be associated with coping abilities as well as affective and neuroendocrine aspects of stress-sensitivity also might be explained in terms of gene × environment interactions.

A prominent and well-studied example within this line of research are genetic predictors of serotonin system activity, due to the high relevance of serotonin in affect regulation and stress reactivity (Ressler and Nemeroff, 2000; Nugent et al., 2011). Regarding stress sensitivity, one of the most widely studied genetic factors is the serotonin transporter-linked polymorphic region (5-HTTLPR), a functional polymorphism in the promoter region of the serotonin reuptake transporter gene. Several studies came to the conclusion that carrying an s-allele of the 5-HTTLPR constitutes a genetic sensitivity factor toward adverse/stressful environmental factors, which reduces one's capacity to adaptively cope with stressors (for review, see Caspi et al., 2010). Moreover, a direct link between this increased sensitivity and the etiology of psychopathology (especially affective disorders) or intermediate phenotypes (e.g., increased reactivity toward negative affective stimuli, amygdala reactivity or dysregulation of stress hormones) has been demonstrated (for review, see Caspi et al., 2010). Accordingly, many authors interpret the s-allele in the sense of a genetic vulnerability factor. These results, however, are somewhat controversial, and the question, whether or not the 5-HTTLPR does constitute a modulator between adverse environmental factors and associated negative outcomes, is still discussed rather critically. Therefore, the question was raised whether the s-allele is associated with an overall increased reactivity to environmental stimuli, regardless of their valence (Belsky and Pluess, 2009). Thus, s-allele carriers would be more vulnerable to the negative effects of stressful LEs, but would also benefit disproportionally from positive environmental influences. This notion is supported by several studies indicating that the s-allele might not be a genetic risk factor, but rather constitutes a genetic plasticity factor associated with increased reactivity to environmental stimuli regardless of their valence, thereby modulating personality facets such as neuroticism, which in turn is associated with stress reactivity.

Several further examples underline how dispositional genetic factors interact with environmental conditions in shaping the activity of the HPA as well as stress responsivity. For example, the previously discussed polymorphism within the FKBP5-gene, rs1360780, does not only show direct associations with cortisol reactivity. Rather, these effects are further dependent on environmental stimuli. In a study on the influence of parent – infant attachment relationship on the regulation of HPA axis activity, it was found that in infants the combination of being a T-allele carrier plus having an insecure attachment relationship with the mother resulted in a doubled risk for increased cortisol relations to stress. Regarding the FKBP5-gene, evidence has been found that several SNPs and haplotypes described modulate the association between childhood trauma and PTSD, depression as well as suicide (Luijk et al., 2010).

Similar gene × environment interactions have been shown for genetic variations of the corticotropin-releasing hormone

receptor (CRH1) gene which interact with early life stress to predict adult depression. Furthermore, the association between childhood maltreatment and HPA axis activity, as measured with the dexamethasone/CRH test, was modulated by SNPs within this gene. The authors concluded that this excessive HPA axis activation might be the underlying mechanism of the interaction of genetic risk factors and the etiology of stress-associated diseases such as mood or anxiety disorders (Tyrka et al., 2009).

A vast amount of literature further underlining the complex interplay between genetic and environmental factors in shaping HPA axis might be found in suitable reviews (see reference list).

In conclusion, it is by now apparent that the HPA axis and its psychological, behavioral and physiological correlates are subject to regulation and modification by a complex and widespread network of different inputs, which are of high relevance for physical as well as psychological health.

See also: Anxiety and Fear Conditioning, Neural Basis of; Emotion, Neural Basis of; Epigenetics of Brain Plasticity and Behavior; Fear: Psychological and Neural Aspects; Gene–Environment Interactions in Health and Well-Being; Hypothalamus; Limbic System; Neuroendocrinology; Stress, Coping and Health.

Bibliography

Appel, K., Schwahn, C., Mahler, J., Schulz, A., Spitzer, C., Fenske, K., Stender, J., Barnow, S., John, U., Teumer, A., Biffar, R., Nauck, M., Volzke, H., Freyberger, H.J., Grabe, H.J., 2011. Moderation of adult depression by a polymorphism in the FKBP5 gene and childhood physical abuse in the general population. Neuropsychopharmacology 36, 1982–1991.

Bangasser, D.A., Valentino, R.J., 2012. Sex differences in molecular and cellular substrates of stress. Cellular and Molecular Neurobiology 32, 709–723.

Belsky, J., Pluess, M., 2009. Beyond diathesis stress: differential susceptibility to environmental influences. Psychological Bulletin 135, 885–908.

Ben-Efraim, Y.J., Wasserman, D., Wasserman, J., Sokolowski, M., 2011. Gene-environment interactions between CRHR1 variants and physical assault in suicide attempts. Genes Brain and Behavior 10, 663–672.

Berardelli, R., Karamouzis, I., D'Angelo, V., Zichi, C., Fussotto, B., Giordano, R., Ghigo, E., Arvat, E., 2013. Role of mineralocorticoid receptors on the hypothalamus-pituitary-adrenal axis in humans. Endocrine 43, 51–58.

Caspi, A., Hariri, A.R., Holmes, A., Uher, R., Moffitt, T.E., 2010. Genetic sensitivity to the environment: the case of the serotonin transporter gene and its implications for studying complex diseases and traits. American Journal of Psychiatry 167, 509–527.

DeRijk, R., de Kloet, E.R., 2005. Corticosteroid receptor genetic polymorphisms and stress responsivity. Endocrine 28, 263–270.

Derijk, R.H., de Kloet, E.R., 2008. Corticosteroid receptor polymorphisms: determinants of vulnerability and resilience. European Journal of Pharmacology 583, 303–311.

Faravelli, C., Lo, S.C., Lelli, L., Pietrini, F., Lazzeretti, L., Godini, L., Benni, L., Fioravanti, G., Talamba, G.A., Castellini, G., Ricca, V., 2012. The role of life events and HPA axis in anxiety disorders: a review. Current Pharmaceutical Design 18, 5663–5674.

Feder, A., Nestler, E.J., Charney, D.S., 2009. Psychobiology and molecular genetics of resilience. Nature Reviews Neuroscience 10, 446–457.

Feodorova, Y.N., Sarafian, V.S., 2012. Psychological stress–cellular and molecular mechanisms. Folia Medica (Plovdiv) 54, 5–13.

Groeneweg, F.L., Karst, H., de Kloet, E.R., Joels, M., 2012. Mineralocorticoid and glucocorticoid receptors at the neuronal membrane, regulators of nongenomic corticosteroid signalling. Molecular and Cellular Endocrinology 350, 299–309.

Heim, C., Nemeroff, C.B., 2001. The role of childhood trauma in the neurobiology of mood and anxiety disorders: preclinical and clinical studies. Biological Psychiatry 49, 1023–1039.

Herman, J.P., Cullinan, W.E., 1997. Neurocircuitry of stress: central control of the hypothalamo-pituitary-adrenocortical axis. Trends in Neuroscience 20, 78–84.

Ising, M., Depping, A.M., Siebertz, A., Lucae, S., Unschuld, P.G., Kloiber, S., Horstmann, S., Uhr, M., Muller-Myhsok, B., Holsboer, F., 2008. Polymorphisms in the FKBP5 gene region modulate recovery from psychosocial stress in healthy controls. European Journal of Neuroscience 28, 389–398.

de Kloet, E.R., Joels, M., Holsboer, F., 2005. Stress and the brain: from adaptation to disease. Nature Reviews Neuroscience 6, 463–475.

Levine, S., 2000. Influence of psychological variables on the activity of the hypothalamic-pituitary-adrenal axis. European Journal of Pharmacology 405, 149–160.

Lightman, S.L., 2008. The neuroendocrinology of stress: a never ending story. Journal of Neuroendocrinology 20, 880–884.

Luijk, M.P.C.M., Velders, F.P., Tharner, A., van Ijzendoorn, M.H., Bakermans-Kranenburg, M.J., Jaddoe, V.W.V., Hofman, A., Verhulst, F.C., Tiemeier, H., 2010. FKBP5 and resistant attachment predict cortisol reactivity in infants: gene–environment interaction. Psychoneuroendocrinology 35, 1454–1461.

McEwen, B.S., 2006. Protective and damaging effects of stress mediators: central role of the brain. Dialogues in Clinical Neuroscience 8, 367–381.

Morris, M.C., Compas, B.E., Garber, J., 2012. Relations among posttraumatic stress disorder, comorbid major depression, and HPA function: a systematic review and meta-analysis. Clinical Psychology Review 32, 301–315.

Nater, U.M., Gaab, J., Rief, W., Ehlert, U., 2006. Recent trends in behavioral medicine. Current Opinion in Psychiatry 19, 180–183.

Nugent, N.R., Tyrka, A.R., Carpenter, L.L., Price, L.H., 2011. Gene–environment interactions: early life stress and risk for depressive and anxiety disorders. Psychopharmacology 214, 175–196.

O'Connor, T.M., O'Halloran, D.J., Shanahan, F., 2000. The stress response and the hypothalamic-pituitary-adrenal axis: from molecule to melancholia. Quarterly Journal of Medicine 93, 323–333.

Reppermund, S., Zihl, J., Lucae, S., Horstmann, S., Kloiber, S., Holsboer, F., Ising, M., 2007. Persistent cognitive impairment in depression: the role of psychopathology and altered hypothalamic-pituitary-adrenocortical (HPA) system regulation. Biological Psychiatry 62, 400–406.

Ressler, K.J., Nemeroff, C.B., 2000. Role of serotonergic and noradrenergic systems in the pathophysiology of depression and anxiety disorders. Depression and Anxiety 12, 2–19.

Stephens, M.A., Wand, G., 2012. Stress and the HPA axis: role of glucocorticoids in alcohol dependence. Alcohol Research 34, 468–483.

Swaab, D.F., Bao, A.M., Lucassen, P.J., 2005. The stress system in the human brain in depression and neurodegeneration. Ageing Research Reviews 4, 141–194.

Tsigos, C., Chrousos, G.P., 2002. Hypothalamic-pituitary-adrenal axis, neuroendocrine factors and stress. Journal of Psychosomatic Research 53, 865–871.

Tyrka, A.R., Price, L.H., Gelernter, J., Schepker, C., Anderson, G.M., Carpenter, L.L., 2009. Interaction of childhood maltreatment with the corticotropin-releasing hormone receptor gene: effects on hypothalamic-pituitary-adrenal axis reactivity. Biological Psychiatry 66, 681–685.

Ventura-Junca, R., Herrera, L.M., 2012. Epigenetic alterations related to early-life stressful events. Acta Neuropsychiatrica 24, 255–265.

Hypothalamus

Phillip Grant, Justus-Liebig-University Giessen, Giessen, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Hennig, volume 10, pp. 7113–7118, © 2001, Elsevier Ltd.

Abstract

Although only a small part of the vertebrate brain, the hypothalamus has a key regulatory function. Being primarily involved in the regulation and maintenance of internal homeostases of several vital bodily functions and neuroendocrine rhythmicity, the hypothalamus adapts bodily functions to internal and external changes as required, thus being a vital component in the body's ability to adapt to internal and external stressors. It is closely linked to the largest intracranial endocrine gland, the pituitary or hypophysis, through which it controls a majority of other endocrine systems.

The hypothalamus is an integral part of the vertebrate brain. Its key function is one of regulating and maintaining internal homeostases as well as neuroendocrine rhythmicity. It does so by directly controlling the vegetative nervous system, major parts of the endocrine system and even basal behavior. The hypothalamus is a part of the diencephalon ('interbrain'), consists mainly of gray matter – the hypothalamic nuclei – and makes up for less than 1% of total brain volume. It is strongly associated with and directly controls the pituitary; an amalgam of various endocrine cells of ectodermal origin (anterior lobe) as well as peptide-hormone releasing hypothalamic axons (posterior lobe).

Ontogeny

Hypothalamus

Around the sixth week of gestation the diencephalic bubble develops four distinct layers. These are, from dorsal to ventral, the epithalamus, the dorsal thalamus (i.e., the thalamus proper), the ventral thalamus (which develops into the pallidum, the reticular nucleus of the thalamus, the subthalamic nucleus and the zona incerta) and the hypothalamus. At this stage of ontogeny the epithalamus has already developed an anlage for the pineal gland, and the hypothalamus shows a frontal anlage for the infundibulum (hypophysial stalk) and an occipital anlage for the mammillary bodies. The infundibular anlage develops into the neurohypophysis (posterior lobe of the pituitary gland) and the median eminence. The relative abundance of embryonic testosterone in relation to maternal and embryonic estrogens leads to early sex-differences in the development of individual hypothalamic nuclei, wherefore the hypothalamus can be considered as a highly sexually dimorphic part of the brain. During the later stages of embryonic ontogeny the fibers of the optic tract begin to form the optic chiasm frontally of the infundibular anlage. The disproportionate growth of the dorsal thalamus leads to a displacement of the diencephalic subsections, whereby the epithalamus is displaced caudally, the ventral thalamus in major parts is displaced laterally and the hypothalamus begins to form a cone-like shape projecting toward the sella turcica at the base of the skull. The hypothalamus itself develops from progenitor cells that migrate outward and first form the lateral and posterior hypothalamic zone, second, the intermediate zone or hypothalamic core and last, the periventricular or midline zone. This fashion of development is phylogenetically highly conserved.

Pituitary Gland

Around the fourth week of gestation ectodermal epithelial cells from Rathke's pouch (an embryonic depression in the roof of the mouth) ascend and contact the infundibular anlage, which by the same time has emerged as a downgrowth from the diencephalic neuroepithelium. By the sixth week, Rathke's pouch has completely disassociated from the oral epithelium; its dorsal aspect develops into the intermediate part of the pituitary, while its ventral aspect proliferates considerably and forms the anterior lobe of the pituitary, aka the adenohypophysis. The neuroepithelial cells of the infundibular anlage induce the development of Rathke's pouch as well as attract it to ascend toward themselves. Afterward, around the third month of gestation, they develop into pituitary-specific glial cells (pituicytes). Around the same time, nonmyelinated axons from the supraoptic and paraventricular nuclei of the hypothalamus (SON and PVN, respectively) grow into the developing glial tissue and form regular varicosities (Herring bodies) for the release of the neurohypophyseal hormones, antidiuretic hormone (ADH) and oxytocin. Immunolabeling-studies show the production of adenohypophyseal hormones starts around the 7th–12th week of gestation, that of the hormones of the posterior pituitary around the 12th–14th week. Interestingly, however, the production of hypothalamic hormones for the control of the anterior pituitary as well as the development of hypophyseal portal vessels to transport said hormones develop several weeks later. Even though the anlages of the hypothalamic–pituitary axis (HP-axis) develop relatively early during ontogeny, its maturation is not completed until way after parturition.

Morphology, Organization and Localization within the Developed Human Brain

The hypothalamus is bordered medially by the third ventricle, dorsally by the mesencephalic tegmentum and apically by the

Table 1 Overview of the major hypothalamic nuclei

	Lateral zone	*Intermediate/medial zone*	*Midline/periventricular zone*
Preoptic region	Lateral preoptic area (LPA)	Medial preoptic nucleus (MPN)	Periventricular preoptic nucleus (PPN)
Anterior region	Supraoptic nucleus (SON)	Paraventricular nucleus (PVN), anterior hypothalamic nucleus (AHN)	Suprachiasmatic nucleus (SCN), paraventricular nucleus (PVN)
Tuberal region	Lateral hypothalamic area (LHA), tuberomammillary nucleus (TMN)	Ventromedial nucleus (VMN), dorsomedial nucleus (DMN), tuberomammillary nucleus (TMN)	Infundibular/arcuate nucleus (IAN)
Mammillary region		Posterior hypothalamic nucleus (PHN), dorsal & ventral premammillary nuclei (DPM; VPM), medial and lateral mammillary nuclei (MMN; LMN), tuberomammillary nucleus (TMN)	

Modified from Nieuwenhuys, R., Voogd, J., Van Huijzen, C., 2008. The Human Central Nervous System, fourth ed. Steinkopff.

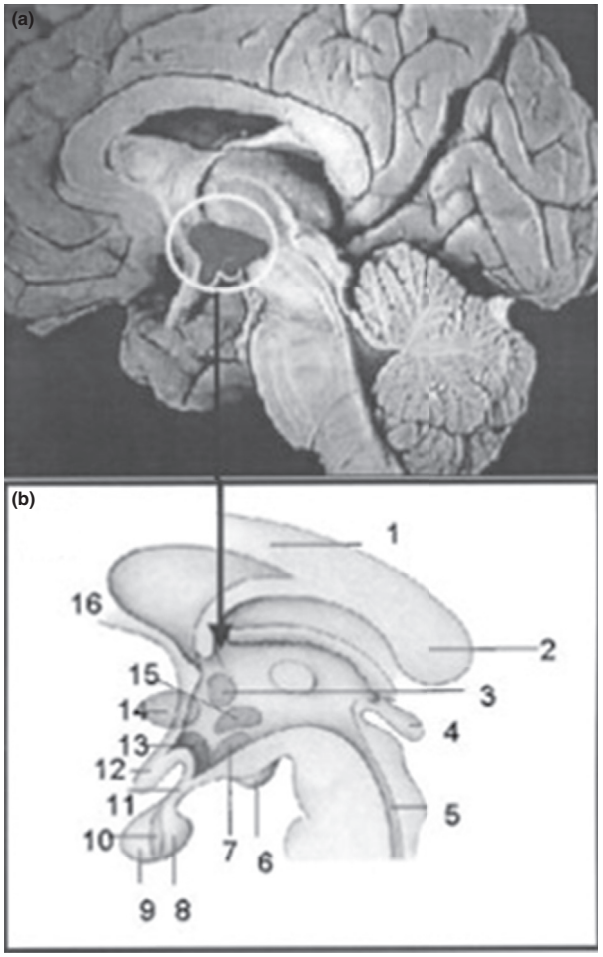

Figure 1 Hypothalamus. Modified from Paulsen, F., Waschke, J. (Eds.), 2010. Sobotta: Atlas der Anatomie des Menschen – Kopf, Hals und Neuroanatomie. Elsevier, Munich.

hypothalamic sulcus. The base of the hypothalamus is the tuber cinereum, which transits laterally into the anterior perforated substance and frontally into the lamina terminalis. The tuber cinereum transitions into the infundibulum toward the sella turcica at the base of the skull, which contains the pituitary. Occipitally, on both sides of the midline, the tuber cinereum borders on the two mammillary bodies.

Apart from the aforementioned distinction between the longitudinally oriented lateral, intermediate and midline zones, the hypothalamus can be divided from rostral to occipital in a coronal fashion into the preoptic, anterior, tuberal and mammillary regions. The distinct hypothalamic nuclei can therefore be grouped within a three-by-four continuum (q.v., Table 1 and Figure 1).

Connections of the Hypothalamus

Pituitary Gland

As described above, the major endocrine output pathway of the hypothalamus is via the pituitary gland, which is connected to the hypothalamus in two distinct fashions. The anterior pituitary, consisting of endocrine cells of ectodermal origin, is connected with and controlled by the hypothalamus via the hypophyseal portal vein. Nuclei of the hypothalamus (mainly the tuberal and infundibular/arcuate nuclei) produce releasing and inhibiting hormones, which are secreted into the portal system and stimulate or inhibit the release of hormones from the anterior pituitary. The posterior pituitary does not, of itself, produce hormones, but consists of axons from the hypothalamic supraoptic and paraventricular nuclei, stabilized by the pituicytes. These axons form regular varicosities (Herring bodies), where the hormones antidiuretic hormone (ADH) and oxytocin are released into the bloodstream. The anterior and posterior lobes of the pituitary are separated by the intermediate part (q.v., Figure 2), and the posterior pituitary continues into the tuberal part, which connects it to the hypothalamus.

Larger magnification of the anterior pituitary reveals the presence of three histologically differently reactive cell-populations. These are the acidophilic, basophilic and chromophobic cells (q.v., Figure 3), whereby the function of the latter is not clear. It is considered that these cells could either be progenitors of either or both of the other cell types, individual cells of either or both of the other two types that have recently released their contents and therefore do not react to histological stainings and antibodies, or even a group of cells producing hormones that are currently unknown and can therefore not be stained with common antibodies.

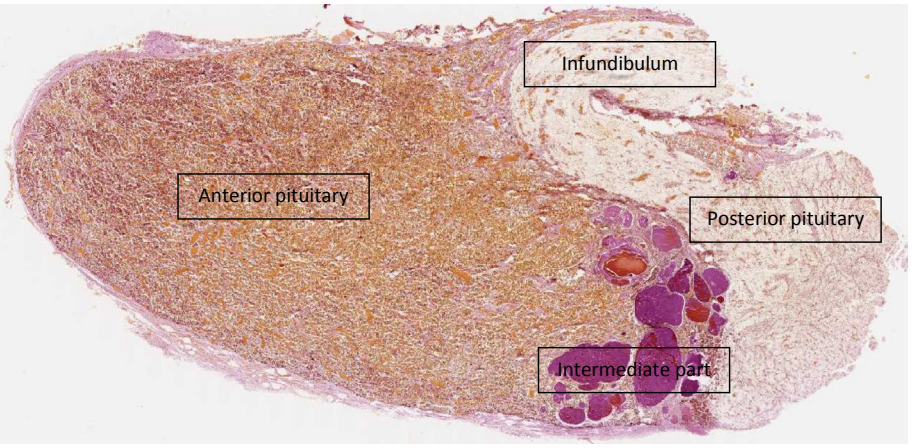

Figure 2 Overview of the human pituitary gland sectioned in the sagittal plane. Image by Grant, taken from histological sections from the Microscopic Anatomy course of the Institute for Anatomy and Cell Biology, University of Giessen.

Acidophilic cells produce the so-called nonglandotropic hormones, which have in common that they stimulate cells that are themselves not endocrine, unlike the glandotropic hormones, which are produced in the basophilic cells and in turn stimulate further endocrine glands; i.e., the thyroid, the adrenal cortex and the gonads (testes or ovaries). The highly feedback-regulated control mechanisms of these latter three endocrine glands are therefore also referred to as the hypothalamic–pituitary–thyroid (HPT), –adrenocortical (HPA) and –gonadal (HPG) axes (q.v., Section Hypothalamic–Pituitary–Thyroid (HPT) Axis). The basophilic cells of the anterior pituitary can therefore be further divided into thyrotropic (producing thyroid-stimulating hormone [TSH], aka thyrotropin), corticotropic (producing pro-opio-melanocortic [POMC], which is posttranslationally processed into adrenocorticotropic hormone [ACTH], aka corticotropin, α-melanotropin [α-MSH] and β-endorphin) and gonadotropic cells (producing the gonadotropins [Gn] luteinizing hormone [LH] and follicle-stimulating hormone [FSH]).

Acidophilic cells can be divided into somatotropic and mammotropic cells, whereby the first produce growth hormone (GH), aka somatotropin, and the latter prolactin. Unlike the basophilic cells, which are stimulated by the hypothalamus but inhibited by the final hormone of their respective HP-axis, acidophilic cells are both stimulated and inhibited by the hypothalamus. Therefore, two hormone classes exist that bind to acidophilic cells, namely releasing and inhibiting hormones (i.e., growth hormone–releasing and –inhibiting hormone (GHRH and GHIH) as well as prolactin–releasing and –inhibiting hormone (PRH and PIH)). GHIH and PIH have been identified to be the same chemical compounds as somatotropin (produced by endocrine cells of the digestive system, especially the δ-cells of the endocrine pancreas) and dopamine (which is usually found to act as a neurotransmitter), respectively. PRH has been shown to be a confusing hormone, since it has become clear that the major factor involved in prolactin release is actually TRH; the exact functions of PRH are therefore unknown.

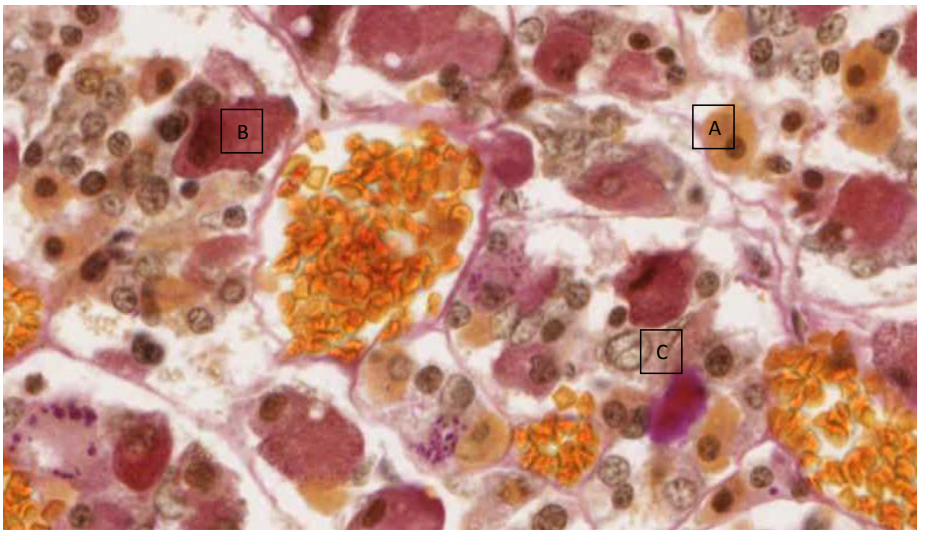

Figure 3 Detailed view of the endocrine cell types of the anterior pituitary (A: acidophilic cell; B: basophil cell; C: chromophobic cell). Image by Grant, taken from histological sections from the Microscopic Anatomy course of the Institute for Anatomy and Cell Biology, University of Giessen.

Table 2 Overview over the major hypothalamic hormones related to the pituitary gland

Hormone	*Main function*
Corticotropin-releasing hormone (CRH)	Stimulation of POMC/ACTH-production
Thyrotropin-releasing hormone (TRH)	Stimulation of TSH- and prolactin-production
Gonadotropin-releasing hormone (GnRH)	Stimulation of LH- and FSH-production
Growth hormone–releasing hormone (GHRH)	Stimulation of GH-production
Growth hormone–inhibiting hormone (GHIH)	Inhibition of GH-production
Prolactin-releasing hormone (PRH)	Unknown
Prolactin-inhibiting hormone (PIH)	Inhibition of prolactin-production
Antidiuretic hormone (ADH)	Increase of water-retention in the kidneys
Oxytocin	Maternal and pair-bonding, prosocial and anxiolytic effects

Table 3 Overview over the hormones of the anterior pituitary

Hormone	*Main function*
Growth hormone (GH)	Stimulation of most bodily cells
Prolactin	Stimulation of lactation, sexual gratification
Corticotropin (ACTH)	Stimulation of adrenal cortex
Thyrotropin (TSH)	Stimulation of thyroid
Luteinizing hormone (LH)	Stimulation of ovaries/testes
Follicle-stimulating hormone (FSH)	Stimulation of ovaries/testes

Regarding the basophilic cells, the hypothalamus only produces releasing hormones, whereby thyrotropin-releasing hormone (TRH) stimulates thyrotropic cells, corticotropin-releasing hormone (CRH) stimulates corticotropic cells and gonadotropin-releasing hormone (GRH) stimulates gonadotropic cells.

Tables 2 and 3 list all hormones of the hypothalamus and the anterior pituitary, including their respective functions.

Hypothalamus–Pituitary–Thyroid (HPT) Axis

The control of the thyroid gland via the HPT-axis begins with the release of hypothalamic TRH, which travels to the anterior pituitary via the hypophyseal portal vein, where TSH is released consequently. This reaches the thyroid via the regular blood stream and here induces the cleavage of the thyroid hormones from the stored colloid. These hormones are triiodothyronine (T_3) and tetraiodothyronine (T_4), aka thyroxine, whereby the latter is produced in greater quantities while the first is the more bioactive hormone. T_4 therefore constitutes a reservoir of T_3 and is deiodinated in the target tissues. Due to the necessity of iodine, which cannot be produced by humans, for the creation of thyroid hormones, the action of the gland is slightly more complex than that of other endocrine glands. The endocrine cells of the thyroid form follicles (wherefore they are mostly referred to as thyroid follicular cells) and produce the peptide thyroglobulin, which is rich in tyrosyl moieties, and release it into the follicular lumen, where it is bound to iodine taken up from food. Thyroglobuline bound to iodine is referred to as colloid. Through this mechanism the body is able to store the otherwise soluble iodine, thereby becoming less dependent on a constant food-intake. Upon stimulation by TSH the stored colloid is then taken back up into the thyroid follicular cell within storage vesicle. These vesicle then fuse with primary lysosomes, whereupon the colloid is cleaved into T_3 and T_4 and peptide residues.

The dependency of the thyroid on iodine intake via nourishment may, in cases of iodine deficiency, lead to compensatory hypertrophy of the gland; a so-called goiter. Due to the lack of iodine, the otherwise healthy thyroid is not able to inhibit the production of TRH and TSH via the HPT-axis, since it lacks iodine to produce it own hormones, which have a negative feedback effect on the HPT-axis. Thus, the gland is constantly overstimulated by the pituitary, wherefore it grows into a pathological size. As thyroid hormones are vital for brain development, iodine deficiency in pregnant women and newborn children leads to severe mental retardation (so-called cretinism). The World Health Organization estimates that a third of the patients suffering from iodine deficiency are of school age, wherefore it considers this condition as the single most preventable cause of mental retardation worldwide. Iodine is naturally found in marine salt, wherefore seafood is an important source of iodine. Additionally, many other food sources, mainly household salt, flour, water and milk, are often added iodine, especially in areas where iodine deficiency is common.

Another common cause of a developing goiter is the autoimmune Graves' disease. Here, autoantibodies against the TSH receptor stimulate said receptor constitutively, wherefore patients suffer from an enlargement of the thyroid but, unlike in iodine deficiency, from hyperthyroidism rather than hypothyroidism (provided, Graves' disease does not occur in patients with additional iodine deficiency). Since TSH receptors, for currently unknown reasons, are also expressed in the retrobulbar orbital fat and the striated muscle of the eye, Graves' disease additionally causes endocrine orbitopathy, the main symptoms of which are a protruding eyeball (exophthalmus) and the inability to focus both eyes on the same point (strabismus).

Apart from idiopathic forms of hypo- and hyperthyroidism, the other common cause of thyroid insufficiency is the T-cell-mediated autoimmune disease Hashimoto's thyroiditis. It is of importance to note that hyper- and hypothyroidism lead to severe affective changes, with hyperthyroidism leading to mania-like symptoms and hypothyroidism being the most important differential diagnosis for depression.

Hypothalamus–Pituitary–Adrenocortical (HPA) Axis

Upon binding of CRH, the corticotropic cells of the anterior pituitary produce POMC, whereof ACTH is an important post-translational product. ACTH stimulates the middle and inner layers of the adrenal cortex (fasciculate and reticular zones) to produce glucocorticoids and androgens, respectively. The former group of hormones, alongside the catecholamines, are important

factors in the body's stress-response, wherefore it major agent in humans (cortisol) is often referred to as a stress-hormone.

Apart from psychological issues like depression and chronic stress, the most common endocrinological disorders involving the HPA-axis are its hyperfunction (Cushing's disease) as well as its hypofunction (Addison's disease). Both conditions can either be caused by over-/underactive adrenocortical cells or by hyper-/hypofunctions of the respective HPA-axis cells in the pituitary (or rarely the hypothalamus).

Hypothalamus–Pituitary–Gonadal (HPG) Axis

Gonadotropic cells of the anterior pituitary are stimulated upon binding of GnRH to produce LH and FSH. While in males it is commonly believed that FSH and LH are released relatively constitutively (although contrary opinions suggest a hormonal cycle in men, somewhat comparable to that of women), in women the plasma levels of both hormones change throughout the menstrual cycle. Acting in synergy, in men both hormones bind to Leydig cells and lead to testosterone production and stimulate spermatogenesis, while in women FSH stimulates follicular growth in the ovaries by binding to granulosa cells (which produce estrogens) and LH stimulates the follicle-surrounding theca cells to produce androgens and hormonal precursors for the production of estrogens. When the ovarian follicle has fully matured (now called a Graafian follicle) it briefly stops estrogen-production, leading to a decrease in negative feedback via the HPT-axis, which in turn leads to a sudden 24–48 h increase in LH-levels. This increase both induces ovulation as well as the conversion of the follicular residue to the corpus luteum. In case of pregnancy, luteal function is supplanted by that of a similar hormone (human chorionic gonadotropin, hCG) produced by the fetal placenta.

Posterior Pituitary

The hormones ADH and oxytocin are produced by the paraventricular and suprooptic nuclei of the hypothalamus, stored in vesicles and transported anterogradely long the axons to the posterior pituitary. Here, they are calcium-dependently released from the Herring bodies into the bloodstream and travel to various organs of the body. The main effect of ADH is the induction of insertion of aquaporin channels into the membranes of the epithelia of the collecting ducts of the kidney. These channels allow for a higher rate of reabsorption of water from the primary urine, wherefore ADH reduces the amount of total urine that reaches the ureters, thus retaining water in the body. In high doses (i.e., when given as a pharmacological agent) or in cases of hypovolemic shock, ADH also increases vasoconstriction, wherefore it is also known under the names of vasopressin or arginine vasopressin (AVP). Additionally, ADH is released directly into the CNS where increasing evidence suggests its role in social, sexual and maternal behavior. It is also believed to have analgesic effects.

Oxytocin is named after its major effects on the uterus during labor (Greek: okys 'fast' and tokos 'birth'), where it assists cervical dilation and uterine contractions. It also causes the injection of milk into the sinuses of the mammary gland (letdown reflex), from where the milk is able to be excreted via the nipple. In both sexes, oxytocin is believed to be vastly important for orgasm, as it acts strongly anxiolytically and induces a feeling of contentment and security with the mate, thereby allowing for orgasm to occur. It is released in high doses during stimulation of the pelvic floor, genital area and areolas. It is therefore believed that maintaining a sufficient pelvic and genital stimulation for oxytocin release in women was one of the major reasons for the extreme evolutionary growth in girth of the male penis in comparison to other primates of similar body size; in consequence of the human species' extreme thickening of the pelvic floor as an adaptation to bipedal gait.

Studies show elevated oxytocin-levels after orgasm, both in cases of self-stimulation as well as coitus, and during/after parturition. It is therefore commonly accepted that oxytocin plays a vital role in pair- and mother/child-bonding. Studies in rodents show that induced genetic dysfunctions of oxytocin-action lead to severe impairment in female offspring-rearing behavior, which can be attenuated by supplication of oxytocin. In humans, oxytocin is also believed to facilitate trust and romantic attraction to potential sexual mates as well as subsequent monogamous pair bonding.

Finally, due to its importance for social emotions and interactions, the oxytocin-system is believed to be causally, at least in part, involved in autism-spectrum disorders.

Afferent and Efferent Neuronal Connections of the Hypothalamus

With the major exception of the pituitary, (dorsal) thalamus and midbrain ventral tegmental area (VTA), which all only receive projections from the hypothalamus, neuronal connections between the hypothalamus and other areas of the brain are mainly reciprocal. Due to its basal function in maintaining of bodily homeostases, the hypothalamus is connected with most parts of the primate brain. Within the telencephalon it communicates with vast areas of the neocortex as well as the septal area/basal forebrain/substantia innominata and the hippocampal formation and extended amygdala (including the bed nucleus of the stria terminalis).

Within the midbrain and pons, the hypothalamus communicates with, inter alia, the periaqueductal gray, monoaminergic centers, nucleus motorius tegmenti, reticular formation and parabrachial nuclei. The myelencephalic regions of interconnectivity with the hypothalamus are vegetative areas of the reticular formation, parasympathetic nuclei of the cranial nerves (via the posterior longitudinal fasciculus), area postrema and the nuclei of the solitary tract.

The hypothalamus is also bidirectionally connected to the layers I and X of the spinal cord, receives projections from the layer V of the spinal cord and the caudal parts of the spinal nucleus of the trigeminal nerve and projects into the intermediolateral column of the spinal cord, which contains preganglionic sympathetic neurons.

Hypothalamic Functions

Regulation of Vegetative Systems and Adaptation to Stress

The hypothalamus directly influences the activity of the preganglionic parasympathetic nuclei of the cranial nerves via

the posterior longitudinal fasciculus. Of these nuclei, the accessory nucleus of the oculomotor (III) nerve, aka Edinger-Westphal nucleus, sends projections to the ciliary ganglion within the orbit. This ganglion controls the activity of all smooth muscles in the eye (i.e., the ciliary muscle and the sphincter and dilatator muscles of the pupil) as well as the superior tarsal muscle within the upper eyelid. The superior and inferior salivatory nuclei project to the pterygopalatine, submandibular and otic ganglia via the intermediate portion of the facial nerve (VII) and the glossopharyngeal nerve (IX). These ganglia control all exocrine glands of the head; i.e., the salivary glands and the lacrimal gland. The last parasympathetic nucleus in the brain is the dorsal nucleus of the vagus nerve (X), which projects onto uncountable ganglia within the walls of all vegetatively innervated organs of the neck, thorax and large parts of the abdomen. The pelvic organs and outer genitalia as well as the last part of the large intestine (i.e., the descending colon, sigmoid and rectum, including the inner anal sphincter) are controlled by the sacral parasympathetic neurons of the spinal cord, which, in turn, are also regulated by the hypothalamus.

Via the aforementioned projections to the intermediolateral column of the spinal cord, the hypothalamus also controls the sympathetic nervous system.

It has to be mentioned, however, that hypothalamic control of most vegetative function is not to be mistaken for an induction of said function. Most bodily systems function autonomously and also have an inherent set of mechanisms for the maintenance of a minimal degree of homeostasis. Notable examples hereof are the expiratory and inspiratory centers within the myelencepahlic reticular formation, the spontaneously firing neurons in the wall of the esophagus, stomach and intestines and last but not least the cardiovascular system, which is, possibly, the most intricately maintained of all peripheral organ systems. Its base activity is controlled through the sinoatrial node, a group of cardiac muscles with the capacity of spontaneous and rhythmic depolarization. The adaptation of the heart and vessels to in–or decreasing cardiac activity is also based on internal reflexes, such as the Frank-Starling reflex of the heart and the Windkessel effect of the large arteries. Finally, the amount of fluid within the cardiovascular system, and thereby largely also the blood pressure, is kept at an optimum through the kidneys, which – through their complex duct system – produce an extensively excess amount of primary urine, whereof, however, averagely up to 98% is reabsorbed specifically depending on the osmotic pressure within the tissue. The role of the hypothalamus is therefore primarily one of integration and adaption to extraordinary circumstances.

An extensively investigated example hereof is the so-called carotid sinus-reflex: The bifurcation of the common carotid artery is richly supplied with baroceptors within a dilated area of the arterial wall (the carotid sinus). The carotid sinus is innervated by a branch of the glossopharyngeal nerve (IX), the sinus nerve of Hering, which relays changes in blood pressure to the nuclei of the solitary tract. This information is then passed on to the hypothalamus, which, in turn, reacts to maintain optimal blood pressure by adapting the dilation of the carotid artery via the dorsal nucleus of the vagus nerve (X). Similar baroceptors also exist in the aortic arch; their information being relayed to the nuclei of the solitary tract via the vagus nerve (X).

Apart from adaptation to extreme changes in internal homeostatic values, the hypothalamus plays a major role in the adaptation to (external) stressors. This is achieved via activation of the sympathetic and inhibition of the parasympathetic nervous system as well as through activation of the HPA-axis. The increases in plasma cortisol and catecholamines thereby work in synergy with each other and the increase in sympathetic activity. While humoral and synaptic catecholamines increase many bodily functions at the expenditure of energy, which in stress-situations is mainly derived from glucose, cortisol increases the amount of glucose available to the body, mainly the striated muscles. Therefore, an important other synergist in the adaptation to stress is the pancreatic hormone insulin, which enables the integration of glucose-transporters into cellular membranes.

Regulation of Water Balance

Although, as mentioned above, the basic control over water balance is performed peripherally by the kidneys, the hypothalamus may regulate kidney function on the one hand and, on the other, is essential for the motivation toward drinking behavior and the development of thirst.

The production and – via the posterior pituitary – release of ADH results from the detection of low blood pressure and high blood osmolarity by the hypothalamus. Herefore, the hypothalamus uses three major mechanisms of detection:

- On the one hand, the ependyma lining the base of the third ventricle, where the hypothalamus is situated, consists partly of specialized ependymal cells, the so-called tanycytes. These cells have long processes that protrude into the hypothalamic tissue and convey signals from the cerebrospinal fluid (CSF) to the CNS. It is believed that this mechanism enables the hypothalamus to monitor not only the osmolarity but also the protein- and metabolic content of the CSF.
- Secondly, the hypothalamus receives projection from the circumventricular organs (CVOs). These regions of the brain are highly vascularized and are characterized by their lack of a tight blood–brain barrier, wherefore they can monitor the contents of the blood stream in a fashion unique within the CNS. The most notable sensory CVOs are the area postrema (AP), the subfornical organ (SFO) and the organum vasculosum of the lamina terminalis (OVLT). The SFO and OVLT, both situated near the third ventricle, relay osmosensory information to the hypothalamus and, due to their modified and less tight blood–brain barrier, are a threshold for peptides and hormones into the CFS.
- The third mechanism of hypothalamic monitoring of the blood is via peripheral sensory (or nonchromaffin) paraganglia. These organs consist of small clusters of chemoceptors and are considered the most richly vascularized organs in the body, wherefore Kummer described their function as, literally, 'tasting the blood.' Two important sensory paraganglia are the carotid and aortic glomera (or bodies), which relay their chemoinformation to the

nuclei of the solitary tract and thereafter the hypothalamus in the same fashion as explained above for the baroceptors of the carotid sinus and the aortic arch.

Apart from the release of ADH for the regulation of water balance, the hypothalamus is involved in a second hormonal system essential for fluid and electrolyte level maintenance, the renin-angiotensin-aldosterone-system (RAAS), whereby is has to be noted that the ADH-system and the RAAS interact with each other. The kidney possesses two monitoring stations for the fluid and electrolyte levels, both within a part of the nephron called the juxtaglomerular apparatus. On the one hand, modified smooth muscle cells of the efferent arterioles in the kidney (the juxtaglomerular cells) detect decreases in blood pressure, and on the other, the macula densa within the distal tubule of the nephron is sensitive to decreases of sodium chloride in the primary urine. Both signals lead to the release of the hormone renin by the juxtaglomerular cells, which converts the freely available prohormone angiotensinogen within the blood plasma to angiotensin I. This is, in turn, converted to angiotensin II by angiotensin-converting enzyme (ACE). Angio II has potent vasoconstrictive effects, leads to the release of the hormone aldosterone from the adrenal cortex (thereby leading to higher sodium and water reabsorption in the kidney) and, finally, induces the release of ADH from the posterior pituitary. Additionally, angio II is detected by the SFO and OVLT and leads the hypothalamus to induce the sensations of thirst and a desire for salt via the median preoptic nucleus.

The area postrema is also involved in water balance. Although primarily responsible for the induction of vomiting behavior when toxins are detected either in the blood or in the intestine, it has also been shown to be sensitive to angiotensin II and, thus, also increases water-seeking behavior through projections to the median preoptic nucleus of the hypothalamus.

The mechanism through which the hypothalamus actually induces the active search for water supplies is believed to be projections to the substantia innominata (SI), possibly the nucleus basalis of Meynert. Through projections from the bed nucleus of the stria terminalis (BST), indirectly from the prefrontal cortex, amygdala and hippocampal formation via the ventral lateral septal nucleus as well as from the supra-chiasmatic nucleus (SCN) it is believed that thirst and induction of water-seeking behavior and drinking can be adapted not only to sleep–wake cycles but also to acute emotional circumstances.

Regulation of Feeding

Due to original studies in rodents from the 1940 and 1950s, it was long believed that the hypothalamus included two discrete areas; one with hyperphagic and the other with hypophagic effects. Recent findings have, however, shown that the major regulation of hunger and food intake is actually performed by the same hypothalamic area, namely the arcuate nucleus. These results are based on the discovery of the gene encoding for the adipocyte-derived hormone leptin. Knockout of the rodent ortholog of the LEP/OB gene (LEP for leptin or OB for obesity) leads to an incapability of the animals to experience satiety, thus leading to obesity. Hence, leptin produced by adipocytes binds to leptin-receptors in the hypothalamic arcuate nucleus

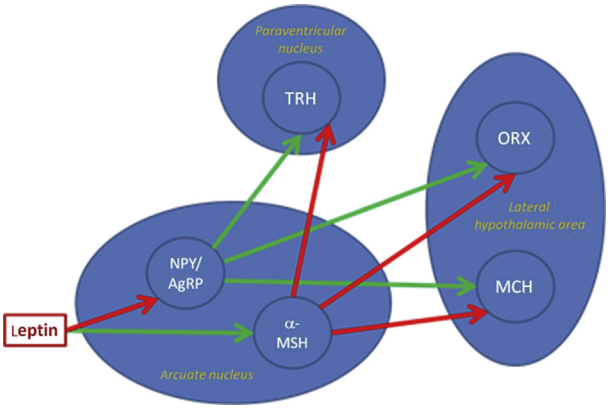

Figure 4 Leptin-action within the hypothalamus (a-MSH: alpha-melanocyte stimulating hormone; AgRP: agouti-related protein; MCH: melanin-concentrating hormone; NPY: neuropeptide Y; ORX: orexin; TRH: thyrotropin (TSH) releasing hormone. Red arrows indicate inhibition and green arrows indicate stimulation).

and here induces satiety. The mechanism of leptin-action is complex, as it acts simultaneously on two populations of neurons within the arcuate nucleus. Of these, neuropeptide Y (NPY) and Agouti-related protein (AgRP) expressing cells are inhibited through leptin, whereas POMC/α-MSH expressing cells are activated. NPY and AgRP are considered orexigenic (i.e., appetite-stimulating) and α-MSH anorexigenic (i.e., appetite-inhibiting). The effects of both peptide-groups is mediated through the lateral hypothalamic area, which produces the peptides orexin (ORX) and melanin-concentrating hormone (MCH), as well as the paraventricular nucleus through the induction of TRH-production. TRH, MCH and ORX activate the metabolism and induce hunger and food-seeking behavior. The production of TRH, MCH and ORX is inhibited by α-MSH and activated by NPY and AgRP. Thus, leptin binding to the neurons of the arcuate nucleus activates α-MSH-producing cells, which, in turn, inhibit TRH, MCH and ORX-production, as well as inhibits NPY/AgRP-producing cells, which otherwise would activate TRH, MCH and ORX-production. Through both pathways, therefore, increases in leptin-concentration lead to reduction of hunger and food-seeking behavior (q.v., Figure 4).

Regulation of Ultradian, Circadian and Infradian Rhythms

The aforementioned hypothalamic neuropeptide orexin is, although shown to be involved in appetite and food-seeking behavior, primarily responsible for the regulation of wakefulness and arousal. The clinical picture of narcolepsy, where patients experience excessive daytime sleepiness and often sudden onset naps that are often not controllable, has repeatedly been linked to dysfunctions in the orexin-system. In humans, it is believed to be caused by autoimmune processes that destroy orexin-producing cells with mutations in the human leukocyte antigen (HLA; aka major histocompatibility complex, MHC) appearing to be the most common biological foundation. In animals, genetic knockout of the genes encoding for orexin or its receptors causes the clinical picture found in human narcolepsy patients.

Research in mice has also shown that hypothalamic neurons of the tuberomammillary nuclei producing histamine show distinct firing patterns during different sleep cycles and transition-phases between sleep and waking. The projections of the hypothalamus to the substantia innominata and nucleus basalis of Meynert are also considered relevant in the regulation of arousal and attention. They comprise a branch of the ascending reticular arousal system, a network of cholinergic projections ascending from the brainstem through the hypothalamus and allothalamus to the striatum and, via the nucleus basalis of Meynert, to the cerebral cortex. Finally, through the HPA-axis, the hypothalamus regulates the levels of the glucocorticoid-production by the adrenal cortex. In humans, the most important of these is the steroid hormone cortisol and is also involved in the induction and upkeep of alertness and, by extension, wakefulness. A common hypothesis, regarding the postwaking sudden increase in plasma cortisol (the cortisol-awakening-reaction, CAR), is that it increases the alertness of a recently awakened individual and, thus, aids in preparation for the possible dangers and increased requirements of waking hours.

The regulation of temporary cycles, which can be approximately the same length as 1 day (circadian), shorter than a day (ultradian) and considerably longer (infradian, even circa-annual) are largely dependent on light. The hypothalamus is therefore strongly connected with specialized ganglion cells in the retina, that project information regarding ambient illumination directly to the hypothalamus via the retinohypothalamic tract. These ganglion cells are receptive for light of blue wavelength, thus constituting primary photoceptors, but do not relay visual information, merely the information how much light acutely falls on the retina. The projections are relayed in the hypothalamic suprachiasmatic nucleus (SCN) and, from there, project to the intermediolateral column of the cervical spinal cord. Via the superior cervical ganglion, which is the most cranial of the paravertebral sympathetic ganglia, the cervical sympathetic fibers activate the pineal gland of the epithalamus. The SCN thus indirectly has control of the conversion of serotonin to melatonin within the pineal gland by influencing the activity of the enzyme serotonin N-acetyltransferase; in a fashion that melatonin production is induced in the dark. Therefore, melatonin levels not only change over the course of a day but also depending on seasonal changes and even an individual's position related to the earth's circles of latitude (aka parallels). Thus, affective disorders related to the melatonin production (seasonal affective disorders, SAD; aka winter depressions) are more common in Scandinavian countries and can be treated by bright light therapy. This is strengthened by findings regarding both native and Canadian Icelandic people who, although being or descending from within Scandinavia, show surprisingly low levels of SAD; a fact which is believed to be explainable through their high nutritional intake of Vitamin D through fish. Vitamin D synthesis is alternatively also induced by exposure to sunlight.

Many processes regulated through the hypothalamus, not only sleep–wake cycles, follow rhythmic patterns. The example of the circadian rhythm of the HPA-axis has also been given. Another important rhythmic system is the infradian regulation of the estrous cycle in females through the HPG-axis. It can therefore be concluded, that the hypothalamus not only maintains internal bodily homeostases, but also aligns these homeostases to external stimuli dependent on regular changes in terrain conditions.

Regulation of Sexual and Reproductive Behavior

As mentioned previously, the brain in general and the hypothalamus in particular are highly sexually dimorphic regions of the mammalian body. Apart from the differences in HPG-axis-regulation and thereby the control over the female estrous cycle, the hypothalamus has also been described above as being involved in parturition and lactation through the hormones oxytocin and prolactin. The latter is a highly versatile signaling agent, but has important function during the induction of lactation but also in postorgasmic phases, where it is believed responsible for the feeling of sexual gratification.

The most commonly examined hypothalamic nucleus regarding sexual behavior and diversity is the sexually dimorphic nucleus (SDN); the human homologue thereof is the third interstitial nucleus of the anterior hypothalamus (INAH3). Together with other regions of the di- and telencephalon (e.g., the amgydala, bed nucleus of the stria terminalis, lateral septal nucleus and the nucleus of the diagonal band of Broca), including the hypothalamic ventromedial, infundibular, ventral premammillary, suprachiasmatic and medial preoptic nuclei, this system is referred to by Nieuwenhuys et al. (2008) as the telodiencephalic reproductive complex (TDRC). The connection with the amygdala, a primary telencephalic center for the interpretation and reaction-decision regarding conspecifics and other socially relevant stimuli, links sexually reproductive behavior with other socially relevant behaviors, such as defense and aggression.

Although this aspect is to extensive to cover in detail, it has been shown in rodents that the areas of the TDRC are differently sized between sexes and that injection of different sexually relevant agents (e.g., aromatase(-inhibitors) or testosterone) into nuclei of the TDRC at an early age leads to significant changes in sexual behavior, such as females mounting other sexually receptive females or males showing female-typical lordosis behavior in the presence of other males. The size of TDRC-areas, especially the INAH3, is also intensively discussed as a possible correlate of human sexual orientation. This notion goes back mainly to a paper published by LeVay (1991), although it implications are often overrated; as also repeatedly maintained by LeVay himself.

Studies using positron emission tomography (Savic et al., 2005; Berglund et al., 2006) also showed that arousal of TDRC-regions to testosterone-derivatives and estrogen-like compounds is consistent with sexual orientation, whereby homosexual men and heterosexual women as well as homosexual women and heterosexual men share arousal to either male or female scent, respectively. It should, however, be mentioned that the accessory olfactory system, which in mammals is responsible for the perception and transduction of pheromonal information, is highly vestigial in humans and the question of its being functional is highly debated.

Regulation of Body Temperature and the Immune System

Through the HPA-axis as well as the projections from the spinal interomediolateral column to the adrenal medulla, the

hypothalamus has considerable influence over the immune response. Important in this context is the notion that immunocytes are not equally influenced by cortisol and catecholamines. Thus, antiinflammatory immune responses are not directly under the influence of the hypothalamus, while proinflammatory immunocytes have several receptors for the aforementioned hormones. Since the pro- and antiinflammatory immunocytes do, however, influence each other, activation of the HPA-axis and the adrenal medulla lead to a shift of the immune response toward the antiinflammatory branch. This has been shown, additionally, to be different for acute verses chronic stress, whereby the first leads to an overall increase in immune response, while the latter will lead to immunosuppression. This has been shown for several factors such as immunocyte concentration and activity, autoimmune regulation, wound healing and susceptibility to clinically induced infection. Importantly, in neoplastic disorders the influence of HPA-axis and adrenal medullary hormones have repeatedly been shown to have severely dire consequences on remission rates and patients' survival. It is therefore essential to note, that a reduction of hypothalamically induced stress-response (e.g., through accompanying psychooncological therapy) has considerable beneficial consequences in cancer patients, whereby the actual form of coping-behavior and therapy is not of relevance, as patients have been shown to 'instinctively choose' the form of coping/therapy that is best suited for them individually.

Closely linked, but not only related to immune response, is the regulation of body temperature by the hypothalamus. In case of bacterial (and a few strains of viral) infection, the hypothalamus raises the index value for body temperature, thereby inducing fever. Apart from its participation in the active immune response, this also influences the motivational and appetitive reactions and behavior in mammals. Taken together, the physiological and behavioral changes during active immune response are called sickness behavior and are interpreted as an attempt to conserve energy and reduce risk of additional dangers (e.g., through other predators). It has been shown, for example, that sickness behavior reduces the motivation toward food-seeking behavior as well as the appetitive component of hunger but not the actual hunger or desire for food intake, as patients given food (by persons of trust) will eat this, even though they will show a reduction in activities necessary for the production of foods.

See also: Autism Spectrum Disorders; Behavioral Medicine; Cancer: Psychosocial Aspects; Cardiovascular Reactivity and Recovery; Chronotype; Clinical Psychology, Psychiatry and Homosexuality; Endocrinology and Psychiatry; Hunger and Eating, Neural Basis of; Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of; Interoception; Neuroendocrinology; Psychoneuroimmunology; Sexual Orientation: Biological Influences; Sleep: Neural Systems; Stress and Cardiac Response; Stress in Adolescence: Effects on Development; Stress in Childhood: Effects on Development; Stress in Infancy and Early Childhood: Effects on Development; Stress, Coping and Health; Stress: Effects on Neonatal and Child Development; Stress: Psychological Perspectives.

Bibliography

Berglund, H., Lindstrom, P., Savic, I., 2006. Brain responses to putative pheromones in lesbian women. Proceedings of the National Academy of Sciences of the United States of America 103 (21), 8269–8274.

Drenckhahn, D. (Ed.), 2003. Benninghoff: Anatomie, sixteenth ed. Urban & Fischer, München Jena.

LeVay, S., 1991. A difference in hypothalamic structure between heterosexual and homosexual men. Science 253, 1034–1037.

Nieuwenhuys, R., Voogd, J., Van Huijzen, C., 2008. The Human Central Nervous System, fourth ed. Steinkopff.

Paulsen, F., Waschke, J. (Eds.), 2010. Sobotta: Atlas der Anatomie des Menschen – Kopf, Hals und Neuroanatomie. Elsevier, Munich.

Savic, I., Berglund, H., Lindstrom, P., 2005. Brain responses to putative pheromones in homosexual men. Proceedings of the National Academy of Sciences of the United States of America 102 (20), 7356–7361.

Standring, S., 2008. Gray's Anatomy: The Anatomical Basis of Clinical Practice, fortieth ed. Churchill Livingstone Elsevier.

Hypothesis Testing in Statistics

Roger L Berger, School of Mathematical and Natural Sciences, Arizona State University, Phoenix, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Casella, R.L. Berger, volume 10, pp. 7118–7121, © 2001, Elsevier Ltd.

Abstract

A statistical hypothesis is a statement about one or more population parameter(s). A hypothesis test is a rule for deciding which of two complementary hypotheses is true, based on some data. Various methods for constructing hypothesis tests (e.g., likelihood, Bayesian, intersection–union) are available. Various measures for evaluating hypothesis tests, most of which are related to the tests' probabilities of making correct or incorrect decisions, are available, also. Construction and evaluation of tests, if a large sample is available, is sometimes simplified by asymptotic results.

Introduction

A *statistical hypothesis* is a statement about one or more population parameter(s), and the two complementary hypotheses in a hypothesis-testing problem are called the *null hypothesis* and the *alternative hypothesis*. They are denoted by H_0 and H_1, respectively.

If θ denotes a population parameter, the general format of the null and alternative hypotheses is $H_0: \theta \in \Theta_0$ and $H_1: \theta \in \Theta_0^c$, where Θ_0 is some subset of the parameter space and Θ_0^c is its complement. A hypothesis test is a rule for deciding whether H_0 or H_1 is true, based on some data. Typically, a hypothesis test is specified in terms of a *test statistic*, $W(X_1, \ldots, X_n) = W(\boldsymbol{X})$, a function of the data, $\boldsymbol{X}$. The sample space for W is partitioned into the *rejection region*, R, and its complement, the *acceptance region*. If $W \in R$ is observed, the null hypothesis H_0 is rejected, and the decision is made that H_1 is true. If $W \notin R$, H_0 is accepted as true.

For example, in a study to examine the factors that influence a student's success at completing a 4-year college, one factor of interest might be socioeconomic status (SES). If, for simplicity, we just have two groups (high SES and low SES), and the success probabilities of the respective groups are p_h and p_l, then a hypothesis test of interest may be $H_0: p_h \le p_l$ versus $H_1: p_h > p_l$.

This is an example of a one-sided test, in which the alternative hypothesis specifies a direction. In contrast, we may consider the two-sided test, $H_0: p_h = p_l$ versus $H_1: p_h \ne p_l$, in which no direction is specified in the alternative. If we let $\widehat{p}_h$ and $\widehat{p}_l$ denote the observed success rates in the two SES classes, the one-sided test would reject the null hypothesis if $\widehat{p}_h - \widehat{p}_l$ is big, while the two-sided test would reject the null hypothesis if $\widehat{p}_h - \widehat{p}_l$ is either big or small.

A hypothesis test of H_0: $\theta \in \Theta_0$ versus H_1: $\theta \in \Theta_0^c$ might make one of two types of errors. If $\theta \in \Theta_0$, but the hypothesis test incorrectly decides to reject H_0, then the test has made a *Type I error*. If, on the other hand, $\theta \in \Theta_0^c$, but the test decides to accept H_0, a *Type II error* has been made. In our example of testing H_0: $p_h \le p_l$ versus H_1: $p_h > p_l$, if we conclude that $p_h > p_l$, but in fact $p_h \le p_l$, we have made a Type I error.

Constructing Tests

There are many methods of deriving test statistics for a hypothesis test, a few of which follow.

Likelihood Ratio Tests

The likelihood ratio method of hypothesis testing is related to maximum likelihood estimators (MLE) (discussed in *Estimation: Point and Interval*). Suppose that we have a sample, $X_1, \ldots, X_n$, where the X_i are independent, each with distribution given by the probability density or mass function, $f(x|\theta)$. The observed values of these random variables are denoted by $\boldsymbol{x} = (x_1, \ldots, x_n)$, with likelihood function $L(\theta|\boldsymbol{x})$ given by

$$L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\theta).$$

The *likelihood ratio test (LRT) statistic* for testing H_0: $\theta \in \Theta_0$ versus H_1: $\theta \in \Theta_0^c$ is

$$\lambda(\boldsymbol{x}) = \frac{\sup\limits_{\Theta_0} L(\theta|\boldsymbol{x})}{\sup\limits_{\Theta} L(\theta|\boldsymbol{x})}$$

An LRT is a test that has a rejection region of the form $\{\boldsymbol{x}: \lambda(\boldsymbol{x}) \le k\}$, where k is any number satisfying $0 \le k \le 1$.

If we interpret the likelihood function as measuring how likely the values of θ are, then we see that the LRT statistic is comparing the plausibility of the θ values in the null hypothesis with those in the alternative. Small values of the LRT statistic are then interpreted as being evidence against H_0 and lead to rejection of H_0.

If the null hypothesis consists of a single value θ_0, and the alternative is everything else, then the LRT statistic is simply $\lambda(\boldsymbol{x}) = \frac{L(\theta_0|x)}{L(\widehat{\theta}|x)}$, where $\widehat{\theta}$ is the MLE of θ.

Example. Let $X_1, \ldots, X_n$ be a random sample from a normal population with mean θ and variance 1 ($n(\theta, 1)$ population). The LRT statistic for testing H_0: $\theta = \theta_0$ versus H_1: $\theta \ne \theta_0$ is

$$\lambda(\boldsymbol{x}) = \frac{L(\theta_0|\boldsymbol{x})}{L(\overline{x}|\boldsymbol{x})} = \frac{(2\pi)^{-\frac{n}{2}} \exp\left[-\sum_{i=1}^{n} \frac{(x_i-\theta_0)^2}{2}\right]}{(2\pi)^{-\frac{n}{2}} \exp\left[-\sum_{i=1}^{n} \frac{(x_i-\overline{x})^2}{2}\right]},$$

where $\overline{x}$ is the sample mean calculated from $\boldsymbol{x}$.

If $T(\boldsymbol{X})$ is a sufficient statistic for θ, then, as with MLE, the LRT statistic is a function of T. That is, $\lambda(\boldsymbol{x})$ depends on $\boldsymbol{x}$ only

through $T(\boldsymbol{x})$. In the previous example $T(\boldsymbol{X}) = \overline{X}$ is a sufficient statistic, and the expression for $\lambda(\boldsymbol{x})$ can be simplified to

$$\lambda(\boldsymbol{x}) = \exp\left[-\frac{n(\overline{x}-\theta_0)^2}{2}\right].$$

Bayesian Tests

The Bayesian paradigm prescribes that the sample information be combined with the prior information using Bayes' theorem to obtain the posterior distribution, $\pi(\theta|\boldsymbol{x})$. All inferences about θ are now based on the posterior distribution. In a hypothesis-testing problem, the posterior distribution may be used to calculate the probabilities that H_0 and H_1 are true.

One way a Bayesian hypothesis tester may choose to use the posterior distribution is to decide to accept H_0 as true if

$$\frac{P(\theta \in \Theta_0|\boldsymbol{X})}{P(\theta \in \Theta_0^c|\boldsymbol{X})} > k,$$

for some constant k, and to reject H_0 otherwise. Equivalently, she can reject H_0 if $P(\theta \in \Theta_0^c|\boldsymbol{X}) \geq 1/(1+k)$.

Example. Let $X_1, \ldots, X_n$ be a random sample from an $n(\theta, \sigma^2)$ population, and let the prior distribution on θ be $n(\mu, \tau^2)$, where σ^2, μ, and τ^2 are known. Consider testing $H_0{:}\theta \leq \theta_0$ versus H_1: $\theta > \theta_0$, and suppose we decide to accept H_0 if $P(\theta \in \Theta_0|\boldsymbol{X}) > P(\theta \in \Theta_0^c|\boldsymbol{X})$. After some calculation, we find that H_0 will be accepted as true if

$$\overline{X} < \theta_0 + \frac{\sigma^2(\theta_0-\mu)}{n\tau^2}.$$

Union–Intersection and Intersection–Union Tests

In some situations, tests for complicated null hypotheses can be developed from tests for simpler null hypotheses. The union–intersection method of test construction might be useful when the null hypothesis is conveniently expressed as an intersection, say $H_0{:}\theta \in \bigcap_{\gamma\in\Gamma}\Theta_\gamma$, where Γ is an arbitrary index set. If tests are available for each of the problems of testing $H_{0\gamma}{:}\theta \in \Theta_\gamma$ versus $H_{1\gamma}{:}\theta \in \Theta_\gamma^c$, where the rejection region for the test of $H_{0\gamma}$ is $\{\boldsymbol{x}{:}\ W_\gamma(\boldsymbol{x}) \in R_\gamma\}$, then the rejection region for the union–intersection test is

$$\bigcup_{\gamma\in\Gamma}\{x{:}W_\gamma(x) \in R_\gamma\}.$$

The rationale is that if any one of the hypotheses $H_{0\gamma}$ is rejected, then H_0 must be rejected, also.

A complementary method, the intersection–union method, may be useful if the null hypothesis is conveniently expressed as a union. Suppose we wish to test the null hypothesis $H_0{:}\theta \in \bigcup_{\gamma\in\Gamma}\Theta_\gamma$, and $\{\boldsymbol{x}{:}W_\gamma(\boldsymbol{x}) \in R_\gamma\}$ is the rejection region for a test of $H_{0\gamma}{:}\theta \in \Theta_\gamma$ versus $H_{1\gamma}{:}\theta \in \Theta_\gamma^c$. Then the rejection region for the intersection–union test of H_0 versus H_1 is

$$\bigcap_{\gamma\in\Gamma}\{\boldsymbol{x}{:}W_\gamma(\boldsymbol{x}) \in R_\gamma\},$$

and H_0 is false if and only if all of the $H_{0\gamma}$ are false, so H_0 can be rejected if and only if each of the individual hypotheses $H_{0\gamma}$ can be rejected.

Example. Returning to our example about success in 4-year colleges, we might argue that there are measures of success other than finishing. For example, yearly income could be used to measure success, and denoting mean incomes in the high and low SES groups θ_h and θ_l, respectively, results in the hypothesis test

$$H_0{:}\theta_h \leq \theta_l \text{ or } p_h \leq p_l \quad \text{versus} \quad H_1{:}\theta_h > \theta_l \text{ and } p_h > p_l,$$

where the high group is considered superior only if H_1 is accepted.

If we have estimators $\widehat{\theta}_h$, $\widehat{\theta}_l$, $\widehat{p}_h$, and $\widehat{p}_l$, a rejection region for the intersection–union test could be given by

$$\left\{\left(\widehat{\theta}_h, \widehat{\theta}_l, \widehat{p}_h, \widehat{p}_l\right){:}\widehat{\theta}_h - \widehat{\theta}_l > k_1 \text{ and } \widehat{p}_h - \widehat{p}_l > k_2\right\}.$$

Thus the intersection–union test decides that the high SES group is superior to the low SES group, that is, H_1 is true, if and only if it decides that each of the individual parameters is better in the high group.

Intersection–union tests provide important tools in acceptance sampling, in which one decides whether to accept a product based on a collection of measurements (see Berger, 1982).

There are many other methods available for constructing hypothesis tests, methods based on invariance, pivots, robust, or large-sample arguments, to name a few. For more on hypothesis testing see Lehmann (1986).

Evaluating Tests

As mentioned previously, a hypothesis test might make one of two different kinds of errors. The probabilities of making these errors can be calculated using the *power function*. If R denotes the rejection region for a test, the *power function* is $\beta(\theta) = P_\theta(\boldsymbol{X} \in R)$. If $\theta \in \Theta_0$, $\beta(\theta)$ is the probability of a Type I error. If $\theta \in \Theta_0^c$, $\beta(\theta)$ is one minus the probability of a Type II error. Hence, a good test has power function near one for most $\theta \in \Theta_0^c$ and near zero for most $\theta \in \Theta_0$.

Example. Let $X_1, \ldots, X_n$ be a random sample from an $n(\theta, \sigma^2)$ population, σ^2 known. The LRT of $H_0{:}\theta \leq \theta_0$ versus $H_1{:}\theta > \theta_0$ rejects H_0 if $\sqrt{n}(\overline{X}-\theta_0)/\sigma \geq k$ and has power function

$$\beta(\theta) = P\left(Z \geq k + \frac{\theta_0-\theta}{\sigma/\sqrt{n}}\right),$$

where Z is a standard normal random variable.

After a hypothesis test is done, the conclusions must be reported in some statistically meaningful way. One method of reporting the results of a hypothesis test is to report the *size* $\left(\alpha = \sup_{\theta\in\Theta_0} P_\theta(\boldsymbol{X} \in R)\right)$ of the test and the decision to reject H_0 or accept H_0. The size of the test carries important information. If α is small, the decision to reject H_0 is fairly convincing, but if α is large, the decision to reject H_0 is not very convincing, because the test has a large probability of incorrectly making that decision. A common practice is to choose the desired small value of α (often 0.05 or 0.01) and then to choose the critical value, k in the previous example, so the resulting test has the specified size, α.

Another way of reporting the results of a hypothesis test, one that is data dependent, is to report the *p-value*. Typically, not one but an entire class of tests is constructed, a different test being defined for each value of α. The p-value for the sample point $\boldsymbol{x}$ is the smallest value of α for which this sample point will lead to rejection of H_0 (*see* Significance, Tests of). Because rejection of

H_0 using a test with small size is more convincing evidence that H_1 is true than rejection of H_0 with a test with large size, the interpretation of p-values goes in the same way. The smaller the p-value, the stronger the sample evidence that H_1 is true.

In the previous example, if data $\boldsymbol{x}$ with sample mean $\bar{x}$ is observed, the p-value is

$$p(\boldsymbol{x}) = P\left(Z \geq \frac{\bar{x} - \theta_0}{\sigma/\sqrt{n}}\right).$$

Many other types of evaluations of tests can be done. The theory of most powerful tests shows how to construct best tests under a variety of conditions (see Lehmann, 1986 or Casella and Berger, 2001: Chapter 8). Hypothesis tests can also be evaluated using risk functions as in Hwang et al. (1992).

Asymptotics

For the LRT statistic, the following general theorem allows the construction of a large-sample test.

Theorem 1. *Let $X_1, \ldots, X_n$ be a random sample from a probability density function or probability mass function $f(\boldsymbol{x}|\theta)$. Under some regularity conditions on $f(\boldsymbol{x}|\theta)$, if $\theta \in \Theta_0$ then the distribution of the statistic $-2\log\lambda(\boldsymbol{x})$ converges to a chi-squared distribution as the sample size $n \to \infty$. The degrees of freedom of the limiting chi-squared distribution is the difference between the number of free parameters specified by $\theta \in \Theta_0$ and the number of free parameters specified by $\theta \in \Theta$.*

The 'regularity conditions' are concerned mainly with the existence and behavior of the derivatives (with respect to the parameter) of the likelihood function, and the support of the distribution (it cannot depend on the parameter). See Casella and Berger (2001: Section 10.6) or Lehmann (1986: Section 8.8) for precise conditions.

Rejection of H_0 for small values of $\lambda(\boldsymbol{x})$ is equivalent to rejection for large values of $-2\log\lambda(\boldsymbol{x})$. Thus, by Theorem 1 the test that rejects H_0 if and only if $-2\log\lambda(\boldsymbol{x}) \geq \chi^2_{\nu,\alpha}$, where ν is the degrees of freedom specified in the theorem, is a test with size approximately equal to α, if the sample size is large.

As might be expected, Theorem 1 has wide applicability. In particular, it is extremely useful in categorical data analysis (*see* Multivariate Analysis: Discrete Variables (Loglinear Models)).

Other large-sample test constructions are based on asymptotic normality of a point estimator (*see* Estimation: Point and Interval). Suppose we wish to test a hypothesis about a real-valued parameter θ, and $W_n = W(X_1, \ldots, X_n)$ is a point estimator of θ based on a sample of size n that satisfies

$$\frac{W_n - \theta}{\sqrt{\mathrm{Var}W_n}} \to Z$$

as $n \to \infty$, where $\mathrm{Var}W_n$ is the variance of W_n and Z is a standard normal random variable. We now have the basis for an approximate test, for example, we would reject $H_0{:}\theta \leq \theta_0$ at level 0.05 if $(W_n - \theta_0)/\sqrt{\mathrm{Var}\,W_n} \geq 1.645$. Note, $\mathrm{Var}\,W_n$ could depend on θ_0, and we can still use it in the test statistic. This type of test, where we use the actual variance of W_n, is called a *score test*.

If $\mathrm{Var}\,W_n$ also depends on unknown parameters, we could look for an estimate S_n^2 of $\mathrm{Var}\,W_n$ with the property that $(\mathrm{Var}\,W_n)/S_n^2$ converges in probability to one. Then, using Slutsky's theorem (see Casella and Berger, 2001: Section 5.5), we can deduce that $(W_n - \theta)/S_n$ converges in distribution to a standard normal distribution, also. The large-sample test based on this statistic is called a *Wald test*.

More details about asymptotic tests may be found in Boos and Stefanski (2013) and Lehmann (1999).

Conclusions

Hypothesis testing is one of the most widely used, and some may say abused, methodologies in statistics. Formally, the hypotheses are specified, an α-level is chosen, a test statistic is calculated, and it is reported whether H_0 or H_1 is accepted. In practice, it may happen that hypotheses are suggested by the data, the choice of α-level may be ignored, more than one test statistic is calculated, and many modifications to the formal procedure may be made. Most of these modifications cause bias and can invalidate the method. For example, a hypothesis suggested by the data is likely to be one that has 'stood out' for some reason, and, hence, H_1 is likely to be accepted, unless the bias is corrected for (using something like Scheffe's method – see Hsu, 1996).

Perhaps the most serious criticism of hypothesis testing is the fact that, formally, it can only be reported that either H_0 or H_1 is accepted at the prechosen α-level. Thus, the same conclusion is reached if the test statistic only barely rejects H_0 and if it rejects H_0 resoundingly. Many feel that this is important information that should be reported, and thus it is commonplace to report the p-value of the hypothesis test, also.

For further details on hypothesis testing see the classic book by Lehmann (1986). Introductions are provided by Casella and Berger (2001) or Schervish (1995), also, and a good introduction to multiple comparisons is Hsu (1996); *see* Hypothesis Tests, Multiplicity of.

Acknowledgment

This article is a revised version of an article on the same topic in the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, which was coauthored with the late George Casella, University of Florida.

See also: Bayesian Statistics; Hypothesis Testing: Methodology and Limitations; Hypothesis Tests, Multiplicity of; Likelihood in Statistics; Linear Hypothesis; Model Testing and Selection, Theory of; Multiple Comparisons, Statistics of; Significance, Tests of; Statistical Sufficiency.

Bibliography

Berger, R.L., 1982. Multiparameter hypothesis testing and acceptance sampling. Technometrics 24, 295–300.
Boos, D.D., Stefanski, L.A., 2013. Essential Statistical Inference, Theory and Methods. Springer, New York.
Casella, G., Berger, R.L., 2001. Statistical Inference, second ed. Wordsworth/Brooks Cole, Pacific Grove, CA.
Hsu, J.C., 1996. Multiple Comparisons, Theory and Methods. Chapman & Hall, London.
Hwang, J.T., Casella, G., Robert, C., Wells, M.T., Farrell, R.H., 1992. Estimation of accuracy in testing. Annals of Statistics 20, 490–509.
Lehmann, E.L., 1986. Testing Statistical Hypotheses, second ed. Springer, New York.
Lehmann, E.L., 1999. Elements of Large-Sample Theory. Springer-Verlag, New York.
Schervish, M., 1995. Theory of Statistics. Springer, New York.

Hypothesis Testing: Methodology and Limitations

TAB Snijders, University of Groningen, Groningen, The Netherlands

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 7121–7127, © 2001, Elsevier Ltd.

Abstract

A test is a statistical procedure to obtain a statement on the truth or falsity of a proposition, on the basis of empirical evidence. After a brief historical account, attention is given to the contrasting approaches of R.A. Fisher's significance tests and J. Neyman and E. Pearson's tests of a null against an alternative hypothesis. The *t*-test is treated as a paradigmatic example and used to illustrate the role of assumptions. The *p*-value and the confidence interval are statistical procedures which are more informative than a test leading merely to the 'reject'/'do not reject' dichotomy. A discussion is presented about problems in the interpretation and use of hypothesis tests, and it is argued that these result to a great extent from our human limitations in reasoning with uncertain evidence. Research results do usually not stand on their own (as assumed in the model of a hypothesis test), but are to be combined with other results, e.g., by meta-analysis.

Hypothesis tests are part of the basic methodological toolkit of social and behavioral scientists. The philosophical and practical debates underlying their application are, however, often neglected. The fruitful application of hypothesis testing can benefit from a clear insight into, the underlying concepts and their limitations.

The Idea of Hypothesis Testing

A test is a statistical procedure to obtain a statement on the truth of falsity of a proposition, on the basis of empirical evidence. This is done within the context of a model, in which the fallibility or variability of this empirical evidence is represented by probability. In this model, the evidence is summarized in observed data, which is assumed to be the outcome of a stochastic, i.e., probabilistic, process; the tested proposition is represented as a property of the probability distribution of the observed data.

Some History

The first published statistical test was by John Arbuthnot in 1710, who wondered about the fact that in human births, the fraction of boys born year after year appears to be slightly larger than the fraction of girls (cf. Hacking, 1965). He calculated that this empirical fact would be exceedingly unlikely (he obtained a probability of 1/483 600 000 000 000 000 000 000) if the probability of a male birth were exactly 0.5, and argued that this was a proof of divine providence, since boys – some of whom will be soldiers – have a higher risk of an early death, so that a higher ratio of male births is needed to obtain an equal ratio of males among young adults. We see here the basic elements of a test: the proposition that the male birth ratio is 0.5, related to data by regarding these as outcomes of stochastic variables, the calculation that the data would be unlikely if the proposition were true, and a further conclusion interpreting the falsity of the proposition.

One of the first statistical procedures that comes close to a test in the modern sense was proposed by Karl Pearson in 1900. This was the famous chi-squared test for comparing an observed frequency distribution to a theoretically assumed distribution. Pearson derived the now well-known statistic to test the proposition, or *hypothesis*, that the probabilities of the various possible (finitely many) outcomes of some random variable are given by certain preassigned numbers. Pearson proved that this statistic has (in a large-sample approximation) the chi-squared distribution; this distribution can therefore be used to calculate the probability that, if the hypothesis holds, the test statistic will assume a value equal to or larger than the value actually observed.

The idea of testing was further codified and elaborated in the first decades of the twentieth century, mainly by R.A. Fisher (1925). In his significance tests, the data are regarded as the outcome of a random variable X (usually a vector or matrix), which has a probability distribution which is a member of some family of distributions; the tested hypothesis, also called the null hypothesis, is an assertion which defines a subset of this family; a test statistic $T = t(X)$, which is a function of X, is used to indicate the degree to which the data deviate from the null hypothesis; and the significance of the given outcome of the test statistic is calculated as the probability, if the null hypothesis is true, to obtain a value of T which is at least as high as the given outcome. If the probability distribution of T is not uniquely determined by the null hypothesis, then the significance is the maximal value of this probability, for all distributions of T compatible with the null hypothesis. The significance probability is now often called the *p*-value (the letter *p* referring to probability). With Fisher originates the convention to consider a statistical testing result as 'significant' if the significance probability is 0.05 or less – but Fisher was the first to recognize the arbitrary nature of this threshold.

A competing approach was proposed in 1928 by J. Neyman and Egon Pearson (the son of Karl). They criticized the arbitrariness in Fisher's choice of the test statistic and asserted that for a rational choice of test statistic one needs not only a null hypothesis but also an alternative hypothesis, which embodies a proposition that competes with the proposition represented by the null hypothesis. They formalized the testing problem as a two decision problem. Denoting the null hypothesis by H_0 and the alternative H_1, the two decisions were represented as 'reject H_0' and 'do not reject H_0.' (The second decision was also

called 'accept H_0,' but it will be discussed below that this is an unfortunate term.) Two errors are possible: rejecting a true H_0, and failing to reject a false H_0. Neyman and Pearson conceived of the null hypothesis as a standard situation, the burden of proof residing with the researcher to demonstrate (if possible) the untenability of this proposition.

Correspondingly, they called the error of rejecting a true H_0 an error of the first kind and the error of failing to reject a false H_0 an error of the second kind. Errors of the first kind are considered more serious than errors of the second kind. The probability of – correctly – rejecting H_0 if H_1 is true, which is 1 minus the probability of an error of the second kind, given that the alternative hypothesis is true, they called the power of the test. Neyman and Pearson proposed the requirement that the probability of an error of the first kind, given that the null hypothesis is indeed true, do not exceed some threshold value called the significance level usually denoted by α. Further they proposed to determine the test so that, under this essential condition, the power will be maximal.

In the Neyman–Pearson formulation, we obtain richer results (namely, specific precepts for constructing good tests) at the cost of a more demanding model. In addition to Fisher's null hypothesis, we also need to specify an alternative hypothesis; and we must conceive the testing problem as a two-decision situation. This led to vehement debate between Fisher on the one hand, and Neyman and E. Pearson on the other. This debate and the different philosophical positions are summarized by Hacking (1965) and Gigerenzer et al. (1989), who also give a further historical account. The latter study also discusses how this controversy was resolved in the teaching and practice of statistics in the social sciences by a hybrid theory which combines elements of both approaches to testing, and which has been treated often as an objective precept for the scientific method, glossing over the philosophical controversies.

Examples of this hybrid character are that, in accordance with the Neyman–Pearson approach, the theory is explained by making references to both the null and the alternative hypotheses, and to errors of the first and second kind (although power tends to be treated in a limited and often merely theoretical way), whereas – in the spirit of Fisher – statistical tests are regarded as procedures to give evidence about the particular hypothesis tested and not merely as rules of behavior that will in the long run have certain (perhaps optimal) error rates when applied to large numbers of hypotheses and data sets. Lehmann (1993) argues that indeed a unified formulation is possible, combining the best features of both approaches.

Instead of implementing the hypothesis test as a 'reject/don't reject' decision with a predetermined significance level, another approach often is followed: to report the p-value or significance probability, defined as the smallest value of the significance level at which the observed outcome would lead to rejection of the null hypothesis. Equivalently, this can be defined as the probability, calculated under the null hypothesis, of observing a result deviating from the null hypothesis at least as much as the actually observed result. This deviation is measured by the test statistic, and the p-value is just the tail probability of the test statistic. For a given significance level α, the null hypothesis is rejected if and only if $p \leq \alpha$.

Hypothesis Tests in Empirical Research

The Role of the Null Hypothesis

In the social and behavioral sciences, the standard use of hypothesis testing is directed at single research questions, practically always fragments of larger investigations, about the existence of some or other effect. This effect could be a difference between two group means, the effect of some explanatory variable on some dependent variable as expressed by a regression coefficient in multiple linear regression analysis, etc. Expressed schematically, the researcher would like to demonstrate a research hypothesis which states that the effect in question, indeed, exists. The negation of this hypothesis, then, is understood as the proposition that the effect is absent, and this proposition is put up as the null hypothesis. The research hypothesis, stating that the effect exists, is the alternative hypothesis. Rejecting the null hypothesis is interpreted as support for the existence of the hypothesized effect. In this way, the burden of proof rests with the researcher in the sense that an error of the first kind is to state that there is evidence supporting the existence of the effect if, in reality, the effect is absent. The usual welcoming phrase for the rejection of the null hypothesis is the statement that a significant result has been obtained.

Example: The *t*-test

Consider, as a major example, the proposition that two groups, arbitrarily labeled A and B, are different with respect to some numerical characteristic X. Available data are measurements X_{Ai} of this characteristic for some individuals i in group A and measurements X_{Bi} of the characteristics for other individuals i in group B. The first step in modeling this is to consider the observed values X as outcomes of random variables, usually with the assumption that they are stochastically independent and have a probability distribution not depending on the individual i. The next step is, usually, to focus on the expected values, i.e., population means in the two groups, denoted here by μ_A and μ_B. The tested proposition is formalized as the statement that μ_A and μ_B are different (two-sided alternative hypothesis) or as the statement that one (e.g., μ_A) is bigger than the other (e.g., μ_B) (one-sided alternative hypothesis). The null hypothesis H_0 is defined as the statement that μ_A and μ_B are equal.

The most commonly used test for this testing problem is *Student's t-test*, called after the pseudonym of W. Gosset, who laid the mathematical basis for this test in 1908. The test statistic is

$$T = \frac{M_A - M_B}{\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right) S^2}}$$

where M_A and M_B are the two sample means, S^2 is the pooled within-group variance, and n_A and n_B are the two sample sizes. (Definitions of these quantities can be found in any statistics textbook.) This formula illustrates the property of many test statistics that an observed effect (here the difference between the two sample means) is compared to a measure of variability (based here on the within-group variance). Student/Gosset showed that, if the variable X has a normal distribution with the same variance in the two groups, then T has, if the

null hypothesis is true, the so-called t distribution on $df = n_A + n_B - 2$ degrees of freedom. For the two-sided alternative hypothesis, H_0 is rejected for large values of the absolute value of T, for the one-sided hypothesis H_0 is rejected for large values of T itself. The threshold beyond which H_0 is rejected is called critical value, and is determined from the t-distribution in such a way that the significance level has the pre-assigned value. The one-sided threshold at significance level α is denoted $t_{df;\,\alpha}$, so that the 'rejection region' for the one-sided t-test is given by $\{T > t_{df;\,\alpha}\}$. The power of the one-sided test is larger than that of the two-sided test for $\mu_A > \mu_B$ (corresponding to the one-sided alternative hypothesis) and smaller for $\mu_A < \mu_B$. This is, in the Neyman–Pearson formulation, the reason for using the one-sided test if the alternative hypothesis is one-sided (in which case parameter values $\mu_A < \mu_B$ are not taken into consideration).

The conventional value of the significance level in the social and behavioral sciences is 0.05. For large values of the combined sample size $n_A + n_B$, the critical value of the t-test approximates the critical value that can be computed from the standard normal distribution. This is because the fact that the variance is estimated and not known beforehand becomes less and less important as the combined sample size gets larger; if the population variance was known beforehand and substituted for S^2, the test statistic would have the standard normal distribution, and the test would be the so-called z-test. For this test, the critical value is 1.645 for the one-sided test and 1.960 for the two-sided test. The power depends on many quantities: the sample sizes, the population means and variances, and the significance level. As examples, the power of the one-sided z-test for $\alpha = 0.05$ is equal to 0.50 for

$$\frac{\mu_A - \mu_B}{\sigma} = 1.645\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

where σ is the within-group standard deviation. The power is equal to 0.95 if $(\mu_A - \mu_B)/\sigma$ is equal to twice this value.

The Role of Assumptions

The probability statements that are required for statistical tests do not come for free, but are based on certain assumptions about the observations used for the test. In the two-sample t-test, the assumptions are that the observations of different individuals are outcomes of statistically independent, normally distributed, random variables, with the same expected value for all individuals within the same group, and the same variance for all individuals in both groups. Such assumptions are not automatically satisfied, and for some assumptions it may be doubted whether they are ever satisfied exactly. The null hypothesis H_0 and alternative hypothesis H_1 are statements which, strictly speaking, imply these assumptions, and which therefore are not each other's complement. There is a third possibility: the assumptions are invalid, and neither H_0 nor H_1 is true. The sensitivity of the probabilistic properties of a test to these assumptions is referred to as the lack of robustness of the test. The focus of robustness studies has been on the assumptions of the null hypothesis and the sensitivity of the probability of an error of the first kind to these assumptions, but studies on robustness for deviations from assumptions of the alternative hypothesis have also been done, cf. Wilcox (1998).

One general conclusion from robustness studies is that tests are extremely sensitive to the independence assumptions made. Fortunately, those assumptions are often under control of the researcher through the choice of the experimental or observational design. Traditional departures from independent observations are multivariate observations and within-subject-repeated measures designs, and the statistical literature abounds with methods for such kinds of dependent observations. More recently, methods have been developed for clustered observations (e.g., individual respondents clustered within groups) under the names of multilevel analysis and hierarchical linear modeling.

Another general conclusion is that properties of tests derived under the assumption of normal distributions, such as the t-test, can be quite sensitive to outliers, i.e., single, or a few, observations that deviate strongly from the bulk of the observations. Since the occurrence of outliers has a very low probability under normal distributions, they are 'assumed away' by the normality assumption. The lack of robustness and sensitivity to outliers have led to three main developments.

First, there are nonparametric tests, which do not assume parametric families of distributions (such as the normal). Most of these are based on the ranks of the observations rather than their numerical values. They are standard fare in statistical textbooks. Second, robust tests have been developed, based on numerical values of the observations, which are less sensitive to outliers or heavy-tailed distributions, e.g., by some kind of automatic downweighting of outlying observations (e.g., Wilcox, 1998). The purpose of such tests is to retain a high power as much as possible while decreasing the sensitivity to deviations from the normal distribution or to departures from other assumptions.

Third, diagnostic means have been developed to spot single observations, or groups of observations, with an undue influence on the result of the statistical procedure (e.g., Cook and Weisberg, 1999; Fox, 1991). The idea behind most of these diagnostics is that most of the observations come from a distribution which corresponds well enough to the assumptions of the statistical model, but that the observations could be contaminated by a small number of poorly fitting observations. Ideally, these observations should also be recognizable by close inspection of the data or data collection process. After deleting such poorly fitting observations, one can proceed with the more traditional statistical procedures, assuming that the remaining data do conform to the model assumptions.

Confidence Intervals

Confidence intervals are closely related to hypothesis tests. They focus on a parameter in the statistical model. Examples of such parameters are, in the two sample situation described above, the difference of the two population means, $\mu_A - \mu_B$, or the within-group standard deviation, σ. To frame this in general terms, consider a one-dimensional statistical parameter denoted by θ. A null hypothesis test is about the question whether the data is compatible with the hypothesis that the parameter θ (in this case μ_A–μ_B) has the particular value 0. Instead, one could put forward the question: what are the values of θ with which the data is compatible? This question can be related to

hypothesis testing by considering the auxiliary problem of testing the null hypothesis $H_0{:}\theta = \theta_0$ against the alternative hypothesis $H_1{:}\theta \neq \theta_0$, for an arbitrary but fixed value θ_0. The data may be said to be compatible with any value of θ_0 for which this null hypothesis would not be rejected at the given level of significance. Thus, the confidence interval is the interval of non-rejected null hypotheses.

Another way of viewing the confidence interval is as follows. A confidence coefficient $1 - \alpha$ is determined by the researcher (a conventional value is 0.95). The confidence interval is an interval with lower boundary L and upper boundary U, calculated from the data and therefore being random variables, with the property that the probability that $L \leq \theta \leq U$, i.e., the interval contains the true parameter value, is at least $1 - \alpha$. It can be proved mathematically that the interval of non-rejected null hypotheses has precisely this property.

The confidence interval can be inverted to yield a hypothesis test for any null hypothesis of the form H_0: $\theta = \theta_0$ in the following way. For the most usual null hypothesis that $\theta = 0$, the hypothesis is rejected if the value 0 is not included in the confidence interval. More generally, the null hypothesis $H_0{:}\theta = \theta_0$ is rejected whenever θ_0 is outside the confidence interval. This shows that a confidence interval is strictly more informative than a single hypothesis test.

Problems in the Interpretation and Use of Hypothesis Tests

While hypothesis tests have been applied routinely by hosts of social and behavioral scientists, there have been continuing debates, sometimes vehement and polemic, about their use, among those inclined to philosophical or foundational discussion. Many contentious issues can be found in Harlow et al. (1997), Nickerson (2000), and Wilkinson and TFSI (1999). Within the context of the present contribution, I can only give a personally colored discussion of some main points which might contribute to a better understanding and a more judicious application of hypothesis tests.

The reason that there has been so much criticism of the nature and use of hypothesis tests, is in my view the difficulty of reasoning with uncertain evidence and the natural human tendency to take recourse to behavioral rules instead of cogitate afresh, again and again, about how to proceed. The empirical researcher would like to conclude unequivocally whether a theory is true or false, whether an effect is present or absent. However, experimental and observational data on human behavior are bound to be so variable that the evidence produced by these data is uncertain. Our human tendency is to reason – if only provisionally – as if our conclusions are certain rather than tentative; to be concise rather than accurate in reporting our arguments and findings; and to use simple rules for steering how we make inference from data to theories. We are tempted to reason as if '$p > \alpha$' implies that the investigated effect is nil, and the presumed theory is false, while '$p \leq \alpha$' proves that the effect indeed exists and the theory is true.

However, such cognitive shortcuts amount to serious misuse of hypothesis testing. Much of the criticism of the use of hypothesis testing therefore is wellfounded. In my view, this implies that better interpretations of hypothesis testing should be promoted, and that tests should be used less mechanically and combined with other argumentations and other statistical procedures. In any case, the user of tests (or other statistical procedures) should realize that these procedures are based on the assumption that there is random variability in the data which cannot be completely filtered out of the results. Whether a test result is significant or not depends partly on chance, and each researcher obtaining a significant result should be aware that this could be an error of the first kind, just as a non-significant result could be an error of the second kind.

Some Misinterpretations

The following list tries to correct misinterpretations that have haunted applications of statistical tests. A more extensive discussion can be found, e.g., in Nickerson (2000).

a. A common misinterpretation is that non-rejection implies support for the null hypothesis. Non-rejection should be interpreted, however, as an undecided outcome: there is not enough evidence against the null hypothesis, but this does not imply that there is evidence for the null hypothesis. Maybe the sample size is small, or error variability is large, so that the data does not contain much information anyway. Usually, the alternative hypothesis contains probability distributions that approximate the null distribution; e.g., when testing $\mu_A - \mu_B = 0$ against $\mu_A - \mu_B \neq 0$, the difference between the two population means can be tiny, while the alternative hypothesis is still true. In other words, the power of practically all tests is quite low in a sizeable part of the alternative hypothesis, so if one of the distributions in this part would prevail, chances would be low of rejecting the null hypothesis. Therefore, non-rejection provides support for those parts of the alternative practically as strongly as for the null hypothesis itself, and non-rejection may not be interpreted as support for the null hypothesis and against the alternative hypothesis.
b. If one wishes to get more information about whether a nonsignificant result provides support for the null hypothesis, a power study is not the answer. Statistical power is the probability to reject the null hypothesis, if a given effect size obtains. Its interpretation cannot be inverted as being the degree of support of the null hypothesis in the case of nonsignificance. Power studies are important in the stage of planning a study. Once the study has been conducted, and one wishes to see in more detail to what extent the null and alternative hypotheses are supported, a confidence interval is the appropriate procedure (also see Wilkinson and TFSI, 1999, p. 596).
c. Test results tell us nothing about the probabilities of null or alternative hypothesis. The probabilities known as the significance level or the power of the test are probabilities of certain sets of outcomes given the condition that the null or, respectively, the alternative hypothesis is true.
d. A significant result is not necessarily important. Very low p-values do not in themselves imply large effect sizes. A small estimated effect size still can yield a low p-value, e.g., if the corresponding standard error is very small, which can be caused by a large sample or by good experimental

control. Investigation of effect sizes is important, and is a different thing than hypothesis testing. The importance of reporting effect sizes is stressed in Schmidt (1996) and Wilkinson and TFSI (1999).

e. The alternative hypothesis is not the same as a scientific theory. If a researcher is investigating some theory empirically, this can be based on tests of which the alternative hypotheses are deduced from the theory, but always under the assumption that the study was well designed and usually under additional technical assumptions such as, e.g., normality of the distributions. Since the alternative hypothesis is only a consequence and not a sufficient condition for the theory, it is possible that the alternative hypothesis is true but the theory is not. To find support for theories, one has to find corroborations of many consequences of the theory under various circumstances, not just of one of its implications. On the other hand, it is possible that the theory is true but the alternative hypothesis deduced from it is not, e.g., because the experiment or observation was ill designed or because the auxiliary technical assumptions are not satisfied.

f. The null hypothesis is just what it says: a hypothesis. The test is not invalidated by the mere fact that there are other reasons for thinking, or knowing, that the null hypothesis is wrong. Even if there may be other reasons for which the null hypothesis is wrong, still it can be sensible to check whether the data at hand are, or are not, compatible with it.

Limitations of the Neyman–Pearson Approach to Hypothesis Testing

The Neyman–Pearson (1928) formulation changed the concept of hypothesis testing because it provided a rational basis for the choice of a test statistic. This was obtained at the cost of a quite narrow frame for the notion of a hypothesis test: a two-decision problem in which one error ('type I') is considered much more serious than the other error ('type II'). This model for decision-making is often inappropriate and, when it is useful, it is more commonly a useful approximation than a faithful representation of the actual decision situation. Some of its limitations are discussed briefly in this section.

a. Even when a dichotomous decision must be taken as a result of a scientific investigation, it is possible that both types of error should be considered equally serious. This is the case, e.g., when two therapies are compared and the most promising one must be chosen for further study or for practical application. If the costs of both kinds of error (choose therapy B, while A is better; or choose therapy A, while B is better) are about the same, then there is no asymmetry between the two competing hypotheses, and a significance test is not in order.

b. Often there are more than two outcomes of the decision situation. For example, in the two-sample situation discussed above, where the population means are μ_A and μ_B, and where the researcher entertains a theory implying that $\mu_A > \mu_B$, it is natural to define three decision outcomes defined as '$\mu_A > \mu_B$ (support for the theory),' 'undecided,' and '$\mu_A < \mu_B$ (evidence against the theory).' Usually, however, such a situation is represented by a Neyman–Pearson testing problem with null hypothesis H_0:'$\mu_A = \mu_B$.' and alternative H_1:'$\mu_A > \mu_B$'. If the t-statistic for $\mu_A - \mu_B$ yields a strongly negative outcome ('significance in the wrong direction') then the researcher only concludes that the null hypothesis is not rejected, whereas the data conveys the message that the alternative hypothesis should be rejected. The data then have to be considered as evidence against the theory embodied in H_1. Such a three-decision formulation is closer to the actual purpose of hypothesis tests in many situations where a one-dimensional parameter is being tested, cf. Leventhal (1999).

c. The Neyman–Pearson approach is formulated as if the data at hand is the only evidence available about the null and alternative hypotheses: it is an *in vacuo* formulation. This is almost never realistic. Often the propositions under study are theoretically plausible or implausible, there is earlier empirical evidence, there may be converging evidence (as in the case of triangulation studies) in the same investigation, and other evidence may become available in the future. In the light of this surrounding evidence, it is unrealistic – perhaps arrogant – to presume that the researcher has to take either of two decisions: 'reject H_0' or 'do not reject H_0.' It is more realistic to see the data at hand as part of the evidence that was and will be accumulated about these hypotheses. This is a major reason why presenting the result of the test as a p-value is more helpful to other researchers than presenting only the 'significant–nonsignificant' dichotomy (also see Wilkinson and TFSI, 1999).

There are several ways in which the results obtained from the data at hand, perhaps summarized in the p-value, can be combined with other evidence. In the first place, an informal combination can be and often will be made by a sensible, nonformalized weighing of the various available pieces of evidence. One formal way ofcombining evidence is provided by the Bayesian paradigm (e.g., Gelman et al., 1995). This paradigm presupposes, however, that the prior and current ideas of the researcher about the plausibility of all possible values of the statistical parameter under study be represented in a probability distribution for this parameter, a very strong requirement indeed. Other formal ways of combining evidence have been developed and are known collectively by the name of *meta-analysis* (e.g., Cook et al., 1992; Schmidt, 1996). These methods require that the several to-be-combined hypothesis tests address the same substantive hypothesis, or that they can be regarded as tests of the same substantive parameter.

Conclusion

The statistical hypothesis test is one of the basic elements of the toolbox of the empirical researcher in the social and behavioral sciences. The difficulty of reasoning on the basis of uncertain data has led, however, to many misunderstandings and unfortunate applications of this tool. The Neyman–Pearson approach is a mathematically elegant formulation of what a hypothesis test could be. Its precise formulation rarely applies literally to the daily work of social and behavioral scientists, very often it is a useful approximation to a part of the research question, sometimes it is inappropriate – but sacrosanct it is not. A

hypothesis test evaluates the data using a test statistic set up to contrast the null hypothesis with the alternative hypothesis, and the *p*-value is the probability to obtain, *if* the null hypothesis is true, outcomes of the test statistic that are at least as high as the outcome actually observed. Low values therefore are evidence against the null hypothesis, contrasted with the alternative hypothesis. It is more conducive to the advance of science to report *p*-values than merely whether the hypothesis was rejected at the conventional 0.05 level of significance.

In the interpretation of test outcomes, one should be aware that these are subject to random variability; and that the probability calculations are based on assumptions which may be in error. Incorrect assumptions can invalidate the conclusions seriously. Nonparametric and robust statistical procedures have been developed which depend less on these kinds of assumption, and diagnostics have been developed for checking some of the assumptions. When reporting empirical results, it is usually helpful to report not only tests but also estimated effect sizes and/or confidence intervals for important parameters. To sum up: 'It seems there is no escaping the use of judgment in the use and interpretation of statistical significance tests' (Nickerson, 2000, p. 256).

Bibliography

Cook, R.D., Weisberg, S., 1999. Applied Regression Including Computing and Graphics. Wiley, New York.

Cook, T.D., Cooper, H., Cordray, D.S., Hartmann, H., Hedges, L.V., Light, R.J., Louis, T.A., Mosteller, F., 1992. Meta-analysis for Explanation. A Casebook. Russell Sage Foundation, New York.

Fisher, R.A., 1925. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, UK.

Fox, J., 1991. Regression Diagnostics. Sage, Newbury Park, CA.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman & Hall, London.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Krüger, L., 1989. The Empire of Chance. Cambridge University Press, New York.

Hacking, I., 1965. Logic of Statistical Inference. Cambridge University Press, Cambridge, UK.

Harlow, L.L., Mulaik, S.A., Steiger, J.H. (Eds.), 1997. What if there were no Significance Tests? Erlbaum, Mahwah, NJ.

Lehmann, E.L., 1993. The Fisher, Neymann-Pearson theories of testing hypotheses: one theory or two? Journal of the American Statistical Association 88, 1242–1249.

Leventhal, L., 1999. Updating the debate on one- versus two-tailed tests with the directional two-tailed test. Psychological Reports 84, 707–718.

Neyman, J., Pearson, E.S., 1928. On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika 20A (175–240), 263–294.

Nickerson, R.S., 2000. Null hypothesis significance testing: A review of an old and continuing controversy. Psychological Methods 5, 241–301.

Schmidt, F.L., 1996. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods 1, 115–129.

Wilcox, R.R., 1998. The goals and strategies of robust methods (with discussion). British Journal of Mathematical and Statistical Psychology 51, 1–64.

Wilkinson L., Task force on statistical inference 1999. Statistical methods in psychology journals. American Psychologist 54, 594–604.

Hypothesis Tests, Multiplicity of

Lyle V Jones, University of North Carolina at Chapel Hill, NC, USA
Charles Lewis, Fordham University, Bronx, NY, USA
John W Tukey[†]**,** Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 10, pp. 7127–7133, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Standard procedures for controlling error rates of statistical inferences assume that only a single inference is drawn from the analysis of a body of data. When an investigation yields a multiplicity of inferences, an adjustment of the error rates from standard procedures is required. Here we distinguish between two forms of multiplicity, 'multiple determination' and 'multiple comparisions.' For both, we set forth techniques for the control of a particular error rate; the rate of false discoveries (as proposed by Benjamini & Hochberg). The techniques are illustrated by applying them to state-by-state results from the National Assessment of Educational Progress. Extensions are discussed, with special attention to whether levels of the independent or classification variables should be considered 'fixed,' 'variable,' or neither.

Standard procedures for controlling the probability of erroneous statistical inferences apply only for a single comparison or determination from a given set of data. Multiplicity is encountered whenever a data analyst draws more than one inference from a data set. With multiplicity, standard procedures must be adjusted for the number of comparisons or determinations that actually or potentially are performed. If some adjustment is not made, the probability of erroneous inferences can be expected to increase with the degree of multiplicity. In other words, the data analyst loses control of a critical feature of the analysis.

Questions arising from problems of multiplicity raise a diversity of issues, issues that are important, difficult, and not fully resolved. This account emphasizes the issues, tries to avoid being trapped in undue detail, and provides examples of the application of a recommended procedure.

It is helpful to recognize several general guidelines:

a. Two approaches to the same data, with different purposes or with different questions being posed, need not yield compatible results.
b. Exact equality of long-run values is very rare and should not play an appreciable role either in one's thinking or one's practice.
c. When posing questions about data, greater realism will be very valuable.
d. The situations faced must not be characterized either as completely 'fixed' or completely 'random.'
e. In selecting a bouquet of techniques for the next decade or the next millennium, the selection must not be cast in stone; further evolution is to be expected (in later months, years, or centuries).

Issues of multiplicity appear in a wide variety of settings, including

a. outcomes, each measured with variability, in several situations:
   1. comparison of all pairs of situations ('multiple comparisons'); and
   2. comparison between each situation and a standard ('multiple determination');
b. selection of one or more elements from a set of alternatives:
   1. selection of how variables are to be expressed for analysis (i.e., what re-expression, if any, should be used?);
   2. selection of which of the available variables are to participate in an analysis (subset selection); and
   3. reduction in size of a set of alternatives in an attempt to reduce their number without losing the best.

Illustrative Settings, Error Rates, and Procedures

Some important issues arise for all of these alternatives. These issues are considered and illustrated for one example of multiple determination and one example of multiple comparisons.

Consider k independent estimates $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$ of, respectively, μ_1, $\mu_2, \ldots$, μ_k and an estimate of variability applicable to all these estimates. This is typical enough of multiplicity problems to render conclusions that are broadly helpful.

Multiple Determination

Defining a standard, λ, and for each μ_i, drawing conclusions about $\mu_i - \lambda$ based on $\bar{y}_i - \lambda$ is, by definition, making a multiple determination. (In some applications, μ_i is a mean change from one time to another. Often in that case, $\lambda = 0$.) Experience and common sense suggest strongly that μ_i and λ are certain to differ in some decimal place, so that a focus on a null hypothesis that $\mu_i = \lambda$ is quite unrealistic and misleading.

With an interest in $\mu_i - \lambda$ for each value of i, it is reasonable to ask for one of these conclusions: (1) to act as if $\mu_i < \lambda$; (2) to act as if the sign of $\mu_i - \lambda$ is not (yet) determined; or (3) to act as if $\mu_i > \lambda$. Here (1) or (3) is a 'definite' result, while (2) is 'indefinite.'

[†] Deceased.

Multiple Comparisons

Consider the $k(k-1)/2$ individual pair-wise comparisons $\gamma_i - \gamma_j$. Using these to draw conclusions about the set of $\mu_i - \mu_j$ is, by definition, making multiple comparisons, although the definition is more general, and applies to any set of linear contrasts among the μ_i. Because μ_i and μ_j are certain to differ in some decimal place, there is no good reason to focus on the null hypothesis that $\mu_i = \mu_j$. Instead, it should be asked, for each (i,j) pair, whether one should (1) act as if $\mu_i < \mu_j$; (2) act as if the sign of $\mu_i - \mu_j$ were not (yet) determined; or (3) act as if $\mu_i > \mu_j$.

In General

Notice the use of '>' or '<' and the avoidance of '=' and '≠.' 'Accepting the null hypothesis' is usually at best misleading. Confounding '>' and '<' as '≠' is unresponsive and confusing. Thus, focusing on direction is important when dealing with only one comparison (or determination), and becomes vital when dealing with several or many comparisons.

In these formulations, an 'error' is asserting one direction when the (unknown) truth is the other direction. (One should forget about the 'null case' of equality.) Asserting that the direction is not yet established may be a wasted opportunity, but is not an 'error.' The task is to control error while minimizing wasted opportunity, i.e., minimizing indefinite results.

Two classical choices for controlling errors consider expected error rates over replications of the study. (See, for example, Hochberg and Tamane, 1987: 7.) They are:

$$E\left\{\frac{\text{number of errors}}{\text{number of comparisons}}\right\} \leq 0.025$$

('individual,' 'unadjusted,' or 'per comparison')

and

$$E\left\{\frac{\text{number of errors}}{\text{number of families}}\right\} \leq 0.025$$

('simultaneous,' 'Bonferroni,' or 'per family').

Here, 0.025 could easily be replaced by another choice. (For consistency with commonly used notation, the chosen value will be denoted by $\alpha/2$.) 'Family' refers to an identified set of comparisons whose simultaneous error rate is to be controlled. Normally, 'number of families' would be taken equal to unity. These two choices for controlling errors differ by a factor of k for multiple determinations and of $k(k-1)/2$ for multiple comparisons, so the difference between the two can be quite large.

There has been a feeling for decades (e.g., Tukey, 1994/1953) that 'individual' or 'unadjusted' is too soft while 'simultaneous' is too severe. More recently, Benjamini and Hochberg (1995) introduced a compromise that is receiving increasing acceptance.

The two parts of their innovation are a definition of an alternative error rate, the false discovery rate (*FDR*), and a moderately simple procedure that, under the simplest circumstances, controls the *FDR*. The definition is the expectation of the ratio of the number of erroneous assertions of direction to the total number of assertions of direction, i.e.,

$$\text{false discovery rate} = \text{FDR} = E\left\{\frac{\text{number of errors}}{\max\{1,\ \text{number of definite results}\}}\right\}$$

Let p_{crit} be the tail area (usually for each of two tails) of the null sampling distribution of the test statistic for any single comparison being judged by a multiplicity-respecting procedure. Each such procedure will stipulate a probability of error or average fraction of error that is bounded by $\alpha/2$ when judging confidence of direction. The value of p_{crit} will depend on the sort of confidence to be attained. Let K be the number of determinations or comparisons in the family under consideration, where $K = k$ for multiple determinations and $K = k(k-1)/2$ for all pair-wise multiple comparisons. Let $h = 1, \ldots, K$ be the rank of the p-value associated with the test statistic for the comparison concerned when ordered from smallest to largest, so that the observed p-values – $p_{(h)}$ for the hth comparison – are weakly increasing from $h = 1$ to $h = K$. Three distinct procedures, two classical choices and one that controls the *FDR*, are defined as follows:

Unadjusted: be confident of the observed direction for comparison h if

$$p_{(h)} \leq p_{\text{crit}} = p_{UNA} = \alpha/2$$

FDR (Benjamini and Hochberg, 1995): be confident of the observed direction of the hth comparison when $p_{(h)}$ is the largest p-value such that

$$p_{(h)} \leq p_{\text{crit}} = p_{\text{B-H}}(h) = (h/K)\alpha/2.$$

Moreover, be confident of the observed direction for all comparisons g with

$$p_{(g)} \leq p_{(h)}, \text{i.e.}, g = 1, \ldots, h.$$

Bonferroni (per family): be confident of the observed direction for comparison h if

$$p_{(h)} \leq p_{\text{crit}} = p_{BON} = (1/K)\alpha/2.$$

Because

$$(1/K) \leq (h/K) \leq 1$$

it must be true that

$$p_{BON} \leq p_{\text{B-H}}(h) \leq p_{UNA}$$

Specifically, $p_{\text{B-H}}(h) = (h)p_{BON}$ and $p_{UNA} = (K)p_{BON}$. The three collections of confident directions corresponding to these three approaches are nested, with the collection for the unadjusted procedure the largest and that for the Bonferroni the smallest.

Examples: The NAEP Trial State Assessment

Data from the National Assessment of Educational Progress (NAEP) Trial State Assessment were subjected to two separate analyses by Williams et al. (1999). The data for the first analysis consist of average eighth-grade mathematics proficiency scores for the 34 states that participated in both the 1990 and 1992 NAEP assessments; data for the second analysis consist of average eighth-grade mathematics proficiency scores for the 41 states that participated in 1992 (Johnson et al. 1993).

Multiple Determination

One way of assessing year-to-year change in state mathematics achievement involves directly testing the significance of change

in each of the 34 states. An ordinary Bonferroni adjustment for control of Type I error per family, as was used by NAEP (Johnson et al, 1993), results in an effective critical *p*- value of $p_{BON} = 0.000\,735$ for $\alpha/2 = 0.025$. Table 1 contains the computed difference between 1990 and 1992 mathematics achievement means for each state and the computed pooled standard error. It also includes the statistical decision about confidence in direction (with confident direction indicated by '*b*' in each column) for each state under the three multiplicity treatments. For this example, critical *p*- values are used as a basis for comparison of the adjustment procedures. The cutoff critical *t* value for each procedure is also displayed in the bottom line of Table 1.

The unadjusted per comparison approach is confident about the direction of 15 differences, whereas the Benjamini–Hochberg (B–H) *FDR* procedure is confident about 11 of these, and the Bonferroni technique is confident about the direction of only four of the differences.

Multiple Comparisons

All pair-wise mean differences between the states' 1992 eighth-grade mathematics achievement scores were considered. Because 41 states participated in the 1992 assessment, the family size is $K = 41 \times 40/2 = 820$ comparisons. Using the Bonferroni adjustment, there are 480 confident directions between states, and the use of the B–H technique results in an additional 159. The unadjusted analysis increases the number of confident directions beyond the B–H technique by only six.

For Both Examples

For each example above, the B–H *FDR* procedure has led to a greater number of confident directions than the Bonferroni procedure. For the examples, we may expect that no more than 0.025 of the confident directions identified by the B–H

Table 1 Mean eighth-grade mathematics achievement change, 1990 to 1992 (and standard error), by state, and *t*, *p*- value, and p_{crit} values for three multiple comparison adjustments, $k = 34$[a]

State	$X_{92} - X_{90}$	(s)	t	p- *value*	p_{UNA}	p_{B-H} (h)	p_{BON}
GA	− 0.323	(1.775 71)	− 0.182	0.428 14	0.025	0.025 000	0.000 735
AR	− 0.777	(1.485 29)	− 0.523	0.301 41	0.025	0.024 265	0.000 735
AL	− 1.568	(2.017 45)	− 0.777	0.220 04	0.025	0.023 529	0.000 735
NJ	1.565	(1.927 28)	0.812	0.209 99	0.025	0.022 794	0.000 735
NE	1.334	(1.527 72)	0.873	0.193 20	0.025	0.022 059	0.000 735
ND	1.526	(1.685 52)	0.905	0.184 45	0.025	0.021 324	0.000 735
DE	1.374	(1.346 51)	1.020	0.155 81	0.025	0.020 588	0.000 735
MI	2.215	(1.847 27)	1.199	0.117 61	0.025	0.019 853	0.000 735
LA	2.637	(2.079 43)	1.268	0.104 82	0.025	0.019 118	0.000 735
IN	2.149	(1.635 56)	1.314	0.096 94	0.025	0.018 382	0.000 735
WI	2.801	(1.962 69)	1.427	0.079 36	0.025	0.017 647	0.000 735
VA	2.859	(1.929 92)	1.481	0.071 87	0.025	0.016 912	0.000 735
WV	2.331	(1.396 39)	1.669	0.050 13	0.025	0.016 176	0.000 735
MD	3.399	(1.923 20)	1.767	0.041 13	0.025	0.015 441	0.000 735
CA	3.777	(2.114 60)	1.786	0.039 56	0.025	0.014 706	0.000 735
OH	3.466	(1.850 22)	1.873	0.032 95	0.025	0.013 971	0.000 735
NY	4.893	(2.531 95)	1.933	0.029 01	0.025	0.013 235	0.000 735
PA	4.303	(2.205 45)	1.951	0.027 86	0.025	0.012 500	0.000 735
FL	3.784	(1.932 66)	1.958	0.027 45	0.025	0.011 765	0.000 735
WY	2.226	(1.096 41)	2.030	0.023 39	0.025[b]	0.011 029	0.000 735
NM	2.334	(1.148 16)	2.033	0.023 25	0.025[b]	0.010 294	0.000 735
CT	3.204	(1.534 43)	2.088	0.020 52	0.025[b]	0.009 559	0.000 735
OK	4.181	(1.754 67)	2.383	0.010 18	0.025[b]	0.008 824	0.000 735
KY	4.327	(1.618 04)	2.674	0.004 82	0.025[b]	0.008 088[b]	0.000 735
AZ	4.994	(1.851 10)	2.698	0.004 52	0.025[b]	0.007 353[b]	0.000 735
ID	2.956	(1.067 75)	2.769	0.003 74	0.025[b]	0.006 618[b]	0.000 735
TX	5.645	(1.887 70)	2.990	0.002 02	0.025[b]	0.005 882[b]	0.000 735
CO	4.326	(1.388 68)	3.115	0.001 41	0.025[b]	0.005 147[b]	0.000 735
IA	4.811	(1.488 05)	3.233	0.001 00	0.025[b]	0.004 412[b]	0.000 735
NH	4.422	(1.353 99)	3.266	0.000 90	0.025[b]	0.003 676[b]	0.000 735
NC	7.265	(1.587 01)	4.578	0.000 01	0.025[b]	0.002 941[b]	0.000 735[b]
HI	5.550	(1.171 34)	4.738	0.000 01	0.025[b]	0.002 206[b]	0.000 735[b]
MN	6.421	(1.352 26)	4.748	0.000 01	0.025[b]	0.001 471[b]	0.000 735[b]
RI	5.097	(0.948 44)	5.374	0.000 00	0.025[b]	0.000 735[b]	0.000 735[b]
t_{crit}					2.00	2.47	3.33

[a]Reproduced from Williams, V.S.L., Jones, L.V., Tukey, J.W., 1999. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. Journal of Educational and Behavioral Statistics 24, 42–69. Copyright 1999 by the American Educational Research Association and the American Statistical Association; reproduced with permission from the publisher.
[b]Confident direction of change.

procedure are in error. (This is what it means to control the *FDR*.) As is generally the case, the relative advantage of the *FDR* procedure becomes greater with increasing numbers of comparisons (Kesselman et al. 1999).

General Considerations

Qualitative Status

In a situation where only a very few of the results are definite, the B–H false discovery procedure performs much like the Bonferroni simultaneous approach. Because these situations include those where 'definite' is purely accidental, as in extreme instances of 'data mining,' it is important that the approach selected resembles a severe procedure, and the *FDR* approach does so.

In a situation where the total number of definite results is a large fraction of the total number of comparisons an algorithm that controls the *FDR* behaves like the unadjusted procedure. If many comparisons deserve a definite result, it is reasonable that the selected approach should not lead to excessively many errors, and again, the *FDR* approach satisfies the criterion.

Family Size

One of the many problems associated with multiplicity is a frequent strong dependence of result on the number of comparisons in the family. Applying 'simultaneous' to Vermont vs. Maine as New England states can give results fairly different from applying it to Vermont vs. Maine as two of the 'lower 48,' with a family more than 75 times as large in the latter case. Controlling *FDR* turns out to be much less dependent on family size than the use of simultaneous control.

If the individual elements, e.g., the individual determinations, are stochastically independent, then it is known (Benjamini and Hochberg, 1995) that using the *FDR* algorithm does control the rate of false discoveries, and may be unnecessarily conservative. It is reasonable to believe that the result of applying the *FDR* algorithm to elements lightly correlated in the sort of way arising from multiple comparisons will be even more conservative.

For making multiple comparisons in the special case where the true values follow an idealized Gaussian distribution, simulations have shown (Lewis and Tukey, 2001) that the algorithm can be made less conservative while still protecting the rate of false discoveries, at least when the number of values being compared is no greater than 25. The modified algorithm follows the B–H approach, but uses $(h/k)\sqrt{(h+1)/2}$ in place of (h/K) in the formula for $p_{\text{B–H}}(h)$, with a caveat that enlargement should stop when the raw modification would otherwise be more liberal than the unadjusted analysis. This modification, here denoted by 'TSR,' results in the following expression for the series of critical values:

$$p_{TSR}(h) = \min\left\{1, (h/K)\sqrt{(h+1)/2}\right\}(\alpha/2)$$

Simulations indicate that a substantial improvement in power may be attained with this modification to the B–H procedure in the situations for which it is appropriate.

For additional recent work on procedures that control the *FDR*, see Shaffer (1999) and Benjamini and Hochberg (2000).

'Fixed' and 'Variable'

A major source of debate associated with multiplicity and factorial design has to do with 'fixed' vs. 'variable,' i.e., fixed versus random. A significant part of the confusion and conflict can be traced to an unwarranted belief, in tests of hypotheses, that mean squares themselves (or what mean squares estimate) are important. With interest just in directionality, each of the individual questions and answers in multiple comparisons involve $\bar{y}_i - \bar{y}_j$, where i and j are specified, and there is *no sense in asking* whether i and j are fixed or variable. (It may make sense, as shown below, to ask if something else is fixed or variable.)

A Prototype

The simplest multiple-comparison situation that illustrates the issues is a two-way table of responses with some replication in the cells and questions directed to differences between rows. Starting with k rows and m columns, each comparison involves only two rows. Replacing two rows of measurements by one row of estimated differences yields a 1-by-m table with some replication. (Readers may choose their favorite identities for rows and for columns; in particular, COLUMNS = blocks and ROWS = varieties or treatments gives appropriate realizations.)

If the columns and calculations are fixed, then the uncertainty about the answer is a consequence of whatever makes replicates within a cell different. An estimate of the corresponding variance is, say, an f-error term.

If the columns or calculations would be different were the data to be collected again, in particular, if different applications would require different weights, then the uncertainty about the overall results would reflect two sources of variability: whatever makes replicates differ, and the long-run differences (for a selected comparison) between one column and another.

(The second would not appear if the columns, etc., would be the same.) The resulting error term will be larger than the f-error term; just how much larger will depend on the extent to which differences in columns and calculations are reflected in the overall comparison.

In one crisply specified case where the columns are a random sample from a specifiable infinite population of columns, (and the calculations are fixed), there is a well-determined error term, say, the r-error term. Clearly

$$\text{r-error term} > \text{f-error term}$$

In practice, it is not likely to have exactly either columns and calculations fixed, or columns that constitute a simple random sample from a population.

Two alternatives, 'fixed' and 'variable,' are not enough.

A good way to provide a reasonable amount of realism is to define 'c' by

$$\begin{aligned}\text{appropriate error term} &= (c)\,[\text{r-error term}] \\ &\quad + (1-C)[\text{f-error term}]\end{aligned}$$

so that 'everything fixed' corresponds to $c = 0$ and 'random' is a particular case of $c = 1$. It pays then to learn as much as possible about values of 'c' in the real world.

Mathematization vs. Realism

Those who believe that data analysis can and should be mathematized are likely to be neither happy about the introduction of 'c' nor highly tolerant of a major role for such a quantity that is not estimable from data. When they recognize, however, that 'c' is about the data analyst or the data analyst's client – about just what question is being asked or about what is the most realistic question that can reasonably be asked – then they may be more willing to think about 'c' and even to work with its default values. Those who recognize the need for more realism today (and even more tomorrow), however, must be prepared for a growing seriousness about 'c' and a growing impact of nonzero values of 'c' on the conclusions to be drawn from the analysis of data. Some may find it useful to give 'c' a Bayesian interpretation, but this is not a necessity.

The only context in which something like being serious about 'c' has passed into a subject-matter field seems to be climatology. The usefulness of a climatological result is clearly measured by its performance over the set of years from its publication until the publication of its next revision. It will be based on what happened, quantitatively, in a set of years before its construction. Weather is a nontrivial time series (surely not a white noise), so neither $c = 0$ nor $c = 1$ can be expected to do well. Empirical study has led to using a base period for climatology of roughly the last 25 years. This is not directly related to a 'c' but is of a similar flavor.

Most data analysis intended to illuminate the real world is rather like climatology. The data, whether experimental or historical, has been collected under some circumstances and in the past, while its analysis has been carried out to estimate what will or what would happen in the future. The strength of the link from past to future is important; judging it effectively is not likely to be easy. Difficult issues need to be faced. If analyses are to be increasingly realistic, decisions about those issues will be required.

Plausible Defaults for 'c'

To begin to face the tasks, some values of 'c' can be set forth for some crudely described situations. Using such a list as a basis for choosing a 'c' in real data analyses is likely to be better than doing what has become customary, i.e., to assume either that $c = 0$ or $c = 1$. Values on this list will be likely to change over time as more is learned. (Here 'columns' refers to whatever versions are involved.)

$c = 0.0$ fixed columns and calculation
$c = 0.1$ fixed columns and somewhat variable calculation
$c = 0.5$ fixed columns and illustrative weights
$c = 1.0$ random columns and fixed weights
$c = 1.4$ somewhat illustrative columns
$c = 1.6$ illustrative columns
$c = 2.0$ very illustrative columns

For simple random sampling without replacement from a finite population, the value of 'c' is 1-(sample size)/(population size).

More Quantitative Conclusions

The focus so far has been on directionality, an appropriate focus when not dealing with crisp measurements, when the hope is to be convinced by the analysis about the definiteness of at least some directions. Especially in some branches of physical science and engineering, the appropriate interest is in the sizes of differences, sizes that may be measured to a small fraction of their nominal value. Instead of directions, the focus will then be upon confidence intervals.

If the concern really is with all the pair-wise differences among k varieties, treatments, or circumstances, or if the pair-wise differences we are concerned about are selected independently of the numbers to be differenced, the Benjamini and Hochberg approach has to regard the results for all differences as definite so that the *FDR* approach could be taken to coincide with the individual unadjusted approach.

If it really is desired to make a simultaneous statement about 'all differences' or to validate a process of focusing on the single difference or the few differences that appear to be most different from a reasonable anticipation, the only viable alternative is to take a simultaneous approach, presumably using percentage points for a Studentized-range distribution.

These two procedures need not yield compatible results. A difference may have a definite direction by *FDR* but a simultaneous confidence interval that includes long-run differences of both signs. Such incompatibility should be regarded only as evidence that the analyst needs to be clear which question is being asked. Is the analyst looking for a few definite directions (or for all the directions that can reasonably be believed to be definite) or is the interest in providing intervals covering the worst that is likely to have happened for any difference?

Other Kinds of Multiplicity

The broad issues presented here in discussing multiple determination and multiple comparisons appear in most of the other situations where multiplicity matters. Details may vary, but the broad issues are parallel. For more comprehensive treatments of multiplicity, see, for example, Miller (1981), Hochberg and Tamane (1987), and Hsu (1996).

See also: Hypothesis Testing in Statistics; Replication.

Bibliography

Bender, Ralf, Lange, Stefan, 2001. Adjusting for multiple testing – when and how? Journal of Clinical Epidemiology 54 (4), 343–349.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57, 289–300.

Benjamini, Y., Hochberg, Y., 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. Journal of Educational and Behavioral Statistics 25, 60–83.

Dudoit, Sandrine, Shaffer, Juliet Popper, Boldrick, Jennifer C., 2003. Multiple hypothesis testing in microarray experiments. Statistical Science 18 (1), 71–103.
Hochberg, Y., Tamane, A.C., 1987. Multiple Comparison Procedures. John Wiley & Sons, New York.
Hsu, J.C., 1996. Multiple Comparisons: Theory and Methods. Chapman & Hall, New York.
Johnson, E.G., Mazzeo, J., Kline, D.L., 1993. Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics. Report 23–ST05. National Center for Education Statistics, Washington, DC.
Keselman, H.J., Cribbie, R., Holland, B., 1999. The pairwise multiple comparison multiplicity problem: An alternative approach to familywise and comparisonwise Type I error control. Psychological Methods 4, 58–69.
Hsueh, Huey-miin, Chen, James J., Kodell, Ralph L., 2003. Comparison of Methods for Estimating the Number of True Null Hypotheses in Multiplicity Testing. Journal of Biopharmaceutical Statistics 13 (4), 675–689.
Lewis, C., Tukey, J W., 2001. Improved multiple comparison procedures for controlling the false discovery rate. Unpublished manuscript.
Miller, R.G., 1981. Simultaneous Statistical Inference, second ed. Springer-Verlag, New York.
Schulz, Kenneth F., Grimes, David A., 2005. Multiplicity in Randomised Trials I: Endpoints and Treatments. The Lancet 365 (9470), 1591–1595.
Shaffer, J.P., 1999. A semi-Bayesian study of Duncan's Bayesian multiple comparison procedure. Journal of Statistical Planning and Inference 82, 197–213.
Tukey, J.W., 1994/1953. The problem of multiple comparisons. In: Braun, H.I. (Ed.), The Collected Works of John W. Tukey, Multiple Comparisons, 1948–1983, vol. VIII. Chapman & Hall, New York, pp. 1–300.
Williams, V.S.L., Jones, L.V., Tukey, J.W., 1999. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. Journal of Educational and Behavioral Statistics 24, 42–69.

Hysteria: History and Critiques

Santosh K Chaturvedi, National Institute of Mental Health and Neurosciences, Bangalore, India
Soumya Parameshwaran, Kasturba Medical College, Mangalore, India

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R.E. Kendell, volume 10, pp. 7133–7138, © 2001, Elsevier Ltd.

Abstract

Hysteria, as a disorder, has been noted since ancient times, has been a condition difficult to understand, explain, and manage. As the ages have gone by, different explanations have been propounded, the most popular being the psychoanalytical explanations. However, the term hysteria went into disrepute and has been replaced by conversion and dissociation symptoms, which unfortunately, are rather inadequate. Hysteria is also considered a form of abnormal illness behavior, and current biological research is exploring the pathophysiology underlying hysteria. The varied presentations, course and outcome, make hysteria a rather charmingly difficult condition to treat. The revisions of the current diagnostic systems, International Classification of Diseases of the WHO, and the fifth edition of Diagnostic and Statistical Manual of mental disorders are likely to propose some clinically relevant modifications in the diagnostic criteria of what was once considered as Hysteria.

Hysteria is one of the enchanting and mysterious disorders of mind, noted since centuries, poorly understood and equally difficult to treat. Due to many debates and controversies, related to its name and phenomenology, the disorder has been rechristened many times. Hence, one does not often come across the term 'hysteria' in current clinical practice. The term 'hysteria' is derived from the Greek word *hystera* meaning womb, reflecting the belief of ancient Egyptian physicians that the phenomena of the disease were due to 'wandering of the womb.' Hippocrates and Galen both accepted this assumption and until the late nineteenth century, hysteria remained a disease of women and of the body (generally the nervous system). A general practitioner working in the nineteenth century described two features, which for a long time have been regarded as central to the diagnosis of hysteria. The first was that hysteria arose as a result of the effect of excess emotion on the nervous system and second, once afflicted, sufferers often stood to benefit from their ongoing illness, a concept now known as 'secondary gain' (Carter, 1853).

Hysteria became a common diagnosis, especially in young women, and often considered to be due to hereditary weakness and cultural degeneration. It was also clearly distinguished from hypochondriasis, which was thought to be more common in men, having a less sudden onset and being accompanied by anxiety and multiple somatic complaints (Crimlisk and Ron, 1999).

The idea that it was a purely psychological disorder dates from Breuer and Freud's *Studies on Hysteria* in 1896 (Breuer and Freud, 1896), which led to the psychoanalytic concept of hysteria. The demise of hysteria, in the West, was due primarily to the fact that the term came to be used in a variety of ways that were only tenuously related, namely:

1. neurological conversion symptoms;
2. psychophysiological reactions, including pain and visceral symptoms;
3. dissociative symptoms, fugues, and amnesic episodes;
4. outbreaks of mass, or epidemic, hysteria;
5. Briquet's syndrome, multiple, unexplained, somatic symptoms, and persistent dysfunction;
6. anxiety hysteria, an obsolete psychoanalytic term for phobic anxiety; and
7. a personality type assumed to be particularly susceptible to develop hysterical symptoms.

Initial Concepts of Hysteria

The concept of hysteria originated in the second millennium BC. Egyptian priest physicians attributed the paroxysmal symptoms of their female patients to wandering of the womb, and treated them either by the local application of fragrant potions to entice the uterus back to its proper station, or by the oral administration of nauseous draughts to drive it back. The European literature of the last 400 years is summarized by Merskey (1979). Thomas Sydenham was probably the first to recognize explicitly the role of emotion in the genesis of hysteria, for in 1697 he said in his *Discourse* that "The *External Causes* of this Disease are either violent motion of the Body, or more frequently vehement Agitations of the Mind by the Assaults of Grief, Anger, Fear and such like Passions." Sydenham was, of course, allowing for 'internal causes' as well, but the identification of these remained elusive. Sir Benjamin Brodie established in the 1830s that the muscles and peripheral nerves of patients with hysterical paralyses or anesthesias appeared quite normal when examined postmortem, and by the 1870s Sir James Paget was convinced that the problem was one of volition rather that of neurological pathology. Even so, the great French neurologists Charcot and Janet still assumed at the end of that century that the fundamental cause of the suggestibility and hypnotizability of hysterics was a hereditary nervous degeneration. According to the psychoanalytic theory that developed from this revolutionary treatise, hysteria was a specific clinical entity arising from sexual conflicts originating in the oedipal phase of psychosexual development. Despite the lack of supporting empirical evidence, and the wartime evidence that widespread and florid hysterical symptoms were precipitated by overwhelming stress rather than by infantile sexual experiences or fantasies, this psychoanalytic concept of hysteria remained dominant until the 1960s.

Alternative Terms for Hysteria and Its Nosological Status

For over a century, physicians have been puzzled by the varied manifestations of the condition, dismayed by its ambiguities, and challenged by the behavior associated with it. The term is now fading into history, mainly due to changes in the nosological systems, though the various syndromes it embraced still exist and still await a satisfactory explanation. The English physician Skey commented as long ago as 1867 that "One good reason that may be assigned to the persistent employment of the term Hysteria, is the difficulty of finding a substitute for it." Those objections and that difficulty have persisted ever since. The most consistent criticisms of hysteria have centered on the multiple meanings of the term and the wide variety of symptoms and behaviors to which it has been applied.

It has been reported that in clinical practice, neurologists use a wide variety of terms for diagnosing patients lacking a neurological basis for their symptoms, like: 'functional,' 'psychogenic,' and 'hysteria.' However, when talking informally, terms 'neurotic,' 'malingering,' and 'supratentorial' become more common, whereas, the terms 'somatoform' and 'conversion,' which are among the preferred official terms, were used by less than 30% of respondents in a survey (Mace and Trimble, 1991).

Due to the criticisms, the term hysteria had been discarded but most of its constituent syndromes survived, however, though with revised titles. Conversion symptoms became conversion disorder and anxiety hysteria became phobic disorder, and both were shorn of their original psychodynamic explanations. Dissociative reactions likewise became dissociative disorders. Briquet's syndrome was renamed somatization disorder, and the hysterical personality was renamed histrionic personality disorder. Mass or epidemic hysteria was not mentioned, presumably because it was assumed that each individual would be categorized separately under the appropriate syndrome. The arguments for abandoning the term 'conversion disorder' include

1. It is based on a questionable psychoanalytical concept that is only one of several competing unproven aetiological hypotheses.
2. It is not widely used by nonpsychiatrists and some assume it is not liked by patients.
3. The disorder should be viewed from a brain as well as a mind perspective, while avoiding an unhelpful brain–mind dichotomy.
4. The current 'conversion' diagnostic criteria require a psychosocial association with symptom onset, which is difficult to find in a minority of patients, and the exclusion of feigning (malingering) which is difficult, if not impossible (Reynolds, 2012).

Diagnosis and Classification

Diagnostic systems have increasingly discouraged the use of the term 'hysteria' because of its ambiguity and current systems have dropped the term altogether. The fourth edition of Diagnostic and Statistical Manual of mental disorders (DSM-IV) classified Conversion disorder, which includes motor and sensory symptoms as well as nonepileptic seizures, under somatoform disorders, whereas dissociative symptoms, such as fugue and multiple personality, come under a separate category, dissociative disorders. The International Classification of Diseases (ICD-10) has combined these two categories under neurotic, stress-related, and somatoform disorders (Crimlisk and Ron, 1999).

Conversion disorder (DSM-IV) is diagnosed when the patient presents with symptoms referred to the central nervous system for which no cause can be found despite adequate investigations, and which are presumably caused by psychological factors. In ICD-10, conversion disorder is grouped with the dissociative disorders. Dissociative disorders are classified according to the mental faculty or psychological function that is impaired. While DSM-IV provides a separate category for dissociative identity disorder and relegates trance and possession disorders to the 'not otherwise specified' category, the converse holds true in ICD-10.

Finally, depersonalization/derealization syndrome, is included as a dissociative disorder in DSM-IV (Chaturvedi and Rajkumar, 2009). Currently, the classifications of psychiatric disorders in DSM-IV and ICD-10 are both under review. In current classifications, the traditional use of the term 'hysteria,' has been replaced by 'conversion disorder,' influenced by the psychoanalytic theory of Freud, and by 'dissociative disorder,' influenced by a related analytical theory of Janet. Within a review of somatoform disorders for DSM-5, proposals have been circulated to replace the term 'conversion disorder' with 'functional neurological disorder.' Similar suggestions are being discussed for ICD-11 (Reynolds, 2012). In DSM-5, dissociative identity disorder will include reference to possession as well as identity fragmentation, to make the disorder more applicable to culturally diverse situations. Dissociative amnesia will include dissociative fugue as a subtype. A dissociative subtype of posttraumatic stress disorder (PTSD), defined by the presence of depersonalization or derealization in addition to other PTSD symptoms, is being recommended, based upon new epidemiological and neuroimaging evidence linking it to an early life history of adversity and a combination of frontal activation and limbic inhibition (Spiegel et al., 2013). Conversion disorder (functional neurological symptom disorder) is likely to remain with the somatic symptom disorders, despite considerable dissociative comorbidity. The ICD-11 is in the process of revising somatoform disorders as Bodily distress disorders, and the dissociative disorders may be subtyped as – somatoform (conversion, bodily loss of function) or psychoform (for cognitive and consciousness involvement) dissociative disorders.

Clinical Presentations

Neurological Conversion Symptoms

Muscular weakness or paralysis, affecting one, two, or all four limbs, or resulting in loss of speech; convulsions simulating epilepsy; or a loss of sensation (commonly blindness or tunnel vision, deafness, or a circumscribed loss of cutaneous sensation) assumed to depend on the psychoanalytic mechanism of conversion. As the extent of the paralysis or sensory loss reflects the patient's ideas rather than anatomical facts, and as it can

often be demonstrated neurologically that the relevant motor or sensory pathways are undamaged, the diagnosis of hysteria can usually be made with some confidence.

Pain and Other Bodily Symptoms

Visceral symptoms like vomiting or otherwise unexplained pain assumed to be produced by the same conversion mechanism. Unlike neurological conversion symptoms, there is usually no easy way of demonstrating that such symptoms must be psychogenic and as a result the diagnosis of hysteria remains uncertain.

Dissociative Symptoms

Episodes of amnesia, fugue states, and rare cases of multiple personality assumed, after Janet, to be based on a restriction or dissociation of consciousness. Like conversion symptoms, hysterical amnesias and fugues usually develop in highly stressful situations and have the effect of providing at least a temporary respite from that stress.

Epidemic or Mass Hysteria

Since the 'dancing manias' of fourteenth-century Europe, over 100 outbreaks of mass hysteria have been recorded, usually in schools, convents, or other communities of girls or young women, and usually at times of obvious stress (Sirois, 1974). Shared apprehensions and suggestion are clearly important as those affected are always in close contact with one another and develop almost identical symptoms. Influential members of the community are generally also the first to be affected. Mass hysteria is still observed off and on in traditional societies, in Asia and Africa, affecting groups of school children or entire village being possessed by spirits, demons, or devils.

Briquet's Syndrome

This well-defined pattern of behavior, first described by the French physician Briquet in 1859 and seen almost exclusively in young women. They repeatedly consult doctors with multiple and constantly changing somatic symptoms for which no adequate cause can be found. They also have unsatisfactory sexual lives and are prone to develop mood disturbances and conversion symptoms. This behavior persists into middle age or beyond, and as a result such women acquire voluminous medical files, have multiple abdominal and pelvic operations, and often become dependent on analgesics or tranquillizers (Guze and Perley, 1963).

The Hysterical Personality

Hysterical personality was regarded as an almost exclusively feminine condition, which was assumed to be particularly susceptible to develop conversion and other hysterical symptoms. The characteristic features are attention-seeking behavior, shallow, labile emotions, and self-centered, demanding, interpersonal behavior. A variety of tactics are used to attract and retain the attention of others, including striking or sexually provocative clothing, flirtatious behavior, flattery, dramatized accounts of their past lives or current circumstances, and extravagant gestures.

Neurobiology/Pathophysiological Mechanisms

Immunological Mechanisms

The pathophysiology of conversion symptoms indicate that women are generally at higher risk for a variety of autoimmune disorders. Recent research in psychoneuroimmunology suggests that the symptoms of conversion and somatoform disorder may be explained by the central effects of inflammatory and immune mediators that are released in response to stress. (Dimsdale and Dantzer, 2007). This model is attractive because (1) it explains the association between stress and the genesis of bodily symptoms and (2) it may account for gender differences in somatization, as women's hormonal and neurochemical milieu may make them more susceptible to develop abnormal immune activation in response to stress (Chaturvedi and Rajkumar, 2009).

Sex Hormone

The high occurrence of somatic symptoms during the premenstrual phase, pregnancy, and the menopause and the preliminary evidence that some hormonally mediated conditions, such as primary dysmenorrhea and secondary amenorrhea, are associated with psychological variables and somatization are evidences to support relation between sex hormones and somatoform or dissociative symptoms (Bitzer, 2003).

Neural Circuitry

Studies looking at cerebral activation in women with somatization showed decreased metabolism in both caudate nuclei, as well as in the left putamen and right precentral gyrus (Hakala et al., 2002). Another study found hypoperfusion in a majority of subjects, with preponderance in the right hemisphere; and right cerebellar hypoperfusion was more frequent in women than men (Garcia-Campayo et al., 2001).

Studies in conversion disorder have also demonstrated decreased activity of the primary motor or sensory cortex, and increased activation of the right orbitofrontal and cingulate cortices, as well as altered basal ganglia function(Vuilleumier et al., 2001). As all these structures are involved in sensorimotor function and are interconnected with the limbic system, further investigations would help in understanding their role in the pathophysiology of somatization, as well as the neural basis of gender differences. A role of the corpus callosum has been hypothesized, based on studies showing a larger anterior corpus callosum in subjects with high hypnotisability, a status linked to conversion (Nash, 2005).

Laterality

The observation that left-sided conversion symptoms appear to be more common has led to the idea that the nondominant hemisphere may mediate 'unconscious' processes. A functional hemispheric disconnection could, therefore, result in left-sided somatic symptoms and abnormal expression of affect (la belle indifference) in away reminiscent of the presence of anosognosia in patients with right hemisphere lesions.

Neuroendocrine Dysregulation

Preliminary evidence suggests that patients with somatization disorder, dissociative disorder, and chronic unexplained pain have dysregulation of the hypothalamic-pituitary-adrenal axis, specifically in the form of elevated cortisol levels (Allen et al., 2001).

Complex Integrative Models

Models that attempt to explain conversion and dissociation have been proposed in recent years, using various perspectives. Certain common factors noted are

1. The importance of the orbitofrontal and cingulate cortex, areas involved in the organization of behavior.
2. The role of stressful and traumatic events in early childhood.
3. The effect of these events on neurodevelopment, leading to deficits in emotional and behavioral regulation.

It seems that whatever the neural substrates of these disorders might be, early life experiences and adversities play a central role in their pathogenesis. Somatoform and dissociative disorders in women represent, par excellence, the value of a biopsychosocial approach in understanding the etiology, pathogenesis, clinical presentation, and treatment of a disorder.

Abnormal Illness Behavior

Contemporary understanding of hysterical behavior has been influenced strongly by the sociological concepts of the 'sick role' and 'illness behavior' and by learning theory, and attempts to reduce both the attractions of the sick role and influences discouraging healthy behavior now have a central role in management. This conceptual model explains the distribution of hysterical behaviors in populations, including the predominance in young women, and many other clinical observations, but it fails to account either for neurological conversion symptoms or for fugues and other dissociative phenomena. The concepts of the 'sick role' and 'illness behavior' originally formulated by Parsons (1951) and Mechanic (1962) provide the starting point for a better understanding of the problem (Pilowsky, 1969). Most illness behavior is generated by recognized disease, but substantial parts of the whole are generated either by fear of disease or by the 'positive reinforcement' provided by the rewards of the invalid role. The sick role becomes attractive only in situations in which their responsibilities are abnormally heavy (soldiers in battle, for example), or when they are in trouble (impending exams, debt, the threat of blackmail or prosecution, and so on), or when the prospect of financial gain increases the attractions of illness still further (so-called compensation neurosis). The very high incidence of hysterical phenomena in exhausted combat troops suggests that everyone is potentially capable of developing hysterical symptoms, just as they are capable of developing epilepsy, but vary in their liability to do so. This variation presumably depends partly on constitutional variables like suggestibility and capacity to dissociate, partly on past experience, and partly on the subject's current social situation.

Relevance of Hysterical Behaviors

Some of the behaviors considered diagnostic features of hysteria and conversion disorders have been found to lack the diagnostic specificity (Lazare, 1981). Secondary gain, once a key feature, has been noted in other medical disorders as well. Similarly, attention-seeking behavior is more a personality trait, nonspecific to any illness. Persons who do not receive due attention to their symptoms and suffering, try to draw attention of their health professionals. *La belle indifference* was considered diagnostic of hysteria for long, but was found to be a nonspecific feature, and a way of coping and response to stress. *La belle indifference* could also be considered as a dissociative phenomenon. There was a lack of evidence that symbolism could explain the hysterical symptoms, since the psychodynamic formulations became less popular. As mentioned above, histrionic personality does not always lead to hysterical symptoms, and persons with this personality could develop other disorders, and persons with other personality traits could develop conversion symptoms (Sharma and Chaturvedi, 1995).

Changing Epidemiology of Hysteria

Few community surveys of the incidence or prevalence of psychiatric syndromes have generated satisfactory measures of hysteria, because the range of phenomena embraced by the term is so vague and variable, and because most hysterical phenomena are short-lived. The incidence of conversion hysteria is falling both in the industrial and the developing world (in India there is epidemiological evidence of this (Nandi et al., 2000)) is probably that in both settings levels of education and sophistication are rising and life becoming less stressful. However, hysteria continues to be prevalent in developing and traditional societies, though the nomenclature has changed (Chaturvedi et al., 2010; Nandi et al., 2000). The reason hysterical symptoms are seen mainly in adolescents and young adults is probably because the adoption of a dependent role comes most readily to those who have only recently forgone the privileges of childhood and are least experienced at coping with the demands of the adult world. The frequent coexistence of hysteria and organic disease is also readily explained, for those who are genuinely ill or handicapped – by epilepsy or multiple sclerosis, for example – are the very people likely to have the most difficulty coping with the demands of everyday life, and with the most vivid and recent awareness of the rewards of the sick role. Finally, the predominance of left-sided limb paralyses and anesthesias, first remarked upon by Briquet, is explained, for in a right-handed person this minimizes the handicap without reducing the social impact.

Choice of Symptoms and Outcome

According to psychoanalytic theory, the choice of symptom is determined by its symbolic meaning for the patient, and doubtless this is sometimes so. Usually, though, it is determined by cultural expectations and the subject's own experience of illness, in themselves or others. The nurse who

develops hysterical blindness, for example, had severe conjunctivitis as a child, and the young man who loses the use of his legs had a grandfather who was housebound for years after a stroke.

Just as the development of hysterical symptoms is encouraged strongly by situations in which the balance of the advantages and disadvantages of sickness outweigh those of health, so recovery is encouraged by a reversal of that balance of advantage. This is why the symptom usually fades away if the situation that provoked its appearance resolves itself. If that situation persists, however, or if the inadequacies of the subject are paramount, the advantages and disadvantages of the chosen symptom become important. This is why disabling symptoms, like blindness or inability to stand or walk, have a comparatively good prognosis, whereas those which can be switched on or off as occasion demands, like fits or vomiting, or cause little disability, like tremors, are more intractable.

Investigation and Treatment

There are two implications for the investigation of patients with suspected hysterical symptoms, both well known. The first is that medical attention, admission to hospital, and investigations may easily potentiate the illness behavior. For this reason it is important that the investigations needed to make sure that the patient's symptoms are indeed hysterical are completed as quickly as possible, and not repeated without good reason. The second is that it is immaterial whether or not the patient is aware that he or she is in some sense feigning illness, because the treatment is the same either way. Although the dilemma usually was avoided by assuming unconscious motivation, the traditional medical or psychoanalytic approach should require a distinction between conscious motivation (which is malingering, and so subject to strong disapproval) and unconscious motivation (which is an illness – hysteria – and so worthy of sympathy and treatment) in spite of the fact that it is generally impossible to distinguish the two. Indeed, the patient's degree of awareness often varies from time to time, so that what is initially unconscious becomes conscious, and vice versa.

There are always two components to the management of hysterical symptoms: an attempt to form a psychotherapeutic relationship with the patient, and an assessment of the elements of their social environment, which are encouraging or discouraging invalid behavior. The former is never straightforward, and inexperienced therapists often allow themselves to be manipulated by flattery or coquetry, or end up angrily rejecting the patient because they resent being manipulated, or indeed do both these things in succession. Occasionally it is possible by gentle degrees to lead the patient into a discussion of the emotional predicament from which their symptoms arose, and of the relationship between that predicament and their symptoms. More often there is a sustained denial of emotional difficulties of any kind, and it becomes increasingly clear that the subject is more interested in being a patient than in being cured, and curiously uninterested in the origins of their symptoms.

Conclusions

Although this contemporary social and behavioral view of hysterical behaviors explains many otherwise puzzling clinical observations, it provides no explanation of the mechanisms of conversion and dissociation. The explanatory theories proposed originally by Freud and Janet over 100 years ago have been abandoned, but nothing has yet been offered in their place and the neurophysiology of both phenomena remains a mystery. Indeed, shorn of the original psychoanalytic assumption that unconscious anxiety was being converted into a neurological deficit, the term conversion disorder has lost all meaning. And hysteria itself has been discarded. But as Lewis (1975) observed a generation ago, "a tough old word like hysteria dies very hard." It may yet come back.

See also: Anxiety and Anxiety Disorders; Anxiety and Fear Conditioning, Neural Basis of; Culture as a Determinant of Mental Health; Freud, Sigmund (1856–1939); Psychoanalysis, History of; Somatoform Disorders.

Bibliography

Allen, L.A., Gara, M.A., Escobar, J.I., et al., 2001. Somatization: a debilitating syndrome in primary care. Psychosomatics 42, 63–67.

Breuer, J., Freud, S., 1896. Studies on hysteria. In: Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 2. Hogarth Press, London.

Bitzer, J., 2003. Somatization disorders in obstetrics and gynecology. Archives of Women's Mental Health 6, 99–107.

Carter, R.B., 1853. On the Pathology and Treatment of Hysteria. Churchill, London.

Chaturvedi, S.K., Rajkumar, R.P., 2009. Somatisation and dissociation. In: Chandra, Prabha, Herrman, Helen, Fisher, Jane, Kastrup, Marianne, Niaz, Unaiza, Randon, M.B., Okasha, Ahmed (Eds.), Contemporary Topics in Women's Mental Health – Global Perspectives in a Changing Society. Wiley Publishers, London, pp. 65–96.

Chaturvedi, S.K., Desai, G., Shaligram, D., 2010. Dissociative disorders in a psychiatry institute in India – a selected review and patterns over a decade. International Journal of Social Psychiatry 56 (5), 533–539.

Crimlisk, H.L., Ron, M.A., 1999. Conversion hysteria: history. Diagnostic Issues, and Clinical Practice Cognitive Neuropsychiatry 4 (3), 165–180.

Dimsdale, J.E., Dantzer, R., 2007. A biological substrate for somatoform disorders: importance of pathophysiology. Psychosomatic Medicine 69, 850–854.

Garcia-Campayo, J., Sanz-Carrillo, C., Baringo, T., Ceballos, C., 2001. SPECT scan in somatization disorder patients: an exploratory study of eleven cases. The Australian and New Zealand Journal of Psychiatry 35, 359–363.

Guze, S.B., Perley, M.J., 1963. Observations on the natural history of hysteria. American Journal of Psychiatry 119, 960–965.

Hakala, M., Karlsson, H., Ruotsalainen, M., et al., 2002. Severe somatization in women is associated with altered cerebral glucose metabolism. Psychological Medicine 32, 1379–1385.

Lazare, A., 1981. Current concepts in psychiatry: conversion symptoms. New England Journal of Medicine 305, 745–748.

Lewis, A., 1975. Survival of hysteria. Psychological Medicine 5, 9–13.

Mechanic, D., 1962. The concept of illness behavior. Journal of Chronic Diseases 15, 189–194.

Merskey, H., 1979. The Analysis of Hysteria. Bailliere Tindall, London.

Mace, C.J., Trimble, M.R., 1991. "Hysteria", "functional'' or "psychogenic"? A survey of British neurologists' preferences. Journal of the Royal Society of Medicine 84, 471–475.

Nandi, D.N., Banerjee, G., Mukherjee, S.P., Ghosh, A., Nandi, P.S., Nandi, S., 2000. Psychiatric morbidity of a rural Indian community over 20 years. British Journal of Psychiatry 176, 351–356.

Nash, M.R., 2005. Salient findings: a potentially groundbreaking study on the neuroscience of hypnotizability, a critical review of hypnosis' efficacy, and the neurophysiology of conversion disorder. Journal of Clinical and Experimental Hypnosis 53, 87–93.

Parsons, T., 1951. The Social System. Free Press, New York.

Pilowsky, I., 1969. Abnormal illness behavior. British Journal of Medical Psychology 42, 347–351.

Reynolds, E.H., 2012. Contribution to classification issues Hysteria, conversion and functional disorders: a neurological. The British Journal of Psychiatry 201, 253–254.

Sirois, F., 1974. Epidemic hysteria. Acta Psychiatrica Scandinavica 51, 7–44.

Sharma, P., Chaturvedi, S.K., 1995. Conversion disorders revisited. Acta Psychiatrica Scandinavica 92, 301–304.

Spiegel, D., Lewis-Fernández, R., Lanius, R., Vermetten, E., Simeon, D., Friedman, M., March 2013. Dissociative disorders in DSM-5. Annual Review of Clinical Psychology 28 (9), 299–326. http://dx.doi.org/10.1146/annurev-clinpsy-050212-185531. Epub 1 February 2013.

Vuilleumier, P., Chicherio, C., Assal, F., et al., 2001. Functional neuroanatomical correlates of hysterical sensorimotor loss. Brain 124, 1077–1090.

Iconicity in Linguistics

John Haiman, Macalester College, St Paul, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Iconicity is a constant factor in generating linguistic 'pantomimes' of thought. Its never less constant enemies are the limitations imposed by the medium of sound itself, and the working of ritualization: routinization (or grammaticalization in the broad sense, which includes both analogy/systematization and sound change/erosion) and decorative bulking.

In spite of sound change, systematization by analogy and systematization via ritualization, iconicity survives both in linguistic form and as a productive force in language change. In fact some of the most fundamental rules and principles in language are iconically motivated. This is in spite of the majority view of students of language today: the doctrine of arbitrariness, that anything can stand for anything. Not only are words arbitrary: the universal of double articulation ensures that, at the phonological level, individual signs are associated with no meaning whatsoever. At the morphosyntactic level, a distinction may be made between the lexicon and grammar. Challenges to the doctrine of arbitrariness thus fall into three groups.

At the phonological level, the notion of sound symbolism is such a challenge. Although submorphemic constant correlations between sound and meaning are marginal in languages like English, it has been argued that they are much more pervasive in many languages of Africa and Southeast Asia. At the word level, the corresponding challenge is the notion of onomatopoeia, which has been acknowledged to be marginal in every language. In the artificial registers of poetry, the sound may indeed be made to be an echo to the sense, but this is not the case in flat-footed prose. At the level of compounds and more complex syntagms, however, diagrammatic iconicity in the sense of Peirce (1932) has always been both pervasive and acknowledged. Among the staples of linguistic theory which have been strongly iconic are

1. The doctrine of semantic compositionality or Saussure's motivation (Saussure, 1916): "The meaning of a complex expression is in some way the sum of the meanings of its parts."
2. Behaghel's first law (Behaghel, 1932: 4): "Conceptual closeness of ideas is reflected in physical closeness of their expression." One of the clearest demonstrations of this is in Bolinger and Gerstman's (1957) discussion of the contrasts between phrases like 'light housekeeper' and 'lighthouse keeper,' in which, as they demonstrated, the perceptual phonetic contrast mirrors the orthographical distinction. Other work in the same vein is Greenberg (1966b) and Bybee (1985) – restatements and elaborations of the traditional insight that derivational affixes are closer to the root than are inflectional ones; Givon (1980) – the degree of embeddedness of a complement clause is a reflection of the degree of control exerted by the subject of the containing clause; and Haiman (1985a), which surveys contrasts between direct and indirect causation, between alienable and inalienable possession, and the semantic consequences of coordination reduction.
3. Related to physical distance is the idea that the degree of physical incorporation of a string into a larger one corresponds to the degree of its conceptual incorporation in the material into the framework represented by the matrix string. Direct quotation resists incorporation into a matrix sentence, indirect quotation does not. So-called 'relevance conditionals' (e.g., "If you're hungry, there's food in the fridge") resist incorporation within a complex sentence, while ordinary conditionals do not: the former do not 'count' as constituents in V/2 languages like German (König and van der Auwera, 1988).
4. "The same form is used for the same meaning." This is the basis of grammatical analogy (Anttila, 1989: 89). The somewhat less familiar corollary, that there are no true synonyms, or that different form is a reflection of different content, is the basis for Bréal's laws of repartition and specialization (Bréal, 1897: 26), and underlies much recent work in descriptive syntax. The 'same is same' heuristic also lies behind all the (never entirely successful) attempts to construct a 'notional' definition for parts of speech and grammatical categories in general (e.g., Jakobson, 1936; Wierzbicka, 1985), as well as for attempts to find explanations for recurrent polysemy of grammatical constructions.
5. "More form reflects more meaning." This is still the most widespread diagnostic for markedness (see Greenberg, 1966a: 26–27). The clearest demonstration in typological research in Benveniste (1946) on the most frequent

expression of the third person singular. As a creative dynamic principle of perception and interpretation, the 'more is more' heuristic is identified as what has led to the frequently attested restructuring of verbal paradigms (e.g., Watkins, 1962). It is probably also what has led to the excrescence of derivational morphology in some less widely attested cases. In Khmer and in Romanian, for example, a reduced pronunciation of a verb was reinterpreted as the verb, the full form being reinterpreted as a derived nominalization or a derived causative (Haiman, 2003).

6. The order of clauses (e.g., in "I came, I saw, I conquered") corresponds to the order of events, other things being equal (Jakobson, 1965). Elaborations on this *post hoc* (*conceptually*), *ergo*, *post hoc* (*linguistically*) heuristic are to be found in the Prague School notion of 'communicative dynamism,' itself a derivative of Behaghel's recognition (1932: 4) in his second law, that the given precedes the new; Greenberg (1966b) on the order of mention of conjuncts, and of the order of protasis and apodosis; and Tai (1985) on the order of constituents within the clause. Hawkins (1994) and others have proposed that Behaghel's second law (that short constituents precede longer ones, as in Malkiel's 'irreversible binomials' (Malkiel, 1959)) may also be reducible to this principle, given the tendency to give reduced expression to what is more familiar or given (frequently old) information.
7. Conceptual symmetry is reflected in formal symmetry in both syntagms and paradigms. In a syntagm, coordinate conjunctions, reciprocals, comparisons, and correlative constructions like 'the more, the merrier' tend to be symmetrically expressed in most languages (Haiman, 1985a: 72–102). More subtly, conceptual symmetry is sometimes expressed by paradigmatic symmetry. The failure of reflexivization in sentences like "I like me" and "I want ME to stay behind" is a reflection of the loss of 'speaker privilege,' and hence, indirectly, a direct reflection of a divided self (Haiman, 1998: 77): 'me' is treated no differently from (i.e., symmetrically with) any other participant on the perceptual stage. So, too, conceptual asymmetry may be reflected in formal asymmetry. Langacker (1985) argues that the structure "there's snow all around (sc. me)" directly reflects the subject's inability to see himself, and thus reflects the asymmetry between the perceiving subject and the perceived object directly.

In spite of the opacity and arbitrariness in much of language, there is a sense in which we do not believe in our bones that it is real. In his short story 'Funes the memorious,' Borges showed what a totally arbitrary code would be:

> In place of 7,013, he would say (for example) *Maximo Perez*; in place of 7,014, *The Railroad*; other numbers were Luis Melian Lafinur, Olimar, suphur, the reins, the whale, the gas, the cauldron, Napoleon, Agustin de Vedia. In place of 500, he would say *nine* …
>
> Borges, 1964: 64

All students of language, however, committed they may be to the idea that 'anything can stand for anything,' will probably find this conception of a language to be strikingly alien. What Borges's imaginative and outrageous reduction forces us to acknowledge is the extreme unlikeliness of such willed original opacity. However opaque the signs and combinations of language may seem to us in any mature grammatical system, this system probably owes its present form to an originally iconic code like a pantomime (Donald, 1991; Fonagy, 2000), modified over the course of time.

Given the acknowledged universality of motivation, analogy, markedness, Behaghel's law, and the extreme implausibility of truly arbitrary 'languages' like the number system of Funes, the interesting question is where does arbitrariness arise? Several factors have been pointed out.

One obscuring factor is that limitations of the medium lead to competing (albeit impeccably iconic) motivations. For example, 'same form' is motivated not only by 'same meaning,' but by symmetry or 'lack of speaker privilege.' 'More form' is motivated not only by 'more meaning,' but also by 'less frequency' (Greenberg, 1966a, on the relative marking of dual and plural), 'less familiarity,' greater difficulty of recovery (Givon, 1983, on the formal bulk of nominal expressions which refer back to more or less distant antecedents), greater surprise (Haspelmath, 1993, on the relative marking of causativity versus intransitivity for inchoative verbs like *freeze* and *break*; Haiman, 1985a: 168–174; Haiman, 1998: Chapter 5, on the relative marking of coreference), and 'more politeness' ('more form' is 'more wrapping,' cf Geertz, 1960; Haiman, 1985a: 151–155). 'First mention' is motivated not only by being 'most familiar,' but also by Jespersen's (1949: 54) "principle of actuality: what is at the moment uppermost in the speaker's mind tends to be first expressed." Similarly 'element incorporation' may be iconically motivated (as is the case for indirect quotation) or economically motivated (as is the case for object incorporation of typical, or frequently occurring objects). Simple repetition may have as many as half a dozen familiar motivations (Haiman, 2014). Overburdened with too many semantic distinctions, a form may come to seem more or less arbitrary (Bolinger, 1975: 110). The first letters of the Library of Congress book cataloging system have absolute values, but the second letters, whose meanings are determined by a large number of context-sensitive rules, are thereby rendered meaningless (Haiman, 1998: 149–150). So far, the most promising demonstration of the reality of iconicity in spite of this limitation is that where there are no competing motivations for the 'limited good' of a particular formal expression, there tends to be agreement (Haiman, 1985a: Chapter 6).

Opacity or the apparent 'autonomy of syntax' can also arise as the result of a number of diachronic factors: erosion, routinization, systematization, and ritualization are perhaps the most important. We can illustrate some of these tendencies in a number system whose iconic origins are still very much in evidence – the Roman numerals.

The first three numbers are of course as iconic as a tally. The numeral V is apparently opaque, but still barely recognizable as a diagram of a human hand. The motivation for the abbreviation was probably no different from the motivation for the sound changes of truncation or assimilation: it is faster and easier to write V than to draw the hand with all five digits. The numeral X can still be seen as a combination of two V's, one on top of the other. The numbers C and M are also icons, but of a different sort: C (entum) and M (ille) are examples of

another kind of abbreviation, and illustrate competing iconic motivations coexisting within the same code.

Implicit in even the simplest compound sign, like II, is a rule which seems so natural that it is hardly recognized at first. It is that the combination XY is to be read as 'X' + 'Y.' This is as true of VI, VII, VIII, …, XXXIII, …, MMCCCVIII, as it is of II. At some point, an unknown genius recognized and applied analogical reasoning to this natural rule: if XY is read as 'X' + 'Y,' then perhaps, under some (very restricted) conditions YX could be read as 'X' − 'Y.' The result, apparent in numerals like IV, IX, XC, and the like, is motivated within the system. But these numbers, even when they are drawn in the most painstaking and redundant detail, can no longer correspond iconically to the quantities they represent. Analogy has led in this case to something that is the precise opposite of diagrammatic iconicity. The signs within a system cease to reflect reality and take on a life of their own when they are manipulated by rules internal to the system, as well as by the reality they still attempt to represent.

Within language, aspects of systematization are recognized as

1. the substitution of value for signification within a paradigm (the meaning of signs lies not in their relationship to things in the world, but in their relationship to each other, cf Saussure, 1916);
2. the analogical extension of fixed structures beyond the contexts for which they were originally motivated;
3. the existence of various kinds of context sensitivity within a syntagm (Haiman, 1998: 149–150);
4. digitization, the conversion of analogical modeling to 'all-or-nothing' representation.

A compact example of systematization in syntax is the stress pattern of thetic sentences such as 'TRUMAN died' (Lambrecht, 1988) in languages like English. A stress pattern should iconically reflect the contrast between old and new information, and a word order pattern should do the same. In English, iconic stress has led to sentences such as 'Truman DIED' where the predicate is new relative to the subject, but routinization has led to a fixed SVO word order, which is emancipated from the motivation of functional sentence perspective. Given that low stress … HIGH stress means 'predicate is new,' perhaps (by an analogical leap similar to the invention of the Roman numeral IV) HIGH stress … low stress means not only 'subject is new' (which it still iconically can do), but also 'everything is new.'

The Roman numeral system is far more iconic than the Arabic system, but it is already fully abstract and digitized. A more primitive system, corresponding to those languages in which there are no numerals higher than some 'highest digit' (frequently 2), resorts to subjective and analogic or iconic means of modeling higher numbers, typically (one, two, …, many). More languages resort to this analogical modeling in the domain of fractions and present series like (one, a half, …, a bit). The digitization process 'colonizes' the analogic domain, substituting fixed quantities for impressionistic assessments like 'many' and 'a bit.' A roughly analogous 'digitalizing colonization' process cannot be observed in the lexicon, or in morphosyntax, but it is observed in the conventionalization of intonation (Léon, 1993; Fonagy, 1971): from being a purely analogical 'more is more' system, it becomes grammaticalized and transformed into an 'all-or-nothing' system – although never completely so (Bolinger, 1961, 1986).

The familiar tension between iconicity and routinization (Haiman, 2011) is apparent in the English auxiliary system, which includes not only the modals (perfective) *have*, and (progressive and passive) *be*, but also auxiliaries like have to, be supposed to, be about to, be going to, just about, keep on, used to, would rather, had better, get to, and an indeterminate number of others. To the extent that these may co-occur, the iconic order of auxiliaries is such that each includes within its scope all of the following (auxiliary) verbs. Thus, *get to try to VP* does not mean the same as *try to get to VP* and the meaning differences in all such cases reliably obey the iconic scope rule.

What distinguishes the familiar English Chomskyan syntagm *Modal* + *Perfective* + *Progressive* + *Passive* from all other combinations of auxiliary verbs is that this order is fixed and thus not subject to scope iconicity. It is only this distinction which separates a class of what could be (falsely) called 'true (Chomskyan) auxiliaries' from all the other complement taking verbs which – irrespective of the nature of the complementizer which they assign to the following verb – occur in scope–iconic order. And even within this set, iconicity is partially maintained: both of the passive auxiliaries *be* and *get* can occur only immediately before the main verb which is within their scope. (Another putative criterion for true auxiliaryhood, the complementary distribution of the Modals, *have*, and *be* with the quasi-auxiliary verb *do* is incoherent, as is well known. By this criterion, perfective *have* is an auxiliary while obligative *have*(*to*) is not, and *be* in all of its functions, whether as a (main verb) copula, the progressive or passive auxiliary, or as part of an auxiliary as in *be going to* is an auxiliary. Only in the imperative and prohibitive, where *do* may or must co-occur with the copula, is the complementary distribution between *do* and strictly auxiliary uses of the verb *be* enforced, cf Haiman, 2010).

Grammaticalization corresponds to both routinization and systematization. Ritualization subsumes both of these factors, which tend to reduce linguistic expression, but also includes stylization, which tends to exaggerate them (in language as in nonlinguistic ritual). Decorative doubling, exemplified in symmetrical compounds like *last and final*, is one example of this semantically unmotivated (and hence noniconic) bulking. Of marginal importance in many Western languages, it is a striking feature of Khmer and other languages of Southeast Asia and elsewhere (Haiman, 2010; Williams, 2013).

See also: Language and Poetic Structure; Ritual and Symbolism, Archaeology of; Ritual; Semantics.

Bibliography

Anttila, R., 1989. Historical and Comparative Linguistics, second ed. Benjamins, Amsterdam.

Behaghel, O., 1932. Deutsche Syntax, Bd. IV. Carl Winter, Heidelberg.

Benveniste, E., 1946. Relations de personne dans le verbe. Bulletin de la Société de Linguistique 43, 1–12.

Bolinger, D., 1961. Generality, Gradience, and the All-or-None. Mouton, The Hague, Netherlands.

Bolinger, D., 1975. Aspects of Language, second ed. Harcourt Brace Jovanovich, New York.
Bolinger, D., 1986. Intonation and Its Parts. Stanford University Press, Stanford, CA.
Bolinger, D., Gerstman, L., 1957. Disjuncture as a clue to constructs. Word 13, 246–255.
Borges, J.L., 1964. Funes the memorious. In: Labyrinths. New Directions, New York, pp. 59–66.
Bréal, M., 1897. Essai de sémantique. Payot, Paris.
Bybee, J., 1985. In: Haiman (Ed.), Diagrammatic Iconicity in Stem-Inflection Relations, pp. 11–47.
Donald, M., 1991. Origins of the Modern Mind. Harvard University Press, Cambridge, MA.
Fonagy, I., 1971. Double coding in speech. Semiotica 3, 189–222.
Fonagy, I., 2000. Languages within Language: An Evolutive Approach. Benjamins, Amsterdam.
Geertz, C., 1960. The Religion of Java. Free Press, Glencoe, IL.
Givon, T., 1980. The binding hierarchy and the typology of complements. Studies in Language 4, 333–377.
Givon, T. (Ed.), 1983. Topic Continuity in Discourse. Benjamins, Amsterdam.
Greenberg, J., 1966a. Universals of Language, with Particular Reference to Feature Hierarchies. Mouton, The Hague, Netherlands.
Greenberg, J., 1966b. Universals of language, with particular reference to the order of meaningful constituents. In: Greenberg, J. (Ed.), Universals of Language, second ed. MIT Press, Cambridge, MA, pp. 73–113.
Haiman, J., 1985a. Natural Syntax. Cambridge University Press, Cambridge, UK.
Haiman, J. (Ed.), 1985b. Iconicity in Syntax. Benjamins, Amsterdam.
Haiman, J., 1998. Talk Is Cheap. Oxford University Press, New York.
Haiman, J., 2003. Explaining infixation. In: Polinsky, M., Moore, J. (Eds.), The Nature of Explanation. Center for the Study of Language and Information, Stanford, pp. 105–120.
Haiman, J., 2010. The creation of new words. Linguistics 48, 547–572.
Haiman, J., 2011. Iconicity versus grammaticalization: a case study. In: Narrog, Heine (Eds.), The Oxford Handbook of Grammaticalization. Oxford University Press, Oxford.
Haiman, J., 2014. Six motives for repetition. In: Whinney, Mac, et al. (Eds.), Competing Motivations. Oxford University Press.
Haspelmath, M., 1993. More on the typology of inchoative/causative verb alternations. In: Comrie, B., Polinsky, M. (Eds.), Causatives and Transitivity. Benjamins, Amsterdam, pp, 87–20.
Hawkins, J., 1994. A Performance Theory of Constituency and Word Order. Cambridge University Press, Cambridge, UK.
Jakobson, R., 1936. Beitrag zur allgemeinen Kasuslehre. In: Jakobson, R. (Ed.), Selected Writings, vol. 2. The Hague, M ounton, pp. 23–72.
Jakobson, R., 1965. Quest for the essence of language. Diogenes 51, 21–37.
Jespersen, O., 1949. A Modern English Grammar on Historical Principles, vol. VII. George Allen and Unwin, London.
König, E., van der Auwera, J., 1988. Clause integration in German and Dutch conditionals, concessive conditionals, and concessive. In: Haiman, J., Thompson, S. (Eds.), Clause Combining in Grammar and Discourse. Benjamins, Amsterdam, pp. 101–133.
Lambrecht, K., 1988. Information Structure and Sentence Form. Cambridge University Press, Cambridge, UK.
Langacker, R., 1985. Observations and speculations on subjectivity. In: Haiman, J. (Ed.), Iconicity in Syntax. Benjamins, Amsterdam, pp. 109–149.
Léon, P., 1993. Traité de phonostylistique. Nathan, Paris.
Malkiel, Y., 1959. Studies in irreversible binomials. Lingua 8, 113–160.
Peirce, C., 1932. Philosophical Writings, vol. 2. Harvard University Press, Cambridge, MA.
Saussure, F.de, 1916. Cours de linguistique Générale. Payot, Paris.
Tai, J., 1985. Temporal sequence and Chinese word order. In: Haiman, J. (Ed.), Iconicity in Syntax. Benjamins, Amsterdam, pp. 49–72.
Watkins, C., 1962. Indo-European Origins of the Celtic Verb. Institute of Advanced Study, Dublin, Ireland.
Wierzbicka, A., 1985. Oats and wheat. In: Haiman, J. (Ed.), Iconicity in Syntax. Benjamins, Amsterdam, pp. 311–342.
Williams, J., 2013. The Aesthetics of Grammar. Cambridge University Press, Cambridge, UK.

Idealization, Abstraction, and Ideal Types

Steve Clarke, Charles Sturt University, Sydney, NSW, Australia; and The University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Abstraction and idealization are mental processes that are employed in order to simplify and organize empirical information to enable the use of laws and models to explain that information. Ideal types are a form of idealization that have been particularly important in the methodology of the social sciences and that are associated most closely with Max Weber.

Abstraction

Abstraction is a psychological action in which usually one, but possibly more, aspect of a thing or process, or a group of things or processes, are mentally separated out from that thing or process, or group of things or processes. All intelligent thought about the sensible empirical world can be understood as being abstractive. The empirical world throws up an ever-changing mélange of details that invades our senses. To render this datum intelligible, inconvenient details and transitory changes need to be abstracted away from, and simple, stable generalizations located. The successful identification of natural and social kinds would seem to require the ability to perform abstractions. Individual police officers, who instantiate the social, kind police officer, vary considerably in personal appearance and in other forms of possible identification. In order to reliably identify an individual as a police officer, the stable and enduring underlying properties that are common to the kind police officer need to be identified.

Because the process of abstraction involves the disregarding of particular features of things or processes in an attempt to develop general concepts, abstractions are sometimes conflated with universals. However, there is no requirement that a successful abstraction need leads to the development of a concept that identifies a universal. A successful abstraction may lead to the identification of relatively invariant generalizations about a kind of thing or process that fall well short of being universally applicable. The universal is conventionally contrasted with the particular, whereas the opposite of the abstract is usually held to be the concrete.

Not everyone would accept that abstraction is as pervasive in the mental lives of individuals as has been suggested here. Plato can be interpreted as denying that the intelligibility of the sensible requires abstraction. It may appear that the mind is abstracting from the experience of concrete entities to the identification of underlying general and abstract kinds. However, Plato would argue that it is actually noticing resemblances between particular entities and the entities found within a world of perfect forms, which it was familiar with before this life. Berkeley and Hume downplay the role of abstractions in everyday thought. They object to an account of abstraction provided by Locke, holding that what cannot be separated in reality cannot be separated in the mind either (Baxter, 1997). Berkeley and Hume see no need for abstraction from the empirical, in order to render the world intelligible, because they do not allow that the sensible is distinct from the intelligible.

Whether or not Berkeley and Hume are right to reject portrayals of abstraction as being all pervasive in thought, it is clear that the process of abstraction plays a central role in the mental lives of individuals. The ability to apply the sophisticated apparatus of modern mathematics and logic to actual situations depends, inter alia, on the capacity to reason abstractly. In addition, the capacity for abstract thinking enables previously unfamiliar experiences to be rendered intelligible, by relating these to the familiar. The nineteenth century empiricist John Stuart Mill presents the example of 'South Sea Islanders,' whose only experience of four-legged animals is of hogs. The islanders are able to make sense of European talk of other four-legged animals by abstracting from the concrete example of hogs to the general concept of the quadruped (Mill, 1950, p. 299).

Idealization

Idealization is a psychological act that, while not as pervasive as abstraction in everyday thought, has been extremely important in modern scientific thinking. While abstraction only involves the mental separation of aspects of things or processes, from those things and processes, when these may not be so separable in reality, idealization, which is enabled by abstraction, involves the mental rearrangement of features of reality so as to assist in the development of explanations. A familiar idealization is the frictionless plane, used to exemplify mechanical explanations. To conceptualize the frictionless plane, real planes that have friction-causing surfaces are abstracted away from. However, frictionless planes are not merely abstractions. It is convenient to imagine things that have surfaces that are frictionless, even though it is generally accepted that all surfaces exert some friction on objects that they come into contact with, however slight. For the purposes of thought experiments, disbelief is suspended and the convenient idealized fiction of a rearranged and simplified reality, which satisfies our explanatory needs, is allowed, a reality in which a surface can come into contact with an object and exert no force of friction on that object whatsoever.

The application of scientific laws and models to real complex situations often requires considerable idealization. In order to practically apply Newtonian mechanics to predict the trajectory of a driven golf ball, the golf ball is idealized as a perfect sphere that is struck with a constant force in the direction in which it is intended to travel. Details of the shape and surface of the actual ball, which in reality will affect its trajectory, are conveniently

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63031-3

ignored. Other potentially relevant causal factors such as wind resistance and humidity are also ignored. The resulting idealized calculation of the golf ball's trajectory will, at best, be approximately true some of the time. The trajectory of a golf ball that is hooked or sliced will not be accurately predicted, nor will the trajectory of an irregularly shaped golf ball or a golf ball that is struck into a strong wind. The idealizations made are useful ones. The utilization of a simple model that exactly describes the motion of an ideal golf ball enables an approximate description of the motion of actual golf balls in a useful range of cases. An explanation of the golf ball's trajectory that did not employ idealizations and that involved an attempt to include all causally relevant factors would be horrendously complex, if it was possible to calculate at all.

Galileo initiated the scientific revolution, in part, by legitimizing use of the abstract language of mathematics to enable the application of idealized explanatory laws and models in the natural sciences. Galileo's *Dialogue Concerning two chief world systems, Ptolemaic and Copernican* (Galilei, 1953) is a polemical work in which a proponent of the new approach to science, which employs idealization, argues against an advocate of the Aristotelianism of Galileo's day that was opposed to idealization in the sciences. In response to the major Aristotelian objection to idealization, that the physical world is not a world of perfect geometrical shapes and therefore cannot be accurately described in geometrical terms, Salviati, Galileo's proponent of idealization, draws an analogy between the 'geometrical philosopher' and a merchant who allows for the weight of packaging materials when calculating the weight of his wares:

> *… when he wants to recognize in the concrete the effects which he has proved in the abstract, [he] must deduct the material hindrances, and if he is able to do so, I assure you that things are in no less agreement than arithmetical computations. The errors, then, lie not in the abstractness or concreteness, not in geometry or physics, but in a calculator who does not know how to make a true accounting.* (Galilei, 1953, pp. 207–8)

The recognition that idealization is endemic to scientific explanation raises problems for the confirmation of theories. An empirical test is rarely, if ever, a decisive test of a theory. Instead it is a test of a conjunction of the theory plus a series of idealizing assumptions. A falsification of this conjunction implies that either the theory is false or that one or more of those idealizing assumptions are to be rejected. Unless the magnitude of the combined effect of all the idealizing assumptions that stand between a theory and empirical reality can be identified, or unless a theory that accurately describes reality without the aid of idealizing assumptions can be constructed, an exact experiment that decisively confirms or disconfirms a theory cannot be performed.

Idealization in Economics

The social science in which idealization is most obviously employed, and most explicitly defended, is neoclassical economics. This is because neoclassical economists make idealizing assumptions that are plainly false and in need of defense from attacks due to critics who object to economist's reliance on false idealizing assumptions. Most, if not all, of the generalizations that form the corpus of neoclassical economics are true only in idealized situations where fully informed, fully rational agents, who are concerned only with maximizing expected utility, trade in conditions of unfettered competition. Economic reality is very different from this idealized situation. The prices of actual goods usually do not vary in the ways in which the neoclassical laws of supply and demand suggest that they should, and real humans do not appear to act so as to always maximize expected utility. Many classes of ordinary human actions bear no apparent relation to the maximization of expected utility, and some classes of ordinary human actions appear to be negatively correlated with expected utility.

The most influential justifications of the use of idealization in economics have been conducted on Galilean lines and appear as part of a defense of economics as being methodologically similar to the natural sciences. A historically influential defense of idealization in economics is presented in John Stuart Mill's *On the Definition of Political Economy; and on the Method of Investigation Proper to It*. According to Mill,

> *That which is true in the abstract, is always true in the concrete with proper allowances. When a certain cause really exists, and if left to itself would infallibly produce a certain effect, that same effect, modified by all the other concurrent causes, will correctly correspond to the result really produced.* (Mill, 1967, pp. 326–7)

Mill notes the unreality of Euclidean geometry for "… lines, angles, and figures, as human hands can construct" (Mill, 1967, p. 327), and draws an analogy to the situation of the 'political economist':

> *No mathematician ever thought that his definition of a line corresponded to an actual line. As little did any political economist ever imagine that real men had no object of desire but wealth, or none which would not give way to the slightest motive of a pecuniary kind. But they were justified in assuming this, for the purposes of their argument; because they had to do only with those parts of human conduct which have pecuniary advantage for their direct and principal object; and because, as no two individual cases are exactly alike, no general maxims could ever be laid down unless some of the circumstances of the particular case were left out of consideration.* (Mill, 1967, p. 327)

Mill's defense of idealization ties its use to a metaphysical picture of the world as being composed of causal tendencies that continue to be efficacious when they are not being interfered with by the activity of other causal tendencies (Hausman, 1992, pp. 33–53). This is a metaphysical picture that has been argued for in contemporary times by Cartwright (1989), whose Millian heritage is quite explicit.

While Mill argued for the general legitimacy of idealization, Marx (1968) went further and attempted to demonstrate the ways in which idealizational explanations in the social sciences could be rigorously tested, or at least this is how he has been interpreted by Leszek Nowak. According to Nowak, Marx should be recognized as 'the Galileo of the social sciences' because his application of Galilean idealization to the social sciences is the first instance of the use of rigorous scientific

reasoning in the social sciences (Nowak, 1980, p. 38). Nowak's Marx relates idealizational generalizations to real situations in which these can be tested, by identifying each interfering cause that affects the expression of an underlying factor represented by an idealizational law, and attempting to calculate the magnitude of the effect of each such interfering cause.

Nowak interprets Marx's presentation of the law of value – a law relating the value of a commodity to the amount of labor devoted to its production – as an attempt to show how this idealized law could be used to explain the actual price of that commodity, in particular economic circumstances. According to Nowak, Marx's law of value is strictly true only when adjustments are made for eight factors that can affect actual economic situations. For example, the law of value applies when the supply and demand of a commodity are exactly balanced, but ordinarily this is not the case and an imbalance between supply and demand will tend to distort the value of commodities. If the magnitudes of all factors that can distort the value of a commodity are identified, then actual market prices can be explained and decisive tests of the idealized law of value produced, or so Nowak's Marx would argue. Marx and Mill are only some of the better known defenders of idealizations in economics. More recent defenses have been provided by Lionel Robbins and Milton Friedman, among others (Hamminga and De Marchi, 1994).

Opposition to the use of idealized explanatory models and laws, in economics and other social sciences, is often tied to broader concerns. In particular, some opponents of the use of Galilean idealization in the social sciences attempt to draw a sharp distinction between the methods appropriate to the social sciences and those appropriate to the natural sciences. One influential way of pursuing this line of reasoning is to hold that naturalistic explanations are inappropriate in the social sciences where interpretive understanding is the appropriate methodological framework. On some such views, explanatory laws and models are out of place in the social sciences, and the fact that these are idealized only serves to demonstrate how unrealistic and inappropriate they are in the social sciences. The possibility of a sharp methodological divide between the natural and the social sciences has been attacked by methodological pluralists who want to recognize the potential legitimacy of a range of methodologies in both the natural and the social sciences (Bohman, 1991). If methodological pluralists are right, then there is no reason to think that idealized nomic explanations are generally inappropriate in the social sciences, although they may be inappropriate in some social sciences, or in some aspects of social science. The legitimacy of idealized nomic explanations in the social sciences is also defended by philosophical naturists, some of whom have provided explicit defenses of the utility of idealized explanatory laws in the social sciences (e.g., Kincaid, 1996; McIntyre, 1996).

A focused attack on the use of idealized explanations in neoclassical economics has been put forward by Hollis and Nell (1975), (pp. 43–6). These authors draw a distinction between the practical application of idealizing explanatory assumptions, such as the idealizing assumption of frictionless motion in physics and the practical application of idealizing explanatory assumptions in neoclassical economics, such as the idealizing assumption that adequate information about the market is available to all participants. In the case of frictionless motion, a clearly defined variable, the coefficient of friction, which can be accurately adjusted for, is assumed away. Actual participants in actual market places often have inadequate information about the market and as – in contrast to the case of frictionless motion – the effects of inadequate information are not well understood, it is practically impossible to adjust real data to precisely test explanatory laws and models that make the idealizing assumption of adequate information.

Hollis and Nell are right that the effects of inadequate information cannot be easily measured and adjusted for, whereas the effect of friction on the motion of objects traveling along a plane can. However, the simplicity of the frictionless plane is atypical of idealizations in the natural sciences, where approximations and adjustments for factors that are not fully understood are frequently made (Cartwright, 1983, 1989; Laymon, 1995). If it is accepted that idealizations are being inappropriately used whenever the effect of interfering factors cannot be quantified – an assumption that Hollis and Nell appear to make – then considerations of consistency suggest that the results of a very large proportion of the natural sciences ought to be rejected on the very same methodological grounds.

Ideal Types

Although it did not originate with him, the notion of an ideal type has come to be associated closely with Max Weber, and it occupies a central place in his thinking about social scientific methodology (Weber, 1949, 1968). Weberian ideal types are a form of idealization that Weber thought to be important in, and distinctive to, the social sciences. Ideal types in the social sciences play many of the explanatory roles that idealizations in the natural sciences play. They are used to identify and separate underlying causal factors that are obscured by the interplay of causes that contribute to the makeup of complex social environments. When messy, complex data about social environments is appropriately simplified by abstraction, useful ideal types can be identified that enable powerful explanatory generalizations to be developed.

Weber is at pains to ensure that ideal types not be understood as averages, or as direct attempts to form empirical social generalizations:

> *It is thus the "idea" of the historically given modern society, based on an exchange economy, which is developed for us by quite the same logical principles as are used in constructing the idea of the medieval "city economy" as a "genetic" concept. When we do this, we construct the concept "city economy" not as an average of the economic structures actually existing in all the cities observed but as an ideal type. An ideal type is formed by the one-sided accentuation of one or more points of view and by the synthesis of a great many diffuse, discrete, more or less present and occasionally absent concrete individual phenomena, which are arranged according to those one-sidedly emphasized viewpoints into a unified analytical construct. In its conceptual purity, this mental construct cannot be found empirically anywhere in reality.* (Weber, 1949, p. 90)

Like idealizations in the natural sciences, ideal types can perform explanatory work at a number of different levels of generality, and can do so in a number of different ways. Weber employs the concept of the ideal type in several different ways

(Burger, 1987, pp. 130–5). First there are 'individual ideal types' which can play a role in clarifying conceptions of specific historical phenomena. Even if there is no actual instance of the pure individual ideal type 'feudalism,' it aids historians in understanding feudal societies, and the era in which feudalism was prevalent, if they can see how and where particular societies approximated to the ideal type feudalism. Second, there are 'generic ideal types' which describe the underlying causal factors that influence many or all societies. Here Weber would include the idealizations that constitute the 'fundamental concepts' of economics (Weber, 1949, p. 100). A third role for ideal types is to describe developmental sequences relating successive historical epochs. Marx's postulated historical sequence leading to the development of a pure socialist society is an example of ideal types playing this role. In all these roles ideal types are not obviously different from idealizations in the natural sciences, which can also be specific, generic, or developmental. However, there is a fourth way in which ideal types are used in an explanation that appears to distinguish them from idealizations in the natural sciences, and this is when they play the role of 'essential' ideal types.

Essential ideal types encapsulate the ideas that are characteristic of the thinking of a particular group of people and that imbue their thought with meaning. If the pre-Raphaelite art movement of the nineteenth century is to be properly understood, for example, it will not be enough only to examine actual instances of pre-Raphaelite art. The ideal type of art that the pre-Raphaelites were striving to perfect will have to be understood. It may well be that no single pre-Raphaelite artwork completely succeeded in capturing the artistic ideals that the pre-Raphaelites aimed to achieve. Nevertheless, there is no hope of understanding the pre-Raphaelite art movement, and its causal interactions with other art movements and with the wider world, unless the expressive aims of its participants are understood. The same is true of broader and more complicated intellectual movements such as Christianity and Communism.

The ideal type is a form of idealization that enables causal explanations to be employed in the social sciences, and it can also be employed as an explanatory form that gives a distinctive place to subjective human interpretation. Because it brings together causal explanation and meaningful human interpretation, the ideal type's use by Weber can be understood as a means to find middle ground – and here Weber anticipates Davidson's subsequent work on reasons and causal explanation (Ringer, 1997, p. 172) – between methodological dualist interpretivists, such as Dilthey, who draw a sharp divide between natural scientific explanation and social scientific understanding, and methodological monist positivists and naturalists, such as Hempel, who typically seek to subsume social scientific concepts under natural scientific categories. The actual role that Weber's discussion of ideal types played in the history of the intense and polemical methodological debates in Germany early in the twentieth century – the 'Methodenstreit' – is much more complex than this story allows, however.

Although ideal types continue to be a subject of considerable historical and sociological interest (e.g., Burger, 1987; Ringer, 1997; Runciman, 1972), they have rarely been discussed in recent philosophy of science and social science (exceptions include Papineau, 1976 and Weinert, 1996). In part this is because philosophical discussion of ideal types has mostly been subsumed into discussion of idealizations and in part it is because idealization has not been a focus of a great deal of discussion in twentieth-century and early twenty-first-century philosophy.

Ongoing Disputes

While it will be conceded by most that there are particular situations where idealizations can be useful in the provision of approximately accurate descriptions, there remains much dispute about the range of the utility of idealizational explanations. If idealized laws and models are to be reliably used in scientific explanation, the ways in which actual situations deviate from ideal ones need to be accounted for. A temptingly simple way of dealing with this problem, which is explicit in Nowak's treatment of idealization, is to understand each idealization as a limit of a series of real cases and make adjustments accordingly (Nowak, 1980). Sometimes this approach is appropriate. The frictionless plane is a limit case of a series of actual planes that are successively less affected by friction. So it is a simple matter to correct for the effect of friction to precisely explain actual cases. However, scientists often have little idea of how to go about correcting for idealizations in order to accurately describe real situations (Cartwright, 1983, 1989; Laymon, 1995).

In the experimental sciences, the gap between idealizations and the reality they are intended to help describe can be reduced by designing more complex experiments that involve fewer idealizations to relate to reality. In addition, experimental situations that eliminate or minimize the effect of interfering causes can be set up, thereby reducing the need to idealize (Laymon, 1995). While these approaches will typically reduce the magnitude of the deviation of idealized explanatory laws and models from reality, there remain a number of problems with their application. First, scientists do not appear to construct complete lists of the causal factors that might interfere in experiments. Indeed, it is not clear that they could do this. Second, causal factors do not always interact in an additive way, so their combined effect resists simple calculation. A third problem is that there is an explanatory cost involved in the use of highly specific experiments. The explanations arrived at are applicable to a narrower band of cases. The cost of accuracy is the loss of generality.

In the social sciences there are moral and practical barriers to experimentation. However, idealizations and explanatory appeals to ideal types remain omnipresent in social science explanations. While the gap between idealized explanations and the real situations that they can be applied to can be reduced in the experimental sciences, there is often little hope of bridging the equivalent gap in the social sciences. Social scientists appear to be stuck with the choice of either admitting that their preferred idealizations may be wrong-headed or insisting, without apparent justification, that these must be accepted a priori. However, idealizations can go wrong (Cartwright, 1995), and it seems a mistake to try to disallow this possibility by appealing to a priori justification. Nowak's Marx attempted to abstract from the actual circumstances of nineteenth century capitalist economies to the law of value.

Neoclassical economists, who deny that economics should concern itself with the concept of value at all, would say that this is a paradigmatic example of idealization gone wrong. What guarantee do neoclassical economists have, or could have, to hold that their preferred idealizations stand on a surer footing?

A further subject of dispute centers on a metaphysical consequence of the prevalence of idealization in science, which has been raised against scientific realists by Nancy Cartwright (1983, 1989), reviving and developing the antirealist thinking of Pierre Duhem, the French physicist and philosopher who wrote at the end of the nineteenth century. Cartwright (1983) agrees with scientific realists that the fundamental laws of science provide the basis for explanatorily powerful generalizations but denies that this in itself is a good argument for their truth-aptness. The fact that fundamental laws are idealized means that they are protected from direct empirical testing by loose approximations, the use of open-ended *ceteris paribus* clauses, and the unrealistic assumption that causal factors are straightforwardly additive. Scientific realists such as Liu (2004) and Weisberg (2007) generally concede that these problems exist, but argue that nonfundamental laws in science, which are special instances of more fundamental laws, are more amenable to testing and can be better confirmed. However, Cartwright denies that actual nonfundamental laws in the natural sciences are literally to be understood as logical derivations of fundamental laws. In a careful study of the construction of amplifiers, Cartwright (1983, pp. 107–12) demonstrates how engineers approximate away from general theory in ad hoc ways so as to find descriptively accurate laws that have only the appearance of being logically derived from general laws.

Scientific realists can accede to all of Cartwright's points and hold out the hope of improving experimental techniques, thereby improving scientific explanations to the point where there is less room for doubt as to the veracity of fundamental laws. However, because of a lack of controlled experimentation in the social sciences, the social scientist who is a realist about particular grand theories in the social sciences is faced with the same problem without apparent hope of a solution. No progress is made toward demonstrating that a grand sociological or economic theory is true by showing how it is applied in a particular situation to provide a more specific theory, which is empirically well established, if the application of the grand theory to the particular situation turns out to depend on ad hoc modifications and is not an instance of logical derivation. However, few, if any, specific social scientific theories appear to be logically derived from grand theories.

See also: Causes and Laws: Philosophical Aspects; Mathematical Models in Philosophy of Science; Problem Selection in the Social Sciences: Methodology; Weber, Max (1864–1920).

Bibliography

Baxter, D.L., 1997. Abstraction, inseparability, and identity. Philosophy and Phenomenological Research 57, 307–330.
Bohman, J., 1991. New Philosophy of Social Science: Problems of Indeterminacy. MIT Press, Cambridge, MA.
Burger, T., 1987. Max Weber's Theory of Concept Formation, expanded ed. Duke University Press, Durham, NC.
Cartwright, N., 1983. How the Laws of Physics Lie. Oxford University Press, Oxford, UK.
Cartwright, N., 1989. Nature's Capacities and Their Measurement. Oxford University Press, Oxford, UK.
Cartwright, N., 1995. False idealization: a philosophical threat to scientific method. Philosophical Studies 77, 339–352.
Galilei, G., 1953. Dialogue concerning the Two Chief World Systems, Ptolemaic and Copernican. University of California Press, Berkeley, CA (trans. Drake S).
Hamminga, B., De Marchi, N., 1994. Idealization and the defence of economics: notes toward a history. Poznan Studies in the Philosophy of the Sciences and the Humanities 38, 11–40.
Hausman, D.M., 1992. Essays on Philosophy and Economic Methodology. Cambridge University Press, Cambridge, UK.
Hollis, M., Nell, E.J., 1975. Rational Economic Man: A Philosophical Critique of Neo-Classical Economics. Cambridge University Press, Cambridge, UK.
Kincaid, H., 1996. Philosophical Foundations of the Social Sciences. Cambridge University Press, Cambridge, UK.
Laymon, R., 1995. Experimentation and the legitimacy of idealization. Philosophical Studies 77, 353–375.
Liu, C., 2004. Laws and Models in a Theory of Idealization. Synthese 138, 363–385.
McIntyre, L.C., 1996. Laws and Explanation in the Social Sciences: Defending a Science of Human Behavior. Westview Press, Boulder, CO.
Marx, K., 1968. Capital, vols. I–III. Progress Publishers, Moscow.
Mill, J.S., 1950. Philosophy of Scientific Method. Hafner, New York.
Mill, J.S., 1967. On the definition of political economy; and on the method of investigation proper to it. In: Mill, J.S. (Ed.), Essays on Economics and Society. University of Toronto Press, Toronto, ON, pp. 309–339.
Nowak, L., 1980. The Structure of Idealization. D. Reidel, Dordrecht, The Netherlands.
Papineau, D., 1976. Ideal types and empirical theories. British Society for the Philosophy of Science 27, 137–146.
Ringer, F., 1997. Max Weber's Methodology: The Unification of the Cultural and Social Sciences. Harvard University Press, Cambridge, MA.
Runciman, W.G., 1972. A Critique of Max Weber's Philosophy of Social Science. Cambridge University Press, Cambridge, UK.
Weber, M., 1949. The Methodology of the Social Sciences. Free Press, Glencoe, IL.
Weber, M., 1968. Economy and Society: An Outline of Interpretive Sociology, vols. I–III. Bedminster Press, New York.
Weinert, F., 1996. Weber's ideal types as models in the social sciences. In: O'Hear, A. (Ed.), Verstehen and Humane Understanding. Cambridge University Press, Cambridge, UK, pp. 73–93.
Weisberg, M., 2007. Three kinds of idealization. The Journal of Philosophy 104, 639–659.

Identity and Identification

Akeel Bilgrami, Columbia University, New York, NY, USA

Published by Elsevier Ltd.

Abstract

A distinction is made between subjective and objective identities and each is explored in the rest of the article, with a view to elaborating its philosophical significance and its relevance to politics, culture, and psychology.

One initial step toward imposing some theoretical order on the notoriously haphazard concept of 'identity' in politics and culture is to distinguish at the outset between its 'subjective' and 'objective' aspects. When a person is said to have a certain identity owing to some characteristics he or she has, and with which he or she identifies, then identity is being thought of in its subjective aspects. If a person is said to have a certain identity owing to some characteristics he or she has, but with which he or she does not necessarily identify, then we are speaking of his or her objective identity.

Subjective Identity, First

Subjective identity is the notion of identity that is most immediately relevant to politics because people sometimes tend to allow themselves to be mobilized in the public arena on its basis. It is not – not by any means – that subjective identity is always mobilized in politics. It is more that, where it exists, it is poised so as to be mobilized if other conditions are present.

What is it to identify with some characteristic one possesses, thereby making it an identity-imparting characteristic in the subjective sense? This is a more complicated question than it might seem.

A first stab at answering it might be to say that someone identifies with a certain characteristic, if she values it. Thus one must value the fact that one belongs to a certain nationality or ethnic community or even a certain profession, if one can be said to have the identity (in the subjective sense) of an Indian or Korean-American, say, or a teacher and writer. That seems to be a minimal initial condition for identifying with it. But it is clearly not sufficient, because one may have values from which one is oneself alienated. This can be a fairly common phenomenon. To be alienated from one's values is structurally akin to being alienated from one's desires. Just as an alcoholic may be disgusted with his addictive desire for alcohol, so also someone may disapprove of his own patriotism or find his pride in his profession intolerably smug.

So a further condition has to be added to our minimal condition before a characteristic imparts subjective identity. One has to endorse one's valuing of that characteristic. That would presumably ensure that one is not alienated from its valuing. There is another reason why endorsement of this kind is necessary apart from the attempt to solve the problem of values one is alienated from. The topic of *subjective* identity is not merely about what one is but also about what one conceives of oneself to be. This idea, therefore, brings with it, in any case, the reflective endorsement of the relevant valued characteristics.

To endorse a value, it is often said, is to have a second-order value. Someone must value the fact that he or she values the characteristic of belonging to a certain nationality or profession, before he or she can be said to have the identity of an Indian or a teacher, in this subjective sense. (We should add that something like a second-order valuing may not be necessary in order to ensure that there is no alienation from one's own first-order values; all that may be needed is something negative: that there is no second-order disapproval of one's first-order value. However, as just said, the second-order level comes in a more positive form, in any case, because subjective identity is unavoidably about reflective matters such as what one takes oneself to be.)

With this second-order valuing in place, have we said something sufficient about subjective identity? Not yet, because one's second-order values can be highly neurotic, and when they are, they can be values that one is also alienated from. For example, someone may feel his second-order value, which disapproves of his first-order valuing of his role in his profession as being too smug, as itself being too prim, too censorious, and too much of a super-ego phenomenon.

What still further condition must now be added to give a sufficient account of subjective identity? One possibility is to conceive of it as requiring a receding hierarchy of orders of value. At each order, one is not alienated from a value if one endorses it at a higher order. So in our example, any neurosis regarding a person's second-order disapproval of his pride in his profession is ruled out, if he has a third-order approval of his second-order disapproval. This solution raises worries about an infinite regress.

In order to avoid such a regress, a second possibility is to conceive subjective identity not as emerging in a receding hierarchy but as requiring a coherence among one's values, no matter what order of value is being considered. So in our example, the second-order disapproval is neurotic, not because there is a lack of third-order approval of the second-order disapproval, but rather because the latter does not cohere with one's other values at all levels, including the first-order values. The second-order disapproval is something that one is not alienated from, something that one identifies with, only if it coheres well with the first-order values and other second-order values one has. Here no infinite regress threatens, but some philosophical account of coherence among values must be worked out to match the coherentist accounts of belief and

knowledge that we already have available with some degree of sophistication. It is worth noting that if this coherentist way of thinking of subjective identity is right, it is very closely tied to rationality in values, since the point of a coherentist account of value is presumably that it is an account of when one's values are rational. In a word, one identifies with one's values to the extent that the value is rational, in the sense of being fortified by coherence with our other values. The conclusion is attractive like any conclusion that allows two seemingly separate themes (rationality and subjective identity) to be united.

The trouble, however, is that if subjective identity is given by a rationality-imparting coherence among our values, then we will be identified with all the values we have that are rational in this sense. But that is of not much help with the idea of identity, because it follows from this that our identity (in the subjective sense) is never going to be anything very distinctive. It will pick out nothing very special or identity-imparting among all our coherent values, such as the ones we have been discussing: one's valuing the fact that one is an Indian, a teacher, a writer, a Muslim; no one of these will be more importantly relevant to one's identity than other characteristics of ours that we value coherently, such as one's weight, one's love of cricket or of dessert. Precisely what seemed attractive about this view is what makes it of no particular help on the subject of identity, with which we are concerned.

Perhaps this difficulty teaches a deflationary lesson. Perhaps it is a sign that there is something inflated about the very idea of identity, that our thinking there is something specially distinctive about some characteristics is misguided, that it is not something we should expect, and that we have come to expect it only because of the recent rise of 'identity' politics that has elevated some characteristics – nationality, ethnicity, linguistic and religious allegiances – beyond anything warranted by or echoed in the actual moral-psychological economies of ordinary citizens. There is some point to this qualm, a point in fact that needs much nuanced development in longer discussions of the subject. But it can also be a point that is too glibly made. For it does seem, at first sight anyway, to be quite accurately descriptive of at least a small, vocal, and influential body of citizens in many polities that they display a strong identification with these very characteristics and allow themselves to be mobilized on its basis. And of them, at least, some notion of identification must be given, which shows why these characteristics are valued more distinctively than the many others that are also coherently and rationally valued by them.

We are, therefore, still lacking a sufficient account of subjective identity. To repeat: what we need is not merely endorsement (of one's own valuing of some characteristic of ours) by some higher order values or by coherence with other values, but some further element that makes the value endorsed more central and distinctive in our psychological economies.

An obvious thought here might be to say that the values that are more central are those that are more *intense* than other values, especially since it does seem as if Muslims, Quebecois, Serbians, and so forth who seem most visible in identity politics value their Muslimness, and so forth, very intensely, more intensely than other things they value. It is a question, however, how theoretically useful it is to plonk down 'intensity' as a primitive and unanalyzable property that values have; but even apart from this problem, the thought is wrong, for in some cases of weakness of will, we act on values that are very intense but that we do not endorse.

A better thought at getting at the required further element is to say that the endorsement of the value must be such that it makes the value concerned, in some sense, more *unrevisable* (rather than more intense) relative to the other values one holds. For it is surely intuitive that a Muslim or Quebecois, whose identity is caught up with his valuing these characteristics of his, is less likely to give up valuing them than the other values he holds, or perhaps – a more subtle variation of the intuition, one that will be developed briefly below – is less likely to *conceive of himself* as giving up the value.

For this thought to be genuinely promising, we need (1) to show what sort of endorsement of a value makes the value relatively unrevisable and (2) to ground the idea of unrevisability in something that would not show the reluctance to revise to be irrational by the lights of the agent herself (in the way cases of weakness of will show that acting on or even holding some of our most intense values is irrational by the agent's own lights).

The task identified as (1) in the preceding paragraph is the most significant: which among all the values that equally cohere with one another are the more unrevisable? As we said, it is only if we answer this question that we would have captured what is distinctively identity-imparting about valuing being a Muslim or Quebecois, for the sort of agent we are concerned with. Here is a way of bringing out why that should be so. Any answer to the question would have an analogous effect to what Quine intended when he argued that some beliefs or propositions are in the very center rather than further out toward the periphery in our physical theory. The idea behind this metaphor was intended to replace misguided traditional ideas of 'analytic' propositions. His point was that for the theory to be the theory it is, is given by the beliefs or propositions that are at the center of our physical theory (e.g., our belief in the transitivity of length), since they are more immune to revision than the ones further out on the periphery, which are more exposed to what Quine called the 'tribunal of experience.' If those central beliefs do get revised, then it is not clear whether the theory has been changed or whether we have changed the subject because the meanings have changed. So also, analogously, we might say that a person's identity is given by his or her relatively unrevisable values – however we characterize them – and if those are given up, then it is not clear whether it is a change in the ordinary sense where the overall identity remains constant, but a change in value takes place, or whether the overall identity itself is changed. This analogy, though inexact, is all the same roughly intuitive and reflects our ordinary talk, when in a fit of nationalist sentiment, we say things such as 'I will lose my sense of self, of who I am, if I betray my country' (or, as in Forster's British schoolboy morality, 'I will lose my sense of self, I will not recognize who I am, if I betray my friend'), the sorts of things we are not likely to say of other things we value. Compare these to, 'I will lose my sense of self, if I give up my love of desserts,' which because of its implausibility shows such first-order values to be more analogous to beliefs or propositions at the periphery for Quine. (These are mere examples intended to convey the structural point of the analogy intuitively. Of course, it is possible,

though perhaps not routine, that someone may value his sweet tooth in the way a stereotypical nationalist or British public schoolboy values country or friendship, that is, analogously to Quine's center of the web rather than the periphery. The possibility in no way spoils the analogy; it merely shows that identities might be eccentric or bizarre on occasion.)

So much for the significance of unrevisability. But task (1) above was the task of defining it. We are seeking to define a way of endorsing one's valuing of some characteristic one possesses, which shows that value to be relatively unrevisable compared to other values one holds. Because we have already seen that coherence cannot provide such a special way of endorsement, let us return to the idea of second-order valuing of a first-order value to explore the sort of endorsement needed. What do we need to add to a second-order value to make the first-order value endorsed (relatively) unrevisable in one's psychological economy? An example may help to make the question and an answer to it less abstract.

Let us take some of the more absolutist Muslims in Iran over the last two decades. They have often urged something that approximates unrevisability of their Islamic values. One way they have done so is to argue that Iran needs to protect itself – in fact not just others in Iran, but even they the absolutists themselves – should protect themselves against their own 'moral' weakening and corruption in the face of the inevitable spread of the pernicious values of modernity in general and the West in particular. And they have argued (like Ulysses did, anticipating the sirens) that this protection should be ensured by entrenching Islamic values so deeply now that were Muslims even to be so weakened, the social, political, and legal institutions would not make it easy for them to shed their Islamic ways of life. Such a form of endorsement of one's Islamic values vividly shows it to be more unrevisable (in a very special sense) than other values one has and endorses having. Consider that the endorsement takes a counterfactual form: we value something in a way that we want ourselves to be living by the value, even if we (counter to present fact) do not value it anymore. At the time of valuing it, then, such a value stands out as very distinctive. The sort of unrevisability here is quite special because it is not so much that the value (at the first-order level) itself is permanent or immutable, but even though we may revise things later and cease to value it, the fact is that at the time of valuing it in this way, one (at the second-order level) yearns for the value to be unrevisable and relatively permanent, unlike all the other values that we endorse in the more ordinary way. That surely makes it part of one's deepest self-conception, because one would so utterly disapprove of oneself if we did not have the value at some future point, that we now try and make sure that one's self, at that future point, will live according to the value. 'Identity-imparting,' 'self-constituting,' and so forth, seem apt descriptions for values held and endorsed with such deep commitment.

Task (2) remains. Not all values that are unrevisable in this way are rational. What needs to be added is that these unrevisable values must also cohere with our other values. Coherence of a value, conceived as unrevisable in this special sense, with one's other values, allows for the rationality of this form of unrevisability, because reluctance to revise could only be irrational if the value one does not revise is one that does not have rational support from one's other values.

Before closing the discussion of subjective identity, one important cautionary remark should be made to protect against a misunderstanding of the notion as just defined. One should not be put off by the specific example given above, to think that the sort of endorsement that generates identity in this way, is a sign of fanaticism or illiberalism, just because we have become used to thinking of Islamic absolutism to be fanatical and illiberal. That would be to allow substantive opinions to blind us to the merits of the theoretical analysis we have come to, that is to say, blind us to the structural feature of the endorsement of a value that generates subjective identity. After all, the special way we value our own basic constitutional rights reflects just such an endorsement of our own identity-generating values *as liberals*, since all it does is reflect the same structure as the Iranian absolutists in the earlier example. We elevate a very few of the many values we more or less coherently hold to fundamental constitutional 'rights' – for example, free speech – precisely because we want to protect our future selves from giving in to any weakening of those values, as when in the face of strong dislike of another's substantive views, we might in the future find ourselves wanting to censor him. What our elevating these values into fundamental rights does is express the fact that we now want that we cannot later censor someone we strongly disagree with if we later have indeed weakened enough to wish to censor him. A liberal's deepest self-conception, that is, a liberal's identity, and a Muslim's identity, therefore – whatever the differences between them on other matters may be – is given by the same counterfactual structures of endorsing and identifying with their cherished values that our theoretical analysis has proposed. Neither is any more fanatical than the other, at least in matters of subjective identity.

Objective Identity, Next

When we turn to the objective aspects of identity, identification on the part of the subject in question with the identity-imparting characteristic(s) is not a necessary condition. Thus, for instance, racial identities when they are thought to be given by characteristics of descent, can be thought of as objective in this sense. Chromosomal criteria for defining gender identity are similarly objective. But biological criteria are not the only criteria that are routinely invoked. Inter-subjective and social criteria are also much favored. Thus, for instance, Marxists often claim that one's identity is given by one's role in a particular economic formation in a given period of history, that is to say, one's class identity, as 'class' is defined by Marx.

Many oppose the purely biological ways of thinking of various kinds of identity, such as racial and gender identity, claiming that these identities are 'socially constructed' by the perceptions and attitudes of one's fellows, by the zeitgeist of a particular period, by the conceptual categories and social institutions at a given time. Foucault and those influenced by him have made much of this and Foucault himself gave detailed historical and social accounts of particular concepts and institutions in Europe as determining identities. In fact, it is interesting that Foucault and his followers claim that it is not only the biological and other scientific criteria that are caught up in social factors of this kind, but the subjective ones we

discussed in the last section as well. These too are shaped by conceptual and institutional formations far removed and hidden from the exercise of our reflective self-understanding, thereby showing the ideals of individual autonomy that we assume to reside in the idea of identification, to be illusory. This article will not take up these issues raised by Foucault's influence.

It will look briefly instead at the motivations for looking at objective factors of identity at all, over and above the subjective ones.

Many subjects may identify with some characteristic that is not what is most salient about them to others, and it is thought important by many political philosophers that it is these latter that may often define their identity, no matter what subjects may conceive themselves to be. A good example of this can be seen in Stalin's well-known definition of a nation, which stresses the importance of historical and economic criteria for national identity, with a view to providing a corrective to what were seen as somewhat premature and ungrounded subjective identifications of nationality found in many secessionist demands in different parts of the world. Here the motivation for objectivist criteria of identity is (at least implicitly) political.

But underlying these is a more interesting theoretical rationale that points to important issues of a more philosophical nature. The claim that agents may have a certain identity even if they do not take themselves to do so, implies that what one takes oneself to be can be mistaken, a kind of self-deception, or at least a self-myopia (which may not always involve the motivated element often associated with self-deception, but involves at least the idea that one may sometimes simply be too deep for oneself – where 'deep' is not intended as a bit of eulogy.)

It would be philosophically clarifying to make a distinction between two different sorts of appeal to objectivist identities that are said to be (possibly) hidden from a subject's own self-conception. (1) One claim, the weaker one, is that subjects often betray signs of a certain identity in much of their behavior, even if they do not endorse and identify with what is reflected in their behavior. (2) The other, stronger claim does not even require that something in the subjects' behavior reflects the identity given by the unendorsed characteristic; rather the characteristic and the identity is given by the deliverances of some (social, political, economic, or biological) theory regarding these subjects.

1. The weaker claim, not surprisingly, is less controversial, because it requires that the characteristics of a subject that are going to define his identity are something that he at least reveals in his behavior. The subject may not endorse them, he may not even acknowledge them, but if the only good explanation of his behavior is that he has those characteristics, and if those characteristics are salient compared to others, then some claim can be made regarding how they impart his identity. Within this view, the more extreme cases will be where the subject does not even acknowledge the characteristics as being revealed in the behavior. Many of the identities that surface in Freudian and psychoanalytic theories make much of this sort of case (Oedipal, narcissistic identities). The less extreme cases will be those where there *is* acknowledgment of the characteristics, but no endorsement of them on the part of the subject. These are likely to be more common. What may be called 'silent' identities, as in 'silent majorities,' often consist of subjects who are not self-identified with a certain pattern of behavior, but will not be in any particular state of denial (as they are in the more extreme cases) about whether their behavior reveals the characteristics they are seen to have. It is very likely, for example, that many ordinary Muslims in Iran and other Islamic countries, who do not identify with absolutist or fundamentalist Islam, may all the same admit that much in their behavior mutedly plays along with the Islamist elements in their societies.
2. The stronger claim very often appeals to biological criteria, but is most interesting when it does not. Because the biological criteria are in any case usually caught up with social factors (see the point made about them above during the brief discussion of 'social construction' of identity), they will be ignored here. Perhaps the most well known, well worked out, and widely discussed of the stronger objectivist version of identity, which is not biologically based, is due to Marx and those influenced by him. What makes for having a class identity, say, a proletarian identity, is not any kind of self-identification with the working class, not even any behavior that suggests certain unacknowledged or unendorsed allegiances to that class, but simply the objective fact of having a certain place and function in the relations of production during the modern capitalist period of economic history. What is remarkable and controversial about this view, more so than anything found in claim above, is that something regarding the self and its identity is being attributed, without any basis or manifestation required in the conscious or unconscious behavior of the selves or agents concerned. A working-class person who exhibits no proletarian consciousness nor any of the solidarity and forms of behavior appropriate to the class, and none of whose behavior reflects an unconscious manifestation of such solidarity or consciousness, is nevertheless said to have proletarian class identity, albeit with a 'false consciousness.' It is only because he has this identity that there can be cause to call such a subject's consciousness 'false.' It is false precisely because he fails to conceive himself aright, fails to see his deepest self, which is determined by objective historical and material relations.

It is such a view of self and identity (where self and self-conception can come so radically apart) that filled Isaiah Berlin with anxiety in his discussion of 'positive liberty,' because what it encourages is the idea that the achievement of *self-realization* of individual citizens – that is, the achievement of their own autonomy and liberty (in the positive sense) – is now left to states or to the 'vanguards' of political parties, which lay claim to greater understanding of what some subject's self really and objectively is. From such a viewpoint, according to those alarmed by it, there is no paradox in the expression 'forcing someone to be free.' (Lenin's 'vanguardism' was frank about denying any such paradox.)

Underlying political anxieties of this kind is a more philosophical issue, which is much discussed in contemporary moral psychology, the issue of external as opposed to internal

reasons. An internal reason is a reason for one to do, believe, or value something, which appeals to some other evaluative element in one's moral-psychological economy. An external reason makes no such appeal to an internal element, it requires only some objective fact that need not even be recognized by the subject for whom it provides a reason. Thus in the orthodox Marxist tradition, a proletarian, given his historically determined identity, has (an external) reason to be a revolutionary even if there is no element in his moral-psychological economy that values it. Berlin's (1969) anxieties about statist tyranny carried out in the name of self-realization and positive liberty, were thus implicitly and more deeply about the very idea of external reasons, even though he never quite articulated them as having that underlying target; however, it becomes very explicit in a denial of the cogency of the very idea of external reasons in a brilliant essay by Bernard Williams (a philosopher much influenced by Berlin), though the point is marred in that essay by a somewhat confused equation of internal reasons with a Humean notion of value and motivation.

This last set of points provides a good resting point for this article on identity, which has distinguished fundamentally between the subjective and objective aspects of the concept. To a considerable extent, which of these two aspects we emphasize in our study of the concept will be a matter of theoretical decision, a decision which, in turn, depends on non-arbitrary philosophical considerations having to do, as we have just seen, with themes at some distance from identity, such as autonomy and moral reasoning. In itself, this is to be expected, because self, freedom, and reason have been closely connected themes in philosophy ever since Kant, both in the analytical and the European traditions of the discipline. Although much more needs to be said in detail to make the links between these themes perspicuous and explicit, it is safe to say that the more inclined we are to be uneasy about the idea of 'external reasons,' the more likely we are to stress the subjective rather than the objective aspects of identity.

See also: Discourse and Indentity; Ethnic Identity and Ethnicity in Archaeology; Ethnic Identity, Psychology of; Identity Movements; Identity in Anthropology; Integration and Multiculturalism: Focus on Western Europe; Intentionality and Rationality; Personal Identity: Philosophical Aspects; Rationality in Society; Religion: Nationalism and Identity; Self and Identity Development During Adolescence across Cultures; Social Constructionism; Social Identity in Sociology.

Bibliography

Berlin, Isaiah, 1969. Two concepts of liberty. In: Four Essays on Liberty. Oxford University Press.
Bilgrami, Akeel, 1992. What is a Muslim. In: Appiah, A., Gates, H.L. (Eds.), Identities. University of Chicago Press.
Elster, Jon, 1979. Ulysses and the Sirens. Cambridge University Press.
Flanagan, Rorty, A. (Eds.), 1993. Identity, Character and Morality. MIT Press.
Foucault, Michel, 1972. Archaeology of Knowledge. Pantheon Books.
Frankfurt, Harry, 1988. The Importance of What We Care about. Cambridge University Press.
Freud, Sigmund, 1963. In: An Outline of Psychoanalysis. W.W. Norton.
Hacking, Ian, 1999. The Social Construction of What? Harvard University Press.
Lenin, V.I., 1975. What is to be done? In: Tucker, R. (Ed.), The Leninist Anthology New York.
Marx, Karl, 1978. The communist manifesto. In: Tucker, R. (Ed.), The Marx-Engels Reader. W.W. Norton.
Rovane, Carol, 1998. The Bounds of Agency. Princeton University Press.
Smith, Anthony, 1991. National Identity. University of Nevada Press.
Stalin, Joseph, 1953. Marxism and The National Question. Progress Publishers.
Taylor, Charles, 1989. Sources of the Self. Harvard University Press.
Williams, Bernard, 1981. Internal and external reasons. In: Moral Luck. Cambridge University Press.

Identity and Identification, Social Psychology of

Nick Hopkins, School of Psychology, University of Dundee, Dundee, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Bilgrami, volume 11, pp. 7148–7154, © 2001, Elsevier Ltd.

Abstract

Identities can be based on our social group memberships. This requires investigation of how people self-define in social categorical terms, how these self-categorizations are produced and reproduced in everyday social practice, and how such social identifications are psychologically and behaviorally consequential. Work addressing these issues is discussed. So too is work concerning the importance of having one's identifications recognized by others, the role of social identities in social and political change, and the ways in which identities may be constructed so as to organize collective action.

The term 'identity' features prominently in everyday and in academic use. It can refer to an individual's personal and idiosyncratic self-conceptualization, but also to a sense of self based on group memberships (e.g., one's national identity). Such group identities can give rise to feelings of pride and joy, and motivate solidarity. However, it is also easy to find evidence of the degree to which people will, in the name of their group, harm others: as Sen (2006) puts it, "identity can also kill – and kill with abandon" (p. 1). People's behavior in terms of their group identities has long fascinated researchers and social psychologists have sought to explain what makes such behavior (whether positive or negative) possible.

Identity and Identification

In some respects, the power of group identities to motivate behavior (sometimes at considerable personal cost) is puzzling: often the bases for membership are not at all apparent. Take national identity. Although a series of necessary and sufficient criteria for national belonging is easily proposed (e.g., a common ancestry, language, culture, etc.), comparative analysis reveals no tangible characteristics essential to the maintenance of national consciousness. Accordingly, many return to Emerson's observation that "the simplest statement that can be made about a nation is that it is a body of people who feel that they are a nation; and it may be that when all the fine-spun analysis is concluded this will be the ultimate statement as well" (1960: p. 102).

Such observations underline the relevance of social psychological theories of group behavior that emphasize the importance of considering how individuals conceptualize themselves in terms of social categories. Indeed, contemporary social psychological theory implies that rather than taking the existence of groups for granted, we need to explore how people self-define in social categorical terms, how these self-categorizations are produced and reproduced in everyday social practice, and how such social identifications are psychologically and behaviorally consequential.

The Social Psychology of Identification

The social psychological concept of social identity developed out of a sense of dissatisfaction with attempts to explain intergroup phenomena in terms of personal and interpersonal processes. Tajfel (1978) drew a distinction between individual-level and group-level processes, which was developed by Turner and elaborated in what became known as self-categorization theory (Turner et al., 1987). This tradition emphasizes the role of social categories in self and other construal. It argues that the self and others may be defined at different levels of abstraction. Sometimes this may be in terms of individual uniqueness. At other times it may be in social categorical terms (e.g., 'us' as Europeans, as Hindus, as supporters of Barcelona football club). This shift in the salience of individual-level and group-level self-definitions is psychologically important: defining oneself in terms of a social category makes group behavior possible and underlies the behavioral shift from individual to group behavior.

Central to Tajfel and Turner's approach is the idea the group behavior arises from our knowledge of our membership of social groups and the meaning that such memberships have for us, and they repeatedly emphasized that this involves processes that cannot be reduced to those involving individuals' personality characteristics. Specifically, group members' behavior is shaped by their social categorical representation of who they are in relation to others such that when one adopts a particular social categorical self-definition, one no longer thinks of oneself in terms of one's unique individuality, but thinks of oneself as a category exemplar. This entails conformity to one's cognitive representation of the beliefs, values, and norms stereotypically associated with that identity.

This logic applies to both the small face-to-face groups that have predominated within social psychological research and also to the mass social phenomena associated with large-scale social categorizations. With regards to the former, whereas previous social psychological theories of group behavior tended to see groups as aggregates of individuals (and accounted for group behavior in terms of the relations between such individuals), contemporary theory conceptualizes group membership in cognitive terms. With regards to the latter, once group behavior is seen to depend upon an act of identification with a social category rather than the prior existence of bonds between individuals, we can understand how what Anderson (1983) calls 'imagined communities' can exist and structure our sense of who we are and how we relate to others. Such identifications are not dependent upon interindividual relationships, nor are they in any sense primordial 'givens': they

too are socially produced self-categorizations. It is important to emphasize that there is no sense in which social identities are to be regarded as superficial. They are simultaneously social yet deeply felt.

Identity-Related Processes

The behavioral outcomes associated with identifying in terms of a social category are diverse. If there may be pride and joy, so there may be hate and derogation of the other, and many have assumed these to be inextricably linked such that the price to be paid for pride in the group is a degree of hostility toward those outside it. However, the diversity in the identity–behavior relationship shows this is simplistic: if some acts of social identification can result in hostility to an out-group, others do not. For example, in the UK there is no simple relationship between participants' level of social identification with the nation and prejudice toward immigrants and asylum seekers: rather, this relationship depends on how the contents of this national identity are defined (Pehrson et al., 2009). This implies that rather than seeking to uncover simple relationships between identifying in terms of a social group membership and particular behavioral outcomes (e.g., ethnocentricism), it is more appropriate to theorize the social psychological processes involved in adopting a social identification and then to explore how, in different contexts, these same processes may support and sustain quite different behaviors. Or to put it another way, rather than seeking generality in terms of the actions that flow from collective identities, generality may better be conceptualized in terms of the generic processes that allow group behavior of any sort (Reicher and Hopkins, 2001).

With regards to these generic processes the most basic is the concept of "depersonalisation" (Turner et al., 1987). This conveys the idea that when acting in terms of a social identification we no longer view ourselves and others in terms of our individual characteristics but see ourselves in terms of our shared social identification. Indeed, people's behavior is no longer guided by their idiosyncratic beliefs and dispositions but conforms to their cognitive representation of the values and norms associated with that collectivity. Moreover, to the degree that a common categorization becomes psychologically salient, others judged as sharing that social self-definition are regarded as one's fellows (even to the extent that we will regard ourselves and our fellows as interchangeable exemplars of that social category).

This has several consequences. To the degree one shares a social identification with others one expects to agree with them and is more likely to be influenced by them than those categorized as out-group (Turner, 1991). Moreover, to the degree that individuals share a common social identification they are more likely to respect each other, cooperate with each other, and help each other. Research also shows that the cognitive and relational changes associated with adopting a shared social identification with others allows those involved to align their goals and interests, and to coordinate their actions so as to enact their social identity. In addition to being psychologically rewarding and empowering, this can feed back into participants' identity commitments (Drury and Reicher, 2005).

The benefits of a shared social identification should not be underestimated. Much research shows the importance of social relationship for well-being, and social identity researchers have shown that one's social identifications contribute to well-being because they lead people to expect support from their fellow group members (Jetten et al., 2012). It has even been found that participating in mass collective events characterized by a shared social identification (such as a pilgrimage event) can bring benefits to well-being (Tewari et al., 2012) and that the emergence of a social identity in an emergency can improve resilience (Drury et al., 2009).

What then about how such processes shape intergroup relations? Tajfel argued group members make intergroup comparisons and seek to establish their group's positive distinctiveness in relation to other group identities. Some researchers have interpreted this implying the derogation of an out-group. However, this is misleading: intergroup differentiation is a process and not a specific behavior, and the behavioral outcomes observed depend upon the norms and values of the relevant identity (Reicher and Hopkins, 2001). Again, we are drawn to social, cultural, and political factors, and to how these shape people's understandings of themselves and others.

Social Identification in Context

As should be apparent by now, an account of identity that emphasizes the importance of addressing an individual's cognitive representation of themselves does not constitute a retreat to a psychologism in which what goes on in people's heads is privileged over what goes on in the social context or in public (Brubaker et al., 2004). Indeed, Brubaker, et al. argue that the promise of a cognitive approach to identity is that it "may help connect our analyses of what goes on in people's heads with our analysis of what goes on in public" (p. 46). For example, it draws our attention to the social practices that support and sustain particular self-definitions. It is through such group-making activities (e.g., the practices of "banal nationalism," Billig, 1995) that self-categories are produced and reproduced, and made psychologically relevant. It also draws attention to how participation in collective events may revivify social bonds and social identities.

Sensitivity to the contingency of social categorization on the social practices associated with any context also helps explain the variability in any individual's self-categorization. In different contexts, different self-categories and group memberships can be psychologically salient (for we can conceptualize ourselves in multiple ways). Doubtless individuals vary in the degree to which they invest in different categories. Yet, situational factors are important in determining the salience of any potential self-categorization (Turner et al., 1987). Indeed, unless there is some match between one's social identification and the context in which one must operate, that identification will have little value or utility in navigating that environment and guiding one's behavior. Moreover, as other people are elements of the social context, it follows that their behaviors are relevant to how one conceptualizes oneself. Take, for example, the degree to which others' behavior

validates one's self-conception. Identity certainly involves an individual's subjective identification with a group. However, if such an identification is to be meaningful and sustainable, it must be validated by others. This insight has a long history in sociological thinking, which routinely emphasizes the importance of the social performance of identity and the recognition of that performance by others. Indeed Jenkins (1996) explains "*social identity is never unilateral*" (p. 21, original emphasis) with the corollary being that to speak of someone as 'having an identity' it is important that others treat that person as having that identity. Being treated by others in terms of a category that is discrepant with one's own self-conception is aversive and people use social interactions to elicit feedback that sustains their self-conceptions (Swann, 1987). For example, people may signal their social identifications (e.g., through their clothing) so as to elicit others' recognition that they are to be taken as members of a particular group (rather than another) and such research shows that this recognition is important for the individuals' subjective experience of that social identification (Hopkins and Greenwood, 2013).

It should also be apparent that as any individual's identifications are multiple and variable, recognition requires sensitivity to the identity definitions that people themselves hold in any specific scenario. It is one thing to willingly identify in terms of a particular identity; it is another to have one's choice over one's identities limited by others. Sen (2006) illustrates this with the observation that while an individual's Jewishness may be of immense personal value and political significance (e.g., in collective action against anti-Semitism) "it would be a long-run victory of Nazism if the barbarities of the 1930s eliminated forever a Jewish person's freedom and ability to invoke any identity other than his or her Jewishness" (p. 8). Moreover, individuals may experience their various social identities as mutually constitutive such that others' failing to recognize one element of one's identity may have implications for one's experience of another. Thus, if one identifies as a British Muslim, others' orientation to just one element of one's self-definition, even if it is deeply valued, may be problematic. For example, others' orientation to one's Muslim identity may limit one's ability to speak and be heard in terms of one's Britishness (Hopkins and Blackwood, 2011). So too, the misrecognition of one's Britishness may be experienced as compromising one's understanding of one's religious identity (e.g., as an identity that can accommodate to different national contexts: Hopkins, 2011).

Identity Boundaries

A social categorical definition of the self and others implies the definition of group boundaries (such that some are defined as in-group and some as out-group). These boundaries vary according to self-category salience. Thus with one social identification (e.g., a supporter of a particular football club) others – e.g., fans of a rival team – will be out-group, but with another social identification (e.g., those who love football), that same rival team's fan may be judged in terms of this more inclusive self-category (and regarded as a fellow). Moreover, such variation in the inclusivity of one's salient social identification has implications: Levine et al. (2005) show that the help extended to the same person depends on whether they are included within the situationally salient social self-categorization or not.

Moreover, the boundaries to any social identification are the products of human imagination, collective opinion, and social practice. This is well illustrated with the example of the nation where it is common to distinguish between more exclusive 'ethnic' conceptions of belonging and more inclusive 'civic' forms of belonging. In different countries different weightings are given to ethnic and civic criteria. However, even within the same nation there is variation in the degree to which particular criteria are adopted and hence different understandings of what defines the nation and who can identify with the nation (Pehrson et al., 2009). Moreover, it is possible to manipulate people's conceptualizations of the boundaries to their self-defining social categorizations with implications for the range of people who may receive the benefits of in-group membership (e.g., being listened to, receiving needed help: Wakefield et al., 2011).

As such boundary definitions are so consequential, it should come as little surprise to find that they are sites of argument and dispute, and this again emphasizes the importance of considering the social, cultural, and political factors that shape people's understandings of themselves and how they relate to others. Consider, for example, the fate of Bulgaria's Jewish population in the era of the Nazi Holocaust. Whereas Jewish communities across Europe were delivered to the Nazi's concentration camps, Bulgarian Jews were not. Rather, Bulgaria's Jews were defined as Bulgarian nationals and many non-Jewish Bulgarians were drawn to large-scale demonstrations on their behalf. This does not mean that there were no Bulgarian anti-Semites. Rather it reveals something of the extent to which the criteria for national inclusion are the products of social and political argument. Whereas the advocates of deportation depicted Bulgaria's Jews as 'community aliens,' others countered that Bulgaria's Jews were members of the national in-group, sharing Bulgarian values (Reicher et al., 2006). Such counterarguments also maintained that to collude with Nazi ideology would threaten Bulgaria's reputation and national interests. In other words, whereas the prodeportation argument depicted Jews as a threat to national identity their opponents construed collusion with the Nazis as identity-threatening. Again, in-group pride is not to be equated with discrimination and violence against minorities: it can also be mobilized in their defense. Which it will be is dependent on the social and political context and which constructions of identity gain popular support.

Representations of Identity

Social identifications are not the product of a lone individual's imagination. What it means to be a Bulgarian, a Muslim, an academic, or whatever, is a social product. For example, it might be tempting to assume that religious identities are defined by their associated texts (e.g., the Bible, Qur'an). However, such texts are always read and interpreted in specific social and economic contexts with the corollary that the meanings of such identifications are contingent upon the social milieu. Yet, even within the same context, the reading of text is contested (Hopkins and Kahani-Hopkins, 2009) with

consequences for group members' understandings of how they should behave. Thus debates about the nature of a Muslim social identification have implications for how British Muslims should relate to non-Muslims in Britain and to non-British Muslims across the globe. Indeed, it is possible to see such identity constructions as organized to promote different forms of behavior. Thus in some constructions of identity, positive engagement with the non-Muslim social order is defined as identity-appropriate, yet in others it is defined as inappropriate, and as identity-threatening.

Such debate could be assumed to reveal a lack of a common social identification. However, unless there is investment in a common identification, debate and contestation have little logic: it is because of a common identification that debate exists, has value and is worth entering into as those with different views on how the community should develop advance different definitions of collective identity. Such debates do not always result in a consensus and the potential for schism is ever-present: yet even schism and its dynamics cannot be understood without reference to the concept of a common identification (Sani and Reicher, 1998). It is also important to note the scope to such debate. Assumptions about the immutability of identity suggest contestation to be limited to topics at the margins of group identity. Yet, debates over the meaning of an identity typically include reference to cultural figures, histories, and traditions that are familiar to all. The reason is simple: invoking such figures and histories in terms that warrant the forms of behavior that one wishes to promote is a powerful means to recruit other group members to one's cause. Indeed, the wider corollary of such insights is that group members' talk about what a social identification means (e.g., the values shared by those adopting a particular social identification) should not necessarily be understood as descriptions of a contemporary social reality. Rather, they can be conceptualized as constructions that are organized to mobilize the behavior that would bring a particular way of living and social order into being. That is, people's talk of identity can be understood as having a performative and constitutive dimension, and as organized to encourage the forms of action that would realize a particular vision of how the social world should be (Reicher and Hopkins, 2001).

Such identity definitions may involve the characterization of in-group identity. However, they may also involve characterization of others' social identifications and how they impact upon one's own group. For example, in India, Hindu nationalists have defined India as essentially Hindu and depicted Muslims as an alien and threatening other (e.g., labeling Muslims as 'Babur ki sanfan' or the children of Babur, the sixteenth-century Muslim leader who conquered lands in present day northern India). Moreover, postcolonial analysts have explored how such contemporary characterizations of Muslim and Hindu identities are only possible because British orientalist 'scholars' (often colonial administrators motivated by a concern to maintain colonial power through a strategy of 'divide and rule') created histories that implied Hindus and Muslims in India are monolithic communities with continuous but separate histories. Such colonial histories provide powerful ideological resources for the contemporary promotion of particular social categorizations and the organization of collective behavior in terms of communal categories.

Representing Identity

With regards to the issue of identity, 'representation' has two meanings. One refers to the way in which people come to understand the meaning of their group membership – i.e., who 'we' are and what 'we' value. The other refers to the issue of who can represent 'our' interests and speak on 'our' behalf. However, these two meanings of the term can be intertwined: the issue of who is best placed to represent us depends on the characterization of our collective identity. The logic to the social identity tradition is that group members seek to act out and conform to the position best exemplifying their collective identity, and that they will be drawn to the person who best encapsulates in-group identity. However, as identity is not fixed but is a site of argument and dispute, we can conceptualize leadership as a site of contestation in which would-be leaders work on the characterization of both the collective and themselves to produce a sense that the latter embodies the former (Haslam et al., 2011). This implies that analyses of leadership cannot only focus on the personal attributes of the leader. Rather we should consider what binds group members to each other and how would-be leaders actively create and maintain their 'followership' through defining themselves as embodying the identity of those they seek to lead (Haslam et al., 2011). If simplistic assumptions about the nature of identity – especially those that naturalize or essentialize identity – fundamentally limit our ability to theorize the construction of this relationship, it is important to note that they have rhetorical value. Indeed, the same features that render essentialized understandings of identity analytically weak render them politically potent. They imply boundaries that are natural and qualities that are given and fixed, and such a naturalization of the socially contingent can be an important ideological resource with which would-be leaders represent (depict) the collective, and themselves as best positioned to be its representative (speaking on its behalf).

Social Change

Just as we need to theorize how social identities are produced and reproduced through a range of social practices such that the social order is maintained, so too we need to theorize social change and how social identities are involved in such change. Much research explores minority group members' beliefs about the legitimacy and stability of group hierarchies, the permeability of group boundaries, and how, in combination, such beliefs encourage strategies of individual social mobility (in which group members attempt to leave their group and enter higher status groups) or enter into collective action to raise the relative status of their group (see Wright, 2003).

Important in such work is minority group members' awareness of alternative social arrangements. Yet, if people act in terms of their collective identities, and these are (re)-produced through wider social practices (e.g., those of 'banal nationalism'), it is not at all clear how the new identity definitions necessary for the transformation of existing social relations arise (Reicher, 1996). What is needed is some sense of agency. Yet, this must avoid a voluntarism in which anything is possible and the oppressed are blamed for the limits to their

political imagination. Moreover, even if individuals can imagine alternative social arrangements, the question of how such 'cognitive alternatives' and dissident ideas can gain currency among a wider collective, and become the basis for new forms of social action, remains. Or to put it another way, how can new understandings of the social world and the new social identifications they imply become the basis for group rather than isolated individual activity?

Insight into such issues comes from research concerning crowd protests. Although many have viewed crowd behavior as irrational and uncontrolled there is much evidence that crowd behavior is socially patterned and violent behavior (if indeed it occurs) should be conceptualized as behavior in accordance with crowd members' common social identification (Reicher, 1984). Recent work on how crowd members develop this common identification highlights the issue of identity change. For example, studies of antiroad protests show that although participants may have originally defined themselves (and acted in terms of) one identity (perhaps as local community members concerned over the loss of an area of natural beauty), they may, because of heavy-handed police action, come to see themselves in terms of a larger scale group identification subscribing to a more radical critique of authority and commercial interests (Drury and Reicher, 2000). Such analyses pay particular attention to how participants' social categorical self-definitions are affected by the response of others (especially authorities such as the police). Where there is an asymmetry between, on the one hand, how crowd members see themselves and, on the other, how they are acted upon by the authorities, crowd participants may begin to entertain (and act in terms of) alternative and radicalized definitions of crowd identity (which initially may have been shunned). Moreover, to the degree that a common social identification facilitates the coordination of behavior and the realization of this 'new' social identification, participants may experience a sense of empowerment which affirms the significance and meaningfulness of that social identification (Drury and Reicher, 2005).

Away from crowds, other research also points to the importance of intergroup interaction in fostering group members' awareness of a mismatch between how they understand themselves and their qualities and how others define them. For example, Werbner (2000) offers an account of a UK Muslim women's group that through interactions with the local male leadership underwent a change in social identification. The group, comprising middle-class Punjabi women, started as a philanthropic organisation engaging in fund-raising for projects in Kashmir and Palestine. Over time, as their campaigning plans were questioned by local male community elders, the women became increasingly aware of a mismatch between their understandings of what they could do and achieve, and what local leaders envisaged for them, and this prompted an exploration of their collective identification as Muslim women. This involved reconstructing hitherto taken-for-granted customs as not genuinely 'Islamic' but 'merely' cultural practices and they took as their role model the Prophet's wife Ayesha (celebrated for her boldness and wisdom).

The creativity and imagination exhibited in these women's self-definitions highlights the limitations of lay and academic theories of identity that imply a fixed essence. Yet it remains important to contextualize these identity constructions as productions arising in the context of intergroup argument. That is, they arose to counter alternative definitions, and, in turn, they prompted further contestation, and so on. In this process, ideas and positions that had once seemed irrelevant or inappropriate became more meaningful as the women's campaign successes highlighted the mismatch between how the women saw themselves and how the traditional male leadership saw them. The women's reformulation of their identity was not immediate nor was the course it would take immediately obvious. Rather, it was contingent on the social dynamic of intergroup dispute in which the women, individually and collectively, engaged with particular discourses of Islam and gender, and so developed new appreciations of their Muslim identification that empowered them as Muslims and as women. Moreover, Werbner describes how their successful activities had unforeseen consequences of empowering them as British citizens.

Conclusion

Everyday talk of identity is problematic. Often it implies that groups have a fixed 'essence' that exists independently of their cognitive representation and the conditions of their social construction. However, this does not mean that the concept of identity is analytically irrelevant – only that certain understandings of this term are problematic – and social psychological theorizing on social identity manages to avoid many of the pitfalls associated with taking identities as givens. It alerts us to the importance of exploring people's self-categorizations and how these social categorical definitions are variable, contingent upon context and social practice. It also alerts us to how such self-defining categorizations are psychologically consequential in making group behavior possible. The inclusivity of such self-defining social categories determines who is judged as in-group and this shapes the range of others with whom one feels a sense of solidarity. In turn, this shared social categorical definition allows group members to coordinate their goals and interests, and to enact the norms and values they associate with their common social identification.

Yet, such identity-defining norms and values are themselves the product of social processes, and again our attention must turn to the social context in which such identity constructions are produced. In turn this requires attention to the contestation over identity. Precisely because social self-categorizations shape individuals' understandings of how they should act, questions relating to which categories are used, their boundaries, their contents, etc., are inevitably the focus of human social activity with important consequences for the making and remaking of the social order. Accordingly, rather than reserving the term 'identity politics' for a distinctive form of political activity, all social and political activity may be analyzed as involving (and being accomplished through) the strategic construction of collective identity (Hopkins and Kahani-Hopkins, 2009; Reicher and Hopkins, 2001). Indeed, analyzing the process of identity construction (and the collective action that such constructions make possible) must be central to the social scientific analysis of people's world-making and world-changing social practices.

See also: Collective Action; Collective Behavior, Social Psychology of; Deindividuation, Psychology of; Intergroup Relations; Leadership; Self-Categorization Theory; Social Categorization; Social Identity in Social Psychology; Social Psychology; Stereotypes in Social Psychology.

Bibliography

Anderson, B., 1983. Imagined Communities. Verso, London.

Billig, M., 1995. Banal Nationalism. Sage, London.

Brubaker, R., Loveman, M., Stamatov, P., 2004. Ethnicity as cognition. Theory and Society 33, 31–64.

Drury, J., Reicher, S.D., 2000. Collective action and psychological change: the emergence of new social identities. British Journal of Social Psychology 39, 579–604.

Drury, J., Reicher, S.D., 2005. Explaining enduring empowerment: a comparative study of collective action and psychological outcomes. European Journal of Social Psychology 35, 35–58.

Drury, J., Cocking, C., Reicher, S.D., 2009. The nature of collective resilience: survivor reactions to the 2005 London bombings. International Journal of Mass Emergency and Disasters 27, 66–95.

Emerson, R., 1960. From Empire to Nation. Harvard University Press, Cambridge, MA.

Haslam, S.A., Reicher, S.D., Platow, M.J., 2011. The New Psychology of Leadership: Identity, Influence and Power. Psychology Press, London.

Hopkins, N., 2011. Dual identities and their recognition: minority group members' perspectives. Political Psychology 32, 251–270.

Hopkins, N., Blackwood, L., 2011. Everyday Citizenship: Identity and recognition. Journal of Community & Applied Social Psychology 21, 215–227.

Hopkins, N., Greenwood, R.M., 2013. Hijab, visibility and the performance of identity. European Journal of Social Psychology 43, 438–447.

Hopkins, N., Kahani-Hopkins, V., 2009. Reconceptualising 'extremism' and 'moderation': from categories of analysis to categories of practice in the construction of collective identity. British Journal of Social Psychology 48, 99–113.

Jenkins, R., 1996. Social Identity. Routledge, London.

Jetten, J., Haslam, C., Haslam, S.A., 2012. The Social Cure: Identity, Health, and Well-being. Psychology Press, London & New York.

Levine, M., Prosser, A., Evans, D., Reicher, S.D., 2005. Identity and emergency intervention: how social group membership and inclusiveness of group boundaries shapes helping behavior. Personality and Social Psychology Bulletin 31, 443–453.

Pehrson, S., Brown, R., Zagefka, H., 2009. When does national identification lead to the rejection of immigrants? Cross-sectional and longitudinal evidence for the role of essentialist in-group definitions. British Journal of Social Psychology 48, 61–76.

Reicher, S., 1984. The St. Pauls' Riot: an explanation of the limits of crowd action in terms of a social identity model. European Journal of Social Psychology 14, 1–22.

Reicher, S., 1996. Social identity and social change: rethinking the context of social psychology. In: Robinson, W.P. (Ed.), Social Groups and Identities: Developing the Legacy of Henri Tajfel. Butterworth-Heinemann, Oxford, pp. 317–336.

Reicher, S., Hopkins, N., 2001. Self and Nation. Sage, London.

Reicher, S.D., Cassidy, C., Wolpert, I., Hopkins, N., Levine, M., 2006. Saving Bulgaria's Jews: an analysis of social identity and social solidarity. European Journal of Social Psychology 36, 49–72.

Sani, F., Reicher, S.D., 1998. When consensus fails: an analysis of the schism within the Italian Communist Party. European Journal of Social Psychology 28, 623–645.

Sen, A., 2006. Identity and Violence: The Illusion of Destiny. Penguin, London.

Swann, W.B., 1987. Identity negotiation: where two roads meet. Journal of Personality and Social Psychology 53, 1038–1051.

Tajfel, H., 1978. Interindividual and intergroup behaviour. In: Tajfel, H. (Ed.), Differentiation between Groups: Studies in the Social Psychology of Intergroup Relations. Academic Press, London, pp. 27–60.

Tewari, S., Khan, S., Hopkins, N., Srinivasan, N., Reicher, S.D., 2012. Participation in mass gatherings can benefit well-being: longitudinal and control data from a North Indian Hindu pilgrimage event. PLoS ONE 7 (10), e47291. http://dx.doi.org/10.1371/journal.pone.0047291.

Turner, J.C., 1991. Social Influence. Open University, Milton Keynes.

Turner, J.C., Hogg, M.A., Oakes, P.J., Reicher, S., Wetherell, M., 1987. Rediscovering the Social Group: A Self-categorisation Theory. Oxford, Blackwell.

Wakefield, J.R.H., Hopkins, N., Cockburn, C., Shek, K.M., Muirhead, A., Reicher, S., van Rijswijk, W., 2011. Impact of adopting ethnic or civic conceptions of national belonging for others' treatment. Personality & Social Psychology Bulletin 37, 1599–1610.

Werbner, P., 2000. Divided loyalties, empowered citizenship? Muslims in Britain. Citizenship Studies 4 (3), 307–324.

Wright, S.C., 2003. Strategic collective action: social psychology and social change. In: Brown, R., Gaertner, S. (Eds.), Blackwell Handbook of Social Psychology: Intergroup Processes. Oxford, Blackwell, pp. 409–430.

Identity in Anthropology

TM Luhrmann, Stanford University, Stanford, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7154–7159, © 2001, Elsevier Ltd.

Abstract

When Erik H. Erikson (for many years the central theorist of identity) wrote about identity, he described it as a kind of consolidation of self, so that when someone achieved her identity, the way she interacted with the world–his or her ability to trust, to work, and to play–was recognized externally by others in a way that was consonant with her own internal understandings. In the closing decades of the twentieth century, that simplicity has been lost. There are at least three reasons for this: the postmodern turn in anthropology and elsewhere in the social sciences; the recent focus on power and agency; and the emergence of what is called 'identity politics.' Each takes a somewhat different perspective. Nonetheless, one can say that 'identity' is no longer conceived as necessarily unitary, and no longer perceived as a match between inner experience and outer acknowledgment, but as a kind of mastery of multiple narratives of self-presentation.

'Identity' is a vexed topic in anthropology. Simply described, a person's identity is his or her psychologically salient individuality, the way he or she feels different from or similar to other people. But what that means to anthropologists has shifted over time.

At one time, anthropologists, and perhaps all social scientists, tended to think of identity as a clear and unambiguous assertion of selfhood, which directed behavior and motivated actions. Self, person, and identity were more or less interchangeable. When Erik Homburger Erikson (for many years the central theorist of identity) wrote about identity, he described it as a kind of consolidation of self, so that when someone acquired her identity, the way she interacted with the world–her ability to trust, to work, and to play–was recognized externally by others in a way that was consonant with her own internal understandings. For Erikson, then, 'identity' was something that had to be achieved, and its achievement was a developmental milestone in the human life cycle. The Eriksonian self stood facing the world as the private awareness of that person which, when mutually recognized and consciously accepted, became her identity. "Identity formation … arises from the selective repudiation and mutual assimilation of childhood identifications and their absorption into a new configuration, which, in turn, is dependent on the process by which a society (often through sub-societies) identifies the young individual, recognizing him as somebody who had to become the way he is and who, being the way he is, is taken for granted" (1963: p. 159). By definition, identity was unitary and more or less coherent.

In the closing decades of the twentieth century, that simplicity has been lost. There are at least three reasons for this: the postmodern turn in anthropology and elsewhere in the social sciences; the recent focus on power and agency; and the emergence of what is called 'identity politics.' Each suggests a somewhat different perspective. Nonetheless, one can say that 'identity' is no longer conceived as necessarily unitary, no longer perceived as a match between inner experience and outer acknowledgment, and no longer understood as a developmental achievement. But before we discuss these changes, some discussion of terms is in order.

Definitions

At least four terms are inextricably intertwined in these discussions: identity, self, agency, and person. Were the intellectual world to be sensibly organized, anthropologists would use these four words to refer to different concepts, but that is not always the case. As a result, the discussions can seem a little murky. It is helpful in this context to turn to discussions in the psychological literature that have been read by anthropologists and to use them to clarify the central concepts.

To begin with the self, more complicated psychoanalytic approaches to the self began to emerge with the psychoanalyst Heinz Kohut's work. In a famous article (1959), he pointed out that the tools of psychoanalysis are introspection and empathy, and he argued that psychological phenomena like the self thus necessarily include introspection and empathy as essential elements. The 'I-experience,' as he called it, captures the way an individual introspects; thus evidence for the self is fundamentally different from evidence for the observable body. Kohut went on to establish his own vocabulary about the 'I-experience,' the most useful phrase probably being the 'self-object,' the other person that an individual may use to soothe himself and restore inner harmony, as a mother can reassure a child that an imagined dragon is not real. His most fundamental contribution, however (at least for the present purpose), was to distinguish between the introspective nature of the self and the differently observable nature of the person.

This is the distinction that the psychoanalyst Roy Schafer later develops as the distinction between the person, as the agent who acts, and the content of the ideas they have about personhood, the person's understanding of his or her 'I.' He calls these 'self-narratives' (but often refers to them as selves). Schafer also argues that from this perspective, there is clearly more than one self, a position that some anthropologists and psychologists also now adopt (Kondo, 1990; Gergen, 1991). There may be the self who is confident, assertive, and effective and the one who is weak, embarrassed, and reactive (it is that 'second' self which the analysand will present in the analytic hour). There is the self we become when speaking French, freer, more sophisticated, we feel, than our English-speaking self (cf Kaplan, 1993). There is

the work self, the play self, the parenting self, the self in different roles, and the self under different circumstances. Schafer (1992: p. 25) quotes a male analysand speaking to his analyst: "I told my friend that whenever I catch myself exaggerating, I bombard myself with reproaches that I never tell the truth about myself, so that I end up feeling rotten inside, and even though I tell myself to cut it out, that there is more to me than that, that it is important to me to be truthful, I keep dumping on myself." There are, Schafer suggests, eight selves of five types in that remark. The types are actual self, ideal self, self as place, self as agent or subject, and self as object. The selves are analysand self, social self (talking to friend), bombarding self, derogated self, exaggerating self, conciliatory advisor self, advisory self, and the defended self with redeeming features. Schafer is perhaps being excessive (it is his exceeding self) but the point is powerful. To call all of what his analysand describes the product of a single self is to lose a good deal of helpful detail.

From this perspective, a self–or more precisely, a self-narrative or self-representation–is a cognitive schema, a bounded collection of conceptions and images that an individual uses to perceive, to categorize, and to experience his world: in short, to think with. Clearly, people are motivated by more than one simple cognitive packet. They are Americans, professionals, parents, neighbors; they are shy in some settings, assertive in others; skilled and aggressive in some games, cautious and awkward in others. Some theorists conceptualize these 'packets' in terms of roles. Mardi Horowitz et al. (1996), more complexly, write of 'role-relationships.' We establish, they argue, particular ways of relating to people that are driven by our concepts of who they are, who we are in relation to them, and how we are to act in that relationship. They see these role relationships as integrated in a sense of personal identity. Claude Steele (1988) provides evidence for that integration in an experiment in which women in Salt Lake City were called by a man posing as a pollster. While the call was ostensibly about conducting a future poll on women's issues, some women were told that it was commonly known that they were uncooperative with community projects. Some were told that it was commonly known that they were cooperative. Some were told that they were bad drivers. Two days later, a woman called them back and asked them to help with a community project. Salt Lake City is a heavily Mormon community, with a strong ethic of community cooperation. Perhaps unsurprisingly, those who were insulted as uncooperative were twice as likely to agree to help out as those who were praised, as if to clear their good names to themselves. What was more remarkable, however, is that those who were called bad drivers were also twice as helpful as those who had been praised. Steele went on to argue that people often respond to damage to one self-representation by affirming success in another. I hang back with embarrassment as a lousy athlete when friends at a beach party decide that we should all play volleyball, but I can make myself feel better by reminding myself, when I drop the ball, that I am competent in the classroom. Those of us who drop volleyballs can enhance our self-esteem by bringing a terrific homemade pie to the next seaside event.

At the same time, the term 'self' seems to imply something more than this tale of loosely integrated but shifting schemas suggests. Cultural psychologists Markus and Kitayama (1994: p. 569) argue that "the model which underlies virtually all current social science views the self as an entity that a) comprises a unique, bounded configuration of internal attributes (e.g., preferences, traits, abilities, motives, values, and rights) and b) behaves primarily as a consequence of those internal attributes." They go on to argue that in the West, for the most part, people assume that the major normative task of the self is to maintain its independence, and in other, particularly Asian, settings, the normative task becomes maintaining interdependence with others. (This proposition has been a subject of active debate in anthropology.) They use the concept of the 'self,' then, to indicate not only schemas of the 'I,' but processes of acting in the world.

The cognitive psychologist Ulric Neisser (1988) makes sense of this confusion in an article widely read by psychologically minded anthropologists. He argues that there are five kinds of self-knowledge, five ways in which we have knowledge of our 'I.' There is the ecological self, the 'I' perceived with respect to the physical environment; the interpersonal self, experienced in emotional communication, the way 'I' behave in a particular human interchange; the extended self, based on memory and anticipation, the 'I' who has done certain things in the past and will likely engage in certain routines in the future; the private self, experienced as the awareness that other people cannot see your thoughts, that (in some ways) only I can know me; and the conceptual self, the wealth of concepts that I believe comprise my self: my roles (professor, friend, dog-owner), my features (I have a liver, a nose, and a mind), and my specific traits (I am tall, short, smart, stupid, brunette, blond). It is important, Neisser argues, to distinguish between kinds of self-knowledge, because otherwise the self is full of apparent contradictions. It is physical yet mental, public yet private, directly perceived yet incorrectly imagined, universal yet culture-specific. "Although there is nothing with which we are more familiar, we are often enjoined to know ourselves better than we do" (1988: p. 35). He resolves the contradictions, however, by speaking of different 'selves.' "They differ in their origins and developmental histories, in what we know about them, in the pathologies to which they are subject, and in the manner in which they contribute to human experience" (1988: p. 35).

The major distinction here is between 'self' and 'person.' 'Self' is used to refer to the introspective experience of the 'I,' and while there are at least two dominant ways of theorizing this experience–the first, as a collection of cognitive schemas and the second, as a more complex integration of those schemas with other psychological processes–the term focuses in on a subject-centered internal experience of the subject. 'Person,' by contrast, is an other-centered term. It evokes an individual's understanding of the shared qualities of individuals like himself or herself–the common qualities of those who have selves. Most anthropological discussion of the person takes its starting point from Mauss's (1938) famous essay. Mauss pointed out that the concept of the person changed considerably over time and across space. It was only in the Roman Empire, for example, that the concept of the person came to refer to a locus of rights, which has remained the legal meaning of person in English and American law today. As Geertz (1983: p. 59) points out, the Western concept of the person is 'a rather peculiar idea within the context of world cultures.'

The distinction between 'identity' and 'agency' redescribes that internal/external opposition in more specific

terms. 'Agency' is the volitional capacity of all persons: their ability to choose, to intend, and to act. When anthropologists talk about 'agency,' they contrast it to 'structure,' by which they tend to mean the external constraints on human freedom institutionalized within a society (the kinship structure, political structure, economic structure, and so forth; see Ortner's (1996) useful discussion).

'Identity' tends to be used for a particular kind of self: the internal, subject-centered experience of unique agency. It would (or should) seem strange to say, 'his whole self is wrapped up in being an anthropologist.' That would seem to imply that even the way he located his body in a room and his memories of childhood were somehow centered on being an anthropologist. But to say, 'his whole identity is wrapped up in being an anthropologist' means that he thinks that the way he currently chooses, intends, and acts has a lot to do with the way he conceives of himself as an anthropologist and not as a chemist. The concept 'self' necessarily entails a sense of not being another person. 'Identity' indexes a much sharper, more conscious sense of difference: having a different kind of agency, making different kinds of choices. The term 'identity,' then, should be used for a subject-centered sense of personal agency. It answers the question, 'who am I?' by asking, 'what can I do in the world?' (See Holland et al., 1998 for a detailed discussion of identity and agency through a dialogic, practice theory orientation.)

Contemporary Influences

There are three dominant influences on contemporary identity theory within anthropology. First is the literary, postmodern turn in the social sciences. This has brought an emphasis on narrative and performance. Identity becomes something which individuals perform, enact, and present; often, the medium of performance is understood as narrative (see Steedly, 1993; Stewart, 1996). Lavie's *Poetics of Military Occupation* (1990) is a good, relatively early, example of this kind of work. The goal of Lavie's ethnography is to describe the way Bedouin have come to understand themselves as actors in the world. She presents this through presenting the performance of different Bedouin roles–the sheikh, the fool, the old woman–and through the stories Bedouin tell about themselves. And she uses photographs which, unlike the photographs of most ethnographies, enact the thesis of the book.

Second is the emphasis on power and agency. Like the literary, postmodern turn, this influence permeates much of contemporary anthropology. Its relevance to identity theory lies in the way agency is understood to act within the constraints of external power. Ethnographies with these concerns focus on those who would seem deprived of power or importance, and then demonstrate how even within the margins of the nation-state, or in the lowest ranks of the factory, people identify themselves as effective actors in the world. For example, Tsing (*In the Realm of the Diamond Queen*, 1993) organizes her ethnography around a woman who is not only marginal in the eyes of most Westerners (she lives in the Indonesian hills) but in the eyes of her own people: she is intermittently psychotic and is the leader of a small, unpopular cult. Tsing uses her speeches as the vehicle to illustrate how those on such margins nevertheless conceive of themselves as effective actors in relationship to the center, and to some extent actually are.

The third and most powerful influence comes from cultural studies, and more specifically from postcolonial studies and identity politics. Postcolonial studies focus on relationships between the colonizer and the colonized, and identity politics on relationships between different races and/or genders, but for both the central issue is an asymmetrical power relationship and its consequences.

Memmi's *The Colonizer and the Colonized* (1965) is a foundational text in postcolonial studies. Memmi argues that he who is colonized models himself on the colonizer, aping his customs and yearning for his appearance. He loses his past; he learns the history of Cromwell but nothing of his own progenitors. He grows embarrassed by his mother tongue. "The first ambition of the colonized is to become equal to that splendid model [of the colonizer] and to resemble him to the point of disappearing in him" (1965: p. 120). The colonized agrees, in Memmi's words, to destroy himself and become what he is not. But then he learns that in the end the colonizer will reject him anyway, and he is filled with shame and self-hatred (1965: pp. 121–122). Ultimately, he has no choice but to recognize the anger in his ambivalence and to rebel. "Revolt is the only way out of the colonial situation, and the colonizer realizes it sooner or later" (1965: p. 127). And yet even if he is successful he remains still tortured, still living in a psyche defined by the colonizer, defined by what he is not. "So goes the drama of the man who is a product and victim of colonialism. He almost never succeeds in corresponding with himself."

Frantz Fanon, in *Black Skin, White Masks* (1967), wrote even more starkly about colonialism and race: "for the black man there is only one destiny. And it is white" (1967: p. 10). The terrible cost of colonialism in Fanon's eyes is that, in yearning to be like the white colonizer, the black colonized man comes to accept the white man's degrading vision of the black man and so to lose himself. He is alienated from himself when he feels most white, and humiliated when he feels most black. And so, perpetually, the black man is torn, rejecting himself to become white, only then to grasp a more terrible vision of his never-to-be-scrubbed-clean skin. "A Negro is forever in combat with his own image" (1967: p. 194).

Identity politics takes this central relationship between a dominant group and a subordinate group, an us and a them, and makes theories out of the attempts to change it. For those confronted by the inferiority of their skin, gender, or accent, should they adopt the mainstream style and hope to be respected, or reject the mainstream in order to build a world in which its values will not matter? Those who write within this arena tend to see the route to political resolution of their problems through the conceptual deconstruction of this basic opposition. West, for example, writes of the 'double consciousness' of the rejection of the mainstream (and its association of blackness with inferiority) while assimilating into it. He writes admiringly of "cultural workers who simultaneously position themselves within (or alongside) the mainstream while clearly aligned with groups who vow to keep alive potent traditions of critique and resistance" (1990: p. 33).

Bhabha, another central voice in identity politics, similarly diagnoses the central problems of the colonized subject as ambivalence and hybridity. The colonial subject identifies with

the colonizer and yet cannot be the colonizer. If he identifies with his nativeness, he confronts his own condemnation; if he identifies with the Westernization he has adopted, he confronts his own alienation. "The taking up of any one position, within a specific discursive form in a particular historical conjecture, is then always problematic" (1990: p. 81). Bhabha uses the word 'hybridity' to describe the solution, or condition, of the multiply defined self which both reiterates binary colonial categories and ultimately subverts them, but through echoes and not direct opposition. Bhabha suggests that the post-colonial subject, the marginal subject, is someone who feels that her capacity for self-declaration has been taken from her repeatedly through a long history in which she has been forced to conform to the dominant culture's prejudices and biases.

As a result, the politics in identity politics become fragmented and complex. Another theorist, Lata Mani, argues that to understand these different perspectives as a unity "sets up problematic chains of equivalences between, say, people of color in the United States, people from the third world, lesbians, gays … It is not as though difference is not acknowledged, for an inventory of difference is crucial to this narrative; rather it is that difference is insufficiently engaged" (1992: p. 393). Political affiliation becomes possible only though acknowledging difference, because these politics are as much about selfhood as about political economy.

Bhabha sees this complexity of identifications as a condition of modernity. This does tend to be a view about elites (or the middle class) by those who are elites (and have money to spend at Tower Records) but, as elites, many theorists agree. The anthropological critic James Clifford (1988: p. 11) diagnoses contemporary identities thus: "intervening in an interconnected world, one is always, to varying degrees, 'inauthentic': caught between cultures, implicated in others. Because discourse in a global system is elaborated vis-à-vis, a sense of difference or distinctness can never be located solely in the continuity of a culture or tradition."

Contemporary Identity

One could argue that the mastery of multiple narratives is the core project of the postmodern politics of identity. This perspective on identity is quite different from Erikson's naturalistic vision of an identity which can be intuited by an objective observer over a limited period of time. The most insistent message of these identity theorists is the refusal to be characterized by someone else's narrative, and to assert their own authority in laying claim to a narrative of agency that ultimately may not be so different from those they have rejected. When Gayatri Spivak (1987), for example, denounces her characterization as a Marxist, a feminist, a deconstructionist (she is famous for being all three), she does so, one suspects, not because she rejects the personal resonance of those categories for her, but because she resists having them imposed. When members of an academic culture are complicit in the dicta that only women can write on feminism, only men may write about the men's movement, only African-Americans can write on black politics, we acknowledge the importance of allowing those who have been disempowered to experience empowerment through the casting of their narratives of self-characterization.

From this perspective on the politics of identity, successful (healthy, appropriate) identity in a modern postcolonial context is less a self-characterizing narrative with a mirroring world than a sense of command over narrative complexity: narratives which one has to some extent chosen, whose mastery gives one a sense of direction, adequacy, and goodness. Rather than locating identity in single narratives, or in narratives matched to external approval, we capture more of the emphasis of these accounts in their experience of comfort and control in slipping from one narrative to another (see Luhrmann, 1996). This comfort is what Schafer evokes in his account of the freeing of the 'imprisoned' analysand: not that the analysand is unconstrained by narrative, but that the analysand comes to feel that her narratives of personhood, complex and contradictory as they might be, are those that she has chosen–that rather than being controlled by the brute oppression of a narrative externally imposed, she is in mastery of the narratives which structure her presentations of personhood to the world.

But, as the decades pass, that formulation will change again.

Bibliography

Bhabha, H., 1990. The other question: difference, discrimination & the discourse of colonialism. In: Ferguson, R., et al. (Eds.), Out There. MIT, Cambridge, pp. 71–88.

Clifford, J., 1988. The Predicament of Culture: Twentieth Century Ethnography, Literature, and Art. Harvard University Press, Cambridge, MA.

Erikson, E.H., 1963. Childhood and Society. Norton, New York.

Fanon, F., 1967. Black Skin, White Masks. Grove, New York.

Ferguson, R., et al. (Eds.), 1990. Out There: Marginalization and Contemporary Cultures. MIT Press, Cambridge, MA.

Geertz, C., 1983. Local Knowledge: Further Essays in Interpretative Anthropology. Basic Books, New York.

Gergen, K.J., 1991. The Saturated Self: Dilemmas of Identity in Contemporary Life. Basic Books, New York.

Grossberg L., et al. (Eds.), 1992. Cultural Studies. Routledge, New York.

Holland, D., et al., 1998. Identity and Agency in Cultural Worlds. Harvard University Press, Cambridge, MA.

Horowitz, M., Stinson, C., Milbrath, C., 1996. Role-relationship models: a person-schematic method for inferring beliefs about identity and social action. In: Jessor, R., Colby, A., Shweder, R. (Eds.), Ethnography and Human Development. University of Chicago Press, Chicago.

Kaplan, A., 1993. French Lessons: A Memoir. University of Chicago Press, Chicago.

Kohut, H., 1959. Introspection, empathy and psychoanalysis. Journal of the American Psychoanalytic Association 7, 459–483.

Kondo, D.K., 1990. Crafting Selves: Power, Gender and Discourses of Identity in a Japanese Workplace. University of Chicago Press, Chicago.

Lavie, S., 1990. The Poetics of Military Occupation. University of California Press, Berkeley, CA.

Luhrmann, T.M., 1996. The Good Parsi: The Fate of a Colonial Elite in a Postcolonial Society. Harvard University Press, Cambridge, MA.

Mani, L., 1992. Cultural theory, colonial texts: reading eyewitness accounts of widow burning. In: Grossberg, L., et al. (Eds.), Cultural Studies. Routledge, New York, pp. 392–404.

Markus, H.R., Kitayama, S., 1994. A collective fear of the collective: implications for selves and theories of selves. Personality and Social Psychology Bulletin 20 (5), 568–579.

Mauss, M., 1938. A category of the human mind, the notion of person, the notion of self, 1985. In: Carrithers, M., Collins, S., Lukes, S. (Eds.), The Category of the Person: Anthropology, Philosophy, History. Cambridge University Press, Cambridge, UK, pp. 1–26.

Memmi, A., 1965. The Colonizer and the Colonized (H Greenfield, Trans). Orian Press, New York.

Neisser, U., 1988. Five kinds of self-knowledge. Philosophical Psychology 1 (1), 35–59.

Ortner, S.B., 1996. Making Gender: The Politics and Erotics of Culture. Beacon, Boston.

Schafer, R., 1992. Retelling a Life: Narration and Dialogue in Psychoanalysis. Basic Books, New York.
Spivak, G.C., 1987. In Other Worlds: Essays in Cultural Politics. Methuen, New York.
Steedly, M., 1993. Hanging without a Rope: Narrative Experience in Colonial and Postcolonial Karoland. Princeton University Press, Princeton, NJ.
Steele, C., 1988. The psychology of self-affirmation: sustaining the integrity of the self. Advances in Experimental Social Psychology 21, 261–301.
Stewart, K., 1996. A Space on the Side of the Road: Cultural Politics in an 'Other' America. Princeton University Press, Princeton, NJ.
Tsing, A.L., 1993. In the Realm of the Diamond Queen: Marginality in an Out-of-the-Way Place. Princeton University Press, Princeton, NJ.
West, C., 1990. The new cultural politics of difference. In: Ferguson, R., et al. (Eds.), Out There. MIT, Cambridge, pp. 19–38.

Identity in Childhood and Adolescence

Jane Kroger, University of Tromsø, Tromsø, Norway; and Western Washington University, Bellingham, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.E. Marcia, volume 11, pp. 7159–7163, © 2001, Elsevier Ltd.

Abstract

Identity development during childhood and adolescence is discussed from the perspective of identity's 'architect,' Erik Erikson. Historical origins of Eriksonian concepts are presented from both classic psychoanalytic as well as object relations theory. The precursors to identity in infancy and childhood are then described, the identity formation process is detailed, and its consequences in adult life are examined. Criticisms of Erikson's work are presented, followed by a review of key findings from studies exploring identity's antecedents, developmental pathways, and consequences. Current directions in identity research are presented and directions for future research are detailed in conclusion.

Identity can be defined as the entity that gives one the power to move with satisfying direction in one's life and to recognize and be recognized as a unique individual by others in the social context. Erik Erikson (1963, 1968) first used the term identity in his clinical practice to describe a psychological entity that seemed missing in the lives of some of the young men returning from World War II. These men knew who they were personally, but their lives seemed to lack a sense of coherence and continuity with what had gone before the War and what might potentially follow. Erikson described identity as a psychosocial phenomenon, rooted both within the individual and also within the social context. He saw adolescence as the time in the life span in which the establishment of an identity came to the forefront of psychological concerns. Today, there are different ways of defining identity as well as different research approaches used to explore identity's many dimensions (see Kroger, 2007, for a review of these approaches). For purposes here, the development of identity in childhood and adolescence will be examined through the psychosocial framework of Erik Erikson, for it is this approach that has generated the greatest body of theoretical work and empirical investigations. I begin with a brief comment on identity's psychoanalytic origins, as well as proposals from object relations theory that describe the development of the self in early childhood, a precursor to the identity formation process of adolescence. I continue with a more complete description of Erikson's identity concept and some of the research that it has generated, and conclude with comments on the ramifications of identity development during and beyond adolescence.

Precursors to Erikson's Concept of Identity

Classic Psychoanalysis

Freud used the term identity only once to refer to his ethnic (Jewish) origins (Erikson, 1959). However, Freud's (1937) work on the basic drives, psychic energy; the roles of the unconscious, preconscious, and conscious thoughts in personality development; the role of the defense mechanisms; and the balances among the id, ego, and superego in personality functioning served as the foundations for Erikson's thinking. Erikson's initial work as a teacher at a small school in Vienna with childhood friend Peter Blos brought Erikson into contact with Freud's inner circle. Invited by Anna Freud to undertake training in child psychoanalysis, Erikson received his psychoanalytic credentials as Hitler rose to power in Europe. As a consequence, Erikson and his family immigrated to Boston, and he became the first child psychoanalyst trained by Freud to work in the United States (Friedman, 1999). As Erikson's thinking evolved, however, he became more strongly focused on the role of the ego in development, and in how one comes to adapt to and engage with one's surrounding context. And while Freud had focused on the roles of biology and conscious and unconscious thought processes to account for personality functioning, Erikson also began to stress the role of the social context in personality development. Several of Erikson's first major works focused on the nature of identity formation from this life span perspective of psychoanalytic ego psychology (Erikson, 1958, 1968, 1969).

Object Relations Theory

Freud's work, based on biology, the drives for pleasure and aggression, and attaining drive satisfaction rose to prominence in Europe and was used by clinicians to understand human development. Soon, however, a rather different approach began to appear in the 1940s with the object relations theorists, who suggested that the motivation for relatedness should either supplement or replace Freud's emphasis on drive discharge as the impetus for human behavior. Object relations thinking is perhaps of greatest relevance for understanding later identity formation, for it details how the development of a sense of an autonomous self emerges during infancy. Object relations theory, in general, is a branch of psychodynamic thinking that focuses on how relationships with others are the basis for the development of a sense of self during life's earliest years and of how these relationships lie at the root of later adolescent and adult personality development, in general, and identity development, in particular.

In the 1960s, psychoanalyst Margaret Mahler and her collaborators began to theorize about how a sense of self develops during infancy and early childhood (Mahler et al., 1975). Working with extensive, video-taped observations of

mothers in interactions with their infants and young children in normal child care settings, Mahler and her colleagues came to propose the course by which one develops an autonomous sense of self during life's earliest years. Mahler used the terms separation and individuation to describe two developmental tracks important to this process. By separation, Mahler referred to the child's psychological emergence from an undifferentiated sense of oneness with the primary caretaker, the symbiotic union. By individuation, she referred to the child coming to assume his or her own unique and individual psychological characteristics. Normatively and ideally, these two developmental tracks occur concurrently, according to Mahler.

Mahler et al. (1975) continued to define a series of phases in what they termed the separation-individuation process of infancy. These phases represent gradual steps by which the young child comes to develop a sense of self, basic to any later sense of identity. Initially, the newborn finds differentiation of the self from its surroundings difficult through cycles of drowsiness, alert inactivity, hungry alertness, feeding, further drowsiness, and sleep. There may be relatively little awareness of where the boundary lies between external and internal worlds. From this normative autistic phase, a gradual awareness of the primary caretaker emerges at about 3–4 months, but this caretaker is perceived merely as an extension of the infant's own self. From this phase of symbiosis, the foundations are set for the beginnings of the separation-individuation process of infancy. More recent research has challenged the existence of these earliest phases, however, suggesting that important infant capabilities for self-differentiation do exist. In response, Pine (2004) has countered that infants experience periods of nondifferentiation that carry an affective significance strong enough to account for merger desires observed in later life. Pine has argued for the continued applicability of Mahler's work, albeit with a recognition of significant accomplishments during infancy and individual differences in the separation-individuation process.

The initial phase of the separation-individuation process, termed differentiation, refers to the time at which the 5- to 10-month-old infant slides from the caretaker's lap and begins to explore the external world. With increased locomotion, practicing, at about 10–15 months, follows for the toddler. This phase is marked by increased exploration and exhilarating engagement with the world, possible as long as the caretaker remains in place to serve as a base for emotional 'refueling.' By about 15–22 months, the child begins to experience his or her primary caretaker as a person in his or her own right, and the crisis of rapprochement ensues. During this time, the reality begins to dawn that one can no longer return to the symbiotic union of earlier times, and there are likely to be repeated attempts to engage the caretaker in external activities to fill what has come to feel like an internal vacuum as the separation-individuation process ensues. A second caretaker here is vital in helping to support the child against the regressive, symbiotic pull. Eventually, an optimal outcome to this crisis results in what Mahler terms object constancy, referring to the child's internalization of the primary caretaker's image that enables the child to function in the physical absence of his or her caretaker; in this way, the child can carry mother (or the primary caretaker) in his or her mind while pursuing independent activities. This intrapsychic structural arrangement will serve the child well until adolescence.

Although Mahler et al. (1975) never discussed the implications of an optimal separation-individuation process beyond childhood, Peter Blos (1967: p. 168) quickly appreciated the application of Mahler's work to understand intrapsychic restructuring during adolescence. He termed this series of developments the second individuation process: "Individuation implies that the growing person takes increasing responsibility for what he does and what he is, rather than depositing this responsibility onto the shoulders of those under whose influence and tutelage he has grown up." Blos compares Mahler's infant differentiation phase to the adolescent process of loosening ties with the internalized parent in order to develop new romantic relationships outside the family; both theorists, however, point to the importance of maintaining strong family relationships in the external world during intrapsychic restructuring. Blos (1967) also finds striking parallels between the normative infant regressive behaviors of the rapprochement crisis to the regressive behaviors of the adolescent during the second individuation process. Subsequent research linking the intrapsychic restructuring described by Mahler and Blos to Erikson's psychosocial concept of identity development will be described in a later section.

Erikson's Identity Concept

Identity was a central theme in Erikson's own life. He never knew his biological father, an issue that haunted him throughout his life (Friedman, 1999). Erikson felt a 'stepson' relationship both to his own family and to his psychoanalytic profession that he entered without ever having earned a single university degree. In response to questioning by Robert Coles (1970: p. 180) about his youth, Erikson remarked, "Yes, if ever an identity crisis was central and drawn out in somebody's life, it was so in my own." Thus, Erikson was intimately aware of identity difficulties, and this experience, along with those of the returning soldiers he saw in his clinical practice, enabled him to understand and elaborate details of identity's elements in ways that others had not previously attempted; such personal experiences also, no doubt, sensitized Erikson to questions about identity's origins and development in the broader, nonclinical population.

Erikson (1963, 1968, 1969) defined identity in different ways over the course of his extensive writings on the topic. Above all, identity, in Erikson's view, is a psychosocial phenomenon, determined in part by one's biology, but also by one's own conscious and unconscious desires, one's psychological needs and wishes in interaction with the demands and responses of one's social environment. Identity gives one's life a feeling of sameness and continuity, despite the different roles one may assume and the different contexts in which one may operate. Identity has its roots in childhood and consequences in adulthood. According to Erikson, identity is an evolving configuration or structure, resulting from successive ego syntheses and resyntheses over the life span.

Erikson described the psychological process of introjection in infancy and identification in childhood as necessary foundations for the identity formation process of adolescence. During infancy, the young child 'takes in' (internalizes) or incorporates the images of significant others from, ideally, satisfying relationships, and these introjects, in turn, serve as the foundation for the development of a sense of self. In childhood, through the process of identification, the child assumes valued traits and characteristics of important others, in wanting to be like them. Erikson (1968) notes, however, that identity formation can only begin when the process of identification ends. Through the identity formation process, the adolescent begins evaluate the fit between previous identifications and his or her own truly unique abilities, interests, wishes, and talents, in efforts to find meaningful forms for their expression within the social context. Erikson noted that identity is not merely the sum of previous identifications, but rather a new structural configuration resulting from syntheses or repudiations of these earlier ways of being. In Erikson's (1968: p. 158) terms, identity is "a new, unique Gestalt which is more than the sum of its parts."

The formation of identity, in Erikson's view, is based on resolutions to a series of hierarchically organized psychosocial tasks met in previous life stages, and identity resolutions, in turn, determine the ways in which life tasks of adulthood can be met. Through his epigenetic principal, Erikson believed that failure to resolve, optimally, earlier life tasks would invariably result in more limited possibilities for the resolution of subsequent life tasks. Finding an optimal balance to the infant's task of Trust versus Mistrust lies at the heart of optimal identity development as well as personality development more generally across the life span; optimal work with this task provides an outlook of hope for the ensuing psychosocial tasks of life. The toddler's resolution to life's second task of Autonomy versus Doubt and Shame optimally provides the toddler with a sense of will, vital to one's later identity development, while the young child's confrontation with the task of Initiative versus Doubt and Shame ideally results in a sense of purpose. The school-age child's optimal resolution to Industry versus Inferiority instills a sense of confidence, vital to the task of Identity versus Role Confusion met normatively during adolescence. Thus, with more or less optimal experiences of hope, will, purpose, and competence, the adolescent is ready to begin exploring and eventually actualizing identity-defining roles and values within his or her social surrounds.

Resolutions to Identity versus Role Confusion also impact resolutions to the successive adult tasks of Intimacy versus Isolation, Generativity versus Stagnation, and Integrity versus Despair. Erikson (1968) was of the opinion that issues of identity must be fairly well resolved in order to experience true psychosocial intimacy. By intimacy, he referred to a counterpointing of identities; where one's identity is uncertain, Erikson found that one might engage in highly stereotyped relationships and/or develop a sense of deep isolation. With a strong sense of identity in place, the ability to love is the optimal outcome of Intimacy versus Isolation in late adolescence or early adulthood. Generativity versus Stagnation is the key psychosocial task of young and middle adulthood, involving the capacity to care for and guide the next generation, to care for one's creative products, and one's broader community. Optimal resolution to this task, possible only with optimal resolutions to Identity versus Role Confusion as well as all preceding stages, results in the ability to genuinely care both about oneself as well as others. During late adulthood, one's main psychosocial task turns to the issue of Integrity versus Despair. As the adolescent struggles to find meaningful identity directions for the future, so the later life adult must come to terms with the meaning of his or her life as is, and "the acceptance of one's one and only life cycle …" (Erikson, 1968: p. 139). Coming to terms with one's own death is a further dimension of this task. Many identity issues again arise in late adulthood, as work and parenting demands diminish and one must redefine one's way of being in the world. Optimal resolution to this final task is that of wisdom. Thus, the establishment of one's identity opens the possibilities for the capacity to love in intimate relationships, to care about the self and others, and to develop wisdom from one's life experiences in adulthood.

Erikson (1968) stresses the importance of exploration of alternatives through a moratorium process of adolescence. He has identified the areas of vocation, ideology, and sexuality to be those identity-defining roles and values generally most expressive of one's identity over the course of adolescent and adult life, at least in many Western contexts, and he stresses the dangers of early foreclosure if opportunities are not provided within one's milieu to enable identity exploration. Erikson (1968) used the term identity crisis to describe the search for a meaningful identity-defining life direction; by this term, he referred to an important developmental phenomenon, critical for optimal identity development. Should this task prove too difficult or overwhelming, Erikson described the possibilities of identity diffusion and negative identity formation as ways of avoiding or bypassing, respectively, the identity formation process. Identity diffusion is the inability to find meaningful life directions and may result in a range of difficulties from a carefree but directionless approach to life to severe psychopathology. Through a negative identity resolution, an adolescent (or adult) will adopt those roles and values that reflect a polar contrast to all those held dear by significant others. Erikson (1968) cites the daughter of an influential southern Negro minister of religion who runs away from home to live among drug addicts in Chicago.

Criticisms of Erikson's concept of identity have come from several directions. Some have taken issue with imprecise or unclear definitions of identity that Erikson has presented. At times, identity is described in structural terms, while at other times it seems to refer to a process; sometimes it is described in phenomenological terms while at others, an unconscious experience. Erikson has rather candidly stated in an interview with Stevens (1983: p. 112): "I think one could be more precise than I am or than I am able to be. I very much feel that scientific training and logic would have helped a lot." Others have sharply criticized his views as androcentric, particularly his comments on how women's identity development is based on the man (or men) she allows into the 'inner space,' as well the ordering of Identity versus Role Confusion and Intimacy versus Isolation in Erikson's epigenetic scheme (e.g., Franz and White, 1985;

Gilligan, 1982). Still others have questioned the cultural universality of Erikson's epigenetic scheme (Arnett, in press). Despite these criticisms, Erikson's (1963, 1968) identity definitions and concepts have been popular themes that researchers have attempted to examine during childhood, adolescence, and the adulthood years.

Empirical Studies of the Identity Formation Process

Erikson's (1963, 1968) writings on identity have generated much research in the social sciences over the past five decades. One general stream of studies from the 1960s through 1980s focused on the role that 'Identity versus Role Confusion' plays in Erikson's eight-stage epigenetic scheme. A further stream of work has focused on the 'Identity versus Role confusion' task in relation to other variables (e.g., Berman, 2009 on identity distress and problem behaviors), while a third approach addresses other dimensions of identity that Erikson describes in his writings (e.g., Van Hoof and Raaijmakers, 2003; on structural integration). For purposes here, it is perhaps the expansions that James Marcia (1966, 1967; Marcia et al., 1993) made to Erikson's model of identity development that have generated the greatest volume of research over the past decades.

Rather than conceptualizing the task of Identity versus Role Confusion in terms of a continuum, with identity being an entity that one has 'more or less of,' Marcia instead proposed qualitatively different pathways by which late adolescents approach the identity formation task. Marcia (1966, 1967) used the variables of exploration and commitment that Erikson had viewed as central to identity development in adolescence to propose two ways by which one might establish identity-defining commitments, and two ways by which one might not do so. Selecting the domains of vocation, ideology, and later sexual roles and values that Erikson (1968) had described as the primary identity concerns of adolescence, Marcia developed the Identity Status Interview to identify which identity approach (or identity status) was most descriptive of the adolescent's way of approach in identity-defining decisions.

The identity-achieved individual has gone through a time of exploration, based on careful consideration of his or her own interests, talents, capacities, and values to form identity-defining commitments that set one's directions in early adult life. Like the identity achieved, the foreclosed individual has equally strong identity commitments, but they have been attained without identity exploration. Most commonly, late adolescent foreclosures will merely assume the values of significant others around them and adopt a type of conferred identity. Moratorium and diffuse individuals both lack firm identity commitments, but the moratorium is in the process of trying to find personally meaningful identity-defining directions, while the diffusion is not. There may have previously been a little identity exploration for the identity diffusion, or none at all. The identity diffuse individual may be unable to adopt meaningful identity commitments for a variety of reasons, ranging from severe psychopathology to a carefree, uninvolved approach to life, just 'going where the wind blows.'

Personality Characteristics

A number of personality characteristics have been associated with those in the various identity statuses (see Kroger and Marcia, 2011, for a review). The identity-achieved individuals have demonstrated resistance to experimental attempts to raise or lower their self-esteem, and they think effectively under stressful conditions. They also demonstrate high levels of ego development and moral reasoning, are not authoritarian in their values, and use an internal locus of control in decision-making. Moratoriums share many of these characteristics with the identity achieved, although that they may be more prone to using an external locus of control than the identity achieved. In addition, they have generally high levels of anxiety relative to the other identity statuses. Foreclosures have demonstrated very high levels of authoritarian values, and they use an external locus of control. They also have very low levels of anxiety relative to the other identity statuses. At the same time, they have shown rather high levels of self-esteem. The diffusions have generally scored high on external locus of control measures and low on measures of self-esteem, moral reasoning, and ego development.

Familial Antecedents

Much research has focused on antecedents to identity development, including different resolutions to the second separation-individuation process, styles of attachment, and styles of family communication. Relatively predictable relationships between the identity statuses and dimensions of the adolescent separation-individuation process have appeared, with the foreclosure status strongly linked to an enmeshed intrapsychic relationship to the internalized parent, while the moratorium and achievement statuses have not shown this high nurturance seeking need (Kroger, 1995). A recent meta-analysis of the relationship between attachment styles and the identity statuses found weak to moderate correlations between attachment styles and the identity statuses, as predicted patterns in the relationships appeared. Secure attachment was far higher among the identity achieved compared to foreclosures and diffusions (Årseth et al., 2009). Many studies have examined family styles of communication in relation identity exploration and commitment variables as well as to the identity statuses themselves. Observational investigations of adolescents and their parents have shown adolescent identity exploration to be positively linked with family communication processes of encouragement for self-assertion, appreciation of separateness, permeability, and mutuality (Grotevant and Cooper, 1985). More recently, positive parenting during late adolescence has been linked with identity achievement and a more positive narrative resolution for a difficult life experience (Dumas et al., 2009). Additionally, achievement and moratorium identity statuses have been positively linked with clearer appropriation of parental voice in adolescent narratives (Mackey et al., 2001). Optimal identity development in adolescence thus points to resolution of the second separation-individuation process, a secure style of attachment, and authoritative communication styles used by parents.

Developmental Patterns of Change

Identity development certainly occurs over the course of adolescent and adult life, and one might question the likelihood of different patterns of change and stability over time. Waterman (1999) has proposed a series of hypotheses central to an understanding of identity status change processes, and a number of Waterman's hypotheses have recently been tested through techniques of meta-analysis (Kroger et al., 2010). Waterman suggested that identity development from adolescence to adulthood involves a preponderance of progressive identity status changes (i.e., diffusion to foreclosure, moratorium, or achievement; foreclosure to moratorium or achievement; and moratorium to achievement). Indeed, meta-analysis of longitudinal identity status investigations conducted between 1966 and 2005 revealed that a mean proportion of 36% of identity status changes would be progressive in nature. Surprising was the mean proportion of 49% for stability in identity status over approximately 3 years during late adolescence.

Waterman (1999) also proposed that over time the moratorium status would be the least stable; results of meta-analyses showed that it was, although the diffusion status was almost equally unstable. From cross-sectional studies, Kroger et al. (2010) also predicted from Waterman's (1999) hypotheses that there would be a decrease in the mean proportion of diffuse and foreclosed individuals and a concomitant increase in the moratorium and achievement identity statuses from mid to late adolescence; however, from late adolescence to young adulthood, we also anticipated an initial drop in achievement and moratorium statuses as youths entered college or adult work roles, with a subsequent increase in these statuses over time. Meta-analyses found general evidence of these patterns. Perhaps of greatest interest was the fact that only about 50% of samples were rated as moratorium or identity achieved in the 23- to 29-year age-group and only 68% were moratorium or achieved in the 30- to 36-year age span. It appears that by young adulthood, many have not established a sense of their own identity – a task that Erikson considers fundamental to adolescence.

Epigenetic Consequences

Identity resolutions have consequences for the remaining Eriksonian stages of adult life: Intimacy versus Isolation, Generativity versus Stagnation, and Integrity verses Despair. Research to date has primarily focused on the relationship between identity and intimacy. A further meta-analytic study by Årseth et al. (2009) examines the relationship between identity and intimacy during young adulthood, statistically combining results from previous studies of this relationship. Erikson's epigenetic theory was generally supported for men. High identity status (moratorium and achievement) was positively associated with high intimacy status (intimate and preintimate), and low identity status (foreclosure and diffusion) was also associated with low intimacy status (pseudo-intimate, stereotypic, and isolate). However, the findings for women were more complicated. Approximately 65% of women who were high in identity status were also high in intimacy status. At the same time, however, women who were low in identity status were almost equally distributed between high and low intimacy status groups. Furthermore, the relationship between identity and intimacy status was much stronger for men than women ($p < 0.001$). For scale measures of intimacy, high identity status participants scored higher on measures of intimacy than low identity groups, with small to moderate effect sizes whether the sample was examined in total, for men only, or for women only. In general, it appears that a foreclosed or diffuse identity resolution during late adolescence restricts the type of intimacy development that is possible during early adult life for men and at least some women. Differing contextual opportunities for the two genders may serve as a moderator for the relationship between identity and intimacy and further research is needed on this issue.

While research on identity and generativity status has been more limited, there now have been investigations supporting Erikson's epigenetic principle on the relationship between identity and integrity status during late adulthood. Perhaps the most recent and comprehensive investigation is that by Hearn et al. (2012). Hearn and his colleagues examined the identity statuses in relation to integrity statuses through the newly developed Self-Examination Interview. Results showed that some 86% of late life adults who were integrated were also identity achieved, while no despairing individuals were identity achieved. Those who were nonexploring (had not examined questions of personal meaning), pseudo-integrated (fit the world into simplistic frames and clichéd meanings), and despairing in integrity status were primarily foreclosed in identity status. Thus, those experiencing nonoptimal resolutions to *Identity versus Role Confusion* appear limited in their resolutions to *Integrity versus Despair* in late adulthood.

Summary and Directions for Future Research

Identity, with its many dimensions, has been an important entity for theory and research in the social sciences since Erikson (1959, 1963, 1968) first began using the concept in his early career. Identity, to Erikson, was a psychosocial construct that focused on the roles of individual biology, psychological processes, as well as the social context in order to account for personality development across the life span. Erikson's fifth task of adolescence, Identity versus Role Confusion, has been the focus of much theory and research over the past five decades. James Marcia (1966, 1967; Marcia et al., 1993) has been instrumental in expanding Erikson's model to describe and trace antecedents, developmental pathways, and epigenetic consequences of his four identity statuses.

Current work with the identity status paradigm has taken a number of new directions. Efforts have been made to further refine some of the statuses. Luyckx et al. (2013) have been able to validate empirically several types of identity diffusion: a 'classical' or troubled diffusion trajectory and a carefree diffusion trajectory. Earlier, Kroger (1995) was able to differentiate a 'firm' (unlikely to change) and 'developmental' foreclosure pathway with measures of adolescent separation-individuation. Luyckx et al. (2008) also have identified

several types of moratorium positions: ruminative (or maladaptive) identity exploration and two forms of reflective (adaptive) identity exploration. Preliminary work is also underway to examine possibly different forms of the identity achievement status as it undergoes change through the years of adult development. Future identity work could fruitfully undertake long-term cohort sequential studies of identity status change and stability with data assessment at relatively frequent time intervals (1–5 years). Future investigations could also usefully examine the relationship between resolutions to identity and their relations to different intimacy, generativity, and integrity statuses in adult life. Narrative strategies might usefully be employed to detail nuances of person–context interactions that are the foundation of identity development across the life span. The field of identity development throughout childhood, adolescence, and adulthood remains a rich avenue for further explorations of and refinements to Erikson's identity concepts.

See also: Erikson's Theory of Psychosocial Development; Human Development, Successful: Psychological Conceptions; Human Development, Theories of; Identity and Identification; Self and Identity Development During Adolescence across Cultures; Self-Development in Childhood and Adolescence; Social Referencing During Infancy and Early Childhood across Cultures; Socialization in Infancy and Childhood.

Bibliography

Arnett, J. Identity development from adolescence to emerging adulthood. In: McLean, K.C., Syed, M. (Eds.), The Oxford Handbook of Identity Development. Oxford University Press, New York, in press.

Årseth, A.K., Kroger, J., Martinussen, M., Marcia, J.E., 2009. Meta-analytic studies of identity status and the relational issues of attachment and intimacy. Identity: An International Journal of Theory and Research 9, 1–32.

Berman, S.L., 2009. The prevalence and incremental validity of identity problem symptoms in a high school sample. Child Psychiatry and Human Development 40, 183–195.

Blos, P., 1967. The second individuation process of adolescence. Psychoanalytic Study of the Child 23, 245–263.

Coles, R., 1970. Erik H. Erikson: The Growth of His Work. Little Brown, Boston.

Dumas, T.M., Lawford, H., Tieu, T.T., Pratt, M.W., 2009. Developmental Psychology 45, 1531–1544.

Erikson, E.H., 1958. Young Man Luther, a Study in Psychoanalysis and History. W.W. Norton, New York.

Erikson, E.H., 1959. Identity and the Life Cycle. Psychological Issues 1 [Monograph 1]. International Universities Press, New York.

Erikson, E.H., 1963. Childhood and Society, second ed. W.W. Norton, New York.

Erikson, E.H., 1968. Identity, Youth and Crisis. W.W. Norton, New York.

Erikson, E.H., 1969. Gandhi's Truth. W.W. Norton, New York.

Franz, C.E., White, K.M., 1985. Individuation and attachment in personality development: extending Erikson's theory. Journal of Personality 53, 224–256.

Freud, S., 1937/1957. In: Rickman, J., Brenner, C. (Eds.), A General Selection from the Works of Sigmund Freud. Doubleday and Company, New York.

Friedman, L.J., 1999. Identity's Architect: A Biography of Erik. H. Erikson. Scribner, New York.

Gilligan, C., 1982. In a Different Voice: Psychological Theory and Women's Development. Harvard University Press, Cambridge, MA.

Grotevant, H.D., Cooper, C.R., 1985. Patterns of interaction in family relationships and the development of identity exploration in adolescence. Child Development 56, 415–428.

Hearn, S., Saulnier, G., Strayer, J., Glenham, M., Koopman, R., Marcia, J.E., 2012. Between integrity and despair: toward construct validation of Erikson's eighth stage. Journal of Adult Development 19, 1–20.

Kroger, J., 1995. The differentiation of 'firm' and 'developmental' foreclosure identity statuses: a longitudinal study. Journal of Adolescent Research 10, 317–337.

Kroger, J., 2007. Identity Development: Adolescence through Adulthood, second ed. Sage, Newbury Park, CA.

Kroger, J., Martinussen, M., Marcia, J.E., 2010. Identity status change during adolescence and young adulthood: a meta-analysis. Journal of Adolescence 33, 683–698.

Kroger, J., Marcia, J.E., 2011. The identity statuses: origins, meanings, and interpretations. In: Schwartz, S.J., Luyckx, K., Vignoles, V.L. (Eds.), Handbook of Identity Theory and Research, vol. 1. Springer, New York, pp. 31–53.

Luyckx, K., Schwartz, S.J., Berzonsky, M.D., Soenens, B., Vansteenkiste, M., Smits, I., Goossens, L., 2008. Capturing ruminative exploration: extending the four-dimensional model of identity formation during late adolescence. Journal of Research in Personality 42, 58–82.

Luyckx, K., Klimstra, T.A., Schwartz, S., Duriez, B., 2013. Personal identity in college and the work context: developmental trajectories and psychosocial functioning. European Journal of Personality 27, 222–237.

Mackey, K., Arnold, M.L., Pratt, M.W., 2001. Journal of Adolescent Research 16, 243–268.

Mahler, M., Pine, F., Bergman, A., 1975. The Psychological Birth of the Human Infant. Basic Books, New York.

Marcia, J.E., 1966. Development and validation of ego identity status. Journal of Personality and Social Psychology 3, 551–558.

Marcia, J.E., 1967. Ego identity status: relationship to change in self-esteem, "general maladjustment," and authoritarianism. Journal of Personality 35, 118–133.

Marcia, J.E., Waterman, A.S., Matteson, D.R., Archer, S.A., Orlofsky, J.L., 1993. Ego Identity: A Handbook for Psychosocial Research. Springer-Verlag, New York.

Pine, F., 2004. Mahler's concepts of "Symbiosis" and separation-individuation: revisited, re-evaluated, refined. Journal of the American Psychoanalytic Association 52, 511–533.

Stevens, R., 1983. Erik Erikson: An Introduction. Open University Press, Oxford.

Van Hoof, A., Raaijmakers, Q., 2003. The search for the structure of identity formation. Identity: An International Journal of Theory and Research 3, 271–289.

Waterman, A.S., 1999. Identity, the identity statuses, and identity status development: a contemporary statement. Developmental Review 19, 591–621.

Identity Movements

Simon Langlois, Université Laval, Québec City, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The specificity of contemporary social movements is a displacement of protest from the economic and the political spheres to the cultural realm. New forms of conflicts imply identity claims and express protest against social control on different aspects of personal life. Social movements combine three kinds of claims: program, identity, and standing. The study of emotion and passion is central in identity movements. Five types of identity movements are identified: national, religious, aboriginal, particularistic, and normative. The Internet has become a key mobilizing resource for identity movements, which raise the issue of universalism versus particularism.

The main objective of social movements is to transform society so as to reflect the social actors' interests and view of the world. Such was the case with the labor movement in the early twentieth century and the decolonization movement in the 1960s (*see* Oberschall, 1983; Chazel, 1992; Buechler and Cylke, 1997). One of the characteristics of contemporary social movements is a displacement of protest from the economic and political spheres to the cultural realm. The identity movements that came to the fore in the last third of the twentieth century are of a different nature, because they express two complementary types of collective demands: (1) the defense of interests and the promotion of the rights of certain groups of individuals who feel discriminated against, and (2) the search for symbolic recognition by a significant other (a good example are aboriginal social movements). These new social movements emerged as non-class based and express a displacement to cultural issues, and to an inclusion of emotions and cognition in collective action.Movements, Social.

New forms of conflicts imply identity claims and express protest against social control on different aspects of personal life. Identity movements defend the interests, world visions, and values of groups of individuals or communities defined by such characteristics as race and ethnicity; sex; language; sexual orientation; mythical origins and ancestral territory, in the case of First Nations in the Americas and the South Pacific; and religion, in the case of certain societies, especially Islamist ones. Melucci (1996) even posits that all social movements have an identity dimension.

Identity movements have four main and very distinct objectives, some of them being more salient than others. First, they denounce injustice toward minorities. Second, they convey the idea that specific cultures must be taken into consideration when public policies are elaborated so that they meet the specific needs of minorities. Third, they demand greater control of their institutions – a demand that sometimes goes as far as self-government. Fourth, identity movements bring back particularistic values in the public and secular spheres, often opposed to by activists and characterized as 'the Western world.' This objective emerged with more salience in the 2000s.

Individuals involved in identity movements promote their own interests (Olson, 1965), but their action is also based on their way of seeing things and their personal knowledge and values. Collective action is therefore marked by instrumental (pursuit of personal interest), cognitive (knowledge), and axiological (values) rationales. Aboriginal movements are promoting a specific link with nature and are often based on the knowledge that the economic situation of the First Nations is problematic.

Boudon and Bourricaud (1982) have shown that the two dimensions of collective behavior described by Smelser (1962), relations with norms and relations with values, are so closely linked as to become interdependent. The shared values of the members of a group or community play a central role in emerging identity movements, at least as important as the pursuit of individual interests. Social movements are concerned with agency but collective action always takes place in context and structure, especially in urban environments. Contemporary perspectives on social movements does not favor either agency or social structure at the expense of the other. This is particularly true for identity movements.

This article will specify what should be understood by identity, define the characteristics of different types of identity movements, and discuss the relationship between universal and particularistic norms that are specific to each movement.

Identity

In the early twentieth century, identity was often defined according to 'objective' traits, such as English tenacity, German discipline, and French ingenuity. What we call identity was then known as the 'soul of peoples' (see Siegfried, 1950). The problem of identity is no longer seen in this light (*see* Social Identity in Social Psychology).

It is no longer possible to separate personal and collective identities. Individuals now belong to several social networks and groups, which, in turn, shape their identity, and they draw from society the elements that they need for self-definition. They are faced with a great number of choices, many more than those of preceding generations. Education, individual and collective fulfillment, social mobility, career mobility, geographical mobility, and life style are some of the social processes that help to build identity. Immigration, urbanization, international trade, unequal regional development, differing interpretations of the past, social conflicts, and

social stratification also affect the construction of collective identities (Staggenborg, 2008).

Taylor (1989) defines the concept of identity in three ways. First, identity defines the moral outlook of each individual, in the sense that individuals are able to determine what is important to them. Individual identity implies self-knowledge and a successful and accepted synthesis of diverse influences, especially those of the individual's family.

The second definition came into being with modernity and has enabled individuals to have a better grip on their own destiny. Theoretically, within the limits of their capacities, people can become what they want to be; the future is no longer dictated by ascribed status in society. Political freedom, education, and economic development have enabled individuals to innovate, and, to a certain point, be responsible for their future. Each individual may have diverging or even conflicting interests, but each will eventually define values that could eventually be expressed through various social movements and collective conduct. Taylor insists that self-affirmation and recognition by a significant other are the two conditions for achieving a successful identity. Social actors involved in identity movements are motivated by what they see as a deficiency in democracy, and by the search for recognition, which involves negotiation and sometimes struggle and which is probably the main feature of such movements (see Hobson, 2003).

Because individuals are also part of a people or nation and therefore also define themselves by their collective past, national identity is the third definition that Taylor gives to the concept of identity. A nation or historical community offers people another perspective made up of a shared culture, life style, and language; nations are also imagined communities (Anderson, 1991). However, most countries are not homogeneous and they have had to deal with the problem of recognition of various nations and minorities of all kind (especially, religious) within their own midst. One can therefore say that the relationship with the other, which is at the heart of identity assertion, is an internal problem in most countries. Dumont (1984) describes the repeated mechanisms that contribute to the building of identities: ideologies, literature, historiography, and media play a key role in the shaping of a common reference. He speaks of conflicting interpretations between various categories of actors involved in the process of building identities, and defines society as a series of interpretive practices.

The Scottish example is relevant here. Language is not a central issue in defining Scottish national identity, despite an interest in Gaelic; more important is the identification with a mythic territory, a common history, and recollection of ancestral wars (Macdonald, 1997). The situation is quite different in Quebec and Catalonia, where promotion of the national language (French and Catalan, respectively) plays a central role in the building of the respective identities.

Identity Movements

For Tilly (2004), social movements combine three kinds of claims: program, identity, and standing. "Identity claims consist of assertions that 'we' – the claimants – constitute a unified force to be reckoned with" (p. 151). There are many identity-related issues in social movements. Five types of identity movements will be characterized by the 'good reasons' (see Boudon, 1999) that inspire their action, the means and field in which the action takes place, and the meaningful others involved in the search for recognition. These types include also different motivational factors, including expectancy of success, and different processes of framing and of collective identification (Pinard, 2011). The study of aspirations, emotions, and passion became central in the analyses of identity movements in the 2000s after having been neglected by the structural and organizational models (Larana et al., 1994; Stryker et al., 2000; Goodwin et al., 2001; Fiffieule et al., 2010).

National Identity Movements

Movements of national affirmation are undoubtedly the best known types of identity movements. Gellner (1997) estimates that there are 15 000 national cultures demanding recognition in a world where there are only a few hundred nation-states. These movements challenge the dominant national reference in their respective countries, and they are determined to achieve official recognition. They are composed of national minorities – not to be confused with ethnic minorities resulting from immigration, who make demands of another nature.

National minorities fight social inequalities of which they see themselves as victims, and many demand that official institutions be modified. These types of protests generally give way to a search for a new constitutional framework within a nation-state. Situations and demands, in accordance with the historical and national contexts, are extremely diversified (Calhoun, 1996). Some identity movements transformed themselves into nationalistic movements and gave rise to political parties fighting for an independent state, as this is the case in Catalona, Québec, or Scotland.

Religious Identity Movements

Identity movements contributed to bring back religion and cultural values in the secular world described by E. Durkheim and M. Weber, especially in the public sphere as defined by J. Habermas. In that perspective, the Islamist veil is an important religious symbol and also a secular claim to recognition of a national and cultural community (see McDonald, 2006, for a discussion of this issue).

Aboriginal Identity Movements

The self-assertion of aboriginal nations in countries shaped by influxes of immigrants must be considered separately. There are approximately 300 million aboriginals living in these immigrant societies throughout the world, mainly in the Americas and the South Pacific. This enormous population is divided into several thousand nations or distinct groups. In Canada, for example, there are 633 different bands and aboriginal nations – some groups numbering only a few hundred people – speaking 57 different languages.

Conquest and colonization by Europeans marginalized aboriginal nations, sometimes moving them to reservations, as is still the case in Canada and the United States. They are

demanding control of a number of institutions that will help them maintain their specific collective identities. Some aboriginal nations are striving for self-government within their respective countries, but their small populations prevent most of them from claiming complete sovereignty.

Aboriginal communities build their action on three main aspects of their value system. First, there is the feeling of a common mythical origin. Second, and probably at the heart of their culture, is their relationship with nature, the land, and ancestral territory. Third is the idea of kinship and blood; aboriginal communities generally do not integrate strangers, for one is either born Native or one is not.

The search for symbolic recognition is particularly important to aboriginal communities, since their culture has been denigrated. The leaders of these communities are striving to build a new sense of identity in order to counteract the feeling of self-contempt. This explains why, for many of them, interpretation and self-definition have become so important (Nagel, 1995). Aboriginal movements illustrate the institutionalization process of social movements when elite and professionals take in charge the claims and negotiate with governments.

Particularistic Identity Movements

Other identity movements do not question existing institutions but aim for the rightful participation of specific groups of individuals, sometimes even imposing various modes of participation suited to their particular characteristics. Such is the case with linguistic or ethnic minorities that insist on having better access to important positions in the civil service and business organizations (Jasper, 1997).

This type of movement includes the great majority of the feminist movements that demand an end to sex discrimination in public institutions and private enterprises. Although the feminist movement has very diverse components, it presents itself as an identity movement because it defends the interests of a portion of the population who feel wronged because they are women, and because of their opposition to a significant other who, in turn, becomes the main point of reference: men. Groups of women now insist on empowerment and their share of all higher appointments in various occupations. The most radical feminist movements go so far as to claim that all social relations are 'gender relations.'

The civil-rights movement for black Americans, which climaxed under the leadership of Martin Luther King, is also an identity movement (Morris, 1984). Even though this movement is typically American, in the second half of the twentieth century, it has taken on a very important symbolic value in most Western countries; it also gave rise to programs of affirmative action in favor of visible minorities in the United Sates.

Normative Identity Movements

Other identity movements set out to modify social norms that affect private, civil, and professional lives of individuals, as old social movements succeeded to do so (working hours, for example). Actors in these movements aim for symbolic acceptance of their differences and want to change social stereotypes. These differences are of a sexual, intellectual, or physical nature, such as the visible characteristics of certain immigrants (Baumann, 1996). Movements based on sexual orientation and in defense of the handicapped are good examples of this type of group, whose members feel that they are the victims of prejudice (Taylor and Whittier, 1997). Their complaints have been legally validated by many countries and have helped to change public attitudes. In most developed countries, homosexuality is no longer seen as a criminal offense and is officially accepted in some institutions, such as the armed forces. A growing number of countries acknowledge same-sex unions and even same-sex marriage.

Internet and Identity Movements

The Internet has become not only a key mobilizing resource for identity movements, but also a means by which identities are promoted and maintained. It has helped movement mobilizations and contributed to the internationalization process of many identity movements, as illustrated by feminist, aboriginal or Islamist social movements. The Internet created virtual spheres that contributed to foster the creation of cultural identities and to support identity movements, especially in urban environment (Castells, 2010). The Internet is at the heart of the diffusion process or diffusion dynamics in social movements, especially along identity issues. It contributes not only to the mobilization process itself, but also to diffuse ideas and values.

As mentioned by Givan et al. (2010), adaptation, political learning, and creative borrowing are vital to social (and identity) movements. Radical Islamists used the Internet to protest against 'the invasion of Western values' while others used it to promote democracy in their countries. Activists against capitalist globalization used the Internet to mobilize people in order to promote global justice (Moghadam, 2009). The Internet explains why some identity movements have a large diffusion outside national borders.

Universalism versus Particularism

Identity movement and the associated communitarian philosophy view values as being interwoven with specific cultures (Touraine, 1997). Symbolic representation is central in identity movements. Rituals, markers, and symbols – such as flags, placards, color (black for activists, red for the leftists), cloths, even music (see Eyerman and Jamieson, 1998) – are important in these movements and collective action and they help create a sense of belonging and establish a distance vis-à-vis others, as it was the case for the union movement. Mineworker locals in the United Kingdom developed their own banners and brass bands.

Some analysts worry about the negative consequences of cultural relativism underlying certain demands put forth by identity movements (see Boudon, 1999; Hardin, 1995). These critics emphasize that communitarian norms easily converge with the interests of the actors involved in identity movements, whereas universal norms seem to be weaker and

take more time to surface because they do not serve the immediate interests of individuals.

For other analysts, countries and modern constitutions today have no option but to accept and make the most of cultural diversity (Tully, 1995). Multiculturalism – which has three different meanings and refers to a sociological reality, a philosophy, or a state policy – must be held up as an example. Multiculturalism does not split society up into isolated subgroups but, on the contrary, offers a common framework within which differences may blossom in a context that respects diversity (Kymlicka, 1989).

See also: Civil Rights Movements; Collective Beliefs, Sociological Explanation of; Counterculture: The Classical View; Ethnic, Racial, and Nationalist Movements; Movements, Social; Social Identity in Social Psychology.

Bibliography

Anderson, B., 1991. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London; New York.

Bauman, G., 1996. Contesting Culture: Identity in Multiethnic London. Cambridge University Press, Cambridge.

Boudon, R., 1999. Le Sens des Valeurs. Presses universitaires de France, Paris.

Boudon, R., Bourricaud, F., 2000 [1982]. Dictionnaire de Sociologie. Presses universitaires de France, Paris.

Buechler, S.M., Cylke, F.K. (Eds.), 1997. Social Movements. Perspectives and Issues. Mayfield Publishing Co, London.

Calhoun, G., 1996. Social Theory and the Politics of Identity. Blackwell, Oxford.

Castells, M., 2010. The Power of Identity. Wiley, Malden (MA).

Chazel, F., 1992. Mouvements Sociaux. In: Boudon, R. (Ed.), Traité de Sociologie. Presses universitaires de France, Paris, pp. 263–312.

Dumont, F., 1984. Les Idéologies. Presses universitaires de France, Paris.

Eyerman, R., Jamieson, A., 1998. Music and Social Movements: Mobilizing Traditions in the Twentieth Century. Cambridge University press, Cambridge.

Fillieule, O., Agrikoliansky, E., Sommier, I., 2010. Penser les Mouvements Sociaux. Conflits et Contestations dans les Sociétés Contemporaines. La Découverte, Paris.

Gellner, E., 1997. Nationalism. New York University Press, New York.

Givan, R.K., Roberts, K.M., Soule, S.A., 2010. The Diffusion of Social Movements. Actors, Mechanisms, and Political Effects. Cambridge University Press, Cambridge.

Goodwin, J., Jasper, J.M., Pollentta, F. (Eds.), 2001. Passionate Politics. Emotions and Social Movements. The University of Chicago Press, Chicago.

Hardin, R., 1995. One for All: The Logic of Group Conflict. Princeton University Press, Princeton.

Hobson, B. (Ed.), 2003. Recognition Struggles and Social Movements. Contested Identities, Agency and Power. Cambridge University Press, Cambridge.

Jasper, J., 1997. The Art of Moral Protest: Culture, Biography, and Creativity in Social Movements. University of Chicago Press, Chicago.

Kymlicka, W., 1989. Liberalism, Community, and Culture. Oxford University Press, Oxford.

Larana, E., Johnston, H., Gusfield, J.R. (Eds.), 1994. New Social Movements: From Ideology to Identity. Temple University Press, Philadelphia.

McDonald, K., 2006. Global Movements. Action and Culture. Blackwell Publishing, Oxford.

Macdonald, S., 1997. Reimagining Culture: Histories, Identities and the Gaelic Renaissance. Berg, Oxford.

Melucci, A., 1996. Challenging Codes. Collective Action in the Information Age. Cambridge University Press, Cambridge.

Moghadam, V.M., 2009. Globalization and Social Movements. Islamism, Feminism, and Global Justice Movement. Rowman & Littlefield publishers, New York.

Morris, A., 1984. The Origin of the Civil Rights Movement: Black Communities Organizing for Change. Free Press, New York.

Nagel, J., 1995. American Indian ethnic renewal: politics and the resurgence of identity. American Sociological Review 60 (6), 947–965.

Oberschall, A., 1983. Social Conflict and Social Movements. Prentice Hall, Englewood Cliffs, NJ.

Olson, M., 1965. The Logic of Collective Action: Public Goods and the Theory of Groups. Harvard University Press, Cambridge.

Pinard, M., 2011. Motivational Dimensions in Social Movements and Contentious Collective Action. McGill-Queen's University Press, Montréal.

Siegfried, A., 1950. L'âme des Peuples. Hachette, Paris.

Smelser, N.J., 1962. Theory of Collective Behavior. Free Press of Glencoe, New York.

Staggenborg, S., 2008. Social Movements. Don Mills, Oxford.

Stryker, S., Owens, T.J., White, R.W. (Eds.), 2000. Self, Identity, and Social Movements. University of Minnesota Press, Minneapolis.

Taylor, C., 1989. Sources of the Self. The Making of the Modern Identity. Harvard University Press, Cambridge.

Taylor, V., Whittier, N.E., 1997. Collective identity in social movement communities. In: Buechler, S.M., Cylke, F.K. (Eds.), Social Movements: Perspectives and Issues. Mayfield Publishing Co, London, pp. 505–523.

Tilly, C., 2004. Social Movements 1768–2004. Paradigm Publishers, Boulder.

Touraine, A., 1997. Pourrons-Nous Vivre Ensemble? Égaux et Différents. Fayard, Paris.

Tully, J., 1995. Strange Multiplicity. Constitutionalism in an Age of Diversity. Cambridge University Press, Cambridge.

Identity Offline and Online

Rob Cover and Steven Doak, School of Social Sciences M257, The University of Western Australia, Crawley, WA, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

There are many approaches to understanding identity, although dominant in critical, cultural, and scholarly contexts are those that understand identity to be constructed or constituted through available discourses and communication methods. Today, that includes the realm of digital Web 2.0 communication. This has complexified the ways in which we can understand the cultivated forms of identity: interactivity, networking, and social networking are all significant factors that change how we perceive the self and its cultural markers. This article provides an overview of the ways in which identity is constituted in off-line and online communicative environments. The article begins with a brief summary of the emerging and current theories of identity and subjectivity; an analysis of the ways in which identity has been discussed as fluid in text-based virtual reality Web 1.0 contexts as well as the shortcomings of such views, and the way in which Web 2.0 sites, emphasizing social networking, present us with new ways in which to think about the production and performance of identity.

Identity and Digital Environments

Identity is one of the most complex facets of everyday communicative engagement, and there are many approaches to understanding the relationship between media, technologies, communicative environments, and subjectivity. However, despite the range and efficacy of theoretical approaches to identity, everyday perceptions of subjectivity and selfhood persist, particularly those which operate within Enlightenment understandings of actions, thoughts, and desires emanating from an inner identity core in which identity is thought and felt to be essentialist, fixed, and innate.

Identity incorporates subjectivity, which is the sense of self as produced through various cultural, political, legal, and linguistic formations, and discrete facets of identity that are intelligible through common demarcations that include gender, race, ethnicity, nationality, sexual orientation, as well as other facets of everyday and political life such as religion, occupation and career, political and partisan affiliation, class demarcations, particular choices in taste and fashion, and the identifications which are produced in everyday relationality with a network of friends, colleagues, acquaintances and significant others in relationships, and in kinship with immediate and extended family.

Core to identity, then, is the broad array of communication methods, channels, forms, texts, interpretations, audiences, and modes of reception. This is the case for two reasons: the cultural and discursive means by which markers and forms of identity are made sensible in discourse, culture and language are mediated (accessed, received, and interpreted through media and communications technologies) and the relationality and forms of mutual identification and recognition that may occur among a group of subjects may at times be face-to-face, embodied and rely on linguistic verbal communication but, at other times in the contemporary era, depend on traditional and new technologies of communication, from the telephone and television to Internet, world wide web, and social networking. Given the increasing contemporary first world ubiquity of digital technologies for communication, there has been some impact on the ways in which we 'do' identity and the ways in which identities are made intelligible to us. However, that is not to say that digital communications technologies and online worlds have wholly changed the methods by which identity is performed; rather, there is a clear – if complexified – articulation between off-line embodied and corporeal communication forms and online digital media technologies and its uses in the context of identity.

Theories of Identity

Modern identity emerged in Western Europe and Great Britain during the fifteenth and sixteenth centuries, culminating in the humanist figure of the free and autonomous individual in the eighteenth century era of Enlightenment. Humanist notions of subjectivity, as Elizabeth Grosz points out, provide the subject with unquestioned certainty, truth, and presence. Central to the contemporary, everyday and 'common sense' understanding of identity and selfhood is the fifteenth century notion of René Descartes (1596–1650) known as the *cogito ergo sum* ('I think, therefore I am'), which operates as a first principle for the idea of an 'I' as a conscious and reasoning individual in which thought or 'mind' is given preference over corporeality and bodily sensations. Significant in western thinking on identity, this notion of the autonomous, coherent, unified self was extended and solidified by numerous writers and thinkers, including John Locke (1632–1704) who posited the liberal and free individual; Jean-Jacques Rousseau (1712–78) who argued for a human individuality grounded in nature; and Immanuel Kant (1724–1804) who equated selfhood and consciousness.

However, although the notion of the unitary, autonomous, coherent and essentialist subject remains the commonsensical and everyday understanding of identity, theories, and philosophies of identity emerging throughout the twentieth century have put this notion of identity in question, and rejected the idea of subjectivity as a self-contained being. Several competing theoretical positions lead to the decentering or rejection of the humanist subject in poststructuralist queer theory. Marxism refuses recognition of the subject as a 'conscious' subject,

attributing it a 'false consciousness' in a capitalist social–economic system. The structuralist–marxist critique of subjectivity undertaken by Louis Althusser (1971) questions the integrity of the universal subject by showing it to be bound by its interpellation through institutions and ideology. In Sigmund Freud's psychoanalytic critique, the discovery of the unconscious fragments the subject at the point of its consciousness, and suggests that one's identity is bound by one's desires. Jacques Lacan (1977), extending Freud's theory through structuralist semiotics, makes clear that desire and subjectivity are inseparable: there cannot be a subject without a concept of desire – in coming in to language in the Symbolic, identity is separated from the unconscious Other, thereby resulting in a subject that is always described as 'split.' For Lacan, desire is constituted by lack as a result of the inability to reacquire the preoedipal and premirror stage *jouissance*, hence the subject (always unsuccessfully) seeks fulfillment by trajecting desire toward the *objet petit a* (another object-of-desire) be it a sexual object, a 'personal goal,' or otherwise – a desire that can never be satiated. Thus, the psychoanalytic approach always posits a subject which is always 'in process.'

From the 1960s and 1970s, the Enlightenment humanist notion of subjective identity was put further in question by theories of social, cultural, and discursive constructionism in which the subject is not born or the result of nature, but produced within the environment, language, and sociality. Both building upon and rejecting the dominant psychoanalytic critique of identity, this anti-subjective structuralist and post-structuralist criticism has become the prevailing understanding in critical and cultural theory of identity, although has by no means resulted in a wholesale rejection of the Enlightenment figure of the subject in contemporary everyday and pedestrian thinking about identity.

Significant among constructionist approaches to identity is the work of Michel Foucault (1926–84) who posits identity not as an 'effect' of power, disciplinarity, and biopolitics, which include processes and techniques of surveillance and normalization. Foucault's theory of subjectivity as a *form* constituted in and by discourse approaches a poststructuralist position – and contributes heavily to it. For Foucault, the subject is inculcated by and through the deployment of power relations, normalized variously in accord with regimentary, disciplinary, and biopolitical discourses. The humanist subject of Descartes as self-existent, coherent, and consciously active is rejected in a Foucauldian position in favor of the disciplining of bodies made 'docile' (Foucault, 1977) and conforming. Three modes of objectification of identity in Foucault's work can be identified: the first is dividing practices, such as the isolation of 'the mad' in asylums. Second is the Foucauldian concept of 'scientific classification' arising from modes of inquiry given the discursive status of science. In the context of Web 2.0 digital media environments, we might similarly refer to this as profile categorization or 'naming'; a discursive practice which in Foucauldian analyses of identity plays a pivotal role in the inculcation of the subject *as* subject. Finally, there is 'subjectification,' the processes 'of self-formation in which the person is active' albeit with conformative regimes. Biopolitics is a technology of power that both analyses and constructs not the individual subject, but whole populations as subjective through a range of techniques from statistical measurement to health promotion to immigration controls; its governance mechanics are generally but not always located within the administration of the nation-state. Emerging slightly later than discipline in the second half of the eighteenth century as part of the further developments of governance for larger states, it is a technology that can be a useful tool for understanding the relationship between online environments and identity in that the use of digital communication tools results in the capacity for large-scale data collection on the level of whole populations, thus producing certain forms of identity which are normalized not along the disciplinary normal/abnormal distinction but in the terms of distance from a norm along a distribution curve of normativities (Foucault, 2007).

Building on the work of Foucault (although at times deploying Lacanian psychoanalysis), Judith Butler projects one of the most useful poststructuralist theorizations of identity by suggesting that the subject is constituted by repetitive performances in terms of the structure of signification that produces retroactively the *illusion* of an inner subjective core (Butler, 1990). Identity/subjectivity becomes a normative ideal rather than a descriptive feature of experience, and is the resultant effect of regimentary discursive practices. The subject, then, is performatively constituted by the very 'expressions' that, from a humanist position, have been considered the essential subject's subsequent conscious actions. As Butler notes, 'performativity' is not a singular, deliberate, and self-conscious act, but a reiterative and citational practice in which discourse produces identity as its effect in a form which stabilizes through repetition over time, but remains always at risk of showing itself up by the constitutive failure of genuine repetition. For Butler, discourse remains the key to subjectivity; but her extension of Foucault's thesis provides a less universalized account of subjectivities. Butler has taken great pains to point out that her theorization of performative identity does not suggest a conscious performance with all the agency that would be accorded in a humanist account (Butler, 1993).

Although Butler's work is complex and wide-ranging, there are four significant elements in understanding how identities can be understood as performative. First, following Nietzsche and Foucault, Butler articulates identity not as static or essentialist but as an *effect* of a performance that is constituted in and through language, discourse, and culture. Secondly, identity or selfhood is performed reiteratively as a process 'in accord' with a discursively given norm or a set of norms. Through the ongoingness of performance of the cited signifier or norm, a subject stabilizes himself or herself by producing a fiction of a fixed, inner, essential selfhood, or subjectivity – this fiction is a retroactive production of an illusion that there is a core identity behind the performance, a doer behind the deed – actions and performances do not stem from an inner essence but retrospectively constitute it. Thirdly, identity is constituted in discourse and can therefore be reconstituted or reconfigured differently in the encounter with different, alternative, emergent, or imaginative discursive arrangements. Finally, the motivation to be articulated as a coherent intelligible self stems from a cultural imperative to perform identity in terms of coherence, intelligibility, and recognizability in order to be permitted belonging and social participation, thus to forge a sense of self and belonging across an array of identity categories or coordinates – which include common axes of

discrimination such as gender, ethnicity, ability, and age, but might also be comprised of spurious experiences that are less easily categorizable and less well demarcated in an identity/difference dichotomy. For Butler, then, identity is manufactured in and through the languages, concepts, and ideas available to a culture at a specific point in time, meaning that in today's media- and digital-saturated western environment, certain practices of communication that are both off-line and online, both transmissional and ritualistic in their use, are highly implicated in the production of identity.

Butler's theoretical account of the performativity of identity puts both liberal-humanist essentialism and postmodern radical constructionism into question in favor of a Foucault-driven anti-foundationalist view of identity. The essentialist/constructionist debates are best delineated by Butler, who points out (1993: p. 8) the ways in which both sides have missed the point of deconstruction by refusing "the constitutive force of exclusion, erasure, violent foreclosure, abjection, and its disruptive return within the very terms of discursive legitimacy." While the essentialist position, for Butler, is both Cartesian and phallogocentric, constructionism leaves a residual essential subject on which cultural construction (sometimes referred to as 'socialization') overlays ideological concepts and gets them 'wrong.' She also identifies an overly radical constructivist position which not only dislodges the essential subject but also makes a claim to the constructive composition of all objects. She argues (1993) that this position understands culture as exhaustively constituting identity, ultimately creating and determining it in its naming that which it names, or, rather, it is that kind of transitive referring which names and inaugurates at once. Her answer to these two positions is to dislodge both the fixity and foundationalism of essentialist views and the determinism of the radical constructivist position in favor of performativity as establishing the subject as *effect* and, more specifically, viewing the subject as a process that stabilizes to produce the effect of fixity over time.

Cultural theorist Fredric Jameson (1985) is correct to point out that in a postliberal or postmodern culture not only is the individual bourgeois subject a thing of the past, but also that in a poststructuralist position "it *never* really existed in the first place; there have never been autonomous subjects of that type." For Jameson, the autonomous subject has been a cultural mystification which sought to persuade people that they actually have an individual and autonomous personal identity. However, despite such theoretical decentering of the subject, the Enlightenment liberal-humanist figure of the coherent essentialist subject remains the key means by which identity is defined in everyday and pedestrian practice in both online and off-line contexts. The mainstream cultural production of identities that are heavily focused on limited practices of individuality manifests within a postmodern framework of problematized self-knowledge, valorized human potential, destabilized social hierarchies, and flexible identity categories. Despite these conditions, a unitary self remains one of the most basic realities premised on notions of substantive identity and perpetuated by dominant medical, legal, and political discourses. A tension thus arises between essentialized notions of identity and more poststructuralist understandings of its constitution; and this discord is reproduced across multiple institutional and theoretical practices, including social network use and its theorization. In effect, identities develop, stabilize, and fracture across multiple, often incongruous discourses, experiences, and technologies.

Virtual Realities and Virtual Identities

Theories of the role of the Internet in critiquing and addressing identity (or in providing opportunities to move beyond Essentialist notions of subjectivity) are many, and emerged in the mid- to late-1990s. Several theoretical approaches that focused on cyberculture sought to examine the role of new technology on the formation of subjects, subjectivity, and identity. The reciprocity of influence between computers and humans gave rise to the emergent field of enquiry known as cyberculture. Early uses of the term were often quite narrow in their definition or utopian in their intent. For example, Mark Dery (1992) focused on the subversive and oppositional forms of identity that emerged in early Web 1.0 Internet use, while Howard Rheingold (1993) presented an idealistic conception of harmonious online virtual communities.

Cyber theorist Sherry Turkle (1999) contends that theories of self that consider the self as largely decentered, such as Mark Poster's approach to anonymous speech on the Internet as being without limits, gender, religion, ethnic, or national requirements, face a constant disjuncture between theorizations of unitary identity as an illusion and, on the other hand, lived experience in which the unitary self is popularly felt to be the most basic reality (Turkle, 1999). The Internet, particularly in the form of early Web 1.0 text-based chat rooms, offered technological objects-to-think-with that allowed postmodern understandings of identity as multiple and always decentered to be explored by facilitating experiential understandings of the illusion of a unitary self (Turkle, 1997). In the often anonymous environments of text-based virtual-reality multi-user domains (MUDs and MOOs), one could invest in and construct oneself in a simulated environment understood to be free of embodied territorial constraints. In effect, one could challenge the idea of unitary identity through the willful experimentation with identity categories. Turkle (1999) gives the example of a user known as Case: a male industrial designer who maintains a 'Jimmy Stewart versus Katharine Hepburn dichotomy' of online personae to exemplify a notion of identity as distributed and heterogeneous, facilitated by text-based online forums. For Case, presenting himself as 'Katharine Hepburn types' online is seen to allow for an externalization of an aspect of himself that he is unable to fully explore in 'real-life' masculine embodiment. Examples of identity play such as Turkle's are problematic for two reasons. First, they maintain the real-world/virtual-world dichotomy that is pervasive in much cyberculture and mainstream discourse of online communication and social networks. The dualism of the human and machine boundary is far less discernibly established and may be more usefully described in terms of a series of performances across a socio-technical network wherein the 'interface' consists of a multiplicity of aligned and dynamic encounters between different configurations, including as persons and machines. Examples of identity play and identity fluidity through the utilization of

online sites, such as Turkle's exploration of Case, rely on acts self-stereotyping in order to maintain online personae. While not explicitly explored, Case's use of film icon stereotypes highlights the theatricality of his online personae, which fall short of the performative approach to identity espoused by Butler. Although celebrated as examples of identity fluidity, these online performances are thus theatrical and maintain the consistency and coherence of off-line embodied identity, rather than being a genuine challenge to prevailing everyday concepts of unitary and essentialist identity.

Resulting from the increased materiality afforded to multiple and divergent identity category expressions online, it is understandable that early cyberculture theorists postulated a future where the unbridled expansion of identity categories might lead to a utopian world wherein 'identity' could be extricated from foundationalist principles. What the success of a social network like Facebook might suggest, though, is that all of the identity 'choices' presented online necessitate a mediatory tool if one is to readily articulate a coherent and intelligible self and present a picture of selfhood for the scrutiny of a plurality of gazes. However, despite the shift from early anonymous text-based communication, Turkle's early work remains helpful in framing the possibility for some technologies to impact 'older centralized ways' of considering identity. Within Turkle's framework, objects anchor experience and memories. So while her example of Case's online personae is problematic, it is nevertheless helpful in highlighting how objects contribute identity construction.

Digital Communication, Social Networks, and Relational Identity

Many current approaches to understanding the relationship between Web 2.0 environments (characterized by YouTube, Wikis, and – most particularly – the social networking sites such as MySpace and Facebook) are grounded in the notion of social network use as emanating from a coherent a priori self (e.g., Donath and Boyd, 2004; Liu, 2007; Livingstone and Brake, 2010), despite the powerful arguments put forward within poststructuralist theory that 'show up' the possibility of a genuinely coherent, intelligible, and unified self (Cover, 2012). Although, from a Foucauldian, Butlerian, and post-structuralist perspective, all performance of identity is relational in that it requires the citation of discursive categories of identity, is performed for recognizability, and operates within constructed 'truth regimes,' the relationality of Web 2.0 digital media and communication environments further implicates identity within a network of others in which mutual surveillance can be understood as the key to identity articulation in an online environment – a surveillance in which online users are always from the beginning complicit. Identifications and categorizations online thus function to assert individuality and often involve a process of stereotyping that allows group members easily to distinguish self, others, and group members (Buckingham, 2008).

Much of the early current literature on social networks describes the conscious, willful *use* of an online technology like Facebook. This position is coterminous with early cyberculture work that maintains the 'real-life' versus 'virtual self' dichotomy and, effectively, the reification of a 'real' identity out of which the 'virtual self' could be voluntarily created. However, several more effective and theoretically-grounded approaches to understanding the relationship between social networking and identity build on the work of Erving Goffman (1959). Goffman's dramaturgical analysis of identity performance offers a convenient framework for the analysis of identity expression only. Social networks, situated in the broader cultural practices of individualism, self-narration, and interconnectivity, offer many parallels to Goffman's work. Goffman points to the self-conscious presentation of selfhood in different contexts as well as the symptomatic elements of performance – that is, what people 'give' and what they 'give off.' An individual, for Goffman, performs identity through both of the forms of intentional and unintentional expression. In the context of social networking and other online forms of relationality, Goffman's approach to the performative expression of selfhood points to the ways in which the online acts of typing, updating, uploading photos and videos, and other activities serve as both intentional and unintentional forms of expression to varying degrees depending on intent and context.

One way in which we can understand the use of social networking through Goffman's concept of 'everyday performance' is through what Hugo Liu (2007) refers to as 'taste performances.' A taste performance is the textual statement of a socioeconomically and aesthetically conditioned taste that conveys 'prestige, differentiation, authenticity, and/or theatrical persona.' Within Liu's Goffman-derived framework, taste performances are voluntarist acts whereby performers are broadly aware of the impressions they foster among viewers. That identity can be cultivated by and through taste performances points to the ways in which the expression of tastes, which can best be seen today through Facebook 'likes' and the listing of particular favorite films, activities, television, and music in Facebook profiles, are not only relational in that they are deliberately communicative but also relational within a neoliberal consumer culture that has brought consumer choices to bear upon the production of self-identity.

Liu's exploration of taste performances on social networks is helpful in two ways. First, by framing identity coherence as a strategically-expedient outcome, Liu highlights a tension that social network use can produce – that is, between expressive coherence and the potentially destabilizing role of an audience that scrutinizes (in)consistencies in the ongoing performance of selfhood through the linearity of taste. Secondly, Liu draws attention to identities increasingly constructed around a growing number of possible consumptive choices – effectively, individual identities are increasingly enfolded in a collapsing between consumption and the performance of taste.

These conclusions, while helpful, are problematic in that Goffman's framework reinforces the notion of (and hides the constructedness of) a core essentialist self from which tastes are seen to emanate. Effectively, taste performances are presented as *external* representations of identity expression, thus maintaining the power mechanism of categorization in identity formation. Taste performances, however, are not merely representations, but can be seen to construct the identity or self [that is] perceived to be making those decisions (Cover, 2012). The stability of a retroactively constituted performative identity

is thus problematized when narrowing boundaries of identifications produce more opportunity for misrecognition and require greater online 'identity work' (performativity) to sustain. Additionally, the implication of Liu's analysis is that as the available system of possible taste performances becomes more complex, so too do identities as they increasingly combine these performances in new and divergent ways.

While Liu explored the expressive potential of listed cultural interests online as taste performances of identity, Donath and Boyd (2004) have investigated how social network users' friending behaviors and the public display of friend networks contribute to identity portraiture and performance where networks of displayed connections between related groups form integral parts of self-identity. They point to the way in which we can understand two features of the architecture of a social network such as Facebook, however, as having implications for the stable performance and coherence of identity – connections that are both 'unnuanced' and 'decontextualized.' Connections can be viewed as unnuanced as there is little, if any, distinction made between, say, a work colleague, a lover, close relatives, and others who are grouped under the category of 'friend.' Connections are also largely decontextualized in that one's network is usually viewable in its entirety. Donath and Boyd (2004) frame identity work on social networks as willful, representational repair work to 'save face.' This implies that 'incompatible texts' – taste performances and connections – stem from an incongruence among external, representational texts that operate from an a priori core. External performance, 'giving off,' is destabilized by contestatory taste performances or incompatible connections, but the prediscursive self from which these performances emanate remains intact.

From a Foucauldian perspective, power mechanisms operating through social networks rationalize a sense of a prediscursive self and mask the unconscious performativity of identity. These mechanisms can be said to include the discursive conditioning of identity categories, self-recognition through a process of interpellation and the internalized practice of self-surveillance, as well as the active (and relational) surveillance of others. Drawing on Judith Butler's theories of identity performativity, Rob Cover (2012) investigated the distinction between profiles and relational groupings in social networking sites, indicating the ways in which the confessional forms of self-disclosure made by users through profile selections and statements conflict at times with the surveillance of social networking 'friends' whereby surveillance is no longer merely disciplinary and modeled on the panopticon but has shifted into a network morphology. However, the fact that social networking sites record a person's history of statements, tastes, relationships, friendships, confessed activities, and responses from others produces an 'archive effect' which is precisely that which opens up the possibility of identity incoherence in cases where the repetition of tastes, performances, relationalities, and articulations is not repeated or consistent with the user's past.

Online Identity and New Technologies

There are three elements to understanding online identity, as divergent from Enlightenment forms of identity experienced prior to the age of the Internet, that emerge from a Foucauldian–Butlerian poststructuralist approach: new relationalities; new forms and processes of self-categorization; and new forms of surveillance. In each case, these indicate the ways in which digital technologies have, on the one hand, opened up new mechanisms for the coherent performance of the (mythical) unitary and autonomous self and, on the other, point to the constructedness of identity.

Relationality

According to Turkle (2011), online networks promote multiple, simultaneous, and mediated interactions that reduce critical reflection and can lead to isolation anxiety; the more we control these interactions, the less we are able to cope with the uncontrolled situations, like a conversation. If an object such as Facebook affects us psychologically, then being connected yet experientially alone means that identity construction is becoming increasingly relational, built on high quantity, low quality interactions. While Facebook reiterates these relations as stemming from a unitary self, and with clearly delineated categories of identity (of say gender), the unconscious performative 'self' in this environment is less critically aware, more connected, more anxious, and more concrete (and controlled) in both the expression and feedback of aspects of self. This further complicates the construction of a unified cohesive self as identity becomes increasingly contingent on technology that simultaneously catalyzes and problematizes identity coherence.

Categorization

One of the central mechanisms in Butler's (1990) analysis of performative identity is the role of categories. The construction of identity categories – gender, race, nationality, and so forth – as coherent and stable, produces identity within regulatory regimes and a reification of identity categories as normative statements. For example, the stability and coherence of categories such as 'male,' 'woman,' and 'heterosexual,' rely on normative discursive frameworks that regulate the field of their intelligibility. While some categories like 'woman' or 'heterosexual' achieve their stability and coherence within the heterosexual matrix and are predominantly naturalized and unproblematically invoked, other identity coordinates like race, class, ethnicity, or nationality are less stable and less accepted. What is important about the mechanism of categorization is that it functions within a representational discourse where categories operate as domains of exclusion and produce 'coercive and regulatory consequences' of their construction.

Social networking sites can be understood as sites through which identity categories are effectively performed. Most social networks are built around individual user profiles. The first step for the creation of any given profile is to engage in a form-filling process; a process that now occurs countless times over one's lifetime. As Butler notes, the repeated act of 'naming' identity categories is an institutionalized form of regimentation that produces and circulates normativities by requiring the construction of identity in accord with dominant discursive arrangements like heteronormativity and naturalized principles of sexual difference. As initial profile creation frequently

involves submitting gender and age, and often includes sexual preference, nationality, and relationship status, it can be viewed as a set of acts that are repeated over time and produce a mythical reality.

Therefore at a deeper level of analysis, the ritualistic practice of form filling, under a new pretense of social network profile creation, cannot be merely the conscious process of 'biographization' (Livingstone, 2008). Rather, the repeated 'naming' of each category operates as a performative that stabilizes both the category itself and its appearance of coherence within an individual. By continuing to operate within the same framework, though, expansive identity categories remain representational social fictions to which one can be held accountable. As Butler (1990) notes, no one category can be fully decontextualized from other axes of power relations – such as race, ethnicity, and class – that together constitute identity and prove that no singular facet or notion of identity can stand for the subject altogether. Thus as the sophistication of differentiation within a system of categorical axes of power increases, so too does the individual requirement to work at their adherence to categorical norms. Social networks therefore exacerbate an existing tension between reified notions of singular identity and the impossibility to achieve complete adherence to categorical norms.

Despite this discord, an effect of this perpetuation of the mechanism of categorization is the normalization and stabilization of identities. As the reiteration of norms stabilizes over time to produce the effect substantial identity, power mechanisms of social networking are able to effectively coordinate the articulation of a coherent identity. As noted, form-filling, profile creation, and network management function to sustain gender as a naturalized foundational 'fact' of one's identity. For example, in order to join Facebook one must select a sex. While instances of misrecognition or nonidentification within the sex binary may occur in individuals who do not self-describe as belonging to either category, the force of disciplinary power within medicolegal discourses maintains the fundamentalism of sexual difference. For a subject who is unable to comply with the 'available' categories, not only is there the off-line risk of being classified as a nonperson, but also the online risk of not being able to be a Facebook user.

Surveillance

However, the existence of categories does not in and of itself ensure that they are accepted. In Foucault's framework of a disciplinary society, dominant discursive arrangements are maintained through processes of exclusion and normalized identity practices that are in part sustained by the mechanism of surveillance. The abstraction of the panopticon offers a useful model to examine how visibility imposes on the observed (and the observer) a regime of self-inspection and self-control. However, in the context of the increased visibility of self, social networks' role in surveillance becomes one that is grounded in the nodal points of a relational network rather than the centrality of the panopticon, even though both encourage self-surveillance as the normative function of disciplinary power (Cover, 2012). Specifically, Facebook can be viewed as a tool par excellence for the stabilization of a coherent identity because it expedites the stylized repetition of highly mediated identity categories and 'templates' under increased surveillance. However, this process is by no means unilateral or uncomplicated, and effectively produces the conditions for its own destabilization. We can view surveillance as an internalized process that operates alongside externalized practices of surveillance. When one participates in a social network, the implicit requirement is to produce knowledge about oneself in the form of profiles, statuses, photos, and so forth. Each instance of this production increases one's visibility within the disciplinary system, and potentially amplifies the effect of that system on the construction of identity.

Identity Futures Online

Inasmuch as users are now cognizant of the problem of the multiplicity of the networked audience in terms of who is accessing information, how it is being utilized, the extent to which one might be required to explain oneself to different friends and different groups (Lewis and West, 2009), it may well be the case that the 'identity project' of social networking has an uncertain future. Recent questions and public commentary on digital privacy and the utilization of social networking accounts by third parties who construct marketing profiles based on portions of identity and taste information and usage have both generated and revealed a number of anxieties around social networking as a tool for identity performativity. One question that remains open for future investigation is whether or not users will grow tired of the additional 'identity work' caused by the potential incompatibilities between the friendship network, the constructed profile and any distinction between online and off-line performativity. Or, on the other hand, if this is the start of a new discursive framework in which a culture begins to embrace the complexity and incoherence of identity, leading to the fruition of post-modern and post-structuralist thinking of the self.

See also: Human–Computer Interfaces; Online Dating; Religion and New Media; Social Identity in Social Psychology; Social Media.

Bibliography

Althusser, L., 1971. Lenin and Philosophy and Other Essays (B. Brewster, Trans.). North London Branch, London.

Buckingham, D. (Ed.), 2008. Youth, Identity, and Digital Media. MIT Press, Cambridge, MA.

Butler, J., 1990. Gender Trouble: Feminism and the Subversion of Identity. Routledge, London; New York.

Butler, J., 1993. Bodies that Matter: On the Discursive Limits of 'Sex'. Routledge, London; New York.

Cover, R., 2012. Performing and undoing identity online: social networking, identity theories and the incompatibility of online profiles and friendship regimes. Convergence 18 (2), 177–193.

Dery, M., 1992. Cyberculture. South Atlantic Quarterly 91, 508–531.

Donath, J., Boyd, D., 2004. Public displays of connection. BT Technology Journal 22 (4), 71–82.

Foucault, M., 1977. Discipline and Punish: The Birth of the Prison (A. Sheridan, Trans.). Penguin, London.

Foucault, M., 2007. Security, Territory, Population: Lectures at the Collège de France, 1977–78. In: Senellart, Michel (Ed.), (G. Burchell, Trans.). Palgrave Macmillan, Hampshire.

Goffman, E., 1959. The Presentation of Self in Everyday Life. Doubleday, Garden City, NY.
Jameson, F., 1985. Postmodernism and consumer society. In: Foster, H. (Ed.), Postmodern Culture. Pluto Press, London, pp. 111–125.
Lacan, J., 1977. The Four Fundamental Concepts of Psycho-Analysis (A. Sheridan, Trans.). Penguin, Harmondsworth.
Lewis, J., West, A., 2009. 'Friending': London-based undergraduates' experience of Facebook. New Media & Society 11 (7), 209–1229.
Liu, H., 2007. Social network profiles as taste performances. Journal of Computer-Mediated Communication 13 (1), 252–275.
Livingstone, S., 2008. Taking risk opportunities in youthful content creation: teenagers' use of social networking sites for intimacy, privacy and self-expression. New Media & Society 10 (3), 393–411.
Livingstone, S., Brake, D.R., 2010. On the rapid rise of social networking sites: new findings and policy implications. Children & Society 24 (1), 75–83.
Poster, M., 2006. Information Please. Duke University Press, Durham, NC.
Rheingold, H., 1993. The Virtual Community: Homesteading on the Electronic Frontier. Addison-Wesley, Reading, MA.
Turkle, S., 1997. Multiple subjectivity and virtual community at the end of the Freudian century. Sociological Inquiry 67 (1), 72–84.
Turkle, S., 1999. Cyberspace and identity. Contemporary Sociology 28 (6), 643–648.
Turkle, S., 2011. Alone Together: Why We Expect More from Technology and Less from Each Other. Basic Books, New York.

Ideological Constraint: History and Current Status of the Concept

John P Connolly, The University of Texas at Arlington, Arlington, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Fifty years ago, Converse (1964) introduced the principle of ideological constraint. While references to this concept are still prevalent in the political science literature, its import largely rests on one's acceptance of Converse's classic formulation of political sophistication, about which there has been a voluminous literature and persistent debate. Here the origins and conceptual foundations of this concept are reviewed to assess its precise relevance to what it means to be politically 'sophisticated.'

My own belief is that many philosophical difficulties and controversies arise from insufficient realization of the differences between different kinds of knowledge, and of the vagueness and uncertainty that characterizes most of what we believe ourselves to know (Russell, 1948: 439).

Introduction

Reflecting on four decades of scholarship on political sophistication, Robert Luskin claimed that use of the term ideological 'constraint' was 'mercifully fading.' This observation should have surprised many coming from a scholar who was once among the most enthusiastic promoters of the concept, which he had once regarded as the *sine qua non* of political sophistication, as well as the one scholar who attempted its formalization. He noted that various other terms were now being used in place of 'constraint,' such as 'expertise,' 'cognitive complexity,' 'ideology,' 'information,' or simply 'knowledge' (Luskin, 2002: 281).

It has been observed elsewhere, however, that the captivating power of certain ideas can long outlive their value (Keynes, 1997: xii). One question at the outset of any discussion on the topic of ideological constraint then is whether the whole concept in question is a defunct notion, kept alive only by scholarly habits. If so, then obvious questions arise given that so much scholarship has been invested in the topic. For example, where did the concept of 'constraint' come from? Further, when, how, and why did it falter? Last, have more useful conceptions been developed in its place? Or is, what has been called political 'sophistication' reducible to plain 'knowledge'?

To preview the conclusions drawn here, it is agreed that the concept of ideological constraint is indeed anachronistic though this does not appear to be as widely recognized as might be intimated from Luskin's claim. Second, it is argued that the alternative conceptualizations that have been advocated in its place are also deficient. This is largely because a close analysis of these ideas reveals them to be little more than variations on the earlier theme and, despite much apparent debate, do not challenge the conceptual foundations which permeate that theme. Last, while it is tempting to give up on the whole idea of political sophistication and adopt less evocative terminology, such as the more prosaic term, 'political knowledge,' it is held here that there is indeed value in developing a conception of the 'sophisticated' over and above that of the 'knowledgeable' citizen. The final section lays out some of the basic requirements that any such new formulation of political sophistication should meet.

The crux of the argument presented here is that the problem with the classic formulation of political sophistication is twofold: first, it requires one to take a dogmatic position on a long unresolved philosophical problem concerning the definition of knowledge; second, it does not address the development and character of goals and values. It is maintained here that both issues need to be addressed for any new formulation to enjoy wider acceptance than its predecessor.

The Conceptual Origins of Ideological Constraint

The concept of ideological constraint dates back to Phillip E. Converse's (1964) seminal essay on political sophistication. If Kenneth Arrow's (1951) 'Social Choice and Individual Values' was the 'big bang' of social choice theory (a comment attributed to Amartya Sen by Poundstone (2008: 57)), Converse's 'The Nature of Belief Systems in Mass Publics' was the analogous foundational study in the field of political behavior. In neither case is it strictly true that what followed came out of nothing. Just as Arrow's findings had been anticipated by others, Converse's formulation of political sophistication had been anticipated in a number of earlier studies (Campbell et al., 1960; Berelson et al., 1954; Lippmann, 1922; Bryce, 1904). However it was not till Converse (1964) that the concepts of political sophistication, ideology, and constraint, were interpreted as essentially referring to the same thing. Opining that "A term like 'ideology' has been thoroughly muddied by diverse uses," he defined one's 'belief system' as a "configuration of ideas and attitudes in which the elements are bound together by some kind of constraint or functional interdependence" (Converse, 1964: 207).

The analogy with Arrow's famous work holds further in that Converse's formulation also led to a paradox of sorts. In everyday parlance, one may distinguish between the 'sophisticate' and the 'sophist,' between one who undertakes the honest process of factual acquaintance versus one who operates completely in rhetorical space, between a truth-seeker and one whose mind is closed to new discovery. Yet such distinctions

are blurred when the 'sophisticate' is equated with the 'ideologue.' Thus Luskin, arguably the leading disciple of the Conversian tradition, was led to assert that "political sophistication has nothing to do with acuity … John Birchers and Trotskyites alike may be highly sophisticated in this sense. Sophistication is a matter of how much and how a person thinks about politics, not what" (Luskin, 1987: 864).

Whether this *reductio ad absurdum* is actually what Converse intended is not absolutely clear. He had, for example, discussed other attributes of a political belief system (PBS), such as its 'range' in addition to the 'centrality' of the political objects that constitute it. But these concepts were quickly passed over with the core of Converse's discussion focusing on ideology and issue constraint. Thus Luskin was only following Converse's lead in giving overwhelming emphasis to the ideological constraint principle and carrying the concept's implications through to their logical end. Likewise, Luskin's (1987: 882) contention that the whole issue of factual truth was traditionally given 'shorter shrift' is wholly consistent with the classic formulation of political sophistication, even if this point is frequently overlooked both by ostensible proponents and opponents of the Conversian tradition.

In Converse (1964), the word 'constraint' was used in two senses, one psychological, the other social. The first referred to what Converse called the 'logical' or 'quasi-logical' connections between the beliefs that comprise one's belief system, while the second referred to the amenability of one's belief system to socially diffused cues. Converse appeared to argue that the distinction between the two was unimportant as a practical matter. As such, his formulation of a belief system was not new, being a blend of the coherence theory of knowledge and pragmatism, compatible with the respective perspectives of Hegel and John Dewey: in short, that no one belief has any intrinsic credibility on its own account but derives all its credibility based on its relation to other beliefs that, taken all together, constitute a closed system.

Yet if this conception of knowledge was not new, it was still *dogmatic*, since the whole question of how one defines knowledge is a long unresolved problem in traditional philosophy, and there are other definitions that accord more importance to data and facts. As such, there are aspects of the Conversian formulation that should have seemed objectionable at the outset. Why, for example, should one accept a formulation of political sophistication that has, as maintained by Luskin, 'nothing to do with acuity'? What is its justification as a dependent variable worthy of study in political science? Further, what normative value does such a formulation offer? As will be shown, a close reading of Converse's own empirical examples of emerging social movements tended to undermine the whole significance of ideological constraint as an analytically important concept.

Indeed this is where the earlier analogy with Kenneth Arrow's parallel work in the field of social choice begins to break down. For while one was compelled to accept Arrow's arguments due to his strictly logical analysis, there is no such obligation to accept the classic formulation of political sophistication, which is an empirically derived conception. Its acceptance depends crucially on its relevance, coherence and, further, it is emphasized here, on its normative value. The rest of this analysis explores whether this concept has ever met these criteria, and, if not, the lessons that could lead to a better formulation.

Reactions to Converse's Formulation of Political Sophistication

Following Converse (1964), a vast literature on the topic of political sophistication ensued in large part reflecting a desire among scholars to disprove implications they found repellant. However, the culmination of this long-running debate was that research and conclusions began to diverge. If there is less overt debate on the topic today it in large part reflects the fact that those in opposing camps have adopted different terminology so that they often talk past each other, the effect of which may create a false picture of consensus. For example, the topic has, to an extent, been rebranded in terms of political 'competence,' rather than political 'sophistication.' Similarly, Converse's use of the terms 'beliefs' and 'belief systems' is often discarded in favor of the term 'attitudes.' By such means, an uneasy peace gradually emerged between a revisionist camp and those more faithful to the old religion, with tensions between the two only occasionally spilling over (e.g., see Lupia, 2006).

It is not the case then that there has been no debate over Converse's (1964) claims. Rather, and notwithstanding the terminological changes just noted, the debate has been almost exclusively concerned with claims concerning the *distribution* of Converse's political sophistication variable, with little or no challenge to the ideological 'constraint' principle which was at the core of the original formulation. In fact, the direction which ostensible critics of the Conversian tradition took in turn led to yet cruder forms of this same principle and, ironically, depictions of the ordinary citizen that are even less flattering. Meanwhile ostensible defenders of the old religion have added confusion to the debate by focusing on the factual awareness of the citizen, apparently unaware that this was never part of the original formulation.

An ironic consequence of these trends is that it is now the disciples of Converse who, by measuring factual knowledge, offer one possible way out of the paradox noted earlier in which all ideologues are indistinguishable, while it is those most vocal critics of the Conversian school of thought who arguably perpetuate the notion of ideological constraint, albeit with different terminology. The means by which the latter is achieved are diverse, reflecting the voluminous reaction to Converse, but a common pattern is discernible. This is a consequence of the exclusive concern among critics to dispute Converse's contention that the distribution of political sophistication is confined to a very small group of citizens, one implication of which is that control of the polity rests on the mass public's whimsical tendencies.

The most common means to refute this conclusion has been to adopt a different starting point. Rather than beginning with a study of political sophistication at the individual level and attempting to infer the aggregate consequences, the alternate approach begins with the postulate that the polity is as functional as it could be and then proceeds to rationalize what must be happening at the individual level for this Panglossian state of affairs to exist. In its extreme form, this argument holds that the very stability of the polity depends on the fickleness of the

mass public. Here the mind of the average citizen is depicted as a *tabula rasa*, conditioned by how a few innovative thought leaders frame issues, as in Stimson's (2004) Humean account of idea association:

> The first time one sees two totally separate issue positions advocated by one party or candidate, they seem to be just that: two ideas espoused by one source. But as time passes and those two same ideas are repeatedly advocated together, for example, liberalism and choice, then they come to be related. As the process runs to ultimate conclusion, we begin to see the two as two facets of the same thing and think that they are inseparable.
>
> Stimson, 2004: 68–69

Stimson's basic thesis is that, while the public may seem to be fickle on the surface, there is still a 'smooth path from here to there' (Stimson, 2004: xvii). Stimson agrees that "the hope of democratic governance [goes] down the drain" with a citizen "who can't know his or her own preferences," but he reconciles this difficulty by contending that individual opinion is 'thermostatic.' Like the "homeowner with a thermostat need not understand the physics of home heating and cooling … All that is required is to notice 'too hot' or 'too cold' and to send the appropriate signal." Further, he argues, it is the framing of issues in this way – more, less, or same – that makes democracy possible (Stimson, 2004: 32–33).

These claims are in one sense remarkable. Half a century earlier, Kenneth Arrow had demonstrated the unavoidable difficulties involved in aggregating each individual's rational preferences into a noncontradictory collective outcome. In Stimson's account, such problems are held to be empirically irrelevant by the ability of elites to induce limited mechanical reactions from the citizenry. In short, while Arrow undertook the challenge of demonstrating how a democratic polity could operate on the basis of autonomous individuals ordering their own ordered preferences, Stimson's solution is based on the logic that only a small number of people pay attention to politics (Stimson, 2004: 15), and who effectively predetermine the possible orderings of everyone else's preferences.

A similar line of thought is pursued by Lupia (2006). There the question is formulated in terms of whether citizens are 'competent' enough to determine which electoral option best serves their interests. How one's interests and goals are determined – in short, the reference point for one's competence – is taken as given and, as such, goes unanalyzed, a topic Lupia opines is an "inherently normative … complex debate." The proposed mechanism for demonstrating competence then leans heavily on the use of heuristic references to guide one's choices.

In taking issue with the widely held contention following Converse that the mass public is politically unsophisticated, Lupia is especially critical of the means by which scholars have attempted to measure knowledge differentials across the public. In this respect, he distinguishes between declarative and procedural knowledge. "Declarative knowledge" refers to knowledge about facts and things (Matlin, 2005: 264) whereas "procedural knowledge" pertains to one's knowledge "about how to do something," such as "how to ride a bicycle" (2005: 129). It is the latter kind of knowledge, argues Lupia, which is relevant to the question of political competence. From a similar perspective, Prior (2009) has argued that assessments of competence from survey data are over reliant on verbal responses, and also pay too little attention to the motivation level of the respondent (Prior and Lupia, 2008).

Further, it is contended that the democratic ideal of the informed citizen is a fiction and that people are more free to the extent they do not have to engage in political deliberation, since "the more we demand of people, the less freedom they have, and freedom is among the most valuable products that democratic societies can produce" (Lupia, 2006). As such, there is no normative value in having any more knowledge than is necessary to exercise one's minimal civic duties. Knowledge over and above this minimum is superfluous to 'efficient' decision-making.

Such arguments have frequently borrowed accounts of decision-making from the field of social psychology and cognitive science, which have then been employed to give stylized accounts of political behavior, a practice which Krosnick (2002) calls 'psychological political science.' Such research could still in theory lead to 'extensions and revisions of the inspiring psychological theories' (Krosnick et al., 2010: 1289), but this appears to have rarely happened. As argued by Druckman et al. (2009), such theories often undergo an evolution within the social psychology literature while only a crude replica of the original formulation continues to be used elsewhere.

The reflexive resort to heuristics to explain political behavior is arguably a case in point. As Druckman argues, "the very point of heuristics is that they function as an alternative *processing* mode to rational decision-making … Because both are procedural, the difference between them does not depend on the outcome of the decision process" (Druckman et al., 2009: 493). That is, whereas heuristics are understood to be a 'computational' shortcut in the social psychology literature, they have instead been frequently interpreted as an 'informational shortcut' in the political science literature. Thus the contention that the use of heuristics bridges informational gaps is problematic on conceptual grounds.

Yet another way of escaping the unsettling implications of Converse (1964) has been to revive Robert Dahl's classic notion of pluralism. Here the key requirement for a smoothly running 'polyarchy' is that citizens have strong attitudes on at least some issues. In turn, the focus on 'attitudes,' as opposed to 'beliefs,' relegates the whole importance of knowledge differentials. Once again, one's goals and values are taken as given and go unanalyzed, serving only in determining the 'importance' of an attitude (Boninger et al., 1995). A kind of political equilibrium then ensues reflecting the public's diverse preferences, here via the interaction of a multitude of interest groups. Converse (1964), himself, had opened the door to this line of thinking with his discussion of 'issue publics' toward the end of his famous essay. Yet attempts to revive the classic concepts of pluralism and polyarchy are somewhat strange in light of misgivings expressed by Dahl, himself, in his later ruminations (Dahl, 2003, 1986).

The foregoing point concerning the use of social psychological theories in the political science literature is especially salient with respect to the concept of ideological constraint. According to Krosnick, Converse's original formulation was

inspired by then contemporary theories concerning cognitive dissonance and cognitive consistency (Krosnick, 2002: 191). In turn, one might depict the political 'sophisticate' as she or he who best performs dissonance reduction. However, the theory of cognitive dissonance has undergone substantial reformulation and qualification since it was introduced by Festinger (1957), with the process of dissonance reduction now understood to be triggered by threats to the self (Kunda, 1990). Yet the corresponding concept of 'constraint' in the political science literature has not undergone any parallel refinement that could potentially rescue it from its own internal ambiguities and contradictions.

A Formalization of Ideological Constraint

Luskin (1987) offered a formalization of ideological constraint which is noteworthy in several respects. First, the implications of this formalization are themselves surprising, in that they seem to make the connection between ideology and constraint somewhat ambiguous. Second, Luskin, arguably the most enthusiastic advocate of ideological constraint as the essence of political sophistication, was later to comment that invocations of the very principle he had attempted to formalize were 'mercifully fading' (Luskin, 2002: 281).

Expounding on the terminology of Converse (1964), Luskin's analytical framework is that of a PBS comprising various cognitions. As such, he defined a PBS as varying along three dimensions: size (the number of cognitions), range (the coverage of the political universe), and constraint, which is "the extent to which the PBS's cognitions are interconnected" (Luskin, 1987: 859). With respect to the individual cognitions, their value is solely in terms of their relationships to other cognitions, not whether there is any factual correspondence to the outside world. This could be viewed as a model of one strain in traditional Western philosophy that defines knowledge solely in terms of connections between beliefs, and not in the truth value of individual beliefs.

What distinguishes a cognition is its centrality; that is, its importance relative to other cognitions. As such, Luskin defines the centrality of the ith cognition in a PBS of N cognitions as follows:

$$m_i = \frac{1}{N-1}\sum_{j=1}^{N} c_{ij}\frac{1}{d_{ij}}$$

where $c_{ij} = 1$ if the ith cognition is directly or indirectly associated with the jth and is zero if not, and d_{ij} is the 'minimum number of intercognitive links to get from the jth cognition to the ith.'

Then Luskin defines the constraint, 'C,' of a PBS, as the proportion of all possible pairs of cognitions in which elements of the pair are directly associated:

$$C = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}$$

where $a_{ij} = 1$ if the ith cognition is directly connected with the jth and is zero if not.

In short, Luskin asserts that a maximally constrained PBS is one in which every cognition has a direct link with every other cognition. This is a somewhat surprising formulation in that it points to a mesh configuration as indicative of the constraint principle, in which all cognitions have the same centrality. As such, the idea of constraint, which notionally seemed to correspond to the concepts of efficiency and organization, turns out to be one whose characteristic feature is redundancy.

Yet what is perhaps more noteworthy here is that all cognitions are equally central in the mind of a sophisticate; in short, the political sophisticate appears to entertain no doubt whatsoever. Such a formulation seems questionable. What one comes to know about politics comes from all different kinds of sources and involves all kinds of different inferences. As such, the connections between some beliefs may come close to the status of deductive logic while others are likely to be much more tenuous. It seems questionable that any rational individual could possibly think about politics in the way implied by Luskin's model of the sophisticated mind; that is, be equally certain of every idea in one's head.

There is a related ambiguity here. Luskin's model suggests a direct trade-off between having a large, differentiated belief system, as opposed to a 'constrained' one. But it is not clear why the latter should be preferable as a rule. The appropriate level of cognitive abstraction in any given political situation must always be context-dependent, whereas it is implied here that thinking in static, abstract terms is always preferable.

What Did Converse Actually Demonstrate?

The debate following Converse (1964) largely centered on the question of whether the distribution of political sophistication – the essence of which has been understood to be 'ideological constraint' – is largely confined to a small segment of elite thinkers or is instead more widely dispersed across the mass public. Yet, regardless of this question, Converse's own analysis of the role of ideology in the case of two notorious social movements appeared to downplay its whole importance.

In his first example, Converse reviewed the emergence of the Nazi party in early 1930s Germany. His main motivation here, perhaps, was to address the whole question of how an informed citizenry could have let the Nazis take over the country. However, his discussion provides little insight here. One difficulty might be noted at the outset in the wider context of his formulation of political sophistication: If the essence of sophistication is in being ideological, and if ideology is, according to Luskin (parodying Gertrude Stein), "an ideology, is an ideology" (Luskin, 1987: 863), then on what grounds does one critique Nazi ideology? In fact, Converse, himself, ridiculed the whole importance of ideology in the case of the growing popularity of the Nazis, arguing that its supposed relevance 'verges on the absurd' (Converse, 1964: 254), and asserting that "Even had the clientele of the Nazi Party been of average education and political sophistication, there would be strong reason to doubt the degree to which prior awareness of Nazi ideology among its voters could be claimed" (1964: 254).

Converse reasoned instead that the German public was mainly responding to concrete proposals to improve its plight in economically desperate times, and no other political organizations were making such efforts to win support, especially in Germany's rural areas (Converse, 1964: 253): "The Nazis

promised changes in a system that was near collapse. Under comparable stresses, it is likely that large numbers of citizens in any society … would gladly support *ad hoc* promises of change without any great concern about ideological implications." Converse opined, in turn, that history might have been quite different and the key to this is that voters were open to other appeals; however, opposition party elites failed to reach out to the mass public and offer such alternatives. Would it have been better if the German public had been more socially 'constrained'? Converse offered no such argument and, in fact, the whole question seems inherently ambiguous.

While Converse's first empirical example argued that ideology was irrelevant in the case of the emergence of a social movement aimed at *bad* social ends, his next example – the emergence of the abolitionist movement in nineteenth-century America – appeared to show the irrelevance of ideology to the success of a social movement aimed at *good* social ends. In that case he argued that beliefs in abolition played very little role in developing the mass base of the Republican Party citing, again, other more concrete factors, such as the promise of free land in the West and the impact of the railroads. As such he held it doubtful that the ordinary non-Southerner in 1855 betrayed any concern for abolition (Converse, 1964: 250–251).

What these two empirical examples seemed to demonstrate was that ideology played little role among the mass public at the outset of social change, whether for better or worse, and that the ideas providing the impetus for social change originated from a relatively small segment of society. Yet even if this insight has wider application, it would still be inadequate as an account of social change since it overlooks the extent to which some ideas subsequently come to resonate with the public more than others. Do, for example, the ideas, beliefs, and values promulgated by the Nazis have the same cognitive status as those promulgated by the abolitionist movement? If it is true, as suggested by Stimson (2004), that the apparent connections between political issues are arbitrary and ephemeral then the whole idea of progress is arguably moonshine and the question of one's goals and values has no relevance to political sophistication. However, if progress is not illusory, then the nature of one's goals and values, not merely one's attitudes and beliefs, must be equally important.

Conclusion

To summarize the key points thus far: First, Converse's formulation of political sophistication was a purely empirical derivation. His justification for attaching any importance to the variable he discovered was solely in terms of considerable observed variability, claiming that "the variation in constraint as a function of sophistication or involvement is extremely large: There are no other discriminating variables that begin to separate populations so cleanly and sharply as these measures" (Converse, 1964: 233). No explicit concern was expressed as to the consequences of this variability and his empirical examples did not demonstrate any benefit in having a different distribution of ideological constraint across the mass public. As such, to the extent the variable he identified, what has since come to be known as 'political sophistication,' has been ascribed any normative value, a naturalistic fallacy of sorts has been committed.

Second, although there has been much apparent debate in reaction to Converse (1964), it has centered almost entirely on whether Converse's variable is more widely distributed than he initially claimed. As such the notion of ideological constraint, which was central to his formulation, has gone unchallenged. This fact is sometimes disguised by the use of different terminology; in particular, the terms 'competence' and 'attitudes' are often substituted for 'sophistication' and 'beliefs' respectively. Yet despite the diverse terminology, a common pattern is discernible in the reactions to Converse (1964). Whereas Converse began with an analysis at the individual level, the critiques of sophistication scholarship typically start with the assumption of a smoothly running polity and then propose a model of what could be happening at the individual level for this Panglossian state of affairs to exist.

Such critiques are in fact focusing on what Converse regarded as the social sources of constraint. But whereas Converse (1964) had suggested that the mass public was not socially constrained, those on the other side of the debate have proposed mechanisms by which it is. Clearly, such reactions do not then present a challenge to the constraint principle but, arguably, serve as a recapitulation of it. In turn, the critics of sophistication scholarship have pushed a normative conception of the citizen as one largely detached from the world of politics but who can perform his or her duties in an 'efficient' manner. In its extreme form this latter formulation holds that requirements for political participation and involvement are an obtrusion into civil society and therefore an impingement on one's freedom.

Third, since the ideological constraint principle remained substantively intact from the criticisms in reaction to Converse (1964), it follows that there has been no resolution to the *reduction ad absurdum* in which sophisticates are not in any way distinguished by the truth content of their beliefs, and according to which all constrained belief systems are treated as equally indicative of sophistication. As such, the political sophistication literature and its offshoots have failed to offer a satisfying formulation of sophistication that has any normative value at the level of the individual citizen. Critiques of sophistication scholarship have in fact been exclusively concerned with the question of collective outcomes, in the process often depicting the mind of the average citizen as a *tabula rasa*, which responds mechanically to the signals of thought leaders. Whereas Converse argued that the distribution of sophistication was highly skewed, the new variable, political 'competence,' is depicted as having a flat distribution. The latter is an inevitable consequence of presenting purely stylized accounts of political behavior, whether heuristic-based decision-making (Lupia, 2006), or other social psychological accounts, such as online processing (Lodge et al., 1995, 1989). As such, it may be questioned whether critics of sophistication scholarship have offered any new variable worthy of study in its place. It is not clear, for example, who could actually be politically 'incompetent,' under the competence framework that has been proposed.

In addition, just as Converse failed to show what purpose his concept of sophistication serves, his critics have likewise failed to show the value of 'competence,' despite the more

pragmatist rhetoric. For example, the question of competence is couched in terms of making decisions that best serve one's 'interests,' but without tackling the issue of how one knows what these are. Indeed this is where the pragmatist emphasis on 'procedural knowledge,' in which political participation is perceived as analogous to riding a bicycle, shows its limits. One may ride a bicycle to get from A to B for some important purpose, or for no other reason than the fun of it. In either case, one is equally 'competent.'

Toward a New Conception of Political Sophistication

Given the persistence of the debate concerning the classic conception of political sophistication, one appealing idea might be to dispense with it entirely. One could, for example, simply substitute the more neutral concept, political 'knowledge,' and, to avoid being dogmatic, adopt the broadest possible definition. A problem remains however in that this approach would still not allow one to distinguish between the different purposes for which such knowledge can be used and, as such, would still fail to offer a normative conception of value to the citizen.

Rebuking the philosophical idealists of his day, Bertrand Russell argued that the field of logic should no more admit a unicorn than the field of zoology (Russell, 1993: 169), and went on to make the point that the use of abstractions is not license to depart from the real world. *A fortiori*, it might be said that a focus on the abstract and empirical in the field of political behavior should not be license to admit absurdities, such as a formulation of political sophistication which, as conceded by Luskin, has nothing to do with acuity, or a formulation of political competence which suggests one is liberated by political ignorance.

As such, one possible way forward is to take the normative question more seriously by making one's goals and values themselves constitutive of political sophistication. Sophistication then ceases to be a purely instrumental phenomenon, potentially serving destructive purposes. In turn it follows that, for sophistication to have normative appeal, such goals and values must be of a character that offers the possibility of consensus. The essence of what is intended here then by the term 'political sophistication' is captured in the following definition:

$$\text{Political sophistication} = \text{political knowledge} \times \text{values}$$

Again the analogy drawn earlier between Converse's study of political sophistication and Kenneth Arrow's study of social choice is revealing. Having demonstrated his impossibility theorem, Arrow went on to consider ways out of his logical cul-de-sac in which the aggregation of individual 'rational' preferences could not be guaranteed to yield a rational collective outcome. His musings led him to a possible distinction between one's overtly expressed preferences and one's ultimate values (the latter of which may never be fully formulated), or what he alternatively referred to as the distinction between 'social decisions' and 'social ends.' As Arrow observed, "The relation between the two sets of values is not unique; for a given set of ultimate values, there may be differing sets of instrumental values, depending on the greater or lesser knowledge of the individual as to the best means of achieving his ultimate values and as to what his ultimate values are" (Arrow, 1951: 86–87).

Bringing values and goals back into the framework as analytical constituents of political sophistication implies the rejection of Converse's classical formulation based on ideological constraint as well as the rival conceptions based on 'competence.' Both approaches have treated one's goals and values as exogenous and fixed, in turn depicting the role of the citizen as only choosing between predetermined options. In particular, those advocating the political competence framework have in turn ridiculed the classic notion of the informed, engaged citizen and argued that citizens are more 'free' to the extent their participation in the polity is minimal. Under this view the use of heuristics or proxies as a decision-making mechanism contributes to this freedom by allowing one to do 'more with less.'

Yet such a perspective overlooks a distinction made by Sen (1992) between 'well-being freedom' and 'agency freedom.' The latter concept is especially pertinent to politics:

> If a person aims at, say, the independence of her country, or the prosperity of her community, or some such general goal, her agency achievement would involve evaluation of states of affairs in light of those objects, and not merely in light of the extent to which those achievements would contribute to her own well-being.
>
> Sen, 1992: 56

This distinction is problematic for the concept of political 'competence' which pertains only to well-being achievement, while treating one's goals as if they can be objectively determined. Yet if such interests cannot be so determined, the competence criterion does not have any reference point. The concept of 'agency freedom' underlines the point that political freedom is not simply a matter of choosing between predetermined options but also pertains to the choice of which options are to be available to choose from. As such, the whole concept of political 'competence,' rather than increasing one's freedom, assumes away some of its important aspects.

"When the facts change, I change my mind" J.M. Keynes is said to have quipped in response to a critic, asking in turn "What do you do, sir?" The political sophisticate, according to the classic Conversian formulation, would presumably respond along the lines of: "Nothing, I think only in terms of abstract ideas," while the political 'competent' might similarly respond "Nothing," albeit with a different qualification: "Until I receive a signal to do otherwise." The first response conveys an inadequate conception of knowledge while the second indicates an absence of goals and values. Neither response seems worthy of the term 'sophisticated.'

In short, there seems no justification to promote the ideological constraint principle as the basis for political sophistication. Nor are the problems associated with Converse's classic formulation solved by changing the topic to 'competence.' The development of ones goals and values, the maintenance of a belief system that is not closed off from the factual world, and the adaptation of one's norms in light of changing circumstances and insights all seem relevant to what it must mean to

be politically 'sophisticated.' Further, it is contended here, the question of goals and values is every bit as relevant as the question of how much information one holds, the coherence of one's beliefs, or one's level of procedural competence.

The formulation of political sophistication offered here will be objectionable to those who insist that any formulation of political sophistication in the field of political behavior must be strictly empirical. This argument, however, is suspect on several counts. First, Converse's classic formulation of political sophistication was very quickly adopted as a normative ideal in the political science literature (Luskin, 1987). Likewise, the various accounts of voter choice and public opinion formation offered in reaction to it have invariably been justified on normative grounds (Lupia, 2006; Stimson, 2004; Lodge et al., 1989, 1995). Yet perhaps the best defense against this objection is to simply note the irony that political behaviorists, in their efforts to be strictly abstract and empirical, have adopted theories of knowledge, such as pragmatism and coherence theory, which relegate the whole importance of data and facts. Ignoring the normative question has turned out to be self-defeating and no satisfactory formulation of political sophistication seems possible without it.

See also: Attitude Formation and Change; Attitude Measurement; Attitudes and Behavior; Attitudes, Political and Public Opinion; Ideology, Sociology of; Political Science and Polling; Public Opinion: Social Attitudes.

Bibliography

Arrow, K.J., 1951. Social Choice & Individual Values, second ed. Yale University Press, New Haven and London.

Berelson, B.R., Lazarsfeld, P.F., McPhee, W.N., 1954. Voting. University of Chicago Press, Chicago.

Boninger, D.S., Krosnick, J.A., Berent, M.K., 1995. Origins of attitude importance: self-interest, social identification, and value relevance. Journal of Personality and Social Psychology 68 (1), 61–80.

Bryce, J., 1904. The American Commonwealth, third ed. Macmillan, New York.

Campbell, A., Converse, P.E., Miller, W.E., Stokes, D.E., 1960. The American Voter. John Wiley & Sons, New York.

Converse, P.E., 1964. The nature of belief systems in mass publics. In: Apter, D.E. (Ed.), Ideology and Discontent. The Free Press, New York, pp. 206–261 (Chapter 6).

Dahl, R., 1986. A Preface to Economic Democracy. University of California Press, San Francisco, CA.

Dahl, R., 2003. How Democratic Is the American Constitution? Yale University Press, New Haven, CT.

Druckman, J., Kuklinski, J., Sigelman, L., 2009. The unmet potential of interdisciplinary research: political psychological approaches to voting and public opinion. Political Behavior 31 (4), 485–510.

Festinger, L., 1957. A Theory of Cognitive Dissonance. Stanford University Press, Stanford, CA.

Keynes, J.M., 1997. The General Theory of Employment, Interest, and Money. Prometheus Books, New York.

Krosnick, J., Visser, P., Harder, J., 2010. The psychological underpinnings of political behavior. In: Handbook of Social Psychology, Chapter 34. Wiley Online Library.

Krosnick, J.A., 2002. Is political psychology sufficiently psychological? Distinguishing political psychology from psychological political science. In: Kuklinski, J.H. (Ed.), Thinking about Political Psychology. Cambridge University Press, pp. 187–216.

Kunda, Z., 1990. The case for motivated reasoning. Psychological Bulletin 108 (3), 480–498.

Lippmann, W., 1922. Public Opinion. Macmillan, New York.

Lodge, M., McGraw, K.M., Stroh, P., 1989. An impression-driven model of candidate evaluation. American Political Science Review 83 (2), 399–419.

Lodge, M., Steenbergen, M.R., Brau, S., 1995. The responsive voter: campaign information and the dynamics of candidate evaluation. American Political Science Review 89 (2), 309–326.

Lupia, A., 2006. How elitism undermines the study of voter competence. Critical Review 18, 217–232.

Luskin, R.C., 1987. Measuring political sophistication. American Journal of Political Science 31 (4), 856–899.

Luskin, R.C., 2002. From denial to extenuation (and finally beyond): political sophistication and citizen performance. Thinking About Political Psychology 281–301.

Matlin, M.W., 2005. Cognition, sixth ed. Wiley, Hoboken, NJ.

Poundstone, W., 2008. Gaming the Vote: Why Elections Aren't Fair (and what We Can Do about It). Hill and Wang, New York.

Prior, M., 2009. Improving media effects research through better measurement of news exposure. Journal of Politics 71, 893–908.

Prior, M., Lupia, A., 2008. Money, time, and political knowledge: distinguishing quick recall and political learning. American Journal of Political Science 52 (1), 169–183.

Russell, B., 1948. Human Knowledge. Routledge, London and New York.

Russell, B., 1993. Introduction to Mathematical Philosophy. Routledge, London.

Sen, A., 1992. Inequality Reexamined. Russell Sage, New York.

Sen, A., 2000. Development as Freedom. Anchor Books, New York.

Stimson, J.A., 2004. Tides of Consent: How Public Opinion Shapes American Politics. Cambridge University Press.

Ideologies, Institutions, and the New Institutionalism

Dirk Sauerland, Witten/Herdecke University, Witten, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Within the last decades new institutionalisms have been developed in political science, sociology, and economics. The variety of analyses is large: In political science institutionalism a rational approach, a historical and a sociological can be distinguished. Sociology itself has developed a new institutionalism which is similar to the political sociological approach. Finally, New Institutional Economics has become a growing part of economic theory and empirical research. All these approaches have a strong focus on institutions, their logic, their origin, and change in common. However, a common definition of institutions is missing in all of approaches. The discussion in the new institutionalisms becomes even more complex, when the phenomenon 'ideology' is introduced into their analyses. The challenge then is twofold: Ideologies and institutions must be clearly distinguished if both should play a role in theory, and both have to be put to empirical testing, to prove their power in explaining real-world phenomena. Although the new institutionalism in social science are no longer new, there still has to be done a lot of work, especially to understand the logic, origin, and change of ideologies and their impact on the performance of political and economic systems and organizations.

New Institutionalism in Social Sciences

Although the insight that '*institutions matter*' is not new, it was neglected for a long time in social sciences. But during the last decades, it has become more popular again. Recently there are new approaches in sociology, political sciences, and economics with a clear focus on the influence that institutions have on human behavior, on their origin, their change, and their impact on the performance of different organizations, economies, and political systems.

The new institutionalisms have a common interest in institutions but their theoretical and empirical approaches are quite different because of their roots in different sciences. These roots also influence the research objectives: Whereas New Political Institutionalism (NPI) analyses the function and origin of institutions in the political sector, the focus of New Sociological Institutionalism (NSI) is more on the organizational level. New Institutional Economics (NIE) has developed different research areas which refer to organizations within the economy, to the economic system, and to the political system being both parts of the larger societal system. In all of these settings, the distinction between the activities framed by a given set of institutions (choice within rules) and the activities to design such institutions (choice among rules) is an important aspect of institutional analysis.

Over the past decades New Institutional Economics has become a growing and recognized part of economics. Leading scholars of NIE have been awarded with the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel (colloquially the Nobel Prize in Economics), for example Elinor Ostrom and Oliver Williamson (in 2009), George A. Akerlof (in 2001), Douglass C. North (in 1993), Ronald H. Coase (in 1991), and James M. Buchanan (in 1986).

One characteristic of the new institutionalism in economics is its application of the general economic approach to human behavior. In contrast to traditional neoclassical microeconomics, the institutional microeconomic approach does not assume perfect information of all relevant actors. Instead imperfect or asymmetric information is a key assumption in the analysis of institutions.

A similar rational choice approach exists in New Political Institutionalism. However, there are two other branches that have been developed in NPI. Taking a different perspective than the rational NPI, the sociological NPI approach sees institutions not as a consciously chosen means to an end. Institutions are perceived as the result of a cultural development. Accordingly, this branch of political institutionalism has its focus more on the function of institutions than on their origin. The third political institutionalism is labeled as historical and combines parts of the rational and the sociological approach. Institutional development is always seen before a historical background, which is on the one hand culturally embedded but on the other hand also a rational decision of political actors within the given historical and sociological context.

Finally, there is also a NSI. A key element in NSI analyses is the diagnosis that not only institutions do influence individual's behavior. Social ties and relations in which individuals are embedded also have an impact on their behavior. In addition, the sociological institutionalists assume that the existing institutions influence the following changes of institutions by themselves. With this idea the sociological institutionalists argue similarly to the economic institutionalists, but put even more emphasis on the path dependency of institutional development and the persistence of institutions.

Institutions and Ideologies: The Logic of Institutionalisms Compared

For most of the interesting issues in political and economic markets, uncertainty, not risk, characterizes choice-making. Under conditions of uncertainty, individuals' interpretation of their environment will reflect their learning. Individuals with common cultural backgrounds and experiences will share reasonably convergent mental models, ideologies, and institutions; and individuals with different learning experiences (both cultural and environmental)

> will have different theories (models, ideologies) to interpret their environment.
>
> Denzau and North, 1994: p. 3f

This quotation refers to the question, why human beings adapt rules, develop routines, and act according to them in everyday decision. The answer is: to reduce uncertainty and the complexity of the individual decision making. Irrespective of being called institutions or ideologies, these phenomena provide helpful orientation for individuals. It is a major assumption that all new institutionalisms have in common: ideologies and institutions are influencing individuals' behavior.

The concepts of ideologies and institutions have at least one more thing in common: There is no standard definition of both. Not only the different disciplines in social science have different perceptions of both terms, also within political sciences, sociology, and economics different definitions of both terms exist. As a result of this variety of definitions, it is even difficult to distinguish between institutions and ideologies. The next section tries to clarify this definitional complexity.

Institutions in Social Science

Although there is no common definition of institutions in the social sciences literature, they are in general and broadly defined as 'the rules of the game.' Games – in this definition – are played by more than one person. They are collective undertakings in the political arena, in society, in organizations, or in the economy. Therefore, institutions are broadly defined as a set of rules that frame the interaction between members of different size collectives. However, the areas of application vary in the different new institutionalisms. And there are some specific aspects in the science-specific definitions.

Institutions in Political Science

With regard to the definition and analyses of institutions, two basic approaches exist in NPI. The first can be labeled as 'calculus approach,' the second as 'cultural approach' (cf Hall and Taylor, 1996: p. 7f).

The calculus approach in NPI assumes that individuals try to maximize the attainment of a defined set of goals. Being similar to the economic approach to human behavior, the NPI calculus approach has a clear microfoundation. Institutions in this respect may be defined as framing rules for the interaction of different individuals. They affect individuals' behavior and they are relevant because of the strategic interactions of those individuals in a given context. Institutions that are shared by a group of people reduce uncertainty about the potential activities of the group members. They are productive because they enable better reactions on others' actions.

In contrast to that, the cultural approach in NPI sees individuals more as satisficers than as maximizers. Individuals' behavior in this approach is not merely strategic but also influenced by their 'Weltanschauung.' Under this framework, individuals develop a set of behavior patterns to reach their goals set. Institutions in this respect serve as moral or cognitive templates for the interpretation of the world around. Defined in this way, institutions not only provide information for strategic behavior, but they affect individuals' identities and their preferences.

Apart from this broad definition based on functions, the three different institutionalisms in NPI have their own special focus on institutions and own specific definitions.

To start with, in *historical institutionalism* institutions are defined as "formal and informal procedures, routines, norms and conventions embedded in the organizational structure of the polity or political economy" (Hall and Taylor, 1996: p. 6). Historical institutionalists do obviously not only apply the calculus but also the cultural approach (especially when referring to the social embeddedness). In their concept of institutional analysis, historical institutionalists emphasize the asymmetrical distribution of powers resulting from the existing institutions. They also assume path dependencies in institutional development, therewith analyzing persisting institutional settings. Finally, historical NPI is open to incorporate other factors that influence political outcome in its analysis. Examples for such other factors are beliefs and ideologies.

In contrast to this flexible theoretical basis, *rational choice institutionalism* in political science adopts the calculus approach very strictly (cf Hall and Taylor, 1996: p. 10ff). Institutions are again seen as the rules of the political game. In this game they help to solve some problems of collective actions, such as unstable majorities. Institutions do that by lowering transaction costs and enabling gains from cooperation or exchange, respectively. Here the close relationship of rational choice NPI to New Institutional Economics becomes obvious.

Rational choice institutionalists assume – in line with the economic approach to human behavior – a given set of individuals' preferences. Individuals are assumed to maximize the attainment of their preferences. Scholars of the rational choice approach have also a focus on collective action dilemmas. Those can often be solved by adequately designed institutions. Those institutions frame strategic interaction of the individuals concerned.

Also the *sociological institutional* approach in political science has some distinctive features. Sociological institutionalists "tend to define institutions much more broadly than political scientists do to include, not just formal rules, procedures or norms, but the symbol systems, cognitive scripts, and moral templates that provide the 'framing of meaning' guiding human action" (Hall and Taylor, 1996: p. 14). Doing this, the concepts of institutions and culture become more similar than in the other NPI approaches. Some scholars even redefine culture itself as an (informal) institution.

In contrast to the rational choice institutionalists, which take the preferences of the individuals as given, the sociological institutionalists assume that institutions affect the underlying preferences or identities of the individuals by themselves. Consequently, individual choices are not only influenced by institutions but also by cultural specificities which result from the context-related embeddedness of the actors.

Institutions in Sociology

The New Sociological Institutionalism shows, not only by name, similarities to the sociological institutionalism in political science.

NSI has a clear focus on the institutions framing organizations and the analysis of their members' behavior. Institutions obviously matter, but what are the relevant institutions in this setting? Basically again institutions are defined as the rules of the game people play within organizations. Having a closer look at the formal organizational rules – for example, within a firm – the question comes up, why some people follow the rules and some not.

Avner Greif proposed a first answer to questions like these referring to the informal, i.e., not written rules. He argued that informal rules, beliefs, and norms combined with the set of formal rules motivate and guide behavior of individuals in organizations. Taking these informal rules into account, Greif (2006) analyzed in game theoretic models various principal agent relations within organizations. His approach fits well with the definition of institutions that Powell and DiMaggio (1991) suggested: institutions are rules, procedures, organizational standards, and governance structures plus conventions and customs. All of these factors define the actions of a rational actor.

Conventions and customs in this definition refer to another important factor influencing individuals' behavior. Mark Granovetter (1985) highlighted, that the impact of formal institutions on the behavior is often overestimated. Formal institutions, as mentioned before, reduce uncertainty by creating trust. Granovetter added that not only formal institutions create trust in an organization. Social ties and networks that exist within organizations have the same effect. These networks and ties are part of the organization but they are not formally fixed. However, the idea that members of an organization are embedded in such informal social structures was important for the development of NSI.

In addition, no single individual can control the effects of a networks; the social context is therewith beyond the individual's control. Having this in mind, the idea of maximizing individuals' goal attainment becomes obscure. In line with the arguments developed by Herbert Simon (1957) individuals being confronted with formal organizational rules and social embeddedness are facing highly complex decision situation. They cannot maximize their goal attainment but will probably 'just' satisfice their behavior. Again, adapting routines or stable patterns of behavior, respectively, makes their life easier.

Institutions in Economics

According to Douglas C. North, one of the most prominent scholars in New Institutional Economics, institutions are "humanly devised constraints that structure political, economic and social interactions" (North, 1991a: p. 97) or, in short again, 'the rules of the game.' The explicit reference to political and social interactions indicates that NIE analyses are not restricted to the core economic sphere. As NIE analyses the institutions of the economic system, within organizations and in the political sector, its range of application is very broad and gives room for economic imperialism, i.e., competitive analyses of NIE in the typical application fields of NPI and NSI.

From the NIE point of view, institutions that are accepted by the participating players enable productive cooperation. They reduce transaction costs for rational decision makers in various contexts and therewith provide the framework to generate mutual gains from cooperation. This perception is obviously very similar to that of the rational approach in NPI.

Having a closer look at the meaning of institutions and their field of application the New Institutional Economics distinguishes different kind of institutions. First, NIE analyses distinguish between formal and informal institutions. Second, there is a difference between internal and external institutions. Third, we find the distinction between institutions and ideologies, both defined to be shared mental models of different individuals (Denzau and North, 1994).

As mentioned in the earlier quotation by North, NIE analyses a wide range of institutions under different foci. One focus is the general role that institutions play within the economic approach to human behavior. In addition to this positive part of institutional theory, the normative question in institutional design is a major part of ongoing research. The focus in this normative part of NIE lies on the choice among rules. The research question is: how should institutions be designed to generate a certain behavior of the actors subject to that rules?

"The rational choice framework assumes that individuals know what is in their self-interest and make choices accordingly. […] Instead people act in part upon the basis of myths, dogmas, ideologies and 'half-baked' theories. Ideas matter; and the way that ideas are communicated among people is crucial to theories that will enable us to deal with strong uncertainty problems at the individual level." This quotation from the seminal paper *Ideologies and Institutions* by Arthur T. Denzau and Douglass C. North (1994) provides a good insight into the concept of institutions and ideologies. Both phenomena are, in the positive theory, tools to cope with uncertainty, which is – more or less – the omnipresent context for decision making in reality. Under uncertainty rational decision making is more than difficult. And individuals – as Denzau and North point out – are aware of the difficulties they are facing.

To begin with the individual's level, internal institutions are those rules that individuals adapt to restrict voluntarily their own set of decision options. This kind of self-restriction is useful to reduce uncertainty within repeated and similar decisions situations. If decisions taken proved to be successful, rational decision makers will remember in similar situations these beneficial strategies and will react (if not identically) in a similar way. In contrast to that, decisions that resulted in a low utility level will (probably) not be repeated. To reduce their decision problems under uncertainty individuals learn and adapt. Thus, individual learning results in the set of behavioral patterns that will be applied regularly. A prominent example in world literature for an internal self-restricting rule is Ulysses' reaction to the challenge of the sirens (cf Elster, 2000).

Whereas internal institutions are applied on the individual action level external institutions frame the interaction of groups of people. Formal external institutions are typically fixed and transferred in a written format. Prominent examples are constitutions, other legal rules, and private contracts. In contrast to these written format rules, informal institutions do also shape the behavior of the collectives' members, but are not fixed in a written way. Culture is – in NIE – one prominent example for informal rules: it clearly has an impact on the individual's behavior in a larger group of people (as societies in general) and also in smaller groups (e.g., a company). However, it is just orally transmitted for example by education.

As a result, individuals are embedded (in the logic of Granovetter) in informal cultural ties. The logic of formal and informal institutions is that they influence individual's behavior and therewith limit the available set of actions to be chosen from.

Informal institutions are sometimes transferred into formal institutions, especially when group size increases. Examples for this development are the modern, formally fixed corporate governance codices that substitute the traditional informal rules of decent behavior in business activities.

External institutions, either formal or informal, are not per se stable. Mechanisms to sanction an inadequate behavior therefore play an important role. Here not only the value of the penalty is relevant but also the probability to be detected as a rule breaker.

Institutions in this sense are shared by different people. Denzau and North (1994) therefore define institutions as some 'shared mental models.' A mental model is an internal institution that helps the individual to understand the world. It provides a structure to interpret the social and economic interactions. If some people develop similar interpretation structures, they share these mental models. This again helps to develop informal and formal institutions that then structure the social and economic activities, according to those shared mental models. In fact, institutions that shape the behavior in a group of people are an expression of their shared models.

Ideologies in Social Sciences

One of the best known definitions of ideologies in social sciences is that of Karl Marx. He referred to ideologies as a quality of thoughts, especially in the social context, that is illusionary or distorted. Apart from the idea of ideologies as mental misconception of reality and the relationships of individuals to their social surroundings there are more topical conceptions of the term ideology within the discussion about new institutionalisms. And most have one perception in common: ideologies are shared by a group of people, often having similar goals in politics, or in other of different size collectives. In terms of Denzau and North (1994) an ideology is another category of shared mental models.

Ideologies in Political Science

If institutions and ideologies have some similarities and differences, theoretical and empirical analyses become complicated. With this regard Mullins (1972) made a helpful contribution to the discussion 'on the concept of ideology in political sciences' with two statements: First, there must be an accepted definition of the term to distinguish it from other similar (social) phenomena. Second, it is necessary to conceptualize ideologies in order to make it an empirically relevant term. Whereas the first aspect refers directly to the discussion about ideologies and institutions, the latter is a more methodological aspect about empirical relevance as a measure of the quality of theories developed. To define ideology – in distinction to other similar phenomena – Mullins presented four arguments.

The first argument refers to ideology as an indicator for historical consciousness. Ideologies provide the cognitive power to identify recurrent aspects of experiences from the past in applying both, abstraction and generalization. It therewith provides, on a positive theory level, an interpretation scheme of the reality based on learning from history.

The second argument is that ideologies provide structured conceptions of the driving forces within the social world. Doing this, ideologies serve simultaneously as benchmarks for the evaluation of real-world policy making. This has a clear normative aspect.

The third argument develops the evaluation idea further. If, for example, the real-world policies are not according to what the prevailing ideology suggests, evaluations typically result in activities of political change. Ideologies are therefore a major source for political reform acts. These reforms can also address the institutional design. If the existing design is not compatible with the majorities' ideology, the institutions will be changed and become adapted to prevailing ideology. Ideologies define – in their normative tenor – objectives to be aimed. Ideologies also foster communication among group members about evaluation, ideals, and purposes. In doing this, ideologies help to mobilize group members, especially if a certain objective seems to be missed by status quo politics.

The fourth argument refers to logical coherence. This coherence, according to Mullins, distinguishes ideology from mere ideas, attitudes, or feelings about politics. However, the logic of an ideology can be broadly defined. As emphasized also by representatives of NSI and NIE, an ideology must make sense. Logical inconsistencies make ideologies vulnerable to upcoming new competing ideologies. And there are most likely more inconsistencies to be found the more mature and sophisticated an ideology becomes.

Going further in time, the term ideology is nowadays used more sophisticatedly, referring to different groups within society. Analyzing *Ideology and Institutions in North America*, Finbow (1993) distinguishes, for example, two basic kinds of ideologies. Organic (or inherent) ideologies arise from social relations and economic conditions, whereas traditional (or derived) ideologies are more formal propositions drawn from traditions and history. In this strand of literature, institutions and ideologies are interdependent variables: ideologies do not only influence the character of policy and institutions, they interact with institutions and social conditions.

In Finbow's definition, formal, society-related ideologies are beliefs and/or values that are often derived from major events in a nation's history. They provide context for a later evolution of ideologies and institutions and they shape the initial policies, parties, and institutions. Such major events can be, e.g., (civil) wars; such events are deeply rooted in the shared memory of a society.

In contrast to these national ideologies, organic ideologies are group related and developed by social groups in response to their specific circumstances and interests. Looking at the organic ideologies, Finbow differentiates further between popular, central, and public ideologies. These relate to different smaller groups within society, each – more or less – trying to reach its own goals. This argument is similar to that developed by Mancur Olson (1965) in his *Logic of Collective Action*. Following this logic, popular ideologies are based on the experiences of popular groups within a society. These groups, like political parties, have influence on public policy-making by electoral power or other channels. The idea

of a central ideology refers to the economic system as a whole. Central ideologies are created and driven by business and professional groups; they are bound to self-interest or social awareness. This kind of ideology can be the underlying idea of an economic order like the Social Market Economy in Germany.

Finally, the public ideology refers to state actors and linkage agents. They are based on traditional beliefs, for example, about the role of the administration in government, group pressures, such as lobbying activities, or other agency interests.

Ideologies in Sociology

As with NPI and NIE ideologies are far less in the focus of analyses of sociological institutionalism than are institutions. In fact, they have been widely neglected for a long time. But in 2009 some sociologist published a volume of works on *Institutions and Ideology*. One of its purposes was "to encourage re-engagement with the notion of ideology in contemporary institutional research" (Meyer et al., 2009: p. 9). The editors justify their attempt to do this based on the insight that "the silence with regard to ideology is surprising since many of the core concepts of institutional theory, such as institution, legitimation, social categories, institutional logics, and theorization, are, in one way or another, closely related to the notion of ideology" (Meyer et al., 2009: p. 2).

This statement implies – again – that ideologies and institutions are in a close relationship. Institutions shape the individual's behavior and social context. They form social entities, which are the basis for any social order. Ideologies are important for institutional design because institutions incorporate the underlying ideology that is shared by the members of a social entity. Ideologies then have a twofold impact: they legitimate a certain social order and they constitute that order in the first place.

Going back in history, Karl Marx' conception of ideology also plays an important role in sociology (cf Marx and Engels, 1990). A second prominent scholar in sociological research on ideology was Karl Mannheim (1972). He saw ideology more as a general phenomenon, claiming that all human perception of reality is influenced by the socio-historical circumstances. Human thoughts are therewith context related and always influenced by ideologies (as the before mentioned 'Weltanschauung'). Ideology is then a system of beliefs shared by a group of individuals. Following this tradition, Simons and Ingram (1997: p. 784) define ideologies as "a set of beliefs about how the world operates, including ideas about what outcomes are desirable and how they can best be achieved." In this definition, ideologies have again a positive and a normative aspect.

Of course, NSI also has to solve the problem of distinguishing ideologies from institutions. As there is no common definition of both terms, there are different suggestions for distinction.

Some scholars argue that ideology is the link between the framing institutions and individuals' behavior. Following this logic, differences in the underlying ideology that individuals share can help to explain differences in individual behavior under a given set of institutions. Accordingly, Vogel (2009: p. 93) in the volume mentioned defines ideology as "a system of beliefs shared by a particular social group for whose members it has immediate behavioral implications." Vogel also defines four essential features of such an ideology.

First, ideologies are a cognitive framework that helps to understand (positive aspect) and to 'design' the world (normative aspect). In addition – like in the political sciences – ideologies have a certain systematic structure to interconnect the different parts of belief elements. This is similar to the idea of ideologies being logical structures, which we discussed in the previous paragraph. Ideologies, according to Vogel, also create group consciousness by formulating consensual positions of the different group members. Finally, ideologies themselves influence individuals' behavior providing legitimization for certain behavior or nonbehavior.

Taking this focus, ideologies are complementary to the formal set of institutions. They form the nonobservable part of the institutional system and, at the same time, are the starting point of institutions origin and change.

Looking at the logic of ideologies in New Institutional Economics, similar ideas can be found.

Ideologies in Economics

As institutions, the ideology people share has an influence on their behavior. Compared to informal and formal institutions, ideologies are not the rules of the game but – in a positive theory – the 'ideas of the game.' And ideologies can also be ideas about how the game should be structured (normative approach). Denzau and North (1994: p. 20) define both, ideologies and institutions, as different classes of shared mental models: "Mental models are shared by communication, and communication allows the creation of ideology and institutions in a co-evolutionary process."

An ideology provides a positive mental model that is – according to Denzau and North – focused on action and valued outcome. In this respect, an ideology enables people in a society (or other collectives) to choose compatible behavior patterns. It defines a set of possible actions with which an individual can reach his or her utility driven aim best. From a normative point of view, ideologies provide a pre-selection of possible actions. They "identify the aspects of reality that are crucial for achieving one's goals" (Denzau and North, 1994: p. 15).

Ideologies and Institutions: Differences and Similarities in Concepts

Until today there is no generally accepted definition of the terms institution and ideology in social sciences. There is also no accepted approach to distinguish both phenomena. And there are a lot of other phenomena and terms, that are also present within the discussion about ideologies and institutions. **Figure 1** summarizes the phenomena discussed earlier and highlights some of their relevant interrelations.

Apart from the differences in NPI, NSI, and NIE, there seem to be some core elements (almost) all institutionalism have in common.

1. All institutionalisms see institutions and ideologies generally as means oriented.
2. All share the idea that institutions and ideologies help to structure human interaction, therewith reducing complexity.

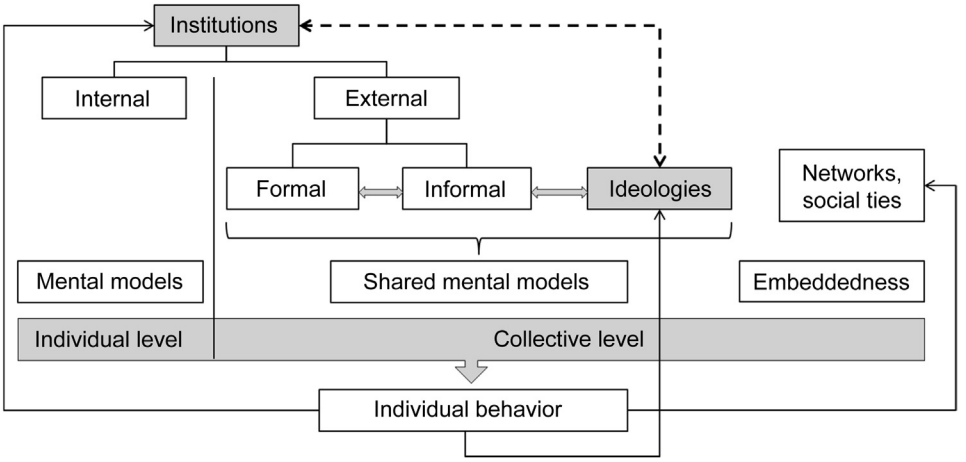

Figure 1 Institutions, ideologies, and embeddedness – an overview.

3. Ideologies and institutions – shared by a group of people – create trust among the group members; but they are not the only sources for creating trust.
4. Both help to make the reactions of the other group members better foreseeable, therewith reducing uncertainty.
5. Especially the institutions assign responsibilities – in the relevant political, organization, or economic context – and therewith allocate power.
6. Prevailing ideologies are an underlying logical structure for the set of institutions shared by a group of people; institutions and ideologies are interdependent.

Vogel (2009: p. 94) underlines the last aspect: "Ideologies are the non-observable and ideal part of an institutional order that is, and must be, related to material practices."

Looking at the three new institutionalisms in social science, the major differences between them result from the methodology they apply. Some approaches clearly apply a rational choice approach: this is very prominent in NIE, but also in parts of the NPI and NSI. On the other hand, the cultural approach is fundamental for most NSI approaches, but makes also its way into some NPI and NIE approaches.

This basic difference in the perception of institutions (and ideologies) has an important impact on the way institutional and ideological origin and change is analyzed. Facing competing explanations of institutional impact on individual behavior, empirical testing of the concepts may be helpful to distinguish the relevance of explanations.

Unfortunately, one of the major challenges for all three institutionalisms still is to conceptualize ideologies and to have, for example, a precise distinction between ideologies and informal institutions. If there is no clear definition of both terms and if ideologies are, as the quotation earlier indicates, the nonobservable part of institutional systems the empirical testing of both concepts faces serious challenges. However, there are empirical analyses on the impact of ideologies. We mention some in Section Ideologies and Institutions: Empirical Relevance.

Ideologies and Institutions: Origin and Change

Apart from the mere logic that institutions and ideologies help to reduce uncertainty in social, political, and economic interactions and relationships two other major questions in social science are about the origin and change of both, institutions and ideologies.

Generally speaking, there are two different perceptions of the origin of ideologies and institutions in new institutionalism literature. The first is to see both social phenomena evolving in an evolutionary process, sometimes as the unintended collective consequences of activities based on individuals' intentions. This perspective is typically assumed by the proponents of the more cultural approach on ideologies and institutions. Accordingly, the researchers applying a rational choice approach typically have a different focus. They assume especially institutions to originate as a result of rational decisions, either of the people affected, e.g., in democratic processes, or of owners and managers, e.g., in hierarchical structures.

In addition, some institutionalisms have a focus on the origin and change of institutions, others take them as somehow given and just analyze the effects of the given set of institutions and its change.

Ideologies and Institutions: Origin

To give a short survey about the different foci and perceptions, we start with the NPI approaches. *Rational choice NPI* has a focus on the origin of institutions. In this strand of literature, institutions originate on purpose because of the function they have to fulfill in the political arena. Consequently, institutional change is also modeled as a conscious choice among rules (Hall and Taylor, 1996: p. 19f). Change will be initiated by the relevant actors if the costs of change are lower than its benefits, which again refer to the purpose of the institutions. *Historical and sociological NPI* both have no focus on the origin of institutions but assume that an initial set of institutions is given before new institutions are adapted or created (Hall and Taylor, 1996: p. 20ff). Taking the perspective of *sociological NPI*

these initial institutions shape the origin of new institutions, because they influence the mental models of the institutional entrepreneurs. The origin of institutions in this perspective is not just means-end related but is the results of interactive processes between the collective members. By communicating with each other the individuals develop similar ideas about how the social world works and how institutions could be designed to structure the social world adequately. *Historical NPI* often focuses on the distribution of powers resulting from the initial set of institutions. This distribution determines the ability to create new institutions in the course of time.

The *NSI perspective* on the origin of institutions is, not surprisingly, similar to that of sociological NPI and also applies a more cultural than a rational choice approach.

Finally, in Denzau and North's *NIE concept*, ideologies help individuals to build a mental model and develop a set of internal institutions. Because ideologies are a 'blueprint' for building these models, individuals that use a preformed ideology are able to develop a mental model faster than those who rely on their own experiences, built on individual-learning processes. However, these individual-learning processes in both cases have an impact on internal institutions, mental models, ideologies, and external institutions.

The Evolutionary Approach: On Ideologies and Institutions

To see how communication and interaction between members of a collective can lead to the evolutionary origin of ideologies and institutions, we will briefly sketch the logic of an NIE approach in this paragraph. To explain that logic, we refer to a statement by North (1993: p. 20): "Formal rules are an important part of the institutional framework but only a part. To work effectively they must be complemented by informal constraints (conventions, norms of behavior) that supplement them and reduce enforcement costs. If the formal rules and informal constraints are inconsistent with each other the resulting tension is going to induce political instability. But we know very little about how informal norms evolve."

Decision making under uncertainty is omnipresent, for political as much as for economic decision makers. Only very rarely the outcome of a decision in a social context is certain. Learning means, as mentioned earlier, dealing with complexity and uncertainty. If individuals realize that a decision was for their benefit, this pattern will become part of their typical behavior. Decisions with a nonsatisficing impact on their target will not be repeated. The learning process results in a set of successful behavioral patterns that serve as internal rules for decision. These *internal rules* or routines, respectively, reduce uncertainty and – in terms of transaction cost economics – and therewith reduce the costs of decision making. As an individual's mental model, they provide orientation in a complex environment. From a rational choice perspective, individuals will stick to their internal rules or institutions if the benefits of a restricted set of alternatives are assumed to be higher than the costs of making wrong (or utility decreasing) decisions (Heiner, 1983).

Having a look at **Figure 1**, the individual mental models become shared mental models via communication. In other words, communication is the basis for the evolutionary origin of *ideologies*. The process of ideology origin may be accelerated by ideological entrepreneurs, taking a leadership role in ideology development. As Denzau and North (1994) pointed out, ideologies and institutions evolve in a co-evolutionary process.

If the decision making of an individual shows regularities the individual's behavior becomes predictable for other individuals who interact with him or her. If, for example, a businessman (-woman) is known to be fair-minded, this is beneficial for all his or her cooperation partners. Build upon reciprocity, to stick to the rule of 'be fair-minded' creates trust in each other's behavior and therewith produces benefits for the group of interacting business people. Adopting and sharing this *informal external rule* in a group enhances the predictability of individual behavior and lowers transaction costs. More gains from cooperation can be realized. Adequately designed institutions, evolving out of an evolutionary learning process (Mantzavinos et al., 2004), are therewith productive.

However, informal external rules are, as mentioned earlier, not per se stable. Taking the game theoretical construction of the prisoners' dilemma as an example, the noncooperative position of the rule breaker can become a beneficial option for one of the group members. To prevent this, sanction mechanisms have to be implemented. In smaller groups these are typically social sanctions like exclusion from the relevant group.

As groups get larger, things become more complicated: rule breakers can no longer be easily detected and the costs for the others to find the rule breakers are high compared to the costs and individual rule breaker causes. Again, the members of the collective are facing the prisoners' dilemma situation. As a result, individuals could choose to set up *formal rules* and to have them stabilized by explicit sanction mechanisms (cf North and Weingast, 1989). At this time, the origin of institutions is no longer merely evolutionary. It becomes an act of rational decision making to ensure mutual gains from cooperation.

The Rational Choice Approach: On Institutions

The second way to think about the origin of institutions is therefore grounded in the rational choice tradition, which we find in political science and especially in NIE.

If institutions are means to solve collective (decision) problems, rational decision makers will search for adequately designed institutions to solve the relevant problems. In the NIE literature, there are examples of this theoretical approach for three different institutional settings: The formal rules framing the democratic system of a state are in the focus of Constitutional Economics. On the level of the rules for the economy as a whole, the German Ordnungstheorie provides an example for some conscious design of institutions. Finally, Transaction Cost Economics provides arguments for a rational design of rules, or institutions, respectively, on the organizational level.

Constitutional Economics, as a prominent part of NIE, models the foundation of a protective and productive state as a conscious and rational decision of all people affected (Buchanan, 1975). The logic of the argument starts again with a dilemma structure. It is the problem to find stable forms of living together peacefully in a larger group of people – the society. From the economics perspective, such a peaceful society is productive: trust is built and gains from cooperation can be realized. The rules framing this cooperative game therewith provide benefits for all people

accepting them. To overcome the free rider problems caused by potential rule breakers, there needs to be a new player in the game that monitors the rule acceptance and puts up sanctions for a breaking of the rules. This is, in short, the logic of a protective state.

The protective state ensures the property rights in a society and is founded in a social contract. This contract contains the rules of the game that all people affected can accept. Similar to Rawls' (1971) *Theory of Justice*, consensus on the rules is reached under a veil of ignorance. This construct helps to find acceptable rules for every member of the community because it prevents the citizens to pursue their individual interest in choosing the basic rules for the society. So, the theoretical model of Constitutional Economics is a rational choice among rules for the constitution of a state, but it is a choice under uncertainty. Built on this protective state, the next step – in theory – is the productive state. This is also framed, or constituted, by a set of formal rules that ensure, e.g., the provision of public goods and services by the state. In Constitutional Economics theory, the underlying ideology is not a central theme; the theory is just about the setting up of institutions.

But there are not just theories of conscious decisions about an adequate constitutional design. A prominent example for a real live experiment for the choice among rules-situation is the German Ordnungstheorie and their transformation into practical politics. This institutional approach, was established in the 1930s in south-west Germany. A team of economists and law scholars also known as the Freiburg School in a joint work developed a framework for a post-war economic constitution (Eucken, 1952) that later became the basis for the so-called 'Social Market Economy' in Germany. The ultimate goal of the theoretical and practical undertaking was to develop a framework for a workable and humane economic order ('funktionsfähige und menschenwürdige' Wirtschaftsordnung, Eucken, 1952: p. 14).

In a historical perspective, the institutional design developed by the scholars of the Freiburg School was affected by the shared ideology of those economists and jurists and the corresponding ideology of the political leaders in the post-war Germany. This ideology was a result of the experiences shared by exponents of the Ordnungstheorie. Based on their experiences with a totalitarian regime and the economic events before the war a 'social-liberal' ideology evolved. This started – in Finbows terminology (cf Section Ideologies in Political Science) – as a popular and public ideology of the people in academia and politics and then became a formal, society-related ideology for post-war Germany.

Two central ideas formed that social-liberal ideology: first, to ensure individual freedom (protective state) and second, to provide protection mechanisms against economic risk associated with unemployment, illness, and age (productive state). Built on this ideology, the scholars in Freiburg developed a set of rules that should be applied by the state to create a humane and workable economic order for the benefit of all people in Germany. These rules were designed to ensure a market economy shaped by workable competition processes in the first place ('konstituierende Prinzipien'). Those rules should be complemented by social security systems to correct for unintended effects of those competitive processes ('regulierende Prinzipien'). The security systems were meant to support the workability of the market economy. Correspondingly, the set of principles framing this order were labeled as constitutive and regulating.

The idea behind this institutional design is simple: The government should set the rules of the game called 'market economy' (choice among rules), whereas the private businesses and households should choose their best available moves within the given set of rules. The logic is similar to that of Constitutional Economics: if the government sets those rules and supervises their fulfillment, the outcome of the social market economy will be in favor of all citizens.

In the German Ordnungstheorie also the idea of a wide acceptance of the rules is a prominent one. In contrast to Constitutional Economics' theory, which assumes an explicit consensus about the founding rules, German Ordnungstheorie assumes a more implicit but broad approval as an indicator for the quality of the rules proposed.

The idea of a consensus or wide acceptance about the institutions is typical for a collective that is organized in a democratic way, i.e., providing participating rights for all members (Sauerland, 2013). This mode of legitimacy is different in hierarchical systems, like firms. Here Transaction Cost Economics (TCE), another important part of NIE theory, has its roots.

The design of institutions in organizations is perceived to be a rational choice, too. As owners and managers have clear ideas of the goal that 'their' organization has to fulfill, institutions in organizations are designed as a means to an end. From the TCE perspective, institutions are therefore designed to be efficient, with reference to the goals to be achieved.

One of the main research questions in TCE is to find the best organizational structure for different kind of transactions. In the governance approach, initiated by Oliver Williamson (1975, 1985), specificity of investment plays an important role when deciding which organization structure to choose for a transaction. Based on TCE arguments, markets (as institutions) are a very efficient organizational structure when generic transactions have to be performed. In contrast to that, a vertical integration is, ceteris paribus, the best organizational structure if highly specific transactions are involved. In this situation, the investor is facing the risk of being exploited by his cooperating partner when using markets to coordinate economic activities. The 'value at risk' in a hold-up situation involving market coordination is the so-called quasi rent. This is the difference in returns on investment in the first best utilization compared to the second best.

To ensure specific transactions from being exploited, institutional precautions are implemented – as a rational choice decision of the people affected. These institutions, e.g., explicit formal contracts, create trust. Investor and partner benefit from this institutional safeguarding because they can realize mutual gains from cooperation.

As in Constitutional Economics, ideologies do not play any explicit role in the TCE line of arguments. Individuals' behavior is framed by (formal) institutions alone, or as Granovetter (1985: p. 489) puts it, the mere focus on formal institutions suggests 'a conception of action and decision carried out by atomized actors.' It is this atomized actor that Granovetter refers to when he emphasizes the relevance of social embeddedness, thus making a point against this mere 'efficiency of

institutions' view. Granovetter (1985: p. 49) emphasizes this point: "Williamson vastly overestimates the efficacy of hierarchical power ('fiat,' in his terminology) within organizations."

Sociological institutionalists in the Granovetter tradition stress the point that not only hierarchical formal and informal structures frame the behavior of people working in an organization and therewith influencing the organization's performance. They point out the relevance of social networks and relations within an organization (cf Section Institutions in Sociology). As people working in organizations are embedded in various social relationships these also have an impact on their behavior. Social relationships and networks are also creating trust and they can also reduce transaction costs. But they are not necessarily created as rational decisions of utility maximizing actors. As discussed in the previous section, the origin of social ties is often evolutionary, the ties being an unintended result of other intended actions. From the trust-building and transaction cost-reducing perspective hierarchical structures and social embeddedness have similar effects. Not surprisingly Williamson (1994: p. 85) admitted: "Transaction cost economics and embeddedness reasoning are evidently complementary in many respects."

Similar to the two examples before, the institutions in the TCE line of arguments originate because of conscious decisions, not just evolutionary. Differently from democratic settings, in this organizational 'choice among rules,' a broad participation (bottom–up) in the decision-making process is not necessary. The institutions in organizations are typically chosen in a hierarchical top–down approach. However, the acceptance of the rules is also relevant for the success of the institutions in this case. If formal institutions are not accepted, they will neither reduce uncertainty nor create trust. They will also not provide a positive contribution in terms of economic success of the organization. Disapproved institutions will not reduce transaction costs and they will not be stable over a longer period of time. Nonacceptance is therefore a source of institutional change.

Except for the differences in the details describing the origin of institutions in the rational choice perspective, another important distinction between constitutional economics and the other theories mentioned refers to the rule setters. As similar the logic of political institutions and economic institutions on the societal level may seem to be, the challenges for the rule setters are completely different. Rules or institutions, respectively, are set for the economy by political decision makers. The actors in the economy have to adapt those rules. Obviously rule setters and potential rule followers are not identical.

Looking at the political institutions (and the institutions framing economic transactions in or between organizations) this is different. In the post-constitutional perspective changes in the constitutions have to be made by the same political decision makers that then have to play according to the new rules. With this perspective, political rules are more like internal rules. Although they are formally fixed, they follow the logic of self-restrictions. This may be one reason why constitutional changes are not often seen in reality, at least in a 'normal mode' of political activity.

Changes in the rules of the political game are more often seen in times of exogenous shocks in the society or in the economy. Crises are often a time for change even on the constitutional level (Acemoglu and Robinson, 2012).

Ideologies and Institutions: Stability and Change

Analyses of institutional change typically refer to changes in the general constitution and/or in the constitution of the economic system. In terms of Acemoglu and Robinson (2012) the specific political and economic institutions are responsible for the economic development and wealth in a certain country. Correspondingly, institutional changes can have a positive or a negative effect on both factors. Theories about institutional change typically distinguish between slow and disruptive changes. Causes for both can be either exogenous or endogenous. A disruptive change in institutions may be caused by endogenous ideological shifts. The corresponding line of arguments runs as follows:

Assumed there is a stable set of political and economic institutions in a country. This situation can be labeled, in terms of NIE, as an institutional equilibrium. Characteristic for such equilibrium is the compatibility between formal and informal rules and the underlying, dominating formal ideology. If these institutions exist for a longer period of time, the structure will result in a self-enforcing stable situation.

Individual actors in the political sector and in the economy will be able to better adapt their behavior to the existing set of rules. They will gain experiences in the handling of the institutions; uncertainty with regard to the logic and application of the institutional structure decreases. In this situation, the existing institutional structure becomes more productive over time, reducing uncertainty in interactions and therewith reducing transaction costs. From an NIE perspective, individuals learn to use the institutions properly; they invest in learning and in building up of successful patterns of behavior.

This process results in a path dependency of the institutional system. The institutions are stable and change is not very attractive for those who have invested in the knowledge of the system. The idea of a path dependency in the institutional structure is not only prominent in NIE analyses (esp. in the work of North) but also in NPI, especially in the historical and the rational approach.

However, institutional changes happen in reality. North argues that institutions' change is caused by a change in relative prices. These changes can be exogenous, caused by, e.g., wars or other catastrophic events for a country. A change in the price structure can also be endogenously caused, e.g., by major technical or organizational innovations.

Changes in relative prices mean a disruption in equilibrium. New technologies disturb the initial equilibrium between the private rate of return and the societal rate of return, which a set of political and economic institutions produces. In the light of the innovation, the private rate of return could be higher now whereas societal rate of return remains stable. In other words, the existing institutional structure limits the benefits for some (innovative) actors. Benefits would be higher if institutions were adapted to the new technological situation. As a result, those who would benefit more have strong incentives to try and change the economic (and sometimes also the political) institutions. In democracies there are established procedures to

enable and manage such institutional changes in the rules of the economic system in a peaceful way.

Based on the argument sketched at the beginning of this section, the new formal institutions must be complementary to the existing informal institutions and the dominating formal ideology. Those represent the guidelines for a successful slow change of the set of formal (economic) institutions (North, 1991b). In contrast to that, disruptive changes are typically caused by exogenous shocks such as wars or the sudden clearance of backlogs of overdue reforms.

A basic prerequisite for a peaceful and successful change in economic institutions are stable workable political institutions. Those ensure conscious negotiations on the political level about changes in the economic constitution. This logic highlights again the importance of a workable democratic system, which is typically characterized by adequately designed political rules to manage (not only institutional) change processes. A lack of workable institutions for political negotiations often results in a reform backlog. In this setting, necessary adaptions of institutions in the light of technological innovations are not going to happen. This may result in a blockade of necessary reforms. The only way to overcome such a blockage is a revolution and therewith a disruptive change in political and economic institutions.

Acemoglu and Robinson (2012: p. 96ff) take up this argument and look for critical junctures in history, where political institutions and economic institutions are not compatible, become unstable, and are changed. Their argument runs as follows: political and economic institutions can be extractive or inclusive. They are distinguished by the allocation of powers that result from those institutions. According to Acemoglu and Robinson (2012: p. 80), extractive institutions allocate the power in politics and economy to a small group of people: "If the distribution of power is narrow and unconstrained, then the political institutions are absolutist, as exemplified by the absolutist monarchies reigning throughout the world during much of history." Inclusive institutions, in contrast, spread the powers broadly and thus enable broad participation in political and economic decisions: "Inclusive economic institutions, […], are those that allow and encourage participation by the great mass of people in economic activities that make best use of their talents and skills and that enable individuals to make the choices they wish." (Acemoglu and Robinson, 2012: p. 74).

Complementary political and economic institutions develop their full potential of productivity if both are inclusive. A combination of extractive and inclusive institutions is generally assumed to be unstable. In this perspective, a lack of compatibility results in instability and change. Change, which is often a disruptive change, is triggered by critical junctures in history and depends on the pre-existing institutions: "The outcomes of the events during critical junctures are shaped by the weight of history, as existing economic and political institutions shape the balance of power and delineate what is politically feasible. The outcome, however, is not historically predetermined but contingent. The exact path of institutional development during these periods depends on which one of the opposing forces will succeed, which groups will be able to form effective coalitions, and which leaders will be able to structure events to their advantage" (Acemoglu and Robinson, 2012: p. 110).

This perspective combines a historical embeddedness of institutional change with an idea of path dependencies and political powers; in other words: it is a combination of a historical NPI- and an NIE-perspective. Apart from that, there is another interesting aspect in Acemoglus and Robinsons book in the context of this article: the term ideology does not play any role in their argument. This is different in some other approaches, for example in NSI (e.g., Vogel, 2009) and also in NIE (e.g., Denzau and North, 1994).

In these lines of arguments ideologies can be a trigger of institutional change: if ideologies change, this can result in an adaptation of institutions to the new majority ideology. Now what may be reasons for a change in ideologies?

Ideologies develop over time. After their origination they get adapted to the reality and to changes in reality. The older an ideology is, the more elaborated it will become. Elaboration means that ideologies become more sophisticated; they are adjusted, for example, if some logical inconsistencies are detected by followers of or opponents to an ideology. Denzau and North as Vogel suggest that a more sophisticated ideology becomes more vulnerable to criticism. This relates to the complexity of an established ideology: The more complex it is, the more likely will inconsistencies be detected, putting the established ideology under pressure. In this logic, the success of an ideology is at the same time its greatest risk.

Both, Denzau and North as Vogel refer to an analogy of this process to Kuhn's idea of shifts in paradigms in sciences. Taking this argument, shifts in ideologies are a result of logical incoherence. Ideological entrepreneurs (Denzau and North, 1994: p. 26) use these incoherencies and develop new, less sophisticated but more consistent competing ideologies. If ideological entrepreneurs are successful, because of support by the media or other opinion leaders, they can cause a shift in ideologies. And as ideologies are the underlying structure of the institutions of a society, the disruptive ideological shift means a crisis for the existing intuitions and therewith triggers institutional change, too. As Vogel (2009: p. 108) puts it: "[…] the failure of ideologically legitimate practices – after a certain point in time – becomes more likely with progressing institutionalization. Put simply, ideologies have an expiration date."

Ideologies and Institutions: Empirical Relevance

When it comes to the discussion about ideologies and institutions in new institutionalism, it is not only a question of theoretical reasoning. As mentioned in Section Ideologies in Social Sciences theories about ideologies and institutions must not only clearly distinguish between both phenomena but also have to put their logics to empirical testing. However difficult it is to model both phenomena properly in empirical analysis, testing is necessary to see a real-world impact of ideologies and institutions.

Since the beginning of new institutionalisms a lot of empirical work has been performed. There are not only qualitative analyses, like case studies but also quantitative, e.g., cross-country comparisons. For example, there is a growing literature in NIE on the impact of economic institutions that ensure freedom to choose within the economic system and also

in the political system. These studies often show that the degree of freedom provided by a set of institutions has a positive effect on the economic development and wealth of nations. Other analyses search for the critical factors that characterize successful constitutions, again with regard to growth and wealth and political stability as a prerequisite for both. Other studies analyze the impact of politics institutions and ideologies on real policy making.

A selection of interesting studies and overviews are listed here:

Qualitative Analyses (Case Studies):

Acemoglu, D., Robinson, J., 2012. Why Nations Fail: The Origins of Power, Prosperity, and Poverty, Crown Businesses, New York.

Elster, J., Preuss, U.K., 1998. Constitutional politics in Eastern Europe. In: Elster, J., Offe, C., Preuss, U.K. (Eds.), Institutional Design in Post-communist Societies: Rebuilding the Ship at Open Sea. Cambridge University Press, Cambridge, pp. 63–108.

Finbow, R., 1993. Ideology and Institutions in North America. Canadian Journal of Political Science 26, 671–697.

North, D.C., 1991. Institutions, Institutional Change and Economic Performance, Cambridge University Press, Cambridge.

Ostrom, E., 1990. Governing the Commons. Cambridge University Press, Cambridge.

Quantitative Analyses

Acemoglu, D., Johnson S., Robinson, J.A., 2005. Institutions as a fundamental cause of long-run growth. In: Aghion, Ph., Durlauf, St. N. (Eds.), Handbook of Economic Growth, vol. 1A, pp. 386–472.

Besley, T., Case, A., 2003. Political institutions and policy choices: evidence from the United States. Journal of Economic Literature 41, 7–73.

OECD, 2014. Economic Policy Reform 2014: Going for Growth Interim Report. OECD Publishing. http://dx.doi.org/10.1787/growth-2014-en.

Persson, T., Tabellini, G., 2003. The Economic Effects of Constitutions, Cambridge.

The quantitative studies often refer to databases about institutional factors in different countries. Some of the databases cited most in NIE are listed below:

http://www.freetheworld.com/ – Economic Freedom of the World Index published by The Fraser Institute.

http://www.freedomhouse.org/ – Freedom in the World Index published by Freedom House.

http://www.heritage.org/index/ – Index of Economic Freedom published by The Heritage Foundation.

Compared to institutional databases, there are only few datasets about ideologies available. One was developed by Berry et al. for the American states. Here are the references:

Berry, W.D., Ringquist, E.J., Fording, R.C., Hanson, R.L., 1998. "Measuring Citizen and Government Ideology in the American States, 1960–93." American Journal of Political Science 42, 327–348. Updated data 1960–2010: http://rcfording.wordpress.com/state-ideology-data/.

Summary

New institutionalisms in social sciences have a clear focus on institutions – and far less on ideologies. However, there are attempts to incorporate ideologies into theoretical and empirical analyses. A major challenge for both, theoretical and empirical work is to have a proper definition of institutions and ideologies – and to distinguish between them. Defining and distinguishing are basic prerequisites to make both concepts fruitful in explaining real-world phenomena.

While there is broad consensus about the logic behind institutions, the picture is more blurred when it comes to ideologies. However, ideologies – referring to Denzau and North – can be seen as a different kind of shared mental models. They can be modeled as the underlying nonformal but logically consistent set of ideas and belief about the structure of the world – and how the world should be structured. Therewith, ideologies are some kind of evolutionary foundation for the set of formal institutions framing the society, the economy, or even organizations. Accordingly, changes (or shifts, respectively) in the underlying ideology can trigger institutional change.

Although there are first results that suggest the importance of formal (and informal) institutions on the performance of political and economic organizations, more work has to be done to explore if and how relevant ideologies are for the new institutionalisms in political science, sociology, and economics.

See also: Cognitive Neuroscience; Contractarianism; Rational Choice Explanation: Philosophical Aspects.

Bibliography

Acemoglu, D., Robinson, J., 2012. Why Nations Fail: The Origins of Power, Prosperity, and Poverty. Crown Businesses, New York.

Buchanan, J.M., 1975. The Limits of Liberty: Between Anarchy and Leviathan. The University of Chicago Press.

Denzau, A.T., North, D.C., 1994. Shared mental models: ideologies and institutions. Kyklos 47, 3–31.

Elster, J., 2000. Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints. Cambridge University Press.

Eucken, W., 1952. Grundsätze der Wirtschaftspolitik. J.C.B. Mohr, Tübingen.

Finbow, R., 1993. Ideology and institutions in North America. Canadian Journal of Political Science 26, 671–697.

Granovetter, M., 1985. Economic action and social structure: the problem of embeddedness. The American Journal of Sociology 91 (3), 481–510.

Greif, A., 2006. Institutions and the Path to the Modern Economy: Lessons from Medieval Trade. Cambridge University Press, Cambridge.

Hall, P.A., Taylor, R.C.R., 1996. Political Science and the Three New Institutionalisms. MPIFG Discussion Paper 96/6.

Heiner, R.A., 1983. The origin of predictable behavior. The American Economic Review 73 (4), 560–595.

Mannheim, K., 1972. Ideology and Utopia. Routledge & Kegan Paul, London.

Mantzavinos, C., North, D.C., Shariq, S., 2004. Learning, institutions, and economic performance. Perspectives on Politics 2 (1), 75–84.

Marx, K., Engels, F., 1990. Die Deutsche Ideologie. MEW Band 3. Dietz, Berlin.
Meyer, R.E., et al., 2009. Ideology and institutions: introduction. In: Meyer, R.E., Sahlin, K., Ventresca, M.J., Walgenbach, P. (Eds.), Institutions and Ideology, Research in the Sociology of Organizations, vol. 27. Emerald Group Publishing Limited, pp. 1–15.
Mullins, W.A., 1972. On the concept of ideology in political science. The American Political Science Review 66 (2), 498–510.
North, D.C., 1991a. Institutions. Journal of Economic Perspectives 5, 97–112.
North, D.C., 1991b. Institutions, Institutional Change and Economic Performance, Cambridge. Cambridge University Press.
North, D.C., 1993. Institutions and credible commitment. Journal of Institutional and Theoretical Economics 149, 11–23.
North, D.C., Weingast, B.R., 1989. Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century England. The Journal of Economic History 49 (4), 803–832.
Olson, M., 1965. The Logic of Collective Action: Public Goods and the Theory of Groups. Harvard University Press.
Ostrom, E., 1990. Governing the Commons. Cambridge University Press, Cambridge.
Powell, W.W., DiMaggio, P.J. (Eds.), 1991. The New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago.
Rawls, J., 1971. A Theory of Justice. Harvard University Press.
Sauerland, D., 2013. Zur Beziehung von Akzeptanz, Gerechtigkeit und Leistungsfähigkeit der sozialen Marktwirtschaft. In: Theurl, Teresia (Ed.), Akzeptanzprobleme der Marktwirtschaft: Ursachen und wirtschaftspolitische Konsequenzen, Schriften des Vereins für Socialpolitik, vol. 336. Duncker & Humblot, Berlin, pp. 49–90.
Simon, H.A., 1957. Models of Man: Social and Rational. John Wiley and Sons, New York.
Simons, T., Ingram, P., 1997. Organization and ideology: Kibbutzim and hired labor, 1951–1965. Administrative Science Quarterly 42, 784–813.
Vogel, R., 2009. Paradigm shifts as ideological changes: a Kuhnian view of endogenous institutional disruption. In: Meyer, R.E., Sahlin, K., Ventresca, M.J., Walgenbach, P. (Eds.), Institutions and Ideology, Research in the Sociology of Organizations, vol. 27. Emerald Group Publishing Limited, pp. 85–113.
Williamson, O.E., 1975. Markets and Hierarchies: Analysis and Antitrust Implications. The Free Press, New York.
Williamson, O.E., 1985. The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting. Free Press.
Williamson, O.E., 1994. Transaction cost economics and organization theory. In: Smelser, N., Swedberg, R. (Eds.), The Handbook of Economic Sociology. Russell Sage Foundation, New York, pp. 77–107.

Ideology, Sociology of

Michel Dubois, GEMASS – CNRS, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The sociology of ideology is part of the general sociology of social beliefs and knowledge. Ideology singularizes itself in the general realm of beliefs not only by the centrality it accords to the nature and the becoming of societies but also by its relational dimension traditionally defined in terms of antagonism and struggle. If the sociological analysis of ideologies was initially developed on a macrosociological level, sociologists have tended progressively to abandon the idea of ideology as a homogeneous and unified belief system for a more localist approach.

Sociological concepts such as *ideology* have multiple lives. Once considered as crucial for the analysis of political and social systems, the concept has since been severely criticized from different perspectives. For some, coming mainly from post-structuralism and postmodernism, it implies a much too rigid demarcation between knowledge and belief, and a too naive conception of truth as correspondence. For others, coming from varieties of comprehensive sociology, the idea of 'false conscious' prevents from understanding the deep feeling of conviction that generally characterizes the relation of any individual to his own beliefs. And most contemporary sociologists agree that the archipelago of ideologies is much more complex than anticipated. The traditional map of large-scale 'systems of interpretation of the historico-political world' (Aron, 1964) does not fit anymore with the complexity and multiplicity of contemporary collective beliefs systems. So should we, as sociologists, definitely forget about ideology? Although the 'end of ideology' (Bell, 1960; Lipset, 1960) has been regularly announced since the end of the twentieth century, it is obvious that sociologists still use the concept to address research questions related to various areas: gender identities (Davis, 2007), values and interests in entrepreneurial science (Abraham and Balling, 2012), professional demarcation (Gieryn, 1983, 1999) and professional politicization (Coulangeon et al., 2012), public opinion and organizational competition (Barnett and Woywode, 2004), religious or political radicalism (Rhys, 1996; Rydgren, 2007), and terrorism and warfare (Turk, 2004; Tosini, 2009).

The first section briefly returns to the origins of the sociological use of the concept of ideology. If the concept preexists to sociology as discipline through the theory of the *idola* developed by the English philosopher Francis Bacon or the French tradition of the *Idéologues* led by Destutt de Tracy, its elaboration starts with Marx's writings and their immediate readers (among them Karl Mannheim). The second section focuses on the specificity of the ideological phenomenon. Just like many other big terms in sociology (*action*, *society*, *community*, etc.), Ideology seems to defy a precise and unique definition: process of production of meanings, signs, and values in social life; process of naturalization of social life; forms of thought motivated by social interests, propaganda, and systematically distorted communication; action-oriented set of beliefs, etc. These are only few examples from a long list of possible definitions (Eagleton, 1991; Van Dijk, 1998). The discussion of classical and contemporary sociological literature allows us to specify the ideological phenomena through three interdependent aspects. The last section focuses on the variety of modes of sociological explanation of ideology. Only few explanatory schemes give real access to the social mechanisms behind its acceptance and diffusion. They generally suppose to adopt toward the ideological beliefs a similar attitude as M. Weber for the magic beliefs or, in a certain extent, Durkheim for the religious ones: reconstruct the 'meaning' of the ideological belief for the social actor (Boudon, 1989).

Origins of the Sociology of Ideology (Marx, Mannheim)

The origin of the sociological use of the concept of ideology can be found in Marx's writings on the coup d'état of Louis Bonaparte (Marx, 1852). The important part of his demonstration consists in showing how, once the riots of June 1848 were over, dissensions within the *parti de l'ordre* – the right wing of the French parliament – helped make this coup d'état unavoidable. This party was composed of many fractions that all, Marx affirms, have a "superstructure of impressions, illusions, ways of thinking and some philosophical conceptions." To understand the origin of these 'ways of thinking' it is essential to identify the 'competition of interest,' the will of each fraction 'to restore its own supremacy.' The landed aristocracy and the industrial bourgeoisie develop different political representations: the former theorizes the benefits of the absolute monarchy, the latter those of parliamentary monarchy. In both cases, Marx suggests, they tend to give a general value to 'ways of thinking,' which actually constitute only the formulation of objective conditions that ensure the domination of a social group over another. Hence the strong relation between analysis of ideology and analysis of the struggle between social groups. Defined as prevailing ideas, ideologies are identified to those of a group seeking to preserve its social domination through their naturalization. The analysis of Marx does not end with this comparison between particular 'interests' and 'ways of thinking,' it also underlines the way in which certain individuals embody ideologies. Describing the 'ideologists' consists mainly for Marx in studying the

'writers,' the 'press,' and more largely 'representatives' of the bourgeoisie. In the present case, the inability of these ideologists to express the real opinion of those they should represent, shortly before the coup d'état, Marx points out, "the representatives and represented had become strangers to each other."

The argument is simple and easily acceptable: when an institution – here a political mode – appears in conformity with the particular interest of a social actor, individual, or collective, this actor generates, directly or indirectly, a whole set of ideas, representations, and theories intended to make the value of this institution acceptable by the majority – even if this value objectively makes real sense only for the social actor considered. When K. Mannheim looks back to this argument in *Ideology and Utopia* (1929), he presents it as a fundamental stage toward the formation of the sociology of knowledge: the first attempt to stress "the role of the position of class and the interests of class in the thought." This attempt rests however on a confusion between two conceptions of the ideology. In the first – particular conception – the 'ways of thinking' are 'mystification more or less conscious of the real nature of a situation whose exact recognition would not be in agreement with the interests' of the ideologist; they are reducible to a psychology of the interests, concern primarily the individuals, and are connected with lies on the moral level, or errors on the epistemological level. In the second – total conception – the same 'ways of thinking' are not 'illusions' deliberately maintained by the individuals according to their more or less immediate interests but elements of a total 'system of significance,' 'knowledge grown out of our experience in the real situations of the life.' One of the rare examples of ideological 'knowledge' in this 'total' conception suggested by Mannheim is the moral judgment related to the loan of money with interest. Formulated and assimilated by the Church like an ethical standard, the morally reprehensible character of the loan with interest becomes ideological, affirms Mannheim, since the social actors try to maintain it out of the social framework to which it is 'adapted': a precapitalist society based, economically and socially, on personal relations of intimacy. Through its attachment to this judgment, the Church expresses its difficulty of thinking social reality independently of a 'system of significances' adapted to a disappeared social framework from now on.

It is possible to read Mannheim's classical analysis on *Conservative thought* (1927) as an exercise of sociology of knowledge. Mannheim connects the spread of conservatism in Germany to three main factors: First, the affinity between the content of conservative thought – defined as a 'style of thought' oriented toward the concrete and qualitative dimensions of reality – and the aspirations of groups opposed to the advent of bourgeois capitalist society. Second, the phenomenon of social polarization reinforced by a middle class too weak to produce a 'synthesis' between extreme views. Third, the socially 'unattached' and economically unstable situation of German romantic intellectuals; a situation that renders them particularly receptive to political and social radicalization. Mannheim's study on *Conservative thought* can be seen as the German counterpart of the one conducted by Alexis de Tocqueville (1856) for France and its intellectuals.

Ideologies as Social Beliefs

This brief detour about the origins of the sociology of ideology suggests that traditionally sociologists associate the concept of ideology with products or properties of thinking shared by social groups. These ideas or systems of ideas have multiple possible forms (principle, theme, theory, myth, style, representation, doctrine, etc.). But whatever their forms, these are products or properties of thinking about which individuals collectively develop a deep feeling of conviction or commitment. Hence their general characterization as social beliefs or clusters of social beliefs. The difficulty that consequently arises is that of the specificity of the ideological phenomenon inside the general realm of social beliefs.

Ideologies and the Legitimate Social Order

Some recent examples of sociological studies of ideologies give us a good starting point for our discussion. Focusing on the interwar period in Austria, Barnett and Woywode (2004) analyze the Viennese newspapers over the period 1918–38. Newspapers are defined as 'organizations' and their relation as an 'ideological competition.' The sociological analysis distinguishes two main dimensions. First, the nature of ideology defined as the 'basis of political identity' of social groups in the Austrian society. Three general ideological 'positions' (*lager*) are identified corresponding roughly to political left (left radical, communist, radical socialist, etc.), center and bourgeois (democratic, Christian social, civil liberal, etc.), and right (patriotic, right radical, national socialist, etc.). Second, the evolving nature of the relation between organizations defined as 'ideological competitors.' The general thesis is that "the strongest ideological competition is likely to take place among close substitutes – adjacent ideologies that are too different to enjoy *esprit de corps* but similar enough to threaten each other's identity claims" (p. 1453). The tension between organizations is described as an ambivalent process merging two opposing tendencies: strong oppositional identity claims and regular attempts of frame alignment on competitors to attract their followers. Leaving the interwar for a more recent period, Rydgren (2007) recalls the emergence of the radical right in Western Europe over the last 20 years. He describes the 'ideological core themes' of this radical right mainly as sociocultural themes. The new radical right gives priority to 'national identity' and builds on the idea of 'ethno-pluralism' elaborated by the French *Nouvelle Droite*: "the notion of ethno-pluralism states that, to preserve the unique national characters of different peoples, they have to be kept separated. Mixing of different ethnicities only leads to cultural extinction (…) contrary to the traditional conception of racism, the doctrine of ethno-pluralism, as such, is not hierarchical: different ethnicities are not necessarily superior or inferior, only different, incompatible, and incommensurable" (p. 244). The emergence of this radical right manifests itself also by an attempt to recode the political space in terms of a struggle against the collusion of all established parties. This oppositional strategy leads to an apparent paradox: generally described as authoritarian, the radical right parties, such as the *Front National* in France, generally present themselves as the new champions of

'true democracy,' which take into account the interests of the common man.

These two brief examples show how the concept of ideology singularized itself in the general realm of beliefs. First, by the centrality it accords to the nature and becoming of society. Ideologies are beliefs or cluster of beliefs about the legitimate political and sociocultural order. Second, by the relational dimension attached to it and generally defined in terms of 'antagonism,' 'struggle,' or 'opposition.' The ideologist and the ideological organization oppose 'what is' to 'what ought to be.' They propose more than a simple interpretation of society: a program to transform it. This transformation may have multiple directions. When Mannheim defines *ideology* vs. *utopia*, by making the former an exclusive vector of conservatism where the latter would be progressivist, he only preserves the initial views of Marx. This vision of the ideology is, however, too restrictive. Not only are the ideologies not the prerogative of the conservatives – Parsons (1959, 1964) distinguishes four types of ideologies: conservative, counter-, reform, and revolutionary – but one ideology can, according to social and historical circumstances, serve various social purposes. Liberalism, socialism, or various contemporary forms of nationalism (Gellner, 1983; Birnbaum, 1997) appear, at a given time of their history, as vectors of social emancipation as well as general principles of immobilism or conservatism.

This double referent – sociopolitical and practical – makes it possible to distinguish between ideology and other types of beliefs, religious ones in particular. The ideologist speaks in the name of reason. Its truth claims positive values of rationality and is diffused by suggestion, persuasion, but much more rarely by revelation. Many sociologists (Bendix, 1964; Gouldner, 1976) thus establish a strong parallelism between the development of the ideologies at the end of the eighteenth century and the emergence of modernity defined as a 'rational' social order in opposition to a 'traditional' social order. This by no means implies that the religious elements disappear from the ideological discourse. As religion exists as a 'cultural system' it may constitute a 'shaping force for political life' (Rhys, 1996). In France, the recent public debate on the reform of bioethics laws (Dubois, 2012) or the public demonstrations against the legalization of gay marriage show that coalitions of religiously based activists regularly mix religious and political arguments. They use the cultural force of religious symbols to attempt to modify the political agenda. From a more radical perspective, the ideologies of numerous fundamentalist movements around the world build on selective and 'innovative' interpretation of sacred texts. Such interpretations contribute to produce a culture of political violence in the name of creating or restoring a legitimate social order (Turk, 2004).

Disassembling Ideology: from Macro- to Microsociology

If the sociological analysis of ideologies was initially developed on a macrosociological level (Apter, 1964; Geertz, 1964), it is becoming increasingly microsociological. This is explained by the fact that sociologists have tended progressively to abandon the idea of ideology as a homogeneous and unified belief system. When Aron (1964) analyses the contemporary political modes, he characterizes ideology as an explicit system of beliefs developed around a core of descriptive and prescriptive proposals. The dominant features of this system are unity, coherence, and totalizing character – the capacity to potentially bring a response to any interrogation. Whatever its heuristic value, this representation of ideology as a 'monolithic' unified block rarely fits with empirical data. Ideological beliefs contain generally irreducible tensions and contradictions. This conclusion is drawn from several perspectives. First, from observation of ideological production itself. Sociologists insist on the 'compromises' and the 'adjustments' necessary to hold in a same unit theoretical logic and the principle of action (Seliger, 1976). They describe the production of ideology as a 'tinkering' process during which the ideologist tries to satisfy the demand of meaning of his public without ever being able to exert a true control on his resources (Bourricaud, 1980). Second, it is drawn from observation of the organization of ideological components. As recently emphasized by Achterberg and Houtman (2009), sociologists cannot presuppose the coherence of ideological belief systems: "knowing someone's values on economic matters does not lead to a correct prediction of what one will think about cultural matters. There is no or very little coherence between the two value dimensions" (p. 1650). Beside this multidimensionality, the degree of coherence of ideology is not socially uniform. Research demonstrates that ideological components are more consistently ordered within the cultural, intellectual, and political elite than in other social groups. Finally, it is drawn from observation of the evolution of ideologies. Modern ideologies are different from the ones observed a century ago. For Boudon (1999), modern ideologies evoke the image of archipelago, while the old ideologies evoked rather the idea of a continent. "We have ideologies, writes Boudon, as to what should be done about unemployment, educational opportunities, fighting against crime or drug addiction and on a myriad of subjects, as well as how it should be done. But these theories are weakly related to one another. We have ceased to believe that they could be derived from an all-encompassing theory. We have all kinds of local ideologies; we no longer believe in general ideologies."

These local ideologies are sometimes defined in terms of professional belief systems. In his study on 'boundary work' in science, Gieryn (1983) observes that when scientists confront the public, they endow science with characteristics selected for an ability to advance 'professional interests.' They use available 'cultural repertoires' to produce profitable 'ideological self-descriptions.' Hence the risk for the sociologist to be too naive about these distorted self-descriptions. More recently, following the Straussian definition of professional arenas as 'political space for debate,' Coulangeon et al. (2012) consider as crucial not to limit the analysis of professional ideologies to political or trade union arbitrations or publicized moments of collective action. Using a questionnaire survey of French police officers of all ranks, they identified the 'ideological dissensions' within the French police on how police work should be carried out. Three general ideological positions are discussed – repressive, median, preventive – and the observed dominance of the first

(repressive) substantiates, according to the authors, "an ideological closeness between profession and the (political) right" (p. 375).

Ideology and Scientific Knowledge

As Gieryn stresses the difference between 'self-description' and the objective reality of practices and values, he reaffirms the traditional distinction between ideology and knowledge. Ideology involves elements of distortion from reality. Well-known authors such as Marx, Mannheim, Parsons, Althusser, Aron, etc. converge to consider ideologies as forms of 'distorted knowledge,' 'false beliefs,' and 'deviation compared to scientific objectivity.' The ideologist produces a representation of the social order, he wishes to be socially considered as the unique depository of a true meaning, but both representation and meaning are objectively illusory. Despite this relative consensus on the demarcation between ideology and knowledge, the reality of the ideological phenomenon is more complex. The ideologist shares in fact with the social scientist the will to produce a general discourse on society. It is not infrequent that certain sociologists consider their analysis as the methodical prolongation of a preexistent ideological point of view. The historical examples are numerous and well known: openly conservative, Frederic Le Play theorizes the traditional form of family – *famille souche* – as 'the social unit per excellence'; convinced liberal Herbert Spencer proposes an evolutionary model closely associated with the diffusion of 'social Darwinism,' etc. Inversely, a sociological theory or paradigm conceived independently of any ideological commitment can, when it is exploited out of its immediate context of validity, generate ideological skid. The mechanism is once again well known: it consists mainly to project on the theory or sociological paradigm the totalizing character of ideologies and by doing so asks those theories and paradigms to answer any possible interrogation.

The significant development, since the beginning of the 1970s, of the social studies of sciences made it possible to highlight the importance of economic, cultural, and political factors on the course of scientific research. These studies show on the one hand how the scientific institution satisfies a social demand, and on the other hand how the scientists integrate in the course of their cognitive practices different cultural elements, in particular, extrascientific 'interests.' These studies are now part of a general program of political sociology of sciences (Frickel and Moore, 2006). Recent works by Kinchy et al. (2008), Lave et al. (2010), and Abraham and Ballinger (2012) try to show how new variants of liberalism (neoliberalism) influence directly or indirectly the content of scientific expertise in the biotechnology area. For the latter, the evolution of carcinogenicity testing systems coincides much more with the 'interests' of the pharmaceutical industry and drug regulatory agencies than with public-health protection. Oreskes and Conway (2010) claim that controversies over tobacco smoking, acid rain, the hole in the ozone layer, or global warming are not only strictly rational disputes about the robustness of empirical proof but also ideological controversies within the scientific community.

Explanatory Schemes of Ideologies

How should we explain the production and diffusion of ideologies? Sociologists have proposed numerous explanatory schemes. Without trying to be exhaustive, it is useful to briefly describe four of the most common schemes.

One key aspect of the production and diffusion of ideology is frequently related to the socialization process. Socialization is a learning process that begins shortly after birth for its primary dimension. Studies on 'gender ideology' (Bolzendahl and Myers, 2004; Davis, 2007) – that is, cluster of beliefs regarding the appropriate roles of women and men in society – emphasize the importance of personal and familial dimensions. Young women with high self-esteem or personal confidence have more probability to break away from traditional definitions of 'gender roles.' But children's exposure to egalitarian behavior models seems as much as important as strictly personal factors: growing female labor participation contributes to socialize children to more egalitarian methods of dividing household responsibilities between women and men. The socialization process is also explicitly referred to in its secondary dimension by sociologists of professions. Observing the unequal distribution of ideological positions within the French police, Coulangeon et al. (2012) consider that the daily exercise of a job acts as a socialization process through which opinions are progressively reconfigured: "giving orders, maintaining public order, making arrests and driving retain a positive effect on adherence to a repressive ideology and a repellent effect on adherence to a preventive ideology, while contact with the public places police officers on the preventive side" (p. 371).

A second factor regularly cited by sociologists of ideology is *social interest*. As previously reminded, the Marxist tradition associates the production of ideology with the ruling class interests that are supposed to dominate the conscience of other social groups. This causal approach of beliefs claims that if the dominated class accepts blindly social representations that contribute to reproduce its subordination, the dominant class generates social representations which maintain its domination. The example of the study of the 18 *brumaire* shows, however, that Marx develops, in certain empirical writings, a less simplistic representation of the production process of ideologies. They are described as more than strategical theories imposed to social actors: representations of limited validity inspired by reality such as it can be observed from a given social position. Although the influence of Marxism in sociology has greatly declined, the notion of social interest associated with economic and cultural position remains sociologically useful. Focusing on scientific professions, Gieryn (1983, 1999) observes that scientists have at their disposal multiple repertoires to produce ideological descriptions of science. The role of interests is to guide the selection between those repertoires: "Ideologists are able to endow science with just those characteristics needed to achieve professional and institutional goals, and to change these attributed characteristics as circumstances warrant" (1983: p. 792). Ideologies are conceived mainly as 'professional means' to gain authority, power, and resources. In their analysis of the ideological competition between Austrian newspapers, Barnett and Woywode (2004) expressly relate the

ideological positions to different kinds of interests. During the interwar period, the political center, for example, included secular liberal (according to the nineteenth century definition) capitalist interests. And a centrist organization such as a purely bourgeois newspaper shows a general tendency to sacrifice its political role in order to pursue its material interests. Building on Converse's theory (1964), Achterberg and Houtman (2009) consider that economic and cultural interests interact to affect ideological (in) coherence: "If a person is in a weak position (which means: low level of education, low income, high risk of unemployment) and suffers from economic insecurity, he or she is more inclined to be pro-welfare state, pro-state intervention in the economy, pro-financial redistribution and have egalitarian values" (p. 1643).

A third factor that drives the production and the diffusion of ideologies is conceived by sociologists in terms of 'passions' or 'feelings.' In his *Trattato di sociologia generale*, Pareto (1916) considers that the ideologies – he speaks of 'derivations' – are primarily the effect of deep feelings remaining inaccessible to the conscience. More precisely Pareto affirms that the social actor wants to be convinced of the value of his passion, and produces derivations for this purpose. The ideologies would be thus the epiphenomenal expression of dominant passions. If the Paretian model of ideology remains generally ignored by the contemporary sociological tradition, sociologists still use the category of feeling in various contexts. The emergence and success of contemporary radical ideologies are sometimes described as the consequence of relative deprivation feelings. As emphasized by Merton (1964), the concept of relative deprivation comes from the study of the 'sentiments and attitudes' of American soldiers. Relative deprivation theory focused on feelings caused by disappointing comparisons with one's own past or with social reference groups. Rydgren (2007), for example, considers that the feeling of frustration is related to the emergence of the new radical right-wing parties in Europe. The 'losers' in the postindustrialization processes are described as 'anxious,' 'bewildered,' 'insecure,' and 'resentful': "sentiments that may be channeled into support for policy proposals that stress the need to return to the traditional values of the status quo ante" (p. 248). Following the same general path, sociological studies of hate crimes and terrorism indicate that ideologically radical behaviors are not a direct response to poverty or social disadvantage but much more to "feelings of indignity and frustration developed in repressive political environments" (Turk, 2004).

Finally, a fourth general type of sociological explanation consists mainly to adopt toward the ideological beliefs a similar attitude as M. Weber for the magic beliefs or, in a certain extent, Durkheim for the religious ones: reconstruct the 'meaning' of the ideological belief for the social actor by seeking the system of reasons or preferences that this actor produces to justify to himself his adhesion. The rationality considered by the sociologist is not reducible to an instrumental one but is also quite different from the 'derivations' described by Pareto. The sociological analysis opens on a set of 'subjective reasons' that represent the necessary conditions of the ordinary thought. These reasons are of general but nonuniversal validity and manifest themselves on different levels: utilitarian, axiological, traditional, or epistemic. Inspired by studies of Downs (1957) and Simon (1982), the restricted theory of the ideology elaborated by Boudon (1989) stresses the necessary 'comprehensive' dimension of the sociological analysis. When some people adopt statements or ideas that appear to the sociologist as false or doubtful, the sociologist has to consider that this may derive from the fact that, given the situation, these beliefs are to them the conclusions of reasons they perceive as 'strong.' This strategy is directly linked to an active theory of ordinary knowledge: the social actor produces knowledge, not simply by contemplating reality, but by interpreting it from its social position, and more generally from questions that are only seldom directly provided by reality itself. This last approach has a certain number of objective advantages. In particular, it makes it possible for the sociologist to give an account of important subjective data: the feeling of conviction and not of interiorization that generally characterizes the relation of any individual to his or her own beliefs.

See also: Collective Beliefs, Sociological Explanation of; Ideological Constraint: History and Current Status of the Concept; Mannheim, Karl (1893–1947); Marxism in Contemporary Sociology.

Bibliography

Abraham, J., Balling, R., 2012. The neoliberal regulatory state, industry interests, and the ideological penetration of scientific knowledge: deconstructing the redefinition of carcinogens in pharmaceuticals. Science Technology and Human Values 37 (5), 443–477.

Achterberg, P., Houtman, D., 2009. Ideologically illogical? Why do the lower-educated Dutch display so little value coherence? Social Forces 87 (3), 1649–1670.

Apter, D. (Ed.), 1964. Ideology and Discontent. The Free Press of Glencoe, London.

Aron, R., 1964. La lutte de classes: Nouvelles leçons sur les sociétés industrielles. Gallimard, Paris.

Barnett, W., Woywode, M., 2004. From Red Vienna to the Anschluss: ideological competition among Viennese newspapers during the rise of the national socialism. American Journal of Sociology 109 (6), 1452–1499.

Bell, D., 1960. The End of Ideology. Free Press, Glencoe, IL.

Bendix, R., 1964. The age of ideology. Persistent and change. In: Apter, D. (Ed.), Ideology and Discontent. The Free Press, Glencoe, IL.

Birnbaum, P., 1997. Sociologie des nationalismes, first ed. Presses Universitaires de France, Paris.

Bolzendahl, C., Myers, D., 2004. Feminist attitudes and support for gender equality: opinion change in women and men, 1974–1998. Social Forces 83, 759–790.

Boudon, R., 1989. The Analysis of Ideology. University of Chicago Press, Chicago.

Boudon, R., 1999. Local vs general ideologies: a normal ingredient of modern political life. Journal of Political Ideologies 4 (2), 141–161.

Bourricaud, F., 1980. Le Bricolage Idéologique. Presses Universitaires de France, Paris.

Converse, P., 1964. The nature of belief systems in mass publics. In: Apter, D. (Ed.), Ideology and Discontent. The Free Press, Glencoe, IL.

Coulangeon, P., Pruvost, G., Roharik, I., 2012. Professional ideologies: a latent class analysis of police officers opinions on the role of the police. Revue Française de Sociologie 53 (3), 347–380.

Davis, S., 2007. Gender ideology construction from adolescence to young adulthood. Social Science Research 36, 1021–1041.

de Tocqueville, A., 1856. L'Ancien Régime et la Révolution. Garnier-Flammarion, Paris.

Downs, A., 1957. An Economic Theory of Democracy. Harper, New York.

Dubois, M., 2012. Social Dynamics of Biomedical Research. An Actionist Perspective on the Sociology of Science. Bardwell Press, Oxford.

Eagleton, T., 1991. Ideology. An Introduction. Verson, London.

Frickel, S., Moore, K., 2006. The New Political Sociology of Science. Institutions, Networks, and Power. The University of Wisconsin Press, Madison, WI.
Geertz, C., 1964. Ideology as a cultural system In: Apter, D. (Ed.), Ideology and Discontent. The Free Press, Glencoe, IL.
Gellner, E., 1983. Nations and Nationalism. Cornell University Press, Ithaca, NY.
Gieryn, T., 1983. Boundary-work and the demarcation of science from non-science: strains and interests in professional ideologies of scientists. American Sociological Review 48 (6), 781–795.
Gieryn, T., 1999. Cultural Boundaries of Science: Credibility on the Line. University of Chicago Press, Chicago, IL.
Gouldner, A., 1976. The Dialectic of Ideology and Technology. Seabury Press, New York.
Kinchy, A., Kleinman, D., Autry, R., 2008. Against free markets, against science? Regulating the socio-economic effects of biotechnology. Rural Sociology 72 (3), 147–179.
Lave, R., Mirowski, P., Randalls, S., 2010. STS and neoliberal science. Social Studies of Science 40 (5), 659–675.
Lipset, Seymour M., 1960. Political Man: The Social Bases of Politics. Doubleday, Garden City, NY.
Mannheim, K., (1929) 1991. Ideology and Utopia. Routledge and Kegan Paul, London.
Mannheim, K., (1927) 1952. Conservative thought. In: Kecskemeti P. (Ed.), Essays on the Sociology of Knowledge by Karl Mannheim. Routledge and Kegan Paul Ltd, London.
Marx, K., 1852. The Eighteenth Brumaire of Louis Bonaparte in Marx and Engels, Collected Works, vol. 11. Lawrence & Wishart, London.
Merton, R., 1964. Social Theory and Social Structure. The Free Press of Glencoe, New York.
Oreskes, N., Conway, E., 2010. Merchants of Doubt. Bloomsbury Press, New York.
Pareto, V., 1916. Trattato di Sociologia Generale. G. Barbéra, Firenze.
Parsons, T., 1959. An approach to the sociology of knowledge: Transactions of the Fourth Congress of Sociology, Milan and Sressa. In: Curtis, J., Petras, J. (Eds.), 1970 The Sociology of Knowledge: A Reader. Gerald Duckworth, London, pp. 25–49.
Parsons, T., 1964. The Social System. Free Press, Glencoe, IL.
Rydgren, J., 2007. The sociology of the radical right. Annual Review of Sociology 3, 241–262.
Rhys, W., 1996. Religion as political resource: culture or ideology? Journal for the Scientific Study of Religion 35 (4), 368–378.
Seliger, M., 1976. Ideology and Politics. Allen & Unwin, London.
Simon, H., 1982. Models of Bounded Rationality. MIT Press, Cambridge, MA.
Tosini, D., 2009. A sociological understanding of suicide attacks. Theory, Culture & Society 26 (4), 67–96.
Turk, A., 2004. Sociology of terrorism. Annual Review of Sociology 30, 271–286.
Van Dijk, T., 1998. Ideology: A Multidisciplinary Approach. Sage Publications, London.

Ignorance, History of Concept

Barbara Hoenig, Innsbruck University, Innsbruck, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term 'ignorance' designates nonknowledge, what is unknown, or absent in empirically valid knowledge. Starting from a distinction of kindred concepts in different scientific disciplines, particularly the concept's role in theories of social action, in theories of knowledge, and in organizational studies are considered. Finally, current developments in agnotology and in the sociology of ignorance are discussed.

Proverbs like "ignorance is bliss" and "what you don't know can't hurt you" reflect the fact that one should be aware of what not to know. In scientific discourse, ignorance designates what is unknown, and absent in empirically valid knowledge, in the sense of nonknowledge. Thus, ignorance is held distinct both from the act of 'ignoring' what is known, and from 'error', 'risk', and 'false beliefs' as forms of imperfect or limited knowledge (Moore and Tumin, 1949: p. 788); rather, ignorance can be regarded as a "more or less hidden component of situations usually discussed in other terms" (Moore and Tumin, 1949: p. 795). Ignorance's dynamic role in social change seems to depend on the recognition of its existence and a subsequent searching for more adequate knowledge. In a similar vein, the specification of ignorance has been regarded as prerequisite to the acquisition of knowledge (Merton, 1987). As subsequently will be shown, ignorance is related to kindred concepts in different scientific disciplines; more specifically, the author takes account of the role of ignorance in theories of social action and in theories of knowledge and science studies. After outlining how ignorance has been reflected as strategic resource in organizational contexts, some recent developments in agnotology and in the sociology of ignorance are considered.

Ignorance and Kindred Concepts

Ignorance and related concepts form an important phenomenon in several neighbor disciplines of the social and behavioral sciences. Think of Freud's 'unconscious' as path-breaking notion in psychology and cognitive science till today: He coined the term in trying to understand which type of nonconscious knowledge is involved in forgetting a word, producing a linguistic lapse, or understanding a joke (Freud, 1901). Ignorance is close to terms such as oblivion, secrecy, censorship, and taboo as well, which were so central in the works of some social anthropologists (Douglas, 1966). Complex relationships between ignorance and actors' rational choices are reflected in economics, where incomplete knowledge and asymmetric information embodied in situations of interdependent decision making creates dilemmas and opportunities for strategic action (Schelling, 1960). In political science, the role of discretion for bureaucratic autonomy has been a topic in discussing how politicians make use of laws to shape the policy process (Huber and Shapin, 2002). In philosophy, skeptical reasoning about knowledge resulted in critically articulating a remarkable amount of ignorance implied in ordinary language (Unger, 1975). In political theory, the concept 'veil of ignorance' has been used in a thought experiment for granting fair, unanimous decisions, assuming that in a fictional situation of decision making, rational actors' reasoning on social order would be oriented toward what anyone would wish, do, or agree with, under the condition that actors would not know about their own social position in it (Rawls, 1971). In science and technology studies, the assessment of more expectable, calculable 'risks' has given rise to accounting for ignorance as well, while kindred 'uncertainty' has been understood as a subbranch of it (Smithson, 1989: p. 9). According to Michael Smithson's taxonomy, ignorance is related both to relevance and error, the latter indicating distortions or incompleteness in knowledge; in the formal sciences, incompleteness is reflected in terms of absence, ambiguity, probability, vagueness, fuzziness, and nonspecificity (Smithson, 1989: p. 9). Bearing on ignorance as a social scientific concept, distinctions such as 'structural' (Moore and Tumin, 1949), 'specified' (Merton, 1987), 'science based' (Ravetz, 1987), or 'strategic ignorance' (McGoey, 2012) have been suggested. Recently, Andrew Abbott has focused on the need to specify the quality of ignorance one is subject to in order to enjoy the positive function of several 'varieties of ignorance' (Abbott, 2010). Investigating ignorance as manifest in different sources of scientific knowledge, he has drawn a distinction between three types of ignorance connected with corresponding actor groups involved in knowledge production: the 'amateur's ignorance' of central cognitive skills, the 'professional's ignorance' of collateral literature, and the experts' 'synthetic ignorance' of facts, as an inability to perceive them other than in the light of one's own paradigm or theory (Abbott, 2010). While most social scientific approaches toward understanding ignorance were independently developed by individual scholars, in the following the author discusses some of them either as part of social theories of action or within the frame of a sociology of knowledge.

Ignorance in Theories of Social Action

Very often in life, one takes action on the basis of rather little knowledge, concerning both structural conditions and consequences of one's activities. Ignorance affects one's capacities to act in a social structure constraining one's choices; it is also

bound to unknown effects, contingent outcomes, and unanticipated consequences of action, in contrast to calculable risks, controllable effects, and desirable aims. Different schools or paradigms of social theory have found divergent answers to how ignorance is related to social action, and mostly they have dealt with it as a rather implicit issue of concern (Schneider, 1962; Gross, 2007). As Louis Schneider (1962) has shown, the 'classics' of social theory have reflected ignorance mainly in terms of mystification and in sharp contrast to notions of enlightenment, truth, and knowledge; While Marxism and in part Karl Mannheim's sociology of knowledge focused on the cognitive distortion of scientific truth by false consciousness and ideology, Max Weber regarded science as a way of demystifying the magic ignorance of the world by intellectualization and rationalization (Louis Schneider, 1962). In contrast to the 'savage' who 'knows incomparably more about his tools', although for whom magical powers existed as well, today one usually does not have much knowledge about, but rather 'count on', the behavior of technical means such as the streetcar, believing that "one can, in principle, master all things by calculation" (Weber, 1917/1919: p. 87). Georg Simmel reflected ignorance as part of an increasing discrepancy of objective and subjective culture; as one of few classical sociologists, he dealt with ignorance as inevitable part of the 'secrecy' inherent in social relations (Simmel, 1908).

Wilbert E. Moore and Melvin Tumin (1949) acknowledged the inescapable character of ignorance intrinsic to social organization with the aim of examining its structural functions in varying institutional contexts. According to them, one function of ignorance is maintaining privilege in asymmetric power relations as between lays and experts, specialists and potential competitors, and in ethnic and class stereotyping as well. As specific kind of institutionalized ignorance they mentioned the case where ignorance of future chances is institutionally used "as a device to create anxieties and spur activity in a competitive system" such that a worker is "expected to be motivated by his insecurity" (Moore and Tumin, 1949: pp. 793f.). However, they also acknowledged institution's ignorance of an individual's irrelevant personal characteristics as precondition for preserving fair competition, for instance, in recruitment by bureaucratic organizations or in peer review procedures of science. Apart from that, the attractiveness of any 'new experience' in life depends on one's ignorance or uncertainty of its outcome (Moore and Tumin, 1949: pp. 793f.). Robert K. Merton is one of the most important theorists focusing on the role of ignorance in social action: in discussing unanticipated consequences of purposeful action (Merton, 1936), in reflecting the role of serendipity (Merton and Barber, 2004), and 'specified' ignorance in scientific progress (Merton, 1987). Unanticipated consequences may arise through the limited knowledge about bases of social action and a correct anticipation of its consequences. As barriers to correct anticipation, Merton mentioned not only ignorance, but also error, imperious immediacy of interest, basic values, and self-defeating prediction (Merton, 1936). Moreover, he dealt with the role of ignorance in the dynamics of scientific discovery and in the relationship between empirical research and sociological theory, as indicated by the 'serendipity pattern' in science, designating "the fairly common experience of observing an unanticipated, anomalous and strategic datum which becomes the occasion for developing a new theory or for extending an existing theory" (Merton, 1957: p. 104).

Symbolic interactionists and phenomenological sociologists have produced interesting accounts of ignorance as well (Berger and Luckmann, 1967; Schütz and Luckmann, 1973). Underscoring the idea that social interaction and communication are made possible by actors sharing a stock of knowledge, meanings of significant symbols, and negotiated definitions of situation, they also have mentioned 'structures of ignorance' as part of relative or even fundamental opacities in everyday life and as 'potential knowledge' (Schütz and Luckmann, 1973). However, the empirical case of nonshared experience and asymmetries in knowledge seems to posit an analytical problem for phenomenological approaches (Smithson, 1989: p. 235). In contrast to that, Erving Goffman's (1959) metaphor of the backstage accounts for the fact that some types of social interaction are only possible when social actors are at least partially ignorant of other's intentions. Ethnomethodologists (Garfinkel, 1967) and social constructivists tried to explicate actors' practical competencies of taking action in everyday life, underscoring its nondiscursive, 'tacit dimension' (Polanyi, 1966) as precondition for successfully accomplishing it. Thus, they regarded ignorance as basic characteristic of microsocial interaction and of social reality in general.

Ignorance in Theories of Knowledge

Similar to his distinction of manifest and latent functions of social action, Merton (1987) has characterized ignorance's function in science as a social institution, drawing on the importance of its recognition that turns formerly unrecognized then to 'specified' ignorance. Contrary to the normative expectation that an individual scientist shows humility through confessing ignorance by acknowledging one's knowledge limitations, Merton clarified the cognitive value of specifying what exactly is not known: According to him, specification of ignorance precedes scientific problem finding, a process of selectively perceiving complex phenomena in a dialectic of awareness of some and negligence of other aspects (Merton, 1987: p. 8). He illustrated that practice by discussing successive stages of theory development in the sociology of deviance from the 1920s onward. Moreover, he was interested in whether scientific disciplines differ in their extent of routinely specified ignorance and how this affects the growth of knowledge in these (Merton, 1987: p. 8). At quite the same time, the term 'science-based ignorance' has been coined by philosopher John Ravetz (1987) who developed a similar thesis in talking about relevant, usable forms of nonknowledge.

The classical paradigm of the sociology of knowledge, formulated by Karl Marx and Karl Mannheim and specified by Merton, focuses on the relationship of cognitive or 'mental productions' to designated 'existential bases' as given by social structure. In trying to articulate a parallel paradigm for the sociology of nonknowledge, David Weinstein and Michael Weinstein asked "why and how people exclude certain cognitions and judgments from their awareness" (Weinstein and Weinstein, 1978: p. 152). They defined their suggested metatheory in terms of 'ground' (epistemologies or general theories

of knowledge), 'resultant' (types of actor's nonknowledge involved), and an intervening process between both, in which the distorted or inadequate cognition arises (Weinstein and Weinstein, 1978: p. 154). Concerning the epistemological ground dimension for nonknowledge, they made a distinction between 'need-grounded' and 'possibility-grounded' conceptions of nonknowledge, the first term indicating universal motivators of social action, the second term bearing upon not necessarily universalistic options, purposes, and goals of human action. Resultant types of nonknowledge they interpreted in analogy to these types of grounds, as 'need-grounded misinterpretations', on a continuum ranging from falsification to distortion presupposing some knowable true description of social life, versus 'possibility-grounded inattentions', neglect, and inadvertence, assuming a 'standard field of possibilities' rather than only one true description of a situation. The dynamic component of the suggested paradigm referred to "a process through which mistaken cognition or inattention takes place" (Weinstein and Weinstein, 1978: p. 156); as examples of explaining mistaken cognition or the narrowing of possible alternatives, Freudian ego-defense mechanisms or social sanctions in structural-functionalism were given. Although in Weinstein and Weinstein's variegation of the Mertonian paradigm for the sociology of knowledge particularly the conceptualization of knowledge bases is confusing, it becomes apparent that the sociology of nonknowledge proposed shall not restrict itself to institutional analyses of social structure but to directly make judgments on the meaningfulness of nonknowledge's cognitive content (Weinstein and Weinstein, 1978: p. 159).

The authors' explicit cognitive interest and their strong focus on describing ubiquitous processes of not-knowing characterizes more recent constructivist sociology of scientific ignorance as well. For instance, Karin Knorr-Cetina (1981) emphasized that knowledge and ignorance, and the criteria by which they are established, are 'socially constructed' in interaction processes and negotiations among actors that are highly contingent, contextually and situationally determined, and to a large extent dependent on the tacit, inexplicable, and implicit character of practical know-how. The interest in the formal structure of that social fabric has inspired lots of empirical studies on knowledge production in microsocial scientific contexts such as laboratories. Within the wide range of the sociology of knowledge, the implied epistemological relativism of that position remains questionable; in addition, the focus on local practices of knowledge production often neglects the influence of institutional contexts and communal procedures of knowledge evaluation on what counts as knowledge. From the perspective of an institutionalist sociology of the sciences, Richard Whitley (1984) has dealt with the question of how ignorance, uncertainty, and the organization of scientific fields affect one another, distinguishing between 'technical' versus 'strategic uncertainty'. While the former refers to scientists' disagreement on the reliability and understanding of research techniques, strategic uncertainty indicates disagreement about and pluralism of intellectual priorities, significance, and ways of tackling research problems. According to Whitley, high technical uncertainty usually entails high strategic uncertainty, but the reverse does not generally hold. Since one of ignorance's most important aspects refers to its function in institutional contexts, here the author takes a short digression accounting for the role of ignorance and its strategic uses as a resource in organizations.

Strategic Ignorance as an Organizational Source

When Schneider (1962) has particularly drawn attention on structural functionalism's usefulness for analyzing the functions of ignorance in institutional contexts, he did so by developing what Abbott (2010: p. 186) later has called a 'liberating' view on ignorance: aiming at specifying those conditions where ignorance might lead to positive, more democratic results. Apart from taking ignorance's protective function exercised in the context of immediate interaction, or its consequences for specific persons in a situational interaction context into account, like in parent–child interaction or a physician–patient relationship, Schneider thought of more institutionalized examples of ignorance's consequences for a wider social structure. Thus, he mentioned Burton Clark's (1960) discovery of the 'cooling out function' of higher education institutions, as informal practices of discouraging and 'adjusting' students' expectations toward aspiring less ambitious goals, and Peter Blau's (1955) study on social dynamics in bureaucratic institutions, in which 'gain through indirection', as one form of ignorance, is often an implicit practice of concern. In his study, Blau outlined how institutional requirements can be reached by bureaucratic mechanisms neutralizing possible bias of actors that, unwittingly or not, might influence decision-making practices by prejudices against minorities. "In functional terms, a social pattern with the latent function of impartiality, statistical records, was more effective than a different pattern with the same manifest function, antidiscrimination rules. Since latent function is defined as an unintended consequence, the concept implies that the social behavior involved is motivated by factors other than the function under consideration" (Blau, 1955: p. 81). To this insight Schneider added the relevance of ignorance involved, conditioning more effective, nondiscriminatory placements, when bureaucratic actors "simply did not know that their action was bringing those things about" (Schneider, 1962: p. 501).

In the sociological study of organizations, ignorance has been considered as fundamental resource of any bargaining relationship between strategic actors, characterizing social power as their structural capacity to act by controlling 'zones of uncertainty' (Crozier and Friedberg, 1977). James March and Herbert Simon identified organizational norms for dealing with ignorance and uncertainties, anomalies, and risk, in constructing a self-confirming environment by 'absorbing' particular questions, inquiries, or communications about these (March and Simon, 1958). Taking the 'organized anarchy' of decision making at the university as an empirical example, Michael Cohen et al. (1972) formulated their 'garbage can model of organizational choice'. The authors underlined that under conditions of uncertainty, indicated by unclear technology, fluid participation, and problematic preferences, decision-making practices in organizations are more complex than suggested by classical analyses of rational action in these structural contexts. Rather, they suggested conceiving these as

contingent outcomes of four intersecting variables: a stream of choice, of problems, of energy from participants, and a rate of flow of solutions. Although Smithson's attempt to reconcile formal with social and behavioral sciences sometimes appear as rather abstract analysis, those interested in further research on ignorance in organizational research may find useful hints there (Smithson, 1989: pp. 239ff.).

Similar to these early findings in the sociology of organizations, recent science and technology studies have attempted to identify organizational strategies for constructing ignorance by the denial, dismissal, diversion, and displacement of 'uncomfortable knowledge' (Rayner, 2012) in tension or outright contradiction with self-consistent institutional definitions of situations. Linsey McGoey (2012) has investigated the strategic use of ignorance as part of organizational pressures where organizations' members seek to preserve and cultivate ignorance rather than to dispel it. She has analyzed strategies how members mobilize strategic unknowns of a situation as a tool of social control, command of resources, denial of liability, and assertion of expert control in the face of unpredictable outcomes. In contrast to Abbott's (2010) emphasis on different qualities of ignorance, and complementary to his remark that scholarly practices of articulating knowledge claims are accompanied by those of acknowledging ignorance, McGoey underscored the need to take scales of ignorance into account in order to recognize its value as an organizational resource: "(S)trategic ignorance is often more institutionally advantageous the more widely it is individually mobilized … because of the shared willingness of individuals to band together in dismissing unsettling knowledge" (McGoey, 2012: p. 570).

Agnotology and the Current Sociology of Ignorance

In public debates on science and technology, for a long time ignorance has mainly been considered as the lay's counterpart in relation to an expert's knowledge. Since the 1970s, however, both new social movements' critique of technological risks and endogenous developments in science studies have contributed to an emerging view of ignorance as inherent part of scientific development (Wehling, 2001: p. 470). Meanwhile, social scientific research on ignorance has rapidly grown to an interdisciplinary research field, 'agnotology', with the explicit intention to establish scientific accounts of formerly unrecognized knowledge (Proctor and Schiebinger, 2008). It is constantly fed by an abundance of empirical studies, many of them inspired by a constructionist sociology of knowledge. These have analyzed the emergence, discovery and consequences of scientific ignorance, like in human genetics and biotechnology, in public incidents with substances like chlorofluorocarbons, dichlorodiphenyltrichloroethane, and dioxin, or consequences of new technologies (for instance, Böschen et al., 2010). In addition, one can recognize an increasing interest in the societal dimensions of ignorance as inevitable part of modern knowledge societies (Luhmann, 1992; Ungar, 2008). Three aspects of these rather recent developments in a sociology of ignorance shall be considered here: the 'knowledge–ignorance paradox', the 'right not to know', and different 'epistemic cultures of nonknowledge' as dimensions of an increasing politicization of ignorance.

Particularly, Ungar has characterized a so-called knowledge–ignorance paradox evolving from contradictory dynamics of knowledge societies and its consequences as well. He indicated that, as inevitable by-product of the specialization pressures modern knowledge workers are subject to, their production of more specialized knowledge results in more general ignorance as well, entailing increasing entry costs and speech barriers to other knowledge domains (Ungar, 2008). The "shifting balance between personal expertise and common knowledge" goes hand in hand with an "increased reluctance to pursue additional information and a concomitant decline in the stock of general knowledge" (Ungar, 2008: p. 312). Focusing on functional knowledge deficits as cases where ignorance definitely is no bliss, Ungar also identified 'chains of ignorance', presumably consequential classes of ignorance that would require reversals throughout these chains in order to critically combat an overall debasement of public discourse (Ungar, 2008: p. 314). As part of the marketing of ignorance as a social problem, particularly successful in coalescence with social movements or moral panics, he identified claims referring to embedded ignorance found in specialty domains, ignorance of medical information, and the special kind of ignorance of the young; young people as a group are more often than others targeted by 'do not die for ignorance' actions, like in antitobacco campaigns, with significant moral undertones; as Ungar remarked, similar to the case of medical ignorance, the problem of ignorance here is often secondary to the real social problem. Ungar's interest in the societal dimension of ignorance within knowledge societies could fruitfully be deepened by future research taking the social stratification of access to knowledge, and nonknowledge, into account.

However, one can also imagine potential situations in which ignorance indeed may be bliss. Both predictive genetic testing and genetic engineering of children indicate areas where it might not only be 'better not to know' (Ungar, 2008: p. 314), but where there also exists the explicit 'right not to know' (Chadwick et al., 1997). Ungar described that although efforts to purposively shelter ignorance for moral reasons may be largely doomed to failure, revelations of uncertainties potentially undermining the authority of science become more commonplace as well. Parallel to scientific progress in human medicine, genetics, and biotechnology, public discourses on 'the right not to know' have emerged because of its preventive function (Popitz, 1968) from taking action, for instance, in the birth of a child with possible disorder. Seen from the formal structure of the argument, a 'right not to know' might entail some similarities to that 'liberal' view on ignorance encountered above, insofar it builds on the opportunity of nondiscriminatory effects of social action mechanisms implying nonknowledge. From a sociological perspective, the challenge of the debate lies in the array of implications of that fragile judicial construct for the role of actors, professionals, and the media in societies usually taking a positive stance toward knowledge (Levitt, 1997). Actors, expected to take responsibility for their health and those of their children, in reality do have limited opportunities and resources to exercise choice and be in control of their lives; thus, it is unsurprising that the topic arises "as the technology to predict risks for individuals runs ahead of both available treatments and the human ability to cope with such knowledge" (Levitt, 1997: p. 35).

Based on a broad study on the role of ignorance in various scientific disciplines, and discussing two empirical cases in greater detail, Stefan Böschen et al. (2010: pp. 790f.) identified three different 'epistemic cultures of nonknowledge' as equally valid, rational ways of dealing with nonknowledge: oriented toward control, toward complexity, or toward case-based experience. They characterized the control-oriented approach toward nonknowledge by an epistemic focus on controlling experimental conditions and avoiding disruptive factors, like in molecular biology. In contrast, in ecology they identified a complexity-oriented culture that obtained a high degree of openness toward unanticipated events and uncontrollable, context-sensitive settings. The remaining culture of nonknowledge concentrating on single case-based experience has been developed in professional expertise such as in medical case histories, characterized by a low degree of decontextualization and a tendency to define nonknowledge as problem of an individual expert. The authors suggested that a full recognition of diverse cultures of nonknowledge is required as prerequisite for socially legitimate, robust decision- making under conditions of politicized scientific nonknowledge (Böschen et al., 2010).

While the increasing politicization of scientific nonknowledge is an interesting topic for the sociology of science in its own right, the broader recognition of ignorance both as part of structural conditions of social action and of human knowledge processes enhances social sciences' efforts to cope with what is necessarily implied in advancing the growth of knowledge.

See also: Afrocentric Approaches to Social Work; Functionalism, History of; Game Theory; Knowledge Society, History of; Science and Technology Studies, History of; Unconscious: History of the Concept; Unintended Consequences: History of the Concept.

Bibliography

Abbott, A., 2010. Varieties of ignorance. The American Sociologist 41 (2), 174–189.

Berger, P.L., Luckmann, T., 1967. The Social Construction of Reality. Doubleday, New York.

Blau, P.M., 1955. The Dynamics of Bureaucracy: A Study of Interpersonal Relationships in Two Government Agencies. University of Chicago Press, Chicago.

Böschen, S., Kastenhofer, K., Rust, I., Soentgen, J., Wehling, P., 2010. Scientific nonknowledge and its political dynamics: the cases of agri-biotechnology and mobile phoning. Science, Technology, & Human Values 35 (6), 783–811.

Chadwick, R., Levitt, M., Shickle, D. (Eds.), 1997. The Right to Know and the Right Not to Know. Avebury, Aldershot.

Clark, B.R., 1960. The cooling out function in higher education. American Journal of Sociology 65 (6), 569–576.

Cohen, M.D., March, J.G., Olsen, J.P., 1972. A garbage can model of organizational choice. Administrative Science Quarterly 17 (1), 1–25.

Crozier, M., Friedberg, E., 1977. L'Acteur et le Système. Edition du Seuil, Paris.

Douglas, M., 1966. Purity and Danger. An Analysis of Concepts of Pollution and Taboo. Routledge, London.

Freud, S., 1901. Psychopathologie des Alltagslebens: Über Vergessen, Versprechen, Vergreifen, Aberglaube und Irrtum. Verlag von S. Karger, Berlin (1914. Engl. Psychopathology of Everyday Life. Fisher Unwin, London.).

Garfinkel, H., 1967. Studies in Ethnomethodology. Prentice-Hall, Englewood-Cliffs, NJ.

Goffman, E., 1959. The Presentation of Self in Everyday Life. Doubleday, Garden-City, NY.

Gross, M., 2007. The unknown in process: dynamic connections of ignorance, nonknowledge and related concepts. Current Sociology 55 (5), 724–759.

Huber, J., Shapin, S., 2002. Deliberate Discretion? The Foundations of Institutional Autonomy. Cambridge University Press, Cambridge.

Knorr-Cetina, K., 1981. The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science. Pergamon Press, Oxford.

Levitt, M., 1997. Sociological perspectives in the right to know and the right not to know. In: Chadwick, R., Levitt, M., Shickle, D. (Eds.), The Right to Know and the Right Not to Know. Avebury, Aldershot, pp. 29–36.

Luhmann, N., 1992. Ökologie des Nichtwissens. In: Beobachtungen der Moderne. Westdeutscher Verlag, Opladen, pp. 149–220 (1998. Engl. Ecology of ignorance. In: Observations on Modernity. Stanford University Press, Stanford, pp. 75–114.).

March, J.G., Simon, H.A., 1958. Organizations. Wiley, New York.

McGoey, L., 2012. The logic of strategic ignorance. The British Journal of Sociology 63 (3), 553–576.

Merton, R.K., 1936. The unanticipated consequences of purposive social action. American Sociological Review 1, 894–904.

Merton, R.K., 1957. The bearing of empirical research on sociological theory. In: Merton, R.K. (Ed.), Social Theory and Social Structure. The Free Press, Glencoe, IL, pp. 102–120.

Merton, R.K., 1987. Three fragments from a sociologists' notebook: establishing the phenomenon, specified ignorance, and strategic research materials. Annual Review of Sociology 13, 1–28.

Merton, R.K., Barber, E., 2004. The Travels and Adventures of Serendipity: A Study in Sociological Semantics and the Sociology of Science. Princeton University Press, Princeton and Oxford.

Moore, W., Tumin, M., 1949. Some social functions of ignorance. American Sociological Review 14 (6), 787–796.

Polanyi, M., 1966. The Tacit Dimension. Doubleday, New York.

Popitz, H., 1968. Über die Präventivwirkung des Nichtwissens: Dunkelziffer, Norm und Strafe. Mohr, Tübingen.

Proctor, R.N., Schiebinger, L. (Eds.), 2008. Agnotology: The Making & Unmaking of Ignorance. Stanford University Press, Stanford.

Ravetz, J.R., 1987. Uncertainty, ignorance and policy. In: Brooks, H., Cooper, C.K. (Eds.), Science for Public Policy. Pergamon Press, Oxford, pp. 77–89.

Rawls, J., 1971. A Theory of Justice. Belknap Press of Harvard University Press, Cambridge, MA.

Rayner, S., 2012. Uncomfortable knowledge: the social construction of ignorance in science and environmental policy discourses. Economy and Society 41 (1), 107–125.

Schelling, T., 1960. The Strategy of Conflict. Harvard University Press, Cambridge, MA.

Schneider, L., 1962. The role of the category of ignorance in sociological theory: an explanatory statement. American Sociological Review 27 (4), 492–508.

Schütz, A., Luckmann, T., 1973. Structures of the Life-World. Northwestern University Press, Evanston.

Simmel, G., 1908. Das Geheimnis und die geheime Gesellschaft. In: Simmel, G. (Ed.), Soziologie. Untersuchungen über die Formen der Vergesellschaftung. Duncker & Humblot, Leipzig, pp. 337–402 (1906. Engl. The sociology of secrecy and secret societies. American Journal of Sociology 11, 441–498.).

Smithson, M., 1989. Ignorance and Uncertainty: Emerging Paradigms. Springer, New York.

Ungar, S., 2008. Ignorance as an under-identified social problem. The British Journal of Sociology 59 (2), 301–326.

Unger, P., 1975. Ignorance: A Case for Scepticism. Clarendon Press, Oxford.

Weber, M., 1917/1919. Wissenschaft Als Beruf. Politik Als Beruf. In: Weber-Gesamtausgabe, M. (Ed.), Bd. 17. Mohr, Tübingen (Engl. Science as a Vocation).

Wehling, P., 2001. Jenseits des Wissens? Wissenschaftliches Nichtwissen aus soziologischer Perspektive. Beyond Knowledge? Scientific Ignorance from a Sociological Point of View. Zeitschrift für Soziologie 30 (6), 465–484.

Weinstein, D., Weinstein, M.A., 1978. The sociology of nonknowledge: a paradigm. In: Jones, R.A. (Ed.), Research in the Sociology of Knowledge, Sciences and Art, vol. 1. Jai Press, New York, pp. 151–166.

Whitley, R., 1984. The Intellectual and Social Organization of the Sciences. Oxford University Press, Oxford.

Illich, Ivan (1926–2002)

Chen Hon-Fai, Lingnan University, Hong Kong

© 2015 Elsevier Ltd. All rights reserved.

Abstract

As a Catholic priest cum public intellectual, Ivan Illich is best known for his critique of modern industrial society and its institutions, in particular his controversial deschooling proposal set forth in the 1970s. What is often neglected is the significant shift in his subsequent intellectual development, in which Illich turned to cultural history for questioning the modern certainties concerning our practices and needs. On that basis, Illich suggested that the origin of modern culture and its malaises can be traced back to Christianity and its perversion.

Biography

Ivan Illich (1926–2002) is arguably one of the most prominent social critics and cultural historians of the twentieth century. His works are at once influential and controversial, covering a wide range of topics, such as education, medicine, development, gender, literacy, and Christianity. Illich was born in 1926 to a family with maternal Jewish origin in Vienna. In the 1940s, he studied theology and philosophy at the Gregorian University of Rome, and received his doctorate in history from the University of Salzburg. Soon after graduation, Illich was ordained to the Catholic priesthood in Rome. In 1951, Illich moved to the United States, where he became first an assistant pastor in the Puerto Rican migrant community in New York, and, later, the Vice Rector of the Catholic University of Puerto Rico. With the purpose of training young priests for foreign languages and missionary works, Illich cofounded the Center for Intercultural Formation in Cuernavaca, Mexico in 1961, later to be renamed as the *Centro Intercultural de Documentacion* (Center for Intercultural Documentation (CIDOC)) in 1966.

In a series of articles written in this formative period, Illich openly criticized the Roman Catholic Church for its involvement in President Kennedy's development program for Latin America, the so-called 'Alliance for Progress.' In utilizing the CIDOC as an institutional and intellectual platform to reflect upon the Church's role and deflect from Pope XXIII's call to send 10% of the United States clergy to Latin America, Illich was eventually put on interrogation by the Sacred Congregation for the Doctrine of the Faith in Rome. Illich's break with the Vatican was sealed with his renunciation of official priestly services (though not priesthood) in return for its recognition of CIDOC's independence. After the incident, in the 1970s Illich published four spectacularly successful and popular works earning him international fame, namely *Deschooling Society* (1971), *Tools for Conviviality* (1973), *Energy and Equity* (1974), and *Medical Nemesis* (1976).

With the closure of the CIDOC in 1976, Illich's intellectual life entered into a new stage. In the 1980s Illich had largely retreated from the public scene, which signified a general reorientation of his thought from reform to 'renunciation.' Illich's initial plan was to learn oriental languages in India and East Asia, an exposure that he supposed could shed a new light on the cultural assumptions of Western modernity. When such a plan proved to be impractical, Illich returned to Europe and served as a guest lecturer in various universities in Germany and later the United States. With his erudite learning of the Middle Ages, from the 1980s onward, Illich had been publishing studies in Western cultural history. While his later works enjoyed less popularity than those of the 1970s, Illich never failed to gather an intimate circle of friends and students devoted to the same line of intellectual and public inquiry.

Illich died at his friends' home in Bremen, Germany in 2002. True to the spirit of his devastating critiques of schooling and medicalization, Illich never took up any position of full-time professorship in his lifetime. Neither did he seek any extended medical treatment to the cancer like lump growing on his face and staying with him for almost 20 years. In his peculiar fashion, Illich is a priest living up to his faith, though his was a religion of freedom and friendship rather than dogmatism and discipline.

Major Works

In *Celebration of Awareness* (1971), a collection of his early essays in the 1960s, Illich calls for an 'institutional revolution' in which the social arrangements and myths pertaining to modernity are subject to radical questioning and thereby fundamental restructuring. At the outset, Illich is highly critical of United States' program to steer Third World development in its image. According to Illich, the 'war on poverty' in Latin America is no less violent than the war in Vietnam in its zealous, quasimissionary attempt to impose American values and its middle-class ethos upon the world. He criticizes the hypocrisy and condescending ignorance on the part of volunteers, developers, and above all the Roman Catholic Church, whose 'charity' and good intention inadvertently play into the hands of political ideology and, worse still, cause irrevocable damages to the traditional way of life on which the dignity of the poor precariously rests (Illich, 1971: 25). 'Underdevelopment' is thus resulted, which for Illich refers to a state of mind whereby insatiable mass demand for commodities and services are imposed on the otherwise self-sufficient people. Echoing the maxim that the road to hell is paved with good intention, Illich urges the American volunteers to renounce their self-ordained power and duly recognize their inability to 'help' the poor.

Published after his break with the Vatican, *Deschooling Society* (1971) can be regarded as Illich's most influential and groundbreaking work. Based upon his observations in Puerto Rico and Mexico, Illich contends that universal education, in the form of compulsory schooling, had proved to be infeasible and undesirable. Nothing short of 'deschooling' could overcome the inherent bias and fallacy of the existing system. To begin with, Illich considers his assaults on schooling as part of a general critique of modern institutions, in which the human values of health, learning, and autonomy are all rendered into operative goals to be accomplished with resources mobilized (and eventually monopolized) by organized agencies such as hospital, school, and transportation system. For Illich, the growing prevalence of modern institutions and the industrial logic they embody represents a global and universal degradation, as nonmaterial needs come to be manufactured and manipulated as self-perpetuating and ever-rising demands for professional services. But the 'institutionalization of values' is deemed to be self-defeating, as schooling becomes a hindrance to learning, medicine induces illness, and transportation produces traffic jams (cf Illich, 1974).

Illich detects several drawbacks in the field of mass education. First, the disproportionately high cost and low return of the school system makes it *economically infeasible*. Despite rocketing educational expenses, dropout rate remains resilient among the lower class in Latin America and the United States itself. The poor institutional performance, however, justifies ever more costly investment of money and 'good faith' in universal education, thus setting a vicious circle in motion (Illich, 1971: 10). Besides, universal schooling can heighten *social polarization*. With their coerced or willed dependence on educational institutions, the poor must live up to the middle-class standards of learning, consumption, and lifestyle that are deliberately imposed and promulgated by the minority of local elites, new professional classes, and foreign nations (Illich, 1971: 9). Finally, obligatory schooling incurs a generalized *psychological impotence* among the average consumers and particularly the poor, who are deprived of the traditional means of self-education once available in the form of everyday life learning embedded in local community (Illich, 1971: 3).

The school system also exercises a profound impact on the minds of students. School attendance and graded promotion constitute modern 'rites of passage' through which students are calibrated and assigned to occupational, social roles with differential privileges and status honor (Illich, 1971: 44). Together with standardized curriculums and credentials, these rituals convey the mythic belief that the 'innocent children' desperately need to be educated within the institutional settings of the school. Insofar as education is confused with schooling, and progress confused with progression, educationalists can legitimize the necessity and expansion of the school system, whereas the teachers can justify the total submission of students to their moral and professional authority (Illich, 1971: 26–32).

Illich accordingly suggests that the malaises of mass education can be overcome only with deschooling, which refers not to a wholesale elimination of school, but rather its disestablishment. In much the same way as the constitutional disestablishment of the Church in the First Amendment, Illich argues that we should stop allocating public expenditure and granting social privileges to the school achievers (Illich, 1971: 11; see also Illich's foreword to Hern, 1996). In advocating for the abolition of legally enforced school attendance as a prerequisite of employment and citizenship, what Illich has in mind is the spiritual monopoly once assumed by the Church and now being revived by the secular school, which exhibits the same tendency of reducing the values of salvation or education into administrated, profiteering services.

Illich sets forth the 'learning Web' as an alternative to compulsory schooling. The idea is to provide open access to educational and community resources for students and the public to pursue their own (rather than predesignated) learning goals. Credentials are to be replaced by an 'edu-credit card,' with which every citizen can freely utilize reference services and educational tools, consult specialists of various fields, join centre of skill exchanges, and seek out learning partners (Illich, 1971: 78–79). For Illich, these deschooling measures are not only intended to resolve the immediate crisis of compulsory education. His broader purpose is to resurrect what he calls 'Epimethean man,' a man of hindsight who retains hope and faith in the natural goodness of tradition and community, in contrast to the modern 'Promethean man' who attempts to control and manipulate the environment only at one's own peril (Illich, 1971: 105–107).

Illich's critique of modern institutions is further elaborated in *Medical Nemesis* (1976), in which he takes to task the threats to health posed by the medical profession. To this he coins the term *iatrogenesis*, which refers to diseases and epidemics inflicted by the physicians. It occurs when patients are subject to clinical conditions and medical treatments either unnecessary or detrimental to their recovery and health. Illich first demonstrates that the decline of mortality rate in the postwar era was a result of the general improvement of physical environment and hygienic awareness, rather than the effectiveness of doctors and antibiotics (Illich, 1976: 6). Medicalization rather gives rise to *counterproductivity*, the point at which increasing outputs of medical technologies and institutions bring about unwanted side-effects such as expensive but ineffective treatments.

At a deeper level, Illich alerts us to the social pressures exercised by standardized health care and hospitalized treatment, which define any self-healing, dying-at-home, or suffering outside institutionalized sick role as deviance and 'irresponsibility.' When the medical institution appropriates the means of healing and becomes a *radical monopoly*, the majority can have no choice but become patients in order to receive its packaged services (Illich, 1976: 33–35). Medicine eventually becomes a religion, as it seeks to indoctrinate an unquestioned faith in treatment, therapy, and health protection. The myth of medicalization, just like the myth of schooling, legitimizes the expansion of health care and costly investment in high technology by inducing public obsession with diagnosis and preventive measures (Illich, 1976: 106).

Hence for Illich, the dominance of medical profession reinforces iatrogenesis as it destroys culturally specific ways of coping with suffering and death. While medicine purports to 'kill' or eliminate pain, vernacular culture inculcates an 'art of suffering' with which a person can live through the necessity of illness and death in grace and dignity. Insofar as pain and death are reduced from cultural performances of the body into mere

clinical symptoms, these episodes of the human condition are devoid of personal meaning and virtue, and become utterly intolerable and meaningless to the spoiled consumers (Illich, 1976: 129–130). In line with his renunciation of the Promethean man in *Deschooling Society*, Illich asserts that the overall outcome of iatrogenesis is 'medical nemesis,' that is, divine vengeance on human beings' 'hubris' or presumption to become gods in their own right. Instead of top-down reform by the medical profession, medical nemesis can be dispelled only with a reversal of heteronomous professional monopoly and resumption of self-care among laypersons.

Alongside his critiques of educational and medical institutions, Illich outlines his ideal society in *Tools for Conviviality* (1973). For Illich, 'conviviality' is the exact opposite of industrial productivity. It connotes the autonomous, creative intercourse of persons with each other and with their environment (Illich, 1973: 11). What is going to supersede the industrial age is a 'convivial society' in which all social members can enjoy equal, open access to physical means and institutional arrangements that enable and facilitate their activities. For Illich, the task is to regain democratic control over these 'tools' from technical experts, such that they can be put in the service of genuine human needs, creative joy, and individual freedom. In place of technocratic decisions, we shall need to revive deliberative procedures through which autonomous individuals can engage in the political process of resolving differences and negotiating the proper usage and reasonable scope of social tools.

Starting from the 1980s, Illich's intellectual development had been undergoing a radical reorientation. *Gender* (1983) can be taken as a transitional work in this period, where Illich sets forth the unconventional argument that modern industrial society can be understood in terms of the transition from the age of gender to that of sex. For Illich, 'gender' and 'sex' are crucial, since they represent two basic principles of social organization and modes of social existence. In the first place, 'gender' represents the reign of *duality* in vernacular culture, which consists in an unequal albeit complementary relationship that permeates and distinguishes "places, times, tools, tasks, forms of speech, gestures, and perceptions that are associated with men from those associated with women" (Illich, 1983: 3). Every specific practice, perception, and position is thus 'gendered' or marked in definite, proportional relationships within the Great Chain of Being. Being cosmos and social milieu at once, gender connotes not so much inequality than hierarchy. It constitutes a 'commons' that defines the general, proper limits of concerted actions and shared activities on the part of both men and women, who accordingly have recognizable (albeit differential) claims to their respective shares in the overall subsistence of a kin-based community.

In contrast to vernacular gender, 'sex' is a regime of economic *scarcity* at once equalizing and polarizing. From the late eighteenth century onward, common characteristics are attributed to all human beings *in abstraction from* their dual relatedness in gendered traditional culture (Illich, 1983: 4). The notion of universal humanity is essentially genderless or 'unisex,' as biological and sexual constitutions become exogenous to the definition of every single individual. On the basis of this cultural and ontological assumption, both male and female are supposed to exhibit the same set of abstract human needs, and are accordingly subject to the same life order and work requirements. From this 'genderless individualism' follows the economic theory of scarcity and competition, which legitimizes women's equal opportunity *and* disadvantageous position vis-à-vis the men (Illich, 1983: 9–10).

Insofar as culturally embedded gender is replaced by disembedded economic neuter, the modern institutions of industrial production and universal education are set into motion, both of which operating under the assumption of scarcity and to the effect of social polarization. In contrast to the loosely regulated social milieu of gender, the hidden injuries of women, the poor and other disadvantaged categories are without limits under the modern regime of sex (Illich, 1983: 11–13). Illich thus concludes that no remedies can be provided to sexual and social exploitation as long as one stay within the formally free but substantively repressive order of modern institutions and languages. The common parlance of 'role' and 'need' in contemporary culture is imbued with a conception of society bifurcating in conformity and deviance from predefined rules or standards. On the other hand, the gender metaphors in vernacular tradition outline the general limits or 'necessities' of action, within which social agents (*qua* persons and bodies) can enjoy the liberty to pursue proper goals, cultivate personal virtues, and engage in the concrete, convivial life of their community.

In hindsight, it is perhaps not surprising that *Gender* provoked fierce criticisms from mainstream feminists like Arlie Hochschild (Bruno-Jofre and Zaldivar, 2012: 581). In fact, Illich composed the book with the incommensurability of gender and sex, or the principles of proportionality and equality in mind. Instead of the feminists' (and modernists') preoccupation with individuality and universality, what Illich advocated for was a recovery of the commons and its vernacular virtues. In this light, the controversy had partly to do with Illich's previous role of social critic, which still lingered in the academic public and obscured his repositioning as cultural historian. Instead of promising the redemption of convivial *tools* and *institutions*, Illich's new task was to question those 'modern certainties' permeating our practices, *discourses* and *senses*, and thereby to reclaim the cultural alternatives long buried in our passage to modernity.

The intellectual reorientation in *Gender*, together with other exploratory works such as *H_2O and the Water of Forgetfulness* (1985) and *ABC: The Alphabetization of the Popular Mind* (1988), eventually leads Illich to fundamental reconsiderations of his longstanding critiques of education and medicine. In his foreword to a volume (Hern, 1996) dedicated to reexamining the deschooling thesis, Illich confesses that he was 'barking up the wrong tree.' For Illich, it proves to be futile to call for education reform by disestablishing school, since its critical potential can be readily assimilated by educationalists in the institutionalization of alternative and lifelong learning programs. While Illich pointed his finger at the inadequacy of school system as means, institutions or tools, he never questioned the desirability of education as an end, need, or value. Hence his early critique of schooling still borne the modernist assumption that 'educational needs' are a given of human nature. Instead of the social effects of schooling, the right question to ponder should be the *ideology* of educational needs.

What Illich wants to problematize is the historicity of *homo educandus*, whose imputed needs laid down the basis of the doctrine of scarcity as applied to educational means and values, against which the school system and its professionally managed alternatives are legitimized.

From where did the need to be socialized, educated, and examined arise in the first place? Illich ventures to seek an answer from the history of reading and literacy. In his highly original work, *In the Vineyard of the Text* (1993), Illich argues that the practice of reading had undergone a significant change in the twelfth century. Prior to that period, the monks used to read biblical texts in an uninterrupted fashion, that is, by reading aloud and following through the unbroken string of words from beginning to end. When the page was invented in the High Middle Ages, spaces were inserted between letters and chapters, such that the reader no longer had to wander around the 'vineyard' but could rather distance oneself from the text. The modern practice of reading thus generates a corresponding inner text or space on the part of individual subject, a 'conscience' that is composed of words and deeds and constitutes a life record susceptible to examination and judgment by God, the Church, and the School.

Illich (1996) then sets out to question the modern certainties of 'life,' 'health,' and 'death' as the ideological underpinnings of medical profession. Alongside the rise of alternative education, Illich proposes that we are entering into a new era of relentless medicalization, in which the physicians not only cure diseases but progressively take up the management of life from womb to tomb. But the widening and deepening of medical system would not be possible without a corresponding 'sacralization of life,' a myth that reifies survival and physical health as something we bear the indubitable responsibility to preserve at all costs (Illich, 1996: 18). According to Illich, the modern Western notion of life has its deepest root in Christianity, which originally conceives of it as a gratuitous gift from Jesus, a revelation that life always means *being a person*. With the subsequent death of nature in scientific revolution and the rise of possessive individualism in capitalist economy, however, 'life' (along with 'health') comes to be redefined as a substance to be maintained, manipulated, and managed (Illich, 1996: 19).

According to the later Illich, the only way out of this self-incurred predicament of the human condition is *askesis* or renunciation, which consists in the suspension of beliefs and attachments to the modern certainties and values of life, health, and death, and with them the responsibility and 'right' to health care (Illich, 1996: 26). As an epistemological guideline and ethical maxim, renunciation is to restore our genuine liberty by duly recognizing the *existential* limits of our body and life, within which we can deliberately cultivate the arts of suffering, dying and loving, and in this way live out our personal meanings. From this perspective, conviviality acquires a renewed, quasi-Christian meaning of catering to our neighbors' genuine needs and lives.

The central significance of 'person' and its root in Christianity receives final exposition in Illich's (2005) last conversations with David Cayley. There Illich sums up his lifelong investigations into modern society and its institutions with the old Latin phrase, *corruptio optimi quae est pessima*, 'the corruption of the best is the worst.' By this Illich proposes that modernity can be best understood in relation to the cultural legacies of Christianity, not as its realization or antithesis but rather *perversion*. Here Illich refers to the story of the good Samaritan told by Jesus in answering the question, "who is my neighbor?" According to Illich's reading, the story conveys the core Christian message that the neighbor can be anyone with whom we voluntarily relate. In this light, the Gospel reveals our untrammeled freedom to love and to see the Incarnation of God in the bodily presence of *this* stranger we happen to encounter. To engage in a voluntary, concrete, and personal relationship with others, and thereby with Jesus, is a call rather than an obligation, since for Christians love always stands above law, custom, language, and any indeed rule or code. Accordingly Illich stresses that 'sin' in its original sense, entails the denial of this personal call to love one's neighbor, rather than the violation of any universal law or moral rule (Illich, 2005: 189).

The early Christian community exemplifies a being-togetherness that fulfills the calling of universal albeit corporeal, interpersonal love. Apart from the Eucharist in which Christians eat, drink, and share the mystical body of Jesus, Illich tells us that *conspiratio* or 'respiring together' was once practiced among the apostles, a ritual of mouth-to-mouth kissing in which the bodies were thereby tied together and suffused with the Holy Spirit (Illich, 2005: 85). The profound corporeality and mutuality of the early Christian community, which for Illich constituted a commons par excellence, was eventually lost with the establishment of the Church, which purported to convert love and charity from personal freedom and virtues to institutionalized rules and values. With the building of parish and the writing of conscience inaugurated by the medieval church around the twelfth century, sin was 'criminalized' as it became an offense against moral duty instead of an incapacity to answer personal call (Illich, 2005: 92). Virtue, friendship, and the good were accordingly transfigured into categorical imperative and moral responsibility, a betrayal of Jesus' promise to free us from the law. The moment of perversion in Christianity occurred when the established Church (and later the school and the clinic) assumed the role of inculcating and sanctioning the spiritual needs of its flock. Illich's cultural history of modernity thus suggests that the deepest source of its evils has to be sought in the process of institutionalization that turns Christianity into its exact opposite, the blasphemy of life that is the 'Anti-Christ.'

Impacts

Among his manifold contributions to the critical interrogations on modern industrial society, Illich's impact has been most strongly felt in the field of education. Shortly after the publication of *Deschooling Society*, educationalists of all stripes issued their critical responses and largely polarized judgments on Illich's position. After a brief period of serious attention amid the crisis of the modern school system in the 1970s, Illich's challenges were soon assimilated into the mainstream under the name of 'alternative education' (Botsford, 1993). Into the 1990s, however, Illich's ideas had begun to regain influences among scholars directly involved in the deschooling movement (Hern, 1996), and those working in the closely related

fields of critical pedagogy (Finkel and Arney, 1995) and adult education (Finger and Asún, 2001). More recently, attempts are made to reconsider Illich's educational thought in the context of his overall intellectual development, thus signaling a new stage of Illich scholarship (Zaldivar, 2011; Bruno-Jofre and Zaldivar, 2012).

In much the same fashion as deschooling, Illich's notion of 'medical nemesis' elicited an extended rejoinder (Horrobin, 1978) that consisted in a largely positive review of Illich's general position, together with a careful assessment of some of his empirical claims and specific contentions. A general revival of interest in Illich is in place as his pioneering critique of medicalization is subject to contemporary reappraisal in light of recent medical advances (Bunker, 2001). Illich's wide-ranging explorations in other problem areas are likewise taken up by scholars and practitioners, such as postdevelopment studies (Rahnema, 1997), peace studies (Kaller-Diestrich, 2011), environmental movement (Mancini, 1998), community work (Schwartz, 1997), and design and architecture (Mitcham, 2003).

Symposiums and conferences were held to pursue certain aspects of Illich's thought, including a symposium on law (Levi, 2012) and a conference on contemporary commons movement (Bollier, 2013). No less significant was the publication of an anthology of personal recollections and cultural–historical investigations in the spirit of Illich (Hoinacki and Mitcham, 2002). The founding of the *International Journal of Illich Studies*, in 2009, furnishes a major site of online publication and open access to secondary works on Illich. Finally, while the religious and philosophical dimension of Illich's social thought had received scant attention in the earlier literature, some commentators begin to draw our attention to the theological background in Illich's works (Inchausti, 2005; Bruno-Jofre and Zaldivar, 2012).

In conclusion, Illich's major contribution is to trace the origins of modern evils in institutional counterproductivity and the perversion of Christian ethics. His works have come full circle from denouncing the good intention of zealous missionaries to revealing the corruption of religious visions. At any rate, Illich's critiques cannot be conveniently dismissed as naive antimodernism. At least for the early Illich, industrial tools (and postindustrial one such as the Internet) can be reinvented in politics and contribute to the making of post-traditional, convivial society. Despite certain affinity of his later works with the postmodernists, in particular the exposé of modernity as a discursive, disciplinary regime à la Michel Foucault, as a man of faith Illich would positively recommend interpersonal, unalienated love as an antidote to the corrosive will to power, which should be radically renounced rather than affirmed in any Nietzschean fashion. In this regard, Illich seems to urge us to go beyond the modern polarities of discipline and resistance, by *choosing* to live in the small but beautiful world of friendship and proportionality. While his final commentaries on Christianity and Western culture remain underdeveloped, Illich's profound insights can open promising lines of intellectual inquiry and political intervention on modernity as a distinct civilizational and institutional order.

See also: Christianity in Latin America; Education in Latin America and the Caribbean: Systems and Research; Medicalization: Sociological and Anthropological Perspectives; Social Inequality and Schooling.

Bibliography

ABC: The Alphabetization of the Popular Mind, 1988. North Point Press, San Francisco.
Bollier, David, 2013. The Quiet Realization of Ivan Illich's Ideas in the Contemporary Commons Movement. Retrieved from: http://bollier.org/blog/quiet-realization-ivan-illichs-ideas-contemporary-commons-movement.
Botsford, David, 1993. Ivan Illich and the Deschooling Movement. Libertarian Alliance.
Brave new biocracy: a critique of health care from womb to tomb. In: Chesworth, Jennifer (Ed.), The Ecology of Health. Sage Publications, Thousand Oaks.
Bruno-Jofre, Rosa, Zaldivar, Jon Igelmo, 2012. Ivan Illlich's late critique of deschooling society: "I was largely barking up the wrong tree". Educational Theory 62 (5), 573–592.
Bunker, John, 2001. Ivan illich and the pursuit of health. In: Heller, Tom, et al. (Eds.), Working for Health. Sage, London.
Celebration of Awareness, 1971. Calder & Boyars, London.
Deschooling Society, 1971. Calder & Boyars, London.
Energy and Equity, 1974. Calder & Boyars, London.
Finger, Matthias, Asún, José Manuel, 2001. Adult Education at the Crossroads. Zed Books, London; NIACE, Leicester.
Finkel, Donald L., Arney, William R., 1995. Educating for Freedom. Rutgers University Press, New Brunswick.
Gender, 1983. Marion Boyars, London and New York.
Hern, Matt (Ed.), 1996. Deschooling Our Lives (Foreword by Ivan Illich). New Society Pub, Gabriola Island, BC.
Hoinacki, Lee, Mitcham, Carl, 2002. The Challenges of Ivan Illich. State University of New York Press, Albany.
Horrobin, David, 1978. Medical Hubris. Edinburgh, London and New York. Churchill Livingstone.
Illich, Ivan, 1971. Celebration of Awareness. Calder & Boyars, London.
Illich, Ivan, 1973. Tools for Conviviality. Calder & Boyars, London.
Illich, Ivan, 1974. Energy and Equity. Calder & Boyars, London.
Illich, Ivan, 1976. Medical Nemesis. Marion Boyars, London.
In the Vineyard of the Text, 1993. University of Chicago Press, Chicago.
Inchausti, Robert, 2005. Subversive Orthodoxy. Brazos Press, Grand Rapids, Mich.
Kaller-Diestrich, Matrina, 2011. Peaces, a gift of grace, turned into modern horror: Ivan illich visionary of the twentieth century. In: Dietrich, Wolfgang, et al. (Eds.), The Palgrave International Handbook of Peace Studies: A Cultural Perspective. Palgrave Macmillan, Houndsmill, Basingstoke, Hampshire; New York.
Levi, Jennifer, 2012. Symposium: radical nemesis: re-envisioning ivan Illich's theories on social institutions. Western New England Law Review 34 (2), 341–350.
Mancini, Joe, 1998. Using Tools for the Common Good – Revisiting the Ideas of Ivan Illich. Retrieved from: http://www.theworkingcentre.org/resource-students/2395-using-tools-common-good-revisting-ideas-ivan-illich.
Medical Nemesis, 1976. Marion Boyars, London.
Mitcham, Carl, 2003. In memoriam Ivan Illich: critic of professionalized design. Design Issues 19 (4), 26–30.
Rahnema, Majid (Ed.), 1997. The Post-Development Reader. Zed Books, London; New Jersey.
Schwartz, David, 1997. Who Cares? Rediscovering Community. Westview Press, Boulder.
The International Journal of Illich Studies. Retrieved from: http://ivan-illich.org/.
The Rivers North of the Future (With David Cayley), 2005. House of Anansi Press, Toronto.
Tools for Conviviality, 1973. Calder & Boyars, London.
Zaldivar, Igelmo, L., 2011. Revisiting the critiques of Ivan Illich's deschooling society. International Journal of Cross-disciplinary Subjects in Education. Special Issue 1 (1), 618–628.

Illiteracy, Sociology of

Narayana Jayaram, Tata Institute of Social Sciences, Mumbai, India

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Girod, volume 11, pp. 7182–7185, © 2001, Elsevier Ltd.

Abstract

This article provides a sociological analysis of functional illiteracy, a problem widespread in the less developed countries, but is invisibly present in the developed countries as well. After clarifying the conceptual basket of illiteracy, the article discusses trends and patterns in functional illiteracy, and concludes with a discussion on the social context in which functional illiteracy needs to be understood.

The noun *illiteracy* (or analphabetism) is a derivative of the noun *illiterate*, which refers to a person who is unable to read or write, other derivatives being the noun *illiterateness*, adjective *illiterate*, and adverb *illiterately*. Thus, lexicologically, illiteracy, in all its inflections, is defined as the antonym of *literacy*, the ability to read and write. In reality, however, between absolute illiteracy and total literacy, different levels of abilities may be distinguished, including *quasi-literacy* and *functional illiteracy*. Empirical evidence shows that the majority of adults, both men and women, in highly developed as well as the less developed countries rank somewhere among these in-between levels. The situation appears to depend mainly on the fact that the way of life and work for large numbers of the population, even in the developed countries, does not require more than rudimentary literacy skills.

The nature and implications of being functionally illiterate, that is, lacking the literacy skills necessary for coping with most jobs and daily situations, is not the same either across different countries or across different strata or categories of the population within a given country. It is well known that lower social strata are more vulnerable to the risks of functional illiteracy, although the rest of society is not immune. In some societies, historically, sections of the population, for example, indigent castes in India and isolated communities in China, were prevented from acquiring literacy skills, which were the exclusive privilege of the literati castes and communities. Even in the literati castes and communities, literacy was the privilege of men, a fact that explains the continued gender skewedness of literacy rates unfavorable for women.

Many tribal and village communities in Africa and Asia were predominantly oral and many linguistic communities did not have a script in which to write their language with. Only when these communities became integrated with the mainstream literate communities or started interacting with them did the need for acquiring reading and writing skills arise. For many such communities, acquiring literacy skills meant acquiring a new language, too. In some cases, they adopted the script of an alien language, most often the language of their colonizers or of the societies into which they came to be integrated. Obviously, the sections of the population that had least resistance to literacy or understood its benefits as a social capital had a lead advantage in such transition.

Furthermore, being functionally literate in one language, often the mother tongue, even if it has a script, is not enough. In multilingual societies, not all the languages enjoy the same functional status. One or two languages may be given the status of the national or official language, which thereby become the dominant functional language of the society. So, in which language one is (il)literate determines how effectively he/she can engage himself/herself in the wider economic and political life of the society. Thus, in multilingual societies, (il)literacy has different connotation and implications, as compared to a monolingual society. Often, regional, religious, and cultural factors prevent individuals from acquiring literacy in a language other their own.

Definition of literacy, and thereby what it means to be illiterate, has also been changing. Today, being fully literate means not just the ability to read and write one's own name; it refers to the ability to read for information and interest, write legibly and coherently, and reflect about the written word. It encompasses the complex set of abilities to understand and use all the symbol systems relevant to a particular community. Thus, in the era of information technology, the concept could be expanded to include the multimedia and electronic text, in addition to conventional alphabetic and numeric systems. These abilities vary across social and cultural contexts, according to need and demand. Thus, both the meaning and the implications of illiteracy is more an empirical issue rather being merely a conceptual issue.

Some experts are of the opinion that literacy is juxtaposed to orality (see Olson and Torrance, 1991). Orality characterizes "societies that have relied on oral communication without the use of writing and are against privileging literacy over orality"; it seeks "to identify a given type of language used in oral communication … and a given type of consciousness, which is supposed to be created by orality or expressible in orality" (Havelock, 1991: 11). While orality and literacy are "sharpened and focused against each other" in many societies in the less developed countries, they are "still interwoven" in the developed world. The recognition of the importance of orality challenges the privileging of literacy and theorizing about the advantage of literacy (Pattanayak, 1991: 105).

Experts on literacy are not convinced that literacy affects intellectual and social change in any straightforward fashion. As Olson, Hildyard, and Torrance observe, "it is misleading to think of literacy in terms of consequences. What matters is what people do with literacy, not what literacy does to people" (1985: 14). Nevertheless, throughout the world, policy makers

have grouped illiteracy with poverty, malnutrition, and lack of education and health care. Total literacy is a cherished value and an avowed objective of many a development plan, especially in the less developed countries many of which have been struggling to achieve it.

Countries that have successfully addressed the scourge of mass illiteracy have done so in the campaign mode. Most of such campaigns since the beginning of the twentieth century have been undertaken as part of an ongoing, larger structural or socioeconomic transformation. The campaigns in the former Soviet Union, the People's Republic of China, the Socialist Republic of Vietnam, and Cuba are cases in point. A more recent illustration is that of India's mass literacy campaigns (see Athreya and Chunkath, 1996). What distinguishes these campaigns from earlier educational movements (in Germany, Scotland, and Sweden, for instance) is their short, time-bound period and the sense of urgency that briefed them.

Part of the great success achieved by the twentieth century campaigns for eradicating mass illiteracy, as in former Soviet Union and Peoples' Republic of China, can be partly attributed to the more effective centralization of political power in these countries, as also the revolution in the technologies necessary for spreading literacy. However, besides the constraints of democracy and federalism and the politics of regionalism, as in the case of India, the existence of complex alphabetical systems (with elaborate symbols) of some languages has been a stumbling block. Even otherwise, in contexts where minimal literacy skills are sufficient, functional illiterates generally cope without much difficulty.

Illiteracy Statistics

Thirty-nine of the 198 countries for which data are available have no illiteracy or less than 1% of illiteracy and, in all, 79 countries have literacy rate of more than 95%. Besides most of the Western and Eastern European countries, these include Australia, Canada, Israel, Japan, New Zealand, the United States or, for instance, Argentina, Barbados, Chile, Cuba, Mongolia, North Korea, the Philippines, Qatar, South Korea, Taiwan, Tajikistan, Trinidad and Tobago, and Uruguay. Leaving out these, all other countries lag somewhat or further behind: 62 countries have more than the world average of illiteracy (16%); in 13 countries, more than 50% of the population is illiterate, with Burkina Faso having the highest percentage (78%) of illiteracy.

Over the twentieth century, the absolute number of completely illiterate adults increased in the less developed countries, mostly due to population growth. In terms of percentages, however, adult illiteracy declined everywhere; the world illiteracy halved between 1970 and 2005. Table 1 illustrates this point as regards six countries typical of different degrees of economic development (Girod, 2001: 7182).

The official statistics of illiteracy among adults in many countries are not precise and reliable. They give little or no information about the real literacy capacities of the adults who are not counted as completely illiterate. Not all countries collect literacy statistics regularly and this imposes restrictions on comparison, across countries or within a country over time.

Table 1 Percentage of illiterate adults from 1900 to 2010 in six countries (UNESCO estimates and national statistics)

	1900	*1950*	*1970*	*1990*	*2010*
USA	11	3	1	–	1
Italy	50	14	6	–	1
Mexico	78	43	25	12	7
Brazil	65	51	32	–	10
India	94	81	66	52	26
Mali	–	–	94	–	69

Functional Illiteracy

The less developed countries are aware of the extent of functional illiteracy of their adult population, and they are all struggling to improve the literacy levels. In developed countries, on the other hand, thanks to decades of universal compulsory free education, the problem of functional illiteracy was thought to be a concern of the past. But many of these countries have been rudely shaken by the findings of empirical research that functional illiteracy was frequent in modern societies, although the problem is not as serious as in the less developed countries. The first international survey of literacy skills was conducted in 1994–2005 under the auspices of the Organization of Economic Co-operation and Development (OECD, 1995, 1997). The survey covered 12 countries, 7 in Western Europe, and Australia, Canada, New Zealand, Poland, and the United States. A uniform set of reading, writing, and arithmetic exercises typically required in daily life and work was administered to evaluate the literacy level of adults aged 16 to 65 years. This survey revealed interesting facts about the distribution of the adult population of the developed countries by literacy levels at the turn of the twentieth century (see Table 2) (Girod, 2001: 7182).

The four categories of Table 2, which excludes immigrants, correspond to the following levels of literacy (Girod, 2001: 7183):

1. *Good:* Ability to at least understand newspapers and nontechnical articles written for the general public, write a simple letter, perform elementary arithmetic, read a road map, etc. These skills correspond roughly to the kind of reading, writing, and arithmetic abilities that compulsory schooling tries to instill.
2. *Mediocre*: Between nearly good and nearly low levels.

Table 2 Percentages of adult population aged 16–65 years (born in the country) by literary levels. Three examples: Germany, Sweden, and the United States in 1994

Literacy levels	*Germany*	*Sweden*	*United States*
Good	19	36	23
Mediocre	41	40	34
Low	31	19	27
Quasi-illiteracy	9	5	16
Total (%)	100	100	100

Organization of Economic Co-operation and Development (OECD), 1995. Literacy, Economy and Society. OECD, Paris.

3. *Low*: At most can understand parts of short, very easy articles; complete a basic application for employment; interpret a pay stub, etc.
4. *Quasi-illiteracy*: At most can understand or write isolated phrases, add or subtract small numbers; often can only understand or write isolated words and numbers. Exceptionally, complete illiteracy.

Sweden ranked top in the survey, although its adults found to have 'good' level of literacy did not make up the majority. In all other countries surveyed, the percentage of adults in the 'good' level of literacy was lower. In Sweden, approximately one in four adults was found to be functionally illiterate (categories (3) and (4)). In the remaining 11 countries, the rate of functional illiteracy was much higher. In the absence of comparable data, landmarks of change are extremely difficult to identify (Girod, 2001: 7183).

Coulombe et al. (2004) derive synthetic time-series over the 1960–95 period on the literacy level of labor market entrants from the age structure of the 1994 International Adult Literacy Survey. They then use these data as a measure of investment in education in a panel data analysis of cross-country growth for 14 OECD countries, namely, Belgium, Canada, Denmark, Finland, Germany, Ireland, Italy, Netherlands, New Zealand, Norway, Sweden, Switzerland, the United Kingdom, and the United States. Their results show that, overall, human capital indicators based on literacy scores have a positive and significant effect on the transitory growth path, and on the long-run levels of gross domestic product per capita and labor productivity. Thus, in contrast to most previous findings, human capital accumulation is found to matter for the long-run well-being of developed nations.

Social Contexts and Illiteracy

Sociological analyses of the reasons for high rates of complete or functional adult illiteracy in less developed countries provide valuable insights. First, a significant proportion of the children does not enroll in or attend primary school at all. The proportion of such cases is decreasing due to the increase in the number of schools and governmental efforts at canvassing enrollment in these schools. Also, more families are becoming aware of the possibility of sending their children to school. However, most families lack the cultural capital to appreciate the importance of education per se, and this is reflected in the poor motivation among the children. Not surprisingly, the children enrolled in the schools drop out; it is estimated that approximately one-fifth of the children on average drop out of the first year of primary school and still many before the end of the fourth year (Girod, 2001: 7183).

Sociologists attribute the high dropout rates in less developed countries to the lack of fit between schooling and the skills necessary for the milieu in which these children live and grow. Their parents tend to think that their children are wasting time at school. In many instances, they would prefer to have the children around for helping with their own job or working independently to supplement the family income.

With jobless growth being a characteristic of the less developed countries during the last few decades, there has been what one could call *education inflation*. With ever more people available with educational credentials, employers tend to escalate the educational requirements for jobs that people with primary school education did earlier. Under these circumstances, primary school education has lost its significance in the job market. For families living on the margins of poverty, the opportunity cost of sending their children to even primary schools would be high. Primary school carries no immediate upward socioeconomic mobility significance to such families (Tanguiane, 1990).

As noted earlier, many adults in less developed countries belong to social strata where illiteracy is the rule and almost all communication is oral. In such cases, the skills acquired during the few years of schooling or through adult education or mass literacy campaigns are never or rarely used. In most cases, they deteriorate or disappear altogether. Thus, many a new literate in less developed countries lapse back into complete illiteracy or close to it (Girod, 2001: 7183).

The problem of functional illiteracy, however, is not unique to the less developed countries. Although not as striking, the presence of a large number of functionally illiterate people in the developed countries, including in Sweden, the United States, and Germany, revealed by the international surveys (see **Table 2**) points to the invisibility of functional literacy in most modern nations. With the possibility of a section of the adult population, even in developed countries, slipping into functional illiteracy, there hardly appears any scope for complacency.

What is serious about the quasi-illiterate and low-literate adults is that they are generally not themselves even aware of having reading, writing, and computing problems. Very often, they do not admit to having poor literacy skills. If the limited literacy abilities of a significant portion of the adult population prevent them from coping normally at work, at home, in a market-place, on the road, etc., then they would constantly make mistakes and have difficulties in the world around them. However, this is not the case, for reasons identified by Girod (2001: 7184).

1. Most quasi- and low-literate adults are capable of performing very simple or at least rudimentary literacy tasks like reading and writing, recognizing numbers, and making simple arithmetic calculations.
2. Not all or even most members of the society need possess sophisticated reading, writing, and arithmetic skills. Routine housework, many jobs, and leisure-time activities can be performed with the levels of literacy skills possessed by quasi-illiterates or low-literates. More and more push-button or touch-screen technological devices are designed keeping in mind lower levels of literacy skills.
3. Reading for pleasure is no more a habit among most adults; writing even less so. With television becoming a popular medium of information and entertainment, a large percentage of adults have ceased or almost so to read books. Electronic calculators and computing devices have relegated paper and pencil to the margins of household stationery.
4. If reading, writing, or computing is useful or necessary, it is always easy to find help at home or at work.

5. The prestige that literacy once bestowed when the society was predominantly illiterate has waned with the spread of literacy.

Girod's analysis reveals that notwithstanding the difference between the less and most developed countries, the basic mechanisms at work as regards literacy and illiteracy are the same. They influence individual motivations and achievements both at a young age and later in life. "These mechanisms, on the whole, form rather clear tendencies towards an approximate equilibrium between, on the one hand, the common subjective and objective needs of the individual's social and occupational milieu and, on the other hand, their reading, writing and computing capacities" (Girod, 2001: 7184).

To reiterate Girod's conclusion, "As long as they cope with the limited demands inherent in their roles at work and in society, with respect to reading, writing, and computing, functional illiterates are generally not inclined to improve their literacy skills" (2001: 7184–7185). Here lies the formidable obstacle to the success of adult education programs and mass literacy campaigns.

See also: Chinese Literacy Acquisition: A Multidimensional Puzzle; Cognitive Styles and Learning Styles; Educational Institutions, History of; Educational Learning Theory; Literacy Education; Literacy and Illiteracy, History of.

Bibliography

Arnove, R.F., Graff, H.J. (Eds.), 1987. National Literacy Campaigns: Historical and Comparative Perspectives. Plenum Press, New York.

Athreya, V.B., Chunkath, S.R., 1996. Literacy and Empowerment. Sage Publications, New Delhi.

Coulombe, S., Tremblay, J., Marchand, S., 2004. International Adult Literacy Survey: Literacy Scores, Human Capital and Growth across Fourteen OECD Countries. Statistics Canada, Ottawa. http://www.nald.ca/library/research/oecd/oecd.pdf (accessed 13.07.12.).

Girod, R., 2001. Illiteracy, sociology of. In: International Encyclopedia of the Social and Behavioral Sciences. Elsevier Science Ltd., Oxford, pp. 7182–7185.

Goody, J. (Ed.), 1968. Literacy in Traditional Societies. Cambridge University Press, Cambridge.

Havelock, E., 1991. The oral–literate equation: a formula for the modern period. In: Olson, D.R., Torrance, N. (Eds.), Literacy and Orality. Cambridge University Press, Cambridge, pp. 11–27.

Olson, D.R., Hildyard, A., Torrance, N., 1985. Literacy, Language, and Learning: The Nature and Consequences of Reading and Writing. Cambridge University Press, Cambridge.

Olson, D.R., Torrance, N. (Eds.), 1991. Literacy and Orality. Cambridge University Press, Cambridge.

Organization of Economic Co-operation and Development (OECD), 1995. Literacy, Economy and Society. OECD, Paris.

Organization of Economic Co-operation and Development (OECD), 1997. Literacy Skills for the Knowledge Society. OECD, Paris.

Organization of Economic Co-operation and Development (OECD), 2000. Literacy in the Information Age: Final Report of the International Adult Literacy Survey. OECD, Paris.

Pattanayak, D.P., 1991. Literacy: an instrument of oppression. In: Olson, D.R., Torrance, N. (Eds.), Literacy and Orality. Cambridge University Press, Cambridge, pp. 105–108.

Tanguiane, S., 1990. Literacy and Illiteracy in the World: Situation, Trends and Prospects. International Bureau of Education and UNESCO, Paris.

UNESCO. National adult literacy rates (15+), youth literacy rates (15–24) and elderly literacy rates (65+). http://stats.uis.unesco.org/unesco/TableViewer/tableView.aspx (accessed 13.07.12.).

Illness and Dyadic Coping

Guy Bodenmann, University of Zurich, Zürich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

While coping was viewed as an individual process for a long time, dyadic coping is an attempt to conceptualize stress and coping as an interpersonal phenomenon. This means that a stressor always affects both partners and thus represents a shared problem. On the other hand, it also means that couples have resources to cope together against these adversities. In the context of illness this perspective invites to consider both partners as part of the trouble ('we-disease') as both, the ill one and the other, are suffering from the situation. Both are, however, also capable to mutually engage in joint efforts to deal with the illness (joint coping) and thus make us of their synergies. More recent studies illustrate that dyadic coping plays an important role in understanding relationship functioning and well-being of patients in close relationships.

Within the last two decades, the individual-oriented perspective of viewing illness as a personal issue has been enlarged by considering the social context. Thus physical and psychological disorders have no longer been studied by excluding variables such as social support, family resources, and relationship functioning. The bio-psycho-social model of illness became more recognized and accepted and scholars as well as practitioners are more aware that physical as well as psychological disorders are embedded in social context (Revenson and Lepore, 2012). An increasing amount of studies supports the strong association between relationship functioning and psychological (Whisman and Uebelacker, 2009) as well as somatic well-being (Robles and Kiecolt-Glaser, 2003). The interplay between close relationship functioning and health (Proulx et al., 2007) is well established (*see* Japan: Sociocultural Aspects). Among relationship variables relevant for health issue, (1) critical, hostile or over-protective attitudes of the partner (see concept of expressed emotion, Hooley, 2007), and (2) negative or ambivalent interaction behavior (Saxbe and Repetti, 2010) or support processes (Uchino, 2006) are discussed. Since almost 20 years, the notion of dyadic coping became an increasingly important variable within the context of health. Several theoretical approaches of dyadic coping or social support in close relationships have been proposed and a great deal of studies addressed this concept in the context of well-being.

Theoretical Concepts of Dyadic Coping

A dyadically enlarged perspective of coping in couples has been suggested by a number of theorists and researchers proposing to define coping in dyads as an interpersonal process involving both partners yielding to different conceptualizations of dyadic coping as (1) an *effort to establish and maintain close relationships* (e.g., Coyne and Smith, 1991; DeLongis and O'Brien, 1990), (2) the *congruence of individual coping efforts within both partners* (Revenson, 1994), and (3) an *interpersonal coping process* by involving both partners in order to reduce problem- and emotion-focused stress (Berg and Upchurch, 2007; Bodenmann, 1995, 2005).

Dyadic Coping as Relationship-Focused Coping

The theorists of the first approach assume that partners do not only have to engage in problem- or emotion-focused coping facing an important chronic or life threatening illness but that they also need to manage the mutual emotional needs of both partners and that they have to deal with the impact of the illness on the close relationship as well. Thus Coyne and Smith (1991) proposed to introduce *relationship-focused coping* as a third coping category, which includes the partners' efforts to reduce stress in the partner and to guarantee his or her physical and psychological well-being. The authors suggested three main strategies: (1) *active engagement* (involving the partner in constructive problem solving, problem-oriented discussions, and helping to explore the feelings of the partner), (2) *protective buffering* (hiding concerns, denying worries, avoiding disagreements), and (3) *overprotectiveness* (dominant, aggressive or submissive strategies to avoid strong emotional involvement).

DeLongis and O'Brien (1990) proposed a somewhat similar approach by focusing on dyadic efforts to maintain the quality of the relationship also in situations of high stressful encounters (e.g., the dementia of one partner). They assume that relationship-focused coping preserves and enhances social relationships (or disrupt dysfunctional relationships) and emphasize the importance of *empathic coping* (affective awareness of the partner's problem and communicating him/her this understanding in an appropriate manner) within this form of coping.

Dyadic Coping as Coping Congruence between the Partners

Theorists referring to dyadic coping as *coping congruence* or *discrepancy* between the partners argue that the compatibility of problem- and emotion-focused coping strategies within both partners may be predictive of marital adjustment. This approach emphasizes the use of similar or complementary coping strategies of the partners facing a common stress event assuming that a high congruence on the applied

individual coping strategies of both partners may be functional, especially with respect to emotion-focused and dissimilar to problem-focused coping (Revenson, 1994).

Dyadic Coping as an Interpersonal Process

In the conceptualization of dyadic coping as an interpersonal process (Bodenmann, 1995, 2005), a systemic transactional model (STM) of stress and coping in close relationships is proposed, defining dyadic coping as an interplay of stress signals of one partner and coping reactions of the other partner to these signals.

Depending upon the stress event at stake, in dyadic coping the stress management resources of both partners are activated in order to maintain or restore a state of homeostasis among both partners as individuals, within the couple as a unit and in regard to other social partners, since it is assumed that one partner's well-being and satisfaction highly depends upon the other's well-being and satisfaction as well as social integration in their social environment. Thus, both partners are motivated (if they are willing to invest in the relationship) to help one another deal with stressful encounters. According to the profile of the situation, the individual and dyadic appraisals, causal attributions and goals, *common*, *supportive*, or *delegated dyadic coping* can be differentiated. These different forms of dyadic coping can be problem-centered or emotion-centered and may be either positive or negative (hostile, ambivalent, superficial) in nature. Whereas in *common or joint dyadic coping* both partners participate in the coping process more or less symmetrically in order to handle a problem-focused or emotion-focused issue relevant to the dyad (by means of joint problem solving, equal division of tasks, joint information seeking, mutual solidarity, joint relaxation exercises, common religious activities), *supportive dyadic coping* is provided by one partner in an attempt to support the other partner who is generally or presently less equipped with the coping resources required to achieve the maintenance or restoration of an adaptive state (by concrete help with daily tasks, supplying relevant information, practical advice, interest to his/her problems, helping the partner to reframe the situation, helping the partner to relax and slow down, manifesting belief in the partner's capabilities, expressing solidarity with the partner). In *delegated dyadic coping*, one partner takes over certain tasks and duties of the other in an effort to reduce the stress experienced by the partner (e.g., one partner does the shopping instead of the other in order to contribute to reducing the partner's stress). According to STM each form of dyadic coping is displayed in accordance to (1) which partner (or both) is primarily concerned by the stress situation, (2) who has coping resources, and (3) what appraisals and goals are at stake. Bodenmann et al. (2011) showed that STM has a significantly higher power than the congruence approach in predicting relationship quality and well-being.

Dyadic Coping and Physical Illness

Although the concept of dyadic coping is relatively recent, an increasing amount of empirical studies support its relevance in couples' and health research. Dyadic coping has been revealed to be a significant and important predictor of relationship functioning and lower stress experience. As relationship functioning is a significant and important predictor of psychological and physical well-being, the assumption is that dyadic coping contributes directly to a better relationship functioning and by this mean to a better well-being (Badr et al., 2010; Bodenmann et al., 2011). The association between dyadic coping and relationship quality is strong and robust (Bodenmann et al., 2011, 2006; Herzberg, 2013; Papp and Witt, 2010).

Findings not only suggest concurrent but also longitudinal associations between dyadic coping and relationship satisfaction (Bodenmann et al., 2006) or divorce (Bodenmann and Cina, 2006). Furthermore, dyadic coping buffers negative effects of stress on aggression in couples (Bodenmann et al., 2010). All these effects assume that dyadic coping exerts a positive effect on well-being by improving relationship satisfaction, which is one key factor in understanding the health of both partners (Robles and Kiecolt-Glaser, 2003). Thus, more indirect effects are expectable. Numerous studies, however, also support positive direct associations between dyadic coping and adjustment in the context of chronic illness (Acitelli and Badr, 2005; Badr et al., 2010; Berg and Upchurch, 2007; Berg et al., 2008; Coyne et al., 2001; Kayser, 2005; Kuijer et al., 2000; Hagedoorn et al., 2008; Manne et al., 2005; Rohrbaugh et al., 2002). Bodenmann et al. (2011) also reported direct associations between dyadic coping and physical well-being.

In a study by Jenewein et al. (2011) in the context of Chronic obstructive pulmonary disease (COPD), low positive and high negative dyadic coping were associated with poorer quality of life and higher psychological distress among patients and their partners. Delegated coping (assistance with daily tasks) was significantly higher among partners. High delegated partner coping (frequent provision of support provided by the partner) and low delegated dyadic coping was associated with poorer quality of life for both patient and partner. COPD patients suffering from depression were supported by their partner considerably more often and attributed deficits in dyadic coping primarily to themselves, while partners with higher scores of depression assessed both their own dyadic coping and that of the patient more negatively.

A study by Whisman and Sbarra (2012) further shows that even wound healing (measured by the biomarker cytokine (IL-6) is positively related to partners' support, especially in women. Uchino (2006) reports various cardiovascular, neuro-endocrine, and immune functional effects of partners' support.

Dyadic Coping and Mental Disorders

Nowadays, only a few studies have investigated dyadic coping in couples with mental disorders. Nevertheless, there is some empirical evidence for the importance of dyadic coping also in this context. Dyadic coping is significantly (negatively) related to psychological disorders such as depression (Bodenmann et al., 2004), compulsive obsessive disorders in couples (Foerster, 2009), or anxiety disorders in couples (Forberg, 2010). Studies show that negative dyadic coping is usually higher in couples where one partner is diagnosed with a psychological disorder.

We-Disease as a New Definition of Disorders

Within a bio-psycho-social view of disorders, the concept of 'we-disease' is particularly noteworthy (Badr et al., 2010; Kayser et al., 2007). 'We-disease' offers a new perspective on support and coping processes in couples dealing with psychological and physical disorders. In the concept of 'we-disease' couples are considered to face the illness as a shared 'we-event' and a 'we-experience' rather than an individual problem of one partner requiring support from the other. This means that the illness of one partner is no longer viewed as his/her illness but is appraised as 'a shared problem' as it concerns both partners as a whole. From this perspective, since the illness affects both partners they may have a vital interest to overcome the illness together since health impacts both. The concept of 'we-disease' also means that the treatment should always include both partners, the one with the illness as well as the other as both partners suffer but also have resources and can contribute to the coping process. Inclusion of the patient and his/her partner should not only happen at the beginning and end of a treatment but regularly during the intervention. By this means not only the remission of the illness is at stake but also the recovery of the whole system and the quality of the close relationship.

See also: Social Integration, Social Networks, and Health; Social Support and Recovery from Disease and Medical Procedures.

Bibliography

Acitelli, L.K., Badr, H.J., 2005. My illness or our illness? Attending to the relationship when one partner is ill. In: Revenson, T.A., Kayser, K., Bodenmann, G. (Eds.), Couples Coping with Stress. American Psychological Association, Washington, DC, pp. 121–136.

Badr, H., Carmack, C., Kashy, D., Cristofanilli, M., Revenson, T., 2010. Dyadic coping in metastatic breast cancer. Health Psychology 29 (2), 169–180.

Berg, C.A., Upchurch, R., 2007. A social contextual model of couples coping with chronic illness across the adult life span. Psychological Bulletin 133, 920–954.

Berg, C.A., Wiebe, D.J., Butner, J., Bloor, L., Bradstreet, C., Upchurch, R., Patton, G., 2008. Collaborative coping and daily mood in couples dealing with prostate cancer. Psychology and Aging 23, 505–516.

Bodenmann, G., 1995. A systemic-transactional conceptualization of stress and coping in couples. Swiss Journal of Psychology 54, 34–49.

Bodenmann, G., 2005. Dyadic coping and its significant for marital functioning. In: Revenson, T., Kayser, K., Bodenmann, G. (Eds.), Couples Coping with Stress: Emerging Perspectives on Dyadic Coping. American Psychological Association, Washington, DC, pp. 33–50.

Bodenmann, G., Widmer, K., Charvoz, L., Bradbury, T.N., 2004. Differences in individual and dyadic coping in depressed, non-depressed and remitted persons. Journal of Psychopathology and Behavioral Assessment 26, 75–85.

Bodenmann, G., Cina, A., 2006. Stress and coping among stable-satisfied, stable distressed and separated/divorced Swiss couples: a 5-year prospective longitudinal study. Journal of Divorce and Remarriage 44, 71–89.

Bodenmann, G., Pihet, S., Kayser, K., 2006. The relationship between dyadic coping, marital quality and well-being: a two year longitudinal study. Journal of Family Psychology 20, 485–493.

Bodenmann, G., Meuwly, N., Bradbury, T.N., Gmelch, S., Ledermann, T., 2010. Stress, anger, and verbal aggression in intimate relationships: moderating effects of individual and dyadic coping. Journal of Social and Personal Relationships 27, 408–424.

Bodenmann, G., Meuwly, N., Kayser, K., 2011. Two conceptualizations of dyadic coping and their potential for predicting relationship quality and individual wellbeing. European Psychologist 16, 255–266.

Coyne, J.C., Smith, D., 1991. Couples coping with a myocardial infarction: a contextual perspective of wives' distress. Journal of Personality and Social Psychology 61 (3), 404–412.

Coyne, J.C., Rohrbaugh, M.J., Shoham, V., Sonnega, J.S., Nickla, J.M., Cranford, J.A., 2001. Prognostic importance of marital quality for survival of congestive heart failure. The American Journal of Cardiology 88, 526–529.

DeLongis, A., O'Brien, T., 1990. An interpersonal frame-work for stress and coping: an application to the families of Alzheimer's patients. In: Stephens, M., Crowther, J., Hobfoll, S., Tennenbaum, D. (Eds.), Stress and Coping in Later-life Families. Hemisphere Publishing Corporation, New York, pp. 221–239.

Foerster, V., 2009. Dyadisches Coping und Zwangsstörungen [Dyadic coping and compulsive disorders]. Dr. Müller Verlag, Hamburg.

Forberg, A., Die ängstliche Persönlichkeit in der Partnerschaft. Eine Untersuchung über den Einfluss von Ängstlichkeit auf das dyadische Coping bei Paaren. [The anxious peronality in couples: The impact of anxiety on dyadic coping]. University of Jena, Germany, Jena, unpublished master thesis.

Hagedoorn, M., Sanderman, R., Bolks, H.N., Tuinstra, J., Coyne, J.C., 2008. Distress in couples coping with cancer: a meta-analysis and critical review of role and gender effects. Psychological Bulletin 134, 1–30.

Herzberg, P.Y., 2013. Coping in relationships: the interplay between individual and dyadic coping. Anxiety, Stress, & Coping: An International Journal 26, 136–153.

Hooley, J.M., 2007. Expressed emotion and relapse of psychopathology. Annual Review of Clinical Psychology 3, 349–372.

Jenewein, J., Meier, C., Mörgeli, H., Bodenmann, G., 2011. Dyadic coping, quality of life, and psychological distress among COPD patients and their partners. Journal of Psychosomatic Research 70, 597.

Kayser, K., 2005. Enhancing dyadic coping during a time of crisis: a theory-based intervention with breast cancer patients and their partners. In: Revenson, T.A., Kayser, K., Bodenmann, G. (Eds.), Couples Coping with Stress: Emerging Perspectives on Dyadic Coping. American Psychological Association, Washington, DC, pp. 175–194.

Kayser, K., Watson, L.E., Andrade, J.T., 2007. Cancer as a "we-disease": examining the process of coping from a relational perspective. Families, Systems and Health 25 (4), 404–418.

Kuijer, R.G., Ybema, J.F., Buunk, B.P., De Jong, G.M., Thijs-Boer, F., Sanderman, R., 2000. Active engagement, protective buffering, and overprotection: three ways of giving support by intimate partners of patients with cancer. Journal of Social and Clinical Psychology 19, 256–275.

Manne, S.L., Ostroff, J., Winkel, G., Fox, K., Grana, G., 2005. Partner unsupportive responses, avoidant coping, and distress among women with early stage breast cancer: patient and partner perspectives. Health Psychology 24, 635–641.

Papp, L.M., Witt, N.L., 2010. Romantic partners' individual coping strategies and dyadic coping: implications for relationship functioning. Journal of Family Psychology 24, 551–559.

Proulx, C.M., Helms, H.M., Buehler, C., 2007. Marital quality and personal well-being: a meta-analysis. Journal of Marriage and Family 69, 576–593.

Robles, T.F., Kiecolt-Glaser, J.K., 2003. The physiology of marriage: pathways to health. Physiology & Behavior 79, 409–416.

Revenson, T.A., 1994. Social support and marital coping with chronic illness. Annals of Behavioral Medicine 16, 122–130.

Revenson, T.A., Lepore, S.J., 2012. Stress and coping processes in social context. In: Baum, A., Revenson, T.A., Singer, J.E. (Eds.), Handbook of Health Psychology, second ed. Psychology Press, New York, pp. 193–217.

Rohrbaugh, M.J., Cranford, J.A., Shoham, V., Nicklas, J.M., Sonnega, J.S., Coyne, J.C., 2002. Couples coping with congestive heart failure: role and gender differences in psychological distress. Journal of Family Psychology 16, 3–13.

Saxbe, D., Repetti, R.L., 2010. For better or worse? Coregulation of couples' cortisol levels and mood states. Journal of Personality and Social Psychology 98 (1), 92–103.

Uchino, B.N., 2006. Social support and health: a review of physiological processes potentially underlying links to disease outcomes. Journal of Behavioral Medicine 29, 377–387.

Whisman, M.A., Uebelacker, L.A., 2009. Prospective associations between marital discord and depressive symptoms in middle-aged and older adults. Psychology and Aging 24, 184–189.

Whisman, M.A., Sbarra, D.A., 2012. Marital adjustment and interleukin-6 (IL-6). Journal of Family Psychology 26, 290–295.

Illness Behavior and Care-Seeking

Howard Leventhal, Jessica Yu, and Elaine A Leventhal, Rutgers University, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Illness behavior, the belief that one is threatened by illness and in need of protective action, is typically initiated by changes in somatic experience and physical function that are interpreted as symptoms of an underlying threat to health. This interpretation involves matching of symptoms to schemas of illness. Schemas are products of personal experience, observations of and communication with others, and media exposure. The conviction and emotional worry that one is sick and the overuse of medical care in the absence of objective signs of disease may be sustained by the interplay between somatic sensation and emotional activation. We propose that a common mechanism underlies illness behavior for acute and chronic conditions and for somatoform disorders, and that the biological processes underlying the condition and the social processes surrounding the patient differentiate the observed differences in illness behaviors.

What Is Illness Behavior?

The phrase 'illness behavior' differentiates disease, a biological phenomenon (e.g., cancer, congestive heart failure, diabetes, etc.), from the behaviors associated with disease (Mechanic and Volkart, 1961; Parsons, 1975). These 'behaviors' range from the private experiences of symptoms, the meaning of the disease, and the emotions accompanying it, to the visible responses of 'sickness' such as fatigue, and the behaviors for cure and control such as the use of home remedies, medical care, being excused from work and household duties, and other behaviors involved in the sick role (Charmaz, 1999). Some investigators have differentiated illness behaviors, instigated and sustained by the presence of physical symptoms, from 'health behaviors,' actions taken due to concerns about the possibility of symptoms or under social pressures, and in the absence of physical symptoms (Kasl and Cobb, 1996). This distinction is questionable when we recognize that action can be taken to control the consequences of serious, diagnosed conditions that are asymptomatic, for example, using medication to control hypertension to avoid stroke or to control blood glucose to avoid diabetes. The distinction does, however, point to an extremely important aspect of human cognition: illnesses are represented at both an abstract, conceptual level, and an experiential or perceptual behavioral level (Leventhal et al., 2011); a distinction can be applied to mental representations of virtually all objects and events (Andrews et al., 2009). 'Illness behavior' seems best defined, therefore, as any action taken to cope with and manage anticipated threats of disease as well as the experience of diagnosed disease with and/or without symptoms. 'Sickness behavior,' a third phrase, refers to a syndrome of less visible somatic (e.g., fever) and overt symptoms (e.g., fatigue) generated by the immune systems response to pathogens (Dantzer, 2006); though not the focus of our presentation, the distinction may be critical for understanding illness behavior in the face of unexplained symptoms. Behaviors that affect health, positively or negatively, but not seen to be health relevant, would not be regarded as illness or health behaviors.

Why Is Illness Behavior Important?

Illness behaviors are important because of their impact on the daily functioning and quality of life of individuals, families, communities, and health care resources. The monetary impact and effects on quality outcomes have been a major focus of public discussion and political controversy and suggestions for resolving these problems vary widely both across and within Western and non-Western nations (Emanuel et al., 2012). Overuse of care in the absence of diagnosable disease is often cited as a cause of the high cost of medical care, including conditions labeled as somatoform disorders, somatic symptom disorder, chemical sensitivity, chronic fatigue syndrome, and most recently, Gulf War syndrome. In each case, patients attribute symptoms to an underlying disease that is currently not attributable to an identified, organic factor and symptoms and resulting use of care are attributed to the interplay of severe life stresses and personal vulnerabilities (Brown, 2004; Young, 2004). The 'worried well,' a phrase often applied to elderly women who are frequent participants in blood pressure screening programs, form another class of resource user, whose worrying drives behavior in the absence of disease and symptoms. The magnitude of the costs generated by these behaviors has been exaggerated, however, as they mainly involve the use of primary care services, and the entire primary sector consumes little more than 20% of the total health care budget.

In contrast to behaviors directly favoring overuse, illness behaviors that delay use of diagnostic and treatment regimens can add to the cost of health care when they result in medical crises and hospitalization. Failures in effective management are common for virtually all chronic conditions including asthma, coronary diseases, diabetes, depression, and cancers. These conditions along with ineffective and aggressive treatment in the last months of life are the major drivers of the high costs of health care. Chronic conditions challenge our everyday beliefs and perceptions of what it means to be sick; they are often asymptomatic, permanent, and require constant treatment and lifestyle changes. Dysfunctional illness behaviors can lead to failure to initiate and sustain rehabilitative regimens post-hospitalization, documented by data for myocardial infarction, strokes, and cancer (Phillips et al., 2001; WHO, 2003).

Nonparticipation may be due to individual and socially supported but mistaken beliefs such as "complete rest is essential to avoid death from heart attacks," or to misinterpretation of the pain from the healing of the surgical wounds as pain from failure to repair the underlying disease, such as the pain attributed to the effect of physical activity on a damaged heart.

Determinants of Illness Behavior and Care-Seeking

Given the number of syndromes used to describe illness behavior such as somatization, hypochondriasis, and body dysmorphic disorder, and the potential overlap with ineffective management for both chronic conditions and everyday illnesses, we will focus on the mechanisms common to all of these conditions that may promote or undermine effective self-management.

Symptoms and Care-Seeking

Everyday common sense tells us that symptoms, changes in somatic sensations and function, are stimuli for seeking care. As survey respondents report experiencing one new symptom each week of the year (Demers et al., 1980) and few individuals seek care 52 times in a year, it is clear that important as symptoms may be for care-seeking care, they are not sufficient prods for doing so. Cameron et al. (1993) interviewed clinic patients immediately prior to seeing their doctor and interviewed a matched control (age; family size; gender) very soon thereafter and reported that new symptoms were reported by 100% of the clinic patients and by 30% of the matched controls; symptoms mattered. In addition, the symptom reports of the clinic visitors differed from those of the matched controls; visitors in the latter group were less likely to have labeled or talked to someone about the symptoms and reported it was less disruptive of daily life. Symptoms prod care-seeking when they are perceived, elaborated upon, and represented as indicators of an underlying illness. The representation may range from that of a minor, self-limited and self-treatable problem to a potentially serious problem needing further evaluation, or a clear sign of a potentially painful and/or life-threatening illness that demands immediate attention from a qualified practitioner (Chrisman and Kleinman, 1983; Leventhal et al., 1980). Care-seeking typically follows an interpretation of somatic and functional deviations from the usual self-suggesting that one is 'ill,' unable to manage the condition on one's own, and therefore, in need of medical attention.

Symptom Interpretation: A Social–Psychological Process

The underlying proposition of our model of the cognitive and emotional processes that elaborate the meaning of illnesses, the common-sense model (CSM), asserts that somatic sensations and functional changes felt as deviations from ones 'normal' self are compared and matched to schemata of underlying conditions. If not in all instances, and certainly in the vast majority, the deviations are detected automatically, i.e., they are not the result of a conscious, 'executive driven' search. The generation of the initial, often provisional hypotheses regarding 'what is wrong' is partly if not completely automatic (Leventhal, et al., 2011). In many though not all instances, the detected deviation and the match becomes a conscious event that integrates the concrete experience and its schemata into a mental model or illness representation. The core, empirically supported proposition underlying CSM, is that illness representations are combinations of experience and abstract concepts, words and labels; for example, a painful lump integrated with a schema and labeled as a cancer, a painful joint integrated with a schema and labeled as arthritis, or an infectious sore integrated with a schema of concept of contact with dirt. That the linkage is likely automatic is supported by data showing that telling study participants that their blood pressure is elevated, false feedback for which they are subsequently debriefed, led to immediate reports of the same symptoms reported by patients with hypertension (Baumann et al., 1989), a sign of tight connectedness or symmetry between labels and symptoms, though symptoms are unrelated to objective readings of blood pressure. Representations have features or attributes in five areas, though they may not be complete at a given instant: (1) *Identity*, the location and pattern of the 'symptoms' combined with a label (initially self-diagnosed); (2) *Cause*, the experienced symptom or function is attributed to an external factor, observed (sharp object; dirt) or nonobserved external (virus, stress, luck), or an internal factor (genes); (3) *Time Lines*, rates of onset, duration, and time to resolve, each of which may be anchored in objective (clock) or perceived time; (4) *Consequences*, ranging from missed work and social contacts, costly treatments and unwelcome hospitalizations, etc. each of which can involve amalgams of concrete experience with concepts; (5) *Controllability*, both objective and subjective evidence of control by treatment. Thus, the variables in each area consist of abstraction, i.e., words and symbols, and concrete experience.

Contextual factors have a major impact on the elaboration of illness representations. Chrisman and Kleinman (1983) posit three such sets: (1) the popular sector, which comprises the individual's immediate social contacts, family, and community, who interpret the person's somatic changes in light of a shared belief system; (2) the folk sector, consisting of nonprofessional specialists, ethnic healers, religious or faith healers; and (3) the professional sector, or culturally defined specialists such as physicians, osteopaths, chiropractors, etc., the latter most salient in Western countries such as Europe and the USA. Initial hypotheses elaborating the representation of a possible illness usually come from the individual experiencing the change, then shared with, and elaborated by individuals in the popular sector (Cameron et al., 1993). The folk and professional sectors typically come into play when the shared, popular evaluation suggests a need for expert help. Interactions of an individual with people in each of the three sectors both provide possible interpretations of symptoms and most importantly, they reinforce the process by which deviations from 'normal self' are evaluated. In typical medical visits, the physician's review of the patients complaint may begin with "What bothers you?" followed by "Where is it?" "How long has it been going on?" "What did you do?" and "What happened when you did that?" The questions are consistent with the process involved in the patient's self-appraisal, a process that may have been conducted automatically, with little conscious awareness, but is validated as appropriate for

conscious and deliberate action. In summary, the initial self-evaluation process seeks to match deviations, i.e., symptoms and function, from the 'normal' self with illness schemata, generating hypotheses and/or representations of illnesses. These initial appraisals are typically discussed with family or friends (public sector) and followed on occasion by visits to a practitioner, either from the folk or professional sector. Each of these exchanges involves the interplay of intraindividual and interpersonal exchanges that are the focus of the "Common-Sense Model of Self Regulation," or CSM (Leventhal, et al., 2011).

The recommendations for management emerging from these intra- and interpersonal exchanges lead to 'action plans,' sometimes simple, at other times complex multistep plans for initiating treatment, ranging from do nothing (tincture of time), to short-term interventions, such as antibiotics or simple prostheses (splints) to promote healing, to alterations of lifestyle (diet and physical activity) and medical treatment that may last a lifetime. Each behavioral step initiates yet another set of implicit or explicit feedback loops for evaluating treatment efficacy. The evaluative process continues the intraindividual and interpersonal processes opened in the initial appraisal (popular; folk; professional).

Factors Affecting When the Interpretive Process Leads to Care-Seeking

Although care-seeking is highly likely in response to painful, disabling symptoms perceived as beyond individual control, they are not the factors most frequently leading to care-seeking. Many if not the majority of care-seeking episodes are initiated by specific factors in one or more of the five domains of the representation. For example, a symptom or functional change may elicit worry and concern and lead to seeking advice and care, if it exceeds the expected *time line* for minor and ignorable events and/or fails to match the schemata for a familiar, manageable illness. Rapid rate of onset, location (chest or arm; head) and pattern (sharp pain, dizziness, difficulty moving or speaking) can match the respective schemata for heart attack and stroke. Schemata for gastrointestinal ulcers, breast cancer, migraines, and head colds are defined by patterns in these domains; though the schemata will vary by culture and by individual, each disorder will have a set of features defining its *identity, time lines, control* (*response to treatment by self, consequences, and cause*) (Bishop, 1991). Symptoms that last longer than expected, are unfamiliar or match a schema for a life-threatening illness are likely to activate worry or fear, and whether the affective response leads to care-seeking or avoidance and delay, depends upon their perceived source and its properties. Care-seeking is likely when a disease is feared because its consequences are perceived as life-threatening or a lifelong reduction in function. The possibility of delay in care-seeking increases when a disease is perceived to be uncontrollable and when the perceived benefits of treatment do not outweigh its concerns (Horne et al., 1999). Consistent with 'commonsense' concerns likely increase for treatments with aversive side effects experienced with HIV treatment (Horne et al., 2004); inconsistent with common sense is the data showing that aversive side effects are associated with treatment benefit (Cameron et al., 2002).

Investigators have examined contextual modifiers of care-seeking that are not directly related to illness representations; life stress is a frequently studied example. It has been proposed and data have supported the hypothesis that life stresses are associated with increased probability of care-seeking (Mechanic, 1979). Moderators such as stress seldom work alone, however, and assuming they have consistent, indirect effects on care-seeking ignores the array of factors affecting the care-seeking process. Gortmaker et al. (1982) found that mothers were less likely to seek medical care for an ill child on the day of a stressful life event. A day later, however, when the event resolved and its barrier to care-seeking was removed, care-seeking frequency increased above that recorded in the absence of such events. How the relationship between life stressors and symptoms are interpreted adds further complexities to the process of care-seeking. Ambiguous symptoms such as fatigue and muscle aches are likely to be interpreted as signs of stress and not merit seeking care if they appear in the context of life stressors that have lasted for a few days. When similar symptoms appear in the context of life stressors that have lasted four or more weeks, the likelihood of care-seeking increases as symptoms in that context are likely to be interpreted as signs of stress-induced illness (Cameron et al., 1995). In this latter case, the rates of care-seeking are approximately the same as that recorded for symptoms that are distinct, nonstress signs of illness. It is worth noting that the duration for a stressor to influence the shift from not seeking to seeking care for ambiguous symptoms is roughly the same duration of exposure to life stresses that Cohen and colleagues (Cohen et al., 1998) reported for stress to increase susceptibility to cold viruses.

Social comparisons provide another set of contextual influences on the processes involved in self-diagnosis and care-seeking. For example, people are less likely to perceive symptoms as serious and in need of medical attention if other individuals (the popular sector) report similar symptoms (Croyle and Jemmott, 1991). The effect is likely when the underlying cause is seen as relatively benign, such as gastrointestinal distress by multiple members of one's party following an outing. This assumes that the distress is time limited and the trajectory for change is improving. Social comparisons also can moderate emotional response during treatment; patients awaiting major surgery are less distressed when waiting in a ward with revering, postsurgical patients in comparison to others awaiting treatment (Kulik and Mahler, 1997). Togetherness with patients in a common state of anxious anticipation, waiting for surgery, appears to be an occasion for emotional contagion; it does not have a calming effect, as some social–psychological models would predict. Waiting for surgery is less threatening when surrounded with postsurgical patients who are talking, alive, and seemingly doing well. The complexity of effects of social context on care-seeking is evident from findings of substantial increases in care-seeking and treatment following observation of symptoms and treatment in others. For example, survey respondents who observed symptoms in family and friends during a threatening flu epidemic were more likely to have been inoculated (Rosenstock et al., 1960). The seemingly contradictory relationships reflect interactions between the social context, the level of disease-associated threat, and the perceived efficacy (benefit

exceeding concerns) of treatment. The cognitive and affective processes underlying people's commonsense perceptions of the identity, cause, controllability, and consequences of disease make good sense psychologically though they may be consistent or inconsistent with biomedical interpretations.

Automatic and Deliberative Processes in Abnormal Illness Behavior

'Somatoform disorders' and 'abnormal illness behaviors' are terms encompassing an array of syndromes defined by the reporting of symptoms and the conviction that one is ill in the absence of medically definable disease. Brown's (2004) detailed analysis of the cognitive processes underlying many of these disorders proposes that symptom experiences are a product of a primary perceptual process, and are experienced as real, objective events. The analysis is consistent with the automatic matching of inputs to schemata in CSM, the basic perceptual process for perceiving deviations from the 'normal' self; they too are perceived as real. The outcome of the match may or may not be biomedically valid. Both Brown and the CSM differentiate these automatic perceptions from the deliberative, secondary, executive systems that attend to and attempt to make sense of and act to regulate symptom experience. Somatoform disorders therefore, involve the following two steps: (1) Patients' perceptual experiences are experienced as real and interpreted and continually updated by the cognitive and affective mechanisms underlying the patient's commonsense framework. And the resulting mental models created by these processes may differ from the interpretations based on the conceptual framework of biomedicine the latter is abetted by a variety of 'objective' measures. (2) The patients' mental models and behaviors are labeled as a somatoform disorder because the behavioral state, the symptoms, emotions, and actions, persist in the absence of identifiable pathology and are seemingly impervious to medical explanation.

A number of factors have been proposed to account for the persistence of symptom experience and physical and cognitive dysfunction in the absence of supportive, biomedical evidence. These factors are believed to focus attention on and magnify the salience of benign somatic sensations thereby converting them into symptoms of illness; the result being repeated care-seeking and doctor shopping (Kellner, 1990; Kirmayer and Robbins, 1991; Brown, 2004). *Trait negative affect* (NA) (a relatively stable feature of personality characterized by self-reports that one's mood or emotional state is usually worried, anxious, or depressed), is a frequently mentioned candidate as individuals scoring high on measures of this trait report a larger number of somatic complaints than do low-scoring individuals. There is little evidence, however, that persons high on trait NA make greater use of health care (Watson and Pennebaker, 1989), and it is likely that the vast majority of individuals high on trait NA are well aware of the difference between specific symptoms of illness and somatic signs of emotional distress; that they do so is validated in several studies (e.g., Mora et al., 2002; Rabin et al., 2001). It is also clear that the percentage of the population with somatoform disorders, fewer perhaps than 4%, is far less than the percentage of individuals with high scores on trait NA (Robins et al., 1984). Thus, while individuals in both categories may report worry about health, worry is not equivalent to or sufficient for interpreting and attributing sensations to illness.

There is reason to believe, however, that personal factors such as trait NA and social factors such as expression of concerns about own and others health by family members and friends, provide a context for the formation of somatoform disorders. These factors may come into play by first focusing attention on somatic sensations and by then activating emotion and worry and linking the two by shifting attention to the activated emotional state (Nakao and Barsky, 2007). It may take but little encouragement to convert the emotional state into sustained worry, the encouragement arising either from within the individual's own history, as is often the case post-treatment for breast cancer, or of observation of illness in another, e.g., observation of cancer and/or cardiac death in a family member or close friend (Nakao and Barsky, 2007).

Thus, momentary states of worry or fear can readily become repetitive and chronic, if activated by the ongoing monitoring of the self-system that can automatically detect somatic and functional deviations from the normal self. The repeated scanning and activation, therefore, can convert momentary concerns into repetitive, chronic conditions; literally a self-sustaining state of ruminative worry. The linkage of emotional activity with somatic sensations may be facilitated by the overlap between their respective properties, e.g., symptoms such as fatigue and restlessness are shared by emotional upset and sickness behaviors. In addition to the major role played by emotional activation or worrying in eliciting and maintaining the search for causes and cures characteristic of illness behavior, emotional activation can also increase muscle tension and amplifying somatic sensations eliciting the affective process. Different moderators may further differentiate these response patterns leading to multiple diagnostic categories (Brown, 2004). While the hypothesized mechanism remains to be tested, it is worth noting that it appears in many ways the inverse of the procedures to deactivate emotional responses during medical procedure, e.g., expecting, normalizing, and being response ready while attending to concrete sensory experiences during noxious medical examination (Johnson, 1975) or during childbirth (Leventhal, 1986), and desensitization therapies (Lang et al., 1983).

Chronic Illness, the Self, and Illness Behavior

If, as we propose, the same cognitive and affective mechanisms underlie illness behaviors associated with acute and chronic conditions and somatoform disorders, it is essential to recognize that each specific acute or chronic disease creates a unique physiological, mechanical, and anatomical context. Different demands are made on the behavioral system with variation in the input from the disease process, from its pattern and location of existent sensations (identity), time lines, causes, means of control, and consequences. Three other factors come into play for chronic diseases. First, they are diseases that last and one must live with for a lifetime. Second, chronic conditions require adjusting not only to the illness itself, but to the interaction of the illness with the cognitive and physical changes of aging. Third, most elderly live with multiple chronic conditions. These conditions impact physical function, restrict

participation in daily activities, can have severe impacts on memory and thought, the various dementias, and these constraints are often associated with increases in NA and declines in positive (Zeiss et al., 1996). Managing many, largely intractable, diseases can contribute to a sense of hopelessness, an integral component of depression, and lower quality of life (Vileikyte et al., 2005). The evidence indicates, therefore, that functional decline associated with both advanced age and chronic illness carries additional, psychological hazards, and these psychological changes reinforce the very functional deficits that initiated them.

Illness Behavior and the Management of Chronic Disease

Many of the chronic conditions prevalent in the later years of life are largely asymptomatic; elevations in blood glucose and blood pressure in hypertension are asymptomatic until visible in systemic damage. The genetic changes and proliferation of cancers and formation of arterial plaques can progress in complete silence until a tumor is detected and/or the individual faces a myocardial infarct or stroke. And the asthmatic patient, who experiences episodes of sometimes violent and potentially life-threatening attacks, may be largely unaware of the silent inflammatory process that underlies the disease. All of these conditions require action based upon action plans independent of and at times inconsistent with the concrete experiences associated with everyday, acute conditions that have stimulated illness behavior for most of life.

Unlike somatoform disorders, where somatic cues conditioned to affective worry both amplify somatic sensations and produce persistent illness behavior, conceptual processes need to be the major drivers of illness behaviors for chronic illnesses. Individual, focus group, and qualitative data suggest how the transition might occur. For example, after narrating how she manages her asthma, an interviewee was asked, "Do you think you can always tell you have asthma?"; somewhat taken aback she replied, "I guess from what I've told you I must think it is always symptomatic." Focus groups of patients with congestive heart failure, whose left ventricles eject 40% or less of the blood at a contraction, may fail to perceive problems breathing lying in bed and chronic tiredness and swollen legs as symptoms of coronary disease. As they explain, feeling "kind of tired … just didn't seem to be anything out of the ordinary" and fail to see symptoms distributed at different times as in need of care or part of their heart conditions: "… Oh, I guess that I could have gone to the doctor after I had that collapse on the hallway floor." Their symptoms do not match the schema for cardiovascular disease: "When you hear about having heart problems, … you're supposed to feel maybe a pain in your left arm, maybe a pain in your chest, or pressure … it would have been more clear to me if I had chest pain then I would have said, okay, I'll call and say I'm having chest pain …" (Horowitz et al., 2004). Data from over 400 patients interviewed soon after they were hospitalized for a myocardial infarction support the qualitative reports (Bunde and Martin, 2006). Delay in care-seeking was brief among patients experiencing chest and arm pain and patients who had a prior history of cardiovascular disease. Delay was lengthier among those reporting fatigue, sleep disturbance, and symptoms perceived as gastrointestinal. Location, pattern, rate of onset, etc., matched a coronary event for quick care seekers and aging or an alternative benign condition for those who took longer, endangering their lives and recovery.

A variety of cognitively based beliefs about the self and treatments have been identified for their impact on illness behaviors. The belief that it is important to 'conserve resources' as one ages (Leventhal and Crouch, 1997) encourages levels of inactivity that can be maladaptive for living with many chronic illnesses. For example, it is associated with increases in pain over a 3-year time span in rheumatoid arthritis patients (van Lankveld et al., 2000), and among coronary patients, it encourages avoidance of the physical activity needed for improved cardiac function. A contrasting cognitive factor, the belief that a strategy of vigorous engagement is necessary for good health, can lead cardiac and cancer patients to adopt behaviors such as jogging and the use of herbal supplements in order to strengthen the disease-fighting capacity of their immune system. Although exercise is effective in reducing risk of mortality (Bijnen et al., 1999) few herbals deliver on their promises for health gains and some generate risk as they may interact adversely with chemotherapies of proven value, and many patients using supplements fail to communicate their use to their physicians. People's beliefs about their bodies' tolerance for medication (e.g., "Even small amounts of medicines can upset my body") are associated with negative views of medicine and lower levels of adherence. This has been found in patients treated for HIV and hypertension, a sample of undergraduates and international travelers who held these beliefs reported more symptoms following vaccination.

Illness Behavior: The Self and Social Context

Each and many other examples not here cited illustrate the operation of the cognitive, affective, and behavioral mechanisms involved in managing acute and chronic conditions and somatoform disorders. Illness behaviors that are adaptive and maladaptive as seen from the perspective of biomedicine can be seen as the product of a system in which behavior is controlled by concrete perceptual experience at one or more points in time, and by abstract conceptual factors at others. The alternation of control reflects the complexity and hierarchical structure of the mental models that are generated by the match of symptoms, diagnoses, and observation of illness in others to illness schemata; these processes are central to human cognition (Andrews et al., 2009).

Although the focus of our presentation on the mechanism critical for the immediate perception and conscious evaluation of symptoms for acute, chronic, and somatoform disorders is biased toward the 'intrapsychic,' it is important to reiterate that the cognitive, affective, and behavioral facets are embedded in individual history and social context. The sense of self, physical, cognitive, and social, represents the integration of the interaction of life experience with biological processes and affect both how self-perception and the formation of schemata for acute, chronic, and somatoform disorders. Illness and illness behavior will feedback to and influence the self-system: how individuals define themselves, and how they act in relation to others, and to the tasks of daily life. The functional and emotional impact of chronic illness and its permanence presents a challenge to the very definition of the self. Thus, chronic

illness represents a challenge both from the bottom-up – its symptoms, signs, dysfunctions, and effects – and from the top-down – the meaning of these signs and symptoms (Leventhal et al., 1999). Prior self-identities and competencies and the social context (the popular (family), folk, and professional sectors) both directly influence and moderate these changes in self. Intellectual, emotional, and social resources can support illness behaviors, which are effective for the management of disease even when disease is not curable and can have negative effects of health outcomes by undermining individual autonomy. Critical and hostile interpersonal exchanges within couples can impact the expression of symptoms and maladaptive illness behaviors (Burns et al., 2013). Illness behaviors effective for disease management can strengthen the individual's sense of competence, create new and strengthen existent identities, and enhance the individual's ability to participate in life despite the drastic impact of illness on the physical self. Illness behavior at its best, i.e., the performance of new actions in managing the self with disease, can be the source of new, stable identities that define the value of life.

See also: Cancer Screening; Chronic Illness, Psychosocial Coping with; Control Beliefs: Health Perspectives; Doctor–Patient Interaction in the West: Psychosocial Aspects; Explanatory Style and Health; Gender and Physical Health; Health Behavior, Psychosocial Theories of; Health Risk Appraisal and Optimistic Bias; Health Risk Perception; Health and Illness: Mental Representations in Different Cultures; Interoception; Psychoneuroimmunology; Social Integration, Social Networks, and Health; Stress, Coping and Health; Symptom Perception, Awareness and Interpretation.

Bibliography

Andrews, M., Vigliocco, G., Vinson, D., 2009. Integrating experiential and distributional data to learn semantic representations. Psychological Review 116, 463–498.

Baumann, L., Cameron, L.D., Zimmerman, R., Leventhal, H., 1989. Illness representations and matching labels with symptoms. Health Psychology 8 (4), 449–469.

Benyamini, Y., Idler, E.L., Leventhal, H., Leventhal, E.A., 2000. Positive healthiness: an important influence on self-assessment of health: expanding our view beyond illness and disability. The Journal of Gerontology 55B (2), 107–116.

Bijnen, F.C.H., Feskens, E.J.M., Caspersen, C.J., Nagelkerke, N., Mosterd, W.L., Kromhout, D., 1999. Baseline and previous physical activity in relation to mortality in elderly men: the Zutphen elderly study. American Journal of Epidemiology 150, 1289–1299.

Bishop, G.D., 1991. Understanding the understanding of illness: lay disease representation. In: Skelton, J.A., Croyle, R.T. (Eds.), Mental Representation in Health and Illness. Springer-Verlag, New York, pp. 32–59.

Brown, R.J., 2004. Psychological mechanisms of medically unexplained symptoms: an integrative conceptual model. Psychological Bulletin 130, 793–812.

Bunde, J., Martin, R., 2006. Depression and prehospital delay in the context of myocardial infarction. Psychosomatic Medicine 68, 51–57.

Burns, J.W., Peterson, K.M., Smith, D.A., Keefe, F.J., Porter, L.S., Schuster, E., Kinner, E., August 6, 2013. Temporal associations between spouse criticism/hostility and pain among patients with chronic pain: a within-couple daily diary study. Pain. http://dx.doi.org/10.1016/j.pain.2013.07.053 pii:S0304–3959(13)00434-X [Epub ahead of print].

Cameron, L.D., Leventhal, E.A., Leventhal, H., 1993. Symptom representations and affect as determinants of care seeking in a community-dwelling adult sample population. Health Psychology 12, 171–179.

Cameron, L.D., Leventhal, E.A., Leventhal, H., 1995. Seeking medical care in response to symptoms and life stress. Psychosomatic Medicine 57, 37–47.

Cameron, L.D., Leventhal, H., Love, R.R., 2002. Trait anxiety influences tamoxifen-induced changes in lumbar bone mineral density and SHBG levels. Psychosomatic Medicine 64 (4), 612–620.

Charmaz, K., 1999. From the 'sick role' to stories of self: understanding the self in illness. In: Contrada, R.J., Ashmore, R.D. (Eds.), Self, Social Identity, and Physical Health. Oxford, New York, pp. 209–239.

Chrisman, N.J., Kleinman, A., 1983. Popular health care, social networks, and cultural meanings: the orientation of medical anthropology. In: Mechanic, D. (Ed.), Handbook of Health, Health Care, and Health Professions. Free Press, New York, pp. 569–590.

Cohen, S., Frank, E., Doyle, W.J., Skoner, D.P., Rabin, B.S., Gwaltney, J.M., 1998. Types of stressors that increase susceptibility to the common cold in healthy adults. Health Psychology 17, 214–223.

Croyle, R.T., Jemmott III, J.B., 1991. Psychological reaction of risk factor testing. In: Skelton, J.A., Croyle, R.T. (Eds.), Mental Representation in Health and Illness. Springer-Verlag, New York, pp. 85–107.

Dantzer, R., 2006. Cytokine-induced sickness behavior – mechanisms and implications. Annals New York Academy of Sciences 933, 222–234.

Demers, R.Y., et al., 1980. An exploration of the dimensions of illness behavior. Journal of Family Practice 11, 1085–1092.

DiMatteo, M.R., Lepper, H.S., Croghan, T.W., 2000. Depression is a risk factor for noncompliance with medical treatment: meta-analysis of the effects of anxiety and depression on patient adherence. Archives of Internal Medicine 160, 2101–2107.

Emanuel, E., Tanden, N., Altman, S., Amrstrong, S., Berwick, D., de Brantes, F., et al., 2012. A systematic approach to containing health care spending. New England Journal of Medicine 367 (10), 949–954.

Gortmaker, S.L., Eckenrode, J., Gore, S., 1982. Stress and the utilization of health services: a time series and cross-sectional analysis. Journal of Health and Social Behavior 23, 25–38.

Horne, R., Weinman, J., Hankins, M., 1999. The beliefs about medicines questionnaire: the development and evaluation of a new method for assessing the cognitive representation of medication. Psychology and Health 14, 1–24.

Horne, R., Buick, D., Fisher, M., Leake, H., Cooper, V., Weinman, J., 2004. Doubts about necessity and concerns about adverse effects: identifying the types of beliefs that are associated with non-adherence to HAART. International Journal of STD and AIDS 15, 38–39.

Horne, R., Fraasse, K., Cooper, V., Diefenbach, M.A., Leventhal, H., Leventhal, E.A., Petrie, K.J., 2012. The perceived sensitivity to medicines (PSM) scale: an evaluation of validity and reliability. British Journal of Health Psychology, 1–13. On Line.

Horowitz, C.R., Rein, S.B., Leventhal, H., 2004. A story of maladies, misconceptions and mishaps: effective management of heart failure. Social Science and Medicine 58, 631–643.

Idler, E.L., Benyamini, Y., 1997. Self-rated health and mortality: a review of twenty-seven community studies. Journal of Health and Social Behavior 38, 21–37.

Johnson, J.E., 1975. Stress reduction through sensation information. In: Sarason, I.C., Speilberger, C.O. (Eds.), Stress and Anxiety, vol. 2. Hemisphere, Washington, DC, pp. 361–373.

Kasl, S.V., Cobb, S., 1996. Health behavior, illness behavior and sick-role behavior. Archives of Environmental Health 12, 531–541.

Kellner, R., 1990. Somatization: theories and research. Journal of Nervous and Mental Disease 178, 150–160.

Kirmayer, L.J., Robbins, J.M. (Eds.), 1991. Current Concepts of Somatization: Research and Clinical Perspectives. American Psychiatric Press, Washington, DC.

Kulik, J.A., Mahler, H.I.M., 1997. Social comparison, affiliation, and coping with acute medical threats. In: Buunk, B.P., Gibbons, F.X. (Eds.), Health, Coping, and Well-Being: Perspectives from Social Comparison Theory. Erlbaum, Mahwah, NJ, pp. 227–261.

Lang, P.J., Levin, D.N., Miller, G.A., Kozak, M.J., 1983. Fear behavior, fear imagery and the psychophysiology of emotion: the problem of affective response integration. Journal of Abnormal Psychology 92, 276–306.

Leventhal, E.A., Leventhal, H., Shacham, S., Easterling, D.V., 1989. Active coping reduces reports of pain from childbirth. Journal of Consulting and Clinical Psychology 57, 365–371.

Leventhal, H., 1986. Health psychology: a social psychological perspective. In: Berkowitz, L. (Ed.), A Survey of Social Psychology, third ed. Holt, Rinehart, Winston, New York, pp. 449–476.

Leventhal, E.A., Crouch, M., 1997. Are there differences in perceptions of illness across the lifespan? In: Petrie, K.J., Wienman, J.A. (Eds.), Perceptions of Health and Illness: Current Research and Applications. Harwood Academic Press, London, pp. 77–102.

Leventhal, H., Idler, E.L., Leventhal, E.A., 1999. The impact of chronic illness on the self system. In: Contrada, R.J., Ashmore, R.D. (Eds.), Self, Social Identity, and Physical Health: Interdisciplinary Explorations. Oxford University Press, New York, pp. 185–208.

Leventhal, H., Leventhal, E.A., Breland, J.Y., 2011. Cognitive science speaks to the "common sense" of chronic illness management. Annals of Behavioral Medicine 41, 152–163.

Leventhal, H., Meyer, D., Nerenz, D., 1980. The common sense representation of illness danger. In: Rachman, S. (Ed.), Contributions to Medical Psychology. Pergamon, New York, pp. 7–30.

van Lankveld, W., Naring, G., van't Pad Bosch, P., van de Putte, L., 2000. The negative effect of decreasing the level of activity in coping with pain in rheumatoid arthritis: an increase in psychological distress and disease impact. Journal of Behavioral Medicine 23, 377–391.

Mechanic, D., 1979. The stability of health and illness behavior: results from a 16-year follow up. American Journal of Public Health 69, 1142–1145.

Mechanic, D., Volkart, E., 1961. Stress, illness behavior, and the sick role. American Sociological Review 26, 51–58.

Mora, P.A., Robitaille, C., Leventhal, H., Swigar, M., Leventhal, E.A., 2002. Trait negative affect relates to prior week symptoms, but not to reports of illness episodes, illness symptoms and care seeking among older people. Psychosomatic Medicine 64, 436–449.

Nakao, M., Barsky, A.J., 2007. Clinical application of somatosensory amplification in psychosomatic medicine. BioPsychoSocial Medicine 17, 1–7.

Parsons, T., 1975. The sick role and role of the physician reconsidered. Milbank Memorial Fund Quarterly 53, 257–278.

Phillips, L.S., Branch, W.T., Cook, C.B., et al., 2001. Clinical inertia. Annals of Internal Medicine 135, 825–834.

Rabin, C., Leventhal, H., Ward, S., Schmitz, M., 2001. Explaining retrospective reports of symptoms: anxiety, initial symptom experience, & and post-treatment symptoms. Health Psychology 20 (2), 91–98.

Robins, L.N., Helzer, J.E., Weissman, M.M., Orvaschel, H., Gruenberg, E., Burke, J.D., Regier, D.A., 1984. Lifetime prevalence of specific psychiatric disorders in three sites. Archives of General Psychiatry 41, 949–958.

Rosenstock, I.M., Hochbaum, G.M., Leventhal, H., 1960. Epidemic impact of the general population in two cities (Publication No. 766). In: The Impact of Asian Influenza on Community Life: A Study in Five Cities. The U.S. Department of Health, Education, and Welfare, the Public Health Service, pp. 53–77.

Vileikyte, L., Leventhal, H., Gonzalez, J.S., Peyrot, M., Rubin, R.R., Garrow, A., Waterman, C., Ulbrecht Cavanagh, P.R., Boulton, A.J.M., 2005. Diabetic peripheral neuropathy and depression: the association revisited. Diabetes Care 28, 2378–2383.

Watson, D., Pennebaker, J.W., 1989. Health complaints, stress, and distress: exploring the central role of negative affectivity. Psychological Review 96, 234–254.

WHO, 2003. Adherence to Long-Term Therapies: Evidence for Action. WHO, Geneva, pp. 7–9.

Young, J.T., 2004. Illness behavior: a selective review and synthesis. Sociology of Health & Illness 26, 1–31.

Zeiss, A.M., Lewinsohn, P.M., Rohde, P., Seeley, J.R., 1996. Relationship of physical disease and functional impairment to depression in older people. Psychology and Aging 11, 572–581.

Imaging Methods in Clinical Psychology

Sara V Tupak and Thomas Straube, Institute of Medical Psychology and Systems Neuroscience, University of Münster, Münster, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Neuroimaging of structural and functional alterations in mental disorders is a rapidly growing field of research with strong relevance for neurobiological theories of mental disorders. Findings concerning the neural correlates of psychopathological symptoms and associated alterations in cognitive and emotional processing, however, strongly depend on specific neuroimaging methods, experimental designs, and analytical procedures. The present article highlights past achievements in the field, the suitability and limitations of imaging methods and experimental designs, and their informational value for clinical psychology and future research.

Introduction

The use of neuroimaging methods in the field of biological psychiatry and clinical psychology has exponentially increased during the past 20 years, thereby promoting scientific theories and models about neurobiological factors contributing to psychopathology. The first studies investigating the relationship between psychopathological symptoms and brain metabolism emerged in the early 1980s by the use of single-photon emission computed tomography (SPECT) and positron emission tomography (PET; Buchsbaum et al., 1982). A new area of brain imaging started with the development and increased usage of functional magnetic resonance imaging (fMRI) in the mid-1990s (Wenz et al., 1994). Today, we have access to a multitude of structural and functional imaging studies highlighting the neural correlates of all kinds of mental disorders and clinical symptoms. Several of those studies had significant influence on our understanding of psychopathology and disorder-related neural processing. For instance, a crucial early fMRI study of Dierks et al. (1999) showed that auditory hallucinations in schizophrenic patients activated the same auditory brain regions as recorded speech does. Due to this study, the sensory quality of hallucinations became sort of 'visible' in terms of neurobiological parameters. Additionally, it highlighted the difficulty to differentiate between real speech and imagined speech psychotic patients typically experience. This was one of the first studies that demonstrated not only how functional imaging can contribute to the neurophysiological understanding but how it can also provide evidence to support specific psychological theories of mental illness symptoms.

The present article gives an overview of (1) the relevant imaging methods, (2) the general aims and scope of neuroimaging research of mental disorders, (3) experimental designs and analytical strategies, and (4) future challenges.

Imaging Techniques in Clinical Psychology

The technical progress in the field of brain imaging methods is immense (for a detailed review, see Bandettini, 2009), providing more and more sophisticated techniques to disentangle the neural correlates of mental disorders. In a narrower sense, all methods that enable the regional or systemic structural or functional localization of dysfunctional processing are considered as brain imaging techniques. These are primarily methods with moderate to high spatial resolution, which are based on local brain tissue, neurovascular, and metabolic measures, such as (f)MRI, functional near-infrared spectroscopy (fNIRS), PET, and SPECT. Computed tomography also belongs to this category, although rather rarely used. In a broader perspective and even though not further discussed in the present article, electrophysiological methods such as electroencephalography (EEG) and magnetoencephalography (MEG) are also considered as imaging methods despite lower spatial resolution. Instead, they offer detailed insight into the temporal dynamics of processing. A comparison of these methods with regard to spatial and temporal resolution and possible research fields is presented in **Table 1**.

Apart from their varying spatial and temporal resolution, functional imaging methods can be subdivided according to the measures they rely on and the processes they refer to. Direct recording of neural functioning is possible with electrophysiological methods, i.e., EEG and MEG. Hemodynamic measures such as fMRI and fNIRS rely on the principle of neurovascular coupling, assessing changes in regional hemoglobin concentrations. Therefore, they are indirect measures of brain function because they refer to processes that are closely related to neural functioning, i.e., oxygen consumption and regional cerebral blood flow. PET and SPECT reflect the biochemical processes that take place during neural activation and also rely on the principle of measuring regional cerebral blood flow changes. However, with these techniques it is possible through the use of different radioactive tracers to have a closer look at metabolic processes other than oxygen consumption (i.e., glucose metabolism or neurotransmitter-related alterations). As a consequence, molecular investigations of neural functioning in humans are only possible using PET and SPECT and despite the use of radiation, on cellular level, neurotransmitter functioning can only be assessed by these methods.

The use of imaging techniques, however, usually requires a great deal of time and effort. Especially with increasing spatial resolution, methods become more expensive and measurements are time consuming because of increasing preparation

Table 1 Overview of neuroimaging and electrophysiological methods used in clinical psychology

Technique	*Spatial resolution*	*Temporal resolution*	*Potential research areas*
MRI	Relatively high	N/A	Morphology Volume
CT	Moderate/ High	N/A	Morphology Volume
SPECT	Moderate	Very low	Functional metabolism Regional metabolism Connectivity
PET	Moderate	Very low	Functional metabolism Regional metabolism Connectivity
fMRI	High	Moderate	Functional hemodynamics Regional hemodynamics Connectivity
fNIRS	Fair	Moderate	Functional hemodynamics Regional hemodynamics Connectivity
MEG	Low	Very high	Temporal processing Lateralization
EEG	Poor	Very high	Temporal processing Lateralization

This is a slightly modified version of the table presented in our article 'Imaging Methods in Psychotherapy' (see Imaging methods in psychotherapy).
CT: computed tomography; EEG: electroencephalography; (f)MRI: (functional) magnetic resonance imaging; fNIRS: functional near-infrared spectroscopy; MEG: magnetoencephalography; PET: positron emission tomography; SPECT: single-photon emission computed tomography.

time. Less expensive techniques such as EEG and fNIRS are advantageous in their ecological validity. They can be applied in a quiet and more natural measurement setting and have no side effects. In contrast, bigger apparatuses are accompanied by loud noise, a narrow and uncomfortable measurement setting, or undesirable side effects due to radioactive tracers or X-rays. With increasing psychopathology or anxiousness, their application in clinical populations becomes difficult. Particularly anxious and claustrophobic individuals are less compliant with such measurements. Symptoms of a disorder itself, such as paranoid thinking in schizophrenia, can also be problematic in brain imaging studies. As a consequence and because all of these methods are more or less susceptible to motion artifacts, the signal-to-noise ratio can decrease significantly in clinical studies. Further problems for the interpretation of hemodynamic functional imaging studies are, for example, changes in baseline activation due to state-induced alterations in overall blood flow and oxygen concentration (e.g., hypocapnia due to hyperventilation in anxious individuals; Giardino et al., 2007).

The systems-level approach of primarily fMRI, fNIRS, EEG, and MEG offers a multitude of detailed research questions and allows for the study of neural clusters, modules, projections, structures, connectivity, functional specificity, and temporal aspects of processing. Very sophisticated acquisition and analyzing techniques exist (e.g., pattern classification; Haxby, 2012) and are constantly developed to enable even more elegant and noninvasive imaging of the human brain (Bandettini, 2009). Advanced structural imaging techniques such as diffusion tensor imaging allow for a more detailed study of the neural pathways and their alterations in mental illness. However, all mentioned functional imaging methods only provide information about correlational brain–behavior relationships. Causal inferences can only be drawn from brain stimulation studies using transcranial magnetic stimulation (TMS) or transcranial direct-current stimulation (tDCS) in varying spatial resolution.

Aims and Scope of Neuroimaging Research of Mental Disorders

Neuroimaging research in the field of mental disorders is an interdisciplinary research area, which brings together psychiatrists, clinical psychologists, and neuroscientists from other professions. The objectives of this research are manifold. Even though not mutually exclusive, they can broadly be divided into (1) those that aim to improve the basic knowledge about the structural and functional neural alterations that accompany mental illness and (2) those that aim to improve the prevention, diagnostics, treatment, and prediction of relapses of clinical symptoms. The potential research questions and resulting conclusions of such research, however, always depend on the neuroimaging methods and the experimental and analytical procedures they rely on.

Regarding the first aim, it is clear that the understanding of mental illness includes a detailed investigation of patients displaying a certain psychopathological behavior. However, it is primarily the study of psychological processes that manifest themselves in a dysfunctional way in patients, which is of high relevance for the field of clinical psychology. Nowadays, disorder-related characteristics are rather seen on a continuum with regard to their frequency and intensity instead of a binary differentiation between either being disordered or not. As such, characteristics that are associated with certain disorders are also present to some degree in healthy subjects either in the form of a current state or a personality trait. For instance, anxiety sensitivity, a trait that predominantly indicates how anxious an individual reacts to habitually occurring bodily sensations, seems to be normally distributed over the average population but is significantly increased in patients with panic disorder (McNally, 2002). In healthy subjects, heightened anxiety sensitivity is related to the experience of single spontaneous panic attacks and is supposed to constitute a risk factor for panic disorder (McNally, 2002). For clinical neuroscience research, the comparison of healthy individuals displaying low and high anxiety sensitivity may help to disentangle the neural correlates of increased fear of bodily reactions, a trait that represents a key symptom in panic disorder. Taken together, although the investigation of patients is of crucial relevance, the study of healthy individuals marks a substantial part for the field of clinical neuroscience.

Further, to increase the knowledge about neurobiological underpinnings of mental illness, it is important to gain insight into the associated mechanisms of dysfunctional processing and to distinguish them from otherwise normal processing. For example, for the successful localization of auditory hallucinations in the schizophrenic brain, it is also interesting to find out whether patients show general alterations in auditory processing and in how far the hallucinatory neural patterns differ from those of listening to externally (Dierks et al., 1999) or internally

generated speech (Simons et al., 2010). In general, psychiatric disorders are accompanied by alterations in many domains of processing such as perception, attention, cognition, language, memory, emotion, and motor function. This entails a vast variety of open research questions and experimental designs.

With respect to the second aim, the relevance of imaging studies for the clinical practice, many studies are dedicated to find the predictors of developing a certain disorder or factors forecasting relapse. A long-term goal of clinical imaging research is the prediction of the treatment response. Recent studies tried to identify neural patterns that can predict clinical outcome (Siegle et al., 2006) or correlate with global improvement (Buchheim et al., 2012). The idea behind such prospective studies is a more individualized and effective treatment by getting a better estimate of the course of disease, thereby lowering costs and efforts for the health care system. However, neuroimaging is relatively expensive and time-consuming prospective studies are rare (Vannini et al., 2007; Chan et al., 2003). Instead, recent studies increasingly focused on neural processes that are associated with clinical outcome (e.g., Siegle et al., 2006) or investigated etiological factors such as certain risk genotypes and how they influence brain functioning (Domschke et al., 2011). Finding neurobiological markers may further support the diagnostic process and may help to exclude differential diagnoses (e.g., Hahn et al., 2010). The identification of such markers may in turn represent a potential target for treatments. Neurofeedback, for instance, has been successfully used as a treatment strategy in attention-deficit hyperactivity disorder (ADHD) using EEG (Fuchs et al., 2003). Ongoing developments in real-time imaging techniques using fMRI and fNIRS are also promising (Weiskopf, 2012; Mihara et al., 2012). In addition, the use of brain stimulation techniques such as TMS, tDCS, or deep brain stimulation received increasing attention as potential treatments for depression and anxiety during recent years (Mayberg et al., 2005; Zwanzger et al., 2009). Brain stimulation, however, relies on the localization of critical target regions within the brain that can only be identified by neuroimaging. Furthermore, brain imaging techniques are increasingly used to investigate the dynamics of medical and psychotherapeutic treatment and clinical improvement. For detailed information about the monitoring of therapy progress by means of imaging techniques, the reader is referred to our article 'Imaging Methods in Psychotherapy' (see Imaging Methods in Psychotherapy).

Neuroimaging of mental disorders covers the whole spectrum of disorders with an exponentially increasing number of published studies. This is illustrated in **Figure 1** by providing estimated numbers of present review articles on fMRI studies related to mental disorders (last updated in February 2014). While disorders such as schizophrenia and depression received most attention in the beginning of human brain imaging research, other conditions such as anxiety disorders (Dresler et al., 2013), eating disorders (van Kuyck et al., 2009), addiction (Goldstein and Volkow, 2011), and ADHD (Bush et al., 2005) have increasingly come into focus during the past decade. In contrast, personality disorders, with the exception of borderline disorder (Driessen et al., 2004), have been widely disregarded until now. Psychiatric neuroscience equally refers to the assessment of adult populations as well as children and adolescents (Bush et al., 2005) and the age of the population of interest naturally varies with the respective disorder (e.g., ADHD vs dementia). Moreover, brain imaging research in clinical psychology is not restricted to mental disorders but extends into other fields such as clinical neuropsychology (Reichenberg and Harvey, 2007) and medical psychology (e.g., Gianaros and Sheu, 2009), which, however, cannot be discussed here.

Neuroimaging research of mental disorders is related to both structural and functional alterations. While functional measures represent stimuli-related, task-related, or so-called resting-state-related (Broyd et al., 2009) brain activation (metabolism, hemodynamics, electrical activity; see above), structural measurements inform about alterations in gray or white matter. Both may be relevant for clinical psychology. Structural alterations may represent a general factor for psychopathology. For structural imaging, several studies have found regional or overall alterations in brain tissue. Most prominent here are the studies on patients suffering from posttraumatic stress disorder (PTSD). Several studies reported smaller hippocampal volume in these patients; however, it is discussed whether this finding represents a cause rather than a consequence of the disorder (Gilbertson et al., 2002). Similarly, overall brain atrophy is a characteristic of various conditions such as Alzheimer's disease and alcohol dependence (Chan et al., 2003; Bjork et al., 2003). However, this research seems to be less relevant for clinical psychologists than functional imaging. In contrast to structural imaging, functional imaging research refers to the mapping of single symptoms or characteristics in processing of the disorder or disorder-related state or trait rather than mapping the disorder as a whole. It is important to keep in mind that every psychopathological syndrome represents a collection of clinical symptoms that can vary with regard to their frequency, intensity, and compilation. Furthermore, neuroimaging findings are also strongly dependent on experimental settings, imaging methods, and analytical procedures. This poses limitations to the generalizability of results from clinical imaging studies and their interpretation as discussed in the following article.

Experimental and Analytical Procedures in Functional Imaging of Mental Disorders

The informative value of functional imaging studies depends to a large degree on the experimental setting. For example, for panic disorder, previous studies reported divergent findings regarding prefrontal functioning. Many studies found hypoactivation in patients compared to controls while others observed the contrary (for a review, see Dresler et al., 2013). Without a consideration of the experimental task and the assumed underlying processes, no conclusion can be drawn since symptom provocation studies are known to activate partly similar prefrontal areas that were found to be hypoactive in emotion regulation studies in patients (Etkin et al., 2011). Similarly, many disorders share certain symptoms or information processing biases. Selective attention or avoidance behavior, for instance, is characteristic of many disorders such as anxiety and trauma-related disorders, depression, or eating disorders. As such, findings from the investigation of such

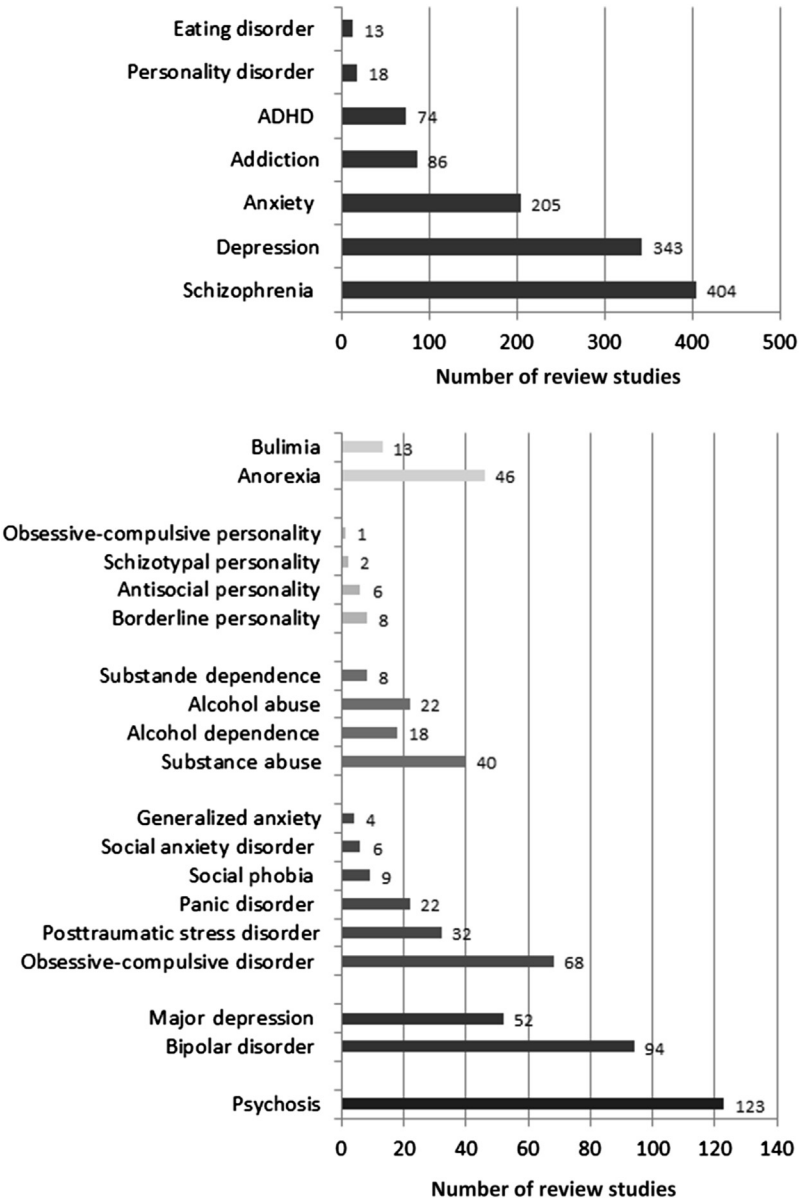

Figure 1 Present number of review studies about psychopathological conditions investigated by using fMRI. The upper panel shows the number of currently available review articles about fMRI research studies on the main conditions that are encountered in clinical psychology and psychiatry. Results were obtained through a literature search on PubMed searching for titles or abstracts comprising the keywords 'review' and 'fMRI' in addition to the disorder name as depicted above (latest update on 25 February 2014). The lower panel presents how fMRI review studies are divided into certain clinical disorders.

particular processes using adequate experimental paradigms such as the dot-probe task or approach-avoidance task might be transferable from one psychopathological condition to the other. In turn, the validity of the task must be scrutinized since it is essential that the symptom or the process that is intended to be mapped in the brain is actually targeted by the experimental manipulation. To conclude, there is no experimental task that covers all psychopathological alterations of a given disorder. Thus, it seems to be essential to investigate what kind of information processing is affected at the brain level and how the temporal features of these alterations look like, which, of course might differ in different brain regions (Schulz et al., 2013).

The variety of experimental designs that has been tested in clinical neuroimaging is immense. Now, there are hundreds of studies focusing on differential aspects of psychological processing, using a multitude of different stimuli and experimental settings. The currently available literature in the field covers all domains of general psychology, i.e., emotion, cognition, perception, sensation, motor function, memory, language, and learning. These domains, in turn, split up in even more specialized and detailed research fields, with an often very

specific focus on a particular disorder. Findings from basic research are instantaneously transferred to clinical studies and different patient groups. Various types of learning, for instance, have received a great deal of attention during past years (Maren et al., 2013). Here, and in contrast to higher order cognitive functions, findings and experimental designs can be almost directly transferred from animal studies to the human model. Due to such translational approaches, we got a thorough understanding about the basic neural processes accompanying different types of learning such as appetitive conditioning or fear learning (Moscarello and LeDoux, 2013; Vanelzakker et al., 2014). And still, research is ongoing, discovering even more finer details of those particular processes underlying human feeling and thinking.

The used experimental designs can be broadly subdivided in two categories: unspecific processes and symptom provocation/disorder-related tasks. Regarding the first point, a great deal of studies investigated general changes in neural functioning that accompany mental illness and are not primarily associated with the symptomatology of a given disorder. These unspecific processes, however, vary depending on the disorder. For example, studies on cognitive control are symptom-based in disorders such as schizophrenia but rather unrelated when it comes to anxiety disorders, whereas alterations in emotional processing are primarily linked to mood and anxiety disorders. Both cognitive functioning and emotional responsiveness have been studied extensively with regard to many mental disorders and, although not representing core symptoms, many of these studies found differences in neural functioning when compared to controls. Verbal fluency, for instance, a task requiring both linguistic and cognitive skills, has been associated with prefrontal hypofunction in a broad range of disorders (Dieler et al., 2012). This finding reflects the advantage of neuroimaging over neuropsychological tests because even in the absence of any behavioral performance deficits, neural alterations can be detected. Similarly, contextual learning plays an important role in many disorders and might lead to changes within the involved neural networks (Maren et al., 2013). To date, fear learning, context conditioning, and fear extinction have been primarily of interest in anxiety disorders because it is supposed that they reflect the neural processes that take place during the acquisition and recovery of clinical symptoms.

With the theoretical shift from the initial assumption that single neural modules underlie particular types of processing to the theory that such complex behavior is rather based on entire neural networks, also resting-state studies received increasing attention during recent years. Temporal correlations between spontaneous fluctuations in the measurement signal are supposed to reflect functional connectivity across different brain regions. Changes in this resting-state network have been predominantly reported for patients suffering from Alzheimer's disease and also for other clinical conditions (Broyd et al., 2009).

The other broad category of tasks focuses on particular symptom-related or disorder-relevant processes primarily during presentation of emotional stimuli. Symptom provocation designs (e.g., Boehme et al., 2013), information processing bias tasks (e.g., Luijten et al., 2011), the presentation of disorder-related stimuli without clear symptom provocation (e.g., during different task conditions; Straube et al., 2004), emotional learning (Maren et al., 2013), and emotion regulation tasks (e.g., Goldin et al., 2009b) belong to this category. The experimental tasks used for symptom provocation and disorder-related alterations may vary completely with the particular range of symptoms/assumed psychological alterations of a certain disorder. Also, the clinical picture may vary to a great extent depending on the disorder. While anxiety disorder patients often exhibit relatively homogenous patterns of clinical symptoms such as fear responses and avoidance behavior to aversive stimuli, mood disorders such as depression, or bipolar disorder are characterized by a more heterogeneous clinical picture. Consequently, symptom-based brain imaging in those disorders requires a finer selection of subsamples according to present symptoms (e.g., rumination, thought suppression). Other disorders, in contrast, completely differ with respect to their clinical manifestation but share certain symptoms. Impulsive behavior, for example, represents a common symptom in ADHD, addiction, or borderline personality disorder.

Thus, the provocation and testing of clinical symptoms in neuroimaging represents a great challenge. As variable as clinical symptoms are, the different are experimental designs and stimuli. Studies trying to catch spontaneously occurring symptoms such as hallucinations or panic attacks are rare (Dierks et al., 1999) and reports often represent rather incidental findings (Dresler et al., 2011). Instead of such passive designs, the majority of studies use stimuli for active and experimentally controlled symptom provocation. Particularly, anxiety disorders are suited for the investigation of disorder-related symptoms during brain imaging because anxiety symptoms can be provoked relatively easily by means of a range of different experimental manipulations. Typical stimuli that have been used include several kinds of pictorial or auditory stimuli (**Table 2**).

Chemosensory stimuli are less frequently used because of their rather minor relevance for symptom provocation in most disorders. However, a study of Vermetten et al. (2007) demonstrated the suitability of odors to provoke symptoms in PTSD. The authors exposed combat-related PTSD patients to the smell of diesel while undergoing PET and found an activation increase in fear-related brain regions (i.e., amygdala, insula, medial prefrontal cortex, and anterior cingulate cortex). Apart from sensory stimulus presentation, there are also pharmacological challenge studies, in which a pharmacological agent is administered to actively induce panic symptoms during the measurement. In those studies, an often used panicogenic agent is cholecystokinin-tetrapeptide (CCK-4), which leads to a profound increase in physiological arousal and anxiousness. CCK-4 reliably provokes panic attacks in the majority of both patients with panic disorder and healthy subjects and allows for a relatively precise timing of symptoms (Zwanzger et al., 2013; Dresler et al., 2013).

In contrast to pharmacological provocation, many studies work with experimental manipulations similar to psychotherapeutic exposition techniques. In panic disorder, for instance, *in vivo* exposition during brain imaging is achieved by interoception or hyperventilation (Dresler et al., 2013). For patients suffering from social anxiety disorder (SAD), public speaking or being observed by others represents a highly threatening situation, which has been successfully used to elicit symptoms during a PET (Tillfors et al., 2001) or fMRI measurement

Table 2 Stimuli being used for symptom provocation or emotional processing in anxiety disorders

	Pictorial	*Auditory*	*Chemosensory*	*Pharmacological*	*Other*
General	Pictures Videos Words	Sounds Imagery scripts	Odors Tastes	Psychotropic agents	Conditioned stimuli
PD	Biofeedback (e.g., heart rate, skin conductance, respiration, blood pressure)	Biofeedback (e.g., heart rate or respiration)		Panicogenic agents (e.g., CCK-4, caffeine)	Interoception Hyperventilation
SAD	Facial stimuli Performance feedback Video feedback of own behavior or reactions	Angry voices Performance feedback Listening to own voice			Being under observation (e.g., being filmed) Performance situations under observation (e.g., Trier social stress test)
PTSD	Trauma-related situations or trigger items	Trauma-related situations or trigger items Voices (e.g., yelling, angry voices, crying) Startle sounds	Trauma-related odors or tastes		

The table presents different stimulus classes that have been used in neuroimaging research to provoke clinical symptoms in anxiety disorders. Particular types of stimuli are listed for three frequently investigated clinical conditions: panic disorder (PD), social anxiety disorder (SAD), and posttraumatic stress disorder (PTSD).

(Boehme et al., 2013). The latter study provides clear evidence that activations in specific regions show different temporal profiles, with an initial response of the amygdala, but more sustained responses in other regions (**Figure 2**).

Apart from exposure to certain auditory or pictorial triggers, the recall of traumatic memories marks an effective *in sensu* strategy to provoke feelings of anxiousness and helplessness in PTSD patients undergoing fMRI (Driessen et al., 2004). Similarly, anxiousness or dissociative states can be elicited by script-driven imagery (Lanius et al., 2002). In contrast to pharmacological challenges, these kinds of manipulations are noninvasive. On the other hand, it is difficult to control for the extent of actual exposure since the effectiveness of the task depends on the cooperation of subjects and willingness to actively engage in the experimental task. Similar as during the pictorial presentation of aversive stimuli, subjects might disengage from the task and try to avoid exposure by looking away or even dissociating.

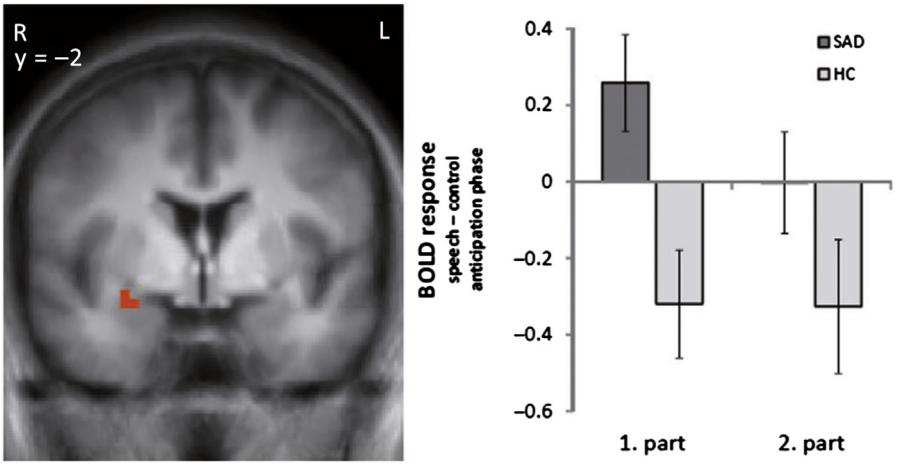

Figure 2 The figure displays findings from a fMRI study by Boehme et al. (2013) investigating the neural correlates during anticipation of a public-speaking task in a group of SAD patients compared to healthy controls. The results showed that during the first 20 s of the aversive anticipation period, patients displayed a profound increase in amygdala activation compared to the control task, whereas a decrease in amygdala activation was found for the same contrast in the control group. This activation increase was only present during the earlier phases of anticipation and diminished over the course of the entire anticipation period (40 s in total). Furthermore, anticipation of the speech task versus control condition led to an additional activation increase in the insula and a simultaneous decrease in the ventral striatum in patients compared to controls over the entire anticipatory period. The decrease in the latter region was negatively correlated with the reported anticipatory anxiety in the patient group. The figure has originally been published in Boehme, S., Ritter, V., Tefikow, S., Stangier, U., Strauss, B., Miltner, W.H.R., Straube, T., 2013. Brain activation during anticipatory anxiety in social anxiety disorder. Social Cognitive and Affective Neuroscience. http://dx.doi.org/10.1093/scan/nst129. Reproduced with kind permission from Oxford University Press.

While studies on anxiety disorders use primarily emotional stimuli for symptom provocation and changed information processing, other disorders are rather related to difficulties in other domains such as cognition, attention, or memory. Schizophrenic patients, for example, experience a range of deficits in executive functioning that can be triggered experimentally by cognitive tasks. The core symptoms of ADHD are attentional deficits, which can be easily assessed during functional imaging by means of conventional neuropsychological tests of attention such as the continuous performance task. The neurodegenerative processes of dementia are associated with a continuous aggravation of working memory performance particularly in early stages of the disease (i.e., mild cognitive impairment) during which experimental testing can be still accomplished by the patients.

Other tasks intend to investigate processes that are related to (dysfunctional) learning and emotion regulation. Several studies refer to strategies used also in cognitive-behavioral treatment such as exposure or cognitive restructuring. Direct exposure to disorder-related stimuli leads to the elicitation of clinical symptoms as described above, whereas repeated and prolonged exposure causes habituation, a process marking the core element of behavioral treatment. Therefore, and because extinction learning can be easily performed in other species allowing for a translational research approach, many imaging studies focused on (impaired) extinction learning in both healthy and clinical populations (e.g., Vanelzakker et al., 2014). A broad field of imaging studies concentrated on emotional regulation (e.g., Goldin et al., 2009b) and the automaticity and effectiveness of different regulatory strategies. Here, knowledge about the neural underpinnings helps to understand why patients engage in disorder maintaining behavior. Distraction from phobogenic stimuli, for example, is an often used avoidance strategy by anxious patients and leads to an attenuation of fear. On the neural level, mental distraction is characterized by a decrease in amygdala activation and an increase in prefontal functioning (Kanske et al., 2011), mirroring the release of fear patients might experience during avoidance. In contrast, cognitive reappraisal also leads to a decrease in amygdala activation but recruits a differential prefrontal network including more ventrally located brain regions (Kanske et al., 2011). However, in contrast to healthy subjects, this regulatory activation seems to be delayed in patients with SAD and encompasses fewer prefrontal areas. Although both groups showed an inverse correlation between the left dorsolateral prefrontal cortex and the amygdala, healthy subjects displayed similar correlational patterns in additional prefrontal areas (Goldin et al., 2009a).

Conclusions and Future Challenges

Clinical imaging studies have contributed to our basic knowledge and understanding of brain function in both single regions and functional networks in mental disorders. Although the neural markers of mental disease are still not clear and display a heterogeneous pattern across disorders, symptoms, and even between patients, they have supported psychobiological models and theories. During the past decades, neurobiological factors were increasingly accepted as etiological contributors to mental illness, thereby helping to minimize social stigmata associated with disorders such as schizophrenia and depression. In clinical practice, the explanation of metabolic and functional changes is nowadays often part of the treatment and increases patients' compliance to participate in either pharmacological or psychotherapeutic treatment.

However, the theoretical construct behind such research and the deliberate selection of appropriate methods are of particular importance. The theoretical conclusions regarding the underlying molecular or cellular processes depend on the imaging method. Likewise, experimental tasks can be differentiated according to the emotional and cognitive processes they are assumed to rely on. Typical studies in the field have focused on the processing of disorder-relevant stimuli, symptom provocation (*in vivo* and *in sensu*), or processes that are supposed to play a role in psychotherapy such as emotional regulation or extinction learning. Depending on the homogeneity of symptoms among patients with the same disorder, some disorders received more attention than others in previous research. Finding intelligent research designs for those disregarded psychopathological conditions represents a potential challenge for future research. Instead of mapping entire syndromes, the mapping of single symptoms by focusing on their underlying emotional or cognitive mechanisms seems more adequate. The systematic investigation of the neural correlates associated with a particular symptom that might be shared among different disorders and that might also be subclinically present among healthy individuals may lead to more specific neurobiological theories and might help to explain some of the controversial findings of past research. Future studies investigating factors that predict response to treatment or epidemiological factors such as genetic contributors and gene-environment interactions will further help to improve clinical treatment and the willingness to seek help. The finding of biomarkers, however, remains a huge challenge for future research. Until now, only a small percentage of the entire complexity of functional neural networks including their molecular, cellular, and structural underpinnings has been discovered but ongoing technical developments will provide additional sophisticated imaging methods to investigate the development, course, and recovery of mental disorders. The slow but increasingly adopted shift from the investigation of single neural modules to entire networks associated with particular types of processing may further help to unravel contradictory findings in the past. The use of specific symptom-related experimental tasks, presenting highly salient and specific stimulus classes, presumably signifies a first step toward this aim. In turn, the study of basic functions such as language or memory in patient groups represents an informative field of research also from a basic neuroscientist's perspective, providing important information about the fundamental neural mechanisms accompanying those processes. Determining the conditions under which a certain deviating neuronal pattern occurs may help to further elucidate common phenomena, such as hypofrontality, which can be observed across many patient groups with little specificity.

An often stated purpose of neuroimaging research is the aim to improve the early diagnosis and treatment of mental disorders. However, prospective studies in the field are rare because of the immense costs and efforts. It is also not clear, how a preventive treatment in those potential patients at risk should

look like. More reasonable, however, is the use of neuroimaging data for the prediction of treatment response: either psychotherapeutically, pharmacologically, or both. There is good evidence according to single studies (e.g., Siegle et al., 2006) but the reliability remains to be tested. Second, brain stimulation techniques rely on the exact determination of target regions in the brain. Likewise, neurofeedback and brain–computer interfaces (e.g., for severe neurological conditions) depend on clinical neuroimaging studies.

With the ongoing improvement of methods and analyzing techniques the knowledge about the neural correlates of mental disease will further contribute to the field of clinical psychology. The use of neuroimaging techniques may further help to disentangle the neural bases of the predispositions of mental disorders, their prodromal states, full-blown psychopathology, the course and treatment of these disorders, recovery, and the predicting factors of relapse. In the long run, imaging research in clinical psychology might be fundamental to provide reliable biological theories about mental illness thereby diminishing the social stigma associated with mental disorders and increasing patients' willingness to seek treatment.

See also: Anxiety and Fear Conditioning, Neural Basis of; Fear: Psychological and Neural Aspects; Imaging Methods in Psychotherapy; Limbic System; Magnetic Resonance and Molecular Imaging in Psychiatry; Near-Infrared Spectroscopy; Prefrontal Cortex; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported.

Bibliography

Bandettini, P.A., 2009. What's new in neuroimaging methods? Annals of the New York Academy of Sciences 1156, 260–293.

Bjork, J.M., Grant, S.J., Hommer, D.W., 2003. Cross-sectional volumetric analysis of brain atrophy in alcohol dependence: effects of drinking history and comorbid substance use disorder. The American Journal of Psychiatry 160, 2038–2045.

Boehme, S., Ritter, V., Tefikow, S., Stangier, U., Strauss, B., Miltner, W.H.R., Straube, T., 2013. Brain activation during anticipatory anxiety in social anxiety disorder. Social Cognitive and Affective Neuroscience. http://dx.doi.org/10.1093/scan/nst129.

Broyd, S.J., Demanuele, C., Debener, S., Helps, S.K., James, C.J., Sonuga-Barke, E.J.S., 2009. Default-mode brain dysfunction in mental disorders: a systematic review. Neuroscience and Biobehavioral Reviews 33, 279–296.

Buchheim, A., Viviani, R., Kessler, H., Kächele, H., Cierpka, M., Roth, G., George, C., Kernberg, O.F., Bruns, G., Taubner, S., 2012. Changes in prefrontal-limbic function in major depression after 15 months of long-term psychotherapy. PLoS ONE 7, e33745.

Buchsbaum, M.S., Ingvar, D.H., Kessler, R., Waters, R.N., Cappalletti, J., van Kammen, D.P., King, A.C., Johnson, J.L., Manning, R.G., Flynn, R.W., Mann, L.S., Bunney Jr, W.E., Sokoloff, L., 1982. Cerebral glucography with positron tomography: use in normal subjects and in patients with schizophrenia. Archives of General Psychiatry 39, 251–259.

Bush, G., Valera, E.M., Seidman, L.J., 2005. Functional neuroimaging of attention-deficit/hyperactivity disorder: a review and suggested future directions. Biological Psychiatry 57, 1273–1284.

Chan, D., Janssen, J.C., Whitwell, J.L., Watt, H.C., Jenkins, R., Frost, C., Rossor, M.N., Fox, N.C., 2003. Change in rates of cerebral atrophy over time in early-onset Alzheimer's disease: longitudinal MRI study. The Lancet 362, 1121–1122.

Dieler, A.C., Tupak, S.V., Fallgatter, A.J., 2012. Functional near-infrared spectroscopy for the assessment of speech related tasks. Brain and Language 121 (2), 90–109.

Dierks, T., Linden, D.E.J., Jandl, M., Formisano, E., Goebel, R., Lanfermann, H., Singer, W., 1999. Activation of Heschl's Gyrus during auditory hallucinations. Neuron 22, 615–621.

Domschke, K., Reif, A., Weber, H., Richter, J., Hohoff, C., Ohrmann, P., Pedersen, A., Bauer, J., Suslow, T., Kugel, H., Heindel, W., Baumann, C., Klauke, B., Jacob, C., Maier, W., Fritze, J., Bandelow, B., Krakowitzky, P., Rothermundt, M., Erhardt, A., Binder, E.B., Holsboer, F., Gerlach, A.L., Kircher, T., Lang, T., Alpers, G.W., Strohle, A., Fehm, L., Gloster, A.T., Wittchen, H.U., Arolt, V., Pauli, P., Hamm, A., Deckert, J., 2011. Neuropeptide S receptor gene – converging evidence for a role in panic disorder. Molecular Psychiatry 16, 938–948.

Dresler, T., Guhn, A., Tupak, S., Ehlis, A.-C., Herrmann, M., Fallgatter, A., Deckert, J., Domschke, K., 2013. Revise the revised? New dimensions of the neuroanatomical hypothesis of panic disorder. Journal of Neural Transmission 120, 3–29.

Dresler, T., Hahn, T., Plichta, M., Ernst, L., Tupak, S., Ehlis, A.-C., Warrings, B., Deckert, J., Fallgatter, A., 2011. Neural correlates of spontaneous panic attacks. Journal of Neural Transmission 118, 263–269.

Driessen, M., Beblo, T., Mertens, M., Piefke, M., Rullkoetter, N., Silva-Saavedra, A., Reddemann, L., Rau, H., Markowitsch, H.J., Wulff, H., Lange, W., Woermann, F.G., 2004. Posttraumatic stress disorder and fMRI activation patterns of traumatic memory in patients with borderline personality disorder. Biological Psychiatry 55, 603–611.

Etkin, A., Egner, T., Kalisch, R., 2011. Emotional processing in anterior cingulate and medial prefrontal cortex. Trends in Cognitive Sciences 15, 85–93.

Fuchs, T., Birbaumer, N., Lutzenberger, W., Gruzelier, J., Kaiser, J., 2003. Neurofeedback treatment for attention-deficit/hyperactivity disorder in children: a comparison with methylphenidate. Applied Psychophysiology and Biofeedback 28, 1–12.

Gianaros, P.J., Sheu, L.K., 2009. A review of neuroimaging studies of stressor-evoked blood pressure reactivity: emerging evidence for a brain-body pathway to coronary heart disease risk. NeuroImage 47, 922–936.

Giardino, N.D., Friedman, S.D., Dager, S.R., 2007. Anxiety, respiration, and cerebral blood flow: implications for functional brain imaging. Comprehensive Psychiatry 48, 103–112.

Gilbertson, M.W., Shenton, M.E., Ciszewski, A., Kasai, K., Lasko, N.B., Orr, S.P., Pitman, R.K., 2002. Smaller hippocampal volume predicts pathologic vulnerability to psychological trauma. Nature Neuroscience 5, 1242–1247.

Goldin, P.R., Manber-Ball, T., Werner, K., Heimberg, R., Gross, J.J., 2009a. Neural mechanisms of cognitive reappraisal of negative self-beliefs in social anxiety disorder. Biological Psychiatry 66, 1091–1099.

Goldin, P.R., Manber-Ball, T., Hakimi, S., Canli, T., Gross, J.J., 2009b. Neural bases of social anxiety disorder: emotional reactivity and cognitive regulation during social and physical threat. Archives of General Psychiatry 66, 170–180.

Goldstein, R.Z., Volkow, N.D., 2011. Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications. Nature Reviews Neuroscience 12, 652–669.

Hahn, T., Marquand, A.F., Ehlis, A.-C., Dresler, T., Kittel-Schneider, S., Jarczok, T.A., Lesch, K.-P., Jakob, P.M., Mourao-Miranda, J., Brammer, M.J., Fallgatter, A.J., 2010. Integrating neurobiological markers of depression. Archives of General Psychiatry 68, 361–368.

Haxby, J.V., 2012. Multivariate pattern analysis of fMRI: the early beginnings. NeuroImage 62, 852–855.

Kanske, P., Heissler, J., Schönfelder, S., Bongers, A., Wessa, M., 2011. How to regulate emotion? neural networks for reappraisal and distraction. Cerebral Cortex 21, 1379–1388.

van Kuyck, K., Gérard, N., Laere, K.V., Casteels, C., Pieters, G., Gabriëls, L., Nuttin, B., 2009. Towards a neurocircuitry in anorexia nervosa: evidence from functional neuroimaging studies. Journal of Psychiatric Research 43, 1133–1145.

Lanius, R.A., Williamson, P.C., Boksman, K., Densmore, M., Gupta, M., Neufeld, R.W.J., Gati, J.S., Menon, R.S., 2002. Brain activation during script-driven imagery induced dissociative responses in PTSD: a functional magnetic resonance imaging investigation. Biological Psychiatry 52, 305–311.

Luijten, M., Veltman, D.J., den Brink, W. v., Hester, R., Field, M., Smits, M., Franken, I.H.A., 2011. Neurobiological substrate of smoking-related attentional bias. NeuroImage 54, 2374–2381.

Maren, S., Phan, K.L., Liberzon, I., 2013. The contextual brain: implications for fear conditioning, extinction and psychopathology. Nature Reviews Neuroscience 14, 417–428.

Mayberg, H.S., Lozano, A.M., Voon, V., McNeely, H.E., Seminowicz, D., Hamani, C., Schwalb, J.M., Kennedy, S.H., 2005. Deep brain stimulation for treatment-resistant depression. Neuron 45, 651–660.

McNally, R.J., 2002. Anxiety sensitivity and panic disorder. Biological Psychiatry 52, 938–946.

Mihara, M., Miyai, I., Hattori, N., Hatakenaka, M., Yagura, H., Kawano, T., Okibayashi, M., Danjo, N., Ishikawa, A., Inoue, Y., Kubota, K., 2012. Neurofeedback using real-time near-infrared spectroscopy enhances motor imagery related cortical activation. PLoS ONE 7, 1–13.

Moscarello, J.M., LeDoux, J.E., 2013. The contribution of the amygdala to aversive and appetitive pavlovian processes. Emotion Review 5, 248–253.

Reichenberg, A., Harvey, P.D., 2007. Neuropsychological impairments in schizophrenia: integration of performance-based and brain imaging findings. Psychological Bulletin 133, 833–858.

Schulz, C., Mothes-Lasch, M., Straube, T., 2013. Automatic neural processing of disorder-related stimuli in social anxiety disorder (SAD): faces and more. Frontiers in Psychology 4, 282.

Siegle, G.J., Carter, C.S., Thase, M.E., 2006. Use of fMRI to predict recovery from unipolar depression with cognitive behavior therapy. American Journal of Psychiatry 163, 735–738.

Simons, C.J.P., Tracy, D.K., Sanghera, K.K., O'Daly, O., Gilleen, J., Dominguez, M.-d.-G., Krabbendam, L., Shergill, S.S., 2010. Functional magnetic resonance imaging of inner speech in schizophrenia. Biological Psychiatry 67, 232–237.

Straube, T., Kolassa, I.-T., Glauer, M., Mentzel, H.-J., Miltner, W.H.R., 2004. Effect of task conditions on brain responses to threatening faces in social phobics: an event-related functional magnetic resonance imaging study. Biological Psychiatry 56, 921–930.

Tillfors, M., Furmark, T., Marteinsdottir, I., Fischer, H., Pissiota, A., Långström, B., Fredrikson, M., 2001. Cerebral blood flow in subjects with social phobia during stressful speaking tasks: a PET study. American Journal of Psychiatry 158, 1220–1226.

Vanelzakker, M.B., Kathryn Dahlgren, M., Caroline Davis, F., Dubois, S., Shin, L.M., 2014. From Pavlov to PTSD: The extinction of conditioned fear in rodents, humans, and anxiety disorders. Neurobiology of Learning and Memory 113, 3–18.

Vannini, P., Almkvist, O., Dierks, T., Lehmann, C., Wahlund, L.-O., 2007. Reduced neuronal efficacy in progressive mild cognitive impairment: a prospective fMRI study on visuospatial processing. Psychiatry Research: Neuroimaging 156, 43–57.

Vermetten, E., Schmahl, C., Southwick, S.M., Bremner, J.D., 2007. Positron tomographic emission study of olfactory induced emotional recall in veterans with and without combat-related posttraumatic stress disorder. Psychopharmacology Bulletin 40, 8–30.

Weiskopf, N., 2012. Real-time fMRI and its application to neurofeedback. NeuroImage 62, 682–692.

Wenz, F., Schad, L.R., Knopp, M.V., Baudendistel, K.T., Flömer, F., Schröder, J., van Kaick, G., 1994. Functional magnetic resonance imaging at 1.5 T: activation pattern in schizophrenic patients receiving neuroleptic medication. Magnetic Resonance Imaging 12, 975–982.

Zwanzger, P., Fallgatter, A.J., Zavorotnyy, M., Padberg, F., 2009. Anxiolytic effects of transcranial magnetic stimulation—an alternative treatment option in anxiety disorders? Journal of Neural Transmission 116, 767–775.

Zwanzger, P., Zavorotnyy, M., Gencheva, E., Diemer, J., Kugel, H., Heindel, W., Ruland, T., Ohrmann, P., Arolt, V., Domschke, K., Pfleiderer, B., 2013. Acute shift in glutamate concentrations following experimentally induced panic with cholecystokinin tetrapeptide—a 3T-MRS study in healthy subjects. Neuropsychopharmacology 38, 1648–1654.

Imaging Methods in Psychotherapy

Sara V Tupak and Thomas Straube, University of Münster, Münster, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Imaging the neural changes that accompany psychotherapeutic treatment of mental disorders received increasing attention in recent years. Although the number of empirical studies in this field of research is relatively small, the informative value that can be gained from such studies is immense for both psychiatric neuroscience and evidence-based psychotherapy. This article gives an overview of important theoretical and methodological issues, discusses the potential mechanisms of action on brain level, and provides information about the present state of research and future directions for brain imaging research on psychotherapy-induced changes in neural functioning.

Introduction

In clinical psychology and psychiatric practice, psychotherapeutic interventions are often part of the standardized treatment approach for many of the mental disorders. Moreover, psychotherapy has been shown to be equally effective than or even superior to pharmacotherapy for many disorders and cases, but most often a combined treatment is recommended (Bandelow et al., 2007). Compared to pharmacological treatment, however, theoretical models of psychotherapeutic effects often lack a sufficient neurobiological basis. Whereas much research has focused on the neural correlates that accompany psychopathology, only little is known about the functional and structural changes that take place during psychotherapy-related recovery. During the past 20 years, however, an increasing amount of studies started to systematically assess the neurobiology behind evidence-based forms of psychotherapy such as cognitive-behavioral therapy (CBT) or psychodynamic theraphy by means of brain imaging techniques. Earlier studies primarily compared brain function before and after therapy by using imaging methods that enable the measurement of changes in overall brain metabolism such as positron-emission tomography (PET). With the increasing use of functional magnetic resonance imaging (fMRI) around the beginning of the new millennium, more and more studies investigated the specific functional activation changes that accompany information processing of disorder-related stimuli and tasks. However, there still exist relatively few studies leveraging functional and structural brain imaging techniques for the study of psychotherapeutic mechanisms. One of the foremost reasons for the sparse amount of studies in the field are the costs and efforts those studies bring about. Like pharmacological treatment studies, psychotherapy studies are quite time consuming and due to the face-to-face character of the treatment they require vast staff resources. Nevertheless, the informative value of such research for both clinical practice and neurobiological models of mental disorders is immense and warrants all of its efforts.

Understanding the Neurobiological Mechanisms Behind Psychopathology by Use of (Functional) Imaging Research

Psychopathology is linked to a range of functional impairments and sometimes also to structural reorganizations on brain level reflecting the disorder-related behavioral, emotional, and cognitive symptoms. Likewise, clinical improvement and recovery are accompanied by fundamental changes in behavior and subjective experience, which might in turn originate from neural modulations (Carrig et al., 2009). While the former (i.e., the acute disorder-related state) has been addressed by many neuroimaging studies, the neural correlates of clinical improvement lack sufficient neurobiological evidence despite the existence of several theoretical models. The missing biological foundation might in turn account for a range of prejudices against psychotherapeutic treatment in the general population, although there is plenty of evidence for the efficacy of cognitive-behavioral and psychodynamic therapy. However, understanding the neural correlates of psychotherapy and their effects on clinical improvement may not only help to increase compliance among patients but primarily serves for the testing and development of previous and future neurobiological and psychological models explaining the modulation of psychopathological behavior that takes place during psychotherapy and recovery (Linden, 2008; Etkin et al., 2005). Furthermore, the combination of brain and psychotherapy research allows for the detailed investigation of functional associations between single brain regions and certain clinical symptoms (Straube and Miltner, 2011). Even more, detailed information about the curative mechanisms of psychotherapy on brain level and the potential individual factors that either support or debilitate these effects can aid the selection of treatment, forecast of clinical improvement (Straube and Miltner, 2011; Linden, 2008; Roffman et al., 2005), or even lead to the development of new treatment protocols (Linden, 2008). Recent attempts in this area have been made by using neurofeedback, transcranial magnetic stimulation, transcranial direct current stimulation, or deep-brain stimulation to target and modulate brain regions associated with disorder-related symptomatology.

Also, the understanding of the complementing mechanisms behind the combined use of pharmacotherapy and psychotherapy on neural level may support the planning of individual treatment strategies, thereby preventing frustrating and time-consuming trial-and-error approaches (Roffman et al., 2005). Evidence exists, for example, showing that medication and psychotherapy affect partly the same and partly distinct brain areas in depressive patients (Martin et al., 2001). Others found that the degree of activation in one specific region

before treatment is associated with a better response to either behavior therapy or medication (Brody et al., 1998). Brain imaging, however, can help to disentangle the differential effects of both kinds of treatment, thereby providing important information for neurobiological processes that take place during recovery.

Because psychopathology is primarily associated with functional changes in neural information processing, especially functional imaging research may be helpful to identify biological markers of certain mental disorders or classes of symptoms which are common among several distinct disorders (e.g., avoidance behavior in anxiety disorders; Linden, 2008; Etkin et al., 2005).

Finally, imaging techniques may serve as an objective method to monitor progress during therapy, potentially even indexing the treatment duration and potential need for pharmacological add-on treatment that is necessary to achieve a satisfactory clinical outcome. On the contrary, they may also serve to detect and prevent potential relapses in advance (Etkin et al., 2005).

A summary of the aims and potential benefits of imaging research on psychotherapy is presented in **Table 1**. Taken together, previous and future imaging studies on psychotherapy and therapy-induced clinical improvement support and complement our understanding of psychopathology from early prodromal signs of illness to complete recovery of function. They may support interdisciplinary clinical practice and research between psychology and psychiatry by providing a neurobiological basis for psychotherapeutic effects, and they may further highlight the great potential of psychotherapy as an evidence-based treatment of mental disorders. The following sections provide an overview of the potential mechanisms of action on neural level, past experimental approaches and findings in the field, and the particular limitations and improvement opportunities for future research.

Mechanisms of Action

How can we determine if psychotherapy changes the brain and if so, what does that mean? In comparison to healthy control subjects, structural and functional alterations in brain activation have been found for many of the mental disorders that benefit from psychotherapeutic interventions such as mood disorders, anxiety disorders, or eating disorders. It seems reasonable that clinical improvement and an attenuation of disorder-related symptoms might in turn also be accompanied by functional or even structural changes on brain level, a process that is commonly referred to as neuroplasticity. If both the development of psychopathological symptoms and their remission are linked to neuroplastic processes, there are several potential mechanisms of how therapeutic interventions may affect neural functioning. On a macrolevel, improved symptomatology might result from a normalization of brain function, compensatory brain activation, or changes in functional connectivity (Linden, 2008; Straube and Miltner, 2011). Regarding the first, the simplest mechanism of action might be the 'reversion' of disorder-related functional, but also structural, alterations back to a 'normal' level that is undistinguishable from nonclinical samples. Overly increased prefrontal recruitment in patients with major depressive disorder (MDD), for instance, may return to nonclinical baseline levels over the course of treatment. According to this idea, the development of psychopathological symptoms is understood as a completely reversible process on both behavioral (symptoms) and physiological (brain) levels.

If the initial homeostasis may not be regained, compensatory processes may support and index clinical improvement. Instead of back-to-normal prefrontal functioning, the remission of depressive symptoms may be associated with compensatory increases or decreases in brain regions other than the disorder-related ones. In this view, disorder-related neural alterations are rather understood as a physiological trait instead of state of the temporarily present clinical symptoms. An efficient way to compensate for the functional impairment in certain brain regions might be the recruitment of other nonimpaired brain regions, a process which is often observed in patients with brain lesions and which might also apply to psychiatric patients.

According to a connectionist's view, therapy-related changes may not primarily take place in certain parts of the brain but rather in entire neural networks. There are, for instance, several studies showing deviations in functional connectivity between the amygdala and prefrontal cortex (PFC) for various anxiety disorders (Etkin et al., 2009; Kim et al., 2011). Functional connectivity, i.e., the joint or antagonistic interplay within one single or between two or more brain regions, is a direct result of neuroplasticity. Neuroplasticity in turn builds the basis for normal and pathological forms of learning, memory, and mental functions. In the context of the neural mechanisms of action underlying psychotherapeutic effects, improved symptomatology might thus result from neuroplastic changes on a microlevel, including the molecular, cellular, and systems level. On a molecular basis, the learning or relearning of certain stimulus–response associations, which is the principal

Table 1 Benefits of neuroimaging research on psychotherapy

Benefits for basic research	*Benefits for clinical practice*
Testing of theoretical constructs and models	Increase compliance
Identification of biological markers	Psychoeducation
Functional link between symptoms and neural structure	New treatment protocols
Conceptualization	Selection of treatment
Complementing effects of medication and psychotherapy	Prognosis and clinical outcome
	Prevention
	Monitoring progress

concept of behavior therapy, may induce functional or structural alterations in receptor systems, the synthesis and depletion of neurotransmitters, enzymes, ion channels, chromatin, or gene expression. On cellular level, clinical improvement may be accompanied by several pre- and postsynaptic processes but also by synaptogenesis or (hippocampal) neurogenesis. Particularly, the long-term potentiation (LTP) of particular projections, e.g., the top-down control of the PFC over subcortical limbic structures, may improve functional connectivity also in the long run thereby preventing future relapses. In contrast, long-term depression (LTD) can help to 'normalize' responses in brain regions that are assumed to react hyperactive or overly sensitive, as it has commonly been postulated for the amygdala and many anxiety disorders (Rauch et al., 2003; Etkin and Wager, 2007). On a systems level, a general reorganization of activation clusters and projections in gray and white matter may lead to improved functional connectivity following psychotherapy.

Taken together, there are several potential approaches to measure clinical improvement by means of functional or structural brain imaging (**Table 2**). Because the direct assessment of the effects on microlevel, particularly molecular and cellular changes, are difficult or ethically impossible to investigate in humans, most research in the field focused on indirect measures of molecular and cellular functioning primarily by the use of imaging techniques relying on metabolic indices of neural activity. Most evidence, however, comes from studies that elucidated therapy-induced changes on the systems level with a particular interest in the processes that are changing on the macrolevel, as mentioned earlier. Especially functional changes in disorder-related brain regions or functional connectivity between those areas have been studied in past research (e.g., Straube et al., 2006; Buchheim et al., 2012; Kircher et al., 2013). However, one has to keep in mind that the immediate effects of psychotherapeutic interventions or single sessions have never been studied so far. Most present evidence is related to measures that were conducted after the completion of therapy or, in rare cases, between single sessions (i.e., a case study of Schiepek et al., 2009). As a result, the neural findings from such studies are generally related to overall clinical improvement following psychotherapeutic treatment. In addition, online measures of brain activation during treatment are complicated and the standardization of such measures is virtually impossible. To reflect the immediate effects of certain therapeutic strategies (e.g., systematic desensitization or direct exposure), particular experimental designs may help to overcome methodological limitations and to gain a better understanding of the processes that may take place during according therapy sessions. Potential approaches for efficient past and future research are described in detail in the following section. To sum up, however, it is important to point out that psychotherapy by itself is very difficult to standardize. Even evidence-based manualized treatment varies between individual patients and therapists and depends on many individual factors such as present situation, symptoms, or coping strategies. The combination of such a nonstandardizable treatment on the one hand, and the aspirations of imaging research for a strict scientific and methodological practice on the other hand, has been a great challenge for past and will still remain a challenge for future studies.

Table 2 Potential effects of psychotherapy on molecular, cellular, and system levels and the according imaging techniques which enable the study of these effects

	Level				
			System		
	Molecular	*Cellular*	*Structural*	*Functional*	*Temporal*
Mechanisms of action (normalization, compensation, connectivity)	Receptor systems Synthesis/depletion Enzymes Ion channels Chromatin Gene expression	Pre-postsynaptic processes Neurotransmitter function Synaptogenesis Neurogenesis LTP and LTD	Reorganization Projections Clusters/modules Structure/volume Gray matter White matter	Reorganization Projections Cluster modules Functional connectivity	Temporal processing
Methods	PET SPECT	PET EEG	PET MRI CT	PET fMRI fNIRS EEG MEG	EEG MEG (fMRI) (fNIRS)
Specific analyzing techniques	Depending on radiopharmacon: e.g., dopamine, glucose, rCBF Postmortem	Postmortem Depending on radiopharmacon: e.g., dopamine, glucose, rCBF	DTI VBM ASL	ERP DTI FFT/wavelet ASL	ERP FFT/wavelet

ASL, arterial spin labeling; CT, computed tomography; DTI, diffusion tensor imaging; EEG, electroencephalography; ERP, event-related potential; (f)MRI, (functional) magnetic resonance imaging; (f)NIRS, (functional) near-infrared spectroscopy; FFT, fast Fourier transformation; LTD, long-term depression; LTP, long-term potentiation; MEG, magnetoencephalography; PET, positron-emission tomography; rCBF, regional cerebral blood flow; SPECT, single-photon emission tomography; VBM, voxel-based morphometry; methods that are presented in parentheses are less suited but still able to infer valuable information.

Studying the Neural Correlates of Psychotherapy

A common way to investigate neural changes related to psychotherapy includes at least either a pre–posttreatment comparison or the posttreatment comparison of a treatment group to a comparable control group (e.g., patients on a waiting list). A combination of both strategies is obviously advantageous: First, because the former cannot control for effects of time (e.g., potential habituation effects to the measurement procedure), and second, because the latter provides only little information about treatment-related changes and may reflect only general trait-related differences between groups in case of a healthy control sample. The experimental approaches regarding time points of measurement, psychotherapeutic treatments, control groups, methods, and experimental tasks vary from study to study.

With respect to time, the majority of studies decided for a pre–post design (e.g., Straube et al., 2006) while others included also intermediate (Schiepek et al., 2009) and follow-up measurements (Schienle et al., 2009). Treatments mostly referred to evidence-based forms of psychotherapy such as (cognitive) behavioral (CBT) and psychodynamic therapy or related therapeutic strategies such as behavioral activation (Dichter et al., 2010) or interpersonal therapy (IPT) (e.g., Brody et al., 2001; Martin et al., 2001), respectively. Apart from those general treatment approaches, there are also studies that investigated other specific forms of therapy such as eclectic behavioral therapy (Lindauer et al., 2008) and eye movement desensitization and reprocessing (EMDR; Levin et al., 1999) which are particularly tailored to patients with posttraumatic stress disorder (PTSD), systemic couple treatment (Schiepek et al., 2009), neuropsychological and cognitive remediation treatment for schizophrenic patients (e.g., Wykes et al., 2002), or sleep deprivation as an antidepressive treatment (Gillin et al., 2001).

Regarding the selection of one or more control groups, several options exist which in turn influence the informative value that can be drawn from the results. A great deal of studies included a healthy control group which particularly allows for the investigation of 'normalization' processes following psychotherapy. Other studies compared the treatment group to patients who received no treatment between the time points of measurement (e.g., patients on a waiting list; Furmark et al., 2002; Straube et al., 2006) or received psychopharmacological medication as an alternative treatment (e.g., McClure et al., 2007; Furmark et al., 2002). A third possibility is the post hoc comparison of treatment responders and nonresponders (Furmark et al., 2002). Generally, the inclusion of a patient control group seems recommended to control for potential overall disorder-related alterations in brain metabolism which may confound the results. There are only few studies including both a patient control group and healthy subjects.

The variety of available functional and structural imaging methods offers a multitude of possibilities to study the effects of psychotherapy on the brain. The selection of methods depends, however, pretty much on the particular research questions and hypotheses at hand (**Table 3**). Structural imaging techniques such as MRI are particularly suited to investigate whether morphological between-subject differences can predict later treatment responses. It has been shown, for instance, that PTSD patients who benefitted from CBT had larger rostral anterior cingulate cortex volume before treatment than nonresponders (Bryant et al., 2008b). Until now, however, most research concentrated on treatment effects on the overall metabolic activity or functional activation in certain disorder-related brain regions using PET and fMRI, respectively. Variations in regional cerebral blood flow (rCBF) were mostly assessed by means of PET, only a minority of studies used other methods such as computerized tomography (CT; Nakatani et al., 2003) or single-photon emission tomography (SPECT; Martin et al., 2001; Lindauer et al., 2008). For the investigation of neural changes in information processing following psychotherapy, fMRI has been the method of choice for most studies. In addition, fMRI allows for specific analyses regarding functional connectivity and the predictive value of pretreatment brain responses in disorder-related brain regions.

Table 3 Overview over neuroimaging and electrophysiological methods suited for the study of psychotherapy effects

Technique	*Spatial resolution*	*Temporal resolution*	*Potential research questions*	*Example study*
MRI	Relatively high	N/A	Morphology Volume	Bryant et al. (2008b)
CT	Moderate/high	N/A	Morphology Volume	Nakatani et al. (2003)
SPECT	Moderate	Very low	Metabolism	Lindauer et al. (2008)
PET	Moderate	Very low	Metabolism	Furmark et al. (2002)
fMRI	High	Moderate	Connectivity Functional activation Regional activation	Straube et al. (2006)
fNIRS	Fair	Moderate	Functional activation Regional activation	N/A
MEG	Low	Very high	Temporal processing Lateralization	Adenauer et al. (2011)
EEG	Poor	Very high	Temporal processing Lateralization	Thase et al. (1997)

CT, computed tomography; EEG, electroencephalography; (f)MRI, (functional) magnetic resonance imaging; fNIRS, functional near-infrared spectroscopy; MEG, magnetoencephalography; PET, positron-emission tomography; SPECT, single-photon emission tomography.

Surprisingly, until now imaging research on psychotherapy focused exclusively on those relatively expensive methods, which are partly accompanied by undesirable side-effects due to radiation (i.e., PET and SPECT), although there are a range of more convenient measurement techniques such as electroencephalography (EEG) or functional near-infrared spectroscopy (fNIRS). Actually, to date there exist relatively few electrophysiological studies in this field of research, although the high temporal resolution of such methods (e.g., EEG or magnetoencephalography (MEG)) allows for a far more detailed insight into the temporal characteristics of information processing when compared to methods that rely on hemodynamic (i.e., fMRI) or metabolic activation (i.e., PET).

In contrast to resting state measures, the performance of experimental tasks allows for a better control of the mental processes underlying the according brain scans. Depending on the task, specific aspects of information processing and how this processing is affected by psychotherapy can be studied. Here, previous experiments predominantly focused on activation changes to disorder-related stimuli or tasks, following exposure or symptom provocation. Also the general theoretical foundations of certain treatments can be mapped on brain level to gain an even more detailed insight into the particular processes that are assumed to accompany psychotherapy. More specifically, experiments on fear conditioning and particularly fear extinction, for instance, may reflect maintaining factors of symptomatology and the effects of exposure therapy on brain level in patients with anxiety disorders (e.g., Kircher et al., 2013).

All in all, the selection of methods and experimental design determines what informative value can be drawn from imaging studies on the treatment effects of psychotherapy. The planning and realization of such studies is accompanied by considerable costs and efforts. On the other hand, this kind of research offers an invaluable opportunity to monitor how psychotherapy, and thus cognitive and emotional restructuring of information processes, affect the brain in a limited time interval. Furthermore, research on psychotherapy may provide insight into the supporting and debilitating neural factors impacting on treatment responses and as such therapy success.

Status Quo

The first studies that compared pre- and posttreatment changes in cerebral metabolism date back to the early 1990s. Since then, most research focused mainly on MDD and anxiety disorders with the largest number of studies investigating patients with obsessive–compulsive disorder (OCD; **Figure 1**). Only a limited amount of data are available for patients suffering from schizophrenia or eating disorders and there are only single studies on other disorders (e.g., chronic fatigue syndrome: de Lange et al., 2008; or alcohol dependence: Feldstein Ewing et al., 2011). In the following, two lines of research are presented to exemplify how imaging techniques help to understand the impact of psychotherapy on the neurobiology of psychopathology, such as depression or anxiety. An overview of the amount of studies on various forms of psychotherapy and with respect to the type of imaging method is displayed in **Figure 1**. More specific information about existing studies, their experimental designs, and findings is presented in **Table 4**.

Major Depressive Disorder

Studies on psychotherapy effects on brain activation in MDD patients observed functional alterations or overall metabolic changes primarily in prefrontal cortical regions. Following psychotherapy, activation decreases were found in widespread prefrontal areas including dorsolateral, ventral, orbitofrontal, and medial PFC (Buchheim et al., 2012; Goldapple et al., 2004). Furthermore, alterations were consistently observed

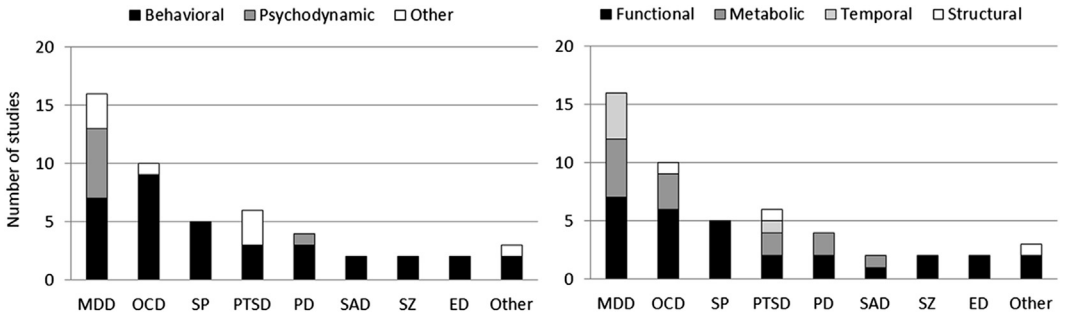

Figure 1 Number of imaging studies with respect to type of psychotherapy and imaging technique. Both figures indicate the number of imaging studies on the treatment of different psychopathological conditions. The left panel shows how these studies divide into those that included a behaviorally oriented psychotherapeutic treatment and those which applied a rather psychodynamically oriented form of therapy. Behavioral treatment mainly referred to (cognitive-) behavior therapy or related techniques, psychodynamic theraphy mostly referred to interpersonal therapy. The right panel displays the use of different imaging techniques in relation to the disorder of interest. Functional methods referred exclusively to functional magnetic resonance imaging, metabolic techniques included positron-emission tomography and single-photon emission tomography, temporal methods referred to electroencephalography and magnetoencephalography, and structural methods included magnetic resonance imaging and computed tomography. Studies were selected based on a literature search in relevant databases by using combinations of keywords including a specific disorder and '(psycho-) therapy,' 'CBT,' 'behavior therapy,' or 'psychodynamic.' ED, eating disorder; MDD, major depressive disorder; OCD, obsessive-compulsive disorder; PD, panic disorder; PTSD, posttraumatic stress disorder; SAD, social anxiety disorder; SP, specific phobia; SZ, schizophrenia; Other: alcohol dependence, chronic fatigue syndrome, generalized anxiety disorder.

Table 4 Overview of experimental designs and main findings of imaging studies on anxiety disorders, depression, and schizophrenia

	Anxiety disorders	*Depression*	*Other*
Disorders or predominant symptomatology	OCD SP PTSD PD SAD GAD	MDD Depressive episode	Schizophrenia Eating disorders Chronic fatigue Alcohol dependence[a]
Type of task	None (structural) Resting state Symptom provocation Exposure Emotional processing Fear conditioning Memory Subliminal processing Executive functions	Resting state Sleep Emotional processing Executive functions	None (structural) Self-perception Executive functions Working memory
Posttherapy (main findings)	Decreased: limbic system, OFC and caudate nucleus (OCD) Increased: PFC, connectivity between PFC and limbic and paralimbic areas	Activation changes: limbic system, PFC (ACC, mPFC, and DLPFC)	Increased: PFC (schizophrenia), PFC volume (chronic fatigue) Activation changes: extrastriate body area (eating disorder)
Predictors	OFC (OCD), visual cortex (SAD), ACC volume (PTSD), superior temporal and superior/middle frontal gyrus (PTSD)	vmPFC, DLPFC, anterior temporal lobe, abnormal EEG sleep profile	Schizophrenia: DLPFC and DLPFC-cerebellum connectivity, PFC, thalamus, precuneus

ACC, anterior cingulate cortex; DLPFC, dorsolateral prefrontal cortex; ED, eating disorder; EEG, electroencephalography; GAD, generalized anxiety disorder; MDD, major depressive disorder; mPFC, medial prefrontal cortex; OCD, obsessive-compulsive disorder; OFC, orbitofrontal cortex; PD, panic disorder; PFC, prefrontal cortex; PTSD, posttraumatic stress disorder; SAD, social anxiety disorder; SP, specific phobia; vmPFC, ventromedial prefrontal cortex.
[a]This study examined situational aspects of motivational interviewing and not directly therapy-related neural changes. Therefore, results are not included.

across studies in cingulate cortex or neighboring regions such as subgenual or para-cingulate cortex (Martin et al., 2001; Siegle et al., 2006; Dichter et al., 2010). Neural treatment responses were also found to vary between psychotherapeutic and psychopharmacological interventions. Whereas both treatments caused increased basal ganglia activity, IPT affected also cingulate areas while antidepressant medication led to activation increases in temporal areas (Martin et al., 2001). Other studies found diverging effects between pharmacotherapy and psychotherapy within medial PFC or laterality of effects (Brody et al., 2001; Goldapple et al., 2004). Apart from resting state, clinical improvement following psychotherapy was associated with functional changes in the respective prefrontal regions and decreases in amygdalar and hippocampal activation (Buchheim et al., 2012; Dichter et al., 2010; Fu et al., 2008).

Anxiety Disorders

Anxiety disorders are particularly suited to investigate therapy-induced changes in information processing of anxiety-related stimuli because symptoms can be elicited relatively easily by exposing the patient to fear-provoking stimuli. Confronting spider phobics with videos sequences of spiders, for instance, led to profound activation of the fear network in comparison to healthy controls. Following CBT, activation in those areas (i.e., anterior cingulate cortex (ACC) and insula) decreased profoundly in contrast to patients assigned to a waiting-list group (**Figure 2**; Straube et al., 2006). Because exposure or systemic desensitization are core CBT techniques for many anxiety disorders, symptom provocation studies exposing subjects either directly or indirectly to fear-provoking stimuli allow for a deeper view on the underlying neural processes of abnormal information processing which are present before treatment and how these processes are affected by psychotherapy. Depending on the particular disorder and symptomatology at hand, the degree of exposure between such studies varied from exposure *in sensu* (i.e., imagery or recall of traumatic events in PTSD; Levin et al., 1999; Lindauer et al., 2008) to exposure *in vivo* (i.e., public speaking task for patients with social anxiety disorder; Furmark et al., 2002). To date, the number of imaging studies on psychotherapy effects on anxiety disorders clearly outweighs those focusing on other disorders such as mood or eating disorders (**Figure 1**). Here, OCD has been the most extensively studied disorder followed by studies on PTSD and specific phobia, whereas only few studies investigated panic disorder and social anxiety disorder. As yet, there is only a singly pediatric study on generalized anxiety disorder (GAD; McClure et al., 2007).

With respect to the mechanisms of action, many studies on OCD found activation changes toward a 'normalized' level following psychotherapy in the orbitofrontal cortex (OFC) and striatum (Freyer et al., 2011; Nabeyama et al., 2008). Particularly, activation in the caudate nucleus decreased after successful treatment and the degree of change in both OFC and caudate nucleus was related to the treatment response (e.g., Brody et al., 1998; Nakatani et al., 2003). Overall, functional and resting state activation of the frontostriatal

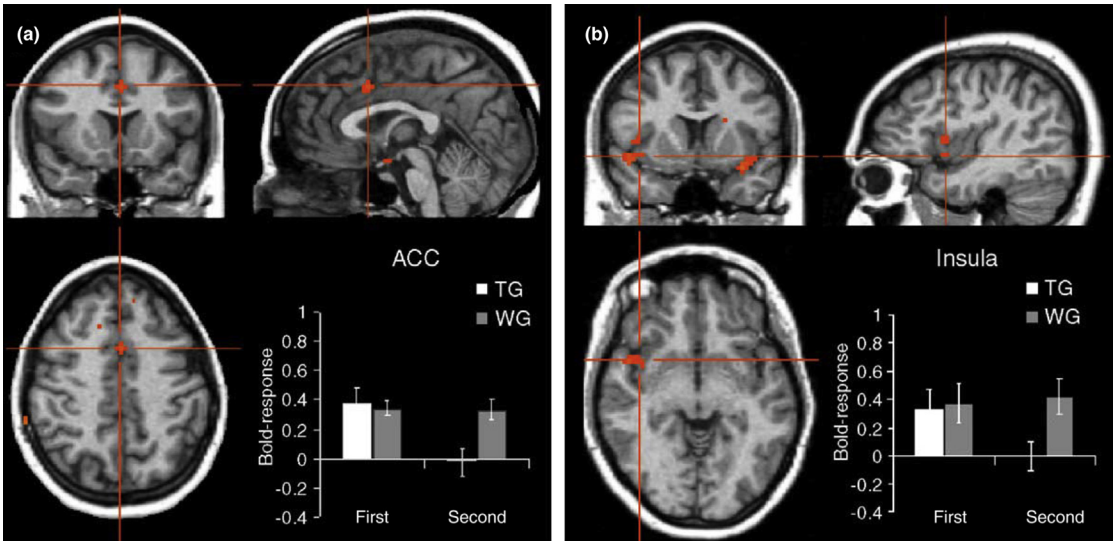

Figure 2 Psychotherapy effects on anterior cingulate cortex (ACC) and insula in a group of spider phobics. The figure displays activation in anterior cingulate cortex (ACC; left panel) and insula (right panel) to phobic stimuli in spider phobics before and after cognitive-behavioral theraphy. Compared to the waiting-list group (WG) and compared to the first scanning session, the therapy group (TG) showed significant reductions in activation in both areas following treatment during the second scanning session. The figure has originally been published in a functional magnetic resonance imaging study performed by Straube, T., Glauer, M., Dilger, S., Mentzel, H.J., Miltner, W.H., 2006. Effects of cognitive-behavioral therapy on brain activation in specific phobia. Neuroimage 29, 125–135 and is reproduced with kind permission from Elsevier (© 2006 Elsevier Inc. All rights reserved).

network was found to be attenuated following behavioral therapy (Baioui et al., 2013).

Overall, anxiety disorders that are primarily associated with an overactivation of the fear network, mainly of the amygdala, were found to exhibit decreased recruitment of those areas following psychotherapy (e.g., Furmark et al., 2002). Moreover, after therapy, regulatory prefrontal regions such as the OFC and ACC were activated more strongly (Schienle et al., 2009; Bryant et al., 2008b; Levin et al., 1999). In contrast, PTSD patients who responded poorly to CBT displayed greater amygdala activation to fearful faces at baseline (Bryant et al., 2008a). Also, functional connectivity within the fear network and particularly fear-regulating prefrontal regions increased after therapy (Kircher et al., 2013).

Taken together, it was shown that psychotherapy affects neural functioning in anxiety disorders through activation of all three suggested mechanisms of action. Following therapy, responders displayed attenuated activity in formerly hypersensitive limbic regions (normalization), additional recruitment of regulatory prefrontal areas (compensation), and increased overall connectivity between both limbic and top-down brain regions (functional connectivity).

Past Challenges and Future Directions

Conducting imaging studies along the course of psychotherapeutic treatment definitely poses a challenge with respect to costs and efforts. The combination of repeated measures designs and clinical samples is relatively time consuming and the high risk of drop-outs during treatment represents an additional challenge. In addition, the selection of a proper control group can be difficult due to ethical constraints. Randomly assigning patients to a waiting list in order to control for treatment effects clearly represents a drawback for those patients and holds the risk of a potential aggravation of symptoms. On the other hand, in clinical practice, patients are often put on a waiting list until they can start psychotherapeutic treatment and this waiting time most often exceeds that of a controlled clinical treatment study. Depending on the disorder and severity of symptoms, however, this proceeding might be inappropriate thereby probably excluding those subjects with the most distinct symptoms in experimental designs. For disorders that are accompanied by such severe symptoms that do not allow for the postponement of treatment (e.g., severe forms of MDD) the comparison of two or more evidence-based psychotherapies represents a valuable alternative. Nevertheless, the informative value of studies with a waiting list design definitively warrants their costs and efforts in contrast to those including a nonclinical control group.

The results of previous studies on the neural correlates of psychotherapy have shown the diversity of mechanisms through which neural functioning is affected by psychotherapeutic treatment of anxiety disorders. Similar widespread effects are conceivable for other disorders and highlight the need for further and extended research on disorders that particularly benefit from psychotherapy such as mood and eating disorders. Studies with larger sample sizes, sophisticated experimental designs, and high levels of standardization may help to better understand the functional and structural neural processes that accompany positive treatment outcomes. A recent study on panic disorder refined these experimental demands by the inclusion of a very large sample size in comparison to earlier studies, manualized treatment, and

functional imaging of specific impaired aspects of information processing (Kircher et al., 2013). In addition to such manual-based treatment, group therapy may further help to increase within-subjects comparability and to lower costs and efforts of future research.

Despite the substantial amount of existing studies, there are still a lot of opportunities for future studies particularly regarding methods and clinical populations. Apart from the lack of studies on disorders other than MDD and anxiety disorders, only single studies (e.g., Huyser et al., 2011; McClure et al., 2007) focused on children or adolescents although particularly the younger brain is much more plastic due to developmental processes. Research methods which can be performed in a more natural setting and are free of side effects (e.g., EEG) might help to investigate such potentially more sensitive subjects. Overall, electrophysiological research in this field is scarce but may provide much more detailed insight into the temporal characteristics of how information processing is affected by psychotherapy. The combined use of methods with high temporal and high spatial resolution is even more tempting and the ongoing refinement of analyzing techniques may stimulate future research.

Because only a portion of the existing studies investigated functional activation in response to specific tasks and stimuli, much more research is needed to elucidate how various aspects of disorder-related stimulus processing such as explicit vs implicit or subliminal vs supraliminal emotional processing are modulated over the course of therapy. Furthermore, symptom provocation studies (e.g., exposure or anticipation of disorder-related stimuli and symptoms) provide direct information about the neural correlates that accompany clinical improvement. New research and treatment methods such as exposure by means of a virtual reality environment even allow for online neurophysiological measures during single therapy sessions (Wiederhold and Wiederhold, 2008). In the long run, follow-up studies may then test whether the degree of activation changes in certain areas predict better clinical outcome.

Overall, the investigation of psychotherapy-related alterations in brain activation still offers a multitude of open questions to answer. Future research should therefore particularly aim to disentangle the neural factors that predict the treatment response to individual therapies and clinical outcome. These studies may in turn help to select the appropriate therapy and number of sessions to meet the patient's individual needs, thereby minimizing the risk of drop-outs and later relapses.

Conclusion

Thus far, existing imaging studies have shown that psychotherapy significantly affects neural and metabolic functioning in a range of mental disorders with most evidence for MDD and anxiety disorders. Particularly, the recovery from anxiety-related psychopathology has been extensively studied by means of different experimental designs and imaging techniques, demonstrating that clinical improvement is associated with several neural processes. The potential mechanisms of action that have been identified so far include a normalization of activity, compensatory recruitment of regulatory brain regions, and increased connectivity within disorder-related neural networks, but also more general processes such as neuroplasticity are strongly linked to the fundamental ideas of particular therapies such as CBT.

Generally, it is important to keep in mind that the neural correlates of psychotherapy are not directly measurable since treatment and clinical improvement represent processes that take more time than a single experimental measurement session. Pre- and postmeasures of single therapeutic sessions or even an entire therapy can only provide insight into the neural changes that accompany the multitude of mechanisms that are activated by psychotherapy. These mechanisms are in turn dependent on the type of therapy and may not be uniform across patients and disorders. Imaging studies on psychotherapeutic effects, whether in general or focusing on specific therapeutic strategies, however, can provide valuable information about the underlying neural substrates of psychopathology and how these substrates are changing from acute illness to complete recovery of function. More detailed insight into the effects of particular therapeutic strategies can be obtained from experimental designs that aim to mirror the psychological processes that occur during single therapeutic sessions (e.g., experiments on extinction learning for the investigation of neural processes that take place during exposure sessions).

Although the number of studies in the field is still relatively small, previous research has given a promising impression of the numerous possibilities with respect to experimental design, hopefully stimulating future research. In the long run, such research may lead to individualized treatment protocols and better prevention of nonresponses, dropouts, and remission of symptoms.

See also: Anxiety and Anxiety Disorders; Behavior Therapy: Background, Basic Principles, and Early History; Behavior Therapy: Psychiatric Aspects; Evidence-Based Psychiatry; Evidence-Based Psychotherapy; Imaging Methods in Clinical Psychology; Interpersonal Psychotherapy; Magnetic Resonance and Molecular Imaging in Psychiatry; Manual Based Treatments; Psychodynamic Therapy: Short-term.

Bibliography

Adenauer, H., Catani, C., Gola, H., Keil, J., Ruf, M., Schauer, M., Neuner, F., 2011. Narrative exposure therapy for PTSD increases top-down processing of aversive stimuli - evidence from a randomized controlled treatment trial. BMC Neuroscience 12, 127.

Baioui, A., Pilgramm, J., Kagerer, S., Walter, B., Vaitl, D., Stark, R., 2013. Neural correlates of symptom reduction after CBT in obsessive-compulsive washers—An fMRI symptom provocation study. Journal of Obsessive-Compulsive and Related Disorders 2, 322–330.

Bandelow, B., Seidler-Brandler, U., Becker, A., Wedekind, D., Rüther, E., 2007. Meta-analysis of randomized controlled comparisons of psychopharmacological and psychological treatments for anxiety disorders. The World Journal of Biological Psychiatry 8, 175–187.

Brody, A.L., Saxena, S., Schwartz, J.M., Stoessel, P.W., Maidment, K., Phelps, M.E., Baxter Jr., L.R., 1998. FDG-PET predictors of response to behavioral therapy and pharmacotherapy in obsessive compulsive disorder. Psychiatry Research 84, 1–6.

Brody, A.L., Saxena, S., Stoessel, P., Gillies, L.A., Fairbanks, L.A., Alborzian, S., Phelps, M.E., Huang, S.-C., Wu, H.-M., Ho, M.L., Ho, M.K., Au, S.C., Maidment, K., Baxter, L.R., 2001. Regional brain metabolic changes in patients

with major depression treated with either paroxetine or interpersonal therapy: preliminary findings. Archives of General Psychiatry 58, 631–640.

Bryant, R.A., Felmingham, K., Kemp, A., Das, P., Hughes, G., Peduto, A., Williams, L., 2008a. Amygdala and ventral anterior cingulate activation predicts treatment response to cognitive behaviour therapy for post-traumatic stress disorder. Psychological Medicine 38, 555–561.

Bryant, R.A., Felmingham, K., Whitford, T.J., Kemp, A., Hughes, G., Peduto, A., Williams, L.M., 2008b. Rostral anterior cingulate volume predicts treatment response to cognitive-behavioural therapy for posttraumatic stress disorder. Journal of Psychiatry & Neuroscience 33, 142–146.

Buchheim, A., Viviani, R., Kessler, H., Kächele, H., Cierpka, M., Roth, G., George, C., Kernberg, O.F., Bruns, G., Taubner, S., 2012. Changes in prefrontal-limbic function in major depression after 15 months of long-term psychotherapy. PLoS One 7, e33745.

Carrig, M.M., Kolden, G.G., Strauman, T.J., 2009. Using functional magnetic resonance imaging in psychotherapy research: a brief introduction to concepts, methods, and task selection. Psychotherapy Research 19, 409–417.

de Lange, F.P., Koers, A., Kalkman, J.S., Bleijenberg, G., Hagoort, P., van der Meer, J.W.M., Toni, I., 2008. Increase in prefrontal cortical volume following cognitive behavioural therapy in patients with chronic fatigue syndrome. Brain 131, 2172–2180.

Dichter, G.S., Felder, J.N., Smoski, M.J., 2010. The effects of brief behavioral activation therapy for depression on cognitive control in affective contexts: an fMRI investigation. Journal of Affective Disorders 126, 236–244.

Etkin, A., Phil, M., Pittenger, C., Polan, H.J., Kandel, E.R., 2005. Toward a neurobiology of psychotherapy: basic science and clinical applications. The Journal of Neuropsychiatry and Clinical Neurosciences 17, 145–158.

Etkin, A., Prater, K.E., Schatzberg, A.F., Menon, V., Greicius, M.D., 2009. Disrupted amygdalar subregion functional connectivity and evidence of a compensatory network in generalized anxiety disorder. Archives of General Psychiatry 66, 1361–1372.

Etkin, A., Wager, T.D., 2007. Functional neuroimaging of anxiety: a meta-analyses of emotional processing in PTSD, social anxiety disorder, and specific phobia. The American Journal of Psychiatry 164, 1476–1488.

Feldstein Ewing, S.W., Filbey, F.M., Sabbineni, A., Chandler, L.D., Hutchison, K.E., 2011. How psychosocial alcohol interventions work: a preliminary look at what-fMRI can tell us. Alcoholism, Clinical and Experimental Research 35, 643–651.

Freyer, T., Klöppel, S., Tüscher, O., Kordon, A., Zurowski, B., Kuelz, A.K., Speck, O., Glauche, V., Voderholzer, U., 2011. Frontostriatal activation in patients with obsessive–compulsive disorder before and after cognitive behavioral therapy. Psychological Medicine 41, 207–216.

Fu, C.H.Y., William, S.C., Cleare, A.J., Scott, J., Mitterschiffthaler, M.T., Walsh, N.D., Donaldson, C., Suckling, J., Andrew, C., Steiner, H., Murray, R.M., 2008. Neural responses to sad facial expressions in major depression following cognitive behavioral therapy. Biological Psychiatry 64, 505–512.

Furmark, T., Tillfors, M., Marteinsdottir, I., Fischer, H., Pissiota, A., Långström, B., Fredrikson, M., 2002. Common changes in cerebral blood flow in patients with social phobia treated with citalopram or cognitive-behavioral therapy. Archives of General Psychiatry 59, 425–433.

Gillin, J.C., Buchsbaum, M., Wu, J., Clark, C., Bunney, W., 2001. Sleep deprivation as a model experimental antidepressant treatment: findings from functional brain imaging. Depression and Anxiety 14, 37–49.

Goldapple, K., Segal, Z., Garson, C., Lau, M., Bieling, P., Kennedy, S., Mayberg, H., 2004. Modulation of cortical-limbic pathways in major depression: treatment-specific effects of cognitive behavior therapy. Archives of General Psychiatry 61, 34–41.

Huyser, C., Veltman, D.J., Wolters, L.H., de Haan, E., Boer, F., 2011. Developmental aspects of error and high-conflict-related brain activity in pediatric obsessive–compulsive disorder: a fMRI study with a Flanker task before and after CBT. Journal of Child Psychology and Psychiatry 52, 1251–1260.

Kim, M.J., Loucks, R.A., Palmer, A.L., Brown, A.C., Solomon, K.M., Marchante, A.N., Whalen, P.J., 2011. The structural and functional connectivity of the amygdala: from normal emotion to pathological anxiety. Behavioural Brain Research 223, 403–410.

Kircher, T., Arolt, V., Jansen, A., Pyka, M., Reinhardt, I., Kellermann, T., Konrad, C., Lueken, U., Gloster, A.T., Gerlach, A.L., Ströhle, A., Wittmann, A., Pfleiderer, B., Wittchen, H.-U., Straube, B., 2013. Effect of cognitive-behavioral therapy on neural correlates of fear conditioning in panic disorder. Biological Psychiatry 73, 93–101.

Levin, P., Lazrove, S., van der Kolk, B., 1999. What psychological testing and neuroimaging tell us about the treatment of posttraumatic stress disorder by eye movement desensitization and reprocessing. Journal of Anxiety Disorders 13, 159–172.

Lindauer, R.J.L., Booij, J., Habraken, J.B.A., van Meijel, E.P.M., Uylings, H.B.M., Olff, M., Calier, I.V.E., den Heeten, G.J., van Eck-Smit, B.L.F., Gersons, B.P.R., 2008. Effects of Psychotherapy on Regional Cerebral Blood Flow during Trauma Imagery in Patients with Post-traumatic Stress Disorder: A Randomized Clinical Trial. Cambridge University Press, Cambridge, ROYAUME-UNI.

Linden, D.E.J., 2008. Brain imaging and psychotherapy: methodological considerations and practical implications. European Archives of Psychiatry and Clinical Neuroscience 258, 71–75.

Martin, S.D., Martin, E., Rai, S.S., Richardson, M.A., Royall, R., 2001. Brain blood flow changes in depressed patients treated with interpersonal psychotherapy or venlafaxine hydrochloride: preliminary findings. Archives of General Psychiatry 58, 641–648.

McClure, E., Adler, A., Monk, C., Cameron, J., Smith, S., Nelson, E., Leibenluft, E., Ernst, M., Pine, D., 2007. fMRI predictors of treatment outcome in pediatric anxiety disorders. Psychopharmacology 191, 97–105.

Nabeyama, M., Nakagawa, A., Yoshiura, T., Nakao, T., Nakatani, E., Togao, O., Yoshizato, C., Yoshioka, K., Tomita, M., Kanba, S., 2008. Functional MRI study of brain activation alterations in patients with obsessive–compulsive disorder after symptom improvement. Psychiatry Research 163, 236–247.

Nakatani, E., Nakgawa, A., Ohara, Y., Goto, S., Uozumi, N., Iwakiri, M., Yamamoto, Y., Motomura, K., Iikura, Y., Yamagami, T., 2003. Effects of behavior therapy on regional cerebral blood flow in obsessive–compulsive disorder. Psychiatry Research 124, 113–120.

Rauch, S.L., Shin, L.M., Wright, C.I., 2003. Neuroimaging studies of amygdala function in anxiety disorders. Annals of the New York Academy of Sciences 985, 389–410.

Roffman, J.L., Marci, C.D., Glick, D.M., Dougherty, D.D., Rauch, S.L., 2005. Neuroimaging and the functional neuroanatomy of psychotherapy. Psychological Medicine 35, 1385–1398.

Schienle, A., Schäfer, A., Stark, R., Vaitl, D., 2009. Long-term effects of cognitive behavior therapy on brain activation in spider phobia. Psychiatry Research 172, 99–102.

Schiepek, G., Tominschek, I., Karch, S., Lutz, J., Mulert, C., Meindl, T., Pogarell, O., 2009. A controlled single case study with repeated fMRI measurements during the treatment of a patient with obsessive-compulsive disorder: testing the nonlinear dynamics approach to psychotherapy. The World Journal of Biological Psychiatry 10, 658–668.

Siegle, G.J., Carter, C.S., Thase, M.E., 2006. Use of fMRI to predict recovery from unipolar depression with cognitive behavior therapy. The American Journal of Psychiatry 163, 735–738.

Straube, T., Glauer, M., Dilger, S., Mentzel, H.J., Miltner, W.H., 2006. Effects of cognitive-behavioral therapy on brain activation in specific phobia. Neuroimage 29, 125–135.

Straube, T., Miltner, W.H.R., 2011. Ängste und Phobien. In: Schiepek, G. (Ed.), Neurobiologie der Psychotherapie. Schattauer GmbH, Stuttgart, Germany, pp. 385–397.

Thase, M.E., Buysse, D.J., Frank, E., Cherry, C.R., Cornes, C.L., Mallinger, A.G., Kupfer, D.J., 1997. Which depressed patients will respond to interpersonal psychotherapy? The role of abnormal EEG sleep profiles. The American Journal of Psychiatry 154, 502–509.

Wiederhold, B., Wiederhold, M., 2008. Virtual reality with fMRI: a breakthrough cognitive treatment tool. Virtual Reality 12, 259–267.

Wykes, T., Brammer, M., Mellers, J., Bray, P., Reeder, C., Williams, C., Cornder, J., 2002. Effects on the brain of a psychological treatment: cognitive remediation therapy: functional magnetic resonance imaging in schizophrenia. The British Journal of Psychiatry 181, 144–152.

Imitation during Infancy and Early Childhood

Rebecca A Williamson, Georgia State University, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Imitation is an efficient way for infants and young children to benefit from others' knowledge. By copying observed acts, children learn novel behaviors for interacting with their environment and they establish interactions with social partners. This article provides an introduction to the development of imitation by reviewing what components of behaviors children are able to imitate at different ages. Then, this article explains how functional and social factors regulate when children take up others' examples through imitation. Throughout, theoretical and practical examples illustrate the interdisciplinary nature of imitation research.

Others' acts are a repository of accumulated human knowledge. Consider watching someone preserve food after a harvest. The canner will examine each vegetable for size and ripeness, use tools and specific behaviors to chop them, and measure and combine ingredients to achieve a desired taste. Carefully copying each behavior can have multiple benefits. For example, copying how the canner boils glass jars will reduce the risk of food contamination and of getting burned. Observational learning and imitation specifically saves a learner from the painstaking (and sometimes dangerous) process of trial and error learning (Bandura, 1977). It allows a learner to take advantage of the knowledge hard-won by past generations.

This article provides an overview of research on young children's imitation, i.e., copying others' acts. I begin with a description of the development of imitation through infancy and into the toddler and preschool years. This article highlights formative findings in research on social learning and provides a timeline of developments in what children imitate from a behavior. The second section describes the flexibility of imitation with a review of the functional and social considerations that influence children's learning.

Throughout this article, I provide implications of imitation research, with the goal of illustrating the breadth of this field. Imitation is readily observable, making it an important research tool. Studies of imitation contribute to basic theoretical debates, such as understanding what is unique about human learning. This basic research also has practical implications, including identifying efficient educational and clinical practices.

Imitating Components of Action

Neonatal Imitation

From their first days, infants reproduce simple body acts. In a seminal study, neonates (aged 12–21 days) saw an adult produce a distinctive movement (Meltzoff and Moore, 1977), such as a tongue protrusion. Then, the adult assumed a passive expression. During this test period, infants were more likely to reproduce the body movement that the adult had most recently shown (e.g., tongue protrusion) versus other target acts (e.g., mouth opening). Children also showed more target acts after seeing the adult produce the target act versus in baseline trials where the adult only ever showed a passive face. Both of these comparisons indicate imitation.

Many subsequent studies have replicated these findings (for review, see Meltzoff and Moore, 1997). One tested infants that were less than 72 h old (Meltzoff and Moore, 1983), with children who had never looked in a mirror or seen pictures of themselves. They also had never been reinforced for copying behaviors. The fact that these infants matched the adult's facial movements suggests an innate mapping between one's own and others' bodies.

Immediate and Deferred Imitation of Acts on Objects

Over the course of the first year, infants' engagement with objects increases (Bakeman et al., 1990), and their imitation parallels these developments. From the age of 6 months, infants imitate simple acts on objects (Barr et al., 1996), such as removing and shaking a puppet's mitten. Over the next year and a half, children become better able to imitate more acts and to maintain the temporal sequence in which acts are demonstrated (Bauer and Mandler, 1989, 1992).

From 6 to 12 months of age, infants also begin to imitate acts after substantial delays. To test this *deferred imitation*, children observe a model produce target acts on objects. After a delay, children are given the objects, and their production of the target acts is scored. Control groups can include different children who do not see an adult act on the materials (a baseline control) or who see an adult manipulate the materials in a different way (an object manipulation control). Using such procedures, researchers have demonstrated that infants aged 6–12 months can remember and reproduce acts after a 24 h delay (Barr et al., 1996; Collie and Hayne, 1999). By age of 14 months, infants remember multiple novel acts after 4 months (Meltzoff, 1995a). Increasing the number of demonstrations or the amount of viewing time improves deferred imitation (e.g., Barr et al., 1996, 2007a,b).

Implications for Theory

Deferred imitation has been important for developmental theory. For one, reports of deferred imitation challenge Piaget's (1962) seminal theory of cognitive development. Because deferring imitation involves the ability to mentally represent actions, Piaget thought it should arise between the

ages of 18–24 months. The clear evidence for deferred imitation in much younger children supports proposals that Piaget underestimated infants' cognitive abilities (Meltzoff, 1988a).

Deferred imitation is also important for considering what neural processes might underlie imitation. Certain cells in the premotor cortex of monkeys have been shown to fire when an act is either observed or produced (di Pellegrino et al., 1992; Rizzolatti et al., 1996); these cells react, for example, both when a monkey reaches or sees another individual reach. Thus, these *mirror neurons* create a link between self and other in the brain. One very simple way that mirror neurons could underlie imitation is for the observation of an act to automatically elicit the same motor program and lead the observer to produce the act (Rizzolatti et al., 1999). However, the delays involved in deferred imitation studies do not support this most basic mechanism. Neuroscientists continue to generate and test more complex neural accounts for imitation (e.g., Huber et al., 2009; Marshall et al., 2011; Paulus et al., 2013; Reid et al., 2011).

Imitation of Subgoals and Means

Many studies of imitation involve causing physical changes to objects, such as turning on lights or opening boxes. This emphasis on outcomes is understandable; infants are more likely to reproduce acts that produce observable effects versus those that do not (for review, see Elsner, 2007). However, young children imitate more than physically effective acts. In one demonstration of this, 14-month-olds saw an adult turn on a light using a touch panel (Meltzoff, 1988a). Instead of reaching out with his hand, the adult bent and touched the panel with his head. When allowed to act, two-thirds of the children also imitated the unusual bending action they had seen the adult use to illuminate the light. Thus, their imitation included the demonstrated outcome as well as the adult's method for producing the result.

Implications – Theories of Human Learning

The distinction between copying outcomes and means has proven important for distinguishing human and nonhuman social learning (for review, see Want and Harris, 2002). Children readily copy the specific means that others use. In fact, they imitate even causally irrelevant acts, a phenomenon called 'overimitation' (e.g., Horner and Whiten, 2005; Lyons et al., 2007; McGuigan et al., 2011). In contrast, although folk wisdom tells us that 'monkey see, monkey do' – really, monkeys don't. Carefully controlled studies show limits in nonhuman primates' social learning relative to humans' (for review, see Whiten et al., 2004). Specifically, nonhuman primates are less likely to copy others' specific means (e.g., Nagell et al., 1993; Whiten et al., 1996).

Imitation may be crucial for the development of complex culture (Boyd and Richerson, 1996; Meltzoff, 1988b; Tomasello, 1999). One account is that imitation is a mechanism that holds behaviors in place within a community (Tomasello, 1999). The idea is that individuals will sometimes innovate; an inventor might devise a new food processing technique. With imitation, that innovation will spread quickly to other members of the group. Thus, if the inventor dies, the innovation is not lost to the community. Imitation should also improve progress within a group. Effective social learning will lead more individuals to use the advanced behavior, and each may enhance it with new innovations.

Peer-to-peer learning studies illustrate the power of imitation for transmitting information within groups (Flynn and Whiten, 2008; Horner et al., 2006; Nielsen et al., 2012a). One procedure, called chain diffusion, is essentially a game of telephone – information is passed through a group, with each learner becoming the teacher for the next individual. These studies show impressive imitation among peer groups, with some target acts maintained across chains of up to eight 3-year-olds (McGuigan and Graham, 2010). These studies illustrate the plausibly of imitation as a mechanism for the efficient spread of information through communities.

Imitating Inferred Goals

Both the outcomes produced and the means used to attain those outcomes are observable from others' behaviors. However, young children's imitation is not bound to what they can see. Children also infer and imitate the goals underlying behavior. In one experiment, 18-month-olds saw an adult try, but fail, to complete acts with objects (Meltzoff, 1995b). For example, the adult pulled at the ends of a barbell toy, but never separated the two pieces. When given the objects to manipulate, children who saw the failed demonstration were as likely to produce the intended outcome as were children who saw the adult complete the act. Thus, children did not imitate the observed surface behaviors; they were not limited to slipping their fingers from the toy. Instead, 18-month-olds inferred and reproduced the intentions underlying the behaviors. Even younger children imitate the goals underlying simple acts. For example, 7-month-olds will selectively reach for an object that an actor previously tried, but failed, to reach (Hamlin et al., 2008; see also Bellagamba and Tomasello, 1999).

Children's ability to infer intentions depends on the identity of the actor. Meltzoff (1995b) contrasted two different unsuccessful attempts at the barbell-pulling task. One group of children saw a human actor fail to pull apart the object, as described above. A second group saw a machine grip the barbell with two pincers. The pincers slid from the edge, to recreate the adult's unsuccessful attempts. Children who saw the adult were significantly more likely to pull apart the object than were children who saw the machine perform the carefully matched acts. Thus, children attributed intentions to the person, but not the mechanical pincers.

Modifications of this imitation paradigm have helped to identify when children classify actors as agents (with goals and intentions) versus nonagents. For one, children are unlikely to imitate the unfulfilled goals of actors that have little similarity in body morphology to humans, such as machines and boxes (Slaughter and Corbett, 2007; Johnson et al., 2001; Mahajan and Woodward, 2009). However, young children will imitate intermediate actors (e.g., puppets or robots) that have human-like physical characteristics, such as a face and arms (Itakura et al., 2008; Johnson et al., 2001). A second factor that influences the imitation of unfulfilled goals is social cues to animacy. When a human confederate engages a nonhuman actor as if it is animate (e.g., naming it, waving to it), children are more likely

to imitate it (Johnson et al., 2001). When the actor responds contingently to a human confederate, imitation is also boosted (Itakura et al., 2008). Understanding how agency is inferred contributes to our understanding of what representations underlie subsequent theory of mind development (Johnson, 2003).

Imitating Inferred Strategies

In addition to inferring the general goal underlying others' acts, children can also reproduce the *rules or strategies* that guide others' behaviors. This is a way that adults may commonly use observational learning; Studying the way a good speaker presents information can be as informative as the content of the talk. One type of abstract information that children learn through imitation is the temporal order of an action sequence (Subiaul et al., 2007a,b). Preschoolers who saw an adult press series of pictures in a particular order on a touch screen (like entering an ATM pin code) needed fewer trials to input the same sequence than did children who had to learn through trial and error. Importantly, the location of the pictures varied between each trial. This meant that the children had to learn the order in which to press the pictures, not the adult's motor movements.

Preschoolers also imitate the hierarchical organization that structures action (Flynn and Whiten, 2008; Whiten et al., 2006). In these tasks, children watched an adult assemble and use a series of keys to unlock a box and retrieve a reward. The adult either assembled and used one key before moving to the next key, or he completed one assembly step with all of the keys before moving to the next step. The children maintained the adult's hierarchical strategy when opening the box themselves; e.g., children who saw the adult take a stepwise approach also completed each step in turn.

Children also imitate abstract sorting rules from others' behaviors (Williamson et al., 2010). In one experiment, children saw an adult sort a set of objects into two piles based on their colors. Children were given a chance to sort the same objects and a second set of objects. The second set included different kinds and colors, and the adult never sorted them. This set was used to test whether children would generalize the sorting rule. Three control groups were used to pinpoint the critical elements of a demonstration for promoting sorting. In the most stringent control, the adult brought out the objects presorted by color, picked up each object, and returned each to its starting location. These behaviors were carefully matched to the experimental group, except the adult never sorted the objects.

The first finding from this experiment was that children who saw the adult's sorting behaviors were more likely to reproduce the color sorting strategy than were children in the control groups (see **Figure 1(b)**). Sorting behaviors were critical for eliciting imitation. Closely matched demonstrations that did not include sorting did not prompt children to imitate. A second finding was that children who saw the adult's sorting behaviors applied the color sorting rule to the second set of object, even though they were not shown what to do with those objects. This illustrates an important benefit of imitating abstract strategies and rules. Unlike specific behaviors on objects, rules are generalizable across materials and situations.

Interim Summary

During the first few years of life, children's imitation advances in complexity. From reproducing body movements and acts with objects, children progress to copying more complex sequences of acts and specific means used to attain outcomes. Children also infer and imitate the goals and rules that guide overt behaviors. Examinations of what children can imitate have contributed to our understanding of how human social learning compares and contrasts to the social learning of other primate species. These examinations have also helped chart the development of memory (e.g., Bauer et al., 2000) and early theory of mind (Carpenter et al., 1998; Meltzoff, 1995a). With this developmental timeline as a foundation, this article now turns to what regulates children's imitation.

Regulating Imitation

Functional Considerations

Although children can reproduce many elements of a demonstration, they are not bound to imitate. A substantial body of research indicates that children use imitation flexibly. Uzgiris

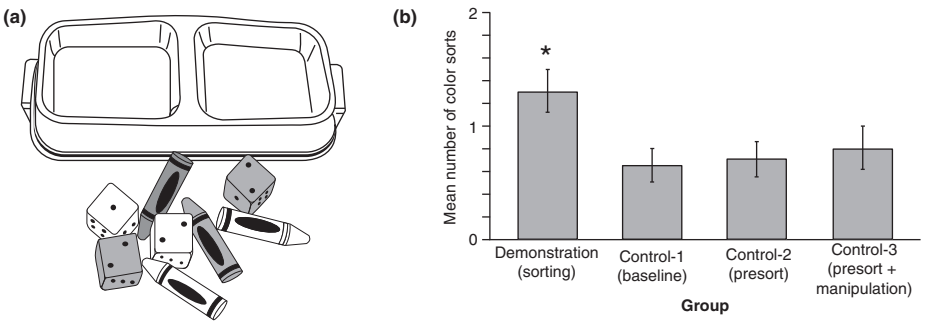

Figure 1 Imitation of an abstract rule. (a) One set of materials used to assess children's imitation of a sorting strategy. (b) 36-month-old children showed elevated rates of sorting by color only after witnessing an adult's sorting demonstration. From Williamson, R.A., Jaswal, V.K., Meltzoff, A.N., 2010. Learning the rules: observation and imitation of a sorting strategy by 36-month-old children. Developmental Psychology 46 (1), 57–65, reprinted with permission.

(1981) thought one reason to imitate is to learn how to do new things. If so, an efficient imitation strategy may be to reproduce the overall outcome of a demonstration, but to exclude acts that may reflect personal style (Williamson and Markman, 2006). In contrast, if the purpose of a behavior is not clear, imitating in precise detail may be beneficial; copying even subtle means will improve the likelihood that the imitator will reproduce the demonstration's critical elements.

Several experiments show that when acts lead to a clear outcome, children are less likely to copy others' precise means (Bekkering et al., 2000; Carpenter et al., 2005). In one, 3-year-olds saw a model place an object on a mat using a distinctive series of acts (Williamson and Markman, 2006). For example, she used two fingers to place a penlight horizontally, and then rotated the light so that it was oriented vertically relative to the child. One group saw only this unusual behavior. In a second group, the adult provided a reason for the placement of the light. After using the unusual acts to place the object, she added extra pieces so the penlight formed the stem of a flower. Those children who saw the extra pieces reproduced fewer of the adult's distinctive acts (e.g., the two-fingered grip, the rotating) relative to children who did not have this context provided.

Another factor that influences children's imitation is the efficacy of the means leading to specific ends. When means are shown to be ineffective, children are less likely to imitate that strategy (Nielsen, 2006; Want and Harris, 2001). Additionally, when children themselves have difficulty attaining an end using their own approach to a task, they are more likely to precisely imitate the means another used to complete the task (Pinkham and Jaswal, 2011; Harnick, 1978; Schulz et al., 2008). In one demonstration, 36-month-old children were given experience with a straightforward task – for example, they were told to open a drawer to retrieve a toy (Williamson et al., 2008). In one group, children could simply pull open the drawer. For the other group, however, the task was rigged to be difficult to complete. (A brace of putty was inserted so that more force was needed to open the drawer.) After this initial self-experience, children saw an adult perform a novel act before taking a turn at the task. For example, the adult pressed a button before opening an identical drawer.

For the test phase, children had another chance to complete the task. The critical question was whether children's own prior experience would influence their imitation on this final attempt. In contrast to children who had an easy experience with the task, children who had a hard time were almost twice as likely to take up the adult's example and perform the novel act; Those children who struggled to open the first drawer were more likely to push the button before trying to open the final drawer. A variation on this procedure indicated that preschoolers also take others' efficacy with a task into account when imitating (Williamson and Meltzoff, 2011).

Implications for Parenting and Education

Studying imitation in controlled contexts can provide insights about children's learning processes more generally. Presenting children with an easy problem may encourage them to explore and try different approaches for solving it. In contrast, letting children struggle with a difficult problem may help them realize that they need help. This may make them more likely to take up another's example. An additional finding from these studies is encouraging to me, as a parent and an educator. Children are usually given only minimal instruction during the test period of imitation tasks. For example, they might be told, 'Now it is your turn,' and given the materials. However, in several of these studies children often reproduced more of a model's demonstration when they were explicitly instructed to imitate ('Do what I did'). Thus, a lack of imitation does not necessarily indicate a lack of learning.

Social Considerations

Although considering imitation as a functional process (as a way to acquire new skills and behaviors) is a useful framework, it does not explain all of the variation in children's imitation. As highlighted by Uzgiris (1981), imitation is an inherently social process. Reproducing others' acts is a way for young children with limited language skills to engage in rudimentary social exchanges with peers (Nadel, 2002). Copying others' acts also highlights similarities between individuals. Thus, imitation is also a way to conform or affiliate with others. Indeed, when 5-year-olds were primed to feel excluded from a social group, they showed increased imitation relative to children who did not feel excluded (Over and Carpenter, 2009). Considering the social situation helps to explain when and why children imitate.

Opportunities for imitation are often embedded in a social interaction. If an adult does not verbally engage or make direct eye contact, children are less likely to imitate (Nielsen, 2006; Brugger et al., 2007). Additionally, although infants and young children do imitate novel acts from behaviors they see others perform on TV, live demonstrations are often more effective than are televised ones (e.g., Barr and Hayne, 1999; Hayne et al., 2003; Meltzoff, 1988c; Nielsen et al., 2008). This *video deficit effect* is likely partially explained by difficulties transferring 2D representations to 3D objects (for review, see Barr, 2010). However, studies that manipulate the contingency of the filmed actor indicate that reduced social interaction matters too. When a televised actor reacted to children's behaviors via a closed-loop TV system, 24-month-olds showed improved imitation relative to when they saw the same acts presented by a traditional, passive television display (Nielsen et al., 2008).

Within a social context, children respond to cues provided by the model as to the importance or intentionality of a behavior. For example, when toddlers were shown two acts leading to an outcome, one overtly marked as intentional ('There!') and one marked as accidental ('Woops!'), children were more likely to imitate the one that appeared purposeful (Carpenter et al., 1998). Children's imitation is also influenced by a specific set of pedagogical cues that parents spontaneously employ when teaching their children novel behaviors (Williamson and Brand, 2014). The modifications parents make, called 'motionese,' include closer proximity to the partner, use of a larger range of motion, more repetition, and more gazes to the partner (Brand et al., 2002, 2007). In one experiment, 2-year-olds who saw research assistants demonstrate a series of novel acts on four objects using motionese produced significantly more of the target behaviors than did children who saw an unembellished adult-directed action style (see **Figure 2**).

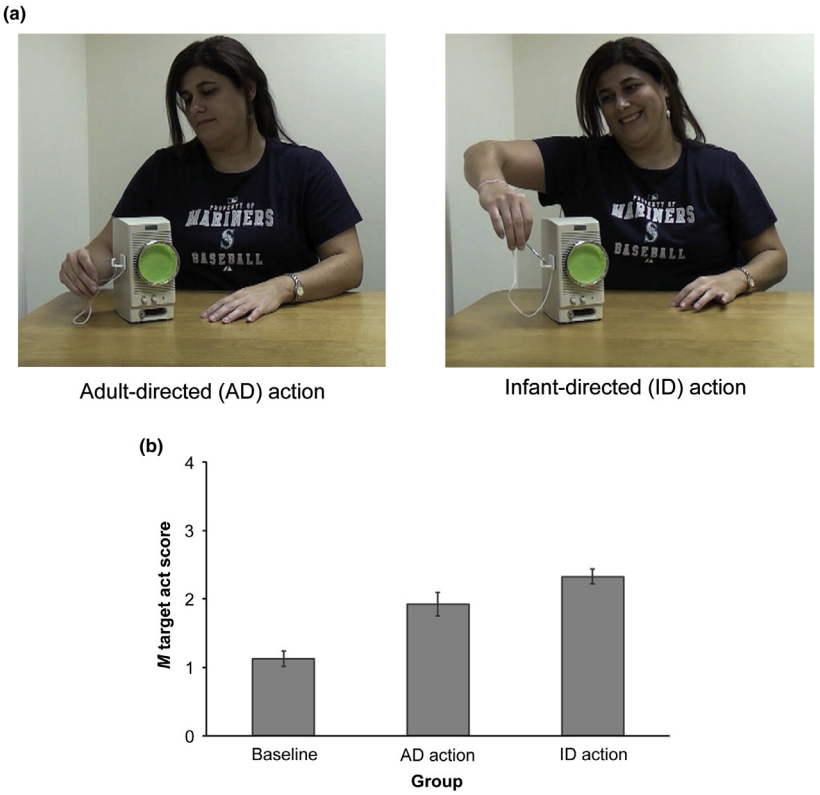

Figure 2 Motionese influences imitation. (a) Examples of the adult-directed and motionese (infant-directed) action style used in demonstrations to children. (b) Children who saw the embellished infant-directed action imitated more than did children who saw the more subtle adult-directed action, who in turn produced more target acts than children who did not see a demonstration (baseline). From Williamson, R.A., Brand, R.J., 2014. Child-directed action promotes 2-year-olds' imitation. Journal of Experimental Child Psychology 118, 119–126, reprinted with permission.

There are several ways that motionese may enhance children's imitation. For one, the cues used in motionese may help establish what Csibra and Gergely (2006) call a pedagogical context. By this account, humans have evolved to give and respond to cues (such as eye gaze) that indicate what information is new and should be learned. In addition to establishing a pedagogical context, motionese may draw children's attention and may help children identify important parts of acts. It should be noted, however, that this rich pedagogical information is not necessary for children to imitate. In Williamson and Brand's (2014) study, even children who did not see the pedagogical embellishments (those in the adult-directed action group) showed some imitation. Additionally, children can learn new behaviors and when to imitate from observing interactions between other people (Repacholi and Meltzoff, 2007; Herold and Akhtar, 2008; Nielsen et al., 2012b).

The identity of the model also matters for children's imitation. Although several studies indicate that infants and young children are more likely to imitate novel acts demonstrated by an adult versus a child (Seehagen and Herbert, 2011; Wood et al., 2012; Zmyj et al., 2012), when familiar acts are demonstrated, peers are more effective models than are adults (Ryalls et al., 2000; Zmyj et al., 2012). Adults who use the child's own language (Buttelmann et al., 2013) or accent (Wagner et al., 2014) are more likely to be imitated by children than are adults using a nonnative language or accent. These interactions suggest that the identity of the model may lead to different motivations for imitating. On one hand, there are benefits to learning from individuals that are likely to be knowledgeable (e.g., adults). On the other hand, it may also be motivating or easier to imitate someone like you, such as peers or members of your own (versus other) social groups.

Practical Implications

Considering the social context and ramifications of imitation has important implications for considering developmental disorders involving atypical social cognition. In particular, Autism spectrum disorder (ASD) is characterized by social, behavioral, and cognitive problems (Baio, 2012). Children on the autism spectrum are less likely to imitate than are typically developing children (e.g., Charman et al., 1997; Rogers et al., 2003, 2008), and one hypothesis for this difference is that children with ASD are particularly unlikely to reproduce acts that are used to engage or affiliate with others (Hobson and Lee, 1999; Ingersoll, 2008; McDuffie et al., 2007; Rogers et al., 2003). A precise characterization of how the imitation of children with ASD varies from that of typically developing children has the potential to contribute to early screens for this disorder. Further, understanding how imitation contributes to theory of mind (Meltzoff, 2007) will improve our understanding of what abilities and experiences may underlie this disorder.

Conclusions

Humans are born into rich social systems that children must master to become competent, integrated social actors. One powerful tool for accomplishing this daunting task is imitation. Even early in infancy, children have the capacity to quickly learn from behavioral examples and to retain the information over substantial delays. Through imitation, children can take up and reproduce others' body acts as well as specific behaviors used on objects. Children can also infer the underlying goals and strategies directing observed surface behaviors. Thus, with imitation, children are well-equipped to acquire the tools, practices, and customs of their cultures.

Historically, imitation has been an important phenomenon for researchers studying child development. It has been (and continues to be) used as a measure of young children's basic cognitive functions, such as memory and action understanding. It also has been used as an indicator of social understanding and a motivation to interact with others. Within psychology more generally, imitation research is helping identify how human learning differs from that of other animal species. Finally, this area of research has extended to other disciplines, for example, by informing models in neuroscience and promoting educational practices and theory. In the twenty-first century, imitation research continues to grow with theoretical and practical implications for a variety of disciplines across the social and behavioral sciences.

See also: Autism Spectrum Disorders; Mirror Neurons, Theory of; Neonatal Imitation; Piaget's Theory of Cognitive Development; Short- and Long-term Memory During Infancy and Early Childhood across Cultures; Theory of Mind During Infancy and Early Childhood Across Cultures, Development of.

Bibliography

Baio, J., 2012. Prevalence of autism spectrum disorders: autism and developmental disabilities monitoring network, 14 sites, United States, 2008. MMWR Surveillence Summary 61, 1–19.

Bakeman, R., et al., 1990. !Kung infancy: the social context of object exploration. Child Development 61 (3), 794–809.

Bandura, A., 1977. Social Learning Theory. Prentice Hall, Englewood Cliffs, NJ.

Barr, R., 2010. Transfer of learning between 2D and 3D sources during infancy: informing theory and practice. Developmental Review 30 (2), 128–154.

Barr, R., Dowden, A., Hayne, H., 1996. Developmental changes in deferred imitation by 6- to 24-month-old infants. Infant Behavior and Development 19 (2), 159–170.

Barr, R., Hayne, H., 1999. Developmental changes in imitation from television during infancy. Child Development 70 (5), 1067–1081.

Barr, R., Muentener, P., Garcia, A., 2007a. Age-related changes in deferred imitation from television by 6- to 18-month-olds. Developmental Science 10 (6), 910–921.

Barr, R., et al., 2007b. The effect of repetition on imitation from television during infancy. Developmental Psychobiology 49 (2), 196–207.

Bauer, P.J., Mandler, J.M., 1989. One thing follows another: effects of temporal structure on 1- to 2-year-olds' recall of events. Developmental Psychology 25 (2), 197–206.

Bauer, P.J., Mandler, J.M., 1992. Putting the horse before the cart: the use of temporal order in recall of events by one-year-old children. Developmental Psychology 28 (3), 441–452.

Bauer, P.J., et al., 2000. Parameters of remembering and forgetting in the transition from infancy to early childhood. Monographs of the Society for Research in Child Development 65 (4), i–213.

Bekkering, H., Wohlschlager, A., Gattis, M., 2000. Imitation of gestures in children is goal-directed. The Quarterly Journal of Experimental Psychology Section A 53 (1), 153–164.

Bellagamba, F., Tomasello, M., 1999. Re-enacting intended acts: comparing 12- and 18-month-olds. Infant Behavior and Development 22 (2), 277–282.

Boyd, R., Richerson, P.J., 1996. Why culture is common, but cultural evolution is rare. Proceedings of the British Academy 88, 77–93.

Brand, R.J., Baldwin, D.A., Ashburn, L.A., 2002. Evidence for 'motionese': modifications in mothers' infant-directed action. Developmental Science 5 (1), 72–83.

Brand, R.J., et al., 2007. Fine-grained analysis of motionese: eye gaze, object exchanges, and action units in infant- versus adult-directed action. Infancy 11 (2), 203–214.

Breazeal, C., Scassellati, B., 2002. Robots that imitate humans. Trends in Cognitive Sciences 6 (11), 481–487.

Brugger, A., et al., 2007. Doing the right thing: infants' selection of actions to imitate from observed event sequences. Child Development 78 (3), 806–824.

Buttelmann, D., et al., 2013. Selective imitation of in-group over out-group members in 14-month-old infants. Child Development 84 (2), 422–428.

Carpenter, M., Akhtar, N., Tomasello, M., 1998. Fourteen- through 18-month-old infants differentially imitate intentional and accidental actions. Infant Behavior and Development 21 (2), 315–330.

Carpenter, M., Call, J., Tomasello, M., 2005. Twelve- and 18-month-olds copy actions in terms of goals. Developmental Science 8 (1), F13–F20.

Charman, T., et al., 1997. Infants with autism: an investigation of empathy, pretend play, joint attention, and imitation. Developmental Psychology 33 (5), 781–789.

Collie, R., Hayne, H., 1999. Deferred imitation by 6- and 9-month-old infants: more evidence for declarative memory. Developmental Psychobiology 35 (2), 83–90.

Csibra, G., Gergely, G., 2006. Social learning and social cognition: the case for pedagogy. In: Johnson, M.H., Munakata, Y. (Eds.), Processes of Change in Brain and Cognitive Development. Attention and Performance XXI. Oxford University Press, Oxford, pp. 249–274.

Dickerson, K., et al., 2013. Age-related changes in learning across early childhood: a new imitation task. Developmental Psychobiology 55 (7), 719–732.

Elsner, B., 2007. Infants' imitation of goal-directed actions: the role of movements and action effects. Acta Psychologica 124 (1), 44–59.

Elsner, B., Aschersleben, G., 2003. Do I get what you get? Learning about the effects of self-performed and observed actions in infancy. Consciousness and Cognition 12 (4), 732–751.

Flynn, E., Whiten, A., 2008. Imitation of hierarchical structure versus component details of complex actions by 3- and 5-year-olds. Journal of Experimental Child Psychology 101 (4), 228–240.

Hamlin, J.K., Hallinan, E.V., Woodward, A.L., 2008. Do as I do: 7-month-old infants selectively reproduce others' goals. Developmental Science 11 (4), 487–494.

Harnick, F.S., 1978. The relationship between ability level and task difficulty in producing imitation in infants. Child Development 49 (1), 209–212.

Hayne, H., Herbert, J., Simcock, G., 2003. Imitation from television by 24- and 30-month-olds. Developmental Science 6 (3), 254–261.

Herold, K.H., Akhtar, N., 2008. Imitative learning from a third-party interaction: relations with self-recognition and perspective taking. Journal of Experimental Child Psychology 101 (2), 114–123.

Hobson, R.P., Lee, A., 1999. Imitation and identification in autism. The Journal of Child Psychology and Psychiatry and Allied Disciplines 40 (4), 649–659.

Horner, V., Whiten, A., 2005. Causal knowledge and imitation/emulation switching in chimpanzees (*Pan troglodytes*) and children (*Homo sapiens*). Animal Cognition 8 (3), 164–181.

Horner, V., et al., 2006. Faithful replication of foraging techniques along cultural transmission chains by chimpanzees and children. Proceedings of the National Academy of Sciences of the United States of America 103 (37), 13878–13883.

Huber, L., et al., 2009. The evolution of imitation: what do the capacities of non-human animals tell us about the mechanisms of imitation? Philosophical Transactions of the Royal Society B: Biological Sciences 364 (1528), 2299–2309.

Ingersoll, B., 2008. The social role of imitation in autism: Implications for the treatment of imitation deficits. Infants & Young Children 21 (2), 107–119.

Itakura, S., et al., 2008. How to build an intentional android: Infants' imitation of a robot's goal-directed actions. Infancy 13 (5), 519–532.

Johnson, S.C., 2003. Detecting agents. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 358 (1431), 549–559.

Johnson, S.C., Booth, A., O'Hearn, K., 2001. Inferring the goals of a nonhuman agent. Cognitive Development 16 (1), 637–656.

Lyons, D.E., Young, A.G., Keil, F.C., 2007. The hidden structure of overimitation. Proceedings of the National Academy of Sciences of the United States of America 104 (50), 19751–19756.
Mahajan, N., Woodward, A.L., 2009. Seven-month-old infants selectively reproduce the goals of animate but not inanimate agents. Infancy 14 (6), 667–679.
Marshall, P.J., Young, T., Meltzoff, A.N., 2011. Neural correlates of action observation and execution in 14-month-old infants: an event-related EEG desynchronization study. Developmental Science 14 (3), 474–480.
McDuffie, A., et al., 2007. Developmental correlates of different types of motor imitation in young children with autism spectrum disorders. Journal of Autism and Developmental Disorders 37 (3), 401–412.
McGuigan, N., Graham, M., 2010. Cultural transmission of irrelevant tool actions in diffusion chains of 3- and 5-year-old children. European Journal of Developmental Psychology 7 (5), 561–577.
McGuigan, N., Makinson, J., Whiten, A., 2011. From over-imitation to super-copying: adults imitate causally irrelevant aspects of tool use with higher fidelity than young children. British Journal of Psychology 102 (1), 1–18.
Meltzoff, A.N., 1988a. Infant imitation after a 1-week delay: long-term memory for novel acts and multiple stimuli. Developmental Psychology 24 (4), 470–476.
Meltzoff, A.N., 1988b. Imitation, objects, tools, and the rudiments of language in human ontogeny. Human Evolution 3 (1–2), 45–64.
Meltzoff, A.N., 1988c. Imitation of televised models by infants. Child Development 59 (5), 1221–1229.
Meltzoff, A.N., 1995a. What infant memory tells us about infantile amnesia: long-term recall and deferred imitation. Journal of Experimental Child Psychology 59 (3), 497–515.
Meltzoff, A.N., 1995b. Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. Developmental Psychology 31 (5), 838–850.
Meltzoff, A.N., 2007. 'Like me': a foundation for social cognition. Developmental Science 10 (1), 126–134.
Meltzoff, A.N., Moore, K.M., 1977. Imitation of facial and manual gestures by human neonates. Science 198 (4312), 75–78.
Meltzoff, A.N., Moore, M.K., 1983. Newborn infants imitate adult facial gestures. Child Development 54 (3), 702–709.
Meltzoff, A.N., Moore, M.K., 1997. Explaining facial imitation: a theoretical model. Early Development and Parenting 6 (3–4), 179–192.
Meltzoff, A.N., Prinz, W., 2002. The Imitative Mind: Development, Evolution and Brain Bases. Cambridge University Press.
Nadel, J., 2002. Imitation and imitation recognition: functional use in preverbal infants and nonverbal children with autism. In: Meltzoff, A.N., Prinz, W. (Eds.), The Imitative Mind: Development, Evolution, and Brain Bases. Cambridge University Press, New York, USA, pp. 42–62.
Nagell, K., Olguin, R.S., Tomasello, M., 1993. Processes of social learning in the tool use of chimpanzees (*Pan troglodytes*) and human children (*Homo sapiens*). Journal of Comparative Psychology 107 (2), 174–186.
Nielsen, M., 2006. Copying actions and copying outcomes: social learning through the second year. Developmental Psychology 42 (3), 555–565.
Nielsen, M., Cucchiaro, J., Mohamedally, J., 2012a. When the transmission of culture is child's play. PLoS ONE 7 (3).
Nielsen, M., Moore, C., Mohamedally, J., 2012b. Young children overimitate in third-party contexts. Journal of Experimental Child Psychology 112 (1), 73–83.
Nielsen, M., Simcock, G., Jenkins, L., 2008. The effect of social engagement on 24-month-olds' imitation from live and televised models. Developmental Science 11 (5), 722–731.
Over, H., Carpenter, M., 2009. Priming third-party ostracism increases affiliative imitation in children. Developmental Science 12 (3), F1–F8.
Paulus, M., Hunnius, S., Bekkering, H., 2013. Neurocognitive mechanisms underlying social learning in infancy: infants' neural processing of the effects of others' actions. Social Cognitive and Affective Neuroscience 8 (7), 774–779.
di Pellegrino, G., et al., 1992. Understanding motor events: a neurophysiological study. Experimental Brain Research 91 (1), 176–180.
Piaget, J., 1962. Play, Dreams, and Imitation in Childhood. Norton, New York.
Pinkham, A.M., Jaswal, V.K., 2011. Watch and learn? Infants privilege efficiency over pedagogy during imitative learning. Infancy 16 (5), 535–544.
Reid, V.M., Striano, T., Iacoboni, M., 2011. Neural correlates of dyadic interaction during infancy. Developmental Cognitive Neuroscience 1 (2), 124–130.
Repacholi, B.M., Meltzoff, A.N., 2007. Emotional eavesdropping: infants selectively respond to indirect emotional signals. Child Development 78 (2), 503–521.
Rizzolatti, G., et al., 1999. Resonance behaviors and mirror neurons. Archives Italiennes de Biologie 137 (2–3), 85–100.
Rizzolatti, G., et al., 1996. Premotor cortex and the recognition of motor actions. Cognitive Brain Research 3 (2), 131–141.
Rogers, S.J., et al., 2003. Imitation performance in toddlers with autism and those with other developmental disorders. Journal of Child Psychology and Psychiatry 44 (5), 763–781.
Rogers, S.J., et al., 2008. Deferred and immediate imitation in regressive and early onset autism. Journal of Child Psychology and Psychiatry 49 (4), 449–457.
Ryalls, B.O., Gul, R.E., Ryalls, K.R., 2000. Infant imitation of peer and adult models: evidence for a peer model advantage. Merrill-Palmer Quarterly 46 (1), 188–202.
Schulz, L.E., Hooppell, C., Jenkins, A.C., 2008. Judicious imitation: children differentially imitate deterministically and probabilistically effective actions. Child Development 79 (2), 395–410.
Seehagen, S., Herbert, J.S., 2011. Infant imitation from televised peer and adult models. Infancy 16 (2), 113–136.
Slaughter, V., Corbett, D., 2007. Differential copying of human and nonhuman models at 12 and 18 months of age. European Journal of Developmental Psychology 4 (1), 31–45.
Subiaul, F., et al., 2007a. Cognitive imitation in typically-developing 3- and 4-year olds and individuals with autism. Cognitive Development 22 (2), 230–243.
Subiaul, F., et al., 2007b. Cognitive imitation in 2-year-old children (*Homo sapiens*): a comparison with rhesus monkeys (*Macaca mulatta*). Animal Cognition 10 (4), 369–375.
Tomasello, M., 1999. The Cultural Origins of Human Cognition. Harvard University Press, Cambridge, MA, USA.
Uzgiris, I.C., 1981. Two functions of imitation during infancy. International Journal of Behavioral Development 4 (1), 1–12.
Wagner, L., Dunfield, K.A., Rohrbeck, K.L., 2014. Children's use of social cues when learning conventions. Journal of Cognition and Development 15 (3), 479–494.
Want, S.C., Harris, P.L., 2001. Learning from other people's mistakes: causal understanding in learning to use a tool. Child Development 72 (2), 431–443.
Want, S.C., Harris, P.L., 2002. How do children ape? Applying concepts from the study of non-human primates to the developmental study of 'imitation' in children. Developmental Science 5 (1), 1–14.
Whiten, A., et al., 1996. Imitative learning of artificial fruit processing in children (*Homo sapiens*) and chimpanzees (*Pan troglodytes*). Journal of Comparative Psychology 110 (1), 3–14.
Whiten, A., et al., 2006. Imitation of hierarchical action structure by young children. Developmental Science 9 (6), 574–582.
Whiten, A., et al., 2004. How do apes ape? Animal Learning & Behavior 32 (1), 36–52.
Williamson, R.A., Brand, R.J., 2014. Child-directed action promotes 2-year-olds' imitation. Journal of Experimental Child Psychology 118, 119–126.
Williamson, R.A., Jaswal, V.K., Meltzoff, A.N., 2010. Learning the rules: observation and imitation of a sorting strategy by 36-month-old children. Developmental Psychology 46 (1), 57–65.
Williamson, R.A., Markman, E.M., 2006. Precision of imitation as a function of preschoolers' understanding of the goal of the demonstration. Developmental Psychology 42 (4), 723–731.
Williamson, R.A., Meltzoff, A.N., 2011. Own and others' prior experiences influence children's imitation of causal acts. Cognitive Development 26 (3), 260–268.
Williamson, R.A., Meltzoff, A.N., Markman, E.M., 2008. Prior experiences and perceived efficacy influence 3-year-olds' imitation. Developmental Psychology 44 (1), 275–285.
Wood, L.A., Kendal, R.L., Flynn, E.G., 2012. Context-dependent model-based biases in cultural transmission: children's imitation is affected by model age over model knowledge state. Evolution and Human Behavior 33 (4), 387–394.
Zmyj, N., et al., 2012. Fourteen-month-olds' imitation of differently aged models. Infant and Child Development 21 (3), 250–266.

Immigration and Crime in the United States

Richard Stansfield, Rutgers University, Camden, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Immigration has recently taken its place front and center of American politics again, with efforts to achieve comprehensive immigration reform and establish a path to citizenship for undocumented immigrants already in the country. The media discussion generated has focused on how to treat illegal immigrants in the country. Missing from public debate on the topic is an examination of what effect immigration has had on American society. To date, scholars have been in agreement that communities within which immigrants have settled are both safer and economically stronger. In large gateway areas, for example, immigration has been associated with lower crime rates, economic revitalization, and reductions in concentrated poverty. While the increase in Hispanic populations cannot be unequivocally tied to these positive trends, immigration has fundamentally changed the character of urban America. This article will briefly examine the theoretical connections between immigration and crime, before discussing evidence searching for a relationship between these elements, the importance of contexts and potential areas of growth for the immigration-crime literature.

Introduction

The United States witnessed significant and rapid changes in society brought about by immigration from all over the globe. Immigration reached historically high levels during the 1990s, particularly from Central and Southern America, so that now 15% of the population is Latino American. With very visible changes in the social landscape of America, immigration was accompanied by vehement antiimmigrant public sentiment and an introduction of tough bans on illegal immigration in some states. At the heart of public outrage was the belief that recently arrived immigrants, especially undocumented immigrants, represented a substantial criminal threat (Longazel, 2013). This belief tended to be strongest among native-born individuals, especially those with an inflated view of the relative size of the immigrant population. The antiimmigrant public opinion mirrored that witnessed in Western Europe, where the response to growing racial heterogeneity often took repressive and antidemocratic forms, including defining excolonial subjects and new immigrants as threats to the national culture.

Missing from public debates on the topic has been the examination of what effects immigration has had on the communities of their new destinations. The value-laden term 'immigrant' as measured in American studies of immigration and crime has referred to persons born outside of the United States (synonymous with the term 'foreign born'). This immigrant status was typically determined by a question on birthplace in the census questionnaire, and conceptually distinct from whether a person has become a naturalized US citizen. Since the early 1990s, as the immigrant population (especially the undocumented population) increased sharply to historic highs, the rates of violent crimes and property crimes in the United States decreased significantly, in some instances to historic lows. Moreover, data from the census and a wide range of other empirical studies showed that for every ethnic group without exception, incarceration rates among young men were lowest for recent immigrants, even those who were the least educated. Subsequently, the growth of immigration was argued to be one of the key reasons for the crime decline in the United States since the early 1990s. This article will briefly examine the theoretical connections between immigration and crime, before discussing evidence searching for a relationship between these elements, the importance of contexts, and potential areas of growth for the immigration-crime literature.

Theory Connecting Immigration and Crime

Sociological and criminological studies of the immigration and crime relationship are not new. In many ways, today's debates are greatly similar to debates about the arrival of European immigrants to the United States at the turn of the twentieth century. Edwin Sutherland was one notable sociologist of the early twentieth century who saw the utility of using social science to examine the immigrant–criminality link. Sutherland (1947) witnessed that any link between immigration status and crime was in fact a function of the social change going on around immigrants in their new destination cities, wherein immigrant youth faced an increased exposure to conditions of poverty, poor health, and Americanized cultural norms favorable to criminal offending. Sutherland argued that there was not anything inherently criminal about America's newcomers, rather any connection that did exist between immigrant status and crime could be explained by environmental conditions.

This line of thinking was fleshed out by recent scholars of community crime under the label of *social disorganization theory*. Social disorganization refers to the inability of a community structure to realize the common values of its residents and maintain effective social control (Bursik, 1988). That is, social disorganization addresses the structural barriers that impede the development of the formal and informal ties that promote the ability of the community to solve common problems. As the United States underwent significant change in the early twentieth century from small stable farming communities to large industrial cities, Shaw and McKay (1931) hypothesized that inner city Chicago was characterized by residential instability, ethnic heterogeneity, and poverty. Poverty was thought to be criminogenic as it prevented communities from mobilizing formal and informal social control mechanisms.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45062-2

Residential instability was seen as a barrier to social network development, important for developing and maintaining a system of informal social control. And ethnic heterogeneity, accompanied by fear and mistrust, was thought to impede communication. As the post-1965 wave of Latino immigrants settled in predominantly urban neighborhoods subject to these same community conditions, the theory of social disorganization provided a framework for scholars to suggest that recent immigration would contribute to community crime because of the high population turnover and increasing population heterogeneity that accompany the immigration process (Shihadeh and Barranco, 2013).

The behavior of specific ethnic groups and their incorporation into specific societies have been examined in the context of a number of other sociological and criminological theories too. Some of these assumed that the motivation for crime preceded arrival to the host society. The *importation model*, as Wortley (2009) described it, assumed explicitly that the individuals chose to leave their land of origin to achieve criminal objectives in the destination society. This model has explained the activities of terrorist organizations that have operated within the United Kingdom in the past decade or the operation of Mexican drug cartels within the United States. In an interesting analysis of the viability of the importation model, Varese (2006) examined how northern Italian mafias migrate. Varese's analysis of two deliberate attempts to transplant mafia branches in new territories highlighted that supply on its own was not sufficient. Transplanting a criminal group, in this case, was only successful when combined with the presence of a local demand for criminal protection. Thus while the importation perspective has wide conservative appeal (Wortley, 2009), it is important to consider environmental conditions and opportunities in the new destination.

Higher rates of criminal involvement, for example, may be related to the level of blocked opportunities and other social circumstances. The viability of these perspectives is also dependent on the underlying motivation behind immigration. The motivation for immigration to the United States is complicated and nuanced. Patterns of immigrant settlement varied historically based on industrial shifts, economic needs, and policies of both sending and receiving areas (Durand and Massey, 2004). In general, however, the reason most individuals from Central and Southern America chose to enter the United States, legally or illegally, was because of the gap between life aspirations and the local realities of achieving them, realizing that their country of origin may stay poor for several generations to come (Portes and Rumbaut, 2006). Thus for the most ambitious and the most resourceful, economic opportunities and the sustained demand for immigrant labor by US employers, made immigration to the United States an attractive way to overcome this contradiction. *Economic theories*, therefore, also proved persuasive in predicting an increase in crime in communities with large concentrations of new immigrants. As an example, new arrivals were thought to initially be faced with limited opportunities entering into the labor market due to language barriers, a lack of well-paying jobs within close proximity, and discrimination (Waldinger, 1993). In the face of restricted legitimate opportunities, recently arrived immigrants were expected to turn to economic acquisition crimes, including instrumental violence, to fulfill their aspirations (Lee et al., 2001).

While all of these theories suggested why immigrants themselves may have increased the likelihood of crime, other scholars have turned this question on its head. Instead, some scholars have asked whether the arrival of immigrants spurred a violent or criminal reaction among existing residents in the host society through changing the structure of the local labor market. To the degree that immigration floods the market with unskilled workers it may weaken the labor market positions, and thus increase the criminal propensities, of both immigrant and native workers alike (Wadsworth, 2010). Communities within which recently arrived Hispanic immigrants and low-income Blacks share the same residential neighborhoods have long been marked with conflict over the ability of Hispanics to obtain jobs, and a perception among Blacks that undocumented workers took jobs away from Americans. And employment trends confirmed the emergence of recently arrived Hispanic immigrants in industry sectors traditionally occupied by low-skill African-American employees (Bean and Stevens, 2003). With a reputation as hardworking employees, newly arrived immigrants may have attracted employers, increasing the level of competition over low-wage jobs.

Empirical Evidence

Despite these theoretical grounds, the evidence has been overwhelming that today's immigrant gateways, in the United States at least, saw some of the largest reductions in crime. As Sampson (2008) discussed, crime in Chicago dropped as immigration surged in the late 1990s. And this was not just limited to recently arrived first-generation immigrants. Evidence also suggested that the second generation integrated successfully into Chicago life, even while maintaining much of their ethnic heritage (Kasinitz et al., 2008). Chicago aside, a growing body of empirical research has considered the relationship of recent immigration and violent crime in contemporary American society with remarkably consistent findings that the presence of new Latino immigrants in neighborhoods or cities resulted in either a negligible or a negative effect on criminal violence (Lee et al., 2001; Martinez and Valenzuela, 2006; MacDonald and Sampson, 2012). Ramiro Martinez and Matthew Lee sum this up nicely in a (2000: p. 496) report for the National Institute of Justice: "the major finding of a century of research on immigration and crime is that immigrants … nearly always exhibit lower crime rates than native groups." While many studies continued to support the notion that the mere presence of recently arrived immigrants has not increased violence, recent scholarship has moved beyond this. Martinez and Stowell (2012), for example, considered the different contexts of an emerging majority Latino city in San Antonio and a heavily Cuban-origin city since the 1980s in Miami. While the authors considered different points in time, different types of homicide, and different types of regional concentration, they still found clear results across time and place, and at both the national and local level – that more immigrants did not mean more homicide.

A smaller number of studies assessed the effects of recent immigration on aggregate property crime rates, with more

mixed results. Hagan and Palloni (1998) first suggested that immigrants may be more likely to engage in property, rather than violent offenses, due to the economic motivation underlying immigration. The authors constructed measures of both legal and illegal immigration within 34 Standard Metropolitan Statistical Areas in 5 southwestern states, and found that the size of the immigrant population was positively related to rates of property crime arrests. Since then, however, empirical results of the immigration-property crime relationship have exhibited more parity with the immigration-violence findings. Reid et al. (2005) examined the relationship of recent immigration and crime across 150 metropolitan areas in the United States. After controlling for relevant demographic and economic influences, they found that the foreign-born population did not increase rates of burglary or larceny. Similar results were found in a 2008 Public Policy Institute of California study of 29 California cities. Finally, in an effort to explicitly examine property offending at the neighborhood level, Stansfield et al. (2013) examined the impact of recent immigration on rates of serious property crime across communities in the fast-growing immigrant gateway of Austin, Texas. Negative binomial regression models with corrections for spatial autocorrelation indicated that recent immigration was not associated with an increased rate of burglary, larceny, or motor vehicle theft once important structural predictors of crime were controlled for.

Why the Crime Protection Effect

While the evidence within the American context suggested a crime protective effect of recent immigration on crime, the next question was why. First, it is important to reiterate that the recent waves of immigrants who arrived in America were not an accurate representation of the sending society, rather they were a selective, economically motivated group with higher aspirations. This notion of *selective immigration* suggests lower criminal involvement in itself (Kubrin and Ishizawa, 2012). It follows that Hispanic immigrants to the United States have been more likely to occupy positions in the labor market (albeit at the low end) than other minorities (Martinez and Valenzuela, 2006), providing stability and regulation in the daily routine, as well as economic sustenance.

The lack of a relationship between new immigration and crime could also be a result of the economic revitalization that results from immigration. There is a growing consensus that immigration revitalized and stabilized local communities (Lee et al., 2001; Stansfield, 2013; Stowell, 2007). Revitalization lessened criminal outcomes, not just among the immigrant population, but also among all population groups through business development and the increase in jobs made for native-born Americans in serving the needs of the expanding population. Many of the jobs created that required more responsibility and skill were likely to be filled by native-born workers (Reid et al., 2005). The past decade witnessed a number of examples of the revitalization of stagnant economies. Queens, New York, has witnessed economic rejuvenation driven largely by the success of Black immigrants from the Caribbean. Similar rejuvenation of inner city areas occurred in Los Angeles, Miami, and other Latino immigrant hotspots (Sampson, 2008).

Additionally, the number of Hispanic-owned businesses has grown at rates far exceeding the growth rate for all US businesses over the past two decades. Data from the 2007 US Census Bureau's Survey of Business Owners showed that numbers of Hispanic-owned businesses in the United States increased at more than double the national rate between 2002 and 2007, reaching a total of 2.3 million Hispanic-owned businesses. Almost half of these, about 45.8%, were owned by the people of Mexican origin. Findings of a University of California Los Angeles (UCLA) study commissioned by the Latino Business Chamber of Greater Los Angeles, where Latino-owned businesses make up the largest share of minority-owned businesses, suggested that Latino-owned businesses have been the key to bringing jobs and revenue back into the economy. One recent study by Stansfield (2013) explicitly examined the role of Hispanic-owned businesses in the relationship between immigration and total property and violent crime rates, confirming that cities with higher rates of new immigration were not associated with higher violent crime rates, and had significantly lower property crime rates. Additional analyses examined whether this relationship was mediated by the presence of Hispanic-owned businesses also yielded empirical support.

In addition to economic resources that Hispanic immigrants brought with them, there is also evidence that speaks to the importance of cultural and structural resources that may have protected immigrants against the forces of low incomes, poor educational and occupation opportunities, and discrimination. Although a diverse group, it is believed that Hispanics do share some core cultural values, some of which may act as a buffer against stresses associated with acculturation and limited opportunity in the United States (Gallo et al., 2009). Theories have been developed, for example, that suggest that features of Hispanic culture, including family stability and religiosity, may have enhanced individual resilience in the face of difficult community conditions (Marin and Marin, 1991). Specifically, Marin and Marin utilized the concept of *'familism'* that represents a mentality in which the needs and objectives of the family are placed over the specific needs of the individual. Familism may have provided the social support network with its associated benefits in the context of adverse circumstances (Gallo et al., 2009). Communities with more immigrants, and in turn more intact families and profamily cultural orientations, may therefore have less crime. Findings from a city-level study by Ousey and Kubrin (2009) supported this argument.

While overall these results of immigration seem largely positive, there are two critical caveats. First, cities vary considerably in how favorable their policies are toward immigrants – something that can have important implications for the economic and community benefit of immigration. In some urban areas such as Austin, Texas, the police department collaborated with local banks to encourage immigrants to open bank accounts in an effort to reduce the risk of robberies. Political efforts to accommodate immigrants and encourage integration also provided strong community benefits in the Delmarva Peninsula (Dunn et al., 2005) and St. James, Minnesota (Carr et al., 2012), for example, whereas draconian measures in the form of local ordinances in Hazleton, Pennsylvania, curtailed the integration of new arrivals. These local contexts are important considerations in the successful integration of newcomers. Second, recent evidence suggests that

prior reporting on the benefits of concentrated immigrants has depicted the Latino community too broadly (Shihadeh and Barranco, 2013). This is discussed further below.

New vs Old Destinations

The economic and social decline of nonmetropolitan America has received less public and academic attention than its metro counterpart, as America's trend over the past century has been one of urbanization. But as small towns have lost their populations to the cities and suburbs, a new influx of immigrants, predominantly Hispanic, has gone the other way, accounting for almost half of the nonmetropolitan population growth in the US since 2000. Understanding the complex interplay between immigration, internal migration, and population growth, and understanding the geographic context of immigration are important tasks for studies of immigration and crime. The majority of existing work on immigration and crime has focused on what the Brookings Institute (2011) terms *major continuous gateways*. New York and Chicago were quintessential immigrant destinations, with large and established immigrant populations for the best part of a century. As job markets in these cities have become saturated, however, immigrants have settled in new destinations that had very small or nonexistent Latino populations up until 1990. This included: *emerging gateways* whose foreign-born populations grew faster than the national rate during the last three decades of the twentieth century, and their immigrant population share has exceeded the national average since 1990 (e.g., Phoenix); *preemerging gateways* like Greensboro and Nashville that had smaller immigrant populations than the other twenty-first century gateways but where immigrant growth has been much faster than the national average; and smaller rural communities that suffered from dwindling populations and economic downturns in recent years.

Existing friendship, professional, and employment networks in major continuous gateways have protected newly arriving immigrants from outside discrimination and against engaging in violence (Sampson, 2008). New destinations, however, have not had significant established ethnic populations (e.g., networks of Latino families and businesses preexisting the recent wave of immigration) and as such lack "the social control umbrella of traditional destinations" (Shihadeh and Winters, 2010: p. 631), depriving new immigrants of a significant source of social capital. This difference in community context was considered an important element for successful integration. Neighborhoods in new destinations have differed in other important ways for recently arrived immigrants too. Segregation and housing research highlighted that Latinos were more segregated in new destinations, with neighbors who were more likely to be poor, contributing to the concentration of poverty. As such, immigrants in new destinations have faced tougher economic and social contexts for adaption, integration, and assimilation (Massey, 2008).

Recent studies by Edward Shihadeh and colleagues examined the effect of Latino immigration to new destinations. Shihadeh and Winters (2010) suggested that Latinos were able to offset some of the lack of existing immigrant community by relying on local religious ecology to keep crime low and community strong. Nevertheless, results also suggested that high Latino immigration areas in new destinations were not the safe havens they were in traditional destinations. Instead Latinos were murdered at an exceedingly high rate (Shihadeh and Barranco, 2010). This relationship between immigration and serious violence existed for only the most recent waves of Latino immigrants, however, those who arrived in the US since 1990 (Shihadeh and Barranco, 2013). The recent body of research from Shihadeh and colleagues also suggested that in these new destinations, the theory of social disorganization (discussed earlier) was apt for understanding the relationship between Latino immigration and crime, exacerbated in conditions of economic deprivation. This has led to the more refined conclusion, that while immigration was paradoxically associated with lower crime, this was not the case for the most recent wave of immigration.

Civic Engagement

Work in this literature must also begin to take seriously the differences among the immigrant population in religiosity and civic engagement. The study of civic engagement, voluntarism, and the strength of community organizations has been a staple of political scientists and social science researchers alike. Seminal work of Robert Putnam (2000) illustrated the critical place of these conditions in community stability and individual lives. The influence of Putnam's concept of 'bowling alone' brought light to the importance of a decline in social memberships and levels of interpersonal trust in the United States. This decline was argued to be particularly pronounced among America's minority populations. The low levels of civic engagement were thought to be a reflection of stagnating social cohesion, trust, and interpersonal social networks, which have been causally associated with crime.

The association between civic engagement and crime can be summarized briefly. Within this framework, Lee and Bartowski (2004) defined social capital as the integration within social networks that promotes norms of reciprocity, trust, and an ethic of civic engagement. Therefore at the individual level, participation in networks and community organizations is an important mechanism of social capital that should deter crime. Furthermore, following from prior research on civic engagement (Sanchez-Jankowski, 2002), attitudes toward civic engagement and knowledge of civic engagement may also be important elements to consider in the association with crime. Following this line of inquiry, scholars within the sociological study of religion have examined the mechanisms through which religious-held beliefs and engagement in religious practice promote civic engagement, social capital, and benefits to the community and a deterrent effect on crime, particularly those crimes that are explicitly condemned by their faith (Lee and Bartowski, 2004).

Immigration has certainly influenced the local religious and civic landscape, yet studies of recent immigration and religion and civic engagement have only just begun to emerge. A rich discussion of immigration, religiosity, and civic engagement in the United States can be found in the work of Stepcik et al. (2009), who considered the immigrant population of Miami in particular. As they noted, many if not most of these

immigrants are highly religious, making religion an essential component in understanding the multifaceted influence immigrants are having on American life. Faith is central to many immigrants' lives as it helps sustain ties to the homeland culture as a mechanism for easing the psychological disturbance inherent in the migration process. Religion also profoundly influences the broader civic agenda of their destination towns and cities, acting as a vanguard for direct engagement with civil society, community politics, and participation in nonprofit organizations. And religious congregations can be a source of social support in and of themselves. Shihadeh and Winters (2010) have demonstrated the importance of a proreligious Catholic presence in new Latino settlement areas, which lowered violence against Latinos. The association of civic engagement and crime has both empirical and theoretical grounds in explaining variation in individual- and community-level violence for Hispanic immigrants especially, yet remains an underutilized theoretical framework for studies of immigration and crime.

Victimization

A wealth of research has explored the contradiction between the public and political rhetoric criminalizing immigrants and the reality of immigration and crime. Recently, scholars have turned their attention to the way that the rhetorical relationship between immigration and crime has increased the risk of immigrants becoming victims of crime. Zatz and Smith (2012) argued that the moral panic about immigration and crime, and the resulting antiimmigrant laws, policies, and local ordinances, made immigrants and their neighborhoods less safe. They specifically identified three key areas that make immigrants, especially undocumented immigrants, vulnerable to violence, abuse, and exploitation. First, Zatz and Smith suggested that the policies designed to curtail the flow of immigrants into the country led to immigrants choosing to cross the border in remote areas where the journey is far more treacherous. This counteractive effect of border control policies, leading to increasingly more dangerous border crossings and less willingness to return back across the border at a later point in time, has been well documented in prior research using the Mexican Migration Project (Durand and Massey, 2004). With widespread fear and belief that the government are not doing enough, scholars have also documented a growth in civilian vigilante groups taking up arms along the border, as part of the public border response that stemmed from a belief that Latino immigrants were 'invading' US society (Chavez, 2008).

Zatz and Smith also highlighted the unique vulnerability of day laborers and domestic workers, with victimization ranging from wage theft, not being paid for overtime or double shifts, assaults, and other forms of violence. Capeda et al.'s (2012) study in New Orleans highlighted how violence can erupt along multiple dimensions: with clashes between displaced African-American workers who feel that they are losing jobs to cheaper migrant labor; clashes between Mexican and other Central American immigrant groups who may have experienced violence and theft during their journey through Mexico; and violence by smugglers and traffickers who are transporting the day laborers across the border (see also Zatz and Smith, 2012).

Finally, there is a patterning of violence in the home and communities against women. Domestic violence scholars have increasingly noted how it is important to examine the intersections of not only gender and race, but also citizenship and nationality. This followed a rise in cases of immigrant women in the United States who were victims of domestic violence, and without legal residence. Furthermore, their abusers used their residency status as a bargaining tool, threatening to call immigration services and request deportation if their partner seeks to escape. The Violence Against Women Act (VAWA) and the Victims of Trafficking and Protection Against Violence Act were created for this reason. Ostensibly these laws allowed victims to break free of their abusers without relinquishing residency or legal status rights. Roberta Villalon (2010) demonstrated the importance of how these intersections are relevant for understanding the experiences of battered women with her study of the Organization for Refugees of America (ORA) based in Texas. And the benefit of these laws and addressing the intersections can be seen in her model case study of Angela. After years of abuse and torture from her husband, Angela visited ORA for their legal services. After 8 months, she was granted residency for her and her three kids by the federal government under the VAWA provisions.

Unfortunately, however, her case was not always matched. Women whose husbands were not US citizens were less fortunate, pointing to a hole in the laws that were created for this very reason. Furthermore, through interviews with foundational lawyers and archival analysis – Villalon demonstrated how the foundational priorities of the organization had shifted into ones of expediency and efficiency. Thus with the power workers had to select parameters for successful cases, the organization in many ways perpetuated the problem faced by those with the most disadvantaged and least privileged identities. While some who appear to be good clients can navigate through the system, those with language barriers, emotional, or mental health problems, or even those without a fixed address faced an uphill struggle being successful in getting their case seen to by ORA. Villalon's work helped illuminate not just the importance of examining the intersection of citizenship with race and gender and how these areas create unique experiences in the realm of domestic violence, but she also highlighted how the very organizations and laws designed to tackle the problem, must attend to even more intersections, including language and abusers of non-US citizenship.

In all these forms of victimization, there is greater risk within new destinations. As discussed earlier, where ethnic enclaves and institutional supports have not fully developed, the immigration and crime relationship may take on very different forms. This is also true for victimization, where a lack of institutional support and expertise in the local criminal justice system may make victims more reluctant to report abuse.

Discussion

Subsequently, a number of important works emerged that examined the potentially fascinating implications of immigration

in new destinations including small rural areas and midsize cities in America's south. Carr et al. (2012), for example, looked at immigration in Hazleton, Pennsylvania, and St. James, Minnesota. In this article, they argued that immigration may be critical for saving dwindling populations and the attendant economic decline in rural areas – if handled correctly from a policy standpoint. They contrasted the different approaches in these two locales to accommodating immigrants, and demonstrated the importance of an approach that considers the economic needs of rural America. In Hazleton, scores of people have used draconian measures in the form of local ordinances to curtail the integration of new arrivals, whereas the conscious effort in St. James to encourage integration between native and nonnative communities provided strong community benefits. As America's immigrant population continues to venture into these new locales, local politics, and immigrant incorporation will become critical issues.

While these areas for future research have an American focus, the lessons are nonetheless applicable to the immigration-crime question in other areas of the world. Immigration to areas in Southern Europe has been far more problematic (Melossi et al., 2009). The newness of these destinations, the lack of institutional support, and the intense racialization of newcomers that mirrors the criminalization and racialization of African-Americans, are all likely suspects in explaining the higher rates of crime involving recently arrived immigrants. Public opinion and the lack of welcoming institutional support seem particularly damning in this case. It is therefore worth reiterating the point that in America, the protective effect of immigration on crime especially, in the United States seems limited to the traditional destinations with high levels of integration, institutional support, and opportunity. Difficulties are far more pronounced in newer destinations. Yet the literature seems quite clear in that the explicit policy efforts of these new destinations go a long way in determining the success of integrating new immigrants (Dunn et al., 2005). In large gateway areas and new destinations that have welcomed immigration, immigration has been associated with economic revitalization, reductions in concentrated poverty, and lower crime rates.

See also: Communities and Crime; Concentrated Disadvantage; Control, Social; Crime, Geography of; Crime, Immigration, and Ethnicity; Crime, Race and Ethnicity; Ethnicity, Sociology of; Immigration Policy; Migration History; Migration and Development; Migration and Social Capital; Racial Relations; Residential Segregation: Geographic Aspects; Urban Violence.

Bibliography

Bean, F., Stevens, G., 2003. America's Newcomers and the Dynamics of Diversity. Russell Sage Foundation, New York.

Bursik, R.J., 1988. Social disorganization and theories of crime and delinquency: problems and prospects. Criminology 26 (4), 519–552.

Butcher, F., Piehl, A.M., 2008. Crime, corrections and California: what does immigration have to do with it? Public Policy Institute of California 9 (3).

Carr, P., Lichter, D., Kefalas, M., 2012. Can immigration save small town America? Hispanic boomtowns and the uneasy path to renewal. Annals, AAPSS 641, 38–56.

Cepeda, A., Negi, N., Nowotny, K., Arango, J., Kaplan, C., Valdez, A., 2012. Social stressors, special vulnerabilities and violence victimization among Latino immigrant day laborers in post Katrina New Orleans. In: Kubrin, C., Zatz, M., Martinez, R. (Eds.), Punishing Immigrants. NYU Press, New York.

Chavez, L., 2008. The Latino Threat: Constructing Immigrants, Citizens, and the Nation. Stanford University Press, Stanford, CA.

Crowley, M., Lichter, D.T., 2009. Social disorganization in new Latino destinations? Rural Sociology 74 (4), 573–604.

Dunn, T.J., Aragones, A.M., Shivers, G., 2005. Recent migration to the rural Del-Mar-Va peninsula. In: Zuniga, V., Hernandez-Leon, R. (Eds.), New Destinations of Mexican Immigration in the United States. Russell Sage Foundation, New York, pp. 155–183.

Durand, J., Massey, D.S., 2004. Crossing the Border: Research from the Mexican Migration Project. Russell Sage Foundation, New York.

Gallo, L.C., Pendo, F., De los Monteros, K.E., Arguelles, W., 2009. Resiliency in the face of disadvantage: do hispanic cultural characteristics protect health outcomes? Journal of Personality 77, 1707–1746.

Hagan, J., Palloni, A., 1998. Immigration and crime in the United States. In: Smith, J.P., Edmonston, B. (Eds.), The Immigration Debate. National Academy Press, Washington, DC.

Hall, M., Singer, A., DeJong, G.F., Graefe, D.R., 2011. The Geography of Immigrant Skills. Metropolitan Policy Program, Brookings, Washington, DC.

Kasinitz, P., Mollenkopf, J.H., Waters, M.C., Holdaway, J., 2008. Inheriting the City: The Children of Immigrants Come of Age. Harvard University Press, Cambridge, MA.

Kubrin, C., Ishizawa, H., 2012. Why some immigrant neighborhoods are safer than others: findings from Los Angeles and Chicago. Annals, AAPSS 641, 148–173.

Lee, M.R., Bartowski, J.P., 2004. Love they neighbor? Moral communities, civic engagement, and juvenile homicide in rural areas. Social Forces 82 (3), 1001–1035.

Lee, M.T., Martinez Jr., R., Rosenfeld, R., 2001. Does immigration increase homicide rates? Negative evidence from three border cities. The Sociological Quarterly 42, 559–580.

Longazel, J., 2013. Moral panic as racial degradation ceremony: racial stratification and the local-level backlash against Latino/a immigrants. Punishment and Society 15, 88–119.

MacDonald, J., Sampson, R.J., 2012. The world in a city: Immigration and America's changing social fabric. Annals, AAPSS 641, 6–15.

Marin, G., Marin, B.V., 1991. Research with Hispanic Populations. Sage, Newbury Park, CA.

Martinez Jr., R., Lee, M.T., 2000. On immigration and crime. In: Criminal Justice 2000: The nature of crime: Continuity and Change, vol. 1. National Institute of Justice, Washington, DC, pp. 485–524.

Martinez Jr., R., Stowell, J., 2012. Extending immigration and crime studies: national implications and local settings. Annals, AAPSS 641, 174–191.

Martinez Jr., R., Valenzuela, A., 2006. Immigration and Crime: Race, Ethnicity, and Violence. NYU Press, New York.

Massey, D.S., 2008. New Faces in New Places: The Changing Geography of American Immigration. Russell Sage Foundation, New York.

Melossi, D., De Giorgi, A., Massa, E., 2009. The 'normality' of 'second generations' in Italy and the importance of legal status: a self-report delinquency study. In: McDonald, W. (Ed.), Immigration, Crime, and Justice. Emerald, New York, pp. 47–65.

Ousey, G.C., Kubrin, C.E., 2009. Exploring the connection between immigration and violent crime rates in U.S. cities, 1980–2000. Social Problems, 56447–56473.

Portes, A., Rumbaut, R.G., 2006. Immigrant America: A Portrait, third ed. University of California Press, Berkeley, CA.

Putnam, R., 2000. The Collapse and Revival of American Community. Simon & Schuster, New York.

Reid, L.W., Weiss, H.E., Adelman, R.M., Jaret, C., 2005. The immigration-crime relationship: evidence across US metropolitan areas. Social Science Research 34, 757–870.

Sampson, R.J., 2008. Rethinking crime and immigration. Contexts 7, 28–33.

Sampson, R.J., 2012. Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago Press, Chicago.

Sanchez-Zankowski, M., 2002. Minority youth and civic engagement: the impact of group relations. Applied Developmental Science 6 (4), 237–245.

Shaw, C.R., McKay, H.D., 1931. Social Factors in Juvenile Delinquency: Report on the Causes of Crime. National Commission of Law Observance and Enforcement. Government Printing Office, Washington, DC.

Shihadeh, E.S., Barranco, R.E., 2010. Latino immigration, economic deprivation, and violence: regional differences in the effect of linguistic isolation. Homicide Studies 14, 338–356.

Shihadeh, E.S., Barranco, R.E., 2013. The imperative of place: homicide and the new Latino migration. The Sociological Quarterly 54, 81–104.
Shihadeh, E.S., Winters, L., 2010. Church, place, and crime: Latinos and homicide in new destinations. Sociological Inquiry 80, 628–649.
Stansfield, R., 2013. Safer cities: a macro-level analysis of recent immigration, hispanic owned businesses, and crime rates in the United States. Journal of Urban Affairs. http://dx.doi.org/10.1111/juaf.12051.
Stansfield, R., Akins, S., Rumbaut, R.G., Hammer, R., 2013. Assessing the effects of recent immigration on serious property crime in Austin, Texas. Sociological Perspectives 56 (4), 647–672.
Stepcik, A., Rey, T., Mahler, S.J., 2009. Churches and Charity in the Immigrant City: Religion, Immigration, and Civic Engagement in Miami. Rutgers University Press, New Brunswick, NJ.
Stowell, J., 2007. Immigration and Crime: Considering the Direct and Indirect Effects of Immigration on Violent Criminal Behavior. LFB Scholarly, New York.
Sutherland, E., 1947. Principles of Criminology. Lippincott, Philadelphia, PA.
Varese, F., 2006. How mafias migrate: the case of the 'Ndrangheta in northern Italy. Law & Society Review 40 (2), 411–444.
Villalon, R., 2010. Violence Against Latina Immigrants: Citizenship, Inequality and Community. NYU Press, New York.
Wadsworth, T., 2010. Is immigration responsible for the crime drop? An assessment of the influence of immigration on changes in violent crime between 1990 and 2000. Social Science Quarterly 91, 531–553.
Waldinger, R., 1993. The ethnic enclave debate revisited. International Journal of Urban and Regional Research 17, 444–452.
Wortley, S., 2009. The immigration-crime connection: competing theoretical perspectives. International Migration and Integration 10, 349–358.
Zatz, M., Smith, H., 2012. Immigration, crime and victimization: rhetoric and reality. Annual Review of Sociology 8, 141–159.
Zuniga, V., Hernandez-Leon, R., 2005. New Destinations: Mexican Immigration in the United States. Russell Sage Foundation, New York.

Relevant Websites

http://www.brookings.edu/about/programs/metro/stateofmetroamerica – Brookings Institute Metropolitan Policy Program.
http://www.pewhispanic.org/ – The Pew Hispanic Research Center.

Immigration, Applied Research on

Maria E Enchautegui, Urban Institute, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The growing number of immigrants and its importance in the demographic changes being experienced in the United States have resulted in a surge in immigration research in the last two decades. This article is an effort to tap into the multidisciplinary literature on immigration to bring the issues that are capturing the attention of the most recent scholarship, pointing to data limitations and gaps. We identify family-based immigration, integration, enforcement, documenting the undocumented, the effects of immigration on the wages of native workers, and inequality as the most pressing issues in current immigration research.

Overview of US Immigration

As of the writing of this article, there were 232 million international migrants in the world – 3.2% of the world's population (Migration Policy Institute Data Hub http://www.migrationpolicy.org/programs/data-hub). At 40.8 million foreign-born in 2012, United States has by far the largest number of world immigrants (Migration Policy Institute Data Hub Migration Policy Institute Data Hub http://www.migrationpolicy.org/programs/data-hub). According to Census Bureau's projections, mid-twenty-first century will be the first time that immigration will surpass natural increase as the main driver of US population growth (U.S. Census Bureau Population Projections: http://www.census.gov/population/projections/data/national/2012.html). As shown in **Figure 1**, the percentage foreign-born in the United States has been growing since 1970 standing at 13% in 2012.

The foreign-born population residing in the United States is composed of naturalized citizens, lawful permanent residents, aliens with temporary residence permits, and those residing illegally. The American Community Survey provides information on whether a foreign-born person is a US citizen or not. In the 2012 American Community Survey, 46% of the foreign-born persons not of American parents reported to be naturalized citizens (IPUMS, American Community Survey 2012 https://usa.ipums.org/usa/). The remaining of the foreign-born population are noncitizens. These data do not provide information about immigration status of noncitizens. The Department of Homeland Security (DHS) produces estimates of the number of legal permanent residents and nonimmigrants on temporary visas. The latest estimates show that there were 13.3 million legal permanent residents in the United States and 1.9 in temporary statuses (DHS, Immigration Statistics Publications: http://www.dhs.gov/immigration-statistics-publications).

DHS and The Pew Hispanic Trends Project produce estimates of the unauthorized population. Estimates from DHS for 2011 based on the American Community Survey put the number of unauthorized immigrants at 11.5 million (DHS, Immigration Statistics Publications: http://www.dhs.gov/immigration-statistics-publications). The most recent estimate from Pew for March 2012 was 11.7 million (Pew Research Hispanic Trends Project: http://www.pewhispanic.org).

A distinct characteristic of the US immigration system in comparison to other countries is the preponderance of family ties in visa allocation. Family reunification has long been the foundation of US immigration policy. Between Fiscal Years (FY) 2003 and 2012, United States awarded an average of 1.03 million legal permanent resident visas annually (DHS Immigration Statistics: http://www.dhs.gov/immigration-statistics). About two-thirds of these visas were awarded based on family relationships to US citizens and legal permanent residents (**Figure 2**).

In addition to family reunification, the granting of permanent residence visas is also driven by humanitarian, economic, and diversity reasons. In FY 2012, 154 614 or 15% of all permanent resident visas went to refugees and asylees (noncitizens who have been granted asylum). Another 14% or 139 000 of visas were awarded based on employment and skills. But when counting only the worker, not his or her family, the percent of all visas awarded in the basis of employment and skills was only about 6%.

The United States also awards permanent visas to diversify the country of origin of the immigrant population. A lottery is conducted among intended immigrants from countries with a low number of people admitted to the United States. Diversity visa participants should have at least a high-school

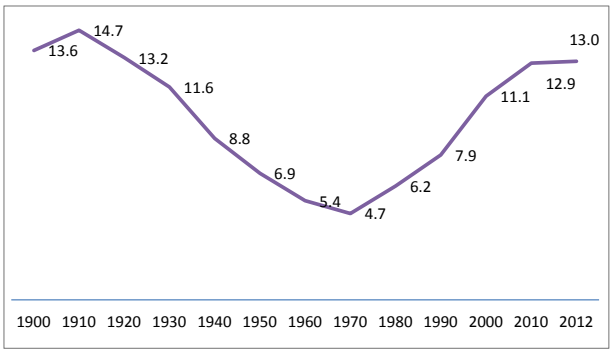

Figure 1 Percentage of foreign-born in United States, 1990–2012. US Census. America's Foreign-Born Population in the Last 50 Years http://www.census.gov/how/infographics/foreign_born.html; 2012 figure from: http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_1YR_S0501&prodType=table.

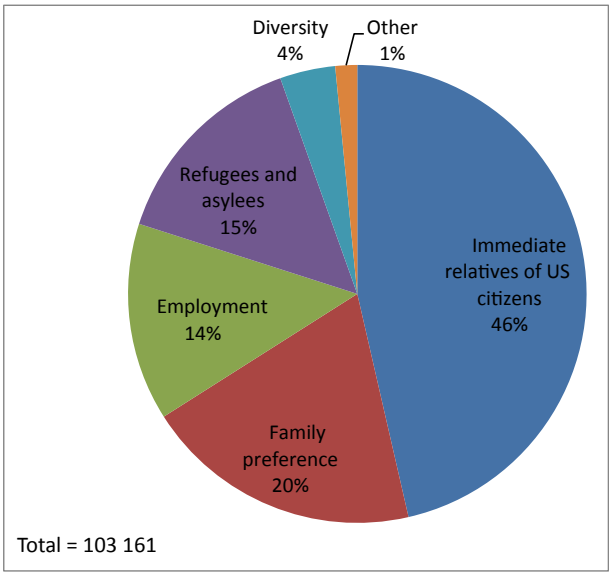

Figure 2 Percentage distribution of legal permanent visas awarded in fiscal year 2012. Department of Homeland Security, Immigration Statistics, Immigration Yearbook, 2012, Table 6. (http://www.dhs.gov/yearbook-immigration-statistics.

diploma or have, within the past 5 years, a minimum of 2 years of experience working in a profession requiring at least 2 years of training or experience (Migration Policy Center, How the US Immigration System Work http://www.immigrationpolicy.org/just-facts/how-united-states-immigration-system-works-fact-sheet). This group accounted for 4% of all legal permanent visas awarded in FY 2012.

The Merits of Family-Based Immigration

The predominance of family ties in awarding legal permanent visas has generated a debate on whether United States is losing international competitiveness by not putting more emphasis on skills in awarding visas, as other economically developed countries do. Examples of this concern can be found in various bills presented at Congress, which stress putting more attention to skills in awarding legal permanent visas of which the most recent are such as S. 744 and H.R. 2131 (Govtrack.us: https://www.govtrack.us).

Researchers have compared family and employment immigrants and have tried to answer whether class of admission makes a difference in terms of the economic standing and economic progress of immigrants. A further indictment on immigration policy is the fact that researchers have also examined if the labor market quality of immigrants have declined across successive entry cohorts.

Since there is no national data containing information on the class of admission under which an immigrant gained legal permanent residence, most work comparing family and employment immigrants use data from United States Citizenship and Immigration Office on specific cohorts that were granted permanent residence. The immigrant-level data are not public and their use has to be procured through special permit from the DHS. Moreover, information contained in these data is quite limited. It does contain information on class of admission, year visa was granted, and a few other variables such as occupation, age, gender, and marital status, but not earnings. With these data and imputing earnings based on the average earnings of the occupation from other data sources, studies have found that employment-based immigrants have higher earnings than family-based immigrants (Akresh, 2006; Barrett, 1998; Jasso and Rosenzweig, 1995).

Another glimpse into the merits, or lack thereof, of family-based immigration comes from the 1996 New Immigrant Survey Pilot – a pilot study of immigrants who became legal permanent residents in 1996, with retrospective information on occupations (Akresh (2006)). The study reports that family immigrants have a lower occupational index than employment immigrants and experience more downward occupational mobility than employment immigrants.

While it appears that employment immigrants have higher earnings, due to the paucity of longitudinal data on immigrants we do not know if family immigrants become like employment immigrants with time, nor how long this would take. The soon to be released second wave of the 2003 New Immigrant Survey will shed new light on the performance of immigrants by class of admission. This survey is a random sample of immigrants who became legalized in 2003 and were interviewed in that same year and again in 2007. It contains information on a host of social and economic variables as well as their class of admission.

Given the paucity of data on class of admission, the merits of the current immigration policy system is being discussed within the banner of immigrant quality. The large representation of persons with low levels of education in the immigration flow and the shift in sending countries after enactment of the 1965 Immigration Act have prompted discussions on whether recent immigrants are of lower labor-market quality than prior immigrants. This question is not at all new since even economists' writings from the late-1800s bring concerns about declining quality of immigrants (see references in Card, 2009). But the availability of microlevel Census data brings empirical precision to this question.

Researchers have compared the entry earnings of successive cohorts of immigrants, relative to native workers and most have found that the labor market quality of immigrants has declined (Borjas, 1985, 2013; Funkhouser and Trejo, 1998; LaLonde and Topel, 1991). But Jasso et al. (2000) claim that to assess the quality of immigrants and to make assessments about immigration policy we must look at legal immigrants and Census data used in other studies that include unauthorized immigrants. Their data on legal husband immigrants from 1972 to 1995 do not show declines in quality, measured by the average occupational earnings. This finding reveals the importance of differentiating between legal and nonlegal immigrants in immigration research.

As more data points become available through the Census and American Community Survey, it has become evident that there is no permanent feature to the labor market quality of US immigrants. Borjas and Friedberg (2009) find an improvement in quality for the immigrants that entered after 1990 attributing it to the increase in the number of employment immigrants resulting from the 1990 Immigration Act. Changes in quality are also evident in the USCIS microdata on different

cohorts of immigrants who were awarded permanent resident visas (Barrett, 1998; Jasso et al., 2000).

Are Immigrants and Their Children Integrating to the US Economy and Society?

The integration of immigrants into US society and economy has been a long-standing theme in the immigration scholarship. The word integration (the term assimilation is also often used) broadly refers to the extent to which, with time in the United States, immigrants' outcomes resemble those of natives on indicators such as income, employment, occupations, fertility, household structure, English language abilities, and civic engagement. Assimilation in earnings takes place because, with time in the United States, immigrants invest in human capital such as education, learn the English language, and accumulate US labor market experience, and these investments increase earnings (Chiswick, 1978). Alternatively, assimilation can be the product of a learning process or job search. With time in the United States, immigrants learn about the US labor market and this knowledge produces a better matching between jobs and skills and a better pay (Daneshvary et al., 1992). Sociologists refer to Anglo conformity, predicting convergence of immigrants to the outcomes of the middle class (see citations in Valdez, 2006). In 1993 Portes and Zhou introduced the concept of segmented assimilation. According to Portes and Zhou "the question is into what particular sector of American society a particular immigrant group assimilates" (p. 82). These authors offer that immigrants assimilate to different segments of the society in three main pathways: assimilation into the middle class; assimilation into the urban underclass, and the preservation of immigrants' culture and values. This perspective has generated a proliferation of research about what assimilation paths are immigrants following, especially the second generation (Greenman and Xie, 2008; Hirschman, 2001; Rumbaut, 1997; Valdez, 2006).

One integration outcome that has received a lot of attention is earnings. Economists have addressed if the earnings of immigrants converge to those of natives. Due to a lack of longitudinal data, this question has been addressed with synthetic cohorts from Census data. The work of LaLonde and Topel (1992) exemplifies this work, finding that the earnings gaps between immigrants and natives are closed within a period of 10 years in the United States.

A problem with synthetic cohorts is that they cannot address selectivity of emigration. If people who emigrate are those with lower earnings. (Hu, 2000; Lubotsky, 2007; Schwabish, 2011), then assimilation patterns from synthetic cohorts may exaggerate the speed at which immigrants catch up with natives. Recently, researchers have used earnings data from the Social Security Administration to follow immigrants' earnings for a long period of time and a few other studies have matched CPS respondents across interview rotations to follow immigrants' earnings for up to 2 years. These studies paint a more dire picture of earnings assimilation, showing little if any earnings growth across time, at times even declines, and assimilation rates much lower than those from synthetic cohorts (Hu, 2000; Schwabish, 2011; Lubotsky, 2007; Kim, 2013).

Most recently and with the advent of data on youth, research on integration has been expanded by looking at the performance of the children of immigrants and testing theories of segmented assimilation. With Adolescent Health Data, Greenman and Xie (2008) analyze a broad set of outcomes and find that assimilation, measured as years of residence in the United States, generation, ethnic composition of the neighborhood, and ethnicity of friends, have a positive effect on some outcomes and negative in others. Using data from CPS on place of birth of mothers and fathers, Card (2005) finds that the wage and the educational gaps with natives are almost closed by the second generation.

The second generation may be the key yardstick by which to assess immigrant integration (Card, 2005). Data sources on immigrant children and youth such as Children of Immigrants (Rumbaut, 1997), Education Longitudinal Survey, and Adolescent Heath allow for examination of a wide variety of outcomes to test hypotheses about integration of the second generation, opening ample opportunities for future research.

Enforcement and the Well-Being of Immigrant Families

Since the 1990s, the levels of immigration enforcement have increased and the range of enforcement practices have widened with the cooperation between federal and local governments, and with state-of-the-art technology (Meissner et al., 2013). Hagan et al. (2011) describe three laws passed between 1996 and 2000 making it easier to deport an immigrant by increasing the type of deportable offences, weakening judicial review, and reducing the ability of the migrant to appeal: the 1996 Illegal Immigration Reform and Immigrant Responsibility Act (IIRIRA); the Antiterrorism and Effective Death Penalty Act of 1996; and the USA Patriot Act of 2001. Enforcement escalated after 9/11. Between FY 2005 and 2012, 2.8 million persons were removed from the country, almost twice the amount from the previous 8 years (DHS Yearbook of Immigration Statistics: http://www.dhs.gov/yearbook-immigration). Most of the deported are from Mexico and Central America.

In the last few years, researchers have started to put out basic information about deportees and analyzing the consequences of enforcement for immigrant families, particularly children. The report *In the Shadow of the Wall* (http://las.arizona.edu/sites/las.arizona.edu/files/UA_Immigration_Report2013web.pdf), sampled 1113 persons who were deported to Mexico. Deportees were interviewed in ports of entry in Mexico immediately following deportation and in migrant shelters in Mexican cities. This study found that 82% of the deportees were men and their median age was 31 years old; 22% had citizen children and 42% were the sole provider. In a survey of 300 deportees taken in El Salvador, Hagan et al. (2008) also found that most were men, in the ages of 18–30 and that 30% had spouses or children in the United States.

Most studies on the impacts of enforcement on families and children are qualitative and based on small samples but several patterns are evident from these studies. It is clear that there is family separation after deportation and that this affects the income of the family left behind (Capps et al., 2007; Hagan et al., 2011). Children's mental health is also affected

(Capps et al., 2007; Hagan et al., 2011). This research also shows that increased enforcement has injected a fear of deportation among unauthorized immigrants (Hagan et al., 2011; Dreby, 2012). Dreby (2012) articulates that it is not the deportation itself but the possibility of deportation that has affected an even greater number of children, even among children whose parents are not at risk of being deported. Dreby finds that increased enforcement has caused children to conflate immigration with illegality, and Menjivar and Abrego (2012) describes how deportation fears spread even among documented immigrants.

Immigration is a social process (Massey, 2004). Having a migrant parent, a migrant sibling, and a migrant child all separately increase the odds of leaving on a first undocumented trip substantially and significantly (Massey and Riosmena, 2010). But enforcement has had an effect on unauthorized immigration. Several studies using the Mexican Migration Project find that border enforcement reduces undocumented immigration (Angelucci, 2012; Hanson and Spilimbergo, 1996; Lessem, 2012; Thom, 2010). Angelucci (2012) concludes that during the most recent period of high enforcement in her data (1997–2003) enforcement actions were a significant deterrent to in-migration, while the effect on average duration of stay became smaller, leading to a decline in the net flow of undocumented immigrants.

Another aspect of enforcement is happening at the local level. Researchers are starting to sort out the effects of local policies on mobility, access to services, and informal work (Amuedo-Dorantes et al., 2013; Lofstrom et al., 2011).

There is a dearth of research about enforcement and employers. Bringing in the employers perspective can enhance our knowledge about effectiveness and consequences of immigration enforcement.

Documenting the Lives of Undocumented Immigrants

With the rise in undocumented immigration, researchers have recognized the need to document the lives of undocumented immigrants. For a long time, the only information available on unauthorized immigrants was population estimates nationally and for the largest states. But this changed with the Mexican Migration Project (http://mmp.opr.princeton.edu/home-en.aspx). The Mexican Migration Project is possibly the earliest concerted effort to collect data on undocumented immigrants and their households. Data have been collected yearly since 1983, though this is not a nationally representative sample of US undocumented immigrants. The sample is drawn from households in 22 Mexican states.

In the last decade there have been efforts to collect data on undocumented immigrants in the United States, and this has been done primarily through place-based surveys in large immigration cities such as Chicago, New York, and Los Angeles. There is also a good number of ethnographic studies of the undocumented immigrants (Bhimji, 2010, Cleaveland, 2012; Menjivar and Abrego, 2012; Salcido and Adelman, 2004). Most recently, a few studies have surfaced using state-based administrative data such as state wage records in the state of Georgia (Hotchkiss and Quispe-Agnoli, 2013) and records from use of emergency rooms in North Carolina hospitals. (DuBard and Massing, 2007).

These different data sources have documented many aspects of undocumented immigrant lives such as food insecurity (Hadley et al., 2008), domestic violence (Salcido and Adelman, 2004), wages and employment (Hotchkiss and Quispe-Agnoli, 2013; Mehta et al., 2002), type of emergency care service used (DuBard and Massin, 2007), effects of economic trends on migrants lives (Cleaveland), labor rights violations (Bernhardt et al., 2009), the job quality (Enchautegui, 2008), and daily laborers (Bhimji, 2010), among others.

Researchers are also looking at the undocumented children and the citizen children of undocumented parents. Gonzales (2011) brings insights about the experiences of undocumented youth in American schools as they transit from an environment where illegality does not matter to one where illegality becomes central to their lives. Growing up as a child of undocumented parents also has implications for many dimensions of child development (Yoshikawa and Kalil, 2011). An emerging group in the undocumented population and one that will likely generate future research is unaccompanied children, children without lawful immigration status with no parent or legal guardian in the United States. The Office of Refugee Settlement, Program of Unaccompanied Alien Children reported serving an average of 8000 children annually and this figured jumped to 13 000 in FY 2012 and to 25 000 in FY 2013 (https://www.acf.hhs.gov/sites/default/files/orr/unaccompanied_childrens_services_fact_sheet.pdf).

Given the importance of undocumented immigrants in the US population and that immigration status is important in determining eligibility for government programs, the Government Accounting Office (2006) has issued recommendations about how to collect immigration status data on government statistics, but these recommendations have not been adopted by the Current Population Survey, the American Community Survey, nor government health surveys. Currently, two government-sponsored data sets collect information on undocumented status. These are the National Agricultural Workers Survey conducted by Bureau of Labor Statistics and the Survey of Income and Program Participation from the Census Bureau.

Impacts of Immigrants on Native's Wages

The impacts of immigration on the wages of native workers continue to be one of the most researched and most debated topics in immigration. Methodology twists and new evidence keeps bringing new insights to this area of research. Earlier studies estimated the effects of immigrants on native wages through variations in the share of metro-area population that are immigrants; Card (2005) and LaLonde and Topel (1991) exemplify this approach. Most of these studies find small or zero effects of immigration on native wages. But these studies are criticized on the basis that natives in local areas respond to immigration by moving to lower immigration areas (Borjas, 2003). Recent work by Cadena and Kovak (2012) adds a further angle to the limitations of the spatial approach when finding a high mobility rate of Mexican Americans and that this high mobility protects low-skilled natives from negative local-demand shocks such as in the

Great Recession. This research points out to the limitations of the spatial approach in uncovering the effects of immigration on the wages of natives.

More recent studies use a structural approach, departing from a production function. The unit of observation is the relative quantities at the national level of workers by skills, experience, and nativity. With this approach Borjas (2003) and Borjas et al. (2010) find large negative effects of immigration on native wages, especially for workers with no high school graduates-with no further education, but effects on college-educated are also found. Borjas' modeling assumes, and examination of occupations suggests, that immigrants and natives within education–experience combinations are perfect substitutes. Ottaviano and Peri- O&P (2012) using a similar approach find negative but small effects of immigration in the wages of natives. The key difference is that O&P assume that the low-skilled labor is composed of high-school dropouts and high-school graduates and that these two groups are substitutes. Borjas and coauthors, however, assume that the low-skilled group is composed of only high-school dropouts. O&P joining of high-school dropouts and high school graduates-with no further education under the low-skilled group is substantiated by evidence that the substitutability between these two groups is almost infinite (Card, 2009). Therefore, the aggregation of high-school dropouts and high school graduates-with no further education under one skill group is critical in these estimations, reducing the effects of immigration on natives' wages. Although immigration has affected the relative supply of high-school dropouts, it has not affected the relative supply of the combined group of dropouts and high-school diplomas.

It must be noted that this research does not distinguish between documented and undocumented immigrants and we do not know if the lack of competition is due to undocumented immigrants not competing openly in the labor market.

Immigration and Inequality

Immigration has also made its way into current discussions about increased inequality. Immigration could increase inequality directly and indirectly. Directly it would increase inequality if immigrants' wages are more unequal than natives' and if immigrants' average income is different from natives (Blau and Khan, 2012). For instance, a high share of poor immigrants will increase the population at the bottom, while a high share of rich immigrants will increase the population at the top. If immigrants are a small share of the population and their skills are similar to those of natives, effects on income inequality will be small (Blau and Khan, 2012). Indirectly, immigrants could increase inequality if it reduces the wages of low-skilled workers, or if it helps raising the wages of those at the top. A study by Card (2009) concluded that immigration contributed only to a small percent (5%) of the increase in inequality experienced between 1980 and 2000. Blau and Khan (2012) reached similar conclusions showing that immigration increased inequality by 6% among men and by 3% among women.

This literature is quite small now but future research may start looking at the role of second-generation immigrants on inequality, similar to what has happened in integration research.

See also: Ethnicity and Migration in Europe; Ethnography in Applied Social Research; Global Migration; Immigration Policy; Immigration and Crime in the United States; Immigration, Emigration, and Citizenship; Immigration: Political Aspects; International Migration; Location Theory; Migration History; Migration and Development; Migration and Social Capital; Migration, Theory of; Migration: Cultural Aspects; Migration: Motivations; Qualitative Research; Sample Survey Methodology, History of.

Bibliography

Akresh, I.R., 2006. Occupational mobility among legal immigrants to the United States. International Migration Review 40, 854–884.

Amuedo-Dorantes, C., Puttitanun, T., Martinez-Donate, A.P., 2013. How do tougher immigration measures affect unauthorized immigrants? Demography 50, 1067–1091.

Angelucci, M., 2012. US border enforcement and the net flow of Mexican illegal migration. Economic Development and Cultural Change 60, 311–357.

Barrett, A., 1998. The effect of immigrant admission criteria on immigrant labour-market characteristics. Population Research and Policy Review 17, 439–456.

Bernhardt, A., Milkman, R., Theodore, N., Heckathorn, D., Auer, M., DeFilippis, J., Spiller, M., 2009. Broken Laws, Unprotected Workers. National Employment Law Project, Washington, DC.

Bhimji, F., 2010. Undocumented immigrant day laborers coping with the economic meltdown in Los Angeles. Cultural Dynamics 22, 157–178.

Blau, F., Khan, L., 2012. Immigration and the Distribution of Incomes. IZA Discussion Paper # 6921, Bonn, Germany.

Borjas, G.J., 2013. The Slowdown in the Economic Assimilation of Immigrants: Aging and Cohort Effects Revisited Again. National Bureau of Economics Research Working Paper, w19116, Cambridge, MA.

Borjas, G.J., 2003. The labor demand curve is downward sloping: re-examining the impact of immigration on the labor market. Quarterly Journal of Economics 118, 1335–1374.

Borjas, G.J., 1985. Assimilation, changes in cohort quality, and the earnings of immigrants. Journal of Labor Economics 3, 463–489.

Borjas, G.J., Grogger, J., Hanson, G.H., 2010. Immigration and the economic status of African-American men. Economica 77, 255–282.

Borjas, G.J., Friedberg, R.M., 2009. Recent Trends in the Earnings of New Immigrants to the United States. National Bureau of Economic Research, w15406, Cambridge, MA.

Cadena, B.C., Kovak, B.K., 2012. Immigrants Equilibrate Local Labor Markets: Evidence from the Great Recession. National Bureau of Economic Research, w19272, Cambridge, MA.

Capps, R., Castaneda, R., Chaudry, A., Santos, R., 2007. Paying the Price: The Impact of Immigration Raid Son America's Children. Urban Institute, Washington, DC.

Card, D., 2005. Is the new immigration really so bad? The Economic Journal 115, F300–F323.

Card, D., 2009. Immigration and inequality. American Economic Review Papers and Proceedings 99, 1–21.

Chiswick, B.R., 1978. The effect of Americanization on the earnings of foreign-born men. The Journal of Political Economy 86 (5), 897.

Cleaveland, C., 2012. "In this country, you suffer a lot": undocumented Mexican immigrant experiences. Qualitative Social Work 11 (6), 566–586.

Daneshvary, N., Herzog Jr., H.W., Hofler, R.A., Schlottmann, A.M., 1992. Job search and immigrant assimilation: an earnings frontier approach. The Review of Economics and Statistics 74, 482–492.

Dreby, J., 2012. The burden of deportation on children in Mexican immigrant families. Journal of Marriage and Family 74, 829–845.

DuBard, C.A., Massin, M.W., 2007. Trends in emergency medicaid expenditures for recent and undocumented immigrants. The Journal of the American Medical Association 297, 1085–1092.

Enchautegui, M., 2008. The job quality of undocumented immigrants. Industrial Relations: A Journal of Economy and Society 47, 108–113.

Funkhouser, E., Trejo, S.J., 1998. Labor market outcomes of female immigrants in the United States. In: Smith, J.P., Edmonston, B. (Eds.), The Immigration Debate: Studies on the Economic, Demographic, and Fiscal Effects of Immigration. National Academy Press, Washington, DC, pp. 239–288.

Gonzales, R.G., 2011. Learning to be illegal: undocumented youth and shifting legal contexts in the transition to adulthood. American Sociological Review 76, 602–619.

Greenman, E., Xie, Y., 2008. Is assimilation theory dead? The effect of assimilation on adolescent well-being. Social Science Research 37, 109–137.

Hadley, C., Galea, S., Nandi, V., Nandi, A., Lopez, G., Strongarone, S., Ompad, D., 2008. Hunger and health among undocumented Mexican migrants in a US urban area. Public Health Nutrition 11, 151–158.

Hagan, J., Eschbach, K., Rodriguez, N., 2008. U.S. deportation policy, family separation, and circular cigration. International Migration Review 42, 64–88.

Hagan, J.M., Rodriguez, N., Castro, B., 2011. Social effects of mass deportations by the United States government, 2000–10. Ethnic and Racial Studies 34 (8), 1374–1391.

Hanson, G.H., Spilimbergo, A., 1996. Illegal Immigration, Border Enforcement, and Relative Wages: Evidence from Apprehensions at the US-Mexico Border. National Bureau of Economic Research, w5582, Cambridge, MA.

Hirschman, C., 2001. The educational enrollment of immigrant youth: a test of the segmented-assimilation hypothesis. Demography 38, 317–336.

Hotchkiss, J.L., Quispe-Agnoli, M., 2013. The expected impact of state immigration legislation on labor market outcomes. Journal of Policy Analysis and Management 32, 34–59.

Hu, W.Y., 2000. Immigrant earnings assimilation: estimates from longitudinal data. The American Economic Review 90, 368–372.

Jasso, G., Rosenzweig, M.R., Smith, J.P., 2000. The changing skill of new immigrants to the United States: recent trends and their determinants. In: Borjas, G.J. (Ed.), Issues in the Economics of Immigration. University of Chicago Press, Chicago, IL, pp. 185–226.

Jasso, G., Rosenzweig, M.R., 1995. Do immigrants screened for skills do better than family reunification immigrants? International Migration Review 29, 85–111.

Kim, S., 2013. Wage mobility of foreign-born workers in the United States. Journal of Human Resources 48, 628–658.

LaLonde, R.J., Topel, R.H., 1991. Immigrants in the American labor market: quality, assimilation, and distributional effects. The American Economic Review 81, 297–302.

LaLonde, R.J., Topel, R.H., 1992. The assimilation of immigrants in the US labor market. In: Borjas, G.J., Freemman, R.B. (Eds.), Immigration and the Workforce: Economic Consequences for the United States and Source. University of Chicago Press, Chicago, IL, pp. 67–92.

Lessem, R., 2012. Mexico-US Immigration: The Effect of Wages and Border Enforcement. Working Paper, Tapper School of Business. Carnegie Mellon University, Pittsburg, PA.

Lofstrom, M., Bohn, S., Raphael, S., 2011. Lessons from the 2007 Legal Arizona Workers Act. Public Policy Institute of California, San Francisco, CA.

Lubotsky, D., 2007. Chutes or ladders? A longitudinal analysis of immigrant earnings. Journal of Political Economy 115, 820–867.

Mehta, C., Theodore, N., Mora, I., Wade, J., 2002. Chicago's Undocumented Immigrants: An Analysis of Wages, Working Conditions, and Economic Contributions. University of Illinois, Center for Urban Economic Development, Chicago, IL.

Massey, D.S., 2004. Social and economic aspects of immigration. Annals of the New York Academy of Sciences 1038, 206–212.

Massey, D.S., Riosmena, F., 2010. Undocumented migration from Latin America in an era of rising US enforcement. The Annals of the American Academy of Political and Social Science 630, 294–321.

Meissner, D., Kerwin, D.M., Chishti, M., Bergeron, C., 2013. Immigration Enforcement in the United States: The Rise of a Formidable Machinery. Migration Policy Institute, Washington, DC.

Menjivar, C., Abrego, L.J., 2012. Legal violence: immigration law and the lives of Central American immigrants. The American Journal of Sociology 5, 1380–1421.

Ottaviano, G., Peri, G., 2012. Rethinking the effect of immigration on wages. Journal of the European Economic Association 10, 152–197.

Portes, A., Zhou, M., 1993. The new second generation: segmented assimilation and its variants. The Annals of the American Academy of Political and Social Science 530, 74–96.

Rumbaut, R.G., 1997. Assimilation and its discontents: between rhetoric and reality. International Migration Review 4, 923–960.

Salcido, O., Adelman, M., 2004. "He has me tied with the blessed and damned papers": undocumented-immigrant battered women in Phoenix, Arizona. Human Organization 63, 162–172.

Schwabish, J.A., 2011. Identifying rates of emigration in the United States using administrative earnings records. International Journal of Population Research, 2011, 1–17.

Thom, K., 2010. Repeated Circular Migration: Theory and Evidence from Undocumented Migrants (Unpublished Manuscript). Department of Economics, New York University, New York.

U.S. Government Accounting Office, 2006. Estimating the Undocumented Population, a Group Answer Approach to Surveying Foreign-Born Respondents. Washington, DC.

Valdez, Z., 2006. Segmented assimilation among Mexicans in the Southwest. The Sociological Quarterly 47, 397–424.

Yoshikawa, H., Kalil, A., 2011. The effects of parental undocumented status on the developmental contexts of young children in immigrant families. Child Development Perspectives 5, 291–297.

Immigration, Citizenship, and Integration: Social Work Connections

Ransford Danso, School of Social Work, Ryerson University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Immigrants encounter formidable challenges to integration, including exclusionary policies that deny them access to citizenship rights. Deep-level analyses of the unequal relationship between the state and immigrants as well as integration processes and outcomes would enable social workers to develop effective tools to serve the needs of immigrant communities. New social, cultural, economic, political, and legal realities in the post-9/11 era demand that migration studies be given focused attention and moved to a central position within social work. This article introduces an innovative framework to guide social work practice with immigrants.

Introduction

In an increasingly globalizing and diasporic world, migration has transformed the ethnocultural map of nations in profound ways. Countries that decades ago could be described as culturally homogeneous no longer fit that description. Emerging realities indicate that migration will retain its transformative influence in the future, creating new opportunities for social work to engage more meaningfully with migration studies as well as develop tools responsive to the needs of migrant populations.

The corpus of scholarship on globalization and international social work (e.g., Dominelli, 2010) links social work practice to immigrant integration, issues of access, and citizenship (Nash et al., 2006). Social work's connection to immigration goes back to the settlement movement that emerged in nineteenth-century England (Shier et al., 2011). Despite this long history, deep-level analyses of immigrant settlement issues have generally remained a blank spot on the social work scholarly map. This may be due to many social workers' lack of awareness of the challenges immigrants face and the fluid environments in which immigrants navigate their integration today. Lack of training and the feelings of disengagement many social workers experience with immigrant clients exacerbate the problem (Bhuyan et al., 2012). This article offers a synopsis of states' responses to immigration and the role social workers play in policing immigration laws. States' use of neoliberal policies to disenfranchise and exclude immigrant populations is highlighted for the purpose of illustration. I then propose an innovative theoretical framework to guide empowerment social work practice with immigrants.

The State, Immigration, and Social Work

States exploit immigrants' vulnerable position to exclude them from membership by erecting structural hierarchies of belonging and citizenship (Dunn and Jacqueline, 2011). In the post-9/11 era, state laws have led to the gradual erosion of access to social rights for immigrant populations. Even second- and third-generation immigrants, who may be living in the only country they have ever known, are denied rights and entitlements, and remain indefinitely without citizenship (Brubaker, 2010).

In Europe, political and policy discourses deny immigrants access to social and citizenship rights (Davidov and Meuleman, 2012). Restrictive policy measures implemented in Europe over the past decade have led to significant reduction in immigration levels as well as cuts to welfare rights and entitlements for newcomers (Lesińska, 2014). Some Asian countries have instituted exclusionary nation-building programs based on uniform, majoritarian forms of citizenship in defiance of calls to implement inclusive and empowering policies toward immigrant communities (Miller, 2011). Similarly, in Australia, antiforeigner attitudes toward immigrants have led to discrimination in employment and other areas (Nelson et al., 2011). Historically, Australia excluded nonwhites, especially Asians, from entering the country through its White Australia policy (Mann, 2013; Yu, 2014). Before the 1960s, Canada also sought to create a 'white' man's country by preventing the immigration of people it deemed 'undesirable.' Canada implemented a series of policies to stop Chinese people from immigrating to Canada. Beginning with a $50 'head tax' on Chinese immigrants in 1885, the racist practices culminated in the passage of the *Chinese Exclusion Act* in 1923 (Go, 2012).

Social work's relationship with the state on migration is contradictory. Social workers experience ethical dilemmas and value conflicts in which practitioners either obey state laws or violate professional values (Furman et al., 2012). While they fight discriminatory policies and practices, social workers have also been state collaborators. Many injustices are committed against immigrants as a result of social work practices that are sanctioned by government policies on the basis of nation-state boundaries (Moosa-Mitha, 2014). Social workers police state policies in the areas of child protection and welfare services (Martinez-Brawley and Zorita, 2011). Mandatory reporting laws require child protection workers to remove children from families, placing them under state care without due regard for the impact of their actions on families. Immigrant families are often represented in childcare protection in significant numbers in many countries (Putnam-Hornstein et al., 2013).

Social workers also enforce immigration laws regarding undocumented migrants (Furman et al., 2012). State laws require social service providers to report people who may have entered or remained in the country unlawfully. Since

September 2001, America and other Western countries have, under the guise of rooting out international terrorism, instituted stringent national security and immigration laws that oppress immigrant populations (Meissner et al., 2013). Social workers should play a leading role in changing the unequal relationship between the state and immigrants and debunking myths and xenophobic attitudes toward immigrants. They should also ensure that unexamined feelings and biases do not influence their work with immigrants. Practitioners should be knowledgeable about the historical role the state has played in oppressing immigrant communities. Knowledge of the historical antecedents to struggles immigrants face today would equip social workers to develop responsive interventions and influence state policies to better meet the needs of immigrants.

Neoliberalism vis-à-vis Immigrant Integration

Emerging in the 1980s, neoliberalism has become a major driving force behind immigration and social policies in many Western countries (Triadafilopoulos, 2011). Neoliberalism is an approach to social, political, and economic life that discourages publicly funded welfare services (Baines, 2011; Walsh, 2014). Neoliberalism promotes individual skill, self-enterprise, and reliance on the market system to provide individuals with the incomes required to support themselves. People who seek assistance from the government are stigmatized as living off the toil of hardworking citizens.

Many countries have instituted policies that threaten the successful integration of immigrant populations. In Western Europe, state policy on immigrant settlement and integration are increasingly converging in many areas, including the mandatory civic integration courses and tests newcomers must take (Joppke, 2007; Lesińska, 2014). In the Netherlands, prospective immigrants must pass an integration test before they can enter the country, effectively making integration a condition for admission rather than of legal residency and citizenship (Joppke, 2007, 2012). Western European countries have also pursued liberal goals "with illiberal means, making it an instance of repressive liberalism" (Joppke, 2007: p. 1). Canada's neoliberal policies have resulted in welfare retrenchment (Johnstone and Lee, 2014) and the closure of settlement agencies with a long history of serving immigrants (Pashang, 2012) as well as cuts to funding for settlement services that immigrants desperately need to establish themselves.

Exclusion through Disenfranchisement

Voting is a fundamental right and defining principle of both the practice and formal status of citizenship. Citizenship enhances participation and helps immigrants overcome socialization experiences that hinder political engagement (Just and Anderson, 2012). The franchise provides a mechanism for people who are subject to the laws of a state to participate in the creation of those laws (Rodríguez, 2010). Some countries, including Ireland, the Netherlands, and the Scandinavian countries allow noncitizens voting rights irrespective of their nationality. Any citizen of a signatory state of the European Union (EU) can vote in another EU country. New Zealand has one of the most inclusive voting rights for noncitizens; immigrants can vote in local and national elections after 1 year of legal residence (Song, 2009).

The above examples are exceptions as the majority of countries deny immigrants the right to participate in political decision-making processes (Song, 2009). In Canada, state and local laws deny noncitizen immigrants the right to vote (Siemiatycki, 2007), although they may have resided in the country for many years. Immigration policies exploit the economic and human capital of immigrants to give Canada a competitive advantage in the global economy. Similarly, cultural diversity policies showcase Canada as an egalitarian country. Yet, none of these policies actively promotes the political incorporation of immigrants, leading to low levels of political representation particularly for 'immigrants of color.'

The foregoing highlights the need for social workers never to underestimate the power of the state to decide who and how to exclude. The systemic barriers immigrants experience suggest that while globalization and advances in communication technology and transportation infrastructure have rendered national borders more porous, they have not completely incapacitated the modern nation-state. The state has lost much of its past power to control capital (Hanley, 2012) but it has not lost the capacity to regulate the flow of persons across its borders. Membership and access to citizenship have not been recast in a way that bypasses or transcends the nation-state (Brubaker, 2010). Opportunities abound for social workers to fight for inclusive policies and settlement services to facilitate the integration of immigrant populations.

Integration: a Theoretical Overview

The integration of immigrants has become a priority issue on policy and political agendas in many countries. The mix of changing immigration flows, increased ethnoracial and religious diversity, and the politics of national security have intensified debates about social integration, social cohesion, cultural diversity, the collective rights of immigrants, and the state's obligation to protect these rights and promote immigrants' cultures (Banting and Soroka, 2012).

Integration is the process by which immigrants participate fully in the social, economic, cultural, political, and legal fabric of the adopted country. For many immigrants, integration also involves adaptation to new climatic conditions (Phillips, 2010). Integration is a two-way process of engagement, if asymmetrical relationship between newcomers and established residents. Integration fosters a bidirectional flow of views, values, lifestyles, and skills that results in adaptation processes in both settled and immigrant populations (Valtonen, 2008). The newcomers desire to adapt to the new country while the receiving society and its institutions must quickly facilitate the process of making them full citizens with all attendant political and social rights.

Immigrant integration is shaped by four sets of factors: first, the personal characteristics and prior experiences of immigrants; second, the nature and cultural practices of the country of origin; third, the challenges immigrants encounter while in transit; and fourth, the contextual factors within the adopted

country, including state policies and attitudes toward foreigners. For instance, during the twentieth century, the British government developed guidelines on how refugees in Britain should adapt and behave there. They were warned not to appear too 'foreign' and not to congregate together in too conspicuous a way, but conform to a constructed form of Englishness (Manz and Panayi, 2012).

Although shaped by many factors, as integration occurs after migration, contextual factors are likely to exert the most influence on it; the receiving society provides the context within which integration is negotiated. Thus, while integration is considered a mutual process of accommodation between immigrants and the settled population, the 'burden of integration' falls more on the newcomers than on the receiving society. The receiving society exists, a priori, and the newcomers are expected to fit into it. The greatest part of the work of integration is undertaken by the immigrants themselves, although they might be assisted by the adopted country (CCR, 1998).

A Framework of Immigrant Integration

Immigrant integration is often seen as a linear event proceeding from acclimatization through adaptation to incorporation into the new society. Acclimatization is the initial process of immigrants familiarizing themselves with the new society. Immigrants' experiences with culture shock coupled with disorientation vis-à-vis local institutions may indicate the onset of the adaptation phase of integration. Successful negotiation of the acclimatization and adaptation stages ushers immigrants into the integration stage. While the three stages overlap, the speed and degree of integration vary significantly within each stage. For instance, immigrants who integrate economically would be expected to integrate socially and culturally (CCR, 1998).

Danso (2009) has developed a simplified, two-dimensional model (**Figure 1**) to illustrate the complex process of immigrant integration. The model conceptualizes integration as a contested process of conflict and negotiation between immigrants and the receiving society. Political organizing, interrogating dominant-group privileges, struggles over power and acts of resistance are considered types of conflict. Negotiation typically takes the form of advocacy, dialogue, mediation, lobbying, and partnerships. Negotiation is more likely to produce an outcome that is 'positive' while conflict might produce an outcome that is 'negative.' The nature of the conflict or negotiation and the kind of relationship between the immigrants and the receiving society would produce different integration outcomes (Danso, 2009).

To predict a likely integration outcome, the characteristics of the receiving society and the immigrant group are superimposed on each other. The vertical axis in **Figure 1** represents the receiving society and how welcoming or open it is to foreigners, while the horizontal axis represents the migrant community. Both axes range in scale from 'closed' to 'open.' This analysis recognizes the multidimensional nature of integration but concentrates on only two sets of factors: the personal characteristics and prior experiences of the immigrants and the structural characteristics of the receiving society. The model assumes that integration outcomes are largely a function of these two sets of factors.

According to the model, the site of most conflict or negotiation is the intersection of the two axes (point X). Point X is considered the site of 'indeterminacy' or 'uncertainty.' At that point, the receiving society is not sure or does not have clearly defined strategies for facilitating the incorporation of its immigrant populations. At the same point, an

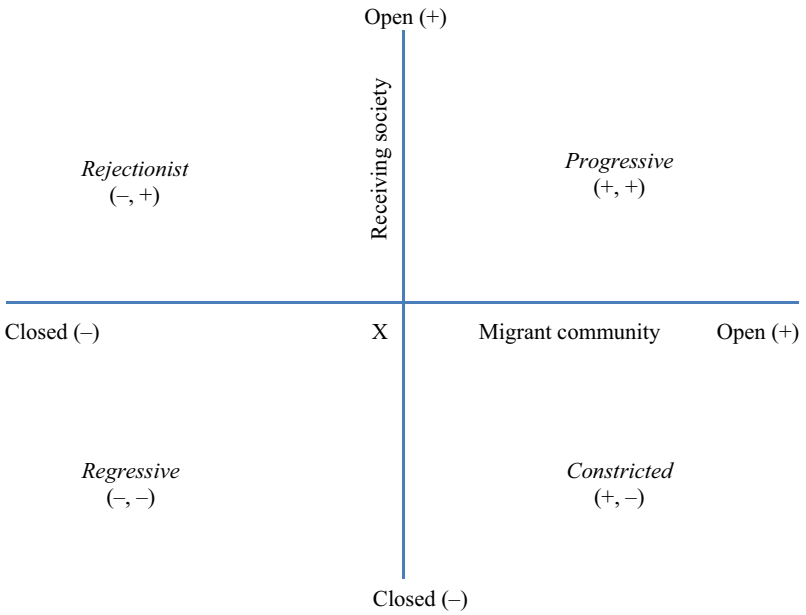

Figure 1 A framework of integration outcomes. Adapted from Danso, R., 2009. Restoring Human Functioning: Impact of Social Policy and Social Networks on the Resettlement and Integration of Refugees in Western Societies. Lambert Academic Publishing, Köln, Germany.

immigrant community is also not sure, at least at the initial stages of settlement, as to how their integration would proceed or what the outcomes would be. While X may constitute a point of no resolution for both groups, it is never a dormant site, but where most action occurs. Any activity at point X reflects a strategy adopted by the receiving society and the immigrant community, which would lead to some predictable integration outcome, and require social work intervention.

Integration Outcomes

Much social science research has focused on the economic and housing experiences of immigrants to the near-neglect of other aspects of immigrant integration. Figure 1 illustrates different outcomes that can result from different permutations of different aspects of the personal characteristics and resources of immigrants and the social structures in the receiving society.

One outcome is labeled 'progressive' or 'double-positive' integration because in responding to their respective concerns, the positive characteristics of the immigrants and the receiving society combine to ensure a healthy compromise that produces proper integration (Danso, 2009). This outcome is based on the assumptions that there are no significant systemic barriers within the receiving society to impede integration; and the personal experiences and characteristics of the immigrants predispose them to integration. Both the receiving society and the immigrants are assumed to be very 'open' such that the positive characteristics of the two groups foster progressive integration.

A second outcome, dubbed 'regressive' or 'double-negative' integration, assumes that the personal characteristics and prior experiences of the immigrants impede integration. Inability to speak the local language, lack of occupational skills, and traumatic life experiences are considered negative characteristics that can drag the integration process. When the negative characteristics and experiences of the immigrants intersect with systemic barriers within the receiving society, these would produce regressive integration outcomes for the immigrant community. In regressive integration, the newcomers and the receiving society are considered 'closed' to each other. According to the model, regressive integration is the polar opposite of progressive integration.

In situations where the immigrants are 'open' and the characteristics of the immigrant community also predispose them to integration, we should expect a 'constricted' integration outcome if the receiving society is 'closed' and less accepting of immigrants. In this case, the immigrants find that systemic barriers within the adopted country frustrate their efforts toward full membership and integration. The degree of constriction would depend on the degree of closure in the receiving society.

An outcome similar to the constricted outcome may be labeled 'rejectionist' integration. This type of integration assumes that the receiving society is 'open' and accepting of immigrants as full members and citizens. Ironically, the offer of welcome and assistance extended by the receiving society is literally rejected by the immigrants. However open the new country is, and however permitting the conditions in the country are, the 'negative' characteristics and experiences of the immigrants hamper their integration. This typically occurs when the immigrants, especially refugees, have experienced excruciating acts of violence resulting in posttraumatic stress disorder for the immigrants. Strong attachment to the past could also prevent newcomers from identifying fully with the adopted country.

As in the case of constricted integration, the degree of rejection would depend on which of the two sets of characteristics exerts the most negative influence as well as the strength of that influence. However, since the receiving society has more powers than the immigrant community, we would expect the structural barriers within the receiving society to outweigh the negative circumstances of the immigrants when there is constricted or regressive integration.

Figure 1 suggests that any outcome that departs from the progressive option necessarily constitutes 'dysfunctional integration' for immigrants. It is so labeled because there is always a negative dimension to dysfunctional integration, regardless of the socio-political, historical, or geographical context within which it occurs. Dysfunctional integration is a product of the wider processes of social exclusion, ethnocentrism, and racialized relations and is typically indexed by spatial segregation, poor labor market participation, disenfranchisement, and denial of social rights for immigrant groups. Dysfunctional integration necessitates social work intervention.

Practice Implications

The model presented above has great potential for social work practice; it positions migration and settlement as sites for social work intervention. The model can facilitate the development of theoretical ideas and practice tools for training cross-cultural and social justice-oriented social workers. It can help social workers to identify areas for advocacy, research and policy development with respect to issues such as civic engagement and access to employment and settlement services. Efficacious application of the model requires social workers to demonstrate the ability to capture the essence of complex social processes involved in immigrant integration. Critical assessment of the issues underlying integration processes would equip practitioners better to understand the nature of the relationship between the receiving society and immigrant communities. Practitioners would then be able to explain how particular integration outcomes emerge from particular social relations between the state and specific immigrant populations. The model also has the potential to inform immigration policy decision making and the design and delivery of culturally responsive settlement services.

'Idiosyncratic Consciousness' and Empowerment Social Work

I introduce the term *idiosyncratic consciousness* to denote social workers' ability to demonstrate deep awareness, acceptance and effective engagement with the 'idiosyncratic self' of the client. Idiosyncratic identities emerge from a person's cultural upbringing, the mannerisms, ways of doing things, modes of

behavior and social interaction that are unique to an individual. Idiosyncrasies are integral aspects of the self; therefore, it is contrary to the goals of empowerment social work to expect people to divest themselves of these important elements of their identity and conform to the norms and expectations of the dominant culture.

Personal idiosyncrasies can hinder the development of a good helping relationship with clients (Heydt and Sherma, 2005). However, the idiosyncratically perceptive social worker embraces them as strengths, rather than as deficiencies, to empower clients. The worker creatively and skillfully uses them to determine the most appropriate intervention approaches to pursue with the client (Diller, 2011; Sisneros et al., 2008). The inability to accept an individual's idiosyncratic self, defined here as idiosyncratic ineptitude, often leads to prejudicial attitudes and ethnocentric behaviors, producing negative client outcomes in the helping relationship.

Applied to migration and settlement, idiosyncratic consciousness requires social workers to be skillful at developing deep understanding of the integration strategies immigrants pursue to bridge the gap between their heritage cultures and dominant social culture. Immigrants engage with two contemporaneous processes: 'acculturation' and 'deculturation' (**Figure 2**), as they negotiate the entanglements of their integration. As immigrants acculturate – by adapting to the receiving culture – they also 'deculturate'; they discard those aspects of their heritage culture that could trigger a cultural tug-of-war with the receiving culture. Such conceptualization broadens the definition of idiosyncratic consciousness to encompass social workers' ability to critically assess how immigrants handle the challenges of their integration in maintaining a balance between acculturation and deculturation. Immigrants develop 'amalgamated identities,' which are the new identities formed as negotiated outcomes between the twin processes of acculturation and deculturation (see Cheng, 2014; Geleta, 2014).

Idiosyncratic consciousness also requires social workers to demonstrate cultural humility and the capacity to integrate diversity principles in practice. Cultural humility describes an interpersonal stance, in which a social worker maintains an other-oriented, rather than a self-focused, ethnocentric perspective that involves respect and lack of superiority toward an individual's cultural background, personal idiosyncrasies and life experiences (Hook et al., 2013; Isaacson, 2014). Cross-cultural humility, competencies, and skills (Diller, 2011; Sisneros et al., 2008) would enable idiosyncratically conscious practitioners to avoid the pitfalls associated with the expectation that immigrants' identities should be supplanted by or absorbed into those of the dominant group. They could also help to liberate social workers from expectations of cultural expertise about immigrants, and to actively engage immigrant communities, inclusive of their cultural differences, in integration and service delivery processes (Ortega and Faller, 2011). Idiosyncratically conscious practitioners eschew ethnocentric inclinations and instead identify ways for responding appropriately to the personal and 'cultural idiosyncrasies' of immigrants. Idiosyncratic consciousness thus extends the notion of cultural competence to include critical analyses of the processes by which immigrants forge new identities as they integrate into the adopted society.

Idiosyncratic consciousness is linked to empowerment social work practice. The empowerment 'paradigm' emerged to counter radical social work's arguments about therapeutic approaches being the only concern of social work practice (Bransford, 2011; Dominelli, 1997). Empowerment social work is emancipatory and transformative in that it strives to change social structures and relations that oppress people (Healy, 2005). It enables oppressed groups to engage with policy decision-making practices and take control of their own lives. Policies are usually developed by people whose lives are least likely to be impacted by the policies; policies are made to serve the interests and needs of the dominant group and to maintain its power (Mullaly, 2007).

Besides requiring practitioners to fight for egalitarian social relations, empowerment social work also requires that practitioners keep abreast with current developments in immigration

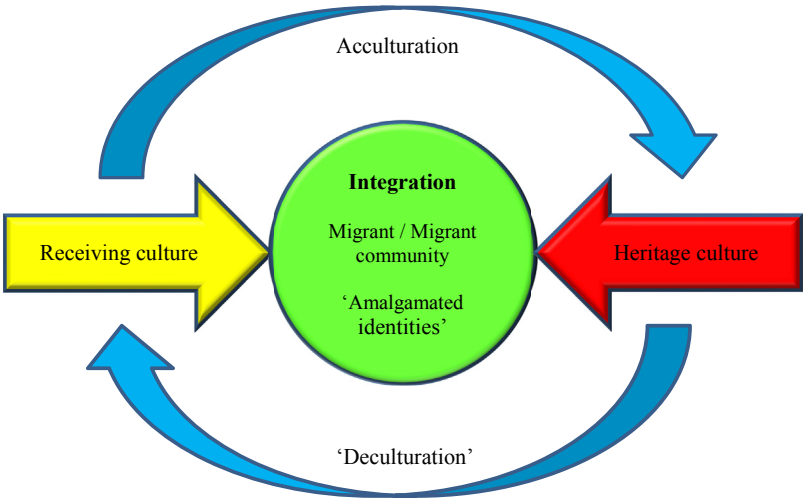

Figure 2 'Amalgamated identity' formation.

regulations so they can create responsive interventions with clients to address challenges presented by policy changes. Empowerment social workers partner with clients and social justice-oriented elements within dominant groups (Gibson, 2010; Sisneros et al., 2008) to change unjust policies and interrogate the unequal power relations that deny immigrants access to resources and opportunities.

Transformative social workers can play four key roles in the struggle for social change. Practitioners can be 'initiators' by starting a change process. Initiators do not wait to be encouraged to act; they start a change process whenever necessary. While they recognize the strength in numbers, transformative social workers do not always seek unanimous decisions on issues affecting people's lives. For transformative social workers, consensus is a necessary but not always a sufficient condition for initiating change (Chappell, 2010; Ezell, 2001). Historically, many social change processes were sparked by individuals with a vision to make a difference in society.

Another role for transformative social workers is that of a 'social catalyst.' A worker acting as a catalyst precipitates an action while also undergoing change, for instance, through constant self-reflection. Social transformation is a life-impacting process; therefore, as social workers attempt to create change they must expect to be changed by the very change they want to create. Otherwise, they perpetuate the oppressive status quo they seek to dismantle.

Transformative social workers could also function as 'sustainers' by ensuring that the change process is carried to successful conclusion. As sustainers, practitioners monitor the change efforts to ensure desired goals are achieved and necessary changes implemented as the process unfolds. Social workers can also be 'evaluators' by assessing whether the change process achieved the goal that triggered the process. Evaluation provides practitioners with an overall picture of the change effort as well as opportunities to develop more effective change strategies and plans for future projects.

Conclusion

Few processes transform human society in more profound ways than migration. Arguably, migration will continue as a powerful dynamic in national development, global social change and international relations. Post-9/11 developments at all geographical scales are recasting the state's responses to migration and integration through policy practices that impact negatively upon access to citizenship and inclusion. These developments demand that social work jettison its tenuous relationship with migration and settlement and move migration studies to a central position within the discipline. The emerging realities present challenges that require critical analyses to produce effective tools for addressing the challenges facing immigrant populations. At the same time, they open up vast frontiers of opportunities for social work to make distinctive, path-breaking contributions to migration scholarship. This article argues for the recognition of migration, citizenship, and integration as important sites for social work intervention.

See also: Ethnicity and Migration in Europe; Globalization: Welfare and Social Work; Immigration: Social Psychological Aspects; Migration and Social Work; Neoliberalism; Oppression; Social Action and Social Justice; Social Work Theory; Social Work with Racialized Groups: Frameworks for Practice; Trafficking of Women and Children in Latin America.

Bibliography

Baines, D., 2011. An overview of anti-oppressive social work practice: neoliberalism, inequality, and change. In: Baines, D. (Ed.), Doing Anti-oppressive Practice: Social Justice Social Work, second ed. Fernwood Publishing, Halifax, pp. 25–47.

Banting, K., Soroka, S., 2012. Minority nationalism and immigrant integration in Canada. Nations and Nationalism 18 (1), 156–176.

Bhuyan, R., Park, Y., Rundle, A., 2012. Linking practitioners' attitudes towards and knowledge of immigrants with their social work education. Social Work Education 31 (8), 973–994.

Bransford, C.L., 2011. Reconciling paternalism and empowerment in clinical practice: an intersubjective perspective. Social Work 56 (1), 33–41.

Brubaker, R., 2010. Migration, membership, and the modern nation-state: internal and external dimensions of the politics of belonging. Journal of Interdisciplinary History 41 (1), 61–78.

CCR (Canadian Council for Refugees), 1998. Best Settlement Practices: Settlement Services for Refugees and Immigrants in Canada. Canadian Council for Refugees, Montreal.

Chappell, R., 2010. Social Welfare in Canadian Society, fourth ed. Nelson Education, Toronto.

Cheng, Y.J., 2014. Bridging immigration research and racial formation theory to examine contemporary immigrant identities. Sociology Compass 8 (6), 745–754.

Danso, R., 2009. Restoring Human Functioning: Impact of Social Policy and Social Networks on the Resettlement and Integration of Refugees in Western Societies. Lambert Academic Publishing, Köln, Germany.

Davidov, E., Meuleman, B., 2012. Explaining attitudes towards immigration policies in European countries: the role of human values. Journal of Ethnic and Migration Studies 38 (5), 757–775.

Diller, J.V., 2011. Cultural Diversity: A Primer for the Human Services, fourth ed. Brooks/Cole, Belmont, CA.

Dominelli, L., 1997. Sociology for Social Work. Palgrave Macmillan, Basingstoke.

Dominelli, L., 2010. Social Work in a Globalizing World. Polity Press, Cambridge, UK.

Dunn, K., Jacqueline, K.N., 2011. Challenging the public denial of racism for a deeper multiculturalism. Journal of Intercultural Studies 32 (6), 587–602.

Ezell, M., 2001. Advocacy in the Human Services. Brooks/Cole, Belmont, CA.

Furman, R., Ackerman, A.R., Loya, M., Jones, S., Negi, N., 2012. The criminalization of immigration: value conflicts for the social work profession. Journal of Sociology and Social Welfare 39 (1), 169–185.

Geleta, E.B., 2014. The politics of identity and methodology in African development ethnography. Qualitative Research 14 (1), 131–146.

Gibson, D.M., 2010. Advocacy counseling: being an effective agent of change for clients. In: Erford, B.T. (Ed.), Orientation to the Counseling Profession: Advocacy, Ethics, and Essential Professional Foundations. Pearson Education, Boston, MA, pp. 340–358.

Go, A., 2012. A race-based analysis of Canada's immigration policy. In: Pashang, S. (Ed.), Unsettled Settlers: Barriers to Integration. de Sitter Publications, Whitby, ON, pp. 13–27.

Hanley, D., 2012. Squeezed from above and pressured from below: the nation state in the era of globalisation. Journal of Contemporary European Studies 20 (4), 419–421.

Healy, K., 2005. Social Work Theories in Context: Creating Frameworks for Practice. Palgrave Macmillan, Basingstoke.

Heydt, M.J., Sherma, N.E., 2005. Conscious use of self: tuning the instrument of social work practice with cultural competence. The Journal of Baccalaureate Social Work 10 (2), 25–40.

Hook, J.N., Davis, D.E., Owen, J., Worthington Jr, E.L., 2013. Cultural humility: measuring openness to culturally diverse clients. Journal of Counseling Psychology 60 (3), 353–366.

Isaacson, M., 2014. Clarifying concepts: cultural humility or competency. Journal of Professional Nursing 30 (3), 251–258.

Johnstone, M., Lee, E., 2014. Branded: international education and 21st-century Canadian immigration, education policy, and the welfare state. International Social Work 57 (3), 209–221.
Joppke, C., 2007. Beyond national models: civic integration policies for immigrants in Western Europe. West European Politics 30 (1), 1–22.
Joppke, C., 2012. The Role of the State in Cultural Integration: Trends, Challenges, and Ways Ahead. Migration Policy Institute, Washington, DC.
Just, A., Anderson, C.J., 2012. Immigrants, citizenship and political action in Europe. British Journal of Political Science 42 (3), 481–509.
Lesińska, M., 2014. The European backlash against immigration and multiculturalism. Journal of Sociology 50 (1), 37–50.
Mann, J., 2013. "Leavening British traditions": integration policy in Australia, 1962–1972. Australian Journal of Politics & History 59 (1), 47–62.
Manz, S., Panayi, P., 2012. Refugees and cultural transfer to Britain: an introduction. Immigrants and Minorities 30 (2/3), 122–151.
Martinez-Brawley, E.E., Zorita, P.M.-B., 2011. Immigration and social work: contrasting practice and education. Social Work Education 30 (1), 17–28.
Meissner, D., Kerwin, D.M., Chishti, M., Bergeron, C., 2013. Immigration Enforcement in the United States: The Rise of a Formidable Machinery. Migration Policy Institute, Washington, DC.
Miller, M.A., 2011. Ethnic minorities in Asia: inclusion or exclusion? Ethnic and Racial Studies 34 (5), 751–761.
Moosa-Mitha, M., 2014. Using citizenship theory to challenge nationalist assumptions in the construction of international social work education. International Social Work 57 (3), 201–208.
Mullaly, B., 2007. The New Structural Social Work, third ed. Oxford University Press, Don Mills, ON.
Nash, M., Wong, J., Trlin, A., 2006. Civic and social integration: a new field of social work practice with immigrants, refugees, and asylum seekers. International Social Work 49 (3), 345–363.
Nelson, J., Dunn, K.M., Paradies, Y., 2011. Australian racism and anti-racism: links to morbidity and belonging. In: Mansouri, F., Lobo, M. (Eds.), Migration, Intercultural Relations and Social Inclusion. Ashgate Publishing, Surrey, pp. 159–175.
Ortega, R.M., Faller, K.C., 2011. Training child welfare workers from an intersectional cultural humility perspective: a paradigm shift. Child Welfare 90 (5), 27–49.
Pashang, S., 2012. Unsettled settlers. In: Pashang, S. (Ed.), Unsettled Settlers: Barriers to Integration. de Sitter Publications, Whitby, ON, pp. 1–10.
Phillips, D., 2010. Minority ethnic segregation, integration and citizenship: a European perspective. Journal of Ethnic and Migration Studies 36 (2), 209–225.
Putnam-Hornstein, E., Needell, B., King, B., Johnson-Motoyama, M., 2013. Racial and ethnic disparities: a population-based examination of risk factors for involvement with child protection services. Child Abuse & Neglect: The International Journal 37 (1), 33–46.
Rodríguez, C.M., 2010. Noncitizen voting and the extraconstitutional construction of the polity. International Journal of Constitutional Law 8 (1), 30–49.
Shier, M.L., Engstrom, S., Graham, J.R., 2011. International migration and social work: a review of the literature. Journal of Immigrant and Refugee Studies 9 (1), 38–56.
Siemiatycki, M., 2007. Invisible city: immigrants without voting rights in urban Ontario. Our Diverse Cities 4 (Fall), 166–168.
Sisneros, J., Stakeman, C., Joyner, M.C., Schmitz, C.L., 2008. Critical Multicultural Social Work. Lyceum Books, Inc., Chicago, IL.
Song, S., 2009. Democracy and noncitizen voting rights. Citizenship Studies 13 (6), 607–620.
Triadafilopoulos, T., 2011. Illiberal means to liberal ends? Understanding recent immigrant integration policies in Europe. Journal of Ethnic and Migration Studies 37 (6), 861–880.
Valtonen, K., 2008. Social Work and Migration: Immigrant and Refugee Settlement and Integration. Ashgate Publishing Ltd, Aldershot.
Walsh, J.P., 2014. Watchful citizens: immigration control, surveillance and societal participation. Social and Legal Studies 23 (2), 237–259.
Yu, N.G., 2014. Ableism and economic rationalism in Australian immigration. International Journal of Social Welfare 23 (3), 254–261.

Immigration, Emigration, and Citizenship

David Cook-Martín, Grinnell College, Grinnell, IA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The global expansion of capitalism and the concurrent rise of an international system of nation-states are the backdrop against which this article defines emigration, immigration, and citizenship. The article identifies the major features of modern mass migration, outlines an analytic approach to avoid the pitfalls of nationally based biases in studying immigration, emigration, and citizenship, and presents a case study to illustrate the benefits of this approach.

Migration

Migration refers to the movement of people from one place to another, for a time and a distance long enough to significantly disrupt links to the original point of departure. It is the fundamental process underlying immigration and emigration. Citizenship or political membership becomes salient when people cross borders and hence is often examined in discussions of immigration and emigration. A consideration of migration, especially over the last two centuries, is, therefore, a necessary backdrop to defining immigration, emigration, and citizenship.

The process of migration has been driven by the search for daily sustenance and reproductive partners since *homo sapiens* spread from Africa throughout the Eastern Hemisphere beginning about 60 000 years ago (Manning, 2005). By modern standards, these early migrations consisted of small movements that happened over long periods of time. Notwithstanding differences of size and rate, migrations then and in more recent years continue to have in common the drive to find resources. The advent of modern capitalism, however, has driven and transformed migration into the mass phenomenon with which we are familiar today. Mass migration, in turn, has changed social organization, economic exchange, culture, and politics in sending and receiving contexts.

Mass Migrations

Geographic mobility has been a fundamental human activity, but large demographic movements are a relatively modern development. The scope of migrations that started in the fifteenth century was unprecedented in world history. Columbus' voyage marked the beginning of modern, large-scale migration. In the first three centuries, after he sailed from Europe, about two million Europeans crossed the Atlantic and settled in the Americas (Manning, 2005) as part of processes of colonization and the search for better economic prospects. In the same period, traffickers brought eight million Africans, mostly as slaves, to the Western Hemisphere.

Driven by the expansion of capitalism, massive movements of people accelerated dramatically between the 1850s and 1950s, and were comprised of three main circuits. First, about 58 million Europeans and 2.5 million Asians and Africans migrated to the Americas. Second, about 52 million migrants moved from India and southern China to Southeast Asia, the Indian Ocean Rim, and the South Pacific. Another four million moved from Africa, Europe, Northeastern Asia, and the Middle East to Southeast Asia, the Indian Ocean Rim, and the South Pacific. Finally, somewhere between 46 and 51 million people migrated from Northeastern Asia and Russia to Manchuria, Siberia, central Asia, and Japan (McKeown, 2004).

Since World War II, global migration has increased steadily as a result of several factors: population growth (an almost threefold increase in the 60 years preceding 2009), relaxed migration controls between the 1960s and 1990s in major destination countries, refugee flows, and the expansion of the global economy (Moya and McKeown, 2011; see also Castles et al., 2014 for an overview). Absolute numbers of contemporary migrants are from double to triple, what they were in the 1910s and late 1920s. Migration *rates*, however, have been roughly similar to those of the previous 'age of migration.' Compared to migrations before World War II, postwar migrations have been more regulated, women have become nearly half of all international migrants, and the sources of migration have changed. Post-World War II migration has been more global in origin, as only the major destinations were not sending significant numbers of emigrants to other regions.

Global Revolutions

Four global transformations joined to produce mass migration through the 1940s (Moya, 1998). The *demographic 'revolution'* resulted in population growth in Europe, despite the departure of many of that continent's natives. In the economic domain, the combined effect of the *agricultural and industrial revolutions* created surplus labor in some regions and demand for workers in others. As McKeown (2004: 155) notes, "the rice fields and rubber plantations of Southeast Asia were as much a part of the industrial processes transforming the world as the factories of Manchester and the wheat fields of North America."

A *political revolution* resulted in the consolidation of an international system of nation-states, which interacted with a widespread acceptance of the freedom to emigrate to produce legal and organizational structures that, however, variable in the capacity to actually control people flows, affirmed universally accepted ideas and practices with long-lasting consequences. The idea that nation-states should have sovereignty over their borders and be able to select who entered the national territory or its community, for instance, became a fundamental political tenet by the nineteenth century. This

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86150-4

notion had been preceded, however, by the 'exit revolution,' which reflected a consensus on the part of states that emigration could serve their interests and hence should not be restricted as it had into the early nineteenth century (Zolberg, 2007). Allowing exit had become a useful tool to relieve domestic pressures, first as a form of benign neglect and later a liberal policy toward population movement. Those who could vote with their feet, local officials who felt firsthand the pressures of demographic growth and labor demands, and free-market liberal politicians and economists, all supported the exit revolution. In 1922, the International Labor Organization noted in its reaction against post-World War I restrictions on migration from Europe to the Americas that "during the whole of the nineteenth century … [a]lmost all countries kept an open door both for emigrants and immigrants" (1922: 13). The truth was more complicated, since major receiving countries in the Americas and in English-speaking settler states selected immigrants by origin, and sending states like Japan and China sought to avoid humiliation on the global stage by acquiescing to the pressures of receiving states to control emigration (FitzGerald and Cook-Martín, 2014).

Finally, a revolution in transportation technologies changed the scope and reach of migration. The advent of steamships and trains that could transport the inputs of emerging industrial capitalism – coal, iron, timber, cotton, wheat – and workers to harvest, collect, or move these commodities, spurred larger movements of people over longer distances than ever before. Refrigeration technologies in tandem with new modes of transportation could, for instance, make possible the delivery of perishable commodities from the Southern Cone to markets in the major industrializing cities of England. The rise of transportation as an industry with its own political interests also drove mass migration and illustrates the interconnection among the global transformations described (Feys et al., 2013).

These four revolutions were, in fact, entwined and mutually reinforcing. Population growth put pressure on local resources often because of institutional property arrangements. Local authorities had to contend with a growing number of residents unsure of how to make a living under new circumstances compounded by the impact of capitalist modes of production. Authorities relaxed constraints on emigration and pressed national authorities to reflect this approach in policy. A political culture of free movement ensued. Shipping companies, for example, in Hamburg and Bremen, viewed emigrant flows as an important source of revenue, and in turn became political actors in their own right.

Features of Modern Mass Migrations

An important concomitant of these revolutions has been the rise of cities as hubs of domestic and international migrations. Migration has been part and parcel of *urbanization*. Three percent of the world's population lived in cities in 1800, but this proportion rose to 14% by 1900, 30% by 1950, and over 50% at the time of writing (Moya and McKeown, 2011). Cities have been the launching sites and ports of arrival for mass migration over the last two centuries; even if countries are the typical unit of analysis.

Migrations since the mid-nineteenth century have primarily been free rather than coerced, as they had been in the case of the chattel slavery system that predominated beginning in the late seventeenth century (Castles et al., 2014) or the trade in Chinese indentured laborers that emerged in the wake of abolition (Meagher, 2008). 'Freedom,' however, has consisted of (1) severing historical ties to local communities and a commons such that people were free, in the Marxian sense, to move elsewhere in search of other means of livelihood and (2) an ideological accomplishment of capitalism that fashioned subjects as workers who transacted the sale of their labor with employers through contracts entered on putatively equal footing. From prospective migrants' point of view, freedom consisted of having opportunities to find work.

Migrations have been not only among countries, but also *domestic*. Rural to urban migrations within national borders often preceded large migrations from Europe to the Americas, although migrants also moved directly from rural areas in their countries of origin to urban areas in destination countries. Migration has also happened within major migration receiving countries. For instance, after emancipation but during the height of Jim Crow segregation, a movement of blacks began from the south to the rest of the United States. All told, this involved a net migration of over six million people by 1970 (Tolnay, 2003). This particular movement demonstrates the limits of the political ideology of free movement – since white southerners waged a daily struggle to prevent the defection of black workers in search of the 'warmth of other suns' (Wilkerson, 2011) – but also similarities among domestic and international migration. And while these movements were partly motivated by a comparison of labor prospects in the south and elsewhere, the extreme castelike system of the Jim Crow south undoubtedly moved blacks to vote with their feet. Another example of domestic migration is the movement of the U.S. mid-Westerners from the 'Dust Bowl' to California in the 1930s and 1940s, famously depicted in John Steinbeck's novel *The Grapes of Wrath*.

Circularity is another feature of modern mass migration. Seasonal differences, for instance, allowed Italian migrants to work in harvests in the Northern (Europe) and Southern Hemispheres (the Southern Cone) on a cyclical basis. The intentions of even these 'sparrow' workers changed and they settled in Argentina and Brazil. The cycle migration to a new destination, return to a sending country, and remigration to a new destination can also be stretched out over years in the life cycle of an individual or family, or even across generations (Bhachu, 1985). The tightening of border controls since the 1990s and especially since 2001, however, has significantly impacted seasonal and circular flows.

Migration movements have also been deeply *gendered* (Calavita, 2006; Cook-Martín, 2006). To the extent that migration decisions have been made at the household rather than the individual level, as suggested by 'new economics of migration' theories (Massey et al., 1993), they have reflected heteronormative and patriarchal ideas and strategies. In pre-1945 migration, immigration policies gave these ideas the force of law often defining the category 'immigrant' as a male head of household. Emigration policies put different constraints on the departure of young men (e.g., completion of military service obligations) and women (e.g., through morality protection laws that forced them to be accompanied by a male head of household). The postwar feminization of

migration has been linked by scholars to neoliberal market reforms, the massive movement of women into the workplace in the Global north, and implicitly, men's unwillingness to contribute labor in the household economy (Salazar Parreñas, 2005). More recently, legally inscribed conventional gender categories and sexualities have been challenged by the migration of same-sex partners and of individuals who defy traditional expressions of sexuality (Luibhéid, 2008).

Immigration, Emigration, and Citizenship

The preceding overview of migration is a necessary precursor to defining immigration, emigration, and citizenship. If not for the global revolutions described earlier – especially the consolidation of an international system of nation-states as an organizational reality as well as an idea – immigration, emigration, and citizenship would not resonate as they do with scholars and the public at large. *Immigration* refers to entering and settling in a national territory, whatever the intention of people engaging in this process. *Emigration* refers to departure from a national territory to a specific destination.

These terms are not merely descriptive of the direction of movements across national territories, but have a strong moral valence. From the perspective of historical destination countries, immigration has been a primary means of nation making and, therefore, a process to be carefully regulated since it affects the character and prospects of the putative imagined community (Anderson, 1991). The assumption on the part of officials in receiving countries is that desirable immigrants will want to stay and that their settlement is permanent. From the perspective of sending countries, emigration has challenged nationalist assumptions that citizenry and territory should match closely. A response has often been to construct notions of the nation that transcend national borders like the Greater Italy (Choate, 2008) or a Salvadoran or Mexican nation of emigrants (Bibler Coutin, 2007; FitzGerald, 2009). The assumption on the part of state agents in the sending country is that their citizens are abroad for a time and that they will return.

Citizenship refers to the legal link between a state organization and an individual – which entails rights and obligations – and the affective tie between individuals and a nation (Cook-Martín, 2013). Migration has challenged normative ideas of citizenship that see it as a permanent and enduring link. In the contemporary world, citizenship may have more than one state and national referent.

Immigration, emigration, and citizenship, however, issue from the foundational assumption of a world organized into discrete states whose territories and people, more or less match. The assumption is so strong that even self-aware analysts have trouble escaping what Gerard Noiriel (1991) has called the 'tyranny of the national' or the tendency to view migration from the vantage point of a particular state.

A transnational political field approach offers a corrective to nation-based assumptions that shape scholarship on migration. A 'political field' refers to a social space in which individual and collective actors compete over valued symbolic or material resources, such as an imagined nation, workers, or potential citizens (cf Bourdieu, 1991). 'Transnational' refers to an analytic global view rather than one privileging the perspective of a single player in the field. Political fields change over time as organizations develop and institutionalize competitive strategies and hence it is also critical to take a long view. A transnational approach is a useful heuristic device to define and understand the concepts of immigration, emigration, and citizenship. This interpretive framework yields three key insights.

First, terms like immigration and emigration that refer to the direction of migration flows often hide nationalist assumptions. One country's immigration is another's emigration. Moreover, today's immigration may be the source of tomorrow's emigration, depending on changes in economic and network factors causing and/or sustaining migration, as well as on the social, political, and institutional ties that bind people at each end of a particular flow. For instance, the legal and organizational frameworks developed by governments to manage migration at one moment in time can facilitate or hinder future migration often in unanticipated ways. When Italian policy makers at the turn of the twentieth century fashioned nationality laws that extended citizenship over several generations, they created a framework that allowed the descendants of Italians in the Americas to 'reclaim' Italian citizenship after several decades (Cook-Martín, 2013).

Second, analysts are more like to assume that assimilation should be the desired outcome for immigrants if they adopt the standpoint of a particular country rather than a global perspective. The flip side of this problematic view is when sending countries assume that the ties that bind them to 'their' citizens abroad are unchanging and enduring, sometimes over several generations. In the first instance, policy makers, the public, and some scholars view assimilation – or in the European version, integration – as a mostly straightforward and unidirectional process that is inevitable given enough time and facilitating conditions in the receiving context. A healthy debate in the literature on assimilation suggests outcomes are to be empirically explored rather than assumed (Alba and Nee, 2003; Alba, 2005; Brubaker, 2001). The assumptions of sending states with respect to the ties that bind emigrants and their descendants have been explored, but not as systematically.

Finally, the dynamics revealed by a transnational political field approach and illustrated by the examples cited suggest that links among people at two ends of a migration flow cannot be presumed to transcend a particular nation nor can identifications be assumed to endure over time. These phenomena require empirical verification, especially of the causal relationship between formal laws and a sense of belonging. Scholars continue to explore whether laws recognize and consolidate existing affective identifications or whether identifications drive adherence to naturalization and citizenship acquisition (e.g., Bloemraad, 2006; Cook-Martín, 2013).

Migration and Selectivity

A transnational political field approach recommends awareness of the politics surrounding terms like immigration, emigration, and citizenship. A case study of how immigration and nationality laws have been instruments to select 'good' workers and citizens illustrates the promise of this analytical approach. In the Americas, for most of the last 200 years,

desirable workers and citizens have been selected by ethnicity. Racist laws have been more common among countries with greater political inclusion, whether democratic or populist (FitzGerald and Cook-Martín, 2014). Taking a long view of economic and ideological factors on a domestic plane – comprising interest group struggles within a country – as these intersect with those on an international plane – including struggles between and across countries – it becomes clear that there is no iron law linking ethnically selective laws to ideologies like liberal democracy and racism, or to domestic struggles among labor and capital.

On the ideological front, conventional wisdom has maintained that racism has no place in liberal thought and institutions (de Tocqueville, 1994; Higham, 1994; Freeman, 1995; Joppke, 2005). Historical instances of racism in law and policy have been characterized as aberrations to be worked out of the democratic body politic or vestiges of earlier, more primitive social arrangements like slavery (Fukuyama, 1989).

While the received wisdom may sound appealing to those who are philosophically sympathetic to the tenets of liberal democracy, it is empirically wrong and fails to explain the emergence and decline of racist immigration laws. The United States, the main global exemplar of liberal democracy, took the lead in excluding blacks and Asians beginning with its nationality laws in 1790 and immigration laws in 1803.The U.S. laws limited naturalization to whites and banned Chinese immigration *because* it was a democracy. Policy makers in the United States, most notably U.S. President Theodore Roosevelt, argued that whiteness indicated capacity for self-government.

Moreover, undemocratic regimes in Cuba, Argentina, Chile, Uruguay, Paraguay, and Mexico reversed discriminatory immigration and nationality laws and pioneered the explicit deracialization of immigration policies beginning in the late 1930s, a generation or more before Canada, the United States, Australia, and New Zealand. Racial equality has not been inherent to liberalism and is especially fragile in democratic and populist political systems.

Critical race theorists reject the premise of a liberal democracy whose practices have not caught up to its ideals. On this view, much more consistent with historical realities of the last centuries, racism and liberalism have been mutually constitutive, fully compatible, and clearly conjoined in practice (e.g., Mills, 1997, 2000). Yet this view cannot explain the decline of ethnically selective laws and the massive ethnic transformations of countries like the United States and Canada that were historically among the most racially selective. It also misconstrues the factors that lead to the decline of racially selective laws around the time of World War II.

Liberalism was not inherently racist, but neither was it inherently antiracist. Historically its immigration and nationality laws have had a color, and it has been white. Liberalism and racism were not simply two independent and cooccurring traditions (cf Rogers Smith); they were intimately linked. The rising aspirations and decolonization of peripheral countries during the geopolitical conflicts of World War II and the Cold War eventually forced powerful white countries to forego racial discrimination in law; the dismantling of the racist regimes was not an attempt by countries with racist laws to reflect liberal democratic principles or primarily a domestic movement for equality.

On the economic front, a transnational and historically comparative view shows that class conflict has been an important but insufficient explanation for patterns of ethnic selectivity, that exclusively class-based accounts have difficulty explaining the malleability of class preferences in the same countries over time, and that there are numerous instances of economic groups supporting overly racist laws even in the absence of an economic incentive and sometimes against economic self-interest.

The pattern of racial selectivity described in the previous paragraphs becomes evident only by adopting a systemic view that captures relations in an international domain, puts domestic migration politics in that geopolitical context, and examines legal-institutional developments over the long run. A transnational political field approach that is historically relational addresses the serious methodological problem of focusing primarily on Western liberal democracies in the second half of the twentieth century, which yields a skewed picture of which countries modeled the move away from ethnic selectivity. This approach allows analysts to understand how the politics of today shape immigration and nationality policy, how the policies of yesteryear 'tick' within the policies of today, and how policies in one country are shaped by the policies of another. From this perspective, immigration and emigration policies are interactive. Indeed, countries of emigration have played an underappreciated role in shaping the policies of immigration countries (FitzGerald and Cook-Martín, 2014: 27).

Citizenship in an Interconnected World

Concerning citizenship, a systemic approach allows us to see that even nationality law, widely conceived as the domain most subject to sovereignty concerns, is shaped in an international political field that unfolds over time. Chinese were first banned from entry to the United States in the 1880s, and then from naturalization; but during World War II, when the Allies needed Chinese support, the naturalization ban was lifted. Citizenship laws in major emigration countries like Italy and Spain were instruments to counter the 'assimilatory' laws of receiving countries like Argentina, Brazil, and the United States (Cook-Martín and Viladrich, 2009).

Moreover, past configurations of migration and nationality laws can have important consequences in the present. The *jus sanguinis* nationality laws of sending countries offer a means by which the descendants of past emigrants can make claims on ancestral homeland states. Generations after the last Italians arrived in New York, Buenos Aires, Caracas, or Melbourne, their descendants can make claims to Italian citizenship. The descendants of Japanese migrants to Brazil and Peru have had access to a privileged immigration status in Japan, thanks to ethnic affinity policies (Tsuda, 2009). Indeed, the combined effect of *jus solis* and *sanguinis* policies have generated a large population eligible for a second nationality, and a potential revolution in perceptions about multiple nationalities. Over 2.5 billion people live in countries that *de facto* allow for multiple legal citizenships, a number that has grown fivefold since 1960 (Sejersen, 2008). In other words, the writing of laws by states in reaction to the movement of people has resulted in the possibility of multiple affiliations.

Countries vary significantly, of course, in the degree to which they have institutionalized control over migration as well as in their capacity to regulate flows and establish legal membership. Sadiq (2009) has shown convincingly that migrants from neighboring countries in India have made claims to legal citizenship through fraudulent use of authentic documents or through counterfeit papers. Historically, migrants in other latitudes also made claims to a status as immigrants or as citizens through similarly fraudulent means. Traffic in official documents has in fact been at the heart of a 'paper industry' to such an extent that it has caused some to observe the fetishism of papers (Hernández-León, 2013; Suárez-Navaz, 2004).

Conclusions and Implications

Immigration, emigration, and citizenship are contested, political terms that encompass far more than the direction of migratory flows or legal links between a state and an individual. A transnational political field perspective uncovers the complexity and political weight of the terms. 'Immigration' and 'emigration' subsume normative assumptions about the desirability of migrants as workers and citizens, about expected relationships and allegiances with sending and receiving communities, and about the longevity of links between states and migrants.

Moreover a transnational analytic approach has implications for explanations of migration. Scholars of migration explain the movement as the result of 'cumulative causation' by primarily economic and social network factors at points of migrant departure and destination (Massey et al., 1993). Structural- and individual-level economic factors have indeed been central to the movement of people across regions, borders, and oceans. Individuals, families, and their broader networks move in search of better economic options and as a means to diversify risk in an unpredictable capitalist system. There is more to the story, however, than the 'standard economic calculus narrative' (Hollifield and Wilson, 2011; Massey, 1999). The advent of national–territorial state organizations that fence in markets and workers is the very historical innovation that makes human mobility a subject of study. The development of a full-blown international system of sovereign states has made migration a key concern of states that feel such movements challenge their dominion (Zolberg, 1999). The very terms immigration and emigration reflect the interest of state organizations in exercising control over who comes and goes. Demography, law, and the social sciences developed in the context of state organizations and with national interests in mind.

While economic interests have been paramount, the impact of who comes and goes on the perceived community or communities associated with a state (nation or nations) has also been a key concern. Citizenship has referred to the relationship that binds individuals to particular states and their nations. This relationship has been challenged by migration and hence is inextricably bound to its understanding. As Hannah Arendt famously remarked, "sovereignty is nowhere more absolute than in matters of emigration, naturalization, nationality, and expulsion" (1973: 278). Terms like immigration and emigration say something about our interests and position relative to particular states and their respective imagined communities. A more detached transnational approach helps us make sense of the movement of people across territorial borders without succumbing to the tyranny of our particular national affiliation and affective ties.

See also: Citizenship, Anthropology of; Citizenship, Historical Development of; Citizenship, Sociological Aspects of; Citizenship: Political; Crime, Immigration, and Ethnicity; Immigration Policy; Immigration, Applied Research On; Immigration: Political Aspects; Immigration: Social Psychological Aspects; Law and Development; Law and Geography; Migration and Social Capital; Migration: Anthropological Perspectives; Migration: Cultural Aspects; Migration: Motivations; Transnational Citizenship.

Bibliography

Alba, Richard, 2005. Bright vs. blurred boundaries: second-generation assimilation and exclusion in France, Germany, and the United States. Ethnic and Racial Studies 28 (1), 20–49.

Alba, Richard, Nee, Victor, 2003. Remaking the American Mainstream: Assimilation and Contemporary Immigration. Harvard University Press, Cambridge, MA.

Anderson, Benedict, 1991. Imagined Communities: Reflections on the Origin and Spread of Nationalism, second ed. Verso, New York.

Arendt, Hannah, 1973. The origins of totalitarianism. In: A Harvest book, new ed. Harcourt Brace Jovanovich, New York.

Bhachu, Parminder, 1985. Twice migrants: East African Sikh settlers in Britain. Tavistock Publications, London.

Bloemraad, Irene, 2006. Becoming a Citizen: Incorporating Immigrants and Refugees in the United States and Canada. University of California Press, Berkeley.

Bourdieu, Pierre, 1991. Language and Symbolic Power. Harvard University Press, Cambridge, MA.

Brubaker, Rogers, 2001. The return of assimilation? Changing perspectives on immigration and its sequels in France, Germany, and the United States. Ethnic and Racial Studies 24 (4), 531–548.

Calavita, Kitty, 2006. Gender, migration, and law: crossing borders and bridging disciplines. International Migration Review 40 (1), 104–132. http://dx.doi.org/10.1111/j.1747-7379.2006.00005.x.

Castles, Stephen, de Haas, Hein, Miller, Mark J., 2014. The Age of Migration: International Population Movements in the Modern World, fifth ed. Guilford Press, New York.

Choate, Mark I., 2008. Emigrant Nation: The Making of Italy Abroad. Harvard University Press, Cambridge, MA.

Cook-Martín, David, 2006. Soldiers and wayward women: gendered citizenship, and migration policy in Argentina, Italy and Spain since 1850. Citizenship Studies 10 (5), 571–590.

Cook-Martín, David, 2013. The Scramble for Citizens: Dual Nationality and State Competition for Immigrants. Stanford University Press, Stanford, CA.

Cook-Martín, David, Viladrich, Anahí, 2009. The problem with similarity: ethnic affinity migrants in Spain. Journal of Ethnic and Migration Studies 35 (1), 151–170.

Coutin, Susan Bibler, 2007. Nations of Emigrants: Shifting Boundaries of Citizenship in El Salvador and the United States. Cornell University Press, Ithaca.

Feys, Torsten, International Maritime Economic History Association, 2013. The Battle for the Migrants: The Introduction of Steamshipping on the North Atlantic and Its Impact on the European Exodus, Research in Maritime History. International Maritime Economic History Association, St. John's, NL.

FitzGerald, David, 2009. A Nation of Emigrants: How Mexico Manages Its Migration. University of California Press, Berkeley.

FitzGerald, David S., Cook-Martín, David, 2014. Culling the Masses: The Democratic Origins of Racist Immigration Policy in the Americas. Harvard University Press, Cambridge, MA.

Freeman, Gary P., 1995. Modes of immigration politics in liberal democratic states. International Migration Review 29 (4), 881–902.

Fukuyama, Francis, 1989. The end of history? The National Interest 16, 3–18.

Hernández-León, Rubén, 2013. The Migration Industry: Brokering Mobility in the Mexico-US Migratory System. UCLA, Los Angeles.
Higham, John, 1994. Strangers in the Land: Patterns of American Nativism, 1860–1925. Rutgers University Press, New Brunswick, NJ.
Hollifield, James, Wilson, Carole, 2011. Rights-based politics, immigration and the business cycle. In: Chiswick, Barry (Ed.), High-Skilled Immigration in a Global Labor Market. AEI Press, Washington, DC.
Joppke, Christian, 2005. Selecting by Origin: Ethnic Migration in the Liberal State. Harvard University Press, Cambridge, MA.
League of Nations International Labour Office, 1922. Emigration and Immigration: Legislation and Treaties. League of Nations, Geneva.
Luibhéid, Eithne, 2008. Sexuality, migration, and the shifting line between legal and illegal status. GLQ: A Journal of Lesbian and Gay Studies 14 (2–3), 289–315.
Manning, Patrick, 2005. Migration in World History, Themes in World History. Routledge, New York.
Massey, Douglas, 1999. International migration at the dawn of the twenty-first century: the role of the state. Population and Development Review 25 (2), 303–322.
Massey, Douglas S., Arango, Joaquin, Hugo, Graeme, Kouaouci, Ali, Pellegrino, Adela, Edward Taylor, J., 1993. Theories of international migration: a review and appraisal. Population and Development Review 19 (3), 431–466.
McKeown, Adam, 2004. Global migration, 1846–1940. Journal of World History 15 (2), 155–189.
Meagher, Arnold J., 2008. The Coolie Trade: The Traffic in Chinese Laborers to Latin America 1847–1874. Xlibris.
Mills, Charles W., 2000. Race and the social contract tradition. Social Identities: Journal for the Study of Race, Nation, and Culture 6 (4), 441–462.
Mills, Charles W., 1997. The Racial Contract. Cornell University Press, Ithaca, NY.
Moya, José C., McKeown, Adam, 2011. World Migration in the Long Twentieth Century, Essays on Global and Comparative History. American Historical Association, Washington, DC.
Moya, José C., 1998. Cousins and Strangers: Spanish Immigrants in Buenos Aires, 1850–1930. University of California Press, Berkeley.
Noiriel, Gérard, 1991. La tyrannie du national le droitd'asile en Europe, 1793–1993, Histoire: Histoire (Calmann-Lévy (Firm)). Calmann-Lévy, Paris.
Parreñas, Rhacel Salazar, 2005. Children of Global Migration: Transnational Families and Gendered Woes. Stanford University Press, Stanford, CA.
Sadiq, Kamal, 2009. Paper Citizens: How Illegal Immigrants Acquire Citizenship in Developing Countries. Oxford University Press, Oxford, New York.
Sejersen, T., 2008. "I vow to thee my countries" – the expansion of dual citizenship in the 21st century. International Migration Review.
Suárez-Navaz, Liliana, 2004. Rebordering the Mediterranean Boundaries and Citizenship in Southern Europe, New Directions in Anthropology: New Directions in Anthropology. Berghahn Books, New York.
Tolnay, Stewart, 2003. The African American "great migration" and beyond. Annual Review of Sociology, 209–232.
Tsuda, Takeyuki, 2009. Diasporic Homecomings: Ethnic Return Migration in Comparative Perspective. Stanford University Press, Palo Alto.
de Tocqueville, Alexis, 1994. Democracy in America. Knopf, New York.
Wilkerson, Isabel, 2011. The warmth of other suns: the epic story of America's Great Migration. Vintage Books, New York.
Wilkerson, Isabel, Miles, Robin, 2011. The Warmth of Other Suns: The Epic Story of America's Great Migration. Brilliance Audio, Sound Recording, Grand Haven, MI.
Zolberg, Aristide R., 1999. Matters of state: theorizing immigration policy. In: Hirschman, Charles, Kasinitz, Philip, DeWind, Josh (Eds.), The Handbook of International Migration: The American Experience. Russell Sage, New York, pp. 71–93.
Zolberg, Aristide R., 2007. The exit revolution. In: Green, Nancy, Weil, Francois (Eds.), Citizenship and Those Who Leave: The Politics of Emigration and Expatriation. University of Illinois Press, Urbana.

Immigration: Health and Mortality

Robert A Hummer and Jennifer E Melvin, University of Texas at Austin, Austin, TX, USA
Monica He, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Over the past 50 years, the immigrant population in the United States and many other high-income countries grew rapidly and became increasingly diverse. Consequently, the study of immigrant health and mortality has become an important area of social science and epidemiological research. Immigrants tend to be healthy at time of arrival, but favorable health profiles tend to erode with increased duration in host countries and continue to deteriorate in the second generation. There is, however, considerable heterogeneity in these overall patterns.

Introduction

The study of immigrant health and mortality has never been as important and interesting as it is now. The most straightforward reason is that migration has increased tremendously over the past half-century, thus creating larger immigrant populations living in many countries throughout the world. For example, the most recent US population estimate suggests that 40.4 million individuals are immigrants, accounting for over 13% of the total population; the next largest immigrant population is in Russia, with just over 12 million immigrants (Pew Hispanic Center, 2013). Overall, it is estimated that 215.8 million people in the world are now living outside of the country in which they were born (Migration and Remittances Factbook, 2011). But beyond the issue of the increasingly high volume of migration, the study of immigrant health and mortality is more compelling than ever because: (1) immigrants are generally a very healthy group of people upon arrival in their new country; (2) immigrants face a challenging process of adjustment to life in a new country that is at the same time potentially harmful and beneficial to their health; and (3) immigrants are a very diverse group, characterized by differences in country/region of origin, skin color and racial identity, socioeconomic status, type of visa, legal and citizenship status, age of entry, duration in the host country, geographic location in the host country, and more. As such, the study of immigrant health and mortality focuses on diverse groups from a wide range of sending countries and measures their health and mortality patterns both shortly after arrival and throughout the life course.

This article reviews basic definitions that are important in this area of study, briefly examines the historical development of immigrant health scholarship, discusses methodological issues and challenges in studies of immigrant health and mortality, reviews current knowledge in the area, and closes with a discussion of future research directions. We focus our discussion on immigrant health and mortality in the US, both because of our own expertise and because the rapidly increasing number of immigrants living in the US since 1965 has resulted in an explosion of immigrant health research in this country. We also incorporate literature on immigrant health and mortality from outside the US as appropriate. Our focus is largely population-based research for the whole US, rather than specific small-scale studies of immigrant health among immigrant groups in particular communities, because our intent is to provide an overview of scholarship and findings in the area rather than diverting attention to particular geographic communities or groups of immigrants. Finally, we devote our attention to mortality, physical health, and health behavior; due to space limitations, we do not address research on mental health.

Key Definitions

This area of study includes several key concepts that are used on a frequent basis and which therefore deserve definitional clarity prior to the discussion of key research findings.

Migration: A general concept that characterizes individuals or groups who make permanent changes in residence across some political or administrative boundary (Swanson and Stephan, 2004). There are a number of types of migration, of which immigration is just one.

Immigration: The permanent movement of individuals across national boundaries to live in a receiving (or host) country; such persons are termed immigrants (Brown and Bean, 2005).

Emigration: The movement of individuals who leave a particular country and cross-national boundaries to permanently live elsewhere (Brown and Bean, 2005). Emigrants are not usually a focus of research because they no longer live in the country that is being studied and it is therefore very difficult to gather data on them.

Nativity: The distinction between persons who are born outside of the country in which they are living (i.e., foreign-born) and persons who are born in the country in which they are living (i.e., native-born) (Swanson and Stephan, 2004). In this article and most literature in this area, the terms foreign-born and immigrant are used interchangeably.

Health selection: The immigration of especially healthy individuals to a destination country, as well the emigration of especially unhealthy individuals back to their home countries (Akresh and Frank, 2008).

Epidemiologic paradox: The unexpected finding in which the health and mortality patterns of US Hispanics were found to be much more similar to those of Whites than Blacks

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31058-3

although socioeconomically, the status of Hispanics was closer to that of Blacks (Markides and Coreil, 1986). The epidemiologic paradox is now often referred to as the Hispanic paradox or Hispanic health paradox.

Immigrant incorporation: The extent to which immigrants are socially, economically, and culturally integrated into the host society (Treas and Batalova, 2009).

Historical Development

There is a long history of studies in public health, epidemiology, and demography that examines the health and mortality patterns of immigrants, some of which make comparisons to both nonmigrants in the origin countries and native-born persons in the host country. Kasl and Berkman (1983) provided an early review of the area. Perhaps most notable, they describe the classic Japanese–American Coronary Heart Study, which found that Japanese immigrants living in California had significantly higher rates of myocardial infarction (i.e., heart attack) and death from coronary heart disease than Japanese immigrants living in Hawaii, who in turn had significantly higher rates than Japanese individuals living in Japan. This study, as well as other similar epidemiologic studies of particular immigrants groups living in particular communities, provided some of the best early evidence that the initial health patterns of immigrants tend to be quite favorable, but less so with increased duration of residence in the host country.

As the number of US immigrants from Latin America (particularly Mexico) grew larger following the Immigration and Nationality Act of 1965, the scientific community gave increasing attention to the health of the Hispanic population. A classic paper by Markides and Coreil (1986) reviewed the literature on Hispanic health and identified more favorable health patterns and lower rates of mortality among this population than might be expected based on their low socioeconomic status; they termed this unexpected pattern the epidemiologic paradox. They further suggested that health selectivity among the immigrant portion of the Hispanic population might be a key explanatory factor for this paradox. The epidemiologic paradox has continued to be the defining theme of literature on Hispanic health in the US, and immigrant health selectivity continues to be a powerful, albeit partial, explanation for the relatively favorable health and mortality patterns of the Hispanic population (Markides and Eschbach, 2011). A great deal of research work in the US continues to focus on the health of Hispanic immigrants, as well as on their second and third generation descendants.

More recent strands of public health, epidemiologic, and demographic research have increasingly given attention to the health and mortality patterns of Asian, Middle Eastern, African, Caribbean, and Eastern European immigrants as well, reflecting the diversity of immigration to the US and other countries (Singh and Hiatt, 2006; Sole-Auro and Crimmins, 2008). This literature also has increasingly taken into account diversity in terms of the geographic settlement of immigrants, the type of visa (e.g., work, family) by which immigrants enter their new countries, and how the gender and socioeconomic status of immigrants influences their incorporation into the host society and their health behaviors and health patterns over time (Treas and Batalova, 2009). Innovative research designs also increasingly take into account data from both sending countries as well as the host society (Landale et al., 2000; Riosmena et al., 2013). In short, this literature has become progressively sophisticated over the last two decades, reflecting an increasingly diverse immigrant population living in a very complex social, legal, and health context. Recent systematic reviews of the immigrant health literature are available in several outlets (Kandula et al., 2004; Loue and Sajatovic, 2012; Treas and Batalova, 2009).

Methodological Issues

The study of immigrant health and mortality patterns is very challenging for a number of reasons. First, by definition, immigrants are people who are geographically mobile. But in addition to the immigration process itself, immigrants may also move a number of times before becoming settled in the host country; moreover, immigrants may continue to travel back and forth to their country of origin and even decide to emigrate at some point in their lives. The study of mobile populations is always more difficult than the study of nonmigrants, simply because geographic mobility makes it more difficult for researchers to survey such individuals and follow them over time. An example of geographic mobility that poses a challenge for mortality researchers is the fact that some immigrants to the US may return to their country of origin when they become elderly and/or sick. This pattern of emigration results in such individuals dying in their home countries instead of in the US. Because the deaths of such emigrants are not recorded by the US vital statistics system, US immigrants have lower mortality rates than what would actually be the case if such individuals had not emigrated. Some researchers have proposed that this selective out-migration process, termed a 'salmon bias', is the primary reason why Mexican immigrants have lower mortality rates than non-Hispanic White native-born adults in the US (Palloni and Arias, 2004). However, more recent research has demonstrated that this pattern of emigration is not nearly common enough to explain the patterns of low infant mortality (Hummer et al., 2007) or adult mortality (Turra and Elo, 2008) among Hispanic immigrants in the United States.

Second, language issues make the study of immigrant health challenging. Large-scale social and health surveys need to accommodate the various language backgrounds and skills of immigrants; if not, such surveys may either exclude some immigrant populations and/or put the validity of the survey at risk. Further, even when individuals are given language choices when answering a survey (e.g., English or Spanish), questions and response categories may be interpreted in different ways depending on the language used. A classic example is in the Spanish vs English interpretation of the response categories used in the 'self-reported health' question that is commonly asked on social and health surveys. Spanish-speaking Hispanics in the US are more likely to report 'fair' while English speaking individuals are more likely to report 'good', even when they have similar health backgrounds (Angel and Guarnaccia, 1989). Thus, researchers have become increasingly concerned with not only providing surveys in different languages, but also

careful attention to the responses to such surveys across immigrant groups who answer in different languages. More generally, difficulty in the assessment of physical health through standard survey-based research techniques has increasingly led researchers to collect health data through biomarker (e.g., blood, DNA) measures. In the immigrant health literature, Crimmins et al. (2007) showed that biomarker assessed health patterns are not fully consistent with survey assessed health patterns among Mexican immigrants in the US; biomarker assessed health for this group appears to be somewhat worse than survey assessed health.

A third methodological challenge in the immigrant health literature is the assessment of health among undocumented immigrants. Below, we briefly discuss that this is a key area for future research, given the large size (i.e., 11 million people in the US), low socioeconomic status, and precarious status of this subpopulation of immigrants. To date, the documentation of the health, health behavior, and mortality patterns of this immigrant subpopulation has been nearly impossible on a large scale because of the legal and ethical difficulties in contacting, surveying, and accurately measuring them.

Finally, a fourth methodological challenge lies in understanding and measuring the dynamic character of the immigrant population. While 'the immigrant population' has been and will continue to be defined by the fact that they are individuals who were not born in the country under study but now permanently reside there, this does not mean that this population is static. Two key points are worth considering. First, the geographic, demographic, socioeconomic, legal, and health characteristics of new immigrant arrivals is dynamic; in other words, new cohorts of immigrant arrivals have the potential to be quite different from previous cohorts – a dynamism that researchers must consider when assessing health and mortality patterns of 'the immigrant population'. Feliciano (2005), for example, provides a thorough documentation of the changing educational composition of immigrant arrival cohorts to the US during the twentieth century. Second, in a similar vein, there is no single trajectory that adequately portrays the life course of immigrants. For example, immigrants may experience an economic context in the host society that is flourishing or stagnating, or they may encounter a generally receptive or hostile receiving community/country. The dynamic characteristics of immigrants along with the dynamic context of life in the host country both need to be taken into consideration when assessing the health and mortality patterns of the immigrant population. Effectively doing so is a very tall order.

Key Findings

Immigrants face numerous challenges throughout the migration process that potentially impact their health in the host country. Even before embarking on their journey, immigrants, who most often originate from low-income countries and migrate to high-income countries, are exposed to health risks endemic to their countries of origin. Once they reach their destination, immigrants may experience difficulties accessing healthcare and other health-related resources because of barriers in socioeconomic status, legal status, and language skills. Despite these challenges, US immigrants nonetheless tend to fare better than the native-born population across a variety of health outcomes, including overall life expectancy. In the following sections, we review current evidence regarding mortality, physical health, and health behavior patterns.

Mortality

Mortality rates are important indicators of population health because they are objective and generally well measured. Literature in this area most commonly shows that immigrants exhibit lower mortality rates compared to the native-born population. Infant mortality rates are lower among women who are foreign-born compared to those who are native-born. Hummer et al. (2007) showed that infants born to Mexican immigrants in the US have a 10% lower mortality rate than infants born to US-born White women and about a 20% lower mortality rate than infants born to US-born Mexican-origin women. Similarly, infants of Black, White, and several major Asian immigrant groups also have lower mortality rates than their US-born counterparts (Hummer et al., 1999). The largest infant mortality advantage is held by Black and Filipino immigrant women, whose infant mortality rates are 44 and 40% lower, respectively, than their native-born coethnics.

However, as foreign-born women reside longer in the United States, their risk of infant mortality tends to increase. In an innovative analysis of infants born to Puerto Rican women both on the island of Puerto Rico and on the US mainland, Landale et al. (2000) demonstrated that infants born to island-born Puerto Rican women have a very low risk of infant mortality when the island-born women first migrate to the mainland US – in fact, much lower than infants born to mainland-born Puerto Rican women. However, after island-born women reside on the mainland for 15–20 years, their infants are characterized by a mortality rate just as high as that of the mainland-born Puerto Rican women. Clearly in this case, the infant mortality advantage found among new Puerto Rican immigrants wears away after women live on the mainland for a couple of decades and becomes a disadvantage among the second and higher generation Puerto Rican women living on the mainland.

A general immigrant advantage in all-cause US adult mortality is also well documented. Singh and Hiatt (2006) found that immigrants have a 19% lower all-cause, age-adjusted adult mortality rate than their native-born counterparts. As such, immigrants are expected to live about 3 years longer, on average, than US-born individuals. The nativity differential is largest for Blacks and Hispanics: compared to their native-born counterparts, foreign-born Blacks and Hispanics display 52 and 26% lower all-cause adult mortality rates, respectively (Singh and Siahpush, 2002). For Black and Hispanic immigrant men, these lower mortality rates translate into life expectancy levels that are 8.1 and 3.8 years longer, respectively, than their US-born counterparts (Singh and Hiatt, 2006). The immigrant advantage in mortality differs to some degree by gender and cause of death. Singh and Siahpush (2006) demonstrated that foreign-born men exhibit a 23% lower all-cause mortality rate and live 4.0 years longer, on average, than US-born men. The immigrant advantage for women is somewhat narrower: foreign-born women exhibited a 16% lower mortality rate than US-born women, translating

into a 2.6-year life expectancy advantage. Finally, immigrant mortality advantages tend to be widest for causes of death associated with cigarette smoking, such as lung cancer, respiratory diseases, and cardiovascular diseases. Immigrant mortality rates also tend to be significantly lower than native-born persons for deaths due to homicide and suicide (Singh and Siahpush, 2002).

Evidence of an immigrant advantage in mortality and life expectancy is not unique to the US. For example, ethnic Germans who migrated from the former Soviet Union to Germany display significantly lower mortality rates compared to that of the general German population (Ronellenfitsch et al., 2006). In Canada, immigrants experience significantly lower mortality rates than those of native-born Canadians. Canadian male immigrants have an all-cause mortality rate about 23% lower than native-born men, while foreign-born Canadian women exhibit a 17% lower mortality rate than their native-born counterparts (Ng, 2011). However, age-adjusted mortality rates among Canadian immigrants are higher among persons with longer duration in the new country. For instance, immigrant men living in Canada for less than 10 years have an all-cause, age-adjusted mortality rate about 20% lower than those who have been in the country for 10–20 years and about 30% lower than immigrants who have been in Canada more than 20 years. Similar patterns were found for Canadian immigrant women. Although the mortality rates of the most established Canadian immigrants were still lower than those of similarly aged native-born Canadians, the mortality rates for immigrants can be described as converging to those of the native-born with increased duration in the country.

Immigrant mortality patterns in the UK are not as consistently favorable as found in the US and Canada. An influential paper by Marmot et al. (1984) found that foreign-born men from Poland, Italy, India, and the Caribbean have lower mortality rates than men born in the UK. In contrast, Irish immigrants and all immigrant women except those born in Italy were found to have higher mortality rates than native-born UK women. More recent research from the UK found that foreign-born Indian men, Eastern European women, and both female and male immigrants from China and Hong Kong exhibited significantly lower mortality rates than the national averages. However, immigrants from Ireland and Africa, men from Bangladesh, and women from India and Pakistan had mortality rates that were significantly higher than the national average (Wild et al., 2007). Thus, while immigrant mortality patterns are generally favorable to native-born populations in most contexts that have been examined to date, researchers must also consider the diversity of patterns exhibited among national origin groups and by gender.

Physical Health and Health Behaviors

A large body of literature documents nativity differences across a variety of physical health and health behavior measures for individuals of different ages who immigrated from different countries. We touch only briefly on key findings here; indeed, it is a very large body of literature. In general, immigrants have quite favorable physical health and health behavior patterns upon entry into receiving countries, but those favorable patterns tend to dissipate with increased number of years spent in the country. Moreover, the health status of second and third generation persons tends to be less favorable than that of the immigrant generation.

Hamilton et al. (2011) comprehensively investigated whether the prevalence of various childhood health conditions differed for immigrant children in comparison to children of native-born parents within each major US racial/ethnic group. Their results point to a clear pattern of increasing rates of asthma, allergies, learning disabilities, and developmental problems from first (i.e., children born outside of the US) to second (i.e., US-born children of an immigrant parent(s)) to third generation children (US-born children of US-born parents). For instance, only 7.7% of first generation Hispanic children reported asthma compared to 11.9% of second generation and 18.8% of third generation children. Antecol and Bedard (2006) demonstrated similar results for US adults: immigrants were characterized by more favorable patterns of several general measures of health compared to the native-born population.

Within the immigrant generation itself, results from many US studies show that the generally favorable health of new immigrant arrivals tends to become less favorable with increased time spent in the country. Indeed, several groups of researchers have shown that the relatively healthy body mass index (BMI) of immigrant men and women in comparison to the native-born population becomes less favorable with increased duration spent in the United States (Antecol and Bedard, 2006; Singh et al., 2011). The favorable health patterns for newly arrived Asian immigrants (Frisbie et al., 2001) and Hispanic immigrants (Cho et al., 2004) also converge to those of the general population with increased duration in the US.

Singh and Miller (2004) examined several measures of health – including activity limitations, days spent sick in bed, and number of chronic diseases – among a number of racial/ethnic groups by nativity and immigrant duration of residence in the US. Although they found immigrants to have drastically lower odds of poor health compared to their US-born counterparts, immigrants who had been in the United States for fewer than 15 years had even lower odds relative to more established immigrants. For instance, recent Mexican immigrants had 70% lower disability risk than US-born Whites, whereas longer-term Mexican immigrants had 50% lower disability risk. More recently, several studies have examined rates of disability among older immigrants both in the US and in a number of European countries. Interestingly, health and disability patterns for older immigrants both in the US (Markides et al., 2007) and in 10 European Union countries (Sole-Auro and Crimmins, 2008) appear to be less favorable than native-born individuals in those countries, even though the immigrant mortality rates in those countries are lower than those of the native-born population. Such a pattern could occur if older immigrants living in these countries have spent many years working difficult jobs that result in disabilities in old age, while continuing to exhibit low rates of mortality from chronic conditions such as heart disease and cancers.

New immigrants also tend to have more favorable health behaviors than the native-born population, which may be a key contributor to their general health and mortality advantages. One of the most well-documented health behaviors is cigarette smoking. In the US, cigarette smoking prevalence is much

lower among immigrants than the US-born population. Lariscy et al. (2013) recently found that foreign-born Hispanic men and women are 65 and 85% less likely, respectively, than their US-born White counterparts to have ever used cigarettes. The same pattern holds true for other immigrant groups as well. According to research by Singh and Siahpush (2002), Hispanic, Black, and Asian immigrants are 63, 75, and 46%, respectively, less likely to smoke cigarette than US-born Whites. Foreign-born Blacks and Hispanics are also significantly less likely than their US-born counterparts to use cigarettes. Blue and Fenelon (2011) recently estimated that low levels of smoking among US immigrants accounted for 50–75% of the immigrant mortality advantage relative to US-born adults. Moreover, US immigrants are also less likely to drink alcohol at all levels of consumption than their native-born counterparts (Lopez-Gonzalez et al., 2005). Lower levels of alcohol use among immigrants, particularly at high levels of consumption, could help explain lower levels of mortality due to certain cancers, homicide, and suicide in comparison to the native-born US population.

While extensive literature points to a general immigrant health advantage in comparison to the native-born population, the health of immigrants – particularly in the United States – tends to deteriorate as they become more incorporated into the host country. In the population-based literature, length of stay in the new country is often used as a proxy for immigrant incorporation, although some studies also use indicators such as whether or not immigrants have naturalized (i.e., obtained US citizenship). But no matter what indicator of incorporation is used, most studies find that recent immigrants have the most favorable health and mortality patterns while those who are more incorporated into the host society have health and mortality profiles that are much closer to or even worse than those of the native-born population. The health of immigrants has also been shown to deteriorate as they adopt detrimental health behaviors over time. Immigrants who have been in the United States longer are more likely to be cigarette smokers and heavy consumers of alcohol. Lopez-Gonzalez et al. (2005) also show that the association between immigrant incorporation and health behavior is stronger for immigrant women than men; that is, the very healthy behavior of new immigrant women compared to US-born women becomes much more like US-born women for immigrants who have resided in the US for more than 15 years, are naturalized citizens, or both.

Other research has focused on patterns of BMI as immigrants become incorporated into the United States. While immigrants in general arrive in the United States with BMIs significantly lower than that of the native population, their BMI increases with longer duration in the new country. Antecol and Bedard (2006) found that the BMI of US immigrants who have been in the country for 15 years is similar to that of the native-born population. This pattern is more apparent for immigrant women, especially Hispanics, than men, who never fully reach the BMI of US-born men. Moreover, increases in BMI among US immigrants appear to be especially concentrated among those who are poor and arrive from low-income countries (Frank and Akresh, 2013). Thus, as the evidence suggests, immigrant advantages in health diminish with incorporation, but perhaps especially so for those who live in poverty in the United States.

Theoretical Explanations

Health Selection

Positive health selection offers one explanation for the health and mortality advantage of immigrants. According to this theory, individuals who choose to migrate are healthier than those who remain in sending countries. Health selection is commonly assessed by comparing the health status of immigrants in the receiving country and nonmigrants in the country of origin. In the case of positive health selection, we would expect immigrants to have much better health status than those who did not migrate. Several studies have found support for immigrant selectivity, particularly among men, in Europe and the United States. In the most comprehensive study of immigrant health selectivity in the US to date, Akresh and Frank (2008) found that immigrants from all countries and regions studied reported more favorable health in comparison to persons in the sending countries or regions from which they originated. The early classic study by Marmot et al. (1984) echoed this finding in the UK by documenting lower mortality rates among almost all groups of male immigrants in comparison to same aged men in their countries of origin.

Health selection, as with other forms of selection among immigrants, is influenced by a variety of micro- and macro-factors ranging from individual resources to economic climate and national policy. Thus, the degree of health selection varies quite extensively across immigrant groups and a couple of key patterns have emerged in this literature. First, immigrant human capital is important to the health selection process. In a theoretical model of health selection, Jasso et al. (2004) stipulate that potential migrants evaluate the costs and benefits of migration such that the gains – primarily, economic – need to outweigh the costs associated with migration. As such, those with greater human capital resources – such as high levels of educational attainment – are often able to reap more benefits from migrating. Superior health status may also impact this decision-making process by increasing the real or perceived earning potential of immigrants, thus acting as an impetus for healthier individuals to capitalize on greater monetary gains through migration. Indeed, Akresh and Frank (2008) found that immigrants who came to the US with employment visas were more likely to have been selected on good health than those who came to the US under the family reunification and refugee visa programs. More directly, Feliciano (2005) examined education selectivity among immigrants from the top 32 sending countries to the United States. Her results indicate that there is positive educational selectivity among immigrants from 31 out of the 32 migrant-sending countries (the one exception is Puerto Rico, a US territory). Given the well-documented association between education and health, the relationship between educational selectivity and health selectivity may also be closely intertwined.

Second, health selectivity among immigrants is especially salient among those who come from geographically distant sending countries. Akresh and Frank (2008) found that Mexican immigrants have the lowest level of positive health selection compared to immigrants from more distant countries, while immigrants from several African and Asian countries were found to report much better health than the average level in their sending countries. However, that is not to say Mexican

immigrants do not exhibit positive health selectivity. Crimmins et al. (2005) used height, an anthropometric measure that acts as a proxy for childhood health, and found that Mexican immigrants in the US are significantly taller than nonmigrants in Mexico. Nevertheless, geographic distance between the sending and receiving country may very well be related to individual migration costs that have been theorized to influence the positive health selectivity of immigrants.

The positive health selectivity of immigrants is further reinforced through host country policies that screen and potentially reject the entry of immigrants who have health problems. Perhaps even more important, the immigration process can be physically taxing for those migrating on an undocumented basis. Simply put, unhealthy people are much less likely to undertake an arduous and dangerous journey.

Immigrant Incorporation

Immigrant incorporation has been used in this article to refer to the difficult process of social, legal, cultural, and economic adjustment to the host society. Most often in the immigrant health literature, the process of incorporation involves the downward health trajectory of immigrants after their selectively healthy arrival. As documented above, immigrants have been found to adopt the poor health behaviors common in the host country (Lopez-Gonzalez et al., 2005). After a decade or so in the United States, for example, the initially lower levels of cigarette smoking, alcohol consumption, and obesity that characterize immigrants converge to those of the native-born population. These health behaviors are certainly important contributors to the worsening mortality and morbidity outcomes among immigrants with longer duration of time spent in the host country.

However, the process of incorporation is much more complicated than health behavior changes. Increasingly, researchers in this area are considering the often poor working conditions of immigrants: the legal, social, and neighborhood contexts in which immigrants live; the social and neighborhood contexts in which immigrants are embedded; and the racial identity and socioeconomic status that characterize different groups of immigrants (Abraido-Lanza et al., 2006). Some of these dimensions of incorporation, such as a hostile legal context, can have negative health effects, while others, such as the social support found in neighborhoods with high concentrations of immigrants, can have positive health effects. The idea that the immigrant incorporation process is not linear or straightforward and involves a complex set of dimensions is increasingly gaining a foothold in this area of study.

Jasso et al. (2004) further posed a serious challenge to the theoretical model of negative immigrant incorporation. They suggest that the health of immigrants may actually improve over time through the economic gains achieved through the labor market. Using longitudinal data, the authors examined the relationship between income and short-term health changes for new legal immigrants in the United States. They found that legal immigrants experienced a substantial income gain after moving to the US, and that this income change has a significant positive impact on their self-reported health status. While these short-term results do not capture longer-term incorporation effects that may be both positive and negative, the authors argue that health benefits resulting from positive income change may not only persist but actually increase for new immigrants over time. Clearly, then, the incorporation process that immigrants experience must be considered as multidimensional and complex; it cannot simply be measured through simple indicators like duration of residence in the host society.

Future Research Directions

As discussed above, the immigrant population is not only growing in the US and many countries, but it is also very diverse and dynamic. Future research must increasingly consider this diversity because patterns of health, health behavior, and mortality for the entire immigrant population may vastly differ for various subpopulations. Consider, for example, highly educated and skilled immigrants coming to work in professional occupations in comparison to poor refugee immigrants from a war-torn country who are simply moving to survive. The heterogeneity in immigrant backgrounds must be considered in order to develop an accurate picture of immigrant health and mortality.

Second, very little is known about the health, health behavior, and longevity of the estimated 11 million undocumented immigrant population in the US (Pew Hispanic Center, 2013). Research on the undocumented immigrant subpopulation is understandably thin because they may be less likely to trust and participate in private-, government-, or university-based survey initiatives than the general public. That said, undocumented persons do participate in population-based surveys, usually without being asked to divulge their legal status, and the population research community has gotten quite skilled in using such surveys to estimate the size and basic characteristics of the undocumented population. However, because of the relatively young age structure of the undocumented immigrant population, their health and mortality patterns have not been a topic of much research. Clearly, that should change. Indeed, many or most undocumented immigrants are thought to work in very difficult jobs with low pay, live very stressful lives with high levels of fear, have limited access to social and/or health services, and may face serious discrimination in social institutions and at the individual level (Massey, 2013). Those are not favorable characteristics for long-term good health or for the health of native-born children of undocumented immigrants. Many social scientists worry that the difficult and fearful social and legal context in which undocumented immigrants and their children live will one day have dire health consequences for the lives of these individuals, almost all of whom are living and working in the host country in the attempt to better their lives while serving as a source of needed labor. A better understanding of the health, health behavior, and mortality patterns of the undocumented immigrant population should be an important topic of research in this area in the near future, even though such research will undoubtedly be very challenging to conduct.

Finally, as discussed over 30 years ago in a seminal article (Kasl and Berkman, 1983), the ideal research design for the study of immigrant health and mortality patterns involves a three-way comparison of immigrants in the destination

country to both persons in the origin country(ies) who did not migrate as well as to native-born persons in the host country. Most often, research designs do not make such three-way comparisons and instead only focus on a comparison between immigrants and the native-born population in the host society. This unfortunately ignores nonmigrants in the origin country(ies), from where the pool of immigrants came. However, a few innovative studies in this area have made such three-way comparisons, allowing researchers to understand both the ways that immigrants differ from demographically similar people in origin countries who do not migrate as well as to how they differ from demographically similar native-born people in the host country (e.g., Landale et al., 2000; Riosmena et al., 2013). But we also urge researchers to think about research designs in a longitudinal framework that follow individuals over time to best determine how the health behaviors, health trajectories, and mortality risk of individuals in these three subgroups unfold. It is only in the context of such a longitudinal, three-way design that the research community will truly be able to understand how the health and mortality patterns of immigrants differ from those of their nonmigrant and native-born counterparts, both at the time of immigration and for many years thereafter.

See also: Adult Mortality in Industrialized Societies; Assimilation of Immigrants; Health Behaviors, Assessment of; Health Behaviors; Immigration Policy; Immigration, Applied Research On; Migration and Social Capital; Migration, Theory of; Migration: Cultural Aspects; Mortality Differentials: Selection and Causation; Population Composition by Race and Ethnicity: North America.

Bibliography

Abraido-Lanza, Ana, Armbrister, Adria N., Florez, Karen R., Aguirre, Alejandra N., 2006. Toward a theory-driven model of acculturation in public health research. American Journal of Public Health 96, 1342–1346.

Akresh, Ilana Redstone, Frank, Reanne, 2008. Health selection among new immigrants. American Journal of Public Health 98 (11), 2058–2064.

Angel, Ronald J., Guarnaccia, Peter J., 1989. Mind, body and culture: somatization among Hispanics. Social Science and Medicine 29, 85–94.

Antecol, Heather, Bedard, Kelly, 2006. Unhealthy assimilation: why do immigrants converge to American health status levels? Demography 43 (2), 337–360.

Blue, Laura, Fenelon, Andrew, 2011. Explaining low mortality among US immigrants relative to native-born Americans: the role of smoking. International Journal of Epidemiology 40, 786–793.

Brown, Susan K., Bean, Frank D., 2005. International migration. In: Poston, Dudley L., Micklin, Michael (Eds.), Handbook of Population. Springer Publishers, New York, pp. 347–382.

Cho, Youngtae, Frisbie, William P., Hummer, Robert A., Rogers, Richard G., 2004. Nativity, duration of residence, and the health of Hispanic adults in the United States. International Migration Review 38 (1), 184–211.

Crimmins, Eileen M., Soldo, Beth J., Kim, Jung Ki, Alley, Dawn E., 2005. Using anthropometric indicators for Mexicans in the United States and Mexico to understand the selection of migrants and the 'Hispanic paradox'. Social Biology 52, 164–177.

Crimmins, Eileen M., Kim, Jung Ki, Alley, Dawn E., Karlamangla, Arun, Seeman, Teresa, 2007. Hispanic paradox in biological risk profiles. American Journal of Public Health 97 (7), 1305–1310.

Feliciano, Cynthia, 2005. Educational selectivity in U.S. immigration: how do immigrants compare to those left behind? Demography 42 (1), 131–152.

Frank, Reanne, Akresh, Ilana Redstone, 2013. Social patterning in body mass index (BMI) among contemporary immigrant groups: the emergence of a gradient. Demography 50 (3), 993–1012.

Frisbie, William P., Cho, Youngtae, Hummer, Robert A., 2001. Immigration and the health of Asian and Pacific Islander adults in the US. American Journal of Epidemiology 153 (4), 372–380.

Hamilton, Erin R., Cardoso, Jodi Berger, Hummer, Robert A., Padilla, Yolanda C., 2011. Assimilation and emerging health disparities among new generations of U.S. children. Demographic Research 25 (25), 783–818.

Hummer, Robert A., Biegler, Monique, De Turk, Peter B., Forbes, Douglas, Frisbie, W. Parker, Hong, Ying, Pullum, Starling G., 1999. Race/ethnicity, nativity, and infant mortality in the United States. Social Forces 77 (3), 1083–1117.

Hummer, Robert A., Powers, Daniel A., Pullum, Starling G., Gossman, Ginger L., Frisbie, W. Parker, 2007. Paradox found (again): infant mortality among the Mexican-origin population in the United States. Demography 44 (3), 441–457.

Jasso, Guillermina, Massey, Douglas S., Rosenzweig, Mark R., Smith, James P., 2004. Immigrant health: selectivity and acculturation. In: Anderson, Norman B., Bulatao, Rodolfo A., Cohen, Barney (Eds.), Critical Perspectives on Racial and Ethnic Differences in Health in Late Life. National Academies Press, Washington DC, pp. 227–266.

Kandula, Namratha R., Kersey, Margaret, Lurie, Nicole, 2004. Assuring the health of immigrants: what the leading health indicators tell us. Annual Review of Public Health 25, 357–376.

Kasl, Stanislav V., Berkman, Lisa, 1983. Health consequences of the experience of migration. Annual Review of Public Health 4, 69–90.

Landale, Nancy S., Oropesa, R.S., Gorman, Bridget K., 2000. Migration and infant death: assimilation or selective migration among Puerto Ricans? American Sociological Review 65, 888–909.

Lariscy, Joseph T., Hummer, Robert A., Rath, Jessica M., Villanti, Andrea C., Hayward, Mark D., Vallone, Donna M., 2013. Race/ethnicity, nativity, and tobacco use among U.S. young adults: results from a nationally-representative survey. Nicotine and Tobacco Research 15 (8), 417–426.

Lopez-Gonzalez, Lorena, Aravena, Veronica C., Hummer, Robert A., 2005. Immigrant acculturation, gender and health behavior: a research note. Social Forces 84 (1), 581–593.

Loue, Sana, Sajatovic, Martha (Eds.), 2012. Encyclopedia of Immigrant Health. Springer Publishers, Dordrecht, Netherlands.

Markides, Kyriakos S., Coreil, Jeanne, 1986. The health of Hispanics in the Southwestern United States: an epidemiologic paradox. Public Health Reports 101, 253–265.

Markides, Kyriakos S., Eschbach, Karl, 2011. Hispanic paradox in adult mortality in the United States. In: Rogers, Richard G., Crimmins, Eileen M. (Eds.), International Handbook of Adult Mortality. Springer Publishers, Dordrecht, Netherlands, pp. 227–240.

Markides, Kyriakos S., Eschbach, Karl, Ray, Laura A., Peek, M. Kristen, 2007. Census disability rates among older people by race/ethnicity and type of Hispanic origin. In: Angel, Jacqueline L., Whitfield, Keith E. (Eds.), The Health of Aging Hispanics: The Mexican-Origin Population. Springer Publishers, New York, pp. 26–39.

Marmot, M.G., Adelstein, A.M., Bulusu, L., 1984. Lesson from the study of immigrant mortality. Lancet, 1455–1457.

Massey, Douglas S., 2013. Comment: building a better underclass. Demography 50 (3), 1093–1095.

Migration and Remittances Factbook, 2011. Available at: www.worldbank.org/prospects/migrationandremittances (accessed on 20.06.13.).

Ng, Edward, 2011. The healthy immigrant effect and mortality rates. Health Reports 22 (4), 25–29.

Palloni, Alberto, Arias, Elizabeth, 2004. Paradox lost: explaining the Hispanic adult mortality advantage. Demography 41 (3), 385–415.

Pew Hispanic Center, 2013. A Nation of Immigrants: A Portrait of the 40 Million, Including 11 Million Unauthorized. Pew Research Center, Washington, DC. Available at the Pew Hispanic Center website: www.pewhispanic.org (accessed 19.06.13.).

Riosmena, Fernando, Wong, Rebeca, Palloni, Alberto, 2013. Migration selection, protection, and acculturation in health: a binational perspective. Demography 50 (3), 1039–1064.

Ronellenfitsch, Ulrich, Kyobutungi, Catherine, Becher, Heiko, Razum, Oliver, 2006. All-cause and cardiovascular mortality among ethnic German immigrants from the former Soviet Union: a cohort study. BMC Public Health 6, 16.

Singh, Gopal K., Siahpush, Mohammad, 2002. Ethnic-immigrant differentials in health behaviors, morbidity, and cause-specific mortality in the United States: an analysis of two national data bases. Human Biology 74 (1), 83–109.

Singh, Gopal K., Miller, Barry A., 2004. Health, life expectancy, and mortality patterns among immigrant populations in the United States. Canadian Journal of Public Health 95 (3), I14–I21.

Singh, Gopal K., Hiatt, Robert A., 2006. Trends and disparities in socioeconomic and behavioural characteristics, life expectancy, and cause-specific mortality of native-born and foreign-born populations in the United States, 1979–2003. International Journal of Epidemiology 35, 903–919.

Singh, Gopal K., Siahpush, Mohammed, 2006. Widening socioeconomic inequalities in US life expectancy, 1980–2000. International Journal of Epidemiology 35 (4), 969–979.

Singh, Gopal K., Siahpush, Mohammad, Hiatt, Robert A., Timsina, Lava R., 2011. Dramatic increases in obesity and overweight prevalence and body mass index among ethnic-immigrant and social class groups in the United States, 1976–2008. Journal of Community Health 36, 94–110.

Sole-Auro, Aida, Crimmins, Eileen M., 2008. Health of immigrants in European countries. International Migration Review 42 (4), 861–876.

Swanson, David A., Stephan, G. Edward, 2004. Glossary. In: Siegel, Jacob S., Swanson, David A. (Eds.), The Methods and Materials of Demography, second ed. Elsevier Academic Press, Amsterdam, Netherlands, pp. 751–778.

Treas, Judith, Batalova, Jeanne, 2009. Immigrants and aging. In: Uhlenberg, Peter (Ed.), International Handbook of Population Aging. Springer Publishers, Dordrecht, Netherlands, pp. 365–394.

Turra, Cassio, Elo, Irma T., 2008. The impact of salmon bias on the Hispanic mortality advantage: new evidence from social security data. Population Research and Policy Review 27 (5), 515–530.

Wild, S.H., Fischbacher, C., Brock, A., Griffiths, C., Bhopal, R., 2007. Mortality from all causes and circulatory disease by country of birth in England and Wales 2001–2003. Journal of Public Health 29 (2), 191–198.

Immigration Policy

Sofia A Perez, Boston University, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Immigration policy includes all of those policies whereby states control the influx of persons who want to establish residence within their borders: rules regarding rights of access to the territory (entry and residence), permission to participate in the labor market (work permits), the rights of asylum seekers and refugees, the rights of immigrants to bring family members (family reunification), and rules for the acquisition of citizenship by immigrants and their family members (naturalization). Such rules may be linked to the notion of humanitarian 'rights' of individuals (as in the case of asylum and family reunification law) or respond to selective state objectives (such as demographic and labor market needs). Immigration policies can be traced back to rules whereby modern states defined who held citizenship rights in the aftermath of the French Revolution. The most important distinction here is that between states followed the rule of *jus solis* (birthright citizenship) versus those following *jus sanguinis* (acquisition of citizenship by descent). Labor shortages that emerged during the industrial revolution led to the introduction of labor recruitment schemes by the United States and several European countries during the late nineteenth century and again after World War II. Together with decolonization, they provided a strong impetus for migration from poorer to richer countries starting in the mid-twentieth century. Following the economic crisis of the 1970s, many countries ended these labor recruitment schemes. The crisis in the former Yugoslavia also led to a tightening of rules regulating applications for asylum. Family reunification hereby became the dominant avenue for immigration into advanced industrialized countries. The 1990s and 2000s saw the introduction of new selective immigration policies aimed at recruiting specialized labor and high-skill professionals in a number of industrialized countries. The overall tenor of immigration policy is difficult to measure because similar rules will cause higher rates of immigration into some countries and lower rates into others depending on other factors that push people to migrate out or their countries of origin or attract them to migrate given conditions in destination countries (referred to as 'push' and 'pull' factors in the literature on migration).

Introduction

The term *immigration policy* encompasses all those measures by which a state regulates the influx of persons, in particular those who want to establish residence in their territory either temporarily or permanently. Rules that governed the *right of abode* of free tradesmen in medieval times and those that served to define *citizenship* (also referred to as *nationality*) in Europe following the French Revolution represent early points of reference for such policy (Brubaker, 1992). Today, immigration policy is commonly divided into four categories: rules that allow work-based immigration, those related to humanitarian concerns such as political asylum and family reunification, and those that govern the acquisition of citizenship through naturalization. Policies aimed at integrating immigrants into host societies (immigrant integration policy) may be distinguished analytically from strict immigration policy (those that govern entry and abode). But such policies, along with many other aspects of social, educational, and labor market policies, can also affect migration flows.

Types of Immigration Policy

The legal and administrative framework of different countries affects the ways in which immigration policies are categorized. Different historical trajectories including the timing of state formation, its relationship to the extension of political and social rights (citizenship) and to socioeconomic and geopolitical developments behind major migratory flows all influence how modern immigration policy is framed and regulated across countries. Nonetheless, post-World War II developments such as international humanitarian conventions and European integration mean that immigration policies are typically categorized in terms of whether they relate humanitarian rights (to political asylum or family life), labor recruitment (work-based immigration), or the acquisition of full citizenship rights by immigrants (naturalization). These categories are also commonly used to quantify immigration statistics (see **Figure 1**). Whereas much early migration was driven by labor recruitment schemes, historical developments (see below) have meant that family reunification has become the dominant basis for immigration into most advanced industrialized economies with work-based migration playing an important role in only a few OECD countries and asylum-based migration subject to increasing restrictions. However, the last set of enlargements of the EU have also produced large-scale migration based on free movement within the area (as illustrated in **Figure 1**). Acquisition of citizenship rules varies substantially with some countries facilitating a path to citizenship for most permanent immigrants over a relatively short number of years and others imposing far lengthier residency requirements that are harder to meet. Naturalization rates thus also vary significantly among the top destination countries for migrants (see **Figure 2**).

Historical Background

Starting in the nineteenth century, many states began to institute laws to regulate the status of foreigners within their territory, attract immigrant labor, or restrict migratory flows across their borders. Such laws were driven both by political and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.75022-7

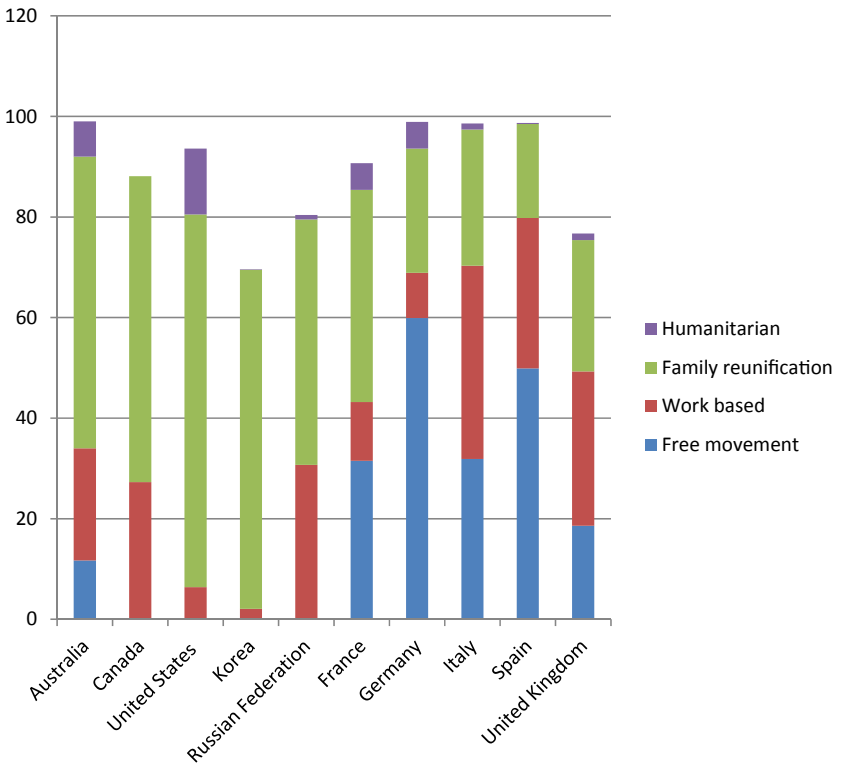

Figure 1 Structure of permanent migration inflows by immigration policy type. (Top 10 immigrant receiving countries, 2010.) Source: Own elaboration from country tables, International Migration Outlook 2013 (OECD).

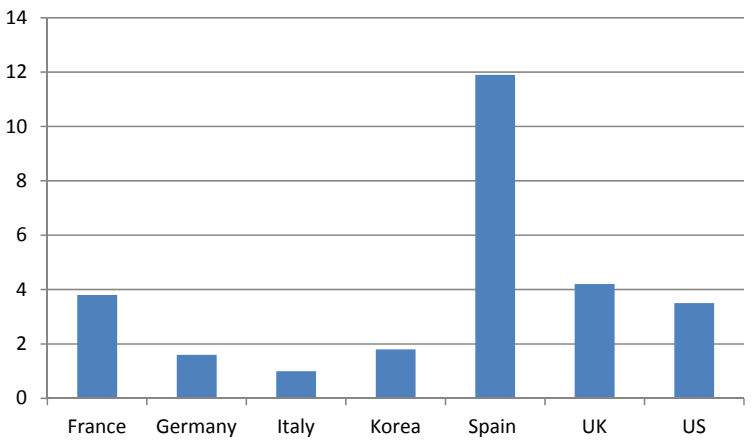

Figure 2 Average naturalization rates for select destination countries. (Percentage of Foreign Population for years 2006–10.) Source: Authors elaboration based on International Migration Outlook 2013 (OECD).

geopolitical reasons and by new demographic needs linked to socioeconomic developments (Brubaker, 1992; Bade, 2003: pp. 129–164). Early postcolonial states that put claim to large expanses of land with relatively limited populations – Argentina, the United States, Australia, Canada, or New Zealand – sought to attract European immigrants who brought capital or specific skills to establish their control over territory and promote economic development (Waisman, 1987; Freeman, 1995). Yet the influx of new ethnic groups often produced political backlash, giving rise to *nativist* political movements that led to the adoption of restrictions to labor immigration along ethnic and racial lines. The Chinese Exclusion Act of 1882 and the 1924 National Origins Formula, which defined immigration policy in the United States for many decades, are prominent examples (Zolberg, 2008). In Europe, the seasonalization of agriculture and the industrial revolution produced large cross-border migratory flows that reached a new peak in the late nineteenth century and early

twentieth century. Rapidly industrializing Prussian Germany soon became the main destination country for migrants in Europe, second only to the United Sates, while Italy and the British Isles long held the status as the main country of emigration (Bade, 2003: Chapter 2). During this time, governments introduced early labor recruitment schemes, such as the Prussian state's regulation of labor migrants from Poland starting in the 1880s and France's interwar regime of bilateral treaties to import workers for key economic activities from Southern and Eastern European states (Bade, 2003: pp. 157–164; Noiriel, 1988; Lewis, 2007). In most Western countries, policies turned more restrictive during the course of the interwar period as national labor unions gained power and economic conditions worsened.

Reconstruction and growth during the post-World War II decades in Europe also hinged on migratory flows from South and South Eastern Europe to the main Continental economies (in particular France, Germany, and the Netherlands). This period marked the heyday of foreign labor recruitment schemes – commonly known as *guest worker* programs – which allowed domestic employers to recruit foreign workers without envisioning a path to permanent settlement or the acquisition of citizenship. Germany and Switzerland, in particular, enforced strict rotation of workers as a means to limit *de facto* settlement. Decolonization was the second main driver of immigration into Europe during this period, as former colonial powers (in particular Britain and France) extended citizenship to former colonial subjects. In the case of France, whose citizenship laws were based on the principle of *jus solis*, the children of immigrants typically acquired nationality at birth or once they had reached legal age. (Nonetheless, French nationality law has been changed repeatedly since the 1980s to raise the residence requirements for children born in France who do not also have one parent who holds French citizenship.) In other countries whose citizenship traditions were based on the tradition of *jus sanguinis*, notably Germany, Switzerland, and Austria, courts subsequently extended rights of residence (although not of citizenship) based on humanitarian considerations. Such rights include the right of political asylum, enshrined in the 1951 Geneva convention on the Rights of Refugees, and the principle of family reunification, which was included in Article 16 of the United Nations' 1948 Universal Declaration of Human Rights and in many national constitutions.

In the United States, where birthright citizenship was enshrined in the Constitution, the domestic struggle for civil rights and equality led to a new immigration law in 1965 that abolished the country of origins quota system established in the 1920s (Safran, 1997). The quota system had heavily favored immigration from Northern Europe (as opposed to Southern and Eastern Europe, Asia, and Africa). The 1965 act allowed for permanent immigration based on work, but it also imposed an overall annual quota system requiring that 74% of visas be reserved for family reunification cases. In the case of Japan, citizenship by descent was the norm, and immigrants who were not born to Japanese parents (including conscripted laborers from China and Korea brought over during World War II) were left in a legal limbo by the 1952 Japanese Immigration law. Nonetheless, Japan did allow for a significant number of refugees from Indochina in the late 1970s and ratified the Geneva Convention on the status of Refugees in 1981. Canada and Australia both maintained liberal immigration laws to encourage the settlement of foreign citizens during the postwar period.

Recent Developments

If the overall trend in the 1950s and 1960s was toward relatively liberal immigration policies, the period following the first OPEC oil shock was marked by abrupt moves to restricted immigration into many advanced industrialized countries. Germany suspended its labor recruitment schemes in 1973, followed by France, Sweden, and the Netherlands, and thereafter immigration into most of Europe was largely restricted to family reunification and asylum cases (Perez, 2011). In the United States, the system adopted in 1965 – allowing for work-based immigration but prioritizing family reunification – was maintained. Political backlash to large-scale immigration from Mexico and Central America led individual states to pass restrictions on the rights of immigrants starting in the 1990s. The hallmark of this trend was California's proposition 187, a ballot initiative passed by public referendum in 1994 that denied health care and public education to undocumented immigrants in the state. The measure was subsequently repealed by a federal court. Nevertheless, Proposition 187 marked the beginning of a wave of efforts by several other states in the United States to encourage so called self-deportation of immigrants through laws that made life inhospitable for them.

In the case of Europe, the break up of Yugoslavia and the massive exodus of refugees from the Balkans at the end of the 1980s led to a second wave of restrictions during the 1990s. The German parliament reformed the country's Basic Law in 1993 restricting the right of asylum so that individuals that had passed through a third country that was deemed 'safe' by the German legislature could no longer apply for asylum. The 'safe third country' rule was in turn adopted by other EU member states. Following the 1997 Treaty of Amsterdam – which opened the way for EU wide immigration policy – it was also adopted as EU law. It has since that time also become part of asylum law in many other countries beyond the EU.

One consequence of the restriction of legal immigration since the 1990s has been the rise in the number of undocumented – also referred to as irregular or illegal – immigrants in many advanced industrialized countries. This includes not only persons who have entered clandestinely but also individuals whose visas have lapsed and who continue to remain in a country. Undocumented immigrants live precarious lives, facilitating the growth of the informal economy and presenting a threat to wages and labor standards. Some states have opted to address the situation of undocumented immigrants through regularization (also known as 'legalization,' 'normalization,' or 'amnesty') allowing those who are deemed to be deserving for humanitarian reasons or those who have long stays to legalize their status. Regularization can occur either through ongoing mechanisms that allow persons with long stays to apply for permanent residence or as one-off programs that allow persons who entered prior to a particular date to apply for legal status (Baldwin-Edwards and Kraler, 2009; OECD, 2011; Brick, 2011). The latter approach, however, is often criticized by those who believe regularization campaigns reward illegal behavior and serve as an implicit inducement to illegal

immigration (Levinson, 2005). Many scholars have also noted the growing tendency by governments and the media to frame immigration policy in terms of its relationship to national security and crime since the 1990s, a phenomenon referred to as the *securitization* of immigration policy (Huysman, 2000; Kostakopoulou, 2000; Muller, 2004; Tsoukala, 2005). This tendency was accentuated by the 11 September 2001 attacks in the United States (Faist, 2001; Levy, 2005; Rudolph, 2006; Messina, 2007).

Another development in the past two decades has been the gradual move away from 'rights'-based immigration policies based on humanitarian and entitlement principles to more 'selective,' 'points-based,' or 'demand-based' immigration policies, which seek to change the profile of immigration in favor of migrants that have specific, and in particular higher, skills (Balch, 2010; Cerna, 2009; Perez, 2011). Canada, a country that pursued a very open immigration policy in the postwar period, launched a *points*-based system to give priority to skilled workers as early as 1967. Japan adopted a skills-based immigration law in 1989, allowing for entry of high-skilled workers and professionals while officially not allowing the entry of low-skilled workers and, in 1993, also introduced a trainee system under which unskilled workers are able to migrate to Japan to work in agriculture, fishery, construction, and other sectors facing labor shortages. In Europe, the Labor government of Tony Blair introduced a 'managed migration' during the late 1990s to promote the immigration of persons that met specific labor market needs, and in 2007, Gordon Brown's Labor government passed a new 'points-based' immigration law that was largely modeled on the Canadian example. These policy changes allowed for a significant rise in work-based immigration into the United Kingdom during the 2000s following two decades in which net migration growth had been held close to zero by Conservative governments.

The two principal destination countries into the EU in the late 1990s and 2000s by far, however, were Italy and, in particular, Spain (Perez, 2011), two countries that had previously served as major suppliers of immigrant labor to the rest of Europe. Clandestine immigration from the African continent and the Balkan region into these two states received a great deal of media attention both domestically and internationally. But most immigrants into Spain and Italy arrived by normal border crossings, either by land or air. By contrast to the rest of Western Europe where immigration was virtually limited to family reunification since the early 1990s, Spain and Italy, along with the United Kingdom, also had a much larger component of work-based immigration and pursued specific labor market objectives with their policies (Perez, 2011). However, following the 2004 and 2207 EU enlargements, immigration from outside the EU was displaced by intra-EU migration from new to old member states, and such intra-EU migration does not lend itself to selectivity as member states are committed by law to allow free movement of workers within EU borders. The EU Commission has consistently advanced the view that EU member states need to adopt more open, if targeted access to third country nationals to address skill shortages and the problem of aging populations and high dependency ratios. With this in mind, it submitted a proposal for the issuance of new EU-wide permanent work permits – called Blue Cards – to compete with the U.S. Green Card program and skills-based immigration schemes in Canada and Australia. The proposal was adopted as an EU directive in 2009, but it leaves room for member state governments to suspend the issuance of such work permits on the basis of national needs.

In the United States, where immigration continues to be heavily dominated by family reunification, there have also been proposals to strengthen the component of skills-based immigration. The more important debate in the 2010s, however, has centered on the status of a very large number of undocumented workers. Immigrants who apply for work-based permits and seek to obtain permanent residence in the United States often end up in undocumented status because they face an interruption in their employment status during the course of completing the work-based residency applications. Together with clandestine entries and people who overstay tourist visas, undocumented workers featured prominently in the 2008 presidential election and in the subsequent congressional elections, with some politicians offering support for a regularization effort. Proposals for immigration reform that would allow persons with long stays, those who reached the United States as minors, or those who belong to 'mixed-status' families to regularize their status have been the subject of long political debate and became a major issue of contention in the 2012 presidential and congressional elections.

Key Debates: Immigrant Integration and Implications for Welfare States and Labor Markets

Immigration policy across the world has also been strongly influenced by the perceived failures to integrate the children of poorer immigrants in advanced industrialized countries, particularly children of Turkish immigrants in Germany, those of North African descent in France, and hispanic immigrants to the United States (Schain, 2008). 'Downward assimilation' (Portes, 2001), or the tendency for the children of poor immigrants with lower educational backgrounds to get trapped in poor American cities where public education resources are stretched, has been a particular concern in the United Sates. Policies to address integration have ranged along many dimensions (Bauböck, 1995; Favell, 1998; Guiraudon, 2000; Koopmans et al., 2005; Jopke, 2007). One distinction that has received particular attention, however, is that between a multiculturalist approach to integration, consisting of cultural accommodation of new immigrants through such measures as bilingual education or legal accommodation of religious practices, and a more traditional approach geared toward cultural assimilation. Proponents of multiculturalism argue that it serves to combat social exclusion and thereby furthers social integration (Kymlicka, 1995; Kymlicka and Banting, 2006) while opponents believe that it can obstruct integration by leading to self-segregation and the fragmentation of society into ethnic subgroups (Barry, 2001; Goodhart, 2004) or even dependence on welfare support (Koopmans, 2010). One particular argument holds that multiculturalist integration policies tend to become a barrier to socioeconomic integration in countries with more generous welfare spending while they may be of less consequence in countries with less generous systems of social support (Koopmans, 2010). The Netherlands once seen as the paragon of the multiculturalist

approach to integration, but also Sweden and parts of Belgium are often cited as examples of such a 'progressive dilemma' (Goodhart, 2004) because they show less successful employment outcomes among immigrants than do other European countries. However, other authors have observed that countries that pursue multiculturalist accommodation may allow immigrants to develop social capital by way of ethnic networks (ethnic capital) that supports economic integration and that rates of civic and political participation by immigrants in countries pursuing multiculturalist policies are also higher (Fennema and Tillie, 1999; Jacobs and Tillie, 2004). The desirability of such mode of integration, however, is questioned by authors who regard it as a challenge to universalist, liberal values (for instance Jopke, 2008). Yet other authors who place immigration policy in a broader structural context distinguish various *modes of incorporation* which are the result not only of policy but also the particular structure of immigrant communities and their historical relationship to the host country (Portes and Zhou, 1993; Portes, 1995; Bodovski and Benavot, 2006).

The perceived failures to integrate immigrants have led many European countries to impose language requirements as a condition for immigration or the renewal of residence permits. The United States, which does not have an official national language, has not imposed such restrictions, although individual states have passed so-called English only laws that either establish English as the state's official language or seek to prohibit state and local governments from using another language. The policy debate on immigrant integration in the United States has centered heavily on the best way to promote the acquisition of English as a second language among the children of immigrants who have limited English language proficiency, with approaches ranging from education of immigrant children in their native language (commonly referred to as bilingual education) to sheltered language immersion programs (Rossell and Kuder, 2005).

Among economists, there has also been an ongoing debate over the benefits of immigration for overall economic growth and its impact on the wages of different categories of domestic workers, with some arguing that the existing pattern of low-wage immigration, in particular into the United States, hurts poorer domestic workers disproportionally (Borjas, 2003) while others argue that low-wage immigration tends to benefit overall economic growth and allows lower paid native workers to move into better paid work (Card, 2004; Mayda and Patel, 2004; Peri and Sparber, 2008; D'Amuri and Peri, 2010). Another concern raised by both economists and political scientists is the potential impact that immigration may have on support for national welfare states. While some authors have argued that the increased cultural heterogeneity that immigration brings about may undermine support economic redistribution (Freeman, 1995; Alesina and Glaeser, 2004), others have argued that such an outcome is unlikely (Banting and Kymlicka, 2006; van Oorschot and Uunk, 2007).

Problems of Measurement

Efforts to measure the overall tenor of immigration policy – how liberal or restrictive it is across countries – have tended to be elusive because of the multidimensionality character of immigration. Indices covering several policy dimensions, such as work visa and asylum policy requirements, as well as measures of policy change in a more or less restrictive direction have been constructed (Mayda and Patel, 2004; Mayda, 2009; Ortega and Peri, 2009), but they tend to raise questions of validity as they require the coding of cumulative laws across countries and the reference points for immigration laws are often different from country to country. Yet another index developed by the British Council and the Migration Policy group – the Migration and Integration Policy Index (MIPEX) – measures the degree of legal equality that immigrants attain once they have arrived. However, immigration can often be encouraged or discouraged through measures that are not captured in formal laws, such as the time it takes to obtain visas or obstacles to the accreditation of educational degrees. Because immigration policy involves so many different elements, aggregation of policy measures almost invariably involves subjective decisions. More importantly, as in the case of trade policy, a given immigration rule may represent a more or less restrictive depending on factors such as the demand for labor in a particular destination country, transportation costs, or differences in wages between two countries. This complexity has led some authors (see Boussichas and Goujon, 2010) to argue for the use of 'revealed' immigration policy measures that compare actual immigration flows to those that we would expect from nonpolicy, structural determinants (growth, employment, trade openness, and other identified drivers of migration) to tease out the effect of immigration policy.

Conclusion

Immigration policy has gained growing importance and electoral salience over the past two decades in the advanced industrialized world (Givens and Luedke, 2005; Schain, 2008), and political and academic debates regarding the consequences of immigration have become increasingly charged. At the same time, countries with aging populations and high dependency ratios (in particular in Europe but also Japan) face demographic trends that may only be resolved through further immigration. The trend toward more open but selective, work-based migration policies, as typified by the EU Blue Cards program and skills-based immigration schemes in Canada and Australia, reflects such underlying demographic imperatives. However, rising migratory flows from poorer to richer countries also produce strong political reactions, as exemplified by the rise of anti-immigrant parties in much of Europe and the hard-line stance of the conservative Tea Party wing of the Republican party in the 2012 elections in the United States. The great recession has further fed such reaction. In the EU, the impact of EU enlargement allowing for large-scale free movement from new to old member states in the period since 2007 has been to encourage renewed popular resistance to immigration from third countries and to place long-term concerns regarding demographic growth on a back burner. At the same time, the crisis in the Eurozone has reversed long-time trends, leading to renewed emigration (principally of skilled labor) from the European countries most affected by the crisis – some of which paradoxically had become major destination countries for

immigrants in the prior decade – to the rest of the EU as well as to emerging market economies. These most recent developments suggest that the long-term outlook for immigration patterns and immigration policies is likely to depend heavily on broader economic developments in the EU and elsewhere.

See also: Adoption, Demography of; Ethnicity and Migration in Europe; Migration, Economics of; Public Policy; Trade Policy; Welfare Retrenchment.

Bibliography

Alesina, Alberto, Glaeser, Edward, 2004. Fighting Poverty in the US and Europe: A World of Difference. Oxford University Press, New York.

Bade, Klaus J., 2003. Migration in European History. Blackwell Publishing, Oxford.

Baldwin-Edwards, Martin, Kraler, Albert, 2009. REGINE Regularisations in Europe: Study on Practices in the Area of Regularisation of Illegally Staying Third-Country Nationals in the Member States of the EU. ICMPD (International Centre for Migration Policy Development), Vienna.

Balch, Alex, 2010. Managing labour migration in Europe: ideas knowledge and policy change. Center for Comparative Immigration Studies (CCIS). Working Paper 184.

Banting, Keith, Kymlicka, Will, (Eds.), 2006. Multiculturalism and the Welfare State: Recognition and Redistribution in Contemporary Democracies. Oxford University Press, Oxford.

Barry, Brian, 2001. Culture and Equality. Polity Press, Cambridge.

Bauböck, Rainer, 1995. Transnational Citizenship. Membership and Rights in International Migration. Edward Elgar, Aldershot.

Bodovski, Katerina, Benavot, Aaron, 2006. Unequal educational outcomes among first-generation immigrants: the case of youth from the former Soviet Union in Israel. Research in Comparative and International Education 1 (3).

Borjas, George J., 2003. The labor demand curve is downward sloping: reexamining the impact of immigration on the labor market. The Quarterly Journal of Economics 118 (4), 1335–1374.

Boussichas, Matthieu, Goujon, Michael, 2010. A quantitative indicator of the immigration policy's restrictiveness. Economics Bulletin, AccessEcon 30 (3), 1727–1736.

Brick, Kate, December 2011. Regularization in the European Union. Migration Policy Institute Insight. Migration Policy Institute, Washington, DC.

Brubaker, Rogers, 1992. Citizenship and Nationhood in France and Germany. Harvard University Press, Cambridge, MA.

Card, David, 2004. Is the New Immigration Really so Bad? CREAM Discussion Paper Series CDP No. 02/04. Centre for Research and Analysis of Migration, London.

Cerna, Lucie, 2009. The varieties of high-skilled immigration policies: coalitions and policy outputs in advanced industrial countries. Journal of European Public Policy 16 (1), 144–166.

D'Amuri, Francesco, Peri, Giovanni, 2010. Immigration and Occupations in Europe. CREAM Discussion Paper Series 26/10. Centre for Research and Analysis of Migration, London.

D'Amuri, Francesco, Peri, Giovanni, 2011. Immigration, Jobs and Employment Protection: Evidence from Europe. NBER Working Papers 17139. National Bureau of Economic Research.

Faist, Thomas, 2002. Extension du domaine de la lutte: international migration and security before and after 11 September 2001. International Migration Review 36 (1), 7–14.

Favell, Adrian, 1998. Philosophies of Integration. Immigration and the Idea of Citizenship in France and Britain. Macmillan, Basingstoke.

Fennema, Meindert, Tillie, Jean, 1999. Political participation and political trust in Amsterdam: civic communities and ethnic networks. Journal of Ethnic and Migration Studies 25 (4), 703–726.

Freeman, Gary P., Winter, 1995. Modes of immigration Politics in liberal democratic states. International Migration Review 29 (4), 881–902.

Givens, Terri, Luedtke, Adam, 2004. The Politics of European Union Immigration Policy: institutions, salience, and harmonization. Policy Studies Journal 32 (1), 145–165.

Givens, Terri, Luedtke, Adam, 2005. European immigration policies in comparative perspective: issue salience, partisanship and immigrant rights. Comparative European Politics 3 (1), 1–22.

Goodhart, David, February 2004. Too Diverse? Prospect (United Kingdom), 30–37.

Guiraudon, Virginie, 2000. Les Politiques d'Immigration en Europe. L'Harmattan, Paris.

Huysmans, Jef, 2000. The European Union and the securitization of migration. Journal of Common Market Studies 38 (5), 751–777.

Jacobs, Dirk, Tillie, Jean, 2004. Introduction: social capital and political integration of migrants. Journal of Ethnic and Migration Studies 30 (3), 419–427.

Jopke, Christian, 2007. Transformation of immigrant integration: civic integration and antidiscrimination in the Netherlands, France, and Germany. World Politics 59, 243–273.

Jopke, Christian, 2008. Immigration and the identity of citizenship: the paradox of universalism. Citizenship Studies 12 (3), 533–546.

Koopmans, Ruud, 2010. Trade-offs between equality and difference: immigrant integration, multiculturalism and the welfare state in cross-national perspective. Journal of Ethnic and Migration 36 (1), 1–26.

Koopmans, Ruud, Statham, Paul, Giugni, Marco, Passy, Florence, 2005. Contested citizenship. Immigration and cultural diversity in Europe. Minneapolis and London: University of Minnesota Press Studies 36 (1), 1–26.

Kostakopoulou, Theodora, 2000. The 'protective union'; change and continuity in migration law and policy in post-Amsterdam Europe. JCMS: Journal of Common Market Studies 38, 497–518.

Kymlicka, Will, Banting, Keith, 2006. Multiculturalism and the Welfare State Recognition and Redistribution in Contemporary Democracies. Oxford University Press, London.

Kymlicka, Will, 1995. Multicultural Citizenship: A Liberal Theory of Minority Rights. Clarendon Press, Oxford.

Lahav, Gallya, 2004. Immigration and Politics in the New Europe: Reinventing Borders. Cambridge University Press, New York.

Levinson, Amanda, 2005. Why Countries Continue to Consider Regularization. Migration Policy Institute, Washington, DC.

Levy, Carl, 2005. The European Union after 9/11: the Demise of a liberal democratic asylum regime? Government and Opposition 40 (1), 26–59.

Lewis, Mary D., 2007. Boundaries of the Republic Migrant Rights and the Limits of Universalism in France, 1918–1940. Stanford University Press.

Mayda, Anna M., 2009. International migration: a panel data analysis of the determinants of bilateral flows. Journal of Population Economics 23 (4), 1249–1274.

Mayda, Anna M., Patel, Krishna, 2004. OECD Countries Migration Policy Changes. Georgetown University. http://www.georgetown.edu/faculty/amm223/.

Messina, Anthony M., 2007. The Logics and Politics of Post-WWII Migration to Western Europe. Cambridge University Press, New York.

Muller, Benjamin J., 2004. (Dis)qualified bodies: securitization, citizenship, and 'identity management'. Citizenship Studies 8 (3), 279–294.

Noiriel, Gerard, 1988. Le Creuset français. Histoire de l'immigration (XIXe – XXe siècle). Seuil, Paris.

OECD, 2011. International Migration Outlook 2010. SOPEMI, Paris.

Ortega, Francesc, Peri, Giovanni, 2009. The Cause and Effects of International Migrations: Evidence from OECD Countries 1980–2005. NBER Working Paper 14833.

van Oorschot, Wim J.H., Uunk, Wilfred J.G., 2007. Welfare spending and the public's concern for immigrants: multilevel evidence for eighteen European countries. Comparative Politics 40 (1), 63–82.

Perez, Sofia A., 2011. Immigration and the European left. In: Cronin, James E., Ross, George, Schoch, James, (Eds.), What's Left of the Left: Democrats and Social Democrats in Challenging Times. Duke University Press, pp. 265–289.

Peri, Giovanni, Sparber, Chad, 2008. Task Specialization, Immigration, and Wages. CREAM Discussion Paper 02/08. Centre for Research and Analysis of Migration, London.

Portes, Alejandro, 1995. Children of immigrants: segmented assimilation and its determinants. In: Portes, Alejandro (Ed.), The Economic Sociology of Immigration. Russell Sage Foundation, New York, pp. 248–280.

Portes, Alejandro, 2001. Legacies: The Story of the Immigrant Second Generation. University of California Press.

Portes, Alejandro, Zhou, Min, 1993. The new second generation: segmented assimilation and its variants. Annals of the American Academy of Political and Social Science 530, 74–96.

Rossell, Christine, Kuder, Julia, 2005. Meta-murky: a rebuttal to recent meta-analyses of bilingual education. In: Söhn, Janina (Ed.), The Effectiveness of Bilingual School Programs. AKI Research Review 2 (Wissenschaftszentrum Berlin für Sozialforschung). Germany, Berlin, pp. 43–76.

Rudolph, Christopher, 2006. National Security and Immigration: Policy Development in the United States and Western Europe since 1945. Stanford University Press, Stanford, CA.

Safran, William, 1997. Citizenship and nationality in democratic systems: approaches to defining and acquiring membership in the political community. International Political Science Review 18 (3), 313–335.

Schain, Martin, 2008. The Politics of Immigration in France, Britain, and the United States: A Comparative Study. Macmillan, London.

Tsoukala, Anastassia, 2005. Looking at migrants as enemies. In: Bigo, Didier, Guild, Elspeth, (Eds.), Controlling Borders: Free Movement into and within Europe. Ashgate, Burlington, VT, pp. 161–192.

Waisman, Carlos H., 1987. Reversal of Development in Argentina: Postwar Counter-revolutionary Policies and Their Structural Consequences. Princeton University Press, New Haven, CT.

Zolberg, Aristide R., 2008. A Nation by Design: Immigration Policy in the Fashioning of America. Harvard University Press, Cambridge, MA.

Immigration: Political Aspects

Christian Joppke, University of Bern, Bern, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews recent cross-disciplinary academic work on the political aspects of immigration, a previously understudied aspect of international migration. A distinction is made between immigration policy (or immigrant selection) and integration policy, reviewing some of the most important theory-guided, explanatory, and comparative works on these topics.

Immigration is the permanent movement of individuals across states, seen from the perspective of the receiving (rather than sending) state. While a 'transfer of jurisdictions' (Zolberg, 1981: p. 5) is thus constitutive of immigration, traditional theories of immigration have surprisingly ignored the state, focusing instead on the economic motivations and social ties that initiate and perpetuate migration flows. The traditional view can be bound, for instance, in this encyclopedia's 1968 predecessor, which discussed 'migration' only under its 'social' and 'economic,' but not political aspects (Petersen, 1968; Thomas, 1968). What since its publication in the late 1990s has remained the state-of-the-art review of migration theories and of the world's major 'migration systems' (Massey et al., 1998) indirectly defends the traditionally nonpolitical migration theories, in arguing that states are in principle incapable of controlling social network-based migration flows (see also Massey, 2001).

While, especially in the United States, the study of immigration continues to be a domain of sociologists and demographers, a major development in the past few decades has been a growing interest by political scientists and political sociologists in the central role of the state, and increasingly of nonstate actors also, in the soliciting, control, and integration of immigrants and immigration. This reflects not only a disciplinary evolution and renewed interest in the state and 'governance' across the social sciences, but also some changing feature of contemporary migration for which the state has become increasingly central. The classic notion of immigration was coined in the 'stateless' period before World War I, when almost 50 million Europeans moved largely unhindered by national entry controls to the Americas and the Antipodes (but for the importance of subfederal controls in nineteenth-century America, see Neuman (1993)). After a brief restrictionist interlude in the inter-War period, the image of 'stateless' immigration was revived after World War II, when the classic settler nations abolished racially motivated restrictions on immigration, and when some northwestern European countries underwent large-scale labor migrations from the South. States already figured prominently for these flows because a conscious change of policy (as in North America or Australia, which abolished its 'White Australia' policy in 1971), bilateral agreements with the governments of sending countries (as in Germany or Switzerland's guestworker migration), or a geopolitically motivated abstention from controlling postcolonial flows (as in Britain, France, or the Netherlands) were necessary to initiate or allow them. However, the involvement of states was invisible because this immigration was wanted or at least passively accepted and considered unproblematic. A typical example of this period is France, where until the late 1960s, immigration from the South proceeded spontaneously and outside the official channels of the *Office nationale d'immigration* but was legitimized after the fact by the routine distribution of residence and work permits (see Viet, 1998: pp. 233–254).

The picture of 'stateless' immigration – though always a fiction – changed radically in the mid-1970s, when under the impact of the first oil crisis and the most serious economic recession after World War II, European states shifted to zero-immigration policies, and when seemingly 'uncontrolled' illegal immigration from Mexico galvanized a restrictionist turn in the United States. Now, 'states' began to matter, either as bulwarks against unwanted immigration (as in the rhetoric of political elites) or as sieves that chronically failed to live up to their restrictionist intentions (as in the account of most academics; see, for instance, the influential 'gap hypothesis' by Cornelius et al. (2004)). In this historical moment, when the central role of the state in international migration became problematic and therefore visible, political sociologists and political scientists discovered 'immigration,' suggesting models to classify and explain states' immigration policies and how the latter derived from – or violated – their sovereignty (Section Explaining Immigration Policies) and investigating the impact of immigration on the membership component of nation-states, which is the topic of a new literature on integration and citizenship (Section Integration and Citizenship). When the previous edition of this encyclopedia went into print, leading demographers and sociologists, the traditional 'owners' of the immigration domain, still thought that the relationship between state policies and international migration, while perhaps considered increasingly important, remained 'under-theorized and little studied' (Massey et al., 1998: p. 286). While this view was already mistaken a decade ago, it appears even less plausible today, as I shall demonstrate in the remainder of this article.

Explaining Immigration Policies

A state's decision to admit immigrants, without which there is either no immigration or at best (or worst) 'illegal' immigration, is formally a function of its sovereignty, as codified in international law, according to which states are free to reject or to expel noncitizens (within due process constraints). This

linkage between immigration policy and sovereignty is expressed in the standard notion, propagated by governments around the world that immigration policy is in the 'national interest.' The problem is that, particularly in liberal states, the 'national interest' is no fixed thing, standing above group interests and notoriously short of organizational manifestation. In a pathbreaking essay, which continues to be the single most influential political science treatise on immigration policy, Gary Freeman (1995) explained in rational-choice terms why 'special' instead of 'national' interests shape the immigration policies of liberal states. According to Freeman, the costs of immigration are diffusely spread over the population, while its benefits are concentrated (for employers and co-ethnics). It follows from the logic of collective action that the beneficiaries of concentrated benefits win over the bearers of diffuse costs, in a constellation that Freeman calls (following J.Q. Wilson, 1980) 'client politics.' This simple model explains the paradox that liberal states, despite their restriction-minded elite rhetoric and public opinion, have remained largely open to (certain kinds) of immigrants.

A number of authors have quickly picked up and revised or amended Freeman's model. In a cogent political science analysis, Jeannette Money (1999) has argued that because of the spatial concentration of immigrant settlement also the costs of immigration are concentrated, thus remedying the free-rider problem in these geographic areas and giving rise to a constellation that Wilson (1980) had called 'interest-group politics.' The restrictionist impulse, first picked up by local politicians, can go nationwide if the respective district is crucial for maintaining or gaining a national electoral majority. Demonstrating this swing-vote dynamics in the cases of France, Great Britain, and Australia, Money's revision of Freeman's model helps explain why some states remain open to new immigration, while in others immigration control is put on the national political agenda. Joppke (1998b, 1999) has argued that the logic of client politics is more firmly established in classic immigrant nations than in European states, where pro-immigrant lobbies are much weaker, and where nominal 'zero-immigration policies' fit more easily with 'national interest' rhetoric. In addition, he stressed (much like Hollifield, 1992) that liberal norms, which invest immigrants with rights, are a second constraint on state policy, next to interest-group pressure. Reference to liberal norms, backed by independent courts, helps explain why European states, despite the absence of an ethnic lobby there, have remained open to certain kinds of immigrants, most notably the family members of labor migrants. Once the role of judicial constraints and 'self-limited sovereignty' of the constitutional state (Joppke, 1998b) was on the map, the next round was to qualify that. Discussing the case of Japan, Amy Gurowitz (1999) argued that in a non-Western state, the liberal-constitutional infrastructure is much weaker, so immigrant activists resort to global human rights norms and regimes (that Joppke (1998b) had found less important than domestic-level legal norms and actors). Comparing guest-worker policies in Switzerland and Germany, Antje Ellermann (2013) even argued that political factors are more important than legal factors to explain the success or failure of these policies. Switzerland, where there is no constitutional court watching over the executive, managed to carry through the 'rotation' of guest workers because of past (negative) policy-learning and the lacking insulation of an (liberal-leaning) executive from populist pressures (due to its direct democracy); Switzerland is the case of a liberal state *not* having to accept 'unwanted' immigration. By contrast, West Germany saw guestworker migration turn into permanent settlement because of the nonpossibility of policy-learning (due to the state collapse in 1945) and 'rotation' kicking in as policy objective too late in time, in a context of greater executive state insulation from popular and business pressure, and after settlement had already occurred. In sum, not legally based 'self-limited sovereignty' but the institutional setup of the recruitment systems and more narrowly political factors matter for understanding the fate of guestworker policies. Also note that, in the 1990s, Germany devised a second-generation guest-worker policy, this time with Poland, that strictly enforced the temporariness of this migration from the start so that these 'guests' would not stay (see Castles, 2006).

When linking the analysis of immigration policy to the larger question of the resilience or transformation of sovereignty in an era of globalization (competing views on this are laid out in Joppke (1998a); also Guiraudon and Lahav, 2000), one must strictly distinguish formal state authority from empirical state capacity. As conceptualized in international relations theory and implemented under international law, 'sovereignty' refers to the formal last-instance authority of the state within 'its' territory. In this formal sense, even ineffective immigration controls say nothing about the presence or absence of state sovereignty over immigration. Instead, the crucial test is if actors other than the state have acquired formal last-instance authority over immigration matters. With some important exceptions, such as the free movement regime within the European Union (which is mostly not emulated by other regional regimes), this test will almost always be negative. One must even register that the constitutive quality of the state to be a container of people has greatly increased in the current era of globalization, where the state has been in retreat on so many other fronts. An altogether different question is whether the presence of unwanted migrants within their territories (such as illegal migrants or asylum seekers) signals a declining empirical capacity (as against formal authority) of states to control immigration. The effectiveness of Western states to bring down, first, the number of unwanted asylum seekers in the 1990s through a combination of national-level and supranational (EU-level) restrictions, and, second, to impose new restrictions on family migration once the 'action' had shifted to that front a decade later, but, above all, thirdly, to digitally control borders and territories with new technologies suggests that there is little evidence for a decline of empirical state capacity (see Guiraudon and Joppke, 2001; Mau et al., 2012).

But 'state capacity' is no static thing. As Antje Ellermann (2009) showed in an important study of deportation policies in Germany and the United States, the state capacity to deport immigrants varies not only across countries but, above all, according to the phase of the policy cycle. Accordingly, law-making on deportation is facilitated by institutional linkages with the (generally restriction-minded) public, while implementation is more effective when bureaucrats are insulated from the public gaze, which now works in favor of

the attention- and sympathy-catching victim of this most brutal form of 'social regulation.'

Two new developments in immigration policy in the past decade have been, firstly, the selective opening of developed European and East Asian states to high-skilled immigration, which went along with a narrowing of previous differences between classic immigration countries and the various newcomers (or rather latecomers) in the developed world; and, secondly, the growing pressure for building an international (or 'global') regime for managing migration, which, however, tends to be permanently frustrated. With respect to the first, after the liberal delegitimation of selecting immigrants by their ethnic or racial origins (see Joppke, 2005), the only selectivity left for states is on the grounds of migrants' skill profile. While there is dramatic demographic (and sometimes economic) need for large numbers of low-skilled immigrants, particularly in Europe and East Asia, no democratically accountable government dares acknowledging this need, for populist reasons. In almost all states of the world, low-skilled immigration, if admitted at all, is strictly limited to seasonal, nonpermanent migration (the most recent craze being about policies that may allow 'circular' migration). With respect to high-skilled immigration, the story is dramatically different. No developed state in the world, not even the notorious 'no-immigration countries' Germany and Japan, dares abstaining from the 'race for talent' (Shachar, 2006) that marks the contemporary global economy (for Japan, see Skrentny et al., 2012). Germany, since its new immigration law of 2004, grants instant permanent legal residence permits for high-skilled immigrants, thus radically departing from the old European logic of only gradually increasing residence rights (*Aufenthaltsverfestigung*) (but it keeps this inflow small through asking for a job offer at hand and high-income requirements).

The new turn to high-skilled immigration directs attention to the different possibilities that states have to steer this migration (see Sachverständigenrat, 2013: pp. 16–24). First, there is the possibility of employer-centered selection, whereby a prior work contract is required for admission. The model for this is the American H1B visa, which has channeled America's need for high-skilled immigrants within an immigration system that heavily prioritizes family migration. Interestingly, Germany has followed the American H1B model in its recent opening to high-skilled immigration, especially in the IT sector. A second, though less utilized, possibility is sector-specific programs, where labor-scarce sectors or professions are provided with quota that may be filled by immigrants even without a prior work contract. Thirdly, the so-called points systems of Canada or Australia prize human capital irrespective of concrete labor market needs, only requiring immigrants to pass a numerical threshold of a scale rewarding youth, education, and language abilities. In reality, however, one can observe the emergence of 'hybrid' systems of recruiting high-skilled immigrants, which greatly reduces the previous variety of national approaches (see Papademetriou and Sumption, 2011). Canada, for instance, increasingly issues only temporary visa, even for high-skilled immigrants, who only gradually may 'earn' the right to permanence, European-style, and it layered onto its points system a prior 'scarce profession' proviso or the requirement of a job offer at hand, which integrates elements of employer- and sector-driven immigrant selection. Conversely, Germany, in terms of a recently reformed Paragraph 18 of its Residence Law, now allows non-EU academics to enter Germany in search of a job, thus obliterating the prior requirement of a formal job offer (for a similar comparison of Spain and Canada, see Finotelli (2012)). The reality of immigrant selection is highly complex and in flux, defying stereotyped national models or old types of immigration or nonimmigration country, and academic accounts are often outdated before they go into print.

A second recent trend has been pressure for building an international regime for managing migration such as the UN-sponsored Global Forum on Migration and Development. Parallel to this, academics have mapped the few elements pointing toward an international regime, normatively stipulating the need for one that is far more elaborate than at present (see Betts, 2011). Simple interest arithmetic shows why no international regime exists or is likely to come into existence anytime soon. The main problem is fundamental interest 'asymmetry' with respect to low- and high-skilled immigrants in sending and receiving countries (Bhagwati, 2004). With respect to low-skilled migration, supply greatly exceeds demand, and there are few incentives on the part of underdeveloped sending states to restrict these moves (the only exception being bilateral agreements that may exchange restricted emigration against development aids). With respect to high-skilled migration, the opposite constellation prevails, the demand for this migration greatly exceeding supply, with sending countries showing little motivation to condone massive brain drain. No international regime is ever to emerge in this totally divergent interest constellation.

Integration and Citizenship

A second line of new and innovative scholarship on the politics of immigration has opened up over the impact of immigration on states as 'membership associations' (Brubaker, 1992). A variety of authors have concluded from the upgrading of immigrant rights across liberal states that 'citizenship' as unitary membership form and focus of collective identity in modern states has been devalued and set on a path of decline, with a dissociation between 'postnational membership,' protected by a global human rights regime, and parochial or larger-than-national identities (Soysal, 1994). At the same time, migrants' 'transnational' practices and 'diasporic' forms of association are said to increasingly bypass the nation-state, thus rendering irrelevant old models of one-directional integration and assimilation (e.g., Glick Schiller et al., 1992).

Two decades later, it is obvious that 'postnational' and 'transnational' approaches have underestimated the resilience and adaptability of the national citizenship construct (for an overview of recent trends, see Joppke, 2010a). The vulnerabilities of noncitizens became clear, first, in the US mid-1990s' welfare reforms that restricted federal social benefits for legal immigrants, and, post-2001, in antiterror laws that primarily targeted noncitizens. In Europe, immigration has triggered not a decline but a transformation of citizenship. The

continent that two centuries ago had been a pacesetter of the postfeudal, 'modern' principle of family- and descent-based citizenship (jus sanguinis) has undergone a reverse trend toward territorial citizenship (jus soli), originally a feudal principle that had been found wanting in the dawning era of nation-states. By the 1990s, most member states of the European Union granted second- and third-generation immigrants a right to citizenship either at birth or at majority age, old assimilation requirements for naturalization were waived, and there was increasing pragmatic acceptance of dual citizenship. Interestingly, new restrictions after the terrorist attacks of 2001 only entailed a tightening of naturalization rules, by means of citizenship tests and new civic integration requirements, while the trends toward jus soli citizenship and the toleration of dual citizenship were not reversed.

As citizenship obviously continued to be the lodestar of immigrant integration, a new literature inquired into the determinants of liberalizing or restricting the access to it. In a 25-country comparison, Patrick Weil (2001) argued that three factors had to be in place for a liberal opening of citizenship: a liberal-democratic legacy; noncataclysmic, past immigration that is accepted as irreversible; and a congruence of national and state boundaries. While a parsimonious model that hardly meets a case that does not fit, it is also functionalist and ex-post, saying little about the political dynamics underlying the evolution of citizenship. Joppke (2003) argued that this evolution was actually in two opposite directions, conditioned by party politics, a 'de-ethnicizing' direction under left-wing governments (facilitating access to citizenship for immigrants), but a 're-ethnicizing' direction under right-wing governments (strengthening ties with emigrants and expatriates abroad) (see also Janoski (2010), confirming the de-ethnicizing trend under left-wing governments). In a compelling comparison of post-Enlargement EU states, Marc Morjé Howard (2009) found the left-right hypothesis wanting. What instead explains liberal reform is the absence of mobilizing a latent anti-immigrant disposition among mass publics, while the presence of such mobilization, often instigated by far-right parties, may lead to new restrictions, even under left-wing governments.

If one considers the traditional closure of ethnic nation-state citizenship to immigrants (for which there had been a complement in the Americas in terms of racial exclusion), the de-ethnicization of citizenship, particularly in Europe, is clearly the more astonishing trend. It went along with an intransitive mode of immigrant integration across liberal states. According to this intransitive mode, immigrants are not the object of forced assimilation, as in the 'Americanization' campaign in the early twentieth century, but are the subjects of freely willed integration. In the process, 'integration' tends to be favored over 'assimilation,' if the latter is not shunned altogether (particularly in Europe). Against a backdrop of weakened nationalist semantics post-World War II, it means that immigrants are expected to adhere to the procedural values that constitute liberal-democratic states; they are not expected to become 'assimilated,' that is to take on the substantive life forms of a particular nation-state. Assimilation may well be the popular expectation, but it just cannot be enforced by means of public policy that is constrained by liberal-democratic norms. In this minimal sense, which could not be derailed by the new wave of citizenship tests and civic integration requirements since the early millennium (see Joppke, 2007; Bauböck and Joppke, 2010), Nathan Glazer's (1997) pronouncement that "we are all multiculturalists now" remains valid. The fundamental change of attitude is visible, for instance, in Germany's move away from individually tested cultural assimilation (*Hinwendung zu Deutschland*) in liberalized naturalization rules in the early 1990s, which even introduced a right to citizenship under certain conditions. And the reassertion of 'Americanization' in the mid-1990s United States stressed that "ethnic and religious diversity based on personal freedom is compatible with national unity" (US Commission on Immigration Reform, 1997: p. 25).

As carriers of legitimate cultural difference, immigrants cannot but undermine the principle of modern nationhood that Ernest Gellner famously identified in the congruence of political and cultural boundaries (1983). This inevitable *de facto* multiculturalism, however, is to be distinguished from policies of official multiculturalism, which have come under fire almost everywhere (see Joppke, 2004). Multiculturalism's most concise philosophical formulation (Kymlicka, 1995) wisely concedes that immigrants, by voluntarily leaving their society of origin, have 'waived' the right to have their homeland resurrected in the receiving society. The retreat from multiculturalism and turn to 'civic integration' implicitly follows this reasoning, allowing immigrants as individuals to pursue their religious and cultural ways of life, while denying them state support to prosper as groups and instead defining the task of public policy as binding immigrants into mainstream institutions.

Indeed, the most important policy innovation with respect to immigrant integration in the past 15 years has been the rise of 'civic integration' and new citizenship tests. In the most detailed comparative account available today, Sara Wallace Goodman has argued, "the new requirements do not signal departures from national approaches to citizenship, but rather fortify them" (2014). The metaphor of 'fortification' is brought against that of the 'lightening' of citizenship, suggested by Joppke (2010b), who argued that citizenship in liberal states is increasingly 'lite' – that is, easy to access, increasingly less exclusive with respect to rights, and capped by thin, procedural identities that are exchangeable across states. Against this diagnosis, Goodman's main claim is that national variations or 'models' of citizenship and immigrant integration are not overcome by new 'convergent' civic integration policies (as first argued in Joppke, 2007), but that nationally distinct citizenship traditions continue to prosper in the context of 'resilient nation-states' (Goodman, 2014). She thus refutes an entire prong of recent scholarship that has found the old distinction of 'national models' – such as French Republicanism, British multiculturalism, or German-style nonintegration – wanting to understand changing integration and citizenship policies.

Goodman's unquestionable contribution is to show that civic integration is not all of one cloth but differing according to 'scope,' 'sequencing,' 'density,' and – above all – 'purpose' (see Goodman, 2012, 2014). With respect to 'purpose,' she argues that the new policies may be liberal or restrictive,

depending on the two variables of existing citizenship legacies and political conflict surrounding them. Within this frame, Goodman finds much variation across Europe. However, she underrates the degree of *change*, often in departure from existing citizenship traditions, which was brought about by the new civic integration policies. For instance, 'restrictive continuity' (a pattern identified for Austria and Denmark) may be a good description for Austria, but it is not for Denmark, which was once as 'liberal' as its Nordic neighbors and only recently turned into a paragon of restrictiveness, under the pressure of one of Europe's most extreme right-wing parties (the Danish People's Party). Secondly, contrary to her claim, there has been no unambiguous 'liberal continuity' in the UK. Only consider the 'probationary' and 'earned' citizenship schemes proposed in the early millennium by the late Labor government under Tony Blair, which were meant to make citizenship more exclusive, even 'virtuous,' and which were scrapped only last minute, under the new Tory government of David Cameron. In sum, in Britain there seems to be less straightforward continuity in citizenship policy than oscillation between liberal-inclusive and restrictive approaches. Thirdly, qualifying Germany's stance as 'restrictive moderation' belittles the significant departure from its 'ethnic' citizenship legacy in the great citizenship reform of 1999, where Germany introduced conditional jus soli citizenship for the children of settled immigrants, following the European trend. It is a stretch to call the purpose of German-style civic integration (which mostly consists of requiring immigrants to learn German, in order to increase their labor market chances) as 'leaving a restrictive citizenship relatively insulated from change' (Goodman, 2014). In reality, there has been real liberalizing change in 1999, as there had been in a previous reform, that is, the 1992 liberalization of ultrarestrictive naturalization rules. If one further considers the introduction of a distinctly nonmoral, noncultural federal citizenship test in 2008 with a near 100% pass rate, what is Germany's 'national approach to citizenship' that is supposed to be maintained here? Finally, it is problematic to group France and the Netherlands in one category of 'liberal moderation.' The French *contrats d'accueil et d'intégration* are as soft and service oriented as they could be, almost Canadian style. It is implausible to square them with the nastier 'integration from abroad' schemes concocted in the Netherlands, which require immigrants to be familiar with Dutch language and history even before they enter the Netherlands, and which are really an immigration control policy under a different name, aiming to restrict unwanted family migration from Muslim societies.

It is correct to argue that by means of the new civic integration policies, the modern nation-state 'reasserts' its 'community-defining competence' (Goodman, 2014). And it is indeed novel that integration is now 'assessed' through explicit and objective criteria that allow no second guessing, whereas in the past, integration was simply 'inferred' from residence time and left to the individual to fill with meaning (Goodman, 2014). But this is no replication of classic nation-building, and it is certainly exaggerated to claim that the nation-state is 'stronger than ever,' 'a resilient and flourishing entity of political order' (Goodman, 2014). This is an odd picture for Europe, where European Community law has bitten significant holes into the possibility of exclusive nation-state citizenship (see Joppke, 2010b). If one examines instead the contents of civic integration requirements and their cognitive and normative dimensions (largely ignored by Goodman), one discovers a reality of *changing* state identities today. A comparison of four West European countries and the United States found that the civic knowledge expected of new immigrants and citizenship applicants is mostly procedural and related to the principles of liberal democracy, thus remaining within the confines of 'political liberalism' à la John Rawls (Michalowski, 2011). Civic integration is no return to cultural assimilation. It has more in common with a general turn of social policy from welfare to workfare (see Gilbert, 2002), in seeking to make immigrants (much like the nonimmigrant rest) functioning participants in society's productive core activities, especially the labor market – note that one of the strange peculiarities of Europe, unknown in America or Australia, is that immigrants are at a minimum double as likely to be unemployed as natives (see Hansen, 2012). This is not to deny that there are repressive possibilities under the mantle of liberalism, in terms of scrutinizing people for their 'loyalty' and personal identification with liberal-democratic principles, which amounts to liberalism hyped up as identity. However, 'repressive liberalism' raises eyebrows precisely for the strange 'repressive–liberal' pairing, examples for which remain rare (see Joppke, 2010a: pp. 137–142). If one scrutinizes the 'new parameters of belonging under the banner of liberalism' (Goodman, 2014), one discovers a reality not different from the one arrived at from the vantage point of 'lightened' citizenship, only that the latter has the advantage of factoring in more explicitly the uses of liberalism in reformed citizenship schemes (Joppke, 2010a: Chapter 4).

See also: Citizenship, Anthropology of; Citizenship, Historical Development of; Citizenship, Sociological Aspects of; Ethnicity and Migration in Europe; Immigration Policy; Immigration, Citizenship, and Integration: Social Work Connections; Immigration, Emigration, and Citizenship; Integration and Multiculturalism: Focus on Western Europe; Integration, Social; Migration: Cultural Aspects; Multiculturalism, Anthropology of; Multiculturalism; Transnational Citizenship.

Bibliography

Bauböck, R., Joppke, C. (Eds.), 2010. How Liberal Are Citizenship Tests? Working paper. Robert Schuman Centre for Advanced Studies, European University Institute, Florence.

Betts, Alexander (Ed.), 2011. Global Migration Governance. Oxford University Press, Oxford.

Bhagwati, J., 2004. In Defense of Globalization. Oxford University Press, New York.

Brubaker, Rogers, 1992. Citizenship and Nationhood in France and Germany. Harvard University Press, Cambridge, MA.

Castles, Steven, 2006. Guestworkers in Europe: a resurrection? International Migration Review 40 (4), 741–766.

Cornelius, Wayne, Martin, Philip, Hollifield, Jim (Eds.), 2004. Controlling Immigration, second ed. Stanford University Press, Stanford.

Ellermann, Antje, 2009. States against Migrants. Cambridge University Press, New York.

Ellermann, Antje, 2013. When can liberal states avoid unwanted immigration? World Politics 65 (3), 491–538.

Finotelli, Claudia, 2012. Change of paradigms? A comparison of Canadian and Spanish labour migration models. Journal of Comparative Policy Analysis 15 (4), 1–18.

Freeman, Gary, 1995. Modes of immigration politics in liberal democratic states. International Migration Review 29 (4), 881–913.
Gellner, Ernest, 1983. Nations and Nationalism. Cornell University Press, Ithaca, NY.
Gilbert, Neil, 2002. Transformation of the Welfare State. Oxford University Press, New York.
Glazer, Nathan, 1997. We Are All Multiculturalists Now. Harvard University Press, Cambridge, MA.
Glick Schiller, N., Basch, L., Cristina, B.-S., 1992. Transnationalism. Annals of the New York Academy of Sciences 645, 1–24.
Goodman, Sara W., 2012. Fortifying citizenship. World Politics 64 (4), 659–698.
Goodman, Sara W., 2014. Immigration and Membership Politics in Western Europe. Cambridge University Press, New York.
Guiraudon, V., Joppke, C. (Eds.), 2001. Controlling a New Migration World. Routledge, London.
Guiraudon, V., Lahav, G., 2000. A reappraisal of the state sovereignty debate. Comparative Political Studies 33 (2), 163–195.
Gurowitz, Amy, 1999. Mobilizing international norms. World Politics 51 (3), 413–445.
Hansen, Randall, 2012. The Centrality of Employment in Immigrant Integration in Europe. Working paper. Migration Policy Institute, Washington, D.C.
Hollifield, Jim, 1992. Immigrants, Markets, and States. Harvard University Press, Cambridge, MA.
Howard, Marc Morjé, 2009. The Politics of Citizenship in Europe. Cambridge University Press, New York.
Janoski, Thomas, 2010. The Ironies of Citizenship. Cambridge University Press, New York.
Joppke, Christian (Ed.), 1998a. Challenge to the Nation-State: Immigration in Western Europe and the United States. Oxford University Press, Oxford.
Joppke, Christian, 1998b. Why liberal states accept unwanted immigration. World Politics 50 (2), 266–293.
Joppke, Christian, 1999. Immigration and the Nation-State: The United States, Germany, and Great Britain. Oxford University Press, Oxford.
Joppke, Christian, 2003. Citizenship between de- and re-ethnicization. European Journal of Sociology 44 (3), 429–458.
Joppke, Christian, 2004. The Retreat of Multiculturalism in the Liberal State: Theory and Policy. British Journal of Sociology 55 (2), 237–257.
Joppke, Christian, 2005. Selecting by Origin: Ethnic Migration in the Liberal State. Harvard University Press, Cambridge, MA.
Joppke, Christian, 2007. Beyond national models: civic integration policies for immigrants in Western Europe. West European Politics 30, 1–22.
Joppke, Christian, 2010a. Citizenship and Immigration. Polity, Cambridge, UK.
Joppke, Christian, 2010b. The inevitable lightening of citizenship. European Journal of Sociology 51 (1), 9–32.
Kymlicka, Will, 1995. Multicultural Citizenship. Oxford University Press, Oxford.
Massey, D., 2001. March of folly. American Prospect, 22–33. March–April.
Massey, D., et al., 1998. Worlds in Motion. Oxford University Press, New York.
Mau, Steffen, et al., 2012. Liberal States and the Freedom of Movement. Palgrave Macmillan, Basingstoke.
Michalowski, Ines, 2011. Required to assimilate? The content of citizenship tests in five countries. Citizenship Studies 15 (6/7), 749–768.
Money, J., 1999. Fences and Neighbors: The Political Geography of Immigration Control. Cornell University Press, Ithaca, NY.
Neuman, Gerald, 1993. The lost century of American immigration law. Columbia Law Review 93 (8), 1833–1901.
Papademetriou, D., Sumption, M., 2011. Rethinking Point Systems and Employer-Selected Immigration. Migration Policy Institute, Washington, D.C.
Petersen, W., 1968. Migration, Social Aspects. International Encyclopedia of the Social Sciences. Macmillan, New York.
Sachverständigenrat deutscher Stiftungen für Integration und Migration, 2013. Die pragmatisch-technokratische Wende in der Migrations- und Integrationspolitik als Ausgangspunkt für unabhängige wissenschaftliche Politikberatung. Sachverständigenrat, Berlin.
Shachar, Ayelet, 2006. The race for talent. New York University Law Review 81 (148), 206.
Skrentny, J., Gell-Redman, M., Lee, J.J.G., 2012. Japan, the United States, and the philosophical bases of immigration policy. American Behavioral Scientist 56 (8), 995–1007.
Soysal, Yasemin, 1994. Limits of Citizenship. University of Chicago Press, Chicago.
Thomas, B., 1968. Migration, Economic Aspects. International Encyclopedia of the Social Sciences. Macmillan, New York.
U.S. Commission on Immigration Reform, 1997. Becoming an American: Immigration and Immigrant Policy. Government Printing Office, Washington, D.C.
Viet, V., 1998. La France immigrée. Fayard, Paris.
Weil, Patrick, 2001. Access to citizenship. In: Aleinikoff, A., Klusmeyer, D. (Eds.), Citizenship Today. Carnegie Endowment for International Peace, Washington, D.C.
Wilson, J.Q., 1980. The Politics of Regulation. Basic Books, New York.
Zolberg, A., 1981. International migrations in political perspective. In: Kritz, M., Lin, L.L., Slotnik, H. (Eds.), International Migration Systems. Oxford University Press, Oxford.

Immigration: Social Psychological Aspects

Eva GT Green and Nicole Fasel, University of Lausanne, Lausanne, Switzerland
Oriane Sarrasin, University of Sussex, Brighton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article introduces a social psychological perspective for understanding phenomena related to immigration in culturally diverse societies. The viewpoint of both national majorities and immigrants is considered. In the first section, we examine how immigration affects identification processes of national majorities and immigrant minorities, and the consequences of identification for intergroup attitudes. In the second section, the role of threat perceptions, stereotypes, and worldviews in national majorities' stances toward immigrants is presented. Moreover, the psychological consequences of discrimination for immigrants are overviewed. In the final section, both intergroup contact between the national majority and immigrants and multicultural norms are discussed.

Throughout the history of humankind, people have migrated. In more recent history, people have left their countries of origin voluntarily or involuntarily for different reasons varying from colonialism, war, political persecution, economic scarcity, family reunification, and to search for new and better life experiences. The numbers of individuals immigrating increase every year: From 2000 to 2010, the total number of international immigrants increased from an estimated 150 million to 214 million (IOM, 2010). The geography of migration also changes over time. For example, migration takes place not only from the 'South' to the 'North,' but also more and more within African, Asian, and Latin American regions. Moreover, former emigrant countries in Europe have become receiving countries during the twentieth century.

While migration benefits the world economy and is an important source of income for developing countries, it also poses societal challenges ranging from human trafficking, 'irregular migration' due to restrictions in immigration laws to marginalization of immigrant communities. Indeed, immigrants are often confronted with negative or even hostile reactions from the national majority of the receiving country (i.e., the citizens with no foreign roots). Frequently immigration is debated and politicized in the culturally diverse societies of today. For instance, radical and populist right-wing parties rely on negative representations of immigrants (e.g., immigrants abuse the welfare system) to mobilize voters among the national majority.

This article provides a social psychological perspective to immigration, presenting insights both on national majorities and immigrant minorities. It should be kept in mind however that the distinction between immigrant and ethnic groups is not always clear-cut. Individuals may be born in a given country but have foreign roots. They may also have mixed cultural and ethnic heritage. Finally, despite the fact that indigenous and aboriginal populations are not immigrant minorities, they also often face strong discrimination. Thus, social psychological research on immigration overlaps to some degree with research on national and ethnic minorities.

The article is divided into three sections. First, we discuss how immigration affects identification processes of national majorities and immigrant minorities, and the consequences of identification for intergroup attitudes. Second, we examine the role of threat perceptions, stereotypes, and worldviews in national majorities' stances toward immigrants, and the psychological costs and consequences of discrimination for immigrants. Finally, we discuss the role of intergroup contact between the national majority and immigrants, and of policies and norms in facilitating or hampering harmonious coexistence in culturally diverse societies.

Identification: Defining the In-Group and Relations to Out-Groups

When immigrants settle in a receiving country, both the national majority and immigrants as newcomers create and negotiate boundaries between their in-group and the out-group. This is reflected in in-group identification as well as in attitudes toward members of the other group ranging from tolerant to hostile, and from straightforward to ambivalent. But identities and group boundaries resulting from immigration are not set in stone: They fluctuate over time and as a function of group and individual status.

How the National Majority Defines Itself

Compared to immigrants, national majorities have typically paid less attention to their ethnicity and nationality. However, the increasingly diversified composition of receiving societies has rendered majorities' own ethnic and national background visible, and made them aware of group boundaries. The way majority members define belongingness to the national in-group is a way to evaluate how they react to ethnic diversity. These definitions of nationhood are generally divided into two categories (Smith, 1991). An *ethnic definition*, based on supposed shared ancestry, traditions, and culture, implies an impermeable understanding of the nation: Individuals endorsing this vision believe that immigrants cannot become 'true' members of their receiving country. This may not even be a conscious thought. For instance, white Americans, and to some extent members of ethnic minorities, implicitly and automatically associate being American with being white,

a characteristic that nonwhite individuals can hardly acquire (Devos and Banaji, 2005). Based on the notion of citizenship, a *civic definition* of nationhood implies more permeable group boundaries than the ethnic definition. For example, by learning the local language or acquiring citizenship, immigrants can be accepted as full members of the national in-group.

The way members of national majorities define the national in-group's boundaries is closely related to the nature of their attachment to the nation. Those endorsing an ethnic definition generally express a blind and uncritical allegiance to the nation, perceived as superior to other countries. This form of national attachment – nationalism – frequently underlies negative reactions toward immigrants. Those with a civic vision of the nation, in turn, tend to express a more critical attachment to the nation. This form of attachment – patriotism – involves being proud of the nation's democratic institutions and values, and monitoring whether democracy is actually enacted. Finally, although a civic vision implies a perception of the nation's boundaries as more permeable than an ethnic vision, it nevertheless entails restricting access to the national in-group.

Identities among Immigrant Minorities

Ethnic identity is defined as a sense of belonging based on one's ancestry, cultural heritage, values, traditions, rituals, and often language and religion. Ethnic identity is thus composed of more or less connected facets that can vary across immigrant groups and across situations. For instance, religion may play a more central role in defining identity in some immigrant groups, as it has been found for Muslim immigrants in various countries. Another important distinction has to be made between national and ethnic identity, since for immigrants coming from multiethnic nations, and especially when there are tensions between the different ethnic groups, being attached to this nation holds a different meaning than being attached to one's own ethnic group (Deaux, 2006).

Beyond their relation to the (ethnic, religious, and national) in-group, immigrants also form bonds with the receiving country. The intersection of the desires to maintain the culture of origin and to have contact with members of the receiving country (Berry et al., 2006) or to adopt its culture (Bourhis et al., 1997) forms four *acculturation* strategies. *Separation* occurs when immigrants want to maintain their culture and avoid contacts with the receiving country. On the contrary, *assimilation* implies renouncing the culture of origin and embracing the one of the receiving country. *Integration* denotes the strategy of immigrants who want and do both, while *marginalization* implies that immigrants adopt neither. Refining the notion of marginalization, Bourhis et al. (1997) distinguish *individualism* – immigrants' own decision not to identify with a group because they see themselves as individuals above all – from *cultural alienation* – when immigrants, for various reasons, cannot identify with any group. Integration has been shown to be most beneficial for the psychological well-being of immigrants (e.g., less stress, less anxiety, higher self-esteem).

The adoption of an acculturation strategy is not only guided by one's desires, but also by external factors. Indeed, real or perceived expectations of the receiving society reflected in immigration policies, the history of immigration, the interactions with members of the national majority, as well as the degree of similarity (e.g., in language, culture, or ethnicity) between their country or culture of origin and the receiving country influence immigrants' strategies (Brown and Zagefka, 2011; see Section Policies and Norms in Culturally Diverse Societies). Finally, a distinction between different domains of acculturation is necessary for understanding immigrants' acculturation strategies in the receiving country. For instance, immigrants may favor assimilation in the professional and public sphere but prefer separation for more private matters such as family and religious beliefs.

Beyond an Immigrant vs Receiving Country Citizen Dichotomy

The increasingly diversified flows of immigrants and the long history of migration in many receiving countries have multiplied the landscape of possible identities, creating complex feelings of belonging that cannot be captured by a mere immigrant vs nonimmigrant dichotomy. More and more individuals with an immigrant background are born in the receiving country (the so-called second or third generations) or immigrate at a very young age with their parents (the so-called 1.5 generation). Children and adolescents often acculturate at a faster pace than their parents. As parents are generally keen to transmit the values and culture of the country of origin, differences in acculturation strategies and pace of acculturation may engender intergenerational conflicts within families.

Moreover, individuals with an immigrant background are described as bicultural when more than one cultural mind-set informs and influences their behavior (Hong et al., 2000). Bicultural individuals, and especially those who perceive their two cultures as compatible, interpret the same event differently as a function of the mind-set that is salient at a given moment. For example, Chinese students in the US have been shown to attribute more internal causes (e.g., personal agency) to an event when guided by a Western, individualistic mind-set, and more external and group-related causes when guided by an Eastern, collectivistic mind-set.

Intergroup Conflicts in the Context of Immigration

The arrival and settling of immigrants can give rise to conflicts and resentment between immigrants and members of the receiving society. In the media, in political discourse, and in everyday conversations, immigrants are often portrayed as a threat to the receiving society and its members. Such threats provide a fertile ground for the national majority to express hostility toward immigrants, and legitimize their exclusion from economic, social, and political life (e.g., Wagner et al., 2010). Immigrants on the other hand, face the challenge of finding their place in the receiving society, for example, due to discrimination at school or the workplace. Immigrants therefore develop strategies to overcome these difficulties and voice their claims.

Intergroup Conflict from a Majority Perspective

A perception that immigrants threaten the national majority is often at the origin of negative stereotypes and anti-immigrant prejudice. Such threats can be grouped into two major

categories: realistic threats and symbolic threats (Riek et al., 2006). Realistic threats concern immigrants competing with the majority for tangible resources such as jobs and housing. As immigrants frequently have a lower socioeconomic status and more unstable job prospects than the national majority, they can be seen as a potential financial burden by taking advantage of welfare benefits. Initiatives to acquire rights and to participate in decision making (e.g., by claiming voting rights) may also be perceived as a threat, namely to the existing power arrangements privileging the national majority. The underlying assumption is that whenever groups compete for limited resources, they develop hostile attitudes toward each other. The resources are thereby viewed as a zero-sum, that is, the gain of immigrants inevitably implies a loss for the majority (Esses et al., 1998). Perceived *similarity*, such as equal educational qualifications, between immigrants and the national majority on the job market can fuel perceived realistic threat (Zárate et al., 2004). This is one explanation why opposition to immigration is generally greater among members of the national majority with a low socioeconomic status. Symbolic threats, in turn, refer to the perception that immigrants undermine the values, customs, and traditions of the majority. These threats refer to *differences* between immigrants and the majority, especially when values and beliefs are perceived as incompatible with each other. Those perceiving different cultures as irreconcilable generally express a greater opposition to the incorporation of immigrants into the national in-group (e.g., by opposing lenient citizenship policies), because they may fear that the founding features of the national in-group (e.g., its religion) will lose their distinctiveness. In addition, they often oppose immigrants' attempts to maintain their cultural and religious practices.

Immigrant groups trigger different reactions in majority members. While overall stereotypes of immigrants are negative, stereotypes of specific immigrant groups are more nuanced, and often ambivalent (Lee and Fiske, 2006). The perceived status of immigrants and their country of origin determine emotional reactions toward them. In the US, for instance, while high-status immigrants (such as Asian immigrants) tend to elicit envy or admiration, low-status immigrants (such as undocumented immigrants) tend to elicit pity or contempt. Whether immigrant groups are valued or not also influences the relations that majority members wish to entertain with them. While immigrants from 'valued' (e.g., wealthy and perceived as culturally close) countries are often invited to integrate into the receiving society, immigrants from 'devalued' (e.g., poor and perceived as culturally distant) countries are more frequently expected to assimilate, or are even excluded (Montreuil and Bourhis, 2001; see also Sections Identities among Immigrant Minorities; Policies and Norms in Culturally Diverse Societies). With growing international mobility, the cultural and religious diversity of immigrant groups has increased in many countries. Policies that explicitly attract some immigrant groups while denying entry or family reunification to others are likely to further encourage double standards in attitudes toward immigrants. In addition, the perceived status of immigrants can change over time. Recently arrived immigrants are more likely to be targets of prejudice than older immigrant groups, which are gradually seen as part of the national in-group. Indeed, with increasing attachment to the receiving country and competition for resources, longer established immigrant groups may develop exclusionary attitudes toward more recent immigrant groups.

Moreover, individual differences in worldviews influence majority members' attitudes toward immigrants. Especially individuals who value coercive social control, obedience to authorities, and conformity to norms easily feel threatened by immigrants (Right Wing Authoritarianism, RWA, Altemeyer, 1988). Their attitudes toward immigrants fit their view of the world as dangerous and threatening, where social cohesion, moral values, and collective security are easily undermined by deviants and outsiders. Moreover, individuals who prefer hierarchical rather than egalitarian relations between groups (high on Social Dominance Orientation, SDO; Pratto et al., 1994) often express negative attitudes toward immigrants. Such individuals tend to view the world as ruthlessly competitive, and are quick to judge immigrants as competing with the national majority for limited resources, status, and power. Given their different values and concerns, individuals high on RWA or SDO are sensitive to different characteristics associated with immigrants: Due to their focus on conformity, individuals with RWA tendencies will be threatened by immigrants perceived as different from the majority or unwilling to assimilate. Individuals with high SDO tendencies, in turn, will feel threatened by immigrants perceived as too similar to the national majority, because of the potential decrease in status asymmetry between immigrants and the national majority.

The wider socioeconomic context also shapes individuals' attitudes toward immigrants (e.g., Green and Staerklé, 2013). Indeed, intergroup competition and threat thrive in deprived economic conditions (e.g., high unemployment). Though a high proportion of immigrants (e.g., in a neighborhood or in a region) may increase threat (see however Section Improving Intergroup Relations with Intergroup Contact), the *perception* of a strong immigrant presence is sufficient to trigger threat. Threat is also actively fueled by social and political narratives in the media and political debates, which frequently have a greater influence on attitudes toward immigrants than the actual presence of immigrants (Wagner et al., 2010).

Intergroup Conflict from a Minority Perspective

When settling into a new country, immigrants come across a number of challenges. These challenges can be practical, such as finding a job or learning the language, or involve a legal or political dimension, such as securing residency and achieving political rights. Discrimination by the national majority, including unfair treatment, restricted opportunities, and disrespectful and hostile behavior, creates or strongly contributes to these challenges. Discrimination can be directly experienced in interactions with the majority, or it can be institutionalized and reflected in policies. Experiences of discrimination have severe consequences for the lives of immigrants. They cause psychological symptoms such as distress, low self-esteem, anxiety, and depression. The resulting general and chronic health issues may then hamper immigrants' integration and performance on the job market. Such health issues can be amplified by the segregation of immigrants in poor and unsafe neighborhoods offering restricted health care access. Discrimination is not always direct, however.

The mere awareness of the negative stereotypes that the majority holds of immigrants can render them self-fulfilling prophecies. For instance, stereotype threat – the awareness of others' stereotypes influencing the evaluation of one's performance – also affects immigrants. In the US, second-generation West Indian immigrants, when identifying with black Americans, underperformed at school when they believed academic ability to be measured (Deaux, 2006). Similarly, in the UK, young black men adopted gang-style behaviors generally attributed to their group by the white majority (Howarth et al., 2013). These examples show how majorities' representations of ethnic and immigrant groups – also known by the groups themselves – are likely to maintain and intensify the status gap between immigrants and the majority.

Immigrants can counteract negative stereotypes and discrimination in different ways. According to the Rejection Identification Model, reinforcing identification with the in-group may help immigrants cope with discrimination (Schmitt and Branscombe, 2002). Indeed, increased identification and strong bonds with the in-group is a source of self-esteem that buffers the consequences of discrimination. Strong bonds with in-group members are also primordial for enabling collective action. Immigrants embedded in ethnic networks, but with social ties to the majority, engage more readily in protest requesting rights and entitlements of immigrants (Klandermans et al., 2008). Experiences of discrimination and feelings of injustice can also lead to less-accepted forms of politicization. For example, British Muslims, who thought that the way British authorities treated Muslims abroad was illegitimate and unjust, were more willing to support both normative (e.g., signing a petition) and nonnormative (e.g., violence against Western troops in Muslim countries) forms of collective action (Tausch et al., 2011).

Living Together in a Diverse Society

In culturally diverse societies, everyday interactions between members of immigrant minorities and national majorities are commonplace. The resulting contacts influence the attitudes of both parties. Moreover, policies and societal norms shape the way in which individuals perceive and think about the relations between immigrants and the national majority. In this section, the notion of *multiculturalism* has several meanings (Berry and Sam, 2013). Multiculturalism implies the presence of cultural diversity. It is also an ideology denoting the willingness of the national majority to maintain and promote cultural diversity and to have contact with immigrants. Finally, multiculturalism also refers to a public policy response to cultural diversity.

Improving Intergroup Relations with Intergroup Contact

Positive intergroup contact has been considered one of the most effective ways of reducing prejudice, especially among majority group members. In the context of immigration this implies that anti-immigrant prejudice of national majority members reduces when they have positive interactions with immigrants. The classic intergroup contact hypothesis states that the optimal conditions for reducing prejudice in an intergroup contact situation include equal status between the groups, common goals, cooperation in achieving these goals, and a supportive institutional climate, for example, in the form of integration policies (Allport, 1954). When immigrants have similar social status to the national majority, reaching these conditions is more plausible. However, as immigrants are frequently at the bottom of the societal hierarchy in the receiving countries, equal status and common goals (due to different needs and interests) are rarely reached in a relationship between a national majority member and an immigrant. Nevertheless, there is now extensive evidence that contact improves attitudes toward immigrants even when these optimal conditions are not fulfilled (Pettigrew and Tropp, 2006).

To have practical value, it is crucial that the impact of positive contact generalizes beyond a specific contact experience to new situations, to the entire immigrant group, and even to other immigrant groups. Three different models explain how contact generalizes to positive attitudes toward immigrants (for overviews Brown and Hewstone, 2005; Hewstone and Swart, 2011). First, positive interpersonal interactions between an immigrant and a national majority member can render salient the *personal* characteristics of the immigrant, instead of category membership. Immigrants are seen as distinct individuals, rather than as a homogenous group, thereby blurring the boundaries between the two categories. Such *decategorization* can result in more positive attitudes toward immigrants. Second, contact can also lead to redrawing of group boundaries, or *recategorization*, where immigrants and members of the receiving society are associated with the shared superordinate group (e.g., 'we are all Europeans'). Embracing this common in-group identity then reduces prejudice. A third model suggests, in turn, that the *distinction* between immigrants and the majority should be *salient*: It is necessary to conceive the situation as *intergroup* contact for the effects of positive contact to generalize to the immigrant out-group as a whole. These models can be considered complementary insofar as contact requires both interpersonal and intergroup significance for positive immigrant attitudes to generalize beyond the contact situation.

How does contact reduce anti-immigrant prejudice? Different social psychological processes (mediational mechanisms) have been put forward (Brown and Hewstone, 2005). First, contact enables one to learn about the life of immigrants, and this *new knowledge* then reduces prejudice. Second, agreeable interactions between immigrants and the receiving society majority *increase positive* and *decrease negative effect* toward immigrants. Typically positive contact with immigrants reduces intergroup threat and anxiety. Close contact, and in particular friendships, also allows taking the perspective of immigrants and empathizing with their concerns, which then improves intergroup attitudes. Finally, *self-disclosure* (revealing important aspects of oneself to another person) contributes to the development of interpersonal relationships. Interactions between an immigrant and a national majority member involving self-disclosure thus further improve intergroup attitudes.

Even indirect or distal contact with immigrants reduces anti-immigrant prejudice. First, simply knowing that fellow in-group members befriend immigrants can improve attitudes

toward immigrants even in the absence of direct personal contact. Observing such friendships, in other words experiencing *extended contact*, conveys tolerant in-group norms regarding intergroup relations and increases empathy toward and knowledge about immigrants, which then ameliorate attitudes toward immigrants. Second, contact reduces prejudice against immigrant groups that were not directly involved in the initial interaction. This phenomenon, labeled a *secondary transfer effect*, implies improved attitudes toward one immigrant group that results from positive contact with members of another immigrant group. Third, merely *imagining positive intergroup interactions* has the potential for improving immigration attitudes. Finally, living in culturally diverse societal contexts (depicted with demographic indices such as a high proportion of immigrants in the whole population) provides more contact opportunities. Such opportunities can give rise to positive intergroup contact, for example, through friendships, which then decrease perceived threat and antagonistic attitudes toward immigration (however, for an opposite prediction of the link between proportion of immigrants and increased feelings of threat, see Section Intergroup Conflict from a Majority Perspective).

Intergroup contact also has downsides and can even be counterproductive for improving the situation of immigrants. Positive, more profound interactions indeed reduce threat perceptions and anti-immigrant prejudice. However, superficial, everyday contacts are more frequent (Dixon et al., 2005) than positive, more profound contacts. Superficial contact, especially when it is negative, may consolidate and even heighten anti-immigrant prejudice. The prejudice-increasing effect of negative contact is frequently greater than the prejudice-reducing effect of positive contact. This means that negative experiences (e.g., hostile interactions in the street) with and negative impressions of immigrants have more impact on increasing prejudiced attitudes than pleasant encounters and positive impressions have decreasing them. Also one cannot assume that cultural diversity and actual contact go hand in hand. Indeed, ethnic and immigrant minorities may be informally segregated, and they may even voluntarily do so themselves. For example, actual contact is minimized if immigrants use public space differently to the national majority (e.g., sitting arrangements in classrooms, use of leisure areas). These factors may explain the persistence of prejudice in culturally diverse communities.

Although classical intergroup contact research has advanced the understanding of the processes involved in prejudice reduction, more recent research examines how the way immigrant and minority groups interpret intergroup contact affects strategies for seeking social change and recognition. Positive attitudes toward the national majority can conflict with the interests of immigrants. Pleasant intergroup interactions can actually reinforce the unequal status relations between immigrants and national majorities by disempowering immigrants and thus efforts to realize social change. Positive *interpersonal* contact attracts attention away from perceptions of *intergroup* inequality between disadvantaged immigrants and privileged national majorities, thereby reducing collective action as well as support for policies aimed at combating inequalities and increasing emancipation (e.g., Wright and Baray, 2012 for this phenomenon). Thus, the reduction of prejudice should not be the sole aim in culturally diverse societies. In turn, collective action opposing societal inequality and reducing power differences among immigrant minorities and the national majority should be encouraged.

Policies and Norms in Culturally Diverse Societies

Countries differ in the way they deal with diversity, recognition of immigrant minorities, and their incorporation in society. Multicultural and assimilationist integration policies, and the underlying ideologies, are frequently differentiated. Multicultural policies have two principal goals (e.g., Berry and Sam, 2013): On the one hand, they involve supporting the maintenance of immigrants' culture and tradition and on the other hand, they focus on promoting the inclusion and participation of immigrants in society thereby reducing discrimination, increasing equality, and improving the economic and societal position of immigrants. This is the case in Canada, where multicultural policies have included both the recognition of cultural diversity and the equity and inclusion of immigrants and ethnic minorities. In contrast, in the US and UK policies have mainly focused on the recognition of diversity. Assimilation policies, in turn, are based on the idea that immigrants should adopt the language, customs, and values of the national majorities, and abandon their own cultural heritage. Assimilationist policies thus aim to homogenize the population and to reduce cultural diversity. France has been described as an assimilationist country that expects immigrants to adopt French culture. For example, the controversial law banning visible religious symbols such as the headscarf at school can be interpreted as an assimilationist expectation. In short, while multicultural policies emphasize appreciating differences between immigrants and the national majority, assimilationist policies focus on de-emphasizing these differences.

Multiculturalism and assimilationism are also social psychological concepts referring to *individuals'* acceptance of cultural diversity and support for policies enabling and promoting cultural diversity (Verkuyten, 2007, for parallels with immigrants' acculturation strategies, see Section Identities among Immigrant Minorities). A multicultural view implies awareness and appreciation of differences between immigrants and the national majority. These views are related to proimmigration attitudes, endorsing supportive integration policies, as well as to efforts to improve the societal status of immigrants. Assimilationist views, in turn, oppose such goals. Frequently, multiculturalism appeals more to immigrants than to the national majority who may prefer assimilationism. But not all aspects of multiculturalism are supported to the same degree nor are they applied to all immigrant groups. There may, for example, be high agreement for promoting equal participation in society, whereas encouragement and the provision of financial aid for cultural maintenance gain less support.

Immigrants and national majorities need to *share* the principles of mutual recognition, because a mismatch in expectations of immigrants and the national majority impedes the attainment of a diverse and harmonious society. The interactive acculturation model (Bourhis et al., 1997) defines the ways intergroup relations play out between immigrants and the

national majority depending on the fit between immigrants' and the majorities' attitudes toward the way diversity is organized in society. A consensual fit between the immigrants' strategies and the receiving country's expectations engenders positive intergroup outcomes (see Section Identities among Immigrant Minorities; Brown and Zagefka, 2011). In contrast, a marked mismatch increases negative and conflictual intergroup relations. For example, religious diversity, in particular the role of Islam in Western societies, is currently fiercely debated. Assimilation expectations and the rejection of Muslim values made salient in such debates may threaten Muslim identification, and increase the need to defend and reaffirm one's religion (Verkuyten, 2007). This can then lead to a spiral of mutual disrespect and rejection between the national majority and Muslim immigrants.

Importantly, multicultural and assimiliationist attitudes *develop in individuals' everyday environments* providing normative reference knowledge that guides thinking about immigration (Green and Staerklé, 2013), for example, through the awareness of national policies that condemn or justify anti-immigrant stances. Thus, multicultural or assimiliation policies define the *institutionalized* normative context within a country that orient, in turn, people's attitudes toward immigrants. These policies (such as the multicultural policies in Canada) can influence personal attitudes toward multiculturalism and assimilation. Moreover, a shift toward more inclusive and tolerant integration policies also translates into heightened support for multiculturalism among the national majority. The impact of policies is not necessarily direct. For instance, policies affect perceived norms related to multiculturalism and assimilation, that is, how individuals perceive cultural diversity is dealt with in their country and what is perceived as the expected way to view and treat immigrants (see Guimond et al., 2013). In contrast, by blaming immigrants for societal issues such as unemployment or welfare expenditures, right-wing parties convey an exclusionary normative climate regarding immigration. Such anti-immigrant stances thereby serve important social functions to consolidate and legitimize the position of political leaders and to sustain power relations. Intolerant beliefs widely shared by the national majority also create a climate that fosters anti-immigrant prejudice.

While national policies shape perceived norms on a macro level, the influence of multicultural education takes place on a more proximal level, frequently in a classroom context (see Verkuyten and Thijs, 2013). Indeed, the very objectives of multicultural educational practices are to improve equality between immigrants and the national majority by conveying and establishing antiprejudice norms as well as increasing recognition of existing cultural differences. Learning to accept and recognize cultural differences and prescribing how to treat members of other groups are seen as a route to tolerant attitudes that will help building harmonious intergroup relations. However, despite good intentions, multicultural education is not without pitfalls. Highlighting differences can backfire, for example, if negative stereotypes are reinforced, thereby leading to an increase in prejudiced attitudes. Moreover, sometimes antiprejudice norms are learned superficially without internalization. Only when norms are internalized, that is when individuals personally endorse them, tolerant stances will generalize beyond the classroom.

See also: Acculturation; Authoritarian Personality; Ethnic Identity, Psychology of; Ethnicity and Migration in Europe; Intergroup Relations; Racism: Social Psychological Perspectives; Social Dominance Orientation; Social Psychology; Xenophobia: Social Psychological Aspects.

Bibliography

Allport, G.W., 1954. The Nature of Prejudice. Doubleday, New York.
Altemeyer, B., 1988. Enemies of Freedom: Understanding Right-Wing Authoritarianism. Jossey-Bass, San Francisco, CA.
Berry, J.W., Phinney, J., Sam, D.L., Vedder, P. (Eds.), 2006. Immigrant Youth in Cultural Transition: Acculturation, Identity and Adaptation across National Contexts. Lawrence Erlbaum, Mahwah, NJ.
Berry, J.W., Sam, D.L., 2013. Accommodating cultural diversity and achieving equity. European Psychologist 18, 151–157.
Bourhis, R.Y., Moïse, L.C., Perreault, S., Senécal, S., 1997. Towards an interactive acculturation model: a social psychological approach. International Journal of Psychology 32, 369–383.
Brown, R., Hewstone, M., 2005. An integrative theory of intergroup contact. Advances in Experimental Social Psychology 37, 255–343.
Brown, R., Zagefka, H., 2011. The dynamics of acculturation: an intergroup perspective. In: Olson, J.M., Zanna, M.P. (Eds.), Advances in Experimental Social Psychology, vol. 44. Academic Press, Burlington, pp. 129–184.
Deaux, K., 2006. To Be an Immigrant. Russell Sage Foundation, New York.
Devos, T., Banaji, M.R., 2005. American = white? Journal of Personality and Social Psychology 88, 447–466.
Dixon, J., Durrheim, K., Tredoux, C., 2005. Beyond the optimal contact strategy: a reality check for the contact hypothesis. American Psychologist 60, 697–711.
Esses, V.M., Jackson, L.M., Armstrong, T.L., 1998. Intergroup competition and attitudes toward immigrants and immigration: an instrumental model of group conflict. Journal of Social Issues 54, 699–724.
Green, E.G.T., Staerklé, C., 2013. Migration and multiculturalism. In: Huddy, L., Sears, D.O., Levy, J. (Eds.), Oxford Handbook of Political Psychology. Oxford University Press, Oxford, pp. 852–889.
Guimond, S., Crisp, R.J., De Oliveira, P., Kamiejski, R., Kteily, N., Kuepper, B., Lalonde, R.N., Levin, S., Pratto, F., Tougas, F., Sidanius, J., Zick, A., 2013. Diversity policy, social dominance, and intergroup relations: predicting prejudice in changing social and political contexts. Journal of Personality and Social Psychology 104, 941–958.
Hewstone, M., Swart, H., 2011. Fifty-odd years of inter-group contact: from hypothesis to integrated theory. British Journal of Social Psychology 50, 374–386.
Hong, Y.-Y., Morris, M.W., Chiu, C.-Y., Benet-Martinez, V., 2000. Multicultural minds: a dynamic constructivist approach to culture and cognition. American Psychologist 55, 709–720.
Howarth, C., Wagner, W., Magnusson, N., Sammut, G., 2013. 'It's only other people who make me feel black': acculturation, identity, and agency in a multicultural community. Political Psychology. Advance online publication.
IOM, 2010. World Migration Report 2010. International Organization for Migration.
Klandermans, B., Van der Toorn, J., Van Stekelenburg, J., 2008. Embeddedness and grievances: collective action participation among immigrants. American Sociological Review 73, 992–1012.
Lee, T.L., Fiske, S.T., 2006. Not an outgroup, not yet an ingroup: immigrants in the stereotype content model. International Journal of Intercultural Relations 30, 751–768.
Montreuil, A., Bourhis, R.Y., 2001. Host majority acculturation orientations towards 'valued' and 'devalued' immigrants. Journal of Cross-Cultural Psychology 32, 718–739.
Pettigrew, T.F., Tropp, L.R., 2006. A meta-analytic test of intergroup contact theory. Journal of Personality and Social Psychology 90, 751–783.
Pratto, F., Sidanius, J., Stallworth, L.M., Malle, B.F., 1994. Social dominance orientation: a personality variable predicting social and political attitudes. Journal of Personality and Social Psychology 67, 741–763.
Riek, B.M., Mania, E.W., Gaertner, S.L., 2006. Intergroup threat and outgroup attitudes: a meta-analytic review. Personality and Social Psychology Review 10, 336–353.

Schmitt, M.T., Branscombe, N.R., 2002. The meaning and consequences of perceived discrimination in disadvantaged and privileged social groups. In: Stroebe, W., Hewstone, M. (Eds.), European Review of Social Psychology, vol. 12. Wiley, Chichester, pp. 167–199.

Smith, A.D., 1991. National Identity. Penguin, London.

Tausch, N., Becker, J.C., Spears, R., Christ, O., Saab, R., Singh, P., Siddiqui, R.N., 2011. Explaining radical group behavior: developing emotion and efficacy routes to normative and nonnormative collective action. Journal of Personality and Social Psychology 101, 129–148.

Verkuyten, M., 2007. Social psychology and multiculturalism. Social and Personality Psychology Compass 1, 1–16.

Verkuyten, M., Thijs, J., 2013. Multicultural education and inter-ethnic attitudes. European Psychologist 18, 179–190.

Wagner, U., Christ, O., Heitmeyer, W., 2010. Anti-immigration bias. In: Dovidio, J.F., Hewstone, M., Glick, P., Esses, V.M. (Eds.), Handbook of Prejudice, Stereotyping, and Discrimination. Sage, Los Angeles, CA, pp. 361–376.

Wright, S.C., Baray, 2012. Models of social change in social psychology: collective action or prejudice reduction? Conflict or harmony? In: Dixon, J., Levine, M. (Eds.), Beyond Prejudice: Extending the Social Psychology of Conflict, Inequality and Social Change. Cambridge University Press, Cambridge, pp. 225–247.

Zárate, M.A., Garcia, B., Garza, A.A., Hitlan, R.T., 2004. Cultural threat and perceived realistic group conflict as dual predictors of prejudice. Journal of Experimental Social Psychology 40, 99–105.

Imperialism

David A Lake, University of California, San Diego, La Jolla, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Imperialism is a form of international *hierarchy* in which one political community effectively governs or controls another political community. It is one of the oldest known political institutions, characterizing relations between peoples in ancient Mesopotamia, China, and Rome through modern Europe. It includes rule both within relatively contiguous areas and overseas colonies. Major explanations for imperialism include metrocentric theories, which focus on the internal characteristics of imperial states; pericentric theories, which emphasize conditions within the colonial polities; systemic theories, which highlight competition between the great powers; and relational contracting theories, which explain imperialism by contrast with other possible institutions.

Imperialism

Imperialism is a form of international *hierarchy* in which one political community, or polity, effectively governs or controls another political community. It is one of the oldest known political institutions, characterizing relations between peoples in ancient Mesopotamia, China, and Rome through modern Europe. It includes rule both within relatively contiguous areas – as in the Habsburg and Ottoman empires – and the overseas colonies held by various European states after the Age of Discovery.

The term has a long and tortured history. It was apparently first used as an invective against the expansionist policies of Napoleon I and has been employed most frequently to refer to the colonial practices of the European states in the late nineteenth and early twentieth centuries, attempts to revise the international territorial status quo, and the economic domination of one country by another (also referred to as 'neocolonialism,' see below). Used as a tool of political rhetoric, the term is highly malleable and often devoid of any general meaning.

As an analytic concept, imperialism refers to the effective domination of one political community by another. According to Michael Doyle (1986, 19), "(e)mpires are relationships of political control imposed by some political societies over the effective sovereignty of other political societies …. Imperialism is the process of establishing and maintaining an empire." By domination we mean the ability of the dominant polity – the metropole – to decide policy for the subordinate policy – the colony. This ability may entail substantial delegation of decision-making authority to elite members of the colony, but the metropole retains the power to decide what gets delegated and how, and when this authority is revoked.

Three corollaries are important. First, imperialism occurs only where distinct political communities exist. Subordinate communities may have a prior history of independence, or a new political consciousness may emerge that creates a relationship of imperialism. Without the possession of a distinct political identity, however, class conflict or unequal political opportunities may exist, but not imperialism. The question of what constitutes a distinct political community introduces significant ambiguity about any particular imperial relationship, particularly in previously integrated and territorially contiguous states. Conversely, it is the clarity of the communities that makes imperialism so easily identified with the European overseas empires constructed during the nineteenth century.

Second, the colony lacks an international political 'personality'; that is, while it possesses an identity as a distinct polity, it does not interact with other states as a sovereign equal. Finally, exploitation of the weak by the strong is not inherent in imperialism, but it is an often natural outgrowth of effective domination. The affinity between domination and exploitation explains the often pejorative status of the term.

Imperialism is an extreme form of international hierarchy in which the colony is, in principle, a subject of the dominant state. If imperialism forms one end of a continuum of international hierarchies and 'anarchic' relations between sovereign equals form the other, we can identity a range of increasingly hierarchical relationships, all of which are sometimes called imperialistic. In *spheres of influence*, the subordinate members remain independent but are constrained by dominant powers from forming relationships, such as alliances, with other great powers. Latin America under the *Monroe Doctrine* is a classic example. In *protectorates*, subordinate states yield control over their foreign and defense policies to dominant powers; although subordinates remain independent, they transfer control over specific areas of policy to other states. With continuing responsibility for their defense, the *United States* today retains protectorates over the Federated States of *Micronesia* and the Republic of the *Marshall Islands*. In *informal empires*, subordinates are subject to imperial states across wide ranges of policy but retain their international personalities and interact with third parties on the basis of sovereign equality. *Eastern Europe* under the *Soviet Union* is a particularly clear example. Protectorates and informal empires are often grouped with formal empires as forms of imperialism.

Neocolonialism is a hierarchy produced through the functioning of an impersonal international market. In this variant, the dominant state need not intend to control the subordinate, but the latter is sufficiently dependent upon the former economically that it has little choice other than to comply with (and even anticipate) the metropole's desires. In this case, economic dependence produces political dominance – the core

of imperialism – but the mechanism of control is indirect. For some, this is a virulent form of modern imperialism. For others, the absence of intent negates the political relationship. Neocolonialism remains a contested concept.

Theories of Imperialism

The major explanations for imperialism can be grouped into three general categories. Metrocentric theories focus on the dispositions or internal characteristics of imperial states. Writing in 1902[1965], for instance, John Hobson grounded the motivation for overseas expansion in the necessity for advanced capitalist states to export their surplus capital. This theme was later the foundation for V.I. Lenin's famous monograph, *Imperialism: The Highest Stage of Capitalism*, first published in 1917. *Neo-Marxists* later argued that the *military-industrial complex* and other features of capitalist states actually produced a need for capital, leading states to create colonial and neocolonial relations with developing regions in order to extract wealth (see Magdoff, 1969). Offensive realist theories similarly ground imperialism in the dispositions of great powers and their inherent desires to pursue power over others (Mearsheimer, 2001).

Pericentric theories emphasize conditions within the colonial polities. Where metrocentric theories focus on the push behind expansion, pericentric theories draw attention to the forces that pull imperialists into hierarchical relationships. John Gallagher and Ronald Robinson's *The Imperialism of Free Trade* (1953) set the direction for much research in this tradition. Positing a constant drive for domination and a preference for indirect rule wherever possible, Gallagher and Robinson explained variations in imperialist outcomes by conditions in the periphery. In particular, where peripheral polities possessed stable regimes and effective collaborators, they argued, imperialists could govern indirectly through informal empires. Only where the peripheral societies were unstable politically or lacked elites willing to protect their interests would metropoles be forced to create formal empires and govern directly. As it was effectively ruled by landed interests tied to the British market, in their view, *Argentina* escaped the need for recolonization but was nonetheless caught within Britain's free trade web. Later work on neocolonialism further developed these insights.

Systemic theories of imperialism, typically part of larger defensive *realist* theories of international relations, highlight competition between the great powers (see Cohen, 1973; Waltz, 1979). The struggle for survival and influence between great powers creates an ever-widening gyre of competition, in this perspective, that both leads metropoles to seize territories to augment their resources and allows them to compensate one another using peripheral territories to maintain an effective balance of power. The classic case of imperialism driven by systemic competition was the so-called *Race for Africa* in the late nineteenth century.

The best explanations of imperialism have always combined ideas from more than one of these traditions. Even Hobson, and Gallagher and Robinson, for instance, augmented their theories with a focus on systemic competition, which they both saw as accentuating the metropolitan or peripheral causes of imperialism. More recently, synthetic works integrating all the three approaches have appeared that provide relatively complete explanations of imperialism (see Doyle, 1986; Smith, 1981; Blanken, 2012). Although different authors emphasize different dimensions and different episodes, any single instance of imperialism carries traces of all the three sets of factors. The 'state of the art,' therefore, recognizes and builds upon metrocentric, pericentric, and systemic insights.

A fourth explanation is also being developed that draws upon *neo-institutionalist theory* and especially theories of *relational contracting* as developed in economics. In this approach, imperialism is understood as simply another form of organizational hierarchy. Emphasis is placed on explaining why hierarchy, in general, and this form, in particular, is superior for obtaining the goals of the actors rather than alternative institutions such as *confederations*, protectorates, and state-to-state relations. Three considerations are central. First, the larger the benefits from pooling resources and efforts between the units, in this case the metropole and colony, the more risks or costs the units are willing to accept in building a relationship. Second, the greater the expected costs of *opportunistic* behavior by the subordinate polity, the more important it is to the dominant state to control the actions of its partner. Especially significant here are assets that are specific to the two parties and that would be costly to the metropole if they were withdrawn. In the absence of some *specific assets*, we would expect an 'arm's-length' relationship between independent units, as the benefits of cooperation could then be captured through market exchange. Only where there are substantial assets that are specific to the two parties is imperialism necessary. Finally, there are costs to governing any relationship, and these are likely to escalate with greater hierarchy. Imperialism is typically costly both for the colony, which gives up valued freedom, and for the metropole, which must either bind its own hands to limit its ability to exploit its partner or use coercion to impose its rule. Thus, states avoid empire if control is either unnecessary or can be achieved in some less hierarchical fashion. In this approach, empire is most likely when there are large benefits from pooling resources between the units, highly specific assets at risk, and the costs of governance do not rise sharply with greater hierarchy.

Employing this approach in the case of classic imperialism, Jeffry Frieden (1994) argues that formal empires arose in the age of *plantation agriculture* and *raw materials extraction* as a means of preventing local elites from appropriating site-specific assets in production: once foreign investors opened a copper mine, for instance, this site-specific asset was then subject to expropriation by the local government. Knowing this, foreigners would not invest in such assets without the control made possible by formal empire. As *multinational corporations* have increasingly taken the place of agricultural or raw materials investors, Frieden continues, formal empires have become obsolete. Multinational corporations possess firm-specific assets such as technology or brand names that cannot be easily appropriated by host governments; a government can seize an axle assembly plant, for example, but it is of little value without access to the multinational's worldwide production and marketing network. Thus, formal empire is less necessary to protect corporate assets. David A. Lake (1999, 2009), Alexander Cooley (2005), and others have applied similar arguments to hierarchies motivated by national security

needs. In its emphasis on comparative institutions, this approach brings an additional dimension to the study of empires.

The End of Empire?

Decolonization was one of the most significant events in modern world history. At the dawn of the twentieth century, most of the globe was ruled from Europe. By 1963, the overseas empires had seemingly evaporated. The strong norm against imperialism as affirmed in numerous United Nations declarations strongly suggests that the age of empire is over. Its imprint endures, however. In the periphery, imperial rule destroyed local structures of governance. In some areas, such as *India*, the foundations of parliamentary rule were erected in their place. In other regions, especially Africa, only weak state structures existed when independence arrived, producing political instability and continuing impediments to economic progress (see Jackson, 1990). In the imperial states themselves, political elites underwent a wrenching process of accommodation to their now-diminished international political and economic status (see Kahler, 1984).

Although scholarly interest waned with decolonization, imperialism nonetheless remains a vibrant issue in the contemporary world. Weakly institutionalized groups within *multinational states*, such as the Chechens in *Russia* or the Tibetans in *China*, are pressing demands for independence, declaring in words and actions that the states within which they are embedded are empires. Other groups are rediscovering nationalist identities, such as the Scots in *Great Britain*, and asserting new political rights – transforming intrastate politics into imperial politics. Even as the overseas empires have ended, a new age of imperial struggle is emerging.

This new age will be politically more complex than the old. The principle of national self-determination, articulated by President Woodrow Wilson at the *Versailles Peace Conference*, could be easily employed in the fight against overseas empires. Today, this principle directly contradicts contemporary notions of sovereignty as an inviolable whole. As demonstrated in the *NATO* intervention in *Kosovo*, Western nations abhorred the political domination of the local Albanians by their *Serb* rulers, but they could condone the breakup of a sovereign state only as a last resort. As both old and new imperial relationships within multinational states become contested, politicians and analysts alike will need to rethink issues of political identity, statehood, and empire and perhaps experiment with new forms of political hierarchy.

Interest in imperialism as a term and concept has also waxed with the rise of US unipolarity after the end of the Cold War and, especially, the Iraq War of 2003. As in its first use as a critique of Napoleon's ambitions in Europe, critics of the US war against Iraq charged the Bush administration with imperialism. From a rare term in political discourse, seldom uttered since the Vietnam War, the phrase 'American empire' appeared over 1000 times in US newspapers between December 2002 and the start of the war in May 2003. The term was also taken up by a slew of authors (among others, see Bacevich, 2002; Ferguson, 2004). In describing the United States as imperialist, critics sought to characterize the war as an illegitimate use of power against another sovereign state. Yet, the term was also embraced by the Bush administration, as when an unidentified but high-ranking official told reporter Ronald Suskind, "We're an empire now, and when we act, we create our own reality." With the end of the Iraq War, the use of the term has once again waned. Nonetheless, its long history as both practice and invective ensures that imperialism will remain a robust political relationship into the future and a deep source of disagreement and discord in world politics.

See also: Institutionalism; Marxism–Leninism: The Ideology of Twentieth-Century Communism; Power; Realism/Neorealism; State Formation, Theory of; War and Nationalism; Wars among Nation-States: Patterns and Causes.

Bibliography

Bacevich, Andrew J., 2002. American Empire: The Realities and Consequences of U.S. Diplomacy. Harvard University Press, Cambridge, MA.

Blanken, Leo J., 2012. Rational Empires: Institutional Incentives and Imperial Expansion. University of Chicago Press, Chicago.

Cohen, Benjamin J., 1973. The Question of Imperialism: The Political Economy of Dominance and Dependence. Basic Books, New York.

Cooley, Alexander, 2005. Logics of Hierarchy: The Organization of Empires, States, and Military Occupations. Cornell University Press, Ithaca, NY.

Doyle, Michael W., 1986. Empires. Cornell University Press, Ithaca, NY.

Ferguson, Niall, 2004. Colossus: The Price of America's Empire. Penguin Press, New York.

Frieden, Jeffry A., 1994. International investment and colonial control: a new interpretation. International Organization 48, 559–593.

Gallagher, John, Robinson, Ronald, 1953. The imperialism of free trade. Economic History Review 6, 1–15.

Hobson, J.A., 1902[1965]. Imperialism: A Study. University of Michigan Press, Ann Arbor.

Jackson, Robert H., 1990. Quasi-states: Sovereignty, International Relations and the Third World. Cambridge University Press, New York.

Kahler, Miles, 1984. Decolonization in Britain and France: The Domestic Consequences of International Relations. Princeton University Press, Princeton, NJ.

Lake, David A., 1999. Entangling Relations: U.S. Foreign Policy in Its Century. Princeton University Press, Princeton, NJ.

Lake, David A., 2009. Hierarchy in International Relations. Cornell University Press, Ithaca, NY.

Mearsheimer, John J., 2001. The Tragedy of Great Power Politics. W.W. Norton, New York.

Magdoff, Harry, 1969. The Age of Imperialism: The Economics of U.S. Foreign Policy. Monthly Review Press, New York.

Smith, Tony, 1981. The Pattern of Imperialism: The United States, Great Britain, and the Late-industrializing World since 1815. Cambridge University Press, New York.

Waltz, Kenneth N., 1979. Theory of International Politics. Addison-Wesley, Reading, MA.

Imperialism, History of

Simon C Smith, University of Hull, Hull, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H.L. Wesseling, volume 11, pp. 7226–7232, © 2001, Elsevier Ltd.

Abstract

The word 'imperialism' is one of the most powerful concepts of our time. Originally coined as a term to describe and analyze the expansion of the European powers over the rest of the world in the last quarter of the twentieth century, the term was soon employed to also characterize a particular stage in the development of capitalism, especially by Marxist thinkers, and later developed into an invective for describing the activities of one's political adversaries. In scholarly writing, the word mostly indicates the extension of what was largely European formal or informal political rule over Asian and African countries in the late nineteenth and twentieth centuries. It also serves more generally to describe some other forms of Western predominance during and after the colonial period. Prior to the end of the Second World War, an economic interpretation of imperialism was prevalent, but after that time, other theories were brought forward giving more attention to other aspects, particularly political ones. More recently, the primacy of economic interpretations, especially of British imperialism, has been reasserted. In the vast literature on British, French, German, Italian, Portuguese, and Dutch imperialism that developed from the 1960s onward, the various national forms of this general European phenomenon have been studied. The rather different cases of Russian, American, and Japanese expansion have also been analyzed from the perspective of the theory of imperialism. In modern scholarly literature, imperialism is no longer conceived of as a unilateral process imposed by a dominating power but as a process of interaction between the metropole and the periphery.

Introduction: The Problem of a Definition

"Imperialism is not a word for scholars", Sir Keith Hancock remarked a long time ago, and he was right (see Wesseling, 1997: p. 74). More recently, Bernard Porter has written that "'Imperialism' is a much vaguer term than 'the empire'. Its extent and limit are less definite, and for that reason open to interpretation and dispute" (Porter, 2012: p. 4). Scholars have to make clear what they mean when they use certain concepts or terms, and therefore have to provide definitions. This, however, is problematic with the word 'imperialism'. The difficulty is not that there is no single definition of imperialism. Rather, there are about as many definitions of imperialism as there are authors who have written on the subject. They vary from those that refer to one specific form of imperialism, mostly Europe's nineteenth-century colonial expansion, to others that give a very general meaning to the word, such as the one in Webster's Dictionary: 'any extension of power or authority or an advocacy of such extension'. Clearly, such a definition can cover almost any situation involving a stronger and a weaker power. Not surprisingly, therefore, the word has often simply been used pejoratively in order to criticize the policy of another country. With some justification, P.J. Marshall has noted: "'imperialism' has tended to become a term of abuse for any supposed domination which the speaker happens to dislike" (Marshall, 1982: p. 49).

Used in this context, the term imperialism is of little scholarly value. However, in academic studies, the word has always had a more limited meaning. The problem is exactly how limited its meaning should be. Sometimes the word is used in a universal historical way in order to characterize the politics of a dominant power. Thus, some historians have spoken of Roman or even Assyrian imperialism, but this is highly exceptional. John Darwin has succinctly defined imperialism as the "sustained effort to assimilate a country or region to the political, economic or cultural system of another power" (Darwin, 1997: p. 614). In historical studies, imperialism generally refers to the policy of European countries, and primarily of Britain during the nineteenth and twentieth centuries, aiming at the expansion of their power and influence over other continents. It is in this context that the term imperialism originated and began to be used as a political and historical concept. Historically speaking, the word imperialism is therefore obviously closely associated with colonialism. While colonialism was only used to refer to one specific form of alien rule, namely, the colonial one, imperialism acquired a wider meaning and included various other forms of influence over alien nations and states, for example, the financial influence of France and Germany in the Russian and Ottoman Empires, or such ideas as British 'gunboat policy' and American 'dollar diplomacy'.

After the end of the colonial empires, the word 'colonialism' could only be used to refer to a phenomenon from the past and thus began to fall out of use. 'Imperialism', however, continued to be used, and from then on also indicated those forms of domination that were formally different from, but factually comparable to, those formerly practiced by the colonial powers. For a while, the word 'neocolonialism' was also used for this purpose, but somehow this term was less successful. By the end of the Second World War, America had become the new superpower. Accordingly, imperialism was now mainly applied to describe the foreign policy of the United States vis-à-vis other countries, in particular in Latin America, Asia, and Africa (Ferguson, 2005). There was also an attempt to make the concept applicable to the policy of the Soviet Union with regard to the Central and Eastern European countries that came under its influence after 1945 (Seton-Watson, 1961), but this was not very successful. The reason for this is that historically speaking, imperialism has connotations with capitalism and not with communism, and with overseas possessions, and not

with adjacent countries. Although there clearly was a Soviet Empire, it was not considered to be an example of imperialism but of traditional power politics. Only in its very general meaning as another word for all forms of power politics or simply as an invective, was it also used to describe communist countries such as the Soviet Union and China. After the end of the Cold War, this use of the word imperialism lost much of its earlier attraction.

In this article, imperialism is used in the sense of its initial meaning, that is to say as a term to indicate the extension of formal or informal, mostly European, rule over Asian and African countries in the late nineteenth and early twentieth centuries as well as, more generally, for some other forms of Western predominance during and after the colonial period.

Imperialism: The History of a Concept

Like 'colonialism', which was probably first used in the title of a book of a French socialist critic of the phenomenon, Paul Louis' *Le Colonialisme* from 1905, 'imperialism' was originally a French word. It was from the 1830s onward that the terms *impérialiste* and *impérialisme* came into use in France. They referred to the empire of Napoleon and to the imperial pretentions of his nephew Louis Napoleon, later known as Napoleon III. The colonial connotation only came after the word had begun to be used in the United Kingdom in the 1860s. Then, of course, the empire it referred to was no longer the continental one of France but the overseas empire of Great Britain (Koebner and Schmidt, 1964).

Although the word imperialism was already used in the United Kingdom in the 1860s, the historical concept only appeared in 1902 with the publication of J.A. Hobson's *Imperialism: A Study* (Hobson, 1938). Hobson, a radical but not a socialist, was deeply influenced by the South African War (1899–1902). In 1900, he published a book on this subject, *The War in South Africa. Its Causes and Effects*, in which he argued that power in South Africa had fallen into the hands of a small group of financiers "chiefly German in origin and Jewish in race" (Hobson, 1900: p. 189). In his famous book *Imperialism: A Study*, he elaborated this vision into a general theory of imperialism, and used the term imperialism to indicate the "expansion of Great Britain and of the chief continental Powers" (Hobson, 1938: p. 27). The word expansion referred to the fact that over the previous 30 years, a number of European nations, Great Britain first and foremost, had "annexed or otherwise asserted political sway over vast portions of Africa and Asia, and over numerous islands in the Pacific and elsewhere" (Hobson, 1938: p. 15). For Hobson, the meaning of the word imperialism was very clear: it was the establishment of political control. He was also explicit about the forces behind it. Various people such as an "ambitious statesman, a frontier soldier and an overzealous missionary" might play some role in it, "but the final determination rests with the financial power" (Hobson, 1938: p. 59). Thus Hobson offered us a definition (imperialism is the expansion of political power of European countries over the non-European world), a periodization (imperialism took place over the previous 30 years, thus between 1870 and 1900), and an explanation (it was the result of the workings of the financial powers). In order to explain their behavior, Hobson argued that, as a consequence of the capitalist system, the British economy suffered from underconsumption. As a result of this, surplus capital could no longer be invested profitably in England itself. Therefore, the capitalists were "seeking foreign markets and foreign investments to take off the goods and capital they cannot sell or use at home" (Hobson, 1938: p. 85). "It is not too much to say that the modern foreign policy of Great Britain has been primarily a struggle for profitable markets of investment", adumbrated Hobson, who concluded that "It is this economic condition of affairs that forms the taproot of imperialism" (Hobson, 1938: pp. 53, 81).

The originality of Hobson's explanation for imperialism has been questioned by Gregory Claeys who contends that the economic side of his ideas was "if more sophisticated, methodical and carefully crafted, none the less the stock-in-trade of both Positivist and much socialist argument, built up over the previous half-century" (Claeys, 2010: p. 246). However, as Hobson's theory implied a criticism of capitalism, it had a palpable attraction for Marxist thinkers. As a result of this, a new Marxist theory of imperialism was born. While originally Marx and Engels had considered colonialism as an 'objective' progressive force (Avineri, 1968), now Marxist theorists such as Karl Hilferding and Rosa Luxemburg scorned late-nineteenth-century imperialism as a form of exploitation and suppression. The Marxist theory of imperialism became very influential when it was appropriated by a man who was not only a theorist but also a practical politician, Lenin. In 1916, he published his famous brochure *Imperialism. The Highest Stage of Capitalism* (Lenin, 1916). While Hobson perceived imperialism as a malfunctioning of capitalism, Lenin saw it as a sign of its ultimate and inevitable decay.

Lenin's ideas were mostly based on the work of the already-mentioned Marxist authors, who in turn had been inspired by Hobson's theory. It was therefore understandable that a direct link was seen between Hobson's and Lenin's theories, so much so that it became fashionable to speak of the 'Hobson–Lenin thesis'. There are, however, two important differences between Hobson and Lenin. First, for Hobson, the flight of capital from the metropolis to the overseas world was a consequence of the development of capitalism, but not a necessary consequence. The origin of the problem was underconsumption. Therefore, theoretically, it should also be possible to solve the problem by increasing the purchasing power of the working classes. Indeed, Hobson remarked: "If the consuming public in this country (Great Britain) raised its standard of consumption to keep pace with every rise of productive powers, there could be no excess of goods or capital clamorous to use Imperialism in order to find markets" (Hobson, 1938: p. 81).

Second, and more importantly, Hobson and Lenin tried to explain two different things. Hobson, who wrote his book during the South African War, wanted to explain the division of the world, and more specifically of Africa, in the late nineteenth century. Lenin, who wrote in 1916, tried to explain the redivision of the world of which the First World War was the most spectacular outcome. The word Africa hardly appears at all in Lenin's brochure. The period he referred to was also different from the one dealt with by Hobson: not 1870–1900 but thereafter. He wrote explicitly about this: "I have tried to show

in my pamphlet that it (imperialism) was born in 1898–1900, not earlier" (see Stokes, 1969: p. 289). Thus Lenin parted ways with Kautsky and Luxemburg, for whom imperialism was little more than another word for colonialism (Stokes, 1969: p. 297). For Lenin it was something else: not the highest stage of colonialism but of capitalism.

Although the capitalist theory of imperialism was not generally accepted, and alternative interpretations were launched and had some influence, some form of economic interpretation became the standard explanation of imperialism during the 1920s and the 1930s. Imperialism was considered as having originated from economic problems in Europe that were characteristic of the late nineteenth century, in particular the need to guarantee the flow of raw materials to the industrialized countries, and the protection of overseas markets for the sale of their industrial products. This consensus broke down after the Second World War under the influence of decolonization and the rise of the American Empire. The new world political situation also had an impact on the theory of imperialism. In a celebrated article, 'The Imperialism of Free Trade', two Cambridge historians, Jack Gallagher and Ronald Robinson, developed the concept of 'informal empire' (Gallagher and Robinson, 1953). They argued that the real zenith of the British Empire was not to be found in the late nineteenth century but rather in the mid-Victorian period of informal British economic hegemony. For Britain, the entire nineteenth century was one of expansion. It was an imperial century. Britain's imperial expansion manifested itself in various forms: emigration, trade, overseas investments, the establishment of naval bases, etc. The extension of political authority over foreign people was only one form of imperialism, and not even the most important one. The mid-Victorian Empire was comparable to the informal American Empire that came into being after 1945. It worked with informal means because that was the best way of doing things. The maxim of British policy makers was: informal empire if possible, formal only if necessary. Due to foreign competition and rivalry, however, the late Victorians were forced to formalize their Empire, and they did so, willy-nilly.

While Gallagher and Robinson discovered imperialism before empire, other theorists also discovered imperialism after empire. This resulted not so much from a reflection on the rise of the American Empire, but from a reassessment of decolonization. While after the First World War the European powers had increased their territorial possessions – for example, by the division of parts of the Ottoman Empire – and stabilized their colonial rule, the situation was very different after the Second World War. In Asia, the process of decolonization started immediately after the war and was followed later in Africa. Thus, by the 1960s, most of the former colonies had become independent politically. But political independence did not automatically bring an end to the social problems, nor to the economic dependency of the excolonies. Some of the new states became even more dependent on the Western-dominated world system than they had been before. For many observers, it was clear that the end of empire was not at the same time also the end of imperialism. Some theorists worked this out in the theory of dependency. According to the *dependencianistas*, imperialism not only was the extension of political control but also included the dependency of less developed parts of the world on the industrial powers. Empire was only one form of imperialism, one stage in the history of Western dominance.

However, why one form of imperialism was replaced by another remained a question, an answer to which was also given by Gallagher and Robinson in their famous book on the partition of Africa: *Africa and the Victorians* (Gallagher and Robinson, 1961). Here, they argued that changes in the periphery, that is, in the overseas world rather than in the mother countries, were responsible for the changes in the ways and means of imperialist control. Although *Africa and the Victorians* dealt primarily with British policy, the theories developed here had a wider meaning. While the theory of the imperialism of free trade was typically a theory about British imperialism, the peripheral theory was applicable to the imperialist activities of other nations as well. In many cases, changes in the non-Western world were decisive in determining imperialist action. Egypt's financial problems, for example, led to increasing foreign interference, and this, in turn, led to a 'nationalistic', or rather proto-nationalistic, reaction that plunged Egypt into an internal political crisis that led again to foreign intervention and occupation. The discovery of minerals in South Africa, to give another example, led to a complete change in the balance of power in that part of the continent.

Ronald Robinson later elaborated this interpretation into a more general theory based on the observation of the important role of the African and Asian partners of the imperialist rulers. In this so-called collaborationist theory, imperialism is conceived of as a system of collaboration between European and non-European forces before, during, and after colonial rule. The changing forms of imperialism are considered as changing forms of collaboration that resulted from changes in the bargaining positions of the various parties (Robinson in Owen and Sutcliffe, 1972: pp. 117–142).

The Gallagher and Robinson theories were followed by a greater number of studies on the economic significance of the British Empire and the role of economic factors in British imperialism. Foremost among these were a number of seminal contributions by P.J. Cain and A.J. Hopkins. "Explanations of imperialism", they contended, "ought to begin with a close study of economic structure and change in Britain" (Cain and Hopkins, 1987: p. 17). In particular, they argued that nonindustrial forms of capitalism, especially those associated with financial and commercial services based in the City of London, had been underestimated. They proceeded to assert that it was "in the second half of the nineteenth century, and especially after 1870, that Britain's expanding financial power created a world-wide 'invisible empire' which compensated for her dwindling economic influence in the United States and Europe" (Cain and Hopkins, 1987: p. 11). In Cain and Hopkins' interpretation, the essential link between Britain's expanding financial sector and imperialism was provided by 'gentlemanly capitalists', landowners, and financiers from southern England who boasted the same social background and inculcated similar notions of national interest as the aristocratic governing class. In his quasi biography of Hobson, Cain has maintained that "Where Hobson went wrong was in his forthright view of the relationship between finance and politics: finance was often crucial to expansion, financial conspiracy was not. Even in Egypt, where the financial motives

for occupation were obvious, there is no evidence to support Hobson's continuously reiterated assertion that ruthless financiers manipulated much less visionary and hesitant politicians to serve their clearly understood ends" (Cain, 2002: pp. 277–278). Although Cain and Hopkins have drawn criticism, not least their tendency to focus on gentlemanly capitalists to the exclusion of other important influences on British imperialism, their work has also found admirers (Webster, 1998).

While examining British imperialism beyond a simple concentration on gentlemanly capitalist interests, John Darwin places economics at the heart of the explanation for British imperialism: "It was coal, cotton and capital, not derring-do or district officers, on which Britain's world empire was built" (Darwin, 2012: p. 394). He also seeks to downplay the role of the 'official mind' in the growth of the British Empire. "British expansion", he insists, "was driven not by official designs but by the chaotic pluralism of British interests at home and of their agents and allies abroad" (Darwin, 2009: p. 3). More specifically, Darwin observes that "Since the prime motive was profit not the glory of Church or king, this was an empire of commercial experiments not an empire of rule by design" (Darwin, 2012: p. 389). As such, he points out that the "hallmark of British imperialism was its extraordinary versatility in method, outlook and object" (Darwin, 2012: p. 388). The impact of imperialism on Britain itself is a contested issue that has drawn contrasting interpretations regarding the depth of its imprint on the metropole (Porter, 2004; Thompson, 2005).

The important place of Britain in the debate on imperialism is understandable, because Britain was the imperial power par excellence. Some historians have suggested that British imperialism 'made' the modern world or even acted as a progenitor of modern globalization (Ferguson, 2004; Darwin, 2008). Nevertheless, Britain was by no means the most typical imperial power. Rather, it was atypical and therefore the discussions in other European countries on imperialism have followed different lines and focused on different questions. Chronologically speaking, however, the European revisionist theories were developed in the same years as the British: the debate started in the 1960s and continued well into the 1980s and beyond.

National Articulations

In France, Henri Brunschwig's *Mythes et Réalités de l'Impérialisme Colonial Français*, 1871–1914, which appeared in 1960, set the tone for the debate on French imperialism (Brunschwig, 1960). According to Brunschwig, the causes of French imperialism were not to be found in economic demands but in the development of French nationalism after the defeat in the Franco-Prussian war of 1870. The protectionist factor was a myth; political factors were decisive. Given the specific intellectual climate that existed in France after the Second World War, and in which Marxism played such an important role, it was to be expected that Brunschwig's book would lead to great controversy, as it did. But the Marxists could hardly deny the fact that the French colonial empire had been of little economic importance in France. In order to rescue the Marxist theory of imperialism, they therefore argued that French imperialism was not to be found in the French colonies but elsewhere, in the Russian and Ottoman Empires. They argued that French colonialism was not imperialist, and French imperialism was not colonial (Bouvier and Girault, 1976). In 1984, an important study by Jacques Marseille, based on an extensive data bank on French colonial trade, threw new light on the question of economic interest. His conclusion was that in the beginning the colonies were useful to French industry from an economic point of view, but subsequently they became a burden (Marseille, 1984).

There was also a strong connection between imperialism and nationalism, but it is not altogether clear to what extent imperialism was a result of nationalism. This is because the decision to found a German colonial empire was very much the decision of one man, Chancellor Bismarck. Therefore, in Germany, the discussion on imperialism has always been concentrated on Bismarck and his motives. There were two main interpretations: a foreign political one (imperialism as a move in Germany's international relations) and one in terms of domestic policy, such as electoral success, financial pressure groups, etc. The discussion was reopened when H.U. Wehler (1969) added new elements to this debate. Although he stressed the economic background of imperialism, he agreed that, as in the case of France, the German colonial empire had not been very profitable. In his view, the link between economics and empire must be sought on a different level. He emphasized the social problems of the *Reich* (its lack of legitimation because of its creation *von oben*, by force) and considered Bismarck's bid for colonies as a shrewd political move intended both as part of a general, more or less anti-cyclical, economic policy and of a social policy seeking to unite the Germans around issues of foreign policy, and thus to overcome internal tensions. Thus Wehler's emphasis was more on the domestic than on the diplomatic motives of German imperialism under Bismarck (Wehler, 1969). Here, the debate on German imperialism touched upon a wider discussion, the one on the problem of continuity and discontinuity in German foreign policy, the so-called German *Sonderweg*, and the place of national socialism in German history.

Italian imperialism was also studied from a special perspective. It was not very successful during the classical period of imperialism but it continued during the interwar years, under the influence of fascism. The French historian Jean-Louis Miège has emphasized not only the demographic factor in Italian imperialism but also its political and ideological dimensions – the nationalistic reaction to the loss of population as a consequence of emigration – comparing it in this respect to Spanish imperialism (Miège, 1968). The interpretation of Portuguese imperialism was long dominated by Hammond's theory of an 'uneconomic', that is to say a primarily nationalistic, form of imperialism. Gervase Clarence-Smith later challenged this view by making a strong case for an economic interpretation of Portuguese imperialism. He argued that economic motives went hand in hand with other ones such as missionary zeal and nationalism (Clarence-Smith, 1985).

The case of Belgium is very special, because in the nineteenth century, Belgium was an anti-colonialist country, but in spite of this, it was eventually to acquire one of the biggest

European colonies in Africa, the Belgian Congo (later called Zaire, now Congo again). That this happened was due to the extraordinary zeal, tenacity, ruse, and ruthlessness of one man, King Leopold II. Jean Stengers has analyzed the singular nature of the king's imperialism, which was one of old-fashioned economic exploitation and in this respect inspired by the example of the Netherlands (Stengers in Owen and Sutcliffe, 1972: pp. 248–276).

In the Netherlands, the historical discussion on imperialism began rather late. The most important contribution to the debate came from a book by Maarten Kuitenbrouwer (Kuitenbrouwer, 1991). In this he argued that the Dutch case was roughly analogous to others, and that the Netherlands followed more or less the general pattern. It has also been noted, however, that Dutch imperialism was defensive rather than offensive, reluctant rather than enthusiastic. In this respect it was comparable to that of Britain. In both cases, there was more continuity than discontinuity, and what discontinuity there was, derived from a change in circumstances, not in policy (Wesseling, 1997).

The historical debate on imperialism was mainly about the traditional colonial powers of Western Europe but observations have also been made about other countries. Russian imperialism poses interesting questions and offers paradoxical aspects because, on the one hand, Russia was an object of Western European financial imperialism, but on the other, it was itself also acting as an expansionist power by extending its empire to the East and eventually to the shores of the Pacific (Geyer, 1977; Le Donne, 1997). The case of Japan is particularly noteworthy because it is the only Asian nation that became an imperial power. In keeping with other Asian countries, it was first confronted with Western influence but it reacted in a very different way to this challenge. After having been forced to 'open' the country in 1853, it accepted Western notions and techniques very rapidly, so much so that already by the 1890s it had started its expansion into China. Japanese imperialism was continued in the 1930s and of course during the Second World War. Some analysts have also considered Japan's economic expansion after 1945 as a form of informal imperialism (Mommsen and Osterhammel, 1986: pp. 53–82).

The concept of American imperialism is a very complicated one, not only because of its origins as a former colony that engaged in a violent struggle to disengage from the British Empire but also because of the fact that it has prided itself on its anti-colonial rhetoric. The foremost historian of American imperialism is arguably Niall Ferguson. Ferguson in many ways sees America as the postwar successor to British imperialism, arguing that "in terms of economic resources as well as of military capability the United States not only resembles but in some respects exceeds the last great Anglophone empire" (Ferguson, 2005: p. 19). For Ferguson, the American Empire resembles its mid-nineteenth-century British predecessor since it "preferred indirect rule to direct rule and informal empire to formal empire" (Ferguson, 2005: p. 12). Also seeking to link British and American imperialisms, Bernard Porter observes that "The continuity does not lie in the imperialisms *per se* – America simply inheriting Britain's imperialism from her … – but in the underlying process that carried both of them. That was the 'progress' of the world towards globalization, free markets, consumerism, liberalism and the all the rest: the 'end of history', *perhaps*. This is the fundamental historical trend. Modern imperialisms are just corks, bobbing along the surface of them" (Porter, 2012: p. 316). For Ferguson, nevertheless, the American empire is a curious form of imperialism; as he puts it: "The United States has acquired an empire, but Americans themselves lack the imperial cast of mind. They would rather consume than conquer" (Ferguson, 2005: p. 29). As such, Ferguson depicts American imperialism as "an empire in denial" (Ferguson, 2005: p. 294).

For John Darwin, the preeminent historian of global empires, it is the case that "Imperial power has usually been the rule of the road" (Darwin, 2008: p. 23). As such it is axiomatic that the United States has become an imperial power. "After 1990", he asserts, "it became the only world empire. A state with the means to intervene forcibly in almost any part of the world, with such a massive advantage in military power over any possible rival, and with an advanced economy more than twice the size of its nearest competitor was such an empire de facto" (Darwin, 2008: p. 482). That the United States has lacked the colonial possessions of its Victorian predecessor, insists Darwin, is of "trivial importance" (Darwin, 2008: p. 483).

Explanation: Motives and Means

The rich literature on imperialism that has been published from the 1960s onward has led to a revision of the traditional views on the origins and meaning of late-nineteenth-century imperialism. Transformations not only in Europe but also in the overseas world have received attention as factors that can explain the new imperialist attitude. The main distinction is between European interpretations on the one hand, which underline economic, political, strategic, and ideological motives, and peripheral interpretations, which give special attention to activities and developments in the overseas world and in particular to the 'frontiers' of European influence. The new research has also given attention to such topics as the ecological aspects of imperialism (Crosby, 1986), cultural imperialism (Said, 1978, 1993), the impact of imperialism on the sciences (Petitjean et al., 1992), and so on.

Much of the debate on imperialism concerned the motives of the imperialists. In order to understand the origins of imperialism, however, attention has also to be given to another aspect, not the motives but the means. The development of imperialism cannot be understood by looking only at transformations in Europe and the overseas world, and the incentives for imperialist actions that were created by these. What also was essential for such action was the possession of the necessary means (Headrick, 1981, 1988). It had always been virtually impossible for Europeans to survive in the environmental conditions of tropical Africa. New developments in the medical sciences, such as the prophylactic use of quinine (as from the 1840s), made it possible for Europeans not only to live but also to work and even to fight under such conditions. The development of new means of transportation (steamships, railways), the opening of new sea routes (like the one via the Suez Canal), and the revolution in the means of communication (the telegraph, and later on the telephone and wireless communication) made the extension of imperial rule possible.

Finally, but perhaps most importantly, the development of new weapons, and in particular of the machine gun, gave the Europeans an enormous advantage in their battles with non-European nations. Colonial wars became successful almost by definition and the European colonial armies became 'ever-victorious' armies. Entire continents could be conquered at very small cost to the conquerors.

Thus the great technological superiority of the Europeans came into existence during the later part of the nineteenth century due to the so-called Second Industrial Revolution that took place in Western Europe and created rivals for British trade. These technological transformations not only offered the means for imperial expansion but also led to new demands in the European societies, which had their effects on foreign and colonial policy. Social and economic questions assumed increasing importance. State welfare provisions expanded. The import of tropical products at affordable prices was considered as a matter of public concern. This called for sustained economic exploitation, which presupposed the existence of peace and order; in other words, effective authority.

The balance of power in Europe also changed dramatically in the 1870s. In the early nineteenth century, from Napoleon to Bismarck, Europe had found itself in an exceptional political situation. Germany and Italy did not yet exist. Britain had eliminated France as a maritime and colonial rival. Spain, Portugal, and the Netherlands, the old colonial powers, had had their day. Hence, Britain enjoyed *de facto* world supremacy, much as the United States was to do after the Second World War. All these factors began to be compromised from the 1870s. After its defeat in 1870, France sought compensation for its loss as a continental power by strengthening its overseas role. Germany and Italy, newcomers both, claimed a place under the sun.

Internal political factors also played a part. European governments were faced with a new phenomenon: they had to take the wishes of their electorate into account. Economic growth and social harmony became declared objectives of government policy. The Paris Commune of 1871 accentuated the danger of social revolution and hence the importance of social issues. The state was asked to do more things for more people. Conversely, technical progress, economic growth, and growing political involvement by the citizens of Europe created the conditions for a strong state. The military might of the European powers reached unprecedented heights. At the same time in the overseas world, transformations took place, which often proceeded from previous contacts with Europe or European settlers, and which changed the existing internal balance of power. Thus in many respects, a new situation came into being after 1870, in Europe as well as in the overseas world, and therefore, after all forms of revisionism, it is still justified to speak of the period of 1870–1914 as an age of imperialism, as Hobson did when he introduced the concept over a century ago. That imperialism is still with us, either in the form of contemporary globalization as writers such as Michael Hardt and Antonio Negri suggest (Hardt and Negri, 2000), or in the rise of an informal American Empire in the postwar world as Darwin, Ferguson, and others contend, demonstrates the vitality of the concept and its ongoing relevance to the contemporary world.

See also: African Studies: History; Ancient Greece and Rome, History of; Capitalism; Colonialism, Anthropology of; Colonization and Colonialism, History of; Spirit Possession, Anthropology of; Structuralism.

Bibliography

Avineri, S. (Ed.), 1968. Karl Marx on Colonialism and Modernization. Doubleday, Garden City, New York.
Bouvier, J., Girault, R., 1976. L'Impérialisme Français d'Avant 1914. Mouton, Paris.
Brunschwig, H., 1960. Mythes et Réalités de l'Impérialisme Colonial Français, 1871–1914.
Cain, P.J., 2002. Hobson and Imperialism. Radicalism, New Liberalism, and Finance 1887–1938. Oxford University Press, Oxford.
Cain, P.J., Hopkins, A.G., 1987. Gentlemanly capitalism and British expansion overseas. II: new imperialism, 1850–1914. Economic History Review, Second Series 40, 1–26.
Cain, P.J., Hopkins, A.G., 1993a. British Imperialism. Innovation and Expansion, 1688–1914. Longman, London.
Cain, P.J., Hopkins, A.G., 1993b. British Imperialism. Crisis and Deconstruction, 1914–1990. Longman, London.
Claeys, G., 2010. Imperial Sceptics: British Critics of Empire, 1850–1920. Cambridge University Press, Cambridge.
Clarence-Smith, G., 1985. The Third Portuguese Empire, 1825–1975. A Study in Economic Imperialism. Manchester University Press, Manchester.
Colin, P., 1966. The Myth of French Imperialism, 1870–1914. Pall Mall Press, London.
Crosby, A.W., 1986. Ecological Imperialism. The Biological Expansion of Europe, 900–1900. Cambridge University Press, Cambridge.
Darwin, J., 1997. Imperialism and the victorians: the dynamics of territorial expansion. English Historical Review 112, 614–642.
Darwin, J., 2008. After Tamerlane. The Rise & Fall of Global Empires 1400–2000. Penguin, London.
Darwin, J., 2009. The Empire Project: The Rise and Fall of the British World-System, 1830–1970. Cambridge University Press, Cambridge.
Darwin, J., 2012. Unfinished Empire: The Global Expansion of Britain. Allen Lane, London.
Davis, L.E., Huttenback, R.A., 1986. Mammon and the Pursuit of Empire. The Political Economy of British Imperialism, 1860–1912. Cambridge University Press, Cambridge.
Dumett, R. (Ed.), 1999. Gentlemanly Capitalism and British Imperialism: The New Debate on Empire. Longman, London.
Ferguson, N., 2004. Empire. How Britain Made the World. Penguin, London.
Ferguson, N., 2005. Colossus: The Rise and Fall of the American Empire. Penguin, London.
Gallagher, J., Robinson, R.E., 1953. The imperialism of free trade. Economic History Review, Second Series 6, 1–15.
Gallagher, J., Robinson, R.E., 1961. Africa and the Victorians. The Official Mind of Imperialism. Macmillan, London.
Geyer, D., 1977. Der russische Imperialismus. Studien über den Zusammenhang von innerer und auswärtiger Politik, 1860–1914. Vandenhoeck & Ruprecht, Göttingen, Germany [1987. Russian Imperialism: The Interaction of Domestic and Foreign Policy, 1860–1914. Berg, Leamington Spa].
Hardt, M., Negri, A., 2000. Empire. Harvard University Press, Cambridge, MA and London.
Headrick, D.R., 1981. The Tools of Empire. Technology and European Imperialism in the Nineteenth Century. Oxford University Press, New York.
Headrick, D.R., 1988. The Tentacles of Progress. Technology Transfer in the Age of Imperialism, 1850–1914. Oxford University Press, New York.
Hobson, J.A., 1900. The War in South Africa. Its Causes and Effects. Nisbet, London.
Hobson, J.A., 1938. Imperialism. A Study, third rev ed. Allen & Unwin, London.
Koebner, R., Schmidt, H.D., 1964. Imperialism. The Story and Significance of a Political Word, 1840–1960. Cambridge University Press, Cambridge.
Kuitenbrouwer, M., 1991. The Netherlands and the Rise of Modern Imperialism. Colonies and Foreign Policy, 1870–1902. Berg, Oxford.
Le Donne, J.P., 1997. Russian Empire and the World, 1700–1917. Oxford University Press, New York.
Lenin, V.I., 1916. Imperialism. The Highest Stage of Capitalism. Progress, Moscow.
Marseille, J., 1984. Empire Colonial et Capitalisme Français, Histoire d'un Divorce. Michel, Paris.

Marshall, P.J., 1982. European imperialism on the 19th century. History Today 3, 49–51.
Miège, J.L., 1968. L'Impérialisme Colonial Italien de 1870 à nos Jours. Société d'édition d'enseignement supérieur, Paris.
Mommsen, W.J., Osterhammel, J. (Eds.), 1986. Imperialism and After. Continuities and Discontinuities. Allen & Unwin, London.
Owen, R., Sutcliffe, B. (Eds.), 1972. Studies in the Theory of Imperialism. Longman, London.
Petitjean, P., Jani, C., Moulin, A.M. (Eds.), 1992. Science and Empires. Historical Studies about Scientific Developments and European Expansion. Kluwer, Dordrecht, The Netherlands.
Porter, A., 1994. European Imperialism, 1860–1914. Macmillan, Basingstoke.
Porter, A., 1999. The Oxford history of the British Empire. In: The Nineteenth Century, vol. 3. Oxford University Press, Oxford.
Porter, A., 2004. Religion versus Empire? British Protestant Missionaries and Overseas Expansion, 1700–1914. Manchester University Press, Manchester.
Porter, B., 2004. The Absent-Minded Imperialists: Empire, Society, and Culture in Britain. Oxford University Press, Oxford.
Porter, B., 2012. The Lion's Share: A History of British Imperialism 1850 to the Present, fifth ed. Pearson, Harlow.
Said, E.W., 1978. Orientalism. Routledge & Kegan Paul, London.
Said, E.W., 1993. Culture and Imperialism. Chatto & Windus, London.
Seton-Watson, H., 1961. The New Imperialism. Bodley Head, London.
Smith, S.C., 1998. British Imperialism 1750–1970. Cambridge University Press, Cambridge.
Stokes, E., 1969. Late nineteenth-century colonial expansion and the attack on the theory of economic imperialism: a case of mistaken identity. Historical Journal 12, 285–301.
Thompson, A., 2005. The Empire Strikes Back? The Impact of Imperialism on Britain from the Mid-nineteenth Century. Pearson, Harlow.
Webster, A., 1998. Gentlemen Capitalists. British Imperialism in South East Asia, 1770–1890. Tauris Academic Studies, London.
Webster, A., 2006. The Debate on the Rise of the British Empire. Manchester University Press, Manchester.
Wehler, H.U., 1969. Bismarck und der Imperialismus. Kiepenheuer & Witsch, Cologne, Germany.
Wesseling, H.L., 1997. Imperialism and Colonialism. Essays on the History of European Expansion. Greenwood Press, Westport, CT.

Implementation: Political

Eugene Bardach, Goldman School of Public Policy, University of California at Berkeley, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Once adopted, a policy must survive the rigors of the implementation process. These often lead to delay, the perversion or diminution of goals, and costs higher than expected. The causes generally lie in the many, and sometimes complex, elements that have to be brought together in a workable fashion, and in the many actors, with diverse priorities and objectives, who control those elements. Countervailing strategies involving leadership, momentum, and strategic retreat may somewhat offset these difficulties.

Implementation studies are inevitably normative. In 1973, the field was created virtually overnight by Jeffrey Pressman and Aaron Wildavsky, in whose book *Implementation* the guiding research questions were: "How well was this authoritative mandate (law, regulation, program, official policy pronouncement) implemented?" and "How might it have been better implemented?" Later researchers redefined the normative standard as achieving the values implicit in some mandate rather than in executing its prescriptive details. Then the guiding question became: "How can a complex implementation process bring out the values implicit in, or improve upon the raw materials provided by, some authoritative mandate?" Such normative questions have provided an intellectual focus for a field unified by little else.

'Successful Implementation' Defined

For simplicity, I restrict discussion to legislative mandates. The raw materials are typically (1) a legislative mandate, (2) some new budgetary resources, and (3) a designated responsible agency. In most circumstances these three assets are not enough and other resources are needed. Mandates defining regulatory programs will often require compliance resources from businesses or ordinary citizens. Service-delivery programs will often need participation from citizens who perceive the availability of the program, are willing to make the effort to establish eligibility, and can manage to sustain continued participation. The responsible agency may also wish to involve relevant nonprofit agencies. Other program agencies with overlapping jurisdictions or responsibilities must sometimes be asked to cooperate. Overhead agencies must supply permission to recruit personnel, lease office space, enter into contracts, and so on. 'Implementation' is therefore a process of constructing an operational program out of both newly mandated and extant resources. 'Successful' implementation means that the result is faithful to at least a minimalist interpretation of 'the mandate,' is at least minimally cost-effective, and is produced without unacceptable delay. (For a time scholars were divided into 'top-down' and 'bottom-up' camps, depending on where in the loose hierarchy of actors in the implementation process they preferred to look for the more important causes of failure or success and for the normative interpretation of 'the mandate' being implemented. This division is no longer meaningful, however (O'Toole Jr., 2004).)

Explaining Success and Failure

The early implementation literature, inspired by the troubles of Great Society programs, focused on documenting and explaining implementation failure. The (partial) subtitle of *Implementation* was "How Great Expectations in Washington Are Dashed in Oakland; Or, Why It's Amazing that Federal Programs Work at All…" The principal explanatory framework in the early literature highlighted the institutions and practices of American pluralist and federalist democracy that were unfriendly to implementation, e.g., the large number of actors that had to be brought into the process, the susceptibility of programs to continuing political conflict over their goals and resources, weak or nonexistent interagency and interjurisdictional coordinating systems, and cumbersome accountability procedures and heavy-handed overhead agencies. In this world of fragmented political power, the many resources required for program construction are controlled by different actors – from front-line regulatory enforcers and service professionals to personnel officials writing civil service job descriptions – with only partially overlapping interests and with highly variable abilities and motivations to induce cooperation among one another (Nakamura and Smallwood, 1980). Furthermore, these actors normally – and often quite properly – have different priorities than those of the advocates for the new mandate. They have ongoing responsibilities, after all, some of which are poorly fulfilled as it is, given normal governmental resource constraints.

Consider these institutional problems in system terms. Constructing a complex operational system out of many, varied, and interdependent elements is vulnerable to failure from many quarters. Bureaucratic resistance, prolonged bargaining, political sabotage, weak technical capacity, and a host of other conditions could affect many of these elements (Bardach, 1977). Failure seems overdetermined. Yet failure appears to be much less frequent than this model suggests (O'Toole Jr., 2012). The model omits the possibility of adaptation, i.e., strategically sacrificing controversial or overly ambitious goals and creatively improvising with respect to means. It also omits momentum, the fact that success with

respect to securing one system element helps to beget success with respect to others.

The capacity to plan and execute effective adaptation and momentum-building strategies often takes the form of committed political leaders, entrepreneurial program managers, clever legislative staff and their principals, and the like (Mazmanian and Sabatier, 1983; O'Toole Jr., 2004). This, at least, is the impression conveyed by many of the case-study reports of successful implementation. Although lacking a methodological and theoretical rationale, such case studies may, nevertheless, have illuminated an important phenomenon. The current theoretical challenge of greatest significance is to explicate the process of adaptation and to identify factors that increase the likelihood of success or failure. The corresponding empirical challenge is to identify the conditions under which skilled adaptive leadership is drawn into certain implementation processes but not into others.

Design Failures and Implementation Failures

The early literature usefully distinguished failures of implementation from three common design-stage mistakes that simply show up *during* the course of implementation and which might have been aggravated (or occasionally ameliorated) by the implementation process. First, design-stage expectations are usually too high. To survive the legislative process, a policy or program mandate will often have had its merits oversold and its limitations understated. Second, designers sometimes commit fundamental conceptual errors, e.g., supposing that price controls in competitive markets can be implemented without causing shortages. Third, they may choose an inappropriate policy instrument. Arguably, the programs studied by Pressman and Wildavsky, concerning the subsidized creation of low-skill jobs in a poor city, represented failures of instrument selection. The chosen instrument was a subsidy to firms promising to hire from a general target population. With no easy way to enforce the promises, however, the implementing agency probed the prospective employer firm for months on end, trying to clarify what might count as a legitimate excuse for underperformance and trying to test the firm's good faith and resolve. A much better choice of instruments would have been an ex-post-facto wage subsidy for workers actually hired.

The Characteristics of Generic Governmental Instruments

In the literature 'implementation problems' also include those that beset most operational governmental programs. Although the early implementation literature focused on 'control' problems, the focus is now broader, encompassing the standard problems of public administration and public management, e.g., how to recruit good personnel and motivate quality performance, how to keep a focus on results rather than on personal or careerist objectives or simply on following procedures and staying out of trouble, how to protect the program from ideological opponents in the legislature or within the cadre of higher-level political appointees, how to build trust and communications channels within and across agencies, and how to create a culture of continuous improvement. The 'reinventing government' movement has been a modest source of ideas and inspiration about how to cope with general management problems.

But the distinctive contribution of implementation research regarding such problems has been to analyze the forms they take when expressed in generic governmental instruments (Salomon, 1989). Two of the most important of these are regulation (e.g., sanitary code enforcement and electric utility rate setting) and pure service delivery (e.g., employment and training) either directly or through provider subsidies.

Regulation

Command-and-control (or 'social') regulation is typically a response to information failures – regarding unsafe bus maintenance, say, or filth in restaurant kitchens – which prevent consumers or employees from protecting themselves through either market action or the threat of lawsuits. Command-and-control programs typically have four main components: a set of standards, a cadre of inspectors and auditors to detect noncompliance, a set of sanctions with which to threaten noncompliant firms or citizens, and the climate of deterrence (or prevention) that these three elements create in combination. This sort of regulation is vulnerable to errors of underregulation or overregulation or to both simultaneously. Political pressures and scientific uncertainty can lead to overly stringent or overly lax standards. Standards can also be too prescriptive or too ambiguous. The enforcement process is vulnerable to budgetary strictures, overconcentration on symbolic targets, go-by-the-book rigidity, and in some cases corruption. Deterrence can be weak and pale or else overweaning and overly broad. And overall, regulatory strategies do not necessarily take account of the capacity, or lack of it, of the targeted parties to actually do what is nominally required.

Economic regulation includes natural monopolies such as local telephone service and (former) oligopolies such as long-distance freight hauling. Most scholars observe that the regulators tend to get 'captured' by the more powerful firms within the regulated industry. Arguably, the regulators are as much captured by a sense of responsibility for the economic well-being of the industry they regulate, however, and allow 'well-being' to be interpreted through the self-interested lenses of the major firms.

Technological change and more realistic economic analysis have undermined the rationale for much traditional economic regulation, and many regulatory activities, along with a few agencies, have been abolished. A new addition, however, has been environmental regulation. Although it often uses command-and-control methods, it also counts as 'economic' regulation because implementers are concerned with how their decisions on standards and enforcement actions affect inter-firm, interindustry, and interregional competitive balances.

Service Delivery

The chief implementation challenges in service-delivery programs are developing technically effective services (as in effective teaching of third-graders); suppressing fraud, waste,

and abuse (as in $2 billion in fraudulent Medicare billings); and preventing inequitable distribution to constituencies favored by politics or by the preferences of service professionals (as in middle-class neighborhoods, tending to attract better teachers to their schools). A fourth challenge is keeping service quality at a high level while being unable to rely on traditional market-based incentives such as the threat of consumer exit.

When government uses subsidies to enlist private providers, such as hospitals treating Medicaid patients or nonprofit organizations to provide vocational training, it typically relies on regulations and audits to attempt to ensure that the providers are targeting their services correctly, maintaining a minimal acceptable quality, and not engaging in fraud or abuse (McDermott, 2004). Many of the problems of command-and-control regulation are thereby imported into service-delivery programs.

Service programs aimed at 'deviant' populations such as drug abusers, criminal offenders, the mentally ill, and welfare recipients typically have a regulatory, paternalistic component as well as a pure service component. Implementing personnel usually have considerable discretion over how to balance the two components (May and Winter, 2007; Lipsky, 1980). Large-scale ideological trends, and more immediate political pressures, also affect the balance.

The Practitioner's View of Implementation

Because implementation studies are normative, their results can be helpful in guiding practice. The literature advises policy designers to: (1) 'map' the functional requirements of a projected operational system and how they are each to be obtained (Elmore, 1979), (2) write 'scenarios' regarding possible implementation pitfalls en route to such a system (Bardach, 1977), and (3) estimate the probability of 'compliance from target groups' whose cooperation is needed for critical tasks (Mazmanian and Sabatier, 1983). O'Toole (2004) has recommended heuristics from game theory and from 'contextual interaction' theory, which focuses on the objectives, information, and power of the 'basic sets of actors.' Designers can also consider institutional solutions, e.g., contracting with an organization in the private sector to take on the responsibility of starting and running a new AIDS clinic or building and operating a new jail.

Implementers can also view the experience of operational implementation, whether successful or otherwise, as a way to improve long-run policy design and to create a mandate of sorts for what they are doing. Consider some program domains such as welfare-to-work, and construe all the 'implementers,' from the Secretary of Health and Human Services through the state and county senior managers down to the lowest-level case workers as a sort of network through which ideas and effort can flow and can become organized. These ideas concern both policy design and the choice of implementation instruments. From the mid-1980s through the mid-1990s this network produced a good deal of lore, and a bit of experimental evidence, that a more work-oriented welfare policy might be effective and would not be unduly harmful to the poor. Counter-claims were also produced. But in the end, the experience of many implementers combined with conservative Republican political leadership and very widespread support in public opinion created the federal welfare reform legislation of 1996. The preferred implementation instruments were a high degree of devolution to the states with regard to means but rather sharp financial incentives to compel adherence to federal government objectives. Deficiencies in the legislation will no doubt be brought to light in the course of implementation and further change will occur.

In some sense, then, 'the policy system is learning,' or 'evolving,' and the implementation process is a part of that system (Goggin et al., 1990; Majone and Wildavsky, 1979). To be sure, 'learning' and 'evolving' are implicitly favorable ways to characterize a policy change. Many observers would object to this characterization of welfare reform. But eschewing value-laden language altogether would denature a phenomenon like implementation, which is so intimately connected to democratic policy choice and to the challenges of creating public value for the citizenry.

See also: Voting and Direct Democracy.

Bibliography

Bardach, Eugene, 1977. The Implementation Game: What Happens after a Bill Becomes a Law. MIT Press, Cambridge, MA.

Elmore, Richard, 1979. Backward mapping: implementation research and policy decisions. Political Science Quarterly 94 (4), 601–616.

Goggin, Malcolm L., Bowman, Ann O'M., Lester, James P., O'Toole Jr., Laurence J., 1990. Implementation Theory and Practice: Toward a Third Generation. Scott, Foresman, Glenview, IL.

Lipsky, M., 1980. Street-Level Bureaucracy. Russell Sage Foundation, New York.

Majone, Giandomenico, Wildavsky, Aaron, 1979. Implementation as evolution. In: Pressman, Jeffery L., Wildavsky, Aaron (Eds.), Implementation. University of California, Berkeley.

Mazmanian, Daniel A., Sabatier, Paul A., 1983. Implementation and Public Policy. Scott, Foresman, Glenview, IL.

McDermott, K.A., 2004. Incentives, capacity, and implementation: evidence from Massachusetts education reform. Journal of Public Administration Research and Theory 16, 45–65.

May, P.J., Winter, S.C., 2007. Politicians, managers, and street-level bureaucrats: influences on policy implementation. Journal of Public Administration Research and Theory 19, 453–476.

Nakamura, Robert T., Smallwood, Frank, 1980. The Politics of Policy Implementation. St. Martins, New York.

O'Toole Jr., L.J., 2004. The theory-practice issue in policy implementation research. Public Administration 82 (2), 309–329.

O'Toole Jr., L.J., 2012. Interorganizational relations and public policy. In: Peters, B.G., Pierre, J. (Eds.), SAGE Handbook of Public Administration, second ed. Sage, London, pp. 292–304.

Salomon, Lester M., 1989. Beyond Privatization: The Tools of Government Action. Urban Institute, Washington, DC.

Implementation Science

Dean Fixsen, Karen Blase, Allison Metz, and Melissa Van Dyke, FPG Child Development Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Implementation science has developed rapidly and has established frameworks for summarizing a variety of influences on intended outcomes. The frameworks are useful for guiding the work of implementation in typical human service practice settings and for generating useful hypotheses for social sciences to explore. Evaluations of successful efforts to make full and effective use of social science innovations to benefit citizens and society have led to rapid expansion of implementation science.

Implementation Defined

> Implementation is defined as a specified set of activities designed to put into practice an activity or program of known dimensions. According to this definition, implementation processes are purposeful and are described in sufficient detail such that independent observers can detect the presence and strength of the 'specific set of activities' related to implementation. In addition, the activity or program being implemented is described in sufficient detail so that independent observers can detect its presence and strength. When thinking about implementation the observer must be aware of two sets of activities (intervention-level activity and implementation-level activity) and two sets of outcomes (intervention outcomes and implementation outcomes).
>
> Fixsen et al., 2005, p. 5

Implementation is universal and applies to any attempt to use innovations in practice. The implementation activities designed to put into practice an activity or program may be haphazard and unintentional (do the best you can), or they can be purposeful and specified well (do what is most effective). Using well-specified implementation supports has a substantial impact on intervention outcomes in practice. Active implementation frameworks help define effective implementation supports and are outlined in this article.

Innovation Defined

Implementation is in service to a well-defined innovation. An innovation is anything new to an individual, organization, or human service system (Rogers, 1995). The innovation may be a therapeutic intervention, instruction method, evaluation practice, management practice, clinical guideline, policy directive, improvement initiative, or other activity or program.

As noted in the definition of implementation quoted above, the extent to which an innovation is operationalized affects the ability to use the intervention in practice. Usable innovation criteria are outlined in this article.

Human Services Defined

The phrase human services is used to denote the full spectrum of services where one human being (e.g., a therapist; teacher; community organizer) is interacting with another human being (e.g., a student; neighborhood resident) in a way that is intended to be helpful. Human service domains include child welfare, community development, corrections, education, health, global health, mental health, public health, social services, substance abuse treatment, and others. Literature and examples drawn from human services are used in this article along with examples from business, manufacturing, and computer science.

Rationale for Implementation Science

The interest in implementation has grown with the interest in evidence-based innovations. The evidence-based intervention movement began with the documentation of intervention and policy failures in human services. Eysenck (1952) was the first of many who reviewed a series of outcome studies and declared various therapies to be ineffective. Rittel and Webber (1973) and Van Meter and Van Horn (1975) were among those who summarized public policies designed to improve human services and noted their general lack of intended impact. These seminal reviews led to great debates and to a new focus on developing evidence-based (effective) approaches to human service interventions of all kinds.

As the evidence-based program movement has developed over the ensuing decades, it has become clear that innovations are not self-implementing in human services. Evidence-based practices and programs are celebrated when they are published. Disappointment follows when attempts generally fail to make use of them outside of experimental trials. This has become known as the science to service gap and the quality chasm (Institute of Medicine, 2001). Implementation is now recognized as the link between science and service and is studied in its own right. Evaluations of successful efforts to make full and effective use of innovations to benefit citizens and society have led to rapid expansion of implementation science.

Diffusion, Dissemination, and Implementation

Implementation science is a blend of three strands of knowledge development. First, diffusion theory as described by Rogers (1995) originated in the 1950s in agriculture and

focused on communication patterns related to adoption of innovations in farming (e.g., use of hybrid seed; crop rotation). Diffusion theory has been generalized to a wide variety of fields and is the best known of the three strands. The outcome of diffusion of information is a decision to adopt an innovation.

The second strand is dissemination theory or, more accurately, dissemination theories. Brownson et al. (2012) provide a summary of dissemination findings and Tabak et al. (2012) list over 60 theories pertaining primarily to dissemination. Dissemination theories typically focus on helping practitioners, managers, policy makers, and others understand research findings so they may be more likely to use those findings in their work. The outcome of dissemination is understanding an innovation and attempting to make use of the innovation in practice.

The third and most recent strand is implementation theory. Implementation theory is based on evaluations of attempts to use innovations in practice. Applied research data are accumulating rapidly as common language, common measures, and guiding frameworks become available. Meyers et al. (2012) provide a comparative summary of 25 implementation frameworks. The outcome of implementation is full, effective, and sustained use of an innovation in practice.

The field is on the verge of a unifying theory of implementation that includes diffusion and dissemination along with active approaches to moving science into service in order to realize the promise of evidence-based programs and other innovations.

Using Innovations in Practice

The evidence-based program movement has permeated all social sciences and human services in the past three decades. There is great appeal to the idea that governments and service agencies funded by governments should make use of innovations demonstrated to be effective. Why invest in using what does not work when effective (evidence-based) alternatives are available for use?

Problems arise because attempts to use innovations (evidence-based or otherwise) in typical service or living environments bring into play a variety of personal, organization, and system factors. Change stimulates concern and concern is a powerful inducement to maintain the comfort of the status quo, even when the evidence persistently demonstrates the need for change. In the United States, literacy outcomes for students have remained at a consistently modest level since 1971, hovering around a score of 215 on a 500-point scale for 9-year-old children. Thus, the fundamental elements of education instruction in school, district, state, and federal systems have maintained an unacceptable status quo in spite of decades of tremendous changes in policy, massive increases in funding, countless legislative mandates, huge investments in education science and technology, and the rise of compliance-oriented data systems.

These problems are noticeable because the pervasive growth of the evidence-based program movement globally has raised expectations for achieving socially significant outcomes in human services. The science to service gap has exposed the challenges related to fidelity, sustainability, and scalability. Implementation science has developed in part in response to these issues. Schofield (2004), Vernez et al. (2006), and others note that achieving socially significant outcomes on a useful scale will require scaling implementation supports needed to make use of effective innovations.

A Science of Implementation

The science of implementation has lagged behind the development of best practices in implementation. Logically, we cannot do research on purported effective implementation practices unless those practices are being used. Thus, the purposeful and active use of implementation best practices amounts to building a better laboratory in which research on implementation can be done. If we continue to use implementation practices that already have been studied and demonstrated to be insufficient, the best we can hope for is a replication of findings that show little or no effect. If we begin to use practices that are predicted to be effective by the active implementation frameworks, then we can hope for different (and perhaps improved) outcomes.

The active implementation frameworks (Fixsen et al., 2005; Blase et al., 2009) can be considered a mid-level theory of implementation (ICEBeRG, 2006). As such, the frameworks can provide guidance for action planning and the development of testable hypotheses to guide research. The active implementation frameworks fit the 14 dimensions described by Meyers et al. (2012) in their review of extant implementation frameworks, and incorporate many dimensions of diffusion and dissemination theories reviewed by Brownson et al. (2012) and Tabak et al. (2012).

The active implementation frameworks are based on the best available evidence. A thorough review and synthesis of transdisciplinary implementation evaluation articles was conducted, and the research base informed each component of the frameworks (Fixsen et al., 2005). The review of the literature has continued since 2005 and has resulted in modifications and additions to the frameworks (Blase et al., 2012; Fixsen et al., 2013). The active implementation components described in 2005 benefitted from systematic knowledge capture (using concept mapping and nominal group process; each an evidence-based approach to collecting qualitative data from groups) by experienced purveyors and users of a range of evidence-based programs (Blase et al., 2005), and by representatives of national racial, ethnic, family, and cultural advocacy groups (Blase and Fixsen, 2003). The information from these meetings expanded and operationalized best practices associated with the active implementation frameworks.

These methods comport with best practices in other fields for arriving at guidelines for clinical practice. Thus, the best available evidence and expert guidance in implementation science provide the foundations for the active implementation frameworks.

A Formula for Success

It is widely acknowledged that implementation science is complex and includes many influences operating simultaneously.

What is known about implementation science can be summarized in a formula for success that accounts for multiple levels of influence over time:

Effective Innovations × Effective Implementation × Enabling Contexts = Socially Significant Outcomes

To explain the formula for success, the following frameworks for implementation will be outlined in this article: (1) usable interventions, (2) implementation stages, (3) implementation drivers, (4) improvement cycles, (5) implementation teams, and (6) enabling contexts.

Effective Innovations: Usable Interventions

If the goal of innovations (evidence-based or otherwise) is to produce change, potential users need to know what the innovation is. Summaries of the literature since 1991 consistently point to the lack of specification of the independent variable (the innovation) in randomized control trials and other evaluations. Naleppa and Cagle (2010) provide a summary of the literature and of the methodological challenges related to this issue. From an implementation perspective, it is not enough for innovations to be 'evidence-based,' they also must be usable. Fixsen et al. (2013) described the criteria for interventions usable in practice.

1. Clear description of the innovation; philosophy, values, and principles; inclusion and exclusion criteria that define the population for which the innovation is intended.
2. Clear description of essential functions that must be present to say that an innovation exists in a given location (essential functions sometimes are referred to as core intervention components, active ingredients, or practice elements).
3. Operational definitions of the essential functions; the core activities that allow an innovation to be teachable, learnable, and doable in practice.
4. A practical assessment of the performance of practitioners who are using the innovation; the performance (fidelity) assessment is highly correlated with intended outcomes.

Innovations that meet these criteria are more likely to be teachable, learnable, doable, and assessable in practice. The evidence of effectiveness is established in criterion #4 that calls for continual assessment of the strength of the innovation as it is used in practice and continual assessment of intended outcomes. This requirement recognizes the complexity of human interactions and the mutual influences people exert on one another. It is expected that the use of innovations will vary within practitioners across time and experience, and across practitioners. To produce socially significant outcomes reliably, innovations need to be used within an acceptable range of performance.

When implementation difficulties are encountered, the lack of specification of an innovation is the first place to look for a solution. Vague ideas, general philosophies or principles, and statements of goals make it difficult for practitioners and others to know what to do in practice to realize the intended benefits. Innovations that meet the usable intervention criteria provide a strong foundation for teaching, learning, doing, and assessing innovations in practice.

Innovations are more than names and claims. In a compliance-oriented environment, it is tempting to simply say an innovation is being done in order to meet a requirement and pass a compliance review. For example, when new standards were issued for quality improvement in health care, many hospitals claimed they were using Total Quality Management (TQM) to meet these new requirements for accreditation. TQM is the application of complex Six Sigma and Lean Manufacturing principles and methods to health care. Westphal et al. (1997) conducted a study of 2721 hospitals and found that few of those that claimed using TQM actually used TQM as intended. The authors concluded, "If organizations can minimize evaluation and inspection of their internal operations by external constituents through adoption alone, they may neglect implementation altogether, decoupling operational routines from formally adopted programs" (p. 371).

In outcome-oriented environments, it is necessary to actually use the innovation in practice and show evidence that intended outcomes are being realized (fidelity of delivery, improvements for intended beneficiaries). Names and claims do not produce socially significant benefits.

Effective Implementation

Innovations (usable interventions) describe *what* to do but the evidence is overwhelming that knowing what to do is necessary but not sufficient for achieving socially significant outcomes (Kessler and Glasgow, 2011). The question is *how* to make use of innovations as intended in typical service settings outside the laboratory.

The status quo supports the existing ways of work. Innovations are, by definition, new ways of work for practitioners, organizations, and systems. Thus, practitioners must learn innovations, and organizations and systems must change to accommodate and support the new ways of work embodied in those innovations. Eventually, an innovation becomes the new standard way of work. After that, any new innovations must undergo the same process of changing the new status quo if they are to be effective and sustained. The process of change is never done as human service systems seek to continually improve socially significant outcomes. Implementation frameworks provide guidance for *how* to incorporate innovations into existing systems effectively and efficiently to realize socially significant outcomes.

Effective Implementation: Stages

The full and effective use of an innovation occurs over time. Stages of implementation have been identified. They are not linear although there are some time dimensions and sequences involved in the staged processes. The stages of implementation are exploration, installation, initial implementation, and full implementation.

During *exploration* information is exchanged and leads to a decision to attempt to use an innovation or not. Given the difficulties inherent in changing the status quo, an implementation team helps organizations and systems carefully consider the need for change, the availability of innovations

and implementation supports to fulfill the need, and the preparation of leadership for initiating and managing the change process over a few years. Convening groups of leaders, stakeholders, and partners is an important part of making the decision to implement an innovation. The understanding and support developed during the exploration stage helps organizations and practitioners negotiate the rocky shoals of change when later attempting to use innovations in complex human service environments. The availability of an implementation team ensures that the human resources are available to shepherd the innovation from an idea to full and effective use.

Installation preparations begin in earnest once a decision is made. Innovations typically have inclusion/exclusion criteria for participants and therefore affect referral processes. Supports for innovations also require extra resources for start-up, planning for training and coaching, and selecting staff who are ready and willing to try to make effective use of an innovation. Immediate impacts on organization and system functions, roles, and structures are considered and accommodations are agreed upon to help create a hospitable environment for practitioners to begin doing the new ways of work.

Some accumulation of resources (preparation of contracts, assignment of staff time) may begin during the exploration stage. During the installation stage, some resources promised during exploration may be difficult to obtain and may lead to further exploration work to bring more people to the table to have them be part of the innovation implementation process.

Implementation is a dynamic process and the stages are overlapping and nonlinear. Implementation teams are essential to expediting the process of exploration and anticipating the needs for installation stage resources. Lack of readiness and lack of resources frequently are found on lists of impediments to success. Experienced implementation teams can help focus attention, anticipate common issues, and avoid wasted time and effort during these early stages.

Initial implementation begins when the first newly trained practitioner attempts to use an innovation in his or her interactions with intended beneficiaries. Many attempts to use innovations fail at this point. As attempts commence to use an innovation in actual practice, the status quo is disturbed. When the use of an innovation requires changes in fundamental operations of an existing organization, the disturbance is more substantial and impacts the status quo more dramatically. Without engaged leadership and persistent implementation supports, the innovation likely will fail (Marzano et al., 2005; Nord and Tucker, 1987).

The availability of skilled implementation teams is important in each stage of implementation, and especially helpful for negotiating the multiple issues that arise during initial implementation. Given their expertise and experience, implementation team members anticipate reactions by the status quo and help organizations quickly change their standard operating methods to accommodate and support the innovation. If serious issues are encountered during initial implementation, it may require reengaging in exploration or installation stage activities to secure agreements or resources found lacking during the initial implementation attempt. The goal of an implementation team is to support the use the innovation as intended and to help overcome barriers to realizing the benefits of the innovation in each organization setting.

Full implementation is achieved when half or more of the intended users of an innovation are using it as intended (with fidelity). Without the support of a skilled implementation team, very few innovations reach full implementation. Estimates of 5–15% are common (Green, 2008; Vernez et al., 2006; Wiltsey Stirman et al., 2012). This means that few innovations such as evidence-based programs actually produce socially significant benefits in practice. Without implementation teams, efforts to support the use of innovations in various practice settings likely will be haphazard and reactive instead of purposeful and proactive.

Skilled implementation teams greatly improve the odds of producing higher rates of success (60–80%) in shorter periods of time (Saldana and Chamberlain, 2012). Implementation teams are not commonly available or purposely constituted in human services but are growing in number and sophistication as their value is recognized.

Effective Implementation: Drivers

Implementation drivers are the heart of the change processes that support the full, effective, and sustained use of innovations in complex human service environments.

In their review of the diffusion and dissemination literature, Greenhalgh et al. (2004) noted three categories of implementation-related activities: letting it happen, helping it happen, and making it happen. Letting it happen summarizes many attempts to use innovations that rely on diffusion of information through networks and communication patterns (e.g., publication of articles; conferences; champions). Others are helping it happen by providing ready access to useful summaries of the literature, Web sites aimed at practitioner and policy audiences, and persuasive communications in the form of social marketing. The idea is to get relevant information into the hands of those who could be using the innovation so they are better informed and more likely to find ways to use the innovation in their daily practice.

Letting it happen and helping it happen approaches result in about 5–15% use of innovations as intended. Making it happen approaches are quite different. They offer purposeful, active, and persistent supports for using innovations as intended and producing promised results in practice. Most of the 25 frameworks reviewed by Meyers et al. (2012), including the active implementation frameworks described in this article, fit the making it happen category.

A review of the implementation evaluation literature (Fixsen et al., 2005) found evidence in support of several factors that seem to drive full and effective uses of innovations. As shown in **Figure 1**, there are competency drivers, organization drivers, and leadership drivers associated with consistent use of innovations in typical human service settings and the production of intended benefits (Fixsen et al., 2009). As noted in the center of the triangle, the various components (e.g., coaching, facilitative administration) are integrated and focused on using the innovation as intended (e.g., fidelity). Furthermore, strengths in some components can be used purposefully to compensate for weaknesses in other components to still realize the goal of

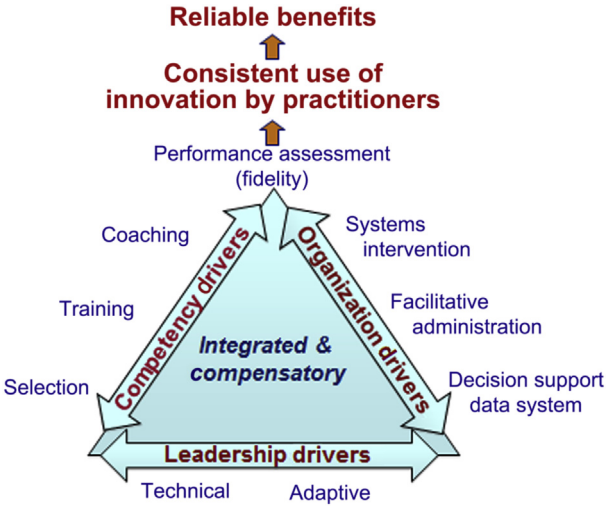

Figure 1 Implementation drivers. Reproduced from Fixsen, D.L., Blase, K.A., Duda, M., Naoom, S.F., Van Dyke, M., June 2008. Establishing an infrastructure for implementing substance abuse programs. Invited presentations and workshop, OASAS, Albany, NY; used with permission.

producing use of the innovation as intended and therefore, reliable benefits. Again, the value of an implementation team is apparent to assure the use of implementation best practices to realize the intended benefits of innovations in typical human service settings.

The logic connecting implementation and innovations is clear in Figure 1. By definition, an innovation is new to practitioners and organizations. Thus, practitioners (and others) need to learn the new ways of work embodied in an innovation. Staff selection, training, coaching, and feedback from performance assessments frequently are cited in the literature as important ways to assure skilled use of innovations. Mentioned less often, but equally important, is the integration of these components. Too often practitioners leave training and find their new knowledge and skills are not supported by their supervisor ("that's not how we do things in my unit"). In an integrated approach, coaches (supervisors) are prepared to support and expand the knowledge and skills introduced in training.

Given the use of an innovation by practitioners, organizations need to change to accommodate and support the new ways of work if they are to be sustained. Creating a hospitable environment is the work of administrators who are alert for ways to change standard operating procedures to facilitate use of the innovation. It is also their task to remove impediments that are discovered as the innovation moves from initial implementation to full implementation. Facilitative administrators rely on data to inform their decisions as they seek to continually improve supports for practitioners and improve innovation outcomes. Given the multilevel impact of using innovations in existing systems, organizations need to help surrounding systems change to accommodate and facilitate the new ways of work (e.g., referrals, funding, accreditation, scheduling).

None of this will happen without engaged leadership (Marzano et al., 2005). Adaptive leadership is needed for solving difficult problems that arise when legitimate and competing interests collide (including wicked problems as defined by Rittel and Webber, 1973). For example, union interests to protect members from unnecessary intrusions in their workspace may conflict with requirements to conduct frequent performance (fidelity) assessments to assure use of an innovation. Adaptive issues require convening meetings to discover the issues underlying a problem, finding common ground with agreed upon outcomes, and arriving at consensus solutions that can be tried in practice and modified as needed to resolve the issue. Given the tentative identification of the problem and search for a solution, wicked problems requiring adaptive leadership often defy any permanent solution. They are said to be resolved at any point in time with the expectation that there will be an opportunity to resolve the problem again in the future.

Technical leadership is needed for managing daily operations of the organization, including assuring the operation of each of the implementation drivers at a high level or performance. These tame problems (Rittel and Webber, 1973) can be solved more readily since the problems are easier to identify, are agreed to be problematic, staff have the skills to solve them, and work groups can be formed and disbanded as needed to modify procedures, realize a solution, and establish the new methods as part of daily practice. Leaders need to engage in both technical and adaptive leadership skills. Applying technical leadership approaches to resolving adaptive issues can make the problems worse. Using adaptive leadership approaches to solve tame problems is inefficient. Implementation team members are alert to identifying the types of challenges they are facing and are involved in supporting the appropriate leadership strategies.

Effective Implementation: Improvement Cycles

In the early days of attempting to use evidence-based programs in typical practice settings, there was debate about the need to replicate an innovation exactly versus the need to adapt an innovation to local circumstances to achieve acceptance and use. There was also the notion that since an evidence-based innovation had research to support its effectiveness those results could be expected when employing it in practice. While these quaint ideas still have some currency in intellectual debates, the issues have been settled in practice: there is no such thing as exact replication, and no laboratory-developed innovation delivers in practice all the benefits one expects. In addition, the (sometimes wild) swings in combinations of public policy, funding, workforce availability, leadership, personal and family dynamics, and community values challenge even the most agile organizations to continue delivering excellent service with excellent outcomes in full implementation mode.

In implementation practice and science, innovations provide a place to begin. Reasonably so, evidence-based innovations offer higher expectations for good outcomes. Innovations that have been demonstrated to be effective and meet the usable intervention criteria have the practical information and supports needed to get started and then get better by using performance (fidelity) assessment and outcome data.

Adjustments and improvements are expected in the beginning and over time and are part of the continuing work of an implementation team.

Implementation teams make use of the plan, do, study, act (PDSA) improvement cycle designed to detect and correct errors and strengthen facilitators *en route* to realizing excellent outcomes. PDSA cycles sometimes are referred to as trial and learning cycles and fit well with the need to make changes in operations while preserving and improving intended outcomes. As used in implementation work, PDSA cycles take three forms (rapid-cycle, usability testing, practice-policy communication) that are distinguished by time frames (rapid, medium, longer term) and scope of issues (small, medium, large).

The first PDSA improvement cycles originally were used in manufacturing and business settings (Shewhart, 1939) and have since been employed in a wide variety of human service settings. Rapid-cycle uses of PDSA require frequent opportunities to do what was planned, study what has been done and the results, and make changes to improve the next plan. Note that rapid-cycle use of PDSA includes the assumption of clear accountability for achieving the desired outcomes. This is a good fit with the making it happen approach inherent in the active implementation frameworks. Rapid PDSA cycles are well suited to making improvements in teacher instruction behavior or therapist behavior from 1 day or week to the next. In these cases, plans can be adjusted and results known on a frequent basis and improvements can be made quickly. Implementation teams make use of rapid-cycle PDSA processes to develop usable interventions. Coaches make use of PDSA cycles as they work to expand the skills and improve the outcomes of practitioners.

Usability testing in human services is PDSA work with medium time frames and issues with more interactive components. Usability testing is done with iterative groups of four or five individuals (Akin et al., 2013; Genov, 2005). For innovation development or capacity building, the number of iterations equals the opportunities to learn, adjust, and learn again. For example, assume 20 people need to be trained and training capacity needs to be developed. Capacity to do training well means having skilled trainers who:

- know the material well,
- can deliver it effectively to adult learners,
- have established behavior rehearsal scenes,
- know how to carry out the behavior rehearsal leader and confederate roles so that trainees have the opportunity to practice essential skills (see usable intervention criteria), and
- have created useful pre-post training tests of knowledge and skill development.

If training is provided to all 20 people at one time, there is one opportunity to develop training capacity and any lessons learned have to wait until next time. If the focus broadens from not only training 20 people but also to developing effective and durable training capacity, then training is done with iterative groups of four or five people. The usability testing format provides more opportunities for the trainers to build their capacity (plan training materials, do training, see the pre-post results, make changes for improved training materials and methods, do it again), while still providing training for all 20 people. The difference is substantial when developing usable interventions and establishing effective implementation supports.

Practice–policy communication cycles are another example of the application of PDSA logic in human services (Fixsen et al., 2013). For example, implementation team members report monthly to the executive managers regarding the systemic facilitators for and barriers to achieving and sustaining excellent outcomes. The implementation team is highly involved in reality at the practice level, can use their first-hand knowledge and data to describe to policy makers what is working or not, and can offer suggestions for how to change roles, functions, and structures in the organization/system to support improved outcomes. Based on this information, policy makers and executive leaders have the authority to make changes to facilitate the full, effective, and sustained use of innovations to achieve socially significant outcomes. Iterative use of such feedback and problem-solving loops can result in a transformed organization or system that is increasingly designed to support effective interventions and effective implementation. It has become an enabling context for supporting the innovation and realizing excellent outcomes.

The use of improvement cycles in implementation work is as important as implementation drivers or stages. Every attempt to do something is an opportunity to learn how to do it better next time. Many people plan and do. Few study and act, and fewer yet use improvement cycles on purpose and persist until they realize intended outcomes.

Effective Implementation: Implementation Teams

Implementation teams are essential to full, effective, and sustained use of innovations. They are *who* do the work of implementation. Like surgical teams, sports teams, financial accounting teams, and others, implementation team members work together to achieve intended outcomes – innovation outcomes and implementation outcomes – in spite of the complexities faced in practice. The functions of implementation teams have been noted throughout this article. In summary, implementation team members are experts in the processes to establish usable innovations, know and contribute to the science of implementation, and participate in organization and system change processes to develop enabling contexts for effective innovations and effective implementation (see the formula for success). Implementation teams greatly increase the likelihood of successful use of innovations and shorten the time for reaching full implementation (Fixsen et al., 2005; Nord and Tucker, 1987; Saldana and Chamberlain, 2012).

Enabling Contexts: Creating Hospitable Environments

It is a truism that all organizations and systems are perfectly designed to achieve exactly the results they obtain ("If you do what you've always done …"). This statement does not impute intention to the design. Many organizations and systems are the accumulated product of decades of good ideas, leadership

plans, well-meaning reforms, and responses to political mandates and local needs. As noted in the formula for success, enabling contexts are needed to achieve and sustain socially significant outcomes. Effective innovations and effective implementation methods interact with the context. Thus, the context needs to change to facilitate the work of innovation practitioners and implementation teams if new (and improved) results are to be realized and sustained.

The process of creating enabling contexts is complex and involves a broad view of interactions among organization and system components within and between multiple layers. Fundamental to enabling contexts is the use of knowledge. Kenneth Arrow (1963) won the Nobel Prize for introducing the idea of information economics. The premise is that knowledge-based activities gain in value with use, unlike atom-based resources that are depleted with use. Thus, knowledge-based organizations and systems make use of information to enable the work in increasingly precise and flexible ways that in turn produce more and better information for ongoing improvements. One glance at the amazing progress of digital products and information systems over the past five decades gives credence to Arrow's premise.

Not every organization enables knowledge-based solutions; if that were the case, every organization would be performing at peak effectiveness and efficiency. According to Von Krogh et al. (2000), intended outcomes can be amplified by designing enabling contexts that purposefully:

- grapple with unique cultural and interpersonal conditions,
- dismantle as many individual and organizational barriers as possible (organization structures can facilitate or hinder enabling interactions within the organization),
- share tacit knowledge with microcommunities (such as implementation teams),
- align contexts, strategies, and structures.

They state that traditional organizations and systems can no longer coordinate activities "in a world where boundaries are fuzzy, relationships are ever more complex, and the competitive environment is in constant flux" (p. 177).

Other contributions come from complexity theory. Complexity theory is useful for understanding how even loosely interrelated components of systems can influence one another as they interact, and how small changes in a few parts of a system can result in massive changes in system functions and outcomes. The practice–policy communication cycle is built on complexity theory and relies on implementation teams to convey meaningful information and act on that information within a human service system (Fixsen et al., 2013). It provides a venue where "thinking can be challenged, issues about authority and the exercise of power candidly explored, and where participants can continue to learn and adapt to ever-changing circumstances" (Onyett et al., 2009, p. 3). For implementation work designed to use innovations at scale, the improvement cycles are essential for quickly finding solutions to thorny problems that inevitably arise in organization and system contexts.

The implementation field has developed well enough to begin research on these challenges. A longitudinal study of organizations that are attempting to use evidence-based programs provides a view of the factors influencing initial adoption, deadoption, readoption, and full implementation of innovations (Panzano et al., 2004). Recently, Glisson et al. (2010) carried out a seminal study of interactions among layers of services, organizations, and systems. This study combined an evidence-based approach to service delivery and an evidence-based approach to system change in a 2×2 experimental design where US counties were randomized to conditions. The results showed that each intervention was effective when used alone. However, the best outcomes for youths resulted when the interventions were combined, that is, when the evidence-based service was operating in a changed system.

The Glisson et al. (2010) experiment is the first to combine simultaneously the three components of the formula for success – evidence-based innovations supported by skilled implementation teams in contexts that were more or less enabling. Socially significant outcomes were the result of attention to each component. Note that the study would not have been possible to conduct in the absence of skilled implementation teams who were well prepared to produce high-fidelity examples of each evidence-based intervention. This study is a good example of how implementation done well in practice creates a better laboratory in which to study implementation variables.

The lack of implementation teams and implementation expertise is a major impediment to realizing goals in human service systems (Schofield, 2004). Human behavior is complex, interactions among humans add to the complexity, and provider organizations and systems often lack the information and skills needed to make meaningful change. Human service systems simultaneously are ever changing and intractable. Improved practices to initiate and mange change and conduct research on system change is the next great frontier in implementation science.

Conclusion

Implementation science is a new science and is reaching a point of geometric growth. Implementation frameworks provide guidance for implementation practice, and improved practice provides a better laboratory for studying implementation done well. This creates a virtuous cycle that feeds on itself to rapidly develop better practices and better science. The value of virtuous cycles is seen in the last century in the computer industry where crude silicone-based integrated circuits were used to develop and manufacture more effective and efficient integrated circuits. Those improved versions then were used to create the next generation of even better integrated circuits. Implementation science is on the verge of following this path with the development of active implementation frameworks to guide the next generation of improved frameworks.

The long-time frames for implementation slow the process of discovery and improvement. Thus, there is a premium on using improvement cycles on purpose, recording data at each point of the journey, and improving the efficiency of the process as well as its effectiveness. The complexity and long-time frames inherent in implementation present challenges related to measures, research designs, and theory development. The transdisciplinary involvement in implementation makes language and conceptual thinking difficult and frustrates efforts

to find and summarize the literature(s). Nevertheless, implementation science has come a long way and is on the brink of advancing rapidly in the coming decades.

Acknowledgments

Preparation of this article was supported, in part, by funds from the US Department of Education Office of Special Education Programs (H326K070003). The views expressed are those of the authors and should not be attributed to the funding source. We would like to thank our National Implementation Research Network colleagues who provide continual inspiration and delight.

See also: Applied Social Research, History of; Cost-Benefit and Cost-Effectiveness Analyses in Evaluation Research; Evidence-Based Practice; Policy Analysis; Prevention Research.

Bibliography

Akin, B.A., Bryson, S.A., Testa, M.F., Blase, K.A., McDonald, T., Melz, H., 2013. Usability testing, initial implementation, and formative evaluation of an evidence-based intervention: lessons from a demonstration project to reduce long-term foster care. Evaluation and Program Planning 41 (0), 19–30. http://dx.doi.org/10.1016/j.evalprogplan.2013.06.003.

Arrow, K.J., 1963. Uncertainty and the welfare economics of medical care. The American Economic Review 53, 941–973.

Blase, K.A., Fixsen, D.L., 2003. Evidence-Based Programs and Cultural Competence. National Implementation Research Network, Louis de la Parte Florida Mental Health Institute, University of South Florida, Tampa, FL pp. I–20.

Blase, K.A., Fixsen, D.L., Naoom, S.F., Wallace, F., 2005. Operationalizing Implementation: Strategies and Methods. University of South Florida, Louis de la Parte Florida Mental Health Institute, Tampa, FL.

Blase, K.A., Van Dyke, M., Fixsen, D.L., 2009. Evidence-based programming in the context of practice and policy. Social Policy Report 23 (3), 13–14.

Blase, K., Van Dyke, M., Fixsen, D., Wallace Bailey, F., 2012. Implementation science: key concepts, themes, and evidence for practitioners in educational psychology. In: Kelly, B., Perkins, D. (Eds.), Handbook of Implementation Science for Psychology in Education. Cambridge University Press, London, pp. 13–34.

Brownson, R.C., Colditz, G.A., Proctor, E.K. (Eds.), 2012. Dissemination and Implementation Research in Health. Oxford University Press, New York.

Eysenck, H., 1952. The effects of psychotherapy: an evaluation. Journal of Consulting Psychology 16, 319–324.

Fixsen, D.L., Blase, K.A., Naoom, S.F., Wallace, F., 2009. Core implementation components. Research on Social Work Practice 19 (5), 531–540.

Fixsen, D.L., Naoom, S.F., Blase, K.A., Friedman, R.M., Wallace, F., 2005. Implementation Research: A Synthesis of the Literature. University of South Florida, Louis de la Parte Florida Mental Health Institute, National Implementation Research Network. (FMHI Publication No. 231), Tampa, FL.

Fixsen, D., Blase, K., Metz, A., Van Dyke, M., 2013. Statewide implementation of evidence-based programs. Exceptional Children 79 (2), 213–230 (Special Issue).

Genov, A., 2005. Iterative usability testing as continuous feedback: a control systems perspective. Journal of Usability Studies 1 (1), 18–27.

Glisson, C., Schoenwald, S.K., Hemmelgarn, A., Green, P., Dukes, D., Armstrong, K.S., Chapman, J.E., 2010. Randomized trial of MST and ARC in a two-level evidence-based treatment implementation strategy. Journal of Consulting and Clinical Psychology 78 (4), 537–550.

Green, L.W., 2008. Making research relevant: if it is an evidence-based practice, where's the practice-based evidence? Family Practice 25, 20–24.

Greenhalgh, T., Robert, G., MacFarlane, F., Bate, P., Kyriakidou, O., 2004. Diffusion of innovations in service organizations: systematic review and recommendations. The Milbank Quarterly 82 (4), 581–629.

ICEBeRG Group, 2006. Designing theoretically-informed implementation interventions. Implementation Science 1, 4.

Institute of Medicine, 2001. Crossing the Quality Chasm: A New Health System for the 21st Century. National Academy Press, Washington, DC.

Kessler, R.C., Glasgow, R.E., 2011. A proposal to speed translation of healthcare research into practice: dramatic change is needed. American Journal of Preventive Medicine 40 (6), 637–644.

Marzano, R., Waters, T., McNulty, B., 2005. School Leadership that Works: From Research to Results. Association for Supervision and Curriculum Development (ASCD), Alexandria, VA.

Meyers, D.C., Durlak, J.A., Wandersman, A., 2012. The quality implementation framework: a synthesis of critical steps in the implementation process. American Journal of Community Psychology 50 (3-4), 462–480. http://dx.doi.org/10.1007/s10464-012-9522-x.

Naleppa, M.J., Cagle, J.G., 2010. Treatment fidelity in social work intervention research: a review of published studies. Research on Social Work Practice 20 (6), 674–681. http://dx.doi.org/10.1177/1049731509352088.

Nord, W.R., Tucker, S., 1987. Implementing Routine and Radical Innovations. D.C. Heath and Company, Lexington, MA.

Onyett, S., Rees, A., Borrill, C., Shapiro, D., Boldison, S., 2009. The evaluation of a local whole systems intervention for improved team working and leadership in mental health services. The Innovation Journal: The Public Sector Innovation Journal 14 (1), 1018.

Panzano, P.C., Seffrin, B., Chaney-Jones, S., Roth, D., Crane-Ross, D., Massatti, R., Carstens, C., 2004. The innovation diffusion and adoption research project (IDARP). In: Roth, D., Lutz, W. (Eds.), New Research in Mental Health, vol. 16. Ohio Department of Mental Health Office of Program Evaluation and Research, Columbus, OH, pp. 78–89.

Rittel, H.W.J., Webber, M.M., 1973. Dilemmas in a general theory of planning. Policy Sciences 4, 155–169.

Rogers, E.M., 1995. Diffusion of Innovations, fourth ed. The Free Press, New York.

Saldana, L., Chamberlain, P., 2012. Supporting implementation: the role of community development teams to build infrastructure. American Journal of Community Psychology 50 (3–4), 334–346. http://dx.doi.org/10.1007/s10464-012-9503-0.

Schofield, J., 2004. A model of learned implementation. Public Administration 82 (2), 283–308.

Shewhart, W.A., 1939. Statistical Method from the Viewpoint of Quality Control. Dover Publications, New York.

Tabak, R.G., Khoong, E.C., Chambers, D.A., Brownson, R.C., 2012. Bridging research and practice: models for dissemination and implementation research. American Journal of Preventive Medicine 43 (3), 337–350.

Van Meter, D.S., Van Horn, C.E., 1975. The policy implementation process: a conceptual framework. Administration & Society 6, 445–488.

Vernez, G., Karam, R., Mariano, L.T., DeMartini, C., 2006. Evaluating Comprehensive School Reform Models at Scale: Focus on Implementation. RAND Corporation, Santa Monica, CA.

Von Krogh, G., Ichijo, K., Nonaka, I., 2000. Enabling Knowledge Creation. Oxford University Press, New York.

Westphal, J.D., Gulati, R., Shortell, S.M., 1997. Customization or conformity? An institutional and network perspective on the content and consequences of TQM adoption. Administrative Science Quarterly 42 (2), 366–394.

Wiltsey Stirman, S., Kimberly, J., Cook, N., Calloway, A., Castro, F., Charns, M., 2012. The sustainability of new programs and innovations: a review of the empirical literature and recommendations for future research. Implementation Science 7 (1), 7–17. http://dx.doi.org/10.1186/1748-5908-7-17.

Implicit Association Test

Sarah Teige-Mocigemba and Karl Christoph Klauer, University of Freiburg, Freiburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Implicit Association Test (IAT) is one of the most often used indirect measures of attitudes, stereotypes, or personality traits in psychology. It is indirect in that participants are not asked to report on their attitudes, stereotypes, or dispositions. Instead, this information is inferred from participants' performance in the IAT. We review the IAT's psychometric properties in terms of reliability and validity. Theories accounting for IAT effects are briefly discussed. Different approaches to deal with confounding influences on IAT effects are considered.

The Implicit Association Test (IAT) was first proposed in 1998 by Anthony Greenwald and colleagues (Greenwald et al., 1998). It is a categorization task that provides a flexible framework within which indirect measures of attitudes, stereotypes, and diverse behavioral dispositions can be designed. The measures are indirect in the sense that participants undergoing an IAT are not asked to report on their attitudes, stereotypes, or dispositions. Instead, they are asked to work as fast and accurately as possible on the categorization task, and attitudes, stereotypes, or dispositions are inferred from their performance in the task. Even when participants may suspect that attitudes, stereotypes, or dispositions are to be measured, it is not immediately obvious to them how task performance is used to assess these variables. Indirect measures thereby aim at overcoming important limitations of direct measures such as self-reports, in particular the susceptibility to self-presentational distortions (e.g., social desirable responding) and introspective limits (e.g., Nisbett and Wilson, 1977). In contrast to most other indirect measures, the IAT measures were found to exhibit satisfactory reliability estimates when reliability is assessed in terms of the internal consistency of the measure. This and the procedure's flexibility promoted the widespread use of the IAT in diverse psychological research domains.

This article is organized as follows: first, we describe the IAT procedure. This is followed by a review of the IAT's psychometric properties in terms of reliability and validity. Theories accounting for IAT effects are then briefly discussed along with mathematical process models based on these theories. Finally, different approaches on how to deal with confounding influences on IAT effects are considered.

Procedure

The IAT requires participants to assign stimuli of four categories, two target categories (e.g., Black vs White) and two attribute categories (e.g., negative vs positive), as fast and accurately as possible to two response keys. A typical IAT consists of seven blocks (for details, see **Table 1**), three of which are practice blocks acquainting participants with the categorization rules, the target stimuli (e.g., portraits of Blacks vs Whites), and the attribute stimuli (e.g., negative vs positive adjectives). In the remaining diagnostically relevant test blocks, participants are to categorize stimuli of all four categories, using only two response keys. The typical finding is that participants respond faster and more accurately, when two overlapping categories share one response (e.g., Black and negative vs White and positive for a prejudiced person) than when two categories with little overlap share one response (e.g., Black and positive vs White and negative for a prejudiced person). This performance difference between the two differently combined categorization tasks is reflected in the IAT effect, which is interpreted as revealing size and direction of the association strength between target categories and attribute categories.

Table 1 Example of a racial attitude Implicit Association Test: task sequence and recommended implementation details

Task sequence				
			Response key assignment	
Block	N *of trials*	*Task*	*Left key*	*Right key*
1	20	Target discrimination	Black	White
2	20	Attribute discrimination	Negative	Positive
3	20	Initial combined task	Black, negative	White, positive
4	40	Initial combined task	Black, negative	White, positive
5	20 or 40	Reversed target discrimination	White	Black
6	20	Reversed combined task	White, negative	Black, positive
7	40	Reversed combined task	White, negative	Black, positive

Recommended implementation details	
Instruction	Respond as fast and accurately as possible!
Correction of errors	Required, error cue: red 'X' or 'error' below stimulus
Category labels	Displayed in the corresponding upper screen corners
Intertrial interval	150 ms–750 ms (most often used: 250 ms)
Number of stimuli	Five to six per category (at least two)
Trial order in combined blocks	Alternation between target and attribute stimuli, otherwise randomized if group differences are the main focus of the experiment, but fixed for all participants in correlational studies

Note: Nosek et al. (2007) recommended increasing the number of trials in the fifth block as an effective means to reduce order effects of the combined tasks (see the section Handling Confounding Influences on the IAT Effect). Therefore, some IAT procedures present 40 instead of 20 trials in the block, in which participants practice the reversed target discrimination task.

For example, if performance is faster and more accurate in the Black/negative–White/positive block (mapping 'Black' and 'negative' on one key and 'White' and 'positive' on the other key) than in the White/negative–Black/positive block (mapping 'White' and 'negative' on one key and 'Black' and 'positive' on the other key), then this is considered to indicate the existence of implicit racial preferences favoring Whites over Blacks. The IAT effect is thus a relative measure: A racial attitude IAT, for example, provides only information about an individual's preference for one category over the other (e.g., a preference for Whites over Blacks or vice versa). In contrast, conclusions about an individual's evaluation of a single category (e.g., Blacks) cannot be derived (for a review of indirect measures that allow for the assessment of single categories, see Teige-Mocigemba et al., 2010).

For the implementation of IATs (for details, see **Table 1**), software packages such as E-Prime, Inquisit, and Direct-RT, among others, can be used and sample programs are available on the Internet (see list of Relevant Websites). Among the many scoring procedures that can be found in the literature, the so-called conventional IAT score and the so-called *D* measures are used most often to calculate IAT effects (for details, see Greenwald et al., 2003). The conventional IAT score is based on the difference between log-transformed, correct response latencies of the two combined tasks (e.g., sixth and seventh blocks minus third and fourth blocks in **Table 1**), with latencies smaller than 300 ms or greater than 3000 ms being recoded to 300 ms and 3000 ms, respectively. For descriptive statistics, however, IAT scores based on raw response latencies are often reported.

The *D* measures, on the other hand, use modified upper and lower tail treatment of latencies, include latencies of both correct and incorrect responses (with the latter being increased by an error penalty), and apply an individual standardization similar to that in Cohen's effect size measure *d*. Due to the ongoing discussion about which scoring algorithm is to be preferred (e.g., Wentura and Rothermund, 2007), many researchers report analyses based on both algorithms.

Psychometric Properties of the IAT

Reliability

To date, only few indirect measures reach the reliability levels that have been repeatedly found for the IAT. Split-half reliabilities and internal consistencies (Cronbach's alpha) of IAT effects typically range from .70–.90 with slight differences depending on the method of calculation (Nosek et al., 2007). The IAT's test–retest reliability, on the other hand, reaches less satisfactory levels of .25–.69 (mean and medium estimates of about .50; Lane et al., 2007) with little variation as a function of the retest interval. This discrepancy between satisfactory internal consistency and unsatisfactory test–retest reliability has been attributed to the impact of context effects on IAT scores (e.g., Blair, 2002). Such context effects might indicate the IAT's sensitivity to changes in the construct of interest (i.e., IAT as a state measure) and/or the IAT's susceptibility to changes in additional sources of variance not related to the construct to be measured (i.e., IAT as an insufficiently reliable trait measure) such as test taking strategies, learning effects, or other component response processes that do not reflect associations, per se. We come back to such contaminants of IAT effects below.

Validity

Group Level

Evidence for the IAT's validity comes from experiments at the group level and studies at the individual level. At the group level, firstly, the IAT has been shown to reflect universal attitudes for which the direction can be assumed on a priori grounds (e.g., a preference for flowers over insects, Greenwald et al., 1998). Secondly, the IAT has been shown to reflect to-be-expected differences between specific groups. For example, homosexuals and heterosexuals differed in their homosexuality attitude IAT effects (Banse et al., 2001), and White and Black participants differed in their racial attitude IAT effects (Nosek et al., 2002). Groups differing with regard to stigmatized behavior, on the other hand, were not consistently distinguished by an IAT (e.g., see Swanson et al., 2001). Thirdly, the IAT has been shown to reflect newly formed attitudes (Olson and Fazio, 2001), indicating its ability to capture recently acquired preferences. In sum, the empirical evidence at the group level confirms the validity of the IAT. Note, however, that this (quasi)-experimental validation approach does not exclude the possibility that uncontrolled third variables account for the expected findings. For instance, instead of universal attitudes, specific characteristics of the selected stimulus sets may be responsible for the observed IAT effects (e.g., see Bluemke and Friese, 2006). Likewise, groups may differ not only with regard to the construct of interest (e.g., racial attitude), but also with regard to their cognitive abilities, which in turn might influence IAT effects (e.g., see Klauer et al., 2010).

Individual Level

At the individual level, firstly, the IAT has been shown to correlate with self-report measures with correlations ranging from about .24 (according to a meta-analysis of Hofmann et al., 2005) to .37 (according to a large-scale analysis of Nosek, 2005) with differences in correlations depending on the construct of interest (e.g., whether it addresses a socially sensitive domain or not), on the corresponding self-report measure (e.g., relative thermometer scale or questionnaire), and on the construct variability in the sample, among other influencing variables. These low-to-moderate correlations have been interpreted as evidence for the IAT's validity, although the to-be-expected size of indirect–direct correlations is still debated: Assuming that an IAT and a self-report measure tap the cognitive representation of one and the same construct in different ways, the IAT's correlations with self-report measures should be moderate to high (e.g., Nier, 2005). On the other hand, assuming that the IAT taps the cognitive representation of an 'implicit' construct that is independent of, or only weakly related to, the cognitive representation of an 'explicit' construct captured by self-report measures, the IAT's correlations with self-report measures should be low (e.g., Wilson et al., 2000).

In any event, at least the IAT's correlations with other indirect measures should be high because different indirect measures targeting the same construct should indeed tap the same, possibly implicit construct. Indirect–indirect correlations,

however, have been shown to be weak at best (e.g., Rudolph et al., 2008). These findings are often attributed to (1) insufficient reliability of indirect measures other than the IAT which limits indirect–indirect correlations (e.g., Teige et al., 2004) and/or (2) to structural misfit (i.e., different task demands) of indirect measures which has been shown to decrease correlations (see Payne et al., 2008).

Finally, the IAT has been shown to predict behavior over and above direct measures (e.g., Asendorpf et al., 2002). Particularly in socially sensitive domains such as prejudice, the IAT proved to be a better behavioral predictor than self-report measures, whereas it performed worse than self-report measures in domains that are less likely to be subject of self-presentational tendencies (Greenwald et al., 2009) or in health-related domains (e.g., Spruyt et al., 2007).

In sum, the empirical evidence at the individual level also confirms the validity of the IAT. As for the group level, however, uncontrolled third variables might in part account for the IAT's correlations with external criteria. For instance, as elaborated on below, cognitive abilities have been shown to contribute to the size of IAT effects (e.g., Klauer et al., 2010). Small, but significant correlations between IATs and behavior may thus not rely on interindividual differences in the construct of interest (e.g., aggressiveness), but on interindividual differences in cognitive skills that influence both an (e.g., aggressiveness) IAT and the likelihood to show a specific (e.g., aggressive) behavior. To assess the extent to which individual differences in IAT effects reflect variance due to confounded variables not related to the construct to be measured, it is helpful to gain an understanding of the psychological processes that shape the IAT effect and that permit confounded variables to exert their influences on IAT effects.

Process Models of the IAT

Stimulus–Response Compatibilities

De Houwer (2001, 2003) proposed that IAT effects rely on differences in stimulus–response compatibilities. According to this model, response keys acquire the meaning of the categories they are assigned to in the course of conducting an IAT (e.g., left key acquires negative meaning, right key acquires positive meaning). A presented IAT stimulus, on the other hand, is assumed to preactivate the response whose meaning is compatible with the stimulus features resulting in better performance, if the preactivated response is the correct response (i.e., facilitation), and in worse performance, if the preactivated response is the false response (i.e., response conflict). In a racial attitude IAT, for example, a prejudiced person should thus respond fast and accurately, if the categorization task requires pressing the negative key for a Black stimulus (i.e., stimulus–response compatibility). Conversely, the same prejudiced person should respond slow and less accurately, if the categorization task requires pressing the positive key for a Black stimulus (i.e., stimulus–response incompatibility). IAT effects are thus assumed to rely on differences of stimulus–response compatibilities between the test blocks (e.g., high stimulus–response compatibility for the Black/negative–White/positive block, low stimulus–response compatibility for the White/negative–Black/positive block for a prejudiced person).

Importantly, in De Houwer's account, not only may so-called 'relevant' stimulus features (i.e., construct-related features such as category membership) preactivate responses in an IAT, but also so-called 'irrelevant' stimulus features (i.e., features not related to the construct to be measured such as perceptual form). As discussed by De Houwer et al. (2005), this opens the door for a host of different variables to exert an influence on IAT effects because each and every feature according to which stimulus and response in an IAT can be compatible might impact on the IAT effect via stimulus–response compatibilities.

Random-Walk Processes

Brendl et al. (2001) proposed that IAT effects rely on differences in random-walk processes. According to this model, accumulation of stimulus information drives an internal counter toward one of two response thresholds. As soon as one response threshold is reached, the respective response is initiated (e.g., left or right key press). In an IAT, it is assumed that both information about target categories (such as category membership) and information about attribute categories (e.g., valence) drive the counter for each stimulus. If associated categories are assigned to one response key, the target information and the attribute information push the accumulation process in the same direction, resulting in fast and accurate performance. If, on the other hand, associated categories are assigned to different response keys, target and attribute information move the accumulation process in opposite directions, resulting in slower and less accurate performance.

For instance, if a Black stimulus is presented to a prejudiced person in the Black/negative–White/positive block, the information about category membership of the Black stimulus (i.e., a Black stimulus belongs to the category Black) and the information about valence of the Black stimulus (i.e., Black stimuli are negative for a prejudiced person) push the accumulation process toward the same response (i.e., to the Black/negative response key). In the White/negative–Black/positive block, on the other hand, category and valence information of the same Black stimulus move the accumulation process toward opposite responses, namely, to the (White/)negative key and to the Black(/positive) key. In most cases, information about category membership will have the stronger impact (i.e., responses will be correct in most trials). However, net accumulation rates of the test block with incompatible mapping will be lower than net accumulation rates of the test block with compatible mapping, leading to worse performance in the incompatible block. Such differences in net accumulation rates are thus assumed to contribute to performance differences between the two test blocks that constitute the IAT effect.

As a further process, Brendl et al. (2001) proposed that differences in response criteria contribute to the observed performance differences between test blocks: Because participants perceive the test block with incompatible mapping (referred to as incompatible block in the following) as more difficult than that with compatible mapping (referred to as compatible block in the following), they would adopt a more conservative response criterion in the incompatible block leading to slower responses. Brendl et al. further assumed that the first mechanism, different rates of information

accumulation, produces construct-related variance in the IAT effect, whereas the second mechanism, different speed-accuracy trade-offs mediated by different response criteria, constitutes a source of variance driven by strategic, situational, and/or trait-related influences on the IAT effect.

These assumptions were broadly confirmed by diffusion-model analyses of IAT data (Klauer et al., 2007). A diffusion model can be seen as an implementation of a random-walk model. The diffusion-model analysis allows one to decompose observed IAT effects into a component due to rates of information accumulation, tapping construct-related variance, and a component due to response criteria, tapping systematic variance due to speed-accuracy settings that is method-related (i.e., shared by IATs measuring different constructs), but not construct-related (i.e., does not covary with the to-be-measured variable).

Task-Switching Costs

Mierke and Klauer (2001, 2003) proposed that differences in task-switching costs contribute to IAT effects. Remember that a typical IAT task requires participants to apply either a target task (e.g., Does the presented stimulus belong to the category Black or White?) or an attribute task (e.g., Is the presented stimulus negative or positive?) depending on whether a target or an attribute stimulus is presented. In order to apply the correct task-set, participants thus have to switch between task-sets because target and attribute stimuli typically alternate from trial to trial. Task-set switching, however, is costly in that it requires time and cognitive resources. Any asymmetry of task-switching costs between the compatible and incompatible block may thus directly contribute to an IAT effect.

This is exactly what the task-switching account postulates. In the compatible block, it is possible to simplify the task, omitting many task-set switches, because in most cases it will be sufficient to apply the attribute task to both attribute and target stimuli. For instance, if a prejudiced person has to categorize a Black stimulus in the Black/negative–White/positive block, applying the attribute task (i.e., respond with 'negative') to the Black stimulus will lead to the same correct response as applying the target task (i.e., respond with 'Black') because the associated categories 'negative' and 'Black' are mapped onto the same response key. In the White/negative–Black/positive block, on the other hand, responding to a Black stimulus according to its valence will lead to an incorrect response. In the incompatible block, it is therefore required to identify and to apply the correct task-set in order to respond correctly. This asymmetry in the necessity of task-set switching implies that task-switching costs will be higher in the incompatible than in the compatible block. The differential task-switching costs thereby directly contribute to the IAT effect. Because there are interindividual differences in task-switching ability, the task-switching account also predicts a cognitive skill confound of IAT effects.

Note that in principle, the task-switching account allows for different confounded variables to exert an influence on IAT effects. This is because simplifying the IAT task in the compatible block is not necessarily based on the attribute dimension (e.g., valence), but can be based on any arbitrary feature shared by the attribute and the target stimuli. Accordingly, inasmuch as participants derive their responses in the compatible block from an arbitrary feature other than the attribute dimension, the IAT effect would reflect construct-unrelated rather than construct-related variance. Furthermore, the more participants tend to simplify the task in the compatible block, the higher is the to-be-expected contribution of individual differences in task-switching ability to IAT effects.

Figure-Ground Asymmetries

Rothermund and Wentura (2001, 2004) proposed that IAT effects rely on differences in the salience of stimulus categories. Similarly to the task-switching account, the figure-ground asymmetry account postulates that participants simplify the compatible block, but not the incompatible block of an IAT. In contrast to the task-switching account, however, it is assumed that the feature according to which participants categorize stimuli in simplifying the task is always salience. Specifically, Rothermund and Wentura argue that participants recode the attribute and the target tasks as figure-ground discriminations in the compatible block. For instance, negative stimuli should be more salient than positive stimuli because of the attention-grabbing power of negative information. Likewise, Black stimuli are assumed to be more salient than White stimuli because Black stimuli such as Black faces or names are less familiar (in Western societies). Accordingly, if the more salient stimuli (e.g., Black and negative stimuli) are to be mapped onto the same response key and the less salient stimuli onto the other as is the case in the compatible block, the complex 4-to-2 categorization task can be reduced to a binary figure-ground discrimination task of whether a presented stimulus belongs to the salient (i.e., figure) category or to the less salient (i.e., ground) category. Responses in the compatible block are therefore fast and accurate. If, however, the salient stimuli require different responses as is the case in the incompatible block, recoding the tasks as figure-ground discriminations cannot be used to simplify the task, leading to overall slower and less accurate responses. Figure-ground asymmetries, not valence associations, are thus assumed to constitute the size of IAT effects.

Note, however, that this does not exclude the possibility that the IAT effect comprises construct-related proportions of variance. That is, inasmuch as the valence of the target categories is a determinant of salience, salience may serve as a mediator between valence and IAT scores, thereby allowing for construct-related variance in IAT effects. For instance, inasmuch as Black stimuli are salient for a prejudiced person because they are negative (not because they belong to an unfamiliar category), the IAT effect may comprise construct-related proportions of variance.

Based on the idea that participants simplify the compatible block, Meissner and Rothermund (2013) have recently proposed a multinomial model, the ReAL model. The ReAL model uses the accuracy data of an IAT to disentangle an estimate for a component of the IAT effect that reflects simplifying processes and an estimate for a component mediated by residual associations that remain once the simplifying

processes have been factored out. Applying the model will, however, typically require the use of a modified IAT procedure.

The Quad Model

Like the diffusion model and the ReAL model, the Quad model (Conrey et al., 2005) aims to decompose observed IAT effects into different process components offering estimates of the contribution of these different processes. Specifically, the Quad model estimates (1) the ability to perform the separate categorization tasks and (2) the influence that associations between attributes and categories have on task performance, an influence that is additionally modulated by (3) the ability to overcome biased responses suggested by activated associations. Finally, the Quad model estimates (4) a guessing parameter that is assumed to drive responses in the absence of other information. Like the ReAL model, the Quad model is a multinomial model that uses only the accuracy data of the IAT. Unlike the ReAL model, it can be applied to the conventional IAT procedure. Both the ReAL and the Quad model require, however, a sufficient number of errors to be applicable.

Handling Confounding Influences on IAT Effects

The just reviewed process models suggest that the IAT effect is shaped by additional variables over and above the associations to be measured. Operating via different processes such variables include cognitive abilities (as predicted by the task-switching account), salience (as predicted by the figure-ground asymmetry account), and in fact any relevant or irrelevant feature that (1) allows for recoding the four categories of an IAT into two categories in the compatible IAT block (as predicted by the task-switching account), or that (2) generates a compatibility relation between stimuli and response keys (as predicted by the stimulus–response compatibility account). Finally, the random-walk account suggests an influence of the adoption of different response criteria for the compatible and incompatible blocks (i.e., differential speed-accuracy trade-off settings).

Indeed, there is evidence that cognitive abilities (e.g., Klauer et al., 2010), salience (e.g., Rothermund and Wentura, 2004), perceptual form (e.g., De Houwer et al., 2005), familiarity (e.g., Kinoshita and Peek-O'Leary, 2006), biased stimulus selections forcing redefinitions of the category labels (e.g., Govan and Williams, 2004), extrapersonal knowledge (cf Olson and Fazio, 2004), faking strategies (e.g., Röhner et al., 2011), and order of compatible and incompatible blocks (e.g., Nosek et al., 2007), among others (for a more detailed review, see Teige-Mocigemba et al., 2010) contribute to IAT effects.

In the absence of a comprehensive process model that takes all variables influencing an IAT effect into account, it is difficult to exactly determine the relative contribution of construct-related and confounding influences on the IAT effect. However, different methods exist to safeguard against the possibility that an observed effect or correlation reflects the operation of confounds instead of being caused by the construct to be measured. We will discuss four approaches that can be taken.

One method is to use the mathematical models just reviewed in the attempt to disentangle construct-related contributions from effects and correlations due to confounding factors as operationalized in each model. A second method is to obtain measures of confounding variables such as cognitive abilities or speed-accuracy trade-off settings using conventional regression analyses to attempt to control for their potential influence. A third method is to complement the IAT of interest with a maximally similar IAT that is, however, designed to measure an unrelated construct (R. Greifeneder, personal communication, 20 June 2013). For example, in addition to an IAT measuring racial prejudice (i.e., an IAT designed to measure the association between negative and Black and between positive and White), an IAT designed to measure the association between positive and persons with white shirts on the one hand and between negative and persons with black shirts on the other hand might be administered. Any effects or correlations that are shown by both these IATs would be attributed to confounded factors, whereas effects and correlations limited to the racial attitude IAT could be more safely attributed to an effect or correlation due to implicit racial preferences.

A final, related method would be to administer an alternative indirect measure designed to tap the same construct using different procedures (see Teige-Mocigemba et al., 2010; for a review). Inasmuch as the alternative measure and the IAT would share construct-related variance, but be less similar in terms of their susceptibility to the same confounding variables, effects and correlations shown by both dependent variables could be more safely attributed to the construct to be measured.

Conclusion

With the publication of the IAT 15 years ago, Greenwald and colleagues introduced a measure that became one of the most often used and most controversially discussed indirect measures. In this article, we have reviewed research on strengths and shortcomings of the IAT. Despite convincing evidence for the IAT's ability to reliably assess associations between categories (i.e., construct-related variance), several contaminants of IAT effects (i.e., systematic variance not related to the construct to be measured) have been revealed operating via different processes. Like any measure, the IAT is thus not process-pure and does not provide direct, undistorted access to associative structures. Instead, performance in the IAT is shaped by a conglomerate of different processes opening the door for different variables to influence IAT effects. Taking this into account, research using the IAT will certainly continue to play a central role in contributing to our understanding of human experience and behavior in a surprising diversity of domains.

Acknowledgment

The preparation of this article was supported by a grant from the Baden-Württemberg-Stiftung (Eliteprogramme for Postdocs) to Sarah Teige-Mocigemba.

See also: Aggression, Social Psychology of; Attitude Formation and Change; Attitude Measurement; Attitudes and Behavior; Consumer Psychology; Control Behavior: Psychological Perspectives; Diagnostic Processes in Clinical Psychology and Psychotherapy; Diffusion and Random Walk Processes; Discrete-State Models of Information Processing; Five Factor Model of Personality, Assessment of; Forensic Psychology; Health Self-Regulation, Motivational and Volitional Aspects of; Implicit Social Cognition; Personality Assessment, Faking and; Personality Assessment, Forced-Choice; Personality Assessment: Overview; Personality Assessment; Personality and Self; Personality, Trait Models of; Prejudice and Discrimination; Projective Methods in Psychology; Self-Concept: From Unidimensional to Multidimensional and Beyond; Self-Esteem; Self-Regulated Learning; Social Cognition; Social Psychology: Research Methods; Social Psychology.

Bibliography

Asendorpf, J.B., Banse, R., Mücke, D., 2002. Double dissociation between implicit and explicit personality self-concept: the case of shy behavior. Journal of Personality and Social Psychology 83, 380–393.

Banse, R., Seise, J., Zerbes, N., 2001. Implicit attitudes toward homosexuality: reliability, validity, and controllability of the IAT. Zeitschrift für Experimentelle Psychologie 48, 145–160.

Blair, I.V., 2002. The malleability of automatic stereotypes and prejudice. Personality and Social Psychology Review 6, 242–261.

Bluemke, M., Friese, M., 2006. Do features of stimuli influence IAT effects? Journal of Experimental Social Psychology 42, 163–176.

Brendl, C.M., Markman, A., Messner, C., 2001. How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. Journal of Personality and Social Psychology 81, 760–773.

Conrey, F.R., Sherman, J.W., Gawronski, B., Hugenberg, K., Groom, C., 2005. Separating multiple processes in implicit social cognition: the quad-model of implicit task performance. Journal of Personality and Social Psychology 89, 469–487.

De Houwer, J., 2001. A structural and process analysis of the Implicit Association Test. Journal of Experimental Social Psychology 37, 443–451.

De Houwer, J., 2003. A structural analysis of indirect measures of attitudes. In: Musch, J., Klauer, K.C. (Eds.), The Psychology of Evaluation: Affective Processes in Cognition and Emotion. Erlbaum, Mahwah, NJ, pp. 219–244.

De Houwer, J., Geldof, T., De Bruycker, E., 2005. The Implicit Association Test as a general measure of similarity. Canadian Journal of Experimental Psychology 59, 228–239.

Govan, C.L., Williams, K.D., 2004. Changing the affective valence of the stimulus items influences the IAT by re-defining the category labels. Journal of Experimental Social Psychology 40, 357–365.

Greenwald, A.G., McGhee, D.E., Schwartz, J.L.K., 1998. Measuring individual differences in implicit cognition: the Implicit Association Test. Journal of Personality and Social Psychology 74, 1464–1480.

Greenwald, A.G., Nosek, B.A., Banaji, M.R., 2003. Understanding and using the Implicit Association Test: I. An improved scoring algorithm. Journal of Personality and Social Psychology 85, 197–216.

Greenwald, A.G., Poehlman, T.A., Uhlmann, E., Banaji, M.R., 2009. Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. Journal of Personality and Social Psychology 97, 17–41.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., Schmitt, M., 2005. A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. Personality and Social Psychology Bulletin 31, 1369–1385.

Kinoshita, S., Peek-O'Leary, M., 2006. Two bases of the compatibility effect in the Implicit Association Test (IAT). Quarterly Journal of Experimental Psychology 59, 2102–2120.

Klauer, K.C., Schmitz, F., Teige-Mocigemba, S., Voss, A., 2010. Understanding the role of executive control in the Implicit Association Test: why flexible people have small IAT effects. Quarterly Journal of Experimental Psychology 63, 595–619.

Klauer, K.C., Voss, A., Schmitz, F., Teige-Mocigemba, S., 2007. Process components of the Implicit Association Test: a diffusion-model analysis. Journal of Personality and Social Psychology 93, 353–368.

Lane, K.A., Banaji, M.R., Nosek, B.A., Greenwald, A.G., 2007. Understanding and using the Implicit Association Test: IV. What we know (so far). In: Wittenbrink, B., Schwarz, N. (Eds.), Implicit Measures of Attitudes. Guilford Press, New York, pp. 59–102.

Meissner, F., Rothermund, K., 2013. Estimating the contributions of associations and recoding in the Implicit Association Test: the ReAL model for the IAT. Journal of Personality and Social Psychology 104, 45–69.

Mierke, J., Klauer, K.C., 2001. Implicit association measurement with the IAT: evidence for effects of executive control processes. Zeitschrift für Experimentelle Psychologie 48, 107–122.

Mierke, J., Klauer, K.C., 2003. Method-specific variance in the Implicit Association Test. Journal of Personality and Social Psychology 85, 1180–1192.

Nier, J.A., 2005. How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach. Group Processes and Intergroup Relations 8, 39–52.

Nisbett, R.E., Wilson, T.D., 1977. Telling more than we can know: verbal reports on mental processes. Psychological Review 84, 231–259.

Nosek, B.A., 2005. Moderators of the relationship between implicit and explicit evaluation. Journal of Experimental Psychology: General 134, 565–584.

Nosek, B.A., Banaji, M.R., Greenwald, A.G., 2002. Harvesting implicit group attitudes and beliefs from a demonstration website. Group Dynamics 6, 101–115.

Nosek, B.A., Greenwald, A.G., Banaji, M.R., 2007. The Implicit Association Test at age 7: a methodological and conceptual review. In: Bargh, J.A. (Ed.), Social Psychology and the Unconscious: The Automaticity of Higher Mental Processes. Psychology Press, New York, pp. 265–292.

Olson, M.A., Fazio, R.H., 2001. Implicit attitude formation through classical conditioning. Psychological Science 12, 413–417.

Olson, M.A., Fazio, R.H., 2004. Reducing the influence of extrapersonal associations on the Implicit Association Test: personalizing the IAT. Journal of Personality and Social Psychology 86, 653–667.

Payne, B.K., Burkley, M.A., Stokes, M.B., 2008. Why do implicit and explicit attitude tests diverge? The role of structural fit. Journal of Personality and Social Psychology 94, 16–31.

Röhner, J., Schröder-Abé, M., Schütz, A., 2011. Exaggeration is harder than understatement, but practice makes perfect! Faking success in the IAT. Experimental Psychology 58, 464–472.

Rothermund, K., Wentura, D., 2001. Figure-ground asymmetries in the Implicit Association Test (IAT). Zeitschrift für Experimentelle Psychologie 48, 94–106.

Rothermund, K., Wentura, D., 2004. Underlying processes in the Implicit Association Test (IAT): dissociating salience from associations. Journal of Experimental Psychology: General 133, 139–165.

Rudolph, A., Schröder-Abé, M., Schütz, A., Gregg, A.P., Sedikides, C., 2008. Through a glass, less darkly? Reassessing convergent and discriminant validity in measures of implicit self-esteem. European Journal of Psychological Assessment 24, 273–281.

Spruyt, A., Hermans, D., De Houwer, J., Vandekerckhove, J., Eelen, P., 2007. On the predictive validity of indirect attitude measures: prediction of consumer choice behavior on the basis of affective priming in the picture – picture naming task. Journal of Experimental Social Psychology 43, 599–610.

Swanson, J.E., Rudman, L.A., Greenwald, A.G., 2001. Using the Implicit Association Test to investigate attitude-behaviour consistency for stigmatised behaviour. Cognition and Emotion 15, 207–230.

Teige, S., Schnabel, K., Banse, R., Asendorpf, J.B., 2004. Assessment of multiple implicit self-concept dimensions using the extrinsic affective simon task (EAST). European Journal of Personality 18, 495–520.

Teige-Mocigemba, S., Klauer, K.C., Sherman, J.W., 2010. A practical guide to Implicit Association Test and related tasks. In: Gawronski, B., Payne, B.K. (Eds.), Handbook of Implicit Social Cognition: Measurement, Theory, and Applications. Guilford Press, New York, pp. 117–139.

Wentura, D., Rothermund, K., 2007. Paradigms we live by. A plea for more basic research on the IAT. In: Wittenbrink, B., Schwarz, N. (Eds.), Implicit Measures of Attitudes. Guilford Press, New York, pp. 195–215.

Wilson, T., Lindsay, S., Schooler, T.Y., 2000. A model of dual attitudes. Psychological Review 107, 101–126.

Relevant Websites

'http://faculty.washington.edu/agg – News from Anthony Greenwald's Research Group Including SPSS Syntaxes.

'https://implicit.harvard.edu/implicit/ – Project Implicit Comprising Several Sample Tests.

'http://www.millisecond.com/ – Software Package Inquisit Comprising Sample Programs.

Implicit Memory

Chad J Marsolek, University of Minnesota, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Implicit memory is said to be operative when people do not deliberately or explicitly recollect information from the past yet they exhibit memory for a previous experience in an indirect manner. It has different properties compared with explicit memory, which is expressed when people consciously and deliberately recollect information from the past. The incremental theoretical progress in explaining implicit memory, especially one form of it labeled repetition priming, is summarized.

Introduction

Human memory is not unitary. This is dramatically evidenced by the phenomenon of medial temporal amnesia following brain damage or neurosurgery. In amnesia, some aspects of memory can be severely impaired while other aspects of memory remain relatively intact, and this has greatly impacted our understanding of human memory. The most famous case was that of the patient H.M. (Milner et al., 1968). Following surgical resection of the medial temporal lobes to relieve severe epilepsy, H.M. exhibited a severe impairment for forming new long-term memories of the sort that can be declared or put into words. Nevertheless, he retained the ability to hold information in short-term memory and to learn new skills of the sort that can be expressed through perceptual-motor tasks.

Testing of additional amnesic patients indicated that, even within the domain of long-term verbal memory, different forms of memory could be dissociated. Patients can be severely impaired in relatively direct memory tests like recall or recognition of previously presented words, yet they remain intact in their abilities to exhibit memory for words in more indirect ways (Warrington and Weiskrantz, 1970). For example, a common indirect memory test is word-stem completion, in which participants are given three-letter word beginnings that can be completed to form many common English words (e.g., TAB___). When asked to think of the first word completion that comes to mind for each stem, even severely amnesic patients produce words that were read previously (e.g., TABLE) at rates higher than that might be attributable to chance.

Subsequent testing of normal participants has uncovered additional ways in which direct and indirect memory performance can be experimentally dissociated. Following deep or meaningful encoding of one list of words (e.g., rating how much each word is liked) and shallow or nonmeaningful encoding of another list of words (e.g., comparing the vowels in one word with the vowels in the preceding word), the direct memory test of free recall indicates significantly greater memory for words that were deeply encoded than for words that were shallowly encoded. However, in the indirect memory test of word-stem completion, no difference is observed between those conditions (Graf et al., 1982). Another dissociation yields the opposite pattern of results. Following visual encoding of one list of words and auditory encoding of another list, the direct memory test of old/new recognition of visual words is little affected by whether the words were previously encoded visually or auditorily. However, the indirect memory test of visually identifying very briefly presented words indicates significantly greater performance for previously visually encoded words than for previously auditorily encoded words (Jacoby and Dallas, 1981). Such findings sparked great interest in what would later be termed as 'implicit memory.'

The terms 'implicit memory' and 'explicit memory' were put forward as descriptive concepts by Graf and Schacter in 1985, but the distinction was not particularly new. Schacter (1987) reviewed effects that fit the label of implicit memory, and some of the reviewed findings extend back to the 1800s. 'Explicit memory' is exhibited when participants consciously and deliberately attempt to recollect information from the past, such as in recall and recognition tests. Implicit memory is exhibited when participants do not deliberately attempt to recollect information from the past yet memory is still expressed in the performance of a task, such as word-stem completion. Roediger (2003) argued that this definition of implicit memory ultimately is too broad to be useful for the field of human memory. Assuming that memory can be defined as "after-effects of stimulation at one time that manifest themselves subsequently at a later time" (Tulving, 1983: p. 7), implicit memory would include all such manifestations that do not happen to involve deliberate recollection. The relatively unbounded nature of the term 'implicit memory' means that it would encompass too vast a set of diverse empirical observations (e.g., perhaps even including changes that occur in the immune system). In line with this concern, the present article focuses on 'long-term repetition priming' as the best-defined and best-studied example of 'implicit memory.' Long-term repetition priming refers to a facilitation in processing a stimulus due to previous processing of that stimulus or a highly similar one, such as in priming of word-stem completions (Tulving and Schacter, 1990). Although other phenomena such as 'short-term masked priming' (e.g., Forster and Davis, 1984) and 'semantic/associative priming' (e.g., Meyer and Schvaneveldt, 1976) may have some properties similar to those of long-term repetition priming, they also happen to have different properties and are not as such included in the focus of this article. During the last quarter century, incremental theoretical progress has been made in understanding long-term repetition priming and its dissociations with explicit memory.

Multiple Memory Systems Theory

The fundamental idea in the multiple memory systems perspective is that repetition priming and explicit memory rely on two different long-term memory stores (e.g., Squire, 1987). In other words, separate representations in separate memory systems exist to support repetition priming for an item and explicit memory for that item. The reason why depth of encoding of a word does not affect repetition priming for it but does affect explicit memory for it (Graf et al., 1982), and the reason why match of perceptual input modality (e.g., auditory and visual) between encoding and test affects repetition priming for an item but not explicit memory for an item (Jacoby and Dallas, 1981), is that separate memory systems support the two kinds of memory effect. The two systems have different representational properties, with one being affected by variables like depth of encoding and the other being affected by variables like perceptual input modality.

An initial suggestion was that repetition priming and explicit memory are supported by the procedural and declarative memory systems, respectively. The procedural memory system supports 'knowing how to do' things, and it involves memory contained in learned skills, cognitive operations, and perceptual abilities. The declarative memory system supports 'knowing that' as opposed to 'knowing how,' and it involves memory for information that can be declared or put into words (Cohen and Squire, 1980). Repetition priming typically involves enhancing perceptual abilities like visual word recognition, thus it may be supported by the procedural memory system. Explicit memory typically involves declaring what words were presented earlier in an experiment, thus it may be supported by the declarative memory system. Another suggestion was that the two effects are supported by the semantic and episodic memory systems, respectively. The semantic memory system stores general world knowledge of the sort that is shared among most people. The episodic memory system stores personal autobiographical memories, such as memories for personal events and experiences (Tulving, 1983). Because repetition priming typically is expressed through general abilities like word reading, it may be supported by the semantic memory system. Because explicit memory typically involves retrieving information from previous experiences in one's personal autobiographical history, it may be supported by the episodic memory system.

Single-System Processing Theory

The fundamental idea in the single-system processing perspective is that both repetition priming and explicit memory (and other forms of memory as well) are supported by a single undifferentiated memory system and the differences between them are that they engage different processes within the unified system (e.g., Roediger, 1990). To account for different patterns of results across repetition priming and explicit memory, appeals are made to the principle of transfer-appropriate processing. When the cognitive operations during encoding overlap with the cognitive operations during test, memory test performance is high. When the cognitive operations do not overlap well, memory test performance is low (Morris et al., 1977). By this theory, the reason why depth of encoding does not affect repetition priming but does affect explicit memory (Graf et al., 1982) is that repetition priming tests typically require only perceptually driven processing and not conceptually driven processing of the sort that matches deep encoding tasks, whereas explicit memory tests require conceptually driven processing of the sort that matches deep encoding tasks. The reason why match of perceptual input modality between encoding and test affects repetition priming but not explicit memory (Jacoby and Dallas, 1981) is that repetition priming tests typically require perceptually driven processes that may be modality-specific, whereas explicit memory tests require conceptually driven processing that is not modality-specific.

An Initial Resolution

A bit of a theoretical quandary ensued. Both the multiple memory systems theory and the single-system processing theory could account for the same functional dissociations between repetition priming and explicit memory. The main advantage for the multiple systems theory was that it more plausibly accounted for why amnesic patients are severely impaired in explicit memory but not in repetition priming. One system can be impaired independently of another. In addition, an appeal by single-system processing theorists to the idea that conceptually driven processing is selectively impaired in such patients was not convincing. Amnesic patients are capable of effective conceptually driven processing in situations other than testing of explicit memory. On the other hand, the main advantage for the single-system processing theory was that it avoided the problem that additional experimentation could inevitably lead to an 'explosion' of theorized memory systems. For example, when the priming test requires conceptually driven processing as opposed to typical priming tests that require only perceptually driven processing, Blaxton (1989) found that depth of encoding affects repetition priming, and Srinivas and Roediger (1990) found that perceptual-modality shifts between encoding and test do not affect repetition priming. Those results are opposite to those found when the priming test requires only perceptually driven processing (Graf et al., 1982; Jacoby and Dallas, 1981). Should such a pattern of results necessitate postulating yet another division of new memory systems? If so, an explosion of systems may cause theoretical chaos.

An initial resolution to the debate was developed by Schacter (1994). Functional dissociations between two memory effects are not sufficient to posit separate memory systems. Instead, additional dissociations from other methods and levels of analysis must provide independent but converging evidence for postulating separate memory systems (Schacter, 1992). From this perspective, Schacter (1994) hypothesized that most repetition priming effects are supported by the perceptual representation system (PRS). The PRS is a small collection of subsystems that are domain-specific and represent perceptual form information but not meaning-based or associative information. The PRS is functionally and neurally dissociable from the systems that underlie

explicit memory. Evidence for this distinction came from left and right cerebral hemispheric differences (Marsolek et al., 1992; Marsolek et al., 1996), complementary impaired and intact memory abilities following damage to different areas of the brain (e.g., Gabrieli et al., 1995), and neural activity in different areas of the brain as measured with functional neuroimaging (e.g., Schacter et al., 1996; Squire et al., 1992). The general understanding that repetition priming takes place within perceptual systems led several researchers to treat repetition priming (a memory effect) as a useful tool for testing the properties of perceptual systems (e.g., Biederman and Cooper, 1991; Bowers, 1996; Marsolek, 1999; Marsolek et al., 1992, 1996).

Bias Theory of Repetition Priming

An important aspect of the PRS theory is that repetition priming occurs when previous visual encoding of a word or object enhances the subsequent visual processing of that word or object at test. Previous visual encoding is purported to enhance the ability "to extract visual information from the test cue" (Schacter, 1994: p. 237). In other words, the previous perceptual encoding increases subsequent sensitivity or the ability to discriminate the test item from other possible items. An alternative to this perspective came from Ratcliff and McKoon (1996, 1997). They argued that repetition priming is not due to enhanced processing or increased sensitivity, but instead it is due to an increase in bias to make a response that corresponds to a previously processed item.

To support the bias theory, Ratcliff and McKoon (1996, 1997) used test tasks in which a word (e.g., 'lied') or a drawing of an object (e.g., a hot-air balloon) is flashed very briefly, and then two visually similar alternative words (e.g., 'lied' and 'died') or two visually similar alternative objects (e.g., a hot-air balloon or a typical light bulb) are presented. In a forced-choice test, participants must choose which of the two words or two objects had just been flashed. When the flashed item had been processed in a preceding task (e.g., 'lied' or hot-air balloon), identification of that item as the one that was flashed increased compared against a condition in which neither alternative had been processed in a preceding task. An additional finding was important for implicating bias as an explanation. When the nonflashed alternative item had been processed in a preceding task (e.g., 'died' or light bulb), identification of the flashed item (e.g., 'lied' or hot-air balloon) decreased compared against a condition in which neither alternative had been processed in a preceding task. In other words, there was a benefit in identifying the flashed item when it had been processed previously, and there was a cost in identifying the flashed item when the nonflashed item had been processed previously.

Ratcliff and McKoon (1996, 1997) developed models of word and object identification to account for these results in a concrete manner, one in which bias is implicated and not sensitivity. In these models, previous processing of an item does not enhance perceptual processing of that item at test. In particular, previous processing of an item does not increase the amount of diagnostic visual information that is extracted from the primed item during the forced-choice test. Instead, previous processing of an item increases the subsequent bias to interpret ambiguous perceptual information in a way that is consistent with the processing in a preceding task. In other words, previous processing of an item only increases a bias to make a subsequent choice that corresponds to that previous experience; it does not increase the subsequent ability to perceptually process or distinguish that item per se. By this explanation, when neither of the two alternatives in the forced-choice identification task had been processed in a preceding task, there is no difference in bias to report either item as the flashed item. However, when one of the alternatives had been processed in a preceding task, participants are biased to report that item as their response, whether or not it is the correct response.

Another finding in the initial research on bias effects turned out to be important (Ratcliff and McKoon, 1997). When both items in the forced-choice identification task (e.g., both 'lied' and 'died') had been processed in a preceding task, performance is equivalent to performance in the condition in which neither alternative had been primed. This was taken as evidence that in the both-primed condition the bias to respond with one alternative cancels out the bias to respond with the other alternative. However, subsequent research indicated that pattern of results is found only when high-frequency words are used. When low-frequency words are used, performance in the both-primed condition is greater than performance in the neither-primed condition (Bowers, 1999; Wagenmakers et al., 2000; Zeelenberg et al., 2002). This is a pattern of results that does not fit the original bias theory, because it indicates an increase in discrimination ability for both words in the both-primed condition that cannot be due to bias (the increased bias for each word should cancel each other out). Ratcliff and McKoon (2000) modified their initial model to account for the results obtained when low-frequency stimuli are used. The modifications enable both increased sensitivity and bias effects to play roles in repetition priming.

Cognitive Neuroscience Theory: Antipriming Accompanies Priming

Contemporary theory in cognitive neuroscience is taking advantage of the positive aspects of the foregoing theories – as well as integrating behavioral, neural, and neurocomputational evidence – to account for repetition priming (Marsolek, 2003). A notion adopted from the PRS theory is the importance of integrating evidence from multiple methods and multiple levels of analysis, to avoid theoretical ambiguity when considering only behavioral experimentation. A notion adopted from bias theory is that concrete models of the changes that take place in perceptual identification systems are crucial for understanding repetition priming (see also many of the chapters in Bowers and Marsolek, 2003). The general notion that previous experiences can have subsequent costs in addition to subsequent benefits also is adopted from the bias theory. Moreover, additional integrative evidence from the neural implementation of perceptual systems in neocortex and neurocomputational modeling of such systems enable a theory of repetition priming that can explain both

what it is and why it is (Marsolek, 2003). An important prediction from this theory is that antipriming is a phenomenon that should accompany observations of repetition priming.

The theory is strongly constrained by neuroscientific and neurocomputational evidence (Marsolek, 2008). The representations of different visual objects in neocortex are distributed and partially overlapping (Haxby et al., 2001; Tanaka, 2003), and they undergo learning-related changes after each time they are accessed and used to identify objects (Kobatake et al., 1998; Sigala and Logothetis, 2002). Neurocomputational models that simulate these properties in biologically plausible ways indicate that antipriming should accompany repetition priming (Marsolek et al., 2006). Antipriming occurs when a visual object representation is strengthened after it is used to identify an object (e.g., a piano), and the changes that take place also entail weakening of some other representations with which the strengthened representation overlaps (e.g., the representation for desk). Strengthening one visual object representation not only enhances the subsequent ability to identify that object (e.g., repetition priming for piano), but it also impairs the subsequent ability to identify some other objects (e.g., antipriming for desk), relative to a baseline condition in which objects are neither primed nor antiprimed. This pattern of results has been observed in young adults, amnesic patients, and their matched control participants (Marsolek et al., 2006).

By this theory, the integration of evidence from neural, functional, and computational levels of analysis (Marr, 1982) helps to explain why priming and antipriming take place (Marsolek, 2008). They both reflect a dynamic interplay of maintenance relearning of overlapping representations in visual object identification systems. Strengthening one representation entails weakening some others. When the previously weakened representations are used to successfully identify antiprimed objects, the strengthening of those objects' representations is of a greater than normal magnitude, helping to maintain them within a system of overlapping representations. This is evidenced by the finding that greater neural activity is elicited by antiprimed objects than by either primed objects or objects that are neither primed nor antiprimed. In addition, the timing of this activation difference is after objects are identified rather than during identification (Marsolek et al., 2010).

A Reconceptualization of Neuroimaging Effects

These recent results help to reconceptualize the neural repetition effect. This ubiquitous effect in neuroimaging is observed when a repeated stimulus elicits lower magnitude neural activity than a nonrepeated stimulus. The standard interpretation is that this reflects a decrease in activity due to habituation, sharpening of activity, or faster activations for the repeated stimuli (Grill-Spector et al., 2006). The reconceptualization is that nonrepeated stimuli are actually antiprimed stimuli, and antiprimed stimuli elicit an increase in activity that reflects a greater than normal magnitude of maintenance relearning (Marsolek et al., 2010).

The reconceptualization of the neural repetition effect is important. Traditionally, it has been understood as a neural correlate of the behavioral effect of repetition priming (Grill-Spector et al., 2006). By this understanding, the magnitude of the neural repetition effect for a participant should correlate with the magnitude of the behavioral repetition priming effect for that participant. However, in perceptual cortical areas (but not prefrontal areas), such correlations are not observed (Maccotta and Buckner, 2004; McMahon and Olson, 2007; Schacter et al., 2007; see also Dobbins et al., 2004). The reason may be that the neural repetition effect in perceptual areas is a marker of the magnitude of maintenance relearning after identifying an object rather than a marker of the improved processing of an object during identification (Marsolek et al., 2010). Future research is needed to more strongly test this possibility.

See also: Amnesia: General; Cognitive Neuroscience; Declarative Memory, Neural Basis of; Episodic Memory; Learning and Memory, Neural Basis of; Non Declarative (Procedural) Memory; Semantic Memory; Visual Memory, Psychology of; Visual Perception, Neural Basis of.

Bibliography

Biederman, I., Cooper, E.E., 1991. Priming contour-deleted images: evidence for intermediate representations in visual object recognition. Cognitive Psychology 23, 393–419.

Blaxton, T.A., 1989. Investigating dissociations among memory measures: support for a transfer appropriate processing framework. Journal of Experimental Psychology: Learning, Memory, and Cognition 15, 657–668.

Bowers, J.S., 1996. Different perceptual codes support priming for words and pseudowords: was Morton right all along? Journal of Experimental Psychology: Learning, Memory, and Cognition 22, 1336–1353.

Bowers, J.S., 1999. Priming is not all bias: commentary on Ratcliff and McKoon (1997). Psychological Review 106, 582–596.

Bowers, J.S., Marsolek, C.J. (Eds.), 2003. Rethinking Implicit Memory. Oxford University Press, New York.

Cohen, N.J., Squire, L.R., 1980. Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how versus knowing that. Science 210, 207–210.

Dobbins, I.G., Schnyer, D.M., Verfaellie, M., Schacter, D.L., 2004. Cortical activity reductions during repetition priming can result from rapid response learning. Nature 428, 316–319.

Forster, K.I., Davis, C., 1984. Repetition priming and frequency attenuation in lexical access. Journal of Experimental Psychology: Learning, Memory, and Cognition 10, 680–689.

Gabrieli, J.D.E., Fleischman, D.A., Keane, M.M., Reminger, S.L., Morrell, F., 1995. A double dissociation between memory systems underlying explicit and implicit memory in the human brain. Psychological Science 6, 76–82.

Graf, P., Mandler, G., Haden, P., 1982. Simulating amnesic symptoms in normal subjects. Science 218, 1243–1244.

Graf, P., Schacter, D.L., 1985. Implicit and explicit memory for new associations in normal and amnesic subjects. Journal of Experimental Psychology: Learning, Memory, and Cognition 11, 501–518.

Grill-Spector, K., Henson, R., Martin, A., 2006. Repetition and the brain: neural models of stimulus-specific effects. Trends Cognitive Science 10, 14–23.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.

Jacoby, L.L., Dallas, M., 1981. On the relationship between autobiographical memory and perceptual learning. Journal of Experimental Psychology: General 110, 306–340.

Kobatake, E., Wang, G., Tanaka, K., 1998. Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. Journal of Neurophysiology 80, 324–330.

Maccotta, L., Buckner, R.L., 2004. Evidence for neural effects of repetition that directly correlate with behavioral priming. Journal of Cognitive Neuroscience 16, 1625–1632.

Marr, D., 1982. Vision. W.H. Freeman, San Francisco, CA.

Marsolek, C.J., 1999. Dissociable neural subsystems underlie abstract and specific object recognition. Psychological Science 10, 111–118.

Marsolek, C.J., 2003. What is priming and why? In: Bowers, J.S., Marsolek, C.J. (Eds.), Rethinking Implicit Memory. Oxford University Press, Oxford, pp. 41–64.

Marsolek, C.J., 2008. What antipriming reveals about priming. Trends in Cognitive Sciences 12, 176–181.

Marsolek, C.J., Deason, R.G., Ketz, N.A., Ramanathan, P., Bernat, E.M., Steele, V.R., Patrick, C.J., Verfaellie, M., Schnyer, D.M., 2010. Identifying objects impairs knowledge of other objects: a relearning explanation for the neural repetition effect. NeuroImage 49, 1919–1932. (Editors' Choice Award winner (Cognitive Neuroscience Section), NeuroImage, 2010.)

Marsolek, C.J., Kosslyn, S.M., Squire, L.R., 1992. Form-specific visual priming in the right cerebral hemisphere. Journal of Experimental Psychology: Learning, Memory, and Cognition 18, 492–508.

Marsolek, C.J., Schacter, D.L., Nicholas, C.D., 1996. Form-specific visual priming for new associations in the right cerebral hemisphere. Memory & Cognition 24, 539–556.

Marsolek, C.J., Schnyer, D.M., Deason, R.G., Ritchey, M., Verfaellie, M., 2006. Visual antipriming: evidence for ongoing adjustments of superimposed visual object representations. Cognitive, Affective, and Behavioral Neuroscience 6, 163–174.

McMahon, D.B.T., Olson, C.R., 2007. Repetition suppression in monkey inferotemporal cortex: relation to behavioral priming. Journal of Neurophysiology 97, 3532–3542.

Meyer, D.E., Schvaneveldt, R.W., 1976. Meaning, memory structure, and mental processes. Science 192, 27–33.

Milner, B., Corkin, S., Teuber, H.L., 1968. Further analysis of the hippocampal amnesic syndrome: 14 year follow-up study of H. M. Neuropsychologia 6, 215–234.

Morris, C.D., Bransford, J.D., Franks, J.J., 1977. Levels of processing versus transfer appropriate processing. Journal of Verbal Learning and Verbal Behavior 16, 519–533.

Ratcliff, R., McKoon, G., 1996. Bias effects in implicit memory tasks. Journal of Experimental Psychology: General 125, 403–421.

Ratcliff, R., McKoon, G., 1997. A counter model for implicit priming in perceptual word identification. Psychological Review 104, 319–343.

Ratcliff, R., McKoon, G., 2000. Modeling the effects of repetition and word frequency in perceptual identification. Psychonomic Bulletin & Review 7, 713–717.

Roediger, H.L., 1990. Implicit memory: retention without remembering. American Psychologist 45, 1043–1056.

Roediger, H.L., 2003. Reconsidering implicit memory. In: Bowers, J.S., Marsolek, C.J. (Eds.), Rethinking Implicit Memory. Oxford University Press, Oxford, pp. 3–18.

Schacter, D.L., 1987. Implicit memory: history and current status. Journal of Experimental Psychology: Learning, Memory, and Cognition 13, 501–518.

Schacter, D.L., 1992. Understanding implicit memory: a cognitive neuroscience approach. American Psychologist 47, 559–569.

Schacter, D.L., 1994. Priming and multiple memory systems: perceptual mechanisms of implicit memory. In: Schacter, D.L., Tulving, E. (Eds.), Memory Systems 1994. MIT Press, Cambridge, MA, pp. 233–268.

Schacter, D.L., Alpert, N., Savage, C., Rauch, S., Alpert, M.S., 1996. Conscious recollection and the human hippocampal formation: evidence from positron emission tomography. Proceedings of the National Academy of Sciences USA 93, 321–325.

Schacter, D.L., Wig, G.S., Stevens, W.D., 2007. Reductions in cortical activity during priming. Current Opinion in Neurobiology 17, 171–176.

Sigala, N., Logothetis, N.K., 2002. Visual categorization shapes feature selectivity in the primate temporal cortex. Nature 415, 318–320.

Squire, L.R., 1987. Memory and Brain. Oxford University Press, New York.

Squire, L.R., Ojemann, J.G., Miezin, F.M., Petersen, S.E., Videen, T.O., Raichle, M.E., 1992. Activation of the hippocampus in humans: a functional anatomical study of memory. Proceedings of the National Academy of Sciences USA 89, 1837–1841.

Srinivas, K., Roediger, H.L., 1990. Testing the nature of two implicit tests: dissociations between conceptually driven and data-driven processes. Journal of Memory and Language 29, 389–412.

Tanaka, K., 2003. Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. Cerebral Cortex 13, 90–99.

Tulving, E., 1983. Elements of Episodic Memory. Clarendon Press, Oxford.

Tulving, E., Schacter, D.L., 1990. Priming and human memory systems. Science 247, 301–305.

Wagenmakers, E.-J.M., Zeelenberg, R., Raaijmakers, J.G.W., 2000. Testing the counter model for perceptual identification: effects of repetition priming and word frequency. Psychonomic Bulletin & Review 7, 662–667.

Warrington, E.K., Weiskrantz, L., 1970. Amnesia: consolidation or retrieval? Nature 228, 628–630.

Zeelenberg, R., Wagenmakers, E.-J.M., Raaijmakers, J.G.W., 2002. Priming in implicit memory tasks: prior study causes enhanced discriminability, not only bias. Journal of Experimental Psychology: General 131, 38–47.

Implicit Social Cognition

Adam Hahn, University of Cologne (Universität zu Köln), Köln, Germany
Bertram Gawronski, University of Texas at Austin, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes the field of implicit social cognition. Mental association is introduced as a core concept that serves as the basis for the definition of several psychological constructs and the development of nonreactive, computerized measurement instruments designed to capture the automatic activation of mental associations. We further discuss the meaning of the term implicit and review research and theorizing on the relation between implicit and explicit measures; prediction of meaningful outcomes; and the formation, change, and contextualization of mental associations. The article concludes with a brief review of mathematical modeling in implicit social cognition and its impact on applied areas.

The term *implicit social cognition* refers to research in social psychology and related disciplines that uses a particular class of nonreactive, computerized measurement instruments to assess thoughts and feelings without directly asking participants to report on them (for a comprehensive review, see Gawronski and Payne, 2010). A central characteristic that distinguishes these tools from other kinds of unobtrusive measures is that they reduce participants' ability to strategically control their responses. These measures are often referred to as *implicit measures*, whereas traditional self-report measures are described as *explicit measures*.

Mental Association as a Core Concept

Based on the two central features of implicit measures, it is often claimed that they (1) overcome the well-known problems of socially desirable responding and (2) capture thoughts and feelings that are outside of conscious awareness, and thus inaccessible to self-report. Over the last few years, evidence conflicting with these assumptions has led many researchers to prefer more agnostic interpretations in terms of mental associations. Such interpretations are based on the idea that many important constructs of social cognition can be defined as associations between concepts in memory (Greenwald et al., 2002). For example, the construct of *attitude* has been defined as association between an object and a particular evaluation. Based on this definition, *prejudice* can be defined as evaluative associations involving a social group, and *self-esteem* as evaluative association involving the self. Similarly, *stereotypes* can be defined as semantic associations between a social group and stereotypical attributes, whereas the *self-concept* refers to semantic associations between the self and its attributes. In general, the concept of mental association is applicable to any kind of target objects (e.g., consumer products, political candidates) and their evaluative and semantic attributes. Although some theorists have proposed alternative frameworks that reject the notion of mental associations (Hughes et al., 2011), associative theorizing has been a driving force in the development of implicit measures and still serves as a conceptual core in research on implicit social cognition.

Measurement Procedures

Implicit measures are based on the idea that activation of a mental concept can spread to other associated concepts in memory. To the extent that the associative link between two concepts is sufficiently strong, spread of activation is assumed to occur automatically (i.e., unintentionally, unconsciously, efficiently, and uncontrollably). Implicit measures make use of such automatic processes by assessing the effect of stimuli or stimulus features on participants' performance (e.g., response times, error rates) in responding to other stimuli or stimulus features (for a review, see Gawronski and De Houwer, 2014).

Sequential Priming Tasks

The first type of implicit measures is based on the notion of sequential priming in cognitive psychology (for a review, see Wentura and Degner, 2010). In a typical sequential priming task, participants are briefly presented with a prime stimulus (e.g., a picture of a person) which is followed by a target stimulus. Participants' task is to categorize the target stimulus as quickly as possible into two alternative categories. Depending on the type of measure, participants may be asked to (1) classify the target stimulus as good or bad (i.e., *evaluative decision task*), (2) classify the target stimulus in terms of its semantic meaning (i.e., *semantic decision task*), or (3) decide whether the target stimulus is a meaningful word or a meaningless letter string (i.e., *lexical decision task*). The basic idea underlying sequential priming tasks is that responses to the target stimulus are facilitated (i.e., lower response times, higher accuracy) when it is preceded by a prime stimulus that is mentally associated with either the target stimulus itself (e.g., sequential priming with lexical decision tasks) or the feature that is relevant for categorization of the target stimulus (e.g., sequential priming with evaluative or semantic decision tasks). For example, using a variant of sequential priming involving evaluative target decisions, a person with negative associations toward Black people should be faster in responding to negative words and slower in responding to positive words when the target words are preceded by Black faces compared to neutral baseline primes (e.g., Fazio et al., 1995). Similarly, using a variant of sequential priming involving semantic target

 http://dx.doi.org/10.1016/B978-0-08-097086-8.24066-X

decisions, a person with gender–stereotypical associations may be faster in responding to male and female pronouns after being presented with stereotype-congruent prime words (e.g., doctor-him, nurse-her) compared to stereotype-incongruent prime words (e.g., doctor-her, nurse-him) (e.g., Banaji and Hardin, 1996). Finally, using a variant of sequential priming involving lexical target decisions, a person may be faster in identifying certain adjectives (e.g., competent) as meaningful words after being exposed to a political candidate to the extent that this person has strong associations between the candidate and the attributes described by the target words (e.g., Wittenbrink et al., 1997).

Implicit Association Test

The implicit association test (IAT) consists of two binary categorization tasks that require participants to sort stimuli representing two opposing target concepts (e.g., Black and White faces) and stimuli representing two opposing attribute concepts (e.g., positive and negative words) as quickly as possible by pressing one of two response keys (Greenwald et al., 1998). The central assumption underlying the IAT is that fast and accurate responses will be facilitated when mentally associated concepts are mapped onto the same response key, and impaired when mentally associated concepts are mapped onto different keys. For example, to measure preferences for Whites over Blacks, participants might be asked to sort pictures of White and Black faces and words depicting pleasant and unpleasant objects. In one block of the task, participants may be asked to respond to White faces and positive words with one key and to Black faces and negative words with the other key. In another block of the task, participants may be asked to respond to Black faces and positive words with one key and to Whites faces and negative words with the other key. The relative difference in participants' performance (i.e., speed, accuracy) in the two blocks is typically interpreted as an index of participants' preference for Whites over Blacks, or vice versa. IAT scores are inherently relative (i.e., relative preference for one group over another), and the blocked presentation of association-congruent and association-incongruent trials has been linked to various sources of systematic measurement error. To overcome these limitations, researchers have developed a number of modified IAT variants that are amenable for assessing associations of a single target concept or a single attribute, and variants that avoid blocked presentations of association-congruent and association-incongruent trials (for a review, see Teige-Mocigemba et al., 2010).

Go/No-Go Association Task

In the go/no-go association task (GNAT; Nosek and Banaji, 2001), participants are asked to show a *go* response to different kinds of target stimuli (e.g., by pressing the space bar) and a *no-go* response to distracter stimuli (i.e., no button press). In one block of the task, the targets include stimuli related to the target concept of interest (e.g., Black faces) and stimuli related to one pole of a given attribute dimension (e.g., positive words); the distracters typically include stimuli related to the other pole of the attribute dimension (e.g., negative words). In a second block, the classification of the particular attribute poles as targets and distracters is reversed (e.g., *go* for Black faces and negative words, and *no-go* for positive words). GNAT trials typically include a response deadline such that participants are asked to show a *go* response to the targets before the expiration of that deadline. Error rates are analyzed by means of signal detection theory, such that differences in sensitivity scores (d′) between the two pairings of *go* trials (e.g., Black-positive vs. Black-negative) are interpreted as an index of associations between the target concept of interest and the respective attributes.

Extrinsic Affective Simon Task

In the Extrinsic Affective Simon task (EAST; De Houwer, 2003), participants are presented with colored words representing a target object and white words representing associated attributes. Participants are instructed to categorize the words in terms of their valence when they are shown in white, and to categorize them in terms of their color when they are colored. For example, in an EAST designed to measure evaluative associations of different kinds of beverages, participants may be presented with positive and negative words in white and with names of beverages that are presented in yellow on some trials and in blue on others. Participants' task is to press one key when they see either a white word of negative valence or a word printed in blue and to press another key when they see either a white word of positive valence or a word printed in yellow. To the extent that participants show faster (or more accurate) responses to a colored word (e.g., Coke) when the required response to this word is combined with a positive as compared to a negative response, it is inferred that participants have positive associations with the object depicted by the colored word (or vice versa). Although the EAST was originally designed as a measure of evaluative associations, a number of studies have demonstrated its applicability to semantic associations (e.g., Teige et al., 2004).

Affect Misattribution Procedure

The affect misattribution procedure (AMP; Payne et al., 2005) adopts the basic logic of sequential priming, in that it measures the effects of prime stimuli on responses to subsequently presented targets. However, the AMP differs from traditional variants of sequential priming by (1) using target stimuli that are ambiguous with regard to the target response and (2) relying on participants' actual judgments of the targets instead of speed or accuracy data. The basic idea underlying the AMP is that the prime stimuli activate thoughts and feelings that may be misattributed to the ambiguous target stimuli (Gawronski and Ye, 2014). For example, in an AMP to measure evaluative associations of racial groups, participants may be presented with Black and White faces as primes and neutral Chinese ideographs as targets. Participants' task is to indicate if they consider the Chinese ideograph as visually more pleasant or visually less pleasant than the average Chinese ideograph. The modal finding is that participants show more favorable responses to the Chinese ideographs when they were primed with a pleasant stimulus than when they are primed with an unpleasant stimulus. Such priming effects tend to emerge even

when participants receive detailed information about the operation of the task and are explicitly instructed to avoid any potential influence of the primes (Payne et al., 2005). Although the AMP has originally been designed to measure evaluative associations, modified variants of the task have been shown to be amenable for the measurement of semantic associations (e.g., Sava et al., 2012).

What Is 'Implicit' about Implicit Social Cognition?

A common source of confusion in implicit social cognition is the meaning of the term *implicit*. Whereas some researchers use the term to describe features of measurement procedures, others use it to describe features of psychological constructs (e.g., implicit attitudes). The first interpretation emerged from the goal of overcoming the effects of socially desirable responding on self-report measures (e.g., Fazio et al., 1995). Research in this tradition was inspired by earlier work in cognitive psychology investigating the role of voluntary and involuntary processes in attention (see Payne and Gawronski, 2010). The second interpretation emerged from the goal of capturing unconscious thoughts and feelings that are inaccessible to self-report (e.g., Greenwald and Banaji, 1995). Research in this tradition was inspired by earlier work on implicit memory, suggesting that prior experiences can influence performance on psychological tasks even when participants are unable to explicitly recall the relevant experience (see Payne and Gawronski, 2010).

Drawing on the notion of implicit memory, Greenwald and Banaji (1995) defined *implicit cognition* as "introspectively unidentified (or inaccurately identified) trace of past experience that mediates responses" (p. 5). Although this definition was meant to imply unawareness of the sources of mental contents, it has often been interpreted to imply unawareness of the mental contents themselves (e.g., unawareness of the source of an attitude vs. unawareness of the attitude itself). However, the latter interpretation conflicts with a considerable body of evidence, suggesting that the psychological constructs captured by nonreactive, computerized measures are consciously accessible and thus not unconscious (e.g., Hahn et al., 2014).

More recently, De Houwer et al. (2009) proposed an alternative conceptualization to overcome the common confusion regarding the meaning of the term implicit. According to this conceptualization, the terms *implicit* and *explicit* describe the processes by which a psychological attribute (e.g., attitude) influences measurement outcomes (which may be described as *measures* in the sense of measurement scores), rather than the procedure itself or the underlying attribute. Specifically, measurement outcomes may be described as *implicit* if the to-be-measured attribute influences participants' responses in an automatic fashion (e.g., unintentionally, uncontrollably). Conversely, measurement outcomes may be described as *explicit* if the to-be-measured attribute influences participants' responses in a controlled fashion (i.e., intentionally, controllably). Measurement procedures, on the other hand, may be described as *direct* if their measurement outcomes are based on participants' self-assessment of the to-be-measured attribute (e.g., when participants' racial attitudes are inferred from their self-reported liking of Black people). Conversely, measurement procedures may be described as *indirect* if their outcomes are not based on a self-assessment (e.g., when participants' racial attitudes are inferred from their reaction times to positive and negative words after being primed with Black faces) or when it is based on a self-assessment of attributes other than the to-be-measured attribute (e.g., when participants' racial attitudes are inferred from their self-reported liking of a neutral object after being primed with Black faces).

Relations between Implicit and Explicit Measures

In a meta-analysis on the IAT, Hofmann et al. (2005) found an average correlation of 0.24 between IAT scores and self-report measures. Similar findings have been reported by Cameron et al. (2012) for different procedures based on sequential priming. However, correlations in both meta-analyses varied considerably as a function of content domains as well as procedural and method-related factors. In general, implicit–explicit correlations tend to be higher when participants rely on intuitive, affective bases (e.g., feelings elicited by an object) as opposed to deliberate, cognitive bases (e.g., reasons for liking or disliking an object) when reporting an explicit judgment. Implicit and explicit measures also show higher correlations when they are similar in terms of their dimensionality (e.g., absolute evaluations of Blacks vs. relative preference for Whites over Blacks) and content (e.g., evaluations of affirmative action policies vs. evaluations of Black and White faces).

A variety of theories have been developed to explain variations in the relation between implicit and explicit measures. According to the MODE (*M*otivation and *O*pportunity and *DE*terminants) model (for a review, see Fazio, 2007), the relation between implicit and explicit measures depends on the motivation and opportunity to engage in effortful processing of judgment-relevant information. To the extent that people are motivated and have the opportunity to engage in effortful processing when making an explicit judgment, they are assumed to scrutinize the available information about the target object. In this case, correlations between implicit and explicit measures should be low. If, however, either the motivation or the opportunity to engage in effortful processing is low, people are assumed to rely on their automatically activated associations when making an explicit judgment. In this case, correlations between implicit and explicit measures should be high. These hypotheses are consistent with research showing that correlations between implicit and explicit measures of racial attitudes tend to be higher for participants low in motivation to control prejudiced reactions compared to participants high in motivation to control (e.g., Fazio et al., 1995). Other research supporting the MODE model includes studies showing that correlations between implicit and explicit measures tend to be higher when participants are asked to provide explicit judgments under time pressure than when they have unlimited time (e.g., Ranganath et al., 2008).

Another theory that explains varying correlations between implicit and explicit measures is the associative-propositional evaluation (APE) model (Gawronski and Bodenhausen, 2006). The APE model conceptualizes implicit and explicit evaluations as the behavioral outcomes of two qualitatively distinct – yet mutually interacting – mental processes. Whereas

implicit evaluations are assumed to be the outcome of associative processes, explicit evaluations are conceptualized as the outcome of propositional processes. Associative processes are further defined as the *activation* of mental associations on the basis of observed co-occurrences between stimuli in the environment and the similarity between features of environmental stimuli and existing representations. Propositional processes are defined as the *validation* of the information implied by activated associations on the basis of their consistency with other momentarily considered information. An important difference between the two processes is that associations may be activated regardless of whether a person considers these associations as a valid basis for an explicit judgment. According to the APE model, the central determinant of perceived validity is the consistency of the evaluation implied by activated associations with all other momentarily considered information. To the extent that this evaluation is consistent with other momentarily considered information, it may be regarded as valid and therefore used as a basis for an explicit judgment. If, however, the evaluation implied by activated associations is inconsistent with other momentarily considered information, the implied inconsistency may lead to a rejection of the implied evaluation as invalid. Whereas in the former case implicit and explicit measures are assumed to show corresponding evaluations, the two kinds of measures are assumed to show diverging evaluations in the latter case. Although the APE model shares many assumptions with the MODE model, an important aspect of the APE model is that effortful processing may not necessarily reduce the relation between implicit and explicit measures when the additionally considered information is consistent with the evaluation implied by activated associations. This assumption is supported by research showing that high motivation to control prejudiced reactions does not reduce the relation between implicit and explicit measures of racial prejudice when consistency can be restored in a way that does not imply a rejection of activated associations (e.g., by denying discrimination of the target group; see Gawronski et al., 2008).

Prediction

Implicit measures are often used as tools to predict meaningful psychological outcomes (e.g., behavior, decisions), and recent meta-analyses tend to support their predictive validity (e.g., Cameron et al., 2012; Greenwald et al., 2009). However, to justify the use of such resource-intensive tasks, many researchers have become concerned with unique aspects of behavior that are difficult to predict with explicit measures. According to Perugini et al. (2010), implicit measures may contribute to the prediction of outcomes over and above explicit measures in various ways, including (1) additive patterns, (2) double-dissociation patterns, (3) moderation patterns, and (4) interactive patterns.

Additive patterns involve cases in which implicit and explicit measures of the same construct jointly predict a particular outcome. Such cases tend to emerge when implicit measures are able to capture particular aspects of the outcome that are not captured by the explicit measure. Thus, additive patterns are best described in terms of explained variance, such that implicit measures may increase the total proportion of explained variance in the outcome measure compared to the variance that is explained by the explicit measure alone.

Although additive patterns have been obtained in a few studies, a more common finding is a double-dissociation in the prediction of different kinds of outcomes. In line with the assumptions of various dual-process theories (e.g., Fazio, 2007), implicit measures have been shown to outperform explicit measures in the prediction of spontaneous behavior, whereas explicit measures tend to outperform implicit measures in the prediction of deliberate behavior. For example, nonverbal behavior in interracial interactions typically shows stronger relations to implicit when compared to explicit measures, whereas verbal behavior has been shown to reveal stronger relations to explicit when compared to implicit measures (e.g., Dovidio et al., 2002).

Other research has focused on various moderators that determine whether a given outcome is predicted by implicit or explicit measures. Such moderating patterns have been shown for various situational factors and their individual difference counterparts (for a review, see Perugini et al., 2010). For example, implicit measures have shown stronger relations to eating behavior than explicit measures under conditions of cognitive depletion, whereas explicit measures showed stronger relations to eating behavior under control conditions (Hofmann et al., 2007). Correspondingly, implicit measures have been shown to outperform explicit measures in the prediction of eating behavior for people low in working memory capacity, whereas explicit measures outperformed implicit measures for people high in working memory capacity (Hofmann et al., 2008). Such findings are consistent with dual-process theories assuming that motivation and opportunity to engage in effortful processing are central moderators of the predictive validity of implicit and explicit measures (e.g., Fazio, 2007).

Deviating from approaches in which implicit and explicit measures are seen as competitors in the prediction of behavior, several studies have investigated interactive relations between the two kinds of measures. The general assumption underlying these studies is that discrepancies between implicit and explicit measures are indicative of an unpleasant psychological state that people aim to reduce. In line with this assumption, people showing large discrepancies between implicit and explicit measures of a particular psychological attribute (e.g., attitude, self-concept) have been shown to process discrepancy-related information more extensively than people with small discrepancies (e.g., Briñol et al., 2006). In a similar vein, combinations of high self-esteem on explicit measures and low self-esteem on implicit measures have been shown to predict various kinds of defensive behaviors (e.g., Jordan et al., 2003).

Formation, Change, and Context Effects

Several early theories of implicit social cognition assumed that implicit measures capture highly overlearned associations that have their roots in long-term socialization experiences (e.g., Wilson et al., 2000). Consistent with this assumption, several studies have shown that implicit measures are meaningfully related to early childhood experiences. However, although

long-term socialization experiences can be an important source of the associations captured by implicit measures, such experiences do not seem to be necessary. The latter conclusion is supported by research showing that recently formed associations that are based on minimal experiences can have a strong impact on implicit measures (e.g., Gregg et al., 2006). Resonating with the dual-process distinction of the APE model (Gawronski and Bodenhausen, 2006), these experiences may involve either descriptive information about a target object (i.e., propositional learning) or incidental pairings of a target object with other stimuli (i.e., associative learning). In the domain of attitudes, propositional learning plays a central role in research on persuasive communication, involving the processing of verbal arguments implying either a positive or negative evaluation of a given target object. Associative learning is prominently reflected in research on evaluative conditioning, which refers to the change in the evaluation of a formerly neutral conditioned stimulus due to its pairing with a positive or negative unconditioned stimulus. The currently available evidence suggests that implicit measures are able to capture newly formed associations regardless of whether they are product of associative or propositional learning (for a review, see Gawronski and Sritharan, 2010).

In addition to the question of how mental associations are formed, a central question is how existing associations can be changed. Several early studies suggested that, although associations captured by implicit measures can be formed relatively quickly as a result of minimal experiences, they seem to be much more resistant to change (e.g., Gregg et al., 2006). However, in contrast to this conclusion, a substantial body of research indicates that implicit measures of existing associations often show evidence for change even when explicit measures are unaffected (for a review, see Gawronski and Bodenhausen, 2006). According to the APE model, such patterns occur when repeated pairings of stimuli in the environment change the structure of associations in memory, and the information implied by the new association is rejected as a basis for an explicit judgment (e.g., Gawronski and LeBel, 2008). Moreover, the reverse pattern is predicted to occur when newly acquired information leads to a rejection of existing associations as a basis for an explicit judgment because of their inconsistency with the newly acquired information (e.g., Gawronski and Strack, 2004). Whereas the former case is assumed to produce changes on implicit but not explicit measures, the latter case is assumed to produce changes in explicit but not implicit measures. In addition, the APE model implies that implicit and explicit measures should show corresponding changes if (1) repeated pairings of stimuli in the environment change the structure of associations in memory and the information implied by the new association is accepted as a basis for an explicit judgment and (2) newly acquired descriptive information is accepted as valid and this information leads to formation of new associations in memory (e.g., Whitfield and Jordan, 2009).

Although implicit measures have been shown to be susceptible to various manipulations to induce long-lasting changes in underlying associations, some effects in the literature are more appropriately interpreted as reflecting temporary shifts that may dissipate over time. In Contrast to initial claims that implicit measures may be resistant to contextual influences, a considerable body of research has shown that they are in fact highly malleable (for a review, see Gawronski and Bodenhausen, 2006). For example, several studies have shown that the same target person can elicit different responses on implicit measures depending on the context in which the target is encountered (e.g., Wittenbrink et al., 2001). Such findings have fueled theoretical debates as to whether evaluations captured by implicit measures indeed reflect stable underlying representations (e.g., Fazio, 2007) or instead are constructed on the spot on the basis of momentarily accessible information (e.g., Schwarz, 2007). More recently, Gawronski et al. (2010) have proposed an integrative account that explains context effects by specifying the contextual conditions under which implicit measures reflect (1) initially acquired information, (2) subsequently acquired information that is inconsistent with the initial information, or (3) a mixture of both. According to their account, exposure to expectancy-violating information enhances attention to the context, which leads to an integration of the context into the newly formed representation of the expectancy-violating information. Thus, the expectancy-violating information is assumed to dominate responses only in the context in which this information has been acquired, whereas the previously acquired information is assumed to dominate responses in any other context. This account not only explains a wide range of context effects; it also includes several novel predictions about patterns of stability and change that have been empirically confirmed.

Lack of Process Purity

Implicit measures are often assumed to provide direct proxies for mental associations. However, although the impact of mental associations on implicit measures is rarely disputed in the field of implicit social cognition, a considerable body of research suggests that implicit measures do not provide process-pure reflections of mental associations. To disentangle the contribution of multiple, qualitatively distinct processes to implicit measures, several theorists have developed mathematical modeling procedures to quantify these processes, including applications of process dissociation, multinomial modeling, and diffusion modeling (for a review, see Sherman et al., 2010). The most prominent example is Conrey et al.'s (2005) quad model, which distinguishes between four qualitatively distinct processes underlying observed responses on implicit measures: (1) activation of an association; (2) detection of the correct response required by the task; (3) success at overcoming associative bias; and (4) guessing. Research using the quad model has provided more fine-grained insights into the mechanisms underlying previous findings obtained with implicit measures. Whereas some effects have been shown to be genuinely related to underlying associations, others stem from nonassociative processes, such as the ability to inhibit activated associations. For example, whereas extended training to associate racial groups with positive or negative attributes has been shown to influence associative bias, alcohol-related increases in implicit measures of racial bias stem from impaired inhibitory control rather than genuine changes in the underlying associations (for a review, see Sherman et al., 2008).

Applications

Although implicit measures have their origin in social psychology, they have been applied in virtually all areas of psychology (for an overview, see Gawronski and Payne, 2010). For example, a large body of research in clinical psychology has used implicit measures to investigate the processes underlying various psychopathologies and the effectiveness of different kinds of treatments (for a review, see Teachman et al., 2010). Similarly, research in health psychology has adopted implicit measures to gain deeper insights into the mechanisms underlying health-related behavior, including alcohol consumption, eating behavior, smoking, and sexual health behavior (for a review, see Wiers et al., 2010). Forensic psychologists have used implicit measures to study various psychopathologies in violent and sexual offenders as well as the determinants of recidivism (for a review, see Snowden and Gray, 2010). Research in consumer psychology has gained valuable insights from using implicit measures to study the determinants of product preferences and consumer choices (for a review, see Perkins and Forehand, 2010). Finally, research in political psychology has adopted implicit measures to study the sources of political preferences and determinants of voting decisions (for a review, see Gawronski et al., in press).

See also: Alcohol Use among Young People; Attitude Formation and Change; Attitude Measurement; Attitudes and Behavior; Cognitive Dissonance; Consumer Psychology; Decision Making, Psychology of; Decision Making: Nonrational Theories; Health Behaviors; Health Psychology; Implicit Association Test; Implicit Memory; Intergroup Relations; Knowledge Representation; Personality Assessment; Persuasion Theories; Political Psychology; Psychological Treatment, Effectiveness of; Recidivism; Self-Concept: From Unidimensional to Multidimensional and Beyond; Self-Esteem; Self-Regulated Learning; Sexual Risk Behaviors; Signal Detection Theory; Smoking and Health; Social Categorization; Social Cognition; Social Psychology: Research Methods; Social Psychology; Stereotypes in Social Psychology; Unconscious: History of the Concept; Voting, Explanations of: Social Class.

Bibliography

Banaji, M.R., Hardin, C.D., 1996. Automatic stereotyping. Psychological Science 7, 136–141.

Briñol, P., Petty, R.E., Wheeler, S.C., 2006. Discrepancies between explicit and implicit self-concepts: consequences for information processing. Journal of Personality and Social Psychology 91, 154–170.

Cameron, C.D., Brown-Iannuzzi, J.L., Payne, B.K., 2012. Sequential priming measures of implicit social cognition: a meta-analysis of associations with behavior and explicit attitudes. Personality and Social Psychology Review 16, 330–350.

Conrey, F.R., Sherman, J.W., Gawronski, B., Hugenberg, K., Groom, C., 2005. Separating multiple processes in implicit social cognition: the quad-model of implicit task performance. Journal of Personality and Social Psychology 89, 469–487.

De Houwer, J., 2003. The extrinsic affective Simon task. Experimental Psychology 50, 77–85.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., Moors, A., 2009. Implicit measures: a normative analysis and review. Psychological Bulletin 135, 347–368.

Dovidio, J.F., Kawakami, K., Gaertner, S.L., 2002. Implicit and explicit prejudice and interracial interaction. Journal of Personality and Social Psychology 82, 62–68.

Fazio, R.H., 2007. Attitudes as object-evaluation associations of varying strength. Social Cognition 25, 603–637.

Fazio, R.H., Jackson, J.R., Dunton, B.C., Williams, C.J., 1995. Variability in automatic activation as an unobtrusive measure of racial attitudes: a bona fide pipeline? Journal of Personality and Social Psychology 69, 1013–1027.

Gawronski, B., Bodenhausen, G.V., 2006. Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. Psychological Bulletin 132, 692–731.

Gawronski, B., De Houwer, J., 2014. Implicit measures in social and personality psychology. In: Reis, H.T., Judd, C.M. (Eds.), Handbook of Research Methods in Social and Personality Psychology, second ed. Cambridge University Press, New York, pp. 281–308.

Gawronski, B., Galdi, S., Arcuri, L. What can political psychology learn from implicit measures? Empirical evidence and new directions. Political Psychology, in press.

Gawronski, B., LeBel, E.P., 2008. Understanding patterns of attitude change: when implicit measures show change, but explicit measures do not. Journal of Experimental Social Psychology 44, 1355–1361.

Gawronski, B., Payne, B.K. (Eds.), 2010. Handbook of Implicit Social Cognition: Measurement, Theory, and Applications. Guilford Press, New York.

Gawronski, B., Peters, K.R., Brochu, P.M., Strack, F., 2008. Understanding the relations between different forms of racial prejudice: a cognitive consistency perspective. Personality and Social Psychology Bulletin 34, 648–665.

Gawronski, B., Rydell, R.J., Vervliet, B., De Houwer, J., 2010. Generalization versus contextualization in automatic evaluation. Journal of Experimental Psychology: General 139, 683–701.

Gawronski, B., Sritharan, R., 2010. Formation, change, and contextualization of mental associations: determinants and principles of variations in implicit measures. In: Gawronski, B., Payne, B.K. (Eds.), Handbook of Implicit Social Cognition: Measurement, Theory, and Applications. Guilford Press, New York, pp. 216–240.

Gawronski, B., Strack, F., 2004. On the propositional nature of cognitive consistency: dissonance changes explicit, but not implicit attitudes. Journal of Experimental Social Psychology 40, 535–542.

Gawronski, B., Ye, Y., 2014. What drives priming effects in the affect misattribution procedure? Personality and Social Psychology Bulletin 40, 3–15.

Greenwald, A.G., Banaji, M.R., 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. Psychological Review 102, 4–27.

Greenwald, A.G., Banaji, M.R., Rudman, L.A., Farnham, S.D., Nosek, B.A., Mellott, D.S., 2002. A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. Psychological Review 109, 3–25.

Greenwald, A.G., McGhee, D.E., Schwartz, J.L.K., 1998. Measuring individual differences in implicit cognition: the implicit association test. Journal of Personality and Social Psychology 74, 1464–1480.

Greenwald, A.G., Poehlman, T.A., Uhlmann, E., Banaji, M.R., 2009. Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. Journal of Personality and Social Psychology 97, 17–41.

Gregg, A.P., Seibt, B., Banaji, M.R., 2006. Easier done than undone: asymmetry in the malleability of implicit preferences. Journal of Personality and Social Psychology 90, 1–20.

Hahn, A., Judd, C.M., Katz, H.K., Blair, I.V., 2014. Awareness of implicit attitudes. Journal of Experimental Psychology: General 143 (3), 1369–1392.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., Schmitt, M., 2005. A meta-analysis on the correlation between the implicit association test and explicit self-report measures. Personality and Social Psychology Bulletin 31, 1369–1385.

Hofmann, W., Gschwendner, T., Wiers, R., Friese, M., Schmitt, M., 2008. Working memory capacity and self-regulation: towards an individual differences perspective on behavior determination by automatic versus controlled processes. Journal of Personality and Social Psychology 95, 962–977.

Hofmann, W., Rauch, W., Gawronski, B., 2007. And deplete us not into temptation: automatic attitudes, dietary restraint, and self-regulatory resources as determinants of eating behavior. Journal of Experimental Social Psychology 43, 497–504.

Hughes, S., Barnes-Holmes, D., De Houwer, J., 2011. The dominance of associative theorising in implicit attitude research: propositional and behavioural alternatives. The Psychological Record 61, 465–498.

Jordan, C.H., Spencer, S.J., Zanna, M.P., Hoshino-Browne, E., Correll, J., 2003. Secure and defensive high self-esteem. Journal of Personality and Social Psychology 85, 969–978.

Nosek, B.A., Banaji, M.R., 2001. The go/no-go association task. Social Cognition 19, 625–666.

Payne, B.K., Cheng, C.M., Govorun, O., Stewart, B., 2005. An inkblot for attitudes: affect misattribution as implicit measurement. Journal of Personality and Social Psychology 89, 277–293.

Payne, B.K., Gawronski, B., 2010. A history of implicit social cognition: where is it coming from? Where is it now? Where is it going? In: Gawronski, B., Payne, B.K. (Eds.), Handbook of Implicit Social Cognition: Measurement, Theory, and Applications. Guilford Press, New York, pp. 1–15.

Perkins, A., Forehand, M., 2010. Implicit social cognition and indirect measures in consumer behavior. In: Gawronski, B., Payne, B.K. (Eds.), Handbook of Implicit Social Cognition: Measurement, Theory, and Applications. Guilford Press, New York, pp. 535–547.

Perugini, M., Richetin, J., Zogmeister, C., 2010. Prediction of behavior. In: Gawronski, B., Payne, B.K. (Eds.), Handbook of Implicit Social Cognition: Measurement, Theory, and Applications. Guilford Press, New York, pp. 255–277.

Ranganath, K.A., Smith, C.T., Nosek, B.A., 2008. Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. Journal of Experimental Social Psychology 44, 386–396.

Sava, F.A., Maricutoiu, L.P., Rusu, S., Macsinga, I., Virga, D., Cheng, C.M., Payne, B.K., 2012. An inkblot for the implicit assessment of personality: the semantic misattribution procedure. European Journal of Personality 26, 613–628.

Schwarz, N., 2007. Attitude construction: evaluation in context. Social Cognition 25, 638–656.

Sherman, J.W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T.A., Groom, C., 2008. The self-regulation of automatic associations and behavioral impulses. Psychological Review 115, 314–335.

Sherman, J.W., Klauer, K.C., Allen, T.J., 2010. Mathematical modeling of implicit social cognition: the machine in the ghost. In: Gawronski, B., Payne, B.K. (Eds.), Handbook of Implicit Social Cognition: Measurement, Theory, and Applications. Guilford Press, New York, pp. 156–175.

Snowden, R.J., Gray, N.S., 2010. Implicit social cognition in forensic settings. In: Gawronski, B., Payne, B.K. (Eds.), Handbook of Implicit Social Cognition: Measurement, Theory, and Applications. Guilford Press, New York, pp. 522–534.

Teachman, B.A., Cody, M.W., Clerkin, E.M., 2010. Clinical applications of implicit social cognition theories and methods. In: Gawronski, B., Payne, K. (Eds.), Handbook of Implicit Social Cognition: Measurement, Theory, and Applications. Guilford Press, New York, pp. 489–521.

Teige, S., Schnabel, K., Banse, R., Asendorpf, J.B., 2004. Assessment of multiple implicit self-concept dimensions using the Extrinsic Affective Simon Task. European Journal of Personality 18, 495–520.

Teige-Mocigemba, S., Klauer, K.C., Sherman, J.W., 2010. A practical guide to Implicit Association Tests and related tasks. In: Gawronski, B., Payne, B.K. (Eds.), Handbook of Implicit Social Cognition: Measurement, Theory, and Applications. Guilford Press, New York, pp. 117–139.

Wentura, D., Degner, J., 2010. A practical guide to sequential priming and related tasks. In: Gawronski, B., Payne, B.K. (Eds.), Handbook of Implicit Social Cognition: Measurement, Theory, and Applications. Guilford Press, New York, pp. 95–116.

Whitfield, M., Jordan, C.H., 2009. Mutual influences of explicit and implicit attitudes. Journal of Experimental Social Psychology 45, 748–759.

Wiers, R.W., Houben, K., Roefs, A., De Jong, P., Hofmann, W., Stacy, A., 2010. Implicit cognition in health psychology: why common sense goes out of the window. In: Gawronski, B., Payne, B.K. (Eds.), Handbook of Implicit Social Cognition: Measurement, Theory, and Applications. Guilford Press, New York, pp. 463–488.

Wilson, T.D., Lindsey, S., Schooler, T.Y., 2000. A model of dual attitudes. Psychological Review 107, 101–126.

Wittenbrink, B., Judd, C.M., Park, B., 1997. Evidence for racial prejudice at the implicit level and its relationships with questionnaire measures. Journal of Personality and Social Psychology 72, 262–274.

Wittenbrink, B., Judd, C.M., Park, B., 2001. Spontaneous prejudice in context: variability in automatically activated attitudes. Journal of Personality and Social Psychology 81, 815–827.

Imprisonment: Sociological Aspects

Tanya N Whittle, University of Delaware, Newark, DE, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by N. Christie, volume 11, pp. 7248–7251, © 2001, Elsevier Ltd.

Abstract

Although incarceration of prisoners is found worldwide, imprisonment rates and conditions vary dramatically. This article provides an overview of the general types, goals, and conditions of imprisonment. Variation in prison conditions and imprisonment rates are discussed along with the social causes of variation and the larger sociological effects of imprisonment.

Introduction

Imprisonment is generally understood to refer to involuntary or coerced physical confinement; however, imprisonment takes many different forms, and imprisonment rates vary significantly over time and geographical area. For instance, the United States of America, the world's imprisonment leader, incarcerates approximately 743 people per 100 000 inhabitants while over half of the countries and territories of the world imprison less than 150 (Walmsley, 2011). Variation in imprisonment rates is not simply a result of variation in crime rates; there are many different causes of variation, such as different goals of imprisonment, politics, and economics.

Types of Imprisonment

Although imprisonment is usually understood in the traditional sense of physical confinement of criminals, control can take many forms. Imprisonment in Western nations usually refers to confinement in small jail and prison cells; however, "Imprisonment is not only in small rooms. Whole societies, or segments of them, can experience the meaning of being in prison" (Christie, 2001: 2). For instance millions of people in open and closed Jewish ghettos during World War II and arguably in modern day North Korea (the Democratic People's Republic of Korea) experienced imprisonment despite formal incarceration in an official jail or prison. Outside of these extreme cases, millions of people worldwide are imprisoned in work camps, boot camps, or are confined to small sections of the community through house arrest, segregation, and/or GPS monitoring that will notify the police or probation officers if the prisoner strays outside of an authorized location.

Goals of Imprisonment

There are many reasons for and goals of imprisonment. From a criminological standpoint, imprisonment is used for retribution, deterrence, incapacitation, or rehabilitation. Retributive imprisonment is punishment that is considered morally just and deserved (i.e., 'just desserts'). Imprisonment of criminals can also prevent more crimes by incapacitating individual criminals while they are imprisoned and may deter future crimes by the incarcerated person (specific deterrence) or others who are made aware of criminals' imprisonment (general deterrence). Incarceration of criminals is also sometimes used as part of a rehabilitative or correctional treatment program aimed at altering prisoners' cognitive thinking, vocational skills, education, etc.

In addition to lowering crime rates or punishing criminals, imprisonment may have political and economic motives. Political leaders may incarcerate segments of the population who are considered 'undesirable' (e.g., economic, racial, ethnic, and religious minorities) in order to better manage and control them. For example, racial and ethnic minorities are disproportionately incarcerated, and minority population size is positively associated with imprisonment and sentencing disparities, presumably due to the dominant population viewing minorities as economic and political threats (Bridges and Crutchfield, 1988; Ulmer and Johnson, 2004; Blumer, 1958). Additionally, people who are unable to or refuse to participate in consumer society may be imprisoned (Bauman, 1998). Some prisoners are incarcerated as a result of state officials' attempts of quelling political movements. For instance, opposition leaders from Taiwan's movement for independence were imprisoned for challenging the standing regime, and thousands of African National Congress and Pan-African Congress members were imprisoned in South Africa for their efforts to end apartheid (Buntman and Huang, 2000).

Prisoners also have economic value. They are sometimes used as part of bribes and negotiations. Additionally, prisoners have historically been used as slave labor and as part of the war industry. For instance, prisoners of war were critical to the Germans and Soviets war efforts during World War II: "In their captured position, they are also prizes—they are here given a use function as a raw material for the control industry (Christie, 2000)" (as cited in Christie, 2001: 6). More recently, prisoners have become a commodity within the prison–industrial complex, particularly with the proliferation of private prisons.

Imprisonment Rates and Understanding Variance

In 2011, according to the International Centre for Prison Studies, 10.1 million people were incarcerated worldwide (Walmsley, 2011). Over half of those people were imprisoned in just three countries: the United States (2.29 million), China (1.65 million), and Russia (0.81 million). In order to account

for large variation in population size, imprisonment *rates* are more useful for comparative purposes. For instance, although China incarcerates twice as many people as Russia, because the Chinese population is much larger, Russia's incarceration rate it much greater (568 per 100 000 inhabitants) than China's (122 per 100 000). Approximately 146 per 100 000 of the world's population are incarcerated.

Countries that are otherwise similar can have very different incarceration rates. For instance, the United States of America incarcerates at a much higher rate (743 per 100 000 persons) than Canada (117 per 100 000) despite similar crime rates, culture, and economic structures and sharing a very long boarder (see Table 1). Similarly, Table 2 shows the great variation that exists between Eastern and Western European counties with former USSR nations, such as Russia and Georgia having much higher imprisonment rates. Imprisonment rates in Asia vary widely with much lower incarceration rates on the Indian subcontinent (42 per 100 000 average) compared to Eastern Asian nations (155.5 per 100 000 average) (see Table 3).

The great variation in imprisonment rates from nation to nation is not simply a reflection of variance in crime rates. Imprisonment rates may vary independently of crime rates due to changes in public sentiment, politics, law enforcement practices, and sentencing guidelines. For instance, the imprisonment rate in Russia decreased from 1400 per 100 000 in 1950 to 353 per 100 000 in 1989 and has since increased back up to 568 per 100 000. These changes were the result of changing politics – not a dramatic change in crime rates. Similarly, although Japan has a much lower imprisonment rate compared to other industrialized countries, its rate has grown over the past two decades due to longer sentences, an increase in the percent of one's sentence offenders must complete before release, and a decline in the parole rate (Shinkai, 2010). Finland and Germany both dramatically decreased their imprisonment rates intentionally by shifting from a reliance on imprisonment to punishments such as fines and community service (Weiss and MacKenzie, 2010). Paradigm shifts like these are the result of changing sentiment and policy.

The United States also underwent a paradigm shift during the same time frame; however, instead of lowering incarceration rates, the US increased the criminalization of drugs with the War on Drugs instead of treating drug use as a public health issue and despite relatively stable drug use rates. Conversely, many European nations, Canada, and Mexico, understand drug use to be a public health issue and have recently adopted public health focused harm-reduction measures in place of punishment. Decriminalization of all drugs in Portugal and heroin-assisted treatment programs around the world are related to

Table 1 Prisoners per 100 000 inhabitants in North and Central America

Belize	439
Bermuda	428
Canada	117
Costa Rica	249
El Salvador	391
Guatemala	77
Greenland	340
Honduras	154
Mexico	200
Nicaragua	119
Panama	349
The United States of America	743

Source: Walmsley, Roy, 2011. World Prison Population List, ninth ed. International Centre for Prison Studies, London; Table 2.

Table 2 Prisoners per 100 000 inhabitants in select European Nations

Austria	103
Belgium	97
Denmark	74
England and Wales	153
France	96
Georgia	547
Germany	85
Greece	102
Iceland	60
Ireland	100
Italy	111
Netherlands	94
Northern Ireland	90
Norway	73
Poland	218
Portugal	113
Russia	568
Scotland	153
Spain	159
Sweden	78
Switzerland	79
Turkey	168

Source: Walmsley, Roy, 2011. World Prison Population List, ninth ed. International Centre for Prison Studies, London; Table 4.

Table 3 Prisoners per 100 000 inhabitants in select Asian Nations[a]

Afghanistan	62
China	122[b]
Hong King (China)	141
Iran	291
Iraq	101
Israel	319
India	32
Japan	58
Macau (China)	170
Mongolia	269
Pakistan	40
Republic of Korea (South Korea)	94
Saudi Arabia	178
Syria	58
Taiwan	278
Thailand	328

[a]Figures not available for the Democratic People's Republic of Korea (North Korea).
[b]Not including the 650 000 presentenced detainees in 'detention centers.'
Source: Walmsley, Roy, 2011. World Prison Population List, ninth ed. International Centre for Prison Studies, London; Table 3.

a decrease in violent and property crimes and, in turn, a decrease in incarceration of drug users (Greenwald, 2009; Loebermann and Verthein, 2009).

Prison Life

Prison life and conditions very considerably depending on the prison's location, goals, and classification and the community's perception of crime, criminals, and rights. Most prisons are 'closed,' total institutions that have power structures and social hierarchies unique from society outside of the prison walls (see Goffman 1961). Prisoners are under the control of guards with limited access to legal assistance and little recourse for living conditions or abuse. Prisoners, who are incarcerated for political or sexual crimes face the greatest abuses and harshest living conditions. Sex offenders in particular are at the bottom of prison hierarchy and are frequently victimized by guards and other prisoners alike.

Prisons are generally classified by the risk assessment of those housed within them and the fortification techniques used. Minimum and low security prisons are generally filled with inmates who are not consider violent or at great risk of escape attempts, medium security prisons house inmates who are considered moderate risks, and maximum security prisons are used to confine and closely control risky inmates. Some facilities, such as Category D prisons in England and Wales are 'open,' meaning that trusted prisoners may leave for home visits or employment. At the far extreme are Supermax prisons. These prisons are intended to house 'the worst of the worst' – those prisoners who are considered extremely dangerous to society in general or to other inmates and guards – and to 'break down all resistance' (Vervaet, 2010: 33). Inmates of Supermax prisons have no contact with other prisoners, are served food through a slot in the door, and are only allowed out of their cells for an hour a day. Isolation via solitary confinement – either due to incarceration in Supermax prisons, need for protection, or administrative punishment for violation of prison rules – is believed to cause social and psychological problems such as cognitive impairment, hallucinations and/or delusions, sensitivity to stimuli, depression, and anxiety (Haney, 2003).

Abuse and unsafe living conditions are rampant in prisons worldwide. Overcrowding, abuse, and health issues are common problems in prisons. Illness and diseases, such as HIV, hepatitis, and tuberculosis, are much more prevalent among inmates compared to the general population. Prison overcrowding prevents adequate sanitary conditions to be maintained and health care to be provided. Prisoners in Haiti the Ukraine must sleep in shifts because of overcrowding, and some prisoners in Eritrea are housed in unventilated storage containers and basements because regular prisons are over capacity (Bureau of Democracy, Human Rights and Labor, 2013). California prisons are operating at 180% capacity resulting in approximately 34 000 more inmates being housed in California prisons than the facilities are designed to house. A 2009 court decision orders the prison population to be reduced to at least 137.5% capacity because overcrowding was deemed a violation of prisoners' rights as it negatively affected health and safety conditions of the prison.

Broader Sociological Impacts of Imprisonment

Imprisonment has sociological implications that reach beyond prison walls. While the construction and maintenance of prisons provides jobs and may improve the economic well-being of communities around prisons – which are often located in very poor and marginalized areas – imprisonment can hurt the economic and social stability of the communities prisoners are taken from and returned to. Before imprisonment, prisoners are members of communities and often serve as employees, parents, friends, etc. Incarceration may reduce crime by incapacitating criminals and deterring future crimes; however, removal of members from society greatly damages social bonds and increases stain and stigma. Loss of communal members, income, and political voice can result in diminished collective efficacy increased social disorganization, and increased crime rates. Even when prisoners are released back to their communities, the sociological effects of incarceration, such as stigma and disenfranchisement persist. As mentioned above, prisoners are disproportionately members of minority populations; therefore, the imprisonment cycle disproportionately impacts minority families and communities. Members of the dominant group in society, however, are more likely to be employed by prisons and therefore often benefit from the incarceration of others, hence reinforcing existing social disparities.

See also: Collective Efficacy and Crime; Control, Social; Crime and Class; Crime, Immigration, and Ethnicity; Crime, Race and Ethnicity; Criminal Justice, Sociology of; Deterrence Theory: Crime; Deterrence: Sanction Perceptions; Discrimination: Racial; Mass Imprisonment and Its Consequences; Minorities; Oppression; Police: A Sociology of Knowledge Approach; Political Sociology; Prejudice in Society: Psychological Perspectives; Prejudice in Society: Sociological Perspectives; Prison Life, Sociology of: Recent Perspectives from the United Kingdom; Prison Life, Sociology of; Prisons and Imprisonment; Racism: Social Psychological Perspectives; Re-entry and Reintegration After Incarceration; Rehabilitation of the offender; Sentencing, Trends, and Disparities: Crime; Sex offenders; Social Justice: Historical and Theoretical Considerations; Stigma, Social Psychology of.

Bibliography

Bauman, Zygmunt, 1998. Globalization: The Human Consequence. Polity Press, Cambridge, UK.

Blumer, Herbert, 1958. Race prejudice and a sense of group position. Pacific Sociological Review 23, 3–7.

Bridges, George S., Crutchfield, Robert, 1988. Law, social standing, and racial disparities in imprisonment. Social Forces 66, 699–724.

Buntman, Fran, Huang, Tong-Yi, 2000. The role of political imprisonment in developing and enhancing political leadership: a comparative study of South Africa's and Taiwan's democratization. Journal of Asian and African Studies 35 (1), 43–66.

Bureau of Democracy, Human Rights and Labor, 2013. Report on International Prison Conditions. United States Department of State, Washington, DC. Available at: http://www.state.gov/documents/organization/210160.pdf.

Christie, Nils, 2001. Imprisonment: sociological aspects. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of Social and Behavioral Sciences, first ed. Elsevier, New York.
Goffman, Erving, 1961. Asylums: Essays on the Social Situation of Mental Patient and Other Inmates. Anchor Books, New York.
Greenwald, Glenn, 2009. Drug Decriminalization in Portugal: Lessons for Creating Fair and Successful Drug Policies. The Cato Institute, Washington, DC.
Haney, Craig, 2003. Mental health issues in long-term solitary and 'supermax' confinement. Crime & Delinquency 49 (1), 124–156.
Loebermann, Rebecca, Verthein, Uwe, 2009. Explaining the effectiveness of heroin-assisted treatment on crime reduction. Law and Human Behavior 33 (1), 83–95.
Shinkai, Hiroyuki, 2010. After amendment of prison law: current Japanese correctional administration. International Journal of Comparative and Applied Criminal Justice 34 (2), 331–349.
Ulmer, Jeffery T., Johnson, Brian, 2004. Sentencing in context: a multilevel analysis. Criminology 42, 137–177.
Vervaet, Luk, 2010. The violence of incarceration: a response from Mainland Europe. Race & Class 51 (4), 27–38.
Weiss, Douglas B., MacKenzie, Doris L., 2010. A Global perspective on incarceration: how an international focus can help the United States reconsider its incarceration rates. Victims and Offenders 5, 268–282.
Walmsley, Roy, 2011. World Prison Population List, ninth ed. International Centre for Prison Studies, London.

Relevant Websites

http://www.prisonerhealth.org – The Center for Prisoner Health and Human Rights.
http://www.prisonstudies.org – International Centre for Prison Studies.

Incest, Inbreeding, and Their Consequences

Alan H Bittles, Duke University, Durham, NC, USA; Murdoch University, Perth, WA, Australia; and Edith Cowan University, Perth, WA, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The global human population can be divided into societies in which legal marriage between close family members traditionally has been favored and practiced, and those in which it is avoided or precluded. Consanguineous unions are mainly chosen on social and economic grounds, whereas the reasons for their avoidance are based on religious teachings and the prevention of adverse health effects. Incest is the closest form of sexual relationship; currently it is a criminal offense in almost all the societies. Given the stringent legal prohibitions, information on the prevalence and adverse genetic effects of incest is scarce, and most of the available data are significantly biased.

Introduction

The terms 'inbreeding' and 'consanguinity' often are used interchangeably to describe sexual unions between couples known to share at least one common ancestor within a recent generation. In practice, they usually are applied to relationships between individuals who are genetically related as second cousins or closer. Other than in specific immigrant communities, 'consanguineous marriage,' a term derived from the Latin *con sanguineus* (of the same blood), is now rare in Western societies. However, marriage between close biological kin remains preferential in many parts of the world, including North and sub-Saharan Africa, the Middle East, and Central and South Asia. According to the cross-cultural ethnographic tabulations published by G.P. Murdock, 353 of the 763 societies examined either permitted or favored first and/or second cousin marriage (Murdock, 1987). Although the rates and types of consanguineous marriage typically vary according to religious norms, marital unions between first cousins are especially common, and in some countries (such as Pakistan) they can account for over 50% of all marital unions. By comparison, an incestuous relationship is a sexual union between biological relatives that is prohibited by law. Most commonly, incest is defined as mating between first-degree relatives (i.e., father–daughter, mother–son, or brother–sister) who predictably have 50% of their genes in common. However, in some countries the definition of incest has been widened to include half-sib and uncle–niece unions.

History of Human Consanguinity

It has been estimated that the number of breeding individuals in the out-of-Africa migration of early humans 60 000–70 000 years ago ranged from as few as 700 to a maximum of 10 000 adults (Bittles, 2012). These early groups followed a largely hunter–gatherer existence, and they quite rapidly dispersed through Asia and Europe and later, the Americas, in small kindred groupings. Therefore, their available mating choices must have been limited, which in turn led to the inevitability of at least some level of close-kin mating. Even after the establishment of settled communities and with early urbanization, by AD 1000 the global population is estimated to have totaled only 310 million persons; that is, less than 5% of the present world population. Geographical distance, poor communications, and the increasing influence of social stratification meant that marriage partner choices remained highly restricted with populations characterized by marital endogamy, that is, marriages confined within ethnic, geographical, religious, or social subcommunities (Bittles, 2012).

The early biblical texts contain many examples of reproductively successful consanguineous unions, including the marriage of Abraham with Sarah, his half-sister (Genesis 20:12), and Amram and Jochebed, the parents of Aaron and Moses, who were aunt and nephew (Exodus 6:20). At a later date the permitted degrees of marital relationships were extensively defined, with marriages up to and including uncle–niece permitted, but not aunt–nephew (Leviticus 18:7–18). Within Christianity, the Levitical guidelines were generally adopted by the post-Reformation Protestant churches, although in most denominations the closest form of approved martial union is between first cousins. By comparison, diocesan dispensation is required by the Roman Catholic Church for marriage between first cousins, and, with the exception of the Coptic Church, the Christian Orthodox faiths have a similarly proscriptive stance on consanguineous marriage. Attitudes toward close-kin marriage are generally favorable or at least neutral among the other major world religions, including Buddhism, Confucianism, the Dravidian branch of Hinduism, Islam, Judaism, and Parsi/Zoroastrianism. By comparison, in the Indo-European branch of Hinduism and in the Sikh religion, pedigrees may be examined for up to seven generations for males and five generations for females to ensure avoidance of a consanguineous union.

Attitudinal Changes toward Consanguineous Marriage in Western Societies

Prior to the mid-nineteenth century, first cousin marriages were quite commonly contracted in most Western societies, including countries such as Italy and Spain where a large majority of the population was Roman Catholic. From the 1850s, an often acrimonious debate ensued within the scientific and medical communities of North America and Western Europe on the biological effects of close-kin marriage.

In Europe, and in most countries that had been colonies of the major European powers, this debate resulted in few if any legislative changes. By comparison, within the United States the eventual outcome was the passage of laws at the state level to restrict or ban first cousin unions, even though the information on which these laws were promulgated was largely anecdotal rather than factual in nature. Restrictions on first cousin marriage were subsequently rescinded in the states of Michigan, New Jersey, and Ohio, but legislation on consanguinity was introduced in Maine in 1985 and Texas in 2005, with the net result that first cousin marriage is currently illegal or a criminal offense in 31 of the 50 states (Ottenheimer, 1996; Bittles, 2012) – despite a unanimous recommendation by the National Conference of Commissioners on Uniform State Laws in 1970 that all state laws on first cousin marriage should be repealed. Globally, the only other countries in which first cousin marriage is banned are the People's Republic of China and Taiwan, the Republic of Korea and the Democratic People's Republic of Korea, and the Philippines (Bittles, 2012).

Contemporary Societies in Which Consanguineous Marriage Is Preferential

As indicated in the preceding section, the current legal situation in the United States, China, the Koreas, and the Philippines differs markedly from the many other regions of the world where consanguineous unions are strongly preferential. The reasons for this preference are mainly social and economic in origin and include the strengthening of family relationships and the maintenance of family property and landholdings. Prenuptial arrangements are greatly simplified, female status and autonomy are enhanced, and the security of marrying a partner whose entire family background is known is perceived as a major benefit in ensuring both the initial success and the lasting stability of the union (Bittles, 2012). The existence of additional customary influences is, however, indicated by variations in the specific types of consanguineous marriage contracted. For example, while father's brother's daughter first cousin marriage is preferred in Arab Muslim communities, it is proscribed for Dravidian Hindus in South India and the Han Chinese, both of whom may instead contract marriage between a man and his mother's brother's daughter.

The Genetic Effects of Inbreeding

Because of the pejorative social connotations of the term 'inbreeding,' in human populations marriages between biological relatives usually are referred to as being consanguineous. In consanguineous unions the partners share genes inherited from one or more common ancestor, and, for example, in first cousin marriages, the spouses are predicted to have 12.5% of their genes in common. This means that on average their progeny will be homozygous at 6.25% of gene loci; that is, they will have inherited identical gene copies from each parent at these sites in their genome. Consequently, if the same mutant gene is inherited from both parents, the individual will express the associated disorder, either at birth or later in life depending on the nature and site of the mutation, thus contributing to the phenomenon of inbreeding depression that describes the adverse genetic effects of consanguinity.

The coefficient of inbreeding (F) is a numerical estimate of the degree of consanguinity of an individual, and so for first cousin offspring $F = 0.0625$, for example, as previously stated, on average they will have inherited identical gene copies from each parent at 6.25% of all gene loci. Similarly, the coefficient of inbreeding for the progeny of uncle–niece unions is $F = 0.125$, while for second cousin offspring $F = 0.0156$ (Bittles and Black, 2010; Bittles, 2012). In many communities there is a long, unbroken history of consanguineous unions, and the cumulative level of inbreeding may be significantly higher than the value calculated for a single generation. Under these circumstances, where a detailed, multigenerational pedigree is available, a correction can be applied to account for the effects of ancestral inbreeding using the formula:

$$F = \sum (1/2)^n (1 + F_A)$$

where F_A is the ancestor's inbreeding coefficient, n is the number of individuals in the path connecting the parents of the individual, and the summation (Σ) is taken over each path in the pedigree that goes through a common ancestor.

Incest

With a few notable exceptions the avoidance of sexual relationships between close family members is, and appears to have been, common to virtually all societies. Various explanations have been advanced for incest avoidance. One is Freudian guilt, based on the supposed attraction of a child for their parent of the opposite sex. Theories ascribed to Lévi-Strauss and others argue that incest serves both to maintain the existence of the family and to encourage the establishment of affinal relationships through marriage with other kin groups. According to the Westermarck hypothesis (Wolf and Durham, 2004), incest avoidance arises principally because of negative imprinting against close associates of early childhood. Studies conducted on individuals raised under mixed-sex, child-rearing regimes in Israeli *kibbutzim* have been cited in support of the negative imprinting theory, since these individuals rarely intermarried or entered into sexual relationships. Perhaps more convincingly, a detailed investigation of *sim-pua* marriage in Taiwan, whereby a girl is adopted with the expectation that she will marry the son of the adoptive family, demonstrated apparent aversion on the part of some intended spouses. Where a marriage did occur, *sim-pua* unions often were characterized by lower mean fertility and reduced marital stability (Wolf and Durham, 2004).

There have, however, been a number of societies in which dynastic incest was practiced over multiple generations, including the eighteenth and nineteenth Egyptian dynasties and the Ptolemaic and Roman periods in Egypt, Zoroastrian Persia, the Incas, and the royal families of Hawaii (Middleton, 1962; Shaw, 1992; Ager, 2005). In pharaonic Egypt, brother–sister or half-sibling unions were regarded as potent means of maintaining and strengthening the royal house and bloodline, although the pharaoh also had reproductive access to other females who were nonrelatives. Incestuous marriages additionally were contracted outside the ruling classes during the Roman period from the first to the third centuries AD, and in

the city of Arsinoe in Lower Egypt, brother–sister unions accounted for 19.6% of marriages with a further 3.9% between half-siblings (Scheidel, 1997). These marriages may have been atypical and in part stemmed from the restricted numbers of potential marriage partners in the minority Greek settler community. The underlying rationale for incestuous unions in Zoroastrian Persia appears to have been based on specific religious beliefs termed *xvétxvét* in Persian, and to have emerged just prior to the Islamic Arab invasion of AD 636. All three types of incestuous union – father–daughter, brother–sister, and mother–son – were advocated, and in each case they were described as being of especial religious merit, with their practice a means of expiating mortal sin.

Despite the claimed universality of the incest taboo, the prevalence of incestuous relationships in Western societies was estimated to be 5000 cases per million, but with much a higher prevalence reported in detailed studies conducted in the United States and Finland (Bittles, 2012). Most reports of incest have emanated from case studies conducted either on persons examined because of intellectual disability or in psychiatric clinics, which has led to the general conclusion that incestuous relationships are highly detrimental to those classified as victims. It has, however, been suggested that the outcome need not necessarily be psychologically damaging, with some women in stable relationships acknowledging prior father–daughter sexual relationships. Since incest is both illegal and widely regarded as morally reprehensible, it would be difficult to determine how common this latter response might be without exposing respondents to the risk of self-incrimination.

Biological Outcomes of Consanguinity and Incest

The initial major study into the biological effects of inbreeding was conducted by Dr Samuel Bemiss of Louisville, Kentucky, in 1858, based on reports forwarded by medical colleagues on the outcomes of unions ranging from incest to third cousin marriages. Literally thousands of additional studies have since been undertaken using a variety of sampling protocols, including pedigree analysis, household surveys, questionnaires administered to hospital in- and outpatients, Roman Catholic dispensation records, and surname analysis. In populations such as Dravidian South India, where uncle–niece ($F = 0.125$) and first cousin unions ($F = 0.0625$) are preferential and jointly account for some 30% of marriages, unions beyond second cousins ($F < 0.0156$) are of limited biological significance (Bittles et al., 1991). Conversely, in a country such as Finland where consanguineous unions are rare, or in endogamous communities in which close cousin marriage has been proscribed on religious grounds, more remote relationships including third cousin unions and beyond ($F \leq 0.0039$) can result in moderate levels of cumulative inbreeding through time and give rise to adverse clinical outcomes (Bittles, 2009).

Measuring the Biological Outcomes of Inbred Marriages

For several decades the detrimental health effects of consanguinity commonly were assessed in terms of lethal gene equivalents – genes that when homozygous contribute to early mortality. This method has largely been replaced by more direct methods of measurement, for example, comparing the rates of infant deaths among first cousin and non-consanguineous progeny in specific study settings (Bittles, 2012) and by genomic analysis to estimate the percentage of loci at which an individual is homozygous. Genome-wide analysis has been used to identify the presence of uninterrupted runs of homozygosity (ROH; extended segments of the genome at which all loci are homozygous), from which an individual's level of homozygosity can be directly calculated (McQuillan et al., 2008). With the introduction of whole genome and exome (i.e., the coding region of the genome) sequencing, estimates of the proportion of gene loci at which individuals have inherited identical gene copies from a common ancestor have become increasingly more accurate.

The Biological Outcomes of Incest

As first-degree relatives have on average 50% of their genes in common, the progeny of an incestuous union will be homozygous at 25% of gene loci, and so $F = 0.25$. Given the strongly antagonistic legal and social attitudes to incest, direct assessment of associated biological outcomes may be subject to significant ascertainment bias, since a detailed examination including determination of paternity may only be initiated if a child shows obvious symptoms of physical and/or intellectual disability. The net result is an overall paucity of data that may or may not reflect the incidence of incestuous pregnancies that successfully proceed to term. Problems also can arise in attempting to control for the potentially adverse effects of nongenetic variables, which include maternal and paternal age, parental disease, and unsuccessful attempted interruption of the pregnancy.

Direct Investigation of the Progeny of Incest

Table 1 comprises data from the five most comprehensive published studies of incest. The first and obvious observation is that very few cases have been systematically investigated, which is surprising, given the probable prevalence of incest and its prominence as a topic in the social and behavioral literature. Three of the studies were prospective, and in one investigation strenuous efforts were made to recruit matched nonincestuous controls (Adams and Neel, 1967). No control subjects were assessed in two of the other studies (Carter, 1967; Baird and McGillivray, 1982), and the brief report by Carter (1967) was further restricted by a lack of information on the physical and mental status of the parents or their socioeconomic background. In several studies the period of follow-up was brief, and so disorders with a later age of onset may not have been detected.

By far the largest survey was that of Seemanová (1971), which was conducted retrospectively on incestuous births reported in Czechoslovakia between 1933 and 1970. An ingenious approach to the problem of controls was employed, with the outcomes of incestuous matings compared to those of pregnancies undertaken by a subset of the same women and unrelated partners. Unfortunately, the methodology could not overcome the fact that the physical and mental status of these women, who formed 33% of the original total, was

Table 1 Health status of the progeny of incestuous relationships ($F = 0.25$)

Country of origin	*Sample size*	*Follow-up (years)*	*Deaths/severe defects*	*Other defects/mild intellectual disability*	*Normal*	*Authors*
United States	18	0.5	6	5	7	Adams and Neel (1967)
United Kingdom	13	4–6	4	4	5	Carter (1967)
Czechoslovakia	161	1–37	65	18	78	Seemanová (1971)
United Kingdom	23	4–25	11	5	7	Bundey (1979)
Canada	21	0.5–19	9	4	8	Baird and McGillivray (1982)
Total	236		95 (40.3%)	36 (15.3%)	105 (44.5%)	

significantly superior to the remainder of the study group. In addition, many of the incestuous pregnancies occurred when the women had just completed menarche; that is, became physiologically capable of reproduction, with mean and modal ages of 18.9 and 16 years in the father–daughter matings and 19.9 and 14 years in the brother–sister matings, versus 24.9 and 21 years in the married pregnancy group (Bittles, 2012). As very young maternal age is significantly associated with an adverse pregnancy outcome, failure to control for the age discrepancies between the test group and their half-sibs significantly detracts from the validity of the study findings. With such small numbers significant bias can be introduced by a single individual; for example, the Seemanová study in which an alcoholic male had fathered five children with three daughters. All five children were born with some form of abnormality, and three died within 10 days of birth.

How then should these five data sets be assessed? Clearly, the mean total rate of death and physical or mental disability of ~55% is very high. But, as previously indicated, the collection of data on this subject is extremely difficult, and, as acknowledged by several authors, control for important nongenetic variables may have been seriously incomplete. In the two studies for which unrelated reference groups were available, 9 of the 113 control children died or had a serious defect (8.0%). Thus the mean excess level of death and severe defect in the offspring of incestuous unions was 32.3%, not all of which was necessarily genetic in origin.

Comparison of Incest with Other Categories of Consanguineous Union

A second method of estimating the probable levels of defect in the progeny of incestuous matings is based on information gained from legal consanguineous unions. For example, it is assumed that as the progeny of an incestuous mating will have a coefficient of inbreeding of $F = 0.25$, their levels of inherited disease leading to disability or death will be four times larger than in first cousin offspring ($F = 0.0625$). The main advantage of this approach is that information on first cousin unions is plentiful, and as a legal form of martial union it is less likely to be subject to bias during data collection. A potential disadvantage is that the relationship between the level of consanguinity and the occurrence of deaths and defects may be nonlinear, with disproportionately larger increases in the prevalence of deaths and defects at closer degrees of inbreeding. This situation could arise if incest was associated with increased rates of conditional lethals; that is, genes that are only expressed under especially stressful circumstances, as might be encountered in an incestuous pregnancy. To date, there has been no evidence that conditional lethals are expressed in the progeny of close-kin unions. Therefore, the excess mortality rate of 3.7% derived for deaths among first cousin progeny (Bittles, 2012) can be applied to estimate the prevalence of lethal defects in an incestuous pregnancy; that is, ~15% excess deaths. Likewise, based on first cousin survey data, the mean excess morbidity among incestuous offspring would predictably be ~13–16% (Bittles, 2012), suggesting a total excess death and major disability in children born to first-degree relatives of approximately 30%.

Conclusions

The data currently available on the outcomes of both legal consanguineous marriages and incestuous relationships merit further examination and updating. Since the major studies undertaken by Schull and Neel and their colleagues in Japan from the late 1940s to the 1970s, initially in Hiroshima and Nagasaki (Schull and Neel, 1965), and later on the island of Hirado, no comparably detailed investigations have been attempted, despite the remarkable epidemiological transition that has occurred during the last two generations. Across Asia, genetic disorders are becoming increasingly obvious as deaths due to infectious diseases decline in number and proportion, and early childhood mortality is replaced by extended morbidity. This change is especially important among couples in consanguineous unions, where higher rates of rare genetic disorders are to be expected involving both autosomal recessive and polygenic disorders. In fact, because of better access to diagnostic and health care facilities, the burden of genetic disease will probably be most obvious and of greatest significance among migrants from Asia and Africa now resident in North America, Western Europe, and Australasia, who have chosen to continue their tradition of close-kin marriage. With respect to incest, it is curious that the voluminous social and behavioral literature is still so inadequately matched by comparable reports on biological outcomes. Studies to identify the precise nature, extent, and timing of any genetic defects that may be expressed are needed, since in a proportion of cases the children of incestuous unions may be offered for adoption. The advent of whole genome analysis and exome sequencing will greatly facilitate this process and provide information of previously unparalleled insight and sophistication (Rehder et al., 2013). In the interim, appropriate caution should be exercised in judging and reporting the biological outcomes of all categories of close-kin union.

See also: Abortion, Demography of: United States; Ancestors, Anthropology of; Caste; Clan; Community Genetics; Familial Studies: Genetic Inferences; Family and Kinship, History of; Genealogy in Anthropology; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Screening for Disease; Genetics and Anthropology; Intelligence, Genetics of: Cognitive Abilities; Kinship in Anthropology; Kinship, Formal Models of; Marriage; Religion, Family, and Kinship; Science and Law; Social Stratification; Tribe.

Bibliography

Adams, M.S., Neel, J.V., 1967. Children of incest. Pediatrics 40, 55–62.

Ager, S.L., 2005. Familiarity breeds: incest and the Ptolemaic dynasty. Journal of Hellenistic Studies 125, 1–34.

Baird, P.A., McGillivray, B., 1982. Children of incest. Journal of Pediatrics 101, 854–857.

Bittles, A.H., 2009. Consanguinity, genetic drift and genetic diseases in populations with reduced numbers of founders. In: Speicher, M., Antonarakis, S.E., Motulsky, A.G. (Eds.), Human Genetics – Principles and Approaches, fourth ed. Springer, Heidelberg, pp. 507–528.

Bittles, A.H., 2012. Consanguinity in Context. Cambridge University Press, Cambridge.

Bittles, A.H., Black, M.L., 2010. Consanguinity, human evolution and complex diseases. Proceedings of the National Academy of Sciences of the United States of America 107, 1779–1786.

Bittles, A.H., Mason, W.M., Greene, J., Appaji Rao, N., 1991. Reproductive behavior and health in consanguineous marriages. Science 252, 789–794.

Bundey, S., 1979. The child of an incestuous union. In: Wolkind, S. (Ed.), Medical Aspects of Adoption and Foster Care. Spastics International Medical Publications, London, pp. 36–41.

Carter, C.O., 1967. Risk to offspring of incest. Lancet 289, 436.

McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenes, A., Macleod, A.K., Farrington, S.M., Rudan, P., Hayward, C., Vitart, V., Rudan, I., Wild, S.H., Dunlop, M.G., Wright, A.F., Campbell, H., Wilson, J.F., 2008. Runs of homozygosity in European populations. American Journal of Human Genetics 83, 359–372.

Middleton, R., 1962. Brother–sister and father–daughter marriage in Ancient Egypt. American Sociology Review 27, 603–611.

Murdock, G.P., 1987. Ethnographic Atlas. University of Pittsburgh Press, Pittsburgh.

Ottenheimer, M., 1996. Forbidden Relatives – the American Myth of Cousin Marriage. University of Illinois Press, Chicago.

Rehder, C.W., David, K.L., Hirsch, B., Toriello, H.V., Wilson, C.M., Kearney, H.M., 2013. American College of Medical Genetics and Genomics: standards and guidelines for documenting suspected consanguinity as an incidental finding of genomic testing. Genetics in Medicine 15, 150–152.

Scheidel, W., 1997. Brother–sister marriage in Roman Egypt. Journal of Biosocial Science 29, 361–371.

Schull, W.J., Neel, J.V., 1965. The Effects of Inbreeding on Japanese Children. Harper and Row, New York.

Seemanová, E., 1971. A study of incestuous matings. Human Heredity 21, 108–128.

Shaw, B.D., 1992. Explaining incest: brother–sister marriage in Graeco-Roman Egypt. Man (New Series) 27, 267–299.

Wolf, A., Durham, W. (Eds.), 2004. Inbreeding, Incest, and the Incest Taboo. Stanford University Press, Stanford.

Relevant Website

http://www.consang.net – Global Consanguinity Website.

Incest Prohibition, Origin and Evolution of

Arthur P Wolf, Stanford University, Stanford, CA, USA

Published by Elsevier Ltd.

Abstract

Why do people avoid sexual relations with close relatives? And why do they punish and condemn as inhuman the few exceptions? The many theorists who have addressed these questions fall naturally into one of two mutually hostile camps – constitutionalists and conventionalists. The former include, most notably, the Finnish scholar Edward Westermarck who argued that the proximate cause of incest avoidance is early association and the ultimate cause, the dangers of inbreeding. The latter include E.B. Tylor, Leslie White, Claude Levi-Strauss, and a motley collection of Marxists. They attribute incest avoidance to the incest taboos and argue that the taboos were instituted to secure various social advantages. They follow Freud in assuming that human beings are naturally inclined to mate and marry in the family. After more than a 100 years of intense debate most scholars now agree that Westermarck was right in attributing incest avoidance to the dangers of inbreeding, but there is still no generally accepted explanation of the incest taboos. The conventionalist view has been eroded by a flow of evidence indicating that the taboos are not necessary to achieve whatever social advantages exogamy may entail, but constitutionalists have failed to provide a convincing alternative. All they have to offer is the suggestion that the taboos are by-products of incest avoidance.

What is commonly called the incest problem consists of two separable parts: incest avoidance and the incest taboos. Thus the problem raises two questions about human behavior: Why is it that, for the most part, humans avoid sexual relations with their close relatives? and Why is it that they disapprove of and often punish people who do not avoid sexual relations with their close relatives? The second question is complicated by the fact that the scope of the incest taboo varies from society to society. In most societies the taboo is extended beyond the family, but the direction of the extensions varies widely. Where many societies prohibit marriage with paternal relatives but allow, and in some cases prefer, marriage with maternal relatives, just the opposite is the case in many others.

Most attempts to answer these questions belong to one or the other of two radically different traditions, what are best called the 'constitutionalist tradition' and the 'conventionalist tradition.' Constitutionalists typically emphasize the dangers of inbreeding; conventionalists, the social advantages of a rule that requires marriage outside of the family. The best known of the constitutionalist arguments was the work of the Finnish philosopher *cum* anthropologist Edward Westermarck. Following the train of thought put in motion by Charles Darwin, he argued that the dangers of inbreeding made human beings naturally averse to sexual relations with anyone with whom they were reared. Humans typically avoided sexual relations with their close relatives because these were the people with whom or by whom they were reared.

The most notable conventionalist contribution is the independent, but almost simultaneous work of the American materialist, Leslie White, and the famous French structuralist, Claude Levi-Strauss. Arguing that inbreeding is only dangerous for inferior stock, they rejected Westermarck's suggestion that early association inhibits sexual attraction, insisting instead that Freud was right in claiming that (1953[1920]: pp. 120–121). Both authors then argued that the incest taboos were instituted to force people to marry outside of the group in which they were reared, their purpose being to create larger, more effective social units by binding together families that might otherwise remain isolated self-perpetuating units. Because this required suppressing a natural inclination, Levi-Strauss once said of the incest taboos, "It is there, and only there, that we find a passage from nature to culture, from animal to human life." (1960: p. 278).

The controversy aroused by these two approaches to the incest problem highlights the importance of three questions: Does inbreeding have deleterious consequences? Is incest common? and Does early association inhibit sexual attraction? The answer to the first question was not finally agreed on until researchers took advantage of the fact that cousin marriage is common in many parts of the world and calculated the excess mortality and major disability rates of the offspring (Bittles, 2005). Using this evidence, they estimated that for the offspring of brother/sister, father/daughter, and mother/son unions, the combined rates would be somewhere between 22 and 36%, more than sufficient to generate a strong evolutionary bias favoring any behavior that reduced the frequency of incest (Wolf, 1995). Evidence from all the societies for which we have evidence says that incest is rare among sexually mature adults. The female is almost always a child and the male a stepfather or a father who was rarely at home when his daughter was a child, or if he was, did not help care for her. Initiated when the female is a child, relationships of this kind are typically terminated by her at puberty, the result being that most of what is counted as incest is the abuse of children.

Evidence relevant to the third question has been slow coming because children reared together are usually siblings, making it difficult to determine if they avoid one another because of early association or because of a barrier imposed by the incest taboo (Wolf, 2014) that has finally displaced Freud in favor of Westermarck comes from two unlikely sources, the early Israeli kibbutzim where communal child rearing was the norm, and South China where it was once customary for families to give away most of their female children shortly after birth and rear in their places wives for their sons. Though they were reared together as intimately as brothers and sisters, kibbutz children were not constrained by an incest taboo. To

the contrary, they were encouraged to marry by elders anxious to keep them in the collective. The fact is, however, that as long as their association began before age three or four, they never married among themselves and, according to reliable observers, never engaged one another in premarital sex. In China, by contrast, the children who were reared together as future mates were forced to marry, the result being abnormally low fertility and abnormally high divorce and adultery rates. Thus, taken together, the Israeli and Chinese cases argue that children who are reared together will not marry if they are not forced to do so, and, if they are forced, will behave as one would expect if early association inhibited sexual attraction.

As the answers to these three questions emerged in the 1960s and 1970s, scholarly opinion shifted from a marked preference for the conventionalist approach to the incest problem to a general if often reluctant acceptance of the constitutionalist approach. The view that the incest taboos are necessary to suppress a natural preference of the kind posited by Freud could not withstand the combined impact of evidence showing that incest is biologically deleterious, rare among sexually mature adults, and generally inhibited by early association. What most constitutionalists have failed to recognize, however, is that while this evidence grounds their explanation of incest avoidance, it does not explain the incest taboos. All contemporary constitutionalists like E.O. Wilson have to say about the taboos is that they "seem to have arisen from the Westermarck effect" (1978: p. 37) or are "by products" of the "deeper, more urgent causes behind incest avoidance" (1998: p. 180).

A great deal more is needed for a complete solution to the incest problem, most importantly an explanation of why people condemn behavior that does themselves no harm and why the range of people vulnerable to such condemnation varies from society to society.

See also: Fertility and Social Networks; Incest, Inbreeding, and Their Consequences; Parenting Attitudes and Beliefs across Cultures; Sexuality: Cultural Aspects.

Bibliography

Arnhart, Larry, 2005. The incest taboo as Darwinian natural right. In: Wolf, Arthur P., Durham, William H. (Eds.), Inbreeding, Incest, and Incest Taboo. Stanford University Press, Stanford, CA, pp. 190–217.

Bittles, Alan H., 2005. Genetic aspects of inbreeding and incest. In: Wolf, Arthur P., Durham, William H. (Eds.), Inbreeding, Incest, and Incest Taboo. Stanford University Press, Stanford, CA, pp. 38–60.

Erickson, Mark T., 2005. Evolutionary thought and the current clinical understanding of incest. In: Wolf, Arthur P., Durham, William H. (Eds.), Inbreeding, Incest, and Incest Taboo. Stanford University Press, Stanford, Calif., pp. 161–189.

Fox, Robin, 1962. Sibling incest. British Journal of Sociology 13, 128–150.

Freud, Sigmund, 1953[1920]. A General Introduction to Psychoanalysis (Joan Riviere, Trans.). Pocket Books, New York, pp. 120–121.

Hopkins, Keith, 1980. Brother-sister marriage in Roman Egypt. Comparative Studies in Society and History 20, 303–354.

Levi-Strauss, Claude, 1960. The family. In: Shapiro, Harry L. (Ed.), Man, Culture, and Society. Oxford University Press, New York, p. 278.

Levi-Strauss, 1969. The Elementary Structures of Kinship (James H. Bell and John R. von Sturmer, Trans.). Beacon Press, Boston, pp. 12–25, (Chapter 2).

Packer, Craig, 1979. Inter-troops transfer and inbreeding avoidance in Papio anubis. Animal Behaviour 27, 1–36.

Pusey, Anne E., 1980. Inbreeding avoidance in chimpanzees. Animal Behaviour 28, pp. 453–452.

Scheidel, Walter, 2005. Ancient Egyptian sibling marriage and the Westermarck effect. In: Wolf, Arthur P., Durham, William H. (Eds.), Inbreeding, Incest, and Incest Taboo. Stanford University Press, Stanford, CA, pp. 93–108.

Shepher, Joseph, 1971. Mate selection among second generation kibbutz adolescents and adults. Archives of Sexual Behavior 1, 293–307.

Talmon, Yonina, 1964. Mate selection in collective settlements. American Sociological Review 29, 491–508.

White, Leslie, 1948. The definition and prohibition of incest. American Anthropologist 50, 416–434.

Westermarck, Edward, 1922. The History of Human Marriage, fifth ed., vol. 2. The Allerton Book Co, New York (Chapters 19 and 20), pp. 82–239.

Wilson, Edward O., 1978. On Human Nature. Harvard University Press, Cambridge, MA.

Wilson, Edward O., 1998. Consilience: The Unity of Human Knowledge. Knopf, New York.

Wolf, Arthur P., 1995. Sexual Attraction and Childhood Association: A Chinese Brief for Edward Westermarck. Stanford University Press, Stanford, CA.

Wolf, Arthur P., 2014. Incest Avoidance and the Incest Taboos: Two Aspects of Human Nature. Stanford University Press.

Income Distribution and Inequality: Measurement Issues

David S Johnson, U.S. Bureau of Economic Analysis, Washington, DC, USA
Timothy M Smeeding, University of Wisconsin–Madison, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A.B. Atkinson, volume 11, pp. 7265–7271, © 2001, Elsevier Ltd.

Abstract

Recent research establishes strong evidence that economic inequality in the US has increased over the past few decades, but that the amount of increase varies by the measure used. In this article, we will examine the components of measures of inequality – the 'who, what, where, when, why, and how' of inequality measures. We show that what resource measure, which data source, and whose standard of living all affect the measurement of both the level of trend in inequality. We also provide examples of how these choices affect the increase in inequality in the US.

The significance of the distribution of resources and its importance to overall economic well-being was recognized by Adam Smith in his characterization of the invisible hand. Smith (1761) stated: "The rich only select from the heap what is most precious and agreeable. They consume little more than the poor, and in spite of their natural selfishness and rapacity … They are led by an invisible hand to make nearly the same distribution of the necessaries of life, which would have been made, had the earth been divided into equal portions among all its inhabitants, and thus without intending it, without knowing it, advance the interest of the society, and afford means to the multiplication of the species." David Ricardo (1821) continued along this theme in discussing how the "… produce of the earth …" is "… divided among the three classes of community …" and concludes "To determine the laws which regulate this distribution, is the principal problem in Political Economy …."

The increased attention to inequality in the twentieth century was stimulated by Simon Kuznets who emphasized the importance of the distribution of income in his seminal paper (Kuznets, 1955). He concludes: "Without better knowledge of the trends in secular income structure and of the factors that determine them, our understanding of the whole process of economic growth is limited; and any insight we may derive from observing changes in countrywide aggregates over time will be defective if these changes are not translated into movements of shares of the various income groups."

After a period in the 1960 and 1970s, where one economist remarked that "studying inequality was like watching the grass grow" (Aaron, 1978), subsequent and continual increases in inequality have brought it to a point where it is being brought 'back in from the cold' and has re-entered mainstream economic thought (Atkinson, 1997). This interest continues today, with an increase in the evaluation and estimation of inequality. The *2012 Economic Report of the President* stated: "The confluence of rising inequality and low economic mobility over the past three decades poses a real threat to the United States as a land of opportunity." Isabel Sawhill (2012) stated that "2011 was the year of inequality" and Paul Krugman (1992) introduced the "staircase vs picket fence" analogy, where he concluded that the change in income gains between 1973 and 1993 followed a staircase pattern with income growth rates increasing with higher income quintiles.

From the beginning, Pareto (1897) demonstrated the importance of individual record empirical data in evaluating the distribution of income. Watkins (1908) and Persons (1909) also used household data to estimate levels of income inequality. Since then, the study of income inequality has progressed, with new measures, new data, examining the impact of taxes, benefits, and so on. Corrado Gini (1921), Hugh Dalton (1920), Jan Pen (1971), Max Lorenz (1905), Henry Theil (1967), and Anthony B. Atkinson (1970) have each created particular and complimentary measures of inequality. The Gini coefficient has become one of the most used measures of inequality as it summarizes the Lorenz curve in an easily understood one-digit index. Not only does the Census Bureau (and CBO) release an annual measure of the Gini for income, this measure is also used in international work (see OECD, Luxembourg Income Study, World Bank, and others).

Some indices are more sensitive to changes at the top (Theil, 1967) or bottom (Atkinson, 1970) of the distribution, but the Gini is the most used measure in the empirical literature.

Other, more recent measures include the share of income received by the top 1 or 5% (see Piketty and Saez, 2003), the decile ratio (90th to 10th percentile values), the ratio of the value at the 90th percentile to the middle (median or 50th percentile) value, and the ratio of the 50th percentile to the 10th (Voitchovsky, 2005; Brandolini and Smeeding, 2009). While all of these measures demonstrate the recent increase in inequality, they all show different levels and rates of increase, with some focusing on changes in the bottom or the top of the distribution as opposed to a single number like the Gini. In this article, we will examine the components of measures of inequality – the 'who, what, where, when, why, and how' of inequality measures, and the ways in which these choices impact the measurement of inequality.

The Who, What, Where, When, Why, and How of Inequality

To determine the economic measurement of inequality, we need to decide whose standard of living we are measuring and

how we are going to measure it and update it over time. The various summary statistics described earlier are only good in so far as one begins with the basics of the measure itself. Far too often in measurement of economic inequality, the beauty of the summary measure far exceeds the accuracy of the concept one is measuring. The following questions must be answered:

- What: Which resource measure is going to be used?
- Who: Whose standard of living is being measured?
- When: What time period is used?
- Where: Do the measures differ by geographic location?
- Why: What is the purpose of these measures?
- How: Which summary statistics and data are going to be used?

What: The Choice of Resource Measure

The most inclusive concept of income and consumption derives from the suggestions of Haig and Simons. Haig (1921) stated that income was "the money value of the net accretion to one's economic power between two points of time" and Simons (1938) defined personal income as "the algebraic sum of (1) the market value of rights exercised in consumption and (2) the change in the value of the store of property rights between the beginning and end of the period in question."

Economists have used the equation that income (Y) equals consumption (C) plus the change in net worth (ΔW) as the working definition of Haig–Simons income ($Y = C + \Delta W$). No household survey, however, has the necessary variables to create a full measure of Haig–Simons income. In an attempt to relate all three components, the Canberra Group Handbook on Household Income Statistics (2011) states: "Household income receipts are available for current consumption and do not reduce the net worth of the household through a reduction of its cash, the disposal of its other financial or non-financial assets, or an increase in its liabilities."

Most studies of income and its distribution include money income, but do not examine changes in assets, and only a few examine the impact of capital gains (e.g., CBO, 2011, 2012; Piketty and Saez, 2003; Smeeding and Thompson, 2011; Wolff et al., 2012). The 'more complete income' (MCI) concept in Smeeding and Thompson (2011) is based on Haig–Simons income and estimated using Survey of Consumer Finance data. They define MCI as earnings and net transfers and include that portion of capital income received as capital gains and royalties. They subtract reported interest, rent, and dividends and include an imputed return to all forms of net worth in order to capture the concept of the change in net worth. Using MCI, Smeeding and Thompson (2011) find a larger concentration of income at the top of the distribution.

There is agreement that inequality increased during the past three decades using various measures of income, but the magnitude of the increase depends on the income measure used and the unit of observation. For example, the Census Bureau estimates that the inequality, using the Gini coefficient, of pre-tax cash income for households (adjusted for family size) increased 19% between 1979 and 2010. CBO (2012) estimates an increase of 21.6% for market income inequality (between 1979 and 2009), and Saez (2012) shows that the share of taxable income obtained by the top 1% doubled between 1979 and 2009. See also Kenworthy and Smeeding (2014) for a summary and comparison of these measures.

Many economists have suggested that consumption may be a more appropriate indicator of permanent income (see Danziger and Taussig, 1979; Cutler and Katz, 1991; Slesnick, 1991, 1994). Johnson and Shipp (1997) were among the first to show different trends in income and consumption inequality. Slesnick (1991) and Cutler and Katz (1991) demonstrated that consumption inequality was lower than income inequality, and that the increase in both income and consumption inequality was similar during the 1980s. Later, Krueger and Perri (2006) identify the divergent trends in income and consumption inequality during the 1990s. Most recent research shows that consumption inequality is less than income inequality, and its increase is less than the increase in income inequality (see Krueger and Perri, 2006; Heathcote et al., 2010; Meyer and Sullivan, 2013; Fisher et al., 2013).

Then wealth or net worth itself may be used as the measure of inequality. Here the evidence is more sketchy and dependent on valuing both private and public pensions as well as businesses and financial holdings. For recent studies of wealth inequality, see Wolff (2012) and Pfeffer et al. (2013).

Who: Whose Standard of Living?

To examine the distribution of income, researchers need to determine whose income to evaluate or the unit of analysis. Many measures of the standard of living use household (or family) resources, while others use individual earnings or income. Households with more members may need additional income to be equally well-off as smaller households. Most studies use family or household income and adjust it by an 'equivalence scale' to obtain equivalent family resources. Some studies using income tax data are limited to tax units, which suggest a much smaller unit for income sharing (Piketty and Saez, 2003). A measure of well-being for individuals is obtained by using the family size (multiplied by the household's sample weight) as a weight. This method assumes that all members of the household have and share the same level of resources (Johnson, 2004). Earlier research has shown that the choice of equivalence scale can have dramatic effects on both the level of and trend in the standard of living (Coulter et al., 1992; Johnson et al., 1998). This literature has focused on the effect that equivalence scales have on both inequality and poverty measures.

One of the most common scales is the single-parameter constant-elasticity equivalence scale reviewed by Buhmann et al. (1988) and Ruggles (1990), which is used most often in international comparisons of inequality (see Atkinson et al., 1995). In general, the constant elasticity scales are given by (family size)^e, in which e is the scale elasticity. Notice that if the elasticity equals one, then the scale equals family size, there are no assumed economies of scale in living arrangements, and the equivalent resources are simply the per-capita resources. Alternatively, if the elasticity equals zero then there is no adjustment for family size, there are complete economies of scale in living and the marginal cost

of another person is zero. Following the suggestions by Ruggles (1990) and Buhmann et al. (1988), much research uses an elasticity of 0.5. This scale indicates that the resources for a two-person family must be 41% more than that of a single-person family for the two-person family to have the same standard of living as the single-person family.

Some scales adjust for more than just the number of adults and children in the family. For instance, the scales implicit in the poverty thresholds adjust for the age of the household head, and the new three-parameter scale (see Betson, 1996; Short et al., 1999) used in the Census's Supplemental Poverty measure provides more economies of scale between singles and childless couples, which are more similar to those in the poverty scale. Slesnick (1994) presents scales that depend on family size, the age, race, and gender of the household head, and regional location and the farm/nonfarm status of the household.

The choice of equivalence scale can impact the level and trend of inequality. Coulter et al. (1992) demonstrate that the constant elasticity equivalence scale yields a U-shaped relationship between inequality and the scale parameter, *e*. Johnson (2004) compares alternative scales and shows that the choice of scale has a dramatic effect on the relative standard of living of different families. A lower equivalence scale implies that the family's resources will be adjusted upward, hence, increasing their equivalent resources. For example, single adults will have a higher standard of living relative to the reference group under the three-parameter scales (compared to the poverty scales). In addition, the Slesnick scales produce a higher standard of living for single parents than any of the other alternatives. Jenkins (1994) and Johnson (2004) also show that different constant elasticity scales yield different changes in inequality over time.

When: Choosing a Time Period and Adjusting the Standard of Living over Time

Any measure of the standard of living must be based on particular reference period and be adjusted by the changes in the cost-of-living that occur during the time period. In discussing poverty measures, Foster (1998: p. 337) claims that "the key distinction between absolute and relative thresholds is not seen in the specific values at a given date, but in how the values change as the distribution changes." Most studies use annual measures of income. However, consumption measures that are more similar to permanent income (or average income over a period of years) yield less inequality. Alternatively, monthly or quarterly income yields higher inequality. Hence, the period of time matters. The Canberra Report (2011) suggests using annual income, which has become the norm, even if it is likely too long a period for low-income families and too short for measures of permanent income, especially for higher-income households (Gottschalk and Smeeding, 1997).

When measuring 'real' living standards, not just income shares, cost of living adjustments are frequently constructed using a price index such as the CPI. The literature on the CPI suggests that it may be over adjusting the cost of living, that is, the real changes in the standard of living may be larger than those found using the CPI (see Boskin et al., 1996; Meyer and Sullivan, 2011). Other research shows that absolute poverty rates and changes in median resources are also affected by the price index used (see Slesnick, 2001; Meyer and Sullivan, 2011).

While different price indexes (or cost of living adjustments) lead to a variety of changes in the *level* of the median and changes in the levels of various quintiles, they do not affect the *trend* in inequality. Slesnick (2001), however, shows that using household-specific price indexes can have a small effect on the trend in inequality. Johnson (2004) and Fixler and Johnson (2013) show the impacts of price indexes on the changes in the median income. However, inequality will be impacted only by using price indexes that vary with income. Recent suggestions in Broda and Weinstein (2008) and Gordon and Dew-Becker (2008) suggest that price changes may be lower for lower-income groups, which would lead to a smaller increase in inequality, but there is a question as to whether they properly measure changes in prices for services which are consumed by the poor.

Where: How Are the Measures Adjusted for Differences in Geography?

Similar to the adjustment for family size, it could be that some areas in the country (or even different countries) could have different costs. Slesnick (2001) uses equivalence scales that depend on the region and the urbanity of the geographic area. The Bureau of Economic Analysis (BEA) has constructed a regional price parity index to adjust personal income and show that the range across counties narrows, and hence, inequality across the country is less than using the unadjusted income. Moretti (2013) and Deaton (2010) discuss other implications of using geographically adjusted income in comparing wages across the US and income internationally. Especially in making international comparisons, the recent Canberra Report (2011) suggests using price index adjustments, and highlights an example from Canada, which uses interarea price indexes.

How: Which Summary Statistic and Data Source?

As we have suggested, much research focuses on the levels and trends in the Gini coefficient to evaluate inequality, but Atkinson (1970), Sen (1997), and Theil (1967) evaluate a variety of inequality measures and their properties. As shown in the Census Current Population Report (DeNavas-Walt et al., 2013), the changes over the past 30 years have been similar for all four measures. Others focus on the ratio of particular percentiles (e.g., 90th/10th percentile ratio, or 90th/50th). Often, techniques are used to estimate differences for the changes at the alternative ends of the distribution, such as a Pareto distribution for the upper tail of income which is often under-reported in surveys. In some comparisons across countries or survey years, arbitrary high and low values are imposed on survey data. For example, incomes may be bounded from below at zero (or at a very small number) thereby omitting negative incomes (as in self-owned businesses which suffer a loss) or bounded above at some multiple of median or mean incomes to protect confidentiality in the top of the distribution (Atkinson et al., 1995).

The final choice is that of data set. The Census Bureau (see DeNavas-Walt et al., 2013) and CBO use the Current Population Survey (CPS) to create a comprehensive income measure, and several other top income data sources can be used to estimate inequality trends. These include income tax records from the IRS (analyzed by Piketty and Saez, 2003; Saez, 2012; Atkinson et al., 2011) and the Federal Reserve Board's Survey of Consumer Finances (SCF). Analysis using these data sources suggests that income inequality has risen dramatically at the very top of the distribution. The analysis by Saez (2012) of IRS data through 2010 finds that the share of federal pre-tax Adjusted Gross Income held by the richest 1% of households more than doubled between 1979 and 2007. The CBO 'comprehensive income' measure (not adjusted for taxes) shows that the top 1% share of total income rose from 9.3 to 19.4% over the 1979–2007 period. Smeeding and Thompson (2011) use the SCF data to calculate a 'more comprehensive income' (MCI) measure which combines standard income flows with imputed income to assets. They show that the top 1% share of MCI rose from 18% in 1989 to 22% in 2007. Armour et al. (2013) compute implicit capital gains and losses each year from 1989 to 2007 and show that inequality declines and the top share falls using their techniques.

Impacts of Choices

Figure 1 compares the changes in the income received by quintiles using six measures of income. This figure illustrates how many of the measurement choices discussed earlier impact the level and distribution of income, including the choice of resource measure, equivalence scale, inequality measure, and price index. The level of inequality is often measured using the Gini index or a variant of the Lorenz curve using the share of income for each quintile. This figure shows the increase in mean income for each quintile and demonstrates an increase in inequality since higher quintiles experienced larger increases in income during this time period (portraying a staircase pattern). Many others have shown similar results (DeNavas-Walt et al., 2013; CBO, 2012; Kenworthy and Smeeding, 2014; Burkhauser et al., 2010).

The six measures in Figure 1 include the Census money income and equivalized income measure (DeNavas-Walt et al., 2013), two measures of the CBO after-tax income with alternative methods for valuing health care benefits (see CBO, 2012), and two measures from Burkhauser et al. (2010). The extent of the increase in inequality differs across the six measures of income, with the largest increase occurring for the CBO after-tax income measure with the fungible value of Medicare and Medicaid.

Looking at the changes for the middle quintile illustrates the variety of results regarding the economic well-being of the nation. Burkhauser et al. (2010) and Meyer and Sullivan (2011) suggest that the median income has increased more than what others suggest, mainly referring to the changes presented by the US Census Bureau (DeNavas-Walt et al., 2013). Evaluating the changes for the middle quintile highlights the impacts of using equivalence scales. Due to the fall in average household size during this period, the increase in equivalized income is larger here than for household income. In addition, because of the

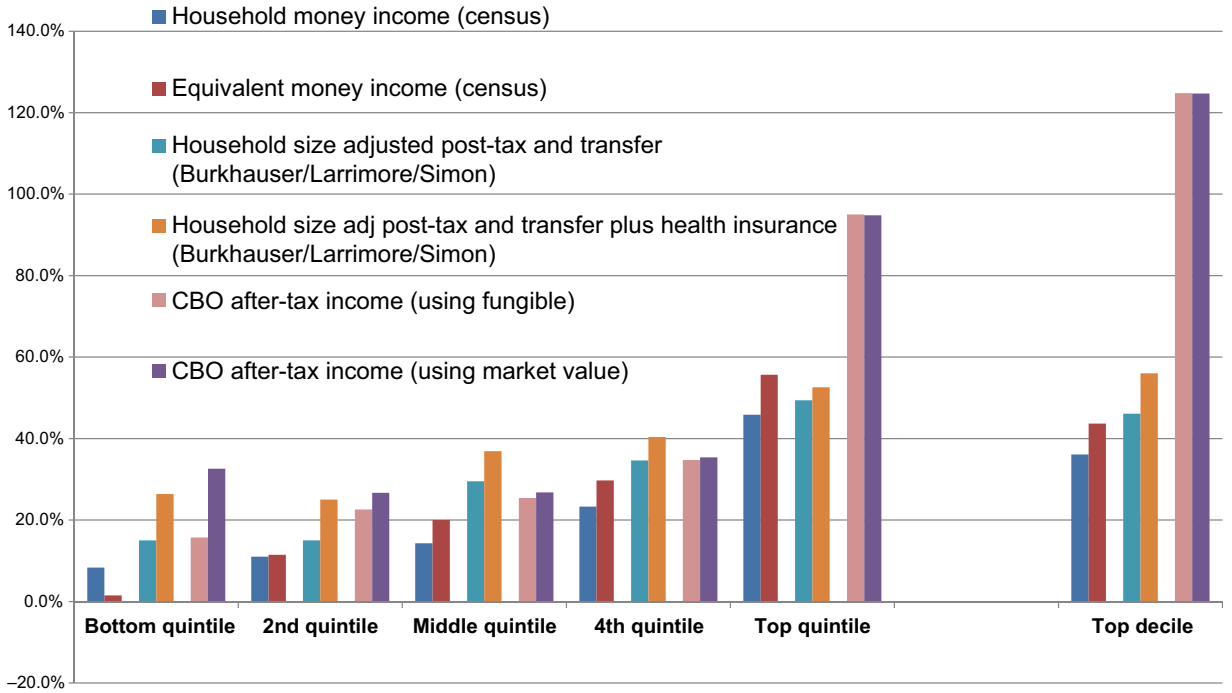

Figure 1 Percentage growth in mean quintile income using alternative definitions of income (adjusted for inflation using CPI-U-RS), 1979–2007. Congressional Budget Office, 2012. The Distribution of Household Income and Federal Taxes, 2008 and 2009. CBO, report; Burkhauser, R., Larrimore, J., Simon, K., 2010. A "Second Opinion" on the Economic Health of the American Middle Class. NBER Working Paper 17164, National Bureau of Economic Research; DeNavas-Walt, C., Proctor, B.D., Smith, J.C., 2013. Income, Poverty, and Health Insurance Coverage in the United States: 2012. US Census Bureau, Current Population Reports, P60–245, US Government Printing Office, Washington, DC.

increases in Medicare and Medicaid as well as in employer subsidized insurance, including these in income at their cost yields a larger increase in income at the median. Finally, using a price index that shows less inflation than the CPI-U-RS (such as the PCE deflator used in CBO, 2012) would yield a larger increase in the median income.

This figure also shows changes in inequality. All measures display a staircase pattern, with each higher quintile showing a larger increase in income, regardless of concept or measure. However, there are differences. For example, the equivalized income shows a larger increase in inequality during this period (see also DeNavas-Walt et al., 2013; Kenworthy and Smeeding, 2014). Since Medicaid and Medicare benefits are received by households with lower income, and since employer health insurance subsidies are constant across all covered workers in a company, these subsidies are a larger fraction of compensation for lower wage employees. It follows that there are larger increases in resources at the lower quintiles and hence, a smaller increase in inequality when one includes the actual cost of health insurance in such estimates. The method used to account for Medicare and Medicaid benefits affect the change in inequality. Recently, the CBO switched to using the cash value (or actual cost) of benefits instead of the fungible value, which places a lower value of benefits for lower-income households.

Finally, the data source matters, especially with respect to the treatment of very high income households. As can be seen in the changes for the top decile, the CBO measures show much larger increases. This is because the other four measures use the CPS household survey while the CBO adjusts the household survey with tax records to account for the under-representativeness of high-income households.

To convert this information into an inequality measure and show the changes in inequality over time, one would use the share of income for each quintile. The changes in these shares would show much more dramatic differences by quintile, with the top quintile increasing its share of income between 1979 and 2007. One can use these shares to compare the Gini coefficients over time. **Figure 2** shows five alternative Gini coefficients. Two of them use similar data shown in **Figure 1** – the CPS and the CBO measures. The CBO measure is the before-tax measure more comparable to the Census money income. While the changes between 1979 and 2009 are similar for these two measures, between 1979 and 2007 the CBO measure increases more than the CPS measure because of the treatment of high-income households.

The other three are from two different data sources. The MCI measure is from Smeeding and Thompson (2011) and represents a comprehensive after-tax measure that includes financial gains, while the disposable income and consumption measures are from Fisher et al. (2013). As shown in **Figure 2**, the trends in the after-tax measures are similar, and over the entire 30-year period the trends in all income measures are similar. As expected, the Gini for consumption is lower than that for income, but increases at about 75% the rate of income inequality.

Why: The Purpose of the Measure

Each of the choices – who, what, where, when, and how – have different implications for the levels and trends of inequality. As

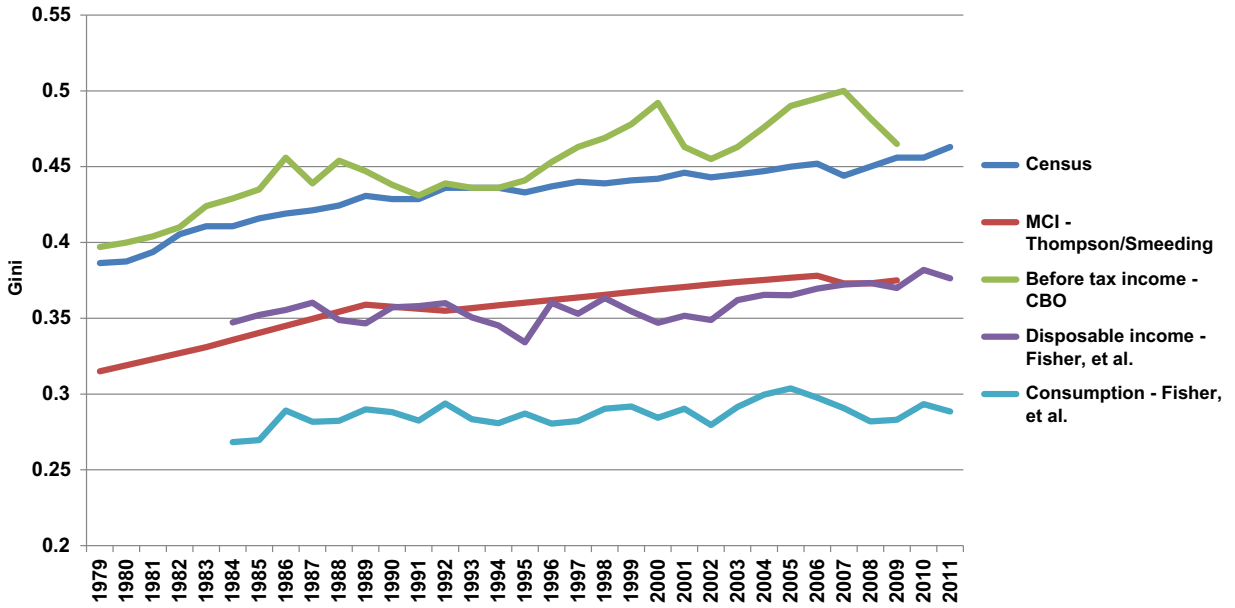

Figure 2 The Gini coefficient using various income measures and a consumption measure. DeNavas-Walt, C., Proctor, B.D., Smith, J.C., 2013. Income, Poverty, and Health Insurance Coverage in the United States: 2012. US Census Bureau, Current Population Reports, US Government Printing Office, Washington, DC, pp. 60–245; Smeeding, T.M., Thompson, J.P., May 2011. Recent trends in the distribution of income: labor, wealth and more complete measures of well-being. Research in Labor Economics, 1–49; Congressional Budget Office, 2011. Trends in the Distribution of Household Income between 1979 and 2007. CBO, report; Congressional Budget Office, 2012. The Distribution of Household Income and Federal Taxes, 2008 and 2009. CBO, report; Fisher, J., Johnson, D., Smeeding, T., 2013. Measuring the trends in inequality of individuals and families: income and consumption. American Economic Review 103(3), 184–188.

a result, researchers need to decide on the purpose of their evaluation when making these choices, whether it is the overall changes in inequality, the redistributive impacts of government taxes and transfers, or the differences between annual and permanent income. Each of these questions may require different data and inequality measures. One should never mix and match measures, as this will provide a misleading picture of the levels and trends in inequality. For example, Leonhardt (2012) combined Census, CBO, and the Saez measures of income to evaluate changes in particular percentiles in one diagram.

Many researchers attempt to make cross-national comparisons of inequality (see OECD, 2011; Brandolini and Smeeding, 2009). As mentioned earlier, the Canberra report recommends a common definition of income along with similar choices in equivalence scales and price indexes in order to make international comparisons. The LIS (see http://www.lisdatacenter.org/) Income Measure is similar in spirit to the Canberra recommended measure. However, many of the alternative choices presented earlier are not all available for international comparisons and researchers need to be aware of these differences.

One common thread is to try to understand the forces which drive inequality: trade, skill-biased technical change; the growth of the capital share of national income relative to the labor share; 'winner take all' markets; and so on. Simple generalizations on causes are not easily found in this literature. Another purpose is to evaluate the effects of inequality on things we care about, such as health, voting, lifespan, happiness, economic growth, equality of opportunity, and intergenerational mobility. Kenworthy and Smeeding (2014) provide a very brief summary of this growing literature.

Conclusion

In this brief survey, we have attempted to convey some of the issues involved in measuring economic inequality and its trends. Longer treatments can be found in the *Oxford Handbook of Economic Inequality* (Salverda et al., 2009) and the *Elsevier Handbook of Income Distribution* (Atkinson and Bourguignon, 2000). Suffice it to say that with inequality rising in most rich nations (for example, see OECD, 2011) and despite the widespread coverage of the top 1% of income earners, we still have much to learn about inequality, its source, and its origins. It is indeed time to bring inequality back into the fold of mainstream economics as Atkinson (1997) has suggested.

Disclaimer

Any views expressed in this article are those of the authors and not necessarily those of the US Bureau of Economic Analysis or the Institute for Research on Poverty.

See also: Earnings Inequality; Education and Income Distribution; Inequality Measurement; Social Inequality and Schooling; Social Inequality in History (Stratification and Classes).

Bibliography

Aaron, H.J., 1978. Politics and the Professors. Brookings Institution, Washington, DC.

Armour, P., Larrimore, J., Burkhauser, R.V., 2013. Deconstructing income and income inequality measures: a crosswalk from market income to comprehensive income. American Economic Review 101 (3), 173–177.

Atkinson, A., Piketty, T., Saez, E., 2011. Top incomes in the long run of history. Journal of Economic Literature 49 (1), 3–71.

Atkinson, A.B., Bourguignon, F.J., 2000. Handbook of Income Distribution. North-Holland/Elsevier, Oxford.

Atkinson, A.B., Rainwater, L., Smeeding, T., October 1995. Income Distribution in OECD Countries: Evidence from the Luxembourg Income Study (LIS). Social Policy Studies, vol. 18. OECD, Paris.

Atkinson, A.B., 1970. On the measurement of economic inequality. Journal of Economic Theory 2 (3), 244–263.

Atkinson, A.B., 1997. Bringing income distribution in from the cold. The Economic Journal 107 (441), 297–321.

Betson, D., 1996. Is Everything Relative? The Role of Equivalence Scales in Poverty Measurement. University of Notre Dame, Memo.

Boskin, M., Dulberger, E., Gordon, R., Grilliches, Z., Jorgenson, D., 4 December 1996. Toward a More Accurate Measure of the Cost of Living. Final Report to the Senate Finance Committee.

Brandolini, A., Smeeding, T.M., 2009. Income inequality in richer and OECD countries. In: Salverda, W., Nolan, B., Smeeding, T.M. (Eds.), Oxford Handbook of Economic Inequality. Oxford University Press, Oxford, UK, pp. 71–100.

Broda, C., Weinstein, D.E., 2008. Prices, Poverty, and Inequality: Why Americans are Better off than You Think. AEI Press, Washington, DC.

Buhmann, B., Rainwater, L., Schmauss, G., Smeeding, T.M., 1988. Equivalence scales, well-being, inequality, and poverty: sensitivity estimates across ten countries using the Luxembourg income study database. Review of Income and Wealth 34, 115–142.

Burkhauser, R., Larrimore, J., Simon, K., 2010. A "Second Opinion" on the Economic Health of the American Middle Class. NBER Working Paper 17164. National Bureau of Economic Research.

Canberra Group, 2011. Handbook on Household Income Statistics, second ed. United Nations Economic Commission for Europe.

Citro, C.F., Michael, R.T. (Eds.), 1995. Measuring Poverty: A New Approach. National Academy Press, Washington, DC.

Congressional Budget Office, 2011. Trends in the Distribution of Household Income between 1979 and 2007. CBO report.

Congressional Budget Office, 2012. The Distribution of Household Income and Federal Taxes, 2008 and 2009. CBO report.

Coulter, F.A.E., Cowell, F.A., Jenkins, S.P., 1992. Equivalence scale relativities and the extent of inequality and poverty. The Economic Journal 102, 1067–1082.

Council of Economic Advisors, 2012. Economic Report of the President. Government Printing Office.

Cutler, D., Katz, L., 1991. Macroeconomic performance and the disadvantaged. Brookings Papers on Economic Activity 2.

Dalton, H., 1920. The measurement of the inequality of income. Economic Journal 30, 348–361.

Danziger, S., Taussig, M., 1979. The income unit and the anatomy of income distribution. Review of Income and Wealth 25 (4), 365–375.

Deaton, A., 2010. Price indexes, inequality, and the measurement of world poverty. American Economic Review 100 (1), 5–34.

DeNavas-Walt, C., Proctor, B.D., Smith, J.C., 2013. Income, Poverty, and Health Insurance Coverage in the United States: 2012. U.S. Census Bureau, Current Population Reports, pp. 60–245. U.S. Government Printing Office, Washington, DC.

Fisher, J., Johnson, D., Smeeding, T., 2013. Measuring the trends in inequality of individuals and families: income and consumption. American Economic Review 103 (3), 184–188.

Fixler, D., Johnson, D., 2013. Accounting for the distribution of income in the U.S. national accounts. In: Jorgenson, D., Landefeld, J.S., Schreyer, P. (Eds.), Measuring Economic Stability and Progress. University of Chicago Press, Chicago, IL.

Foster, J., 1998. Absolute versus relative poverty. American Economic Review 88 (2), 335–341.

Gini, C., 1921. Measurement of inequality of incomes. Economic Journal 30, 124–126.

Gordon, R., Dew-Becker, I., 2008. Controversies about the Rise of American Inequality: A Survey. NBER Working Paper, No. 13982. National Bureau of Economic Research.

Gottschalk, P., Smeeding, T.M., 1997. Cross-national comparisons of earnings and income inequality. Journal of Economic Literature 35, 633–687.
Haig, R.M., 1921. The concept of income – economic and legal aspects. In: Haig, R.M. (Ed.), The Federal Income Tax. Columbia University Press, New York, pp. 1–28.
Heathcote, J., Perri, F., Violante, G., 2010. Unequal we stand: an empirical analysis of economic inequality in the U.S., 1967–2006. Review of Economic Dynamics 13, 15–51.
Jenkins, S.P., 1994. Accounting for inequality trends: decomposition analyses for the UK, 1971–86. Economica 62, 29–63.
Johnson, D., Shipp, S., 1997. Trends in inequality in the United States using consumption expenditures: the U.S. from 1960–1993. Review of Income and Wealth 43 (2), 133–152.
Johnson, D., 2004. Using expenditures to measure the standard of living in the United States: does it make a difference? In: Wolff, E.N. (Ed.), What has Happened to the Quality of Life in America and Other Advanced Industrialized Nations. Edgar Elgar Press, Northhampton, MA, pp. 27–48.
Johnson, D., Short, K., Garner, T., October 1998. Poverty Measurement Research at the Bureau of the Census and the Bureau of Labor Statistics. Paper presented at the Annual Research Conference of the Association for Policy Analysis and Management.
Kenworthy, L., Smeeding, T.M., 2014. The United States: high and rapidly rising inequality. In: Brian, N., et al. (Eds.), Changing Inequalities and Societal Impacts in Rich Countries: Thirty Countries' Experiences. Oxford University Press.
Krueger, D., Perri, F., 2006. Does income inequality lead to consumption inequality? Evidence and theory. Review of Economic Studies 73, 163–193.
Krugman, P., 1992. The right, the rich and the facts. The American Prospect 11, 19–31.
Kuznets, S., 1955. Economic growth and income inequality. The American Economic Review 45 (1), 1–28.
Leonhardt, D., 6 August 2012. A Slowdown in Growth, an Increase in Income Inequality. New York Times.
Lorenz, M.O., 1905. Methods for measuring concentration of wealth. Journal of the American Statistical Association 9 (70), 209–219.
Meyer, B., Sullivan, J., 2013. Consumption and income inequality and the great recession. American Economic Review 103 (3), 178–183.
Meyer, B., Sullivan, J., 2011. The Material Well-being of the Poor and Middle Class Since 1980. AEI Working Paper #2011-04. American Enterprise Institute.
Moretti, E., 2013. Real wage inequality. American Economic Journal: Applied Economics 5 (1), 65–103.
OECD, 2011. Divided We Stand: Why Inequality Keeps Rising. OECD Publishing.
Pareto, V., 1897. Cours D'economie Politique. Rouge and Pichon, Lausanne and Paris.
Pen, J., 1971. Income Distribution. Penguin, Harmondsworth.
Persons, W., 1909. The variability in the distribution of wealth and income. The Quarterly Journal of Economics 23 (3), 416–449.
Pfeffer, F.T., Danziger, S.H., Schoeni, R.F., 2013. Wealth disparities before and after the great recession. Annals of the American Academy of Political and Social Science 650 (1), 98–123.
Piketty, T., Saez, E., 2003. Income inequality in the United States, 1913–1998. Quarterly Journal of Economics 118 (1), 1–39.
Ricardo, D., 1821. On the Principles of Political Economy and Taxation. John Murray, London.
Ruggles, P., 1990. Drawing the Line. The Urban Institute Press, Washington, DC.
Saez, E., 2012. Striking it Richer: The Evolution of Top Incomes in the United States (Updated with 2009 and 2010 Estimates). Manuscript.
Salverda, W., Nolan, B., Smeeding, T.M., 2009. The Oxford Handbook of Economic Inequality. Oxford University Press, Oxford.
Sawhill, I.V., 2012. 2011: The Year that Income Inequality Captured the Public's Attention. Brookings Institution. Up Front blog post, April 19.
Sen, A., 1997. On Economic Inequality. Clarendon Press, Oxford.
Short, K., Garner, T., Johnson, D., Doyle, P., 1999. Experimental Poverty Measures: 1990 to 1997. U.S. Census Bureau. Current Population Reports, Consumer Income. U.S. Government Printing Office, Washington, DC, pp. 60–205.
Simons, H., 1938. Personal Income Taxation: The Definition of Income as a Problem of Fiscal Policy. University of Chicago Press, Chicago, IL.
Slesnick, D.T., 1991. The standard of living in the United States. Review of Income and Wealth 37 (4), 363–386.
Slesnick, D.T., 1994. Consumption, needs and inequality. International Economic Review 35 (3), 677–703.
Slesnick, D.T., 2001. Consumption and Social Welfare: Living Standards and Their Distribution in the United States. Cambridge University Press, Cambridge.
Smeeding, T.M., Thompson, J.P., May 2011. Recent trends in the distribution of income: labor, wealth and more complete measures of well-being. Research in Labor Economics, 1–49.
Smith, A., 1761. The Theory of Moral Sentiments. A. Millar, London.
Theil, H., 1967. Economics and Information Theory. Rand McNally and Company, Chicago, IL.
Voitchovsky, S., 2005. Does the profile of income inequality matter for economic growth? Journal of Economic Growth 10 (3), 273–296.
Watkins, G.P., March 1908. An interpretation of certain statistical evidence of concentration of wealth. Publications of the American Statistical Association 11 (81), 27–55.
Wolff, E., Zacharias, A., Masterson, T., Eren, S., Sharpe, A., Hazell, E., 2012. A Comparison of Inequality and Living Standards in Canada and the United States Using an Expanded Measure of Economic Well-being. Levy Economics Institute of Bard College Working Paper #703.
Wolff, E., 2012. The Asset Price Meltdown and the Wealth of the Middle Class. NBER Working Paper No. 18559. National Bureau of Economic Research.

Income Inequality

Dennis Gilbert, Hamilton College, Clinton, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article first defines income and describes the sources of income data, the issues surrounding the measurement of household income, and the most common measures of income inequality. It then examines household income trends in affluent countries since the 1970s and shows that the distribution of household income has become more unequal in most, but especially in the US and UK. It includes that Skill Biased Technological Change, globalization, and the waning influence of wage-setting institutions, along with shifts in household composition and behavior, are driving the rising inequality of household incomes in affluent countries.

Income may be defined as the amount of money and other resources received by an individual or household over a specified period of time (e.g., $2000 per month, $500 000 in 2015). It is distinguished from wealth, which is the value of assets held at a point in time (e.g., $250 000 on 30 June 2015). In recent years, there has been increasing political and social science interest in the distribution of income, stimulated by evidence of growing inequality among households in affluent countries, the central focus of this article. Concern with the distribution of income is, for many, rooted in questions of social justice: should societies permit a few to live opulently, while others struggle to attain the basics of food, clothing, shelter, health care, and education? Does the skewed distribution of income fairly reflect differences in the contributions individuals make to the common welfare? But interest in income distribution is also encouraged by more practical concerns with the possible effects of rising income inequality on economic growth, democratic governance, social mobility, physical and mental health, family stability, and rates of violent crime (Wilkinson and Pickett, 2009; Hacker and Pierson, 2010).

Sources of Income

Households receive income from one or more of the following sources: (1) earnings from jobs held by one or more members of the household; (2) self-employment income from small businesses and professional practices; (3) capital income in the form of interest, dividends, rents, royalties, or capital gains derived from financial assets and real property owned by the household; (4) public transfers such as social security benefits, farm subsidies, and assistance to low-income households; (5) private transfers including pension payments, alimony, and gifts. Job earnings are the major income source for most nonelderly households, but the weight of job earnings and other income sources varies across the distribution. If all the households in a typical developed country were lined up in a parade, from lowest to highest income, it would be found that those at the beginning of the parade are the most dependent on public transfers and hold one job or none; that households in the middle and toward the end of the parade are likely to have two employed members; and that households at the very end of the parade receive a significant, if not the predominant share of their income from self-employment and capitalist sources.

Income Data and Measurement

A household, the basic unit of analysis in most income studies, may consist of a family, unrelated persons residing together, or an individual living alone. The two basic sources of income data are household surveys, such as the US Current Population Survey or the French Household Budget Survey, and administrative records, especially those based on income tax returns. Annual government surveys are the most comprehensive income data source. They are typically based on large samples of the national population and collect essential demographic and economic information from households. A subset of surveys, such as the Panel Study of Income Dynamics in the US and the European Community Household Panel, track smaller samples of households over a period of years to understand long-term variation in income.

Surveys tend to underestimate income at the upper end of the distribution, presenting a problem for studies of income inequality. Tax records, based on legally defined tax reporting units, provide better coverage of high incomes, but are otherwise less comprehensive, since not everyone pays income taxes and those who do are not asked to provide detailed information about themselves. Synthetic estimates of the income distribution, such as those developed by the US Congressional Budget Office, integrate data from both sources (www.cbo.gov/publication/44604).

The LIS Cross-National Data Center (formerly, the Luxembourg Income Study) collects national income data sets from upper and middle-income countries and makes it available to researchers in a compatible, microdata format. LIS is the source of much of the data employed in comparative studies of income distribution. The LIS website (www.lisdatacenter.org) also provides comparative summary income statistics ('Key Figures'), which will be used in this article.

Measuring income for research purposes presents a series of conceptual and practical issues, including the appropriate time frame, unit of analysis, and elements to be included in the definition of income. The guiding assumption in these matters is that income is intended as a measure of living standard or

relative welfare. Annual household income is the key variable in most studies. The use of annual income is problematic because incomes are known to vary more from year to year than household consumption does. Income averaged over several years (available in some cases from panel studies) might be a superior indicator, but the data that most researchers depend on are collected in official annual surveys. Household income, pooled from multiple sources, is regarded the best available indicator of the welfare of household members, even if it cannot be assumed that income is shared equally among them. Some studies, including those based on LIS data, adjust incomes for household size, assuming that, at a given income level, the members of larger households will have a lower standard of living; adjustment factors allow for economies of scale – families big and small, for example, only need one kitchen.

Analysts may distinguish between market income and disposable income. In broad terms, the former is the sum of income from job earnings, self-employment, and capital sources, while the latter – believed to be a better welfare measure – is market income plus public transfers, less direct taxes. Disposable income may incorporate other items, including the value of noncash transfers, such as food stamps or subsidized housing, and the imputed value of rent for owner-occupied housing.

Measuring Inequality

There are several common ways of measuring income inequality. One approach looks at the proportional distribution of income shares across segments of the population. For example, in the United States in 2012, the bottom quintile of households received about 3% of all household income, while the richest quintile absorbed 51%; only the fourth quintile, at 23%, received close to a proportionate share (U.S. Census, 2013: 40). A telling measure of concentration at the top is the share of income flowing to the richest 1% of households. Another inequality measure, percentile ratios, permit comparisons for specific ranges of the distribution. For example, the ratio between incomes at the 90th and 50th percentiles in United Kingdom rose from 1.764 to 2.126 between 1974 and 2010, indicating growing inequality in the upper half of the distribution (LIS Key Figures).

The Gini index is the most frequently used of several summary measures of inequality derived from the Lorenz curve, which plots cumulative income shares against cumulative percentage of the population. A hypothetical Lorenz curve following an upward slanting 45° diagonal would indicate perfect equality (the bottom 10% receives 10% of all income, the bottom 50% receives 50%, etc.) The Gini coefficient is a measure of the degree of deviation from the diagonal, indicating the extent of income inequality. It ranges from 0 (perfect equality) to 1 (perfect inequality). The Gini coefficients for Denmark, the United States, and Mexico in 2004, were, respectively, 0.228, 0.372, and 0.457 (LIS Key Figures).

Trends in Income Inequality

Three broad generalizations can be made about the changing distribution of income in affluent countries. First, income inequality, by diverse measures, has increased in most high-income countries since the 1970s. The extent, timing, and exact pattern of rising inequality vary among countries, but the general direction of change is unmistakable. Second, this growing inequality has typically been marked by income polarization, with fewer households in the middle-income range (commonly defined as incomes between 75 and 125% of the median income) and more at the higher or lower levels (Some have characterized this phenomenon as the disappearance of the middle class, a notion that conflates middle class with middle income. In fact, many workers with what would be considered middle-class occupations enjoy incomes well beyond the median). Third, the rising inequality of recent decades is a reversal of the declining inequality of the post-World War II era – a shift often described as the great U-turn. Here the evidence is thinner but suggestive, since comparable income data for the earlier period is often lacking.

Table 1 records income inequality levels and trends for 15 affluent countries for which comparable LIS statistics were available. Countries are ranked in descending order by Gini coefficient for the last year of the period indicated. These numbers are based on disposable income, adjusted for household size. Corresponding market income coefficients would, of course, be higher, since they would not reflect the equaling effects of progressive taxation and government provided social benefits including retirement support. For example, the market income Gini coefficient for 24 mostly high-income OECD countries in the mid-2000s was 0.450, compared to a disposable income coefficient of 0.300 (OECD, 2008: 53107).

The right hand column in the table shows the percentage change in each country's Gini index from a point in the late-1970s or 1980s to the mid-2000s. Among the nations listed, the English-speaking countries, especially the US and UK, have high levels of inequality and the Scandinavian countries have relatively low levels. All except 4 of 15 countries have experienced rising inequality over the last two to three decades. In most of these countries, the proportion of the population in the middle-income range, as defined above, shrank between

Table 1 Inequality of disposable income in affluent nations

	Gini	*Percentage Change*
United States 1979–2007	0.378	26
United Kingdom 1979–2007	0.357	34
Ireland 1987–2007	0.328	−9
Spain 1980–2007	0.318	−4
Australia 1987–2003	0.312	11
France 1978–2005	0.312	−10
Italy 1986–2008	0.306	1
Canada 1975–2007	0.289	1
Luxembourg 1985–2007	0.276	16
Austria 1987–2004	0.269	19
Finland 1987–2007	0.264	28
Germany 1978–2007	0.263	1
Norway 1979–2004	0.256	14
Denmark 1987–2004	0.255	−11
Netherlands 1983–2007	0.252	1
Sweden 1981–2005	0.197	20

Source: LIS Key Figures.

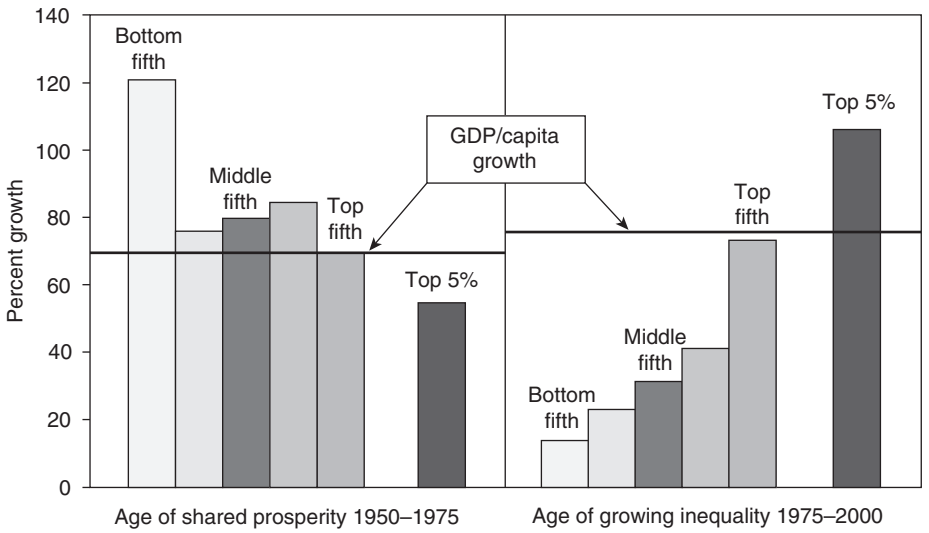

Figure 1 Growth of family income and GDP per capita, 1950–2000. Reproduced from Gilbert, D. 2015. The American Class Structure in an Age of Growing Inequality. Sage, Los Angeles, p. 95.

1985 and 2004 (Atkinson and Bandolini, 2013: 85; Alderson and Doran, 2013).

The Anglophone countries, with their high and growing income inequality, were especially notable for the rising share of income claimed by the top 1% since 1970. The extreme case was the United States, where the income share of the richest percentile climbed from under 10% in the mid-1970s to over 20% by the beginning of the new century. In contrast, the share of the top percentile was relatively flat in continental Europe and Japan over that same period (Piketty and Saez, 2013).

Figure 1 describes the great U-turn phenomenon for the United States with the shifting pattern of income growth before and after 1975. It shows that the prosperity of the postwar period was broadly shared among families, with the bottom quintile experiencing the fastest growth and the top 5% the slowest. The pattern of the second period, beginning in 1975, is an exaggerated mirror image of the first. The growth of family incomes became steeply stratified, with the top 5% enjoying a distinct advantage. Remarkably, the growth of GDP per capita during both of these 25-year periods was about the same, close to 75%, but the benefits of growth were distributed very differently. Sweden, Finland, West Germany, and, to a more modest degree, Canada also experienced a U-turn, traced by the path of the Gini coefficient, with declining inequality through the early 1970s into the early 1980s and rising inequality thereafter; France, an outlier, saw declining income disparities from the early 1970s to the end of the twentieth century (Brandolini and Smeeding, 2009: 88–89; Alderson and Nielsen, 2002).

Growing Inequality: Labor Market Factors

Increasing wage inequality among men has been the primary driver of the growing inequality in household incomes. From 1980 into the twenty-first century wage dispersion increased for both men and women, but especially for men, in most developed countries. It appears that technological change and globalization have narrowed prospects for lower-skilled workers in these countries, while opening new opportunities for those with sophisticated skills. In manufacturing, digitally controlled robotic technology, outsourcing of labor-intensive intermediate goods, and cheap imports from low-wage countries have sharply reduced demand for semiskilled production workers and constrained their wages. Computerization of routine office tasks and the 'offshoring' of some through the Internet have had a similar, if less dramatic effect on demand for clerical workers. Many unskilled workers were channeled into the low wage, routine, service sector jobs that could not be offshored. In contrast, at the high end of the labor market, those with valued technical, analytic, and managerial skills – from corporate executives to software engineers and financial analysts – have been the winners in a changing economy. The most commonly cited indicator of these trends is the growing wage gap between high school educated workers and those with college and graduate degrees, especially notable in the US and the UK. Another indicator is the widening skill disparity in unemployment rates, especially among the young in some continental European countries (Katz and Autor, 1999; Blau and Kahn, 2009; Liu and Grusky, 2013).

Economists commonly describe this transformation of the labor market in terms of 'Skill Biased Technological Change' (SBTC) and give special emphasis to the spread of digital technologies as the catalyst for SBTC. Although globalization and SBTC can be regarded as complementary forces, both 'biased' against the unskilled, there is no consensus among economists about the relative importance of globalization in growing wage dispersion. Research has focused on the effects of international trade. But globalization may be influencing the labor market in less obvious ways, from the offshoring of services to the threat, often explicit, that aggressive wage demands will drive production abroad (Katz and Autor, 1999; Lemieux, 2008; Wood, 1994; Alderson and Nielsen, 2002; Freeman, 2009).

The growing inequality in earnings also reflects the declining influence of wage-setting institutions, such as labor unions, minimum wage laws, centralized bargaining, and wage indexation mechanisms, such as Italy's *scala mobile* (abandoned in 1992). Where they are strong these institutions shield workers from market forces and tend to compress earnings at the lower end of the distribution. But their strength is highly dependent on state support, which has been waning in many countries. In the United States, for example, the real value of the minimum wage, periodically reset by legislation, sank 12% from 1967 to 2011; private sector union membership declined from about 35% in the mid-1950s to under 7% in 2012 (Mishel et al., 2012: 279–280; Gilbert, 2015: 217). Between 1980 and 2000, the coverage of the labor force by collective bargaining agreements declined in six of twelve affluent countries, remained unchanged in three, and modestly expanded in three others. The loss of coverage was greatest in the UK, the US, and New Zealand, countries that also experienced substantial increases in wage inequality. In contrast, France, where coverage was high and became slightly higher, wage disparities shrank (Blau and Kahn, 2009: 189; Visser and Checchi, 2009). While all developed economies were subject to the global and technological forces described above, these differences in labor market institutions help explain the considerable variation among countries in outcomes (Lemieux, 2008; Katz and Autor, 1999).

Growing Inequality: The Top 1%

Various explanations have been offered for the steeply rising share of income flowing to the top 1% in the United States and other Anglophone countries since the 1970s. Capital sources have traditionally accounted for most income at this level, and probably continued to do so for the very wealthy, the top 0.1%. The expansion of financial markets and the increased international mobility of capital, the growth of international trade, and the emergence of profitable new sectors associated with information technology all contributed to a healthy rise in capital income. But by 2000 the majority of the income of the top 1% came in the form of job earnings (Leigh, 2009: 161; Piketty and Saez, 2013). In part, this phenomenon reflected the broader growth in rewards to individuals with highly valued skills. Some observers have noted a tendency, especially in the United States, to confer outsized rewards on top performers – so-called 'superstars' – in fields as diverse as finance and medicine (Rosen, 1983; Frank and Cook, 1995). The most salient example is the compensation of the CEOs of major American corporations, often climbing into the tens of millions of dollars. Rising scale and changing social norms have been suggested as explanations for the spread of superstar rewards beyond their accustomed arenas of entertainment and sports. CEO compensation has clearly increased with the scale of global business. But the CEOs of European and Japanese corporations, operating on a similar scale, are not so richly rewarded as their American peers, and it is hard to see how scale has contributed to the emergence of superstar physicians or superstar university presidents. It may well be that, in some countries, social norms regarding appropriate or just compensation levels have shifted since the 1970s – at least among those with power over pay decisions.

Cross-national differences in tax regimes, reflecting broader political differences, have also influenced the fortunes of the wealthy in developed countries. For the Anglophone nations, reductions in top marginal tax rates since 1970 contributed an estimated one third to one half of the rise in the income shares claimed by the top 1% – first by increasing the amount of market income retained after taxes and, in the longer term, by boosting the accumulation of income producing capital (Leigh, 2009: 165; Hacker and Pierson, 2010).

Growing Inequality: Household Factors

Explanations of rising income inequality in affluent nations, usually focus on market income. But market income is channeled through households, whose composition and behavior affect the ultimate distribution. Three household level factors deserve attention: the increasing proportion of female-headed families; the growing labor force participation and earnings of wives; and 'assortive mating,' the rising tendency for husbands and wives to have similar levels of education and income-earning potential (Espying-Andersen, 2007; McCall and Percheski, 2010; Burtless, 2009; Blank, 2011; Harkness, 2013).

Evidence is best for the first of these factors. In affluent countries generally, female-headed families – especially single-mother families – have become increasingly common. By the first decade of the twenty-first century, the proportion of children in single-mother families was over 20% in the US and the UK and generally high in northern Europe, though notably lower in Spain, Italy, and Greece (LIS Key Figures). Single-mother families typically have lower than average incomes because they are likely to be single-earner families, because women's wages tend to be lower than men's, and because lone mothers must balance competing work and family obligations on their own. Cross-national research indicates that the increasing proportion of female-headed families has measurably contributed to increasing income inequality in affluent countries since the 1970s, except in Scandinavia, where easy access to child care supports the labor participation of single mothers (Espying-Andersen and Myles, 2009: 646–647).

The growing contribution of wives to household earnings is also a factor in rising inequality. A widening gap separates the rising numbers of husband-wife, dual earner households from those with single-earners, especially if they are single parents. But even among married-couple families, wives' increasing contribution may be enlarging income disparities because women's wages, like men's, are growing more unequal, because better educated women are more likely to work than their less educated sisters, and because assortive mating at the top increasingly links high-income earners. The effect of wives' earnings is greatest in southern Europe, where women's labor force participation rates remain low, except among highly educated women with strong earning potential. Its effect is less significant in the US, UK, and Scandinavia.

Conclusion

A combination of labor market and household factors has been reshaping income distribution in high-income countries since the 1970s. Their effect has, in varying degrees, been intensified or attenuated by political/policy influences, evident in the strength of wage-setting institutions, marginal tax rates, and access to state-funded child care. While there is broad agreement among researchers about the growth of income inequality in most developed countries, there is less consensus regarding the causes of this trend and their relative importance.

See also: Capitalism: Global; Class: Social; Income Distribution and Inequality: Measurement Issues; Inequality Analysis: Overview; Social Stratification; Work and Industry, Sociology of.

Bibliography

Atkinson, A., Bandolini, A., 2013. In: Gornick, J. (Eds.), On the Identification of the Middle Class.

Alderson, A., Doran, K., 2013. Has income inequality grown? the reshaping of the income distribution in LIS countries. In: Gornick, Jantti (Eds.), Income Inequality: Economic Disparities and the Middle Class in Affluent Countries. Stanford University Press, Stanford, pp. 51–74.

Alderson, A., Nielsen, F., 2002. Globalization and the great U-Turn: income inequality trends in 16 OECD countries. American Journal of Sociology 107, 1244–1299.

Blank, R., 2011. Changing Inequality. University of California Press, Berkeley.

Blau, F., Kahn, L., 2009. Inequality and earnings distribution. In: Salverda, et al. (Eds.), The Oxford Handbook of Economic Inequality. Oxford University Press, Oxford.

Burtless, G., 2009. Demographic transformation and economic inequality. In: Salverda, et al. (Eds.), The Oxford Handbook of Economic Inequality. Oxford University Press, Oxford, pp. 435–454.

Brandolini, A., Smeeding, T., 2009. Inequality in richer and OECD countries. In: Salverda, et al. (Eds.), The Oxford Handbook of Economic Inequality. Oxford University Press, Oxford, pp. 71–100.

Espying-Andersen, G., 2007. Sociological explanations of changing income distributions. American Behavioral Scientist 50, 639–658.

Espying-Andersen, G., Myles, J., 2009. Economic inequality and the welfare state. In: Salverda, et al. (Eds.), Oxford Handbook of Economic Inequality. Oxford University Press, Oxford, pp. 639–664.

Freeman, 2009. Globalization and inequality. In: Salverda, et al. (Eds.), Oxford Handbook of Economic Inequality. Oxford University Press, Oxford, pp. 575–589.

Frank, R., Cook, P., 1995. The Winner-take-all Society. Simon & Schuster, New York.

Gilbert, D., 2015. The American Class Structure in an Age of Growing Inequality. Sage, Los Angeles.

Gornick, J., Jantti, M., 2013. Income Inequality: Economic Disparities and the Middle Class in Affluent Countries. Stanford University Press, Stanford.

Hacker, J., Pierson, P., 2010. Winner-Take-All Politics: How Washington Made the Rich Richer and Turned its Back on the Middle Class. Simon & Schuster, New York.

Harkness, S., 2013. Women's employment and household income inequality. In: Gornick, Jantti (Eds.), Income Inequality: Economic Disparities and the Middle Class in Affluent Countries. Stanford University Press, Stanford, CA, pp. 207–233.

Katz, L., Autor, D., 1999. Changes in the wage structure and earnings inequality. In: Ashenfelter, Orley, Card, David (Eds.), Handbook of Labor Economics, vol. 3A. Elsevier, New York.

Leigh, A., 2009. Top incomes. In: Salverda, et al. (Eds.), Handbook on Economic Inequality. Oxford University Press, Oxford, pp. 150–176.

Lemieux, T., 2008. The changing nature of wage inequality. Journal of Population Economics 21, 21–48.

Liu, Y., Grusky, D., 2013. Payoff to skills in the third industrial revolution. American Journal of Sociology 118, 1330–1374.

McCall, L., Percheski, C., 2010. Income inequality: new trends and research directions. Annual Review of Sociology 36, 329–347.

Mishel, L., et al., 2012. The State of Working America, twelfth ed. Cornell University Press, Ithaca, NY.

OECD, 2008. Growing Unequal: Income Distribution and Poverty in OECD Countries.

Piketty, T., Saez, E., 2013. Top incomes and the great recession. IMF Economic Review 61, 456–478.

Rosen, S., 1983. The economics of superstars. American Scholar 5, 449–461.

Salverda, W., et al., 2009. The Oxford Handbook of Economic Inequality. Oxford University Press, Oxford.

U.S. Census, 2013. Income, poverty and health insurance coverage in the United States: (2012). Current Population Reports, P60–P245.

Visser, J., Checchi, D., 2009. Inequality and the labor market: labor unions. In: Salverda, et al. (Eds.), The Oxford Handbook of Economic Inequality. Oxford University Press, New York.

Wilkinson, R., Pickett, K., 2009. The Spirit Level: Why Greater Inequality Makes Societies Stronger. Bloomsbury Press, New York.

Wood, A., 1994. North-South Trade, Employment, and Inequality. Clarendon Press, Oxford.

Income Taxes

Agnar Sandmo, Norwegian School of Economics, Bergen, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Following a discussion of the nature of income taxation, the article reviews theory and evidence of the effects of the income tax on individual behavior with respect to labor supply, saving, portfolio choice, and tax evasion. It also discusses the role of the tax in income redistribution and the design of an optimal tax system that combines the concerns for economic efficiency and distributive justice. It also reviews a number of recent proposals for tax reform in different countries.

In the eyes of the taxpayer, the income tax is the most visible form of taxation and therefore also politically the most controversial. Controversies surround the issues of how the income tax affects individual behavior in areas such as labor supply, saving, home production, and tax evasion. Even more controversial, obviously, is the use of the income tax to redistribute income, which raises the more normative question of how progressive the income tax ought to be.

The Nature of Income Taxation

The income tax is the prime example of a direct tax; it is levied directly on the incomes of the individual consumers in the economy, and it can be designed to take account of the individual circumstances of the taxpayer. In contrast, indirect taxes are levied on anonymous market transactions and are paid at the same rate irrespective of the taxpayer's income, number of children, age, health – to mention just a few of the taxpayer characteristics that have over time been taken into account in the design of the personal income tax. It is this tailoring of the tax to the individual's ability to pay, which makes it a potentially important instrument for the redistribution of income and for reducing differences in the standard of living among individuals and families.

The definition of taxable income varies in the legal systems of different countries and over time. Labor income – wages and salaries – forms the central element of the tax base. To this is added income from capital in the form of interest, dividends, etc. However, income from capital is a more elusive concept than income from labor. Thus, countries differ in the extent to which capital gains are subject to taxation, and they differ in the extent to which noncash income, such as the imputed income from owner-occupied housing, is subject to tax. Perhaps even more importantly, rules differ regarding the deductibility of negative capital income, such as interest payment on mortgage loans. Tax laws also differ in the extent to which they allow the deduction of particular expenses, such as the cost of travel to work, the cost of child care, trade union fees, etc.

These differences in the tax base may reflect different perceptions among policy makers of what is a just and fair tax system; they may also reflect differences in the balance of political power among various special interests. What is clear is that different legal definitions of the tax base create different economic incentives for the taxpayers as well as different consequences for the after-tax distribution of income.

The income tax may be progressive, proportional, or regressive. A proportional tax is one where the average tax rate is constant; the proportion of income paid in tax is the same for all levels of income. Note that this implies that the marginal tax rate – the percentage tax paid out of a small increase in income – is also constant. In a progressive tax system the average tax rate increases with income, and this implies that the marginal rate is higher than the average rate. Finally, in a regressive tax system the average tax rate is decreasing, with the marginal tax rate being lower than the average rate.

Countries differ considerably both in their legal definitions of the tax base and in the degree of progressivity of their income tax schedules. There is also significant variation between countries as regards the importance of taxes on income as a source of revenue. To illustrate, in 2010 according to official Organisation for Economic Co-operation and Development statistics, the share of income taxes (as levied on all sources of income and inclusive of social security taxes) in total tax revenue varied from about 45% in Chile, Mexico, and Israel to around 70% in countries like Norway, Switzerland, and the United States. Of course, these percentage shares are reflected in a correspondingly high share of revenue from indirect taxes in the former group of countries and a low share in the latter.

The Income Tax and Economic Behavior: Labor Supply

How do variations in the income tax rate affect the labor supply of individuals and households? The economic theory of labor supply takes its point of departure from a model of a single individual who faces certain market opportunities. In the simplest version of this model, the wage rate per hour worked is constant, so that earnings are proportional to hours worked. In deciding how many hours to work, the individual takes account both of the fact that higher earnings make it possible to consume more, but that it also costs something in the form of less leisure. At the optimal number of hours worked, it must be the case that the subjective marginal value of leisure equals the real after-tax wage rate. For longer hours of work the subjective value of leisure is higher than the wage rate; at shorter hours it is less.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71043-9

Now consider what happens if the marginal tax rate increases. Two effects can be identified. First, since the reward to labor at the margin falls, there is an incentive to work less; this is called the substitution effect. Second, since the tax increase means that the taxpayer is less well off than before, his marginal valuation of leisure falls; this is what is known as the income effect, which in itself tends to increase work effort. Clearly, the two effects pull in opposite direction, so that it is not possible to formulate any clear-cut hypothesis about the connection between the marginal tax rate and labor supply.

A common objection to this theory is that there are in fact few individuals who are able to choose their hours of work freely, and that the theory, therefore, tends to focus attention on a problem of little importance. In the short run this is certainly true for a large number of employees. However, the theory can be interpreted in a number of different ways to make it applicable to many more aspects of labor supply. Consider, for example, a young worker who is entering the labor force and considering different job opportunities. Some jobs are offering long hours and high earnings, while others promise more leisure time at the cost of lower earnings. The occupational choice between job alternatives can in fact be modeled in a way that is formally similar to the hours of work model, and the substitution and income effects of taxation emerge in this setting as well. The same is true of models of the choice of retirement age; a tax increase decreases the reward to a further extension of one's working life and creates an incentive to earlier retirement (the substitution effect), but it also makes it necessary to work longer than before to achieve a given standard of living during retirement (the income effect).

The consideration of retirement raises the more general issue of labor force participation. For some people, taxation may influence their decision on whether or not to participate in the labor market. Income from labor, provided that they decide to work, will be compared to the income that they can count on if they do not work. Alternative sources of income may be income from capital, earnings by family members, income derived from black market work or other illegal activities as well as social assistance and other forms of public support. The tax rate that is relevant for the participation decision is not so much the marginal as the average rate. The crucial point is how much income from labor will be reduced by the total amount of tax paid on this income.

A number of attempts have been made to estimate models of labor supply and taxation using econometric methods. Since the nature of the data sets employed in these studies differs considerably, one might expect that the results would be highly specific with respect to the country, the time, and the nature of the industry to which the data pertain. This is true to some extent, but there are also some general results that characterize a large number of studies. One is that the labor supply of women tends to be more elastic with respect to tax changes than that of men. When women's marginal tax rates go up, they tend to work significantly less, whereas for men one typically finds very small effects. Attempts have also been made to follow the theoretical model and estimate the substitution and income effects separately. One then frequently finds that the reason the total effect for men is small is not that both effects are insignificant, but rather while each is of substantial magnitude, they tend to cancel each other. Moreover, women's participation rate tends to be more sensitive to taxation than that of men.

How is one to interpret the result that the labor supply of women is more elastic than that of men? The most natural explanation is that in most data sets, women tend to be the secondary breadwinners in the family; they tend to work shorter hours and have lower earnings than their male partners. Variations in family labor supply in response to tax changes will, therefore, most naturally take the form of adjustments in the hours of work and labor force participation by women.

In the theoretical framework sketched above, leisure was implicitly defined as the amount of time not spent working in the labor market. But as a matter of fact people spend a lot of time in household work; studies of time use tend to show that household work takes up about as much time on the average as work in the labor market. The substitution effect of higher taxes in favor of leisure is accordingly also in favor of household work; in other words, the income tax 'subsidizes' labor used in household production.

The tax rates that are relevant for the theoretical and empirical study of labor supply are most naturally considered to be simply the statutory rates of the tax law. However, in some contexts it is necessary to take a broader view. Suppose that an individual or household receives a transfer, in cash or in kind, which is income contingent; in order to qualify for it, one's income must not exceed a certain limit. Imagine now that the income of a particular household increases from somewhere below this limit to somewhere above. This increase is subject to the statutory rate of tax, which in the case of a low-income household, can be supposed to be relatively low. But in addition the household loses its transfer, and this loss is in economic terms exactly like a tax increase. The combined effect of the statutory rate and the loss of the transfer payment results in an effective tax rate, which could be much higher and might even exceed 100%, thus creating a powerful disincentive to work among low-income families, often referred to as a 'poverty trap.' This is not just a theoretical possibility but has been identified as an empirical fact in the tax and transfer systems in a number of countries. Clearly, it is very important to consider the interaction between tax and transfer systems in order to avoid unintended and unfortunate consequences of their combined effects.

The Income Tax and Economic Behavior: Saving and Portfolio Choice

The base on which the income tax is levied includes income from capital as well as from labor. A simple extension of the theoretical framework so far is to imagine that the individual lives in two periods, and that he saves some of his labor income in the first period. This will then become available – with interest added – in the second period. Interest income is subject to tax. What happens to saving when the tax rate increases?

As in the labor supply case, we can identify a substitution and an income effect of an increase in the marginal tax rate. When the tax rate goes up, the after-tax interest rate falls and there is an incentive to save less; this is the substitution effect. On the other hand, the lower the net-of-tax interest is, the more it is necessary to save in order to achieve a given level of future

consumption, and this is the income effect. The substitution and income effects clearly work in opposite directions. But we also need to take account of the fact that some people are not net savers in the first part of their life. Instead they borrow with the loan to be repaid with interest in the later part of their life cycle. If the interest on the loan is tax deductible, a higher rate of tax provides an incentive to borrow more, i.e., to save less. This, again, is the substitution effect. But now the lower interest rate makes it possible to borrow more than before without having to reduce future consumption; thus the income effect on saving for a borrower is negative, not positive. Since some people are borrowers and some lenders, the income effect for all consumers together might not be very significant and the substitution effect would tend to dominate. Empirical studies also tend to show that there is in fact a negative tax effect on saving, although the magnitude is not always very strong.

This picture of tax effects on saving seems to assume that all saving takes the effect of lending or borrowing in some kind of homogeneous asset. In reality, of course, people hold a portfolio of assets (housing, bank deposits, mortgage debt, shares, etc.,), and the income tax can therefore be expected to influence not only the aggregate amount of saving but also the composition of the savings portfolio. This is particularly so since the income tax law in many countries does not provide for the taxation of different asset returns at a uniform rate. While interest on bank deposits is often taxed at the statutory rate, interest on some types of bonds may be tax exempt, capital gains may be tax free provided one has held the stock for a sufficiently long time, and the returns from collector's assets like art, antique furniture, rare books, and stamps are usually not taxed at all. Empirical studies also seem to indicate that portfolio composition is indeed very sensitive to tax effects.

In the academic literature there is one type of tax effect on portfolio choice which has received special attention by economic theorists. Some assets are relatively safe in the sense that their return can be predicted with a high degree of certainty; others are more risky. Could the income tax be expected to change the balance between safe and risky assets in savers' portfolios? Suppose that there is full loss offset, so that any negative returns on risky assets can be deducted against other income. Then one can deduce, basically from the assumption that the individual saver is averse to risk, that an increase in the rate of income tax will induce him to invest more in the risky asset than before. The explanation is that although the government takes a share of his expected return, it also shares in the risk, thereby protecting him from the full consequences of low or negative returns. This theoretical insight has had a considerable influence on the way that economists view the bias of investment incentives related to the income tax, but the hypothesis is not an easy one to test on empirical data. The major reason for this is, as argued above, that there are so many other tax incentive effects on portfolio composition that the risk substitution effect becomes very hard to trace.

The Corporate Income Tax

All income ultimately accrues to private individuals, so that the tax on corporate income or profits is in the last instance also a tax on individual consumers. In fact, some economists have suggested that it might be a good idea to abolish the tax on corporations and instead collect all income from corporations at the hands of the individual stockholders. But abolishing the tax on corporations might not be a politically popular issue, and in addition there are some real economic arguments why it should be preserved. One is that collecting the tax at the corporate level helps to reduce incentives for tax avoidance and evasion; another is that it is an efficient way to collect some tax revenue from foreign owners of domestic companies who are not subject to the domestic tax on personal income.

In the theoretical literature there is a further argument in favor of the corporate income tax. Assume that corporations maximize profits and that markets are efficient. Then it would be desirable to have taxes that do not lead to any distortions of production and investment decisions in corporations. But if corporations are subject to a (say) 40% tax on profits, the decisions that would maximize profits in the absence of tax are obviously the same as would maximize 40% of profits. Hence the corporate income tax is neutral; it collects revenue for the government without leading to any distortions of decisions at the firm level and leaves the efficient market outcome unaffected.

However, this picture is very much simplified. The main reason for this is that the tax definition of corporate income is in several respects different from the theoretical concept of profits. One important source of deviation between the two lies in the treatment of real capital depreciation. The real depreciation of a piece of machinery or a building or another type of capital equipment during a year is part of the costs of the corporation, and in order for the corporate income tax to be neutral, real capital depreciation should count as part of the costs in arriving at a definition of taxable income that corresponds to true profits. But a corporation may employ thousands of different types of real capital, and it is not possible for the tax authorities to allow for the deductibility of true depreciation for all of these. Instead they typically allow standard depreciation rates for large groups of capital assets such as machines, buildings, transportation equipment, and so on. But this means that for some types of capital the corporation is allowed to deduct less than the true depreciation, for others more. In its interaction with the tax depreciation rules the corporate income tax in fact works as a selective tax on some types of capital equipment and as a subsidy on others. Similar issues may arise regarding interest deductibility, stock valuation, and capital gains both on real and financial assets. It is, therefore, not surprising that empirical studies, e.g., Jorgenson and Yun (1991), tend to show that there are substantial tax effects on production and investment decisions in corporations.

Income Tax Evasion

Tax evasion is an illegal activity, which aims to hide taxable income from the view of the tax authorities. It should be distinguished from tax avoidance, which consists in trying to reduce one's taxable income by exploiting the tax law while staying inside its boundaries. While the distinction between the two may sometimes be unclear, there are at least a large

number of cases that definitely belong in the evasion category; these are sometimes referred to as the hidden or black economy.

Tax evasion is a risky activity. By holding back information from the tax authorities, there is always a risk of being discovered, in which case one typically faces a penalty rate on the unreported income. The gain on a dollar of unreported income, i.e., the regular rate of income tax, must be balanced against the probable loss, which is the probability of detection times the penalty rate of tax. When the regular income tax rate goes up the gain increases, and this tends to encourage evasion e.g., in the form of work in the hidden economy. Theory seems to support the common concern that the high-tax level encourages cheating and dishonesty. However, this should not lead us to expect that high-tax countries necessarily have more tax evasion than countries with lower tax levels. The extent of evasion depends not only on the level of income taxation, but also on a number of additional social and economic factors, such as the level of tax enforcement and efficiency of tax collection as well as the general attitude among taxpayers toward government and the public sector. All of these factors are likely to vary considerably among countries.

Theoretical hypotheses in this area are particularly hard to confront with empirical data; in the nature of things, there are no official statistics for the hidden economy, and survey research faces some special difficulties in making individuals answer questions truthfully. A number of the empirical studies that have been made uses very indirect methods, like deducing the size of the hidden economy from the public's holding of cash, and the reliability of this kind of approach is highly disputed; see e.g., Cowell (1990). More standard approaches tend to yield estimates which indicate that the hidden economy has a volume of 2–10% of the gross national product of the official economy. Many studies also indicate that evasion is higher for self-employed than for salaried employees. This is in good accordance with the theory, for wages and salaries are usually reported directly by employers, so the probability of evasion being detected is much higher for employees than for the self-employed, implying that evasion is a less risky gamble for the latter.

It should be added that the emphasis on rational risk-taking behavior as an explanation of tax evasion does not imply that in the economist's view of the world all taxpayers are completely amoral, reporting their income truthfully only when a cost–benefit calculation has shown them that evasion does not pay. There are strong indications that many taxpayers report their true income even when underreporting could easily go undetected, and at least part of the explanation for this are likely to be found in peoples' moral aversion to lying about their true income. However, it remains a fact that a large number of people do engage in tax evasion activities, and these are likely to consider the risk-taking aspects of their activities in a manner which is rational, given their preferences and attitudes to risk. For further discussion see e.g., Sandmo (2012).

The Income Tax and the Redistribution of Income

It is widely accepted in most countries that the income tax should take account of the taxpayer's ability to pay, and the most common interpretation of this principle is that the tax should be progressive. This raises the question of whether the income tax in fact has led to a situation where posttax or disposable income is more evenly distributed than pretax or factor income. In most empirical studies the answer to this question is an affirmative one, although the estimated magnitude of the redistributive effect varies substantially. A general feature of empirical studies is that the redistributive effect is smaller than the formal progressivity of the income tax schedule would lead one to expect, and the main reason for this is the role played by exemptions or tax credits. As an example, suppose that mortgage interest payments are fully deductible and that their share of gross income tends to increase with income. Then, as gross income increases, taxable income increases less than disposable income in percentage terms, and this mitigates the increase in the effective average tax rate. The importance of this effect has been well documented in a number of studies, such as Wagstaff et al. (1999).

From a theoretical point of view one would expect that the progressivity of the tax system would have an effect on pretax incomes. The markets for at least some occupations would react to high marginal tax rates by increasing gross incomes; this is the problem of tax incidence. Since the gross incomes that would have existed had there been no income tax or a purely proportional tax, are unobservable, this observation makes it much more difficult to study the redistributive impact of the income tax. In this connection it is interesting to observe that countries with a high degree of pretax inequality also tend to be the ones with high posttax inequality, and moreover that the redistributive impact of the income tax is high in some countries, which have a low degree of inequality in pretax incomes (Wagstaff et al., 1999). However, it is not clear whether this means that the incidence problem is in fact of little importance or whether there is some other explanation. If public opinion is dominated by egalitarian social and political values, these may be reflected not only the progressivity of the tax system, but also in the structure of pretax earnings.

The Income Tax, Economic Efficiency, and the Equality–Efficiency Trade-Off

The income tax has always been under attacks from critics for leading to distortions of economic incentives and consequently to inefficient social outcomes. The criticism has been especially focused on the distortion of work incentives, which leads to incentives to work less than would be desirable from a social point of view. As pointed out in Section The Income Tax and Economic Behavior: Labor Supply above, there is clear evidence for such an effect in the case of women, while for men the net effect tends to be very small. However, in considering the efficiency loss from income taxation, it is the substitution effects, and not the total effects, that should concern us. The reason is that the income effects would always be present, whatever the form of the tax system (and given the amount of revenue to be collected). The substitution effects are the direct consequences of the distortion of wages as efficient market prices, and if our sole concern is with efficiency, we should attempt to design the tax system in such a way as to make the substitution effects as small as possible. Similar concerns are relevant in judging the

effects of income taxation on saving and investment behavior. The taxation of interest income might lead consumers to save less and to invest in the assets or industries with the most favorable tax rules rather than in those with the highest rates of return.

However, efficiency is not the only standard by which to judge the income tax. As demonstrated in the progressive structure of the income tax in most countries, one of the goals of income taxation is to redistribute income from the rich to the poor, and the gains on the side of redistribution must be set against the cost of a less efficient allocation of resources. This is a recurrent problem in the design of economic policy; there is a need to choose a trade-off between equality and efficiency. However, if more equality is a real goal of economic policy, it should presumably be worth some cost in terms of less economic efficiency in order to achieve it. These ideas have been formalized in the theory of optimal income taxation as first formulated by Mirrlees (1971). This theory attempts to show how the optimal schedule of income taxation – in particular, the taxation of income from labor – depends both on the efficiency cost, as measured by the elasticity of labor supply, and by the social attitude to inequality. On the one hand, the more elastic labor supply is with respect to the marginal tax rate, the lower should the tax rates be. On the other hand, the more pronounced society's taste for equality is, the stronger is the case for a progressive tax schedule. Clearly, what economic theory can contribute to this issue is not to produce exact numbers, but to help establish a logically consistent way of thinking about a complex issue that involves both economics and ethics (Sandmo, 1999). By varying the assumptions made about value judgments, the structure of labor supply and the distribution of earnings, the theoretical analysis helps to clarify how conclusions about the desirable form of income taxation can be derived from a consistent framework of economic analysis.

Proposals for Tax Reform

A number of proposals have been advanced for basic reform of the income tax system, frequently motivated by the desire to improve the terms of the equality–efficiency trade-off. For example, is it possible to achieve the same degree of redistribution at a lower efficiency cost? Tax reforms in a number of Western countries during the last couple of decades have tried to move in this direction by decreasing the formal degree of progressivity in the tax system, while at the same time eliminating many of the deductions and exemptions that had contributed to a lower degree of effective progressivity. Studies of the American (Auerbach and Slemrod, 1997) and Scandinavian (Agell et al., 1996; Sørensen, 1998) reforms of the 1980s and 1990s, do in fact conclude that their overall effects have been to leave the redistributive effects of the tax system virtually unaffected, while removing many of the tax distortions both in labor and capital markets. The Scandinavian countries were pioneers in introducing the principle of dual income taxation whereby labor income is taxed at progressive rates while income from capital is taxed at a flat rate (the same for personal and corporate income), which is considerably lower than the marginal tax on labor income. The main motivation behind this version of the income tax has been to reduce the opportunities for tax arbitrage with respect to capital income, thereby leading to a system, which is arguably both more efficient and more equitable.

A more radical proposal for reform is the adoption of a so-called flat tax. There are several versions of this, but one that was proposed by Hall and Rabushka (1995) and has been much discussed in the United States involves imposing a constant marginal tax rate on all income above a certain exemption level. Although the name of the tax indicates that the major point in its favor is the simplification of the tax structure, probably its more important features are the elimination of a large number of deductions and a more uniform taxation of income from capital. In this, it is in line with the general trends in the reforms of the 1980s and 1990s.

Another reform, whose intellectual history goes back a long time and which was much discussed during the late 1970s and early 1980s was the possible adoption of an expenditure tax (Meade Committee, 1978; Kay and King, 1990). Although the expenditure tax is often seen as an alternative to the income tax, it may also be regarded as simply a particular form of income taxation. Briefly, an expenditure tax is an income tax with full deduction for all saving, but with negative saving being added to income. What is taxed is accordingly expenditure or consumption, not income. There are two main claims for the expenditure tax. First, since differences in the standard of living between persons are primarily related to their consumption, not income, consumption is the more logical base on which to level a progressive tax. Second, the treatment of saving and dissaving leads to a more uniform and, therefore, more efficient taxation of capital income. Although neither the expenditure tax nor the flat tax have so far achieved any direct political success, the political and economic debates over these proposals may indirectly have had considerable influence on the reforms that were actually adopted in the following years.

A recent ambitious attempt at the analysis of tax reform is the Mirrlees Review in the United Kingdom. Its final report (Mirrlees Review, 2011) proposes a series of reforms of the UK tax system, including a more transparent and coherent schedule of progressive income taxation with rates that reflect the available evidence on behavioral responses to taxation. The report is based on a collection of research articles by an international group of leading experts on the economics of taxation, including several that have a direct bearing on the design of the income tax (Mirrlees Review, 2010). In its discussion of the relationship between the results of economic research and the design of policy, the report is of interest to economists and policy makers everywhere, not just in the United Kingdom. Income tax reform and design is an area where economic analysis and research can make – and indeed has made – significant contributions to policy formation.

Conclusion

Among all forms of taxation, the income tax has traditionally been the one that has been most surrounded by political controversy. To the individual taxpayer, the tax is much more visible than other forms of taxation, and it is one that more obviously affects his own economic situation. To the academic

economist, the income tax provides a fascinating area of application of his theoretical and empirical tools of analysis and it also gives him the opportunity to contribute to the public debate about tax policy. The trade-off between efficiency and equity concerns, which is basic to many aspects of economic policy, is nowhere more prominent than in the design of the income tax. The question of what weights one should assign to the benefits of redistribution versus the costs in efficiency is ultimately one that must be decided by the political system. What the economist can contribute to the political debate is a framework for systematic analysis and a sense of the empirical magnitudes involved both as regards incentives and redistribution.

See also: Corporate Taxation; Income Distribution and Inequality: Measurement Issues; Income Inequality; Inequality, Social; Informal and Underground Economics.

Bibliography

Agell, Jonas, Englund, Peter, Södersten, Jan, 1996. Tax reform of the century – the Swedish experiment. National Tax Journal 49, 643–664.

Auerbach, Alan J., Slemrod, Joel, 1997. The economic effects of the Tax Reform Act of 1986. Journal of Economic Literature 35, 589–632.

Cowell, Frank A., 1990. Cheating the Government. MIT Press, Cambridge, Massachusetts.

Hall, Robert E., Rabushka, Alvin, 1995. The Flat Tax, second ed. Hoover Institution Press, Stanford, CA.

Jorgenson, Dale W., Yun, Kun-Young, 1991. Tax Reform and the Cost of Capital. Oxford University Press, Oxford.

Kay, John A., King, Mervyn A., 1990. The British Tax System, fifth ed. Oxford University Press, Oxford.

Meade Committee, 1978. The Structure and Reform of Direct Taxation. Allen & Unwin for Institute for Fiscal Studies, London.

Mirrlees, James A., 1971. An exploration in the theory of optimum income taxation. Review of Economic Studies 38, 175–208.

Mirrlees Review, 2010. Dimensions of Tax Design. Oxford University Press for Institute for Fiscal Studies, Oxford.

Mirrlees Review, 2011. Tax by Design. Oxford University Press for Institute for Fiscal Studies, Oxford.

Sandmo, Agnar, 1999. Asymmetric information and public economics: the Mirrlees-Vickrey Nobel Prize. Journal of Economic Perspectives 13, 165–180.

Sandmo, Agnar, 2012. An evasive topic: theorizing about the hidden economy. International Tax and Public Finance 19, 5–24.

Sørensen, Peter B. (Ed.), 1998. Tax Policy in the Nordic Countries. Macmillan, London.

Wagstaff, Adam, van Doorslaer, Eddy, van den Burg, Hattem, et al., 1999. Redistributive effect, progressivity and differential tax treatment: personal income taxes in twelve OECD countries. Journal of Public Economics 72, 73–98.

Indicator: Methodology

Kenneth A Bollen, University of North Carolina, Chapel Hill, NC, USA
Shawn Bauldry, University of Alabama, Birmingham, AL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by K.A. Bollen, volume 11, pp. 7282–7287, © 2001, Elsevier Ltd.

Abstract

The term 'indicator' refers to a variable that is an imperfect representation of a social science concept. First, this article distinguishes different types of indicators (effect, causal, and composite indicators). Second, it describes the construction of measurement models, which are models that specify the links between latent variables (representations of social science concepts) and their indicators. Attention is given to the major questions that must be answered when formulating a measurement model. Third, this article reviews the properties of continuous and categorical indicators. Finally, the article provides an overview of the guidelines for developing and/or selecting indicators in empirical analyses.

Indicators and related terms (e.g., proxies, scales) refer to variables that are imperfect representation of social science concepts. The focus of this article is the treatment of indicators in social science research. It describes the construction of measurement models to link the latent variables that represent concepts to their indicators that contain measurement error. The article reviews the major questions that must be answered when formulating a measurement model. The properties of the indicators are the next topic. Finally, the article gives an overview of the guidelines to employ when developing or selecting indicators.

Concepts in the social sciences are often abstract. Current knowledge or data allow only imperfect empirical representations of them. Abstract ideas such as intelligence, the values of goods, individual attitudes, or social class are important to social science research and need to be empirically tracked even if the means of doing so are limited. 'Indicator' is a common term to refer to the variables that social scientists use to detect these concepts empirically. Other terms are *proxy variables*, *items*, *measures*, *scores*, *scales*, *indices*, or *observed variables*. Occasionally, some of the terms have a 'part–whole' relation where, for instance, a scale or an index might refer to the sum of items or the sum of indicators. But there is no complete consistency in these uses so that it is common for researchers to refer to a Likert scale even though it is a single item or an indicator. Furthermore, the frequency of the use of these terms varies by discipline. Economists, for example, regularly refer to proxy variables while educational testing psychologists frequently refer to items or scales. Regardless of the name, they all refer to a variable that is directly associated with a latent variable such that differences in the values of the latent variable mirror differences in the values of the indicator.

Latent variables are variables that represent concepts. There should be a separate latent variable for each dimension of a concept and a minimum of one indicator per latent variable. Indicators of latent variables are typically observed variables, but in special circumstances a latent variable might serve as an 'indicator' of another latent variable (e.g., higher-order factor analysis). The indicator's importance derives from what it reveals about the latent variable. The indicator informs about the magnitude of the latent variable across cases or overtime for the same cases. The key idea is that the indicator provides information on the latent variable. One might, for example, be interested in socioeconomic status (SES) of individuals where SES is defined as the overall standing of an individual in the social and economic stratification systems of a community and use a single latent variable to represent this overall standing. Income and education are two indicators of SES, but they do not fully capture SES. Or one could have a test designed to tap the anxiety levels of individuals, but the item or overall test score would not perfectly measure anxiety. So the use of the term 'indicator' and other related terms gives recognition to the fact that the variables social scientists often have to work with are not identical to their respective latent variables.

Measurement Strategies

When faced with latent variables for which there is no ideal measure, researchers typically follow one of three strategies: (1) ignore the latent variable, (2) include indicators but ignore their measurement errors, or (3) include indicators but take account of their measurement error. An example of strategy (1) is a study that is interested in the educational attainment of individuals. Suppose that the researcher recognizes that an individual's motivation or drive plays a role in educational attainment, but given the difficulty of measuring motivation she or he ignores it. The danger of this all too common decision is that the omitted latent variable is likely to bias the estimates of the impact of the other included explanatory variables. In the educational attainment empirical analysis, part of the effect attributed to say, parental educational attainment might actually be due to the unmeasured motivation variable.

In strategy (2) one could include an indicator of motivation but ignore its measurement error. In this case, the indicator for motivation allows one to partially control for motivation and estimate the impact of motivation, but because measurement error is ignored both the estimate of motivation and estimates for variables associated with motivation are likely to be biased. Contrary to conventional wisdom disregarding measurement error is not a conservative approach. In general, random measurement errors in two or more explanatory variables will not always lead to underestimates of effects of motivation or the effects of the other variables in the analysis (Bollen, 1989).

A natural question is whether one is better off including the indicator and ignoring the measurement error as in strategy (2) or ignoring the variable altogether as in strategy (1). Those researchers who have explored this question generally find that it is better to include an imperfect indicator than to ignore the latent variable altogether (Aigner, 1974; McCallum, 1972; Wickens, 1972).

Strategy (3), including indicators and taking account of their measurement errors, is generally the best choice. Here one could include indicators of motivation and control for the influence of the measurement error in these variables. This approach would halt the spread of bias due to omitting the latent motivation variable or treating the motivation indicators as if they were perfect measures. A requirement for taking account of the measurement error is the formulation of a measurement model.

Measurement Models

Measurement models refer to the implicit or explicit models that relate the latent variable to its indicators. Several basic questions require consideration prior to building a measurement model: (1) Are there *causal* or *effect indicators*? (2) Are there *multiple* or *single indicators* of the latent variable? (3) Is the latent variable continuous or noncontinuous? (4) Are the indicators continuous or noncontinuous? (5) Does more than one latent variable influence the indicator?

The first question concerns whether the indicators influence the latent variable (causal indicators) or whether the latent variable drives the indicators (effect indicators). The overwhelming majority of social science work on measurement has focused on effect indicators where the indicators depend on the latent variable. Sometimes these are called 'reflective indicators.' Differences in verbal ability, for instance, determine the differences in values for a test of verbal ability. Or an antitax attitude should determine the responses that an individual gives to a question on attitudes toward taxes. Effect indicators are implicitly assumed in factor analysis and in much of measurement theory in the social sciences.

Blalock (1964) called attention to causal indicators, in which the indicator influences the latent variable rather than the reverse. Some refer to these as 'formative indicators,' though this term is better reserved for composite indicators, a different type of indicator discussed below. Education and income of an individual, for instance, influence SES making these variables likely to be causal indicators. The amount that a child watches violent television shows, plays violent video games, or views violent movies are causal indicators of the latent variable of exposure to media violence. Often researchers do not consider that an indicator may be causal and treat all variables as effect indicators. This is not an inconsequential decision since causal and effect indicators have different properties and their incorrect classification can bias estimates from a model (Bollen and Lennox, 1991; Bollen and Bauldry, 2011). So a researcher needs to decide whether indicators are causal or effect indicators or some mixture of types.

The primary method for making this determination is performing mental experiments where the researcher imagines whether shifts in an indicator are likely to result in shifts in a latent variable as would occur with causal indicators or vice versa as would be the case for effect indicators. For instance, increases in time spent watching violent television shows increases the overall exposure to media violence. Increased overall exposure to media violence does not necessarily lead to increased watching of violent television. In some cases, empirical tests based on the presence or absence of vanishing tetrads can test for causal versus effect indicators (Bollen, 1989; Bollen and Ting, 2000). Tetrads refer to differences between the products of pairs of covariances among four random variables. Tetrads that equal zero are said to be vanishing tetrads. Models with effect indicators often imply vanishing tetrads that are not implied by models with causal indicators. These differences in vanishing tetrads are key to empirical tests.

Two complications arise when dealing with causal indicators. First, it is more challenging to specify an identified model with causal indicators. Identification is too complicated to cover here, however, suffice it to say that it is often easier to identify models with effect indicators than those with causal indicators. Second, causal indicators can be mistaken for composite (formative) indicators or covariates due to the fact that causal indicators, composite indicators, and covariates all affect latent variables. Bollen and Bauldry (2011) distinguish causal indicators as having conceptual unity and stable effects on the latent variable they measure across different outcomes of the latent variable. By contrast, composite indicators need not have conceptual unity, perfectly determine the 'latent variable' (i.e., there is no residual variance for the latent variable), and their effects on the latent variable may differ across outcomes of the latent variable. Covariates have similar properties to composite indicators, but they are not intended as indicators of a latent variable and generally do not perfectly determine the latent variable.

As an illustration of these differences, race and gender are common *covariates* that influence many different latent variables without being considered measures of those latent variables. A composite variable of childhood background effects might be formed as the weighted sum of *composite indicators* of single-headed household, recipient of welfare, record of juvenile delinquency, and frequency of moving. Finally, education, income, and occupational prestige could serve as *causal indicators* of SES.

The second characteristic of measurement models is whether there are multiple indicators of a latent variable or just one. Multiple indicators are generally preferable to single measures. In the case of effect indicators, multiple indicators are often essential in 'identifying' (permitting unique estimates of parameters in) the measurement model, are helpful in testing the number of latent variables underlying the measures, and enable assessments of the quality of the indicators. Multiple causal indicators can lead to a fuller description of the latent variable's core meaning. For instance, using education, occupational prestige, and income as causal indicators of SES provides a more comprehensive picture of SES than would income alone. In addition, leaving out causal indicators usually will bias the estimates of the effects of the remaining causal indicators on SES. So for both types of indicators it is better to have more rather than less.

The third and fourth distinctions from above are whether the latent variables and indicators are continuous or not. Latent

or indicator variables that are best thought of as approximating a variable that has infinite gradations of magnitude can be thought of as continuous. Indicators at best can approximate continuous variables since the limitation of measuring instruments will not permit infinite gradations. However, indicators approximate continuity in measurement to varying degrees. Population size, number of crimes, or counts of correct answers on a test are strictly speaking noncontinuous since the measures have interval units. However, in many situations, little harm is done in treating such indicators as if they were continuous. Alternatively, if one uses a three-point scale to measure an attitude, it is difficult to consider the indicator continuous even if the corresponding latent variable is continuous.

One source of confusion when considering whether a latent variable is continuous or not is that researchers sometimes confound the indicators' properties with those of the latent variable. Consider the situation of a survey question on abortion that asks whether the respondent supports or opposes a woman's right to an abortion. Though the response categories for the question might permit only two answers, this does not mean that the abortion attitude *latent variable* is a dichotomy. It seems far more likely that the latent attitude is continuous such that not everyone who agrees to the right to abortion has the same degree of support. Some might be just above a threshold on the continuous abortion attitude and this leads them to choose the support category whereas others might be way above the threshold when they give their support response. In this and related situations, the latent variable is continuous even though the design of the measure forces the responses into categories. The researcher needs to consider whether it makes sense to consider the latent variable as continuous or categorical without confounding the nature of the indicator with that of the latent variable.

Theoretically, it is possible to have any combination such as a continuous latent variable and noncontinuous indicators or a noncontinuous latent variable and noncontinuous indicators. In practice some combinations are far more frequent than others. Consider effect indicators first. Treating both the latent variable and its effect indicators as continuous is particularly common. A typical equation for a measurement model for this case is

$$y_j = \alpha_j + \lambda_{j1}\eta_1 + \varepsilon_j \qquad [1]$$

where y_j is the jth effect indicator that is approximately continuous and the right hand side of the equation includes an intercept (α_j), a factor loading coefficient (λ_{j1}) that gives the expected difference in the indicator for a one unit difference in the continuous latent variable (η_1), and the continuous disturbance (ε_j). The disturbance has a mean of zero and is uncorrelated with the latent variable. This model captures many factor analysis (Mulaik, 1972) and structural equation model formulations (Bollen, 1998), though factor analyses often deviate the observed variables from their mean so as to eliminate the intercept term.

An increasingly popular but less common formulation is to have a continuous latent variable with one or more noncontinuous indicators. With two exceptions, the same equation captures this situation. One is that a y_j^* replaces y_j where the y_j^* is the jth *latent* continuous indicator of the latent variable. The second difference is that there is an auxiliary measurement model that links the continuous y_j^* latent indicator to the noncontinuous observed indicator, y_j. For instance, if y_j is a dichotomous 0,1 variable the auxiliary equation might represent it as

$$y_j = \begin{cases} 1 & \text{if } y_j^* \geq 0 \\ 0 & \text{if } y_j^* < 0 \end{cases} \qquad [2]$$

An exact nonlinear function would link the continuous latent indicator to the observed noncontinuous effect indicator. This is easily generalized to ordinal and censored effect indicators. This formulation is common in structural equation models (Muthen, 1984). Item Response Theory (IRT) provides a related approach to having continuous latent variables and noncontinuous indicators (Lord, 1980; Hambleton and Swaminathan, 1985).

Less frequent are instances where both the latent variable and its effect indicators are categorical. For instance, one might have several informants who report whether the head of the household is receiving welfare or not. The latent variable is welfare status with only two categories: receiving or not. The informant indicators have the same two categories though each informant might make a classification error. Latent structure or latent class analysis (Lazarsfeld and Henry, 1968) provides a means to handle this situation. Other combinations such as a categorical latent variable in combination with continuous effect indicators or a mixture of continuous and noncontinuous effect indicators are also possible, though they are rare in practice.

Causal indicators are far less studied than effect indicators. However, incorporation of continuous and noncontinuous causal indicators in a model with a continuous latent variable is possible provided the latent variable has sufficient outcomes to permit model identification (Jöreskog and Goldberger, 1975). Perhaps the least studied case is that of having causal indicators (continuous or not) combined with a categorical latent variable.

A final consideration is whether more than one latent variable influences an effect indicator. It could be that two indicators derived from a questionnaire are from similarly worded questions such that the wording is a method latent variable that influences both items in addition to the substantive variable that they tap. Or indicators deriving from the same judge might share a common bias beyond their common dependence on the same latent substantive variable. It is also possible that two or more substantive latent variables help to determine the values of an indicator. For instance, a measure of the democracy of a political system might measure both democracy and the stability of the system. An indicator of depression might reflect both negative affect and sleep disorders. In these situations, the measurement model should allow two or more latent variables to affect the indicator. Generally researchers try to design indicators so that they primarily represent a single latent variable, but this is not always feasible.

Properties of Indicators

Properties of the indicators are useful to both current and future researchers who plan to use them. Among the two most important properties are the validity and the reliability of the

indicators. Validity and reliability are properties that have received their greatest attention in the case of measurement models with continuous latent variables and approximately continuous effect indicators. Indicator *validity* concerns whether the indicator really measures the latent variable it is supposed to measure. Construct validity, criterion validity, and content validity are types of validity that researchers sometimes examine. Construct validity, for instance, assesses whether the indicator is associated with other constructs that it is supposed to relate to and not associated with those that it should not. Criterion validity compares the indicator to some standard variable that it should be associated with if it is valid. Content validity examines whether the indicators are capturing the concept for which the latent variable stands. See Nunnally and Bernstein (1994) for further discussion.

Reliability focuses on the consistency or 'stability' of an indicator in its ability to capture the latent variable. It is distinct from validity in that one can have a reliable indicator that does not really measure the latent variable. A general definition of the reliability of an indicator is the proportion of the 'true' (latent variable) variance divided by the total indicator variance. A variety of statistics to estimate reliability exist. One of the most popular method to measure the reliability of several combined effect indicators is Cronbach's (1951) alpha. It also makes a number of assumptions that might be difficult to satisfy in practice. Alternative measures of reliability built from less restrictive assumptions also are available (Bollen, 1989).

IRT assumes a continuous latent trait and a categorical effect indicator, usually dichotomous or ordinal. IRT focuses on other properties of categorical items or indicators such as item discrimination and item difficulty (Hambleton and Swaminathan, 1985). IRT is similar to the traditional treatments of reliability and validity in that it too focuses on effect indicators. Latent class or latent structure analysis (Lazarsfeld and Henry, 1968) also deals with effect indicators. It applies when we have latent categorical variables with categorical indicators. The degree of classification error of the observed categorical variables provides information on the accuracy of the indicator. The combination of a latent categorical variable with continuous effect indicators are less extensively developed than are the cases of continuous latent variables with continuous or categorical effect indicators.

The measurement properties of causal indicators are less discussed. An important point is that use of the causal indicator assumes that it is the causal indicator that directly influences the latent variable. If the causal indicator itself contains measurement error, then this needs to be part of the measurement model.

Indicator Development and Selection

A salient component of indicator methodology is the development or the selection of indicators. The most critical and most neglected aspect of this task is providing a clear *theoretical definition* of the concept that a researcher seeks to measure. The theoretical definition assigns the meaning to the concept. It should explain in as simple of terms as possible what it is that is to be tracked by the indicators. A clear theoretical definition should identify the number of distinct aspects or *dimensions* of the concept. Each dimension requires a separate latent variable. Furthermore, the theoretical definition should clarify whether the latent variable is continuous or not. Each latent variable is ideally measured with several indicators.

Once a clear theoretical definition is in hand, the researcher collecting original data can design indicators that correspond to this definition. If only secondary data are available, the theoretical definition provides guidance in the selection of indicators. With the indicators selected to represent each latent variable, the next step is to decide whether each indicator is a causal or an effect indicator. That is, an investigator needs to decide whether the indicator influences or is influenced by the latent variable to which it corresponds.

Next, the researcher should construct a model summarizing the relation between the indicators and the latent variable(s). The model should take account of the continuous or noncontinuous nature of both the latent variable(s) and the indicators. Once the measurement model is constructed, a researcher can then explore ways of estimating it. If, for instance, both the latent variables and effect indicators are essentially continuous, then confirmatory factor analysis applies. After estimation, the researcher can assess the overall fit of the model and the properties of the indicators. This check might reveal that modifications to the model are required. It might be the case that some indicators prove faulty or the overall fit of the model might be poor, raising questions about the original specification. The analyst might then implement minor or major changes ranging from the introduction of additional parameters to the model to the formulation of quite a different model. The researcher would then estimate the new model and again assess its adequacy.

Summary

Indicators and related terms give recognition to the gap that exists between social science concepts and measures. During the last half of the twentieth and the first part of the twenty-first century, a variety of techniques and statistical models have been developed to enable researchers to study and to take account of the measurement error in the indicators. It is also possible to estimate relations between different latent variables while controlling for measurement error. Though the measurement models cover many situations, further development of techniques to handle some combinations of categorical and continuous latent variables and indicators is required. However, a far more serious problem is the prevalent practice by social scientists of ignoring measurement error in their indicators or ignoring altogether latent variables that are difficult to measure.

See also: Classical (Psychometric) Test Theory; Computerized Test Construction; Data Collection: Interviewing; Dimensionality of Tests: Methodology; Instrumental Variables in Statistics and Econometrics; Measurement Theory: History and Philosophy; Measurement, Representational Theory of; Mediating Variable; Psychometrics; Reliability: Measurement.

Bibliography

Aigner, D.J., 1974. MSE dominance of least squares with errors of observation. Journal of Econometrics 2, 365–372.

Blalock Jr., H.M., 1964. Causal Inferences in Nonexperimental Research. University of North Carolina Press, Chapel Hill, NC.

Bollen, K.A., 1989. Structural Equations with Latent Variables. Wiley, New York.

Bollen, K.A., 1998. Structural equation models. In: Armitage, P., Colton, T. (Eds.), Encyclopedia of Biostatistics. Wiley, Chichester, UK.

Bollen, K.A., Bauldry, S., 2011. Three Cs in measurement models: causal indicators, composite indicators, and covariates. Psychological Methods 16, 264–284.

Bollen, K.A., Ting, Kwok-fai, 2000. A tetrad test for causal indicators. Psychological Methods 5, 3–22.

Bollen, K.A., Lennox, R., 1991. Conventional wisdom on measurement: a structural equation perspective. Psychological Bulletin 110, 305–314.

Cronbach, L.J., 1951. Coefficient Alpha and the internal structure of tests. Psychometrika 16, 297–334.

Hambleton, R.K., Swaminathan, H., 1985. Item Response Theory: Principles and Application. Kluwer-Nijhoff Publishers, Boston.

Jöreskog, K.G., Goldberger, A., 1975. Estimation of a model with multiple indicators and multiple causes of a single latent variable. Journal of the American Statistical Association 70, 631–639.

Lazarsfeld, P.F., Henry, N.W., 1968. Latent Structure Analysis. Houghton Mifflin, New York.

Lord, F.M., 1980. Applications of Item Response to Theory to Practical Testing Problems. Erlbaum, Hillsdale, NJ.

McCallum, B.T., 1972. Relative asymptotic bias from errors of omission and measurement. Econometrica 40, 757–758.

Mulaik, S., 1972. The Foundations of Factor Analysis. McGraw-Hill, New York.

Muthén, B., 1984. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika 49, 115–132.

Nunnally, J.C., Bernstein, I.H., 1994. Psychometric Theory, third ed. McGraw-Hill, New York.

Wickens, M.R., 1972. Use of proxy variables. Econometrica 40, 759–761.

Indigeneity

Jeremy Beckett, University of Sydney, Pyrmont, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Indigenous peoples are the original inhabitants of lands that were subsequently occupied by new comers, in relation to whom they remain to some degree culturally different, and politically and economically inferior. Such situations are to be found in European settler colonies in the Americas, Australasia, and elsewhere, although the original inhabitants are not always recognized as indigenous. Since the mid-twentieth century, there has been local and international recognition of indigenous peoples, and a corresponding political mobilization, and cultural revival. Instances of these developments are included from Mexico and Australia.

Introduction

Indigeneity has as its literal meaning 'belonging to a place,' but in its current use it differentiates the original inhabitants of a territory from those with whom they now share it. Culturally, indigenous difference is defined in terms of not only adherence to tradition and a keeping of modernity at arm's length, but also as having an inferior and deprived status in the society of which they are a part. Conceptions of indigeneity are poised uneasily between these two sets of criteria, as political and economic objectives for either group may or may not be compatible. Settler colonialism, in which the original inhabitants have been displaced by latecomers, whether a majority or a minority, has given rise to many, although not all situations of indigeneity.

As it is experienced, and as understood by social scientists, indigenous difference has also been seen in terms of racial, ethnic, or class difference. Ethnicity is the term often applied to indigenous people in the discourse of multiculturalism, but it differs in having "a different relationship to national time and space" (Povinelli, 2002: p. 480). From around the middle of the twentieth century, the emphasis has shifted to the cultural, with revaluation of indigenous traditions, or acceptable elements of them, and recognition of the disadvantageous effects of settlement upon indigenous ways of living. At the same time, inspired in part by struggles against racism and colonialism elsewhere, indigenous groups have found the political and cultural space to make claims on their own account, to rearticulate fractured traditions, and even to reindigenize where these have been lost.

In the new millennium, indigenous or 'First' Peoples have, as Francesca Merlan insists (2009: p. 306), emerged as a construction, based on a kind of difference that, despite its diversity, is notionally shared vis-à-vis the rest of the world (Niezen, 2000). Although rarely powerful at the level of the nation states in which they are situated, they have belatedly established a presence at the United Nations (UN).

After a brief historical review, mainly of the European settler colonies, this article will focus on the years following World War 2 up to the present. It is suggested that any discussion of indigeneity as a global category has to be read from the bottom up against the innumerable local instances, where indigenous identities may be embedded in other kinds of relations, political, economic, and cultural. A distinction must also be made in terms of the number of the indigenous sector vis-à-vis the nonindigenous, the capacity of national governments to control the indigenous sector, combined with their vulnerability to international opinion.

While gains have been made on the global, and certain local stages, economic globalization and technological change have subjected surviving indigenous resources to a new round of pressures. However, resistance to the Norwegian government's plans to build a new hydroelectric project, induced the Saami to identify as indigenous, although they had not previously been identified as such (Paine, 1982).

Imperial Beginnings

The contemporary notion of indigeneity has its beginnings in the European colonization of the New World, and the appearance of settlers intending to make their homes in places that were already inhabited. It can be conveniently dated from Columbus's landfall in the Americas in 1492, although events subsequently followed a similar course in Australasia, parts of Africa, and some Pacific islands. Settler societies commonly identified indigenous peoples using generic terms such as 'natives,' in the Americas 'indians,' aborigines, kaffirs, kanakas – not the names by which the subjects had hitherto identified themselves – thus defining them in relation to the settlers rather than recognizing differences that might have existed among themselves. Settler colonization entailed the subjugation of the indigenous population, followed by the extraction of precious metals and the establishment of plantations and farms. These often entailed the expropriation and displacement of the indigenous population. In some places, they were required to work for the colonists; in others, they were consigned to the margins of the emerging settler society, which had little use for them or their land. In some colonies, these arrangements took the form of formal treaties, not always observed; in others, the indigenous groups were confined to church missions or government settlements.

In the heyday of settler colonialism, indigenous populations were of lower status, an opinion reinforced by theologians, philosophers, and scientists: not merely different, but as inherently inferior, or in the words of early anthropology,

'primitive.' The fact that they were not, unlike the colonists, Christians, provided the opportunity for missionaries to 'convert the heathen,' although rarely with certainty that their flocks had become 'truly Christian.'

Indigenous Peoples in Settler States

Starting with the American war of independence from Britain in the eighteenth century, followed by the revolt of Spain's colonies, most of the world's settler colonies had either won or, as in the case of Canada, Australia, and New Zealand, been granted independence from the metropolitan power by the turn of the twentieth century. (French and Italian settlers were driven out of North Africa mid-century, leaving only South Africa until the end of the millennium.)

Often enough, this independence gave the Creole element greater license to exploit or dispossess the indigenous population. However, while some remained setter states, many imagined themselves as nations (Anderson, 1991: Ch 4), whether a nation of citizens, which at least in principle if not in practice, included indigenous people, or a system of stratification. In the latter case, indigenous difference was formally recognized, in law or treaty, often with a special government agency having charge of their affairs. As such, they were consigned to a peripheral role in national affairs or excluded altogether.

Given the dislocation and the social and cultural disruption suffered by indigenous populations under the impact of settler colonialism, the original boundaries between indigenous groups, and between indigenous and nonindigenous were often disrupted. There emerged categories whose status was ambiguous, hybrid, or otherwise undefined, with terms such as mestizo, ladino, half breed, half caste, etc. Whether or not recognized in national ideology, such groups may now constitute a majority of the population, but only some will be classified as indigenous; in other classifications, anyone of indigenous descent is classified as indigenous.

Such definitions, or the lack of them, become significant when national governments formulate and implement policies specifically for the government and policing of indigenous minorities. Initially, the solution was often to perpetuate the regimes of the departing colonial administration, with bureaucracies or missions taking responsibility for 'native affairs,' whether under the guise of 'protection' of groups who seemed to be dying out, or controlling the behavior of the survivors. Later there might be attempts to 'assimilate' or 'integrate' indigenous groups, eliminating cultural differences, so that they might gradually become citizens, notwithstanding physical or racial difference. Finally, in the postcolonial period (The term 'postcolonial' has been used mainly to refer to the dissolution of the European colonies, and the new states that emerged. Ironically, the European settler colonies have retained at least some of the features of colonialism, even though they might like to shed the colonial reproach in their relations with the new states.), there might be attempts to recognize indigenous difference, or some version of it, as acceptable to the dominant group.

Government programs for indigenous peoples depend on the resources available, the number of the indigenous population as a proportion of the total population, as well as the resources that the indigenous group still occupy or can claim. According to official estimates (These figures are not all based on the same sets of criteria. As reproduced here, they simply indicate the number of indigenous people a state recognizes as living within its borders.), the indigenous population exceeds 40% in the Andean republics and Guatemala, and ranges between 5 and 20% in New Zealand, Chile, Venezuela, Mexico, Panama, and Nicaragua, and between 1 and 4% in Australia, the United States, Canada, Colombia, Norway, Malaysia, and the Philippines (The Philippines has lately declared certain groups living in Mountain Province as 'indigenous' partly it seems because they were so identified during the American colonial period. Muslim groups living in the south of the archipelago are not classified as indigenous.). The percentage is even smaller for Brazil, Sweden, Russia, and Japan.

However, the size of an indigenous population is no guide to its visibility on the national and global arena. In the second half the of twentieth century and in the new millennium, liberal and left opinion – although not necessarily a majority of the nonindigenous populations – has formulated alternative views on the rights and entitlements of indigenous people, in particular and in general. International bodies including the UN and numerous nongovernment organizations (NGOs) have pressured national governments. This climate has at least intermittently created a space for indigenous mobilization, which governments variously attempt to channel or contain. What happens varies according to the relative powers of indigenous groups, the governments who rule them, and global bodies such as the UN and various NGOs.

Indigenous Peoples in National and Global Perspective

The old colonialist formulation of indigenous peoples as a general category, overriding the particular ways in which they were identified by themselves or local others, underwent a transformation in the mid-twentieth century and into the new millennium, with the idea that all such people were entitled to political, economic, and cultural rights, which had all too often been denied in recent history and were still under threat. Colonial and national governments had in earlier years been left to deal with their indigenous peoples as they thought fit, whereas international bodies and eventually the UN came to provide an arena in which such matters might be raised.

In the context of the Cold War, and the replacement of the old imperial domains by nonaligned new nations, 'colonial attitudes' and racial discrimination became reproaches that most national governments preferred to avoid. The new medium of television transformed what were once domestic issues such as the African-American civil rights movement in the United States or the Bantu struggle against apartheid in South Africa into global dramas, on which the world could pass judgment and which they could relate their to own situations.

The UN made its Declaration of Human Rights in 1948, and this provided the milieu in which in the years that followed, the rights of indigenous people could become an exclusive agenda, leading to the establishment of a Working Group, and in 1993 to the submission of a draft declaration, which would eventually come before the General Assembly. At the 1992 UN Conference on Human Rights in Vienna, where racism was emerging as an issue, it was agreed to appoint a Special Rapporteur on indigenous peoples. Important in this process was the participation of indigenous peoples, with or without the backing of their national governments, whose travel was funded by the UN, the first such occasion being the UN NGO Conference on Discrimination against Indigenous Populations in 1977. Negotiations with national governments followed, and the final Declaration did not come before the UN until 2007 (Eide, 2009: p. 32).

The UN was but one of the opportunities for indigenous people to air their needs and grievance to the world at large. To mention a few key dates – 1971 World Council of Churches' conference on the Liberation of the Indian, the 1975 convening of the World Council of Indigenous Peoples in British Columbia, the meeting of the same body in Australia in 1981, and the World Trade Organization meetings in Cancun, Mexico in 2003. One would have to add to this list, numerous NGOs such as Cultural Survival, Survival International, and the International Working Group for Indigenous Affairs. This has also been a time when various indigenous groups and spokespersons have come together to share experiences with one another and with nonindigenous sympathizers.

It should be noted that in varying degrees, indigenous activism received nonindigenous support from national as well as international intelligentsias, moved by new approaches to social justice, religion, the environment, and changing notions of national identity.

The UN Declaration included some 40 articles, which, briefly summarized, called upon governments to protect indigenous peoples from any kind of discrimination, enable them to practice and restore their cultures, protect their land and resources from encroachment, facilitate the return of lands usurped, and achieve self-determination. In the event, the Declaration received overwhelming support from the General Assembly. It was, however, opposed by Australia, Canada, New Zealand, and the United States, because of the implications of the return of land and the possible implications of self-determination. However, the declaration was 'nonbinding' on the governments who approved it, and they all subsequently gave their support. It is difficult to assess to what extent it has affected the practices of the governments with whom, in most cases, indigenous peoples have still finally to deal.

Merlan in her overview 'Indigeneity, Local, and Global' (2009) argues that, "the impetus for internationalisation of indigeneity has come from liberal democratic 'political cultures'" "the impetus for internationalisation of indigeneity has come from liberal democratic" (p. 304), such as Australia, New Zealand, Canada, the United States, and the Scandinavian countries, whose Lappish population were renamed Saami following their reidentification as indigenous in the 1970s. She later suggests that the "volatile and often authoritarian regimes of Latin America and elsewhere have not provided such a favourable climate for the accommodation of indigenous aspirations." To argue this, however, is to ignore recent developments in many parts of Latin America (Stephen, 1997: pp. 4–5). Moreover, though many of Latin America's governments have been 'volatile,' this does not apply to Mexico, which has seen an orderly succession of single-term presidents since 1920. While Merlan may be right to differentiate the liberal democracies Australia, Canada, the United States, and New Zealand from those of Latin America, this approach weighs her analysis in favor of the center, at the expense of the view from below, where the action often is. This raises not only the question of how strong central government is, and how much space it leaves for local mobilization, but the relative numbers of indigenous and nonindigenous involved. As indicated earlier, the numbers of people recognized as indigenous in most of the Latin American states far exceed that of the liberal democracies, except for New Zealand, where the Maori have long been integrated in national and economic affairs. Moreover, in Latin America they are often concentrated in particular regions, and to the extent that central government has to work through local alliances – whether *caciques* or parties – so that there can be some measure of local empowerment, whether the center wants it or not (cf Rubin, 1996). By comparison, in Australia, Canada, and the United States, indigenous people rarely have much influence even on local events, and even when their affairs are broadcast in the national media, these are dealt with by specialized arms of government. In the discussion that follows, Australia and Mexico are taken, not as prototypical of indigeneity in general, but as case studies that reveal some of its features as well as some of its diversities.

Indigenous Mobilization and Political Process: Mexico and Australia

Charles Hale, in commenting on Merlan has argued that while "Liberal democracies have surely developed their own recognition projects for indigenous peoples….this should not negate the central role of collective indigenous agency and self making…" (Hale, 2009: p. 323). A Latin Americanist himself has pointed up the greater role of indigenous mobilization in these countries than in the Liberal democracies, where the tendency has been top-down. With the exception of New Zealand, it has been possible for the national and regional governments to separate off indigenous affairs from other branches of government, since the indigenes in question, being few in number, have had little in the way of political leverage.

In much of Latin America, there has been a long tradition of indigenous activism, although indigeneity has often been left implicit, in favor of class, regional, or populist party alignments. Marxism, however, attracted national and – particularly from the United States – international repression, and Warren and Jackson suggest that during the 1970s and 1980s, "mobilizing switched to the new discourses of collective rights and cultural diversity." They note that in 16 Latin American countries "constitutional reforms have challenged the dominant imaginary of the nation by redefining the status of indigenous communities and transforming the meaning of citizenship"

(2002: p. 13). However, Mexico is not among the countries on which they base this claim, and it appears to be somewhat different.

Mexico classifies some 15% of population as indigenous, with use of one of the six indigenous languages among the criteria. The nation had a long history of recognizing the indigenous, to the extent of celebrating the achievements of the pre-Columbian civilizations, notably in the Anthropological Museum in Mexico City, although this did not necessarily benefit their living descendants. In earlier years, the national government also promulgated a doctrine of *mestizaje*, according to which indigenous difference would in due course be absorbed into the nation. Among the tasks of the National Indigenist Institute was the setting up of schools in which indigenous children could learn Spanish. This assimilationist approach came under increasing criticism from scholars and students of the generation of the 1960s and after (Lomnitz, 2001). In 1992, the Mexican state revised its constitution to declare itself a pluricultural nation. It was also among those nations that endorsed the UN Declaration.

The relations between Mexican government and its indigenous peoples are more complex, however. Following the revolution of 1910–20, indigenous and other peasant groups had benefited by the land distributions in the 1930s, but national politics from that time, and with only few interruptions up to the present, have been dominated by a single ruling party (Partito Revolucionario Institucional (PRI)), which in complex and variable ways mediated between the center, local power holders, and the populace at large, often to the disadvantage of the last.

Mexico's indigenous population is concentrated in the country's southwest. None of the six ethnolinguistic groups mobilizes as a bloc, and locally they have often been divided by a class aligned with the center, but going back to the nineteenth century, local groups have played a significant part in local politics, sometimes allied with nonindigenous parties, sometimes on their own account. Rubin, in his study of Juchitán, a town in the Isthmus of Tehuantepec, emphasizes that nation and regime exist not only outside of the town and Zapotec culture, but in the very fabric of local discourse and institutions as well (Rubin, 1996: p. 25). Rubin's study is focused on the local Coalition of Workers, Peasants, and Students of the Isthmus (COCEI), which in the 1970s and the 1980s cut across a local indigenous elite, and for a decade dominated local government. Although indigeneity was omitted from its title, the confederation "successfully appropriated Zapotec culture – including its elaboration and leadership – for a poor people's movement. This ethnic appropriation was a crucial ingredient in COCEI's success. COCEI's claim to Zapotec ethnicity … made available for radical politics a common language, a remembered history, dense market and ritual network, and abundant artistic and cultural representation that reconfigured cultural boundaries between 'popular' and high culture" (1997: pp. 42–43). In somewhat similar terms, Howard Campbell argues that, "Rather than opposing ethnicity and class, we need to examine their interplay … by abandoning this artificial division, we can understand how ethnicity is simultaneously a weapon against outsiders and a contested terrain of class interests within indigenous communities" (Campbell et al., 1993: p. 216). Although COCEI at times suffered violent reprisals from the national government, they could also enter into agreement, for example, during the *sexenio* of President Salinas (Rubin, 1996: p. 190). However, the Salinas presidency ushered in a neoliberal economy, including the North American Free Trade Agreement, and an impoverishment of the rural sector (Gledhill, nd).

While COCEI was mostly able to avoid direct confrontation with government, the very name of Chiapas-based Zapatista Liberation Army (Ejército Zapatista de Liberación Nacional (EZLN)), which emerged in 1994, suggested a more radical distancing from central government, and it was met at times by violent reprisals. Contrary to Warren and Jackson's generalization, it is identifiable as Marxist by the internationally diffused pronouncements of its nonindigenous leader, Subcomandante Marcos. The identification with the agrarian hero of the 1910–20 revolution, coinciding with land invasions, appeals to a broader constituency. Gledhill does not deny that "the cultural politics of indigenous identity has salience in the Selva. What needs to be stressed, however – given the diverse origins of those living in the Selva (Stephen, 1997: pp. 7–9) – is that it is a new cultural politics that is reappraising culture history and identity in the light of discourses on human rights" (nd: p. 24). Gledhill continues, "The strength of the Zapatista version of indigenous rights politics lies in its creative qualities and ability to engage universal values." This is "a twentieth century innovation" (Gledhill, nd: p. 24). One can add that its considerable international appeal – the news reached the world at large as soon as it did Mexico City – was almost certainly enhanced by the fact that it seemed to be an indigenous movement, and, following military repression, its retreat into the Lacandón Jungle.

Australian and North American governments have not experienced insurrectionary challenges to their authority in recent times – although there have been symbolic displays recalling such events. Dealing with much smaller indigenous populations, and being wealthier countries, they have been able to contain the indigenous problem through legal and economic means, seemingly approving in general terms the global indigenous rights movement, funding delegations to international conferences and festivals, but keeping it at arm's length when forming policy.

To turn to the case of Australia, the Aborigines have in the last 50 years been able to move from a position of virtual disempowerment, to a position of formal recognition in terms of cultural difference, but increasing economic and administrative dependence on government. This has left them collectively unequal vis-à-vis the majority population (Austin-Broos, 2011: pp. 9–12). Mobilization began in the 1960s and 1970s as a mainly urban 'protest movement.' It took as its model the African-American Black Power movement of the period, only later connecting with Native American Red Power, and other indigenous movements. Among the first such events was a 'Freedom Ride,' exposing widespread discrimination against Aboriginal people in country towns. Reversing the deference characteristic of the previous generation of spokespersons, many of whom had been nurtured by Christian churches, they communicated to the rest of Australia anger and alienation in deliberately confronting words. Taking advantage of the television cameras to stage what

Paine has called an 'ethnodrama,' one such gesture consisted of the 'Aboriginal Embassy,' a cluster of tents and a camp fire on the lawn of Parliament House, as a reminder of the conditions under which many Aboriginal people still lived (Lothian, 2004). The period saw increasing concern about the situation of indigenous people in a prosperous Australia, with liberal scholars as well as the political left drawing attention to discrimination and deprivation, which to many looked uncomfortably like the remnants of colonialism in a postcolonial world.

By degrees, the Federal government took nationwide responsibility for what became known as 'Aboriginal Affairs,' replacing or overriding state governments and church missions. The policies of the 1970s through the 1990s were intended to recognize and respect Aboriginal difference. One prime minister took personal charge of land rights; another made a formal apology for the removal of children under the former assimilation policy.

Although still to be heard, the urban mobilization became secondary to developments in Northern and Central Australia. Although the official definition of Aboriginal included anyone of Aboriginal decent, the difference, physical and cultural, of those living remote from the urban centers seemed more apparent to government and the country at large. A group of Aboriginal pastoral workers had struck for higher wages in the mid-1960s, but when government legislated for equal, the employers' response was to mechanize the industry in such a way as to deprive Aborigines of their main source of employment, leaving them dependant on welfare. Thereafter, the main form of political action was in relation to land rights.

Foremost among the claims of the indigenous mobilization has been for return of at least some of the land that settlement had originally taken from them without any recognition of prior ownership. Partly through government, and partly through the activism of nonindigenous lawyers, appealing to a sympathetic judiciary, it became possible for indigenous groups to claim certain categories of land, through a legal process, requiring their 'elders' to testify to connection from before settlement, but at the same time establish boundaries separating them from other Aboriginal groups. Processes, often prolonged over years and requiring the services of lawyers and anthropologists, are funded by government. Often at the time of reclaiming it, the land returned had no value to nonindigenous interests, but in recent years has proved to have minerals, which industry now wants to exploit. Its offers of royalties and often of jobs in return for access, may be opposed by the owners, but may also divide them along new lines.

Aboriginal Australians had only been able to vote since 1962, and as subsequent events proved, their numbers were rarely sufficient to affect the outcome of a contest in the nation at large, or even one of the states. Government instead set up indigenous representative bodies, notionally to meet the demand for self-determination. Established in 1990, *The Aboriginal and Torres Strait Islander Commission* was the first attempt to surrender a measure of policy-making power to indigenous people, ahead of Norway's Saami Parliament in 1989 and Canada's Nunavut in 1999. However, in 2004, the body was dissolved (Richardson, 2004). The attempt to run a body for the whole of indigenous Australia, disregarding the diversities, was probably a mistake. The need for executive office holders with a degree of sophistication tended to favor urban Aborigines, who were not necessarily in sympathy with these in the remote regions. Subsequently, the government has limited decision making powers mainly to local communities. 'Consultation' preferably with 'elders' is nevertheless an important feature of government practice, albeit often beyond the capacity of busy bureaucrats to achieve (Lea, 2008). With customary practices failing among an increasingly young population, one may wonder whether cultural difference may recede, revealing a stark inequality (Beckett, 2010: p. 43; Austin-Broos, 2011).

In the event, the national government's assumption of responsibility for Aboriginal affairs has committed it to an ever-expanding budget, since it coincided with the closure of the kinds of occupations in which indigenous people had previously made their living (albeit underpaid), while a return to traditional hunting and gathering was no longer an option. At the same time, governments had pledged to 'close the gap' between indigenous and other Australians in the areas of health, housing, education, and incarceration rates. This has left a large part of the indigenous population dependent on welfare or some other government services – a condition that some critics have called welfare colonialism – and that leaves them vulnerable to bureaucratic intervention in their domestic lives.

For a small but growing number of educated indigenous people, there is a parallel dependency on bureaucratic jobs in some branch of Aboriginal affairs. Under such circumstances, the mobilization of this sector, if not quite controlled by government, is diverted along channels laid down by government service.

In addition to the bureaucratic jobs available to school leavers, government has brought into being a small indigenous intelligentsia, including academics. Various arms of government have also sponsored indigenous artists and performers, sending them with their creations to international events. It is mainly from this sector that ideas of a pan- rather than local Aboriginality, its qualities, rights, and future, are articulated. Likewise, it is from this sector that the media, academe, and government seek opinion on policy matters, although there is considerable disagreement on the social and economic viability not only of the remote community, but also of the desirability of employment in the mainstream economy (Langton, 2010).

The preceding thumbnail sketches of the situation of some indigenous people in Mexico and Australia reveal substantial differences, not merely the relative sizes of the two country's indigenous populations and the resources available to the two governments to assist them if they wish to, but the differing emphases on indigeneity, as compared to other forms of engagement with the encompassing society. Australia's Aborigines base their appeal largely on their indigenous status, and their nonindigenous support base seems to view them in the same way, whereas the Mexican groups discussed above seem often to leave their indigeneity implicit – a basis perhaps for internal solidarity – but situate themselves in terms of class (as in the case of COCEI) or national and Latin America-wide revolutionary traditions (as in the case of the Zapatistas). Both groups have been ready to negotiate with other dissident parties within Mexican political life, and occasionally reach agreement with the central government, as these bodies are with them.

Indigeneity, Indigenism, and Reindigenization

The idea of the indigenous evokes in the Western mind people occupying the places they have 'since time immemorial,' passing on to each generation traditional knowledge of pursuits that keep them close to nature, as well as a close community life and forms of spirituality that are in tune with this. The history of indigenous peoples has in greater or lesser degree been characterized by encroachments on this space, and if they can no longer disengage, their lives become hostage to the schemes of the majority. However, some of these groups have become adroit at keeping the encroaching world at arm's length enabling them to preserve vestiges of an indigenous space, perhaps grounded in a distinctive mode of livelihood, relatedness, or religious practice, albeit distorted by the engagement.

The renewed valuation of indigenous cultures – or elements of them on the part of the nonindigenous sector – has given their bearers the space to identify or reidentify, but as Li argues, "as a positioning that draws upon historically segmented practices, landscapes, and repertoires of meaning, and emerges through patterns of engagement and struggle" (Li, 2000: p. 151). With the passing of time, not all such things may still be accessible, nor in the contemporary situation may they be practicable or desirable. Those that are accessible may carry a different kind of baggage in the contemporary situation, becoming reified in the process of 'positioning.' Some practices marked as indigenous, may, at least when compared with earlier accounts, prove to be modeled after prior practices, rather than as continuations of them, and some are simply reinvented.

The renewed acceptance of indigenous cultures has given rise to an ironic rapprochement that Sider and Dombrowski (2001) have called indigenism, referring to the political and the author would add cultural middle ground, which arises out of the engagement of indigenous and nonindigenous meanings, in which understandings are sometimes reached, but sometimes left contested or unresolved. Thus, spirituality – a common trope of indigenism – may be expressed in the form of Christianity, into which a people may have been converted in colonial times, but that has been incorporated into the life of a community so that it can be seen as significantly different.

Arts and crafts are areas where indigenous people are able and increasingly encouraged to practice, by state and commercial sponsorship with creations often modeled after traditional forms rather than replicating them, and directed at the international high art market, as in the case of Aboriginal Western Desert art (Myers, 2002: pp. 252–254), or tourism. National museums both collect and display indigenous creations, and send them overseas. Whether state sponsored or not, these developments may amount to a moment of reindigenization, in the sense of a self-conscious attempt to revive tradition in particular areas of life, if not the whole.

Scholarly Studies of Indigenous Peoples

The discipline of anthropology emerged in the wake of European colonization, as a study of the peoples who had been conquered and their lands exploited and settled by others. The original aim of the research was to document the societies and cultures of the conquered and displaced – 'before it was too late' – which is to say before depopulation or modernity took what was believed to be their inevitable toll. Although responding to newer trends in anthropology, some studies remain of this kind.

However, with the recovery of indigenous populations, anthropology turned its efforts to advising governments on how surviving indigenous groups should be treated. In some countries, national anthropologies emerged, which started with the assumption that indigenous groups were conationals, but while Mexican anthropology of the *indigenismo* period involved forging citizenship "by indigenising modernity and modernizing the Indians" (Lomnitz, 2001: pp. 228–262), Australia in the same period focused on the modernizing agenda, only later coming to some tokenistic idea of indigenizing the nation.

Other anthropologists, along with ethnohistorians, have maintained distance from government in monitoring the impact of settlement on indigenous peoples (Wolf, 1982). These studies, too numerous to mention individually, are usually grounded in the experience of a local community, although attention has recently shifted to the situation of indigenous peoples living in urban environments, away from their home communities. Anthropologists have also used their local knowledge to act as advocates in processes for the recovery, of land, and pursuing cases of racial discrimination.

Studies at the level of the nation state have been rather the province of historians, political scientists, and lawyers, some of whom have also worked comparatively, and in relations between indigenous peoples and the UN. Almost alone among anthropologists, Ronald Niezen has attempted an overview of indigeneity in the human rights framework, although Charters and Stavenhagen have edited a collection detailing the events leading up to the World Decoration on Human Rights (2009).

See also: Americas, Sociocultural Overviews: North America, Native Americans; Arctic Archaeology and Prehistory; Australian Aboriginal Society and Culture: An Overview; Ethnicity, Sociology of; Ethnicity: Anthropological Aspects; Human Rights, Anthropology of; Inequality: Comparative Aspects; Multiculturalism.

Bibliography

Anderson, B., 1991. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, New York.

Austin-Broos, D., 2011. A Different Inequality: The Politics of Debate about Remote Aboriginal Australia. Alan and Unwin, Sydney.

Beckett, J., 2010. National anthropologies and their problems. In: Altman, J., Hinkson, M. (Eds.), Culture Crisis, Anthropology and Politics in Aboriginal Australia. UNSW Press, Sydney.

Campbell, et al., 1993. Zapotec Struggles: Histories, Politics and Representation from Juchitan, Oaxaca. Smithsonian Institute, Washington.

Charters, C., Stavenhagen, R., 2009. Making the Declaration Work: The United Nations Declaration on the Rights of Indigenous Peoples. Copenhagen, Document No. 127.

Dombrowski, K., 2001. Against Culture: Development, Politics and Religion in Indian Alaska. University of Alaska Press, Lincoln.

Eide, A., 2009. The indigenous peoples, the working group on indigenous populations and the adoption of the UN declaration on the rights of indigenous peoples. In: Charters, C., Stavenhagen, R. (Eds.), Making the Declaration Work: The United Nations Declaration on the Rights of Indigenous Peoples. Copenhagen, Document No. 127.

Gledhill, J., nd. Getting New Bearings in the Labyrinth: The Transformation of the Mexican State and the Real Chiapas. University of Manchester (epub), England. [Published in Spanish as Gledhill, J., 2002. Una nueva orientación para el laberinto: La transformación del estado Mexicano y el verdadero Chiapas. Relaciones 90, 201–257.]

Hale, C., 2009. Comment on 'Indigeneity: local and global.' Current Anthropology 50, 3.

Langton, M., 2010. The shock of the new: a postcolonial dilemma for Australian anthropology. In: Altman, J., Hinkson, M. (Eds.), Culture Crisis: Anthropology and Politics in Aboriginal Australia. University of New South Wales Press, Sydney.

Lea, T., 2008. Bureaucrats & Bleeding Hearts: Indigenous Health in Northern Australia. University of New South Wales Press, Sydney.

Li, T., 2000. Articulating indigenous identity: resources politics and the Tribal Slot. Comparative Studies in Society and History 42 (1), 149–179.

Lomnitz, C., 2001. Deep Mexico Silent Mexico: An Anthropology of Nationalism. Minnesota University Press, Minnesota.

Lothian, K., 2004. Moving blackwards, black power and the tent embassy. In: Mast, I., Hannah, M. (Eds.), Transgressions, Critical Australian Indigenous Histories. Australian National University e-press, Canberra.

Merlan, F., 2009. Indigeneity: local and global. Current Anthropology 50 (3), 303–333.

Myers, F., 2002. Painting Culture: The Making of an Aboriginal High Art. Duke University Press, Durham.

Niezen, R., 2000. The Origins of Indigenism: Human Rights and the Politics of Identity. Blackwell, Oxford.

Paine, R., 1982. Dam a River, Damn a People? Saami (Lapp) Livelihood and the Alta/Kautokeino Hydro-Electric Project and the Norwegian Parliament. IWGIA document 45, Copenhagen.

Povinelli, E., 2002. The Cunning of Recognition: Indigenous Alterities and the Making of Australian Multiculturalism. Duke University Press, Durham.

Richardson, S. (Ed.), 2004. The abolition of ATSIC: silencing indigenous voices? Dialogue, Academy of the Social Sciences in Australia 23 (2), 4–61.

Rubin, J., 1996. Decentering the Regime: Ethnicity, Radicalism, and Democracy in Juchitǎn, Mexico. Duke University Press, Durham.

Stephen, L., 1997. The Zapatista opening: the movement for indigenous autonomy and state discourses on indigenous rights in Mexico, 1970–1996. Journal of Latin American Anthropology 2 (2), 2–41.

Warren, K., Jackson, J., 2002. Indigenous Movements, Self Representation and the State in Latin America. Texas University Press, Austin.

Wolf, E., 1982. Europe and the People Without History. University of California Press, Berkeley.

Indigenous Child Welfare Practice: The Canadian Case

Naadli T Ormiston, Indigenous Studies Program, Camosun College, Victoria, BC, Canada; and School of Social Work, University of Victoria, Victoria, BC, Canada
Kundoqk J Green, School of Social Work, University of Victoria, Victoria, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article will explore child welfare practice from an Indigenous perspective. In the article, we begin by centering Indigenous ways of caring for our children. We also provide an overview of how state-controlled jurisdiction over Indigenous children had intergenerational and devastating effects of legislation and policies on Indigenous families. Finally, we review how Indigenous peoples are reclaiming jurisdiction to deliver their own distinct methods of child and family services, in an effort to reclaim, regenerate, and strengthen Indigenous child well-being in British Columbia, Canada.

Introduction

This article explores child welfare practice from an Indigenous perspective. Both authors have personal and professional experience within Indigenous child welfare. Both authors have been instrumental in contributing, developing, and implementing Indigenous child welfare courses and programs at the University of Victoria, British Columbia, Canada.

This article is organized in three sections. Section one provides an introduction to Indigenous ways of caring for our children. Section two provides an overview of state control and jurisdiction over Indigenous children and the devastating effects of legislation and policies on Indigenous families, communities, and nations. Finally, section three examines ways in which Indigenous peoples are reclaiming jurisdiction over ways in which our people deliver their child and family services, in an effort to building healthy individuals, healthy communities.

Traditional First Nations Child Raising

(Throughout this article, we will use terminology of First Nations, Aboriginal, and Indigenous interchangeably depending on what we are speaking to. Many communities in Canada refer to themselves as First Nations people. Many other communities and people refer to themselves as Indigenous people and finally, Aboriginal, is terminology defined by the state.) Before settler contact, Indigenous nations ('Nations' is closely linked to the modern European concept of a 'state.' For Indigenous people, decision-making structure and methods are resolved within Indigenous sovereign states/nation.) were responsible for their own child welfare, and every culture dealt with their child-raising practices differently. There was recognition that children belonged to the community, and that children were the most precious people of the community. For many Indigenous nations, the old people and/or elders sustained and lived traditional and ceremonial practices that strengthened and enhanced the well-being of children and families. For example, rites of passage rituals involved practices that honored all stages of life and in this practice; family members such as aunts, grandmothers, and traditional teachers were involved in raising children. Rites of passage have been important for children, because these not only acknowledge varying life stages, but also the child learns who his or her caregivers are and knows that these caregivers will be nurturing and will provide support throughout their life. Another cultural practice important for the nurturing of Indigenous children and families include learning about Indigenous ceremony. In this, learning important ceremonies and rituals were done through storytelling. In this practice, children learned not only about different forms of ceremonial practices, but children also learned about their histories, landscapes, relations to animals, about other Indigenous communities, and about natural resources that sustained their way of life. This way of life and cultural practices were and are Indigenous knowledge, philosophies, and praxes once utilized to enhance the well-being of children and families.

Ceremonies were and are methods that encapsulate Indigenous living. Hence, ceremony is a way that our people view Indigenous law and policy. As Indigenous peoples, our belief is that our inherent rights, responsibilities, and laws flow from natural laws of what we refer to as inherent and from the Creator. Inherent knowledge or rights for our people instill our sense of belonging, connectedness to the earth and the responsibilities that we have to care for the earth. Importantly, in knowing ceremony and rights of passage, we understand the importance of connecting natural laws to our social responsibilities to take care of our children, families, and communities. In presettler times, ceremony and inherent knowledge was a method practiced by Indigenous peoples to care 'holistically' for children, families, and communities. Prior to settler contact, cultural practices, ceremonies, and rituals strengthened families and the caregiving of children, now known as child welfare.

Commonalities for Indigenous Approaches to Caring for Children

Prior to settler contact in Canada, and perhaps for other Indigenous peoples throughout the world, children were and are viewed as precious and the heart of communities. In this, there were common practices about caring for children that

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28042-2

were nonintrusive to the family and community. For example, if a family were experiencing difficulties in raising their child, another family member would intervene and raise the child. In doing so, the 'other' family members who would intervene, would care for the children and this was done to provide space and time for parents to address their issues. For many Aboriginal nations, the first born was always raised by their grandmother and not by the 'nuclear' parent. In other situations, ceremony was utilized as a practice method to provide support to the family in need. Ceremony is a method that connects the family to all aspects of their being, such as their spiritual, physical, emotional, and mental state. The belief here is that when we collectively look after all aspects of our being, families recognize where there is 'unbalance' and also can see 'issues' they may need to work on. Traditionally, Aboriginal peoples in Canada have a broad definition of family membership and our children were cared for and raised by the extended family.

Storytelling has also been an important practice historically. Children learned the values of respect, honor, love, sharing through stories. They learned how to act in the world around them. Thomas and Green (2009) remind us that stories provided children with the skills needed to survive. "Traditionally storytelling was the primary source of educating our children. Our children were considered gifts from the Creator and, as such, were protected by all in our communities through stories" (p. 101).

Interruptions for Indigenous Child Caring

Disruptions in Indigenous family life in Canada began shortly after contact with European people. During early contact times, the relationships between European and Indigenous peoples of Canada were one of respect. However, over the years and due to development of lands for economic expansions, these relationships began to deteriorate. Relationships succumbed to the notion of 'power over' by the Canadian government toward Indigenous people. This power difference is described historically as the 'protection' phase of relations between settler Canadians and Aboriginal people. One of the earliest documented legislation intended to 'protect' Indigenous peoples is the Royal Proclamation of 1763. The *Royal Proclamation* (The *Royal Proclamation* was issued by King George III in 1763 to officially claim British territory in North America after Britain won the Seven Years' War over the French.) is a document that sought to protect Aboriginal (Throughout this article, we will use terms interchangeably. Indigenous is a term used by international group of Indigenous peoples. Aboriginal is a term used by government and Indian is a term within the Indian Act. Native is also used to describe a group characteristic of indigeneity.) people by providing instructions for European settlement of Aboriginal territories in what is now North America. The Royal Proclamation explicitly states that Aboriginal title has existed and continues to exist, and that all land would be considered Aboriginal land until ceded by what is now known as 'treaty.' The Proclamation provided the first policy preventing individual settlers from claiming land from Aboriginal occupants. The Royal Proclamation further sets out that only the Crown (Canada) can purchase land from Aboriginal people. By way of this law, the concept of the Crown's 'fiduciary' (One who holds anything in trust, or who holds a position of trust or confidence with respect to someone else.) responsibility with Indigenous people was set in motion, which continues to be in effect today. The Proclamation and its fiduciary responsibilities began to set the framework for the relationship between Indigenous people and Canada. This imposed relationship also affected how Indigenous people related to and learned about land teachings, and in turn affected cultural teachings and practices for Indigenous families.

Many more laws and policies emerged in a manner that continued to confine and control an Indigenous way of being. One such act that subjugated and oppressed Aboriginal people was *the Indian Act*, federal legislation passed in Canada in 1876. *The Indian Act* collected all "previous legislation pertaining to Indians into a body of law comprising over a hundred sections" (Lawrence, 2004: p. 33). The Indian Act defines who is Indigenous and who is not, according to the federal government. This act also articulated a reserve system where Indigenous peoples would reside and specified policies on how Indigenous peoples would be governed. This *act* is still in place today. Lawrence (2004) talks of the lingering effects of the 'Indian Act' and its policies:

> The Indian Act is much more than a body of laws that for over a century has controlled every aspect of status Indian life, it provides a conceptual framework that has organized contemporary First Nations life in ways that have been almost entirely naturalized, and that governs ways of thinking about Native identity.
>
> p. 34

Ormiston (2010) also contends that this act of colonialism intends to force the "colonized group into dominant society; adopting policies that suppress … transform or destroy Native values … orientations and way of life … manipulation and management of the colonized … justified by an ideology of racism" (p. 2).

Within the Indian Act, the goal of Canada was to eliminate traditional teachings, values, and significant cultural ceremonies through the banning of the Potlatch, Longhouse, and Sundance ceremonies as well as traditional governance systems. Cultural knowledge, philosophies, and practices were all outlawed as they provided a threatening opportunity for First Nations people to share their history and ways of doing, knowing, being into the future (Thomas, 2005: pp. 244–245). The federal government exercised jurisdiction and control over Indigenous peoples in Canada.

Affects of Colonial Practices

Resulting from colonial governing systems, Indigenous peoples were deemed to be inferior and uncivilized. Nowhere is this more evident than through the creation of residential schools, which resumed the federal government's role of 'protecting the child.' The government of Canada also sought to 'civilize' and assimilate Aboriginal people through residential schools.

Beginning in the late 1800s, the government and the church's attitude was that European culture was superior to Native cultures. Therefore, every First Nation's child was deemed to be in need of protection and it became compulsory in 1880 for First Nations children to attend residential schools (Powderface, 1984). Children as young as 5 years of age were removed from their families and communities for 10 months of the year or longer. In residential school, children were prevented from speaking their languages and taught to reject their traditional lifestyle. As Harold Cardinal (1999) wrote: "These schools were nothing less than state sponsored programs of cultural genocide aimed at Indian First Nations. Residential schools were an integral component of a systematic, intergenerational, state planned program of brainwashing aimed at removing the 'Indian' from the minds and souls of Indian children" (p. 18). The last residential school in Canada closed in 1996.

Most Canadian citizens (settler nonIndigenous) are not aware that in some areas of Canada, almost half of the children who attended residential schools died while in those schools (Mission School Syndrome, 1985). Death of young children at residential school was due to tuberculosis, malnutrition, and other forms of physical, mental, emotional, and spiritual abuse. The residential school of ideology dictated that religion and domestic duties comprised the 'education' deemed important for Native children. In residential schools, the common experience was religious practices, followed by domestic duties. Girls would learn household domestic skills while the boys would learn trades and agricultural skills. Only a portion of each afternoon of a school day was preserved for 'education.' Civilizing and assimilating Native people were the goals of this 'doomed to fail' law and policy.

Protection, civilizing, and assimilating policies in Canada remain a devastating and torturous factor for Indigenous people. Protective and assimilative laws and policies nearly destroyed the cultural knowledge and practices, stories, experiences, and knowledge of the land for First Nations. These laws were harmful, and Indigenous peoples resisted and confronted these injustices. Consequently, throughout the past half century, many Canadian laws have been amended to correct past injustices.

Provincial Involvement

Through the Indian Act in 1876, the federal government of Canada exercised jurisdiction over child welfare for Indigenous peoples in Canada. The way they enforced this was through compulsory education for Indigenous children, as articulated above 'residential schools.' The provinces in Canada became involved in child welfare in 1951 when revisions to Section 88 of the federal *Indian Act* extended provincial laws of general application to Indians and lands belonging to Indians. At that time, provinces were not involved in the 'well-being' or welfare of Native children. The amendment extended provincial child welfare activities and practices to children on reserves. Without Section 88, provincial legislation would not apply to Indians because Canada's constitution gives the federal government the exclusive power to legislate laws and policies regarding Indians. The changes were somewhat controversial, the federal government feeling that any developing provincial laws would be deemed unconstitutional while Indigenous peoples fought against this, feeling that the federal government was reneging on their fiduciary responsibility.

By the 1970s, First Nations peoples across Canada had started to confront and challenge these issues. Although amendments were intended to address Native issues, public policy continued to promote assimilation of First Nations children into the dominant culture. The federal government began to dismantle the residential schools in the 1960s and 1970s and replaced them with 'day schools' (which are still controlled and operated by Indian Act civil servants). As residential schools were 'phased out,' government continued to assimilate Indigenous peoples through practices of child removal, which is remembered by many Indigenous people in Canada as the '60s scoop.' (A term used to describe the highest numbers of adoptions, which took place in the decade of the 1960s. These were so described because, in many cases, children were literally scooped from their homes and communities without the knowledge or consent of families and communities (http://www.originscanada.org/aboriginal-resources/the-stolen-generation/).) This government practice marked a dark period of history for Aboriginal people because the adolescents who returned from residential schools never experienced or received love and affection within these schools. The only parenting they knew for up to 10 years was from the religious order (nuns and priests) and they had been deprived of opportunities to develop positive cultural parenting skills. When these residential school survivors became adults and parents, the government ultimately labeled them as incompetent and unable to raise their own children, and deemed them to be in social need. In many circumstances, this was the government's justification for removing these children and placing them into provincial care. Throughout the 1960s and 1970s, the government continued to apprehend Aboriginal children in the province of British Columbia. According to Patrick Johnston (1983):

> In 1955, there were 3433 children in care of B.C.'s Child welfare branch. Of that number, it is estimated that 29 children, or less than one percent of the total were of Indian ancestry. By 1964, however, 1466 children in care in B.C. were of Indian extraction. That number represented 34.2 percent of all children in care.
>
> p. 23

Today the statistics are even worse. Newly released data from the National Household Survey suggests that "of the approximately 30,000 children in care in Canada in 2011, 14,225 (48 percent) were aboriginal" (Woods and Kirkey, 2013). Clearly these staggering statistics show that almost half of the children in care in Canada are Aboriginal even though Aboriginal people comprise only 3% of the overall Canadian population.

There are many reasons 'defined by government' why Native children are placed into foster care. The most common rationale for the removal of children is neglect, poverty, housing, and substance misuse. These variables are all areas in which the child welfare system takes action for removal, as articulated by

Cindy Blackstock, Executive Director of the First Nations Child and Family Caring Society of Canada:

> "What we have here is a very dire statistic for children who, just like their parents in many cases, are being removed from their families because of state neglect", Blackstock said. "The government is simply not giving these children the same opportunity to grow up with their families that all other Canadian children enjoy."
> Quoted in Woods and Kirkey, 2013: np

For over a century, the government of Canada has maintained and enforced laws, which confined Indigenous people to small parcels of land, defined who is or is not a Native person, and outlawed many ceremonies and Indigenous governance systems, which had been in place. Moreover, Canadian law forced Native people to adapt and live according to European standards. Indigenous peoples became subservient to colonial laws and policies, a term described by Antonio Gramsci (1971) as 'hegemony.' Hegemony is a way of thinking and being, and it occurs when oppressed groups take on dominant group thinking and ideas uncritically. Hegemonic thinking and living is the definitive way to colonize people; the colonized continue the cycle of colonizing themselves (Smith, 2003).

Ultimately, as a result of these laws, Indigenous people throughout the country including those living on their reserves became poverty stricken. Due to this forced assimilation, which marginalized our people further, Native people face racism and stereotyping, which affects our ability to successfully and with confidence to partake in employment, education, and governance.

Interim Measures – A Step in the Right Direction?

Over the last half century, Indigenous leaders have confronted both the provincial and Canadian governments and demanded that child welfare for Indigenous children be centered within Indigenous policy, practices, and worldviews. This includes the necessity for autonomy and authority over the governance of child welfare for Indigenous communities. According to Armitage (1995), "The vision of independent jurisdiction with its own laws is clear, but the process of putting it into place and the form that it will take are not yet defined" (p. 131). In the 18–23 years since the publication of *Liberating our Children, Liberating Our Nations* (Community Panel, 1991) and since Armitage's identification of First Nations' vision for independent jurisdiction, the self-determining vision remains unfulfilled. Currently the Provincial government in British Columbia, the Ministry for Children and Family Development (MCFD) provides authority over voluntary services, guardianship, and child protection through a broad range of child welfare services, including services that protect children from abuse and neglect, services for children with disabilities, adoption services, and family support programs. MCFD delegates these services to Indigenous agencies and communities under the authority of the *Child, Family, and Community Service Act* and the *Adoptions Act*. Through delegation within this governing system, approximately 140 of the 198 First Nation communities in British Columbia have been delegated by the province to provide child welfare for their communities. Communities that choose this type of 'delegated responsibilities' provide unique cultural services that meet the needs of their people. Yet, the overall policies and laws are still under the auspices of the provincial government.

Analysis of This Policy Story

Today, some Indigenous peoples view 'delegated' models of child welfare as a tool of colonialism, based on dominant culture's assumption that the "values of other cultures are abnormal and in need of correction" (Community Panel, 1991: p. 5). Brown et al. (2002) have also identified British Columbia's provincial delegated authority as neocolonial and go on to say:

> Delegated services are founded on the notion of 'giving' authority to deliver child welfare services, rather than recognizing First Nations inherent authority to care for their children. Further, it imposes a way of thinking about and practicing child and family services that is based on mainstream concepts, beliefs, and practices. While delegation allows for some adaptation to accommodate culture and promote community empowerment, it doesn't challenge the fundamental beliefs that construct concepts of child protection. Is delegation wrong? Not entirely. It provides a mechanism for community empowerment …. It is about working inside a box constructed by someone else.
> p. 146

Although the neocolonial landscape of delegated authority currently persists, First Nations people have, in fact, demanded the recognition of the "rights of our people to take responsibility for our own lives" and the "inherent right of our people to govern ourselves" (Community Panel, 1991: p. 29). There is a strong belief among Indigenous people that jurisdiction and legislation over Indigenous child and family services should be returned to them and that legislation must recognize the paramountcy of family and child law enacted by Indigenous Nations.

The Royal Commission on Aboriginal Peoples (RCAP) was a federal government research project where a panel of Indigenous researchers visited many Aboriginal communities between 1991 and 1995. The mandate of the Commission was to study the evolution of the relationship between Aboriginal peoples, the government of Canada and Canadian society as a whole. The Commission heard from many First Nations people who expressed the need for Aboriginal-specific legislation at the federal level in the area of child welfare. RCAP noted:

> In our view, the authority for legislating child welfare and regulating practices should rest with the First Nations people. Native people argued that with Aboriginal jurisdiction in place, there will be greater flexibility to introduce practices in keeping with Aboriginal culture and community realities.
> Royal Commission Report, 1997: p. 376

The Spallumcheen Experience

In 1980, the Spallumcheen people in British Columbia established the first Aboriginal child welfare agency, through

a band-by-law model, (Within the Indian Act, 'Band' is used to identify a First Nations community in Canada. First Nations Bands are given a Band number that identifies their community. In addition to allocating a Band number, each member of the Band receives their own individual status number, identified by the federal government.) which releases them from provincial laws and standards (Royal Commission on Aboriginal Peoples, 1996; Union of BC Indian Chiefs, 2002).

Spallumcheen passed a written bylaw, declaring their own jurisdiction over their children. The Minister of Indian Affairs, who has a veto power over band bylaws under the *Indian Act*, did not veto the bylaw. Spallumcheen people sought out provincial government's approval by organizing a march and demonstration at the then provincial Minister of Human Resources, Grace McCarthy's house, where almost 3000 people including children from the Spallumcheen community and surrounding First Nations people attended. This action pressured the province into recognizing Spallumcheen bylaw. McCarthy signed a handwritten agreement stating Spallumcheen children in British Columbia were recognized through the Spallumcheen bylaw. This bylaw was the first contemporary case of an Aboriginal nation-governing system to reassert its authority and jurisdiction over its children. As of 2009, according to Chief Wayne Christian, Spallumcheen continues to be the only First Nations community in Canada whereby they have their own legislation for the care of their children. In this model, Spallumcheen social workers are authorized by their own community governance to provide child and family services to their membership and to take full protective action when warranted. This is an example of full jurisdictional and legislative authority over child welfare services resting in the hands of Indigenous government.

This governance model for child care allows the Spallumcheen people a greater degree of authority over child and family service matters involving their people. Although this form of community governance has been effective for Spallumcheen people, the federal government stated that they would never sign or approve another governing system as this. Hence, Spallumcheen is the only First Nations community in Canada to implement this bilateral agreement. It should be noted, however, that the Canadian government indicated in 1999 that it would not recognize any other First Nation bylaws guaranteeing sole jurisdiction over Aboriginal Child and Family Services in British Columbia to First Nations (Storey, personal communication, 2002). The Union of BC Indian Chiefs (2002) perceives this as detrimental and states:

> A child welfare bylaw would allow the federal government to recognize Indigenous Peoples' partial re-assumption of jurisdiction for child welfare without requiring new legislation, and would represent an improvement over the current system where Canada requires that Indigenous Peoples seek delegated provincial authority.
>
> p. 38

The Spallumcheen movement led to a number of First Nations communities across Canada looking for more formalized relationships in which they could take over their own child and family services. Green et al. (2012) expand on the Spallumcheen experience with a call to other Indigenous leaders to rewrite policy in the struggle to reduce apprehensions and reclaim Indigenous children: "It is time for our programs, our leaders, to think outside their 'policy box' and outside our circle and to be creative in child welfare development" (p. 110). Clearly, the passing of a Band Bylaw resolution, as was the case for Spallumcheen, led to a return of full jurisdiction and authority over the well-being of their children!

Delegation

The province of British Columbia is similar to other jurisdictions in Canada, in that they may delegate their authority for delivering child welfare services to specific groups in accordance with provincial legislation. In British Columbia, this is done so through the *Child Family and Community Service Act*, (1996) ((CFCSA) or the 'Act'). Social workers are required to receive provincially legislated training for three designated areas in order to provide the necessary services. Caring for First Nations Children Society (an Aboriginal run agency) trains many of the Aboriginal and non-Aboriginal social workers hired by Aboriginal child welfare agencies in British Columbia. The legal authority of the British Columbian government to enter into delegation agreements is expressed in section 90 of the CFCSA.

For the purposes of this act, the Minister may make an agreement with any of the following:

1. an Indian band or a legal entity representing an Aboriginal community;
2. the government of Canada, the government of a province of Canada, or the government of a jurisdiction outside Canada, or an official agency of any of these governments; and
3. any person, or group of persons (1996).

The acceptance of the delegated model by First Nations agencies, communities may be seen as a capacity building step toward self-determination, one which will allow the Aboriginal agency to develop the infrastructure, services, and human resources to provide quality service for Indigenous peoples. A recurring theme in the document *Liberating Our Children: Liberating Our Nations*, was that "the paramountcy of Aboriginal law is necessary to properly recognize the inherent Aboriginal right to the care for Aboriginal children" (Community Panel, 1991: p. 18). The Aboriginal people's council questioned whether the delegation model is a step toward the eventual recognition and affirmation of inherent Aboriginal rights regarding children, or whether it is government's way of absolving its financial responsibilities (p. 3).

In 2002, Ormiston interviewed 16 workers from 4 Indigenous agencies in British Columbia to identify and analyze the effectiveness of the child welfare delegation model currently operating in British Columbia, from an Indigenous perspective. What clearly flowed from the research were the following recommendations:

1. Communities should begin exploring the idea of developing and implementing separate legislation in the area of child and family services.

2. Preventative services in Aboriginal communities should be the priority, with a strong emphasis on culture and traditions.
3. More focus on competency-based hiring criteria for social workers; less emphasis on formal education.
4. Providing education and awareness on a regular basis within the community as part of the agency's mandate.
5. Ongoing awareness/education is essential for ministry staff in terms of how delegation works and the role of Aboriginal agencies.
6. Agencies should be encouraged to negotiate 'blanket' funding arrangements, which will allow them more autonomy in choosing how to allocate their funds.
7. There should be a clear division between child welfare agencies and band councils (Ormiston, 2002).

Current and Future Strategies of Child Welfare Resurgence for Our People

In 2008, MCFD was in the process of transferring legal authority for child and family services to regional community governance structures. They planned to establish parallel Regional *Aboriginal* Authorities by the end of the fiscal year 2008–09. While there was strong support for the future establishment of Regional Aboriginal Authorities, these authorities never came to fruition. The government did not have the political will in the end for this to get passed in the legislature. The British Columbia Assembly of First Nations (BCAFN), along with other Aboriginal policy tables, continued participating in the struggle for jurisdiction over Indigenous children in this province. This was stated in their declaration as:

> That the Ministry of Children and Family Development enter into consultations with regional and provincial planning bodies, duly organized by Aboriginal communities, to begin drafting child welfare legislation specific to Aboriginal children that mandates Aboriginal approaches to child and family development.
>
> BCAFN, 2011: np

Paul Lacerte, Executive Director of the BCAFN expands on the importance of this imperative:

> We are involved in policy discussions with the Ministry on shifting provincial financial resources from paying non Aboriginal people to care for our children, primarily by way of foster homes, towards investments on family supports, family violence and neglect in our communities.
>
> Personal Communication, 2013

Lacerte speaks to the importance of making families resilient and self-determining once again, as opposed to the removal of children, which is what child protection in British Columbia currently does and shifting resources to prevention and early intervention work.

The Indigenous Child Welfare Research Network (ICWRN) was developed through the University of Victoria in 2008. Its vision is to "establish a Research Institute in the province of British Columbia to ensure our children and families are nurtured with stories, traditional teachings and practices relevant to our diverse territories, languages and teachings" (Indigenous Child Well-being Research Network, n.d.: np). This network provides research training to communities throughout British Columbia. Research training here is defined, developed, and implemented from diverse and distinct cultural teachings. In essence, these training sessions teach that 'stories' and 'cultural knowledge' is knowledge that can serve to develop programs, policies, and laws to strengthen children, families, and their communities.

In Conclusion

We will close with our personal Indigenous teachings that inform the importance of cultural knowledge as a foundation for contemporary child welfare practices. Ormiston who is Tlingit, speaks to how Tlingit values inform best practices. Tlingit people are located in the northern part of Canada and Alaska and for Tlingit people they used the 'consensus' process embedded through teachings of the potlatch to honor relations with one another and a method to make important decisions, based on philosophies of respect. When reflecting on traditional teachings and ways of 'living' today within community and in the field of 'social work,' we realize that Tlingit philosophies are critical when developing and implementing new ideas, relations, programs, and the overarching policies that guide these formations. Westerners may critique this 'consensus process' as being time consuming or resource driven, but Indigenous traditional teachings say that it really is the process, the time taken in building relationships and ensuring decisions are made collectively, which is important and reflective of an Indigenous paradigm.

Green who is Haisla, speaks to notions of 'helping' from a Haisla perspective. For Haisla people who are located in the northwest coast of British Columbia, there is a term used 'gywaghlaab,' which means helping one another. The philosophy here is that when a family notices another family faced with challenges or problems, they stop what they are doing to help the family. If there is a community in distress, the Haisla community gathers to help them. A notion of 'helping' was and is a way of life – there were no rules of how to or when to help. People knew when to help. For Tlingit and Haisla teachings and philosophies, cultural knowledge and practice was policy that strengthened and maintained Indigenous peoples for many years. There are important Western systems and Indigenous traditional child-rearing practices that can be used 'together' to develop and implement appropriate services for Indigenous children, families, and communities. It is essential for all social workers to understand the history of how Canadian laws and policies have confined Indigenous people. In doing so, we must collaborate on ethical and cultural practices to ensure that Indigenous knowledge; ceremony and culture are central to enhancing the well-being of children and their communities.

See also: Child Protection: International Issues; Child Protection; Children and Families in Social Work; Fatherhood; Genetics and Indigenous Communities: Ethical Issues;

Historical Thought and Historiography: Indigenous Cultures in the Americas; Indigenous Knowledge and Intellectual Property Rights; Indigenous Psychology; Indigenous Rights; Indigenous Social Work; Motherhood; Risk Assessment in Social Work; Social Work Theory.

Bibliography

Armitage, A., 1995. Comparing the Policy of Aboriginal Assimilation: Australia, Canada, and New Zealand. UBC Press, Vancouver, BC.

British Columbia Assembly of First Nations, 2011. Governance Toolkit: A Guide to Nation Building. British Columbia Assembly of First Nations, Vancouver, BC.

Brown, L., Haddock, L., Kovach, M., 2002. Watching over our families and children: lalum'utul' smun'eem child and family services. In: Wharf, B. (Ed.), Community Work Approaches to Child Welfare. University of Toronto Press, Toronto, ON, pp. 131–151.

Cardinal, H., 1999. The Unjust Society. Douglas & McInytre, Vancouver/Toronto.

Community Panel, 1991. Liberating Our Children, Liberating Our Nations. Queens Printer, Victoria.

Finney, S., Green, J., Brown, L., 2009. Towards transformational research for and with indigenous communities: the new BC indigenous child welfare research network. First Peoples Child & Family Review (An Interdisciplinary Journal Honoring the Voices, Perspectives and Knowledges of First Peoples through Research, Critical Analyses, Stories, Standpoints and Media Reviews) 4 (2), 161–164.

Gramsci, A., 1971. Selections from the Prison Notebooks of Antonio Gramsci. International Publishers, New York.

Green, J., Kovach, M., Montgomery, H., Thomas, R., Brown, L., 2012. Witnessing wild woman: resistance and resilience in aboriginal child welfare. In: Foster, L., Wharf, B. (Eds.), People, Politics and Child Welfare in B.C. UBC Press, Vancouver, pp. 97–116.

Green, J., 2009. Decolonizing social work practice through oolichan fishing. In: Sinclair, R., Hart, M.A., Bruyere, G. (Eds.), Wicihitowin: Indigenous Social Work in Canada. Fernwood Press.

Green, J., Thomas, R., 2005. Learning through our children, healing for our children: best practice in first nations communities. In: Dominelli, Lena (Ed.), Communities in a Globalising World: Theory and Practice for Community Empowerment. Ashgate Publications, Aldershot.

Indigenous Child Well-being Research Network, n.d. University of Victoria. Retrieved from: http://icwrn.uvic.ca/.

Istchenko, V., Atkinson, J., 1988. The Mission School Syndrome (Motion Picture). Northern Native Broadcasting, Canada.

Johnston, P., 1983. Native Children and the Child Welfare System. James Lorimer and the Canadian Council on Social Development, Toronto.

Lawrence, B., 2004. From sovereign nations to "a vanishing race". In: "Real" Indians and Others: Mixed Blood Urban Native Peoples and Indigenous Nationhood. UBC Press, Vancouver, pp. 25–44.

Ormiston, T., 2002. Aboriginal Child & Family Services Agencies: A First Nations Analysis of Delegated Services (Unpublished masters thesis). University of Victoria, Victoria, BC.

Ormiston, N.T., 2010. Re-conceptualizing research: an indigenous perspective. First Peoples Family and Child Review 5 (1), 50–60. Retrieved from: http://www.fncaringsociety.com/volume-5-number-1-2010 (accessed 04.10.13.).

Powderface, S., 1984. Self government means biting the hand that feeds us. In: Little Bear, L., Boldt, M., Long, J.A. (Eds.), Pathways to Self-determination: Canadian Indians and the Canadian State. University of Toronto Press, Toronto, pp. 164–168.

Royal Commission on Aboriginal Peoples, 1996. The Report on the Royal Commission on Aboriginal Peoples. Retrieved from: http://www.ainc-inac.gc.ca/ap/rrc-eng.asp (accessed 03.12.13.).

Smith, G.H., 2003. Indigenous Struggle for the Transformation of Education and Schooling. Speech Presented to the Alaskan Federation of Natives (AFN) Convention, Anchorage, Alaska.

Thomas, R., 2005. Honoring the oral traditions of my ancestors through storytelling. In: Brown, L., Strega, S. (Eds.), Research as Resistance: Critical, Indigenous and Anti Oppressive Approaches. Canadian Scholars Press, Toronto, pp. 237–254.

Thomas, R., Green, J., 2009. Indigenous children in the centre: indigenous perspectives on anti-oppressive child welfare perspectives. In: Strega, S., Carrier, J. (Eds.), Walking This Path Together: Anti-racist and Anti-oppressive Child Welfare Practice. Fernwood Publishing, Halifax.

Union of BC Chiefs, 2002. Calling Forth Our Future: Options for the Exercise of Indigenous Peoples' Authority in Child Welfare. Union of BC Indian Chiefs, British Columbia. Retrieved from: http://www.ubcic.bc.ca/files/PDF/UBCIC_OurFuture.pdf (accessed 03.12.12.).

Woods, M., Kirkey, S., 2013. Tragic' Number of Aboriginal Children in Foster Care Stuns Even the Experts. Canada.com. From: http://www.canada.com/health/Tragic+number+aboriginal+children+foster+care+stuns+even+experts/8354098/story.html (accessed 05.12.13.).

Indigenous Knowledge and Intellectual Property Rights

Jane E Anderson, Department of Anthropology and Museum Studies, New York University, NY, USA; and Institute for Intellectual Property and Social Justice, Howard University of Law, Washington DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article aims to give both an overview as well as to develop certain critical thinking tools that can help a reader engage with the complexity and diversity of Indigenous knowledge and intellectual property rights. It considers the cultural specificity of intellectual property law as well as looking at sites of contemporary struggle. Embedded in the article are questions of knowledge and power as well as how historical legacies of colonialism continue to affect the way we interpret the world and its cultures. The article also raises questions about legal authority and the way frameworks for participation and interpretation in legal contexts are produced and made available to parties that have been marginalized from such spaces.

> Research within late-modern and late-colonial conditions continues relentlessly and brings with it a new wave of exploration, discovery, exploitation and appropriation. Researchers enter with good will in their front pockets and patents in their back pockets, they bring medicine into villages and extract blood for genetic analysis. No matter how appalling their behaviors, how insensitive and offensive their personal actions may be, their acts and intentions are always justified as being for the 'good of mankind'.
>
> Smith, 1999: p. 24

> Indigenous peoples have the right to maintain, control, protect, and develop their cultural heritage, traditional knowledge, and traditional cultural expressions, as well as the manifestations of their sciences, technologies, and cultures, including human and genetic resources, seeds, medicines, knowledge of the properties of fauna and flora, oral traditions, literatures, designs, sports and traditional games, and visual and performing arts. They also have the right to maintain, control, protect, and develop their intellectual property over such cultural heritage, traditional knowledge, and traditional cultural expressions.
>
> Article 31, *United Nations Declaration on the Rights of Indigenous Peoples*, September 2007

Introduction

The intersection of Indigenous knowledge with intellectual property law has produced one of the more complicated legal problems in the contemporary present. This is partly because the issues regarding the protection of Indigenous knowledge are not only legal in nature but also raise questions of historical injustice and colonialism, and at their core reveal that intellectual property law is a culturally specific body of law that was designed for a very specific purpose. Despite its claims to universality, there exist knowledge systems that intellectual property law was never designed to incorporate or protect.

Another complicating factor is that discussions about this issue are not only located within the legal discipline, but are spread across the sciences and humanities. This means that multiple discourses, with very different objects of inquiry, produce different and sometimes contradictory interpretations. Moreover, the problems that intellectual property law raises for non-Western knowledge systems are directly experienced in communities across the globe. Unfortunately these experiences are not always included in the disciplinary analyses that tend to focus more on abstract elements like definitions, boundaries of knowledge systems, and legal rationalities. This results in multiple registers of meaning – those produced through experience and those constructed through disciplinary foci. This leads to different interests and interpretations of what makes this problem to start with, and importantly, what options for adequately dealing with it are available.

Indigenous representatives in multiple countries and contexts including (but not only) Australia, the United States, Indonesia, Brazil, the Philippines, India, Canada, Vanuatu, Cook Islands, Kenya, Peru, Namibia, Mexico, Jordan, Russia, and Sweden argue that they have legitimate rights to control, access, and utilize in any way, including restricting other's access to, knowledge or information that derives from unique cultural histories, expressions, practices, and contexts. A key concern in the intersection of Indigenous knowledge with intellectual property law has also been the inevitable translation of Indigenous knowledge systems into frameworks of western property, complete with principles of alienation, exclusion, and exclusive possession. Such property systems are often not culturally compatible, as the translation into western paradigms requires foundational transformations in how knowledge is understood and how it is shared. Such transformation often sheers off the complexity, interconnection, and intelligibility of these knowledge systems in order to render them legible for property-style protection. This translation and transformation raises very real concerns from Indigenous peoples.

One of the most significant challenges for the protection of Indigenous knowledge systems is that intellectual property has a unique European derivation, and this informs its modes of classification, identification, and operation (Rose, 1993; Drahos, 1996; Sherman and Bently, 1999; Deazley, 2004). Intellectual property law promotes cultural interpretations of knowledge, ownership, authorship, and property. For intellectual property law, the individual as author, genius, owner, and creator hold a central position. These frameworks do not necessarily correspond or compliment Indigenous people's understandings about the role and function of knowledge, or the role of individuals within communities and the joint responsibilities for collectively developed and held knowledge. As intellectual property law has developed from a Eurocentric location and logic, alternative conceptualizations have been very difficult for this area of law to acknowledge and absorb.

There is no easy compatibility between intellectual property law and Indigenous knowledge systems. Their intersection raises issues that involve both legal and nonlegal components, are almost always political, sometimes controversial and can produce conflict in ideals, values, and desired outcomes. Problems, concerns, and claims are not always commercial in nature, and can involve ethical, cultural, historical, religious/spiritual, and moral dimensions. For example, inappropriate use of sacred cultural artifacts, symbols, or designs may not cause financial loss but can cause considerable offence to the relevant community responsible for the use and circulation of that artifact, symbol, or design.

With a field this large and spread across so many different community and disciplinary locations, it is not possible to give voice to every instance, concern, or interpretation. In this regard the article has specific limits. This article is arranged without presuming prior knowledge of Indigenous rights or intellectual property law. By necessity it is also deeply interdisciplinary.

The discussion begins with the one central theme that underpins this issue at very basic and functional levels. This is the role of colonialism, law, and the treatment and construction of Indigenous peoples. The issue of intellectual property and Indigenous knowledge only arises because of this history. Therefore it must be understood in relation to these dynamics especially as they continue to fold into the issues themselves. To not address these is to perpetuate a profound silence of dispossession, presumed superiority, and discrimination: a past that nonindigenous peoples have largely benefited from and that still has significant effects today. This article then moves to address definitions and classificatory frameworks as well as considering the way in which intellectual property law has been made and remade over time. This is important for appreciating the way in which intellectual property law is not ahistorical and timeless, but is constantly evolving as new concerns and problems arise. What makes intellectual property law so powerful and important in our present society is precisely the way it has been able to adapt and develop in reach and scope, and to incorporate new technological and other developments.

The three specific sites to be explored where intellectual property and Indigenous knowledge issues play themselves out include: in art, in archives and museums, and in relation to biodiversity. All are illustrative of the tensions and challenges faced by Indigenous peoples and the very real difficulties for developing new forms of legal and social responsibility and respect for cultural differences in the use and circulation of Indigenous knowledge and knowledge products. This article will then detail some of the developments at an international level raising questions about how, in the international context, the debate itself is constructed and with what kind of participation and by whom. The final section of this article will focus on several recent developments that are designed to enhance Indigenous positions and strategies in both legal and educative ways.

Colonialism, Law, and Indigenous Peoples

Concerns for the protection of Indigenous knowledge have increased in visibility, circulation, and recognition in the last 30 years. Over this period, expectations and strategies for explaining and phrasing concerns have shifted and changed. At the core however, remain the structural, classificatory, and discriminatory legacies of colonialism. These have contributed to the ways in which Indigenous peoples were viewed and treated, and consequently how their knowledge systems have been treated. To be clear, because Indigenous peoples have historically been constructed as 'primitives,' 'savages,' and less than human; as homogenous groups of people despite different languages, customs, practices, and cultures; and as objects of curiosity for study, collection, and research, Indigenous knowledge systems have been treated in the same way. It is this lack of respect and dignity for Indigenous peoples as peoples and as part of sophisticated cultures with inherent rights to their cultures and their lands that make this problem in the first place. Importantly, the failure to account and accept difference on its own terms rather than through the narrow Christian and European lenses offered, continue to play out in many ways. Two of these include how Indigenous concerns in the present can be recognized, heard, and treated as legitimate; and how the same legal systems that facilitated Indigenous dispossession can be rehabilitated to include Indigenous interests.

Colonialism is one of those words that can cause unease. It is also something that many assume they already know enough about – it refers to a period in the past that is increasingly further and further away. Yet colonialism continues to condition the present in multiple ways. It is what makes modernity possible and it continues to function and underpin our modern social order (Quijano, 1999; Mignolo, 1995, 2000, 2011; Moraña, Dussel, and Jáuregui, 2008). Colonialism has many components and many different histories, trajectories, and effects. It is full of conflict and contradiction.

Central to the function and instrumentality of colonialism is law, as it is law that establishes the frameworks and the conditions for colonial conquest, domination, and control. The development of colonial legal systems foundationally affected the ways in which Indigenous peoples, as colonial subjects/objects as well as their laws, customs, and systems of governance, were recognized and treated within the colonial order of things. As Sally Merry explains:

> Colonialism typically involved the large-scale transfer of laws and legal institutions from one society to another, each of which had its own distinct socio-cultural organization and legal culture. The result was a dual legal system: one for the colonized peoples and one for the colonizers … [t]he language of transfer and duality, however, ignores the central feature of colonialism: It was a process in which one society endeavored to rule and to transform another … [t]hus, law, along with other institutions of the colonial state, transformed conceptions of time, space, property, work, marriage, and the state. The role law played in the colonizing process is an instance of its capacity to reshape culture and consciousness.
>
> Merry, 1991: p. 891

Colonial legalities in new territories sought to render already existing laws, customs, and forms of governance as inferior to these newly transferred laws and institutions. Intrinsically they sought to change and reshape social expectations and behavior.

With imposed legal systems underpinned by new Christian moral value systems, those peoples subjected to these new frameworks were also positioned within hierarchical networks of relations that placed the colonists at the top and those 'other' peoples underneath. This 'positional superiority' naturalized discriminatory practices and embedded entitlement to conduct and govern according to a Eurocentric logic (Said, 1979). As colonialism advanced across and into new territories, so too did new governing techniques and new Enlightenment forms of classification and categorization, all of which targeted colonized peoples.

Alongside control of lands, control of knowledge became central to achieving colonial governing ambitions. This rendered many colonized peoples open to invasive and destructive regulatory and research practices. As Linda Tuhiwai Smith observes:

> Indigenous peoples were classified alongside the flora and fauna; hierarchical typologies of humanity and systems of representation were fuelled by new discoveries; and cultural maps were charted and territories claimed and contested by the major European powers. Hence some Indigenous peoples were ranked above others in terms of such things as the belief that they were 'nearly human', 'almost human' or 'sub-human'. This often depended upon whether it was thought that the peoples concerned possessed a 'soul' and could therefore be offered salvation and whether or not they were educatable and could be offered schooling. These systems for organizing classifying and storing new knowledge and for theorizing the meaning of such discoveries constituted research. In a colonial context however, this research was undeniably also about power and domination. The instruments or technologies of research were also instruments for legitimizing certain colonial practices.
>
> Smith, 1999: p. 60

The legacies of colonialism are apparent in how legal frameworks sought to position colonized and Indigenous peoples and their concerns (Benton, 2001; Keal, 2005). Colonial law sought to manage how these 'subjects' were recognized and translated in a myriad of contexts. While there was no 'grand plan' or homogenous colonialism, the 1452 Papal 'Doctrine of Discovery' certainly presented a consistent template for engaging with Indigenous peoples. It expressly gave Christian explorers the right to claim discovered lands that were not inhabited by Christians. In the attempts to recognize and deal with the peoples already existing on lands subject to colonial ambitions there were notable differences across jurisdictions, for example in Northern America, initially Indigenous peoples were treated as distinct entities and formal legal relationships (treaties), based on the prevailing international law, were entered into. This was later replaced by more oppressive ideological doctrines like Manifest Destiny, which subverted Indigenous rights and further justified the taking of lands and territories across the Americas. In other contexts like Australia, which was colonized under the doctrine of *terra nullius* (meaning 'no man's land'), the mere presence of Indigenous peoples was harder to deal with because the doctrine posited that they actually were not there. This resulted in classifications of Aboriginal people as subhuman, and later, childlike, which had obvious consequences in terms of how Aboriginal people were placed under extreme regulation and governmental control well into the 1970s.

The reason behind why all this matters for the present discussion is intellectual property law folds into itself prevailing colonial logics about the value of Indigenous peoples and their knowledges. The development of intellectual property did not happen in some kind of political, social, and cultural vacuum, but was slowly assembled during the same colonial periods and was therefore affected by prevailing rationalities and assumptions about non-European peoples.

Definitions and Terminology

Almost all the terms in the debates on the protection of Indigenous knowledge are contested, have problematic histories, and/or defy simple definition. Indigenous, knowledge, local, community, tradition, cultural heritage, public, culture, property, are a sample of the most prominent of these. As Oguamanam perceptively explains:

> It is the norm that the majority of writings on this complex subject devote substantial effort to the clarification of terms. Without doubt, clarification of key terms is congruous to analytical integrity and guards against misleading assumptions. In some contexts, however it may portend some form of red herring or escapist expedition from the substantive questions, especially if every given phrase or term is a contested one.
>
> Oguamanam, 2009: p. 35

The confusion in definitions and terminology moves in multiple directions and has diverse politics. The starting point for many of the definitional concerns is around 'Indigenous peoples.' For example, contests over who is or who is not Indigenous (which implicitly affects the identification of Indigenous or traditional/local knowledge) derives from both early Enlightenment thinking that sought to classify the stages of humanity and in colonial intervention and governance that subsequently created the means for the identification of peoples as Indigenous, Aboriginal, Indian, and so forth. As many Indigenous scholars have noted, these were not terms that Indigenous/Aboriginal/First Nations/Native American people used for themselves, but were created and became imposed forms of legal and social identification that erased the particularity of specific cultural and community names and their significance to identity making. The modern concept of indigeneity is produced through colonial relations (Popova-Gosart, 2012). The emergence of this concept, as an attempt to identify and classify a new category of peoples also, subsequently, informed the first legal and international documents addressing Indigenous rights. (The first international document addressing Indigenous rights and the responsibilities of States toward these peoples was the 1957 International Labor Organization's Convention 107).

The modern appropriation and current mobilization of national and international Indigenous polities draws from this past but has also been graphically subverted and extended in multiple ways, especially through realignments of citizenship and identity. The 2007 *Declaration on the Rights of Indigenous Peoples* is significant precisely because it recognizes the distinctive nature of Indigenous needs in relation to language, education, culture, health, housing, governance, and representation. As James Anaya explains Indigenous peoples, in a legal context, are now defined as "the living descendants of

pre-invasion inhabitants of lands now dominated by others" (Anaya, 1996: p. 3).

The problem of defining Indigenous peoples inevitably extends into Indigenous knowledge. The search for definitions is embedded in a classificatory logic that seeks clarity in the boundaries. Unfortunately such clear boundaries are impossible. This makes many parties very uneasy. While it might not always be clear what exactly Indigenous knowledge is, there must be a strong resistance to the binary of 'Indigenous' versus 'scientific' knowledge, because it upholds very specific power relations (Agrawal, 2002; Anderson, 2009). Knowledge, whether Indigenous, scientific or other, is notoriously difficult to define and identify and as Indigenous knowledge does not really offer itself as a special case to this problem, we need to be wary of situating certain knowledges as oppositional and easier to demarcate boundaries to than others, as this can reinforce cultural prejudices and colonial biases.

Intellectual Property and '*Gnaritas Nullius*'

Intellectual property is a term that often generates confusion. The potential for literal interpretation and the capacity for everybody who thinks to have some kind of intellectual property of their own only increase the problem. Intellectual property is actually an umbrella term used to cover specific laws that are loosely united in their efforts to manage the relationships between an idea and the tangible expression of that idea (a book, a photograph, an artwork, a sound-recording, a design on fabric, an invention). There is no specific intellectual property law named as such. Rather, the independent legislation for copyright, patents, designs, trademarks, trade secrets, confidential information together constitute the 'laws of intellectual property.' They are grouped under this term 'intellectual property' because they are seen to share some dimension of the problematic of determining legally recognized and justifiable rights in the expression of ideas and treating this expression as some kind of 'property.'

Copyright, patents, designs, etc., are specific laws that evolved slowly and haphazardly in Europe from the late seventeenth century. This slow evolution was in response to cultural, political, social, and economic shifts that occurred throughout this period. Some scholars point to the development of the printing press by Johannes Gutenburg in 1450, increasing the possibility of the circulation of texts, as marking the moment where new laws governing the production and circulation of texts began (Eisenstein, 1980). Others point to the increasing development of nation-states following the Treaty of Westphalia in 1648 that saw the increased need for agreements and the development of reciprocal trading rights within and across newly emerging countries leading to the distinct patent and copyright statutes in Britain in 1623 and 1710, respectively. (Statute of Monopolies, 1623 and Statute of Anne, 1720. For some of the work on these see Deazley and also the Web site). Still others point to the emergence and maturity of concepts of property, like those espoused by John Locke, that helped establish the crucial relationships between property and individuals that continues to underpin current intellectual property law (Rose, 1993). Whatever the focus, it is clear that intellectual property law developed because of a combination of all these elements and was a result of significant political and economic changes occurring in Europe.

The early developments do not feature the terminology 'intellectual property.' Nevertheless the early cases make clear that the challenges in determining this kind of property arise from its intangibility: that it constitutes a product of the mind and therefore cannot be easily demarcated and identified in the same ways as tangible forms of property like land and livestock. For new ideas of property to be developed, law needed to create new categories for identifying the characteristics of this new kind of property. This was because of problems about what exactly the property was, how it could be identified, how it could be measured, how loss could be recognized and compensated, what labor the individual exerted to 'make' a work, and generally how a right to something that was intangible could be justified (Anderson, 2009).

For copyright law, the two most important categories developed were authorship and originality. The making of the category of the author within copyright law, and by implication within society, begins most clearly with the literary property cases in Great Britain in the seventeenth century. Yet it was ostensibly relations between booksellers and publishers that pushed the law to consider the category of the author. What this means is that intellectual property law, in various ways, has been influential in shaping some of the most foundational categories that we use in our society. The author as an individual, a genius and someone that exerts certain property rights over cultural products like writing is completely naturalized: to the extent that it is hard to imagine another configuration. But this norm is produced at the expense of appreciating that not all other cultures make this connection, do not have the same ideas about property nor have these concepts permeate their societies in the same way as they have come to dominate Western contexts. This begins to point to some of the incommensurable concepts that arise in the intersection of intellectual property and Indigenous knowledge – that is, Indigenous knowledge is not easily individualized nor easily abstracted, but derives from and maintains meaning in the contexts in which it was developed and redeveloped. Moreover, there is often not a single identifiable author, rather a community of knowledge holders that act as custodians and keepers of the knowledge but not necessarily owners.

In 2010, Indigenous scholar Greg Younging wrote a paper titled *Gnaritas Nullius (No One's Knowledge): The Public Domain and the Colonization of Indigenous Knowledge* (Younging, 2010). In this paper, Younging explains the complex and varied knowledge systems and knowledge products that Indigenous peoples across the Americas have produced for over 500 years from architecture to irrigation systems, from pharmacology to arts and intricate governance systems. Yet despite these significant innovations, the paradigms that newly arrived colonial peoples viewed Indigenous peoples, their cultures, and inevitably their knowledge systems and products, were so impoverished that they were unable to see Indigenous peoples' civilizations as profound and advanced. Like the land Indigenous peoples were living upon which was, because of the discriminatory and myopic colonial legal doctrines, made into the possession and property of the colonists, so too Indigenous knowledge systems were rendered into either curiosities to be collected and gazed upon or taken and appropriated to be

repositioned as someone else's idea and property. If the land upon which Indigenous peoples had lived for hundreds of years could be constructed into the fiction of *terra nullius* – no one's land, so too was Indigenous knowledge and innovation constructed into the fiction of *gnaritas nullius* – no one's knowledge. And as no one's knowledge it was never offered any kind of respect and certainly no legal protection. Intellectual property law, concerned with offering all kinds of different and varied protections never considered Indigenous knowledge valuable. It actively supported the contexts and nonindigenous individuals that appropriated, stole, and/or made this knowledge into colonial and recognizable forms of (intellectual) property.

Sites of Struggle

So how did this actually occur, what was the role of intellectual property in facilitating the dispossession of Indigenous knowledges, and what can this history tell us about present claims and problems? What follows is also an illustration of the inherent cultural biases built into the law in three different contexts. The way law presents itself as political, ahistorical, and acultural is herein undone. The next three examples show more clearly how intellectual property law has served one cultural vision of the world at the expense of others.

Copyright and Aboriginal Art

Art, in its broadest sense, is one area where intellectual property law effectively functions. From literary texts to painting, music, photography, and the performing arts, all now utilize elements of copyright to protect such works. As mentioned above the key relationship that functions at the core of copyright is that the author is the originator and owner of the work, and as the owner, the author has a range of rights that can be asserted over the work; including distribution, copying, licensing, and publishing.

From the moment nonindigenous people first encountered the cultural practices of Indigenous peoples, they were not considered art proper, but more cultural curiosities. Some products of Indigenous cultural practice, like beadwork and weaving in the Americas became highly sought after, but they were never considered 'art' in the sense that deserved protection. From a colonial perspective there were no clear authors and the works were religious, ceremonial, and functional within Indigenous societies. Such works circulated within an alternative regime of value that was unfamiliar to nonindigenous peoples. Instead, Indigenous cultural products were collected and placed in ethnographic museums and adorned curiosity shops. In other instances they became trophies from frontier battles or were used as items to trade with colonists.

In Australia, Aboriginal art is currently estimated to be worth several million dollars within the Australian economy. The making of this art industry is complex and depends not only on interactions between individual and governmental initiatives (Myers, 2000), but also involves the recognition of Aboriginal art as 'art' and therefore open to the same copyright protections as other works. Like in other contexts, prior to this transformation, Indigenous artistry was constructed by anthropological and ethnographic lenses as 'objects' of culture and constituted as 'nonart' captured through the term 'folkore.' As unprotected and valued for its strangeness, Aboriginal cultural products were collected and displayed in private collections and museums, while the persons responsible for making and painting the Aboriginal works were rarely documented. In the 1920s emerging new Australian artists were encouraged to copy the designs, styles, and colors – everything about the works was open to being copied, reproduced, and reused – without concern whether these uses were sensitive or culturally appropriate. The works were considered 'raw material' for other artists to benefit directly from.

In the 1970s there started to be a visible and voiced concern from Aboriginal spokespeople about the large-scale copying of Aboriginal designs and works onto fabrics, tea towels, and carpets. For the Aboriginal producers (not yet recognized as artists), these copies were not only offensive because they failed to acknowledge the original painters and designers, but they also disrupted important cultural stories and narratives that were embedded in the works. As a result, the 1980s and 1990s saw the Australian courts grapple with a series of cases that focused on inclusion of Aboriginal art as a legitimate copyright work. These cases have been repetitively written about in numerous national and international contexts, partly because of their novelty and partly because they really function as the foundational case law in the entire area of intellectual property and Indigenous knowledge. They showed that peoples whose knowledge was historically considered inferior, timeless and without 'authors' could assert legal title to the works and could be recognized as 'artists' in their own right. Due to the strategic work of several key individuals working on these cases, Aboriginal art was included within the intellectual property paradigm and Aboriginal producers were recognized as individual artists who had legally defensible rights in relation to their works.

Through these cases, intellectual property law was shown to be capable and responsive to the inclusion of new subject matter. This deflected attention away from the conditions that led to the exclusions to start with – namely culturally specific categories (of authorship and ownership for example) that privileged some forms of cultural production and not only ignored others, but also endorsed the wholesale copying for nonindigenous artists' benefit. This cultural specificity and hence bias was really only addressed in the margins of the cases, but the law was unable to expand its own presumed neutral categories to include other cultural interpretations of artistry, and importantly, where that artistry came from if it did not come from the individual 'genius.' The complexity of Aboriginal artistry was flattened into the already existing copyright categories in order to secure protection. For instance, copyright law could not cope with the reality that Aboriginal rights to paint certain images came from rights to land, and that Aboriginal artists painted according to complex governance systems and obligations within their cultural contexts, and were not considered exclusive 'owners,' nor were the works considered 'property.' Fitting an entirely different paradigm of cultural production into the narrow terms of Western intellectual property law has come with costs for Aboriginal artists, and while many artists work around some of these confines, the bias within the law still

perpetuates forms of exclusion while at the same time presenting apparent inclusion.

Archives, Museums, Libraries, and Indigenous Collections

While copyright and Aboriginal art prompted some of the first case law in this field, it is also an example of the specific national contours of the emergence of Indigenous knowledge issues more generally. The actions and reactions in Australia were unique and have opened up a significant area of discussion about Indigenous knowledge rights in a range of contexts. Addressing Indigenous intellectual property rights is now integral to governmental funding initiatives and research collaborations which engage with Indigenous people and their perspectives, authority and governance frameworks.

At least since the 1970s, another area of contest and struggle over intellectual property and Indigenous knowledge has been within archives, libraries, and museums. These contests are international because across contexts the issues tend to be remarkably similar. Cultural institutions that hold Indigenous collections are at the forefront of Indigenous peoples' efforts to achieve access and control of materials relating to their families and communities. These issues are understandably acute in settler colonial contexts, where relations with Indigenous peoples continue to be difficult and historical legacies of discriminatory treatment remain remarkably present.

The struggle for control over Indigenous collections within cultural institutions exists because these sites are primary repositories for enormous amounts of materials collected over the colonial period on the lives, cultures, and languages of Indigenous peoples. This was almost always without the consent of those being recorded, and certainly without any clear explanations of where this material would be kept, who would be able to access it, and that it would often be used to prove the inferiority of Indigenous peoples and their practices. Alongside the troubling history of how these collections were acquired and their uses, one main point of concern is that these collections are also not legally owned by Indigenous peoples but rather, by the person, usually a researcher, governmental official or curiosity seeker, who came to the community and made the documentation – either in sound, in film, through photographs, or even written transcription.

In this context, intellectual property law again finds itself on the side of maintaining unequal relationships. For the nonindigenous peoples, rights to this material, regardless of how the material was collected, are upheld. Moreover, through specific exceptions and limitations within the copyright legislation, cultural institutions are given specific options for making the material accessible, especially material that is no longer protected and now exists in the 'public domain.' This is despite Indigenous claims that some of this public domain material should not be circulated, does not have a time period for circulation, and is to be governed according to complex Indigenous authority and management systems. Again copyright is beholden to its own cultural contingencies. In dismissing any role in the social conditions that led to these collections, copyright law nevertheless endorses those very conditions by vesting the ownership and the property rights with those people who 'made' the works. This legal entitlement granted to the nonindigenous recorders assumes that it is they that have the creativity, the genius, and authorship. Those peoples in the recordings and documents remain passive subjects whose representations remain controlled by others. It is this unequal terrain that makes Indigenous peoples so concerned about the collections within cultural institutions. That the law cannot be retroactively undone also means these legacies – with copyright protection extending far into the future (death of the author plus 70 years in almost all jurisdictions that hold this material) – remain very present and very difficult to negotiate (Anderson, 2012).

Biodiversity and Biopiracy

While the ongoing problems over Indigenous collections are probably one of the most substantial areas of concern given the sheer amount of Indigenous materials within collections, they are not very visible in the literature. The area that is the most discussed, and fiercely debated, is over rights to biodiversity and threats of biopiracy – where concerns oscillate around circumstances where Indigenous knowledge about plants, animals, and medicines, is taken and converted into a 'scientific' language of synthesized distinct properties that can be isolated and patented.

This area appears much more complicated than the copyright issues partly because of the different area of intellectual property law that is involved. Patents themselves are difficult because of the extent of the rights, what exactly it is that is protected and the necessary steps to secure protection. Patenting requires access to sophisticated scientific equipment, access to technical language, and access to extensive financial resources to file the patents and to protect them. In this sense alone there are very specific protections for very specialized research endeavors.

Patents over bioresources have been controversial in many other contexts besides Indigenous ones. The first patents over live organisms and, subsequently, cell lines have produced significant commentary ranging from justification to ethical questions to future implications. The opportunities to patent live organisms and biocompounds have corresponded with significant technological developments in the biotechnology arena, and this has led many researchers to remote and bioresource-rich contexts: contexts often inhabited by Indigenous peoples.

There are multiple examples of patents that have provoked significant condemnation because they are based upon already existing Indigenous or traditional knowledge and therefore should not meet the standard of 'novelty' for the patent grant. The most discussed patents in this context are turmeric, neem, and ayahuasca – and significant efforts have been made to have these patents revoked. Anthropologists like Cori Hayden and Shane Green have provided much needed analysis in tracing the ways in which researchers come to gather botanical samples from Indigenous communities, and the challenges in translation and understanding about what is exactly going on and for whose benefit (Hayden, 2003; Greene, 2004).

One of the most infamous cases of bioprospecting and patenting involve the San people of South Africa, Namibia, and Botswana and the Hoodia plant known for its appetite suppressant properties. A case that has produced a significant amount of literature, it raises issues about the early documentation of Indigenous medicinal knowledge in the colonial

period, how knowledge about these properties was transmitted to researchers who began developing new drugs based on this knowledge, how this activity happening in labs and involving significant financial interests and pharmaceutical companies, was made known to the San and the subsequent negotiations over consent and benefit sharing (Wynburg et al., 2009).

The problem for intellectual property is, again, the conditions embedded in the grant of the property right that privilege certain actors over others. Those that benefit are those that have access to the language, technology, and financial resources. Indigenous peoples are seldom in these positions, and their knowledge is regarded as 'raw material' from which others can extend and benefit from. In this paradigm, there is very limited space to recognize the long and substantial work Indigenous peoples have made to refining this knowledge, nor acknowledgement of the labor that is involved in determining which plants have special properties and which do not. For those seeking to utilize Indigenous environmental knowledge, it is almost assumed that Indigenous peoples developed knowledge about the properties of plants by chance, rather than through innovation, invention, and hard work. This treatment of Indigenous knowledge is completely consistent with the embedded colonial assumptions about Indigenous peoples as backwards and limited. Thus the patent system reinforces this view of Indigenous peoples and their knowledge by not according them legitimate status or authority in the discovery and husbandry of these plants and their medicinal and health-based properties.

International Interest and the World Intellectual Property Organization

Key agencies like the World Intellectual Property Organization (WIPO) have been instrumental in establishing new frameworks for the interpretation of Indigenous interests in intellectual property. This includes setting new parameters for identifying Indigenous knowledge in need of protection. For example, the three discrete areas to refine discussion and policy development are now conceived as: 'genetic resources,' 'traditional knowledge,' and 'traditional cultural expressions/folklore.' In international contexts, these new phrases do more than ameliorate contest. They streamline the discussion in important ways so that the international framework assumes an authoritative position and starts to function as the point of departure for any discussion of the issue. This has the additional effect of enhancing the reputation and authority of the body that is circulating them to start with.

The WIPO has come to hold a central position in relation to debates around intellectual property and Indigenous interests therein. This was not always the case. There is sufficient evidence to suggest that countries like Australia, India, and Canada would have had the debates about intellectual property and Indigenous knowledge regardless of international attention and reworked interagency agendas. (This is because the claims in these countries arise out of the particular colonial politics arising from land rights. Australia's interest predates the WIPO focus by at least 20 years).

Even though discussions about the protection of Indigenous knowledge (specifically biological knowledge) have also been channeled through the Convention on Biological Diversity (CBD) and UNESCO retains expertise in relation to cultural property or tangible and intangible cultural property, WIPO remains the key site for making meaning about the legal possibilities for the protection of Indigenous knowledge through intellectual property law. This speaks to the modern significance and importance of intellectual property law itself. The politics of intellectual property law, and the diverse range of contexts that it reaches into, affects any contemporary discussions about knowledge protection. Importantly, and because of this, it also mobilizes a range of interested parties that might otherwise have no reason to pay attention to Indigenous knowledge issues.

Future Options and Strategies

Despite the historical use of intellectual property law against Indigenous peoples and their knowledge systems, Indigenous peoples also see potential opportunities within this body of law. This forward-looking approach recognizes the power of intellectual property law. But intellectual property law is not a panacea. It still remains firmly grounded in Eurocentric visions of the world, which include ideas of property, the concept of individual creation and genius that happens in a context void of human relationships, and that it exists to foster economic return. In circumstances where Indigenous expectations are malleable and can meet these embedded assumptions, intellectual property law can work quite effectively. When the needs differ, for instance in asserting community rather than individual authorship, or when there should be no time period for protection, intellectual property law is not necessarily the best option.

The problem for Indigenous peoples is that they are experiencing a range of problems in multiple areas of knowledge control and knowledge governance, especially when dealing with nonindigenous people and other third party interests. It is unlikely that one specific legislative development could comprehensively solve all the issues currently being encountered. While the development of an international approach to Indigenous intellectual property issues is important, there is also an urgent need to develop local strategies that are appropriate to both community and context and are more immediately accessible.

The extent of current proposals reflects the complexity of the issues for law, policy, and local/national/regional/international governance. Due to space constraints these cannot all be outlined here but can be found elsewhere (Anderson, 2010). While they are neither uniformly coherent nor necessarily applicable for every instance or in every context, they can be loosely grouped into six separate categories:

- Proposals that modify the current intellectual property framework;
- Proposals that utilize other areas of critical intellectual property discourse;
- Proposals that target private law solutions;
- Combined approaches;
- Alternative regimes; and
- International treaties and conventions.

That the proposals range from international and overarching frameworks (treaties), to localized and targeted strategies (private law making), reflect the different opinions about

what the nature of the problem is and what kinds of solutions might be developed that would provide immediate relief or remedy. Given the complicated dimensions informing this area, a singular solution is perhaps inappropriate. Importantly, any successful intervention in this area is dependent upon the capacity for Indigenous people and communities to make informed decisions about what options are available and appropriate for each problem and each context.

Research Agreements and Protocols

In order to deal directly with the problem of researchers within Indigenous communities and the embedded assumption that the researcher comes, does his/her research and then leaves with all the material gathered, research agreements and protocols have been developed. Protocols have played an important role in establishing new kinds of relationships between Indigenous people, Indigenous communities, and other organizations and/or nonindigenous people.

In general protocols can be understood as context-driven policy. They can be developed to address specific problems and provide guidance in relation to appropriate behavior when it is required. Protocols can incorporate community perspectives, and targeted to particular issues. For example, protocols have been developed for libraries and archives, for visual artists, for collaboration between filmmakers and increasingly for relationships around access to environmental knowledge.

A protocol is a code of conduct, a guideline, or a set of manners that explains how people should behave in certain circumstances. Protocols make new kinds of negotiation possible. They can be used to set community standards around knowledge circulation and use for outsiders as well as help change attitudes and set new standards. Generally, protocols are flexible and can change over time. It is important to see them as a tool to help achieve certain goals that other areas of law have been unable to fulfill. As formal or informal guidelines for behavior, protocols can help build relationships and make new ones possible.

Protocols are prescriptive and they offer Indigenous people a range of options. This is especially in situations when Indigenous people may not want to engage with formal law mechanisms, but rather find means for articulating customary law, or local laws specific to the context. Protocols provide conditions for Indigenous agency in the sense that they can embolden already existing practices rather than imposing new ones. This is one of the central reasons why they have been adopted and are increasingly found across all areas involving negotiations around Indigenous knowledge use.

International Legal Instruments

For the last 12 years, the WIPO has been working on developing an international legal instrument or instruments in this area. Currently there are three instruments being negotiated: one for 'genetic resources,' one for 'traditional knowledge,' and one for 'traditional cultural expressions.' The proposals for international instruments in this area provide an opportunity to streamline activity, ask fundamental questions about what is at stake and for whom, and develop an appropriate approach. Importantly they could offer an overarching authoritative framework for negotiating equitable relationships in the use of Indigenous knowledge.

Yet questions of jurisdiction, enforceability, and constructing foundational concepts remain as significant challenges. The highly contested political nature of Indigenous knowledge debates also slow the progress of development. Other questions about consensus, Indigenous participation in design, voting, and whom the instruments would be between again engage historical legacies that are not easy to resolve. Sensitivity to the range of complex interests, as well as the unequal negotiating positions that Indigenous people occupy in practice will directly affect the success and the usefulness of any new instrument/s.

Within the UN CBD, there has been sustained effort to develop mechanisms that recognize the value and significance of Indigenous knowledge within the life sciences and biotechnology industries. In particular, the CBD has been working to develop an international regime on access and benefit sharing of biological resources. This scheme seeks to develop equitable means for Indigenous communities to derive benefits from any knowledge that they grant access to in the context of genetic resources. Such benefits for the Indigenous community are not necessarily dependent upon the future financial benefits that may be developed. Rather, the scheme aims to create conditions where the knowledge of the source community is recognized and valued at the point of transfer/exchange. Access and benefit sharing is dependent upon the making of agreements between parties that recognize the contribution that each makes as well as the value of that contribution.

The development of an access and benefit-sharing scheme recognizes the value of Indigenous people's knowledge, especially in areas where it has become highly desired. Not only do Indigenous people have the right to be involved in any decision making within research projects that involve themselves or their knowledge, but they also have the right to derive benefits from any products that are developed as a result of these participatory projects. However, the use of contract law, while offering a legal framework to secure Indigenous rights, still requires translation and legal advice. The primary way in which access and benefit-sharing agreements are being articulated is through protocols (Raven, 2006). There remain significant concerns about unequal negotiating positions – not only between a community and a researcher/researchers, but also between a national government and a community. This is particularly the case because what is also being negotiated in this context is the recurring issue of sovereignty over land and hence biological resources.

New *Sui Generis* Legislation

Owing to the perceived difficulties of building new mechanisms that directly address Indigenous needs and expectations about knowledge use and control within the current intellectual property framework, suggestions for an altogether different approach have been made. *Sui generis* law means of its own kind; that is, it is a unique law complete unto itself and often created when current and existing laws are inadequate. (Examples of *sui generis* legislation in relation to traditional/Indigenous knowledge include the USA *Indian Arts and Crafts*

Act 1990; *Panama Law No. 20* (26 June 2000) and *Executive Decree No. 12* (20 March 2001). Another example of *sui generis* legislation relating to Indigenous rights is the Australian *Native Title Act* 1993 (Cth)).

A significant issue in this area is that Indigenous people must consistently be translating and transplanting their concepts into a framework of rights that are not necessarily appropriate or address the expressed needs. To the extent that Indigenous knowledge can be protected through laws of intellectual property, this is only possible through concepts including property, ownership, works, monopoly privilege, exclusive rights, originality, and individual authorship. Proposals for *sui generis* legislation for the protection of Indigenous knowledge and Indigenous rights are being slowly crafted. Countries like Peru and Panama have been at the forefront of developing national *sui generis* legislation. As Tobin and Swiderska comment:

> The role of a *sui-generis* regime could therefore be to establish a bridge between Indigenous/local community and national and international legal systems, in order to secure the effective recognition and protection of rights which derive from customary law and practice.
>
> Tobin and Swiderska, 2001: p. 47

There remains a range of questions about how *sui generis* legislation should be developed. For instance, is it appropriate for international agencies to assist in the development of *sui generis* legislation or should each nation state take responsibility for the development of a *sui generis* approach that is appropriate to the circumstances of each context? Other issues such as the diversity of subject matter, the difficulty of identifying owners/custodians and applicability, enforceability, and transferability across diverse cultural contexts raise challenges and responses to these will affect what approach is taken.

Traditional Knowledge Licenses and Labels

The Traditional Knowledge Licenses and Labels initiative takes its initial cue from the success of Creative Commons. (This initiative has been developed by Jane Anderson, Kim Christen, and Michael Ashley with initial support from the Canadian Social Science and Humanities Research Council *Intellectual Property in Cultural Heritage: Law, Policy, Ethics* Project, and the WIPO). Creative Commons harnesses the capacity of copyright owners to license their works. In providing more detail to the expectations and desires of copyright owners regarding the use of their works, users of copyright material are provided with more information that helps them make better decisions about the use of such works. This is also the point of departure for the Traditional Knowledge Licenses and Labels. In short, they are designed to address two key problems being experienced by Indigenous peoples in relation to the management of their cultural resources. Firstly, how to protect individual, community, and familial rights to knowledge according to community based rules and regulations while simultaneously enabling the sharing of that knowledge with others beyond the community. Secondly, how to create a new informational and educational paradigm for documented cultural materials that Indigenous peoples can no longer control – for example the extensive material produced through the colonial encounter that resides in the public domain.

Addressing these two issues constitutes the foundational aims and ambitions of the Traditional Knowledge Licenses and Labels initiative. They are not being designed to provide blanket licenses for all Indigenous or traditional knowledge – but rather licenses for very specific contexts and situations, especially where current license options like Creative Commons fall short. The Labels function as an educative strategy to help users make more informed decisions about material that might generically be deemed 'public domain' but that still has Indigenous rules and obligations about access and use. The initiative is developed as a twofold strategy because it is not just legal protection that is required, but mechanisms that facilitate and support the communities from which this knowledge derives and is transformed. The hope is that this initiative will contribute to a shift in attitudes, understanding, and behavior about what Indigenous peoples consider to be fair and equitable use of their knowledge resources.

Conclusion

> Although Indigenous peoples are now recognized as key actors in this global discourse, it will need to be expanded to encompass a wider range of principles and priorities, which will eventually encompass political commitment to indigenous people's rights of self-determination. Only when Indigenous peoples are full partners in this dialogue, with full juridical standing and only when their cultural world views, customary laws, and ecological practices are recognized as fundamental contributions to resolving local social justice concerns will we be engaged in anything we can genuinely call dialogue.
>
> Coombe, 2001: pp. 284–285

One of the substantive difficulties with this particular contemporary problem for law is that the issues are constituted through and by multiple discourses and historical moments – not all of which are located or properly accommodated (or recognized) within intellectual property law. This raises questions then about what intellectual property law is being asked to do and can reasonably achieve. It is not equipped, or even sufficiently self-reflective, to function as an agent of reconciliation. Moreover, intellectual property law poses specific kinds of problems for Indigenous people. Not only does it conveniently ignore its own role in the historical exclusion of Indigenous interests, but upon inclusion it demands that Indigenous people, as well as their cultural expressions, conform to normative identifications of liberal individualism, authorship, labor, property, creativity, identifications of what a work is, periods of protection, and the like. The incommensurability that this has produced exposes the differing cultural values regarding ownership, access, and control of knowledge that inform the law. Through the inclusion of this new 'subject matter' intellectual property law's own historical methods for justifying a 'right' in something that is intangible have become more visible. With Indigenous issues, the position and politics of 'culture' have become much more apparent, but it would be a mistake

to read the difficulties for law as only ones about the capacity of law to recognize culture and accommodate cultural difference (Anderson, 2009).

Questions around Indigenous knowledge protection are unlike any other issues that intellectual property law has had to consider. This is because Indigenous concerns stretch across every part of the intellectual property axiom. Issues can include legal questions involving copyright, patents, trademarks, designs, and/or confidential information. Moreover, how the problems manifest and are experienced are not always straightforward and often overlap into other areas of law, policy, and ethics. Therefore there is not yet an international consensus about how Indigenous rights to the protection of cultural knowledge systems can be secured, either within an intellectual property regime or through some other overarching legislative or policy framework.

Indigenous peoples are looking to this body of law for several interrelated but sometimes opposing reasons. Attention to intellectual property law is about both the way intellectual property law has been a tool to facilitate the exploitation and expropriation of Indigenous knowledge, wherein nonindigenous peoples become the owners, authors, and holders of property rights in Indigenous knowledge and how it can be used to protect and function as a tool of leverage to control Indigenous knowledge systems. To deal with both sides of intellectual property law – its historical use against Indigenous peoples and its contemporary use for Indigenous peoples – it requires understanding the history and inherent limitations as well as its potential in certain select circumstances.

See also: Cultural Heritage; Cultural Resource Management: Conservation of Cultural Heritage; Ethnohistory; Human Rights and Social Work; Indigeneity; Intercultural Legal Discourse: Philosophical Aspects; Law and Economics; Law: Anthropological Aspects; Recognition, Politics of: Perspectives from Social Work and Sociology; Rights: Legal Aspects.

Bibliography

Agrawal, A., 2002. Indigenous knowledge and the politics of classification. International Social Science Journal 54 (173), 208–211.

Anaya, J., 1996. Indigenous Peoples in International Law. Oxford University Press, New York.

Anaya, J., August 30, 2012. The Situation of Indigenous Peoples in the United States of America. United Nations Human Rights Council.

Anderson, J., 2009. Law, Knowledge, Culture: The Production of Indigenous Knowledge in Intellectual Property Law. Edward Elgar Publishers, Cheltenham.

Anderson, J., 2010. Intellectual Property and Indigenous Knowledge: Issues Paper. Centre for the Public Domain. Duke University, Durham.

Anderson, J., 2012. Anxieties of authorship in colonial archives. In: Chris, C., Gerstner, D. (Eds.), Media Authorship. Routledge Press, New York.

Benton, L., 2001. Law and Colonial Cultures: Legal Regimes in World History 1400–1900. Cambridge University Press, Cambridge.

Coombe, R., 2001. The recognition of indigenous peoples' and community traditional knowledge in international law. St Thomas Law Review 14, 275–285.

Deazley, R., 2004. On the Origin of the Right to Copy: Charting the Movement of Copyright Law in Eighteenth Century Britain 1695–1775. Hart Publishing, Oregon.

Drahos, P., 1996. A Philosophy of Intellectual Property. Dartmouth Press, Sydney.

Eisenstein, E.L., 1980. The Printing Press as an Agent of Change. Cambridge University Press, Cambridge.

Greene, S., 2004. Indigenous peoples incorporated: culture as politics, culture as property in pharmaceutical bioprospecting. Current Anthropology 45 (2), 211–237.

Hayden, C., 2003. From market to market: bioprospecting's idioms of inclusion. American Ethnologist 30 (3), 359–371.

Keal, P., 2005. European Conquest and the Rights of Indigenous Peoples: The Moral Backwardness of International Society. Cambridge University Press, Cambridge.

Merry, S.E., 1991. Law and colonialism. Law and Society Review 25 (4), 889–922.

Mignolo, W., 1995. The Darker Side of Renaissance: Literacy, Territoriality and Colonization. University of Michigan Press, Ann Arbor.

Mignolo, W., 2000. Local Histories/Global Designs: Coloniality, Subaltern Knowledges and Border Thinking. Princeton University Press, Princeton.

Mignolo, W., 2011. The Darker Side of Western Modernity: Global Futures, Decolonial Options. Duke University Press, Durham and London.

Moraña, M., Dussel, E., Jáuregui, C., 2008. Coloniality at Large: Latin America and the Postcolonial Debate. Duke University Press, Durham and London.

Myers, F., 2000. The Making of an Aboriginal High Art. Duke University Press, Durham and London.

Oguamanam, C., 2009. Local knowledge as trapped knowledge: intellectual property, culture, power and politics. The Journal of World Intellectual Property 11, 29–57.

Popova-Gosart, U., 2012. Indigenous peoples: attempts to define. In: Berthier-Foglar, S., Collingwood-Whittick, S., Tolazzi, S. (Eds.), Biomapping Indigenous Peoples: Towards an Understanding of the Issues. Rodopi, Amsterdam.

Quijano, A., 1999. Coloniality and modernity/rationality. In: Therborn, G. (Ed.), Globalizations and Modernities. Forksningsradnamnden, Stockholm.

Raven, M., 2006. Protocols and ABS: recognising indigenous rights to knowledge in Australian Bureaucratic Organizations. Indigenous Law Bulletin 39. http://austlii.law.uts.edu.au/au/journals/ILB/2006 (last accessed 15.11.09.).

Rose, M., 1993. Authors and Owners: The Invention of Copyright. Harvard University Press, Cambridge.

Said, E., 1979. Orientalism. Vintage Books, New York.

Sherman, B., Bently, L., 1999. The Making of Modern Intellectual Property: The British Experience 1760–1911. Cambridge University Press, Cambridge.

Smith, L.T., 1999. Decolonizing Methodologies: Research and Indigenous Peoples. Zed Books, London, New York.

Tobin, B., Swiderska, K., 2001. Speaking in Tongues: Indigenous Participation in the Development of Sui-generis Regimes to Protect Traditional Knowledge in Peru. International Institute for Environment and Development, London.

Wynburg, R., Schroeder, D., Chennells, R., 2009. Indigenous Peoples, Consent and Benefit Sharing: Lessons from the San-hoodia Case. Springer Press, New York.

Younging, G., 2010. Gnaritas nullius (no one's knowledge): the public domain and the colonization of indigenous knowledge. Paper presented at the Seventeenth Session of the World Intellectual Property Organizations Intergovernmental Committee on Intellectual Property and Genetic Resources, Traditional Knowledge and Folklore. WIPO/GRTKF/IC/17/INF/5 (a).

Indigenous Knowledge: Science and Technology Studies

Kavita S Philip, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 7, pp. 7292–7297, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Philosophers of science, activists, and practitioners of indigenous knowledge have debated the status of indigenous knowledge with respect to science. 'Indigenous knowledge' is often used synonymously with 'local knowledge' or 'ethnoscience,' indicating knowledge specific to a culture. Science and Technology Studies scholarship, however, suggests that all knowledge, including scientific knowledge, is specific to its context. The study of indigenous knowledge has had practical results, shaping the politics of colonialism, nationalism, postcolonialism, and globalization. Issues at the intersection of science studies and indigenous knowledge, such as the imbrication of nature and culture, the construction of scientific certainty, and the historical contingency of systems of knowledge, have cultural and political implications for peoples negotiating their interrelationships in a globalized world.

Introduction: Convergence in Definitions

Science and Technology Studies (STS) and the study of indigenous knowledge are both interdisciplinary areas of investigation rather than singular disciplines. Their social and institutional origins were distinct, and early development in these areas occurred largely in isolation from each other. STS has disciplinary origins in history, sociology, policy, and philosophy, while the study of indigenous knowledge originated in anthropology and grew to overlap with agrarian studies and environmental history. Recent scholarship in both areas, however, suggests a convergence in their definitions of knowledge and their methodological assumptions.

'Indigenous knowledge' has long been associated with the terms 'local knowledge' or 'ethnoscience,' indicating knowledge systems that are specific to cultures or groups in particular historical or social contexts (see Richards et al., 1989). Adding the prefix 'indigenous' or 'ethno' to the terms 'knowledge' and 'science' often served to mark the embeddedness and context dependence of indigenous truths, in implicit contrast to the context-free, singular truth of science. The distinct meanings and uses of the terms 'indigenous knowledge' and 'science' both depended, thus, on a dichotomy that lay at the heart of modern definitions of knowledge, separating universal, value-free, static truth from situated, value-laden, changing cultural beliefs. Through the 1980s and 1990s, STS scholars questioned this dichotomy. They argued that all knowledge, including scientific knowledge, is specific to its particular cultural context. If this were correct, all knowledge might be considered 'indigenous.' The methods of study of the two fields should converge, then, since all knowledge should be studied with attention to its cultural and historical contexts. The distinct and separate foci of the two fields (science being conventionally studied in First World locations, and indigenous knowledge in the Third World) would then lose the logic of their original delineation. By the second decade of the twenty-first century, this blurring was evident, with anthropologists of science bringing 'indigenous' fieldwork experiences to First World laboratories, and sociologists of Euro-American techno-science extending their models to postcolonial contexts.

The status of the separation between 'indigenous' and 'universal' knowledge, the conditions of its construction, and the consequences of its erasure have long been issues of critical debate in STS. In a polemical and influential essay, *We Have Never Been Modern* (1993), Bruno Latour argued that if 'modernity' signifies the policing of boundaries between nature and culture, belief and reason, humans and nonhumans, then anthropological study reveals that we have never in fact been modern, for our practices have always involved hybrids of these categories. Latour and others, revisiting western knowledge from a distanced perspective, adopted and popularized ethnomethodological research. STS practitioners pictured the communities they studied as 'tribes' with characteristic beliefs and rituals. Western knowledge is, in this method of analysis, 'estranged' from its context or 'othered' (Douglas and Wildavsky, 1982). Conversely, nonwestern knowledges are presumed to be no less 'rational' or no more 'context dependent' than their western counterparts (Watson-Verran and Turnbull, 1995; Turnbull, 2000).

Studying the cultural and political role of scientific knowledge in different societies, scholars note that nature and culture are produced by networks of contingent histories; we encounter them not in their 'pure' states, but as hybrids –'socionatures' (Callon, 1995, p. 58) or 'natures–cultures' (Latour, 1993, p. 104). Anthropological observers of indigenous knowledge have always been interested in the ways in which boundaries were (or failed to be) drawn between nature and culture, and in the political weight of certain constructions of nature (see, for example, Lévi-Strauss, 1969). Late twentieth-century shifts in disciplines such as peasant studies, cultural geography and anthropology, and development studies also explored how ideas of nature shape and are historically shaped by social, political, and economic relationships (Peet and Watts, 1996; Bryant and Bailey, 1997). It became common to see interlinked analyses of the changing representations of nature, the production of ideologies of progress and modernity, and the functional relationship between the domination of indigenous groups and the production of scientific knowledge about them. The study of indigenous knowledge gradually shifted away from the province of primitivist ethnographers who tended to value it in opposition to science, assuming it to embody

primitive expressions of folklore that defy rationality and mythologize social relations.

Since the 1980s, anthropologists and interdisciplinary social scientists have viewed indigenous knowledge as coherent systems of belief and explanation that are homologous with scientific thought and that can often offer better accounts of local phenomena than those theorized by universalist science. This sort of investigation has prompted the reassessment of operative assumptions in medicine, pharmacology, resource management, and science education, as well as in nonscientific fields such as international law, policy, and development studies. Anthropologists, in the wake of postcolonial critiques of the imperial construction of ethnographic power, became increasingly self-reflexive and critical of their own representations of indigenous knowledge. Anthropological studies exposing ideological constructions of indigenous knowledge have become inseparable from efforts (in theoretical, pedagogical, and political realms) to alter power differentials in favor of marginalized groups. In the STS landscape at the turn of the twentieth century, histories of canonical European sciences (such as physics) began to share a common disciplinary space with the study of nonwestern knowledge.

Meanwhile, in the political sphere, the study of indigenous knowledge generated a host of policy implications, ranging from support for collective cultural rights and intellectual property rights for indigenous communities, to the reexamination of political assumptions about identity, modernity, and global citizenship. Nonwestern histories of the nineteenth and twentieth centuries, analyzed through the lenses of colonialism, nationalism, postcolonialism, and globalization, influenced both activism and scholarship on indigenous knowledge (Skaria, 1999; Sundar, 1997; Li, 1999). The study of indigenous knowledge facilitated, and in turn was enhanced by, the gaining of an influential international voice by indigenous rights groups and environmental nongovernmental organizations. These dynamics also helped shape emerging issues such as the politics of multiculturalism, the ownership of knowledge in the information economy, and the role of nationalism and group identity in a globalized world (Barker, 2011; Baviskar, 1995; Beteille, 1998). These are large, interdisciplinary topics in themselves, but even as they grow wider in scope and diffuse in their disciplinary underpinnings, the issues raised by the study of indigenous knowledge continue to intersect with questions about the authority of science and its representations of nature, about the politics of science and technology as interpreted by international institutions and policymakers, and about the role of technology and development in the forging of a more equitable world. In proportion to the seriousness with which it has grappled with these disparate discourses, STS research has yielded both theoretical insights and practical results.

Theoretical Results: Indigenous Knowledge and Epistemology

Epistemological concerns are central to the study of indigenous knowledge. Scholars of indigenous knowledge argue that if we take nonwestern knowledges seriously, we are quickly led to abandon a naïve realism with respect to scientific truth. That is, if we investigate indigenous knowledges for their internal coherence, social role, and explanatory power, rather than as instances of nonrational narratives, then we accumulate strong support for the argument that explanation, evidence, and truth are socially constructed and context dependent. This suggests, one argument continues, that what we see in the world, within and among cultures, are competing rationalities, rather than a singular mode of rationality and reason (commonly held to be the province of the west, and its sciences), versus nonrational or mystical modes of thinking that are antithetical to science (commonly held to be found outside the west). Such an argument requires a complex and situated definition of rationality, rather than a simple one assessing syllogistic consistency or the efficient choice of technical means for particular instrumentalist ends (Wilson, 1970; Sen, 2000).

Disagreement persists about how best to characterize indigenous rationalities with respect to western science, as illustrated in a well-known debate between anthropologists Marshall Sahlins (1995) and Gananath Obeyesekere (1992). Sahlins, holding that different cultures possess different rationalities, suggested that anthropologists ought to cultivate a "regard for cultural difference" (p. 14). He called on anthropologists to respect the worldwide resurgence of particular cultures as "distinct forms of life" (p. 13) and to value their opposition to a dominant, utilitarian, western modernity. In his own work, Sahlins suggested that Hawai'ian thought, for example, does not distinguish the practical from the mythical, nor the observable from the fictional, although these oppositions characterize European epistemologies (p. 6). Gananath Obeyesekere challenged this view, maintaining that all cultures possess a 'practical rationality' – that is, a flexibility and creativity expressed through manipulating and interpreting the world (p. 19). Obeyesekere's interpretation of Hawai'ian history via the attribution of a 'practical rationality' to the Hawai'ians was, in Sahlin's view, an example of 'symbolic violence' to the culture, and amounted to a neocolonialist model that reduced all cultural practices to western logic and rationality (p. 14). Obeyesekere contended that the cross-cultural interpretation of practical rationality, although complex and difficult, is "necessary if one is to talk of the other culture in human terms" (p. 21).

More recently, Eduardo Viveiros de Castro has suggested that the terms of the debate be shifted away from the realism/relativism argument. Arguing, through Amerindian cosmologies, for 'multinaturalism' and 'perspectivism' to replace multiculturalism, he insists on retaining difference and incommensurability, but on refusing the "relativism [which] is often implied in the ethnographic characterization of Amerindian cosmologies" (Castro, 2004b, p. 471). Castro argues that "Amazonian ontologies postulate difference rather than identity as the principle of relationality" (Castro, 2004a, p. 18). The task of translation, he suggests, involves an equivocation that is at the heart of relationality – a task whose inaccuracies and contingency we should embrace, rather than seek to erase. "Anthropology, then, is about misunderstandings," suggests Castro (2004a, p. 11). His advocacy of Amerindian perspectivalism takes him closer to a Deluezian than a Latourian position:

> Because there are no points of view onto things, things and beings are the points of view themselves (as Deleuze would say, 1988: 203). The question for Indians, therefore, is not one of knowing

> "how monkeys see the world" (Cheney and Seyfarth, 1990), but what world is expressed through monkeys, of what world they are the point of view. I believe this is a lesson from which our own anthropology can learn.
>
> Castro, 2004a, p. 11

Rather than view STS or anthropology as saving indigenous knowledge from obscurity, a range of new work from the perspective of indigenous practices has suggested that we could transform our disciplinary discourses by taking seriously the possibility of incommensurable ontologies and natures. Initially central to STS scholars' focus on epistemology, indigenous knowledge later became a window to new questions about ontology. It is key to an emerging emphasis in twenty-first-century STS on thinking through things and exploring pluralist, historically shaped ontologies (Henare et al., 2006).

Practical Results: Indigenous Knowledge and Politics

The categories of these theoretical debates recur in the realm of activism. For example, many activists, from Canada to India, have found it useful to employ a relativist argument that rejects the subordination of indigenous knowledge to science, while retaining the postulate of incommensurable difference. That is, they value indigenous knowledge over scientific reason, arguing that narratives of re-enchantment are needed to restore the integrity of indigenous cultures, lost or rendered inauthentic through violent encounters with European Enlightenment science. This position sees an incommensurability between tradition (seen as authentically indigenous) and modernity (seen as ethnocentrically European). Although phrased in the vocabulary of the anticolonial struggle, it is reminiscent of eighteenth- and nineteenth-century Euro-American romanticist and primitivist discourses in its desire to value a nature-centered spirituality over a technicist instrumentality. In both industrialized and developing countries, however, romantic and biocentric attempts to preserve the authenticity of indigenous knowledge have been critiqued from perspectives suspicious of epistemological relativism and biocentric romanticism.

Take, for example, the diversity of arguments from India. STS scholar Meera Nanda, supporting practical, rational 'people's science' movements in India, argues that the nostalgia for authentic indigeneity has oppressive overtones of nativism and cultural nationalism (Nanda, 1997). Anthropologist Akhil Gupta acknowledges the legacy of romanticism, but also links its persistence to the efficacy of political tactics: "the effectiveness of 'indigenous' identity depends on its *recognition* by hegemonic discourses of imperialist nostalgia, where poor and marginal people are romanticized at the same time that their way of life is destroyed" (Gupta, 1998, p. 18). Sociologist Amita Baviskar points out that these political tactics could shift historically. Observing changes in central Indian tribal and indigenous politics from the 1990s to the 2000s, she observes that, as the idea of cultural rights to specific lands was legislatively marginalized, the new politics of Hindu nationalism offered an alternative: "Adivasis could now imagine themselves as part of a larger cultural landscape – the Hindu nation, rather than as a subaltern presence incarcerated in a state of nature" (Baviskar, 2005, p. 5110).

Analyzing the politics of indigeneity, political scientist Ronald Niezen argues that, rather than signifying primordial identities, the category of indigenous people should be understood as a recent, historically shaped one (Niezen, 2000, p. 121). It may be understood through another growing subfield in STS: the histories of resource extraction, bioprospecting, and transnational economics (Li, 1999, 2000).

The notion of indigenous knowledge holds practical allure for biotechnologists, who believe that local knowledge of nature might hold the key to decades of future research into medicine and pharmaceuticals. It represents an ethical dilemma for the architects of globalization, heavily criticized for seeming to obliterate diversity in the interest of a homogeneous and efficient global marketplace. Advocates of indigenous knowledge often use the term synonymously with 'eco-friendly' or environmentally correct knowledge. But, as diverse studies show, the term 'indigenous people' has a somewhat diffuse reference and can range from small peasants, linked to domestic and global markets, to tribes, isolated from other groups.

In the last two decades of the twentieth century, there was a surge in support, from many different interest groups, for indigenous peoples' rights to their knowledge, resources, and cultural practices. Historical changes have often altered indigenous communities' relationship with their environment, or altered the environment itself, so that methods once sustainable might not be so in a changed context. Nor are all indigenous communities necessarily committed to preserving natural environments. For example, the Kissia of the Republic of Guinea, who use forest resources, tend to conserve forest growth, while other tribes in the same region, who are not dependent on trees for their survival, try to reduce forest growth wherever possible (Fairhead and Leach, 1997).

Practical struggles facing indigenous communities have often involved rights of access to water, fuel, and other elements of their natural surroundings that are crucial to their survival. Globally successful activism in the 1990s defended the rights of indigenous people to manage their natural surroundings independently from 'top-down' projects that seek to impose universalist models of productivity and resource management. In these arguments, indigenous knowledge was seen as having both cultural and pragmatic significance. The state and the global economy, representing hegemonic modes of resource extraction and cultural homogenization, were often seen as antagonistic to the interests of indigenous communities. Summing up the importance of law, politics, and the state in understanding the category of indigenous people, Ronald Niezen noted, "Although referring to some three hundred million original peoples worldwide who maintain attachments to 'timeless' original traditions, the political reality of 'being indigenous' is a product of the past several decades, originating in the terminology of international law and broadening to become a new form of group identity" (Niezen, 2000, p. 120).

Studying Indigenous Knowledge in Historical and Ecological Context

Antagonism between the state and indigenous communities dates back to the colonial era, when the expansion of European

territories was in large part an effort to control not only more markets and labor power, but also natural resources. Colonial environmental management often met with strong resistance from indigenous communities whose control over their environment was weakened or destroyed by colonial policies (Ranger, 1989; Peet and Watts, 1996). Soon after the mid-twentieth-century's wave of decolonization, in the 'development decades' of the 1960s, postcolonial nations struggled to catch up with the productivity of the industrialized west. Postindependence development involved, in many countries, shifts in economic policy to maximize resource extraction for exports, large conversions from food crop to cash crop cultivation, and increasing industrialization. The implementation of these policies (advocated by large international development agencies such as the World Bank) often caused growing pollution problems and the depletion of soil, water, and other natural resources. This most dramatically affected small farmers and indigenous groups, displacing them from their livelihood and threatening their modes of survival. Indigenous environmental movements, such as the antidam movement in India's Narmada valley, have often characterized the World Bank and the postcolonial nation-state as perpetuating colonial exploitation by subordinating local knowledge and experience to the rationalizing forces of science and economics (Rich, 1995; Scott, 1998).

The rapid growth, through the 1970s, of environmental ethics and economics did not always bring support for indigenous communities (see Harvey, 1996). Neo-Malthusian and neo-Hobbesian ecological scholarship identified overpopulation, poverty, and weakly governed communities as responsible for much environmental degradation and led to policies promoting the privatization of common property or advocating a global Leviathan. Influential theorists such as Garrett Hardin and William Ophuls predicted that the combination of resource scarcity with social complexities would lead to an ecological crisis (Hardin, 1968; Ophuls, 1977). Many saw these looming crises as legitimating more authoritarian state-led solutions, giving primacy to private property and modern resource management techniques. Since most indigenous communities are not affluent and are dependent on jointly governed common resources, these theories often pitted them against the prevailing expert wisdom.

Hardin's 'tragedy of the commons' thesis spurred detailed studies of indigenous environmental knowledge, increasingly critical of Hardin's assumptions. New scholarship on the commons, combining historical and political economic analysis of ecology, discovered empirical evidence that commons use did not inevitably lead to tragedy (Ostrom, 1990; Carney, 1996). Indigenous use of commons land was found to be more complex than simply opening access to all users, including irresponsible 'free riders.' Research in diverse communities found that successful commons use was invariably tied to complex, differentiated sets of communities with separate but overlapping rights and responsibilities, facilitating individual as well as group uses of lands and resources. In order to recognize and maintain these delicate social relations of resource use, both indigenous participants and outside administrators needed to be aware of the complexities of social stratification, and historical changes in social relationships among indigenous communities, and between them and the state. In many developing countries, however, colonial and postcolonial administrations altered the meanings of land-holding and resource use traditions through successions of new legal definitions and administrative hierarchies, over time squeezing indigenous systems into models that privileged private over communitarian interests. Histories of disparate indigenous groups reveal a continual flux of local, regional, and global knowledges and practices, rendering unviable nostalgia for a pure localism, or notions of pristine authenticity (McCay and Acheson, 1987). An emerging field of 'Third World Political Ecology' argued that support for rights of indigenous peoples should be inseparable from the development of more nuanced historical analyses of environmental and social relations.

Since the 1980s, the importance of indigenous knowledge has been widely acknowledged by influential international institutions. In 1982, recognizing that disparate indigenous movements all over the developing world were articulating concerns similar to each other, the United Nations established the Working Group on Indigenous Populations, annually bringing together participants from indigenous communities around the world. Indigenous groups, working in solidarity with each other and developing an international alliance facilitated by the annual working group meetings in Geneva, prepared a joint statement for the 1992 Earth Summit at Rio, asserting indigenous rights as a fundamental requisite for the future of sustainable development. Recognizing the growing importance of biotechnology and its links with traditional knowledge of medicinal plants, it stated: "[W]e demand guaranteed rights to our intellectual property, and control over the development and manipulation of this knowledge" (Article 44, Charter of the Indigenous-Tribal Peoples of the Tropical Forests, reprinted in International Alliance of Indigenous-Tribal Peoples of the Tropical Forest and the International Working Group for Indigenous Affairs, 1996, p. 9).

The 1993 UN Declaration on the Rights of Indigenous Peoples recognizes indigenous peoples' rights to the preservation of all forms of their knowledge, including cultural, spiritual, medical, scientific, and technological. It also recognizes the right of indigenous peoples to protect their intellectual property, and to the restitution of any such property that was taken from them in violation of their laws or customs. In 1998, the World Intellectual Property Organization established the Global Intellectual Property Issues Division, recognizing indigenous peoples as the 'most important' beneficiaries of intellectual property (Roulet, 1999, p. 129).

The successes of indigenous peoples' movements in the late twentieth century were accompanied by a popularization, at both expert and lay levels, of the terms and categories of 'indigeneity.' These successes were linked, paradoxically, to a tendency critiqued by political ecologists. Critics pointed out that increased funding for the protection of indigenous groups often went along with romanticized descriptions of such groups. In a study of South Asian environments and ethnicities, Sumit Guha suggested that the 'poetic vagueness' in the definitions of indigenous groups feeds into a tendency to regard such groups as static remnants of original racial groups, supporting policies that seek to preserve them as museumized 'endangered species' (Guha, 1999, pp. 4–5). Anthropologist Lisa Malkki has critiqued the 'sedentarist metaphysics' of indigeneity for essentializing,

rather than historicizing, the link between indigenous people, knowledge, and land (Malkki, 1992).

International recognition and support for indigenous knowledge grew dramatically through the new millennium. Nevertheless, as conflicts between indigenous rights and international trade grew rather than diminished, it remained to be seen how far courts would go to enforce 'soft law' elements such as international conventions, to impose regulations on transnational corporations and states, protecting indigenous communities. The future of indigenous knowledge in this age of globalization may well be worked out in policy debates and resource conflicts rather than in philosophical texts. There will always, however, be an important link between academic and political life in this area. Interpretations of the various assumptions in environmental law and the indigenous knowledge will depend on real-world manifestations of the construction of scientific uncertainty, the historical contingency of systems of knowledge, the historical ontologies, and social epistemologies of nature and culture. In this way, the theory and practice of STS and indigenous knowledge may become more closely interlinked and may have significant consequences for the shape of future transnational indigenous worlds.

See also: Cultural Studies of Science; Environmental Sciences; Genetics and Indigenous Communities: Ethical Issues; Historical Thought and Historiography: Indigenous Cultures in the Americas; Indigenous Knowledge and Intellectual Property Rights; Postcolonial Science Studies; Science and Technology Studies, Ethnomethodology of; Situated Knowledge, Feminist and Science and Technology Studies Perspectives.

Bibliography

Barker, Joanne, 2011. Native Acts: Law, Recognition, and Cultural Authenticity. Duke University Press, Durham.

Baviskar, Amita, 1995. In the Belly of the River: Tribal Conflicts over Development in the Narmada Valley. Oxford University Press, Delhi.

Baviskar, Amita, 2005. Adivasi encounters with Hindu nationalism in MP. Economic and Political Weekly 40 (48), 5105–5113.

Beteille, Andre, 1998. The idea of indigenous people. Current Anthropology 39 (2), 187–191.

Bryant, Raymond, Bailey, Sinead, 1997. Third World Political Ecology. Routledge, London.

Callon, Michel, 1995. Four models of the dynamics of science. In: Jasanoff, S., Markle, G., Petersen, J., Pinch, T. (Eds.), Handbook of Science and Technology Studies. Sage, London, pp. 29–63.

Carney, Judith A., 1996. Converting the wetlands, engendering the environment: the intersection of gender with agrarian change in Gambia. In: Peet, R., Watts, M. (Eds.), Liberation Ecologies: Environment, Development, Social Movements. Routledge, London, pp. 289–308.

Castro, Eduardo Viveiros de, 2004a. Perspectival anthropology and the method of controlled equivocation. Tipití 2 (1), 3–22.

Castro, Eduardo Batalha Viveiros de, 2004b. Exchanging perspectives: the transformation of objects into subjects in Amerindian ontologies. Common Knowledge 10 (3), 463–484.

Cheney, Dorothy L., Seyfarth, Robert M., 1990. How Monkeys See the World: Inside the Mind of Another Species. University of Chicago Press, Chicago.

Deleuze, Gilles, 1988. Le pli. Leibniz et le baroque. Minuit, Paris.

Douglas, Mary, Wildavsky, Aaron, 1982. Risk and Culture. University of California Press, Berkeley.

Fairhead, James, Leach, Melissa, 1997. Culturing trees: socialized knowledge in the political ecology of Kissia and Kuranko forest islands of Guinea. In: Seeland, K. (Ed.), Nature Is Culture: Indigenous Knowledge and Sociocultural Aspects of Trees and Forests in Non-European Cultures. Intermediate Technology Publications, Southampton, pp. 7–18.

Guha, Sumit, 1999. Environment and Ethnicity in India, 1200–1991. Cambridge University Press, Cambridge.

Gupta, Akhil, 1998. Postcolonial Developments: Agriculture in the Making of Modern India. Duke University Press, Durham.

Hardin, Garrett, 1968. The tragedy of the commons. Science 162, 1243–1248.

Harvey, David, 1996. Justice, Nature, and the Geography of Difference. Blackwell, Oxford.

Henare, Amiria, Holbraad, Martin, Wastell, Sari (Eds.), 2006. Thinking through Things: Theorising Artefacts Ethnographically. Routledge, London.

International Alliance of Indigenous-Tribal Peoples of the Tropical Forest and the International Working Group for Indigenous Affairs, 1996. Indigenous Peoples, Forests, and Biodiversity: Indigenous Peoples and the Global Environmental Agenda Denmark. IWGIA Document No. 82.

Latour, Bruno, 1993. We Have Never Been Modern (Catherine Porter, Trans.). Harvard University Press, Cambridge, MA.

Lévi-Strauss, Claude, 1969. The Raw and the Cooked (John, Doreen Weightman, Trans.). Harper & Row, New York.

Li, Tania M. (Ed.), 1999. Transforming the Indonesian Uplands: Marginality, Power and Production. Harwood Academic Press, Amsterdam.

Li, Tania M., 2000. Articulating indigenous identity in Indonesia: resource politics and the tribal slot. Comparative Studies in Society and History 42 (1), 149–179.

Malkki, Liisa, 1992. National Geographic: the rooting of peoples and the territorialisation of national identity among scholars and refugees. Cultural Anthropology 7 (1), 24–44.

McCay, Bonnie, Acheson, James (Eds.), 1987. The Question of the Commons: The Culture and Ecology of Communal Resources. University of Arizona Press, Tucson.

Mohanty, Satya, 1998. Literary Theory and the Claims of History: Postmodernism, Objectivity, Multicultural Politics. Oxford University Press, Delhi.

Nanda, Meera, 1997. Against social de(con)struction of science: cautionary tales from the Third World. In: Meiksins Wood, E., Bellamy Foster, J. (Eds.), In Defense of History: Marxism and the Postmodern Agenda. Monthly Review Press, NY.

Niezen, Ronald, 2000. Recognizing indigenism: Canadian unity and the international movement of indigenous peoples. Comparative Studies in Society and History 42, 119–148.

Obeyesekere, Gananath, 1992. The Apotheosis of Captain Cook: European Mythmaking in the Pacific. Princeton University Press, Princeton.

Ophuls, William, 1977. Ecology and the Politics of Scarcity: Prologue to a Political Theory of the Steady State. W. H. Freeman, San Francisco.

Ostrom, Elinor, 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, Cambridge.

Peet, Richard, Watts, Michael (Eds.), 1996. Liberation Ecologies: Environment, Development, Social Movements. Routledge, London.

Ranger, Terence, 1989. Whose heritage? The case of the Matobo National Park. Journal of Southern African Studies 15 (2), 217–249.

Rich, Bruce, 1995. Mortgaging the Earth: The World Bank, Environmental Impoverishment, and the Crisis of Development. Beacon Press, Boston.

Richards, Paul, Slikkerveer, Jan L., Phillips, Adedotun O., 1989. Indigenous Knowledge Systems for Agriculture and Rural Development: The CIKARD Inugural Lectures. Studies in Technology and Social Change, No. 13. Iowa State University, Ames.

Roulet, Florencia, 1999. Human Rights and Indigenous Peoples: A Handbook on the UN System. IWGIA Document No. 92, Copenhagen.

Sahlins, Marshall, 1995. How "Natives" Think: About Captain Cook, for Example. University of Chicago Press, Chicago.

Scott, James C., 1998. Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed. Yale University Press, New Haven.

Sen, Amartya, 20 July, 2000. East and West: The Reach of Reason. New York Review of Books.

Skaria, Ajay, 1999. Hybrid Histories: Forests, Frontiers and Wildness in Western India. Oxford University Press, Delhi.

Sundar, Nandini, 1997. Subalterns and Sovereigns: An Anthropological History of Bastar, 1854–1996. Oxford University Press, Delhi.

Turnbull, David, 2000. Masons, Tricksters and Cartographers: Comparative Studies in the Sociology of Scientific and Indigenous Knowledge. Harwood Academic Publishers, Amsterdam.

Watson-Verran, Helen, Turnbull, David, 1995. Science and other indigenous knowledge systems. In: Jasanoff, S., Markle, G., Petersen, J., Pinch, T. (Eds.), Handbook of Science and Technology Studies. Sage, London, pp. 115–139.

Wilson, Bryan (Ed.), 1970. Rationality. Harper and Row, Evanston, IL.

Indigenous Management Styles

Charmine EJ Härtel, UQ Business School, The University of Queensland, Brisbane, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Examples of indigenous perspectives infusing management practices and research are scarce in the mainstream management literature. This article summarizes different approaches to defining indigenous, provides examples of indigenous management perspectives, and describes the concept of indigenous capitalism. Collectively the article demonstrates the diversity of philosophies underpinning management practice and highlights opportunities for advancing management theory and practice by including indigenous research approaches, knowledge, and practices.

Given the vast international literature on management and global market practices, one could be forgiven for thinking that our understanding of non-Western concepts of management was well formed. Quite the opposite is true, however, with Western perspectives embedded in colonial traditions prevailing (Holtbrügge, 2013; Rodrigues et al., 2012; Suddaby et al., 2011; Verbos and Humphries, 2012). Although the need and the legitimacy of indigenous management theory and practice is contested by some, there are a growing number of voices calling for research on indigenous management practices and the use of indigenous knowledge and research approaches to build new management theory (Leung, 2012; Pio et al., 2013; Suddaby et al., 2011). Advocates' rationale for embracing indigenous perspectives on management is illustrated by this quote from Marsden in 1991 (p. 32, cited in Holtbrügge, 2013, p. 2): "strategies based on European scientific techniques ignore the rich resources, both practical and intellectual, which exist in non-Western societies (…). 'Local', 'traditional' or 'folk' knowledge is no longer the irrelevant vestige of 'backward' people who have not yet made the transition to modernity, but the vital well-spring and resource bank from which alternative futures might be built."

Notions of Indigenous

The term indigenous is used in several different ways in the academic literature (Jackson, 2013). As noted by Van de Ven and Runtian (2012), the meaning of indigenous is culturally sensitive and has many possible definitions, but all referencing in some way belonging to, originating from, or identifying with a specific people or place. Scholars such as Van de Ven and Runtian (2012) take a broad view of the term, arguing that all research and practice is indigenous in that they emerge from a specific culture and context. Others reserve the term indigenous knowledge, indigenous research, and indigenous management for that which relates to first nations peoples (e.g., Bunten, 2011). Still others use the term indigenous to refer to the cultural, historical, and ideological traditions of non-Western contexts. The United Nations Permanent Forum on Indigenous Issues defines indigenous as "peoples who are the inheritors and practitioners of unique cultures and ways of relating to other people and to the environment" (UNPFII, 2007). Using this definition, they identify "more than 370 million indigenous peoples in some 90 countries worldwide" (UNPFII, 2014).

Examples of Indigenous Management Perspectives

Regardless of which definition of indigenous we adopt, it is useful to acknowledge the synergy between Eastern and indigenous traditions of wisdom, interconnectedness, and holism (Holtbrügge, 2013; Pio et al., 2013). These notions in particular have been drawn on by scholars to argue that management theory and practice would be strengthened by connecting the traditional notions of wisdom associated with indigenous and Eastern traditions with the economic focus of Western management perspectives (Pio et al., 2013). Examples of indigenous management perspectives are informative in differentiating indigenous ideology from Western ideology as well as identifying possible bridges between the two worlds.

Like all indigenous peoples, Australian Aboriginal peoples hold a relational view of the world whereby they recognize themselves and their place as part of a greater cosmology in which all creation is purposeful and interconnected as kin (Miley and Read, 2013; Royal, 2002). Relational knowledge accepts there are different points of view and allows differences to coexist in an atmosphere of trust, collaboration, and respect (Verbos and Humphries, 2012). A focus of indigenous management practices therefore is on inclusion and valuing all employees' contributions as equally worthwhile (Miley and Read, 2013). The relational perspective is also evident in traditional problem-solving processes, which involve the whole kinship structure, including children, in a public discussion (Pearson and Chatterjee, 2010). Face-to-face decision-making that openly considers and balances the views of work, family, and community is a common hallmark of indigenous management practices in Australia (Pearson and Chatterjee, 2010).

Australian Aboriginal peoples' traditional thinking style has been described as patterned in contrast to Western triangulated thinking which is based in an ideology of hierarchy and power, where success is defined according to monetary wealth, property, rulership, and ownership (Miley and Read, 2013; Pearson and Chatterjee, 2010). Pattern thinking refers to viewing

everything as interconnected and belonging within the larger pattern of the universe (Miley and Read, 2013). Accordingly, Aboriginal law promotes harmony, bonding all things together in equal relationships. Likewise, their belief systems are centered on universality, kinship, and 'country' (Miley and Read, 2013). The notion of 'caring for country' is a manifestation of pattern thinking and leads to indigenous management practices associated with inclusion of environmentally responsible outcomes in key performance indicators (Miley and Read, 2013).

The violent colonization of Australia displaced its traditional Aboriginal peoples physically, spiritually, culturally, and psychologically (Appo and Härtel, 2003, 2005). The associated cultural upheaval is manifest in identity confusion, loss of traditional languages, and disintegration of cultural traditions (Appo and Härtel, 2003, 2005). To illustrate, of the 250 languages spoken preinvasion, just 60 remain as a living language (Korff, 2014). Moreover, in 2008, just 62% of Aboriginal adults reported identifying with a clan, language, or tribal group (Korff, 2014). The result is a complex and heterogeneous picture of contemporary Aboriginal peoples, where, in many instances, only fragments of traditional connections to people, place, and language remain, and personal experiences and ways of life vary. The implication of this is that not all Aborigines identify or follow an indigenous perspective.

In general terms, an Aboriginal worldview shares the focus on relatedness of other indigenous perspectives. All aspects of reality, past, present, and future, are viewed as interlinked with one another and to a spiritual world (Miley and Read, 2013). Thus, all of creation is shared and all is for the benefit of others in the relationship. One cannot simply take, therefore, one must ensure that what one takes is transformed into benefit for others and sustaining positive relationships. As summarized by Roy (2014, p. 118), "Aboriginal worldviews are characterized by their relational, communalistic, eco-centric and cosmo-centric focus."

A number of lessons can be drawn from examining the interface of European management practices with Aboriginal worldviews. First, indigenous management practices not only reflect a particular indigenous worldview; in today's world they almost always must operate within the parameters established by a larger economic system founded on strikingly different cultural assumptions. Thus, to engage in indigenous management practices also requires cross-cultural competencies on the part of indigenous and nonindigenous practitioners (Farrelly and Lumby, 2009). Across a range of discipline areas, there is consensus that these competencies include enacting cultural sensitivity, cultural awareness involving an understanding of Aboriginal history and philosophy as well as one's own culture and views of Aboriginal people (Sue, 2006) and the concept of cultural safety. The latter is especially critical in that it creates "an environment that is spiritually, socially and emotionally safe, as well as physically safe for people; where there is no assault challenge or denial of their identity, of who they are and what they need. It is about shared respect, shared meaning, shared knowledge and experience of learning together" (Williams, 1999). Too often, indigenous people's nonindigenous counterparts lack an accurate understanding of their culture and history. To illustrate, recent figures estimate that 91% of white Australians have little to no contact with Aboriginal people, with the majority citing the media as their sole source for developing impressions of Australian Aboriginal peoples (Moreton-Robinson, 2011, cited in Duthie et al., 2013). Genuine contact and experience with Aboriginal peoples is essential to develop accurate cultural knowledge.

A second lesson relating to interfacing indigenous and nonindigenous business practices is recognizing that "it is an economic imperative … for businesses to identify the ways in which traditional culture can be integrated into their organizational practices, such as in the community consultation processes …" (Härtel et al., 2013, p. 540). For example, to respect cultural norms of involving all community members, a business would adopt extended decision-making time frames. Also, to integrate the Aboriginal relational worldview, nonindigenous businesses would focus on building and sustaining long-term mutually beneficial partnerships where the indigenous stakeholder is not only the specific indigenous business partner, but also the larger Aboriginal community of which it is a part (Härtel et al., 2013). This represents a real shift for most nonindigenous organizations in that it means human resource management activities must not just be internally focused, but also must be externally focused on communities, and that a lifespan view of the employment process be taken (Härtel et al., 2013, p. 538). Thus, competencies in community relations and community development are essential features of nonindigenous organizations that work respectfully and effectively with indigenous organizations (Härtel et al., 2013). It means adopting "a view of social inclusion as incorporating community consultation in corporate activity from pre-entry to exit, good corporate-community relations, community self-determination, and promotion of community and regional sustainability" (Härtel et al., 2013, p. 538). Further, the considerable funds provided to support Aboriginal employment and businesses within the Australian context has led to new forms of exploitation that need to be guarded against, for example, tokenism. An example of this is the organizations that pose as Aboriginal organizations when in fact they do not represent Aboriginal economic and cultural interests. Again, working closely with Aboriginal communities and respecting their cultural authority and rights to self-governance addresses this issue by letting them identify which claimants are indeed Aboriginal organizations (Härtel et al., 2013).

Illustrative of pattern thinking, the Māori worldview, termed *Te Ao Māori*, is one of interconnectedness between the physical and the metaphysical. The relational perspective in Māori businesses in New Zealand encompasses values of respect, care, well-being, and high-quality relationships (Spiller et al., 2011). These traditions are referred to as *Kaupapa Māori*, or 'the Māori way' (Ruwhiu and Cone, 2013).

A core Māori value is *kaitiakitanga* or stewardship, where humans are born with the obligation and power to create *mauri ora* or well-being for all (Spiller et al., 2011a,b). *Mauri* refers to the spiritual life force that is seen to exist in all aspects of creation and humans have the purpose and agency to care for and conserve it (Spiller et al., 2011). This perspective places humans in a reciprocal relationship with others, the spiritual, the ancestral, and the environmental and thus one's own well-being is inextricably linked with the well-being of all the others in the relationship (Pio et al., 2014;

Spiller et al., 2011a,b). Accordingly, the Māori view business as having the goal of creating well-being (*mauri ora*), which includes "healing and sustaining life" (Spiller et al., 2011, p. 225). Thus Māori organizations are differentiated from organizations founded in the Western individualistic philosophy in that they see "profit and economic well-being as a means to serve broader social, cultural, environmental, and spiritual well-being goals" rather than as an end in themselves (Spiller et al., 2011, p. 225).

Springing from the focus on relationships, Māori management practices prioritize relationships, operating from the stance that "You do business because you've got a relationship. The relationship leads to business" (Ruwhiu and Cone, 2013, p. 32). The focus on relational well-being and wealth provides a unique competitive advantage for Māori organizations in the areas of "forging long-term supplier arrangements and joint venture partnerships with other global firms" (Spiller et al., 2011, p. 224). Moreover, the Māori holistic philosophy of interconnectedness and well-being in and through relationships manifests in corporate social responsibility as a deeply embedded and central value.

Māori organizations ensure culturally relevant management practices by focusing on relational processes. Management decisions are assessed according to Māori values, most notably respect (*manaaki*), empathy (*aroha*), fostering unity (*kotahitanga*), stewardship of resources (*kaitiakitanga*), seeking knowledge and understanding (*mātauranga*), just behavior (*tika*), relational/interconnected view of all creation (*whakapapa*), and uplifting others (*hāpai*) (Spiller et al., 2011).

In Chinese indigenous research, one of the central concepts is the Confucian notion of ethics (Pan et al., 2012). This perspective reveals a relational view of ethics where people are viewed as sacred which creates respectful relations and participation in rituals is seen as key to civility (Kim and Strudler, 2012). Other Chinese indigenous concepts receiving growing attention in the management literature are *guanxi* (good relationships) and *xin* (mutual trust), both of which derive from the Confucian tradition of reciprocity (Härtel et al., 2010; Pan et al., 2012). The type of relationship it entails is an instrumental one, involving the repeated and reciprocal exchange of favors with members of the significant groups to which one belongs (Härtel et al., 2010; Jiang et al., 2013). Guanxi emerges and is maintained through the continual demonstration of *xin* (trust) and *xinyong* (trustworthiness, credibility, reliability, and reputation) in one's network (Härtel et al., 2010). Managerial *guanxi* practices have been separated into *inward guanxi practices* and *outward guanxi practices* to reflect differences in managers' use of their guanxi ties to bring resources into their organization as a result of favor receiving or moving resources outward as a result of favor giving (Jiang et al., 2013).

The indigenous concept of *ubuntu* is frequently evident in management practices in South Africa. The Bantu word *ubuntu* conveys a relational perspective where people are bound together as is their well-being (Holtbrügge, 2013). Like the traditional decision-making practices of Aboriginal Australians, *ubuntu* emphasizes community participation and consensus in decision-making.

The Hindu concept of *vanaprastha ashrama* manifests in the service orientation, social responsibility, and focus on a purposeful life of managers in India (Cappelli et al., 2010). Stemming from the Vedic religion, the concept of *lokasangraha* or 'prosperity for all' underpins Vedic economic philosophy, which focuses on harmony, cooperation, and frugality. The Hindi concept of *jugaad* refers to the mentality of improvising when resources are lacking or obstacles are encountered (Cappelli et al., 2010).

Indigenous Capitalism

The fundamental features that differentiate indigenous corporations and capitalism from other forms are the embedding of indigenous values, law, culture, and community accountability into the corporate mission and business culture of the organization (Bunten, 2011). Indigenous capitalism therefore is defined as the accumulation and redistribution of wealth by indigenous corporations within indigenous communities, with ecological, social, spiritual, and cultural stewardship as the overarching objective (Bunten, 2011; Spiller et al., 2011). Indigenous capitalism prioritizes relationships with people and the environment, emphasizing stewardship and sustainability rather than economic growth (Spiller et al., 2011). The development of indigenous corporations has been described as a means for indigenous groups to participate in global and political economies while investing in their own communities, land, and governance (Bunten, 2011). The relational logics of indigenous perspectives share with liberal feminist perspectives an emphasis on relationships, relational processes, and ethics (Verbon and Humphries, 2012). Relational logics may provide the foundation required to transform institutions from promoting marginalization and discrimination to encouraging inclusiveness, integrity, and justice (Verbos and Humphries, 2012).

Conclusion

Indigenous management practices are embedded in traditional or local knowledge, and as such have long been marginalized in a Western dominant management literature (Nkomo, 2011, p. 378). They are necessarily context and culture bound, which is not to say they do not change or cannot generalize (Leung, 2012). Most importantly to the identification of indigenous management practices is participation of those deemed as indigenous so that descriptions and theories are tied to the people, their place, and the time, and overgeneralization is avoided (Appo and Härtel, 2003; Holtbrügge, 2013). Ultimately, the advancement of management practices and theories depends on decolonizing the management literature so that Western philosophies and methods are not privileged, portrayed as superior, or playing the role of enlightening the rest of the world.

See also: Attitude Formation and Change; Attitudes and Behavior; Collective Action; Collectivism and Individualism: Cultural and Psychological Concerns; Cooperation and Competition; Cross-Cultural Psychology; Cross-Cultural

Training in Industrial and Organizational Psychology; Cultural Influences on Interpersonal Relationships; Cultural Intelligence and Competencies; Cultural Psychology; Diversity in organizations; Epistemological Beliefs and Motivation; Ethnic Identity, Psychology of; First Language Acquisition, Developmental Psychology of; Identity and Identification, Social Psychology of; Indigenous Psychology; Moral Development, Cultural Differences in; Motivation in Australian Aboriginal Populations; Motivation: Life Course and Sociological Perspectives; Prosocial Behavior, Cultural Differences in; Self in Culture: Early Development; Spirituality; Subjective Wellbeing and Culture; Values Across Cultures, Development of; Values, Social Psychology of; Wisdom, Psychology of.

Bibliography

Appo, D., Härtel, C.E.J., 2005. On the pathogenicity of culture: questioning the rationality of 'development' programs for perpetually dysfunctional aboriginal groups. Cross-cultural Management: An International Journal 12 (1), 4–30.

Appo, D., Härtel, C.E.J., 2003. Questioning management paradigms that deal with aboriginal development programs in Australia. Asia Pacific Journal of Human Resources 41 (1), 36–50.

Becker-Ritterspach, F., Raaijman, T., 2013. Global transfer and Indian management. Management International Review 53 (1), 141–166.

Bunten, A.C., 2011. A call for attention to indigenous capitalisms. New Proposals: Journal of Marxism and Interdisciplinary Inquiry 5 (1), 60–71.

Cappelli, P., Singh, H., Singh, J., Useem, M., 2010. The India way: lessons for the U.S. Academy of Management Perspectives 24 (2), 6–24.

Duthie, D., King, J., May, J., 2013. Raising awareness of Australian aboriginal peoples reality: embedding aboriginal knowledge in social work education through the use of field experiences. The International Education Journal: Comparative Perspectives 12 (1), 197–212.

Farrelly, T., Lumby, B., 2009. A best practice approach to cultural competency training. Aboriginal and Islander Health Worker Journal 3 (5), 14–22.

Härtel, C.E.J., Appo, D., Hart, B., 2013. Inclusion at societal fault lines: aboriginal peoples of Australia. In: Ferdman, B.M., Deane, B.R. (Eds.), Diversity at Work: The Practice of Inclusion. Jossey-Bass Publishers, SIOP Professional Practice Series, pp. 520–546.

Härtel, C.E.J., Ma, R.M.M., As-Saber, S., 2010. When East meets West: managing Chinese enterprise relationships through Guanxi-based diversity management. In: Syed, J., Özbilgin, M.F. (Eds.), Managing Cultural Diversity in Asia: A Research Companion. Edward Elgar, Cheltenham, New York.

Holtbrügge, H., 2013. Indigenous management research. Management International Review 53 (1), 1–11.

Jiang, H., Cannella, A., Gao, L., Jiao, J., 2013. Inward–outward guanxi contagion and employees' responses to managerial guanxi practices. Journal of Management & Organization 19, 613–634.

Jackson, T., 2013. Reconstructing the indigenous in African management research, implications for international management studies in a globalized world. Management International Review 53 (1), 15.

Kim, T.W., Strudler, A., 2012. Workplace civility: a confucian approach. Business Ethics Quarterly 22 (3), 557–577.

Korff, J., June 28, 2014. Aboriginal Languages. http://www.creativespirits.info/aboriginalculture/language/ (accessed 07.07.14.).

Leung, K., 2012. Indigenous Chinese management research: like it or not, we need it. Management and Organization Review 8 (1), 1–5.

Li, P.P., 2012. Toward an integrative framework of indigenous research: the geocentric implications of Yin-Yang Balance. Asia Pacific Journal of Management 29 (4), 849–872.

Miley, F., Read, A., 2013. Indigenous spirituality at work: Australia. In: Neal, J. (Ed.), Handbook of Faith and Spirituality in the Workplace. Springer, New York, pp. 175–194.

Moreton-Robinson, A, 26 September 2011. Representations of Aboriginality in Australian media. Oodgeroo Noonuccal Public Lecture. Queensland University of Technology, Brisbane, Australia.

Nkomo, S.T., 2011. A postcolonial and anti-colonial reading of 'African' leadership and management in organization studies: tensions, contradictions and possibilities. Organization 18 (3), 365–386.

Pan, Y., Rowney, J.A., Peterson, M.F., 2012. The structure of Chinese cultural traditions: an empirical study of business employees in China. Management and Organization Review 8 (1), 77–95.

Pearson, C.A.L., Chatterjee, S.R., 2010. Extending business education beyond traditional boundaries: a case study in negotiated problem resolution is a remote regional indigenous community in Australia. Journal of Teaching in International Business 21, 307–328.

Pio, E., Tipuna, K., Rasheed, A., Parker, L., 2014. Te Wero – the challenge: reimagining universities from an indigenous world view. Higher Education 67, 675–690.

Pio, E., Waddock, S., Mangaliso, M., McIntosh, M., Spiller, C., Takeda, H., Syed, J., 2013. Pipeline to the future: seeking wisdom in indigenous, eastern, and western traditions. In: Neal, J. (Ed.), Handbook of Faith and Spirituality in the Workplace. Springer, New York, pp. 195–219.

Roy, A., April 2014. Aboriginal worldviews and epidemiological survey methodology: overcoming incongruence. International Journal of Multiple Research Approaches 8 (1), 117–128.

Ruwhiu, R., Cone, M., 2013. Pragmatic leadership: a return to wisdom. Journal of Management & Organization 19, 25–43.

Rodrigues, S.B., Duarte, R.G., Carrieri, A.P., 2012. Indigenous or imported knowledge in Brazilian management studies: a quest for legitimacy? Management and Organization Review 8 (1), 211–232.

Royal, T.A.C., February 21, 2002. Indigenous Worldviews: A Comparative Study. Report for Ngāti Kikopiri, Te Wānanga-o-raukawa, Te Puni Kōkiri. Winston Churchill Memorial Trust, Fulbright New Zealand.

Spiller, C., Erakovic, L., Henare, M., Pio, E., 2011a. Relational well-being and wealth: māori businesses and an ethic of care. Journal of Business Ethics 98 (1), 153–169.

Spiller, C., Pio, E., Erakovic, L., Henare, M., 2011b. Wise up: creating organizational wisdom through an ethic of Kaitiakitanga. Journal of Business Ethics 104, 223–235.

Suddaby, R., Hardy, C., Huy, Q.N., 2011. Introduction to special topic forum: where are the new theories of organization? Academy of Management Review 36 (2), 236–246.

Sue, D.W., 2006. Multicultural Social Work Practice. Wiley, Hoboke, NJ.

UNPFII, 2007. United Nations Permanent Forum on Indigenous Issues – Declaration on the Rights of Indigenous Peoples. http://undesadspd.org/IndigenousPeoples/AboutUsMembers/History.aspx (accessed 07.07.14.).

UNPFII, 2014. United Nations Permanent Forum on Indigenous Issues – History of Indigenous Peoples and the International System. http://undesadspd.org/IndigenousPeoples/AboutUsMembers/History.aspx (accessed 07.07.14.).

Van de Ven, A.H., Runtian, J., 2012. Indigenous management research in China from an engaged scholarship perspective. Management and Organization Review 8 (1), 123–137.

Verbos, A.K., Humphries, M.T., 2012. Decoupling equality, diversity, and inclusion from liberal projects: hailing indigenous contributions to institutional change. Equality, Diversity and Inclusion: An International Journal 31 (5/6), 506–525.

Williams, R., 1999. Cultural safety: what does it mean for our work practice? Australian and New Zealand Journal of Public Health 23 (2), 213–214.

Relevant Websites

http://seniorsecondary.tki.org.nz/Social-sciences/Business-studies/Maori-business/Culture-and-values – Māori Culture and Values in Business.

http://nni.arizona.edu/ – Native Nations Institute.

http://www.saubuntu.co.za/ – South African Ubuntu Foundation.

http://undesadspd.org/IndigenousPeoples/AboutUsMembers/History.aspx – United Nations Permanent Forum on Indigenous Issues.

http://www.firstfound.org/david%20miller.htm – Western and Mäori Values for Sustainable Development.

Indigenous Psychology

Rogelia Pe-Pua, University of New South Wales, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Indigenous psychology (IP) is both an approach or perspective and a movement in psychology. There have been several attempts to define it. The core elements of this approach are indigenous cultures used in explaining behavior and mental processes, using indigenous interpretations, and discovery or revalidation of psychological concepts and theories that emerge from within the culture. These concepts and theories may not necessarily have emerged or been emphasized in current psychological literature that has been pointed out as having a predominantly North American or European character or 'bias.' This article will examine the different definitions of IP, and outline its main features. It will outline the historical background of the perspective and movement, followed by a portrayal of the development and current status in the main societies where this has advanced significantly. Particular attention will be given to contributions to theory and methods. The article will end with an outline of ongoing debates, future prospects as well as challenges for IP.

Definition of Indigenous Psychology

Indigenous psychology (IP) is both an approach or perspective and a movement in psychology. It has been defined by a number of scholars, most of whom were the founders of the IP movements in their respective countries. Here are key examples of how IP was defined:

Psychology that is anchored on the thought and experience of the indigenous people, as understood from an indigenous perspective (adaptation from the Philippines' Virgilio Enriquez's definition of *Sikolohiyang Pilipino*, the Philippine IP, 1975).

Culturally appropriate psychology, or a route to 'appropriate' psychology (Japan's Hiroshi Azuma, 1984: 53).

"Psychological knowledge that is native, that is not transported from another region, that is designed for its people. In other words, indigenous psychology is understanding rooted in a particular sociocultural context" (South Korea's Uichol Kim, 1990: 145).

Psychology that emerges from cultural tradition; lies in daily, mundane activities of people; understands and interprets in terms of indigenous and local frames of reference and culturally derived categories (Berry et al., 1992: 380–381); and embodies psychological knowledge that is relevant and is designed for its people – reflecting the sociocultural reality of its society (India's Durganand Sinha, 1997: 132).

"Various definitions express the same basic goal of developing a scientific knowledge system that effectively reflects, describes, explains, or understands the psychological and behavioral activities in their native contexts in terms of culturally relevant frames of reference and culturally derived categories and theories" (Taiwan's Kuo-Shu Yang, 2000: 245–246).

Denotes three distinct phenomena: (1) the actual culturally organized psychology of an individual, (2) the local understanding of an individual's psychology, and (3) cultural–psychological theory and methodology (Ratner, 2008).

A number of elements of IP can be drawn from these definitions, for example:

Primacy of the indigenous or local or culturally derived perspective, understanding, categories, interpretation, explanation, frames of reference.

Relevance to the indigenous/native culture/people, reflecting their sociocultural reality.

The indigenous culture as the source of concepts and theories, rather than a set of imposed theories and knowledge.

'Indigenous compatibility,' which refers to the extent by which "the researcher's concepts, theory, methods, tools, and results adequately represent, reflect, or reveal the natural elements, structure, mechanism, or process of the studied phenomenon embedded in its context" (Yang, 2000: 250).

A movement or potentially a discipline.

The development of IP as a movement in various regions of the world was the subject of a survey conducted by Allwood and Berry in 2006. Their article included 15 contributions that focused on global development and characteristics, country development and local characteristics in each of the countries that they surveyed. They provided the following useful summary (abstract):

> Post-colonial reactions to mainstream psychology, and the belief that it was not an efficient aid to solving local social problems, were seen as important reasons for developing IPs. IPs were generally seen as attempts to produce a local psychology within a specific cultural context. Different views about what methods are legitimate in IPs were present (from experiments to various more 'humanistic' methods). IPs were commonly seen as being able to open up, invigorate, and improve mainstream psychology. The style of theorizing in the IPs was felt by many to be to build theories from the 'bottom up' on the basis of local phenomena, findings, and experiences. Some contributors saw the IP as a kind of cultural psychology, and a few noted that IP and cross-cultural psychology have an interactive mutually enriching relationship. Nearly half of the contributors emphasized the critical reaction to their work on IP by colleagues working more in the line of mainstream psychology. Many contributors felt that IP could contribute to the development

> of a more general universal psychology. Different indications of heterogeneity in the IPs were found among the contributors, for example, with respect to the role given to religion in the local IP. Sometimes the presence of different IPs within the same country was reported. This also indicates heterogeneity in the IPs.
>
> Allwood and Berry, 2006: 243

The term 'Indigenous psychology' begs the question, "What is the indigenous in indigenous psychology?" Borrowing from the attempt of Enriquez to clarify this during the early days of IP in the Philippines, it is possible to clarify three uses of 'indigenous': *Indigenous psychology* refers to psychology that is anchored on the thought and experience of the indigenous people, as understood from an indigenous perspective. *Psychology of the indigenous people* refers to any analysis of psychological behavior of the indigenous people, whether from an indigenous perspective or not, i.e., whether using indigenous or nonindigenous theories and methods. *Psychology in the indigenous country or culture* would then refer to the development of psychology in the particular culture or country, embodying both indigenous and nonindigenous perspectives and research.

But the term 'indigenous' needs further elaboration. Indigenous Maori (from New Zealand) psychologist Linda Nikora and her colleagues (in Allwood and Berry, 2006: 254) gave two meanings for the term 'indigenous': (1) Fourth World peoples (Nikora et al. (in Allwood and Berry, 2006: 254) referred to "Fourth World nations, defined here as indigenous communities positioned within First and Second World nations, for example, Hawai'ians, Aboriginals, and Maori – the original inhabitants of the lands in which they dwell."); and (2) all peoples residing in a society. In countries such as Australia, New Zealand, and Canada, there are distinct and widely recognized/acknowledged Indigenous peoples who are the first inhabitants of the land. In these countries, the word 'Indigenous' (with a capital 'I') in IP clearly refers to these peoples. On the other hand, in countries such as the Philippines, China/Taiwan, and India, 'indigenous' in the term IP refers to the people residing in those societies. The term has also come to be used in contrast to the 'exogenous' (that is, theories and methods that have been imported from the West). This has to be understood in terms of a historical background of colonization that these countries underwent (by Spain, Japan, and the United States in the case of the Philippines; by British rule in the case of India; and academic colonization by the West in the case of all the three countries).

Indigenization

IP as a movement has been referred to as the "indigenization movement in psychology" (Hwang, 2013: 716), which Hwang traced back to Wilhelm Wundt and his first laboratory at Leipzig in 1879. D. Sinha (1997) defined indigenization as the process of developing IP, the process or the strategy for obtaining an IP. This process evolves through stages, beginning with an acknowledgment of the limitations of Western theories and methods, which are then corrected by adapting them to suit local realities, or through the discovery of indigenous concepts and methods arising from the local culture.

Adair (2006) also talked about stages, identifying indigenization as one of the four stages in the development of IP: importation, implantation, indigenization, and autochthonization. "Autochthonization refers to the processes leading to the emergence of a self-perpetuating discipline independent of its imported source, the culmination of the indigenization process" (p. 472).

Pickren (2013) defined indigenization as "the process whereby a local culture or region develops its own forms of knowledge and practice, either by developing them from within that culture, or by importing knowledge and practices developed elsewhere and combining them with local concepts" (p. 698). Pickren argued that "American psychology was indigenized from German, French, and English developments that were imported to North America and mixed with home-grown metaphysics, religion, and healing practices (e.g., mesmerism), each of which has its own history of indigenization" (p. 698). And this North American (indigenous) psychology was and continues to be dominant in the training of psychologists in many Asian and African countries that lead many to experience the lack of fit to their own cultures. Because of this, they "began a self-conscious indigenization process" (p. 698).

Long before the above definitions of indigenization were offered, Virgilio Enriquez, the forefather of IP in the Philippines, had been very much involved with the process since the early 1970s. This hands-on experience with IP led him to clarify two pathways of indigenization. The first path is *indigenization from without* (Enriquez, 1987, 1992) or *indigenization of the exogenous* (D. Sinha, 1997). Using the exogenous culture as the source, the flow of indigenization is inward, i.e., the indigenous culture is the target. The goal is to create or produce an indigenous version of the exogenous or imported materials, adapting the foreign material to the indigenous context. Typical examples are translation of Western theories, modification of psychological tests or other measures, and replication of Western studies. The second path of indigenization is *indigenization from within* or *cultural revalidation* (Enriquez, 1987, 1992) or *internal indigenization* (D. Sinha, 1997). Here, the source of concepts and methods is the indigenous culture. This path involves processes such as semantic elaboration, indigenous codification or recodification, and systematic explication of implied theoretical frameworks, to produce knowledge that reflects the indigenous experience and realities (Enriquez, 1987, 1992). This strategy of indigenization also pays attention to the application and use of the generated knowledge to benefit the indigenous. The outward flow indicates that there is every intention to apply or relate this outside the indigenous culture. The term 'indigenization' does not seem to fit this second pathway since, how do you indigenize something that is already indigenous? (Pe-Pua and Protacio-Marcelino, 2000); thus the term 'cultural revalidation' seems more appropriate.

India's J.B.P. Sinha offered five overlapping trends of indigenization: purist endogenous trend, endogenous indigenization, purist exogenous trend, exogenous indigenization, and integrative indigenization "in which Western and Indian concepts and methods were integrated to produce hybrid concepts and theories" (J.B.P. Sinha, 2003; also in Allwood and Berry, 2006: 256). Vohra (2004) argues for a combination of these trends to help and ensure the sustainability of the indigenization movement.

As borne out by the development of IP in various countries, there is no single path for developing IP. Each country has their particularities in terms of national history, policies, priorities, and positioning in the academic environment. These tend to influence the path taken by the majority and the minority of psychologists, as well as attitudes toward IP in general.

Cross-Indigenization

One of the end goals of IP is to contribute to universal psychology. In contrast to the cross-cultural approach of comparing cultures, Enriquez proposed 'cross-indigenization' as the strategy for consolidating various indigenization efforts. "In this model [cross-indigenization], the different cultures of the world are tapped as sources of cultural knowledge. The resulting pool may then be called 'cross-cultural knowledge.' More aptly, it is cross-indigenous knowledge…" (Enriquez, 1992: 86).

Development of IP in Various Parts of the World

As mentioned earlier, Allwood and Berry (2006) provided a good overview of the development of IP in various parts of the world. Notable among these countries in terms of advancement are those where 'indigenous' refers to the 'nonexogenous' such as the Philippines, India, and Taiwan; and those where 'indigenous' refers to the Indigenous peoples, such as New Zealand. These developments are outlined briefly below.

The development of IP in the Philippines is well documented (Pe-Pua and Protacio-Marcelino, 2000; Pe-Pua and Perfecto-Ramos, 2012; Yacat, 2013; Gastardo-Conaco, 2005). In fact, there is a wide recognition of the advanced development of IP in this country than any other (D. Sinha, 1997).

The indigenization movement in psychology also has a long history in Taiwan, spearheaded by Taiwanese psychologists Yang Kuo-Shu and Hwang Kwang Kuo (see Hwang, 2012 for a comprehensive history).

The IP movement in India has also been the subject of a number of writings (Sarawathi, in Allwood and Berry, 2006; J.B.P. Sinha, in Allwood and Berry, 2006; D. Sinha, 1997).

There are a number of similarities in the development of IP in the three countries. (1) All started IP development in the 1970s. The Philippines started in 1971 when its founder Virgilio Enriquez (1942–94) returned from the United States. Taiwan started in the late 1970s, with the founder Yang Kuo-Shu influenced by the development of IP in the Philippines. In the early 1970s, India's Durganand Sinha made a call "to make psychology relevant to the Indian context and to serve the needs of the Indian people" (Vohra, 2004). However, it was not until the 1980s that a clear articulation of IP was made. (2) There are recognized 'founding fathers' of IP in these countries: Enriquez (the Philippines), Yang (Taiwan), and D. Sinha (India). Both Enriquez and Yang were trained in the United States; both started thinking about IP even before they returned to their home countries. And after their return, they embarked on discussions with colleagues and planting the seeds for the development of IP. (3) All countries have professional organizations that are dedicated to the advancement of IP. (4) All countries have published large quantities of work on IP, mostly in the native languages. (5) The use of the native language in research and writing/publishing has been prevalent from the beginning of the IP movement in the Philippines and Taiwan. During the last 10 years, there have been pressures for these scholars to publish in English due to institutional pressure to publish in internationally refereed journals (in order to get academic promotion) and due to the need to promote IP internationally. (6) In all the three countries, indigenization from without has been the more popular path to indigenization.

Particular to the Philippines is the identification of four lines of filiations of psychology in the country: academic–scientific psychology, academic–philosophical psychology, ethnic psychology (which is the major basis of *Sikolohiyang Pilipino*), and the psychomedical system (Salazar, 1985).

Taiwan's IP development saw the determination "to refrain from uncritically or even blindly, applying American psychologists' concepts, theories, methods, and tools to the study of Chinese behaviour. Instead, they have based their studies on Chinese historical, cultural, social, and language traditions, especially the Confucian, Taoist, and Buddhist ones, and have sought to develop an indigenous Chinese psychology suitable for people in all the Chinese societies" (Yang, in Allwood and Berry, 2006: 251).

India's history of IP is no different from the Philippine and Taiwanese in terms of realizing the limitations of Western psychological theories and tools. But if the Philippines has its ethnic psychology as main basis, and Taiwan has its Confucian, Taoist, and Buddhist inspirations, Indian IP is "closely entwined with religion. Hinduism, which is described more as a way of life than as a religion, prescribes a code of conduct for human behaviour, recognizing the changing developmental tasks in different life stages" (Sarawathi, in Allwood and Berry, 2006: 255). Therefore, Indian IP draws on Hindu philosophy (Sarawathi, in Allwood and Berry, 2006: 255). In addition, it has a strong basis on folklore practices (D. Sinha, 1997).

The Philippines, Taiwan, and India are countries where indigenization has emerged in reaction to Western imported knowledge. The term 'indigenous' refers to the non-exogenous (aka non-Western), leading some scholars to suggest that the term 'national psychology' is more accurate than IP. Be that as it may, the use of the 'indigenous' perspective is what psychologists in these countries would hold on to in claiming that theirs is 'indigenous psychology.'

On the other hand, New Zealand, Canada, and Australia represent countries where the 'indigenous' in IP refers only to their original inhabitants. (We will discuss New Zealand only here since it has a clear history of indigenization in psychology.) Colonization of the Indigenous peoples in these countries is quite different from the three other countries discussed previously in the sense that the colonization process is ongoing as the Indigenous people continue to live side by side with their colonizers. As Sanson (2000) compares, in countries like the Philippines and India, "such a psychology [IP] would pertain to people of non-Western, largely non-industrialized nations, who have experienced colonization and domination by a Western power; but who nevertheless form the majority in their own country" (p. 13). In contrast, the Indigenous peoples in Australia and New Zealand form the minority of

the population, with the 'colonizer' background forming the majority. Furthermore, there are other cultural background groups in both countries, some of whom number more than the Indigenous groups. Thus, the struggle of indigenization of the discipline of psychology in these countries occurs hand in hand with the political struggle of the Indigenous peoples.

New Zealand's indigenization movement has focused on the relevance of psychology to the Indigenous people of the country, i.e., the Maori.

> Indigenous psychology in A/NZ [Aotearoa/New Zealand] has always been part of how Maori approach wellness, health, and being, stemming from a world-view that values balance, continuity, unity, and purpose … Perhaps it is best referred to by the Maori term *tikanga*, or customary practice – those behaviours, values, ways of doing things, and understanding actions that have always (sic) and will continue to be with us.
>
> Nikora et al., in Allwood and Berry, 2006: 254

Michelle Levy's (2007) PhD thesis *Indigenous Psychology in Aotearoa: Realising Maori Aspirations* provides a comprehensive account of the history of IP in A/NZ,

> "the current status of Maori development in psychology, the importance of the critical mass and the notion of collective responsibility … [It also] identifies 'reaching the point of irreversible change' as the next phase of indigenous psychology development in Aotearoa. This is the point at which indigenous psychology development becomes self-sustaining"
>
> (p. ii).

Just like the three countries previously discussed, psychology as a discipline in A/NZ was 'imported' from the West, having been introduced in 1874. Nikora and colleagues (in Allwood and Berry, 2006) pointed out that interest in Maori between the 1940s and the 1960s led to research that "was criticized for being 'on' Maori, rather than 'with' Maori (p. 254). It was not until around 1987 that psychologists became involved in the 'Maori development agenda,'" inspired and led by Mason Durie, a professor of Maori Studies, also a psychiatrist and a psychologist. Articulating *tikanga* in a clear and concise way, Durie's work has influenced the health and welfare sectors. The ongoing agenda for psychology is "to create psychologies to meet the needs of Maori people in a way that maintains a unique Maori heritage, and makes for a better collective Maori future… towards Maori self-determination… [This requires] the development of a critical mass of indigenous psychologists capable of developing robust *tikanga*-based psychological frameworks" (p. 255). In terms of action in promoting a Maori-centered psychology, Linda Nikora is regarded as the prime mover, with the help of the Maori and Psychology Research Unit based at the University of Waikato in Hamilton, New Zealand.

Achievements of IP

What do indigenous psychologists see as the achievements of IP? A common theme is decolonization. For example, IP in the Philippines started as a movement at a time when the country was rebelling against a dictatorial government, and rejecting 'imperialistic' influence from the West (aka America). Decolonization in psychology was about being critical of Western theories and methods, and discovering their 'own' psychology. The decolonization in Taiwan was also in terms of breaking away from Western psychology. Both countries felt the inadequacy of Western theories in explaining their realities. Decolonization was also an important component of IP history in A/NZ but this was focused on how psychology can help solve the impact of colonization on the well-being of the Maori people.

Another achievement of IP has been in redefining psychology or highlighting what is important as foci of the discipline. Enriquez (1975) took into account the study of emotions and experienced knowledge (*kalooban* and *kamalayan*), awareness of one's surroundings (*ulirat*), information and understanding (*isip*), habits and behavior (another meaning of *diwa*), and the soul (*kaluluwa*) which is the way to learn about people's conscience. Yang (1999) defined psychology as "an empirical science that investigates, from an individualistic-orientation and collectivist-orientation perspective, the contents, structure, mechanism, and processes of human mind and behavior conditioned by physical, biological, social, cultural, and historical characteristics at the intrapersonal, interpersonal, intragroup, and intergroup levels." The A/NZ perspective would focus on *tikanga* and related concepts as what constitute psychology.

Psychological testing is another area where both the Philippines and Taiwan made a contribution in terms of indigenization. This is not just the translation of foreign-made tests but starting with indigenous personality traits then moving to develop the tests.

Research methods is an area where IP in the Philippines has made a significant contribution. The advances in this area are discussed by Pe-Pua (2006). These have their humble beginnings in *pakapa-kapa*, introduced by Santiago (1977) and later defined by Torres (1982: 171) as "a suppositionless approach to social science investigations … characterized by groping, searching, and probing into an unsystematized mass of social data to obtain order, meaning and directions for research." The articulation of *pakapa-kapa* was significant, representing a turning point, paving the way to scrutinize research topics so that they could be more relevant to the needs of Filipinos. It dared the social scientists to question Western methods. "The resulting methods were considered indigenous – not imported nor invented, but natural or existing patterns of behaviour (not methods), discovered and developed as research methods" (Pe-Pua, 2006: 112). Aside from describing some of the indigenous methods that have emerged in the Philippine IP, Pe-Pua explained some guiding principles in undertaking psychological research the IP way. These are: (1) how the quality of research data is affected by the level of social interaction between researcher and participants; (2) the importance of treating participants as equal; (3) valuing the welfare of participants more than just getting data from them; (4) choosing methods on the basis of appropriateness, and adapting these to existing norms; and (5) the language of the participants should be the language of research (Pe-Pua, 2006).

Indigenous Theorizing

IP has made some inroads in identifying and developing indigenous concepts and theories – not necessarily to be unique to one culture only, but unique in the sense of not having been tapped or articulated as culturally sensitive in mainstream psychology. Some examples are given below.

Enriquez's book *From Colonial to Liberation Psychology* (1992) provides a good overview of the early development and philosophy of *Sikolohiyang Pilipino* (indigenous Filipino psychology). Enriquez discovered *kapwa* as a core Filipino indigenous concept which turned out to be the antithesis of the Western individual-centered 'self.' The English 'others' is used in opposition to the self, implying the separation of the self as a separate identity. In contrast, *kapwa* is a recognition of shared identity, an inner self shared with others; the unity of 'self' and 'others'; or fellow human beings (Enriquez, 1978, 1994). In Filipino psychology, *ako/sarili* (I/self) is associated with selfishness or egocentricism. There are words in the Filipino language for 'I' (*ako*) and 'others' (*ibang tao*) but both are included in *kapwa* psychology, so much so that a Filipino would say, "*Hindi ako iba sa aking kapwa* (I am no different from others)" (Enriquez, 1994). Several studies and writings have focused on *kapwa*, including its place in indigenous research approaches (see Pe-Pua, 2006 for a summary), and *kapwa* as value and social interaction (Enriquez, 1978, 1994). The concept of *kapwa* reflects the flexibility by which Filipinos navigate the domains of inclusion–exclusion or membership/non-membership. These domains are relevant to Filipinos, but in a more flexible way. While Filipinos make a distinction between the *ibang-tao* ('not one of us') and the *hindi-ibang-tao* ('one of us'), both categories are accommodated in the concept of *kapwa*.

The Taiwanese Yang Kuo-Shu's (1995) theory of Chinese social orientation is quite advanced and has potential application to other cultures. This is a theory of Chinese psychological functioning centering on the key concept of social orientation as a comprehensive system as well as a synthetic pattern of Chinese dispositional characteristics. Yang (1999: 194) defined social orientation as "a person's tendency to establish and maintain a harmonious relationship with, and merge into, the surroundings so that collective and relational goals can be effectively achieved." Yang's theory points to four closely related orientations: familistic orientation, other orientation, relationship orientation, and authoritarian orientation. The relationship orientation is particularly elaborate, consisting of three compartmentalized categories of relationship: *jiaren* (family members); *shuren* ('cooked' persons: relatives outside the family, people in the same village, neighbors, friends, colleagues, and classmates); and *shengren* ('raw' persons like strangers). These categories require distinct modes of interaction in accordance with different principles: *jiaren* (*zeren* or responsibility); *shuren* (*renqing* or interpersonal favor and affect); *shengren* (*lihai* or an eye on possible gains and losses) (Yang, 1999).

Another Taiwanese psychologist Hwang (2012) has been advocating for culture-inclusive theories in psychology following Shweder's (Shweder et al., 1998) principle of 'one mind, many mentalities.' He has urged the IP movement to have as its mission "to initiate a scientific revolution by constructing a series of theories on the presumption of relationalism in replacing Western theories of individualism, so as to help people of non-Western countries solve the various problems they encounter in their daily lives" (Hwang, 2013: 717). He integrated the various theoretical contributions of Chinese IP, as well as the history and philosophy of Chinese IP in his book *Foundations of Chinese Psychology: Confucian Social Relations* (Hwang, 2012).

Hwang's theoretical model of face and favor is a major contribution in this direction of indigenous theorizing. This is basically a theory in interpersonal relationship that is related to four key indigenous concepts: *guanxi* (relationship), *renqing* (favor), *mianzi* (face), and *bao* (reciprocation) (Hwang, 2012). Hwang identified four kinds of interpersonal ties: expressive ties, mixed ties, instrumental ties, and vertical relationship. He then related them to Fiske's (1991) model of elementary forms of social behavior. He then illustrated how certain forms of social behavior are given more credence depending on the orientation of the particular society on the individualism ideal (Hwang, 2013).

There are more examples from the Philippines and Taiwan where IP has developed a quite advanced history. Enriquez's theory of *kapwa*, Yang's theory of social orientation, and Hwang's theory of face and favor present an opportunity for developing a cross-indigenous (it will be premature to speculate 'universal' at this point) theory of the social self. There are a number of similarities across the three theories, which will be the subject of future writings.

The relational orientation was also prominent in Indian indigenous theorizing. Instead of the duality of man and society in Western conceptualization, the Indian perspective values 'man-society' where the two cannot be separated (D. Sinha, 1981). "The individual's identity is conceived not so much in terms of his/her personal qualities and achievements, but in the context of his/her family, caste, place, institutions, and a whole web of interpersonal roles and relationships. It is in this perspective that social behavior is analyzed and interpreted" (D. Sinha, 1997: 151). This relational model is also congruent with Indian healing practices (Kakar, 1984). Incidentally, several indigenous Indian concepts and practices have permeated and have been embraced by Western psychotherapy, including *Yoga*, *Ayurveda*, and so on. These are clear contributions to the discipline of psychology.

There is a Maori concept that comes close to relationalism – *whanangatanga*, meaning relationship, relatedness – that is quite prominent in A/NZ IP development.

Japanese IP is also a source of indigenous theorizing that has potential international application. One such concept that has gained international attention is *amae*, which is defined by Yamaguchi (1999, cited in Yamaguchi and Ariizumi, 2006) as

> …presumed acceptance of one's inappropriate behavior or request. In close relationships, one is able to presume that one's inappropriate behavior or request will be accepted due to the positive or at least non-negative attitude of one's counterpart. This definition of *amae* implies that two components, the inappropriate behavior or request and the presumption of acceptance, are involved in *amae* episodes.
>
> p. 165

There are indeed several developments in indigenous theorizing that could be pulled together through systematic research in order to illustrate the potential contribution to psychological theory in different fields of the discipline.

Issues and Debates on IP

Psychologists undertaking IP work face a number of challenges and issues. One challenge is related to the language of research and publication. In countries where the first language is not English, most of the publications on IP work are in the local languages, mainly due to the philosophy behind IP that they uphold, i.e., IP should be developed and is best understood and articulated using the local language. In fact, in India, one of the reasons identified for the slow progress in developing IP is the reduced usage of the local languages, which is explained in terms of the dilemma of diversity in languages, religions, and grouping in the country. In some countries, the lack of publication in English, which would make IP more accessible internationally, is due to their lack of English proficiency. A subtle reason also is the attitude held by some psychologists that spreading IP knowledge internationally is not a priority; what is important is that the national development benefits from IP knowledge. Another obstacle to publishing, in general, is the heavy teaching load in many universities and the lack of support (or pressure) to conduct research and have this recognized as part of their academic workload.

Indigenous theorizing is an important undertaking if IP is to be taken seriously by the wider international community of psychologists. Any theorizing, however, is neither a simple nor a straightforward process. There is even a question of whether indigenous theories are useful, i.e., are these just confined to the indigenous cultures or do they have potential to be universal? Stevenson (2000) speculated on 'indigenous theories' failure to flourish,' offering the following reasons: (1) lack of resources for research to advance theorizing; (2) the disciplinal bias for topics that are interesting to a more universal audience; (3) difficulty with linguistic translation; (4) the use of a mere translated version of Western instruments that may not be relevant to the indigenous population; and (5) 'newness of the considerations' – there are simply not enough studies done.

The reality is that there is already a wealth of indigenous concepts and theories out there that are in various stages of development. A more systematic and well-resourced stocktaking is required to bring these out. The body of knowledge needs to be synthesized, and future directions for research are to be guided by this synthesis.

Cross-indigenization would seem to provide a good process for extending the applicability or relevance of indigenous concepts and theories to a more international context. For example, some Japanese psychologists provide a good example of cross-indigenization without necessarily being aware of it. When they identify a 'Japanese' (they do not readily use the word 'indigenous') concept, they try to understand it from the indigenous perspective, relate it immediately to the Western concepts closest to it, and try to collect empirical data, not only with Japanese subjects, but also with non-Japanese subjects. In many cases, part of the comparison is with Western subjects.

Cross-indigenization is a pathway to discovering universals, if there is indeed such a thing. For example, Yamaguchi and Ariizumi (2006) speculate, as a result of an initial investigation of the Japanese concept of *amae* across three cultural groups, that *amae* could be an etic construct. At the same time that researchers of IP need to be supported in contributing to universal psychology, the wider scholarly community also needed more acceptance and valuing of IP. Developing IP and cross-indigenization help us move away from 'intellectual prejudice,' 'entrenched intellectual habits,' and 'cultural myopia.' Cross-indigenization provides a balance of flow of knowledge between East and West. Not only that, the flow of knowledge within the East is also important. Cross-indigenization also helps us bridge the gap between 'emotional overreactions' and blind assumption that "scientific psychology is a monopoly of the West" (Enriquez, 1992).

In terms of future directions of IP, Church and Katigbak (2002) offered seven pressing needs in relation to Philippine IP that could very well apply to other IPs as well. These are:

> "the formulation of indigenous theory; objective consideration informed by empirical data; continuing development and validation of indigenous measures; systematic investigation of the comparative and convergent validity of various indigenous and imported research methods; institutional/structural improvements leading to growth and stability of the indigenous research culture; maintenance of an appropriate balance between the pursuit of an independent psychology and the avoidance of insularity; and eventually, increased efforts to relate cross-indigenous approach toward a universal psychology"
>
> (p. 141).

By way of concluding, it is important to highlight that IP has been the subject of a number of edited volumes that are inclusive of different countries. The earliest of these was Kim and Berry's (1993) book *Indigenous Psychologies: Research and Experience in Cultural Context* (Sage). A more recent one, Kim et al.'s (2006) volume, *Indigenous and Cultural Psychology: Understanding People in Context* (Springer) was the result of an intensive seminar-workshop in Taiwan that explored the latest developments of the time. This book was then reviewed by a number of authors, with the reviews published in a special issue of the journal *Pastoral Psychology* (volume 56, 2007), a Springer journal. A few journals have also promoted discussion and publication on IP. For example, volume 3, issue 3 (2000) of the *Asian Journal of Social Psychology* provoked a healthy exchange in clarifying the distinctions between cultural, cross-cultural, and indigenous psychology. Another issue of the journal, volume 8, issue 1 (2005) was dedicated to "responses to the epistemological challenges to indigenous psychologies." The *International Journal of Psychology* also devoted a "Special issue on the indigenous psychologies" (volume 41, issue 4, 2006).

See also: Acculturation; Critical Psychology; Cross-Cultural Psychology; Cultural Influences on Interpersonal Relationships; Cultural Psychology; Cultural Views of Life

Phases; Culture and Emotion; Culture and the Self: Implications for Psychological Theory; Ethnic Identity, Psychology of; Indigenous Management Styles; Indigenous Social Work; Prosocial Behavior, Cultural Differences in; Social Psychology: Research Methods; Subjective Wellbeing, Psychology of; Values Across Cultures, Development of.

Bibliography

Adair, J., 2006. Creating indigenous psychologies. In: Kim, U., Yang, K.-S., Hwang, K.-K. (Eds.), Indigenous and Cultural Psychology: Understanding People in Context. Springer US, New York, pp. 467–485.

Allwood, C.M., Berry, J.W., 2006. Origins and development of indigenous psychologies: an international analysis. International Journal of Psychology 41 (4), 243–268.

Azuma, H., 1984. Psychology in a non-western country. International Journal of Psychology 19 (1–4), 45–55.

Berry, J.W., Poortinga, Y.H., Segall, M.H., Dasen, P.R., 1992. Cross-cultural Psychology: Research and Applications. Cambridge University Press, Cambridge.

Church, A.T., Katigbak, M.S., 2002. Indigenization of psychology in the Philippines. International Journal of Psychology 37 (3), 129–148.

Enriquez, V.G., 1975. Mga batayan ng Sikolohiyang Pilipino sa kultura at kasaysayan (The bases of Filipino psychology in culture and history). General Education Journal 29, 61–88.

Enriquez, V.G., 1978. Kapwa: a core concept in Filipino social psychology. Philippine Social Sciences and Humanities Review 42 (1–4).

Enriquez, V.G., 1987. Decolonizing the Filipino psyche: impetus for the development of psychology in the Philippines. In: Blowers, G.H., Turtle, A.M. (Eds.), Psychology Moving East: The Status of Western Psychology in Asia and Oceania. Westview Press, Boulder & London, pp. 265–287.

Enriquez, V.G., 1992. From Colonial to Liberation Psychology: The Philippine Experience. University of the Philippines Press, Quezon City.

Enriquez, V.G., 1994. Pagbabangong-dangal: Indigenous Psychology and Cultural Empowerment. Akademya ng Kultura at Sikolohiyang Pilipino, Quezon City.

Fiske, A.P., 1991. Structures of Social Life: The Four Elementary Forms of Human Relations. The Free Press, New York, NY.

Gastardo-Conaco, M.C., 2005. The development of a Filipino indigenous psychology. Philippine Journal of Psychology 38 (2), 1–17.

Hwang, K.K., 2012. Foundations of Chinese Psychology: Confucian Social Relations. Springer, New York.

Hwang, K.K., 2013. Indigenous psychology. In: Keith, K.D. (Ed.), The Encyclopedia of Cross-cultural Psychology. John Wiley & Sons, Chichester, UK, pp. 716–718.

Kakar, S., 1984. Psychological counselling: is there an Asian model?. In: Proceedings of the 5th Biennial Conference-Workshop of the Asian Association of Psychological and Educational Counsellors of Asia. Bangalore.

Kim, U., 1990. Indigenous psychology. In: Brislin, R.W. (Ed.), Applied Cross-cultural Psychology. Sage Publications, California, pp. 143–160.

Kim, U., Berry, J.W. (Eds.), 1993. Indigenous Psychologies: Research and Experience in Cultural Context. Sage, Newbury Park.

Kim, U., Yang, K.S., Hwang, K.K. (Eds.), 2006. Indigenous and Cultural Psychology: Understanding People in Context. Springer, New York.

Levy, M., 2007. Indigenous Psychology in Aotearoa: Realising Māori Aspirations (Ph.D.). University of Waikato, Hamilton, New Zealand.

Pe-Pua, R., 2006. From decolonizing psychology to the development of a cross-indigenous perspective in methodology: the Philippine experience. In: Kim, U., Yang, K., Hwang, K. (Eds.), Indigenous and Cultural Psychology: Understanding People in Context. Springer, New York, pp. 108–137.

Pe-Pua, R., Perfecto-Ramos, P., 2012. Philippines. In: Baker, D.B. (Ed.), The Oxford Handbook of the History of Psychology: Global Perspectives. Oxford University Press, New York, pp. 395–411.

Pe-Pua, R., Protacio-Marcelino, E., 2000. Sikolohiyang Pilipino (Filipino psychology): a legacy of Virgilio G. Enriquez. Asian Journal of Social Psychology 3 (1), 49–71. http://dx.doi.org/10.1111/1467-839X.00054.

Pickren, W.E., 2013. Indigenization. In: Keith, K.D. (Ed.), The Encyclopedia of Cross-cultural Psychology. John Wiley & Sons, Chichester, UK, pp. 698–699.

Ratner, C., 2008. Cultural Psychology, Cross-cultural Psychology, and Indigenous Psychology. Nova Science Publishers, New York.

Salazar, Z., 1985. Four filiations in Philippine psychological thought. In: Aganon, A., David, M. (Eds.), Sikolohiyang Pilipino: Isyu, Pananaw at Kaalaman (New Directions in Indigenous Psychology). National Bookstore, Manila, pp. 194–214.

Sanson, A., 2000. Commentary: an Australian perspective on indigenous psychology. International Society for the Study of Behavioural Development Newsletter 1 (37), 12–13.

Santiago, C.E., 1977. Pakapa-kapa: Paglilinaw ng isang konsepto sa nayon (Pakapa-kapa: clarifying a concept in a rural setting). In: Pe-Pua, R. (Ed.), Sikolohiyang Pilipino: Teorya, Metodo at Gamit (Filipino Psychology: Theory, Method and Application), 1982 ed. Surian ng Sikolohiyang Pilipino, Quezon City, pp. 161–170.

Sinha, D., 1981. Social psychology in India: a historical perspective. In: Pandey, J. (Ed.), Perspectives on Experimental Social Psychology in India. Concept, New Delhi, pp. 3–17.

Sinha, D., 1997. Indigenizing psychology. In: Berry, J.W., Poortinga, Y.H., Pandey, J. (Eds.), Handbook of Cross-cultural Psychology, second ed., vol. 1. Allyn & Bacon, Boston, MA, pp. 129–169.

Sinha, J.B.P., 2003. Trends towards indigenization of psychology in India. In: Yang, K.S., Hwang, K.K., Pederson, P., Diabo, I. (Eds.), Progress In Asian Psychology: Conceptual and Empirical Considerations. Praeger, West Port, CT & London, pp. 11–28.

Shweder, R.A., Goodnow, J., Hatano, G., LeVine, R., Markus, H., Miller, P., 1998. The cultural psychology of development: one mind, many mentalities. In: Damon, W. (Ed.), Handbook of Child Psychology: Theoretical Models of Human Development, vol. 1. John Wiley & Sons, Inc., New York, pp. 865–937.

Stevenson, H., 2000. Commentary: indigenous theories' failure to flourish. International Society for the Study of Behavioural Development Newsletter 1 (37), 10.

Torres, A., 1982. "Pakapa-kapa" as an approach in Philippine psychology. In: Pe-Pua, R. (Ed.), Sikolohiyang Pilipino: Teorya, Metodo at Gamit. Surian ng Sikolohiyang Pilipino, Quezon City, pp. 171–174 (Filipino Psychology: Theory, Method and Application).

Vohra, N., 2004. The indigenization of psychology in India: its unique form and progress. In: Setiadi, B.N., Supratiknya, A., Lonner, W.J., Poortinga, Y.H. (Eds.), Ongoing Themes in Psychology and Culture. International Association for Cross-cultural Psychology, Melbourne.

Yacat, J., 2013. Filipino psychology (Sikolohiyang Pilipino). In: Keith, K.D. (Ed.), The Encyclopedia of Cross-cultural Psychology. John Wiley & Sons, Chichester, UK, pp. 551–555.

Yamaguchi, S., Ariizumi, Y., 2006. Close interpersonal relationships among Japanese: Amae as distinguished from attachment and dependence. In: Kim, U., Yang, K., Hwang, K. (Eds.), Indigenous and Cultural Psychology: Understanding People in Context. Springer, New York, pp. 163–174.

Yang, K.-S., 1995. Chinese social orientation: an integrative analysis. In: Lin, T.-Y., Tseng, W.-S., Ye, Y.-K. (Eds.), Chinese Societies and Mental Health. Oxford University Press, Hong Kong, pp. 19–39.

Yang, K.-S., 1999. Towards an indigenous Chinese psychology: a selective review of methodological, theoretical, and empirical accomplishments. Chinese Journal of Psychology 41 (2), 181–211.

Yang, K.-S., 2000. Monocultural and cross-cultural indigenous approaches: the royal road to the development of a balanced global psychology. Asian Journal of Social Psychology 3, 241–263.

Indigenous Rights

Jacquelyn A Jampolsky, University of Colorado, Boulder, CO, USA
Kristen A Carpenter, University of Colorado Law School, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M. Asch, volume 1, pp. 1–5, © 2001, Elsevier Ltd.

Abstract

Indigenous communities and nation-states understand indigenous rights in different and contested ways. However, as indigenous people increasingly appeal to the international community and international law to secure their rights against encroaching state and private interests, the term 'indigenous rights' has come to refer to a specific set of rights codified within the corpus of international law or set forth as international standards. This article discusses the historical origins and development of the contemporary indigenous rights regime through instruments of international law and policy.

Introduction

Indigenous communities and nation-states understand indigenous rights in different and contested ways. However, as indigenous people increasingly appeal to the international community and international law to secure their rights against encroaching state and private interests, the term 'indigenous rights' has come to refer to a specific set of rights codified within the corpus of international law or set forth as international standards. This article discusses the historical origins and development of the contemporary indigenous rights regime through instruments of international law and policy.

There is no universal definition of what it means to be 'indigenous.' The term 'indigenous' did not gain significance until after the onset of European exploration, when settler societies needed a way to collectively identify the hundreds of culturally diverse Native groups that typically self-identified by a name of their own language (Bruyneel, 2007: p. ix). Today a shared global identity around being indigenous is common. Within the indigenous rights framework, "[t]he rubric of indigenous peoples includes the diverse Indian and aboriginal societies of the Western Hemisphere, the Inuit and Aleut of the Arctic, the aboriginal peoples of Australia, the Maori of Aotearoa (New Zealand), Native Hawaiians and other Pacific Islanders, the Sami of the European far North, and at least many of the tribal or culturally distinctive nondominant peoples of Asia and Africa" (Anaya, 2009: p. 1). This list is not exhaustive of the rich multiplicity of cultures and lifeways of the more than 370 million indigenous people living throughout the world (Permanent Forum Factsheet). The United Nations (UN) describes indigenous people as sharing the characteristics of self-identification, strong communal ties tracing to precolonial and presettler Indigenous groups, strong cultural ties with traditional lands, and as practicing distinct linguistic, social, political, economic, and religious ways that differ from the dominant society (Permanent Forum Factsheet).

This article provides a general introduction to indigenous rights. First, it outlines the historical development of indigenous rights as a subject of international law. Second, it discusses the treatment of indigenous rights in binding international treaties and in international jurisprudence. Third, it discusses the United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP), as well as other legal sources of indigenous rights. Finally, it concludes by highlighting some of the critical debates in indigenous rights scholarship and practice.

Historical Development of Indigenous Rights

Colonialism and the Doctrine of Discovery

Indigenous peoples have a long history of internal, intertribal, and even international legal traditions. Some early encounters between Native peoples and Europeans were relatively peaceful, reflecting indigenous diplomacy and giving rise to mutually beneficial alliances for trade and political purposes (Williams, 1990). Others were not. The origin of indigenous rights as part of the state-based international legal order can be traced to the legal and political systems that governed European exploration, conquest, and colonization of the Western hemisphere. Papal law, dating as early as the eleventh century as the justification for the Holy Wars (Crusades) in Europe, granted the Church universal jurisdiction over all non-Christian lands and encouraged exploration and religious conversion of Native peoples (Miller, 2005). Violence between indigenous peoples and Europeans followed, along with enslavement, conquest, and devastation of Native peoples.

In the fifteenth century, European legal theorists began questioning the moral and legal right for Church-backed monarchs to assume control over indigenous lands. Natural law principles postulating moral and ethical ideals above the interests of the monarch suggested that indigenous peoples, like the rest of humanity, possess an innate set of rights. Spanish legal theorists and Dominican clerics Bartolmé de las Casas (1474–1566) and Francisco de Victoria (1486–1547) challenged the callous treatment and enslavement of indigenous peoples by affirming that indigenous peoples were indeed human beings, and thus entitled to certain rights possessed by all humans, such as control over property (Anaya, 2004a, b).

Although natural law theorists challenged the legality of conquest by discovery alone, they also argued that conquest could be justified if for the benefit of the Native peoples, severely limiting the scope of indigenous rights at its inception. De Victoria wrote that Native peoples are "unfit to found or administer

a lawful State up to the standard required by human and civil claims … therefore, be maintained that in their own interests the sovereigns … might undertake the administration of their country … so long as this was clearly for their own benefit" (de Victoria, 1557: p. 160). These ideas were incorporated into the articulation and application of the Doctrine of Discovery, which explicitly governed relations among competing European nations, and marked the first instantiation of indigenous rights as an issue of international law. The Doctrine of Discovery granted the first Christian European power to arrive on foreign soil the exclusive right to purchase or otherwise conquer Native lands against all other Christian European nations. "When European countries, and later the United States, set out to exploit new lands in the fifteenth through twentieth centuries, they justified their claims over these territories and over indigenous peoples with the Doctrine" (Miller, 2011: p. 330).

The Formation of the Modern International Legal System

Indigenous rights receded from the international arena and became part of the domestic affairs of states in the late seventeenth century. The 1648 Treaty of Westphalia ushered in a new era of international law defined by the Law of Nations. Emmerich de Vattel (1714–69) wrote that individual nation-states are equal, independent, and hold their own rights. Vattel's premise was solidified in the late nineteenth and early twentieth century by legal positivists who argued that international law could only govern state rights, and that the principle of mutual sovereignty shielded the internal affairs of individual states from international interference without consent. Post-Westphalian principles of sovereignty shielded domestic treatment of indigenous rights from international scrutiny, and indigenous peoples became subject to the unmitigated will of states to eliminate Native cultures, take Native lands, and strip Native peoples of political mobility (Anaya, 2009).

Many of the emerging domestic indigenous rights regimes incorporated the Doctrine of Discovery. In the United States, the Doctrine was famously invoked in 1823 by Supreme Court Justice John Marshall in the decision of *Johnson v. McIntosh* where the Court held: "[D]iscovery gave title to the government by whose subjects, or by whose authority, it was made against all other European governments, which title might be consummated by possession" (Johnson v. McIntosh, 1823: p. 573). In Marshall's formulation, the Doctrine of Discovery meant that the 'courts of the conqueror' would only recognize property transfers from indigenous nations to the national government, and not to private citizens. However, the Doctrine also recognized that indigenous peoples retained usufructuary, or use rights by permitting the continued use and occupation of traditional lands even after discovery, unless or until the United States chose to extinguish aboriginal title by purchase or conquest. This system of ownership established by the Doctrine of Discovery formed the foundation of the trustee relationship and the backdrop for domestic indigenous rights regimes in the United States, Australia, New Zealand, and Canada that still exist today (Miller and Ruru, 2008; Williams, 2012).

Although the Doctrine of Discovery and positivist developments in international law weakened the power of indigenous peoples internationally, the sovereign status of indigenous peoples remains viable largely due to the simultaneous practice of treaty making. Beginning in the early 1600s and continuing through as late as the early 1900s, English, Dutch, Spanish, and French colonists made treaties with tribal governments, recognizing the sovereign prerogative of indigenous peoples to make and enact binding instruments of international law that established conditions of peace and secured indigenous rights to property, governance, and natural resource use to name a few (Cohen and Newton, 2009: §1.01). Today, treaties continue to form the basis for domestic indigenous rights regimes in the United States, Canada, New Zealand, and Australia (Williams, 1999).

The horrors of World War I brought indigenous rights back into the international spotlight as part of the nascent human rights movement and the establishment of the League of Nations in 1919. Despite the inability of nonstates to participate in the League of Nations, indigenous peoples persisted in attempting to lobby for their rights. For example, the representative of the Six Nations of the Grand River and chief of the Younger Bear Clan of the Cayuga Nation Levi General Deskaheh lived in Geneva from September of 1923 until the end of 1924, while attempting to gain a hearing at the League of Nations. Deskaheh sought to resolve a dispute with the Canadian government over the 1876 Indian Act, which forced tribal peoples to form elected band councils for the purpose of eliminating traditional indigenous governments. Deskaheh argued that he had a right to a hearing because he represented a sovereign state, based in part on the Haldiman Treaty of 1784 in which King George III reserved the Grand River territory to the Iroquois who fought for the British during the American Revolution. While Deskeheh never obtained an official hearing at the League of Nations, he garnered support from at least six other nations and renewed the presence of indigenous peoples in international law (Niezen, 2003).

The League of Nations dissolved by the late 1930s but gave rise to the establishment of the United Nations (UN) in 1945 after the human rights violations of WWII. The UN provided a new forum for staging indigenous rights as part of the broader human rights movement in international law. Like the League of Nations, the UN is based on the founding principle of state sovereignty. However, in the aftershock of Nazi-genocide, the UN recognized the need to protect all human rights even if that meant that sovereignty could no longer be an absolute shield from international intervention in domestic affairs. This shift demonstrated the resurgence of principles of natural law in international law by championing the 'equal rights and self-determination of all peoples' including respect for fundamental freedoms regardless of race, sex, language, or religion (UN Charter, 1945: art. 2). The rights introduced in the UN Charter were articulated in 1948 with the adoption of the Universal Declaration of Human Rights, which were later codified in a series of binding international treaties and covenants setting the framework for contemporary indigenous rights as a part of the broader human rights movement (Niezen, 2003).

Binding International Law and Treaty Jurisprudence

International Treaties

In the mid-to-late twentieth century, the continued presence of indigenous peoples in international forums, coupled with the

recognition of individual and group rights as distinct from those of states, inspired emergence of a series of international organizations and binding covenants that form the basis for indigenous rights today (Henderson, 2008). The major instruments providing the backdrop for indigenous rights are the International Covenant on Economic, Social, and Cultural Rights (ICESCR) and International Covenant on Civil and Political Rights (ICCPR) passed at UN General Assembly in 1966 and the Convention on the Elimination of All Forms of Racial Discrimination (Convention) held in 1969. Although none of these instruments explicitly mentions indigenous people, they each establish a set of substantive rights as well as monitoring bodies to investigate state conduct with respect to rights guaranteed to all minority peoples. The 1989 International Labour Organization (ILO) Treaty No. 169 Concerning Indigenous and Tribal Peoples in Independent Countries (ILO 169) was the first treaty to explicitly mention indigenous rights. The primary monitoring bodies established by these treaties are the Human Rights Committee, now called the Human Rights Council (HRC), of the ICCPR and the Committee on the Elimination of Racial Discrimination (CERD) of the Convention. States were required to specifically adopt the monitoring bodies as optional protocols.

The treaties most frequently cited in indigenous rights cases are the ICCPR and ILO 169. The ICCPR provides the backdrop for indigenous rights by establishing the rights of 'peoples' in Article 1 and the rights of 'minorities' in Article 27. Article 1 of the ICCPR, which is identical to article 1 of the ICESCR, granted all peoples the right to self-determination. It holds that "[a]ll peoples have the right of self-determination … [and that] … By virtue of that right they freely determine their political status and freely pursue their economic, social and cultural development" (ICESCR, 1966: art. 1; ICESCR, 1966: art. 1). Indigenous peoples often insist that they fall under article 1 and that their right to self-determination is equal to that of all peoples and that self-determination is the basis for a myriad of correlative rights. "Self-determination is not separate from other human rights norms; rather, self-determination is a configurative principle or framework complemented by the more specific human rights norms that in their totality enjoin the governing institutional order" (Anaya, 2004b: p. 99). In other words, the right to self-determination informs all other indigenous rights.

The substantive content of the right to self-determination is contested as it pertains to indigenous peoples. Traditional interpretations argue that the right to self-determination embodies a twofold right to both 'internal' and 'external' self-determination (Anaya, 2004b). Internal self-determination refers to self-governance, or the right to practice independent political, social, and cultural rule as distinctive groups within the colonizing government. External self-determination refers to the ability to participate as a distinct sovereign amongst other governments. The question of whether indigenous peoples enjoy the right to both internal and external self-determination has inspired debate. Some settler states fear that recognizing a right to external self-determination grants a right of secession thereby threatening the fundamental international law tenet of states' territorial sovereignty (Morgan, 2004; Cassese, 1995; Coulter, 2010). Yet, from another perspective, indigenous peoples often articulate a vision of self-determination in which they have the rights to govern their members and territories, to be recognized and represented as indigenous peoples in state political processes, and to be protected from the expansion of colonist regimes over lands, political systems, and cultures (Anaya, 2009).

Whereas article 1 protects the right to internal self-determination for 'peoples,' article 27 of the ICCPR protects the rights of 'minorities' to practice their unique cultures within the political fabric of the state. Article 27 holds that "States in which ethnic, religious, or linguistic minorities exist, persons belonging to these minorities shall not be denied the right, in community with the other members of their group, to enjoy their own culture, to profess and practice their own religion, or to use their own language" (ICCPR, 1966: art. 27). Although the ICCPR does not specifically mention indigenous people, article 27 has been interpreted to protect the rights of indigenous people as minorities, to exercise their distinct cultural traditions as a group, free from discrimination by the state (see, e.g., Ominayak, 1990 *infra*). Additionally, the ICCPR and ICESCR affirm a series of human rights necessary for the equal and dignified treatment of all peoples. For example, the ICCPR affirms the right to life (ICCPR, 1966: art. 6), the right to liberty, security, and freedom of movement (ICCPR, 1966: arts. 9–12), the right to not be enslaved or subject to compulsory labor (ICCPR, 1966: art. 7), the right to equal protection before the law (ICCPR, 1966: art. 16), and the right to freedom of thought, opinion, and religion (ICCPR, 1966: arts. 18–19). The ICESCR affirms the right to work, unionize, and strike (ICESCR, 1966: arts. 6–8), the right to education (ICESCR, 1966: arts. 13–14), and the right to culture (ICESCR, 1966: art. 15).

The ICCPR, ICESCR, and the Convention inspired the UN to declare the Decade for Action to Combat Racism and Racial Discrimination in 1980, which premised the formation of the Working Group on Indigenous Populations (Working Group) in 1982. The Working Group was established by the UN Economic and Social Council for indigenous leaders and advocates from around the world to participate in yearly forums to voice the challenges facing Native peoples both as part of, and distinct from, the challenges of all minority peoples (Charters and Stavenhagen, 2009). This renewed presence of indigenous peoples at the UN resulted in the first major treaty to incorporate indigenous participation in the formation of indigenous rights.

In 1989, the ILO set important precedent for the promulgation of indigenous rights when it passed ILO 169 treaty as a revision of ILO No. 107 signed in 1975. ILO No. 107 initially addressed indigenous working conditions but failed to address issues of culture or self-determination and had a distinct assimilationist flavor. ILO 169 was the first major treaty to expressly denounce past policy in reference to indigenous rights and considers that "developments in the situation of indigenous and tribal peoples in all regions of the world, have made it appropriate to adopt new international standards on the subject with a view to removing the assimilationist orientation of the earlier standards…" (ILO 169, 1989, preamble). Today, 22 countries have ratified ILO 169, including most Central and South American countries, but excluding Canada, the United States, and most African nations. The basic philosophy of ILO 169 is to establish respect for indigenous

cultures and lifeways through the process of consultation in national decisions that affect them (Anaya, 2009).

For example, article 2 provides that governments "have the responsibility for developing, with the participation of the peoples concerned, coordinated and systematic action to protect the rights of these peoples … promoting the full realization of the social, economic, and cultural rights of these peoples with respect for their social and cultural identity, their customs and traditions and their institutions" (ILO 169, 1989: art. 2). Read with article 6, which mandates that consultation be carried out "in good faith … with the objective of achieving agreement or consent…", ILO 169 affirms the right to meaningful consultation with national governments over decisions that would affect the ability of indigenous peoples and individuals to maintain distinct cultural systems and governance (ILO 169, 1989: art. 6). Although ILO 169 clearly orders some form of consultation, and even consent in some articles, the mechanisms and extent of consultation with indigenous peoples remains largely undefined in international treaties. No state understands 'consultation with the objective of consent' to grant veto power to indigenous peoples, however, and the degree of consultation practiced by states varies.

Perhaps the most important and contentious right secured to indigenous peoples in ILO 169 is the right to land, as developed in articles 13–19. Article 14 provides that indigenous peoples retain "the rights of ownership and possession of the peoples concerned over the lands which they traditionally occupy … [and that] … [g]overnments shall take steps as necessary to identify the lands which the peoples concerned traditionally occupy, and to guarantee effective protection of their rights of ownership and possession" (ILO 169, 1989: art. 14). This article asserts that indigenous peoples have existing rights to their traditional territories, and mandates that settler states recognize and protect those rights against encroaching interests at least as well as for properties owned by non-Native citizens. Whether the right to own and occupy traditional lands extends to the natural resources above and below the surface right is unclear. At the very least, article 15 specifically secures the right to consultation over natural resource development on traditional territories. Article 15 provides "[t]he rights of the peoples concerned to the natural resources pertaining to their lands shall be specially safeguarded. … In cases in which the State retains the ownership of mineral or subsurface resources or rights to other resources pertaining to lands, governments shall establish or maintain procedures through which they shall consult these peoples…" (ILO 169, 1989: art. 15). The indigenous right to consultation over the use of natural resources on traditional lands is echoed in the right against removal or relocation. Article 16 provides that if indigenous peoples must be removed from their land, it shall only take place with "their free and informed consent … [and] … [w]here their consent cannot be obtained, such relocation shall take place only following appropriate procedures…" and with compensation in the form of equivalent lands or monies (ILO 169, 1989: art. 16).

Because of the unique relationship indigenous people maintain with their traditional territories, violations of the right to land lead to violations of many other rights affirmed in ILO 169 such as labor rights that secure equal treatment of all citizens in terms of pay and employment conditions, access to vocational training, as well as the right to health and education (ILO 169, 1989: arts. 20–32). Some scholars have been critical of how ILO 169 and other treaties elevate the relationship of indigenous peoples to their land as something that is culturally distinct and separable from broader modern society, arguing that this approach runs the risk of limiting indigenous land claims and essentializing indigenous cultures (Engle, 2009). The contentious nature of the indigenous right to land and the ambiguity of the right to consultation have spurred the majority of disputes over indigenous rights.

International Treaty Jurisprudence

Disputes over indigenous rights are reviewed and decided by the treaty monitoring bodies, especially the HRC, and increasingly by international tribunals. The initial disputes over indigenous rights proceeded under the HRC pursuant to the jurisdiction of the ICCPR and its optional protocol. In the majority of these cases, indigenous peoples challenged state or private encroachment on indigenous lands and resources under the article 1 right to self-determination and the article 27 minority rights provision, which gave rise to violations of other articles (see, e.g., *Sandra Lovelace v. Canada*, No. 24/1977, U.N. Doc. CCPR/C/OP/1 (1984), *Kitok vs. Sweden*, No. 40/1988 U.N. Doc. CCPR/A/43/40 (1988); *Bernard Ominayak, Chief of the Lubicon Lake Band vs. Canada*, No. 167/1984, U.N. Doc. CCPR/C/38/D/167/1984 (1990); *Lansman et al. vs. Finland*, No. 511/1992, U.N. Doc. CCPR/C/52/D/511/1992 (1999); *Francis Hopu and Tepoaitu Bessert vs. France*, No. 549/1993, U.N. Doc. CCPR/C/60/D/1993/Rev.1 (1997); *Mikmaq Tribal Society vs. Canada*, No. 205/1986, U.N. Doc. CCPR/C/39/D/205/1986 (1990); *Apirana Mahuika et al. v. New Zealand*, No. 547/1993, U.N. Doc. CCPR/C/70/D/547/1993 (2000)).

For example, in *Chief Bernard Ominayak and the Lubicon Lake Band vs. Canada*, the Lubicon Lake Band of Cree Indians alleged that the Government of Canada violated the Tribe's article 1 "right to self-determination and by virtue of that right to determine freely its political status and pursue its economic, social and cultural development, as well as the right to dispose freely of its natural wealth and resources and not to be deprived of its own means of subsistence," when it allowed the provincial government to lease the Tribe's traditional lands for oil, gas, and timber development (Ominayak, 1990: paras. 2.1 and 2.3). These developments precluded the Lubicon Band from practicing traditional ways of subsistence and degraded the environment to such an extent that they risked survival as a Tribe altogether. The HRC found that "the author, as an individual, could not claim under the Optional Protocol to be a victim of a violation of the right to self-determination … which deals with the rights conferred upon peoples, as such" (Ominayak, 1990: para. 13.3). However, the HRC did rule that the Canadian government violated a series of other articles that could be covered under a violation of article 27. It wrote, "[h]istorical inequities, to which the State party refers, and certain more recent developments threaten the way of life and culture of the Lubicon Lake Band, and constitute a violation of article 27 so long as they continue" (Ominayak, 1990: para. 33). This decision confirmed the indigenous right to

culture while recognizing the connection of subsistence practices on traditional lands to indigenous survival.

Contrastingly, in *Apirana Mahuika et al. v. New Zealand*, the HRC declined to find that the government of New Zealand violated article 27 when it entered into the 1992 Settlement Act (Settlement Act) with a group of Maori people to settle all fisheries disputes between the Maori and the state through a quota system. The case was brought by a second group of Maori people who claimed that the contracting group did not represent them, that they were not adequately informed nor did they participate in the negotiation, and therefore that New Zealand had violated article 27 and a series of associated rights enumerated in articles 6 to 27 which it restricted Maori fishing rights under the Settlement Act. The HRC reasoned that "the acceptability of measures that affect or interfere with the culturally significant economic activities of a minority depends on whether the members of the minority in question have had the opportunity to participate in the decision-making process in relation to these measures and whether they will continue to benefit from their traditional economy," and ruled that New Zealand did properly engage the Maori people negotiation of the Settlement Act (Apirana, 2000: para. 9.5). It cautioned, however, that New Zealand is still bound by article 27, and that "measures affecting the economic activities of Maori must be carried out in a way that the authors continue to enjoy their culture, and profess and practice their religion in a community with other members of their group" (Apirana, 2000: para. 9.9).

Some scholars and advocates argue that the HRC has proven insufficiently sympathetic to indigenous causes, criticizing the HRC's apparent reluctance to treat indigenous peoples as 'peoples' under article 1 and its tendency to find article 27 violations in only the most extreme situations such as "forcible relocation, severe environmental degradation and denial of access to subsistence areas and areas of cultural and religious significance" (MacKay, 2001: p. 13). From another perspective, however, the HRC has laid important groundwork for treating indigenous peoples as eligible for human rights under the ICCPR (Anaya, 2009).

Regional Treaty Jurisprudence

The Organization of American States (OAS) is a regional human rights regime, consisting of the United States, Canada, Mexico, the Caribbean, and all countries in Central and South America, which has adopted a series of human rights instruments that affirm norms similar to those in the UN treaties and declarations. Most notably in the indigenous context, the OAS adopted the American Declaration of the Rights and Duties of Man in 1948 and the American Convention on Human Rights (American Convention), which came into force in 1978. The American Convention has subsequently been supplemented with various additional protocols including the Additional Protocol to the American Convention on Human Rights in the Area of Economic, Social, and Cultural Rights (Protocol of San Salvador); Protocol to the American Convention on Human Rights to Abolish the Death Penalty; Inter-American Convention to Prevent and Punish Torture; Inter-American Convention on Forced Disappearance of Persons; Inter-American Convention on the Prevention, Punishment, and Eradication of Violence Against Women (Convention of Belém do Pará); and the Inter-American Convention on the Elimination of Discrimination Against Persons with Disabilities, to name a few. The OAS is also working on a draft of the proposed American Declaration on the Rights of Indigenous Peoples. Like the UN instruments, the American Convention also establishes monitoring bodies for treaty compliance by ratifying states. The American Commission on Human Rights (American Commission) initially reviews alleged disputes, and either resolves or recommends those disputes to be decided by the Inter-American Court on Human Rights (Inter-American Court). Notably, the United States has not submitted to the jurisdiction of the Inter-American Court.

In 2001 the Inter-American Court set international precedent when it upheld the indigenous right to property in *The Case of the Mayagna (Sumo) Awas Tingni Community v. Nicaragua*. In this case, the Nicaraguan government granted a logging concession to a foreign company on the traditional lands of the Indigenous Awas Tingni community. The Awas Tingi challenged the Nicaraguan government in the Inter-American Court alleging that it had violated the community's right to property and other related human rights, as well as the right to judicial protection. The Inter-American Court cited to ILO 169, the American Convention, and other Inter-American instruments of international law including the draft of the United Nations Declaration on the Rights of Indigenous People (discussed *infra*) to rule in favor of the Awas Tingni. It found that Nicaragua violated the Awas Tingni right to property and violated the community's right to judicial protection, and mandated that Nicaragua pay monetary damages. Additionally, the Inter-American Court found that "the State must adopt in its domestic law … the legislative, administrative, and any other measures necessary to create an effective mechanism for delimitation, demarcation, and titling of the property of Indigenous communities, in accordance with their customary law, values, customs and mores" (Awas Tingni, 2001: para. 143).

The ruling in *Awas Tingni* ushered in a new way of interpreting indigenous rights by considering them in light of community-specific traditions and group rights. Most notably, the Inter-American Court recognized a communal property right as derived from traditional tenure practices as equivalent to the individual title based system recognized by the state. In finding that the human right to property "embraces the communal property regimes of indigenous peoples as defined by their own customs and traditions," the Inter-American Court adopted an "evolutionary method of interpretation, taking into account normative international developments both within and outside the Inter-American system" (Anaya and Grossman, 2002: p. 12). Following *Awas Tingni*, the American Commission and the Inter-American Court have affirmed the group indigenous right to property and corollary rights such as the right to life and cultural integrity in a growing number of cases favoring indigenous plaintiffs (see, e.g., *Mary and Carrie Dann vs. United States*, Inter-Am. Ct. Hum. Rts. Report No. 75/02, 27 Dec. 2002, OEA/Ser.L/V/II.117 Doc. 5 rev. 1 (2003), the case *Maya Indigenous Communities vs. Belize*, Inter-Am. Ct. Hum. Rts. Report No. 40/04, Oct. 12, 2004; *Yaka Axa Indigenous Community vs. Paraguay*, Inter-Am. Ct. Hum. Rts. (Ser. C) No. 125 (17 June 2005); *Miowana Community vs. Suriname*,

Inter-Am. Ct. Hum. Rts. (Ser. C) No. 145 (8 Feb. 2006); *Sawhoyamaxa Indigenous Community vs. Paraguay*, Inter-Am. Ct. Hum. Rts. (Ser. C) No. 146 (29 Mar. 2006); *Saramaka vs. Suriname*, Inter-Am. Ct. Hum. Rts. (Ser. C) No. 172 (28 Nov. 2007)).

Other regional human rights bodies similar to the OAS exist in Africa, under the auspices of the Organization of African Unity (OAU) and in the African Commission (African Commission) and Court on Human and Peoples' Rights (African Court); and in Europe, as the Council of Europe and in the European Court on Human Rights. Each regional body has its own set of regional human rights treaties, judicial precedents, and protocols with varying treatment of indigenous rights as a specific area of law.

For example, in the 2010 case of *Endorois Welfare Council v. Kenya*, the African Commission found that the Kenyan government violated several rights under the African Charter on Human and Peoples' Rights when it evicted the Endorois people from their ancestral lands to create a game reserve. The African Commission cited to the AU human rights instruments, as well as precedent from the OAS, ILO 169, and the Draft United Nations Declaration on the Rights on the Rights of Indigenous Peoples (discussed *infra*) to find that Kenya had violated the Endorois' collective right to property, the right to free disposition of natural resources, the right to religion, the right to cultural life and the right to development. Additionally, the African Commission recognized Endorois ownership over their traditional land, granted unrestricted access to their land for grazing and practicing cultural and religious traditions on traditional cites, and ordered the Endorois be compensated for their losses (Endorois, 2010, recommendations). This case demonstrates the continued evolution of indigenous rights as inclusive of communal property claims and customary norms following the *Awas Tingni* case in other regional international jurisprudence.

Elsewhere around the world, human rights regimes are evolving from nascent origins. The UN Permanent Forum on indigenous peoples recognizes seven international regions (Africa; the Arctic; Asia; Eastern Europe, Russian Federation, Central Asia and Transcaucasia; Central, South America & the Caribbean; North America; and the Pacific) that each have representation in the Permanent Forum and contribute important, if still developing, work to the domestic and regional implementation of indigenous rights, respectively.

The United Nations Declaration on the Rights of Indigenous Peoples

After more than 20 years of activism and organizing by indigenous peoples and advocates, the UN General Assembly adopted the United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP) on 13 September 2007, with 143 member-states voting in favor of the UNDRIP, 11 abstentions, and four no-votes from the United States, Australia, New Zealand, and Canada. By December of 2011, all four countries voting against it in 2007 had officially endorsed the UNDRIP.

As a matter of international law, declarations such as the UNDRIP are nonbinding statements, standard-setting documents that offer guidance to states on international standards and norms. Although the UNDRIP is not a binding treaty or covenant, it can be viewed as a remedial and interpretive instrument whose effect is to make clear that the rights afforded to other individuals and peoples through treaties, customary international law, and international principles, extend to indigenous individuals and peoples as well. The UNDRIP could also be seen as an independent source of customary international law, international legal principles, or of persuasive effect due to a demonstrated consensus among the majority of state actors (Anaya, 2009). Although the specific legal effect of the UNDRIP remains unsettled, it articulates a clear set of indigenous rights to both gage and guide the treatment of Native communities by colonizing governments.

Substantive Provisions of the UNDRIP

The UNDRIP consists of 24 preamble paragraphs and 46 substantive articles affirming the equal rights of indigenous peoples to those of all peoples, while preserving the cultural diversity of native groups and enhancing "harmonious and cooperative relationships between the state and indigenous peoples based on principles of justice, democracy, respect for human rights, nondiscrimination, and good faith" (UNDRIP, 2007, preamble). Articles 1 and 2 of the UNDRIP affirm both the individual and collective right of indigenous peoples to all protections afforded in international human rights law and affirm that indigenous peoples are "free and equal to all other peoples" (UNDRIP, 2007: art. 2). Articles 1 and 2 of the UNDRIP thus affirm that indigenous peoples are 'peoples,' and thus entitled to the same rights of all peoples afforded under the ICCPR and ICESCR, and add to the remedial strength of the UNDRIP as a source of international law.

The following summarizes the UNDRIP by grouping samples of individual articles into five broad categories of indigenous rights reflected in the literature: the right to self-determination; the right to property; the right to culture; the right to health, education and work; and the right to free, informed, prior consent.

The Right to Self-Determination

Article 3 of the UNDRIP imports the exact language of article 1 of the ICCPR and ICESCR, granting indigenous peoples the right to self-determination. Various articles affirm rights to self-governance and customary decision-making, and participation in the political landscape of the state. For example, articles 4 and 5 elaborate on the right to self-determination, specifically securing the right to autonomous self-governance (UNDRIP, 2007: art. 4), and "the right to maintain and strengthen their distinct political, legal, economic, social and cultural institutions, while retaining their right to participate fully, if they so choose, in the political, economic, social and cultural life of the State" (UNDRIP, 2007: art. 5). Article 35 recognizes "the right of indigenous peoples to make their own decisions about responsibilities given to members of their community," which includes the right to determine tribal membership (UNDRIP, 2007: art. 33) and to develop their own governmental structures and institutions (UNDRIP, 2007: art. 34). Furthermore, the right to self-determination secured by the UNDRIP is likely internal and not external,

because it prohibits a right to succession. Article 46 of the UNDRIP provides that "Nothing in this Declaration may be interpreted as implying for any State, people, group or person any right to engage in any activity … which would dismember or impair, totally or in part, the territorial integrity or political unity of sovereign and independent States."

The Right to Property

The UNDRIP provides that "Indigenous peoples have the right to own, use, develop and control the lands, territories and resources that they possess by reason of traditional ownership or other traditional occupation or use, as well as those which they have otherwise acquired," and that states shall recognize and grant legal protection over those traditional lands and resources (UNDRIP, 2007: art. 26). Included in the indigenous right to property is the "right to maintain and strengthen their distinctive spiritual relationship with their traditionally owned or otherwise occupied and used lands, territories, waters and coastal seas…" (UNDRIP, 2007: art. 25), the right not be removed from their lands without their free, informed, prior consent, the right to proper compensation in the form of equivalent lands or money should they be removed from their lands (UNDRIP, 2007: art. 10; UNDRIP, 2007: art. 28), the right to be free from military activity on indigenous lands without free, informed, prior consent (UNDRIP, 2007: art. 30), the right to "the conservation and protection of the environment and the productive capacity of their lands or territories and resources" (UNDRIP, 2007: art. 29); and the right to develop their lands and resources according to local needs (UNDRIP, 2007: art. 32).

The Right to Culture

The UNDRIP affirms the indigenous right to cultural integrity free from discrimination. This includes the right to practice and protect traditional forms of knowing and being in the world. For example, article 8 affirms the right to culture and mandates that states implement domestic policy that prevents and provides redress for "(a) Any action which has the aim or effect of depriving them of their integrity as distinct peoples, or of their cultural values or ethnic identities; (b) Any action which has the aim or effect of dispossessing them of their lands, territories or resources." Included in the Indigenous Right to culture is the right to maintain ceremonial practices and protect cultural sites, artifacts, and designs (UNDRIP, 2007: arts. 11–12), maintain and control traditional knowledge, cultural expressions, and the tangible manifestations of those "including human and genetic resources, seeds, [and] medicines…" (UNDRIP, 2007: art. 31).

The Right to Health, Education & Work

The UNDRIP secures a series of rights related to the general health and individual economic and social well-being of indigenous peoples. For example, article 7 affirms that "Indigenous individuals have the rights to life, physical and mental integrity, liberty and security of person." This includes rights to education (see, e.g., Graham, 2010), health, and employment. Article 21 provides that "Indigenous peoples have the right, without discrimination, to the improvement of their economic and social conditions, including, inter alia, in the areas of education, employment, vocational training and retraining, housing, sanitation, health and social security." This includes the right to nondiscriminatory and culturally appropriate media (UNDRIP, 2007: arts. 15–16), and the right to equal and equitable employment (UNDRIP, 2007: art. 17). The UNDRIP affirms that "[p]articular attention shall be paid to the rights and special needs of indigenous elders, women, youth, children and persons with disabilities" in the implementation of each of the articles in the Declaration (UNDRIP, 2007: art. 22).

The Right to Free Informed Prior Consent

Incorporated throughout the UNDRIP is the indigenous right to free, prior, informed consent (FPIC) in decisions that may affect their sociocultural life, lands, or resources. The FPIC language appears in articles 10 (relocation), 11 (cultural property), 19 (legislative or administrative decision making), 29 (environmental protection), 30 (military activity), and 32 (development). The exact meaning, content, and processes for achieving affirmative consent is now developing in international and domestic forums.

The UNDRIP in International and Regional Human Rights Tribunals

On 24 July 2012, in *Sarayaku v. Ecuador*, the most recent case interpreting the UNDRIP after it had been adopted by the UN and endorsed by the majority of member-states, the IACHR ruled in favor of the Sarayaku people of Ecuador in a landmark decision against the state. The decision culminated a 10-year legal battle in the Inter-American system stemming from the Ecuadorian government's authorization of oil exploitation by a foreign company on Sarayaku land beginning in the 1990s. The oil development degraded Sarayaku land, and the use of explosives threatened the lives of community members. As a result, the Sarayaku alleged that the Ecuadorian government violated their right to FPIC, their right to property, their right to life and personal integrity under the ICCPR, ILO 169, the American Convention, the UNDRIP, and previous decisions of the Court including Awas Tingni.

The Inter-American Commission on Human Rights (IACHR) found that the Ecuadorian government failed to properly consult with the Sarayaku people, which led to a violation of the right to property and accompanying rights to personal life, culture, and personal integrity. It ordered $1 340 000 damages and the removal of all explosives, and that Ecuador incorporate the right to free, prior, and informed consent into domestic law. In so ruling, the court elaborated on the substantive meaning of the right to free, informed, prior consent by affirming that consultation is not a mere formality, but a true instrument of participation, "carried out in good faith, through culturally adequate procedures with the intent to arrive at an agreement … in the first phases of the development plan or investment" (Sarayaku, 2012: para. 177).

Other Instruments and Institutions of Indigenous Rights

In addition to the ICCPR, ICESCR, ILO 169, and the UNDRIP, indigenous rights have been incorporated in a variety of other

important international laws and processes involving the indigenous right to culture, and specifically the use of traditional knowledge in conservation and development.

For example, article 8(j) of the Convention on Biological Diversity states that subject to national legislation, contracting parties shall "respect, preserve and maintain knowledge, innovations and practices of indigenous and local communities embodying traditional lifestyles relevant for the conservation and sustainable use of biological diversity and promote their wider application with the approval and involvement of the holders of such knowledge, innovations and practices and encourage the equitable sharing of the benefits arising from the utilization of such knowledge innovations and practices." Similarly, the United Nations Educational, Scientific and Cultural Organization (UNESCO) has recognized the need to protect Indigenous cultures through a variety of instruments including the 2001 Declaration on Cultural Diversity, the 2003 Convention for the Safeguarding of the Intangible Cultural Heritage, and the 2005 Convention for the Protection of the Diversity of Cultural Expression. The World Intellectual Property Organization (WIPO) is currently engaging in multigovernmental negotiations to address the use and protection of indigenous traditional knowledge, cultural expressions, and genetic resources (WIPO, 2013).

As a result of indigenous advocacy, the international community has recognized that indigenous peoples are disproportionately affected by the negative effects of climate change, and uniquely situated to address mitigation and adaptation through indigenous experience and knowledge. Accordingly, the UN Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries (UN-REDD) engages local Indigenous communities in REDD project formation and implementation in participating countries around the world. Although REDD recognizes the role of indigenous peoples in climate change policy, the actual potential of REDD to incorporate indigenous rights on the ground remains unclear (Llanos and Feather, 2011; Baez, 2011).

World Bank funding for climate change strategy and other development projects similarly acknowledges indigenous rights in project management and funding criteria. For example, the 2005 World Bank operational policy 4.10 requires that a borrower set forth an action plan "[f]or all projects that are proposed for Bank financing and affect Indigenous Peoples … the Bank provides project financing only where free, prior, and informed consultation results in broad community support to the project by the affected Indigenous Peoples" (World Bank, 2005). This World Bank policy which acknowledges indigenous peoples as stakeholders distinct from the state demonstrates the increasing influence of indigenous rights, specifically the right to FPIC, in domestic and international law and policy.

Conclusion

The story of indigenous rights is unfolding before our eyes, with recent and ongoing advances in the recognition of indigenous rights as a matter of international and domestic law (Carpenter and Riley, forthcoming). Yet, as scholars have argued, it remains unclear exactly how indigenous rights will impact indigenous peoples on the ground, and in particular whether legal regimes have the capacity to bring about positive changes in the lives and wellbeing of indigenous peoples (Wainright and Bryan, 2009; Badger, 2011; Singel, 2012). Through these developments, Native peoples have demonstrated an insurmountable will to survive in the face of overwhelming pressures from states, and have begun to change the shape of international human rights law through the indigenous rights movement.

See also: Empirical Legal Studies; Indigenous Knowledge: Science and Technology Studies; Judicial Review; Natural Law; Postcolonial Law; Race and the Law; Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective.

Bibliography

Allen, S., Xanthaki, A. (Eds.), 2011. Reflections on the UN Declaration on the Rights of Indigenous Peoples. Hart Publishing, Oxford.

Anaya, S.J., Spring 2004a. International human rights and indigenous peoples: the move toward the multicultural state. Arizona Journal of International & Comparative Law 21, 13–61.

Anaya, S.J., 2004b. Indigenous Peoples in International Law. Oxford University Press, New York.

Anaya, S.J., 2009. International Human Rights and Indigenous Peoples. Aspen Publishers, New York.

Anaya, S.J., Grossman, C., Spring 2002. The case of Awas Tingni v. Nicaragua: a new step in the International Law of Indigenous Peoples. Arizona Journal of International & Comparative Law 19, 1–14.

Apirana Mahuika et al. v. New Zealand, Communication No. 547/1993, U.N.H.R.C., U.N. Doc. CCPR/C/70/D/547/1993 (2000).

Armstrong, D., 2009. Routledge Handbook of International Law. Routledge, New York.

Badger, A., 2011. Collective v. individual human rights in membership governance for indigenous peoples. American University International Law Review 26 (2), 485–514.

Baez, S., 2011. The "right" REDD framework: national laws that best protect indigenous rights in a global REDD regime. Fordham Law Review 80, 821–875.

Bruyneel, K., 2007. The Third Space of Sovereignty: The Postcolonial Politics of US-indigenous Relations. University of Minnesota Press, Minneapolis.

Carpenter, K., Riley, A., forthcoming. Indigenous peoples and the jurisgenerative moment in human rights. 102 California Law Review.

Cassese, A., 1995. Self-Determination of Peoples: A Legal Reappraisal. Grotius, Cambridge.

Charters, C., Stavenhagen, R. (Eds.), 2009. Making the Declaration Work: The United Nations Declaration on the Rights of Indigenous Peoples. IWGIA, Copenhagen.

Chiefs Bernard Ominayak and the Lubicon Lake Band vs. Canada, Communication No. 167/1984, Report of the Human Rights Committee, U.N. GAOR, 45th Sess., Suppl. 40, vol. 2. U.N. Doc. A/45/40, Annex IX (A) (1990).

Cohen, F.S., Newton, N.J., 2009. Cohen's Handbook of Federal Indian Law. Matthew Bender, Newark, NJ.

Convention on Biological Diversity, opened for signature Jun. 5, 1992, 1760 U.N.T.S. 79 (entered into force Dec. 29, 1993).

Coulter, R.T., 2010. The law of self-determination and the United Nations declaration on the rights of indigenous peoples. UCLA Journal of International Law and Foreign Affairs 15, 1–18.

Coulter, R.T., Center, I.L.R., 2002. Indigenous Peoples and the Law of Self-determination: A Possible Consensus. Indian Law Resource Center.

de Victoria, F., 1557. The First Relectio of the Reverend Father, Brother Franciscus de Victoria, On the Indians Lately Discovered, A Translation by John Pawley Bate, Appendix A in Scott, J. 2000. The Spanish Origin of International Law, Francisco de Vitoria and his Law of Nations. The Lawbook Exchange, Ltd., Union.

de Vitoria, F., 1917. Francisci de Victoria De Indis et De ivre belli relectiones. The Carnegie Institution of Washington, Washington.

Eide, A., 2009. The Indigenous Peoples, the Working Group on Indigenous Populations, and the Adoption of the UN Declaration on the Rights of Indigenous Peoples. Making the Declaration Work. The United Nations Declaration on the Rights of Indigenous Peoples. IWGIA, Copenhagen.

Endorois Welfare Council v. Kenya, 276/2003, Afr. Comm'n on Human & Peoples' Rights, Feb. 4, 2010.

Engle, K., 2009. Indigenous rights claims in international law: self-determination, culture, and development. In: Armstrong, D. (Ed.), Routledge Handbook of International Law. Routledge, Oxford, p. 331.

Graham, L., November 2010. The Right to Education and the UN. In: Declaration on the Rights of Indigenous Peoples, 2. Suffolk University Law School Research Paper No. 10–61. Available at SSRN http://ssrn.com/abstract=1701913 or http://dx.doi.org/10.2139/ssrn.1701913.

Henderson, J.Y., 2008. Indigenous Diplomacy and the Rights of Peoples: Achieving UN Recognition. Purich Publishers, Saskatoon.

International Covenant on Civil and Political Rights, opened for signature Dec. 16, 1966, 999 U.N.T.S. 171 (entered into force Mar. 23, 1976).

International Covenant on Economic, Social and Cultural Rights, opened for signature Dec. 16 1966, 993 U.N.T.S. 3 (entered into force Jan. 3, 1976).

International Labour Organization Treaty No. 169, opened for signature 27 Jun. 1989, C169 (entered into force Sept. 5, 1991).

Indigenous Peoples, Indigenous Voices. Indigenous Peoples, Indigenous Voices Fact Sheet. United Nations Permanent Forum on Indigenous Peoples. http://www.un.org/esa/socdev/unpfii/documents/5session_factsheet1.pdf.

Johnson v. McIntosh, 1823. 21 U.S. 543.

Kichwa Indigenous People of Sarayaku v. Ecuador, Inter-Am Ct. Hum. Rts. (Ser. C) No. 245, June 27, 2012.

Llanos, R.E., Feather, C., 2011. The Reality of REDD+ in Peru: Between Theory and Practice. Interethnic Association for the Development of the Peruvian Amazon (AIDESEP), Lima, p. 63.

MacKay, F., 2001. A Briefing on Indigenous Peoples Rights and the United Nations Human Rights Committee, Forest Peoples Programme. Moreton-in-Marsh, UK.

Miller, R.J., 2005. Doctrine of discovery in American Indian law. Idaho Law Review 42, 1–96.

Miller, R.J., 2011. American Indians, the Doctrine of Discovery, and Manifest Destiny. Wyoming Law Review 11, 329–349.

Miller, R.J., Ruru, J., 2008. An indigenous lens into comparative law: the Doctrine of Discovery in the United States and New Zealand. West Virginia Law Review 111, 849–917.

Morgan, R., 2004. Advancing indigenous rights at the United Nations: strategic framing and its impact on the normative development of international law. Social & Legal Studies 13 (4), 481–500.

Niezen, R., 2003. The Origins of Indigenism: Human Rights and the Politics of Identity. University of California Press, Berkeley.

Pasqualucci, J.M., Spring 2009. International indigenous land rights: a critique of the jurisprudence of the inter-American court of human rights in light of the United Nations declaration on the rights of indigenous peoples. Wisconsin International Law Journal 27, 51–98.

Robert, C.T., Spring 2010. The law of self-determination and the declaration on the rights of indigenous peoples. UCLA Journal of International Law and Foreign Affairs 15, 1–27.

Singel, W.T., 2012. Indian tribes and human rights accountability. San Diego Law Review 49, 567–624.

The Case of the Mayagna (Sumo) Awas Tingni Community v. Nicaragua, Inter-Am. Ct. Hum. Rts. (Ser. C) Case No. 79 (Aug. 31, 2001).

U.N. Declaration on the Rights of Indigenous Peoples, G.A. Res. 61/295, art. 8, U.N. Doc. A/RES/61/295 (Sept. 13, 2007).

United Nations, Charter of the United Nations, 24 October 1945, 1 UNTS XVI.

Wainright, J., Bryan, J., 2009. Cartography, territory, property: postcolonial reflections on indigenous counter-mapping in Nicaragua and Belize. Cultural Cartographies 16 (2), 153–178.

Williams, R., 1990. The American Indian in Western Legal Thought: The Discourses of Conquest. Oxford University Press, New York.

Williams, R., 1999. Linking Arms Together: American Indian Treaty Visions of Law and Peace, 1600–1800. Routledge, New York.

Williams, R., 2012. Savage Anxieties: The Invention of Western Civilization. Palgrave Macmillan, New York.

World Bank. World Bank OP 4.10-Indigenous Peoples. Retrieved from: http://web.worldbank.org/WBSITE/EXTERNAL/PROJECTS/EXTPOLICIES/EXTOPMANUAL/0,contentMDK:20553653~menuPK:4564185~pagePK:64709096~piPK:64709108~theSitePK:502184,00.html (adopted July 1 2005).

World Intellectual Property Organization (WIPO), 2013. Traditional knowledge, genetic resources and traditional cultural expressions/folklore. http://www.wipo.int/tk/en/.

Youngblood, H., 2008. Indigenous Diplomacy and the Rights of Peoples: Achieving UN Recognition. Purich Publishers, Saskatoon.

Indigenous Social Work

Michael A Hart, University of Manitoba, Winnipeg, MB, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents an overview of indigenous social work. It begins with a presentation of indigenous epistemes and helping practices as a foundation to working with indigenous peoples. It outlines the diversity of indigenous peoples and impacts of colonialism on the people. It provides an overview of indigenous social work and closes with an example of a particular indigenous approach.

Introduction

Professional social work has been around for a century, while indigenous social work is a relatively new development. Partially emerging in response to the concerns of how social work reflects nonindigenous ideas, perspectives, and practices, it has developed with a strong connection to indigenous epistemes and ways of being and indigenous helping practices (Hart, 2002; Sinclair et al., 2009). Some challenges that have emerged during its development have been the continued impact of colonialism that indigenous peoples face. This has not only resulted in a diversity of indigenous perspectives but also influenced social work practice by creating a need for diverse approaches. All these points contribute to the explanation of indigenous social work. The following sections provide an overview of indigenous social work, beginning with an overview of indigenous epistemes and ways of being that are a foundation for indigenous social work. Then, helping practices of indigenous peoples generally to provide some comparative context for indigenous social work is outlined. The impact of colonialism on indigenous peoples is reviewed as an additional factor influencing the context of indigenous social work. The last half of this article focuses on an overview of indigenous social work. This overview includes a detailed example of one approach within indigenous social work, namely, the *mino-pimatisiwin* approach (Hart, 2002, 2008, 2009). This approach relies on indigenous knowledge about helping stemming from indigenous peoples in the central area of North America. It has the foundational concepts of focusing on the whole, relationships, balance, harmony, healing, growth, and *mino-pimatisiwin* (a Cree term literally meaning good life). Following indigenous ways of being, it has particular views of people, helping processes, and helping techniques.

Indigenous Epistemes and Ways of Being

Episteme is understood in a broad sense to mean the taken-for-granted presuppositions that are the primary socializing force and act as the foundation to our values, attitudes, and perceptions of the world (Koukkanen, 2007). Indigenous peoples' epistemes and ways of being in the world share some commonalities. These commonalities include the centrality of spirit, connection to land and other life, the recognition of cycles and flux, the importance of elders, communalism, egalitarianism, respectful individualism, and ceremonies (Couture, 2011; Hart, 2009; Koukkanen, 2007; Little Bear, 2000). These are important to recognize because these provide the foundations of indigenous helping theories and practices and are key to developing an understanding of how to work with indigenous peoples when one functions from alternate epistemes like those rooted in European ways of thinking and being.

Centrality of Spirit

As explained by Mark and Lyons (2010), "spirituality is defined variously as existential reality, connectedness, and energy, or as a dimension of a person that is concerned with the transcendent and life's meaning" (p. 1756). Spirit in many indigenous peoples' cultures plays a prominent role (Hart, 2008, 2009; Mark and Lyons, 2010). This prominence is evident by the significant attention paid to the metaphysical world in everyday activities alongside having specific events and/or activities that guide indigenous people's interactions with the spirit realm.

Holistic Perspective: Connection to Land and Other Life

Indigenous epistemes are not the bioreductionist perspectives of modernist societies. They are all-encompassing because attention is given to the mental, physical, emotional, and spiritual aspects of people in relation to the lands and life on the lands where the people are located (Hart, 2002). This holistic perspective directs the actions of indigenous people whose efforts are focused on being in relation with other life as opposed to exercising control over other life. Well-being is then reflected in how people give attention to all aspects within themselves, as well as in how people interconnect to the land and life around them (Hart, 2002; Mark and Lyons, 2010).

Recognition of Cycles and Flux

Many indigenous languages are verb based (Little Bear, 2000). Working from this base, indigenous peoples focus on actions and inactions and patterns of these actions and inactions. The patterns are studied to recognize the cycles that are evident, e.g., cycles within individuals, between individuals, and in nature generally. Examples of such cycles include the times of high and low energy a person regularly experiences each day, the communication patterns that two individuals

share, and impact of the moon or seasons on individuals. These cycles are also looked at in relation to one another. The observations are made during various lengths of time ranging from short time spans to observations made over generations. When significant changes occur in these patterns, observations continue to see if the patterns have changed permanently or whether they are a continuation of larger ebbs and flows. All observations are made to determine how any, or all, individuals fit with the patterns and how to adjust to the fluxes that occur. The concept of flux is grounded within indigenous perspectives that see the world and life as in constant motion and changing (Little Bear, 2012). As the world and life are constantly changing – in flux, individuals are constantly adjusting to these changes. This focus emphasizes the beliefs that people are part of the world and not controllers of it, and that they do best by fitting in with the cycles and adjusting to the fluxes.

The Importance of Elders

Elders are recognized as the individuals who consciously carry indigenous knowledge for their peoples and who demonstrated their knowledge by how they live. They are seen as the living embodiment of indigenous cultures and knowledge (Couture, 2011). They play a primary role in these knowledge systems, particularly since they are the link in the intergeneration development of indigenous knowledge. This key role ensures that they are regularly turned to for guidance, support, and understanding.

Community, Egalitarianism, and Respectful Individualism

Indigenous societies rest on a strong sense of community where groups of individuals are interdependent. Within these societies, individuals are considered participants of the group's activities. Such participation includes the commitment to support other members while recognizing that they will receive support when needed. This form of egalitarianism emphasizes less hierarchical forms of organization and supports respectful individualism (Gross, 2003; Hart, 2008, 2009). Respectful individualism gives individuals significant freedom in self-developing their personal gifts under a shared understanding that they are committed to contributing to group well-being.

Ceremonies

To fit in the patterns of life, including personal development of one's gifts, indigenous peoples have developed particular processes referred to as ceremonies. These ceremonies regularly require multigenerational participation, and necessitate the knowledge of elders and other informed individuals on how to carry out the ceremonies. Ceremonies then become part of the group's life patterns that connect them to the land and other life entities surrounding them. These play a central role in maintaining and reestablishing wellness.

Values

Certain values emerge from indigenous epistemes, including a heightened and expanded understanding of respect, an emphasis on sharing, and important social responsibilities (Hart, 2002, 2008, 2009; Mafile'o, 2008; Weaver, 2008). Respect is about treating others with dignity, esteem, and in the manner that they wish to be treated, or do unto others, as others want done unto them. It requires people to learn about the other person. Respect stems from understanding that all beings have spirit, and that all spirit is to be respected. To be disrespectful means to go against a fundamental aspect of life. Hence, all beings are to be respected since no one species of life is held as more important than another. For indigenous people, respect is reflected in practices such as how individuals treat one another, plants, the raising and gathering of crops, and animals that are herded or hunted. It is also reflected in the ceremonies that give positive attention to all forms of life.

Regarding people, a great deal of respect is given to children by providing them with lots of freedom to develop and learn through modelling processes, and to older people who have knowledge and experience. All genders are well respected in recognition that no gender is subservient to another and that all people play a role in the well-being of the community. Respect is important for peoples' well-being.

Another important value within indigenous epistemes is sharing (Hart, 2002). It is a fundamental aspect of indigenous cultures, and directly reflects the importance of relationships. Sharing is the distribution of what one has in their possession with others. It helps people maintain relationships and goodwill and is indicative of the concepts of community and interconnection. It is so pervasive in some indigenous cultures that a wide variety of things are shared, including land, food, supplies, ideas, ceremonies, and songs. Sharing runs counter to notions of the 'survival of fittest' since life is not about competition where the mightiest wins, but about providing to others through commitment and even sacrifice.

A third value worthy of note is social responsibility (Weaver, 2008). In direct relation to respectful individualism and communalism, indigenous people provide significant space for individuals to develop their own gifts with the understanding that as these develop they will be contributing back to the community. This understanding is driven by the value of social responsibility where people hold a personal sense of commitment and accountability to their families, communities, and all life. This responsibility is a key link in the peoples' interconnections to one another and to all life and is central to indigenous helping practices.

Helping Practices within Indigenous Cultures

Indigenous peoples always had their own ways of helping. These practices were and are manifestations of indigenous epistemes and ways of being in and experiencing the world. They are not based on the theoretical premises of many European/American theories that emphasize individualism, nuclear families, the Cartesian split of the mind and body, and objective distancing of the helper (Lettenberger-Klein et al., 2013; Little Bear, 2012; Liu, 2011). Nor are these based on the avoidance of spiritual considerations that drives positivistic theories and empiricism followed by nonindigenous professional helpers.

Traditional indigenous helping processes embed the spirit conceptually (Hart, 2002; Mark and Lyons, 2010; Weaver, 2008)

in interconnections between the individual and other life. These processes consider how an individual, family, or community fits within the cycles around them, and how they may have worked against or disconnected from these cycles. Ceremonies are relied upon to facilitate the reconnection of individuals, families, and communities to these cycles. Traditional indigenous helping processes recognize that changes to the cycles can impact individuals, families, or communities and that helping processes can assist people in readjusting to these changes. Furthermore, many trickster stories in indigenous cultures indicate that these processes acknowledge that there are forces that can cause significant fluctuations in the cycles, seriously impacting upon people and requiring intervention. These interventions, whether for supporting individuals to reestablish themselves with the cycles around them or addressing destructive forces, often involve ceremonies. These ceremonies can range from being focused narrowly on the individual and/or issue to a wider scope of incorporating whole communities as an indirect means of addressing the issue. Most often, these ceremonies address people's physical aspects while giving significant consideration to the emotional, mental, and spiritual dimensions of participants and healers, ensuring that both are closely involved in ceremonial processes. These indigenous ways of helping differ considerably from practices stemming from European epistemes. These differences were and continue to form the bases of the colonial oppression of indigenous peoples.

Colonial Impacts

Colonialism has greatly impacted upon indigenous peoples and their helping practices and requires recognition. In many parts of the world, indigenous peoples are oppressed and prevented from living in ways that emerge from their epistemes and cultural practices. This oppression varies throughout the world, ranging from forms that marginalize indigenous voices and practices to outright violent persecution of indigenous peoples.

Theorists indicate that oppression has disconnected indigenous peoples from their own epistemes and ways of experiencing and being in the world in varying degrees (Akena, 2012; De Souza and Rymaez, 2007; Fanon, 2008/1951; Hart, 2002; Memmi, 1991/1965; Simpson, 2008). Some indigenous peoples have maintained a strong connection with their epistemes to remain the dominant driver in how their peoples move forward in their daily lives. These individuals include elders who maintain knowledge of the traditional ceremonial practices of their nation (Couture, 2011). Others have internalized cultures stemming from oppressive forces to the extent of incorporating various dominant epistemes and practices. For another segment, this incorporation has come by force and impacted upon them to the degree that they have internalized oppression against their own people and now actively work against ways of living that reflect indigenous epistemes. An example of internalized oppression is demonstrated through indigenous gangs that attempt to rule their own communities through fear and violence (Comack et al., 2013). One result of imposed colonialism has been increased diversity within indigenous communities, spanning from individuals maintaining a way of being based in their indigenous episteme to individuals who separate themselves from indigenous ways of being.

Indigenous Diversity

Indigenous peoples' experiences with colonial oppression have created significant diversity between and within indigenous cultures. This diversity can be seen along one axis stemming from deeply immersed within indigenous cultures to deeply immersed within colonial cultures. This axis would have a split on one end with those consciously choosing not to identify with the colonizing culture on one side and those who have unconsciously internalized their oppression on the other. This diversity is further complicated in that colonizing cultures are heterogeneous, e.g., significant diversity among French, Spanish, Dutch, English, and American cultures.

Indigenous peoples' cultures and practices also have to be recognized as significantly diverse on other grounds. Recalling that a key aspect of indigenous epistemes is their connection to land of their territories, indigenous peoples' ideas and cultural practices will reflect diversity in the land. There are approximately 370 million indigenous peoples in 90 countries across the world (United Nations Department of Economic and Social Affairs, 2009); there is considerable diversity in indigenous peoples' cultures, languages, ideas, and practices.

With such great diversity between indigenous cultures and peoples' experiences of colonialism, talking about indigenous cultural helping practices with a presumption of these applying to indigenous peoples throughout the globe is problematic and challenging. Nonindigenous people have to maintain constant awareness of this, remembering that any work with indigenous peoples requires learning from them directly to address their diversity. This point is central to indigenous social work.

Indigenous Social Work

Indigenous social work stems firmly from indigenous epistemes and ways of life and includes a crucial stance against colonialism that is informed by indigenous epistemes, the diversity of indigenous nations, indigenous knowledge stemming from these nations, and critical perspectives that address colonial oppression. Indigenous social work acknowledges the more common epistemic elements of indigenous cultures. Hence, it tends to be holistic with strong recognition of the relationships between individuals, families, communities, and the land. It recognizes interdependence between people, and peoples and other life. It looks at these relationships for patterns, cycles, and disruptions to patterns and cycles. It is based in concepts of community, respectful individuals where individuals are seen as inherently good and oriented to developing their gifts and capable of contributing to their communities. These matters are examined all at the same time to reassure the well-being of all community members as well as the living environment around them.

Being based in indigenous epistemes requires indigenous social work to reflect not only the common elements among indigenous epistemes (noted above) but also the perspectives of particular indigenous communities. In other words, indigenous

social work is global in scope as it relates to many indigenous nations, but is local in focus as it directly reflects the communities from which it stems. Thus, it avoids a 'pan-indigenous' stance that sees and treats all indigenous peoples as stereotypically the same (Baikie, 2009). Instead, it stems from and reflects the people and lands being served. This ability allows indigenous social work to pull teachings from various indigenous nations and have them considered by the people being served for their appropriateness and/or adaptation. Significant in these processes is the recognition that the power to determine the fit of a practice lies with indigenous peoples.

Indigenous social work practice recognizes diversity and must accommodate this diversity in practice. Thus, there is no one practice, theory, or approach that covers indigenous social work. This diversity has been reflected upon in various ways (Baikie, 2009; Hart, 2003). Practices must cover a range of approaches that may incorporate other indigenous forms, producing social work practices based in indigenous cultures, but resonating across many indigenous cultures. For example, Green and Baldry (2008) have outlined key indigenous concepts relevant to many indigenous peoples throughout Australia. These include collective self-determination, interdependence, reciprocity, obligation, land, and family. There are indigenous social work practices that while specific to particular indigenous culture(s) have been transferred to nonindigenous cultures. An example is the Maori constructs that have been adopted by European (Pakeha) social work professionals in Aotearoa/New Zealand (Munford and Sauders, 2011). These concepts include *homai ki te tangta, te kanohi kit e kanohi* – give it to a person face to face, *mamaakitanga* – hospitality with genuine care and support, and *whanaungatanga* – understanding the responsibility of belonging to a *whanau*/family. Some indigenous helping practices require greater skill development than that provided in mainstream social work educational systems. An example includes the use of the sweat lodge in social work (Schiff and Pelech, 2007; Voss et al., 1999). These practices are specific to a particular culture and rest in the hands of indigenous healers or elders. They may be used by social workers trained in these practices, but more often they are practices that social workers need to be aware of to make appropriate referrals.

Indigenous social work relies on indigenous knowledge that rests in keepers such as elders and traditional healers. Hence, people working from indigenous social work stances seek their guidance, support, and teachings. These individuals and teachings may be incorporated into formal education, but are more likely additional resources that an individual would have to learn to access. As indigenous knowledge has an element of subjectivity and is localized, indigenous social work includes the necessity to develop long-term relationships with local elders and healers. It also requires a basic level of learning and understanding about indigenous helping knowledge. Overall, indigenous social work avoids the subjugation and marginalization of indigenous knowledge and recognizes these and their holders alongside any other professionals and their education. Indigenous knowledge processes are also valued in indigenous social work (Baikie, 2009), recognizing that social work practices can be supported through inductive means of learning where people focus on the immediate context and slowly broaden their knowledge base with each circumstance they encounter. This process of developing theory from practice values pragmatic, tacit, and/or culturally implicit knowledge.

Additionally, indigenous knowledge and indigenous social work incorporate critical theories that address oppression, inequities, the influence of social structures and policies on individuals, and social justice. In particular, indigenous social work encompasses literature and practices on colonialism, anticolonialism, and decolonization (Hart, 2009). In some circumstances, material stemming from postcolonial literature is considered as well, although there are limitations to this literature for practice applications in social work (Hart, 2007). A related area of theoretical knowledge stemming primarily from academic sources is that of racism. Racism is a central component to maintaining colonial oppression (Memmi, 1991/1965) and evident in cultural, structural, and personal realms (Dominelli, 1988; Mullaly, 2007). By addressing colonialism and racism, indigenous social work recognizes that indigenous peoples are caught in relationships that distribute power in unequal and harmful ways. Thus, indigenous social work pays attention to critical issues including the distribution of power, privilege, and relationships between indigenous and nonindigenous peoples in colonial processes. By addressing these relationships and power distributions, matters of rights, including indigenous rights, human rights, and ecological justice, are given considerable attention.

From this foundation, many approaches and theories have emerged. Baikie (2009) has theorized indigenous-centered social work using postcolonial and anticolonial thoughts. Mafile'o (2008) has outlined an approach to Tongan social work practice. McKenzie and Morrissette (2003) have addressed a way to approach social work with aboriginal peoples in Canada. Weaver (2008) has written on indigenous social work in the United States. Kee (2008) has outlined the development of culturally appropriate social work practice in Malaysia, making indigenous social work vast and developing. These various approaches and theories have been progressing for differing lengths of time and to various degrees, but much of it has emerged in the past two decades.

An Example of Indigenous Social Work Based within Indigenous Epistemes

One approach that has been developed over two decades is the *mino-pimatisiwin* approach initially developed in 1997 and later (Hart, 1997, 2002, 2006, 2008, 2009). At its commencement, there were no indigenous approaches to social work relevant to practice with indigenous peoples in the central part of North America in the literature. This aimed to reflect the epistemes and helping practices of indigenous peoples, particularly the Anishinaabe, Cree, Dakota, Dene, and Oji-Cree peoples, who reside in the central part of Ininiwi-Ministik (North America). Hence, it developed from the perspectives of indigenous peoples in this territory to support social workers to practice in a culturally relevant way. It has several key concepts and values, and has a particular orientation to viewing people and helping processes.

Three aspects of indigenous epistemes are the primary foundation of the *mino-pimatisiwin* approach. First, it has the

central pillar of spirituality where *manito*, spirit, is recognized as a positive force that is found within, and shared between all life entities. Second, the *mino-pimatisiwin* approach rests on the stance of respectful individualism where people are given extensive room and freedom for inner exploration and self-development. Third, it rests on a sense of community that is tied together through biological and adapted family relations, and a commitment to the community by family members as demonstrated by their sharing, support, and interest in the well-being of one another.

The key concepts of the *mino-pimatisiwin* approach are holism, relationships, balance, harmony, growth, healing, and *mino-pimatisiwin*. Holism is a recognition that when working with an individual, attention must be given to all aspects of that individual, including their emotional, mental, spiritual, and physical dimensions, including the many aspects around that person, e.g., their family, communities, nation, and the natural and built environment. This approach holds relationships as a key concept where all elements within individual and those aspects around them are understood as being connected to one another. These connections are to be supported and nurtured since positive connections are viewed as enhancing health and well-being. In particular, individuals are to maintain or bring balance to their life by giving attention to the four fundamental aspects of people: their physical, mental, emotional, and spiritual selves. Further, individuals are understood to be in the contexts of their families, communities, nations, and larger environment. When people are focused on maintaining balance and honoring relationships between the various aspects within and around them, and when each element is fulfilling its obligation in relation to all others, harmony is the result. This harmony is the ideal condition for growth where growth is a daily and lifelong process of developing all aspects of one's self. Healing is similar to growth in that it is orientated to the restoration of relationship, balance, and harmony, where disconnection, imbalance, and disharmony emphasize the need for attention. Ultimately, the intent is to strive for *mino-pimatisiwin*, the good life where people are living as fully as possible with positive, balanced, and harmonious interconnections among all aspects of their lives. While *mino-pimatisiwin* is the aim, and on occasions people feel they have reached it successfully, it is not a permanent state. It is a lifelong goal as life is constantly changing.

This approach acknowledges that values are frequently highlighted as guides for various indigenous peoples. While many values have been identified, two values are highlighted in this approach, namely, respect and sharing. Respect is a universal value among diverse cultures. When this value is followed within the context of respectful individualism and community, it takes the form of honor, esteem, deference, and courtesy to all aspects of life, and particularly of the person being helped and the relationship being developed with the helpers. It also includes refraining from imposing one's views onto others whenever possible. Thus, the focus is on supporting the individuals through their own self-discovery of the issue(s) at hand. Respect is necessary for maintaining positive and healthy relationships, including that between the helper and the person seeking help.

Sharing is strongly emphasized in this approach. Such deep sharing helps to build and maintain positive relationships and demonstrates that everyone is important. The role of the helper includes sharing knowledge, understanding, guidance, support, and encouragement. It is also important that helpers recognize, respect, and acknowledge that the person who they are helping is also sharing, including their time, experiences, trust, faith, and life.

The *mino-pimatisiwin* approach has a specific orientation toward people. The basic nature of people is seen as good, but recognizes that people can learn negative attributes that focus on selfishness, breaking relationships, and hurting life. The *mino-pimatisiwin* approach sees people in a state of being, where being is understood as the here-and-now relationship with that which surrounds them. People are also seen as being in a state of being-in-becoming, which emphasizes people identifying their sense of purpose as including *mino-pimatisiwin*, and making efforts to strive toward that sense of purpose. From the perspective of this approach, people are seen as dependent on those who have lived before them and on other life since people are social beings. This perspective implies that people's histories are as important as their connections with others. To support these relationships, people are guided by good conduct that is learned through such practices as participation in ceremonies, the guidance of elders, and deep, honest self-reflection.

The *mino-pimatisiwin*'s orientation to helping rests heavily on indigenous epistemes. Recognizing the importance of relationship, this approach focuses on the relationships of the person seeking help, including the relationship between the helper and the person seeking help. While helpers may have experience and knowledge about relationship dynamics, they are not considered experts. The subjective experiences of the person seeking help remain the paramount source of information and understanding. Thus, helpers focus on establishing and strengthening the relationships between themselves and the people they are serving. By highlighting the positives within their relationships, helpers are able to extract practical examples that would strengthen the abilities of the persons seeking help. Further, by developing this relationship, helpers will be provided with a deeper sense of the experiences of the people seeking help so as to positively mirror these back to them for reflection and consideration. To facilitate this development of and strengthen this relationship, helpers speak from the heart, meaning not only with their 'professional' knowledge but also with a holistic sense of their emotional experiences, present feelings, 'gut reactions,' and intuition. While not dismissing evidence-based practice or similar stances, working from a *mino-pimatisiwin* approach requires practice-based evidence where helpers utilize their experiences that may arise holistically through their cognitive understandings, feelings, and physical awareness and reactions. Finally, it is highly relevant that helpers recognize that the helping process is not a one-way experience. By honestly recognizing that they will be impacted, grow, and even develop as people, helpers can begin to honor the relationship between the person seeking help and themselves and recognize the gifts that they are receiving.

To approach helping in this manner, helpers are expected to address their own growth and healing. Recognizing that growth and healing are lifelong activities as people seek *mino-pimatisiwin*, helpers regularly reflect on their own lives,

how their practices are a continued reflection of their life experiences and learning, and how they are impacting upon relationships that they have with the people they serve. From this approach, there is an understanding that helpers are striving to live the life that they wish for others and that they are incorporating the key concepts and values of this approach within their own lives. Helpers are expected to have excellent listening skills, and be deeply empathetic, accepting of silence and long pauses. Such skills are not only appropriate to this approach but also serve as foundations for good practice and communication skills in social work generally. Furthermore, social workers utilizing this approach should be able to utilize skillfully nondirect means of working, such as the use of personal stories, role modeling, and practice examples. Other techniques include eliciting the support and direction of elders, incorporating humor, facilitating connections and incorporating culturally based rituals and ceremonies in their practice and as a source of additional services. These rituals and ceremonies include such activities as sharing circles, family group conferencing, and sweat lodges.

Conclusions

Indigenous social work is a broad concept reflecting the diversity of indigenous nations and peoples, as well as their experiences with nonindigenous peoples in social work. Despite this diversity, there are some aspects of indigenous epistemes and ways of being that are common to many indigenous peoples. These aspects not only serve as a foundation for cultural helping practices within indigenous cultures but also provide a unifying element to indigenous social work across the globe. They also speak to nonindigenous peoples who want to work in ways reflective of indigenous perspectives, such as antioppressive and strength-based approaches (Dominelli, 2004; Mullaly, 2010). While these common elements demonstrate the global nature of indigenous social work, it is important to recognize its local focus where the practices, values, and epistemes of specific indigenous peoples are supported by indigenous social work. Recognition of this diversity among indigenous peoples in relationship with one another as they work to strengthen their peoples and cultures epitomizes indigenous social work.

See also: Ethnicity, Race, Oppression, and Social Work: The Canadian Case; Indigenous Child Welfare Practice: The Canadian Case; M?ori Social Work; Orientalism, Oppression, and Contemporary Social Work; Social Work Education; Social Work Theory.

Bibliography

Akena, F.A., 2012. Critical analysis of the production of western knowledge and its implications for indigenous knowledge and decolonization. Journal of Black Studies 43 (6), 599–619.

Baikie, G., 2009. Indigenous centre social work: theorizing a social work way of being. In: Sinclair, R., Hart, M.A., Bruyere, G. (Eds.), Wicihitowin: Aboriginal Social Work in Canada. Fernwood Publishing, Winnipeg, pp. 42–61.

Comack, E., Deane, L., Morrissette, L., Silver, J., 2013. Indians Wear Red: Colonialism Resistance and Aboriginal Street Gangs. Fernwood Publishing, Winnipeg.

Couture, J.E., 2011. The role of native elders: emergent issues. In: Long, D.A., Dickason, O.P. (Eds.), Visions of the Heart: Canadian Aboriginal Issues, third ed. Oxford University Press, Don Mills, pp. 18–34.

De Souza, M., Rymarz, R., 2007. The role of cultural and spiritual expressions in affirming a sense of self, place, and purpose among young urban, Indigenous Australians. International Journal of Children's Spirituality 12 (3), 277–288.

Dominelli, L., 1988. Anti-Racist Social Work. Macmillan, London.

Dominelli, L., 2004. Social Work: Theory and Practice in a Changing World. Polity Press, Cambridge.

Fanon, F., 2008/1951. Black Skin, White Mask. Grove Press, New York.

Green, S., Baldry, E., 2008. Building Indigenous Australian social work. Australian Social Work 61 (4), 389–402.

Gross, L.W., 2003. Cultural sovereignty and Native American hermeneutics in the interpretation of the sacred stories of the Anishinaabe. Wicazo Sa Review 18 (3), 127–134.

Hart, M.A., 1997. An Ethnographic Study of Sharing Circles as a Culturally Appropriate Practice Approach with Aboriginal People. M.S.W. University of Manitoba, Winnipeg, MB.

Hart, M.A., 2002. Seeking Mino-Pimatiswin: An Aboriginal Approach to Helping. Fernwood Publishing, Halifax.

Hart, M.A., 2003. Am I a social missionary? Reflections of a Cree social worker. Native Social Work Journal 5 (November), 299–313.

Hart, M.A., 2006. An aboriginal approach to social work practice. In: Heinonen, T., Spearman, L. (Eds.), Social Work Practice: Problem Solving and Beyond. Nelson, Toronto, pp. 235–260.

Hart, M.A., 2007. Cree Ways of Helping: An Indigenist Research Project (Ph.D. thesis). University of Manitoba, Winnipeg, MB.

Hart, M.A., 2008. Critical reflection on an aboriginal approach to helping. In: Gray, M., Coates, J., Yellow Bird, M. (Eds.), Indigenous Social Work Around the World: Towards Culturally Relevant Education and Practice. Ashgate Press, Aldershot, pp. 129–139.

Hart, M.A., 2009. Anti-colonial indigenist social work: reflections on an aboriginal approach. In: Sinclair, R., Hart, M.A., Bruyere, G. (Eds.), Wicihitowin: Aboriginal Social Work in Canada. Fernwood Publishing, Winnipeg, pp. 25–41.

Kee, L.H., 2008. The development of culturally appropriate social work practice in Sarawak, Malaysia. In: Gray, M., Coates, J., Yellow Bird, M. (Eds.), Indigenous Social Work around the World: Towards Culturally Relevant Education and Practice. Ashgate Press, Aldershot, pp. 97–106.

Koukkanen, R., 2007. Reshaping the University: Responsibility, Indigenous Epistemes, and the Logic of the Gift. UBC Press, Vancouver.

Lettenberger-Klein, C.G., Fish, J.N., Hecker, L.L., 2013. Cultural competence when working with American Indian populations: a couple and family therapist perspective. The American Journal of Family Therapy 41 (2), 148–159.

Little Bear, L., 2000. Jagged worldviews collide. In: Battiste, M. (Ed.), Reclaiming Indigenous Voice and Vision. UBC Press, Vancouver, pp. 77–85.

Little Bear, L., 2012. Traditional knowledge and the humanities: a perspective by a Blackfoot. Journal of Chinese Philosophy 39 (4), 518–527.

Liu, J.H., 2011. On the limited foundations of western skepticism towards indigenous psychological thinking: pragmatics, politics, and philosophy of indigenous psychology. Social Epistemology 25 (2), 133–140.

Mafile'o, T., 2008. Tongan social work practice. In: Gray, M., Coates, J., Yellow Bird, M. (Eds.), Indigenous Social around the World: Towards Culturally Relevant Education and Practice. Ashgate Press, Aldershot, pp. 251–282.

Mark, G.T., Lyons, A.C., 2010. Maori healers' views on wellbeing: the importance of mind, body, spirit, family, and land. Social Science & Medicine 70 (11), 1756–1764.

McKenzie, B., Morrisette, V., 2003. Social work practice with Canadians of aboriginal background: guidelines for respectful social work. In: Al-Krenawi, A., Graham, J.R. (Eds.), Multicultural Social Work in Canada: Working with Diverse Ethno-Racial Communities. Oxford University Press, Don Mills, ON, pp. 251–282.

Memmi, A., 1991/1965. The Colonizer and the Colonized. Beacon Press, Boston.

Mullaly, B., 2007. The New Structural Social Work. Oxford University Press, Don Mills.

Mullaly, B., 2010. Challenging Oppression and Confronting Privilege, second ed. Oxford University Press, Don Mills.

Munford, R., Sanders, J., 2011. Embracing the diversity of practice: indigenous knowledge and mainstream social work practice. Journal of Social Work Practice 25 (1), 63–77.

Schiff, J.W., Pelech, W., 2007. The sweat lodge ceremony for spiritual healing. Journal of Religion & Spirituality in Social Work: Social Thought 26 (4), 71–93.

Simpson, L., 2008. Our elder brothers: the lifeblood of resurgence. In: Simpson, L. (Ed.), Lighting the Eighth Fire: The Liberation, Resurgence, and Protection of Indigenous Nations. Arbeiter Ring Publishing, Winnipeg, pp. 73–88.

Sinclair, R., Hart, M.A., Bruyere, G., 2009. Wicihitowin: Aboriginal Social Work in Canada. Fernwood Publishing, Halifax.

United Nations Department of Economic and Social Affairs, 2009. The State of the World's Indigenous Peoples. United Nations, New York. http://www.un.org/esa/socdev/unpfii/documents/SOWIP_web.pdf (accessed 15.11.12.).

Voss, R.W., Douville, V., Soldier, A.L., Twiss, G., 1999. Tribal and Shamanic-based social work practice: a Lakota perspective. Social Work 44 (3), 228–241.

Weaver, H.N., 2008. Indigenous social in the United States: reflections on Indian tacos, Trojan horses, and canoes filled with indigenous revolutionaries. In: Gray, M., Coates, J., Yellow Bird, M. (Eds.), Indigenous Social around the World: Towards Culturally Relevant Education and Practice. Ashgate Press, Aldershot, pp. 71–82.

Individualism versus Collectivism: Philosophical Aspects

Philip Pettit, Princeton University, Princeton, NJ, USA and Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Individualism as a doctrine in social ontology – in the theory of what sorts of entities are important in the social world – is usually contrasted with collectivism. Collectivism comes in two varieties, hard and soft. The hard variety of collectivism argues that social-structural laws and forces, as revealed in social science, undermine the image of the human being as an autonomous agent. Individualism denies that this is so, while admitting that agency is subject to various limitations and failures. The softer variety of collectivism argues only that social science shows that human agency is not quite as central as common sense may suggest, illustrating the point in a variety of different theses. This sort of collectivism need not be rejected by individualists.

The issue between individualism and collectivism, as taken in this article, is the ontological issue as to whether the regularities of social life, in particular the regularities that have come to light in social science, cast doubt on the status of human beings as more or less autonomous intentional subjects. This issue should be distinguished from the methodological issue as to whether the explanation of social phenomena ought to invoke only the properties of individuals, as it should be distinguished from the moral issue as to whether social institutions should always be assessed from the point of view of their impact on individuals. Émile Durkheim, the great French sociologist, was the most influential defender of a hard form of collectivism in its ontological sense. He holds or at least suggests that individuals, as seen from the perspective of social science, lack the autonomy that they ascribe to themselves in common sense. Just as astronomy taught us that the sun does not literally cross the sky, and physics that solid objects are not solid in any literal sense, so the collectivist idea is that the social sciences teach us that individuals are not what they seem; they are not the self-determining centers of decision making that they take themselves to be. Hard collectivism in this sense looks implausible, as we shall see; only some soft versions of the doctrine – versions that sideline or downgrade intentional psychology without undermining it – look to have any serious support (Pettit, 1993).

The Emergence of Collectivism

The sociological way of thinking that brought Durkheim to this conclusion had its origins in the rise of social statistics in the early part of the nineteenth century. Responsive to electoral and utilitarian pressures, the governments of the newly industrializing countries sponsored the statistical mapping of a great variety of phenomena. "It was necessary to count men and women and to measure not so much their happiness as their unhappiness: their morality; their criminality; their prostitution; their divorces; their hygiene; their rate of conviction in the courts. With the advent of laws of statistics one would find the laws of love or, if not that, at least the regularities about deviancy" (Hacking, 1981, 1990).

This statistical mapping of large-scale society, this avalanche of numbers, had an important intellectual impact, because it revealed that social aggregates display surprising constancies, constancies that are often invariant over changes in the individuals, and in the mentalities of the individuals, involved. The statisticians constructed indices of divorce, suicide, crime, sickness, prostitution, and a variety of other phenomena. It soon became clear from their work that there were constancies over time in many such indices, and there were constancies in the correlation of different indices with one another.

At the beginning of the nineteenth century it would have been natural to expect aggregate statistics to be as variable as individual preference and choice. The revelation of unexpected pattern encouraged a sort of statistical determinism, associated in the beginning with French writers like AM Guerry and Adolphe Quetelet. Already by 1832, Guerry could write: "We are forced to recognise that the facts of the moral order are subject, like those of the physical order, to invariable laws" (Hacking, 1990). Such determinism came to be described as an astronomical conception of society, a conception in which people were driven by forces akin to a cosmic force like gravity; the fact that Quetelet was an astronomer lent credence to this characterization of the view (Hacking, 1990). The astronomical conception soon spread abroad, receiving a powerful impetus in the English-speaking world, and indeed more broadly, with the publication in 1857 of T. H. Buckle's *History of Civilisation in England*.

The astronomical conception of society received perhaps its most dramatic expression in Tolstoy's *War and Peace*, which was written between 1863 and 1869. One passage will serve to give the tenor of the view "Ever since the first person said and proved that the number of births or crimes is subject to mathematical laws, that certain geographical and politico-economical laws determine this or that form of government, that certain relations of the population to the soil lead to migrations of peoples – from that moment the foundations on which history was built were destroyed in their essence. By disproving those new laws, the old view of history might have been retained. But without disproving them, it would seem impossible to continue studying historical events, merely as the arbitrary product of the freewill of individual men. For if a certain type of government is established, or a certain

movement of peoples takes place in consequence of certain geographical, ethnographical, or economic conditions, the freewill of those persons who are described to us as setting up that type of government or leading that movement cannot be regarded as the cause" (Tolstoy, 1972).

The avalanche of numbers and the statistical determinism associated with it led to the development of the new science of sociology. True, the person who invented the term 'sociology', Auguste Comte, was averse to seeing the discipline as a merely statistical enterprise. However, the rise of statistical mapping had a crucial impact on Émile Durkheim, who was responsible for establishing sociology as a respectable science. The first empirically based study that he undertook was an investigation of suicide – published in 1897 – and this reveals both how deeply he was impressed by the statistical constancies of suicide and how naturally he moved toward the conclusion that those constancies were independent of the whirl and flow of individual choice. "Each society is predisposed," he remarked, "to contribute a definite quota of voluntary deaths" (Durkheim, 1951).

Durkheim accepted that aggregate constancies have a life of their own, remaining surprisingly invariant over relevant changes at the individual, intentional level. He postulated aggregate, social regularities to explain those constancies. An aggregate regularity may explain a constancy by subsuming it, as when a constancy in the correlation between two indices – say, the divorce rate and the suicide rate – is taken to reveal a regularity. An aggregate regularity may explain a distinct constancy, revealing a relationship between it and some independent factor, say, a relationship between an aggregate index or correlation of indices on the one side and a background factor like increased population density, an established current of opinion, growing urbanization, or even the fact that the factor in question – say, the correlation – is socially functional in some way. Durkheim gives expression to the sorts of regularities that he saw at the origin of statistical constancies, when he writes in the study on suicide: "the relations of suicide to certain states of social environment are as direct and constant as its relations to facts of a biological and physical character were seen to be uncertain and ambiguous. Here at last we are face to face with real laws" (Durkheim, 1951).

In order to understand the sorts of regularities that Durkheim envisaged, here are some examples. They are not chosen on the grounds that Durkheim actually endorsed them or even considered them – he did not – but because they represent a fair spread of the sort of sociological doctrine with which he was associated.

1. Increased unemployment leads to a rise in crime.
2. Urbanization leads to a decline in religious practice.
3. Policies for increasing employment cause inflation.
4. Economic stratification comes about in order to secure personnel for crucial positions.
5. Capitalism is stable when it is optimal for developing productive power.
6. States act internationally in their best economic interest.
7. States decline in political influence as they fall behind in economic capacity.
8. Companies maximize expected returns.
9. The protestant ethic facilitates the rise of capitalism.
10. Capitalism generates a breakdown in community values.

Regularities of these kinds are social-structural in character. They are social regularities, because they each demand that a number of individuals display certain intentional responses: they require that individuals commit crime in a certain proportion, give up attending church, and so on. They are structural, because they are discontinuous with intentional regularities. It is not supposed that unemployment leads to crime, or urbanization to religious decline, because it causally impacts in some intentionally intelligible way on the antecedents of intentional regularities; it is not supposed, for example, that it is because people become aware of unemployment or urbanization that they produce the behavior required for the consequent.

Durkheim and most of those who followed him took the sorts of regularities illustrated to give the lie to intentional psychology. He saw them as pointing us toward constraints that undermine or limit human autonomy. He saw them as vindicating, in essence, something like the so-called astronomical conception of society. The constraining factors to which social-structural regularities pointed, in Durkheim's view, range from the morphological features of a society, like the density of its population, to the norms or rules institutionalized there, to the currents of opinion and the enthusiasms that take over from time to time (Durkheim, 1938; Lukes, 1973). He described all these as social facts and argued that social facts constrain intentional agents from without. "Currents of opinion, with an intensity varying according to time and place, impel certain groups either to more marriages, for example, or to more suicides, or to a higher or lower birth rate, etc. A social fact is to be recognised by the power of external coercion which it exercises or is capable of exercising over individuals" (Durkheim, 1938).

In giving social facts this constraining role, Durkheim was explicitly challenging the individualist view – a view he naturally presents as an extreme position – that individuals enjoy intentional autonomy. Thus, he wrote of social facts that: "when we define them with this word 'constraint,' we risk shocking the zealous partisans of absolute individualism. For those who profess the complete autonomy of the individual, man's dignity is diminished whenever he is made to feel that he is not completely self-determinant" (Durkheim, 1938).

The Implausibility of Hard Collectivism

There are two different ways in which social-structural regularities might prove to compromise the intentional autonomy of individuals in the sort of measure that Durkheim apparently had in mind: in the measure, as we may take it, that would give the lie to individualism and support a position truly deserving of the title of hard collectivism. Individuals will be autonomous in the relevant sense so far as they form their beliefs and desires – their intentional attitudes – and act on those beliefs and desires in a manner that satisfies the general expectations that we hold in relation to a rational or approximately rational agent. One way in which social regularities would compromise individual autonomy is by overriding those intentional regularities, the other by outflanking them (Pettit, 1993).

The social regularities would override the intentional if they sometimes conflicted with them, predicting different

outcomes for the same situation, and if they proved in those cases to be the more reliable laws. The social regularities would outflank the intentional if it transpired that things are organized at the intentional level – intentional agents and attitudes are selected – so that the social regularities are assured of being satisfied. In the overriding case, the social regularities would predetermine what happened; in the outflanking case, they would predestine it.

The social sciences have not produced results robust enough to support collectivism in the literal or hard sense associated with the overriding or outflanking thesis. Hume argued that the testimonial evidence that a natural miracle occurred – that a law of nature was suspended – is almost always going to be outweighed by the day-to-day evidence of the reliability of the laws of nature. We might argue in parallel that the social scientific evidence in favor of an intentional miracle – in favor of the overriding of intentional regularities in otherwise normal agents – is almost always going to be outweighed by the day-to-day evidence of the reliability of intentional laws.

As it is with the overriding thesis, so it is with the outflanking one. This thesis would hold that in cases where human beings act so as to ensure that a social regularity obtains, that is no accident: were they going to have attitudes that would lead them not to support that regularity, then they would not have been there; they would not have been selected. The case envisaged parallels that of natural selection. Assume that mammals have been selected so as to favor the regularity whereby they can digest certain foods. That means if there were certain creatures going to have a system that did not support that regularity, then they and their ancestors would not have been there; they would not have been selected. While there is evidence of certain forms of social selection, there is no evidence for the powerful sort of selection – parallel to natural selection – that would be needed to support the outflanking alleged.

It ought not to cause any embarrassment for social scientists, however, that the overriding and outflanking theses – and the hard collectivism they would entail – are not borne out empirically. For if we recur to the examples given earlier, we can see that social laws of the sort that impressed Durkheim can be interesting and significant, quite consistently with not entailing hard collectivism. That those laws obtain is fixed in part by the fact that human beings are intentional systems who obey standard intentional laws, and in part by contingent facts about the attitudes of those individuals. That does not mean that just knowing about individuals and their intentional psychology will give us knowledge of those laws, not at all. It may take considerable data collection and data analysis before it begins to become plausible, say, that increased unemployment leads to a rise in crime or increased urbanization to a decline in religious practice. Such a regularity will be plausible in good part, because we can see how it fits with our knowledge of intentional psychology: we can see, for example, how unemployment might lead to crime by increasing both the opportunity and the incentive for criminal acts. However, that does not mean that we can derive the law from armchair reflection on our psychology.

Social laws not only will be interesting in this sense, even if they do not entail hard collectivism, but can also be of explanatory importance. Explanatory or methodological individualists hold that explanations that go to the level of individual psychology are to be preferred in social science to higher level, structural or aggregate explanations. Jon Elster assumes that just the fact of getting nearer the detail of production makes such an explanation intrinsically superior: "a more detailed explanation is also an end in itself" (Elster, 1985). Michael Taylor (1988) agrees: "A good explanation should be, amongst other things, as fine-grained as possible." However, the rejection of hard collectivism does not necessarily entail the adoption of such explanatory individualism.

Assume that explanation usually involves giving information on causal process. The reason that social laws are explanatorily important is that reference to such a law may give information on the causal process at its origin that is not available just from the story about the detailed psychological profiles of the individuals involved. Thus, we might be told of the individual psychological etiologies that produced the level of crime in a given year, and be told that that level is higher than earlier, without being given the information provided by the law, as we may assume it to be, which relates increased unemployment to increased crime. Indeed, we might not even have noticed that most of the new or more frequent offenders were unemployed. The reference to that law provides the information that the increased crime would materialize, not only in the presence of the precise psychological profiles of actual offenders but also under any condition where there is an increase of unemployment. It servesas an explanatory purpose that is neglected in the more detailed, individual-level stories (Jackson and Pettit, 1992; Pettit, 1993).

The Plausibility of Soft Collectivism

There is also another line of thought that ought to reconcile many social scientists to the failure of hard collectivism. This is that the argument just given is consistent with the truth of some claims that have a collectivistic ring but that do not involve either the overriding or the outflanking of intentional law. Any one of these claims might be thought to support a soft collectivism: that is, a position that is collectivistic so far as it makes intentional psychology somewhat less important than it may have seemed but that is consistent, nevertheless, with the soundness of intentional psychology. I shall mention six such claims, many of which are quite plausible theses. The claims address a somewhat different question from that to which hard collectivism provides an answer.

The first question that is important to distinguish from the hard collectivism issue is the question that divides those who are well described as atomists and holists: the question as to whether the individual agent noncausally depends on relations with other people for the appearance of distinctive human capacities (Pettit, 1993). Holists maintain that individual people are dependent on one another for the possession of distinctive human capacities: say, for the capacity to think. However, this belief need not lead them to hold that people cannot therefore have the autonomy of subjects who act more or less rationally in the light of more or less rationally held beliefs and desires. Intentional psychology is silent on the question of whether a capacity like that of thought requires communal resources. The holist assertion that thought does

require those resources cannot, therefore, undermine that psychology.

The second issue that is distinct from the issue that divides individualists and hard collectivists is whether groups can ever have a more or less autonomous status as corporate agents or legal persons or whether 'singularists' are right to say that the only real agents are individuals. The question is not whether any groups can perform as if they were individual agents or persons, which corporations and churches and states clearly can. Rather it is whether they have such a unity as agents that in practice the only way to understand and interact with them properly is to abstract from their particular members, with their particular attitudes, and treat them as agents with minds of their own: agents with judgments and preferences that form and update, and ultimately shape behavior, on a pattern is not discernible just on the basis of the corresponding attitudes of their members.

There have been thinkers who defend the reality of groups on in a collectivist manner, postulating the appearance of a group mind or spirit that displaces the minds of individual members (Runciman, 1997). However, a commitment to group agents of the kind described need not involve collectivism. A recent theory supports such a commitment, not on collectivistic grounds, but on grounds related to problems in the aggregation of individual attitudes into group attitudes (List and Pettit, 2011). The reason why group agents enjoy a certain autonomy, on this approach, is that it is effectively impossible for a group to be sure of displaying a sensitivity to rational constraints like consistency – a precondition for performing like an individual person – while allowing its particular attitudes to be determined in systematic response to the corresponding attitudes of individuals. This shows that it is possible to reject a singularism about group agents, as it is possible to reject atomism about people in society, without rejecting individualism.

The third question that must be distinguished from the hard collectivism issue has come to the fore in recent thought. Let us agree that people intentionally or nonintentionally produce – and reproduce – a variety of social entities by the patterns manifested in their attitudes and actions. They constitute social groups and instrumentalities, and they constitute the social properties that attach to certain individuals: the status enjoyed by this person, the power enjoyed by that. Here now is a question raised by such entities. Do the entities exercise a reciprocal influence on individual agents? Do they serve to determine, and indeed rationally determine, the opportunities, the motives and the perceptions of those agents, and thereby shape their further intentional responses? There are some who think, or who think certain individualists have to think, that the answer to this question is, no (Ruben, 1985). Those who defend an affirmative answer, therefore, will appear in the guise of collectivists. This means, in effect, that most contemporary social theorists will be given a collectivist cast (Giddens, 1976; Coleman, 1990).

The affirmative answer to this question is surely the correct one; social entities exercise a reciprocal influence on agents, at least so far as agents come to form beliefs and desires that involve those entities in their contents. Admitting the reciprocal influence of social entities in this manner does not in any way threaten the standing of our intentional psychology and does not involve a commitment to hard collectivism. I may be influenced by the objects in my social world, as I am influenced by the objects in my physical milieu, coming to form beliefs and desires in respect of them, but this does nothing to compromise my capacity as an intentional subject.

A fourth proposition that is collectivistic in spirit but not damaging to intentional psychology is a thesis of revisability. It is the thesis that the regularities identified in social science often challenge intentional psychology in the sense of pointing us toward unsuspected disturbing factors. Like the holistic and reciprocity doctrines, this revisability thesis is perfectly plausible. Consider the sort of regularity that much of Goffman's work is given to supporting: that people, in his language, are extremely loath to break frame, that they readily establish a set of shared assumptions about what is say-able and do-able in any interaction, and that they generally conform to those assumptions, even at the cost of considerable frustration of their other desires or ideals (Goffman, 1975). Anyone who has read the work of Goffman is likely to conclude, quite reasonably, that here we have been alerted to a factor that is often going to disturb people's intentional performance. The work of such social scientists bears out the truth of the revisability thesis ascribed to collectivists, that the identification of social regularities is likely to point us to conditions that disturb the operation or revelation of people's intentional processes, and in that sense that it is likely to affect our intentional psychology.

Although the revisability thesis is plausible, it does not serve to displace intentional psychology in a collectivist manner. The frame law used in illustration leaves intentional psychology fundamentally intact. Intentional psychology is an open-ended practice; it remains possible at any stage for its practitioners – who are of course also its subjects – to revise their conception of favorable conditions. If social science can interact with intentional psychology to force such revisions, in particular revisions that can be internalized by participants in their practice, then that does not involve any displacement of intentional psychology; on the contrary, it serves as a way of enriching the practice. Consider the effect of the discovery in optics that a rotating disc with a certain black–white pattern looks to be colored in different ways at different speeds. This does not undermine the intentional-psychological story of color perception but simply points us to a perturbing factor – rotation – that generally goes unnoticed but that is of a kind with factors – say, to do with lighting conditions – that are commonly recognized. A similar gloss will apply to the deliverances of social science, under the revisability thesis.

So much for the antiatomism and antisingularism, the reciprocity, and the revisability claims, which should be distinguished from hard collectivism. The fifth claim that should equally be distinguished from it is a thesis about the dispensability of individual contributions to social life, the sixth a claim about the inevitability of those contributions.

The dispensability thesis would say that while any socially significant event is going to come about because of the actions of this or that individual or set of individuals, still no individual makes an indispensable contribution. Had that individual not existed or not acted appropriately, still the phenomenon in question would have occurred. It is either the case (1) that the individual's contribution was not a necessary part of the total cause of the phenomenon or (2) that in the absence of that individual some other agent would have taken its place.

Durkheim was motivated by the existence of certain social constancies – the suicide rate, the divorce rate, and the like – to defend a thesis of type (1). The fact that these constancies survive variation in the individuals and the mentalities of the individuals involved shows that no one individual contribution is a necessary part of the relevant total cause. In defending this sort of dispensability claim, Durkheim was doing nothing to displace intentional psychology, so that the claim could be defended equally well by an individualist. The social constancies in which he was interested are characterized at such a high level that the absence of one individual contribution would not make a difference to the obtaining of the constancy. An artifact of aggregation and abstraction accounts for the dispensability of individual contributions, not any radical questioning of intentional psychology.

What of the other sort of dispensability thesis: the type (2) thesis that even with social phenomena that do actually depend on particular individuals – say, something like a political revolution – that phenomenon would still have occurred even in the absence of a relevant contributing party; it would still have occurred, because there would have been someone else available to take the vacated place? This type of claim is not associated with Durkheim particularly but is to be found in other, more historically oriented traditions of social science. The main source is the Marxist tradition and here Engels is quoted as making a telling remark about Napoleon Bonaparte "That such a man, and precisely this man, arises at a determined period and in a given country is naturally pure chance. But, lacking Napoleon, another man would have filled his place" (Sartre, 1963). Is this sort of dispensability thesis inimical to our intentional psychology? I think that it is not inimical, although I doubt that it is generally true (James, 1984). All that is required for it to be the case that an individual's contribution is dispensable in this sense is that a number of individuals have the capacity and motivation required to make the sort of contribution involved but that at most one individual can have the opportunity. To think that this was true of all individually necessary contributions to socially significant phenomena might be dogmatic, but it would not involve rejecting intentional psychology.

The sixth and last sort of thesis that should be distinguished from hard collectivism fits better with less empirical parts of Durkheim's work than his study of suicide: for example, with his analysis of the division of labor characteristic of our sort of society. I describe it as an inevitability thesis, as distinct from a dispensability thesis. What it says is that various features of social life limit the opportunities available to individual agents – they are structural constraints on feasible optio.ns – in such a way that it is inevitable that agents will act so as to sustain certain social constancies: "In an extreme version this would mean that the constraints jointly have the effect of cutting down the feasible set to a single point; in a weaker and more plausible version that the constraints define a set which is so small that the formal freedom of choice within the set does not amount to much" (Elster, 1979). This could be what collectivists have in mind, at least some of the time, when they denigrate the place of the intentional subject. Certainly it is a thesis that many social scientists defend.

The inevitability thesis, even in the weaker version, is generally unsound. However, that is beside the point in our present discussion. More important, it does no more than the dispensability thesis to undermine intentional psychology. It just might be the case, as the inevitability thesis has it, that our opportunities are tailored to the production of certain social constancies: that we are so limited in our range of options that it is inevitable we will act so as to sustain such constancies. Even if it were true, that would cast no serious doubt on the soundness of intentional psychology or on our status as autonomous subjects. It would not mean that we individual subjects were compromised by social laws in a manner that undermined our view of ourselves as intentional subjects. Our intentional psychology represents us as more or less rational subjects but it says nothing on the range of opportunities that are actually available to us as agents.

It may come as a surprise that hard collectivism is not entailed by the inevitability thesis or the dispensability thesis. Many self-described individualists would certainly want to deny the general truth of such propositions. Even if that is so, the matters in question do not go to the heart of the issue between individualism and collectivism: they do not have an intimate bearing on the status of our intentional psychology. They support a soft sort of collectivism but not the hard kind that would provide a really significant challenge to our self-image.

See also: Atomism and Holism: Philosophical Aspects; Causation: Physical, Mental, and Social; Causes and Laws: Philosophical Aspects; Collective Beliefs, Sociological Explanation of; Collectivism and Individualism: Cultural and Psychological Concerns; Colonialism, Anthropology of; Economics, Philosophy of; Individuals in Society: History of the Concept; Methodological Individualism in Sociology; Methodological Individualism: Philosophical Aspects.

Bibliography

Coleman, J.S., 1990. Foundations of Social Theory. Belknap Press of Harvard University Press, Cambridge, MA.
Durkheim, É., 1938. The Rules of Sociological Method. Free Press, New York.
Durkheim, É., 1951. Suicide. Free Press of Glencoe, New York.
Elster, J., 1979. Ulysses and the Sirens. Cambridge University Press, Cambridge, UK.
Elster, J., 1985. Making Sense of Marx. Cambridge University Press, Cambridge, UK.
Giddens, A., 1976. New Rules of Sociological Method. Hutchinson, London.
Goffman, E., 1975. Frame Analysis: An Essay on the Organization of Experience. Penguin, Harmondsworth, UK.
Hacking, I., 1981. The taming of chance by an army of statistics. The Times Higher Education Supplement 12–13.
Hacking, I., 1990. The Taming of Chance. Cambridge University Press, Cambridge, UK.
Jackson, F., Pettit, P., 1992. In defence of explanatory ecumenism. Economics and Philosophy 8, 1–21.
James, S., 1984. The Content of Social Explanation. Cambridge University Press, Cambridge, MA.
List, C., Pettit, P., 2011. Group Agency: The Possibility, Design and Status of Corporate Agents. Oxford University Press, Oxford.
Lukes, S., 1973. Émile Durkheim: His Life and Work: A Historical and Critical Study. Penguin, Harmondsworth, UK.
Pettit, P., 1993. The Common Mind: An Essay on Psychology, Society and Politics. Oxford University Press, New York.
Ruben, D.-H., 1985. The Metaphysics of the Social World. Routledge & Kegan Paul, London.
Runciman, D., 1997. Pluralism and the Personality of the State. Cambridge University Press, Cambridge.
Sartre, J.-P., 1963. The Problem of Method. Methuen, London.
Taylor, M., 1988. Rationality and Collective Action. Rationality and Revolution. Cambridge University Press, Cambridge, UK.
Tolstoy, L., 1972. War and Peace. Heinemann, London.

Individuals in Society: History of the Concept

Ron Harré, University of Oxford, Oxford, UK; and Georgetown University, Washington, DC, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7306–7310, © 2001, Elsevier Ltd.

Abstract

Philosophical dimensions of the distinction between individuals and collectives include the problem of universals (how are similarities among members of collectives to be explained?), of emergent properties (do collectives have properties that are different from those of their members?), and of internal and external relations (are the properties of individuals brought into being by membership of collectives?). In political philosophy the problem of the authority of the state as a collective led to accounts by Hobbes, Locke, and Rousseau in terms of a (mythical) moment of a social contract between members to create a sovereign. By contrast Vico saw the state as a development of the existing social structures of families. Utilitarians assumed, problematically, that properties of groups of people were merely aggregates of attributes of individuals. Mill defended a radical distinction between individual and collective as the ground of political freedom. Hegel and Marx made collectives fundamental. Individuals acquired different attributes by virtues of membership of social aggregates which changed as the history of humanity unfolded. Durkheim too assumed that there were social facts not reducible to attributes of individuals. Sociology, as the study of collectives, can focus on attributes of individual members (idiography) or attributes of groups (nomothetic research). The debates around ontological and methodological individualism continue still.

The concepts 'individual' and 'society' are intimately related. Societies, whatever else they may have been thought to be, are collections or groups of individuals. Many philosophers and psychologists have thought that individuals only took on the characters they did by virtue of the influence of the society in which they grew up. But which had priority? Which was the cause and which the effect? The problem of how the concepts of individual and society are related has been part of a more extensive philosophical debate on the nature of classes and their relations with their own individual members. Some have argued that we can dispense with class concepts in favor of ones referring only to individuals, others that any such attempt will necessarily fail.

Philosophical Aspects of the Individual/Society Contrast

Kinds of Groups

The problem of how to interpret the concepts of individual person and social collective is embedded in a more ample philosophical issue, the general question of how to understand the contrast between individuals and groups and the interrelations between them. There are two main kinds of collectives or groups that are of relevance to social science. There are taxonomic groups in which the members share some set of qualifying properties but may have no real relations with one another. Two people may be members of an ethnic minority solely on the basis of their mother tongue, or their religious affiliation. This concept of a collective is particularly appropriate to the study of groupings like social classes. Then there are structured groups like families or institutions, the members of which may be biologically or socially related to one another. In structured groups members are usually different from one another with respect to group membership. Mum and Dad are members of the nuclear family just insofar as one is female and the other male. In recent sociology and social psychology such groups have been analyzed by the use of the concept of role. It seems particularly appropriate to the study of institutions and teams.

The Related Problem of Universals

Philosophical problems associated with taxonomic groups have a history that reaches back through Medieval attempts to deal with the problem of universals into antiquity and the Platonic forms. How could one account for the fact that many individual entities were the same color? Or the same shape? According to the Ancient Greek philosophers this fact had to be explained by some version of the form and matter distinction. Clearly similar things shared a common form. How did that form exist? Plato built his account on the principle that the property displayed by each individual entity was a realization in the material world of a form that transcended it. The Platonic account has often been rejected. In the later Middle Ages, William of Ockham argued that the only common element in a group of things of the same color, for example, is the use of the same *word* to describe each of them.

Emergent Properties

The idea that collectives are somehow more than the sum of their members has persisted, albeit in rather different ways, down to the present day. The notion of an 'emergent property' has been coined to describe cases in which a collective of parts, say a brain composed of neurons, has a property, a capacity for thought, that is not possessed by any neuron taken singly. Emergent properties contrast with aggregative properties such as weight. Both brains and their component cells have weight. An important debate in social science centers on the issue of whether among the emergent properties of social structures are causal powers. Some have argued that agency is an attribute only of the human beings who create these structures.

Internal and External Relations

The third main issue concerning the relation between individuals and collectives is tied in with the distinction between internal and external relations. Two beings are internally related if their natures are partly or wholly determined by their standing in those relations. Marriage is an internal relation in that a person becomes a husband or a wife by virtue of being related by marriage to someone else. Spatial relations are commonly cited as examples of external relations between people, though they may not be for material things in general. The general question is whether people are the products of the societies in which they have a place or are the societies' products of the activities of people? Contemporary structuration theory takes a middle way, seeing people as shaped by their social relations and in turn, in their subsequent ways of living, recreating those social relations, albeit imperfectly.

Hobbes and Locke: The Genesis of the Sovereign

Leviathan: Individuals Covenant to Maintain a Sovereign

In Protestant Europe the individualism of the renaissance soon evolved into the idea of the state as a corporation of citizens. What sort of collective could such a corporation be? Where is the sovereign to be found in such a picture of the state? From what source does the state derive authority, when the Divine Right of monarchs is no longer accepted? The overt philosophical problem, which is to ask in what way a corporation exists other than as an aggregate of its members, was displaced by the more pressing question of how a group could come to have an identity to which a citizen owed a duty of obedience. The social contract, an agreement to give up some of one's individual freedoms for the benefits of civil society, in particular personal security, was invoked by Hobbes, Locke, and later Rousseau as a necessary precondition for a social order to arise out of individual aspirations and intentions.

This was the moment at which the collective emerged from the aggregate of individuals. It was also the ground of the state's authority. Both aspects of the social contract appear very clearly in Hobbes's *Leviathan* of 1651.

> [the agreement that creates civil government] is more than Consent or Concord; it is a real Unity of them all, in one and the same Person [the Sovereign], made by Covenant of every man with every man ... [This creates a Commonwealth, the essence of which is] One Person, of whose Acts a great Multitude, by mutual Covenants one with another, have made themselves everyone the Author, to the end he may use the strength and means of them all, as he shall think expedient, for their Peace and Common Defence. (Hobbes 1651, Part 2, Chap. 17)

Yet it is not the collective that is sovereign, since the purpose of the creation of the unity is the delegation of powers to an individual. Each person is the 'author' of the acts of whosoever is the sovereign.

The General Will

Rousseau, writing a century later, dispensed with the sovereign as such. Here we have the submergence of the individual in the collective.

> Each one of us puts into the community his person and all his powers under the supreme direction of the general will; and as a body we incorporate every member as an indivisible part of the whole. (Rousseau 1764 [1968], Book I, Chap. 6)

Nevertheless he insisted that the legislative role of the social group was necessarily distinct from the power to execute the general will.

The metaphysical question of the kind of reality that the collective or group might have, by way of contrast with that of the members taken individually, is not directly addressed by these authors. For instance no one in the high era of the social contract paused to ask for the criteria of identity and individuality of a society, nation, or institution. Nor is the issue of the status of emergent properties raised so far as I know. It is certainly implicit in Rousseau's version of the social contract in that the 'general will' of the collective is explicitly distinguished from the aggregate of personal wishes and intentions, the 'will of all.'

Vico and the Role of Unintended Consequences

The question of how a group of individuals can come to have a collective identity troubled Giambattista Vico (1668–1774). Unlike his distinguished predecessor, Machiavelli, he did not see the collective as the product of the power and will of the sovereign. He remarks that

> It is true that men have themselves made this world of nations …. But this world, without doubt, has issued from a mind often diverse, at times quite contrary, and always superior to the particular ends that men had proposed to themselves; which narrow ends, made means to serve wider ends, it has always employed to preserve the human race upon this earth. (Vico 1725, p. 1108)

How is this possible? How can one account for the transition from anarchy to the institutions of the state? It cannot be by the accidental coordination of individual motivated acts. The superior 'mind' Providence employs individual intentions to bring about social order.

Providence is not God, but should be interpreted, I believe, as a metaphor for the flux of unintended consequences of individual acts. There is no need for a social contract to bring the conditions of order into being in Vico's view. Social order comes to be because the conditions for its emergence are already in existence.

> Yet chance did not divert them nor fate force them out of this natural order. For at the point when the commonwealths were to spring forth, the matter was all prepared and ready to receive the form, and there came forth the commonwealths composed of body and mind. The prepared materials were these men's own religions, their own languages, their own lands, their own nuptuals, their own names, their own arms ... and because these things were their own they were completely free and therefore constitutive of true commonwealths …. The fathers at this juncture, by uniting themselves in an order, came to produce the sovereign civil power, just as in the state of nature they had held the family powers …. (Vico 1725, p. 630)

By bringing about the birth of families, each with its own language and rules of life, Providence had at the same time brought into being the natural law to be used in the management of commonwealths.

The identity of a nation was identical with the identity of a family, only writ large. Vico's collectives were structural not taxonomic.

The Problematic Status of Collectives

Vico's account of the genesis of the state neatly avoids the issue of how individuals can jointly constitute a collective entity the properties of which are not just aggregates of the properties of individuals. He denies individual intentions an explicit role and at the same time traces the means of maintaining institutions back to simpler institutions in which individual intentions directly create social order. In essence Vico assumes the native character of institutions. They are just there.

Utilitarians and Collectivists

Individuals and Freedom

The Utilitarians, Jeremy Bentham (1789), and the Mills, father and son, had no difficulty with the concept of collectives of human beings, for rather the opposite reason from Vico. For them the properties of groups were just aggregates of the attributes of the individual members. The happiness of all was the arithmetical sum of the happinesses of each. The Utilitarian account of the relation between individual and collective is sometimes used to illustrate the fallacy of composition. To infer from the fact that each of the 20 soldiers in a platoon is six feet in height that the platoon is 120 feet high is an obvious fallacy. Is it not equally fallacious to give a measure of the happiness of a society as the sum of the happinesses of individual members?

John Stuart Mill's (1859) *On Liberty* presupposes a sharp distinction between individual and society, at least when 'society' takes the form of government, public opinion and so on. He argued for the fundamental moral propriety of resisting those who claim to be the voice of the mass. Freedom and individuality are more or less the same thing for Mill. There is a contrast for him between the self-subsisting power of the individual to resist and the power of the state to coerce.

We would now tackle the problem of picking out which properties could be aggregated and which could not by invoking the idea of an emergent property, and using examples of the difference between the properties of constituent cells and the organs of which they are components to illustrate the point. What could one make of the idea of a group happiness as an emergent property of the individual activities of the members of a society?

Collectives and Mass Societies

Running more or less parallel to Mill's individualistic utilitarianism, Hegelian idealism presented a very different picture. History is a collective progress since the world spirit, *Geist*, is embodied in the human race as material beings, more or less as individual minds are embodied in individual bodies. All relations between all beings whatsoever are internal and constitutive. As the realization of the one World Spirit human beings have a unity as a kind of supra-individual.

Marx and Durkheim

Marx's sociology, the structure of which is drawn more or less directly from Hegel's metaphysics, pitted class against class in a dialectic of contradictory interests (Marx 1867 [1965]). The this-ness of individuals gave way to the individual class member as relevant to sociology only as the vehicle for the necessary attributes of class membership. Borrowing from Hegelian Idealism, Marxian 'Man' was the nexus of internal relations, the relations of the social necessities of certain means of the production of the material necessities of life. However, Marx's two main social classes were taxonomic, in the sense I have defined above. To be a member of the working class or of the bourgeoisie was to exemplify a certain set of attributes, though those attributes were possessed by virtue of standing in certain internal relations to other people.

Durkheim's plea for sociology as the science of 'social facts' comes close to Marx's unpsychological sociology insofar as Durkheim's social facts are attributes of groups. In his *Elementary Forms of the Religious Life* (Durkheim 1895 [1982]), the ceremonies of the Native Australians are to be interpreted as presentations of collective facts about the local social order through participation in which loyalty and commitment to that order are renewed. Again the collective is thought of in aggregational rather than structural terms.

What Sort of Science Is Sociology?

Windelband and the Nomothetic/Idiographic Distinction

The individual/collective issue appears not only in metaphysics and social theory, but in methodology as well. That there are two kinds of human sciences was the claim of Windelband (1894 [1998]). There are studies of individuals as such, an idiographic science. There are studies of people in groups or collectives, a nomothetic science. The analogue of a law of nature had no place in idiographic science while, perhaps with the help of statistical analysis, it could have a place in nomothetic science. This distinction is related to but not identical to that between two kinds of empirical enquiry. There is the intensive design, the study of a group by the study of a typical member and there is the extensive design, the aggregation of information about the members of a population into a statement about the population as a whole. The results of studies based on the extensive design will tend to ignore emergent properties, since they are not aggregates of the properties of individuals.

The Reduction of Collectives to Individuals

In the twentieth century the issue has surfaced again, this time in the guise of the problem of explanation. Are the collective aspects of human societies to be accounted for as just patterns of the unintended consequences of individual actions according to individual intentions? Or are they mass social effects of mass social causes? If we take the former view, explanations of mass

phenomena must be sought not in other mass phenomena, but in the intentions of individual actors as they ramify through a population and interact with the intentions of other actors. This proposal has been called 'methodological individualism.' It has been revived by Popper (1961) and others. The idea is an old one. Adam Smith (1776) thought that individual choices in the marketplace 'added up' to the illusion of the 'hidden hand.' It had an important place in Menger's (1883 [1981]) advocacy of microeconomics, the study of the grounds of individual acts of choice as opposed to statistical analyses of vast masses of data. It cannot be a wholly individualistic story since some reference to structure seems to be essential to understand how individual actions could aggregate into something that could be perceived as orderly.

Adam Smith (1776) and Carl Menger (1883 [1981]) did not feel the need to offer arguments for methodological individualism, since each thought he had shown its methodological advantages. In the 1950s a more recognizably philosophical debate began, forcing the positions to be more clearly differentiated and the arguments' pro and con more carefully formulated. Three lines of argument in favor of collectivism can be discerned. A metaphysical argument emerged to the effect that individuals are simply products of social processes (Lukes 1973). Then there was an epistemological argument that the tactic of studying only individuals will not allow the recovery of properties of groups. Finally there was a conceptual argument that collectivist concepts such as 'nation' or 'army' cannot be replaced by individualistic ones like 'citizen' and 'soldier' with preservation of meaning.

I think it is fair to say that the issue has not been resolved to anyone's entire satisfaction. The degree to which one sees the ideal form for a society along the polar contrast between collectivist and individualist is not just a question of one's political or religious affiliations. It is also a matter of how far one sees oneself as a product of ready-made social orders, and how far one thinks of oneself as a crusader against the tyranny of the majority.

See also: Bentham, Jeremy (1748–1832); Consequentialism (Including Utilitarianism); Control, Social; Durkheim, Emile (1858–1917); Freedom: Political; Individualism versus Collectivism: Philosophical Aspects; Marx, Karl (1818–83); Methodological Individualism in Sociology; Methodological Individualism: Philosophical Aspects; Mill, John Stuart (1806–73); Norms; Rousseau, Jean-Jacques (1712–78); Social Influence, Sociology of; State and Society: History of the Concept.

Bibliography

Bentham, J., 1789 [1948]. Introduction to the Principles of Morals and Legislation. Blackwell, Oxford, UK.
Durkheim, E., 1895 [1982]. The Rules of Sociological Method. Macmillan, Basingstoke, UK.
Giddens, A., 1984. The Constitution of Society. Polity Press, Cambridge, UK.
Hobbes, T., 1651 [1946]. Leviathan. Blackwell, Oxford, UK.
Lukes, S., 1973. Individualism. Blackwell, Oxford, UK.
Marx, K., 1867 [1965]. Capital. Lawrence & Wishart, London.
Menger, C., 1883 [1981]. Problems of Economics and Sociology. New York University Press, New York.
Mill, J.S., 1959. On Liberty. Longman Green, London.
Ockham, W., 1990. Ockham: Philosophical Writings (P. Boehner, Ed.). Hackett, Indianapolis, IN.
Popper, K.R., 1961. The Poverty of Historicism. Routledge and Kegan Paul, London.
Rousseau, J-J., 1764 [1968]. The Social Contract. (M. Cranston, Trans.). Penguin Books, Baltimore, MD.
Smith, A., 1776 [1976]. Inquiry into the Nature and Causes of the Wealth of Nations. Clarendon Press, Oxford, UK.
Vico, G., 1725 [1948]. Scientia Nova [The New Science]. (T.G. Bergin and M.H. Fisch, Trans.). Cornell University Press, Ithaca, NY.
Windelband, W., 1894 [1998]. Geschichte und Naturwissenschaft (History and natural science). (J.T. Lamiell, Trans.). Theory and Psychology 8 (1), 5–22.

Indo-European Languages

Philip Baldi, Penn State University, University Park, PA, USA
Pierluigi Cuzzolin, Università degli Studi di Bergamo, Bergamo, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article the criteria for establishing language families are reviewed, and are then applied to a diagnostic set of morphological and phonological data from the Indo-European family of languages. Each of the 11 subgroups of attested Indo-European languages is outlined, with essential dates and earliest monuments mentioned for each group. An outline of the phonological system of the protolanguage is provided, as well as sample information on its morphological and syntactic systems.

Language Families

Inheritance

Before we begin our discussion of the Indo-European (IE) family of languages, we should make explicit what exactly we mean by the term 'language family.' The term 'language family' is used in linguistics to characterize a genealogical model of the relationship between languages. A 'family' of languages is a set of languages which derives from a common ancestor or 'parent' language, e.g., the IE family of languages consists of the 'daughter' languages Greek, Latin, Sanskrit etc., all of which developed out of the Proto-Indo-European (PIE) ancestor. Languages that have a common ancestor can be shown to share common traits, which identify them as members of the same family.

Curiously, the genealogical analogy extends only as far as the node of the single parent, usually the 'mother' language. Under normal circumstances of language evolution, there is only one parent. When there is more than one parent present, we are dealing with a 'creole' situation, which is not generally considered to be normal language evolution. Under such circumstances speakers may cobble a language together using parts from two existing languages that they know (Dixon, 1997: 13). The creole model is particularly useful when linguistic families exhibit scant evidence for a chronologically deep reconstruction.

The notion of the language family is usually restricted to units comprising agreed-upon languages and language subgroups. Beyond the family, there is the 'phylum.' A phylum is a grouping of languages, which is much more speculative than is a family: a phylum is a more remote level of grouping, where a clear ancestor is not known. The term 'macro-phylum' is used for even less certain groupings of languages.

For example, one can find the term 'Australian family' (of Aboriginal languages), even though there is no clear parent language and it is not even clear that the languages in this 'family' are related to each other. For this grouping the term 'phylum' or 'macro-phylum' would be a better choice.

Evidence

The notion of a language family is founded on the idea that two or more languages may display a noticeable similarity to each other. They may share common elements of lexicon (vocabulary), phonology (sound patterns), and morphology (word structure) and those similarities may be an indication that the languages are descended from a common ancestor. If the shared features are not due to chance, to borrowing through contact or to universal properties of language in general, then the languages in question have most likely inherited these features from a common source, a mother language. It follows from this that the languages are in the same family.

The relationships of the members of a language family are typically represented in *family tree* diagrams such as the one below.

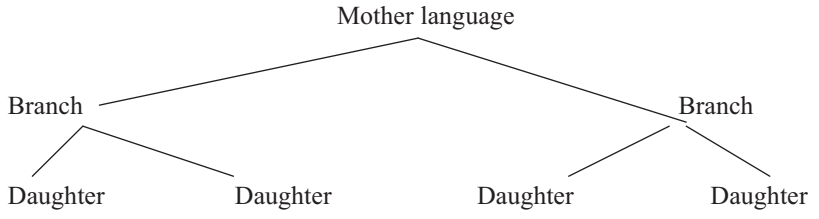

We should point out that family tree diagrams are much more complicated as they typically have a number of subgroupings. Continuous branching is the norm, with subfamilies continuing to split until every node on the tree ends with a specific language. It is assumed that features which are shared via contact have been isolated and eliminated from genealogical consideration (fully articulated trees of the IE family can be found in Baldi, 2002: 22; Fortson, 2004: 10).

With the foregoing discussion in mind, let's return to the question: how do we know that languages are related to each other, that they stem from a common ancestor? In other words,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.52046-7

how do we know that they are in the same family? We can start by reviewing some morphological data from several languages that share some similarities that we want to explain. Remember, this is only one part of the proof of genealogical relatedness. Similar types of arguments can be made by comparing phonology and syntax.

Linguists know from over two hundred years of scientific study that while languages freely borrow vocabulary it is less frequent that they borrow elements of grammar or word formation. It would be unlikely, for example, for a language to borrow the plural ending on nouns, especially a rare or irregular one, like the *-en* on *oxen*, or the *-ren* on *children*, or a vowel-changing plural like that in *foot/feet*. Nor would we expect to find borrowed verbs of the form seen in *sing/sang*; or *bring/brought*. Each of these forms marks a highly specific lexical function and it would be very unusual for another language to borrow them.

The reason languages don't borrow these types of forms is that more general mechanisms already exist within each language that perform these functions, forming the plural of nouns and the past tense of verbs. Furthermore, specific endings such as *-ren* or *vowel change* are from a layer of exceptions that exist in languages and are unique to the words mentioned. For example, the word *foot* has a plural form, *feet*. This type of plural in English is restricted to a few nouns such as *goose/geese*, *tooth/teeth*, *mouse/mice*, *louse/lice*, and so on. If English were to borrow or coin a word *schnub*, meaning 'a box full of foxtails,' the plural of this word would no doubt be *schnubs*, not ***schnuben*,***schnubren*, or ***schnub-schnibrin*.

We can illustrate further with a less abstract example. The verb to *sauté* is borrowed into English from French. The past tense of *sauté* in English is formed by adding the normal past tense marker *-ed*, *sautéed*. In other words, the past tense is formed by using the English ending *-ed*, not the French past tense formation (which would be **I have sauté*), or some irregular English marker such as *sauté/*seuté*, which would be like *fall/fell*. Similarly again with plurals: *typhoon* is borrowed into English from Chinese. If there is more than one of them, they are *typhoons*. English uses its most common plural and doesn't follow the Chinese pattern (which would be *typhoon*), or some other English plural marker such as ***typhoonen*, ***typhoonren*, or ***typhoon-typhin*. (Note: Possible but nonoccurring words are marked with ** in this article).

But this is all negative evidence – what we wouldn't expect to find, not what we would expect to find, if two languages were related. Let's consider some data from German and English, two well-known European languages that we already know to be in a genealogical relationship. Based on our observations of living languages, we would not expect to find a set of irregular adjectives to be borrowed from one language to the other, but yet we find a near-perfect match between two such forms in both English and German, viz., Eng. *good/better/best* and *much/more/most* paralleled by Germ. *gut/besser/best* and (*viel*)*/mehr/meist*. Each is an exception to the general rule of adjectival comparison in present-day English and German (i.e., add the suffix *-er* and *-est* on the base of an adjective: *tall*, *taller*, *tallest*; *dunkel*, *dunkler*, *dunklest*). This would be a most unusual borrowing; indeed, the only practical explanation for this shared similarity is that the two languages inherited the pattern from some earlier stage in their history, when they were undifferentiated:

Here is a summary of some formations shared by English and German, which can only be explained by common inheritance:

German has a (rare) plural suffix in *-s*, which matches the one in English (German: *Schmidt/Schmidts* = English: *Smith/the Smiths*).

In German, *-en* is found in plural formations such as the word for *woman* (*Frau/Frauen*). It is the same as that seen in the plural of the English *ox/oxen*.

In German, there is a vowel-changing plural in *Fuss/Füsse* (pronounced *foos/fees-uh*), which is paralleled exactly by the English *foot/feet*.

In German, the irregular comparative formation *gut/besser/best* is a perfect match to the English *good/better/best* and the German *viel/mehr/meist* is a partial match (the latter two words) with the English *much/more/most*.

In German, the verbal forms *bringen/brachte* match the English *bring/brought*, and the German *singen/sang/*(*ge-*)*sung*(*-en*) matches the English *sing/sang/sung*.

The presence of such a high degree of similarity in these specific parts of German and English is a robust indicator that the two languages are related and that they have inherited features from some common ancestor. The common ancestor in this case is Proto-West-Germanic.

The Comparative Method

In addition to the similarities which are made obvious by morphological comparison, we also evaluate the data according to the comparative method, which is a method used by linguists to test the phonology for patterns of similarity and consistency among matching segments. We have already seen how morphological data are used to establish filiation between languages.

Consider the following data, from four West-Germanic languages and Latin:

English	*Dutch*	*Latin*	*Yiddish*	*Old English*
mouse	muis	mūs	moyz	mūs
seven	zefen	septem	ziebn	seofon
wind	wind	ventus	vint	wind
goose	gans	(h)ānser	gandz	gōs
winter	winter	XXX	vinter	winter
goat	geit	haedus	XXX	gāt 'she-'goat'
wade	waden	vādere 'to go'	XXX	wadan
whore	hoer	cārus 'dear'	hor	hore
mother	moeder	māter	muter	mōdor
six	sechs	sex	zeks	siex
sister	zuster	soror	shvester	sweoster

XXX indicates that a cognate is lacking.

The overall composition of these words, in effect, satisfies the morphological criterion that the vocabulary must bear some similarity to each other. Our basic intuition about these words is that the members of each line are, in a sense, 'the same word.' However, making this assertion and proving it are two different things. To prove it, we must establish that there is a systematicity to the phonological patterns that we claim to have found.

That systematicity is to be found in the *regularity* with which one sound represents a parallel sound in another language. For example, the *m-* in 'mouse' is consistent in the initial position in our chart ('mouse,' 'mother'), and the same is true of the *-n-* in 'wind' and the *-r-* in 'whore.' Note also that the initial *g-* in 'goose' and 'goat' are matched by an *h-* in Latin (the initial *h-* of (*h*)*anser* is optionally written, showing the weak articulation of the *h-*). The medial *-d-* of 'wade' is a match. Note that the vowels seem to have a less-than-regular pattern, at least when compared to the consonants. But this is an illusion: in fact, the vowels show as much systematic correspondence as do the consonants, though the demonstration of these correspondences would take us way beyond the purposes of this survey. What we should stress, however, is that sounds need not be identical, or even phonetically similar, in order to fit into a sound correspondence. What matters is that they be systematic in their correspondences. An extreme, but instructive example of this is to be found in Armenian *erk-*, which systematically continues *dw-*, even though the clusters are phonetically very different. The somewhat problematic final *-n* in 'seven' can easily be shown to derive from an earlier *–*m*. The final *-n* in 'wade' appears only in two of the five words, where it is a morphological marker indicating "infinitive,' and hence is not a part of the phonological equation, and not subject to regularity. The final *-d* in 'wind' is matched by two *t*'s in Latin and Yiddish, and by a final *-d* in Dutch and Old English. All these can be traced back to a final *–*t*, with Dutch and Old English giving way to a forward (progressive) assimilation of voice in the combination *-nt*, which results in *-nd*. The *g-* of 'goose' and 'goat' is consistently derived from an earlier reconstructed **gh-*, with the normal outcome of this sound being *g-* in Germanic, and *h-* in Latin.

With the exception of the words marked with XXX, where there is no cognate form, we can readily see that there is a systematicity of correspondence in these data. That is, wherever we find a *g-* in the initial position in English, we are sure to find an *h-* in Latin. And so it is with all the other words in Germanic which begin with *g-*, such as 'guest' (Eng. *guest* = Lat. *hostis*). Note once again that we are not claiming that the sounds must be the same from language to language, but only that they match, or correspond. That is, if there is a sound X in environment A in Language 1, then there will consistently be a sound Y in environment A in Language 2 if the languages are genealogically related (Campbell, 2004).

The IE Family of Languages

By the term 'Indo-European' we are speaking of a set of genealogically related languages which, by the second millennium, were written and/or spoken over a large part of Europe and parts of India, stretching from the Indian subcontinent in the east and Europe in the west at the time of their identification as a family in the eighteenth and nineteenth centuries. Of course, modern discoveries such as Anatolian (discovered in modern Turkey) and Tocharian (found in western China), as well as post-Columbian colonial expansions, now suggest a different name, but the original 'Indo-European' (and its German equivalent *Indogermanisch*) has stuck through the centuries.

Thus, though the term is in origin a geographical one, the name *Indo-European* is primarily linguistic, referring as it does to linguistic branches of the language family, such as the 'Italic branch' or the 'Germanic languages.'

The dating and location of the protolanguage from which all the IE languages have descended are controversial but the most widely held opinion among specialists puts the protolanguage in the area of the Pontic-Caspian steppes north of the Black and Caspian Seas at around 3500 BCE. Around this time the protolanguage began to diversify into the descendant subgroups through a number of phases and stages. Proposals have been made for a later date (*c*2500 BCE), also in the steppes, and for a much-earlier beginning for the protolanguage, from *c*7000–10 000 BCE, in what is now modern Turkey. Still other proposals project the protolanguage back to *c*12 000 BCE, for the macro-phylum called *Nostratic*, which included, besides IE, the Semitic, Caucasian, Uralic, Altaic, and Dravidian families. Most scholars are not persuaded by either of the earlier dates or at least by the availability of methodology which can take us back that far. This is the position that we adopt in this article.

In comparison with some of the other 250–300 language families of the world, the IE family is relatively small. It contains about 140 languages (many extinct), more than 90 of which belong to the Indo-Iranian family. These 140 or so languages are classified into 11 subgroups, namely Anatolian, Tocharian, Indo-Iranian, Baltic, Slavic, Armenian, Albanian, Germanic, Celtic, Italic, and Hellenic (Mallory, 1989).

By contrast, the Austronesian language family (Pacific) has some 800 languages in a large number of subgroups, and the Bantu family (Africa) has as many as 400 languages. Of course, it is important to distinguish the number of languages in a family from the number of speakers, or their social/geopolitical importance. By these latter criteria the IE family, specifically through global languages such as French, Spanish, and especially English, has an almost unique standing among the language families of the world.

When we say that such and such a word is IE, it means that we are claiming that the word can be traced all the way back to PIE and that it has evolved in the daughter languages as a descendant of PIE. When we claim that a word is Germanic (or Baltic, etc.), we are claiming that the word is only found in the Germanic subfamily more or less in that form. If we say that a word has been borrowed from A into B, we are saying that the word is not original in B.

We now discuss each IE subgroup and the dates of its most important linguistic monuments.

The IE Subgroups

Anatolian. Completely extinct, the Anatolian languages were unknown until archaeological excavations in Boğazköy,

Turkey, in the early twentieth century uncovered the royal archives of the ancient Hittite city of Hattušaš. The original trove of about 10 000 clay tablets (now numbering 25 000), dating from the seventeenth to the thirteenth centuries BCE, was deciphered from its cuneiform script and shown to be representing an IE language now called Hittite. Other languages were also represented in the tablets, both IE (e.g., Palaic and Luwian) and non-IE (e.g., Sumerian and Akkadian). The main language of the Hattušaš materials, however, was Cuneiform Hittite, the chief representative of the Anatolian branch. Other Anatolian languages which have been uncovered and classified since that time are Palaic, Hieroglyphic and Cuneiform Luwian (and its offshoot Lycian), Carian, and some other fragmentary languages.

Indo-Iranian. This subgroup contains two closely related subdivisions, namely Indic (also known as Indo-Aryan) and Iranian. Because of the extremely early dates of attestation for the oldest languages of Indo-Iranian, especially Sanskrit, and their generally conservative (preserving many features of the protolanguage) character, these languages have figured heavily in our understanding of the protolanguage and of the IE language family.

Indic (sometimes called Indo-Aryan). The languages of the Indic group are classified into three historical periods, namely Old Indic (1500–600 BCE), Middle Indic (600 BCE–CE 1000), and Modern Indic (since CE 1000). The most ancient language is Vedic, an archaic form of Sanskrit whose oldest documents are usually dated to about 1200–1000 BCE. A closely related form of Vedic is Classical Sanskrit (*c*500 BCE). The oldest Middle Indic texts are in Pāli (sixth to fifth century BCE), followed by the Aśoka inscriptions (*c*250 BCE) and some Jainist religious writings from about the same period. Modern Indic is one of the largest and most heterogeneous of the IE subgroups, with perhaps as many as 90 different languages. Among the best known of them are Hindi/Urdu, (Hindi is written in the [Hindu] *devanāgarī* script, while Urdu is written in the [Moslem] Arabic script, though the languages are essentially the same), Marathi, Punjabi, and Gujarati. It should be noted that the majority of the modern Indic languages derive from the *Prakrits*, which are spoken varieties of Indo-Iranian that must be kept distinct from the written *Sanskrit*, which was used for religious purposes.

Iranian. Ancient Iranian has two important representatives of IE. The first of these is Avestan (with its oldest variety preserved in the Gāthās), dating from about 600 BCE, possibly earlier. The second important member of Ancient Iranian is Old Persian, which may date to as early as 500 BCE. Modern descendants of Iranian are, among others, Modern Persian (Fārsi), Pashto, and Kurdish.

Greek. The Greek branch contains some of the oldest testaments of the IE family. Attested in inscriptions from as early as the eighth century BCE, Greek also has textual monuments in the Homeric epics, *The Iliad* and *The Odyssey*, which may be as old as 800 BCE. Even older than these are the Linear B tablets from Crete, Pylos, and other ancient locales, which represent a form of Greek called Mycenaean and may date from as far back as the fourteenth century BCE.

Italic. The Italic subgroup of the IE consists of many genetically connected languages from ancient Italy that share certain distinctive characteristics. There are two main Italic subdivisions. The more important of the two, Latin–Faliscan, is represented chiefly by Latin, one of the most important IE languages and arguably the most important language in the development of Western Civilization. Latin is identifiable in some short inscriptions from the seventh century BCE. However, the first literary texts stem from the third century BCE. Faliscan is known only from inscriptions, the oldest of which dates to the early seventh century BCE. Latin survives in the form of the modern Romance languages that developed from spoken varieties of the language in various parts, at different times, and in social circumstances in the history of the Roman Empire and beyond. The best known of the Romance languages are Italian, French, Spanish, Portuguese, Catalan, and Rumanian; less well known are Dalmatian, Rhaeto-Romansch, Ladino, Sicilian, Sardinian, Occitan, and many other local and social varieties. The second Italic subdivision is called Sabellic or Sabellian, represented chiefly by the Oscan and Umbrian, which in the *Iguvine Tables* represents the only foundation ritual we possess in any IE language.

Germanic. The Germanic subgroup, which includes English among its members, is widespread geographically. The oldest attestations of Germanic are the Scandinavian Runic inscriptions, the oldest of which dates from the first century CE. The Germanic languages are conventionally separated into three geographic subdivisions.

The first, East Germanic, contains only a single well-attested language, Gothic, which died out, but traces of which were still found in Crimea in the sixteenth century. The second subdivision of Germanic is North Germanic, whose principal older representative is Old Icelandic from the ninth century, Modern North Germanic languages are Icelandic, Faroese, and Norwegian in one group, with Danish and Swedish in another. The final group, West Germanic, is the most expansive and internally diverse of the Germanic languages.

Five West-Germanic subgroups are recognizable:

Old High German, which descends into Middle High German and on to Modern German and Yiddish, a 'fusion' language based on three components, viz. (the dominant) Germanic, Semitic, and modern dialects such as those from Slavic (Katz, 2000: 496).

Old Saxon, which eventually develops into Low German.

Old Low Franconian, which evolves ultimately into Modern Dutch, Flemish, and Afrikaans.

Old Frisian, which becomes Modern Frisian.

Old English (Anglo-Saxon) develops through Middle English to Modern English, with great influence of French after the Battle of Hastings in CE 1066.

It should be noted that 'Modern English' is not a single language, meaning that it has many varieties, some profoundly different from each other. As the most important language in the world, from an economic and political point of view, English is the most commonly acquired second language. Because English is a second language for many, this leads to countless numbers of varieties of English that have developed in different locales around the world, many of them mutually unintelligible. By way of illustration, consider the profound differences among English spoken in Africa, India, Hong Kong, Australia, Ireland, and Scotland, to name just a few.

Celtic. The languages of the Celtic subgroup are divided into two main geographical sections, the Continental and the Insular. The Continental group, made up of Celtiberian [Hispano-Celtic], Lepontic, and Gaulish, is extinct. The oldest written material from the Continental group is from the sixth century BCE. The Insular Celtic languages appear somewhat later. Split into two groups, Goidelic and Brittanic (Brythonic), the Insular languages are first attested in some Ogham Irish inscriptions from around CE 300. The Goidelic group is made up of Irish, Scots Gaelic, and the extinct Manx. The Brittanic group comprises Welsh, the most robust of the modern Celtic languages, Breton, spoken in Brittany (France), and the extinct Cornish.

Tocharian. Discovered in archaeological excavations around the turn of the twentieth century in Chinese Turkestan (Xīnjiāng), the two varieties of Tocharian (usually called simply 'A' and 'B') can be called a significant find, in the sense that they had a serious impact on IE studies, mostly in the form of disputed phonological and morphological archaisms which they preserved, and thereby confirmed for the protolanguage. The documents of the languages, mostly religious and some commercial, are relatively late, stemming from the period around CE 500–700.

Baltic. The Baltic subgroup survives in two modern languages, Lithuanian and Latvian (Lettish), which together make up the East Baltic subdivision. Many other Baltic languages have become extinct, including a language called Old Prussian, which was spoken until the early eighteenth century and represents the West Baltic subdivision.

Slavic. The Slavic languages fall into three geographical subdivisions. The first, South Slavic, comprises Bulgarian, Macedonian, Serbian, Croatian, Slovenian, and the extinct Old Church Slavic, in which the bulk of the oldest (tenth century) Slavic materials are written. The second Slavic subdivision is West Slavic, which comprises Czech, Slovak, Polish, Kashubian, and some others. And finally there is East Slavic, made up of Russian, Ukrainian, and Belarusian.

Armenian. Armenian is first attested in religious documents and translations from the fifth century CE. It shows a great deal of influence from neighboring languages, including Greek, Arabic, Syriac, and Persian. Persian was such a strong influence that Armenian was, at first, misclassified as a dialect of Iranian. Armenian is written in a script developed for Classical Armenian in the fifth century.

Albanian. Unknown linguistically until the fifteenth century CE, Albanian shows a great deal of influence from neighboring languages such as Greek, Slavic, and Turkish, as well as from Latin. This made its identification as a branch of the IE family somewhat problematic when the IE languages were being classified in the nineteenth century. The first document in Albanian is a baptismal formula from the fifteenth century.

Fragmentary Languages

In addition to these 11 subgroups of IE languages, there are also a number of 'minor' languages which are known only from fragments, glosses, inscriptions, and other unpredictable sources. The IE character of some of these languages has been called into doubt, but scholars generally agree on the basic 'IE-ness' of the following languages: Ligurian (Mediterranean region), Lepontic (with possible Celtic affiliations), Sicel (Italic?), Thraco-Phrygian (Armenian and Albanian?), Illyrian (prevalent along the Dalmatian coast), Messapic (connected with Albanian or Italic?), and Venetic (probably connected with Italic). None of these languages exists in sufficient material detail to be of much help in the reconstruction of PIE.

Reconstruction

When we say that we are comfortable with such and such a reconstruction, or that the earlier form of a word or sound was x or y, what exactly are we saying? Our claim is that the protolanguage from which our attested languages have descended is not attested, but it is still discoverable, by reconstruction. Reconstruction is a method used in historical linguistics in which a hypothetical system of sounds or forms representing an earlier, nonextant state of a language (proto forms in a protolanguage) is established deductively from the attested sounds and forms of extant texts. The process of comparative reconstruction is dependent on good written records of several known related languages. In the case of IE, these written languages usually turn out to be Greek, Latin, and Sanskrit, whereas the role of Anatolian is still debated. When such records do not exist, as in many American Indian and African languages, it is still possible to hypothesize about the historical development of these languages by conjecturing about the structural regularities and irregularities of their contemporary states and deducing underlying forms which might reveal earlier states – a process of internal reconstruction (Crystal, 2007: 388–389). Reconstructed items in this article are marked with an *.

The Structure of PIE

Based on the intensive comparison of the IE languages, the IE protolanguage has been reconstructed in considerable detail. In this part of the article, we will identify many of the more prominent features of the protolanguage, concentrating on the phonology, morphology, and syntax.

Phonology

The following phonological system is one of many that have been reconstructed for PIE (for two current ones, see Szemerényi, 1996; Fortson, 2004: 48–62). This system is the least controversial in that it does not incorporate changes brought about by the 'glottalic theory' proposed by Gamkrelidze and Ivanov (1995), now largely discredited.

The phonological system represented in **Figure 1** is based on the comparative analysis of the daughter systems. Some of the sounds have continued to be controversial, such as the set of palatal stops, while most others are secure, being squarely based on the comparative method. This system is maximally balanced: that is, for every sound that occupies a given slot in the system, there is an opposing sound which anchors it. For example, there is a voiceless stop /p/ which is opposed by a /b/, (which occurs only rarely), its voiced counterpart;

Obstruents	p	b	b^h	
	t	d	d^h	
	ḵ	ĝ	$ĝ^h$	
	k	g	g^h	
	k^w	g^w	g^{wh}	
Fricatives	s			
Nasals	n	m		
Semivowels	i̯	u̯		
Liquids	l	r		
Syllabic liquids and nasals	ḷ	ṛ	ṃ	ṇ
Laryngeals	h_1	h_2	h_3	

Vowels

i			u
	e	o	
		a	
	ē	ō	
		ā	

Figure 1 The IE phonological system.

furthermore, the plain voiceless stops such as /p/ and /t/ are opposed by voiceless palatal stops such as /ḵ/ and /ĝ/. Note that all the vowels except /i/ and /u/ have a short and a long variety: the long varieties /ī/ and /ū/ can be accounted for as sequences of short vowel plus laryngeal; and that the voiceless stops are opposed not only by voiceless palatal stops such as /ḵ/ and /ĝ/, but also by plain voiced stops such as /b/ and /d/.

The resulting phonological system is one which is 'tight' from the comparative point of view, and also from the typological standpoint. That is, the phonological system proposed here is one which occurs within the languages of the world, and contains no segments or oppositions which are unusual in the world's languages.

Morphology

The Noun

In all, eight or nine cases have been proposed for PIE: besides the nominative and accusative, there is the genitive (used primarily for expressing a relationship between two nominals, such as *Bill's hat*); the dative (used mainly to express the notion of *indirect object*, as in Give the book *to Bill*); the ablative (which expresses *separation from*, as in Sara took the book *from Hannah*); the locative (which expresses *stationary location*, as in Maxine is *in the room*); the instrumental (which is used primarily to express *instrumentation* or *means*, as in Charley chopped down the tree *with an axe*); the vocative (used mainly to express *direct address* as in Hey, *Jack*!), and finally, though not unanimously, the directional (mainly used in statements expressing *motion toward*, as in *to the house*). In addition, the protolanguage had three numbers (singular, plural, and dual), as well as three genders (masculine, feminine, and neuter). Adjectives agreed with their head nouns in gender, number, and case, and generally followed the same pattern of inflection as nouns. Pronouns, on the other hand, had their own more or less unique inflectional endings, and were unmarked for gender in the first and second persons. A sample paradigm of a noun in PIE is given in Figure 2. These data represent the following words: Hitt. *ḫūmant-* 'every,' Ved. *pad-*, Gk. *pod-* 'foot,' Av. *kəhrp-* 'body,' *nar* 'man,' and Lat. *vōx* 'voice.' The PIE column records only the inflectional endings (Figure 3).

The Verb

The PIE verb is a complex combination of meaningful elements that indicate certain categories. Specifically, the PIE verb marks *voice*, which indicates the role of the subject; two voices, active and middle, from which the passive developed, are usually postulated; *tense*, which locates the verbal action temporally; the following tenses are usually postulated: present, imperfect, aorist, perfect, and possibly a future; and the *mood*, which indicates the factual content of the utterance from the speaker's point of view. At least the indicative, imperative, and optative are generally reconstructed, and often the subjunctive. The voice, tense, and mood markers are attached to different stems, which indicate aspectual categories, e.g., whether the action is continuous or punctual. The entire verbal complex is indexed to the subject by person and number markers. Finally, verbs can be transitive or intransitive, depending on whether they can take an object (*Mary sees Bill*), or not (*Mary sleeps late*). Although there is no specific marker of transitivity on the IE verb, there are morphemes which can transitivize (like causative *-éye/o-) or detransitivize (like *-eh-) in some roots.

	Hitt.	Ved.	Av.	Gk.	Lat.	PIE
Nom.	*ḫūman-za*	*pā́d (=pā́d -s)*	*kərəf-š*	*poús (=pód-s)*	*vōx (=vōc-s)*	*-s
Acc.	*ḫūmant-an*	*pā́d -am*	*kəhrp-əm*	*pód-a*	*vōc-em*	*-m
Gen.	*ḫūmant-aš*	*pad-ás*	*kəhrp-ō*	*pod-ós*	*vōc-is*	*-(é)s
Dat.		*pad-é*	*nair-e*		*vōc-ī*	*-ei
Instr.		pad-*ā́*	k*əhrp-a*			*-(e)h_1
Loc.	*ḫūmant-i*	*pad-í*	*nair-i*	*pod-í*	*vōc-e*	*-i

Abbreviations: Hitt.=Hittite, Ved.=Vedic, Av.=Avestan, Lat.=Latin, Lith.=Lithuanian, PIE=Proto-Indo-European. Nom.=nominative, Acc.=accusative, Gen.=genitive, Dat.=dative, Abl.=ablative, Loc.=locative (after Fortson 2004: 104).

Figure 2 A sample noun paradigm from PIE, with supporting data; singular only.

Vedic	Greek	Latin	Gothic	Old Church Slavic	PIE
bhárāmi	phérō	agō	baíra	berǫ	*bhér-o- h_2
bhárasi	phéreis	agis	baíris	bereši	*bhér-e-si
bhárati	phérei	agit	baíriþ	beretŭ	*bhér-e-ti
bhárāvas			baíros	berevě	*bhér-o-u̯e-
bhárathas	phéreton		baírats	bereta	*bhér-e-to-
bháratas	phéreton			berete	*bhér-e-to-
bharāmas(i)	phéromen	agimus	baíram	beremŭ	*bhér-o-me-
bháratha	phérete	agitis	baíriþ	berete	*bhér-e-te(-)
bháranti	phérousi	agunt	baírand	berǫtŭ	*bhér-o-nti

Figure 3 The present of the verb 'to bear, carry' in reconstructed PIE, with supporting data. Data are arranged according to first, second, and third persons, singular, dual, and plural. From Fortson, IV, Benjamin, W., 2004. Indo-European Language and Culture. Blackwell, Oxford, p. 89.

Adjectives

In PIE, the lexical category of adjective is well documented and is inflected like nouns, usually showing agreement in case, gender, and number: Lat. *bonī cīvēs* (masc.) 'good citizens,' *bonae mulierēs* (fem.) 'good women,' and *bona dōna* (neut.) 'good gifts.' In its oldest stage, PIE did not exhibit any difference between an attributive and a predicative adjective as in *A good friend is a treasure* vs *This man is good*. This distinction is late and occurs only in some branches like Baltic, Slavic and, partly, Germanic. PIE also possessed other adjectival types, such as demonstratives, indefinites, and interrogatives.

Pronouns

PIE possessed a series of personal pronouns that are rather easily reconstructed for the first and second person singular, dual, and plural, whereas for the third person the pronoun was created independently in each branch, mainly from the demonstrative stem.

The relative pronoun was formed basically on two different stems: a stem $*k^{w}o$-/$*k^{w}i$- as in Latin *quī*, *quae*, *quod* or Eng. *who*, *what*; and a stem **jo-* as in Vedic *yas*, *yā́*, *yat*.

Numerals

For PIE, the series of numerals is reconstructed on the base of 10, but there is good evidence that, originally, numeration had a base of 4: the form reconstructed for the number 8 is **(h_3)ókt̄o(w)*, a clear dual form meaning 'twice four.' Moreover in Vedic the word *(h)āsta-*, which derives from **(h_3)ókto-*, means 'hand' probably the so called 'small hand' (four fingers only, excluding the thumb).

It also deserves mention that originally the numbers two and three had different forms for the masculine and the feminine, still preserved in Vedic and Celtic.

Syntax

Even though the syntax of PIE has been intensively investigated during the last five decades, it still remains a field particularly open to new research and is not as well established as is, for instance, morphology. These current limits notwithstanding, there is a small set of features that can be confidently reconstructed for PIE.

PIE as a 'be' Language

The IE protolanguage is generally considered to be a 'be' language rather than a 'have' language, that is, it expresses the possessive relationship such that the *possessum* is the subject in the nominative case, the verb agrees with the subject, and the *possessor* is in the dative case, as in Lat. *mihi est liber* 'the book is at/to me' (I have a book). Similar archaic constructions are found in Slavic, Celtic, Sanskrit, and elsewhere. The alternative, the 'have' language, predicates the possessive relationship with a verb 'to have,' as in Eng. 'John has a book,' found in Germanic, Latvian, Greek, and elsewhere (Baldi and Nuti, 2010).

Subject–Object

PIE was a language in which alignment was of the nominative–accusative type: the subject regularly occurred in the nominative case whereas the object occurred in the accusative, independently of the verb: Lat. *Antoniu-**s*** (Nom.) *libru-**m*** (Acc.) *legit* 'Anthony reads a book.' There are some very old patterns where the 'subject' occurs in the accusative or dative, as in Lat. *mē pudet* 'I regret,' OE *mē likeþ* 'I like,' which probably witness an older stage in which the protolanguage was of the ergative or accusative type. But this is still matter of debate and we leave it aside in this article.

Word Order

In PIE the basic word order, i.e., the order in which the Subject (S), the Verb (V), and the Object (O) occurred in a positive declarative sentence like *John throws the stone*, was SOV. This is not certain beyond any doubt but there is a considerable amount of evidence that points to it: for instance, SOV is the regular word order of the declarative sentence in the oldest languages, such as Hittite GU_4-*ma* AMAR-*ŠU mimmas* 'The cow rejected her calf,' but literally 'cow-but calf-her rejected.

Moreover, it is characteristic of SOV languages that whatever modifies an element usually precedes the element that is modified. This order is well documented, at least in the oldest attested languages, with the genitive before the head noun: Lat. *terrae mōtus* 'earth-quake,' with the adjective preceding the noun: Lat. *rāra avis* 'rare bird'; in the comparisons: Gr. *mélitos glukíōn* 'sweeter than the honey'; and even in the case of the relative clauses: Lat. *cūius regiō, eius religiō* 'whatever one's region, so too his religion.' The order modifier–modified element is clearly documented also by the oldest compounds like Ved. *rājaputrá-* 'son (*putrá-*) of the king (*rāja-*).'

Pronominal Subjects

PIE was a language in which the subject pronouns were not obligatorily expressed, as is the case in many of world's languages but not in Germanic as evidenced by modern English or German. Subject pronouns were employed in contrastive or emphatic contexts only.

Clitics

For PIE, scholars have reconstructed a set of words without their own accents which occurred immediately after the first accented word of the sentence. These words are labeled clitics 'words that lean (on the preceding word).' They are attested in the oldest languages: for instance the conjunction *-kwe, giving Lat. -que as in *senatus populusque* 'The senate and the people' or Gr. *pólemoí te mákhai te* 'and wars and battles,' where the clitic *te* occurs after both nouns.

Subordination

The system of subordinates that is reconstructed for PIE is rather poor: subordination was expressed by means of connecting particles rather than by explicit subordinators. This only means that a true system of subordinate clauses, such as those found in modern languages, had not yet developed in PIE. It is interesting to observe that a huge number of such subordinate clauses were introduced by inflected forms of the relative or the demonstrative stem. Moreover, it should be noted that in PIE there is a small set of verbal nouns (including infinitives and supines) and adjectives (including participles) that is used to express hypotaxis.

Among the different syntactic patterns employed to express subordination with verbal nouns and adjectives, the so-called absolute constructions play a special role. They represent a pattern attested in almost all the IE languages that can be traced back to the protolanguage, despite the fact that the morphological case in which the verbal participle occurs varies from language to language: Lat. *victis hostibus, dux triumphum celebravit* 'after defeating the enemies, the commander celebrated his triumph.' In this example, the participle *victis* occurs in the ablative; but in Greek it occurs in the genitive, in Vedic in the locative, and in Slavic in the dative, just to mention a few examples.

Bibliography

Baldi, Philip, 2002. The Foundations of Latin. Mouton de Gruyter, Berlin.

Baldi, Philip, Nuti, Andrea, 2010. Possession in latin. In: Baldi, Philip, Cuzzolin, Pierluigi (Eds.), New Perspectives on Historical Latin Syntax, vol. 3. Mouton de Gruyter, Berlin, pp. 239–387.

Campbell, Lyle, 2004. Historical Linguistics: An Introduction. MIT Press, Cambridge.

Crystal, David, 2007. A Dictionary of Linguistics and Phonetics. Blackwell, Oxford.

Dixon, Robert M.W., 1997. The Rise and Fall of Languages. Cambridge University Press, Cambridge.

Fortson IV, Benjamin W., 2004. Indo-European Language and Culture. Blackwell, Oxford.

Gamkrelidze, Thomas V., Ivanov, Vjacheslav V., 1995. Indo-european and the Indo-europeans (Part 1, the text; part 2, bibliography, indices). (Trends in Linguistics: Studies and Monographs 80). Mouton de Gruyter, Berlin.

Katz, Dovid, 2000. Yiddish. In: Price, Glanville (Ed.), Encyclopedia of the Languages of Europe. Blackwell, Oxford.

Mallory, James P., 1989. In Search of the Indo-europeans. Thames and Hudson, London, New York.

Szemerényi, Oswald, 1996. Introduction to Indo-european Linguistics. Oxford University Press, Oxford.

Relevant Website

www.utexas.edu/cola/centers/lrc/links.html – The University of Texas at Austin – Offsite Web Links.

Industrial Cluster Dynamics

Max-Peter Menzel, Institute of Geography, Universität Hamburg, Hamburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Regional clusters are a pervasive phenomenon. Yet, clusters evolve and exhibit different forms in different places. Clusters are defined by interconnections of geographically concentrated firms in a particular field. This article will describe how these interconnections affect the propensity and shape of clusters. First, it describes the particularity of interconnections between geographically clustered firms and shows how firms substitute permanent colocation by temporarily generated proximity. Second, the article describes how these processes are connected to the emergence and decline of clusters. Third, the article analyzes how socioeconomic contexts affect these interconnections, resulting in variegated clusters.

Introduction

Industries are geographically fragmented and distributed unevenly in space. Geographical agglomerations of interconnected firms and industries are commonly termed as 'clusters.' Clusters like the ICT cluster in Silicon Valley or the Hollywood movie cluster can dominate whole industries. Beside these exceptional examples, clusters are a regular phenomenon of economic activity and important elements in regional economic development.

Porter (2000, p. 15) defines cluster as:

> geographic concentrations of interconnected companies, specialized suppliers, service providers, firms in related industries, and associated institutions (e.g., universities, standards agencies, trade associations) in a particular field that compete but also cooperate.

These interconnections between firms in a particular field define the clusters. The emphasis on interconnections and interdependencies indicate that clusters not only consist of firms in the same industry, but often form along a value chain. The steel and mining cluster in the Ruhr area, for example, also produced machines for steel production and mining (Grabher, 1993). As such, clusters differ from agglomerations of firms of the same industry and are difficult to detect by standard industry classifications. Initiatives to map clusters like www.clustermapping.us for the US or www.clusterobservatory.eu for the EU therefore use bundles of related industries, whereby a regional concentration of different related industries indicates a cluster (cf Porter, 2003).

Another element that distinguishes a cluster from firm agglomerations is its specific institutional environment. Special education infrastructures, industry organizations, and rules of interaction support the activities of firms in the cluster. The specific institutional setting also marks the boundaries of the cluster. Maskell (2001, p. 936) defines the boundaries of the cluster as the "fit between the economic activities carried out by the related companies of the cluster on the one hand and the particular institutional endowment developed over time to assist these activities on the other." These boundaries usually do not necessarily align with administrative boundaries. The geographical extent of clusters might range from a few blocks and streets as found for many media clusters to large areas like in the rural Danish wind turbine cluster. Clusters even pervade country borders like the pharmaceutical cluster around Basel that ranges from Switzerland to Germany and France.

Regional clusters are a pervasive phenomenon. Yet clusters evolve and take different forms in different places. This article will describe the dynamics that affect these processes and in doing so the propensity and shape of clusters. The remainder of the article is structured as follows. The next section will give theoretical explanations for clustering processes. Additionally, it will describe how firms substitute the benefits of permanent colocation by generating temporary forms of geographical proximity. The third section describes the evolution of clusters, i.e., how clusters come into being and how their benefits might turn into disadvantages. The fourth section analyzes how the specific context affects form and extent of spatial clustering, and how changing context conditions over time and space results in variegated clusters. The fifth section concludes and gives an outlook on recent trends in the evolution of regional clusters. The theoretical accounts will be illustrated on the ICT cluster in Silicon Valley as probably the world's most prominent and investigated cluster.

Cluster and Geographical Proximity

Firms in clusters experience stronger growth and faster innovation than those outside clusters (Audretsch and Feldman, 1996; Baptista and Swann, 1998). There are several explanations why geographical proximity between firms might have these effects and the overwhelming ones focus on externalities, i.e., processes taking place outside and between firms (for another explanation see Klepper (2007)). The importance of externalities depends on the possibilities to divide labor. Accordingly, fields like ICT that provide more possibilities of division of labor are stronger clustered than industries like oil refinery (Steinle and Schiele, 2002).

Classical accounts on agglomeration economies already described by Marshall (1920) argue that colocated firms in the same industry benefit from a pooled labor market, knowledge spillovers, and localized supply of nontraded specialized inputs. Contemporary explanations mostly focus on interactions between firms and how interactions in geographical

proximity enhance firm performance. Transaction costs approaches, for example, explain the concentration of firms in geographical space as input–output relations in geographical proximity that reduce transaction costs, either in the form of omitting transport costs (Gordon and Mccann, 2000) or facilitated coordination, especially when firms have to rely on flexible supplier relations to react to volatile markets. Yet, the observations that competition between firms increasingly takes place via innovation than via prices and that innovative activities are increasingly spatially concentrated (Florida, 2005) brought the relation between clusters and innovative activities into focus. This relation is conceptualized by learning via interaction, whereby firms exchange, integrate, and utilize knowledge.

Malmberg and Maskell (2006) elaborate on the particularities of localized learning. First, localized learning describes local capabilities that are "some forms of knowledge creation and exchange that are still very much rooted in the cultural, institutional, and social structures of particular places" (Malmberg and Maskell, 2006, p. 3). In this vein, firms benefit from access to knowledge deeply embedded into a regional context. The more knowledge is codified and globally accessible, the scarcer becomes the embedded knowledge that is hard to codify. The global pharmaceutical industry is a good example of how firms try to get access to these scarce resources. Zeller (2004) describes how the Swiss companies Roche and Novartis established research facilities in the most important clusters of the industry in order to tap into the regional embedded knowledge.

Second, spatial proximity facilitates localized learning on several levels. On the dyadic level, geographical proximity improves learning processes by facilitating face-to-face interaction. Face-to-face interaction allows rapid feedback especially under conditions of vagueness and uncertainty and it enables the transfer of tacit knowledge in form of mimics, gestures, etc. It is "verbal, physical, intentional, and unintentional" (Storper and Venables, 2004, p. 355). This multidimensionality is essential for the transfer of complex knowledge.

Different ways of knowledge exchange, often in the form rumors or recommendations, caused by intended and accidental contacts on various occasions between firm owners and employees within a cluster can create a steady and pervasive stream of information. This stream of information is described with the word 'buzz' (Bathelt et al., 2004; Storper and Venables, 2004). Local buzz allows a firm to monitor other firms, observe what competitors are doing, assess the behavior and results of other firms' activities, and combine their ideas with their own knowledge. Local buzz enables colocated firms to perceive actions of other firms or market developments without deliberately scanning their environment.

Because of the manifold ways of interactions, the particularity of localized learning does not lie in transaction costs or in the efficiency of knowledge transfer. It is rather the diversity of knowledge sources that are perceived and accessible and the type of knowledge that is transferred between actors when they are in the same location (Malmberg and Maskell, 2006). Yet, accessing external knowledge is important for the longevity of a cluster, and efficient knowledge exchanges do not depend on geographical proximity. Mutual understanding between actors and efficient knowledge exchange can also take place within long-term relationships, which often form in geographical proximity, but endure in geographical distance. Firms strategically generate these relations to firms outside the cluster. Trust and efficient knowledge exchange in geographical distance is enabled by a "set of procedural rules involving a sequence of transactions and interactions" (Bathelt et al., 2004, p. 41), which additionally is supported by temporarily generated spatial proximity within projects and meetings (Torre, 2008).

Also 'local buzz' can be substituted. Industry fairs create a temporarily and spatially limited context, where actors from different locations meet, exchange knowledge, give recommendations, and exchange assessments on markets, products, or firms. Fairs thus can be termed as 'temporary clusters' (Maskell et al., 2006), which do not depend on permanent colocation of firms. Accessing knowledge produced elsewhere spills over to other firms in a cluster. In doing so, external relations enhance the quality of local buzz as other firms also get more information about the environment outside the cluster.

How a cluster gives affects the way how firms innovate and produce is exemplary described by Saxenian's (1994) comparison of the computer firms in the Silicon Valley and the Boston area. She described how in the 1990s computer firms in the Silicon Valley by applying firm strategies based on externalities, i.e., low levels of vertical integration and dependence on manifold collaborations with other firms, outperformed Boston computer firms. The strategies of two workstation producers will serve as respective examples (the following is based on Saxenian (1996)).

Apollo was formed in Boston in 1980. Sun was formed in the Silicon Valley 2 years later. Apollo started to establish a proprietary technology. To monopolize its technology and secure its knowledge, Apollo relied to a high degree on in-house research. It developed and produced large parts of the workstation itself, especially the chipsets and operation system. Additionally, it strongly penalized labor mobility. This strategy resulted in a comparatively high degree of vertical integration. Yet, this strategy was also a result of the particular context. As the typical computer producer in Boston was highly vertically integrated, there were few possibilities to vertical disintegrate.

In contrast, Sun adopted a more collaborative and network approach. Its workstation bases on the open UNIX standard to adhere to technologies and components from other firms. Sun only produced prototypes and assembled their workstations. Components were produced by other manufacturers. This network-based model impedes the monopolization of knowledge, but in turn it allowed Sun to easily integrate innovations made by other firms in the Silicon Valley.

The different approaches were also reflected by the founders. The founder of Apollo was 46 years old and its team already had respective experiences from a previous formation. Sun was founded by a team of PhD students in their 20s. Such formations by talented but inexperienced founders were facilitated in an environment like the Silicon Valley where they could rely on external resources. The high propensity to form a firm (and shut it down if not working out) in connection with high rates of labor mobility in the Silicon Valley further contributes to the diffusion of knowledge.

It should be kept in mind that this is a story from the 1990 and more recent studies show that Boston firms also adopted network modes of organization with large reliance on regional

collaboration (Bathelt, 2001). However, these contrasting examples show that the benefits of regional clusters do not simply stem from agglomerations of firms but how firms use and produce externalities, both along and at the same level of a value chain. Geographical proximity is only an enabler of such processes.

Evolution of Clusters

The economic advantages that stem from cluster dynamics are not permanent. In fact, the decline of clusters seems to be caused by factors that were advantages in the past. Grabher (1993) describes the decline of the coal and steel cluster in the Ruhr Area. A conglomerate of unions, firms, and parties tried to sustain the established development path based on mass production despite structural changes in the global industry toward flexible forms of steel production. The Ruhr Area is an example for a region that "became 'locked-in' by the very socioeconomic conditions that once made these regions 'stand out against the rest' [… and …] fell into the trap of 'rigid specialisation'" (Grabher, 1993, p. 256). Other examples are the old Lancashire textile district in the UK and the Detroit automobile cluster. Despite the decline, competitive pressure can lead to high innovation rates. These innovations, however, take place within the existing and exhausted technology path. Long existing, closed, and homogeneous networks were not able to renew the cluster by accessing and integrating new and often external knowledge.

While the decline of clusters is often connected with a too strong inward focus, how clusters emerge, and why they emerge at particular places is a highly disputed topic. Usually, emergence of clusters is connected to the upcoming of new industries. Apart from generic factors like a strong science base, an entrepreneurial culture, and supportive institutions, the requirements of future industries are unknown a priori. No place provides a specific supportive environment for the new industry, which makes it impossible to predict where a new cluster will form (Storper and Walker, 1989). This is reflected by the spatial patterns of firms in new industries. Firms in many regions enter new industries, but clusters develop only in few places (Klepper, 2007). As a result, scholars started to conceptualize cluster emergence not as logical outcome of purposeful actions, but as a stochastic process. "Some seemingly historical accidents" (Krugman, 1991, p. 35) increase the number of new firms at a certain location and once a certain threshold is reached, agglomeration economies will start to work (Arthur, 1994).

Yet, there are some patterns observable regarding cluster emergence. The first is that regional firm entries exhibit some path dependencies. Regions that already provide industries that are related to the new industry exhibit larger amounts of entries. For example, the automobile industry evolved in places with a bicycle and coach industry (Boschma and Wenting, 2007). Firms in related industries are a source of entries, as they can diversify and their employees can spin-off into the new industry. Additionally, related industries provide a possible pool of labor, which allows firms in a new industry to utilize localized resources. Also biotech firms formed in places where 'star scientists'(Zucker and Darby, 1996) were affiliated to the local university and research programs and post-graduate education created a labor pool for the new industry.

A second pattern that is connected to cluster emergence are pervasive intraindustry spin-off processes. From the perspective described above, clusters would emerge at the place where agglomeration effects start first, i.e., the places with most entries. But Romanelli and Feldman (2006) show that this did not take place in the US biotherapeutics industry. Instead, only those regions developed clusters where firms that formed in the early phase spawned spin-offs. Also Klepper (2007) showed the importance of spin-off processes for the emergence of the Detroit automobile cluster. Most spin-offs can be traced to only three firms: Olds, Cadillac, and Buick. Furthermore, while the first wave of entries in a new industry is comparatively spatially dispersed, new industries start to concentrate due to spin-off processes that only occur in some places. Therefore, these spin-off processes that Romanelli and Feldman (2006) term 'second generation growth' are an essential feature for the emergence of clusters.

Apart from emergence and decline, clusters are not bound to a particular field but can change, evolve, and transform. Grabher (1993) illustrates how a new cluster in environmental technologies is emerging in the old coal and steel cluster in the Ruhr Area. The companies of the steel and coal complex acquired the necessary competencies correcting environmental damage that was caused by their industry. Tappi (2005) illustrates how the accordion cluster in Marche/Italy by integrating electronics in their previously traditional musical instruments subsequently moved into electronics. Firms in the cluster got their knowledge on electronics by expatriates from Marche who had lived in the US (Tappi, 2005).

These examples also show that clusters change and move into new fields by integrating new knowledge. Spin-off processes serve to diffuse knowledge that often already is available in the cluster, yet not necessarily accessible. Many structural transformations of clusters are therefore connected with spin-offs that venture in fields deviating from the established path of the cluster.

The evolution of the Silicon Valley is a good example for these dynamics. When the semi-conductor industry emerged in the 50s, the technology was invented in established centers like the Bell Labs in New Jersey. Firms formed at quite different places: Texas Instruments in Dallas, Motorola in Phoenix, DEC in Boston, and Shockley Semiconductor in the area later known as Silicon Valley. These places differ in their locational factors. Places like Boston were equipped well with a scientific infrastructure and industrial base at that time, while for the later Silicon Valley Saxenian (1994) observed that "the only local industry was small-scale food processing and distribution" (Saxenian, 1994, p. 11). As a result, scholars like Arthur (1994) interpret the emergence of the Silicon Valley as an accidental process:

> Certain key persons – the Packards, the Varians, the Shockleys of the industry – happened to set up near Stanford University in the 1940s and 1950s, and the local labor expertize and interfirm markets they helped to create in Santa Clara County made subsequent location there extremely advantageous for the thousands or so that followed them. If these early entrepreneurs had had other predilections, Silicon Valley might well have been somewhere else.
>
> Arthur, 1994, p. 100

Yet, what distinguished the early computer industry in the Silicon Valley were the extensive spin-off processes. Fairchild Semiconductor spun out of Shockley Semiconductor, which again was the parent of several firms like Intel. The milieu these entrepreneurs created made it easy for other entrepreneurs to form a firm.

In the late 1980s, the Silicon Valley also faced a decline. Due to increasing rigidity, large mass-producing chipmakers like Intel and AMD had difficulties to react to a changing market and competition from Japanese semiconductor producers, resulting in 27 000 layoffs in semiconductor and related firms. Yet, this downturn led to a new wave of spin-offs producing small batches of highly specialized chips, which renewed the computer and semiconductor cluster (Saxenian, 1990).

Clusters in Different Socioeconomic Conditions

The previous section described how clusters evolve over time. Yet the processes taking place between firms in clusters are not independent from their socioeconomic environment. The socioeconomic conditions change over time and space and affect the benefits of colocation, the propensities of firms to cluster, and the patterns of cluster evolution.

Clusters in Time

At the beginning of the twentieth century, scholars like Marshall (1920) and Weber (1909) gave first explanation on spatial concentration of firms. Weber (1909) described the economic landscape of his time as one of increasing spatial concentration and agglomeration, which was accelerated by decreasing transportation costs. Marshall (1920) described a shared labor pool, knowledge spillovers, and specialized inputs (Krugman, 1991) as reasons for firms forming 'industrial districts.' These industrial districts were shaped by mostly small- and medium-sized firms. These firms specialized both horizontally, i.e., they produced different products for the same markets, and vertically, i.e., they focused on different functions of the value chain. End producers often sold their products internationally, but gained their supplies from a local supplier base.

Yet, this time already saw the upcoming of the vertical integrated corporation as dominant organizational form. The introduction of organizational inventions like the assembly line and scientific management in connection with increasing standardization of end products allowed dividing production into even smaller steps (Robertson and Langlois, 1995). Due to the constant demand for supply created by the assembly line and in order to internalize productivity gains, manufacturers internalized their backward commodity chains (Sturgeon, 2002). The local form of such a production was the highly vertically integrated factory complex. The paradigmatic example of this developmental model was Ford's River Rouge Plant, developed from 1917 on with up to 100 000 workers, which had coal and iron ore as input and manufactured cars as outputs.

From the 1970s, this production model also came into crisis. Reasons were an increasingly heterogeneous and volatile demand which could not be met by Fordist mass production; an exhaustion of productivity gains by deeper technical division of labor; and increased competition from firms that applied more flexible production models, like Japanese car producers.

In the 1980s, Italian economists recognized that particular regions in Northeastern and Central Italy unexpectedly exhibited a positive economic performance. Those regions like the Emilia Romagna were shaped by geographic concentrations of vertically disintegrated small and medium firms that focused on highly specific fields like sofa or shoe manufacturing and performed a craft-based and neo-artisanal production (Becattini, 2002). Upon these examples, Marshall's ides of Industrial district were rediscovered. Yet, compared to Marshall's industrial districts where coordination took place via markets, the Italian examples exhibited manifold nonmarked-based interconnections between firms (Piore and Sabel, 1984). Scholars described economic relations embodied in personal relations, embeddedness of actors in regional networks that penalized opportunistic behavior, and regional institutions that guided interactions as distinct factor of the Italian industrial districts (Becattini, 2002).

This setting enabled trust-based relationships between firms that were necessary to adjust products, supplies, and respective relations to volatile and fragmented markets. Similar forms of localized production and innovation systems around a value chain that performed above average were also found in other places like Baden-Württemberg/Germany or the Silicon Valley. Scholars like Piore and Sabel (1984) saw the flexibility of network-based interfirm relations as a paradigm for a new and post-Fordist form of production. In doing, the foundation for the contemporary concept of the regional cluster was laid.

Clusters in Varieties of Capitalism

The previous section described how the activities of firms and their propensity to cluster changed with changing socioeconomic conditions. These conditions differ not only in time, but also in space. Institutional approaches like the Varieties of Capitalism approach (Hall and Soskice, 2001) argue that firms choose forms of coordination that are institutionally supported. Institutional differences between countries would lead to different behaviors of firms and accordingly to different forms of clusters.

The Varieties of Capitalism approach distinguishes between two archetypes of capitalism: liberal (LME) and coordinated (CME) market economies. In LMEs like the US, coordination between economic actors mainly takes place via markets, competitive relations, contracting, and internal corporate hierarchies. Arm-length relations between firms allow firms to quickly adjusting and switching processes and resources toward new activities. This quality benefits the firms in LMEs in the exploitation of technological developments independent of established fields and outside of existing paths or paradigms.

In CMEs like Germany, nonmarket institutions support strategic interactions and collaborations to address and align the needs of different stakeholders. In such an entangled environment, firms benefit from investing in assets "whose returns depend heavily on the active cooperation of others" (Hall and Soskice, 2001, p. 17). As a result, firms benefit from incrementally and cumulatively building up resources within or in connection to established fields.

This difference in connection to established resources and establishing distinct resources also affects cluster evolution. While the Silicon Valley as example for a cluster in LMEs quickly built up its distinct resources, clusters in LMEs evolve with stronger connection to established fields (Storper, 2009). When the Silicon Valley emerged in the US, the paradigmatic clusters that formed in Europe were the industrial districts in Italy or Baden-Württemberg/Germany, which yet used new technologies but formed around traditional industries like machine or furniture production.

In Asia, cluster evolution again follows a different pattern. The particularities of Asian economies can hardly be grasped by the categories developed within the Varieties of Capitalism approach. Furthermore, the varieties among Asian economies are even larger than among Western economies (Witt and Redding, 2013). China, for example, presented a state-led capitalism with direct interventions and indirect interventions by its state-owned companies, mediated by informal networks, with again different regional varieties (Zhang and Peck, 2014). This institutional variety results in variegated forms of cluster evolution, whose different forms cannot be described in this article. But an example for cluster evolution in the Chinese province Guangdong is described in Rubini et al. (2015).

Conclusion

The intention of this article was to describe the dynamics of regional industrial clusters. Clusters are not only a geographically concentrated group of firms, but rather defined by their interconnections. The quality and characteristics of these interconnections affect the performance of clustered firms. Clusters also change via these interconnections.

The paper started with the particularity of interconnections between spatially clustered firms and conceptualized them as localized learning (Malmberg and Maskell, 2006). Additionally, it showed that firms can substitute permanent physical proximity by organizing temporarily generated proximity like 'temporary clusters' (Maskell et al., 2006) in the form of industry fairs and projects. Learning within and outside the cluster are mutually enhancing, as knowledge relations outside the cluster improve knowledge flows within the cluster. In this vein, access to external knowledge is mandatory for the longevity of regional clusters. When firms rely too much on knowledge flows within the cluster, the tightly regionally embedded firms fail to perceive developments outside the cluster. As a result, the benefits of knowledge exchange in spatial proximity turn into disadvantages and the cluster might decline.

Institutions guide forms of localized learning. In different institutional environments and under different socioeconomic conditions, localized learning exhibits different qualities and characteristics. Clusters from the time of Marshall differ from those of today and nowadays clusters vary between countries and regions. Even in the same industry, clusters are time and place specific.

The intention of the article was to shed some light onto the dynamics that affect cluster evolution. But the rationales under which clusters evolve in space constantly change. Contemporary literature mainly focuses on clusters as entity to improve the innovativeness of firms. Yet, some clusters seem to evolve in a different way and might give examples for the evolution of other clusters.

The Silicon Valley is surely an adequate starting point for such an analysis. Sturgeon (2003) described in the Silicon Valley ICT industry an increasing trend toward modularization of supplier relations. Modularity describes a product architecture where the components of a product are interchangeable without adjustments in other components. Particular components can be integrated in different versions and generations of a product. A modular product architecture also affects supplier relations. Standardization of component interfaces and codification of product specifications disembed relations between manufacturers and suppliers from their social and special context. Suppliers become interchangeable (Baldwin and Clark, 1997) and Silicon Valley manufacturers increasingly outsourced their production to contract manufacturers on other regions and countries, especially China and Taiwan. Yet, to mediate the volatility of modular organized interactions, key suppliers moved their (regional) headquarters to the Silicon Valley. In doing so, the ICT cluster changed its developmental rationale from innovation of new products to local organization of global value chains (Sturgeon, 2003).

Modularization of production affects interconnections between firms within and outside clusters. Resembling processes can be detected in many industries from automobiles to aircraft production. A deeper implementation and broader diffusion of modular production in different industries will therefore affect the general dynamics behind cluster evolution and in doing so the general fragmentation of economic activity in space.

See also: Boundaries and New Organization Forms; Business Models; Diversification, Relatedness, and the New Logic of Co-Creation; Hierarchies and Markets; Innovation; International Business; Interorganizational Relationships and Networks; Modularity and Organizations; Multinational Corporations; Special Economic Zones and Cluster Dynamics: China; Supply Chain Management.

Bibliography

Arthur, W.B., 1994. Increasing Returns and Path Dependence in the Economy. University of Michigan Press, Ann Arbor.

Audretsch, D.B., Feldman, M.P., 1996. Innovative clusters and the industry life cycle. Review of Industrial Organization 11, 253–273.

Baldwin, C., Clark, K., 1997. Managing in an age of modularity. Harvard Business Review 75, 84–93.

Baptista, R., Swann, P., 1998. Do firms in clusters innovate more? Research Policy 27, 525–540.

Bathelt, H., 2001. Regional competence and economic recovery: divergent growth paths in Boston's high technology economy. Entrepreneurship & Regional Development 13, 287–314.

Bathelt, H., Malmberg, A., Maskell, P., 2004. Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. Progress in Human Geography 28, 31–56.

Becattini, G., 2002. From Marshall's to the Italian "industrial districts". A brief critical reconstruction. In: Curzio, A.Q., Fortis, M. (Eds.), Complexity and Industrial Clusters: Dynamics and Models in Theory and Practice. Physica-Verlag, Heidelberg; New York, pp. 82–106.

Boschma, R., Wenting, R., 2007. The spatial evolution of the British automobile industry: does location matter? Industrial and Corporate Change 16, 213–238.
Florida, R., 2005. The world is spiky. The Atlantic Monthly 296, 48–51.
Gordon, I.R., Mccann, P., 2000. Industrial clusters: complexes, agglomeration and/or social networks? Urban Studies 37, 513–532.
Grabher, G., 1993. The weakness of strong ties. The lock-in of regional development in the ruhr area. In: Grabher, G. (Ed.), The Embedded Firm. Routledge, London, pp. 255–277.
Hall, P., Soskice, D. (Eds.), 2001. Varieties of Capitalism: The Institutional Foundations of Comparative Adavantage. Oxford University Press, Oxford.
Klepper, S., 2007. Disagreements, spinoffs, and the evolution of detroit as the capital of the US automobile industry. Management Science 53, 616–631.
Krugman, P., 1991. Geography and Trade. MIT Press, Cambridge MA.
Malmberg, A., Maskell, P., 2006. Localized learning revisited. Growth and Change 37, 1–18.
Marshall, A., 1920. Principles of Economics: An Introductory Volume. Macmillan, London.
Maskell, P., 2001. Towards a knowledge-based theory of the geographical cluster. Industrial and Corporate Change 10, 921–943.
Maskell, P., Bathelt, H., Malmberg, A., 2006. Building global knowledge pipelines: the role of temporary clusters. European Planning Studies 14, 997–1013.
Piore, M.J., Sabel, C.F., 1984. The Second Industrial Divide: Possibilities for Prosperity. Basic Books, New York.
Porter, M.E., 2000. Location, competition, and economic development: local clusters in a global economy. Economic Development Quarterly 14, 15–34.
Porter, M.E., 2003. The economic performance of regions. Regional Studies 37, 549–578.
Robertson, P.L., Langlois, R.N., 1995. Innovation, networks, and vertical integration. Research Policy 24, 543–562.
Romanelli, E., Feldman, M., 2006. Anatomy of cluster development: emergence and convergence in the US human biotherapeutics, 1976–2003. In: Braunerhjelm, P., Feldman, M. (Eds.), Cluster Genesis: Technology-Based Industrial Development. Oxford University Press, Oxdord, pp. 87–112.
Rubini, L., Barbieri, E., Ditomaso, M.R., 2015. This volume. Special economic zones and industrial cluster dynamics. In: Wright, J.D. (Ed.), The International Encyclopedia of the Social and Behavioral Sciences. Elsevier.
Saxenian, A., 1990. Regional networks and the resurgence of Silicon Valley. California Management Review 33, 89–112.
Saxenian, A., 1994. Regional Advantage: Culture and Competition in Silicon Valley and Route 128. Harvard University Press, Cambridge, MA.
Saxenian, A., 1996. Clusters are geographic concentrations of interconnected companies, specialized suppliers, service providers, firms in related industries, and associated institutions (e.g., universities, standards agencies, trade associations) in a particular field that compete but also cooperate. (Porter 2000: 15). Cityscape 2, 41–60.
Steinle, C., Schiele, H., 2002. When do industries cluster? A proposal on how to asses an industrie's propensity to concentrate at a single region or nation. Research Policy 31, 849–858.
Storper, M., 2009. Roepke lecture in economic geography—regional context and global trade. Economic Geography 85, 1–21.
Storper, M., Venables, A.J., 2004. Buzz: face-to-face contact and the urban economy. Journal of Economic Geography 4, 351–370.
Storper, M., Walker, R., 1989. The Capitalist Imperative: Territory, Technology, and Industrial Growth. Basil Blackwell, Cambridge, MA.
Sturgeon, T.J., 2002. Modular production networks: a new American model of industrial organization. Industrial and Corporate Change 11, 451–496.
Sturgeon, T.J., 2003. What really goes-on in Silicon Valley? Spatial clustering and dispersal in modular production networks. Journal of Economic Geography 3, 199–225.
Tappi, D., 2005. Clusters, adaptation and extroversion – a cognitive and entrepreneurial analysis of the Marche Music Cluster. European Urban and Regional Studies 12, 289–307.
Torre, A., 2008. On the role played by temporary geographical proximity in knowledge transmission. Regional Studies 42, 869–889.
Weber, A., 1909. Über den Standort der Industrien. Verlag von J.C.B. Mohr, Tübingen.
Witt, M.A., Redding, G., 2013. Asian business systems: institutional comparison, clusters and implications for varieties of capitalism and business systems theory. Socio-Economic Review 11, 265–300.
Zeller, C., 2004. North Atlantic innovative relations of Swiss pharmaceuticals and the proximities with regional biotech arenas. Economic Geography 80, 83–111.
Zhang, J., Peck, J., 2014. Variegated capitalism, Chinese style: regional models, multi-scalar constructions. Regional Studies 1–27 (online advanced access).
Zucker, L.G., Darby, M.R., 1996. Star scientists and institutional transformation: patterns of invention and innovation in the formation of the biotechnology industry. Proceedings of the National Academy of Sciences 93, 12709–12716.

Industrial Dynamics: Patenting Perspectives

Mei-Chih Hu, National Tsing Hua University, Hsinchu, Taiwan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article is based on the premise that the dynamics of an industry are not only driven by patenting activities at the levels of country, industry, and firm, but have also coevolved along with the various developmental stages of the industrial cycles in each country. Consequently, most merger and acquisition activity, manpower recruitment, patent licensing and lawsuits, and antitrust prosecutions take place in the Trilateral Patent Offices of the United States, Japan, and Europe. The most prolific and fastest-growing patentees (in the United States, Japan, Germany, Korea, Taiwan, and China) are those that are closely linked with information and communication technologies and the biotechnology industries. Emerging policy challenges and future research directions are also elaborated in order to better describe the role of patenting in the evolving industrial dynamics of the twenty-first century.

Introduction

The study of industrial dynamics goes beyond the traditional disciplines of industrial organization studies to encompass a wide range of characteristics of firms within a country or an industry, in terms of size, entry and exit, innovation performance, research and development (R&D) expenditure, and innovation strategies, and it deals with comparisons of one period or geographic region with another (Malerba and Orsenigo, 1996). Industrial dynamics is considered to be among the most interesting and rewarding fields of research, since its nature is knowledge- and strategy-centric, and it is constituted by a number of networks and driven by various types of industrial actors (including new entrants, incumbents, governments, academy, business associations, and regulatory agencies). The ultimate goal is promoting the growth of industrial companies and building stronger competitiveness for them, through the creation and diffusion of knowledge within and across the industry.

Given that knowledge diffusion is critical for building innovation capability in a dynamic industry, the development of knowledge sources, delivery processes, and interactions as dynamic elements has become essential, as has the contribution of technological and institutional actors. Knowledge creation and diffusion is effected through a range of channels, including foreign direct investment (FDI), joint ventures, international trade, technological licensing, imports of capital goods, cooperative research projects, researcher networks, movement of personnel, and the scientific literature. In this complex situation, codified knowledge such as patents can play a role as legal documents, since they contain reliable and traceable records of knowledge creation and diffusion. This makes patenting activity one of the best and most accessible proxies for exploring the evolution of industrial dynamics, especially in knowledge-intensive high-tech industries. This article, based on patent perspectives, is intended to summarize the state of current knowledge on industrial dynamics and to discuss the major challenges presented by current intellectual property systems, through which emerging policy implications and future research directions may be developed.

Worldwide International Patenting Activity

A patent is a temporary monopoly awarded for the commercial exploitation of a newly invented device. The grant of a patent implies that such a device would not appear obvious to a skilled practitioner of the relevant technology. While researchers usually use industrial production, value chain networks, and market share to evaluate an industry, the volume of patenting applied for and approved can be a further important indicator of industrial dynamics.

According to the cooperative Trilateral Patent Office, most patent applications are to the USPTO, the JPO, and the EPO, while worldwide patent demand rose from 2 586 045 in 2000 to 6 082 067 in 2009, which is more than a twofold increase in a decade, corresponding to an average annual increase of about 13%. Within this overall total, there has, however, been a marked difference in patenting propensity between the advanced countries and latecomer countries. First filings accomplished in the advanced countries over the period 2003–09 registered a rise of just 2.6% (USPTO) and 0.97% (EPO) and a fall of 3.8% (JPO). In comparison, patenting activity in the emerging latecomer countries, such as that to China's SIPO, Taiwanese IPO, and Indian IPO exhibit an approximately 20% increasing rate of first filings over the same decade, indicating the rapid diffusion and internationalization of patent knowledge spillovers in these emerging countries.

The patent document is considered to be one of the major indicators of cutting-edge technology, embedding as it does great commercial value (World Intellectual Property Organization (WIPO), 1999). WIPO reported that over 90% of R&D achievements were disclosed in the patent gazettes, while 80% of them were not revealed in academic journals or the technical literature. Moreover, the world's No. 1 patentee, IBM, earns over US$1 billion in royalties per year. Patents and patenting information represent some of a company's most valuable assets and decisive factors in company strategy, protecting R&D achievements, raising research performance, creating profits and new business models, and maintaining competitive advantage in the market.

Patent Statistics as Indicators of Innovation and Industrial Dynamics

While technological innovation cannot be precisely measured, patenting rates have long been considered a well-grounded proxy for level of innovation. Though the sources of patents may introduce bias, the advantages associated with using international patents are described in many prior studies (e.g., Soete, 1987). In addition, a strong link between foreign patent grants and domestic R&D investments, as well as between foreign patents and FDIs demonstrates that patenting activity is an important factor in international trade. While some researchers have emphasized that involvement in global production networks is an important source of knowledge for advancing export performance in late industrializing countries, other studies have explored the importance of the different relationship between technology and international business trade performance (e.g., Dosi et al., 1990). To various degrees, all the literature supports the positive relationships between knowledge, innovation activity, and economic or industry performance in both advanced and latecomer countries.

On the country and industry levels, researchers associated with the National Bureau of Economic Research have developed methodologies for studying the flow of knowledge from one country or industry to another, using patent statistics. In particular, data from patents taken out in the USPTO by both American and foreign firms have enabled researchers to trace the influence of knowledge spillovers and degrees of specialization associated with industrial clustering (Jaffe and Trajtenberg, 1999). Studies have used this USPTO patenting activity to examine the experiences of specific institutions, such as biological resource centers (Furman and Stern, 2004), individual industries, such as solar photovoltaic (Wu and Mathews, 2012), and international knowledge flow, such as that between Taiwan and South Korea (Hu and Jaffe, 2003). Many have argued that only certain types of innovation can be measured, and that many important innovations are not patented, for instance in food and tobacco; some patents are awarded for very modest discoveries (e.g., in basic metals and automotive techniques). Nevertheless, patent activities are still widely recognized as an appropriate indicator in the context of high-tech industries.

On the level of the firm, patent data are a valuable information source allowing decision makers to assess their own technology portfolio in comparison with their competitors within and across the industry, including the ability to assess the attractiveness of technologies, especially those posing a threat, or new opportunities for the existing business; to recognize strategic changes in the firm's competitive environment; to identify and assess external sources for knowledge generation (e.g., merger and acquisitions (M&A) and R&D alliances); to assess the status of patent appropriation in new business areas; to evaluate important market partners, especially customers and suppliers to determine whether the firm's R&D strategy is in alignment with that of its major customers and competent suppliers; and to improve human resource management pertaining to leading inventors in specific technological fields. In summary, integrated patent information can act as a quick and efficient monitoring system for capturing the industrial dynamics through timely review and update of competitive portfolio analysis, technology trends, market segmentations, and business opportunities within and across industries, as shown in Figure 1.

Worldwide Industrial Dynamics: Patenting Perspectives

The US industrial market is widely recognized as the largest and the most technologically advanced in the world and the US patent statistics are reliable indicators because of the general screening procedures imposed by the USPTO. With all their deficiencies, then, patenting rates, and particularly data on what the USPTO calls 'utility patents' (which represent real inventiveness, rather than minor improvements or other 'noise' in the overall data), represent a uniform and time-robust measure of innovative output with an immediate economic significance, since the time and trouble involved in seeking patents in the USPTO system has the effect of filtering out trivial or idiosyncratic innovations. Accordingly, worldwide industrial dynamics will be explored in terms of patenting activity at the levels of country, industry, and firm, as follows.

Country-Level Patenting Comparisons

Many studies examine the extent of innovation by using patent data for advanced countries (mainly in the Organization for

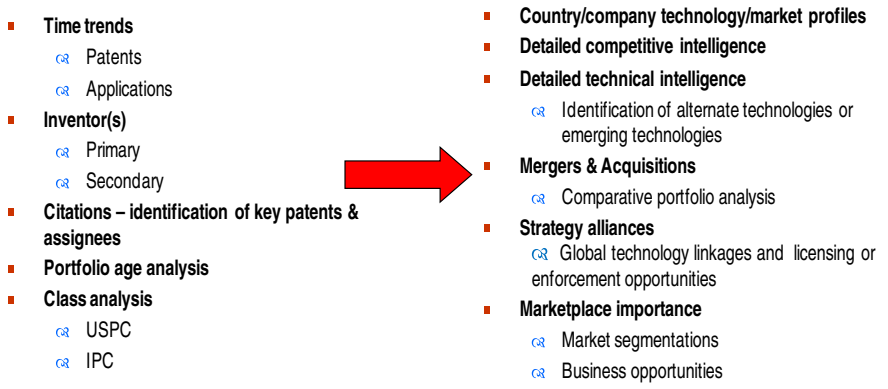

Figure 1 Patent information as monitoring systems for industrial dynamics.

Table 1 Country-level patenting comparisons: G7 vs East Asian 5 and others

	Number of patents				*Annual growth rate %*			
	1963–80	*1981–90*	*1991–2000*	*2001–10*	*1963–80*	*1981–90*	*1991–2000*	*2001–10*
G7								
USA	825 866	403 623	640 541	858 394	0.9	3.0	6.3	3.1
Japan	69 071	134 805	248 862	353 443	**20.5**	**11.2**	5.2	4.3
Germany	85 277	68 350	77 706	103 107	6.6	3.3	3.4	2.8
UK	48 675	24 957	28 065	35 435	3.0	2.2	3.3	2.4
France	33 068	24 524	31 627	35 504	6.7	2.8	3.2	2.7
Canada	19 127	13 893	24 286	35 522	4.5	6.4	6.5	4.4
Italy	11 074	9784	13 287	15 345	6.1	5.6	3.7	1.3
East Asia 5								
Taiwan	245	2841	21 573	60 864	9.2	**30.2**	**20.8**	6.5
Korea	85	738	17 345	60 232	8.0	**44.9**	**33.2**	14.0
Singapore	37	75	853	4171	2.6	**50.1**	**36.7**	12.4
Hong Kong	212	320	976	3032	28	11	16.6	10.2
China	93	181	643	8556	25.4	78.4	12	**38.6**
Reference group								
India	226	127	550	4954	10.3	**36.3**	**21.5**	**24.9**
Switzerland	20 385	12 286	11 966	12 528	4.7	0.6	0.6	3.1
Sweden	12 604	7845	9414	13 162	6.5	0.02	8.3	0.2
Israel	1193	1981	4985	11 869	13.5	11.7	10.8	7.6
Finland	1190	1999	4413	8675	18.3	11.3	8.2	7.6

Source: USPTO, patent statistics general reports, http://www.uspto.gov/about/stats/index.jsp.

Economic Cooperation and Development (OECD)) while tending to ignore the rapidly increasing patenting activity in emerging latecomer countries. A few studies have indicated that the overall patenting activity of the newly industrializing countries had been quite low in early times, but has gone up substantially in the past decade, and that the growth in patenting has reflected a rise in technological capability over time in these countries. The recent trend is more dramatic, as shown in Table 1, which shows that the advanced countries (particularly the United States, Japan, and Germany) enjoyed the highest accumulation of innovation capacity throughout the period of 1963–2010, while Taiwan and Korea are listed, respectively, as the USPTO's fourth and fifth most active patentees in the 2000s. Figure 2 further indicates that Taiwan and Korea, with their highest compound average growth rate in the 2000s, aggressively built their innovation capabilities since the 1990s, while fast-growing China and India are positioned at a lower base, but secured the highest annual growth rate starting in the 2000s. This massive increase in innovative capabilities, in terms of patenting activity, in these latecomer countries reflected the effect of international knowledge spillovers in the direction of the newly industrialized economies and their achievements on the frontier of technological progress in certain areas (Wu and Mathews, 2012). The three advanced countries (the United States, Japan, and Germany) and the three latecomer countries (Korea, Taiwan, and China)

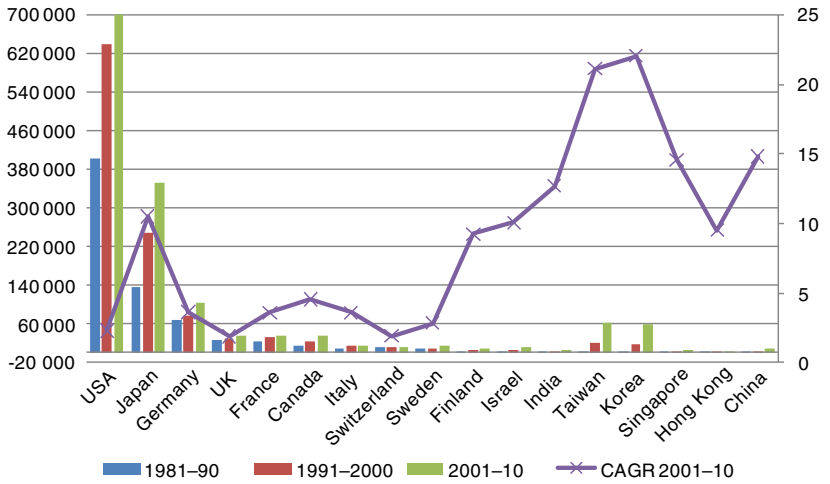

Figure 2 Patenting activity and compound average growth rate: G7 vs East Asian 5 vs reference countries, 1981–2010 by 10 years. Source: USPTO and compiled by this study.

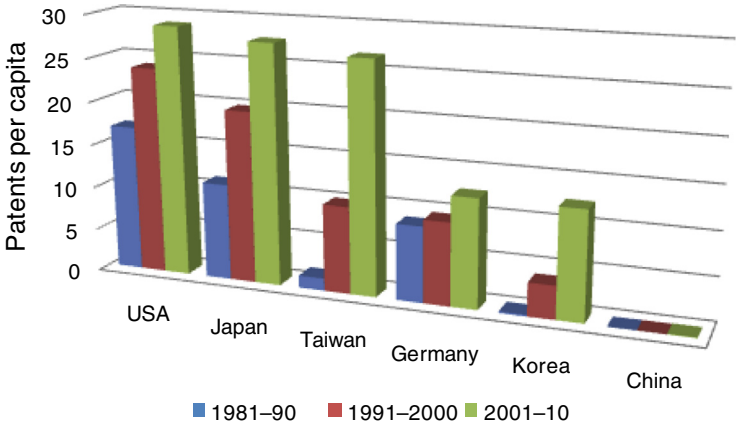

Figure 3 Comparisons of patent per capita in the USPTO, 1981–2010 by 10 years. Source: USPTO and World Bank and calculated by the author.

were the most prolific or fast-growing patentees in the USPTO in the 2000s. (Even though India presents a high annual growth rate during the last three decades (Table 1), this article does not include India in further discussions of its biotechnology specialization area due to its low absolute number of patents. However, see a 2007 special issue for Indian pharmaceutical industry at *Technology Analysis & Strategic Management*, volume 19, issue 5, for the effect of Indian patenting activity on the development of its biotechnology industry.) It is thus worthwhile examining how the patenting activities with respect to these six countries have influenced the global industrial dynamics in their specialized areas.

Since the 1990s, the top five patentees in the USPTO, in terms of patents per capita, have been the United States, Japan, Taiwan, Germany, and Korea, as shown in Figure 3. With country size acting as one of the growth-driven impetuses, China secured the position as the fastest growing patentee in the 2000s, but the rate of patent per capita is overshadowed by its huge population (1.3 billion), implying an even greater opportunity for potential growth and innovation capacity.

Industry-Level Patenting Comparisons

Table 2 demonstrates that the top patenting activities in the Trilateral Offices (USPTO, EPO, and JPO) essentially converge and are concentrated in the high-tech and knowledge-intensive industries, namely medical and biotechnology (A61), semiconductors (H01), communications (H04), and opto-electronics (G01 and G06), while renewable energies are also starting to attract interest. It can also be noted that the patenting activity of EPO is also prolific in the automobile industry (B60, mostly smart cars) since the second half of the 2000s. Indeed, these high-tech and knowledge-intensive industries have been found increasingly important in focusing on intellectual property appropriation (particularly of patents), which then in turn lead to a rapidly changing and transforming dynamic industry. This will be discussed in relation to the situation in the global light-emitting diode (LED) industry.

The above argument is further reflected in country-level patenting activity, as shown in the six prolific patentees in the USPTO (the United States, Japan, Germany, Taiwan, Korea, and China) in Table 3. In terms of international patent classifications (IPCs), the technology specialization is reflected perfectly in the foci of economic growth in each prolific patenting country (the IPC is an international standard for classifying technologies in patent applications; it divides technologies into eight sections with approximately 70 000 subdivisions (see WIPO for details, www.wipo.int/classifications/ipc/en/general/preface.html). The details of IPCs description and search can be accessed at http://v3.espacenet.com/eclasrch?classification=ecla&locale=en_V3). The US patenting activity was most productive in the biotechnology area (IPC: A61K) in the 1990s and switched to a fundamental physics area for multicomputer digital processing (G06F) in the 2000s, while other countries retained their primary focus across the 1990s and 2000s: Japan paid particular attention to semiconductor devices (H01L); Germany takes biopharmaceuticals, especially organic compounds (C07D), as the priority; Taiwan and Korea as latecomers followed Japan in a 'flying geese' model in developing semiconductors (H01L) as national campaigns; and China has focused on general and operational chemical and physical processes (B01J) appropriate to its lower level of economic development.

It is interesting to note that the patenting rates in the semiconductor and information technology sectors in Japan, Taiwan, and Korea effectively retain the flying geese model, in which knowledge spillovers and technology progress is led by Japan and transferred to Taiwan and Korea, the Asian latecomer countries. Indeed, the importance of knowledge diffusion is particularly recognized in these East Asian latecomer countries, by way of their evolving development of strategic high-tech industries, starting with personal computers in the 1980s and progressing to semiconductors in the 1990s, optoelectronics (particularly liquid crystal displays (LCDs)) in the 2000s, and recently, starting in mid-2005, moving onto LED and solar photovoltaic cells (Mathews et al., 2011). Meanwhile, the flying geese model for patenting activity has been the transfer mechanism for China only since the 2000s, when semiconductor (H01L) and information technologies (H05K and G06F) emerged significantly, despite that the industrial innovation priority in that country is focused on general physical and chemical processing equipment (B01J) due to its lower capability for innovation and its restricted economic base. The patenting datasets cited here not only reflect the

Table 2 The top five two-digit IPCs granted in the USPTO, JPO, and EPO; utility patents, 2003–10

	USPTO	*JPO*	*EPO*	*IPC description*
2010	**G06**	**H01**	**A61**	A61: Medical or veterinary science; hygiene
	H01	**H04**	**H04**	B60: Vehicle in general
	H04	G06	H01	C07: Organic chemistry
	A61	A61	B60	C08: Organic macromolecular compounds; their preparation or chemical working-up
	G01	G01	G01	
2009	**G06**	**H01**	**A61**	G01: Measuring; testing
	H01	**H04**	**H04**	G06: Computing; calculating; counting
	H04	G06	H01	H01: Basic electric elements
	G01	A61	B60	H04: Electric communication technique
	A61	G01	G01	
2008	**G06**	**H01**	**A61**	
	H01	**H04**	**H04**	
	H04	G06	H01	
	G01	G01	B60	
	A61	A61	G01	
2007	**G06**	**H01**	**H04**	
	H01	**H04**	**A61**	
	H04	G06	H01	
	G01	G01	B60	
	A61	B60	G01	
2006	**G06**	**H01**	**H04**	
	H01	**H04**	**A61**	
	H04	G01	H01	
	A61	G06	B60	
	G01	A61	G01	
2005	**H01**	**H01**	**A61**	
	G06	**H04**	**H04**	
	H04	G01	C07	
	A61	G06	H01	
	G01	A61	B60	
2004	**H01**	**H01**	**A61**	
	G06	**H04**	**H04**	
	H04	G01	C07	
	A61	G06	H01	
	G01	A61	C08	
2003	**H01**	**H01**	**A61**	
	G06	**H04**	**C07**	
	A61	G01	H04	
	H04	G06	H01	
	G01	C08	G01	

Source: Trilateral Cooperation Office.

evolving industrial competitive advantage in each country but also illustrate the national strategies toward industrial development (Hu and Mathews, 2005, 2008).

Firm-Level Patenting Comparisons

The top patentees listed in the Trilateral Cooperation Office present an even clearer picture to further verify the dominance of information and communication technologies (ICTs) in the global market. The top 10 organizations listed in both USPTO and JPO over the first decade of the 2000s are overwhelmingly dominated by international electronics and communication companies. In comparison, the EPO is also led by the European ICT giants Siemens (No. 1) and Phillips (No. 2) with two exceptions, BASF and Bayer (both German Corporations) from chemical and medical industries, which are listed as No. 3 and No. 10, respectively in 2010. In the USPTO as shown in Table 4, IBM has been the No. 1 patentee for more than 20 years and Samsung become the No. 2 prolific patentee for the first time in 2006 while Microsoft jumped into the top 10 list. Indeed, the patenting rates for both IBM and Samsung are seen to have doubled in a decade, so as to secure the No. 1 and No. 2 positions, respectively. It is interesting to note that, of the Top 10, four are US firms, four are Japanese *keiretsus*, and two are Korean *chaebols*, while no European firm has penetrated to the top echelons throughout the investigation period until 2010.

The significance of patenting activities for securing technology and business model appropriation in the global market has induced both international leaders and technology latecomers to put a great deal of effort into pursuing stronger patent portfolios. While the risks and costs of developing cutting-edge products are rapidly increasing, the efforts of securing intellectual property appropriation are closely tied to

Table 3 The top five three-digit IPCs granted to the prolific patenting countries by the USPTO, 1991–10, by 10 years

	Ranking	*1991–2000*	*2001–10*	
US	1	A61K	G06F	A01N: Chemistry: molecular biology and microbiology
	2	H01L	H01L	A61K: Drug, bioaffecting, and body-treating compositions
	3	C07D	A61K	B01J: Chemical or physical processes in general
	4	G06F	C07D	C07C: Acyclic or carbocyclic compounds
	5	C07C	H04L	C07D: Heterocyclic compounds
Japan	1	H01L	H01L	C08G: Macromolecular compounds obtained
	2	G11B	G11B	G02B: Optical elements and systems
	3	H04N	H04N	G02F: Optical devices
	4	C07D	G06F	G06F: Multicomputer data transferring
	5	G06F	G02B	G11B: Dynamic information storage
Germany	1	C07D	C07D	G11C: Static information storage
	2	C07C	H01L	H01L: Semiconductor or active solid-state devices
	3	A61K	C07C	H01R: Basic electric connectors
	4	C08G	A61K	H04L/H04N: Multiplex communications and telecommunications
	5	A01N	G06F	
Taiwan	1	H01L	H01L	H05K: Electric techniques such as printed circuits
	2	G06F	H01R	
	3	H01R	G06F	
	4	G11C	H05K	
	5	H04N	G11C	
Korea	1	H01L	H01L	
	2	H04N	G11C	
	3	G11B	G02F	
	4	G11C	G11B	
	5	G06F	H04N	
China	1	**B01J**	B01J	
	2	C07C	H01L	
	3	H01L	H05K	
	4	C07D	A61K	
	5	A61K	G06F	

Source: USPTO.

Table 4 The top 10 patentees in the USPTO, by organizations, 2000–10

Rank	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*
1	IBM (2886)	IBM (3411)	IBM (3288)	IBM (3415)	IBM (3248)	IBM (2941)	IBM (3621)	IBM (3125)	IBM (4169)	IBM (4887)	IBM (5866)
2	NEC (2021)	NEC (1953)	Canon (1893)	Canon (1992)	Matsushita (1934)	Canon (1829)	**Samsung** (2451)	Samsung (2723)	Samsung (3502)	Samsung (3592)	Samsung (4518)
3	Canon (1890)	Canon (1877)	Micron (1833)	Hitachi (1893)	Canon (1805)	HP (1790)	Canon (2366)	Canon (1893)	Canon (2107)	Microsoft (2901)	Microsoft (3086)
4	Samsung (1441)	Samsung (1450)	NEC (1821)	Matsushita (1774)	HP (1775)	Matsushita (1688)	**Microsoft** (2229)	Matsushita (1910)	Microsoft (2026)	Canon (2200)	Canon (2551)
5	Lucent (1411)	Matsushita (1440)	Hitachi (1601)	HP (1707)	Micron (1760)	Samsung (1641)	Intel (2099)	Intel (1864)	Intel (1772)	**Panasonic** (1759)	Panasonic (2443)
6	Sony (1385)	Sony (1363)	Matsushita (1544)	Micron (1592)	Samsung (1604)	Micron (1561)	Toshiba (1959)	Microsoft (1637)	Toshiba (1575)	Toshiba (1669)	Toshiba (2212)
7	Toshiba (1232)	Fujitsu (1166)	Sony (1434)	Intel (1353)	Intel (1601)	Intel (1549)	Sony (1771)	Toshiba (1519)	Fujitsu (1475)	Sony (1656)	Sony (2130)
8	Motorola (1196)	Toshiba (1149)	GE (1416)	Philips (1313)	Hitachi (1513)	Hitachi (1271)	HP (1732)	Micron (1476)	Matsushita (1469)	Intel (1534)	Intel (1652)
9	Fujitsu (1147)	Lucent (1109)	HP (1373)	Samsung (1311)	Toshiba (1311)	Toshiba (1258)	Hitachi (1672)	HP (1466)	Sony (1461)	Seiko Epson (1328)	**LG** (1488)
10	Matsushita (1137)	Motorola (778)	Mitsubishi (1328)	Sony (1302)	Sony (1305)	Fujitsu (1154)	Fujitsu (1610)	Sony (1454)	HP (1422)	HP (1269)	HP (1480)

Note: The parenthesis indicates the number of granted patents.
Source: USPTO, various years.

the involved multifaceted character of the innovation that has boomed since the 2000s. Consequently, patenting licensing activities, multifarious patent trolls, and frequent patent litigations have come to be a critical part of the business strategies in the creation of legal and value-creating assets. These resultant business activities not only challenge the usage of patent (intellectual property) for innovation and knowledge creation but also drive the evolution of industrial dynamics (Reichman, 2009; Fink and Maskus, 2005). These challenges derived from industrial dynamics are discussed in the context of global LED industrial development, with the intent that future emerging directions in theory and research can be illuminated.

Case: Evolution of the Global LED Industry Driven by Technological Appropriability (Patents)

The first commercialized low-brightness LED was developed by the US firm General Electric (GE) as early as 1962, but the LED lighting era did not kick off until the US Hewlett–Packard and Japanese Toshiba successfully codeveloped the first high brightness LED in 1991. During the 1990s, the appropriation of LED technology was dominated by first movers in the United States (GE, RCA, and Cree) and Europe (Osram and Phillips). However, the technological appropriation was shifted and shaped by the innovative Japanese latecomer Nichia in the mid-1990s, when it achieved a technological breakthrough for the high-brightness indium gallium nitride blue LED (InGaN, a III–V compound semiconductor material). This technological breakthrough immediately led the industry to move into the introductory stage, while the world's first white LED came onto the market in 1996. This technological breakthrough enabled Nichia to monopolize the global LED market through its strong patent portfolios while winning every patent lawsuit up until 2002. However, the patent portfolio built by Nichia was, to a certain degree, still based on the previous accumulated knowledge created by US and European first movers. In 2002, Nichia's patent lawsuit failed for the first time to prevent infringement by the other four leaders Cree, Phillips Lumileds, Toyota Gosei, and Osram. Thereafter, the five leaders reached cross-licensing agreements with one another in order to build a strong entry barrier against technological latecomers and led the industry move into the earlier growth stage.

The global LED industry remains restricted by the appropriability of technology (through patents) leading to market competition in an oligopoly. To push the industry into the rapid growth stage, a round-table discussion on the LED manufacturing roadmap convened by the US Department of Energy reported in July 2010 (Department of Energy, 2010); the round-table experts, after several discussions, identified seven priority tasks for LED manufacturing, some associated with luminaire manufacturing and others to do with LED chips and packages. These tasks have the immediate goals of increasing emitting efficiency through technological improvement in the manufacturing process (e.g., mechanical/thermal, packaging, and drivers), through the setting of industrial standards in order to drive down the current LED manufacturing costs by 80–90%, by the 2020s (Department of Energy, 2010: p. 6). For those firms at the technological frontier (Cree, Nichia, TG, Phillips, and Osram), these manufacturing capabilities depend neither on their expertise nor on their embedded resources; nevertheless, they are necessary to closely connect the technological fast-followers (the Korean, Taiwanese, and Chinese players who specialize in mass production) with the first movers via technological appropriability. This evolution thus exhibits a double-edged challenge for the LED technological fast followers. On one hand, the latecomer disadvantage is constrained by the intellectual property rights (mainly patents) possessed by technological first movers, such as patented phosphor materials and high-end equipment. On the other hand, this constraint, in its turn, provides an opportunity for technological fast followers to build the subsequent capabilities complementary with those at the technological frontiers (mainly patents and expertise) in the industrial growth stage (Hu, 2012).

Consequently, the global market share of the top five international leaders in LED (Nichia, Osram, Cree, Phillips Lumiled, and Toyota Gosei) has been reduced from 80% in 2004 to 39% in 2009, while that of the top five latecomers (Korean Samsung LED and Seoul Semi, and Taiwanese Epistar, Everlight, and Optotech) increased from 17% in 2007 to 30% in 2010. The increased market share of the technology latecomers is derived through links with the technological appropriability of the five international leaders, and by extending them through OEM licensing, patent licensing, joint ventures, and manpower recruitment, as shown in Figure 4.

It is clear that the current technological regime is dominated by technological leaders in the advanced countries, illustrated by the patenting activities of the USPTO, EPO, and JPO. However, this industrial dynamic is continuously evolving, while a different latecomer strategy (from Korea and Taiwan) has been evident in the development of the LED industry in China. The shortcomings of technological appropriation (through intellectual property), as reflected in Table 2 and Figure 3, have seriously damaged the status of Chinese LED players in the global market. Even so, China's LED industry has been burgeoning, due to government subsidies and the official inauguration of the 12th Five Year Plan for National Economic and Social Development from 2011 to 2015. Taking advantage of the huge domestic market, China's solid government support of a large number of public projects has provided the opportunity for growth for numerous small Chinese LED firms and, since it is focused on the domestic market (without intellectual property issues), defers the effect of intellectual property on technology acquisition, transfer, diffusion, and internalization, while it helps the endogenous Chinese firms to build and accumulate their innovation capabilities and patent portfolios based on setting new domestic industrial standards, product specifications, and application systems. This effect has been demonstrated by the rapid growth of China's patenting activity in the Trilateral patent offices (and particularly in the USPTO, as shown in Table 1 and Figure 2). The goal of China's strategy has been to shift the focal of industrial dynamics away from the existing technological regime built by the international leaders and to mitigate the latecomer disadvantage by creating a new technological regime.

The development of the global LED industry has clearly vindicated the patenting practices explored in the previous sections, where intellectual property appropriation (particularly of patents) has not only led to a rapidly changing and transforming dynamic industry, but has also coevolved with it.

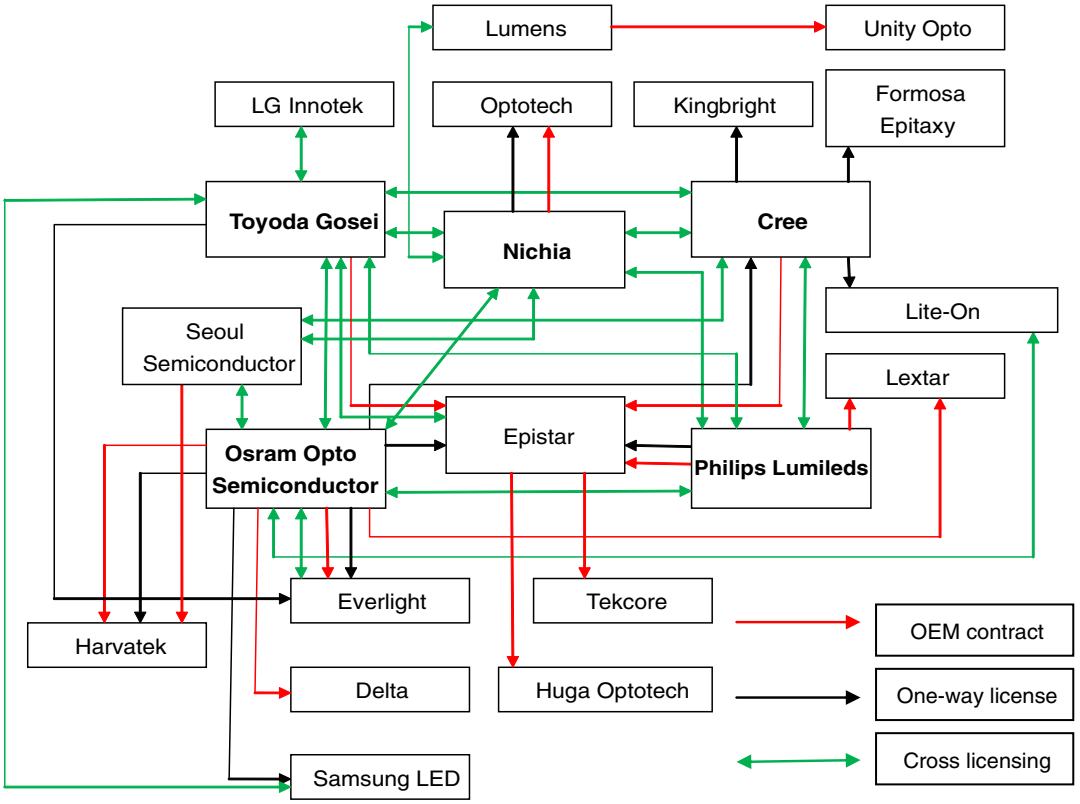

Figure 4 Major players in the global LED industry: linked by technological appropriability, up to the end of 2010. Source: Compiled by the author from various media reports.

Concluding Remarks

The ultimate goal of intellectual property protection is to facilitate economic growth derived from innovations, so as to increase social welfare as a whole. However, due to the differing stages of industrial development in each country, the aggressive series of patenting races in the so-called high-tech industries have challenged the existing system for both the advanced and latecomer countries, as demonstrated in the case of the global LED industry described above. In advanced countries such as the United States, the cumulative cost of patent litigations has compromised the benefit of patented innovations, particularly for upstream research. Reichman (2009) has thus suggested wider use and transparency of publicly funded research results and nonexclusive licensing, in the spirit of antiblocking or public interest, in which the legal schemes for manpower mobility, such as *Non-Compete Covenants*, *Inter-Institutional Agreement*, and the *Material Transfer Agreement* require careful specification. In latecomer countries such as China and India, the intellectual property regime is designed to stimulate industrial development through rapid knowledge diffusion and by building indigenous innovation capability (Kim et al., 2012). For example, compulsory licensing is enforced for those inventors who are unable or unwilling to practice their inventions in China, while minor adaptive invention is outlawed in India in order to ensure the novelty of innovation which successfully upgraded its bio-pharmaceutical industry onto the global market since the 2000s. The link between patents and trade policy was made explicit by the international TRIPS framework, adopted and enforced as part of the creation of the WTO in 2001.

These examples demonstrate that the various development stages of industrial dynamics evolve with the industrial cycle while the patenting and industrial activities between technological leaders and latecomers are closely linked, as shown in Figure 5. During the embryonic and introductory stages, IP wars emerge among various types of technology breakthrough and competition before specific industrial standards can be established by technological leaders to steer the industry through its initial growth stages. Based on the guidelines, dominating industrial standards, building stronger IP portfolios in the patent race, IP licensing, M&A activity, and manpower recruitment have become the main strategies of technological leaders during the industrial growth stage, while this competition has extended the existing technological regime and has transformed it from a cost basis to a quality basis, with an abundance of differentiated products systematically developed by the technological latecomers. The surviving companies then form an oligopolistic industry in the latter growth stage, while IP lawsuits, antitrust, and fair-trade issues become critical elements in driving the next generation of technology and developing system services and platform strategies in the mature stage. At this stage, the latecomers (with limited technological appropriability) must then either focus on new technology generation and applications (as in Korea and Taiwan) or try to shift the focal of

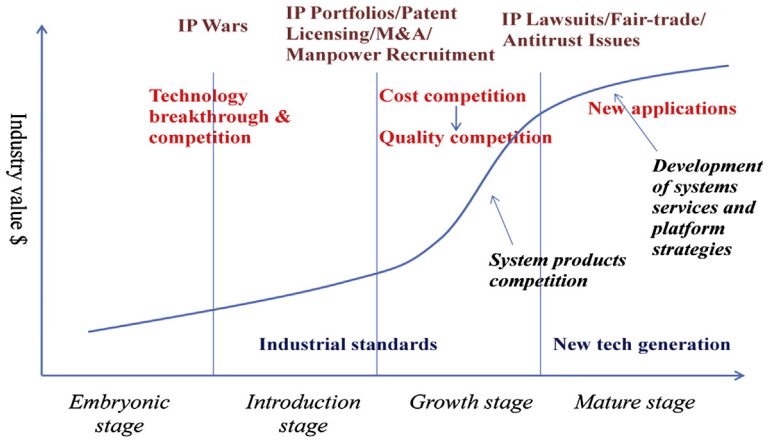

Figure 5 Industrial dynamics evolved over the industrial cycle and intellectual property system.

industrial dynamics away from the existing technological regime (as in China) in order to compete with the incumbents and international leaders. Driven by the appropriability of intellectual property rights (and especially patents), these coevolving and cyclical industrial dynamics have been evidenced in many studies (for example, Hu (2012)) and are especially apparent in the knowledge-intensive ICT and biotechnological industries as found in the Trilateral patent offices.

Even though the role of intellectual property is considered and strategized differently for technological leaders and latecomers in the context of industrial- and firm-level dynamics, the changing modes of innovation across the high-tech industrial dynamics suggest that policy makers at country-level should pay attention not only to the traditional approach of merely building physical knowledge infrastructure (e.g., hardware Internet accessibility and competition regulation) but should also explore socially accessible knowledge mechanisms (e.g., fluid manpower flows, open data for publicly funded research, and networks, as shown in the LED case) along the journey from innovation to the market. Based on the WTO framework, this may include the secondary design of an adaptive intellectual property legal system so as to meet the indigenous needs of different stages of industrial development over time, as well as establishing and facilitating transparent markets for technology and open collaboration. Along with this, future research may need to further explore the extent and magnitude of patent licensing, patent litigation, and fair-trade issues over the various development stages of the industrial cycle, particularly for their different impact toward technological leaders and latecomers, as demonstrated in last section.

See also: Cyclical Industrial Dynamics; Innovation; International Trade: Economic Integration; Organizational Path Dependence.

Bibliography

Dosi, G., Pavitt, K., Soete, L., 1990. The Economics of Technical Change and International Trade. Columbia University Press, New York.

Fink, C., Maskus, K., 2005. Intellectual Property and Development. World Bank: Oxford University Press.

Furman, J.L., Stern, S., 2004. A Penny for Your Quotes?: The Impact of Biological Resource Centers on Life Sciences Research (Chapter 4) in Biological Resource Centers: Knowledge Hubs for the Life Sciences. Brookings Institution Press, Washington, DC.

Hu, A.G.Z., Jaffe, A.B., 2003. Patent citations and international knowledge flow: the cases of Korea and Taiwan. International Journal of Industrial Organization 21 (6), 849–880.

Hu, M.C., 2012. Technological innovation capabilities in the thin film transistor-liquid crystal display industries of Japan, Korea, and Taiwan. Research Policy 41 (3), 541–555.

Hu, M.C., Mathews, J.A., 2005. National innovative capacity in East Asia. Research Policy 34 (9), 1322–1349.

Hu, M.C., Mathews, J.A., 2008. China's national innovative capacity. Research Policy 37 (9), 1465–1479.

Jaffe, A., Trajtenberg, M., 1999. International knowledge flows: evidence from patent citations. Economics of Innovation and New Technology 8 (1–2), 105–136.

Kim, Y.K., Lee, K., Park, W., Choo, K., 2012. Appropriate intellectual property protection and economic growth in countries at different levels of development. Research Policy 41 (2), 358–375.

Malerba, F., Orsenigo, L., 1996. The dynamics of the evolution of industries. Industrial and Corporate Change 5, 51–87.

Mathews, J.A., Hu, M.C., Wu, C.Y., 2011. Fast-follower industrial dynamics: the case of Taiwan's emergent solar photovoltaic industry. Industry and Innovation 18 (2), 177–202.

Reichman, J.H., 2009. Intellectual property in the 21st century: will the developing countries lead or follow? Houston Law Review 46 (4), 1116–1185.

Soete, L., 1987. The impact of technological innovation on international trade patterns: the evidence reconsidered. In: Freeman, C. (Ed.), Output Measurement in Science and Technology. North-Holland, Amsterdam.

World Intellectual Property Organization (WIPO), 1999. Patent Documents as a Source of Information for R&D Organizations and for Inventors. Document Prepared by Mr. Christopher Kiige, Examination Officer. African Regional Intellectual Property Organization (ARIPO), Harare.

Wu, C.Y., Mathews, J.A., 2012. Knowledge flows in the solar photovoltaic industry: insights from patenting by Taiwan, Korea, and China. Research Policy 41 (3), 524–540.

Industrial Ecology

Reid Lifset and Thomas E Graedel, School of Forestry & Environmental Studies, New Haven, CT, USA

Published by Elsevier Ltd.

Abstract

Industrial ecology is a rapidly growing field that examines the flow of materials and energy at various scales as part of the study and pursuit of sustainable production and consumption. Originating in part from an analogy to resource flows in ecosystems with an emphasis on the environmental gains from cyclical use of resources, the field emphasizes the importance of a systems perspective in environmental analysis, management, and policy. It includes tools such as life cycle assessment (LCA), material flow analysis (MFA), input-output analysis, and design for environment (DfE). Concepts such as industrial symbiosis, eco-efficiency, dematerialization, urban metabolism are central to the field.

Setting out the goals and boundaries of an emerging field is a hapless task. Set them too conservatively and the potential of the field is thwarted. Set them too expansively and the field loses its distinctive identity. Spend too much time on this task and scarce resources may be diverted from making concrete progress in the field.

But in a field with a name as provocative and oxymoronic as industrial ecology, the description of the goals and definitions is crucial. Hence this introductory chapter describes the field of industrial ecology, identifying its key topics, characteristic approaches, and tools. The objective is to provide a map of the endeavors that comprise industrial ecology and how those endeavors relate to each other. In doing so, we seek to provide a common basis of discussion, allowing us then to delve into more conceptual discussions of the nature of the field.

No field has unanimity on goals and boundaries. A field as new and as ambitious as industrial ecology surely has a long way to go to achieve even a measure of consensus on these matters, but, as we hope this article shows, there is much that is coalescing in research, analysis, and practice.

Defining Industrial Ecology

The very name *industrial ecology* conveys some of the content of the field. Industrial ecology is *industrial* in that it focuses on product design and manufacturing processes. It views firms as agents for environmental improvement because they possess the technological expertise that is critical to the successful execution of environmentally informed design of products and processes. Industry, as the portion of society that produces most goods and services, is a focus because it is an important but not exclusive source of environmental damage.

Industrial ecology is *ecological* in at least two senses. As argued in the seminal publication by Frosch and Gallopoulos (1989) that did much to coalesce this field, industrial ecology looks to nonhuman 'natural' ecosystems as models for industrial activity.[1] This is what some researchers have dubbed the 'biological analogy' (Wernick and Ausubel, 1997; Allenby and Cooper, 1994). Many biological ecosystems are especially effective at recycling resources and thus are held out as exemplars for efficient cycling of materials and energy in industry. The most conspicuous example of industrial reuse and recycling is an increasingly famous industrial district in Kalundborg, Denmark (Ehrenfeld and Gertler, 1997: Chapter 27). The district contains a cluster of industrial facilities including an oil refinery, a power plant, a pharmaceutical fermentation plant, and a wallboard factory. These facilities exchange byproducts and what would otherwise be called wastes. The network of exchanges has been dubbed 'industrial symbiosis' as an explicit analogy to the mutually beneficial relationships found in nature and labeled as symbiotic by biologists.

Second, industrial ecology places human technological activity – industry in the widest sense – in the context of the larger ecosystems that support it, examining the sources of resources used in society and the sinks that may act to absorb or detoxify wastes. This latter sense of 'ecological' links industrial ecology to questions of carrying capacity and ecological resilience, asking whether, how, and to what degree technological society is perturbing or undermining the ecosystems that provide critical services to humanity. Put more simply, economic systems are viewed, not in isolation from their surrounding systems, but in concert with them.

Robert White, the former president of the US National Academy of Engineering, summarized these elements by defining industrial ecology as … "the study of the flows of materials and energy in industrial and consumer activities, of the effects of these flows on the environment, and of the influences of economic, political, regulatory, and social factors on the flow, use, and transformation of resources" (White, 1994).

This broad description of the content of industrial ecology can be made more concrete by examining the core elements or foci in the field:

- the biological analogy,
- the use of systems perspectives,
- the role of technological change,
- the role of companies,
- dematerialization and eco-efficiency, and
- forward-looking research and practice.

The Biological Analogy

The biological analogy has been applied principally at the level of facilities, districts, and regions, using notions borrowed from ecosystem ecology regarding the flow and especially the cycling

of materials, nutrients, and energy in ecosystems as a potential model for relationships between facilities and firms. The archetypal example is the industrial symbiosis in Kalundborg, but the search for other such arrangements and even more conspicuously the effort to establish such symbiotic networks is emblematic of industrial ecology – so much so that many with only passing familiarity of the field have mistakenly thought that industrial ecology focused only on efforts to establish eco-industrial parks.

This analogy has been posited more generically as well, not merely with respect to geographically adjacent facilities. Graedel and Allenby (1995) have offered a typology of ecosystems varying according to the degree to which they rely on external inputs (energy and materials) and on release of wastes to an external environment. Expressed another way, the ecosystems vary according to the linearity of their resource flows as shown in **Figure 1**: type I is the most linear and reliant on external resources and sinks; type III stands at the other extreme, having the greatest degree of cycling and least reliance on external resources and sinks. The efficient cycling of resources in a biological system is held out as an ideal for industrial systems at many scales. This framework thus connects the biological analogy to strong emphasis in industrial ecology on the importance of closing material cycles or 'loop closing.'

The biological analogy has been explored in other ways. The ecological analogy has, for example, been applied to products as a source of design inspiration (Benyus, 1997), as a framework for characterizing product relationships (Levine, 1999), and as a model for organizational interactions in technological 'food webs' at the sector or regional levels (Hardy and Graedel, 2002; Frosch et al., 1997).

The analogy to ecology is suggestive in other respects (Ehrenfeld, 1997). It points to the concepts of community and diversity and its contribution to system resilience and stability as fundamental properties of ecosystems – and as possible models of a different sort for industrial activity. These dimensions of the analogy may point to ways to integrate organizational aspects of environmental management more deeply into the core of industrial ecology, but they have not been as extensively explored as the use of ecosystem ecology with its emphasis on flows and cycling of resources. As Andrews (2000) points out, there are long-standing bodies of scholarship that apply the ecological notions directly to social, as opposed to technological, dimensions of human activity including organizational, human, and political ecology. The biological analogy is not confined to ecological similes. A more quantitative embodiment of the biological analogy is the metabolic metaphor that informs materials flow analysis (see below) by analogizing firms, regions, industries, or economies with the metabolism of an organism (Ayres and Simonis, 1994; Fischer-Kowalski, 1998; Fischer-Kowalski and Hüttler, 1998). Whether or not there is a significant difference between the ecological and metabolic metaphors is a matter of friendly dispute. For one view, see Erkman (1997).

Systems Perspective

Industrial ecology emphasizes the critical need for a systems perspective in environmental analysis and decision making. The goal is to avoid narrow, partial analyses that can overlook important variables and, more importantly, lead to unintended consequences. The systems orientation is manifested in several different forms:

- use of a life cycle perspective,
- use of materials and energy flow analysis,

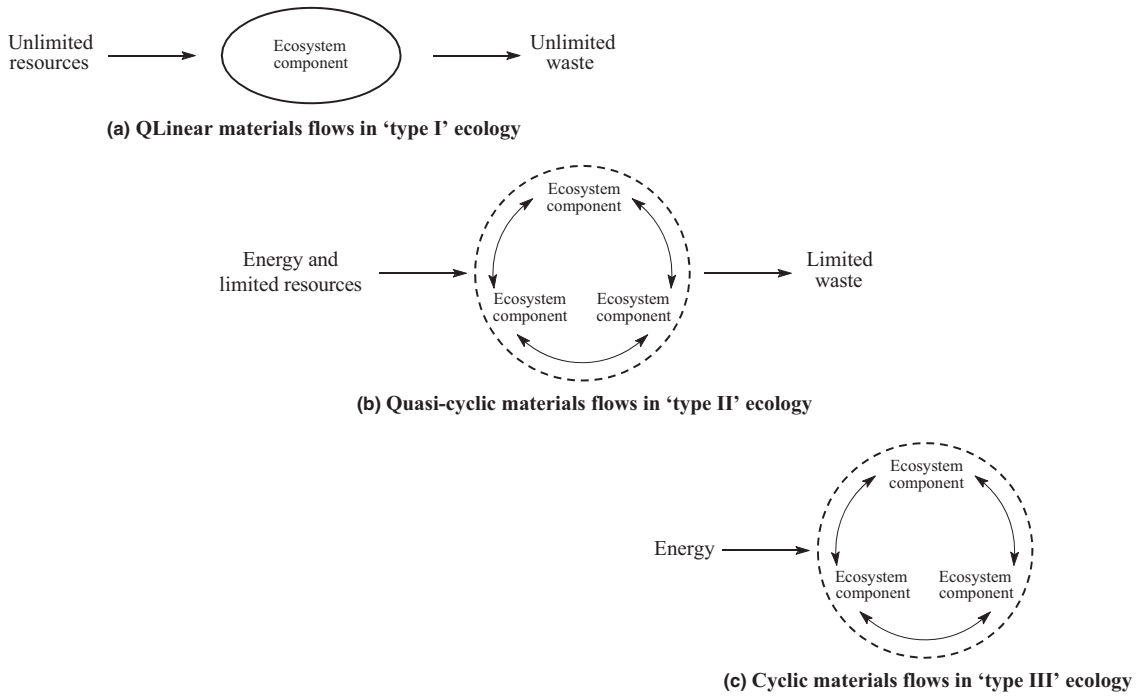

Figure 1 Typology of ecosystems.

- use of systems modeling, and
- sympathy for multidisciplinary and interdisciplinary research and analysis.

The effort to use a life cycle perspective, that is, to examine the environmental impacts of products, processes, facilities, or services from resource extraction through manufacture to consumption and finally to waste management, is reflected both in the use of formal methods such as life cycle assessment (LCA) and in attention to approaches that imply this cradle-to-grave perspective and apply it in managerial and policy settings as well as in research contexts. This latter group includes product chain analysis (Wisberg and Clift, 1999), integrated product policy (IPP, also known as product-oriented environmental policy) (Jackson, 1999), greening of the supply chain (Sarkis, 1995), and extended producer responsibility (EPR) (Lifset, 1993).

Analysis of industrial or societal metabolism, that is, the tracking of materials and energy flows on a variety of scales is also motivated by a system orientation. Here reliance of research in industrial ecology on mass balances – making sure that inputs and outputs of processes add up in conformance with the first law of thermodynamics – reflects an effort at comprehensiveness. Because of the use of mass balances on these different scales, industrial ecology often involves the mathematics of budgets and cycles, and stocks and flows. By tracking chemical usage in a facility (Reiskin et al., 1999), nutrient flows in a city (Björklund et al., 1999), flows of heavy metals in river basins (Stigliani et al., 1994), or bulk materials in national economies (Adriaanse et al., 1997), industrial ecology seeks to avoid overlooking important uses of resources and/or their release to the environment. The tracking of materials and energy is sometimes embedded in the consideration of natural, especially biogeochemical cycles and of how anthropogenic activities have perturbed those flows. For example, the study of anthropogenic perturbations of the nitrogen cycle is an important contribution of industrial ecology (Ayres et al., 1994).

This same effort to examine human–environment interaction from a holistic perspective is manifested in formal systems modeling including dynamic modeling (Ruth and Harrington, 1997), use of process models (Diwekar and Small, 1998), and integrated energy, materials, and emissions models such as MARKAL MATTER (2000) and integrated models of industrial systems and the biosphere (Alcamo et al., 1994). Such systems modeling not only increases the comprehensiveness of environmental analysis; it can also capture some of the interactions among the factors that drive the behavior of the system being studied (for example, Isaacs and Gupta, 1997). Conceptual discussions of the nature of industrial ecology and sustainable development have highlighted the importance of nonlinear behavior in human and environmental systems and argued that chaos theory and related approaches hold out potential for the field (Ruth, 1996; Allenby, 1999a), but little such work has been done to date.

Finally, the imperative for systems approaches is also reflected in a sympathy for the use of techniques and insights from multiple disciplines (Lifset, 1998a; Graedel, 2000). There have been some notable successes (Carnahan and Thurston, 1998; van der Voet et al., 2000), but multidisciplinary analysis – where several disciplines participate but not necessarily in an integrative fashion – is difficult and interdisciplinary analysis – where the participating disciplines interact and shape each other's approaches and results – is even more so. Interdisciplinarity remains an important challenge for not only industrial ecology, but all fields.

Technological Change

Technological change is another key theme in industrial ecology. It is a conspicuous path for pursuing the achievement of environmental goals as well as an object of study (Ausubel and Langford, 1997; Grübler, 1998; Norberg-Bohm, 2000; Chertow, 2001). In simple, if crude, terms, many in the field look to technological innovation as a central means of solving environmental problems. It should be noted, however, that while that impulse is shared widely within the field, agreement as to the degree to which this kind of innovation will be sufficient to solve technological problems remains a lively matter of debate (Ausubel, 1996; Graedel, 2000).

Ecodesign (or design for environment – DFE) is a conspicuous element of industrial ecology. By incorporating environmental considerations into product and process design *ex ante*, industrial ecologists seek to avoid environmental impacts and/or minimize the cost of doing so. This is technological innovation at the microlevel, reflecting technological optimism and the strong involvement of academic and professional engineers. Ecodesign frequently has a product orientation, focusing on the reduction in the use of hazardous substances, minimization of energy consumption, or facilitation of end-of-life management through recycling and reuse. Implicitly, ecodesign relies on the life cycle perspective described earlier by taking a cradle-to-grave approach. Increasingly, it also strives for a systems approach, not only by considering impacts throughout the product life cycle, but also by employing comprehensive measures of environmental impact (Keoleian and Menerey, 1994).

Ecodesign is complemented by research that examines when and how technological innovation for environmental purposes is most successful in the market (Preston, 1997; Chertow, forthcoming). The focus on technological change in this field also has a macro version, examining whether technological change is good for the environment or how much change (of a beneficial sort) must be accomplished in order to maintain environmental quality. Here the IPAT equation (*I*mpact *P*opulation × *A*ffluence × *T*echnology) has provided an analytical basis for parsing the relative contributions of population, economic growth (or, viewed in another way, consumption) and technology on environmental quality (Wernick et al., 1997b; Lifset, 2000; Chertow, 2001). The equation provides a substantive basis for discussion of questions of carrying capacity implicit in the definition of industrial ecology offered earlier.

Role of Companies

Business plays a special role in industrial ecology in two respects. Because of the potential for environmental improvement that is seen to lie largely with technological innovation, businesses as a locus of technological expertise are an important agent for accomplishing environmental goals. Further,

some in the industrial ecology community view command-and-control regulation as importantly inefficient and, at times, as counterproductive. Perhaps more significantly, and in keeping with the systems focus of the field, industrial ecology is seen by many as a means to escape from the reductionist basis of historic command-and-control schemes (Ehrenfeld, 2000a). Regardless of the premise, a heightened role for business is an active topic of investigation in industrial ecology and a necessary component of a shift to a less antagonistic, more cooperative and, what is hoped, a more effective approach to environmental policy (Schmidheiny, 1992).

This impulse to view business as a 'policy-maker rather than a policy-taker' (Socolow, 1994, p. 12) is reflected in a diverse set of analyses and initiatives that explore the efficacy of beyond-compliance environmental strategies and behavior. These include product take back (Davis, 1997), microeconomic rationales for beyond-compliance behavior (Reinhardt, 1999), corporate environmental innovation pursued to maintain autonomy (Sharfman et al., 1998), corporate strategy and sustainable development (Hart and Milstein, 1999), and macro-level analyses of the effectiveness of voluntary policy schemes (Harrison, 1998).

Dematerialization and Eco-Efficiency

Moving from a type I to a type II or III ecosystem entails not only closing loops, but using fewer resources to accomplish tasks at all levels of society. Reducing resource consumption and environmental releases thus translates into a cluster of related concepts: dematerialization, materials intensity of use, decarbonization, and eco-efficiency. Dematerialization refers to the reduction in the quantity of materials used to accomplish a task; it offers the possibility of decoupling resource use and environmental impact from economic growth. Dematerialization is usually measured in terms of mass of materials per unit of economic activity or per capita and typically assessed at the level of industrial sectors, regional, national, or global economies (Wernick et al., 1997a; Adriaanse et al., 1997). Decarbonization asks the analogous question about the carbon content of fuels (Nakicenovic, 1997). Inquiry in this arena ranges from analysis of whether such reductions are occurring (Cleveland and Ruth, 1998), whether dematerialization per se (that is, reduction in mass alone) is sufficient to achieve environmental goals (Reijnders, 1997) and what strategies would be most effective in bringing about such outcomes (Weizsacker et al., 1997). The intersection between investigation of dematerialization on the one hand, and other elements of industrial ecology such as industrial metabolism with its reliance on the analysis of the flows of materials on the other is clear. There is also overlap with industrial ecology's focus on technological innovation. This is because investigations of dematerialization often lead to questions about whether, at the macro- or sectoral level, market activity and technological change autonomously bring about dematerialization (Cleveland and Ruth, 1998) and whether dematerialization, expressed in terms of the IPAT equation, is sufficient to meet environmental goals.

At the firm level, an analogous question is increasingly posed as a matter of ecoefficiency, asking how companies might produce a given level of output with reduced use of environmental resources (Fussler, 1996; OECD, 1998; DeSimone et al., 1997). Here, too, the central concern is expressed in the form of a ratio: output divided by environmental resources (or environmental impact). The connection between this question and industrial ecology's focus on the role of the firm and the opportunities provided through technological innovation is conspicuous as well.

Forward-Looking Analysis

One final element of this field is worth noting. Much of research and practice in industrial ecology is intentionally prospective in its orientation. It asks how things might be done differently to avoid the creation of environmental problems in the first place, avoiding irreversible harms and damages that are expensive to remedy. Ecodesign thus plays a key role in its emphasis on anticipating and designing out environmental harms. More subtly, the field is optimistic about the potential of such anticipatory analysis through increased attention to system-level effects, the opportunities arising from technological innovation and from mindfulness of need to plan and analyze in and of itself. This does not mean that history is ignored. Industrial metabolism, for example, pays attention to historical stocks of materials and pollutants and the role that they can play in generating fluxes in the environment (Ayres and Rod, 1986). However, industrial ecology does not emphasize remediation as a central topic in the manner of much of conventional environmental engineering.

Putting the Elements Together

There are (at least) two ways in which these themes and frameworks can be integrated into a larger whole. One is to view industrial ecology as operating at a variety of levels (**Figure 2**): at the firm or unit process level, at the interfirm, district, or sector level and finally at the regional, national, or global level. While the firm and unit process are important, much of industrial ecology focuses at the interfirm and interfacility level, in part, as described above, because a systems perspective emphasizes unexpected outcomes – and possibly environmental gains – to be revealed when a broader scope is used and because pollution prevention, a related endeavor, has already effectively addressed many of the important issues at the firm, facility, or unit process level.

Another way to tie the elements together is to see them, as in **Figure 3**, as reflecting the conceptual or theoretical aspects of industrial ecology, on the one hand, and the more concrete, application-oriented tools and activities on the other. In this framework, many of the conceptual and interdisciplinary aspects of the field comprise the left side of the figure, while the more practical and applied aspects appear on the right side.

The Goals of Industrial Ecology

Given this overview of the elements of industrial ecology, it is possible to entertain more complicated questions about this field. One set of especially notable and knotty questions revolve around the goals of industrial ecology. Clearly, the field is driven by concerns about human impact on the biophysical environment. Put simplistically, the goal is to improve and

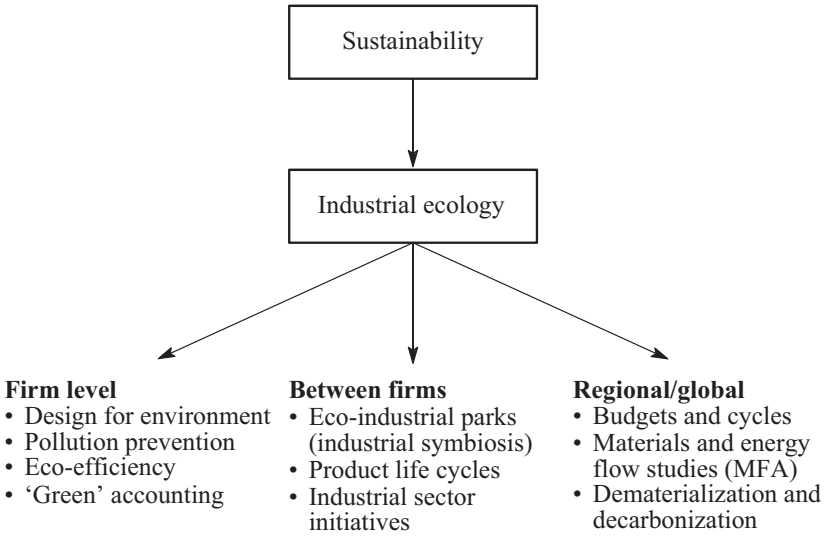

Figure 2 The elements of industrial ecology seen as operating at different levels.

maintain environmental quality. Just as clearly, such a statement of goals does not begin to speak to the multiple dimensions of the research or practice in this field.

Reducing Risk versus Optimizing Resource Use

Industrial ecology emphasizes the optimization of resource flows where other approaches to environmental science, management, and policy sometimes stress the role of risk. For example, pollution prevention (P2) (also known as cleaner production or CP) emphasizes the reduction of risks, primarily, but not exclusively, from toxic substances at the facility or firm level (Allen, 1996). Underlying this focus is an argument that only when the use of such substances is eliminated or dramatically reduced can the risks to humans and ecosystems be reliably reduced. In contrast, industrial ecology takes a systems view that typically draws the boundary for analysis more broadly – around groups of firms, regions, sectors, and so on – and asks how resource use might be optimized, where resource use includes both materials and energy (as inputs) *and* ecosystems and biogeochemical cycles that provide crucial services to humanity (Ayres, 1992). In concrete terms, this means industrial ecology will sometimes look to recycling where P2 will emphasize prevention (Oldenburg and Geiser, 1997). The differences between industrial ecology and P2 are not irreconcilable either conceptually or practically (van Berkel et al., 1997). In conceptual terms, P2 can be seen as a firm-level approach that falls under the broader rubric of industrial ecology (as shown in **Figure 2**). In concrete terms, the difference in actual practices by operating entities may not be great, although careful empirical work documenting how these two frameworks have differed in shaping decision making has not

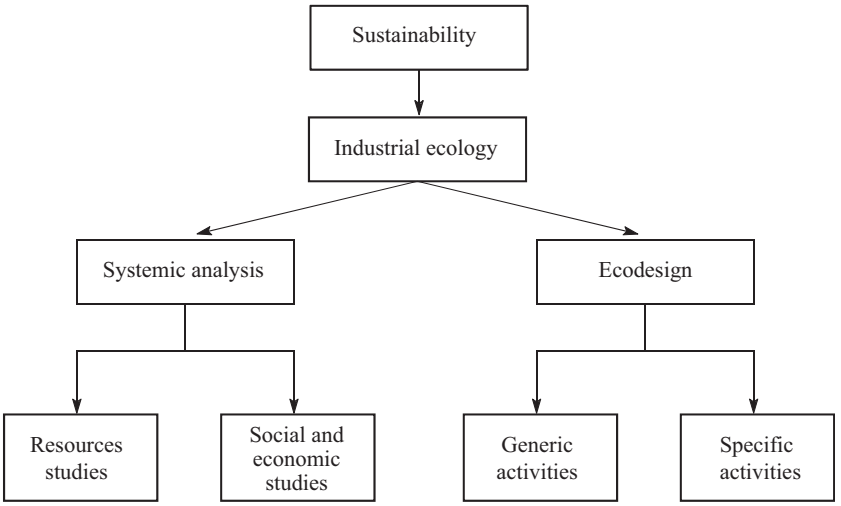

Figure 3 Industrial ecology conceptualized in terms of its system-oriented and application-oriented elements.

been conducted. However, some interesting analysis has been conducted of the risks posed by the recycling of hazardous materials, asking whether it is indeed possible to recycle such substances in an environmentally acceptable manner (Socolow and Thomas, 1997; Karlsson, 1999).

This is not the only way in which industrial ecology differs from allied fields in its orientation toward risk. The focus of industrial ecology on the flows of anthropogenic materials and energy is not often carried further than the point of release of pollutants into the environment. In contrast, much of traditional environmental science focuses precisely on the stages that follow such release – assessing the transport, fate, and impact on human and nonhuman receptors. Similarly, risk assessment and environmental economics focus on the damages to humans and ecosystems, only sometimes looking upstream to the source of pollutants and the human activities that generate them. In this respect, industrial ecology can be seen as providing a complementary emphasis to these fields by concentrating on detailed and nuanced characterization of the sources of pollution. In a related vein, research in industrial ecology often examines perturbations to natural systems, especially biogeochemical cycles, arising from anthropogenic activities. The impacts of such perturbations can be construed in terms of risks to human health and economic well-being as well as to ecosystems, but the analysis of perturbations differs from the manner in which risk assessment – typically focused on threats to human health – is often conducted. This is not to suggest that industrial ecology ignores questions of risk, fate and transport or environmental endpoints. The intense work on methodologies for life cycle impact assessment (Udo de Haes, 1996) is but one example of the field's efforts to systematically incorporate questions of environmental impact. Further, there is work in the field that integrates fate and transport into such analyses (Potting et al., 1998; Scheringer et al., 1999).

Another aspect of the focus on flows and releases rather than damages and endpoints is that the threats posed by releases – especially of persistent pollutants – endure and the receptors can change in a manner that later causes harms that may not be captured in a typical risk assessment. For example, cadmium deposition to agricultural soils that takes place as a result of naturally occurring cadmium contamination of phosphate fertilizers may not cause significant human health or ecological damage as long as fields are limed and thereby kept alkaline. If the fields are taken out of production, liming is likely to end. Soil pH will thereby increase, and cadmium may become biologically available and environmentally damaging (Stigliani and Anderberg, 1994: Chapter 40).

Positive and Normative Analysis

One apparent tension related to the goals of industrial ecology relates to whether the field is positive (descriptive) or normative (prescriptive). If it is positive, then industrial ecology seeks to describe and characterize human–environment interactions, but not necessarily to alter them. On the other hand, if industrial ecology is normative, then some degree of human or environmental betterment is intrinsic to the goals of the field. This tension is reflected in multiple meanings accorded to key terms in the field. For example, the phrase 'industrial ecosystem' refers to facilities or industries that interact in a biophysical sense. Often it is a label for industrial districts like Kalundborg, where residuals are exchanged among colocated businesses. Leaving aside an especially loose usage that denotes any group of facilities, firms, or industries, the question arises as to whether an industrial ecosystem necessarily refers to a desirable arrangement – where, for example, the participating firms extensively exchange residuals and thereby minimize releases of pollutants into the environment – or to a neutral description of a network of firms, which might constitute either good industrial ecosystems (with considerable closing of loops and little pollution) or bad industrial ecosystems (with linear flows of resources and large amounts of pollution)[2]. The point is not the ambiguity in the terminology, but the difference in the emphasis that it reflects. Are there desirable end states that are integral to the notion of industrial ecology? Can one be an 'industrial ecologist' if one does not think that, more often than not, the closing of materials loops brings about environmental improvement? Is it necessary to think that the environmental situation is quite severe to engage in research in or the application of industrial ecology?

The ambiguity, in general, of the boundary between positive and normative endeavors plays a role here. Clearly, some fields exist at one pole or the other. Physics seeks to describe the physical universe without reference to ethical or other prescriptive principles. Theology, on the other hand, is obviously concerned with what ought to be. Economics, like many social sciences, represents a complex middle ground. On the one hand, it conceives of itself as value neutral and engaged in the study of markets and the allocation of scarce resources. Yet, on the other, the field frequently offers advice of the sort 'if the goal is x, then the appropriate choice is y.' In this conception, the assertion of the value or importance of 'x' originates from outside the discipline, maintaining the apparent value neutrality of the analysis. But economists typically argue that the goal should be maximization of utility (at the individual level) and of social welfare (at the societal level) in a manner that sharply constrains what 'x' might be and carrying with it an implicit set of value choices. Further, the analysis that economics uses to deduce y from x is argued by many to be value laden (see, for example, Frank et al., 1993).

Yet this tension may be overstated. Contemporary engineering sciences fuse the positive and the normative without destroying the distinctiveness of each (Ehrenfeld, 2000a). In practice, most industrial ecologists appear to enter the field out of concern for the potential environmental implications of production and consumption and the opportunities for improvement. It is thus regarded by most practitioners – those operating on the right side of **Figure 3** – as a normative endeavor, one that is informed by the positive analysis generated by the activities on the left side of the figure. This characterization does not resolve all disputes about the normative status of field (Allenby, 1999b; Boons and Roome, 2000), but it does, we think, narrow the purview of the disagreements to questions about whether the field has teleological (i.e., goal-oriented) characteristics (Ehrenfeld, 1997, 2000b) and to matters of meta-analysis such as by whom and, by what criteria, these sorts of debates are decided (Boons and Roome, 2000).

Transformative and Incremental Change

Once some degree of normative content in the field is acknowledged, it is easy to entertain a related question of goals: is the environmental improvement that is sought large, transformative, and discontinuous or is it incremental and continuous with current practice and infrastructure? Much of the conceptual discussion in industrial ecology looks to transformative change through the development and/or implementation of radically innovative technology, changes in consumption patterns, or new organizational arrangements. This type of transformative change can range from shifts to a hydrogen (Marchetti, 1989) or carbohydrate economy (Morris and Ahmed, 1992), factor 4 or factor 10 reductions in materials throughput (Weizsacker et al., 1997), a shift to the use of services in lieu of products (Stahel, 1994; Mont, 2000) or new political–economic structures. At the same time, much of the practical work in ecodesign and life cycle management aims at more modest changes in product design protocols, materials choice, interfirm relationships, or environmental policy. Still other analyses assert that current arrangements meet the test of industrial ecology, employing the practices characteristic of the field, and need no significant improvement (Linden, 1994). Yet the tension between transformative and incremental change may be overdrawn to the extent that, in many circumstances, the two paths are not mutually exclusive: the more modest changes can be pursued while the more ambitious ones are debated, refined, and implemented. To the extent that such tensions do exist, they frequently reflect differing assessments of the severity of the current environmental situation as well as ideological differences about the degree to which market economies and current political institutions can and do achieve environmental goals.

The Boundaries of Industrial Ecology

Like the question of goals, the boundaries of the field are subject to varying interpretations. They cannot be defined deductively from first principles; there is no authoritative epistemology in industrial ecology. At the same time, views on what are part of or outside this field shape, what is published in the *Journal of Industrial Ecology*, what is included in compendia such as this handbook and what projects receive funding from sources focusing on the development of industrial ecology.

One of the considerations with respect to the boundaries is whether industrial ecology should address not only 'what' but 'how' (Andrews, 2000). Investigations of 'what' inform our understanding of the character of technological and natural systems – characterizing the manner in which these systems behave and interact, and under what circumstances environmental situations that humans deem preferable (for example, the absence of ozone holes in the atmosphere) might occur, much as in the definition of industrial ecology quoted in the beginning of this article. In that respect, the 'what' investigations include the what-if questions described above with respect to the normative aspects of industrial ecology: 'What if different materials were used for packaging; would carbon dioxide emissions decrease and global warming slow?' Some industrial ecologists, however, argue that the field must also embrace social, political, and economic questions of 'how.' That is, given the identification of a preferred outcome, what strategies should be employed to bring about that outcome (Andrews, 2000; Jackson and Clift, 1998).

Thus the 'how' questions are largely a province of the social sciences. Sociology, economics, anthropology, psychology, political science, and related fields have the potential to help identify strategies that are more likely to succeed. Nonetheless, the social sciences are not confined to 'how' questions (Fischhoff and Small, 1999). They can also indicate what is happening as when, for example, social scientists investigate the quantity and character of consumption in households and how it drives production and waste management activities and therefore environmental outcomes (Duchin, 1998; Noorman and Schoot Uiterkamp, 1998).

Most industrial ecologists would agree that such knowledge is crucial, but some would argue that that knowledge should remain lodged in allied fields, otherwise the boundaries and identity of industrial ecology will become so expansive as to be diffuse. Further complicating this tension are questions of whether these different sorts of inquiry can be construed as modular. That is, can they be pursued independently and *subsequently* melded to generate reliable insights? Or does their intellectual and organizational separation inevitably mean that the modular inquiries will be impoverished, incapable of integration, or even fundamentally misleading (Lifset, 1998b)? Put more simply, *must* the questions that industrial ecology seeks to answer be pursued on an interdisciplinary basis to produce reliable answers? Ultimately, it will be the productivity of the various approaches in generating conceptual insights and practical knowledge that will determine their adoption[3].

Conclusion

As a new field, industrial ecology is a cluster of concepts, tools, metaphors, and exemplary applications and objectives. Some aspects of the field have well-defined relationships, whereas other elements are only loosely grouped together, connected as much by the enthusiasm of the proponents as by a well-articulated intellectual architecture. We do not see this looseness as a fatal flaw in an emerging field, but rather as an opportunity for creativity and constructive discourse, and as a challenge.

See also: Eco-Cities and Sustainable Urbanism; Energy Flow Analysis; Greening of Business; Material Flow Analysis; Sustainability and Sustainability Science; Sustainable Production and Consumption.

Notes

1. We put 'natural' in quotation marks because there are many ways in which the notion of natural ecosystems is complicated or contested. Many analysts argue that there are no longer any ecosystems unaffected by humankind, although clearly, even in this view, there is wide variation in the degree to which human activity dominates nonhuman ecosystems.

More subtly, the notion of 'natural' is socially constructed and subject to varying interpretations across cultures (Williams, 1980; Cronon, 1996).

2. Multiple meanings extend to other terms in the field. 'Industrial ecology' is variously used to mean (1) a field of study, (2) a set of environmentally desirable practices, and (3) the same practices as in meaning (2), but viewed neutrally. Such plurality of meanings is not unusual, however: 'history' refers both to past events and to the discipline that systematically studies those events.
3. Disagreement about industrial ecology's boundaries is exacerbated by more pedestrian conflation of the ethics and values, the social sciences, and public policy analysis. In particular, nonsocial scientists sometimes do not realize that the social sciences have a primarily positive/analytical focus, characterizing how humans behave, whereas it is the humanities that investigate and debate matters of values. Public policy analysis is often instrumental, asking how effectively certain strategies accomplish a set of public goals. Few industrial ecologists would suggest that the field offers powerful tools for adjudicating disputes over values, even if those disputes are important to the field.

Addendum

1. Addendum to the Lifset and Graedel article, "Industrial Ecology: Goals and Definitions" for *International Encyclopedia of Social and Behavioral Sciences*.
2. In the 15 years since this chapter was written, industrial ecology has evolved from an emerging to a rapidly growing field. Methodologies have matured, concepts have evolved, and empirical understanding has grown substantially.
3. Industrial symbiosis, the sharing and exchange of resources by geographically proximate facilities, has moved from an academic curiosity to an important component of environmental policy in countries such as China, Korea, and the United Kingdom. Research has shifted from identification and characterization of individual instances of industrial symbiosis to theorizing (Chertow, 2007) and hypothesis testing and statistical analysis (Paquin et al., 2014).
4. The concept of a 'Type III' industrial ecology in which resources are reused indefinitely (as described in the chapter) originated in the 1990s as a vision of a sustainable future. It gathered considerable force in 2006 when it was included in China's 11th five-year plan under the rubric of the 'circular economy'(Yuan et al., 2006). The focus of the circular economy concept, which has considerable popular appeal (Ellen MacArthur Foundation, 2012), has gradually shifted from the 3Rs (reduce, reuse, recycle) to a broader approach that encompasses upstream (supply chain) considerations, cleaner production, eco-industrial parks, and a number of other practices.
5. Despite the attractiveness of the circular economy concept, it has been pointed out that circularity has its limits (Kral et al., 2013). Basic thermodynamics, dissipative uses such as fertilizer and paint, and human frailties all conspire against a perfectly circular economy. Allwood (2014) argues for an analytically grounded and technologically sophisticated approach – efforts in product life extension, lightweight design, and component reuse. The combination of the circular economy vision with engineering innovations holds out the possibility of more effectively optimizing the use of materials.
6. Quantification of resource flows – circular or otherwise – has remained an important foundation for industrial ecology. In a little more than a decade, industrial ecologists used material flow analysis (MFA) to quantify the cycles of many of the elements, and to do so at both country and global levels (Chen and Graedel, 2012). As a result, production, flows into use, recycling, loss, and often trade flows as well, are now much better understood than was formerly the case. MFA has also been extended to incorporate material stocks, generating insights such as the possibility that anthropogenic stocks may saturate on a per capita basis in developed countries (Müller et al., 2006). Economy-wide MFA, in which large categories of material use are quantified in bulk, has been institutionalized (Eurostat, 2014) and used, for example, to quantify material footprints of nations to challenge the notion that the use of natural resources has been decoupled from economic growth (Wiedmann et al., 2013) and to generate theories regarding large-scale socioecological transition (Fischer-Kowalski and Haberl, 2007).
7. Input–output analysis (IOA) has become a key tool in industrial ecology, providing a means of assessing economy-wide environmental impacts. IOA is also frequently used in 'footprinting' (quantitative tracing of the impact of human activity through the economy and then into the biophysical environment) of carbon, water, biocapacity, and other resources (Lifset, 2014).
8. Material flow and stock analysis are increasingly conceptualized as metabolic processes, extending industrial ecology's use of the biological analogy. Research on the metabolism of cities has blossomed (Kennedy et al., 2007), assisting cities in their efforts to quantify and manage their carbon emissions. Urban carbon footprinting has revealed the crucial role – and ambiguity – in assigning responsibility for carbon emissions from infrastructure and trade beyond the urban boundary (Ramaswami et al., 2012).
9. The manufacture of products for international trade, and the role of consumption in driving resource flows and emissions, has become a central focus of industrial ecology in multiple domains. Lin et al. (2014), for example, have used trade statistics and air pollution models to demonstrate that 17–36% of air pollutants emitted in China relate directly to exports, with a noticeable shift in air quality from consuming countries to producing countries. Motivated by the dramatic increase in attention to climate change in the past decade and a half, carbon accounting and carbon footprinting have emerged as a means of quantifying the global warming impacts of the production and consumption of specific products and activities at various scales. At the national and international level, carbon footprinting and multiregional input–output analysis have allowed researchers to quantify greenhouse

gas (GHG) emissions associated with the final consumption of goods and services, e.g., for 73 nations and 14 aggregate world regions (Hertwich and Peters, 2009). By assessing GHG emissions according to a region's consumption rather than activities within its territorial boundaries, this approach provided an influential, alternative perspective on the drivers of global warming.

10. As global populations grow and as richer mineral deposits are depleted over time, concerns have arisen as to the long-term availability of mineral resources in sufficient quantities (Nakamura and Sato, 2011). The subjects for such discussions have ranged from phosphorus, essential for food production (Cordell et al., 2009), to gold, widely employed in investment, jewelry, and technology (Kerr, 2012). An industrial ecology approach to this issue has been to quantify specific resources in ore deposits, to compile production, demand, and recycling statistics, and to compare that information with anticipated demand (Northey et al., 2014). Substantial effort has also gone into assessing 'criticality' by determining supply risk and vulnerability to supply disruption for a number of elements (Nassar et al., 2012). The result of this research has been to embed industrial ecology firmly in discussions of the human life style and of sustainability.
11. Among the most high-profile developments in industrial ecology has been the widespread adoption of life cycle thinking as a form of systems-oriented environmental analysis, and the associated tool, LCA. Use of LCA to assess products, services, and technologies by business, government, and academe to identify environmental 'hotspots' and to support environmental decision-making has become common globally. LCA, for example, is the basis for regulation of the global warming emissions from transportation fuels in California (the Low Carbon Fuel Standard) and the source of considerable debate about the indirect impacts of biofuels production on land use in agriculture (Searchinger et al., 2008).
12. LCA has undergone significant methodological development including the incorporation of input–output analysis (Hendrickson et al., 2006). Distinctions have emerged between attributional LCA which provides a snapshot of potential environmental impacts and consequential LCA which seeks to model the consequences of a decision or change in the phenomenon being studied. Efforts are growing to incorporate social impacts into LCA under the rubric of 'social LCA' (Norris, 2012), part of a broader trend to incorporate social science in many of the subfields within industrial ecology.
13. The maturing of industrial ecology is reflected not only in the research output sketched above, but also in its role in the fifth Assessment of the Intergovernmental Panel on Climate Change (IPCC, 2014), where industrial ecologists played central roles in the working group on industry (WG3) among others. Studies such as those mentioned above demonstrate the transition of industrial ecology from basic conceptualizing and exploratory analysis to substantiated and increasingly theoretically embedded generalization and sophisticated model development.

Bibliography

Adriaanse, Albert, Bringezu, Stefan, Hammond, Allen, Moriguchi, Yuichi, Rodenburg, Eric, Rogich, Donald G., Schütz, Helmut, 1997. Resource Flows: The Material Basis of Industrial Economies. World Resources Institute, Washington, DC.

Alcamo, Joseph G., Kreileman, Eric J.J., Krol, Maarten S., Zuidema, G., 1994. Modeling the global society-biosphere-climate systems part 1: model description and testing. Water, Air and Soil Pollution 76, 1–35.

Allen, David T., 1996. Pollution prevention: engineering design at macro- meso- and micro-scales. Clean Technology 2 (2), 51–59.

Allenby, Braden R., 1999a. Industrial Ecology, Policy Framework and Implementation. Prentice-Hall, Englewood Cliffs, NJ.

Allenby, B.R., 1999b. Culture and industrial ecology. Journal of Industrial Ecology 3 (1), 2–4.

Allenby, Braden R., Cooper, William, Spring 1994. Understanding industrial ecology from a biological systems perspective. Total Quality Environmental Management 3, 343–354.

Allwood, J.M., 2014. Squaring the circular economy: the role of recycling within a hierarchy of material management strategies. In: Worrell, E., Reuter, M.A. (Eds.), Handbook of Recycling. Elsevier, Amsterdam, pp. 445–477.

Andrews, Clinton J., 2000. Building a micro foundation for industrial ecology. Journal of Industrial Ecology 4 (3), 35–51.

Ausubel, Jesse H., 1996. Can technology spare the earth? American Scientist 84 (2), 166–178.

Ausubel, Jesse H., Langford, H. Dale (Eds.), 1997. Technological Trajectories and the Human Environment. National Academy Press, Washington, DC.

Ayres, Robert U., 1992. Cowboys, cornucopians and long-run sustainability. Ecological Economics 8, 189–207.

Ayres, Robert U., Rod, Samuel R., 1986. Patterns of pollution in the Hudson-Raritan basin. Environment 28 (4), 14–43.

Ayres, Robert U., Simonis, Udo E. (Eds.), 1994. Industrial Metabolism: Restructuring for Sustainable Development. United Nations University Press, Tokyo; New York; Paris.

Ayres, Robert U., Schlesinger, William H., Socolow, Robert H., 1994. Human impacts on the carbon and nitrogen cycles. In: Socolow, Robert H., Andrews, Clinton J., Berkhout, Frans G., Thomas, Valerie M. (Eds.), Industrial Ecology and Global Change. Cambridge University Press, Cambridge, UK and New York, pp. 121–156.

Benyus, Janine M., 1997. Biomimicry: Innovation Inspired by Nature. William Morrow and Co, New York.

van Berkel, René, Williams, Ester, Lafleur, Margie, 1997. The relationship between cleaner production and industrial ecology. Journal of Industrial Ecology 1 (1), 51–67.

Björklund, Anna, Bjuggren, Charlotte, Dalemo, Magnus, Sonesson, Ulf, 1999. Planning bio-degradable waste management in Stockholm. Journal of Industrial Ecology 3 (4), 43–58.

Boons, Frank, Roome, Nigel, 2000. Industrial ecology as a cultural phenomenon: on objectivity as a normative position. Journal of Industrial Ecology 4 (2).

Carnahan, James V., Thurston, Deborah L., 1998. Tradeoff modeling for product and manufacturing process design for the environment. Journal of Industrial Ecology 2 (1), 79–92.

Chen, W.-Q., Graedel, T.E., 2012. Anthropogenic cycles of the elements: a critical review. Environmental Science & Technology 46, 8574–8586.

Chertow, M.R., 2000. Accelerating commercialization of environmental technology in the United States: theory and case studies. Thesis (Ph D) thesis. Yale University.

Chertow, Marian R., 2001. The IPAT equation and its variants: changing views of technology and environmental impact. Journal of Industrial Ecology 4 (4), 13–29.

Chertow, M., 2007. "Uncovering" industrial symbiosis. Journal of Industrial Ecology 11 (1), 11–30.

Cleveland, Cutler J., Ruth, Matthias, 1998. Indicators of dematerialization and the materials intensity of use. Journal of Industrial Ecology 2 (3), 15–50.

Cordell, D., Drangert, J.-O., White, S., 2009. The story of phosphorus: global food security and food for thought. Global Environmental Change 19, 292–305.

Cronon, W., 1996. Uncommon ground: rethinking the human place in nature. Pbk. ed. W.W. Norton, New York.

Davis, Gary A., 1997. Book review. Journal of Industrial Ecology 1 (4), 131–133.

DeSimone, Livio D., Popoff, Frank, The world business Council for sustainable development, 1997. Eco-Efficiency. The Business Link to Sustainable Development. MIT Press, Cambridge, MA.

Diwekar, Urmila M., Small, Mitchell J., 1998. Industrial ecology and process optimization. Journal of Industrial Ecology 2 (3), 11–13.

Duchin, Faye, 1998. Structural Economics: Measuring Change in Technology, Life-Styles and the Environment. Island Press, Washington, DC.

Ehrenfeld, John R., 1997. Industrial ecology: a framework for product and process design. Journal of Cleaner Production 5 (1–2), 87–95.

Ehrenfeld, John R., October 6, 2000a. Personal Communication.

Ehrenfeld, John R., 2000b. Industrial ecology, paradigm shift or normal science? American Behavioral Scientist 44 (2), 229–244.

Ehrenfeld, John R., Gertler, Nicholas, 1997. Industrial ecology in practice: the evolution of interdependence at Kalundborg. Journal of Industrial Ecology 1 (1), 67–79.

Ellen MacArthur Foundation, 2012. Towards the Circular Economy, vol. 1. Cowes, United Kingdom.

Erkman, Suren, 1997. Industrial ecology: an historical overview. Journal of Cleaner Production 5 (1–2), 1–10.

Eurostat, 2014. Material flow accounts. http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Material_flow_accounts. Accessed November 3 2014.

Fischer-Kowalski, Marina, 1998. Society's metabolism. The intellectual history of material flow analysis, Part I, 1860–1970. Journal of Industrial Ecology 2 (1), 61–78.

Fischer-Kowalski, M., Haberl, H. (Eds.), 2007. Socioecological Transitions and Global Change. Edward Elgar, Cheltanham, UK.

Fischer-Kowalski, Marina, Hüttler, Walter, 1998. Society's metabolism. The intellectual history of material flow analysis, part II, 1970–1998. Journal of Industrial Ecology 2 (4), 107–136.

Fischhoff, Baruch, Small, Mitchell J., 1999. Human behavior in industrial ecology modeling. Journal of Industrial Ecology 3 (2), 4–7.

Frank, Robert H., Gilovich, Thomas, Regan, Dennis T., 1993. Does studying economics inhibit cooperation? Journal of Economic Perspectives 7 (2), 152–171.

Frosch, Robert A., Gallopoulos, Nicholas E., 1989. Strategies for manufacturing. Scientific American 261 (3), 94–102 (special issue on 'Managing Planet Earth'.).

Fussler, Claude, 1996. Driving Eco-Innovation: A Breakthrough Discipline for Innovation and Sustainability. Pitman Publishing, Washington, DC.

Frosch, R.A., Clark, W.C., Crawford, J., Sagar, A., Tschang, T., Webber, A., 1997. The industrial ecology of metals: a reconnaissance. Philosophical Transactions of the Royal Society of London. Series A: Mathematical. Physical and Engineering Sciences 355 (1728), 1335–1347. http://rsta.royalsocietypublishing.org/content/355/1728/1335.abstract.

Graedel, Thomas E., 2000. Evolution of industrial ecology. Environmental Science and Technology 34 (1), 28A–31A.

Graedel, Thomas E., Allenby, Braden R., 1995. Industrial Ecology. Prentice-Hall, Upper Saddle River, NJ.

Graedel, T.E., Cao, J., 2010. Metal spectra as indicators of development. Proceedings of the National Academy of Sciences of the United States 107, 20905–20910.

Grübler, Arnulf, 1998. Technology and Global Change. Cambridge University Press, Cambridge, UK.

Hardy, C., Graedel, T.E., 2002. Industrial ecosystems as food webs. Journal of Industrial Ecology 6 (1), 29–38.

Harrison, Kathryn, 1998. Talking with the donkey: cooperative approaches to environmental protection. Journal of Industrial Ecology 2 (3), 51–72.

Hart, Stuart L., Milstein, Mark B., 1999. Global sustainability and the creative destruction of industries. Sloan Management Review 41 (1), 21–33.

Hendrickson, C.T., Lave, L.B., Matthews, S., 2006. Environmental Life Cycle Assessment of Goods and Services: An Input-Output Approach. Resources for the Future, Washington, DC.

Hertwich, E.G., Peters, G.P., 2009. Carbon footprint of nations: a global, trade-linked analysis. Environmental Science & Technology 43 (16), 6414–6420.

Intergovernmental Panel on Climate Change, 2014. Fifth Assessment Report. Paris, France: Intergovernmental Panel on Climate Change. www.ipcc.ch.

Isaacs, Jacqueline A., Gupta, Surendra M., 1997. Economic consequences of increasing polymer content for the US automobile recycling infrastructure. Journal of Industrial Ecology 1 (4), 19–33.

Jackson, Tim, 1999. Book review. Journal of Industrial Ecology 3 (2), 181–182.

Jackson, Tim, Clift, Roland, 1998. Where's the profit in industrial ecology? Journal of Industrial Ecology 2 (1), 3–5.

Karlsson, Sten, 1999. Closing the technospheric flows of toxic metals: modeling lead losses from a lead-acid battery system for Sweden. Journal of Industrial Ecology 3 (1), 23–40.

Kennedy, C., Cuddihy, J., Engel-Yan, J., 2007. The changing metabolism of cities. Journal of Industrial Ecology 11 (2), 43–59.

Keoleian, Gregory A., Menerey, D., 1994. Sustainable development by design: review of life cycle design and related approaches. Journal of the Air and Waste Management Association 44, 644–668.

Kerr, R.A., 2012. Is the world tottering on the precipice of peak gold? Science 335, 1038–1039.

Kral, U., Kellner, K., Brunner, P.H., 2013. Sustainable resource use requires "clean cycles" and safe "final sinks". Science of the Total Environment 461-462, 819–822.

Levine, Stephen H., 1999. Products and ecological models: a population ecology perspective. Journal of Industrial Ecology 3 (2), 47–62.

Lifset, Reid, 1993. Take it back: extended producer responsibility as a form of incentive-based policy. Journal of Resource Management and Technology 21 (4), 163–175.

Lifset, Reid, 1998a. Taking stock and improving flow. Journal of Industrial Ecology 2 (1), 1–2.

Lifset, Reid, 1998b. Setting the boundaries? Journal of Industrial Ecology 2 (2), 1–2.

Lifset, Reid, 2000. Industrial ecology and forestry. Journal of Forestry 98 (10), 12–13.

Lifset, R., 2014. Frontiers in footprinting. Journal of Industrial Ecology 18 (1), 1–3.

Linden, Henry R., 1994. Energy and industrial ecology. In: Allenby, Braden R., Richards, Deanna J. (Eds.), The Greening of Industrial Ecosystems. National Academy Press, Washington, DC, pp. 38–60.

Lin, J., Pan, D., Davis, S.J., Zhang, Q., He, K., Wang, C., Streets, D.G., Wuebbles, D.J., Guan, D., 2014. China's international trade and air pollution in the United States. Proceedings of the National Academy of Sciences of the United States 111, 1736–1741.

Liu, G., Bangs, C.E., Müller, D.B., 2013. Stock dynamics and emission pathways of the global aluminium cycle. Nature Climate Change 3, 3338–3342.

Marchetti, Cesare, 1989. How to solve the CO_2 problem without tears. International Journal of Hydrogen Energy 14 (8), 493–506.

MARKAL MATTER, 2000. MATerials Technologies for Greenhouse Gas Emission Reduction. http://www.ecn.nl/unit_bs/etsap/markal/matter/index.html (accessed 10.10.00.).

Mont, Oksana, 2000. A Recent Review of Theory, Research and Practice of Product Service Systems. AFR report no. 288. Swedish Waste Research Council (AFR), Stockholm.

Morris, David, Ahmed, Irshad, 1992. The Carbohydrate Economy: Making Chemicals and Industrial Materials from Plant Matter. The Institute for Local Self Reliance, Institute for Local Self-Reliance, Washington, DC.

Müller, D.B., Wang, T., Duval, B., Graedel, T.E., 2006. Exploring the engine of anthropogenic iron cycles. Proceedings of the National Academy of Sciences 103 (44), 16111–16116.

Nakamura, E., Sato, K., 2011. Manging the scarcity of chemical elements. Nature Materials 10, 158–161.

Nakicenovic, Nebojsa, 1997. Freeing energy from carbon. In: Ausubel, Jesse H., Langford, H. Dale (Eds.), Technological Trajectories and the Human Environment. National Academy Press, Washington, DC, pp. 74–88.

Nassar, N.T., Barr, R., Browning, M., Diao, Z., Friedlander, E., Harper, E.M., Henly, C., Kavlak, G., Kwatra, S., Jun, C., Warren, S., Yang, M.-Y., Graedel, T.E., 2012. Criticality of the geological copper family. Environmental Science & Technology 46, 1071–1078.

Noorman, Klaas J., Schoot Uiterkamp, A.J.M. (Eds.), 1998. Green Households? Domestic Consumers, Environment and Sustainability. Earthscan Publications, London.

Norberg-Bohm, V., 2000. Review of Innovation Effects of Environmental Policy Instruments, vols. 5 and 6, by Paul Klemmer, Ulrike Lehr, and Klaus Lobbe; Policies for Cleaner Technology: A New Agenda for Government and Industry, by Anthony Clayton, Graham Spinardia, and Robin Williams; Adoption of Environmental Innovations: the Dynamics of Innovation as Interplay between Business Competence, Environmental Orientation and Network Involvement, by Koos Van Dijken. Journal of Industrial Ecology 4 (3), 123–126. http://dx.doi.org/10.1162/108819800300106429.

Norris, C.B., 2012. Social life cycle assessment: a technique providing a new wealth of information to inform sustainability-related decision making. In: Mary Ann Curran (Ed.), Life Cycle Assessment Handbook. John Wiley & Sons, Inc, Hoboken.

Northey, S., Mohr, S., Mudd, G.M., Weng, Z., Giurco, D., 2014. Modelling future copper ore grade decline based on a detailed assessment of copper resources and mining. Resources, Conservation, and Recycling 83, 190–201.

Oldenburg, Kirsten V., Geiser, Ken, 1997. Pollution prevention and/or industrial ecology. Journal of Cleaner Production 5 (1–2), 103–108.

Organization for Economic Cooperation and Development (OECD), 1998. Eco-Efficiency. OECD, Paris.

Paquin, R.L., Tilleman, S.G., Howard-Grenville, J., 2014. Is there cash in that trash? Journal of Industrial Ecology 18 (2), 268–279.

Potting, José, Schöpp, Wolfgang, Blok, Kornelis, Hauschild, Michael, 1998. Site-dependent life-cycle impact assessment of acidification. Journal of Industrial Ecology 2 (2), 63–87.

Preston, John T., 1997. Technology innovation and environmental progress. In: Chertow, Marian R., Esty, Daniel C. (Eds.), Thinking Ecologically, the Next Generation of Environmental Policy. Yale University Press, London and New Haven, CT, pp. 136–149.

Ramaswami, A., Chavez, A., Chertow, M., 2012. Advances in carbon footprinting of cities and implications for coupled analysis of urban material-energy flows. Journal of Industrial Ecology 16 (6), 783–785.

Reijnders, Lucas, 1997. The factor X debate: setting targets for eco-efficiency. Journal of Industrial Ecology 2 (1), 13–22.

Reinhardt, Forest, 1999. Market failure and the environmental policies of firms: economic ratio-nales for 'beyond compliance' behavior. Journal of Industrial Ecology 3 (1), 9–22.

Reiskin, Edward D., White, Allen L., Johnson, Jill K., Votta, Thomas J., 1999. Servicizing the chemical supply chain. Journal of Industrial Ecology 3 (2/3), 19–31.

Ruth, Matthias, 1996. Evolutionary economics at the crossroads of biology and physics. Journal of Social and Evolutionary Systems 19 (2), 125–144.

Ruth, Matthias, Harrington Jr., Thomas, 1997. Dynamics of material and energy use in US pulp and paper manufacturing. Journal of Industrial Ecology 1 (3), 147–168.

Sarkis, Joseph, 1995. Supply chain management and environmentally conscious design and manufacturing. International Journal of Environmentally Conscious Design and Manufacturing 4 (2), 43–52.

Scheringer, Martin, Halder, Daniel, Hungerbühler, Konrad, 1999. Comparing the environmental performance of fluorescent whitening agents with peroxide bleaching of mechanical pulp. Journal of Industrial Ecology 3 (4), 77–95.

Schmidheiny, Stephan, 1992. Changing Course: A Global Business Persepctive on Development and the Environment. MIT Press, Cambridge, MA.

Searchinger, T., Heimlich, R., Houghton, R.A., Dong, F., Elobeid, A., Fabiosa, J., Tokgoz, S., Hayes, D., Yu, T.-H., 2008. Use of U.S. croplands for biofuels increases greenhouse gases through emissions from land-use change. Science 319 (5867), 1238–1240.

Sharfman, Mark, Ellington, Rex T., Meo, Mark, 1998. Conoco and the vapor recovery project: using innovation to preserve autonomy. Journal of Industrial Ecology 3 (1), 93–110.

Socolow, Robert H., 1994. Six perspectives from industrial ecology. In: Socolow, Robert H., Andrews, Clinton J., Berkhout, Frans G., Thomas, Valerie M. (Eds.), Industrial Ecology and Global Change. Cambridge University Press, Cambridge, UK, pp. 3–16.

Socolow, Robert H., Thomas, Valerie M., 1997. The industrial ecology of lead and electric vehicles. Journal of Industrial Ecology 1 (1), 13–36.

Stahel, Walter R., 1994. The utilization-focused service economy: resource efficiency and product-life extension. In: Allenby, Braden R., Richards, Deanna J. (Eds.), The Greening of Industrial Ecosystems. National Academy Press, Washington, DC, pp. 178–190.

Stigliani, William M., Anderberg, Stefan, 1994. Industrial metabolism at the regional level: The Rhine Basin. In: Ayres, Robert U., Simonis, Udo E. (Eds.), Industrial Metabolism: Restructuring for Sustainable Development. United Nations University Press, Tokyo, pp. 119–162.

Stigliani, William, Jaffé, Peter F., Anderberg, Stefan, 1994. Metals loading of the environment: cadmium in the Rhine Basin. In: Socolow, Robert H., Andrews, Clinton J., Berkhout, Frans G., Thomas, Valerie M. (Eds.), Industrial Ecology and Global Change. Cambridge University Press, Cambridge, UK, pp. 287–296.

Su, B., Heshmati, A., Geng, Y., Yu, X., 2013. A review of the circular economy in China: moving from rhetoric to implementation. Journal of Cleaner Production 42, 219–227.

Udo de Haes, Helias A., 1996. Towards a Methodology for Life Cycle Impact Assessment. Society of Environmental Toxicology and Chemistry, Brussels.

van der Voet, Ester, Guinée, Jeroen B., Udo de Haes, Helias A., 2000. Heavy Metals: A Problem Solved? Methods and Models to Evaluate Policy Strategies for Heavy Metals. In: Environment and Policy, vol. 22. Kluwer Academic Publishers, Dordrecht, The Netherlands; Boston, MA, London.

Weizsacker, E. von, Lovins, A.B., Lovins, L.H., 1997. Factor Four: Doubling Wealth, Halving Resource Use. Earthscan, London. Earthscan Publications Ltd, London.

Wernick, Iddo K., Ausubel, Jesse H., 1997. Research Needs for Industrial Ecology. Lawrence Livermore National Laboratory for the US Department of Energy. http://phe.rockefeller.edu/ie_agenda.

Wernick, Iddo K., Herman, Robert, Govind, Shekhar, Ausubel, Jesse H., 1997a. Materialization and dematerialization: measures and trends. In: Ausubel, Jesse H., Langford, H.D. (Eds.), Technological Trajectories and the Human Environment. National Academy Press, Washington, DC, pp. 135–156.

Wernick, Iddo K., Waggoner, Paul E., Ausubel, Jesse H., 1997b. Searching for leverage to conserve forests: the industrial ecology of wood products in the United States. Journal of Industrial Ecology 1 (3), 125–146.

White, Robert, 1994. Preface. In: Allenby, Braden R., Richards, Deanna J. (Eds.), The Greening of Industrial Ecosystems. National Academy Press, Washington, DC.

Wiedmann, T.O., Schandl, H., Lenzen, M., Moran, D., Suh, S., West, J., Kanemoto, K., 2013. The material footprint of nations. Proceedings of the National Academy of Sciences. http://dx.doi.org/10.1073/pnas.1220362110, 3, 2013.

Williams, R., 1980. The Idea of Nature. In: Problems in Materialism and Culture: Selected Essays. Verso, London.

Wisberg, Nicoline, Clift, Roland, 1999. The CHAINET concerted action. Journal of Industrial Ecology 3 (4), 8–9.

Yuan, Z., Bi, J., Moriguichi, Y., 2006. The circular economy: a new development strategy in China. Journal of Industrial Ecology 10 (1–2), 4–8.

Industrial Geography

Amy K Glasmeier and Christa Lee-Chuvala, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

From early craft production methods to the rise of the knowledge economy, spatial dynamics have shaped the organizational structure of industry. However, the global industrial landscape has been created not only by the location decisions of firms but also by historical processes related to technological change, social interaction, and cultural influences. Current industrial geography reflects the manifestation and evolution of these processes through successive eras.

Industrial geography is the study of the global industrial landscape, which serves as the background context for studies of industrial location. Industrial geography is "not simply the summation of individual location; it is the consequence of historical processes that transcend the ability (and often the perception) of any individual firm" (Harrington and Warf, 1995: p. 115). Industrial geography represents the accretion of material reality and human ingenuity, which are in turn shaped by the dominant mode of wealth creation. The geographic approach to understanding industrial structures emphasizes how the organizational structure of industry manifests across space and interacts with and helps shape the lived experience of people. As such, the industrial geography of today is the sum of past practices, technology eras, and social relationships.

Industrial geography is related to but distinct from industrial location. Industrial location is the study of the location decision-making processes of firms and reflects the trade-off between access to natural resources and access to markets. Early work in this field focused on the fixed-location competitive firm (Weber, 1971). Subsequent work incorporated concerns including the nature of the market (Lösch, 1954), the economic spillover effects of firms in the same location (Hoover, 1948), and the importance of and variability in pools of labor (Storper and Walker, 1989). In contrast to the study of industrial location, this article focuses on the historical development of the industrial landscape.

The Industrial Geographic Landscape

What most people see today in the industrial landscape is the sum of multiple periods of economic development. In the American and Western European settings, the old mill town is a physical representation of a type of industrial landscape that evolved prior to or in conjunction with the industrial revolution. By beginning from the late eighteenth and early nineteenth centuries, a considerable period of the economic history that forms the bedrock of the contemporary industrial landscape is left out. Under feudalism, for example, industrial activity was focused on the manor and the creation of goods for local consumption. Occasional surpluses were produced that were traded in market towns (Rosenberg and Birdzell, 1986). With the decline of feudalism and the ascendancy of urban centers of trade and production, the industrial landscape began to emerge in the form of urban workshops. Independent manufacturers were only one source of manufactured goods, however. Merchants were key to early industrialization; they organized not only the markets for manufactured goods, but also production outside the influence of urban guilds. The 'putting out' system of rural manufacturing extended the influence of the industrial landscape significantly beyond urban centers. Rural manufacturing provided a check on urban wages but eventually declined when merchants could no longer guarantee or control output from rural manufacturers.

Prior to the industrial revolution, the production of manufactured goods unfolded within a system of handicraft manufacture under the purview of a master craftsman. One or more journeymen or apprentices assisted an exceedingly clever and dexterous person, the master.

The regulation of this form of production rested with guilds, production organizations that regulated skills and access to skills, and to a much lesser extent access to markets (Landes, 1969). By restricting access to skills, craftsmen were able to preserve their markets and ensure public belief in the credibility of their products. Guilds acted as brakes on the expansion of output and thus overproduction. In later periods, guilds also thwarted technological change, a detail that led ultimately to their own demise.

The movement from craft production to the putting out system was an outgrowth of the merchants' increasing control over producers. The expansion of the market and the lack of access to and knowledge of its extensiveness placed artisans at a decided disadvantage. Market reach and knowledge of the tastes and preferences of distant customers was a clear advantage of merchants who could accumulate demand and periodically fill it (Glasmeier, 1990).

Putting out, while highly efficient in an era with few economic alternatives, grew less and less viable as an organizing system over time. As workers grew wealthier and additional avenues of income creation became available to them, merchants had increasing difficulty controlling the output and quality of cottage industry goods. During periods of market growth, meeting new demand became harder as workers began to assert more control over their lives and labor process. The need to find an alternative, more controllable system of manufacture became overwhelming as markets grew and demand burgeoned. In Britain, geographic range had essentially been reached and in some cases exceeded. Transport costs became a real impediment.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.72026-5

As control of workers' efforts lessened, and the limits of geographic expansion were reached, a slow but steady transformation took place as machines began to replace human labor, and the organization of labor within specific locales and structures provided the control employers sought. The movement to a factory-based system of manufacture was neither rapid nor complete. The slow demise of the putting out system provides one explanation for the emergence of factories: the need to control labor and the production process. Volumes have been written on this subject, from a wide variety of perspectives and with an equally broad range of emphases. What is important here is the way the movement from a more decentralized to a more centralized system removed control of production from workers laboring in the manufacture of a good, and turned it over to the capitalist. Of additional importance is the recognition that the process of change did not occur in the same way simultaneously in all relevant locations.

The ways in which mechanization unfolded across industrial sectors are also relevant to understanding industrial geography. The reasons for early and rapid mechanization in some sectors and for resistance of machine support in others can be reduced to several intervening factors. The nature of the good largely determined its susceptibility to mechanization. This attribute went hand in hand with the extent of the market for the good. If the good was a high-value/low-volume one and of a luxurious nature, labor costs were relatively insignificant and thus there was a minimal stimulus to change, unless of course the transition to machines was twinned with a movement toward a broader market and a more widely disseminated good. In this case, market shifts could simply require change in order to hold a longer term market position. Institutional factors could also serve as impediments to change. The number of possible explanations supporting the lack of investment is certainly sufficient to see that the prospect of a smooth functioning factor market acting as a stimulus for change was inherently contingent on elements other than just the short-term costs of production.

Early centers of industrialization were highly specialized – a result of natural resource location as well as of built-up competencies (e.g., in the local labor force) – the initial impetus of which was often accidental. Trade was essential to the growth of industrialization in place. The formation of institutions and social relationships combined with key developments in technology facilitated and ultimately encouraged trade among regions. Fundamental to the success of trade across long distances was the development of shipping technology that allowed the movement of large volumes of goods in reasonable time. Simultaneously, in order for trade to occur, mechanisms were needed to reduce risk and to finance the time between the manufacture and the sale of a good. With the creation of a number of institutional innovations such as the banking, credit, and insurance sectors, industrialization in the late eighteenth and early nineteenth centuries moved beyond local production for local consumption and toward production for trade.

The industrial revolution marks a critical moment in the formation of the industrial landscape. Essential to the pervasive diffusion of the industrial revolution was the ability to harness power sources in order to concentrate economic activities in space. Simultaneously, the development of new technologies led both to an enhanced capacity to organize industrial activity and to the creation of new manufactured goods themselves. Finally, the organization of production within factories allowed the achievement of economies of scale and task specialization across the production process.

Before the beginning of the eighteenth century, the primary source of power used in production was human and animal (Rosenberg and Birdzell, 1986). This dependence significantly restricted the size, scale, and location of industrial activity. With the ability to harness inanimate power sources, industrial activity moved toward a larger scale and a greater degree of specialization. First water, then charcoal, and finally coal became the defining power sources of the early industrial revolution. The industrial landscape began to spread out as manufacturing became unshackled from water power sources. Timber and charcoal were an intermediate solution; both, however, quickly declined in importance when most of the trees were cut down and charcoal became uneconomic to manufacture. With the discovery of large supplies of coal, production spread further and moved quickly beyond a relatively few industrial centers.

Technological change is a critical marker of the evolving industrial landscape. In addition to the ability to utilize inanimate energy sources, a hallmark of the industrial revolution was the development of machines that replaced human action. Metal cutting and metal shaping machines were vital to the formation of the industrial equipment industry. New materials and the understanding of the chemical basis of new materials made possible more complex manufacturing processes and new products. This in turn led to specific spatial concentrations of firms in the same industry, as divisions of labor formed and fanned out in independent enterprises.

The Organization of Industry in the Factory Era

The organization of capitalist enterprises into factories occurred as both merchants and manufacturers sought to control and coordinate an entire production process. During the craft era, workers exhibited some control over the labor process (declining through time). Capital's increasing span of control had important implications for the volume of output and therefore the size of markets that could be served, the level and source of innovations, and the sources and control of power. The need to harness these capabilities into a more centralized system of command and control had major implications for how industrial systems evolved. According to Lazonick (1991: p. 25), "The transition to the era of proprietary capitalism brought the integration of asset ownership with managerial control and with it an internalization of aspects of the craft system, in particular, in order for firms to generate (what were by the standards of the time) high-quality products at low unit costs." Further, "firms relied on an ample supply of highly skilled and well-disciplined labor. Senior workers—who eventually came to be known collectively as the aristocracy of labor—not only provided their own skills to the building and operating of machinery, but also recruited junior workers whom they trained and supervised on the shop floor."

The heavy reliance on skilled labor had obvious spatial implications. Firms were more or less bound to a particular

place. This dependence on labor is one of the prime determinants of spatially bound production systems, aptly captured by such well-known observers as Marshall (1997). The external economies within such a complex were substantial. Materials, labor, and services (distribution) were sufficiently honed that spatial proximity enabled work in progress to flow through its vertically specialized branches and to market its output (Lazonick, 1991: p. 26). These regional concentrations facilitated high levels of specialization. Consequently, employers' ongoing reliance on skilled labor to organize work on the shop floor diminished the need for business firms to invest in the development of managerial structure and organizational capabilities (Lazonick, 1991: p. 26). It also meant that pressures to cut costs were resolved through market transactions rather than through internal restructuring.

The system was self-perpetuating. Because proprietors had access to this complementary system of transactions, it was unnecessary to invest in or develop complex managerial structures. Instead, in the early phases of factory industrialization, proprietors tended to own single establishments and rely on the market for the provision of necessary inputs. Firms mostly remained small, and new establishments were formed to fill niches. The manufacturing process was inherently labor intensive and reliant on a cadre of skilled and unskilled workers to manipulate the raw material into final, finished form. The advantages of this supply system, combined with the benefits of a labor force that was self-reproducing, resulted in low fixed costs. Although problems of coordination would eventually become disabling, Britain, and other countries including Switzerland, built a globally competitive industry in the early nineteenth century based on just such a system of organization.

What unseated this remarkably flexible system was the emergence of mass production and the manufacture of volumes of products, all of more or less the same quality and capability. In the early era of mass production, the American system became the model to emulate. Unfolding in fits and starts, this system has been characterized as "the sequential series of operations carried out on successive special-purpose machines that produce interchangeable parts" (Hounshell, 1984: p. 15). Implicit in this description was the move away from a dependence on skilled labor and from a system based around fragmented and decentralized production producing highly variable, relatively high-cost goods (Hounshell, 1984). Costs had not been the only concern. Equally important were similarity of parts, interchangeability, and rapid production.

The Emergence of Managerial Capitalism and Factories

The transition to machine-based, vertically integrated, large-scale industrial manufacturing occurred over a long period of time. There are examples of metal cutting machines assisting in the rough shaping of parts for final finishing as early as the middle of the eighteenth century. However, it was the middle of the nineteenth century before widespread implementation of this technological and organizational innovation occurred. The movement toward this form of production was neither pervasive nor persistent. There were many obstacles to overcome, not least the construction of the necessary machinery and the resolution of the inherent conflicts presented by machine-based output to established labor relations and hence to the production system itself (Lazonick, 1991). Ultimately, the movement toward the American system caused turmoil throughout the entire system as machine-made parts and large-scale assembly yielded an unprecedented volume of goods. Many problems were associated with the flow of goods, and spilled over into the acquisition of consistent inputs and, equally important, the development of an effective means of distribution. Goods were no longer produced with a proximate market in mind. Indeed, the international supply of foreign markets became an increasingly important part of firm strategy. To produce at large scale necessarily meant distribution at large scale, ushering in the move toward managerial capitalism.

The evolution toward machine-based manufacturing and the ensuing complementary innovations in scale and scope of production, distribution, and management of the modern industrial enterprise arose rather gradually, starting in the 1870s (Chandler, 1991). The rise of the multiunit, capital-intensive corporation was the result of three investments. First, production levels had to be large enough to utilize economies of scale and scope of new technological innovations. Second, marketing and distribution capabilities had to be of a scale large enough to sell the goods produced by the new processes of production in the volume in which they were made. Third, a managerial hierarchy was needed to manage and coordinate the operations of large-scale production and distribution and to allocate resources for future production and distribution (Chandler, 1991: p. 432). That this new paradigm first took hold primarily in the United States and Germany has much to do with the relatively late industrialization of the two countries compared with Britain, and hence the absence of a preexisting labor process and managerial system to overcome. In the United States, the shortage of skilled labor placed a premium on developing machine-based alternatives to industrial production. Preceding the American Civil War and strongly afterward, American firms in a wide range of industries pursued mechanical solutions to the expansion of production to achieve economies of scale and serve burgeoning domestic and international markets.

The rise of the managerial corporation brought with it the integration of multiple functions, multiple units, multiple products, and all the attendant physical and human resources. Although each unit had independent capabilities to execute and manage the enterprise, coordination was offered and accomplished by middle managers who were themselves managed by top officials of the corporation. Quite strikingly, the emergence of this mode of organization and coordination arose rather quickly at the end of the nineteenth and the beginning of the twentieth centuries and was most observed in a subset of all industries comprising the US economy. These industries tended to be concentrated in certain industrial sectors, to consist of firms that were vertically integrated within the same industry group, and to be characterized by high-volume, capital-intensive production processes. In other words, what linked these firms were product characteristics that yielded the need for a uniform managerial structure designed to confront and control problems of production and distribution while benefiting from the potential associated with specific sectoral characteristics.

The sectors and firms most successful in executing the full complement of attributes of the modern corporation were those that benefited from the scale and scope linked to standardization of the production process. At the same time, however, there were other prerequisites to effective implementation, particularly relating to market control. To support the huge capital costs required to achieve economies of scale, a relatively few players had to exercise market power. The emergence of oligopolies in such sectors as oil, steel, and heavy machinery went hand in hand with the rise of the modern corporation. Control was both seized and achieved through investments in up- and downstream linkages, a capability that most smaller, more consumer-oriented goods producers could not make.

With the emergence of the managerial corporation, firms were now in a position to own and control resources and markets distributed across space. The spatial organization of the national economy evolved from production for local consumption, to regional markets of production and consumption, and finally to national markets and corporate structures of spatially decentralized production systems. The ability to standardize products and to produce at mass scale, twinned with the development of national brands and brand recognition, resulted in huge corporations that controlled vast business assets.

The Rise of Collective Capitalism

The model of market coordination embodied in managerial capitalism remained an effective organizing system until the 1960s. The lack of cutthroat international competition, and the presence therefore of stable markets, buoyant growth, and rising standards of living, provided hospitable conditions for the cultivation and persistence of the large, vertically integrated corporation. The seeds of turmoil that would come to destabilize this system of production, consumption, and market regulation were sown much earlier in the century, however, and laid dormant until after World War II. The challenge to this dominant system arose, ironically, as a new competitor – in this case Japan – coupled long-standing and underappreciated competency in the production of goods with state-supported (yet private-sector-led) manufacture of new products capable of competing directly in the world's most sophisticated industries and its most developed markets (Conroy, 1995). The national pursuit of both import substitution and export-led development policies provided infant industries with the protection required to achieve economies of scale and hence low overall costs. Accustomed to developing technological competency through technology transfer, the Japanese successfully adopted and adapted major innovations, primarily from the United States, to take the lead in key consumer and producer products over the postwar period.

The emergence of the Japanese as effective competitors in many consumer and producer product markets dates back to before World War II. In the last three decades of the nineteenth century, Japan had begun to develop the competency to produce tradable goods, primarily through the purchase of foreign know-how and technology. By the 1920s, in several key sectors, Japan was second if not the first in the world in terms of exports (Glasmeier et al., 1993). Although cut off from international trade in the 1930s, Japan continued to build up its productive capacity until just before the declaration of war. This time the emphasis was placed on heavy and chemical industries.

The emergence of collective capitalism as a distinct mode of organization is both a creature of history and a more modern institutional form. At the end of World War II, Japan had little choice but to pursue an export-led development strategy (McCraw, 1986; Samuels, 1994). The country lacked resources and had a small land base with a large population. In order to stave off poverty and economic despair, the nation had to develop export competencies in a variety of sectors. For reasons that can only be understood retrospectively, the Japanese did not choose to pursue products that were labor intensive despite its large labor force; rather, it moved rapidly into capital-intensive production as part of economic recovery in the 1950s.

The second major stimulant in the formation of collective capitalism was in part a holdover of the structure of the national economy that had developed since the beginning of the twentieth century. Japan's economy is composed of interlocking business organizations that enjoy both vertical and horizontal affiliations. Furthermore, the system of integration reaches down to the level of select blue-collar workers in first-tier firms who are offered lifetime employment in return for a life of dedication to the firm. In this way, guaranteed their jobs, workers willingly accept technological change as a part of labor–management relations.

Although many distinctions are attributable to the Japanese system, of particular importance is the role of the enterprise groups – networks of firms – that form the modern-day extensions of the *zaibatsu*, the family-controlled businesses that led the development of the Japanese economy from the end of the nineteenth century until the US occupation after World War II. These enterprise groups consist of a central business firm around which reside vast numbers of both affiliate and nonaffiliate subcontractors. This disintegrated system allows the core firm to enjoy the benefits of vertical integration, especially in terms of financing, technology transfer, and internal markets, without the disadvantages of ownership and bureaucratic control. The core firm is able to ensure quality, quantity, and timely and effective delivery of parts and services due to long-term contracts and investments made by the core firm into subcontracted operations (Lazonick, 1991).

The emergence of collective capitalism is essential in understanding the industrial geography of the early twenty-first century. Along with recognition of differences in the organization of industry, there was increasing interest in how culture – defined as the rules and practices, identities, and aspirations of individuals at the most intimate level – helps to shape economic space. Throughout the life of an industry, different conventions and rules of behavior form the basis for collective action. Organizational behavior and prevailing institutional norms structure the range of possibilities of firms and regions. This interaction forms the industrial landscape. Observers began to focus on elucidating the importance of culture as a determinant of corporate behavior and the ways culture manifests within people, which in turn shapes firms within regions (Schoenberger, 1997).

The role of culture in the diffusion of the industrial landscape became clearer as developing nations asserted their

positions as industrial powerhouses. Since the late 1980s, select countries in the Global South have experienced significant growth in manufacturing scope and export capacity. Countries such as South Korea and Taiwan over 30 years pursued strategies based on industrial upgrading and explicit price setting to achieve industrial policy targets, and exploited second-mover advantages to become supply chain leaders in increasingly sophisticated products including cell phones, advanced machine tools, and computer products. In contrast to the model of industrial supply chains composed of myriad small firms, Taiwanese industrialization took the form of gigantic vertically integrated firms of the likes of Foxcom, the singular supplier of many of Apple's most coveted products. South Korea and China among others have followed similar paths in their quest to become dominant global suppliers.

Each country's path to integration in the world economy has been distinct, varying by the direction of state-led policy and access to critical assets. After the collapse of the Soviet Union, Russia leveraged its vast natural resources, specifically in oil and gas, to rebuild its wealth despite its troubled transition away from a centrally planned economy. Brazil also benefitted from access to raw materials in advancing its steel industry, but since the stabilization of its economy in the mid-2000s, national leaders have turned their attention to strengthening industrial policies that have served the country well in the past. Low wages are facilitating the expansion of India's labor-intensive manufacturing sectors, although the country's initial transition from an agrarian society centered on noncapital-intensive industries of a service nature. Challenges related to infrastructure and land access hinder the speed of India's industrialization. China has built its economic dominance on its reputation as 'the workshop of the world,' focusing its state-driven efforts on maximizing exports and import substitution while placing restrictions on direct investment. Other emerging players, such as South Africa, are building on their strengths and attracting the attention of profit-seeking multinational corporations.

The Existence and Importance of the Creative Economy

Starting in the 1960s, increasing global competition in product markets, dramatic advances in information sharing and technology, and ongoing mechanization and outsourcing of traditional manufacturing jobs in industrialized countries accelerated the evolution of the knowledge economy. Industries using knowledge as a factor of production and those that created knowledge-intensive goods became key drivers of economic growth. One result of this shift was a change in the skill profiles that employers were seeking in both developed and emerging economies. Firms began to move away from the standardized and easily transferrable competencies of the assembly line and toward the tacit – unspoken or implicit – knowledge required for technology-heavy and highly interpersonal work.

At the turn of the century, the discussion of the knowledge economy narrowed to the concept of creativity as an economic growth engine. Richard Florida (2002) is credited for bringing the economic importance of creative workers to the forefront of popular dialogue. His 'creative class' broadly consisted of those developing new ideas and innovative materials in fields ranging from science and engineering to arts and entertainment. As the most critical input of production for the twenty-first century, theory dictated that the urban areas most effective in drawing employees from this class would increase their chances for successful development. City leaders quickly embraced strategies associated with constructing and promoting the amenities prized by creative workers. This approach was particularly enticing for declining industrial locations looking to recover from devastating losses of middle-class manufacturing employment.

The emphasis on creativity expanded to the global context, motivating some countries to adopt national policies that enhanced their appeal to creative industries, firms, and employees. Measures of creativity became a barometer for evaluating growth potential, and intergovernmental organizations advocated bolstering the creative economies of developing nations to provide income and employment while promoting social equity (United Nations, 2013). Cities ranging in scale from Rio de Janeiro to Kuala Lumpur have implemented interventions designed to encourage creativity and innovation (Evans, 2009).

Despite the vigorous transfer of creativity policies among cities across the globe, these approaches have not benefited all adherents equally (Peck, 2005). In a manner reminiscent of their traditional industrial predecessors, agglomerations of creative industries display patterns of geographical unevenness based on existing humans, built environment, and natural resources. Not all cities are able to parlay their assets into a setting attractive to a footloose sector of society not always uniform in its tastes. Concentrations of creative firms tend to form in urban areas that have well-developed horizontal producer networks, a diversified labor market with a mix of high- and low-skilled workers, and a culture supporting seamless knowledge exchange (Scott, 2006). Determinants of these critical elements are often associated with cities' historical functional specializations and locational endowments.

Today's industrial landscape is thus the sum of long-evolving processes, the residue of which is reflected in the structure and formation of business enterprises, industrial clusters, and systems of industrial activity. In this way, industrial geography is not only the study of a physical place per se but also an examination of the reservoir of experiences and actions that come to represent actors in a locality and its economy.

See also: Economic Geography; Economic Restructuring: Geographic Aspects; Globalization: Geographical Aspects; Historical Geography; Industrial Ecology; Industrialization, Typologies and History of; Industrialization; Knowledge Society, History of; Location Theory; Rural Geography; Spatial Interaction.

Bibliography

Chandler, A., 1991. Big Firms and the Wealth of Nations. Harvard University Press, Cambridge, MA.

Conroy, M., 1995. Industrial strategies, the newly industrializing economies, and new international trade theory in Latin America. Environment and Planning A: Commentary 27 (1), 1–10.
Evans, G., 2009. Creative cities, creative spaces and urban policy. Urban Studies 46, 1003–1040.
Florida, R., 2002. The Rise of the Creative Class. Basic Books, New York.
Glasmeier, A.K., 1990. A missing link: the relationship between distribution and industrial complex formation. Entrepreneurship and Regional Development 2, 315–333.
Glasmeier, A.K., Thompson, J., Kays, A., 1993. The geography of trade policy: trade regimes and location decisions in the textile and apparel complex. Transactions of British Geographers 18, 19–35.
Harrington, J.W., Warf, B., 1995. Industrial Location: Principles, Practice and Policy. Routledge, New York.
Hoover, E., 1948. The Location of Economic Activity. McGraw Hill, New York.
Hounshell, D.A., 1984. From the American System to Mass Production, 1800–1932. The Development of Manufacturing Technology in the United States. Johns Hopkins University Press, Baltimore.
Landes, D., 1969. The Unbound Prometheus. Cambridge University Press, London.
Lazonick, W., 1991. Business Organization and the Myth of the Market Economy. Cambridge University Press, New York.
Lösch, A., 1954. The Economics of Location, second ed. Yale University Press, New Haven, CT.
Marshall, A., 1997. Principles of Economics. Prometheus Books, New York.
McCraw, T., 1986. America versus Japan. Harvard Business School, Boston.
Peck, J., 2005. Struggling with the creative class. International Journal of Urban and Regional Research 29, 740–770.
Rosenberg, N., Birdzell, L., 1986. How the West Grew Rich. Basic Books, New York.
Samuels, R., 1994. Rich Nation, Strong Army, National Security and the Technological Transformation of Japan, first ed. Cornell University Press, Ithaca, NY.
Schoenberger, E., 1997. The Cultural Crisis of the Firm. Blackwell, Cambridge, MA.
Scott, A.J., 2006. Creative cities: conceptual issues and policy questions. Journal of Urban Affairs 28, 1–17.
Storper, M., Walker, R., 1989. The Capitalist Imperative. Blackwell, Oxford, UK.
United Nations Development Program, 2013. Creative Economy Report, 2013 Special Edition: Widening Local Development Pathways. UNDP, New York.
Weber, A., 1971. The Theory of the Location of Industries. Atheneum, New York.

Industrial Relations, History of

Patrick Fridenson, EHESS, Paris, France

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7344–7347, © 2001, Elsevier Ltd.

Abstract

In a society of individuals and of globalization, is the already-old US concept of industrial relations still valid? Is collective bargaining compatible with flexibility and competitiveness? On the other side, does it take into account stress and suffering at work or the social costs of unemployment? This article first makes the point that different types of industrial relations systems have developed: Britain, continental Europe, and the USA-plus-Japan model. Then it shows that wage-earners have always had various ways of expressing themselves within firms and in the public sphere. Trade unions are only one of them. Unions, which have recently undergone a severe crisis almost everywhere, cannot be stable partners, as was initially argued in the USA, but they remain significant whenever they consider the diversity of the labor force and the interests not only of the firm but of society at large. Therefore the recent critiques of the concept of industrial relations require us to revisit it.

In most developed countries, wages and working conditions, forms of employment, bonuses, and allowances are regulated, either by law or by collective agreements. The latter are themselves bordered by juridical or customary rules which define their beginning, process, and results. Their coherence does not merely reflect the logic of the national institutions and of the economy. It takes into account the social actors themselves and the aggregation of their decisions, which constitute industrial relations systems. They vary throughout the course of history. They also differ amongst North America, Japan, and Europe and within Europe itself. But everywhere these systems have been deeply altered or even challenged since the 1980s.

The Concept of Industrial Relations

The term 'industrial relations' was born in the USA. It is the product of both changing relations between companies and wage-earners and of academic attempts to instil some order into the turmoil of social change.

Commons

John R. Commons, an economist at the University of Wisconsin (Madison), produced, during the first third of the twentieth century, the first major interpretation of social relations in 'American industrial society.' He argued that US workers were more wage-conscious than class-conscious. If companies became aware of these fundamental tendencies, there was room for contracts between management and trade unions. Conflicts could be avoided, minimized, or solved. Government was welcome to promote economic growth by supporting the contractualization of industrial relations. Commons and his followers stressed the importance of institutions for organizing society. Accordingly, they considered trade unions to be a source of social progress. They privileged collective bargaining as fitted to US exceptionalism and adapted to unionization by trade, not by industry.

Commons and his followers were not pure and simple academics. They actively contributed to the liberalization of US social legislation which developed between 1918 and 1940.

Dunlop

John T. Dunlop, a sociologist at the University of Wisconsin, working between the 1950s and the 1970s, went further in the same direction. He systematized the idea of systems of industrial relations at local, regional, or national levels. Thus he stressed the interdependence between the key elements of industrial relations: for instance, between the definition of a trade union, its representativity, and the nature of a negotiation. This interdependence is characterized not only by a common juridical doctrine, but also by reciprocal strategies carried out by managers and unions. Moreover, rules in industrial relations need to be interpreted by reference to the policies of unions, companies, and trade associations. Social actors create rules and bind each other by such regulations, at least for a specific period of time. Like Commons and his followers, Dunlop and his colleagues worked as experts for public administrations and played a significant part in the evolution of collective bargaining and of legislation till the first oil crisis. Unlike Commons, they moved from the idea of US exceptionalism to its apparent opposite: a possible universal convergence between the national systems of industrial relations. 'Industrialism' would bring about a civilization of social relations everywhere by generating stable processes of collective bargaining. Meanwhile, Dunlop recognized the plurality of national and continental systems.

Three Models in Europe

At the end of the nineteenth century and during the twentieth, Europe developed specific features in its industrial relations systems, which are still in force. Collective bargaining is first and foremost the responsibility of employers' associations and trade unions by trade or sector, coordinated by confederations. Bargaining at company level came later, is growing, but remains under the aegis of federations. A distinction is made between collective bargaining, which deals mostly with wages and working time, and participation of wage-earners, which includes work conditions, welfare, and the adaptation of a broader collective agreement to a firm; also, more recently,

information on the firm's strategy and its impact on employment. In Germany, participation extends to codetermination, introduced under the Weimar Republic and generalized in the aftermath of World War II, by laws passed in 1951–52. Governments are involved in the national regulation of industrial relations. Intra-European variety may be reduced to three types of patterns. The patterns are quite different on other continents.

The UK and the Republic of Ireland

This is the oldest industrial relations system. Trade unions are closely related to trades and collective bargaining is basically at the level of the firm or the workshop. Conversely, strikes are frequent at this level, yet they do not foster political radicalization. Labour legislation is light. Politics and society tend to privilege the freedom of action of employers and unions.

Central and Northern Europe

This group comprises Germany, Austria, Switzerland, and the Scandinavian nations. It is characterized by a trend towards uniqueness in several forms: uniqueness of employers' associations – powerful, structured, obeyed; uniqueness of trade unions – strong and connected to political power; uniqueness of social conflicts – abiding in written or implicit rules. Collective bargaining is centralized by branch. So are strikes, limited to the period of negotiation. In these countries government intervenes by legislating on industrial relations. However, employers and unions have moved to restrict its interference. The Netherlands and Belgium may be included in this group, although both have a plurality of trade unions.

The Latin Countries

This group covers France, Italy, Spain, Portugal. The keyword here is pluralism: pluralism of conflicts, with a significant percentage of preventive strikes and of wildcat strikes; pluralism of trade unions, with a pole of socialist origin, another of Christian origin, and one which calls itself autonomous or independent; pluralism of political influences among wage-earners. In addition, there is limited authority and even, for many years, pluralism of employers' organizations themselves. Collective bargaining long proceeded only at the national level (for one branch or for the general economy), except at the end of local strikes. Bargaining at company level became a more regular practice in the mid-1950s or even later. Government regulates a variety of issues, except in Italy.

North America and Japan

North America

The US labor movement long featured a pluralism of organisations (in fact till 1955), and employers were as reluctant as elsewhere to negotiate with unions. Things changed with the New Deal. The 1935 Wagner Act, however deeply modified by the 1947 Taft–Hartley Act, made collective bargaining a right for employees, but a union had to be accredited by the vote of a majority of them as the sole counterpart of management in collective bargaining. Law deemed collective bargaining to progress 'in good faith,' i.e., to aim at an agreement. Thanks to the sitdown strikes of the mid-1930s, bargaining spread all over the USA and its contents broadened. The areas of bargaining are now wages, working hours, and work conditions. They may include other matters which have consequences for these elements. The closed-shop clause has been declared illegal. As for Canada, it generally followed the US trends.

Japan

Up to Japan's defeat at the end of World War II, collective bargaining was exceptional in Japanese companies and would materialize only at the end of strikes. Practices and legislation changed in the late 1940s, under the influence of Japanese liberals and US occupying forces. Collective bargaining became a constitutional right for wage-earners. A company has to bargain with any union which asks for it, and even with any group of employees which does not constitute a union, provided it can express a common will. The employer must present counterproposals to the demands of the employees and negotiate 'loyally.' The contents of collective bargaining are extremely broad. A wave of strikes enabled both the spread of unions and of collective bargaining in postwar Japan. However, radical unions were finally defeated; worker activism and protest gave way to management victory in the 1960s and 1970s. Company unions took the upper hand in pursuit of economic affluence. Collective bargaining became a type of cooperation with management, and often quite an influential one. This evolution laid the foundations for a corporate-centered society.

In North America and in Japan, when there is no union in the workplace or no agreement is reached, the rights of the employee may be in fact minimal and management keeps its unilateral powers.

Recent Changes

The End of an Era

Many authors consider the stages reached by national industrial relations systems after World War II to have been parts of social settlements built after the strains of the international economic depression of the 1930s and of the wars of the early 1940s, paving the way for a new era of growth. When growth decelerated in the time of oil crises in the 1970s, when national tariff barriers were lowered, and neoliberal ideas made a spectacular headway in the West, industrial relations systems came under attack. At the same period in a number of countries, though not all, membership of unions started to decline. The breakdown of the population changed from a majority of blue-collar workers to an increasing proportion of white-collar employees, in keeping with a growing proportion of women at work and a shift from industry to services. Unemployment simultaneously rose.

A Shift in Industrial Relations

On a world scale, trade unions entered an era of concessions to employers in terms of wages. The flexibility of work conditions

grew. Mechanisms of social protection weakened. Companies moved to a greater individualization of wages, careers, even contracts, and to a more intense involvement of each employee in the workplace. Integration quite often became their motto. Japan was the last country to undergo this new trend and to try and adjust to the challenge. In many countries militancy and the yearly number of strikes and strikers declined. Yet both large companies and governments stressed the importance of collective bargaining to accompany such far-reaching changes and to stabilize the emerging global competitive order. In some countries ways were found to integrate small and medium-sized enterprises in industrial relations systems. Outside the USA the reduction of working time became a target for collective bargaining and legislation. In each country some of the trade unions began to adapt and modernize their strategies and structures. Some were able to regain members in new working groups. But altogether their societal position has not yet been stabilized and the balance between individual and collective has been altered in significant aspects, in favor of the former.

Current Debates

The Impact of New Practices on the Paradigm

These changes in industrial relations themselves have caused new debates among scholars about Dunlop's paradigm. The development of framework agreements, of agreements on discussion methods, of agreements setting a policy and targets rather than rights, in short of agreements which are not contracts, is quite different from the explicit forms and the strict obligations characterizing most of collective bargaining in the USA during the 1950s and the 1960s. Collective actors appear less neatly defined, more unstable, and the representativity of unions is under redefinition in some countries. In short, actors may become more defined by action itself. A number of scholars argue that such developments do not invalidate Dunlop's paradigm, as it is based on autonomous actors able to create rules and to keep their word.

Recurrent Critiques

Simultaneously older critiques of the paradigm find a greater audience. A number of authors have stressed the importance of unorganized conflict, expressed by absenteeism and turnover. They also stress that the US model of the social conflict as a simple form of adjusting industrial rules may reduce its dimensions and that trade unions are not everywhere limited to the provision of personal services to wage-earners. These authors doubt the possibility of a convergence of national industrial relations systems, despite the development of multinational companies and of continental economic unions. Is it still so easy to say that these systems are functional to industrialization? Are the frontiers of these systems clearly defined, as they indeed produce part of their economic or technological 'context' themselves? Are they not less stable than their earliest analysts implied? What are the multiple sources of their dynamics? Therefore, research on industrial relations is coming back on the agenda of social sciences, as well as of organizations and of individuals.

See also: Business History; Class: Social; Economic History; Industrialization, Typologies and History of; Labor Supply; Labor Unions; Trade Unions and Social Work: Lessons from Canada; Trade Unions, Economic Behavior of; Work, History of; Working Classes, History of.

Bibliography

Dunlop, J.T., 1970. Industrial Relations Systems, second ed. Southern Illinois Press, Carbondale, IL.

Giraud, O., Lallement, M., 1997. L'institutionnalisation des relations professionnelles en RFA. Entreprises et Histoire 6 (October), 36–47.

Gordon, A., 1998. The Wages of Affluence: Labor and Management in Postwar Japan. Harvard University Press, Cambridge, MA.

Morin, M.L., 1994. Le Droit des Salariés à la Négociation Collective, Principe Général du Droit. Librairie générale de droit et de jurisprudence, Paris.

Reynaud, J.D., Eyraud, F., Paradeis, C., Saglio, J. (Eds.), 1990. Les Systèmes de Relations Professionnelles. Examen Critique d'une Théorie. Editions du CNRS, Lyon, France.

Slomp, H., 1995. Between Bargaining and Politics. An Introduction to European Labor Relations. Praeger, Westport, CT.

Industrial Societies

Alf Hornborg, Lund University, Lund, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of 'industrial society' is generally used for a type of social organization based on mass production of commodities. This type of societies began to dominate Western Europe toward the end of the eighteenth century, reflecting a strategy to expand export production in order to appropriate embodied labor and resources from elsewhere in the world-system. Perceived as progress and 'development,' this transition has shaped hegemonic worldviews and visions of a desirable future for all humankind, but current concerns with inequalities, resource depletion, and climate change suggest that such aspirations are not globally sustainable. Industrialization and urbanization tend to fundamentally transform sociocultural conditions and human–environmental relations. Anthropological perspectives can illuminate how the cultural inclinations and self-representations of industrial societies are geared to global political economy, sociology, and ecology.

The concept of 'industrial society' is generally used for a type of social organization that emerged in Western Europe after the so-called Industrial Revolution toward the end of the eighteenth century. It has affinities with concepts such as 'modernity' and 'capitalism' and is often thought of as based on Enlightenment ideals of rationality and progress. As it tends to reflect a cultural self-representation of economically affluent or 'developed' nations, the concept deserves critical cultural analysis and needs to be contextualized in time and space. This article approaches the phenomenon of industrial societies through the macro-level perspectives of world history, political economy, and global ecology, while indicating how a cultural critique rooted in anthropology can complement the accounts of history and sociology.

World-Historical Prerequisites

Industrial societies are often contrasted against 'preindustrial,' generally agricultural societies with regard to fundamental economic strategies. The most central connotation of the word *industrial* is undoubtedly an inclination toward mass production of commodities. The Industrial Revolution in England was essentially a strategy to radically expand the production of textiles for export, originally woolens but increasingly also cotton cloth. This strategy should be understood in relation to both the global historical context and the local ecological conditions in England at the time. English textile manufacturers had been competing for market shares with those of other nations such as the Netherlands. Rather than continue to export raw wool to continental Europe, England in the eighteenth century increasingly turned to manufacturing woolen textiles for export. The rising demand for wool prompted large land owners to expel peasants deriving their subsistence from small-scale agriculture in order to replace them with sheep. The rising numbers of landless people provided a source of (wage) labor for the expanding textile industry. At the same time, the incentive to expand production at lower cost per unit of cloth encouraged technological rationalization, including the large-scale use of water mills. The same incentive stimulated the development of the steam engine, which was the first machine capable of converting fossil fuels (coal) to mechanical energy. Britain's abundant deposits of coal had long been used as a source of heat for both domestic use and manufacturing, in large measure as a less expensive alternative to increasingly depleted sources of firewood, but the Industrial Revolution represented a decisive historical discontinuity largely defined by the adoption of fossil fuels as a source of mechanical energy. The full potential of this new mode of production was realized as British textile manufacturers turned increasingly to cotton fiber from the colonies, which, like the subterranean coal deposits, relaxed the pressure on domestic land. By the end of the nineteenth century, the displacement of Britain's environmental footprint through imports of cotton, sugar, timber, and a number of other products amounted to several times its total land area (Pomeranz, 2000).

The expansion of industrial production in England in the late eighteenth and early nineteenth century was importantly geared to the Atlantic slave trade. First, African slaves provided the bulk of the labor in the colonial plantations that supplied the British cotton textile industry with raw material. Second, the colonial slave owners were major buyers of cotton textiles, as they required inexpensive clothing for their slaves. Third, the cotton textiles from Britain were able to oust the rival Indian textiles that had traditionally been used in trading with slave dealers along the west coast of Africa. Fourthly, the rational organization of production in the colonial plantations served as a model for the design of industrial production processes in England. In other words, the products of the cotton textile industry were used to both purchase and clothe the labor force that provided it with raw material and served as a model for the organization of production. The steam engine, the very rationale of which was founded on the triangular trade between Britain, West Africa, and the American colonies, was the crucial factor in giving Britain the capacity to oust the Indian textile industry on the world market. Later on, it propelled the railways and steamboats with which Britain was able to consolidate its colonial empire. These new modes of rapid and reliable transportation were used not only to move materials, commodities, and labor, but also to deploy troops and apply military coercion in different parts of the world (Headrick, 2010).

Cultural Perspectives on Technology and Economics

Although the incentive to increasingly mechanize production is conventionally elevated to a general principle of technological

progress, it is important to recognize this conception as, to a large extent, a reflection of the historical experience of core sections of the capitalist world system (Adas, 1989; Marsden and Smith, 2005). Technological progress in certain parts of the world frequently hinges on environmental load displacement to other parts, as illustrated by British industrialization in the nineteenth century. Mechanization and industrialization should thus be understood as socioecological strategies rather than mere feats of engineering, particularly since the very feasibility of a given technological system is determined by relative prices of labor and resources in different parts of the world. Although there is a general assumption of continuous evolution, over several millennia, of ever more efficient technologies, the cultural image of technology that organizes conventional historiography reflects a mere three centuries of developments, which are currently being reevaluated in terms of global inequalities and environmental repercussions. To the extent that technological developments in core nations have been based on asymmetric appropriation of embodied labor and land from their peripheries, mainstream visions of 'development' can obviously not be accepted as models for the so-called 'developing nations.' This conclusion is underscored by the apparent, biophysical impossibility of universalizing the patterns of consumption currently enjoyed in 'developed' nations. It has been estimated that if such patterns of consumption were universalized across the globe, the consequent resource demands would require the land surfaces of several additional planets. A similar conclusion is evident from the unlikely prospects of universally implementing industrial technologies in agricultural production. The application of fossil fuel-based technologies to food production has established profitable export industries in several parts of the world, but their dependence on nonrenewable resources, wasteful energy balances, and the degradation of soils and ecosystems precludes their universalization as components of a sustainable development.

The industrial mass production of commodities for export was not an invention of the eighteenth century. There are plenty of historical antecedents going back to the earliest urban civilizations in China, India, and Mesopotamia, but the application of fossil-fuelled steam technology in combination with access to American land and African labor enabled Britain to transcend its own ecological constraints in a self-reinforcing process of capital accumulation that became foundational to industrial society as we know it. The societal transformations generated by these processes were unprecedented, both in the European core nations and in their overseas peripheries. Various attempts to interpret these transformations constitute the bulk of social science over the past two centuries. Fields such as economics, sociology, anthropology, political science, and development studies would be inconceivable without them. To this day, their mainstream understandings of growth, modernity, and development are fundamental to globally shared views of the world.

Economic theory was fundamentally transformed by the historical experience of industrialization (Polanyi, 1944). Prior to the Industrial Revolution, schools of thought such as the Physiocrats perceived agricultural land as the only factor of production capable of yielding a net value. While this economic worldview persists in some parts of the global periphery, industrializing core countries increasingly experienced land as a dispensable asset. Whereas Thomas Malthus worried about how access to land would ultimately constrain economic growth, Adam Smith and David Ricardo argued that, with free trade, labor and capital could compensate for a lack of land. This new economic worldview reflected the experience of Britain in the nineteenth century in two important ways. First, the requirements on land had been disconnected from the need for transport energy, which could now be derived from subterranean deposits of fossil fuels rather than fodder for draft animals. Second, the products of an increasingly mechanized industry could be exchanged on the world market for increasing quantities of land-intensive commodities from abroad. Karl Marx, too, was generally optimistic about the prospects of labor and technology, rejecting Malthus's worries about ecological constraints. In contrast to Smith and Ricardo, however, Marx emphasized that industrial capitalism was founded on accumulation, unequal exchange, and class conflict. These three strains of social thought have continued to provide the basic framework of social philosophy up to the present. Smith's and Ricardo's understanding of the world has largely guided mainstream economics, whereas Marxian and neo-Malthusian (ecological) critiques have continued to challenge it.

Contributions from economic anthropology, such as Stephen Gudeman's (1986), approach economic worldviews as cultural systems that can be analyzed in terms of tacit metaphors and implicit assumptions. From this perspective, the science of economics that emerged in tandem with industrial society can be viewed as this society's dominant mode of self-representation. Its inclination toward formalism and abstraction thus reflects the ambition of nineteenth-century Europe to encompass the cultural diversity of the world it was set to conquer. Its faith in the benevolence of what Smith called the 'invisible hand' of the free market suppressed and replaced concerns about inequality and conflict, such as are expressed in the 'image of limited good' that tends to predominate in peasant societies. This denial of unequal exchange is related to its exclusive concern with monetary exchange values, which ignores the asymmetric flows of materials, energy, embodied land, and embodied labor on which industrial production is ultimately dependent. Marxian analyses of capital accumulation and the neo-Malthusian concerns of ecological economics have revealed how such assumptions of mainstream economic science reproduce a specific worldview that is not only cultural but also clearly ideological.

Of central significance in this worldview is the faith in technological progress. While conventionally perceived as a progressive increase in efficiency propelled by human ingenuity, current concerns about environmental degradation, climate change, and global inequalities raise serious doubts about mainstream cultural conceptions of technology. Rather than a set of purely material engineering strategies for locally solving given tasks, technologies may be viewed as thoroughly social strategies for redistributing time and space in global society (Hornborg, 2011). For a given social category, access to a specific technology may be tantamount to saving time and space, e.g., through increasing speed and a more efficient utilization of land, but citizens of industrial societies tend to be unaware of the extent to which this local saving of time and space would be unthinkable without significant losses

of human time and natural space elsewhere in the world system. The social dimensions of such technological arrangements protrude most clearly when we consider how they are consistently mediated by the market. The rationality of any technology is contingent upon relative prices of labor and resources in different parts of the world. The apparent cornucopia of technological efficiency in certain sectors of the world system is thus continuously subsidized by relatively low-paid labor and resources in other sectors.

To understand industrialism as a specific social form, no less than to understand other types of societies, it is important to conceptualize how cultural and metabolic phenomena interact. Whereas human societies for many thousands of years had been organized around the harvesting of solar energy in the form of food and fodder, the experience of European merchant capitalists in the sixteenth century promoted monetary exchange value as the central incentive of economic life. In other words, money at this time became as widely coveted a vital essence as energy. The difference between money and energy, of course, is that money is ultimately only a cultural concept without any intrinsic metabolic potential. Nevertheless, the accumulation of money in Western Europe prior to the Industrial Revolution was destined to completely reorganize the metabolism of societies throughout the world. Its immediate effects were to set masses of African slaves working on American plantations and landless Europeans toiling in mines and factories, transforming societies and landscapes on both sides of the Atlantic. Its long-term effects are yet to be determined, but contemporary concerns with global inequalities, sustainability, peak oil, and climate change suggest that the future of industrial society poses very significant challenges for humankind.

The interaction of cultural and metabolic phenomena can be illustrated by processes of capital accumulation. In his classical account of industrial capital accumulation in nineteenth-century England, Karl Marx observed that the growth of material infrastructure in the British textile industry was intertwined with the flows of exchange values. The scope for investments in further accumulation of real capital was determined by the success of its owners in amassing monetary profits from selling textiles and other industrial products at prices higher than their costs for wage labor and other inputs. Rather than refer to a labor theory of economic value, it suffices to observe that, for industrial capitalism to be feasible, the price of labor must be less expensive than the price of its products. This observation could be extended to include other inputs such as fuels and raw materials. The central point here, however, is that the accumulation of material infrastructure hinges on social relations of exchange, and that these exchange relations are in turn mediated by cultural conceptions such as 'market price' and 'wage.' This illustrates how cultural concepts can be essential to the formation of physical infrastructure, although our Cartesian inclination to distinguish mind from matter tends to preclude such conclusions. It also enables us to propose a crosscultural definition of capital accumulation because it can be argued that it is a universal feature of productive infrastructures – whether textile factories or agricultural terraces – to intertwine, as requisite to their very existence, cultural concepts and material metabolism. For example, the agricultural terraces of ancient Inca society were built and maintained through ritual exchanges of services mediated by shared conceptions of obligations between the emperor and his subjects. Concepts of 'price' and 'wage,' it can be argued, are no less cultural than these ancient Andean conceptions. In both cases, cultural concepts orchestrate and justify material flows of labor and other forms of energy.

Industrialization and the Sociocultural Condition of Modernity

Industrial society has from the very start implied a radical transformation of human–environmental relations. Urbanization and employment in industry have in various ways detached humans from the ecological systems from which they continue to derive their fundamental metabolic requirements such as food, drink, oxygen, and a comfortable temperature. The modern, urban perception of 'nature' or the 'environment' represents a more abstract and distant view of such material connections than the eco-cosmologies encountered by anthropologists among nonindustrial societies intimately engaged in drawing their sustenance from their immediate surroundings. The physical alienation of humans from natural ecosystems thus generates not only material, ecological problems anticipated in the Marxian identification of a metabolic rift between town and countryside but also a cognitive or cultural incapacity to grasp the connectedness between humans and the remainder of the biosphere. Modern people are generally unable to perceive or estimate the social and ecological consequences of their consumption patterns, which means that the reach of their moral responsibility is much more restricted than their increasingly globalized economic and ecological reach. This predicament was acknowledged already in the Marxian concept of commodity fetishism, but has recently often been referred to as 'consumer blindness.' Such cognitive constraints, of course, aggravate the difficulties of identifying and addressing the mounting, global environmental problems of industrial society. It is, for instance, difficult for individual consumers to sense a connection between their purchases and the abstract threat of climate change.

Sociological theory was born in the nineteenth century as a mode of reflecting on the conspicuous transformations within European society. Large shares of the population shifted from rural community life to a more anonymous, urban existence in the industrial centers and submitted to a rigorous regimentation of everyday work (Frykman and Löfgren, 1987). Social theorists such as Emile Durkheim, Max Weber, Ferdinand Tönnies, and Georg Simmel developed conceptual frameworks for grasping this shift and its many social and existential implications. Karl Marx's concept of alienation was also a contribution to this discourse. In the twentieth century, sociological understandings of the 'modern' condition became increasingly sophisticated. In recent decades, sociologists such as Anthony Giddens (1990) and Zygmunt Bauman (1991) have synthesized the most salient features of this condition. Although it must be remembered that sociological models are ideal types that do not claim to be universally applicable, there are undoubtedly characteristic features of modern life that contrast in significant ways with the kind of life generally experienced in pre- or non-modern contexts. Central aspects of the modern condition include an unprecedented individualism and

reliance on abstract systems of trust such as money and expert knowledge.

The social condition of modernity is generally also reflected in a pervasive worldview, including components such as mind-body and culture-nature dualism, a submission to standardization and discipline (including clocks, work schedules, and calendars), and an orientation toward the future. These cognitive orientations toward dualism, homogenization, and progress have continuously been challenged by dissidents emphasizing the contrary values of monism, diversity, and stasis. Thus, in recent decades, ecofeminists have challenged the Cartesian inclination to objectify nature; postcolonial theorists have challenged the notion that European definitions of order can be universalized; and ecological economists have advocated steady state economics and even degrowth. Such cultural countercurrents are in a sense 'antimodern,' but can be traced historically as pervasive subcultures of modernity from their very beginnings. A matter of contention is the extent to which the Enlightenment should really be viewed as an historical discontinuity. The conventional view is that Enlightenment ideals of rationality and efficiency liberated the potential of modern economics, technology, and politics to disengage from the cultural constraints that had fettered these domains and practices in premodern societies. However, as we have seen, there is currently a good basis for arguing that modern conceptions and practices of economics, technology, and politics remain thoroughly embedded in culture, even if this is not recognized in mainstream ideology.

The extent to which modernization also implies secularization is not self-evident, in part because significant proportions of the population in some modern nations adhere to a religious or at least spiritual worldview, in part because some components of the modern condition (e.g., consumption, or faith in economic and technological progress) arguably have ritual or even religious aspects. This suggests that the so-called Enlightenment did not represent a decisive historical discontinuity in the sense we are accustomed to believing. Beginning with concepts such as 'money fetishism' and 'commodity fetishism' introduced by Karl Marx in the mid-nineteenth century, several critiques of industrial civilization imply that modern phenomena such as economy and technology contain elements that have affinities with the 'magic' that is commonly dismissed as premodern (Taussig, 1980; Graeber, 2011). As Marx realized, the specific cultural idiosyncrasies of industrial capitalism are most clearly revealed through comparison with contrasting cultural conceptions and practices, a method referred to as 'defamiliarization by cross-cultural juxtaposition' (Marcus and Fischer, 1986).

The centrally modern phenomenon of consumption, for instance, can only be understood through cultural analysis (Sahlins, 1976). Jean Baudrillard (1972) and Pierre Bourdieu (1984) have shown that specific patterns of consumption serve to communicate messages about a person's identity. The scope for consumption, viewed as a distinct social activity, has widened with increasing commercialization and affluence, while also serving as a modern source of meanings compensating for a general loss of community, tradition, and existential coherence. In this sense, the driving forces of consumption, and thus also of economic growth, appear to be connected to the existential predicament of individualism and alienation. In preindustrial societies, conspicuous patterns of consumption signaled elite status, whereas in industrial societies the communication of social identity through consumption has been generalized to all classes. Particular types or brands of commodities can thus be broadly associated with particular identities or even occupations. However, the communicative and semiotic dimensions of consumption present a shifting field of meanings, where different social groups continuously attempt to emulate or distinguish themselves from each other. Such cultural processes have been of central significance for patterns of world trade, production, and environmental impacts throughout history, but the affluence and mass consumerism of modern industrial societies have intensified their economic and ecological repercussions (Mintz, 1985).

Development, Underdevelopment, and Sustainability

The expansion of affluent, industrial societies in Europe in the nineteenth century was inextricably intertwined with the social and ecological impoverishment of their extractive peripheries (Wolf, 1982). The development of modern societies and lifestyles in some parts of the world system thus implicated the often detrimental historical trajectories of other parts. Although specific developments in the periphery were frequently induced by interventions from and relations with the core, the resultant differences were generally interpreted as representing separate developmental or evolutionary stages in time. The discipline of anthropology emerged as an auxiliary of colonialism in the late nineteenth century. Its first decades were dominated by an evolutionist outlook that represented industrial civilization as the apex of development, indicating a desirable future for the less advanced nations in the periphery. While much anthropology in the twentieth century has continued to advocate 'development' in former colonies, it is difficult to clearly distinguish the extent to which this ambition remains an heir to Europe's civilizing mission, and the extent to which it represents concern over the detrimental consequences of colonialism (Ferguson, 1999).

The recognition that societies or social systems encompass wider fields of interaction than bounded nation states has been most explicitly articulated in the so-called dependency theory of Andre Gunder Frank (1967) and the world-system analysis of Immanuel Wallerstein (1974–1989). Such global perspectives on national trajectories have been particularly influential in sociology, political science, economic history, and development studies. The common denominator of these perspectives is the recognition that conditions within a given nation are connected in significant ways to its history of exchanges with other nations. Frank and Wallerstein have demonstrated that the economic and technological expansion of early modern Europe was conditioned by global conjunctures, and several scholars have challenged the conventional Eurocentric historiography that accounts for this expansion in terms of uniquely European features of culture, society, or geography. A recurrent explanation of European expansion since the sixteenth century focuses on the unusual alliance between merchant capital and national governments, which provided long-distance traders with military support. The non-Eurocentric historians instead emphasize Europe's position as middleman between New World resources and Old World markets, particularly its ability to use American silver in its

trade with China. As we have seen, the sixteenth-century European conquest of the New World and the subsequent trans-Atlantic slave trade were undoubtedly essential prerequisites of the Industrial Revolution.

The precarious predicaments of industrial society have generated influential political discourses that often serve to avert from the undeniable dilemmas of mounting inequalities, resource depletion, environmental degradation, and climate change. These discourses permeate the research agendas of sciences such as economics, engineering, environmental science, and development studies. One by one, the solutions that have been proposed have generally been dismissed by other researchers as components of the utopian, modernist mythology of industrial society. For example, the faith in a postindustrial dematerialization of the economy has been dismissed as neglecting the extent to which it tends to rely on a displacement of material-intensive industries to other parts of the world system, and the faith in responsible consumerism and recycling has largely dissolved in view of the discouraging records of such policies. Similarly, the faith in substituting for fossil fuels with technologies such as agro-fuels and photovoltaic cells has widely been rejected with reference to the vast land requirements, low energy returns, doubtful environmental benefits, and/or prohibitive costs of these technologies. The recent financial crises of industrial societies in North America and Europe contribute to a widespread anticipation of a major shift in the cultural worldview that has accompanied the global expansion of industrial societies over the past two centuries. A crucial component of such a shift would be to promote policies to distinguish economic globalization from globalization in a purely communicative sense. The continued integration of world society as a network of information flows would not necessarily entail the continued intensification of global flows of matter and energy. To rethink conventional economics, engineering, and political science with this objective in mind is currently a central challenge of industrial society.

See also: Anthropology and History; Anthropology at Home; Center–Periphery Relationships; Civilizations; Colonialism, Anthropology of; Cultural Critique, Anthropological; Development: Social-Anthropological Aspects; Economic Anthropology; Energy, Anthropology of; Fetishism; Globalization, Anthropology of; Markets: Anthropological Aspects; Modernity: Anthropological Aspects; Political Economy in Anthropology; Postcoloniality; Postindustrial Societies; Slavery: Comparative Aspects.

Bibliography

Adas, M., 1989. Machines as the Measure of Men: Science, Technology, and Ideologies of Western Dominance. Cornell University Press, Ithaca.
Baudrillard, J., 1972. For a Critique of the Political Economy of the Sign. Telos, St. Louis.
Bauman, Z., 1991. Modernity and Ambivalence. Cornell University Press, Ithaca.
Bourdieu, P., 1984. Distinction: A Social Critique of the Judgment of Taste. Harvard University Press, Cambridge.
Ferguson, J., 1999. Expectations of Modernity: Myths and Meanings of Urban Life on the Zambian Copperbelt. University of California Press, Berkeley.
Frank, A.G., 1967. Capitalism and Underdevelopment in Latin America. Monthly Review Press, New York.
Frykman, J., Löfgren, O., 1987. Culture Builders: A Historical Anthropology of Middle-Class Life. Rutgers University Press, New Brunswick.
Giddens, A., 1990. The Consequences of Modernity. Polity, Cambridge.
Graeber, D., 2011. Debt: The First 5,000 Years. Melville House, New York.
Gudeman, S., 1986. Economics as Culture: Models and Metaphors of Livelihood. Routledge & Kegan Paul, London.
Headrick, D.R., 2010. Power over Peoples: Technology, Environments, and Western Imperialism, 1400 to the Present. Princeton University Press, Princeton.
Hornborg, A., 2011. Global Ecology and Unequal Exchange: Fetishism in a Zero-Sum World. Routledge, London.
Marcus, G.E., Fischer, M.M.J., 1986. Anthropology as Cultural Critique: An Experimental Moment in the Human Sciences. The University of Chicago Press, Chicago.
Marsden, B., Smith, C., 2005. Engineering Empires: A Cultural History of Technology in Nineteenth-Century Britain. Palgrave, Houndmills.
Mintz, S.W., 1985. Sweetness and Power: The Place of Sugar in Modern History. Penguin, New York.
Polanyi, K., 1944. The Great Transformation: The Political and Economic Origins of Our Time. Beacon, Boston.
Pomeranz, K., 2000. The Great Divergence: China, Europe, and the Making of the Modern World Economy. Princeton University Press, Princeton.
Sahlins, M.D., 1976. Culture and Practical Reason. The University of Chicago Press, Chicago.
Taussig, M., 1980. The Devil and Commodity Fetishism in South America. University of North Carolina Press, Chapel Hill.
Wallerstein, I., 1974–1989. The Modern World System I–III. Academic Press, New York.
Wolf, E.R., 1982. Europe and the People without History. University of California Press, Berkeley.

Industrialization

Richard Biernacki, University of California, San Diego, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sociohistorical investigations of industrialization in the past decade have overthrown inherited models of the uniqueness of early-modern commerce and technology in Europe, path-dependent development, and technological determinism. The global economy has entered a novel stage of institutional experiment and organizational diversity within nation-states.

The past decade has witnessed profound shifts in sociohistorical studies of industrialization. Industrialization is still defined as the process of applying mechanical, chemical, and electrical techniques to reorganize production with inanimate sources of energy. This criterion of industrialization in no way entails a technological determinism. On the contrary, comparative research now accentuates four paradoxes of industrial growth: the composite character of technological practice; the centrality of accidental conjunctures of political and military events; the simultaneous success of institutions that appear to follow incompatible strategies; and delayed positive effects from arrangements that deterred industrialization at earlier historical moments.

The inaugural case of industrialization, usually termed an 'industrial revolution,' occurred in Britain from approximately 1760 to 1860. Britain's coal output rose more than 10-fold in this period, due in part to the installation of steam-driven water pumps in mines. Pig iron production increased more than 16-fold. By far the largest portion of increase in the country's overall productivity came from the textile sector after small innovators such as James Hargreaves and Edmund Cartwright facilitated the mechanization of cotton spinning and weaving in factories. Yet the record is studded with puzzles. In textiles, virtually none of the individuals or firms making initial technological breakthroughs were rewarded via patents or exceptional returns (Clark, 2007). Compared to other times and places, manufacturers in general dedicated little of their profit to new equipment and expansion. Their conservatism is all the more remarkable in view of the uniquely secure demand from colonial markets that British entrepreneurs enjoyed, which lessened the risks of fresh capital investment. It was less the rate of economic growth than its sustained accrual that distinguished Britain from all predecessors. Few investigators maintain that overall global development would look much different today had Britain not led the advance (Horn et al., 2010).

Processes of industrialization across capitalist societies as well as across state-dominated societies with central planning, such as the former Soviet Union, have shared essential similarities. Initially industrialization is marked by massive transfers of labor out of agriculture and into factories that have concentrations of capital equipment. Increases in the productivity of the labor devoted to manufacture come to balance increases in demand for goods, however, and employment in the service sector increases more rapidly than in manufacture after initial industrialization. In the leading industrial societies of Europe, Asia, and North America, employment growth has increasingly concentrated in services, expert professions, and finance. Even in industrialized countries that are net exporters of manufactures, such as Germany, the absolute number of workers in manufacturing has usually declined since the 1970s. Indeed, some national economies, including that of the UK, have become net importers of manufactured goods and net exporters of design know-how and of financial services that put technology to work elsewhere.

Technology as a Composite

Apart from basic principles familiar since Galileo, the machinery of the first industrial revolution emerged independently of scientific discovery. Elites of the European Enlightenment diffused an ethos of invention, but actual technical improvements in Britain drew almost entirely on artisanal experience and its incremental consolidation via the shopfloor and vernacular publications. Technology has continued to operate as a hybrid system of tools combined with responsive know-how, both woven together by broader practices of education and family socialization. Well into the era of mechanized cloth production, one manufacturer in Europe might succeed with a device that proved unworkable for neighbors. Throughout the nineteenth century, centralized facilities with steam power demanded greater hand inputs, including custom equipment and reengineered local installation. Mechanization of simpler processes, such as spinning, incongruously enlarged demand for manual household enterprise that remained resistant to mechanization, such as handweaving.

Britain's jump into steam-powered textile mills appears quixotic in comparative perspective. Already in the seventeenth century, the Chinese had diffused a piston/cylinder steam engine as serviceable as James Watt's. Until Hargreave's spinning jenny and Kay's flying shuttle, Chinese textile equipment developed as dynamically and efficiently as Britain's. Traditions of statecraft led Chinese scholars to disseminate practical manuals on improved farming and technological best practice. The new scholarly consensus is that core regions of China in the late eighteenth century enjoyed mobile and disciplined labor, secure property rights in dense markets, flexible financial instruments, high agrarian productivity, and elastic household formation equal to that of pioneering northwestern Europe. The wider distribution of literacy and of self-motivated

tinkering in Britain probably offers a distinguishing key to its divergent take-off.

Yet explanations of industrialization that highlight the supply of the inputs of labor and technology on one side, as well as wide demand for new consumer goods on the other side, can prove inattentive to what lies between: the social institutions of industrial work itself. What led entrepreneurs to centralize work in mechanized factories? Unlike merchants who gave work out to domestic workers, factory employers could directly accelerate throughput to meet shorter delivery deadlines, increase the obligatory hours of work, regulate and reorganize work procedures to ensure quality or flexibility, and prevent competitors from gaining knowledge of novel improvements in products or technology enclosed behind factory walls. Early industrialization was driven as much by employers' need to steer labor at the point of production as by their need to save on labor as a market input. On the eve of industrialization, the commercial blossoming of domestic putting-out industries across many regions of Europe supported demographic expansion by facilitating early marriage and rapid family formation among younger adults. With galloping rises in population, employers lacked compelling incentives to invest in centralized machine production purely for the sake of economizing on labor.

Indeed, manufacturers and engineers in Britain's initial lead sectors of industrialization rarely justified technological investment by the savings in labor effort. They claimed that such classic inventions as the Crompton spinning mule and the use of chlorine in bleaching works increased turnover and thereby reduced the outlays on supplies and land per unit produced. The management goal, as a textile mill proprietor put it, was to "produce the goods on time." Not unexpectedly, therefore, variation in regional economic variables – such as labor supplies and the level of commercialization of household crafts – correlated very inexactly with variation in the transition to factory systems. Many regions at the forefront of dynamic networks of commerce and putting-out systems by 1750, such as East Anglia, were bypassed in the early Industrial Revolution. These areas show that industrialization did not represent the logical culmination of trade, capital accumulation, separation of labor from independent means of subsistence, and formation of a nonagricultural work force. Entrepreneurial interest in coordinating the production process played a role of its own. When more complex machinery demanded repertoires of skills beyond those of a household shop, factories emerged as centralized knowledge storehouses to communicate informal technical detail to employees while excluding outsiders (Mokyr, 2002).

In today's science-driven chemical, electronic, and biotechnology industries, the composite character of technology in use still guides export economies. Germany, overall the global leader in merchandise exports during the first decade of the current century, maintained its historically intensive apprenticeship systems in middling companies making specialized instruments and capital goods. Even when China recently became the global export leader, shipments of such German machinery continued to widen Germany's trade surplus with China. Social scientists now recognize that China's rapid industrialization stemmed from the late Maoist era of mass education and of state-sponsored rural industry. These earlier undertakings crash-trained a pool of 'mindful hands': adaptable workers who had 'learned by doing' (Bramall, 2007).

Conjunctural Coincidences

Comparative research on industrialization has isolated the massive effects of political and geographical accidents in relation to endogenous commercial evolution. Britain's island focus on naval power combined with the necessity of import substitution during the Napoleonic disruption of Continental markets was crucial for sustaining the country's gradual technological improvements. The vast expansion of material inputs, especially cotton, from North America helped Britain overcome the ecological constraints of its land size. In China, by contrast, promising macroeconomic conditions in the eighteenth century, such as high agrarian productivity and capital stocks, were dissipated due to the Qing empire's experience of overstretch and its lack of incentives to steer ambitious overseas commerce. China's coal reserves in the north were accidentally distant from its most dynamic markets and from southern centers cultivating iron production. More astute imperial settlement and transport policies might have united Chinese mineral deposits with iron making. The more closely social scientists examine proximate conditions, the more they enlarge alternative possibilities for Britain to have fallen into an economic cul-de-sac or for East Asia to have escaped from one two centuries ago (Pomeranz, 2000).

In our day, China's breathtakingly swift rise to manufacturing preeminence also relied on international contingencies. US imports from China increased 10-fold from 1991 until the Great Recession of 2008. Industry-specific studies suggest that Chinese imports account for as much as 40% of a decline in US manufacturing employment in this period. The diversion of US funds to military campaigns after 9/11 and accompanying US divestment from domestic research, education, and infrastructure facilitated a shift of initiative to China.

Conflicts between Diverse Institutional Logics

Karl Marx, perhaps the most famed theorist of industrialization, inferred a single path of factory organization based on cumulative investments in machinery. He expected that advanced firms would mass produce standard wares and simplify work tasks to reduce continually labor inputs per product manufactured. The assembly line made famous by the automaker Henry Ford, as well as the state-run firms of socialist regimes, attempted to embody a similarly rigid logic.

Yet examination of historical diversity has undermined the assumption that industrial development depends on integrating mass production in ever larger facilities. Networks of small enterprises in districts such as Solingen or Neuchâtel in the nineteenth century, and Bologna or Silicon Valley in the 1980s, prospered by creating complementary resources for short runs of custom products, a strategy termed 'flexible specialization.' In place of vertically integrated, hierarchical firms, smaller producers clustered to specialize in discrete stages of shifting production processes or they drew jointly

upon specialized pools of labor and particular supplies for innovative products. The hundreds of modest cutlery firms in the nineteenth-century Solingen, for example, obtained crude blades from mechanical forges and on purchasers' demand custom finished them with steam-powered grindstones into literally thousands of types of cutlery. By 1914, Solingen had outstripped foreign competitors by focusing as much on variety in design as on cost minimization. This alternate course of development appears attractive for manufacturers in mature economies today. Usually with the aid of computers, producers who bear high wage costs may try to compete in the world market by rapidly adapting technology to add value to products through revamped or specialized designs (Sabel and Zeitlin, 1997).

Whether flexible specialization works effectively as an industrial strategy under capitalism depends on political events and on the infrastructure of public institutions, not purely on economic conditions. Firms that remain modest in size seem more adept at adjusting technology to changes in the needs of market niches. Much of the employment growth and dynamism in mature economies is consequently focused in small-scale enterprises. Yet these smaller organizations typically lack sizable pools of funding for research into fundamentally new kinds of technology, and they risk stagnation if they vary products only minutely. For enterprises to share the costs of basic advances in technology, flexible specialization typically calls for a complement of cooperative research centers, universities with ties to business, or other institutes for promoting local technological improvement. The creation of these supporting institutions can grow out of municipal campaigns for local entrepreneurs, out of guild programs for promotion of the trade, and out of scientific provisions by central states. Flexible specialization also requires safeguards against sweatshop wage reductions, so workers have reason to invest in their own skills and to contribute their ingenuity to adapting equipment in changing markets. Labor relations supporting this path of industrialization may emerge unintentionally out of struggle between labor and capital at key political junctures. In the cutlery trade of Solingen, for instance, employers were 'forced into their luck' by powerful unions in the 1870s that successfully pushed for codified procedures to divide productivity gains between workers and employers.

An accelerating need for firms to deploy fundamentally new scientific discoveries has further undermined the accustomed models of factory work. Frederick Winslow Taylor and other 'scientific' managers at the beginning of the twentieth century believed that the division of work into fixed elementary tasks should take primacy. By contrast, corporations embedded in fast-changing scientific communities today find their primary challenge is to respond to novel trends and to speed collective learning of sophisticated skills. The tasks in the nineteenth-century mass production could be carried out separately and their outputs added up to assemble a standard product. Contemporary managers focused on design innovation often find that work processes are more interactive than additive and that the challenge is to encourage mutual adjustments among design and production groups as they address qualitatively new assignments. Firms dependent on continual transformations of knowledge have therefore preferred collegial and participatory work styles over bureaucratic and hierarchical forms.

Intensified global exchange has conversely reinvigorated opportunities for mass production in low-wage regions, because it widened economies of scale for making durable goods parts as well. The consequent outsourcing of simpler manufacturing and services to developing economies has had surprising repercussions. In more mature economies, all but the smallest of companies are negotiating diverse kinds of 'vertical disintegration' (Herrigel, 2010) Two principal causes prevail. The model of high-wage, knowledge-based economies establishing blueprints that are manufactured at arms-length under contract in low-wage offshore regions generally proves infeasible with time. The final products from electronic gadgets to automobiles increasingly take shape as integral wholes. Production groups inform designers of emerging technical blockages and anticipate technological opportunities so that designers incorporate this practical information into a resculpted architecture for products as wholes. With speedups in global competition to innovate, design personnel have to cooperate on the ground with producers and tend to shift location to the offshore sites for more forward-looking integration. Second, products that are engineered for marketing to consumers in developing countries require local reconfiguring expertise. As these two causes ripple together, the developing countries may become better situated than the core countries to cross-fertilize interrelated technologies. The stark divide between developing and developed economies seems to be eroding in favor of islands of advance in each kind of economy. Whether the net effect is enhancement of a developed country's relative ability to add high value to products appears uncertain. Vertical disintegration makes it feasible for new entrants to penetrate quickly the higher-priced segments of an industry (Contractor et al., 2011). Automobile companies that spin off the manufacturing of whole modules become vulnerable to the formidable offshore suppliers of their own creation.

Capitalist firms today encounter an unprecedented flux between competing strategies of manufacture. No single scheme for organizing enterprise can become normative within nations or sectors. Once popular theories about coherent, path-dependent continuity or staged evolution in industrial practice due to national regulatory environments or cultural habit are becoming less credible to social scientists. On the one hand, intensifying global competition may lead even the largest of firms to concentrate narrowly by Adam Smith's model of the division of labor. Companies may strip themselves down to create a single core 'product,' sometimes amounting to only an intermediary assembly or final coordination between components. Digital communication may assist such a contractual parceling of tasks by codifying computerized templates by which product ingredients will be combined (Arrighi, 2007). Computer networks incessantly condense previously untradeable intellectual services into commodity objects. Accordingly, in the 1990s the Boeing corporation undertook to slash expenses by subcontracting not only the manufacture but also the engineering of critical parts for its heralded 787 jetliner. The corporation intended to manage only the overall concept. In the end, however, Boeing with prohibitive delays was left holding dozens of incomplete planes that required expensive fixing prior to sale a decade later. If products are constantly morphing as

integral wholes, firms must safeguard quality by supervising all phases of production and must broaden their palette of activities to cross-fertilize possible contributors to innovation. It looks telling that Boeing recomposed its work in-house by purchasing former suppliers and by retraining the remaining contractors.

Such contradictions within global manufacturing today appear to have made the task of governments that would promote development less calculable. The administrations of Japan, Korea, and Taiwan once were prominent among those that guided local industry into more rewarding stations in the global hierarchy of labor. Such states had the capacity to sponsor industrial leapfrogging in the late twentieth century due to their autonomy from narrowly self-interested business groups and due to their cadres of career servants recruited and remunerated by meritocratic standards. In Japan, the state's amalgam of bureaucratic autonomy and informal social connections to private business managers was key to its stewardship of the domestic computer industry. The central state not only offered tax incentives for domestic computer engineering while discouraging imports, but it also acquired in-depth knowledge of the industrial players to allocate patent information and funds for shrewdly chosen research projects (Evans, 1995). Across historical eras, however, comparative industrial advance cannot be linked to a particular kind of state intervention. Today Japan's ministries of trade and industry help foreign companies set up operations in Japan. Pathways toward technological superiority are increasingly unforeseeable by single governments because progress depends on transnational scientific collaboration.

Positive Legacies of Layered Histories

In the ambiguous and problematic terrains of current economic environments, corporations define their own challenges, and correlatively they experiment outside the institutional architectures of their own nations. Whereas equipment builders in the US traditionally relied on market mechanisms alone to find parts suppliers, today they try out collaboration with state governments and local colleges to assist suppliers at large in labor training and cost reduction (Herrigel, 2010). In these settings of self-reflexive problem solving, the cultural repertoires inherited from largely superceded, residual institutions sometimes function as adaptable resources for coping with uncertainty. Economic historians once assessed as an obstacle to capitalist take-off the historic tendency of extended rural households in China to remain attached to parcels of land rather than to separate completely from the means of subsistence. This Chinese family pattern appeared most particularly to discourage single women from migrating to factories and cities as women did in Western economies. Nowadays researchers suspect that China's lagged maintenance of rural households functions as a cultural reservoir. It sustained traditions of creative and disciplined task rotation as well as of independent household management that have infused factory labor in turn. Large Chinese factories by some accounts operate with thinner managerial staffing than is common in their Western analogs (Arrighi, 2007). Likewise, if only in comparative perspective, investigators suggest that the Maoist legacy of collective provisioning buffers many Chinese industrial workers from the reduction in education and in human welfare that initially accompanied the formation of factory workforces in Europe and North America.

Bedrock Universals

Concentrating workers in factories has everywhere triggered the elaboration of novel ideologies to legitimate the employers' exercise of authority (Bendix, 1974). The mobilization of resources for industrialization also entails new definitions of responsibilities and rights over the factors of production, including guidelines for financing or profit allocation and for protecting workers' health in some measure. State ownership of enterprise alters merely the details of these regulatory problems. In all settings industrialization has called forth new statutory frameworks for governing the flow of labor across nations, cities, and firms. Industrialization leads eventually to procedures for crediting intellectual discovery and for singling creators out for reward, whether in the form of patents and copyrights or through other kinds of recognition. Recognizing acute variations in industrial organization should by no means obscure the overarching epochal transformations.

Bibliography

Arrighi, G., 2007. Adam Smith in Beijing: Lineages of the Twenty-First Century. Verso, London.

Bendix, R., 1974. Work and Authority in Industry: Ideologies of Management in the Course of Industrialization, first ed. University of California Press, Berkeley, CA.

Bramall, C., 2007. The Industrialization of Rural China. Oxford University Press, New York.

Clark, G., 2007. A Farewell to Alms: A Brief Economic History of the World. Princeton University Press, Princeton.

Contractor, F., Kumar, V., Kundu, S., Pedersen, T. (Eds.), 2011. Global Outsourcing and Offshoring. Cambridge University Press, Cambridge.

Evans, P., 1995. Embedded Autonomy: States and Industrial Transformation. Princeton University Press, Princeton, NJ.

Herrigel, G., 2010. Manufacturing Possibilities: Creative Action and Industrial Recomposition in the United States, Germany, and Japan. Oxford University Press, New York.

Horn, J., Rosenband, L.N., Smith, M.R. (Eds.), 2010. Reconceptionalizing the Industrial Revolution. M.I.T. Press, Cambridge, MA.

Mokyr, J., 2002. The Gifts of Athena: Historical Origins of the Knowledge Economy. Princeton University Press, Princeton.

Pomeranz, K., 2000. The Great Divergence: Europe, China, and the Making the Modern World Economy. Princeton University Press, Princeton.

Sabel, C.F., Zeitlin, J. (Eds.), 1997. World of Possibilities: Flexibility and Mass Production in Western Industrialization. Cambridge University Press, Cambridge, UK.

Industrialization, Typologies and History of

PK O'Brien, University of London, London, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7360–7367, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Industrialization includes a rapid transformation in the significance of manufacturing in relation to all other forms of production and work undertaken within national or regional economies. The process is accompanied by technological and organizational change which leads to higher levels of productivity, rising living standards, population growth, urbanization, cultural changes, and shifts in the balance of power among nations. When, how, and why particular societies passed through the process continues to preoccupy social scientists, who for more than two centuries, have elaborated in historical depth and analytical sophistication on 'connections' between techniques of production, transport facilities, financial intermediation, state policies, foreign trade, urbanization, institutions, legal systems, and cultures on the one hand and the observed pace and pattern in different regions of the world economy on the other. Alas these connections cannot be specified, investigated, and weighted in ways that might give rise to a general theory of industrialization for Europe or other continents. There is only the complexity of history.

The Process of Industrialization

Industry is distinct from agriculture, forestry, fishing, mining, quarrying, transportation, finance, trade, and the myriad services that make up the separate but classifiable tasks performed by the workforces of modern economies. Throughout history, men and women manufactured commodities for use or for trade and sale. No society has operated without producing some range and levels of industrial output. Industrialization refers, however, to economic change that is recent and different in scale and scope from the manufacture of artifacts. As a socioeconomic process, industrialization includes a rapid transformation in the significance of manufacturing activity in relation to all other forms of production and work undertaken within national (or local) economies.

Economists, historians, and sociologists have measured and compared industrialization in statistical form as it appeared in national accounts and evolved historically for a large number of countries. Their data shows that as industrialization proceeds, the share of national workforces employed in agriculture and other primary forms of production, as well as the national outputs emanating from them, decline; and the shares of employment and output that are classified as industrial, increase. Output and employment emanating from the third macro-sector of national production, services, goes up or down in relative terms. The term 'services' includes all forms of noncommodity output that are sold (and/or supplied) either to consumers (for instance, health care) or utilized as 'inputs' (e.g., distribution, legal advice, accountancy, etc.) in order to sustain both manufacturing and primary forms of production. Clearly as and when industry grows more rapidly than other forms of commodity output, the allocation of services shifts toward manufacturing and shifts away from farming, fishing, forestry, and mining (Kuznets, 1965).

For sustained development there is no substitute for industrialization, which can be measured as the reallocation of a nation's stock of capital (embodied in the form of buildings, machines, equipment, tools, infrastructure, communications, and distribution networks) away from primary production, toward industrial production. But macro-data is more likely to be available in respect of the foreign trade of nations, which allows observers to measure the progress of industrialization over the long run through the study of predictable shifts in the composition of a country's exports and imports. Sales of domestically produced manufactured exports normally grow in significance; and purchases of foreign manufactures diminishes as a share of total imports (Jones, 1988).

Economic data has been classified in heuristic ways and disaggregated into numerous activities and functions in order to expose the extent, pattern, and pace of industrialization over time, across regions, and among countries. These essentially taxonomic exercises help to define and make concrete a process that has proceeded on a global scale for centuries. They expose national variations from more general or regional patterns and contribute to the understanding of major economic variables that historically have fostered or restrained industrialization in various parts of the world (Kuznets, 1965).

Industrialization has been an important process for the welfare of humankind because the reallocation of labor, capital, and other national resources toward industry has usually been accompanied by technological and organizational change, which led to higher levels of output per person hour; rising living standards; population growth; urbanization; cultural changes; and shifts in the balance of power among nations. Thus, industrialization can also be defined in social, cultural, and political terms, and sociology depicts the history of industrial societies in terms of a set of interconnected characteristics, hegemonic values, and legal systems, represented as functional for the development of modern industry.

How, when, and why particular societies moved from preindustrial to industrial norms, motivations, status, family systems, and modern institutions that characterize industrial society remains difficult to explain. Sociological theories elaborate upon the type of changes required in individual behavior and social institutions for the modern industry to succeed. They contrast 'traditionalistic patterns of action' that are defined as

ascriptive, multidimensional, communitarian, familial, and authoritarian with the types of individualistic, achievement-orientated, mobile, entrepreneurial attitudes and behavior that became dominant in cultures as industrialization took hold. This approach depends upon the vocabularies and concepts drawn from sociology, psychology, and cultural anthropology, and on analyses inspired by Max Weber that continue to be preoccupied with value systems (derived ultimately from religions) that have 'motivated' the 'drive to industrialize' in different national and cultural settings (Parsons, 1971).

Unfortunately, historical records are not clear on whether social changes precede or accompany industrialization. Until the late twentieth century, sociological approaches to industrialization were more concerned with its disruptive, dislocative, and potentially negative consequences for families, communities, villages, and regions, than with its nature, origins, and positive effects on living standards. Read as a social process, industrialization often leads to differentiation flowing from the division of labor, class formation, and uneven regional development. As industry diffuses from country to country, it becomes associated with diminishing returns, deindustrialization, unemployment, and the economic decline of some nations. Pessimistic writings about industrialization are often derived from Marx. Fortunately modern sociological understanding of industrialization combines several traditions and schools of theory with historical enquiry and a more process-centered global perspective, which has exposed how complex, multifaceted, and variable the process of industrialization has become since Marx, Comte, Durkheim, and other canonical social scientists wrote their critiques.

The British Industrial Revolution and the Diffusion of Modern Industry

Modern industry, concentrated in particular regions and towns, housed in factories, organized in corporations and firms, and using machinery and inanimate forms of energy evolved gradually. That evolution has preoccupied generations of economic and social historians who have explained why industry became dominant in some national economies before others and analyzed major forces that carried industrial sectors forward from one stage to another. Their writings are heavily concentrated upon the epoch which opens with the beginnings of the British Industrial Revolution in the mid-eighteenth century and closes with World War I. Latterly, historical literature, dealing with proto-industrialization in Europe and Asia before the First Industrial Revolution and with the delayed industrialization characteristic of Russia, Eastern Europe, East Asia, and Latin America, has emerged. The bibliography of literature on the histories of firms, industries, manufacturing sectors located in villages, towns, regions, countries, and continents, includes entire libraries of monographs, essays, and articles. The volume of research covering millennia of time and every conceivable set of spatial arrangements defies summary. Thus the task for historians (and social scientists) and the aim of this article is to review the theories, patterns, and typologies that have been imposed upon this mass of otherwise incomprehensible local and historical detail (O'Brien, 1998).

Shortly after 1945, when the causes of the wealth of nations emerged to dominate the agenda for the social sciences, historians began to formulate models to make sense of Europe's precocious industrial development. At that time the most accessible heuristic device was to refer 'continental' industrialization to a paradigm case, the British Industrial Revolution, as it supposedly proceeded from the 1750s to the 1840s. On the assumption that Britain's lead could be ascribed to the early diffusion of superior industrial technology, it followed that the economic history of Europe could be understood as the spread of techniques of production and new forms of industrial and commercial organization from the first industrial nation to more backward economies on the mainland. In this perspective, British methods constituted not merely initial practice but best practice with universal application. Diffusion then became a major concept and the central problem for European economic history, at least for the nineteenth century, was to explain Britain's early start and the rate at which other European economies adopted British ways (Landes, 1969).

Yet European statesmen, businessmen, and other contemporary observers rarely perceived of their national goals in terms of a race to catch up with Britain. Indeed many expressed doubts about the 'British way' and historians of European economies have not allowed their accounts of the industrialization of economies on the mainland to be dominated by a model inspired largely by Anglo-American perceptions and experience. Diffusion is now perceived as a complex process of creative adaptation rather than mere emulation. The transfer of technology only succeeds when a matrix of related capacities into which machinery can fit has been built up and is receptive to change (Berg and Bruland, 1998). Once in place, European firms and technologies spawned their own problems, solutions, and national trajectories for future innovation (Mokyr, 1990).

Furthermore, late twentieth-century historical research has restored perceptions of the British Industrial Revolution as a drawn-out process in which the upswings in national income and industrial production of the eighteenth century are no longer represented as 'discontinuities' or 'take offs.' Britain's technological lead has emerged as less novel, more slowly diffused, and eventually successful because it rested upon the evolution of a commercialized economy throughout the British Empire, over several generations before the American and French Revolutions (Crafts, 1985). Comparative historical research which now brings Asian as well as European economies into the frame, has also elaborated upon the foundations of Britain's early start, which were specific to that well-endowed offshore island. For example, the relatively high level of energy consumed per capita and per worker, obviously reflected abundant and accessible supplies of cheap coal (Wrigley, 1988). Located at the hub of a rapidly growing Atlantic economy, when waterborne transport was the cheapest way to conduct commerce, a small island was more likely to reap larger gains from oceanic trade than continental rivals. That locational advantage was, moreover, safeguarded and extended by extraordinarily high levels of public investment in sea power, which over the long run lowered transaction costs for British merchants, provided favored access to colonial and foreign markets, and weakened the economies of Britain's

Iberian, Dutch, and French rivals. Writing long after Trafalgar (where the Royal Navy finally won the centuries-old battle for the Atlantic) and sanctified by an ideology of liberalism, British historians tended to overlook the role of the Royal Navy in securing for British business 'disproportionate' shares of world trade, shipping, shipbuilding, and commercial services (O'Brien and Quinnault, 1993).

Thus the First Industrial Revolution can be represented as something of a special case rather than a paradigm case in the economic history of Europe (Mokyr, 1985).

Regions, Towns, and Proto-Industrialization

Industrialization was a regional process. Modern industry certainly did not spread randomly over national maps but 'clustered' within defined geographical boundaries. Insights are gained by observing the factors and circumstances which prompted industry to locate, grow, and decay in some regions rather than others; and in some places before others (Pollard, 1981). Clearly the geographical matrix for the analysis of economic change must be specified in terms of what the historian is trying to explain; and the rise and decline of regions forms an important component of national economic histories. Where and when markets for commodities and factors of production are bounded in a regional environment rather than in national or multinational spaces, economies constitute the obvious units for study.

Research into European regions of proto-industrialization over three centuries before 1800 promises to create the evidence required for a typology which might help historians explain the long-term evolution of major branches of manufacturing (particularly textiles, but also leather goods, finished metal wares, clocks and watches, and furniture) into mechanized factory industries over the nineteenth and twentieth centuries (Musgrave, 1999). However, no single model has emerged to account for the complex range of cases that historians have uncovered of tendencies for industries to concentrate geographically, to specialize, and to engage in long-distance external and internal trade. Histories of proto-industrialization usually elaborate upon a set of initial conditions which contributed to the availability and willingness of local workers (males, females, and children) to turn from dependence upon agriculture to become involved (either part- or full-time) in manufacturing. These include infertile land, inheritance systems, the sizes of farms, communal rights to land, the cost of alternative supplies of food, the productivity of local agricultures, regional specialization, and competition in the agricultural sector. Populations living close to subsistence levels on relatively small holdings, were affected by the high densities of labor to cultivable land, and they were among the first to embrace the opportunities provided by industrial work to supplement family incomes; and the landless and seasonally underemployed laborers from richer and more commercialized agricultural areas also seized these opportunities.(Kriedte et al., 1981).

But mercantile organization and capital were also required and they usually emanated from towns. Towns housed the majority of educated and innovative people. Productive knowledge was more readily diffused across economies where populations became spatially concentrated into the 'networks' or 'hierarchies' of cities and towns. The development of towns improved the efficiency of labor markets, promoted the commercialization of agriculture, and replaced barter with monetary transactions. Europe's maritime cities established imperial outposts overseas, which fostered international trade and specialization (de Vries, 1984).

By the eighteenth century, the map of industrial Europe could be colored into fairly well-defined regions of proto-industrialization, specializing in the production of textiles, clothing, boots and shoes, clocks and watches, tools, weapons, furniture, household utensils, and pottery. Several of these urbanized regions had survived and adapted to competition over long periods of time. Others had recently emerged as loci of concentrated production. All these industries faced limits to improvements in labor productivity through further refinements to the division of labor or from organizational improvements to the 'putting-out' system.

During the nineteenth century as mechanical innovations came onstream and markets widened through innovations in transport, the long-term development of some regions endowed them with several advantages for transitions to factory systems of production. For example, their sales outlets had been long established; they contained supplies of labor experienced in manufacturing for distance markets; capital and managerial skills were available from urban merchants; and all kinds of other external economies flowed from the prior concentrations of industry in the hinterlands of towns. Local agricultures had long demonstrated their capacity to respond to the demands of industrial populations for food and raw materials (Braudel, 1981–1984).

Yet despite and sometimes because of these favorable historical traditions, many regions of advanced proto-industrialization did not move rapidly forward to embrace new techniques of production and modes of organization. In some cases traditions proved to be more of a hindrance than help, and several regions lacked some essential ingredient for mechanization, such as supplies of cheap energy or engineering skills.

Research into proto-industrialization constitutes a program with explanatory potential, but it only covers certain sectors of modern industry. Producers' goods (iron, steel, ships, transport equipment, bricks, glass, and minerals) as well as some important consumer goods, such as flour, beer, and soap, were not manufactured under conditions that even approximate to a putting-out system. Furthermore, if the research program is to fulfill its promise, we still need to know rather more about the varieties and adaptabilities of organizational forms that are encompassed under the heading of proto-industrialization and also far more about the causes of success and failure, both before, and more importantly, during the transition to the factory system (Cipolla, 1973–1974).

Patterns and Typologies of European Industrialization

Largely because he conceived of European industrialization as a process of diversity rather than uniformity, Gerschenkron produced a heuristic model, which focused research upon variations in the manner in which European economies passed

through the decisive stage of accelerated industrial growth or the 'great spurt.' In his view, Britain forged ahead because the conditions for 'spontaneous industrialization' developed early – and the problem for 'relatively backward economies' on the European mainland was to find effective substitutes for missing (British) prerequisites. Delays exhibited by continental economies in their endeavors to catch up with Britain are imputed to a range of 'prerequisites' for successful industrialization that were available to those who transformed the British economy, but were in shorter supply elsewhere in Europe; and the time that other 'continental' countries took to develop substitutes for their cultural, institutional, and technological deficiencies also added to the delay (Gerschenkron, 1962).

Gerschenkron (like Rostow before him) neglected the concept that natural endowments and the role of the state were factors that affected the British Industrial Revolution, and he concentrated upon other elements of industrialization, which he hypothesized would vary systematically with the timing and degree of backwardness exhibited by 'follower' countries, as they passed through clear discontinuity in the growth rate of industrial production (Rostow, 1965).

Gerschenkron predicted that when historians examined the buildup of modern industry in Europe, country by country, they would observe that the more backward or older the economy was when it began that process, the more rapid would be the rate of acceleration in the growth of industrial production, and the more pronounced would be its reliance on capital as distinct from consumer goods. He anticipated that among 'latecomers,' new industrial enterprises would emerge as being technically sophisticated and capital intensive, larger in scale, and more prone to monopoly. Such enterprises would rely to a great degree upon external sources (banks and the state) for both investible funds and entrepreneurial initiatives. For their markets and supplies of cheap inputs they would depend far less than their British counterparts had upon rising productivity in domestic agricultures. Historical research (inspired by his hypothesis) has not merely widened the sample to include potentially deviant cases from the Austrian Empire, the Balkans, and Scandinavia, but has also deepened the knowledge of British, Dutch, French, Italian, and even Russian industrialization in ways that appear inimical to the survival of Gerschenkron's typology as a usable heuristic device to generalize about European, and Asian, industrialization (Sylla and Toniolo, 1991).

The theoretical and statistical problems involved in delineating phases of trend acceleration seem to be insurmountable and as historians turned away from imperfect and partial indices of industrial and national production to investigate the regional and urban proto-industrial foundations of modern industry, the drama of 'take-offs,' 'spurts,' and 'discontinuities' in macroeconomic history has faded from view.

Expectations that historical records might reveal distinctively national patterns of industrial organization cannot be tested because the data required to quantify and systematically compare the prevalence of larger-scale, more capital intensive, and monopolistic firms across countries cannot be constructed. Furthermore, typologies, which juxtapose continental deviations against the 'stylized' version of British industrialization, now look anachronistic. Late twentieth century research on the British industrial organization reveals variance in scale, capital intensity, and departures from competitive norms across industries throughout the First Industrial Revolution (Floud and McCloskey, 1994). Variations in plant size, competitiveness, and factor intensity are not systematically correlated to backwardness, but are related to legal conditions for incorporation, the development of capital markets, changing technology, and above all to product mix. Long-term changes in the composition of manufacturing output do not expose any systematic or predictable shifts from the production of consumer to capital goods or vice versa (Pollard, 1990).

Nevertheless, transitions from simple to complex processes of manufacturing activity are normally orderly. Thus the initial mechanization and standardization of several consumer-goods industries tended to be located within well-defined regions of specialization, utilizing the investible funds, skilled labor, and entrepreneurial talents available in established proto-industrial regions. Factories emerged on a small scale with relatively low capital–labor ratios to engage in the linear transformation of raw materials into familiar products sold on well-established markets. In many cases resources could be mobilized from within the traditional economy in order to replace imported manufactures or the products of handicraft industry. As manufacturing expanded, as the agricultural sector embarked on mechanization and farmed with chemicals, and as transport networks began to cover national terrains, demand and supply constraints on the expansion of domestic capital goods sectors gradually diminished. At that stage a 'second wave' of industrialization carried economies forward (Mathias and Davis, 1990–1998).

Deviations from a path-dependent process of diversification from simple to complex systems of manufacturing occurred, but these deviations are explicable in terms of country size, natural endowments, comparative advantages within an evolving world economy, direct foreign investment, and state policies. Typologies of European industrialization which pay insufficient attention to geography and foreign trade miss essential elements of the process. Europe developed over the nineteenth and twentieth centuries as a continent of large and small political units, with a skewed distribution of natural resources. As possibilities for foreign trade expanded, different patterns of industrial growth evolved from it, varying from country to country (Aldcroft and Sutcliffe, 1999). For example, at comparable levels of per capita income very few European economies achieved anything like Britain's comparative advantage in the export of manufactures. As late as 1913 the shares of 'manufactured commodities' to total exports varied from under 10% for Greece, Denmark, Norway, and Portugal to over 75% for Britain, Belgium, Switzerland, and Germany, with the rest of Europe somewhere in between. Systematic comparisons of trade data across countries and studies of long-term trends in the composition of net imports (exports minus imports) for major manufactured commodities offers historians prospects for the statistical delineation and comprehension of the patterns of industrialization undertaken by different European economies before and after 1914 (Bairoch, 1976).

General propositions about the role of banks have also proved difficult to force into the format of generalizations. Modern research on financial intermediation has fruitfully shifted attention toward the problem of analyzing responses by banks to industrialization as it proceeded in particular national

and regional contexts. Cameron summarized the results of international comparisons by admitting he lacked criteria to measure their success or efficiency. Although he was prepared to laud the contribution of banks in Scotland and Sweden and to deplore their sparseness in France – an economy which apparently managed to achieve respectable levels of per capita income and industrial output without much help from its banks (Cameron, 1967). For Germany, (once the paradigm example for the positive role of banks) there is now a debated hypothesis that the overpraised Kreditbanken misallocated capital and slowed overall growth (Fremdling and Tilly, 1976). In the Czech Crownlands (Bohemia, Moravia, and Upper Silesia) risk-averse banks did little except supply a modicum of working capital. After 1873 Austrian banks "lost most of their interest in industrial enterprises for years to come." For Hungary it was not the banks but the "unencumbered access to the Austrian capital market" that appears crucial to the beginning of industrialization (Good, 1984). Norwegian commercial banks showed little initiative and failed to play any leading role in the process of economic growth. In general it is difficult to find cases where banks clearly played the progressive and entrepreneurial role in national industrialization (Teichova et al., 1997). In analyzing connections between the state and industrialization Gerschenkron suggested that before 1914 several European governments supplied substitutes for deficiencies in private enterprise, private capital markets, and household demand. Although his observation seems plausible and is based upon historical examples (drawn largely from Russia and Eastern Europe), it has not been possible to subject it to statistical testing because the scale and intensity of governmental intrusion into an otherwise market-led spontaneous process cannot be measured (Gerschenkron, 1962). One macro-economic indicator – the ratio of public consumption to gross national expenditure – shows no tendency to vary in any systematic way across countries (O'Brien, 1994).

Instead economic historians have moved on to classify and to attempt to evaluate potential effects on private industrial investment (domestic and foreign) flowing from the different fiscal, tariff, monetary, and exchange rate policies pursued by European governments over the nineteenth and twentieth centuries; and for this topic there is a bibliography of sophisticated studies which apply macro-economic theory to history (cited in O'Brien, 1994).

Nevertheless to confine investigations into the direct and indirect role of governments to their macro-economic policies would be to neglect the core functions performed by states to promote industrialization, namely the creation and reform of laws and institutions which underpinned the operation and spread of markets for commodities and the factors of production. Such markets operated within a framework of rules, which regulated cooperation and competition among individuals and enterprises engaged in economic activities. One of the central traditions of 'old' economic history was to study the laws, regulations, property rights, and enforcement procedures which molded and constrained economic behavior over those long spans of history before the era of the French Revolution, when 'inefficient' sets of institutions are perceived to have held European economies inside their own concepts of the boundaries of production possibility. That program (associated with the German historical school) has been revived with the aid of theories and a vocabulary borrowed from modern economics (Hoselitz, 1961). Neo-liberal institutional economics has been applied to the history of early modern Europe and suggests that frameworks for the operation of commodity and factor markets emerged in some societies (Holland and England) long before others (Spain and Russia) (North and Thomas, 1973).

Clearly it remains important to understand how governments constrained the progress of industrialization by defining and enforcing inefficient systems of rules and property rights. The general problem resides in comparing legal frameworks for economic competition and cooperation across countries, in analyzing alterations to laws and property rights over time, and in systematically linking specified variations in law to economic growth. Although the research program is concentrated on the early modern period when states established (or failed to establish) conditions for subsequent specialization and the widening of markets in the eighteenth and nineteenth centuries, its relevance to the later era of the 'Second Industrial Revolution' (which witnessed the rise of new industries, such as chemicals, electricity, and motorized transportation, and the concomitant appearance of 'big government' to regulate the operations of corporations, unions, and financial intermediaries) is also clear (Mathias and Pollard, 1989). Historians have work to do and are still some way from theorizing or generalizing about connections between states and industrialization (Tilly, 1990).

Conclusions

A long line of distinguished social scientists and historians (including St. Simon, Schmoller, Weber, Bloch, and Braudel) have offered students of European economic history, theories, concepts, approaches, and typologies designed to help them organize, clarify, and generalize from the vast accumulation of information now in print about the history of European industries. At present when the British Industrial Revolution has reemerged as a special case, the potential for histories of European industrialization constructed around the concept of technological diffusion has been degraded (Landes, 1969).

From philosophers of the Enlightenment, through Marx to the German historical school, scholars of European industrial development have found it difficult to organize their analyses without recourse to stages of growth. Rostow's stage theory certainly provided a memorable vocabulary but his schema is not considered to be a particularly illuminating way to write industrial history (Rostow, 1965). Conceptually, the whole notion of a sharp discontinuity no longer attracts historians of economic growth inclined to analyze that process with biological rather than aeronautical metaphors (like 'take-offs') and to explain accelerations (if, when, and where they might be measured) as the product of fortuitous political events, or basically as the final outcome of a protracted process of institutional change and the prior accumulation of physical and human capital (Braudel, 1981–1984).

Gerschenkron's typology inspired generations of scholars and survives as the best frame of reference for economic historians concerned with placing their research within wider and potentially deeper perspectives that emanate from comparisons across Europe. But more often than not, the

heuristic value of typologies seems to reside in the organization of evidence that simply fails to fit the country or case into a pattern (Sylla and Toniolo, 1991). Meanwhile, and as research upon economies, regions, towns, industries, firms, and factors of production proliferates, no scholar with Gerschenkron's awesome learning and capacity for generalization has emerged to reconstruct an alternative typology which might serve to draw the long history of industrialization into a European frame of reference (Senghaas, 1982).

That is why the majority of economic historians continue to pursue research within the framework of nation-states and to offer their scholarship as a basis for comparative history. Indeed that program is exemplified by most textbooks in the field, which tend to be surveys of foreign scholarship arranged country by country, industry by industry, input by input. Readers of these analytically sophisticated books in economic history are implicitly asked to make intellectual leaps from national histories moving through long centuries of industrialization to comparative history unfolding across Europe. They will certainly find scholarly and detailed elaborations on 'connections' between techniques of production, transport facilities, types of financial intermediation, and state policies, foreign trade, urbanization, institutions, legal systems, and cultures, on the one hand; and the pace and pattern of industrialization, on the other. Alas, their connections can rarely be specified, investigated, and quantified in ways that might suggest how observed variations across regions and countries in the manifold factors promoting or restraining industrialization through time might be 'tested' for significance.

Meanwhile (and after two generations of postwar research on the economic histories of Africa, the Near East, and South East Asia) another meta-narrative concerned with the 'Great Divergence' (or why Europe industrialized ahead of economies located on other continents) is beginning to generate syntheses, generalizations, and exciting controversies for historians to pursue (Landes, 1998; Frank, 1998; Pomeranz, 2000).

See also: Economic History; Industrialization; Modernity: History of the Concept; Modernity; Modernization and Modernity in History; Modernization, Sociological Theories of.

Bibliography

Aldcroft, D., Sutcliffe, A. (Eds.), 1999. Europe in the International Economy 1500–2000. Elgar, Cheltenham, UK.

Bairoch, P., 1976. Commerce Exterieur et Developpement Economique de l'Europe an XIXe Siecle. Mouton, Paris.

Bensel, R.F., 2000. The Political Economy of American Industrialization, 1877–1900. Cambridge University Press, Cambridge, England.

Berg, M., Bruland, K. (Eds.), 1998. Technological Revolutions in Europe. Elgar, Cheltenham, UK.

Bramall, C., 2007. Industrialization of Rural China. Oxford University Press, Oxford.

Braudel, F., 1981–1984. Civilization and Capitalism 15th–18th Century. Harper & Row, New York.

Cameron, R., 1967. Banking in the Early Stages of Industrialization. Oxford University Press, New York.

Chibber, V., 2006. Locked in Place: State-Building and Capitalist Industrialization in India. Princeton University Press, Princeton, NJ.

Cipolla, C.H. (Ed.), 1973–1974. The Fontana Economic History of Europe, vols. 2, 3. Collins 4(I), 4(II), London.

Crafts, N.F.R., 1985. British Economic Growth during the Industrial Revolution. Oxford University Press, Oxford, UK.

Delfino, S., Gillespie, M., 2008. Technology, Innovation, and Southern Industrialization: From the Antebellum Era to the Computer Age. University of Missouri Press, Columbia.

Floud, R., Johnson, P., 2003. The Cambridge Economic History of Modern Britain. Cambridge University Press, Cambridge.

Floud, R., McCloskey, D. (Eds.), 1994. The Economic History of Britain since 1700, vol. 1. Cambridge University Press, Cambridge, UK, pp. 1700–1860.

Frank, G., 1998. ReOrient: Global Economy in the Asian Age. University of California Press, London.

Fremdling, R., Tilly, R., 1976. German banks and German growth and economic history. Journal of Economic History 36, 416–424.

Gallagher, K., Porzecanski, R., 2010. The Dragon in the Room: China and the Future of Latin American Industrialization. Stanford University Press, Stanford, Calif.

Gerschenkron, A., 1962. Economic Backwardness in Historical Perspective. Harvard University Press, Cambridge, MA.

Good, D., 1984. The Economic Rise of the Habsburg Empire. University of California Press, Berkeley, CA.

Hicks, J., 1969. A Theory of Economic History. Clarendon Press, Oxford, UK.

Horn, J., 2006. The Path Not Taken: French Industrialization in the Age of Revolution, 1750–1830. MIT Press, Cambridge, Mass.

Hoselitz, B., 1961. Theories of Economic Growth. Free Press, New York.

Jones, E., 1988. Growth Recurring. Economic Change in World History. Oxford University Press, Oxford, UK.

Kohli, A., 2004. State-Directed Development: Political Power and Industrialization in the Global Periphery. Cambridge University Press, Cambridge, UK.

Kriedte, P., et al., 1981. Industrialization before Industrialization. Cambridge University Press, Cambridge, UK.

Kuznets, S., 1965. Economic Growth and Structure. Norton, New York.

Landes, D.S., 1969. The Unbound Prometheus. Technological Change and Industrial Development in Western Europe. Cambridge University Press, Cambridge, UK.

Landes, D.S., 1998. The Wealth and Poverty of Nations. W. W. Norton, New York.

Mathias, P., Davis, J. (Eds.), 1990–1998. The Nature of Industrialization, vols. 1–6. Blackwells, Oxford, UK.

Mathias, P., Pollard, S. (Eds.), 1989. Cambridge Economic History of Europe, vol. 8. Cambridge University Press, Cambridge, UK.

Meyer, D.R., Center for American Places, 2003. The Roots of American Industrialization. Johns Hopkins University Press, Baltimore.

Mokyr, J. (Ed.), 1985. The Economics of the Industrial Revolution. Rowan & Allanheld, Totowa, NJ.

Mokyr, J., 1990. Lever of Riches: Technological Creativity and Economic Progress. Oxford University Press, New York.

Musgrave, P., 1999. The Early Modern European Economy. St. Martin's Press, New York.

North, D.C., Thomas, R.P., 1973. The Rise of the Western World. Cambridge University Press, Cambridge, UK.

O'Brien, P.K. (Ed.), 1994. The Industrial Revolutions in Europe, 2 vols. Oxford, UK: Blackwell.

O'Brien, P.K. (Ed.), 1998. Industrialization. Critical Perspectives on the World Economy, 4 vols. Routledge, London.

O'Brien, P.K., 2001. Urban Achievement in Early Modern Europe: Golden Ages in Antwerp. Cambridge University Press, Amsterdam and London. Cambridge.

O'Brien, P.K., Quinnault, R. (Eds.), 1993. The Industrial Revolution and British Society. Cambridge University Press, Cambridge, UK.

Parsons, T., 1971. The System of Modern Societies. Prentice-Hall, Englewood Cliffs, NJ.

Pollard, S., 1981. Peaceful Conquest. The Industrialization of Europe 1760–1970. Oxford University Press, Oxford, UK.

Pollard, S., 1990. Typology of Industrialization Processes in the Nineteenth Century. Harwood Academics, New York.

Pomeranz, K., 2000. The Great Divergence: China, Europe and the Making of the Modern Economy. Princeton University Press, Princeton, NJ.

Rostow, W.W., 1965. The Stages of Economic Growth. Cambridge University Press, Cambridge, UK.

Senghaas, Dieter, 1982. The European Experience: A Historical Critique of Development Theory. Leamington Spa, UK.

Stearns, P., 1993. The Industrial Revolution in World History. Westview Press, Boulder, CO.

Sylla, R., Toniolo, G. (Eds.), 1991. Patterns of European Industrialization. Routledge, London.

Teich, M., Porter, R. (Eds.), 1996. The Industrial Revolution in National Context. Cambridge University Press, Cambridge, UK.

Teichova, A., Van Hentenryk, G.K., Ziegler, D., 1997. Banking, Trade and Industry. Cambridge University Press, Cambridge, UK.
Tilly, C., 1990. Coercion, Capital and European States, AD 990–1990. Oxford, UK: Blackwell.
de Vries, J., 1984. European Urbanization 1500–1800. Methuen, London.
Wangwe, S.M., 2003. Exporting Africa: Technology, Trade, and Industrialization in Sub-Saharan Africa. Routledge, London.
Weber, M., 1961. General Economic History. Collier Books, New York.
Winch, D., O'Brien, P., British Academy, 2002. The Political Economy of British Historical Experience, 1688–1914. Published for The British Academy by Oxford University Press, Oxford.
Wrigley, A., 1988. Continuity, Chance and Change: The Character of the Industrial Revolution in England. Cambridge University Press, Cambridge, UK.
Yun, C.B., O'Brien, P., 2012. The Rise of Fiscal States: A Global History, 1500–1914. Cambridge University Press, New York.

Industrial–Organizational Psychology: Science and Practice

Wayne F Cascio, University of Colorado Denver, Denver, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by K. Kraiger, volume 11, pp. 7367–7371, © 2001, Elsevier Ltd.

Abstract

Industrial and organizational (I–O) psychology is the study of the behavior of men and women in work settings. I–O psychologists seek to understand and optimize the effectiveness, health, and well-being both of individuals and of organizations. This article considers the three major areas that comprise the broad domain of the field (personnel psychology, organizational psychology, and human engineering); an overview of the evolution of the field; the scientist–practitioner model and the role of research in guiding science and practice; and finally, factors that are driving demand for research and expertise in I–O psychology.

This article is divided into four main sections. The objective of the first section is to describe the domain of industrial and organizational (I–O) psychology, three major focal areas that comprise it, and to identify typical settings where I–O psychologists work. The second section presents an overview of the evolution of I–O: How the field developed to where it is today. The third section presents the scientist–practitioner model and the role of research in I–O psychological science and in practice. Finally, the fourth section identifies several key factors that are driving the growing demand for research and expertise in I–O psychology.

The Domain of I–O Psychology

Psychology is the scientific study of behavior. I–O psychologists focus on a key aspect behavior, namely, the behavior of men and women in work settings. This brief definition suggests three broad areas that comprise the domain of I–O psychology, the person (i.e., the worker), the work that is performed, and the context in which the work is performed (Zedeck, 2011).

With respect to context, the term 'work setting' may extend far beyond work, and with it, the domain of (I–O) psychology. While most of our lives are spent at work, what happens at work can have a huge impact on people and their families. Likewise, what happens while people are away from work, or in a family setting, can have a significant impact on one's behavior at work. Consider just one example of such 'spillover', the care of children. Research shows that men are taking more overall responsibility for the care of their children (providing one-on-one care, as well as managing childcare arrangements), according to themselves and their wives/partners. This has led to increased work-life conflict, as 59% of fathers in dual-earner couples report experiencing some or a lot of conflict today, up from 35% in 1977 (Aumann and Galinsky, 2009; Halpern and Murphy, 2005).

The influence of personality on the behavior of men and women in work settings is even more significant, because personality is often influenced by events that took place long before an individual began employment. Personality traits relevant to organizations are those psychological characteristics that predict consistent work-related thoughts, motivations, behavior, and other outcomes across situations and over time. Not surprisingly therefore, personality characteristics (e.g., emotional stability, extroversion, conscientiousness, agreeableness, openness) predict important outcomes, such as occupational attainment and advancement, overall managerial effectiveness, the performance of expatriates, learning and skill acquisition during training, counterproductive work behaviors, team cohesion and teamwork, and job and career satisfaction (Oswald and Hough, 2011).

I–O psychologists seek to understand and optimize the effectiveness, health, and well-being, both of individuals and of organizations (Rogelberg, 2007). With respect to individuals, I–O psychology seeks to understand the reasons for behavior, as well as intentions, attitudes, emotions, motives, values, and beliefs in work settings. It also seeks to understand, and to some extent to predict, how people can become effective, satisfied, engaged in their work, committed to it, fulfilled, and rewarded, and how these outcomes can be maintained. With respect to organizations, that is, collections of people working together in a division of labor to achieve a common purpose (Hitt et al., 2011), I–O psychology seeks to understand how they can be sustained and developed. I–O psychology therefore focuses on interactions between people and the settings in which they work in order to gain a deeper understanding of how their behavior can be influenced and enhanced to benefit the individual, the organization, and society as a whole. I–O psychology examines issues at the individual level, at the level of groups or teams, and also at the broader organizational level. It is therefore a multilevel and multiattribute study of behavior (Zedeck, 2011).

Traditionally, I–O psychology has been divided into three major areas: personnel psychology, organizational psychology, and human engineering. The following sections briefly consider each of these. Although we treat them separately, they often overlap considerably.

Personnel Psychology

Personnel psychology is a subfield within I–O psychology. It is an applied discipline that focuses on individual differences in behavior and job performance and on methods of measuring

and predicting such differences. Some of the major areas of interest to personnel psychologists include job analysis and job evaluation; recruitment, screening, and selection; training and development; and performance management.

Job analysis is the study of the work to be done and the personal characteristics necessary to do the work. Job evaluation is the process of rank-ordering jobs in terms of their relative worth to an organization. Recruitment is a set of activities designed to attract talent to an organization. It may be internal to a firm (e.g., across departments or functional areas) or external (e.g., job fairs, university recruiting, company and specialized Web sites, social media). In the hiring process, screening refers to a rough, initial examination (e.g., based on minimum qualifications) designed to identify candidates for further consideration. Selection may involve a variety of methods and instruments, such as interviews, written or performance tests, personality measures, biographical information, reference checks, or background checks. In selecting among applicants for jobs, personnel psychologists are concerned with developing valid and reliable assessment methods that are fair to members of all groups. Their objective is to select applicants who are predicted to perform well on the job. Training and development activities involve learning experiences, planned by the organization, and designed to improve performance at the individual team, or organizational levels. Methods may include, for example, formal classroom instruction, online training, on-the-job training, outdoor training, role-playing exercises, team building, or behavioral simulations. Finally, performance management is a continuous process of identifying, measuring, and developing the performance of individuals and teams, and of aligning performance with the strategic goals of an organization (Cascio and Aguinis, 2011).

Personnel psychology also represents the overlap between psychology and human resource management (HRM). HRM is concerned with the management of staffing, retention, development, adjustment, and change in order to achieve both individual and organizational objectives (Cascio, 2013). As a subfield of HRM, personnel psychology excludes, for example, such topics as labor and compensation law, organization theory, industrial medicine, collective bargaining, and employee benefits. Psychologists have already made substantial contributions to the field of HRM; in fact, most of the empirical knowledge available in such areas as motivation, leadership, staffing, and performance management is due to their work.

Organizational Psychology

Organizational psychology addresses the motivational and emotional side of work. To do that, it typically combines research and ideas from social psychology (the study of power and influence, attitudes and attitude change, communication in groups, and individual and group social behavior) with organizational behavior (a field devoted to understanding, explaining, and ultimately improving the attitudes and behavior of individuals and groups in organizations) (Colquitt et al., 2011).

Organizational psychology typically includes topics such as aging and retirement, turnover, retention, attitudes (e.g., job satisfaction, organizational commitment, employee engagement), fairness, organizational justice, motivation, decision making, stress, leadership, teams, the interaction of work with family (and vice versa), as well as organizational and work design. Managing change is also a key focus within organizational psychology. Such changes may stem from a variety of initiatives, such as mergers and acquisitions, restructuring, downsizing, or various efforts to enhance productivity and quality. For change-management initiatives to be effective, I–O psychologists need to understand thoroughly the theoretical bases for the interventions they use, as well as research to assess the impact of the interventions on people and organizations.

Human Engineering

Human engineering (also known as human factors psychology or engineering psychology) is the study of the capacities and limitations of people with respect to their work environments. The goal is to develop an environment that is compatible with the worker, whether that worker is a pilot interacting with the controls and displays in an airplane cockpit, a computer user navigating a graphical user interface, or a truck driver who uses controls and displays to operate a vehicle.

In a systems approach to human engineering, the human operator is the central figure within a complex sociotechnical system. That system is composed of the operator, the technology (e.g., tools, computers) necessary to complete the required tasks, and the work environment (the organization, physical setting, and task requirements). One approach to system design has been the sociotechnical systems approach, in which design is driven by analysis of the environment, the social system, and the technology system. The objective is to maximize the fit among all three components (Paley and Grier, 2007; Salvendy, 2006).

More recently, human engineering, like much of psychology, has been greatly influenced by advances in cognitive science. To appreciate this, consider two other approaches to system design: user-centered design and participatory ergonomics. With user-centered design, the primary objective is to design systems that explicitly satisfy both the task and information-processing requirements of the human operator. Participatory ergonomics incorporates user involvement throughout the design process, but it goes further, allowing users, with expert support, to complete the design.

Cognitive–perceptual principles guide the design of user interfaces, the primary means by which users interact with systems. As an example, consider the design of a control panel in a power plant. Here are four such principles (Paley and Grier, 2007):

- Logical groupings of displays,
- Consistency of look and function (like items look and function similarly),
- Labels are legible, visible, meaningful, and discriminable, and
- Controls are reachable and movable by the population in the work environment.

Clearly, user-interface designs and system designs can have a major impact on the fit between humans and the environments in which they work.

Having reviewed the broad domain of I–O psychology, it should come as no surprise that I–O psychologists work in a variety of settings (Society for Industrial and Organizational Psychology, Inc., 2014). In academia, they may be full, associate, or assistant professors of psychology, management, organizational behavior, or industrial relations. In other settings they may hold the titles of corporate vice president, director, manager, or staff member of organizational development, management development, human resources research, employee relations, training and development, or leadership development. Finally, they may be presidents, vice presidents, or directors of private research organizations, consulting companies, or other organizations.

A Brief Overview of the Evolution of I–O Psychology

The objective of early industrial psychology was to improve organizational goals (productivity and efficiency), primarily by applying psychology with an emphasis on individual differences through employee selection and training (Koppes, 2007). In 1913, Hugo Munsterberg's book, *Psychology and Industrial Efficiency*, described experiments in selecting streetcar operators, ship's officers, and telephone switchboard operators. His contributions to I–O psychology are noteworthy for his emphasis on the analysis of jobs in terms of (1) the abilities required to do them and (2) the development of testing devices – aptitude as well as work sample tests – that were able to forecast subsequent job performance (Munsterberg, 1913).

Munsterberg's 1917 book, *Business Psychology*, included the results of questionnaires sent to hundreds of laborers. The questionnaires requested anonymous data on attitudes toward their work, involvement in the labor movement, and leisure activities.

Several decades earlier, the scientific management movement began with the goal of bringing order and rationality to the practice of management. Frederick Winslow Taylor was the prophet of scientific management, and the stopwatch was his 'bible' (Bell, 1972). Taylor began his experiments in the steel industry in 1885. His objective was the systematic analysis and breakdown of work into its smallest mechanical elements, and then their rearrangement into the most efficient combination. The objective was to identify the 'one best way' to perform work. In addition to the scientific study of the task itself ('time and motion' study), Taylor argued that individuals selected to do the work should be as perfectly matched physically and mentally to the demands of the job as possible, and that overqualified individuals should be excluded.

Taylor believed that employees should be trained carefully by supervisors (whose own work was also divided into specialties) to ensure that they perform the work exactly as specified by prior scientific analysis. In no case, however, should employees ever be called upon to work at a pace that would be detrimental to their health.

Finally, to provide an incentive for the employee to follow the detailed procedures specified (and closely supervised by line supervisors on the shop floor), Taylor felt that employees should receive an addition of between 30 and 100% of their ordinary wages when the task was done right and within the time limits specified. The first piece-rate incentive system had arrived.

Taylor was interested in the social aspects of work, although he saw little good emerging from the social interaction within work groups. He felt that work groups fostered a level of individual efficiency equal to the level of the least productive worker in the group. In other words, he believed that the efficiency of the group would not be any greater than the efficiency of the least productive member. Thus, Taylor tried to institute a kind of social physics into the workplace.

Both Munsterberg and Taylor were concerned with how workers were selected for jobs, as well as with the individual worker's feelings and aspirations. Although they differed on several issues (notably, on how social structure affected productivity), Munsterberg's writings clearly suggest that he felt that the two emerging fields of personnel selection and social psychology (the study of how people affect and are affected by one another) were complementary (Landy, 1992; Moskowitz, 1977).

When the United States entered World War I in 1917, psychologists Walter Dill Scott, Walter VanDyke Bingham, and Robert Yerkes collaborated in developing intelligence tests that could be used to select and classify recruits. The Army Alpha was administered to individuals with English language skills, while the Army Beta was administered to recruits who were illiterate or did not have strong English language skills. These tests paved the way for large-scale intelligence testing and for later expansion of psychological testing into government, industry, and education (Vinchur and Koppes, 2011).

Other developments at about the same time included the use of rating sheets for interviewers, advances in checking references given by workers, and in statistical methods for estimating validity (the extent to which selection devices accurately forecast job performance). Other kinds of psychological measures also appeared during and after World War I, such as measures of aptitude, interest, and personality. The measurement of attitudes was made possible by the scaling techniques of L.L. Thurstone in 1927, and improved further by Rensis Likert in 1932. Likert-type rating scales (e.g., 'strongly agree, agree, neither agree nor disagree, disagree, strongly disagree') are still popular today.

The Hawthorne Experiments

In Chicago, in 1923, the Hawthorne Works of the Western Electric Company provided the setting for one of the most famous behavioral research efforts of all time. The purpose of the research was to identify variables that affect worker productivity. Experiments by professors Elton Mayo, Fritz Roethlisberger, and T. North Whitehead of Harvard Business School indicated that productivity was directly related to the degree of group teamwork and cooperation. The level of teamwork and cooperation, in turn, seemed to be related to the interest of the supervisor and the researchers in the work group, the lack of coercive approaches to productivity improvement, and the participation afforded the workers in changes affecting them (Pennock, 1930; Roethlisberger and Dickson, 1939). In short, the researchers came to view an organization of workers as a *social system*, in contrast to Taylor's view of an organization as a *technical–economic system*. Interest and research in topics such as leadership, motivation, group processes, and job

satisfaction have direct antecedents in the Hawthorne research (Vinchur and Koppes, 2011).

World War II brought new challenges for I–O psychology, and also many new developments. These included the Army General Classification Test, situational stress tests, and the selection and simulation training of pilots to fly warplanes. In addition, building on work done by German psychologists and that of the War Office Selection Board of the British army in the early 1940s, the US Office of Strategic Services used the assessment center method to select spies and sabotage agents during World War II. Each candidate had to develop a cover story that would hide his identity during the assessment. Testing for the ability to maintain cover was crucial, and ingenious situational tests were designed to seduce candidates into breaking cover (McKinnon, 1975; Office of Strategic Services, 1948).

Post–World War II, the US economy thrived, jobs were plentiful, and developments in I–O psychology focused on job analysis, job evaluation, staffing, training, performance appraisal, labor relations, equipment design, and reducing accidents to enhance safety. The conviction that group behavior and workers' feelings were associated with morale and productivity characterized much of the research and theorizing in what became known as the 'human relations movement' for the next two decades. Influential thinkers such as Douglas McGregor (Theory X and Theory Y managers), Chris Argyris (modern organizations are in conflict with the personalities of mature adults), and Rensis Likert (the 'linking pin' model of integrating small groups into the organization) focused on the fit between organizations and individuals. Their ideas can be summed up concisely: organizations and people need each other, and a good fit between the two benefits both (Shafritz and Ott, 1996).

The 1960s and 1970s

The Vietnam War, baby boomers in the workplace, and the rise of foreign competition led a new generation of employees to question the authority of organizations, which stimulated concerns about democracy, discrimination, and autonomy in the workplace. Major civil rights legislation was passed, including the 1963 Equal Pay Act (outlawing gender-based discrimination in pay); the 1964 Civil Rights Act (Title VII of which prohibited discrimination in employment on the basis of race, color, religion, sex, or national origin); and the Age Discrimination in Employment Act (which forbids discrimination against employees 40 years of age and over). This legislation coupled with major court decisions, such as *Griggs v. Duke Power Company* (1971), significantly influenced research in I–O psychology concerned with job analysis, test validation, and fair employment practices. After publication of the federal *Uniform Guidelines on Employee Selection Procedures* (1978), I–O psychologists became heavily involved in defending (or challenging) tools and techniques used for personnel decisions. Validity generalization (see below) and meta-analysis were introduced as approaches to support generalizing the findings of validity studies across jobs and organizations. At the same time, cognitive-based theories began to emerge in areas such as leadership and motivation. Finally, in 1976 Marvin Dunnete published a one-volume *Handbook of Industrial and Organizational Psychology* to chronicle the evolution, development, and future challenges facing the field.

From the 1980s Forward

With the fall of Communism, the end of the Cold War, and the passage of the North American Free Trade Agreement, a global and diverse workforce began to emerge. Downsizing white-collar workers became more common, along with the widespread restructuring of organizations (also known as process reengineering), and mergers and acquisitions. Passage of the Americans with Disabilities Act (ADA) in 1990 created the need to identify essential job functions, physical requirements of jobs, and brought renewed interest in job design. Likewise, passage of the Civil Rights Act of 1991, which amended the Civil Rights Act of 1964, provided the sanctions for violations of the ADA and outlawed the practice of 'race norming'. Using that procedure, percentile scores on a selection procedure were computed relative to one's own race/ethnic group, and then merged with those from other race/ethnic groups into a single list. The prohibition against race norming subsequently sparked renewed interest in research on fair personnel decisions.

Over the past 30 years, there have been many developments in I–O psychology research and practice. Some of these include, for example, a cognitive perspective on performance appraisals; a shift toward performance management (which incorporates ongoing coaching and feedback); organizational justice theory; the use of personality tests in employment; computerized adaptive testing; and widespread use of the Internet and social media for recruitment, assessment, and selection. Two other noteworthy developments were the Occupational Information Network to replace the former Dictionary of Occupational Titles, the second edition of the four-volume *Handbook of Industrial and Organizational Psychology* in 1991, and, more recently, a three-volume *APA Handbook of Industrial and Organizational Psychology* in 2011.

The Scientist–Practitioner Model and the Role of Research in Guiding Science and Practice

There is a strong connection between research that is conducted using scientific methods and the practice of I–O psychology. For decades, the professional association for I–O psychologists, namely, the Society for Industrial and Organizational Psychology, has endorsed the scientist–practitioner model (Bass, 1974; Cascio and Greene, 2012; Dunnette, 1990; Murphy and Saal, 1990; Rupp and Beal, 2007). That model discourages both practice that has no scientific basis and research that has no clear implications for practice.

I–O psychology is an applied discipline, as reflected in the title of the premier journal in the field, *Journal of Applied Psychology*. At the same time, however, what differentiates I–O psychology from many other applied disciplines is its reliance on scientific methods to produce basic research and theory. That basic research and theory often lead to strategies and systems that are relevant to individuals, teams, organizations, and society. As an example, consider the theory of validity generalization (VG).

A traditional belief among personnel psychologists was that validity coefficients, that is, correlations of scores on preemployment tests with measures of job performance, were highly specific to each employment situation. Hence, an empirical study would be required in each situation, and the ability to generalize validities across similar situations would not be possible.

Schmidt and Hunter (1977) and Hunter and Schmidt (1990) hypothesized that the variability across studies in raw validity coefficients, even when jobs and tests appear to be similar or essentially identical, might be artifactual in nature. In developing a model to test this hypothesis they identified seven potential sources of artifactual, between-study variance in observed validity coefficients, the most important of which is sampling error. Applying a Bayesian statistical model, they demonstrated that when jobs are similar in different employment situations, it is possible to generalize validity to new settings without carrying out an empirical validation study in each new setting.

The more than 500 VG studies and technical refinements of VG procedures that now exist in the literature in applied psychology (for reviews see Murphy, 2003; Schmidt and Hunter, 2003) provide disconfirming evidence that validity is situationally specific, and indicate that VG is a robust phenomenon. Perhaps the most important implication of this work is that it has called attention to the fact that the mean of several validity coefficients may be a better basis for inferring a valid relationship between a predictor (e.g., an employment test) and a criterion (a measure of job performance) than any one coefficient (Society for Industrial and Organizational Psychology, Inc., 2003).

The implications of VG research have direct implications for practice, and several authors have identified key considerations for practitioners to consider when assessing VG evidence (Schmitt, 1996; Wanous et al., 1989). The many studies that VG has enabled have made it possible to develop general principles and theories that have helped to propel the field beyond a mere technology to the status of a science (Guion, 1976).

Some Factors Driving Demand for Research and Expertise in I–O Psychology

Dramatic changes are occurring in the social, legal, political, and economic environments everywhere. Some of these issues include finding and keeping top talent in tight labor markets, and capitalizing on the information revolution in ways that will benefit organizations, such as social media and other Internet-based recruitment sources. Other challenges include managing the performance of geographically dispersed employees and teams (virtual teams), and providing training and development opportunities on demand.

Many of the ways of working that individuals have taken for granted are disappearing – working 9–5, spending an entire career with only one company, or working in an office everyday. Similar changes are happening in organizations. For example, the idea that hierarchy is the best way to manage information flows, that most people will work with team members in the same office, or that the majority of talent will be held within the boundaries of an organization. All of this is shifting, and what is coming in its place is less knowable and less understandable – accelerating developments in technology, accelerating globalization, accelerating demographic changes, and accelerating social trends (Cascio and Aguinis, 2008; Gratton, 2011).

Indeed, this is today's paradigm of work: anytime, anywhere, in real space, or in cyberspace. This presents ongoing challenges in areas such as work-life fit; developing a total rewards package that will attract, retain, and motivate employees at all levels; measuring social networks to identify trusted connections; promoting collective leadership; and promoting and enabling ambidextrous leadership that will identify new growth businesses, while remaining focused on the execution of existing ones (Cascio and Boudreau, 2011). I–O psychology will contribute to the bottom-line success of organizations and to the betterment of employee welfare by developing research-based responses to the workplace challenges wrought by this new paradigm.

See also: Attitude Measurement; Family and Work; Female Power at Work; Five Factor Model of Personality, Assessment of; Human Cognition, Evolution of; Human Factors and Ergonomics; Human Resource Management, Psychology of; Law: History of Its Relation to the Social Sciences; Personality and Values at Work; Personnel Selection, Psychology of; Technology and Organization; Work Motivation.

Bibliography

Aumann, K., Galinsky, E., 2009. The 2008 National Study of the Changing Workforce, the State of Health of the American Workforce: Does Having an Effective Workplace Matter? Families and Work Institute, New York.

Bass, B.M., 1974. The substance and the shadow. American Psychologist 29, 870–886.

Bell, D., 1972. Three technologies: size, measurement, hierarchy. In: Davis, L.E., Taylor, J.C. (Eds.), Design of Jobs. Penguin, London, pp. 54–78.

Cascio, W.F., 2013. Managing Human Resources: Productivity, Quality of Work Life, Profits, ninth ed. McGraw-Hill, New York.

Cascio, W.F., Aguinis, H., 2008. Research in industrial and organizational psychology from 1963–2007: changes, choices, and trends. Journal of Applied Psychology 93, 1062–1081.

Cascio, W.F., Aguinis, H., 2011. Applied Psychology in Human Resource Management, seventh ed. Prentice-Hall, Upper Saddle River, NJ.

Cascio, W.F., Boudreau, J.W., 2011. What's around the corner in HR? Organizational dynamics. Special Issue on the Future of HR 40 (4), 243–245.

Cascio, W.F., Greene, R.J., 2012. The employee-organization relationship and the scholar-practitioner divide. In: Shore, L.M., Coyle-Shapiro, J.A., Tetrick, L.E. (Eds.), The Employee-Organization Relationship: Applications for the 21st Century. Applied Psychology Series, Psychology Press/Routledge, New York, pp. 553–573.

Colquitt, J.A., LePine, J.A., Wesson, M.J., 2011. Organizational Behavior: Improving Performance and Commitment in the Workplace, second ed. McGraw-Hill, New York.

Dunnette, M.D., 1990. Blending the science and practice of industrial and organizational psychology: where are we and where are we going? In: Dunnette, M.D., Hough, L.M. (Eds.), Handbook of Industrial and Organizational Psychology, second ed., vol. 1. Consulting Psychologists Press, Palo Alto, CA, pp. 1–27.

Gratton, L., 2011. Workplace 2025 – what will it look like? Organizational Dynamics 40, 246–254.

Griggs v. Duke Power Company, 1971. 401 U.S. 424.

Guion, R.M., 1976. Recruiting, selection, and job placement. In: Dunnette, M.D. (Ed.), Handbook of Industrial and Organizational Psychology. Rand McNally, Chicago, IL, pp. 777–828.

Halpern, D.F., Murphy, S.E., 2005. From Work-Family Balance to Work-Family Interaction: Changing the Metaphor. Erlbaum, Mahwah, NJ.

Hitt, M.A., Miller, C.C., Collela, A., 2011. Organizational Behavior, third ed. Wiley, Hoboken, NJ.

Hunter, J.E., Schmidt, F.L., 1990. Methods of Meta-analysis: Correcting Error and Bias in Research Findings. Sage, Newbury Park, CA.

Koppes, L.L., 2007. History of industrial/organizational psychology in North America. In: Rogelberg, S.G. (Ed.), Encyclopedia of Industrial and Organizational Psychology, vol. I. Sage, Thousand Oaks, CA, pp. 312–317.

Landy, F.J., 1992. Hugo Munsterberg: victim or visionary? Journal of Applied Psychology 77, 787–802.

McKinnon, D.W., 1975. Assessment centers then and now. Assessment and Development 2, 8–9.

Moskowitz, M.J., 1977. Hugo Munsterberg: a study in the history of applied psychology. American Psychologist 32, 824–842.

Munsterberg, H., 1913. Psychology and Industrial Efficiency. Houghton Mifflin, Boston, MA.

Murphy, K.R. (Ed.), 2003. Validity Generalization: A Critical Review. Lawrence Erlbaum, Mahwah, NJ.

Murphy, K.R., Saal, F.E., 1990. What should we expect from scientist-practitioners? In: Murphy, K.R., Saal, F.E. (Eds.), Psychology in Organizations: Integrating Science and Practice. Erlbaum, Hillsdale, NJ, pp. 49–66.

Office of Strategic Services (OSS) Assessment Staff, 1948. Assessment of Men. Rinehart, New York.

Oswald, F.L., Hough, L.M., 2011. Personality and its assessment in organizations: theoretical and empirical developments. In: Zedeck, S. (Ed.), APA Handbook of Industrial and Organizational Psychology, vol. 2. American Psychological Association, Washington, DC, pp. 153–184.

Paley, M.J., Grier, R., 2007. Engineering psychology. In: Rogelberg, S.G. (Ed.), Encyclopedia of Industrial and Organizational Psychology, vol. 1. Sage, Thousand Oaks, CA, pp. 206–209.

Pennock, G.A., 1930. Industrial research at Hawthorne and experimental investigation of rest periods, working conditions, and other influences. Personnel Journal 8, 296–309.

Roethlisberger, F.J., Dickson, W.J., 1939. Management and the Worker. Harvard University Press, Cambridge, MA.

Rogelberg, S.G. (Ed.), 2007. Encyclopedia of Industrial and Organizational Psychology, vol. 1. Sage, Thousand Oaks, CA, p. xxxv.

Rupp, D.E., Beal, D., 2007. Checking in with the scientist-practitioner model: how are we doing? Industrial and Organizational Psychologist 45, 35–40.

Salvendy, G., 2006. Handbook of Human Factors and Ergonomics, third ed. Wiley, New York.

Shafritz, J.M., Ott, J.S. (Eds.), 1996. Classics of Organization Theory, fourth ed. Wadsworth, Belmont, CA.

Schmidt, F.L., Hunter, J.E., 1977. Development of a general solution to the problem of validity generalization. Journal of Applied Psychology 62, 529–540.

Schmidt, F.L., Hunter, J.E., 2003. Meta-analysis. In: Schinka, J.A., Velicer, W.F. (Eds.), Handbook of Psychology: Research Methods in Psychology, vol. 2. Wiley, Hoboken, NJ, pp. 533–554.

Schmitt, N., 1996. Validity generalization. In: Barrett, R.S. (Ed.), Fair Employment Strategies in Human Resource Management. Quorum Books/Greenwood Publishing Group, Westport, CT, pp. 94–104.

Society for Industrial and Organizational Psychology, Inc., 2003. Principles for the Validation and Use of Personnel Selection Procedures, fourth ed. Author, Bowling Green, OH.

Society for Industrial and Organizational Psychology, Inc., 2014. Building Better Organizations, Part 2. Retrieved from: http://www.siop.org/visibilitybrochure/building2revised.pdf.

Uniform guidelines on employee selection procedures. Federal Register 43, 1978, 38290–38315.

Vinchur, A.J., Koppes, L.L., 2011. A historical survey of research and practice in industrial and organizational psychology. In: Zedeck, S. (Ed.), Handbook of industrial and organizational psychology, vol. 1. American Psychological Association, Washington, DC, pp. 3–36.

Wanous, J.P., Sullivan, S.E., Malinak, J., 1989. The role of judgment calls in meta-analysis. Journal of Applied Psychology 74, 259–264.

Zedeck, S., 2011. Introduction. In: Zedeck, S. (Ed.), APA Handbook of Industrial and Organizational Psychology, vol. 1. American Psychological Association, Washington, DC, pp. xxi–xxxiv.

Inequality Analysis: Overview

Guillermina Jasso, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Inequality is pervasive in the social life, and understanding it is a central task for the social sciences. This article provides an overview of inequality analysis, discussing the framework with its questions and building blocks and surveying the twin branches of theoretical inequality analysis and empirical inequality analysis, the latter only briefly given that dozens of other articles discuss empirical studies of inequality. Substantively, this article identifies two main types of inequality – inequality between persons and inequality between subgroups – and highlights the fundamental questions concerning the determinants and consequences of inequality. Methodologically, it summarizes the major tools for measuring both kinds of inequality, including both scalar measures and whole-distribution approaches. To illustrate theoretical inequality analysis, the article reviews recent examples of a priori explorations, including mechanisms for altering inequality.

Introduction

For all of recorded history, questions about inequality have commanded human attention. From Plato (*c*.428–348/7 BC) and Aristotle (384–322 BC), through John Chrysostom (*c*.350–407) and Gregory the Great (540–604), to Rousseau (1712–78) and Mill (1806–73), and finally to social scientists since the nineteenth century, studying inequality has been central to analyses of the human condition. For example, Plato (*Laws*, Book V) insight that social relations and societies suffer when the largest income exceeds five times the smallest income, John Chrysostom's (Homily 66.3, On the Gospel of Matthew) effort to estimate the proportions poor and rich in the city of Antioch, Jerome's concern that marriage hurts women (and his celebrated advice, "do not marry"), and Gregory the Great's (*Moralia in Job*, xxi) observation, "where there is no sin, there is no inequality," have helped to shape social thought and social science. The measurement tools provided in the twentieth century by Gini (1913–1914), Theil (1967), Atkinson (1970), and others brought new rigor to the study of inequality, laying the foundation for modern inequality analysis.

Framework for Inequality Analysis

At its most general, inequality refers to variability in a characteristic among the units of a group or population. The units may be at any level of analysis, for example, persons, cities, countries. Of course, not all variability rises to the level of inequality, as Aquinas ([1265–1274]1952) argued against Gregory, but certainly variability in social goods does.

There are two main kinds of inequality and two core questions in inequality analysis. The two main kinds of inequality are inequality between units of the population (as in earnings inequality between persons or resource inequality between countries) and inequality between subsets of a population (as in the race earnings gap or the resource gap between subsets of countries). The two core questions address the determinants and consequences of both types of inequality. The starting building blocks pertain to quantitative and qualitative characteristics of the units. For simplicity and clarity, this article is framed in terms of persons but extension to other units (such as cities or countries) is straightforward. Accordingly, we begin with personal quantitative and qualitative characteristics.

Basic Building Blocks

Personal quantitative characteristics are characteristics of which there can be more or less, or on which a person may rank higher or lower. The first kind – of which there can be more or less – is called *cardinal*; examples are wealth, land, head of cattle. The second kind – on which a person may rank higher or lower – is called *ordinal*; examples include beauty, intelligence, athletic skill.

Personal qualitative characteristics are unorderable, categorical personal characteristics. They may be dichotomous – like gender – or polytomous – like race, ethnicity, or religious affiliation. Personal qualitative characteristics give rise to groups and subgroups, with these groups and subgroups defined by the category of a personal qualitative characteristic. For example, the group "residents of the United States" is defined in terms of a category of the personal qualitative characteristic "country of residence," and the subgroups "men" and "women" are defined in terms of the categories of the personal qualitative characteristic "gender."

Both quantitative and qualitative characteristics are at the heart of inequality, and it would be impossible to study either type of inequality without them. The first type of inequality – called *personal inequality* or *overall inequality* – measures inequality in a quantitative characteristic X among a group specified by a qualitative characteristic – for example, income inequality among residents of Illinois or wealth inequality among members of a social club.

Thus, the simple operation of measuring personal inequality – for example, calculating the Gini coefficient or one of the Theil coefficients – requires both a personal quantitative characteristic and a personal qualitative characteristic.

Similarly, the second type of inequality – called *subgroup inequality* – refers to the discrepancy between the distribution (most often the average or other measure of central tendency) of X across two subgroups of a group or population, as in the race wage gap or the gender wage gap, and thus also requires both a personal quantitative characteristic and a personal qualitative characteristic.

The distinction between quantitative and qualitative characteristics, long appreciated in mathematics and statistics (e.g., Allen, 1938:10–11), has since the pioneering work by Blau (1974) come to be seen as structuring behavioral and social phenomena in a fundamental way. Here, we see that quantitative and qualitative characteristics play key and distinctive parts in the study of inequality.

This setup gives rise to a set of basic quantities in the study of inequality. Consider a group or population of size N. Each person has an amount x of the quantitative characteristic X – for simplicity, henceforth income. Arranging everyone in ascending order of their x amount yields the absolute rank i. Three new quantities are immediate: the arithmetic mean μ, the individual's relative rank α (approximated by $[i/(N+1)]$), and the individual's relative income (defined as x divided by the arithmetic mean). These quantities are important both substantively and methodologically. Substantively, both the relative rank and the relative amount are involved in behavioral mechanisms, such as status and fairness. Methodologically, they appear in the formulas for inequality measures and in important tools such as Pen's Parade (1971), which visualizes income as a function of relative rank.

Measures of Inequality

Table 1 presents some convenient formulas for four major measures of overall inequality – the Gini coefficient, one of Atkinson's (1970) measures (defined as 1 minus the ratio of the geometric mean to the arithmetic mean), Theil's (1967: 125–127) mean logarithmic deviation (MLD), and the coefficient of variation (CV) – whose properties are well known (see, e.g., Cowell, 1977; Dagum, 1983; Kleiber and Kotz, 2003). Formulas are provided for both observed and mathematically specified distributions. As shown, the formulas include the quantities introduced above, namely the individual's income amount x, absolute rank i, and relative rank α. As well, the relative amount appears in many inequality formulas (Cowell, 1977; Firebaugh, 1999; Jasso and Kotz, 2008: 71).

Table 1 Some Convenient Formulas for Four Inequality Measures in Observed and Mathematically Specified Distributions

Inequality measure	*Observed distributions*	*Mathematically specified distributions*
Gini coefficient	$\frac{2\sum_{i=1}^{N} i x_i}{\mu N(N-1)} - \frac{N+1}{N-1}$	$1 - \frac{2}{\mu}\int_0^1\int_0^\alpha Q(\alpha)d\alpha\, d\alpha$
Atkinson's measure	$1 - \left(\frac{G}{\mu}\right)$	$1 - \frac{\exp\left\{\int_0^1 \ln[Q(\alpha)]d\alpha\right\}}{\mu}$
Theil's MLD	$\ln\left(\frac{\mu}{G}\right)$	$\ln(\mu) - \int_0^1 \ln[Q(\alpha)]d\alpha$
Pearson's coefficient of variation	$\frac{\sigma}{\mu}$	$\left\{\int_0^1 [Q(\alpha)]^2 d\alpha - \mu^2\right\}^{1/2} \Big/ \mu$

Notes: The letters x, i, and N denote the magnitude of X, the absolute rank (arranged in ascending order), and the population size, respectively. The letters μ, G, and σ denote the arithmetic mean, geometric mean, and standard deviation, respectively. Formulas for mathematically specified distributions are expressed in terms of the quantile function (QF) and the mean μ.
Source: Jasso and Kotz (2008:38).

Probability Distributions

For theoretical inequality analysis, there is no more helpful tool than classical probability distributions. Theoretical analysis using probability distributions begins with selection of a set of modeling distributions for the quantitative characteristic X. A useful set includes distributions with different combinations of lower and upper bounds, such as distributions with and without a safety net and distributions with and without a maximum amount of X. Here, we use three two-parameter continuous distributions whose properties are well known – the lognormal, Pareto, and power-function (Johnson et al., 1994(1995); Kleiber and Kotz, 2003). Together this set includes some desired combinations. The Pareto combines a safety net with no maximum. The lognormal lacks an upper bound, like the Pareto, but, unlike the Pareto, it has no safety net. The power function has a maximum but no safety net. All three are two-parameter distributions and expressed in terms of the mean μ and a second parameter, denoted c, which Jasso and Kotz (2008), building on the literature pioneered by Pareto (1897) and more recent contributions such as Cowell (1977: 95) and Kleiber and Kotz (2003: 78), suggest operates as a general inequality parameter governing all measures of overall inequality – a parameter of which all measures of overall inequality are monotonic functions. In the lognormal, inequality increases as c increases, while in the Pareto and power function, inequality decreases as c increases.

Formulas for the main associated functions of these variates – the probability density function (PDF), the cumulative distribution function (CDF), and the quantile function (QF) – are presented in Table 2. Figure 1 provides visual illustration. Relations among these functions in continuous univariate distributions are well known. The PDF is the first derivative of the CDF, and the CDF and QF are inverses of each other. Visually, the graph of the QF can be glimpsed by turning over the graph of the CDF and rotating it 90°.

For analyzing inequality, the view of choice is the one afforded by the QF (portrayed in the bottom row of Figure 1), which expresses x as a function of the relative rank. And there is no more evocative imagery than that introduced by Jan Pen (1971). Imagine an hour-long parade of all members of society, with each person passing at an equal pace (visible for $1/N$th of the hour) and each person's height proportional to his or her magnitude of the variable. If there is perfect equality, all persons would have the same height, but if there is inequality Pen's Parade would start with dwarfs and end with giants. The QF thus provides a whole-distribution tool for assessing inequality.

Theoretical Inequality Analysis

Some theoretical analyses establish properties of and relations between inequality measures and their decompositions. Others

Table 2 Three continuous univariate distributions and associated functional characteristics

Variate	*Cumulative distribution function*	*Probability density function*	*Quantile function*
Lognormal $x > 0$, $c > 0$	$F_N\left\{\frac{\left[\ln\left(\frac{x}{\mu}\right)+\frac{c^2}{2}\right]}{c}\right\}$	$\frac{1}{xc\sqrt{2\pi}}\exp\left\{-\frac{\left(\frac{c^2}{2}+\ln\frac{x}{\mu}\right)^2}{2c^2}\right\}$	$\mu\exp\left[cQ_N(\alpha)-\frac{c^2}{2}\right]$
Pareto $x > \frac{\mu(c-1)}{c}$, $c > 1$	$1-\left[\frac{\mu(c-1)}{cx}\right]^c$	$\left[\frac{\mu(c-1)}{c}\right]^c cx^{-c-1}$	$\frac{\mu(c-1)}{c(1-\alpha)^{1/c}}$
Power-function $0 < x < \frac{\mu(c+1)}{c}$, $c > 0$	$\left[\frac{xc}{\mu(c+1)}\right]^c$	$\left[\frac{c}{\mu(c+1)}\right]^c cx^{c-1}$	$\frac{\mu(c+1)\alpha^{1/c}}{c}$

Notes: For all variates, $x > 0$; other restrictions as indicated. The expressions $F_N(\cdot)$ and $Q_N(\cdot)$ denote the cumulative distribution function and the quantile function, respectively, of the standard normal variate:

$$F_N(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt;$$

$$Q_N(\alpha) = \sqrt{2}\,\mathrm{erf}^{-1}(2\alpha - 1)$$

where erf denotes the error function. Inequality is a decreasing function of c for the Pareto and the power-function variates and an increasing function of c for the lognormal distribution.
Source: Jasso, Guillermina, Samuel Kotz, 2008. Two types of inequality: inequality between persons and inequality between subgroups. Sociological Methods and Research 37, 36.

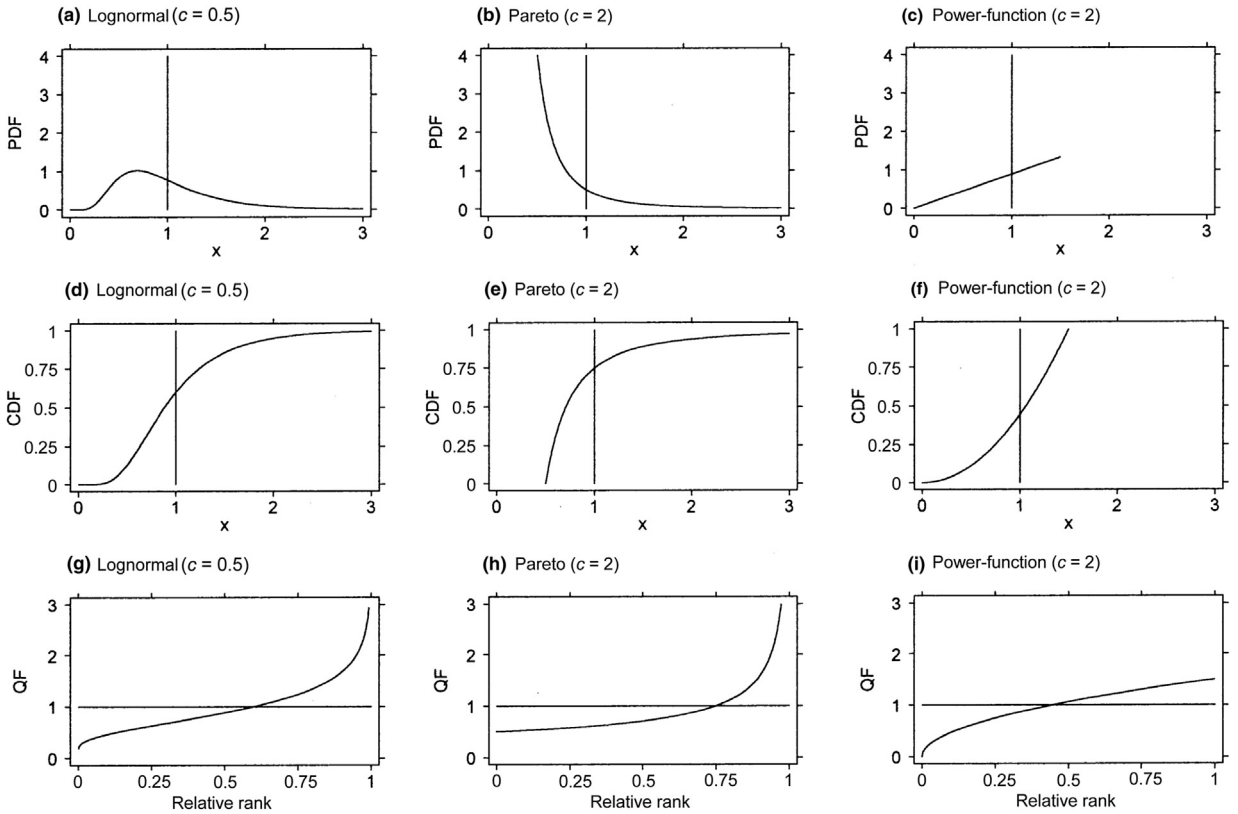

Figure 1 Basic associated functions (PDF, CDF, and QF) in three continuous univariate distributions: lognormal, pareto, and power-function. Arithmetic mean fixed at unity and displayed as either a vertical line (PDF and CDF) or horizontal line (QF). PDF = probability density function; CDF = cumulative distribution function; QF = quantile function. Source: Jasso, Guillermina, Samuel Kotz, 2008. Two types of inequality: inequality between persons and inequality between subgroups. Sociological Methods and Research 37, 37.

investigate determinants of inequality, and still others, consequences. This section provides some examples.

Theoretical Inequality Analysis – Properties and Relations of Inequality Measures

Two major sets of properties and relations pertain to (1) operations that increase or decrease inequality, and (2) subgroup structures and decompositions of inequality measures.

Altering Inequality – Transfers. The basic idea underlying transfer operations is that a rank-preserving transfer from a richer to a poorer person reduces inequality, while a rank-preserving transfer from a poorer to a richer person increases inequality (Dagum, 1983). The archetypal case involves transfers from one extreme of the distribution to the other, for example, from the richest person to the poorest person. There is a large literature building on the basic idea, proposing a principle of transfers that measures of inequality should satisfy, together with weak and strong versions of the principle of transfers.

To visualize transfers, Figure 2 presents graphs of the QFs of two members of each of the three variates in Table 2 and Figure 1 – the lognormal, Pareto, and power-function. For simplicity, the distributions are of relative amounts (x/μ), so that their mean is fixed at one. The point where the two curves intersect is the boundary between the set who receive transfers and the set who give transfers. In these two-parameter distributions, the curves intersect only once, and there is a tight link between inequality and transfers. When the leftmost – or poorest – receive transfers, inequality decreases. Thus, the distribution with the less inequality is the distribution whose QF is higher in the left region – the distribution in which the poorest are better-off. For example, in panel C, the Pareto that has the higher minimum income also has the lower inequality (as Rawls (1971) would recognize). Behaviorally, when the income distribution changes from one to another of a within-variate pair, as in Figure 2, and the poorest favor the lower-inequality distribution, it is impossible to know whether they are selfish (maximizing own income) or altruistic (reducing inequality).

But transfers can simultaneously include both rich-to-poor transfers and poor-to-rich transfers. As is well known, when that happens, different inequality measures may yield different orderings. To visualize these situations, Figure 3 reports the QFs of distributions drawn from different variate families. As in Figure 2, the distributions are of relative amounts. In the cases depicted, the two QFs cross twice, so that the economic fate of the poorest and richest is linked, and either both gain or both lose, always opposite to the fate of the middle region. For example, in panel C, the poorest and the richest are better-off in the Pareto than in the power-function distribution, while the middle is better-off in the power-function; similarly, in panel D, the poorest and the richest are better-off in the Pareto than in the lognormal distribution, while the middle is better-off in the lognormal. The pair of distributions in panel D is especially interesting, because the median person is not in the majority – which comprises the poorest 49% and the richest 1.3%.

Subgroups and Decompositions. A large literature examines decompositions of inequality measures as well as subgroup structures. For example, Liao (2006) develops a model-based cluster approach to estimate classes of individuals who share similar levels of income and then extends the Gini coefficient to enable a decomposition into individual inequality and class inequality. In a second example, Jasso and Kotz (2008) show that in a subset of societies, characterized by special income distributions (exemplified by continuous two-parameter distributions or distributions with nonintersecting Lorenz curves), as overall inequality increases, so does subgroup inequality; in such societies, when inequality attacks, it attacks on two fronts.

Theoretical Inequality Analysis – Determinants of Inequality

Wage-Setter Model. Recent work formulates a model in which N wage-setters recommend wages for R workers and each worker's final wage y is a weighted average of the recommended amounts x:

$$Y_r = \sum_{i=1}^{N} x_{ir} w_{ir}$$

where w denotes the wage-setter's weight and the weights sum to one (Jasso, 2009). Thus, the final wage distribution is a linear combination of the recommended wage distributions:

$$Y = w_1 X_1 + w_2 X_2 + \cdots + w_N X_N,$$

where X denotes each wage-setter's recommended wage distribution, Y denotes the final wage distribution, and, as above, w denotes each wage-setter's weight. Fundamental theorems in statistics provide the variance of the final wage distribution:

$$Var(Y) = \sum_{i=1}^{N} w_i^2 [Var(X_i)] + 2 \sum_{i=1}^{N} \sum_{k>i}^{N} w_i w_k Cov(X_i, X_k)$$

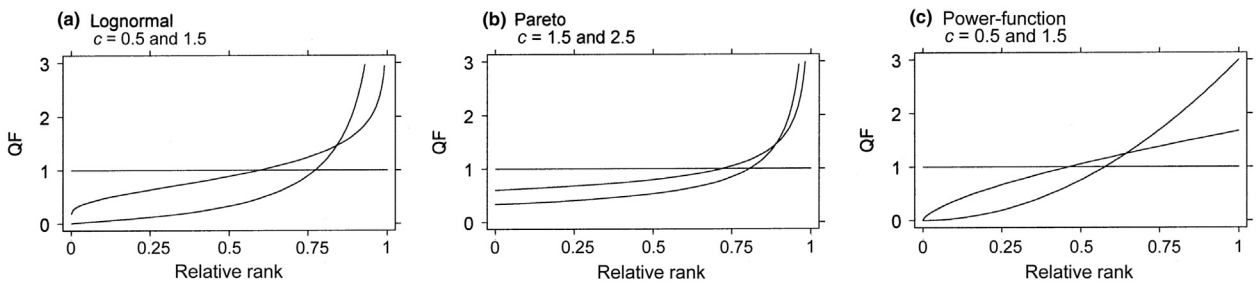

Figure 2 Quantile Functions (QFs) of Two Members of Three Variate Families. For simplicity, the distributions are of relative amounts (x/μ). Each plot also includes the graph of the equal distribution, shown as the horizontal line at the value 1. The point where the two members of each family intersect is the boundary between transfer groups.

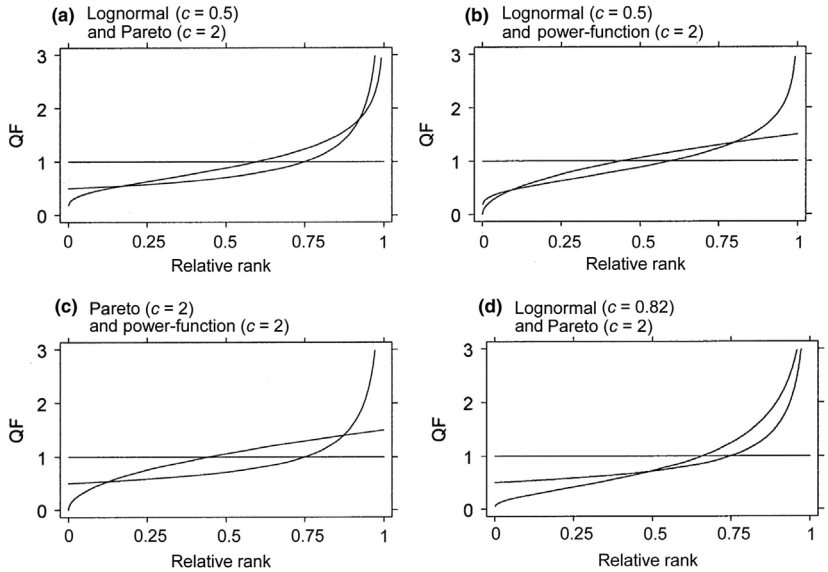

Figure 3 Quantile Functions (QFs) of Pairs of Distributions Drawn from Different Variate Families. For simplicity, the distributions are of relative amounts (x/μ). Each plot also includes the graph of the equal distribution, shown as the horizontal line at the value 1. The points where the two curves intersect is the boundary between transfer groups. In Panel D, the two curves intersect at α equal approximately 0.49 and 0.987 (the latter not visible in the graph).

Two analytic results follow immediately: First, inequality in the final wage distribution declines as the covariances among the recommended wage distributions go from positive to zero to negative. Second, inequality declines as the number of wage-setters increases.

The two analytic results are particularly easy to appreciate visually in the special case where the recommended wage distributions are identical and equally weighted:

$$\frac{Var(X)}{N} + \frac{2}{N^2} \sum_{i=1}^{N} \sum_{k>i}^{N} Cov(X_i, X_k)$$

In this case, when the recommended wage distributions are perfectly positively associated, the variance of Y reduces to the original variance of X. When they are independent, the covariances are zero and the variance reduces to the original variance divided by N – a steep decline. And when the sum of the covariances is negative, $Var(Y)$ declines still further.

The model can be applied both literally and metaphorically. For example, political parties may be thought of as wage-setters, given that their party platforms operate to produce what may be thought of as recommended wage distributions. Thus, the model implies that the greater the disagreements among parties, the lower the inequality; and the larger the number of parties, the lower the inequality. The model also yields the result that in the special case of two political parties, if their recommended wage distributions satisfy certain conditions (for example, they have equal finite variances and are either independent or identical and perfectly negatively associated), inequality is minimized when the two parties are equally weighted (i.e., share equal support among the electorate).

In both literal and metaphorical applications, the model raises new questions about determination of decision rules (such as number of wage-setters) and diversity of thought (such as how wage-setters come to disagree with each other). More broadly, these results may be part of larger fields. For example, the result on the number of wage-setters may be part of a larger class going back to Condorcet's Jury Theorem; similarly, the result on covariances may contribute to the nascent idea that diversity of thought reduces inequality and that seeking consensus may be counterproductive.

Wage Function with Multiple Criteria. Consider now a model in which each workplace unit has a wage function that it uses to establish the wages for its employees. The wage function may differ across workplace units in the number of employee characteristics it rewards and their weights. For example, some units may reward only a single factor (say, gross sales or output); other units, however, may reward additional factors (say, timeliness or courtesy). In contemporary vocabulary, we might say that the mix of skills and traits rewarded varies across workplace units:

$$Z = \beta_1 Q_1 + \beta_2 Q_2 + \cdots + \beta_K Q_K$$

where Z denotes the outcome (say, ln(wage)), Q denotes the reward-relevant traits and skills, of which there may be up to K, each with an associated coefficient β. Using the same reasoning as in the wage-setter model and the same fundamental results from statistics, the variance of Z is given by:

$$Var(Z) = \sum_{i=1}^{K} \beta_i^2 [Var(Q_i)] + 2 \sum_{i=1}^{K} \sum_{k>i}^{K} \beta_i \beta_k Cov(Q_i, Q_k)$$

Two analytic results are again immediate. First, as the covariances among rewarded skills and traits go from positive to zero to negative, wage inequality in the workplace units declines. Second, as the number of wage-relevant traits and skill increases, inequality decreases. This reasoning underlies a recent application in which the Q characteristics

are specified as relative ranks, so that Z has an exponential distribution in the case of one Q, and becomes Erlang, general Erlang, or mirror-exponential, depending on the number of wage-relevant characteristics Q and their associations (Jasso and Kotz, 2007: 320–321). Substantively, this model has deep sociological roots in the pioneering work of Berger et al. (1966: 44), who argued that status inequality declines when the characteristics generating status are negatively associated.

As in the wage-setter model, this model raises new questions, including questions about the process by which skills and traits come to be independent or dependent and the process by which employers decide which characteristics to reward (and how much). Further, this model points to a new direction in studying the shape of the income distribution, which has long defied easy characterization, namely that it may be a mixture of several orderly distributions specific to particular contexts (such as workplace units, departments of a firm, etc.) and that the key elements may be the number of characteristics valued by employers and the associations among the characteristics.

Assortative Mating. It has been known since Plato (*Laws* V) that positive assortative mating perpetuates social inequality while negative assortative mating reduces it, a powerful insight that informs contemporary social science (Jencks et al., 1972: 74, 211, 233–234, 271–273; Becker, 1991: 228). For example, the variance in the couples' income distribution declines as mating goes from positive assortative to random to negative assortative. What has been less appreciated is that under perfect positive assortative mating, husband and wife are equals, but in the other cases there are varying amounts of marital inequality (Jasso, 2011). Plato (*Laws* VI) quotes the saying, "equality makes friendship," but it was Aristotle (*Nicomachean Ethics*, Book VIII) who would note more trenchantly that love is possible only to equals, thus signaling the price of negative assortative mating. As **Table 3** shows, social inequality and marital inequality are on a collision course. What is good for society is bad for marriage, and what is good for marriage is bad for society.

Table 3 Variance in the married-couple income distribution and in the marital-inequality distribution, under three assortative mating regimes, for the case of identical spousal distributions

X_H and X_W Identical	Association between X_H and X_W		
	Perfect positive	*Independent*	*Perfect negative*
Married couple	$4\mathrm{Var}(X)$	$2\mathrm{Var}(X)$	$[2\mathrm{Var}(X)] \times (1+\rho)$
Marital inequality	0	$2\mathrm{Var}[\ln(X)]$	$\{2\mathrm{Var}[\ln(X)]\} \times (1-\rho)$

Notes: The correlation between the X_H and X_W distributions is denoted by ρ. The correlation ρ is always negative in the perfect negative association case, so that the right-hand factor attenuates the variance in the married-couple distribution and increases it in the marital-inequality distribution. The ordering of the variances across the row is opposite for the two distributions, decreasing in the married-couple distribution and increasing in the marital-inequality distribution, as the association between the two spouses' distributions goes from perfect positive to independent to perfect negative.
Source: Jasso, Guillermina, 2011. Plato: seven sociological ideas for the happy life. In: Edling, Christofer, Rydgren, Jens (Eds.), Sociological Insights of Great Thinkers: Sociology through Literature, Philosophy, and Science. Praeger, Santa Barbara, CA, p. 39.

Theoretical Inequality Analysis – Consequences of Inequality

Both inequality between persons and inequality between subgroups are pervasive. And it is widely believed that their consequences are equally pervasive. Theoretical inequality analysis provides a disciplined a priori approach, one that yields and makes visible a secret garden of testable inequality effects. The garden, however, requires a key. The key is a theory of some corner of behavior, a theory which turns out to yield inequality effects.

This section summarizes one logically necessary inequality effect and several predicted inequality effects. In every case, the theory has nothing about inequality in its assumptions; yet derivation yields outcomes with inequality in them. And in every case yielding predictions, the predictions are testable and contain no visible terms from the assumptions. In this type of deductive theory, the postulates are "genuine guesses about the structure of the world" (Popper, 1963: 245) and the predictions display the "marvelous deductive unfolding" of the theory (Popper, 1963: 221). The postulates' fruitfulness is evident in the "derivations far afield from its original domain," which "permit an increasingly broad and diversified basis for testing the theory" (Danto, 1967: 299–300). The promise of hypothetico-deductive theory is that a small set of Popperian guesses about deep mechanisms will generate a large and disparate set of effects, including novel predictions, and, conversely, that tests of the predicted effects will help assess the guesses' fidelity to the real world.

The theoretical consequences selected for this summary are derived from a branch of justice theory whose starting assumption is the justice evaluation function. The justice evaluation J represents the assessment that someone (self or other) is justly or unjustly rewarded and if unjustly rewarded whether underrewarded or overrewarded and to what degree. Perfect justice is represented by zero, unjust underreward by negative numbers, and unjust overreward by positive numbers. The justice evaluation function – the Popperian guess – specifies J as a function of the logarithm of the ratio of the actual reward A to the just reward C:

$$J = \theta ln\left(\frac{A}{C}\right)$$

where θ denotes the signature constant, whose sign (called the framing coefficient) is positive for goods and negative for bads and whose absolute value (called the expressiveness coefficient) measures expressiveness (Jasso, 1999: 138). Here interest centers on the experience of (in)justice about income, and thus θ vanishes.

It is important to know the properties of the keys used in theoretical analysis, in order to gauge their appropriateness and their price. Some properties of the justice evaluation function are: First, it provides an exact mapping from the combination of the actual reward and the just reward to the justice evaluation; for example, when A is less than C, J is negative. Second, the justice evaluation it yields is in justice units, not in reward units (as it would be, for example, if the functional form were a simple difference between A and C). Third, it

integrates the two main rival conceptions, the ratio favored by Homans (1961) and others and the difference favored by Berger et al. (1972); this is accomplished via the property of logarithms that the logarithm of a ratio equals the difference between two logarithms. Fourth, it embeds the property that deficiency is felt more keenly than comparable excess (now more commonly known as loss aversion), long considered central to matters of justice (Brown, 1986: 78; Homans, 1961: 75–76; Wagner and Berger, 1985: 719); for example, if the just reward is 10, an actual reward of 5 will have a larger absolute magnitude than an actual reward of 15 (−0.693 vs 0.405). These properties were discussed when the justice evaluation function was introduced (Jasso, 1978: 1401–1403, 1416–1417) and subsequently by Berger (1992: 8) and Wagner and Berger (1985: 719), among others.

Inequality, Poverty, and the Sense of Justice. To begin, that is, before wandering into Danto territory and obtaining predictions from which the justice in the assumptions (via the justice evaluation function) has seemingly vanished, consider a property that is close to the original Popperian guess. Readers who have looked hard at Atkinson's inequality measure in **Table 1** and the justice evaluation function in (7) will not be surprised to learn that the average of the justice evaluations can be expressed as the log of the ratio of the geometric means of the actual rewards A and the just rewards C and thence decomposed into injustice due to inequality and injustice due to poverty:

$$E(J) = ln\left[\frac{E(A)}{E(C)}\right] + ln\left[\frac{1 - I(A)}{1 - I(C)}\right]$$

where E denotes expected value and I denotes Atkinson's inequality (Jasso, 1999: 148–150). Equation 8 also yields further properties; for example, as inequality in the actual rewards increases, the average of the justice evaluations goes deep into underreward territory; and as variability in the just rewards increases, the average of the justice evaluations goes deep into overreward territory (another effect of diversity of thought).

Inequality Effects on Conflict between Subgroups. Consider a group or population in which everyone values income, and the members fall into two subgroups, such as persons from two races or ethnicities or of both genders. In the simplest case, the two subgroups are nonoverlapping on the valued good; for example, the poorest person in the advantaged subgroup is richer than the richest person in the disadvantaged subgroup. The group is characterized by the justice evaluations of its members and each of the two subgroups is characterized by the average of the justice evaluations among its members. Using the Newtonian idea that mass accumulates at a point, conflict between the two subgroups is represented by the absolute difference between the averages of the subgroup-specific justice evaluations. It is then straightforward to analyze the predicted conflict severity (Jasso, 1993).

The theoretical results in the income case indicate that the effect of the proportions in the two subgroups differs across distributional forms for income but that inequality always intensifies the conflict. If income is Pareto, conflict increases as the proportion in the bottom subgroup increases; if income is power-function, conflict decreases as the proportion in the bottom subgroup increases; and if income is lognormal, conflict first decreases, then increases, as the proportion in the bottom subgroup increases, reaching its nadir when the two subgroups are of equal size. As noted, regardless of the variate-specific pattern of the effect of the proportions in the two subgroups on conflict, conflict always increases with income inequality. **Figure 4** provides visualization of these theoretical predictions.

Inequality Effects on the Public Benefit of Religious Institutions. Suppose a subset of the population withdraw from society to

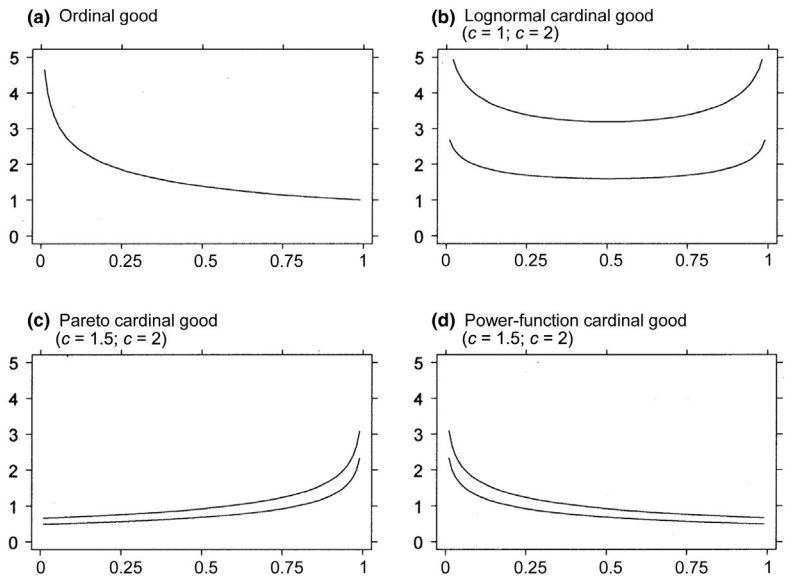

Figure 4 Conflict severity as a function of the proportion in the bottom subgroup, by valued good and cardinal-good's distributional form. In the three cardinal-good societies, the greater the inequality in the good's distribution, the greater the conflict severity (and the higher the vertical placement of the graph). Source: Jasso, Guillermina, 1993. Analyzing conflict severity: predictions of distributive-justice theory for the two-subgroup case. Social Justice Research 6, 372.

a cloister and take a vow of poverty, so that they assume the lowest places in the society's income hierarchy. The benefit to society of the religious institution may be represented by the difference between the precloister average of the justice evaluations and the postcloister average among the noncloistered. The public benefit then depends jointly on the proportion cloistered and income inequality. In every variate investigated, as inequality increases, the public benefit of religious institutions increases. This result suggests that religious institutions are more valuable to the social life when inequality is high, and as well that vocation rates will be higher in high-inequality societies. The model also suggests that the rise of mendicant institutions early in the thirteenth century – the Franciscans and Dominicans – was a response to a switch from valuing ordinal goods, like birth and nobility, to valuing cardinal goods, like money (Jasso, 1999: 165).

Effects of Husbands' and Wives' Wage Inequality on Marital Cohesion. Suppose that marital cohesion depends on the difference between the two spouses' justice evaluations. It is then straightforward to derive several predictions about marital wage dynamics and the further prediction at a higher level of analysis that in societies in which husbands earn more than their wives, marital cohesion increases with husbands' wage inequality and decreases with wives' wage inequality, and conversely (Jasso, 1999: 164). **Table 4** provides a handy visualization of these predicted effects.

Preventing Inequality Effects among a Family's Children. Justice theory not only yields predictions about inequality, it also yields predictions about preventing or mitigating inequality effects. One such prediction states that parents of two or more children who do not all have the same birthday will, to spare them pain, spend more of their toy budget at an annual giftgiving occasion than at the children's birthdays (Jasso, 1999: 164). A further implication is that keeping the sibship in harmony being a universal parental challenge, societies will invent an annual giftgiving occasion. In some societies, it will be New Year's Day, in others the Sixth of January, in still others Christmas or Chanukah or Eid.

Empirical Inequality Analysis

It could be said that the whole of sociology is about inequality. And thus this *Encyclopedia* contains dozens of articles on inequality, most of them empirical. This article will not attempt to summarize them. Instead, we invite readers to read the two other articles that have inequality in their title – Income Inequality; Inequality, Social – and to search for the larger set of articles whose keywords include inequality and the even larger set which in one way or another address inequality matters. Concomitantly, we invite readers to use the tools in this article to classify the inequality matters discussed and perhaps gain deeper understanding. Do the articles treat inequality between persons or inequality between subgroups? Or, at a larger scale, inequality between larger social entities like countries or inequality between subsets of those larger social entities? What are the quantitative characteristics in which inequality is explored or assessed? What are the qualitative characteristics which define the populations and subsets?

As for empirical assessment of the theoretical predictions discussed above, the testing task is in its infancy. But already there are early results. Bellou (2013) reports support for the proposition that as inequality increases in the male wage distribution, the divorce rate declines, consistent with the prediction linking husbands' wage inequality and marital cohesion. Similarly, the recent increase in vocations to the priesthood in the United States is consistent with the predictions on the public benefit of religious institutions. Finally, the perennial surge in US toy sales in the fourth quarter is consistent with the prediction that parents, in order to prevent inequality effects among their children, will spend more of their toy budget at an annual giftgiving occasion rather than at the children's birthdays.

Table 4 Effects on marital cohesion of husbands' and wives' wage inequality

	$X_H > X_W$	$X_W > X_H$
Wives' wage inequality	Increases	Decreases
Husbands' wage inequality	Decreases	Increases

Bibliography

Allen, Roy G.D., 1938. Mathematical Analysis for Economists. St. Martins Press, New York.

Aquinas, St Thomas, [1265–1274]1952. Summa Theologica. Translated by Fathers of the English Dominican Province. Britannica, Chicago.

Aristotle, [384–322 BCE] 1952. The Works of Aristotle, 2 vol. (W.D. Ross, Trans.) Britannica Press, Chicago.

Atkinson, Anthony B., 1970. On the measurement of inequality. Journal of Economic Theory 2, 244–263.

Becker, Gary S., 1991. A Treatise on the Family, enlarged ed. Harvard University Press, Cambridge, MA.

Bellou, Andriana, 2013. Male Wage Inequality and Marital Dissolution: Is There a Link? IZA Discussion Paper No. 7331.

Berger, Joseph, 1992. Expectations, theory, and group processes. Social Psychology Quarterly 55, 3–11. Based on 1991 Cooley-Mead Award address.

Berger, Joseph, Cohen, Bernard P., Zelditch Jr., Morris, 1966. Status characteristics and expectation states. In: Berger, Joseph, Zelditch Jr., Morris, Anderson, Bo (Eds.), Sociological Theories in Progress, vol. 1. Houghton Mifflin, Boston, pp. 29–46.

Berger, Joseph, Zelditch, Morris, Anderson, Bo, Cohen, Bernard P., 1972. Structural aspects of distributive justice: a status-value formulation. In: Berger, Joseph, Zelditch Jr., Morris, Anderson, Bo (Eds.), Sociological Theories in Progress, vol. 2. Houghton Mifflin, Boston, pp. 119–246.

Blau, Peter M., 1974. Presidential address: parameters of social structure. American Sociological Review 39, 615–635.

Brown, Roger, 1986. Social Psychology: The, second ed. Free Press, New York, NY.

Cowell, Frank A., 1977. Measuring Inequality: Techniques for the Social Sciences. Wiley Halsted, New York, NY.

Dagum, Camilo, 1983. Income inequality measures. In: Kotz, Samuel, Johnson, Norman L., Read, Campbell B. (Eds.), Encyclopedia of Statistical Sciences, vol. 4. Wiley, New York, pp. 34–40.

Danto, Arthur C., 1967. Philosophy of science, problems of. In: Edwards, Paul (Ed.), Encyclopedia of Philosophy, vol. 6. Macmillan, New York, pp. 296–300.

Firebaugh, Glenn, 1999. Empirics of world income inequality. American Journal of Sociology 104, 1597–1630.

Gini, Corrado, 1914. Sulla Missura della Concentrazione e della Variabilità dei Caratteri. Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti 73, 1203–1248.

Gregory the Great, St., [*c*.579–604]1849. Patrologiae Cursus Completus. Series Latina, vol. 75–79. In: Migne, J.P. (Ed.).

Homans, George C., 1961. Social Behavior: Its Elementary Forms. Harcourt, Brace and Company, New York.

John Chrysostom, St., [*c*.386–407]1860. Patrologiae Cursus Completus. Series Graeca, vol. 47–64. In: Migne, J.P. (Ed.).

Jasso, Guillermina, 1978. On the Justice of earnings: a new specification of the justice evaluation function. American Journal of Sociology 83, 1398–1419.
Jasso, Guillermina, 1993. Analyzing conflict severity: predictions of distributive-justice theory for the two-subgroup case. Social Justice Research 6, 357–382.
Jasso, Guillermina, 1999. How much injustice is there in the world? Two new justice indexes. American Sociological Review 64, 133–168.
Jasso, Guillermina, 2009. A new model of wage determination and wage inequality. Rationality and Society 21, 113–168.
Jasso, Guillermina, 2011. Plato: seven sociological ideas for the happy life. In: Edling, Christofer, Rydgren, Jens (Eds.), Sociological Insights of Great Thinkers: Sociology through Literature, Philosophy, and Science. Praeger, Santa Barbara, CA, pp. 31–44.
Jasso, Guillermina, Kotz, Samuel, 2007. A new continuous distribution and two new families of distributions based on the exponential. Statistica Neerlandica 61, 305–328.
Jasso, Guillermina, Kotz, Samuel, 2008. Two types of inequality: inequality between persons and inequality between subgroups. Sociological Methods and Research 37, 31–74.
Jencks, Christopher, Smith, Marshall, Acland, Henry, Jo Bane, Mary, Cohen, David, Gintis, Herbert, Heyns, Barbara, Michelson, Stephan, 1972. Inequality: A Reassessment of the Effect of Family and Schooling in America. Basic Books, New York, NY.
Jerome, St., [374–420]1845–1846. Patrologiae Cursus Completus. Series Latina, vol. 22–30. In: Migne, J.P. (Ed.). Britannica, Chicago.
Johnson, Norman L., Samuel Kotz, Balakrishnan, N. 1994(1995). Continuous Univariate Distributions, vol. 1 (2), second ed. Wiley, New York, NY.
Kleiber, Christian, Kotz, Samuel, 2003. Statistical Size Distributions in Economics and Actuarial Sciences. Wiley, Hoboken, NJ.
Liao, Tim Futing, 2006. Measuring and analyzing class inequality with the Gini index informed by model-based clustering. Sociological Methodology 36, 201–224.
Mill, John Stuart, [1869]1997. The Subjection of Women. Dover Thrift Editions, Mineola, NY.
Pareto, Vilfredo, 1897. Cours DEconomie Politique. Rouge, Lausanne, Switzerland.
Pen, Jan. 1971. Income Distribution. Allen Lane, London, UK.
Plato., [*c.*428–348/7 BCE]1952. The Dialogues of Plato. (Benjamin Jowett, Trans.).
Popper, Karl R., 1963. Conjectures and Refutations: The Growth of Scientific Knowledge. Basic Books, New York.
Rawls, John, 1971. A Theory of Justice. Harvard University Press, Cambridge, Massachusetts.
Rousseau, Jean Jacques, [1755]1952. A Discourse on the Origin of Inequality (G.D.H. Cole, Trans.). In: Rousseau, (Ed.). Britannica, Chicago, pp. 323–366.
Theil, Henri, 1967. Economics and Information Theory. North-Holland, Amsterdam.
Wagner, David, Berger, Joseph, 1985. Do sociological theories grow? American Journal of Sociology 90, 697–728.

Inequality: Comparative Aspects

Gerald D Berreman[†], University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7377–7382, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article offers a panoramic overview of equality and inequality in human societies throughout history, from an anthropological perspective. Discussing power, status, and economy as major dimensions, it covers a wide range of societal forms, from small-scale groupings of foragers to contemporary large-scale societies, and it also discusses race and caste, colonialism, gender, and age as they have related to social inequality. Types of authority such as big men and chiefs are also noted.

Understanding social inequality requires that it be examined comparatively through time (developmentally) and space (cross-culturally): in prehistoric, historic, and contemporary societies; among those described as 'egalitarian,' kin-ranked, and stratified; with foraging, horticultural, agro-pastoralist, and agro-industrial economies; with diverse types of social and political organization; in all habitable regions of the world; with systems of inequality based on principles of kinship, social roles, sex/gender, age, class, status, power, ethnicity, 'race,' caste, estate, and colonial hegemony.

An analytical framework based on key variables is presented. Because societies continually change, typological pigeonholes and evolutionary 'stages' are avoided, whereas nodes and junctures are identified along many paths, from relatively small, homogeneous, egalitarian societies to large, complex, stratified ones. Recent archeological and ethnographical research on transitions informs this discussion. As the analysis turns to comparison of inequalities among stratified societies, it moves from developmental issues to institutional ones – to the contrasting and interacting principles that underlie such systems and the practices they entail. It concludes with reactions generated among those who experience these systems from above and below, and the role of ideology in their consequences.

Inequality is defined here as sanctioned differential access to valued things, experiences, and conditions in human societies. Comparative aspects are similarities and differences found among systems of inequality as a result of comparing societies in all of their variety: through time (developmentally) and space (cross-culturally), in societies small and large, homogeneous and complex, prehistoric, historical, and contemporary, throughout the world.

Social Inequality

Social inequality has been debated from the times of Aristotle and Plato to those of Rousseau, Karl Marx, Max Weber, and innumerable contemporary authors (Beteille, 1987). The classic triumvirate, 'class' (economic order), 'status' (honor), and 'party' (power) are widely cited as its essential dimensions, with power – the ability to prevail over others – as the virtual independent variable. Powerful people are privileged; powerless people are rarely so, and then only under sponsorship of the powerful. Their perspectives on inequality reflect their places in the system: those who look from below see and experience a different world than those whose gaze and experience it from above, and the latter are the ones in a position to define and explain the system, as well as to enforce it.

This article draws on understandings of social inequality obtained, primarily by anthropologists, through cross-cultural and cross-temporal research focused on peoples who have been otherwise largely ignored: those commonly labeled, if noticed at all, as primitive, aboriginal, tribal, native, indigenous, heathen, nomadic, peasant, untouchable, colonial, minority, ethnic, immigrant, and refugee – in short, those rendered vulnerable for whatever reason, or unreason.

Egalitarian (Unranked) Societies

Dominance and differential evaluation exist in every society, but in some – those often loosely described as 'egalitarian' – they are limited to the elemental roles found in all societies: those of family, sex/gender, age/seniority, specific activities or skills, and cherished or stigmatized personal characteristics. If additional inequalities occur, they are more symbolic than material or political. Inequalities in such societies have received increasing attention from ethnographers and archaeologists in recent years (Flanagan, 1989).

These societies are small, mobile bands of kinsmen, with households as the basic social and economic units. Theirs are subsistence economies derived by foraging wild species, without privileged or curtailed access to resources or the means to obtain them. They share their economic pursuits and rewards without expectation of specific returns in what has been termed 'generalized reciprocity' (Sahlins, 1968). Requisite knowledge and skills are present in every household, extrafamilial authority or occupational roles are rare or absent, and sexual division of labor is complementary, with little or no differential evaluation.

Assertively Egalitarian Societies

Even in 'egalitarian' societies, equality of power, possessions, and prestige is in jeopardy. The Hadza of Tanzania are foragers

[†] Deceased.

who consume the products of a successful foray before embarking on another. Woodburn (1982) describes this as an 'immediate-return system.' He characterizes societies with such economies as being 'assertively egalitarian,' in that their members characteristically resist and sanction attempts to exert authority or to consume resources without sharing.

Relatively Egalitarian Societies

By contrast, foraging economies in which long-term collective investments of resources are employed (e.g., construction and maintenance of corrals for game drives), Woodburn terms 'delayed-return systems': the return on investment of labor is delayed rather than immediate. Although they are egalitarian to an extent found among few non-foragers, and are therefore 'relatively egalitarian,' they entail authority and accumulation untenable in immediate-return societies as a result of the need for coordination of delayed-return activities and preservation of products for use during periods of delay.

All societies whose economies and organization are more complex than those of 'assertively egalitarian' foragers have delayed-return systems. Those with economies and social systems more complex than those of 'relatively egalitarian' foragers are inegalitarian, or 'rank, societies' (Fried, 1967).

Inegalitarian (Rank) Societies

Inegalitarian societies are those wherein members are individually, or more often collectively, hierarchically ordered in their access to valued things, conditions, and experiences. All societies discussed here are, in this sense, inegalitarian. This presentation moves from those of relatively 'small scale' (small in numbers and relatively homogeneous) to those of increasingly 'large scale' (large in population and socially complex). This sequence reflects and extends the strong correlation that exists between scale and inequality.

Affluent Foragers and Farmers: Transition to Assertive Inequality

Some cultural anthropologists and many archaeologists have turned to supplementing largely static extant typologies with analyses of transitions, especially those from unranked to, and within, rank societies (Price and Gebauer, 1995; Kirch and Marshall, 1992).

Rank society emerges when circumstances permit some members to acquire privileged or, in the extreme, exclusive access to essential resources. This has traditionally been attributed to the invention of domestication – primarily of plant species, hence agriculture – and the resulting opportunity for some to control land and command the labor of landless others to till and guard it. Similar conditions obtain for pastoralists' control of their herds.

It has long been recognized that some foraging societies share with early farmers the inequalities characteristic of agricultural and pastoral peoples (Koyama and Thomas, 1981). The most widely cited example is the peoples of the northwestern coast of the United States, who developed maritime economies devoid of domesticated species, combined with a complex and highly inegalitarian social structure (Lightfoot, 1993). This seeming anomaly is explained by the annual runs of salmon which, accompanied by technologies of preservation and storage, provide a food supply as predictable and productive as any domestic crop or herd. Inequality results from the opportunities for some to control the best fishing sites and equipment and to mobilize personnel to exploit them. This system entails ranked kin groups (clans) and their chiefs. The latter accumulate goods provided by their clansmen for subsequent ceremonial, status-enhancing distributions to rival chiefs and clans – the storied potlatch.

Other examples of such systems are found primarily in horticultural societies, which employ two distinct kinds of leadership role and social organization: big man and chief (Sahlins, 1968).

Big Men, Big Women

The simpler of the two structures is the 'big man' (rarely 'big woman') society. The big man's role is an informal, acquired, or achieved status. Through personal charisma and persuasion, he recruits a volunteer following whose number contribute valued goods to a fund that he will ceremonially distribute to a competing big man and his followers (as in a potlatch). Recipients thereby incur a debt to be paid off, and if possible exceeded, in a reciprocal distribution. The magnitude of the distribution is the measure of a big man's stature, while his followers receive credit (and when next he is on the receiving end, goods) proportional to their contributions.

Chiefs

A chief differs from a big man in that his is an office ascribed to him as a result of his paramount position in his clan. Fellow clansmen are his obligatory supporters and he their hereditary headman. As chief, he must behave in a chiefly manner, ever responsible and generous to his clan. His behavior follows from his role; the opposite of the big man, whose role follows from his behavior.

Big man organization may evolve into chieftainship when big men find that kinsmen are more reliable supporters than others and therefore preferentially, and eventually exclusively, recruit them. As kinship becomes the de facto principle of organization, the followers of a charismatic big man may become in effect, and ultimately in fact, clansmen of their paramount kinsman, and he their chief.

Chiefs to Kings, Kinship to Class: Transition to Complex Society

The offices of both chiefs and kings are ascribed on the basis of kinship with their predecessors. But they differ fundamentally: the relevance of kinship for the king stops with his coronation; for the chief, investiture is only the beginning. The king becomes sovereign over all within his domain, beholden to no one; the chief is paramount only among his kinsmen and, as kinsman, is as beholden to them as they are to him. The king rules, accumulates, and consumes; the chief proposes, coordinates, accumulates, and redistributes.

Kingship thus represents the demise of kinship as the basis for social structure, replaced by economic stratification: ownership of productive resources, including land, territory as the dimension of governance, authority and force as its method, and depersonalization as a result of dramatic increases in scale. There appear for the first time, ranked economic classes and cultural categories.

Because reciprocity is the economic hallmark of unranked society, and accumulation for conspicuous redistribution is the hallmark of the big man's and the chief's organization, accumulation for conspicuous consumption is a hallmark of kingship and the state.

Stratification

Contemporary nation states are socio-economically stratified in a variety of complex ways, their populations distributed along continuums of differential access to necessities and rewards. These distributions are often perceived as comprising distinctive, ranked categories – social strata – of people similarly positioned.

An issue to which comparative analysis has made major contributions is the nature and function of stratification: Does it primarily serve to select the best qualified for the most important roles in society by offering them the highest rewards? Or does it serve mainly to withhold opportunities from vulnerable segments of society, thus assuring that there will be people with no choice but to perform cheaply and tractably the necessary tasks shunned by others? Caste in India, and US race/ethnic relations, are examples that support the latter interpretation (Berreman, 1960), as is the status of Nepalese and Brazilian immigrant labor in Japan (Yamanaka, 2000).

Dimensions of Stratification

Power is key to stratification, but class and status are closely linked to it; the three vary together. Stratification systems differ in significant ways, the analysis of which is facilitated by comparison along three axes of inequality. One defines the nature of the ranking, contrasting status strata ('honor and privilege') at one pole and class strata (the 'economic order') at the other. A second axis describes the locus of characteristics that determine an individual's status, contrasting those regarded as intrinsic to individuals (ascribed) with those regarded as extrinsic (acquired or achieved). A third axis addresses the composition of strata: whether they comprise named, organized, acting groups or analytically discernible but unorganized categories. Variations along these axes, together with the cultural explanations that rationalize each system, constitute the significant differences among them.

Acquired Economic Status: Categories as 'Classes,' Groups as 'Sodalities'

Acquired status is based on possessions and privileges, its inequalities defined by economic factors (income, net worth) and their accompanying lifestyles (patterns of consumption, education, occupation). These are classes in which it is useful to contrast: (1) those composing objectively discernible categories, useful to an analytical observer but amorphous, unorganized, and not necessarily recognized even by those who compose them; and (2) those who compose groups that are publicly recognized, members who are aware of their shared situation and its sources, are organized, and are acting in their common interests. Marx recognized this distinction when he described the first as 'groups in themselves' and the second as 'groups for themselves.'

Ranked economic *categories* are here designated simply classes. Ranked, economy-based groups, such as labor unions and professional associations, are designated associations or sodalities.

Feudalism and Estates: Groups with Ascribed Economic Status

Feudal systems are complex agricultural societies comprising hierarchical relations among landed nobility, various levels of commoners, and rural peasants. Because patron–client relationships with respect to land are a defining characteristic, such systems fall well within the 'economic order.' They therefore constitute systems of class stratification (Stavenhagen, 1975). But these are classes with a difference, for they are explicitly ranked, with specified rights and obligations. As long as a person belongs to a given category, he or she shares its rank. However, mechanisms exist – usually economic – by which an individual may, with difficulty, move to another category, thereby acquiring its rank. Status is thus intrinsic to the category but not to the individual. Such named, bounded, and jurally defined categories can be labeled estates so as to distinguish them from classes.

Ascribed Honor: Caste, 'Race,' and Ethnicity

Ascribed statuses of honor are distinguished by the fact that rank is assigned on the basis of membership in differentially privileged and esteemed groups or categories rather than on the basis of activities or possessions. They are also distinguished by the fact that, as a result of ascription, those who share them constitute publicly recognized, self-acknowledged groups. Often they are culturally distinct, cohesive – sometimes corporate – groups *for* themselves. Weber called them status groups. Four types will be contrasted below: 'caste/race,' 'ethnicity,' 'sex/gender,' and age/seniority.

Caste/Race

Status assigned to culturally distinct, socially segregated categories of people on the basis of shared and differentially evaluated characteristics claimed to be innate, is unalterable. 'Race,' as it is socially constructed in the United States, and 'caste' (*jati*), as it is constructed in India, are classic examples. Castes are localized, internally organized, endogamous, and often corporate groups ranked according to conceptions of ritual purity. Most low-ranking castes lay claim to higher status than that accorded them – claims that are routinely ignored or brutally quashed by high castes. Attempts to validate their claims entail status emulation: adoption of high-caste rituals and behaviors, a process familiar in Indian social science and known as *Sanskritization* (Srinivas, 1966). They are successful with exceeding rarity, and then only if backed by unprecedented power.

Stigmatized 'races' in the United States differ from the low castes of India in that they are nationwide social categories rather than localized groups, but they share with low castes their birth-ascribed subjection to disparagement, discrimination, and exploitation, rationalized by accusations of intrinsic worthlessness. Similarly disparaged 'outcaste' groups have included the *Burakumin* of Japan and the *Hutu* of Rwanda and Burundi. All respond to their oppression with resentment and resistance. The only consensus on inequality in these systems is the pragmatic one of who has the power and under what circumstances and with what effects it will be used. When this consensus is breached, the systems are often vigorously challenged. These fundamental similarities are here deemed to be sufficiently significant to justify the combined term 'caste/race,' except where their differences are at issue.

Ethnicity

Ethnicity and *ethnic group* are terms for ascribed identity bestowed on groups or categories that are defined by their distinctive cultural attributes, such as language or religion. They are commonly, but not necessarily, ranked within their societies. If ranked, their status is not wholly ascriptive for, although the ethnic group's ranking is ascribed, the defining attributes of its individual members are acquired through shared socialization – learning – and what is learned can be rejected or forgotten. Therefore, as in the case of estates, rank is intrinsic to the group but extrinsic to individuals, so that individual mobility within the system is possible. If circumstances permit, people can acquire or construct other ethnicities, other statuses (Barth, 1969).

Sex/Gender

Sex is usually a birth-ascribed and immutable category of difference which, in all societies, serves as a basis for division of labor and, in most complex societies, is a basis for inequalities. Gender is the social construction of its nature and significance. Many societies define other genders as well, among them the *berdache*, the 'men–women' of several Native American societies and the 'manly-hearted women' of some of them; Polynesian *fa'a fahini* ('in the manner of women'), and India's 'third sex,' the *hijara*. These, and lesbian, gay, bisexual, and transsexual genders of Euro–US and other complex societies, are but some of the diverse gender roles worldwide. Some are stigmatized, but many are not and occupy conventional or exceptional roles in their societies.

The relative status of the sexes in human societies has been long debated. Universal female subordination is often taken for granted. To challenge this, one has only to look across the full range of human societies, including especially those that resemble the societies in which 99.9% of our ancestors, through time, have lived (Leacock, 1978). Among more complex and inegalitarian societies, one must consider those whose systems of descent, inheritance, and postmarital residence follow the female line. The highest incidence of such systems are found in Southeast Asia and Melanesia, among horticultural peoples in which women own the land, do the lion's share of cultivation of staple crops, and enjoy a high to superior degree of autonomy, authority, and respect.

Gender status correlates positively not only with the prominence of its role in descent, residence, and horticulture, but also with its contribution to the household or kin group economy, whether foraging, horticulture, agriculture, pastoralism, wage labor, or profession. In complex societies, it correlates most highly with the economic contribution made from outside of the household, or 'domestic sphere' (Caulfield, 1981).

In discussing gender inequalities, it must not be overlooked that, although every society has distinctive female and male cultures, the genders are largely socialized in the intimacy of family and kin of both sexes, sharing class, ethnic, caste/race statuses, and environments. They are therefore familiar with one another to an extent that people of other ranked identities are not. However, the extent to which such experiences contribute to respect and empathy cannot be predicted from the fact of contact. As studies of US 'race' relations have demonstrated, contact does not necessarily reduce prejudice and may exacerbate it by seeming to confirm stereotypes, unless it occurs in status-equal circumstances.

US women have long been discriminated against, while sharing class and ethnic status with male family members. In India, women share the ritual and economic status of their male caste fellows, while being subordinated to males. Circumscription of women in activities outside the household is an important symbol of high caste rank. *Sanskritization*, caste-status striving, entails increasing subordination of women (Berreman, 1993).

Age/Seniority

Age categories and cultures, like those of gender, are ascribed, socially constructed, familiar to all, and highly variable. As a basis for inequality, age and seniority are at their minimum in unranked societies. Societies organized in lineal kin groups or clans commonly place great importance on seniority, and the first born in a senior (most often male) lineage becomes the paramount.

A number of societies, the most thoroughly reported of which are East African cattle herders, are organized into ranked, age-based groups, more often male than female, whose members share distinctive ritual, economic, military, or other privileges and responsibilities.

Comparative studies of age in other kinds of societies include Mead's classic *Coming of Age in Samoa* (1928), which challenged the universality of adolescence as a period of trauma, and Henry's *Personality and Aging* (1973), which argued that individuals are regarded as persons to the degree that they possess symbols of attachment to their social system, and that in the United States such symbols (gainful employment, license to drive) are often withdrawn from the elderly, leading to the experience of diminished personhood and resulting in social suffering, hostility, and depression. This is precisely the kind and consequence of denial of full humanity inflicted by racism, casteism, ethnic, and gender discrimination.

Colonialism: International Hegemony

Relationships between colonizer and colonized are among the most powerful of the world's structures of domination and exploitation (Asad, 1973). In its classic context, colonialism

has referred to nations and peoples annexed for their natural, manufactured, or human resources by one of the 'great', usually European, maritime powers. Subsequently, the concept has been directed to other contexts, as in 'third-world colonialism' (hegemony over developing nations, often by their erstwhile colonizers), and 'internal colonialism' (domination of minorities within developed nations, e.g., African Americans, Palestinians in Israel).

By analogy, 'academic colonialism' refers to the dominant role well-funded researchers from 'advanced' nations play in 'developing' nations, while local scholars are scantily funded, if at all. It is most damaging and resented when indigenous scholars are evaluated by the standards and fashions of colonial countries, ratified by foreign degrees, and by publication in foreign journals, in colonial languages. Social and behavioral scientists, perhaps more often than others, may find themselves strangers in their own academies and virtual pariahs in world scholarship (Saberwal et al., 1967).

Full Circle: Modern Egalitarian Societies?

Egalitarian social systems require egalitarian economies. Both have been on the agenda of utopian communities and ideal societies studied by social and behavioral scientists: Soviet and Chinese communes, Israeli kibbutzim, socialist societies, and anarchist communities (Barclay, 1982), as well as the American countercultural communities of the 1960s–70s and various egalitarian cults. Many live in small, self-contained communities, combining interdependence and individual/familial autonomy bound by ties of kinship (usually fictive) and sharing ideology and communal activities, while avoiding insofar as possible, participation in the complex society, economy, and politics of the world system. They often pursue subsistence economies through cooperation, sharing, and reciprocity; their production is for use rather than profit; they abjure private property, accumulation of wealth, hierarchies of authority or privilege, and status symbols. That is, they reproduce the social and economic practices of the small-scale, 'assertively' or 'relatively' egalitarian societies with which human societies began and for so very long survived. Prehistoric, recent, and current exemplars have virtually all succumbed to the appetites, blandishments, and requirements of stratified, state-organized producer–consumer society.

Bodley closed *Victims of Progress* (1975) with the observation that small-scale, egalitarian culture 'is gone – after a very long and successful tenure,' and it seems that the presently dominant, stratified "culture of consumption … too, will disappear … but after a very brief and preposterous career."

In the early twenty-first century, that remains to be seen. Recent times have seen a decline in a number of the varieties of human inequality discussed here. Colonialism of the conventional historical–political type is mostly gone (although it may remain in other forms); as a matter of human rights, the tendency in much of the world has been to try to abolish at least official discrimination on the basis of race, ethnicity, gender, and age. With increasing global interconnectedness, on the other hand, large-scale transnational migration and refugee streams from crisis regions have led to the formation of large vulnerable populations whose links to states, markets, and local communities are extremely tenuous. At the same time, dominant market conditions have tended to increase class inequality in much of the world, including North America and Europe. These conditions appear to be volatile, however, and it is not obvious that contemporary modes or degrees of inequality will be permanent.

See also: Caste; Discrimination and the Law; Discrimination, Economics of; Discrimination: Racial; Equality and Inequality: Legal Aspects; Equality of Opportunity; Equality: Philosophical Aspects; Income Inequality; Inequality, Social; Migration: Anthropological Perspectives; Social Inequality and Schooling; Social Inequality in History (Stratification and Classes); Social Stratification; Social Work and Refugees.

Bibliography

Asad, T. (Ed.), 1973. Anthropology and the Colonial Encounter. Ithaca Press, London.
Barclay, H., 1982. People without Government: An Anthropology of Anarchy. Kahn and Averill, London.
Barth, F., 1969. Introduction. In: Barth, F. (Ed.), Ethnic Groups and Boundaries, The Social Organization of Culture Difference. Little Brown, Boston, pp. 9–38.
Berreman, G.D., 1960. Caste in India and the United States. American Journal of Sociology 66, 120–127.
Berreman, G.D., 1993. Sanskritization as female oppression in India. In: Miller, B.D. (Ed.), Sex and Gender Hierarchies. Cambridge University Press, Cambridge, UK, pp. 366–392.
Beteille, A., 1987. The Idea of Natural Inequality and Other Essays. Oxford University Press, New Delhi, India.
Bodley, J.H., 1975. Victims of Progress, first ed. Mayfield, Mountain View, CA, p. 174.
Caulfield, M.D., 1981. Equality, sex and mode of production. In: Berreman, G.D. (Ed.), Social Inequality: Comparative and Developmental Approaches. Academic Press, New York, pp. 201–219.
Flanagan, J.G., 1989. Hierarchy in simple egalitarian societies. Annual Review of Anthropology 18, 245–266.
Fried, M., 1967. The Evolution of Political Society. Columbia University Press, New York.
Henry, J., 1973. Personality and ageing with special reference to hospitals for the aged poor. In: On Sham, Vulnerability and Other Forms of Self-destruction. Vintage Books, New York, pp. 16–37.
Kirch, P.V., Marshall, S., 1992. Anahulu: The Anthropology of History in the Kingdom of Hawaii. University of Chicago Press, Chicago.
Affluent foragers. In: Koyama, S., Thomas, D.H. (Eds.), Senri Ethnological Studies 9. National Museum of Ethnology, Osaka, Japan.
Leacock, E., 1978. Women's status in egalitarian society, implications for social evolution. Current Anthropology 19, 247–275.
Lightfoot, K., 1993. Long-term developments in hunter–gatherer societies: recent perspectives from the coast of North America. Journal of Archaeological Research 1, 167–201.
Mead, M., 1928. Coming of Age in Samoa: A Psychological Study of Primitive Youth for Western Civilization. Morrow, New York.
Price, T.D., Gebauer, A.B., 1995. Last Hunters, First Farmers: New Perspectives on the Prehistoric Transition to Agriculture. School of American Research, Santa Fe, NM.
Saberwal, S., et al., 1967. Academic colonialism: a symposium on the influences which destroy intellectual independence. Seminar 112, 9–40.
Sahlins, M., 1968. Tribesmen. Prentice-Hall, Englewood Cliffs, NJ.
Srinivas, M.N., 1966. Social Change in Modern India. University of California Press, Berkeley, CA.
Stavenhagen, R., 1975. Social Classes in Agrarian Societies. Anchor Books, Garden City, NY.
Woodburn, J., 1982. Egalitarian societies. Man 17, 431–451.
Yamanaka, K., 2000. Nepalese labour migration to Japan: from global warriors to global workers. Ethnic and Racial Studies 23, 62–93.

Inequality Measurement

Frank A Cowell, London School of Economics and Political Science, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the core issues of inequality measurement and explains the relationship between underlying theoretical concepts and practical judgments. The main tools of distributional analysis are used to show the social welfare underpinnings of standard ranking principles and to derive families of inequality indices. Developments that focus on income differences and reference income levels are also examined.

Introduction

Inequality measurement is principally concerned with the comparison of statistical distributions of a specific economic or social indicator. The comparisons usually, but not necessarily, involve distributions among persons or households and involve quantitative methods. The motivation for considering inequality measurement in this way is both analytical and practical: the principal concepts reviewed in this article are of concern to theoretical and applied economists, to other social scientists and also to policy analysts. The subject touches on questions in several disciplines other than economics.

Inequality measurement in its modern form derives some of its intellectual heritage from statistics and is considered by some to be a branch of welfare economics. The subject is related to, but distinct from, the measurement of concentration, poverty, or relative deprivation. Inequality measurement in economics is perhaps most often applied to the study of income distributions but is also used in connection with the study of wealth, expenditure, and other indicators of well-being.

To introduce some of the main issues consider the example in Tables 1 and 2, taken from an easily accessible source and giving a convenient snapshot of what happened to the distribution of income in the United States over 40 years. From Table 1 it is clear that the bottom decile of the income distribution experienced a 12.8% growth over the period (in real terms) while the top decile grew by almost four times as much (48.7%). Table 2 describes what happened to the average incomes of particular *groups*. The average income of households in the bottom fifth of the distribution grew by just 8.5% over the 40 years while the average income of households in the top fifth grew by almost 60%. We return to the use of the concepts of quantiles and shares after introducing some of the technical equipment needed for analyzing income distributions.

This brief discussion suggests a substantial increase in inequality in the United States in the late twentieth century and early twenty-first centuries. But by how much did inequality increase? Can the intuitive method of inequality comparisons suggested in the example be made precise and interpreted within the context of standard economic analysis? The purpose of this article is to provide a succinct overview of the role played by economic theory and other abstract principles in this class of problem and how to make precise and to interpret inequality comparisons such as those suggested in the example.

Tables 1 and 2 also illustrate some of the practical issues that have to be resolved when implementing the principles of

Table 1 Quantile incomes and growth, United States 1971–2011

	q-*Quantile*		
q	*1971*	*2011*	*Growth*
10%	$10 637	$12 000	12.8%
20%	$18 818	$20 262	7.7%
50%	$44 707	$50 054	12.0%
80%	$75 271	$101 582	35.0%
90%	$96 564	$143 611	48.7%
95%	$119 532	$186 000	55.6%

Note: Columns two and three give the upper limit of the bottom 10%, 20%, … of the population. Incomes are in 2011 dollars; the income-receiving unit is the household.
Source: Reproduced from Navas-Walt, C., Proctor, B.D., Lee, C.H., 2012. U.S. Census Bureau, Current Population Reports, P60–243, Income, Poverty, and Health Insurance Coverage in the United States: 2011. U.S. Government Printing Office, Washington, DC, (Appendix Table A2). Online. Available at: www.census.gov/prod/2012pubs/p60-243.pdf (accessed 31.07.13.).

Table 2 Growth in average incomes for the five quintile groups and overall. United States, 1971–2011

	Average income		
Group	*1971*	*2011*	*Growth*
First	$10 360	$11 239	8.5%
Second	$27 381	$29 204	6.7%
Third	$44 392	$49 842	12.3%
Fourth	$63 113	$80 080	26.9%
Top	$111 830	$178 020	59.2%
Overall	$51 415	$69 677	35.5%

Note: Columns two, three give the average incomes of the bottom fifth, second fifth…. Incomes are in 2011 dollars; the income-receiving unit is the household.
Source: Reproduced from Navas-Walt, C., Proctor, B.D., Lee, C.H., 2012. US Census Bureau, Current Population Reports, P60–243, Income, Poverty, and Health Insurance Coverage in the United States: 2011. U.S. Government Printing Office, Washington, DC, (Appendix Table A2). Online. Available at: www.census.gov/prod/2012pubs/p60-243.pdf. (accessed 31.07.13.).

inequality measurement: Should the focus of analysis be on families, households, or individuals? How the income defined and what problems may the definition of income present for formal inequality analysis?

We deal first with the principles underlying these basic issues (Section Basics) and then we examine three theoretical approaches to the problem of inequality measurement (Sections Social Welfare and Inequality and Income Differences).

Basics

Components of the Problem

The framework adopted here is not the most general approach, but one that is suitable for setting out the key ideas. We begin by considering the basic building blocks and then show how to assemble the constituent parts.

Income and Income Distribution

At the center of the problem there is some entity, which we will call 'income,' but which in practice could be wealth, expenditure, or some other economic quantity, the distribution of which is of particular interest. It is important that this be cardinal (like height or weight) otherwise many of the concepts used below will not be well defined. There are cases where one wants to discuss inequality among distributions over ordinal variables such as self-reported health but these raise special problems, which are not discussed here – see Allison and Foster (2004).

Income is distributed among a number of income receivers, which we will refer to as 'persons,' although the income receiver in practice may be a family or household (we return to households below). Suppose that there is a known number of income receivers n and that person i has income x_i. The *income distribution* is then simply the vector

$$\mathbf{x} = (x_1, x_2, \ldots, x_n) \qquad [1]$$

The set of all possible income distributions X is a subset of n-dimensional real space. The nature of X is going to depend in practice upon the precise definition of 'income': is it logically possible to have a zero value of x_i, for example? Or a negative value? As a working assumption we will take it that X consists of all vectors (eqn [1]) such that $x_i \geq \underline{x}$ and leave open the specification of the lower bound $\underline{x}$ for particular instances of the inequality-measurement problem. Representations of the income distribution other than eqn [1] will appear later in the article.

Indices

The topic of inequality measurement suggests that there ought to be an inequality measure (or measures). A natural interpretation of an 'inequality measure' is as an index I that, given a particular income distribution $\mathbf{x}$, yields a number that is taken to be the amount of inequality exhibited by the distribution. In some respects, the index I behaves like other well-known summary statistics of distributions, such as the mean

$$\mu(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad [2]$$

and the variance

$$\mathrm{var}(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} [x_i - \mu(\mathbf{x})]^2 \qquad [3]$$

In fact a simple transformation of the variance can be used as an inequality index, known as the *coefficient of variation*:

$$I_{\mathrm{CV}}(\mathbf{x}) := \frac{\sqrt{\mathrm{var}(\mathbf{x})}}{\mu(\mathbf{x})} \qquad [4]$$

One of the most commonly used indices in practice is the *Gini coefficient* defined as

$$I_{\mathrm{Gini}}(\mathbf{x}) := \frac{1}{2n^2\mu(\mathbf{x})}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j| \qquad [5]$$

This has the natural interpretation of a normalized average of the average income differences between pairs of people in the population. The inequality measures eqns [2] and [5] have the advantage that they can be used also in cases where 'income' is negative (a situation that can often arise when analyzing wealth distributions).

There are many more inequality indices in common use, but rather than running through an exhaustive list, it is more useful to examine the principles that have usually been applied to construct indices. We do by considering a priori what constitutes a 'suitable' inequality measure, the issue addressed in Section Axioms.

Ranking and Dominance

An apparently more flexible interpretation of the idea of inequality measurement is the idea of an inequality *ranking*. This is a partial ordering that picks up the general flavor of the kind of comparisons that we suggested in the introduction; the partial ordering is typically captured by a simple representational tool. Consider three of these, each of which is defined for negative as well as positive values of income.

The first of these tools is *Pen's parade*, named after the famous parable introduced by Pen (1974, ch. 3), which is simply the inverse of the empirical distribution function. To depict it, let $x_{[i]}$ denote the ith smallest component in the vector eqn [1] – the ith smallest income. Then take the collection of points

$$\left(\frac{i}{n}, x_{[i]}\right), \quad i = 1, 2, \ldots, n \qquad [6]$$

From this simple definition, we can also introduce the idea of *dominance*. Take two distributions $\mathbf{x}'$ and $\mathbf{x}''$ in X, where $\mathbf{x}' = (x'_1, x'_2, \ldots, x'_n)$ and $\mathbf{x}'' = (x''_1, x''_2, \ldots, x''_n)$. If it is true that $x'_i > x''_i$ for all $i = 1, 2, \ldots, n$ then we say that $\mathbf{x}'$ strictly Parade-dominates $\mathbf{x}''$.

The resulting graph plots income quantiles against population proportions: $x_{[i]}$ is the quantile corresponding to the bottom q percent of the population where $q = 100\frac{i}{n}$. To illustrate the concept, we use the information in Table 1 to produce a graph that looks like Figure 1. In Pen's parable, we imagine the whole population (seen as individuals rather than households) arranged in order on the (0, 1) interval where each person's height has been altered in proportion to his/her income; the average-height income recipient in 1971 is located at position 0.58 in Figure 1 (in other words, at a point 58%

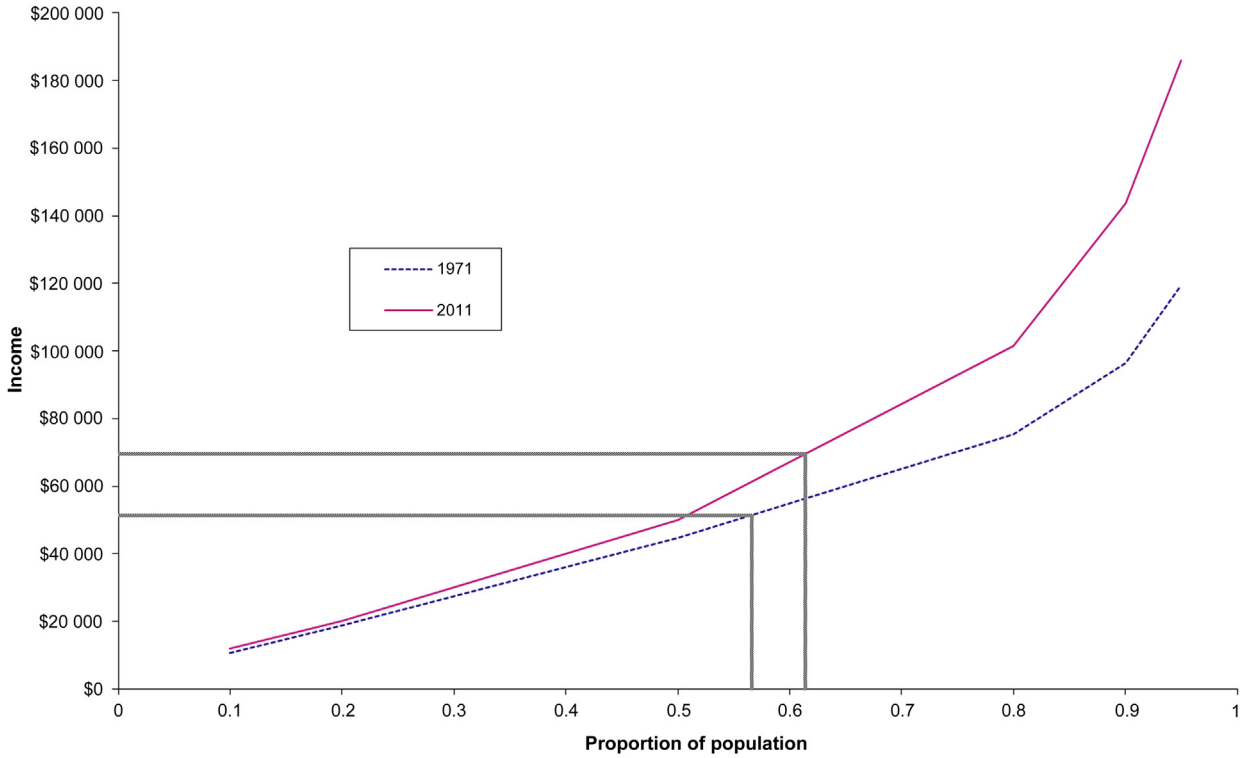

Figure 1 Parade diagram corresponding to **Table 1**.

along the horizontal axis the height of the Parade is exactly mean income), but in 2011 the average-height income recipient is located at position 0.61. Although the distribution of 2011 Parade-dominates the distribution in 1971, it is clear from Table 1 that overall the Parade shifted upward in an uneven fashion over the 40 years with the incomes of the very rich (95% quantile) growing more than four times faster than those of the poor (10% quantile); this shift suggests increased inequality over the period. However, by itself the Parade does not tell us much about inequality directly, although concepts closely related to it are widely used to characterize inequality comparisons. It is common to use *quantile ratios* for distributional comparisons; for example, the popular '90–10 ratio' is given by $x_{[k]}/x_{[j]}$, where j and k are, respectively, the smallest integers satisfying $j/n \geq 10\%$ and $k/n \geq 90\%$. In the example above, this ratio increased from 9.08 to 11.97. Furthermore, there is an important interpretation of the Parade that will be discussed in Section Ranking Distributions below.

For the second and third concepts, we use the $x_{[i]}$ to derive the normalized income cumulations; for any, $i = 1, 2, \ldots, a$ these are

$$c_i := \frac{1}{n}\sum_{j=1}^{i} x_{[j]} \qquad [7]$$

Then the *generalized Lorenz curve* (GLC) is given by the graph of

$$\left(\frac{i}{n}, c_i\right), \; i = 1, 2, \ldots, n \qquad [8]$$

Again we have a natural definition of dominance: for two distributions $\mathbf{x}'$ and $\mathbf{x}''$ in X if it is true that $c_i' > c_i''$ for all $i = 1, 2, \ldots, n$ then we say that $\mathbf{x}'$ strictly GLC-dominates $\mathbf{x}''$. For the example we used earlier, the GLC is illustrated in Figure 2, derived from Table 2. The definitions of Parade- and GLC-dominance can be extended to cases where the two distributions do not have the same number of incomes – this step makes use of the 'population principle' defined in Section Axioms.

The GLC plots the normalized income of the bottom $100q$ percent of the population against q and, although the 2011 distribution GLC-dominates 1971, it is clear that over the period the growth of these group averages was not evenly distributed – the higher was q, the higher was the growth over 1971–2011. (This is easily inferred from Table 2: for example, the average income of the top 20% grew almost seven times as fast as the average income of the bottom 20%.) Once again, although the GLC does not give information about inequality comparisons directly, there is an important welfare-economic interpretation (in Section Ranking Distributions). In addition, a small modification of the GLC yields one of the central concepts of distributional analysis. Dividing c_i in eqn [7] by the mean $\mu(\mathbf{x})$ gives the income share of the bottom $100\frac{i}{n}$ percent of the population. The graph of the (population–proportion, income–share) pairs

$$\left(\frac{i}{n}, \frac{c_i}{\mu(\mathbf{x})}\right), \; i = 1, 2, \ldots, n \qquad [9]$$

gives the *Lorenz curve*. Also, for two distributions $\mathbf{x}'$ and $\mathbf{x}''$, if it is true that $c_i'/\mu(\mathbf{x}') > c_i''/\mu(\mathbf{x}'')$ for all $i = 1, 2, \ldots, n-1$, then we say that $\mathbf{x}'$ strictly Lorenz-dominates $\mathbf{x}''$. In the case of the example using US data this is illustrated in Figure 3: the Lorenz curve plots the income share of the bottom $100q$ percent of the

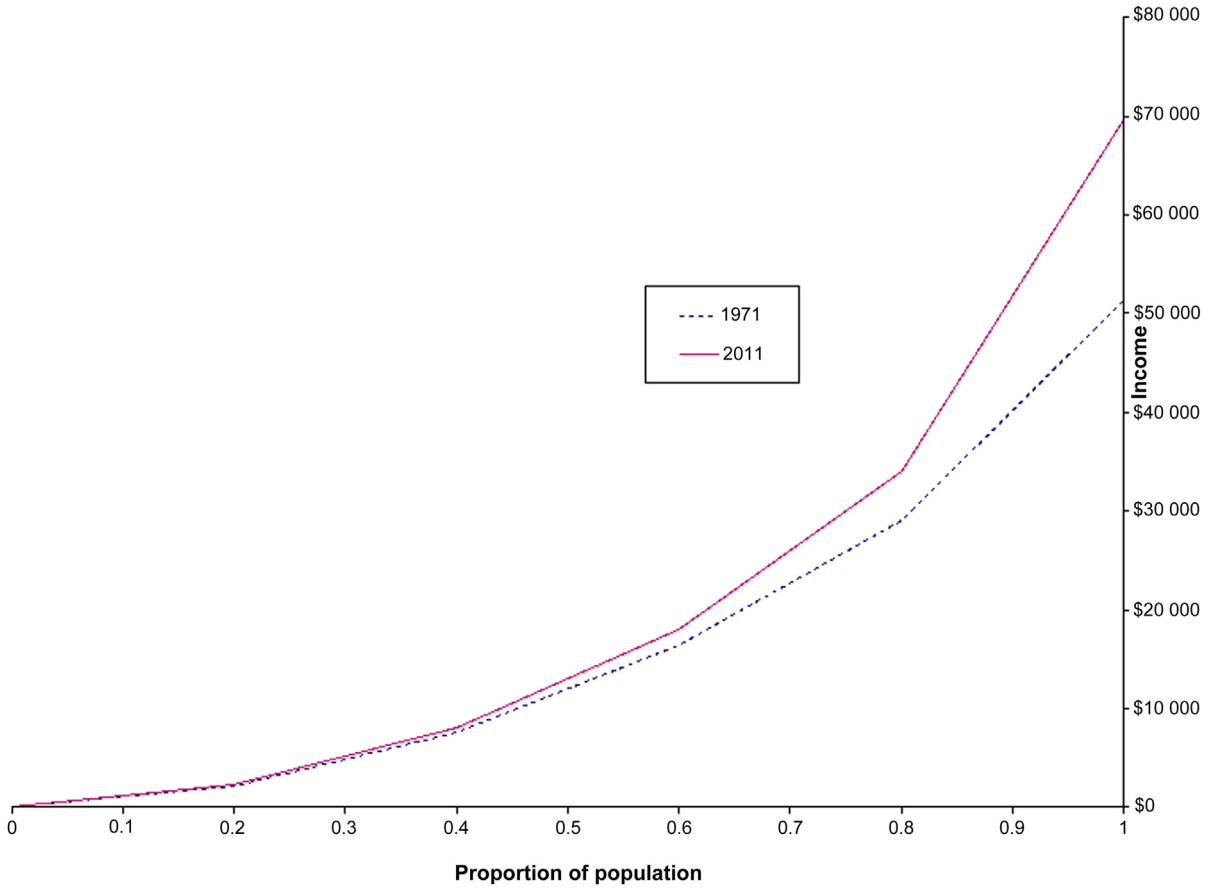

Figure 2 Generalized Lorenz curve. Source: Reproduced from Navas-Walt, C., Proctor, B.D., Lee, C.H., 2012. U.S. Census Bureau, Current Population Reports, P60–243, Income, Poverty, and Health Insurance Coverage in the United States: 2011. U.S. Government Printing Office, Washington, DC (Appendix Table A2). Online. Available at: www.census.gov/prod/2012pubs/p60-243.pdf (accessed 31.07.13.).

population against q and the diagonal line depicts a hypothetical distribution of perfect equality. It is clear that for each q the share was smaller in 2011 than it was in 1971 – the 1971 distribution Lorenz-dominates that for 2011. This simple intuitive notion of greater inequality conforms exactly to a fundamental principle to be explained below.

A further simple result emerges from this diagram. Take the area trapped between the Lorenz curve and the equality diagonal. Using eqns [7] and [9], we can show that the ratio of this area to the area of the whole triangle is given by the weighted sum $\sum_{i=1}^{n} \kappa_i x_{[i]}$, where the weights are $\kappa_i = [2i - 1 - n]/[n\mu(\mathbf{x})]$. This is exactly the Gini coefficient eqn [5].

Axioms

An inequality index I is in some ways like a utility function in consumer theory: it is a representation of an inequality ordering on the members of X and is usually taken to be continuous and ordinal – although there is often a 'natural' cardinal representation of a particular index, a formal argument for one representation rather than another is not usually provided (why not use the logarithm of the Gini coefficient, for example?). Ordinality is sufficient for making comparing income distributions, the primary task of inequality analysis. Axioms are essentially formal statements of the principles of assessment that are used to give meaning to the ordering represented by I. The treatment here does not claim to generality; rather, it focuses on those principles that are central to modern approaches to inequality. Rather than presenting the axioms as formal statements, however, it is more useful here to introduce the underlying key principles discursively.

Assume that everywhere in the following discussion, the vector $\mathbf{x}$ in eqn [1] is any arbitrary member of the set X.

- It is reasonable that the labeling of the components of $\mathbf{x}$ be irrelevant: it does not matter which income receiver gets which income. This means that I has the *symmetry* property:

$$I(x_1, x_2, \ldots, x_n) = I(x_2, x_1, \ldots, x_n) = I(x_3, x_1, \ldots, x_n) = \ldots \quad [10]$$

We will always assume that this holds and we may therefore adopt the convention that incomes have been labeled such that $x_1 \leq x_2 \leq \ldots \leq x_{n-1} \leq x_n$.

- Second, it is necessary to have a coherent way of characterizing inequality in different-sized populations. Perhaps the most obvious assumption is that simple replications of

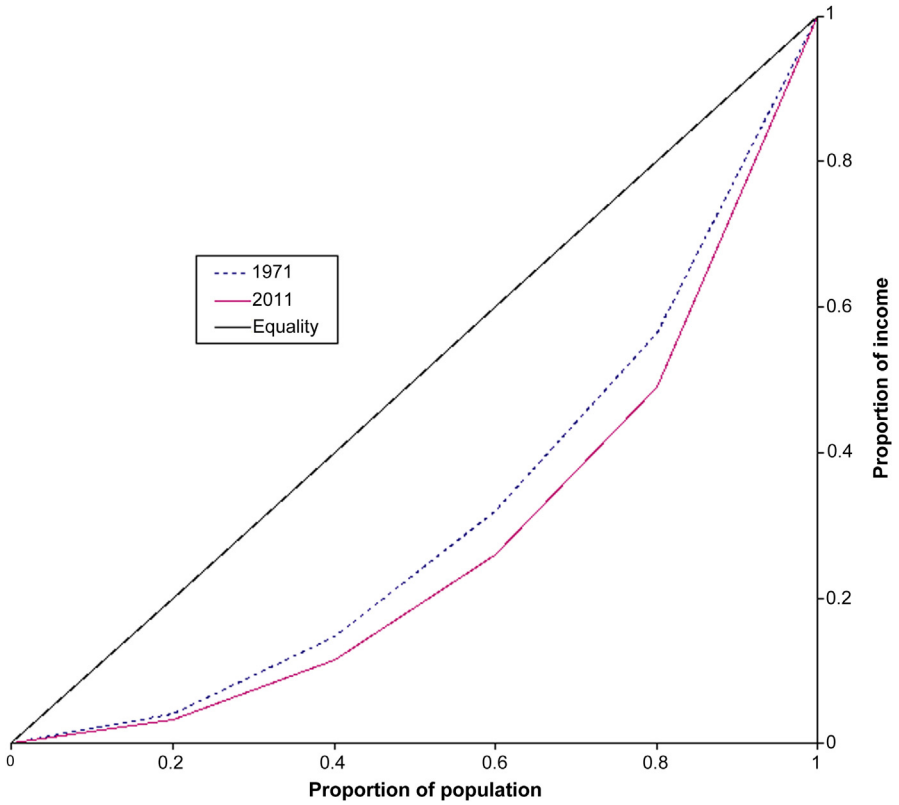

Figure 3 Lorenz curve. Source: Navas-Walt, C., Proctor, B.D., Lee, C.H., 2012. U.S. Census Bureau, Current Population Reports, P60–243, Income, Poverty, and Health Insurance Coverage in the United States: 2011. U.S. Government Printing Office, Washington, DC, (Appendix Table A2). Online. Available at: www.census.gov/prod/2012pubs/p60-243.pdf (accessed 31.07.13.).

an income vector eqn [1] leave inequality unchanged. This is the *population principle*:

$$I(x_1, x_2, \ldots, x_n) = I(x_1, x_1, x_2, x_2, \ldots, x_n, x_n) = I(x_1, x_1, x_1, x_2, x_2, x_2, \ldots, x_n, x_n, x_n) = \ldots \quad [11]$$

Taken in conjunction with symmetry this allows one to represent distributions purely in terms of a distribution function.

- A key assumption that is commonly invoked focuses on the effect on inequality of a hypothetical small income transfer. Suppose $x_i < x_j$ and consider some positive number δ such that $x_i - \delta \geq \underline{x}$, then the *principle of transfers* (Dalton, 1920) requires that

$$I(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_n) < I(x_1, \ldots, x_i - \delta, \ldots, x_j + \delta, \ldots, x_n) \quad [12]$$

  - a poorer-to-richer income transfer will always increase inequality.
- As a counterpart to the assumption relating to different sizes of population (eqn [11]), it is useful to have an assumption relating to different amounts of total income. The standard assumption is that of *scale independence*. This requires that for any scalar $\lambda > 0$,

$$I(\lambda x_1, \lambda x_2, \ldots, \lambda x_n) = I(x_1, x_2, \ldots, x_n) \quad [13]$$

  - double all incomes or halve all incomes and inequality is left unaltered. An alternative assumption that is sometimes used is *translation independence*. Take any real number δ such that $x_1 + \delta \geq \underline{x}$; then

$$I(x_1 + \delta, x_2 + \delta, \ldots, x_n + \delta) = I(x_1, x_2, \ldots, x_n) \quad [14]$$

  - add or subtract one dollar from every income and inequality is left unaltered.

Clearly this brief list raises some important questions. Why use these particular axioms? Some of them appear to be quite strong; for example, although scale independence seems attractive if the 'incomes' x_i here are measured in dollars and we consider just dividing through by some rate of exchange so as to work with incomes in some other monetary units, it may seem less attractive if one wishes to examine the impact on inequality of redistribution policies at different stages of economic growth: a rearrangement of income shares that constitutes a reduction in inequality in a low-income society might not be considered as a reduction in inequality if the whole population is prosperous. Furthermore, the axioms captured by eqns [10]–[13], for example, are satisfied by both eqns [4] and [5] as well as other important classes of inequality measures; on the other hand, the axioms captured by eqns [10]–[12] and [14] are satisfied by eqn [3] and another rich class of inequality measures. Following on from this question, what more is required to get a specific index or well-defined

family of indices that is both theoretically appropriate and practical to implement?

To answer this, we need to be precise about what it means to say that one distribution is more unequal than another and the intellectual basis used for making such comparisons. The meaning of inequality can be further clarified through one of several routes: this article will analyze three of these in turn, namely, social welfare, decomposition, and income differences.

Social Welfare and Inequality

The welfare-economic approach to the subject starts from the position that distributional equality is socially desirable so that inequality is essentially the opposite of welfare. If we adopt this approach then a natural definition of inequality follows. The idea is similar to the conventional measurement of economic waste and the basis for a simple but precise model can be established with only a little more theorizing. Here 'income' is used as shorthand for a comprehensive measure of individual well-being and is assumed to be non-negative.

The *social welfare function* (SWF) is a real-valued function W defined on the space of distributions X. The social welfare associated with a particular income distribution eqn [1], given by

$$W(x_1, x_2, \ldots, x_n), \qquad [15]$$

is to be interpreted as follows: suppose we are given a specific SWF $W(\bullet)$ and that for two separate income distributions $\mathbf{x}'$ and $\mathbf{x}''$, we have $W(\mathbf{x}') > W(\mathbf{x}'')$; then social welfare associated with the distribution $\mathbf{x}'$ is higher than the social welfare associated with the distribution $\mathbf{x}''$. In principle, W is an ordinal function so that the scale of measurement of welfare levels can be subjected to arbitrary monotonic increasing transformations.

This basic specification raises a number of important questions:

- Why express social welfare as a function of income? Income defined how?
- What particular form should W take?
- What is the relation between the functions I and W?

The answer to the first question helps to pin down the relationship between inequality measurement as conventionally practised and standard welfare economics – see Section Welfare and Income. The answers to the last two questions will determine the form of a class of inequality measures and permit us to establish some important welfare-economic results: these are addressed in Sections Social Welfare and Inequality Measures and Ranking Distributions.

Welfare and Income

Let us examine an issue that was sidestepped in the discussion of the example of Tables 1 and 2: how to do the trick of passing from a distribution of dollar income among households to a standard welfare analysis that is typically concerned with the levels of economic well-being of individuals. The standard approach is as follows. We require a method of appropriately capturing the relationship between the living standard that is attainable by an individual and the income that he/she is presumed to have access to within the household. This is conventionally done by defining a function $\nu(\bullet)$ that has as its argument a list of nonincome attributes $\mathbf{a}$ that might include household size, age, and sex of household members and health status; $\nu(\mathbf{a})$ determines the number of *equivalent adults* in the household with attributes $\mathbf{a}$ such that

$$x = \frac{y}{\nu(\mathbf{a})}, \qquad [16]$$

where y is nominal income and x is *equivalized income* that is taken to be comparable across different household types. Note that the equivalization function ν is typically specified as independent of income although this simplification is not essential; of course, the way in which the function ν is determined – from ethical considerations or econometric studies – is an important issue in its own right, but one that lies outside the present discussion. The function ν transforms a distribution of dollar incomes among n households

$$\mathbf{y} = (y_1, y_2, \ldots, y_n) \qquad [17]$$

into a distribution of equivalized incomes by households given by eqn [1]. In order to complete the welfare interpretation, we need to recognize that social welfare considerations are usually represented in terms of individuals rather than households, and so, for example, households consisting of couples should receive more weight in social welfare evaluations than households consisting of single individuals. Therefore, if the income-receiving units consist of households of differing size, we might want to represent this by introducing a corresponding set of population weights w_i for the observations, so that the distribution becomes an ordered list of pairs:

$$((w_1, x_1), (w_2, x_2), \ldots, (w_n, x_n)) \qquad [18]$$

where w_i is the number of persons in household i divided by the number of persons in the whole population. There is little analytical complication in using eqn [18] rather than eqn [1] as a representation of the distribution of equivalized incomes by individuals. Typically, it is just a matter of a minor redefinition of formulas for inequality measures and the like; for example, the coefficient of variation eqn [4] would now be written

$$\sqrt{\sum_{i=1}^{n} w_i \left[\frac{x_i}{\mu} - 1\right]^2} \qquad [19]$$

where μ is the appropriately redefined mean $\sum_{i=1}^{n} w_i x_i$. The method generalizes any inequality measure that can be written in the form $\Phi\left(\frac{1}{n}\sum_{i=1}^{n}\phi(x_i), \mu\right)$ just needs to be rewritten as $\Phi(\sum_{i=1}^{n} w_i \phi(x_i), \mu)$.

However, having introduced this important theoretical qualification, we will now set it aside – for expositional purposes it is convenient to assume (1) that the population consists of isolated individuals that are identical in every relevant respect other than income and (2) that income appropriately represents individual welfare. So, from here on, i indexes individuals or households and the distinction between x and y is dropped.

Social Welfare and Inequality Measures

The idea of the SWF was introduced without discussing specific properties of the function W. Some properties must be imposed on W if we require there to be a specific relationship

between social welfare and inequality and we impose specific assumptions on the function I. However, in addition it is particularly important to be explicit about how W should respond to an increase in one or more incomes. This is the usual principle that is applied:

- Suppose we consider any income distribution $(x_1, x_2, \ldots, x_n)$ and some positive number δ. Then *monotonicity* requires that

$$W(x_1, x_2, \ldots, x_i + \delta, \ldots, x_n) > W(x_1, x_2, \ldots, x_i, \ldots, x_n) \qquad [20]$$

On the assumption that monotonicity holds and that W is a continuous function, the SWF can itself be used to derive a family of inequality measures. There are several ways of doing this, but a standard approach is to represent social welfare using a money metric: we can always do this in view of the ordinal nature of W and the requirement that it be monotonic and continuous. The *equally distributed equivalent* (EDE) income is a real number ξ such that for any $(x_1, x_2, \ldots, x_n)$ in X:

$$W(\xi, \xi, \ldots, \xi) = W(x_1, x_2, \ldots, x_n) \qquad [21]$$

Clearly the relationship eqn [21] can be used to derive EDE as a function of the income distribution, $\xi(\mathbf{x})$; so the function $\xi(\cdot)$ is itself a valid way of representing social welfare.

Suppose we require that the principle of transfers apply to W; this by analogy with eqn [12] means that a mean-preserving poorer-to-richer income transfer will *de*crease social welfare. Then it is always true that $\xi(\mathbf{x}) \leq \mu(\mathbf{x})$ and the normalized gap between ξ and μ provides a natural basis for an inequality index

$$1 - \frac{\xi(\mathbf{x})}{\mu(\mathbf{x})} \qquad [22]$$

It is clear that this index is bounded between zero and one and that if there were perfect equality then we would have $\xi(\mathbf{x}) = \mu(\mathbf{x})$ and inequality in eqn [22] would be zero.

Furthermore, if the scale-independence property eqn [13] is also satisfied, then EDE income takes the form of a generalized mean:

$$\xi(\mathbf{x}) := \left[\frac{1}{n}\sum_{i=1}^{n} x_i^{1-\varepsilon}\right]^{\frac{1}{1-\varepsilon}}, \varepsilon > 0 \qquad [23]$$

and eqn [22] gives the class of *Atkinson indices*:

$$I_{\mathrm{A}}^{\varepsilon}(\mathbf{x}) := 1 - \left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i}{\mu(\mathbf{x})}\right]^{1-\varepsilon}\right]^{\frac{1}{1-\varepsilon}}. \qquad [24]$$

(The limiting forms of eqns [23] and [24] as $\varepsilon \to 1$ are, respectively, $\xi(\mathbf{x}) = \exp\left(\frac{1}{n}\sum_{i=1}^{n} log(x_i)\right)$ and $I_{\mathrm{A}}^{1}(\mathbf{x}) = 1 - exp\left(\frac{1}{n}\sum_{i=1}^{n} log(x_i)\right)/\mu(\mathbf{x})$.)

The number ε is a parameter that characterizes individual members of the class of inequality measures; it is the degree of (relative) inequality aversion associated with the particular member of the class. For any given unequal income distribution, the larger is ε the larger is the Atkinson inequality index – there is an example of this in Table 3 below. There is a close analogy with a class of risk indices in the case of constant relative risk aversion. This is unsurprising since this approach

Table 3 Inequality indices for the example in Table 2

	1971	*2011*
$I_{\mathrm{A}}^{0.25}$	0.068	0.101
$I_{\mathrm{A}}^{0.5}$	0.138	0.198
$I_{\mathrm{A}}^{0.75}$	0.214	0.300
I_{A}^{1}	0.309	0.443
I_{Gini}	0.396	0.477
I_{GE}^{0}	0.370	0.585
I_{GE}^{1}	0.273	0.422

was explicitly founded on the formal similarity between distributional comparisons in terms of inequality and of risk (Atkinson, 1970).

If, instead of the scale-independence property, we required I to satisfy translation independence eqn [14], then we would obtain a different class of indices

$$I_{\mathrm{K}}^{\beta}(\mathbf{x}) := \frac{1}{\beta}\log\left(\frac{1}{n}\sum_{i=1}^{n} e^{\beta[x_i - \mu(\mathbf{x})]}\right) \qquad [25]$$

where $\beta > 0$ is a sensitivity parameter indexing members of the class (Kolm, 1976). The connection of eqn [25] with constant absolute risk aversion is evident. There are other 'regularity' assumptions that could be used, apart from than scale or translation independence – see Bossert and Pfingsten (1990).

Ranking Distributions

As noted earlier, there are important results available about welfare and inequality comparisons that do not require the usage of specific indices. They follow from standard first- and second-order dominance results that are familiar from finance and other disciplines. Take the special class of *additive* welfare functions where W in eqn [15] can be written in the form $\sum_{i=1}^{n} u(x_i)$ for some function u. If W is additive and satisfies the monotonicity axiom, then u must be a strictly increasing function; if, furthermore, W satisfies the principle of transfers, then u must be strictly concave. Then the following powerful results are available for any two distributions $\mathbf{x}'$ and $\mathbf{x}'' \in X$:

- *First order*: $\mathbf{x}'$ strictly Parade-dominates $\mathbf{x}''$ if and only if $W(\mathbf{x}') > W(\mathbf{x}'')$ for any additive W that satisfies the principle of monotonicity.
- *Second order*: $\mathbf{x}'$ strictly GLC-dominates $\mathbf{x}''$ if and only if $W(\mathbf{x}') > W(\mathbf{x}'')$ for any additive W that satisfies monotonicity and the principle of transfers (Shorrocks, 1983).

From the statement of these results, it is clear that the Parade- and GLC-dominance criteria are formally equivalent to first- and second-order stochastic dominances in the analysis of probability distributions.

A version of the second-order result applies to the conventional Lorenz curve and it accords with the intuitive argument presented in the introduction. Take the class of SWFs that satisfies the principle of transfers (they do not have to be

additive). Then, for two distributions $\mathbf{x}'$ and $\mathbf{x}''$ that have the same mean, the statement $W(\mathbf{x}') > W(\mathbf{x}'')$ 'for any W in this class' is true if and only if $\mathbf{x}'$ strictly Lorenz-dominates $\mathbf{x}''$. Furthermore, under these circumstances for any inequality index I that satisfies the principle of transfers it must be the case that $I(\mathbf{x}') < I(\mathbf{x}'')$. The implication of this is that all inequality measures that satisfy the principle of transfers 'go the same way' if one distribution Lorenz-dominates the other. This is illustrated in Table 3 (which again uses the distribution of household income by households). Rows one to four give the results for the Atkinson indices: notice that in each case measured inequality is closer to one (the maximum), the higher is the degree of inequality aversion. The indices in the last two rows of Table 3 are discussed in the next section.

Decomposition

The axioms discussed in Section Axioms induce some structure on inequality measures. By introducing the idea of *decomposing* inequality, we can impose more structure and thereby obtain a useful class of indices. There are two principal types of decomposition: by subgroups of the population (regions, age groups, etc.) and by components of income (labor income, income from capital, etc.) – see Cowell and Fiorio (2011). Here we focus on the population-subgroup issue.

Imagine that the population of n persons can be partitioned into a collection of m groups so that any individual falls into just one of these m groups. Each group j could be considered as a subpopulation of size n_j in its own right (the groups are exhaustive so that $\sum_{j=1}^{m} n_j = n$) and one could compute inequality within this subpopulation as

$$\iota_j = I(\mathbf{x}_j) \qquad [26]$$

where $\mathbf{x}_j$ is the income distribution consisting of just the members of subgroup j. The essence of the decomposition problem is to represent inequality overall as a function of inequality in each group $j = 1, \ldots, m$.

$$I(\mathbf{x}) = F(\iota_1, \iota_2, \ldots, \iota_m; \pi_1, \ldots, \pi_m, s_1, \ldots, s_m) \qquad [27]$$

where F is an aggregation function and the terms after the ';' show that aggregation may depend on the groups' shares of the population $\pi_j = n_j/n$ and the groups' shares of total income $s_j = n_j\mu(\mathbf{x}_j)/n\mu(\mathbf{x})$. A consistency requirement on eqn [27] is that, if the income distribution within subgroup j changes so as to increase ι_j in eqn [26], all other things remaining the same, then inequality overall should increase. Insisting on this requirement on F for all logically possible partitions induces a type of separability on the function $I(\cdot)$ so that the index must be of the general form mentioned just after eqn [19] above. If we also require that scale-independence hold, then the inequality index must take the specific form

$$I_{\mathrm{GE}}^{\alpha}(\mathbf{x}) := \frac{1}{\alpha^2 - \alpha}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i}{\mu(\mathbf{x})}\right]^{\alpha} - 1\right] \qquad [28]$$

or some monotonic transform of it, where α is a real number. The 'GE' used in the labeling of eqn [28] stands for the *generalized entropy* class, which is a generalization of the two indices introduced by Theil (1967). Theil's two indices are those corresponding to the special forms in the cases $\alpha = 0$ and 1, namely

$$I_{\mathrm{GE}}^{0}(\mathbf{x}) := -\frac{1}{n}\sum_{i=1}^{n}\log(x_i/\mu(\mathbf{x})) \qquad [29]$$

and

$$I_{\mathrm{GE}}^{1}(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n}[x_i/\mu(\mathbf{x})]log(x_i/\mu(\mathbf{x})) \qquad [30]$$

The values of these indices for the US example are given in the last two rows of Table 3. The α in eqn [28] is a parameter that characterizes different members of the GE class: a high positive value of α yields an index that is very sensitive to income transfers at the top of the distribution; specifying a negative value will produce an index that is sensitive to income transfers among the poor. Members of the GE class can only handle negative incomes if α is an integer greater than or equal to 2; if α is greater than or equal to 1 the GE index can handle zero incomes.

Finally note that there is a functional relationship between the class eqn [24] and the class eqn [28]. If we set $\varepsilon = 1 - \alpha$ then, long as $\alpha < 1$, we have

$$I_{\mathrm{A}}^{\varepsilon}(\mathbf{x}) = 1 - \left[1 + \alpha[\alpha - 1]I_{\mathrm{GE}}^{\alpha}(\mathbf{x})\right]^{1/\alpha} \qquad [31]$$

Income Differences

The third way forward from the basic argument outlined in Section Axioms focuses on fundamental income differences. This is one of the key ways in which one can motivate usage of the very well-known inequality indices mentioned in Section Indices. The variance and the coefficient of variation eqn [4] can be thought of as a representation of the averaged squared difference between each income x_i and the mean. A compelling argument for the Gini coefficient is that it is the (normalized) expected value of the absolute difference between any two randomly selected incomes in the population.

However, there are other types of income difference that are of special relevance to inequality measurement. Just as some poverty indices can be characterized as a kind of average distance of individual incomes from a reference income level – the poverty line (many poverty indices can be written in the form $\frac{1}{n}\sum_{i=1}^{n} p(z - x_i)$, where z is the poverty line and $p(\cdot)$ is a nondecreasing function that is zero for all $x_i \geq z$) – so also some inequality measures use the idea of a reference level income. In the case of inequality, the reference income level has been suggested as either that of the best-off person in society, or of the average income of all those who are better off than any given person i (Temkin, 1993). In each of these cases, application of standard axioms about the structure of inequality orderings leads to a class of inequality indices that bears a functional similarity to poverty indices and to indices of relative deprivation (Cowell and Ebert, 2004).

Implementation

The practical issues associated with the exposition of the example in Tables 1 and 2 highlight some of the problems in

implementing inequality measures and associated tools – the definition of income, income receiver, etc. Given the way in which income data are usually obtained, issues of sampling and measurement error usually need to be treated carefully. Furthermore, the special nature of income and wealth distributions and the sensitivity of inequality indices to very high or very low incomes usually require that particular attention be paid to the problem of outliers. Finally, it should be noted that it is still sometimes the case that the data required for estimating inequality indices are made available only in grouped form rather than as microdata so that special techniques may be required for interpolation within income intervals and for modeling the tails of the distribution.

See also: Earnings Inequality; Income Distribution and Inequality: Measurement Issues; Poverty: Measurement and Analysis; Wealth Distribution.

Bibliography

Allison, R.A., Foster, J.E., 2004. Measuring health inequality using qualitative data. Journal of Health Economics 23, 505–552.

Atkinson, A.B., 1970. On the measurement of inequality. Journal of Economic Theory 2, 244–263.

Atkinson, A.B., 1983. The Economics of Inequality, second ed. Clarendon Press, Oxford.

Bossert, W., Pfingsten, A., 1990. Intermediate inequality: concepts, indices and welfare implications. Mathematical Social Science 19, 117–134.

Cowell, F.A., 2000. Measurement of inequality. In: Atkinson, A.B., Bourguignon, F. (Eds.), Handbook of Income Distribution. North Holland, Amsterdam.

Cowell, F.A., 2011. Measuring Inequality, third ed. Oxford University Press, Hemel Hempstead.

Cowell, F.A., Ebert, U., 2004. Complaints and inequality. Social Choice and Welfare 23, 71–89.

Cowell, F.A., Fiorio, C., 2011. Inequality decompositions – a reconciliation. Journal of Economic Inequality 9, 509–528.

Dalton, H., 1920. Measurement of the inequality of incomes. Economic Journal 30, 348–361.

Kolm, S.-C., 1976. Unequal inequalities I. Journal of Economic Theory 12, 416–442.

Lambert, P.J., 2001. The Distribution and Redistribution of Income, third ed. Manchester University Press, Manchester.

Navas-Walt, C., Proctor, B.D., Lee, C.H., 2012. U.S. Census Bureau, Current Population Reports, P60–243, Income, Poverty, and Health Insurance Coverage in the United States: 2011. U.S. Government Printing Office, Washington, DC. Online. Available at: www.census.gov/prod/2012pubs/p60-243.pdf (accessed 31.07.13.).

Pen, J., 1974. Income Distribution, second ed. Allen Lane, Penguin Press, London.

Sen, A.K., Foster, J.E., 1997. On Economic Inequality, second ed. Clarendon Press, Oxford.

Shorrocks, A.F., 1983. Ranking income distributions. Economica 50, 3–17.

Temkin, L.S., 1993. Inequality. Oxford University Press, Oxford.

Theil, H., 1967. Economics and Information Theory. North Holland, Amsterdam.

Inequality, Social

Douglas S Massey, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although humans have existed for around 200 000 years, for most of their time on earth their societies were not characterized by much inequality. Despite some differentiation on the basis of age, gender, and kinship, people enjoyed relatively equal access to resources. Inequality begins to rise with the emergence of cities around 10 000 years ago, which yielded the first stratified class-based societies. Inequality rose quite dramatically with the industrialization and urbanization of society around 200 years ago, as more human interactions occurred in the context of markets. Because markets are socially embedded, different architectures for exchange, production, and consumption yield widely discrepant levels of inequality. Variations in equality reflect the degree to which categorical mechanisms of stratification, such as exploitation and opportunity, are built into societies and the markets they house, and the number and range of people who are subjected to nonmarket processes of discrimination and exclusion.

The Evolution of Inequality

Prior to the agricultural revolution, human societies generally were not stratified. People living in stratified societies are characterized by unequal access to material, symbolic, and emotional resources. Material resources include food, goods, income, and various accumulations of physical and financial capital. Symbolic resources include power, social capital, and cultural capital, which are often signified by physical tokens of status, prestige, and social standing. Emotional resources include warmth, attraction, liking, and love, both physical and emotional.

Until around 10 000 years ago the vast majority of human beings lived in small foraging groups of no more than 150 people. Foraging peoples sustain themselves through a combination of hunting and gathering, with the precise mix being determined by what food sources are available in the local environment (Kelly, 1995). In foraging societies, people are differentiated on the basis of age, gender, and kinship but not much else. Although foraging societies evince some stratification with respect to these characteristics, resources are generally distributed equally, especially when flora and small fauna constitute the main food sources. In societies where big-game hunting serves as a key food source, males typically garner more power and prestige within the group (Sanday, 1981); but even in these societies a variety of informal leveling mechanisms limit inequalities, enforcing equality through mechanisms such as shaming, ridicule, and ostracism on the negative side, and praise, kudos, and respect on the positive side.

The agricultural revolution created the first settled populations and led to the formation of cities and the first truly stratified societies. According to Fiske (1991), humans are capable of engaging in four fundamental relationships: communal sharing, equality matching, authority ranking, and market pricing. In communal sharing relationships, people have a common sense of identity, express weak individuality, and share resources with others on the basis of need. Equality matching occurs when peers are recognized as equals but are treated as distinct individuals who engage in tit-for-tat exchanges over time, such that any one person's inputs to the collective are ultimately matched by contributions from others.

Whereas communal sharing and equality matching predominate in foraging societies, the advent of cities brought about the proliferation of authority-ranking relationships (Fiske, 1991). Authority ranking occurs within a hierarchical social structure characterized by progressive upward inclusion within a relation of dominance. As one moves up the hierarchy, members at each successive rank dominate a larger number of subordinates occupying lower levels. A ranked category within any social structure is generally referred to as a class, and as noted earlier, members of different classes are characterized by unequal access to material, symbolic, and emotional resources.

Urbanism also created the potential for a fourth kind of human relationship – market pricing – in which value is determined by competition between people who supply and demand scarce goods, services, and resources (Fiske, 1991). Relationships in a market system are structured according to price, a universal metric by which people are able to compare themselves to each other and to judge the value of the things they exchange, no matter how dissimilar they might otherwise be. People who interact through market pricing value each others' actions, services, and products according to the price at which they can be exchanged.

Although urbanism created the potential for markets, market pricing was rather slow to evolve as a common human relationship. Although cities emerged around 10 000 years ago and quickly evolved quarters dedicated to barter (Chant and Goodman, 1999), the first coins were minted only 2700 years ago (Davies, 2002). With the invention of money, market pricing came into existence; but markets remained limited in scope until quite recently, being confined to the exchange of a small range of goods and services within urban areas. Prior to industrialization, only about 5% of any society's inhabitants lived in cities and most human relations occurred in small peasant villages with relationships dominated by communal sharing, equality matching, and authority ranking (Massey, 2005).

Although markets began to expand in the sixteenth century, it was not until the advent of industrialism and the urbanization of society in the nineteenth century that market pricing became a common form of human interaction. From AD 1800 onward, well-functioning markets for land, labor, capital,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32183-3

goods, information, and a myriad of services expanded within and between nations, ultimately creating a global economy that exposed people to market pricing on a daily basis. The emergence of the market economy was accompanied by a dramatic increase in material inequality, as the gap between the richest and poorest members of society expanded to unprecedented levels (Kuznets, 1955). Between AD 1850 and 1950, for example, the total value of goods and services produced throughout the world rose from $939 trillion to $5336 trillion (Maddison, 2001) and the largest private fortune in the United States rose from $1 million to $1.6 billion (Phillips, 2002).

Markets generate inequalities in income and wealth because of differences in productivity between individuals and variations in the scarcity of the things exchanged through markets. Differences in productivity reflect different endowments of human, social, physical, and financial capital whereas differences in scarcity reflect variation in the abundance of goods, services, and resources. More productive people earn higher incomes and owners of scarce commodities receive higher prices, thus producing inequalities with respect to income and wealth.

As indicated by the foregoing history, however, markets are a human invention, not a state of nature, and quite a recent one at that. Being of recent origin, markets evolved out of and are embedded within preexisting social structures, yielding a variety of different market 'architectures' (Hall and Soskice, 2001; Fligstein, 2001). Depending on how markets are socially constructed, widely divergent levels of inequality may be achieved. The degree of inequality ultimately produced by a market economy is thus socially determined. Across all human societies, inequality is produced by a combination of two simple but powerful social mechanisms: (1) the allocation of people to socially defined categories and (2) the creation of mechanisms that allocate resources unequally across categorical boundaries, producing what Tilly (1998) calls 'categorical inequality,' – a form of stratification that is remarkably 'durable' in the sense that it is reproduced over time and across generations (Massey, 2007).

The Psychological Foundations of Inequality

Stratification begins psychologically with the creation of cognitive boundaries that separate people into social groups. Before categorical inequality can be implemented socially, categories must be created cognitively to classify people based on observed or inferred characteristics. Categorical thinking is a characteristic feature of human cognition. Since human memory is finite, if the brain is to remember more it must combine pieces of information into sets of conceptual categories known as schemas, using common properties to classify a large number of people, objects, or experiences into a smaller number of readily identifiable mental constructs.

Schemas do not exist simply as neutral mental representations, but are typically associated with emotional valences. The human brain is composed of two parallel processors which, while interconnected, function independently (Carter, 1998; Panksepp, 1998). The 'emotional brain' is rooted in a set of neural structures that are common to all mammals and known collectively as the limbic system, whereas the 'rational brain' is centered in the prefrontal cortex and is by far the most developed in humans (Damasio, 1994, 1999). The two portions of the brain, (labeled System 1 and System 2 by Kahneman, 2011) are interconnected, but the numbers and speed of the neural connections running from the limbic system to the neocortex are greater than in the reverse connections. Therefore the emotions stored in the limbic system, which are typically unconscious or implicit, greatly affect how humans feel about categories that exist within the rational, conscious brain (LeDoux, 1996; Zajonc, 1998).

Emotions stored in the limbic system may be positive or negative but when they are associated with particular categories of people they contribute to prejudice, which is a predetermined emotional orientation toward members of other social categories (Fiske, 2003). A prejudicial orientation for or against a social group thus contains both conscious and unconscious components (Bargh, 1997). On the one hand, a person may be a principled racist who consciously believes that African-Americans are inferior. Another person may sincerely and overtly believe in racial equality but he may harbor a variety of anti-black sentiments and associations that are unconscious. These associations are created through conditioning in society; – for example, the repeated visual pairing of violent crime scenes with black perpetrators seen on television, creates an unconscious association between African-Americans and criminality.

All human beings whether they think of themselves as prejudiced or not, mentally construct schemas that classify people into categories based on age, gender, race, and ethnicity (Stangor et al., 1992; Fiske, 1998). Although stereotypes are always present, people are especially likely to fall back on them in making judgments when they feel challenged, face uncertainty, or experience a sensory overload (Bodenhausen and Lichtenstein, 1987). In making stereotypical judgments about other social actors, humans evaluate people along two fundamental dimensions: warmth and competence (Fiske et al., 2002). Warmth refers to how likable and approachable a person is. We are attracted to people we view as high on the warmth dimension and seek to interact and spend time with them. We find people who are low on the warmth dimension to be off-putting and generally avoid them, seeking to minimize the number and range of social contacts with them; we don't like them and find them 'cold.'

In addition to these subjective feelings of attraction and liking, we also evaluate people in terms of competence – their ability to act in a purposeful manner to get things done. We may or may not like people who are highly competent, but we generally respect them and admire their ability to achieve. These two dimensions of social perception define the stereotype content model, which holds that stereotyping involves the cognitive placement of groups and individuals in a two-dimensional social space defined by the intersection of the dimensions of warmth and competence, yielding four basic combinations that determine the content of social perceptions (Fiske et al., 2002).

People perceived as high in both warmth and competence include members of one's own social group as well as members of groups perceived to be similar to one's own. People who are warm and competent are approachable and worthy of respect and the relevant emotion associated with them is esteem or

pride. In addition to those perceived to be in and similar to the 'in-group,' the model defines three basic kinds of 'out-groups.' Those groups and individuals perceived to be competent but not warm are respected but not liked and the relevant emotion people feel toward them is envy. This category includes professionals and the rich, but also the classic middlemen minorities, such as Jews in Medieval Europe, the Chinese in Malaysia, and Indians in East Africa. In a stable social structure, people show public respect and defer to members of envied out-groups, but if the social order breaks down they may become targets of communal violence. Groups and individuals perceived to be warm but not competent include the disabled, the elderly, the blind, the mentally retarded – people who have experienced some misfortune but who are otherwise perceived as 'people like us' – so the relevant emotion for them is pity. In a stable social structure pitied out-groups are cared for; but in times of social disorder or scarcity they may suffer from neglect, though they generally do not become targets of communal violence. Finally, individuals and groups perceived as low in both warmth and low in competence include social outcasts such as junkies, drug dealers, sex offenders, and the homeless and the relevant emotion associated with them is disgust or scorn.

Recent work in neuroscience has implicated a particular region of the brain as central to the process of social cognition (see Harris and Fiske, 2006). Whenever individuals perceive a stimulus to be a human being and therefore a potential social actor, an area of the brain known as the medial prefrontal cortex lights up when observed under functional magnetic resonance imagery (fMRI). Harris and Fiske (2006) pretested a number of photographic images of social actors to establish the quadrant into which they fell; then they showed these images to experimental subjects so that each person saw a total of 80 images – 20 images of esteemed in-group members; 20 images from envied out-groups; 20 images of pitied out-group members; and 20 images of people from despised out-groups. As they viewed the various social images, the brains of subjects were scanned under fMRI and the centers of activity recorded.

As expected, the investigators found that images of people representing the in-group or others like them, triggered a clear reaction in the medial prefrontal cortex, as did images of envied out-groups and pitied out-groups. Startlingly, however, images of despised out-groups did not light up this brain region (Harris and Fiske, 2006). Whereas out-groups triggering feelings of pity and envy were instantly perceived as human beings and social actors, those that were despised were not seen in social terms at all at the most fundamental level of cognition. Despised out-groups are thus dehumanized at the neural level, giving other human beings license, in their own minds, to treat them as if they were animals or objects.

The Social Construction of Inequality

The position of a group within the social space defined by warmth and competence is not fixed but malleable, varying across time and space as well as across. Although social categories are ultimately constructed and maintained by individuals within their own minds, the process by which boundaries are expressed in society is quintessentially social. Group identities and boundaries are negotiated through repeated interactions that establish working definitions of the categories in question, including both objective and subjective content, a process that sociologists have labeled boundary work (Lamont and Molnar, 2002). When social actors succeed in establishing the limits and content of various social categories in the minds of others, psychologists refer to the process as framing (Kahneman and Tversky, 2000). In essence, boundary work involves defining social categories in the social structure and framing involves defining them in human cognition.

People naturally favor boundaries and framings that grant them greater access to materialistic, symbolic, and emotional resources and they seek to convince others to accept their favored version of social reality (see Lakoff, 2002). In general, social actors who control more resources in society – those toward the top of the stratification system – have the upper hand in framing and boundary work (Fiske, 2011). Thus whites historically have perpetuated negative stereotypes of African-Americans as unintelligent, violent, hypersexual, and shiftless; and rich people likewise have promoted a view of the poor as lazy, unmotivated, indisciplined, and undeserving, cognitively rendering both groups despicable and not worthy of consideration. To the extent that negative stereotypes become a part of everyday social cognition, individual members of 'cold' and incompetent out-groups will tend to experience mistreatment and inequality in society.

Given any set of cognitively framed and socially constructed categories and a population of people distributed among them, inequality is generated and perpetuated by two basic mechanisms: exploitation and opportunity hoarding (Tilly, 1998). Exploitation occurs when people in one social group expropriate a resource produced by members of another social group and prevent the latter from realizing the full value of the effort involved in producing it. Opportunity hoarding occurs when actors in one social group restrict access to a scarce resource, either through outright denial to members of other social groups or by exercising monopoly control that requires members of other groups to pay a rent in return for access (Tilly, 1998).

Within any social structure, exploitation and opportunity hoarding are, in turn, reinforced by two other social processes that work over time to institutionalize categorical distinctions and lock them into place (Tilly, 1998). The first is emulation, whereby one group of people copies a set of social categories and stratifying mechanisms from another group, or transfers the categories and mechanisms from one social setting to another. The second is adaptation, whereby social relations and day-to-day behaviors become oriented toward a set of ranked categories, so that decisions about who to befriend, who to help, who to share with, who to live near, who to court, and who to marry are made in ways that assume the existence and importance of asymmetric distinctions. According to Tilly (1998: 10), "exploitation and opportunity hoarding favor the installation of categorical inequality, while emulation and adaptation generalize its influence."

Social Inequality in Practice

In the developed societies of today, inequality is produced by categorical mechanisms of exploitation and opportunity

hoarding embedded within social structures that organize and shape markets, thereby conditioning the market outcomes experienced by different social groups. In the United States, categorical distinctions on the basis of class, race, and gender have historically served to generate inequality, though the precise nature of these dimensions have shifted over time (Massey, 2007). Until the 1960s the starkest inequalities were produced along the lines of race by socially structured exclusion and discrimination, which were built into US markets and infused into American social practices.

The institution of slavery represents the harshest possible system for socially constructing exploitation and opportunity hoarding. In the United States, persons of African origin historically have been defined socially as a despised and scorned out-group, thus rendering them cognitively suitable for systematic discrimination and exclusion. Until AD 1865 they were subject to a set of formal and informal social institutions – public laws and private practices as well as shared expectations and individual prejudice – that enslaved Africans and compelled them to work for little more than sustenance while excluding them from virtually all sources of well-being in society, including income, health, wealth, and all forms of capital.

After the Civil War, slavery in the South was replaced by a new system of exploitation and opportunity hoarding known as 'Jim Crow.' To replace slavery an alternative set of formal laws and informal practices were established to confine African-Americans to a system of debt peonage and exclude them all but the most menial occupations, while paying them below-market wages, subjecting them to separate and unequal circumstances across all venues, and controlling their behavior through high levels of violence, both vigilante and state-sanctioned (Packard, 2003). In the North, segregation was *de facto* rather than *de jure*, as African-Americans were residentially confined to ghettos created through institutionalized practices of discrimination and through exclusion from real estate and banking services (Massey and Denton, 1993). As the state assumed a larger role in the political economy during the 1930s, African-Americans were also systematically excluded from the resources and protections of the New Deal (Katznelson, 2005).

During the 1960s and 1970s, of course, Civil Rights legislation eliminated the legal foundations of Jim Crow and overturned the racial exclusions of the New Deal. The 1964 Civil Rights Act outlawed discrimination in labor markets and the provision of services, and promoted school desegregation; the 1965 Voting Rights Act enfranchised black voters; the 1968 Fair Housing Act banned discrimination in the rental and sale of housing; the 1974 Equal Credit Opportunity Act prohibited discrimination in lending markets; and the 1977 Community Reinvestment Act forbade the denial of credit to black neighborhoods. At the same time, President Lyndon Johnson's 'War on Poverty' channeled vast resources to formerly excluded groups (Quadagno, 1994).

As a result of these legislative reforms, black employment, earnings, and educational achievement rose and levels of residential segregation declined. Despite these improvements, however, gaps in socioeconomic status persist; and progress toward racial equality eventually stagnated in the wake of a white counter-reaction (Bobo et al., 1997; Sharkey, 2012). Although segregation declined in small metropolitan areas with small black populations, it remains high in large areas with large black populations and racism continues to play an active role in promoting black spatial isolation (Rugh and Massey, 2013). At the same time, new mechanisms of racial exclusion have evolved, with restrictive density zoning emerging as the primary barrier to integration across metropolitan areas (Rothwell and Massey, 2009).

Although civil rights laws have eliminated overt racial discrimination in most markets, surreptitious discrimination nonetheless continued to be documented by audit studies (Blank et al., 2004). Moreover, beginning in the 1980s, a new, racialized system of criminal justice evolved to produce extreme rates of incarceration for African-Americans, especially those who were young, male, and lacked a college education (Alexander, 2010). The rising prevalence of criminal records among young black males, in turn, put them at a severe disadvantage competing in markets and gaining access to social resources, while straining families and social organizations in the black community (Western, 2006; Pager, 2007).

Other racial and ethnic groups also have historically experienced exploitation and exclusion from opportunity in the United States, though not to the same extent as African-Americans. Asian immigrants were excluded by discriminatory immigration legislation for most of the twentieth century; were denied full access to citizenship (Zolberg, 2006); were confined to their own residential enclaves (Tsui, 2009); and under certain circumstances interred in camps (Daniels, 1993). Immigrants from Southern and Eastern Europe also faced restrictive immigration quotas after 1920 (Zolberg, 2006) and ensuing generations struggled to be accepted as White (Roediger, 1991). Following the Treaty of Guadalupe Hidalgo in 1848, persons of Mexican origin were subjected to systematic exploitation and exclusion throughout the American Southwest, resulting in high levels of residential segregation, low levels of education, meager earnings (Grebler et al., 1970), and widespread skin color discrimination in markets (Telles and Murguía, 1992).

Although exploitation and exclusion directed at persons of Southern European, Eastern European, and Asian origin have faded in recent years to yield rough parity with mainstream Americans (Alba and Nee, 2003), Mexicans paradoxically have come to experience higher levels of discrimination and exclusion in recent years, owing to the rise of the US immigration system as a powerful race-making institution (Massey, 2013). The curtailment of opportunities for legal entry after 1965, produced a rise in illegal migration and the propagation of a 'Latino Threat Narrative' in public discourse and the media (Chavez, 2008). The demonization of Latinos hardened public opinion against Mexicans in particular and Hispanics in general and spurred a militarization of the Mexico–US border, a policy that only served to cut off circulatory migration and increase the size of the undocumented population in the United States (Massey and Pren, 2012). The rising number and increasing geographic spread of illegal immigrants in the context of recession and unemployment unleashed a new wave of repressive policies and negative framing that steadily pushed Latinos into the despised quadrant of social cognition (Lee and Fiske, 2006), down the socioeconomic hierarchy (Massey, 2007), and toward higher levels of segregation (Rugh and Massey, 2013).

Gender has also served historically as a prominent dimension of stratification in the United States. Women were denied the right to vote until 1920, could not enter into private contracts without a husband's approval in many states until the 1960s, were systematically excluded from many occupations until well into the civil rights era (Massey, 2007), and they continue to earn less than men with similar qualifications, especially in female-dominated occupations (England et al., 2001; Bobbit-Zeher, 2011). Although overt discrimination against women is now illegal, covert gender discrimination has persisted in many markets (Darity and Mason, 1998; Goldin and Rouse, 2000) and the earnings gap increases with rise in occupational status (2001). Meanwhile, occupational segregation by gender remains quite high (Charles and Grusky, 2004) and the differential distribution of men and women across occupational categories explains a large part of the gender gap in earnings (Blau and Kahn, 2006).

Finally, class is the critical third dimension of categorical inequality in the United States and it has also undergone significant change in recent decades. Whereas inequalities by gender and race have generally decreased, though at times quite slowly, class divisions have unambiguously widened. Residential segregation by income and education have increased (Massey et al., 2009) and inequality with respect to wealth and income have skyrocketed since the 1970s (Piketty and Saez, 2003; Wolff, 2010; Alvaredo et al., 2013) in response to deliberate policy actions that have served to harm low and middle income people while benefiting the rich (Bartels, 2008). The real value of the minimum wage has been allowed to decline to the lowest level in decades; public spending on benefits and services to the poor has flagged; federal income transfers and food subsidies have been severely curtailed;, and consumer debt has risen and surged to widen the gap between overall interest rates and credit card interest rates. At the same time, tax rates on top incomes and capital gains have been reduced to the lowest levels since the 1920s (Massey, 2007). As a result, whereas pretax inequality in the United States is no greater than in other developed nations (and in several cases is less) once taxes are taken into account, income inequality in the United States is by far the highest (Smeeding, 2005) – thus pointing to the social origins of inequality in the United States (Massey, 2009).

See also: Class Consciousness; Class: Social; Economic Sociology; Elites: Sociological Aspects; Equality of Opportunity; Income Inequality; Mobility, Social; Money, Sociology of.

Bibliography

Alba, R.D., Nee, V., 2003. Remaking the American Mainstream: Assimilation and Contemporary Immigration. Harvard University Press, Cambridge, MA.

Alexander, M., 2010. The New Jim Crow: Mass Incarceration in the Age of Colorblindness. New Press, New York.

Alvaredo, F., Atkinson, A.B., Piketty, T., Saez, E., 2013. The top 1 percent in international and historical perspective. Journal of Economic Perspectives 27, 3–20.

Blank, R.M., Dabady, M., Citro, C.F. (Eds.), 2004. Measuring Racial Discrimination. National Academies Press, Washington, DC.

Bargh, J.A., 1997. The automaticity of everyday life. In: Weyer Jr., R.S. (Ed.), The Automaticity of Everyday Life: Advances in Social Cognition. Erlbaum, Mahwah, NJ, pp. 1–61.

Bartels, L.M., 2008. Unequal Democracy: The Political Economy of the New Gilded Age. Princeton University Press, Princeton, NJ.

Blau, F.D., Kahn, L.M., 2006. The gender pay gap: going, going…but not gone. In: Blau, F.D., Brinton, M.C., Grusky, D.V. (Eds.), The Declining Significance of Gender? Russell Sage Foundation, New York, pp. 25–66.

Bobbit-Zeher, D., 2011. Gender discrimination at work: connecting gender stereotypes, institutional policies, and gender composition of workplace. Gender & Society 25, 764–786.

Bobo, L.D., Kluegel, J.R., Smith, R.A., 1997. Laissez-faire racism: the crystallization of a kinder, gentler, antiblack ideology. In: Tuch, S.J., Martin, J.A. (Eds.), Racial Attitudes in the 1990s: Continuities and Change. Praeger, Westport, CT, pp. 15–44.

Bodenhausen, G.V., Lichtenstein, M., 1987. Social stereotypes and information processing strategies: the impact of task complexity. Journal of Personality and Social Psychology 52, 871–880.

Bodenhausen, G.V., Wyer, R.S., 1985. Effects of stereotypes on decision making and information processing strategies. Journal of Personality and Social Psychology 48, 267–282.

Carter, R., 1998. Mapping the Mind. University of California Press, Berkeley, CA.

Chant, C., Goodman, D., 1999. Pre-industrial Cities and Technology. Routledge, London.

Charles, M., Grusky, D.B., 2004. Occupational Ghettos: The Worldwide Segregation of Women and Men. Stanford University Press, Stanford, CA.

Chavez, L.R., 2008. The Latino Threat: Constructing Immigrants, Citizens, and the Nation. Stanford University Press, Stanford, CA.

Damasio, A.R., 1994. Descartes' Error: Emotion, Reason, and the Human Brain. Putnam, New York.

Damasio, A.R., 1999. The Feeling of What Happens: Body and Emotion in the Making of Consciousness. Harcourt Brace, New York.

Daniels, R., 1993. Prisoners without Trial: Japanese Americans in World War II. Hill and Wang, New York.

Darity Jr., W.A., Mason, P.L., 1998. Evidence on discrimination in employment: codes of color, codes of gender. Journal of Economic Perspectives 12, 63–90.

Davies, G., 2002. A History of Money: From Ancient Times to the Present Day. University of Wales Press, Cardiff, CA.

England, P., Thompson, J., Aman, C., 2001. The sex gap in pay and comparable worth: an update. In: Berg, I., Kalleberg, A. (Eds.), Sourcebook on Labor Markets: Evolving Structures and Processes. Plenum, New York, pp. 551–566.

Fiske, A.P., 1991. Structures of Social Life: The Four Elementary Forms of Human Relations. Free Press, New York.

Fiske, S.T., 1998. Stereotyping, prejudice, and discrimination. In: Gilbert, D.T., Fiske, S.T., Lindzey, G. (Eds.), The Handbook of Social Psychology, fourth ed., vol. I. McGraw Hill, New York.

Fiske, S.T., 2003. Social Beings: A Core Motives Approach to Social Psychology. Wiley, New York.

Fiske, S.T., 2011. Envy up, Scorn Down: How Status Divides Us. Russell Sage Foundation, New York.

Fiske, S.T., Cuddy, A.J.C., Glick, P., Xu, J., 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology 82, 878–902.

Fligstein, N., 2001. The Architecture of Markets: An Economic Sociology of Twenty-first Century Capitalist Societies. Princeton University Press, Princeton, NJ.

Goldin, C., Rouse, C., 2000. Orchestrating impartiality: the impact of 'blind' auditions on female musicians. American Economic Review 90, 715–741.

Grebler, L., Moore, J., Guzman, T., 1970. The Mexican-American People: The Nation's Second Largest Minority. Free Press, New York.

Hall, P.A., Soskice, D., 2001. Varieties of Capitalism: The Institutional Foundations of Comparative Advantage. Oxford University Press, Oxford, UK.

Harris, L.T., Fiske, S.T., 2006. Dehumanizing the lowest of the low: neuro-imaging responses to extreme outgroups. Psychological Science 17, 847–853.

Kahneman, D., 2011. Thinking, Fast and Slow. Farrar, Straus, and Giroux, New York.

Kahneman, D., Tversky, A., 2000. Choices, Values, and Frames. Cambridge University Press, New York.

Katznelson, I., 2005. When Affirmative Action was White: An Untold History of Racial Inequality in Twentieth-Century America. W.W. Norton, New York.

Kelly, R.L., 1995. The Foraging Spectrum: Diversity in Hunter-Gatherer Lifeways. Smithsonian Institution Press, Washington, DC.

Kuznets, S., 1955. Economic growth and income inequality. American Economic Review 49, 1–28.

Lakoff, G., 2002. Moral Politics: How Liberals and Conservatives Think. University of Chicago Press, Chicago, IL.
Lamont, M., Molnar, V., 2002. The study of boundaries in the social sciences. Annual Review of Sociology 28, 167–195.
LeDoux, J., 1996. The Emotional Brain: The Mysterious Underpinnings of Emotional Life. Simon and Schuster, New York.
Lee, T.L., Fiske, S.T., 2006. Not an outgroup, not yet an ingroup: immigrants in the stereotype content model. International Journal of Intercultural Relations 30, 751–768.
Maddison, A., 2001. The World Economy: Historical Statistics. Organization for Economic Cooperation and Development, Paris.
Massey, D.S., 2005. Strangers in a Strange Land: Humans in an Urbanizing World. W.W. Norton, New York.
Massey, D.S., 2007. Categorically Unequal: The American Stratification System. Russell Sage Foundation, New York.
Massey, D.S., 2009. Globalization and inequality: explaining American exceptionalism. European Sociological Review 25, 9–24.
Massey, D.S., 2013. The racialization of Latinos in the United States. In: Tonry, M., Bucerius, S. (Eds.), The Oxford Handbook on Ethnicity, Crime, and Immigration. Oxford University Press, New York, pp. 21–40.
Massey, D.S., Denton, N.A., 1993. American Apartheid: Segregation and the Making of the Underclass. Harvard University Press, Cambridge, MA.
Massey, D.S., Pren, K.A., 2012. Unintended consequences of US immigration policy: explaining the post-1965 surge from Latin America. Population and Development Review 38, 1–29.
Massey, D.S., Rothwell, J., Domina, T., 2009. Changing bases of segregation in the United States. Annals of the American Academy of Political and Social Science 626, 74–90.
Packard, G.M., 2003. American Nightmare: The History of Jim Crow. St. Martin's Press, New York.
Pager, D., 2007. Marked: Race, Crime, and Finding Work in an Era of Mass Incarceration. University of Chicago, Chicago, IL.
Panksepp, J., 1998. Affective Neuroscience: The Foundations of Human and Animal Emotions. Oxford University Press, New York.
Phillips, K., 2002. Wealth and Democracy: A Political History of the American Rich. Broadway Books, New York.
Piketty, T., Saez, E., 2003. Income inequality in the United States, 1913–1998. Quarterly Journal of Economics 158, 1–16. Updated to 2012 at: http://elsa.berkeley.edu/∼saez/TabFig2012prel.xls.
Quadagno, J., 1994. The Color of Welfare: How Racism Undermined the War on Poverty. Oxford University Press, New York.
Rugh, J.S., Massey, D.S., 2013. Segregation in the post-civil rights era: stalled integration or end of the segregated century? The DuBois Review Online http://dx.doi.org/10.1017/0S1742058X13000180.
Roediger, D.R., 1991. The Wages of Whiteness: Race and the Making of the American Working Class. Verso Books, New York.
Rothwell, J., Massey, D.S., 2009. The effect of density zoning on racial segregation in U.S. urban areas. Urban Affairs Review 44, 799–806.
Sanday, P.R., 1981. Female Power and Dominance: On the Origins of Sexual Inequality. University of Pittsburgh Press, Pittsburgh, PA.
Sharkey, P., 2012. Stuck in Place: Urban Neighborhoods and the End of Progress toward Racial Equality. University of Chicago Press, Chicago, IL.
Smeeding, T.M., 2005. Public policy, economic inequality, and poverty: the United States in comparative perspective. Social Science Quarterly 86, 955–983.
Stangor, C., Lynch, L., Duan, C., Glass, B., 1992. Categorization of individuals on the basis of multiple social features. Journal of Personality and Social Psychology 62, 207–218.
Telles, E.E., Murguía, E., 1992. The continuing significance of phenotype among Mexican Americans. Social Science Quarterly 73, 120–122.
Tilly, C., 1998. Durable Inequality. University of California Press, Berkeley, CA.
Tsui, B., 2009. American Chinatown: A People's History of Five Neighborhoods. New Press, New York.
Western, B., 2006. Punishment and Inequality in America. Russell Sage Foundation, New York.
Wolff, E.N., 2010. Recent Trends in Household Wealth in the United States: Rising Debt and the Middle-class Squeeze-an Update to 2007. Working Paper No. 589. Levy Economics Institute of Bard College, NY. http://www.levyinstitute.org/pubs/wp_589.pdf (accessed 28.12.13.).
Zajonc, R.B., 1998. Emotions. In: Gilbert, D.T., Fiske, S.T., Lindzey, G. (Eds.), The Handbook of Social Psychology. McGraw-Hill, Boston, MA.
Zolberg, A.R., 2006. A Nation by Design: Immigration Policy in the Fashioning of America. Russell Sage Foundation, New York.

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 12

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser—Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 12

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 12 *xlv*

VOLUME 12

Infancy and Childhood: Emotional Development 1
Ross A Thompson

Infancy and Human Development 7
Marc H Bornstein

Infant and Child Mortality in Industrialized Countries 14
Markéta Pechholdová

Infant and Child Mortality in the Less Developed World 21
Magali Barbieri

Infant Gaze Behaviors 27
Rechele Brooks

Infectious Diseases: Psychosocial Aspects 32
Andrew Steptoe and Lydia Poole

Inferences in Discourse, Psychology of 37
Leo GM Noordman and Wietske Vonk

Infertility: Demographic Aspects 45
Marida Hollos

Informal and Underground Economics 50
Bruno S Frey and Friedrich Schneider

Informal Sector 56
Karen T Hansen

Informal Work in Developing Countries 60
Arnab K Basu and Nancy H Chau

Information and Communications Technology in Education 66
Saïd Assar

Information and Knowledge, Organizational 72
Patrizia Porrini and William H Starbuck

Information Processing Architectures: Fundamental Issues 77
James T Townsend, Haiyuan Yang, and Trisha Van Zandt

Information Society 83
Alistair S Duff

Information Society, Geography of 90
Taylor Shelton and Matthew Zook

Information Structure in Linguistics 95
Dejan Matić

Information, Economics of 100
SS Lippman and John J McCall

Information Theory 106
Laura Martignon

Informed Consent of Research Subjects 110
Paul S Appelbaum

Infrastructure Investment 116
David C Perry

Infrastructures of Society, Anthropology of 119
Jörg Niewöhner

In-House Counsel 126
Robert E Rosen

Injuries and Accidents: Psychosocial Aspects 130
Sharon Clarke

Injustice: Legal Aspects 135
Austin Sarat

Inner City 140
William J Wilson, James M Quane, and Bruce H Rankin

Innovation 145
Roy Green, Renu Agarwal, and Danielle Logue

Innovation and Employment 152
Marco Vivarelli

Innovation and Technological Change, Economics of 160
Wesley M Cohen

Innovation, Theory of 169
Jon Sundbo

Instinct, History of 175
Daniel R Huebner

Institutionalism 181
Sven Steinmo

Institutionalism and Neo-institutionalism: History of the Concepts 186
Markus Hadler

Institutions 190
Thomas R Voss

Instructional Design 196
Franz Schott and Norbert M Seel

Instructional Psychology 201
Erik De Corte

Instrumental Variables in Statistics and Econometrics 205
James H Stock

Insurance 210
Christian Gollier

Insurance and the Law 215
Shauhin A Talesh

Insurgency 221
Erica Chenoweth and Jonathan Pinckney

Integrated Attachment Theory 228
Ella Levert-Levitt and Abraham Sagi-Schwartz

Integration and Multiculturalism: Focus on Western Europe 235
Tariq Modood

Integration, Social 243
Richard Münch

Intellectual Functioning, Assessment of 249
Leslie Atkinson

Intellectual History 255
Cesare Cuttica

Intellectual Property Rights: Ethical Aspects 262
Joseph Straus

Intellectual Property: Law in Context 266
Christoph B Graber and Jessica C Lai

Intellectuals: History of the Concept 273
Christophe Charle

Intellectuals, Sociology of 277
Robert J Brym

Intelligence: Assessments of 283
Adam W McCrimmon, Emma A Climie, and Donald H Saklofske

Intelligence: Central Conceptions and Psychometric Models 290
Florian Schmiedek and Shu-Chen Li

Intelligence, Genetics of: Cognitive Abilities 297
Frank M Spinath and Juliana Gottschling

Intelligence: Historical and Conceptual Perspectives 303
Robert J Sternberg

Intelligence: History of the Concept 309
John Carson

Intelligence in Cultural, Social and Educational Context 313
Andreas Demetriou

Intelligence, Prior Knowledge, and Learning 323
Elsbeth Stern

Intensification and Specialization, Archaeology of 329
Kathleen D Morrison

Intentionality and Rationality 333
David Braddon-Mitchell

Intentionality during Infancy and Early Childhood, Development of 340
Dankert Vedeler

Intentionality in Language and Communication, Emergence of 347
Aliyah Morgenstern

Interactionism, Symbolic 352
Alex Dennis and Greg Smith

Intercultural Communication: An Anthropological Perspective 357
Pál Nyíri and Joana Breidenbach

Intercultural Legal Discourse: Philosophical Aspects 362
Otfried Höffe

Interest Groups 368
Kay Lehman Schlozman

Interest Groups, History of 371
Hans-Jürgen Puhle

Interest, Psychology of 378
K Ann Renninger and Rose K Pozos-Brewer

Interest: History of the Concept 386
Johan Heilbron

Interests in Sociological Analysis 391
Pierre Demeulenaere

Intergenerational Justice 396
Paul Kelly

Intergenerational Solidarity 402
Valeria Bordone

Intergroup Relations 409
Jorge Vala and Rui Costa-Lopes

Interhemispheric Interaction in Language and Cognitive Processes 416
Dimitrios S Kasselimis and Andreas Nidos

Internal/External Frame of Reference Model 425
Herbert W Marsh, Jens Möller, Philip Parker, Man K Xu, Benjamin Nagengast, and Reinhard Pekrun

Internal Migration: Developing Countries 433
Randall Kuhn

Internal Migration in Industrialized Countries 443
Michael J Greenwood

Internal Reconstruction in Linguistics 447
Anthony Fox

Internal Validity 450
Charles S Reichardt

International Advertising 455
C Samuel Craig

International and Comparative Criminology 462
Amy Nivette

International Arbitration 470
Bryant G Garth

International Association of Schools of Social Work 474
Vimla V Nadkarni and Tetyana Semigina

International Business 480
Mauro F Guillén

International Business: Geographic Aspects 484
Sharmistha Bagchi-Sen and Torsten Schunder

International Communication: History 490
Daniel Headrick

International Council on Social Welfare 495
Sergei Zelenev

International Issues in Child Protection 501
Jonathan Parker

International Law and Treaties 507
Mark W Janis

International Migration 511
Giovanni Facchini, Tommaso Frattini, and Anna M Mayda

International Migration and Remittances 519
Dean Yang

International Migration by Ethnic Chinese 525
Emmanuel Ma Mung

International Migration by Ethnic South and Southeast Asians 530
Graeme Hugo

International Organisations and Education 538
Pradeep A Dhillon

International Social Problems: A Systems Perspective 542
Heather K Manion

International Social Survey Programme 550
Tom W Smith

International Social Work 557
Bala R Nikku

International Social Work Student Exchanges 562
Rosemary J Link

International Trade: Commercial Policy and Trade Negotiations 568
Andreas Dür

International Trade: Economic Integration 574
Anthony Venables

Internationalization: Patterns in Business Firms 580
Ivo Zander

Internet and Culture 587
Marc Verboord and Susanne Janssen

Internet and Privacy 593
Andreas Busch

Internet and Social Media: Anthropological Aspects 600
Paula Uimonen

Internet Research: Ethical Concerns 606
Annette N Markham and Elizabeth A Buchanan

Interoception 614
André Schulz

Interorganizational Relationships and Networks 621
Mark Ebers

Interpersonal Attraction, Psychology of 626
Shelly Zhou, Marie B Chelberg, and Arthur Aron

Interpersonal Psychotherapy 631
Denise E Wilfley and Allison L Shore

Interpersonal Trust across the Lifespan 637
Ken J Rotenberg

Interpretation in Anthropology 641
Johannes Fabian

Interpretive Methods: Macromethods 645
R Stryker

Interpretive Methods: Micromethods 648
Norman K Denzin

Interrupted Time Series Quasi-Experimental Designs 652
Richard McCleary

Intersectionality 657
Stephanie G Guittar and Nicholas A Guittar

Intersectionality and Crime 663
Brenda S Blackwell and Jennifer Cruze

Interviewing, History of 670
Raymond M Lee

Intimate Partner Abuse, Applied Research on 677
Jordana N Navarro

Intimate Partner Violence 685
Jennifer T Buntin

Intracarotid Sodium Amytal Test and Language 689
Caleb M Pearson and Alexander I Tröster

Intrinsic Motivation: A Cultural Perspective 696
Qian Wang

Introspection: History of the Concept 702
Kurt Danziger

Invariance in Statistics 705
Morris L Eaton

Ion Channels and Molecular Events in Neuronal Activity 710
Bryan Kolb

IPAT (Impact, Population, Affluence, and Technology) 716
Geoffrey McNicoll

Irrationality: Philosophical Aspects 719
Richard Samuels and Stephen P Stich

Irrigation Societies 724
John S Lansing and Florence Durney

Islam: Asia 727
Mark R Woodward

Islam: Middle East 731
Charles Lindholm

Islam: Sub-Saharan Africa 737
Peter B Clarke

Islamic Fundamentalism 741
Youssef M Choueiri

Islamic Law 746
Lena Salaymeh

Islamic Social Work: Hope and the Commitment to Change 754
Nada Eltaiba

Isms and History 759
Cesare Cuttica

Issue Evolution 765
Edward G Carmines

Issue Networks: Iron Triangles, Subgovernments, Policy Communities, Policy Networks 769
Mark Thatcher and Juergen Braunstein

Jackson, John Hughlings (1835–1911) 774
W Ian McDonald and Hans-Joachim Freund

Jacobs, Jane (1916–2006) 776
Mary P Corcoran

Jakobson, Roman (1896–1982) 782
Linda R Waugh

James, William (1842–1910) 786
M Brewster Smith

Janet, Pierre (1859–1947) 791
John I Brooks, III

Japan: Sociocultural Aspects 794
Emiko Ohnuki-Tierney

Jeffreys, Sir Harold (1891–1989) 798
Arnold Zellner

Jewish Diaspora 801
Sergio DellaPergola

Job Analysis and Work Roles, Psychology of 811
Heiner Dunckel

Job Design and Evaluation: Organizational Aspects 816
Alan J Arthurs

Job Displacement Insurance 819
Donald O Parsons

Job Satisfaction 833
Howard M Weiss and Kelsey L Merlo

Joint Action 839
Margaret Gilbert

Joint Attention during Infancy and Early Childhood Across Cultures 844
Jessica L Kinard and Linda R Watson

Journalism 851
Daniel C Hallin

Journalism and Journalists 857
Slavko Splichal

Journals, Scholarly 862
Mark Ware

Judaism 869
Stephen Sharot

Judaism and Gender 875
Laura S Levitt

Judges 879
José J Toharia

Judicial Behavior 885
Michael J Nelson

Judicial Independence 891
Michael J Nelson

Judicial Review 897
Jesse Merriam and Joel B Grossman

Jung, Carl Gustav (1875–1961) 902
Angela Graf-Nold

Juries 907
Shari S Diamond

Just City 913
Susan S Fainstein

Justice, Access to 919
Earl Johnson

Justice and Health 928
Johannes Siegrist

Justice and Law 932
David Luban

Justice, Conflicts, and the Justice of Conflict Resolution 937
Leo Montada

Justice: Philosophical Aspects 943
Thomas Pogge

Justice: Social Psychological Perspectives 949
Ali Kazemi, Kjell Törnblom, and Gerold Mikula

Juvenile Justice: International Law and Children's Rights 956
Barry Goldson and John Muncie

This page intentionally left blank

CONTRIBUTORS TO VOLUME 12

Renu Agarwal
Management Discipline Group, UTS Business School, Ultimo, NSW, Australia

Paul S. Appelbaum
Columbia University, New York, NY, USA; and New York State Psychiatric Institute, New York, NY, USA

Arthur Aron
Stony Brook University, New York, NY, USA

Alan J. Arthurs
University of Bath, Bath, UK

Saïd Assar
Telecom Ecole de Managment, Institut Mines-Telecom, Evry, France

Leslie Atkinson
Ryerson University, Toronto, ON, Canada

Sharmistha Bagchi-Sen
State University of New York — Buffalo, Buffalo, NY, USA

Magali Barbieri
Institut National d'Études Démographiques, Paris, France

Arnab K. Basu
Charles H. Dyson School of Applied Economics and Management, Ithaca, NY, USA; Center for Development Research (ZEF), Bonn, Germany; and Institute for the Study of Labor (IZA), Bonn, Germany

Brenda S. Blackwell
Georgia Southern University, Statesboro, GA, USA

Valeria Bordone
Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/ÖAW, WU), International Institute for Applied Systems Analysis, Laxenburg, Austria

Marc H. Bornstein
Eunice Kennedy Shriver *National Institute of Child Health and Human Development, National Institutes of Health, Public Health Service, Bethesda, MD, USA*

David Braddon-Mitchell
University of Sydney, NSW, Australia

Juergen Braunstein
London School of Economics, London, UK

Joana Breidenbach
Betterplace Lab, Berlin, Germany

John I. Brooks
Fayetteville State University, Fayetteville, NC, USA

Rechele Brooks
Institute of Learning & Brain Sciences, University of Washington, Seattle, WA, USA

Robert J. Brym
University of Toronto, Toronto, ON, Canada

Elizabeth A. Buchanan
Center for Applied Ethics, University of Wisconsin-Stout, Menomonie, WI, USA

Jennifer T. Buntin
Lewis University, Romeoville, IL, USA; and One University Parkway, Romeoville, IL, USA

Andreas Busch
Georg-August-Universität Göttingen, Göttingen, Germany

Edward G. Carmines
Indiana University, Bloomington, IN, USA

John Carson
University of Michigan, Ann Arbor, MI, USA

Christophe Charle
Professeur d'histoire contemporaine à l'Université de Paris I, Institut d'Histoire Moderne et contemporaine, Paris, France

Nancy H. Chau
Charles H. Dyson School of Applied Economics and Management, Ithaca, NY, USA; Center for Development Research (ZEF), Bonn, Germany; and Institute for the Study of Labor (IZA), Bonn, Germany

Marie B. Chelberg
Stony Brook University, New York, NY, USA

Erica Chenoweth
Josef Korbel School of International Studies, University of Denver, Denver, CO, USA

Youssef M. Choueiri
School of Arts, Languages and Cultures, University of Manchester, Manchester, UK

Peter B. Clarke
King's College, London, UK

Sharon Clarke
University of Manchester, Manchester, UK

Emma A. Climie
Werklund School of Education, University of Calgary, Calgary, AB, Canada

Wesley M. Cohen
Duke University, Durham, NC, USA

Mary P. Corcoran
National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland

Rui Costa-Lopes
Institute of Social Sciences — University of Lisbon, Lisboa, Portugal

C. Samuel Craig
Stern School of Business, New York University, New York, NY, USA

Jennifer Cruze
Georgia State University, Atlanta, GA, USA

Cesare Cuttica
Université Paris 8-Vincennes, Saint-Denis, France

Kurt Danziger
York University, Toronto, ON, Canada

Erik De Corte
University of Leuven, Leuven, Belgium

Sergio DellaPergola
The A. Harman Institute of Contemporary Jewry, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem, Israel

Andreas Demetriou
University of Nicosia Research Foundation, University of Nicosia, Nicosia, Cyprus

Pierre Demeulenaere
Université de Paris-Sorbonne, Paris, France

Alex Dennis
University of Sheffield, Sheffield, UK

Norman K. Denzin
University of Illinois at Urbana-Champaign, IL, USA

Pradeep A. Dhillon
University of Illinois, Urbana-Champaign, IL, USA

Shari S. Diamond
Northwestern University School of Law, Chicago, IL, USA; and American Bar Foundation, Chicago, IL, USA

Alistair S. Duff
Edinburgh Napier University, Edinburgh, Scotland, UK

Heiner Dunckel
International Institute of Management, University of Flensburg, Flensburg, Germany

Andreas Dür
University of Salzburg, Salzburg, Austria

Florence Durney
School of Anthropology, University of Arizona, Tucson, AZ, USA

Morris L. Eaton
University of Minnesota, Minneapolis, MN, USA

Mark Ebers
Faculty of Management, Economics and Social Sciences, University of Cologne, Koeln, Germany

Nada Eltaiba
The University of Western Australia, Perth, WA, Australia

Johannes Fabian
University of Amsterdam, Amsterdam, The Netherlands

Giovanni Facchini
University of Nottingham, Nottingham, UK; and University of Milan, CEPR, CES-Ifo, CReAM, IZA and LdA, Milan, Italy

Susan S. Fainstein
Harvard University, Cambridge, MA, USA

Anthony Fox
University of Leeds, Leeds, UK

Tommaso Frattini
University of Milan, CReAM, IZA and LdA, Milan, Italy

Hans-Joachim Freund
Heinrich Heine University, Dusseldorf, Germany

Bruno S. Frey
University of Warwick, Coventry, UK; and CREMA (Center for Research and Economics Management and Arts), Zurich, Switzerland

Bryant G. Garth
University of California, Irvine, CA, USA

Margaret Gilbert
University of California, Irvine, CA, USA

Barry Goldson
The University of Liverpool, Liverpool, UK

Christian Gollier
University of Toulouse, Toulouse, France

Juliana Gottschling
Saarland University, Saarbruecken, Germany

Christoph B. Graber
University of Lucerne, Lucerne, Switzerland

Angela Graf-Nold
Zurich, Switzerland

Roy Green
UTS Business School, Ultimo, NSW, Australia

Michael J. Greenwood
University of Colorado, Boulder, CO, USA

Joel B. Grossman
Johns Hopkins University, Baltimore, MD, USA

Mauro F. Guillén
The Wharton School of the University of Pennsylvania, Philadelphia, PA, USA

Nicholas A. Guittar
Valdosta State University, Valdosta, GA, USA

Stephanie G Guittar
Valdosta State University, Valdosta, GA, USA

Markus Hadler
Macquarie University, Sydney, NSW, Australia

Daniel C. Hallin
University of California, San Diego, La Jolla, CA, USA

Karen T. Hansen
Northwestern University, Evanston, IL, USA

Daniel Headrick
Roosevelt University, Chicago, IL, USA

Johan Heilbron
Centre Européen de Sociologie et de Science Politique de la Sorbonne (CESSP-CNRS-EHESS), Paris, France; and Erasmus University, Rotterdam, The Netherlands

Otfried Höffe
Universität Tübingen, Tübingen, Germany

Marida Hollos
Brown University, Providence, RI, USA

Daniel R. Huebner
University of North Carolina at Greensboro, Greensboro, NC, USA

Graeme Hugo
The University of Adelaide, Adelaide, SA, Australia

Mark W. Janis
University of Connecticut, Hartford, CT, USA; and University of Oxford, Oxford, UK

Susanne Janssen
Erasmus University Rotterdam, Rotterdam, The Netherlands

Earl Johnson
California Court of Appeal, CA, USA; and University of Southern California Law School, Los Angeles, CA, USA

Dimitrios S. Kasselimis
University of Crete, Rethymno, Greece; and National and Kapodistrian University of Athens, Greece

Ali Kazemi
University of Skövde, Skövde, Sweden

Paul Kelly
London School of Economics, London, UK

Jessica L. Kinard
The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Bryan Kolb
Department of Neuroscience, University of Lethbridge, Lethbridge, AB, Canada

Randall Kuhn
Josef Korbel School of International Studies, Denver, CO, USA

Jessica C. Lai
University of Lucerne, Lucerne, Switzerland

John S. Lansing
School of Anthropology, University of Arizona, Tucson, AZ, USA; Santa Fe Institute, NM, USA; and Stockholm Resilience Center, Stockholm, Sweden

Raymond M. Lee
Royal Holloway University of London, Surrey, UK

Ella Levert-Levitt
Center for the Study of Child Development, University of Haifa, Haifa, Israel

Laura S. Levitt
Temple University, Philadelphia, PA, USA

Shu-Chen Li
TU Dresden, Dresden, Germany

Charles Lindholm
Boston University, Boston, MA, USA

Rosemary J. Link
Simpson College, Iowa, IA, USA

S.S. Lippman
University of California, Los Angeles, CA, USA

Danielle Logue
Management Discipline Group, UTS Business School, Ultimo, NSW, Australia

David Luban
Georgetown University Law Center, Washington, DC, USA

Emmanuel Ma Mung
CNRS, University of Poitiers, Poitiers, France

Heather K. Manion
Royal Roads University, Victoria, BC, Canada

Annette N. Markham
Aarhus University, Aarhus N, Denmark

Herbert W. Marsh
Australian Catholic University, Strathfield, NSW, Australia; King Saud University, Riyadh, Saudi Arabia; and University of Oxford, Oxford, UK

Laura Martignon
Ludwigsburg University of Education, Ludwigsburg, Germany

Dejan Matić
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Anna M. Mayda
Georgetown University, CEPR, CReAM, IZA and LdA, Washington, DC, USA

John J. McCall
University of California, Los Angeles, CA, USA

Richard McCleary
University of California, Irvine, Irvine, CA, USA

Adam W. McCrimmon
Werklund School of Education, University of Calgary, Calgary, AB, Canada

W. Ian McDonald[†]
Royal College of Physicians, London, UK

Geoffrey McNicoll
Population Council, New York, NY, USA

Kelsey L. Merlo
Georgia Institute of Technology, Atlanta, GA, USA

Jesse Merriam
Johns Hopkins University, Baltimore, MD, USA

Gerold Mikula
University of Graz, Graz, Austria

Tariq Modood
University of Bristol, Bristol, UK

Jens Möller
University of Kiel, Kiel, Germany

Leo Montada
University of Trier, Mozartstr, Germany

Aliyah Morgenstern
Sorbonne Nouvelle University, Paris, France

Kathleen D. Morrison
University of Chicago, Chicago, IL, USA

Richard Münch
Otto-Friedrich-Universität Bamberg, Bamberg, Germany

John Muncie
The Open University, Milton Keynes, UK

Vimla V. Nadkarni
International Association of Schools of Social Work, School of Social Work, Tata Institute of Social Sciences, Deonar, Mumbai, India

Benjamin Nagengast
University of Tübingen, Tübingen, Germany

Jordana N. Navarro
Cookeville, Tennessee Technological University, TN, USA

Michael J. Nelson
The Pennsylvania State University, University Park, PA, USA

Andreas Nidos
Department of Neurology, Metropolitan Hospital, Neo Faliro, Greece

Jörg Niewöhner
Humboldt University Berlin, Berlin, Germany

Bala R. Nikku
School of Social Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia; and Nepal School of Social Work, Kathmandu, Nepal

Amy Nivette
Nuffield College, Oxford, UK

Leo G.M. Noordman
Tilburg University, Tilburg, The Netherlands

[†] Deceased.

Pál Nyíri
Vrije Universiteit, Amsterdam, The Netherlands

Emiko Ohnuki-Tierney
University of Wisconsin, Madison, WI, USA

Jonathan Parker
Bournemouth University, Bournemouth, UK

Philip Parker
Australian Catholic University, Strathfield, NSW, Australia

Donald O. Parsons
George Washington University and IZA Bonn, Washington, DC, USA

Caleb M. Pearson
Barrow Neurological Institute, Phoenix, AZ, USA

Markéta Pechholdová
University of Economics, Prague, Prague, Czech Republic

Reinhard Pekrun
University of Munich, Munich, Germany

David C. Perry
University of Illinois, Chicago, IL, USA

Jonathan Pinckney
Josef Korbel School of International Studies, University of Denver, Denver, CO, USA

Thomas Pogge
Yale University, New Haven, CT, USA

Lydia Poole
University College London, London, UK

Patrizia Porrini
Long Island University, Brookville, NY, USA

Rose K. Pozos-Brewer
Swarthmore College, Swarthmore, PA, USA

Hans-Jürgen Puhle
Institut für Politikwissenschaft, Goethe-Universität Frankfurt am Main, Frankfurt am Main, Germany

James M. Quane
Harvard University, Cambridge, MA, USA

Bruce H. Rankin
Koç University, Sarõyer, Istanbul, Turkey

Charles S. Reichardt
University of Denver, Denver, CO, USA

K. Ann Renninger
Swarthmore College, Swarthmore, PA, USA

Robert E. Rosen
School of Law, University of Miami, Coral Gables, FL, USA

Ken J. Rotenberg
School of Psychology, Keele University, Staffordshire, UK

Abraham Sagi-Schwartz
Center for the Study of Child Development, University of Haifa, Haifa, Israel

Donald H. Saklofske
Social Science Centre, University of Western Ontario, London, ON, Canada

Lena Salaymeh
University of California (UC) Berkeley School of Law, Berkeley, CA, USA

Richard Samuels
The Ohio State University, Columbus, OH, USA

Austin Sarat
Jurisprudence, and Social Thought, Amherst College, Amherst, MA, USA

Kay Lehman Schlozman
Boston College, Chestnut Hill, MA, USA

Florian Schmiedek
German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany

Friedrich Schneider
Johannes Kepler University, Linz, Austria

Franz Schott
Institute for Educational and Developmental Psychology, Technische Universitaet Dresden, Dresden, Germany

André Schulz
University of Luxembourg, Walferdange, Luxembourg

Torsten Schunder
State University of New York — Buffalo, Buffalo, NY, USA

Norbert M. Seel
University of Freiburg, Freiburg i.Br., Germany

Tetyana Semigina
School of Social Work, National University of "Kyiv-Mohyla Academy", Kyiv, Ukraine

Stephen Sharot
Ben-Gurion University of the Negev, Beer-Sheva, Israel

Taylor Shelton
Clark University, Worcester, MA, USA

Allison L. Shore
School of Medicine, Washington University, St. Louis, MO, USA

Johannes Siegrist
Heinrich-Heine-University of Duesseldorf, Duesseldorf, Germany

Greg Smith
School of Nursing, Midwifery, Social Work and Social Sciences, University of Salford, Salford, UK

M. Brewster Smith
University of California, Santa Cruz, CA, USA

Tom W. Smith
NORC at the University of Chicago, Chicago, IL, USA

Frank M. Spinath
Saarland University, Saarbruecken, Germany

Slavko Splichal
University of Ljubljana, Ljubljana, Slovenia

William H. Starbuck
University of Oregon, Eugene, OR, USA

Sven Steinmo
European University Institute, Fiesole, Italy

Andrew Steptoe
University College London, London, UK

Elsbeth Stern
ETH Zürich, Zürich, Switzerland

Robert J. Sternberg
Office of Academic Affairs, Oklahoma State University, Stillwater, OK, USA

Stephen P. Stich
Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

James H. Stock
Harvard University, Cambridge, MA, USA

Joseph Straus
Max Planck Institute for Innovation and Competition, Munich, Germany; and NIPMO-UNISA Chair for Intellectual Property, University of South Africa (UNISA), Pretoria, South Africa

R. Stryker
University of Minnesota, Minneapolis, MI, USA

Jon Sundbo
Roskilde University, Roskilde, Denmark

Shauhin A. Talesh
University of California Irvine, Irvine, CA, USA

Mark Thatcher
London School of Economics, London, UK

Ross A. Thompson
University of California Davis, Davis, CA, USA

José J. Toharia
Metroscopia, Madrid, Spain

Kjell Törnblom
ETH Zurich, Zurich, Switzerland

James T. Townsend
Indiana University, Bloomington, IN, USA

Alexander I. Tröster
Barrow Neurological Institute, Phoenix, AZ, USA; and Barrow Center for Neuromodulation, Phoenix, AZ, USA

Paula Uimonen
Stockholm University, Stockholm, Sweden

Jorge Vala
Institute of Social Sciences — University of Lisbon, Lisboa, Portugal

Trisha Van Zandt
Ohio State University, Columbus, OH, USA

Dankert Vedeler
Norwegian University of Science and Technology (NTNU), Trondheim, Norway

Anthony Venables
London School of Economics, London, UK

Marc Verboord
Erasmus University Rotterdam, Rotterdam, The Netherlands

Marco Vivarelli
Università Cattolica del Sacro Cuore, Piacenza, Italy; IZA, Bonn, Germany; and University of Sussex, Brighton, UK

Wietske Vonk
Max Planck Institute for Psycholinguistics and Radboud University Nijmegen, Nijmegen, The Netherlands

Thomas R. Voss
University of Leipzig, Leipzig, Germany

Qian Wang
The Chinese University of Hong Kong, Hong Kong SAR, China

Mark Ware
Mark Ware Consulting, Bristol, UK

Linda R. Watson
The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Linda R. Waugh
University of Arizona, Tucson, AZ, USA

Howard M. Weiss
Georgia Institute of Technology, Atlanta, GA, USA

Denise E. Wilfley
School of Medicine, Washington University, St. Louis, MO, USA

William J. Wilson
Harvard University, Cambridge, MA, USA

Mark R. Woodward
Arizona State University, Tempe, AZ, USA

Man K. Xu
University of Cambridge, Cambridge, UK

Dean Yang
Department of Economics, Gerald R. Ford School of Public Policy, University of Michigan, Ann Arbor, MI, USA

Haiyuan Yang
Indiana University, Bloomington, IN, USA

Ivo Zander
Uppsala University, Uppsala, Sweden

Sergei Zelenev
International Council on Social Welfare, Bronx, NY, USA

Arnold Zellner
University of Chicago, Chicago, IL, USA

Shelly Zhou
Stony Brook University, NY, USA; University of Toronto, Toronto, ON, Canada

Matthew Zook
University of Kentucky, Lexington, KY, USA

This page intentionally left blank

Infancy and Childhood: Emotional Development

Ross A Thompson, University of California Davis, Davis, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R.A. Thompson, volume 11, pp. 7382–7387, © 2001, Elsevier Ltd.

Abstract

Emotional development is a uniquely integrative and psychologically constructive feature of psychological growth in infancy and childhood. The role of emotions in behavior and development, and the nature of emotion itself, are discussed in relation to structuralist and functionalist approaches to emotional development. The focus then turns to considering 'what is emotional development the development of?' and the contributions of psychobiological and neurobiological growth, perceptual processes, emotion understanding, the growth of self-awareness and self-understanding, understanding of emotional display rules, and the development of emotion regulation are considered. The article concludes with consideration of the importance of social relationships to early emotional development and future directions for research in this field.

Emotion is a familiar but psychologically complex human experience. Consequently, the growth of emotion integrates diverse developmental processes in infancy and childhood, including psychobiological and neurobiological growth, perceptual processes, emotional and social understanding, self-awareness and self-understanding, and the growth of self-regulation. The study of emotional development has also been stimulated by basic questions concerning the nature of emotion itself and its role in behavior and development. This article explores the issues that have contributed to considerable research interest in emotional growth early in life.

The Nature of Emotional Development

Emotions color the life experience of people of all ages. For infants and children, emotions have especially powerful motivational consequences because emotion understanding is rudimentary, capacities for emotion self-regulation are limited, and the young child's engagement or withdrawal from events is primarily emotional in quality. Emotional development is central to understanding the growth of social competence, intellectual achievement, self-understanding, and the earliest origins of mental health or psychopathology because of the importance of emotion to the organization and regulation of behavior. Emotional development provides a window into the psychological growth of the child as, from infancy to late childhood, emotions become integrated into the broader fabric of behavioral competence, become objects of thought and analysis, can be strategically expressed or concealed in social situations, are shaped by the child's experience of significant relationships, motivate prosocial as well as aggressive tendencies, become part of self-understanding (especially in feelings of pride and shame), and color the individuality of personality.

Emotions and Development

Emotional development was vigorously studied earlier in the twentieth century and has received renewed attention owing to several new theoretical approaches to understanding emotional development. First, emotions are regarded as centrally important to understanding development in infancy and childhood, owing in part to how emotions provide organization and motivation to early developmental achievements. Young children's experience of important social relationships, their emergent understanding of themselves, their motivation to achieve and their responses to success and failure, their understanding of cultural values, and many other central facets of early development are infused with emotion. Second, emotional development offers a uniquely integrative perspective on human development. Understanding emotional growth requires synthesizing biological perspectives on development (including neurobiological growth and neuroendocrine functioning) with social perspectives (including relational and family influences) and the child's developing representations of emotion (including understanding the causes and consequences of emotions, and of cultural norms). Third, emotions are viewed as having constructive as well as disorganizing features. Although emotions are capable of undermining effective functioning (which accounts for current scientific interest in the growth of emotion regulation), developmental researchers also emphasize how emotions motivate and organize young children's understanding of the social world, appraisals of the self, judgments of approved and disapproved conduct, engagement in social conflict resolution, and many other developing skills and capabilities. This is consistent with a new appreciation for the constructive, organizational role of emotions in other areas of psychological study.

Contemporary study of emotional development in infancy and childhood is also motivated by clinical concerns – most notably, a growing realization of the early emergence of emotionally related disorders in young children, including depression, anxiety disorders, and behavioral and conduct problems (Luby, 2006). Thus the study of emotional development is informed by the field of developmental psychopathology and studies of the origins of affective disorders in troubled parent–child relationships, temperamental vulnerability, environmental and genetic risk factors, and individual capacities for emotion self-regulation (e.g., Vasey and Dadds, 2000). This research further underscores the complex, integrative quality of emotional growth and the various risks that can exist for emotional dysfunction, as well as alternative avenues to prevention and remediation.

Developmental research on emotions has required a broad range of methodologies. These include fine-grained analysis of videotaped facial expressions, naturalistic observations of social interaction at home or school, conversations with young children about their comprehension of emotional events, detailed study of the responses of children to emotionally evocative laboratory vignettes, parent-report measures, psychobiological measures of heart rate or cortisol, or brain imaging during emotionally arousing situations. Developmental scientists are increasingly integrating these methodological strategies in broadly based longitudinal research that enables them to study the emergence of emotional capacities over time and in the context of family processes, psychobiological maturation, and other influences.

What Is Emotion?

Contemporary study of emotion in behavior and development is accompanied by debate over the nature of emotion itself (Izard, 2010). Two broad perspectives can be distinguished. The structuralist perspective defines emotion in terms of discrete emotional states, each with unique patterns of subjective feeling, cognitive appraisal, physiological arousal, and facial expression (Izard and Ackerman, 2000). These emotion components are tightly interconnected and are deeply rooted in human evolution because of their relevance to reproductive success (e.g., feelings of anger involve visceral, perceptual, subjective, and expressive components relevant to self-defense). This view is consistent with the everyday tendency to distinguish basic emotions like anger, fear, sadness, and joy as readily dissociable states.

Other theorists advocating functionalist perspectives to emotional development have questioned whether emotional life is so discretely packaged and have offered an alternative view (Saarni et al., 2006). In this perspective, emotion is defined in terms of the quality of the transactions between the person and the environment on matters of significance to the individual. In this regard, what defines an emotion is not a network of internal subjective, physiological, and other components, but rather the constellation of a person's goals, appraisals, action tendencies, and other capacities in relation to environmental inventives, obstacles, and opportunities. As a result, emotional experience is not confined to a predefined set of emotion categories, but can encompass the rich variety of subtly nuanced emotional blends characteristic of everyday experience, as well as the breadth of how emotion is conceptualized in diverse sociocultural groups based on how specific person–environment transactions are conceived.

A functionalist approach is also well suited to developmental analysis. One way that emotions develop, for example, is in how the generalized arousal conditions of early infancy mature into the more specific, organized, and refined emotional states apparent in older children and adults. Young infants do not necessarily experience or express emotions in ways that fit into the discrete emotion categories by which adults conceptualize emotion, but appear instead to experience more diffuse affective systems that become progressively better differentiated as a result of early social experiences and socialization, cognitive growth, and psychobiological maturation. Studying this developmental process requires methodological creativity and breadth, as functionalist theorists suggest. Rather than relying exclusively on facial expressions as primary indicators of underlying subjectivity, for example, developmental researchers increasingly index emotion multimodally, integrating facial, vocal, and other behavioral indicators into an organizational interpretation in relation to the child's interaction with environmental opportunities and challenges. However, the functionalist perspective faces the challenge of a conceptualization of emotion that may be unduly inclusive: it is not entirely clear how emotion differs from other motivated incentives, such as needs and desires. Perhaps for this reason, contemporary emotion researchers have continued to embrace the organized portrayal of emotion coming from structuralist theory, while enlisting the functionalist understanding of these emotions.

What is Emotional Development the Development of?

The primary challenge of developmental analysis is understanding why developmental changes in emotion occur, and how emotional growth is associated with concurrent advances in social, cognitive, linguistic, and physical development. Exploring the various components of emotional development reveals how emotional experience integrates significant aspects of psychological development in infancy and childhood.

Neurobiological and Psychobiological Growth

Emotions are biologically basic features of human functioning. They are predicated on widely distributed and richly interconnected brain structures and hormonal influences that organize the interplay of activation and regulation that constitutes emotional reactions. The young infant's capacity to exhibit distress, fear, surprise, happiness, and rage reflects the early functioning of these biologically deep-rooted neurological and psychobiological systems. Emotional development is based, therefore, on neurophysiological, neuroendocrine, and other biological processes that emerge early and develop over an extended period in childhood (LeDoux, 2000; Panksepp and Biven, 2012). For example, the developing stability of sympathetic and parasympathetic nervous system functioning in infancy helps to account for the growing predictability, manageability, and environmental responsiveness of the infant's emotional responding (Gunnar and Davis, 2003). Later in childhood, as higher brain regions (especially in the prefrontal neocortex) progressively mature and become integrated with early developing regions of the limbic system and other subcortical structures, emotional life becomes increasingly characterized by more acute emotion appraisals, complex emotional blends, and growing capacities for emotion self-regulation. Because an apparently simple emotional reaction is actually a surprisingly complex psychobiological and neurobiological event, developmental changes in multiple brain regions, hormonal functioning, and neurotransmitter regulation are involved in emotional growth (*see* Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences).

It is important to recognize, however, that lower brain regions (such as the limbic system) governing emotional

activation and higher regions (such as the prefrontal cortex) associated with emotion regulation are typically coactive in any circumstance in which an infant or adult experiences an emotional response. This has two important implications (Thompson, 2011). First, there is never any situation in which emotion is 'unregulated' because regulatory systems are constantly coactive with activation systems. Second, emotional experience arises not only from the 'top-down' regulation of neurobiologically lower activation processes by higher regulatory centers, but also by the 'bottom-up' influence of lower systems on higher cortical systems. The latter can occur, for example, as immediate emotion appraisals arising from limbic structures feedforward to influence the processing of emotional circumstances by cortical systems and their regulatory influence.

Perceiving and Responding to Emotions in Others

Throughout life, emotions are evoked by one's perception of emotion in others. Infants are capable of discerning the emotional meaning in an adult's facial and vocal expressions before the end of the first year, and there is evidence that infants respond resonantly to the sound of another infant's crying even earlier. Moreover, by the end of the first year, infants not only attend to the emotional expressions of adults, but also incorporate the meaning of these expressions into their appraisal of people or circumstances about which the infant is uncertain. This phenomenon is called 'social referencing,' and it is analogous to how people enlist the emotional expressions of others into their assessment of their circumstances (Saarni et al., 2006). In a comparable manner, when a 1-year old encounters an unfamiliar person or object, the sight of a trusted adult's reassuring smile or terrified look (especially if it is accompanied by appropriate vocalizations and other behavior) significantly influences the child's approach or withdrawal tendencies. It appears that the infant 'reads' appropriately the meaning of the adult's emotional expression and its referential relevance to the unfamiliar event.

In the years that follow, advances in emotion understanding significantly expand the young child's capacity to respond appropriately and insightfully to the perception of emotion in others. By the end of the second year, for example, toddlers are more consciously aware of the subjectivity of emotional experience (i.e., another person can feel differently than oneself) and of the causes for others' emotions (Wellman, 2011). This is often manifested when young children are observed comforting a distressed adult or peer, teasing a sibling, during pretend play, or even in humor (Denham, 1998).

In early childhood, moreover, genuinely empathic responses to another's distress can be observed. Initially, young children respond with concerned attention to the sight and sound of another's distress, but empathic responses increase in sophistication and scope in the years following. Early empathy is not reliably accompanied by prosocial initiatives toward the distressed person because empathic arousal is a motivationally complex experience for young children (Zahn-Waxler and Radke-Yarrow, 1990). Young preschoolers have been observed to laugh, tease, become distressed, and even hit a distressed parent, for example, as well as offering comfort. With increasing age and greater understanding of their own efficacy as prosocial agents, empathy becomes somewhat more regularly associated with helping others.

Emotion Understanding

The advances in emotion understanding accompanying conceptual growth enable children to better understand others' emotions, and also their own (Thompson and Lagatutta, 2006). Consistent with growth in their developing 'theory of mind,' for example, 2–3 year olds appreciate how emotion is associated with the satisfaction or frustration of desires (which vary for different people), and 4–5 year olds understand the more complex associations between emotion and thoughts, memories, beliefs, and expectations (Wellman, 2011). These conceptual achievements are important because they contribute to the growing realization that (1) emotions are complexly associated with other mental and motivational states in a person, (2) emotions can sometimes be based on false beliefs or appearances (a peer anticipating a delicious dessert that is actually missing from her lunch bag), and (3) outward emotional appearance does not necessarily reveal underlying emotional reality. People can, in other words, conceal what they are feeling.

In middle childhood, children begin to understand the dynamics of emotional experience, such as how emotional intensity gradually dissipates over time, how particular emotions are related to specific antecedent causes, and how a person's unique background, experiences, and personality can yield unique emotional reactions (Saarni, 1999). By late childhood, children begin to appreciate how multiple emotions can be simultaneously evoked by a single event, such as feeling happy and afraid when performing before a group (Saarni et al., 2006). Despite these conceptual advances, some important features of emotion understanding do not emerge significantly until adolescence, such as an awareness of the psychological causes of emotion in relational experience, the experience of ambivalence or emotional conflict, and a grasp of complex emotional states (e.g., poignancy) (*see* Childhood and Adolescence: Developmental Assets).

Throughout this developmental process, cultural values are integrated into children's emotion understanding and their knowledge of how to interpret, express, and regulate emotional reactions (Cole and Tan, 2007). Because of how emotions are perceived in them and interpreted to young children by parents and others, children from different cultures derive different understandings of the antecedent causes of emotion and the appropriate emotions to experience and express in different social situations. They learn, in a sense, how and what to feel, as well as how and when it is appropriate to express their feelings, and how to understand the feelings of others. In this respect, emotion understanding is part of broader cultural belief systems that children incorporate into their thinking.

Emotion, Self-Awareness, and Self-Understanding

Emotional development is complexly tied to the growth of self-awareness and self-understanding in infancy and childhood. Prior to a dawning awareness of the self as a psychological entity, for example, a young child's emotional repertoire does

not include self-referent emotions such as pride, guilt, shame, and embarrassment. Late in the second and during the third years of life, however, as young preschoolers experience increasing psychological self-awareness (e.g., calling self by name, insisting on 'doing it myself'), these self-referential emotions begin to appear (Lagattuta and Thompson, 2007). As young children begin to see the self as an object of analysis by others, and also by oneself, the simple joy of success becomes pride in achievement, and in response to conspicuous attention, young children increasingly respond with gaze aversion, smiling, and other signs of embarrassment.

Emotional reactions to the self are significantly shaped by the reactions of others to the child. The development of self-referential emotions is advanced not only by growing psychological self-awareness, but also by the young child's increasing understanding of behavioral standards and the ability to apply those standards to an evaluation of personal behavior (Thompson et al., 2006). Guilt derives, for example, from an awareness of how personal behavior has departed from a behavioral standard. Parental reactions are crucial to the young child's dawning awareness and application of these standards of conduct, and indeed the earliest expressions of self-referential emotions commonly occur in contexts where young children can explicitly perceive, or anticipate, parental reactions to their accomplishments or misconduct. With increasing age, children appropriate parental standards that become the child's own internalized expectations for conduct. Thus the manner in which parents convey, reinforce, and attach emotional and relational significance to a young child's achievements or misbehavior is an important influence on the child's emotional life, and developing self-esteem and self-concept.

This underscores the complexity of the association between self-understanding and the growth of self-referential emotions. A dawning psychological self-awareness is necessary, in other words, before self-referential emotions become part of a young child's emotional repertoire. Once these emotions have developed, however, the circumstances in which they are elicited and their relational consequences are significant influences on further growth in self-concept and self-understanding.

Emotion and Emotional Displays

One consequence of young children's developing understanding of the difference between belief and reality is the growing realization that emotional expressions may not accurately reveal underlying emotional reality. This awareness emerges at the same time that children discover the social expectations governing emotional expressions. A person is supposed to show delight when opening a gift in the presence of the gift giver, for example, even if the gift is undesirable. This and other emotional 'display rules' governing the expression of emotions in social situations prescribe how emotions are used to avoid hurting others' feelings, protect self-esteem, and preserve relationships.

Young children do not understand display rules, but they can be observed using them as young as age four. Not until middle childhood do children grasp the nature and purposes of emotional display rules, but once they do so they can enlist them in their interactions with family members, peers, and others (Saarni, 1999). As they do so, children also begin to appreciate the privacy of emotional experience: emotions need not necessarily be conveyed to others if one can manage emotional experience and its expression.

The Development of Emotion Regulation

One of the important differences between infants and adolescents is in their capacities for emotion regulation. Emotion regulation is distinct from emotional display rules because although the latter regulate the *expression* of emotion, the former manages emotional *experience*. People have many reasons for managing their emotions: to feel better under stress, elicit support from others, act courageously, resist bullying, enhance motivation, think better, and for many other reasons. Although the goals underlying emotion regulatory efforts are easily understood in very young children, these goals become increasingly complex and nuanced in the more complex social situations of older children. Competent emotion regulation may be observed, for example, when a child loudly protests a peer's provocations when adults are nearby, but quietly tolerates the bully when adults are absent. However, in conditions of risk for developmental psychopathology, such as family violence or child abuse, the goals underlying a child's emotion regulatory efforts may be complex, mutually inconsistent, and thus, in some senses, self-defeating (Thompson and Goodman, 2010).

In infancy, emotions are primarily regulated by others, especially parents. Parents manage a baby's emotions by directly intervening to soothe the child, and also by regulating the emotional demands of familiar settings like home or child care (often in accord with their perceptions of the child's temperament), altering how the child construes an emotional arousing situation (such as by smiling reassuringly when a friendly, but unfamiliar, adult approaches the child), and later by actively coaching young children on expectations or strategies of emotion management. More broadly, the security that young children derive from their emotional attachments to caregivers is an important resource for emotion regulation because of the confidence it inspires that events will not be uncontrollable and that the adult will intervene helpfully when necessary (Thompson, 2011). Emotion regulation is fostered by others, in other words, not only by their immediate interventions, but also by the reassurance provided by the child's relationship with a caring adult.

From an early age, however, young children are also developing rudimentary skills for managing their own emotional experience (Thompson and Goodman, 2010). This can be observed initially in the comfort-seeking of a distressed infant or toddler, and skills of emotion self-regulation expand considerably in early childhood. Young children can be observed making active efforts to avoid or ignore emotionally arousing events, changing or substituting goals that have been frustrated, redirecting activity or attention, using distraction, engaging in encouraging or reassuring self-talk, obtaining further information about the situation, or enlisting other simple behavioral strategies. In the years that follow, moreover, older children become increasingly proficient at enlisting psychologically more sophisticated strategies of emotion

regulation, such as using emotion-blunting ideation, redefining the situation (or attributions for behavior), or acting in a manner inconsistent with felt emotion. The growth of strategies for emotion self-regulation occurs in concert with the development of emotion understanding and children's increasing psychological self-awareness, and contributes to developing social competence with peers and adults.

Relational Influences on Emotional Development

Emotions are not just the outward expression of an internal experience, but are deeply influenced by the social contexts in which they occur. Emotional development illustrates how early in life, intrinsically human and biologically fundamental response processes take shape in a social context.

Social influences on emotional development are especially apparent in children's close relationships with caregivers and other adults. Many things happen in their relationships with adults that guide the course of children's emotional growth (Dunn, 2003). Early in life, the responsiveness of caregivers to a baby's emotions influences the frequency, persistence, and intensity of the infant's emotional signaling, as well as the baby's developing confidence in the adult's helpful assistance. The adult's emotional expressions influence the child's emotional and behavioral responses to events, in a phenomenon earlier described as social referencing. Parental behavior is often also a catalyst for empathic responding, especially when parents fix the attention of offspring on another's distress and carefully explain the causes of that person's feelings. As earlier noted, the behavioral standards that parents convey in their responses to good behavior and misbehavior shape the earliest experiences of pride, guilt, shame, and embarrassment, as well as developing self-concept and self-esteem. More broadly, the simple conversations shared by parents and their young offspring about the day's events help to color and socialize emotion understanding by how the parent describes and explains emotional experiences, attributes motives and emotions to people, and interprets the emotional experiences they have shared (Thompson et al., 2003). Among the many things children learn from these conversations are the sociocultural rules governing emotional expressions in social situations. Finally, adults in close relationships are influential as they directly manage the emotional arousal of young children, and as they coach (and model) the child's use of personal strategies for emotion self-regulation.

Beyond these specific influences, the broader quality of the parent–child relationship provides a context of security and confidence, or of insecurity and uncertainty, that colors emotional growth and the development of emotion self-regulation (Sroufe et al., 2005). In the security or insecurity of the parent–child attachment relationship that begins in infancy and continues throughout childhood, young children experience reassurance that emotions are constructive, enlivening, and manageable facets of life experience through the support of others, or instead learn that emotions are potentially overwhelming and undermining experiences that are faced alone.

Because of the importance of these relational influences, the 'emotional climate' of the home is a significant influence on early emotional development (Gottman et al., 1997). The emotional climate is significantly shaped by sociocultural values, for example, that guide how emotions are socialized and how children learn to interpret their emotional experiences (*see* Socialization in Infancy and Childhood). In more extreme conditions, unfortunately, the home climate is characterized by the emotional turmoil created by marital dysfunction, coercive or abusive parental conduct, or the adult's affective or substance abuse problems. In these circumstances, caregivers not only fail to be sources of emotional support for the child, but they also contribute to emotional demands that most young children find difficult to manage and which are, for many, potentially overwhelming (Thompson and Goodman, 2010). As a consequence, children are themselves at risk for emotional disturbance resulting from their inability to cope with the emotional demands they face, together with their enmeshment in the emotional turmoil of their caregivers, their heightened sensitivity to cues of emotional arousal around them, and their exposure to a compelling model of emotional dysfunction in the adults they live with. Sadly, studies of such children illustrate further the importance of close relationships to healthy emotional growth.

In more constructive family environments, by contrast, parents contribute to the child's growing ability to understand and enlist emotions constructively in social relationships, and enjoy emotional experience personally. In other words, they contribute to the growth of emotional competence (Saarni, 1999).

Future Directions

This aricle reflects two themes that will guide future research on emotional development. The first concerns children's developing representations of emotional experience, especially as they predict children's responses to emotionally arousing events. How do children understand emotion and its influence upon them, and how does this affect their efforts to manage emotion and respond adaptively to situations that evoke emotion? How do these representations change with age and in concert with changes in other representational abilities? The second theme concerns the developing neurobiological foundations of emotion and emotion regulation in children. In what ways are developing neurological, neurohormonal, and psychobiological processes relevant to the child's evolving capacities related to emotional experience, emotional responding, and the management of emotional arousal. Future research on both questions will advance understanding of emotional development and its applications to clinical and other practical issues.

See also: Early Emotional Development and Cultural Variability; Early Memory Development and Cultural Variations; Emotions, History of; Emotions, Psychological Structure of; Harlow, Harry Frederick (1905–81); Identity in Childhood and Adolescence; Play During Infancy and Early Childhood: Cultural Similarities and Variations; Problem Solving During Infancy and Early Childhood, Development of; Self-Regulation During Early Childhood Across Cultures, Development of; Socialization in Infancy and Childhood.

Bibliography

Cole, P.M., Tan, P.Z., 2007. Emotion socialization from a cultural perspective. In: Grusec, J.E., Hastings, P.D. (Eds.), Handbook of Socialization: Theory and Research. Guilford, New York, pp. 516–542.

Cummings, E.M., Davies, P., 1994. Children and Marital Conflict. Guilford, New York.

Denham, S.A., 1998. Emotional Development in Young Children. Guilford, New York.

Dunn, J., 2003. Emotional development in early childhood: a social relationship perspective. In: Davidson, R.J., Scherer, K.S., Goldsmith, H.H. (Eds.), Handbook of Affective Sciences. Oxford University Press, New York, pp. 332–345.

Gottman, J.M., Katz, L.F., Hooven, C., 1997. Meta-Emotion: How Families Communicate Emotionally. Erlbaum, Mahwah, NJ.

Gunnar, M.R., Davis, E.P., 2003. Stress and emotion in early childhood. In: Lerner, R.M., Easterbrooks, A. (Eds.), Handbook of Psychology, Developmental Psychology, vol. 6. Wiley, New York, pp. 113–134.

Izard, C.E., 2010. The many meanings/aspects of emotion: definitions, functions, activation, and regulation. Emotion Review 2, 363–370.

Izard, C.E., Ackerman, B.P., 2000. Motivational, organizational, and regulatory functions of discrete emotions. In: Lewis, M., Haviland-Jones, J.M. (Eds.), Handbook of Emotions, second ed. Guilford, New York, pp. 253–264.

Lagattuta, K., Thompson, R.A., 2007. The development of self-conscious emotions: cognitive processes and social influences. In: Robins, R.W., Tracy, J. (Eds.), Self-Conscious Emotions, second ed. Guilford, New York, pp. 91–113.

LeDoux, J.E., 2000. Emotion circuits in the brain. Annual Review of Neuroscience 23, 155–184.

Luby, J. (Ed.), 2006. Handbook of Preschool Mental Health. Guilford, New York.

Panksepp, J., Biven, L., 2012. The Archaeology of the Mind: Neuroevolutionary Origins of Human Emotions. Norton, New York.

Saarni, C., 1999. The Development of Emotional Competence. Guilford, New York.

Saarni C., Campos, J.J., Camras, L.A., Witherington, D., 2006. Emotional development: action, communication, and understanding. In: Damon, W., Lerner, R.W. (Eds.), Eisenberg, N. (vol. Ed.), Handbook of Child Psychology, Social, Emotional, and Personality Development, vol. 3, sixth ed. Wiley, New York, pp. 226–299.

Sroufe, L.A., Egeland, B., Carlson, E.A., Collins, W.A., 2005. The Development of the Person. Guilford, New York.

Thompson, R.A., 2006. The development of the person: social understanding, relationships, conscience, self. In: Damon, W., Lerner, R.W, (Eds.), Eisenberg, N. (vol. Ed.), Handbook of Child Psychology, Social, Emotional, and Personality Development, vol. 3, sixth ed. Wiley, New York, pp. 24–98.

Thompson, R.A., 2011. Emotion and emotion regulation: two sides of the developing coin. Emotion Review 3, 53–61.

Thompson, R.A., Goodman, M., 2010. Development of emotion regulation: more than meets the eye. In: Kring, A., Sloan, D. (Eds.), Emotion Regulation and Psychopathology. Guilford, New York, pp. 38–58.

Thompson, R.A., Lagatutta, K., 2006. Feeling and understanding: early emotional development. In: McCartney, K., Phillips, D. (Eds.), The Blackwell Handbook of Early Childhood Development. Blackwell, Oxford, UK, pp. 317–337.

Thompson, R.A., Laible, D.J., Ontai, L.L., 2003. Early understanding of emotion, morality, and the self: developing a working model. In: Kail, R.V. (Ed.), Advances in Child Development and Behavior, vol. 31. Academic, San Diego, pp. 137–171.

Thompson, R.A., Meyer, S., McGinley, M., 2006. Understanding values in relationship: the development of conscience. In: Killen, M., Smetana, J. (Eds.), Handbook of Moral Development. Erlbaum, Mahwah, NJ, pp. 267–297.

Vasey, M.W., Dadds, M.R. (Eds.), 2000. The Developmental Psychopathology of Anxiety. Oxford University Press, Oxford, UK.

Wellman, H.M., 2011. Developing a theory of mind. In: Goswami, U. (Ed.), Wiley-Blackwell Handbook of Childhood Cognitive Development, second ed. Wiley-Blackwell, New York, pp. 258–284.

Zahn-Waxler, C., Radke-Yarrow, M., 1990. The origins of empathic concern. Motivation and Emotion 14, 107–130.

Infancy and Human Development

Marc H Bornstein, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Public Health Service, Bethesda, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Infancy, which encompasses only a small fraction of the average person's expected life span, is a period characterized by astonishing biological, psychological, and social changes. This article first briefly overviews the history of theory and research in infancy and then discusses whether infancy is a 'stage' of life. Challenges infants pose to their study are next reviewed. Although infants pose special challenges to research, human infancy has been studied as a platform for hypothesis and theory testing; as a major period of physiological and psychological adjustment; as a time of socialization; for infants' effects on adults; as a basis of phylogenetic and ontogenetic comparison; and as a setting point for the balance of the life course. This article presents a catalog of multiple types of infancy studies and concludes with a consideration of the general significance of human infancy.

Introduction

The period in the life cycle called *infancy* has been variously defined by educators, lawyers, and social and behavioral scientists. However, the Latinate root of infant, *in + fans*, ties the definition to a psychosocial status, viz the absence language, rather than to educational requirements or vagaries of legal opinion. Infancy roughly brackets the first 12–18 months of postnatal life. Several poignant and telling observations have been made about this period: All of us have been infants, yet mysteriously the mental, emotional, and social life of our infancy is completely lost to us. Infants are irresistible to our senses (*Who is not enchanted by the smell of a baby's head?*), engaging of our intellect (*What are you thinking little one?*), and moving to our emotions (*Who can resist responding to a baby's cry?*). Infants are completely dependent and under our care, and the results of our actions toward them become embodied in them. Infancy represents a deeply resonant beginning in which much is invested theoretically, psychologically, and personally.

Infancy encompasses only a small fraction of the average person's expected life span, but this period is characterized by astonishing biological, psychological, and social changes, many of which are evident even on casual observation. Within this brief phase, infants' length and weight multiply and they move from feeding dependence to being weaned; infants change from immature beings unable to move limbs in a coordinated manner to toddlers who can purposefully control complicated sequences of muscle contractions and flexions necessary to grasp and walk; the infant matures from unintelligible babbler to sentient being who can make needs and wishes abundantly clear verbally. Explaining these many transformations from infancy to maturity constitutes one of the challenging tasks confronting philosophy, biology, and psychology.

A major impetus to the formal and scientific study of infancy came from Charles Darwin who published in *Mind* "A Biographical Sketch of an Infant," an account of the infancy of his firstborn son William Erasmus (Doddy). Darwin's foremost roles in the origins and descent of infancy studies, developmental research, and comparative science sparked continuing interest. In the years following Darwin's (1877) publication, a plethora of baby biographies appeared, and the study of infancy came into its own (Bornstein et al., 2014).

Baby biographies that followed on Darwin's heels in the late nineteenth and early twentieth centuries provided a wealth of ideas about infants, and the evidence developed from them had two important implications. First, baby biographies recorded that infants of different ages were competent at different tasks, and, second, baby biographies revealed the existence of considerable individual variation among infants of the same age. On these observations, scientists eventually developed standardized assessments for infants, and their sequences, scales, and schedules helped further define normal early human development. A landmark example was Nancy Bayley, who in the 1920s began to measure mental and motor growth in infancy systematically. Bayley developed the *California First Year Mental Scale*, now the widely used *Bayley Scales of Infant Development*. Around the middle of the twentieth century, mothers also began to bring their infants to university and governmental laboratories for study. Among the notable progenitors of scientific infancy studies was Robert Fantz, who instigated two techniques – habituation and novelty preference – which essentially broke the communication barrier that separates infants from the rest of us and ignited a revolution in understanding infancy.

Our ever-burgeoning knowledge about infancy since has been compiled periodically, and over the last half-century articles, essays, chapters, monographs, and books on infancy have appeared at a quickening pace. Milestones in the literature on infancy include Kessen et al.'s (1970) *Infancy: A Bibliography and Guide*, a survey that appended more than 2000 references to works on infants appearing at that time. Stone et al.'s (1973) *The Competent Infant* excerpted eclectic important works to that date. Among the notable contemporary treatments focused on academic (not trade) approaches to infancy are Fogel's (2010) *Infant Development*, Bremner and Wachs's (2010) *Handbook of Infant Development*, and Bornstein et al.'s (2014) *Development in Infancy: A Contemporary Introduction*. Several scientific journals, *Infancy*, *Infant Mental Health Journal*, and *Infant Behavior*

and Development, are now devoted exclusively to reports about infancy.

Is Infancy a Stage of Life?

Theory in the biological and social sciences is undecided about whether human development over the life course is gradual and continuous or saltatory and best described by discontinuous stages. Do human beings develop in the way an acorn becomes an oak or as a caterpillar becomes a butterfly? For most developmentalists, the answer to this question turns on predilection and emphasis. In Medieval times, infants were conceived of as miniaturized adults, and the American philosopher psychologist William James (1890: p. 237) opined that human development is continuous, "without break, crack, or division." In development it seems often to be the case that the new absorbs and builds on the old, and so ontogeny seems more like a scale.

At the same time, some constructs, structures, functions, and processes in infants appear to undergo discrete changes. A stage is thought to represent a unique pattern of interrelated features and to differ qualitatively from other patterns that identify other stages. Stages constitute a kind of organizing principle and are used conventionally to compare and contrast assorted aspects of development at one period or between periods. Dividing the life cycle into stages (variously called ages, phases, and periods) is challenging and subtle, and both theoretical and historical currents shape professional and popular opinion on the topic. Stages help make sense of life; they indicate that there are developments and experiences that merit attention in their own right and in terms of their contributions to subsequent development. Many outstanding developmental theorists – Freud, Piaget, Erikson, and Werner – have championed stage theories of development, and all identified infancy as one. Thus, in the minds of many, human infancy possesses an independent status and quality, even if its starting and end points are a bit uncertain. For them, the core of infancy defines a life stage based on biological, cognitive, and social data. Infants do not speak, children do; infants are fed, children are weaned; infants creep and crawl, children walk and run. Social convention also has had important things to say about the distinctiveness of infancy. "… a primary function of culture in shaping human experience is the division of the continuum of human development into meaningful segments, or 'stages.' … All cultures, so far as we know, recognize *infancy* as a stage of human development" (Harkness and Super 1983: p. 223). Indeed, infancy itself has been subdivided into stages. Perhaps, the best known are the six stages described by Piaget, but many developmental scientists since have hypothesized that 'biobehavioral shifts,' 'psychic organizations,' 'developmental crises,' and the like take place during infancy.

All that said, some theorists question the utility of the stage concept, noting that stage theories are descriptive rather than explanatory; stages downplay individual differences and may not contribute to understanding individual development; and stages are often general and idealistic, whereas life is replete with variation and inconsistency. As stated, idiosyncratic prediction and emphasis prevail in this domain.

Challenges to Studying Infants

How much we have learned about infants' behavior and development since Darwin is testimony largely to the creativity, patience, and persistence of researchers in meeting and overcoming the many formidable challenges posed by infants. First and foremost, infants are (by definition) nonverbal (but not, of course, noncommunicative), and they are also, especially in the early months of life, subject to rapid fluctuations in behavioral state, motorically incompetent, and emotionally labile. Unlike adults, infants are not motivated to perform for researchers. Infants' short attention spans, limited response repertoires, and inherent variability are just a few of the other problems that vex investigators.

Yet, infants are also important and attractive to study, for a host of reasons. The impediments infants present have stimulated researchers to develop strategies geared to overcome the communication barrier (as well as other challenges) that naturally and infelicitously separates infants from the rest of us. Through decades of intensive investigation, infants have slowly divulged many of their secrets. In the view of some, this glimpse into the heretofore private world of infancy, and what it portends for the future, represents one of the impressive achievements of modern behavioral and social science.

Domains of Infancy Studies

Infants contribute to biological, psychological, and social knowledge in several major ways.

A Platform for Theory Testing

Because infancy is a period of life that precedes the onset and influence of (much) experience, traditional 'nature–nurture' debates about mental and moral development have persistently turned to infancy for thought experiments and field studies, and different grand theories made specific (and sometimes) testable predictions about infants. What do we know before we have any experience in the world? What kinds of knowledge require what kinds of experience? Infancy has long been a platform for theorizing and hypothesizing.

Philosophers from Plato to Jean-Jacques Rousseau and W.V. Quine directly appealed to infancy for answers to questions that seek the origins and nature of human knowledge. Extreme views on epistemology were proposed by empiricists, who asserted that all knowledge derives from the senses and grows by way of experience; and by nativists, who reasoned that some kinds of knowledge cannot possibly rely on experience and thus human beings must enter the world with a mental apparatus equipped (at the very least) to order and organize their percepts and cognitions.

Empiricists contend that there is no endowed knowledge at birth, that knowledge is learned, and that development proceeds through experience and association. The empiricist's view of the nature of the mind early in life can be characterized by two separate, though conceptually related, schools of thought. One derived from John Locke (1632–1704), who in *An Essay Concerning Human Understanding* (Book II, Chapter 1.2) described the mind at birth as a 'white paper': Mental life

begins 'without any ideas,' and understanding the world depends wholly on the accumulation of experiences. A slightly different empiricist view can be attributed to William James (1842–1910), who in *Some Problems of Philosophy* (1924: p. 50) described sensible life as a big 'blooming buzzing confusion' out of which experience organizes and creates knowledge and order. According to empiricist beliefs, the naive infant does not share the same world of the experienced adult. Empiricism is inherently developmental because, by whatever mechanism postulated, human beings grow from immature to mature.

By contrast, extreme nativists postulated that human beings are not 'created' mindless and that the knowledge that humans possess cannot be achieved by learning alone in so short a span of time as infancy or even childhood. As a consequence, philosophers like René Descartes (1596–1650) and Immanuel Kant (1724–1804) conceived of humans as endowed from birth with 'ideas' or 'categories' of knowledge. The nativist program holds that the human mind naturally and from the beginning of life imposes order on experience, and infant and adult share many capacities. For those abilities that are congenital, nativism is not a developmental view; for those abilities that mature, however, nativism is developmental in outlook.

When psychology turned from nineteenth-century philosophical speculation to twentieth-century empirical investigation, Myrtle McGraw's studies of Johnny and Jimmy; Arnold Gesell's *Atlas* and *Embryology of Behavior*; Jean Piaget's attentiveness to his own three infants: Jacqueline, Lucienne, and Laurent; John Watson's immortal pronouncements on the determinism of association; B.F. Skinner's demonstrations of his daughter in a box; and John Bowlby and Mary Ainsworth's identification of pervasive and enduring effects of attachment drew serial notice to infants as proofs for one or another theory. This tradition continues in comparisons of monozygotes and dizygotes reared apart or treated differently from infancy, in tracing the long-term consequences for children removed from the dire straits of orphanages before or after the end of infancy, and so forth. Infancy continues to provoke enduring philosophical, biological, social, and juridical debate and question. When is the developing fetus a child? When do legal rights of an individual begin?

A Period of Multiple Adjustments

A principal task of infancy is radical physiological and psychological adjustment from the close comfort and complete fulfillment of the intrauterine environment to the deprived, kaleidoscopic extrauterine universe. Save death, the transition from fetus to infant is one of the most abrupt and thoroughgoing adaptations in the life cycle, invoking changes in a constellation of pulmonary, thermoregulatory, and vegetative capacities – not to mention the need to cope with and make sense of the torrent of new sights, sounds, and smells. In many ways, infants are well prepared for this transition to new people and things and for their novel requirements; in a host of other ways, babies are surprisingly ill-prepared. On the one hand, just the way infants look, not to mention their spontaneous and reflexive cries and smiles, draw caregivers near and adaptively maintain caregiver proximity, and infants are notoriously quick learners. Most infants are competent – the transition has been completed successfully by every reader of this sentence. On the other hand, human infants are altricial, and not precocial, animals. As the analyst D.W. Winnicott (1965: p. 39) wrote, "There is no such as an infant." Infants do not and cannot exist outside the infant–caregiver symbiotic system. Of course, infancy is still only a single step on the lifelong path of continuing adjustments.

One concern of infancy study is therefore delineating the status of different characteristics (constructs, structures, functions, and processes) at the beginning of extrauterine life. Focusing on characteristics so close to the beginning of life, investigators of infancy are also naturally interested in how biological and experiential forces fuse to shape our origins as well as inflect the course of our future development. For example, the sensory systems are brain matter, and they do not lie dormant until suddenly 'switched on' at birth; rather, they begin to function before birth. Determining how and approximately when the senses begin to function is important for several reasons. One is that knowing about development of the sensory systems enlarges our understanding of the general relation between structure and function. A second reason is that it is practically (as well as theoretically) important to learn whether and how early brain development is influenced by experience. Preterm babies, born before their expected due date, are exposed to environmental stimulation from which they would normally be protected. If the sensory systems are not functional prenatally, preterms would be protected; if the sensory systems function, however, then preterm babies might be adversely affected by the hyperstimulating environment of the Neonatal Intensive Care Unit.

A Time of Socialization

Adults wield extraordinary and nearly limitless control over infants in part because infants are helpless and in part because infants are plastic to experience. Aside from their genetic bequest, biological parents (just as do social parents) exercise continuing and powerful sway over infants in their roles as ministrators, socializers, and educators. Adults in the infant's immediate environment govern the child's early biological, perceptual, cognitive, and social development by organizing the child's environment, by stimulating, educating, and commanding contingencies. Proximally, many parental decisions of everyday life, from breastfeeding to bowel training, are wholly out of the baby's control but are likely determinative of infant (and possibly future) life nonetheless. Distally, parental sociodemographic characteristics, such as their educational accomplishment, social class, ethnicity, culture, and nationality, determine the lifelong possibilities and prospects for infants, not to mention a child's very chances of surviving infancy. It has been observed that the most significant thing a parent does for an infant is to determine where the child is born (Weisner, 2002). These considerations all betray a perspective of infants as lacking agency, and infants do, but only in proactive degree.

An Exhibition of Agency

Human infants are actually powerful stimuli to adults. Breaking out of a narrow diode top–down view of socialization

focused singularly on adult effects on infants, Richard Bell (1968) inevitably identified many reciprocal 'offspring effects' on parents. Infants feature prominently in this revisionist view. Alone and as a species, infants exert primary direct effects and exhibit secondary indirect effects. The baby's very stature and physiognomy (as mentioned) are compelling. The ethologist Konrad Lorenz (1943) observed that *Kindchenschema* ('babyishness'), the infant's exaggerated head in proportion to the remainder of a short body, bulbous forehead, large eyes, and rounded cheeks, constitute sign stimuli that automatically elicit solicitude in adults. Experimental studies in neuroscience and behavior since support this assertion: The brains of adults (even nonparents who are strangers to the infants depicted) respond quickly and unconsciously to pictures of human infants as opposed to human adults and infant and adult nonhumans as positive and special (Caria et al., 2012; Esposito et al., 2014; Senese et al., 2013). Equally potent and direct are infants' smiles and cries, the one promotes proximity and triggers positive emotions, and the quality, tone, and resonance of the other brooks no excuse or delay in responsiveness. Infant cry has specific and direct effects on adult nervous systems causing blood pressure and skin conductance changes (LaGasse et al., 2005), and eliciting their own unique responses deep within the adult brain (Ranote et al., 2004; De Pisapia et al., 2013; Seifritz et al., 2003). On these bases, infants are unabashedly pressed into service around the world to elicit sympathy, as dependable aids to begging, for example. When Lewis and Clark undertook their renowned nineteenth-century 'Corps of Discovery' of the American west, they included Sacajawea – a Shoshone teenager with a 2-month-old baby strapped on her back. Why? William Clark recorded in his journal that native tribes they met along the way would be inclined to believe that their party was friendly when they saw Sacajawea because a war party would never travel with a woman with a baby (Ambrose, 1997). On a grander scales, almost no society (unless under extreme economic duress) fails to provide infants with the care and attention requisite to their survival. Not infrequently, infants have prodded historical or sociopolitical events; the birth of a child ensures "the line" (Bauer, 2010). Infants likewise powerfully affect women and men in their artistic and scientific work: Infants stimulate song, poetry, literature, humor, art as well as laboratory work. On an individual scale, a newborn constitutes evidence of womanhood and manhood, rousing pride and psychological security. Babies also reorganize the family, provoking mutual commitment and radically altering family economy.

Finally, infants affect their own development. Insofar as people are agents in their own lives (Elder et al., 2015), individual differences in intellect and temperament among infants change those who interact with them.

For purely heuristic purposes, this article separates the otherwise inextricable mutuality between infants and adults into putative 'effects on' and 'effects of.' However, transactions rule life. Modern constructivist theories posit the intricate and permanent interplay of action and experience in the world. How much the infant suckles influences a mother's lactation as much as her lactation influences how much the infant suckles; the level at which parents pitch the complexity of their speech to infants depends on the infant's display of understanding, and infants' understanding depends on the complexity of speech addressed to them.

A Basis of Comparison

Infancy is a natural basis of comparison, and comparisons of many kinds have occupied the attention and interests of infancy investigators. Phylogenetic comparisons contrast different species, and two kinds of infant phylogenetic comparisons are common. One straightforward ethological motive compares human infant abilities, behaviors, and development to those of infants of other species. For example, the rate of brain development in monkey infants outstrips that of human infants. Monkey infants develop faster than human infants, in about a 4:1 ratio (monkeys do in 1 month what a human does in 4 months; Gunderson et al., 1993).

Because many sorts of experimentation with human infants fall outside of ethical bounds, the other phylogenetic comparison uses infant animal models to manipulate contingencies on development, as for example to explore the roles of different kinds of experience in development. Harlow's (1958) renowned studies with infant monkeys illustrate this application. Harlow raised infant monkeys in isolation with one of two mother 'surrogates': One of these 'mothers' was made from wire mesh and had an attached milk bottle from which the monkey could nurse, and the other 'mother' had a soft toweling surface but did not deliver any food. When they were frightened, infant monkeys preferred the cloth mother and held it tightly – in spite of the fact that the wire mother was their only source of nourishment. When these infant monkeys were later put into an unfamiliar environment, they would use the cloth mother as a 'base for operations.' Likewise, experiments based on other sorts of deprivation and cross-fostering can be undertaken with infant monkeys, but not humans.

Other types of comparisons focus on human infants per se and include babies born with or into different circumstances (preterm vs term, adopted vs nonadopted, breastfed vs formula fed). Finally, direct infant–adult ontogenetic comparisons are common (Bornstein et al., 2011); ontogenetic comparisons contrast different stages of life. These include developmental comparisons of mature versus immature status or function. Studies of infancy also provide baseline data against which the normal course of maturation as well as effects of experience into adulthood can be assessed. The human infant is recognized today as 'perceptually competent'; determining just how perception functions in infancy helps to specify the effective perceptual world of babies. If a substance tastes sweet to adults, we may readily suppose that it tastes that way to infants, and even that infants will like it; in fact, however, taste receptors for sweetness may not even be present in infants or, if present, may not function or signal the same perception or quality to infants as to adults.

A Setting Point for the Life Course

Last, infancy has been studied as a basis of stability and change of future development. Development is most readily associated with change and growth (Wohlwill, 1973). As the child rapidly emerges out of the infant, growth and change in many

developmental characteristics are prominent and observable. The complements to growth and change in development are stability and continuity (Bornstein and Bornstein, 2008). Somewhat paradoxically, consistency is more parsimonious and orderly than is transformation. Many features of human development remain (more or less) consistent over time. Biological forces generally tend to reinforce homeostasis in the individual, and the consistent social network in which infant development normally transpires also contributes to stability. Thus, individual–environmental relational processes tilt to promote stability.

Transformation and constancy across time occur at both group and individual levels (Hartmann et al., 2011), and group mean-level consistency or inconsistency (continuity/discontinuity) are complemented by individual-order consistency or inconsistency (stability/instability). The measured mean value of a given characteristic plotted across age defines the developmental function of that characteristic. It may change or remain the same. By contrast, stability describes consistency in the relative ranks of individuals in a group with respect to the expression of some characteristic over time (Bornstein and Bornstein, 2008). For example, a stable characteristic would be one that some infants demonstrate at a relatively high level at time 1 and again at time 2. There are different types of stability. Homotypic stability describes maintenance of rank-order status on an identical characteristic from time 1 to time 2 (e.g., vocabulary size at 12 months and at 4 years). Heterotypic stability describes maintenance of rank-order status between two related, but not identical, characteristics, often related because they presumably share the same underlying process (e.g., vocabulary at 12 months and storytelling at 4 years).

Stability and change tell us about the nature and overall developmental course of a characteristic. Whether individuals maintain their order on some characteristic through time informs not only about individual variation, but contributes to a deeper understanding of the origins, nature, and future of the characteristic as well. Insofar as a characteristic is ontogenetically stable, we know that infants who do well or poorly in that characteristic at one time are likely to do well and poorly again later. Moreover, stable early characteristics can be expected to shape later emerging ones. Infants who know more words at 1 year tend to know more words at 2 years, and 2-year-olds who know more words may be at a long-term advantage because knowing more words speeds learning to read, improves verbal comprehension, and eventuates in better oral language skills (Marchman and Fernald, 2008).

Thus, stability and change have meaning for the more general significance of infancy. The life span perspective in developmental science specifies that human beings are open systems, and the plastic nature of psychological functioning ensures both stability and change across the life course. Significantly, immaturity after birth is marked in humans, and the elongation of infancy permits ample time to develop skills and sample experiences from the environment. In turn, both promote greater behavioral flexibility and adaptive function. Many developmental processes are therefore Janus-like with stable and unstable, continuous and discontinuous aspects. Both stability and change confer advantage. Consider how life might best engineer a human infant. Certainly, we would not want to see the infant's intelligence, personality, what have you fixed in advance of any actual experience. It would be shortsighted and counterproductive to rule out possible effects of experience because adequate development mandates the ability to adapt successfully to the effective surrounding environment. By the other side of the same token, however, we would not want to leave development wholly to the whims of experience because individual experiences can vary so much. A fundamental conceptual issue that has framed debates in theory and research across the history of developmental science has concerned questions of stability and change. In fact, both of these design criteria appear to be met in the evolution of development from human infancy.

The Significance of Infancy

Whether infancy is 'significant' has engendered intense debate between two polar points of view. Some contend that infancy is not particularly important because the status of the infant or experiences in infancy have little (if any) long-term predictive meaning. That is, infant characteristics and experiences are supplanted later in development. This position emphasizes discontinuities and instabilities between infancy and maturity. In this view, infancy is disconnected from the balance of the life course, and infant characteristics and experiences are peripheral or ephemeral, in the sense that they exert little or no enduring effect on later development (Clarke and Clarke, 2000; Kagan, 1998; Lewis, 1997). Empirical support for models that invest more in contemporary status and experience typically consists of failures of early experiences and interventions to show sustained effects as well as demonstrations of the recovery of function from early enrichment or deprivation.

However, other theories contend that infancy is part of a seamless or continuous lifeline and that characteristics and experiences in infancy are important in themselves and are also crucial to later development. Since the time of Plato (*c*.350 BC), philosophers and scientists alike have considered infancy to be uniquely important, and folk wisdom (*As the twig is bent, so grows the tree*) and poetry (*The Child is Father to the Man*) alike express the conviction that early life experiences shape later development. Arguments for the specialness of infancy in this way derive from a diverse and impressive array of theoretical starting points, including psychoanalysis, behaviorism, constructivism, ethology, neuropsychology, attachment, and systems theory. For these theorists, social orientations, intrapersonal proclivities, and intellectual predilections established in infancy set influential, if not enduring, patterns (Bornstein, 2014). For example, Sigmund Freud (1949) focused attention on infancy, suggesting that the ways people are treated in infancy establish lifelong personality traits, and Erik Erikson (1963) also contended that experiences in infancy have implications for the way a person negotiates successive stages of development. For learning theorists like John Watson, B.F. Skinner, and John Dollard and Neal Miller, experiences in infancy are important because they occur first, have no competing propensities to replace, and thus promote

easy and rapid learning. Moreover, they asserted that early and simpler behavior patterns laid down in infancy underlie later and more complex behavior patterns acquired in maturity. What is experienced or learned first often has lasting influences, and infants seem to command an extraordinary facility for learning. Jean Piaget opined that more advanced developmental capacities build on more elementary ones constructed in very early life. Likewise, ethologists and embryologists from Konrad Lorenz and Niko Tinbergen to Stephen J. Gould and Gilbert Gottlieb emphasized the lifelong reverberations of early experiences (as in sensitive periods; Bornstein, 1989) that tend to occur in infancy. The immature nervous system appears to be especially plastic; it can be shaped by experience and its infant character affects the subsequent trajectory of development. Neoteny (the prolongation of infancy, especially in human beings) may have important adaptive significance. John Bowlby, Mary Salter Ainsworth, and their successors (Sroufe et al., 2005) hypothesized that attachment experiences in infancy augur future personality, social, and cognitive development. Even modern systems theorists contend that development consists of hierarchically organized characteristics that incorporate earlier emerging ones that increasingly differentiate over time (Lewontin, 2005). All of these perspectives imply that infant state and experience may disproportionately affect the course of subsequent development.

The development that takes place during the short 635 days after conception – roughly, the prenatal period and the first year – is as or more dramatic than any others in the human life span. The most remarkable developments involve the changing shape and capacity of the body; the complexification of the nervous system; the dawning of sensory and perceptual capacities; the increasing abilities to make sense of, understand, and master things in the world; the achievement of communication; the emergence of characteristic personal and social styles; and the formation of specific social bonds. At no other time is development so fast paced or thoroughgoing in so many different spheres of life. Yet, like Theseus's ship, at the core of the infant and later the toddler, child, adolescent, and adult, is the same individual, and some stability from infancy is expectable.

Summary

On first observation, newborns and infants appear to be erratic and disorganized. At every moment, babies seem to be constantly moving all body parts without apparent purpose; and observed over longer periods, they seem to move unpredictably between sleep and alertness. However, infants are not quite so random. Close and consistent inspection shows that infant activity is actually organized and rhythmical. Some infant actions regularly cycle at high frequencies, perhaps once or more every second. Heartbeats, breathing, and sucking illustrate fast biological rhythms that maintain life, and kicking and rocking exemplify other fast-cycling behaviors. General movements of the body cycle at intermediate rates, once every minute or so. States of waking, quiet sleep, and active sleep cycle at low frequencies, in periods of up to 24 h. By observing over extended times and carefully decomposing infant activity states, it is possible to detect regularity underlying infants' seeming randomness. What we see in the infant at any one time is only apparent irregularity; much of infant behavior is characterized better as underlying regularity, the simultaneous and independent cycling of many complex rhythms.

Understanding infancy better has several implications. First, it has meaning for developmental science in terms of more fully describing the growth of individuals. Second, it is important for more completely understanding the nature and structure of biological and psychological characteristics. Infant development is governed by genetic and biological factors in combination with experiences and environmental influences. Thus, the status and development of any characteristic are attributable to factors in the individual that transact with factors in the environment. Third, measures of infants serve as screens for early detection of risk status or for delayed development. Finally, infancy predicts important things in life like education, occupation, social mobility, and health, so knowing more about infancy is valuable.

Deconstructing infancy in the way contributes to elaborating and to explaining infancy qua a dimension of scientific inquiry, human fascination, and cultural interest. The ineffable romance of infancy may haunt us in many ways, but in other ways we now see that infancy contributes concretely to our understanding ourselves. The fact that infants are unvolitional, uncooperative, unstable, nonverbal, and motorically inept once warded off all but philosophical speculation about them. Scientists have now overcome infants' formidable and intractable posture to extract information of all sorts from and about them. We now know a great deal about babies' perceptions, thoughts, and feelings.

In modern times, a revolution has taken place in infancy studies, fueled by technological, methodological, and theoretical advances. This revolution has shown ever more faithfully how endogenous and exogenous forces coordinate to bring infants in line with their physical and social reality. Infancy is the first phase of life lived outside the womb. Characteristics developed and acquired in infancy may be fundamental to the balance of the life course; some may endure, at least in part, and they as well as others that do not, are still those which later maturation and experiences build on or modify. Infancy is only one phase in the life span, however, and so our nervous system maturity, physical development, motor capacities, perceptual abilities, cognitive competencies, idiosyncratic personalities, and social styles are also shaped by experiences and development after infancy. The start does not fix the course or outcome of development, but it clearly exerts an impact on both.

Acknowledgments

This article summarizes selected aspects of my research, and portions of the text have appeared in previous scientific publications cited in the references. Supported by the Intramural Research Program of the NIH, NICHD. I thank A. Dovidio. Correspondence to: Marc H. Bornstein, Child and Family Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Rockledge 1, Suite 8030, 6705 Rockledge Drive, MSC 7971, Bethesda, MD 20892–7971, USA. E-mail: Marc_H_Bornstein@nih.gov.

See also: Attachment During Infancy and Early Childhood: Cultural Variations, Development of; Cognitive Development During Infancy and Early Childhood across Cultures; Early Emotional Development and Cultural Variability; Intellectual Functioning, Assessment of; Motor Skills: Development in Infancy and Early Childhood; Self in Culture: Early Development; Social and Emotional Development in the Context of the Family; Temperament Development, Theories of; Temperament and Human Development.

Bibliography

Ambrose, S., 1997. Undaunted Courage. Simon and Shuster, New York.

Bauer, S.W., 2010. The History of the Medieval World: From the Conversion of Constantine to the First Crusade. W. W. Norton & Company, New York.

Bell, R.Q., 1968. A reinterpretation of the direction of effects in studies of socialization. Psychological Review 75, 81–95.

Bornstein, M.H., 1989. Sensitive periods in development: structural characteristics and causal interpretations. Psychological Bulletin 105, 179–197.

Bornstein, M.H., 2014. Human infancy … and the rest of the lifespan. Annual Review of Psychology 2014 (65), 121–158.

Bornstein, M.H., Arterberry, M.E., Lamb, M.E., 2014. Development in Infancy: A Contemporary Introduction, fifth ed. Psychology Press, New York.

Bornstein, M.H., Mash, C., Arterberry, M.E., 2011. Perception of object-context relations: eye-movement analyses in infants and adults. Developmental Psychology 47, 364–375.

Bornstein, M.H., Bornstein, L., 2008. Psychological stability. In: Darity Jr., W.A. (Ed.), International Encyclopedia of Social Sciences, vol. 8. Macmillan Reference USA, Detroit, MI, pp. 74–75.

Bremner, J.G., Wachs, T.D. (Eds.), 2010. The Wiley-Blackwell Handbook of Infant Development, second ed. Wiley-Blackwell, Malden, MA.

Brockelman, W.Y., Schilling, D., 1984. Inheritance of stereotyped gibbon calls. Nature 312, 634–636.

Caria, A., de Falco, S., Venuti, P., Lee, S., Esposito, G., Rigo, P., Birbaumer, N., Bornstein, M.H., 2012. Species-specific response to human infant faces in the premotor cortex. NeuroImage 60, 884–893.

Clarke, A.M., Clarke, A.D.B., 2000. Early Experience and the Life Path. Jessica Kingsley Publishers, London, England.

Darwin, C.R., 1877. A biographical sketch of an infant. Mind 2, 286–294.

De Pisapia, N., Bornstein, M.H., Rigo, P., Esposito, G., De Falco, S., Venuti, P., 2013. Sex differences in directional brain responses to infant hunger cries. NeuroReport 24, 142–146.

Elder, G.H., Shanahan, M.J., Jennings, J.M., 2015. Human development in time and place. In: Lerner, R.M., Bornstein, M.H., Leventhal, T. (Eds.), Ecological Settings and Processes in Developmental Systems, Handbook of Child Psychology and Developmental Science, seventh ed., vol. 4. Wiley, Hoboken, NJ.

Erikson, E., 1963. Childhood and Society. Norton, New York.

Esposito, G., Nakazawa, J., Ogawa, S., Stival, R., Kawashima, A., Putnick, D.L., Bornstein, M.H., 2014. Baby, you light-up my face: Culture-general physiological responses to infants and culture-specific cognitive judgments of adults. PlosOne.

Fogel, A., 2010. Infant Development. Sloan Publ., Cornwall-on-Hudson, New York.

Freud, S., 1949. An Outline of Psycho-analysis. Norton, New York.

Gunderson, Virginia M., Yonas, Albert, Sargent, Patricia L., Grant-Webster, Kimberly S., March 1993. Infant macaque monkeys respond to pictorial depth. Psychological Science 4 (2), 93–98. http://dx.doi.org/10.1111/j.1467-9280.1993.tb00467.x.

Harkness, S., Super, C.M., 1983. The cultural construction of child development: a framework for the socialization of affect. Ethos 11, 221–231.

Harlow, H.F., 1958. The nature of love. American Psychologist 13 (12), 673–685.

Hartmann, D.P., Pelzel, K.E., Abbott, C.B., 2011. Design, measurement, and analysis in developmental research. In: Bornstein, M.H., Lamb, M.E. (Eds.), Developmental Science: An Advanced Textbook, sixth ed. Psychology Press, New York, pp. 109–197.

James, W., 1890. The Principles of Psychology. Henry Holt, New York.

Kagan, J., 1998. Three Seductive Ideas. Harvard University Press, Cambridge, MA, US.

Kessen, W., Haith, M.M., Salapatek, P.H., 1970. Human infancy: a bibliography and guide. In: Mussen, P. (Ed.), Carmichael's Manual of Child Psychology. Wiley, New York, pp. 287–445.

LaGasse, L.L., Neal, A.R., Lester, B.M., 2005. Assessment of infant cry: acoustic cry analysis and parental perception. Mental Retardation and Developmental Disabilities Research Reviews 11, 83–93.

Lewis, M., 1997. Altering Fate: Why the Past Does Not Predict the Future. Guilford Press, New York.

Lewontin, R., 2005. The Triple Helix. Harvard Univ. Press, Cambridge, MA.

Lorenz, K., 1943. Die angeborenen Formen möglicher Erfahrung. (Innate form of potential experience). Zeitschrift für Tierpsychologie 5, 235–309.

Marchman, V.A., Fernald, A., 2008. Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. Developmental Science 11, F9–F16.

Ranote, S., Elliott, R., Abel, K.M., Mitchell, R., Deakin, J.F., Appleby, L., 2004. The neural basis of maternal responsiveness to infants: an fMRI study. Neuroreport 15, 1825–1829.

Seifritz, E., Esposito, F., Neuhoff, J.G., Luthi, A., Mustovic, H., Dammann, G., von Bardeleben, U., Radue, E.W., Cirillo, S., Tedeschi, G., Di Salle, F., 2003. Differential sex-independent amygdala response to infant crying and laughing in parents versus nonparents. Biological Psychiatry 54, 1367–1375.

Senese, V.P., De Falco, S., Bornstein, M.H., Caria, A., Buffolino, S., Venuti, P., 2013. Human infant faces provoke implicit positive affective responses in parents and non-parents alike. PlosOne 8, e80379, 1–7.

Sroufe, L.A., Egeland, B., Carlson, E.A., Collins, W.A., 2005. The Development of the Person: The Minnesota Study of Risk and Adaptation from Birth to Adulthood. Guilford, New York.

Stone, L.J., Smith, H.T., Murphy, L.B., 1973. The Competent Infant: Research and Commentary. Basic Books, New York.

Weisner, T.S., 2002. Ecocultural understanding of children's developmental pathways. Human Development 45, 275–281.

Winnicott, D.W., 1965. The Maturational Processes and the Facilitation Environment: Studies in the Theory of Emotional Development. Hogarth Press, London.

Wohlwill, J.F., 1973. The Study of Behavioral Development. Academic, New York.

Infant and Child Mortality in Industrialized Countries

Markéta Pechholdová, University of Economics, Prague, Prague, Czech Republic

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Monnier, volume 11, pp. 7398–7404, © 2001, Elsevier Ltd.

Abstract

Infant mortality is a sensitive indicator of social progress and disparities. Until the beginning of the twentieth century, the chances of surviving the period of infancy would rarely exceed 80%. General improvement in nutrition and childcare, followed by key medical discoveries and public health action, resulted in an unprecedented secular decline in infant mortality, which took place in all industrialized countries and which continues until present. Before 1950, reductions resulted mainly from management of malnutrition and infection in older infants. Mortality of newborns started to decline only after intensive development of obstetrics and neonatology in the 1960s. Nowadays, infant mortality primarily depends on the neonatal survival of high-risk babies born preterm or with low birth weight.

Infant mortality, measuring survival of newborns to their first birthday, holds a special position in demographic analysis. Shortly after birth, humans experience adaptation difficulties and are extremely vulnerable to environmental conditions. High infant mortality, very low child and juvenile mortality, and regularly increasing mortality as of age 30 are typical traits of the J-shaped curve of human age-specific mortality risk pattern. In some industrialized countries, infant mortality levels have recently fallen to historically minimal levels providing evidence that most infant deaths are preventable and avoidable. The levels of infant mortality therefore also indicate the success or failure of medicine and society. It is generally agreed that more than any other demographic indicator, infant mortality reflects the societal health and well-being progress of the population.

Level of infant mortality is prominently expressed by the infant mortality rate (IMR), defined as the number of deaths under age 1 divided by the number of live births within the given calendar year. As the number of births may strongly vary from 1 year to another, the precision of the indicator can be improved by adjusting the denominator (relating the deaths to the appropriate birth cohort or, in absence of cohort-specific data, by applying correction based on the assumption of stable share of infants from the given cohort dying during the next calendar year).

Historical and Geographic Perspective

Limited information from historical populations suggests that under natural reproduction regimes, one-third of infants would die before the first anniversary, with important surges in times of crises (war, famine, epidemics, heat waves, etc.). As the numbers of births and deaths were similar (natural growth was small), infant deaths also represented one-third of all yearly deaths. Medieval newborns were at high risk not only at the moment of delivery or during epidemics but also due to persistence of harmful feeding and childcare practices (such as swaddling, early intentional exposure to cold, early weaning, administration of raw animal milk or improper milk substitutes, use of soporifics, etc.), while overlay and smothering were also frequent. Among the poorest, infanticide was not negligible. Finally, also a death of a mother would significantly decrease the chances of a newborn to survive. Of the first-year survivors, another one-tenth or more would die before reaching the age of 5 from some of the major infectious threats (pneumonia, diphtheria, tetanus, typhoid fever, whooping cough, dysentery, smallpox, or cholera). In the earliest stages of urbanization, infant mortality in the cities was higher than that in rural areas due to higher concentration of poverty in the cities, worse housing and sanitary conditions, and lower quality of urban milk and food. In some populations, infant mortality started to decline since the beginning of the nineteenth century. General economic progress (McKeown, 1976) and spread of hygienic and childcare know-how (Preston and Haines, 1991) are believed to be the main factors thereof.

Secular Decline

The end of the nineteenth century put public health on a more rational basis. A revolution in preventive medicine occurred after discovering infectious agents of major diseases (cholera, diphtheria, typhoid fever). Etiologic advances were quickly followed by efficient treatments or vaccines, while life-saving value of pasteurizing the milk was ultimately recognized. During the twentieth century, gradual improvement of hygiene, sanitation, and nutrition, together with next medical milestones (sulfonamides in the 1930s, penicillin in the 1940s) and targeted public health campaigns including public infant milk depots (such as the *Goutte de lait* in France), further contributed to unprecedented and spectacular mortality decline. **Figure 1** provides long-term time series of IMR for five European countries. The onset of the decline varied from the earliest in Sweden to the latest in France or the Czech Republic; from the beginning of the twentieth century, however, the infant mortality decreased in all industrialized countries. The decline took place in two stages: before 1950, the trend was steep but irregular, interrupted by temporary fluctuations, while since 1950, the rates have smoothly reduced by one-third every 10 years. In the first decade of the twenty-first century, the IMR dropped below 3 deaths per 1000 live births in Finland, Japan,

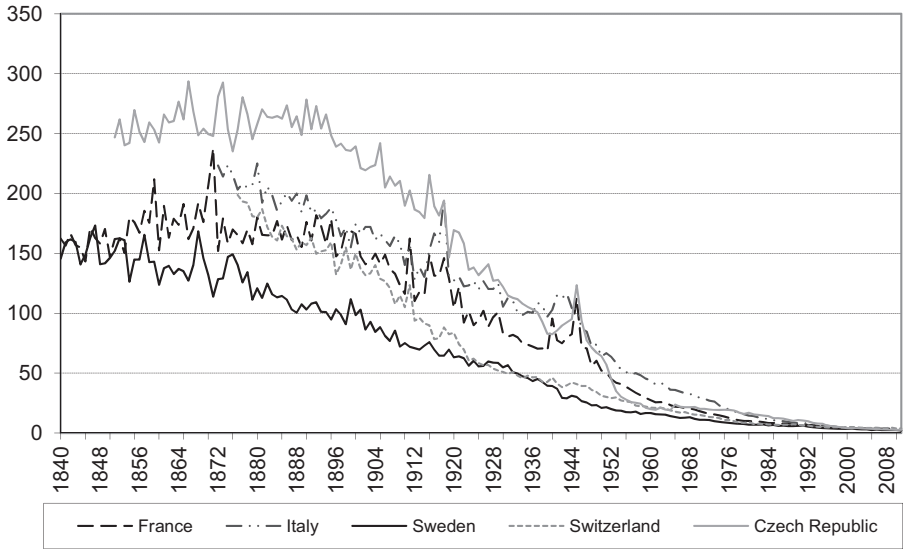

Figure 1 Long-term trends in IMR (per 1000 live births). Source: Human mortality database, Czech Statistical Office.

Sweden, Iceland, Slovenia, and the Czech Republic. A 100-fold historical decline has been achieved.

Regional Disparities

As Figure 1 suggests, important disparities exist in the timing and initial levels of the secular decline. Table 1 provides an overview of IMR levels across most of the today's industrialized countries, covering – where possible – the entire twentieth century. Around 1900, infant mortality was by far the lowest in Northern Europe, with Norway being the leading country (82 deaths per 1000 live births). England and Wales, together with Finland, were the worst in this region, with levels exceeding 150 deaths per 1000 live births. In Western Europe, 150 was the minimum (in Switzerland), followed by 166 in France, 199 in Germany, and up to 216 in Austria. These historical regional disparities can be attributed to different feeding practices (breastfeeding highly prevalent in the north, while rare, e.g., in Austria), climatic conditions, and different urbanization degree (via the negative urban effect on infant mortality). In Central and Eastern Europe, with exception of Poland and Bulgaria, infant mortality exceeded 200 deaths per 1000 live births. Concerning Russia, infant mortality around 1900 could have been as high as 250 per 1000 live births, but varied from 116 to 298 depending on ethnicity and religion. High levels were explained by overall underdevelopment, while religious differentials were attributed mainly to different feeding practices (Zakharov, 1996). Due to incomplete registration of vital events in the past, the earliest rates as provided by Table 1 can be underestimated, especially in less developed countries of Southern and Eastern Europe.

Until 1950, the rate was reduced to 30 or less not only in Scandinavia, the Netherlands, and Switzerland but also outside of Europe (Australia, New Zealand, and the United States), while Austria, Belgium, France, Germany, Southern Europe, and Japan lagged behind. In the East, and especially in the Southeast, the disadvantage was much clearer – except for Slovenia, Bulgaria, Hungary, and the Czech Republic, IMRs still exceeded 100 deaths per 1000 live births.

Between 1950 and 2010, infant mortality decreased in all regions by roughly 90% regardless of initial levels. In the traditionally industrialized countries, the decrease was regularly distributed in time, and in spite of very low current levels, the relative gains have not diminished. Reaching the lowest-low levels has, however, proven particularly difficult in North America, which has led to worsening of the relative position of United States and Canada. Southern Europe was the last one to take part in the infant mortality decline, but has caught up since the 1970s. In Central Europe, the rates halved between 1950 and 1960 as a direct consequence of universal access to free-of-charge health care. Socialist health care systems were quite efficient in reducing infectious mortality, as the antibiotics and vaccines were relatively cheap, accessible, and easily controllable. Since 1960s, these systems could not keep abreast with advances in highly specialized intensive (and expensive) medicine and the progress slowed down. Eastern Europe – former European republics of the USSR and Russia – followed a segmented path with an initial decline followed by stagnation or even deterioration in the 1970s. A part of the deterioration can be attributed to the adoption of the new definition of live birth in 1974, but there has also been a real increase in infant mortality from infectious diseases in the 1970s (Meslé and Vallin, 2012). These unfavorable trends drove the Soviets to stop publishing the mortality statistics between 1974 and 1987. The infant mortality in the ex-USSR also stagnated during the posttransitional socioeconomic crisis in the 1990s and in spite of recent reduction nowadays, represents maximum with an average value of 7 infant deaths per 1000 live births.

The ranking of the countries was remarkably stable. Scandinavian countries were the leading countries in 1900 and are among the first ones in 2010. Maximum values in Albania, Moldova, and Romania have also been observed consistently.

Table 1 Infant deaths per 1000 live births in the twentieth century

	1900	*1930*	*1950*	*1960*	*1970*	*1980*	*1990*	*2000*	*2010*
Northern Europe									
Denmark	128	80	31	22	14	8	7	5	3
Finland	153	75	44	21	13	8	6	4	3
Iceland	127	45	22	13	13	8	6	3	2
Ireland	–	68	46	29	20	11	8	6	3
United Kingdom	159[a]	63	31	22	18	12	8	6	4
Sweden	99	55	21	17	11	7	6	3	3
Norway	82	42	26	18	12	8	7	4	3
Western Europe									
Austria	216	104	66	38	26	14	8	5	4
Belgium	178	100	53	31	21	12	8	5	3
France	166	83	52	27	18	10	7	4	3
Germany (West)	199	85	60	34	24	13	7	4	4
Germany (East)	199	85	72	39	18	12	7	4	3
Netherlands	170	51	25	17	12	9	7	5	4
Switzerland	150	51	31	21	15	9	7	5	4
Southern Europe									
Greece	–	99	35	40	30	18	10	5	3
Italy	174	106	64	44	30	15	8	4	3
Portugal	168	144	94	78	58	24	11	6	4
Spain	201	124	70	44	28	12	8	4	3
Albania	–	–	121	83	98	50	28	18	13
Bosnia and Herzegovina	–	–	126	107	69	32	15	8	7
Croatia	–	–	118	70	34	21	11	7	4
Serbia	–	–	117	85	54	33	23	11	7
Slovenia	–	–	81	35	25	15	8	5	3
Central Europe									
Bulgaria	161	138	95	45	27	20	15	13	9
Czech Republic	239	121	64	20	20	17	11	4	3
Hungary	223	152	86	48	36	23	15	9	5
Poland	163	142	108	56	33	21	19	8	6
Romania	200	176	117	76	49	29	27	19	10
Slovakia	207	–	106	29	26	21	12	9	4
Eastern Europe									
Belarus	–	–	–	35	19	16	12	9	4
Estonia	–	–	–	31	18	17	12	9	3
Latvia	–	–	–	27	18	15	14	10	6
Lithuania	–	–	–	38	19	14	10	9	4
Moldova	–	–	–	49	24	35	19	18	12
Russia	–	–	–	37	23	22	18	15	8
Ukraine	–	–	–	30	17	17	13	12	9
Outside Europe									
Australia	–	47	24	20	18	11	8	5	4
Canada	–	89	42	27	19	10	7	5	5
Japan	–	125	60	31	13	8	5	3	2
New Zealand	–	35	28	23	17	13	8	6	5
USA	–	65	29	26	20	13	9	7	6

[a]England and Wales.
Source: Human mortality database, United Nations, WHO mortality database, national statistical offices.

In contrary, Austria and Germany passed from the worst to the best group. Positive reverse also occurred in some Central European countries (namely, Czech Republic and Slovenia), in Estonia, and in Japan, where the rates are now among the lowest in the world. In 1950, the range between minimum and maximum level of IMR was 105, while today it is only 11. The homogenization is, however, only apparent: when relative variability measures are used (such as coefficient of variation), the overall geographic disparities are as great as in 1950.

Components of Infant Mortality

Understanding the history and the disparities of infant mortality requires viewing it as a differentiated and irregular pathological process. A simple observation of infant deaths by time elapsed since birth shows that infant mortality is the highest shortly after the delivery. The so-called neonatal and postneonatal deaths are customarily distinguished. Neonatal mortality includes deaths occurring during the first month of

life (0–27 days) and is further decomposed into early neonatal (0–6 days) and late neonatal (7–27 days). As the infant mortality declines, deaths are increasingly concentrated to the very start of life. It is therefore recommended to additionally measure the first-day mortality (within 24 h after birth). Newborns dying in the early neonatal period are exposed to similar risk factors as the stillbirths. To aggregate this intra-uterine/intrapartum causation, and to eliminate possible effect of misclassification between live- and stillbirths, perinatal mortality rate can be computed as a sum of stillbirths and early neonatal deaths related to the sum of live- and stillbirths. Stillbirth is the delivery of a presumably viable fetus (delivery of nonviable fetus is abortion). Postneonatal mortality then encompasses deaths occurring between 28 and 364 days after birth.

Based on age at death, a method proposed in the 1950s introduced a dual pseudoetiological classification of infant deaths into endogenous and exogenous (Bourgeois-Pichat, 1951). Endogenous deaths occurring shortly after the birth are linked with congenital viability and as such are hardly preventable, while exogenous deaths occur due to failure in treatment and are – at least theoretically – fully avoidable. According to Bourgeois-Pichat, all postneonatal plus a quarter of neonatal deaths are exogenous, while virtually all first-day and most early neonatal deaths can be attributed to endogenous etiology. These estimates are no more valid for recent data, but the concept of endogenous and exogenous causes is still used.

In the nineteenth century, postneonatal component represented around two-thirds of all deaths in infancy and was a great source of mortality variation, especially between urban and rural areas or between legitimate and illegitimate children (Woods et al., 1988). Reductions in postneonatal mortality, resulting mainly from the control of malnutrition and infection, have also driven the majority of the infant mortality decline in the first half of the twentieth century. As a result, postneonatal deaths formed only one half of all infant deaths in 1950, and in the mid-1960s, their share decreased to 30% (Lantoine and Pressat, 1984). After 1960, neonatal mortality advances gained more importance due to intensive development of neonatology (Masuy-Stroobant, 1997). **Figure 2** shows the neonatal and postneonatals component of infant mortality change since the 1970s in European regions and in the United States. Neonatal mortality declined regularly everywhere as a result of better prenatal screening and prevention (e.g., prenatal ultrasound diagnostics, folic acid administration, *in utero* blood transfusion, managing high-risk pregnancies, therapeutic abortions) and intensive obstetric and neonatal health care (heart rate and acid–base balance monitoring during delivery, infant respirators, intravenous alimentation, phototherapy). The resulting progress in neonatal survival rates was so intense that it enforced changes to viability criterion (from >1000 g to > 500 g). In Eastern, Southern, and Central Europe, the neonatal mortality decline started later, but since 1970, sustained progress in both components has been observed. In North and West of Europe, postneonatal mortality stagnated between the mid-1960s and 1990s. The stagnation has been explained as a consequence of postponing neonatal deaths to the post-neonatal period, and the emergence of the sudden infant death syndrome (SIDS) as the leading postneonatal cause of death in developed world (Barbieri, 1998). Since 1990, the progress was resumed, dividing the postneonatal mortality by three in the following two decades. In the United States, the decline of both components was slower than in the North and the West of Europe and has virtually stopped since 2000. Several explanations have been put forward, among them higher registration of very preterm and very low birth weight newborns as live births (applying pragmatic rather than definition-based criterion of live birth), but most of the

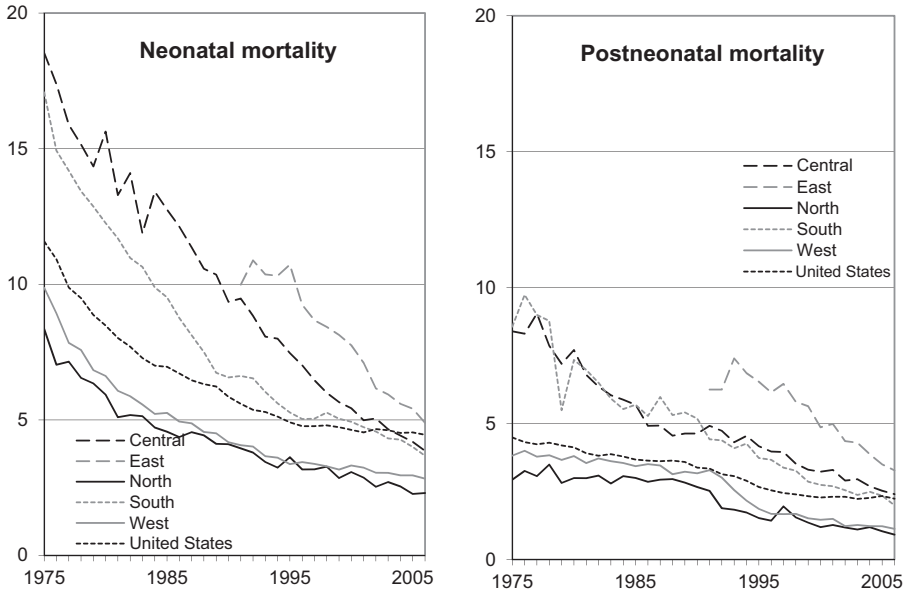

Figure 2 Components of infant mortality change since 1970, per 1000 live births. Regions of Europe and the United States. Note: due to unavailability of data prior to 1990, Albania is not included in the data set. Source: World Health Organization, National Center for Health Statistics.

observed stagnation is due to higher occurrence and lack of survival progress of high-risk newborns (MacDorman and Mathews, 2008).

Table 2 shows the components of infant mortality in 2006. Early neonatal deaths represent the largest infant mortality component, reflecting extreme frailty of newborns on one side and the quality and availability of neonatal health care on the other side. Countries with lowest IMRs usually have the lowest neonatal rates (Sweden, Finland, Spain, Greece, France, Czech Republic, Slovenia, and Estonia). Neonatal mortality is considerably higher in most former socialist countries, and also in Denmark, the Netherlands, Switzerland, the United Kingdom, and the United States. Diverse registration practices (pragmatic, birth weight/gestational age criterion, or with additional condition of a minimum survival time) are at source of some part of the observed disparities in early neonatal mortality, most sensitive to distinction between live- and stillbirths (Gourbin and Masuy-Stroobant, 1995; Mohangoo et al., 2011). The shares of postneonatal deaths are currently under 30% in countries with lowest IMRs (less than 4), and close to 40% in countries with highest IMR levels. High post-neonatal mortality in the United States (2.2 per 1000 live births), unrelated to the differences in perinatal registration, points to the relative deterioration of infant health.

Table 2 Components of infant mortality in 2006

	Infant deaths in 2006 per 1000 live births			
	Early neonatal (0–6 days)	*Late neonatal (7–27 days)*	*Postneonatal (28–364 days)*	*Infant (0–364 days)*
Sweden	1.4	0.4	1.0	2.8
Finland	1.6	0.5	0.8	2.9
Czech Republic	1.4	0.9	1.0	3.3
Slovenia	1.9	0.6	0.9	3.4
Norway	2.0	0.7	0.7	3.4
Spain	1.5	0.8	1.2	3.5
France	1.6	0.8	1.3	3.6
Austria	2.0	0.6	1.1	3.6
Italy	1.8	0.8	1.0	3.6
Greece	1.6	0.9	1.2	3.7
Denmark	2.5	0.6	0.7	3.8
Germany	2.0	0.6	1.3	3.8
Ireland	2.2	0.6	1.1	3.9
Belgium	1.9	0.7	1.5	4.1
Netherlands	2.5	0.8	1.1	4.4
Switzerland	3.0	0.4	1.0	4.4
Estonia	1.7	1.0	1.8	4.4
United Kingdom	2.6	0.9	1.5	5.0
Croatia	3.2	0.7	1.3	5.2
Hungary	2.8	0.9	2.0	5.7
Poland	3.3	1.1	1.6	6.0
Slovakia	2.4	1.1	3.1	6.6
USA	3.6	0.9	2.2	6.7
Lithuania	2.9	1.0	2.9	6.8
Serbia	4.0	1.2	2.2	7.4
Latvia	3.6	1.1	3.0	7.6
Bulgaria	3.8	1.6	4.4	9.7
Russian Federation[a]	4.8	1.6	4.7	11.0
Republic of Moldova	5.4	1.6	4.8	11.8
Romania	5.5	2.3	6.2	14.0

[a]Data available for 2005.
Source: World Health Organization, National Center for Health Statistics.

Medical Causes of Death

The World Health Organization (WHO) international classification of diseases devotes two separate chapters to conditions of infancy, entitled 'Certain Conditions Originating in the Perinatal Period' (XVI) and 'Congenital Malformations, Deformations and Chromosomal Abnormalities' (XVII). One of the earliest available cause-of-death statistics show that 40–50 infant deaths per 1000 live births in France in 1930 out of 80–100 in total was caused by perinatal and congenital disorders, while an equal share of deaths were due to infectious diseases (Vallin and Meslé, 1988). In the Czech Republic in 1920, out of 169 infant deaths per 1000 live births in total, 95 were due to infections diseases, while 50 (similar as in France) were diagnosed as due to congenital weakness or preterm birth complications. Nowadays, 75% or more of all infant deaths are due to perinatal and congenital causes. Due to extended survival of high-risk infants, both perinatal and congenital causes of death are becoming an important component of postneonatal and child mortality.

Figure 3 shows the trends in IMRs for main groups of causes of death per 100 000 live births in France since 1950. For the

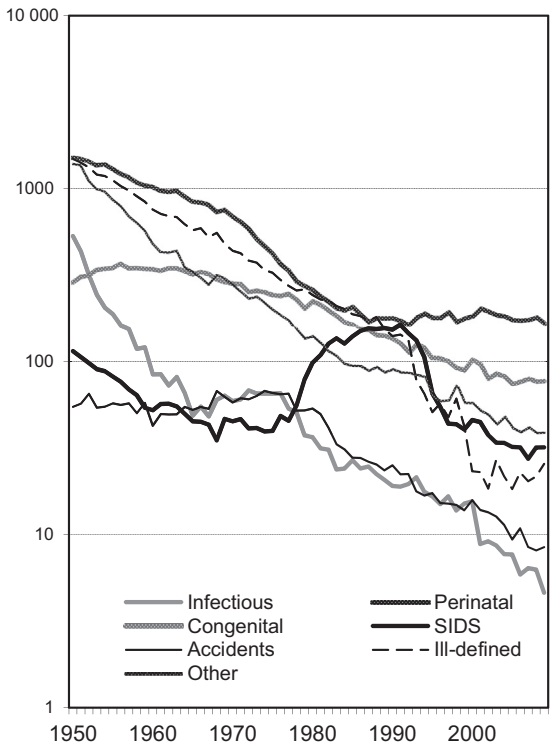

Figure 3 Causes of infant mortality in France since 1950, per 1000 live births. Source: Reconstructed time series of causes of death (Vallin, J., Meslé, F., 1988. Les causes de décès en France de 1925 à 1978. Presses Universitaires de France, Paris).

whole period, perinatal causes of death (typically hypoxia, asphyxia, birth trauma, hemorrhage, and prematurity) were the main cause of infant deaths but declined rapidly. On the contrary, congenital anomalies underwent much slower decline and are now the second leading cause of infant deaths. Deaths from infectious diseases, as well as from other diseases (where pneumonia and gastroenteritis are prominent), became rare. Due to better protection and care, and owing to numerous safety measures and products, accidental mortality has declined since the mid-1970s. In France as in many industrialized countries, an increase in SIDS, formerly called crib or cot death, a diagnostic category gathering sudden deaths in infancy where the initial cause of death remains unknown even after thorough investigation and autopsy, was observed in the 1970s. Deaths from SIDS peak between the second and the fourth month of life and are more likely to occur in boys and in infants sleeping on the belly or on the side. Among alleged risk factors (such as the mother's age, maternal smoking, or preterm birth), sleeping position was targeted by public health 'Back to sleep' campaigns, resulting in rapid decrease of the SIDS incidence (Barbieri, 1998). SIDS, together with congenital malformations and infection, remains the leading cause of postneonatal mortality in most low-mortality countries. Similar analysis for Ukraine for the period 1965–2006 (Meslé and Vallin, 2012) has shown virtually no improvement in perinatal and congenital causes of death, and even an increase in injury and poisoning. Only postneonatal causes (infectious and respiratory diseases) have declined since the 1980s, pointing to large reserves in the health care system in the ex-USSR countries.

Factors of Infant Mortality

Infant mortality is a highly stratified, multivariate, and multilevel process. As infant death prevention was among the first public health target, its risk factors have received long-term attention. Initially, two main determinants were identified: nutritional status (of the mother and the child) and cleanliness of the environment. Today, a variety of factors affecting infant mortality are identified and customarily classified as biological, socioeconomic, and environmental:

1. Environmental factors include national or local level of economic development, medical awareness, access to health care, access to education, climatic conditions, and air pollution.
2. Socioeconomic factors are measured by the position of the household (individual indicators of education, occupation, income, marital status) or by area-based indicators of deprivation. In research of a general theory of infant mortality decline, mother's educational status is believed to play central role (Masuy-Stroobant, 2002);
3. Biologic factors encompass individual characteristics on the mother's and the child's side (gestation length, birth weight, sex, race, birth order, birth interval, multiple birth, mother's age, genetic predisposition, maternal stress, nutrition status, etc.).

Because of a great deal of interaction, it is impossible to clearly distinguish the separate effect of each factor. For example, social disadvantage, independent of its measure, increases both the risk of infant death and that of preterm or low birth weight delivery (Weightman et al., 2012). Typically, the factors operate in a multiplicative sequence: socioeconomic deprivation would increase the risk of maternal smoking and/or poor nutritional status, which is associated with higher risk of low birth weight or preterm birth, which multiplies the risk of infant death. Birth weight and gestational age – indicators of intrauterine development – are now the strongest predictors of neonatal and infant survival. Births before 37 weeks of gestation are classified as preterm (as very preterm before 32 weeks), while the accepted dividing line for low birth weight is 2500 g and 1500 g for very low birth weight. The neonatal mortality regularly increases both with lower and the higher-than-average birth weight, and this pattern remains unchanged in spite of large recent reductions (Wilcox, 2001). Gestational-age-specific mortality data are scarcer than birth-weight-specific data, but follow similar gradient of increasing risk with shorter gestation duration (Mohangoo et al., 2011). The resulting IMRs are thus affected by (1) survival of low birth weight and preterm babies and (2) the share of these high-risk newborns in the population. The latter increases in many industrialized countries due to more frequent multiple births, postponement of childbearing, increase in delivery management techniques (induced labors, Cesarean sections), and worsening of social disparities. Lowest shares of low birth weight neonates (under 5%) are currently found in Scandinavian countries, while in Hungary, Portugal, Spain, Greece, Bulgaria, United States, and Japan, the proportions exceed 8% (OECD, 2011). Low birth weight babies have about 20 times higher infant mortality than heavier ones (Wilcox, 2001). The survival of low birth weight neonates has, however, significantly improved: in the United States, a drop from 694 to 422 deaths per 1000 live births was observed for the very low birth weight category (500–1499 g) and from 94 to 36 for low birth weight category (1500–2499 g) between 1960 and 1980 (Buehler et al., 1987). Important country-specific differences in low birth weight survival exist between Central and Eastern Europe and between Western countries, especially for very low birth weight and very preterm births (Rychtaříková, 1999; Masuy-Stroobant, 2002; Mohangoo et al., 2011). The birth-weight-specific neonatal mortality rate nowadays ranges from 0.5 deaths per 1000 live births for the best-performing category (3500–3999) to approximately 400 deaths per 1000 live births for the first 'viable' category (500–749 g) (based on the US data, Alexander et al., 2003). Birth weight also predicts mortality at different gestational age: thus, even in perfectly term births (40–41 weeks of gestational age), mortality risk is considerably higher for low birth weight babies. This disproportion has been attributed to the intrauterine growth restriction syndrome. Lowest neonatal mortality rates are often found for low birth weight babies in disadvantaged populations, which is known as the low birth weight paradox. The paradox has been observed in small babies of smokers, African-Americans, or less educated and single women (MacMahon et al., 1965; Yerushalmy, 1971; Syrovátka and Rychtaříková, 1984; Wilcox and Russel, 1990). The existence of the paradox stresses the importance of distinction between birth weight and gestational age: among low birth weight babies, only the small and preterm ones are at highest mortality risk (Wilcox, 2001). As preterm births are now responsible for up to two-thirds of all infant deaths (based on US data, MacDorman and Mathews, 2008), preventing

spontaneous preterm birth and intrauterine growth retardation, together with reducing of social inequalities, is a necessary condition for further infant mortality decline.

Less significant biological factors include sex of the child (consistently higher risk of infant death in males), age of the mother (J-shaped risk pattern with lowest risk in the mid-childbearing age), birth order (associated with higher death risk especially if at inappropriate age, e.g., third child at age 20–24 Rychtaříková, 1999), and multiple birth (positively correlated with both low birth weight/prematurity and death risk). The biosocial stratification of infant mortality follows a similar pattern in all industrialized countries, but is generally stronger in Central and Eastern Europe than in traditionally industrialized countries.

Child Mortality

In worldwide concepts such as the WHO development strategies, child mortality is usually expressed as the probability of dying under age 5 years (the so-called under-5 mortality rate). In demographic terminology, child mortality refers only to the period between the first and the fifth birthday and as such is measured by the (postinfantile) child mortality rate: the sum of all deaths at age 1–4 years related to the average number of children alive within the same interval. During the eighteenth and the nineteenth centuries, child mortality rate was approximately five times lower than IMR. The twentieth-century victory over infectious diseases resulted in more pronounced drop in child mortality compared to infant mortality and resulted in an increase of this ratio. In 1950, lowest child mortality (less than 2 deaths per 1000 alive) was found in Nordic countries, Austria, Switzerland, the United States, and Canada. Highest values were observed in Portugal (13), Japan (10), Spain (7), or Bulgaria (8). Nowadays, child mortality is as low as 0.15 deaths per 1000 alive in the Nordic countries; 0.2 deaths per 1000 alive in Western and Central Europe, Canada, and Japan; 0.3 deaths per 1000 alive in the Baltic countries and in the United States; and 0.5 deaths per 1000 alive in Ukraine and Russia. Main causes of child death are nontraffic accidents (mainly accidental drowning and falls), diseases of infancy (perinatal and congenital), cancer, and road accidents. All these causes have regularly decreased since the 1960s due to better protection, prevention, and health care provided to increasingly precious children. Similar to infant mortality, socioeconomic differentials are observed in child mortality. As the health status between conception and 5 years sets the human capital in later life, reducing inequalities in infant and child health would help to equalize health status through the remaining life course.

See also: Abortion, Demography of: United States; Children, Value of; Early Life Influences on Health and Mortality in Adulthood; Infant and Child Mortality in the Less Developed World; Mortality, Biodemography of; Mortality, Epidemiological, and Health Transitions; Mortality: The Great Historical Decline.

Bibliography

Alexander, G.R., Kogan, M., Bader, D., Carlo, W., Allen, M., Mor, J., 2003. US birth weight/gestational age-specific neonatal mortality: 1995–1997 rates for whites, hispanics, and blacks. Pediatrics 111 (1), e61–e66.

Barbieri, M., 1998. La mortalité infantile en France. Population 4, 813–837.

Bourgeois-Pichat, J., 1951. La mesure de la mortalité infantile, I: principe et méthodes. Population 6 (2), 233–248.

Buehler, J.W., Strauss, L.T., Hogue, C.J.R., Smith, J.C., 1987. Birth weight-specific cause of infant mortality, United States, 1980. Public Health Reports 102 (2), 162–171.

Gourbin, C., Masuy-Stroobant, G., 1995. Registration of vital data: are live births and stillbirths comparable all over Europe? Bulletin of the World Health Organization 73 (4), 449–460.

Lantoine, C., Pressat, R., 1984. Nouveaux aspects de la mortalité infantile. Population 39 (2), 253–264.

MacDorman, M.F., Mathews, T.J., 2008. Recent trends in infant mortality in the United States. NCHS Data Brief 9, 1–8.

MacMahon, B., Alpert, M., Salber, E.J., 1965. Infant weight and parental smoking habits. American Journal of Epidemiology 82 (3), 247–261.

Masuy-Stroobant, G., 1997. Infant health and infant mortality in Europe: lessons from the past and challenges for the future. In: Corsini, C.A., Viazzo, P.P. (Eds.), The Decline of Infant and Child Mortality: The European Experience, 1750–1990. Martinus Nijhoff, Den Haag, pp. 1–34.

Masuy-Stroobant, G., 2002. Les déterminants de la santé et de la mortalité infantiles. In: Caselli, G., Vallin, J., Wunsch, G. (Eds.), Démographie: Analyse et Synthese, vol. 3. Institut National d'Études Démographiques, Paris, pp. 129–144.

McKeown, T., 1976. The Modern Rise of Population. Arnold, London.

Meslé, F., Vallin, J., 2012. Mortality and Causes of Death in 20th-Century Ukraine. Springer, Dordrecht.

Mohangoo, A.D., Buitendijk, S.E., Szamotulska, K., Chalmers, J., Irgens, L.M., Bolumar, F., Nijhuis, J.G., Zeitlin, J., 2011. Gestational age patterns of fetal and neonatal mortality in Europe: results from the Euro-peristat project. PLoS One 6 (11).

OECD, 2011. Infant health: low birth weight. In: Health at a Glance 2011: OECD Indicators. OECD Publishing.

Preston, S.H., Haines, M.R., 1991. Fatal Years: Child Mortality in Late Nineteenth Century America. Princeton University Press, Princeton.

Rychtaříková, J., 1999. Social and biological factors of infant mortality [Article in Czech] Demografie 41 (2), 95–104.

Syrovátka, A., Rychtaříková, J., 1984. Naissances vivantes et décès de moins d'un an selon le poids à la naissance en République socialiste tchèque entre 1950 et 1980. Population 39 (3), 515–539.

Vallin, J., Meslé, F., 1988. Les causes de décès en France de 1925 à 1978. Presses Universitaires de France, Paris.

Weightman, A.L., Morgan, H.E., Shepherd, M.A., Kitcher, H., Roberts, C., Dunstan, F.D., 2012. Social inequality and infant health in the UK: systematic review and meta-analyses. BMJ Open 2, e000964.

Wilcox, A.J., Russel, I.T., 1990. Why small black infants have a lower mortality rate than small white infants: the case for population-specific standards for birth weight. The Journal of Pediatrics 116, 7–10.

Wilcox, A.J., 2001. On the importance—and the unimportance— of birthweight. International Journal of Epidemiology 30 (6), 1233–1241.

Woods, R.I., Watterson, P.A., Woodward, J.H., 1988. The causes of rapid infant mortality decline in England and Wales, 1861–1921, Part 1. Population Studies 42 (3), 343–366.

Yerushalmy, J., 1971. The relationship of parents' cigarette smoking to outcome of pregnancy–implications as to the problem of inferring causation from observed associations. American Journal of Epidemiology 93 (6), 443–456.

Zakharov, S.V., 1996. The second world war as a turning point of infant mortality decline in Russia. In: Masuy-Stroobant, G., Gourbin, C., Buekens, P. (Eds.), Santé et mortalité des enfants en Europe: inégalités sociales d'hier et d'aujourd'hui. Academia-Bruylant, Louvain-la-Neuve, pp. 311–333.

Infant and Child Mortality in the Less Developed World

Magali Barbieri, Institut National d'Études Démographiques, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

About 7 million children below 5 years of age die every year around the world. The situation is particularly dire in sub-Saharan Africa where infant and child mortality rates are 15 times higher than in high-income regions. Over two-thirds of all child deaths are attributable to a small number of infectious and parasitic diseases aggravated by malnutrition, and most could be avoided with the mass implementation of simple and low-cost interventions. Government commitment along with the support of international organizations has led to tremendous improvements over the past few decades; however, in the poorest countries, the challenges of AIDS and the resurgence of diseases such as malaria and cholera have slowed the positive trends observed since the 1950s.

In 2014, about 7 million children below 5 years of age will die around the world. Of which, 3.4 million deaths will occur in sub-Saharan Africa, 2.3 million in South Asia, and fewer than 100 000 in the developed world. Globally, this death toll represents more than one child death every 5s. The vast majority of these children die from diseases preventable or treatable with simple and low-cost medical techniques. Despite this depressing toll, tremendous progress has been achieved since the 1950s. Worldwide, the proportion of children dying below 5 years of age declined from about 200 per 1000 births around 1950 to 120 per 1000 in 1980–85, and further to about 55 per 1000 in 2005–10. Compared to the conditions of the 1950s, improvements in health care and sanitation have resulted in the survival of nearly 20 million children who would have died every year.

Mortality Measures

Numbers of child deaths are rough estimates rather than exact measures because accurate data on child health and mortality in developing countries are not easy to attain. The gold standard in estimating the level of child mortality is to compute the ratio of the number of deaths in a given year to the number of births reported to the vital statistics system. However, only one-quarter of the world's population lives in countries with complete vital registration, and these are countries with the lowest mortality. To circumvent this problem, major data collection efforts have been undertaken since the 1980s and techniques for estimating child mortality have improved so remarkably that it is now possible to make reliable comparisons among all the world's regions and to better understand disease processes in developing countries.

In countries where vital registration data are incomplete or unreliable, nationally representative household surveys constitute the main source of statistical information. Starting in the 1970s, international programs, such as the World Fertility Surveys and the Demographic and Health Surveys funded by the United States Agency for International Development, have become the main sources of such data. More recently, a complementary program of data collection, the Multiple Indicator Surveys, has been developed by UNICEF to monitor progress toward the Millenium Development Goals. These surveys often provide birth history data that can easily translate into estimates of child mortality.

The Millenium Development Goals are eight major development targets established during a United Nations sponsored summit of government leaders in 2000 to improve the conditions of life of the most economically deprived populations around the world. The fourth of these goals is to reduce the mortality of children under 5 years of age by two-thirds by 2015 compared to 1990. Two measures are traditionally used by demographers to characterize child mortality: the infant mortality rate and the under-five mortality rate. The infant mortality rate is the ratio of deaths below 1 year of age during a given calendar year to the number of live births during that same calendar year. The under-five mortality rate is the proportion of children who die before reaching their fifth birthday in a 1-year cohort of live births. Both rates are usually provided per 1000 births.

Geographic Disparities

The mortality of young children has declined everywhere around the world; however, the favorable trend observed at the regional level conceals a wide diversity of situations (Figure 1). About half of the 50 least developed countries still experience an under-five mortality rate above 100 per 1000 and differences between regions and among countries in each continent continue to widen.

Latin America was the first developing region to benefit from medical progress. Among developing regions, the decline in child mortality in Latin America is more uniform than in Asia or Africa. Improvements in child mortality began in the early decades of the twentieth century and have been steady since. Currently, the overall probability of dying below five is around 20 per 1000. Even the poorest countries (namely Haiti, Bolivia, and Honduras) exhibit child death rates well below 100 per 1000, and in many countries children's survival chances are now close to those observed in Europe or North America. This is particularly true in the Caribbean where under-five mortality rates around 10 per 1000 are not uncommon.

Asia exhibits the largest internal diversity. Overall the average probability of dying below five is now less than 45 per 1000 – but it is three times higher in South Asia than in East

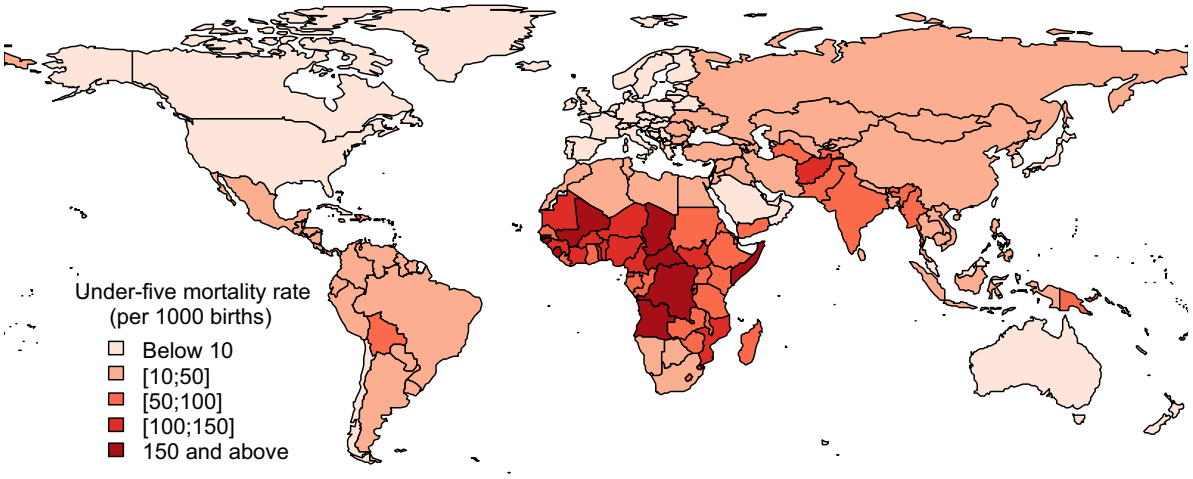

Figure 1 Under-five mortality rate in 2011. Source: Map produced by the author using the R worldmap package (A. South, 2013) and data in Level and Trends in Child Mortality. Report 2011. Estimates developed by the UN Inter-agency Group for Child Mortality Estimation (UNICEF, WHO, World Bank, UN DESA, UNDP).

Asia. Variations are even wider among countries within the same continental subregion. For example, the infant mortality rate reaches 75 and 72 per 1000 in Yemen and Pakistan, respectively, more than 15 times the level recorded in Cyprus and Israel. Similarly, in Southeast Asia, it is over 40 per 1000 in Laos and Cambodia, compared to less than 5 in Japan and in the Republic of Korea. The overall level of mortality in Asia is largely driven by the two most populated countries of the planet, China and India, which experience rates of infant mortality of 13 and 47 per 1000, respectively, and an under-five mortality of 15 and 60 per 1000. Nearly a quarter of all deaths under 5 years of age, or a total of 1.7 million, occurs in India because of the country's size and its relatively high level of child mortality.

Africa remains by far the region with the highest overall level of infant and child mortality. One in 15 children dies during his/her first year and one in 10 before reaching his/her fifth birthday. Wide disparities are observed between Northern and Southern Africa, on the one hand, and Western, Central, and Eastern Africa on the other. In Northern and Southern Africa, the under-five mortality rate is around 30–40 per 1000 while it is above 80 per 1000 in nearly every African country outside of these two regions. Several of these countries exhibit extremely high levels of under-five mortality: in Angola, the Central African Republic, Chad, the Democratic Republic of Congo, Guinea-Bissau, Mali, Sierra Leone, and Somalia, more than 15% of all children die before their fifth birthday. Even there however, survival has improved, with a decline in the under-five mortality rate on the order of 25% over the past 15 years or so.

Intraregional disparities are more the result of diverging trends over time than variations in the initial levels of mortality. Indeed, historical research suggests that the infant mortality rate was well above 200 per 1000 in preindustrial Europe as well as in developing countries prior to the twentieth century. By 1950, in contrast to Asia and Africa, Latin America had achieved an average level of mortality significantly below the pretransitional level. During the following decades, while mortality continued to decline steadily in Latin America, the other two main continents of the developing world experienced sharp drops in infant and child mortality (Figure 2), with the decline in Asia during 1955–75 being especially marked.

While both regions exhibited similar levels of infant mortality around 1950 (about 180 per 1000), the proportion of children dying during their first year fell by 40% in Asia (from 160 to 100 per 1000) in 20 years, a remarkable achievement largely credited to substantial improvements in China. During the same period, the proportional decline was about 30% in Latin America and 25% in Africa.

Medical Causes of Infant and Child Deaths

As socioeconomic development proceeds, countries move through the epidemiological transition. The concept was coined by Omran (1971) to describe the shift in disease composition from infectious and parasitic diseases to congenital and degenerative diseases. This trend parallels a change in the age structure of mortality, with an increase in the average age at death. More than 60% of all deaths are experienced by children below 5 years of age in high-mortality populations compared with 3% in industrialized nations. As the level of mortality declines, deaths tend to concentrate increasingly in the months, weeks, and days immediately after birth.

The Rising Share of Neonatal Mortality

Neonatal mortality, which refers to mortality during the first 4 weeks of life, now represents over 40% of the overall mortality of children under 5 years of age worldwide. A quarter to half of these early deaths takes place within 24h of delivery. They are mostly attributed to intrapartum conditions, especially the complications of prematurity and low birth-weight (for about one-third of them) as well as birth asphyxia (for another 23%), and to infectious diseases (35%), sepsis,

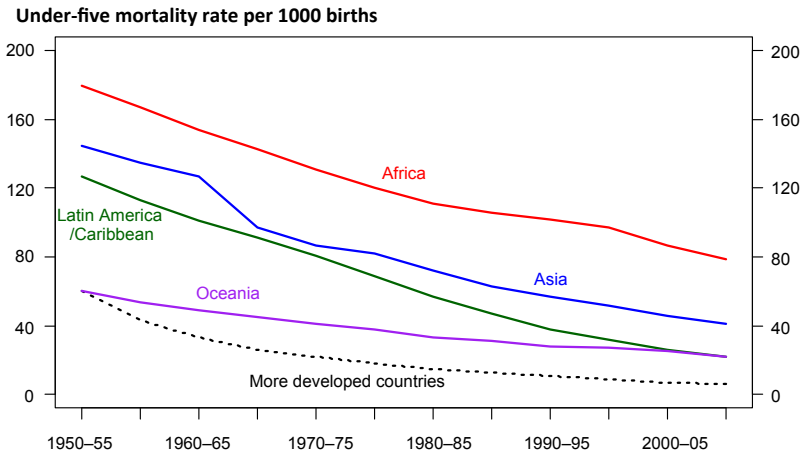

Figure 2 Infant mortality rate by region: 1950–55 to 2005–10. Source: World Population Prospect 2010, United Nation Population Division, 2011.

meningitis, pneumonia, tetanus, and diarrhea in particular (Lawn et al., 2006). The vast majority of neonatal deaths could be prevented without intensive care, through the implementation of low-technology interventions such as prenatal checkups and maternal immunization, hygienic practices during birth, thermal care, treatment for infections, and breastfeeding support. Because these measures must be taken sequentially, before, during, and after delivery, they have been promoted through the concept of a continuum of care. Continuity in care is more costly than most of the interventions directed at older children though, since a more sustainable health system with a larger workforce is required, but this approach would also avert part of the nearly 3 million stillbirths and a large proportion of the 290 000 maternal deaths occurring every year. Because progress has been slower for neonatal than for postneonatal and early childhood mortality, the proportion of child deaths occurring during the first weeks of life is higher where the overall level of under-five mortality is lower, from about 55% in some regions of Asia (Western Pacific in particular) to less than 25% in Africa, and neonatal conditions now represent the leading cause of death among all children worldwide.

The Predominance of Infectious and Parasitic Diseases

After the first month of life, deaths are predominantly attributed to infectious and parasitic diseases. Over 60% are due to only three diseases or disease categories (Johnson et al., 2010; Liu et al., 2012). The highest toll is caused by acute respiratory infections (Williams et al., 2002). Pneumonia alone directly results in 1 million deaths of children between 1 and 60 months of age in developing regions. Pneumonia is also believed to indirectly contribute to an additional 2 million child deaths for which another underlying cause is reported. It is particularly deadly in South and Southeast Asia with more than 30% of all child deaths due to this cause in some areas. Diarrheal diseases are the second leading cause in less developed regions, with almost one in five of all child deaths. Malaria continues to kill more than half a million children every year. It is the leading cause of mortality among children in Africa (Rowe et al., 2006). The other main causes of child death are external causes (injury), meningitis, AIDS, and measles. Most of these diseases also represent a high proportion of morbidity in industrialized countries. They are, however, much less lethal there because of differences in the immune system of children born in these countries: malnutrition is a major factor in the weakening of natural resistance, while immunizations are fundamental to the building of artificial resistance.

The Role of Malnutrition

The major role of nutrition in child mortality processes is now well documented. The effect of malnutrition on mortality was overlooked for a long time because malnutrition is rarely the direct cause of death. Fewer than 200 000 deaths of children below 5 years of age are attributed to either marasmus or kwashiorkor, the most acute forms of malnutrition. However, recent research has demonstrated that malnutrition is an initial or aggravating factor in one-third to half of all child deaths. Malnutrition is usually the result of a combination of inadequate dietary intake and infection. It affects over one-quarter of all the children under 5 years of age in the developing world.

There exist different kinds of malnutrition. Malnutrition not only results from a deficiency of protein and calories but also from a lack of vital dietary minerals and vitamins. International organizations single out deficiencies of four micronutrients (iron, iodine, vitamin A, and zinc) as being particularly common and of special concern for children in developing countries. The general effect of malnutrition is to lower the defenses in children's immune systems, resulting in a greater incidence, severity, and duration of common childhood diseases. Vitamin A deficiency makes children especially vulnerable to infections and worsens the course of many infectious diseases. Death rates in children lacking vitamin A are 20–30% higher than in other children. Vitamin A deficiency has been shown more particularly to reduce children's resistance to diarrhea and measles. Not only does malnutrition increase the probability of infectious and parasitic diseases, but

diseases also speed nutrient loss and suppress appetite, provoking a vicious circle ultimately leading to death. For instance, in parts of Africa where malaria is common, studies show that about one-third of the cases of child malnutrition are caused by malaria.

In an attempt to reduce the incidence of malnutrition in the earlier months of life, international organizations such as the WHO and UNICEF have been actively promoting breastfeeding. Though breastfeeding durations had declined significantly during the 1980s and early 1990s due to the combined effects of commercial pressure for the use of formula and the increased participation of women in the labor force, such efforts have proved useful in encouraging breastfeeding in the poorest countries and the proportion of newborns exclusively breastfed has increased in a number of countries over the past decade or two (Cai et al., 2012). Breastmilk contains all the nutrients, antibodies, hormones, and antioxidants an infant needs to thrive during the first 6 months of life. Breastfed children not only show better immune responses to immunizations, but they also exhibit natural resistance against a number of infectious diseases, gastrointestinal and respiratory in particular. By comparison, commercial infant formulas are an expensive and inferior substitute for breastfeeding. They are frequently overdiluted in order to save money and mixed with unsafe water before being fed to children in unsterilized bottles capped with unclean nipples. Research indicates that in countries where infant mortality rates are high or moderately high, a bottle-fed baby in a poor community is 14 times more likely to die from diarrheal diseases and 4 times more likely to die from pneumonia than a baby who is exclusively breastfed. Unfortunately, only one in three infants below 6 months of age is currently in the latter situation.

Devastating Epidemics

Old curses have recently intensified. Epidemics of plague and upsurges in cholera have demonstrated a surprising rebound in parts of Africa, Asia, and Latin America. Malaria has also showed frustrating resistance to the many approaches tried to combat the disease. The parasite is no longer vulnerable to some of the most powerful antimalarial compounds available. For a time, the widespread use of pesticides reduced the incidence of the anopheles mosquito transmitting the parasite, but it, too, has developed resistance.

In addition to such historically deadly diseases, AIDS has become a major killer of young children in parts of the developing world and it is now the sixth leading cause of child deaths worldwide (UNAIDS, 2012). The AIDS epidemic is a main public health issue in Africa and in parts of Asia, threatening millions of children with premature death. Its real impact is very difficult to measure because of a lack of statistical data and problems of diagnosis in the countries most affected by the epidemic. The burden of the disease has been steadily increasing for children over the 1980s and 1990s but appears to have decreased since. According to recent estimates from the Joint United Nations Program on HIV/AIDS (UNAIDS, 2012) and WHO, about 200 000 children below 15 years of age are now dying every year from the disease, compared to 500 000 children 15 years ago. The number of newly infected children, 90% of them born in sub-Saharan Africa, is estimated as 330 000 in 2011. This is however less than half the number in 2003. The number of new infections has declined everywhere but in Angola, Congo, Equatorial Guinea, and Guinea-Bissau. The majority of infected children have acquired the infection directly from their mothers. Mother-to-child transmission of HIV, also known as vertical transmission, occurs during pregnancy, during labor and delivery, and through breastfeeding. Research has showed that a child stands the greatest risk of vertical transmission (about 20%) at the end of the pregnancy and during labor. There is an additional 15% risk that an infant will become infected through breastmilk. Consequently, more than one in three of the babies born to mothers infected with HIV stand at risk of infection. The vast majority of those infected die before the 5 years of age (Sutcliffe et al., 2008). The cumulative number of children infected with HIV/AIDS had reached more than 6 million by the end of the twentieth century, and the mortality of young children had been increasing up to 50% in countries most affected by the epidemic, mainly located in Eastern and Southern Africa such as Namibia or South Africa. In Zimbabwe, AIDS was the biggest single killer of the nation's children under 5 years of age at the turn of the twenty-first century. Because of the very slow changes in sexual behavior and the difficulty of developing an effective vaccine, prevention of mother-to-child transmission through measures such as artificial infant feeding rather than breastfeeding as well as the increasing use of antiretroviral drugs have played an essential role in lowering the incidence of the disease among children and increasing the survival odds of those who become infected. Though nearly 6 in 10 pregnant women infected with HIV are now receiving antiretroviral prophylaxis in sub-Saharan Africa, not all subregions are doing well and transmission rates still reach about 30% in Central and Western Africa (UNAIDS, 2012).

HIV/AIDS remains a substantial burden in countries with poor health care systems. AIDS is an expensive disease. In communities with limited resources, the epidemic diverts a tremendous proportion of the available budget away from basic health care toward intensive care. In countries with mature epidemics, such as Uganda, Zambia, and Zimbabwe, some doctors report that up to three-quarters of the beds in hospital pediatric wards have at times been occupied by children ill from AIDS. Another indirect effect of the epidemic is the dramatic increase in orphanhood. Since the beginning of the epidemics, nearly 16 million children have been orphaned because of the disease, more than 95% of them in sub-Saharan Africa, and the survival odds of orphan children are significantly lower than those of others. Efforts will need to expand quickly for the UNICEF goal of zero new child infection by 2015 to be achieved.

Economic Determinants

Poverty is viewed as the underlying factor for high child mortality in developing countries. At the family level, economic privation clearly affects the health of children. Low purchasing power is related to poorer nutrition, the lack of

ready access to a safe water supply and proper sanitation, and unhygienic conditions in and around homes. Such conditions are typically associated with a high prevalence of childhood diarrhea and they have significant implications for the spread of infectious and parasitic diseases.

However, at the national level, the relationship between national income and infant and child mortality is much less systematic. Wealthier countries are clearly those in which children are less likely to die. As indicated in Figure 3, countries with a GNP per capita above $5000 have a level of infant mortality below 25 per 1000 and this has changed very little over the past three decades.

Nonetheless, the relationship does not appear to be determinative since countries faring the worst in terms of health and mortality are not always the poorest. There is a wide range of situations among poor countries with similar levels of national income per capita. For example, while GNP per capita in Indonesia is more than twice than in Vietnam ($3500 vs $1400), infant mortality is lower in the latter (at 17 per 1000 compared with 25 per 1000 in Indonesia). The contrast is even more striking in sub-Saharan Africa where the ratio of infant mortality reaches three to one in countries with the same level of GNP per capita ($500–550), such as Rwanda or Madagascar on the one hand (at 50 and 60 per 1000, respectively), and the Central African Republic on the other (at over 160 per 1000). Equatorial Guinea is an extreme case with a GDP per capita higher than that of Portugal and Greece and close to that of Spain and Italy thanks to its oil exports, but a level of under-five mortality that is among the highest in the world, with nearly 12% of children dying before reaching their fifth birthday. Poverty per se is thus insufficient to explain geographic variations in infant and child mortality across the world and other factors must be taken into account.

Political Commitment

Experience and research demonstrate that without a commitment at the government level the health of children improves very slowly. Poor countries experiencing a low level of mortality relative to their level of socioeconomic development are those where a primary health care strategy was adopted early and implemented steadily at the national level, following recommendations first developed at the 1978 WHO conference in Alma-Ata. The strategy recommended direct medical intervention (such as immunizations and oral rehydration therapy), nutrition programs, and activities indirectly affecting health, such as hygiene, education, and family planning campaigns. China, Cuba, Vietnam, and Costa Rica are well-known examples of poor countries where the primary health care strategy has been successfully implemented. Even countries that have not been able to implement the strategy as a whole but only in part have benefited from a decline in child mortality. Among the main achievements of the strategy are the declines recorded over the past decade in the mortality attributable to pneumonia (−22%), diarrhea (−18%), and measles (−18%) due to low-cost interventions like the distribution of antibiotics, oral rehydration salts and zinc and immunization, as well as better hygiene practices, improved water and sanitation facilities, and adequate nutrition, including exclusive breastfeeding for children below 6 months of age.

In many regions of the world, the international community has played an important role in promoting primary health care services. WHO, UNICEF, and many other public or private organizations have worked with governments toward the goal of immunizing the world's children against six preventable diseases. Due to very active campaigns carried out during the 1980s, rates of child immunization against measles, tetanus, poliomyelitis, tuberculosis, pertussis (whooping cough), and

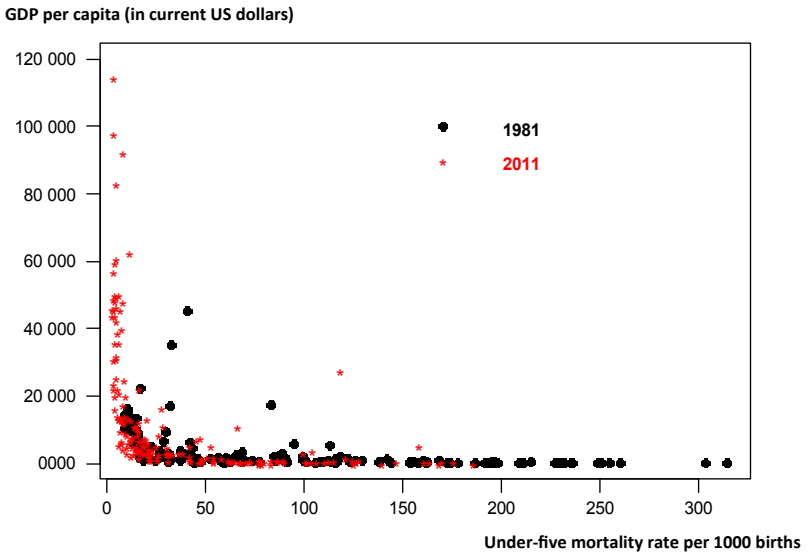

Figure 3 Relationship between the probability of dying below five and national income per capita in 1981 and in 2011. Source: For the under-five mortality rate, Level and Trends in Child Mortality. Report 2011. Estimates Developed by the UN Inter-agency Group for Child Mortality Estimation (UNICEF, WHO, World Bank, UN DESA, UNPD); for the GDP per capita: World Bank national accounts data, and OECD National Accounts data files. Both series downloaded by the author from http://data.worldbank.org/indicator on 29 January 2013.

diphtheria have risen from about 10% in the mid-1970s to 30% 10 years later and 80% around 1990. The proportion of children vaccinated leveled out over the 1990s but increased again after 2000 and immunization rates against the main six targeted diseases now reach 85 to 90% overall (75–85% in sub-Saharan Africa). This achievement has meant the prevention of up to 3 million child deaths a year (http://www.who.int/gho/immunization/en/index.html, consulted 6 January 2014).

In conclusion, dramatic progress has been accomplished since the 1970s, so that children in many Asian and Latin American countries do not experience a higher risk of dying than those born in industrialized regions. There remain sources of serious concern for the future of child health in the most deprived areas of the Third World, however, and inequalities continue to increase among subgroups of national populations and among countries. As of 2012, only half of the world developing regions and 15 of the 66 countries with over 40 under-five deaths per 1000 live births are on track to achieve this Millenium Development Goal for 2015 (UNICEF et al., 2012). The challenge is to circumvent economic and political difficulties in order to implement the health interventions known to accelerate the reduction of infant and child mortality.

See also: Infant and Child Mortality in Industrialized Countries; Mortality and the HIV/Aids Epidemic; Mortality, Epidemiological, and Health Transitions; Mortality: The Great Historical Decline.

Bibliography

Boschi-Pinto, C., Velebit, L., Shibuya, K., 2008. Estimating child mortality due to diarrhea in developing countries. Bulletin of the World Health Organization 86 (9), 710–717.

Cai, X., Wardlaw, T., Brown, D.W., 2012. Global trends in exclusive breastfeeding. International Breastfeeding Journal 7 (12). http://www.internationalbreastfeedingjournal.com/content/7/1/12.

Johnson, H.L., Liu, L., Fischer-Walker, C., Black, R.E., 2010. Estimating the distribution of causes of death among children age 1–59 months in high-mortality countries with incomplete death certification. International Journal of Epidemiology 39, 1103–1114.

Lawn, J.E., Wilcynska-Ketende, K., Cousens, S.N., 2006. Estimating the causes of 4 million neonatal deaths in the year 2000. International Journal of Epidemiology 35, 706–718.

Liu, L., et al., 2012. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. The Lancet. http://dx.doi.org/10.1016/S0140-6736(12)60560-1. Published online 11 May.

Omran, A.R., 1971. The epidemiologic transition: a theory of the epidemiology of population change. Milbank Memorial Fund Quarterly 49 (4), 509–538.

Rowe, A.K., et al., 2006. The burden of malaria mortality among African children in the year 2000. International Journal of Epidemiology 35, 691–704.

Sutcliffe, C.G., van Dijk, J.H., Bolton, C., Persaud, D., Moss, W.J., August 2008. Effectiveness of antiretroviral therapy among HIV-infected children in sub-Saharan Africa. The Lancet 8, 477–487.

UNAIDS Joint United Nations Programme on HIV/AIDS, 2012. Global Report: UNAIDS Report on the Global AIDS Epidemic 2012. UNAIDS.UN Inter-agency Group for Child Mortality Estimation (UNICEF, WHO, World Bank, UN DESA, UNPD), New York. Level & Trends in Child Mortality. Report 2012.

UNICEF, 2008. The State of the World's Children 2008. Oxford University Press, Oxford, UK.

UNICEF, World Health Organization, The World Bank, United Nations, 2012. Levels and Trends in Child Mortality. Report 2012. Estimates Developed by the UN Inter-agency Group for Child Mortality Estimation. New York.

United Nations, 2010. World Population Prospects: The 2010 Revision, Volume I. Comprehensive Tables. Department of Economic and Social Affairs, Population Division. United Nations, New York.

Williams, B.G., Gouws, E., Boschi-Pinto, C., Bryce, J., Dye, C., January 2002. Estimates of world-wide distribution of child deaths from acute respiratory infections. The Lancet Infectious Diseases 2, 25–32.

World Bank, 2013. Online Indicator Database. http://data.worldbank.org/indicator/NY.GDP.PCAP.CD (accessed 25.01.13.).

Relevant Websites

http://www.measuredhs.com/ – Demographic and Health Surveys.

http://www.healthmetricsandevaluation.org/ghdx/record/child-mortality-estimates-and-mdg-4-attainment-country-1990-2011 – Institute for Health Metrics and Evaluation.

http://www.unicef.org/statistics/index_24302.html – Multiple Indicator Cluster Surveys.

http://www.unaids.org/en/dataanalysis/ – UNAIDS Statistics.

http://www.childinfo.org/ – UNICEF Statistics.

http://www.un.org/millenniumgoals/ – United Nations, Millenium Development Goals project.

http://www.un.org/esa/population/ – United Nations Population Division.

http://data.worldbank.org/indicator/ – World Bank Indicators.

http://www.who.int/gho/child_health/en/index.html – World Health Organisation, Child Health.

http://www.who.int/healthinfo/statistics/mortality_neonatal/en/index.html – World Health Organisation, Neonatal Mortality.

Infant Gaze Behaviors

Rechele Brooks, Institute of Learning & Brain Sciences, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The perception of eye gaze is a crucial component of social understanding from infancy to adulthood. Individuals can use eye gaze as a nonverbal means to communicate interests, intentions, and other mental states. Gaze behaviors increase in complexity with development. This article reviews the emergence of four gaze behaviors: visual scanning, gaze cueing, alternating gaze, and gaze following. A deep understanding of infants' gaze behaviors has implications for language, autism spectrum disorders, robotics, and theory of mind.

Although many human social interchanges occur via mobile phones and other technologies, communication in person is still a key context for social interaction. In face-to-face conversations, perception of eye gaze is a crucial factor in social understanding. By using eye gaze, individuals can convey that they are attracted to a person, reveal that they are lying, or indicate that they have completed their conversational turn (Kleinke, 1986). In short, gaze behaviors serve many functions in social exchanges.

Eye gaze is both an entry point for understanding other people's minds and a way to determine what information is relevant in the environment. From the adult view, eye gaze is an essential component to understanding a person's mental states, including emotions and intentions. However, a person's eye gaze can also refer to something in the external world. For example, if a person looks up into the sky, onlookers follow his or her gaze. The gaze followers want to look at what that person sees above them. Adults realize that people are in visual contact with objects and obtain information by looking at them. The developmental question is when does a psychological understanding of eye gaze emerge.

Theoretical Background

Infants' abilities to make inferences about people's mental states – including perception, emotion, and beliefs – remain intensely debated (Flavell, 2004; Metcalfe and Terrace, 2013). This debate extends into the interpretations of infants' gaze behavior. Though infant acts often appear outwardly the same as adult ones, these actions may have different underlying meanings. On one side of the debate, theorists argue eye gaze already has psychological meaning to young infants (e.g., Baron-Cohen, 1995). On the opposing side, others suggest that infants' gaze behaviors reflect simpler mechanisms based on behavioral explanations, such as noticing contingencies between head/eye orientation and external objects (e.g., Moore, 2008). However, we should consider a third perspective (Brooks and Meltzoff, 2002): Infants' understanding of gaze may change. Infants gradually come to understand the psychological nature of eye gaze.

Gaze behaviors have implications for development, because they can advance infant cognitive and social learning. When infants follow the gaze of others, infants can learn about desires, emotions, and perceptions of others, as well as identifying significant information in their surroundings. When infants alternate their gaze to and from an object, they have a nonverbal means to initiate and maintain social interactions. The importance of these actions is highlighted by the fact that many gaze behaviors are impaired in children with autism spectrum disorders (ASD) (Baron-Cohen, 1995; Toth et al., 2006). Other research has established that gaze following predicts subsequent language acquisition for children with typical development and atypical development (Brooks and Meltzoff, 2008; Kasari et al., 2008; Mundy et al., 2007), in part because this ability enables the child to coordinate what a person is looking at with what he or she is labeling.

Current Knowledge

There are at least four distinct gaze behaviors that are widely studied: visual scanning, gaze cueing, alternating gaze, and gaze following. These gaze behaviors are reviewed in order of complexity and with attention to their functional significance. Developmental changes in gaze following will be used to discuss infants' emerging understanding of eye gaze.

Visual Scanning of the Face

Neonates scan the eye region of the face (Johnson et al., 2008). They prefer to look at faces with eyes oriented forward rather than ones with eyes averted to the side. By 2 months of age, infants fixate on the eyes more than the mouth; declines in scanning the eyes by 6 months of age predict developmental outcomes for children at risk for autism (Jones and Klin, 2013). These early scanning patterns may reflect an automatic reaction to eyes, which may lay the foundation for understanding mutual gaze. However, detecting eye gaze may rely on the physical properties of the display, such as the direction of high-contrast spots. As a minimum definition, visual scanning of the face simply requires that infants look at facelike stimuli.

Gaze Cueing

Gaze cueing refers to when infants shift their own gaze after being cued by a centrally positioned stimulus, such as a face with averted eyes. In the typical gaze-cueing paradigm, a screen

displays a digitized human face with eyes that shift laterally before a target becomes visible on the left or the right of the screen (Hood et al., 1998). The experimental manipulation is whether the eyes move to the same side as the upcoming target ('cued' target trials) or whether the eyes move to the opposite side ('uncued' target trials). In most studies, the face disappears before the onset of the targets. The results show that infants and adults have a faster response to the cued than the uncued targets (Johnson et al., 2008). For example, infants look faster to the left target after the eyes shift to the left. This effect is even found with newborns who are shown simplified, schematic faces.

However, there are reasons to suggest gaze-cueing responses are different than following the gaze of actual people. First, the cueing effect is disrupted when the face (and eyes) remains visible on the screen (Hood et al., 1998). Second, presenting high-contrast black spots (instead of eyes) results in a similar cueing effect for newborns (Johnson et al., 2008). Third, the timing of newborns' and young infants' gaze shifts to the target suggests that they are automatically, rapidly processing the central cue. Finally, infants' target looks could occur because of displayed motion. When they respond to the apparent movement, they need to only slightly shift their gaze to the materializing targets. This is in direct contrast to live interactions. When a mother looks at an object, her face does not disappear; nor do objects suddenly appear immediately near her face. Thus, these gaze-cueing responses could be explained by different mechanisms than those that underlie intentionally following a person's gaze. As a minimum definition of cueing, infants shift their own gaze in response to a lateral shift of spots.

Alternating Gaze

When infants look at an object and then make eye contact with their parents, infants are 'alternating gaze.' Instances of this gaze shift behavior may occur during play when infants look up from their toy to their parent's face. Infants typically begin to use gaze in this manner by 9 months of age.

The likelihood that infants will engage in alternating gaze varies with the social context. When infants are already facing an adult tester and a toy, 9-month-olds alternate their gaze from an object to the adult's eyes as often as 18-month-olds do (Mundy et al., 2007). In contrast, during unstructured play with their parents, 9-month-old infants briefly and rarely initiate this behavior even though they are playing near their parent (Carpenter et al., 1998). Infants' gaze alternation significantly increases with age during naturalistic play. By 18 months, infants more frequently shift their eye contact from an object to a person, but they still spend less than half of their unstructured play during periods of alternating gaze (Adamson et al., 2004). In both contexts, alternating gaze can serve a similar function of showing interest in their social partner and an object. The use of alternating gaze also occurs when infants face an unfamiliar, ambiguous object. When infants alternate their gaze in this context, they are said to be 'social referencing' (*see* Social Referencing during Infancy and Early Childhood across Cultures). After looking to their parent (or other adult), infants regulate their own behavior, such as by modifying their approach toward the ambiguous object.

From the adult perspective, eye gaze creates a psychological connection between social partners and an object. Infants may alternate their gaze to indicate or refer to an object. To find evidence of infants' understanding, researchers examine whether infants *initiate* eye contact with their social partner after looking at an object (Carpenter et al., 1998; Mundy et al., 2007). In this manner, infants' gaze helps them to establish an object as a topic of conversation. They can use their own gaze to 'show' an object to their parent, who in response may comment on it. However, when making eye contact, infants may be simply checking what their parent is doing (Moore, 2008). Infants may look at their parent's face out of habit, based on an established social routine without awareness of shared visual attention. As a minimum definition of alternating gaze, infants disengage their attention from an object and shift their gaze to a person's eyes.

Gaze Following

The prototypical example of gaze following is when an individual looks where another person just looked. However, this should exclude instances when two individuals simultaneously look at the same object, such as after hearing a noise. A full understanding of gaze following incorporates a psychological concept of perception: An adult wants to *see* what the looker *perceives*. The developmental question is when does this understanding begin.

Before their first birthdays, infants track the looking behavior of people. However, in most situations, a person turns her *head* and her eyes toward a target object. Because of the multiple cues from the person, it is unclear which one infants follow. Infants could follow salient head movement, which could pull their attention in a direction and to the target by accident. By using motion salience, infants would not be gaze following at all. They would be responding to the person without understanding the importance of the person's eyes or perception.

To find evidence of gaze following, Brooks and Meltzoff (2002) tested whether infants follow eye gaze or head motion with an open- versus closed-eyes test. As is typical in gaze-following procedures, an adult directly faced the infant and then turned toward one of two targets. In this study, there were two conditions: open versus closed eyes. In one group, the adult made eye contact with the infant, silently turned her head, and looked at the object with open eyes. In the other group, she acted in the same manner, except that she had her eyes closed as she turned her head. This action provides a control for head motion. The manipulation helps to address an issue fundamental to perception because we see with our eyes and not our head. An important step toward the psychological interpretation of seeing is to recognize that eyes are essential for looking. If infants recognize that eyes allow visual contact with an object, they should only look at the target when the adult looks at it (with open eyes). If infants follow head motion, they should look at the target in both conditions – without regard for the adult's eyes.

The test revealed that infants at 18, 14, and even 12 months of age noticed the perceptual status of the adult's eyes (Brooks and Meltzoff, 2002). Infants at all three ages differentiated the adult's turns with open eyes from the ones with closed eyes. They rarely looked at the target in the closed-eyes condition even though the adult's nose and head moved toward the target. By 12 months of age, infants selectively looked at the

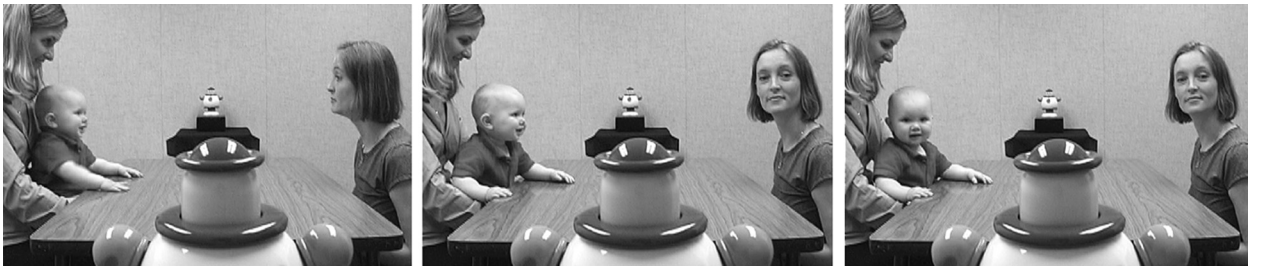

Figure 1 An example of gaze following: The adult silently turns to the target and then the 12-month-old infant follows her look. Reprinted from Meltzoff, A.N., Brooks, R., 2007. Intersubjectivity before language: three windows on preverbal sharing. In: Bråten, S. (Ed.), On Being Moved: From Mirror Neurons to Empathy. John Benjamins, Philadelphia, PA, p. 156, Reprinted with permission.

adult's target when her eyes were open: Infants were following eye gaze (**Figure 1**).

Infants also coordinated their own gaze with other social behaviors. They vocalized and pointed at the target. This combination is particularly striking given that children with ASD often have deficits with each of these behaviors (Kasari et al., 2008). The typically developing infants used these acts more often in the open-eyes than the closed-eyes condition. In addition, though the objects in the room remained the same, they acquired a special valence when the adult looked at them. Infants looked at the target longer when the adult could see it than when she could not. By 12 months of age, infants treated the adult's gaze like a psychological spotlight on the distal object.

When does gaze following emerge? To examine this issue Brooks and Meltzoff (2005) tested infants at 9, 10, and 11 months of age with the open- versus closed-eyes test. The results were intriguing because there was a developmental shift. At 9 months, infants did not discriminate open eyes from closed eyes. The 9-month-olds followed the adult's turns when her eyes were closed just as often as when her eyes were open. In contrast, at 10 and 11 months of age, there was a significant effect of eye status. For the closed-eyes group, 10- and 11-month-old infants refrained from following the adult's turns. In the open eyes group, they looked at the adult's target. These findings show that gaze following emerges by 10 months of age. By 10 months, but not 9 months of age, infants were truly following the gaze of the adult.

The evidence helps explain the behavior of 9-month-old infants and younger ages. Though infants as young as 3–4 months of age react to adults' looking behavior (Moore, 2008), they are not specifically following eye gaze. They are responding to the person regardless of whether her eyes are open or closed (Brooks and Meltzoff, 2005). The pattern suggests that young infants look in the direction indicated by head motion or postural changes. The findings support the conclusion that young infants are 'body following' and construe people as orienting in a direction.

In contrast to the younger ages, by 10 months of age, infants selectively follow the looks (not just the head turns) of a person to distal targets (Brooks and Meltzoff, 2005). This is a significant developmental change in infants' understanding of eye gaze. Just 1 month earlier, at 9 months of age, infants do not make this distinction. By 10–12 months of age, infants treat others as visual perceivers. The converging evidence suggests that infants link the looker to the target object (Brooks and Meltzoff, 2002; Flavell, 2004). The infants make this connection specifically when the person has open eyes, but not closed eyes. This is a crucial step in recognizing people as intentional perceivers. By 10 months of age, infants have achieved at least a minimum definition of gaze following because they look where a person's eyes point.

Change in Focus

In recent years, many researchers have introduced findings based on eye tracking software. In most studies, the eye tracker works with a 2-D display of prerecorded images of a person's looking behavior. The systems provide time-accurate coding of latency and duration of gaze fixations. The presentation format provides new opportunities to test how different conditions affect infants' gaze behaviors. For example, Senju and Csibra (2008) used an innovative 2-D design to manipulate the cues for social communication (e.g., eye contact or infant-directed speech) in a gaze-following procedure. As more research uses 2-D paradigms, it is important to consider how well the 2-D findings will generalize to the 3-D world. There may be differences in infants' reactions to these experiences. Infants may more readily learn from a real person than a 2-D image of a person (e.g., as with language learning, Kuhl et al., 2003). Future research should explore whether the same factors identified in 2-D tests hold true for in-person interactions.

Current and Emerging Direction in Theory and Research

Language Development

Research demonstrates that infants' gaze behaviors relate to their language development. Two behaviors – gaze following and alternating gaze – are significant factors. The other reviewed gaze behaviors, visual scanning and gaze cueing, have rarely been tested as predictors of language acquisition. Gaze following provides a potent means to learn words, because infants who follow gaze can identify what object is being labeled. (This works well because parents often look at what they are naming.) The empirical evidence shows that gaze following reliably predicts receptive and productive language. For example, infants with high gaze-following scores at 10–11 months of age later have rapid productive vocabulary growth through 24 months of age (Brooks and Meltzoff, 2005,

2008). This gaze-language relationship remains significant even after accounting for infants' cognitive abilities (Mundy et al., 2007).

Alternating gaze also has a link to vocabulary development. As infants look from an object to the parent, they can monitor what their parent talks about. Their gaze alternation also invites parents to name an object of the infants' choice (e.g., often called 'follow-in labeling'), and thus, infants' gazing promotes increased word-learning opportunities. The observational research reveals significant effects for both infants' alternating gaze and parents' follow-in labeling (Carpenter et al., 1998). Of interest, one-year-olds' alternating gaze relates to their later vocabulary, even after accounting for maternal labeling. In experimental tests of infants' word learning, by 16 months of age, infants look up at the adult's face during object naming and are able to learn words under conditions of follow-in labeling. By 18 months of age, they learn labels for objects whether or not the adult uses follow-in labeling (Sage and Baldwin, 2010). The converging findings demonstrate that infants' own gaze behaviors play a role in word learning.

Autism Spectrum Disorders

Gaze behaviors are also of interest because they are often an area of deficiency for children with ASD. Children with ASD generally have less interest in sharing attention. They often have other social deficits that make it difficult for them to learn about others' intentions and mental states (Baron-Cohen, 1995; *see* Autism Spectrum Disorders).

A concern is whether children with ASD notice and follow the gaze of others. In a prospective longitudinal study, infants later diagnosed with ASD scanned faces at 2 months of age as much as a typical infant, but they looked less at faces by 6 months of age (Jones and Klin, 2013). This reduced attention to the eyes after 2 months of age could lead to less experience with the social information that eyes express. Even by 3- to 4-years of age, children with ASD are less likely to respond to or initiate joint attention than children with typical development (Toth et al., 2006). Nevertheless, gaze following and alternating gaze are strong predictors of language for children with ASD. Given that gaze behaviors emerge early in typical development (e.g., before expressive language), assessments of and interventions on gaze understanding can affect long-term outcomes of the development of children with ASD (Kasari et al., 2008). The work in this field highlights the interconnectedness of gaze behaviors, language, and the development of an understanding of mental states.

Robotics

Infants' gaze behaviors are also instrumental in the field of robotics. The modeling of gaze behaviors is part of this growing interdisciplinary field of research. Engineers use developmental data to build models and to train robots to follow people's looks (Sage and Baldwin, 2010). For developmental psychologists, on the other hand, gaze following provides a test of what cues inform infants about agency. For instance, work with socially interactive robots allows for a detailed investigation of when an object becomes a social entity (Meltzoff et al., 2010). The robot in this study was metallic and had black circles for eyes (i.e., the edges of camera lenses). To an infant, would this piece of metal ever become a social agent with a gaze to follow? To test this question, infants observed the robot and an adult under different social conditions. In the experimental group, infants watched as the robot and the adult socially engaged (e.g., playing imitation games). In control groups, the robot or the adult used the same actions, but the two did not have a social interchange (e.g., the robot moved but the adult simply watched). The results showed that infants were more likely to follow the 'gaze' of a robot after they saw the robot's social exchange with a person than when a social interaction was not achieved. Other work has shown that when robots contingently interact with young children, the robot is treated like an autonomous social agent. These examples and others demonstrate that infants' gaze behaviors can help evaluate social-cognitive development (Meltzoff et al., 2009).

Theory of Mind

Gaze-following paradigms also provide insights into the development of the understanding of mental states, often called 'theory of mind' (*see* Theory of Mind during Infancy and Early Childhood and Cultural Variation, Development of). For example, to adults, eye closure and other barriers equivalently block line of sight. However in infancy, eye closure is different from other obstructions (e.g., blindfolds or walls). As previously discussed, at 12 months of age, infants rarely follow the head turns of an adult who has closed eyes. In contrast, 12-month-olds make an interesting 'mistake' – they follow the head turns of an adult wearing a blindfold just as often as one wearing a headband (Brooks and Meltzoff, 2002). The findings suggest that infants understand eye closure in advance of other barriers. There is an experiential difference between the different visual obstructions. Infants have extensive firsthand knowledge of eye closure from opening and closing their own eyes. These daily experiences may help them understand how eye closure affects their own and others' visual perception. If true, infants' own experiences could powerfully change what infants understand about other people's visual perception.

Meltzoff and Brooks (2008) put the self-experience proposal to a test by using a blindfold to block infants' views. During a gamelike activity the infant looked at toys on a table. The experimenter periodically interposed the cloth blindfold between the infant and a toy. This repeatedly gave 12-month-old infants the direct experience of being unable to see when the opaque blindfold was lifted in front of their eyes during the training experience. In a control condition, the blindfold cloth had a window in the center, so the cloth did not block the infants' view. As an additional control, there was a baseline group in which infants had no training with the blindfold's visual properties. Thus, only the opaque blindfold group had the firsthand experience of having the blindfold block their view of the world. To assess infants' learning, all infants saw the experimenter wear the blindfold during a gaze-following test. This was the first time that the infants saw the experimenter put the blindfold on her eyes. The blindfolded experimenter turned toward distal targets. The blindfold outwardly looked the same on the experimenter. Would the infants interpret the adult looking as a function of their own experiences?

The self-experience had a significant effect on how the 12-month-old infants interpreted the adult's behavior (Meltzoff and Brooks, 2008). Infants in the control groups (i.e., window and baseline groups) still followed the blindfolded adult's turns. These infants acted like the naïve 12-month-olds in the previous study. In contrast, infants in the opaque blindfold group stopped following the 'gaze' of the blindfolded adult. Infants generalized their experiences with the opaque cloth to their social partner. After being unable to see when a blindfold was in front of their own eyes, they subsequently inferred that the adult could not see in a similar situation. Hence, they refrained from following the adult's turns. Their experiences had an immediate effect and completely changed how they treated the adult. The implication is that infants' firsthand experiences helped them infer what others could (or could not) see.

Given the laboratory findings that infants' own experiences influence their social cognition (Meltzoff and Brooks, 2008), this leads to the theoretical question of what naturally occurring (daily) experiences affect development. Cultures vary in how they structure their social interactions, which could modify the frequency and type of social cues available to infants (Akhtar and Gernsbacher, 2008 for a review; *see* Joint Attention during Infancy and Early Childhood Across Cultures). As patterns of social interaction differ by culture, will infants differ in which skills they learn? For example, the infant's experience of being carried facing away or toward their mother could affect how much they follow their mother's gaze versus posture. In another vein, when cultural practices emphasize learning through observing others (Rogoff et al., 2003), these practices potentially improve how readily infants monitor the gaze and the actions of other people. They may become adept at observing the gaze of others from afar (i.e., even when the person is not directly interacting with them). Thus, the study of cultural practices would further illuminate the effects of children's own experiences on their understanding of others.

Conclusions

Examining infants' gaze behaviors helps illuminate social-cognitive development. The reviewed work demonstrates that infants gradually change their gaze behaviors. The proposed interpretation is that infants alter their behavior because they are developing a psychological interpretation of gaze. As infants decode the looks of other people, they can learn about what others see as well as what they talk about. Missing these essential gaze cues is especially noteworthy for children with ASD. Deficits in gaze behaviors may contribute to subsequent difficulties in language and social interactions for children with ASD. The gaze behaviors of infants with typical and atypical development are potentially interconnected to other social behaviors, such as pointing, social referencing, and prosocial understanding. The field of robotics offers opportunities to examine how acts become construed as socially meaningful by infants with typical and atypical development. Future work should further explore the interrelationships between the understanding of gaze and other aspects of social cognition, including face perception, goal detection (e.g., intentionality), and theory of mind.

See also: Autism Spectrum Disorders; Joint Attention During Infancy and Early Childhood Across Cultures; Social Referencing During Infancy and Early Childhood across Cultures; Theory of Mind During Infancy and Early Childhood Across Cultures, Development of.

Bibliography

Adamson, L.B., Bakeman, R., Deckner, D.F., 2004. The development of symbol-infused joint engagement. Child Development 75, 1171–1187.

Akhtar, N., Gernsbacher, M.A., 2008. On privileging the role of gaze in infant social cognition. Child Development Perspectives 2, 59–65.

Baron-Cohen, S., 1995. Mindblindness: An Essay on Autism and Theory of Mind. MIT Press, Cambridge, MA.

Brooks, R., Meltzoff, A.N., 2002. The importance of eyes: how infants interpret adult looking behavior. Developmental Psychology 38, 958–966.

Brooks, R., Meltzoff, A.N., 2005. The development of gaze following and its relation to language. Developmental Science 8, 535–543.

Brooks, R., Meltzoff, A.N., 2008. Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: a longitudinal, growth curve modeling study. Journal of Child Language 35, 207–220.

Carpenter, M., Nagell, K., Tomasello, M., 1998. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. Monographs of the Society for Research in Child Development 63 (4). Serial No. 255.

Flavell, J.H., 2004. Development of knowledge about vision. In: Levin, D.T. (Ed.), Thinking and Seeing: Visual Metacognition in Adults and Children. MIT Press, Cambridge, MA, pp. 13–36.

Hood, B.M., Willen, J.D., Driver, J., 1998. Adult's eyes trigger shifts of visual attention in human infants. Psychological Science 9, 131–134.

Johnson, M.H., Grossmann, T., Farroni, T., 2008. The social cognitive neuroscience of infancy: Illuminating the early development of social brain functions. In: Kail, R.V. (Ed.), Advances in Child Development and Behavior, vol. 36. Elsevier, San Diego, CA, pp. 331–372.

Jones, W., Klin, A., 2013. Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. Nature 504, 427–431.

Kasari, C., Paparella, T., Freeman, S., Jahromi, L.B., 2008. Language outcome in autism: randomized comparison of joint attention and play interventions. Journal of Consulting and Clinical Psychology 76, 125–137.

Kleinke, C.L., 1986. Gaze and eye contact: a research review. Psychological Bulletin 100, 78–100.

Kuhl, P.K., Tsao, F.-M., Liu, H.-M., 2003. Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. Proceedings of the National Academy of Sciences 100, 9096–9101.

Metcalfe, J., Terrace, H.S. (Eds.), 2013. Agency and Joint Attention. Oxford University Press, New York.

Meltzoff, A.N., Brooks, R., 2008. Self-experience as a mechanism for learning about others: a training study in social cognition. Developmental Psychology 44, 1257–1265.

Meltzoff, A.N., Brooks, R., Shon, A.P., Rao, R.P.N., 2010. "Social" robots are psychological agents for infants: a test of gaze following. Neural Networks 23, 966–972.

Meltzoff, A.N., Kuhl, P.K., Movellan, J., Sejnowski, T.J., 2009. Foundations for a new science of learning. Science 325, 284–288.

Moore, C., 2008. The development of gaze following. Child Development Perspectives 2, 66–70.

Mundy, P., Block, J., Vaughan Van Hecke, A., Delgado, C., Venezia Parlade, M., Pomares, Y., 2007. Individual differences and the development of joint attention in infancy. Child Development 78, 938–954.

Rogoff, B., Paradise, R., Mejía Arauz, R.M., Correa-Chávez, M., Angelillo, C., 2003. Firsthand learning through intent participation. Annual Review of Psychology 54, 175–203.

Sage, K.D., Baldwin, D., 2010. Social gating and pedagogy: mechanisms for learning and implications for robotics. Neural Networks 23, 1091–1098.

Senju, A., Csibra, G., 2008. Gaze following in human infants depends on communicative signals. Current Biology 18, 668–671.

Toth, K., Munson, J., Meltzoff, A.N., Dawson, G., 2006. Early predictors of communication development in young children with autism spectrum disorder: joint attention, imitation, and toy play. Journal of Autism and Developmental Disorders 36, 993–1005.

Infectious Diseases: Psychosocial Aspects

Andrew Steptoe and Lydia Poole, University College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Infectious diseases, also known as communicable diseases, are illnesses caused by invasive pathogens such as viruses and bacteria. Psychosocial factors (including life stress, mood, social support, and coping responses) have been shown to influence the acquisition of infections such as the common cold, and the activation of latent infections such as genital herpes. Coping styles have also been shown to affect the course of infectious illness and the duration of symptoms. Several mechanisms are involved in mediating these effects, including the suppression of the individual's immune defenses, changes in the reporting of symptoms, and health-related behaviors such as smoking and physical activity.

Infectious diseases are caused by invasive pathogens, such as viruses, bacteria, fungi, and parasitic infestations. When the germ theory of disease was established by Pasteur and Koch in the nineteenth century, a mechanistic view of infectious disease became dominant, with illness depending on exposure to the agent and the virulence of the infection. However, the ascent of immunology in the mid-twentieth century shifted the balance toward recognition that the strength of bodily defenses plays a role in susceptibility to infectious disease.

Psychosocial factors are relevant to the experience of infectious disease in two principal ways:

1. Psychosocial factors may influence the acquisition of infection or the reactivation of latent infection. Much recent research on upper respiratory illnesses such as the common cold and on herpes virus infections (genital and oral herpes) has been directed at this possibility.
2. Psychosocial factors may affect the course of infectious illness once it is developed, influencing the duration and severity of symptoms.

Readers interested in more detailed overviews of psychosocial aspects of infectious disease are referred to more comprehensive reviews (Pedersen et al., 2010; Segerstrom, 2012).

Mediating Processes

There are four general processes through which psychosocial factors such as acute and chronic life stress, social support, and coping may influence the acquisitional course of infectious disease.

Exposure to Pathogens

Since infectious diseases are caused by pathogens, any changes in exposure that are stimulated by psychosocial factors may influence the chances of becoming ill. For example, some people respond to distressing events in their lives by seeking social support and mobilizing social networks. This might have the unwanted effect of exposing the individual to more respiratory infections. Alternatively, people may respond by social withdrawal, reducing contact with airborne pathogens. Differences in exposure to infections resulting from crowding and other factors probably account for much of the socioeconomic gradient observed in conditions such as infectious mononucleosis (glandular fever). Exposure to pathogens is a key issue in the prevention of outbreaks of serious waterborne, airborne, and sexually transmitted infections in the developing world, but little psychosocial research have yet touched on this topic. One study, which has addressed this issue, used data from the National Health and Nutrition Examination Survey III to examine the association between sociodemographic factors and pathogen burden in a sample of over 19 000 adults (Zajacova et al., 2009). The pathogens studied included *Helicobacter pylori*, cytomegalovirus, herpes simplex virus-1, and hepatitis B. These authors found that individuals with less education, lower income, and ethnic minority status were more likely to have been infected by these pathogens, probably because of greater exposure earlier in life than their more privileged counterparts.

Changes in Behavior

Health behaviors and lifestyle have an influence on infectious disease. Smoking, alcohol consumption, and physical activity all influence vulnerability to infection and the duration of symptoms. Malnutrition is probably the principal cause of diminished immunity and susceptibility to infection across the world. Although severe protein energy malnutrition is rare in developed countries, undernutrition is common among the elderly and impairs immune resistance. Other health behavior parameters such as sleep duration and physical exercise appear to influence the effectiveness of vaccination against infections (Edwards et al., 2008; Prather et al., 2012). If the pattern of health behaviors is altered in relation to psychosocial factors such as life stress, then the acquisition and course of infectious disease may be modified. Behaviors can have direct effects on infectious disease, such as when sexual behavior stimulates the reactivation of latent herpes virus, or the acquisition of human papillomavirus and risk of cervical cancer. Behaviors related to health service utilization are associated with socioeconomic status in the developing world, and play an important role in determining risk of diseases such as malaria (Onwujekwe, 2005). Moreover, behaviors surrounding health decision-making processes

such as uptake of the human papillomavirus vaccination have been found to be dependent on psychosocial factors such as cultural and religious beliefs (Marlow et al., 2009). Research investigating the influence of ethnic differences in such health programs is important, since greater insights will help tackle health inequalities.

Symptom Reporting

People vary substantially in the extent to which they attend to physical sensations and recognize them as symptoms of illness. Research on illness behavior has shown that the same infection may elicit quite different levels of complaint, work absence, and health service utilization, depending on levels of psychosocial stress. Negative affect also influences symptom reporting and illness behavior, with more symptoms being reported in distressed individuals independently of objective markers of illness (Howren and Suls, 2011).

Suppression of Bodily Defenses

The fourth possibility is that psychosocial factors influence susceptibility and the course of infectious disease by modifying physiological defenses. Depending on the infectious agent, skin abrasions, immune processes, or mucosal barriers may be involved. One of the principal interests in psychoneuroimmunology research has been the possibility that psychosocial factors reduce immune defense, allowing infections to become established that would otherwise be resisted. These hypothetical associations are illustrated schematically in Figure 1. The healthy response to infections involves the individual mounting an adequate defense that excludes the infection and prevents illness (upper panel). In contrast, the impaired response arises when exposure to the infectious agent takes place against a backdrop of compromised immune function due to adverse psychosocial stimulation (lower panel). Under these circumstances, illness may develop. This framework highlights the stress response as a key determinant of the acquisition and progression of infectious diseases via modulation of the immune system (Padgett and Glaser, 2003; Segerstrom and Miller, 2004).

These mediating processes are not mutually exclusive, and one of the difficulties in research on infectious disease lies in teasing out the various pathways through which the acquisition or duration of infection are affected.

Methodological Aspects

The processes mediating between psychosocial factors in infectious disease outlined in Section Mediating Processes have important implications for how research on this topic is carried out.

Verification of Infection

Interpretation of associations is enhanced greatly by objective verification of infection. If infection is assessed merely by the subject or a family member, then it is not possible to rule out symptom reporting biases, and spurious associations between psychosocial factors and unfounded symptoms may be identified. Objective verification allows the timing of infection in

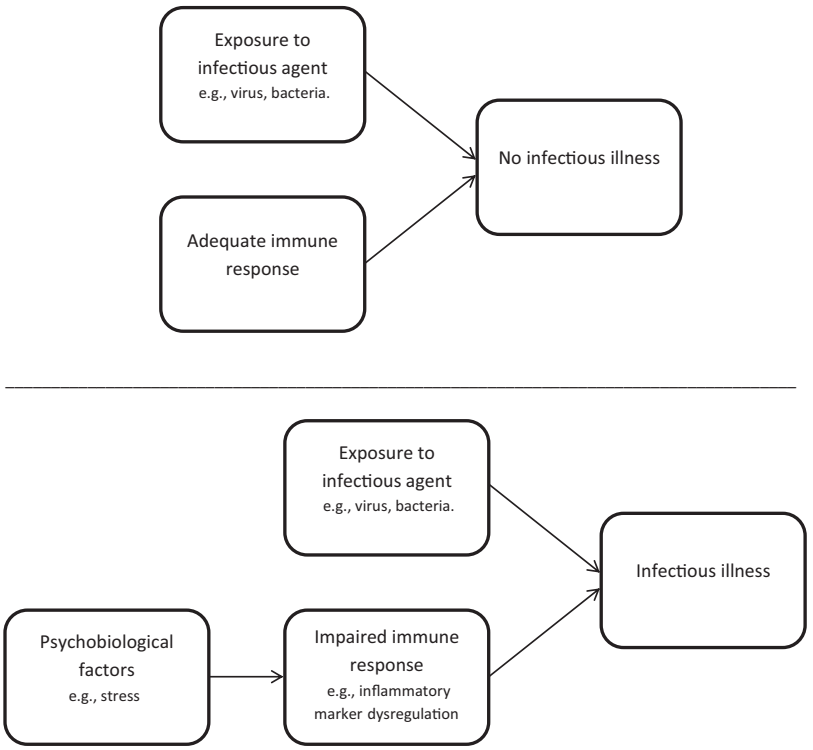

Figure 1 Schematic outline of the links between infection, immune defenses, psychosocial factors, and illness.

relation to psychosocial stimuli to be established, so that causal sequences can be described. The most precise verification is possible in experimental studies of infection, where volunteers are infected deliberately with metered doses of virus or bacteria. The acquisition of infection then is assessed by culturing samples obtained some days later. Unfortunately, biochemical verification of infection is difficult and invasive in naturalistic studies, requiring repeated sampling of blood or nasal washes (for upper respiratory infection) from the individual under investigation. As an alternative, clinical examination by health professionals or trained observers is valuable for objectively assessing the occurrence of infection.

Standardized Measures of Symptoms

Since symptom reporting can be influenced by psychosocial factors independently of objective infection, it is important to reduce variability by employing standardized measures. These measures detail the number, type, and duration of symptoms. Work on upper respiratory infection, for example, typically assesses 10–14 standard symptoms of colds. Such indices are particularly useful in assessing the duration and severity of symptomatic episodes.

Measures of Health Behavior and Biological Factors

One of the most important implications of the discussion of mediators in Section Mediating Processes is that adequate measures are taken of social activity, smoking, alcohol consumption, sleep patterns, diet, and other health behaviors. Analysis of these variables will establish whether changes in health behaviors account for variations in vulnerability to infections. These measures often rely on self-report and so are subject to reporting and recall bias. Technological innovations have improved objective assessment of some behaviors such as physical activity and sleep (using accelerometers). Measurement of hormonal and immunological factors via blood and saliva samples is, of course, extremely interesting in developing an understanding of mediating biological processes (Steptoe and Poole, 2010).

Longitudinal Designs

Many studies of infectious disease have utilized cross-sectional or retrospective designs, with assessments of the frequency of infection in the past, along with measures of stress and other factors. Such studies have been used to find out whether the number of infections over the past month or year was greater among individuals who experience more life stress or have poorer social support than others. More recent work has focused on a life course approach to the study of infectious illness, trying to disentangle the role of early life exposure to psychosocial adversity on later disease vulnerability (Pace et al., 2006; Taylor, 2010). Cross-sectional and retrospective studies are limited in the inferences that can be drawn, since selective recall biases may operate. For example, it is possible that individuals with depressed mood will recall more episodes of infection, and also report more negative life experiences or poorer social support than others. As with many topics in health psychology, the most convincing data derive from longitudinal studies in which psychosocial factors are assessed prior to illness, so that prospective links can be established. Studies involving repeated measures of psychosocial, behavioral, and biological factors are particularly strong, since they allow the time course of vulnerability to be established.

Experimental Studies

Many studies have investigated factors such as stress in relation to infection in animal models. Animal studies do not suffer from many of the methodological problems inherent in human research, but are limited in terms of the subtlety of psychosocial investigation (Moynihan and Ader, 1996).

Experimental studies of infection in humans are of course rare, and are only justified for mild forms of disease. But an important program of work on colds and other upper respiratory tract infections has been conducted by Sheldon Cohen and colleagues in the USA and UK that throws considerable light on the role of psychosocial factors. In these studies, volunteers free of infection are inoculated with standard doses of the virus responsible for the common cold or influenza. If an appropriate dose is administered, not everyone will become ill. Differences in the subsequent rates of infection and illness must, therefore, be due to differences in vulnerability. Social activities are limited by quarantine, and health behaviors such as smoking are controlled, so psychosocial impairment of bodily defenses can be evaluated.

Cohen et al. (1991) demonstrated with this paradigm that susceptibility to common cold viruses was associated directly with a psychological stress index created through an amalgam of measures of major life events over the past 12 months, perceived stress and negative affect. The dose–response association was independent of age, allergic status, and other risk factors. Later work has established that chronic stressors such as long-term work problems and relationship difficulties are particularly relevant. By contrast, social networks have a protective effect, with lower susceptibility to experimental infection among volunteers with more extensive and diverse social networks. Cohen et al. (1997) found that the rates of infection and illness were more than four times greater in volunteers with limited social networks, compared with high social networks, after adjustment of control variables. It has also been demonstrated that susceptibility to infectious illness is reduced in people from low socioeconomic status background in childhood independently of adult social status (Cohen et al., 2004), and in people who report poor sleep (Cohen et al., 2009), while positive affect is protective (Cohen et al., 2006).

These studies have established clearly that psychosocial factors can influence vulnerability to infectious disease and can account in part for interindividual variability in susceptibility. They have also begun to identify psychobiological mediators, with evidence that associations between psychological stress and the severity of symptoms are related to the levels of the inflammatory cytokine interleukin-6 production in the nasal passages following influenza administration. This suggests that inflammatory mediators may be involved in symptom production in stressed individuals. Cellular aging is also

implicated, with an association being described between short leukocyte telomere length and venerability to experimentally administered rhinovirus (Cohen et al., 2013).

Colds and Upper Respiratory Illness

Upper respiratory illnesses are caused by many different types of microorganism. The common cold is a viral infection, and more than 200 different viruses have been implicated. Acute sore throats are commonly caused by streptococcal bacterial infection, while influenza is a viral disease. Naturalistic studies of psychosocial factors in upper respiratory illness involve assessing spontaneous illnesses in people living their everyday lives. They have shown similar effects to those identified in experimental studies, albeit with less control over confounding factors (Pedersen et al., 2010). For example, some studies have followed volunteer populations over several months with periodic measures of life stress, coping, health behavior, and illness. It has been demonstrated in both adults and children that susceptibility to colds is greater in people who experience high rather than low life stress. Interactions between life-events stress and social support have also been observed, with lower incidence of colds among individuals reporting high support. In addition, episodes of upper respiratory infection have been shown to be preceded for 1 or 2 weeks by high levels of perceived stress or a high frequency of minor life events (hassles). Since the acquisition of an infection is followed by several days of incubation, studies of this kind suggest that the pathways outlined in **Figure 1** (lower panel) may be operating. However, obtaining detailed neuroendocrine or immunological data in naturalistic studies remains a methodological challenge to researchers, and it is possible that some of the associations observed are due to changes in exposure to pathogens rather than psychobiological influences on bodily resistance.

Psychosocial Influences on Symptoms

The interaction between the brain and the immune system is two way, with behavioral function being affected by immune responses as well as psychosocial factors influencing the immune system. It is now recognized that the behavioral concomitants of infectious disease such as lethargy, loss of appetite, increased sleep, and loss of interest in the outside world, collectively termed 'sickness behavior,' are stimulated in part by immune products such as interleukin-1 (Dantzer et al., 2008). These cytokines are released as part of the biological response to infection, and help to coordinate the pattern of behavior that conserves and protects the organism, and may thereby be adaptive in promoting survival.

In humans, psychosocial factors not only affect susceptibility, but also the duration and severity of symptoms. For instance, one study assessed the duration of symptoms of naturally occurring colds in a sample of children (Turner Cobb and Steptoe, 1998). Prior to the study, patterns of coping had been assessed with standardized measures. It was found that children who typically dealt with stress using problem-focused coping strategies tended to suffer from colds for a relatively short time. In contrast, children who stated that they used avoidant coping strategies such as distraction and social withdrawal experienced colds of longer duration. These effects remained significant after statistical adjustment for gender, age, family composition, social class, parental smoking, and parental alcohol consumption. It is possible that those who used distraction and avoidant coping strategies failed to respond directly to their illnesses, and did not take the necessary steps in terms of self-care and activity restriction that might prevent lengthy episodes of symptomatic disease (Falagas et al., 2010).

Recurrence of Herpes Virus Infections

Herpes simplex viruses cause a range of diseases, the most common of which are oral herpes (cold sores) and genital herpes. Different types of herpes virus cause the two diseases. Both forms of herpes infection are common and typically benign, though they may cause considerable distress. In more than half of cases, the virus is not eliminated following first infection, but becomes permanently established in sensory ganglia. In some individuals the virus is reactivated periodically, and this leads to a recurrence of infection. Reactivation is partly under immunological control, and may take place when immune surveillance is impaired by psychosocial factors.

Studies of both types of herpes infection have been carried out which suggest that, in some cases, recurrence of infection is preceded by periods of negative mood or stress. Colds also seem to be important in the triggering of recurrent oral herpes. A meta-analysis of 11 prospective studies indicated that psychosocial stress was associated with recurrence of herpes simplex virus infection (Chida and Mao, 2009). However, the models of stress that have been employed in this area have been very varied, including exposure to negative life events, mood, and daily hassles. Greater work is needed to evaluate the interaction between stressors and psychosocial coping resources.

Another type of herpes virus is Epstein–Barr virus (EBV), responsible for infectious mononucleosis (which may present as a sore throat). A few studies have shown that stress and negative moods are associated with this type of infection. Interestingly, there is also psychoneuroimmunological work that throws light on the role of EBV. It has been shown that antibody titers for EBV are raised by life stress, for example, during the period prior to important academic examinations, or in people experiencing marital conflict (Kiecolt-Glaser et al., 1988). A raised antibody titer suggests that EBV has been reactivated and allowed to replicate due to failure of immune control. If unchecked, the reactivation may lead to fresh episodes of illness.

Other Infections

One important type of infection with extensive psychosocial ramifications is human immunodeficiency virus; this will not be discussed here, since it is examined elsewhere in the Encyclopedia. Other types of infectious illness such as urticaria, streptococcal throat infection, and tuberculosis have also been studied, but data are too limited to draw general conclusions. A large range of more severe infections such as poliomyelitis, rabies, and salmonella have been investigated in animals, with

evidence for increased vulnerability with exposure to psychological stressors (Rabin, 1999). However, there are important variations with species, age, concomitant disease and prior exposure, and these make it difficult to draw simple inferences for humans.

Conclusions and Future Directions

It is likely that research on psychosocial influences on infectious disease will become increasingly important over the next decade. One reason is that it is now recognized that infection and inflammatory processes play an important role in 'noninfectious' diseases such as coronary heart disease. Smeeth et al. (2004) observed that risk of acute myocardial infarction or stroke was elevated in the first 3 days following the diagnosis of a systemic respiratory tract infection, and it is possible that influenza vaccination is protective (Warren-Gash et al., 2009). However, limited psychosocial research on this and other conditions from the perspective of infection has been carried out. Indeed, most research on psychosocial aspects of infection has been conducted on a limited group of infections, and much remains to be learned about resistance to fungi, parasites, and other microorganisms. The recognition of the importance of psychosocial factors in vulnerability and adaptation to infectious disease may lead to new and more effective methods of helping people cope with these illnesses.

See also: Illness Behavior and Care-Seeking; Interoception; Mortality and the HIV/AIDS Epidemic; Psychoneuroimmunology.

Bibliography

Chida, Y., Mao, X., 2009. Does psychosocial stress predict symptomatic herpes simplex virus recurrence? A meta-analytic investigation on prospective studies. Brain, Behavior, and Immunity 23, 917–925.

Cohen, S., Tyrrell, D.A.J., Smith, A.P., 1991. Psychosocial stress and susceptibility to the common cold. New England Journal of Medicine 325, 606–612.

Cohen, S., Doyle, W.J., Skoner, D.P., Rabin, B.S., Gwaltney, J.M., 1997. Social ties and susceptibility to the common cold. Journal of the American Medical Association 277, 1940–1944.

Cohen, S., Doyle, W.J., Turner, R.B., Alper, C.M., Skoner, D.P., 2004. Childhood socioeconomic status and host resistance to infectious illness in adulthood. Psychosomatic Medicine 66, 553–558.

Cohen, S., Alper, C.M., Doyle, W.J., Treanor, J.J., Turner, R.B., 2006. Positive emotional style predicts resistance to illness after experimental exposure to rhinovirus or influenza A virus. Psychosomatic Medicine 68, 809–815.

Cohen, S., Doyle, W.J., Alper, C.M., Janicki-Deverts, D., Turner, R.B., 2009. Sleep habits and susceptibility to the common cold. Archives of Internal Medicine 169, 62–67.

Cohen, S., Janicki-Deverts, D., Turner, R.B., Casselbrant, M.L., Li-Korotky, H.-S., Epel, E.S., Doyle, W.J., 2013. Association between telomere length and experimentally induced upper respiratory viral infection in healthy adults. Journal of the American Medical Association 309, 699–705.

Dantzer, R., O'Connor, J.C., Freund, G.G., Johnson, R.W., Kelley, K.W., 2008. From inflammation to sickness and depression: when the immune system subjugates the brain. Nature Reviews Neuroscience 9, 46–56.

Edwards, K.M., Burns, V.E., Adkins, A.E., Carroll, D., Drayson, M., Ring, C., 2008. Meningococcal A vaccination response is enhanced by acute stress in men. Psychosomatic Medicine 70, 147–151.

Falagas, M.E., Karamanidou, C., Kastoris, A.C., Karlis, G., Rafailidis, P.I., 2010. Psychosocial factors and susceptibility to or outcome of acute respiratory tract infections. International Journal of Tuberculosis and Lung Disease 14, 141–148.

Feldman, P.J., Cohen, S., Doyle, W.J., Skoner, D.P., Gwaltney, J.M., 1999. The impact of personality on the reporting of unfounded symptoms and illness. Journal of Personality and Social Psychology 77, 370–378.

Howren, M.B., Suls, J., 2011. The symptom perception hypothesis revised: depression and anxiety play different roles in concurrent and retrospective physical symptom reporting. Journal of Personality and Social Psychology 100, 182–195.

Katon, W., Lin, E.H.B., Kroenke, K., 2007. The association of depression and anxiety with medical symptom burden in patients with chronic medical illness. General Hospital Psychiatry 29, 147–155.

Kiecolt-Glaser, J.K., Kennedy, S., Malkoff, S., Fisher, L., Speicher, C.E., Glaser, R., 1988. Marital discord and immunity in males. Psychosomatic Medicine 50, 213–229.

Marlow, L.A.V., Wardle, J., Forster, A.S., Waller, J., 2009. Ethnic differences in human papillomavirus awareness and vaccine acceptability. Journal of Epidemiology and Community Health 63, 1010–1015.

Moynihan, J.A., Ader, R., 1996. Psychoneuroimmunology: animal models of disease. Psychosomatic Medicine 58, 546–558.

Onwujekwe, O., 2005. Inequities in healthcare seeking in the treatment of communicable endemic diseases in Southeast Nigeria. Social Science & Medicine 61, 455–463.

Pace, T.W.W., Mletzko, T.C., Alagbe, O., Musselman, D.L., Nemeroff, C.B., Miller, A.H., Heim, C.M., 2006. Increased stress-induced inflammatory responses in male patients with major depression and increased early life stress. American Journal of Psychiatry 163, 1630–1633.

Padgett, D.A., Glaser, R., 2003. How stress influences the immune response. Trends in Immunology 24, 444–448.

Pedersen, A., Zachariae, R., Bovbjerg, D.H., 2010. Influence of psychological stress on upper respiratory infection – a meta-analysis of prospective studies. Psychosomatic Medicine 72, 823–832.

Prather, A.A., Hall, M., Fury, J.M., Ross, D.C., Muldoon, M.F., Cohen, S., Marsland, A.L., 2012. Sleep and antibody response to hepatitis B vaccination. Sleep 35, 1063–1069.

Rabin, B., 1999. Stress, Immune Function, and Health. Wiley-Liss, New York.

Segerstrom, S.C. (Ed.), 2012. Oxford Handbook of Psychoneuroimmunology. Oxford University Press, New York.

Segerstrom, S.C., Miller, G.E., 2004. Psychological stress and the human immune system: a meta-analytic study of 30 years of inquiry. Psychological Bulletin 130, 601–630.

Smeeth, L., Thomas, S.L., Hall, A.J., Hubbard, R., Farrington, P., Vallance, P., 2004. Risk of myocardial infarction and stroke after acute infection or vaccination. New England Journal of Medicine 351, 2611–2618.

Steptoe, A., Poole, L., 2010. Use of biological measures in behavioral medicine. In: Steptoe, A. (Ed.), Handbook of Behavioral Medicine. Springer, New York , pp. 619–632.

Taylor, S.E., 2010. Mechanisms linking early life stress to adult health outcomes. Proceedings of the National Academy of Sciences of the United States 107, 8507–8512.

Turner Cobb, J.M., Steptoe, A., 1998. Psychological influences on upper respiratory infectious illness in children. Journal of Psychosomatic Research 45, 319–330.

Warren-Gash, C., Smeeth, L., Hayward, A.C., 2009. Influenza as a trigger for acute myocardial infarction or death from cardiovascular disease: a systematic review. Lancet Infectious Disease 9, 601–610.

Zajacova, A., Dowd, J.B., Aiello, A.E., 2009. Socioeconomic and race/ethnic patterns in persistent infection burden among US adults. Journals of Gerontology. Series A, Biological Sciences and Medical Sciences 64, 272–279.

Inferences in Discourse, Psychology of

Leo GM Noordman, Tilburg University, Tilburg, The Netherlands
Wietske Vonk, Max Planck Institute for Psycholinguistics and Radboud University Nijmegen, Nijmegen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

An inference is defined as the information that is not expressed explicitly by the text but is derived on the basis of the understander's knowledge and is encoded in the mental representation of the text. Inferencing is considered as a central component in discourse understanding. Experimental methods to detect inferences, established findings, and some developments are reviewed. Attention is paid to the relation between inference processes and the brain.

When listeners or readers understand discourse, they understand much more than what is stated explicitly in the sentences. On the basis of their knowledge of the language and the world, they are able to understand what has been left implicit in the text but is intended to be communicated.

Consider the following text (a):

(a) There were municipal elections yesterday. Because the majority of the lower town voted for the local party, there was a shift toward the left in the city council. But the right-wing party was not completely disappointed. It had expected to lose much more.

In text (a), readers can infer that the local party is a left-wing party; that there is a causal relation between many people voting for a particular party and that party getting a stronger representation; that there are height differences in the town and that the town is probably located on a river bank; that the city council used to be more right oriented; that there is a contrast between the shift toward the left and not being completely disappointed, and consequently that the shift is a reason for being disappointed and that there are other reasons (worse expectations) why the right-wing party was not so disappointed; and that 'it' refers to the right-wing party and not, for example, to the city council. All these pieces of information are not stated explicitly in the text, but readers do understand them. They are called inferences.

An inference is defined as information that is not expressed explicitly by the text but that can be derived from the text on the basis of the comprehender's knowledge and that is encoded into the representation the comprehender constructs of the text. In this article, the notion of inference is restricted to the derivation of implicit information that occurs in spontaneous language processing. These inferences are distinguished from logical deductions in reasoning and problem solving (Johnson-Laird, 1983; see also *Reasoning with Mental Models*), such as the logical deduction 'Some artists are chemists' from the premises 'Some beekeepers are artists. All beekeepers are chemists.' Logic is concerned with the assessment of the validity of arguments in reasoning. Although listeners and readers certainly make logical deductions, and some inferences can be considered as logical inferences, everyday language behavior is not characterized by the evaluation of the validity of the arguments in reasoning.

Two kinds of inferences might be distinguished. The first kind is an inference that is the derivation of new information. This is what in ordinary language is called an inference. For example, from the sentence 'John selected his stranston shoes because there was much mud,' the reader can infer that apparently stranston is a material or brand that has advantages if there is much mud. Given that the reader is not familiar with stranston, the reader cannot know that the conjunction 'because' is correctly used. However, by assuming that the sentence makes sense, the reader can derive the inference as new information. The second kind of inference is an activation of available knowledge. Examples from text (a) are that 'it' refers to the right-wing party, and that, given the contrastive relation indicated by 'but,' a shift toward an opponent party is in general a reason for disappointment. This kind of inference in general is not called an inference in everyday language, but most of the psycholinguistic research focuses on this kind of inference.

Issues in Inference Research

There are several issues that make inferences an important topic of research in language understanding. In understanding a discourse, the number and variety of inferences that may be derived from the discourse seem to be almost unlimited. This may lead to a computational explosion. Yet, the human processing system has limited capacity, and comprehension is accomplished too quickly for many time-consuming inferences to be made. Therefore, the first issue is how to account for the control of inferences: Which inferences are made and which ones are not made? This issue will be discussed in Section Some Established Findings in Inference Research. The second issue concerns the inference as a process. What constitutes the process of making an inference; how does the inferred information get activated and deactivated; how does the information in the text interact with the reader's knowledge? Some aspects of the process of inferring will be discussed in Section Toward a More Differentiated View on Inferences. Two additional issues will be discussed that reflect the development in inference research in the last decade. First, until recently, inferences were considered as a rather isolated phenomenon, while in current research inferences are considered as one of the

components of the comprehension process (Section Inferencing as One Component of Discourse Comprehension). Second, the advent of methods to measure brain activity leads to a better understanding of the relation between inferences and the brain (Section Inferences and the Brain). This overview focuses on inferences in reading, but most of the results apply to listening as well.

Methods in Inference Research

Whether an inference is made and when it is made can be investigated by measuring specific aspects of the reading behavior. The measurements are made during the reading of the words or sentences that trigger the inference (*online methods*) or after reading those words or sentences (*off-line methods*).

Off-line Methods

The assumption underlying off-line methods is that the inferences are encoded into the mental text representation, similar to information that is expressed explicitly by the text. Therefore, they can be investigated in tasks such as reproduction, recognition, and verification. In a reproduction task, the reader is required to reproduce the text after having read it. Information that was not stated explicitly in the text but that is reproduced by the reader is supposed to be inferred by the reader. In a recognition task, the reader has to judge whether particular words or sentences occurred in the text. The rationale of the recognition task is similar to that of the reproduction task. Information that is incorrectly judged as having appeared in the text is assumed to be inferred. Many people who had studied the sentences 'John was trying to fix the birdhouse. He was pounding the nail when his father came out to watch him and to help him do the work' incorrectly recognized the sentence 'John was using the hammer to fix the birdhouse when his father came out to watch him and to help him do the work.' This indicates that an instrumental inference (hammer) was made (Bransford and Johnson, 1973). Reproduction and recognition tasks can indicate whether the inferences are made but in general cannot distinguish whether the inferences are made during reading (at encoding) or during the measurement (at retrieval). In a verification task, readers judge whether the content of a sentence is true or false with respect to the content of the text. Not only the accuracy, but also the reaction time for the verification or the recognition is important. If inferred information is recognized or verified as quickly as explicit information, this is an indication that the inference is made during reading, and not at the moment of the off-line task. Off-line methods are used frequently in combination with online methods to answer the question of when inferences are made.

Online Methods

Online methods are employed during reading and aim to detect the ongoing inference process immediately. The assumption is that inference processes require time. Therefore, they are detected by a long reading time at the moment the inference is made, relative to a control condition in which no inference is made. Reading times are measured in a self-paced reading task in which the reader by pressing a button exposes successive units of text (words or clauses) in a window on a computer screen. The interval between button presses is defined as the reading time for the unit of text. For example, for understanding the second sentence in 'Mary got the picnic supplies out of the car. The beer was warm,' the inference is required that there was beer among the picnic supplies. Therefore, the reading time for the second sentence in this sequence should be longer than in the sequence 'Mary got the beer out of the car. The beer was warm' (Haviland and Clark, 1974). Reading times are also measured in eye-tracking recordings that reveal what the reader looks at and how long. This makes it possible to measure fairly exactly when the inferences are made. Just and Carpenter (1978) found that the gaze duration on 'killer' in 'The killer left no clues for the police to trace' is longer if this sentence is preceded by the sentence 'The millionaire died on a dark and stormy night' than if it is preceded by the sentence 'The millionaire was murdered on a dark and stormy night.' Reading the word 'killer' requires the information that the person was murdered; this information has to be inferred after 'die' and is explicit in 'murdered.'

Another online method is measuring event-related potentials (ERPs), changes in the electric encephalogram due to activities in the brain. This technique is very time sensitive and therefore can reveal ongoing reading processes. Van Berkum et al. (2005) showed that hearers of a discourse make predictions of an upcoming noun. The noun was preceded by a gender-marked adjective whose suffix matched or mismatched the predicted noun. For prediction-inconsistent adjectives, a positive inflection in the ERP wave was observed 50–250 ms after onset of the inflection (before the noun), indicating that listeners anticipate an upcoming word. In addition, a negative inflection was observed at about 400 ms from noun onset, which is generally observed when a word does not fit very well in the context.

Other methods are administered immediately after the reading of the word or sentence that triggers the inference and can also be considered as online methods. In these tasks, a word is presented as a probe during reading, at or shortly after the moment that the inference is supposedly made. The probe is related to the information that is presumably inferred. An example is the probe word 'break' after the sentence 'No longer able to control his anger, he threw the delicate porcelain vase against the wall.' The task may be to decide whether or not the string of letters of the probe forms a word (*lexical decision*), to decide whether or not the probe word had been presented in the previous sentence(s) (*probe recognition*), or to pronounce the probe word (*naming*). The assumption is that if the inference is encoded into the text representation, the lexical decision time for the probe and its naming time are shorter and the time needed to indicate that 'break' did not occur in the text is longer than in a control condition (see Section The Process Character of an Inference) in which the inference is not made. The status of the different lexical techniques is not quite undisputed. In some investigations, it is claimed that these lexical tasks are sensitive to transient activations (i.e., inferences that are activated only for a short time) rather than to inferences. In particular, naming is sometimes considered a task that does not

reflect the nature of the text representation and, accordingly, does not reflect inferences that are encoded in the text representation. In other studies, it is suggested that lexical decision and probe recognition, in contrast to naming, register the fit of the probe with the preceding context rather than the encoding of inferences.

Some Established Findings in Inference Research

The issue that has stimulated much psycholinguistic research on inferences since the early 1970s is the following question: Which inferences are made and which ones are not made? This issue was investigated by looking for classifications of inferences, on the assumption that some kinds of inferences are made during reading and some other kinds are not.

A common distinction is between *necessary* and *elaborative inferences*. Inferences are called necessary if without them the text representation is not coherent. Inferences that are not necessary for coherence are called elaborative, optional, or embellishing. Two aspects of coherence can be distinguished. *Referential coherence* is achieved by the fact that a sentence in a text deals with entities that are expressed earlier in the text. Linguistic devices to indicate referential coherence are referential (or anaphoric) expressions, such as pronouns and definite noun phrases that refer to an antecedent in the text (*see Anaphora Resolution*). *Relational coherence* is achieved by the fact that the content of a sentence has a conceptual relation with other sentences, such as a causal, contrastive, or concessive relation. Relational coherence can be expressed by conjunctions. If, in text (a) above, the inference about the coreferentiality between 'it' and 'party' or the inference about causality between 'voting' and 'shift' is not made, there is no coherence between the sentences. The pronoun 'it' is considered as an instruction to find a particular referent; the conjunction 'because' is considered as an instruction to find a causal relation in the context. If such instructions are not executed, the representation is not coherent. In this way, inferences triggered by referential expressions and conjunctions contribute to the coherence of the representation.

The notion of *necessary inference* is sometimes defined in a different way and contrasted with *pragmatic inference*. An inference is necessary if it follows logically from the propositions in a text. From 'John is taller than Pete and Pete is taller than Bill,' it follows that John is taller than Bill. The causal inference between the propositions with 'voting' and 'shift' in text (a) is not only necessary for the coherence, but also necessarily follows from the text. It can be deduced from the conjunction 'because' and is called a conventional implicature.

Inferences that do not follow logically from the text are called pragmatic inferences. An example is the inference that the vase broke from the sentence 'John slipped on the wet floor and dropped the vase.'

Another distinction that is made frequently is that between *backward* and *forward inferences*. Backward inferences relate the current part of the text to an earlier part. Forward inferences anticipate information that might be expressed in the subsequent text. Backward inferences contribute in general to the coherence of the text and are in that sense necessary inferences, while forward inferences are not. Backward inferences are much more restricted than forward inferences; the reader can anticipate a great number of things.

Inferences can also be classified with respect to their content. This yields a list of inferences that can be extended *ad libitum*. It includes inferences about instruments, causes, consequences, goals, time, place, and protagonists.

Established findings in inference research (Singer, 1994, 2007; Van den Broek, 1994; Garrod and Sanford, 1994) are that necessary inferences – that is, inferences that achieve coherence between sentences – are made during reading. These inferences are in general backward inferences. Elaborative, embellishing inferences are in general not made during reading, except when they are highly constrained (O'Brien et al., 1988; Garrod et al., 1990); they are not required for comprehension. They are in general forward inferences. There are many possible forward inferences that can be made, and since they do not achieve coherence, there is no motivation for making them. Inferences that can necessarily be derived from the text are far from always made; in general, it depends on whether they contribute to the coherence or not.

Toward a More Differentiated View on Inferences

The research discussed in the previous section can be characterized by two views on inferences. First, inferences were considered as dichotomous entities: They are made or not made. Second, inferences were described with respect to their function in the text: Whether they are made or not depends on whether they contribute to the coherence of the text. Since the late 1980s and early 1990s, there has been a gradual change in inference research. These changes will be discussed in two subsections. First, the dichotomous view on inferences is being replaced by a process view on inferences: What happens if an inference is made? Second, attention has shifted from the function of inferences in the text to the availability of information and knowledge as determinants of inferences, and in this way to the role of the reader.

The Process Character of an Inference

The shift in attention toward a process view on inferences was stimulated by research on forward inferences. Presumably, the reason was that backward inferences were well established and that the idea that forward inferences are not made was not quite convincing. Several studies (McKoon and Ratcliff, 1986) demonstrated that forward inferences about very predictable events can be made, but are only partially or minimally encoded. After presentation of several sentences, among them, 'The director and the cameraman were ready to start shooting when suddenly the actress fell from the 14th floor,' the word 'dead' was presented as a probe for recognition. If an inference about dying were encoded in memory, it would interfere with the correct answer 'no.' When the target word was preceded by another word from the same sentence, an interfering effect did indeed occur, but when the target word was preceded by a neutral word (e.g., 'ready'), no interference occurred. This result was interpreted as evidence for a minimal encoding of the inference, such as 'something bad happened,' because a prime word from the same sentence

was necessary to strengthen the match between the probe word and the memory representation.

Other research demonstrated that inferences are built up and decay over time. First, forward inferences are not generated until some time after the presentation of the words the inference is based on. Experiments in which the time between the context that primes the inference and the presentation of a target word to detect the inference is varied indicate that a delay of about 1 s is necessary for the inference to be built up (Calvo et al., 1999). This observation may explain why no evidence for forward inferences was obtained in most of the earlier studies. If, in a reading experiment, subsequent information that does not support the inference is presented very quickly, the inference is not built up. However, if readers can read at their own pace or if subsequent information supports the inference, forward inferences can be built up. Second, an inference can be made for a brief period of time. It need not remain active. Keefe and McDaniel (1993) found that the naming latency for the word 'break' was shorter after 'One day, no longer able to control his anger, he threw a delicate porcelain vase against the wall' than after the control sentence 'One day, unable to control his impulses, he went out and purchased a delicate porcelain vase' when the probe word was presented with a short delay after the target sentence, but there was no facilitation when there was an intervening sentence between target sentence and probe. They argued that predictive inferences are drawn but are quickly deactivated. The likelihood that predictive inferences are encoded seems to increase if the information to be inferred is foregrounded and salient, and if the inference concerns a causal relation (Klin et al., 1999). In addition, the specificity of predictive inferences increases as the contextual support increases (Lassonde and O'Brien, 2009). With respect to the predictive inferences found by Van Berkum et al. (2005), one might hypothesize that since listening is slower than reading, more inferences are made during listening than during reading.

Availability of Information and Reader's Knowledge as Determinants of Inferences

Whether inferences are made or not depends not only on the function of the inference in the *text*, but also on whether the *reader* has information and knowledge available. This has been acknowledged in two well-known theories of inferences: the *minimalist theory* and the *constructionist theories*. In the minimalist theory, language processing and inferencing are described largely in terms of bottom-up processing of the information in the text. In the constructionist theories, the reader's search for meaning and the reader's knowledge play an important role, so that understanding entails a considerable amount of top-down processing.

In their minimalist theory on inferences, McKoon and Ratcliff (1992) argue that only two kinds of inference are drawn spontaneously during reading: inferences that serve to achieve *local coherence* and inferences that are based on *easily available information*. Local coherence refers to coherence between propositions that are not farther apart in the text than one or two sentences. Information is easily available if it is expressed in the current two or three sentences or if it is well-known general knowledge, such as instances of categories (collie – dog). In this way, the availability of information in the text is defined in terms of the linear structure of the text.

In constructionist theories, the knowledge of the reader plays a much more important role than in minimalist theory. According to constructionist theories, readers construct a *mental model* or a *situational model*. Such a model contains not only information in the propositions of the text but also information that is constructed by the reader, including elaborative inferences and global inferences (Johnson-Laird, 1983; Garnham and Oakhill, 1996; Zwaan and Radvansky, 1998). Many participants (Bransford and Johnson, 1973) who heard a sentence such as 'Three turtles rested on a floating log and a fish swam beneath them' recognized, incorrectly, the sentence 'Three turtles rested on a floating log and a fish swam beneath it,' but participants who had heard the sentence 'Three turtles rested beside a floating log and a fish swam beneath them' did not often recognize incorrectly the sentence with 'it.' Apparently, comprehension requires the activation of knowledge about spatial relations.

The central idea in the constructionist theory of Graesser et al. (1994) is that reading is considered as a search for meaning. The readers' goal and knowledge guide the reading process. Readers try to construct a meaning representation that is coherent both at the local and at the global level. They try to explain the actions, events, and states mentioned in the text; that is, readers try to answer 'why' questions with respect to the text. According to this theory, inferences concerning causal antecedents and superordinate goals are made because they explain actions in the text, whereas inferences about causal consequences and subordinate goals and states are not made because they do not contribute to the explanation. In evaluating minimalism and constructionism, one should keep in mind that even local inferences are not made if they do not address information that is part of the reader's knowledge (Noordman and Vonk, 1992).

Inferencing as One Component of Discourse Comprehension

In the studies reviewed up till now, inferences were considered as a rather isolated phenomenon. The main question was whether inferences are made or not, and the answer was considered to depend on the kind of inferences: bridging, forward, elaborating, instrumental, etc. In an earlier overview on inferences in discourse, Vonk and Noordman (2001) concluded that progress in inference research will be made only if inferences are no longer considered as a rather isolated phenomenon, but a component in discourse comprehension. Inferences should then be considered in relation to models of discourse comprehension and attention should be paid to the crucial role of world knowledge. This is indeed what happened to certain extent in the last decade. Once one considers inferences as a component in the comprehension process, it becomes clear that minimalism and constructionism are compatible. Discourse comprehension is, like all cognitive activities, *memory-based*: The processes take place in working memory and make

contact with information in long-term memory, which comprises information both from the previous discourse and from world knowledge. A crucial concept in the minimalist theory of inferences is 'easily available information.' But since easily available information includes previous discourse, world knowledge, knowledge about the topic of the text, about the goals of the protagonist (who can be referred to by a pronoun, testifying for its accessibility), and about the text structure, it does not follow from this notion that only a few inferences are made. For example, given that immediately available information includes world knowledge related to a specific word, it is not clear why minimalism would predict that the inference 'spoon' is not made when reading a sentence 'John stirred the coffee.' Interesting in this respect are data from *embodied cognition* showing that when reading sentences that describe a particular action similar activity is observed in the brain as when performing that action. Why would minimalism claim that the instrumental inference for 'stirring coffee' is not made, while inferences about the waiter, the menu, and the bill are made in a restaurant script story? Whether the inferences are made or not is, of course, an empirical question, but the prediction does not follow from the minimalist position. Other examples are studies by Albrecht and O'Brien (1993) and Myers et al. (1994). Given that the protagonist and his or her attributes are active during the whole story, it is not surprising – and it does not refute minimalism – that there is an increase in reading time for 'she ate a cheeseburger' if the protagonist was introduced earlier in the discourse as a vegetarian. To further strengthen the conciliation between minimalism and constructionism: it is not clear why some findings are considered to support constructionism and to argue against minimalism. Consider the beginning of a story 'Valerie left early for the birthday party' vs 'Valerie left the birthday party early' followed by a number of sentences including 'She spent an hour shopping at the mall' (Graesser et al., 1994). World knowledge about 'left for' and 'left' makes clear that the birthday party is an upcoming episode or a terminated episode, respectively. Consistent with the view that readers keep track of the temporal development in a story (Bestgen and Vonk, 1995; Zwaan and Radvansky, 1998), the birthday script is relevant and available only in the 'left for' version, which was evidenced by a test sentence. This explanation of the results is not in conflict with minimalism at all. It shows that the constructionist theory is not incompatible with memory-based processing. And this claim can be made stronger. In the constructionist view, reading is considered as a search after meaning and explanations. The goal dictates what kind of knowledge the reader is after and activates specific knowledge and makes it available. It determines what information is in focus. The reader's goal can affect the processes in working memory and interact with long-term memory. The search after meaning of this explanation-based approach can be accommodated in the memory-based approach. The position of Gerrig and O'Brien (2005) is in agreement with our proposal. "There is no need to define categories of inferences that readers will typically encode and those that readers are less likely to encode. Inferences are encoded to the extent that information in active memory makes contact with relevant or necessary information from inactive portions of the discourse model and general world knowledge" (p. 236). They also suggest that the search-after-meaning of the explanation-based view can be accommodated in the memory-based view: "within the memory-based view, the concept of search-after-meaning is not actually a search of memory; instead it is an attentional, resources-consuming … process" (p. 237). What is less clear, however, is how this memory-based view on inferences can be invalidated, since all cognitive processes are memory processes.

Less progress has been made in relating inferences to models of discourse comprehension and to the crucial role of world knowledge. Most of the models of discourse incorporate world knowledge, but not in a computational way (Frank et al., 2008). In a memory-based model, world knowledge of the previous discourse and world knowledge are supposed to resonate. The Resonance model (Myers and O'Brien, 1998) is the example *par excellence* of memory-based processing. But the model does not incorporate world knowledge and does not claim to do so. The Construction–Integration model (Kintsch, 1988, 1998) has been a fruitful framework for the interpretation of experimental data and the generation of further research. The model accommodates inferences, but the implementation of world knowledge is rather ad hoc and restricted to text-relevant knowledge. This also is true for the Landscape model (Van den Broek et al., 1999) and the model by Langston and Trabasso (1999). On the other hand, the Distributed Situation Space model (Frank et al., 2003) encodes world knowledge based on events in a microworld. Therefore, world knowledge in this model is not introduced ad hoc and the model predicts inferences in the microworld in a computational way. The question, of course, is whether the model can be scaled up to more realistic amounts of knowledge. This problem applies to the other models as well.

Inferences and the Brain

The advent of techniques that measure brain activity is an important step in the study of discourse processing. One such technique is measuring ERPs. The ERP signal is a pattern of positive and negative peaks in the EEG. Hagoort et al. (2004) showed that when a word in a sentence violates world knowledge, this is reflected by the N400 in the brain wave. In the study by Van Berkum et al. (2005) cited earlier, differential ERP effects were found when an upcoming word matched vs mismatched a prediction.

Kuperberg et al. (2011) investigated inference processes in understanding sentences that have different degrees of causal relatedness. Lexicosemantic co-occurrence was held constant across the conditions. Critical words in the causally unrelated sentences evoked a larger N400 than critical words in both the highly related and intermediately related sentences. This was the case both when the critical words occurred before and at the end of the sentences. These data demonstrate that readers are immediately sensitive to coherence breaks. Causal coherence at the situation level immediately influences inferencing during semantic processing of incoming words. At many of the electrode sites, the amplitude of the N400 to critical words in intermediately related sentences fell in

between the amplitude of the N400 in the unrelated and highly related conditions. Interestingly, the linear relation between causal relatedness and the amplitude of the N400 differs from curvilinear relations obtained in a functional magnetic resonance imaging (fMRI) study by Mason and Just (2004) in brain activation and Myers et al. (1987) in recall. One possible interpretation, according to Kuperberg et al. (2011), might be the difference in temporal resolution between ERP and fMRI.

An interesting topic is the mapping of cognitive processes to regions in the brain. In general, it is assumed that processes in language understanding, such as word coding and syntactic analysis, activate the left hemisphere. There is increasing evidence, however, that the right hemisphere is also involved in language processes and in particular in inferencing. Beeman et al. (2000) showed in a visual hemifield study that predictive inferences activate the right hemisphere. Reading the sentence 'The space shuttle sat on the ground, waiting for the signal' might lead the reader to the predictive inference that the shuttle is taking off. This would lead to facilitation of the target word 'launch.' When this word was presented to the left visual hemifield (corresponding to the right hemisphere), the naming latency for this word was shorter than for an unrelated control word. But presenting the target word to the right visual hemifield, there was no difference between target and control word. According to the coarse semantic coding theory (Beeman et al., 1994), the right hemisphere is involved in the representation of wider connections of word meanings and less central meanings of words. This representation includes distant associations between words. Since these associations play a role in the processing of inconsistencies and in inferencing, in particular when there is no direct overlap between adjacent pieces of text, it is supposed that the right hemisphere is involved in inferences. Interestingly, for causal inferences that require bridging inferences, Beeman et al. (2000) found that the left hemisphere is responsible for the inferences after the shuttle sat on the ground. When the shuttle disappeared into space,' the word 'launch' is processed faster than a control word when it is presented to the right visual hemifield, but not when it is presented to the left visual hemifield.

Virtue et al. (2006b) investigated in a visual hemifield study bridging and predictive inferences that were either strongly or weakly constrained by the context. A target word related to the inference was presented in one of the visual fields for lexical decision. For strongly constrained bridging and predictive inferences, facilitation (shorter lexical decision time for target words in the inference condition than in a control condition) was obtained in both hemispheres. For weakly constrained bridging and predictive inferences, there was greater facilitation in the right hemisphere than in the left hemisphere, but for predictive inferences there was some facilitation in the left hemisphere as well.

Mason and Just (2004) used fMRI to study the neuroanatomical basis for inference processes in reading. This is a technique that measures activity in specific brain areas by detecting changes in blood flow. It is less sensitive to temporal resolution than ERPs. Mason and Just (2004) used sentence pairs of Keenan et al. (1984) and Myers et al. (1987) but only in three versions: the distantly related, moderately related, and highly related versions. In the right hemisphere, there was greater activity for the moderately related sentence pairs than for the highly related and the distantly related pairs. In the left hemisphere, there was no difference between the three versions and the activation for the highly related and distantly related versions was greater than in the right hemisphere. These results were interpreted to mean that the left hemisphere is concerned with processing the sentences, while the right hemisphere is responsible for establishing the coherence and presumably for making the inferences.

Kuperberg et al. (2006) used in an fMRI study similar items as Mason and Just, but they had items of three sentences long instead of two. In addition, they changed the experiment in some methodological ways (e.g., increasing the number of items and testing only on the last sentence instead of on two sentences). In the intermediately related condition, processing required more time (participants had to judge the coherence of the sentences). For the intermediately related sentences (compared to the other two conditions) they found increased activity in both hemispheres (left lateral temporal/inferior parietal/prefrontal cortices, the right inferior prefrontal gyrus, and bilateral superior medial prefrontal cortices). Apparently, understanding information that is not explicitly stated in the discourse is accompanied by the activation of a large cortical network in both hemispheres.

Interesting as it is to identify the areas that are activated, more interesting is what localization data can contribute to a theory of inferencing. A distinction is frequently made between the generation of inferences and the integration of the inferred information in long-term memory (Mason and Just, 2004). Is there experimental evidence from brain studies for this distinction? Mason and Just suggest that these processes correspond to two large-scale cortical networks: a reasoning system in dorsolateral prefrontal cortex for generating inferences and specific parts of the right-hemisphere for integrating the possible inferences that have been generated. Kuperberg et al. (2006) are even more specific in relating cognitive processes to brain regions. They state that inferencing is accompanied by the activation of many regions in the brain. These regions act in consort, but building upon previous studies, Kuperberg et al. (2006) suggest that parts of this network may have distinct roles in causal inferencing. The temporal regions are involved in the activation of stored semantic information; the inferior prefrontal regions may mediate the retrieval and selection of semantic information; the posterior inferior prefrontal regions and posterior dorsolateral prefrontal regions may play a role in the maintenance of semantic information in working memory as it is integrated in long-term memory; activity in the superior medial prefrontal regions may reflect directed search for meaning and the examination of temporal relationships between events to generate inferences. Furthermore, they suggest that the activation in the right temporal and inferior prefrontal regions play a role in detecting incoherence in the discourse. These suggestions are speculative, but they indicate how imaging studies may advance theories about inference processes.

The neural correlates of discourse comprehension processes have only recently been investigated. This is a promising line of research. But we need more relevant empirical data, as well as

replications of earlier data, because consistency and convergence are important issues. This is particularly relevant since studies differ in many respects: the method, the task for the reader/listener, the materials that are used in the experiment. To illustrate this point: A topic of quite a few studies is coherence, including lack of coherence, coherence breaks, and inferences to establish coherence. Different studies find quite different locations of activations; they differ even with respect to whether there is activation in the left or right hemisphere. But that is not so surprising if one considers that under the guise of (in)coherence processing quite different operationalizations of coherence and quite different tasks are hid, that give rise to quite different cognitive processes.

In Mason and Just (2004), for example, coherence is operationalized as different degrees of causal relatedness between two consecutive sentences. Moderate coherence resulted in greater activation in specific areas of the right hemisphere (middle and superior temporal gyrus, inferior temporal gyrus, inferior frontal gyrus, and inferior parietal area). The task was to read the texts and to answer probe questions for filler texts.

Robertson et al. (2000), on the other hand, operationalized coherence as presence of a definite or an indefinite article. The indefinite article produced more activity than the definite article in the right hemisphere, in particular in the prefrontal region, but not in the left hemisphere. The task was to read sets of sentences.

Ferstl and Von Cramon (2001), for their part, operationalized coherence in terms of (un)related sentences. The task was to judge whether the two sentences have something to do with each other. They found that the coherence manipulation resulted in activation in the left frontal gyrus. They did not find any additional activation in the right hemisphere (see also Ferstl et al., 2008).

In Virtue et al. (2006a), coherence break is operationalized as a place in the text where an inference is required: A sentence had to be connected to the preceding sentence; in one condition an unspecific verb ('got to work') occurred in the preceding sentence, in the other condition a specific verb (started ironing). The task was to listen to the text and answer a question, not related to the coherence break. They found increased activation for the unspecific over the specific event in the right and left superior temporal gyrus and inferior frontal gyrus.

Kim et al. (2012) used strong coherence and weak coherence stories of three sentences in a positron emission tomography study. This technique, as fMRI, measures brain activity on the basis of local blood flow. The last sentence in the weak condition required an inference, but the last sentence in the strong condition did not. The task was to read the sentences and to judge the plausibility of the last sentence. In the strong coherence condition, the dorsomedial prefrontal cortex was activated relative to a control condition. This was interpreted as reflecting coherence processing (but in other studies activation at this location is obtained in low coherence conditions). In the weak coherence condition, the left middle temporal gyrus was activated. This was interpreted as reflecting a bridging inference.

The operationalization of coherence in these studies might very well lead to quite different cognitive processes. The use of indefinite articles in the Robertson et al. (2000) study violates (Gricean) maxims of conversation, in particular of referring. Moreover, some of the references are ambiguous. The reader has to deal with these problems. It is not very likely that establishing reference in this task is done on the basis of world knowledge activation. Quite different are the Mason and Just (2004) and the Virtue et al. (2006a) studies. In these studies, coherence can be achieved by activating world knowledge, making bridging inferences, and integrating them into the discourse representation. And indeed, Mason and Just argue that these processes take place and that they can be distinguished as located in different brain networks. In the Ferstl and Von Cramon (2001) study, participants perform a metalinguistic decision task. To reach a criterion that two sentences have nothing to do with each other, in comparison with the sets of sentences that are connected, it is not necessary that participants engage in world knowledge activation and inference generation, as in the Mason and Just study. Finally, in the Kim et al. (2012) study, the metalinguistic task of judging the plausibility of a sentence introduces an additional task and probably introduces additional processes.

The upshot is that different materials and different tasks may lead to different processes in coping with 'incoherence.' These examples illustrate the need for carefully conducted empirical studies on inference processes. Only then will brain studies increase the understanding of the different subprocesses in inferencing and how they are related to each other. Brain research will then not only reveal which areas in the brain are activated in specific tasks, but may also yield data that contribute to a theory of inferences and discourse comprehension.

See also: Psychology of Inferences; Sentence Comprehension, Psychology of.

Bibliography

Albrecht, J.E., O'Brien, E.J., 1993. Updating a mental model: maintaining both local and global coherence. Journal of Experimental Psychology: Learning, Memory, and Cognition 19, 1061–1070.

Beeman, M.J., Bowden, E.M., Gernsbacher, M.A., 2000. Right and left hemisphere cooperation for drawing predictive and coherence inferences during normal story comprehension. Brain & Language 71, 310–336.

Beeman, M.J., Friedman, R.B., Grafman, J., Perez, E., Diamond, S., Lindsay, M.B., 1994. Summation priming and coarse coding in the right hemisphere. Journal of Cognitive Neuroscience 6, 26–45.

Bestgen, Y., Vonk, W., 1995. The role of temporal segmentation markers in discourse processing. Discourse Processes 19, 385–406.

Bransford, J.D., Johnson, M.K., 1973. Considerations of some problems of comprehension. In: Chase, W.G. (Ed.), Visual Information Processing. Academic Press, New York, pp. 383–438.

Calvo, M.G., Castillo, M.D., Estevez, A., 1999. On-line predictive inferences in reading: processing time during versus after the priming context. Memory and Cognition 27, 834–843.

Ferstl, E.C., Neumann, J., Bogler, C., Von Cramon, D.Y., 2008. The extended language network: a meta-analysis of neuroimaging studies on text comprehension. Human Brain Mapping 29, 581–593.

Ferstl, E.C., Von Cramon, D.Y., 2001. The role of coherence and cohesion in text comprehension: an event-related fMRI study. Cognitive Brain Research 11, 325–340.

Frank, S.L., Koppen, M., Noordman, L.G.M., Vonk, W., 2003. Modeling knowledge-based inferences in story comprehension. Cognitive Science 27, 875–910.

Frank, S.L., Koppen, M., Noordman, L.G.M., Vonk, W., 2008. World knowledge in computational models of discourse comprehension. Discourse Processes 45, 429–463.

Garnham, A., Oakhill, J., 1996. The mental models theory of language comprehension. In: Britton, B.K., Graesser, A.C. (Eds.), Models of Understanding Text. Erlbaum, Mahwah, NJ, pp. 313–339.

Garrod, S.C., O'Brien, E.J., Morris, R.K., Rayner, K., 1990. Elaborative inferences as an active or passive process. Journal of Experimental Psychology: Learning, Memory, and Cognition 16, 250–257.

Garrod, S.C., Sanford, A.J., 1994. Resolving sentences in a discourse context: how discourse representation affects language understanding. In: Gernsbacher, M.A. (Ed.), Handbook of Psycholinguistics. Academic Press, San Diego, CA, pp. 675–698.

Gerrig, R.J., O'Brien, E.J., 2005. The scope of memory-based processing. Discourse Processes 39, 225–242.

Graesser, A.C., Singer, M., Trabasso, T., 1994. Constructing inferences during narrative text comprehension. Psychological Review 101, 371–395.

Hagoort, P., Hald, L., Bastiaansen, M., Petersson, K.M., 2004. Integration of word meaning and world knowledge in language comprehension. Science 304, 438–441.

Haviland, S.E., Clark, H.H., 1974. What's new? Acquiring new information as a process in comprehension. Journal of Verbal Learning and Verbal Behavior 13, 512–521.

Johnson-Laird, P.N., 1983. Mental Models. Cambridge University Press, Cambridge, UK.

Just, M.A., Carpenter, P.A., 1978. Inference processes during reading: Reflections from eye fixations. In: Senders, J.W., Fisher, D.F., Monty, R.A. (Eds.), Eye Movements and the Higher Psychological Functions. Erlbaum, Hillsdale, NJ, pp. 157–174.

Keefe, D.E., McDaniel, M.A., 1993. The time course and durability of predictive inferences. Journal of Memory and Language 32, 446–463.

Keenan, J.M., Baillet, S.D., Brown, P., 1984. The effects of causal cohesion on comprehension and memory. Journal of Verbal Learning and Verbal Behavior 23, 115–126.

Kim, S., Yoon, M., Kim, W., Lee, S., Kang, E., 2012. Neural correlates of bridging inferences and coherence processing. Journal of Psycholinguistic Research 41, 311–321.

Kintsch, W., 1988. The role of knowledge in discourse comprehension: a construction-integration model. Psychological Review 95, 163–182.

Kintsch, W., 1998. Comprehension: A Paradigm for Cognition. Cambridge University Press, Cambridge, UK.

Klin, C.M., Guzmán, A.E., Levine, W.H., 1999. Prevalence and persistence of predictive inferences. Journal of Memory and Language 40, 593–604.

Kuperberg, G.R., Lakshmanan, B.M., Caplan, D.N., Holcomb, P.J., 2006. Making sense of discourse: an fMRI study of causal inferencing across sentences. NeuroImage 33, 343–361.

Kuperberg, G.R., Paczynski, M., Ditman, T., 2011. Establishing causal coherence across sentences: an ERP study. Journal of Cognitive Neuroscience 23, 1230–1246.

Langston, M.C., Trabasso, T., 1999. Modeling causal integration and availability of information during comprehension of narrative texts. In: Van Oostendorp, H., Goldman, S.R. (Eds.), The Construction of Mental Representations during Reading. Erlbaum, Mahwah, NJ, pp. 29–69.

Lassonde, K.A., O'Brien, E.J., 2009. Contexual specificity in the activation of predictive inferences. Discourse Processes 46, 426–438.

Mason, R.A., Just, M.A., 2004. How the brain processes causal inferences in text. Psychological Science 15, 1–7.

McKoon, G., Ratcliff, R., 1986. Inferences about predictable events. Journal of Experimental Psychology: Learning, Memory, and Cognition 12, 82–91.

McKoon, G., Ratcliff, R., 1992. Inference during reading. Psychological Review 99, 440–466.

Myers, J.L., O'Brien, E.J., 1998. Accessing the discourse representation during reading. Discourse Processes 26, 131–157.

Myers, J.L., O'Brien, E.J., Albrecht, J.E., Mason, R.A., 1994. Maintaining global coherence during reading. Journal of Experimental Psychology: Learning, Memory, and Cognition 20, 876–886.

Myers, J.L., Shinjo, M., Duffy, S.A., 1987. Degree of causal relatedness and memory. Journal of Memory and Language 26, 453–465.

Noordman, L.G.M., Vonk, W., 1992. Reader's knowledge and the control of inferences in reading. Language and Cognitive Processes 7, 373–391.

O'Brien, E.J., Shank, D.M., Myers, J.L., Rayner, K., 1988. Elaborative inferences during reading: do they occur on-line? Journal of Experimental Psychology: Learning, Memory, and Cognition 14, 410–420.

Robertson, D.A., Gernsbacher, M.A., Guidotti, S.J., Robertson, R.W.W., Irwin, W., Mock, B.J., Campana, M.E., 2000. Functional neuroanatomy of the cognitive process of mapping during discourse comprehension. Psychological Science 11, 255–260.

Singer, M., 1994. Discourse inference processes. In: Gernsbacher, M.A. (Ed.), Handbook of Psycholinguistics. Academic Press, San Diego, CA.

Singer, M., 2007. Inference processes in discourse comprehension. In: Gaskell, M.G. (Ed.), The Oxford Handbook of Psycholinguistics. Oxford University Press, Oxford, pp. 343–359.

Van Berkum, J.J.A., Brown, C.M., Zwitserlood, P., Kooijman, V., Hagoort, P., 2005. Anticipating upcoming words in discourse: evidence from ERPs and reading times. Journal of Experimental Psychology: Learning, Memory, and Cognition 31, 443–467.

Van den Broek, P.W., 1994. Comprehension and memory of narrative texts: inferences and coherence. In: Gernsbacher, M.A. (Ed.), Handbook of Psycholinguistics. Academic Press, San Diego, CA, pp. 539–588.

Van den Broek, P., Young, M., Tzeng, Y., Linderholm, T., 1999. The landscape model of reading: inferences and the online construction of a memory representation. In: Van Oostendorp, H., Goldman, S.R. (Eds.), The Construction of Mental Representations during Reading. Erlbaum, Mahwah, NJ, pp. 71–98.

Virtue, S., Haberman, J., Clancy, Z., Parrish, T., Beeman, M.J., 2006a. Neural activity of inferences during story comprehension. Brain Research 1084, 104–114.

Virtue, S., Van den Broek, P., Linderholm, T., 2006b. Hemispheric processing of inferences: the effects of textual constraint and working memory capacity. Memory and Cognition 34, 1341–1354.

Vonk, W., Noordman, L.G.M., 2001. Inferences in discourse, psychology of. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Elsevier, Amsterdam, pp. 7427–7432.

Zwaan, R.A., Radvansky, G.A., 1998. Situation models in language comprehension and memory. Psychological Bulletin 123, 162–185.

Infertility: Demographic Aspects

Marida Hollos, Brown University, Providence, RI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by U. Larsen, J. Menken, volume 11, pp. 7435–7439, © 2001, Elsevier Ltd.

Abstract

Infertility is a condition that affects individuals and couples in developing and industrialized countries alike. In developing countries, infertility is relatively high; it is predominantly secondary infertility (the inability to conceive or bear a second child), due to STIs. In industrialized countries, it is mostly due to delayed childbearing by which time women's fecundity has decreased significantly. Thus, it is usually primary infertility (the inability to conceive or bear a first child). In developing countries, not having a child is a devastating problem for women, particularly in the high fertility context of sub-Saharan Africa. Assisted reproductive technologies (ARTs) are few in developing countries and only affordable to a small portion of the population. ARTs exist in practically all the industrialized countries and are available to most women and are often subsidized. Voluntary childlessness, however, is growing in Europe.

Definitions

The clinical definition of infertility is the absence of conception after 12 months of regular unprotected intercourse. The World Health Organization (WHO) definition specifies a period of 24 months. Demographers define infertility as the inability of a sexually active woman who is not using contraception to have a live birth. Demographers have shifted the endpoint from conceptions to live births because it is difficult to collect complete data about conceptions in population-based studies.

Primary infertility is defined as the inability to have one child, and secondary infertility occurs when a woman who has borne at least one child is unable to have another. According to the demographic definitions, primary infertility is approximated by childlessness and secondary infertility by the proportion of subsequent infertility measured for women with children. The level of infertility in a population is usually measured by the proportion of reproductive age women who are infertile.

It should be noted that while these definitions are useful in Western settings and also allow for the comparison of cross-national or regional comparisons of fertility and infertility rates or across time, researchers who are interested in local meanings and perceptions of infertility try to utilize local definitions. For example, Leonard (2002) describes Chadian women's perceptions of fertility problems as including difficulty in proving one's fertility shortly after marriage, the loss of reproductive potential (through miscarriage, stillbirth, abortion), and bearing "too few children."

Developing Countries

Until recently, infertility as a public health or demographic concern has largely been neglected in developing countries (Inhorn and van Balen, 2002; Feldman-Savelsberg, 2002; Hollos and Larsen, 2008). It has been obscured by the high fertility rate of many regions, particularly sub-Saharan Africa, that gave rise to a "concern over population growth that is not conducive to the perception of infertility as a real problem" (Frank, 1983: p. 142). There is now an increasing recognition in demography and the social sciences in general that infertility is a devastating problem for women, particularly in the high fertility context of sub-Saharan Africa (e.g., Boerma and Mgalla, 2001; Inhorn and van Balen, 2002). Worldwide, an estimated 580 million people experience infertility at some point of their reproductive lives. Of these, about 372 million people reside in low- and middle-income countries. Africa shares the largest burden of infertility in the world (Okonofua and Obi, 2009).

Childlessness in a pronatalist society with high fertility rates has a great effect on women, as women have responsibility for social and biological reproduction (Gerrits, 1997; Sundby and Jacobus, 2001; van Balen and Gerrits, 2001; Cornwall, 2001; Silva, 2009; Hollos, 2003; Dyer, 2007). This is particularly the case in patrilineal societies, in which the father's family has sole legal rights over children.

To document the increased attention paid to the problem of infertility, the 1994 United Nations International Conference on Population and Development called for reproductive health rights that seek to provide women with the right to have the "capacity to reproduce and the freedom to decide if, when and how often to do so … and provide couples with the best chance of having a healthy infant" (United Nations, 1994: Section 7.2).

Following this attempt, the WHO Department of Reproductive Health and Research proclaimed that infertility is an 'unmet need' in family planning research and program initiatives (WHO, 2004). Since the WHO convened an international meeting on infertility in 2001, they have published several reports and papers regarding advanced reproductive technologies in developing countries, including widely disseminated guidelines on standards, glossaries, ethics, and even a laboratory manual on "standard reference values for human semen" (WHO, 1999).

The proportion of infertility in currently married women ages 25–49, in first union, married more than 5 years that are childless ranges from 1.7 to 2.5% in Asia, from 0.9 to 2.8% in Latin America, from 2.2 to 2.8% in North Africa, and from 0.9 to 2.3% in sub-Saharan Africa (Larsen, 2003). Secondary infertility is more variable but highest in sub-Saharan Africa,

due largely to the presence of a threatening, high prevalence 'infertility belt' stretching across Central Africa from Southwestern Sudan to Gabon. In some Central African populations, infertility problems occur in 30–50% of all couples attempting to conceive. As a result, the WHO sought to obtain baseline statistical information in country- and region-wide infertility prevalence patterns throughout the continent (WHO, 1987). Based on this data, Larsen (2000) estimated infertility for 28 countries that conducted world fertility surveys or demographic and health surveys between 1978 and 1997 in Africa. The most striking aspect of infertility in Africa is its wide variation, which can occur within as well as between countries. For example, according to the 1990 Nigeria DHS, approximately 4% of women aged 30 years and older have never borne a child. A recent population-based survey of women of reproductive age in Ile-Ife, Nigeria, however, revealed that up to 20% of women have secondary infertility (Snow et al., 1995). Other community-based data suggest that up to 30% of couples in some parts of Nigeria may have proven difficulties in achieving a desired conception after 2 years of marriage without the use of contraceptives (Adetoro and Ebomoyi, 1991).

There is also evidence (Larsen, 2003) that there is an increase with age in the African pattern of infertility, which is in part due to biological aging, but most of it must be attributed to disease. High rates of secondary infertility are attributed to very early initiation of sexual intercourse and sexually transmitted diseases. The primary infertility-causing diseases appear to be gonorrhea and chlamydia (Okonofua and Obi, 2009), which can, if untreated, lead to occlusion of the fallopian tubes. Evidence is accumulating that infertility today may be related to HIV infection.

The rates of infertility are variable even within the same region, and local levels and sociodemographic patterns of infertility have not been widely documented .There are a number of important studies, however, that focus on infertility in developing countries, most of them in Africa. Inhorn (1994) looks at the relentless quest by poor Egyptian urban women for remedies for their infertility. In a later work, Inhorn (1996) looks at gender relations and the importance of fertility/infertility in the context of patriarchy. Some other studies focus on different aspects of childlessness, including local perceptions of the causes of infertility, its social consequences, individual treatment seeking behaviors, and community-based mechanisms for dealing with barrenness.

Local perceptions of the causes of infertility are documented in a number of works. These include witchcraft and spirit possession (Kielmann, 1998; Upton, 1999; Ebin, 1982; Okonofua, 2003). Other locally imputed causes of infertility include sexual promiscuity by women (Upton, 1999; Kielmann, 1998), nonobservance of prescribed behavior (Ebin, 1982), blood incompatibility (Roth, 2001), abortion (Okonofua et al., 1997), and 'mchango' (Roth, 2001; Gijsels et al., 2001).

There is ample documentation on the social consequences of infertility for women, who bear the brunt of it, especially in Africa. As Inhorn (1994, 1996) has shown for Egypt, regardless of the medical causes of infertility, it is women who are ostracized, suffer personal grief and frustration, social stigma, and often serious economic deprivation. In the Cameroon, infertility is a cause for divorce (Feldman-Savelsberg, 1999) as it is in Nigeria (Hollos, 2003). Bodies of infertile women are buried on the outskirts of town among the Yoruba (Ademola, 1982) and among the Aowin in Ghana (Ebin, 1982). Fertility has been shown to be important in the progression of women through life stages and thus in the creation of their identity as mature persons (Hollos, 2003).

More recently, Hollos and her colleagues (Hollos and Larsen, 2008; Hollos et al., 2009) documented differences in the rates of infertility and the perception of the condition in two Nigerian communities. They found a relationship between community treatment of women and the historical prevalence of childlessness.

Industrialized Countries

It is not possible to provide accurate country level estimates of infertility in populations with widespread contraception. In these societies, fertile women historically had the number of children they wanted and subsequently used contraception to prevent further births. Nevertheless, several investigators in Europe provided a wide range of estimates of the prevalence of infertility, ranging from 6 to 20%. A population-based survey of infertility and subfecundity in Denmark, Germany, Italy, Poland, and Spain (Karmaus and Juul, 1999) found that the percentage of sterile women ranged from 7% in Denmark to 1.8% in Poland, with the rest of the countries between these figures. The proportion of sterile couples presented a similar profile, with Denmark at 9.3% and Poland at 2%. According to a report by INED (Sardon, 2000: p. 298) permanent infertility in Western Europe rose rapidly among the post-World War II cohorts and has topped 20% since the 1960 cohorts in England and Wales, 1963 in Austria and Italy, and 1965 in Finland and Ireland. Portugal's rate remained low at under 8% of infertile women.

In the United States, estimates from a large national sample suggest that 10–15% of nonsurgically sterile American women ages 15–44 experience fertility problems (Chandra and Stephen, 1998).

According to Philipov and Dobritz (2002) involuntary infertility in Eastern and Central Europe has risen during the preceding years. Before the start of the political transition, infertility was close to the average level observed in Western Europe, around 6%. During the 1990s, it increased in many countries of the region. A study by Philipov (1998) indicates that in Western Siberia, 16.7% of the couples could not achieve pregnancy for at least 12 months; 3.8% of the couples suffered from primary infertility and 12.9% from secondary infertility. The Bulgarian Association of Sterility and Reproductive Health states that infertility is between 9 and 15.3% in that country. In the Czech Republic, it was found to be 12%, in Romania in 1999 it was 11.0%, and in Ukraine 10.6% (Philipov and Dorbritz, 2002: p. 123).

There is persistent fear that infertility in industrialized countries is increasing even further. There are a number of factors accounting for this trend, most important of which is the increasing delay in childbearing. Whereas previously women had their children at younger ages, currently there is a delay in childbearing, which means that more women are hoping to have a child at older ages when their fecundity may

have declined. Kohler et al. (2002) call this a "postponement transition," i.e., "a shift from an early to a late pattern of first birth" (Sobotka, 1989: p. 280). This increase in the mean age at the birth of the first child increased suddenly and significantly by about 1.5–2 years toward 2000. The rise in age was steepest in the ex-Yugoslav countries, the Czech Republic, Hungary, and the Slovak Republic. Western European countries have experienced a similar rise in the mean age at first birth. Births of the first child decreased significantly. This secular trend toward later childbearing has contributed to the decline in period fertility rates. The total fertility rate (TFR) Europe wide of first order births was around 0.9 and higher before the start of the transition, thus indicating universality of the birth of at least one child. Toward the end of the 1990s, the rate dropped below 0.9. According to Philipov (2003) this trend indicates a break in the universality of having children. Voluntary childlessness came on the rise, as well as a rise in the infertility rates of individuals who did desire children.

The reasons for the postponement and subsequent declining birth rates predominantly reflect social and economic factors that are to a large degree voluntary. These social and economic factors include an expanded role for women in the workforce, higher educational opportunities for women, as well as the availability of modern contraceptives and safe and legal abortions. The result of these social and economic trends is that women in industrialized countries suffer predominantly from primary infertility.

A consequence of the delay in childbearing and the difficulty of conceiving is the growth of assisted reproductive technology clinics in Europe, Australia, and New Zealand, and in the United States. Following is a brief summary of the results generated from European registers by the European Society for Human Reproduction and Embryology (ESHRE) and published in 2006. This is the tenth annual report. Reporting began in 1997.

Data in 2006 was collected from 32 countries, up from 26 in the previous year covering *in vitro* fertilization (IVF), intracytoplasmic sperm injection (ICSI), frozen embryo replacement, egg donation, *in vitro* maturation, pooled data on preimplantation genetic diagnosis, and frozen oocyte replacements. The number of clinics in the various countries ranged from 1 (Latvia, Iceland) to 202 in Italy, 182 in Spain, 122 in Germany, and 102 in France. The number of treatment cycles in total was 458 759, 40 648 more than in 2005. The distribution of clinics according to the number of cycles varied considerably among the countries. For example, in Italy, 46% of the clinics provided fewer than 100 cycles annually, whereas in Belgium and Sweden, 50% of the clinics performed more than 1000 cycles a year. In the Netherlands, 62% of the clinics performed more than 1000 cycles annually. The largest number of assisted reproductive technology (ART) cycles was reported from France (66 000), Germany (55 000), Spain (50 000), the United Kingdom (44 000), and Italy (42 000). None of these reach the number of reported cycles from the United States (138 198). Australia and New Zealand reported 53 543 cycles.

The average number of treatment cycles per million inhabitants ranged from 122 in Latvia to 2268 in Denmark. ART cycles in women of reproductive age (15–49 years) also show large differences from 475 cycles per million women in Latvia to 10 132 in Denmark. The percentage of newborns conceived through ART varied from 0.8% in Montenegro to 4.1% in Denmark.

The age distribution of women treated with IVF varied across countries. In several countries, more than 20% of women were aged 40 years of more (Greece, Ireland, Italy, Macedonia, Montenegro, Serbia, and Switzerland), whereas it was less than 10% in Bulgaria, Czech Republic, Lithuania, Poland, Norway, and Ukraine. As expected, pregnancy rates declined with age. Pregnancy rates in Europe remain lower than in the United States where 35% of cycles from nondonor cycles resulted in pregnancy. However, pregnancy rates in Europe are very similar to what is achieved in Australia and New Zealand where the pregnancy rate per transfer was 30.6% after fresh cycles and 21.5% after frozen embryo transfers.

In summary, the ESHRE report shows a continuing expansion of clinics and treatment cycles. Pregnancy rates from IVF and ICSI showed marginal increases compared with 2005.

Childlessness

Childlessness is an important and growing phenomenon and has increasingly become the focus of research by sociologists and policy makers, primarily in Western Europe. As McAllister and Clarke (2000: p. 189) point out, "as the average age of motherhood has increased in many Western societies, the issue of distinguishing between those who postpone parenthood and those who renounce it altogether has become of key interest."

Historically, childlessness is not new. It was quite common until the early part of the twentieth century. According to Wrigley and Schoefield (1983), in eighteenth-century England, up to one quarter of women remained unmarried and the majority of these had no children. However, in the past, childlessness among married women was primarily due to poverty and consequent malnutrition. Childlessness today (see, for example, Dykstra and Hagestad, 2007), while similar in magnitude, is very different from the past. It occurs among healthy, sexually active women who live in relative prosperity.

The prevalence of lifetime childlessness has been shown by demographers to be far from uniform both within and between countries, and according to Coleman (1993), it is likely to generate considerable differences in fertility levels. While exact figures are difficult to obtain, many sources provide differing estimates and projections of current trends suggesting that in many countries from one-fifth to one-quarter of women will remain childless throughout their lifetimes (Eurostat, 1997; Morgan and Chen, 1992; Coleman, 1998; Rowland, 2007). In addition to Germany, England and Wales, Canada, Finland, and Austria, this includes various Eastern European countries (Coleman, 1998). Researchers agree that primary infertility cannot account for this level of childlessness (Rindfuss et al., 1988; Ruthstein and Shah, 2004; Vaessen, 1984).

Evidence from a number of studies (Veevers, 1983; Campbell, 1985; Gerson, 1985; Morell, 1994; Beets, 1995) suggests that there is fluidity to being childless. Decisions about childbearing may be very different at younger ages than toward the end of a woman's reproductive years (McAllister and Clarke, 2000). While the desire to have children at younger

ages may be stronger, the postponement of first births may result in increasing fertility problems after age 35; and delaying the decision to have children may lead to infecundity. Campbell (1985) and MacInnes (2006) also argue that becoming childless is a process and that many couples may 'drift' into childlessness through postponement. However, an analysis of fertility intentions, using Eurobarometer data, shows that across Europe there is a large variation in the intention to remain childless. This may, of course, depend on cultural values and norms regarding parenthood (Bernardi, 2005). The 1996 Italian Fertility and Family Survey, for example, suggested that "childlessness, rather than being driven by deliberate choice, was an outcome of a series of life circumstances that were beyond women's control" (Bonarini et al., 1999).

Conclusion: Two Worlds of Infertility

As shown in the previous discussion as well as in a review of recent literature on infertility by Greil et al. (2010), there are salient differences in the experience of infertility between developing and industrialized societies. They propose that it may be justifiable to think in terms of two worlds of infertility. Developing and industrialized countries tend to differ in prevailing assumptions about childlessness.

In industrialized societies, voluntary childlessness is viewed as a viable legitimate option and women without children are often assumed to be voluntarily childfree. However, in developing countries, having children is central to women's role in society. The option of voluntary childlessness does not exist. In these societies, it is impossible to hide infertility.

Another difference between infertility in developing and industrialized societies has to do with folk models for making sense of the condition. In industrialized societies, the acceptance of the biomedical model is hegemonic, while in other societies biomedical interpretations of infertility coexist and interact with traditional interpretations.

Finally, ARTs are widely available in industrialized countries and are few and mostly beyond the reach of individuals in developing countries.

See also: Demographic Measurement: General Issues and Measures of Fertility; Family Planning Programs: Feminist Perspectives; Fertility Theory; Fertility and Social Networks; Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective; Sexually Transmitted Infections: Social Network Analysis.

Bibliography

Ademola, A., 1982. Changes in the patterns of marriage and divorce in a Yoruba town. Rural Africana 14, 1–24.

Adetoro, O.O., Ebomoyi, E.W., 1991. The prevalence of infertility in a rural nigerian community. African Journal of Medicine and Medical Sciences 20, 23–27.

Araoye, M.O., 2003. Epidemiology of infertility: social problems of the infertile couple. West African Journal of Medicine 22, 190–196.

van Balen, F., Gerrits, T., 2001. Quality of infertility care in poor resource areas and the introduction of new reproductive technologies. Human Reproduction 16 (2), 215–219.

Beets, G., 1995. Does the increasing age at first birth lead to increases in voluntary childlessness?. In: EAPS, European Population Conference 1995. Evolution or Revolution in European population, vol. 2. Franco Angeli, Milano, pp. 15–30.

Bernardi, F., 2005. Public policies and low fertility. Rationales for public intervention and a diagnosis for the Spanish case. Journal of European Social Policy 15, 123–138.

Bonarini, F., Castiglioni, M., Rosina, A., 1999. Infecondita, sterilita e tempi di attesa del concepimento. In: DeSandre, P., Pinnelli, A., Santini, A. (Eds.), Nuzialita e fecondita in trasformazione: percorsi e fattori di cambiamento. Il Mulino, Bologna, pp. 667–682.

Boddy, J., 1989. Wombs and Alien Sprits: Women, Men and the Zar Cult in North Sudan. University of Wisconsin Press, Madison.

Boerma, J. Ties, Mgalla, Zaida (Eds.), 2001. Women and Infertility in Africa: A Multidisciplinary Perspective. Royal Tropical Institute Press, Amsterdam.

Campbell, E., 1985. The Childless Marriage: An Exploratory Study of Couples Who Do Not Want Children. Tavistock, London.

Coleman, D., 1993. Contrasting age structures of Western Europe and of Eastern Europe and the former Soviet Union: demographic curiosity or labor resource? Population and Development Review 19 (3), 523–555.

Coleman, D., May 1998. Reproduction and Survival in an Unknown World. The Hofstee Lecture presented under the auspices of the Social Science Council of the Royal Netherlands Academy for Arts and Sciences. Amsterdam.

Council of Europe, 1998. Recent Demographic Developments in Europe. Council of Europe Press, Strasbourg.

Chandra, A., Stephen, E.H., 1998. Impaired fecundity in the United States 1982–1995. Family Planning Perspectives 39, 34–41.

Cornwall, Andrea, 2001. Looking for a child: coping with infertility in Ado-Odo, Southwestern Nigeria. In: Tremayne, Soraya (Ed.), Managing Reproductive Life: Cross-Cultural Themes in Sexuality and Fertility. Berghahn Books, New York, pp. 140–156.

Dyer, S.J., 2007. The value of children in African countries: insights from studies of infertility. Journal of Psychosomatic Obstetric Gynecology 28, 69–77.

Dykstra, P.A., Hagestad, K., 2007. Roads less taken: developing a nuanced view of older adults without children. Journal of Family Issues 28 (10), 1275–1310.

Ebin, V., 1982. Interpretations of infertility: the Aowin people of Southwest Ghana. In: MacCormack, Carol P. (Ed.), Ethnography of Fertility and Birth. Academic Press, London, pp. 147–159.

Eurostat, 1997.

ESHRE, 2006. Assisted reproductive technology Europe 2006: results generated from European registers by ESHRE (European Society for Human Reproduction). Human Reproduction 25 (8), 1852–1862.

Feldman-Savelsberg, P., 1999. Plundered Kitchens and Empty Wombs. University of Michigan Press, Ann Arbor.

Feldman-Savelsberg, P., 2002. Is infertility an unrecognized public health and population problem? The view from the Cameroon Grassfields. In: Inhorn, M., van Balen, F. (Eds.), Infertility around the Globe. University of California Press, Berkeley.

Frank, O., 1983. Infertility in Sub-Saharan Africa. Estimates and implications. Population and Development Revew 9, 137–144.

Gerrits, T., 1997. Social and cultural aspects of infertility in Mozambique. Patient Education and Counseling 31, 39–48.

Gerson, K., 1985. Hard Choices: How Women Decide about Work, Career and Motherhood. University of California Press, Berkeley.

Gijsels, M., Mgalla, Z., Wambura, L., 2001. 'No child to send': context and consequences of female infertility in northwest Tanzania. In: Ties Boerma, J., Mgalla, Z. (Eds.), Women and Infertility in Sub-Saharan Africa. Royal Tropical Institute, The Netherlands.

Greil, Arthur L., Slauson-Bevins, K., McQuillan, J., 2010. The experience of infertility: a review of recent literature. Sociology of Health and Illness 32 (1), 140–162.

Hollos, M., 2003. Profiles of infertility in Southern Nigeria: women's voices from Amakiri. African Journal of Reproductive Health 7, 46–56.

Hollos, M., Larsen, U., 1992. Fertility differentials among the Ijo in Southern Nigeria: the role of women's education and place of residence. Social Science & Medicine 35, 1199–1210.

Hollos, M., Larsen, U., 2008. Motherhood in Sub-Saharan Africa: the social consequences of infertility in an urban population in Northern Tanzania. Culture, Health and Sexuality 10 (2), 159–173.

Hollos, M., Whitehouse, B., 2009. (In)fertility and the female life cycle in two Southern Nigerian communities. Ethnology 47 (1), 23–43.

Hollos, M., Larsen, U., Obono, O., Whitehouse, B., 2009. The problem of infertility in high fertility populations: meanings, consequences and coping mechanisms in two Nigerian communities. Social Science & Medicine 68 (11), 2061–2068.

Inhorn, M., 1994. Quest for Conception: Gender, Infertility and Egyptian Medical Traditions. University of Pennsylvania Press, Philadelphia.

Inhorn, M., 1996. Infertility and Patriarchy: The Cultural Politics of Gender and Family Life in Egypt. University of Pennsylvania Press, Philadelphia.

Inhorn, M., van Balen, F. (Eds.), 2002. Infertility around the Globe: New Thinking on Childlessness, Gender, and Reproductive Technologies. University of California Press, Berkeley.

Kielmann, K., 1998. Barren Ground: contesting identities of infertile women in Pemba, Tanzania. In: Lock, Margaret, Kaufert, Patricia A. (Eds.), Pragmatic Women and Body Politics. Cambridge University Press, Cambridge, pp. 127–163.

Karmaus, W., Juul, S., 1999. Infertility and subfecundity in population-based samples from Denmark, Germany, Italy, Poland and Spain. European Studies of Infertility and Subfecundity in and EC/COST Concerted Research Program.

Kohler, H.-P., Billari, F.C., Ortega, F.C., 2002. The emergence of the lowest low fertility in Europe during the 1990s. Population and Development Revew 28 (4), 641–680.

Larsen, U., 2000. Primary and secondary infertility in Sub-Saharan Africa. In International Journal of Epidemiology 29, 285–291.

Larsen, U., 2003. Infertility. In: Demeny, Paul, McNicoll, Geoffrey (Eds.), Encyclopedia of Population. Macmillan Reference, USA.

Leonard, L., 2002. Problematizing fertility: "Scientific" accounts and Chadian Women's narratives. In: Inhorn, Marcia, Balen, Frank van (Eds.), Infertility Around the Globe: New Thinking on Childlessness, Gender and Reproductive Technologies. University of California Press, Berkeley, pp. 193–214.

MacInnes, J., 2006. Voluntary childlessness, fertility 'plans' and the 'demand' for children. Evidence from Eurobarometer Surveys. Centre d'Estudis Demografics.

McAllister, F., Clarke, L., 1998. Choosing Childlessness. Family Policy Studies Center, London.

McAllister, F., Clarke, L., 2000. Voluntary childlessness: trends and implications. In: Bentley, G.R., Mascie-Taylor, C.G.N. (Eds.), Infertility in the Modern World: Present and Future Prospects. Cambridge University Press, Cambridge.

Morgan, S.P., Chen, R., 1992. Predicting childlessness for recent cohorts of American women. International Journal of Forecasting 8, 477–493.

Morell, C.M., 1994. Unwomanly Conduct: The Challenges of Intentional Childlessness. Routledge, London.

Okonofua, F., 2003. New reproductive technologies and infertility treatment in Africa. African Journal of Reproductive Health 7 (1), 7–11.

Okonofua, F., Obi, H., 2009. Editorial. Specialized versus conventional treatment of infertility in Africa: time for a pragmatic approach. African Journal of Reproductive Health 13 (1), 9–11.

Okonofua, F.E., Harris, D., Odebiyi, A., Kane, T., Snow, R.C., 1997. The social meaning of infertility in Southwest Nigeria. Health Transition Review 7, 205–220.

Papreen, N., 2000. Living with infertility: experiences among urban slum populations in Bangladesh. Reprod Health Matters 8, 33–44.

Philipov, D., 2003. In: Kotowska, J., Jozwiak, J. (Eds.), Fertility in Times of Discontinuous Social Change. Population of Central and Eastern Europe: Challenges and Opportunities. Statistical Publishing Establishment, Warsaw, pp. 665–689.

Philipov, D., Dorbritz, J., 2002. Demographic Consequences of Economic Transition in Countries of Central and Eastern Europe. Council of Europe Pubishing. Population Studies, No. 39. Geneva.

Philipov, D., 1998. Major Trends Affecting Families in Central and Eastern Europe. Vienna Institute of Demography.

Rebecca, U., 1999. "Our Blood Does not Agree": Negotiating Infertility in Northern Botswana. Ph.D. Dissertation. Department of Anthropology, Brown University.

Rindfuss, R.R., Morgan, S.P., Swicegood, G., 1988. First Births in America; Changes in the Timing of Parenthood. University of California Press, Berkeley.

Roth, A.D., 2001. Mchango, menses and quality of eggs: women's perception of fertility risks. In: Ties, Boerma J., Mgalla, Z. (Eds.), Women and Infertility in Sub-Saharan Africa. Royal Tropical Institute, The Netherlands.

Rowland, D.T., 2007. Historical trends in childlessness. Journal of Family Issues 28 (10), 1311–1337.

Ruthstein, S.O., Shah, I.H., 2004. Infecundity, Infertility and Childlessness in Developing Countries. DHS Comparative Reports No. 9. ORC Macro and the World Health Organization, Calverton, Maryland, US.

Sardon, Jean-Paul, 2000. The demographic situation in Europe and the developed countries. Overview: an annual report. Population: An English Selection 12, 293–328.

Silva, S., 2009. Mothers of solitude: childlessness and intersubjectivity in the upper Zambezi. Anthropology and Humanism 34, 179–202.

Skramstad, H., 1997. Coping with childlessness: the Kanyaleng Kafos in Gambia. In: IUSSP Seminar on Cultural Perspectives on Reproductive Health, Rustenburg, South Africa, June 16–19, 1997.

Snow, R.C., Okonofua, F.E., Kane, T., Farley, T.M.M., Pinol, A., 1995. Prevalence and determinants of infertility in Ile-Ife, Nigeria. Contraception, Fertility and Sex. Special Supplement No. 23, 1–544.

Sobotka, T., 1989. Fertility in central and Eastern Europe after 1989. Collapse and gradual recovery. Historical Social Research/Historische Sozialforschung 36 (2), 246–296.

Sundby, J., Jacobus, A., 2001. Health and traditional care for infertility in the Gambia and Zimbabwe. In: Boerma, J. Tiers, Mgalla, Zaida (Eds.), Women and Infertility in Sub-Saharan Africa. Royal Tropical Institute, The Netherlands, pp. 257–268.

Suggs, D.N., 1987. Female status and role transition in the Tswana life cycle. Ethnology 26 (2), 107–120.

United Nations, 1994. Programme of action of the 1994 international conference on population and development. (A/CONF, 171/13). Reprinted in Population and Development Review 21 (1), 187–313, 21 (2), 437–61.

Veevers, J.E., 1983. Childless by Choice. Butterworths, Toronto.

Vaessen, M., 1984. Childlessness and Infecundity. World Fertility Survey Comparative Studies No. 31. International Statistics Institute, Voorburg.

Wiersema, N.J., Drukker, A.J., Dung, M.B.T., Nhu, G.H., Lambalk, B., 2006. Consequences of infertility in developing countries: results of a questionnaire and interview survey in the South of Vietnam. Journal of Translational Medicine 4, 54–61.

World Health Organization, 1987. Infections, pregnancies and infertility: perspectives on prevention. Fertility and Sterility 47, 944–949.

World Health Organization, 1999. WHO Laboratory Manual for the Examination of Human Semen and Sperm-cervical Mucus Interaction, fourth ed. Cambridge University Press, Cambridge, UK.

World Health Organization, 2000. Unsafe Abortion: Global and Regional Estimates of the Incidence Unsafe Abortions and Associated Mortality, fourth ed. Geneva, Switzerland.

World Health Organization, 2004. A Tabulation of Available Data on Prevalence of Primary and Secondary Infertility. WHO/MCH/91.9, Geneva.

Wrigley, E.A., Schoefield, R.S., 1983. English population history from family reconstructions. Summary results 1600–1799. Population Studies 37, 157–184.

Informal and Underground Economics

Bruno S Frey, University of Warwick, Coventry, UK; and CREMA (Center for Research and Economics Management and Arts), Zurich, Switzerland
Friedrich Schneider, Johannes Kepler University, Linz, Austria

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by B.S. Frey, volume 11, pp. 7441–7446, © 2001, Elsevier Ltd.

Abstract

This article presents estimations of the shadow economies of 17 OECD countries, 21 developing countries, and 14 transition countries for the mid-1990s and mid-2000s. Additionally the reasons why people work in the shadow or underground economy are presented. Also the effect of the underground economy on the official economy is discussed and the presence of a shadow economy tends to overestimate the inflationary effects from a fiscal or monetary stimulus, and tends to understate the respective effects on unemployment. Finally, policy consequences are shown. One major result is that a positive approach to raise the motivation to stay in the official economy by improving the efficiency of public services, reducing the tax and social security burden and/or by raising the civic virtue, has been used only rarely, but is a highly efficient policy tool.

There is a widespread feeling that the underground economy is of substantial size and, compared to the official economy, has been growing rapidly. This holds in particular for developing and transition economies, but also for developed economies. The major concern is that government revenue is thereby lost and the cement of society is undermined. On the other hand, underground economies are value-creating and introduce a dynamic element into the economy. The underground economy is variously defined, but the predominant definition is to comprise all at present not officially recorded productive (i.e., value-adding) activities, which should be part of gross national income. This definition excludes activities not in the gross national product (GNP) (e.g., housework), and it is not identical to tax evasion (which is mostly redistributive) nor to illegal activities (some activities are legal but not subject to taxes, and therefore tend to escape being recorded). Economists have devised a considerable number of approaches to measure the size of the underground economy. The direct approaches are based on surveys and the examination of tax returns. The indirect approaches analyze the discrepancies between what would be 'normal' and what is actually observed with respect to spending, employment, the use of money, and physical input (such as electricity).

A more refined approach constructs a model identifying the incentives to work in the underground economy: the burden of taxation and social security contributions, expected punishment, and the moral cost of being active in the illegal economy. Empirical estimates of the size and the development of the underground economy relative to official recorded GNP are presented for developed, transition, and developing economies. In most countries and periods, the underground economy has been negatively evaluated and therefore considerable efforts have been expended in fighting it. Most policies seek to increase deterrence by stepping up controls and raising punishment, often with doubtful success. More promising policies are to lure people from the underground economy into the official economy by legalizing part of the respective activities, improving the efficiency of public services and the relationship of the state to its citizens, thereby raising civic virtue, and thus reducing the incentive to work in the underground economy.

There is a widespread feeling that a substantial and increasing share of value-creating activities take place outside the official economy. This holds, in particular, for developing and transition economies, but also for high-income economies. Such activities are not recorded by the system of national income accounting, which has become the accepted standard in all countries of the world.

The existence and increase of an underground economy gives rise to four major sets of concerns. The economic and social conditions of individuals, households, and countries are evaluated in a biased way if one relies on the official statistics. Thus, the official number of unemployed persons may hide that a (unknown) share of them actually work and receive wage income. As a consequence, the macroeconomic policies are likely to be too expansionary and social policy too excessive. A second concern is the loss of tax revenue, as underground activities escape taxation. A third concern interprets the underground economy as an indicator of an unhealthy relationship between citizens and government. The taxpayers are dissatisfied with the public services they get for their contributions and seek to redress the balance by escaping to the underground economy. It is feared that such a reaction makes a government unable to finance the public goods necessary for an economy and society. A fourth concern goes even further, and fears that an underground economy undermines the cement of society because it favors everyone acting in his or her own interests, while solidarity is renounced. In contrast to these concerns, the existence of an underground economy has also been seen in a positive light. Opponents of extensive state activity welcome the limits imposed on taxation and regulation, and applaud the dynamism and flexibility in the underground sector.

Denominations and Definitions

The phenomenon is known, and has been discussed in literature, under many different names: informal, unofficial,

irregular, parallel second underground, subterranean, hidden, invisible, unrecorded, and shadow economy or moonlighting. In several languages, the term most often used is 'black economy' ('le travail au noir' in French, 'Schwarzarbeit' in German, and 'svarta sektor' in Swedish).

No single definition exists, but it depends on the purpose. The most precise and predominantly used definition seeks to relate the underground economy to officially measured national income: it comprises all presently not recorded productive (i.e., value-adding) activities which should be in the GNP. This definition allows us to compare and to add the underground economy to GNP.

This definition excludes two major activities:

1. Production that by convention is not part of GNP, in particular private household activities. The evaluation of its size has engendered a research area of its own. Depending on the approach and measurement technique in developed economies, the household sector comprises between 30 and 50% of GNP.
2. Tax evasion is not value-adding but redistributional and is therefore not included as such in the above definition (e.g., when taxes on interest payments are evaded). However, in general, no taxes are paid on underground activities (such as moonlighting for house building) which are value-adding. Thus, underground activities and tax evasion are related, but certainly not identical. Tax evasion has also become a research area of its own. For the United States, between 1973 and 1992 for example, it has been estimated that 17 of total taxes owned have been evaded.

The underground economy should not be identified with illegality. Some activities are perfectly legal but are not subject to taxes (e.g., because of their small size), and therefore escape measurement in official statistics. Other activities are legal as such, but taxes are evaded. Finally, on illegal activities (such as drug production and distribution), no taxes are paid.

Broader definitions of the informal and underground economy, depending on their purpose, include private household production and redistributional activities. In the following, the more narrow definition considering unrecorded productive activities will be focused on.

Measurement Approaches

The major effort of economists has been to measure the size of the underground economy relative to official GNP; its size compared to other countries, and its development over time. The other social sciences, especially sociology, have put more emphasis on the structure: who participates where? The respective studies often provide highly detailed analyses of particular regions, industries, and types of underground activities. A considerable part (it varies strongly from country to country) of informal activity is undertaken as a second job and part-time. This allows underground workers to be eligible for social security benefits and makes detection more difficult. Not surprisingly, casual work is dominant (if not the only kind of work) for illegal immigrants. In developed economies, casual work is most prevalent in the following sections: in agriculture (especially seasonal labor), in construction, and in the services sector, particularly in repairs (automobiles and other) and cleaning, in hotels and restaurants, as well as in wholesale and retail trade. Accordingly, among the occupations with most illegal work are cleaning persons, waiters, people working at home, and construction and mechanical workers.

In the 1980s and 1990s, the quest to measure the size of the underground economy led to interesting methodological innovations, i.e., the challenge to 'measure the invisible' resulted in new measurement methods. Three general sets of measurement approaches can be distinguished.

Direct Approaches

An obvious way to analyze the underground economy is to undertake surveys among (supposed) suppliers and demanders of casual services. This method allows us to gather a detailed picture of the structure of this sector. However, because a substantial part of these activities is illegal, the interviewees may not be prepared to disclose their involvement. Somewhat surprisingly, research suggests that disclosure is not a serious problem. Typically, men are more involved in the underground economy than women, the young more than the old, people without work more than employed people, and the most prominent sectors are construction and all kinds of services. In contrast, where substantial capital assets (which are visible) are needed, less underground activity is observed.

A second direct approach is based on auditing of tax returns undertaken by tax collection and social security administrations. A sample of tax-payers is scrutinized in depth and under threats of sanctions for failing to collaborate. This approach provides detailed information on the biggest evaders, in particular the self-employed, who have better opportunities for concealment. Capital income also lends itself easier to evasion than labor income. Tax audits have proved unable to reveal all tax evasion, and are limited to taxable activities. As a large part of tax evasion is redistributional, that part is not relevant to the narrow definition of the underground economy.

The two direct methods (surveys and tax auditing) have the disadvantage of permitting point estimates only. Moreover, it is unlikely that they capture all 'shadow' activities, so they can be seen as providing lower bound estimates. They are unable (at least at present) to provide estimates of the development and growth of the shadow economy over an extended period of time. They have, however, one considerable advantage – they can provide detailed information about shadow economy activities and the structure and composition of those who work in the shadow economy.

Indirect or Discrepancy Approaches

The underground economy is reflected in discrepancies showing up in various markets. Persons working in the unofficial sector are able to *spend* more than their officially recorded income. The discrepancy between the two may be observed at the level of individual households, as well as in the aggregate national accounts. This approach is questionable, as this expenditure–income discrepancy may either be due to measurement errors (which is indeed often the case) or to reasons unrelated to the underground economy (e.g., the use of credits, or reductions in wealth).

Another discrepancy may be observable in the *labor market*. A decline in official participation rates, or a low participation rate compared to other countries, may be an indication of unofficial work. But again, the discrepancy may be related to other factors. Moreover, this approach is unable to isolate those persons who are at the same time active in the official and the unofficial economy (which seems often to be the case, according to other approaches).

A third discrepancy may be visible in the *monetary market*. The dominant approach starts with the assumption that underground transactions are paid in cash in order to make detection more unlikely. The size of the underground economy is reflected in the amount of cash used in a country beyond that used for official transactions. This approach is elegant and easily applicable because the amount of currency is well documented. The assumption that unofficial activities are transacted in cash is, however, questionable. Empirical research suggests that between 20 and 30% of the unofficial activities are not paid in cash, i.e., either by payment in kind or via a bank. This fact is especially bothersome for the currency demand approach when the share of cash payments changes over time and differs between countries. A significant portion of some currencies is held outside the country issuing it; thus the US dollar is widely used in South America and Asia. Again, the discrepancy is influenced by many factors unrelated to the underground economy, such as the use of credit cards. Finally, it is problematic to infer the size of the unofficial sector from currency transactions, because the velocity of cash circulation may differ between the official and the unofficial sector. Some of the difficulties just mentioned have been successfully addressed by more recent research. Thus, the use of credit cards and the amount of currencies outside a country have been taken into account. Most importantly, instead of comparing the actual use of cash to the one deemed necessary for the official economy, a cash demand function is empirically measured, i.e., econometrically estimated. This allows us to check for influences (such as changes in the interest rate or the increasing use of cash substitutes) unrelated to the underground economy. Moreover, the extra use of cash has been directly attributed to causal factors, in particular to an increase in the tax and social security burden. Figures for the size and development of the shadow economy can be calculated by comparing the difference between the development of currency when the direct and indirect tax burden and government regulations are held at their lowest value, and the development of currency with the current (higher) burden of taxation and government regulations. The currency demand approach is one of the most commonly used approaches.

The most recent discrepancy approach looks at *physical inputs*, in particular the use of electricity. How much electricity would normally be used to produce the official national income is calculated. The excess use can be attributed to the underground economy. This approach again has the great advantage of relying on easily available data, which is a distinct advantage for developing and transition economies. However, not all underground activities use much, if any, electricity, and the relationship between production and electricity used may change over time, or differ between countries, due to substitution and technical progress.

A general problem of all discrepancy approaches is that one has to assume a base year without underground economy. Only then is it possible to attribute the existence and rise of a discrepancy to the underground economy.

The Model or Multiple Indicator and Multiple Causes Approach

This method focuses on the causes and effects of the underground economy. By constructing a model, it seeks to identify the unobserved sector in between.

Three sets of factors are taken to be the main motivators of whether to engage in the unofficial economy:

1. The burden of taxation and social security contributions as well as government regulations provide incentives to exit the official economy. The latter hinder or even prohibit activities in the official economy (e.g., because of lacking work permits, especially for foreigners), which provides an incentive to undertake them unofficially. This holds, of course, also for illegal activities such as the production and distribution of drugs. An important reason for exiting the official sector is the restrictions imposed on working time. Low work hours, long holidays, early retirement, and, above all, unemployment, are important reasons for moonlighting.
2. The expected punishment of working in the underground economy provides a disincentive for doing so. It consists of the probability of being caught and the size of punishment. For persons outside the established society (for example, illegal immigrants), or for self-employed persons, expected punishment is lower than for other persons, and therefore a higher rate of participation in the underground sector is likely.
3. The moral costs are another disincentive to be active in the underground economy. A good citizen has moral qualms about undertaking a forbidden activity. These moral costs are closely related to 'tax morale' which motivates citizens to pay their dues to the state.

The effects of the underground economy are reflected in the traces visible in the labor, money, and product markets discussed earlier. A specific econometric technique called 'unobservable variables approach' allows us to estimate the size of the underground economy. This technique seeks to identify the size of the underground sector, which cannot be observed itself, but whose causes (e.g., tax and social security burden, unemployment) and whose effects (e.g., fall in the participation ratio) can be measured. This approach is the most comprehensive and builds on a well-structured behavioral model, but it requires a large amount of data. As these are often not available (in particular not for developing and transition economies), this approach is not generally applicable over a long (extended, 30 years) period of time; for 162 countries data are available over 1999–2008. The estimation technique moreover tends to be statistically fragile (i.e., small changes in specification and values of variables affect the estimates).

Empirical Estimates

The following tables serve to indicate approximate magnitudes of the size and development of the underground economy, defined as productive, value-adding activities, which should be

part of GNP. Table 1 presents a rough comparison of the size of the underground economies relative to GNP for a selection of Western European countries, Japan, and the United States for the mid-1990s, using the currency demand approach and for the mid-2000s the Multiple Indicator and Multiple Causes (MIMIC) approach.

The South European countries (Greece and Italy) have an underground economy almost one-third as large as the officially measured GNP: followed by Spain, Portugal, and Belgium, having a shadow economy between 20 and 24% of (official) GNP. According to these estimates, the Scandinavian countries also have a sizable unofficial economy (between 18 and 20% of GNP), which is attributed mainly to the high fiscal burden. The 'central' European countries (Ireland, The Netherlands, France, Germany, and Great Britain) have a smaller underground economy (between 13 and 16% of GNP), probably due to a lower fiscal burden and moderate regulatory restrictions. Smaller underground economies are estimated to exist in countries with relatively low public sectors (Japan, the United States, and Switzerland) and comparatively high tax morale (United States and Switzerland).

Table 2 provides a rough comparison of the size of the underground economy relative to official GNP for a selection of developing and transition economies for the beginning of the 1990s, using the physical input (electricity) demand approach and for the mid-2000s the MIMIC approach. Some of these countries (Nigeria, Egypt, and Thailand) are estimated to have an underground sector nearly three-quarters the size of officially recorded GNP. In many countries, the size is one-quarter to one-third of GNP. In Asian countries, with a comparatively low public sector, high tax morale, or high expected punishment (Hong Kong, Singapore), the underground economy is estimated to be similar to that in many 'Northern' European countries.

Transition economies are estimated to often have substantial unofficial activities, many around one-quarter of GNP. An exception is ex-Czechoslovakia where, according to these estimates, the underground sector is clearly around 10% of GNP.

Table 1 Size of the underground economy relative to GNP in various OECD countries, mid-1990s (estimation based on the currency demand approach for the mid-1990s and the MIMIC method for the mid-2000s)

	Mid-1990s	*Mid-2000s*
Greece	27–30%	26–28%
Italy		23–25%
Spain		20–22%
Portugal	20–24%	20–22%
Belgium		19–21%
Sweden		16–18%
Norway	18–23%	16–18%
Denmark		15–17%
Ireland		13–15%
France	13–16%	12–14%
Netherlands		10–12%
Germany		15–17%
Great Britain		11–12%
Japan		9–10%
United States	8–10%	7–8%
Austria		9–11%
Switzerland		8–9%

Source: Compiled from Schneider, F., Enste, D., 2000. Shadow economies: size, causes and consequences. Journal of Economic Literature 38, 77–114; Schneider, F., Buehn, A., Montenegro, Cl., 2010. New estimates for the shadow economies all over the world, International Economic Journal 24 (4), 443–461 and Feld, L., Schneider, F., 2010. Survey on the shadow economy and undeclared earnings in OECD countries. German Economic Review 11 (2), 109–149.

Table 2 Size of the underground economy relative to GNP in various developing and transition countries (for the mid-1990s, estimates are based on the physical input (electricity) demand approach and for the mid-2000s on the MIMIC method)

Developing countries	*Mid-1990s*	*Mid-2000s*
Africa		
Nigeria	68–76%	53–55%
Egypt		33–35%
Tunisia	39–45%	31–33%
Morocco		16–17%
Central and South America		
Guatemala		49–51%
Mexico	40–60%	28–30%
Peru		55–58%
Panama		60–64%
Chile		18–19%
Costa Rica	25–35%	25–26%
Venezuela		32–35%
Brazil		37–39%
Paraguay		37–38%
Columbia		35–37%
Asia		
Thailand	70%	48–50%
Philippines		39–42%
Sri Lanka	38–50%	42–44%
Malaysia		30–31%
South Korea		25–27%
Hong Kong	13%	15–16%
Singapore		12–13%
Transition economies		
Central Europe		
Hungary	24–28%	23–24%
Bulgaria		33–35%
Poland	16–20%	26–27%
Romania		30–32%
Slovakia	7–11%	17–18%
Czech Republic		17–18%
Former Soviet Union countries		
Georgia		63–65%
Azerbaijan	28–43%	53–58%
Ukraine		46–48%
Belarus		44–46%
Russia		41–43%
Lithuania	20–27%	29–31%
Latvia		27–29%
Estonia		29–31%

Source: Compiled from Schneider, F., Enste, D., 2000. Shadow economies: size, causes and consequences. Journal of Economic Literature 38, 77–114, for the mid-1990s, and Schneider, F., Buehn, A., Montenegro, Cl., 2010. New estimates for the shadow economies all over the world. International Economic Journal 24 (4), 443–461, for the mid-2000s.

Table 3 Growth of the underground economy relative to GNP for selected West European countries and the United States, 1960–95 (estimates based on the currency demand approach and for the mid-2000s on the MIMIC method (rounded figures))

	1960	*1995*	*Percentage point increase from 1960 to 1995*	*2005*	*Percentage point increase/ decrease from 1995 to 2005*
Sweden	2%	16%	+14.0%	17.5%	+1.5%
Denmark	4.5%	17.5%	+13.0%	16.5%	−1.0%
Norway	1.5%	18.0%	+16.5%	17.6%	−0.4%
Germany	2%	13.2%	+11.2%	15.4%	+2.2%
United States	3.5%	9.5%	+6.0%	8.2%	−1.3%
Austria	0.5%	7%	6.5%	10.3%	+3.3%
Switzerland	1%	6.7%	5.7%	9.0%	+2.3%

Source: Compiled from Schneider, F., Enste, D., 2000. Shadow economies: size, causes and consequences. Journal of Economic Literature 38, 77–114 and Feld, L., Schneider, F., 2010. Survey on the shadow economy and undeclared earnings in OECD countries. German Economic Review 11 (2), 109–149, for 2005.

Table 3 reports estimates of the growth of the underground economy (relative to GNP) for selected Western countries and the United States, using the currency demand and the MIMIC approach.

The Scandinavian countries (Sweden, Norway, and Denmark) and the German-speaking countries (Germany and Austria) exhibit a sizable increase of the underground economy within the 35 years covered (1960–95) and tend to understate the respective effects on unemployment. Also the countries with a low share in the beginning (Switzerland and the United States) show a significant increase, with the US share more than doubled. Sizable increases have been estimated, with few exceptions, for all types of countries and all kinds of approaches: the increasing importance of the underground relative to the official economy is a robust phenomenon. It has been attributed to many different causes, but the most important and most often cited ones are the rise in the burden of taxes and social security contributions, the increase in public regulations, especially on labor markets, as well as a long-term fall in civic virtue in the population.

The Effects of the Underground Economy on the Official Economy

In order to study the effects of the underground on the official economy, the underground economies have been integrated into macroeconomic models; this yields an extended macromodel of the business cycle, as well as tax and monetary policy linkages with the shadow economy. As a result, it becomes clear that these effects should be taken into account for tax and regulatory policies. The presence of a shadow economy tends to overstate the inflationary effects of a fiscal or monetary stimulus and tends to understate the respective effects on unemployment. When the growth of the shadow economy and the 'official' economy are positively related (which is likely to be the case when entry costs into the shadow economy are low due to a low probability of enforcement), an expansionary fiscal policy has a positive stimulus for both the formal and informal economies. It has also been found that the US productivity slowdown over the period 1970–89 was vastly overstated, as the under-reporting of income, due to the more rapid growth of the US shadow economy during this period, was disregarded.

The underground economy is beneficiary insofar as it responds to the economic environment's demand for urban services and small-scale manufacturing. The casual sector provides the economy with dynamic and entrepreneurial spirit and can strengthen competition, increase efficiency, and put effective limits on government activities. The casual sector contributes to the creation of markets, increases financial resources, and transforms the legal, social, and economic institutions necessary for accumulation. Moreover, a substantial part (up to 70%) of the earnings gained in the shadow economy is quickly spent in the official sector and thus boosts the official economy. These expenditures tend to raise consumer expenditures as well as (mostly indirect) tax revenues.

Policy Consequences

The growth of the underground economy over the few last decades and its effect on the official economy, both in general perception and scholarly research, has prevailingly been evaluated as a negative development, which should be counteracted. In particular, politicians and public officials have pointed out that the state's capacity to provide the desired public services is undermined because tax revenue is lost. But not all commentators share this view. Some see it as a welcome and effective limit on the tendency of governments to continually raise the fiscal burden and to impose more and more bureaucratic restrictions on the economy and society. More generally, the rise of the underground economy is interpreted to indicate that the relationship between the state and the citizens is fragile and needs improvement. It has also been argued that many developing and transition economies would break down or function at a much lower level of production and welfare if the underground economy did not exist.

The fight against the underground economy is a recurrent theme in many countries. The dominant method is to increase deterrence. The probability of being caught is raised by more regular and intensive controls, often by the police. Punishment is raised by imposing higher fines and, in severe cases, prison sentences. Both consumers and suppliers of goods and services produced in the underground sector are targeted.

The success of such deterrence policies is rather doubtful. In the case of the underground sector related to drugs, prostitution, and alcohol prohibition, the effect has turned out to be at best weak, if not counterproductive. Many persons active in the underground economy move more deeply into illegality in order to make detection more difficult. As a result, the underground activities become more criminalized, and more difficult to observe and to influence. Well-organized and ruthless, organized crime (often called the Mafia) becomes more dominant.

A deterrence policy need not always be a failure, but the successes tend to be short run. Provided the demand for

underground goods and services remains intact, the profit opportunities in the underground economy become so large that supply reappears and the underground economy recuperates. A less oppressive policy is to lure people into the official economy by legalizing parts of the underground economy, and by facilitating the move into the official economy (e.g., by granting an amnesty). Such measures have proved to be only moderately successful.

A positive approach to raise the motivation to stay in the official economy by improving the efficiency of public services, reducing the tax and social security burden imposed on labor, and/or by raising civic virtue has been used only rarely. Many policy makers doubt whether such measures work at all. In any case, they are effective only in the longer run. One possibility is to adjust public supply more closely to what the citizens desire, and to lower cost for a given quantity and quality of public supply. This can be attempted by streamlining public sector activities (e.g., by New Public Management). A more fundamental way is to improve the political process by opening the political arena to contending interests, and by strengthening the democratic participation rights of the citizens.

See also: Work, History of; Work, Sociology of.

Bibliography

Andreoni, J., Erard, B., Feinstein, J., 1998. Tax compliance. Journal of Economic Literature 36, 818–860.

Cowell, F.A., 1990. Cheating the Government. The Economics of Evasion. MIT Press, Cambridge, MA.

De Soto, H., 1989. The Other Path: The Invisible Revolution in the Third World. Harper and Row, New York.

Feige, E.L. (Ed.), 1989. The Underground Economies. Tax Evasion and Information Distortion. Cambridge University Press, Cambridge, UK.

Feld, L., Schneider, F., 2010. Survey on the shadow economy and undeclared earnings in OECD countries. German Economic Review 11 (2), 109–149.

Frey, B.S., Werner, W., Pommerehne, March 1984. The hidden economy: state and prospect for measurement. Review of Income and Wealth 30, 1–23.

Gaertner, W., Wenig, A. (Eds.), 1985. The Economics of the Shadow Economy. Springer-Verlag, Berlin.

Mars, G., 1994. Cheats at Work. An Anthropology of Workplace Crime. Dartmouth, Brookfield, CT.

Pozo, S. (Ed.), 1996. Exploring the Underground Economy: Studies of Illegal and Unreported Activity. W. E. Upjohn Institute for Employment Research, Kalamazoo, MI.

Pyle, D.J., 1990. The Informal Economy. Kluwer Academic Publishers, Utrecht.

Rogoff, K., 1998. Blessing or curse? Foreign and underground demand for the Euro notes. Economic Policy: The European Forum 26, 261–304.

Schneider, F., Enste, D., 2000. Shadow economies: size, causes and consequences. Journal of Economic Literature 38, 77–114.

Schneider, F., Buehn, A., Montenegro, CI, 2010. New estimates for the shadow economies all over the world. International Economic Journal 24 (4), 443–461.

Simon, C.P., Witte, A.D., 1982. Beating the System – the Underground Economy. Auburn House Pub. Co., Boston.

Tanzi, V. (Ed.), 1982. The Underground Economy in the United States and Abroad. Lexington Books, Lexington, MA.

Thomas, J.J., 1992. Informal Economic Activity. Wheatsheaf, London.

United Nations, 1993. Inventory of National Practices in Estimating Hidden and Informal Economic Activities for National Accounts. United Nations, Geneva.

Weck-Hannemann, H., Pommerehne, W.W., Bruno, S.F., 1984. Schattenwirtschaft. F. Vahlen, Munich.

Informal Sector

Karen T Hansen, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A processual notion of informality is replacing that of an informal sector, which emerged in the early 1970s from research on urban development as a name for activities and enterprises escaping formal economic regulation. It and other terms for nonregulated economic activities have since then been applied to research in the West, yet the bulk of work continues to be undertaken in the developing world. Characterizing organizational processes as informal recognizes complex linkages between informal and formal activities across all economic levels rather than within clearly bounded sectors. The focus has changed from descriptions of informal activities in different settings to questions about their potential for growth to the dynamics of informalization in general.

Meanings of 'Informal'

In social science scholarship since the early 1970s, including policy research and development economics, and in applied work, the term 'informal' has been used variously for a segment of the labor market, a set of activities, a nonregulatory status, a form of insertion in labor markets, a type or size of firm, self-constructed housing, and a general characterization of the economy.

Although the notion of an informal sector was new, the activities to which it referred were not. Many had long histories in the West as well as the developing world. In effect, in the 1970s, the notion of an informal sector gave name to economic activities that generated income for large portions of urban populations in developing countries. With the unprecedented increase of urbanization in the developing world since the opening decade of the twenty-first century, more than half of the populations of many of these cities make a living and obtain shelter informally. Such processes represent persistent rather than provisional forms of urbanism, and they have complex links with the rest of the city, the region, and the global economy.

Informality: Sector or Process

In the early 1970s in field research among migrants in Accra, Ghana, Hart coined the term 'informal economy' (1973), contrasting their economic activities to formal, contractual wage labor. Since then, the informal sector has been a subject of extensive debate and research. The term 'informal' is more of a residual notion than a concept, defined in relationship to something else, that is, to what it is not. The chief questions raised in this scholarship concerned the content, structure, and dynamics of the informal sector, that is, what to include and exclude, how precisely to differentiate it from the formal sector, and what growth potentials it might have. Some scholars suggested alternative terms and frameworks, such as casual work, petty commodity production, or householding and provisioning (Bromley and Gerry, 1979). In francophone scholarship, informality is often referred to as *systeme D* (from *l'economie de la debrouillardise*), meaning self-reliance or do-it-yourself. Still, the 'informal sector' notion continues to be widely used for lack of any better descriptive term.

Despite their conceptual misgivings, most scholars agree that discussions about economic life must take account of informal economic development in its relationship to wider economies. For example, MacGaffey (1991) found informal activities to be so central to efforts at making a living in Zaire (now the Democratic Republic of Congo) during the last half of the 1980s that she characterized them as the 'Real Economy.' The idea of 'popular' or 'informal economies' is another way of highlighting the pervasiveness of nonregulated economic practices, recognizing their overall importance (Guyer et al., 2002). Thinking processually about 'informalization' avoids the arbitrariness of sectoral dichotomies (Hansen and Vaa, 2004). It also helps bridge the conceptual gap between production and reproduction, directing attention to household strategies and the ways in which household composition, stage of development cycle, and the gender and age of members introduce cultural assumptions into the division of labor. Women and children are highly visible segments of many informal markets, and development practitioners and human rights groups have recently begun to pay attention to the relationships of age and gender to informal sector developments, specifically with regard to trafficking and forced labor.

The 'Informal Sector,' Urbanization, and Development

Research into informal economic development in the developing world has clustered around several specific policy approaches aimed at promoting economic growth. The first cluster of work was influenced by Hart's (1973) research among migrants in Ghana that identified an 'informal economy' of self-employment, which was legitimate yet illegal, as opposed to regularized wage labor. International development agencies were at this time concerned about the failure of industrialization efforts to improve economic growth. Hart's observations were quickly absorbed into development policy directed at satisfying basic needs by harnessing informal income opportunities. In the International Labour Office's (ILO) Kenya mission, for example, such activities were explained as 'a way of doing things' (ILO, 1972), characterized by ease of entry, reliance on indigenous resources and family ownership, smallness of scale,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12094-X

being labor intensive with adopted technology, using skills acquired outside the formal school system, and participating in unregulated and competitive markets. Here, as elsewhere in the developing world, governments were encouraged to take a positive role, ceasing to demolish unauthorized housing, changing trade and licensing procedures, and providing technical and outreach support to informal sector activities.

The second cluster of work appeared in the wake of the International Monetary Fund's (IMF) structural adjustment programs, particularly from the late 1980s onward. Many of the policy measures (e.g., employment entrenchment in the public and private sectors; the introduction of value-added taxes; the removal of subsidies on food, fertilizer, and housing; and the introduction of user fees in schools and hospitals) advocated by the IMF and World Bank to transform state-controlled economies into market economies had adverse effects on already vulnerable groups. In fact, the volume and differentiation of informal sector activities increased during the 1980s. Strapped for revenue and burdened by accelerating international debt, many governments began encouraging interventions by local and external nongovernmental organizations (NGOs). The centerpiece of this renewed policy attention is 'microenterprises' and their enhancement through credit arrangements and the provision of basic training in technical skills. Perhaps more than any other scholar, King (1996) has documented, with rare longitudinal insights, the local effects of these changing policy interventions on the actions and achievements of informal sector artisans in Kenya.

Some observers who were disenchanted with the lackluster government performance in developing countries over previous decades viewed informal sector activities as examples of free market initiatives. The third, and more recent, approach to informality is represented by de Soto (1989, 2000), who offers a vision of informal entrepreneurship akin to heroism with people spontaneously and creatively responding to the state's inability to deliver basic services. In this view, legalization will enable people to secure and trade their assets, thus becoming part of the formal system. Property ownership, according to de Soto, will lift people from poverty. Attractive to the World Bank and some development agencies, de Soto's thesis has not found much empirical support.

The 'Informal Sector' and the West

While visually evident in streets and markets in developing countries, informal economic activities take place everywhere. Throughout most of the West today, small-scale enterprises that bypass official regulation and use off-the books labor practices put people to work providing goods and services (Portes et al., 1989). In some regions of the United States, such activities are associated with immigrant communities, whereas in parts of Europe, for example, they are based on long histories of artisanal enterprises and putting-out systems (Mingione, 1991). In the second or parallel economies of the former socialist countries in Eastern Europe, producers and consumers supplied, swapped, and bartered goods and services to alleviate market shortages, avoiding the bureaucratic bottlenecks of centrally planned economies. Also, informal sector activities in the West are not exclusively associated with a bygone era of economic history but keep reappearing in different places, operating in a variety of politico-economic contexts along with globalization (Sassen, 1998).

Policy Concerns

Since the turn of the last millennium, informality has come back on the agenda of international development organizations, urban planners, and researchers. In the renewed attention to informality, definitional issues continue to be debated as they have important implications for the identification of target groups and types of intervention. If the informal sector provides an economic safely net, it does so in different ways, depending on the type and level of activity, the resources brought to bear on its operation, and the local cultural context that shapes the participation of actors (Rakowski, 1994). Whether a growth potential in fact exists requires understanding of the nature both of the activities in question and of the social relations that structure them. In addition, the growth potential is not only a technical challenge or problem for urban management but also, in many places, a complex political struggle that provokes the use of force and violence.

Some researchers and development organizations avoid definitional discussions by using descriptive terms like 'microenterprises' and highlighting the role of informal networks in creating social capital. But microenterprises encompass a range of activities that draw on widely diverse access, require different types of resources and inputs, and, above all, put the young and the old and women and men to work on unlike terms. While some microenterprises may well contribute to economic empowerment and popular political participation, many others may fragment into networks of accumulation and survival that marginalize the efforts of the very economic actors who are pursuing them (Meagher, 2010).

Unless the economic and social nature of informal economic activities is accounted for in more specific terms, development assistance risks being restrictive, thus jeopardizing the very goal of expanding employment and income generation that it seeks to promote. Anticipating that persons with few means may forfeit loans, some credit programs, for example, target better-off participants, in this way accentuating already existing socioeconomic differentiation rather than reducing it.

Policy planning by government agencies and development organizations easily excludes informal economic activities because of their very nature: unregulated, always changing, and often mobile. Assessing their scale and scope is difficult because of the lack of comparable statistical information with which to document and analyze the characteristics of activities, enterprises, their built environment, the institutional or organizational context in which they operate, as well as the needs of their participants in terms of gender and age. In recent years, participatory and bottom-up approaches have begun to play local and transnational roles in mobilizing poor urban residents with regard to informal economic activity and informal housing. The mapping and documentation by residents of informal settlements may help to give them a voice in

lobbying, for example, about water, sanitation, and land issues. Being counted and showing that you have an address are part of the rhetoric of claiming a right to the city by populations who tend to be ignored (Patel and Baptist, 2012). Across the global urban South, such community-led exercises are sometimes linked with the establishment of local savings groups, and peer exchanges with the support of local and transnational NGOs (Lindell, 2010).

The Market Meets the 'Informal Economy'

Since the term 'informal sector' was coined in the early 1970s, the economic and political contexts in which research on this topic first emerged have changed. Shifts in the global economy are altering production processes in an international division of labor that is affecting both the industrial West and the developing world, but differently so. Neoliberalism in both politics and markets almost everywhere is giving rise to new research questions concerning the experiences of ongoing transformations. Socioeconomic differentiation has increased, juxtaposing wealth and poverty in new forms of inclusion and exclusion that are conspicuous in the urban-built environment: gated communities and upscale shopping malls alongside crowded informal settlements. Policy focus on decentralization, deregulation, economic retrenchment, and the privatization of public services has intensified economic hardship. Faced with weak or ineffective state and local governments, the absence of social safety nets, and few opportunities for secure employment, more and more urban residents turn to informal activities for their living. Their economic ventures connect with the overall economy in complicated ways, crisscrossing public and private space. This is why it may make constructive sense to approach informality as an organizing logic rather than as a specific sector of the economy (Roy, 2005). In this view, informality also plays out as practice in the growth of networks and associations in the organization of urban life (Lindell, 2010). In the recent global economic recession, especially since 2008, such processes are likely to become adversely affected and constrained, hampering the ability of informal actors to break through their economic marginalization in social and legal terms.

Beyond the Informal Sector?

Regardless of the difficulties involved in defining and assessing its scope, informality has returned not only on the scholarly development agenda but also as a topic with popular appeal from two different but equally sensationalist angles: a relentless involution of economic activity and housing comprising an urban nightmare bent on self-destruction (Davis, 2005), and a self-reliant, innovative entrepreneurship employing half of the world's population (Neuwirth, 2005, 2011). Whether one is championing its creative livelihood potential or advocating its banning, there is no doubt that an informal economic culture has taken complicated roots and that its operational practices have become pervasive across the globe. Today, the most relevant questions to explore about informalizations concern the effects of political and economic liberalizations, on members of the many households whose livelihoods continue to be made from informal activity.

While the term 'informal sector' may remain part of our descriptive vocabulary, new scholarship is likely to continue to disentangle the complexities of informal activities that need to be understood in their own right if we are to establish policy programs to support them. The enthusiastic reception in some circles of de Soto's praise of the resourcefulness of the urban poor must be tempered by a recognition of the heterogeneity and change of informal sector activities as well as of the exploitative conditions under which much of this sector's work is carried out, particularly, by women and children. Lessening this state of affairs on a global scale requires more than self-help by small-scale entrepreneurs, assistance from participatory networks and organizations, and intermittent intervention by local and transnational NGOs. In effect, it also invites governmental action on many fronts to reshape the institutional structures that continue to reproduce livelihoods in the interstices of the economy.

See also: Apprenticeship and School Learning: Lessons from Germany; Economic Anthropology; Informal and Underground Economics; Social Capital; Urban Anthropology.

Bibliography

Bromley, Ray, Gerry, Chris, 1979. Casual Work and Poverty in Third World Cities. Wiley, New York.

Davis, Mike, 2005. Planet of Slums. Verso, London.

de Soto, Hernando, 1989. The Other Path: The Invisible Revolution in the Third World. Harper and Row, New York (1986. El Otro Sendero: La Revolucion Informal. Editorial El Barranco, Lima, Peru).

de Soto, Hernando, 2000. The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else. Basic Books, New York.

Guyer, Jane I., Denzer, LaRay, Agbaje, Adigun (Eds.), 2002. Money Struggles and City Life: Devaluation in Ibadan and Other Urban Centers of Southern Nigeria, 1986–1996. Heinemann, Portsmouth.

Hansen, Karen T., Vaa, Mariken (Eds.), 2004. Reconsidering Informality: Perspectives from Urban Africa. Nordic Africa Institute, Uppsala.

Hart, Keith, 1973. Informal income opportunities and urban employment in Ghana. Journal of Modern African Studies 11, 61–89.

International Labour Office (ILO), 1972. Employment, Incomes and Inequality: A Strategy for Increasing Productive Employment in Kenya. ILO, Geneva.

King, King, 1996. Jua Kali Kenya: Change & Development in an Informal Economy 1970–95. James Currey, London.

Lindell, Ilda (Ed.), 2010. Africa's Informal Workers: Collective Agency, Alliances and Transnational Organizing in Urban Africa. Zed Books, London.

MacGaffey, Janet, 1991. The Real Economy of Zaire: The Contribution of Smuggling & Other Unofficial Activities to National Wealth. University of Pennsylvania Press, Philadelphia.

Meagher, Kate, 2010. Identity Economics: Social Networks & the Informal Economy in Nigeria. James Currey, London.

Mingione, Enzo, 1991. Fragmented Societies: A Sociology of Economic Life beyond the Market Paradigm. Blackwell, London.

Neuwirth, Robert, 2005. Shadow Cities: A Billion Squatters, a New Urban World. Routledge, New York.

Neuwirth, Robert, 2011. Stealth of Nations: The Global Rise of the Informal Economy. Pantheon Books, New York.

Patel, Sheela, Baptist, Carrie, 2012. Editorial: documenting the undocumented. Environment and Urbanization 24, 3–12.

Portes, Alejandro, Castells, Manuel, Benton, Lauren (Eds.), 1989. The Informal Economy: Studies in Advanced and Less Developed Countries. John Hopkins University Press, Baltimore.

Rakowski, Cathy (Ed.), 1994. Contrapunto: The Informal Sector Debate in Latin America. State University of New York Press, Albany.
Roy, Ananya, 2005. Urban informality: toward an epistemology of planning. Journal of the American Planning Association 71, 147–158.
Sassen, Saskia, 1998. Globalization and Its Discontents: Essays on the New Mobility of People and Money. New Press, New York.

Relevant Websites

www.homenetsouthasia.net – Homenet South Asia. Network of home-based workers in South Asia, focusing on advocacy, skills upgrading, data collection, and market linkages. Established in 2000.

www.inclusivecities.org – Inclusive Cities. Network of membership-based organizations advocating to include informal activities within urban planning processes. Established in 2008.

www.sdinet.org – Slum Dwellers International (SDI). A global network of slum dwellers' federations working to improve the property rights and lives of the poor through advocacy, research, and documentation. Established in 1996.

www.streetnet.org.za – StreetNet International. Provides resources and information to enable street vendors to organize to improve their livelihoods. Established in 2002.

www.wiego.org – Women in Informal Employment: Globalizing and Organizing (WIEGO). Network of informal organizations, improving skills, collecting data and documentation. Established in 1997.

Informal Work in Developing Countries

Arnab K Basu and Nancy H Chau, Charles H. Dyson School of Applied Economics and Management, Ithaca, NY, USA; Center for Development Research (ZEF), Bonn, Germany; and Institute for the Study of Labor (IZA), Bonn, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides a selective survey of the large literature on informal work in developing countries. The literature spans a wide range of issues, including definition, measurement, impact on workers and aggregate productivity, as well as policy linkages. We highlight three sets of issues where a renewal in research seems particularly warranted: informal work and aggregate productivity, minimum wages and informal labor market outcomes, and the labor market impact of trade liberalization in the presence of an informal sector.

How important is informal work in developing country labor markets? What are some well-established facts concerning its scale, persistence, implications on policies? What are some areas of research that warrant further scrutiny? These are the objectives of this brief survey.

As preface, it should be noted that many excellent and up-to-date surveys on informal sector activity already exist, with various emphases covering definitional issues, measurement concerns, time trends, linkages with input markets other than labor, and linkages with other ongoing macroeconomic processes and/or policies. For example, Chen (2012) classifies the many types of jobs in the informal sector, summarizes the various definitions of the informal sector, and examines the important role the informal sector plays in a developing country. Schneider (2014) articulates the various measurement issues in informal sector research and argues in favor of the use of a multiple indicators multiple causes (MIMIC) estimation method due to its flexibility. In terms of overall trends, OECD (2009a,b) provide extensive overviews of the trends in informal employment around the world. Batini et al. (2010) survey the literature on informal labor and informal credit markets and argue for the need to use a dynamic general equilibrium approach to simultaneously capture both sets of issues. A recent joint International Labour Organization (ILO) and World Trade Organization (WTO) study (ILO–WTO, 2009) takes a comprehensive look at the link between globalization (trade reforms, foreign direct investment, and economic growth) on the expansion and contraction of informal jobs in developing countries.

In this survey, our focus will be on informal work in the developing world. It is thus less ambitious in scope than studies that attempt to define, measure, and assess the full gamut of informal sector activities, their determinants, and their consequences. Yet, the definition of 'informal work' itself can be challenging – what are the various types of work that come under the umbrella term informal work, and what difference does a change in definition alone make when it comes to the causes and consequences of informal work? In what follows, we begin with a broad stroke summary of the various definitions of informal work in developing countries and highlight some of the aspects of informal work that have received particular attention in the academic literature.

Our narrower focus relative to existing surveys also allows us to zone in on a few selective sets of controversies about informal work about which we believe the verdict is very much still out, and are thus potentially fruitful areas for future theoretical and empirical research. In each case, we introduce one or more features of developing country labor markets that can potentially modify/overturn standard predictions and results. We do this for three areas of inquiry. The first concerns the relationship between the productivity of informal relative to formal work. To this active area of research we introduce three oft overlooked issues, unobserved heterogeneity, imperfect competition in the labor market, and the different shades of informality that in fact exist in what is commonly referred to as the formal sector.

Next, we will revisit a large and growing literature on the impact of minimum wage legislation on wages and employment, both formal and informal. To this literature, we will introduce the role of imperfect enforcement by government enforcement agencies, imperfect commitment by policy makers, and imperfect competition in the labor market.

Finally, we examine the issue of welfare implication of trade liberalization in the presence of an informal sector. Here, we argue that research on both the theoretical and empirical fronts is nascent, and that there are a number of fruitful areas that warrant future research. In the rest of this survey, the terms work, jobs, and employment will appear interchangeably. All three terms are synonymous with paid work in informal sector enterprises including self-employment in the developing world.

Informal Work: Definition, Measurement, and Macroeconomic Correlates

According to Chen (2012), informal employment can be divided into two broad categories: self-employment and wage employment. Each of these two broad categories can be further divided into subcategories according to an individual's status in employment. For instance, informal self-employment includes (1) employers in informal enterprises; (2) own-account workers in informal enterprises; (3) contributing family workers (in informal and formal enterprises); and (4) members of informal producers' cooperatives. Correspondingly, informal wage employment includes employees hired without social protection contributions by formal or informal enterprises or as paid domestic workers by households. Chen (2012) associates informal jobs as those consisting of (1) employees of informal

 http://dx.doi.org/10.1016/B978-0-08-097086-8.94028-5

enterprises; (2) casual or day laborers; (3) temporary or part-time workers; (4) paid domestic workers; (5) contract workers; (6) unregistered or undeclared workers; and (7) industrial outworkers (also called homeworkers). These informal jobs can be also viewed as a hierarchical pyramid with unpaid family workers at the bottom rung followed by industrial outworkers and homeworkers. These two bottom layers of the pyramid are predominantly women workers. The next two (middle) layers of the pyramid are respectively casual wage workers and own-account operators with both men and women populating these two types of jobs. Finally, regular wage workers and employers, both predominantly men, constitute the two topmost layers of the pyramid. Corresponding to this job pyramid, the pay is the lowest and the risk of poverty is the highest at the bottom layers (family workers and home-based industrial outworkers) with the pay rising (and corresponding risk of poverty falling) as one moves to the top of the pyramid where employers and regular informal wage workers reside.

Estimates of informal employment based on four selected categories of (1) enterprises with less than five workers; (2) own-account workers; (3) contributing family workers; and (4) domestic workers are provided in the ILO–WTO study (2009) for a selected group of Latin American countries. Data from 2006 show that the share of each of these informal employment categories in total informal activity ranges between 22.1 and 41.5% for enterprises with less than five workers; between 38.1 and 56% for own-account workers; between 2.4 and 16.9 for contributing family workers; and between 3.9 and 18.5% for domestic workers.

Alternatively, OECD (2002) presents a system of accounts adopted by the United Nations in 1993. In this definition, informal (or underground) work is defined as work that produces all market-based goods and services that are

> deliberately concealed from public authorities for any of the following reasons: (i) to avoid payment of income, value added or other taxes, (ii) to avoid payment of social security contributions; (iii) to avoid having to meet certain legal labor market standards, such as minimum wages, maximum working hours, safety standards, and (iv) to avoid complying with certain administrative procedures, such as completing statistical questionnaires or other administrative forms.
>
> OECD, 2002, pp. 139

Studies by the OECD (2009a,b) and reported by Schneider (2014) give general estimates of informal work across the developing world that are in line with this alternative definition. These studies first estimate the incidence of informal employment in the nonagricultural sectors of the developing world between 1990 and 2007 at 900 million workers. Inclusion of the agricultural sector more than doubles this number to 2 billion workers. The OECD studies further conclude that 1.8 billion people work in informal jobs, compared to 1.2 billion who benefit from formal contracts and social security protection. Specifically, informal economic activity in the nonagricultural sector of developing countries accounts for 75% of all jobs in sub-Saharan Africa; more than 66% in South and Southeast Asia; about 50% in Latin America, the Middle East, and North Africa; and nearly 25% in transition countries. If agriculture is included, the informal share of the economy in the above mentioned regions is even higher.

Depending on how informal jobs are defined, and the number of countries in a region included, data on the time trend of informal employment across global regions vary from study to study. For instance, the OECD (2009a,b) studies show that the share of informal employment (as percent of total nonagricultural employment) has increased over time globally with the informal employment share of South- and Central American countries rising from 32.4% during 1985–89 to 50.1% over 2000–07. Similarly, the share of informal employment in Asian countries rose from 55.9% during 1985–89 to 70.2% between 2000 and 2007, while for African countries the share of informal employment rose from 40.3% during 1985–89 to 60.5% between 2000 and 2007. By contrast, ILO–WTO (2009) finds that informal employment (as a share of total employment, not total nonagricultural employment as in OECD (2009a,b)) has remained relatively stable for Latin America from the early 1990s till the 2000s at 50–52%, the corresponding numbers for sub-Saharan Africa show a decline from 63.6% in the late-1990s to around 55.7% in the 2000s. Asia, however, shows an interesting pattern with 78% of the workforce in informal employment both in the early 1990s and in the 2000s, but a drop in the late-1990s to 68.5%. These trends are interesting since the period under study between early 1990s and mid-2000s also coincide with significant trade liberalization in the developing world.

Based on these and other related estimates, researchers have attempted to ascertain the relationship between the size of the informal sector and the stage of a country's economic development, measured, for example, by the per capita GDP. To wit, the aforementioned ILO–WTO (2009) study shows that the average rate of informal employment (informal employment as a share of total employment averaged over all countries in the relevant country grouping) declines from 70% in low-income countries to 63.8% in intermediate-income countries and to 46.1% in high-income countries. This negative relationship is further borne out by La Porta and Shleifer (2014) who groups 185 countries by per capita income quartiles. Using percentage of the labor force in self-employment as a proxy measure of informality they find that the bottom income-quartile countries have 46.4% of their population as self-employed, the second income-quartile has 35.7%, the third income-quartile has 23.1%, and the top income-quartile countries have only 13.3% of self-employed people in the economy. Furthermore, La Porta and Shleifer (2014) estimate that doubling of the GDP for a country is associated with almost a 5% point reduction in self-employment. There is, however, a caveat. If economic growth is accompanied by a faster growth in the labor force then the decline in self-employment will be much slower. Thus, the transition to formality is driven by economic growth as long as the labor force does not grow too fast.

As opposed to viewing informality as an outcome determined by the stage of development of an economy, a number of studies consider informality as a potential contributor to economic growth (e.g., growth in per capita GDP). To this end, Loayza (1996) shows that for Latin American countries there exists a negative relationship between the size of the informal sector and real GDP growth. Loayza (1996) attributes the differential productivity of informal and formal sector workers as a potential reason for this result. Perry et al. (2007) argue that a large informal sector can pose a challenge to a country's public

finance and impede the country's ability to deliver public goods, raise human capital, and facilitate technological change. One potential consequence is a slowdown in economic growth.

These broad stroke summaries conceal a much more nuanced diversity in the features that labor markets with formal and informal employment exhibit. In what follows, we look at three such sets, including labor productivity, response to policies such as the minimum wage, and the impact of trade liberalization on the informal sector.

Productivity of Informal Sector Firms

De Soto (1989) famously views the informal sector as a dynamic sector where much productive potential can be harnessed if government regulations and taxes can be removed and property rights secured. This is in sharp contrast to an alternative view (Farrell, 2004), which advocates the 'hidden dangers' of the informal economy, as the informal sector is seen here as unfairly competing with legal formal sector firms. Hence, policy measures that scale-down the informal sector will be rewarded with overall productivity improvements. A third view characterizes informal sector firms as different from their formal counterparts, employing a vast reservoir of lower-skilled workers, paying correspondingly lower wages, and operating in smaller establishments. La Porta and Shleifer (2014) refer to these three views respectively as the romantic view, the parasite view, and the dual economy view. This third view sees the informal sector as an employment of last resort for the poor, and increases in the formal employment opportunities will decrease informal employment over time.

What are some potential contributors to the productivity differences between the informal and formal sector workers as seen in these distinctive views of the informal sector? We argue that empirical work in this area has not fully accounted for one or more of following issues: unobserved heterogeneity; the possibility of employer market power in the formal and/or the informal sector; and the different shades of informality that already exists within the formal sector.

To gauge the difference in labor productivity between the formal and informal sectors, two approaches are often employed. The first is a direct method which relates productivity to value added per worker (sales net of expenditures on raw materials and energy per worker). The second is an indirect method, which deduces relative productivity from data on wages received by workers in informal firms relative to workers in formal firms. (For example, based on the ratio of value-added definition, the informal productivity gap ranges from 99% (in the Democratic Republic of Congo) to 30% (in Cape Verde) with the median around 85% (in Uganda and Guatemala). Based on the relative productivity definition, on average informal wages are about one-half of small formal firms and one-third of large formal firms (La Porta and Shleifer, 2014, 2008).)

The reliability of the first approach – using the ratio of observed formal and informal value-added – depends critically on available information on the other inputs employed in formal and informal firms. These tend to vary significantly across countries and across inputs. For example, La Porta and Shleifer (2008, 2014) do not find any perceptible differences in the education level of formal and informal sector workers, instead, a huge difference in managerial human capital is found – 76% of managers in formal firms have a college degree compared to only 7% of managers in the informal sector. Other studies, however, find that although skill (or education) levels in the formal and informal sectors overlap, the average skill (or education) level is higher in the formal sector (Maloney, 1999, 2004; Thomas, 1992; Gong and Van Soest, 2002; Gong et al., 2004; ILO–WTO, 2009.). Furthermore, depending on data source, critical data on inputs are oftentimes missing. The World Bank Enterprise Surveys, for example, have data on the use of electricity generators by formal and informal firms, but not the actual physical capital input employed by the firms. In addition, the composition of inputs – whether they are from formal or informal sources – is not often clearly specified.

Specifically, many formal firms hiring a certain percentage of the formal workforce do not have regular contracts that pay social security and other worker benefits. Such a mix of long-term regular and short-term contract workers (through labor contractors, for example) has been observed in Argentina and other Latin American countries (Bosch and Maloney, 2006; Perry et al., 2007); in horticultural firms in South Africa (Barrientos and Kritzinger, 2004); in the garment industry in Jordan; and the fruit industry in Chile (Barrientos, 2008). Some motives that are often associated with the hiring of temporary workers include (1) cost savings; (2) flexibility in the presence of output demand shocks; (3) specialty skills that are not available on regular basis; and (4) the need to screen workers' ability. Empirical result in this area is nascent and includes only developed country studies (Beckmann and Kuhn, 2009). Soundararajan (2014a) is the only exception and finds using data from Indian manufacturing firms that (1) contract worker is a critically important source of labor input for Indian formal manufacturing firms, and (2) the productivity impact of the contract worker intensity is highly nuanced and nonlinear.

The reliability of the second approach – using the ratio of observed formal to informal wages – depends critically on whether the wage-determination mechanisms in the two sectors are comparable. There is a fair amount of evidence of a formal–informal sector wage gap in developing countries. For instance, Bargain and Kwenda (2010) demonstrate the existence of an informal wage penalty using evidence from Brazil, Mexico, and South Africa. Charlot et al. (2011) present evidence from Brazil indicating a wage premium in the formal sector after controlling for individual characteristics. In particular, if labor markets in the formal and informal sectors are perfectly competitive, then observed wages can be expected to be a good proxy for labor productivity. This is of course a strong assumption. To this end, Basu et al. (forthcoming) highlight two features of informal economy – the imperfect enforceability of contracts and the competitiveness of the labor market – and show that the two are in fact incompatible with one another in a model of the formal and informal sectors with search frictions and self-enforcing contracts. The model is consistent with a host of salient observed features of informal labor markets, including wage dualism, and nonmonotonic unemployment response to a minimum wage hike. Thus, without further information about the deviation from marginal productivity pricing, wage differences are in general a poor proxy for formal and informal sector productivity differences.

Minimum Wage Legislations and Informal Work

Standard models of the informal labor market typically aruge that an artificially high minimum wage in the formal sector gives rise to informal sector employment. Within this general framework, there are two classes of models each with different views on the determinant of informal sector employment. Dual' labor market models following Fields (1975) view the informal sector as a by-product of urban unemployment that arises as a result of rural–urban migration generated by the downward rigidity of a legislated minimum wage in the urban sector. This class of models does not view informal enterprises as deliberately avoiding regulations but rather as employers of last resort for otherwise unemployed workers. On the other hand, the 'voluntarist' school views informal sector firms as deliberately avoiding labor market regulations like the minimum wage, social security contributions, and worker benefits that increase labor hiring costs for employers. Some employers, especially in countries with low-enforcement intensities, circumvent these legislations by choosing to operate in the shadow/informal economy (Johnson et al. (1997) and Johnson et al. (1998) as examples).

These models tend to generate very different predictions concerning the impact of a minimum wage hike on (1) subminimum wages, (2) informal sector employment, and (3) unemployment. For some time, these predictions have met with mixed confirmation in the empirical literature, where a number of salient features of informal labor markets have begun to take root. Basu et al. (forthcoming) highlight four such features:

- There is widespread noncompliance to minimum wage legislation both in the formal and informal sectors.
- There is coexistence of compliant and noncompliant employers.
- There are broad ranges of noncompliance, where a spike at the official minimum, typically occur alongside a dispersion of subminimum wages in covered sectors.
- Observed unemployment response to the minimum age is mixed ranging from positive, to insignificantly different from zero, to negative.

These observations challenge both the theoretical and empirical literature on informal work in a number of important ways. First, in terms of theory, these observations call for a unified theoretical model that is capable of reconciling these diverse empirical findings. Basu et al. (forthcoming) take a first stab and write a model combining an imperfectly competitive labor market, an imperfectly enforced minimum wage, and imperfect commitment to the minimum wage policy by policy makers. In doing so, the paper questions why authorities turn a blind eye to the very legislation that they have themselves passed. The model generates an inverted-U relationship between changes in the minimum wage and employment. It explains the selection into compliant and noncompliant firms among a continuum of heterogeneous firms in terms of productivity. The model also explains the spike at the minimum wage as the marginal cost of labor exhibits a discrete jump at the legislated minimum wage. Finally, the model explains how in the presence of imperfect enforcement, employment response to a minimum wage hike can remain the same as the setting with perfect enforcement, exhibit a muted response, or exhibit the opposite sign.

Soundararajan (2014b) presents the first empirical study to formally test the above results. Using a repeated cross-sectional data set from the nationally representative employment surveys of India for the years 2004–12, the study estimates the interactive effect of minimum wages and enforcement on employment, wages, and days of work in the construction industry – the second largest industry in India after agriculture. The study finds an inverted-U relationship between minimum wage and employment after controlling for imperfect enforcement of the minimum wage.

Many interesting areas for future research in the area remain. For example, labor market regulations are not the only regulations faced by firms that operate in the formal sector of developing countries. Indeed, in addition to minimum wages, formal firms are often obliged to pay government authorities a tax on profits earned and withhold taxes on wage income to be transferred directly to authorities. By comparison, informal firms and informal workers are free from both the state-mandated regulation on wages, the need to pay taxes on profits or wages, and operate in an environment where entry and exit may be seen as more or less cost-free. In the presence of multiple regulations imperfectly enforced by different agencies (tax auditors and labor inspectors), Basu et al. (2012) set out to answer a number of additional questions: (1) how do profit tax, personal tax, and the intensity of tax enforcement impact tax evasion and minimum wage noncompliance? (2) How do a minimum wage hike and stricter enforcement of the minimum wage impact the extent of either of the profit or personal tax evasion or exit by formal firms? It is shown that tax evasion is increasing with the profit tax rate and the minimum wage, and decreasing in the personal income tax rate. Meanwhile, noncompliance with minimum wage on the other hand is independent of the profit tax rate and increasing in the personal income tax rate. The next set of open issues is accordingly an empirical investigation into the role of multiple policies (e.g., minimum wage and taxes) on the multiple types of informal activities (e.g., tax evasion, informal employment).

Trade Liberalization and the Informal Sector

The national welfare implication of trade liberalization is a subject of longstanding importance in the economics literature. Importantly, in a whole range of possible settings including neoclassical trade models featuring perfect competition in input and output markets (Chau and Färe, 2011), or the latest class of models incorporating heterogeneous firms (Melitz, 2003), trade liberalization is expected to bring unambiguous welfare gains. The former operates through more efficient allocation or resources as trade barriers are removed, while the latter operates through the exit of lower productivity firms as trade liberalization promotes competition between domestic and foreign firms.

None of these settings, however, incorporate possible complications when the informal sector is included in the analysis. Despite the importance of the informal sector in developing countries, to date there are in fact only a handful of empirical and theoretical studies on the subject. Already, however, what we have learned suggest that the welfare implications of trade liberalization are highly nuanced as soon as considerations are given to the informal sector.

Consider, therefore, two popular indicators of trade liberalization – trade openness (sum of exports and imports) as a share of GDP and average foreign direct investment liabilities. ILO–WTO (2009) demonstrates that among five Asian countries (China, India, Indonesia, Thailand, and Sri Lanka), informal employment in countries with low-, intermediate-, and high-trade openness are 80%, 74.6%, and 72.4%, respectively. This negative relationship is preserved with a larger set of countries and also when openness is measured based on average foreign direct investment liabilities. The reasons for these differences are often seen to stem from the exit of least-productive informal firms as a consequence of trade liberalization, particularly when a productivity gap between formal and informal sector firms prevails (La Porta and Shleifer, 2008; Hsieh and Klenow, 2009; McCaig and Pavcnik, 2013). Fugazza and Fiess (2010) likewise produce mixed findings and show that the results depend in particular on whether the empirical analysis employs cross-country or cross-firm evidence. Complementing these findings relating informal sector employment with trade liberalization, Soares (2004) provides evidence on the impact of trade reforms on wages in the informal sector in Brazil between the 1980s and 1990s. The results suggest a reduction in the wage gap between formal and informal sector workers driven by a fall in the very high wages enjoyed by formal sector workers prereform. By contrast, Bargain and Kwenda (2011) find a wage penalty for lower-tier informal sector workers vis-à-vis the formal sector counterparts in Brazil over 2002–07 time period and South Africa over the 2005–08 time period.

These and other related empirical evidence to date notwithstanding, there continues to be very little theoretical developments in the area to organize these results. An exception is Becker (2014), who proposes a trade model with heterogeneous firms incorporating the informal sector, wage dispersion due to productivity heterogeneity, and a concern for fair wages on the part of workers (Akerlof and Yellen, 1990). The model is uniquely suited to analyze firm exits out of and entry into exporting, the formal sector and the informal sector subsequent to trade liberalization. The model produces findings consistent with Nataraj (2011), for example, which shows that the productivity improvement observed subsequent to the Indian trade liberalization in 1991 arise in large part due to the exit of the least-productive informal firms. Furthermore, the model also generates results on wage responses to trade liberalization.

Conclusion

Despite the many new advances summarized here in studies related to informal work in developing countries, many issues remain. For example, the informalization of formal sector work, such as through the subcontracting of formal work, is an area of research that has received very little attention due both to the paucity of data and the lack of a theoretical basis with testable hypothesis, and associated policy prescriptions. The appropriateness of a minimum wage policy depends both on the extent to which the minimum wage is enforced and the extent to which employers exhibit market power. In both of these areas, more empirical observations in a wider range of countries are needed. Finally, ongoing work on the role of trade liberalization on informality has focused almost exclusively on goods-trade liberalization, while the impact of an influx of foreign investment, and as such foreign competition for domestic labor in formal and informal sectors remains largely an open question.

See also: Agriculture, Economics of; Development Studies; Development and Employment: A Mutual Relationship; Informal and Underground Economics; Minimum Wages; Shadow Work: Measurement.

Bibliography

Akerlof, George A., Yellen, Janet L., 1990. The fair wage-effort hypothesis and unemployment. Quarterly Journal of Economics 105, 255–283.

Bargain, Olivier, Kwenda, Prudence, 2010. Is Informality Bad? Evidence from Brazil, Mexico and South Africa. IZA Discussion Papers 4711, Institute for the Study of Labor (IZA).

Bargain, O., Kwenda, P., 2011. Earnings structures, informal employment, and self-employment: new evidence from Brazil, Mexico, and South Africa. Review of Income and Wealth 57, 100–122.

Barrientos, Stephanie, 2008. Contractor labour: the achilles heel of corporate codes in commercial values chains. Development and Change 39, 977–990.

Barrientos, S., Kritzinger, A., 2004. Squaring the circle – global production and the informalisation of work in South African fruit exports. Journal of International Development 16, 81–92.

Basu, Arnab K., Chau, Nancy H., Kanbur, Ravi, 2010. Turning a blind eye: costly enforcement, credible commitment and minimum wage laws. The Economic Journal 120, 244–269.

Basu, Arnab K., Chau, Nancy H., Kanbur, Ravi, 2011. Contractual dualism, market power and informality. IZA Discussion papers 5845, Institute for the study of labor (IZA), forthcoming in The Economic Journal.

Basu, Arnab K., Chau, Nancy H., Siddique, Zara, 2012. Tax evasion, minimum wage non-compliance and informality. Research in Labor Economics 34, 1–53.

Batini, Nicoletta, Kim, Young-Bae, Levine, Paul, Lotti, Emanuela, 2010. Informal Labour and Credit Markets: A Survey. IMF Working Paper WP/10/42, Washington, DC.

Becker, Dennis, 2014. Heterogeneous Firms and Informality: The Effects of Trade Liberalization on Labor Markets. Working Paper WP 2014-13, Charles H. Dyson School of Applied Economics and Management, Cornell University, Ithaca, NY.

Beckmann, M., Kuhn, D., 2009. Temporary Agency Work and Firm Performance: Evidence from German Establishment-Level Panel Data. Discussion Paper No. 09–17, German Economic Association of Business Administration, Vallendar.

Bosch, M., Maloney, W., 2006. Gross Worker Flows in the Presence of Informal Labor Markets. The Mexican Experience 1987–2002. Mimeo, The World Bank.

Charlot, Olivier, Malherbet, Franck, Terra, Cristina, 2011. Product Market Regulation, Firm Size, Unemployment and Informality in Developing Economies. IZA Discussion Papers 5519.

Chau, Nancy H., Färe, Rolf, 2011. Shadow pricing market access: a trade benefit function approach. Journal of Economic Theory 146 (4), 1631–1663.

Chen, Martha, 2012. The Informal Economy: Definitions, Theories and Policies. Women in Informal Employment Globalizing and Organizing (WEIGO) Working Paper No 1. www.weigo.org.

De Soto, Hernando, 1989. The Other Path: The Invisible Revolution in the Third World. Harper and Row, New York.

Farrell, Diana, 2004. The hidden dangers of the informal economy. McKinsey Quarterly 3, 27–37.

Fields, Gary, 1975. Rural-urban migration, urban unemployment and underemployment, and job-search activity in LDCs. Journal of Development Economics 165–187.

Fugazza, Marco, Fiess, Norbert, 2010. Trade Liberalization And Informality: New Stylized Facts. UNCTAD Blue Series Papers 43, United Nations Conference on Trade and Development.

Gong, Xiaodong, Van Soest, Arthur, 2002. Wage differentials and mobility in the urban labour market: a panel data analysis for Mexico. Labour Economics 9 (4), 513–529.

Gong, Xiaodong, Van Soest, Arthur, Villagomez, Elizabeth, 2004. Mobility in the urban labor market: a panel data analysis for Mexico. Economic Development and Cultural Change 53 (1), 1–36.

Hsieh, Chang Tai, Klenow, Peter J., 2009. Misallocation and manufacturing TFP in China and India. Quarterly Journal of Economics 124 (4), 1403–1448.

International Labour Organization (ILO), World Trade Organization, 2009. Globalization and Informal Jobs in Developing Countries (Prepared by Marc Bacchetta, Ekkehard Ernst, Juana P. Bustamante. Geneva).

Johnson, Simon, Kaufmann, Daniel, Shleifer, Andrei, Fall 1997. The Unofficial Economy in Transition. Brookings Papers on Economic Activity, pp. 159–240.

Johnson, Simon, Kaufmann, Daniel, Zoido-Lobaton, Pablo, 1998. Regulatory discretion and the unofficial economy. American Economic Review 88 (2), 387–392.

Loayza, N.V., 1996. The Economics of the Informal Sector: A Simple Model and Some Empirical Evidence from Latin America. Policy Research Working Paper 1727. World Bank, Washington, DC.

Maloney, William F., 1999. Does informality imply segmentation in urban labor markets? Evidence from sectoral transitions in Mexico. The World Bank Economic Review 13, 275–302.

Maloney, William F., July 2004. Informality revisited. World Development, Elsevier 32 (7), 1159–1178.

McCaig, Brian, Pavcnik, Nina, 2013. Moving Out of Agriculture: Structural Change in Vietnam. Working Paper 19616. National Bureau of Economic Research.

Melitz, Mark J., 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. Econometrica 71 (6), 1695–1725.

Norbert, Fiess, Fugazza, Marco, 2010. Trade Liberalization and Informality: New Stylized Facts. Policy Issues in International Trade and Commodities Study Series No. 43. UNCTAD, Geneva.

Nataraj, Shanthi, 2011. The impact of trade liberalization on productivity: evidence from India's formal and informal manufacturing sectors. Journal of International Economics 85 (2), 292–301.

OECD, 2002. Measuring the Non-observed Economy: A Handbook. OECD, Paris.

OECD, 2009a. Is Informal Normal? Towards More and Better Jobs. OECD, Paris.

OECD, 2009b. Policy Round Tables, Competition Policy and the Informal Economy. OECD, Paris.

La Porta, Rafael, Shleifer, Andrei, 2014. Informality and development. Journal of Economic Perspectives 28 (3), 109–126.

La Porta, Robert, Shleifer, Andrea, 2008. The unofficial economy and economic development. Brookings Papers on Economic Activity 275–352.

Perry, Guillermo E., Maloney, William F., Arias, Omar S., Fajnzylber, Pablo, Mason, Andrew D., Saavedra-Chanduvi, Jaime, 2007. Informality : Exit and Exclusion. World Bank, Washington, DC.

Schneider, Friedrich, 2014. The Shadow Economy and Shadow Labor Force: A Survey of Recent Developments. IZA Discussion Paper No. 8278, Bonn, Germany.

Soares, Fabio Veras, 2004. Some stylized facts of the informal sector in Brazil in the past two decades. In: Proceedings of the 32th Brazilian Economics Meeting.

Soundararajan, Vidhya, 2014a. Contract Workers and Firm Productivity in the Indian Manufacturing Sector. Mimeo, Charles H. Dyson School of Applied Economics and Management, Cornell University, Ithaca, NY.

Soundararajan, Vidhya, 2014b. Minimum Wage Effects at Different Enforcement Levels: Evidence from Employment Surveys in India. Working Paper WP 2014–12, Charles H. Dyson School of Applied Economics and Management, Cornell University, Ithaca, NY.

Strobl, Eric, Walsh, Frank, 2001. Minimum wage and compliance: the case of Trinidad and Tobago. Economic Development and Cultural Change 51 (2), 427–450.

Thomas, Jim J., 1992. Informal Economic Activity. LSE Handbooks in Economics. The University of Michigan Press.

Information and Communications Technology in Education

Saïd Assar, Telecom Ecole de Managment, Institut Mines-Telecom, Evry, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

As Information and Communication Technologies (ICT) gradually permeate daily life, they are profoundly changing the way education is conceived and delivered. Teachers play a key role in this transformation process; their beliefs, pedagogical practices, and teaching skills are continuously challenged. ICT integration in the educational process can significantly enhance traditional courses and Internet-based education, i.e., e-learning, is becoming a serious alternative to traditional, face-to-face courses. To be used as a lever for pedagogical innovation and institutional transformation, teacher ICT competencies need to go beyond skills in ICT use per se, and enclose contextual knowledge about technology, pedagogy, and content. On the other hand, learners' engagement with ICT in education depends on their expectations and conceptions of learning and required assessment. A learner's experience with ICT in education is linked to his perception of systems' ease of use and usefulness in achieving learning goals. Adaptive learning systems open new potentialities for a personalized instruction which is tailored to the learner's characteristics. At the edge of a new era, schools, colleges, and higher education struggle to seize opportunities and overcome obstacles.

Introduction

Education is existential for human societies; it is a fundamental leverage for social preservation and economic growth. For centuries, the way education is performed remained the same: the teacher talks and acts, while pupils listen, watch, and write. Printing machines made textbooks widely available and knowledge rapidly spread; educational institutions appeared and developed all over the world. However, the way education was performed did not change so much and remained teacher centered. With the emergence of new technological devices for handling picture and sound, new possibilities for knowledge transmission emerged. Instead of handmade drawings, photos could be used to present study subjects, for example, in biology or physics courses. Using audio recordings, pupils learning foreign languages could, at their convenience, listen to native speakers and practice their lessons. With animated pictures and movies, a chemistry course could be illustrated with the visual transcription of a complex experiment. Television took education a step further as lectures and seminars could be broadcasted. The *Gyan Darshan* educational TV channel in India, for example, aired on a large-scale programs originating from institutions all over the country for the benefits of millions of Indian students (Rani, 2006). However, it is the rise of computers and personal computers – combined with the rapid expansion of Internet – that opened a wide area of new applications and usages in education.

Fast development in Information and Communication Technologies (ICT) is shaping a new world in which education at all levels can no longer be assimilated to a group of learners in a classroom listening and watching a teacher with a textbook following a fixed curriculum. With ICT, learning is shifting from teacher centered to learner centered and can potentially be undertaken anytime and anywhere, from classrooms to homes and offices. E-learning refers to learning via the Internet; it provides learners with a flexible and personalized way to learn (Zhang and Nunamaker, 2003). It offers learning-on-demand opportunities and can significantly contribute to reduce teaching and learning costs. E-learning is the enabler for a massive transformation in the education world; new teaching and learning opportunities are continuously challenging traditional schools, colleges, and higher education (Oliver, 2002; Collins and Halverson, 2009). However, effective and innovative use and integration of ICT in education is a complex and multifaceted problem. Complexity lies in the intertwining of technology, pedagogy, user adoption, and institutional policies (John and Sutherland, 2004). ICT encompasses many different things, and can address multitude challenges; in fact, to teach and to learn with ICT remains partially understood, and all its benefits are still not fully exploited (Cuban, 2001; Kirkwood, 2009). Furthermore, from primary schools to higher education, scientific evidence is still insufficiently available about the effectiveness of ICT integration in education (U.S. Dept. of Education, 2010).

Digital Resources for Education

There are three general and complementary categories of ICT artifacts to be used in teaching.

- The first category is stand-alone digital components to be used to support learning inside a course for didactic and/or assessment purposes. Called also Digital Learning Material (DLM) or Learning Objects (Polsani, 2006), it includes video clips (e.g., a YouTube fragment), illustrations (e.g., photos and drawings), simulations (e.g., simulation of an organizational process or an electronic circuit), and interactive assessment resources (e.g., quiz) (Kreijns et al., 2013).
- The second category includes general tools for communication and information diffusion such as Really Simple Syndication (RSS), blogs, chat and voice over IP, peer-to-peer file sharing, wikis, web conferencing, and social networks. These tools were not specifically designed for teaching; however, they can support the teaching purpose by providing communication facilities in pedagogical scenarios.

- The third category resembles software systems to run a course. Called also Learning (or Course) Management System (LMS or CMS), these complex tools are educational software packages for online course administration and delivery (Morgan, 2003). Their main functions include course content organization and presentation, students recording and tracking, class activities management, teachers and learners' communication, student assessment tools, and grade books tools. Other terms used to denote such systems are Web-Based Learning Systems (WBLS), Virtual Learning Environments (VLE) or, more simply, learning platforms. Blackboard and Moodle are common examples of learning platforms in higher education (see Web references).

Simple learning objects (e.g., a video sequence or a quiz) can be composed into more complex educational resources and made available on their own in the form of a CD-ROM or interactively online. They can also be part of complex software applications dedicated to providing learners with specific assistance while learning a particular subject, e.g., mathematics (cf *see* Computer Assisted Instruction). For primary and secondary school, simple and composite learning objects are a common way to integrate ICT in teaching. In higher education, this approach is particularly common as well: learning objects are integrated in face-to-face lectures; they can also be part of courses provided online through a learning platform. Learning platforms can be used either for blended learning where face-to-face teaching alternates with and complements online teaching, or for fully online learning where no face-to-face interaction occurs (i.e., virtual learning). In both cases, different communication and information diffusion tools (e.g., blog, wiki) can be used in combination with the learning platform.

Teachers and ICT

Teachers are critical change agents at the academic work floor; they are the instrument by which changes in education will become true. In an extensive literature review, Mumtaz (2000) summarized what influence teachers' attitude toward technology use in school: available and easy-to-use digital resources, incentives to change and support from colleagues and school managers, clear and understandable school and national polices, and background in formal computer training. A worldwide educational survey at primary and secondary education levels confirmed these early indications. Pelgrum (2001) assessed ICT integration in teaching and ranked three factors as most significant barriers: (1) computers insufficiently available, (2) teachers' lack of ICT knowledge and skills, and (3) difficulties in integrating ICT in instruction in a relevant manner. Furthermore, teachers tend to ignore the full potential of ICT to power the learning environment. For example, in the Netherlands, a survey in the highest grade primary schools found that 60–80% of teachers in the sample never used CD-ROM encyclopedias and information search on the Web (Smeets, 2005). In another survey, in higher educational institutions, 51% of teachers considered that the quality of students' learning in their course did not improve using Internet; and 27% think that e-learning environments have no added value for their course (Mahdizadeh et al., 2008: p. 148).

Nonetheless, teachers' adoption of ICT use in teaching cannot be restricted to merely technology-related factors: "integration of educational computer use in professional competencies of teachers implies a more complex approach" (Tondeur et al., 2008: p. 2542). Beyond factors related to technology resource and education management, teachers' beliefs about teaching and learning with ICT are essential and need to be fully taken into account (Mumtaz, 2000; Cloke and Sharif, 2001). Teachers' beliefs are understandings, premises, or propositions felt to be true about the educational purpose and the educational process. They are rooted in teachers' perception of their role and mission as knowledge owners and knowledge transmitters. Established typologies distinguish between 'traditionalistic' or behaviorist, teacher-centered, and more 'progressive' or constructivist, student-centered beliefs (Tondeur et al., 2008: p. 2544). These two types are not exclusive; research in the last decades on the dimensionality of educational beliefs has acknowledged a multidimensional vision of the belief system. For example, Tondeur et al. (2008) found that most frequent adoption of all types of computer uses is amid teachers with relatively high constructivist beliefs and also high traditionalist beliefs.

To make teachers adopt a more constructivist perspective, they primarily need to gain a broader and deeper understanding of what is expected from ICT integration in class. They also need to acquire higher self-confidence through improving their ICT self-efficacy and their awareness of ICT potential. Furthermore, encouraging teachers' experimental behavior and training them with pedagogy-oriented ICT skills can be strong determinants of ICT adoption in teaching (Kreijns et al., 2013). In fact, the knowledge and skills that teachers need to acquire will differ depending on the content that is taught and the pedagogical goal. This can vary from improving learning effectiveness in school subjects to promoting the development of specific skills such as lifelong learning and learning to learn.

Pedagogical Issues

ICT alone does not improve teaching; rather it is the ways in which ICT is incorporated into the various learning activities that is of crucial importance (John and Sutherland, 2004). This will depend on the teacher and on his pedagogical approach in presenting content knowledge. Shulman (1987: p. 8) defines pedagogical content knowledge as "the blending of content and pedagogy into an understanding of how particular topics, problems, or issues are organized, represented and adapted to the diverse interests and abilities of learners, and presented for instruction." The challenge with ICT lies in exploiting its vast possibilities to incorporate adequate digital resources – created from scratch or reused from existing material – into his teaching process in order to support the pedagogical content knowledge he intends to present. Learning theories can be helpful in guiding this process to build ICT-based curriculums (Ally, 2008).

To illustrate the contribution of learning theories in shaping ICT usage, Mödritscher (2006) conducted an experiment in

which he designed a virtual course (i.e., fully online) according to three well-known theories: behaviorism, cognitivism, and constructivism (cf *see* Learning Theories and Educational Paradigms). These three courses had identical knowledge goal and content, but pedagogically, were developed differently. The first (i.e., behaviorism) was structured as three modules provided sequentially to the students, and the learning process was assessed by typical behaviorist elements like multiple-choice questions. The second (i.e., cognitivism) course was organized in two phases: first, three groups of four students each worked on the course's objectives; second, the groups (dispatched differently) had to restructure the result of the first phase and synthesize the content in a wiki environment. Assessment was done by grading the result of each phase based on quality and quantity. The third course (i.e., constructivism) was organized in such a manner to lead students to build the knowledge content collectively. The four groups were given all material, and the task was to create a document for mediating the course's learning objectives to colleagues. The three members of each group had then to compare the works of the other groups and evaluate them. Results showed that each of the three courses varied in various aspects such as learners' and teacher's effort and effectiveness. The behaviorist and constructivist approaches showed better results in teaching effectiveness and earned better rating from students. Furthermore, the best efficiency in knowledge transfer was obtained through those tasks that students had to complete on their own (Mödritscher, 2006).

Beyond pedagogical approaches to ICT integration in teaching, there is a serious issue that arises: is it teaching 'with ICT' or teaching 'about ICT?' When teaching, for example, a basic statistical technique using a spreadsheet (e.g., Excel), there is always a risk that the focus is on the tool and not on the subject. According to Selinger (2001), we need to teach ICT skills in ways that "enable students to perceive the benefits and potential of using computers to support their work" (p. 144). Because it is difficult to teach ICT tools without their usage contexts, both should be taught simultaneously: the content material with the help of ICT (e.g., accounting), and effective ICT skills (e.g., how to use Excel). However, among teachers, ICT integration in teaching practices is still difficult to understand. For example, in a survey among primary school teachers in Tondeur et al. (2007) found that, although the governmental ICT curriculum centers on the integrated use of ICT within the learning and teaching process, "teachers in primary education still stress to a large extent technical ICT skills" (p. 972). Teachers need to better understand how ICT can be integrated in teaching, and what kind of skills they require (cf *see* Teacher Education).

A Framework for Teacher Knowledge

The Technology, Pedagogy, and Content Knowledge (TPACK) framework is a proposal to develop teacher knowledge in the ICT-based education era (Harris et al., 2009). It is intended to help teachers successfully integrate technology into their teaching practice. It is based on a recurring observation: "different disciplines have differing organizational frameworks, established practices, ways of acknowledging evidence and proof, and approaches for developing knowledge" (Harris et al., 2009: p. 395). It claims that it is insufficient to expose teachers to particular technological tools and resources and their possible curriculum-based educational use. There are complex relationships among content, pedagogy, technology, and context that teachers need to understand and conceptualize in order to put these technological resources into practice. The TPACK framework emphasizes "the connections among technologies, curriculum content, and specific pedagogical approaches"; it is expected to demonstrate "how teachers' understandings of technology, pedagogy, and content can interact with one another to produce effective discipline-based teaching with educational technologies" (Harris et al., 2009: p. 396).

Activity types are the building concept in the TPACK framework. An activity type is a content-oriented teaching pattern which conceptualize notions of pedagogical content knowledge mentioned before (Shulman, 1986, 1987). It acknowledges the need for particular pedagogical techniques when teaching a particular content. It is a structured representation of pedagogical techniques that can be associated with classroom activities, interactions, and recognizable material for a particular content. For example, one of the first proposals of such teaching pattern mentioned in educational literature is the Mehan (1979) I–R–E sequence: teacher Initiation, student Reply, teacher Evaluation (cited in Harris et al., 2009: p. 404). Other well-known pedagogical techniques, such as 'to have a discussion,' 'to brainstorm an issue' or 'to make an experiment,' can be further refined in order to form activity types. However, because of the interdependency between pedagogy and content, an activity type has to be defined in the context of a specific teaching domain. 'To make an experiment,' for example, would be defined differently depending on the content: in a chemistry course, it would mean going to the laboratory and manipulating material and measurement tools; in a foreign language course, it would mean composing linguistic constructs and evaluating them.

Concretely, for a given content, the TPACK framework gradually and incrementally matches relevant activity types with adequate technological resources (see Web references). For example, in the mathematics domain, a typical interpretation activity type is to 'develop an argument.' Candidate technologies to support this activity would be concept mapping software, presentation software, and specialized word processing tools (Grandgenett et al., 2011: p. 3). Another possible activity in teaching mathematics is to 'describe an object or concept mathematically'; candidate technologies would be specialized tools such as Logo graphics, engineering visualization software, or Mathematica software package (Grandgenett et al., 2011: p. 4).

Learners and ICT

Learner acceptance is a key issue when using ICT in education. It will depend on two sets of intertwined factors: the first concerns ICT's role in the educational process and how it contributes – directly or indirectly – to better performance, the second relates to the learner's own experience while using ICT for educational purposes. For example, when instructors make

available to students recorded lectures as audio and video podcasts to download, learners' attitudes and opinions will depend primarily on the extent to which this facility can contribute to successful grading in the course. In turn, this will depend on podcasts' content, and how it complements – or supplements – face-to-face lectures. Beyond traditional lectures, availability as podcasts of short revision summaries were, for example, highly appreciated by students (Copley, 2007). On the other hand, learners' attitudes will also depend on their experience in downloading, storing, and using the audio and video digital files. File size, file format, text size, and video resolution then play a key role; in addition, the ease of use and flexibility of podcasts will contribute to learner satisfaction.

Let's illustrate how multiple factors intertwine in shaping learners' perception according to educational contexts and scenarios. Kay and Knaack (2008) report a large survey to better understand how teachers and students perceive learning objects in secondary schools. Although students were less positive than teachers, they rated positively visual support, ease of use, and animation with learning objects; they also felt better engaged with the course. Negative remarks mentioned inadequate level of details in certain graphics and the use of different wording and explanation methods than those used in class. Kirkwood (2008) investigated how undergraduate learners use online resources. He finds that most learners use the internet for "personal, domestic, social and employment purposes as well as for educational goals" (p. 381), but seeking information resources is most frequent when it is directly useful for their ongoing studies, particularly in relation with assessment. He insists however on the importance of adequate ICT literacy in terms of copyright and plagiarism issues, identifiers and passwords management, and protection against malicious software.

Nonetheless, it is the student's perception of distance education through the Internet (i.e., e-learning) that received highest attention. Researchers regularly investigate the extent to which e-learning is appropriate for all students and is an effective manner to provide education. In 2004, Allen et al. analyzed previous empirical studies; results showed that distance learning slightly outperformed face-to-face classes on the basis of performance (i.e., ability to master content and skills). The highest improvement was for teaching specific contents, e.g., foreign languages (Allen et al., 2004: p. 413). Concannon et al. (2005) surveyed students in an undergraduate blended course – lectures and laboratory sessions were supplemented with online course content. They found that e-learning adoption depended on the learner's general attitude and skills with computers, awareness of online resources availability, peer influence, and instructor support. Complaints concerned difficulties in navigating through the resources and technical problems in accessing the platform. Hence, authors issue some tips to ease students' use of learning platforms: email notification when new material is online (students often did not notice that new material was published on the platform), high-quality tutorials on how to use the system, and feedback on students' questions along with comments. Globally, although students "experienced a real improvement in the quality of their education experience" (p. 511) and considered ICT as a valuable support to the learning process, 81% of them claimed "traditional lectures and tutorial groups result in more effective student learning than a pure e-learning environment" (p. 508). A more fine-grained study of students' perception of e-learning showed that students favored face-to-face learning over online learning when confronted with acquisition of certain specific knowledge and skills, e.g., "conceptual knowledge in the subject matter, skills in the practical application of one's knowledge, knowledge and skills in using scientific work routines, or in communication" (Paechter and Maier, 2010: p. 296). Another recent study found comparable results when analyzing students' attendance (Larkin, 2010): "contrary to popular belief, Generation Y students in general, do not aspire to replace lectures with downloadable, online versions" in fact, they "valued the opportunity for interactive learning provided by face to face teaching" (Larkin, 2010: p. 238).

Indeed, the actual tendency is to align learner's adoption models with general models of technology adoption, i.e., Technology Adoption Model (TAM) (Lee et al., 2003). Hence, perceived system quality, e.g., platform ease of use and learner's interface user-friendliness, is the first global factor that directly impacts learners' adoption (Chang and Tung, 2008; Shee and Wang, 2008). The second factor measures the global usefulness of technology-based educational process for attaining learners' goals. Paechter et al. (2010) found "achievement goals proved to be more important than other course characteristics, e.g., the design of the learning material or the user friendliness of the learning platform" (p. 227). Thus, they suggest that instructors should increase learners' motivation by adapting instruction accordingly, e.g., clarifying learning objectives, providing self-tests to assess progress all along the course.

To conclude this section, it is worthwhile to mention adaptation and personalization techniques. As learning is a complex and sophisticated process in which individual traits and characteristics can interfere, adaptive learning systems seek to adjust the content, the appearance, or the process to learner's knowledge level, goals, and other characteristics (Papanikolaou et al., 2002). The aspiration is to break the 'one size fits all' educational paradigm and to provide personalized content, to protect learner from cognitive overload and disorientation, and to better assist learner through the learning process. The basic ideas behind such approaches is to classify learner according to a certain user model, and then match learner profile with adequate content, representation, or process (De Bra et al., 2009). A basic and widely known user model is prior knowledge: students are classified into groups that have similar ability and knowledge. Adaptation (i.e., automatic personalization) in such case is relatively straightforward: a quiz or a test determines learner level, and accordingly, directs learner to specific learning material. However, most of the existing research on adaptive learning is devoted to model learner according to his learning preferences, i.e., learning style. Learning style is the person's preferred way to learn and the way he/she learns best. More complex definitions include cognitive characteristics, in addition to affective and psychological behaviors (cf *see* Cognitive Styles and Learning Styles). Adequately tailoring learning material to student learner style is a complex and open problem, and actual results seem – for the moment – contradictory and controversial (Brown et al., 2009).

However, it is widely acknowledged that personalization and adaptability in technology-enhanced learning will have enormous potential for improving the user experience.

Conclusion

ICT is no doubt profoundly transforming education. In order to fully benefit from this ongoing revolution, academics need to reassess their own beliefs and pedagogical practices concerning teaching and assessment and their potential impact on learners' experiences. Both teachers and learners need to understand how ICT should be integrated in education, why e-learning activities are to be undertaken, and what expected rewards can be derived. ICT literacy for education, i.e., necessary ICT skills and knowledge for educational purposes, at individual and organizational levels, is the foundation for modern societies to reinvent teaching and learning in the age of technology.

See also: Cognitive Styles and Learning Styles; Computer-Assisted Instruction; Learning Theories and Educational Paradigms; Teacher Education.

Bibliography

Allen, M., Mabry, E., Mattrey, M., Bourhis, J., Titsworth, S., Burrell, N., 2004. Evaluating the effectiveness of distance learning: a comparison using meta-analysis. Journal of Communication 54 (3), 402–420.

Ally, M., 2008. Foundations of educational theory for online learning. In: Anderson, T. (Ed.), Theory and Practice of Online Learning, second ed. Athabasca University Press, pp. 15–44.

Brown, E.J., Brailsford, T.J., Fisher, T., Moore, A., 2009. Evaluating learning style personalization in adaptive systems: quantitative methods and approaches. IEEE Transactions on Learning Technologies 2 (1), 10–22.

Chang, S.-C., Tung, F.-C., 2008. An empirical investigation of students' behavioural intentions to use the online learning course websites. British Journal of Educational Technology 39 (1), 71–83.

Cloke, C., Sharif, S., 2001. Why use information and communications technology? Some theoretical and practical issues. Journal of Information Technology for Teacher Education 10 (1–2), 7–18.

Collins, A., Halverson, R.R., 2009. Rethinking Education in the Age of Technology: The Digital Revolution and Schooling in America. Teachers' College Press, New York.

Concannon, F., Flynn, A., Campbell, M., 2005. What campus-based students think about the quality and benefits of e-learning. British Journal of Educational Technology 36 (3), 501–512.

Copley, J., 2007. Audio and video podcasts of lectures for campus-based students: production and evaluation of student use. Innovations in Education and Teaching International 44 (4), 387–399.

Cuban, L., 2001. Oversold and Underused: Computers in the Classroom. Harvard University Press.

De Bra, P., Kay, J., Weibelzahl, S., 2009. Introduction to the special issue on personalization. IEEE Transactions on Learning Technologies 2 (1), 1–2.

Grandgenett, N., Harris, J., Hofer, M., 2011. Mathematics Learning Activity Types. School of Education, College of William and Mary. Learning Activity Types Wiki. Retrieved from: http://activitytypes.wmwikis.net/Mathematics.

Harris, J., Mishra, P., Koehler, M., 2009. Teachers' technological pedagogical content knowledge and learning activity types: curriculum-based technology integration reframed. Journal of Research on Technology in Education 41 (4), 393–416.

John, P.D., Sutherland, R., 2004. Teaching and learning with ICT: new technology, new pedagogy? Education, Communication & Information 4 (1), 101–107.

Kay, R.H., Knaack, L., 2008. An examination of the impact of learning objects in secondary school. Journal of Computer Assisted Learning 24 (6), 447–461.

Kirkwood, A., 2008. Getting it from the web: why and how online resources are used by independent undergraduate learners. Journal of Computer Assisted Learning 24 (5), 372–382.

Kirkwood, A., 2009. E-learning: you don't always get what you hope for. Technology, Pedagogy and Education 18 (2), 107–121.

Kreijns, K., Van Acker, F., Vermeulen, M., Van Buuren, H., 2013. What stimulates teachers to integrate ICT in their pedagogical practices? The use of digital learning materials in education. Computers in Human Behavior 29 (1), 217–225.

Larkin, H.E., 2010. "But they won't come to lectures…" the impact of audio recorded lectures on student experience and attendance. Australasian Journal of Educational Technology 26 (2), 238–249.

Lee, Y., Kozar, K.A., Larsen, K.R.T., 2003. The technology acceptance model: past, present, and future. Communications of the AIS 12 (1), 53.

Mahdizadeh, H., Biemans, H., Mulder, M., 2008. Determining factors of the use of e-learning environments by university teachers. Computers & Education 51 (1), 142–154.

Mehan, H., 1979. Learning Lessons: Social Organization in the Classroom. Harvard University Press, Cambridge, MA.

Mödritscher, F., 2006. E-learning theories in practice: a comparison of three methods. Journal of Universal Science and Technology of Learning 3–18.

Morgan, G., 2003. Faculty Use of Course Management Systems, vol. 2. EDUCASUE Center for Applied Research p. 94. Retrieved from: http://www.educause.edu/ir/library/pdf/ers0302/rs/ers0302w.pdf.

Mumtaz, S., 2000. Factors affecting teachers' use of information and communications technology: a review of the literature. Journal of Information Technology for Teacher Education 9 (3), 319–342.

Oliver, R., 2002. The Role of ICT in Higher Education for the 21st Century: ICT as a Change Agent for Education. HE21 Conference. Curtin University of Technology, Sarawak Campus, Miri, Malaysia.

Paechter, M., Maier, B., 2010. Online or face-to-face? Students' experiences and preferences in e-learning. The Internet and Higher Education 13 (4), 292–297.

Paechter, M., Maier, B., Macher, D., 2010. Students' expectations of, and experiences in e-learning: their relation to learning achievements and course satisfaction. Computers & Education 54 (1), 222–229.

Papanikolaou, K.A., Grigoriadou, M., Magoulas, G.D., Kornilakis, H., 2002. Towards new forms of knowledge communication: the adaptive dimension of a web-based learning environment. Computers & Education 39 (4), 333–360.

Pelgrum, W., 2001. Obstacles to the integration of ICT in education: results from a worldwide educational assessment. Computers & Education 37 (2), 163–178.

Polsani, P.R., 2006. Use and abuse of reusable learning objects. Journal of Digital Information 3 (4).

Rani, N.U., 2006. Educational Television in India: Challenges and Issues. Discovery Publishing House, New Delhi, India.

Selinger, M., 2001. Learning information and communications technology skills and the subject context of the learning. Journal of Information Technology for Teacher Education 10 (1–2), 143–156.

Shee, D.Y., Wang, Y.S., 2008. Multi-criteria evaluation of the web-based e-learning system: a methodology based on learner satisfaction and its applications. Computers & Education 50 (3), 894–905.

Shulman, L.S., 1986. Those who understand: knowledge growth in teaching. Educational Researcher 15 (2), 4.

Shulman, L.S., 1987. Knowledge and teaching: foundations of the new reform. Harvard Educational Review 57 (1), 1–23.

Smeets, E., 2005. Does ICT contribute to powerful learning environments in primary education? Computers & Education 44 (3), 343–355.

Tondeur, J., Hermans, R., Van Braak, J., Valcke, M., 2008. Exploring the link between teachers' educational belief profiles and different types of computer use in the classroom. Computers in Human Behavior 24 (6), 2541–2553.

Tondeur, J., Van Braak, J., Valcke, M., 2007. Curricula and the use of ICT in education: two worlds apart? British Journal of Educational Technology 38 (6), 962–976.

U.S. Dept. of Education, 2010. Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies. Washington, DC.

Zhang, D., Nunamaker, J.F., 2003. Powering e-learning in the new millennium: an overview of e-learning and enabling technology. Information Systems Frontiers 5 (2), 207–218.

Relevant Websites

General Information

http://www.worldbank.org/ – ICT and Education, World Bank.
http://www.unesco.org/new/en/unesco/themes/icts/ – ICT in Education, UNESCO.

Digital Learning Resources

http://www.education.uoit.ca/lordec/ – LORDEC (Learning Object Research, Development and Evaluation Collaboratory), Canada.
http://www.merlot.org/merlot/ – MERLOT (Multimedia Educational Resource for Learning and Online Teaching), USA.
http://ocw.mit.edu – MIT OpenCourseWare, USA.
http://www.ndlrn.edu.au/ – National Digital Learning Resource Network, Australia.

Course Management Systems

http://www.blackboard.com/ – Blackboard, Course Management System.
http://moodle.org/ – Moodle, Course Management System (Open Source).

TPACK Framework

http://tpack.org/ – About TPACK.
http://activitytypes.wmwikis.net/ – Learning Activity Types Wiki.

Virtual Schools and Universities

http://www.academicearth.org/ – Academic Earth, USA.
http://www.coursera.org – Coursera, USA.
http://www.ignouonline.ac.in/ – Indira Gandhi National Open University (IGNOU), India.
http://openlearn.open.ac.uk/ – The Open University Learning Space, UK.

Information and Knowledge, Organizational

Patrizia Porrini, Long Island University, Brookville, NY, USA
William H Starbuck, University of Oregon, Eugene, OR, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Information is data that have meaning. In organizational contexts, meaning derives from sense-making frameworks that reflect organizations' properties, including communication within and between organizations. Organizations transmit different messages internally than externally.

Knowledge is an accumulation of information. Whereas information changes rapidly, knowledge changes incrementally. Distinctions between data, information, and knowledge are often ambiguous, as sense making depends on knowledge, which derives from information.

Information and knowledge have grown very important in developed economies. Some developing countries have been investing in advanced education, so they have been replacing developed countries as suppliers of expertise.

Explicit efforts to manage knowledge yielded some successes and many disappointments.

Properties of Information and Sense Making by Organizations

Information is data that have meaning. Meaning arises as organizational sense-making frameworks interpret data. This sense making depends on what people already know as well as what data people can access.

Organizational information derives its meaning from the sense-making frameworks that characterize specific organizations. Many organizational members need information in order to fulfill their responsibilities, so other members gather data and convert it into information. Organizations can acquire data by scanning environmental data sources, by ingesting new members, or through internal activities such as accounting and research. Many organizations may have explicit strategies to acquire information and to maintain and archive information and knowledge. Most organizations include personnel who specialize in data acquisition or information processing, and large organizations have departments that focus on such specialties. Organizations convert data into information by discussing the implications of data, by adding to databases, or by feeding data into decision processes. They process information by filtering it, integrating it, disseminating it internally, and interpreting its implications. Nearly all organizations incorporate rules and procedures that cause personnel to initiate actions or to halt actions when they receive information that satisfies certain conditions.

Strangely, both managers and scholars paid little attention to the centrality of information in organizations' activities until the last half of the twentieth century. Not until the 1950s did articles and books begin to portray managers and work groups as decision makers. In 1958, March and Simon pointed out that information sources and channels affect organizations' perceptions, intraorganizational conflict, and goal coherence. They emphasized the human limitations that prevent people and organizations from acquiring and processing unlimited amounts of information. They also characterized people and organizations as using programs to process information, an analogy to computers.

Weick (1969) offered the first comprehensive analysis of organizations as information-processing systems. Weick treated information processing as the essence of organized activity. Indeed, he argued that organizations are not static systems and that organizing is a never-ending process through which organizations continuously evolve. This evolution occurs primarily in ideas, perceptions, data, beliefs, and communications, and people are endlessly choosing whether to follow standardized routines. Whereas Weick described evolution within one organization, Nelson and Winter (1982) described processes of imitation and knowledge transmission that foster evolution by populations of business firms. Evolutionary ideas have subsequently guided studies of technological development and the diffusion of knowledge.

Hodgkinson and Sparrow (2002) reviewed multitude psychological studies and sought to identify the key elements of 'strategic competence,' the ability of an organization to learn what it needs to learn and to remember what it ought to remember. They emphasized that humans have limited capacities for information processing, so they rely on simplified schemas to navigate complex, data-rich environments. These schemas incorporate biases and filters that can both facilitate action and foster errors.

Beginning late in the twentieth century, many organizations turned away from meetings in which managers pool their judgments and develop perceptions collectively, and instead, these organizations are educating managers about actual properties of organizations and market environments. This trend has propelled the SAS Institute to more than three decades of double-digit growth. The institute's product integrates an Executive Information System with data warehousing, data analysis, client-server computing, applications development, graphics, report writing, project management, and decision support (Taylor, 1996).

The trend toward more objectivity and less subjectivity has involved data gathering. Data warehouses are massive databases that describe past transactions. Plummeting costs of data storage allow organizations to simply record everything, without having to decide in advance what information to

 http://dx.doi.org/10.1016/B978-0-08-097086-8.73074-1

gather and retain, so earlier decisions about what to record do not limit later analyses. For example, a data-warehouse analysis enabled a large cosmetics company to avoid making a costly pricing blunder. This company buys data recorded by checkout scanners in chain stores, and it gets daily reports from sales personnel who record data from retail outlets and transmit them to headquarters nightly. When a major competitor introduced a new product, this cosmetics firm planned to make a defensive price reduction on its own high-priced product to keep the new competitor from taking too much business away from that product. However, analyses of scanner data revealed that the new product was not actually competing with the company's high-priced product; it was competing with a product that had a much lower price.

On the other hand, organizations may gather data or process information so poorly that they damage performance. For example, a rapidly growing health-care insurance firm processed receipts from clients much faster than it did claims from doctors and hospitals. As one result, this firm greatly overestimated its revenues, underestimated its costs, and made unrealistic forecasts about its profits … until the manager in charge of information processing reported that bill payments had fallen in arrears by nearly a year.

Cohen and Levinthal (1990) proposed that organizations differ in absorptive capacity, which they defined as an ability to see the potential value of new information, to assimilate it, and to apply it. They pointed to ways in which research and development activities enhance absorptive capacity.

Organizations acquire some of their sense-making frameworks from commercial markets. Publishers release many books every year that promise to provide better ways to analyze situations and better solutions for problematic situations. Organizations hire consultants who claim that they can improve organizations' performance by analyzing economic environments, potential customers, competitors, employees' capabilities, unexploited resources, and so forth. Gurus assert that their distinctive insights can transform mediocre performance into superb performance. However, the chances are remote of coming up with a genuinely effective new idea, and several studies have found that claims about the effectiveness of new sense-making frameworks have had weak support. Organizations have adopted new jargon but either have not attempted to make changes or have obtained poor results from changes they made.

Probably because ineffectiveness saps demand, the markets for sense-making frameworks have been quite faddish (Calhoun et al., 2011). Yet, there have been long-run trends arising from workers' higher educational levels, increasing availability of data, increasing technological complexity, and globalization.

Communication within Organizations and between Them

Communications in organizations reflect the statuses and aspirations of organizations' members. People in hierarchies talk upward and listen upward. They send more messages upward than downward, they pay more attention to messages from their supervisors than to messages from their subordinates, and they try harder to establish rapport with supervisors than with subordinates. People also bias their upward messages to enhance good news and to suppress bad news. However, the studies that support these patterns occurred before modern telecom technology came into widespread use. Newer technology has been flattening organizational hierarchies and facilitating lateral communication.

Formal communications in organizations are systematically troublesome. Organizations bias their formal reports to win the support of employees, customers, investors, and the public, and the members of organizations use formal reports to promote their careers or other interests. As a result, formal reports contain many misrepresentations and biases, and organizations that take formal reports seriously are prone to run into trouble (Altheide and Johnson, 1980). Such issues raise the importance of informal communication. For instance, more profitable business firms pay less attention to formal reports and rely more strongly on informal communication, and more profitable firms base their analyses on information from more diverse sources.

Organizations differ in terms of which communication modes are beneficial. Military organizations may be more effective when they restrict communication channels, whereas research institutes may benefit from open channels and random interactions. Informal communication such as work interrupted by phone calls, random chats on the way to meetings, unplanned meetings among employees, and unexpected interactions in halls can be just as important as formal communication to people's willingness to take action. Message recipients assess information quality by properties such as timeliness, richness of detail, and verifiability through two-way conversations.

Galbraith (1973) proposed an approach to organization design that focuses on the communication needed to coordinate tasks. He pointed out that behavioral programs can substitute for communication. If tasks can be adequately defined as programs and if personnel follow these programs, personnel can coordinate their actions with little communication. Programs are effective, however, only when organizations face familiar situations. If organizations continually encounter new situations, old programs become ineffective and personnel have to coordinate by communicating with each other. Galbraith pointed out that organizations can make communication less necessary by separating tasks into self-contained clusters and by lowering performance goals, thus leaving some resources underutilized. Organizations can also substitute direct horizontal communication across departmental boundaries for vertical communication up and down hierarchies; this moves decision making down to lower-level people who possess needed information rather than transporting information up to higher-level decision makers.

Innovations in communication technology have had mixed effects. Until very recently, technology has compromised informal communication by failing to support facial expressions, voice inflections, and two-way interactions. The emerging availability of video support for communication alleviates these limitations. Moreover, technology has also been promoting more informal, ad hoc, and choice-oriented communication. In global organizations, technology has been facilitating more timely coordination and bringing lower-level employees closer to higher-level managers.

Managers send quite different messages to internal audiences than to external audiences. For example, they may describe a current situation as an opportunity to internal audiences while characterizing it as threatening to external audiences. Competing organizations emit misleading information and they hide their planned actions.

Properties of Knowledge

Knowledge is an accumulation of information. Whereas information comprises a flow that changes rapidly, knowledge is a stock that changes only incrementally. Some knowledge is explicit; other knowledge is tacit. Activities such as teaching, consulting, and practicing law draw on extensive knowledge without processing large amounts of current information. Other activities such as bookkeeping and data entry process much information without using much knowledge. However, distinctions between data, information, and knowledge are often difficult to draw, as the conversion of data into information depends on knowledge that derives from data and information. Since the meaning of knowledge arises in part from its structure, knowledge reflects the intentions of the people who compile or organize it, and different people often extract different knowledge from a shared flow of information.

An organization's knowledge is both more and less than the sum of the knowledge held by its individual members. Organizations can store knowledge in routines and social norms that they transmit to new members, so organizational knowledge can include the discoveries of former members who have departed. Organizations can also purchase databases and computer programs from other organizations. Nonaka (1994) argued that organizations can only share and organize explicit knowledge, whereas individuals often acquire knowledge tacitly and their tacit knowledge must become explicit before it can become organizational. However, Cook and Yanow (1993) described a production process in which workers transmitted techniques to each other without explicit communication. Nonaka and Takeuchi (1995) described knowledge development as spiral in which employees' interactions produce repeated conversions of knowledge between its tacit and explicit forms.

Unshared knowledge can be a source of power for individual members (Starbuck, 1992). Some knowledge produces strong bonds between specific members and external constituents, and these bonds remain the property of the individuals rather than their organizations.

Scholars have debated the circumstances in which organizations should exploit current knowledge versus searching for new knowledge. However, such a choice seems to be more academic than practical, because people make unclear distinctions between use and search and they apply confusing labels to the transformations of information and knowledge. People who perform very similar tasks may describe their activities rather differently as applying old knowledge to new problems, as creating new knowledge, or as preserving knowledge (Starbuck, 1992). It can be difficult to distinguish between creating knowledge and applying it, because research may have direct applicability and efforts to apply knowledge may yield fundamentally new insights. Indeed, to create valid knowledge about systems as complex as a human body or an economy, people have to apply their current knowledge. Similarly, applying and preserving knowledge complement each other. Merely storing knowledge does not preserve it for long periods, because people must relate it to their current problems and activities.

Because all knowledge is imperfect and volatile, how knowledge evolves is more important than what knowledge exists. Fads and fashions serve as media for knowledge development and dissemination. They draw on aspirations, enthusiasm, fear, greed, mass media, social influence, and social pressure to encourage people to try new ideas. They make people aware that knowledge deteriorates, and they facilitate the discarding of obsolete ideas. The prevalence of many fads and fashions provides constant reminders that knowledge is transitory (Abrahamson and Fairchild, 1999). Change processes both produce knowledge and make it obsolete. People take new looks at what they have been doing, they develop new perceptions and make discoveries, some of which have lasting value. As a result, efforts to exploit new knowledge can improve organizational performance even if the new knowledge per se is not a significant improvement. People who are struggling with very difficult problems need visions of possible improvement to keep them going.

Changes in Information and Knowledge

Studies of organizational knowledge began to appear in the 1990s and they began to attract attention around 2000. The main stimuli for rising attention have been changes in evolving activities in developed economies, geographic migrations of economic activities, and technological innovations. Computers not only facilitate the processing of information but also provide metaphors for information processing by both individual people and social systems. Thus, equipped with an appropriate sense-making framework, managers and scholars began to notice the importance of information processing.

Information and knowledge have been growing in importance in developed economies and this trend seems likely to continue. The trend involves several components: (1) The developed economies are shifting away from manufacturing and toward services. For instance, in Europe, the telecommunications sector has been growing just under 14% annually, and the software and computing services sector has been growing by 15–20% annually. These changes have been occurring in part because agriculture and manufacturing have been migrating to developing societies. (2) The fastest-growing occupations have been clerical, professional, and technical workers, including managers and administrators. Professional-service organizations such as health-care agencies, law firms, and public-accounting firms have become much more important, grown much larger, and expanded globally. Knowledge workers and information workers are replacing manual workers. For example, machine-tool operators are giving way to technicians who monitor computer-controlled machines. (3) Workers of all kinds are gaining more education and more information-processing skills. Education, especially the higher levels of education that became prevalent during the last half of the twentieth century, has meant that more workers have valuable knowledge and more of them want jobs that allow

them to display or create knowledge. In the United States in 1900, only 14% of adults had completed 12 years of schooling. By 1998, 83% of adults had completed 12 years of schooling. Organizations have also been providing more education for their employees. Only 40% of the Americans who are receiving education are doing so in schools, and 60% are receiving it in their employing organizations. (4) Business is investing heavily in equipment for information work. In the United States in 2011, information-related equipment accounted for over 50% of capital investment. (5) Knowledge-intense and information-intense organizations are employing millions of people (Starbuck, 1992). (6) Corporations are depending on communications and information processing to keep them coherent as they globalize. Transoceanic communications have been growing about 35% annually, with the faster growth being in the Pacific region. (7) Firms are forming alliances and coalitions that rely on shared databases and electronic communication instead of face-to-face meetings. Business-to-business commerce has grown very rapidly during the early years of the twenty-first century.

Current thinking about strategic management emphasizes the usefulness of valuable resources that are rare and difficult to imitate (Barney, 1991). Both managerial and academic writings highlight the importance of knowledge to business success (Kogut and Zander, 1992; Nonaka and Takeuchi, 1995). Many organizations, not only the knowledge-intensive ones, seek to accumulate valuable and proprietary knowledge that they can exploit strategically. For instance, chemical and pharmaceutical companies employ skilled scientists and spend heavily on research aimed at developing patentable, proprietary products. Some manufacturing firms assert that cadres of experienced and highly trained managers comprise their key assets.

The policies of national governments have been changing the global distributions of expertise (Lewin et al., 2009). China, India, Singapore, South Korea, and Taiwan have been building capacity for knowledge-intensive work. For instance, China multiplied doctoral education in science and engineering by a factor of six between 1995 and 2003. Global corporations have been taking advantage of these plentiful and lower-cost resources by shifting their product development activities to these countries and away from the United States and Europe. At the same time, other countries – such as Canada, Europe, Latin America, Mexico, the Philippines, and the United States – have not built their capacity for knowledge-intensive work and they have not attracted it.

Knowledge Management within Organizations

Around 2000 managers and scholars began to pay much more attention to methods for managing organizational knowledge. Organizations have been trying to create knowledge management systems in order to store, create, transfer, and apply knowledge. Academic research in operations, information technology, economics, marketing, and management has proposed and analyzed ways to support the creation of new knowledge and to store and codify knowledge. Many firms have appointed learning officers who try to identify useful knowledge and to disseminate it to personnel who can use it. The University of Pennsylvania is offering a masters degree for people who aspire to become learning officers. Krogh et al. (2012) inferred that knowledge creation requires leadership that is distributed throughout an organization and that provides appropriate resources and structures.

Studies have suggested that three knowledge-management methods may be yielding benefits. All three methods focus on communication. First, some organizations are attempting to foster the conversion of tacit knowledge to explicit knowledge by organizing employees into on line communities that identify and discuss shared issues. Second, some organizations are attempting to disseminate knowledge by creating multiple channels for knowledge transfer and creating organizational cultures that advocate sharing. Third, some organizations are attempting to overcome the reluctance to share knowledge by adopting incentive pay systems that take account of overall organizational performance.

However, other knowledge-management methods face serious challenges that have caused them to disappoint their users (Calhoun et al., 2011). Several studies have documented knowledge-management efforts that proved unsuccessful. One central problem is the ambiguity surrounding knowledge and its value. What knowledge deserves keeping? Databases can easily grow so large that searching through them becomes very difficult. Indexing terms that appear useful today may make knowledge inaccessible tomorrow (Schultze and Leidner, 2002). A second central problem is that knowledge management techniques build on existing knowledge, so they are useful only as long as technological and social changes follow continuous, mainly linear trajectories. When technologies and societies make abrupt, nonlinear changes, knowledge-management methods are likely to impede organizations' efforts to adapt. A third central problem, related to the second, is that organizations, especially large ones, change slowly, so they need time to exploit new knowledge. To try to take advantage of the knowledge that is benefiting other organizations today is likely to yield disappointment because competitors are already exploiting that knowledge, so latecomers can hope only to catch up. Although current trends generally persist for a few periods, short-term benefits are likely to be small benefits. To extract large benefits from knowledge, organizations need to predict what knowledge will turn out to be beneficial a few years in the future.

Powerful organizations that collaborate with each other have advantages in predicting the medium-term future. Barnett et al. (2003) evaluated the accuracy of 3142 forecasts about US manufacturing industries that forecasters made during the 1970s. The studied forecasts concerned events that would not occur for at least 8 years into the future. The analysis found that forecasts about highly concentrated industries proved to be more accurate, especially forecasts about variables over which companies could exert stronger unilateral control. However, differences in the accuracy of forecasts were small in comparison with the wide ranges of errors that are typical of longer-range forecasts.

See also: Absorptive Capacity (of Organizations); Design and Form: Organizational; Forecasting; Innovation; Internationalization: Patterns in Business Firms; Learning: Organizational; Professions in Organizations; Research and

Development in Organizations; Strategic Management; Strategizing; Technology and Organization.

Bibliography

Abrahamson, E., Fairchild, G., 1999. Management fashion: lifecycles, triggers and collective learning processes. Administrative Science Quarterly 44, 708–740.

Altheide, D.L., Johnson, J.M., 1980. Bureaucratic Propaganda. Allyn & Bacon, Boston, MA.

Barnett, M.L., Starbuck, W.H., Pant, P.N., 2003. Which dreams come true? Endogeneity, industry structure, and forecasting accuracy. Industrial and Corporate Change 12 (4), 653–672.

Barney, J., 1991. Firm resources and sustained competitive advantage. Journal of Management 17, 99–120.

Calhoun, M.A., Starbuck, W.H., Abrahamson, E., 2011. Fads, fashions and the fluidity of knowledge: Peter Senge's *The Learning Organization*. In: Easterby-Smith, M., Lyles, M.A. (Eds.), Handbook of Organizational Learning and Knowledge Management, second ed. Wiley, Chichester, pp. 225–248.

Cohen, W.M., Levinthal, D.A., 1990. Absorptive capacity: a new perspective on learning and innovation. Administrative Science Quarterly 35, 128–152.

Cook, S.D.N., Yanow, D., 1993. Culture and organizational learning. Journal of Management Inquiry 2, 373–390.

Galbraith, J.R., 1973. Designing Complex Organizations. Addison Wesley Longman, Reading, MA.

Hodgkinson, G.P., Sparrow, P., 2002. The Competent Organization: A Psychological Analysis of the Strategic Management Process. Open University Press, Buckingham.

Kogut, B., Zander, U., 1992. Knowledge of the firm, combinative capabilities, and the replication of technology. Organization Science 3, 383–397.

Krogh, G., Von, Nonaka, I., Rechsteiner, L., 2012. Leadership in organizational knowledge creation: a review and framework. Journal of Management Studies 49, 240–277.

Lewin, A.Y., Massini, S., Peeters, C., 2009. Why are companies offshoring innovation? The emerging global race for talent. Journal of International Business Studies 40, 901–925.

March, J.G., Simon, H.A., 1958. Organizations. Wiley, Cambridge, Mass.

Nelson, R.R., Winter, S.G., 1982. An Evolutionary Theory of Economic Change. Harvard University Press, Cambridge, MA.

Nonaka, I., 1994. A dynamic theory of organizational knowledge creation. Organization Science 5, 14–37.

Nonaka, I., Takeuchi, H., 1995. The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, Oxford.

Schultze, U., Leidner, D., 2002. Studying knowledge management in information systems research: discourses and theoretical assumptions. MIS Quarterly 26 (3), 213–242.

Starbuck, W.H., 1992. Learning by knowledge-intensive firms. Journal of Management Studies 29, 713–740.

Taylor, P., May 1, 1996. Survey – FT IT: R&D Keeps Software at the Leading Edge. Financial Times.

Weick, Karl E., 1969. The Social Psychology of Organizing, second ed. Addison-Wesley, Reading, MA. in 1979.

Information Processing Architectures: Fundamental Issues

James T Townsend and Haiyuan Yang, Indiana University, Bloomington, IN, USA
Trisha Van Zandt, Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article starts with a brief history of information-processing architectures, emphasizing a traditional experimental paradigm, the additive factors method, and the classic problem of model mimicry. Several solid approaches to identifying mental architecture are introduced with a discussion of the necessary assumptions for these tests. Other fundamental issues attached to information-processing systems are also presented. A brief discussion with regard to challenges for the future concludes the article.

In the late nineteenth century, budding experimental psychology began to investigate perceptual and mental operations that would be considered part of the contemporary purview of cognition. Such interests reappeared in the 1960s in the emerging information-processing approach to elementary cognition. Over the last several decades, considerable progress has been made in this tradition through the combination of mathematical modeling and experimental design, an approach becoming known as metamodeling. This section reviews a number of selected findings within this approach, which is relatively novel in the social sciences. One benefit of the strategy is that potential dilemmas involving model mimicking, where models based on strikingly different principles can imitate each other's predictions, are brought to light. In particular, problems and experimental solutions associated with testing among parallel (simultaneous) versus serial (one-at-a-time) processing, stopping rules (logical basis for cessation of processing), limited versus unlimited versus Supercapacity (effects of increasing workload on processing speed), and dependence (issue of stochastic independence of channels or events) relationships are described. The foregoing issues are fundamental in almost all information-processing situations since virtually any operating system must take them into account. The section concludes by pointing to future challenges within this approach.

The term 'information-processing architecture' refers broadly to the arrangement of mental subsystems that are hypothesized to be active in the performance of one or more psychological tasks. For instance, the simplest, most prototypical, and opposed types of architectures are serial (one-at-a-time) versus parallel (simultaneous) arrangement of two or more separate subsystems or processes. More complex arrangements are mentioned below (*see Additive Factor Models*). Further, there are a number of other aspects of perceptual and cognitive processing that are often included under the 'architecture' label including the questions of the basis on which cognitive processes will cease, various kinds of independence and dependence, and processing capacity. These will be described and discussed.

An early researcher especially pertinent to this review was Donders, a nineteenth-century Dutch physiologist. Donders believed that he could uncover the durations taken by various thought processes through his method of subtraction. The method of subtraction was based on the idea that complicated mental activities were compounded in a simple sequential fashion from less complex parts. Let mean response time be written as $\overline{RT}$ and response time in general as simply RT. Then, for instance, the scientist might engage the subject in a task requiring both perception and decision and compare $\overline{RT}$ from that condition with $\overline{RT}$ from a task requiring only perception. The difference in $\overline{RT}$s was interpreted as the mean duration of the decision operation.

The issues selected for review here seem elemental in the following sense. The construction of almost any system intended to carry out the processing (e.g., detection, search, comparison, recognition, recall, analysis of various kinds, and so on) of a finite number of tasks or items would have to make decisions on each of the studied issues. In addition, they are somewhat unique in having been subjected to quite general theoretical, quantitative, and methodological analysis, perhaps more than any other such concepts in the field. The issues will be introduced in the context of a popular experimental paradigm and then discussed in more detail.

An Experimental Paradigm and the Major Issues

The concepts to be defined below have figured prominently in studies on short-term memory and visual display search and we shall employ that type of paradigm for illustration. Considerable impetus was given to the information-processing approach by several pioneering studies in the 1960s focusing on short-term memory and visual processing, using RT as the dependent variable (Atkinson et al., 1969; Egeth, 1966; Sternberg, 1966). In the roughly 50 years of the interim, scores and perhaps hundreds of studies in memory and visual search have been carried out. We focus on Sternberg's (1966) short-term memory search paradigm for illustration. Short-term memory is specified operationally by the tasks requiring the retention of a small number of items for anywhere from a few seconds to several minutes. In this paradigm, a varying number (less than or equal to the number that can be maintained in short-term memory without error) of items, for instance, randomly arranged letters, is presented to the experimental participant. This is called the memory set. Then, a few seconds later, that participant is presented a so-called probe item and the task is

to indicate very quickly, but at a very high level of accuracy, whether or not the probe was contained within the memory set. As introduced above, RT is the primary dependent variable of interest here, although a substantial history of work with accuracy also exists.

The first, and in some ways the most critical question, referred to the temporal structure of comparison of the probe item with the memory items. A hallmark of many early information-processing models was seriality of operation, that is, each subsystem or component could operate only when preceding mechanisms were completed and only one could operate at any given time. Thus, *seriality* has the connotation that no overlap of processing times occurs, in addition to the sequentiality of starting times. This latter type of constraint is referred to as *discrete flow*. The alternative possibility arose early on that all the items might be searched simultaneously, that is, in parallel. This issue of processing architecture can refer to the manner of treatment with regard to items within a single stage of processing, for instance memory search, as in the present case, or to the arrangement of larger components of the system, such as stimulus encoding versus memory storage.

Of course, arrangements other than serial versus parallel are possible, for instance, a system being serial part of the time and parallel part of the time. When all operations under focus begin at the same time but can finish at different times, the system (or mode of processing) is of course simply parallel. However, when operations can feed into other components, but with overlapping processing times, the system is said to have continuous flow, rather than discrete flow. A great deal has been learned about discrete-flow architectures of considerable complexity over the past two decades (*see Additive Factor Models*). We will confine ourselves to serial and parallel systems and other processing characteristics that these can involve.

The primary RT durations under study in the memory search task are those associated with comparison of the probe with the memory items. Nonetheless, other time intervals, such as those associated with early sensory processing and late motor components, must also be taken into consideration. It is usual, but by no means uncontested assumption, that these so-called residual times are serially arranged with and stochastically independent from the process under study (Dzhafarov, 1992; Luce, 1986; Townsend and Ashby, 1983).

Besides the architecture issue, independence versus dependence among the processing times of the items is also an important question. For instance, in serial processing, if an earlier item takes more time than usual, a subsequent item might consume less duration due to more preparation time, resulting in a negative dependence among processing times. On the other hand, positive dependences are possible. An example would be a positive dependence created somewhat artificially if *attention* available on a trial were a random variable, for then the items would all tend to be processed quickly or slowly more or less in unison. Analogously, parallel systems can either be constituted with independent channels (e.g., Bundesen and Habekost, 2008; Eidels et al., 2010) or with positively (e.g., Townsend and Wenger, 2004b; Wenger and Townsend, 2001) or negatively (e.g., Usher and McClelland, 2003) interactive channels.

Another critical concept is that of *workload capacity* (or 'capacity' for short if the meaning is clear), which refers to how processing times are affected by the number of things to be worked on (Townsend and Ashby, 1983). This is most easily illustrated with a version of parallel processing where the system slows down when the number of items that are being processed increases (i.e., limited capacity). However, limitations in capacity that are indirect, even with serial processing, can be conceived. For instance, a serial processor might speed up as it goes through the items, due to warm-up effects, or slow down due to inertial factors or fatiguing of the processor. Even though capacity and independence are logically separate notions, they can interact. For instance, an important type of parallel system, one that can mimic serial processing, assumes that as the processing of each item is completed, its processing capacity is reallocated to remaining items (Townsend and Ashby, 1983). This obviously affects the overall RT, but also creates a positive dependence among the item processing times. Townsend and Wenger (2004b) built a theory and associated methodology featuring the capacity theme.

Yet another important notion is that of stopping rule (e.g., Sternberg, 1966; Townsend, 1974). Depending on the task, it may or may not be necessary for the system to process all of the items in order to make a correct response. In the memory search paradigm, if the probe item is present and located in the current stimulus set, the processing can cease at that instant, without finishing the remaining items. This event is known as self-termination. However, because short-term memory search consumes only a few hundred milliseconds it is possible that the system nevertheless completes all items. This event is called *exhaustive* processing. On probe-absent trials, it is necessary to process all of the memory items to be sure of correctly making a 'no' response, that is, exhaustive processing must occur. In some experimental designs, all the items are probes. These need not be physically identical. For instance, the task might be to say 'yes' if any of the items is a vowel and a target-present trial could contain all vowels. This latter case permits the possibility of *first-terminating* or *minimum time* processing. Of course, it is an empirical question as to whether any kind of self-termination can actually take place in high-speed perceptual or cognitive operations, and must be addressed experimentally in each case.

Space constrains the scope of mathematical detail in the present article, but we provide some fundamental, if simplified, depiction. Although the serial and parallel classes of models both contain an infinite class of possibilities, the serial notion has traditionally been attached to a particular serial model that assumes identical processing time random variables on each item, independent of the number of items in memory (i.e., the load) and of the order of processing the items. It is also often assumed that the individual processing times are themselves mutually stochastically independent. We call this the *standard serial model*. Let the density of processing time for each item be designated $f(t)$ and that for the independent residual processing time be $r(t)$. Then the density on a probe-absent trial, $p(\mathrm{RT} = t)$, or $p(t)$ for short, is just the mathematical convolution of the n processing densities and the residual time density:

$$p(T|EX) = r * f_1 * f_2 * \ldots * f_n$$

where each f_i, $i = 1,\ldots, n$, is equal to f. The expected or average theoretical RT is $E(\mathrm{RT}) = n \cdot E(T) + E(R)$, where T and R are the respective random variables for individual item-processing time and residual-processing time. Notice that processing is exhaustive, as designated by EX in the left-hand side, in this case. In the case of self-terminating (ST) processing on a probe-present trial, the number of items finished before the probe is found is itself a random variable. Under usual conditions, the probability that the probe will be found in the i-th processing position is just $1/n$, so the average density for this case is just.

$$E[p(RT|ST)] = [(1/n)\sum_{j=1}^{n} f_1 * f_2 * \ldots * f_{i-1} * f_i] * r$$

Similarly, the expected RT, with $\overline{\mathrm{RT}}$ as the sample statistic, is just

$$\begin{aligned} E[(\mathrm{RT}|\mathrm{ST})] &= (1/n)\cdot \sum_{j=1}^{n} jE(T) + E(R) \\ &= [(n+1)/2]\cdot E(T) + E(R), \end{aligned}$$

which yields the time-honored result that this line has half the slope of that for the exhaustive serial RT.

The concept of parallel processing, like serial processing, is usually confined in the experimental literature to a very special kind of parallel processing. Processing is often said to be parallel if average RT is invariant across increases in load. When exhaustive processing has to occur, as in the probe-absent trials of memory search, this stipulation actually implies a very unusual kind of parallel processing, unlike the situation with the standard serial model. For instance, consider a set of parallel models with independent processing times, with the additional provision of unlimited capacity, in the sense that the probability distribution on completion times for each item does not change as n is increased. This class of models serves as a prototype of parallel processing in an analogous way to the standard serial model class just presented. Let us call this class the standard parallel model.

Let $g(t)$ be the fixed density function on processing time for each of the n items. Then $E[\mathrm{RT}|\mathrm{EX}] = \mathrm{E}\{\mathrm{MAX}[T_1, T_2, \ldots T_{n-1}, T_n]\} + E(R)$, where R is the residual time as above and T_i is the processing time of the item in position i. We drop reference to R which plays no role in the function shape, letting PT stand for processing time, that is, $\mathrm{PT} = \mathrm{RT} - R$. Then $E[(\mathrm{PT}|\mathrm{EX})] = \int_{-0}^{\infty} [1 - G(t')]^n \mathrm{d}t'$, where $G(t)$ is the cumulative distribution function associated with g. It is straightforward to show from this formula that mean RT is indeed an increasing, concave function of n (Townsend and Ashby, 1983, p. 92). Note that $g(t)$ is invariant across values of n. Hence, many investigators use a very restrictive and typically unrealistic criterion for parallelism when they demand flat rather than increasing functions $\overline{\mathrm{RT}}$ of n, even under exhaustive processing conditions.

It can also be observed that if mean exhaustive $\overline{\mathrm{RT}}$s are flat functions of n, then mean ST (again, a single position exists in the memory list that contains the probe) processing times would actually be decreasing, within the same model – a strong prediction which apparently has never been checked in studies using this logic. Thus, in the rare cases where flat exhaustive $\overline{\mathrm{RT}}$ functions are actually found with exhaustive processing, the implications are quite strong and in favor of very supercapacity parallel processing. Models that can make such predictions are considered in Townsend and Ashby (1983, Chapter 4).

In contrast, in the case of ordinary ST processing when a single position contains the probe, standard parallel models do predict a flat $\overline{\mathrm{RT}}$ function. This can be readily intuited, since only the channel that processes the probe matters in ST processing, and it is independent of all the rest (and is of unlimited capacity into the bargain). Finally in this model, mean first-terminating times should decrease with n.

Important Related Research Topics

Although the issues considered here are applicable to virtually any information-processing system or task, there are some classical or contemporary subject areas for which they are particularly germane.

When applied within the confines of a single subprocess, such as search within a memory or display list, the concept of attention is evoked. How attention is deployed, when it ceases, whether its application is independent across the various items, and its strength as a function of the workload, all call upon the critical processing issues introduced above.

The topic of *automatic processing* or automaticity, traditionally discussed within the realm of attention, likewise overlaps our present issues (e.g., Schneider and Shiffrin, 1977). The most obvious correlate within our processing issues is that of capacity. Although writers are sometimes rather vague concerning detailed quantitative accounts of what automaticity means with regard to workload capacity, it seems apparent that parallel processing is definitely mandated. In addition, we have proposed that capacity should at least be at the unlimited level. That is, each subsidiary task should see its operations proceed at the same speed as if it were functioning alone. These and related matters are discussed further in Neufeld et al. (2007).

Another traditional theme also seeing resurgence is that of perceptual and cognitive *holism*, a topic of long standing interest in philosophy and, in the twentieth century, of *Gestalt psychology*. Although much has been learned about psychological holism, there has been little accomplished with regard to definitions, theory, and experimental investigation of information processing characterizations of dynamic holistic operations. In that spirit, a set of working axioms which portray holistic perception in terms of the present issues have been proposed and developed (e.g., Wenger and Townsend, 2001; Wenger and Ingvalson, 2003; Fific and Townsend, 2010). The accompanying essentials of holistic perception can be informally expressed as involving highly interactive, supercapacity, exhaustive and parallel processing.

The Mimicking Dilemma of Serial versus Parallel Processing

As mentioned earlier, one of the driving forces behind mathematical metamodeling in this area was the discovery that

mathematical representations of diametrically opposed psychological principles could nevertheless sometimes mimic one another, even to the point of mathematical equivalence of their defining probability laws (Murdock, 1971; Townsend, 1971; Townsend and Ashby, 1983). A historical account of this dilemma and its resolution is offered by Townsend et al. (2011). Hence, we will take some time here to outline the state of the art with regard to such questions within the parallel versus serial question.

Consider a model for the processing times of n items under fixed experimental conditions. With regard to the parallel–serial issue, suppose no assumptions are made other than the probability mixture of generalized convolutions in the case of seriality and joint distributions on processing times in the case of parallelism. Then the parallel and serial classes of models are equivalent, in the sense that mappings can be provided that homeomorphically (this is mathematical jargon for a one-to-one and bicontinuous function relating the two types of models) carry the serial probability distribution into the parallel probability distribution and vice versa (Townsend and Ashby, 1983).

Nevertheless, over time, accumulating theoretical results have demonstrated that if the scientist is willing to make further restrictive, but sometimes still very general (and reasonable), assumptions about the models, and/or more complex experimental designs are utilized, the parallel–serial issue can be decided. For instance, certain rather fundamental differences between serial and parallel processing can be explored in experimental methods designed to exploit those differences (Townsend and Ashby, 1983; Townsend and Wenger, 2004a).

Factorial Methodologies for Architectures and Systems Factorial Technology

One of the most promising and general approaches to identifying mental architecture (i.e., serial versus parallel processing) is that based on the notion of *selective influence* of experimental factors, a notion first employed in tests of strict seriality by Sternberg (1969) in his well-known additive factors method (*see Additive Factors Models*). All factorial methodologies, like the original Sternberg strategy, depend on the selective influence assumption, namely that distinct experimental factors affect distinct processing components (i.e., subsystems). It can be assumed that $\mathrm{RT}(X+\Delta X,\ Y)$ refers to the case where the X factor has prolonged RT but Y is at base level, and so on. Basically, the fundamental statistic for the original method and for most extensions was the *mean interaction contrast* (MIC). The MIC is defined as

$$\begin{aligned}\mathrm{MIC} = {} & \overline{\mathrm{RT}}(X+\Delta X, Y+\Delta Y) - \overline{\mathrm{RT}}(X+\Delta X, Y) \\ & - [\overline{\mathrm{RT}}(X, Y+\Delta Y) - \overline{RT}(X, Y)].\end{aligned}$$

Schweickert (1978) in his *latent mental network theory* contributed the first major extension of the additive factors method involving more complex architectures under the assumption of selective influence. Townsend and Ashby (1983) found that the MIC distinguished parallel and serial stochastic models when selective influence was assumed and Schweickert and Townsend (1989) produced general theorems for Schweickert's latent networks, within a stochastic setting, assuming exhaustive processing (see *Additive Factor Models*; Network Models of Tasks).

Although the early theorems in all this work were accomplished in the context of exhaustive processing, analogous results can be found in the case of ST and first-terminating processing times (e.g., Townsend and Nozawa, 1995; Townsend and Wenger, 2004a). Because Sternberg's original ideas inherent in his additive factors method have been extended in so many new directions, it has been suggested that the general approach be referred to as systems factorial technology (Townsend and Thomas, 1994). For instance, one novel strategy has been to enlist entire RT distributions in providing more powerful tests of parallel versus serial processing or other related architectures (Dzhafarov and Schweickert, 1995; Roberts and Sternberg, 1993; Townsend, 1990; Townsend and Ashby, 1983; see also Balakrishnan, 1994). For instance, in analogy to the MIC, the scientist can form a contrast function composed of the double difference (corresponding to the double difference in mean RTs in the usual case) of cumulative distribution functions. This new statistical function turns out to be very helpful in assaying mental architecture (Townsend and Nozawa, 1995). Another example of the use of entire distributions will be reviewed below.

The assumption of selective influence is critical to the legitimacy of systems factorial technology. Much has been learned in recent years about its foundational underpinnings, what may go awry if it is violated, and about certain experimental indications of its failure (Dzhafarov, 1997; Dzhafarov and Kujala, 2010; Townsend and Schweickert, 1989; Townsend and Thomas, 1994).

Discriminability Results on Stopping Rule, Capacity, and Dependence

The question may be raised as to whether processing issues other than parallel versus serial processing also suffer significant problems in identification within the search or related paradigms. With regard to the stopping rule, mathematical investigations have shown, somewhat ironically, that the same types of memory search data that were incapable of deciding the parallel–serial question could in many instances prove that processing was ST rather than exhaustive. Interestingly, it is more difficult to prove that processing is exhaustive in that ST models can mimic exhaustive processing but not vice versa (Townsend and Colonius, 1997; Van Zandt and Townsend, 1993).

Moreover, recent theoretical and empirical discoveries utilizing the entire RT distributions rather than means provide much strengthened tests of architecture (parallel versus serial) and in addition allow one to firmly distinguish stopping rules at the same time. In fact, within systems factorial technology and assuming selective influence at the distributional level, Townsend and Nozawa 1995 (see also Townsend and Wenger, 2004a) proposed an experimental design that they called the *double factorial paradigm* in which investigators can test architecture and stopping rule as well as capacity within the same block of experimental trials.

Additionally, the important matter of workload capacity seems to possess little in the way of mimicking predicaments.

Whether capacity increases, decreases, or is invariant in the face of workload alterations is immediately captured by the current statistical measures.

The most difficult issue of the ones under discussion is that of stochastic dependence, even though it plays a vital role in processing systems. Interestingly, at this point in time, dependence is more readily and more directly assessed within experimental designs based on accuracy rather than RT (Ashby and Townsend, 1986; Kadlec and Townsend, 1992).

Challenges for the Future

The bringing together of mathematics and experimental methodology has lead to a useful stratagem that we call metamodeling. Systems factorial technology provides an example of metamodeling. Metamodeling facilitates the development of experimental methodologies that are capable of testing broad classes of models, rather than highly specific models, against one another. This approach is especially helpful when attempting to settle well-defined issues, for example, diametrically opposed concepts such as parallel versus serial processing. It may be more difficult to use this strategy with highly complex and detailed models of phenomena (e.g., Van Zandt and Ratcliff, 1995).

Metamodeling has arguably led to striking advances in the ability to decide experimentally a number of elementary, but major, issues in the purview of human information processing over the last several decades of the twentieth century. Some of these were outlined above. However, many challenges remain. For instance, long-standing challenges relate to the somewhat pesky residual time component introduced earlier. It is difficult to know in many cases how successful our various strategies can be until its nature is firmly tied down. As noted earlier the residual time is usually assumed to be in series with, and stochastically independent of, the other processing components (Luce, 1986; Smith, 1995; Townsend and Ashby, 1983). However, this assumption may be and has been challenged. Thus, it has been contrasted with the extreme opposite assumption of perfect correlation, but still within a series arrangement of the residual component with the other stages of processing (Dzhafarov, 1992; see also a specific case of apparent nonindependence, Diederich and Colonius, 1991).

The most time-honored approach has been to attempt to separate the residual time component from those under study employing Fourier analysis, assuming stochastic independence of these components. This has proven to be a tricky and arduous strategy. Luce (1986) and Sheu and Ratcliff (1995) provide useful reviews and commentaries on this approach. Another kind of powerful, if often difficult, approach regarding the residual time component has been to prove that experimental predictions hold true regardless of its presence (Dzhafarov and Schweickert, 1995; Roberts and Sternberg, 1993; Townsend and Ashby 1983).

The residual time problem also abuts a more general possibility, namely that processes in the same forward sequence may not satisfy discrete flow, but rather obey continuous flow properties. For instance, almost all systems based on differential and integral equations would fall into the continuous flow class. Some intriguing progress, both theoretical and empirical, has been made on certain subclasses of such systems (McClelland, 1979; Miller, 1993; Schweickert and Mounts, 1998), but much more remains to be done.

Another evident opportunity for development is the extension of factorial methods in general, and systems factorial technology in particular, to other dependent variables than response times. The most palpable extension would be to accuracy or accuracy and response times. Another interesting and topical subject matter would be *confidence ratings*. In this regard, it has been pointed out that it constitutes a grave error to simply assume that predictions for response times will also apply to other observable variables (Townsend, 1984). Theorems that are appropriate for specific observables within an information processing milieu must be proven and tested in the experimental crucible.

As a final example that is offered is that of implementing the concept of a theoretical and methodological *sieve*. Within such a sieve, the investigator first establishes the major processing characteristics of information processing in a cognitive task, and then proceeds to propose and test more computationally detailed accounts that are in accord with the earlier findings. The methodological sieve is related to, but is more general than, the concept of *strong inference* put forth by Platt (1964) (e.g., Massaro, 1998).

See also: Discrete-State Models of Information Processing; Network Models of Tasks.

Bibliography

Ashby, F.G., Townsend, J.T., 1986. Varieties of perceptual independence. Psychological Review 93, 154–179.

Atkinson, R.C., Holmgren, J.E., Juola, J.F., 1969. Processing time as influenced by the number of elements in a visual display. Perception and Psychophysics 6, 321–326.

Balakrishnan, J.D., 1994. Simple additivity of stochastic psychological processes: tests and measures. Psychometrika 59, 217–240.

Bundesen, C., Habekost, T., 2008. Principles of Visual Attention. Oxford University Press, Oxford, UK.

Diederich, A., Colonius, H., 1991. A further test of the superposition model for redundant signals effect in bimodal detection. Perception and Psychophysics 50, 83–86.

Dzhafarov, E.N., 1992. The structure of simple reaction time to step-function signals. Journal of Mathematical Psychology 36, 235–268.

Dzhafarov, E.N., 1997. Process representations and decompositions of response times. In: Marley, A.A.J. (Ed.), Choice, Decision, and Measurement: Essays in Honor of R. Duncan Luce. Erlbaum Associates, Mahwah, NJ, pp. 255–278.

Dzhafarov, E.N., Kujala, J.V., 2010. The joint distribution criterion and the distance tests for selective probabilistic causality. Frontiers in Quantitative Psychology and Measurement 1, 151.

Dzhafarov, E.N., Schweickert, R., 1995. Decompositions of response times: an almost general theory. Journal of Mathematical Psychology 39, 285–314.

Egeth, H., 1966. Parallel versus serial processes in multidimensional stimulus discrimination. Perception and Psychophysics 1, 245–252.

Eidels, A., Donkin, C., Brown, S.D., Heathcote, A., 2010. Converging measures of workload capacity. Psychonomic Bulletin and Review 17, 763–771.

Fific, M., Townsend, J.T., 2010. Information-processing alternatives to holistic perception: identifying the mechanisms of secondary-level holism within a categorization paradigm. Journal of Experimental Psychology: Learning, Memory, and Cognition 36, 1290–1313.

Kadlec, H., Townsend, J.T., 1992. Signal detection analysis of dimensional interactions. In: Ashby, F.G. (Ed.), Multidimensional Models of Perception and Cognition. Erlbaum, Hillsdale, NJ, pp. 181–228.

Luce, R.D., 1986. Response Times: Their Role in Inferring Elementary Mental Organization. Oxford University Press, New York.

Massaro, D.W., 1998. Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. MIT Press, Cambridge, MA.

McClelland, J.L., 1979. On the time relations of mental processes: an examination of systems of processes in cascade. Psychological Review 86, 287–330.

Miller, J., 1993. A queue-series model for reaction time, with discrete-stage and continuous-flow models as special cases. Psychological Review 100, 702–715.

Murdock Jr., B.B., 1971. A parallel-processing model for scanning. Perception and Psychophysics 10, 289–291.

Neufeld, R.W.J., Townsend, J.T., Jetté, J., 2007. Quantitative response time technology for measuring cognitive-processing capacity in clinical studies. In: Neufeld, R.W.J. (Ed.), Advances in Clinical Cognitive Science: Formal Modeling and Assessment of Processes and Symptoms. American Psychological Association, Washington, D.C.

Platt, J.R., 1964. Strong inference. Science 146, 3642.

Roberts, S., Sternberg, S., 1993. The meaning of additive reaction time effects: tests of three alternatives. In: Meyer, D.E., Kornblum, S. (Eds.), Attention and Performance XIV. MIT Press, Cambridge, MA, pp. 611–654.

Schneider, W., Shiffrin, R.M., 1977. Controlled and automatic human information processing: I. Detection, search, and attention. Psychological Bulletin 84, 1–66.

Schweickert, R., 1978. A critical path generalization of the additive factor method: analysis of a stroop task. Journal of Mathematical Psychology 18, 105–139.

Schweickert, R., Mounts, J., 1998. Additive effects of factors on reaction time and evoked potentials in continuous flow models. In: Roberts, F.S., Dowling, C., Theuns, P. (Eds.), Progress in Mathematical Psychology. Erlbaum Associates, Mahwah, NJ, pp. 311–328.

Schweickert, R., Townsend, J.T., 1989. A trichotomy method: interactions of factors prolonging sequential and concurrent mental processes in stochastic PERT networks. Journal of Mathematical Psychology 33, 328–347.

Sheu, C.F., Ratcliff, R., 1995. Fourier deconvolution of reaction time data: a cautionary note. Psychological Bulletin 116, 285–299.

Smith, P.L., 1995. Psychophysically principled models of visual simple reaction time. Psychological Review 102, 567–593.

Sternberg, S., 1966. High-speed scanning in human memory. Science 153, 652–654.

Sternberg, S., 1969. The discovery of processing stages: extensions of Donders' method. In: Koster, W.B. (Ed.), Attention and Performance, vol. 2. North Holland Press, Amsterdam.

Townsend, J.T., 1971. A note on the identifiability of parallel and serial processes. Perception and Psychophysics 10, 161–163.

Townsend, J.T., 1974. Issues and models concerning the processing of a finite number of inputs. In: Kantowitz, B.H. (Ed.), Human Information Processing: Tutorials in Performance and Cognition. Erlbaum Press, Hillsdale, NJ, pp. 133–168.

Townsend, J.T., 1984. Uncovering mental processes with factorial experiments. Journal of Mathematical Psychology 28, 363–400.

Townsend, J.T., 1990. The truth and consequences of ordinal differences in statistical distributions: toward a theory of hierarchical inference. Psychological Bulletin 108, 551–567.

Townsend, J.T., Ashby, F.G. (Eds.), 1983. Stochastic Modeling of Elementary Psychological Processes. Cambridge University Press, Cambridge, UK.

Townsend, J.T., Colonius, H., 1997. Parallel processing response times and experimental determination of the stopping rule. Journal of Mathematical Psychology 41, 392–397.

Townsend, J.T., Nozawa, G., 1995. On the spatio-temporal properties of elementary perception: an investigation of parallel, serial and coactive theories. Journal of Mathematical Psychology 39, 321–360.

Townsend, J.T., Schweickert, R., 1989. Toward the trichotomy method: laying the foundation of stochastic mental networks. Journal of Mathematical Psychology 33, 309–327.

Townsend, J.T., Thomas, R.D., 1994. Stochastic dependencies in parallel and serial models: effects in systems factorial interactions. Journal of Mathematical Psychology 38, 1–34.

Townsend, J.T., Wenger, M.J., 2004a. The serial-parallel dilemma: a case study in a linkage of theory and method. Psychonomic Bulletin and Review 11, 391–418.

Townsend, J.T., Wenger, M.J., 2004b. A theory of interactive parallel processing: new capacity measures and predictions for a response time inequality series. Psychological Review 111, 1003–1035.

Townsend, J.T., Yang, H., Burns, D.M., 2011. Experimental discrimination of the world's simplest and most antipodal models: the parallel-serial issue. In: Colonius, Hans, Dzhafarov, Ehtibar (Eds.), Descriptive and Normative Approaches to Human Behavior in the Advanced Series on Mathematical Psychology. World Scientific, Singapore.

Usher, M., McClelland, J.L., 2003. The time course of perceptual choice: the leaky, competing accumulator model. Psychological Review 108, 550–592.

Van Zandt, T., Ratcliff, R., 1995. Statistical mimicking of reaction time distributions: mixtures and parameter variability. Psychonomic Bulletin and Review 2, 20–54.

Van Zandt, T., Townsend, J.T., 1993. Self-terminating vs. exhaustive processes in rapid visual, and memory search: an evaluative review. Perception and Psychophysics 53, 563–580.

Wenger, M.J., Ingvalson, E.M., 2003. Preserving informational separability and violating decisional separability in facial perception and recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition 29, 1106–1118.

Wenger, M.J., Townsend, J.T., 2001. Faces as Gestalt stimuli: process characteristics. In: Wenger, M.J., Townsend, J.T. (Eds.), Computational, Geometric, and Process Perspectives on Facial Cognition: Contexts and Challenges. Erlbaum, Mahwah, NJ, pp. 229–284.

Information Society

Alistair S Duff, Edinburgh Napier University, Edinburgh, Scotland, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by F. Webster, volume 11, pp. 7464–7468, © 2001, Elsevier Ltd.

Abstract

The concept of the information society is a master key for both social-scientific interpretation of contemporary society and sociophilosophical articulation of normative and public-policy issues. It expresses 'informatization' – translation of the Japanese original, 'johoka' – a manifold sociotechnical process involving postindustrial economic shifts, opening of institutions, information and knowledge explosions, and, especially, the rapid diffusion and phenomenal impact of new information and telecommunication technologies. This article examines canonical sources of what became known as the information society thesis, notably the prescient ideas of the late polymath Daniel Bell, criticisms of the thesis, and a selection of salient information policy challenges facing twenty-first century democracies and global organizations.

Introduction: Informatization and the Information Society

The idea of the 'information society' has emerged as a master key capable of unlocking numerous salient developments in the entire arc – economy, culture, and polity – of postindustrial societies, and indeed of the world, conceived as a global, interconnected community. The underlying claim is essentially very simple, that information in its various dimensions, modes, and phases, including, but not confined to, information and communication technologies (ICTs), has played a leading role in the transformation of human engagement with almost all domains of life, to such an extent that it is now valid to refer to advanced nations as information societies. While information has of course always been vitally important to social existence, it is now somehow center stage. The sense, not yet complete but apparently well underway, of a reconfigured sociotechnical order, a new form of being in the world, perhaps even an embryonic information-conscious global civilization, is what gives increasing purchase to the smart catchphrase 'the information society.'

A closely related concept is 'informatization.' Underutilized in the West, this awkward neologism has the advantage that it denotes a process rather than a product, thereby facilitating modulated, evidence-driven discussion about the extent and character of the changes afoot, without necessarily enforcing endorsement of stark propositions about a new society. 'On the one side,' as Frank Webster observes in *Theories of the Information Society*, "are subscribers to the notion of an information society, while on the other are those who insist that we have only had the *informatisation* of established relationships" (2006: p. 6). The information society concept is likely to remain controversial deep into the millennium, pitting believers against sceptics, revolutionists against evolutionists, optimists against pessimists, technological determinists against social determinists, even political right against left. Some of these axes will be examined below.

The process of informatization is extensive, multifaceted, and conflictual. It includes not just the spread of new ICTs in line with standard models of diffusion (Rogers, 2003), but the superseding of manufacturing industry by economies geared predominantly to services and myriad information-based occupations; explosive growth in news and other media content and outlets; the worldwide official trend in favor of disclosure, open institutions, and freedom of information; a less visible, gradual, and in many ways countervailing movement toward the commodification, classification, or encryption of data; unceasing expansion of scientific journals and other scholarly documentation; the phenomenon of 'grassroots' or 'wild' information flows, of which Wikileaks has been the preeminent exponent; and no doubt other, as yet unregistered, information-related developments quietly gathering strength behind the scenes.

One could argue that informatization, broadly understood in such ways, is a fundamental social principle, a motor of modern economic and political development. This is why scholars have been led to refer to the 'informatization of the world view' (de Mul, 1999) and to posit the existence of an 'info sphere' (Floridi, 2007: p. 59) – in more antique language, informatization seems to represent 'the spirit of the age.' It should be noted that the tendency to equate informatization and information technology is unhelpful, since the relation of many aspects of informatization to computers or telecommunications is contingent rather than logically necessary. Nevertheless, there can be no doubt that electronic and specifically digital technologies have played a decisive part in the contemporary informatization process, and that the arrival of personal computing and then the Internet is what precipitated the widespread consciousness of a transition to an information society.

At the root of the information society debate is the basic meaning of the word 'information.' There have always been two separate – overlapping, but essentially different – conceptions. The first is the technical conception seminally expounded in Claude Shannon and Warren Weaver's *Mathematical Theory of Communication* (1949). More commonly known as information theory, this field is dedicated to analyses of the conditions of efficient telecommunication, of getting information from senders to receivers with a minimum of erosion or 'noise.' Information here means message, any kind of communicable signal, but its content is immaterial. It must be measurable – must be capable of being cashed into and back out of binary digits – but it need not be intelligible. "Information," Shannon

and Weaver confirmed (1949: pp. 8, 31), "must not be confused with meaning," because "semiotic aspects of communication are irrelevant to the engineering problem." Information theory is recognized as an intellectual pillar of the postwar epoch, one that has assisted the development not only of computing and telecommunications, but also of cryptography, the theory of stock market investments, and numerous other branches of applied knowledge. Indeed, in this catholic sense, information has installed itself as a preoccupation of fundamental disciplines at a great remove from the social and behavioral sciences, such as physics and biology; it is arguably even, as one thinker vividly puts it, 'the new language of science' (von Baeyer, 2004).

However, this is by no means the ordinary-language meaning. In common parlance, information denotes meaningful messages, and, moreover, accurate and useful ones – facts, news, statistics, official reports, astronomical data, bank statements, rail timetables, and the like. When we ask for flight information at the airport desk, we do not expect to receive misdirections, still less to hear, say, a rousing rendition of 'Pretty Vacant.' Unfortunately, much of what cascades through the ubiquitous, ever-expanding channels of the global information infrastructure is information only in the minimal, Shannon–Weaver sense. It consumes disk space and clogs up networks, but as often as not it is substantively uninformative: it is phatics, pointless phone calls, palpable falsehoods, pornographic imagery, or plain gobbledygook. If, therefore, we are going to build an information society on solid foundations, a civilization of which we can be justly proud, we probably need to start with a great deal more clarity about what we wish to count as real information.

Whatever its semantic and other issues, information society rhetoric has swept through policy and official circles at all levels. It is difficult to find a country – even in the developing world – which has not at some point in the past promulgated its own information society plan, usually a wish-list about computerization and the 'wiring' of economic growth. At a regional level, the European Commission since the 1990s has heavily exploited the concept in its drive both to privatize and deregulate telecommunications, and to foster pan-European solidarity among citizens; Brussels now houses a dedicated Information Society and Media Directorate. Similar ventures were launched, although with less success, in Africa and other regions. Global bodies have also been active. The Organisation for Economic Co-operation and Development has invested enthusiastically in the promotion of information society indicators. The G8 flirted. However, the most significant global effort was the World Summit on the Information Society, which in 2003 and 2005 convened senior representatives of the nations to address pressing policy issues, such as Internet governance and the digital divide. The paramountcy of the information society phenomenon can thus no longer be denied; and there will presumably need to be more such summits to debate the future of the only planetary resource whose supply is literally inexhaustible.

A Canon of Information Society Studies

Some, especially the young, perceive the information society to be so natural that it does not need interrogation. But while from infancy they can 'surf' cyberspace and as teenagers some can dazzle with technical jargon, they often fail to grasp how the Internet really works, still less comprehend the invisible drivers of the wider information society. Rather than parading their online transactions, these 'early adopters' (Rogers, 2003) perhaps need to ask whether they are not puppets subject to "the soft compulsion of constant consumption training" (Jurgen Habermas quoted in Webster, 2006: p. 167). They are usually surprised to hear that scholars have been thinking about the information society for at least half a century, spotting trends, anticipating problems, mapping future scenarios. The present section introduces the main pioneers of what became known as the information society thesis, and seeks to explain precisely how each advanced this intellectual cause. While explicit talk of canonicity has not yet engulfed this most ecumenical and interdisciplinary of specialties, volumes of 'key readings' and 'classic texts' are common enough. At risk of premature closure, then, the following can be construed as a canon of information society studies (Duff, 2000; Raban et al., 2011).

Credit for launching the information society thesis must be given to Fritz Machlup. Machlup (1902–1983) was a classically trained economist, an émigré from Vienna who became one of America's most distinguished professors. It was while pursuing more or less routine research on the economics of patents that he had his 'Eureka moment.' Perusing, no doubt for the umpteenth time, some standard statistical tables, he suddenly saw that a great portion of the value of the US economy, and the occupations undergoing the fastest growth, were involved in either the production or the distribution of knowledge – broadly defined. The core of the world's leading economy, according to *The Production and Distribution of Knowledge in the United States* (Machlup, 1962), consisted not in agriculture, automobiles, or oil but in the intangible goods and services of the mind. Education was now a huge 'industry,' worth some $60 000 000 000, if one included not merely classrooms but training on-the-job, Sunday schools, and public libraries. Research and development, although more modest in terms of employment, was still worth $11 000 000 000, even disregarding its multiplier effects on the rest of the economy. The other major knowledge industries were information services, communication media, and information machines, all of them expanding aggressively. IBM was still mainly building typewriters, but Machlup was quick to note the 'ascent' of the computer, 'this amazing machine' (1962: pp. 306–307).

All in all, the knowledge economy accounted for $136 000 000 000, or nearly 30% of the US Gross National Product. It was a revelation, what looked like the makings of a bona fide paradigm shift. Reviewing the book, Kenneth Boulding, himself a leading economist, opined that "the very concept of a knowledge industry contains enough dynamite to blast traditional economics into orbit." In the event, mainstream economics preferred to continue in its normal conservative tracks. However, Machlup's work did eventually lead to the flowering of a new technical specialty called the economics of information, associated now with other luminaries such as the Nobel Laureates Kenneth Arrow and Joseph Stiglitz; and on a wider front, it had a permanent and salutary impact on the academy and beyond. Machlup himself assiduously pursued his new research front, his final service being the editorship of

an unfinished multidisciplinary oeuvre entitled *Knowledge: Its Creation, Distribution, and Economic Significance* (Machlup and Mansfield, 1980–1984).

Marc Porat also belongs in the canon. He is cited almost as much as Machlup, although he did not produce a single book, just an excellent PhD dissertation, which turned into a government report (Porat, 1977). He then disappeared from academic trace, resurfacing later as an information technology entrepreneur. Yet Porat's contribution to information society studies was substantial. He established the vocabulary of 'information economy,' 'information society,' and 'information policy,' pushing aside Machlup's talk of knowledge. He also rationalized the various categories of information jobs. In particular, Porat showed that there are really two types of information economy, the primary, composed of people, such as journalists and librarians, who engage directly with information, and the secondary, comprising workers who deal with it behind the scenes, in industries which are not as a totality classified as information industries. Administrative staff in a garage are illustrative of the latter, but Porat also included less obvious occupations: half of professional nursing, for example, he argued, was devoted to advising, recording, and other information activities. He then forecast, correctly, that workers in the US primary and secondary information economies would by 1980 outnumber manual workers. Finally, writing well after the publication of Machlup's classic, Porat was in a position to give information technology the place it now deserved. 'The foundation,' he declared (Porat, 1977: p. 205), "of the information economy, our *new* central fact, is the computer."

The third and greatest figure, in what was a pioneering western triumvirate, is the late polymath Daniel Bell (1919–2011). A professional journalist before becoming a sociologist, Bell was able not just to offer broader social-scientific perspectives on the technical and economic changes underway, but also to turn them all into a compelling narrative, a feat he accomplished definitively in *The Coming of Post-Industrial Society* (1973). Bell's central claim, first voiced at a European seminar as early as 1959, was extremely bold: the industrial era, no less, was passing away and a new world being born, one based not on manufacturing but on the provision of services, especially those of the human and professional varieties. Whereas labor in developing countries was a harsh 'game against nature' and in the industrial regions a 'game against fabricated nature,' in postindustrial societies life and work were becoming largely a 'game between persons.' And at the center of this socioeconomic shift was information and knowledge. "What counts," he pithily wrote (Bell, 1973: p. 127), "is not raw muscle power, or energy, but information." In particular, the highly refined form of information that is science – 'theoretical knowledge' – is the axial principle driving advanced societies, its progress facilitated by 'intellectual technologies' like cybernetics, forecasting, and decision theory. The computer had a starring role in Bell's account too, especially in later formulations; this information machine was, he wrote in an influential, Machlup- and Porat-citing, essay entitled 'The Social Framework of the Information Society' (Bell, 1980: p. 509), "the 'analytical engine' that has transformed the second half of the twentieth century." And presiding over the whole show, according to one of Bell's most controversial theses, was a new, meritocratic knowledge class, a replacement for the power elites of bygone eras.

In *The Cultural Contradictions of Capitalism* (1976), Bell explored, in masterly prose – he ought to be remembered as one of the finest prose-stylists of his day – the implications of this technoeconomic paradigm shift for the social order. The traditional, heavily ideological struggle between labor and capital, he suggested, was being overwritten by new kinds of conflicts, for example, between the capitalist economy's norms of self-discipline and 'delayed gratification' and the rank hedonism of modern bourgeois culture. Bell accurately foresaw major power-struggles over privacy, the funding of higher education, and intellectual property. He also bravely problematized the western intelligentsia's war on religion and universal values. Yet this prophet's vision was at heart progressive and futuristic – indeed, Bell is also often credited as the father of futures studies – and his optimism has proven to be the prevalent mood among commentators on informatization.

The Japanese are often underestimated in these Sinocentric times, yet they actually coined the terms informatization, a translation of the Japanese word 'johoka,' and information society, or 'joho shakai.' It is in some ways not surprising that Japan, a country reeling from nuclear assaults and enjoying little more than tuna fish in terms of natural resources, should look to rebuild its economy on the intangible, undepletable asset of information. As early as 1968, Japanese researchers were quantifying the so-called information explosion, and from 1975 the government ministry of posts and telecoms began conducting a regular, large-scale information flow census. Using words as their unit, they demonstrated empirically what everyone sensed intuitively, that the supply of information had indeed greatly outstripped our capacity to consume it, and that this was largely due to the dramatic expansion of electronic media (Ito, 1981, 1991). It was also this research programme that first established the existence of a trend from mass communication to point-to-point, personalized information flows, what they called 'bunshu shakai' (segmented information society) – the latter now of course a dominant feature of the social network-laden info-communications landscape of the twenty-first century. The information explosion continues apace, now having to be measured in zettabytes (sextillions, or 10 to the power of 21) and even yottabytes (septillions, power of 24) (Hilbert, 2012).

It is certain, therefore, that the Japanese deserve a niche in the canon. Who else gets in is more difficult. However, one name unlikely to be disputed by any of the disciplines and factions, is that of Manuel Castells. His trilogy *The Information Age: Economy, Society, and Culture*, already into its second edition, has brought much of value to the theory of the information society – or, to be more precise, to the theory of 'informational society,' the heading he reserves for the new social form, on the grounds that all societies throughout history have been to some extent information societies. Castells has unabashedly made 'the information technology revolution' – title of volume one, chapter one (Castells, 2000) – the cornerstone of his analysis. He has correctly highlighted the key role of networks, both digital and offline. He has also internationalized the thesis, moving away from the triumvirate's focus on the United States to encompass developing as well as advanced countries.

This allows him to treat the effects of ICTs on dispossessed groups as diverse as al-Qaeda and the Mexican Zapatistas. The result is a major sociocultural thesis about how modernity and technology are forcing some communities into a resurgent, resistant politics of identity: the Net, as he puts it, vs the Self. Reflecting a more radical political outlook than Bell's, Castells makes the plight of the whole world his parish. In the world's 'back alleys,' he concludes with characteristic verve, "I have sensed the embryos of a new society" (Castells, 2004: p. 428).

Finally, brief mention should be made of the unfinished work of the much-mourned Rob Kling (1944–2003). He founded social informatics, "the body of research and study that examines social aspects of computerization—including the roles of information technology in social and organizational change and the ways that the social organization of information technologies are influenced by social forces and social practices" (Kling quoted in Berleur et al., 2006: p. 1). Although many scholars within social informatics tend to concentrate on the meso- and microlevels, shying away from more complex and demanding macrolevel debates, Kling himself was more adventurous. A disciplinary migrant from computing to library and information science, he oversaw groundbreaking research studies on the societal impact of computerization. Equally importantly, as an energetic editor of *The Information Society*, the first and, alongside *Information, Communication & Society*, the leading journal in the field, he moderated a high-profile space in which a richly textured, politically sophisticated information society debate could materialize.

Critiques of the Information Society Thesis

While an information society perspective, however raw, has gradually imposed itself on much of officialdom, it has not been without opposition in the academy. Some of the objections have been philosophical, others sociological; in addition, sections of the political left have maintained a dogged rear-guard action against central claims surrounding informatization and the information society – for them, the thesis is little more than a smoke screen for contemporary capitalism. The following paragraphs outline the chief criticisms and some standard responses.

The first, philosophical or perhaps sophistical, is about the cutoff. It asks at what point industrial society metamorphoses into information society, and insinuates that any such specification will be unjustifiable. This seems like a version of the ancient 'heap' paradox. There is a heap of, say, 100 stones. Take away one and it remains a heap. Take away two and it is clearly still a heap. But it gets trickier as one goes down, and eventually, the Greeks argued, one is forced into either saying that one stone constitutes a heap or making an arbitrary judgment about when to stop. The response to this is simply that in the context of informatization the tipping point is the unexceptionable yardstick of a majority. A nation becomes an information society when over half of its labor force can be credibly designated as information workers, and this point has already been reached in numerous western nations. If the objections were valid, it would count equally against the industrialization thesis, according to which preindustrial economies graduate eventually into industrial ones. But if no one is disputing the industrial revolution, why should anyone dispute the information revolution?

Another charge is that the information society thesis is trivial, if not downright tautological. The key variable, information, it is contended, is too vague, too all-encompassing. The information sector might well be over 50% of the economy, but only when it is defined so loosely as to be drained of much of its significance. Indeed, sociologists have joked that the definition is so liberal that it includes fortune-tellers and even 'strip-tease' dancers. Yet against this one just has to set the painstaking, blow-by-blow analyses of Machlup, Porat and their school. Exactly which of their categories is being impugned? Surely not education? Surely not R & D? Surely not people who work with computers and other information machines? As for other occupations, was Porat so mistaken in his decision to include half of nursing? Anyone familiar with the intricacies of modern nursing would have no problem with that. Dancers, subsumed in Machlup's category of communication media, along with the rest of the entertainment industry, admittedly seem more brawn than brain. Yet even if we score off this numerically, if not always morally, modest trade, that is hardly going to overturn the entire information society thesis. The accumulation of evidence for informatization, as regards not just the work force but also the exponential communications explosion and other vectors mentioned above, is now so strong that legitimate questions over particular inclusions should not blind us to its essential validity.

A further persistent charge has been that information society exponents are guilty of hyperbole, of outlandish assertions. While claims of partial informatization can be accepted, those alleging social revolution – propositions about a total transformation of society – are erroneous and dangerous. It must be admitted that some have talked in such a manner. "The information epoch to be brought about by computer-telecommunications technology," announced Japan's Yoneji Masuda (1981: p. 51), "does not simply mean that it will have a big socio-economic impact upon contemporary industrial society; it will demonstrate a force of societal change powerful enough to bring about the transformation of society into a completely new type of human society, which is the information society." Bell himself was also often accused of sensationalism and utopianism, despite his having explicitly distanced himself from the 'star-trek sociology' of some fellow travelers. The lesson is simple. It is essential when assessing the information society thesis to make a clear distinction between core and peripheral claims. Bell's core claim about the coming of postindustrial society is, as regards occupational structure and some other aspects of the socioeconomic system, now vindicated. In this sense, the headline 'the information society is here' is no exaggeration. Some of his concomitant assertions, about scientific knowledge driving social policy or the ascendancy of a new political class, are equally clearly mistaken, or at best premature. The same medicine can be applied to other information society theorists.

Many leftists have seen the use of the information society concept as somehow ideological or propagandistic. Behind the progressive technological scenarios – glib talk of

'computopia,' for example – lie, they insist, conservative economic scenarios. Much of the early animus against Bell, according to which this Marxist-turned-social democrat was "advanced capitalism's most persuasive advertising man" (Giddens, 1973: p. 21) and the like, was obviously politically loaded. The entire information society thesis, according to a self-styled 'neo-Luddite' treatise, is a mere 'snow job' for the mercenary powers orchestrating the capitalist system (Webster and Robins, 1986: p. 31). This whole tradition of criticism was more recently epitomized in Scott Lash's aphorism that the information society is "an order in which the principle of 'society' becomes displaced by the principle of 'information'—that *society itself* has been lost" (2002: p. 75). Socialist intellectuals get understandably incensed when right-wing politicians cite the theory of postindustrialism as an excuse for closing mines and factories. And they are absolutely right to insist that the information economy is an inherently fragile affair, that at the end of the day information cooks nobody's dinner. There are elements of genuine substance in the radical critique of the information society thesis, but for its own part, the left must adjust its social analyses to accommodate new realities. An intelligent socialism will first interpret the world, before trying to change it.

Perhaps the most durable charge, finally, is technological determinism, a term that has become something of a bogeyman in social-scientific circles. The idea that technology has single-handedly caused the changes outlined above, the transformation of industrial, energy-driven societies into postindustrial, information-powered ones, is clearly untenable. Many more variables need to enter the equation if one is to grasp what is happening. On the other hand, to deny any causal efficacy to ICTs is to be unable fully to account for a veritable kaleidoscope of surprising events, not least the phenomenal rise of new social powers like Facebook and Google. While technology should not be deified or crudely reified, it should be credited with its own measure of agency, not just explained away as a social construction – and in practice no one really thinks otherwise. Political factors, economic factors, sociological factors, cultural factors, psychological factors, all obviously also play their part; but it cannot plausibly be claimed that seismic occurrences like, say, the Arab Spring or Edward Snowden's megaleak of US state secrets, would have materialized without digital networks and other benefits of the Information Age.

The Information Society Agenda Today

Notwithstanding whatever is valid in the critiques of the thesis, it is safe to say that the information society has in several nonnegligible senses arrived. Thus, one can conclude that its existential crisis is over. This is not at all to suggest a moratorium on empirical studies or social-theoretical analysis of the information society. These need to continue, now focused on probing the nature of informatization in its manifold national, regional, and global manifestations. However, the main research action of the future should be not the empirical but the normative, not the positive but the political (Duff, 2012; Dutton, 1999).

The information society agenda is vast and open ended, and only a few issues will be introduced below. However, a few generic points can be made about its logical structure. The information society agenda will probably not be articulated so much by crude dichotomies, like utopian *vs* dystopian or technologically determinist *vs* socially determinist; its mode will be the continuum, the synthesis. Temporally, the agenda will be increasingly oriented not only to the future but also to the past, to memory and tradition as well as innovation. Spatially, policies will be shaped by considerations of scale and 'granularity.' In addition, the character of information society policy is likely to be a function not only of advances in computer-telecommunications but also of extraneous technological systems, such as nuclear power. And its approach will of course be interdisciplinary, because complex social problems have no respect for artificial borders. Nevertheless, the information society agenda will be governed by the perennial preoccupations of political philosophy, as classically conceived in terms of values, such as freedom, social justice, and fellowship. Indeed, it could be argued that the overarching aim of information society forward thinking today lies in the development of nothing less than what Walter Lippmann (1989) called a 'public philosophy.' Finally, the setting of this agenda will be not only for policy elites and the academy, but for all who feel the impact of informatization – which is to say the entire citizenry.

Certain issues have been prominent. The digital divide, for example, enjoyed a remarkable vogue in the 1990s. The broader problem of information inequality had been festering for decades, but the arrival of the World Wide Web threw it into sharp relief (Norris, 2001). Vice-President Al Gore made access to the 'information superhighway' a front-page issue, championing a raft of socially inclusive initiatives, such as the retooling of schools and public libraries. Subsequent Republican administrations toned down the egalitarian rhetoric, perhaps sensing that this new focus on social divides was unwelcome to their patrons on Wall Street. In some ways, whether by public policy or the invisible hand of the market, the digital divide is lessening. For example, the Oxford Internet Institute (Dutton and Blank, 2011: pp. 8–9) estimates that the proportion of households and individuals using the Internet reached 73% of the British population in 2011. However, it is obviously still a major problem compounding other endemic inequalities within preindustrial and industrializing nations. And at a global level the imbalance of information flows and technologies between rich and poor nations was what prompted the World Summit on the Information Society, in many ways a resumption of an idealistic United Nations initiative of the 1970s, the New World Information and Communication Order (Vincent et al., 1999).

Moreover, it would be a mistake to deduce that the problem of inequality has disappeared entirely even inside the post-industrial world. We still have our own poor, and information poverty is a significant, and usually underestimated, dimension of their predicament. This goes far deeper than lack of technology ownership or access; indeed, postindustrial paupers seem sometimes curiously well equipped in gadgetry. The essence of their information poverty lies rather in their ignorance of 'normal' democratic information, those crucial

facts about legal rights, welfare and political representation that well-heeled citizens take for granted. It might be countered that they could access such data sets through public libraries, if not their own phones and televisions, but the deeper point is that they are in many cases too demoralized even to want to make the effort. And it is difficult to see how anything short of major structural reforms could alleviate such ingrained false consciousness. In short, if policy makers are serious about cracking information inequality they may need to consider putting social engineering back onto their agenda.

The rapid diffusion of ICTs – at far above the rate of most innovations (Rogers, 1986: p. 132) – has made another recurrent information problem much worse, that of content regulation. Here one might say that the agenda is now not so much 'information management' as 'damage limitation.' The incessant deluge of textual, photographic, audio, and video materials, much of it of dubious information-value and some of grave moral questionability, is posing as yet unmet challenges. The harsh reality is that children now look up anything they fancy and parents who try to stop them are like King Canute ordering back the waves. The technical architecture of the Internet defies hierarchical control – ironically, since it was originally set up by the US military establishment – so even when the authorities decide to act, they are typically frustrated by the ability of information to hide or relocate. The traditional tools of media regulation are useless, because cyberspace is not just a medium, or even a supermedium: it is the habitat of a whole new world, a world without end and one with its very own – to continue in a theological vein – demonic dimension. Scientific, technical, and medical information is far more widely available than ever before, but so are sadomasochism and suicide Web sites. Thus the challenge of the future will be to enable information liberation 'within the limits,' to borrow a phrase from Kant, 'of reason alone.'

Information security, a topic seemingly so arcane and dull, will also be at the top of the agenda. It is not very long ago that most people refused to reveal financial and personal information online; now citizenship is impossible without doing so all the time. Yet while security standards have improved, regular scandals about lost card details, leaked search logs, illegally accessed medical records, breached military firewalls, and so on, vindicate the slogan 'Information Wants to be Free.' Better systems will be made widely available, but perfect protection is an unattainable ideal, because the problem is behavioral and 'sociotechnical,' not merely technical. So we must ask not just how R & D can develop information security, but also how society as a whole can assimilate the intrinsic vulnerability of its central asset, information, and the gigantic risks that this entails.

The information society agenda may have to accommodate some explosive new issues. The population census, for example, is one of the oldest and most useful information institutions in the world, but its very familiarity has caused it to pass largely under the research and top-policy radar. That will need to change. Among the issues that the twenty-first century census must address are the following: Is a complete population poll still necessary, given recent advances in sampling and modeling? If the answer is affirmative, then is a purely online census – a highly likely scenario – compatible with the convivial ethos embedded in public practices such as censuses or voting? What further types of questions might a future census seek to ask of informants? And, if private information corporations are going to continue being hired to implement the census, should there be more stringent requirements? For example, the 2011 Scottish census's use of CACI (UK), whose parent company CACI International admitted supplying several interrogators at the notorious Abu Ghraib prison in Iraq, raised profound issues of conscience for citizens – some of who preferred fines or imprisonment to putative dirty hands. If, with the glacial 10-year interval that censuses afford, the authorities cannot get such decisions right, it bodes ill for the general prospects of ethical information policy.

All of these issues point to the central question of the State. Some theorists have not hesitated to write its obituary, as though this were integral to the information society thesis, yet they have so far marshalled little hard evidence. Others have sought with more success to identify a new species, such as the 'informational state,' which Sandra Braman in an important treatise (2006) sees replacing the bureaucratic welfare state. Yet others still strive to keep the focus on long-running fears about technocracy – technology-enabled dictatorship – and surely with good reason. *Nineteen Eighty-Four*, George Orwell's science-fiction homage to individual freedom, is often cited by the public at large. Scholars tend now to reference Jeremy Bentham's concept of the Panopticon, the hidden, all-seeing potentate, especially since it was given (post)modern, antiestablishment expression in the work of Michel Foucault. The sense that one might be being watched by unknown others, hence that one should be just a little more circumspect, is becoming almost definitive of contemporary human experience – both online and offline. The problem for the information society, then, is to work out how there can be effective limits to interference by the State, and not just by the State but by other social powers too, such as corporations, pressure groups, and paparazzi. Should we invent a new institution akin to religious retreats, mooted as far back as the 1960s (Kalven, 1997: p. 249)? An increasingly attractive thought, certainly, yet the main solution will surely not lie in quasi-monastic forms of life but in institutionalizing revitalized norms of privacy inside the quotidian world.

Finally, it perhaps makes sense to begin to speculate about what, if anything, might lie beyond the information society, assuming technocracy can somehow be averted. Some speak of 'knowledge society,' a concept developed with considerable skill by sociological thinkers, such as Nico Stehr (2008). At the opposite end of the cognitive scale, Klaus Jensen (1999) interestingly predicts a shift from information society to 'dream society,' a scenario in which stories and imagination displace the cold light of reason. And with most nations having seemingly already deserted the battle against mass unemployment, people should have plenty of time thus to dream. However, there is no need for a stark either/or. What is more likely is parallel and countervailing developments, a further informatization, scientification, and complexification of political society combined with increasing flights to realms of feeling, intuition, and metaphysics. Thus, the great technoprophet's paradoxical vision of a revival of felt religion, what he called "the great instauration" (Bell, 1996: pp. 146–171), could well come true.

See also: Capitalism: Global; Consumption, Sociology of; Differentiation: Social; New Media, News Production and Consumption; Social Changes: Models.

Bibliography

von Baeyer, H.C., 2004. Information: The New Language of Science. Phoenix, London.

Bell, D., 1973. The Coming of Post-Industrial Society: A Venture in Social Forecasting. Basic Books, New York, NY.

Bell, D., 1980. The social framework of the information society. In: Forester, T. (Ed.), The Microelectronics Revolution. Blackwell, Oxford, UK, pp. 500–549.

Bell, D., 1996[1976]. Toward the great instauration: religion and culture in a post-industrial age. In: Bell, D. (Ed.), The Cultural Contradictions of Capitalism. Basic Books, New York, pp. 146–171.

Berleur, J., Nurminen, M.I., Impagliazzo, J. (Eds.), 2006. Social Informatics: An Information Society for All? In Remembrance of Rob Kling. Springer/International Association for Information Processing, New York, NY.

Braman, S., 2006. Change of State: Information, Policy, and Power. The MIT Press, Cambridge, MA.

Castells, M., 2000. The Information Age: Economy, Society, and Culture, second ed., vol. 1, The Rise of the Network Society. Blackwell, London, UK.

Castells, M., 2004. The Information Age: Economy, Society, and Culture, second ed., vol. 2, The Power of Identity. Blackwell, London, UK.

Duff, A.S., 2000. Information Society Studies. Routledge, London.

Duff, A.S., 2012. A Normative Theory of the Information Society. Routledge, London.

Dutton, W.H. (Ed.), 1999. Society on the Line: Information Politics in the Digital Age. Oxford University Press, Oxford, UK.

Dutton, W.H., Blank, G., 2011. Next Generation Users: The Internet in Britain. Oxford Internet Survey 2011. Oxford Internet Institute, University of Oxford, Oxford, UK.

Floridi, L., 2007. A look into the future impact of ICTs on our lives. The Information Society 23 (1), 59–64.

Giddens, A., 1973. The Class Structure of the Advanced Societies. Hutchinson, London.

Hilbert, M., 2012. How to measure "How much information?" theoretical, methodological, and statistical challenges for the social sciences. International Journal of Communication 6, 1042–1055.

Ito, Y., 1981. The *Johoka Shakai* approach to the study of communication in Japan. In: Wilhoit, G.C., de Bock, H. (Eds.), Mass Communication Review Yearbook, vol. 2. Sage, Beverly Hills, CA, pp. 671–698.

Ito, Y., 1991. *Johoka* as a driving force of social change. Keio Communication Review 12, 35–58.

Jensen, K., 1999. The Dream Society: How the Coming Shift from Information to Imagination Will Transform Your Business. McGraw-Hill Professional, New York, NY.

Kalven, H., 1997[1967]. The problems of privacy in the year 2000. In: Bell, D., Graubard, S.R. (Eds.), Toward the Year 2000: Work in Progress. The MIT Press, Cambridge, MA, pp. 244–250.

Lash, S., 2002. Critique of Information. Sage, London.

Lippmann, W., 1989[1955]. The Public Philosophy. Transaction, New Brunswick, NJ.

Machlup, F., 1962. The Production and Distribution of Knowledge in the United States. Princeton University Press, Princeton, NJ.

Machlup, F., Mansfield, U., 1980–1984. Knowledge: Its Creation, Distribution, and Economic Significance, 4 vols. Princeton University Press, Princeton, NJ. John Wiley, New York.

Masuda, Y., 1981. The Information Society as Post-Industrial Society. World Future Society, Bethesda, MD.

de Mul, J., 1999. The informatization of the worldview. Information, Communication & Society 2 (1), 69–94.

Norris, P., 2001. Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide. Cambridge University Press, Cambridge, UK.

Porat, M.U., 1977. The Information Economy: Definition and Measurement. United States Department of Commerce, Washington, DC.

Raban, D.R., Gordon, A., Geifman, D., 2011. The information society: the development of a scientific speciality. Information, Communication & Society 14 (3), 375–399.

Rogers, E.M., 1986. Communication Technology: The New Media in Society. The Free Press, New York.

Rogers, E.M., 2003. Diffusion of Innovations, fifth ed. The Free Press, New York.

Shannon, C.E., Weaver, W., 1949. The Mathematical Theory of Communication. University of Illinois Press, Urbana, IL.

Stehr, N. (Ed.), 2008. Knowledge and Democracy: A Twenty-first Century Perspective. Transaction, New Brunswick, NJ.

Vincent, R.C., Nordenstreng, K., Traber, M. (Eds.), 1999. Towards Equity in Global Communication: MacBride Update. The Hampton Press, Cresskill, NJ.

Webster, F., 2006. Theories of the Information Society, third ed. Routledge, London.

Webster, F., Robins, K., 1986. Information Technology: A Luddite Analysis. Ablex, Norwood, NJ.

Information Society, Geography of

Taylor Shelton, Clark University, Worcester, MA, USA
Matthew Zook, University of Kentucky, Lexington, KY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E. Sheppard, volume 11, pp. 7464–7468, © 2001, Elsevier Ltd.

Abstract

This article provides an overview of the sociospatial processes associated with the massive growth of information technologies over the past 20 years, as well as the intellectual developments that have sought to describe and analyze these processes. While geographers took on a key role in combating notions of the 'death of distance' in the 1990s, early research on the geography of the information society failed to create a coherent disciplinary approach. Recent work in geography has, however, created a more unified approach to understanding key sociospatial processes in the information society.

The 'information society' refers to the large-scale shift in contemporary societies where processes and practices of information production, analysis, and consumption take on an increasingly dominant role in social life. While the term typically refers to a broad array of information and communication technologies (ICTs), it has become most closely associated with the development of the Internet, the commercialization and popularization of which made the reality of an information society more visible and accessible to a wider range of people. Most recently, the spread of mobile phones throughout the developing world and the explosion of social media have made the previously new and exciting facets of the information society ever more routine through greater integration into the daily lives of people around the world.

The geography of the information society refers to the study of the new technologies and practices centered on information; the focus is on how these technologies and practices both are structured by longstanding sociospatial processes and act to reconstruct and reconfigure such processes. Given the centrality of the Internet in popular imaginations of the spatial expression of the information society, much of this research has focused on the so-called network of networks. Geographers have also been keenly interested in the growth of geographic information science and associated technologies over the past two decades, especially as Global Positioning System-enabled devices and spatially based search have become increasingly common in the past 5 years.

This article situates the concept of the information society in its historical and geographical context, while tracing the development of research on the geographies of the information society from their beginnings in the early 1990s to conceptual and methodological shifts that have taken place since the mid-2000s.

Situating the Information Society

Ultimately, any understanding of the 'information society' must begin from the context in which the concept emerged, which is intimately related to an understanding of what the information society is not, or at least what it is imagined not to be. While there is no single accepted definition of the information society or point at which society made a wholesale transition into the information society, the term is inextricable from the economic context of the 1970s, at which time deindustrialization and crisis were widespread in the developed Western economies (Bluestone and Harrison, 1982).

In response, many commentators pointed to the emergence of an economy, indeed an entire society, organized not around the production, exchange, and consumption of tangible, manufactured goods that were increasingly being produced more cheaply elsewhere, but of information (Touraine, 1971; Bell, 1973). Such ideas have continued to be influential, as more recent commentators such as Robert Reich (1992) and Richard Florida (2003) have argued that the future of Western economies rests on what they respectively refer to as 'symbolic analysts' or the 'creative class.'

The problem with such constructions of an 'information economy,' of course, is that they tend to neglect the fundamental interconnections between industrial and informational production, as well as the role of information in primary-extractive activities (e.g., the importance of information processing in petroleum exploration). Information has always been a crucial element of social and economic life, even in the industrial and preindustrial eras, just as material resources continue to be extracted and factories continue to manufacture goods in the supposed information age. After all, the computers, fiber-optic cables, server farms, and mobile devices that undergird today's information society have clear material presences and often significant environmental footprints as a result of their production, consumption, and disposal (Zook and Shelton, 2012).

So while computers, mobile phones, the Internet, and geographic information systems have all made significant impacts on social life in the last 20+ years, it is necessary to situate these technological developments within broader histories and understand that they exist along a continuum of quantitative intensification of information use, rather than as wholesale shifts or qualitatively different ways of living in and understanding the world. So while the information society as it is known is marked by real and important differences from previous areas, it is also deeply structured by the historical and geographical forces from which it emerged.

The Death of Distance?

The earliest commentaries on the geographies of the information society emphasized the new spatialities of economic activities related to globalization. Popular writing by non-geographers predicted that new ICTs would mean the 'death of distance' (Cairncross, 1997). Of course, such pronouncements were not without historical precedent, as similar claims were made of earlier advances such as the telephone, telegraph, and even of the postal service (Fischer, 1994).

Nonetheless, it was these popular, if fallacious, accounts that provided the impetus for a great deal of research on the geography of the information society, leading to a refutation of the claims of such popular writers. Rather than seeing these new technologies as having inexorable social effects, which in this case amounted to a spatial expression of technological determinism, geographers and other social scientists attempted to situate these new technologies in their social and spatial context, to understand them as contingent outcomes of particular processes with contingent effects. As Kirsch (1995) argues "space is not 'shrinking,' but rather must be perpetually recast" (p. 544).

But perhaps the earliest programmatic statement against such views of the space-transcending power of technology came from Manuel Castells' theorization of the network society, constituted by the divergent space of flows and space of places. Castells (1996) argued that the use of new communications technologies provided the means through which a new social organization could emerge, organized around the network logic of an electronic 'space of flows' overlaid on cities and localities. He argues that "[t]he space of flows is not placeless, although its structural logic is. It is based on an electronic network, but this network links up specific places, with well-defined social, cultural, physical, and functional characteristics …. Both nodes and hubs are hierarchically organized according to their relative weight in the network. But such a hierarchy may change depending upon the evolution of activities processed through the network" (p. 413). Thus, the 'space of flows' associated with the information society can enable a reformulation of place-based interactions into virtual interactions, but is merely a space of opportunity rather than a predetermined outcome. Instead, one sees the continued influence of places and their cultural and economic systems on the structure of the information society. Castells argues that these local factors, the longer standing 'space of places' in opposition to the newer 'space of flows,' determine the role of each place within the network and its importance or irrelevance.

While Castells' work promoted an interdisciplinary spatial perspective on the information society, early research on the geography of the information society tended to be somewhat diverse, with scholars addressing the Internet or associated technologies from a particular disciplinary or theoretical standpoint (Hepworth, 1990; Kellerman, 1993). For instance, while some political geographers studying the Internet might focus on the way social movements incorporated the Internet into organizing or protest strategies, economic geographers were investigating how the Internet was integrated into existing networks of production. Although the influence of such disciplinary divisions made it difficult for geographers to produce a unified body of theoretical and empirical work around the information society, and thus promote a lasting institutional home for this research, it is important to review and understand this work in this context.

Geographies of Urban and Economic Restructuring in the Information Age

Just as the definition of the information society itself tended to revolve around ideas of economic restructuring, so too did the earliest and perhaps most prominent work by geographers on the topic. Although geographers were always quick to point to the lack of nuance in arguments about the end of geography, even the earliest works touching on the geographic implications of new ICT offered ideas such as 'time–space compression' (Harvey, 1989) that highlight the connection between new technologies and globalization.

But further developments have tended instead to emphasize the simultaneous concentration and dispersal of economic activity. That is, while ICTs undoubtedly have allowed for economic transactions to take place over much greater distances, this process has been accompanied by a concurrent concentration of certain activities in particular places. So rather than seeing urban dispersal and dissolution as the result of these new technologies, urban and economic geographers have argued that cities continue to be crucial players in the information society.

On the one hand, the prototypical industries of the information economy have remained largely clustered in a handful of locations that specialize in knowledge production and innovation. Places like Silicon Valley in California, Route 128 in Massachusetts, the Research Triangle in North Carolina, as well as Cambridge in the United Kingdom and Tokyo in Japan, remain home to some of the world's most important electronics and biomedical companies. Instead of representing the death of distance, these clusters testify to the enduring relevance of agglomeration economies. Similarly, it is clear that economic activity in the information society tends to cluster in cities. Rather than exhibiting a placeless character, the infrastructure of the Internet, has emerged from long-standing political and economic processes, meaning that prominent 'global cities' such as New York and London have not been threatened by the Internet, but in many ways have been made even more important (Malecki, 2002). In addition, the information processing industry as a whole, including both those specializing in the development of the technologies themselves and those specializing in producing informational content, also remained highly clustered in a few cities, although in this respect cities like San Francisco were able to gain a relative advantage, solidifying their importance to the global system to which they had previously been peripheral (Townsend, 2001; Zook, 2005).

But the geographic contours of urban and economic restructuring in the information society are not limited simply to the positionality of major cities in the global economy. Indeed, the growth of the information economy affects the internal spatial structure of cities as well, ranging from growth in traffic congestion to changing roles for various

neighborhoods in the aforementioned cities. For instance, Graham (2002) documents how the dot-com boom in San Francisco led to a pattern of 'cyber gentrification' in some neighborhoods, marginalizing poor and minority populations in order to make way for wealthy tech start-ups. Areas zoned for light industry saw the influx of mixed-use loft developments that met the letter of the law while greatly increasing real estate costs, thus forcing manufacturing firms to close or relocate. So while the new activities and capabilities of the information society have enabled the restructuring of economies and city-regions, these new processes demonstrate the persistently variegated and multiscalar processes of uneven geographical development typical of earlier iterations of capitalism.

Globalizing Identities and Protest Movements

Just as urban and economic geographies have brought their own conceptual frameworks to the study of the information society, so too have a number of cultural and political geographers, although their numbers have been considerably smaller and somewhat less prominent. The most common thread woven through political geographies of the Internet have focused on the potential of the Internet to reconfigure the spatial relations of protest, allowing oppositional social movements to 'scale up' and reach broader audiences, which might in turn be able to promote their cause (Adams, 1996). And while there has always been significant skepticism about the ability of the Internet to enact change independent of broader social structures, the Internet continues to be seen as a key tool for social movements in that it offered a relatively unregulated space where oppositional and counterhegemonic discourses could flourish (Warf and Grimes, 1997; Froehling, 1997).

Similarly, cultural geographers have seen the Internet as an important contributor to the formulation and maintenance of identity. In some cases, the Internet has allowed diasporic groups to maintain a level of social cohesion despite the lack of colocation (Adams and Ghose, 2003), while in other examples the Internet has allowed identity to become more fluid in virtual space, perhaps even becoming divorced from corporeality (Roberts and Parks, 1999). Although cultural and political geography approaches to new technologies have become increasingly prominent in recent years, as will be described in the following section, these subdisciplinary perspectives were largely peripheral to debates going on elsewhere in the discipline and in studies of the information society more generally.

Toward New Geographies of the Information Society

While earlier approaches to Internet geographies were largely approached from preexisting subdisciplinary perspectives, with the Internet as a common object of analysis, theoretical and methodological shifts, along with technical developments and ever-expanding access to these technologies, have led most recently to efforts to create a more integrated approach to understanding the geography of the information society, with a series of central themes or problematics. While much work remains to be done, the authors identify four interrelated themes that have emerged within the past decade, and which promise to extend this line of research into the future.

Hybrid Spaces and the Virtual/Material Binary

The first and most important of these themes is the issue of the relationship between virtual and material spaces. Perhaps the most important contribution to this literature has been the conception of space as hybridized, or as variegated mixture of both material and virtual, rather than a so-called real world and an ontologically distinct cyberspace that we can choose to integrate into or remove ourselves from (Graham, 2013; see also Aoyama and Sheppard, 2003). As such, it becomes more and more difficult to conceive of the Internet as an analytically distinct object of analysis, leading one to conceive of practices, processes, and geographies of information as a unified set of phenomena worthy of study. Instead, one should understand "These cyberplaces are not simple floating and static mirrors of the physical world. They are instead often a component of the palimpsest of place. The virtual Earth and digital representations of place are often characterized by a reflexive relationship with their physical counterparts: they are shaped by, and, in turn, shape the physical world" (Graham, 2010: p. 423).

This unity includes everything from how one's movements through urban space are controlled in part by information profiles about him or her as individuals and types (e.g., radio frequency identification tags, biometric scanning, and pattern recognition), to the ways in which one's interpretation of cultural landscapes are shaped and influenced by increasingly available information and commentary about places that can be engaged with *in situ* and in real time. In short, the growing importance of computation to social life, beginning in the 1950s with the processing of government and insurance records, has become so deeply integrated into everyday life that the idea of 'the information society' is somewhat passé. Therefore, one might say that the geography of the information society is the geography of society writ large.

Rethinking Places and Flows

In addition to an understanding of space as hybridized has been a rethinking of the binary logics that underpinned earlier theorization about flows vs places, or of spatial distanciation vs localization. As was mentioned previously, many of the necessary correctives to death of distance thinking still involved an understanding of social relations stretched over wider and wider geographic areas.

In contrast, the last 5–10 years of work on the geographies of information have increasingly focused on how ICTs do not simply lead to a stretching of social relations but in many cases are implicated in the production of more intensely localized social relations, allowing for a closer connection to one's own place as much as, if not more than, allowing one to connect to someone half a world away.

This awareness of the place-based character of ICTs necessitates a rethinking of the classic Castellian binary between places and flows. It is necessary to understand, following the idea of hybrid spaces, that spatial relations are simultaneously

place based and flow based. For instance, the effects of ICTs on economic processes are no longer seen to allow simply for the outsourcing of production facilities or back-office functions around the world. Instead, it is clear that flows of material goods around the planet – a phenomenon in no way limited to the last 30 years of globalization – are fundamentally supported by accompanying flows of virtual information. As commodities are produced and consumed, the global economy is increasingly dependent on information produced about and attached to these commodities in order to connect them to their places of origin, their conditions of production, and their final resting place in the form of waste (Zook and Shelton, 2012). So just as one has been able to move past binaries between the virtual and the material, so too is it necessary that one moves past binaries between the local and the global, or between spaces of places and spaces of flows.

Code/Space

Another important theme in recent writing on the geographies of the information society has focused on the role of software in the remaking of social and spatial processes. Software code embedded in buildings, roads, traffic signals, and automobiles, among other things, acts as a constraint on the range of possible choices, allowing particular spaces to be used only in particular ways while simultaneously enabling certain forms of action. For instance, the failure of software-mediated processes can quickly repurpose a grocery store into a massive warehouse when check-out scanners no longer work or an airport into a simple waiting room when electronic tickets become inaccessible. Similarly, the code embedded in computers, mobile phones, and various search engines and social media applications structures one's ways of seeing the world by automatically sorting the massive amounts of information available to provide one with simplified and algorithmically constructed view of the built environment through which he or she moves (Zook and Graham, 2007).

Kitchin and Dodge (2011) understand the role of software code as a process of simultaneous translation and transduction. While the production of software code entails the inscription of certain commands, values, and social structures into automatically enacted instructions, the use of this code in turn produces new effects in the world, altering the ways that new iterations of code will structure one's experiences in the world. But, as Thrift and French (2002) lament, the incredible power of software in the world has not been subjected to the same critical scrutiny as other similarly powerful technologies. They argue that attention to the spatialities of software has been limited because we are trained not to see it, to ignore its presence in the interstices of our lives and social processes due to its absence from our field of vision. While this research remains relatively new, it is promising insofar as geographers have been at the forefront of such studies, integrating an explicitly spatial perspective into the interdisciplinary field of software studies.

Volunteered Geographic Information and Big Data

But the importance of such information does not come only from being produced and implemented from the top-down. For another key element of the information society has been the shift in who is producing the information at the heart of our society. Increasingly, such information is being produced by amateurs rather than professionals and shared openly through a variety of online platforms, or what has been described as the shift from Web 1.0 to Web 2.0. And as that information is increasingly tied to particular places through the practice of geotagging, it opens up new opportunities for citizens to produce their own digital representations of place through what is known as 'volunteered geographic information' (Goodchild, 2007). This layer of virtual information is a key element in the production of hybrid landscapes – blurring the boundaries of the virtual and material – in which the understandings of the material world are influenced by representations in the virtual world, which in turn influence how one lives in and reproduces those same spaces. While Graham and Zook (2011) outline a methodology for understanding these layers of place-based information and interrogating the ways they do and do not reflect sociospatial processes offline, these methods have similarly been applied to understand topics as diverse as disaster response (Crutcher and Zook, 2009) and religious landscapes (Shelton et al., 2012).

When collected, aggregated, and cross-referenced with other large databases of information, this user-generated content becomes one part of a mass of so-called big data, providing potentially unprecedented insights into social interactions. While big data offers researchers a rich source of contextual data from which to work, equally interesting work has emerged on the social conditions giving rise to the phenomenon of big data and how it is shaping academic research and modes of thinking, as well as the way we as citizens and consumers relate to the public and private sector institutions we are dependent upon (cf Crampton et al., 2013). And while big data remains a nascent and contested idea with a variety of possible futures, it has the potential to reconfigure social and spatial relations in new and interesting ways, just as other rounds of technological innovation in the information society have done.

See also: Citizen Mapping; Geographic Information Systems; Location, Absolute and Relative; Networks, Geography of; Surveillance and Privacy, Geography of; Time-Space in Geography; Urban Geography.

Bibliography

Adams, Paul C., 1996. Protest and the scale politics of telecommunications. Political Geography 15 (5), 419–441.

Adams, Paul C., Ghose, Rina, 2003. India.com: the construction of a space between. Progress in Human Geography 27 (4), 414–437.

Aoyama, Yuko, Sheppard, Eric, 2003. The dialectics of geographic and virtual space. Environment and Planning A 35 (7), 1151–1156.

Bell, Daniel, 1973. The Coming of Post-Industrial Society. Basic Books, New York.

Bluestone, Barry, Harrison, Bennett, 1982. The Deindustrialization of America: Plant Closings, Community Abandonment, and the Dismantling of Basic Industry. Basic Books, New York.

Cairncross, Frances, 1997. The Death of Distance: How the Communications Revolution Will Change Our Lives. Harvard Business Press, Boston, MA.

Castells, Manuel, 1996. The Rise of the Network Society. Wiley-Blackwell, Oxford.

Crampton, Jeremy W., Graham, Mark, Poorthuis, Ate, Shelton, Taylor, Stephens, Monica, Wilson, Matthew W., Zook, Matthew A., 2013. Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. Cartography and Geographic Information Science 40 (2), 130–139.

Crutcher, Michael, Zook, Matthew, 2009. Placemarks and waterlines: Racialized cyberscapes in post-Katrina Google Earth. Geoforum 40 (4), 523–534.

Fischer, Claude, 1994. America Calling: A Social History of the Telephone to 1940. University of California Press, Berkeley.

Florida, Richard, 2003. The Rise of the Creative Class: And How It's Transforming Work, Leisure, Community and Everyday Life. Basic Books, New York.

Froehling, Oliver, 1997. The cyberspace 'war of ink and internet' in Chiapas, Mexico. Geographical Review 87 (2), 291–307.

Goodchild, Michael, 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69 (4), 211–221.

Graham, Mark, 2010. Neogeography and the palimpsests of place: Web 2.0 and the construction of a virtual earth. Tijdschrift voor economische en sociale geografie 101 (4), 422–436.

Graham, Mark, 2013. Geography/internet: ethereal alternate dimensions of cyberspace or grounded augmented realities? The Geographical Journal 179 (2), 177–182.

Graham, Mark, Zook, Matthew, 2011. Visualizing global cyberscapes: mapping user-generated placemarks. Journal of Urban Technology 18 (1), 115–132.

Graham, Stephen, 2002. Bridging urban digital divides? Urban polarisation and information and communications technologies (ICTs). Urban Studies 39 (1), 33–56.

Harvey, David, 1989. The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change. Wiley-Blackwell, Oxford.

Hepworth, Mark, 1990. Geography of the Information Economy. Guilford Press, New York.

Kellerman, Aharon, 1993. Telecommunications and Geography. Belhaven Press, London.

Kirsch, Scott, 1995. The incredible shrinking world? Technology and the production of space. Environment and Planning D: Society and Space 13 (5), 529–555.

Kitchin, Rob, Dodge, Martin, 2011. Code/Space: Software and Everyday Life. The MIT Press, Cambridge, MA.

Malecki, Edward J., 2002. The economic geography of the Internet's infrastructure. Economic Geography 78 (4), 399–424.

Reich, Robert B., 1992. The Work of Nations: Preparing Ourselves for 21st Century Capitalism. Vintage Books, New York.

Roberts, Lynne D., Parks, Malcolm R., 1999. The social geography of gender-switching in virtual environments on the internet. Information, Communication & Society 2 (4), 521–540.

Shelton, Taylor, Zook, Matthew, Graham, Mark, 2012. The technology of religion: mapping religious cyberscapes. The Professional Geographer 64 (4), 602–617.

Thrift, Nigel, French, Shaun, 2002. The automatic production of space. Transactions of the Institute of British Geographers 27 (3), 309–335.

Touraine, Alain, 1971. The Post-Industrial Society: Tomorrow's Social History: Classes, Conflicts and Culture in the Programmed Society. Random House, New York.

Townsend, Anthony, 2001. The Internet and the rise of the new network cities, 1969–1999. Environment and Planning B: Planning and Design 28 (1), 39–58.

Warf, Barney, Grimes, John, 1997. Counterhegemonic discourses and the Internet. Geographical Review 87 (2), 259–274.

Zook, Matthew, 2005. The Geography of the Internet Industry. Blackwell, Oxford.

Zook, Matthew A., Graham, Mark, 2007. The creative reconstruction of the Internet: Google and the privatization of cyberspace and DigiPlace. Geoforum 38 (6), 1322–1343.

Zook, Matthew A., Shelton, Taylor, 2012. The integration of virtual flows into material movements within the global economy. In: Hall, Peter V., Hesse, Markus (Eds.), Cities, Regions and Flows. Routledge, New York, pp. 42–57.

Information Structure in Linguistics

Dejan Matić, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Information structure is a subfield of linguistic research dealing with the ways speakers encode instructions to the hearer on how to process the message relative to their temporary mental states. To this end, sentences are segmented into parts conveying known and yet-unknown information, usually labeled 'topic' and 'focus.' Many languages have developed specialized grammatical and lexical means of indicating this segmentation.

Theoretical Background

The term *information structure* refers to the ways linguistically encoded information is presented relative to the speaker's estimate of the temporary mental state of the receiver of the message (cf Chafe, 1976). Utterances transmit both the information contained in the message and the implicit or explicit instructions on how this information is to be processed and integrated into the hearer's knowledge stock. In order to achieve this, the speaker has to decide which parts of the sentence are 'old' or 'given' and which are 'new' for the hearer. The hearer is led to identify those elements of her existing knowledge ('given elements') which shall be relevant for the processing of the message. The information comes about by relating the 'new' elements (i.e., what the hearer is assumed not to be aware of) to these 'given' elements. The speaker's choice of 'given' and 'new' segments within a sentence depends on her hypotheses about the current state of the hearer's attention and consciousness. Information structure thus relates two major functions of language, as a means to transmit knowledge and as a vehicle of social interaction. Linguistic research on information structure is based on the assumption that natural languages are equipped with formal means of signaling the basic distinction between known and unknown pieces of information, and a number of other distinctions. This research draws from a rich philosophical and psychological tradition and has been established as a separate field of linguistics in the past 50 years.

The way information structure is rendered overt in natural languages is illustrated in the sentence pairs (1)–(2) (small capitals mark the position of the sentence stress).

(1a) Jonathan opened the DOOR. (English)
(1b) JONATHAN opened the door.
(2a) Peter hat gestern das FAHRRAD genommen. (German)
P. has yesterday the bicycle taken
'Peter took the BICYCLE yesterday.'
(2b) Das Fahrrad hat gestern PETER genommen.
the bicycle has yesterday P. taken
"As for the bicycle, PETER took it yesterday."

In (1a) and (1b), the sentence stress indicates what the hearer is to treat as new elements, the locus of information update. This is clearly apparent if one observes different context compatibilities of (1a) and (1b): the former is a felicitous answer to the question "What did Jonathan open?", while the latter answers the question "Who opened the door?". The carrier of sentence stress in English thus seems to indicate what element(s) of the clause carry the desired information update, as it is this part of the sentence that is regularly used to fill the gap in the hearer's knowledge marked by the question word. This kind of element, which brings about information update, is usually called *focus*. Sentences (2a) and (2b) illustrate how 'given' elements work. In German, given elements are usually marked by the clause-initial position. The normal use of these sentences in context reveals the differences in their information structure: while (2a) is fine in a situation in which the interlocutors are talking about Peter and his mischiefs, (2b) would be used if the conversation is about the whereabouts of the bicycle. In the former case, 'Peter' is the segment of the message toward which the attention of the interlocutors is directed; in the latter, this segment is 'the bicycle.' Given elements to which the hearer is expected to relate the information update are called *topic*. Importantly, both examples (1) and (2) show that the information conveyed by the message is intended to increase the hearer's knowledge about the topic by ascribing the denotation of the focus to this topic. In (2b), for instance, the speaker intends to inform the hearer about the bicycle (topic), and does so by ascribing it the property of being taken away by Peter (focus).

This idea of information structure is rooted in a particular model of communication, the incremental model (Stalnaker, 1999). According to this model, communication consists of reducing the differences in the knowledge of the interlocutors by increasing the *common ground* between them, i.e., their stock of shared knowledge. In order to do this, constraints on the input to the common ground have to be taken into account: only such content can be added which relates to the previously existing knowledge. Since the particular portion of the existing knowledge relevant at the current point in communication is not necessarily manifest per se, it often needs to be indicated linguistically, via *presuppositions*. The content that is added, the proposed change in the output common ground, is related to presuppositions via *assertion*. Similarly, other speech acts have effects on common ground, so that, for instance, questions identify which type of content is expected to be added to which presupposed portion of knowledge. By structuring the transmitted information, speakers indicate how this

permanent change of common ground should develop, so that information structure itself has been occasionally defined as a *common ground management* system (Krifka, 2008). Information structure research assumes that this common ground management system operates via a number of linguistic categories. These minimally include topic and focus, but some approaches differentiate no less than half a dozen of categories (see the 'Categories of Information Structure' section).

Categories of Information Structure

Topic (also called *theme*, *link*, *given information*, etc.) has been defined in a number of ways, depending on the general model of information structure. Two major competing definitions are those based on the notions of *givenness* and *aboutness*. Givenness-based definitions are hearer-centered: topic is that part of the utterance that is assumed to be already known to the hearer, present in the common ground of the interlocutors, and/or activated in the hearer's short-term memory, by being mentioned previously, inferable, or given in the extralinguistic context. The alternative view is that topic is that part of the utterance about which this utterance is meant to give information. The focus is here more on the speaker's intentions than on the hearer's state of mind: the speaker determines what she intends to increase the hearer's knowledge about and encodes this element as a topic. As shown by Reinhart (1981), the aboutness-based definition is empirically superior, even though the notion of givenness cannot be fully excluded from definitions of topicality. Speakers seem to be free to choose what segments of the hearer's knowledge they intend to enrich with new content, so as to increase the common ground. However, they are constrained by the considerations of the hearer's processing capabilities. What is chosen to be a topic of a given utterance is generally in one way or another given/old and thus easily accessible to the hearer; otherwise, the hearer is confronted with the double task of identifying inaccessible, inactive knowledge and adding new content to that segment of their knowledge. This usually leads to a breakdown of communication, or at least to infelicitous utterances, as seen in examples in (3).

Context: Why didn't you come to work yesterday?

(3a) I have some elephants I take care of. Yesterday, one little elephant was ill.
(3b) #Yesterday, one little elephant was ill.

The second clause in (3a) is felicitous, since its topic, 'one little elephant,' has been indirectly introduced previously and thus made accessible to the hearer. The answer in (3b) is communicatively bad, since it forces the hearer to accommodate a rather unusual piece of knowledge (the speaker takes care of elephants) and to ascribe it a property of being ill at the same time. Topics are thus elements of the proposition that the utterance is construed about, and that are usually restricted to given, accessible elements.

Focus has been defined in even more different ways, and has been labeled even more variably; other, partly or mostly synonymous labels include *comment*, *rheme*, *new information*, etc. Intuitively, focus is the locus of the common ground update, the element that carries the proffered new content. The two most prominent theoretical approaches to focus are Alternative Semantics and the assertion-based theory of focus; other approaches include Givenness Theory, Structured Meanings Theory, etc. (see Krifka, 2008 for an overview). The assertion-based approach to focus (Lambrecht, 1994) defines focus as that part of the proposition by which assertion differs from presupposition. Importantly, focus is not equated with assertion, as assertion, i.e., update, comes about only by applying the new content to the presupposed content. Focus is rather defined negatively, as the segment that is not presupposed. This definition relies heavily on the incremental model of communication and treats focus as a propositional phenomenon. Alternative Semantics (Rooth, 1992) attempts to develop a denotational semantics for focus, treating it as an operator which generates alternatives to the expression that is focused. Informally, focus indicates alternatives relevant for the interpretation of the focused expression. In processing a focused element, the hearer must assume that this element can be interpreted only with respect to some set of alternative denotations. This is illustrated in (4) and (5).

(4) Who broke the vase? $[\textsc{Peter}]_{focus}$ broke it.
(5) They were quarreling about buying a present. In the end, $[\textsc{Mary}]_{focus}$ bought it.

The question in (4) indicates that there are a number of possible culprits for the destruction of the vase; the focus of the answer, 'Peter,' picks out one of these referents and, by evoking alternative values (e.g., 'John,' 'Mary,' etc.), relates it to the set induced by the question. The same mechanism is at work in (5): the focus on 'Mary' indicates that there were other potential present-buyers.

The assertion-based approach and Alternative Semantics are not mutually exclusive. As shown by Matić and Wedgwood (2013), alternatives necessarily arise whenever something is newsworthy enough to be asserted. Assertion is, as it were, plausible only when things could have been otherwise. However, some important differences remain: while assertion-based theories allow for focusing only at the level of propositions, nothing in Alternative Semantics prevents smaller linguistic entities, such as noun phrases, from having information structure of their own. The two major approaches to focus thus differ in the assumed scope of information structure. Some researchers, mostly those that assume some version of Alternative Semantics, take information structure to be recursive, so that not only the sentence is segmented into topic and focus, but also its component parts (e.g., Rooth, 1992). Others restrict information structure to propositional entities, since only propositions can have truth values and contribute to the update of the common ground (e.g., Lambrecht, 1994). The issue is not fully resolved, and it is sometimes conflated with yet another unresolved aspect of information structure research, that of the number of its dimensions. Researchers differ with respect to the number and kinds of subdivisions and categories involved in informational segmentation of the utterance. The minimal assumption is that all sentences fall into two parts, the topical and the focal one. In some approaches, it is assumed that in addition to topic and focus, there is a transitional informational element that indicates how new content (focus) is to be related to the topic (*transition*

in Firbas, 1992; *tail* in Vallduví and Engdahl, 1996). Other approaches take it that different dimensions of information structure are realized in different structural domains. It is thus often claimed, with different terminologies, that there are two different levels: the topic–comment and the focus-background articulation, which are orthogonal to one another (e.g., Halliday, 1967; Vallduví and Vilkuna, 1998; Kruijff-Korbayová and Steedman, 2003). The topic–comment division corresponds roughly to what assertion-based theories of focus describe as information structure, bringing new content (comment) into relationship with a piece of the existing knowledge (topic), while the focus background partition expresses the existence of alternatives for focus and the lack thereof for background. The former is a sentence-level phenomenon, whereas the latter is recursive. The way these various divisions function is illustrated by example (6), adapted from Kruijff-Korbayová and Steedman (2003).

6. Q: I know that this car is a Porsche. But what is the make of your other car?
   A: (My OTHER car) (is ALSO a Porsche)
   (Background Focus) (Background Focus Background)
   Topic **Comment**

In other systems, what is labeled as topic and comment in (6) would be called topic and focus, and no other subdivisions would be acknowledged; or the comment in (6) would be subdivided into focus ('also') and transition, or tail ('is a Porsche').

Apart from these basic categories, subdivisions, and dimensions, a number of other categories are occasionally mentioned. The most persistently discussed one is *contrast* (Repp, 2010). It is occasionally equated with focus, especially in the form given to the latter in Alternative Semantics (Rooth, 1992); or it is assumed that it is a separate category of information structure, compatible with both topics and foci (Vallduví and Vilkuna, 1998; Molnár, 2002). There are opposing views, too: ever since Bolinger (1961) it has often been argued that contrast is merely a matter of pragmatic inference, not a separate linguistic entity (Zimmermann, 2008).

Research Traditions

The roots of information structure research reach deep into antiquity (good historical overviews are provided by Seuren, 1998; von Heusinger, 1999). The basic concept of the division of sentences into two parts, subject and predicate, goes back to Aristotle, who used it somewhat ambiguously to refer to the dichotomy of judgment at the logical, psychological, and grammatical levels. The rise of linguistics and modern logic in the nineteenth century brought the revival of the subject–predicate discussion. Linguists like Georg von der Gabelentz and Hermann Paul developed a theory of *psychological subject* and *psychological predicate*, applying the traditional Aristotelian dichotomy to the temporary psychological states of the interlocutors. They were also the first to notice the interconnectedness of information structure and discourse context, as shown by Paul's use of the question–answer test (see the analysis of example (1)) to show different configurations of subject and predicate. The growth of Gestalt psychology at the turn of the centuries, with its insistence on the perceptual dichotomy between *figure* and *ground*, triggered further interest in information structure. The idea that human subjects are capable of understanding foregrounded objects only by relating them to their background naturally relates to the theory of psychological subjects and predicates. The decisive move from psychology to linguistics was undertaken by the linguists of the Prague School, especially Vilém Mathesius, who used the categories derived from psychology and philosophy to account for phenomena of word order variation and prosody. The subject–predicate division was replaced by that of theme and rheme (later topic and comment, or topic and focus). The Praguean ideas of information structure were disseminated in the wider linguistic community through the work of Halliday (1967–68), who modified and refined the notion of theme–rheme partition. On his view, information structure (the term was coined by Halliday) is a component of grammar separate from syntax and semantics, but it interacts with both in a number of complex ways.

From the early 1970s, information structure has become an integral part of many grammatical theories and a frequent research topic in descriptive linguistics. Chafe (1976) has developed a framework in which many of the notions discussed above have been systematized for the first time. His work has spawned a number of approaches which share the view that information structure needs to be linked to the communicative and psychological reality of language users, no matter whether it is considered a proper part of grammar or a communicative, pragmatic phenomenon influencing grammar (e.g., Vallduví, 1992; Lambrecht, 1994; Van Valin, 2005). Important developments in this line of research are Vallduví's (1992) application of file change semantics to information-structural phenomena, where knowledge is conceived of as a set of file cards which get activated and deactivated, and Lambrecht's (1994) explicit embedding of information structure in the Stalnakerian model of communication. Another line of research was conceived in the generative framework, most notably by Jackendoff (1972). The principal purpose is to find a way how categories like topic and focus can be represented in grammatical description so as to account for the range of grammatical structures influenced or triggered by information structure in a maximally economical way. A device that has been used almost universally to achieve this aim is the representation of information-structural categories as grammatical features (F-feature for focus was introduced by Jackendoff himself) which trigger word order permutations and determine sentence stress assignment and similar phenomena. Further developments include the postulation of dedicated hierarchical positions for topic and focus (Rizzi, 1997) and optimality-theoretical accounts of the interaction of the focus feature with sentence structure (Büring, 2006). In recent years, important attempts have been made to formalize the relationship between discourse structure and information structure (Roberts, 2012). The basic idea is that the discourse develops through a series of implicit questions under discussion. The information structure relates the utterances to these underlying questions and thus renders the discourse structure transparent.

Information Structure in Grammar

The linguistic interest in the phenomena of information structure stems from the fact that many formal features of language, such as prosody or word order, are intimately connected with information-structural variation. It is thus generally agreed that the assignment of sentence stress in many languages is determined by the position of focus (Ladd, 2006), as illustrated in examples (1a, b). Word order is also affected by information structure in many languages, either via optional rearrangement of constituents, as in German examples (2a, b), or via obligatory movement of information-structurally marked constituents to certain positions in the clause, as in Hungarian, where focus has to be immediately preverbal (compare the neutral sentence (7a) with the focus sentences in (7b) and (7c)). Other expressive means are also attested. Syntactic transformations of different kinds are found in most languages of the world; the best known example is cleft sentences, which in some languages, such as French, play a major role in signaling information structure (8). Many languages of the world have specialized morphology to indicate topic, focus, or some kind of contrast; the Japanese particle *wa* is a well-known example of a topic morpheme (9), while the case marker *–ləŋ* and the agreement suffix *–mələ* in Yukaghir (isolated language spoken in northern Siberia) is an instance of focus-indicating morphology (10).

(7a) János meg-ette az almát. (Hungarian)
J. PERFECTIVE-ate the apple.
"János ate the apple."

(7b) János ette meg az almát.
J. ate PERFECTIVE the apple
"JÁNOS ate the apple."

(7c) János az almát ette meg.
J. the apple ate PERFECTIVE
"János ate the APPLE."

(8) C'est maman qui décide. vs. Maman décide. (French)
it's mother who decides mother decides
"It's MOTHER that decides." vs. 'Mother DECIDES.'

(9) Inu wa hasitte iru.
dog TOP running is
"The dog is RUNNING." (Japanese)

(10) Taŋ köde metin čoγojə-ləŋ tadi-mələ
that man to.me knife-FOC give-OBJECT.FOC.3SG
"that man gave me a KNIFE." (Yukaghir)

It is unclear whether it is possible to establish cross-linguistic regularities which would subsume all these different types of structures. Attempts to establish one universal information-structure-based sentence template out of which variable systems could be derived have not been successful (see Matić and Wedgwood, 2013 for a critique). Recently, there have been noteworthy attempts to reduce the number of possible information-structure types to a limited number (Van Valin, 1999) and to derive the cross-linguistic variability from some more fundamental processing strategies (Büring, 2010), but all these attempts need further elaboration. Irrespective of this, it has been shown that, in one way or another, information structure does play an important role not only in such patent cases as word order variation or prosody, but also in reference resolution, long-distance dependencies, extraction, etc. (Van Valin, 2005, Erteschik-Shir, 2007).

The question of regularities that govern the form of information-structure-related structures in the languages of the world is not the only debated issue. The position of information structure in grammar is also a matter of dispute. The dispute revolves around two interconnected topics. First, is the proper domain of information structure semantics or pragmatics? Second, how is it connected to grammar? The answer to both questions depends on the definition of the semantics/pragmatics distinction one adopts. On one view, semantics is only what is truth-conditionally relevant. In most cases, topic and focus have no truth-conditional relevance whatsoever, so that they can be considered a proper subdomain of pragmatics. However, in some cases, they do seem to trigger truth-conditional effects, as in some kinds of generics and in certain types of conditionals. The best known examples include focus-sensitive expressions, whose interpretation seems to fully depend on focus assignment, so that they are considered *associated with focus* (Beaver and Clark, 2008). Such expressions are, among others, many particles, such as *only*, *also*, *even*, etc. The truth-conditional impact of their association with different foci is illustrated in (11).

(11a) John only likes BANANAS (and nothing else).
(11b) John only LIKES bananas (and has no other relationship to them).
(11c) Only JOHN likes bananas (and nobody else).

This property of information-structural categories has often been adduced as a proof of their semantic nature (e.g., Rooth, 1992). However, one might argue that the differences observed in (11) are a mere product of pragmatic inference arising out of the meaning of the particles in particular contexts. The issue needs further research.

A different idea of the semantics/pragmatics distinction is based on what is linguistically encoded vs what needs to be derived inferentially. From this point of view, the question is whether the structures identified as connected with topic, focus, etc., are indeed dedicated to expressing the meanings of topic, focus, etc., or whether these meanings are rather pragmatic effects derived from quite different denotations. The answer to this question is an empirical one: each information-structurally marked construction in a language needs to be tested for its primary denotation, and, if it turns out that this structure is not primarily an information-structural category, a pragmatic mechanism needs to be defined through which information-structural effects come about. Some first results largely diverge: Beaver and Clark (2008) convincingly argue for focus as the encoded meaning of the sentence stress in English, while Matić and Wedgwood (2013) show that many apparent focus-denoting structures, such as English clefts or Somali focus morphology, can be better analyzed as having distinct (identificational, modal, aspectual, etc.) denotations with interestingly similar information-structural effects. It thus seems that information structure can but need not be an element of grammar, depending on whether the given language has grammaticalized the categories of topic, focus,

etc., or not. Information structure is thus a universal phenomenon of human communication, but it is not necessarily a universal phenomenon in natural language.

See also: Communicative Competence: Linguistic Aspects; Intentionality in Language and Communication, Emergence of; Linguistic Presupposition; Linguistic Typology; Pragmatics, Linguistic; Semantics; Suprasegmentals; Word Order.

Bibliography

Beaver, D., Clark, B., 2008. Sense and Sensitivity. How Focus Determines Meaning. Wiley-Blackwell, Chichester.

Bolinger, D.L., 1961. Contrastive accent and contrastive stress. Language 37, 83–96.

Büring, D., 2006. Focus projection and default prominence. In: Molnár, V., Winkler, S. (Eds.), The Architecture of Focus. De Gruyter, Berlin, pp. 321–346.

Büring, D., 2010. Towards a typology of focus realization. In: Zimmermann, M., Féry, C. (Eds.), Information Structure. Theoretical, Typological, and Experimental Perspectives. Oxford University Press, Oxford, pp. 177–205.

Chafe, W., 1976. Givenness, contrastiveness, definiteness, subjects and topics. In: Li, C.N. (Ed.), Subject and Topic. Academic Press, New York, pp. 27–55.

Erteschik-Shir, N., 2007. Information Structure. Oxford University Press, Oxford.

Firbas, J., 1992. Functional Sentence Perspective in Written and Spoken Communication. Cambridge University Press, Cambridge.

Halliday, M.A.K., 1967–8. Notes on transitivity and theme in English. Journal of Linguistics 3, 37–81, 199–244; 4, 179–215.

Heusinger von, K., 1999. Intonation and Information Structure. Habilitationsschrift. University of Konstanz.

Jackendoff, R., 1972. Semantic Interpretation in Generative Grammar. MIT Press, Cambridge, MA.

Krifka, M., 2008. Basic notions of information structure. Acta Linguistica Hungarica 55, 243–276.

Kruijff-Korbayová, I., Steedman, M., 2003. Discourse and information structure. Journal of Logic, Language and Information 12, 249–259.

Ladd, R.D., 2006. Intonational Phonology. Cambridge University Press, Cambridge.

Lambrecht, K., 1994. Information Structure and Sentence Form. Cambridge University Press, Cambridge.

Matić, D., Wedgwood, D., 2013. The meanings of focus: the significance of an interpretation-based category in cross-linguistic analysis. Journal of Linguistics 49, 127–163.

Molnár, V., 2002. Contrast from a contrastive perspective. Language and Computers 39, 147–162.

Reinhart, T., 1981. Pragmatics and linguistics. An analysis of sentence topics. Philosophica 27, 53–94.

Repp, S., 2010. Defining 'contrast' as an information-structural notion in grammar. Lingua 120, 1333–1345.

Roberts, C., 2012. Information structure: towards an integrated formal theory of pragmatics. Semantics and Pragmatics 5, 1–69.

Rooth, M., 1992. A theory of focus interpretation. Natural Language Semantics 1, 75–116.

Rizzi, L., 1997. The fine structure of the left periphery. In: Haegeman, L. (Ed.), Elements of Grammar. A Handbook in Generative Syntax. Kluwer, Dordrecht, pp. 281–337.

Seuren, P., 1998. Western Linguistics. An Historical Introduction. Blackwell, Oxford.

Stalnaker, R.C., 1999. Context and Content. Essays on Intentionality in Speech and Thought. Oxford University Press, New York.

Vallduví, E., 1992. The Informational Component. Garland Publishers, New York/London.

Vallduví, E., Engdahl, E., 1996. The linguistic realization of information packaging and grammar. Linguistics 34, 459–519.

Vallduví, E., Vilkuna, M., 1998. On rheme and kontrast. In: Culicover, P., McNally, L. (Eds.), The Limits of Syntax. Academic Press, San Diego, pp. 79–108.

Van Valin, R.D., 1999. A typology of the interaction of focus structure and syntax. In: Raxilina, E., Testelec, J. (Eds.), Tipologija i teorija jazyka. Jazyki russkoj kul'tury, Moscow, pp. 511–524.

Van Valin, R.D., 2005. Exploring the Syntax-Semantics Interface. Cambridge University Press, Cambridge.

Zimmermann, M., 2008. Contrastive focus and Emphasis. Acta Linguistica Hungarica 55, 347–360.

Information, Economics of

SS Lippman and John J McCall, University of California, Los Angeles, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7480–7486, © 2001, Elsevier Ltd.

Abstract

This article assesses the current status of the economics of information. The focus is on the acquisition of information (search) and the incentives associated with asymmetric information (noncooperative game theory). The evolution of these two areas also is presented along with simple examples. Their evolutions have a similar structure. The formal, abstract development of search is separate from its informal, concrete growth. At critical points the two paths intersect. The unification and illumination generated by these sudden meetings propel the economics of search to higher levels of understanding. In the economics of asymmetric information, game theory comprises the formal trajectory and industrial organization, broadly construed, is the relatively informal path. Econometric methods, devised shortly after an intersection, have exploited the empirical aspects of the new search paradigm. Experimental economics is performing an analogous function in the economics of asymmetric information. The article concludes with predictions regarding future developments.

Introduction

Prior to the 1980s, the 'economics of information' usually meant the economics of search. Search began in the 1960s with the seminal research by George Stigler, which spawned an enormous and growing literature. Both search and noncooperative game theory are considered in this article. As we will see, they are closely related and much can be gained by fostering a closer relation.

Stigler modified competitive price theory by noting that price is a random variable with a given cumulative distribution function. He thereby enlarged price theory to include the search activity necessary to gather information about prices. He defined search as follows: "A buyer (or seller) who wishes to ascertain the most favorable price must canvas various sellers (or buyers) – a phenomenon I shall term 'search'." The author of the famous text *The Theory of Price* knew that search theory would receive widespread attention because of its intimate link with price. That is, the primary reason for the almost immediate influence of Stigler's work was his focus on price uncertainty. Any such major change in price theory eventually would reverberate throughout all of economics. This is what has occurred since the 1960s.

The game theory meaning of the 'economics of information' centers around the formation, continuation, and dissolution of contractual relationships. The incentives of parties to a contract are determined by the different information possessed by each. To exemplify these asymmetric information problems, consider an auction of a valuable object with the following information structure. There are two buyers and one seller. One buyer is willing to pay more for the object, while the other has more information about the object's relative price. If the buyers can collude, "the Nash program directs us to view this collusion as the outcome of a process of communication where each player has choices about what to say. In this process, the more-informed buyer may try to mislead the less-informed buyer about the value of the object, and the less-informed buyer should rationally take this possibility into account" (Myerson, 1999: p. 1075). All problems of incentive-based 'economics of information' can be formulated as Nash noncooperative games.

While Stigler's information model fits snugly into standard price theory, Nash's noncooperative game theory complements price theory, but is not reducible to it. Stigler's fit is snug because he takes the distribution of prices as exogenously determined. This corresponds to the basic assumption of price theory – consumers accept prices as exogenous. Thus, the equilibrium problem is not addressed. Nash enlarged the scope of price theory so interactive behavior can be studied in almost any social institution. This was apparent only after the contributions by Harsanyi, Selten, Kreps et al., thereby explaining the long period between Nash's discovery and its adoption by the economics profession.

In recent years these two research programs have started to merge. Search is embedded in Abraham Wald's sequential analysis in that it determines when the process of information collection should terminate. Gideon Schwarz (1994) notes that Wald used the theory of two-person zero-sum games to unify statistical inference. Statistical inference was viewed as a game played by the statistician against nature. Statistics from this game theory perspective was called decision theory. One of the most important topics in decision theory is optimal stopping. Thus, game theory is as applicable to search as it is to other informational problems in economics.

Search Theory

Introduction

Modern search theory solves the problem of when to terminate the collection of information and take an action. Search is the confluence of two seemingly unrelated intellectual streams: the formal evolution of optimal stopping rules beginning with Cayley (1875) and the verbal, common sense literature found in Knight (1921) and implicit in Adam Smith. The first stage of this confluence occurred when Stigler (1961) blended these two streams with his fixed-sized-sampling rule. The second stage replaced the fixed-sized sampling rule with an optimal stopping rule, thereby combining the insights of Stigler, Bellman, and Wald.

The third stage seeks the source of an endogenous distribution of prices (wages) to replace the exogenous distribution in the Stigler and sequential models. This large and complex research effort was initiated by Rothschild (1973). The fourth stage replaced the one-sided sequential model with a two-sided

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71042-7

equilibrium problem, where there is competition and interaction among optimizing agents. This ongoing research is based on the matching and bargaining models devised by Diamond, Jovanovic, Maskin, Mortensen, and Pissarides.

In the early work on search, many economists placed a much higher value on the verbal insights of common sense, when they sought a solution to a search perplex. Mathematical formulations of these insights were undervalued. Today, the situation is somewhat reversed in that formal analysis is applauded, while common sense is depreciated. Both attitudes diminish economics. The first does not encapsulate insights in a formal model. Hence, testing and extensions are difficult if not impossible. The second closes off economics from the imaginative intuitions of common sense.

Intuitive, Common Sense Precursors to the Economics of Search

See Lippman and McCall (1993) for further detail.

In 1930, Cannan clearly saw the *moral hazard* problem associated with unemployment insurance. He also noted that the best strategy for a job-seeker was to "Take what you can get now or hold out another week, when something better may turn up." With the advantage of hindsight this looks very much like the reservation wage policy.

In 1932, Hicks recognized the unanticipated unemployment effects of the search which is connected to the presence of imperfect information, when he claims that firms' "knowledge of opportunities is imperfect [whence] it is not surprising that an interval of time elapses between dismissal and reengagement during which the workman is unemployed." His statement that "the unemployment of the man who gives up his job in order to look for a better one" recognizes that voluntary unemployment is joined with search as a productive activity.

An early explicit statement viewing search unemployment as a productive activity was made by Hutt in 1939.

> A worker … may refuse immediately available work…because he feels that to accept it will prevent him from seeking for better openings. …. When actively searching for work … he is really investing in himself …. He is doing what he would pay an efficient employment agency to do – the search for a better opening is worth the risk of immediately foregone income.
>
> Hutt 1977: p. 83

In addition to the labor literature anticipating the job search model, other prominent precursors in the economics of search include Knight, Schumpeter, and Alchian who were concerned with the search for natural resources, the search for inventions, and evolutionary economics. Knight (1921) was keenly aware of the ancient activity of prospecting for natural resources "where the possibility of securing wealth by the discovery of natural resources is known … resources will be attracted into the field of searching for them in accordance with men's estimates of the chances of success in relation to the outlays to be incurred."

Schumpeter (1942) continued the advance toward Stigler's introduction of search in his concern for uncertainty and the opportunistic contractual behavior which uncertainty fosters.

> Practically any investment entails, as a necessary complement of entrepreneurial action, certain safeguarding activities such as insuring and hedging. Long range investing under rapidly changing conditions … is like shooting at a target that is not only indistinct but moving – and moving jerkily at that. Hence it becomes necessary to resort to such protecting devices as patents or temporary secrecy of processes or, in some cases, long-period contracts secured in advance.

In Alchian's (1951) evolutionary view, innovations may result from bungled imitation as well as from conscious efforts to improve. He considers innovation, like prospecting for natural resources, as a form of search.

It should also be noted that Simon (1951) suggested a stochastic theory of the employment relation resembling Wald's sequential analysis rule. In this remarkable paper, Simon also studied credible threats and promises, reputation and cooperation in a pregame theory environment. He thereby suggested a connection between the two meanings of the 'economics of information.'

No one contributed more to the informal literature on the importance of information to economic decisions than Hayek. While not specifying an explicit search rule, Hayek (1978) clearly saw the importance of the concrete knowledge and special decisions made by millions of dispersed individuals. The key which facilitated the transition from *particular* knowledge to a *specific* decision was the price system:

> The recognition that a man's efforts will benefit more people and on the whole satisfy greater needs, when he lets himself be guided by the abstract signals of prices rather than by perceived needs, and that by this method we can best overcome our constitutional ignorance of most of the particular facts, and can make the fullest use of the knowledge of concrete circumstances widely dispersed among millions of individuals, is the great achievement of Adam Smith.

Thus for Hayek the crucial economic problem is the search for and use of dispersed knowledge whose common source is the price system.

The evolution of job search reveals how useful economic models are generated by *both* formal mathematics and common sense. The origin of the mathematical strand may be quite remote from practical applications, while the original intuitive insight may have little if any formal content. Nevertheless, as both strands evolve a threshold emerges. When this threshold is hit, a moment of insight occurs which unifies the two strands into a new economic model, illuminating a whole set of previously invisible phenomena. The success of the new approach is gauged by empirical data analyzed by novel econometric procedures that are frequently novel.

The Formal Development of Search Theory

See Lippman and McCall (1993) for further detail.

Stopping rules emerged from a very practical problem: Johannes Kepler's search for a wife in 1613. His procedure was

a version of the secretary problem, which was formally designed by Arthur Cayley in 1875 and rejuvenated more precisely by Leo Moser in 1956. (See Ferguson (1989) for a scholarly treatment of the secretary problem's evolution.)

In 1961, D.V. Lindley solved the following version of the secretary problem. The decision maker selects from n candidates who are presented in random order, with no recall of rejected candidates. The ability of the candidate is revealed at the presentation. This information tells the decision maker nothing about the abilities of the remaining candidates. The payoff function is 1, if the best secretary is chosen, and 0, otherwise. The optimal policy for large n is to interview and reject n/e of the candidates, and then choose the next candidate with the highest rank observed. The probability of choosing the best candidate is $1/e$, where e equals 2.718.

In 1960, MacQueen and Miller posed and solved the house-hunting problem. Independent, identically distributed offers $X_1, X_2, \ldots$ arrive one at a time. The return from stopping after n observations is $R_n = \max(X_1, \ldots, X_n) - nc$. MacQueen and Miller found the stopping rule N which maximized ER_N, where E is the expectation operator.

In 1961, Stigler formulated his fixed sample size model. The buyer obtains n price quotations for the item he wishes to purchase. A cost c is incurred each time a quote is made. The mathematical formulation of Stigler search is: choose the sample size n to minimize $E\min\{X_1, \ldots, X_n\} - nc$.

In 1970, the house-hunting problem was applied to job search in a simple model by McCall. The fixed sample size of Stigler's model is replaced by a stopping time N and the sequential decision model of search is given by: select a stopping time N to maximize $E\max\{X_1, \ldots, X_n\} - Nc$.

The optimal stopping time is: accept the first offer which exceeds the reservation wage ξ. The reservation wage ξ is the solution to the first-order condition

$$c = \int_x^\infty (y - x)\mathrm{d}F(x) \equiv H(x) \qquad [1]$$

where $H(x)$, the marginal gain from one more offer, is strictly decreasing and convex. The first-order condition has ξ as its unique solution since $H' < 0$. The economic interpretation of the first-order condition is: equate the marginal cost c of taking an additional offer with the expected increase in the return from taking one more search. It has been shown that ξ is the expected return from search (see Lippman and McCall, 1976). The time until an acceptable offer is obtained, a period of frictional unemployment, is a geometric random variable with mean $[1 - F(\xi)]^{-1}$.

Equilibrium Price (Wage) Dispersion

Rothschild (1973) raised the important question of the existence of an equilibrium price dispersion, namely, what gives rise to the nondegenerate price distribution assumed in simple search models. This question provoked an enormous literature on equilibrium search models which sought a positive answer to Rothschild, thereby negating Diamond's (1971) finding that all sellers charge the monopoly price. In Diamond's model, buyers are identical, sellers are identical, and buyers have the same search cost. If the rules are reversed – buyers instead of sellers make take-it-or-leave-it offers and sellers incur search costs – then markets clear at the competitive price. The preferred way out of this dilemma is to introduce heterogeneity into the model. In McMillan and Rothschild (1993), there are several sources of heterogeneity among the 31 papers surveyed. They include differences in sellers' production costs, buyers' search costs, buyers' beliefs about the offer distribution, sellers' time varying prices, buyers' inventory holdings, and advertising which yields differential buyer information. Burdett and Mortensen (1998) obtain a unique equilibrium solution to a game where a continuum of employers establish wage offers and a continuum of workers search sequentially. If workers search when unemployed *and* employed, wage dispersion is achieved.

Two-Sided Search and Bargaining

Since the early 1980s, the early one-sided optimization search models have been generalized to two-sided equilibrium search models. This research combines game theory with search theory by introducing bargaining into the formation of matches. The steady-state equilibrium is the intersection of best reply functions for workers and employers, i.e., a Nash equilibrium, *when utility is nontransferable*. If utility is perfectly transferable, the matched pair bargains until either an agreement or an impasse is reached. The Nash solution is the equilibrium of this cooperative game. See Burdett and Wright (1998).

McMillan and Rothschild (1993) note that the joint study of bargaining and search fortifies each. In the search theory used by Diamond (1971) and in the price dispersion models described above, after a seller is located by the searching buyer, the seller presents a take-it-or-leave-it price offer. But in many circumstances these offers lack credibility, i.e., the seller has an incentive to bargain if his offer is rejected by the buyer. In bargaining theory, it is usually assumed that a single buyer and a single seller are locked in their negotiations. In fact, each often has the option of searching for another trading partner.

Econometric Methods

The sequential search model gave rise to a sequence of novel econometric methods beginning with the work of Heckman and Singer; and Kiefer and Neumann. Structural estimation has been developed to estimate sophisticated models that have grown from the simple sequential model.

The impetus for structural estimation was Lucas' insight that reliance on reduced-form estimation in policy analysis gives rise to systematic forecasting errors. This insight is a consequence of the rational expectations hypothesis, which is a version of the Nash equilibrium. Thus, the estimation of search models relies on the pillar of contract and incentive models. Indeed, most structural estimation has been conducted within Markov decision processes. These processes are natural for modeling stochastic sequential decision making. Control variables d_t and state variables s_t, $t = 1, \ldots, T$ characterize the decision process. The individual is represented by three primitives: a utility function $u(s_t, d_t)$ denoting preferences at time t, a Markov transition probability $p(s_{t+1}|s_t, d_t)$ denoting the agent's beliefs about future states given current states and

decisions, and the discount factor β. The optimal decision rule δ is obtained via dynamic programming. See Gotz and McCall (1980) for an early application.

An error term ε_t is usually imbedded in the discrete decision process. This insures that the econometric model is nondegenerate. The structure of the error term is determined by the specifics of the problem. The error term is a fundamental element of the overall optimization, a remarkable rise in its stature compared to earlier econometric models. It is now considered an unobserved state variable, the usual state variable s_t being decomposed into (x_t, ε_t), where x_t is a subvector observed by the econometrician and ε is observed only by the agent.

Rust (1994) identifies practical and theoretical reasons for the recent interest in structural estimation. Theoretically, "structural estimation is the most direct way to assess the empirical validity of a specific Markov decision process model." On a practical plane, "structural models can generate more accurate predictions of the impact of policy changes than reduced-form models." This practical reason lies at the heart of the Lucas critique: reduced-form estimates are unreliable forecasts of policy changes because they ignore the impact of policy on the 'deep parameters' (u, p, β). Wolpin (1995) is an excellent survey of this literature by one of its founders.

The Effects of Labor Market Policies in Equilibrium Search Models

Over the past decade there has been an increasing interest in economic models that study the flow of jobs and workers among diverse states. Most of this research relies on equilibrium models and emphasizes the frictions in both search and recruiting. Using a rich set of data on flows in labor markets, this research studies job creation, job closure, unemployment, and labor mobility. Matching and equilibrium are crucial to the theoretical approach. Decisions made by employers on the demand side of the market are highlighted. A comprehensive review of this prodigious literature is Mortensen and Pissarides (1999).

Ljungqvist and Sargent (1995a,b) have devised sophisticated extensions of the simple sequential search model to study equilibrium unemployment behavior in European countries. In Sweden, they recommend a smoothing of the distorted incentives in the labor market by adopting less generous unemployment benefits. In European welfare states, their equilibrium search model finds that:

> when the generosity of unemployment insurance increases, the upward pressure on the unemployment rate eventually dominates the reduction in labor mobility due to high marginal income taxes. The higher unemployment rate is caused by unemployed workers reducing their search intensity and increasing their reservation rates. Higher unemployment pushes taxes upward which in turn increases the unemployment rate further, and so on.

Contracts, Incentives, and Asymmetric Information

Introduction

The past two decades of the twentieth century have witnessed dramatic changes in economic theory. Several years ago most graduate texts in economic theory were founded on price theory. Today, in the major economic departments, economic theory revolves around noncooperative game theory. Price theory has been relegated to the back seat. Before the ascendance of game theory, the emphasis in economics was on the usefulness of economic models in explaining empirical phenomena. While this emphasis persists, it has been overshadowed by the quest to apply strategic thinking to resolve economic problems.

The key ingredients of this remarkable transformation are: the Nash equilibrium (NE) concept, Harsanyi's characterization of contracts with asymmetric information as noncooperative games, and Selten's explicit consideration of time and the elimination of equilibria associated with noncredible threats and promises.

Nash's equilibrium is a strategy for each player where each strategy is a best response to the strategies of the other players in the *n*-person noncooperative game. This concept pervades economics. As a simple example, consider the traveler who drives a car first in the US and then in the UK. He drives on the right-hand side of the road in the US as does everyone else. When he visits the UK, his best response to British drivers is to drive on the left side. These Nash equilibria are akin to rational expectation's equilibria.

The NE is not immaculate. Jack Hirshleifer (private communication), sees

> two major problems with NE: (1) Each player's decision is supposed to be a 'best reply' to the opponent's corresponding choice. But the Nash protocol requires simultaneity. Thus, neither side knows the opponents' strategy to which it is supposed to be replying optimally. Without that knowledge, how can someone make a 'best reply'? … (2) … the 'best reply' has to be to the opponent's strategy, and not just to his observed actions or moves. In all but the very simplest cases, only a tiny fraction of a player's full strategy will ever be visible …. So in general the player can never know whether his current strategy is or is not a 'best reply' to what the opponent has in mind. … NE is fine for 'toy worlds' of our textbooks, but I remain skeptical of its general predictive validity.

van Damme and Weibull (1995) observe that "John Harsanyi showed that games with incomplete information could be remodeled as games with complete but imperfect information, thereby enabling analysis of this important class of games and providing a theoretical foundation for 'the economics of information'." Signaling, moral hazard, and adverse selection are prominent members of this class. A game is said to be one of *complete information* if Nature makes the first move and this move is observed by all players. A game has *perfect information* if each *information set* – the set of nodes in the tree such that one is known to be the actual node – is a singleton. (See Rasmusen (1989).) Harsanyi adopts a Bayesian approach assuming that each player may be of several types – a type specifies the information a player possesses regarding the game. The resulting refinement of Nash's equilibrium is called a Bayes–Nash equilibrium.

Reinhard Selten was the first to refine the Nash equilibrium for analysis of dynamic strategic interactions. Such refinement is necessary since many equilibria entail noncredible threats and do not make economic sense. Selten's formalization of the requirement that only credible threats should be considered, the concept of *subgame perfect equilibrium*, is used widely in the

industrial organization literature. It has generated significant insights there and in other fields of economics.

At first it was not clear how the problems of asymmetric information could be formulated as noncooperative games. Indeed, much significant research in signaling and insurance was performed in the 1970s with little reference to game theory. These models generated important economic implications, but were stymied by problems of equilibrium and credibility. This was changed suddenly when fundamental papers by Kreps and Wilson (1982) and others showed that this research could use the deep insights of Harsanyi and Selten to illuminate credibility and equilibrium. The ensuing unification of industrial organization was analogous to the epiphany that occurred in search theory when optimal stopping, dynamic programming, and matching were applied to search problems.

Principal-Agent Models

This section is a brief discussion of three of the most important asymmetric information models: moral hazard, adverse selection, and signaling. All three belong to the class of principal-agent models. In these models, the principal P designs the contract. The agent A either accepts or rejects the contract depending on the expected utility of the contract vis-à-vis the utility from other alternatives. Finally, the agent performs the task specified by the principal. The two parties are opposed in that the revenue for the agent is a cost for the principal and the effort of the agent is costly to him and beneficial to the agent. In determining the optimal contract between principal and agent, this opposition must be resolved. Macho-Stadler and Perez-Castrillo (1997) see this as "one of the most important objectives of the economics of information."

There are two parties to a P–A model. One is informed while the other does not know a crucial piece of information. For example, in the insurance P–A model, the insurance company is the principal and the insuree is the agent. It is usually assumed that the agent knows his health status, whereas the principal is uncertain. The P–A model is a bilateral monopoly. Hence, the nature of the bargaining process must be specified. For simplicity, it is assumed that either P or A receives all of the surplus. For example, P does this by stipulating a 'take-it-or-leave-it' contract, which the agent either accepts or rejects. Salanie (1997) notes that bargaining under asymmetric information is very complicated: "There is presently no consensus among theorists on what equilibrium concept should be used." Salanie observes that: "the P–A game is a Stackelberg game in which the leader (who proposes the contract) is called the principal and the follower (the party who just has to accept or reject the contract) is called the agent."

The actual bargaining in a concrete setting is unlikely to have this 0–1 structure, where the game terminates if either an agreement or disagreement happens. Instead, the bargaining may continue or, more likely, the disgruntled agent will search until he finds an acceptable contract. Such a model would combine the P–A analysis with search theory.

Three Principal-Agent Models with Asymmetric Information

Moral hazard occurs when the agent's actions are not observed by the principal. More specifically, moral hazard is present when the subject's actions are: (1) influenced by the terms of the contract, and (2) in a way that is not fully specified in the contract. These hidden actions are evident in automobile insurance contracts, where the driving behavior of the insuree is not known by the insurance company. The contract usually has a deductible to mitigate the risk to the insurance company.

Adverse selection occurs when the agent is privy to pertinent information before the contract is finalized. This information is hidden from the principal. Adverse selection is evident in health insurance contracts where certain characteristics (health status) of the agent are imperfectly known by the insurance company. An insurance contract is designed for each group with a particular health characteristic (sickly or healthy) so that each member of the group has an incentive to buy it.

Signaling occurs when one of the parties (the agent) to the contract has pertinent information regarding his type. Before entering the contract, the agent can signal the principal that he has this type, i.e., the behavior of the informed party conveys (signals) the type information to the uninformed party.

A key observation is that each of these asymmetric information models can be represented as a noncooperative game. The crucial equilibrium concept is a refinement of the Nash equilibrium. For example, the signaling game has a perfect Bayesian equilibrium. Excellent discussions are given in Gibbons (1992) and Macho-Stadler and Perez-Castrillo (1997).

One of the most important topics in the economics of information is the optimal design of contracts under symmetric and asymmetric information. In the symmetric case, the principal designs the contract so that the expected marginal payoff equals the marginal cost of effort. If the principal is risk neutral, then the optimal contract is one in which the principal accepts all the risk and the agent is, therefore, fully insured, receiving a payoff that is independent of the outcome. In the asymmetric case, the situation is much more complicated. The optimal contract balances the conflict between two opposing goals: efficiency in the allocation of risk between principal and agent and maintaining the incentives of the agent. For lucid presentations of the analysis, see Macho-Stadler and Perez-Castrillo (1997) and Salanie (1997).

Future Developments

It seems clear that the merger between search and game theory will continue. The benefits from this include: strategic thinking will become even more apparent in search models and the econometric testing of game theory models should develop into a cottage industry alongside experimental testing. Search models will continue to be used to evaluate labor policies like cross-country unemployment and other macro issues. Their use in other economic areas should also flourish.

In his evaluation of the applications of Nash equilibrium, Selten claims that we still know little about actual behavior and that empirical research and theories of bounded rationality are needed. Rubinstein (1995) is more emphatic. "Economic theory should deal with the real world. It is not a branch of abstract mathematics even though it utilizes mathematical tools. Since it is about the real world, people expect the theory to prove useful in achieving practical goals. But economic theory has not delivered the goods." Many would dispute the final sentence, but

most would agree that game-theoretic models require extensive econometric testing. This has already begun, with structural estimation being the favored method, but much remains. This is an exciting time for the economics of information.

Bibliography

Alchian, A.A., 1951. Uncertainty, evolution, and economic theory. Journal of Political Economy 58, 211–221.

Burdett, K., Mortensen, D.T., 1998. Wage differentials, employer size, and unemployment. International Economic Review 39, 257–273.

Burdett, K., Wright, R., 1998. Two-sided search with nontransferable utility. Review of Economic Dynamics 1, 220–245.

Cannan, E., 1930. Review of: the post-war employment problem by H. Clay. Economic Journal 40, 45–53.

Cayley, A., 1875. Mathematical questions and their solutions. Educational Times 22, 18–19.

van Damme, E., Weibull, J.W., 1995. Equilibrium in strategic interaction: the contributions of John C. Harsanyi, John F. Nash and Reinhard Selten. Scandinavian Journal of Economics 97, 15–40.

Diamond, P.A., 1971. A model of price adjustment. Journal of Economic Theory 3, 156–168.

Ferguson, T.S., 1989. Who solved the secretary problem? Statistical Science 4, 282–296.

Gibbons, R., 1992. Game Theory for Applied Economists. Princeton University Press, Princeton, NJ.

Gotz, G.A., McCall, J.J., 1980. Estimation in sequential decision making models. Economic Letters 6, 131–136.

Hayek, F., 1978. In: New Studies. University of Chicago Press, Chicago, p. 269.

Hicks, J.R., 1932. The Theory of Wages. MacMillan, London.

Hutt, W.H., 1977. The Theory of Idle Resources. Liberty Press, Indianapolis, IN.

Knight, F.H., 1921. Uncertainty and Profits. Hart, Schaffner and Marx, New York.

Kreps, D., 1990. Game Theory and Economic Modeling. Oxford University Press, Oxford, UK.

Kreps, D.M., Wilson, R., 1982. Sequential equilibria. Econometrica 50, 863–894.

Lindley, D.V., 1961. Dynamic programming and decision theory. Applied Statistics 10, 39–51.

Lippman, S.A., McCall, J.J., 1976. The economics of job search: a survey. Economic Inquiry 14, 155–189.

Lippman, S.A., McCall, J.J., 1993. Search and the development of the economics of information. Estudios de Economia 20, 223–249.

Ljungqvist, L., Sargent, T.J., 1995a. The Swedish unemployment experience. European Economic Review 39, 1043–1070.

Ljungqvist, L., Sargent, T.J., 1995b. Welfare states and unemployment. Economic Theory 6, 143–160.

Macho-Stadler, I., Perez-Castrillo, D., 1997. An Introduction to the Economics of Information. Oxford University Press, Oxford, UK.

MacQueen, H., Miller, R.G., 1960. Optimal persistence policies. Operations Research 8, 362–380.

McCall, J.J., 1970. Economics of information and job search. Quarterly Journal of Economics 84, 113–126.

McMillan, J., Rothschild, M., 1993. Search. In: Aumann, R., Hart, S. (Eds.), Handbook of Game Theory with Economic Applications. North-Holland, Amsterdam.

Mortensen, D.T., Pissarides, C.A., 1999. New developments in models of search in the labor market. In: Ashenfelter, O., Layand, R. (Eds.), Handbook of Labor Economics, vol. 3. North-Holland, Amsterdam, pp. 2567–2627.

Myerson, R.B., 1999. Nash equilibrium and the history of economic theory. Journal of Economic Literature 37, 1067–1082.

Rasmusen, E., 1989. Games and Information. Basil Blackwell, New York.

Rothschild, M., 1973. Models of market organization with imperfect information. Journal of Political Economy 81, 1283–1308.

Rubinstein, A., 1995. John Nash: the master of economic modeling. Scandinavian Journal of Economics 97, 9–13.

Rust, J., 1994. Structural estimation of Markov decision processes. In: Engel, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. 4. North-Holland, Amsterdam, pp. 3082–3139.

Salanie, B., 1997. The Economics of Contracts: A Primer. MIT Press, Cambridge, MA.

Sargent, T.J., 1999. The Conquest of American Inflation. Princeton University Press, Princeton, NJ.

Schumpeter, J.A., 1942. Capitalism, Socialism and Democracy. Harper and Row, New York.

Schwarz, G., 1994. Game theory and statistics. In: Aumann, R.J., Hart, S.M. (Eds.), Handbook of Game Theory, vol. 2. Elsevier Science B.V, Amsterdam.

Simon, H.A., 1951. A formal theory of the employment relationship. Econometrica 19, 293–305.

Stigler, G.J., 1961. The economics of information. Journal of Political Economy 69, 213–225.

Wolpin, K.I., 1995. Empirical Methods for the Study of Labor Force Dynamics. Harwood Academic Publishers, Luxembourg.

Information Theory

Laura Martignon, Ludwigsburg University of Education, Ludwigsburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Information theory is the mathematical treatment of the concepts, parameters, and rules governing the transmission of messages through communication systems. It was founded by Claude Shannon toward the middle of the twentieth century and has evolved into a vigorous branch at the interface between mathematics and communication theory, fostering the enrichment of other scientific fields, such as biology, behavioral science, neuroscience, and statistical mechanics. The techniques used in information theory are probabilistic in nature, and some view information theory as a branch of probability theory. In a given set of possible events, the information of a message describing one of these events quantifies the symbols needed to encode the event in an optimal way. 'Optimal' means that the obtained code word will determine the event unambiguously, isolating it from all others in the set, and will have minimal length, that is, it will consist of a minimal number of symbols. Information theory also provides methodologies to separate real information from noise and to determine the channel capacity required for optimal transmission conditioned on the transmission rate.

Introduction

Radios, telephones, televisions, and, last but not least, computers are devices that store, process, and communicate information; today they belong to the basic equipment of the average citizen in the industrialized world. It is no exaggeration to claim that the twentieth century witnessed the greatest revolution of communication and information technology since the invention of the printed book. This vigorous development created the need for a well-founded theory and a rigorous terminology that would allow discussion about information with scientific precision. In 1948, the American electrical engineer Claude Shannon provided a synthesis of the fundamental ideas of what is now considered the classical theory of communication (1948). In his seminal paper he wrote that the fundamental problem of communication is that of reproducing at one point, either exactly or approximately, a message selected at another point (Shannon, 1948). Information theory is the conceptual framework in which Claude Shannon solved main aspects of this problem.

For the sake of historical precision it should be noted that some of Shannon's ideas had already been put forward, although in less rigorous terms, by Hartley (1928). Yet it was Shannon's achievement to set information theory on firm theoretical grounds. He also discovered the basic mathematical laws of systems that communicate information. Separating relevant information (signal) from irrelevant information (noise) was one of the problems Shannon tackled and solved. Another problem that Shannon essentially solved was the setting of upper bounds on what can be achieved with a given information-carrying medium or channel. For these seminal contributions to information theory, Shannon is considered its founder.

In telephony for instance, there is a source of information, namely a person speaking into a telephone. The message is the sound that is produced, while the transmitter is composed by the microphone and the equipment that changes this sound into an electromagnetic wave, which becomes the signal. The channel is the space between the transmitting and the receiving telephone, that is, the receiver. Any static or other phenomenon distorting the signal is called noise. An unambiguous specification of the particular message chosen by the information source has to be transmitted so the original message can be reconstructed at the receiving point. The receiver reconverts the received signal into an audible sound. Information theory deals with the problem of specifying a quantitative relation between a given message and the set of all possible messages from which it is chosen (see **Figure 1**).

Quantifying Information

Imagine a simple game for two players: one player thinks of a number between 1 and 64 and the other has to guess what the number is by asking only questions that can be answered by 'yes' or 'no.' The strategy followed by a player could be, for instance, to begin by asking the question: Is the number less than or equal to 32? If the answer is, say, yes, then the player

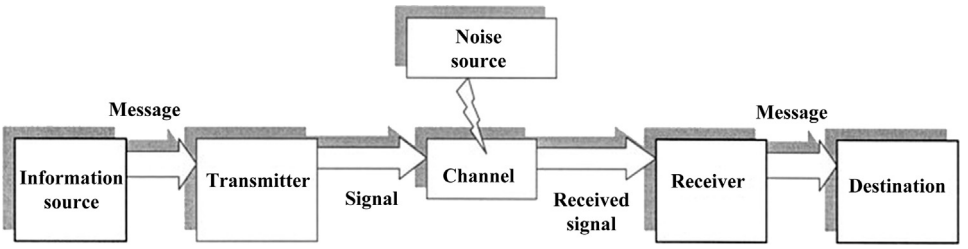

Figure 1 A typical channel for information transmission.

could ask whether it is less than or equal to 16. If the answer is, say, no, then the remaining set of candidate numbers is 17, 18, 19, …, 32. One possible optimal question at this moment – clearly not the only optimal question – would be: Is the number less than or equal to 24? Clearly, the minimum number of yes–no questions required to determine the number for certain is 6. Roughly speaking, the amount of information contained in the message describing an event out of a set of possible events corresponds to the number of yes–no questions needed to isolate that event from all others in the set, following an optimal strategy. If, in the previous game the number to be guessed were, for instance, 23, then the sequence yes-no-no–yes-no-yes, beginning from right to left, is a code for the number 23 that corresponds to the strategy described herein and the specification of the message would then be this ordered sequence of yes and no obtained as answers. The unit for measuring information is the bit. One bit of information corresponds to a message describing one event out of two equally likely events. Two bits correspond to a message describing an event out of four equally likely events. One event out of $N = 2^d$ likely events has d bits of information, or, in a somewhat more suggestive form, it has $-\log_2 1\ N^{-1}$ bits of information (see Figure 2).

If an event with probability N^{-1} has $-\log_2 N^{-1}$ bits of information, it appeared natural to Shannon to extend this measure to an event with probability p by establishing that it must have $-\log_2 p$ bits of information. It was shown by Shannon that the negative logarithm of the probability is the only function satisfying all the properties one expects from information (see also Luce, 1960). Characterizing the properties of this measure for information and the concepts related with it was one of Shannon's breakthroughs. Given a set of N events with probabilities $p_1, p_2, \ldots, p_N$, the amount of information associated with the i-th message is $-\log_2 p_i$ so that the expected value of information per message is given by

$$H = -\{p_1 \log p_1 + p_2 \log p_2 + \cdots + p_N \log p_N\}.$$

This quantity represents the uncertainty contained in the source. Shannon had originally called this expression the average information of the message set. It was John von Neumann who convinced Shannon to introduce the term 'entropy' for average information, seeing the natural connection between average information and entropy in thermodynamics. The term stems from thermodynamics and was introduced by Rudolf Clausius, who formulated the second law of thermodynamics in 1850 (the entropy of an isolated system is always increasing until it reaches its maximum). He composed the term 'entropy' with the Greek roots tropein (turn) and *en* (into) in analogy with the term 'energy.' Entropy takes on its smallest value, zero, when there is one message that is certain to occur (i.e., the described event has a probability 1) and all other messages never occur. Intuitively, there is no information in a message that was a priori certain. Maximal entropy occurs when all messages are equally likely, which corresponds to the situation of greatest uncertainty about outcomes. The ratio of the uncertainty of a source to the maximal entropy possible on the same set of symbols is the relative entropy of the source and

$$\text{“}1 - \text{relative entropy”}$$

is the redundancy of the source. The specification of a message as a (finite) sequence of binary symbols (say, for instance, 0 for no and 1 for yes) is called a 'code word' for the message. An important feature of the average information, H, is that it determines the saving in transmission time that is possible, by proper encoding, because of the statistics of the message source.

Based on the concept of information, Shannon introduced measures for information rate as well as for channel capacity. He stressed the fact that information has all the properties of matter: it can be transported, compressed, and segmented into small pieces. The diagram in Figure 1 can be viewed as a transportation system. The channel can be seen as transporting matter from a source to a receiving site. In this situation, there are two important quantities: the rate at which information is produced at the source and the capacity of the channel. Obviously if the rate of production is greater than the capacity of the channel, it will be impossible to transport the full output of the source. Suppose, however, that there is a device – the encoder – at the source that, just like a mill, cuts

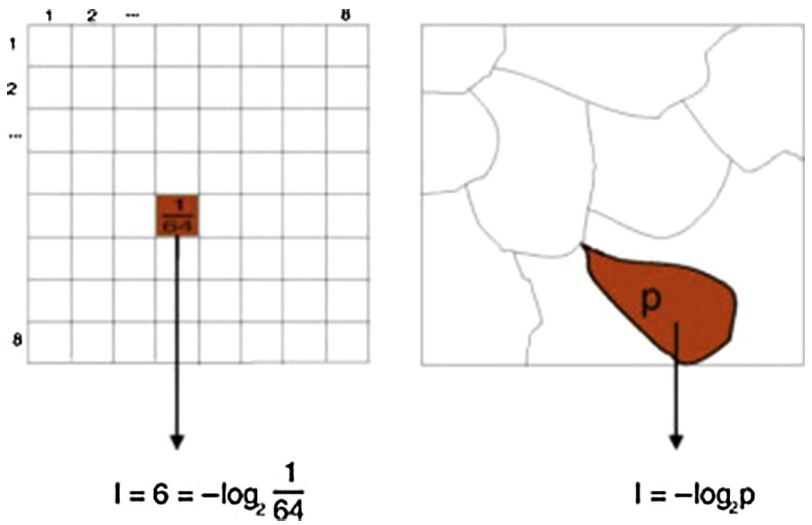

Figure 2 Quantifying information.

information into smaller pieces, then 100% efficiency can be achieved. Shannon proved that if the information rate at the source is less than the channel capacity, then encoders and decoders can be constructed such that the source output is recreated at the destination with as small an error probability as desired. Conversely, if the information rate at the source is greater than the capacity, then the error probability cannot be made arbitrarily small.

Information and related concepts have led to a unified and theoretically sound treatment of uncertainty and reduction of uncertainty in statistics and decision theory. In the set of all distributions on a finite sample space, the distribution that maximizes entropy is the uniform distribution, for which all single elements are equally probable. Laplace's 'uncertainty principle' (1812) advised to assume that a distribution is uniform, whenever no further knowledge is at hand. For far more than one century probabilists and statisticians searched for an adequate extension of the uncertainty principle to the case when partial information on a distribution is available. Information theory provided the ideal extension via the concept of entropy: it was called the 'maximum entropy principle,' or 'MaxEnt.' It states that whenever some of the marginals of the distribution are known, one should assume that the distribution is the one – and only one – that maximizes entropy in the set of all distributions with those marginals. By doing this one guarantees that no additional information is added to the distribution. A family of important algorithms has been developed for the construction of the MaxEnt distribution (Bishop et al., 1975). The MaxEnt principle and algorithms for constructing the MaxEnt distributions form the fundamental ingredients for the designation of a priori distributions when events are not equiprobable.

An excellent, modern survey of information theory and its applications was provided by Cover and Thomas (1991).

Information Theory Applied in Psychology

Information theory has had a major impact on behavioral science and on cognitive neuroscience. Psychologists have discovered important relationships between the amount of information in a stimulus and reaction time to the stimulus. An important finding concerns the adaptation of participants to the statistics of information conveyed by presented stimuli and their ability to recognize redundancy of information in stimuli. This leads to shorter reaction times. Another important finding concerns human short-term memory: Pollack (1953) and Miller (1956) noted that basic information is coded in 'chunks.' The ability to form chunks enables organisms to make better use of their limited capacity short-term memory. Because short-term memory appears to be capable of holding only five to seven chunks, efficient information processing depends on the ability to pack more information into a single chunk. Grouping into doublets (two-bit-chunks) and recoding (e.g., 00 = A, 01 = B, 10 = C, 11 = D) can double the number of bits that can be held in short-term memory at a given time.

Forming triplets or larger groupings further increases the number of memorized bits in a nonlinear way. But, there are constraints to this procedure. Chunking is only effective with sufficiently long presentation times. With presentation times under 0.5 s, no successful grouping takes place (Kleinberg and Kaufmann, 1971). Yet with presentation times of 1 s and longer there is a clear advantage in using chunks. An excellent, by now classical, survey of applications of information theory in psychology was provided by Attneave (1959).

Information theory has allowed a treatment of redundancy in language. For example, the sentence 'Sm mn in ths rm us glsss' can be understood as 'Some men in this room use glasses,' although several vowels have been deleted. Shannon studied the redundancy of English language and discovered that English texts can be reconstructed when 50% of the letters, chosen at random, are deleted. Other important applications of information theory deal with word perception. If the source vocabulary is small, humans are able to perceive words nearly perfectly, whereas perception becomes increasingly difficult as the source vocabulary becomes larger (Miller and Selfridge, 1950).

Information Theory Applied in Neuroscience

In cognitive neuroscience, information theory has become an essential tool for understanding brain function. Modeling the brain as an information processing machine has led to important advances in both theory and practice. The production of action potentials (spiking) of neurons can be discretized and interpreted as a binary process. In this way, a temporal sequence of spikes can be represented as a binary vector of 0 and 1. All information coming from sensory inputs is processed by neurons using this simple binary language. Hebb (1949) conjectured that neural information processing cannot be understood just in terms of the behavior of individual neurons. Rather, the brain must be viewed as a system of so-called cell assemblies, wherein information processing occurs as an emergent property resulting from the interaction of these cell assemblies. Hebb's conjecture received a great deal of experimental confirmation in the last decades of the twentieth century (Gray et al., 1989). The famous binding problem, or problem of identifying an object by recognizing individually and 'binding together' features that characterize the object, appears to be solved by means of synchronization of spiking activity (Singer et al., 1994). Theoretical neuroscientists have developed many different approaches to modeling how information is stored in the brain. The most convincing and persistent model has been the associative memory (e.g., Palm, 1980) that stores similar patterns close to each other, where the closeness between patterns is measured in terms of information overlap.

See also: Probability Theory: Formal; Probability Theory: Interpretations.

Bibliography

Attneave, F., 1959. Applications of Information Theory to Psychology: A Summary of Basic Concepts, Methods and Results. Holt, New York.

Bishop, Y., Fienberg, S., Holland, P., 1975. Discrete Multivariate Analysis. MIT Press, Cambridge, MA.

Cover, T., Thomas, J., 1991. Elements of Information Theory. Wiley Series in Communications, New York.

Gray, C.M., Koenig, P., Engel, A.K., Singer, W., 1989. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. Nature 388, 334–337.

Hartley, R.V.L., 1928. Transmission of information. Bell System Technical Journal 7, 535–563.

Hebb, D., 1949. The Organization of Behavior. Wiley, New York.

Kleinberg, J., Kaufmann, H., 1971. Constancy in short-term memory: bits and chunks. Journal of Experimental Psychology 89, 326–333.

de Laplace, P., 1812. Théorie analytique des Probabilités. Courcier, Paris.

Luce, R.D., 1960. The theory of selective information and some of its behavioral applications. In: Luce, R.D. (Ed.), Developments in Mathematical Psychology. Free Press, IL, Glencoe, pp. 1–119.

Miller, G.A., Selfridge, J.A., 1950. Verbal context and the recall of meaningful material. American Journal of Psychology 63, 176–185.

Miller, G.A., 1956. The magical number seven, plus minus two: some limits on our capacity for processing information. Psychological Review 63, 81–97.

Palm, G., 1980. On associative memory. Biological Cybernetics 36, 19–31.

Pollack, R.B., 1953. The assimilation of sequentially encoded information. American Journal of Experimental Psychology 66, 421–435.

Shannon, C., 1948/1962. The mathematical theory of communication. In: Shannon, C., Weaver, W. (Eds.), The Mathematical Theory of Communication. University of Illinois Press, Urbana IL, pp. 3–91.

Singer, W., 1994. Time as coding space in neocortical processing: a hypothesis. In: Buzsaki, G., et al. (Eds.), Temporal Coding in the Brain. Springer Verlag, Berlin.

Informed Consent of Research Subjects

Paul S Appelbaum, Columbia University, New York, NY, USA; and New York State Psychiatric Institute, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Informed consent' denotes a person's autonomous agreement to a proposed course of action, usually in a treatment or research context. The principle that research requires the informed consent of participants is accepted around the world. Elements necessary for consent to be informed include disclosure of relevant information, adequate decision-making capacity, and absence of coercion or undue influence. Challenging issues characteristic of social and behavioral research include use of deception; necessity of consent for record-based research; informed consent in observational research and fieldwork, and for Internet-based research; and role of community consent.

'Informed consent' denotes a person's autonomous agreement to a proposed course of action, usually in a treatment or research context. Consent is not considered to be informed unless the person has received disclosure of the information that a reasonable person would consider material to the decision at hand; has the capacity to use that information in a rational process of decision making; and reaches a decision in the absence of coercive offers or threats (Berg et al., 2001). Broad agreement exists that informed consent is an indispensable requirement for the ethical conduct of research with human subjects, at least when the research renders them liable to more than minimal risks of harm. However, considerable controversy remains regarding the precise contours of the doctrine and of the circumstances that might constitute exceptions to the general rule. Since the rules have been devised largely with biomedical research in mind, issues more characteristic of social and behavioral science research have been relatively neglected.

History of Informed Consent to Research

The history of informed consent to research is usually traced back to the decision of the court in the Nazi doctors' trial at Nuremberg after World War II, which declared, 'The voluntary consent of the human subject is absolutely essential.' It is clear, however, that the Nuremberg tribunal neither originated nor established the concept of consent to research. References to requirements for consent from research subjects can be found in the writings of physician experimenters and in governmental pronouncements dating back to the late nineteenth century, when experimental medicine grew out of its infancy (Lederer, 1995). Indeed, George Bernard Shaw's *Pygmalion* contains an explicit statement of the principle. Two of the leading instances, ironically, derive from Germany, where the advanced state of empirical medical science made consent an object of at least periodic concern (Grodin, 1992). Nonetheless, however much the Nuremberg court desired to root the principle in contemporary practice, it appears that researchers in many countries, even after the Nazi atrocities came to light, frequently failed to obtain consent from their subjects.

Nor did the Nuremberg decision immediately change these practices. Investigators using human subjects apparently saw the Nuremberg Code, as the 10 principles embodied in the court's decision came to be known, as applying only to the anomalous situation of concentration camp inmates. Thus, the history of the postwar decades is marked by repeated revelations of questionable ethical practices, including studies in which persons were subject to clear harms without their knowledge or consent (for one of the most influential reports of problematic studies, see Beecher, 1966). Pressure built for clearer rules and mechanisms of enforcement.

The World Medical Association responded to these concerns with its *Declaration of Helsinki* in 1964 (World Medical Association, 2008), the first attempt to forge an international professional consensus on the ethics of research. The *Declaration* endorsed a requirement for informed consent, albeit in less sweeping terms than the Nuremberg Code; subsequent iterations of the document have brought increased rigor to the mandate for consent of human subjects. National professional organizations developed similar documents, but their nonbinding character left research subjects with few certain protections.

A move toward legally enforceable rules came in the USA in 1974, in the wake of disclosure of the Tuskegee study, in which African-American men in Alabama, without their knowledge, had been denied treatment for syphilis for more than four decades, ostensibly for the sake of observing the natural course of the disease (Jones, 1981). The federal legislative response to the scandal included the appointment of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, which was charged with generating recommendations for closer oversight of research with human subjects. One of its major products, The Belmont Report (1979), laid out the ethical underpinnings of human subjects research, locating a requirement for informed consent in the principle of 'respect for persons,' which required that "subjects, to the degree that they are capable, be given the opportunity to choose what shall or shall not happen to them." The Commission's reports stimulated the adoption of regulations governing federally funded research. Most of the countries in the world have since followed suit, by way of legislative or regulatory enactments, or binding guidelines (for a summary of the status of research regulation in 103 countries internationally, see Office for Human Research Protections, 2014). The general principle that research should not be conducted

without the informed consent of participants has been accepted around the world.

Current Contours of Informed Consent to Research

The requirements for informed consent to research are rooted in the ethical principle of autonomy, which demands respect for the free, informed choice of a competent person (Faden and Beauchamp, 1986). Valid informed consent requires three elements to be present: disclosure, decision-making capacity, and voluntariness.

Disclosure

The Nuremberg Code (1948) declared that the research subject "should have sufficient knowledge and comprehension of the elements of the subject matter involved as to enable him to make an understanding and enlightened decision." Brody (1998) summarized the international consensus regarding those items that need to be disclosed to achieve this end as follows:

1. "The purpose and the research nature of the project;
2. The duration of the subject's participation, the treatment and procedures that the subject will undergo (with a clear specification of which are experimental), and the subject's responsibilities;
3. The risks of harms and discomforts that might result from participating in the research;
4. The possible benefits to the subject or to society from the research;
5. The alternatives to participating in the research;
6. The right not to participate and to stop participating at any time without penalty or loss of benefits to which the subject is otherwise entitled;
7. The measures to protect confidentiality;
8. The anticipated payments, if any, to the subjects and the anticipated expenses, if any, that will be borne by the subject;
9. The compensation and treatment available for injuries sustained in the research;
10. The identity of a contact person to discuss questions about the research;
11. The circumstances under which the subject's participation will be terminated;
12. The assurance that new information relevant to participation that becomes available during the research will be provided to the subject;
13. The names of the sponsors of the research and the ownership of any commercial results."

This information is generally provided in written form, though often supplemented by verbal disclosures. Frequently, the result is a lengthy form (20–30 page, single-spaced forms are not uncommon in clinical research), filled with a combination of medical/scientific and legal jargon, that is difficult for potential subjects to read and assimilate. Provision of a mass of detailed information means, moreover, that important data, such as those dealing with significant risks, are frequently neglected by subjects. A common resulting problem is the so-called 'therapeutic misconception' – a belief on the part of subjects that procedures are being undertaken for their benefit (including randomization and use of placebos), when they are actually designed solely to advance the research effort (Appelbaum and Lidz, 2008).

A substantial empirical literature documents the failure of this process to result in well-informed research subjects. Consent forms are routinely written at a reading level well beyond the comprehension of the average subject and considerably exceed the amount of information to which most people will attend (Albala et al., 2010). Though investigators are often urged to provide additional disclosures in more understandable formats (e.g., illustrated brochures that subjects can read at their leisure), this still represents the exception rather than the rule. However, there has been a good deal of experimentation with the use of augmented means of disclosure, including iterative educational interactions, hyperlinked computer-based methods, and interactive video disclosures. The data on the efficacy of these approaches are variable, but in at least some studies encouraging results have been achieved (Ryan et al., 2008; Schenker et al., 2011).

Decision-Making Capacity

The subject's capacity (often interchangeably referred to as competence) to give a meaningful consent was considered a *sine qua non* in the Nuremberg Code. Although formulations of the requirements differ somewhat across jurisdictions, decision-making capacity generally involves the abilities to understand the information relevant to a decision, to appreciate its implications for one's own situation, to reason in a rational fashion, and to evidence a choice (Appelbaum and Grisso, 2001).

Capacity is generally presumed to be present, unless there is reason to believe otherwise. Indeed, studies of general medical patients show extremely high rates of capacity and indicate that, as a group, they tend not to differ from the general population in their decision-making abilities. Although studies of populations at elevated risk of incapacity, including persons with severe mental disorders, mild dementia, head injury, and the like, show higher rates of decisional impairment, many of these patients remain capable of making their own decisions about participation in research (Appelbaum, 2006). Particularly when the research may present significant risks to these groups of participants, investigators are increasingly assessing capacity prospectively. This can be done by clinical evaluation or formal screening instruments, with the latter offering a means of structured assessment, standardized scores, and the ability to compare performance across individuals and groups (Dunn et al., 2006). Identification of a potential subject as having impaired capacities does not necessarily mean that the person must be excluded from research participation. Many such persons can be brought to acceptable levels of functioning with additional educational efforts or interventions that address the underlying cause of their impairment, and ultimately allowed to make their own choices about participation.

Were capacity to be an absolute requirement, as the Nuremberg Code suggested, no research could be conducted with children – who lack capacity as a matter of law – or with

many patient groups, including potential subjects who are comatose, suffering from moderate to severe dementia or intellectual disability, or severely psychotic and disorganized. The desirability of allowing research with these groups, so that progress might be made on the disorders from which they suffer, led the drafters of the Declaration of Helsinki, and most subsequent compilers of codes, regulations, and statutes, to make some provision for the entry of incapable subjects into research projects. Typically, a surrogate decision maker appointed by the potential research subject prior to incapacity or otherwise designated to make medical treatment decisions is permitted to provide informed consent on behalf of an incapable subject. Rules governing the inclusion of incapable persons in research differ, but common limits include that the research be directly applicable to their conditions, not able to be performed with capable subjects, hold out the prospect of some therapeutic gain, or present no more than minimal risks of harm. Mechanisms that would allow persons to make advance decisions while competent about whether to enter research projects should they later become incompetent are beginning to be introduced (under the generic rubric of 'research advance directives'), paralleling similar mechanisms that have been developed for treatment settings.

Voluntariness

Overt coercion – the application of threats or unfair offers to a prospective research subject – is universally recognized as vitiating that person's decision and thus rendering consent invalid. The effects of more subtle influences on the legitimacy of subjects' consent to research are a matter of greater controversy. Recent studies suggest that most research participants do not perceive themselves as having been exposed to illegitimate pressure regarding enrollment in a study (Appelbaum et al., 2009). However, some environments may be considered so intrinsically coercive – prisons, for example – that research with the persons detained therein has been severely restricted in many countries. Whether hospitals, institutions for the mentally ill or intellectually disabled, schools and colleges, places of employment, and other entities fall into that category is more difficult to say. Paradoxically, efforts to protect inmates, patients, workers, and others by prohibiting their participation in research deprive them of their residual decision-making power, and may increase their feelings of helplessness and their perceptions of being coerced. Indeed, some recent data suggest that prison inmates do not perceive themselves as vulnerable to coercion with regard to making decisions about research participation (Edens et al., 2011).

Research subjects are often compensated for their participation with money or its equivalent, a practice that has engendered a great deal of confusion among investigators and regulators regarding its appropriate limits. Fair compensation for the time expended in a study is generally considered acceptable. However, to the extent that payment is so substantial as to lead subjects to give insufficient attention to the risks they may run by participating, it may be considered an undue influence on their decisions (a condition different from coercion, with which it is often confused), which renders their choices substantially involuntary. Some commentators have argued that compensation for research participation is similar to payment for any other job, and hence is intrinsically incapable of subverting free choice (Emanuel, 2005). Although this remains a minority view, many research ethicists concur that restrictions on payment of subjects have been unnecessarily rigid, with the unintended result that participants' contributions receive less reward than they deserve.

Special Considerations in Behavioral and Social Science Research

Rules governing informed consent to research, from the Nuremberg Code to most contemporary governmental regulations, typically have been formulated with biomedical research in mind. Issues more characteristic of social and behavioral research have been relatively neglected. Indeed, the federal regulations in the USA categorically exempt many types of behavioral and social research from their purview, including educational research; studies designed to evaluate governmentally provided benefits or services; and testing, surveys, interviews, or observations of public behavior, unless information is recorded in a manner that would render subjects identifiable, and subjects would be likely to suffer harm if the information were released (Code of Federal Regulations, 1991).

Even so, many behavioral and social scientists object to what they perceive as a steady expansion of the scope of regulatory attention – sometimes pejoratively characterized as 'mission creep' – embodied in the USA in requirements for prospective review of their studies by institutional review boards (often referred to as 'research ethics committees' in other countries). Controversies have erupted over the application of regulatory requirements for research, including requirements for informed consent, to oral history, journalism, ethnographic research, and other approaches that seem very different from the biomedical studies that initially motivated development of the regulations (Schrag, 2010). It is sometimes argued that research oversight restricts behavioral and social science investigators' freedom of inquiry and freedom of speech in ways that are both unfair and detrimental to the research enterprise.

Yet, many behavioral and social science studies involving human subjects raise important issues related to informed consent. One of the major questions is when, if ever, it is appropriate to deceive subjects about the purpose of a research study, a frequent technique in psychological research. Other issues concern consent for access to records, consent in observational and field work studies, and the appropriateness of seeking consent – or some equivalent – from entire communities that may be affected by an intervention that is part of a research project.

Consent in Studies Based on Review of Records

Analysis of existing records, often aggregated into large databases, is a frequent technique used by behavioral and social scientists. These records may have been collected originally with subjects' consent but for nonresearch purposes (e.g., medical records) or without subjects' consent and sometimes without their knowledge (e.g., national birth registries,

educational testing results, census information, or tracking data on Internet utilization). When individual subjects are not reasonably identifiable in the database, it is generally agreed that consent need not be obtained prior to granting researchers access. Even with identifiable data, researchers have often been allowed to utilize such records as long as they agree to protect the confidentiality of those persons whose information they are reviewing (e.g., by publishing data only in aggregate form).

This practice has been criticized as insufficiently protective of subjects' interests in both autonomy and privacy. The autonomy argument is predicated on the belief that persons should have the right to control access to information about them, preventing its use, for example, for research purposes of which they may disapprove. Privacy concerns have been given greater weight of late with the growing ability of researchers to integrate information obtained from different databases about the same individuals, thus accumulating large amounts of information that, were it to be used or accessed for unauthorized purposes, could prove embarrassing or harmful to the people involved. These concerns have grown as new types of information, often collected without people's conscious awareness, have come into being, e.g., cellphone tracking data that permit continuous monitoring of persons' locations and movements. Strong arguments have been made for obtaining consent from the subjects of these records prior to their release for research purposes.

Researchers often object to these proposals on the grounds that they would materially increase the costs of research and, insofar as subjects' refusal of consent would bias the resulting data, decrease the validity of their findings (Institute of Medicine, 2009). In many cases, it will simply be infeasible to track down and obtain consents from the persons whose data are in question. No reasonable argument can be made for precluding researchers' access in these circumstances, so long as the study is of some significant value and appropriate safeguards are in place. When smaller and more recent repositories of data are involved, and obtaining consent would not involve unreasonable expenditures of resources, it is clearly more respectful of subjects' interests to seek their agreement. When records are being compiled prospectively with the intent of using them in the future for research purposes, there is no ethical justification for failing to obtain subjects' consent at the time.

Consent in Observational Studies and Field Work

Many of the behavioral sciences have a long tradition of sending researchers into the field to study the behaviors of persons in their natural environments. This often involves requesting access from those to be observed to otherwise private interactions and territory (e.g., discussions between spouses in the subjects' home, physician–patient interactions in medical clinics). Considerable controversy exists over whether and how truly informed consent should be obtained in such cases. Insofar as researchers may observe and report embarrassing or frankly criminal behavior, subjects may be placed at risk of psychological or tangible harm. Thus, the case for obtaining consent in such studies would seem strong, bolstered by the discomfort some investigators report in withholding this information from their subjects.

On the other hand, researchers have an understandable concern that informing subjects clearly of the goals of their studies may lead subjects to alter their behaviors, thus obscuring the very phenomena that are under study. Though many subjects appear remarkably prone to ignore such information, especially once they become accustomed to the observer's presence, and appear to behave as they would were no observers present, this concern needs to be taken seriously. It should usually be possible to accommodate both parties' interests by disclosing the intent of the study in general terms, along with the likely risks to which subjects may be exposed. If some modification of subjects' behavior occurs, this may be an unavoidable price to pay for acquiring a meaningful consent. In addition, after data collection is completed, the researchers can provide more complete information to the subjects regarding the purpose of the research and the questions it was designed to address. Although such *post facto* debriefing does not vindicate subjects' autonomy and privacy interests as would full prospective consent, it does demonstrate respect for the personhood of research subjects and hence should be encouraged.

Consent in Community Intervention Projects

Social science research increasingly involves large-scale interventions directed at particular communities or segments of a community. These may include interventions designed to avoid undesirable outcomes (e.g., juvenile delinquency, teenage pregnancy, illegal drug use) or to promote what are generally seen as salutary ends (e.g., moving persons off welfare and into paying jobs). When such interventions occur at the level of the community as a whole – such as broad-based educational programs concerning drug abuse – it is impossible to obtain informed consent as it is usually conceptualized. Moreover, when studies are based on changes to entitlement programs that have been enacted through the political process, the consent of individual persons may be seen as irrelevant.

There may, nonetheless, be a role in such instances for a modified form of consent interactions with the relevant community. Although each person's consent cannot be obtained, disclosure of the nature and purpose of the study and its risks and benefits can still take place. Consultation with the community at large and its designated representatives may result in suggestions for modifications in the study culminating in a general consensus concerning its desirability. The growing popularity of the community-based participatory research model (Minkler and Wallerstein, 2008), which represents the next step in the evolution of this approach, is premised on the value of going beyond mere consultation with the community, and involving community representatives in all aspects of research from study design to data analysis as equal participants with the researchers.

Consent in Internet-Based Research

The rapid growth of the Internet since the beginning of the new millennium has made it an attractive venue for research that previously could only have been carried out by face-to-face interaction or observation. Investigators have mined publicly available data to assess, for example, networks of social

relationships and consumer preferences, and have engaged in observation and interaction on social media portals ranging from interactive gaming sites to chat rooms. Although some of the issues related to informed consent that arise in these studies are similar to those in traditional participant-observation research (e.g., will obtaining formal consent so alter subjects' behavior as to vitiate the possibility of obtaining valid data?), the ability of Internet-based researchers to disguise their identity or even presence (i.e., 'lurking' in a chat room) heightens concerns about when consent is required and how it should be obtained. In addition, the possibility of identifying participants and linking their data to other data sets enhances the risks associated with such studies. Moreover, while online postings and interactions often take place in nonprivate domains, studies suggest that participants are frequently unaware of the possibility that their statements and behaviors may be open to public view, and scholars have suggested that consent should be required to mine these data (Zimmer, 2010).

Standards for when informed consent must be obtained online and how that can be accomplished are still in evolution, in part because Internet-based technologies themselves are in flux. The media themselves create technical challenges, for example, when participants in a chat room come and go at will during a period of observation, even if the investigator desires to obtain consent from each participant, how can that be accomplished? Other complexities include the real possibilities that some participants are minors, and hence unable to give valid consent on their own behalf, while others may be located in countries different than the one where the investigator is based, which may have different rules governing research and varying requirements for consent. As this domain of research grows into maturity, it seems likely that a clearer perspective will be obtained on participants' views about the desirability of consent, the probable risks they face from various forms of Internet-based research, the feasibility of obtaining their consent, and regulatory considerations.

The Future of Informed Consent to Research

Imperfect as the process of obtaining consent to research is in practice, it is nonetheless widely considered a bulwark of ethical research with human subjects. However, dissatisfaction with the ways in which consent is obtained has led some scholars to suggest that the goals of the informed consent process should be reconceptualized. Rather than seeing the purpose of consent as vindicating subjects' autonomy, for example, Manson and O'Neil (2007) have suggested that it should be understood as a mechanism for waiving rights to which subjects would otherwise be entitled (e.g., privacy of medical information). A practical consequence of this conceptual difference, they argue, would be a more manageable consent process, one focused on effective communication of a limited range of critical information, as opposed to the current tendency toward massive – and often incomprehensible – disclosure. Similarly disturbed by the failures of many research subjects to absorb consent disclosures, Miller and Wertheimer (2011) argue for a 'fair transaction model' of informed consent, whereby investigators would disclose the information necessary for fair consideration of participation in a study, but without an expectation that subjects would necessary grasp the disclosure in its entirety. They, too, believe that it is unrealistic to expect informed consent to result in fully autonomous decision making, but more explicitly rely on the vetting of research studies by research ethics committees to ensure that potential subjects are not being asked to participate in unreasonably harmful research.

At the same time that these arguments for reconceptualizing consent have been put forward, other scholars have pursued empirical studies that promise to shape for the better a debate that has frequently been waged on theoretical grounds alone. Data on the failings of the current system can point to practical ways of improving the consent process that do not require a reorientation of its goals or changes in regulation. Among the areas of investigation have been participants' views regarding the information material to an informed decision; the adequacy of investigators' disclosures and means by which they can be improved; the prevalence of incapacity among various groups of research subjects; subjects' understanding of the information they have received and their appreciation of its implications for their situation; subjects' retention of information over time; and subjects' perceptions of pressures to consent to participation.

Conducting research on informed consent is often challenging. Naturalistic research on the consent process involves overcoming the reluctance of investigators to grant access to the research setting, and of subjects to consent to participation in yet another research study. Subject recruitment can be both difficult and time consuming. The willingness of funding agencies to sponsor such research is often related to the degree of public scrutiny of and concern about the agencies' research portfolio. But such research is likely to continue to increase over time and to play a greater role in helping to shape investigators' practices, official policies, and theoretical understandings of informed consent to research.

See also: Common Rule: US Regulatory Protections for Human Research Participants; Fieldwork: Ethical Aspects; Internet Research: Ethical Concerns; Privacy of Individuals in Social Research: Confidentiality; Research Ethics: Research.

Bibliography

Albala, I., Doyle, M., Appelbaum, P.S., 2010. The evolution of informed consent forms for research: an evaluation of changes over a quarter century. IRB Ethics & Human Research 32 (3), 7–11.

Appelbaum, P.S., 2006. Decisional capacity of patients with schizophrenia to consent to research: taking stock. Schizophrenia Bulletin 32, 22–25.

Appelbaum, P.S., Grisso, T., 2001. The MacArthur Competence Assessment Tool for Clinical Research (MacCAT-CR). Professional Resource Press, Sarasota, FL.

Appelbaum, P.S., Lidz, C.W., 2008. The therapeutic misconception. In: Emanuel, E.J., Grady, C., Crouch, R.A., Lie, R.K., Miller, F.G., Wendler, D. (Eds.), The Oxford Textbook of Clinical Research Ethics. Oxford University Press, New York, pp. 633–644.

Appelbaum, P.S., Lidz, C.W., Klitzman, R., 2009. Voluntariness of consent to research: a preliminary empirical investigation. IRB Ethics & Human Research 30 (6), 10–14.

Beecher, H.K., 1966. Ethics and clinical research. New England Journal of Medicine 274, 1354–1360.

Berg, J.W., Appelbaum, P.S., Lidz, C.W., Parker, L., 2001. Informed Consent: Legal Theory and Clinical Practice, second ed. Oxford University Press, New York, NY.

Brody, B.A., 1998. The Ethics of Biomedical Research: An International Perspective. Oxford University Press, New York.

Code of Federal Regulations (US), Revised June 18, 1991. Protection of Human Subjects, Title 45, Part 46. US Government Printing Office, Washington, DC.

Dunn, L.B., Nowrangi, M.A., Palmer, B.W., Jeste, D.V., Saks, E.R., 2006. Assessing decisional capacity for clinical research or treatment: a review of instruments. American Journal of Psychiatry 163, 1323–1334.

Edens, J.F., Epstein, M., Stiles, P.G., Poythress, N.G., 2011. Voluntary consent in correctional settings: do offenders feel coerced to participate in research? Behavioral Sciences & the Law 29, 771–795.

Emanuel, E., 2005. Undue inducement: nonsense on stilts? American Journal of Bioethics 5 (5), 9–13.

Faden, R.R., Beauchamp, T.L., 1986. A History and Theory of Informed Consent. Oxford University Press, New York.

Grodin, M.A., 1992. Historical origins of the Nuremberg Code. In: Annas, G.J., Grodin, M.A. (Eds.), The Nazi Doctors and the Nuremberg Code: Human Rights in Human Experimentation. Oxford University Press, New York, pp. 121–144.

Institute of Medicine, 2009. Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health through Research. National Academies Press, Washington, DC.

Jones, J., 1981. Bad Blood. Free Press, New York.

Lederer, S.E., 1995. Subjected to Science: Human Experimentation in America before the Second World War. Johns Hopkins University Press, Baltimore, MD.

Levine, R.J., 1986. Ethics and Regulation of Clinical Research, second ed. Urban and Schwarzenberg, Baltimore, MD.

Manson, N.C., O'Neil, O., 2007. Rethinking Informed Consent in Bioethics. Cambridge University Press, New York.

Miller, F.G., Wertheimer, A., 2011. The fair transaction model of informed consent: an alternative to autonomous authorization. Kennedy Institute of Ethics Journal 21, 201–218.

Minkler, M., Wallerstein, N., 2008. Community-Based Participatory Research for Health: From Process to Outcomes. Jossey-Bass, Hoboken, NJ.

National Bioethics Advisory Commission, 1998. Research Involving Persons with Mental Disorders That May Affect Decision-Making Capacity. National Bioethics Advisory Commission, Rockville, MD.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978. The Belmont Report: Ethical Guidelines for the Protection of Human Subjects of Research. US Government Printing Office, Washington, DC.

Office of Human Research Protections, U.S. Department of Health and Human Services, 2014. International Compilation of Human Research Standards. http://www.hhs.gov/ohrp/international/index.html (accessed 08.09.14).

Ryan, R.E., et al., 2008. Audio-visual presentation of information for informed consent for participation in clinical trials. Cochrane Database of Systematic Reviews, CD003717.

Schenker, Y., et al., 2011. Interventions to improve patient comprehension in informed consent for medical and surgical procedures: a systematic review. Medical Decision Making 31, 151–173.

Schrag, Z.M., 2010. Ethical Imperialism: Institutional Review Boards and the Social Sciences, 1965–2009. Johns Hopkins University Press, Baltimore, MD.

The Nuremberg Code, 1948. In: Annas, G.J., Grodin, M.A. (Eds.), The Nazi Doctors and the Nuremberg Code: Human Rights in Human Experimentation. Oxford University Press, New York.

World Medical Association, 2008. Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. http://www.wma.net/en/30publications/10policies/b3/ (accessed 27.05.13.).

Zimmer, M., 2010. "But the data is already public": on the ethics of research in Facebook. Ethics and Information Technology 12, 313–325.

Infrastructure Investment

David C Perry, University of Illinois, Chicago, IL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7486–7489, © 2001, Elsevier Ltd.

Abstract

Infrastructure investment is central to urban development, generally conceived of as the physical foundation of the city or, as one historian graphically put it, the 'technological sinews' built to meet the needs of society. More precisely these are the transportation networks, power girds, utility systems, and telecommunicative linkages that deliver the heat, light, water, and power that are central to urban life and without which modern cities would not exist. This foundational primacy affords infrastructure a rather unique status in the literature: namely it is not described simply as an important physical artifact but as an investment in the city-building process. Understanding the political and economic implementation of infrastructure investment is important to understanding the modern city in both the 'developed' and 'developing' urban world. Infrastructure investment links technology to: (1) the organization and constitution of modern governments; (2) public finance and debt policies; (3) the creation and regulation of capital markets; (4) the articulation of public goods; (5) the privatization of governmental services; and (6) the increasing role of nongovernmental intermediaries in markets and states, thereby creating socio-technical networks of service underservice that contribute to the distributions of power and unevenness of development in cities, worldwide.

The city-building process is a product of infrastructure investment. At the most general level, the term infrastructure describes the underlying foundation of the physical and social facilities of the city (Seeley, 1993: p. 20; Konvitz, 1985: p. 130), both the physical systems and technologies, and the health, education, and welfare facilities that knit the city into a collective fabric. From this perspective, Herbert Muschamp (Perry, 1995: p. 1) observes that, "infrastructure is shorthand for the structural underpinnings of the public realm."

While such an encompassing approach to infrastructure has conceptual resonance, most studies carve a more particular focus (OECD, 1991; Neutze, 1997; Graham, 2000), defining infrastructure as the underlying physical foundation of the city: the material systems, networks, and technologies (roads, water and sewer systems, power grids, and telematic networks) that serve key urban relations; the social flows, economic exchanges, and political relations of the city. This foundational primacy of physical infrastructure affords it a rather unique status in the literature: namely, the bridges, roads, and water systems of the city are not described simply as important physical artifacts but as investments in the city-building process. The returns on such investments are measured in terms of a society's economy, health, and social well-being.

Hence infrastructure scholarship and policy often begin with the normative position that the more an urban population is able or willing to invest in physical infrastructure, the more apt the urban settlement is to prosper economically and benefit socially from the resulting conditions of efficiency, health, and human development. This is perceived to be as true in new or emerging urban centers as it is in the established cities of the developed world, thereby making issues of definition, public or private provision of such systems, finance, and technology equally important to every modern city.

Infrastructure: Evolution from Private to Public Function

The term 'infrastructure' as it is employed here is relatively new: a conception of the late twentieth century that is used to describe the systemic applications of technology to the needs of urban society, producing what historian Joel Tarr (1984: p. 4) calls the 'technological sinews' of the city: its transportation networks, power grids, utility systems, and telecommunicative linkages. For much of the nineteenth and twentieth centuries, urban physical facilities were not viewed systemically as 'physical networked services' (OECD, 1991: p. 19), but as either privately produced, geographically individuated 'internal improvements' or equally decentralized, governmentally generated 'public works.'

For example, the seventeenth-century invention of the British toll road, the late eighteenth-century canal systems, and the early railways were initially proprietary responses to the requirements of the Industrial Revolution. Technological advances in water treatment, electricity, gas distribution, and modes of communication were virtually all developed by private inventors and investors. Only later, as the impact of such inventions on the dramatic and uneven shifts in urban populations, levels of disease, congestion, and fire became apparent, did they come to be viewed as a public responsibilities or goods. Even then, cities "different … business leaders, politicians and professionals" (Tarr, 1984: p. 5) would take years to determine which technology or proprietary service was the right one and more years to decide how to plan, pay for, and ultimately build the service (Konvitz, 1985; Anderson, 1988). For example, the benefits of water and sewage treatment to health and economic development were well known by the middle nineteenth century, but it took US cities until 1940 to build such systems. World wide, hundreds of millions of urbanites live in political-economic environments that still

do not offer reliable water and healthy sewage systems, transportation systems, or power.

As such, the history of successful infrastructure investment is less about the direct impact of technology and more about the interaction between technology and society (Tarr, 1984). Put another way, recent scholars posit that infrastructure networks are both social and technical: they are always contested and their impacts on cities are never neutral (Graham, 2000). No matter how 'public' the good is perceived to be, the impacts of telecommunicative grids, transportation roads, and energy will always privilege some places and people over others (Swyngedouw, 1993).

Public Infrastructure Investment

However convoluted its investment process and uneven its impacts, infrastructure investment became essentially public practice in the twentieth century, carried out either by governments or heavily regulated private firms. Several reasons are offered for this (Neutze, 1997; Perry, 1995). First, urban infrastructure often serves the essential health and well-being of the citizenry, requiring full citizen access to such services at reasonable cost. Second, this implies the provision of goods at a scale and price (given the elasticity of demand of several of these services) that is beyond the capacity or interest of private providers. Third, urban infrastructure goods often constitute natural monopolies, whereby it is less costly to have a single provider than several service delivery systems. For example, it would be very costly to have competing water and sewer companies putting parallel networks of mains and pipes under every street in a city to give consumers a choice of systems. Fourth, some forms of urban infrastructure produce almost exclusively public goods or externalities (or services which can not be used by one consumer without also being made available to all), such as public roads, sidewalks, and parks. Fifth, equity can require that certain key benefits of urban infrastructure should not be withheld from any citizen. (Neutze, 1997: pp. 29–31)

Special Purpose Governments

Just as the conditions of scale, cost, and equity have led to public provision of physical infrastructure, the sheer size and the accompanying obligations of debt and delivery of infrastructure systems has often led one level of government (especially national governments) to offload many of these services on to lower levels of government. Municipal governments in turn have further devolved the delivery of infrastructure, one by one, to special purpose governments or public authorities (Sbragia, 1996; Perry, 1995). In the United States, for example, such governments issue the majority of all bonds used to finance infrastructure investment. It is argued that such decentralized structures work well to:

1. preserve and circumvent local and state debt limits, maintaining the legal requirements of general municipal fiscal balance while allowing special purpose, extra-municipal public agencies to generate debt, raise revenue, and build physical infrastructure networks; and
2. create more direct relationships between the government and the capital markets, thereby better satisfying market requirements for enhanced credit risk management (through municipal debt guarantees, or bond insurance, or policies of securitization, such as the provision of bank letters to offset credit exposure, etc.) in order to pay for infrastructure.

Critics of such public corporations point to their privatistic practices as evidence of the increasing transformation of government into 'business' (Walsh, 1978). Adherents suggest that the privatization of public infrastructure services in general and through special purpose governments and public authorities in particular is part of the century-long move away from the inefficiencies of the state and back to the private sector as the best, most stable source of infrastructure investment (Smith, 1999).

Infrastructure Investment as Public/Private Practice

In many ways, the dominant normative consideration of public infrastructure investment has come full circle, with a twist. Where conditions of scale and cost, and developmental importance made infrastructure investment essentially a public or governmental function at the beginning of the twentieth century, the very same issues are challenging the capacities of the state and forcing new mixes of public and private service delivery at the beginning of the twenty-first century. The government, at national and local levels, is finding it difficult to fully conceive, finance, build, or efficiently maintain modern systems of urban infrastructure. This is held to be equally true for basic infrastructure systems in the emerging urban centers of the developing world (Smith, 1999) and for advanced telematic systems of telecommunication and digital media that are radically transforming the space of urban flows in general (Graham and Marvin, 1996).

Emerging Cites in an Urban World

Nowhere are these arguments more forcefully made than in the literature on infrastructure development in emerging cities. Increasing patterns of economic globalization and migration into cities are central features of global geographic and social development. Sometime in the last decade of the twentieth century, the world became 'urban,' more than 50% of the global population could be found living in towns and cities. The emerging cities of this new urban world face tremendous infrastructural challenges with limited technological and fiscal resources, and yet they represent stable and viable places to ever more people (Clark, 1996).

Just as issues of health, transportation, and scale challenged the United States and Europe at the beginning of the twentieth century, similar patterns, at an even greater scale, challenge the emerging urban centers of Latin American, Asia, and Africa in this century. Millions continue to move to the cities without access to basic infrastructure facilities such as safe water, sanitation, and garbage disposal.

Financial Practices: A Public–Private Mix

While national governments remain the single most important source of infrastructure delivery in the emerging urban world, the costs, which can be as high as 10% of GNP, suggest

that they cannot do it alone. They must attract domestic and international equity capital for private investments in the infrastructure. This requires the move to Western-style financing schemes and patterns of governmental devolution of infrastructure investment, making it a more 'public–private' activity worldwide.

In the past, municipalities (especially in developing countries) were able to combine their own taxes with central government transfers to build public works, but now infrastructure demands make them just as fiscally burdened as their national governments. In addition, they are required by both the national government and international banks to 'forge partnerships with financiers, private operators and constituents' (El Daher, 1997: p. 1). The sources of revenue available to governments to pay for such services generally fall into three categories: (1) taxes, (2) user fees, and (3) other forms of leases (and purchase agreements) and revenue schemes. Depending on their success in using these sources of revenue, cities' infrastructure investments require the organization and constitution of government to accommodate new debt structures, create relationships with the capital markets, and regulate public–private infrastructure agreements. In certain key areas, the infrastructure investment is becoming even more privatized or transferred to the private sector through deregulation, concessions, contracting out, and private management and control agreements.

Access to Private Investment Markets

Considering all of this, a most important signifier of infrastructure investment in cities is their access to the credit markets, that is providing the institutional transparency, strong government regulatory power, debt guarantees, and revenues needed to secure an attractive private investment climate in which to develop infrastructure. While these are common issues of municipal infrastructure finance everywhere in the urban world, in no places are they any more evident than in developing countries where not only can the fiscal strength of the municipality be open to question, but also the very existence of a stable viable domestic credit market can be equally problematic. Cities in developing countries often have more success generating direct private investment in fully privatized public works than in creating the stability of process required of a local municipal bond market.

In a rapidly urbanizing world and uneven investment climate, nongovernmental intermediaries like the World Bank, the IMF, the Inter-American Development Bank, and the Asian Development Bank have become important sources of infrastructure investment policy. They attempt to provide debt guarantee instruments and policy strictures designed to influence the creation of stable domestic capital entities, with the larger purposes of not only institutionalizing long-term markets but also enhancing governmental legitimacy across regimes.

Infrastructure Investment and Urban Development

Infrastructure investment is an important way to view the urban. It represents networks of technology and urban life, binding them together in patterns that are clearly important elements of cities and their development. It comprises a substantial share of the sunk physical capital of cities; materially delivering the networks of heat, light, water, telematics, and power that are central to urban life; without them modern cities do not exist (Graham, 2000: p. 115). The material grids of such components of the urban landscape and their accessibility go a long way toward explaining what is meant by the 'developed' and 'developing' urban world.

The study of infrastructure investment provides an understanding of the state as a site of debt and public regulation and the market as a site of capital and risk-generated requirement. The restructuring of the market into risk niches and intermediaries and the state into devolving agencies of delivery and transfer in both developed and developing urban centers are central characteristics of infrastructure investment as a public-private practice of the new urban world. More generally, the full implementation of infrastructure investment is a key element of the dominant practices of urban political economy: the structuring of government, the creation and regulation of capital markets, the conditions of uneven urban economic development, and the socio-technical networks of service and underservice that represent the distributions of power in the city and globally (Tarr, 1984; Swyngedouw, 1993). In sum, the study of infrastructure investment is the study of urban development.

See also: Telegraph; Telephone; Transportation Planning; Transportation, Supply, and Congestion; Urban Geography.

Bibliography

Anderson, L., 1988. Fire and disease: the development of water supply systems in New England, 1970–1990. In: Tarr, J.A., Dupuy, G. (Eds.), Technology and the Rise of the Networked City in Europe and America. Temple University Press, Philadelphia, PA.

Clark, D., 1996. Urban World/Global City. Routledge, London.

El Daher, S., 1997. Municipal Bond Markets: Prospects for Developing Countries. Urban No. FM-8b. Infrastructure Notes, The World Bank, Washington, DC.

Graham, S., 2000. Introduction: cities and infrastructure networks. International Journal of Urban and Regional Research 24 (1), 114–119.

Graham, S., Marvin, S., 1996. Telecommunications and the City: Electronic Spaces, Urban Places. Routledge, London.

Konvitz, J., 1985. The Urban Millennium: The City-building Process from the Early Middle Ages to the Present. Southern Illinois University Press, Carbondale, IL.

Neutze, M., 1997. Funding Urban Services: Options for Physical Infrastructure. Allen & Unwin, St Leonards, NSW, Australia.

Organization for Economic Co-operation and Development (OECD), 1991. Urban Infrastructure: Finance and Management. OECD, Paris.

Perry, D., 1995. Building the public city: an introduction. In: Perry, D.C. (Ed.), Building the Public City: The Politics, Governance and Finance of Public Infrastructure. Sage, Thousand Oaks, CA, pp. 1–20.

Sbragia, A., 1996. Debt Wish: Entrepreneurial Cities, U.S. Federalism, and Economic Development. University of Pittsburgh Press, Pittsburgh, PA.

Seeley, B., 1993. The Saga of American Infrastructure. Wilson Quarterly. Winter.

Smith, A.J., 1999. Privatized Infrastructure: The Role of Government. Thomas Telford, London.

Swyngedouw, E., 1993. Communication, mobility, and the struggle for power over space. In: Giammopoulos, G., Gillespie, A. (Eds.), Transport and Communications in the New Europe. Belhaven, London.

Tarr, J.A., 1984. The evolution of the urban infrastructure in the nineteenth and twentieth centuries. In: Hanson, R. (Ed.), Perspectives on Urban Infrastructure. National Academy Press, Washington, DC, pp. 4–66.

Walsh, A.H., 1978. The Public's Business: The Politics and Practices of Government Corporations. MIT Press, Cambridge, MA.

Infrastructures of Society, Anthropology of

Jörg Niewöhner, Humboldt University Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of infrastructure refers to the embedded, often invisible technical support structures that help to deliver services to a population or organization, most commonly water, energy, and information. Infrastructures mediate human interaction and shape social organization. Anthropology has developed a relational perspective on infrastructures analyzing them as the ongoing interweaving of embodied social and political choices, moral orders, and technical networks. This approach has much to offer for anthropologists, because it is largely based on ethnographic research, shows a deep commitment to materiality as practice and provides a productive way of thinking through the changing relations of center and periphery. It is an area of research with important intersections into the information sciences and urban studies.

Introduction

There exist a number of topics within anthropology that for long stretches within the history of the discipline, and somewhat surprisingly, suffered the fate of "handmaiden to other anthropological frames and issues" (Munn, 1992, p. 93). Infrastructure is one of them, the body is another. It is as if their ubiquity, the fact that they "[mediate] all reflection and action upon the world" (Lock, 1993, p. 133), renders them moot and implicit in anthropological inquiry. The concept of infrastructure refers first and foremost to modern technical structures such as roads, power lines, or databases that support particular forms of work and social life. Much of anthropological research, however, has taken place in settings where perhaps technical infrastructures did not play a particularly central role in the organization of social life. And where they did, anthropology, for a long time, was not particularly interested in analytically foregrounding material culture and technical artifacts as infrastructure in the modern sense. Hence, the Azande's grain storage vessels featured centrally in Evans-Pritchard's ethnography and his analysis of forms of rational explanation. Yet they were not problematized as part of a technical supply infrastructure that may be implicated in shaping those forms of rational explanation. Infrastructure as an analytical concept in anthropology has thus largely lain dormant until the late 1980s, when research at the interface of ethnomethodology, anthropology, feminist critique, and science and technology studies showed that infrastructure was not simply a technical artifact that supported social life. Rather, infrastructure mediated and shaped social practice (Star and Ruhleder, 1996). This perspective opened infrastructure up as an anthropologically useful concept to think through aspects of the relational work that goes into human interaction and that helps to configure social organization. It has made infrastructure accessible to social and cultural inquiry along new analytic paths, paths that research until then, which had treated infrastructure largely as an object, had not been able to cleave. Neither the structural Marxist analyses of the material basis of the social relations of production nor the detailed attention of the engineers and planners in charge of constructing and maintaining infrastructure as a technical object, turned the concept of infrastructure into one that anthropologists would find particularly useful to think with and to invent around. Today, infrastructure as part of an increasingly relational and practice-oriented thinking is rapidly becoming a well-established topic of anthropological research, particularly at the intersections of sociocultural anthropology, and science and technology studies with important alliances into the information sciences and urban studies.

This article deals with the analytical, epistemological, and methodological issues that have arisen during research on infrastructure since the 1990s charting genealogies where appropriate. The emphasis is on the new and productive ways in which infrastructure has been problematized (Rabinow, 2003). Infrastructure refers to the embedded, often invisible technical support structures that help to deliver services to a population or organization, most commonly water, energy, and information. Yet, these technical support structures are now conceptualized relationally as transient embodiments of social, technical, political, economic, and ethical choices that are building up incrementally over time. Analyzing infrastructure as part of an ecology of infrastructure thus offers a way for the social sciences to inquire into human interaction and social organization, a way that emphasizes the partial connections between structure and agency, and inquires into the 'how' of connecting and its implications. This thought style has much to offer for anthropologists, because it is largely based on ethnographic research, shows a deep commitment to materiality as practice and provides a productive way of thinking through the changing relations of center and periphery.

The article is divided into three core sections. Section Analytical Uses discusses three analytical uses of the concept of infrastructure: ecologies of infrastructure, dys-appearing infrastructure, and infrastructure and statecraft. These are heuristic distinctions that reflect three uses of the concept of infrastructure that are very much aware of and borrow from each other. Section Epistemological Issues discusses epistemological issues arising in and from infra-structural research, most importantly the shift from

infrastructure to infrastructuring. Section Methodological Issues raises a number of methodological points that pertain to research on infrastructuring and that inform a wider debate about the changing practices of ethnography as method.

Infrastructure: Analytical Uses

Ecologies of Infrastructure

Infrastructure is crucial for social organization yet often invisible, physically as well as in social scientific or public discourse. This discrepancy marks the starting point for the research on ecologies of infrastructure that has brought together anthropology, science, and technology studies; history; and information science, and that has been advanced primarily through a focus on informational infrastructures supporting organizations (Bowker and Star, 1999). Ecological infrastructure research has initially developed from the analysis of standards and classifications as practices connecting formal knowledge, social interaction, and organizational structure (Bowker and Star, 1999; Star and Ruhleder, 1996). This thinking evolved into a relational view of infrastructure as a continuous coordinated networking of social organization, moral order, and interweaving layers of technical integration. It refuses to reduce infrastructure to the status of technical object. Following Bateson's argument that one can only ever study relationships and never a 'thing,' ecological research on infrastructure is interested in the manifold relations between technical structure, routines of work, wider scale organizational and technical resources, as well as heterogeneous communities of users – relations always understood as continuously enacted in everyday practice stabilizing a particular albeit contingent, negotiated order (Bowker and Star, 1999; Law, 1994).

Infrastructure seeps into the background, it sediments out and disappears from view. And with it disappear all the technical, social, political, and ethical choices that are necessarily made during the design, construction, implementation and maintenance of any infrastructure. "[Down-stream], the choices made in the building of the […] systems may no longer be obvious or even visible to ordinary users. The civics retreats silently into the infrastructure" (Fischer, 2005). Hence "[u]nderstanding the nature of infrastructural work involves unfolding the political, ethical, and social choices that have been made throughout its development" (Bowker et al., 2010, p. 99) and that often appear at the surface as mere technical decisions.

Born out of this ecological perspective, the following criteria have been established to define infrastructure (abridged from: Star and Ruhleder, 1996, p. 34):

- Embeddedness: Infrastructure is sunk into other structures, social arrangements, and technologies.
- Transparency: Infrastructure does not have to be reinvented each time or assembled for each task.
- Reach: Infrastructure has a temporal or spatial reach beyond a single event or one-site practice.
- Learned as part of membership: The taken-for-grantedness of infrastructure is a *sine qua non* of membership in a community of practice. Strangers and outsiders encounter infrastructure as a target object to be learned about.
- Links with conventions of practice: Infrastructure both shapes and is shaped by the conventions of a community of practice.
- Embodiment of standards: Infrastructure plugs into other infrastructures and tools in a standardized fashion.
- Built on an installed base: Infrastructure does not grow *de novo*; it wrestles with the inertia of the installed base and inherits strengths and limitations from that base.
- Becomes visible upon breakdown: The normally invisible quality of working infrastructure becomes visible when it breaks down.
- Is fixed in modular increments, not all at once or globally. Infrastructure is big, layered, and complex, and hence never changed from above. Changes take time, negotiation, and adjustments.

These criteria are sufficiently generic so as to be applicable to almost any type of infrastructure. They have been widely employed in social scientific research, predominantly in the areas of computer-supported cooperative work and learning environments, ubiquitous computing, and technically mediated forms of health care practices. Recent ethnographic research in this vein reaches from infrastructures of governance and administration in mental health care to urban utilities in the Global South (e.g., Levinson, 2010; Mcfarlane, 2011). More recently, the interest has expanded toward cyber infrastructures (cf Bowker et al., 2010) in domains of knowledge making that have traditionally relied less heavily on technical artifacts, particular the humanities, and in fields that have rapidly risen from local endeavors to truly global networks such as the climate sciences and molecular biology.

Infrastructures build up over time and form layers. It is this temporal dimension of infrastructure, which has interested historians of science in large-scale systems and their infrastructures. Their analyses demonstrate how infrastructures are seldom mere technical solutions to a societal need or want. "We see things and name things differently under different infrastructural regimes" (Star, 1999). Infrastructures mediate the looping between naming and 'ways to be a person' thus shaping how people are moving targets (Hacking, 2006). "Infrastructure does more than make work easier, faster or, more efficient; it changes the very nature of what is understood by work" (Bowker et al., 2010, p. 108). And as the following section will show, it is not only work. Infrastructuring often has more far reaching and existential implications. It is heavily implicated in shaping the very nature of life, labor, and language (Foucault, 1970).

Dys-Appearing Infrastructures

An important strand in anthropology is interested in the 'dys-appearing' of infrastructure (Akrich and Pasveer, 2004). Infrastructure comes to matter in its failing: as acute or chronic infrastructure failure denying access to and fair and equal distribution of the basic means of living. Infrastructure only becomes visible, present, and experienced in its failure to deliver what has been promised by authorities, by development agencies or by narratives of modernity.

This thinking is deeply embedded in the history of anthropology. It has its roots in Marxist historical materialism, translated primarily through the writings of Althusser and Gramsci into the structural Marxism of anthropology in the 1970s, cleansed of its structuralist underpinnings by Bourdieu and particularly Foucault in the 1980s, before opening out into a rich delta of approaches integrating elements of feminist and postcolonial thinking and developing a strong interest in spatial and ecological thinking, particularly in urban studies. The original understanding of infrastructure within this Marxist tradition has been discussed controversially and this is not the place to begin to unravel its complicated genealogy (cf Godelier, 1978). Suffice to say that infrastructure has primarily been discussed as part of the trias superstructure, infrastructure, and ideology, with infrastructure initially understood to be the tangible and material setting within which class struggle needed to operate. Other proponents have conceptualized the differences between super- and infrastructure in terms of function rather than institution. Infrastructure in these terms becomes one particular social form of production analytically integrating elements of the local ecosystem, productive forces, and social relations of production, and in this almost ecological sense forming the place of class struggle. While there is a chronology to these different readings of infrastructure within anthropology, they certainly do not follow each other in evolutionary succession. Rather, the different readings exist side by side emphasizing different analytical and political concerns.

What unites these different readings of infrastructure is the primary concern with processes of subjectivation or interpellation rather than infrastructure per se. What differentiates them is whether they understand this process of making a subject as a process driven by ideology (structural Marxism, Gramsci), by the micropolitics of power/knowledge (post-structuralism, Foucault) or mediated through ecologies of infrastructure (science and technology studies, Star). Empirically, structuralist approaches display a strong tendency to investigate infrastructure as a technical object and a black box. Not its everyday working practices are of interest, but the fact that it historically embodies ideology and thus materially determines the (re)production of an unjust set of relations of production and exchange. Foucault has turned structuralism on its feet concerning himself with the conditions for the emergence of an apparatus, i.e., the physical, administrative and epistemic mechanisms of power. This approach treats infrastructure still as an object, but with an interest in how that object is engaging in forging relations. The concern lies with the mechanisms through which infrastructure as a technical object makes certain connections and denies others, thereby mediating the micropolitics of power/knowledge. This relational thinking paves the way for the last shift, driven primarily by the social construction of technology and material-semiotics: technical infrastructure is drawn into the mangle of practice (Pickering, 1995). Infrastructuring as a material-semiotic practice makes infrastructure available to cultural analysis not only as an object or the carrier of ideology but as a socio-technical phenomenon and practice relating technology, actors, and moral orders in specific ways. Infrastructure ceases to be a black-boxed technical object. It is pried open by practice-theoretical, constructivist inquiry revealing the ongoing work that goes into making it into a seemingly stable, technical object. The sharp Marxist distinction between subject and object, which Foucault already problematized when analyzing technologies of the self as relational manifestations of power, is dissolved altogether in the mangle of practice. This does not mean that research on ecologies of infrastructure is not concerned with interpellation and the making of subjects any longer. Yet interpellation as a matter of practice is not about subjects and objects, where subjects are made and objects are invariant contributors to this process. Rather interpellating as an everyday practice mangles subjects and objects, constantly changing roles, perspectives, and agencies thus rendering the distinction meaningless. Water pipes are objects of design practices one moment and defective agents in water delivery in the next. Utility infrastructuring as a material-semiotic practice thus interpellates 'life as such' (Fassin, 2009), because it embodies and enacts political and ethical choices and helps to shape the "quotidian practices and […] 'microspheres of negotiation' through which [people access resources]" (Gandy, 2008, p. 125).

Infrastructure does not only dys-appear for those marginalized and in the periphery. Infrastructuring instead focuses one's attention on the partial connections between center and periphery (Hannerz, 2001). Infrastructure is commonly designed in the center and here it dys-appears in the design and construction process. Struggles arise over defining the problem that infrastructure is meant to solve, over aligning and legitimizing inevitable sociotechnical choices, and overdealing with uncertainty and ambiguity in long-term planning processes. Yet infrastructure dys-appears in the center always in relation to a representation of the site of its deployment, i.e., some idea of the periphery. In the periphery itself, infrastructure dys-appears differently. It dys-appears because it does not deliver what is needed. It breaks down, it has unintended consequences and side effects, and it is appropriated in ways that change its original intent for better or worse or infrastructure does what the center intended, but that is not what anyone in the periphery wanted or needed. Yet these forms of dys-appearing also relate to design and construction choices made in the center. Infrastructuring then focuses the analytical attention on the partial connections between center and periphery, on the way one sees the other, on the exchanges, connections, collisions, and dependencies between center and periphery on different spatial and temporal scales. In fact, center and periphery only take shape the way they do and for the moments they do, because they are partially connected through processes of infrastructuring in specific ways.

Most of the anthropological research in this trajectory of dys-appearing is concerned with utility infrastructure in the major metropolitan regions in the North and South and convincingly demonstrates just how central infrastructure is "in the construction of the city as 'modern', as a site of capitalist production and expansion, as constitutive of social relations of inequality, and as a space of environmental transformation" (Mcfarlane and Rutherford, 2008). To a degree, this strand of research remains true to its Marxist roots with a continued interest in the hidden materialized networks of fetishized relations of production and demonstrating the instrumental role of infrastructure in mediating power relations. Many recent analyses in this vein oscillate between foregrounding

the political, ethical, and economic choices that go into infrastructure design and operation on the one hand, and on the other hand, the tactics, appropriations, and suffering of those living with the effects of these choices. They oscillate between narratives of top-down power and the manifold resistances against it that are always also articulated in material, i.e., infrastructural terms.

Inspired by the watchful critique of the consequences of increasingly global flows of people, things and knowledge (Appadurai, 1991; Hannerz, 1997), the notion of the assemblage (Collier and Ong, 2005) has advanced a more ecological understanding of infrastructure: "'Infrastructure' designates specific institutional, material, or social conditions through which the functioning of a certain technology, ethical regime, form of regulation, or mode of communication is either enabled or impeded. It inscribes the space and form of limited, finite, and localizable relationships and effects that occupy a certain space and that concretely link – or distinguish and divide – various objects, spaces, techniques, individuals" (Collier and Ong, 2003). This definition firmly establishes infrastructure as a coconstitutive part of cities as urban assemblages and extends the concept from its traditional focus on utilities to include

- the political and media infrastructures, i.e., means and modes of representation, and how they contribute to an infrastructure politics;
- people as infrastructure in constellations where the material and technological means are minimal and forming support networks and communities becomes the most important resource; and
- the emergent informational infrastructures that are rapidly changing from top-down service provision to ubiquitous and increasingly user-generated networks of communication and interaction.

Ubiquitous computing, sensing technologies, cartographic techniques, and manifold forms of 'urban hacking' are rapidly developing into a pervasive information infrastructure woven into the urban fabric and they are involved in reconfiguring modes of representation and participation in processes of infrastructure development and distribution.

Infrastructure and Statecraft

The governance of and control over infrastructure has always been a crucial technique of modern government. In the longue durée of infrastructure development, infrastructure is an integral part of the architectures of modernity, be it as an element of nation-state formation, of countless efforts of modern state governments to impose a singular order on society, or as an element in late modern attempts at governing increasingly complex urban fabrics. The art of governing in modern societies has always depended on "'geopolitics', whereby the state uses technological means to transform the land into both a resource for administration and a symbol" (Mukerji, 1994). If controlling infrastructure is an important technique and strategy of rendering a society governable, it is of little surprise that movements of resistance struggle to (re)gain some control of infrastructures, most recently information infrastructures in movements of political change in North Africa and the Middle East (e.g., Fischer, 2010). Less immediately visible, because it is operating at a much larger temporal and spatial scale, is the intense link between utility infrastructures and the flows of political and economic power at large. The notion of carbon democracy (Mitchell, 2009) focuses historical attention on how the histories of carbon-based fuels, i.e., coal and oil, and Western politics and statecraft are closely aligned throughout the twentieth century. The concentration and control of energy flows shapes democratic possibilities and, during the postwar period, democratic stability and its implicit goal, the growth of markets, have become closely aligned with the circulation of fossil fuels and hence infrastructure and infrastructural decisions.

In anthropology, this historical analysis has been extended into the ethnographic present to investigate energopolitics, i.e., the power over and through energy, in an analogy to bio- and geopolitics (Boyer, 2011). "[…] Power over energy has been the companion and collaborator of modern power over life and population from the beginning" (Boyer, 2011). Particularly as countries all over the world are beginning to face up to the transition to renewable energy sources, existing infrastructures need to be extended or replaced in major national efforts requiring a plethora of political, technical, economic, social, and ethical choices. The transition into renewable energies requires careful infrastructuring in a political and economic field that is dominated by major industrial actors; a technological field that has not settled on key technologies yet and that is still witnessing a multitude of decentralized, often very local initiatives and innovations; and a civil society particularly in the affluent North that so far shows little inclination to reduce energy use or develop other ways of accounting for its externalities. It is an important task of infrastructure research as part of an anthropology of the contemporary to analyze these processes of energy infrastructuring and understand and make visible the social and ethical implications of choices that are often made as technical choices in the here and now but that have significant global consequences far beyond the present.

Infrastructure: Epistemological Issues

Ecological infrastructure research begun to develop in the 1980s, a time in anthropology when important theoretical advances were made through feminist critique, a heightened sense of reflexivity, and the integration of scientific and technological practices into cultural analyses. It is in this intellectual environment that infrastructure ceases to be a technical object or material repository of social and cultural activity. Infrastructure becomes part of material-semiotic practices (Haraway and Jeanne, 1991) and hence part of cultural analysis in very much the same way that science and its laboratories were coming to be understood as intimately and inextricably woven into the fabrics of modern societies (Martin, 1998). Infrastructure becomes legible through the analytical lens of culture; it becomes anthropological strangeness of a second-order, "that of the forgotten, the background, the frozen in place" (Star, 1999, p. 379).

The ecological and relational perspective grants infrastructure a more active role in contributing to material-semiotic

practices. This shift in perspective allows the analysis to take into account that infrastructure never really works the way designers intend. Infrastructure always 'does' unintended and unforeseeable things and it has to be able to do so. Infrastructure is never simply rolled out into everyday life as the term infrastructure implementation suggests. It needs sufficient degrees of freedom to accommodate the heterogeneity and dynamics of social practices. The fit between plan, infrastructure and social practices needs to be produced and continuously reproduced in situated action (Suchman, 1987) and in careful tinkering (Mol et al., 2010). Hence, infrastructure is not any longer passive repository of a homogeneous culture waiting for an anthropologist to take a reading. It is part of material-semiotic practices and is analyzed in its ability to help or hinder the forging of relations. It is analyzed in its ability to stabilize particular configurations, i.e., its ability to shape culture as practice.

Infrastructure in an ecological reading ceases to be a homogeneous technical artifact. Rather it is understood as a fluid technology contributing to practices in multiple ways. Deleuze captures this new sense when writing about the device: "[a multi-linear ensemble] composed of different sorts of lines. And these lines do not frame systems that would be homogenous as such (e.g. the object, the subject, the language). Instead, they follow directions, they trace processes that are always at disequilibrium, sometimes coming close to each other and sometimes getting distant from each other. Each line is broken, is subjected to variations in direction, bifurcating and splitting, subjected to derivations" (cited in Muniesa et al., 2007). It is in this sense that the shift is introduced from infrastructure as a noun to infrastructuring as a verb and as an ongoing practice: "Discussing 'infrastructure' as a noun […] suppresses the variety of material and non-material components of which it consists, the efforts required for their integration, and the ongoing work required to maintain it" (Bossen and Markussen, 2010). Hence, the shift from infrastructure to infrastructuring is not only a methodological shift toward the analysis of infrastructure in the making. It is also an epistemological shift toward infrastructuring as articulation work entangling actors, technologies, and moral orders in specific ways. Infrastructuring is never simply the design and bringing into being of a technical artifact but it is an ongoing attempt at ordering social practices, an engagement in heterogeneous engineering (Law, 1994).

Infrastructure: Methodological Issues

Inversion and Othering

This final section discusses briefly three methodological issues that arise in the context of infrastructural research: infrastructural inversion and othering, relationality as an ethnographic challenge and infrastructure as an ethnographic resource. In the late 1980s, the history of science began to analyze large-scale systems as emergent sociotechnical networks concentrating for the first time on infrastructure and technology rather than the people and ideas these were supposed to enable and advance. Bowker (1994) referred to this shift in focus as 'infrastructural inversion.' Practicing infrastructural inversion means turning figure into ground and vice versa and investigating not an object but relations (Bowker et al., 2010). Particularly in the history of science, this is a powerful methodological shift that puts the emphasis firmly on sociotechnical networks in the making as well as the choices and path dependencies that channel these developments. Yet there is a larger point to figure and ground reversals that anthropology has raised and that is relevant to the analysis of infrastructuring. Marilyn Strathern points out that figure and ground do not relate to each other in a part–whole relationship constituting some kind of totality. "Figure and ground work as two dimensions. They are self-scaling – not two perspectives as it were, but a perspective seen twice, ground as another figure, figure as another ground. Since each behaves as an invariant in relation to the other, the dimensions are not constituted in any totalising way" (Strathern, 1991, pp. 112–113). One important methodological implication is that the ethnographer may try to oscillate between perspectives and attempt different kinds of involvement in the field. Key to such an approach, however, is not the integration of different analyses into a coherent whole, but rather to explore how they relate to each other, always only partially connected, and how the constant failure of any perspective to capture something whole and complete, starts to produce something interesting when put next to other failures.

The analysis of the literature on infrastructure not only in anthropology shows that each discipline names and treats as infrastructure (ground) that which is not its research object (figure). Thus, the social and cultural sciences interested in social organization and human interaction frame the technical environment of particular situations as (technical) infrastructure. The computing and engineering sciences primarily interested in the coordination of technology and algorithms make social interaction their (social) infrastructure. Thus, disciplinary thinking runs the risk of not seeing figure and ground as partially connected but of othering the ground by keeping it invariant, black boxed, and detached from the analysis. Already in 1935, Ludwik Fleck demonstrated how the philosophizing naturalists of the early twentieth century, e.g., the Vienna Circle, saw facts as variable and human thought and logic as fixed, whereas humanist philosophers thought of facts as fixed and of human thought as variable: "It is characteristic that both parties relegate that which is fixed to the region with which they are unfamiliar" (Fleck, 1935/1979, p. 50).

Relationality and Different Modes of Infrastructural Research

The particularly modern epistemological split in attention to either matter or meaning that Fleck points out continues throughout the better part of the twentieth century and is now increasingly dissolved in relational infrastructure research. Researching infrastructure today always attempts to relate three analytical dimensions to each other: "ubiquity, material and semiotic (texture), indeterminacy of the past (identity politics), practical politics of designing infrastructures (how to do it and the (in)visibilities it produces)" (Star, 1999). Relating these dimensions to each other and exploring the connections between them forces anthropologists to continuously cross established modern boundaries of matter and meaning and of disciplinary thought. Anthropological research on infrastructuring thus not only becomes an inherently

transdisciplinary endeavor. It also challenges in different ways the ethnographic method.

First, new models for 'scalable qualitative research' (Bowker et al., 2010) are called for, i.e., the variance in the phenomena being studied ought to be matched by the variance in the size of research teams. While there are many very good reasons for individual fieldwork in many cases, there are equally good reasons in the case of infrastructure phenomena that argue for larger teams able to match the heterogeneity and scale of infrastructure. Infrastructures are often distributed across large spaces, they span center–periphery relationships and they connect many different places. There is also a temporal dimension: changes somewhere in the network may have effects in several other locations at the same time. The spatial distribution and simultaneity of infrastructural phenomena challenges the revered ethnographic principle of co-located i-witnessing.

Second, understanding infrastructural phenomena as material-semiotic practices requires multimodal research (Bowker et al., 2010), i.e., diversity in methods and methodologies each with their own strengths and weaknesses. And it requires analytic means of relating the different outputs together: thick descriptions, narratives, technical data, and maps. These cannot be triangulated in any simple sense to achieve a more comprehensive understanding of the phenomenon. Rather the interesting findings come from understanding how different results partially connect to each other.

Last, new ways of knowing are required if the different methods are not to produce segregated analyses of the social, organizational, technical, and cognitive aspects of infrastructure as if these were somehow sitting on top of and apart from each other and not entangled in highly specific ways. The established textual form of the ethnography has long ceased to be a singular and monolithic narrative. Experiments with perspective, coproduction, and style are widespread. Research on infrastructure encourages anthropology to carry these experiments further into distributed and only partially connected ways of knowing.

Infrastructure as Ethnographic Resource

Another challenge to the co-location of ethnography comes from the 'intimate sciences' such as the humanities (Beaulieu, 2010), where knowledge practices are often difficult to engage ethnographically as they are private and individualized rather than collective and publicly displayed, for example, in laboratories. While Wittgenstein alerted not only anthropology to the fact that language and thinking are also embodied practices, this remains difficult to translate into ethnographic research, not because it is an epistemological impossibility to observe thinking practices, but because most people for many good reasons want to be left alone when thinking. Despite this apparent need for privacy, or maybe because of it, cyber infrastructures are playing an increasingly important role in the humanities strengthening networked knowledge practices. These infrastructures can also be conceived of as a resource for ethnographic research as they offer the possibility to create a co-presence of the ethnographer with the research community in question without requiring co-location. Ethnographic copresence has to do with the alignment of observing and observed practices through information infrastructures, sharing channels of communication; of data management; and of reflexive commentary. It offers possibilities to partake in the practices of knowledge production in asynchronous ways and at a physical distance without being detached. It offers more co-laborative roles to the ethnographer than passive observer, including author, participant, and scholar. The multisitedness of the traveling individual becomes here not a matter of sequential co-location, but of interaction distributed along information infrastructures such as mailing lists, Web sites, or databases.

See also: Center–Periphery Relationships; Modes of Production; Organizational Culture, Anthropology of.

Bibliography

Akrich, Madeleine, Pasveer, Bernike, 2004. Embodiment and disembodiment in childbirth narratives. Body & Society 10, 63–84.

Appadurai, A., 1991. Global Ethnoscapes: Notes and Queries for a Transnational Anthropology.

Beaulieu, Anne, 2010. From co-location to co-presence: shifts in the use of ethnography for the study of knowledge. Social Studies of Science 40, 453–470.

Bossen, Claus, Markussen, Randi, 2010. Infrastructuring and ordering devices in health care: medication plans and practices on a hospital ward. Computer Supported Cooperative Work – The Journal of Collaborative Computing 19, 615–637.

Bowker, Geoffrey C., 1994. Science on the Run: Information Management and Industrial Geophysics at Schlumberger, 1920–1940. MIT Press, Cambridge, MA.

Bowker, Geoffrey C., Baker, Karen, Millerand, Florence, Ribes, David, 2010. Toward information infrastructure studies: ways of knowing in a networked environment. In: Hunsinger, Jeremy (Ed.), International Handbook of Internet Research. Springer, New York, pp. 97–117.

Bowker, Geoffrey C., Star, Susan Leigh, 1999. Sorting Things Out: Classification and Its Consequences. MIT Press, Cambridge, MA.

Boyer, Dominic, 2011. Energopolitics and the anthropology of energy. Anthropology News 52, 5–7.

Collier, Stephen J., Ong, Aihwa, 2003. Oikos/Anthropos: rationality, technology, infrastructure. Current Anthropology 44, 421–426.

Collier, Stephen J., Ong, Aihwa, 2005. Global assemblages, anthropological problems. In: Ong, Aihwa, Collier, Stephen J. (Eds.), Global Assemblages. Technology, Politics, and Ethics as Anthropological Problems. Malden & Oxford Blackwell Publishing.

Fassin, Didier, 2009. Another politics of life is possible. Theory Culture & Society 26, 44–60.

Fischer, Michael, 2005. Technoscientific infrastructures and emergent forms of life: a commentary. American Anthropologist 107, 55–61.

Fischer, Michael M.J., 2010. The rhythmic beat of the revolution in Iran. Cultural Anthropology 25, 497–543.

Fleck, Ludwik, 1935/1979. Genesis and Development of a Scientific Fact. University of Chicago Press, Chicago.

Foucault, Michel, 1970. The Order of Things. Random House, New York.

Gandy, Matthew, 2008. Landscapes of disaster: water, modernity, and urban fragmentation in Mumbai. Environment and Planning A 40, 108–130.

Godelier, Maurice, 1978. Infrastructures, societies, and history. Current Anthropology 19, 763–771.

Hacking, Ian, 2006. Kinds of people: moving targets. British Academy Lecture 10, 1–18.

Hannerz, Ulf, 2001. Center–periphery relationships. In: Smelser, Neil J., Baltes, Paul B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Elsevier Science.

Hannerz, Ulf, 1997. Flows, Boundaries and Hybrids: Keywords in Transnational Anthropology.

Haraway, Donna, J., 1991. Simians, Cyborgs, and Women: The Reinvention of Nature. Routledge, New York.

Law, John, 1994. Organizing Modernity. Blackwell, Oxford, UK; Cambridge, MA, USA.
Levinson, Jack, 2010. Making Life Work. Freedom and Disability in a Community Group Home. Minnesota University Press, Minnesota.
Lock, Margaret, 1993. Cultivating the body: anthropology and epistemologies of bodily practice and knowledge. Annual Review of Anthropology 22, 133–155.
Martin, Emily, 1998. Anthropology and the cultural study of science. Science, Technology & Human Values 23, 24–43.
Mcfarlane, Colin, 2011. The city as assemblage: dwelling and urban space. Environment and Planning D: Society and Space 29, 649–671.
Mcfarlane, Colin, Rutherford, Jonathan, 2008. Political infrastructures: governing and experiencing the fabric of the city. International Journal of Urban and Regional Research 32, 363–374.
Mitchell, Timothy, 2009. Carbon democracy. Economy and Society 38, 399–432.
Mol, Annemarie, Moser, Ingunn, Pols, Jeannette (Eds.), 2010. Care in Practice. On Tinkering in Clinics, Homes and Farms. transcript, Bielefeld.
Mukerji, Chandra, 1994. The political mobilization of nature in seventeenth-century French formal gardens. Theory and Society 23, 651–677.
Muniesa, Fabian, Millo, Yuval, Callon, Michel, 2007. An introduction to market devices. The Sociological Review 55, 1–12.
Munn, Nancy D., 1992. The cultural anthropology of time: a critical essay. Annual Review of Anthropology 21, 93–123.
Pickering, Andrew, 1995. The Mangle of Practice – Time, Agency & Science. Chicago University Press, Chicago.
Rabinow, Paul, 2003. Anthropos Today. Reflections on Modern Equipment. Princeton University Press, Princeton.
Star, Susan L., 1999. The ethnography of infrastructure. American Behavioral Scientist 43, 377–391.
Star, Susan L., Ruhleder, Karen, 1996. Steps toward an ecology of infrastructure: design and access for large information spaces. Information Systems Research 7, 111–134.
Strathern, Marilyn, 1991. Partial Connections. Rowman & Littlefield Publishers, Savage, MD.
Suchman, Lucy A., 1987. Plans and Situated Action. The Problem of Human Machine Communication. Cambridge University Press, Cambridge.

In-House Counsel

Robert E Rosen, School of Law, University of Miami, Coral Gables, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In-house counsel refers to professionally trained employees working for an organization's law department. They may not be the only legally trained individuals working for the organization, nor the only employees responsible for the legal work of the organization. They are defined by working in a legal department and reporting to a chief legal officer/general counsel. Sociological research on in-house counsel has focused on in-house counsel being both organizational and professional actors.

In-house counsel are legally trained employees who have responsibilities for some of the law jobs of the organizations employing them. The phrase 'legally trained' does not necessarily mean 'lawyer' or 'attorney,' as in some parts of the world paralegals or notaries hold these positions too. Organizations having needs for legal services may resolve them internally by employing those externally trained in the law or by allocating authority over law jobs to those who did not externally receive legal training (Llewellyn, 1930; Abbott, 1988). Organizations also can resolve their legal needs externally. There is a market of vendors offering to help businesses meet their legal needs, most notably law firms who serve as 'outside counsel.' From a professional perspective, the choice between an in-house (or inside) counsel and outside counsel is simply a 'make or buy' (Williamson, 1975) business choice (Schwarcz, 2008), but it is a choice that takes place in the presence of conflicts over business prerogatives, professional privilege, and regulatory legitimacy.

Delimiting the Role of In-House Counsel

Many individuals with external legal training work for organizations in other than distinctly legal capacities. In most of the world, university legal study usually does not lead graduates into careers as legal professionals (Twining, 1994). These graduates take a facility in the law into whatever positions they assume. Others may leave a career in law to work for organizations in positions other than that of in-house counsel (Rosen, 2002a). The jobs performed by legally trained employees inside corporations are not coextensive with law-related jobs.

In some countries, including the United States, lawyers may combine being lawyers with careers as organizational actors. Legally trained executives thus may be subject to professional sanction even when they are not in-house counsel. This allows the professional association to sanction lawyers for their organizational activities (see, e.g., In re Nixon, 1976) (disbarment of Richard Nixon for his actions not as a lawyer, but as U.S. President), theoretically encouraging lawyers to act as organizational gatekeepers. On the other hand, corporations can claim benefits – primarily the ability to assert the professional privilege to keep communications confidential – because they have chosen to hire a member of the legal profession (DeStefano, 2009). In most of the world, however, the professional privilege is much more constrained (see, e.g., Akzo Nobel, ECJ, 2010), as are the professional responsibilities of lawyers when they are not acting before a court (Van Houtte, 1999) (European countries).

In-house counsel are not distinguished from other law-trained employees simply because they are primarily employed for their law training and practice experience. Corporate finance departments, for example, may employ law graduates, some with significant careers as lawyers, to perform law work. These individuals, however, have not been considered to be in-house counsel. Although they perform legal work and are law graduates, they report to the corporate finance department, not to the organization's general counsel. An organization's *law department*, managed by the chief legal officer/general counsel, gives birth to and is the home of in-house counsel.

Legal work is performed in organizations also by those who are neither legally trained nor part of the legal department. Executives, staff members, regulatory specialists, managers, and salespeople may be responsible for a number of law jobs. Non-legally trained employees, for example, may negotiate, write, and sign contracts. Organizations also may have such continuing relations with regulators that they develop specialized subunits, separate from the legal department, such as employment or environmental relations offices, with responsibilities for legal compliance and dispute resolution.

There are many professions that compete with the legal one to supply businesses with legal services. Human resources is one such group (Edelman, 1990). The movement of work into the legal department indicates a victory for the legal profession, a recognition of its jurisdictional claims. But work may leave the legal department too; at the time of writing this article, for example, many compliance departments in the United States are emerging in response to new legal regulations that demand compliance activities. These compliance departments may be staffed by non-lawyers.

In sum, with in-house counsel organizations delimit a distinctly legal capability, but in-house counsel's responsibilities are neither coextensive with the organization's law jobs nor comprehensive of organizational law compliance. Legal departments and their employed lawyers provide the organization with a professionally managed legal capability (Mackie, 1989). The scope of the work allocated to the legal department,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86016-X

and consequently the work delimited to inside counsel, depends importantly on the legal regime and its professionalization and regulation projects.

Whereas in the past, "[e]very country had its own formal and informal understandings about what it means to be a lawyer and how law should be practiced, … in an increasingly interdependent world, these norms inevitably come into contact" (Wilkins, 2012: p. 275) and are subject to hegemonic market forces. Thus, there may be emerging from the centers of finance a general understanding of the roles to be played by in-house counsel (Wilkins and Papa, 2013).

In-House Counsel as Organizational Actors

In-house counsel vary in their work responsibilities. Some organizations make in-house counsel rather than buying their legal services across the market to increase the efficient delivery of recurrent predictable legal work. Others may hire in-house counsel because firm-specific knowledge is needed to provide rather routine legal services. Others may hire in-house counsel in imitation of other organizations. Still others risk reducing organizational flexibility and hire in-house counsel to increase the organization's coordination and control of its legal services (Chayes and Chayes, 1985).

The extent to which in-house counsel assume organizational responsibilities has varied greatly (*compare* Miyazawa, 1986 on Japanese corporations in the United States *with* Rosen, 2010b(1984) on California corporations). In-house counsel perform a staff function, but, either individually or organizationally, they also may command executive power. In-house counsel may extend beyond fulfilling a staff function by making professional jurisdictional claims, for example, by demanding and gaining a voice in the choice of remedy for legal problems. In-house counsel also may extend beyond fulfilling a staff function on the basis of personal qualities of the lawyer or a demonstrated history of knowing what the organization needs.

In the United States, in-house counsel make professional claims to organizational power. The rules of the legal profession require in-house counsel and legal departments to have access to the Board of Directors (ABA, Rule 1.13). Some in-house counsel secure these organizational powers by emphasizing their skills in handling legal risk. Others do so by emphasizing their knowledge of the risk levels of the corporate cultures. In exercising executive power, in-house counsel decide or help decide for the corporation. Some in-house counsel, formally or informally, become company managers at the same time as they are members of the legal department. In the United States, organizations 'professionalize' themselves by hiring in-house counsel (Parker, 2010).

Consistent with the success of the law's professionalization project, in the United States, it is not unusual for legal training and practice experience to lead individuals to executive positions in corporations. But whether the CEO holds a JD also depends on how the organization perceives the importance of regulatory compliance, the prerogative of business norms, and competition with employees from other disciplines, such as finance (Rosen, 2002b).

Today, in the U.S. business sector, many in-house counsel are involved in compliance issues and other strategic decisions that gain them access to the ultimate organizational decision-makers. U.S. in-house counsel have won the right to monitor and review legal services produced in departments other than the law department as well as that purchased across the market and produced by outside counsel. Often, in-house counsel even are sought for their judgment about decisions where legal risks do not predominate (Rosen, 1989).

In many other economies, corporate in-house counsel less frequently are ceded executive responsibility (see, e.g., Van Houtte, 1999; Daly, 1997). In many places in the business world today, in-house corporate counsel resembles U.S. corporate counsel of an earlier era. Although situations varied depending on the trust and confidence particular in-house counsel attained, in-house counsel used to perform only scrivener and secretarial tasks, carrying out decisions reached without their significant input (Rosen, 1989). But even in economies where in-house counsel have few responsibilities, global corporations located there use in-house counsel more in line with the U.S. model (Daly, 1997).

David Wilkins outlines strong forces leading toward the export of the U.S. model of in-house counsel (Wilkins, 2012). In the United Kingdom, this has already occurred (Sako, 2011). Foreign direct investment, with its attendant knowledge and capital flows, is fostering this export to other countries. Among the BRICS countries (Brazil, Russia, India, China, and South Africa), Wilkins and his team find the increasing export of the U.S. in-house counsel model, everywhere but Russia. Wilkins' findings are consistent with Hendley's findings in the late 1990s (Hendley et al., 2001; Hendley, 2010) that Russian in-house counsel did low-level work, had a marginal position in the company, and were reluctant to exercise even the most obvious gatekeeping function. Whether these findings are still representative, however, can be questioned (see Martindale-Hubbell, 2009). As Liu notes, the role of lawyers and the legal profession are moving targets and one can only get a partial and incomplete picture at any point in time (Liu, 2012). This makes renders most important the general processes depicted by Wilkins.

In assuming executive power, in-house counsel risk professional dominance. As a matter of professional ethics in the United States, lawyers cede most final forms of control to their client, the organization (ABA, Rule 1.2). Decisional authority over the scope and content of legal work, however, may be allocated within an organization by dynamics driven by other than the norms of the legal profession.

Outside the business sector, in-house counsel for organizations – especially public interest organizations – face the possibility and are cognizant of the attendant responsibilities of performing executive tasks for the organizations or causes they take as clients. Cause lawyers fear making decisions that are not properly theirs to make. They also fear not empowering those with whom they work (Sarat and Scheingold, 1998).

In-House Counsel as Professional Actors

Professions whose professionalization projects stress autonomous employment as a hallmark of the profession – the 'liberal' professions – may denigrate employed lawyers for lacking professional independence. Although organizational

employment had not been derogated in the United States, during the progressive era, in-house counsel began to be called 'kept lawyers' (Rosen, 2010a). In the European Union, unlike self-employed lawyers, in-house counsel are not legally privileged to protect the confidentiality of their work product or client communications (Akzo Nobel, ECJ, 2010). In many civil law countries, a lawyer must resign from the bar before taking salaried employment. In liberal professions, a 'heteronomously employed' legally trained individual who performs legal work, such as an in-house counsel, may not be professionally respected as a 'lawyer.'

The substantive claim in the professional denigration of in-house counsel is their supposed lack of professional independence. Organizational control over the work of professionals, from a legal perspective, creates a potential conflict of interest. The in-house counsel may not pursue professionally proper work, instead, they may be led by organizational interests. Nelson and Nielsen (2000) described three ideal types of inside counsel. One, the 'cop,' would police organizational interests, pursuing legal ends. The second, the 'counsel,' would better align legal services with the interest of the organization and, while not negating legal ends, would maximize corporate ability to 'comply' with the law, raising issues of 'creative compliance' (McBarnet, 2004). Nelson and Nielsen's third ideal type was the inside counsel as 'entrepreneur,' who understood the law (and his or her job) to be a means for making money for the corporation. An entrepreneurial inside counsel may misalign the corporation with legal goals.

The claim that inside counsel lack independence suggests that others have greater amounts of independence. The professional independence of in-house counsel is relative to that which can be bought in the market. Sung Hui Kim compared inside and outside counsel by their (1) 'willingness to interdict' illegal activities, (2) 'willingness to monitor' organizational activity, (3) 'capacity to monitor,' and (4) 'capacity to interdict' (Kim, 2008). Kim argued that inside counsel have increased 'willingness' and 'capacity to monitor,' making them superior to outside counsel as protectors of legality in ordinary cases. On the other hand, obedience pressures in extraordinary cases may make inside counsel less able than outside counsel to interdict corporate wrongdoing. In other work, Sung Hui Kim has examined the psychological pressures that may keep inside counsel from pursuing professional ends even when there is not strong corporate resistance to doing so (Kim, 2005, 2011). Especially as compared with outside counsel, others contend that inside counsel are not only able but are also willing to be carriers of legal independence inside the company (Hackett, 2012) (but see Wallace, 1995).

The law jobs performed by in-house counsel may be bought by the corporation by contracting with external providers. In the marketplace, there are sellers of legal services in professional law firms, multidisciplinary practices, legal process outsourcing services, and competing professions. In-house counsel compete with autonomously employed lawyers who compete with accounting–consulting firms, regulatory engineering specialists, and business advisors of many sorts.

In-house counsel, in the United States, have become the purchasing agents for the corporation in regards to legal services (compare Rosen, 1989 with Rosen, 2002b). This hiring control has altered the distribution of deference within the legal profession, with independently employed lawyers no longer denigrating in-house counsel. In fact, "He who pays the piper calls the tune," describes the current professional hierarchy in corporate law, almost as much as it attests to mere business power. As Donald Langevoort puts it, "many outside lawyers are seeking to channel the values, language, habit, and mind-set of in-house counsel elites in order to become synchronous with their preferences." (Langevoort, 2012: p. 497).

As expert purchasers of the work of outside legal service providers, inside counsel increase the organization's ability to monitor the legal services purchased and consequently may decrease external providers' abilities to be independent. Sometimes, in-house counsel may demand that external providers be independent. Other times, in-house counsel's influence on whom and when to hire in the external market can decrease external providers' willingness and capacity to exert their independence. In-house counsel also decrease the abilities of external providers to be independent by inhibiting personal relations between outside providers and organizational members and by limiting the time frame in which the external providers respond or the problems they are posed. By retaining authority to select external providers, to define the goals of their work and to define the problem which is the subject of legal work, organizations control what legal services they purchase in the market, decreasing the professional independence of external providers (Kagan and Rosen, 1985).

The independence of in-house counsel also is relative to the organizational roles they assume. First, in-house counsel may be delegated or may assume organizational decisional authority. In-house counsel vary in the extent to which they are assigned thick responsibility in their organizations. Second, organizations differ not only in their enacted cultures but also in the extent to which they seek to gain the advantages of in-house counsel's commitment to the organization by supporting job enrichment schemes that may mirror the rents of professional privileges. In-house counsel may become trusted organizational actors and may use that trust to exercise what could be seen as professional independence. And third, organizations differ in their internal politics so that in-house counsel may have opportunities to assert their independence by bargaining and other political activities.

Legal professions differ in where and how they establish the lack of plasticity in professional work that is deemed to warrant the assertion of professional independence. During much of this century, there have been calls for in-house counsel to serve as consciences for their organizations. In the U.S. business sector, where in-house counsel have attained organizational power, this call is seldom heard. Nor on the agenda are limits on what information about the law in action in-house counsel may professionally provide. As risk management replaces law compliance as the legal department's governing philosophy, regulatory inaction and inattention – not only legal norms – become part of the calculus of legal advice. Students of professions thus may study inside counsel to gain information about the plasticity of the professional legal role.

In their work, in-house counsel attest to the law's legitimacy for the organizations they serve and apply professional

knowledge. Yet in-house counsel are employees, and therefore are subject to workplace norms and business priorities that may be determined by executives outside their control. Inside counsel and their work create subjects for research for students of both organizations and professions.

See also: Center–Periphery Relationships; Modes of Production; Organizational Culture, Anthropology of.

Bibliography

ABA (American Bar Association), 2013. Model Rules of Professional Conduct. American Bar Association Press, Chicago.

Abbott, A., 1988. The System of Professions: An Essay on the Division of Expert Labor. University of Chicago Press, Chicago.

Chayes, A., Chayes, A.H., 1985. Corporate counsel and the elite law firm. Stanford Law Review 37, 277–300.

Daly, M.C., 1997. The cultural, ethical and legal challenges in lawyering for a global organization: the role of the general counsel. Emory Law Journal 46, 1057–1111.

DeStefano, M., 2009. The corporate attorney-client privilege: third rate doctrine for third party consultants. Southern Methodist University Law Review 62, 727–802.

Donnell, J., 1971. Corporate Counsel: A Role Study. Bureau of Business Research, Graduate School of Business, Indiana University, Bloomington.

Edelman, L., 1990. Legal environments and organizational governance: the expansion of due process in the workplace. American Journal of Sociology 95, 1401–1440.

Hackett, S., 2012. Corporate counsel and the evolution of practical ethical navigation: an overview of the changing dynamics of professional responsibility in in-house practice. Georgetown Journal of Legal Ethics 25, 317–340.

Hendley, K., 2010. The role of in-house counsel in post-Soviet Russia in the wake of privatization. International Journal of the Legal Profession 17, 5.

Hendley, K., et al., 2001. Agents of change or unchanging agents? The role of lawyers within Russian industrial enterprises. Law and Social Inquiry 26, 685.

Kagan, R., Rosen, R.E., 1985. On the social significance of large law firm practice. Stanford Law Review 37, 399–443.

Kim, S.H., 2005. The banality of fraud: re-situating the inside counsel as gatekeeper. Fordham Law Review 74, 983–1077.

Kim, S.H., 2008. Gatekeepers inside out. Georgetown Journal of Legal Ethics 21, 411–463.

Kim, S.H., 2011. Naked self interest? Why the legal profession resists gatekeeping. Florida Law Review 63, 129–162.

Kolvenbach, W., 1979. The Company Legal Department: Its Role, Function and Organization. Kluwer-Deventer, The Netherlands.

Langevoort, D.C., 2012. Getting (too) comfortable: in-house lawyers, enterprise risk, and the financial crisis. Wisconsin Law Review 2012, 495–519.

Liu, S., 2012. Palace wars over professional regulation: in-house counsel in Chinese state-owned enterprises. Wisconsin Law Review 2012, 549.

Llewellyn, K.N., 1930, 2012. The Bramble Bush: On Our Law and Its Study. Quid Pro, New Orleans.

Mackie, K.J., 1989. Lawyers and Business and the Law Business. Macmillan Press, London.

Martindale-Hubbell LexisNexis, 2009. CCE In-House Counsel Survey.

McBarnet, D., 2004. Crime, Compliance and Control. Ashgate, Surrey.

Miyazawa, S., 1986. Legal departments of Japanese corporations in the United States: a study on organizational adaptation to multiple environments. Kobe University Law Review 20, 97–162.

Nelson, R.L., Nielsen, L.B., 2000. Cops, counsel, and entrepreneurs: constructing the role of inside counsel in large corporations. Law and Society Review 34, 457–490.

Parker, C.E., 2010. The Open Corporation: Effective Self-Regulation and Democracy. Cambridge University Press, Cambridge.

Rosen, R.E., 1989. The inside counsel movement, professional judgment and organizational representation. Indiana Law Journal 64, 479–553.

Rosen, R.E., 2002a. Educating law students to be business leaders. International Journal of the Legal Profession 9, 27–51.

Rosen, R.E., 2002b. "We're all consultants now": how change in client organizational strategies influences change in the organization of corporate legal services. Arizona Law Review 44, 637–684.

Rosen, R.E., 2010a. Rejecting the culture of independence: corporate lawyers as committed to their clients in special issue: law firms, legal culture, and legal practice. In: Sarat, A. (Ed.), Studies in Law, Politics and Society, vol. 52, pp. 33–71.

Rosen, R.E., 2010b(1984). Lawyers in Corporate Decision-Making. Quid Pro Books, New Orleans, Louisiana.

Sako, M., 2011. General Counsel with Power? Novak Druce Centre for Professional Service Firms, Oxford.

Sarat, A., Scheingold, S. (Eds.), 1998. Cause Lawyering: Political Commitments and Professional Responsibilities. Oxford University Press, New York.

Schwarcz, S.L., 2008. To make or to buy: in-house lawyering and value creation. Iowa Journal of Corporation Law 33, 497–575.

Spangler, E., 1986. Lawyers for Hire: Salaried Professionals at Work. Yale University Press, New Haven.

Twining, W., 1994. Blackstone's Tower: The English Law School. Sweet & Maxwell, Andover.

Van Houtte, J., 1999. Law in the world of business: lawyers in large industrial enterprises. International Journal of the Legal Profession 6, 7–25.

Wallace, J.E., 1995. Organizational and professional commitment in professional and nonprofessional organizations. Administrative Science Quarterly 49, 228–255.

Wilkins, D., 2012. Is the in-house counsel movement going global? A preliminary assessment of the role of internal counsel in emerging economies. Wisconsin Law Review 2012, 251.

Wilkins, D., Papa, M., 2013. The rise of the corporate legal elite in the BRICS: implications for global governance. Boston College Law Review 54, 1149.

Williamson, O.E., 1975. Markets and Hierarchies: Analysis and Antitrust Implications, a Study in the Economics of Internal Organization. The Free Press, New York.

Cases

Akzo Nobel Chemicals Ltd and Accrues Chemicals Ltd v European Commission, 14 September 2010. [2010] EUECJ C-550/07.

In Re Nixon, 8 July 1976. 53 A.D.2d 178; 385 N.Y.S.2d 305; 1976 N.Y. App. Div. LEXIS 12500.

Injuries and Accidents: Psychosocial Aspects

Sharon Clarke, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C.W. Roberts, volume 11, pp. 2697–2702, © 2001, Elsevier Ltd.

Abstract

Unintentional injuries are a major public health problem in developed and rapidly developing countries. The causes of injuries and accidents may be studied from an epidemiological or psychological perspective. Psychological research in this area has identified personality traits, error proneness, demographic factors (such as age), risk-taking behavior, general life stress, and work-related stress as underlying factors in accidents. Injury prevention efforts have focused on safety education, modification to the environment, and behavioral change. Further research is needed to explore the most effective means of intervention to prevent accidents and injuries in future.

Accidental or unintentional injuries are a leading cause of death in developed countries (WHO, 2008; US National Centre for Injury Prevention and Control, 2010). For example, in the United States, unintentional injuries result in more than 180 000 deaths, an estimated 2.5 million people hospitalized, and an estimated 31.6 million people treated in hospital emergency departments every year (US National Centre for Injury Prevention and Control, 2010). In addition to the human costs associated with injuries, there are also substantial economic costs in terms of health care and lost productivity, affecting families, communities, and societies. Accidents and injuries can occur across many types of activity, including accidents at work, road traffic accidents, accidents in the home, and during leisure activities. The leading cause of unintentional injuries varies by age group, with young children most likely to be injured in the home (including unintentional suffocation, drowning, and poisoning), while older children and young adults (5–24 years of age) are most likely to be victims of motor vehicle crashes (WHO, 2008).

Advances in safety equipment, technology, and safe practices have combined to reduce the risk of death from unintentional injuries over recent years. For example, over the previous 30 years, child injury rates have declined by up to 50% in many developed countries; nevertheless, unintentional injuries remain a significant problem, still accounting for 40% of all child deaths in the developed world (WHO, 2008). The scientific study of accidents and injuries can help us to further understand the underlying causes of accidents and how to prevent them, as individuals, organizations, and societies.

Scientific Study of Injuries and Accidents

The study of injuries and accidents is multidisciplinary, drawing on research conducted across a number of different scientific disciplines. Scientific research aims to develop the knowledge of how and why injuries result from unintentional events (or accidents) and how they can be prevented or controlled.

Epidemiological Approach

One approach to the study of injuries is based on epidemiology, which identifies the multiple factors leading to injuries. Injuries occur due to the interaction of a person with an environmental hazard that leads to bodily damage within a short period of time (to distinguish injury from disease, which manifests over a longer time period; Haddon, 1980).

The epidemiological approach is focused largely on description, rather than analysis, of the causes of accidents. For example, a cyclist is injured when knocked from his or her bicycle by a car when cycling on a busy road; the injury can be described in terms of the 'host' (e.g., cyclist), the 'agent' (e.g., other road user), and the 'environment' (e.g., condition of the road, volume of traffic, time of day, etc) based on the work of Haddon (1980). Such an approach provides an insight into patterns and trends of injuries, leading to a better understanding of their occurrence, as well as their severity and impact. For example, an understanding of injury mechanisms, such as head trauma, can help in the design of effective helmets to prevent or reduce the severity of injuries (e.g., helmets worn by cyclists, horse riders, or skiers, where the risk of head injury is high).

Psychological and Sociological Approach

A more analytical approach to the study of accidents and injuries focuses on understanding the underlying causes of accidents that lead to injuries. Research in this area has demonstrated the substantial role played by human factors, such as human error, ergonomics (interaction between people and equipment), personality, and risk-taking behavior, as well as broader factors, such as organizational policies and practices, cultural norms, and social climate.

Accident causation theories emphasize that multiple factors combine to cause accidents, rather than having a single cause. For example, Heinrich's domino theory (1932) links accident causation to a sequence of events in which one event triggers the next (like the falling of dominos). Events in the sequence may relate to characteristics of the environment (e.g., unsafe conditions), as well as characteristics of people (e.g., unsafe behavior). More recently, accident theories have broadened their scope to consider the interaction of factors (e.g., socio-technical system theories) and the wider context of the system (e.g., macroergonomic theories) (Khanzode et al., 2012). These theories emphasize the importance of understanding the interaction between people and the wider context (e.g., organizational, environmental, and societal factors) in the

causation of accidents. For example, an employee may injure himself or herself at work, when using a piece of equipment; accident causation theories focus on understanding the role of human factors (such as carelessness, rule-breaking, and lack of attention), ergonomics (such as the design of the equipment and its fit with the person using it), job-related factors (such as time pressure and workload), and wider organizational factors (such as the quality of supervision and the safety culture).

Research concerned with accident causation is largely the domain of psychologists, sociologists, and behavioral scientists. Understanding people's behavior not only provides an insight into how accidents and injuries occur, but also helps develop effective strategies to ensure that people use safety equipment, such as car seat belts and safety helmets, which have been designed to protect them.

Causes of Injuries and Accidents

Accidents were once viewed as 'acts of God,' which cannot be predicted or controlled. Indeed, if we speak of something happening 'by accident' we usually mean it was unintentional, unavoidable, and happened by chance. Although the negative consequences of accidents are unintended, most accidents do involve an element of chance; however, research has shown that accidents are most frequently caused by a combination of multiple factors, many of which can be predicted and controlled. Not all factors can be controlled, but understanding their influence allows us to develop interventions to mitigate their effects.

Accident Proneness

You might expect that, all things being equal, each one of us has the same likelihood of being involved in an accident. However, research has demonstrated that this is not the case. In fact, accident data consistently reveal that a very few people are involved in a disproportionately high number of accidents (Visser et al., 2007). The discovery of 'accident repeaters' led early researchers to hypothesize that some people are accident-prone, meaning that they suffer injuries more often than would be expected by chance, due to a stable set of personal characteristics. However, despite statistical evidence supporting the idea of accident proneness, researchers have had little success in identifying an underlying set of psychological traits or characteristics that make some people accident-prone and others not (Clarke, 2011).

There is evidence to suggest that accident proneness is stable over time (af Wåhlberg and Dorn, 2009), but an enduring set of psychological characteristics that can account for this stability has not been identified. However, psychological research has identified a number of distinct characteristics that are significantly associated with higher accident involvement, including personality traits, decision-making style, and the tendency to suffer cognitive failures, such as forgetting things.

Personality and Other Individual Differences

Information-processing skills are required to avoid accidents, as people must be able to rapidly recall information, identify problems quickly, and react swiftly to unexpected situations. Indeed, analyses of accident causes have found that ineffective processing of information, lack of attention, and poor judgment are the most common precursors of accidents (Brown, 1990). These reflect the psychological phenomenon known as 'cognitive failure,' which is characterized by forgetfulness and distractibility. It relates to the ability to distribute attention across multiple tasks, particularly under stress; this is a cognitive ability that is relatively stable over time, but is not specifically related to intelligence. People with a higher tendency toward cognitive failures (i.e., those who are error-prone) are more likely to have accidents (Wallace and Vodanovich, 2003). They are also more vulnerable to the symptoms of stress, such as depression, anxiety, difficulty sleeping, and general ill health.

Clarke and Robertson (2005) conducted a comprehensive review of studies linking personality traits to accident involvement; they found that people who are more agreeable (pleasant, tolerant, and diplomatic) and conscientious (dependable, compliant, and self-disciplined) are involved in fewer accidents overall. At the opposite pole of these personality traits were those who were most likely to be involved in accidents. These people tended to be more hostile, belligerent (low in agreeableness), impulsive, and noncompliant (low in conscientiousness). Personality traits predispose people to behave in certain ways, which can increase their chances of being involved in an accident. For example, people who are low in agreeableness tend to respond aggressively to situations, take risks, and break rules, increasing their risk of an accident. Clarke and Robertson's (2005) review also looked at work and traffic accidents separately. They found that people who are extraverted (sociable, outgoing, and sensation-seeking) are involved in more traffic accidents, while those who are neurotic (emotionally unstable, anxious, and worried) have more accidents at work. Extraverts tend to be sensation-seekers who are more likely to take risks, such as driving too fast; they are also more prone to fatigue-related driving errors. Neurotics tend to experience more negative emotions and cope less well with stressful situations, making them more likely to be injured at work.

In summary, psychological research has shown that some personality traits affect people's ability to concentrate, manage stressful situations, succumb to fatigue, take risks, and break rules, and so influence their likelihood of having an accident and being injured.

Demographic Factors

Age is a demographic factor associated with accident liability. Both older (>64 years) and younger (<30 years) age-groups have greater frequency of accidents and injuries compared to middle-aged people. It is a common assumption that younger people are accident-involved because they are immature risk-takers, while older people are more accident-involved because of the decline in their mental faculties. However, research suggests that both views may be stereotypes and not fully supported by research.

As we grow older, our liability to suffer accidents and injuries tends to increase. For example, older drivers experience an increased probability of accidental injury of up to 45%

compared to other drivers. This increased injury liability is not necessarily linked to a loss of driving skills, but may be related to other psychological mechanisms associated with aging, such as changes in circadian and sleep systems. For example, older drivers tend to have more accidents in afternoons due to drowsiness than younger drivers (Smolensky et al., 2011). Moreover, the effect of higher accident liability is reduced if driving distance is taken into account.

Higher accident rates among younger drivers (<30 years) are often viewed as resulting from reckless behavior, which tends to be outgrown as individuals mature. Indeed, research indicates that younger drivers and young men, in particular, commit the highest levels of driving violations. However, overtly risky behavior only accounts for a small proportion of young people's accidents, which occur largely as the result of errors (McKnight and McKnight, 2003). Higher crash rates tend to disappear rapidly with increasing experience, suggesting that the majority of accidents are caused by inexperience rather than dangerously risky behavior.

Recent research looking at the effect of aging in the workplace has suggested that 'work ability' is more important than chronological age, as people tend to age at different rates (Costa and Sartori, 2007). Work ability takes into account situational factors that affect a person's work efficiency (e.g., type of work), as well as his/her actual age. Work ability can be augmented as the result of interventions, which allow the work environment to be adapted to better fit a person's capability.

General Life Stress and Work-Related Stress

The analysis of accident statistics has suggested that people may experience temporary 'spells' when they have increased vulnerability to accidents. These 'spells' would suggest a temporary cause, rather than an enduring predisposition for accident involvement. One possible explanation would relate to the experience of significant periods of stress, which can reduce efficiency and increase accident risk (Clarke, 2011). Major life events, such as divorce or bereavement, cause many people to experience high levels of stress, which increase their vulnerability to accidents and injuries. One research study (Legree et al., 2003) found that life stress in drivers was strongly associated with higher involvement in road accidents. Even minor daily hassles and work-related stress are related to traffic accidents, suggesting a spillover effect of stress experienced outside the driving environment on driving behavior (Rowden et al., 2011). Daily hassles were found to increase the likelihood of both errors (such as cognitive failure) and violations (rule-breaking, such as speeding and tailgating). Stress may affect safe driving behavior by acting as a source of distraction (making errors more likely) and also through emotional 'venting' (such as driving too fast).

Clarke (2012) found that work stress had a significant effect on employees' safe behavior and their accident involvement in the workplace. The most negative effect was associated with 'hindrances'; those situational factors that cannot be controlled, such as daily hassles, conflict with supervisors or coworkers, workplace politics, and concerns about job security. These stressors were found to significantly increase the occurrence of injuries at work, by reducing participation in safety activities (such as reporting faults or attending safety training) and compliance with safety rules (such as following safety procedures or wearing protective clothing). Further research has shown that stress at work can have a spillover effect to other activities, such as driving, leading to more traffic accidents (Rowden et al., 2011).

Risk-Taking Behavior

Risk-taking is usually associated with reckless behavior often in violation of safety rules (such as speeding or drunken driving) but can also relate to high-risk leisure activities such as skiing, skydiving, or mountaineering (which involve known hazards). A review of research looking at the effect of risk-taking on the likelihood of sustaining an injury found that both risky behavior and a sensation-seeking personality led to increased injuries (Turner et al., 2004). The personality trait of sensation-seeking (an aspect of extraversion) predisposes people to seek thrills through taking risks and deliberate rule-breaking (violations). This personality trait has been associated with increased crash risk. Although Turner et al. (2004) supported these findings, they also found that for dangerous sports (such as skiing) the personality trait of sensation-seeking was protective of injury, meaning that sensation-seekers were less likely to be injured. This protective effect was found as those who scored highly on sensation-seeking were also the most skilled skiers.

Risk assessment and evaluation are important psychological processes associated with accidents and injuries. Research conducted on traffic accidents has found that the ability to recognize and assess the risk associated with environmental hazards, such as road conditions, visibility, and weather, is critical to safe driving. These abilities become more sophisticated with driving experience. Risk-taking may occur when hazards have been correctly assessed, but no action is taken to mitigate the risk. For example, the driver correctly identifies that thick fog represents a significant hazard, but fails to reduce speed and increase distance to the next vehicle accordingly. In addition to cognitive processes, risk-taking involves emotional processes as well. For example, Legree et al. (2003) found that drivers who were less competent at managing their own emotions were more likely to be involved in traffic accidents during periods of stress (caused by major life events, such as divorce). Drivers with better emotional awareness were able to adjust their driving behavior to reduce crash risk when they felt emotional and under stress.

Overall, risk-taking behavior increases accident and injury risk by increasing cognitive failure and violations. However, risk-taking involves not only cognitive processes, but also emotional processes, such as thrill-seeking, and emotion awareness and management.

Accident and Injury Prevention

There are a number of different approaches to the prevention of unintentional injuries. These range from education (providing information and training) to product or environmental modifications, behavior modification, and enforcement (regulations and legislation). While changes to regulation and legislation can be effective in changing behavior, public education campaigns

are often needed to encourage behavioral change (e.g., UK legislation to enforce the wearing of car seat belts introduced in 1983 was preceded by the 'clunk click' media campaign during the 1970s). Other approaches focus on interventions designed to reduce or eliminate hazards (environmental modification) or change people's behavior (behavior modification).

Health and Safety Education

Health and safety education is a frequently used method for targeting injury prevention; it focuses on educating people about the hazards in a particular environment (e.g., educating pedestrians about road safety) and providing them with strategies to better cope with those hazards (e.g., learning how best to check for oncoming traffic before crossing the road). A review of the effectiveness of education-based interventions to reduce pedestrian injuries in children (Duperrex et al., 2002) found significant effects on changing road safety behavior. However, a lack of data meant they were not able to establish if these behavioral changes led to reduced injuries. Further research has examined educational interventions aimed at the road-crossing behavior of elderly pedestrians (Dommes and Cavallo, 2012). Although there were short-term effects, in the longer term elderly pedestrians still made poor judgments regarding the speed of approaching vehicles, which affected their decision to cross. Where it is not possible to change such behavior, environmental modifications (such as the provision of pedestrian-operated crossings) may be more appropriate.

Environmental Modifications to Achieve Safety

Environmental modification focuses on the elimination or reduction of hazards in order to prevent injuries. This approach is used most frequently in interventions designed to improve safety in the home. It is often targeted at young children (<5 years) and the elderly (>75 years) who are most at risk of injury in the home environment (e.g., falls, burns/scalds, and poisoning). Most interventions are multifactorial, recognizing the role of multiple factors in accidents, and so often involve an element of education and advice, as well as changes to the domestic environment (e.g., installation of stair gates, locks on cupboards, better lighting in hallways and on stairs, etc.).

A recent review examined the effectiveness of home safety interventions (which provided one-to-one face-to-face education in the home and free/discounted safety equipment) and found they were successful in changing safety practices of mothers with young children (Kendrick et al., 2013). Although safety practices may be improved using interventions, the evidence concerning injuries remains inconsistent (Kendrick et al., 2013; Turner et al., 2011). Turner et al. (2011) were not able to find evidence that home safety interventions are effective in reducing injuries in children and the elderly. However, this was due to the lack of systematic evaluation studies, rather than an indication that these interventions do not work.

Behavioral Modification

Behavior modification focuses on specific unsafe behaviors, such as not wearing safety equipment at work, or failing to put on a car seat belt. The objective of behavior modification is to change the way that people behave, rather than to change environmental hazards. It is often used as a way of further reducing the risk of injury, where hazards have already been reduced to the lowest possible level, or ensuring people use safety equipment designed to protect them from injury (e.g., in the workplace). Behavior modification works by changing either (1) the triggers or prompts, or (2) the consequences of behavior. For example, employees may not wear safety gloves because they are uncomfortable and reduce their productivity; by introducing a financial reward for wearing safety gloves, behavior will be changed. Once a new behavior is learned and becomes habitual, the intervention can be withdrawn, but the new behavior will continue (e.g., even though they no longer receive a financial reward, employees continue wearing safety gloves as they put them on as a habit when starting work, they have got used to working with them on and everybody else is wearing them). The behavioral change approach is not suitable in all circumstances, as it tends to work best for very specific behaviors that are easily observed and monitored.

Behavioral science models are extremely useful for guidance in the design of interventions. There is evidence that interventions based on behavioral theory are much more successful in changing behavior, than those that are not. Behavioral modifications are often very effective in changing behavior in the short term, but behavior can drift back unless attitudes to safety and behavioral norms change in the longer term. The most successful changes are achieved when there is not only behavioral change, but also a change in work climate and workers' safety attitudes (Cooper et al., 1994). This means that one person's new behavior is supported by everyone around him or her (e.g., it is easier to keep wearing safety gloves when all your workmates wear them too and believe that they are essential for safety).

Future of Injury and Accident Research and Prevention

The overview of different approaches to injury prevention has shown that each type of intervention can be effective in some circumstances, but there is not one best way to prevent accidents and injuries. In terms of the future, there needs to be more research focused on identifying the right place for the right intervention. There also needs to be more rigorous evaluation of the effectiveness of interventions and further large-scale studies where data are currently lacking; for example, home safety interventions (Turner et al., 2011; Kendrick et al., 2013), safety education (Duperrex et al., 2002), and the use of combined approaches (Dommes and Cavallo, 2012).

Another area for further investigation is the use of technology to improve safety. For example, as many accidents are due to distraction and cognitive failure, automated systems have been designed to reduce the risk of accidents. In-car technologies, such as adaptive cruise control (ACC), are increasingly being introduced into vehicles to reduce driving errors. A research study (Ward, 2000) found that the use of ACC led to improved driving performance by reducing errors, such as following too close. However, it also had the effect of reducing situation awareness (e.g., lane positioning, yielding to

other traffic, slower response times to unexpected events). This research demonstrates that in-car technologies can have unintended consequences, such as diverting attention to the automated system away from controlling the vehicle and monitoring for external hazards. If automated systems are to be successful in reducing accidents, they must be designed around the person, taking psychological effects into account.

See also: Adolescent Health and Health Behaviors; Environmental Health and Safety: Social Aspects; Health Education and Health Promotion; Occupational Health; Personality and Health; Social Exclusion, Social Deprivation and Health; Urban Life and Health.

Bibliography

Brown, I.D., 1990. Drivers' margins of safety considered as a focus for research on error. Ergonomics 33, 1307–1314.

Clarke, S., 2011. Accident proneness: back in vogue? In: Burke, R.J., Clarke, S., Cooper, C.L. (Eds.), Occupational Health and Safety. Gower, Farnham, Surrey, pp. 95–117.

Clarke, S., 2012. The effect of challenge and hindrance stressors on safety behavior and safety outcomes: a meta-analysis. Journal of Occupational Health Psychology 17, 387–397.

Clarke, S., Robertson, I.T., 2005. A meta-analytic review of the big five personality factors and accident involvement in occupational and non-occupational settings. Journal of Occupational and Organizational Psychology 78, 355–376.

Cooper, M.D., Phillips, R.A., Sutherland, V.J., Makin, P.J., 1994. Reducing accidents using goal setting and feedback: a field study. Journal of Occupational and Organizational Psychology 67, 219–240.

Costa, G., Sartori, S., 2007. Ageing, working hours and workability. Ergonomics 50, 1–17.

Dommes, A., Cavallo, V., 2012. Can simulator-based training improve street-crossing safety for elderly pedestrians? Transportation research part F. Traffic Psychology and Behavior 15, 206–218.

Duperrex, O., Bunn, F., Roberts, I., 2002. Safety education of pedestrians for injury prevention: a systematic review of randomised controlled trials. British Medical Journal 324, 1129–1133.

Haddon, W., 1980. Advances in the epidemiology of injuries as a basis for public policy. Public Health Reports 95, 411–421.

Heinrich, H., 1932. Industrial Accident Prevention, first ed. McGraw Hill, New York.

Kendrick, D., Young, B., et al., 2013. Home safety education and provision of safety equipment for injury prevention. Evidence-Based Child Health: A Cochrane Review Journal 8, 761–939.

Khanzode, V.V., Maiti, J., Ray, P.K., 2012. Occupational injury and accident research: a comprehensive review. Safety Science 50, 1355–1367.

Legree, P.J., Heffner, T.S., Psotka, J., Medsker, G.J., Martin, D.E., 2003. Traffic crash involvement: experiential driving knowledge and stressful contextual antecedents. Journal of Applied Psychology 88, 15–26.

McKnight, A.J., McKnight, A.S., 2003. Young novice drivers: careless or clueless? Accident Analysis and Prevention 35, 921–923.

NCIPC, 2010. Web-Based Injury Statistics Query and Reporting System (WISQARS). http://www.cdc.gov/injury/wisqars.

Rowden, P., Matthews, G., Watson, B., Biggs, H., 2011. The relative impact of work-related stress, life stress and driving environment stress on driving outcomes. Accident Analysis and Prevention 43, 1332–1340.

Smolensky, M., Di Milia, L., Ohyayon, M., Philip, P., 2011. Demographic factors and fatigue: chronic medical conditions and sleep disorders. Accident Analysis and Prevention 43, 516–532.

Turner, C., McClure, R., Pirozzo, S., 2004. Injury and risk-taking behavior: a systematic review. Accident Analysis and Prevention 36, 93–101.

Turner, S., Arthur, G., Lyons, R.A., Weightman, A.L., Mann, M.K., Jones, S.J., John, A., Lannon, S., 2011. Modification of the home environment for the reduction of injuries. Cochrane Database of Systematic Reviews (2). http://www.thecochranelibrary.com.

Visser, E., Pijl, Y.J., Stolk, R.P., Neeleman, J., Rosmalen, J.G.M., 2007. Accident proneness, does it exist? A review and meta-analysis. Accident Analysis and Prevention 39, 556–564.

Wallace, J.C., Vodanovich, S.J., 2003. Can accidents and industrial mishaps be predicted? Further investigation into the relationship between cognitive failure and reports of accidents. Journal of Business and Psychology 17, 503–514.

Ward, N.J., 2000. Automation of task processed: an example of intelligent transportation systems. Human Factors and Ergonomics in Manufacturing 10, 395–408.

af Wåhlberg, A., Dorn, L., 2009. Bus driver accident record: the return of accident proneness. Theoretical Issues in Ergonomics Science 10, 77–91.

World Health Organization, 2008. World Report on Child Injury Prevention. http://www.who.int/violence_injury_prevention/child/injury/world_report/en/index.html.

Injustice: Legal Aspects

Austin Sarat, Jurisprudence, and Social Thought, Amherst College, Amherst, MA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7521–7526, © 2001, Elsevier Ltd.

Abstract

This article explores the meaning of injustice by contrasting it with misfortune. It then explores varying ways in which people respond to injustice. It argues that the fact that most people respond to injustices by doing nothing provides an important factor to be considered in the discussion of the so-called 'litigation explosion.' Yet commentators from Plato to Derrida have called law to account in the name of justice, scolding and chiding it for the injustices which it does, promotes, or tolerates. They also recognize that all too often law, in this culture and elsewhere, is a tool of injustice. For every step taken toward realizing the good, an equal, if not greater, number have been taken in the name of evil. These commentators have asked that law provide a language for responding to injustice and demanded that it promotes, insofar as possible, the attainment of a just society. This article examines various responses to this demand and points to the need for contextual and particularlistic examination of both individual responses to injustice and of law's proper role in the promotion of a just society.

What is injustice? Is it simply the absence of justice, the opposite of a utopian achievement? Or does it have a life of its own? Is it possible to disentangle the experience of injustice from the dissatisfactions and disappointments that people experience everyday? How should an understanding of injustice affect our thinking about law? These are the questions to which any student of injustice must ultimately address himself or herself.

The best recent treatment of the subject of injustice and its complex relations to justice and law is found in Judith Shklar's *The Faces of Injustice* (1990). Shklar (1990, p. 19) tries to rehabilitate injustice as a subject of study, arguing that it "should not be treated intellectually as a hasty preliminary to the analysis of justice." According to Shklar (1990, p. 19; also Cahn, 1949; Folger, 1984), "The real realm of injustice is not in an amoral and prelegal state of nature. It does not appear only on those rare occasions when a political order wholly collapses. It does not stand outside the gate of even the best known states. Most injustices occur continuously within the framework of an established polity with an operative system of law, in normal times."

Shklar recognizes that injustice exists at one end of a continuum of human experience and is often confused with other undesirable conditions. To avoid such confusion, the study of injustice must be differentiated from what she calls misfortune. "When is a disaster a misfortune and when is it an injustice?," Shklar asks. For her the answer is to be found in the presence or absence of a responsible human agent. "If the dreadful event is caused by the external forces of nature, it is a misfortune and we must resign ourselves to our suffering. Should, however, some ill-intentioned agent, human or supernatural, have brought it about, then it is an injustice and we may express indignation and outrage" (Shklar, 1990, p. 1).

This analysis treats injustice as a matter of malevolent human action, rather than as a matter of structure or of institutional design. It embodies what Gordon (1996, p. 36) calls the perspective of 'narrow agency,' in which injustice is thought of as a wrong 'done by specific perpetrators to specific victims.' But Gordon suggests that the danger of this perspective is that it will lead us to ignore other kinds of injustices. He reminds us that injustices can be done by collective perpetrators to collective victims, such as the Nazi atrocities against the Jews. This kind of injustice involves what he calls 'broad agency.' In addition, injustice can be attributed to bad structures in addition to bad agents. An interest in this kind of injustice 'takes the form of explanations rather than a search for villains and attribution of blame.'

Responses to Injustice

Whether defined in terms of narrow agency, broad agency, or structure, some believe that in the United States people are too ready and eager to use the language of injustice and to turn to law to redress conditions which in the past would have been regarded as the uncontrollable workings of fate. Typical is Lieberman's (1981, p. xi) contention that ours "is a law-drenched age… Though litigation has not routed out all other forms of fight, it is gaining public favor as the legitimate and most effective means of seeking and winning one's just deserts. So widespread is the impulse to sue that litigation has become the nation's secular religion." In this view, our legal culture is today governed by two principles (Friedman, 1985, p. 43). "The first… is the citizen's expectation of fair treatment, everywhere and in every circumstance … The second… is the general expectation that somebody will pay for any and all calamities that happen to a person …." Together these two principles create a culture of 'total justice' (Friedman, 1985) and what some believe to be a 'litigious society' (Lieberman, 1981; Huber, 1988).

In an article I co-authored in 1980, my co-authors and I argued that responses to injurious events are more complicated than arguments about 'total justice' or the 'litigious society' admit. These responses, we noted, could be understood as occurring in three stages (Felstiner et al., 1980–81). The first, defining a particular experience as injurious, we called 'naming.' The next step in the emergence of a claim of injustice is the transformation of a perceived injurious experience into a grievance. This occurs when a person attributes an

injury to the fault of another individual or social entity' (Felstiner et al., 1980–81, p. 635). It involves 'blaming.' The third step occurs "when someone with a grievance voices it to the person or entity believed to be responsible and asks for some remedy" (Felstiner et al., 1980–81, p. 635). This step we called 'claiming.'

It was our contention that "only a small fraction of injurious experiences ever mature into disputes. Furthermore, we know that most of the attrition occurs in the early stages: experiences are not perceived as injurious; perceptions do not ripen into grievances; grievances are voiced to intimates but not to the person deemed responsible" (Felstiner et al., 1980–81, p. 636). Paying attention to the process through which responses to injustice are formulated and to the high rate of attrition in, or resistance to, the making of claims was, we contended, a way of putting the burgeoning argument about the litigation explosion in context. We argued that the processes through which claims emerge, or through which people decide to 'lump it,' are "subjective, unstable, reactive, complicated, and incomplete" (Felstiner et al., 1980–81, p. 637). Most of what occurs in this domain is cultural, not legal, occurring as it does in the everyday lives of ordinary citizens and in their struggles to make sense of who they are and who they want to be in a world of risk, danger, and injury (Sarat, 2000) (*see* Legal Culture and Legal Consciousness).

Since then, some scholars have studied the way in which people actually react to the injurious events of their lives (Merry, 1990; Yngvesson, 1993; Greenhouse, 1986; Ewick and Silbey, 1998). That scholarship characterizes litigation as a process of contesting meanings and adjudication as the choice and imposition of one structure of meaning over another (Merry, 1990). Contests over meaning in courts or communities thus become occasions for observing the play of power. Meanings which seem natural, or taken for granted, are described as hegemonic, but because the construction of the meaning of injurious events is, in fact, typically contested, scholars show the many ways in which different people respond to injustice. They have found that the tendency to label injurious events in the language of injustice is contingent and complex, varying from one community to the next, from one kind of injury to another.

In this process, law plays a key role. As Merry sees it, the relationship between injustice and law is not one in which law passively responds to expectations for 'total justice.' Law works, she says (Merry, 1990, pp. 8–9), "not just by the imposition of rules and punishments but also by the capacity to construct authoritative images of social relationships and actions, images that are symbolically powerful. Law provides a set of categories and frameworks through which the world is interpreted. Legal words and practices are cultural constructs which carry powerful meanings not just to those trained in law … but to the ordinary person as well." Merry's research shows that responses to injustice involve "a complex repertoire of meanings and categories understood differently by people depending on their experience with and knowledge of the law" (Merry, 1990, p. 5). This work connects the study of injustice to the world of law and is continuous with a long history of scholarship about the connections between law and justice and about the ways law is implicated in injustice.

Injustice and Law

Commentators from Plato to Derrida have called law to account in the name of justice, scolding and chiding it for the injustices which it does, promotes, or tolerates. They also recognize that all too often law, in this culture and elsewhere, is a tool of injustice (see, for example, Cover, 1975; Irons, 1983). For every step taken toward realizing the good, an equal, if not greater, number have been taken in the name of evil. These commentators have asked that law provide a language for responding to injustice and demanded that it promote, insofar as possible, the attainment of a just society.

They have done so, however, in relentlessly abstract and general language, as if the demands of justice and the problem of injustice could only be apprehended accurately outside of history and context, and as if only philosophers were fit to engage in conversation about these subjects. The justice which is spoken about is, as a result, elusive, if not illusory, and disconnected from the embodied practices of law of the kind that Merry emphasizes. As one writer noted, "there is no such thing as justice 'in general,' one can meaningfully discuss only fairly concrete injustices and procedures for their mitigation with existing social machinery or possible ways of improving the overall social organization" (Knight, 1963, p. 1).

The turn to injustice is generally associated with an interest in context and particularity, in which attention is given to individuals and institutions and their contingent histories. In one sense, then, when scholars talk about injustice and try to connect it to law they are often reacting against the vexing difficulty that attends to the effort to examine the connection of law and justice.

In earlier days, however, speaking about law and justice was not so vexing or difficult. Justice (*jus*, meaning 'law') was a legal term, pure and simple. At the outset, then, justice was defined and constituted by laws that were 'given' and held to be unchanging and unchangeable. This ineluctable link between justice and law had the virtue of making the boundaries of justice more or less clear, but it had the considerable vice of labeling even heinous, iniquitous laws just. Attention to the problem of injustice could do no critical/reconstructive work in legal scholarship because it was impossible to think of justice as external to law.

Hobbes (1958) to the contrary notwithstanding, most natural law thinkers have resisted this result by insisting that unjust laws are not law (see, for example, Aquinas, 1988; George, 1992; Finnis, 1890; Hittinger, 1987). Doing so meant the end of any easy identification of positive or human law with 'real' or binding law. The alternative, embraced by perhaps a majority of those who do legal philosophy (Hart, 1961; for an important contrast, see Fuller, 1964), is to cut justice and law free from one another, to insist that justice is more than mere conformity to law and to acknowledge that even unjust laws might nonetheless be law. This has freed students of injustice to direct their attention to the practices of law itself, scrutinizing them for their contributions to injustice in society.

More recently, the distance between law and justice, and the potential for injustice in the name of the law, have been recognized in postmodern theorizing about ethics (Cornell, 1988, 1992). Thus, as Douzinas and Warrington (1994, p. 23)

(see also Keenan, 1990) argue: "Justice has the characteristic of a promissory statement. A promise states now something to be performed in the future. Being just always lies in the future, it is a promise made to the future, a pledge to look into the event and the uniqueness of each situation …. This promise, like all promises, does not have a present time, a time when you can say 'there it is, justice is this or that.' Suspended between the law and the good …, justice is always still to come or always already performed."

Severance of the definitional tie between justice and law has left both notions free (if also bound) to acquire new identities (Balkin, 1994). It has opened the way for an examination of unjust actions or conditions and for an inquiry into the ways law is implicated in them. However, this definitional separation has by no means put an end to the commonplace conviction that the principal home of justice is law and the special concern of law, perhaps its principal business, is justice. As Clarence Morris (1963, p. 170) notes, "Though there can be law without justice, justice is realized only through good law." In fact, law and legal theory continue to be shaped by concerns about justice and injustice, just as understandings of these notions are shaped by an awareness of law and the concerns of legal theory (Lucas, 1980).

From Law to Politics in the Study of Injustice

By what might fairly be regarded as an accident of intellectual history, concerns about injustice as a motive for studying and reforming law have been somewhat muted in some contemporary discourse which seeks to expand interest in injustice beyond the sphere of law. They have been swamped by a preoccupation with distributive justice, a subject whose contemporary pre-eminence can unquestionably be attributed to John Rawls' *A Theory of Justice* (1971). 'Justice,' Rawls (1971, p. 3) contends, "is the first virtue of social institutions, as truth is of systems of thought." But the justice he has in mind is almost exclusively distributive justice, not justice more broadly construed; injustice of a distributive type is not peculiarly the concern of law at all.

Similarly, the philosopher Agnes Heller (1987, p. 54) introduces the notion of a broad, though incomplete, "ethico-political concept of justice." In her view, justice and injustice are not simply about principles of distribution; they concern the perspectives, principles and procedures for evaluating institutional norms and rules. Developing themes in Habermas's (1987) communicative ethics, Heller (1987, p. 54) proposes that justice is "primarily the virtue of citizenship, of persons deliberating about problems and issues that confront them collectively in their institutions and actions, under conditions without domination or oppression, with reciprocity and mutual tolerance of difference" (Young, 1990, p. 33). Domination, oppression, and intolerance are thus preconditions for injustice.

Iris Young (1990, pp. 34, 37) expands on even this notion, concluding that "the concept of justice coincides with the concept of the political," and while it is "not identical with the concrete realization of these values in individual lives," it is attentive principally to "the degree to which a society contains and supports institutional conditions necessary for the realization" of the values of equal worth as these are promoted or confined in a society's basic institutional arrangements. The study of injustice then moves from law to politics. On this account the subject is as inclusive as the political order itself.

On the Continuing Significance of Procedure

Despite these broad and encompassing views of justice and injustice, the fact remains that injustice is generally thought to be involved in matters of procedure, punishment, and recompense (Aristotle, 1985; Lucas, 1980). Thus, while law in its legislative moments might share with other social institutions certain distributional objectives, it is generally thought that law has a separate and distinctive commitment to doing justice, located most conspicuously, though not exclusively, in matters of form and process. This is, of course, symbolized in the universal icon of justice (and invariably associated with law): the statue of Justicia, the blindfolded bearer of sword and balance (*see* Rule of Law (and Rechtsstaat); Due Process in Law).

Despite certain ambiguities, Justicia presumably embodies law's central link to justice, impartiality (Curtis and Resnik, 1987). The instrument of Justicia is a balance, not a scale, so it is only a comparative rather than an absolute measurement that is involved. She is blindfolded to prevent her decision from being inappropriately influenced or determined by forces of fear, bias, or favoritism. Moreover, she is blindfolded – not blind – suggesting a self-willed refusal of vision. If we are to avoid injustice, partial justice, or biased justice, what is required is a denial of sight though not of seeing, a regulation of information though not of knowing, a restriction on what is permissibly attended to though not a deficit of attention (Black, 1989).

According to Curtis and Resnik (1987, p. 1727), "procedure is the blindfold of Justice." The proposition that procedure is itself an aspect of justice, and that procedural irregularities are in themselves instances of injustice, takes us some distance toward the further result that law and justice are ineluctably linked. Thus, to the extent that one supposes that the Rule of Law and the principle of legality are constituted largely by a commitment to regularized, self-limiting procedures, it is but a short step to the conclusion that procedural justice is inextricably tied to law itself.

Along some such path, it seems possible to argue further that procedures themselves must meet additional normative requirements, sounding in fairness and desert, and other dimensions of justice. As Hart (1961, Chap. VIII) has observed, a rule that is regularly, uniformly applied to prohibit certain minorities from using the town park might, despite its unfailing enforcement, be viciously unjust. Legal justice, then, involves more than general observance of the rules that regulate exercises of legal authority, the rules themselves must be fair, eschewing wholly arbitrary distinctions between and among the persons to whom they apply or on whom they finally have effect. The rules must provide suitable forms of what, in the United States, are known as requirements of due process and equal protection.

Acknowledging the relationship of procedure and justice has spawned a large and important empirical literature which

explores what leads people to believe that law is just or unjust (Lind and Tyler, 1988; Tyler, 1989). This literature suggests that procedure is more important than substance, that the way one is treated is more important than the results one gets in determining whether people believe that they have been treated justly or unjustly in the legal system. As Lind (1998, p. 184) explains. "If authorities seem to be making decisions in fair ways, people assume that they can obey orders and follow rules without worrying too much about exploitation. If, on the other hand, authorities seem to be reacting unfairly, exploitation becomes a real possibility and obedience tends to be viewed with more suspicion …. Outcome-oriented theories of justice judgements err because they fail to recognize that people are generally quite ready to accept outcomes that are less than they had hoped for."

Can Law Overcome Injustice?

Returning to the questions with which we started, we have seen that injustice is not simply an absence of justice, that it is a fact and an experience different from, though related to, misfortune, and that it may involve the relations of one person to another, of groups of people, and of the attributes of legal, political, or social structure. Though some would focus the attention of persons interested in understanding injustice fairly broadly, law seems to be at the center of any proper understanding of injustice.

Yet having envisaged ways of connecting the study of injustice and of law, it must be granted that not everyone regards the imagined association as an unalloyed good. Hayek (1982), among others, is persuaded that social justice is a mirage, a misguided affection the pursuit of which not only threatens liberty but, eventually, law itself. A somewhat more moderate concern is that the demands of distributive justice and the Rule of Law are inevitably at odds with one another, so that the effort to cure distributive injustices comes at a high price to law itself (Unger, 1977; Barnett, 1988). A third apprehension is that the meaning of justice is so ambiguous, expansive, and elusive that law invariably will be seen as falling short, as failing to meet justice's demands, as always, then, implicated in injustice. As Balkin (1994, p. 16) argues, laws apportion responsibility, create rights and duties, and provide rules for conduct and social ordering. Law is always, to some extent and to some degree, unjust. At the same time, our notion of justice can only be articulated and enforced through human laws and conventions. We may have a notion of justice that always escapes láw and convention, but the only tools we have to express and enforce our idea are human laws and human conventions. Our conception of the just relies for its articulation and enforcement on the imperfect laws, conventions, and cultural norms from which it must always be distinguished.

Some worry that law's legitimate authority, such as it is, will be dangerously (and unjustly) eroded (Tyler, 1990). Alternatively, but with consequences no less grave, the impossibility of satisfying justice's demands may eventually be seen as the fault of justice, not of law.

Still other qualms dampen enthusiasm about the possibility of connecting law to the study of justice and injustice. Some believe that the embrace of legal justice is, at best, a rhetorically appealing device for quieting alarm about the excesses and inequities social and economic arrangements inevitably arouse (Hay, 1975). So conceived, legal justice is largely a symbolic response to the incorrigible defects and injustices of liberal capitalist societies, not a positive virtue which these arrangements can promote or make possible (Marx, 1978).

The fear that legal justice can be put to work in the name of legitimating social injustice, for example by 'justifying' outcomes that seem patently, outrageously disproportionate has been expressed in a different voice, and on different grounds, by Shklar (1990). She argues that the focus on justice inhibits the development of a community that both recognizes the culturally constructed quality of boundaries between the ideas of 'injustice' and 'misfortune,' and objects to passive injustice as vehemently as it abhors the active violation of rights. But Shklar's dissatisfaction with a narrow interest in legal justice is pressed even further in her contention that contemporary theory tends almost exclusively to regard injustice as "simply the absence of justice" (Shklar, 1990, p. 15; Woozley, 1973).

On this view, "once we know what is just, we know all we need to know," Shklar (1990, p. 15) writes. Or, as Knight (1963, p. 15) observes, "One misses a great deal by looking only at justice. The sense of injustice, the difficulties of identifying victims of injustice, and the many ways in which we learn to live with each other's injustices tend to be ignored, as is the relation of private injustice to the public order." Scholars such as Shklar and Knight contend that we not only lack a full and coherent account of justice; even if we had one, it would not yield an equally full account of injustice, nor would it confront adequately our calloused complacency regarding vast, yet eliminable, misfortunes.

These observations call to mind a variety of uncertainties and ambivalences regarding the relationship of law and justice, but they fall far short of embracing the extreme position that justice be jettisoned from legal discourse. Rather, they remind us of the vastness of the subject of injustice, of the difficulty of constructing a single account capable of holding together its many strands, and of the space that exists to theorize anew about justice and injustice in law and legal theory.

See also: Due Process in Law; Justice and Law; Rights: Legal Aspects; Rule of Law (and Rechtsstaat); Social Justice and Criminal Justice.

Bibliography

Aquinas, T., 1988. Summa Theologica. In: Baumgarth, W., Kegan, R. (Eds.). Hackett, Indianapolis, IN.

Aristotle, 1985. Nicomachean Ethics. In: Irwin, T. (Ed.). Hackett, Indianapolis, IN (trans.).

Balkin, J., 1994. Being just with deconstruction. Social and Legal Studies: An International Journal 3, 393–405.

Barnett, R., 1988. Can justice and the rule of law be reconciled foreward? Harvard. Journal of Law and Public Policy 11, 597–624.

Black, D., 1989. Sociological Justice. Oxford University Press, New York.

Cahn, E., 1949. The Sense of Injustice: An Anthropocentric View of Law. New York University Press, New York.

Cornell, D., 1988. Post-structuralism, the ethical relation, and the law. Cardozo Law Review 9, 1587–1620.
Cornell, D., 1992. The Philosophy of the Limit. Routledge, New York.
Cover, R.M., 1975. Justice Accused: Antislavery and the Judicial Process. Yale University Press, New Haven, CT.
Curtis, D.E., Resnik, J., 1987. Images of justice. Yale Law Journal 96, 1727–1772.
Derrida, J., 1990. Force of law: 'The mystical foundation of authority'. Cardozo Law Review 11, 919–1046.
Douzinas, C., Warrington, R., 1994. The face of justice: A jurisprudence of alterity. Social and Legal Studies: A International Journal 3, 405–426.
Ewick, P., Silbey, S.S., 1998. The Common Place of Law: Stories from Everyday Life. University of Chicago Press, Chicago.
Finnis, J., 1890. Natural Law and Natural Rights. Clarendon Press, Oxford, UK.
Folger, R. (Ed.), 1984. The Sense of Injustice: Social Psychological *Perspectives*. Plenum Press, New York.
Felstiner, W.L., Abel, R.L., Sarat, A., 1980–81. The emergence and transformation of disputes: Naming, blaming, and claiming. Law and Society Review 15, 631–654.
Friedman, L., 1985. Total Justice. Russell Sage Foundation, New York.
Fuller, L.L., 1964. The Morality of Law. Yale University Press, New Haven, CT.
George, R.P., 1992. Natural Law Theory: Contemporary Essays. Clarendon Press, Oxford, UK.
Gordon, R., 1996. Undoing historical injustice. In: Sarat, A., Kearns, T.R. (Eds.), Justice and Injustice in Law and Legal Theory. University of Michigan Press, Ann Arbor, MI, pp. 35–76.
Greenhouse, C.J., 1986. Praying for Justice: Faith, Order, and Community in an American Town. Cornell University Press, Ithaca, NY.
Gruber, G.M.A., 1992. Plato's Republic (trans.). Rev. Reeve C D C. Hackett, Indianapolis, IN.
Habermas, J., 1987. The Theory of Communicative Competence, Vol. 2: Life World and System. Beacon Press, Boston.
Hart, H.L.A., 1961. The Concept of Law. Clarendon Press, Oxford, UK.
Hay, D., 1975. Property, authority and the criminal law. In: Hay, D., Linebaugh, P., Thompson, E.P. (Eds.), Albion's Fatal Tree. Lane, London, pp. 17–63.
Hayek, F.A., 1982. Law, Legislation and Liberty, Vol. II: The Mirage of Social Justice. Routledge, London.
Heller, A., 1987. Beyond Justice. Basic Books, New York.
Hittinger, R., 1987. A Critique of the New Natural Law Theory. University of Notre Dame Press, Notre Dame, IN.
Hobbes, T., 1958. Leviathan. Bobbs-Merrill, Indianapolis, IN.
Huber, P.W., 1988. Liability: The Legal Revolution and Its Consequences. Basic Books, New York.
Irons, P., 1983. Justice at War: The Story of the Japanese American Internment Cases. Oxford University Press, New York.
Keenan, T., 1990. Deconstruction and the impossibility of justice. Cardozo Law Review 11, 1675–1686.
Knight, F.H., 1963. On the meaning of justice. In: Friedrich, C.J., Chapman, J.W. (Eds.), Justice. Atherton Press, New York, pp. 1–23.
Lieberman, J.K., 1981. The Litigious Society. Basic Books, New York.
Lind, E.A., 1998. Procedural justice, disputing, and reactions to legal authorities. In: Sarat, A., Constable, M., Engel, D., Hans, V., Lawrence, S. (Eds.), Everyday Practices and Trouble Cases. Northwestern University Press, Evanston, IL, pp. 177–198.
Lind, E.A., Tyler, T.R., 1988. The Social Psychology of Procedural Justice. Plenum Press, New York.
Lucas, J.R., 1980. On Justice. Clarendon Press, Oxford, UK.
Marx, K., 1978. Critique of the Gotha program. In: Tucker, R. (Ed.), The Marx-Engels Reader, second ed. Norton, New York, pp. 525–542.
Merry, S., 1990. Getting Justice and Getting Even: Legal Consciousness Among Working-Class Americans. University of Chicago Press, Chicago.
Morris, C., 1963. Law, justice, and the public's aspirations. In: Friedrich, C.J., Chapman, J.W. (Eds.), Justice. Atherton Press, New York, pp. 170–190.
Rawls, J., 1971. A Theory of Justice. Belknap Press of Harvard University Press, Cambridge, MA.
Sarat, A., 2000. Exploring the hidden domains of civil justice: 'Naming, Blaming and Claiming' in popular culture. DePaul Law Review 50, 425–452.
Shklar, J.N., 1990. The Faces of Injustice. Yale University Press, New Haven, CT.
Tyler, T., 1989. The psychology of procedural justice: A test of the group value model. Journal of Personality and Social Psychology 57, 830–838.
Tyler, T., 1990. Why People Obey the Law. Yale University Press, New Haven, CT.
Unger, R., 1977. Law and Modern Society. Free Press, New York.
Woozley, A.D., 1973. Injustice. American Philosophical Quarterly 7, 109–122.
Yngvesson, B., 1993. Virtuous Citizens, Disruptive Subjects: Order and Complaint in a New England Court. Routledge, New York.
Young, I.M., 1990. Justice and the Politics of Difference. Princeton University Press, Princeton, NJ.

Inner City

William J Wilson and James M Quane, Harvard University, Cambridge, MA, USA
Bruce H Rankin, Koç University, Sarõyer, Istanbul, Turkey

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A long history of racial segregation, government policy, demographic trends, and industrial shifts in the labor market has contributed to the growth of disadvantaged and under-resourced inner-city neighborhoods especially in older northern industrial cities. The subsequent social and physical isolation of poor families in joblessness inner-city ghettos has disconnected the poor from job opportunities and the socialization processes associated with regular employment. Economic expansion in the 1990s helped to expand the black middle class but the Great Recession reversed much of these gains. Policies that improve the quality of urban public schools and help grow a workforce that is prepared for a modern economy are discussed.

The Rise of the Modern Inner-City Ghetto

One of the legacies of historic racial and class subjugation in the United States is the growing concentration of poor and minority residents in the nation's impoverished inner-city areas. The demand for low-skill blue-collar workers in the North during the early twentieth century, coupled with the rise in southern blacks' economic insecurity brought about by the declining demand for farm labor due to the increasing mechanization of agriculture, helped fuel the migration of blacks to northern industrial cities. In the North, blacks found themselves increasingly segregated in ghettos that previously housed white ethnic immigrants, who themselves were gradually assimilating into American society.

Since 1934, with the establishment of the Federal Housing Administration (FHA), a program necessitated by the massive mortgage foreclosures during the Great Depression, the US government has sought to enable citizens to become homeowners by underwriting mortgages. In the years following World War II, however, the federal government contributed to the early decay of inner-city neighborhoods by withholding mortgage capital and making it difficult for these areas to retain or attract families who were able to purchase their own homes. But the FHA selectively administered the mortgage program by formalizing a process that excluded certain urban neighborhoods using empirical data that suggested a probable loss of investment in these areas. 'Redlining,' as it came to be known, was assessed largely on racial composition.

Although many neighborhoods with a considerable number of European immigrants were redlined, virtually all black neighborhoods were. Home buyers hoping to purchase a home in a redlined neighborhood were universally denied mortgages, regardless of their financial qualifications. This severely restricted opportunities for building or even maintaining quality housing in the inner city, which in many ways set the stage for the urban blight that many Americans associate with black neighborhoods. This action was clearly motivated by racial bias, and it was not until the 1960s that the FHA discontinued mortgage restrictions based on the racial composition of the neighborhood (Bartelt, 1993; Katz, 1993; Kelley, 1993; Sugrue, 1993).

Federal public housing policy contributed to the gradual growth of segregated inner-city ghettos. The federal public housing program's policies evolved in two stages that represented two distinct styles. The Wagner Housing Act of 1937 initiated the first stage. Concerned that the construction of public housing might depress private rent levels, groups such as the US Building and Loan League and the National Association of Real Estate Boards successfully lobbied Congress to require, by law, that for each new unit of public housing erected, one 'unsafe or unsanitary' unit of public housing must be destroyed.

The early years of the public housing program produced positive results. Initially, the program mainly served intact families temporarily displaced by the Depression or in need of housing after the end of World War II. For many of these families, public housing was the first step on the road toward economic recovery. Their stays in the projects were relatively brief because they were able to accumulate sufficient economic resources to move on to private housing.

The passage of the Housing Act of 1949 marked the beginning of the second policy stage. It instituted and funded the urban renewal program, designed to eradicate urban slums, and therefore was seemingly nonracial. However, the public housing that it created "was now meant to collect the ghetto residents left homeless by the urban renewal bulldozers" (Condon, 1991). A new, lower income ceiling for public housing residency was established by the federal Public Housing Authority, and families with incomes above that ceiling were evicted, thereby restricting access to public housing to only the most economically disadvantaged segments of the population.

This change in federal housing policy coincided with the Second Great Migration of African Americans from the rural South to the cities of the Northeast and Midwest, which lasted 30 years – from 1940 to 1970. This mass movement of African Americans was even larger and more sustained than the First Great Migration, which began at the turn of the twentieth century and ended during the Great Depression, and had a more profound impact on the transformation of the inner city. As the black urban population in the North grew, pressure mounted in white communities to keep blacks out. Suburban

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32178-X

communities, with their restrictive covenants and special zoning laws, refused to permit the construction of public housing, and the federal government acquiesced to opposition to the construction of public housing in the neighborhoods of organized white groups in the city. Thus, units were overwhelmingly concentrated in the overcrowded and deteriorating inner-city ghettos – the poorest and least powerful sections of cities and metropolitan areas. In short, public housing became a federally funded institution that isolated families by race and class, resulting in high concentrations of poor black families in inner-city ghettos (Sampson and Wilson, 1995).

In spite of the tremendous obstacles faced by African Americans, inner-city ghetto neighborhoods in the early twentieth century developed a diversified social and institutional structure, characterized by a mix of poor, working, and middle-class blacks and viable institutions (e.g., businesses, schools, churches, and cultural organizations) that fostered a vibrant social and cultural life. This social organization of earlier ghettos helped to buffer the deleterious effects of marginalization and concentrated poverty (Du Bois, 1899; Drake and Cayton, 1962). Despite their disadvantaged status, well into the 1960s most ghetto-dwellers resided in neighborhoods where a majority of adults were employed.

However, in the last quarter of the twentieth century, one of the most significant changes in these neighborhoods was the out-migration of middle-income blacks (Wilson, 1987). Before the 1970s, African American families faced extremely strong barriers when they considered moving into white neighborhoods. Not only did many experience overt discrimination in the housing market, some were violently attacked. Although even fair housing audits continue to reveal the existence of discrimination in the housing market (Massey and Denton, 1993), fair housing legislation has reduced the strength of these barriers. At the same time, middle-income African Americans have increased their efforts to move from concentrated black poverty areas to more desirable neighborhoods throughout the metropolitan area, including white neighborhoods (Quillian, 1999).

This pattern represents an important change in the formation of neighborhoods. In the earlier years, communities undergoing racial change from white to black tended to experience an increase in population density, as a result of the black migration from the South. Because of the housing demand, particularly in the late stages of the succession from white to black, homes and apartments in these neighborhoods were often subdivided into smaller units (Quillian, 1999). However, 1970 marked the end of the great migration wave of blacks from the South to northern urban areas, and two developments affected the course of population movement to the inner cities after that time. Improvements in transportation made it easier for workers to live outside the central city, and industries gradually shifted to the suburbs due to the increased residential suburbanization of the labor force and the lower cost of production. Because of the suburbanization of employment and improvements in transportation, inner-city manufacturing jobs were no longer a strong factor pulling migrants to central cities (Quillian, 1999). Moreover, the increasing suburbanization of employment generated a growing mismatch between the suburban location of employment and the inner-city residence of minority populations (Wilson, 1987). Inner-city joblessness increased as residents had to contend with both fewer employment opportunities nearby and longer commutes to jobs outside the central city (Wilson, 1996).

So with the decline of industrial employment in the inner city, the influx of southern black migration to northern cities ceased and many poor black neighborhoods, especially those in the Midwest and Northeast, changed from densely packed areas of recently arrived migrants to communities gradually abandoned by the working and middle classes (Wilson, 1987, 1996; Quillian, 1999). With the departure of higher income families, the least upwardly mobile in society – mainly low-income people of color – were left behind in neighborhoods with high concentrations of poverty and deteriorating physical conditions. These neighborhoods offer few jobs and typically lack basic services and amenities, such as banks, grocery stores and other retail establishments, parks, and quality transit (Wilson, 1987, 1996; Fox and Treuhaft, 2006). Typically, these communities also suffer from substandard schools, many with run-down physical plants. Two of the most visible indicators of neighborhood decline are abandoned buildings and vacant lots. According to one recent report, there are 60 000 abandoned and vacant properties in Philadelphia, 40 000 in Detroit, and 26 000 in Baltimore (Fox and Treuhaft, 2006).

Segregation in these depopulated inner-city ghettos also aggravated employment problems by weakening informal employment networks and contributing to the social isolation of families and individuals, thereby reducing their chances of acquiring the human capital skills, including adequate educational training, that facilitate social mobility (Massey and Denton, 1993).

The Changing Labor Market for Unskilled Workers

With the decline in manufacturing employment in many of the nation's central cities, most of the jobs for lower skilled workers are now in retail and service industries (e.g., store cashiers, customer service representatives, fast food servers, and custodial work). Whereas jobs in manufacturing industries typically were unionized, relatively stable, and carried higher wages, those for workers with low to modest levels of education in the retail and service industries tend to provide lower wages, be unstable, and lack the benefits and worker protections – such as workers' health insurance, medical leave, retirement benefits, and paid vacations – typically offered through unionization. This means that workers relegated to low-wage service and retail firms are more likely to experience hardships as they struggle to make ends meet. In addition, the local economy suffers when residents have fewer dollars to spend in their neighborhoods (Wilson, 1996; Fox and Treuhaft, 2006).

Two-thirds of job growth has occurred outside the central city as a result of the suburbanization of employment: manufacturing is now over 70% suburban, and wholesale and retail trade is just under 70% (U.S. Department of Housing and Urban Development, 1999). The suburbs of many central cities, developed originally as bedroom localities for commuters to the central business and manufacturing districts, have become employment centers in themselves. For

example, in Detroit, Philadelphia, and Baltimore, less than 20% of the jobs are now located within three miles of the city center (Fox and Treuhaft, 2006).

Accompanying the rise of suburban and exurban economies has been a change in commuting patterns. Increasingly, workers completely bypass the central city by commuting from one suburb to another. "In the Cleveland region, for example, less than one-third of workers commute to a job in the central city and over half (55 percent) begin and end in the suburbs" (Fox and Treuhaft, 2006: p. 32). Sprawl and economic stagnation reduce inner-city residents' access to meaningful economic opportunities and thereby fuel the economic decline of their neighborhoods. For example, in Cleveland, although entry-level workers are concentrated in inner-city neighborhoods, 80% of the entry-level jobs are located in the suburbs and there is little public transportation between these neighborhoods and jobs (Fox and Treuhaft, 2006).

In addition to the challenges in learning about and reaching jobs, there is persistent racial discrimination in hiring practices, especially for younger and less-experienced minority workers. This racial factor affects black males especially seriously. Today, most of the new jobs for workers with limited education and experience are in the service sector, which includes jobs that tend to be held by women, such as waitstaff, sales clerks, and nurse's aides. Indeed, "employment rates of young black women now exceed those of young black men, even though many of these women must also care for children" (Holzer et al., 2003: p. 15). The shift to service jobs has resulted in a greater demand for workers who can effectively serve and relate to the consumer. Many employers, in an extensive study in Chicago conducted by my colleagues and me, felt that, unlike women and immigrants (who have recently expanded the labor pool for service-sector jobs), inner-city black males lack these qualities. Instead, low-skilled black males are perceived as dangerous or threatening (Wilson, 1996).

In the past, all black men had to demonstrate was a strong back and muscles for heavy lifting and physical labor in a factory, at a construction site, or on an assembly line. They did not have to interact with customers. Today, they have to search for work in the service sector, and employers are less likely to hire them because they have to come into contact with the public. Consequently, inner-city black male job seekers face rising rates of rejection (Wilson, 1996).

Concentrated Poverty, Social Organization, and Culture in the Inner-City Neighborhoods

In contrast to socially organized neighborhoods, many segregated inner-city neighborhoods today have few institutional resources and weak and less cohesive social networks. As a result, residents are less able to enforce norms and control undesirable behavior in the community or to act collectively to preserve the quality of neighborhood life. When the effectiveness of formal and informal social control mechanisms drops below a certain threshold, sanctions against nonnormative behavior become ineffective and opportunities for illegitimate activities proliferate in the community (Sampson and Wilson, 1995).

One of the negative implications of living in such highly segregated neighborhoods, which are characterized by widespread joblessness, high rates of crime, and social dislocation, is the exposure to cultural traits that emerge under conditions of chronic racial and economic subordination, which may undermine efforts toward economic and social mobility. In this context, cultural traits refer to the shared outlooks, modes of behavior, values and norms, attitudes, meaning- and decision making, and behavior that are perpetuated and reinforced in highly restrictive social settings. As such, personal agency in these environments is severely limited by the reinforcement of certain behaviors and habits that are acquired over time under conditions of chronic disadvantage. In addition, the stock of social and human capital that one is likely to acquire in these under-resourced neighborhoods is likely to severely limit the range of agency the poor can deploy to alleviate their circumstances. Consequently, "(T)he limitations caused by this disconnect from capital-enhancing opportunities deny them the ability to think strategically about ways to transcend poverty. Instead, their lives are predominantly influenced by more constraining institutions that infuse the social and economic rationalities under which they operate" (Quane and Wilson, 2012: p. 2982).

In sum, the cultural milieu of the inner city is embedded in the structural features of concentrated poverty, racial segregation, social isolation, and the lack of neighborhood social organization. Cut off from employment opportunities and trapped in neighborhoods with little access to conventional role models, social networks that contain social capital (e.g., informal job networks), and institutional resources that support families, poor inner-city families face severe obstacles in their efforts to accomplish socially desired ends such as labor force participation and family stability.

The Latino Paradox: Concentrated Inner-City Poverty and the Rise of Latino Immigration

The declining white population in urban areas has had significant implications for the urban tax base as the median household income of blacks and of the growing number of Latinos is significantly below the median income of white households. (According to the US Census Bureau, in 2000 the median annual household income for Latinos was about $14 000 less than that of white households.) Thus, because of the declining tax base and the simultaneous loss of federal funds, many municipalities have experienced difficulty in raising sufficient revenue to cover basic services such as street cleaning, garbage collection, and police protection. Nonetheless, several notable positive developments for inner-city urban areas have accompanied the rise of the immigrant population.

Foreign immigration, especially from Mexico, increased by more than 50% between 1990 and 2000, resulting in a rapid growth in immigrant enclaves in large cities (Sampson, 2012). According to a report by the PEW Hispanic Center, the most significant growth in immigration occurred in the mid-1990s, reaching its peak at the end of the decade (Passel et al., 2012). Although annual immigrant flows from Mexico declined sharply after 2005, in the wake of a weakening US job market and heightened border enforcement, the number

of immigrants from Mexico remains high at approximately 12 million (Passel et al., 2012).

It is noteworthy, as Robert Sampson's (2012) impressive research so clearly reveals, that areas of concentrated immigration are among the safest in the city, and the process of immigrant diffusion revitalized and increased the population of a number of deteriorating inner-city neighborhoods. Sampson states: "Among the public, policy makers, and many academics, a common expectation is that the concentration of immigrants and the influx of foreigners drive up disorder and crime because of the assumed propensities of these groups to commit crimes and settle in poor, presumably disorganized communities … Yet increases in immigration are correlated with less crime, and immigrants appear to be less violent than those born in America, particularly when they live in neighborhoods with high numbers of other immigrants" (Sampson, 2012: p. 252). Immigration has rejuvenated such inner-city neighborhoods as Queens in New York, the Lower West Side of Chicago, pockets of the North and South Sides of Chicago, Bushwick in Brooklyn, and large strips of south central Los Angeles and has lowered their crime rates. "It follows," states Sampson, "that the spatial 'externalities' associated with immigration are multiple in character and constitute a plausible mechanism to explain some of the ecological variation in crime rates of all groups in the host city" (Sampson, 2012: p. 253).

What should be underlined is that although the links between economic hardship and homicide are similar for blacks, Latinos, and Haitians, and although concentrated disadvantaged and neighborhood social processes predict similar patterns of violence across racial and ethnic groups, immigrants tend to be less violent, especially if they reside in concentrated inner-city immigrant enclaves. The research of Morenoff et al. (2001) revealed that residing in a neighborhood of concentrated immigration, even after adjusting for a series of correlated factors such as an individual's immigrant status and poverty, was directly related to lower rates of violence. In short, it appears that immigration is 'protective' against violence.

Moreover, as Sampson points out, increases in immigration are also linked to crime decreases in black inner-city neighborhoods that are proximate to the center of Mexican immigration such as the middle, as well as near southern section of Chicago and the far north-western areas of the city "that have been predominantly white" (Sampson, 2012: p. 255). Sampson hypothesizes that the spillover 'immigration effects' may eventually spread to other Chicago neighborhoods near immigration centers, including the inner-city black neighborhood of Englewood. In Chicago, "crime dropped in the 1990s as immigration surged" (Sampson, 2012: p. 253).

Because immigration results in the spread of diverse and formerly external cultures, this logically implies that the spatial process of diffusion has a cultural overtone. Referring to the subculture of violence in the black inner city depicted in Anderson's 'Code of the Street' (1999) where shared expectations for 'saving face' and demanding respect in the 'street culture' often lead to violent reactions to otherwise petty encounters such as perceived insults and slights, Sampson theorizes that "if one does not share the cultural attribution or perceived meaning of the event, violence is less likely. Outsiders to the culture, that is, are unlikely to be caught in the vicious cycle of interaction (and reaction) that promote violence" (Sampson, 2012: p. 257). In other words, it may be hypothesized that current immigration is resulting in greater perceptibility of competing nonviolent mores that are diffused through social interaction and not only affect immigrant communities, but also depress violent behavior in general.

However, more research is needed on the lasting effects of immigration on crime. Morenoff et al. (2001) found (after accounting for differences in individual, neighborhood, and family background) that first-generation immigrants (those born outside the United States) in Chicago were 45% less likely to commit violent crimes than third-generation Americans. Accordingly, the *Latino paradox* – "whereby Latinos do much better on various social indicators, including violence, than blacks and apparently even whites, given relatively high levels of disadvantaged" (Sampson, 2012: p. 251) – is driven mainly by first-generation immigrants. If the sharp drop in Latino immigration to the United States that began in 2005 represents a long-term rather than a short-term pattern, rates of violence in Latino communities could increase significantly and the structural and related cultural problems that drive high-crime rates in the inner-city African American community could conceivably over time have similar and notable effects in Latino communities, effects that offset the positive impacts of the first generation on inner-city neighborhoods, including their nonviolent mores.

The Inner-City and Public Policy

The prolonged economy recovery of the 1990s provided much hope for poor inner-city residents. Unemployment declined dramatically from 8.6% in 1992 to 5.1% in 1998, and unemployment in the nation's 50 largest cities fell from 8.6% in 1992 to 4.9% in 1998, a decline that exceeded the decreases in unemployment in the cities' surrounding suburbs (U.S. Department of Housing and Urban Development, 1999).

The Great Recession laid waste to many of these impressive gains and many inner-city families that prospered as a result of their hard work in the construction industry and other occupational sectors that surged during the boom are once again finding themselves in dire financial strains. In an effort to blunt some of the misfortune that disproportionately impacts low-income Americans, the Obama Administration's American Recovery and Reinforcement Act included important supports for low-income families. As part of this stimulus package, $80 billion was allocated to extend, among other things, unemployment benefits, temporarily increase the amount of the Earned Income Tax Credit and expand the Supplemental Nutrition Assistance Program (food stamps) for low-income families and children. In addition, the Patient Protection and Affordable Care Act (commonly known as Obamacare) includes provisions that extended health insurance for children to remain covered under their parents'

health insurance plan until they reach the age of 26, which provides needed coverage for young adults in a tight labor market who are unable to find work. Without such coverage, underskilled or undereducated unemployed young adults would have few if any alternatives to accessing quality health services (Wilson, 2013).

However, as the aftershocks of the Great Recession continue to reverberate around the country there is strong evidence that the gains of the 1990s are being eroded for many blacks and Latino families. A recent report by the Pew Charitable Trust estimates that wealth disparities between white households and those of blacks and Latinos are the largest they have been since the government began publishing these figures. The authors remark that "… in percentage terms, the bursting of the housing bubble in 2006 and the recession that followed from late 2007 to mid-2009 took a far greater toll on the wealth of minorities than whites. From 2005 to 2009, inflation-adjusted median wealth fell by 66% among Latino households and 53% among black households, compared with just 16% among white households" (Kochhar et al., 2011: p. 1).

Just a few weeks into the first legislative session of the 113th session of the United States Congress, there is no indication that any new major initiatives are being considered to tackle the declining conditions of the inner-city poor. Indeed, there is no real evidence that a serious attempt to expand employment options for the hundreds of thousands of unemployed and underemployed adults in the country regardless of race, ethnicity, gender, or age is taking shape. Done well and done properly, such efforts should seek to realign the needs of working adults with those of a vastly changing labor market. Adults with a high school diploma or who dropped out of high school face especially onerous challenges to escaping poverty as adults. The abysmal high school dropout rates among blacks and Latinos, especially in the inner city, makes them particularly susceptible to continuing the intergenerational cycle of poverty. Time will tell as to whether a workforce development policy that is serious about meeting the needs of a more specialized labor market with better prepared workers is able to facilitate training and education programs that engage young adults in ways that the formal education system has been unable to do. Similarly, it remains to be seen whether policy makers can work with the business sector to find ways to expand the earning potential of underskilled workers until such time as they can acquire the competencies to be competitive in the labor market. Finding solutions to long-term joblessness and the concentration of poverty in disadvantaged urban areas are not issues that federal, state, or local governments can tackle on their own. Neither can an increasingly specialized workforce be expected to absorb large numbers of low- or unskilled workers, many of whom are also deficient in essential soft skills, which are prerequisites for many employment opportunities. Real and lasting change will require the sort of public–private partnerships that allow for the targeted training of workers to meet the specialized needs of industry. Working together, the public sector can help underwrite the training component and assistance to trainees while they upgrade their skills in the time that the industry can ensure a smooth transition for trained workers into positions where employees' earning potential provides real home for escaping poverty and economic stability.

See also: Crime, Immigration, and Ethnicity; Crime: Organized; Discrimination: Racial; Poverty, Sociology of; Racism, Sociology of; Residential Segregation in the United States; Underclass; Urban Violence.

Bibliography

Anderson, E., 1999. Code of the Street: Decency, Violence, and the Moral Life of the Inner City. W.W. Norton, New York.

Bartelt, D.W., 1993. Housing the 'Underclass'. In: Katz., Michael B. (Ed.), The "Underclass" Debate: Views from History. Princeton University Press, Princeton, NJ, pp. 118–157.

Condon, M., 1991. Public Housing, Crime and the Urban Labor Market: A Study of Black Youth in Chicago (Working Paper Series No. H-91-3). Malcolm Wiener Center, John F. Kennedy School of Government, Harvard University, Cambridge, MA.

Drake, S.C., Cayton, H.R., 1962. Black Metropolis: A Study of Negro Life in a Northern City. Harper & Row, New York.

Du Bois, W.E.B., 1899. The Philadelphia Negro. University of Pennsylvania Press, Philadelphia, PA.

Fox, R.K., Treuhaft, S., 2006. Shared Prosperity, Stronger Regions: An Agenda for Rebuilding America's Older Core Cities. Report Prepared for Policy Link. Oakland, CA.

Holzer, H.J., Offner, P., Sorensen, E., 30 August 2003. What Explains the Continuing Decline in Labor Force Activity Among Young Black Men? Paper Presented for the Color Lines Conference. Harvard University Press, Cambridge, MA.

Katz, M.B., 1993. Reframing the 'Underclass' Debate. In: Katz, M.B. (Ed.), The "Underclass" Debate: Views from History. Princeton University Press, Princeton, NJ, pp. 440–478.

Kelley, R., 1993. The black poor and the politics of opposition in a new south city. In: Katz, M.B. (Ed.), The "Underclass" Debate: Views from History. Princeton University Press, Princeton, NJ, pp. 293–333.

Kochhar, R., Fry, R., Taylor, P., 2011. Wealth Gaps Rise to Record Highs between Whites, Blacks, Hispanics. Pew Research Center, Washington, DC. Available at: www.pewsocialtrends.org/2011/07/26 (accessed 11.01.13.).

Massey, D.S., Denton, N.A., 1993. American Apartheid: Segregation and the Making of the Urban Underclass. Harvard University Press, Cambridge, MA.

Morenoff, J.D., Sampson, R.J., Raudenbush, S., 2001. Neighborhood inequality, collective efficacy and the spatial dynamics of urban violence. Criminology 39, 517–560.

Passel, J., Cohn, D., Gonzales-Barrera, A., 23 April 2012. Net migration from Mexico falls to zero—and perhaps less. Pew Hispanic Center Report.

Quane, J.M., Wilson, W.J., 2012. Making the connection between socialization and social isolation in low-income urban neighborhoods. Urban Studies 49 (14), 2977–2987.

Quillian, L., 1999. Migration patterns and the growth of high-poverty neighborhoods, 1970–1990. American Journal of Sociology 105 (July), 1–37.

Sampson, R.J., Wilson, W.J., 1995. Toward a theory of race, crime, and urban Inequality. In: Hagan, J., Peterson, R. (Eds.), Crime and Inequality. Stanford University Press, Stanford, CA.

Sampson, R.J., 2012. Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago Press, Chicago, IL.

Sugrue, T.J., 1993. The structure of urban poverty: the reorganization of space and work in three periods of American history. In: Katz, M.B. (Ed.), The Underclass Debate: Views from History. Princeton University Press, Princeton, NJ, pp. 85–117.

U.S. Department of Housing and Urban Development, 1997. The State of Cities. Government Printing Office, Washington, DC.

U.S. Department of Housing and Urban Development, 1999. The State of Cities. Government Printing Office, Washington, DC.

Wilson, W.J., 1987. The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy. Chicago University Press, Chicago, IL.

Wilson, W.J., 1996. When Work Disappears: The World of the New Urban Poor. Knopf, New York.

Wilson, W.J., 9 January 2013. Obama Heads in the Right Direction on Poverty and Jobs. The Root. Root.com.

Innovation

Roy Green, UTS Business School, Ultimo, NSW, Australia
Renu Agarwal and Danielle Logue, Management Discipline Group, UTS Business School, Ultimo, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M.A. West, volume 5, pp. 2895–2900, © 2001, Elsevier Ltd.

Abstract

Innovation, and the understanding of processes of innovation, is fundamentally about social, scientific, and economic change. Innovation is seen as underpinning the productivity and success of firms, communities, and nations. Innovation, as a concept, is increasingly considered as a nonlinear multidimensional phenomenon, which involves a complex relational ecosystem with concurrent processes of value cocreation through collaboration. Innovation can be understood on several broad dimensions – product or process; technological or nontechnological; radical or incremental; and open or closed. The article will explore these dimensions, the historical background to innovation, and current issues and challenges with reference to developments in the social, scientific, engineering, and behavioral sciences, including the cross-disciplinary linkages, which illuminate the nature and significance of innovation for the twenty-first century.

Introduction

Innovation, and the understanding of processes of innovation, is fundamentally about social, scientific, and economic change and is underpinned by human experience and action. Writers from Joseph Schumpeter to Thomas Kuhn have described waves of change, where old paradigms and methods are questioned, tested, and creatively destroyed and new ideas, inventions, and innovations emerge.

Over time, the meaning of innovation has itself changed. Originally associated with science and technology, and the new product development approach for commercializing ideas and inventions, innovation increasingly encompasses interactions among institutional, technological, and human systems. Consequently, the understanding of how innovation occurs has also changed from a traditional linear process of invention followed by commercialization to a multidimensional phenomenon involving a complex relational ecosystem with concurrent processes of value cocreation through collaboration, 'design thinking,' and business analytics. George Lakoff and Mark Johnson say that the essence of innovation lies in the normal pattern of everyday life embodied in the experience of individuals, across cultures and domains (Lakoff and Johnson, 2009).

Innovation is now broadly defined by governments, firms, and international organizations as the mechanism by which economic performance can be improved, and even transformed, and social progress achieved. It is a significant part of the strategy and rhetoric of many organizations, cutting across disciplines and specializations, and is seen as a key not only to increasing productivity and competitiveness, but also to driving environmental sustainability and social inclusion. Managers and organizations increasingly engage in innovation 'ecosystems,' investing both in the tangible aspects of innovation, most notably technological research and development (R&D), and in more intangible aspects, such as social capital, knowledge creation, application and dissemination, systems integration, and business model design. The interaction of these organizations, within markets, with research and educational institutions and public funding agencies has evolved into local and national systems of innovation, with Freeman (1982) and Nelson and Winter (1982) among the first to conceptualize the development of these national systems.

In the context of unprecedented technological advancement, demographic change, environmental pressures, and globalization, this article will provide an updated taxonomy of innovation concepts and practice, with an emphasis on the role of management and organizations, to illuminate the nature and significance of innovation for the twenty-first century. The historical background to innovation, common types of innovation, and current issues and challenges in the management of innovation will be explored with reference to developments in the social, scientific, engineering, and behavioral sciences, including cross-disciplinary linkages.

History

Some of the oldest innovations stretch back to prehistoric times when innovation was closely linked to survival and adaptation. Innovation can be detected in the cave paintings in the Chauvet Cave in France, arguably a precursor to the innovation of filmic representational processes. The paintings, some dating back 35 000 years, depict a number of animals through a technique employing subtle shading and an attempt at perspective. Filmmaker Werner Herzog, in his three-dimensional documentary *Cave of Forgotten Dreams* (2011), conjectures that the paintings, with their attempt at capturing movement with light, prefigure the creation of cinema. The style of the painting was developed by incorporating the shape and texture of the cave's walls to create a sense of movement in conjunction with a flickering light from fire. This exemplifies a characteristic of innovation – it can be a slow and continuously progressed idea, which evolves and takes a final form over time. Even in the contemporary context, film is continuously advanced through new technologies and techniques heightening the filmic experience, while simultaneously transforming the industries around it.

Definition

Innovation is broadly defined as 'an idea, practice, or object that is perceived as new by an individual or other unit of adoption' (Rogers, 2001: 7540). It can apply to any phenomenon (material and immaterial) developed by an individual/collective that is in itself a catalyst for improvement in any area of life. The term 'innovation' when entered in Google currently brings up over 400 million hits. This is ever increasing, with the concept of innovation becoming more pervasive in all aspects of social and behavioral sciences.

Over time, the definition of innovation has widened significantly; originally associated with science and technology and the new product-development approach for commercializing ideas and inventions, innovation increasingly encompasses interactions among technological and human systems that lead to managerial and organizational change. The definition of innovation in the 2005 Oslo Manual as 'the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organizational method in business practices, workplace organization or external relations' is still relevant and valid. Paul Trott (2012) adds a further dimension – innovation as an integrated management process of idea generation, technology development, knowledge application, and commercial exploitation to create a new (or improved) product, process, service, or organizational change.

Innovation is inherently connected to broad social, scientific, and economic changes. Joseph Schumpeter was an early writer on innovation, exploring large-scale change through economics. Originally used by Marx, the term 'creative destruction' was further developed by Schumpeter to describe how the creation of new modes of production destroyed existing modes that had themselves been the result of past innovations. That is, over time one can observe innovation as waves of change. Schumpeter saw technological competition (where successful innovators spurred imitators) as a way to understand economic development. In his work *Capitalism, Socialism and Democracy*, he proposed that large firms with monopoly power were necessary to support innovation, in part a reflection of the transformations that had occurred in the American economy (Schumpeter, 1942).

These broader waves of growth cycles or technological development in Western societies were also termed the 'Kondratieff waves.' They follow the sequence of early mechanization (1770s–80s), steam power and railways (1830s–40s), electrical and heavy engineering (1880s–90s), automobile production (1930s–40s), and information and communication (1980s–90s), extending into the present-day information revolution. Each of these waves of innovation was accompanied by the growth of specific industries led by different countries, in turn leading to global economic shifts over time. For example, while the first wave was led by Britain, France, and Belgium, the second wave was propelled by the United States and Germany, and the final wave is being driven by Asian countries – Taiwan, Korea, Japan, etc. (Trott, 2012). These waves of change are also reflected in Thomas Kuhn's *The Structure of Scientific Revolutions*, which describes knowledge production in terms of paradigm shifts, where sets of understandings, assumptions, and knowledge are created, established, and then eventually displaced by new paradigms (Kuhn, 1962).

This historical brief highlights the heterogeneity of the innovation process across time, across sectors, and across countries, and shows that, although innovation is closely associated with 'invention,' it is much more than invention – it occurs not just in laboratories and universities but across society and the workforce. There is a conceptual interdependence between innovation and invention and 'time lag' is one of the major factors that differentiate the two concepts. Invention is a first occurrence of an idea, and is often associated with technical developments and novelty. Innovation, while it may include technical/technological and nontechnological change, it also involves a broader process that leads to the diffusion and adoption of an invention in practice. Innovation is challenging, and being a 'first mover' does not necessarily secure a strategic position or guarantee success.

From Linear to Open

Innovation has been historically conceptualized as a linear process of development, associated with new product development, and the activity to take it to market. In such models, innovation goes through a series of processes/stages such as research, or a scientific basis, for the innovation, development, production, and marketing. This linear approach leads to an assumption of a generalized chain of causation, and neglects the multiple feedback loops and iterations that occur during the evolution of an invention or innovation.

Although most innovations stem from scientific breakthroughs, this is not always true. An organization's ability to review and combine new and existing knowledge, as well as its ability to interact with external parties, has led to innovation, more recently being considered as an open process, with complex and integrative relationships between consumer, product, firm, and market. In this way, organizational boundaries are porous and firms must leverage outside expertise, and the insights of consumers, in developing new products and services.

The Diffusion of Innovation

The diffusion of innovation, looking at how new products, services, and ideas spread, has long been a topic of research across both the social sciences and natural sciences. The cumulative adoption of innovation (of any type) over time generally follows an S-shaped curve as the product moves through its life cycle. Everett Rogers classified consumers, as a group of adopters based on their demographic and psychographic features, into five segments – innovators, early adopters, early majority, late majority, and laggards, according to how they successively adopted the innovation through the stages of the product life cycle (Rogers, 1995).

Geoffrey Moore extends this by classifying adopters as technology enthusiasts, visionaries, pragmatists, conservatives, and sceptics, who sequentially adopt the innovation as the product life cycle progresses (Moore, 2004). Ordinarily, the rate of adoption of radically new products, services, and ideas is impeded by consumer unfamiliarity and uncertainty.

The Systemic Nature of Innovation

The extent to which innovation is developed and sustained within organizations is contingent upon the external business environment and, in turn, the wider economic, social, and political context in which the organizations operate. Increased importance is placed on defining the linkages between the innovation processes within organizations and the wider system with which these processes are integrated. The systemic frameworks of innovation that have emerged are embedded in the technological, institutional, infrastructural, and economic realms (Freeman and Soete, 1997).

Innovation systems also operate on sectoral, regional, and national levels. In sectoral systems of innovation, organizations belonging to a particular sector engage in innovation through collaborative networks, and leveraging sectoral knowledge and technologies. Regional systems focus on territorial-based innovation systems, leveraging the success of regional clusters and industrial districts, while national systems emphasize interlinked political, economic, social, and cultural factors within a country (Lundvall, 1992). These three perspectives – national, sectoral, and regional – are variants of systems of innovation.

For organizational managers, and policy-makers, the finance, technology, and infrastructure requirements of the innovation system, and the complementarities that exist between these, are crucial. Private and public sector agents such as government bodies, and educational and research institutions, play a critical role in ensuring that innovation systems are supported by the right mix of technological, social, and economic infrastructure.

Types of Innovation

The phenomenon of innovation has historically brought change and transformation. Today, innovation is increasingly seen through a cross-disciplinary lens spanning engineering, economics, sociology, organizational science, and management perspectives.

Product or Process Innovation

Joseph Schumpeter was among the first to classify innovation into five different types: new products, new methods of production or transportation, new sources of supply, new markets, and new ways of organizing business (Schumpeter, 1942). Earlier research on innovation focused on the first two of these: product innovation and process innovation. Product innovation concentrates on the development of new or improved products and services, for example, the introduction by Apple Inc. of the iPod and iPad. Process innovation, on the other hand, involves the development of new ways of producing products and services, for example, lean manufacturing methods, just-in-time (JIT) manufacturing systems, and quality circles.

Technological or Nontechnological Innovation

While the distinction between product and process innovation is useful, the classification of innovation as 'technological innovation' and 'nontechnological innovation' is more relevant to recent developments. Technological innovation involves new or improved technology, for example, new types of machinery, while nontechnological innovation finds new ways of producing, distributing, or organizing work. Nontechnological innovation is associated with organizational innovations such as new or improved organizational structures, systems, or processes. Examples of nontechnological organizational innovations include total quality management, Six Sigma, and business process reengineering.

Business model innovation is a form of organizational innovation where businesses create innovative ways to create and deliver value to their stakeholders. Accordingly, management innovation has been expressed as "*invention and implementation of a management practice, process, structure, or technique that is new to the state of the art and is intended to further organisational goals*" (Birkinshaw et al., 2008: 825). IKEA and Dell are business model innovators who have fundamentally changed the structures and dynamics of their respective industries. IKEA cocreates value with its customers by outsourcing the transport and assembly of products to them. Dell, on the other hand, has innovated its supply chain by eliminating distributors, and instead moved to assembling-to-order, which allows the customization and personalization of products to individual customer needs.

Incremental or Radical Innovation

Innovation can also be categorized as incremental or radical based on how revolutionary and ground-breaking it is (Freeman and Soete, 1997; Christensen, 2003). Incremental innovations are continuous improvements that incrementally change existing products or services, and are regarded as 'sustaining innovations.' Radical innovations, on the other hand, are discontinuous, path-breaking changes that have far-reaching impacts, and often transform the face of the industry and marketplace. Examples of such radical innovations include online music downloads (disrupting CDs), digital film (disrupting photographic film), online news platforms (disrupting traditional newspapers), and online shopping (disrupting the brick-and-mortar models in the retail industry). Such innovations are also referred to as 'disruptive innovations' and can be equated to Schumpeter's concept of creative destruction.

Open or Closed Innovation

Innovation can be open or closed depending on the extent to which the organization interacts with its environment. Historically innovations were mainly closed, involving little communication or collaboration with entities outside of the organization. Today's highly networked, knowledge-based economy increasingly requires open innovations where the innovation activities involve knowledge exchange with external entities, in particular, across the organization's supply chain (Chesborough, 2003). Open innovation leverages external relationships and collaborations to enhance systemic knowledge-sharing and drives innovation outcomes.

User-Led Innovation and Service Innovation

Service innovation is a type of open innovation gaining prominence in recent times. With the growing economic importance of the services sector, service innovation is playing an ever more significant role in driving growth in today's knowledge-intensive economy. This type of innovation responds to increased customer input, interactivity, and specialization; is multidimensional in nature; and includes strategic, productivity, and performance dimensions. Services with higher interactivity with customers provide increased opportunities for innovation through interactions involving enhanced knowledge-sharing and learning processes, and it is these higher-order capabilities that contribute to innovation in services. Examples include Apple's development of the app resulting in value cocreation, and the development in Australia of a virtual critical care unit for patients receiving emergency care in remote locations through real-time digital telehealth.

The Innovation Life Cycle

Innovation opportunities change over time, and the most important innovations are often based on the life cycle of an industry or product. These innovations can involve changes that totally transform economic value. Mature industries tend to focus on product innovation or repositioning their existing products, while newer industry sectors, such as biotech or nanomaterials, have scope to experiment with new product or service concepts. The innovation life cycle has three sequential phases: fluid, transitional, and specific. The fluid phase is characterized by radical product innovation, the transitional phase involves process innovation, and the final specific phase comprises incremental innovation (Abernathy and Utterback, 1978).

Management of Innovation

At the organizational level, the management of innovation is critical to develop and foster sustained innovation as a driver of long-term survival, and a sustainable competitive advantage. Innovation can be managed and facilitated through organizational structure, strategy, processes, leadership and culture, and the connection of these elements to organizational performance.

Structure

Management of innovation in organizations involves the constant pressure to strike a balance between the stable, efficiency-seeking routines of day-to-day operations and the dynamic, creativity-seeking routines that promote innovation. Organizations need to be ambidextrous in order to simultaneously engage in both exploitation – focusing on control, efficiency and productivity, and exploration – embracing creativity, search, and discovery. One way to do this is to design organizational structures with separate subunits focusing on exploration and exploitation as well as internally aligning systems, processes, and cultures to effectively foster both explorative and exploitative activities (O'Reilly and Tushman, 2008). Organizations can cultivate weak ties with some network partners and strong ties with other partners, and this can create a balance between novelty and reliability in innovation processes.

Earlier theories argue that 'organic' structures, which are generally more decentralized, less formalized, and more flexible, are more capable of facilitating innovation compared to 'mechanistic' structures that are more centralized, formalized, and rigid. A network model is another structure that supports openness and communication among partners; and organizational structures that facilitate informal and formal networking, as well as lateral communication and knowledge-sharing can also foster creativity.

Strategy

Two general schools of strategy have implications for the management of innovation: the market-based view and the resource-based view. According to the market-based view, the drivers and inhibitors of innovation within an organization are contingent on external market conditions and innovation is dependent on firms' success in exploiting opportunities in the marketplace. The resource-based view sees the presence of valuable, rare, inimitable, and nonsubstitutable resources and capabilities within an organization as the basis for a sustainable competitive advantage built on enhanced innovation performance (Prahalad and Hamel, 1990). The dynamic capabilities approach is an extension of the resource-based view and contends that firms need to constantly align, modify, and reconfigure their resources and capabilities in a dynamic and volatile environment to ensure sustained innovation and earnings above normal returns (Teece, 2009).

Knowledge Management Processes

The early linear models of innovation mainly focused on the role of technology in stimulating innovation, and then included the role of the marketplace and customer interactions as key to the innovation process. It was not until later that the role of knowledge as a conduit between all stages of innovation, starting from idea generation to commercialization of innovation, was recognized. Terms like 'architecture knowledge' explain how technological knowledge is integrated as a central process in fostering innovation. Different aspects of organizational learning also play into knowledge creation, which in turn feeds into innovation processes. How effective this knowledge acquisition and assimilation is, will impact on knowledge exploitation, a critical component for generating innovative output. Technology also assists in this, enabling better information flow and knowledge exchange, and driving innovation through facilitation of creativity, communication, and operational improvement processes.

Leadership and Organizational Culture

Organizational managers recognize that, in managing innovation, traditional management approaches need to be reframed to address the dynamic, networked, and complex nature of innovation processes. Empowered cross-functional teams, and shared leadership, creates the necessary environment for

innovation to flourish. Innovation in organizations is also more likely to be supported where the leadership is more open to new ideas and solutions. All cultures are dynamic and continuously evolving. Organizational culture is no different, and workplace cultures should reflect this through: fostering employee involvement in decision-making processes, promoting openness to learning and change, and being responsive to customer and stakeholder needs. This is likely lead to an organization's superior innovation performance and advantage over its peers.

Measuring Innovation

Innovation, by definition, involves novelty and includes both the creation of new knowledge and the reinterpretation and application of existing knowledge. This makes innovation inherently difficult to quantify and measure. Indicators such as the European Community Innovation Survey (CIS) are centered on innovation inputs and outputs and produce national measures that have some degree of international comparability. Major indicators used to measure innovation include data on R&D, patent applications and grants, scientific publications and citations (bibliometric data), and technical performance characteristics (technometric data).

CIS works well for manufacturers, but not for the heterogeneous services sector, and its associated intangible inputs and outputs. In response, a framework of seven innovation management measurement has been developed. They are inputs, knowledge management, innovation strategy, organizational culture, portfolio management, project management, and commercialization. Devising metrics to quantify intangible elements involves a multiindicator approach that integrates multiple data sources on innovation activities. These may comprise resources (people and ideas), efforts (measured by innovation expenditure of various sorts), innovation outputs (innovation outcomes including numbers of patents), and performance (including the percentage of sales due to new products and services). These measure innovation on a holistic basis. One problem with this approach, however, is that it does not necessarily encompass services, managerial, and open innovations centered on social and human elements.

National Systems of Innovation

At the heart of any national innovation system is the interaction of organizations with other organizations in the value network, such as suppliers, competitors, and customers. Other networks of institutions including universities, and government and public agencies are also embodied in the system. A national system of innovation includes all economic, social, political, and institutional factors impacting on the development and diffusion of innovations, both nationally and globally.

Innovation is a driver of economic growth and productivity. It evolves out of a mixture of activities, and is influenced by framework conditions but these are not globally uniform. There are big variations in the performance of nations linked to the characteristics of their innovation systems. Nations have different levels of direct investment in R&D and indirect investment in other mechanisms such as tax benefits, subsidies, and incentives for collaboration, skills and education levels, and national policy decisions, all of which act as crucial stimulants for innovation. National systems of innovation seek to provide a well-coordinated governance framework for achieving congenial national conditions that foster innovation as a means of sustaining long-term economic growth. National governments, from a range of countries including Australia, Japan, India, the United States and the UK, and international organizations such as the OECD and the UNCTAD, use policy-making instruments to orient innovation systems toward enhancing innovative performance in national economies. In some countries, for example, educational institutions such as universities drive R&D, while in others, independent research institutes are the key R&D performers.

Emerging Directions and Challenges

Innovation is now viewed via a cross-disciplinary approach, including economics, cognitive sciences, sociology, organizational science, and management and business studies.

Technological Directions

As we move forward into the twenty-first century, the concept of cloud computing is a recent disruptive innovation providing a platform for many innovations in the future. It has the potential to be the next major driver of business model innovation. In the new paradigm of cloud computing, there is an intersection of technology as a complex conglomeration of products and services is offered as a utility, rather than a service. It harnesses various aspects and modes of technology such as social and business networking sites, and personal digital/computer products/applications such as smart phones, e-mail, tablets, etc. Cloud computing requires minimal management by service providers and gives 24-hour access anywhere to personal data and computer resources held remotely.

In effect, this innovation empowers the end user to be free of dependency on software applications and the more prosaic aspects of computer and digital device use. The innovative aspects of cloud computing encapsulate recent developments and implementations of computer-based services and applications, and their use. It is a paradigm shift in the way technology is used and will have a far-reaching impact into the twenty-first century. The ability to capitalize on this technology into the future will depend on how cultural differences affect the likelihood, and pace, of business and IT cloud service adoption.

The New Face of Manufacturing

Additive manufacturing is a process of manufacturing that signals a major shift in the way products are created. It can be defined as technologies linked to a type of manufacturing embracing key elements of digital technology, computer-aided design, and three-dimensional modeling information, in order to create products through a layering of material using innovative manufacturing techniques. Additive manufacturing will be highly relevant to future improvements in medical and dental technologies, and many other areas where creating

tailor-made products will provide precision and convenience. The future directions for this type of technology include the notion of a 'desktop factory'; this has already begun to take place with the sale of three-dimensional printers released in the market since 2010. This makes additive manufacturing a new frontier for the development of prototypes, models, and items for various purposes ranging from engineering and industrial to personal needs, and encompassing functional aspects as well as the area of entertainment. The traditional manufacturing approach of keeping the customer removed from the operation is not possible, or appropriate, for service organizations. The servitization of the manufacturing industry, and the bundling of services with products, has blurred the distinction between product and service. Technological advancements, and increasing globalization, have pushed value cocreation; and open innovation, as a result of customer participation, has prompted business model, social, and service innovations.

Design and Integrative Thinking

In response to 'wicked problems,' a category of problem that cannot be singularly defined nor solved, innovators are turning to new ways of problem-solving that are far different from the traditional methods of rational, linear-based methods. Design thinking is one such alternative method, which encompasses the unwrapping of complex problem-solving processes – it is a user-centered and iterative approach focused on the solution, rather than the problem. It has application for both policy-makers and organizations. Traditional ways of running companies/firms are no longer adequate for the twenty-first century, and it is increasingly recognized that integrative thinking is essential for all organizations. Integrative thinking involves principles of design thinking being ingrained in the innovation processes that are adopted. This engenders a mind-set of creative thinking-in-action and fosters a creative and human-centered organizational culture.

Implications for Teaching Innovation/Management Education

The assumption that organizations, their behavior, culture, values and doctrines, and management are central to stimulating innovation highlights the role of management education and business schools. The increasing importance of human capital in innovation means the role of management education, and in particular, the way business schools might better develop management competencies and attributes, is crucial in fostering innovation.

The recent work on management education by Datar et al. (2010) reviewed a number of leading North American and European business schools and identified unmet needs relating to 'knowing' (identify, analyze, explain), 'doing' (learn skills needed), and 'being' (deeper psychological understanding of self). Management innovation needs to innovate and change future educational approaches and outcomes. The emergence of design and integrative thinking for business and management, and the necessity to solve 'wicked problems,' creates an essential need to build innovation capability in organizations. More broadly, there is a challenge globally for policy-makers and organizations alike to draw on the talent and creativity of workforces to innovate our approaches to value creation, economic efficiency, environmental sustainability, and social progress.

Conclusion

Innovation, as a concept, is increasingly considered as a nonlinear multidimensional phenomenon, which profoundly affects us. Despite its significance, the challenges of innovation are immense. Organizations rarely innovate alone: increasingly, innovation involves a complex relational ecosystem with concurrent processes of value cocreation through collaboration. Innovation can be understood on several broad dimensions – product or process; technological or nontechnological; radical or incremental; and open or closed. Innovation is a relentless process of change and improvement that is inevitably characterized by successes and failures; and its outcomes are always uncertain. Nevertheless, the importance of innovation is clear, not only in driving the growth and profitability of individual enterprises, but also in realizing the social and productive potential of nations and communities.

Acknowledgment

The authors acknowledge the research contributions of Dr Alexandra Pitsis and Ms Krithika Randhawa in preparing this article.

See also: Business Models; Corporate Culture; Diversification, Relatedness, and the New Logic of Co-Creation; Entrepreneurship and Economic Growth; Information and Knowledge, Organizational; Organization: Overview; Social Business; Social Innovation; Social Networks; Strategic Adaptation; Strategic Management; Strategizing.

Bibliography

Abernathy, W.J., Utterback, J., 1978. Patterns of industrial innovation. In: Tushman, M.L., Moore, W.L. (Eds.), Readings in the Management of Innovation. Harper Collins, New York, pp. 97–108.

Birkinshaw, J., Hamel, G., Mol, M., 2008. Management innovation. Academy of Management Review 33 (4), 825–845.

Chesborough, H., 2003. Open Innovation: The New Imperative for Creating and Profiting from Technology. Harvard Business School Press, Boston, MA.

Christensen, C.M., 2003. The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail, third ed. HBS Press, Cambridge, M.A.

Datar, S., Garvin, D., Cullen, P., 2010. Rethinking the MBA: Business Education at a Crossroads. Harvard Business School Publishing, Boston.

Freeman, C., 1982. The Economics of Industrial Innovation, second ed. Frances Pinter, London.

Freeman, C., Soete, L., 1997. The Economics of Industrial Innovation, third ed. Pinter, London.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago, ISBN 0-226-45808-3.

Lakoff, G., Johnson, M., 2009. Metaphors We Live By. The University of Chicago Press, Chicago and London.

Lundvall, B.A., 1992. National Systems of Innovation – Towards a Theory of Innovation and Interactive Learning. Pinter, London.

Martin, R., 2007. The Opposable Minds: How Successful Leaders Win through Integrative Thinking. Harvard Business School Press, Boston USA.

Moore, G.A., 2004. Crossing the Chasm: Marketing and Selling Technology Products to Mainstream Customers, second ed. Capstone, Oxford.

Nelson, R.R., Winter, S.G., 1982. An Evolutionary Theory of Economic Change. Harvard University Press, Cambridge Mass.
OECD and Eurostat, 2005. Oslo Manual: Guidelines for Collecting and Interpreting Data, third ed.
O'Reilly, C.A., Tushman, M.L., 2008. Ambidexterity as a dynamic capability: resolving the innovator's dilemma. Research in Organization Behavior 28, 185–206.
Prahalad, C.K., Hamel, G., 1990. The core competence if the corporation. Harvard Business Review 68 (3), 79–91.
Rogers, E.M., 1995. Diffusion of Innovations, fourth ed. Free Press, New York.
Rogers, E.M., 2001. Theory of innovation. In: Smelser, N.J., Baltes, P.B. (Eds.), The International Encyclopedia of the Social & Behavioral Sciences. Elsevier, Amsterdam.
Schumpeter, J.A., 1942. Capitalism, Socialism and Democracy. Allen and Unwin, London.
Teece, D.J., 2009. Dynamic Capabilities and Strategic Management: Organizing for Innovation and Growth. Oxford University Press, USA.
Trott, P., 2012. Innovation Management and New Product Development, fifth ed. Pearson Education, Essex, England.

Innovation and Employment

Marco Vivarelli, Università Cattolica del Sacro Cuore, Piacenza, Italy; IZA, Bonn, Germany; and University of Sussex, Brighton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, the theoretical and empirical literature on the employment impact of technological change is critically discussed, the relative explanatory power of the competing theories is compared and the macro-, sectoral, and microevidence on the issue is detailed.

Introduction

Historically, the fear of technological unemployment has always emerged in ages characterized by radical technological changes. For instance, the striking response of the English workers to the first industrial revolution was the destruction of machines under the charismatic leadership of Ned Ludd in the industrial areas and of Captain Swing in the countryside (see Hobsbawm, 1968; Hobsbawm and Rudé, 1969).

On the other hand, since its very beginning, the economic theory has pointed out the existence of economic forces which can spontaneously compensate for the reduction in employment due to technological progress. It is interesting that nowadays – *mutatis mutandis* – the present economic debate about the employment impact of innovation is still characterized by a focus on those market compensation forces which can counterbalance the initial labor-saving effect of process innovation.

In this article, the theoretical and empirical literature on the employment impact of technological change will be critically discussed, the relative explanatory power of the competing theories will be compared and the macro- and microevidence on the issue will be detailed.

In particular, the next section will be devoted to the so-called 'theory of compensation,' a classical theory which is still the basis of current economic models which try to understand the employment impact of innovation. In the light of the classical taxonomy, Section Innovation and Employment: the Empirical Evidence will present a survey of the available empirical evidence on the relationship between innovation and employment both at the macro-, sectoral, and microlevel. Section Conclusion will briefly conclude.

The Theory

By definition, technological change allows to produce the same amount of goods with a lower amount of production factors, namely capital and labor. However, this is only the 'direct' effect of technological change, and particularly of process innovation. Indeed, the economic discipline – since its foundation – has tried to dispel the concerns about the direct harmful effects of technological change pointing out both the market mechanisms able to counterbalance the direct impact of process innovation and the job-creating effects of product innovation.

Thus, in the first half of the nineteenth century – while the Luddites were destroying the new machines – economists put forward a theory that Marx later called the 'compensation theory' (see Marx, 1961: vol. 1, Chap. 13, 1969: Chap. 18). This theory is made up of different market compensation mechanisms which are triggered by technical change itself and which can counterbalance the initial labor-saving impact of process innovation (for an extensive analysis, see also Vivarelli, 1995: Chaps 2 and 3; Petit, 1995; Vivarelli and Pianta, 2000: Chap. 2; Pianta, 2005).

1. The compensation mechanism 'via new machines.'

The same process innovations, which displace workers in the user industries, create new jobs in the capital sectors where the new machines are produced (see, for instance, Say, 1964).

2. The compensation mechanism 'via decrease in prices.'

On the one hand, process innovations involve the displacement of workers; on the other hand, these innovations themselves lead to a decrease in the unit costs of production and – in a competitive market – this effect is translated into decreasing prices; in turn, decreasing prices stimulate a new demand for products and so additional production and employment.

Indeed, the compensation mechanism 'via decrease in prices' has been often reproposed in the history of economic thought both by neoclassical economists at the beginning of this century (see Clark, 1907: p. 270; Pigou, 1962: p. 672) and by modern theorists (see Heffernan, 1981; Neary, 1981; Stoneman, 1983a: Chaps 11 and 12; Hall and Heffernan, 1985; Dobbs et al., 1987; Nickell and Kong, 1989; for a detailed analysis of the hypotheses, procedures and conclusions of these models, see Vivarelli, 1995: Chaps 4 and 6).

3. The compensation mechanism 'via new investments.'

In a world where the competitive convergence is not instantaneous, it is observed that during the gap between the decrease in costs – due to technical progress – and the consequent fall in prices, extraprofits may be accumulated by the innovative entrepreneurs. These profits are invested and so new productions and new jobs are created. Originally put forward by Ricardo (1951: vol. I, third ed. 1821), this proposition has also been called forth by marginalist economists like Marshall (1961: p. 542) and Douglas (1930: p. 936) and by recent dynamic models such as those by Hicks (1973) and Stoneman (1983a: pp. 177–181, 1983b).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.94010-8

4. The compensation mechanism 'via decrease in wages.'

As with other forms of unemployment, the direct effect of labor-saving technologies may be compensated within the labor market through a proper price adjustment. In a neoclassical framework – with free competition and full substitutability between labor and capital – a decrease in wages leads to an increase in the demand for labor. The first to apply this kind of argument to the specific case of technological unemployment was Wicksell (1961: p. 137), followed by Hicks (1932: p. 56), Pigou (1933: p. 256) and Robbins (1934: p. 186).

In modern times, the wage adjustment is a component of partial equilibrium models such as those by Neary (1981) and Sinclair (1981) and generally equilibrium analyses such as those by Layard and Nickell (1985), Venables (1985) and Layard et al. (1991, 1994).

5. The compensation mechanism 'via increase in incomes.'

Directly in contrast with the previous one, this compensation mechanism has been put forward by the Keynesian and Kaldorian tradition. In a Fordist mode of production, unions take part in the distribution of the fruits of technical progress. So it has to be taken into account that a portion of the cost savings due to technical change can be translated into higher income and hence higher consumption. This increase in demand leads to an increase in employment which may compensate the initial job losses due to process innovations (see Pasinetti, 1981; Boyer, 1988a,b, 1990).

6. The compensation mechanism 'via new products.'

Technical change is not equivalent to process innovation, but it can assume the form of creation and commercialization of new products; in this case, new economic branches develop and additional jobs are created. Once again, the labor-intensive impact of product innovations was underlined by classical economists (Say, 1964: p. 88) and even the most severe critic of compensation theory admitted the positive employment benefits that can derive from this kind of technical progress (Marx, 1961: vol. 1, p. 445).

In the current debate, various studies (Freeman et al., 1982; Freeman and Soete, 1987, 1994; Vivarelli and Pianta, 2000; Edquist et al., 2001) agree that product innovations have a positive impact on employment since they open the way to the development of either entire new goods or main differentiation of mature goods. In the latter case, the 'welfare effect' (new branches of production) has to be compared with the 'substitution effect' (displacement of mature products; see Katsoulacos, 1984, 1986).

On the whole, economic theory clarifies those market forces which can potentially counterbalance the initial labor-saving impact of process innovation; moreover, it is recognized that the alternative form of technological change – namely the diffusion of entirely new products – can imply an obvious job-creating effect.

On the other hand, compensation mechanisms can be hindered – or even annihilated – by the existence of important drawbacks. Using the same taxonomy which has been proposed above, the main criticisms of the compensation theory can be singled out as follows.

1. With few exceptions (see Hicks, 1973), nowadays this compensation mechanism is not put forward anymore. Indeed, Marx's critique of this mechanism was so sharp that it stopped this first line of reasoning of the compensation theory: he pointed out that a machine can only be employed profitably if it is the outcome of far fewer men than it replaces (Marx, 1969: p. 552).

Moreover, labor-saving technologies spread around in the capital goods sector, as well; so this compensation is an endless story which can be only partial (Marx, 1969: p. 551).

Finally, the new machines can be implemented either through additional investments (see point C) or simply by substitution of the obsolete ones (scrapping). In the latter case – which is indeed the most frequent one – there is no compensation at all (see, for instance, Freeman et al., 1982).

2. As originally noted by Malthus (1964: vol. II, pp. 551–560) and Sismondi (1971: p. 284), the very first effect of labor-saving process innovation is a decrease in the aggregate demand due to the cessation of the demand previously associated with the dismissed workers. So, the mechanism 'via decrease in prices' deals with a decreased demand and has to more than counterbalance the initial decrease in the aggregate purchasing power.

In addition, this mechanism relies on Say's law and does not take into account that demand constraints might occur. Difficulties concerning the components of the 'effective demand' – in Keynes' terms – such as a low value of the 'marginal efficiency of capital' (see Keynes, 1973: Chap. 11) can involve a delay in expenditure decisions and a lower demand elasticity. If such is the case, this compensation mechanism is hindered and technological unemployment ceases to be a temporary problem (for a detailed criticism about compensation based on Say's law, see also Standing, 1984: pp. 131–133).

Finally, the effectiveness of the mechanism 'via decrease in prices' depends on the hypothesis of perfect competition. If an oligopolistic regime is dominant, the whole compensation is strongly weakened since cost savings are not necessarily and entirely translated into decreasing prices (see Sylos Labini, 1969: p. 160).

3. Also the compensation mechanism 'via new investments' relies on the Say's law assumption that the accumulated profits due to technical change are entirely and immediately translated into additional investments. Again, Marx's and Keynes's treatment of Say's law can be used to argue against the full effectiveness of this compensation mechanism. In other words, the theoretical analysis has to take into account the 'animal spirits' and the economic agents' expectations, which can delay the translation of additional profits into 'effective demand' (see Pasinetti, 1981; Freeman and Soete, 1987; Vivarelli, 1995; Appelbaum and Schettkat, 1995; Pianta, 2005).

Moreover, the intrinsic nature of the new investments does matter; if these are capital-intensive, compensation can only be partial (Marx, 1961: vol. I, p. 628).

4. Also the mechanism 'via decrease in wages' collides with the Keynesian theory of effective demand. On the one hand, a decrease in wages can induce firms to hire additional

workers, but – on the other hand – the decreased aggregate demand lowers the employers' expectations and so they tend to hire less workers.

A second criticism can be launched to this mechanism if the cumulative and irreversible nature of modern technical change is properly taken into account (see Rosenberg, 1976; Dosi, 1988). In this view, science and technology have their own rules: along a 'technological trajectory,' 'localized technical progress' occurs and this can give rise to 'locked in' technologies (see Stiglitz, 1987: p. 128; Freeman and Soete, 1987: p. 42). If the cumulative and localized nature of technical change is taken into account, the hypothesis of perfect substitutability between capital and labor assumed by neoclassical models appears to be quite arbitrary (Freeman and Soete, 1987: p. 46).

5. During the 'golden age' of the 1950s and 1960s the Fordist mode of production was based on a relevant change in the labor-wage nexus. Instead of leaving the wage to be regulated by a competitive labor-market, workers were allowed to take possession of a relevant portion of productivity gains due to technical progress. In turn, the increased real wages involved mass consumption and this stimulated investments leading to further productivity gains through technical progress and scale economies (Boyer, 1988c,d). Labor-saving technologies were introduced on large scale, but the Kaldorian 'virtuous circle' allowed an important compensation 'via new incomes.'

Nowadays, the Fordist mode of production is over for many reasons that cannot be discussed here (see Boyer, 1988c, 1990). The distribution of income follows different rules and the labor market has returned to be competitive. On the whole, this compensation mechanism has been strongly weakened in the new institutional context.

6. New products are still the more powerful way to counterbalance labor-saving process innovations. Yet, different 'technological paradigms' (see Dosi, 1982) are characterized by different clusters of new products which in turn have very different impacts on employment. So, the introduction of the automobile had a much higher labor-intensive effect than the diffusion of the home computers. As a matter of fact, in different historical periods and different institutional frameworks, the relative balance between the labor-saving effect of process innovations and the labor-intensive impact of product innovations can considerably vary.

In conclusion, the relationship between innovation and employment can be represented by a very complex picture where the direct labor-saving impact of process innovation, the compensation mechanisms, the drawbacks and hindrances which can severely weaken the effectiveness of such mechanisms and the labor-friendly nature of product innovation can combine in very diverse outcomes.

While the 'end of work' popular approach (see Rifkin, 1995) erroneously underestimates the opportunities for compensation, economists sometimes forget that compensation can only be partial and that it is dependent on particular historical and institutional circumstances.

Indeed, as it emerges from whatever has been discussed so far, economic theory does not have a clear-cut answer about the employment effect of innovation; then, attention should be turned to aggregate, sectoral and microeconomic empirical analyses which should take into account the different forms of technical change, their direct effects on labor, the various compensation mechanisms and the possible hindrances to these mechanisms.

Innovation and Employment: The Empirical Evidence

Whilst theoretical economists may develop clear models about the employment impact of process and product innovation, applied economists have 'to measure' process and product innovation, the compensation mechanisms and the final employment impact of innovation; from this respect, at least three main problems arise.

First, innovation is difficult to measure; traditional indicators such as R&D (input indicator), patents, and relevant innovation (output indicators) are seldom fully available and are often inadequate to fully represent technological progress. For instance, most of process innovation is implemented through 'embodied technological change' (ETC) (The embodied nature of technological progress and the effects related to its spread in the economy were originally discussed by Salter (1960) and Solow (1960) who underlined that technological progress might be incorporated in new vintages of capital introduced either through additional investment or simply by scrapping.), in turn introduced through gross investment. This technological input is often more important than R&D activities: think for instance to traditional sectors and to the Small and Medium Enterprises (SMEs), where in-house R&D has a very limited role and conversely most of innovation is introduced through the acquisition of capital goods. In terms of Pavitt (1984) taxonomy, technological change in the 'supplier-dominated' firms is much more dependent on machineries acquired by the 'specialized suppliers' than from their own R&D expenditures.

R&D and ETC can be considered the two faces of innovation, with R&D mainly correlated with the labor-friendly product innovation and ETC mainly related to the labor-saving process innovation. Indeed, recent microeconometric studies – using data from the European Community Innovation Surveys (CIS) – have confirmed empirically how R&D expenditures are closely linked with product innovation, while innovative investment (especially devoted to new machinery and equipment incorporating ETC) turns out to be related to process innovation (see Conte and Vivarelli, 2005; Parisi et al., 2006).

Second, as discussed in Section The Theory, the final employment impact of innovation depends on institutional mechanisms which can be very different at the micro-, meso-, and macrolevel and can vary in different economic contexts, such as in different countries or in different sectors within the same country. For instance, a different degree of enforcement in the intellectual property rights (IPRs) can induce a different balance in the distribution between R&D and ETC and consequently can imply different employment outcomes.

Third, it is difficult to distinguish the final impact of innovation on employment, since the latter is influenced by many other factors: namely, the macroeconomic and cyclical

conditions, the labor-market dynamics, the trends in working time, and so on.

The macroeconomic empirical analyses of the relationship between innovation and employment are largely affected by the three methodological problems discussed above. While they are attractive, since they can take into account all the direct and indirect effects of innovation (see previous section), they are often severely constrained by the difficulty to find a proper aggregate proxy of technological change and by the fact that the final employment national trend is codetermined by overwhelming institutional and macroeconomic determinants difficult to control for.

Although to a lesser extent, these considerations also apply when empirical analysis is carried out at the sectoral level. Moreover, sectoral analysis can be affected by either a negative or a positive bias according to the observation point of view (manufacturing versus services).

In this context, microeconometric studies have the great advantage to allow a direct and precise firm level mapping of innovation variables, both in terms of inputs and outputs. Indeed, only firm-level data can account for R&D expenditures, capital formation, product and process innovation, and directly relate them to the firm's employment trend. In other words, only the microeconometric empirical analysis can grasp the very nature of firms' innovative activities and throws some light on the actual ways how new products may generate new jobs and labor-saving process innovation may destroy old ones.

Obviously, the need for the microeconometric analysis should also take into account important limitations of firm-level data. First, the microeconomic approach cannot take fully into account the indirect compensation effects (see Section The Theory) which operate not only at the level of the firm but also through intersectoral linkages and aggregate dynamics. Second, a possible shortcoming of this kind of analysis consists in a 'positive bias' which tends to underline the positive employment consequences of innovation. In fact, once the empirical analysis is developed at the level of the single firm, innovative firms tend to be characterized by the better employment performances since they gain market shares because of innovation. Even when the innovation is labor-saving, microeconomic analyses generally show a positive link between technology and employment since they do not take into account the important effect on the rivals, which are crowded out by the innovative firms (this is the so-called 'business stealing' effect; see Section Microeconomic Empirical Evidence below).

Keeping these methodological remarks in mind, now attention can be turned to the single levels of the empirical analysis of the relationship between technology and employment; consistently with whatever has been discussed so far, the macroeconomic and the sectoral studies will be briefly surveyed, while a more detailed attention will be devoted to the microeconometric literature.

Aggregate Empirical Evidence

As far as the aggregate empirical studies are concerned, macroeconometric studies have tried to test the validity of the compensation mechanisms (see Section The Theory) within a partial or general equilibrium framework.

For instance, Sinclair (1981) put forward a macro- IS/LM scheme and concluded that a positive employment compensation can occur if the demand elasticity and the elasticity of factor substitution are sufficiently high. Using estimates based on US data, the author found a strong evidence supporting the mechanism 'via decrease in wages' but not the mechanisms 'via decrease in prices.'

Layard and Nickell (1985) derived a demand for labor in a quasigeneral equilibrium framework and stated that the crucial parameter was the elasticity of the demand for labor in response to a variation in the ratio between real wages and labor productivity; in fact, technical change increases labor productivity and – given an adequate elasticity – proportionally the demand for labor and this can be enough to fully compensate initial job losses. Using data for the UK economy, the authors estimated an elasticity coefficient equal to 0.9 and this was sufficient – in the authors' opinion – to rule out technical change from the possible causes of British unemployment. (See Vivarelli and Pianta (2000): Chap. 1, for an extensive critical discussion about the role of technology and the role of labor-market flexibility in determining employment performance.)

Using a similar approach, Nickell and Kong (1989) focused their attention to the operating of the compensation mechanism 'via decrease in prices' in nine UK two-digit industries. Putting forward a price equation where cost-saving effects of labor-saving technologies were fully transferred into decreasing prices, the authors found out that in seven sectors out of nine a sufficiently high demand elasticity was able to imply an overall positive impact of technical change on employment.

In Vivarelli (1995: Chaps 7–9) the direct labor-saving effect of process innovation, the different compensation mechanisms and the job-creating impact of product innovation have been tentatively represented and estimated through a simultaneous equations model over the period 1960–1988. Running three stages least squares regressions using Italian and US data, the author has shown that the more effective compensation mechanism is that 'via decrease in prices' in both countries, while other mechanisms reveal to be less important. In addition, the US economy has turned out to be more product-oriented (and so characterized by a positive relationship between technology and employment) than the Italian economy where different compensation mechanisms cannot counterbalance the labor-saving effect of prevailing process innovation.

Finally, Simonetti et al. (2000) have applied the same simultaneous equations macroeconomic model, running three stages least squares regressions using American, Italian, French, and Japanese data over the period 1965–1993. The authors have showed that the more effective compensation mechanisms were that 'via decrease in prices' and that 'via increase in incomes' (especially in European countries till the mid-1980s). The other mechanisms revealed to be less significant and conditional on the institutional structures of the different countries; for instance the 'mechanism via decrease in wages' turned out to be relevant in the American flexible labor market. Finally, product innovation significantly revealed its labor-intensive potentiality only in

the technological leader country in the period, namely the US.

Sectoral Empirical Evidence

The sectoral dimension is particularly important in investigating the overall employment impact of innovation; in particular, the compensation mechanism 'via new product' (which in recent times often takes the form of a compensation 'via new services') may accelerate the secular shift from manufacturing to services (see Evangelista and Perani, 1998; Evangelista and Savona, 1998).

On the other hand, in manufacturing new technologies seems to be characterized mainly by labor-saving ETC, only partially compensated by the market mechanisms discussed in Section The Theory.

For instance, Clark (1983, 1987) put forward a supply oriented vintage model investigating UK manufacturing. The author found out that the *expansionary effect* of innovative investments (Keynesian multiplier) had been dominant until the mid-1960s, when the *rationalizing effect* (due to labor-saving ETC incorporated in investments and scrapping) started to overcome the expansionary one.

More recently, Pianta et al. (1996), found out an overall positive relationship between growth in value added and growth in employment; nevertheless, especially in European countries, an important group of sectors appeared characterized by a marked labor-saving trajectory (restructuring sectors), with growth in production and a fall in employment.

In another study based on Italian data, Vivarelli et al. (1996) showed that in Italian manufacturing the relationship between productivity growth and employment appeared to be negative and, in particular, that product and process innovation had opposite effects on the demand for labor, in line with whatever has been discussed in the previous sections.

As already mentioned, the scenario may change if services sectors are taken into account: for instance – using standardized sectoral data derived from national CISs – Pianta (2000) and Antonucci and Pianta (2002) found an overall negative impact of innovation on employment in manufacturing industries across five European countries, while – in contrast – Evangelista (2000) and Evangelista and Savona (2002) found a positive employment effect of technological change in the most innovative and knowledge-intensive service sectors. (While a negative one in the case of financial-related sectors and most traditional services like trade and transport.)

Taking manufacturing and services jointly into account (using CIS cross-sectional sectoral data on relevant innovations for different European countries), Bogliacino and Pianta (2010) found a positive employment impact of product innovation (which turns out particularly obvious in the high-tech manufacturing sectors, see Mastrostefano and Pianta, 2009). Finally, Bogliacino and Vivarelli (2012) – running systemic generalized method of moments (GMM-SYS) panel estimations covering 25 manufacturing and service sectors for 15 European countries over the period 1996–2005 – found that R&D expenditures – mainly fostering product innovation – do show a job-creating effect. (As in the previous study, the labor-friendly impact of R&D expenditures turns out to be particularly significant in the high-tech sectors.)

Microeconomic Empirical Evidence

While the macroeconomic studies about the relationship between technology and employment seem to be confined to the 1980s and1990s, the literature devoted to the microeconometric investigation of the issue has begun later, but it is nowadays flourishing.

Starting from the earlier contributions, Entorf and Pohlmeier (1990) found a positive employment impact of product innovation – measured with a dummy – in a cross section of 2276 West German firms in 1984. The positive employment impact of product innovation in West German manufacturing was confirmed by Smolny (1998) using a panel of 2405 firms for the period 1980–1992.

Using the 1984 British Workplace Industrial Relations Survey, both Machin and Wadhwani (1991) and Blanchflower et al. (1991) found out a negative raw correlation between ICT adoption and employment; however, once controlled for workplace characteristics and fixed effects, the same correlation turned out to be positive.

In contrast with the previous studies, Brouwer et al. (1993) – using a cross-section of 859 Dutch manufacturing firms – discovered an aggregate negative relationship between aggregate R&D expenditures and employment (while the opposite emerged when only product innovation was considered). By the same token, Zimmermann (1991) – using microdata from 16 German industries – concluded that technological change was one of the determinants of the employment decrease in Germany during the 1980s.

Although the employment impact of innovation is not the main object of Doms et al. (1997), the authors found that advanced manufacturing technologies – measured by a set of dummy variables – have implied higher employment growth in U.S. manufacturing plants over the period 1987–1991.

More controversial results came from Klette and Førre (1998). The authors' database comprised 4333 Norwegian manufacturing plants over the period 1982–1992; in contrast with most of the other studies, they did not find any clear-cut positive relationship between net job creation and the R&D intensity of the examined plants.

Most recent studies fully took the advantage of new available longitudinal datasets and applied more sophisticated panel data econometric methodologies.

For example, Van Reenen (1997) matched the London Stock Exchange database of manufacturing firms with the SPRU innovation database and obtained a panel of 598 firms over the period 1976–1982. Running GMM-DIF estimates, the author found a positive employment impact of innovation and this result turned out to be robust after controlling for fixed effects, dynamics, and endogeneity.

By the same token, Blanchflower and Burgess (1998) confirmed a positive link between innovation (roughly measured with a dummy) and employment using two different panels of British and Australian establishments; their results showed to be robust after controlling for sectoral fixed effects, size of firm, and union density.

An interesting panel analysis was conducted by Greenan and Guellec (2000), using microdata from 15 186 French manufacturing firms over the 1986–1990 period. According to this study, innovating firms – defined according to the outcomes of an innovation survey – create more jobs than noninnovating ones, but the reverse is true at the sectoral level, where the overall effect is negative and only product innovation is job-creating. Interestingly enough, the opposite employment impact of innovation at the firm and sectoral level may be due to the 'business stealing effect' discussed above.

However, even when taking the business stealing effect into account (that is, controlling for lagged firms' employment and current sales). Piva and Vivarelli (2004, 2005) found evidence in favor of a positive effect of innovation on employment at the firm level. In particular – applying a GMM-SYS methodology to a longitudinal dataset of 575 Italian manufacturing firms over the period 1992–1997 – the authors provided evidence of a significant, although small in magnitude, positive link between firm's gross innovative investment and employment.

Using firm level data (obtained from the third wave of the Community Innovation Survey, CIS) from four European countries (Germany, France, UK, Spain), Harrison et al. (2008) put forward a testable model – based on Peters (2004) – able to distinguish the relative employment impact of process and product innovation (discrete variables). The authors concluded that process innovation tends to displace employment, while product innovation is basically labor-friendly. However, compensation mechanisms (see Section Introduction) are at work and in the service sectors are particularly effective through the increase in the demand for the new products. Hall et al. (2008) applied a similar model to a panel of Italian manufacturing firms over the period 1995–2003 and found a positive employment contribution of product innovation and no evidence of employment displacement due to process innovation.

More recently, Lachenmaier and Rottmann (2011) put forward a dynamic employment equation including wages, gross value added, years, and industries controls and alternative proxies (dummies) of current and lagged product and process innovation. Their GMM-SYS estimates – based on a very comprehensive dataset of German manufacturing firms over the period 1982–2002 – show a significantly positive impact of different innovation measures on employment. Partially in contrast with previous contributions, the authors found out a higher positive impact of process rather than product innovation. (However, this result can be due to the discrete nature of the adopted measure of process and product innovation (dummy variables). Interestingly enough, once the authors restrict their attention to (important) product innovation which went along with patent applications, they found out a highly positive and significant employment effect.)

Previous literature has very rarely split the empirical analysis according to sectoral belonging. One of the exceptions is the contribution by Greenhalgh et al. (2001), developing fixed effects estimates based on a panel of UK firms over the period 1987–1994. Consistently with most of the other studies here discussed, the authors found out a positive – albeit modest – effect of R&D expenditures on employment. However, once analyzed the data according to firms' sectoral belonging, the positive impact of R&D on employment turned out to be limited to the sole high-tech sectors.

In contrast – once split the sectors into high-tech and nonhigh-tech ones – Lachenmaier and Rottmann (2011) did not find any significant sectoral heterogeneity in the effects that innovation has on employment.

In a very recent study, Coad and Rao (2011) limited their focus on US high-tech manufacturing industries over the period 1963–2002 and investigated the impact of a composite innovativeness index (comprising information on both R&D and patents) on employment. The main outcome of their quantile regressions is that innovation and employment are positively linked and that innovation has a stronger impact for those firms that reveal the fastest employment growth.

Finally, Bogliacino et al. (2011, 2012) – using a longitudinal database covering 677 European manufacturing and service firms over the period 1990–2008 – found that a positive and significant employment impact of R&D expenditures is detectable in services and high-tech manufacturing but not in the more traditional manufacturing sectors.

On the whole – although previous microeconometric evidence is not fully conclusive about the possible employment impact of innovation – most of recent panel investigations tend to support a positive link, especially when R&D and/or product innovation are adopted as proxies of technological change and when high-tech sectors are focused on.

Conclusions

Preliminary conclusions have been already discussed within the previous sections; here it will be sufficient to briefly summarize the main findings of this study.

- Process innovation tend to be labor-saving, while product innovation exhibit a labor-friendly nature.
- According to the classical 'compensation theory,' market forces should assure a complete compensation of the initial labor-saving impact of process innovations; however, in Section The Theory different possible hindrances to the various compensation mechanisms have been critically discussed, reaching the general conclusion that compensation may be only partial, depending on the particular institutional setting and on the specific values of crucial parameters such as demand elasticity, degree of competition, capital-labor substitution, demand expectations, and so on.
- Since economic theory does not have a clear-cut answer about the employment effect of innovation, there is a strong need for aggregate, sectoral and microeconomic empirical analyses able to take into account the different forms of technological change, their direct effects on labor, the various compensation mechanisms, and the possible hindrances to these mechanisms.
- In particular, microeconometric studies have the great advantage to allow a direct and precise firm's level mapping of innovation variables, both in terms of inputs and outputs. Indeed, only firm's level data can account for R&D

expenditures, capital formation, product and process innovation, and directly relate them to firm's employment trends.

- In general, the microeconometric literature and particularly the most recent panel data analyses, tend to support a positive link between technology and employment, especially when R&D and/or product innovation are adopted as proxies of technological change and when high-tech sectors are focused on.

See also: Education and Employment; Employment and Labor, Regulation of; Innovation and Technological Change, Economics of; Innovation, Theory of; Innovation; Research and Development in Organizations; Unemployment: Structural.

Bibliography

Antonucci, T., Pianta, M., 2002. Employment effects of product and process innovation in Europe. International Review of Applied Economics 16, 295–307.

Appelbaum, E., Schettkat, R., 1995. Employment and productivity in industrialized economies. International Labour Review 134, 605–623.

Blanchflower, D., Burgess, S.M., 1998. New technology and jobs: comparative evidence from a two-country study. Economics of Innovation and New Technology 5, 109–138.

Blanchflower, D., Millward, N., Oswald, A., 1991. Unionisation and employment behaviour. Economic Journal 101, 815–834.

Bogliacino, F., Pianta, M., 2010. Innovation and employment. A reinvestigation using revised Pavitt classes. Research Policy 39, 799–809.

Bogliacino, F., Vivarelli, M., 2012. The job creation effect of R&D expenditures. Australian Economic Papers 51, 96–113.

Bogliacino, F., Piva, M., Vivarelli, M., 2011. R&D and Employment: Some Evidence from European Microdata. IZA Discussion Papers, 5908, Bonn.

Bogliacino, F., Piva, M., Vivarelli, M., 2012. R&D and employment: an application of the LSDVC estimator using european data. Economics Letters 116, 56–59.

Boyer, R., 1988a. New technologies and employment in the 1980s: from science and technology to macroeconomic modelling. In: Kregel, J.A., Matzner, E., Roncaglia, A. (Eds.), Barriers to Full Employment. Macmillan, London, pp. 233–268.

Boyer, R., 1988b. Assessing the impact of R&D on employment: puzzle or consensus?. In: Paper Presented at the International Conference on New Technology: Its Impacts on Labour Markets and the Employment System, 5–7 December Berlin.

Boyer, R., 1988c. Technical change and the theory of regulation. In: Dosi, G., Freeman, C., Nelson, R.R., Silverberg, G., Soete, L. (Eds.), Technical Change and the Economic Theory. Pinter, London, pp. 67–94.

Boyer, R., 1988d. Formalizing growth regimes. In: Dosi, G., Freeman, C., Nelson, R.R., Silverberg, G., Soete, L. (Eds.), Technical Change and the Economic Theory. Pinter, London, pp. 608–635.

Boyer, R., 1990. The Capital Labor Relations in OECD Countries: from the Fordist "Golden Age" to Contrasted National Trajectories. Working paper CEPREMAP, 9020, Paris.

Brouwer, E., Kleinknecht, A., Reijnen, J.O.N., 1993. Employment growth and innovation at the firm level: an empirical study. Journal of Evolutionary Economics 3, 153–159.

Clark, J.B., 1907. Essentials of Economic Theory. Macmillan, New York.

Clark, J., 1983. Employment projections and technological change. In: Bosworth, D.L. (Ed.), The Employment Consequences of Technical Change. Macmillan, London, pp. 110–125.

Clark, J., 1987. A vintage-capital simulation model. In: Freeman, C., Soete, L. (Eds.), Technical Change and Full Employment. Basil Blackwell, Oxford, pp. 86–98.

Coad, A., Rao, R., 2011. The firm-level employment effects of innovations in high-tech US manufacturing industries. Journal of Evolutionary Economics 21, 255–283.

Conte, A., Vivarelli, M., 2005. One or Many Knowledge Production Functions? Mapping Innovative Activity Using Microdata. IZA Discussion Paper, 1878, Bonn.

Dobbs, I.M., Hill, M.B., Waterson, M., 1987. Industrial structure and the employment consequences of technical change. Oxford Economic Papers 39, 552–567.

Doms, M., Dunne, T., Troske, K., 1997. Workers, wages, and technology. Quarterly Journal of Economics 112, 253–289.

Dosi, G., 1982. Technological paradigms and technological trajectories. Research Policy 11, 147–163.

Dosi, G., 1988. Source, procedure and microeconomic effects of innovation. Journal of Economic Literature 26, 1120–1171.

Douglas, P.H., 1930. Technological unemployment. American Federationist 37, 923–950.

Dunne, M., Haltiwanger, J., Troske, K., 1997. Technology and jobs: secular changes and cyclical dynamics. The Carnegie-Rochester Conference Series on Public Policy 46, 107–178.

Edquist, C., Hommen, L., McKelvey, M., 2001. Innovation and Employment: Product versus Process Innovation. Elgar, Cheltenham.

Entorf, H., Pohlmeier, W., 1990. Employment, innovation and export activities. In: Florens, J.P., Ivaldi, M., Laffont, J.J., Laisney, F. (Eds.), Microeconometrics: Surveys and Applications. Basil Blackwell, London, pp. 394–415.

Evangelista, R., 2000. Innovation and employment in services. In: Vivarelli, M., Pianta, M. (Eds.), The Employment Impact of Innovation: Evidence and Policy. Routledge, London, pp. 121–148.

Evangelista, R., Perani, G., 1998. Innovation and employment in services: results from the Italian Innovation Survey. In: Paper Presented at the EAEPE Conference, 5–8 November, Lisbon.

Evangelista, R., Savona, M., 1998. Patterns of innovation in services. The results of the Italian Innovation Survey. In: Paper Presented at the VIII Annual RESER Conference, 8–10 October, Berlin.

Evangelista, R., Savona, M., 2002. The impact of innovation on employment in services: evidence from Italy. International Review of Applied Economics 16, 309–318.

Freeman, C., Clark, J., Soete, L., 1982. Unemployment and Technical Innovation. Pinter, London.

Freeman, C., Soete, L. (Eds.), 1987. Technical Change and Full Employment. Basil Blackwell, Oxford.

Freeman, C., Soete, L., 1994. Work for All or Mass Unemployment? Computerised Technical Change into the Twenty-First Century. Pinter, London-New York.

Greenan, N., Guellec, D., 2000. Technological innovation and employment reallocation. Labour 14, 547–590.

Greenhalgh, C., Longland, M., Bosworth, D., 2001. Technological activity and employment in a panel of UK firms. Scottish Journal of Political Economy 48, 260–282.

Hall, P.H., Heffernan, S.A., 1985. More on the employment effects of innovation. Journal of Development Economics 17, 151–162.

Hall, B.H., Lotti, F., Mairesse, J., 2008. Employment, innovation, and productivity: evidence from Italian microdata. Industrial and Corporate Change 17, 813–839.

Harrison, R., Jaumandreu, J., Mairesse, J., Peters, B., 2008. Does Innovation Stimulate Employment? A Firm-Level Analysis Using Comparable Microdata from Four European Countries. NBER Working Papers, 14216, Cambridge, MA.

Heffernan, S.A., 1981. Technological Unemployment. Unpublished Ph.D. Thesis. Oxford University, Oxford.

Hicks, J.R., 1932. The Theory of Wages. Macmillan, London.

Hicks, J.R., 1973. Capital and Time. Oxford University Press, Oxford.

Hobsbawm, E.J., 1968. Industry and Empire: An Economic History of Britain since 1750. Penguin Books, Harmondsworth Middlesex.

Hobsbawm, E.J., Rudé, G., 1969. Captain Swing. London Reader's Union, London.

Katsoulacos, Y.S., 1984. Product innovation and employment. European Economic Review 26, 83–108.

Katsoulacos, Y.S., 1986. The Employment Effect of Technical Change. Wheatsheaf, Brighton.

Keynes, J.M., 1973. The general theory of employment, interest and money. In: The Collected Writings of John Maynard Keynes, first ed. 1936. Macmillan, London.

Klette, T.J., Førre, S.E., 1998. Innovation and job creation in a small open economy: evidence from Norwegian manufacturing plants 1982–92. Economics of Innovation and New Technology 5, 247–272.

Lachenmaier, S., Rottmann, H., 2011. Effects of innovation on employment: a dynamic panel analysis. International Journal of Industrial Organization 29, 210–220.

Layard, R., Nickell, S., 1985. The causes of British unemployment. National Institute Economic Review 111, 62–85.

Layard, R., Nickell, S., Jackman, R., 1991. Unemployment: Macroeconomic Performance and the Labour Market. Oxford University Press, Oxford.

Layard, R., Nickell, S., Jackman, R., 1994. The Unemployment Crisis. Oxford University Press, Oxford.

Machin, S., Wadhwani, S., 1991. The effects of unions on organisational change and employment: evidence from WIRS. Economic Journal 101, 324–330.

Malthus, T.R., 1964. Principles of Political Economy, first ed. 1836. M. Kelley, New York.

Marshall, A., 1961. Principles of Economics, first ed. 1890. Macmillan, Cambridge.

Marx, K., 1961. Capital, first ed. 1867. Foreign Languages Publishing House, Moscow.

Marx, K., 1969. Theories of Surplus Value, first ed. 1905–10. Lawrence & Wishart, London.

Mastrostefano, V., Pianta, M., 2009. Technology and jobs. Economics of Innovation and New Technology 18, 729–741.

Neary, J.P., 1981. On the short-run effects of technological progress. Oxford Economic Papers 32, 224–233.
Nickell, S., Kong, P., 1989. Technical Progress and Jobs. Centre for Labour Economics. Discussion Paper 366. London School of Economics, London.
Parisi, M.L., Schiantarelli, F., Sembenelli, A., 2006. Productivity, innovation and R&D: microevidence for Italy. European Economic Review 50, 2037–2061.
Pasinetti, L., 1981. Structural Change and Economic Growth. Cambridge University Press, Cambridge.
Pavitt, K., 1984. Patterns of technical change: towards a taxonomy and a theory. Research Policy 13, 343–374.
Peters, B., 2004. Employment Effects of Different Innovation Activities: Microeconometric Evidence. ZEW Discussion Papers, 0473, Bonn.
Petit, P., 1995. Employment and technological change. In: Stoneman, P. (Ed.), Handbook of the Economics of Innovation and Technological Change. North Holland, Amsterdam, pp. 366–408.
Pianta, M., 2000. The employment impact of product and process innovations. In: Vivarelli, M., Pianta, M. (Eds.), The Employment Impact of Innovation: Evidence and Policy. Routledge, London, pp. 77–95.
Pianta, M., 2005. Innovation and employment. In: Fagerberg, J., Mowery, D., Nelson, R.R. (Eds.), Handbook of Innovation. Oxford University Press, Oxford, pp. 568–598.
Pianta, M., Evangelista, R., Perani, G., 1996. The dynamics of innovation and employment: an international comparison. STI Review 18, 67–93.
Pigou, A., 1933. The Theory of Unemployment. Macmillan, London.
Pigou, A., 1962. The Economics of Welfare, first ed. 1920. Macmillan, London.
Piva, M., Vivarelli, M., 2004. Technological change and employment: some microevidence from Italy. Applied Economics Letters 11, 373–376.
Piva, M., Vivarelli, M., 2005. Innovation and employment: evidence from Italian microdata. Journal of Economics 86, 65–83.
Ricardo, D., 1951. Principles of political economy. In: Sraffa, P. (Ed.), The Works and Correspondence of David Ricardo, third ed., vol. 1. Cambridge University Press, Cambridge. 1821.
Rifkin, J., 1995. The End of Work. The Decline of the Global Labour Force and the Dawn of the Post-market Era. G.P. Putnam's Sons, New York.
Robbins, L., 1934. The Great Depression. Macmillan, London.
Rosenberg, N., 1976. Perspectives on Technology. Cambridge University Press, Cambridge.
Salter, W.E.G., 1960. Productivity and Technical Change. Cambridge University Press, Cambridge.
Say, J.B., 1964. A Treatise on Political Economy or the Production, Distribution and Consumption of Wealth, first ed. 1803. M. Kelley, New York.
Simonetti, R., Taylor, K., Vivarelli, M., 2000. Modelling the employment impact of innovation. In: Vivarelli, M., Pianta, M. (Eds.), The Employment Impact of Innovation: Evidence and Policy. Routledge, London, pp. 26–43.
Sinclair, P.J.N., 1981. When will technical progress destroy jobs? Oxford Economic Papers 31, 1–18.
Sismondi, J.C.L., 1971. Nouveaux Principes d'Economie Politique ou de la Richesse dans ses Rapports avec la Population, first ed. Calmann- Levy, Paris. 1819.
Smolny, W., 1998. Innovations, prices and employment: a theoretical model and an empirical application for West German manufacturing firms. Journal of Industrial Economics 46, 359–381.
Solow, R.M., 1960. Investment and technical progress. In: Arrow, K.J., Karlin, S., Suppes, P. (Eds.), Mathematical Methods in the Social Sciences. Stanford University Press, Stanford, pp. 89–104.
Standing, G., 1984. The notion of technological unemployment. International Labour Review 123, 127–147.
Stiglitz, J.E., 1987. Learning to learn, localized learning and technological progress. In: Dasgupta, P., Stoneman, P. (Eds.), Technology Policy and Economic Performance. Cambridge University Press, Cambridge, pp. 125–153.
Stoneman, P., 1983a. The Economic Analysis of Technological Change. Oxford University Press, Oxford.
Stoneman, P., 1983b. New technology, demand and employment. In: Bosworth, D.L. (Ed.), The Employment Consequences of Technological Change. Macmillan, London, pp. 82–96.
Sylos Labini, P., 1969. Oligopoly and Technical Progress, first ed. 1956. Harvard University Press, Cambridge MA.
Van Reenen, J., 1997. Employment and technological innovation: evidence from U.K. manufacturing firms. Journal of Labor Economics 15, 255–284.
Venables, A.J., 1985. The economic implications of a discrete technical change. Oxford Economic Papers 37, 230–248.
Vivarelli, M., 1995. The Economics of Technology and Employment: Theory and Empirical Evidence. Elgar, Aldershot.
Vivarelli, M., Pianta, M., 2000. The Employment Impact of Innovation: Evidence and Policy. Routledge, London.
Vivarelli, M., Evangelista, R., Pianta, M., 1996. Innovation and employment in Italian manufacturing industry. Research Policy 25, 1013–1026.
Wicksell, K., 1961. Lectures on Political Economy, first ed. 1901–1906. Routledge & Kegan, London.
Zimmermann, K., 1991. The employment consequences of technological advance: demand and labour costs in 16 German industries. Empirical Economics 16, 253–266.

Innovation and Technological Change, Economics of

Wesley M Cohen, Duke University, Durham, NC, USA

Published by Elsevier Ltd.

Abstract

We review empirical findings on the determination of R&D and R&D performance, as well as key questions that remain unaddressed. This article briefly considers the Schumpeterian hypotheses relating innovation to market structure and firm size, and considers in more depth the role of firm characteristics and industry-level variables – broadly characterized as reflecting demand, technological opportunity, and demand conditions – in affecting firms' innovative activity and performance.

Introduction

The writings of Joseph Schumpeter in the first half of the century pushed economists to appreciate the fundamental role of technological progress in affecting economic growth and social welfare. Since that time, economists have increasingly appreciated the economic significance of technological progress, and it is now common to hear that a firm's, an industry's, or even a nation's capacity to progress technologically underpins its long-run economic performance. Stimulated by Schumpeter's writings and Solow's (1957) subsequent 'discovery' of the contribution of technological change to economic growth, economists have conducted numerous empirical studies on the determination of innovative activity and performance. In this article we review the empirical literature and highlight the questions addressed and progress in the field.

The section 'Empirical Studies in the Schumpeterian Tradition' briefly reviews the 'neo-Schumpeterian' literature that examines the effects of firm size and market concentration upon innovation. The section 'Firm Characteristics' discusses the literature that considers the effect on innovation of firm characteristics other than size. The section 'Industry Characteristics' covers the literature on the three classes of factors affecting interindustry variation in innovative activity and performance: demand, appropriability, and technological opportunity conditions. We conclude in the section 'Conclusion.'

Empirical Studies in the Schumpeterian Tradition

This section briefly considers studies on the relationship between market concentration or firm size and innovative effort elicited by Schumpeter's controversial claims about the key role of large, monopolistic firms in advancing technology. (For a comprehensive review of this voluminous literature, see Cohen (2010).)

While Schumpeter (1942) confounded his discussions of the impacts on innovation of firm size and concentration, Galbraith (1952) explicitly argued that large firm size confers an advantage in innovation. A literal reading of Schumpeter's (1942) classic discussion suggests, however, that he was primarily impressed by the qualitative differences between the innovative activities of small, entrepreneurial enterprises and those of large, modern corporations with formal R&D laboratories. Nonetheless, the empirical literature has interpreted Schumpeter's claim for a large firm advantage in innovation as a proposition that innovative activity increases more than proportionately than firm size. With some exceptions (e.g., Scherer, 1965; Nelson et al., 1967; Pavitt et al., 1987) the Schumpeterian hypothesis about firm size has been tested by regressing some measure of innovative activity (input or output) on a measure of size. A thorough review of this literature (Cohen, 2010) highlights one longstanding, robust finding: a monotonic relationship between firm size and R&D. There is reason to believe that the source of this relationship is the R&D cost-spreading incentive effects of firm size (Cohen and Klepper, 1996). This cost-spreading effect of size itself, however, reflects underlying appropriability conditions that confine firms to exploiting their innovations in their own output, typically limited growth due to innovation, and market segmentation.

Schumpeter also argued that the possession of *ex ante* market power, linked to an *ex ante* oligopolistic (or monopolistic) market structure, also favored innovation. The profession's investigation of a relationship between market structure and R&D remains, however, problematic. Notwithstanding Schumpeter's arguments, no clear theoretical rationale for the relationship has emerged. And empirical results remain weak, with measures of market structure – albeit poor ones – explaining little of the variance in R&D spending. More importantly, the consensus of the empirical literature is that, to the extent that there is a relationship between market structure and R&D activity, it likely reflects the influence of other more fundamental determinants of technical advance, specifically technological opportunity and appropriability conditions, as well as the degree of market segmentation, all discussed below.

Firm Characteristics

This section considers firm-level determinants of innovative effort and performance other than firm size. Applied economists often control for firm size in their studies of innovative effort by using R&D intensity as the dependent variable measuring innovative effort. Employing the FTC Line of Business Program data that distinguish between the firm and its constituent business units, Scott (1984) found that once

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71047-6

one controls for size by expressing R&D intensity as R&D normalized by business unit sales, fixed firm effects explain about 50% of the variance in R&D intensity. Economists have made, however, only modest headway in explaining interfirm differences in R&D intensity and performance since Cohen and Levin (1989) observed that the most widely considered measures of firm characteristics up to that time, namely cash flow and the degree of diversification, jointly explained less than 10% of the variance in business unit R&D intensity explained by fixed firm effects.

One firm-specific variable long considered as a possible determinant of R&D is cash flow, which may be the most thoroughly examined firm characteristic in this literature. Indeed, the claim that cash flow affects R&D constitutes one of Schumpeter's (1942) arguments for an advantage to large firm size. The rationale for considering cash flow as a determinant of R&D effort typically assumes that (1) capital markets are imperfect; (2) those imperfections especially constrain investments whose outcomes are more uncertain; and (3) the returns to R&D, as a class of investment expenditure, are more uncertain than the returns to, say, investment in property, plant, and equipment. The possibility that R&D expenditure is liquidity constrained due to capital market imperfections is important for policy makers because it represents another rationale for public support for R&D beyond the widely recognized market failure affecting R&D investment associated with R&D spillovers.

Many, but not all, of the studies have found that a firm's cash flow is associated with higher levels of R&D intensity or unnormalized R&D effort. Scholars have disagreed over the interpretation of this finding. Some have argued that it is difficult to distinguish cash flow as a measure of liquidity from its possible function as a signal of the future profitability of R&D investment. Indeed, the likelihood that R&D and cash flow both respond to the same demand shocks is likely the most challenging of the empirical issues facing researchers on this topic. Others question whether cash flow encourages R&D or whether it simply reflects the profitability of past R&D expenditures, which tend to be stable over time. Moreover, a positive relationship between cash flow and R&D expenditures may simply reflect their codetermination by a firm's underlying capabilities. Nonetheless, while the methodological challenges that plague this literature are not entirely addressed, the weight of findings suggests that cash flow likely contributes to R&D spending (Hall, 2002).

Another widely studied corporate attribute is diversification. The influence of product diversification upon basic research spending was first suggested by Nelson (1959), who argues that, because the results of basic research tend to be unpredictable, the diversified firm possesses more opportunities for exploiting the new knowledge. This argument implicitly assumes what Arrow (1962) expressed clearly: the market for information is imperfect and appropriability may be best achieved by the internal application of knowledge. The empirical results regarding the role of diversification have, however, been mixed (e.g., Link and Long, 1981). The absence of robust findings concerning an effect of diversification on R&D is unsurprising given the problems that beset this research. For example, it is just as likely, if not more, that the direction of the relationship runs opposite to that posited in the early work (Montgomery, 1994). Perhaps the most revealing work on the relationship between R&D and diversification has focused on the relationship between innovative performance and firm diversification across submarkets within one industry, namely drugs. Henderson and Cockburn (1996) observed a significant, robust positive relationship between the number of research programs pursued by drug firms (where a program is distinguished by therapeutic area), and the productivity of each program's R&D, measured by the number of 'important' patents produced, suggesting the presence of scope economies where diverse R&D activities are related across submarkets.

Differences in R&D activities and innovative performance across firms have focused some economists' attention on the influence on R&D of differences in firms' R&D-related capabilities. Before discussing this literature, it is important, however, to take heed of Scott's (1991) point that one should not presume that heterogeneity in firms' R&D activities or performance necessarily signal differences in capabilities rather than differences in the incentives faced by comparably capable firms. Indeed, in Cohen and Klepper's (1996) R&D cost-spreading model, firms of different sizes may pursue different types of R&D not because of differences in capabilities, but due to the different R&D incentives conditioned by size. Obviously, however, differences in capabilities may contribute to differences in the level and direction of R&D effort.

This modest economics literature on the role of firm-specific capabilities in affecting innovation raises the question of exactly what kinds of capabilities matter for innovative activity and performance. The research broadly distinguishes two sorts. Clark, Henderson, Mansfield, Mowery, Nelson, and others emphasize the role of organizational or procedural capabilities in conditioning the R&D productivity of firms. In this view, firms may pursue similar innovative activities, but some firms are more successful than others in either generating or profiting from innovation due to superior organization of internal processes. In contrast, Jewkes et al. (1958), Cohen and Klepper (1992a, b), and others focus on differences in firms' areas of substantive technological or domain expertise which affect the success of their pursuit of different innovative activities.

The firm characteristic that economists have been long believed to affect R&D performance is the link between R&D and other functions within the firm. Starting with the work of Mansfield (1968), economists such as Teece (1986) and Mowery and Rosenberg (1989) have highlighted the importance of the links across marketing, manufacturing, and R&D in conditioning innovative success. Exploring complementarities between R&D and other functions within the firm or across different activities within the R&D function itself poses, however, a challenge for both methods and measurement (cf Arora, 1996). The issue is how can one estimate such complementarities given the possibility that an apparent complementarity across activities or capabilities may rather reflect the effect of unobserved firm heterogeneity; that, for example, there is some unobserved firm characteristic that is driving what is believed to be complementary activities as well as R&D performance itself.

With notable exceptions (e.g., Cockburn and Henderson, 1998; Henderson and Cockburn, 1994), economists have been slow to take up the call to consider the role of capabilities within the R&D function itself. Although one could treat such

capabilities as observable, the construction of measures of R&D-related capabilities suitable for analyzing innovation in large samples spanning multiple industries is a formidable challenge. Construction of such measures demands detailed knowledge of the nature of the technologies and other activities associated with innovation in each sample industry. Moreover, it is not even clear how such capabilities would be defined or distinguished in a broad cross section. Partly reflecting these difficulties and reflecting economists' inattention to organizational processes and the nature of organizations more generally, only modest progress has been made in conceptualizing differences in firm R&D-related capabilities in ways that yield testable, interpretable hypotheses.

A small number of empirical studies on the role of capabilities in determining innovative performance within industries date from the late 1980s and early 1990s. Clark et al. (1987), for example, identify wide disparities in firms' R&D productivities in the international automobile manufacturing industry. This study offers evidence that these disparities result from differences in the organization of product development in these firms, including the degree to which tasks are subdivided, how the different phases of the development process are coordinated, how technical problems get solved, the relative autonomy of the project leader, and the nature of the links between product development and upstream applied research.

Notwithstanding the type of capability that is thought to matter, many studies of firm capabilities assume that key R&D-related capabilities are exogenously given. This literature, however, does not make a strong theoretical case for this assumption, nor does it, with the exception of the limited evidence offered by Henderson (1993), indicate whether or under what circumstances it may be true. Although partly a question of time horizon, case study, and anecdotal evidence support to some degree the assumption that selected R&D-related capabilities are exogenous by highlighting the fact that a firms' 'core' capabilities are hard to change. The question of whether R&D capabilities are endogenous or not suggests the utility of entertaining the fundamental questions raised by Clark (Clark et al., 1987: 781) about why such differences exist, how they evolve, and why they persist.

Related to the observation that firms differ in their innovative activities and capabilities is the notion that no one firm may pursue all the activities necessary to generate and commercialize a given innovation. This observation should push economists to understand that technical advance within an industry is commonly realized through the interactions of firms that are distinguished by size, expertise, and other attributes, and thus one might consider that technical advance is not promoted by some ideal firm type – distinguished by any particular single or set of characteristics – but by the interactions of firms with differing and complementary features, no less by the interactions between firms and other institutions such as universities. Accordingly, it would be important to understand the range and distribution of attributes that might promote technical advance, as well as factors that promote the efficiency of firms' interactions with one another and with other institutions that support technical advance.

Industry Characteristics

In seeking to understand why industries differ in the degree to which they engage in innovative activity, empirical researchers such as Pakes and Schankerman (1984) have come to classify explanatory variables under three headings: product market demand, technological opportunity, and appropriability conditions. The subsections that follow summarize and interpret what is known about how demand, technological opportunity, and appropriability conditions vary across industries and how they contribute to an explanation of interindustry differences in innovative activity.

Demand

In his seminal work on technological change in various capital goods industries, Schmookler (1962, 1966) made two arguments. The first, general claim was that the direction and rate of technological change could be explained as the outcome of the behavior of purposive, profit-seeking firms, and was thus amenable to economic analysis. To demonstrate the importance of economic incentives in general, he highlighted the role of demand in particular. To substantiate that demand determined the rate and direction of inventive activity, he marshaled evidence indicating that cyclical changes in the output of capital goods and in capital expenditures by downstream industries 'led' cycles in the time series of relevant capital goods patents.

Schmookler's focus on the role of demand sparked a lively debate among economic historians and other economists concerning whether 'demand-pull' or 'technology-push' was the primary force behind technological change. Schmookler's proposition that demand almost alone determines the rate and direction of technical change has not, however, survived empirical scrutiny. The consensus, after dozens of case studies, is that the Marshallian scissors cuts with two blades. Perhaps the most persuasive refutation of Schmookler's proposition is offered by Parker (1972) and Rosenberg (1974), who document several important historical examples (e.g., the mechanization of hand operations in agriculture, the use of coal as an industrial fuel) in which the sequence of particular applications of a 'generic' technological idea was determined not by demand, but by the state of knowledge and inherent technological complexity of particular industrial applications. Offering statistical evidence that both blades matter, Stoneman (1979) showed that the cost of, as well as the demand for inventions conditioned the level of innovative effort. Scherer (1982) offered additional evidence, finding that dummy variables classifying industries by technology (chemical, electrical, mechanical, etc.) and variables representing demand conditions were statistically significant in a regression analysis of line of business patenting activity, but the technology variables explained considerably more variance.

There are two principal respects in which interindustry differences in demand conditions might be expected to affect the incentives to engage in innovative activity. First, as Schmookler himself emphasized, there is the size of the market, which might be represented in static terms by a scale parameter and in dynamic terms by a rate of growth. The argument is straightforward. The (expected) investment required to yield

a given reduction in unit cost or a given improvement in product quality is independent of the level of output that will be produced once the innovation is made. The benefits realized by such investment, however, are proportional to the size of the market in which the innovation is used. More inventive activity would therefore be expected in the larger of two markets, holding constant the cost of innovation; in two markets of equal size, more inventive activity would be expected in the market that is expected to grow more rapidly.

Second, Kamien and Schwartz (1970) suggested that the price elasticity of demand will also affect the marginal returns to investment in R&D. They demonstrated that the gains from reducing the cost of production (process innovation) are larger the more elastic is demand. On the other hand, Spence (1975) demonstrates that the gains from improvement in product quality (product innovation) will, under many circumstances, be larger the more inelastic is demand, since inelastic demand tends to magnify the gains from a rightward shift in the demand curve. Thus, the effect of price elasticity will be ambiguous in empirical studies that do not distinguish between process and product innovation.

The distinction between process and product innovation raises the subtle conceptual and operational question of how to characterize the demand conditions relevant to product innovation. In the case of intermediate products, such as those studied by Schmookler, the demand function for inputs of higher quality can in principle be derived from estimates of final product demand and the downstream production technology. It is more difficult to characterize and estimate the demand for consumer product innovation. A variety of econometric techniques can be used to estimate the demand for routine improvements in some measurable dimension of product quality (e.g., hedonic price models, Lancasterian demand models, and discrete choice models). Such techniques, though useful in particular applications, are unlikely to be fruitful in cross-industry analysis, since they require very detailed data and special modeling efforts for each specific product.

In regression studies of R&D investment, demand conditions, although rarely featured, have often been considered. A variety of categorical variables have been used, presumably as proxies for interindustry differences in price elasticity. Most common are those distinguishing durables from nondurables, as well as those distinguishing consumer goods from material inputs or investment goods. Some researchers have used input–output data on the disposition of industry output (i.e., the shares of output destined for personal consumption, intermediate use, exports, the government sector, etc.). Although these categorical and input–output variables are sometimes statistically significant in regressions explaining a measure of innovative activity, there are no notably robust findings.

To capture market size and growth effects, sales and the rate of growth of sales are typically used, despite the obvious problem that these variables do not measure demand conditions, but the interaction of demand and supply. Moreover, since improved products will have larger markets, sales revenue and growth are endogenous with respect to innovation. Measures of market size that are defensibly exogenous and stand apart from supply conditions have just recently been incorporated into a study of innovation and R&D, though only for the pharmaceutical industry, by Acemoglu and Linn (2004). With a dependent variable being the number of new drug approvals by the US FDA in broad categories distinguished by therapeutic effects, the exogenous component of potential market size is derived from Current Population Survey data on demographic trends combined with differences in the population age profiles of expenditures and use for different types of drugs.

Sutton (1998) highlights submarket homogeneity as another dimension of demand that conditions R&D spending. In his model, submarket heterogeneity, and the interaction between the nature of products and the buyer preferences that give rise to it, condition the degree to which firm size (or, more properly, business unit size) offers an R&D cost-spreading advantage, and, thus, the degree to which firm size increases the incentive to conduct R&D. One might argue that submarket heterogeneity simply begs the question of how one defines a market, though, as discussed at length by Sutton, market definitions need to be empirically implementable.

A limitation of economists' consideration of demand is the assumption, highlighted by Malerba (2007), that buyers are passive participants in the innovation process. This is far from true. von Hippel, 1988 shows, for example, that in semiconductors and scientific instruments, some customers – what he calls 'lead users' – may not only suggest ideas, but may themselves act on them, modifying firms' products, and, in von Hippel's view, anticipate or even shape the future direction of product innovation. Economists would surely benefit from a more detailed institutional understanding of the role of buyers in affecting the rate and direction of technical advance.

Technological Opportunity

Much of the empirical literature takes for granted that technical advance, at prevailing input prices, is 'easier' (less costly) in some industries than in others. Although it is widely accepted that industries differ in the opportunities they face for technical advance, there is no consensus on how to make the concept of technological opportunity precise and empirically operational. In the framework of the standard neo-classical theory of production, technological opportunity can be regarded as the set of production possibilities for translating research resources into new techniques of production that employ conventional inputs. Some theoretical treatments have thus represented technological opportunity as one or more parameters in a production function relating research resources to increments in the stock of knowledge, with the stock of knowledge entering as an argument, along with conventional inputs, in the production function for output (e.g., Griliches, 1979). Related approaches treat technological opportunity as the elasticity of unit cost or unit price–cost margin with respect to R&D spending (Dasgupta and Stiglitz, 1980; Spence, 1984; Sutton, 1998), or as a shift parameter determining the location of an innovation possibility frontier representing the trade-offs in the direction of technical change (Levin, 1978).

Rosenberg (1974) suggested a simple mechanism by which the growth in scientific knowledge encourages innovation; he

claimed that "as scientific knowledge grows, the cost of successfully undertaking any given, science-based invention declines …" (p. 107). Conceptualizing R&D as a stochastic search process, Evenson and Kislev (1976) and Nelson (1982) similarly suggested that 'strong' science affects the cost of innovation by increasing the productivity of applied research. Nelson in particular argued that a strong science base narrows the set of research options and focuses attention on the most productive approaches. The consequence is that the research process is more efficient. There is less trial and error; fewer approaches need to be evaluated and pursued to achieve a given technological end. From this perspective, the contribution of science is that it provides a powerful heuristic guiding the search process associated with technological change. Evenson and Kislev (1976) and Nelson (1982) enrich this characterization of the role of science by suggesting that stronger science may also provide a larger pool of candidate approaches to achieving some technological objective, thereby increasing the expected payoff, although at the cost of broadening the search. Cohen and Klepper (1992b) suggest that a more vital underlying science and technology may also increase the number of technological objectives to be pursued rather than simply increase the number of approaches to pursuing any given objective.

Most attempts to represent technological opportunity as a determinant of innovative activity in regression studies have followed the practice introduced by Scherer (1965), who classified industries on the basis of the scientific or technological field with which each was most closely associated. Scherer's initial classificatory scheme (chemical, electrical, mechanical) was refined in his subsequent work (1967, 1982), and variants have been used by numerous investigators. Although Scherer intended to capture interindustry differences in the vigor of advance of underlying scientific and technological knowledge, he recognized that statistical results obtained with the use of such crudely defined categorical variables would also likely reflect the influence of unspecified industry practices or demand effects not captured by other regressors. Nonetheless, the simple classification of industries into a small number of technology groups has powerful statistical consequences; it has explained a substantial fraction of variance in patenting activity (Scherer, 1965, 1982) and R&D intensity (Scott, 1984).

Using the survey data described in Levin et al. (1987), Klevorick et al. (1995) examine across a broad range of manufacturing industries the general relevance of 11 fields of basic and applied science to each industry's technological progress, as well as the relevance of university-based research for those same fields. Klevorick et al. find that the reported overall relevance of science to technical advance varies considerably across industries and identify the industries to which science appears to be most relevant. For science in general, they found drugs and semiconductors closest to a field of science among the industries for which they received 10 or more responses. While these survey results represent an important first step in examining an important dimension of technological opportunity, empirical economists need to build on these initial efforts and develop data – survey and other data – that are more readily interpretable and speak more directly to some of our theoretical notions of the role of science in promoting technical advance.

Linked to the progress of underlying science and technology that may affect an industry's technological opportunity is the contribution of technical knowledge from sources external to the industry: suppliers, customers, universities, technical societies, government, and independent inventors. A voluminous institutional literature documents the contribution of such extraindustry sources to technological progress. The case studies of Jewkes et al. (1958) contain instances of virtually every type of external influence. A notable example of institutional–empirical work on this subject is von Hippel's (1988) treatment of the contributions of users to technological development in a variety of industries, including scientific instruments and semiconductor process equipment. Klevorick et al. (1995) offer the first broad, cross-industry empirical examination of the contributions to technical advance made by extraindustry sources of embodied and disembodied knowledge. Not surprisingly, they found that what they call sources 'within the industrial chain,' such as buyers and materials and equipment suppliers, contribute more to most industries' technical advance than nonindustrial sources such as universities and government labs.

A key source of technological opportunities that reflects the influence that innovation in one industry may have on downstream and complementary technologies are 'general purpose technologies' – a term coined by Bresnahan and Trajtenberg (1995) to characterize technologies that may enable and stimulate applications across a range of markets (possessed of what they call, 'generality of purpose'), and whose prospective applications typically elicit follow-on R&D, thus exhibiting 'innovational complementarities.' In other words, there are technologies that, once developed, provide significant, and often pervasive technological opportunities for innovation in other industries.

An extensively studied extraindustry influence on technological opportunity is that of government. In numerous sectors, such as agriculture, aircraft, electronics, and medicine, government has contributed to reducing the cost of innovation and has influenced the direction of industrial research by its own research, as well as by both subsidizing and sponsoring private sector research and disseminating technological knowledge developed in its own labs and elsewhere. The distribution of government expenditures on R&D across industries is highly skewed, especially in the United States, where industries supplying the military, and, more recently, universities conducting research in the life sciences, are the principal recipients of R&D support. Although its direct role in creating and disseminating knowledge is substantial in some sectors, its indirect influence is also felt through a variety of other channels that have different impacts across industries. Most important is the impact of government demand on the rate and direction of innovation.

Over the past 30 years, scholars have also studied intensively the effect of university research on industrial innovation. Nelson (1986) and Klevorick et al. (1995), for example, find university research to be an important source of innovation in only selected industries, including especially those associated

with biology. In contrast, also relying upon survey data, but collected over a decade later, Cohen et al. (2002b) found the impact of public research (including university and government sources) to be more pervasive. Using state-level patent, innovation, and patent citation data, respectively, Jaffe (1989b), Acs et al. (1992), and Jaffe et al. (1993) also show that the effects of university research on innovation increase with geographic proximity.

Just as spillovers from extraindustry sources may augment a recipient firm's technological opportunity; so may spillovers within an industry reduce the own R&D required to achieve a given level of technical performance. There have been several econometric attempts to measure the effects of both extra-industry and intra-industry spillovers on firms' productivity, including efforts by Jaffe (1986, 1988, 1989a) using data on the distribution of patents by patent class to measure the technological relatedness of every pair of firms in a sample of over 500 firms. For each firm, he constructed a 'spillover pool,' defined as the sum of all other firms' R&D weighted by the measure of relatedness. Bernstein and Nadiri (1988, 1989) took a more direct approach to estimating the magnitude of spillover effects by including the R&D capital of other firms or industries in the cost function of the receiving firm or industry. They found evidence of large efficiency gains from both intra-industry and extraindustry spillovers. On the basis of these and other studies, Griliches (1992), in a review of the econometric literature on R&D spillovers, concluded that, "R&D spillovers are present, their magnitude may be quite large, and social rates of return remain significantly above private rates" (p. 24).

Thus, most of the work on technological opportunity over the past decade has focused on particular sources – notably that emerging from public research and other firms. With the exception of Klevorick et al. (1995), there is a paucity of work that probes the nature and sources of technological opportunity more generally, or that operationalizes the concept and examines its role across industries, despite the fact that a number of prominent scholars consider technological opportunity to be not only a key determinant of R&D and technical advance, but foundational to the evolution of market structure and entry, as well as key to the links across market structure, entry, and innovation (e.g., Geroski, 1994; Sutton, 1998).

Appropriability

To the extent that new knowledge is transmitted at a relatively low cost from its creator to prospective competitors, and particularly to the extent that such knowledge is embodied in new processes and products that may be copied or imitated at relatively low cost, appropriable rewards may be insufficient to justify innovative effort. In theory, patents solve the problem of imperfect appropriability; the exclusive right granted by society enhances the incentive to invent by sanctioning restriction of an invention's use. To the would-be inventor, the prospect of a patent represents the expectation of ex-postmarket power that Schumpeter claimed was an essential spur to innovation. In fact, however, industries differ widely in the extent to which patents are effective. The evidence suggests that patents are featured as a means of protecting innovation in only a few industries, while, in most industries, firms tend to rely more heavily upon other means of appropriation. In some instances, imitation is costly despite the absence of strong patent protection. In others, investment in complementary capabilities such as marketing and manufacturing can facilitate appropriation when neither strong patents nor technical barriers to imitation are present. Reflecting the Levin et al. (1987) and Cohen et al. (2000) survey results, the consensus is that patents are limited in their effectiveness in most manufacturing industries because patents rarely prevent 'inventing around,' while disclosing information that facilitates such activity (Levin et al., 1987; Cohen et al., 2000).

The Levin et al. (1987) survey revealed that firms in many industries tend to regard mechanisms other than patents as quite effective in appropriating the returns from innovation. In numerous lines of business outside the chemical and pharmaceutical industries, firms reported that the advantages of a head start and the ability to move quickly down the learning curve were more effective means of appropriation than patents. In their survey administered 11 years later in 1994, Cohen et al. (2000) also found that firms typically believed means other than patents to be most effective in protecting innovation. But, in contrast to the Levin et al. results, secrecy was ranked as the most effective mechanism in 17 of 34 three-digit SIC code level manufacturing industries, and thus at least comparable in importance to lead time, which is ranked first in 13 industries.

In addition to differences in appropriability conditions over time and across firms of different sizes, appropriability conditions may also differ across regions. Although Arundel (2001) found appropriability conditions in Europe to resemble that of the United States, Cohen et al. (2002a), found sharp differences between the United States and Japan, where the effectiveness of patents in protecting product innovations were at least comparable to the other major mechanisms, and equivalent to the most effective mechanism, lead time.

Despite the evidence on interindustry – no less cross-national – differences in appropriability conditions, there is, however, no clear empirical consensus about whether greater appropriability encourages innovative activity. The simplest hypothesis, derived from the standard argument supporting the patent system, is that innovative activity will increase monotonically with appropriability because spillovers create a disincentive to innovative effort. In the model of Cohen and Levinthal (1989) the simple 'disincentive effect' of spillovers remains, but there is an offsetting incentive to invest in 'absorptive capacity'-building R&D to make use of them, suggesting that R&D spillovers are not as much of a public good as previously supposed. Levin and Reiss (1988) similarly suggested another countervailing incentive effect. To the extent that own and rival R&D are different but complementary, the knowledge produced by a firm's competitor may raise the marginal product of own R&D. Thus, the same knowledge flows that can diminish appropriability can increase the productivity of other firms' R&D and offer a positive incentive to invest in R&D. In either instance, an increase in spillovers (decrease in appropriability) has an ambiguous effect on industry R&D. If we broadly consider the Levin and Reiss and Cohen and Levinthal models as together positing a complementarity effect of R&D spillovers, then a key question for

understanding R&D incentives and innovation at the industry level is what factors condition the trade-off between spillovers' negative appropriability incentive effect versus their positive complementarity effects. Empirical findings to date do not establish whether the net effect of appropriability on R&D incentives is positive or negative, nor do we yet know the extent to which the net effect varies across industries, no less what may condition different effects observed across industries.

Rather than focusing on appropriability conditions generally, economists over the past 25 years or so have focused on the narrower and directly policy-relevant question of the effects on R&D and innovation of patents in particular. Economists have argued that the impact of patents – and stronger patents – on innovation is not as straightforward as popularly believed. There are theoretical reasons to question whether patents stimulate R&D. For example, although the expectation of monopoly rents should induce inventive effort, the prospect of patent disclosures can offset the private gains from patenting (cf Horstmann et al., 1985). Also, 'stronger' patents mean that, not only any given firm's patents, but also those of its rivals are stronger (cf Jaffe, 2000), possibly dissuading rival innovation. Merges and Nelson (1990) and Scotchmer (1991) argue that, where technologies progress cumulatively and patents are broad, the licensing or enforcement decisions of upstream inventors may retard downstream innovation. Heller and Eisenberg (1998) and Shapiro (2000) also argue that, in some instances, the fragmentation or proliferation across different stakeholders of potentially overlapping patent rights for a given technology can raise transaction costs sufficiently to retard or even block otherwise worthwhile innovation.

A key question given our concern with the drivers of innovative activity is the effect of patents on R&D. The antecedent question, however, is whether the patenting of inventions indeed yields a return, presumably by increasing the value of whatever inventions that are patented. The work of Pakes, Schankerman, Lanjouw, and colleagues, in their examinations of European firms' patent renewal decisions (e.g., Schankerman, 1998; Schankerman and Pakes, 1986; Lanjouw, 1998), suggest, for Europe, that patent protection does yield a return, sometimes substantial, with estimates of the implied subsidy to R&D varying between 10% (Lanjouw, 1998) to as much as 25% (Schankerman, 1998). Using the Carnegie Mellon Survey data for the US manufacturing sector, Arora et al. (2008) estimate the 'patent premium' – the proportional increment to the value of an invention conferred by patenting it – by estimating a structural model linking a firm's R&D effort with its decision to patent, recognizing that R&D and patenting affect one another and are both driven by many of the same factors. Their results indicate that, even though most innovations are not worth patenting, patents are valuable for a subset of innovations and thus do provide incentives for R&D. Moreover, conditional on patenting, the expected premium is substantial, with firms earning an average 50% premium, with this conditional premium ranging from an average of 60% in the health-related industries to about 40% in electronics. Employing the estimated coefficients, the authors also simulate the effect of changing the premium on R&D. They find that an increase in the mean of the patent premium distribution for a typical manufacturing firm in their sample would significantly stimulate R&D, especially in industries where the patent premium is high, such as drugs, biotech, and medical instruments. Thus, consistent with the descriptive findings of Levin et al. (1987) and Cohen et al. (2000), Arora et al. (2008) find that patents are not as central to the protection of inventions as other mechanisms except in a few, selected industries; their estimates confirm that in most industries, patenting the *typical* innovation is indeed not profitable. However, even in these industries, some innovations are profitable to patent.

An important feature and limitation of both the patent renewal and Arora et al. (2008) studies is their focus on the direct, private returns to patenting. By restricting R&D spillovers or potentially creating complex thickets, patents may, however produce aggregate, more system-wide effects not easily discerned by considering the responses of individual firms to changes in the strength of their own patents. Indeed, empirical studies of the relationship between patenting and innovation at the aggregate levels of nations or industries have provided rather more ambiguous results about the impact of patents on innovation, although this ambiguity may be partly due to difficulties in controlling for the endogeneity of patent policy or the joint determination of R&D and patenting. Thus, neither the results of the patent renewal studies, nor those of Arora et al. (2008) imply that patents necessarily yield a net social welfare benefit overall.

The main lesson of the Levin et al. (1987) and Cohen et al. (2000) surveys is that firms use a range of appropriability mechanisms – and patents are not the central means of protection in most US manufacturing industries. Economists have not, however, devoted the energy to the empirical study of the use of the other mechanisms as they have to patents, nor, in their analyses of the effects of patents, have they explicitly considered patents as just one of several possible mechanisms in any given industry. Likely reasons for this inattention include the difficulty of finding suitable data and formulating precise tests to distinguish among competing hypotheses concerning the deployment of the different means of appropriation.

Conclusion

This article attempts to highlight important and robust empirical findings on the determination of R&D, and R&D performance, as well as key questions that remain to be addressed. We have divided our consideration across three broad areas of inquiry: (1) the Schumpeterian hypotheses relating innovation to market structure and firm size; (2) the role of firm characteristics; and (3) the role of industry-level variables broadly characterized as reflecting demand, technological opportunity, and demand conditions. The overall conclusion is that progress in advancing our empirical understanding of the determination of firms' and industries' innovative activity and performance has been uneven.

Despite the consensus regarding the importance of demand, appropriability, and technological opportunity conditions in affecting innovative activity and performance, consideration of the role of these variables has a considerable distance to go.

These three classes of industry-level variables appear to explain a good deal of the variance in firms' innovative activities. Yet, our understanding of how this influence is exercised is limited. We have ideas, some clearer than others, of how different dimensions of demand, technological opportunity, and appropriability may influence innovative activity and performance. Our tests for the presence and importance of these particular effects are, however, often indirect.

In comparison to our understanding of the influence of industry-level variables, our understanding of the role of firm-level variables is more primitive still. Economists began to consider the influence of firm characteristics from the mid-1980s through the mid-1990s, but there has been less work on the topic since. Although it is difficult to collect data on firm characteristics, empirical analysis is made all the more challenging due to the possible endogeneity of many of these characteristics.

Several overarching lessons may be drawn from the empirical literature on innovative activity and performance. First, we need more and better data on the range of variables considered. Better measurement of innovative activity itself is also essential. Consider, for example, the absence of measures of innovative activity that occurs outside of the formal labs and accounting categories encompassed by R&D, which Mansfield (1968) suggests may represent about 50% of manufacturing firms' investments in product and process innovation. In addition, we lack accurate measures of innovation itself – the output of innovative activity. Although progress has been made on measuring innovation with citation-weighted patent counts (Trajtenberg, 1989, 1990), it is understood that such a measure is still flawed given differences in the propensity to patent across industries and firms as well as our limited understanding of citing behavior (cf Alcacer et al., 2009). Finally, another major lacuna in our understanding of the determinants of innovative activity and performance is our ignorance of innovation in the service sector, except for a modest literature on financial services (Frame and White, 2004; Lerner, 2006). This gap is not simply a matter of a lack of data. It speaks to our very notion of innovation, and how to capture innovation when a good deal of it occurs outside of R&D labs, and especially in the context of client-specific relationships.

Perhaps one of the more basic lessons to emerge from the empirical literature is that, although testing loosely motivated hypotheses may yield empirical results, even robust ones, their interpretation can be challenging, and the insight that can be gleaned from such findings is often limited in the absence of underlying theory. However banal, the prescription that modeling can serve as a useful handmaiden to empirical analysis has often been overlooked in studies of innovation.

There is also, however, an important role to be played by more inductive efforts. Consider, for example, that much of our empirical understanding of innovation derives not from the estimation of econometric models, but from the use of other empirical methods. The historical and case study literatures provide a rich array of insights and factual information, and often constitute a source of hypotheses and inspiration for more rigorous approaches. More strikingly, many of the most credible empirical regularities have been established not by estimating and testing elaborate models with published data but by the painstaking collection of original data, often in the form of responses to simple questions. Even as econometric methods advance and the quality of published data improves, it will be important to remain catholic in the application of empirical techniques.

Acknowledgments

This article is a condensation of the author's 2010 review, *Fifty Years of Empirical Studies of Innovative Activity and Performance*, (Cohen, 2010) which updated and drew heavily from a prior survey written by the author (Cohen, 1995), which, in turn, drew extensively from a prior survey written by Richard Levin and the author (Cohen and Levin, 1989).

See also: Evolution: Diffusion of Innovations; Innovation, Theory of; Innovation; Technology and Organization; Technology and Social Control; Technology, Evolution of.

Bibliography

Acemoglu, D., Linn, J., 2004. Market size in innovation: theory and evidence from the pharmaceutical industry. Quarterly Journal of Economics 119, 1049–1090.

Acs, Z.J., Audretsch, D.B., Feldman, M., 1992. Real effects of academic research: comment. The American Economic Review 82 (1), 363–367.

Alcacer, J., Gittelman, M., Sampat, B., 2009. Applicant and examiner citations in US patents: an overview and analysis. Research Policy 38 (2), 415–427.

Arora, A., 1996. Testing for complementarities in reduced-form regressions: a note. Economic Letters 50, 51–55.

Arora, A., Ceccagnoli, M., Cohen, W.M., 2008. R&D and the patent premium. International Journal of Industrial Organization 26, 1153–1179.

Arrow, K.J., 1962. Economic welfare and the allocation of resources for invention. In: Nelson, R.R. (Ed.), The Rate and Direction of Inventive Activity. Princeton University Press, Princeton.

Arundel, A., 2001. The relative effectiveness of patents and secrecy for appropriation. Research Policy 30, 611–624.

Bernstein, J.I., Nadiri, M.I., 1988. Interindustry R&D spillovers, rates of return, and production in high-tech industries. American Economic Review Proceedings 78, 429–439.

Bernstein, J.I., Nadiri, M.I., 1989. Research and development and intraindustry spillovers: an empirical application of dynamic duality. Review of Economic Studies 56, 249–267.

Bresnahan, T., Trajtenberg, M., 1995. General purpose technologies – engines of growth? Journal of Econometrics 65, 83–108.

Clark, K.B., Chew, W.B., Fujimoto, T., 1987. Product development in the world auto industry. Brookings Papers on Economic Activity 3, 729–771.

Cockburn, I., Henderson, R., 1998. Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery. Journal of Industrial Economics 46, 157–182.

Cohen, W.M., 1995. Empirical studies of innovative activity. In: Stoneman, P. (Ed.), Handbook of the Economics of Innovation and Technical Change. Basil Blackwell, Oxford.

Cohen, W.M., 2010. Fifty years of empirical studies of innovative activity and performance. In: Hall, B.H., Rosenberg, N. (Eds.), Handbook of the Economics of Innovation, vol. I. Elsevier, Amsterdam, North Holland.

Cohen, W.M., Goto, A., Nagata, A., Nelson, R.R., Walsh, J.P., 2002a. R&D spillovers, patents and the incentive to innovate in Japan and the United States. Research Policy 31, 1349–1367.

Cohen, W.M., Klepper, S., 1992a. The anatomy of industry R&D intensity distributions. American Economic Review 82, 773–788.

Cohen, W.M., Klepper, S., 1992b. The tradeoff between firm size and diversity in the pursuit of technological progress. Small Business Economics 4, 1–14.

Cohen, W.M., Klepper, S., 1996. A reprise of size and R&D. Economic Journal 106, 925–951.

Cohen, W.M., Levin, R.C., 1989. Empirical studies of innovation and market structure. In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization. North Holland, Amsterdam.

Cohen, W.M., Levinthal, D.A., 1989. Innovation and learning: the two faces of R&D – implications for the analysis of R&D investment. Economic Journal 99, 569–596.
Cohen, W.M., Nelson, R.R., Walsh, J.P., February 2000. Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (Or Not). Working Paper 7552. National Bureau of Economic Research, Cambridge, MA. revised 2004.
Cohen, W.M., Nelson, R.R., Walsh, J.P., 2002b. Links and impacts: the influence of public research on industrial R&D. Management Science 48, 1–23.
Dasgupta, P.S., Stiglitz, J.E., 1980. Industrial structure and the nature of innovative activity. Economic Journal 90, 266–293.
Evenson, R.E., Kislev, Y., 1976. A stochastic model of applied research. Journal of Political Economy 84, 265–281.
Frame, W.S., White, L.J., 2004. Empirical studies of financial innovation: lots of talk, little action? Journal of Economic Literature 42, 116–144.
Galbraith, J.K., 1952. American Capitalism: The Concept of Countervailing Power. Houghton Mifflin, Boston.
Geroski, P.A., 1994. Market Structure, Corporate Performance and Innovative Activity. Clarendon Press, Oxford.
Griliches, Z., 1979. Issues in assessing the contribution of research and development to productivity growth. Bell Journal of Economics 10, 92–116.
Griliches, Z., 1992. The search for R&D spillovers. Scandinavian Journal of Economics 94 (Suppl.), 29–47.
Hall, B.H., 2002. The financing of research and development. Oxford Review of Economic Policy 18, 35–51.
Heller, M., Eisenberg, R., 1998. Can patents deter innovation? The anticommons in biomedical research. Science 28, 698–701.
Henderson, R.M., 1993. Underinvestment and incompetence as responses to radical innovation: evidence from the photolithographic alignment equipment industry. The Rand Journal of Economics 24 (2), 248–270.
Henderson, R.M., Cockburn, I., 1994. Measuring competence: exploring firm effects in pharmaceutical research. Strategic Management Journal 15, 63–84.
Henderson, R.M., Cockburn, I., 1996. Scale, scope, and spillovers: determinants of research productivity in the pharmaceutical industry. RAND Journal of Economics 27, 32–59.
Horstmann, L., MacDonald, J.M., Slivinski, A., 1985. Patents as information transfer mechanisms: to patent or (maybe) not to patent. Journal of Political Economy 93, 837–858.
Jaffe, A.B., 1986. Technological opportunity and spillovers of R&D. American Economic Review 76, 984–1001.
Jaffe, A.B., 1988. Demand and supply influences on R&D intensity and productivity growth. The Review of Economics and Statistics 72, 431–437.
Jaffe, A.B., 1989a. Characterizing the "technological position" of firms, with application to quantifying technological opportunity and research spillovers. Research Policy 18, 87–97.
Jaffe, A.B., 1989b. Real effects of academic research. American Economic Review 79, 957–970.
Jaffe, A.B., 2000. The US patent system in transition: policy innovation and the innovation process. Research Policy 29, 531–558.
Jaffe, A.B., Trajtenberg, M., Henderson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. The Quarterly Journal of Economics 108, 576–598.
Jewkes, J., Sawers, D., Stillerman, R., 1958. The Sources of Invention. Macmillan, London.
Kamien, M.I., Schwartz, N.L., 1970. Market structure, elasticity of demand, and incentive to invent. Journal of Law and Economics 13, 241–252.
Klevorick, A.K., Levin, R., Nelson, R.R., Winter, S., 1995. On the sources and significance of interindustry differences in technological opportunities. Research Policy 24, 195–205.
Lanjouw, J.O., 1998. Patent protection in the shadow of infringement: simulation estimations of patent value. Review of Economic Studies 65, 671–712.
Lerner, J., 2006. The new new financial thing: the origins of financial innovations. Journal of Financial Economics 79, 223–255.
Levin, R.C., 1978. Technical change, barriers to entry, and market structure. Economica 45, 347–361.
Levin, R.C., Klevorick, A.K., Nelson, R.R., Winter, S.G., 1987. Appropriating the return from industrial R&D. Brookings Papers on Economic Activity 783–820.
Levin, R.C., Reiss, P.C., 1988. Cost-reducing and demand-creating R&D with spillovers. Rand Journal of Economics 19, 538–556.
Link, A.N., Long, J., 1981. The simple economics of basic scientific research: a test of Nelson's diversification hypothesis. Journal of Industrial Economics 30, 105–109.
Malerba, F., 2007. Innovation and the dynamics and evolution of industries: progress and challenges. International Journal of Industrial Organization 25 (4), 675–699.
Mansfield, E., 1968. Industrial Research and Technological Innovation: An Econometric Analysis. Norton, New York.
Mansfield, E., 1981. Composition of R and D expenditures: relationship to size, concentration, and innovation output. Review of Economics and Statistics 62, 610–614.
Merges, R.P., Nelson, R.R., May 1990. On the complex economics of patent scope. Columbia Law Review 90 (4), 839–916.
Montgomery, C., 1994. Corporate diversification. Journal of Economic Perspectives 8, 163–178.
Mowery, D.C., Rosenberg, N., 1989. Technology and the Pursuit of Economic Growth. Cambridge University Press, Cambridge.
Nelson, R.R., 1959. The simple economics of basic scientific research. Journal of Political Economy 67, 297–306.
Nelson, R.R., 1982. The role of knowledge in R&D efficiency. Quarterly Journal of Economics 97, 453–470.
Nelson, R.R., 1986. Institutions supporting technical advance in industry. American Economic Review Proceedings 76, 186–189.
Nelson, R.R., Peck, M.J., Kalachek, E.D., 1967. Technology, Economic Growth and Public Policy. The Brookings Institution, Washington, DC.
National Research Council (NRC), 2005. Measuring Research and Development Expenditures in the U.S. Economy. National Academies Press, Washington, DC.
Pakes, Schankerman, M., 1984. An exploration into the determinants of research intensity. In: Griliches, Z. (Ed.), R&D Patents and Productivity. University of Chicago Press for the NBER, Chicago.
Parker, W.N., 1972. Agriculture. In: Davis, L.E., Easterlin, R.A., Parker, W.N. (Eds.), American Economic Growth: An Economist's History of the United States. Harper and Row, New York.
Pavitt, K., Robson, M., Townsend, J., 1987. The size distribution of innovating firms in the UK: 1945–1983. Journal of Industrial Economics 35, 297–316.
Rosenberg, N., 1974. Science, invention, and economic growth. Economic Journal 84, 90–108.
Schankerman, M., 1998. How valuable is patent protection? Estimates by technology field. RAND Journal of Economics 29, 77–107.
Schankerman, M., Pakes, A., 1986. Estimates of the value of patent rights in European countries during the post-1950 period. Economic Journal 96, 1052–1076.
Scherer, F.M., 1965. Firm size, market structure, opportunity, and the output of patented inventions. American Economic Review 55, 1097–1125.
Scherer, F.M., 1967. Market structure and the employment of scientists and engineers. American Economic Review 57, 524–531.
Scherer, F.M., 1982. Demand-pull and technological innovation: Schmookler revisited. Journal of Industrial Economics 30, 225–237.
Scherer, F.M., et al., 1959. Patents and the Corporation, second ed. Boston, privately published.
Schmookler, J., 1962. Economic sources of inventive activity. Journal of Economic History 22, 1–10.
Schmookler, J., 1966. Invention and Economic Growth. Harvard University Press, Cambridge, MA.
Schumpeter, J.A., 1942. Capitalism, Socialism, and Democracy. Harper, New York.
Scotchmer, S., 1991. Standing on the shoulders of giants: cumulative research and the patent law. Journal of Economic Perspectives 5, 29–41.
Scott, J.T., 1984. Firm versus industry variability in R&D intensity. In: Griliches, Z. (Ed.), R&D Patents and Productivity. University of Chicago Press for the NBER, Chicago.
Scott, J.T., 1991. Research diversity and technical change. In: Acs, Z., Audretsch, D. (Eds.), Innovation and Technical Change. University of Michigan Press, Ann Arbor.
Shapiro, C., 2000. Navigating the patent thicket: cross licenses, patent pools, and standard setting. In: Jaffe, A., Lerner, J., Stern, S. (Eds.), Innovation Policy and the Economy, vol. 1. National Bureau of Economic Research and MIT Press, Cambridge.
Solow, R., 1957. Technical change and the aggregate production function. Review of Economics and Statistics 34, 312–320.
Spence, A.M., 1975. Monopoly, quality, and regulation. Bell Journal of Economics 6, 417–429.
Spence, A.M., 1984. Cost reduction, competition, and industry performance. Econometrica 52, 101–121.
Stoneman, P., 1979. Patenting activity: a re-evaluation of the influence of demand pressures. Journal of Industrial Economics 27 (4), 385–401.
Sutton, J., 1998. Technology and Market Structure. MIT Press, Cambridge.
Taylor, C.T., Silberston, Z.A., 1973. The Economic Impact of the Patent System: A Study of the British Experience. Cambridge University Press, Cambridge.
Teece, D.J., 1986. Profiting from technological innovation: implications for integration, collaboration, licensing and public policy. Research Policy 15, 286–305.
Trajtenberg, M., 1989. The welfare analysis of product innovations, with an application to computed tomography scanners. Journal of Political Economy 97 (2), 444–479.
Trajtenberg, M., 1990. A penny for your quotes: patent citations and the value of innovations. The Rand Journal of Economies 21, 172–187.
von Hippel, E., 1988. The Sources of Innovation. Oxford University Press, New York (1994).

Innovation, Theory of

Jon Sundbo, Roskilde University, Roskilde, Denmark

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E.M. Rogers, volume 11, pp. 7540–7543, © 2001, Elsevier Ltd.

Abstract

In sociology, the idea of innovation was primarily introduced by Gabriel Tarde. Tarde's theory has been further developed by diffusion theorists. Entrepreneurship is part of behavioral innovation theory just as technological, social, and behavioral changes. Three different perspectives may be identified in the emerging reborn innovation theories around 2000: Entrepreneurship, technology-economic, and strategic. Innovation as a phenomenon and concept has recently been brought out of the market-economic academic sphere into the public sector and the civil society, conceptualized as social entrepreneurship and social innovation.

An innovation is a renewal of behavioral or material structures or practices. More formally, innovation can be defined as the introduction of a new social, material, or economic phenomenon or a new combination of old ones. Examples include the introduction of new technology, a new type of organization, or a new fad of young peoples' dating behavior. Innovation refers to both behavioral and technological changes. In sociology, the concept of innovation has been used to explain social change, indeed it was only later that it was used in economics to explain economic growth. Recent theoretical developments have again focused on the sociological side of innovation (e.g., Sundbo, 1998; Franz et al., 2012).

This article takes its point of departure in the assumption that innovation is fundamentally behavioral and thus a sociological phenomenon, even when its result may be technological or economic. Innovation is created by individuals and their will to change, however, their behavior is influenced by social and economic structures. This perspective is in line with Giddens' (1984) structuration theory that, put somewhat simplistically, argues that social structures provide the framework in which actors act and whom either uphold or change those structures. This article will discuss different aspects of innovation theory and has a theory-historical perspective, however, the emphasis will be on contemporary theory.

Innovation is about the process and the result, not the idea. An idea or an invention is necessary to create change, but it does not in itself create any change. Innovation as a concept has two meanings. One refers to the process of changing elements or structures, the other, the result of the process. The result may be a new type of behavior or, for example, a new good or service (such as the personal computer or Facebook). Even when the result is a new technology or good, the use of it requires social acceptance and willingness to change habits and attitudes. Even in economic theory, innovation is seen as a new product that is only sold on the market if people buy it. People must want to use it and they will perhaps need to adapt their behavior to be able to use it.

Classic Theory

In sociology, the idea of innovation was primarily introduced by Gabriel Tarde (1890, 1897). He developed a theory stating that societal changes are introduced by individuals who start behaving in a new way or use new tools. These individuals are imitated by others, who again are imitated by others. This is done via communication and social interaction that defuses the change more widely throughout society. Tarde saw this as a complex social process that included resistance to the new phenomena, social conflicts, and problems. The new behavior and the structures they result in are, perhaps, advantageous to some individuals and social groups, but disadvantageous to others.

Joseph Schumpeter is regarded as the father of innovation theory within economics. Schumpeter emphasized the behavioral aspects of innovation processes and he has thus contributed to behavioral theory as much as to economic theory. For the early Schumpeter, some individuals, whom he termed entrepreneurs, have a drive to create new activities and change the market and production systems and thereby create economic growth (Schumpeter, 1934); this theoretical position is often called Schumpeter I as he later changed his view and explained economic growth and social changes in terms of organized structures (Schumpeter, 1939) such as large firm organizations that decide to innovate to get a better competitive position on the market or political organizations that introduce new structures into society; this theoretical position is often called Schumpeter II. In addition, he discussed whether innovation leads to capitalism or socialism and how it influences democracy (Schumpeter, 1943).

Diffusion

Tarde's theory has been further developed by diffusion theorists, particularly Everett M. Rogers (1995) who stated that innovations diffuse through society following an S-shaped acceptance curve. In the earliest phase, a few people accept the innovation (early adopters), it then spreads as more people accept it and finally the least change oriented, the so-called laggards, will accept it. The S-curve theory of diffusion, similar to Tarde, makes an implicit assumption about social stratification. The first adopters are individuals who have a high social status, the last ones are people with low status. Rogers supported his arguments with the case of agricultural diffusion. Those that first accepted new technology were the

landed proprietors, then came the farmers, and last the smallholders. However, this pattern may not always occur. Some behavior and technology is adopted first by the lower status groups. Some entrepreneurship theories point to the opposite cause–effect relationship – that being the first to introduce a new innovation may give someone high status, at least in the long run.

Rogers also emphasized how social structures, for example, governmental regulation or culture, influence the diffusion processes. A common point of departure in an institution, for example, a university, can lead to development of a series of innovations in a local community, as was the case with the emergence of Silicon Valley in California. Following this line of thought the rapidly growing technology business area is a result of entrepreneurs not imitating each other, but using their competencies, social relations, and earlier innovations as the basis for developing new innovations. This can be seen as a diffusion of entrepreneurship within a local community.

Within anthropology, the diffusion of innovations has been a core factor in explaining cultural change in isolated and traditional cultures. New phenomena from outside a culture are, from time to time, adopted by this culture. In such cases, anthropologists talk about assimilation or acculturation of the new phenomena that can be behavioral or material. The central issues in anthropological theory are whether the phenomena will be adopted or rejected, how they are adopted, and what social consequences the adoption will have (cf Barnett, 1953). The new things may be introduced by a member of the community who has traveled abroad, seen something new, and brought it with him back home. The new things may also be introduced if immigrants have come to the community. Sometimes the new thing exists side by side with an old or several new phenomena compete to become dominant.

Entrepreneurship

Entrepreneurship is part of behavioral innovation theory. The entrepreneur is a person. The idea of entrepreneurs as creators of change stems from economics. As early as 1803, Jean-Baptiste Say formulated a law of economic development based on entrepreneurship, and Schumpeter continued this line of thinking. An individual entrepreneur introduces something new (a technology, a product, or a behavior) that penetrates the market and then changes the market conditions. This contrasts with the behavioral sciences where there is no clear idea about one person creating and carrying out the whole change process. In Tarde's theory, several people are involved in the change process. The diffusion of innovations is seen as a process having different stages where different individuals and social subsystems participate in the development of the innovation.

Schumpeter's and the economic theory's way of thinking about entrepreneurship is indeed connected to innovation – the entrepreneur is an innovator. However, in his early work, Schumpeter (1934) emphasized the entrepreneur's activity rather than the innovations (new elements) that were possible outcomes of the activity. The entrepreneur is a person who is extremely active in creating and organizing production and sales activities. These activities may be based on a new idea or invention, but that does not need to be the case. The entrepreneur can take others' ideas or inventions as the basis for his business activities. It is important that he creates activities that change the economic market's 'rules of the game.' The entrepreneur in economic theory refers to a person who creates new business activities, and there is an ongoing debate about whether the activities must contain something new, that is, an innovation, to be called entrepreneurship or not. In some languages, entrepreneurship just means small business.

In the behavioral sciences, such as sociology and anthropology, there has also been some focus on individuals as creators of innovation and social change, however, not as much as in economic theory. Rather innovation has been understood as a systemic phenomenon, or as one in which several individuals are involved (cf Tarde's and Rogers' adopters, imitators, and opponents). This may be a logical consequence of sociology's attempt to create theories of how the development of social systems can be explained by system factors and not by individuals, which was a general understanding before sociology emerged. However, more recently, there has been a fundamental discussion within sociology about whether individuals or the structures themselves determine social change, the so-called actor–structure debate referred to above (e.g., Giddens, 1984). Until recently, this debate had not been connected with the concepts of innovation or entrepreneurship.

Social Change

In the 1950s and 1960s, a theory of social change based on innovation and diffusion emerged in the United States (e.g., Zollschan and Hirsch, 1964). At the time functionalistic theory dominated American sociology and a part of the functionalistic approach was to see society as fundamentally stable. Sociology was inspired by economics and the natural sciences to see societies as systems in, or seeking, equilibrium. This did not mean that there are no differences or conflicts between people and groups; however, a social system was seen as developing by itself toward a state in which differences were institutionalized without open conflicts and social structures becoming permanent. When this state is reached, the social system functions best; every element, relation and structure, has its function in maintaining the optimal equilibrium situation.

Within functionalist sociology, there was an acknowledgment that social systems may change. However, this was seen as an exception and only happening rarely. Social change could therefore be a special discipline within sociology that deals with the special situations, where changes in a social system take place (e.g., Moore, 1967). These changes may be revolutionary such as when a social revolution or a war suddenly breaks out. This happens very rarely. Most changes, when they happen, are changes in parts of the system. Social change was a discipline aimed at explaining why these partial changes happen. Innovation and diffusion became core concepts in this explanation (LaPiere, 1965). The social change theories operate with new elements, may emerge within the equilibrium system, just as happens in, for example, biological equilibrium systems (e.g., fish gain new skin colors or generic mutation). 'Innovation' was introduced to explain why and how these new elements enter the system.

The theory argues that even social equilibrium systems may develop in some cycles or from time to time slowly developing toward a new equilibrium. Social change as a discipline studies these 'changes within the stable social system' by emphasizing innovations and the diffusion of them into and within the system. Innovation is a rare phenomenon. Social change theories were less occupied by explaining why innovations emerge – they more or less based that explanation on the Tardian theory of entrepreneurship and diffusion. It was held that innovation processes were started by individuals (entrepreneurs), who must be characterized as deviants; they have a mentality and behavior outside the norm. This 'deviant' behavior may in the long run be seen as positive by the members of the system, but in the first phase it is often seen as negative – just because it is deviant. The social change theorists were more occupied by discussing what the innovations and their diffusion do to the social system that led to different theories of 'system conformity' dynamics where the single innovations were seen as part of a pattern, not as anarchic changes that point in every direction.

Social change theories may be seen as more 'reformist' in their view on what innovation and entrepreneurship can do to the social system than Schumpeter was in his view on what innovation and entrepreneurship can do to the economic system. Schumpeter talks about swarms of innovations that completely destroy economic equilibrium ('creative destruction' as he called it), which is a radical view in contrast to the mainstream economic equilibrium theory.

Technological and Behavioral Changes

Originally, innovation theory dealt with both behavioral and material changes; and both the macro (society) and micro (organizational) levels and this also characterized the social change theories. Later in the mid-twentieth century, the interest in innovation become focused on the organizational level and on technological changes that are narrower, structured, and science- or engineer-based changes concerning the material aspects of life. The wider behavioral aspects became overshadowed by the focus on technological change and firms' position in the market competition. Innovation theory became an economics and business discipline.

This development due to the realization that World War II had resulted in the development of many new war technologies. After the war it was observed that these war technologies could be used in civil society and create economic growth and welfare. Science-based technological innovations were seen as a basis for the welfare society. Therefore economic innovation theory became a theory on technological innovation.

Technological innovation also became a topic in sociology. Theories on how technology is developed in society were put forward by, for example, Rosenberg (1976). An emerging discipline of science and technology in sociology attempted to explain how research and technology development activities could be understood and explained. In sociology, technological innovation and change was, however, mostly discussed in relationship to behavioral changes. A core theme was whether technological changes create social changes or vice versa. Ogburn (1957) put forward a theory of cultural lag which suggested that technological changes determined by technological innovations come first and change the material structures of society while the necessary cultural changes to utilize the new technologies lag behind and only come after some time.

Ogburn's theory may be interpreted as a kind of technological determinism where technological innovation determines social change. Later, a criticism of technological determinism emerged. In that discussion, some sociological theorists stated that social structures and processes determine the changes, even the material ones. It might be the case that technological innovations are necessary to develop society, but it is social needs and aims that select which technologies are introduced and diffused through society. This position was among others taken by the Science and Technology Studies (STS) approach (e.g., Bijker et al., 1987). This led to a development of theories that by some theorists were accused of expressing a behavioral determinism, where the materiality and science and technological innovation do not have any power to determine behavior or change society. More modified positions such as the actor network theories (ANT, e.g., Latour and Woolgar, 1979) emerged. They started to include material (or technological) innovations as determinants of social change in addition to social innovations.

The interest in innovation in the behavioral sciences decreased after the 1960s (except within the above mentioned technology-social structure determination fight, which did not primarily focus on the concept of innovation). Innovation became a topic dealt with in economics and was synonymous with technological innovation. In economics, the focus was much more on the invention, not the diffusion of innovations. Economic theory and analyses emphasized research and development of technology at macro (societal) and micro (firm) levels mostly concerning large companies. Entrepreneurship was by and large forgotten.

Contemporary Theory

The reappearance of the concept of innovation in behavioral theories started in the late 1980s, and innovation theory has increasingly become an important part of mainstream theories. More recently, innovation theories in economics have broadened their scope thus they are now concerned with entrepreneurship in addition to technological innovation and research and development in large organizations.

Three Perspectives

Around 2000, the reappearance of innovation theory pointed in several directions mixing elements from old and new theories. Political and business interests in society led to a new focus on how innovation could create economic growth and thus increase welfare and provide advantages in market competition. A wish to increase the number of innovations led to an interest in understanding how innovations emerge and how the development of innovations could be influenced. The interest and theorizing has concerned the macro, societal, and micro, organizational levels.

Three different perspectives may be identified in the emerging reborn innovation theories (Sundbo, 1998). They

offer different explanations of innovation and appoint innovation to different agents. The perspectives are summarized in the following model (Figure 1):

One of the perspectives, entrepreneurship, is a revival of the old entrepreneurship-based innovation theories, but it is now more focused on the emergence of entrepreneurs than on diffusion of innovations. Most of the theoretical discussions are related to the economic perspective: how do business entrepreneurs emerge and what are the social and economic conditions for them remaining entrepreneurs and business to grow?

The technology-economic perspective is a continuation of the interest in economic theory for technological innovation, research, and the economic results of technological innovation. However, this interest has been extended into other behavioral sciences such as sociology. In addition to the STS and ANT traditions, a focus on the organizational conditions for development of technological innovations and for their diffusion in the society has developed.

The third, strategic, perspective is new in behavioral and economic theories. It is a more sociological or anthropological perspective. It is focused on understanding the social processes that lead to innovations. Innovation is a goal-oriented and planned activity; it has led to the suggestion of using the notion strategy to characterize this perspective. Innovation is seen as an activity carried out in a social system determined by structural factors as well as actors. The goal, not the invention or idea, is in focus. The goal is one for a total group or social system such as a firm, a social group, or even a society. Ideas and inventions flow around in the social system, however they are selected and some of them are supported. Somebody takes the decision about which ideas to promote. This may be the result of more or less democratic or inclusive processes, but some individuals need to take the final decision or take the responsibility for carrying out the collective decision and reflect on the result of the strategy; if the result is not as desired, other innovations may be introduced (Sundbo and Fuglsang, 2002). The individuals who make the decisions are in contemporary social systems normally managers (which is a wide concept including, e.g., ministers, enterprise owners, and informal leaders of social groups). Even if the goal setting may be a collective process, the managers have a great influence on the innovation processes and the direction of social change. The manager is the main agent of innovation and change.

The three perspectives are metatheoretical ideal types and are often mixed in innovation theories and empirical research. The following presents contemporary innovation theories.

Entrepreneurship Roles and Networks

Entrepreneurship has increasingly been narrowed to entrepreneurial activities that lead to some kind of innovation (e.g., Casson, 2013). The theoretical interest has changed focus from not only looking at how entrepreneurs emerge, but also how they develop innovations. One discussion is whether entrepreneurs are driven by an internal force and the will to establish projects or firms, or whether they are driven by opportunities, for example, in the market.

Another discussion concerns how entrepreneurs act in their attempt to develop and implement the innovation and, maybe, as a consequence, enable their firm (or other organizations) to grow. They often use social networks, and entrepreneurship is increasingly seen as a social activity where the entrepreneur has or establishes relations to other people who again have relations to other people and the entrepreneur thus has access to many people in, often loosely coupled, networks. The networks are both professional – which can give access to knowledge, capital, labor force, and so forth – and social. If there are no positive mutual social attitudes, the networks will rarely work professionally either. Theories even state that innovation can happen in networks where several entrepreneurs collaborate or where one cannot point to any individual being the entrepreneur more than others. Entrepreneurship can be collective and several entrepreneurs can collaborate and engage in group innovation.

Entrepreneurship thus is increasingly seen as a collective activity that may be conditioned by social structures and not only by an individual drive that may stem from the person's

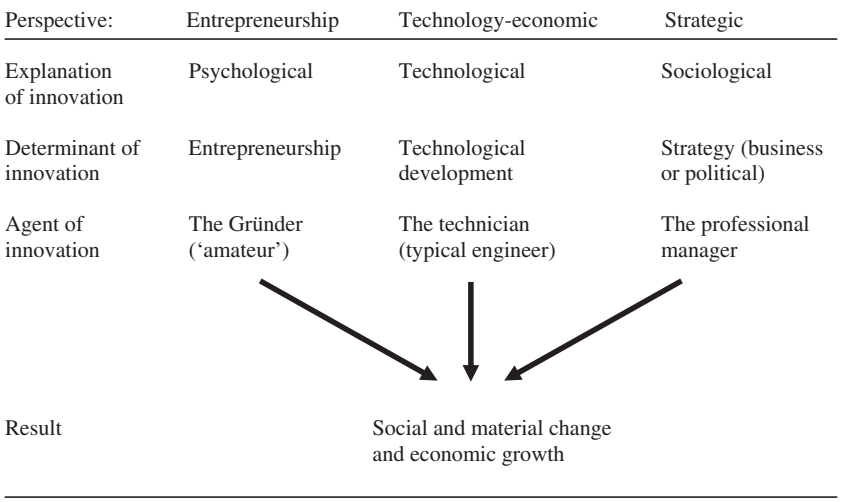

Figure 1 Three perspectives on innovation. Based on Sundbo, J., 1998. The Theory of Innovation. Edward Elgar, Cheltenham. p. 159.

adolescence. A theoretical development, inspired by social psychology, has been to see entrepreneurship as roles in innovation processes. Different roles have been identified, examples include the analyst who contributes knowledge, the actor who makes things happen, the sponsor or business angel who support the innovation, the community entrepreneur who supports networking, the extrapreneur who from within an organization propagates for entrepreneurship in other organizations.

Technology-Economic Innovation: Knowledge Diffusion

In recent research and theory, the interest for technological innovation has changed from a focus on the research and development processes to the diffusion of formal knowledge, typically scientific knowledge, which is the basis for technological innovation. The knowledge bases have increased dramatically as many new universities have been established and old ones have grown. Private knowledge bases have emerged within knowledge services (e.g., via new consultancy and private laboratory industries) and IT Networks such as the Internet extend access to knowledge bases dramatically. Firms do not need to have internal research to get new technology; they can procure that from outside and then internally develop it to create innovations. Knowledge has been seen as the most valuable economic factor (above labor and capital) because it is the basis for innovation, which economically has become the most important factor because it creates economic growth. Schumpeter's theory about organized innovation (Schumpeter II) has become mainstream within economic and welfare policy and in economic theory.

Knowledge diffusion has become a core topic within the technology-economic perspective (e.g., Nonaka and Takeuchi, 1995). The importance of knowledge institutions at a macro level and knowledge management at the micro (organizational) level has become the subject for innovation theories. Themes have concerned, among others, how some knowledge is formal and how some is tacit – informal knowledge that people have internalized, for example, in their work activities. Theories speculate on how formal knowledge can be utilized to create innovation, how tacit knowledge can be converted into formal knowledge, and how firm organizations can store knowledge about innovation processes – called organizational learning. Theories also speculate on how organizations can absorb external knowledge. One specific theory deals with absorptive capacity (Cohen and Levinthal, 1990), which is organizations' ability to process external knowledge and use it for innovation. The absorptive capacity is assumed to be highest if managers and employees have a higher education thus they seek formal external sources (such as research results, consultancy knowledge, and so forth) and are able to transform it into innovations.

Strategic Reflexivity and Innovation in Organizations: Open Systems

The strategic perspective focuses on how innovations are developed in organizational processes (e.g., Mintzberg, 1989; Sundbo, 2001). The development of innovation moves in several directions. One is a top-down process, where the top management of the organization (normally thought of as a firm organization) decides that the firm should develop thus innovation is necessary. The management therefore induces innovation processes in the organization by setting up formal innovation project groups and gives incentives to people in the organization, people outside the project groups also contribute to the innovation processes. Many employees thus have a role in innovation processes. Employees can also be internal entrepreneurs or intrapreneurs (cf Pinchot, 1985), and innovation processes can start from bottom-up. Employees can be entrepreneurs for a period, sometimes often, sometimes once in their lifetime.

Innovation thus becomes a complex process where many actors are active and they play together with, or against, the organizational system. Innovation may even create conflicts because it may lead to organizational changes that can change the power basis of different departments and individuals. Innovation is within this theoretical perspective seen as a broad phenomenon. It can include the development of new technology or new social behavior such as new organizational or management forms or a new market or communication behavior. This complex system must be managed and the task of that belongs to the top management. This is complicated because many ideas for innovation are presented and the most important is whether the environments (in business, the market; in politics, the voters) will accept the innovation. The managers often use a formulated organization policy, strategy, or business plan as the guide for making decisions. However, they need to reflect on how the innovations work in the environments, which may be called strategic reflexivity (Sundbo and Fuglsang, 2002). Employees are often involved in the strategic reflexive process because they encounter the environments (customers and clients) and therefore observe how the innovations work.

Innovating organizations are, in this perspective, considered as open systems (cf Chesbrough, 2006). They interact with the environments to reflect on the introduced innovations and to get new ideas for innovation. The strategic aim is to satisfy the environments (e.g., the market or a political system). Theory and research has therefore been interested in how users of the potential innovations can be involved in the innovation process (e.g., Sundbo and Toivonen, 2011). Firms imitate innovative behavior that has been developed in or by civil society. Some authors, such as Eric von Hippel (2005), claim that this will democratize the innovation processes in society.

Social Innovation

Innovation as a phenomenon and concept has recently been brought out of the market-economic academic sphere. There has, for example, been an increased political focus on innovation in the public sector. The strategic innovation theory perspective has been applied to public organizations, primarily with the aim of studying how innovation can make the public sector more efficient and simultaneously create more user satisfaction, for example, that citizens are more satisfied with public services (e.g., Windrum and Koch, 2008).

This interest has also spilled over to discussions about developments in civil society and the role innovation and

entrepreneurship play. Research suggests that social problems are often solved by innovations outside the formal public sector and are carried out by individuals, who may be characterized as entrepreneurs, and by groups and local networks. A question for future theory development is therefore whether the formal political system will be unable to solve the challenges of an aging population, increased bureaucracy, lifestyle determined health problems, and so forth. The solutions may come from civil society that might be more oriented toward innovation activities. A concept of social entrepreneurship has been developed to explain these activities (e.g., Ziegler, 2011). The entrepreneurship theory has been extended to cover the whole of society to include other sectors of civil society than the social sector, for example, cultural activities, sports activities, and so forth. The theoretical understanding focuses on networks and local communities.

Social entrepreneurship creates innovative solutions to societal problems. The development of the theoretical understanding of innovation is reverting back toward the social change theories of the 1950s and 1960s; however, it has dropped the functionalistic approach of that tradition. The concept of social innovation is increasingly used to characterize and understand innovation processes outside the technological-economic sphere. However, social innovation is not a very clear concept. It is used to characterize innovations within the social sector of civil society (a kind of voluntary social policy effort), but also nontechnological innovations in general and innovations initiated by employees, managers, users, and others outside the professional research and technology sphere (Cajaiba-Santana, 2014).

The Future of Innovation Theories

Innovation theory is facing a challenge to its foundation as a result of the recent developments relating to, for example, discussions concerning strategic and social innovation tendencies in the society. Innovation was introduced as a concept to understand change that was seen as an exception to stability and equilibrium. But, if the society is constantly moving and everything is changing, which many observers claim, change is not an exception rather it is the norm. Has the concept of innovation become meaningless? One may create theories about how changes are happening within particular field and their direction, but does a general theory of innovation have any meaning? Furthermore, it is postulated that all individuals need to adapt to changes and to be active in that process (e.g., if they want to make career and not loose their job). If everybody must carry out some kind of role with respect to change, we may ask whether or not the concept of entrepreneurship is meaningful?

The answer to these theoretical challenges is not yet clear. The answer might be that the concept of innovation will be restricted to technological business renewals – the technology-economic perspective. It might be the case that innovation theory develops into a new fundamental discipline, in line with psychology or sociology, that has the aim of explaining that every system has a driving force to change (maybe not only social systems, but also material systems and individuals) – a general change theory. It might also be that theory will state that people by nature both want change, but also stability and the task is to understand how much change, or innovation, is wanted or optimal in different social systems. It might also be that innovation theory will discuss socio-psychological issues by defining different roles in change process. Whatever, the extension of the use of the innovation concept calls for a, maybe radical, revision of innovation theory reflecting on what the theory fundamentally tries to explain.

See also: Diffusion in Sociological Analysis; Entrepreneurship and Economic Growth; Entrepreneurship and Organizations; Entrepreneurship, Geography of; Knowledge, Sociology of.

Bibliography

Barnett, H., 1953. Innovation: The Basis of Cultural Change. McGraw-Hill, New York.
Bijker, W., Hughes, T., Pinch, T., 1987. The Social Construction of Technological Systems. MIT Press, Cambridge, MA.
Cajaiba-Santana, G., 2014. Social innovation: moving the field forward. A conceptual framework. Technological Forecasting and Social Change 82, 42–51.
Casson, M. (Ed.), 2013. History of Entrepreneurship: Innovation and Risk-Taking 1200–2000. Edward Elgar, Cheltenham.
Chesbrough, H., 2006. Open Innovation. Harvard Business School Press, Boston.
Cohen, W.M., Levinthal, D.A., 1990. Absorptive capacity – a new perspective on learning and innovation. Administrative Science Quarterly 35 (1), 128–152.
Franz, H.W., Hochgerner, J., Howaldt, J., 2012. Challenge Social Innovation, Potentials for Business, Social Entrepreneurship, Welfare and Civil Society. Springer, Berlin.
Fuglsang, L., Sundbo, J., 2005. The organizational innovation system: three modes. Journal of Change Management 5 (3), 329–344.
Giddens, A., 1984. The Constitution of Society. Polity Press, Cambridge.
von Hippel, E., 2005. Democratizing Innovation. MIT Press, Cambridge, MA.
LaPiere, R.T., 1965. Social Change. McGraw-Hill, New York.
Latour, B., Woolgar, S., 1979. Laboratory Life: The Social Construction of Scientific Facts. Sage, London.
Mintzberg, H., 1989. Mintzberg on Management. Free Press, New York.
Moore, W.E., 1967. Order and Change. John Wiley, New York.
Nonaka, I., Takeuchi, H., 1995. The Knowledge Creating Company. Oxford University Press, New York.
Ogburn, W.F., 1957. Cultural lag as theory. Sociology and Social Research 41 (3), 167–174.
Pinchot, G., 1985. Intrapreneuring. Harper & Row, New York.
Rogers, E., 1995. Diffusion of Innovation. Free Press, New York.
Rosenberg, N., 1976. Perspectives on Technology. Cambridge University Press, Cambridge, MA.
Schumpeter, J., 1934. The Theory of Economic Development. Oxford University Press, New York.
Schumpeter, J., 1939. Business Cycles. McGraw-Hill, New York.
Schumpeter, J., 1943. Capitalism, Socialism and Democracy. Unwin, London.
Sundbo, J., 1998. The Theory of Innovation. Edward Elgar, Cheltenham.
Sundbo, J., 2001. The Strategic Management of Innovation. A Sociological and Economic Theory. Edward Elgar, Cheltenham.
Sundbo, J., Fuglsang, L. (Eds.), 2002. Innovation as Strategic Reflexivity. Routledge, London.
Sundbo, J., Toivonen, M. (Eds.), 2011. User-based Innovation in Services. Edward Elgar, Cheltenham.
Tarde, G., 1999[1897]. L'Opposition Universelle. Les Empêcheurs de penser en rond, Paris.
Tarde, G., 2001[1890]. Les Lois de l'imitation. Les Empêcheurs de penser en rond, Paris.
Windrum, P., Koch, P., 2008. Innovation in Public Sector Services. Edward Elgar, Cheltenham.
Ziegler, R., 2011. An Introduction to Social Entrepreneurship. Edward Elgar, Cheltenham.
Zollschan, G., Hirsch, W. (Eds.), 1964. Explorations in Social Change. Routledge and Kegan Paul, London.

Instinct, History of

Daniel R Huebner, University of North Carolina at Greensboro, Greensboro, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of 'instinct' has served as an influential way to conceptualize the recurrent forms, the ultimate motivations, and the possible heritability of human behaviors. Theories of instinct had a strong influence on nineteenth century psychology and social science. By the 1920s, however, instinct theories were the target of considerable criticism, which questioned whether it represented a single, scientifically valid, and observable phenomenon and whether it was relevant to the explanation of complex human behavior. The notion fell out of favor in the social and behavioral sciences by the 1930s, despite sporadic attempts to defend and rehabilitate it.

The concept of 'instinct' has served as an influential way to conceptualize the recurrent forms, the ultimate motivations, and the possible heritability of human behaviors. In the social and behavioral sciences, this notion has been strongly debated, with authors variously arguing that it refers to mental impulses or to physiological reflexes; that it represents the ultimate source of human behavior or is largely irrelevant to its explanation; that it denotes definite responses or amorphous tendencies; that it is fundamentally fixed or malleable; and that, as an explanation, it is scientifically sound or tainted with metaphysics.

The reputation of instinct as an explanatory concept rose strongly as Darwinian evolutionary principles were adopted by the social and behavioral sciences. Theories of instinct had a strong influence on late nineteenth and early twentieth century psychology and social science, but the notion was also the target of heavy criticism and fell out of favor as an explanation of human behavior by the 1930s. As alternative conceptual schemes for the explanation of human motives and conduct have developed in these sciences, attempts to rehabilitate or reformulate the notion of instinct have been largely overshadowed. In general, the modern human sciences have sought to explain the regularities of human behavior by more proximate psychological and environmental causes and to place emphasis on the social or cultural conditioning of behavioral patterns, at the expense of theories of instinct.

The Evolution and Heritability of Behavior

No systematic attempt is made here to trace the popular, philosophical, or theological uses of the term 'instinct.' However, there were several identifiable precursors to the modern biological conceptions of instinct in the eighteenth and nineteenth centuries that were drawn upon by later authors in the social and behavioral sciences. A notion akin to instinct in explaining the behavior of living organisms goes back at least as far as Aristotle, whose *De Anima* argued that the actions of animals are directed by an impulse in accord with the requirements of their natures. Other classical philosophical warrants include René Descartes' epistemological doctrine of 'innate ideas,' denoting those mental states or faculties necessary to explain the basic features of the mind that are apparently not the result of experience, and the political philosophy of 'natural law,' which led Thomas Hobbes and many other authors to ground human rights in often conflicting doctrines of man's innate qualities or needs.

In the popular writings on 'natural history' prior to the spread of modern evolutionary theories, instinct was commonly referred to as a special impulse imparted to a species by an act of divine creation. Each species was therefore thought to have an endowed set of fixed dispositions as part of its particular mode of existence, and humans were thought to be endowed not with instincts, but with reason, to guide their actions. In English-language scholarship, William Paley's (1802) *Natural Theology* was particularly influential in defining instinct as 'a propensity prior to experience and independent of instruction,' and in arguing that such complex, innate behavioral tendencies are evidence of the existence of a divine creator. Other major natural philosophers of the eighteenth century, including Goethe and Buffon, expressed similar views.

The works of Georges Cuvier (1769–1832) and Jean-Baptiste Lamarck (1744–1829) are representative of two of the major scientific directions in understanding instinct prior to the Darwinian revolution and their work served as critical reference points for later theories. Cuvier's (1817) *The Animal Kingdom* formalized the popular conception of instinct, arguing that they are innate, species-particular, 'involuntary mental impulses' that guide the conduct of 'lower' animals. They do not manifest any visible marks in the physiology of the organism, he thought, and they are not subject to fundamental change across generations. The prevalence of instinct in a species varies inversely with its level of intelligence and, consequently, instinct occupies a nearly irrelevant place in explaining human conduct. Cuvier illustrated the fundamental place of 'fixed instinct' in explaining the behavior of lower animals with now-classic examples including the stereotyped formation of honeycomb by bees and nests by birds; and he was an early proponent of the concept of 'social instincts' to explain the behavior of communal species.

In contrast to Cuvier, Lamarck's *Zoological Philosophy* (1809) argued that instincts are subject to change in the course of an individual organism's life. In this view, instincts may be acquired over the course of an organism's experience as 'habits,' these habits can affect the physiological structure of the organism and these 'acquired characteristics' may then be

directly heritable. This early evolutionary theory, he thought, allowed him to better explain the demonstrable fact that apparently innate behavior patterns can be changed in a permanent and heritable way, as in the 'domestication' of animals. The theory seemed to offer an explanation of how complex mental states are reproduced or spread in subsequent generations of a species, but the supposed 'inheritance of acquired characteristics' was controversial from its first printing and later proved incompatible with the biological mechanisms of inheritance elaborated by Darwin and others.

In his foundational *On the Origin of Species* (1859), Charles Darwin (1809–82) admitted that explaining instinct by means of 'natural selection' was one of his theory's most serious challenges and he consequently addressed the topic in detail. He referred to instincts as actions "performed in the same way by many individuals of a species, without experience and without their knowing the purpose." The role of instincts is comparatively less important in 'higher' than in 'lower' animals, he argued, and in man there are few if any fixed instincts. Although Darwin left open the possibility that some instincts may be inherited as Lamarckian 'acquired habits,' he argued that in most cases they could clearly be explained by principles of natural selection. Crucial to his argument, he sought to demonstrate that there are spontaneous variations in instincts manifested by individuals of a species just as there are variations of physical traits, and that these variations may affect the survivability of organisms that exhibit them in the ways predicted by natural selection processes. Darwin examined possible objections to this theory in depth, including apparent cases of inheritance of new habits by the offspring of domesticated animals; the development of instincts that seem too intricate to have developed spontaneously; instincts that appear to go against self-preservation; and others. In each case, he found that the observations could be explained by natural selection or human-directed selective breeding of instinctive tendencies that are already naturally exhibited.

Although he did not address all the possible problems, Darwin outlined the first coherent evolutionary theory to account for both the heritability and variability of complex behaviors. Other evolutionary theorists, especially Alfred Russel Wallace (1823–1913) and August Weismann (1834–1914), sought to eliminate all vestiges of the Lamarckian 'acquired habits' view of instincts that remained in scientific and popular discussions of the topic, including among those heavily influenced by Darwin such as Ernst Haeckel, Herbert Spencer, and Wilhelm Wundt. For Wallace (1875) and Weismann (1889), instincts may, in many cases, be perfected in the course of experience (by the exercise of imitation, memory, or observation) but they are rooted in heritable physiological structures. The modifications that an organism undergoes in life do not directly pass to later generations. Instead, spontaneous genetic variations in organic structure predispose some individual organisms to develop more adaptive behavioral tendencies and those more adaptive 'predispositions of the germ,' as Weismann termed them, are more likely to be inherited. Indeed, Weismann's theory of the 'germ plasm' as the mechanism for inheritance is key to the so-called 'modern evolutionary synthesis.'

Wallace further proposed that the notion of instinct should be used only where it can be conclusively demonstrated that no instruction or observation is necessary for the performance of some complex act. On the basis of this criterion, he argued that most of what was attributed to instinct, including classic examples such as bird nesting, can actually be shown to be dependent on a combination of learning, simple physiological reflexes, and the structure of the local environment. Wallace, along with Haeckel (1834–1919), T.H. Huxley (1825–95), Lewis H. Morgan (1818–81), and others, criticized instinct as a scientific explanatory device, because of its ambiguous and multiple definitions and because it seemed to rely on a 'metaphysical' notion of motive 'impulse,' rather than on observable faculties to explain behavior. As a supposed 'first cause' of behavior, it seemed to stifle inquiry into the nature and causes of instincts, themselves. J. Mark Baldwin (1861–1934), C. Lloyd Morgan (1852–1936), and others, considering the growing number of experimental studies on animal instinct, argued that their authors should make a principled distinction between biological evolution passed through the germ plasm, and 'social heredity,' in which behaviors are passed directly from one individual to another by imitation, instruction, or shared experience. This social heredity, too, could be shown to be functional for the survival of a species, and could help explain the immense, variable array of human behavior patterns.

Instinct in Human Psychology

Prior to Darwin's (1871) own detailed application of his evolutionary principles to humans in *The Descent of Man*, Herbert Spencer (1820–1903), Francis Galton (1822–1911), and others had already commented extensively on this topic. Galton's studies of 'hereditary genius,' for example, purported to demonstrate that human intellectual and moral traits are subject to the same laws of inheritance as physical traits, which helped explain the appearance of families of purportedly 'natural' leaders and savants. Spencer too argued for the inheritance of 'psychical' as well as physical characteristics in his pre-Darwinian (1855) *Principles of Psychology*, which laid the groundwork for one of the most influential understandings of instinct in human psychology. Considered physiologically, instincts are 'compound reflex actions,' according to Spencer, and are formed in the same way as habits: over the course of experience the frequent co-occurrence of certain simple reflexes leads to their coordination, such that complex actions tend to unfold more automatically with repetition. In this way, he developed a coherent physical theory that linked simple neuromuscular reflexes with complex instinctive behaviors and provided an account of instincts that did not rely on teleological mental faculties, such as will or consciousness. Indeed, he argued that intelligence, itself, developed by the further coordination of complex nerve pathways and was therefore a consequence of, rather than in opposition to, instinct. In animals with higher intelligence, the modes of behavior are not stereotyped wholes but rather separable bundles of reflexes that could be recombined in ways that allowed for intelligent adaptability to variable surroundings. This 'chained reflex' understanding of instincts was further developed by biologists Jacques Loeb (1859–1924), H.E. Ziegler (1858–1925), and others and became central to later Behaviorist criticisms of the utility of the instinct concept.

Authors considered the founders of modern human psychology, especially William James (1842–1910) and Wilhelm Wundt (1832–1920), built critically upon the work of Spencer and Darwin regarding instinct. James (1890) viewed instincts as the 'functional correlates' of an organism's 'structure.' In this view, instincts are 'impulses' that naturally but implicitly tend toward expression but which do not have only one fixed goal or form of release. Yet he also favored Spencer's physiological interpretation, which classed instincts with reflexes instead of by means of the purposes they appear to serve to the observer. James' theory consequently sought to combine the seemingly opposed 'mental impulse' and 'physiological reflex' views into a single 'functional' statement.

If, in James' view, each separate impulse is also a coordinated reflex, then there must be an immense number of so-called 'instincts' to account for all the adjustments that complex organisms make to stimuli in their environments. And indeed, contrary to Darwin and Cuvier, James proposed that humans have a greater variety and number of instincts than any other animal, a point on which Spencer and Wundt agreed. These numerous impulses often come into conflict with one another in conduct, leading humans to a life of "hesitation and choice," an "intellectual life," not because they have fewer instincts but because they have so many that they inhibit the expression of one another. James proposed one of the earliest lists of definite human instincts (following William Preyer and G.H. Schneider in this regard), which included actions that definitely require learning (e.g., walking) alongside apparent simple reflexes (e.g., suckling or biting). James' eclectic list emphasized the native diversity of impulses in human action. Additionally, James' and Spencer's emphasis on the inability to physiologically distinguish instinct from habit, in effect, divorced the psychology of behavior from a definite evolutionary mechanism. While James favored the 'natural selection' view of Darwin, other leading psychologists, including Wundt and Theodule-Armand Ribot (1839–1916), drew the opposite conclusion. The phenomena of instinct, they argued, are consistent with a notion of inheritance by 'acquired habits.'

Wundt (1894) further emphasized the inadequacy of a purely mechanical view of human instinct, stemming from Spencer. Like James, he preferred the notion of 'impulse' or 'drive' (German: *Trieb*) to 'instinct' in human psychology; and more than James he emphasized the ways in which the expression of impulse was regulated by its orientation to goals and its association with affective states. Such impulses are, thus, centrally important motive forces in Wundt's 'voluntaristic' psychology. Other psychologists of the late nineteenth and early twentieth centuries, including C. Lloyd Morgan, L.T. Hobhouse (1864–1929), and G.T. Ladd (1842–1921), also elaborated what some called the 'appetitive' nature of instincts. In this view, instincts are not merely the sum of a chain of reflexes but are also a particular state of the organism that persists through the action, and without which the reflexes would not be discharged by appropriate stimuli or combined effectively.

A similar view was also espoused by prominent 'vitalist' philosophers of this period, including Hans Driesch (1867–1941) and, to a certain extent, Henri Bergson (1859–1941). Driesch (1908), in particular, insisted on the specificity of instincts in contrast to simple reflexes and habits. Instincts are peculiar in that, although they may in some cases be improved through repetition, they are complete from their first expression. In addition to being composed of chained reflexes, instincts are impelled toward a more-or-less definite end condition (e.g., a finished spider web) and their course persists even in the face of obstacles that require additional effort or adjustment of operations. For Driesch, these findings provided some of the strongest evidence for the existence of a 'vital principle' at work in the organism that could not be explained on purely mechanical principles. Much of this psychological literature on instinct as a motive force drew upon a broader tradition regarding the irrational sources of human behavior. Driesch, for example, drew inspiration from Schopenhauer, Nietzsche, Eduard von Hartmann, and others. However, in the subsequent history of the human sciences, the most influential and controversial author on these questions has been Sigmund Freud (1856–1939).

Most current understandings of Freud's 'drive theory' do not equate it with a 'biological instinct' theory. This confusion persists in English, however, because influential translations of Freud's work have used the word 'instinct' to translate his notion of 'drive' (*Trieb*). Drives, in Freud's theory, are not prefixed in form or object. Instead, he sought to explain the intricate and problematic processes involved in dynamically structuring these energies, including the formation of repressions and sublimations and the sequences of attachments to external- and self-objects. Freud modified his theory of drives several times, most notably in *Beyond the Pleasure Principle* (1922), where he proposed that in addition to the 'libido' there is also a fundamental 'death instinct' or 'death drive,' that is necessary to explain the persistent enactment of unpleasurable experiences. Most schools of psychoanalysis that have been influential in the social sciences, especially ego-psychology, social psychiatry, self-psychology, and post-structural psychoanalysis, for all their disagreements, share an emphasis on the necessity of structuring the otherwise indeterminate drives, and hence, a largely 'non-instinct' view of Freud's theory of drives.

Kurt Koffka (1886–1941) also proposed a *Gestalt* theory of instincts, which drew upon existing 'appetitive' theories but sought to avoid their 'vitalistic' or 'mentalistic' implications. In Koffka's (1924) view, instincts are continuous courses of movement oriented to an end as well as regulated by a stimulus, and in this way they form 'articulate wholes.' Instincts, unlike mere reflexes, actively seek stimulation and that stimulation is given not primarily from the external environment but from the tensions or needs of the organism itself. Koffka attempted to elaborate a theory that could account for the forward impulse of instincts without implying a guiding will or consciousness, by arguing that they constitute 'open' or 'transitional' situations in the dynamic equilibrium of the organism that continue to operate until they reach an end-state of 'closure,' perhaps prefiguring a more cybernetic or systems-theory view of instincts. After the initial positive reception this theory had little further influence, at least in part, because of the growing body of criticisms of the instinct theory in the 1920s.

Instinct in the Social Sciences

In the social sciences of the late nineteenth century, influential early followers of Darwin placed emphasis on the role of

instinct in explaining the origins of complex human institutions and behaviors. Galton's (1883) observations of herd animals and human crowds led him to claim that humans possess a fundamental 'gregarious' or 'social' instinct that is at the origin of social behaviors. This gregarious instinct, he proposed, had evolved to ensure mutual security in social groups, but in exchange for this stability it tended to retard cultural innovation and progress. Galton argued that these instincts may not be perfect guides to action in modern society and could actually be dysfunctional, pointing to what he saw as the general servility of the vast majority of people. In his view, human instincts were adapted to a precivilizational state and had not had time to adapt to the conditions of industrialization. Galton's conclusions were closely tied to his advocacy of eugenics and he argued that measures should be taken to ensure that the individuals most adapted to advanced civilization, reproduce for the betterment of human progress.

Spencer and others took a somewhat more charitable view in proposing that human instincts, because they are subject to the same laws of evolution as other traits, could be shown to vary in determinate ways across races and historical epochs, according to the dominant conditions to which the human groups had to adapt. Other authors, including Pyotr Kropotkin (1842–1921), Henry Drummond (1851–97), and John Fiske (1842–1901), even argued that social instincts represent a separate evolutionary 'factor,' which shifts selection pressures to the level of the social group rather than the individual organism and are, therefore, productive of continued social development. Spencer, of course, also proposed that human society was analogous to a natural organism, and he went so far as to suggest that something analogous to 'self-preservative instincts' are present, not only in individuals who make up social institutions, but in those institutions themselves. Indeed, in his *Principles of Biology* (1864), Spencer coined the term "survival of the fittest" to describe the working of natural selection, and he made clear in his *Principles of Sociology* (1898) that he considered this a principle that operated at the social as well as the individual level.

Virtually all of these 'Social Darwinist' authors sought to show that human morality could be attributed to the influence of social instincts. In his *The Descent of Man*, Darwin (1871) posited that, although the particular features of morality are learned, the seed of morality is found in the instincts that lead humans to take pleasure in association, to feel sympathy with others, and to provide services for one another. In human conduct there are situations, Darwin argued, in which one instinctive tendency temporarily dominates others; yet memory allows instincts that go unsatisfied to return to the individual's awareness, and the outstanding human capacity for language allows unsatisfied impulses to be expressed and judged communally. Over the course of experience, humans develop self-command to prevent over-expression of self-preservative instincts at the expense of others, and in this way moral conscience arises. Further, Darwin argued that if the source of moral behavior is the 'social instincts,' then behavior need not be explained by means of hedonistic or utilitarian ethics that sought to explain each 'altruistic' action by the pleasure (or avoidance of pain) that it supposedly must give to the individual. Spencer, Haeckel, Karl Pearson (1857–1936), and others sketched out similar theories, acknowledging the variability of human morality by their emphasis on the shaping of instincts by progressively more complex human institutions.

The authors now recognized as founders of social and cultural anthropology, including E.B. Tylor (1832–1917), Franz Boas (1858–1942), and Emile Durkheim (1858–1917), emphasized the variability and adaptability of human customs in contrast to instinct theory. For example, in his foundational *Primitive Culture*, Tylor (1871) argued that phenomena, such as bow hunting or stone tool making, should be explained by selection and gradual improvement of the objects available to humans in their local environments and by imitation of others, rather than by hunting or tool-making instincts. In opposition to this growing ethnological literature, other evolutionists such as Edvard Westermarck (1862–1939) argued that the true motivations for social conduct come from underlying instincts and sentiments. Ignoring these motive powers, Westermarck (1891) argued, leads to a problematic, 'mechanical' treatment of human customs because it seeks to explain social phenomena primarily by social antecedents (a methodology that Durkheim explicitly championed). Consequently, Westermarck's *The History of Human Marriage* relied heavily on notions of parental and sexual instincts to explain the biological foundations of human marriage customs.

The notion that collective behaviors could be explained by attributing them to more-or-less definite instincts was hugely influential in the social sciences of the late nineteenth and early twentieth centuries. Gustave Le Bon (1841–1931), Wilfred Trotter (1872–1939), and others, for example, used notions of instinct to explain the irrational behavior of crowds, and other authors sought to provide an ultimate ground for social behavior in quasi-instinctive, general tendencies to 'imitation,' 'suggestion,' or 'consciousness of kind.' Theories of 'aggressive' or 'pugnacious' instincts proliferated as purported explanations of World War I, although much of this literature was criticized by later social psychologists for the seemingly arbitrary or unprincipled use of 'instinct' as an explanatory device. In economics, authors including Thorstein Veblen (1857–1929), E.A. Ross (1866–1951), and Carleton H. Parker (1878–1918) drew on instinct theory to criticize the structure of industrial production (although they, too, tended to rely on a somewhat metaphorical notion of instinct). They argued that modern economic organization does not permit the greater number of workers sufficient opportunities for the expression of natural impulses such as creativity. Earlier critical anticipations of this argument came from utopian socialist Charles Fourier (1772–1837), who envisioned an emancipatory social order centered on the fulfilling expression of natural impulses for all, and anarchist Mikhail Bakunin (1814–76) who claimed, in opposition to Marx, that men's instinctive need for freedom would impel the masses to social revolution.

The author who sought to ground the social sciences most thoroughly in an analysis of instinct, and who was consequently a frequent target of criticism, was William McDougall (1871–1938). In *An Introduction to Social Psychology* and subsequent works McDougall (1908) argued that social science must be based on an adequate theory of the impulses and motives that sustain and regulate conduct, the so-called "springs of human action." He sought to prevent the social sciences from making the 'ad hoc' assumptions about the nature of the mind or over-generalizing from a particular set of human

motivations as authors had done in the past, and instead to elaborate a synthetic "dynamic, functional, voluntaristic" approach consistent with comparative and evolutionary natural history. He argued that the 'reflex' understanding of instincts only took account of the physiological movements to which instincts give rise, without explaining the 'mental processes' of instincts, themselves. Building from the 'appetitive' or 'impulse' view and from 'faculty' psychology, he argued that instinctive behaviors must be understood to involve cognition of objects, emotion or feeling with regard to the objects, and conative striving toward or away from those objects that regulate the onset and course of behaviors. Thus, the ultimate impetus and end of all human activity, even the most 'passionless' thought, may be found in these instinctive impulses if defined in McDougall's encompassing fashion. McDougall listed what he considered the 'principal' human instincts, which varied somewhat across his work but typically included flight, repulsion, curiosity, pugnacity, self-assertion, self-abasement, parental, reproductive, gregarious, acquisitive, and constructive instincts. Each is directly connected with identifiable emotional states and actions. McDougall explained the vast array of human behavioral tendencies as resulting from the ramification and organization of these instincts in the course of social conduct. His work was drawn upon by a wide variety of writers in the social sciences, psychology, education, and ethics in the early twentieth century, because it appeared to give a principled and comprehensive theory of human motivation.

Criticisms and Alternatives

Among the most prominent early critics of instinct theory in the human sciences were the Behavioral psychologists. In the late 1910s and early 1920s, critiques of instinct were published by Knight Dunlap (1875–1949), Z.Y. Kuo (1898–1970), E.C. Tolman (1886–1959), and Ivan Pavlov (1849–1936), who elaborated themes from the earlier work of Spencer, James, Loeb, and others. Dunlap's work (1919, 1922), in particular, asserted that lists of specific instincts do not represent fundamental, substantively distinct, psychological processes. Instead, these categorizations vary according to the given author's judgment of the classes of 'ends' served by the behaviors (e.g., if a behavior seems to be connected with acquiring sustenance, then it is supposed to result from a feeding instinct). Any list that suits the author's purposes is as valid as any other and no list could be proven to provide a mutually exclusive and exhaustive classification of the observed organism's behaviors. The most systematic analysis of this point was in sociologist L.L. Bernard's (1924) monograph study, which demonstrated from an examination of discussions of instinct in 500 scientific works that no general consensus could be derived.

Second, the observable phenomena explained by instinct could also be explained by notions of 'reflex' or 'habit,' which these authors thought had better claims to being scientifically rigorous concepts. Explaining a behavior as 'instinctive' had the disadvantage of mentalistic or even metaphysical connotations. Any observed behavior was, according to the Behaviorists, obviously the product of both acquired and inherent capacities and a focus on the inherent factors prevents inquiry into the learned and proximate antecedents of the behavior (especially the "conditioning of reflexes" which was the focus of much Behaviorist research). Although John B. Watson (1878–1958) is identified as the primary opponent of instinct theory in many subsequent discussions, he developed his adamant anti-instinct stance as a result of his detailed studies of human infants in the 1920s with Rosalie Rayner, later his wife (1898–1935), and in response to the criticisms of other Behaviorists. Indeed, in his early work on animal psychology, Watson had even proposed his own classification of definite instincts. B.F. Skinner (1904–90) later added to the Behaviorists' critique by arguing that instinct was at best an 'explanatory fiction' or tautology, in that it 'explained' observed behavior by inventing an unobserved faculty or force that tended to produce the corresponding behavior (e.g., humans fight because of an instinct to fight).

An overlapping set of criticisms came from the social science literature of the same period, led by Ellsworth Faris (1874–1953), C.E. Ayres (1891–1972), Charles H. Cooley (1864–1929), John Dewey (1859–1952), and L.L. Bernard (1881–1951). According to these authors, the profound diversity of human customs concerning even such fundamental biological processes as nutrition or procreation, as witnessed in the growing record of cross-cultural studies, could not be explained by a finite set of native instincts. Cooley (1922) in particular, posited that if scientists begin not with the assumptions of individualist psychology but from the premise that human nature is essentially social, there is no need to postulate particular instincts as the special motives of collective behavior. Then, to the extent that there are apparent internal drives motivating human behavior, they could be shown to be products of the personal histories of interactions between individuals and their social environments. Faris and Bernard also pointed out that recourse to instinct tends to mean explaining behaviors in broad, general classes rather than accounting for individual temperamental differences, and such explanations have controversial ethical and pedagogical implications, including an affinity with theories of 'racial characteristics' and the 'standardization' of educational practices. In place of instinct theory, social psychology was dominated from the 1930s by the notion of 'attitude,' especially in the United States, which emphasized explanation by concrete, personalized motives that were more easily recorded, analyzed, and explained, than classes of abstract, uniform instincts.

Since the 1930s, instinct theories have fallen out of favor as explanations in the behavioral and social sciences, while alternative understandings of human motivation have developed. In addition to notions of conditioned reflexes, attitudes, and drives, these include work on the internalization of social expectations; cybernetic feedback theories of behavioral systems; concepts of bounded rationality and uncertainty in decision making; phenomenological or contextualist theories of intentionality; and many others. Alternative vocabularies, without a notion of instinct, have also developed to conceptualize the relationship between the body and social conduct, including terms such as 'body-politics,' 'practice theory,' and 'life-course theory.'

Still, several attempts have been made to rehabilitate a notion of instinct in explaining human behavior. Most

notable among these is the work of 'ethologists,' especially Niko Tinbergen (1907–88) and Konrad Lorenz (1903–89). Tinbergen (1951) and Lorenz (1957) drew upon their work on the native behavioral capacities of lower vertebrate and invertebrate animals to add a new level of sophistication to 'appetitive' theories of instinct. Although humans do not have the same reliance on 'fixed action patterns' as do other animals, they may have some of the same instinctive mechanisms, such as 'imprinting,' that would contribute to the explanation of how human behaviors are learned and develop.

Indeed, the ethologists influenced some work in 'evolutionary anthropology' regarding apparently innate human behavioral tendencies such as territoriality and aggressiveness, as well as work in 'sociobiology' on the biological basis of human prosocial behaviors. In psychology, John Bowlby (1907–90) articulated an influential 'attachment theory,' which draws significantly from the ethologists in arguing that connections between caregivers and infants are based evolutionarily on imprinting like mechanisms. In this work, he noted that as an adjective, 'instinctive' may be useful as a description, even while 'instinct,' as a noun denoting definite behaviors, is problematic. Likewise, Abraham Maslow (1908–70) developed a theory of the 'hierarchy of needs,' positing that human needs are 'instinctoid,' while also arguing that human behavior expresses resistance to, and control of, such impulses in pursuit of higher goals. Indeed, calls to take into consideration the biological nature of human behavior typically have not sought to supplant a one-sided social determinism with biological determinism, but rather to bring these factors together into a broader synthesis.

Another recent literature has developed around theories of psychological 'nativism,' especially the apparently innate aspects of human perception and language capacities. Steven Pinker (1994) explicitly posited that humans possess a unique 'language instinct,' on the basis of the work of Noam Chomsky and others regarding the 'universal grammar' posited to underlie all human language. Still, the growing literature in cognitive science has tended to eschew a notion of 'instinct' and instead has elaborated detailed neurobiological and functional explanations of cognition and behavior. Other lines of investigation have introduced yet another set of explanatory principles based around 'genomics,' and increasingly, 'epigenetics,' to explain the heredity of human behavior and the relationship between biological organisms and the environment. Although these new fields may ultimately revive interest in theories of instinct, they may also have the effect of supplanting that older literature with an entirely different language, viewed as more scientifically productive.

See also: Darwin, Charles Robert (1809–82); Galton, Sir Francis (1822–1911); Habit: History of the Concept; James, William (1842–1910); Lorenz, Konrad (1903–89); Motivation: History of the Concept; Psychology, History of (Twentieth Century); Sociology, History of; Spencer, Herbert (1820–1903); Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Bernard, Luther L., 1924. Instinct: A Study in Social Psychology. Henry Holt, New York.
Cooley, Charles H., 1922. Human Nature and the Social Order, revised ed. Charles Scribner's Sons, New York.
Cuvier, George, 1817–1829. The Animal Kingdom: Arranged in Conformity with Its Organization (Edward Griffith, Trans.). Whittaker, Treacher, and Co., London.
Darwin, Charles, 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. John Murray, London.
Darwin, Charles, 1871. The Descent of Man, and Selection in Relation to Sex. John Murray, London.
Driesch, Hans, 1908. The Science and Philosophy of the Organism: The Gifford Lectures Delivered before the University of Aberdeen in the Year 1907. Adam and Charles Black, London.
Dunlap, Knight, 1919. Are there any instincts? Journal of Abnormal Psychology 14, 307–311.
Dunlap, Knight, 1922. The identity of instinct and habit. Journal of Philosophy 19, 85–94.
Freud, Sigmund, 1922. Beyond the Pleasure Principle (C.J.M. Hubback, Trans.). International Psycho-Analytical Press, London; Vienna.
Galton, Francis, 1883. Inquiries into Human Faculty and Its Development. Macmillan and Co., London.
James, William, 1890. The Principles of Psychology. Henry Holt, New York.
Koffka, Kurt, 1924. The Growth of the Mind: An Introduction to Child-Psychology (R.M. Ogden, Trans.). Harcourt, Brace & Company, New York.
Lamarck, Jean Baptiste de Monet de, 1807 [1914]. Zoological Philosophy: An Exposition with Regard to the Natural History of Animals (Hugh Elliot, Trans.). Macmillan and Co., London.
Lorenz, Konrad, 1957. The conception of instinctive behavior. In: Schiller, C.H. (Ed.), Instinctive Behavior: The Development of a Modern Concept. New York: International Universities Press, Inc., New York, pp. 129–175.
McDougall, William, 1908. An Introduction to Social Psychology. Methuen & Co., London.
Paley, William, 1802. Natural Theology: Or, Evidences of the Existence and Attributes of the Deity. R. Faulder, London.
Pinker, Steven, 1994. The Language Instinct. W. Morrow and Co., New York.
Spencer, Herbert, 1855. The Principles of Psychology. Longman, Brown, Green and Longmans, London.
Spencer, Herbert, 1864. Principles of Biology, vol. I. Williams and Norgate, London.
Spencer, Herbert, 1898. Principles of Sociology, vol. II. D. Applegate and Company, New York.
Tinbergen, Niko, 1951. The Study of Instinct. Clarendon Press, Oxford.
Tylor, Edward B., 1871. Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Art, and Custom. John Murray, London.
Wallace, Alfred, R., 1875. Contributions to the Theory of Natural Selection: A Series of Essays. Macmillan and Co., London.
Weismann, August, 1889. In: Poulton, E.B., Schönland, S., Shipley, A.E. (Eds.), Essays upon Heredity and Kindred Biological Problems. Clarendon Press, Oxford, pp. 251–332.
Westermarck, Edvard, 1891. The History of Human Marriage. Macmillan and Co., London.
Wundt, Wilhelm, 1894. Lectures on Human and Animal Psychology (J.E. Creighton, E.B. Titchener, Trans.). Macmillan and Co., New York.

Institutionalism

Sven Steinmo, European University Institute, Fiesole, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of political institutions has moved back into center stage in political science. The study of political institutions and their consequences for political choice and development has moved to center stage in the study of politics. The article first surveys the history of institutionalist thought, then explores several different types of institutionalist theory, and, finally, suggests the central analytical agendas of modern institutionalist theory. The article examines how and why both history and institutions matter for politics and emphasizes the importance of 'ideas' for understanding political change.

The study of political institutions has moved back into center stage in political science. Whereas only a few years ago, institutions were mostly casually mentioned in most political science research, today they are a central focus of attention. This change is not simply a change of language or the catching on of a new popular phrase or academic fad. Instead, these changes represent an important development in evolution of political science theory and intellectual focus.

The following article surveys the history of institutionalist thought, explores several different types of institutionalist theory, and suggests the central analytical agendas of modern institutionalist theory.

What Are Institutions?

In the broadest sense, institutions are simply rules. As such, they are a foundation for all political behavior. Some are formal (as in constitutional rules), some are informal (as in cultural norms), but without institutions, there could be no organized politics. Indeed, absent institutions, and there could be no organization at all. To understand this point, simply attempt to consider a world in which there were no rules: In this Hobbesian hell, individuals would be forced to 'invent' communication every time they encountered another individual. In this sense, then, if we study social interaction, we study institutions. This does not suggest, however, that all social scientists are 'institutionalists.'

Institutionalism

The 'institutionalism' specifically examines the ways in which institutions structure social and political behavior (North, 1990). This burgeoning body of literature argues that policy, politics, and behavior can only be understood in the context of the institutions in which they take place. Thus, for example, Ellen Immergut argues that variations in national health insurance systems are best explained by variations in national political institutions (Immergut, 1998). Similarly, Bo Rothstein shows that Sweden's high union density is best explained by the 'Ghent' unemployment insurance system, which gives workers powerful incentives to join Swedish unions (Rothstein, 1992). Even more broadly, Douglas North suggests that the very success of Western political economic model is rooted in the peculiar institutions developed in these societies (North, 1990).

The central tenet of this new institutionalism is that institutions are not neutral to policy outcomes. As Peter Hall has suggested, "On the one hand, the organization of policy-making affects the degree of power that any one set of actors has over policy outcomes…. On the other hand, organizational position also influences an actor's definition of his own interests…in this way, organizational factors affect both the degree of pressure an actor can bring to bear on policy and the likely direction of that pressure" (Hall, 1986: p. 12).

In sum, institutions define the rules of the political game, and as such they define who can play and how they play. Consequently, they ultimately can shape who wins and who loses. If politics is the study of who gets what, when, and why, then institutionalists argue that institutions should be at the heart of that study.

A Brief History of Institutionalism

Institutionalism has a long-established tradition among those interested in politics and political outcomes. Plato's *Republic* is a comparative study of institutions. Similarly, Aristotle's central concern in *Politics* is which kinds of political institutions will produce the best outcomes. James Madison must clearly also be seen as an early American 'institutionalist' in that he was specifically concerned with which kinds of institutions would produce the best political outcomes and how the specific design of institutions would shape political outcomes.

Political scientists have also long been interested in institutions. Indeed, in its early years, political science meant the study of political institutions (Wilson, 1891). But, with some important exceptions (Herring, 1940; Key, 1947), early political science was often more descriptive than analytical. 'Comparative politics,' in particular, consisted mostly of detailed configurative studies of different legal, administrative, and political structures. In the immediate post-War years, a new generation of political scientists attempted to make the study of politics more 'scientific.' For many, this effectively meant that political science ought to model itself on the 'hard sciences,' which they believed was fundamentally a deductive process. Thus, rather than studying the details of political life and

inductively uncovering the patterns of behavior and action, political 'science' should be a deductive science that seeks to discover of the general laws and fundamental forces that lie behind political action. Focusing on particular institutions, proponents of this intellectual agenda implied, was 'atheoretical.' Concomitant with the push for more abstract laws, political scientists were disillusioned by the failure of parliamentary institutions in inter-War Weimar Germany (and later in postcolonial Africa) to prevent these polities from devolving into authoritarianism. Clearly, many argued, there were bigger, more important, indeed more fundamental forces at work in politics and development than political institutions. These forces, they argued, should be the focus of political science.

Thus, behavioralist, functionalist, and Marxist perspectives took leading roles in political science theory-building through most of the 1960s and 1970s. As a consequence, institutional analysis diminished in prominence, particularly in comparative politics. For functionalists, political institutions were simply seen as organizations created to fulfill the systemic needs of society. As such, political institutions were neither important nor interesting. For behavioralists and Marxists, political institutions were simply arenas in which political battles were fought out. While Marxists saw the relevant groups as classes and Pluralists saw them as more narrowly defined, neither theoretical perspective paid particular attention to the structure or character of the institutions themselves. The real meat of politics, they argued – and the keys to understanding political outcomes – was found in the articulation of group interests. Though this was rarely explicitly stated, implicit in these theories was the assertion that if politics or policies differed between societies, this difference was surely the result of different constellations of group and class interests or preferences (Almond and Verba, 1963; Verba, 1967; Miliband, 1969).

It was not the case, however, that all political scientists had abandoned the study of institutions. Indeed, many of the most widely read scholars in American politics maintained an explicitly institutional emphasis (Schattschnieder, 1960; Greenstein, 1963). In comparative politics as well, several political scientists continued to examine political institutions and their effects on political outcomes (Ekstein, 1960; Bendix, 1964; Huntington, 1968), even while they sometimes had to defend their 'inductive' approach from the challenge that it was not 'scientific' (Przeworski and Teune, 1970). For these scholars, it was self-evident that if one wanted to understand what government does, one needs to specifically study the institutions through which it acts.

It was probably in comparative politics that the search for 'grand theory' had its most significant impact. It may be for this reason that a self-conscious return to 'institutionalism' was most forcefully articulated here. Among the first group of scholars to move in this direction was Peter Katzenstein and his colleagues, who sought to explain why several advanced capitalist states responded so differently to the oil shock of the 1970s. The answer, they concluded, was found in the differing institutional structures in these polities and the consequent patterns of economic policy pursued in each nation (Katzenstein, 1978). Similarly, Theda Skocpol's study of social revolutions also concluded that one could not explain the course of a country's revolution without examining the nature and structure of the state against at which these revolutions were aimed (Skocpol, 1979).

Soon a new 'state-centered' approach emerged in comparative political inquiry. These scholars forcefully argued against the behavioralist and Marxist 'grand theory' emphasis, then dominant in the study of comparative politics. Instead, they suggested, political outcomes were shaped and structured by the specific actors and their position in the state. One should not treat the state as a neutral 'black box' through which group or class interest was simply translated (Evans et al., 1985). Instead, the state had independent interests and agendas, which were separate and different from the interests and preferences of classes and interest groups that made up society.

Of course, it did not take long for these scholars to discover that 'the state' is too broad a concept and too varied a set of institutions to be 'taken seriously' without being broken down. To make analytic sense out of the insights, scholars began examining institutions more carefully. If state institutions 'matter,' they argued, then why should this not be equally true of institutions outside the boundaries of the executive state? With these questions, 'the new institutionalism' was born.

Three Types of Institutionalism

Today, three different intellectual approaches lay claim to the term 'institutionalist' (Hall and Taylor, 1996). Each grows out of a different academic discipline and attempts to integrate these different analytic traditions into the understanding of politics. Sociological institutionalism, as its name implies, grows out of sociology and the study of organizations (Selznick, 1949). These scholars have been centrally interested in understanding culture and norms as institutions. These scholars emphasize 'folkways,' 'patterns of behavior,' and 'cognitive maps' and argue that these social institutions are critical for understanding the structure of social, political, and economic interactions (March and Olsen, 1989; DiMaggio and Powell, 1991). Building on upon their analyses of complex organizations, these scholars show the relationship between formal institutions and the structure or patterns of behavior and beliefs. They argue that these informal institutions are core to any understanding of the nonrational aspects of human communication and exchange.

'Rational choice (RC) institutionalism,' taking its lead from economics, emphasizes quite different sets of institutions and their effects. RC scholars attempt to apply the formal logic and methods to the study of politics and history and often explicitly eschew the more 'mushy' or less precise variables such as norms and beliefs. Instead, basic assumptions are made about human behavior and motivation in order to uncover and examine the basic laws of political behavior and action. Scholars in this tradition argue that once these laws are discovered, models can be constructed that will help us understand and predict political behavior (Levi, 1988). In their deductive model, RC scholars look to the real world to see if their model is right (test the model). For these scholars, understanding real outcomes is not the first point – creating, elaborating, and refining a theory of politics is (Weingast, 1996).

The implications of this scientific orientation are substantial. Morris Fiorina, a highly regarded RC scholar at Harvard,

put the issue in the following way: "most PTI scholars are not as interested in a comprehensive understanding of some real institution or historical phenomenon, so much as in a deeper understanding of some theoretical principle or logic … [F]or most PTI scholars, breadth trumps depth; understanding 90 percent of the variance in one case is not as significant an achievement as understanding 10 percent of each of nine cases, especially if the cases vary across time and place" (Fiorina, 1995: pp. 110–111).

The third 'new institutionalist' approach emerges out of what might be considered a more traditional political science. Consequently, it has a quite different aim: Historical institutionalists are primarily interested in understanding and explaining real-world events and outcomes. As with the other approaches noted above, scholars working in this tradition also argue that one cannot explain particular historical outcomes without specifically examining the way in which the political institutions have shaped or structured the political process (Steinmo et al., 1992). But, unlike RC scholars, in particular, historical institutionalists came about their 'institutional' arguments inductively after testing a variety of alternative theories (i.e., Marxist, structural functionalist, culturalist, and rationalist) against the outcomes they observed. In other words, historical institutionalists are first interested in explaining an outcome (say, e.g., why France and Britain have pursued such different styles of industrial policy (Hall, 1986) or why some welfare states generate more popular support than others (Rothstein, 1998)); they then proceed to explore alternative explanations for the outcomes they observe.

Historical institutionalists do not argue that institutions are the only important variables for understanding political outcomes. Quite the contrary, these scholars generally see institutions as intervening variables (or structuring variables) through which battles over interest, ideas, and power are fought. Institutions are thus the focal points in critical junctures in a historical path analysis because political battles are fought inside institutions and over the design of future institutions.

Taking *History* Seriously

These insights had important implications, both for what we study and for how we study it. Historical institutionalists study history because they believe history matters, not merely to increase the reference points for analysis (as is done in time series analysis). There are at least three important ways in which history matters. First, political events happen within a historical context, which has a direct consequence on the decisions or events. An early example of this is the seminal work of Alexander Gerschenkron, who argued that *when* a country industrializes necessarily affects *how* it industrializes. He shows us why latecomers cannot go through the same long trial-and-error process followed by early developers. (An example outside politics may prove illustrative. Many of us recognize that firstborn children have a very different developmental experience than second (or later) children. Not only are the parents more experienced after the first child, they are also taking care of more than one child at a time. Finally, and equally importantly, subsequent children grow up in a home where there are older siblings – something the first child, by definition, cannot do.) In other words, the process of industrialization is essentially different for late developers than for early developers. This is a huge insight that is easily missed in large-scale quantitative, cross-national comparisons, which very often pool data across continents and time periods and treat the time/place as inconsequential (or assume that it will 'wash out' of the analysis).

The second reason why history matters is that actors or agents can learn from experience. Historical institutionalists understand that behavior, attitudes, and strategic choices take place inside particular social, political, economic, and even cultural contexts. Rather than treating all political actions as if they are fundamentally the same, irrespective of time, place, or context, historical institutionalists explicitly and intentionally attempt to situate their variables in the appropriate context. Thus, by deepening and enriching their understanding of the historical moment and the actors within it, they are able to offer more accurate explanations for the specific events that they explore than had they treated their variables outside the temporal dimension.

E.E. Schattschneider's early work on tariff policy showed how political choices made at time A have important consequences for time B. In this work, he famously argued that "new policies create new politics." Following Schattschneider, Paul Pierson has shown in several important works, *how and why* policy choices at one point in time affect choices at subsequent points in time. Similarly, Esping-Andersen pointed out in his seminal *Three Worlds of Welfare Capitalism*, how, given the fact that we live in modern welfare states with unemployment insurance, health insurance, pension programs, and the like, the existence of the welfare state is a fact of modern political life that itself *shapes* politics, expectations, and policy in the countries that have developed it (Esping-Andersen, 1990).

Finally, again as Pierson has shown, expectations are also molded by the past (Pierson, 2004). While some might point to America's adventure in Iraq as a simple product of power politics and/or the demand for oil, a historical institutionalist would more likely look to the patterns of past victories for an understanding of why this country reacted in the way it did to the 9/11 attacks. Certainly they were mistaken, but there should be little doubt that America's past successes in Germany and Japan – to say nothing of their perceived victory over Communism at the century's end – led policy makers in the administration to believe that they could assert American power and bring successful capitalism and democracy to a former dictatorship.

In sum, for historical institutionalists, *history is not a chain of independent events*. There is more than the temporal dimension implied in this basic point. Taking history seriously ultimately means that the scholar is skeptical of the very notion of variable independence. Instead, acknowledging the importance of history suggests an explicit awareness that important variables can and often do shape one another. Historical institutionalists, more than political scientists in some other traditions, are explicitly interested in these interactive effects on the interdependence of multiple causal variables.

Historical institutionalists are something like an environmental biologist, who believes that in order to understand the specific fate of a particular organism or behavior, he or she

must explicitly examine that organism or its behavior in the ecology or context in which it lives. This implies a different scientific ontology than that commonly found in the hard sciences of physics and chemistry. At the root of evolutionary biology is the assumption that the objects of analysis – living organisms – are fundamentally different from inanimate matter. While objects in the physical world often adhere to constant 'laws' of nature, biological organisms often defy attempts to reduce them to their essential components because of their complexity. Thus, as eminent evolutionary biologist Ernst Mayr points out, the development of biology as a science has required an investigation of 'additional principles' that applied only to living organisms. He argues, "This required a restructuring of the conceptual world of science that was far more fundamental than anyone had imagined at the time" (Mayr, 2004: p. 26).

Historical institutionalism represents something like this ontological move in social science. In order to understand historically specific events and long-term political outcomes, one could not strictly apply methods and epistemologies drawn from the study of invariant variables that have fixed relationships across space and time. This, of courses, does not mean that it is not science – unless one's definition of science would exclude biology as well – rather, it implies that the scientific methods applied should fit the subject being studied.

Ideas, Institutions, and Change

To argue that 'institutions matter' is by now taken as common sense. Today, it is widely recognized that institutions provide stability in political life because they structure political choices. This is because when we know the rules, we are more likely to know how to interact and how to behave. Indeed, when we know the rules, we adapt our behavior accordingly. The more difficult and interesting questions arise when we ask the following: (1) Why do people follow rules? (2) How do we explain institutional change? (3) if institutions are so important, what role is left for human agency and ideas?

Early institutionalist scholars drew insights from evolutionary theory of 'punctuated equilibrium' to explain institutional change (Steinmo et al., 1992). In recent years, however, there has been growing interest in explaining the *endogenous* sources of institutional change. A number of scholars have mapped out different types of institutional change in different contexts (Streeck and Thelen, 2005; Mahoney and Thelen, 2009). These efforts have led to considerable refinement of our understanding of different *types* of institutional change, but are insufficient for explaining *why* institutions change.

The most important contributions to institutionalist theorizing on change have come from those scholars who specifically have tried to examine relationship between human agency and ideas (Campbell, 2002; Lieberman, 2002a; Blyth, forthcoming; North, 2008). Traditionally, institutions (especially those in the RC school) are created by individuals seeking to maximize their individual interests. Ideas, in this view, were simply epiphenomenal or justification for people's 'real' motivation. But such an approach is unsatisfactory, both at the common sense level (we know that human beliefs and ideas matter for our choices) and because these approaches fail to explain how we move from one equilibrium to another.

The interesting questions arise when we look at the interdependent relationships between ideas, institutions, and interests over time. Since none of these factors are static, only through a historical analysis can we learn how they develop, how they change, and why they are so different in different contexts (Lieberman, 2002b). Building on the work of social psychologists such as Amos Tversky and Daniel Kahneman (Kahneman et al., 1982; Kahneman, 2011), institutionalist scholarship is moving toward a more nuanced understanding of the 'social foundations' of human rule following behavior and is attempting to better understand the ways in which human cognitive biases' shape interacts with institutional structures to influence human behavior. Social psychologists have long understood that human beings are social creatures who are not normally individualistic in a meaningful sense. Instead, we are driven to follow social rules and look for social approval. Human beings are more likely to be normative actors, possessing significant cognitive biases, rather than rational self-interested utility maximizers. Appreciating human motivations in this way forces us to consider the interaction of institutional rules, human choices, and eventually, differences in cultural systems.

Finally, institutionalists have begun to argue that we can understand change in evolutionary terms and thus bring some of the insights from evolutionary theory into the study of institutional evolution (Lewis and Steinmo, 2012). Whereas traditional political science has taken a mechanical approach to the study of politics and human affairs (Hall, 2003), an evolutionary approach to institutional change allows one to integrate contingency into the study as well as be more honest with the fact that there are no truly *independent* variables of interest in human affairs. Almost everything that is really interesting and important is instead the product of the complex interaction of multiple factors that interact with each other over time (Streeck, 2010).

See also: Ideologies, Institutions, and the New Institutionalism; Marxism–Leninism: The Ideology of Twentieth-Century Communism; New Institutionalism in the Analysis of Complex Organizations; Old and New Institutionalism in Economics; Pluralism; Sociology and the New Institutionalism.

Bibliography

Almond, G.A., Verba, S., 1963. The Civic Culture; Political Attitudes and Democracy in Five Nations. Princeton University Press, Princeton.

Bendix, R., 1964. Nation-building and Citizenship. University of California Press, Berkeley.

Blyth, M., 2010. Ideas, uncertainty and evolution. In: Cox, R., Belan, D. (Eds.), Ideas and Politics in Social Science Research. Oxford University Press, Oxford, pp. 83–101.

Campbell, J., 2002. Ideas, politics and public policy. Annual Review of Sociology 28, 21–38.

DiMaggio, P.J., Powell, W.W., 1991. Introduction. In: Powell, W.W., DiMaggio, P.J. (Eds.), The New Institutionalism in Organizational Analysis. The University of Chicago Press, Chicago, pp. 1–38.

Ekstein, H., 1960. Pressure Group Politics: The Case of the Medical Association. Stanford University Press, Stanford, CA.

Esping-Andersen, G.S., 1990. The Three Worlds of Welfare Capitalism. Polity, Cambridge.
Evans, P., Rueschemeyer, D., et al., 1985. Bringing the State Back in. Cambridge University Press, New York.
Fiorina, M., Fall 1995. Rational choice and the new (?) institutionalism. Polity XXXVIII (1), 107–115.
Greenstein, F.I., 1963. The American Party System and the American People. Prentice-Hall, Englewood Cliffs.
Hall, P., 1986. Governing the Economy. Oxford University Press, New York.
Hall, P., 2003. Aligning ontology and methodology in comparative research. In: Mahoney, J., Rueschmeyer, D. (Eds.), Comparative Historical Analysis in the Social Sciences. Cambridge University Press, Cambridge, pp. 373–404.
Hall, P., Taylor, R., 1996. Political science and the three new institutionalisms. Political Studies 44, 936–957.
Herring, P., 1940. The Politics of Democracy; American Parties in Action. W. W. Norton & Company, New York.
Huntington, S., 1968. Political Order in Changing Societies. Yale University Press, New Haven.
Immergut, E., 1998. The theoretical core of the new institutionalism. Politics and Society 26 (1), 5–34.
Kahneman, D., 2011. Thinking, Fast and Slow. Penguin, London.
Kahneman, D., Slovic, P., et al. (Eds.), 1982. Judgement under Uncertainty: Heuristics and Biases. Cambridge University Press, New York.
Katzenstein, P.J., 1978. Between Power and Plenty: Foreign Economic Policies of Advanced Industrial States. University of Wisconsin Press, Madison.
Key, V.O., 1947. Politics, Parties and Pressure Groups. Thomas Y. Crowell Company, New York.
Levi, M., 1988. Of Rule and Revenue. University of California Press, Berkeley.
Lewis, O., Steinmo, S., 2012. How institutions evolve: evolutionary theory and institutional change. Polity 44 (3), 314–339.
Lieberman, R.C., 2002a. Political institutions and the politics of race in the development of the modern welfare state. In: Rothstein, B., Steinmo, S. (Eds.), Restructuring the Welfare State-Political Institutions and Policy Change. Palgrave Macmillan, New York, pp. 102–129.
Lieberman, R.C., 2002b. Ideas, institutions, and political order: explaining political change. The American Political Science Review 96 (4), 697.
Mahoney, J., Thelen, K., 2009. How historical institutionalists explain change. In: Mahoney, J., Thelen, K. (Eds.), Explaining Institutional Change: Ambiguity, Agency, and Power in Historical Institutionalism. Cambridge University Press, New York.
March, J., Olsen, J., 1989. Rediscovering Institutions. Free Press, New York.
Mayr, E., 2004. What Makes Biology Unique?: Considerations on the Autonomy of a Scientific Discipline. Cambridge University Press, Cambridge, UK, New York.
Miliband, R., 1969. The State in Capitalist Society. Weidenfeld & Nicolson, London.
North, D., 1990. Institutions, Institutional Change and Economic Performance. Cambridge University Press, Cambridge.
North, D.C., 2008. What is missing from political economy. In: Weingast, B.R., Wittman, D.A. (Eds.), The Oxford Handbook of Political Economy. Oxford University Press, Oxford, pp. 1003–1009.
Pierson, P., 2004. Politics in Time: History, Institutions, and Social Analysis. Princeton University Press, Princeton.
Przeworski, A., Teune, H., 1970. The Logic of Comparative Social Inquiry. Wiley-Interscience, New York.
Rothstein, B., 1992. Labor market institutions and working class strength. In: Steinmo, S., Thelen, K., Longstreth, F. (Eds.), Structuring Politics: Historical Institutionalism in Comparative Politics. Cambridge University Press, New York, pp. 33–56.
Rothstein, B., 1998. Just Institutions Matter: The Moral and Political Logic of the Universal Welfare State. Cambridge University Press, New York.
Schattschnieder, E.E., 1960. The Semi-sovereign People. Holt Reinhart, New York.
Selznick, P., 1949. TVA and the Grass Roots: A Study in the Sociology of Formal Organization. University of California Press, Berkeley.
Skocpol, T., 1979. States and Social Revolutions: A Comparative Analysis of France, Russia and China. Cambridge University Press, Cambridge.
Steinmo, S., Thelen, K.A., et al., 1992. Structuring Politics: Historical Institutionalism in Comparative Analysis. Cambridge University Press, Cambridge, England; New York.
Streeck, W., 2010. Epilogue. Institutions in history: bringing capitalism back. In: Campbell, J., Crouch, C., Kristensen, P.H., et al. (Eds.), Handbook of Comparative Institutional Analysis. Oxford University Press, Oxford.
Streeck, W., Thelen, K.A., 2005. Introduction: institutional change in advanced political economies. In: Streeck, W., Thelen, K. (Eds.), Beyond Continuity. Oxford University Press, Oxford, pp. 1–39.
Verba, S., 1967. Some dilemmas in comparative research. World Politics 20, 111–127.
Weingast, B., 1996. Political institutions: rational choice perspectives. In: Goodin, H.D.K.R.E. (Ed.), A New Handbook of Political Science. Oxford University Press, Oxford, pp. 167–190.
Wilson, W., 1891. Congressional Government: A Study in American Politics. Houghton, Boston.

Institutionalism and Neo-institutionalism: History of the Concepts

Markus Hadler, Macquarie University, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the history of neoinstitutional theory. It starts with a definition of the term institution and then discusses the development of neoinstitutionalism in economics, political science, and sociology. It also addresses differences between the main strands of neoinstitutionalism and concludes with a brief overview of the discursive institutionalism.

Early institutionalism and neoinstitutionalism share an interest in the effects and origins of 'institutions.' The term 'institution' however is widely used and this article, therefore, starts with exploring its different meanings. The scientific approach of neoinstitutionalism itself is commonly seen as divided into the three strands rational-choice, historical, and organizational neoinstitutionalism. This classification aligns vaguely with the disciplines of economics, political science, and sociology. The second part of this article, therefore, discusses the development of institutionalism in each of these three disciplines. This historical overview is followed by a discussion of the differences between these different strands and a description of the most recent development, the discursive neoinstitutionalism.

Institution and Institutionalism

'Institution' is used in various regards and can, according to Immergut (2011), refer to the degree of formalization of social behavior, ranging from casual habits to formal arrangements, to the extent of the collective binding, ranging from traditions and folkways to sanctioned laws, and to the way they are sanctioned, ranging from being self-reinforced to coercion by other entities. Finally, institution can also label organizations such as institutes of scientific study and financial organizations that are charged with the implementation of 'institutions' such as education and monetary instruments. All these aspects have in common that institutions are considered socially constructed and not biological determinants.

In terms of scientific disciplines, early institutionalism refers to approaches in the late-nineteenth and early twentieth centuries and scholars such as Thorsten Veblen or Max Weber. Their main emphasis was on studying the development of larger formal institutions such as legal, economic, religious, and other societal systems. These approaches, however, lost ground in the early decades of the twentieth century, when behavioristic models gained importance. Once shortcomings and inconsistencies in these behavioristic models became obvious, institutionalism was revived, starting in the 1970s.

These new versions of institutionalism share only some of the views of the early institutionalism, despite being labeled 'neo'-institutionalism. A main difference is that neoinstitutionalism considers institutions broader than the earlier scholars and is rather in line with Peter Berger and Thomas Luckmann's definition that "institutionalization occurs whenever there is a reciprocal typification of habitualized action by types of actors." (1966: p. 54). In this regard, the term institution is not limited to formal institutions any more, but emphasizes the experience of an institutional world as an objective reality. There are, however, significant differences within the contemporary 'neoinstitutionalisms' and these differences will be addressed in more detail in the following sections.

Economics

The early institutionalism in economics developed in opposition to the neoclassical view that individuals are rational economic actors and that a few basic assumptions and laws are sufficient to explain economic processes. These conflicting views found their expression in the German Methodenstreit of the late nineteenth century, in which institutionalists such as Gustav Schmoller emphasized that economic processes are outcomes of specific historic developments and embedded in cultural settings, whereas scholars such as Carl Menger represented the neoclassical views. Early American institutionalists such as Thorstein Veblin and John Commons were influenced by this German discussion and shared the interest in the historical development of economic phenomena.

Thorstein Veblen studied in his book *The Theory of the Leisure Class: An Economic Study of Institutions* (1899) the individual overconsumerism, a behavior that violates the assumption of rational, resourceful, and efficient actions. Veblen identified the stratification system and social classes in feudalism as a possible cause of this utility violating behavior. John Commons, in contrast, focused on the transaction, the basic social interaction of two parties. This interaction follows certain rules of conduct, which are also considered social institutions. Commons was particularly interested in the historical development and origin of these institutions and summarized his views on the collective nature of institutions in his 1934 book *Institutional Economics*.

The early institutionalism lost ground in economics over the first decades of the twentieth century, whereas neoclassic views gained influence at the same time. A possible explanation for its demise can be seen in the institutionalists' overemphasis on historical developments and idiosyncratic explanations with little emphasis on theory development. This diminished influence is also reflected in the fact that Ronald Coase's article from 1937, *The Nature of the Firm*, that focused on transaction costs and is now considered a classic

work by neoinstitutionalists, went rather unrecognized for several decades.

It was only in the 1970s, when Oliver Williamson added specific conditions such as increasing transaction costs under uncertainty and individual opportunism to Coase's work, that Coase's early work became more popular. It was also Williamson (1975) who coined the term 'Economic neo-institutionalism.' Williamson's work, however, shows an important difference to the early institutionalism and highlights its closeness to neoclassic views: Williams assumes given individual preferences and does not consider them as being shaped by specific historical developments and a cultural embedding.

This difference is also generally reflected in the contemporary economic neoinstitutionalism, which shares only some assumptions with the early institutionalism. Similar to the early approaches, it focuses on the effects of institutions and their origin and evolution. It considers "Institutions … written and unwritten rules, norms, and constraints that human devise to reduce uncertainty and control their environment" (Menard and Shirley, 2005: p. 1). It, however, also adopted the neoclassic assumptions of scarcity and competition – without the notion of perfect information and instrumental rationality.

A different strand within economics focuses on the macro-development of societal institutions and thus is close to the historical institutionalism, which will be discussed in the political science section. A prime example in this regard is the economic historian Douglas North, who, for example, discussed the types and forms of economic exchange in different societies and how its patterns and structures have evolved over time (North, 1991).

The historical development of contemporary economic institutions is also addressed in evolutionary economics. Richard Nelson and Sidney G. Winter's book *An Evolutionary Theory of Economic Change* (1982) that describes changes in routines and technology can be named in this regard. Despite not explicitly referring to institutions, their research resonates with the ideas of Thorstein Veblen and particularly Joseph Schumpeter – one of the few economists, who continued to work in the tradition of the early institutionalism.

Political Science

Institutional approaches with a focus on descriptions of the roots and developments of legal systems were the dominant approach in political science in the nineteenth and early twentieth centuries. Prominent scholars in this time were J.W. Burgess and the later US president Woodrow Wilson. These scholars produced dense descriptions of legal systems and other political institutions that were guided by moral philosophical accounts and – similar to the early institutionalism in economics – rather scarce in theory building. This approach lost ground in the first decades of the twentieth century when behavioral perspectives gained importance, accompanied by a topical shift to political behavior, voting behavior, and public opinion.

Ideas of these early institutionalists were revived in a reaction to group theories and structural-functionalist theories that were dominant in the 1960s and 1970s and labeled historical institutionalism. Historical institutionalism sees institutions as "… the formal or informal procedures, routines, norms and conventions embedded in the organizational structure of the polity or political economy" (Hall and Taylor, 1996: p. 938).

Among others, the rise of this theory in political science is associated with Theda Skocpol's book *States and Social Revolutions: A Comparative Analysis of France, Russia and China* (1979), which focuses on the social institution 'state' and its influence on social change. Its interest in the development of modern institutions resonates with the ideas of early institutionalists, but differs as it does not follow normative considerations. Outcomes are not considered predetermined, but often as unintended consequences. Insofar, there is no necessary trend toward democracy, modernity, and other societal traits. Institutions, however, are not considered the only determinant of social change. Factors such as dominant ideas and societal affluence are taken into account as well.

Despite being a reaction to group theories and structural-functionalism, the historical institutionalism nonetheless incorporated ideas of these theories (Hall and Taylor, 1996). It kept the assumption of rivalry and scarcity. In this regard, power differences between groups is a central concern and historical institutionalists frequently focus on the role of institutions in giving some groups more power than others. Similar to structural-functionalists, historical institutionalists consider the polity a system of interrelated parts, but rather consider its influence on individual behaviors than seeing individual traits and beliefs as the basis of this system.

As far as the relationship between the individual and institutions is concerned, historical institutionalism is located between the rational-choice and sociological neo-institutionalism (Hall and Taylor, 1996). Scholars use both a rational-choice related approach, where institutions provide information that can reduce the uncertainty for actors that make rational choices, and a cultural approach, where institutions are considered cognitive templates that frame the choices of an individual. While the former considers individuals as utility maximizers, the latter sees them as individuals that select or satisfy different options.

Alongside with this historical institutionalism, a rational-choice focused neoinstitutionalism that extents economic institutionalism to political science has developed as well. The rise of this rational neoinstitutionalism was inspired by the paradox observation that the voting outcomes and agreements within the US congress were driven by relatively stable majorities, whereas rational-choice theory would predict constantly changing alliances, given the many different interests of legislators (Hall and Taylor, 1996). An explanation of these stable outcomes is 'institutions' as they are considered to "lower the transaction costs of making deals" (Hall and Taylor, 1996: p. 943). These ideas have been employed well beyond the US congressional outcomes and applied to the influence of laws, political institutions, and the evolution of law.

Sociology

Institutions have been discussed within sociology in more detail than in economics and political science and, thus, the perspectives of various classic European and American

sociologists are reflected in contemporary neoinstitutionalism (see Scott, 2001).

Among the European roots, Karl Marx's view that the material world and the associated beliefs and norms determinate the social construction of reality, Emile Durkheim's emphasis on social facts that govern human behavior, and Max Weber's approach that considered economic institutions as specific historical developments, but was also in favor of theory building, can be named. Among the American sociologists, Herbert Spencer's and William Sumner's division of societies into different segments, Everett Hughes' studies of the institutional structures of the workplace, Talcott Parsons' system of action, and Peter Berger and Thomas Luckmann's theory of the social construction of reality can be named.

A first important shift toward the contemporary sociological neoinstitutionalism was the application of these classic thoughts to organizations – an aspect the early institutionalists with their focus on broader social phenomena such as political and economic systems did not consider explicitly. The impetus to focus on organizations came via Robert K. Merton and, particularly, his student Philip Selznick. Selznick's (1949) study of the Tennessee Valley Authority (TVA) can be considered a classic work in this regard. The TVA was founded by the US government to aid the underdeveloped areas between Kentucky and Alabama by acquiring private land to control floods and producing electric power. The organization co-opted local actors, which turned out to undermine its mission. Important features of this study are a focus on unintended consequences, political influences, and struggles, also from outside the organization, and not just a narrow view on the economic environment.

In the consecutive decades, however, organizational sociology was dominated by contingence theory, resource dependency, and population ecology with a narrower focus on the economic environment (Senge, 2007). Contingency theory, for example, focuses on technical and economic circumstances and which form of organization survives; resource dependency on organizations that exchange resources and, even more, resources that have monetary value; and population ecology on the adaption of organizations. In the late-1970s and early 1980s, the interest in institutions gained again importance. This new institutionalism revived perspectives of the classic approaches and added ideas developed in philosophy and social psychology.

The relationship of neoinstitutionalism to philosophy comes from phenomenology, a strand of philosophy that goes back to the German scholar Edmund Husserl in the late-nineteenth century. His ideas were picked up by various scholars and led, for example, to the development of philosophical anthropology, an approach that considers humans as both creatures and creators of their social environments and institutions as a replacement for biological determinants. Phenomenology found its way into sociology via the Viennese scholar Alfred Schütz, whose social phenomenology was based on everyday knowledge. Sociologists such as Harold Garfinkel, Peter Berger, and Thomas Luckmann continued these thoughts and the focus on the semiotic functions of culture with an emphasis on knowledge and belief systems turned out to be central to sociological neoinstitutionalism.

Parallel, neoinstitutionalism also integrated developments of social psychology that emphasize the cognitive processes of individuals. Individuals are considered information processers that use particular 'scripts' and 'schemas' as templates in this interpretation. This view complements the sociological emphasis on the context that sees individuals as rather passively exposed individuals. It is also reflected in the organizational sociology definition of institutions: "Institutions are socially constructed, routine-reproduced (ceteris paribus), program or rule systems. They operate as relative fixtures of constraining environments and are accompanied by taken for granted accounts" (Jepperson, 1991: p. 149).

In sum, the old and the new institutionalism in sociology share various perspectives, but have also clear differences (see DiMaggio and Powell, 1991: p. 12ff). Both are distant to rational actor models of organizations, emphasize the influence of the organizational environment, and consider the role of culture in shaping organizations. As for differences, early institutionalism focused on the political influence and resulting constraints through trade-offs implied by these choices, whereas the new institutionalism emphasizes the influence of common understandings that are taken for granted. Furthermore, the old institutionalism focuses on informal interactions within organizations to explain irrational decisions, whereas the new institutionalism considers the organizational structure itself irrational and decisions as driven by cultural accounts from outside the organization. It has also a broader understanding of organizational environments and does not limit them to the local context. This is reflected in the new view that institutionalization occurs outside the organization, within industrial sectors, or even at the societal level. The classic organizational institutionalism, in contrast, located these processes within the organizations.

Ideas developed within this new sociological institutionalism have been extended beyond the narrow scope of economic organizations. Within sociology, the world society – also world polity – approach can be named in this regard (see Meyer et al., 1997). This approach discusses the diffusion of global ideas and how they influence states, organizations, and other constituents around the world. This view is reflected in the hypothetical example Meyer et al. discuss: A newly discovered island would be immediately visited by consultants of various international organizations and soon develop an organizational structure that resembles the structure of most states around the world.

Differences between the Three Main Areas

An obvious difference between these approaches, first, arises from the nature and topical areas of their disciplines. Less obvious, differences also emerge regarding the role of the individual (Schofer et al., 2012). The economic rationalism operates under a rational actor assumption and asks why individuals enter institutional arrangements and studies the consequences of these arrangements on behaviors. Historical institutionalism focuses on macrodevelopments and considers the evolution of different institutions, but places less emphasis on their influence on the individual (Hall and Taylor, 1996). The sociological neoinstitutionalism sees individuals as

constituted by institutions, thus asserting that even the choices and ideas individuals perceive are shaped by the prevalent institutions (Jepperson, 1991).

Another difference comes from the question of the origins of institutions. Both sociological and historical institutionalisms assume an existing world of institutions. The question then arises how new institutions are drawn from this existing repertoire. As Hall and Taylor (1996) point out, the sociological view places less emphasis on the role of agents in implementing institutions, while the historical institutionalism places more emphasis on existing power relationships that favor some actors over others in the process of creating new institutions. Rational-choice institutionalism, on the other hand, emphasizes the functions of institutions, and the reason why they were kept up.

Discursive Institutionalism

Recently, a fourth strand of neoinstitutionalism has developed that is labeled discursive institutionalism or constructivist institutionalism (Schmidt, 2008). This approach focuses on the role of ideas and is considered a more dynamic approach than the three other neoinstitutionalisms. Rational-choice, historical, and sociological neoinstitutionalism have a tendency to emphasize the external role of institutions in shaping individual behaviors, organizational goals, and other outcomes. The discursive approach, in contrast, considers institutions as a process in which agents create and maintain institutions. In this regard, discursive institutionalism studies the development of ideas, places them in a specific context, and scrutinizes the way they are conveyed. In its application in political science, it is employed to two basic forms of discourse: between politicians (coordinate discourse) and between politicians and the general public (communicate discourse). In studying these processes, research in the area of discursive institutionalism often integrates insights from the other three neo-institutionalisms and thus could be also considered a complementary approach.

See also: Globalization and World Culture; Individuals in Society: History of the Concept; Microeconomics, History of; Political Science, History of; Sociology, History of.

Bibliography

Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Anchor, Garden City, New York.
Coase, R., 1937. The nature of the firm. Economica 4, 386–405.
Commons, J.R., 1934. Institutional Economics. Macmillan, New York.
DiMaggio, P.J., Powell, W.W., 1991. Introduction. In: Powell, W.W., DiMaggio, P.J. (Eds.), The New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago, pp. 1–38.
Hall, P.A., Taylor, R.C., 1996. Political science and the three institutionalisms. Political Studies 44, 936–957.
Immergut, E.M., 2011. Institution/institutionalism. In: Badie, B., Berg-Schlosser, D., Morlino, L. (Eds.), International Encyclopedia of Political Science. Sage, London.
Jepperson, R.D., 1991. Institutions, institutional effects, and institutionalism. In: Powell, W.W., DiMaggio, P.J. (Eds.), The New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago, pp. 143–163.
Menard, C., Shirley, M.M. (Eds.), 2005. Handbook of New Institutional Economics. Springer, Dordrecht.
Meyer, J.W., Boli, J., Thomas, G.M., Ramirez, F.O., 1997. World society and the nation-state. American Journal of Sociology 103, 144–181.
Nelson, R., Winter, S.G., 1982. An Evolutionary Theory of Economic Change. University Press, Harvard.
North, D.C., 1991. Institutions. The Journal of Economic Perspectives 5, 97–112.
Schmidt, V.A., 2008. Discursive institutionalism: the explanatory power of ideas and discourse. Annual Review of Political Science 11, 303–326.
Schofer, E., Hironaka, A., Frank, D.J., Longhofer, W., 2012. Sociological institutionalism and world society. In: Amenta, E., Nash, K., Scott, A. (Eds.), The New Blackwell Companion to Political Sociology. Oxford, Blackwell.
Scott, R.W., 2001. Institutions and Organizations. Sage, Thousand Oaks.
Senge, K., 2007. Was ist neu am Neo-Institutionalismus. OEZS 32, 42–65.
Selznick, P., 1949. TVA and the Grass Roots: A Study in the Sociology of Formal Organization. University of California Press, Berkeley.
Skocpol, T., 1979. States and Social Revolutions: A Comparative Analysis of France, Russia and China. University Press, Cambridge.
Veblen, T., 1899. The Theory of the Leisure Class: An Economic Study of Institutions. Macmillan, New York.
Williamson, O.E., 1975. Markets and Hierarchies, Analysis and Antitrust Implications: A Study in the Economics of Internal Organization. Free Press, New York.

Institutions

Thomas R Voss, University of Leipzig, Leipzig, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Institutions induce regularity and stability in human interaction. All human societies have developed a variety of institutions that deal with recurrent basic problems in social life; for example, institutions regulating kinship and family life and economic and political institutions. There are many different kinds of institutions some of which are described in the article. To explain the emergence of institutions and of institutional diversity is one of the most important tasks in social theory. Functionalist accounts point out the beneficial effects of institutions. It is, however, particularly important to explain why inefficient institutions emerge and persist.

Institutions induce regularity and stability in human interaction. All human societies have developed a variety of institutions that deal with recurrent basic problems in social life; for example, institutions regulating kinship and family life and economic and political institutions. The *concept* institution entails duality in a certain sense. Institutions are, on the one hand, *constraints* on human action. They provide 'rules of the games' actors are playing. On the other hand, institutions are results of human action and interaction. Institutions can then be defined as *outcomes* of social interactions such that the agents' behavior becomes stable because it is consistent with an equilibrium. Thus, there is a 'structural' and an 'agency' aspect in the concept of institutions. Institutions can be defined as shared beliefs that certain rules will be used in specific recurrent situations by almost all other agents. Given these beliefs, a behavioral regularity will be observed. Institutions are relevant in various societal domains and on different levels. Social theory deals with the *explanation* of institutions and of institutional diversity within and across societies. Many sketches of explanations of institutions emphasize the *effects* (or functions) of institutions. However, these (quasi) functionalist accounts are not sufficient to explain the *emergence* of institutions.

The Concept 'Institutions'

In everyday language and in the social sciences, the term 'institutions' is used in a broad and sometimes amorphous way. It is quite common to denote organizations such as governments, business firms, political parties, schools, and public administrations as 'institutions.' For example, Goffman (1961) coined the term 'total institutions' to characterize certain types of organizations like prisons, monasteries, and mental hospitals, which tend to 'mortify' the inmate's self. On the other hand, the term 'institutions' also, more specifically, refers to rules and other constraints of human interaction. 'Institutions' in this sense include conventions (e.g., traffic rules), as well as informal and formal social norms. Furthermore, the term is often used to describe complex systems of interrelated norms and the organizations and other factors that enforce these normative standards (Greif, 2006). Modern law can be considered as a set of formal norms proscribing or prescribing certain actions. Another aspect of the law is that it employs (mainly negative) sanctions to punish and prevent deviant behavior. In accordance with Max Weber's terms (1922), the law induces stable regularities of nondeviant behavior via (threats of) sanctions that are employed by a specialized staff of actors. In this sense, the law may be conceived as an organization that enforces certain formal norms.

There is some disagreement about whether or not organizations should be conceptually demarcated from institutions. Greif (2006: p. 30 – italics omitted) defines institutions as "systems of social factors that conjointly generate a regularity of behavior." Besides norms, rules, and beliefs, Greif considers organizations as institutional elements. Other authors (e.g., North et al., 2009, Chapter 1) admit that organizations comprise institutional structures, namely, rules, norms, and beliefs. However, in contrast to institutions, organizations are sets of actors pursuing their goals via coordinated or even centralized decision making. Many organizations can be conceived as collective or corporate actors (Coleman, 1990). Following this line of reasoning, organizations and institutions are distinct: institutions are elements of organizations but organizations include additional components.

Some 'institutions' and social orders like authority systems have 'legitimacy' in M. Weber's (1922) sense. That is, (some of) the involved individuals share the belief that the institutional rules are valid, fair, or obligatory and there is a kind of moral duty to accept them. In cases like these, the institution's stability may be more robust in comparison to institutions that merely depend on self-interest. It seems, however, inappropriate to define institutions by the criterion of legitimacy. The legitimacy of an institution is in general an empirical question and may be a matter of degree.

Another point of disagreement is related to the question of whether institutions can be represented by the rules that prescribe or prohibit certain actions; that is, by linguistic *statements* with deontic ('ought') operators that specify that certain actions should be performed under specific conditions (Ostrom, 2005, Chapter 5). Analyses of the syntax and grammar of institutional statements may be important in institutional analysis, but one should keep in mind that the existence of normative or other institutional statements per se does not guarantee that a stable regularity of behavior will emerge.

The concept 'institutions' is thus used in a dual way. First, it denotes rules, constraints, or norms of human interaction. Second, it describes the resulting stable patterns of interaction

(regularities of behavior) among a set of agents and the social mechanisms generating this outcome. To allude to the modern choice theoretic framework of game theory, institutions are first represented by the rules of the games actors are playing. These rules define, among other aspects, the set of possible actions (strategies) of the agents. They also specify the consequences of the actions (payoffs) and the available amount of information and shared beliefs about each agent's behavior. Given these rules, constraints, and shared beliefs, agents choose an action (strategy) that is individually optimal. The optimal choices will, under general conditions, yield an equilibrium. An equilibrium (in the Nash sense of classic game theory) is a combination of actions (strategies) such that no agent has an incentive to deviate from his or her chosen actions unilaterally; that is, under the condition that the other actors choose their equilibrium actions. The stability of an equilibrium is due to the fact that no (rational) actor could gain by shifting to another strategy unilaterally. Consider now a population of actors who recurrently interact in a social situation governed by the rules of the game. Then it may happen that a specific equilibrium will eventually evolve. In other words, a stable pattern of social interactions may emerge under the constraints of the rules of the game.

An 'institution' can thus be conceived to be the set of constraints or rules that determine the opportunities, beliefs, and incentives of the relevant actors in a given situation. Second, interactions within these rules generate equilibria, which may be stable over time if the situation is recurrent. These equilibrium outcomes can be considered an integral component of the institution. To illustrate, consider a coordination situation. Imagine a time with considerable road traffic (e.g., horse-drawn carriages in the seventeenth century) but no universally accepted rules with respect to keeping left or right. If a central authority (the government) were to create the rule that pedestrians use the right side of the road and horse-drawn carriages use the left side, one source of potential accidents could be avoided, namely, inconsistent mutual reciprocal expectations with regard to the usage of public roads. The rule would, so to speak, 'solve' a coordination problem. A coordination problem can roughly be defined as an interaction with multiple equilibria. In the absence of any rules, precedents, or 'focal points' (Th. C. Schelling), there would be ambiguity with regard to the different courses of action available to road users: should one use the right or the left side of the road? The actors probably do not have intrinsic preferences with regard to the direction of the rule; that is, whether to go right is better than to go left. But they prefer to adopt a rule that gives clear and unique prescriptions and prevents accidents if universally acknowledged, rather than ambiguous rules or rules that are not easily implemented (e.g., the rule that one should determine by chance whether to go right or left). Given that the government has publicly announced a rule, the participants easily form convergent expectations based on shared beliefs with regard to each other's behavior. In other words, an equilibrium of universal rule-oriented behavior emerges as a consequence of individually adaptive or optimizing behavior under the constraints of the rules of the game and of external constraints created by the government. This gives rise to a 'convention.' Following Lewis (1969), a convention may be defined as a behavioral regularity R in a population of agents who recurrently interact in a coordination situation if R is an equilibrium of the recurrent situation and if almost every member of the population prefers to conform to R on the condition that almost every other member of the population conforms to R too.

Conventions arising in so-called pure coordination situations (with no conflict of interest) are self-enforcing. This is so because any individual who deviates from R would punish himself. There is no incentive to break the rule – even if external monitoring and sanctions were absent. The shared belief (expectation) that the rule is universally accepted is confirmed and reinforced in any instance of the recurrent situation and becomes a kind of self-fulfilling prophecy. In other social situations, for example, of the prisoner's dilemma type, there is an incentive to defect. A prisoner's dilemma is a social interaction such that universal cooperation is profitable to everyone, but it is even more profitable to defect unilaterally from cooperation. Cooperation is not an equilibrium situation in the single-shot prisoner's dilemma. In a population of agents who recurrently interact in a prisoner's dilemma, an institution of mutual cooperation may emerge if there is a policing mechanism such that members of the population of agents expect that nonconformity will be punished by (negative) sanctions. The policing or enforcement of cooperation could be achieved under appropriate social conditions due to sanctions that are created by the agents themselves. Of course, intervention by external, third parties (in the sense of Hobbes' *Leviathan*) would be another mechanism that enforces cooperation. Generalizing Lewis's concept of convention, a social institution can be defined as a regularity in the behavior of the members of a population that "specifies behavior in specific recurrent situations, and is either self-policed or policed by some external authority" (Schotter, 1981: p. 11; italics omitted). The self-enforcement of an institution may be due to the fact that the regularity is an equilibrium of a coordination problem. If the recurrent situation has similarities with a prisoner's dilemma, however, an equilibrium of cooperative behavior may result from self-created sanctions (e.g., of the 'tit for tat' type, cf. Axelrod, 1984) or from policing by an external authority or from other factors. The concept 'social institution' in this sense includes conventions as a special case, and also formal and informal social norms (see Lewis, 1969; Schotter, 1981; Aoki, 2001; Greif, 2006 for more sophisticated elaborations of this line of reasoning).

Some Types of Institutions

There are many different kinds of institutions. First, adopting the intuitive meaning of institutions as rules or constraints of behavior, it is useful to make the following distinctions. There are rules that regulate the personal life of human agents. These *personal* rules (e.g., the rule to go early to bed every night) resemble individual habits and individual routines. Some of these rules may be interpreted as attempts of individuals to commit or bind themselves. Personal rules may be motivated by individuals' long-term interests to achieve self-control. But they are not social institutions in the proper sense. Institutions require that routines and habitualizations are based on reciprocal typifications of interacting actors (Berger and

Luckmann, 1966: pp. 47–91). In contrast to personal rules, *social* rules affect social interactions; that is, recurrent situations of social interdependence. For example, it can be argued that social institutions typically regulate social dilemmas. Social dilemmas are situations where interdependent agents are faced with choices for which individual rationality does not yield a socially optimal outcome. The prisoner's dilemma and public good situations are cases in point. Furthermore, there are *formal* and *informal* institutions as special cases of social rules. Informal institutions do not rely on an external authority's monitoring and policing of the participants' behavior. These institutions include, for example, social norms of self-help among residents in rural communities (see Ellickson, 1991), or, to allude to an example from history, the community norms governing irrigation systems in villages of Tokugawa, Japan (Aoki, 2001, Chapter 2). Other informal institutions are conventions. Although conventions are defined as self-enforcing shared beliefs, they often become codified in formal rules (e.g., time standards or traffic rules). Formal institutions are generally created and arranged by agents who are able to rely on third parties for monitoring and enforcement. Among the most elaborate formal institutions in modern societies one should mention the law. There are also complex formal institutions regulating political committees in political systems and collective decisions in large capitalist firms. However, it is important to notice that formal rules are, in general, embedded in informal norms. Thus, formal institutional constraints are often supplemented by informal rules and vice versa.

Second, it is useful to point out the different levels of rules that can be relevant in institutional analysis (see Ostrom, 1990, Chapter 2). There is first the *constitutional* level. Constitutional rules define the most general context of the games agents are allowed to play. They determine the set of eligible actors and the principles used in crafting or changing more specific rules of the system. Constitutional rules are relevant not only in the arena of international and national politics but also in regional and local contexts. Second, there is the level of *collective choice*. Collective choice rules specify principles of policy making. They help to generate unique collective decisions. Third, there is the *operational* level. Operational rules help to implement the decisions of the higher levels. They determine day-to-day decisions; for example, decisions related to monitoring and sanctioning the participants.

Third, institutions may be classified by the domains or spheres of the society to which they belong and contribute. There are *economic* institutions regulating economic exchange; for example, property rights institutions. Other economic institutions include conventions with regard to media of exchange or the specific rules and constraints affecting economic behavior between and within firms (see Williamson, 1985). *Political* institutions provide basic rules and procedures to achieve collective decisions. The most important political institutions in contemporary societies are, of course, the constitutions of nation-states; for example, the constitution of the United States of America. There are also international political institutions. The European Union is constituted by a number of supranational institutions like, among others, the European Commission, which constrain national polities to a certain degree and provide a kind of federal framework for the set of participating countries. Besides these basic ground rules of international and national politics there are more specific institutions and laws dealing with the operation of the democratic system. Every human society has developed institutions shaping family life and kinship relations. Modern societies have created a differentiated system of educational institutions. There are also cultural and religious institutions of different kinds.

Explaining Institutions

The analysis of institutions has always been considered to be one of the most important tasks in social theory. For example, classic sociologist Emile Durkheim (1895) emphasized that sociology's subject was 'social facts,' in particular social institutions. Social facts are external or exogenous constraints to each individual. A desideratum of sociological analysis, according to Durkheim, is the discovery of the mechanisms, causes, and effects of institutions in the diverse spheres of societal life. There are multiple approaches to explain institutions and institutional change within the social sciences.

Effects of Institutions

A starting point of many explanations is the study of the effects or consequences of institutions. Not only functionalist theorizing in the proper sense but also many other approaches argue that institutions may have beneficial effects on different levels: For the involved agents institutions provide exoneration ('Entlastung' (A. Gehlen)) by reducing the complexity in the psychological process of everyday decision making (see Berger and Luckmann, 1966: pp. 53–54). Routines and habits yield lower costs of information and deciding.

Institutions regularize social life. Rule-oriented behavior reduces uncertainty and increases the predictability of others' actions in situations of social interdependence. The institutionalization of social norms in Parsons' sense is a key to the solution of the Hobbesian problem of order. In a 'state of nature' without an enforcing agency, mutual cooperation, for example, the recognition of private property rights, would be profitable, but there are strong incentives to defect from mutual agreements. The institutionalization of social norms of reciprocity and solidarity via their internalization into the actors' motivational systems would be, according to Parsons, a solution to the Hobbesian problem (Parsons, 1937). Thus, institutions and the institutionalization of appropriate social norms may foster cooperation and increase efficiency (in the Pareto sense of making agents better off without harming anyone in society).

Institutions such as exclusive private property rights with respect to economic resources are prone to enhance a society's aggregate welfare because they reduce transaction costs (North, 1990). If transaction costs within an economic system decrease, the economy will generally become more productive, provided the costs of creating and enforcing these institutions are not too high (see Eggertsson, 1990 for a survey of the literature). Many institutions, such as property rights and, more generally, the rule of law can be said to be part of a society's social capital

(Coleman, 1990, Chapter 12). In this sense, institutions are a productive asset because they enhance the efficiency, the welfare or, in other words, the well-being of most members of a society. Furthermore, they can have the properties of public or collective goods. That is, they are goods such that it is not feasible easily to exclude agents from their consumption if these goods are available to anyone. In other words, institutions with collective goods properties can be profitably 'consumed' by all those actors who wish to benefit from them, not only by those agents who contributed to the costs of their 'production.'

This public good characteristic of many institutions makes it clear that the potential beneficial effects of institutions cannot sufficiently explain the emergence and change of institutions. This is so because public goods create incentives to a free ride. If the formation and maintenance of an institution is costly, the amount of voluntary contributions to these costs will generally be too low, that is, suboptimal (Olson, 1971). At least for this reason, simple functionalist accounts of efficient institutions are incomplete.

Processes of Institutional Change

There are many different types of processes that are prone to generate or change social institutions. First, there is the extreme type of institutional change by conscious design. In particular, formal institutions such as legal norms are in many cases designed purposively. Parliamentary committees in modern democracies regularly propose such changes. However, many of these changes are anticipated, accompanied, or supplemented by changing informal institutions in the larger society. These informal rules often show a higher degree of inertia and change more incrementally.

Second, institutional change may be due to evolutionary forces of an 'invisible hand' (Adam Smith). In this case, the goal-oriented actions of a large number of actors may in the long rule unintentionally produce results that seem to be due to conscious planning by a central authority. But, in fact, the institution is a by-product of each individual's activity to locally adapt to its circumstances. Institutions of this kind are called 'spontaneous orders' (F.A. Hayek). It can be argued that many conventions have emerged as spontaneous orders (Sugden, 1986; Young, 1998) in the first place. Once adopted, they may be included in a cultural tradition and will be taught to children by their parents (e.g., table manners).

To explain institutions by invisible-hand processes (Ullmann-Margalit, 1978) is attractive because it may yield theoretical arguments with some 'depth.' Facts on the collective level of social institutions are by principles of purposive action and by a 'generating mechanism' (Boudon, 1979; Hedström, 2005). Obviously, most empirically observable institutions will not be explicable exclusively by one of these two extreme or ideal types of mechanisms but are probably complex combinations of different types of processes. However, the study of theoretical models representing some aspects of a complex reality in an idealized way may be understood as a theoretical 'thought experiment' (*Gedankenexperiment*). Such thought experiments enhance the clarity of theoretical arguments and provide an antidote against simplistic and empirically dubious 'conspiracy theories' of institutions (K.R. Popper), which argue that most social institutions are explicable in terms of the joint efforts of some group of powerful actors who want to foster their material interests.

Explaining Institutional Diversity

Given the functionalist account, which points out that institutions are or can be beneficial, one may ask whether in fact every institution is efficient. There is indeed some evidence on efficient rules. Wittman (1982) has investigated various microinstitutions like rules of thumb (like 'first come, first served') in certain road traffic situations in the United States and comes to the conclusion that only the most efficient rules will be used. Rules have evolved such that two types of costs are optimized, namely, costs associated with monitoring the rule and inappropriate incentives created by the rule. Similarly, Ellickson (1991) argues that rules that governed potential conflicts among whalers in the eighteenth century whaling industry spontaneously evolved such that they minimized transaction costs and deadweight losses. Rules were adapted to the particular circumstances such as the kind of whales and helped to improve efficiency with respect to defining and enforcing property rights about which ship is entitled to possess a whale in cases where several whalers competed about a catch. However, there is ample evidence on the emergence and persistence of prima facie inefficient institutions (see, for example, Edgerton, 1992). In certain rural areas of southern Italy, observers described the informal institutional pattern of an 'amoralist familism,' that is, family members follow the maxim: 'Maximize the material short-run advantage of the nuclear family. Assume that all others do likewise' (Banfield, 1958: p. 83). The set of shared beliefs implied by this rule obviously will generate a lack of cooperation and trustfulness in social and economic interactions with strangers (or actors outside the own nuclear family). There will be larger transaction costs and deadweight losses than those implied by alternative rules that are seemingly more prevalent in other Italian regions (Putnam, 1993). Moreover, economic historian North (1990: p. 6) points out that the "central puzzle of human history is to account for the widely divergent paths of historical change. How have societies diverged? What accounts for their widely disparate performance characteristics?" It seems safe to say, according to North, that market processes of competition did not weed out inferior institutions but that inefficient institutions have proved to be highly stable. Besides efficiency, distributional aspects of institutions are important. Some authors point out that there are two roughly different types of political institutions: inclusive and exclusive or extractive institutions (Olson, 1982; North et al., 2009; Acemoglu and Robinson, 2012). Extractive institutions concentrate political power in the hands of a small elite with little constraints or control of this power by other groups of society, whereas inclusive institutions tend to guarantee a significant degree of pluralism within the political system and the rule of law within society as a whole. Inclusive institutions foster inclusive economic institutions and therefore efficiency within the economy and economic prosperity of the society.

There are many different arguments that could explain the persistence of inefficient institutions:

1. As far as efficiency gains from institutional change are public or collective goods, there may be a lack of appropriate incentives to contribute to the costs (see Olson, 1971; Ostrom, 2005). In the management of self-governed local common pool resources like common fishing grounds or irrigation systems, some basic design principles must be followed to make possible viable institutions (Ostrom, 1990, Chapter 3): clearly defined boundaries, rules that are adapted to local conditions of the resource, effective decentralized monitoring and sanctioning, and some others. Given these conditions, users of a commons will have appropriate incentives to operate the resource in a sustainable manner. On the other hand, if some or most of these conditions are not met, inefficient institutions will emerge.
2. Inequality in bargaining power can have the effect of selecting institutions that primarily serve the interests of the more powerful agents or groups in a society. Thus, institutions need not foster the aggregate welfare of a society or efficiency (North, 1990, Chapter 6).
3. There are many recurrent interaction situations on different levels with multiple equilibria, for instance, situations resembling coordination games or iterated prisoner's dilemma games. In cases like these, the outcome may depend on specific boundary conditions; that is, historical circumstances and cultural traditions that are associated with different sets of shared beliefs. Thus, there is 'path dependence': Small differences in initial conditions and chance circumstances in structurally similar social interactions yield different paths of historical development and therefore different institutional outcomes (see North, 1990, Chapter 11; Young, 1998; Greif, 2006).
4. Limited rationality and incomplete information of human agents may be obstacles to attaining efficient institutions (North, 1990, Chapter 3). On the other hand, some results from variants of evolutionary game theory seem to demonstrate that stochastic elements and erroneous decision making within a population of boundedly rational agents can in the 'long run' help produce 'better' equilibria. The basic reason is that there are, in general, multiple equilibria that can be attained in the process of institutional change. Given that an inefficient, 'bad' equilibrium has been realized, stochastic shocks or individual mistakes in choosing a best reply to other agents' behavior may open the path to switch to a 'better' equilibrium. These equilibria, however, will in general not persist forever, even if they are efficient (Young, 1998; Aoki, 2001, Chapter 7).
5. Some social scientists, in particular in anthropology, suggest that institutions are generally inexplicable by functionalist or rationalist ideas. According to this approach, (cultural) institutions mainly serve 'symbolic' functions (Sahlins, 1976). Sociologists within the tradition of 'new institutionalism' argue that efficiency (whatever this means) is less relevant than 'legitimacy.' Institutional elements, for example, practices within organizations, will therefore be structurally similar to cultural norms that are common in other domains of society rather than based on the criterion of efficiency (see Powell and DiMaggio, 1991).

The Dynamics of Institutional Change

There are many different approaches with respect to the dynamics of change.

Research on the dynamics of the emergence of institutions has not yet provided a unified theoretical framework. There is, however, some agreement with respect to the criteria that an adequate approach should meet (see Aoki, 2001; Greif, 2006; North et al., 2009). Although it seems obvious that much change is due to exogenous factors (like technological change or other shocks), it is important to be able to explain institutional dynamics endogenously. This means that institutions are studied as self-enforcing systems of incentives in which the motivation is supplied endogenously: institutional elements like rules, conventions, or norms provide the information for each individual about shared beliefs. Given this information, an agent chooses the best reply – that is, to comply with the relevant regularity if the institution is consistent with an equilibrium in the Nash sense. The aggregate of individuals thus stabilizes and reinforces the institution. Second, stability and change must be studied within a single, unified theoretical framework. Thus institutional change has to be analyzed endogenously. There are many different analytic approaches. Classical game theory is arguably the most elegant and precise theoretical tool. However, there is some doubt about the empirical content and empirical accuracy. One problem is that classical game theory is indeterminate with respect to the outcome in situations with multiple equilibria. Game theory therefore seems to be basically a useful heuristic tool to clarify alternative explanatory mechanisms. Generally, the dynamics of institutional change will depend on initial conditions. That is, structurally similar social systems may develop qualitatively distinct institutions with considerable differences with regard to the degree of 'efficiency.' A particularly important aspect of initial conditions are 'cultural beliefs' in Greif's (2006, Chapter 9) sense. These beliefs are embedded into a society's culture and are a product of history. They are employed as equilibrium selection devices in those interaction situations with multiple equilibria. Therefore, different sets of cultural beliefs may yield different institutional outcomes (at least in the medium run). Greif (2006) illustrates this with the help of a comparison of the cultural beliefs in 'collectivist' and 'individualistic' societies, using the historical data on Maghribini and Genoese traders in the Middle Ages. An adequate theory should, in addition, point out the conditions that generate distinct cultural beliefs (and other equilibrium devices such as fairness norms) and the paths of an evolutionary transition from one equilibrium to another (possibly more efficient) equilibrium. In this respect, evolutionary approaches will be attractive.

Some ideas in this context are inspired by evolutionary game theory. Historically, the roots of evolutionary explanations of institutions are ideas from the tradition of the Scottish Moralists (in particular David Hume) and the Austrian School of Economics (Carl Menger and Friedrich A. von Hayek).

In constructing an evolutionary approach, one must keep in mind that there are important differences between biological and social evolution. Theories of social evolution must contain appropriate behavioral assumptions of adaptive behavior or learning processes. The evolutionary approach can be characterized by the following core ideas (see Young, 1998). Institutions are the products of interactions within large populations of actors over considerable periods of time. These agents are boundedly rational and act under conditions of limited information. This means that actors may fail to figure out responses to other people's actions that are optimal. However, *under appropriate circumstances,* the *evolutionary process* may in the long run (i.e., in the presence of stochastic shocks that correspond roughly with mutations in biological evolution) realize equilibrium outcomes such that efficient institutions persist. But this result may well be an exception rather than the rule. An evolutionary approach should, in principle, point out the conditions that generate distinct cultural beliefs (and other equilibrium devices such as fairness norms) and the paths of an evolutionary transition from one equilibrium to another (possibly more efficient) equilibrium. The evolutionary approach is a promising research program, but much further work is needed to fully understand and explain concrete cases of empirically observed institutions.

See also: Collective Action; Collective Behavior, Sociology of; Conventions and Norms: Philosophical Aspects; Durkheim, Emile (1858–1917); Educational Institutions, History of; Family as Institution; Game Theory; Ideologies, Institutions, and the New Institutionalism; Institutionalism and Neo-institutionalism: History of the Concepts; Labor Market Institutions and Labor Market Outcomes in Transition Economies; Law and Society: Development of the Field; New Institutionalism in the Analysis of Complex Organizations; Old and New Institutionalism in Economics; Organization, Geography of; Organizations, Sociology of; Prisoner's Dilemma, One Shot and Iterated; Public Goods: International; Smith, Adam (1723–90); Sociology and the New Institutionalism.

Bibliography

Acemoglu, D., Robinson, J.A., 2012. Why Nations Fail. Crown, New York.
Aoki, M., 2001. Toward a Comparative Institutional Analysis. MIT Press, Cambridge, MA.
Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York.
Banfield, E., 1958. The Moral Basis of a Backward Society. Free Press, New York.
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality. Doubleday, Garden City, NY.
Boudon, R., 1979. Generating models as a research strategy. In: Merton, R.K., Coleman, J.S., Rossi, P.H. (Eds.), Qualitative and Quantitative Social Research: Papers in Honor of Paul Lazarsfeld. Free Press, New York.
Coleman, J.S., 1990. Foundations of Social Theory. Belknap Press of Harvard University Press, Cambridge, MA.
Durkheim, E., 1895. Les règles de la méthode sociologique, eighteenth ed. 1973. Presses Universitaires de France, Paris.
Edgerton, R.B., 1992. Sick Societies. Free Press, New York.
Eggertsson, T., 1990. Economic Behavior and Institutions. Cambridge University Press, Cambridge, UK.
Ellickson, R.C., 1991. Order without Law. Harvard University Press, Cambridge, MA.
Goffman, E., 1961. Asylums. Doubleday, Garden City, NY.
Greif, A., 2006. Institutions and the Path to the Modern Economy. Cambridge University Press, Cambridge, UK.
Hedström, P., 2005. Dissecting the Social – Principles of Analytical Sociology. Cambridge University Press, Cambridge, UK.
Lewis, D.K., 1969. Convention. Harvard University Press, Cambridge, MA.
North, D.C., 1990. Institutions, Institutional Change and Economic Performance. Cambridge University Press, Cambridge, UK.
North, D.C., Wallis, J.J., Weingast, B.R., 2009. Violence and Social Orders. Cambridge University Press, Cambridge, UK.
Olson, M., 1971. The Logic of Collective Action. Harvard University Press, Cambridge, MA.
Olson, M., 1982. The Rise and Decline of Nations. Yale University Press, New Haven.
Ostrom, E., 1990. Governing the Commons. The Evolution of Institutions for Collective Action. Cambridge University Press, Cambridge, UK.
Ostrom, E., 2005. Understanding Institutional Diversity. Princeton University Press, Princeton, NJ.
Parsons, T., 1937. The Structure of Social Action. Free Press, New York.
Powell, W., DiMaggio, P. (Eds.), 1991. The New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago.
Putnam, R., 1993. Making Democracy Work. Yale University Press, New Haven, CT.
Sahlins, M., 1976. Culture and Practical Reason. University of Chicago Press, Chicago.
Schotter, A., 1981. The Economic Theory of Social Institutions. Cambridge University Press, Cambridge, UK.
Sugden, R., 1986. The Economics of Rights, Cooperation and Welfare. Blackwell, Oxford, UK.
Ullmann-Margalit, E., 1978. Invisible-hand explanations. Synthese 39, 263–291.
Weber, M., 1922. Wirtschaft und Gesellschaft, third ed. 1978. J. C. B. Mohr, Tübingen, Germany.
Williamson, O.E., 1985. The Economic Institutions of Capitalism. Free Press, New York.
Wittman, D., 1982. Efficient rules in highway safety and sports activity. American Economic Review 72, 78–90.
Young, H.P., 1998. Individual Strategy and Social Structure. Princeton University Press, Princeton, NJ.

Instructional Design

Franz Schott, Institute for Educational and Developmental Psychology, Technische Universitaet Dresden, Dresden, Germany
Norbert M Seel, University of Freiburg, Freiburg i.Br., Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by F. Schott, volume 11, pp. 7566–7569, © 2001, Elsevier Ltd.

Abstract

This article introduces into the field of instructional design (ID) that is closely related to instructional technology, defined in terms of the application of scientific knowledge to the systematic design of learning systems. First, the relationship between instructional technology, design, and systems development will be described. Then we provide a brief historical overview to characterize relevant aspects of ID. Historically seen, the nine events of instruction as introduced by Gagné in the 1960s resulted in an abundance of models of systematic ID that were integrated in a comprehensive framework centering on five tasks: analysis, design, development, implantation, and evaluation (shortly called the ADDIE model). Along with the approaches of systematic ID some alternative approaches of learning and instruction emerged that were based on cognitive psychology and constructivism. These approaches will be discussed briefly before we describe, in some detail, present theories and applications of ID. Finally, we will discuss some challenges for future theory, research, and development within the realm of ID.

Instructional Design and Its Relation to Instructional Technology

The term instructional design (ID) refers to the systematic and professional planning and implementation of education or training. Considerations regarding planned instruction have been made at least as long as there have been institutions for instruction and training. The term ID itself appeared for the first time in the United States in the mid-twentieth century.

From its very beginning, ID was closely related to instructional technology, which is generally defined as the systematic application of theoretically and practically established knowledge to the development of learning systems, for which the name 'instructional systems development' (ISD) is also used. Often the relationships between ID, ISD, and instructional technology are expressed by the formula 'Instructional Technology = ID + ISD.' In a general sense, ID is defined as the entire process of instructional planning and implementation. However, the term ID is also used to denote a scientific discipline that refers to theory building and research on instruction and its implementation in various educational settings.

More specifically, the term ID is used in a variety of ways with different meanings:

1. The narrowest variant concerns only the planning of a particular instructional treatment.
2. The somewhat wider usage concerns both the planning and construction of that which is needed for instruction in various educational settings.
3. The even wider interpretation adds the implementation of the instructional system or program in a specific context as well as the evaluation of outcomes. The term ID is most often used according to this understanding and sometimes refers to instructional technology with the same broad meaning.
4. The widest interpretation adds to the previous meaning by including maintenance and management of the instructional system or program over an extended period of time; this includes conducting ongoing evaluations, planning, and developing upgrades as well as managing all of these processes in an organized and economical way. This broad interpretation of ID corresponds to the most comprehensive definition of instructional technology.

ID, under any of these interpretations, can refer to these education or training processes with regard to varying degrees of scope (unit of instruction, lesson, curriculum, program, system, etc.). At the *macro level*, ID concerns the instruction of complex curricula, programs, or systems that may cover months or even years of instruction; this is called *large-scale ID*, or, alternatively, curriculum construction and implementation. At the *meso level*, the object of ID is lessons that cover hours or weeks of instruction. The *micro level* concerns only short parts of a lesson, such as exercises or illustrative examples. The most frequent use of ID models refers to the meso level. However, the limits between the three levels are fluid and not always clearly defined.

Some authors tend to reduce ID to the mere planning of instruction as well as to restrict the term instructional technology to hardware and software, but these narrower interpretations do not meet the conditions of efficient instruction, because it has to take into account all relevant factors (e.g., learners, environmental factors) in a systematic way and in accordance with a systemic approach.

Today, education and instruction must increasingly accommodate a diverse set of students who need to learn complex cognitive skills and transfer them to a varying set of complex real-world settings and contexts. Accordingly, these new demands present a serious challenge to the field of ID. In particular, more powerful ID models are needed to deal with today's scientific and societal demands.

A Brief Historical Overview Characterizes Relevant Aspects of ID

The various essays about the short history of ID (e.g., Reiser, 1987) agree on the assumption that the systematic planning of instruction began in the 1920s. Usually, Tyler's (1922)

'Objectives for Learning' and Pressey's (1926) idea of teaching machines are considered as starting points. In accordance with Tyler's view on instruction, the 1930s were characterized by various attempts to establish a systematic planning of instruction, whereas the 1940s saw the rapid development of instructional media and training programs for the military. During these years, instruction as a special field of research and development was mainly stimulated by two general objectives:

1. Training as many learners as possible within a short period of time;
2. Using the most efficient available instructional means (e.g., media, such as written instructions and demonstration films, as well as practical exercises).

The Second World War introduced enormous training problems for the military in the United States. To address these problems, the 'Division of Visual Aids for War Training' of the 'US Office of Education' produced hundreds of films and training handbooks. Military training became an example for the successful connection of research and development to professional practice. Teams for instructional development started to see the use in including instructional designers as well as subject matter experts and a 'producer.' From that time on, the necessity for professional ID standards and professional ID practitioners has been recognized.

The next decade, the 1950s, was mainly characterized by the systematic application of basic learning principles of behaviorism to the development and implementation of programed instruction as well as to task analyses, with a particular emphasis on the adaption of instruction to individual learning conditions.

This educational movement culminated in the 'birth of ID' at the beginning of the 1960s, when Glaser (1966) introduced the term 'instructional design' into the literature and Gagné (1965) published *The Conditions of Learning*, which originally had a behavioristic orientation; later editions, however, provided an orientation on cognitive psychology (Gagné, 1985). In accordance with the idea of a hierarchy of learning processes, Gagné distinguished nine events of instruction that provided a fundamental basis for an instructional theory "to propose a rationally based relationship between instructional events, their effects on learning processes, and the learning outcomes that are produced as a result of these processes" (Gagné, 1985: p. 244). Based on the nine events of instruction, Gagné and Briggs (1979) contributed greatly to the emerging field of ID by answering three basic questions: (1) What is known about human learning that is relevant for ID? (2) How can knowledge about human learning be organized for concrete applications by designers of instruction? (3) What methods and procedures should be followed in applying knowledge of human learning to the design of instruction?

Gagné's theory, in combination with a systemic view (Salisbury, 1989), had unique effects on the development of various prescriptive ID models, such as the well-known model of Dick and Carey (1996) and various derivative models provided by Merrill (1983), Reigeluth (1983), Smith and Ragan (2004), and many others. These models agree on the assumption that learning can be classified in accordance with cognitive operations and processes that are similar by nature (i.e., internal conditions of learning) and can be facilitated by comparable instructional methods and strategies (i.e., external conditions of learning). However, the various models of ID initiated by Gagné do not provide a homogeneous class but rather focus on different components and strategies of ID. For example, Kaufman (e.g., 1988) focused on the particular role of needs assessment and strategic planning at different levels of education. In Kaufman's view, ID should not simply start with the formulation of instructional goals but rather with an analysis of the instructional objectives that are useful for the anticipated audience of particular learning environments. Needs analysis and strategic planning is still one of the most important areas of ID and ISD today.

Another important extension of ID models inspired by Gagné was the idea of automatizing the design of instruction and providing computer-based expert systems for this purpose (Tennyson, 1994). However, the idea of automating ID by means of expert systems can be considered as a failed project.

As a consequence of the rapid development of ID models, a major 'watershed' for the professionalization of instructional designers was reached in the United States in the middle of the 1970s. For the first time, instructional designers took precedence over subject matter experts as responsible contract partners for course development. In Europe this development was delayed.

Parallel to the emergence of ID models, but in contrast to their behaviorist orientation, some psychologists, such as Ausubel, Bruner, Cronbach, Glaser, Wittrock, and others, realized what is occasionally called the 'cognitive revolution' of psychology in the late 1960s (Bruner, 1990). To a great extent inspired by Piaget's epistemology, cognitive approaches of learning and instruction increasingly replaced the earlier behaviorist approaches (Shuell, 1986). This type of 'new' learning theories was called 'educational learning theory' (Bereiter, 1990). The introduction of educational learning theory has been associated with a strong tendency to investigate complex instructional matters.

This tendency as well as new societal and technological demands have challenged ID as a discipline, culminating in the constructivism–objectivism debate of the 1990s (Duffy and Cunningham, 1996). This debate can be considered as the expression of an essential uncertainty of ID theorists concerning the epistemological, psychological, and technological foundations of ID. This debate made essential contributions to the self-conception of ID by assimilating and advancing theories from cognitive psychology, systems theory, and communication technologies. Thus, recent additions to ID have been influenced by constructivism, situated cognition, e-learning approaches to distance education, and information theory. As a result of a Delphi study aiming at the identification of trends that may influence ID, Ritchie and Earnest (1999) point out that "with each iteration, we enhance our understanding of how to impact the performance of individuals and organizations" (p. 35) by means of designed instruction. So, for example, some approaches of ID have tried to integrate more systems thinking into ID. This includes (1) the view on instructional systems, subsystems, and their relationships to the environment; (2) the development of highly systematic, often linear or branched design procedures; and (3) more emphasis on systemic approaches of ID, which takes the

interrelationship between all relevant factors of instruction into account.

If we reconsider the various trends of ID during in the past two decades, we can state that theories and models of ID now come in different types and concern different worlds. In these different worlds, ideas about 'how to help people learn better' lead to different answers to the two basic questions of ID: "what to teach?" and "how to teach it?" (van Merriënboer et al., 2002). In general, the models of ID found in the literature can be classified into three 'generations' of ID. The first generation includes the development of procedural ID models (often illustrated as flowcharts) and is greatly influenced by Gagné and his disciples. The second generation of ID models contains approaches that can be understood as realizations of 'educational engineering' and aims at the automation of parts of the overall design process. The third generation, however, contains approaches which aim at the derivation of theoretically sound and research-based principles of the design of learning environments. The present applications of ID correspond to a large extent to generations (1) and (3).

Present Applications of ID

In accordance with the traditional procedural models of ID, present-day ID and ISD generally comprises the following subtasks:

- Performing a needs assessment,
- Identifying instructional goals,
- Analyzing goals into subgoals and identifying prerequisite knowledge,
- Defining what the learners have to learn to bridge the gap between prerequisite knowledge and subgoals,
- Developing criterion-referenced tests and/or other assessment measures,
- Developing instructional strategies,
- Developing materials and preparing media,
- Conducting formative evaluation, and
- Conducting summative evaluation (a final evaluation of the whole instructional procedure).

It should be noted that these are typical tasks found in many current ID models; variations among professional practitioner models still exist. Moreover, from a systemic point of view these subtasks are interrelated and are usually not performed in a linear way or only a single time; rather, the practice of ID is interactive and responsive to individual situations. However, in view of the vast amount of concurring ID and ISD models, a comprehensive framework – called ADDIE, an acronym for *A*nalysis, *D*esign, *D*evelopment, *I*mplementation, and *E*valuation – is widely used in ISD. Actually, ADDIE exists more as a label than as an ID model and its origins are not clear at all. It seems that the ADDIE model is an umbrella term referring to a family of procedural ID models which share a common underlying structure of ISD. Although there is no original and elaborated ADDIE model, most theorists in the field of ID agree that ADDIE is only an acronym that covers the major processes of the generic ISD process: *A*nalysis, *D*esign, *D*evelopment, *I*mplementation, and *E*valuation.

The phase of *analysis* contains, first of all, a needs analysis for different levels. Furthermore, this phase contains the analysis of objectives and performances as well as analyses of the addressees and the tasks and a cost–benefit analysis of the planned instruction. The *design* phase of ADDIE focuses on the design of a blueprint of the outcomes of instruction. Accordingly, the emphasis of this phase is on the design of a storyboard encompassing the entire structure of the instructional intervention or learning environment. At the center of design is the planning of instruction, and the necessary decisions refer to the orchestration of the external conditions for learning (e.g., methods of instruction, social interactions, media, and the organization of the environment).

The *development* phase of ADDIE contains the concrete realization of the decisions made in the design phase. More specifically, development concerns the production of the concrete learning material. The starting point of the development phase is a decision centered on the alternatives 'Buy it or make it.' In cases of doubt it is more efficient to work with preexisting material instead of producing it from scratch, and often it is more cost-effective to modify and adjust existing material to meet the requirements of a particular instructional situation.

Implementation contains the concrete realization of planned instruction in a real setting, for instance an online setting. Typically, the implementation of instruction occurs under controlled conditions and involves critical examination. This is often closely associated with a formative evaluation during the development phase. The *evaluation* phase of ADDIE refers to the test of whether the instruction was successful. The evaluation helps to measure the instructional impact on learners in terms of the predefined instructional objectives. Usually, a distinction is made between formative evaluation (i.e., a process-oriented evaluation of the development of instruction) and summative evaluation as an instrument of quality management. While it is perhaps the most popular model of ISD, the ADDIE model has some weaknesses, especially with regard to the development of multimedia-based instruction. For, example, ADDIE tends to be inefficient because it is not iterative. Furthermore, its strict sequence of steps tends to work well for static content but may be restrictive when one is dealing with learning outcomes that do not have a predetermined end state (e.g., complex problem solving). Another significant weakness is that there is no accommodation for dealing with faults or good ideas throughout the overall process of ADDIE. The designer must know all of the requirements before s/he develops the content to be learned. Probably the biggest weakness, however, is that ADDIE is nothing other than a mere extrapolation of former ID models that does not take into account new approaches of instructional psychology. As a consequence, the ADDIE model has been criticized as a "metaphor for the lack of clarity in the field of instructional design and technology" (Bichelmeyer, 2004).

These weaknesses have led to a number of spin-offs or variations, such as van Merriënboer's (1997) 4C/ID model or the Decision Oriented Instructional Design (DO-ID) from Niegemann et al. (2008). The four components of the 4C/ID model are authentic *learning tasks*, which are considered as the driving force of the intended learning processes, *supportive*

information (i.e., instructional assistance, coaching), *procedural information* (how to do things), and *part-task practices*. The DO-ID model of Niegemann et al. is an elaboration of ADDIE aiming at the design of multimedia.

For the time being, ADDIE can be considered as the culmination of the development of ID and ISD models in the tradition of Gagné and his disciples. However, in the 1990s alternative approaches emerged within the field of educational psychology. These new approaches aim at deriving theoretically sound, research-based principles for the design of learning environments for complex problem solving and guided discovery learning. They range from the design of learning tasks according to the Cognitive Load Theory (Chandler and Sweller, 1991; Paas et al., 2004) to comprehensive conceptions of instruction, such as anchored instruction, goal-based scenarios, and model-centered learning and instruction (see, for an overview, Seel and Dijkstra, 2004). The approaches of anchored instruction and goal-based scenarios can be considered historical and terminated events, whereas model-centered learning and instruction and the closely related approach of role-based design have developed into an increasingly prospering field of creative ID (Hanke et al., 2011; Hokanson and Miller, 2009; Seel, 2003).

ID approaches can be applied in all fields and situations that deal with education and training, including school settings, technical training, professional training, collaboration with human resource development professionals in industrial and professional organizations, higher education and university settings, as well as distance and lifelong learning situations – using all old and new kinds of media and telecommunications (see Bibliography). In short, ID has matured into a broad profession with many connections to other professions and activities, including personnel and project management.

It is especially noteworthy that the practice of ID often involves a variety of experts, such as text designers, media designers, software programmers, subject matter experts, and learning specialists. Such a team is obviously necessary when the development involves technology-based learning or instruction since the required skills and expertise are likely to be distributed among a variety of specialists. When ID teams are involved, it is critical to have recognized standards and an established methodology.

Challenges for Future Theory, Research, and Development

When we take into consideration the status quo of ID and technology we can state that there has been no significant change with regard to the general principles and methods of ID to date – and they were probably not necessary at all. Nevertheless, there are some important challenges in store for future theory, research, and development within the realm of ID:

- Dealing with integrated sets of learning goals for teaching complex skills and higher order skills. Such transferable skills are becoming more and more important in our society (cf van Merriënboer, 1997).
- An expert can broaden his or her competency by integrating the individual knowledge *and* the resources involved in obtaining, storing, and processing information into computerized tools. To take this into account we need to develop more powerful, integrated techniques for instructional and task analysis.
- Fostering the systematic design of collaborative learning/team training, of new technological developments in distance education, in particular web-based instruction.
- Dealing with highly heterogeneous target groups (differences in prior knowledge and skills, cultural diversity, older learners) and target groups with low motivation (e.g., 'adolescents at risk,' cf Attewell and Seel, 2003).
- Further development of effective forms of assessment. The need of assessment for efficient ID solutions is often underestimated.
- Automating ID processes using techniques of artificial intelligence. This can help to lower their cost, for example, by generating tasks for a given item format and content or producing instructional materials for a variety of products of the same type automatically.
- Last but not least, a better theoretical foundation of ID which takes into account the whole problem of providing *useful* knowledge for human beings in our world of nearly unlimited possibilities, in which we collect immense amounts of information from everywhere.

Ten years ago, an international consortium of researchers was organized by the International Center for Learning, Education, and Performance Systems (ICLEPS) of the Learning Systems Institute at Florida State with the aim of promoting new developments within the realm of ID. The cooperating partners and collaborators include numerous faculties and researchers from different universities around the world (e.g., the Open University of the Netherlands, the University of Freiburg (Germany), the University of Bergen (Norway), the University of Sydney, the University of Central Florida). ICLEPS has organized several workshops and meetings where studies were reported and discussed with an emphasis on integrated research for the development of human learning and performance systems. The stated mission of ICLEPS is to build on cognitive load theory, mental model development, and ID theory to investigate: (1) human cognition and social processes that mediate learning and performance; (2) factors that influence human performance; (3) conditions that promote practical and effective learning; (4) methods to assess individual and team progress; and (5) methods to evaluate programmatic change and diffusion of innovation. However, it seems that ICLEPS does not exist anymore. But this does not affect the future of ID.

See also: Adult Education and Lifelong Learning: The US Experience and Beyond; Co-constructivism in Educational Theory and Practice; Cognitive Development: Mathematics Learning and Instruction; Cognitive Styles and Learning Styles; Computer-Assisted Instruction; Distance Education; Education and Learning: Lifespan Perspectives; Education: Skills Training; Educational Effectiveness Theory and Research: Recent Advances; Educational Evaluation: Overview; Educational Learning Theory; Educational Media: Potentials for Learning; Intelligence, Prior Knowledge, and Learning; Learning Theories and Educational Paradigms; Learning and Instruction:

Social-Cognitive Perspectives; Learning to Learn; Lifelong Learning and Its Support with New Technologies; Metacognitive Development: Educational Implications; Self-Concepts: Educational Aspects; Self-Efficacy: Education Aspects; Self-Regulation During Adolescence: Variations Associated with Individual–Context Relations; Tangible User Interfaces in Learning and Education.

Bibliography

Attewell, P., Seel, N.M. (Eds.), 2003. Disadvantaged Teens and Computer Technologies. Waxmann, Münster.

Bereiter, C., 1990. Aspects of an educational learning theory. Review of Educational Research 60 (4), 603–624.

Bichelmeyer, B., 2004. Instructional Theory and Instructional Design Theory: What's the Difference and Why Should We Care? AECT 2004 IDT Futures Group Presentations. IDT Record. From: http://www.indiana.edu/%7Eidt/articles/documents/ID_theory.Bichelmeyer.html (retrieved 09.11.04.).

Bruner, J., 1990. Acts of Meaning. Harvard University Press, Cambridge, MA.

Chandler, P., Sweller, J., 1991. Cognitive load theory and the format of instruction. Cognition and Instruction 8 (4), 293–332.

Dick, W., Carey, L., 1996. The Systematic Design of Instruction, fourth ed. Harper Collins Publishing, New York.

Duffy, T.M., Cunningham, D.J., 1996. Constructivism: implications for the design and delivery of instruction. In: Jonassen, D. (Ed.), Handbook of Research for Educational Communications and Technology. Simon & Schuster Macmillan, New York, pp. 170–198.

Gagné, R.M., 1965. The Conditions of Learning and Theory of Instruction, first edition. Holt, Rinehart & Winston, New York.

Gagné, R.M., 1985. The Conditions of Learning, fourth ed. Holt, Rinehart and Winston, New York.

Gagné, R.M., Briggs, L., 1979. Principles of Instructional Design. Holt, Rinehart and Winston, New York.

Glaser, R., 1966. The design of instruction. In: Goodlad, J.S. (Ed.), The Changing American School. Rand, Chicago, pp. 215–242.

Hanke, U., Ifenthaler, D., Seel, N.M., 2011. Effects of creative dispositions on the design of lessons. The Open Education Journal 4 (Suppl. 2: M10), 179–185.

Hokanson, B., Miller, C., 2009. Role-based design: a contemporary framework for innovation and creativity in instructional design. Educational Technology 49 (2), 21–28.

Kaufman, R., 1988. Planning Educational Systems. Technomic Publishing Comp., Lancaster, PA.

van Merriënboer, J.J.G., 1997. Training Complex Cognitive Skills: A Four-Component Instructional Design Model for Technical Training. Educational Technology Publications, Englewood Cliffs, NJ.

van Merriënboer, J.J.G., Seel, N.M., Kirschner, P.A., 2002. Mental models as a new foundation for instructional design. Educational Technology 42 (2), 60–66.

Merrill, D., 1983. Component display theory. In: Reigeluth, C.M. (Ed.), Instructional Design Theories and Models: An Overview of their Current States. Lawrence Erlbaum, Hillsdale, NJ.

Niegemann, H.M., Domagk, S., Hessel, S., Hein, A., Hupfer, M., Zobel, A., 2008. Kompendium Multimediales Lernen. Springer, Berlin.

Paas, F., Renkl, A., Sweller, J., 2004. Cognitive load theory: instructional implications of the interaction between information structures and cognitive architecture. Instructional Science 32, 1–8.

Pressey, S.L., 20 March 1926. A simple apparatus which gives tests and scores – and teaches. School and Society 23, 373–376.

Reigeluth, C.M. (Ed.), 1983. Instructional Design Theories and Models: An Overview of their Current States. Lawrence Erlbaum, Hillsdale, NJ.

Reiser, R.A., 1987. Instructional technology: a history. In: Gagné, R.M. (Ed.), Instructional Technology: Foundations. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 11–40.

Ritchie, D., Earnest, J., 1999. The future of instructional design: results of a Delphi study. Educational Technology 39 (1), 35–42.

Salisbury, D.F., 1989. What should instructional designers know about general systems theory? Educational Technology 29, 42–45.

Seel, N.M., 2003. Model-centered learning and instruction. Technology, Instruction, Cognition, and Learning 1 (1), 59–85.

Seel, N.M., Dijkstra, S. (Eds.), 2004. Curriculum, Plans and Processes of Instructional Design: International Perspectives. Lawrence Erlbaum Associates, Mahwah, NJ.

Shuell, T.J., 1986. Cognitive conceptions of learning. Review of Educational Research 56 (4), 411–436.

Smith, P.L., Ragan, T.J., 2004. Instructional Design, third ed. John Wiley, Danvers, MA.

Tennyson, R.D. (Ed.), 1994. Automating Instructional Design: Computer-based Development and Delivery Tools. Springer, Berlin.

Tennyson, R.D., Schott, F., Seel, N.M., Dijkstra, S. (Eds.), 1997.

Instructional design – international perspectives. Theory, Research, and Models, vol. 1. Lawrence Erlbaum Associates, Mahwah, NJ.

Tyler, R.W., 1922/1949. Basic Principles of Curriculum and Instruction, thirteenth ed. The University of Chicago Press, Chicago.

Instructional Psychology

Erik De Corte, University of Leuven, Leuven, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Instructional psychology is the study of learning processes in educational settings, and of the design of environments for eliciting those processes in view of attaining adaptive competence in subject-matter domains. A theory of learning from instruction involves five components: (1) description of desired end states of instruction; (2) description of initial states of learners prior to instruction; (3) explanation of transition processes from initial to desired states; (4) specification of instructional conditions that promote this transition; and (5) assessment of instructional effects. This article focuses on three components: adaptive competence as end state, constructive learning as lever for the transition to desired end states, and learning environments to promote transition.

Introduction

Instructional psychology is the study of the processes and outcomes of human learning in a variety of educational and instructional settings, and of the nature and the design of environments that are appropriate to elicit those learning processes aiming at the attainment of adaptive competence, i.e., the ability to apply meaningfully learned knowledge and skills flexibly and creatively in a variety of situations. The term 'instructional psychology' was introduced in 1969 by Gagné and Rohwer (1969) in the *Annual Review of Psychology* as title of their literature review of a field of inquiry indicated until then as 'educational psychology.' Today one can say that while educational psychology is somewhat broader in scope than instructional psychology – involving, for instance, psychological aspects of child rearing in the family – both terms cover largely the same domain of research, because most work has been done on studying knowledge and skill acquisition in instructional settings.

Since its earnest inception at the beginning of the twentieth century, the scientific study of learning has undergone major paradigmatic shifts (De Corte, 2010). In the behaviorist era, it was conceived of as response strengthening based on reinforcements. A fundamental change occurred in the 1950s with the advent of cognitive psychology: the focus on the central role of information processing by the learner led to the view of learning as the rather passive acquisition of knowledge. Stressing the active role of the learner as a sense maker resulted next in a new metaphor for learning as knowledge construction. In the 1980s, this constructivist view was amended by highlighting the important role of situations in which cognition and learning occur; from that development emerged the socio-constructivist conception of learning as participation and social negotiation. From an educational perspective, this currently constitutes the dominant view of learning. In this approach, the psychological processes evolving in the learner, on the one hand, and the social and situational aspects impacting learning, on the other hand, are considered to be reflexively related (Palincsar, 1998).

A comprehensive review of this extensive field of research is outside the scope of this article. Snow and Swanson (1992) have distinguished five essential components in a theory of learning from instruction: (1) description of desired end states or goals of instruction in a domain; (2) description of goal-relevant initial states of learners prior to instruction; (3) explanation of the transition processes from initial to desired states; (4) specification of instructional conditions that promote this transition; and (5) assessment of performance and instructional effects. This article will focus selectively on three of these components, namely desired end states, transition processes, and instructional conditions. For a more thorough overview of the field, especially from an Anglo-Saxon perspective, the reader is referred to the *Handbook of Educational Psychology. Second Edition* (Alexander and Winne, 2006), *The Cambridge Handbook of the Learning Sciences* (Sawyer, 2006), the *Handbook of Research on Learning and Instruction* (Mayer and Alexander, 2011), and *Educational Psychology: Concepts, Research and Challenges* (Rubie-Davies, 2011). To get a more comprehensive idea of European research, one can consult the periodicals *European Journal of Psychology of Education* (published since 1986), and *Learning and Instruction. The Journal of the European Association for Research on Learning and Instruction* (published since 1991).

Toward Adaptive Competence as Desired End State

Many scholars in the field of education now agree that the ultimate goal of learning and instruction in different subjects consists in acquiring adaptive competence, i.e., the ability to apply meaningfully learned knowledge and skills flexibly and creatively in different situations. This is opposed to routine expertise, i.e., being able to complete stereotype school tasks quickly and accurately without understanding.

Building adaptive competence in a domain requires the acquisition of several categories of cognitive, affective, and motivational components (De Corte, 2010):

1. a well-organized and flexibly accessible domain-specific knowledge base that constitutes the contents of a subject-matter domain;
2. heuristic methods, i.e., search strategies for problem analysis that increase the probability of finding the correct solution, because they induce a systematic approach to the task;

3. metaknowledge, involving knowledge about one's cognitive functioning (metacognitive knowledge) on the one hand, and knowledge about one's motivation and emotions relating to learning in a subject on the other hand;
4. self-regulatory skills, involving skills relating to regulating one's cognitive processes (metacognitive skills or cognitive self-regulation) on the one hand, and skills for regulating one's affective and motivational processes (motivational self-regulation) on the other hand;
5. affective components such as positive beliefs, attitudes, and emotions relating to oneself as a learner, and relating to subject-matter domains and learning and problem solving in those domains.

Prioritizing adaptive competence does not mean that routine expertise is unimportant. It is obvious that mastering certain skills routinely (e.g., basic operations in arithmetic, spelling, technical competencies) is crucial in view to efficient functioning in all kinds of situations. If certain aspects of solving a complex problem can be performed more or less mechanically, it creates room to focus on higher order cognitive activities that are needed to reach a solution.

In the early days of cognitive psychology, laboratory investigations using mainly knowledge-lean tasks showed the important role of general cognitive skills, such as heuristics, in skilled problem solving. But starting in the mid-1970s, expert–novice studies in rich content domains demonstrated the significant role of domain-specific knowledge in expert performance. This has stimulated in the last decades of the twentieth century the study of learning and teaching of school subjects that can be considered as one of the most significant achievements of instructional psychology of the past 25 years (Mayer and Alexander, 2011). One successful field of work in this respect is the research on conceptual change, i.e., the study of how concepts in subject-matter domains change with learning and development (Vosniadou, 2008). Yet research in the 1980s also evidenced that domain-specific knowledge could not fully account for human expertise, and showed that self-regulatory skills are a major determinant of successful learning and skilled thinking. The study of the cognitive, motivational, and emotional self-regulation of learning has since become a very productive area of inquiry (Zimmerman and Schunk, 2011). More or less in parallel, investigations also revealed the important influence on their learning processes and results of students' beliefs about knowledge and learning (Bendixen and Feucht, 2010). The view on adaptive competence conceived as the integrated acquisition of the different components described above, developed thus gradually over the past five decades.

Finally, adaptive competence as defined here implies the ability to transfer one's knowledge and skills to new learning tasks and problem situations, and in that perspective also the willingness to continually develop the breadth and depth of its components. As such, adaptive competence is crucial in view of lifelong learning. Transfer itself was always considered as an important goal of education, and therefore it is not surprising that since the early days of instructional psychology, it has repeatedly been, and still is, a major topic in research on learning and instruction. The available literature shows that transfer is a very complex phenomenon, that transfer effects do not occur spontaneously, and that they are even difficult to obtain deliberately (Alexander and Winne, 2006).

Constructive Learning as Lever for the Transition from Initial to the Desired End State

Due to the primary interest in unpacking the mental processes and knowledge structures underlying performance, the study of learning was rather unpopular in the early days of cognitive psychology. Toward the end of the twentieth century, however, this situation began to change. First, with the substantial progress made in the 1970s and 1980s in understanding knowledge structures, skills, and processes underlying expert performance, there reemerged an interest in the learning processes required to acquire such competence, and consequently in the instructional arrangements that can support this acquisition. Second, the rise of the socio-constructivist perspective that stresses the importance of context and especially social interaction stimulated interest in studying learning in the complex reality of classrooms. From this research a series of characteristics of effective and meaningful learning processes has emerged, which can be summarized in the following definition of learning: it is a constructive, cumulative, self-regulated, goal-directed, situated, collaborative, and individually different process of meaning construction and knowledge building. Only some of these characteristics can be briefly elaborated here (see De Corte, 2010 for a more detailed discussion).

Learning Is Constructive

Research has shown that learning is an effortful and mindful process, in which students actively construct knowledge and skills through reorganization of their already acquired mental structures in interaction with the environment. Some scholars – often inspired by the work of Piaget – take a radical position in this respect, claiming that all knowledge is a subjective and idiosyncratic cognitive construction. Others represent a more moderate point of view that allows for the possibility of mediating learning through appropriate guidance. Notwithstanding the large variety in theoretical perspectives along the radical–moderate dimension, the constructivist view certainly implies that acquiring knowledge and skills requires active cognitive processing from the learner.

Learning Is Self-Regulated

This feature refers to the metacognitive and metamotivational nature of effective learning, and is implied in the constructive perspective. Indeed, self-regulation means that students manage and monitor their own processes of knowledge building (Zimmerman and Schunk, 2011). Skilled self-regulation facilitates appropriate decision making during learning, as well as the monitoring of an ongoing learning process by providing one's own feedback and performance evaluations, and by keeping oneself concentrated and motivated. The more students become self-regulated, the more they assume agency over their learning; consequently they are less dependent on external, instructional support for performing those regulatory activities. As evidenced by Hattie's (2009) extensive meta-analyses, the largest effects on

students' learning are obtained when they display self-regulatory characteristics such as self-monitoring, self-evaluation, and self-teaching. Research also shows that students' motivational beliefs play an important role in self-regulation. This holds especially for their self-efficacy beliefs, i.e., beliefs about one's capabilities to learn and perform at a designated level (Bandura, 1986). An additional argument for promoting students' self-regulation of their learning derives from the finding that high levels of metacognition facilitate transfer: students having good metacognitive awareness and strategies are more able at using what they have learned to approach unfamiliar problems in multiple ways.

Learning Is Situated and Collaborative

The view referred to above that learning and cognition are situated emerged in the 1980s in reaction to the then prevailing mentalistic, information processing view of learning and thinking as highly individual and purely cognitive processes occurring in the head, and resulting in the construction of mental representations. In contrast, situativity theory – influenced by the work of Vygotsky – proposes a contextualized and social conception of learning and thinking: learning is enacted essentially in interaction with the social and cultural context and artifacts, and especially through participation in cultural activities and contexts. In other words, effective learning is not a purely 'solo' activity, but essentially a distributed one, i.e., the learning effort is distributed over the individual students, their partners in the learning environment, and the resources and (technological) tools that are available. Thus, this situated perspective strongly stresses the importance of collaboration in productive learning reflected in such activities as exchanging ideas, comparing solution strategies, and discussing arguments.

Although the literature provides fairly good support for this conception of learning (see, e.g., Bransford et al., 2006), the constructivist perspective has also come in for criticism: approaches based on constructivism rely excessively on discovery learning and provide minimal guidance to students, ignoring thereby the structure of human cognitive architecture and resulting in cognitive overload of working memory. These authors plea for a return to direct instruction. The critics are correct in concluding that pure discovery does not yield the best learning gains. But, they mistakenly equate constructive learning with discovery learning. Learning as an active, constructive, self-regulated process does not at all imply that students' construction of their knowledge and skills cannot and should not be guided and mediated through appropriate modeling, coaching, feedback, worked-out examples, and scaffolding by teachers, peers, and educational media. An extensive meta-analysis of the available research came to the conclusion that direct teaching is better than unassisted discovery, but guided, enhanced, or assisted discovery learning is superior to direct teaching (Alfieri et al., 2011).

Powerful Learning Environments to Promote Transition to the Desired End State

Powerful learning environments create the appropriate instructional conditions to evoke in students the learning processes that facilitate gradually the transition from their initial state toward increasing adaptive competence in a domain. Therefore, a challenging task for instructional psychology research consists of elaborating and validating a coherent framework of principles for the design of such powerful learning environments. Starting from the current knowledge and understanding of adaptive competence in a content domain (as described in Section Toward Adaptive Competence as Desired End State) and of the characteristics of effective acquisition processes (see Section Constructive Learning as Lever for the Transition from Initial to the Desired End State), but also based on observation of the practice of excellent teachers, this challenge has been addressed in a substantial number of design experiments. Such design-based research consist in the development and evaluation in real classrooms of complex and fundamentally new instructional interventions, guided by design principles that embody this present understanding of competence, of effective learning processes, and of successful teaching (see e.g., De Corte et al., 2003). Major examples of such orienting principles for the design of powerful learning environments are the following:

1. Learning environments should initiate and support constructive and self-regulated acquisition processes in all students, thus also in the more passive learners. However, as argued above, students' construction of knowledge can be mediated through appropriate guidance by teachers, peers, and educational media. In other words, the claim that productive learning requires good teaching still holds true. But, this first principle also implies that systematic interventions should gradually be removed, so that students progressively become agents of their own learning.
2. Complementary to the first principle, learning environments should allow for the flexible adaptation of instructional support, especially the balance between self-regulation and external regulation, to take into account individual differences among students in cognitive as well as in affective and motivational characteristics. Moreover, the crucial influence on learning of motivational factors points to the necessity of balancing instructional interventions and affective support.
3. Taking into account the situated and collaborative nature of effective learning, powerful learning environments should embed students' constructive acquisition processes as much as possible in (real life) contexts that have personal meaning for the learners, offer ample opportunities for distributed learning through social interaction and cooperation, and are representative of the tasks to which students will have to apply their knowledge and skills in the future. The progressive acquisition of adaptive competence will require extensive experience and practice with the different components involved in a large variety of situations.
4. Because domain-specific knowledge and domain-general heuristics and self-regulation skills play a complementary role in competent learning and problem solving, learning environments should create opportunities to acquire domain-general skills embedded in different subject-matter domains.
5. A variety of activating instructional methods should be used to initiate and support the intended learning processes. In

this respect, a whole array of methods is available that have been shown to foster students' learning: instruction based on modeling, feedback, scaffolding, worked examples, self-explanation, (computer-based) collaborative learning, inquiry, discussion, reflection, coaching and tutoring, visualizations, and computer simulations (see e.g., Mayer and Alexander, 2011)

Cognitive apprenticeship is a well-known model for the design of innovative learning environments that is largely in accordance with those principles (Collins et al., 1989). Although further research is needed to validate the design principles described above, a series of intervention studies in a variety of content domains that embody those guidelines to some degree have already reported supporting empirical evidence. Often these investigations also integrate educational technology, especially computer programs and multimedia systems, in their learning environments (Schauble and Glaser, 1996; De Corte et al., 2003).

Challenges

This selective, and thus incomplete review shows that over the past decades research in instructional psychology made substantial progress toward the elaboration of a theory of learning from instruction. Indeed, we now have a fairly good understanding of what adaptive competence in content domains involves, major characteristics of effective learning processes, and important principles for designing powerful learning environments to elicit and maintain those learning processes in students. But challenging issues need to be addressed in future inquiry of which the following are major examples. First, the expectations about the potential of research in instructional psychology for the improvement of schooling have not been fulfilled. Indeed, the substantial advances made in our understanding of learning and teaching have not resulted in a proportional innovation and improvement of classroom practices (De Corte, 2010). Bridging the gap between theory/research and practice constitutes thus a major joint priority task for educational researchers and professionals. Second, the high expectations about the contribution of information and communication technologies (ICT) for supporting and fostering student learning have until now not by far been met. Research is needed aiming at proving and realizing the incontestable potential of ICT for learning. Finally, a relatively new, challenging strand of research that emerged especially over the past decade is the interdisciplinary area of neuroscience and education. Currently still in its infancy, this field will in the future certainly contribute to the further deepening of our understanding of learning processes, for instance, by linking neural underpinnings to behavioral aspects of learning. However, in view of being relevant for education this domain of research must be conceived as a two-way cooperation with multiple reciprocal interactions between educational research and neuroscience (Bransford et al., 2006).

See also: Apprenticeships and School Learning: General Considerations; Cognitive Development: Mathematics Learning and Instruction; Cooperative Learning in Schools; Education and Learning: Lifespan Perspectives; Educational Learning Theory; Instructional Design; Intelligence, Prior Knowledge, and Learning; Learning Theories and Educational Paradigms; Learning and Instruction: Social-Cognitive Perspectives; Learning to Learn; Metacognitive Development: Educational Implications; Piaget's Theory of Human Development and Education; School Learning for Transfer; Self-Regulated Learning: Theories, Measures, and Outcomes; Situated Learning (Learning In Situ); Tangible User Interfaces in Learning and Education; Teaching for Metacognition; Vygotsky's Theory of Human Development and New Approaches to Education.

Bibliography

Alexander, Patricia, Winne, Philip (Eds.), 2006. Handbook of Educational Psychology, second ed. Erlbaum, Mahwah, NJ.

Alfieri, Louis, Brooks, Patricia, Aldrich, Naomi, Tenenbaum, Harriet, 2011. Does discovery-based instruction enhance learning? Journal of Educational Psychology 103, 1–18.

Bandura, Albert, 1986. Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice Hall, Englewood Cliffs, NJ.

Bendixen, Lisa, Feucht, Florian (Eds.), 2010. Personal Epistemology in the Classroom: Theory, Research, and Implications for Practice. Cambridge University Press, Cambridge, UK.

Bransford, John, Stevens, Reed, Schwartz, Dan, Meltzoff, Andy, Pea, Roy, Roschelle, Jeremy, Vye, Nancy, Kuhl, Pat, Bell, Philip, Barron, Brigid, Reeves, Byron, Sabelli, Nora, 2006. Learning theories and education: toward a decade of synergy. In: Alexander, Patricia, Winne, Philip (Eds.), Handbook of Educational Psychology, second ed. Erlbaum, Mahwah, NJ, pp. 209–244.

Collins, Alan, Brown, John Seely, Newman, Susan, 1989. Cognitive apprenticeship: teaching the craft of reading, writing and mathematics. In: Resnick, Lauren (Ed.), Knowing, Learning, and Instruction. Essays in Honor of Robert Glaser. Erlbaum, Hillsdale, NJ, pp. 453–494.

De Corte, Erik, 2010. Historical developments in the understanding of learning. In: Dumont, Hanna, Istance, David, Benavides, Francisco (Eds.), The Nature of Learning: Using Research to Inspire Practice. OECD Publishing, Paris, pp. 35–67.

De Corte, Erik, Verschaffel, Lieven, Entwistle, Noel, Van Merriënboer, Jeroen (Eds.), 2003. Powerful Learning Environments: Unraveling Basic Components and Dimensions (Advances in Learning and Instruction Series). Elsevier Science, Oxford, UK.

Gagné, Robert, Rohwer, William Jr., 1969. Instructional psychology. Annual Review of Psychology 20, 381–418.

Hattie, John, 2009. Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Learning. Routledge, London.

Mayer, Richard, Alexander, Patricia (Eds.), 2011. Handbook of Research on Learning and Instruction (Educational Psychology Handbook Series). Routledge, New York.

Palincsar, Sullivan, 1998. Social constructivist perspectives on teaching and learning. Annual Review of Psychology 49, 345–375.

Rubie-Davies, Christine (Ed.), 2011. Educational Psychology: Concepts, Research and Challenges. Routledge, New York.

Sawyer, Keith (Ed.), 2006. The Cambridge Handbook of the Learning Sciences. Cambridge University Press, New York.

Schauble, Leona, Glaser, Robert (Eds.), 1996. Innovations in Learning: New Environments for Education. Erlbaum, Mahwah, NJ.

Snow, Richard, Swanson, Judy, 1992. Instructional psychology: aptitude, adaptation, and assessment. Annual Review of Psychology 43, 583–626.

Vosniadou, Stella (Ed.), 2008. International Handbook of Research on Conceptual Change. Routledge, New York.

Zimmerman, Barry J., Schunk, Dale H. (Eds.), 2011. Handbook of Self-Regulation of Learning and Performance. Routledge, New York.

Instrumental Variables in Statistics and Econometrics

James H Stock, Harvard University, Cambridge, MA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7577–7582, © 2001, Elsevier Ltd.

Abstract

The method of instrumental variables (IVs) is a general approach to the estimation of causal relations using observational data. This method can be used when standard regression estimates of the relation of interest are biased because of reverse causality, selection bias, measurement error, or the presence of unmeasured confounding effects. The central idea is to use a third, 'instrumental' variable to extract variation in the variable of interest that is unrelated to these problems, and to use this variation to estimate its causal effect on an outcome measure. This article describes IV estimators, discusses the conditions for a valid instrument, and describes some common pitfalls in the application of IV estimators.

The method of instrumental variables (IVs) is a general approach to the estimation of causal relations using observational data. This method can be used when standard regression estimates of the relation of interest are biased because of reverse causality, selection bias, measurement error, or the presence of unmeasured confounding effects. The central idea is to use a third, 'instrumental' variable to extract variation in the IV of interest that is unrelated to these problems, and to use this variation to estimate its causal effect on an outcome measure. This article describes IV estimators, discusses the conditions for a valid instrument, and describes some common pitfalls in the application of IV estimators.

The Method of IVs

Common Problems with Standard Regression Analysis of Observational Data

In many cases in the social and behavioral sciences, one is interested in a reliable estimate of the causal effect of one variable on another. For example, suppose a mayor is considering increasing the size of the police force; what is the effect of an additional police officer on the crime rate? Or, what is the effect of an additional year of schooling on future earnings? What will happen to economic growth if the central bank raises short-term interest rates by 1% point? What is the effect of a new medical procedure on health outcomes? These and many other questions require estimates that are causal, in the sense that they are externally valid and can be used to predict the effect of changes in policies or treatments, holding other things constant.

In theory, such causal effects could be estimated by a suitably designed randomized controlled experiment. Very often, however, as in the first three questions, such an experiment could be prohibitively expensive, could be unethical, and/or could have questionable external validity. Even when randomized controlled experiments are available, such as clinical trials of medical procedures, it is of interest to validate the experimental predictions using information on outcomes in the field. Thus, to address such questions empirically typically entails the use of nonexperimental, i.e., observational, data.

Unfortunately, standard regression analysis of observational data can fail to yield reliable estimates of causal effects for many reasons, four of which are particularly salient. First, there could be additional unmeasured effects, leading to 'omitted variable bias'; for example, the educational attainment of parents is correlated with that of their children, so if parents' education facilitates learning at home but is unobserved then the correlation between years of school and earnings could overstate the true, causal effect of school on earnings. Second, there might be reverse causality, or 'simultaneous equations' bias; for example, more police officers might reduce crime, but cities with higher crime rates might demand more police officers, so standard regression analysis of crime rates on the number of police confounds these two different effects. Third, there could be selection bias, in which those most likely to benefit from a treatment are also most likely to receive it; for example, because ambition is one reason for success both at school and in the labor market, the measured correlation between years of school and earnings could simply reflect the effect of unmeasured ambition. Fourth, standard regression estimates of the causal effect are biased if the regressor is measured with error (*see* Measurement Error Models).

IV Regression

IV regression provides a way to handle these problems. The main early application of IV methods was for estimating the parameters of a system of linear simultaneous equations, and this remains a good expositional framework. Let y_i denote the outcome variable of interest (say, log future earnings), let Y_i denote the r treatment variables (years of education), let X_i denote K_1 additional observed control variables, and let Z_i denote the K_2 IVs, where these are all observed for observations $i = 1,\ldots,N$. Also let β and γ be unknown parameter vectors, let Φ and $\prod$ be matrices of unknown parameters, and let β' denote the transpose of β. Suppose that these variables are linearly related as

$$y_i = \beta' Y_i + \gamma' X_i + u_i \qquad [1]$$

$$Y_i = \Phi X_i + \prod Z_i + V_i \qquad [2]$$

Where u_i and V_i are 'error terms' that represent additional unobserved effects, measurement error, etc. The coefficient of interest is β; in the schooling example, this is the percentage

change in future earnings caused by attending school for an additional year.

In this notation, the four problems listed in the preceding paragraph have a common implication, that the correlation between Y_i and u_i is nonzero, and in consequence the ordinary least squares (OLS) estimator of β will be biased and, in large samples, inconsistent. It is assumed that $\text{corr}(X_i, u_i) = 0$; at this level, this assumption can be made without loss of generality, for if this is suspected to be false for some element of X, then it should be listed instead in Y. In the terminology of simultaneous equations theory, y and Y are endogenous variables and X is exogenous.

The key idea of IV methods is that although Y is correlated with u, if Z is uncorrelated with u (that is, if Z is exogenous), then Z can be used to estimate β. Intuitively, part of the variation in Y is endogenous and part is exogenous; IV methods use Z to isolate exogenous variation in Y and thereby to estimate β. More formally, consider the simplest case, in which there is no X in eqns [1] or [2] and which Y and Z are single variables (i.e., $K_1 = 0$ and $r = K_2 = 1$). Then $\text{cov}(y, Z) = \beta\ \text{cov}(Y, Z) + \text{cov}(u, Z) = \beta\ \text{cov}(Y, Z)$. Thus, $\beta = \text{cov}(y, Z)/\text{cov}(Y, Z)$. This leads to the IV estimator, $\widehat{\beta}^{\text{IV}} = s_{yz}/s_{YZ}$, where s_{yZ} is the sample covariance between y and Z. Evidently, if these two sample covariances are consistent and if $\text{cov}(Y, Z) \neq 0$, then $\widehat{\beta}^{\text{IV}} \xrightarrow{\text{P}} \beta$. The availability of the instrument Z thus permits consistent estimation of β.

If more than r instruments are available, it makes sense to use the extrainstruments to improve precision. There are, however, many ways to do this, since each subset of r instruments will produce its own estimate of β.

The most common way to combine instruments is to use two-stage least squares (2SLS). In the first stage, eqn [2] is estimated by OLS, producing the predicted values $\widehat{Y}$. In the second stage, y is regressed against $\widehat{Y}$ and X, yielding the 2SLS estimator $\widehat{\beta}^{\text{2SLS}}$. With only one instrument, this reduces to $\widehat{\beta}^{\text{IV}}$ given above. Under Gaussian disturbances, this provides an asymptotically efficient method for weighting the various instruments that is easy to understand and to implement.

Conditions for Valid Instruments

The question of whether a candidate set of instruments can be used to estimate β is a special case of the more general problem of identification of parameters in multivariate econometric and statistical models. In the context of eqns [1] and [2], when there is a single Y, the requirements for instrument validity are simple. Then a potential instrument Z must satisfy two conditions: (1) Z is uncorrelated with u; and (2) the partial correlation between Y and Z, given X, is nonzero. The first condition states that the instrument is exogenous. The second condition is that the instrument is relevant. Together, these conditions permit using Z to isolate the exogenous variation in Y and thus to identify β. For example, when $r = K_2 = 1$ and there are no Xs, if Z satisfies the two conditions then β is identified by the moment condition derived above, $\beta = \text{cov}(y, Z)/\text{cov}(Y, Z)$.

When Y is a vector (i.e., $r > 1$), the exogeneity condition (1) remains but the relevance condition (2) is somewhat more complicated. For identification of the independent effects of each Y (all elements of β), a sufficient condition for instrument validity is that the covariance matrix of Z has full rank and that $\prod$ has full row rank. Clearly, a necessary condition for $\prod$ to have full row rank is that there are at least as many instruments as Ys ($K_2 \geq r$). The number of excess instruments, $K_2 - r$, is referred to as the degree of over identification of β.

When y is a vector, so that eqn [1] is itself a system of multiple equations, the conditions for identification become intricate (*see* Simultaneous Equation Estimation: Overview; Statistical Identification and Estimability). For a rigorous and complete treatment in the linear simultaneous equations framework, see Rothenberg (1971) and Hsiao (1983).

IV Estimators and Their Distributions

Linear Models

In addition to two-stage least squares, other estimators are available for IV estimation of a single equation that is linear in the unknown parameters. The leading alternative estimator is the so-called limited information maximum likelihood (LIML) estimator. The LIML estimator minimizes the Anderson–Rubin (1949) statistic,

$$F(\beta) = \frac{\tilde{y}'\left(\boldsymbol{M}_X - \boldsymbol{M}_{[X\,Z]}\right)\tilde{y}\Big/K_2}{\tilde{y}'\boldsymbol{M}_{[X\,Z]}\tilde{y}\Big/(N - K_1 - K_2)} \qquad [3]$$

$$\tilde{y} = (\tilde{y}_1 \ldots, \tilde{y}_N)', \quad \text{where } \tilde{y}_i = y_i - \beta' \text{Y}_i, \text{ and}$$
$$M_X = \mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}', \quad \text{where } \boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)'$$

This can be solved analytically as an eigenvalue problem. Many other single-equation IV estimators have been proposed, but these are rarely used in empirical applications; see Hausman (1983) for expressions for the LIML estimator and for a discussion of other estimators.

Under standard assumptions (fixed numbers of regressors and instruments, validity of the instruments, convergence of sample moments to population counterparts, and the ability to apply the central limit theorem), the LIML and 2SLS estimators are asymptotically equivalent and have the same asymptotic normal distribution. That is, $\sqrt{\text{N}}(\widehat{\beta}^{\text{LIML}} - \beta)$ and $\sqrt{\text{N}}(\widehat{\beta}^{\text{2SLS}} - \beta)$ both have asymptotic normal distributions with the same covariance matrix. However, their finite sample distributions differ. There is a large literature on finite sample distributions of these estimators (*see* Simultaneous Equation Estimates (Exact and Approximate), Distribution of). Because these exact distributions depend on nuisance parameters, such as $\prod$, that are unknown and because few IV statistics are (exactly) pivotal, this literature generally does not provide approximations that are useful in empirical applications. However, some general guidelines do emerge from these studies. Perhaps the most important is that when there are many instruments and/or when $\prod$ is small in a suitable sense, LIML tends to exhibit less bias than 2SLS and LIML confidence intervals typically have better coverage rates than 2SLS.

These methods extend naturally to the case that y is a vector and eqn [1] is a system of equations. This entails imposing

restrictions on the coefficients in eqns [1] and [2] that are implied by the model, and then using the available instruments to estimate simultaneously all these coefficients with these restrictions imposed. The system analog of 2SLS is three-stage least squares, in which all the unknown coefficients are estimated simultaneously using an estimate of the variance–covariance matrix of u obtained from the 2SLS estimates. The system analog of LIML is full information maximum likelihood. For further discussion, *see* Simultaneous Equation Estimation: Overview; Simultaneous Equation Estimates (Exact and Approximate), Distribution of.

Nonlinear Models

Many details of the foregoing discussion hinged on the system being linear in the parameters and variables. In many settings, however, theory suggests models that are nonlinear rather than linear. A leading class of examples is consumption-based models of asset prices, in which the equations contain nonlinearities inherited from posited utility functions of representative agents. In such settings, eqn [1] is usefully extended and recast to incorporate these nonlinearities. Specifically, suppose that the model at hand implies,

$$E[h(y_i, Y_i, X_i, \theta)|X_i, Z_i] = 0 \qquad [4]$$

where as before y_i and Y_i are endogenous variables, X_i and Z_i are exogenous variables, and θ is a vector of parameters. For example, in the linear system eqns [1] and [2], eqn [4] is implied by $E(u_i|X_i, Z_i)$ upon setting $h(y_i, Y_i, X_i, \theta_0) = y_i - \beta' Y_i - \gamma' X_i$ and $\theta = (\beta, \gamma)$.

The orthogonality condition eqn [4] corresponds to the first condition for a valid instrument, that Z is exogenous. If in addition Z is relevant, then eqn [4] can be exploited to estimate the unknown parameters θ. The issue of relevance of Z is equivalent to whether θ is identified. The question of identification in nonlinear models is complex and little can be said about global identification at a general level, although conditions for local identification sometimes yield useful insights (*see* Statistical Identification and Estimability).

Given a model [4] and a set of instruments, in general a variety of estimators of θ are available. Estimation is typically undertaken by minimization of a quadratic form in $h(y_i, Y_i, X_i, \theta)$, times the instruments; 2SLS and LIML estimates obtain as special cases when h is linear. In practice, the choice of the weighting matrix in this quadratic form is important, and if there is serial correlation and/or heteroskedasticity this weighting matrix (and the standard errors) must be computed in a way that accounts for these complications. Such estimators are referred to as generalized method of moments (GMM) estimators. For discussions of identification and estimation in GMM, see Newey and McFadden (1994).

A Brief History of IV Methods

S. Wright (1925) and his father P. Wright (1928) introduced the estimator $\widehat{\beta}^{IV}$ (defined in Section IV Regression) and used it to estimate supply and demand elasticities for butter and flaxseed; this work, however, was neglected (see Goldberger, 1972). The development of IV methods in the 1940s stemmed from the attempts of early statisticians and econometricians to solve what appeared to be two different problems: the problem of measurement error in all the variables, and the problem of reverse causality arising in a system of simultaneous equations that describe the macroeconomy. The term 'instrumental variables' was first used in print by ReiersØ1 (1945), who developed IV estimation for the errors in problem of variables and extended work by Wald (1940) which, in retrospect, also can be seen as having introduced IV estimation. Concurrently, work by Haavelmo (1944) and the Cowles Commission emphasized likelihood-based methods of analyzing simultaneous equation systems, which implicitly (through identification conditions) entailed the use of IVs; Anderson and Rubin (1949) developed LIML as part of this research program.

An important breakthrough came with the development of 2SLS by Basmann (1957) and Theil (1958), which permitted computationally efficient estimation in single equations with multiple instruments. Sargan (1958) introduced IV estimation for multiple equation systems and Zellner and Theil (1962) developed three-stage least squares.

The extension of these methods to nonlinear models was undertaken by Amemiya (1974) (nonlinear 2 SLS) and by Jorgenson and Laffont (1974) (nonlinear three-stage least squares). The modern formulation of GMM estimation is due to Hansen (1982). GMM constitutes the dominant modern unifying framework for studying the issues of identification, estimation, and inference using IVs.

Potential Pitfalls

In practice, inference using IV estimates can be compromised because of failures of the model and/or failures of the asymptotic distribution theory to provide reliable approximations to the finite sample distribution. Much modern work on IV estimation entails finding ways to avoid, or at least to recognize, these pitfalls. This section discusses three specific potentially important pitfalls in linear IV estimation. These problems, as well as additional problems associated with estimation of standard errors, also arise in nonlinear GMM estimation, but their sources and solutions are less well understood in the general nonlinear setting.

Some Instruments Are Endogenous

If an instrument is endogenous then $\text{corr}(Z_i, u_i) \neq 0$, and reflection upon the derivation $\widehat{\beta}^{IV}$ and its probability limit in Section IV Regression reveals that $\widehat{\beta}^{IV}$ is no longer consistent for β. More generally, if at least one instrument is endogenous then the 2SLS and LIML estimators are inconsistent. If at least r instruments are exogenous, however, then it is possible to test the null hypothesis that all instruments are exogenous against the alternative that at least one (but no more than $K_2 - r$) is endogenous. In practice, a researcher might have at least r instruments that he or she firmly believes to be exogenous, but might have some additional instruments which are more questionable.

Testing the null that all the instruments are exogenous entails checking empirically the assumption that the instruments are uncorrelated with u_i. This can be done by regressing the IV residual (estimated by 2SLS or LIML) against Z. The R^2 of this regression should be zero; if NR^2 exceeds the desired $\chi^2_{K_2-r}$ critical value, the null hypothesis of joint exogeneity is rejected.

Although this statistic provides a useful diagnostic and rejection suggests that the full instrument list is (jointly) invalid, failure to reject is not necessarily reassuring since the maintained hypothesis is that at least r of the instruments are valid. Moreover, this test can only be implemented if β is overidentified ($K_2 > r$). The hypothesis that at least r instruments are exogenous is both essential and untestable, and thus must be contemplated carefully in IV applications.

Weak Instruments

The condition for instrument relevance states that Z and Y must have nonzero partial correlation (given X). In practice, this correlation, while arguably nonzero, is often small, a situation sometimes referred to as the problem of weak instruments. When the instruments are weak, the usual large sample approximations provide a misleading basis for inference: the 2SLS estimator in particular is biased toward the OLS estimator, and asymptotic LIML and 2SLS confidence regions have coverage rates that can differ substantially from the nominal asymptotic confidence level. When $r = 1$, an empirical measure of the strength of the instruments is the F-statistic testing the hypothesis $\Pi = 0$ in the first-stage regression eqn [2]. The risk of weak instruments is especially relevant when there are many instruments, for even if some instruments have a large partial correlation with Y, if there are many instruments quite a few of them could be weak so taken together this first-stage F-statistic could be small.

Although no single preferred way to handle weak instruments has yet emerged, one solution is to construct confidence intervals by inverting the Anderson–Rubin statistic eqn [3]. This is readily done because $F(\beta_0)$ has an asymptotic $\chi^2_{k_2}$ distribution under the joint null hypothesis that $\beta = \beta_0$ (a fixed vector) and $\text{corr}(X, u) = \text{corr}(Z, u) = 0$ (the instrument relevance condition is not needed for this result).

Heterogeneity of Treatment Effects in Microdata

Many modern applications of IV are to microdata sets with cross-sectional or panel data, that is, data sets where the observation units are individuals, firms, etc. An important consideration that arises in this context is the role of heterogeneity in interpreting what it is that IV methods measure. To make this concrete, consider a modification of eqns [1] and [2], where for simplicity it is assumed that $r = K_2 = 1$ and the only X is a constant. Suppose, however, that there is heterogeneity both in the responses of each individual to the 'treatment' Y and in the influence of the instrument on the level of treatment received. In equations, this can be written as,

$$Y_i = \mathbf{\Phi} + \prod_i Z_i + V_i \qquad [6]$$

where β_i and $\prod_i$ vary randomly across individuals. Suppose that the instrument is distributed independently of (u_i, V_i, β_i, $\prod_i$), and technical conditions ensuring convergence of sample moments hold. Then $\widehat{\beta}^{2SLS} \xrightarrow{p} E[\prod_i \beta_i]/E\prod_i$.

Evidently, with heterogeneity of this form 2SLS can be thought of as estimating $E[\prod_i \beta_i]/E\prod_i$, which differs from the usual estimand $\beta = E\beta_i$ if β_i and $\prod_i$ are correlated. For example, suppose that $\prod_i = 0$ for half the population, that $\prod_i = \prod$ (which is fixed) for the other half, and that the sample is drawn randomly. If β_i differs systematically across the two halves of the population, then 2SLS estimates the mean of β_i among that half for which $\prod_i = \prod$; this differs from the mean of β_i over the full population. More generally, 2SLS can be thought of as estimating a weighted average of the individual treatment effects, where the weights reflect the influence of Z_i on whether the individual receives the treatment. This has been referred to as the 'local average treatment effect'; see Angrist et al. (1996).

Where Do Valid Instruments Come from?

In practice the most difficult aspect of IV estimation is finding instruments that are both exogenous and relevant. There are two main approaches, which reflect two different perspectives on econometric and statistical modeling. The first approach is to use a priori theoretical reasoning to suggest instruments. This is most compelling if the model being estimated is itself derived from a formal theory. An example of this approach is the estimation of intertemporal consumption-based asset pricing models, mentioned in Section Nonlinear Models in which previously observed variables are, under the model, uncorrelated with certain future expectational errors. Thus previously observed variables can be used as instruments.

The second approach to constructing instruments, more commonly found in program evaluation studies (broadly defined), is to look for some exogenous source of variation in Y that either derives from true randomization or, in effect, from pseudorandomization. In randomized experiments in the social sciences, compliance with experimental protocol is usually unenforceable, so a subject's decision whether to take the treatment introduces selection bias; assignment to treatment, however, can be used as an instrument for receipt of treatment. In nonexperimental settings, this reasoning suggests looking for a variable that plays a role similar to random assignment in a randomized experiment. For example, McClellan et al. (1994) investigated the effect of the intensity of treatment (Y) on mortality after 4 years (y) using observational data on elderly Americans who suffered an acute myocardial infarction (heart attack). Standard regression analysis of these data would be subject to selection bias and omitted variables bias, the former because the decision to pursue an intensive treatment depends in part on the severity of the case, the latter because of additional unobserved health characteristics. To avoid these biases, they used as instruments the distance of the patient from hospitals with various degrees of experience treating heart attack patients, for example, the differential distance to a hospital with experience at cardiac catheterization.

If distance to such hospitals is distributed randomly across potential heart attack patients, then it is exogenous; if this distance is a factor in the decision whether to move the patient to the distant hospital for intensive treatment, then it is relevant. If both are plausibly true, then 2SLS provides an estimate of the average treatment effect for the marginal patients, where the marginal patients are those for whom the effect of distance on the decision to treat is most important.

From the humble start of estimating how much less butter people will buy if its price rises, IV methods have evolved into a general approach for estimating causal relations throughout the social and behavioral sciences. Because it requires valid and relevant instruments, IV regression is not always an option, and even if it is, the practitioner must be aware of its potential pitfalls. Still, when they can be applied, IV methods constitute perhaps our most powerful weapon against omitted variable bias, reverse causality, selection bias, and errors-in-variables in our efforts to estimate causal relations using observational data. For recent textbook treatments of IVs issues see Hayashi (2000: Chapters 3 and 4) and Ruud (2000: Chapters 20–22 and 26).

See also: Selection Bias, Statistics of.

Bibliography

Amemiya, T., 1974. The nonlinear two-stage least squares estimator. Journal of Econometrics 2, 105–110.

Anderson, T.W., Rubin, H., 1949. Estimation of the parameters of a single equation in a complete systems of stochastic equations. Annals of Mathematical Statistics 20, 46–63.

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables (with discussion). Journal of the American Statistical Association 91, 444–472.

Basmann, R.L., 1957. A generalized classical method of linear estimation of coefficients in a structural system of stochastic equations. Annals of Mathematical Statistics 20, 46–63.

Goldberger, A.S., 1972. Structural equation methods in the social sciences. Econometrica 40, 979–1001.

Haavelmo, T., 1944. The probability approach to econometrics. Econometrica 12 (Suppl.), 1–118.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.

Hausman, J.A., 1983. Specification and estimation of simultaneous equation models. In: Grilliches, Z., Intrilligator, M.D. (Eds.), Handbook of Econometrics, vol. 1. North Holland, Amsterdam, pp. 391–450.

Hayashi, F., 2000. Econometrics. Princeton University Press, Princeton, NJ.

Hsiao, C., 1983. Identification. In: Grilliches, Z., Intrilligator, M.D. (Eds.), Handbook of Econometrics, vol. 1. North Holland, Amsterdam, pp. 223–283.

Jorgenson, D.W., Laffont, J., 1974. Efficient estimation of nonlinear simultaneous equations with additive disturbances. Annals of Economic and Social Measurement 3, 615–640.

McClellan, M., McNeil, B.J., Newhouse, J.P., 1994. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. Journal of the American Medical Association 272, 859–866.

Newey, W.K., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, vol. 4. North Holland, Amsterdam, pp. 2113–2247.

ReiersØ1, 1945. Confluence analysis by means of instrumental sets of variables. Arkiv for Mathematik Astronomi och Fysik 32, 1–119.

Rothenberg, T.J., 1971. Identification in parametric models. Econometrica 39, 577–595.

Ruud, P.A., 2000. An Introduction to Classical Econometric Theory. Oxford University Press, New York.

Sargan, J.D., 1958. On the estimation of economic relationships by means of instrumental variables. Journal of the Royal Statistical Society, Series B 21, 91–105.

Theil, H., 1958. Economic Forecasts and Policy. North Holland, Amsterdam.

Wald, A., 1940. The fitting of straight lines if both variables are subject to error. Annals of Mathematical Statistics 11, 284–300.

Wright, S., January 1925. Corn and Hog Correlations. US Department of Agriculture Bulletin 1300, Washington, DC.

Wright, P.G., 1928. The Tariff on Animal and Vegetable Oils. Macmillan, New York.

Zellner, A., Theil, H., 1962. Three-stage least squares: simultaneous estimation of simultaneous equations. Econometrica 30, 54–78.

Insurance

Christian Gollier, University of Toulouse, Toulouse, France

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7582–7587, © 2001, Elsevier Ltd.

Abstract

The socially efficient allocation of risks in an economy of risk-averse agents is examined. This allows the potential social value of insurance to be measured. The reasons why most risks can be only partially insured are transaction costs, adverse selection, risqué moral, risk of fraud and limited liability, are then explored. The regulation of the insurance sector is also examined.

There is insurance when one party agrees to pay an indemnity to another party in the event of the occurrence of a predefined random event generating a damage for the latter. The most standard example is the insurance contract where the first party – the insurer – is compensated by being paid a premium by the second party, the policyholder. But many other contracts involve some forms of insurance. For example, in share-cropping contracts, the landlord agrees to reduce the rent of land in case of a bad crop. In many instances, employers agree to pay their employees more than the competitive wage in bad years. In cost-plus contracts, the buyer accepts to pay a larger price if the producer incurred an unexpected increase in cost. In the case of income taxes, the state partially insures the losses of taxpayers by reducing the tax payment when incomes are low.

The shifting of risk is of considerable importance for the functioning of our modern economies. Insurance is desirable for risk-averse agents as a risk-reduction device through the use of the law of large numbers. Moreover, insurance allows for disentangling investment decisions from risk-taking decisions. Without it, there would certainly not have been the experience of the historical economic growth of the twentieth century. Ford, Solvay, Rockefeller, and the others would not have taken the investment risks that they actually took without the possibility to share the risk with shareholders and insurers. Similarly, most consumers would not purchase new expensive cars or houses if they would not be able to insure them.

However, various informational problems (observation of the risk, solvency issues, of the prevention efforts and/or of the loss by the insurer) can make competitive insurance markets inefficient by not providing enough coverage at an acceptable price.

The Social Value of Insurance

There is an added value to insurance only because policyholders are risk-averse, that is they dislike zero-mean risk on wealth. Consider an agent facing a random loss X to their wealth. An insurance contract stipulates a premium P and an indemnity schedule I(.) that determines the indemnity I(X) for each possible loss X. There is full coverage if I(.) is the identity function. The actuarial value of the contract is the expected indemnity EI(X). The insurance premium is said to be actuarially fair if it is equal to the actuarial value of the contract, $P = EI(X)$. Suppose that it is the case. Then, the purchase of a full insurance contract at an actuarially fair premium has the effect of replacing a random loss X by its expectation $P = EX$. The private value of such a contract is equal to the risk premium attached to the initial risk by the policyholder. It is increasing with the policyholder's degree of risk aversion and with the riskiness of the loss. As a first approximation, the riskiness of the loss can be measured by its variance.

In order to determine the social value of insurance, the effect of the transfer of the risk on the insurer's welfare should also be measured. Consider an insurer selling n fair full insurance contracts to n policyholders, each of them bearing a random loss X_i, with $X_1, \ldots, X_n$ being identically and independently distributed. The common wisdom is that the insurer does not bear any risk in the aggregate, if it is able to sell enough such insurance contracts to cover independent risks, i.e., if n is large enough. This is not true, as explained by Samuelson (1963). It is a fallacious interpretation of the law of large numbers, which leads people to believe that accumulating several independent risks generates diversification. Indeed, the aggregate indemnity to be paid by the insurer is the sum of the X_i. Its riskiness measured by its variance equals $n\sigma^2$, where σ^2 is the variance of each individual risk X_i. If the insurer is risk-averse, the risk transfer is not Pareto-improving, since it makes the insurer worse off. The classical view of modern finance on this problem is to recognize that insurance companies are not owned by a single person, but rather by a large set of shareholders. If insurable losses are uncorrelated with financial markets returns, the riskiness of insurance companies has no adverse effect on shareholders. In such a situation, the risk transfer to insurance companies is Pareto improving, and the social value of insurance companies can be measured by the sum of the risk premia associated by policyholders to their random loss.

An alternative insurance scheme is to organize a mutual arrangement among the n risk-averse agents. Consider an arrangement in which the pool guarantees to each of its member a complete coverage of losses against an *ex-post* contribution equaling the average loss in the pool. In such a scheme, the random loss X_i is replaced by $(X_1 + \cdots + X_n)/n$, whose variance is $1/n$ the variance of X_i. In consequence, this mutual arrangement is Pareto-improving. At the limit, when n tends to infinity, individual risks are completely washed out by diversification.

Efficient Risk Sharing with Complete Markets

An important question is to characterize socially efficient allocations of risks in an economy where individuals may have

very different initial risk exposures and risk attitudes. Wilson (1968) solved this question when individuals are expected – utility maximizers. Any efficient allocation of risk must satisfy two conditions. The first one is often called the mutuality principle. It states each *ex-post* individual's consumption only depends upon the ex-post average wealth in the economy. This means that all members raise their consumption in good states, and reduce it in bad states. There is no loser in an improving environment. Notice that this condition is satisfied in the case of the mutual arrangement examined above. The intuition for the mutuality principle is the need for diversification: all individual risks must be gathered in the pool. The second characteristic of a socially efficient risk-sharing rule deals with the sharing of the aggregate risk in the pool. As suggested by the intuition, a socially efficient sharing of the aggregate risk is such that the risk borne by a member of the pool must be inversely proportional to their Arrow-Pratt degree of absolute risk aversion. At the limit, if there is a risk-neutral agent in the economy, this agent should bear all risks, i.e., he should insure everyone else in the economy.

Whether decentralized economies can allocate risks in an efficient way is another question that has been solved by the pioneering works of Arrow (1953), Debreu (1959), and Borch (1962). Suppose that the economic system is such that it is possible to freely insure against any bet, at fixed odds, any amount wished on the occurrence of any event, which will not affect the welfare of the insurers in any way. Markets are said to be complete in this case. Suppose also that there is no transaction cost on these markets, and no asymmetric information of any kind. The standard first theorem of welfare economics can then be applied to obtain that competition on these markets yields a Pareto-efficient allocation of risks in this economy. The competitive allocation must in particular satisfy the mutuality principle. At the competitive equilibrium, all households should sell their entire human and physical capital in exchange for a stake of the aggregate economy. A part of this story can be found in the Capital Asset Pricing Model in which all investors have the same composition of the portfolio of risky assets, the 'market portfolio.'

Townsend (1994) tested the mutuality principle in small rural villages in India. He did not reject the hypothesis of efficient risk sharing within each village. But he rejected the hypothesis of efficient risk sharing across villages. More generally, it is quite obvious that the mutuality principle does not hold in large economies. Some workers may see their income reduced in a booming economy. In case of a labor-saving macroeconomic shock, workers may lose against capital owners. Some classes of retirees may enjoy an increase in their pension benefit even when there is a recession. Some countries may be in recession in a growing world economy. Shiller (1993) recommends that new institutions be created to share these society's large economic risks that are currently inefficiently shared. The welfare impacts of imperfect risk sharing are likely to be very large. In the remainder of this article, the reasons for these market imperfections are examined.

Optimal Insurance with Transaction Costs

Insurance entails transaction costs. In many lines of casualty insurance, transaction costs may be as large as 30% of the premium. Consider an economy with a risk-neutral insurer and consumers with idiosyncratic risks X. Suppose that each euro of indemnity yields k euros of deadweight cost. In a competitive insurance market, the premium associated with insurance contract I(.) must be equal to $(1 + k)\mathrm{EI}(X)$. Mossin (1968) showed that the optimal insurance contract for the policyholder is no longer the full insurance contract. It is always optimal for him to retain part of the risk in order to reduce the insurance cost. This is because risk is a second-order effect relative to transaction costs.

Arrow (1965) examined the optimal form of the risk retention. There are various ways for policyholders to retain a share of the risk. The most standard is to accept a straight deductible, in which the indemnity is either zero, if the loss is less than the prespecified deductible, or the loss minus the deductible otherwise. Alternatively, the insurance contract can contain a coinsurance rule in which case, the indemnity is a prespecified percentage of the loss. Other clauses can be considered, as upper caps on indemnities, disappearing deductibles, etc. Arrow showed that the optimal insurance contract takes the form of a straight deductible. Any nondeductible insurance contract is dominated by a straight deductible contract with the same actuarial value, but yielding a smaller retained risk for the policyholder. Straight deductibles provide the best compromise between the willingness to cover the risk and the limitation of the insurance deadweight cost.

One can link transaction costs to undiversifiable risks. Obviously, many natural, environmental, or technological risks are in the class of large risks that are difficult to eliminate by using the mutuality principle. The insurer's risk-neutrality for these risks may be questioned. Insurance companies will not provide fair insurance premiums for them. Indeed, shareholders will not be able to diversify the risk associated to the dividends paid by insurance companies that cover these large risks. They will ask for a risk premium, which will increase the cost of capital of these companies. This cost will be passed on to policyholders through a larger premium rate for the component of individual risks that is systematic. It will induce them to retain part of their individual risk. In short, the fact that the risk is systematic induces insurance premiums to contain a positive loading that has an effect equivalent to a transaction cost. This is the logic behind larger deductibles for systematic risks.

Adverse Selection

Since the seminal paper by Rothschild and Stiglitz (1976), it is recognized that the fact that insurers face a heterogeneous population of consumers is a source of inefficiency on insurance markets. The classical model presented above allows for a heterogeneous population as long as the characteristics of the risk borne by each agent is common knowledge. For example, the fact that women are safer drivers than men is compatible with full insurance of every driver at the competitive equilibrium with a risk-neutral insurance industry. The premium rate for every category of risk will be fair, thereby inducing each individual to purchase full insurance at the optimum.

A problem arises when the population is heterogeneous, but the observable characteristics of the agents are not perfectly

correlated to the intensity of their risk. The adverse selection problem originates from the observation that if insurance companies calculate the premium rate on the basis of the average probability distribution in the population, the less risky agents will purchase less insurance than riskier agents. In the extreme case, the low-risk agent will find the premium rate too large with respect to their actual probability of loss. They will prefer not to insure their risk. Insurers will anticipate this reaction, and they will increase the premium rate to break even, only on the population of high-risk policyholders. The literature on adverse selection is devoted to characterizing an equilibrium. Insurers will use the fact that low-risk agents and high-risk agents behave differently in the face of a large set of insurance contracts. In particular, low-risk agents could credibly signal their type by selecting a contract with a large deductible, something that high-risk agents dislike.

The presence of high-risk agents generates a negative externality to lower-risk agents who are unable to find an insurance premium at an acceptable premium rate. To illustrate, this is probably why the proportion of households that purchase life insurance is so small, despite the potential severity of the risk. People have private information about their health status that cannot be observed by insurance companies. Then, only those with the lowest life expectancy purchase life insurance.

The standard policy recommendation for improving risk-sharing efficiency under an adverse selection problem is to make public all relevant information about risks. For example, insurers should be allowed to know whether the potential policyholder has some severe illness. They should also be allowed to use genetic testing. Obviously, redistributional reasons may go against such a policy recommendation if the state is not in a position to compensate poor high-risk agents.

Moral Hazard

The population of risks can be heterogeneous not only because agents bear intrinsically different risks, but also because they do not invest the same amount of their energy, wealth, or time in risk prevention. In particular, it has long been recognized that individuals who are better covered by insurance invest less in risk prevention, if the link between the premium rate and the size of these investments is weak. It will be the case, if insurers are not in a position to observe the investment in risk prevention by the policyholder. In that case, the premium rate is not sensitive to the effort made by the policyholder to prevent losses. Obviously, there will be an inverse relationship between risk prevention and insurance coverage; policyholders will not internalize the benefits of their efforts. The level of risk prevention will be inefficient. This is *ex ante* moral hazard, common to all principal-agent problems. Anticipating this low degree of prevention and the higher frequency of losses that it entails, insurers will raise their premium rate. Full insurance will not be optimal for agents. At the limit, no insurance can be an equilibrium. Holmstrom (1979) characterized the equilibrium insurance contract with *ex ante* moral hazard.

To illustrate, this is why it is not possible to insure against the absence of promotion on the workplace, about failure at school or university, about the lack of demand for a new product, or about divorce. To some extent, this is also why it is hard to insure against unemployment, or against environmental and technological risks.

The policy recommendation to fight against *ex ante* moral hazard is the enforcement of norms for risk prevention. This is the case for environmental risks in which ships transporting chemical products have to satisfy various safety requirements that are imposed by regulatory agencies. Automobile driving norms are also standard, as speed limits, alcohol-free driving, etc. Why these norms are mostly organized by a regulatory agency rather than by insurers is not completely clear. One reason is due to the combination of negative externalities and limited liability. If there is more than one principal supervising the implementation of norms, the information among the different principals should be pooled to save on monitoring costs. For example, auto insurers should be allowed to get the information about driver fines by the police.

Another policy recommendation is to allow insurers to discriminate prices among different policyholders that exercise various preventive efforts. Allowing for discrimination is a way to provide incentive to policyholders to invest in risk-reducing activities. In France again, insurers are not allowed to discriminate homeowner premium rates on the basis of natural risks such as earthquake, flood, and storms. The consequences are by now notorious: many households built their houses in areas that were secularly known to be flooded periodically.

Insurance Fraud

The classical model assumed that the size of the loss is observable. There are many instances in which this is at best a crude approximation of the real world. Contracts can be made contingent only upon observable events. The problem here is to give the good incentives to the policyholder to report their actual loss. The difficulty for insurers to verify claims is at the origin of why it is not possible to insure against loss of happiness, or against some forms of suffering that cannot be measured by physicians.

There exist other types of risk for which outcomes can be observed by the insurer only at a relatively high auditing cost. Townsend (1979), Mookherjee and P'ng (1989), and others analyzed the optimal risk-sharing scheme in this case. If there is no limit on the penalty that can be imposed to policyholders that do not declare the actual level of their loss, the first-best solution can be attained. Indeed, insurers should announce that they will audit claims with some probability p that is very low. If the policyholder made a fraudulent claim, an unbounded penalty ('death penalty') would be imposed on him. This is enough to give the incentive not to defraud the insurance contract, even if p is very small. In this case, the fact that there is costly claim verification is not detrimental to welfare, and the risk is insurable in full.

But there are several reasons to believe that an unbounded penalty in case of a fraudulent claim is not a realistic assumption: ethics, limited liability, risk of legal errors, etc. *Ex ante*, it is then Pareto-efficient to limit the size of the penalty. In order to report their loss correctly, the insurer will have to audit claims at a high frequency. This entails additional costs on the insurance contract. If the auditing cost is high, or if the frequency of audit necessary to give the good incentive for the

policyholder to reveal the truth is too high, consumers would be better off by not insuring the risk. Notice that another way to reduce the willingness to submit a fraudulent claim is to limit the indemnity. The maximal indemnity that is compatible with truth-telling is an increasing function of the penalty and of the probability of audit. Consumers would like to announce *ex ante* that they will not submit fraudulent claims *ex-post*. That would allow insurers to save the audit cost, thereby reducing the equilibrium premium rate, but the announcement is not credible.

Is *ex-post* moral hazard an important problem? It is often suggested that the cost of fraudulent claims may well amount up to 10% of premiums paid for such insurance lines as automobile insurance or homeowner insurance. This estimation is the cost of unjustified indemnities to policyholders, not the auditing cost to fight against fraud. This percentage is comparable to the rate of transaction costs, whose effects on insurability have been previously examined.

The policy recommendation is clear from the discussion above; one should impose a larger penalty on policyholders that have been convicted of a fraudulent claim. Several countries in Africa, and to a smaller extent in Europe, have been weak in this area, recognizing fraud as a 'national sport' that should be forgiven. By doing so, the legal system imposes a possibly large cost to society in terms of a loss of insurability. This weakness has been particularly clear for insurance lines where the indemnity payer does not have the good incentives to be tough on fraud. For example, one may question whether social security organizations are fighting fraudulent claims efficiently. This yields a general distrust of the system, which is detrimental to the unemployed themselves.

Limited Liability

An individual can cause damage to others, either in the course of his/her profession (medicine, surgery, house building, etc.) or because of other activities (e.g., driving a car). The same kind of external random effect occurs for firms. In most countries, the agent found liable for damage to others must indemnify them accordingly. This is done to force decision makers to internalize all costs generated by their choice. But indemnification is only possible up to the decision maker's financial capacity. Limited liability is a way to protect risk-takers against an excessive financial distress. But it has long been recognized that limited liability distorts the decision of the risk-taker in a way that is socially inefficient. The US Saving and Loans crisis is often explained by the fact that 'zombie' S and Ls adopted in the early 1980s, a very risky attitude in an attempt to 'bet for resurrection' after some blows were delivered to their portfolio of (real estate) assets. This is because limited liability gives the agent the equivalent of a free put option. Put it in simpler terms, under limited liability, an insolvent agent can only benefit from taking more risk, because he does not bear the burden of losses. Therefore, if the agent is risk-neutral, he will seek to maximize the expectation of a convex function of his wealth. As a result, he will systematically exhibit a risk-loving behavior, and adopt a very risky attitude. This is a kind of moral hazard problem. Risk aversion mitigates this result, but only for agents who are well capitalized.

The effect of limited liability of the policyholder on their demand for insurance is thus unambiguous; if he/she is risk-neutral, it is never optimal to cover a risk of loss, even in the most favorable case where the premium rate is fair. Insuring the risk would yield a sure reduction in wealth equaling the expected loss. Not insuring the risk would yield an expected reduction of wealth that is less than it, since the agent bears only part of the risk of loss. Another way of looking at this problem is that the insurance contract creates a 'deep pocket' where victims can find compensation for their losses. This kind of problem is particularly crucial when examining the demand for insurance by firms for catastrophic environmental risks. Limited liability on the part of the insurance also reduces the demand of insurance, since it makes the indemnity dependent on solvency.

Limited liability thus raises several important questions. How to organize compensation for those who bear the negative externalities? How to build an incentive-compatible mechanism that increases loss prevention by decision makers with limited liability? How to solve the market failure of liability insurance markets? How to force firms not to undercapitalize their subsidiaries, which are in charge of managing the riskiest part of the business? Two routes have been used. The first one is compulsory insurance. This solves the misallocation of risk in the economy and the organization of a system to compensate the victims. But, most of the time, compulsory insurance has been funded by a flat, nondiscriminatory, nonincentive-compatible insurance tariff. The policyholder's investment in loss prevention is not observed by the fund, either because it is difficult to get information on it, or because the fund did not organize an incentive-compatible system.

The second route has been to organize a 'deep pocket' for decision makers. For example, the hospital that employs an uninsured physician may be made liable in case of the physician's insolvency. Under the US Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), when a bank has been relatively closely involved in the monitoring of a firm's activities, it may be considered by the courts as liable for cleaning up the environmental damages generated by the insolvent firm. The objective of this strategy is to force risk-takers to internalize the full cost of potential losses; the hospital will reduce the income of the careless physician, and banks will increase the loan rate of riskier firms. If there is no asymmetric information between the principal (the hospital, the bank) and the agent (the physician, the firm), the agent will select the socially efficient level of care and insurance. There would then be no insurability problem. But, there is no reason to believe that the principal can monitor the agent at no cost. The CERCLA legislation, for example, introduces more asymmetric information on credit markets. Consequently, there will be more credit rationing, the cost of capital will be larger, and the structure of banking contracts for firms will be affected. Is insurability worth these costs?

Regulation of the Insurance Sector

Insurance markets are among the most heavily regulated ones around the world. In the Middle Ages, the Catholic Church was

reluctant to recognize insurance as an ethically acceptable activity. The French Revolution prohibited insurance companies in 1793. States have been very active to control insurance tariffs, in particular, in life insurance. They also imposed compulsory insurance. More recently, they established solvency rules similar to those of the banking sector.

There is a European tradition to use insurance for a redistributive purpose. Insurance implements wealth redistribution if the expected profit of the insurer depends upon the observable characteristics of the policyholders. The expected profit on the low-risk class of policyholders is used to cover the expected loss on the high risk class. This scheme is standard in unemployment, health, and disability insurance, but it can also be found in other insurance sectors like automobile (for young and old drivers), homeowner (for natural catastrophes), and life (for the prohibition of using health factors to price life insurance). However, the solidarity among different risk classes is difficult to organize with decentralized insurance markets without strong regulation. The absence of discrimination on the basis of policyholders' observable characteristics that are correlated with the risk introduces adverse selection. It also generates moral hazards if these characteristics can be affected by the preventive effort of the policyholder. In order to alleviate the adverse selection problem, regulators often impose compulsory insurance, as in social security systems. Pricing regulation is also necessary to react to the natural tendency of competitive markets to discriminate. Finally, insurers should be forced to supply coverage to all customers at the undiscriminating premium. Various reforms of the health insurance sector in OECD countries have been proposed to deal with the problems inherent in such heavy regulation.

There is a form of saving in an insurance contract from the policyholder's point of view. Indeed, the indemnity, if any, may be paid a long time after the premium. This is particularly true for life insurance. The lender, i.e., the policyholder, should check the solvency of the borrower–insurer. In perfect competitive markets, insurers facing a higher risk of insolvency should be penalized by collecting smaller premia, to compensate the risk borne by the policyholders. This is enough to provide good incentives for insurance companies to manage their financial reserves efficiently. However, policyholders may not be in a position to evaluate correctly the probability of failure of the insurers. This asymmetric information calls for centralized solvency regulation. It takes two forms. The first type of regulation is about solvency ratios, similar to those imposed on banks. The second type of regulation is about the kind of risks that can be taken by insurers on the investments they make with their financial reserves. Typically, there is an upper limit on the share of reserves invested in stocks and real estate. Many countries prohibit the use of derivative assets, in spite of their potential benefits for risk reduction. These constraints must be strong enough to fight against the moral hazard problem generated by asymmetric information and limited liability. It should not be too strong in order to allow for efficient asset liability management by insurers.

See also: Behavioral Economics; Insurance and the Law; Regulation, Economic Theory of; Regulatory Agencies.

Bibliography

Arrow, K.J., 1953. Le Rôle des Valeurs Boursières pour la Répartition la Meilleure des Risques. Edition du CNRS, Paris.

Arrow, K.J., 1965[1971]. Aspects of the Theory of Risk Bearing. Yrjo Jahnsson Lectures, Helsinki. Reprinted from: Essays in the Theory of Risk Bearing. Markham Publishing Co., Chicago.

Borch, K., 1962. Equilibrium in reinsurance market. Econometrica 30, 424–444.

Debreu, G., 1959. Theory of Value. Wiley, New York.

Holmstrom, B., 1979. Moral hazard and observability. Bell Journal of Economics 10, 74–91.

Mookherjee, D., P'ng, I., 1989. Optimal auditing, insurance, and redistribution. Quarterly Journal of Economics 103, 399–415.

Mossin, J., 1968. Aspects of rational insurance purchasing. Journal of Political Economy 76, 533–568.

Rothschild, M., Stiglitz, J., 1976. Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. Quarterly Journal of Economics 80, 629–649.

Samuelson, P.A., 1963. Risk and uncertainty: a fallacy of the law of large numbers. Scientia 98, 108–113.

Shiller, R.J., 1993. Macro Markets: Creating Institutions for Managing Society's Largest Economic Risks. Oxford University Press, Oxford, UK.

Townsend, R., 1979. Optimal contracts and competitive markets with costly state verification. Journal of Economic Theory 21, 1–29.

Townsend, R.M., 1994. Risk and insurance in village India. Econometrica 62, 539–592.

Wilson, R., 1968. The theory of syndicates. Econometrica 36, 113–132.

Insurance and the Law

Shauhin A Talesh, University of California Irvine, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by T. Baker, volume 11, pp. 7587–7591, © 2001, Elsevier Ltd.

Abstract

Insurance and the law are interconnected. Legislation, court decisions, and regulations impact and influence the meaning of private and social insurance arrangements in society. While the law shapes and influences what insurance means in society, insurance also exerts a regulatory force over its subjects and acts as a form of governance beyond the state. Drawing from sociolegal scholars who study the gap between the law on the books and the law in action, this article explains the basic forms and functions of insurance in society and explores insurance's intertwined relationship with the law.

As Justice Black of the United States Supreme Court wrote in 1945: "Perhaps no modern commercial enterprise directly affects so many persons in all walks of life as does the insurance business. Insurance touches the home, the family, and the occupation or the business of almost every person in the United States" (United States v. South-Eastern Underwriters Association, 1945).

The law shapes and influences what insurance means, while the law is simultaneously influenced by insurance. Courts interpret the meaning of insurance policies, statutes establish what social insurance and other public sector insurance arrangements are permitted, and insurance regulations attempt to enforce state insurance laws, promulgate rules and regulations, and conduct hearings to resolve disputed matters pertaining to insurance. Private insurance arrangements in particular depend on a well-defined and robust contract law and a regulated market. While the law influences insurance, insurance also exerts a regulatory force over its subjects and acts as a form of governance beyond the state. Moreover, the presence of liability insurance often shapes how tort lawsuits are structured. Insurance is also an ideological construct that influences the relationship between law and regulation, such that much of society is governed through risk, both within the boundaries of insurance but even beyond insurance. Although ideas about what is insurable and what should be insurable vary over time and location, it is clear that risk and insurance impact so many aspects of society.

Drawing from sociolegal scholars who study the gap between the law on the books and the law in action, this article explains the basic forms and functions of insurance in society and explores the intertwined relationship of insurance and the law.

What Is Insurance?

Although there is not one formal definition of what insurance is (Kimball, 1992: 1; Abraham, 2013), insurance is generally thought of as a formal mechanism for sharing costs for misfortune or injurious experiences. Insurance deals with 'risk', that is, the chance that something can happen (often something negative). In the insurance field, risk is often conceptualized in two ways: (1) a person, property, or enterprise that is insured; and (2) the possibility that something can harm that person, property, or enterprise. Insurance essentially protects the persons or things insured against the possibility of harm.

Contemporary insurance arrangements are designed around a formal and organized scheme for the distribution of an economic loss over a large number of persons subject to the risk of a particular loss, with a goal of replacing the uncertain risk of loss with a predictable cost. The loss is often distributed by transferring the risk to an insurer (risk transfer). The loss is distributed in advance, often by charging a fixed premium or by charging an assessment after the event, or by some combination. For the insuring agreement to function well, the incidence of the loss should generally be reasonably predictable for some class of persons from which the participants will be drawn (risk pool), and therefore for any sufficiently large number of such participants taken at random (Abraham, 1986).

Because of the variation in the size and complexity of insurance, insurance can be characterized in multiple ways: (1) by line, which divides insurance into property and casualty insurance (fire, ocean, title, errors and omissions, and various forms of liability insurance) and personal insurance (life, accidental death or dismemberment, disability, and health insurance); (2) by interests protected, which distinguishes first-party insurance (the policyholder insures his or her own interest in a person's life or in property) and third-party insurance (liability insurance, which pays proceeds to a third party to whom an insured becomes liable); (3) by method of marketing, which distinguishes between group versus individual policies, and among insurance entities based on the methods used to sell policies (agents, brokers, etc.); and (4) by insurer organization, which acknowledges the various structures through which insurance entities operate (e.g., stock companies, mutual companies, reciprocal exchanges, Lloyd's associations, and various hospital and medical organizations) (Jerry and Richmond, 2012).

In general, legal scholars tend to divide the insurance law field into two subareas. One focuses on the law concerning the relationship between private insurance organizations and their insureds. This is most often developed from judicially articulated doctrines that regulate the relationship between an insurer and policyholder and is considered a part of contract law (although tort law and agency law principles and

some statutes and administrative regulations are sometimes important to this framework). The other realm of insurance law is largely a series of statutes enacted by state legislatures and administrative regulations developed by agencies that exist in most states. This is considered a part of the law of regulated industries (e.g., Baker, 2010; Jerry and Richmond, 2012; Cousy, 1999).

Insurance as Contract

Historically, insurance law has been considered derivative of contract law (Baker, 2008). Contemporary insurance institutions were derived from two distinct roots: (1) insurance as a mutual benefit protecting a fraternal group or (2) insurance as a commercial enterprise promoting trade and investment. Under either scenario, "insurance was a voluntary undertaking whose obligations were determined almost exclusively according to the rules governing the field of law that came to be known as the law of contracts" (Baker, 2008: 29).

By the mid-nineteenth century, courts in Europe and the United States recognized that insurance contracts were not traditional contracts, that is, voluntary agreements with terms that are negotiated between two parties with equal bargaining power. Rather, insurance policies are largely contracts of adhesion. Rarely do insurance companies and policyholders have equal bargaining power. Insurance companies typically use standard-form contacts with terms that are not subject to negotiation. Standard-form insurance policy provisions are often vague and drafted at a level of generality that leads to differing expectations about the scope of coverage within the policy. Moreover, most individuals do not have much knowledge about precisely what they are purchasing, especially because they often purchase insurance through intermediaries such as agents (who often offer minimal explanation of what the insurance policy provisions mean). On the other hand, insurance companies are often repeat players, wealthy, and the more sophisticated economic entity when dealing with a prospective insured. There is also not much variation in the market; with some exceptions (Schwarcz, 2011), insurance companies offer similar contract terms to consumers with similar profiles and needs. Moreover, insurance companies not only have more information, but better information concerning the meaning and value of the insurance policy than the person seeking to purchase insurance. Due to its gatekeeping function (see 'Insurance as Gatekeeper' later in the article), insurance is often not voluntary but a prerequisite to obtaining other privileges in society.

Although government regulation of insurance began in the United States during the latter half of the nineteenth century, the relationship between insurers and policyholders has remained largely within the domain of contract law. Courts and state regulations attempt to provide some protection to insureds, albeit with varying degrees of success. In particular, state regulations in the United States require premarket approval of some of the terms of insurance policies. Courts use the *contra proferentem* principle to try to protect insureds. Courts invoking this principle hold that ambiguities in a contract should be construed against the drafter. This interpretive principle is often used to supersede rather than interpret insurance policy language. Although regulatory agencies and courts can provide a formal check on the asymmetric relationship and require companies to reform the content of the standard form insurance policy, insurance companies often retain the contractual ability to modify insurance policy provisions going forward (to make policy provisions less ambiguous). By exercising this ability, companies can avoid judicial and regulatory scrutiny. Thus, contract law has and will continue to define the major aspects of the relationship between insurers and policyholders.

Insurance as Regulation

Insurance is also a form of regulation that affects many aspects of our lives. The state engages in the regulation of its citizenry by setting eligibility requirements and benefit levels for social insurance. Insurance companies establish the underwriting criteria, standards, and charge premiums. These mechanisms allow insurance companies to control who can or cannot obtain insurance. Through insurance policy terms and pricing, insurance companies also establish norms of conduct. Private insurance policies for life, health, and property often take the form of private legislation or regulation through exclusions and conditions. Thus, insurance is able to regulate even acts within the home or business, where traditionally individuals have been free from interference (O'Malley, 1991; Baker, 2010; Ben-Shahar and Logue, 2012). The following sections highlight two specific ways that insurance regulates law and society: (1) insurance as gatekeeper and (2) insurance as a specific form of tort regulation.

Insurance as Gatekeeper

Insurance serves a gatekeeping function in society because it is a prerequisite to other activity. Generally a person cannot register an automobile without automobile insurance, obtain a commercial business loan without business-owners' insurance, bid on a government contract without a surety bond, obtain practice privileges at hospitals without medical malpractice insurance, and sign a commercial lease without commercial property and liability insurance (Baker, 2010).

Home ownership in the United States quite clearly illustrates the gatekeeping function insurance plays. Home ownership is traditionally considered one of the fundamental features of the American economy. Thus, the availability of mortgages at a reasonable and affordable rate is often regarded as an important economic indicator. However, homeowners' insurance plays an important role in the availability of mortgages, because one cannot take out a mortgage without homeowners' insurance. In particular, obtaining homeowners' insurance is mandatory for standard mortgages and homeowners' insurers often use their own underwriting concerns that are independent from the screening procedures implemented by the lender (Squires, 1997).

Liability Insurance as Financier for the Tort System

Liability insurance acts as a form of tort regulation and in doing so, finances the civil litigation system. Tort law in action is often

shaped by the absence or presence of liability insurance. Over the past century, tort law has continually sought available sources of recovery and often creates and expands the liability of individuals and businesses that are likely to be covered by or have access to liability insurance (Abraham, 2008). As tort law expands, so does liability insurance. For example, medical liability insurance grew with medical liability and became so intertwined with it that occasional medical liability insurance crises are widely interpreted as medical liability crises (Baker, 2005a). Product liability doctrine has been facilitated by product liability insurance. Insurance companies responded by creating new forms of liability insurance to address the new liabilities when existing insurance was not available. Employment liability insurance rose as employment liability grew in the late nineteenth century. Moreover, state workers' compensation statutes enacted in the early twentieth century addressed the rising number of work-related accidents and the fact that the litigation system was unpredictable, expensive, filled with delays, and often led to small recoveries for people with work-related injuries. Workers' compensation laws allowed individuals suffering work-related injuries to receive compensation relatively quickly and without needing to obtain a lawyer and establish tort liability (Friedman and Ladinsky, 1967; Talesh, 2012). Even the US government took steps after the terrorist attack of 11 September 2001 to design a first-party liability insurance scheme, to compensate victims' families who otherwise would have had to use the tort system to seek relief. The compensation fund established by British Petroleum in response to the oil spill that occurred in the Gulf of Mexico in 2010 provides a more recent example of how insurance-based principles are used to preempt tort liability.

Because of consumer debt and the ability of bankruptcy courts to discharge civil liabilities, liability insurance is the primary asset on which plaintiffs can count when seeking to collect tort judgments. Without liability insurance, many underserved individuals who were innocently injured as the result of a tort by another person would not be compensated through civil litigation. Liability insurance in action means that when an injured victim brings a tort claim against a tortfeasor, it is the liability insurer that often defends the claim and, if necessary, pays the claim (Gross and Syverud, 1996). The presence of liability insurance consequently shapes how plaintiffs and defense lawyers litigate cases (Yeazell, 2001). A plaintiff lawyer's decision to represent an injured victim in a tort case is predicated not merely on proving the elements of a tort but on the defendant's ability to pay (Baker, 2005b). Liability insurance determines who can be sued, for how much, and for what wrongs. For individuals bringing tort-based lawsuits, "liability insurance is a de facto element of tort law, [and a] de facto cap on tort damages" (Baker, 2005b: 13). Even large corporations with significant assets often have liability insurance, but this relationship does not necessarily make litigation more efficient; large corporations involved in tort lawsuits often spend a significant amount of time convincing their own insurance companies to pay (Baker and Griffith, 2010).

In one of the early but influential studies of the gap between the law on the books and the law in action, Ross's (1970) study of automobile insurance adjustors in the late 1960s revealed how liability insurance led insurance claims administrators to focus on interpreting and implementing tort law by primarily managing aggregate costs rather than determining the individual fault of defendants:

> Adjustment of insurance claims compromises the legal mandate for individualized treatment with the need of a bureaucratic system for efficient processing of cases. This compromise can be observed at many points in the processes of investigation and evaluation. Investigation is vastly simplified, for instance, by presumptions as to liability based on the physical facts of the accident. Accidents are thus seldom individualized to an insurance adjuster or a claims attorney. Rather, they are rear-enders, red-light cases, stop sign cases, and the like and the placement of an accident into one of these categories ordinarily satisfies the requirements for investigation of liability….
>
> A large-scale society proceeds by routinizing and simplifying inherently complex and difficult procedures. This is how the work of the world is done. This is the law, as it is experienced by its clients rather than by its philosophers (Ross, 1970: 135).

Thus, under Ross's study, liability insurance is a bureaucratic claim processing mechanism that renders large amounts of tort law into a simple, manageable set of compensation rules and procedures. Although Ross focused on automobile claims, others have demonstrated how tort claiming is highly connected to liability insurance in a number of other tort subfields (Abraham, 2008; Abraham and Liebman, 1993; Baker, 2005a; Vidmar et al., 2005; Baker and Griffith, 2010).

Other forms of insurance also regulate tort law in action. When a statutory insurance scheme is in place, tort liability is often limited. Such statutory schemes have multiple effects: (1) they sometimes create a bar on certain types of tort claims; (2) they alter the manner in which tort claims are resolved by causing plaintiffs' lawyers to frame their claim in a manner that avoids being precluded from recovery; and (3) they lower the potential level of total recovery for a tort by making it more difficult to sue in tort when health, disability, or property insurance are covering an injury. Conversely, the existence of first-party insurance can sometimes (depending on the jurisdiction's collateral source rules) lead to a potentially higher tort recovery because collecting from the first-party insurance carrier shortly after the injury occurs can decrease the need or desire to settle early in the case, which allows litigants to pursue a more aggressive litigation strategy.

In some respects, legal insurance in Europe (and to a lesser extent in the United States) is an attempt to finance the litigation process even further by prospectively offering coverage for the risk and cost of a lawsuit. Legal insurance, legal cost insurance, or simply legal protection insurance is a form of voluntary private insurance that covers the costs of legal proceedings. With some variation, legal insurance in Europe covers all costs related to bringing or defending a claim, though not the claim itself. Although legal insurance covers claims arising out of contracts, social insurance, some family matters, and defense of criminal and misdemeanor cases (among others), "insurance for motoring and traffic accidents has been the main factor behind the expansion [of legal insurance]" (Raiser, 2001: 8639). Although less present in the United States, legal insurance is often reflected in two ways: (1) prepaid legal services offered by trade unions and professional groups to their members and (2) group legal

services plans, which cover workers in a particular branch and are financed primarily by the employers. These legal service plans allow workers and families greater access to legal counseling and, if necessary, legal support in court. Although neither the European nor the American markets are dominated by legal insurance, the existence of these insurance schemes reflects an awareness that litigation is often financed by insurance and provides a market for people to shift legal risk away from themselves.

Insurance as Governance

A series of empirical studies in the past two decades has also helped us understand how insurance institutions, forms, technologies, and visions (Ewald, 1991) act as a form of governance beyond the state (Ericson et al., 2003; Ericson and Doyle, 2004). Beginning with Heimer's (1985) inquiry into how insurers manage moral hazard in property and fidelity insurance relationships and continuing with Ericson et al.'s (2003) comprehensive research on the Canadian insurance industry, scholars have begun to understand how insurance is an institutional force with effects on individuals, organizations, and institutions inside and outside the insurance industry.

Ericson and his colleagues explain that insurance governs society through 'nine-interconnected dimensions' (Ericson et al., 2003). To begin, insurance simultaneously produces knowledge of risk, creates a scheme from which risks can be made objectively calculable, and develops a risk pool. In particular, these three dimensions of insurance classify everything into degrees of chance of harm and make everything calculable and consequently subject to commodification. Through actuarial techniques, concrete facts about objective risks are converted into probabilities and ultimately assigned a cost so that prices can be established. Actuarialism creates a risk pool, that is, a population that has a stake in the identified risks and the specific harms they entail. The risk pool transforms the population into a collective that seeks to minimize loss while also compensating those who suffer a loss. Turning to the fourth dimension, insurance does not protect against the particular event that causes harm to a member of the risk pool, but instead protects against a loss of capital that might result from that event. Their fifth and sixth dimensions identify the hybrid nature of insurance as both managerial and legal. Insurance manages risks in the population by using oversight, surveillance, and auditing while also making its risk pool subject to contract and ultimately adjudication. But insurance also assists the law in assigning liability to the party most capable of distributing the loss through insurance. Addressing the final three dimensions, these scholars explain how insurance's cultural, technological, and political functions impact and influence society in several important ways: "in providing a futures market in security, insurance offers a cultural framework for conceptions of time, destiny, providence, responsibility, economic utility, and justice." They also observe that "insurance is a social technology of justice. It bridges individual and social responsibility through distributive justice (collective sharing of loss) and restorative justice (financial indemnification)." For that reason, they contend, "insurance is … political, combining aspects of collective well-being and individual liberty in a state of perpetual tension" (Ericson et al., 2003: 6).

Scholars adopting an insurance-as-governance framework emphasize the way in which private insurers and governments exercise similar power in society and behave in a similar manner. In particular, the private insurance industry has many of the same goals as the state, including seeking forms of social security and solidarity by pooling risks. The private insurance industry provides technologies and social arrangements for allocating risk across pools of risk takers and provides for a sharing of the risks of misfortune through financial compensation of loss. It also attempts to establish preventative security arrangements that try to minimize harm and loss to its citizens. Thus, the private insurance industry governs through its powers to transfer and distribute risk, and it engages and involves the state only when necessary. The private insurance industry also uses many of the same methods as the state to achieve its goals such as surveillance, underwriting (sophisticated information systems for selecting risks), and claims (compensating loss). It also uses a substantial amount of private policing in the form of technologies, investigators, and inspectors to address fraud and achieve loss reduction and preventative security. Similar to the state, the private insurance industry is subject to many social, economic, and political forces (such as changes to the environment, economic globalization, terrorism, property and violent crime, and advances in medicine and health). The private insurance industry partners with the state to regulate insurance practices.

Consistent with the work of Ericson and his colleagues, there has been an effort in the past 20 years toward both theorizing and empirically exploring how the private insurance industry plays an active role in constructing the meaning of risk and responsibility in different segments of society. Notable governance-as-insurance studies include studying the role of property insurers in governing security in the home (O'Malley, 1991), highlighting the governance role of insurance companies in the motion picture industry in the United States (Hubbart, 1996–97), analyzing the rise of risk management approaches toward campus drinking (Simon, 1994), and examining the tort-settlement factoring industry in the United States (Scales, 2002). More recently, scholars are beginning to think about governance in the health insurance context. Hunter (2008) argues for using 'risk-centered governance' as the model for health law and policy. Hunter highlights how the debate in the United States over universal access to health insurance often links itself to the debate over the rise of actuarial medicine. She argues that the political attractiveness of 'risk' is subject to multiple interpretations and appropriations, including some that could mobilize pressure to use conceptions of risk in a more equitable manner (Hunter, 2008; see also Stone, 2002). Hoffman (2010) examined the Massachusetts mandatory health insurance experiment and identified how the fragmentation of health insurance risk pools that results from actuarial risk rating limits the ability of private insurance markets to carry out the social goal of redistributing health across society.

In sum, the private insurance industry acts as a system of governance through various forms of collaboration with state

governments: "sharing similar goals of security and solidarity through the pooling of risks; using similar techniques for governing at a distance; and collaborating in insurance regimes" (Ericson et al., 2003: 65). Insurance institutions are not just pooling and transferring risk, but instead actively shaping the world within which they operate. As these examples highlight, under a governance conception, insurance does not fit well within a public–private dichotomy. The relationship between insurers and policyholders falls somewhere in between the relationship between a public government and its citizens and a private relationship between two parties to a contract. However, private insurers are not agents of government; they are not subject to the rules, policies, and procedures that limit government discretion. Moreover, they are not required to afford policyholders constitutional protections (such as due process or equal protection) (Abraham, 2010). As the private insurance industry's governance role continues to grow, scholars going forward may want to consider standards for regulating the behavior of private insurers beyond those contained in administrative regulations, civil law doctrines, and contract law doctrines.

Insurance as a Tool for Increasing or Decreasing Social Stratification

Insurance in action often serves a social stratification function. Insurance law scholar Tom Baker argues that the "social stratification function of insurance is a more generalized way of thinking about the dynamic that makes insurance companies gatekeepers" (Baker, 2008: 10). As the prior sections highlight, the relationship between insurance and law in society is often mediated by insurance institutions, namely, insurance companies. Those individuals who cannot obtain insurance from insurance companies occupy a different social position than those who are able to obtain insurance. Moreover, people who have to pay more for insurance have fewer resources to spend on other things. To be clear, insurance companies are not the sole cause of such inequality. However, insurance companies do play an important role and in doing so, simultaneously construct and reflect the broader social conditions that lead to social stratification. In the context of first-party insurance, many life, health, and disability insurance companies have traditionally refused to provide coverage to people who have been treated for mental illness (Baker, 2002). Some insurance companies that are guided by an overzealous desire to insure 'good' risks sometimes 'cream-skim' and insure only 'good' risks and therefore decrease the likelihood of paying out on claims. This has the undesirable effect (from the standpoint of society) of leaving those most in need of insurance without being able to obtain coverage and increases the gap between 'haves' and 'have-nots.'

Conversely, insurance laws and regulatory schemes sometimes allow insurance to ameliorate social stratification and consequently serve the public interest (Talesh, 2012). In addition to liability insurance financing the civil litigation system, state laws require people to purchase automobile insurance and allow employers to opt into state-created workers' compensation insurance schemes. In doing so, insurance laws provide injured persons with insurance mechanisms through which to seek relief. Health insurance laws such as Medicaid and Medicare in the United States assist the poor and the elderly (and other groups) by increasing access to care to those who may not otherwise be able to purchase health insurance. Universal or socialized health care programs in Canada and some European countries (and more recently, the Affordable Care Act in the United States) also attempt to reduce social stratification. Though not perfect, insurance regulation, in addition to court decisions, attempts to ensure that insurance companies follow through on their promises, remain solvent, and avoid unfair business practices. Without insurance laws and regulations, policyholders could not possibly trust insurers to pay claims or remain solvent. As a system of risk spreading, transfer, and distribution, insurance can sometimes promote the public good and often allows individuals to seek compensation without needing to use a lawyer. In this way, insurance law acts in the public interest because it provides a collection mechanism for injured persons seeking compensation and relief for injuries (Talesh, 2012).

Governing through Risk beyond Insurance

Conceptions of insurance impact many social and legal relationships in society beyond the traditional insurance context. Insurance technologies and visions are used to govern risk outside of insurance institutions and risks that cannot or should not be governed by insurance. Baker and Simon refer to this concept as 'governing through risk,' the idea of using formal considerations about risk to direct organizational strategy and resources outside of (beyond) the traditional insurance context (Baker and Simon, 2002). There are numerous examples of governing through risk that draw on insurance technologies and visions but occur outside the traditional insurance context. To name a few, community policing models target high-risk areas (Ericson and Haggerty, 1997), social service agencies focus on at-risk children, environmental engineers conduct risk assessments of hazardous waste sites and other sources of environmental concern (Graham and Weiner, 1995), judges and policymakers debate tort and accident law in terms of allocation and spreading of risk (Calabresi, 1970), and financial analysts develop and structure portfolios based on a risk-reward ratio (Bernstein, 1992, 1996). Actuarial techniques are incorporated and mobilized by a range of institutions, from police departments to social service agencies to financial institutions. All these institutions rely on evaluating events based on probabilities and statistics and conceptions of risk that insurance has helped make possible (Ewald, 1986).

Risk-based principles drawn from insurance are increasingly a template through which to govern other relationships in society. There is at least one important difference, however, between insurance institutions using risk-based principles and noninsurance institutions using risk-based principles. Insurance institutions gather, pool, transfer, and spread risk and thereby eliminate or at least reduce the importance of those day-to-day concerns of people who are exposed to them. However, risks that are not subject to insurance

coverage, such as the examples I provided earlier, are not spread and eliminated in this manner, and therefore have significant consequences for the people exposed. A risk that occurs beyond insurance, therefore, is experienced in a much different manner than a risk that is insured. More recently, public and private insurance institutions are placing a greater emphasis on an individual's responsibility to embrace risk more directly. Scholars continue to engage in philosophical, theoretical, and normative debates concerning what distinguishes an insurable versus and uninsurable risk: "Because of the crucial role of insurance institutions in socializing risk and responsibility, studying risks beyond insurance opens a window on the limits of social responsibility and the role of ideas about individual responsibility in the shaping of insurance institutions and forms" (Baker and Simon, 2002: 12).

As this article highlights, insurance serves multiple functions in society. The forms and functions of insurance and conceptions of risk more broadly can have potentially positive and negative effects in society. Setting aside normative arguments concerning whether insurers' rising role in society is a good or bad thing, there is little doubt that insurance will continue to be intertwined and connected with the law and consequently impact society in multifaceted ways.

See also: Contracts: Legal Perspectives; Health Insurance: Economic and Risk Aspects; Insurance; Law and Risk; Regulation, Economic Theory of; Risk, Sociology and Politics of; Social Insurance: Legal Aspects; Torts in Law.

Bibliography

Abraham, K., 1986. Distributing Risk. Yale University Press, New Haven, CT.
Abraham, K., 2008. The Liability Century: Insurance and Tort Law from the Progressive Era to 9/11. Harvard University Press, Cambridge, MA.
Abraham, K., 2010. Insurance Law & Regulation: Cases and Materials, fifth ed. Foundation, New York.
Abraham, K., 2013. Four Conceptions of Insurance. Pennsylvania Law Review 161, 653–698.
Abraham, K., Liebman, L., 1993. Private insurance, social insurance, and tort reform: toward a new vision of compensation for illness and injury. Columbia Law Review 93, 75–118.
Baker, T., 2002. Risk, insurance and the social construction of responsibility. In: Baker, T., Simon, J. (Eds.), Embracing Risk: The Changing Culture of Insurance and Responsibility.
Baker, T., 2005a. The Medical Malpractice Myth. University of Chicago Press, Chicago, IL.
Baker, T., 2005b. Liability insurance as tort regulation: six ways that liability insurance shapes tort law in action. Connecticut Insurance Law Journal 12, 1–16.
Baker, T., 2008. Insurance Law & Policy, second ed. Aspen, New York.
Baker, T., 2010. Insurance in sociolegal research. Annual Review of Law and Social Science 6, 9.1–9.15.
Baker, T., Griffith, S.J., 2010. Ensuring Corporate Misconduct: How Liability Insurance Transforms Shareholder Litigation. Univ. Chicago Press, Chicago, IL.
Baker, T., Simon, J. (Eds.), 2002. Embracing Risk: The Changing Culture of Insurance and Responsibility. University of Chicago Press, Chicago, IL.
Ben-Shahar, O., Logue, K., 2012. Outsourcing regulation: how insurance reduces moral hazard. Michigan Law Review 111, 197–248.
Bernstein, P., 1992. Capital Ideas. Free Press, New York.
Bernstein, P., 1996. Against the Gods. John Wiley, New York.
Calabresi, G., 1970. The Cost of Accidents. Yale Univ. Press, New Haven.
Cousy, H. (Ed.), 1999. International Encyclopedia of Laws. Insurance. Kluwer Law International, Boston, MA.
Ericson, R.V., Doyle, A., 2004. Uncertain Business: Risk, Insurance and the Limits of Knowledge. Univ. Toronto Press, Toronto.
Ericson, R., Doyle, A., Barry, D., 2003. Insurance as Governance. Univ. Toronto Press, Toronto.
Ericson, R., Haggerty, K., 1997. Policing the Risk Society. Oxford Univ. Press, New York.
Ewald, F., 1986. L'tat Providence. Grasset, Paris.
Ewald, F., 1991. Insurance and risk. In: Burchell, et al. (Eds.), The Foucault Effect. Univ. of Chicago Press, Chicago, IL.
Friedman, L., Ladinsky, J., 1967. Social change and the law of industrial accidents. Columbia Law Review 67, 50–82.
Graham, J., Weiner, J. (Eds.), 1995. Risk v. Risk. Harvard Univ. Press, Cambridge.
Gross, S.R., Syverud, K.D., 1996. Don't try: civil jury verdicts in a system geared to settlement. UCLA Law Review 44, 1–64.
Heimer, C., 1985. Reactive Risk and Rational Action: Managing Moral Hazard in Insurance Contracts. Univ. Calif. Press, Berkeley, CA.
Hoffman, A., 2010. Oil and water: mixing individual mandates, fragmented markets, and health reform. American Journal of Law and Medicine 36, 7–77.
Hubbart, E.O., 1996–97. When worlds collide: the intersection of insurance and motion pictures. Connecticut Insurance Law Journal 3, 267–304.
Hunter, N.D., 2008. Risk governance and deliberative democracy in health care. Georgetown Law Journal 97, 1–60.
Jerry, R.H., Richmond, D., 2012. Understanding Insurance Law, fifth ed. LexisNexis, Bender, New York.
Kimball, S., 1992. Cases and Materials on Insurance Law. Aspen, Boston, MA.
O'Malley, P., 1991. Legal networks and domestic security. Studies in Law, Politics and Society 11, 171–190.
Raiser, T., 2001. Legal insurance. In: Smelser, N., Baltes, P. (Eds.), International Encyclopedia of the Social and Behavioral Sciences, p. 8639.
Ross, H.L., 1970. Settled Out of Court. Aldine, Chicago, IL.
Scales, A.F., 2002. Against settlement factoring? the market in tort claims has arrived. Wisconsin Law Review, 859–964.
Schwarcz, D., 2011. Reevaluating standardized insurance policies. The University of Chicago Law Review 78, 1264–1348.
Simon, J., 1994. In the place of the parent: risk management and the government of campus life. Social and Legal Studies 3, 15–45.
Squires, G.D. (Ed.), 1997. Insurance Redlining: Disinvestment, Reinvestment, and the Evolving Role of Financial Institutions. Urban Institute Press, Washington, DC.
Stone, D., 2002. Beyond moral hazard: insurance as moral opportunity. In: Baker, T., Simon, J. (Eds.), Embracing Risk: The Changing Culture of Insurance and Responsibility.
Talesh, Shauhin, 2012. Insurance law as public interest law. U.C. Irvine Law Review 2, 985–1009.
United States v. South-Eastern Underwriters Association, 1945. 322 U.S. 533.
Vidmar, N., Lee, P., MacKillip, D., McCarthy, K., McGwin, G., 2005. Uncovering the 'invisible' profile of medical malpractice litigation: insights from Florida. DePaul Law Review 54, 315–356.
Yeazell, S., 2001. Re-financing civil litigation. DePaul Law Review 51, 183–217.

Insurgency

Erica Chenoweth and Jonathan Pinckney, Josef Korbel School of International Studies, University of Denver, Denver, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Insurgency is the use of irregular tactics by a weaker foe seeking to impose military costs upon a stronger armed incumbent government or occupying power. Insurgencies have been ubiquitous in the twentieth century, with well over 200 new insurgencies emerging. Although insurgency has been a fairly popular method of conflict, it has not been particularly effective. The theory and practice of insurgency and counterinsurgency leaves many questions about why insurgency occurs, why it succeeds and fails, and how to respond to it. While key concepts and principles of insurgency and counterinsurgency brought to prominence in the mid-twentieth century remain largely visible in the contemporary world, scholarship must continue to evolve as new tactics and technologies revolutionize the nature of insurgency and political violence.

Insurgency Defined

Classical Definitions

Insurgency is as old as war. Ancient texts from the Bible to the histories of Tacitus record the use of irregular tactics by a weaker foe seeking to impose military costs upon a stronger armed incumbent government or occupying power. Indeed, as many have pointed out (Beckett, 2001; Nagl, 2002; O'Neill, 2005), incidences of insurgency and counterinsurgency (COIN) are historically more common than 'conventional' war. Yet traditionally such proto-insurgents were viewed as bandits or brigands, and the politicization of their activities was largely absent. Insurgency in its modern iteration has been defined by a particular series of events and thinkers, as well as by the rise of modern weapons and communications technology which have empowered small groups of insurgents to a level undreamed of in premodern times (Nagl, 2002).

Western thought on insurgency arose in response to the American War of Independence, in which American irregular units such as that commanded by Francis 'Swamp Fox' Marion played a significant part in keeping the largely loyalist southern states from returning to British control. The first European theoretical works on irregular war were written by veterans returning from service in the British Army during the American Revolution (Beckett, 2001: p 4). Revolutionary France's use shortly thereafter of the *levee en masse*, literally calling the entire French nation to arms eroded further the concept of the autonomy of war, and the Spanish nonconventional resistance to French occupying forces gave birth to the term *guerrilla* (literally 'small war'). Clausewitz soon after formulated his famous trinity of 'the people, the army, the state,' emphasizing the political nature of armed conflict and the insufficiency of merely defeating an enemy's armed forces in order to achieve victory.

All of these events shaped Western thought on insurgency. However, the concept's true birth comes in the thought and practice of Mao Tse-tung, whose 'revolutionary war' has served as modern insurgency's blueprint. Mao's primary insight is that in revolutionary war the primary battleground is the people themselves and the primary weapons are political. The revolutionary forces must "be like fish swimming in the water of the population" (quoted in Nagl, 2002: p. 28) – they must be both supported by and indistinguishable from the population at large. To that end Mao enjoined his guerrillas, even as they sought to intimidate the local population, to treat them with dignity and respect and to support local popular political issues. Temporary military defeat would be made irrelevant and could even benefit Mao's cause as long as the Communists maintained the support of the people.

Contemporary Uses

This essential focus on political legitimacy and the people defines contemporary insurgency. Modern insurgency's central characteristic is to undermine the people's belief in and consent to the incumbent government's rule. To this end the insurgency first utilizes political tools such as the distribution of propaganda to undermine the ideological basis for the government's rule and gain the people's support for their activities. Galula writes specifically of the importance of a cause, "The insurgent cannot seriously embark on an insurgency unless he has a well-grounded cause with which to attract supporters among the population" (Galula, 1964: p. 8). The effective dissemination of the insurgent's cause denies the incumbent regime ideological legitimacy.

In tandem with overt political activity, the insurgent engages in acts of violence, typically including terrorism and guerrilla warfare, to sap the government's will to continue the fight and to destroy the government's practical ability to govern and effectively deliver government services. Primarily operating from rural areas, areas of difficult terrain, or the 'urban jungle,' the insurgent slowly wears down its opponent and denies the government's claim to hold a monopoly of force. Insurgencies seek to gain control of territory but do not maintain a static control over a set territory, as such action would cause them to lose their 'fluidity' and put them in a disadvantageous position of conventional war in which the full force of their stronger opponent may be brought to bear. More important is control over the people. In an ideal insurgency, the continued use of violence against government forces and against those who do not cooperate with the insurgents further demonstrates to the people the impracticality of following government orders and eventually tips the balance of power in favor of the insurgency.

While this typical portrait of modern insurgency assumes that the insurgency's goal is regime change, contemporary scholarship has recognized a number of subcategories of

insurgency based on differing goals. O'Neill distinguishes between insurgencies seeking change in the political community (the state itself), the political system (the particular structures of the state), the authorities, and particular state policies (O'Neill, 2005). These varying types of insurgency are linked by their common application of insurgency's more basic indirect, population-focused, political-military strategy.

Insurgency Distinguished from Rebellion, Guerrilla Warfare, and Civil War

Insurgency has been distinguished in various ways from the more general civil conflict categories of rebellion (or revolution), guerilla warfare, and civil war. Galula contrasts insurgency from rebellion by its protracted nature. While revolution is an 'explosive upheaval,' an insurgency is "a protracted struggle conducted methodically, step by step, in order to attain specific intermediate objectives leading finally to the overthrow of the existing order" (Galula, 1964: p. 2). Thus while revolutions may involve long periods of planning, the event itself is characterized by sudden, unexpected social upheaval with widespread popular support. Insurgencies engage in a longer, protracted struggle in order to reach the ultimate goal of controlling the people.

Guerrilla warfare is typically an integral part of that longer, protracted struggle. However, guerrilla warfare in and of itself is merely a method which may be pursued by insurgents or state actors as a part of more traditional warfare. Many insurgencies rely not only on guerrilla tactics but also on clandestine, small-scale, armed operations, or terrorism, and indeed may be too weak to engage in guerrilla warfare at all until late in their development (Byman, 2008). And the contemporary, largely urban insurgencies in Iraq and Afghanistan often rely on a set of battlefield tactics foreign to traditional guerrilla warfare (Kilcullen, 2006).

Aspects of tactics and territory also distinguish insurgency from civil war. One of insurgency's defining characteristics is the insurgent's unwillingness to engage in a direct military conflict with his opponent. Thus while insurgencies seek to control territory, their control remains fluid. Civil war, however, is characterized by two subnational parties, each in control of a specific portion of territory, engaging in conventional conflict. As Galula says: "[Civil War] soon resembles an ordinary international war except that the opponents are fellow citizens" (Galula, 1964: p. 3). As insurgencies gain strength and firmer control over territory, their conflict may transform into a civil war – but this is by no means necessarily so.

The Onset of Insurgency

Researchers have spent decades identifying the causes of rebellion (Gurr, 1970; Goldstone, 1991; Fearon and Laitin, 2003). In macro explanations of substate violence onset, few scholars distinguish between insurgency and civil war. From 1900 to 2006, insurgency was a relatively common phenomenon. During that time period alone, 217 major insurgencies broke out across the globe. (This refers to insurgencies that involved at least 1000 battle deaths and sought to overthrow incumbent governments, achieve territorial secession, or drive out foreign occupation. All data referenced in this article come from the Nonviolent and Violent Campaigns and Outcomes Data Set, version 1.1 (Chenoweth, 2011).) In an average year, two new insurgencies set on worldwide.

As Figure 1 illustrates, the onset of new insurgencies steadily rose until 1980. Since that time, the patterns of onset have been a bit more erratic. Although there were fewer new insurgencies in the 1980s and early 2000s, a large wave of new insurgencies broke out in the 1990s owing largely to the breakup of the Soviet Union.

The vast literature tends to focus on two sets of causal processes – grievance and opportunity – to explain the onset of violent conflict. A scholarly consensus exists that both grievances and opportunities are necessary for insurgent mobilization, but neither alone is sufficient. For example, Jack Goldstone identifies three factors that lead to the onset of revolutions: financial instability, divisions among elites, and high mobilization potential (Goldstone, 1991). He argues

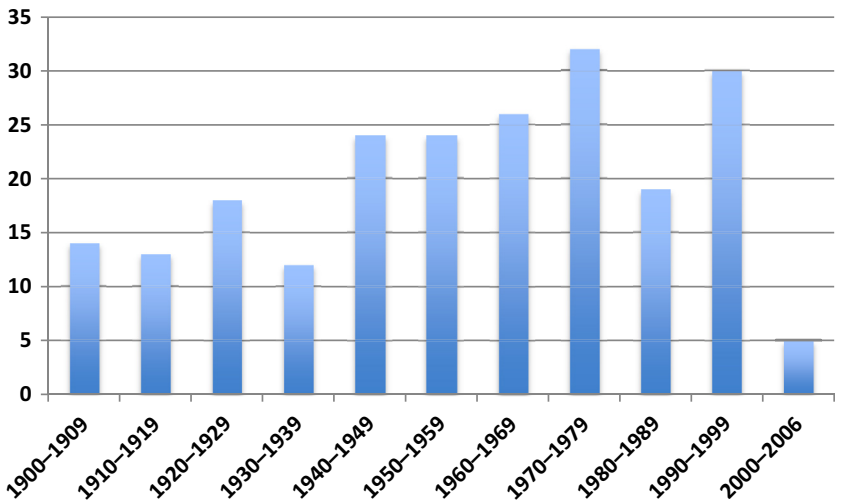

Figure 1 Number of insurgency onsets, 1900–2006.

that when these factors coexist in a society, they provide both the grievances (i.e., financial instability) and opportunities (elite divisions and high mobilization potential) that motivate people to seek the overthrow of their own governments. Figure 2 illustrates the typical grievance-opportunity approach.

Grievances

When scholars isolate the grievances that lead to violent conflict, they typically settle on perceived injustices, such as the unequal distribution of income, chronic poverty, or identity-based deprivations (Gurr, 1970; Collier and Rohner, 2008; Kaufman, 2006). Gurr (1970) finds that real economic grievances only arise when there are significant 'horizontal inequalities' (an uneven distribution of wealth), a finding corroborated by some (Cedermann et al., 2011; Do and Iyer, 2010; Nepal et al., 2011; Vadlamannati, 2011). When significant volumes of people are deprived of economic or political gains relative to others within the same society, a sense of shared grievance can develop against the wealthy or ruling class. Political grievances, such as the lack of political representation or channels for nonviolent expression, are also thought to be potential root causes of insurgency (Bohara et al., 2006; Carey, 2007; Gleditsch and Ruggeri, 2010; Goodwin, 2001; Joshi and Mason, 2008; Walter, 2004). And, in some cases, severe ethnic tensions may be associated with the onset of violent insurgency. Such tensions may arise when certain ethnic groups are excluded from economic and social benefits available to others within a given society (Cedermann et al., 2011; Wimmer et al., 2009). As many scholars point out, however, grievance alone is an insufficient explanation for the outbreak of conflict (Goldstone, 1991; Collier et al., 2009). Many people and nations experience considerable deprivations – both in absolute and relative terms – but do not experience the outbreak of insurgencies. This means that potential insurgents must perceive opportunities for mobilization as well.

Opportunities

According to opportunity-based approaches, violence emerges where the costs of using it are low and the probability of success is high (Fearon and Laitin, 2003; Collier et al., 2009; Hegre and Sambanis, 2006). In terms of domestic conditions, insurgency should be easiest where the state is incapable of repressing dissent because of a weak security apparatus or a lack of familiarity with local populations (Fearon and Laitin, 2003). Thies (2010) finds that primary commodities may have an indirect effect on the onset of armed conflict through the reduction of state capacity. We should also expect to see a higher probability of insurgency onset where weapons are easily accessible, such that large inflows of small arms could signal an impending insurgency. Fearon and Laitin (2003) find that mountainous terrain often facilitates insurgency in that insurgents are better able to take advantage of the terrain to evade government repression. New states, which may not have yet addressed prior grievances, established popular legitimacy, or incorporated armed militias, may also be more likely to experience insurgency (Fearon and Laitin, 2003). Opportunities for insurgency may persist in countries where civil war has recently occurred, draining the government's coercive capacity, allowing potential insurgents to access leftover weapons, and in some cases, providing former rebels with know-how and training (Zukerman Daly, 2012). Some scholars have also found that youth bulges – wherein a disproportionate percentage of the population is between the ages of 15 and 34 – are associated with the onset of civil conflict (Urdal, 2006). Large youth populations facilitate recruitment into armed groups, satisfying the need for willing participants (Goldstone, 1991).

Several external conditions may also increase perceived opportunities for insurgent mobilization. First, if an insurgency is occurring in a neighboring state, the easy availability of weapons, know-how, and additional fighters may convince would-be recruits to mobilize (Kathman, 2010; Gleditsch, 2007; Fearon and Laitin, 2003 Salehyan, 2007). Second, a foreign state's promise of weapons, cash, logistical support, intelligence, or sanctuary may also create opportunities – and incentives – for a new insurgency to set on (Chenoweth and Stephan, 2011; Record, 2006; Byman, 2005).

COIN: Theory and Practice

Intimately connected to the understanding of insurgency, and of primary concern to security studies scholars, and military practitioners, is the study of COIN. COIN theory may be approached from two basic perspectives or approaches. Nagl describes these approaches as the 'direct approach' and the 'indirect approach' (Nagl, 2002: p. 27–28).

The 'direct approach,' sometimes described as the 'French' school, is 'enemy' or 'kinetic' focused. It downplays the distinctive characteristics of insurgency and seeks to achieve victory through direct military action against insurgents. Frisch takes this perspective in his article on Israeli COIN strategy, arguing that the essential factor in preventing violence by Palestinian insurgent groups has been the military disruption of insurgent capabilities by the Israel Defense Forces (Frisch, 2006). Historical examples of this school of COIN in practice include the French COIN campaign in Algeria from 1954 to 1962 and the US 'search-and-destroy' campaign during the Vietnam War.

The 'indirect approach,' sometimes described as the 'British' school, focuses on defeating the insurgent through winning the 'hearts and minds' of the local populace. This is traditionally sought through good governance and effective security. Furthermore, this approach recognizes the importance of

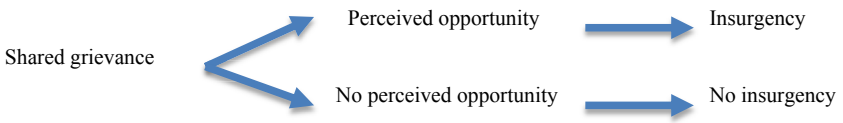

Figure 2 Grievance-opportunity model of insurgency onset.

addressing the political grievances which gave rise to the insurgency. The counterinsurgent seeks to reconcile these political problems on his or her own terms, thus maintaining control while denying the insurgency its raison d'etre. This approach centers on the basic understanding of insurgency as a primary political activity. If people can be convinced both of the effective control and political legitimacy of the counterinsurgent, then the insurgent group will be unable to gain the material and ideological support necessary to continue its operations. Through increasing political irrelevance and waning numbers, the insurgency will ultimately be defeated.

The textbook example of this theory in practice is the British COIN campaign against the Malayan National Liberation Army in Malaya from 1948 to 1960. However, some scholars have argued that even in the case of Malaya, scholarship has overstated the importance of the British 'hearts and minds' campaign, and asserted that British victory in Malaya was assured primarily through the use of overwhelming manpower and continuous control of populated areas. The efforts to win 'hearts and minds' came only when the insurgency's main force had already been effectively defeated (Bennett, 2009; Hack, 2009).

While significant debate remains as to the efficacy of either of these approaches, the academic and military communities have largely embraced the 'British' school of COIN. The political focus on 'hearts and minds' is foremost in U.S. Army FM 3-24 'Counterinsurgency,' written by Generals David Petraeus and James F. Amos, which has served as the primary source of military doctrine for the American COIN campaigns in Iraq and Afghanistan (at least after 2005).

However, the U.S. COIN campaign in Iraq has given rise to a number of criticisms of this COIN strategy. Some have argued that the emphasis on good governance as a means of gaining legitimacy lacks relevance in communities where ethnic allegiances are the primary factor in determining legitimacy (Fitzsimmons, 2008). And COIN theorist David Kilcullen has argued that many contemporary Islamist insurgencies are best understood as theaters of operation aggregated into a larger 'global insurgency.' The insurgents in this global insurgency are motivated not by the traditional insurgent goal of regime change but by a larger vision of systemic change in the world order. This difference in central motive necessitates a radically different COIN response, including disaggregating local and regional grievances from the larger cause of global change (Kilcullen, 2005).

This dispute on the nature of insurgency illustrates one of the primary challenges of effective COIN: the necessity for clear understanding of the insurgency and accurate intelligence on its operations. As O'Neill states, "An effective counterinsurgency program depends on an accurate, substantive, and comprehensive profile of the adversary and the environmental context in which he operates" (O'Neill, 2005: p. 155). A correct understanding of the nature and context of the insurgency gives the incumbent government the necessary background to make correct policy prescriptions in its efforts to deny the insurgency political traction with the people. Many have argued convincingly for the importance of counterinsurgents understanding the culture, language, and local context of the people in which insurgents seek to operate (Kilcullen, 2006; Lyall, 2010). Clear understanding of insurgents' tactical preferences will also dictate varying COIN responses. Urban insurgency characterized by terrorism and the use of explosives necessitates a different response than rural insurgency characterized by guerrilla hit-and-run tactics. Because the insurgency avoids direct, open confrontation, it is essential that the counterinsurgent maintains excellent intelligence throughout a COIN campaign.

A second primary tactical challenge in COIN is distinguishing between insurgents and the people. Because insurgents deliberately conceal themselves within the population, COIN operations are often unable to differentiate between insurgents and civilians, leading to significant collateral damage. This has been particularly true in contemporary cases of insurgency in Iraq and Afghanistan, which have been primarily urban-based, in contrast to the primarily rural base of many 'classical' insurgencies (Kilcullen, 2006). While Lyall argues that in some circumstances indiscriminate violence against civilians may lead to a reduction in violence (Lyall, 2009) the general scholarly consensus is that collateral damage backfires on the counterinsurgent force by creating new grievances against the counterinsurgent and greater sympathy for the insurgent (Kocher et al., 2011; Condra and Shapiro, 2012). Condra and Shapiro in particular argue that civilians punish both insurgents and counterinsurgents for noncombatant casualties through the choice to share or withhold information regarding insurgents' activities from counterinsurgents, significantly hindering the counterinsurgents' ability to obtain effective intelligence on insurgents (Condra and Shapiro, 2012).

One tactical response to this dilemma proposed by Gompert is the use of a 'continuum of force' by COIN forces. Gompert argues that COIN operations may be conducted using a variety of nonlethal weapons such as 'directed sound and light' to separate insurgents from the larger population without causing civilian casualties (Gompert, 2009).

COIN theory and practice continues to evolve in response to the changing nature of modern insurgency. For instance, the rise of the Internet usage by insurgents has led to a reconsideration of the traditional idea of insurgent sanctuaries, and a subsequent emphasis on the importance of denying potential insurgents an 'online sanctuary' through which to build their cause. And doubtless COIN will continue to evolve in response to the United States' continuing 'live experiments' in COIN in Afghanistan and elsewhere around the world.

The Outcomes of Insurgencies

How often do insurgents win? Between 1900 and 2006, there was an average success rate of about 25% among insurgencies that involved at least 1000 battle deaths and sought to overthrow incumbent governments, achieve territorial secession, or drive out foreign occupation (**Figure 3**). Rates of insurgent victory peaked in the 1970s and 1980s, with close to 40% success. Secessionist insurgencies tend to be less successful than antigovernmental or anti-occupation campaigns.

Four major factors contribute to insurgent victory: resources, sanctuary, participation, and opponent response. First, insurgencies often have higher chances of success if they secure sponsorship by an outside state – particularly powerful states – that can help to finance and arm the group (Bob, 2005; Chenoweth and Stephan, 2011; Record, 2006). Second,

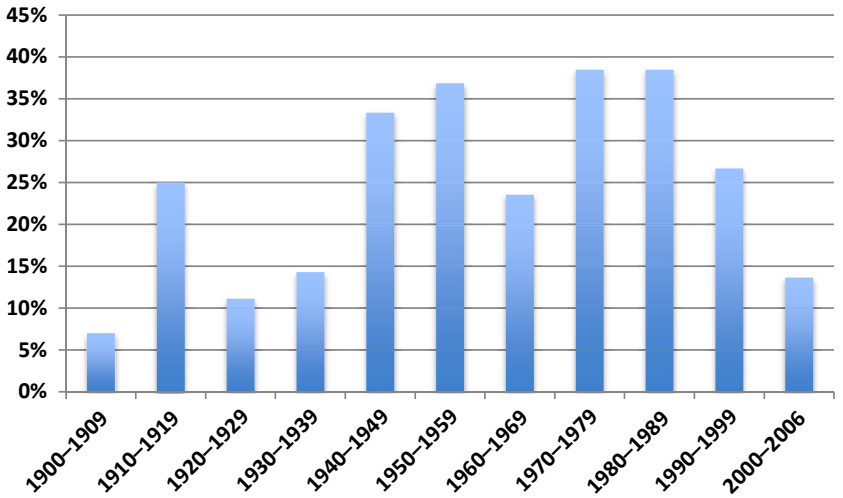

Figure 3 Percentage of victories among insurgencies, 1900–2006.

because insurgent groups are often vulnerable to repression by the state, sanctuary can prove an indispensible asset to many groups – particularly those with territorial goals (Salehyan, 2007). Third, participation is a crucial and oft-examined factor of success (Chenoweth and Stephan, 2011; Weinstein, 2007; DeNardo, 1985). Groups with large numbers are more flexible, more resilient to repression, and more capable of waging durable wars of attrition against their opponents. Fourth, opponent response is an important factor in explaining insurgent outcomes. States that are fragile or particularly incompetent in their response to the insurgency may be incapable of resisting a popular insurgency. Alternatively, selecting an inappropriate method of COIN may doom attempts to halt an insurgent campaign.

Although researchers have identified factors leading to the success and failure of insurgencies, they have often paid little attention to the considerable variation in insurgent success rates over time.

As Figure 4 demonstrates, whereas insurgency proved to be a relatively effective method of coercion during the mid-twentieth century with success rates far above the mean, it appeared far less fruitful pre-1930s and post-1990. What explains the variation in these rates of effectiveness over time?

Lyall and Wilson (2009) may have a partial answer. They argue that because of the mechanization of modern armies, counterinsurgent operations have become far less effective. Instead of using war fighting techniques that involve on-the-ground interactions with the civilian population consistent with the 'indirect approach' discussed above, contemporary armies tend to rely on technologies – aircraft, tanks, and the like – that emphasize the 'direct approach' and therefore provide very little population-centric information about the strategic environment. As such, intelligence – often argued to be the crucial element of COIN – is lacking. This may be true for the mid-twentieth century, where insurgent groups tended to enjoy far more victories than in recent years. Perhaps, modern

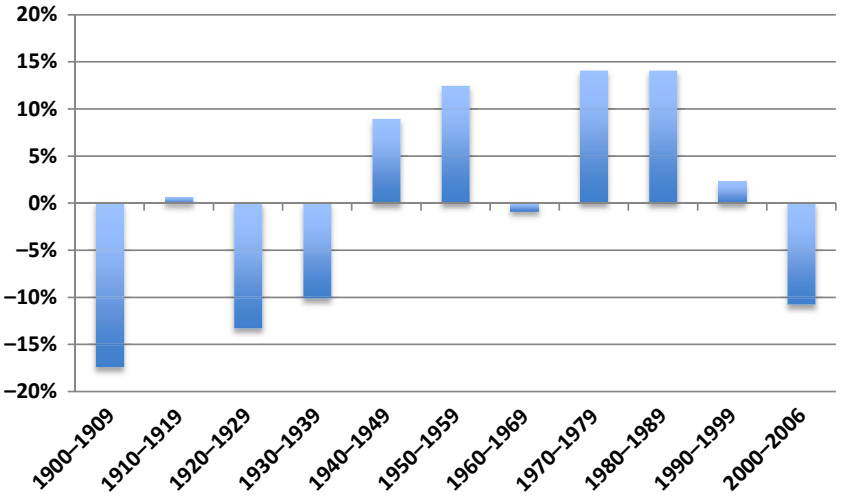

Figure 4 Change in insurgent victory rates, relative to mean, 1900–2006.

militaries have upgraded their understanding of COIN to include more population-centric COIN approaches.

A second potential explanation relates to the rise of nonviolent mass movements as alternatives to violent insurgencies. Recent research suggests that nonviolent resistance campaigns against dictatorships and foreign occupations tend to be more successful than violent insurgencies (Chenoweth and Stephan, 2011). This is especially true since the 1980s, when waves of nonviolent resistance swept through Eastern Europe and the Soviet Union. The decline in the effectiveness of violence may correspond to the rise of effectiveness of nonviolent resistance, suggesting that mass nonviolent movements may be displacing armed uprisings as the resistance method of choice in many contexts.

A third potential explanation – particularly for the decline in insurgent victories since 1990 – is the absence of willing sponsors of insurgent organizations. Although the United States and the Soviet Union avoided direct confrontation during the Cold War, both sides supported proxies throughout their so-called 'spheres of influence' in Latin America, Asia, Africa, Europe, and the Middle East. As such, the 1960s and 1970s saw the highest rate of state sponsorship of insurgencies, with nearly 55% of new insurgencies enjoying the sponsorship of a foreign power (compared with 38% over the entire 1900–2006 era). Incidentally, those insurgencies, which likely concluded in the following two decades, also tended to be the most successful against their opponent states.

Since the end of the Cold War, however, state sponsorship of insurgencies has returned to average rates – or fallen below them, depending on the decade. In the absence of great power rivalry, states may see sponsorship of insurgent groups as unnecessarily costly and provocative to their regional rivals.

Recent events in the Middle East and North Africa, however, may signal a return to the strategy of supporting insurgent movements. The NATO intervention in support of Libyan rebels against Muammar Qaddafi – and alleged Saudi, Qatari, and Turkish support of Syrian rebels against Bashar al Assad – reflects that states are still willing to take some risks in sponsoring violent nonstate actors who oppose their enemies. But, whether such strategies will continue to yield insurgent victories or longer-term strategic gains remains to be explored in future research.

Future Directions for Policy and Research

Important questions remain concerning the causes, outcomes, and responses to insurgency. With regard to the causes, the scholarly emphasis on broad structural conditions (i.e., grievance and opportunity) introduces two core problems: although extant research speaks to why people rebel, they do not answer precisely when or how people rebel. For example, such arguments fail to explain the specific causal mechanisms that drive insurgencies to break out at particular points in time. Populations may have a similar set of grievances and opportunities for long periods of time before the outbreak of insurgency. Why did an insurgency break out in Syria in 2011 but not in 2001 when the conditions of the Syrian population were largely similar?

Regarding the means of rebellion, both grievance and opportunity approaches assume that when potential insurgents do decide to mobilize, they do so using violence. A conventional wisdom persists that aggrieved groups respond – almost automatically – to structural conditions by arming themselves and using violence against the state. But in reality, campaigns of nonviolent resistance often occur before, during, and after insurgencies – and many nonviolent resistance campaigns pursue similar political objectives, and succeed, without an insurgency ever breaking out. Thus, grievances and opportunities may produce different patterns of resistance, which may or may not include insurgency. Further research is needed to clarify the conditions under which aggrieved parties with the capacity to undertake insurgency choose these various patterns of resistance.

With regard to the outcomes of insurgency, key questions also remain. First, researchers must better understand why insurgency is declining in its effectiveness over time. Although some plausible explanations are mentioned above, none of these has been systematically tested. Second, as mentioned above the post-Cold War dynamics of state sponsorship for insurgency, often an essential element for insurgent success, remain murky, with little research done into when, why, and how state actors choose to support violent resistance movements.

Finally, more questions remain about how states can better respond to insurgencies. In particular, most of the literature on the subject has been operational rather than scholarly in nature. Those who have developed 'best practices' in COIN have not yet ascertained whether those practices can easily apply in other cases. This critique particularly applies in the debate regarding collateral damage and the effect of civilian casualties on the success of insurgency. While several studies have proposed possible answers to this question based on specific COIN campaigns (Lyall, 2009; Condra and Shapiro, 2012; Kocher et al., 2011) there is little large-scale research.

Intimately related to this question is the question of how a counterinsurgent's actions may encourage or discourage participation in or support for insurgency. Is it more important – strategically and normatively – for the counterinsurgent to appeal to the peoples' 'hearts and minds' through good governance, political accommodation, and a light military touch? Or is it preferable to compel obedience and cooperation through a powerful kinetic-based COIN campaign?

More research is also needed into how COIN may be prosecuted effectively against insurgencies that do not fit the classical paradigm. What strategies and tactics are most effective against insurgencies that transcend the boundaries of the state, either through their character as the 'global insurgency' (Kilcullen, 2005; Jones and Smith, 2010) or through the use of new media and the shadowy sanctuary of cyberspace? While key concepts and principles of insurgency and COIN brought to prominence in the mid-twentieth century remain largely visible in the contemporary world, scholarship must continue to evolve as new tactics and technologies revolutionize the nature of insurgency and political violence.

See also: Civil Wars; Military, War, and Politics; Sociology of War; Terrorism.

Bibliography

Beckett, Ian F.W., 2001. Modern Insurgencies and Counter-Insurgencies: Guerrillas and Their Opponents since 1750. Routlegde, New York.

Bennett, Huw, 2009. 'A very salutary effect': the counter-terror strategy in the early Malayan emergency, June 1948 to December 1949. Journal of Strategic Studies 32 (3), 415–444.

Bob, Clifford, 2005. The Marketing of Rebellion: Insurgents, Media, and International Activism. Cambridge University Press, New York.

Bohara, Alok K., Mitchell, Neil J., Nepal, Mani, 2006. Opportunity, democracy, and the exchange of political violence: a subnational analysis of conflict in Nepal. Journal of Conflict Resolution 50 (1), 108–128.

Byman, Daniel, 2005. Deadly Connections: States that Sponsor Terrorism. Cambridge University Press, New York.

Byman, Daniel, 2008. Understanding proto-insurgencies. Journal of Strategic Studies 31 (2), 165–200.

Carey, Sabine C., 2007. Rebellion in Africa: disaggregating the effect of political regimes. Journal of Peace Research 44 (1), 47–64.

Cederman, Lars-Erik, Weidmann, Nils B., Kristian, Skrede G., 2011. Horizontal inequalities and ethnonationalist civil war: a global comparison. American Political Science Review 105 (3), 478–495.

Chenoweth, Erica, 2011. The Nonviolent and Violent Campaigns and Outcomes Dataset, Version 1.1. University of Denver.

Chenoweth, Erica, Stephan, Maria J., 2011. Why Civil Resistance Works: The Strategic Logic of Nonviolent Conflict. Columbia University Press, New York.

Collier, Paul, Dominic, Rohner, 2008. Democracy, development, and conflict. Journal of the European Economic Association 6 (2–3), 531–540.

Collier, Paul, Anke, Hoeffler, Dominic, Rohner, 2009. Beyond greed and grievance: feasibility and civil war. Oxford Economic Papers 61 (1), 1–27.

Condra, Luke N., Shapiro, Jacob N., 2012. Who takes the blame? The strategic effects of collateral damage. American Journal of Political Science 56 (1), 167–187.

DeNardo, James, 1985. Power in Numbers: The Political Strategy of Protest and Rebellion. Princeton University Press, Princeton, NJ.

Do, Quy-Toan, Iyer, Lakshmi, 2010. Geography, poverty and conflict in Nepal. Journal of Peace Research 47 (6), 735–748.

Fearon, James D., Laitin, David D., 2003. Ethnicity, insurgency, and civil war. American Political Science Review 97 (1), 75–90.

Fitzsimmons, Michael, 2008. Hard hearts and open minds? Governance, identity, and the intellectual foundations of counterinsurgency strategy. Journal of Strategic Studies 31 (3), 337–365.

Frisch, Hillel, 2006. Motivation or capabilities? Israeli counterterrorism against Palestinian suicide bombings and violence. Journal of Strategic Studies 29 (5), 843–869.

Galula, David, 1964, 2006. Counterinsurgency Warfare: Theory and Practice. Praeger Security International, Westport, Connecticut, CT.

Gleditsch, Kristian S., 2007. Transnational dimensions of civil war. Journal of Peace Research 44 (3), 293–309.

Gleditsch, Kristian S., Ruggeri, Andrea, 2010. Political opportunity structures, democracy, and civil war. Journal of Peace Research 47 (3), 299–310.

Goldstone, Jack, 1991. Revolution and Rebellion in the Early Modern World. University of California Press, Berkeley and Los Angeles, CA.

Gompert, David C., 2009. 'Underkill': fighting extremists amid populations. Survival: Global Politics and Strategy 51 (2), 159–174.

Goodwin, Jeff, 2001. No Other Way Out: States and Revolutionary Movement, 1945–1991. Cambridge University Press, New York.

Gurr, Ted R., 1970. Why Men Rebel. Princeton University Press, Princeton, NJ.

Hack, Karl, 2009. The Malayan emergency as counter-insurgency paradigm. Journal of Strategic Studies 32 (3), 383–414.

Hegre, Havard, Sambanis, N., 2006. Sensitivity analysis of empirical results on civil war onset. Journal of Conflict Resolution 50 (4), 508.

Jones, David M., Smith, M.L.R., 2010. Whose hearts and whose minds? The curious case of global counter-insurgency. Journal of Strategic Studies 33 (1), 81–121.

Joshi, Madhav, Mason, T. David, 2008. Between democracy and revolution: peasant support for insurgency versus democracy in Nepal. Journal of Peace Research 45 (6), 765–782.

Kathman, Jacob D., 2010. Civil war contagion and neighboring interventions. International Studies Quarterly 54 (4), 989–1012.

Kaufman, Stuart J., 2006. Escaping the symbolic politics trap: reconciliation initiatives and conflict resolution in ethnic wars. Journal of Peace Research 43 (2), 201–218.

Kilcullen, David J., 2005. Countering global insurgency. Journal of Strategic Studies 28 (4), 597–617.

Kilcullen, David, 2006. Counter-insurgency redux. Survival: Global Politics and Strategy 48 (4), 111–130.

Kocher, Matthew A., Pepinsky, Thomas B., Kalyvas, Stathis N., 2011. Aerial bombing and counterinsurgency in the Vietnam War. American Journal of Political Science 55 (2), 201–218.

Lyall, Jason, Wilson III, Isaiah, 2009. Rage against the machines: explaining outcomes in counterinsurgency wars. International Organization 63 (1), 67–106.

Lyall, Jason, 2009. Does indiscriminate violence incite insurgent attacks? Evidence from Chechnya. Journal of Conflict Resolution 53 (3), 331–362.

Lyall, Jason, 2010. Are co-ethnics more effective counterinsurgents? Evidence from the second Chechen war. American Political Science Review 104 (1), 1–20.

Merom, Gil, 2003. How Democracies Lose Small Wars: State, Society, and the Failures of France in Algeria, Israel in Lebanon, and the United States in Vietnam. Cambridge University Press, New York.

Nagl, John A., 2002. Counterinsurgency Lessons from Malaya and Vietnam: Learning to Eat Soup with a Knife. Praeger Security International, Westport, Connecticut, CT.

Nepal, Mani, Bohara, Alok K., Gawande, Kishore, 2011. More inequality, more killings: the Maoist insurgency in Nepal. American Journal of Political Science 55 (4), 886–906.

O'Neill, Bard E., 2005. Insurgency & Terrorism: From Revolution to Apocalypse. Potomac Books, Inc., Dulles, Virginia, VA.

Record, Jeffrey, 2006. External support: enabler of insurgent success? Parameters 36 (3), 36–49.

Salehyan, Idean, 2007. Transnational rebels: neighboring states as sanctuary for rebel groups. World Politics 59 (2), 217–224.

Thies, Cameron G., 2010. Of rulers, rebels, and revenue: state capacity, civil war onset, and primary commodities. Journal of Peace Research 47 (3), 321–332.

Urdal, Henrik, 2006. A clash of generations? Youth bulges and political violence. International Studies Quarterly 50 (3), 607–630.

Vadlamannati, Krishna C., 2011. Why Indian men rebel? Explaining armed rebellion in the northeastern states of India, 1970–2007. Journal of Peace Research 48 (5), 605–619.

Walter, Barbara F., 2004. Does conflict beget conflict? Explaining recurring civil war. Journal of Peace Research 41 (3), 371–388.

Weinstein, Jeremy N., 2007. Inside Rebellion: The Politics of Insurgent Violence. Cambridge University Press, New York.

Wimmer, Andreas, Cederman, Lars-Erik, Min, Brian, 2009. Ethnic politics and armed conflict: a configurational analysis of a new global data set. American Sociological Review 74 (2), 316–337.

Zukerman Daly, Sarah, 2012. Organizational legacies of violence: conditions favoring insurgency onset in Colombia, 1964–1984. Journal of Peace Research 49 (3), 473–491.

Integrated Attachment Theory

Ella Levert-Levitt and Abraham Sagi-Schwartz, Center for the Study of Child Development, University of Haifa, Haifa, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article addresses core concepts of John Bowlby's attachment theory, underscoring the mechanism of primary drive reduction in the context of real-life events. We highlight Ainsworth's contribution to attachment theory and present her major advances such as the Strange Situation Procedure. We also review recent research including disorganized attachment, goal-corrected partnerships and attachment assessments through the life span, antecedents of attachment security/insecurity, the concept of maternal sensitivity, continuity and longitudinal correlates of attachment across the life span, attachment and genetics, and attachment-based interventions.

In one of his many influential papers, John Bowlby, father of the attachment theory, states: *Family doctors, priests and perceptive laymen have long been aware that there are few blows to the human spirit so great as the loss of someone near and dear. Traditional wisdom knows that we can be crushed by grief and die of a broken heart, and also that a jilted lover is apt to do things that are foolish or dangerous to himself and others. It knows, too, that neither love nor grief is felt for just any other human being but for one, or a few, particular and individual human beings. The core of what I am terming an affectional bond is the attraction that one individual has for another individual* (Bowlby, 1979: p. 75).

This statement is at the heart of attachment theory, which has since Bowlby's writing gained unprecedented research support and applied-clinical momentum in the field of human development. This wealth of research cannot be fully covered here. However, an interested reader may consult the 'Handbook of Attachment,' which has become the most authoritative reference (Cassidy and Shaver, 2008). Below we focus on what lies at the core of attachment theory, with attention to Bowlby and Ainsworth's perspective.

Bowlby's Trilogy of Attachment Theory

'Attachment,' 'separation,' and 'loss' are core terms of Bowlby's theory, and the names of his famous trilogy. Attachment theory addresses issues in human relationships and development, such as how children become attached to significant caring figures, what maintains close relationships, and how individuals react to separation or loss. Bowlby used simple albeit powerful terms and metaphors to express his ideas, and we continue to use them because their precision underscores forcefully the core tenets of attachment theory.

Attachment

According to attachment theory, all infants are innately biased to become attached to primary caregivers: *To say of a child that he … has an attachment to someone, means that he is strongly disposed to seek proximity to and contact with a specific figure, and to do so in certain situations, notably when he is frightened, tired or ill* (Bowlby, 1969: p. 371). The theory, which has been extensively validated cross-culturally (Van IJzendoorn and Sagi-Schwartz, 2008) emphasizes a 'specific figure' and 'certain situations,' mainly stressful. It is not a theory that refers to all human relations under all circumstances. Note however that the blooming of attachment theory and research in recent decades may have led to many misconceptions about how far Bowlby intended to go.

Primary vis-à-vis Secondary Drive Reduction

Perhaps because of his childhood experiences, and despite his background as a psychoanalyst, Bowlby strongly objected to the psychoanalytic consensus that viewed the infants' tendency to form relationships with their primary caregivers as 'secondary drive reduction.' Bowlby conceived such bonds as stemming from a primary and existential need, and viewed attachment as basic in and of itself: *Mother love in infancy is as important for mental health as are vitamins and proteins for physical health* (Bowlby, 1953: p. 182). In his seminal writing in *Child Care and the Growth of Love* (1953: p. 181), Bowlby asserted that *there is tolerance for conditions of bad mental hygiene in nurseries, institutions and hospitals… If it were found in the field of physical hygiene it would long since have led to public outcry*. He also opposed the view that the strong ties infants form with important adult figures evolves secondarily as an outcome of gratification with food or association with the mother's breast (Bowlby, 1940).

In an attempt to understand the implications of maternal deprivation, Bowlby was inspired by a broad spectrum of fields inside and outside psychoanalysis, including evolutionary biology, ethology, object relations theories, control systems theory, and cognitive and cross-cultural psychology. He was impressed by Spitz's (1946) writings on the effects of institutionalization on children, the work of Harlow (1961) on maternal caregiving and isolation from the mother in rhesus monkeys, the work of Lorenz (1935) on imprinting with geese, and of Hinde (1987) on ethology and social relationship, all of which express the ethological foundations of close relationships as independent primary motivational processes in survival terms (Bowlby, 1969, 1973, 1980; Duniec and Raz, 2011). Bowlby's personal background may also have contributed to his approach.

Edward John Mostyn Bowlby (1907–90) was born in London and grew up in an upper-middle-class family. Like many young children in similar families, he was cared for by

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23198-X

nannies. Occasions on which he was allowed to spend time with his parents were strictly regulated: *My mother held the view that it was dangerous to spoil children… bids for her affection and attention were the opposite of what was required* (Bowlby in Byng-Hall, 1987: p. 19). Encouraged by his father, he went to study medicine at Cambridge. One of his tutors was enthusiastic about Freud's ideas, leading Bowlby to read psychological literature. Later he worked with maladjusted children and delinquent youths, and concurrently on his Ph.D., training in medicine, psychoanalysis, and psychiatry (Van Dijken, 1998).

Fantasies versus Real-Life Events

While training as a psychoanalyst, Bowlby found himself in continuous disagreement with his supervisor, Melanie Klein, and his ideas were not warmly welcomed by the psychoanalytic society in the UK. Klein emphasized the importance of early fantasies, marginalizing real-life events. She argued that the child's internal representations of the mother were strongly shaped by the child's drives, and were the ones that formed the child's psychic reality. Riviere, another one of Bowlby's psychoanalysis trainers, maintained that *[child] analysis has no concern with the real world … it is concerned simply and solely with the imaginings of the childish mind* (Riviere, 1927: pp. 376–377). At the end of World War II, Bowlby was appointed head of the Children's Department at the Tavistock Clinic. Most of the work there had a Kleinian orientation, insufficiently considering life events, and therefore not viewing family interaction patterns as important. For Bowlby, however, maternal deprivation, separation from attachment figures, and unresolved loss of close figures were key determinants of future adaptive difficulties. Consequently, he formed his own research unit to study the implications of mother–child separation.

Bowlby was greatly impressed by the work of social workers, especially Robertson (1953), who emphasized real-life events, the role of the family and environment as formative factors in the child's well-being and psychopathology. Together the two conducted naturalistic observations of hospitalized and institutionalized children who were separated from their parents. Bowlby went beyond infancy, suggesting that *all of us, from the cradle to the grave, are happiest when life is organized as a series of excursions, long or short, from the secure base provided by our attachment figures* (Bowlby, 1988: p. 62). One is reminded of present-day travelers, who immediately upon landing in a foreign country activate their mobile phones to call home for emotional refueling, to hear the calming voice of their loved ones; home is our secure base, which enables us to safely and comfortably explore new territories.

Separation

Bowlby (1944) suggested in '44 Juvenile Thieves: Their Characters and Home Life' that prolonged separation from the main caregiver in the first 5 years of life may render the young child susceptible to maladaptive behavior.

Again, his personal story may be related to his thoughts about separation. When Bowlby was 4 years old, his nanny left the household: *For a child to be looked after entirely by a loving nanny and then for her to leave when he is two or three, or even four or five, can be almost as tragic as a loss of a mother* (Bowlby, 1958: p. 7). At age 7, Bowlby was sent to a boarding school, and his connection with home consisted of letters and one to four visits per year. At age 14 he entered the Royal Naval College at Dartmouth. He was probably unhappy in these settings and as an adult spoke strongly against them (Van Dijken, 1998).

Robertson and his famous documentary filmed observation 'A two-year-old goes to the hospital' (Bowlby and Robertson, 1952) was of formative significance for Bowlby, who contended that when a child is separated from an attachment figure unwillingly, he shows distress. Distress is likely to intensify if the child is exposed to an unfamiliar environment and the care is provided by a stranger. Bowlby described the child's behavior under such circumstances as progressing through a three-stage process: *protest* – expressing separation anxiety, the child uses his or her sensorimotor repertoire to draw his or her caregiver back; *despair* – after protesting, the child realizes that he or she is not able to regain his or her caregiver but remains preoccupied with the caregiver and alert to any signs of return (behavior signifying grief and mourning); and *detachment* – after a while, the child seems to lose interest in the caregiver and becomes detached from his or her as a defense strategy. However, the child does not remain detached indefinitely. After reuniting with his or her caregiver, his or her attachment system reactivates. After experiencing such a process, the child insists on staying close to his or her attachment figure, exhibiting elevated anxiety when suspecting that the caregiver might leave again.

For many years, pediatric wards worldwide did not allow parents to stay with their hospitalized children. The focus was on "medical intervention" per se, ignoring the children's other basic needs, such as attachment needs. Such (mal)practice and its possible ill effects on recovery from the illness have been studied extensively in recent decades, but Bowlby was a pioneer to express these ideas as early as the 1950s.

Loss and Grief

Under prolonged separation, as in the case in loss, grief and mourning take place with despair as the most pronounced behavioral marker. Bowlby's view is consistent with Parkes (2009) who emphasized the important function of working through the loss with successful resolution. Unlike the psychoanalytic tradition, Bowlby viewed depressive responses following the loss of an attachment figure not as outcome of unconscious fantasies but as a reaction to real-life events, namely, death, abandonment, or prolonged separation. According to Bowlby, children separated permanently from their attachment figures, without trustful substitutes – as in the case of many maltreated children placed in group care settings, or of abandoned and orphaned children in institutions – may lack the capacity to love because of an absence of opportunities in their early years to form close ties with an attachment figure.

Ainsworth and Growing Points in Attachment Theory

A fortunate coincidence, an advertisement in the *Times* magazine, connected Bowlby with Mary Salter Ainsworth who became his most influential colleague: Ainsworth (1913–99), a Canadian psychologist, wrote a doctoral dissertation

inspired by Blatz's 'security theory.' According to this theory, during early childhood, parental support of the child's explorations will help him form 'independent security': *where family security is lacking, the individual is handicapped by the lack of a secure base from which to work* (Salter, 1940: p. 45).

When Ainsworth joined Bowlby's team, she became impressed by Robertson's naturalistic observations as a way of recording children's behavior and decided to adopt the method when undertaking her own research. Her husband's anthropological research took Ainsworth to Uganda, where she began documenting normative developmental changes during infancy and individual differences in patterns of mother–infant interactions. On her return to the USA, she received a position at Johns Hopkins University and sought to replicate her findings from the Uganda study in a Western society.

In her famous Baltimore study, Ainsworth followed 26 babies from birth to age 12 months. In her detailed observations of mother–infant dyads, she noticed how mother's sensitive, suitable response correlated with a more harmonious and sensitive relationship at age 8–12 months. Ainsworth also found that maternal responsiveness to the infant's crying during the first months of life is correlated with less crying in the months to come. This finding was met with skepticism, especially among learning theorists, who argued that attending to the baby's cry is likely to reinforce rather than stop the crying, and will not enhance preverbal, vocal, and gestural communication (Bretherton, 2003).

Ainsworth's Strange Situation Procedure

When the Baltimore study's babies reached the age of 12 months, Ainsworth developed a laboratory procedure to examine how they use their mothers as a secure base from which they can explore their environment (Ainsworth et al., 1978). This procedure was inspired by Harlow's study that illustrated how infant rhesus monkeys, when separated from their mothers and placed in a strange environment, were able to find comfort and explore their surroundings if their 'cloth mother' (a 'stepmother' made of wire covered with cloth) was available. But if the monkeys were in the presence of the 'wire mother' (a 'stepmother' made of steel wire that only fed the monkey), the infant monkeys showed elevated levels of fear and did not engage in exploration. Ainsworth's 20-min laboratory procedure became the well-known 'Strange Situation Procedure' (SSP), the gold standard for assessing infant attachment.

In the SSP, the 12–18 months infant and his or her caregiver are invited into a room equipped with toys that encourage play and exploration. During the session, the infant is separated from the caregiver and left in the room with a stranger, then reunited with the caregiver twice. Most of the infants stand near the door, some cry vigorously when the caregiver leaves the room. On her return, they approach her, seeking proximity and contact, without expressing negative behavior. When comforted by their caregiver, the infants resume exploration and play, hence the term *secure* (type B) attachment. Ainsworth found that secure attachment is associated with the attachment figures' sensitive responses throughout the first year of life. But not all infants display secure behavioral pattern. Some avoid the caregiver when she return, look away, or reject her bids for contact. These infants are referred to as having *insecure avoidant* (type A) attachment, masking negative affect and minimizing the importance of the caregiver as a source for comfort and security. Infants use this behavioral strategy in response to recurrent rejection by their caregiver, especially in times of need, which is a defensive strategy to avoid further rejection. A third group of infants, smallest in size, appears angry with the caregiver upon reunion, although they protest and show excessive distress when she leaves. The caregiver could not soothe the infant, who often resists when she tries to comfort him. These infants were named *insecure- resistant* or *insecure ambivalent* (type C). Insecure-ambivalent children exhibit proximity/contact maintenance with their caregivers mingled with excessive motor resistance to gain the inconsistently available caregiver's attention.

Ainsworth had a revolutionizing effect on how we understand infant–mother relationships and human relationships in general. Her insights fundamentally expanded attachment theory. She subjected Bowlby's theory to an empirical test by identifying patterns in mother–infant interaction that were implicated in differences in the quality of attachment relationships. She was first to note that parental sensitivity and responsiveness to the child's signals are key precursors of secure attachment. Indeed, both Bowlby and Ainsworth believed it was necessary to adapt scientific approaches, as they did, in order to validate the central concepts associated with attachment theory.

Other researchers and clinicians have taken these concepts further, proposing the *circle of security* model. According to Ainsworth and Bowlby, feelings of security equip the infant with a *secure base* from which he can depart for exploration and return to a *safe haven* when distressed and in need of emotional refueling. Jointly, secure base and safe haven create a *circle of security*. Social environments that do not provide a secure base or a safe haven generate a sense of *limited security* (Cooper et al., 2005). We suggest that *limited security* vis-à-vis insecure attachment may reflect more accurately the feelings and reality of some normative children who lack the full circle of security. In less favorable but not necessarily pathological environments, children who cannot fully rely on their caregivers' support and emotional availability, develop a sense of limited security.

Insecure attachment or *limited security* should be discussed in the context of what adaptation means. According to Hinde's (1982) position on conditional strategies, both mothers and babies are equipped to elicit not one set of interactions but a wide range of potential relationships compatible with specific environmental requirements. Main (1981) argued that both secure and insecure attachments are adaptive strategies to a given environmental condition, but insecure attachment has a higher adaptive cost for the infant and may be less functional in day-to-day life. Nevertheless, insecurity cannot be equated with psychopathology, and further research is needed to specify more concretely what such costs might be, namely the long-term implications for the well-being of the individual.

Mary Main and Disorganized Attachment

According to Main (1990), Ainsworth's student and collaborator, attachment classifications should be regarded as

behavioral strategies: the infant achieves the goal of maintaining proximity to the caregiver by planning his or her actions based on the caregivers' behaviors and responses. Despite the striking differences between secure and insecure babies, all experience consistent caregiving environments that enable them to develop an *organized attachment* strategy, adaptive to their environment and to the attachment figures' behaviors. But what happens when the attachment figure is also a source of fear, threat, or violence? Infants who are exposed to frightened/frightening parental behavior develop insecure attachment and may find it impossible to form an organized strategy toward their caregivers because they cannot depend on the caregiver's trust and security. These infants may withdraw and show insecure disorganized behaviors.

Main and Weston (1981) noticed that even in a normative sample 13% of infants could not be assigned to any of the three organized categories of attachment. In their parents' presence, unclassifiable infants showed 'contradiction in movement pattern' by expressing a range of disorganized, bizarre, stereotypical, unpredictable, and confused behaviors. Main and Hesse (1990) called it a vicious circle of 'fright without solution' that could lead to the collapse of organized attachment strategies. Disorganized attachment occurs more frequently with maltreated children. The effect of a disorganized attachment may go beyond infancy. Unlike insecure attachment, infant disorganized attachment was correlated with dissociative behaviors later in life, with difficulties in stress management (Van IJzendoorn et al., 1999), and with externalizing problems in childhood (Fearon and Belsky, 2011).

Although the SSP has yielded massive amounts of meaningful findings, its full-scale validity has been debated because of its reliance on laboratory settings alone, and for being restricted to the beginning of the second year of life, overlooking later developmental phases. Waters and Deane (1985) introduced a different method of assessing attachment security in infants and toddlers, known as the Attachment Q-Set (AQS). The AQS evaluates the child's behavior in his or her natural setting, by sorting a large number of cards that portray various behavioral characteristics of the child, with a focus on secure base behaviors. The sorting is compared with a prototypical profile of a secure child provided by experts, leading to a security score.

Assessing Attachment beyond Infancy

Following the development of secure or insecure behavioral patterns in infancy, children develop mental representations regarding their relationship with their caregivers, referred to as 'internal working models' (Bowlby, 1969). This change is facilitated by the development of new cognitive, linguistic, and social capacities, as attachment relationships move to what Main et al. (1985) referred to as 'the level of representations.' Bowlby (1982) termed this phase *goal-corrected partnership* (GCP). As children become increasingly able to express their feelings, thoughts, and motives, they realize that their parents may have goals and expectations different from their own. In such partnerships, both child and parent must consider the other's perspective and negotiate personal perspectives in order to gain a shared view of emotional and relationship issues. Individual differences in attachment security/insecurity determine the extent to which affective negotiations during conversations become congruent (Bowlby, 1973). The openness and autonomy that characterize child–parent conversations may vary in secure and insecure dyads (Bowlby, 1988): open communication enables the child to express a wide range of thoughts, feelings, and concerns, whereas non-open communication restricts his or her ability to express his or her needs and affects, excluding or over emphasizing meaningful emotional themes. Several methods focusing on verbal communication, parent–child dialogs or children's narratives, were developed to assess individual differences in attachment-based representations and dyadic negotiations (see Bretherton et al., 1990; Cassidy, 1988; Etzion-Carasso and Oppenheim, 2000; Gini et al., 2007; Oppenheim et al., 2007).

Parental State of Mind and Adult Attachment Interview

Main et al. (1985) hypothesized that adults' mental representations of their childhood attachment experiences crystallize into a relatively stable 'state of mind,' which may significantly affect the quality of their own child's attachment to them. The Adult Attachment Interview (AAI; George et al., 1984/1985/1996) was developed to capture such states. The AAI is a semi-structured interview in which the interviewee's childhood memories and current perceptions of these experiences are discussed. The interviewees describe attachment-related experiences from their childhood, including situations of emotional upset, illness, and separation from the parents. The AAI captures the speakers' current state of mind concerning their attachment representations stemming from childhood experiences with specified caring figures. Individuals may show a state of mind that is autonomous/free to explore (F), dismissing (Ds), or preoccupied/enmeshed (E). Individuals exhibiting errors in monitoring the discourse, when discussing traumatic or loss events, are rated as unresolved (U) with regard to a specific loss/trauma.

The AAI has been used extensively in studies of intergenerational transmission of attachment (Van IJzendoorn, 1995), demonstrating its potential strengths. Many questions remain open, including the extent to which the AAI coding system captures the underlying attachment-related representations for all life experiences and circumstances (see, for example, Koren-Karie et al. (2003) on the Absence of Attachment Representations (AAR) when experiencing catastrophic parental loss in childhood).

Antecedents of Attachment Security/Insecurity

Maternal Sensitivity

Ainsworth suggested that 'maternal sensitivity' shapes infant's secure attachment. Maternal sensitivity is the ability to read and interpret the baby's signals and to respond promptly and harmoniously. When such capacity is inadequate we may observe misinterpretations of the infant's behaviors resulting in disharmonious responses, conceived as maternal insensitivity.

Sensitive responses are likely to contribute to secure attachment. But not all studies yielded high correlations between attachment and maternal sensitivity (see *meta-analyses* by Goldsmith and Alansky, 1987; De Wolff and van IJzendoorn, 1997). Although maternal sensitivity contributes to secure attachment, it cannot be considered an exclusive factor, leaving a transmission gap, namely, unexplained variance in infant attachment security (Belsky, 2005).

Ainsworth focused on behavioral markers that underlie maternal sensitivity and explored it more broadly, at the level of the mother's state of mind and of her capacity to view the children's experiences from their point of view (Ainsworth et al., 1978): this is the capacity to consider the child's feelings, internal motives, and goals. Oppenheim and Koren-Karie (2009) referred to the mothers' insights regarding their children's internal experiences as 'maternal insightfulness.' They developed a new method for assessing insightfulness, and their recent research supports its discriminant and predictive validity. Mothers' capacity to consider the motives underlying their children's behaviors and emotional experiences in an insightful and open manner contributed significantly to the explained variance in attachment security above and beyond the mothers' interactive sensitivity. A more detailed discussion on the extensive research on sensitivity is beyond the scope of this article. For the latest findings on the subject see also the special issue of *Attachment and Human Development* (Grossmann et al., 2013) – 'Maternal Sensitivity: Observational Studies Honoring Mary Ainsworth's 100th Year.'

Continuity and Longitudinal Correlates: Attachment across the Life Span

As attachment theory became a central theoretical framework, longitudinal correlates of early relations began to occupy attachment research. But as Sagi-Schwartz and Aviezer (2005) noted, available longitudinal studies of attachment were inconsistent in their predictions from infancy to adulthood, concluding that children vary considerably in the degree to which attachment security in infancy remains consistent over time. Therefore, predicting later development based on early attachment experiences is a challenging and complex task that requires consideration of many issues pertinent to both longitudinal research and attachment theory.

An important issue that most longitudinal studies fail to address is the interface between coherent developmental change across time and susceptibility to environmental influences. Thompson (2008) suggested that both early experiences and contemporaneous events may influence developmental outcomes, and that the joint effect of the two may not be similar for different children. Furthermore, more stable living conditions would be more likely associated with continuity in attachment and its outcomes.

Another important issue, addressed by Sagi-Schwartz and Aviezer (2005), is whether early attachment should be conceived as a grand theory-based construct (Waters et al., 2000) predicting a broad spectrum of domains, or whether predictions should be restricted a narrower view, with a focus on attachment-based constructs (Grossmann and Grossmann, 1990). Longitudinal studies have applied both broad and narrow constructs (Goldberg et al., 1999), which may account for the great variability in research findings (Thompson, 2008).

Sagi-Schwartz and Aviezer (2005) also noted the need for a closer study of networks of attachment relations. They propose to explore all of the child's relationships and test whether the cumulative effect of the attachment network is a more powerful predictor of future development than each separate attachment relationship by itself. Specifically, is there a dominant attachment figure within a hierarchy of relationships that has a more pronounced effect on future personality development than do other attachment figures? It is not clear yet whether such influences are organized into a generalized personality construct (and if so, how) or into relationship-specific attachment representations (see more on this issue: Furman and Simon, 2004).

More work is necessary to determine whether and how experiences with different attachment figures exert differential/specialized effects on children, contributing unique domain-specific components to personality development (Grossmann et al., 2005; Oppenheim et al., 1988; Sagi-Schwartz and Aviezer, 2005). In all, existing research suggests that the developmental trajectories from infancy to adulthood are multidetermined and multifaceted. Various conceptual formulations, some contradicting each other, compete for settling the issues of continuity/discontinuity and stability/instability. Sagi-Schwartz and Aviezer (2005) noted that when pioneers in the field began their studies in the early 1970s, they were concerned mainly with the procedure for assessing infant attachment. Longitudinal research evolved as a by-product of these exciting developments and was not based on a priori conceptual or methodological planning. Much of this research was based on hypothesis-generating rather than hypothesis-testing models, and therefore many of the issues mentioned above remain to be addressed. Finally, the recently proposed concept of 'differential susceptibility' (see below) may be used also as an explanatory mechanism for inconsistencies in longitudinal research that often neglect to consider the genetic makeup of both parents and children.

Attachment and Genetics

Two lines of research on the interaction between attachment and genetics are noted. The first follows behavioral genetic models. In one such study, Bokhorst et al. (2003) found that 52% of the variance in attachment security was accounted by environmental factors and 48% by unique environmental factors/measurement errors, leaving no room for genetic effects. But genetic factors played a significant role in explaining 77% in temperamental reactivity, and differences in temperamental reactivity did not correlate with attachment concordance.

The second line of research follows molecular genetics and the study of gene–environment interaction. It examines how similar environments may have differential effects on children because of the children's genetic makeup. The idea of differential susceptibility (Belsky, 2005) which is gaining momentum in current research, suggests that because of

genetic markers some children are more or less susceptible to both positive and negative effects of similar caregiving practices. Bakermans-Kranenburg and van IJzendoorn (2007) found that interactions between specific susceptibility markers of the DRD4 gene and unfavorable environmental factors explain individual differences in negative outcomes, such as risk for disorganized attachment. They also present evidence of more positive outcomes when susceptible children are exposed to favorable environments. How does this relate to Bowlby's work? Bowlby did not use the contemporary notation of genetic X environment interaction. He did pay attention, however, to the possible influences of genetic factors on human development: *Psychoanalysts like the nurseryman should study intensively, rigorously, and at first hand, the nature of the organism, the properties of the soil and the interaction of the two* (Bowlby, 1940: p. 23).

Attachment-Based Interventions

Attachment theory has also inspired the clinical arena and the field of intervention, and various attachment-based interventions became available. For example, Video-Feedback to Promote Positive Parenting and Sensitive Discipline (VIPP-SD), (Juffer et al., 2008; http://www.marinusvanijzendoorn.com/video-feedback-intervention-vipp); Attachment and Biobehavioral Catch-up (ABC) (Dozier et al., 2005) targets parental behaviors among high-risk and foster parents (http://abcintervention.com/); Minding the Baby (MTB) (Sadler et al., 2013), an interdisciplinary, intensive home-visiting program that focuses on first-time, young, at-risk mothers and their families (http://mtb.yale.edu/index.aspx); and Circle of Security (Cooper et al., 2005), a group psychoeducational interventions seeking to alter attachment interaction in high-risk caregiving dyads (http://circleofsecurity.net). Each of these attachment-based interventions has a different focus and requires further research support. In their *meta-analysis*, Bakermans-Kranenburg et al. (2003) showed that short-term interventions with a behavioral focus were more effective than long-term intervention, suggesting that 'less is more.' Finally, interventions that succeeded in improving parental sensitivity were more effective in enhancing attachment security.

Summary

Bowlby proposed his attachment theory based on observations of the difficulties experienced by children exposed to adverse environments, mainly neglect, abandonment, and institutionalization. In the last decades, however, attachment theory has evolved in various directions, including both normative and at-risk children and settings. Today attachment theory reaches into such areas as representations, mentalization, caregiving, marital relationships and divorce, child care, foster care and adoption, religious orientation, intervention, and recently even political psychology (Sagi-Schwartz, 2012). These are extensively reviewed in two editions of the *Handbook of Attachment* (Cassidy and Shaver, 1999/2008). The *Journal of Attachment and Human Development*, under the editorial leadership of Howard Steele and Jude Cassidy, is another important platform for discussing attachment-related issues.

In closing, we wish to call attention to one direction that represents Bowlby and Ainsworth's integrated perspective of attachment theory and research over the past decades. It relates to children at profound risk in their environment. Recently an international group of researchers produced a 'Consensus Statement on Group Care for Children and Adolescents' also endorsed as a policy statement of the American Orthopsychiatric Association. Accordingly, *every child has a basic right and need to grow up in a safe home with a stable continuous relationship with at least one adult who is a trusted, committed parent figure. Group settings should not be used as living arrangements, because of their inherently detrimental effects on the healthy development of children, regardless of age* (Dozier et al., 2014).

Acknowledgments

We thank Dr. Nina Koren-Karie and Dr. David Oppenheim for their invaluable comments.

See also: Attachment During Infancy and Early Childhood: Cultural Variations, Development of; Friendship During Childhood and Cultural Variations; Grandparent–Grandchild Relationships; Mother-Child Communication: Cultural Differences.

Bibliography

Ainsworth, M., Blehar, M., Waters, E., Wall, S., 1978. Patterns of Attachment. Erlbaum, Hillsdale, NJ.

Bakermans-Kranenburg, M.J., van IJzendoorn, M.H., 2007. Research review: genetic vulnerability or differential susceptibility in child development: the case of attachment. Journal of Child Psychology and Psychiatry 48, 1160–1173.

Bakermans-Kranenburg, M.J., Van IJzendoorn, M.H., Juffer, F., 2003. Less is more: meta-analyses of sensitivity and attachment interventions in early childhood. Psychological Bulletin 129, 195–215.

Belsky, J., 2005. Differential susceptibility to rearing influences. In: Ellis, B.J., Bjorklund, D.F. (Eds.), Origins of the Social Mind: Evolutionary Psychology and Child Development. Guilford, New York, pp. 139–163.

Bokhorst, C.L., Bakermans-Kranenburg, M.J., Fonagy, P., Schuengel, C., 2003. The importance of shared environment in mother–infant attachment security: a behavioral genetic study. Child Development 74, 1769–1782.

Bowlby, J., 1940. The influence of early environment in the development of neurosis and neurotic character. International Journal of Psycho-Analysis XXI, 1–25.

Bowlby, J., 1944. Forty-four juvenile thieves: their characters and home life. International Journal of Psychoanalysis 25, 1–57, 207–228.

Bowlby, J., 1953. Child Care and the Growth of Maternal Love. Penguin Books, London.

Bowlby, J., 1958. Can I Leave My Baby? National Association for Mental Health, London.

Bowlby, J., 1969. Attachment and Loss. In: Attachment, vol. 1. Pimlico, London.

Bowlby, J., 1973. Attachment and Loss. In: Separation – Anxiety and Anger, vol. 2. Pimlico, London.

Bowlby, J., 1979. The Making and Breaking of Affectional Bonds. Routledge, London.

Bowlby, J., 1980. Attachment and Loss. In: Loss, vol. 3. Pimlico, London.

Bowlby, J., 1982. Attachment and loss: retrospect and prospect. American Journal of Orthopsychiatry 52, 664–678.

Bowlby, J., 1988. A Secure Base: Clinical Applications of Attachment Theory. Routledge, London.

Bowlby, J., Robertson, J., 1952. A two-year-old goes to hospital: a scientific film. Proceedings of the Royal Society of Medicine 46, 425–427.

Bretherton, I., 2003. Mary Ainsworth: insightful observer and courageous theoretician. Portraits of Pioneers in Psychology 5, 317–331.

Bretherton, I., Ridgeway, D., Cassidy, J., 1990. Assessing internal working models of the attachment relationship: an attachment story completion task for 3-year-olds. In: Greenberg, M.T., Cicchetti, D., Cummings, E.M. (Eds.), Attachment in the Preschool Years: Theory, Research, and Intervention. University of Chicago Press, Chicago, IL, pp. 273–308.

Byng-Hall, J., 1987. Interview with John Bowlby, unedited version.

Cassidy, J., 1988. Child- mother attachment and the self in six-year-olds. Child Development 59, 121–134.

Cassidy, J., Shaver, P.R. (Eds.), 1999/2008. Handbook of Attachment: Theory, Research and Clinical Applications, second ed. Guilford, New York.

Cooper, G., Hoffman, K., Powell, B., Marvin, R., 2005. The circle of security intervention: differential diagnosis and differential treatment. In: Berlin, L.J., Ziv, Y., Amaya-Jeckson, L., Greenberg, M.T. (Eds.), Enhancing Early Attachments. Guilford, New York, pp. 127–151.

Dozier, M., Kaufman, J., Kobak, R.R., O'Connor, T.G., Sagi-Schwartz, A., Scott, S., Shauffer, C., Smetana, J., van IJzendoorn, M.H., Zeanah, C.H., 2014. Consensus statement on group care for children and adolescents. American Journal of Orthopsychiatry 84, 219–225.

Dozier, M., Lindhiem, O., Ackerman, J., 2005. Attachment and biobehavioral catch-up. In: Berlin, L., Ziv, Y., Amaya-Jackson, L., Greenberg, M.T. (Eds.), Enhancing Early Attachments. Guilford, New York, pp. 178–194.

Duniec, E., Raz, M., 2011. Vitamins for the soul: John Bowlby's thesis of maternal deprivation, biomedical metaphors and the deficiency model of disease. History of Psychiatry 22, 93–107.

De Wolff, M.S., van IJzendoorn, M.H., 1997. Sensitivity and attachment: a meta-analysis on parental antecedents of infant attachment. Child Development 68, 571–591.

Etzion-Carasso, E., Oppenheim, D., 2000. Open mother-preschooler communication: relations with early secure attachment. Attachment and Human Development 2, 347–370.

Fearon, R.M., Belsky, J., 2011. Infant–mother attachment and the growth of externalizing problems across the primary-school years. Journal of Child Psychology and Psychiatry 52, 782–791.

Furman, W., Simon, V.A., 2004. Concordance in attachment states of mind and styles with respect to fathers and mothers. Developmental Psychology 40, 1239–1247.

George, C., Kaplan, N., Main, M., 1984, 1985, 1996. Adult Attachment Interview, third ed. Department of Psychology, University of California, Berkeley. Unpublished Protocol.

Gini, M., Oppenheim, D., Sagi-Schwartz, A., 2007. Negotiation styles in mother – child narrative co-construction in middle childhood: associations with early attachment. International Journal of Behavioral Development 31, 149–160.

Goldberg, S., Grusec, J.E., Jenkins, J.M., 1999. Confidence in protection: arguments for a narrow definition of attachment. Journal of Family Psychology 13, 475–483.

Goldsmith, H.H., Alansky, J.A., 1987. Maternal and infant temperamental predictors of attachment: a meta-analytic review. Journal of Consulting and Clinical Psychology 55, 805–816.

Grossmann, K.E., Bretherton, I., Waters, E., Grossmann, K., 2013. Maternal sensitivity: observational studies honoring Mary Ainsworth's 100th year. Attachment and Human Development 15, 443–447.

Grossmann, K.E., Grossmann, K., 1990. The wider concept of attachment in cross-cultural research. Human Development 33, 31–47.

Grossmann, K., Grossmann, K.E., Kindler, H., 2005. Early care and the roots of attachment and partnership representations. In: Grossmann, K.E., Grossmann, K., Watters, E. (Eds.), Attachment from Infancy to Adulthood: The Major Longitudinal Studies. Guilford, New York, pp. 98–136.

Harlow, H.F., 1961. The development of affectional patterns in infant monkeys. In: Foss, B.M. (Ed.), Determinants of Infant Behaviour. Methuen, London, pp. 75–97.

Hesse, E., 2008. The adult attachment interview: protocol, method of analysis and empirical studies. In: Cassidy, J., Shaver, P.R. (Eds.), Handbook of Attachment: Theory, Research and Clinical Applications, second ed. Guilford, New York, pp. 552–598.

Hinde, R.A., 1982. Attachment: some conceptual and biological issues. In: Parkes, C., Stevenson-Hinde, J. (Eds.), The Place of Attachment in Human Behavior. Basic Books, New York, pp. 60–76.

Hinde, R.A., 1987. Individuals, Relationships and Culture: Links between Ethology and the Social Sciences. Cambridge University Press, New York.

Juffer, F., Bakermans-Kranenburg, M.J., Van IJzendoorn, M.H., 2008. Promoting Positive Parenting: An Attachment-Based Intervention. Erlbaum, London.

Klein, M., 1932. The Psycho-Analysis of Children. Hogarth Press, London.

Koren-Karie, N., Sagi-Schwartz, A., Joels, J., 2003. Absence of attachment representations (AAR) in the adult years: the emergence of a new AAI classification in catastrophically traumatized holocaust child survivors. Journal of Attachment and Human Development 5, 381–397.

Lorenz, K.Z., 1935. Der kumpan in der umwelt des vogels (The companion in the bird's world). Journal fur Ornithologie 83, 137–213 (Abbreviated English translation published 1937 in Auk, 54, 245–273).

Main, M., 1981. Avoidance in the service of proximity: a working paper. In: Immelmann, K., Barlow, G., Petrinovich, L., Main, M. (Eds.), Behavioral Development: The Bielefeld Interdisciplinary Project. Cambridge University Press, New York, pp. 651–693.

Main, M., 1990. Cross-cultural studies of attachment organization: recent studies changing methodologies, and the concept of conditional strategies. Human Development 33, 48–61.

Main, M., Hesse, E., 1990. Parents' unresolved traumatic experiences are related to infant disorganized attachment status: is frightened/frightening parental behavior the linking mechanism? In: Greenberg, M., Cicchetti, D., Cummings, E.M. (Eds.), Attachment in the Preschool Years. University of Chicago Press, Chicago, IL, pp. 161–182.

Main, M., Kaplan, N., Cassidy, J., 1985. Security in infancy, childhood, and adulthood: a move to the level of representation. In: Bretherton, I., Waters, E. (Eds.), Growing Points in Attachment Theory and Research, vol. 50. Monographs of the Society for Research in Child Development (1–2, Serial No. 209), pp. 66–104.

Main, M., Weston, D., 1981. The quality of the toddler's relationship to mother and to father: related to conflict behavior and the readiness to establish new relationships. Child Development 52, 932–940.

Oppenheim, D., Koren-Karie, N., 2009. Parents' insightfulness regarding their children's internal worlds: assessment, research, and clinical implications. In: Zeanah, C. (Ed.), Handbook of Infant Mental Health, third ed. Guilford, New York, pp. 266–280.

Oppenheim, D., Koren-Karie, N., Sagi-Schwartz, A., 2007. Emotion dialogues between mothers and children at 4.5 and 7.5 years: relations with children's attachment at 1 year. Child Development 7, 38–52.

Oppenheim, D., Sagi, A., Lamb, M.E., 1988. Infant-adult attachments on the kibbutz and their relation to socioemotional development four years later. Developmental Psychology 24, 427–433.

Parkes, C.M., 2009. Love and Loss: The Roots of Grief and Its Complication. Routledge, New York.

Riviere, J., 1927. Symposium on child analysis. The International Journal of Psycho-Analysis 8, 370–377.

Robertson, J., 1953. Some responses of young children to loss of maternal care. Nursing Care 49, 382–386.

Sagi-Schwartz, A., 2012. Children of war and peace: a human development perspective. Journal of Conflict Resolution 56, 933–951.

Sagi-Schwartz, A., Aviezer, O., 2005. Correlates of attachment to multiple caregivers in kibbutz children from birth to emerging adulthood: the Haifa longitudinal study. In: Grossmann, K.E., Grossmann, K., Watters, E. (Eds.), Attachment from Infancy to Adulthood: The Major Longitudinal Studies. Guilford, New York, pp. 165–197.

Salter, M.D., 1940. An Evaluation of Adjustment Based upon the Concept of Security. Child Development Series. The University of Toronto Press, Toronto.

Sadler, L.S., Slade, A., Close, N., Webb, D.L., Simpson, T., Fennie, K., Mayes, L.C., 2013. Minding the baby: enhancing reflectiveness to improve early health and relationship outcomes in an interdisciplinary home-visiting program. Infant Mental Health Journal 34, 391–405.

Spitz, R.A., 1946. Anaclitic depression. Psychoanalytic Study of the Child 2, 313–342.

Thompson, R.A., 2008. Early attachment and later development: familiar questions, new answers. In: Cassidy, J., Shaver, P.R. (Eds.), Handbook of Attachment: Theory, Research and Clinical Applications, second ed. Guilford, New York, pp. 348–365.

Van Dijken, S., 1998. John Bowlby, His Early Life: A Biographical Journey into the Roots of Attachment Theory. Free Association Books, New York.

Van IJzendoorn, M.H., 1995. Adult attachment representations, parental responsiveness, and infant attachment. A meta-analysis on the predictive validity of the adult attachment interview. Psychological Bulletin 117, 387–403.

Van IJzendoorn, M.H., Sagi-Schwartz, A., 2008. Cross-cultural patterns of attachment: universal and contextual dimensions. In: Cassidy, J., Shaver, P.R. (Eds.), Handbook of Attachment: Theory, Research and Clinical Applications, second ed. Guilford, New York, pp. 880–905.

Van IJzendoorn, M.H., Schuengel, C., Bakermans- Kranenburg, M.J., 1999. Disorganized attachment in early childhood: meta-analysis of precursors, concomitants, and sequelae. Development and Psychopathology 11, 225–250.

Waters, E., Deane, K.E., 1985. Defining and assessing individual differences in attachment relationships: Q-methodology and the organization of behavior in infancy and early childhood. Monographs of the Society for Research in Child Development, 41–65.

Waters, E., Weinfield, N.S., Hamilton, C.E., 2000. The stability of attachment security from infancy to adolescence and early adulthood: general discussion. Child Development 71, 703–706.

Integration and Multiculturalism: Focus on Western Europe

Tariq Modood, University of Bristol, Bristol, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Examining the concept of social integration, I challenge the view that multiculturalism emphasizes difference at the expense of commonality, separatism rather than mixing, group rather than national identities, relativism rather than a defense of democratic values. I argue that multiculturalism is a mode of integration, which can be contrasted with other modes such as assimilation, individualist-integration, and cosmopolitanism, and like the others it is based on the core democratic values of liberty, equality, and fraternity/unity. While multiculturalism involves a respect for minority identities, this is not at the expense of national identities, as long as the latter are remade to include minority as well as majority identities. The discussion relies mainly on examples from Western Europe.

There is much confusion about what multiculturalism is. This situation is partly because 'multiculturalism' is too often defined by its critics, whose sole purpose is to create a straw man to knock down. But it is also because both its critics and some of its defenders falsely oppose multiculturalism with integration; and the confusion also partly stems from the fact that there is more than one form of multiculturalism and they relate to integration in different ways. I will clarify the key terms of assimilation, integration, cosmopolitanism, and multiculturalism. My argument is that discourses of integration and multiculturalism are exercises in conceptualizing post-immigration difference or identities and as such operate at three distinct levels: as an (implicit) sociology; as a political response; and as a vision of what is the whole in which difference is to be integrated.

Integration

The need for integration arises when an established society is faced by some people who are perceived and treated unfavorably from standard members (and typically who also perceive themselves as 'different' though not necessarily in a negative way). This may relate to different areas or sectors of society and policy, such as employment, education, housing, and so on. Someone is integrated into the labor market, for example, when she or he is able to enjoy equality of opportunity in accessing jobs and careers, including accessing the education and training necessary to compete for such jobs. Additionally, integration would require that the labor market not be segmented into different parts with radically different monetary rewards and working conditions for those with broadly similar qualifications and experience. The labor market is particularly relevant, where the segmentation is not, formally or informally, based on criteria such as race, ethnicity, religion, and so on; namely the categories of 'difference'. Avoidance of segmentation of this sort can be more generally applied to other sectors such as education and housing. A core idea of integration is equality of opportunity in an unsegmented society where no channeling into or away from a sector of society takes place based on criteria such as race and ethnicity. Integration has a number of components based on opportunities to participate that are context specific and need to be secured by law and policy initiatives. However, integration also has a subjective and symbolic dimension. This dimension will have some context-specific features, but also has a more general or macro character. It includes how the rest of the country perceives a minority and how members of a minority perceive their relationship to society as a whole. Sectoral integration, however, is not full integration without some degree of subjective identification with the society or country as a whole. This requires a feeling of belonging, and the acceptance by the majority that the minority is a full member of society. This is what the Commission on Multi-Ethnic Britain called 'a sense of belonging' (CMEB, 2000: Introduction).

An individual may choose to integrate or not, and may be given opportunities to participate or not. My interest here is not on individual choices and opportunities themselves but rather the level of groups or society as a whole. A sense of belonging is dependent on how others perceive and treat you, not just as an individual but also as a member of a racial group or ethno-religious community. Each policy area (e.g., whether it is issues of qualification levels or residential segregation) will have its own imperatives and difficulties. (Different groups may integrate to different degrees across sectors. For example, Jews in Britain are highly integrated in relation to employment but are the most segregated religious minority (Peach, 2007).) There is, however, also a general understanding that we have, as members of society, about what our society is and what it is to be a member – a macro-symbolic conception of society and of integration. Such general conceptions inform popular understanding as well as political ideas and the general terms of policy paradigms. Hence, it has been said by a Commission on these topics in Quebec, "the symbolic framework of integration (identity, religion, perception of the Other, collective memory, and so on) is no less important than its functional or material framework" (Bouchard and Taylor, 2008). Indeed, the sense of 'crisis' about multiculturalism and integration is operating at this macro-symbolic level. This is evident when one considers how few in any country are the policies that could be said to be about integration or how small the funds involved compared to the headline importance that the issues regularly achieve. In thinking about policy paradigms of a general ethos or

orientation at a national level it is therefore important to engage at this macro-symbolic level.

I consider this larger, macro-symbolic sense of integration and implied policy paradigms in terms of four modes of integration (summarized in **Table 1**), namely, assimilation, individualist-integration, and two versions of multiculturalism, one of which I will call cosmopolitanism. (The concern here is not primarily in relation to socioeconomic integration, for which see Loury et al. (2005) and Heath and Cheung (2007).) Each offers their own distinctive take on freedom, equality, and civic unity (what might be called 'fraternity' or solidarity), the core values of European democracy. (In doing so I follow Charles Taylor's treatment of the concept of secularism (Taylor, 2009), though without claiming that he would use it as I do in relation to integration.) Different interpretations and prioritizations of these concepts suggest embryonic paradigms. The issue or 'problem' these paradigms are addressing is post-immigration 'difference' (Modood, 2007). 'Difference' primarily refers to how people are identified: how they identify themselves (e.g., as 'white', 'black', 'Chinese', 'Muslim,' etc.), how they identify others, and how they are identified by others.

These identities fall (not necessarily unambiguously or discretely) within the fields of 'race', ethnicity, religion, culture, and nationality as various forms of difference. The problem, then, is how to integrate difference, by which I mean the process whereby difference ceases to be problematic. It is commonly said that we must distinguish between multiculturalism as social description and multiculturalism as political understanding of those social facts (e.g., Malik, 2010). Interestingly, the same could be, but is rarely said about class/Class, nation/Nation, or gender/Gender too are cases which could be distinguished in terms of social description, on the one hand, and as political understanding of those facts on the other hand. Yet, this is rarely done and the distinction is specially applied to multiculturalism. The distinction, however, is of limited value because social description and sociological concepts are not normally politically neutral. The concepts I analyze below, then, are normative and policy oriented but they presuppose an understanding of what the social phenomenon is that demands a political response. The problem, then, is how to integrate difference or to put it another way, the processes whereby difference ceases to be problematic. I shall consider four modes of integration (summarized in **Table 1**).

Modes of Integration

Assimilation, or one-way integration, is where the processes affecting change and the relationship between social groups are largely one way with the result being one where the newcomers do little to disturb the society they are settling in and become as much like their new compatriots as possible. (When the US sociologists use the term 'assimilation', they usually mean what is meant by integration in the UK, as in the 'segmented assimilation' proposed by Portes and Zhou (1993).) By erasing difference it is may be thought that the occasions for discrimination and conflict are not allowed to take root. While 'assimilation' as a term has been dropped in favor of 'integration', even today, some European politicians use the term 'integration' they consciously or not are referring to what I have defined as assimilation (Fekete, 2008). The use of these terms

Table 1 Four modes of integration[a]

	Assimilation	*Individualist-Integration*	*Cosmopolitanism*	*Multiculturalism*
Objects of policy	Individuals and groups marked by 'difference'.	Individuals marked by 'difference', especially their treatment by discriminatory practices of state and civil society.	Individuals marked by 'difference', especially their treatment by discriminatory practices of state and civil society, and societal ideas, especially of 'us' and 'them'.	Individuals and groups marked by 'difference', especially their treatment by discriminatory practices of state and civil society, and societal ideas, especially of 'us' and 'them'.
Liberty	Minorities must be encouraged to conform to the dominant cultural pattern.	Minorities are free to assimilate or cultivate their identities in private but are discouraged from thinking of themselves as minority, but rather as individuals.	Neither minority nor majority individuals should think of themselves as belonging to a single identity but be free to mix and match.	Members of minorities should be free to assimilate, to mix and match or to cultivate group membership in proportions of their own choice.
Equality	Presence of difference provokes discrimination and so is to be avoided.	Discriminatory treatment must be actively eliminated so everyone is treated as an individual and not on the basis of difference.	Antidiscrimination must be accompanied by the dethroning of the dominant culture.	In addition to antidiscrimination the public sphere must accommodate the presence of new group identities and norms.
Fraternity	A strong, homogeneous national identity.	Absence of discrimination and nurturing of individual autonomy within a national/European, liberal democratic citizenship.	People should be free to unite across communal and national boundaries and should think of themselves as global citizens.	Citizenship and national identity must be remade to include group identities that are important to minorities as well as majorities; the relationship between groups should be dialogical rather than one of domination or uniformity.

[a]In all cases it is assumed that a backdrop of liberal democratic rights and values are operative to a large degree and what is highlighted here is in addition or in interaction with them.

in public discourse must not be taken therefore at their face value but critically inspected.

In the three nonassimilative modes of integration, processes of social interaction are seen as two way, where members of the majority community as well as immigrants and ethnic minorities are required to participate; so the latter cannot alone be blamed for failing to, or not trying to, integrate. The established society is the site of institutions – including employers, civil society and the state – in which integration has to take place, and accordingly they must take the lead. The new (prospective) citizens' rights and opportunities must be made effective through antidiscrimination laws and policies. We need, however, to distinguish between *individualist-integration* and *multiculturalism*. Individualist-integration sees the need for institutional adjustments in relation to migrants or minorities but sees these as only individual claimants and bearers of rights as equal citizens (Barry, 2001). Minority communities may exist as private associations but are not recognized or supported in the public sphere.

Multiculturalism is where processes of integration are two way and as involving groups as well as individuals. Moreover, the processes may work differently for different groups (CMEB, 2000; Parekh, 2000; Modood, 2007). In this understanding, each minority group is distinctive, and thus integration cannot consist of applying a single template to them (hence the 'multi' in 'multiculturalism'). The 'culturalism' – by no means a happy term either in relation to 'culture' or 'ism' – refers to that the groups in question are likely not just to be marked by newness or phenotype or socioeconomic location but by certain forms of group identities. The groups are not merely 'newcomers' or 'black' or 'inner-city dwellers' but have a sense of subjective identification, a sense of 'we'. The integration of groups is in addition to, not as an alternative to, the integration of individuals, antidiscrimination measures, and a robust framework of individual rights.

Multiculturalism, like most concepts, takes different form in different contexts and at different times. For example, it has been differently understood in the Netherlands than in Britain (Joppke, 2004; Koopmans et al., 2005) and in Quebec compared to in Anglophone Canada (Bouchard and Taylor, 2008: chapter 6). The meaning of any mode of integration is subject to debate and contestation and its policy originators may start with one meaning, as Roy Jenkins did in relation to race and culture, and others may push it or extend it in other directions by, say, making religion central, as Muslims in Britain have done (Modood, 2005).

Among what is central to multiculturalism is the concept of equality, as indeed it is to other conceptions of integration. The key difference between individualist-integration and multiculturalism is that the concepts of group and of 'multi' are essential to the latter. Postimmigration minorities are groups differentiated from the majority society or the norm in society by two kinds of processes. On the one hand, by negative 'difference' with alienness, inferiorization, stigmatization, stereotyping, exclusion, discrimination, racism, and so on. On the other hand, by the senses of identity that groups so perceived have of themselves. The two together are the key data for multiculturalism. The differences at issue are those perceived both by outsiders or group members – from the outside in and from the inside out – and constitute not just some form of distinctness, but a form of alienness or inferiority that diminishes or makes difficult equal membership in the wider society or polity.

Multiculturalism has recently been defined as "where ethnocultural-religious minorities are, or are thought of, as rather distinct communities, and where public policy encourages this distinctiveness" (Emmerson, 2011). This, however, is only a third of it. Multiculturalism allows those who wish to encourage such distinctiveness to do so; but it also seeks forms of social unity that are compatible with this, what Hartmann and Gerteis (2005) call 'new conceptions of solidarity', grounded in a concept of equality (Bouchard and Taylor, 2008). Each mode of integration must be understood in terms of its interpretation of free choice, equality, and fraternity. Characterizations of multiculturalism that subtract its emphasis on unity are extremely common but incomplete.

Further Unpacking Integration and Multiculturalism

Multicultural accommodation of minorities, then, is different from individualist-integration because they explicitly recognize the social reality of groups, not just of individuals and organizations. There may, however, be considerable complexity about what is meant by social reality of groups or 'groupness' here. Ideas of groups as discrete, homogeneous, unchanging, bounded populations are not realistic when we are thinking of multicultural recognition (Modood, 2007: pp. 93–97). This discussion leads us to cosmopolitanism.

Cosmopolitanism emerges by accepting the concept of difference while critiquing or dissolving the concept of groups (Waldron, 1991). (Here I do not mean the ethical view that one should be a citizen of the world, rather I am characterizing a mode of integration within a country that emphasizes a mixing of people from all over the world as in the expression 'London is a cosmopolitan city'. British sociologists sometimes use the term 'multiculture' but this clearly has not carried over into public discourse. It has been suggested to me that the term 'interculturalism' best fits here but the place where it is most used in relation to national politics, Quebec, it is closer to what here I call 'individualist-integration'. More generally, it is not clear that 'interculturalism' includes anything that is not or cannot be included in multiculturalism (see Meer and Modood, 2012). I did also consider the term 'diversity' but it is either too descriptive and generic and does not pick out a mode of integration, or has been appropriated as 'diversity management' by human resource professionals.) Disagreement about the extent to which postimmigration groups exist and/or ought to exist and be given political status means that there are two kinds of multiculturalism (Modood, 1998; Meer and Modood, 2009). While in public discourse, as well as in academia, one or both are referred to as multiculturalism, I will reserve the term 'multiculturalism' for the sociological and political position in which groups are a critical feature. (This is how the term has been used by the leading political theorists such as Taylor (1994), Kymlicka (1995), and Parekh (2000) and, by the Canadian government; it is also consistent with CMEB (2000) and other exponents of multiculturalism – see Modood (2007: pp. 14–20) for details.)

Cosmopolitanism positively valorizes 'difference' (or pragmatically accepts it) but denies that groups exist or, alternatively, exist but should not be politically recognized. The contention is that in the early stages of migration and settlement, especially in the context of a legacy of racism, colonialism, and European supremacism, forms of social exclusion created or reinforced certain forms of groupness such as white and black. However, as a result of social mixing, cultural sharing, and globalization in which dominant identities of modernity (such as of race and nation) are dissolving, people have much more fluid and multiple identities, combine them in individual ways and use them in context-sensitive ways (Hall, 1992a). For example, the ways that Caribbean-origin Britons have socially blended into a 'multiculture' and have sought conviviality and sociability rather than separate communities may perhaps not be fully captured as a form of individualistic integration (Gilroy, 2000). While remaining economically marginal and overrepresented in relation to the social problems associated with deprived inner city areas, they have become a feature of popular culture in terms of music, dance, youth styles, and sport (Hall, 1998). To the extent that football teams, Olympiads, and television programmes such as *The X Factor* are central to popular and national identities, Caribbean-origin people are placed at the center of British national imaginaries.

Moreover, Britain and most other countries in Western Europe have recently experienced and are experiencing a new wave of immigration and will continue to do so, including from within the European Union. Given the diversity of the locations from where migrants are coming, the result, it is argued, is not communities, but a churning mass of languages, ethnicities, and religions, all cutting across each other and creating a 'super diversity' (Vertovec, 2007). This may be setting a pattern for the future, and it may be allied to a further argument that globalization, migration, and telecommunications have created populations dispersed across countries that interact more with each other, and have a greater sense of loyalty to each other, than they might to their fellow citizens.

In what ways does cosmopolitanism go beyond individualist-integration? Its distinctive ethos is that we should value diversity and create the conditions in which it is individually chosen. We should oppose all forms of imposition of group identities on individuals and therefore the ideas, images, and prejudices by which individuals are inferiorized or portrayed as threatening and so excluded from full membership of society, and we should not require assimilation or conformity to dominant group norms. Yet, a requirement of communal membership can also be oppressive of individuals and their life chances (Appiah, 1994). Inherited or ascribed identities that slot people into pigeonholes not of their choosing should be refused. These ascribed identities not only reduce the options of the kind of person one can be but divide society up into antagonistic groups. (British exponents of this view tend, however, to put some communal identities in a normative, privileged position. This particularly applies to political blackness and to some extent to noncultural and nonreligious political identities generally (Modood, 1994).) Cosmopolitanism, then, is a conception of multiculturalism as maximum freedom, for minority as well as majority individuals, to mix with, borrow, and learn from all (whether they are of one group or another) so individual identities are personal amalgams of bits from various groups and heritages, and there is no one dominant social identity to which all must conform. The result will be a society composed of a blend of cultures, a 'multiculture'.

While this is an attractive image of contemporary society and blends easily with the ideas of liberal democracy, it has only a partial fit with London today, let alone many parts of Britain and continental Europe. In some towns and cities, such as in northern England, there is not a diversity of ethnic groups but often just two (e.g., Asian Muslims and whites) and minority individuals do not float across identities, mixing, and matching, but have a strong attachment to a one or few identities. For example, most British Muslims seem to think of themselves in terms of 'Muslim' and/or 'British' (usually both) (Travis, 2002). The fact of super diversity is emerging alongside rather than displacing the fact of settled, especially postcolonial, communities, who have a particular historical relationship with Britain, and the political significance of such communities. Similarly, there are other communities in other European countries with their own historical significance such as Maghrebians (migrants from the northern countries of Africa) in France and the Turks in Germany. Moreover, some groups continue to be much larger than others, and stand out as groups – in their own eyes and those of others. Muslims, for example, seem to be in this category across much of Western Europe regardless of the degree of conviviality or super diversity that might be present.

That is not to say that such minority identities are exclusive. Successive surveys have shown that most Muslims in Britain strongly identify with being Muslim but the majority also identify as British. Indeed, they are more likely to identify with 'British' and say they have trust in key British institutions than non-Muslims. Several researchers (Heath and Roberts, 2008; Gallup, 2009) found the same in Germany, albeit less so in France though Pew (2006) found much higher levels of national identification in France than other western European countries.

Postimmigration hyphenated identities, such as British-Indian, have become as commonplace in Britain as they have been in the United States for decades. Similarly, diasporic links as described above certainly exist, and are likely to increase, but the net result is not an inevitable erosion of national citizenship – British African-Caribbeans and South Asians have families in their countries of origin and in the US and Canada, but there is little evidence that most branches of those families do not feel British, American, or Canadian (e.g., Modood, 2007; Bloemraad, 2006). Indeed, studies show that the more multiculturalist countries achieve higher levels of national identification among migrants and the second generation (Esses et al., 2006; Wright and Blooemrad, 2012).

An important point of difference, then, between the concepts of individualist-integration and multiculturalism proper is that for the latter, the postimmigration minorities are not of one kind but are a 'multi'. For example, some people will identify with a color identity like 'black' but there will be others for whom national origin identities (like 'Turkish'), or a regional heritage (like 'Berber'), or a religious identity (like 'Sikh') may be much more meaningful, expressing forms of community and ethnic pride and struggling for recognition and

inclusion. And of course these minority identities will interact with wider, societal identities – 'woman', 'working class', 'Londoner', 'British' – in differing ways, expressing the different experiences, locations, and aspirations of different groups. So, both the alternative models of multiculturalism as cosmopolitanism and what may be called ethno-religious communitarianism have some grounding and meet the political aspirations of some minority groups. Neither works as a comprehensive sociological or political model and they should be viewed as complementary (Modood, 1998; CMEB, 2000; Modood and Dobbernack, 2011). Moreover, while recognition of ethnic or religious groups may have a legal dimension, for the most part, it will be at the level of civic consultations, political participation, institutional policies (e.g., schools and hospitals), discursive representations, especially in relation to the changing discourses of societal unity or national identity, and their remaking.

Regardless of the extent to which recognition of minority identities in this way is formal or informal, led by the state or the semiautonomous institutions of civil society, it does not challenge, let alone displace, individual rights and the shared dimensions of citizenship. There may however be genuine concern that some groups at a particular time and in some areas are becoming too inward looking. Where the concern is primarily about a lack of positive mixing and interaction between groups at a local level, community cohesion measures – for example, a Christian school offering places to non-Christians or twinning with a non-Christian school – may be an appropriate response (Cantle, 2001). Where the concern is about self-conceptions and discourses more generally, the issue will be about the national or societal identity. While such inwardness has never been part of any theory or policy of multiculturalism, it is clear that it is a fundamental anxiety of the critics of multiculturalism, many of whom go as far as defining multiculturalism in terms of separatism. (A review of the American social science literature found that "[t]he most common conception of multiculturalism in both scholarly circles and popular discourse is a negative one, having to do with what multiculturalism is not or what it stands in opposition to. Multiculturalism, in this usage, represents heterogeneity as opposed to homogeneity, diversity as a counterpoint to unity" (Hartmann and Gerteis, 2005: pp. 219). They found that if they looked at exponents, as opposed to critics, of multiculturalism, such simplistic dichotomies were unsustainable and they concluded: "multiculturalism is best understood as a critical-theoretical project, an exercise in cultivating new conceptions of solidarity in the context of dealing with the realities of pervasive and increasing diversity in contemporary societies" (pp. 221–222).) It is therefore important to emphasize that accommodation of ethno-religious communities is a mode of integration, and that it, no less than hostility to minorities or other modes of integration, should be examined as a possible contributory cause of exclusion and integration (Banting and Kymlicka, 2008).

Multiculturalism Is Not Dead

This unpacking of what I mean by 'multiculturalism' is also helpful in understanding those who say that multiculturalism has failed (Weldon, 1989; see Presseurop (2010) for Angela Merkel's speech on the failure of *multikulti*) or that multiculturalism is dead (Cameron, 2011). They may mean to endorse assimilation, individualistic integration, or cosmopolitanism. At the same time they are acknowledging and possibly reinforcing the sociological reality of group difference because their lament is that some groups (especially Muslims) are clearly visible as distinct groups when they should not be; they attribute this fact to a separatist tendency in the groups, encouraged by allegedly multiculturalist policies. The accusatory discourse that 'some groups are not integrating' may actually be reinforcing group identities and therefore contributing to the social conditions that gives multiculturalism a sociological pertinence. Moreover, it is not just at the level of sociology that antimulticulturalists may find themselves using multiculturalist ideas; even while deploying an antimulticulturalist discourse politicians may enact multiculturalist policies. Politicians may, for example, continue with group consultations, representation, and accommodation. The latter have actually increased. The British government has found it necessary to increase the scale and level of consultations with Muslims in Britain since 9/11, and, dissatisfied with existing organizations, has sought to increase the number of organized interlocutors and the channels of communication. Avowedly antimulticulturalist countries and governments have worked to increase corporatism in practice, for example with Nicholas Sarkozy's creation of the Conseil Francais du Culte Musulman in 2003 to represent all Muslims to the French government in matters of worship and ritual. Similarly, Islamkonferenz was established in Germany in 2005 as an exploratory body, but has an extensive political agenda. These bodies are partly top-down efforts to control Muslims or to channel them into certain formations and away from others. Nevertheless, such institutional processes cannot be understood within the conceptual framework of assimilation, individualist-integration, or cosmopolitanism. These top-down bodies are normatively less than the best of multiculturalism but they clearly are deploying a multiculturalist mode of integration and their presence offers the possibility of movement toward a more egalitarian multiculturalism.

The analytical framework offered helps us also to understand those who say they welcome diversity but seem to be in agreement with critics of political multiculturalism. Critics point to the public assertion of strong group identities to achieve certain policies and/or to demand differential treatment. Some respond to this criticism by noting that multiculturalism is working in their neighborhoods. Such proponents say that neighborhoods are multiethnic and people do not just live peaceably side by side, but mix freely and value mixing. Yet such views do not imply support for strong group identities and related policies; on the contrary, their success may be seen to be dependent on the absence of the latter. (Hence the irony that antimulticulturalists like President Sarkozy are trying to create corporate representations for Muslims in France; while prodiversity authors call for the cessation of government meetings with Muslim community leaders (Sen, 2006; Malik, 2011).) Hence, while this is a reasonable response in its own terms it does not counter the above criticism of multiculturalism and in fact may share it. It is the group-based

multiculturalism that has become unpopular and is what critics have in mind so that criticism is left standing. The weakness of the rebuttal though is obscured by the fact that advocates of neighborhood 'multiculturalism' are in practice referring to what I call 'cosmopolitanism'. Another way of putting this is to say that cosmopolitans may be as critical of an ethno-religious communitarianism as antimulticulturalists.

Can integration of all postimmigration formations, however, be achieved without group-based multiculturalism (Modood, 1998, 2007)? Moreover, a group-based multiculturalism, where group membership is voluntary, may be part of the future in an unintended way as it is highly compatible with the Prime Minister Cameron's vision of a 'Big Society' in which civil society associations based on locality and faith, including interfaith groups, take over some responsibilities currently undertaken by state agencies. A flagship policy of the Big Society agenda is the state funding to create new community-based nonstate schools called 'free schools'. Over a quarter of these are led by religious groups; those that started in September 2011 included two Jewish, a Hindu, and a Sikh school (Vasagar, 2012). Of the 102 schools approved for funding in July 2012, three were Jewish, one was Hindu, four were Sikh, and four were Muslim (BHA, 2012). It is difficult to see how the new Big Society is a break with what is rejected as 'state multiculturalism' (Cameron, 2011). The same trend is found in France, where three Muslim schools have joined the many thousands (mainly Catholic) of state-supported religious schools (Akan, 2009: pp. 246–247), and in Germany, where there are no state funded religious schools but where Islam is increasingly incorporated in instruction (The Local, 2011; see also DIK, 2009).

The analysis offered here of related macro-symbolic ideas and policy paradigms, each of which consists of a model of society and normative political ideas, includes a sense of unity or fraternity. Modes of integration are not just about sociology or politics but include ideas, however inchoate, of ourselves as a social unity (as displayed at the bottom of **Table 1**). For assimilationists, this consists of a strong, homogeneous national identity. Individualist-integration emphasizes the liberal and democratic character of the national polity. Cosmopolitanism is uneasy with the national, an identity that demands allegiance from all citizens, while creating boundaries between ourselves and the rest of the world. With multiculturalism comes a positive vision of the whole remade so as to include the previously excluded or marginalized on the basis of equality and sense of belonging. It is at this level that we may fully speak of multicultural integration or multicultural citizenship (Taylor, 1994; Parekh, 2000; Modood, 2007). This third level of multiculturalism, incorporating the sociological fact of diversity, groupness, and exclusion, but going beyond individual rights and political accommodation, is perhaps the level that has been least emphasized. Or at least that is how it seems to many whose understanding of multiculturalism, sometimes polemical but sometimes sincere, is that multiculturalism is about encouraging minority difference without a counterbalancing emphasis on cross-cutting commonalities and a vision of a greater good. This has led many commentators and politicians to talk of multiculturalism as divisive and productive of segregation.

Theorists of multiculturalism (Taylor, 1994; Parekh, 2000), related policy documents (CMEB, 2000), and national enactments (citation), are universally regarded as pioneers and exemplars of state multiculturalism and all appealed to and built on an idea of national citizenship. From a multiculturalist point of view, though not from that of its critics, the recent emphasis on cohesion and citizenship, what has been called 'the civic turn' (Mouritsen, 2008), is a necessary rebalancing of the political multiculturalism of the 1990s. This form of multiculturalism largely took the form of accommodation of groups while being ambivalent about national identity (Meer and Modood, 2009). (In the 1990s cosmopolitanism and multiculturalism in Britain began to be linked to a national identity and its modernization, to, for example, 'Cool Britannia' and 'rebranding Britain' (Leonard, 1997) but others welcomed globalization as an era of the 'postnational' (Hall, 1992b; Soysal, 1994).) This does not invalidate the analysis that integration without institutional accommodation is unlikely to be successful. Indeed, for multiculturalists, a renewed national identity has to be distinctly plural and hospitable to the minority identities. It involves 'rethinking the national story' with the minorities as important characters; not obscuring difference but weaving it into a common identity that all can see themselves in and giving all a sense of belonging to each other (CMEB, 2000: pp. 54–56; Modood, 2007: pp. 145–154).

Conclusion

It may be the case that all the attempted models of integration, especially national models, are in crisis. Certainly many perceive each as having failed and none have brought about equality of opportunity measured in terms of socioeconomic indicators, nor full acceptance of minorities as full cocitizens and conationals. We can, however, have a better sense of what the issues are and so what needs to be done if, firstly, we recognize that discourses of integration and multiculturalism are exercises in conceptualizing postimmigration difference and as such operate at three distinct levels: as an (implicit) sociology, as a political response, and as a vision of what is the whole in which difference is to be integrated. Depending upon the sociology in question, certain political responses are possible or not, or, more reasonable or less. The sociological and political assumptions are thus mutually dependent.

Secondly, I have offered a framework in which four distinct political responses (assimilation, individualist-integration, cosmopolitanism, and multiculturalism) can be understood in terms of liberty, equality, and fraternity/unity, thereby bringing out what is of merit in each mode as well as the ways in which it is deficient (summarized in **Table 1**). They illuminate each other and each successive position attempts to include what is thought to be missing from the predecessor. Each position can in principle be attractive from different perspectives: some ethnic minorities may wish to assimilate, some to have the equal rights of integrated citizens, some to maintain the cultural differences of their group identities, and some to be free to choose

cosmopolitan mixed identities suiting the roles they take on in a more diverse society. Equally, host communities may look on different groups of migrants in all these different ways. Assimilation may be more appropriate in terms of national language acquisition before naturalization; individualist-integration may provide the model for nondiscrimination in the labor market; yet multiculturalism may be the basis for supplementing electoral representation (where minorities are underrepresented) and in creating new attitudes of inclusivity and in rethinking national identities. Each approach has a particular conception of equal citizenship but the value of each can only be realized if it is not imposed but is the preferred choice of minority individuals and groups, who, of course, being a 'multi', are bound to choose differently. Thus no singular model is likely to be suitable for all groups. To have a reasonable chance of integrating the maximum number of members of minorities, none of these political responses should be dismissed. Communitarian multiculturalism, then, has to be understood in terms of meeting the liberty, equality, and unity requirements that the other modes of integration fall short on. Yet, perhaps the ultimate meaning of multiculturalism, lets call it, multicultural citizenship, is not as one mode of integration but as the perspective that allows all four modes of integration their due.

Ethno-religious communitarianism may currently be viewed as undesirable by European publics and policymakers but given how central Muslims have become to the prospects of integration on a number of fronts, it is unlikely that integration can be achieved without some element of this approach (Levey and Modood, 2009; Triandafyllidou et al., 2012). Perceptions of Muslims as groups, by themselves and by non-Muslim majorities, are hardening. So the key question is whether they are to be stigmatized as outsiders or recognized as integral to the polity.

Finally, we must not overlook the third analytical level, which in many ways is not primarily about minorities but about the majority. The enlargement, hyphenation, and internal pluralizing of national identities is essential to an integration in which all citizens have not just rights but a sense of belonging to the whole as well as to their own 'little platoon' (Burke, 1982: p. 135). ("To be attached to the subdivision, to love the *little platoon* we belong to in … we proceed toward a love to our country, and to mankind" (Burke, 1982: p. 135).)

Acknowledgments

This article is based on my contribution to the British Academy 'New Paradigms in Public Policy' project and the publications that have come out of it, namely my Post-Immigration "Difference" and Integration: The Case of Muslims in Western Europe, The British Academy, 2012 and a modified version of it in Taylor-Gooby, P., (Ed.), 2013. New Paradigms in Public Policy, Oxford University Press.

See also: Communitarianism; Cosmopolitanism; Discrimination and the Law; Ethnic Conflict, Geography of; European Union Law; Legal Pluralism; Migrants and Migration in Europe; Migration: Anthropological Perspectives; Migration: Cultural Aspects; Multiculturalism, Anthropology of; Prejudice in Society: Psychological Perspectives; Prejudice in Society: Sociological Perspectives; Race and Racism in the Twenty-First Century; Religion: Nationalism and Identity; Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns; Xenophobia and Nativism.

Bibliography

Akan, M., 2009. Laïcité and multiculturalism: the Stasi Report in context. British Journal of Sociology 60 (2), 237–256.

Appiah, K.A., 1994. Identity, authenticity, survival: multicultural societies and social reproduction. In: Gutmann, A. (Ed.), Multiculturalism: Examining the Politics of Recognition. Princeton University Press, Princeton, NJ, pp. 149–164.

Banting, K., Kymlicka, W., 2008. Multiculturalism and the Welfare State: Recognition and Redistribution in Contemporary Democracies. Oxford University Press, Oxford.

Barry, B., 2001. Culture and Equality: An Egalitarian Critique of Multiculturalism. Polity, Cambridge.

BHA, 2012. BHA: Approved 2013 Free Schools Include First 'faith' Special, Alternative Provision Schools. http://www.politics.co.uk/opinion-formers/bha-british-humanist-association/article/bha-approved-2013-free-schools-include-first-faith-special-a (last accessed 12.07.12.).

Bloemraad, I., 2006. Becoming a Citizen: Incorporating Immigrants and Refugees in the United States and Canada. University of California Press, Berkeley and Los Angeles.

Bouchard, G., Taylor, C., 2008. Building the Future: A Time for Reconciliation, Consultation Commission on Accommodation Practices Related to Cultural Differences. Gouvernement du Québec, Quebec.

Burke, E., 1982. Reflections on the Revolution in France. Penguin, Harmondsworth.

Cameron, D., 2011. PM's Speech at Munich Security Conference (online). http://www.number10.gov.uk/news/speeches-and-transcripts/2011/02/pms-speech-at-munich-security-conference-60293 (last accessed on 29.03.11.).

Cantle, T., 2001. Community Cohesion: A Report of the Independent Review Team. Home Office, London.

Commission on Multi-Ethnic Britain (CMEB), 2000. The Future of Multi-Ethnic Britain: Report of the Commission the Future of Multi-Ethnic Britain. Runnymede Trust, London.

DIK (Deustsche Islam Konfrenz), 2009. Islamic Religious Education Trials in Schools. http://www.deutsche-islam-konferenz.de/DIK/EN/DIK/StandpunkteEgebnisse/Personal/ReligionBildung/Schulversuche/schulversuche-node.html (last accessed 16.02.09.).

Emmerson, M., 2011. Interculturalism: Europe and Its Muslims, in Search of Sound Societal Models. Centre for European Policy Studies, Brussels.

Esses, V., Wagner, U., Wolf, C., Preiser, M., Wilbur, C.J., 2006. Perceptions of national identity and attitudes toward immigrants and immigration in Canada and Germany. International Journal of Intercultural Relations 30, 653–669.

Fekete, L., 2008. Integration, Islamophobia and Civil Rights in Europe. Institute of Race Relations, London.

Fox, J., Miller-Idriss, C., 2008. Everyday nationhood. Ethnicities 8 (4), 536–563.

Gallup, 2009. The Gallup Coexist Project: Muslim West Facts Project. http://www.euro-islam.info/wp-content/uploads/pdfs/gallup_coexist_2009_interfaith_relations_uk_france_germany.pdf (accessed 28.03.11.).

Gilroy, P., 2000. Between Camps: Race, Identity and Nationalism at the End of the Colour Line. Allen Lane, London.

Glazer, N., Moynihan, D.P., 1963. Beyond the Melting Pot: The Negroes, Puerto Ricans, Jews, Italians and Irish of New York City. The MIT Press and Harvard University Press, Cambridge, MA.

Hall, S., 1992a. New ethnicities. In: Donald, J., Rattansi, A. (Eds.), 'Race', Culture and Difference. Sage, London, pp. 252–259.

Hall, S., 1992b. The question of cultural identity. In: Hall, S., McGrew, T. (Eds.), Modernity and its Futures. Polity Press, Cambridge, pp. 218–240.

Hall, S., 1998. Aspiration and attitude…reflections on black Britain in the nineties. New Formations 33, 38–46.

Hartmann, D., Gerteis, J., 2005. Dealing with diversity: mapping multiculturalism in sociological terms. Sociological Theory 23 (2), 218–240.

Heath, A., Roberts, J., 2008. British Identity: Its Sources and Possible Implications for Civic Attitudes and Behaviour (online). http://www.justice.gov.uk/docs/british-identity.pdf (last accessed on 28.03.11.).

Heath, A.F., Cheung, S.Y., 2007. Unequal Chances: Ethnic Minorities in Western Labour Markets. published for the British Academy by Oxford University Press, Oxford.
Joppke, C., 2004. The retreat of multiculturalism in the liberal state: theory and policy. British Journal of Sociology 55 (2), 237–257.
Joppke, C., 2010. Citizenship and Immigration. Polity Books, Cambridge.
Koopmans, R., Statham, P., Giugni, M., Passy, F., 2005. Contested Citizenship: Immigration and Cultural Diversity in Europe. University of Minnesota Press, Minneapolis, MN.
Kymlicka, W., 1995. Multicultural Citizenship. Oxford University Press, Oxford.
Leonard, M., 1997. Britain TM: Renewing Our Identity. Demos, London.
Levey, G., Modood, T. (Eds.), 2009. Secularism, Religion and Multicultural Citizenship. Cambridge University Press, Cambridge.
The Local, 2011. Islamic Studies Gain Foothold in State Schools. http://www.thelocal.de/education/20111222-39667.html (last accessed 22.12.11.).
Loury, G.C., Modood, T., Teles, S.M. (Eds.), 2005. Ethnicity, Social Mobility, and Public Policy: Comparing the USA and UK. Cambridge University Press, Cambridge.
Malik, K., 17 March 2010. Multiculturalism undermines diversity. The Guardian. http://www.theguardian.com/commentisfree/2010/mar/17/multiculturalism-diversity-political-policy (last accessed 03.01.14.).
Malik, K., 2011. I am still a critic of multiculturalism, honest. Pandemonium. http://kenanmalik.wordpress.com/2011/02/10/still-a-critic-ofmulticulturalism/ (Accessed on 30.03.11.).
Meer, N., Modood, T., 2009. The multicultural state we're in: Muslims, 'multiculture' and the 'civic re-balancing' of British multiculturalism. Political Studies 57, 473–497.
Meer, N., Modood, T., 2012. How does interculturalism contrast with multiculturalism? Journal of Intercultural Studies 33 (2), 175–196.
Modood, T., 1994. Political blackness and British Asians. Sociology 28, 859–876.
Modood, T., 1998. Anti-essentialism, multiculturalism and the 'recognition' of religious minorities. Journal of Political Philosophy 6, 378–399.
Modood, T., 2005. Multicultural Politics: Racism, Ethnicity and Muslims in Britain. Edinburgh University Press, Edinburgh.
Modood, T., 2007. Multiculturalism: A Civic Idea, 2013, second ed. Polity, Cambridge.
Modood, T., Dobbernack, J., 2011. A left communitarianism? What about multiculturalism? Soundings 48, 54–64.
Mouritsen, P., 2008. Political responses to cultural conflict: reflections on the ambiguities of the civic turn. In: Mouritsen, P., Jørgensen, K.E. (Eds.), Constituting Communities: Political Solutions to Cultural Conflict. Palgrave, London, pp. 1–30.
Parekh, B.C., 2000. Rethinking Multiculturalism: Cultural Diversity and Political Theory. second ed., 2006. Harvard University Press, Cambridge, MA.
Peach, C., 2007. Muslim population of Europe: a brief overview of demographic trends and socio-economic integration. Transatlantic Dialogue on Terror. Berlin Conference.
Pew Research Center, 2006. The Great Divide: How Westerners and Muslims View Each Other (online). http://pewglobal.org/2006/06/22/the-great-divide-how-westerners-and-muslims-view-each-other (last accessed on 28.03.11.).
Portes, A., Zhou, M., 1993. The new second generation: segmented assimilation and its variants. The Annals of the American Academy of Political and Social Science 530, 74–96.
Presseurop, 2010. Mutti Merkel Handbags Multikulti (online). http://www.presseurop.eu/en/content/article/364091-mutti-merkel-handbags-multikulti (accessed on 28.03.11.).
Sen, A., 2006. Identity and Violence. Allen Lane, London.
Soysal, Y., 1994. Limits of Citizenship: Migrants and Post National Membership in Europe. Chicago University Press, Chicago, IL.
Taylor, C., 1994. The politics of recognition. In: Gutmann, A. (Ed.), Multiculturalism and "the Politics of Recognition": An Essay. Princeton University Press, Princeton, NJ, pp. 25–73.
Taylor, C., 2009. Foreword: what is secularism? In: Levey, G.B., Modood, T. (Eds.), Secularism, Religion and Multicultural Citizenship. Cambridge University Press, Cambridge, pp. xi–xxii.
Travis, A., 2002. The Need to Belong – but with a Strong Faith (online). http://www.guardian.co.uk/uk/2002/jun/17/september11.religion1 (last accessed on 02.06.11.).
Triandafyllidou, A., Modood, T., Meer, N. (Eds.), 2012. European Multiculturalisms: Cultural, Religious and Ethnic Challenges. Edinburgh University Press, Edinburgh.
Vasagar, J., 2012. Third of new free schools are religious. The Guardian. http://www.newstatesman.com/uk-politics/2008/08/religious-state-secular (last accessed 13.07.12.).
Vertovec, S., 2007. Super-diversity and Its implications. Ethnic and Racial Studies 30, 1024–1054.
Waldron, J., 1991. Minority cultures and the cosmopolitan alternative. University of Michigan Journal of Law Reform 25, 751.
Weldon, F., 1989. Sacred Cows. Chatto & Windus, London.
Wright, M., Blooemrad, I., 2012. Is there a trade-off between multiculturalism and socio-political integration? Policy regimes and immigrant incorporation in comparative perspective. Perspective on Politics 10 (1), 77–95.

Integration, Social

Richard Münch, Otto-Friedrich-Universität Bamberg, Bamberg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social integration refers, in the first instance, to the extent and intensity of the interlinkages among the constituent parts of a social unit. This definition applies to all types of social units – a small group like a family; a gang or an Internet newsgroup; an organization like a firm, an administration or a nongovernmental humanitarian organization; a community like a neighborhood, a local church community, a local business community, or a local ethnic community; or a society like a local society, a national society, a supranational society, for example, European society or even world society.

Social integration refers, in the first instance, to the extent and intensity of the interlinkages among the constituent parts of a social unit. This definition applies to all types of social units – a small group like a family; a gang or an Internet newsgroup; an organization like a firm, an administration, or a nongovernmental humanitarian organization; a community like a neighborhood, a local church community, a local business community, or a local ethnic community; or a society like a local society, a national society, a supranational society, for example, European society or even world society.

Actor Integration: Cohesion and Breaking Apart, Integration and Disintegration, Harmony and Conflict

What are the constituent parts of social units? On one level, we can say that these are actors as members of social units like groups, organizations, communities, and societies. On a second level, families form constituent parts of clan societies, estates of estate societies, classes of class societies, functional units of functionally differentiated societies, and organizations of organizational societies; groups and divisions are second-level constituent parts of organizations. What is then the extent and intensity of interlinkages between the constituent parts of a social unit? With regard to members of social units as individual actors or corporate actors (groups, organizations, communities, and national societies within world society), it is the number of actors involved in recurrent interaction (extent) and the frequency of interactions (intensity). Such interactions can be more or less harmonious or conflictual. Actors can be extensively and intensively linked by conflict. For example, a couple could live in war from marriage to death. Such a case would be covered by the thus far outlined definition of social integration. Yet, there must be stronger forces of attraction that keep them together than just forces of repulsion (Simmel, 1908/1992, pp. 312–315; Coser, 1956, pp. 81–85).

Here, we enter the field of explaining social integration. One explanation is that they are linked together because both of them have either no alternative at all or alternatives that are too costly so that transaction costs would be too high. For reasons of clarity, we can use the term 'cohesion' in order to denote a type of relationship between the members of a social unit that entails extensive and intensive interlinkage, as there are more forces of mutual attraction than of mutual repulsion. The opposite of social integration is not conflict but social disintegration; both are the extreme poles of a continuum with many gradations in between and a middle zone where it is difficult to precisely determine where integration ends and disintegration begins. We can say that a couple is more integrated the more they share a common life together and the more their action is mutually supportive, and it is more disintegrated the less this is the case. Just as integration has its opposite in disintegration and not in conflict, so conflict has an opposite of its own, namely, harmony. Conflict means that with two or more actors, one actor's gain implies the other actor's loss. They may have different goals that cannot be attained at the same time, or they may have the same goal, which, however, cannot be reached by every actor at the same time. Harmony, instead, implies that every actor's goal can be realized at the same time or that one and the same goal can be reached by every actor at the same time.

Action Integration: Negative and Positive Integration

A more ambitious notion of social integration is at stake, when we look at not only the integration of actors as members of social units, but also the integration of their actions. This is a second step in defining social integration. The integration of actions can be realized in two ways: negatively and positively. The negative way avoids the actions interfering with each other. The ideal-typical case of this integration of actions is embodied by two farmers who live on their own on their demarcated territory with guarded fences so that there is no chance of intrusion in either's domain of freedom. This kind of coordinated freedom can be called negative freedom. It is a freedom that results from preventing others from intruding in one's domain of free action. The corresponding type of integration can be called 'negative integration.' Durkheim used the term 'negative solidarity' for that (Durkheim, 1964). It refers to the mutual acknowledgment of property rights. In the case of negative integration, the actors do not need more than the acknowledgment of their freedom to act by other actors, they do not require their cooperation and support.

The counterparts to negative freedom and negative integration are positive freedom and positive integration, which explicitly rely on mutual cooperation and support (Durkheim, 1964). The most elementary form of positive freedom can be

found in market exchange. As Adam Smith (1937) demonstrated, the utility maximization of one actor does not occur at the cost of other actors here, but even furthers the other actors' own utility maximization. In the founding years of sociology, it was Herbert Spencer (1972) who promoted this idea. The actors are not engaged in a zero-sum game, but in a game with a growing sum of which potentially every actor can make a profit. Their actions are in a complementary relationship to each other. Both exercising negative freedom on one's own space without interference from others and positive integration through market exchange are only possible if the actors share a minimum of rules of the game; otherwise they would end up in struggles for advantages by unlimited means including force and fraud, the Hobbesian war of all against all (Hobbes, 1966). The question is then how it is possible that free actors come to share a minimal set of rules of the game. In a first step, cooperation in market exchange might rely on the human individual's sympathy for other people as Adam Smith, as well as David Hume, assumed (Smith, 1966; Hume, 1777/1980). However, this sympathy mostly does not reach beyond particular communities and does not always guarantee the commitment to the rules of the games. Therefore, there is a need for further arrangements. Here the notion of 'positive freedom' becomes relevant. We move from the Scottish moral philosophers' individualistic theory of market integration to Rousseau's collectivistic theory of integration by the general will based on the republican spirit of citizens (Rousseau, 1973). A further tradition of thought, which takes on significance here, is German idealism. It was founded by Fichte, Schelling, and Hegel in response to Kant's critical philosophy and carried on in transformed ways by Marx and his followers right through to critical theory and Habermas' theory of communicative action (Kant, 1956; Hegel, 1965; Marx, 1962; Habermas, 1984, 1989).

Social Integration and Social Order

At this stage of treatment of social integration, social integration comes close to social order as soon as one asks for the integration of free action and of rights to act. Integration and disintegration then correspond to the pair of order and disorder or anomie in Durkheim's terms (Durkheim, 1964). From this viewpoint, the actions of individuals are integrated, in as much as everybody can rely on normative regulations that determine the space of their actions so that negative interferences are avoided. Social order of this kind is a normative order, as Talcott Parsons has emphasized in particular (Parsons, 1968). It is based on processes of regulation through collectively binding decision making. Its opposite, disorder or anomie, means that individual actors cannot count on such regulations when they encounter other actors and thus have to rely on the power available to them in order to defend their claims for space for action. As everybody has to do so, there is always the danger of the Hobbesian war of all against all.

If actors and groups are free to articulate interests within a wide framework of constitutional rules, and if the actors see themselves dependent on the tolerance and/or cooperation of the other actors, they will engage in bargaining processes in which actors look for the maximization of their interests. Such rules will have better chances of survival, which allow for the growing satisfaction of a growing number of interests. Here the actors conceive of themselves as utility maximizers who learn to cooperate in a way that is beneficial for all. This is how the evolution of cooperation is explained from the point of view of utility maximization. This order is an ongoing process of steady change that serves for one interest some time and another interest at another time, or better, an always changing mixture of them. The evolutionary explanation of social order in economic terms has been particularly demonstrated by Robert Axelrod (1984). Generally speaking, this tradition of thought in sociology goes back to Herbert Spencer (1972) and is represented in contemporary sociology, for example, in the works of George C. Homans (1974), Michael Hechter (1987), and James Coleman (1990). The basic paradigm is rational choice theory. There have been many attempts to transcend the limits of economic thought, beginning with the classics, particularly Adam Smith in his work on moral sentiments (Smith, 1966). One of the more prominent recent attempts is Jon Elster's acknowledgment of the noneconomic, in his view nonrational, emotional sources of norm-regulated behavior (Elster, 1989).

However, there are also other settings that exert a different shaping influence on the emergence and continuation of social order. There may also be unbridgeable conflicts between the goals of major actors. In this case, social order is an outcome of a power struggle with winners who shape social order by their ideology and impose that order on the defeated through the application of monopolized physical force. This is what Dahrendorf (1958) has emphasized in his conflict theory of social order. A more complex model of conflict theory for the explanation of social order, which includes negotiation and ritual, has been constructed by Randall Collins (1975).

A completely different setting is given, when there are well-established institutions of discursive reasoning and professional legislation in the hands of legal and other types of experts. Both interest negotiation and power struggle have a more limited space for influencing the emergence and continuation of social order. This process takes place much more as the steady building up of an all-embracing, consistent, coherent, and detailed set of laws. The major question, which decides on the success of a certain proposal of legislation, is not so much as to what interests are served and what power can be mobilized, but as to what good reasons can be given and as to how it fits in the existing system of law. This type of social order is a steadily growing legal order that covers almost everything that changes in an ongoing process of adjustments to new problems and challenges. In this perspective, we conceive of social order as a norm-building process. There are different approaches to addressing this problem. The tradition of German idealism sees norm building as a process of meaning construction. The modern version of this approach is Habermas' (1984, 1989) discourse ethics, which explains norm building as a procedure of rational argumentation. Another perspective of relevance here is phenomenology as developed by Schütz (1964), Berger and Luckmann (1966), and others. Here, norm building is conceived of as a process of

institutionalization through habitualization, sedimentation, and traditionalization, as well as legitimation from rudimentary explanations to generalized symbolic. From the point of view of Harold Garfinkel's (1967) ethnomethodology and related conversational analysis, normative order is an ongoing process of ordering through the concerted action of individuals in their respective situation. For symbolic interactionism as promoted by Herbert Blumer (1969) – based on the work of George Herbert Mead (1934) – and by Anselm Strauss (1978), for example, order is negotiated in processes of mutually fitting different perspectives on the situation. In the pragmatist perspective developed in the works of William James, Charles Sanders Peirce, John Dewey, and George Herbert Mead, and revitalized by Hans Joas, order is produced and reproduced by the creative acts of individuals who cooperate in the solution of problems (Joas, 1992).

A further factor that exerts an influence on the emergence and continuation of a social order is the building of relationships of solidarity in the sense of mutual support and help. The breakdown of the traditional order and its hierarchical system of privileges for estates was partly caused by the lacking inclusion of the rising bourgeoisie according to its contribution to the working of the society, particularly in terms of tax payments. There was a new class that was inadequately included in the participation in associational, political, and cultural life. The bourgeois revolutions established a new form of inclusion, which was much broader. Inclusion was no longer based on inherited status, but simply on achievement (particularly in terms of market income, but increasingly also in terms of education and professional status). This new form of inclusion served as the door for every further movement aiming at the participation in economic, political, associational, and cultural life and their organizations. The first one to succeed was the labor movement with the inclusion of the working class; it was followed by different kinds of movements, from the women's movement, the civil rights movement, and the movements for any kind of minority rights, right through to the movements for multiculturalism, prochoice, and prolife. The sociological approach that sees the emergence and continuation of social order particularly as a process of solidarity, civic order and citizenship production, reproduction, and transformation has been founded by Emile Durkheim. This approach is carried on, in particular, in the functionalism of Talcott Parsons and in its contemporary revitalization by neofunctionalism. A further major contribution is T.H. Marshall's work on citizenship (Durkheim, 1893/1964; Marshall, 1964; Parsons, 1971; Alexander, 1998).

The Negative Effects of Social Integration

Social integration is not an overall positive state of affairs. In as much as it is limited to a particular unit below world society, it is always accompanied by at least some demarcation from other social units and corresponding disintegration. In this sense, nation-state disintegration is a precondition for supranational integration. Further on, social integration resides at least to some degree on processes of internal homogenization that may include not only the peaceful merging of different regions, ethnicities, and cultures, but also the imposition of a central culture on peripheral cultures, of the domination of a central ethnicity on peripheral ethnicities, and of a central region on peripheral regions. Instruments for this process of 'internal colonization' are the penetration of the center's language, legal system, bureaucracy, school system, and communication system into the peripheral regions (Hechter, 1975). In this way, social integration levels down differences and thus the richness of cultures and social life as well as the potential for innovation and change. Thus, we can say that a balance of integration and at least some disintegration is a necessary prerequisite for preserving the potential for diversity and innovation. This was particularly emphasized by Emile Durkheim (1982). Simmel (1908/1992) pointed out the interrelation of integration and disintegration, harmony and conflict. Lewis Coser (1956) and Donald Levine (1980) revitalized this contribution by Simmel emphatically.

Social Integration and Systems Integration

Social integration becomes increasingly complex and increasingly requires a balance with elements of disintegration the more one proceeds in the evolution of human societies. Clan societies in the early stages of sociocultural evolution are integrated by blood and rules of intermarriage. Traditional estate societies are integrated by every estate serving a specific function for the living of the whole and the attribution of privileges according to the inherited status. Cohesion and the integration of actions are promoted by the hierarchical order of inherited ranks (Weber, 1976; pp. 133–135, 148–155). Modern societies are integrated in a much more complex way. Their cohesion is based on comprehensive inclusion according to achievement and the constitution of citizenship with civil, political, social, and cultural rights as an abstract community beyond any primordial ties of ethnicity, color, gender, or any other kind of group.

Beyond social integration, modern societies are systemically integrated in as far as they have established functionally specified systems like economy, polity, law, science, education, the media, arts and literature, or religion. Money and democratically legitimated political power are examples of media of communication that serve as means of systemic integration in as much as they allow for the coordination of extremely differentiated economic interests or political goals, irrespective of time, place, and persons involved (Lockwood, 1964; Luhmann, 1988).

In the process of globalization, the functionally specialized systems increasingly reach beyond the nation-state with its relatively strong social integration and produce a world society that is much more integrated systemically than socially through solidarity production and a binding social order (Luhmann, 1997). It is disputed whether one approaches worldwide systemic integration without social integration or whether there is a necessity and also a chance of a subsequently growing social integration on the supranational and even global levels. If we understand social integration in terms of the relatively strong positive integration of the nation-state, there are little chances for social integration on the supranational and global levels. If we realize, however, that this process involves a transformation

of social integration in terms of farther-reaching and internally more differentiated solidarity of networks instead of homogeneous social units, which is accompanied by a corresponding abstraction and formalization of justice, we can imagine that supranational and global systemic integration can still be complemented by social integration.

Social Integration Beyond the Nation-State

A new level of inclusion has now been achieved in the process of transcending the system of nation-states by supranational units like the European Union (EU) or even the United Nations. A major driving force behind this transformation of social integration is the international division of labor. In Emile Durkheim's terms, the shrinking of distances through population growth, and, in the highly developed regions, much more through the increasingly faster methods of transport and communication, is the major force that promotes the growing division of labor (Durkheim, 1964). In this situation, competition for scarce resources intensifies and leaves specialization as the only effective means of escaping the struggle of all against all, besides suicide or emigration, which are, however, less attractive for people who love their lives and their country. Emigration is also only effective for people in poorer regions yet ever less feasible because of the crowding of the economic centers. The change from nation-state inclusion to European and worldwide inclusion is a further step from mechanical to organic solidarity in Durkheim's view. The high level of equality of results in the European welfare states can be interpreted as the most advanced level of founding the organic solidarity of labor division on much reconstituted mechanical solidarity within the boundaries of the nation-state. It reserves the equality of results for natives and excludes people in the poorer regions from its accumulated wealth. In Max Weber's terms, in-group morality and out-group morality are strictly differentiated (Weber, 1923; pp. 303–304). Within the boundaries of the EU, this differentiation is fading and being replaced by one comprehensive morality that is now less substantial and concrete and more formal and abstract justice as fairness within and beyond the boundaries of the nation-state (Münch, 2010, 2012).

The question as to what organic solidarity is and how it is constituted is, however, difficult to answer (Parsons, 1967). First of all, it consists of commitments between individuals who gain in independence from their collective of origin – family, class, and nation – through specialization and growing ties of exchange beyond group boundaries. The farther their relationships reach beyond group boundaries and the greater their number, the more they attain autonomy. Pioneers of cross-border networking exert pressure on less active people to follow in their footsteps, because the commitment to one's group of origin does not offer security any longer, and because claims to loyalty and support by that group appear no longer legitimate in the context of growing cross-border exchange and shrinking in-group ties. In a self-strengthening process, the pioneers create cross-border networks of autonomous individuals while, at the same time, older collective solidarities become weak and fragile.

While classic industrial workers were supported by their class and its representation by trade unions and labor parties, far better educated employees become entrepreneurs themselves, who have to invest in the growth of their human capital. Lifelong learning is crucial for improving one's returns on investment (Pongratz and Voß, 2003; Bröckling, 2007; Jakobi, 2009). The inclusion of the entrepreneurial individual in society beyond national borders does not depend on the solidarity of families, classes, or nations any longer, but on the openness and differentiation of the network woven by the international division of labor. Exclusion in terms of lacking access to markets, wealth, and recognition does not affect whole classes, but marginalized individuals. These disadvantaged people cannot improve their position through collective action, but only through likewise individualized forms of coping with their situation such as, for example, through measures of retraining, further education, self-help groups, or drifting toward various forms of retreat from social life. Collective action against exclusion is totally fragmented. It does not unite people across different situations of frustration. Global movements like criticism of globalization and Occupy Wall Street, on the one hand, lack coherence, endurance, and a forceful idea of reform. National movements, on the other hand, are backward oriented and try to revitalize a collective solidarity that has faded away already.

According to Durkheim, 1964, cross-border economic exchange and division of labor serve as a kind of material infrastructure from which ties of transnational solidarity can emerge. Yet, this needs further supporting devices. Within the nation-state, the control of economic competition and exchange by occupational groups was such a device for him. We can say that the collaboration of the government, employer federations, and trade unions in the national welfare state created a kind of corporatist capitalism that kept economic growth and social integration balanced and came close to Durkheim's vision of a socially integrated capitalist society (Streeck, 1999). It was also in line with Polanyi's (1944) idea that the rise of capitalism as disembedding of the economy from moral ties calls for a new re-embedding in such ties in order to restore social integration in the modern capitalist society. Marshall (1964) coined this development in terms of expanding citizenship from civil to political and to social rights. For Talcott Parsons (1971), it was the development of a societal community. Jeffrey Alexander (2006) has continued along this track with his work on the civil sphere as the center of social integration in modern society. International labor division was kept under control by the segmentary differentiation of world society into strong nation-states.

With the liberalization of markets for goods and services, and particularly the liberalization of financial markets, the preconditions for worldwide economic growth have been improved, although the preconditions for social integration have deteriorated. The growth of moral ties beyond the nation-state has not kept pace with the growth of economic transactions. Financial transactions have proceeded even far beyond the production and exchange of goods so that there is not only a growing gap between global economic exchange and global social integration, but also between the exploded financial transactions and actual economic production and exchange. In this situation, transnational financial and economic integration is being accompanied by national

disintegration. While transnational moral ties are growing but remain still weak, national solidarity is eroding (Rueda, 2007; Streeck, 2009; Wacquant, 2008/2009; Palier, 2010). While economic inequality between nations has decreased slightly since 1970, it has increased within nations (Firebaugh, 2003; Alderson et al., 2005; Beckfield, 2006; Bartels, 2007; Hacker and Pierson, 2010). This contrary course of transnational and national integration confirms Durkheim's (1964) statement that growing international labor division reduces differences between nations but increases them within nations because of individual specialization. The latter is increasingly difficult to subject to collective solidarity organized by trade unions. In this process, national collective forces working against inequality and disintegration caused by market competition are weakened (Kim and Sakamoto, 2010).

At this point, we might ask about the chances of reconstituting a kind of corporatist collaboration between governments, employer federations, and trade unions on the international level along the lines of Durkheim's and Polanyi's argument that the capitalist economy needs to be embedded in ties of solidarity in order to keep its socially destructive forces under control. In fact, the growth of cross-border economic transactions has been followed by increasing international activities of governments and nongovernmental organizations in regulating the economy (Boli and Thomas, 1999). The World Trade Organization is expected to keep free trade and the working of national economies balanced. The International Labor Organization is committed to the institutionalization of basic rights of employees everywhere. An enormous number of international nongovernmental organizations (INGOs) are devoted to promote the respecting of human rights all over the world. Multinational corporations are under the observation of INGOs with regard to their compliance with standards of employee rights and ecological balance (Curbach, 2009). In order to demonstrate their willingness to cooperate, they have engaged in self-regulation such as the global compact for realizing basic rights of employees worldwide. Corporate social responsibility has become a major device in the struggle for global recognition (Hiß, 2006). However, this follow-up growth of the moral regulation of global economic competition is far behind the level that had been attained by the national welfare state.

Not even the most advanced supranational political unit – the EU – has been able to reconstitute social integration on the Union level. According to the neofunctionalist theory of integration, growing economic integration is expected to spill over to legal and political integration (Haas, 1958/1968). This is close to Durkheim's thesis that international labor division creates the tracks along which organic solidarity may advance (Durkheim, 1964). In fact, the history of the EU shows increasing legal and political integration at an accelerating pace since the mid-1980s (Fligstein and Stone Sweet, 2002). The Union's legal integration is largely limited to making cross-border economic transactions easier. Its political integration lags far behind the political capacities of the nation-state (Münch, 2001, 2010). Therefore, the EU has still limited capacities to cope with the socially destructive effects of European economic integration and global economic competition. The Economic and Currency Union struggles with the destructive effects of mounting state debts in economically weak member states. As a consequence, there is a large gap between the globally minded elites and the still nationally minded masses both with regard to European and global integration (Fligstein, 2008; Haller, 2008). The elites are advancing transnational integration, while the masses miss the quality of life in security in the transnational context and therefore still expect this security from the nation-state. It is, however, the dilemma of social integration in the multilevel system that transnational integration cannot proceed without some kind of national disintegration and that transnational integration cannot attain the quality of solidarity that existed within the national welfare state (Rodrik, 2011; Münch, 2012).

See also: Anomie; Communities of Practice; Community Sociology; Control, Social; Cooperation: Sociological Aspects; Culture and the Self: Implications for Psychological Theory; Diffusion in Sociological Analysis; Diffusion: Anthropological Aspects; Diffusion: Geographical Aspects; Diversity and Disagreement in Ethics: Philosophical Implications; Groups, Sociology of; Social Stratification; Social Structure; Social System; Values, Sociology of.

Bibliography

Alderson, A.S., Beckfield, J., Nielsen, F., 2005. Exactly how has income inequality changed? Patterns of distributional change in core societies. International Journal of Comparative Sociology 46 (5–6), 405–423.
Alexander, J.C., 1998. Neofunctionalism and After. Blackwell, Oxford.
Alexander, J.C., 2006. The Civil Sphere. Oxford University Press, Oxford.
Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York.
Bartels, L.M., 2007. Unequal Democracy. The Political Economy of the New Gilded Age. Princeton University Press, Princeton.
Beckfield, J., 2006. European integration and income inequality. American Sociological Review 71 (6), 964–985.
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality. Penguin, Harmondsworth.
Blumer, H., 1969. Symbolic Interactionism. Prentice Hall, Englewood Cliffs, NJ.
Boli, J., Thomas, G.M. (Eds.), 1999. Constructing World Culture. International Nongovernmental Organizations since 1875. Stanford University Press, Stanford, CA.
Bröckling, U., 2007. Das unternehmerische Selbst. Suhrkamp, Frankfurt am Main, Germany.
Coleman, J.S., 1990. Foundations of Social Theory. Belknap Press of Harvard University Press, Cambridge, MA.
Collins, R., 1975. Conflict Sociology: Toward an Explanatory Science. Academic Press, New York.
Coser, L.A., 1956. The Functions of Social Conflict. Free Press, New York.
Curbach, J., 2009. Die Corporate-Social-Responsibility-Bewegung. VS Verlag, Wiesbaden, Germany.
Dahrendorf, R., 1958. Toward a theory of social conflict. Journal of Conflict Resolution 2, 170–183.
Durkheim, E., 1893/1964. The Division of Labor in Society (trans. Simpson, G.). Free Press, New York.
Durkheim, E., 1982. In: Lukes, S. (Ed.), The Rules of Sociological Method. MacMillan, London.
Elster, J., 1989. The Cement of Society. A Study of Social Order. Cambridge University Press, New York.
Firebaugh, G., 2003. The New Geography of Global Income Inequality. Harvard University Press, Cambridge, Mass.
Fligstein, N., 2008. Euroclash. The EU, European Identity, and the Future of Europe. Oxford University Press, Oxford.
Fligstein, N., Stone Sweet, A., 2002. Constructing polities and markets: an institutionalist approach of European integration. American Journal of Sociology 107, 1206–1243.
Garfinkel, H., 1967. Studies in Ethnomethodology. Prentice Hall, Englewood Cliffs, NJ.

Haas, E.B., 1958/1968. The Uniting of Europe. Political, Social and Economic Forces, 1950–1957. Stanford University Press, Stanford.
Habermas, J., 1984, 1989 (trans. McCarthy, T.). The Theory of Communicative Action, vol. 2. Beacon Press, Boston.
Hacker, J.S., Pierson, P., 2010. Winner-Take-All Politics. How Washington Made the Rich Richer – And Turned Its Back on the Middle Class. Simon and Schuster, New York.
Haller, M., 2008. European Integration as an Elite Process. The Failure of a Dream. Routledge, London/New York.
Hechter, M., 1975. Internal Colonialism. The Celtic Fringe in British National Development, 1536–1966. Routledge and Kegan Paul, London.
Hechter, M., 1987. Principles of Group Solidarity. University of California Press, Berkeley, CA.
Hegel, G.F.W., 1965. Philosophy of Right. Clarendon Press, Oxford.
Hiß, S., 2006. Warum übernehmen Unternehmen politische Verantwortung? Campus, Frankfurt/New York.
Hobbes, T., 1966. Leviathan. In: Molesworth, W. (Ed.), Collected English Works of Thomas Hobbes, vol. 3. Scientia, Aalen, Germany.
Homans, G.C., 1974. Social Behavior: Its Elementary Forms. Harcourt Brace Jovanovich, New York.
Hume, D., 1777/1980. In: Selby-Bigge, L.A. (Ed.), Enquiries Concerning the Human Understanding and Concerning the Principles of Morals. Greenwood Press, Westport, CT.
Jakobi, A., 2009. International Organizations and Lifelong Learning: From Global Agendas to Policy Diffusion. Palgrave, Houndmills, Basingstoke.
Joas, H., 1992. Die Kreativität des Handelns. Suhrkamp, Frankfurt am Main, Germany.
Kant, I., 1956. Die Metaphysik der Sitten. In: Weischedel, W. (Ed.), Werke in sechs Bänden, vol. 4. Insel Verlag, Frankfurt am Main, Germany.
Kim, C.H., Sakamoto, A., 2010. Assessing the consequences of declining unionization and public-sector employment. A density-function decomposition of rising inequality from 1983 to 2005. Work and Occupations 37 (2), 119–161.
Levine, D., 1980. Simmel and Parsons. Two Approaches to the Study of Society. Arno Press, New York.
Lockwood, D., 1964. Social integration and systems integration. In: Zollschan, G.K., Hirsch, W. (Eds.), Explorations in Social Change. Routledge, London.
Luhmann, N., 1988. Die Wirtschaft der Gesellschaft. Suhrkamp, Frankfurt am Main, Germany.
Luhmann, N., 1997. Die Gesellschaft der Gesellschaft, vols. 2. Suhrkamp, Frankfurt am Main, Germany.
Marshall, T.H., 1964. Class, Citizenship and Social Development. Greenwood Press, Westport, CT.
Marx, K., 1867/1962. Das Kapital vol. 1. In: Marx-Engels-Werke, vol. 23. Dietz, Berlin, Germany.
Mead, G.H., 1934. In: Morris, C.W. (Ed.), Mind, Self and Society. University of Chicago Press, Chicago.
Münch, R., 2001. Nation and Citizenship in the Global Age. Palgrave, Basingstoke.
Münch, R., 2010. European Governmentality. The Liberal Drift of Multilevel Governance. Routledge, London.
Münch, R., 2012. Inclusion and Exclusion in the Liberal Competition State: The Cult of the Individual. Routledge, London.
Palier, B. (Ed.), 2010. A Long Goodbye to Bismarck? The Politics of Welfare Reform in Continental Europe. Amsterdam University Press, Amsterdam.
Parsons, T., 1968/1937. The Structure of Social Action. Free Press, New York.
Parsons, T., 1967. Durkheim's contribution to the theory of integration of social systems. In: Parsons, T. (Ed.), Sociological Theory and Modern Society. Free Press, New York.
Parsons, T., 1971. The System of Modern Societies. Prentice Hall, Englewood Cliffs, NJ.
Pongratz, H.J., Voß, G.G., 2003. Arbeitskraftunternehmer. Erwerbsorientierungen in entgrenzten Arbeitsformen. Edition Sigma, Berlin, Germany.
Polanyi, K., 1944. The Great Transformation. Rinehart Press, New York.
Rodrik, D., 2011. The Globalization Paradox: Democracy and the Future of the World Economy. W.W. Norton, New York and London.
Rousseau, J.J., 1973. The Social Contract and Discourses (trans. Cole, G.D.H.). Dent, London.
Rueda, D., 2007. Social Democracy Inside Out. Partisanship & Labor Market Policy in Industrialized Democracies. Oxford University Press, Oxford.
Schütz, A., 1964. Collected Papers II: Studies in Social Theory. Nijhoff, The Hague, The Netherlands.
Simmel, G., 1908/1992. Soziologie. Untersuchungen über die Formen der Vergesellschaftung. Suhrkamp, Frankfurt am Main, Germany.
Smith, A., 1937. The Wealth of Nations. Modern Library, New York.
Smith, A., 1966. The Theory of Moral Sentiments. A.M. Kelley, New York.
Spencer, H., 1972. In: Peel, J.D.Y. (Ed.), On Social Evolution. Selected Writings. University of Chicago Press, Chicago.
Strauss, A., 1978. Negotiations. Varieties, Contexts, Processes and Social Order. Jossey-Bass, San Francisco.
Streeck, W., 1999. Korporatismus in Deutschland. Zwischen Nationalstaat und Europäischer Union. Campus, Frankfurt/New York.
Streeck, W., 2009. Re-Forming Capitalism. Institutional Change in the German Political Economy. Oxford University Press, Oxford.
Wacquant, L., 2008. Urban Outcasts: A Comparative Sociology of Advanced Marginality. Polity Press.
Wacquant, L., 2009. Punishing the Poor: The Neoliberal Government of Social Insecurity. Duke University Press, Durham, NC.
Weber, M., 1923. In: Hellman, S., Palyi, M. (Eds.), Wirtschaftsgeschichte. Duncker & Humblot, Munich, Germany.
Weber, M., 1976. Wirtschaft und Gesellschaft. Mohr Siebeck, Tübingen, Germany.

Intellectual Functioning, Assessment of

Leslie Atkinson, Ryerson University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Three assessment approaches, each with different conceptualizations of intelligence and why it should be measured, are reviewed. The defining feature of psychometric assessment is its empiricism. The approach rests on strong method but its theoretical origins are slim. Exploiting both these factors, psychometricians bootstrapped a structure of intelligence model, reduced (sub)cultural bias, and increased remedial utility in these tests. The tests are still criticized, however, for their limited sampling of the intelligence domain. Psychometric tests are good classifiers and predictors in academic and work contexts. Focusing not on what an examinee knows but how he/she reasons, neuropsychological approaches rest on theories of brain function to assess the *process* of cognition. Empirical considerations remain important but are subsumed by theory. Although on introduction these tests had severe limitations, second-generation neuropsychologically based intelligence tests work with improved psychometrics and a better match of subtests to theory. Psychometric and neuropsychologically based approaches are showing a trend toward integration, both in terms of utilizing converging theoretical models and as practitioners increasingly rely on cross-battery assessment. This merged approach appears the way of the future; psychometric and neuropsychologically based frameworks have both demonstrated that they can generate novel findings and seamlessly incorporate new cognitive advances. The third approach, dynamic assessment of intellectual functioning, incorporates varied interactive evaluation procedures to assess and modify learning potential. The strength of dynamic assessment lies in its individualized, remediation-based reconceptualization of intellect and its attention to bias. While ambitious and still widely used, research in the approach is subsiding, as is practice. The dynamic approach will likely fade from mainstream assessment, although its emphases on reducing test bias and targeting remediation will remain a legacy.

Oversimplifying somewhat, there are three applied approaches to assessing intelligence (Daniel, 1997). Each rests on a different conceptualization of what intelligence is and why it should be measured. Each approach, the psychometric, the neuropsychological, and the dynamic, is reviewed below.

In addition, several physiological indices (e.g., brain wave tracings, reaction and inspection times, nerve conduction velocity, rate at which glucose is metabolized in the brain) correlate with intelligence measures (Neisser et al., 1996). However, physiological indices are not reviewed here because they have little applied utility. For example, a quantitative review of 172 studies showed only small to moderate correlations between information processing speed (typically involving reaction time to simple stimuli) and psychometric tests of intelligence (Sheppard and Vernon, 2008).

Psychometric Assessment

The psychometric approach has dominated intelligence testing for a century. The defining feature involves its empirical foundation; 'psychometric' simply refers to the quantitative measurement of psychological states/abilities. While quantitative assessment rests on a massive measurement technology, its theoretical origins are shallow. Contemporary intellectual measures emerged directly from studies by Alfred Binet and colleagues in France. In 1904, Binet was directed to devise a means of distinguishing educable from noneducable students in the relatively new universal education system. Having investigated cranial, facial, palmar, and handwriting indices, Binet discovered that the direct measure of complex intellectual tasks involving judgment, comprehension, and reasoning most successfully distinguished among pupils. Based on these pragmatic beginnings, Binet (1905) defined intelligence as the capacity to adopt and sustain a direction, make adaptations for the purpose of attaining a desired end, and monitor performance self-correctively. With little elaboration, this definition still directs the psychometric paradigm.

Typically, modern psychometric instruments consist of varied subtests that tap diverse aspects of the loosely defined intelligence construct. For example, scales may include subtests that sample a broad range of knowledge (e.g., vocabulary, historical facts) and require the examinee to solve unfamiliar problems (e.g., assemble colored blocks to resemble a prespecified design) (Groth-Marnat, 2009). Again, choice of subtests is not driven by theoretical prescription. Subtests are selected because they work – in combination, they serve to rank individuals according to how much they know and how good they are at solving particular problems. The pragmatic selection of subtests is based on Binet's (1905) broad definition of intelligence and Spearman's (1904) conception of intelligence as a general or undifferentiated ability (*g*); in principle, the tasks that tap *g* are interchangeable (Mackintosh, 2011).

At the heart of psychometric testing lies norm referencing (Urbina, 2011). Norm referenced tests are developed by administering items in a standardized manner to a representative sample of the population. The normative sample is considered 'representative' insofar as it is stratified within age groups for variables that might influence performance differentially, such as sex, geographic region, ethnic status, community size, etc. Scores are scaled such that each individual's derived score represents a relative standing within the normative or standardization group. In this sense, psychometric testing is an empirical endeavor in its purest sense: as a comparative construct, there is little need to theorize about the exact nature of intelligence.

As mentioned, most modern psychometric tests include varied tasks. The original intention was to ensure that *g* was comprehensively surveyed. With time, however, clinicians came to exploit the multitask construction of intelligence tests to make intraindividual comparisons (Urbina, 2011). By looking at the variability among subtests or groups of subtests, assessors hypothesized about the examinee's relative intellectual strengths and weaknesses. For example, a particular respondent might prove better on tests of verbal comprehension than on tests involving memory. It is important to note, however, that the analysis of intraindividual differences developed after the fact; such comparisons are driven by the practicalities of what subtests are available.

The empirical base of the psychometric effort implies both strength and weakness. The central strength involves its emphasis on reliability and validity of test scores. 'Reliability' refers to consistency of measurement; the more reliable a measure, the less error involved in estimates derived from it. Many psychometric tests boast extremely high internal reliability (the degree to which each component score of the test correlates with the full test score) and 'test–retest' reliability (an index of stability derived by administering the test to the same sample of individuals more than once). Short-term stability of composite scores typically exceeds 0.90 and can show reasonable stability over a 20-year time span (Groth-Marnat, 2009). The validity of these tests, too, has proven strong (Gottfredson and Saklofske, 2009). 'Validity' refers to the extent to which a test measures what it is designed to measure. Intelligence test scores correlate with the amount of schooling, quality of work produced in school, occupational status, and performance in the work situation, both concurrently and predictively (Gottfredson and Saklofske, 2009). Intelligence tests are also sensitive to neuropsychological deficit (Groth-Marnat, 2009) and, perhaps because life requires continual learning and reasoning, intelligence is related to accidental injury, lower adherence to medical treatment, preventable chronic illness, and premature death (Gottfredson and Saklofske, 2009).

The atheoretical, empirical approach central to psychometric testing, lies at the heart of many test shortcomings. Ironically, however, this same atheoretical, empirical base confers an elasticity that permits adaptive change via renewed test content, testing procedures, and test development/revision strategies, all designed to attenuate test limitations.

To exemplify, the interpretation of intraindividual differences is often emphasized over full-scale score, although the latter is the most reliable and valid indicator in any intelligence test (Gottfredson and Saklofske, 2009). As mentioned above, subtests are pragmatically chosen and the implications of subtest variability are rarely validated empirically. Furthermore, the atheoretical approach to task selection has resulted in restricted and incomplete sampling of the intelligence domain. For example, it has been argued that intelligence tests neglect musical and interpersonal abilities (Chen and Gardner, 2012; Sternberg, 2012) as well as vital cognitive domains such as rational thinking (Stanovich, 2009) and learning potential (Feuerstein et al., 1979).

Despite the pragmatic selection of subtests, however, post hoc analyses of intelligence and other cognitive data reveal that they converge in remarkably similar ways across tests and studies (Schneider and McGrew, 2012). Three hierarchical strata of intelligence have been identified in what is called the Cattell–Horn–Carroll (CHC) theory of intelligence. The CHC model is one of the most important developments in over a century of intelligence testing (Schneider and McGrew, 2012), providing, as it does, an empirically and theoretically convincing model of the structure of intelligence. At the highest stratum is a general factor, *g*. This factor subsumes a second stratum of broad factors, including 'fluid' and 'crystallized' intelligence. Fluid intelligence involves the ability to cope with novelty and think flexibly. Crystallized intelligence involves the storage and use of declarative knowledge such as vocabulary or information. Subsumed under each broad factor is a set of narrow abilities, such as 'induction' and 'reading comprehension.' The CHC model has proven applicable across all ages, sexes, races, and cultures evaluated thus far (Gottfredson and Saklofske, 2009). As such, it assists in interpreting intraindividual cognitive variability. Furthermore, the model could be used to inform one about important vectors neglected by extant intelligence tests (Schneider and McGrew, 2012). The CHC model exerts a profound influence on test construction/revision while imposing little restriction on the type of tasks that can be incorporated into intelligence tests.

It remains true, however, that while psychometric testing has illuminated the structure of intelligence, the CHC model provides little insight into the meaning of intelligence (Urbina, 2011) or the operations, processes, and mechanisms that underlie it (Mackintosh, 2011; Sternberg, 2012). In part, psychometricians circumvent this imperfect comprehension by assessing intellectual function with reference to a norm group; the question an assessor asks is not 'how intelligent is this person?' but 'how well does this person perform compared to other people who are like him/her in important demographic respects?' However, this gives rise to questions about bias because (sub)cultural, ethnic, life-experience, and motivational differences could render the standardization sample unrepresentative of any given examinee. This is not to say that the construct of *g* is false, but that a given item/subscale may not tap into *g* equally for all individuals. This becomes a social issue when examinees from minority groups are compared to a norm sample whose context, values, and learning experiences are different from their own. Testing thereby betrays its original purpose of providing objective data on an individual's intellectual functioning and comes, instead, to discriminate against atypical examinees. While this issue has historically plagued test construction, contemporary test developers typically incorporate a panoply of bias-mitigation procedures, including improved standardization sampling, increasingly sophisticated statistical approaches to bias detection, minority stakeholder reviews, reduction of language demands, and assessment of examiner and examinee features that may influence test performance (Gottfredson and Saklofske, 2009).

Another difficulty with psychometric tests is that, although they usually correlate highly among themselves, this is not always the case. Correlations may be influenced by what tasks are included and how they are weighted. Perhaps a greater problem lies in the fact that even where test scores do correlate highly, the same individual may earn discrepant scores on different instruments because tests are normed on different

standardization groups and at different times. These issues are not entirely resolvable, but improved norming standards and increased understanding of the structure of intelligence, as provided by the CHC model, will likely attenuate these differences (Gottfredson and Saklofske, 2009).

A crucial criticism of psychometric tests is that recommendations derived from these instruments have not been shown to enhance remediation for the examinees (Feuerstein et al., 1979). Again, this can be attributed to the fact that the content of these scales has not been selected according to any theory of intelligence, brain functioning, or pedagogy. Nevertheless, some test developers now incorporate companion guides into their test kits that address the issue of test-based intervention directly (Truch, 2006), although research in the area remain sparse.

To summarize, although there are limitations to psychometric measurement, the approach yields reliable and valid estimates of intellectual functioning. Psychometric tests are accurate classifiers and predictors when used with care in appropriately circumscribed contexts. The empirical nature of the psychometric approach permits flexibility in test improvement, such that the limitations discussed above are theoretically tractable. The CHC model provides insight into the latent structure of intelligence, informing test development without imposing undue restrictions, while the empirical foundations of the approach permit development and refinement of items/subtests that incrementally improve test properties. A century of research on the psychometrics of intelligence amply confirms an early assessment of the field: "simply … measuring promiscuously any large number of different abilities and pooling the results together …, which might seem the most arbitrary and meaningless procedure imaginable, had really a profound theoretical basis and a supremely practical utility" (Spearman, 1930).

Neuropsychologically Based Assessment

The neuropsychological approach to intellectual assessment is represented by several theory-based tests that were developed in reaction to the pragmatics of psychometric assessment. The neuropsychological approach, emerging in the 1980s, reflects an attempt to increase the validity and utility of assessment by basing it on an understanding of brain function and brain–behavior relationships. The emphasis is not on what the examinee knows but on how the examinee reasons; not on observable abilities but on underlying capacities (Sparrow and Davis, 2000). Repudiating the psychometric emphasis on *g*, neuropsychologically oriented test construction operationalizes the notion that the brain is composed of interdependent but separable functional systems (neuromodularity).

First-generation neuropsychological approaches to intelligence testing were almost exclusively influenced by the theories of Alexander Luria (1980), a Russian psychologist. Luria proposed three levels of brain functioning that mediate distinct types of learning and performance. The first, attention, involves selectively attending to particular stimuli while inhibiting attention to competing stimuli over time. Attention is assessed with tasks that demand sustained concentration on one stimulus and sustained inhibition of response to distracting stimuli. The second functional unit involves encoding and processing information. According to Luria, information can be processed simultaneously and/or successively. Simultaneous processing is holistic; it involves organizing stimuli into interrelated groups. Simultaneous processing can be assessed with tasks that require an examinee to recall the spatial locations of stimuli, and construct abstract designs from several identical triangles. Successive processing involves integrating stimuli serially, such that each stimulus is related to the one that precedes and follows it in chainlike succession. Successive processing can be tested with tasks requiring, for example, that the respondent repeat numbers spoken by the examiner or repeat an ordered series of hand movements made by the examiner. Planning and monitoring functions form the third and highest cognitive level, according to Luria. This involves the integration of attention and simultaneous and successive processing, along with acquired knowledge, to permit the formation and execution of plans and ongoing verification of their efficacy. Neuropsychological tests of intelligence include subtests that assess such processes. These scores can be compared to derive a profile of strengths and weaknesses, which, in turn, has direct educational implications. It is also important to note that psychometric considerations such as reliability, validity, and norm referencing are stressed, but they subserve theoretical relevance (Kaufman, 2000).

Second-generation neuropsychological intelligence tests are based on the elaborations of Luria's theory (Naglieri et al., 2012) and neurocognitive findings on cerebral specialization. Of extreme theoretical import, at least one test (Kaufman and Kaufman, 2004) also incorporates the CHC model, simultaneously measuring performance on the CHC factors and on the levels of brain functioning using largely overlapping subtests. Not all neuropsychologically based test developers adopt this approach (Naglieri et al., 2012), however, suggesting that the field is moving toward a continuum of test options representing psychometric and neuropsychological traditions at the extremes, with hybrid instrumentation occupying a middle ground. Furthermore, this trend is augmented by the increasing popularity of cross-battery assessment, which involves the systematic augmentation of a core test with carefully selected components of other tests to extract psychometrically robust and comprehensive understanding of an individual's strengths and weaknesses. Originally founded on the CHC model, the cross-battery approach was later augmented with neuropsychological theory (Flanagan et al., 2013), further integrating psychometric and neuropsychological approaches to the assessment of intellectual function.

The neuropsychological approach to intelligence testing has limitations and strengths. A central shortcoming of these theory-based tests lies in the speculative nature of the theories upon which the tests are based (Gottfredson and Saklofske, 2009). This is compounded by controversy regarding whether test components actually measure the theoretical constructs they purport to measure. This is particularly true where the same subtests are used to represent both levels of brain processing and CHC factors (Mays et al., 2009), contributing some ambiguity to their interpretation. Resolution of issues pertaining to 'construct validity' (whether a test measures what it purports to measure) is crucial.

What of test bias, as discussed in the context of psychometric testing? Although the neuropsychological approach focuses aggressively on minimizing discrepancies between whites and minority groups, early attempts worked by sacrificing breadth of intellectual sampling (Urbina, 2011) and minimizing differences artifactually. However, recent efforts employing psychometric technologies such as listed above show that it is possible to reduce racial discrepancies legitimately, but not necessarily to eliminate them. These discrepancies indicate the need for judicious use of testing and appropriate advocacy for resource allocation to disenfranchised populations.

It has been argued that the theoretical underpinnings of neuropsychological tests of intelligence both restrain and augment the breadth of subtest selection (Das et al., 1994). Which perspective one takes depends on the neuropsychological test one is considering and what one considers the purpose of testing to be. It is true that such tests neglect areas of intelligence such as creativity and wit, but the purpose of these new instruments is primarily to aid in pedagogic planning. Hence, it is the insight these tests afford into a given examinee's strengths and weaknesses with respect to learning that is important, as well as the remedial implications of the testing. Although further research is needed, preliminary evidence indicates efficacy in both these domains, with carefully controlled trials showing hypothesized gains in prespecified areas (Naglieri et al., 2012). For example, students with poor planning improved dramatically in math calculation when instructed in the use of strategies; other students who were good in planning showed more modest improvements when instructed in the use of strategies.

In the past, it has been argued that if these instruments prove successful in directing effective pedagogy (Atkinson, 2001) and if they are seen as more than just alternative or adjunctive to psychometric measures (Sparrow and Davis, 2000), they will have addressed the major shortcoming of psychometric assessment without sacrificing its empirical strengths. These are high standards and, although more research is needed, it is fair to say that these standards are in the process of being met.

Dynamic Assessment

Incorporating a heterogeneous set of evaluation procedures, the defining feature of dynamic assessment is that it evaluates and modifies cognitive functioning simultaneously through mediation and intervention. The product of a dynamic assessment is not an estimate of the examinee's intellectual standing relative to their peers, but an estimate of learning potential. The notion behind the learning potential construct is that individuals with similar starting competencies (e.g., similar IQs) may respond differentially to instruction; what an individual has learned does not necessarily reflect the individual's potential to learn (Grigorenko and Sternberg, 1998). Some theorists equate learning potential with *g* (Fabio, 2005), although they emphasize that *g* is mutable through training (Corman and Budoff, 1973). Dynamic assessment represents a reaction against psychometric testing in particular, with its purported overreliance on previously learned material; neglect of process in favor of content; failure to consider factors other than ability in measuring IQ (e.g., the (sub)cultural, motivational, personality, and social adequacy differences in examinees); and irrelevance to remediation (Fabio, 2005). The theoretical writings of Lev Vygotsky, a Russian psychologist, form the context of dynamic assessment. Vygotsky (1987) argued that what a child can accomplish when assisted by someone more knowledgeable ('zone of proximal development') is more indicative of the child's mental development than what the child can do alone ('zone of actual development'). The zone of proximal development is operationalized in dynamic assessment as a 'mediated learning experience' (Feuerstein et al., 1979) by a collaborative examiner (a 'mediator,' in dynamic parlance). The mediator provides ongoing feedback, titrated such that it is just beyond the examinee's zone of actual development, until the examinee either solves the problem or gives up. Thus, testing and teaching are joined. This is facilitated by emphasizing tasks that stress fluid rather than crystallized reasoning skills. The Vygotskian basis of dynamic assessment is supplemented by information processing concepts in some dynamic approaches. For example, one approach stresses the central role of working memory, the ability to hold old and new information in memory simultaneously, so as to manipulate and transform it (Campione and Brown, 1987). Dynamic instrumentation often includes standardized psychometric tasks, as described above, but they are administered interactively.

Dynamic approaches vary along two inversely related continua: individuality of assessment and psychometric adequacy. At one end of these continua, the assessor works in intensive, individualized interaction, modifying tasks freely, encouraging, prompting, and mediating between child and task at will to derive qualitative (as opposed to quantitative) diagnoses and educational prescriptions (Feuerstein et al., 1979). One commentator described the approach as "a clinician's dream and a psychometrician's nightmare" (Lidz, 1997). At the other end of the continua lies a test–train–retest sequence with standardized mediation and outcome measures. Mediation is delivered through a series of predetermined hints that vary from general to specific. Each hint is offered in response to the youngster's difficulty. Outcome is quantified as the number of prompts required to bring the child to correct response and as the ability to transfer to tasks other than those on which training first occurred (Campione and Brown, 1987). Some dynamic practitioners doubt the efficacy of standardized dynamic procedures, because clues are not individually modified to the child's needs (Lidz, 1997). However, the method's standard, quantifiable format renders it amenable to psychometric study.

The goals of dynamic assessment – to assess learning ability, to identify learning processes, and to teach the individual generalizable strategies for problem solving, i.e., to capitalize on cognitive plasticity (Fabio, 2005) – remain its major strength. No other assessment tradition is so ambitious. Dynamic assessment advocates have sharpened the focus of other assessment approaches, alerting them to the ultimate goal of intellectual evaluation. And in selecting response to remedial efforts as an outcome (as opposed to comparing an examinee with his or her age mates), dynamic assessors propose a powerful means of attenuating experiential bias in intelligence testing. The reasoning is as follows: Suppose there

are two children, one biologically disadvantaged with mental retardation, the other who has experienced atypical living/educational opportunities. One might expect relatively little success in teaching the child with mental retardation. However, the educationally atypical child should make great gains, i.e., there should be a large discrepancy between this youngster's pretest and posttest scores (or the child should require relatively few clues to reach problem solution), whether or not the posttest scores approximate the average test score of this child's educationally advantaged age peers. The point is that pretest scores, conceptually comparable to conventional test results, tap a different aspect of intelligence than do difference scores; pretest scores correspond to developed abilities (and, implicitly, past opportunities to develop abilities), while difference scores approximate learning potential. It is by tapping into learning potential that dynamic assessments attenuate the biases of traditional assessment.

However, research support for the claims underlying dynamic assessment is sparse, ambiguous, and often flawed. This is more the case for approaches that emphasize nonstandardized assessment than for those that stress psychometric adequacy, but weak research foundations are common to all approaches (Grigorenko and Sternberg, 1998). Little validation of the 'zone of proximal development' exists, for example, and much of what does exist has been conducted at the level of theory, ideology, and description, rather than quantitative experiment. And although change is central to the dynamic philosophy, virtually no work addresses the reliability of change scores. Little has been published on internal consistency and there is ambiguous evidence that change scores derived from different tasks do not correlate (Grigorenko and Sternberg, 1998). Although there may be reasons to expect incongruence of change scores, these reasons need to be explored and verified so that systematic predictions can be made. Questions have also been raised about variability among examiners in administering, scoring, and interpreting the less standardized assessments, and the possibility of obtaining different results for a given child as a result of this. Few studies address the issue of retest reliability of change scores, and some studies suggest that retest reliability is weak (Buchel and Schanhorst, 1993). Factor analytic studies, which are used to demonstrate that subtests intercorrelate according to theoretical expectations, are rarely conducted and do not always confirm hypotheses (Grigorenko and Sternberg, 1998). Studies have shown that mediation promotes change from pre- to posttest (Delclos et al., 1992), but it is not always clear how much of the change is attributable to mediation and how much change occurs simply by virtue of practice (Grigorenko and Sternberg, 1998). Importantly, some research shows that dynamic indices augment prediction of posttest performance beyond data derived from pretest performance (Day et al., 1997). However, other studies have not confirmed the independent value of gain indices over traditional measures; results on this aspect of dynamic testing are ambiguous (Grigorenko and Sternberg, 1998). From a long-term perspective, some evidence suggests a relation between ability to benefit from mediation and economic and social independence in adulthood among individuals with low learning potential (Lidz, 1997). However, there have been no demonstrations that dynamic assessment actually changes school or work performance or leads to recommendations that change performance (Grigorenko and Sternberg, 1998). In all, dynamic approaches to assessment represent an ambitious but unproven perspective on human intelligence. Moreover, scientific research and, indeed, the practice of dynamic assessment, have fallen precipitously in the last 15 years, suggesting the approach is in decline, its greatest contributions already made.

Concluding Comments

In an earlier article in this Encyclopedia, it was suggested that "The aforementioned approaches to intellectual assessment have shared and unique strengths. With time, it is reasonable to expect an amalgamation of these strengths such that, with the addition of practical biological measures of intelligence, assessment will become increasingly reliable, valid, fair, and useful" (Atkinson, 2001). In part, this is what has transpired, although the promise of biological measures is over because they do not measure brain–behavior relationships, they neglect the complexity of human behavior, and they do not relate to diagnostic, intervention, or pedagogical strategies. While the dynamic approach to assessment is falling out of the mainstream, its legacy remains in the form of emphases on fair testing and pedagogical relevance. The psychometric and neuropsychological approaches to intelligence testing are in process of merging, based on their joint statistical underpinnings and the flexibility of test construction this permits, increasingly shared paradigms, and amenability to integration through cross-battery assessment. Improved structural models (CHC) will likely result in increasingly accurate measures of *g*, encompassing broader aspects of intelligence. At the same time, greater understanding of CHC and neuropsychological models will move us toward refined assessment of intraindividual differences, with teaching/intervention implications. This integrated approach to intellectual assessment permits successive classification, prediction, and remediation that encompass large groups (e.g., people with average *g*) and the smallest groups (i.e., the individual) alike.

See also: Binet, Alfred (1857–1911); Classical (Psychometric) Test Theory; Cognitive Neuropsychology, Methods of; Cognitive Neuroscience; Emotional Intelligence and Competencies; Intelligence, Genetics of: Cognitive Abilities; Intelligence: Central Conceptions and Psychometric Models; Intelligence: Historical and Conceptual Perspectives; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Personal Intelligence and Competencies; Vygotsky's Theory of Cognitive Development; Wisdom, Psychology of.

Bibliography

Atkinson, L., 2001. Intellectual functioning, assessment of. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of Social and Behavioral Sciences. Elsevier, Amsterdam, pp. 7600–7605.

Binet, A., 1905. New methods for the diagnosis of the intellectual level of subnormals. L'Annee Psychologique 12, 191–244 (Kite, E.S., 1916, Trans.). The Development of Intelligence in Children. Training School at Vineland, Vineland, NJ. http://psychclassics.asu.edu/Binet/binet1.htm.

Buchel, F.P., Schlatter, C., Scharnhorst, U., 1997. Training and assessment of analogical reasoning in students with severe learning difficulties. Educational and Child Psychology 14, 109–120.

Campione, J.C., Brown, A.L., 1987. Linking dynamic testing with school achievement. In: Lidz, C.S. (Ed.), Dynamic Testing. Guilford, New York, pp. 82–115.

Chen, J.-Q., Gardner, H., 2012. Assessment of intellectual profile: a perspective from multiple intelligences theory. In: Flanagan, D.P., Harrison, P.L. (Eds.), Contemporary Intellectual Assessment: Theories, Tests and Issues, third ed. Guilford, New York, pp. 145–155.

Corman, L., Budoff, M., 1973. A Comparison of Group and Individual Training Procedures in the Raven Learning Potential Measure (Rie Print #56). Research Institute for Educational Problems, Cambridge, MA (ERIC Document Reproductive Services No. ED 086924).

Daniel, M.H., 1997. Intelligence testing: status and trends. American Psychologist 52, 1038–1045.

Das, J.P., Naglieri, J.A., Kirby, J.R., 1994. Assessment of Cognitive Processes: The PASS Theory of Intelligence. Allyn and Bacon, Needham Heights, MA.

Day, J.D., Engelhardt, J.L., Maxwell, S.E., Bolig, E.E., 1997. Comparison of static and dynamic assessment procedures and their relation to independent performance. Journal of Educational Psychology 89, 358–368.

Delclos, V.R., Vye, N.J., Burns, M.S., Bransford, J.D., Hasselbring, T.S., 1992. Improving the quality of instruction: roles for dynamic testing. In: Haywood, H.C., Tzuriel, D. (Eds.), Interactive Testing. Springer-Verlag, New York, pp. 317–332.

Fabio, R.A., 2005. Dynamic assessment of intelligence is a better reply to adaptive behavior and cognitive plasticity. The Journal of General Psychology 132, 41–64.

Feuerstein, R., Rand, Y., Hoffman, M., 1979. The Dynamic Assessment of Retarded Performers: The Learning Potential Assessment Device – Theory, Instruments, and Techniques. University Park Press, Baltimore, MD.

Flanagan, D.P., Ortiz, S.O., Alfonso, V.C., 2013. Essentials of Cross-Battery Assessment, third ed. Wiley, Holboken, NJ.

Gottfredson, L., Saklofske, D.H., 2009. Intelligence: foundations and issues in assessment. Canadian Psychology 50, 183–195.

Grigorenko, E.L., Sternberg, R.J., 1998. Dynamic testing. Psychological Bulletin 124, 75–111.

Groth-Marnat, G., 2009. Handbook of Psychological Assessment, fifth ed. Wiley, Holboken, NJ.

Kaufman, A.S., 2000. Tests of intelligence. In: Sternberg, R.J. (Ed.), Handbook of Intelligence. Cambridge University Press, Cambridge, UK, pp. 445–476. Wiley, Hoboken, NJ.

Kaufman, A.S., Kaufman, N.L., 2004. Kaufman Assessment Battery for Children, second ed. American Guidance Service, Circle Pines, MN.

Lidz, C.S., 1997. Dynamic assessment approaches. In: Flanagan, D.P., Genshart, J.D., Harrison, P.L. (Eds.), Contemporary Intellectual Assessment: Theories, Tests, and Issues. Guilford, New York, pp. 28–296.

Luria, A.R., 1980. Higher Cortical Functions in Man, second ed. Basic Books, New York.

Mackintosh, N.J., 2011. History of theories and measurement of intelligence. In: Sternberg, R.J., Kaufman, S.B. (Eds.), The Cambridge Handbook of Intelligence. Cambridge University Press, Cambridge, pp. 3–19.

Mays, K.L., Kamphaus, R.W., Reynolds, C.R., 2009. Applications of the Kaufman assessment battery for children, 2nd edition in neuropsychological assessment. In: Reynolds, C.R., Fletcher-Janzen, E. (Eds.), Handbook of Clinical Child Neuropsychology. Springer, New York, pp. 281–296.

Naglieri, J.A., Das, J.P., Goldstein, S., 2012. Planning, attention, simultaneous, successive. A cognitive—processing-based theory of intelligence. In: Flanagan, D.P., Harrison, P.L. (Eds.), Contemporary Intellectual Assessment: Theories, Tests and Issues, third ed. Guilford, New York, pp. 178–194.

Neisser, U., Boodoo, G., Bouchard, T.J., Boykin, A.W., Brody, N., Ceci, S.J., Halpern, D.F., Loehlin, J.C., Perloff, R., Sternberg, R.J., Urbina, S., 1996. Intelligence: knowns and unknowns. American Psychologist 51, 77–101.

Schneider, W.J., McGrew, K.S., 2012. The Cattell-Horn-Carrol model of intelligence. In: Flanagan, D.P., Harrison, P.L. (Eds.), Contemporary Intellectual Assessment: Theories, Tests and Issues, third ed. Guilford, New York, pp. 99–144.

Sheppard, L.D., Vernon, P.A., 2008. Intelligence and speed of information-processing: a review of 50 years of research. Personality and Individual Differences 44, 535–551.

Sparrow, S.S., Davis, S.M., 2000. Recent advances in the assessment of intelligence and cognition. Journal of Child Psychology and Psychiatry 41, 117–131.

Spearman, C., 1904. "General intelligence," objectively determined and measured. American Journal of Psychology 15, 201–292.

Spearman, C., 1930. Autobiography. In: Murchison, C.A. (Ed.), A History of Psychology in Autobiography. Russell & Russell, New York.

Stanovich, K.E., 2009. What Intelligence Tests Miss: The Psychology of Rational Thought. Yale University Press, New Haven, CT.

Sternberg, R.J., 2012. The triarchic theory of successful intelligence. In: Flanagan, D.P., Harrison, P.L. (Eds.), Contemporary Intellectual Assessment: Theories, Tests and Issues, third ed. Guilford, New York, NY, pp. 156–177.

Truch, S., 2006. The WISC-IV Companion: A Guide to Interpretation of Educational Intervention. Pro-Ed, Austin, TX.

Urbina, S., 2011. Tests of intelligence. In: Sternberg, R.J., Kaufman, S.B. (Eds.), The Cambridge Handbook of Intelligence. Cambridge University Press, Cambridge, pp. 20–38.

Vygotsky, L.S., 1987. The Collected Works of Lev Vygotsky, vol. 1. Plenum Press, New York.

Intellectual History

Cesare Cuttica, Université Paris 8-Vincennes, Saint-Denis, France

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by F. Oz-Salzberger, volume 11, pp. 7605–7612, © 2001, Elsevier Ltd.

Abstract

Intellectual history as an academic discipline is nowadays taken for granted: learned journals are devoted to it and universities in the UK and the US offer a variety of graduate and postgraduate courses in the field. However, this was not always the case. After a brief panoramic account of its historical development on both sides of the Atlantic, this article will focus on the main work carried out in intellectual history in the last few decades. Particular attention will be given to some practitioners (e.g., Quentin Skinner, John W. Burrow, Stefan Collini) and their methodological approach to the study of past ideas. By concentrating on the practice and role of the intellectual historian, the *object* 'intellectual history' will be presented not only as a vibrant scholarly activity but also as a valuable attempt at understanding our society from a plurality of perspectives. In order not to sound too optimistic about this intellectual enterprise, attention will equally be paid to the *French case* as an instance of a country where intellectual history is still seen with suspicion, and of an academic world where it is not that much studied nor practiced.

Introduction

To be asked 'what do you do for a living?' might be an occasion for displaying one's pride in himself/herself, a way of showing one's ambitions, a source of embarrassment or, less dramatically but perhaps more interestingly, a chance of reflecting seriously on how one spends his/her days. Historians are people who are particularly prone to embark upon this self-examining journey: what instruments do they use when studying the past? What kind of material do they work on? Why do they choose one methodological and/or historiographical approach instead of another?

Intellectual history offers the paradox of a field of research whose practitioners, despite constantly addressing this type of questions, often find themselves under attack for pursuing a historical enterprise which is not, and cannot be, clearly defined; for lacking rigor as well as a recognizable identity; and for being amateurs allergic to archives. A frequently heard refrain has it that, because of its shifty nature, intellectual history is like "nailing jelly to the wall" (Parageau, 2010). In order to tackle these claims, and reveal their manifest narrow-mindedness, this article traces some key historical developments in intellectual history and casts new light on the work of some of its most significant practitioners.

A Historical Take

What is now commonly referred to as 'intellectual history' has a precursor. That precursor took the name of 'the history of ideas' with at its head the American philosopher Arthur O. Lovejoy (1873–1962) – founder of the long-standing *Journal of the History of Ideas* (1940) – who first used the phrase in 1919. Before then, one needs to go back a good couple of centuries and consider the lesson of the German *Aufklärung*, Giambattista Vico's elaboration of his *New Science* (1725) and the Eclecticism of Victor Cousin in nineteenth-century France (Kelley, 2002). Cousin (1792–1867) developed the history of ideas as an independent discipline associated with philosophy, but one where the prevalence was of history rather than of ideas in the Platonic sense. The ideal and the real found thus a bridge, first, in the history of ideas and, subsequently and more cogently, in intellectual history. At the discipline's inception, a crucial discussion saw those who regarded ideas as exclusively the product of the mind ('internalist' perspective) opposed to those for whom ideas had to be placed in their cultural, political, social context ('externalist' perspective). This exemplifies the contrast between the notion of the timelessness of ideas and their transcendent validity, and that of their inescapably historical and contextual nature.

On the basis of this brief – and inevitably selective – account, we can agree with Anthony Grafton that the history of ideas was always practiced, even before becoming part of the discipline of history (Grafton, 2006). In this respect, in the current panorama of higher education the existence of academic chairs as well as of undergraduate and postgraduate courses dedicated to intellectual history, especially in British, North American, and Australian universities, might be taken for granted. Similarly given is that recognition has also been conferred upon it by the foundation of specialized publications such as *History of European Ideas* (1980) and the more recent *Modern Intellectual History* (2004) and *Intellectual History Review* (2007). Intellectual history – we may dare say – has never had it so good. And yet this achievement took a long time to be accomplished and was never short of impasses (Mandelbaum, 1965).

The cornerstone of Lovejoy's work was the investigation of conceptual problems derived from, and intertwined with, history and philosophy. He contested William James' proposition that consciousness and ideas do not have an existence, and argued that there were unit-ideas, which remained constant throughout history but to which varying historical circumstances gave different answers. Unit-ideas framed all great works of literature, art, science, and philosophy in the West (e.g., the great chain of being). They constituted an intellectual continuum recoverable thanks to the critical selection of texts (Lovejoy, 1964). For Lovejoy this did not mean to deal exclusively with the great thinkers, but it entailed to extend

research to popular opinions, so as to grasp problems, questions and experiences of the time under scrutiny (Kelley, 2002).

According to Felix Gilbert, in the late 1930s in the US the term "'intellectual history' had not yet become a household word", even though it had 'gradually' entered scholarly parlance. Gilbert identified Perry Miller's *New England Mind* (1939) as the first work, which could be properly defined as "an 'intellectual history'". One distinct feature of the discipline was its 'search for precision' in reaction to accusations that intellectual historians used words like "'influence' or 'dominating trend'" too loosely. On Gilbert's account, focus on 'exactitude' suggested intellectual historians' willingness 'to compete' with social historians. As a result, intellectual history acknowledged the existence of 'political, economic, and social forces' and underscored its relation to them. For the intellectual historian, this meant to work at the 'crucial point of the historical process' where 'long-range factors', on one side, and 'the individual event', on the other, became connected by 'human consciousness' (Gilbert, 1971).

Turning to Britain in the 1930s, we find another famous historian of ideas, Isaiah Berlin (1909–97), experience a large dose of intellectual loneliness in that "[i]ntellectual history fell between two stools, neglected by philosophy, history and politics alike". In fact, at that time "the study of the history of ideas […] barely existed at Oxford". This is why Berlin accepted more and more invitations to lecture in the US, where "intellectual history, especially of the Russian nineteenth century, had academic standing" (Ignatieff, 1998). On the whole, it is undeniable that in Britain a major contribution to the discipline was given by émigré scholars like Berlin himself and Arnaldo Momigliano (1908–87) – and closer to us by, e.g., Istvan Hont (1947–2013). And this in a climate of open hostility to it shown by other émigré scholars like the *anti-intellectual* Lewis Namier and Geoffrey Elton (Young, 2006).

With the 1950s (notably, in the US) came the rise of intellectual history as a prominent discipline until its phase of 'decline' in the late 1960s, followed by a full-on 'crisis' in the 1970s and early to mid-1980s, when a new style in the 'rethinking' of intellectual history and its identity was fostered by scholars like Hayden White (1928–) and Dominick LaCapra (1939–) (Jacoby, 1992). Writing in 1977, Paul. K. Conkin stated that in the US "[i]ntellectual history now seems as dated as narrow ties", to the extent that only "a peculiarly reckless graduate student" could "don the label and move with it into a perilous job market" dominated by fashions (Conkin, 1977). Ten years later *and* in Britain, John W. Burrow too spoke of the lack of 'recognition' accorded to intellectual history (Cuttica, 2014). So much so that he epigrammatically declared that "the fate of an intellectual historian" was "to feel at best marginal in the company of 'proper' historians and, when giving talks, to be playing what Donald [Winch] calls 'away matches'" (Burrow, 2009). In a similar vein, Stefan Collini has pointed out that while historians of philosophy work on past texts, which have been reputed to be *philosophically interesting*, intellectual historians are condescendingly given the task to deal with the rest, that is with the less interesting philosophical writers or ideas of the past (Collini, 2001).

Significantly though, Burrow's contribution to assessing the status of the discipline written almost 20 years later offered a much more optimistic picture. In fact, he spoke of the 'revolution' that had occurred in 'the academic standing' of intellectual history in England (Burrow, 2006). A key role in this important change has no doubt been played by the Cambridge School (a misnomer) principally identified with John Dunn (1940–), John G.A. Pocock (1924–), and Quentin Skinner (1939–) and by the University of Sussex among whose original members were Burrow himself (1935–2009) and Collini (1947–).

The Cambridge School

The 1960s and, above all, the 1970s saw the emergence of the contextualist approach to the history of ideas pioneered by Dunn, Pocock, and Skinner (scholars engaged in the history of political thought). Shifting the traditional focus from the great authors of the past taken as immortal and coherent voices of eternally valid truths, the history of ideas became the history of languages, discourse, paradigms, vocabularies, ideologies in historical context. Instead of being seen proleptically (as the anticipation of other doctrines in the present) or anachronistically (as the projection in the past of concepts elaborated now), ideas were interpreted by placing the texts in which they were conveyed within their intellectual milieu. Dunn adopted this method to study John Locke (1969), while Skinner set forth this way of reading past authors in articles on Hobbes (1960s) and in *The Foundations of Modern Political Thought* (1978); as for Pocock, one needs to consider his influential *The Machiavellian Moment* (1975). Despite sharing a general methodological take on the past, these scholars hold by no means the same views.

In particular, whereas Pocock studied the theories expounded by thinkers at different times by capturing the various languages they spoke (e.g., the language of the ancient constitution, natural law, classical republicanism, politeness, and civility (Pocock, 1985, 2009)), Skinner concentrated on the intentions of the authors. This means to look at texts as the articulation of stances in a dialogue in which a great role is played both by the exchange of ideas and by the constant invention of arguments and counterarguments in controversies. Endorsing this view enables readers to ask which works and which theorists certain texts responded to and why specific languages were chosen to defend a specific political position. Skinner's reelaboration of the theory of 'speech acts' – formulated by J.L. Austin and John Searle – helps to grasp authors' concepts by paying attention to the multifarious ways in which language(s) and discourse(s) function. Moreover, Skinner stressed the importance of considering the doing(s) when exploring the meaning(s) of texts.

The Cambridge School forcefully criticized Lovejoy's 'unit-ideas' for failing to explain how these were used and interpreted by those who did so in their contexts. Ideas (and concepts) mean something for people and their audiences, and are connected to the authors' goals and vocabularies (synchronic approach). Ideas also change over time and patterns of new meanings intervene in different historical phases (diachronic approach). Most importantly, the linguistic approach to political texts pursued by Pocock and Skinner (influenced by authors like R.G. Collingwood and Ludwig

Wittgenstein) differs from the perspective of Reinhart Koselleck's *Begriffsgeschichte* (whose influences lie with Carl Schmitt, Martin Heidegger, and Hans-Georg Gadamer).

German *Begriffsgeschichte*, Michel Foucault, and French Dismissal

Begriffsgeschichte (the history of concepts or conceptual history) originated in the late 1960s and it is generally associated with Otto Brunner (1898–1982), Werner Conze (1910–86), and, above all, Reinhart Koselleck (1923–2006). Their attempt was to connect concepts with social history, examining structural modifications in politics, economy, society. They paid attention to semantic change ('keywords') within social and political contexts. Their *Geschichtliche Grundbegriffe* (1972–97) placed the language used to discuss the state, society, and economy in relation to the strata, groups, and orders that employed or rejected such a language (Richter, 1987). For Koselleck, the fundamental moment in old-Europe German-speaking history was what he called *Sattelzeit* (1750–1850) that signaled the transformation of concepts used in political and social language. This involved the advent of modernity whose key concepts were identified with 'historicization', 'ideologization', 'democratization', and 'politicization' (Cuttica, 2012).

By focusing more on the practical implications of concepts rather than on their linguisticality, Koselleck's perspective incurred the criticism of the Cambridge School. The divergence between the two methodologies recalls what Pocock defined as the problem of the separation/unity of social structure (manifested, for instance, in official languages) and discourse (intended as speech and intertwined with debate) (Pocock, 1987). As for Skinner, *Begriffsgeschichte* detaches concepts from their specific historical settings. It also undermines the role of human agency in the formulation of opinions and theories. According to Skinner, the problem with *Begriffsgeschichte* is its identifying a text as the byword of a static *mentalité* which, in turn, is the expression of a given social formation (Bevir, 2000). Nonetheless, the two approaches share a vigorous rejection of those historiographical currents, which see ideas as passive utterances of a specific social reality or as eternal essences (Tully, 1988; Skinner, 2002).

If this account shows some differences between conceptual and intellectual history, it is also worth recalling that Michel Foucault (1926–84) – well known for his genealogical method – distinguished the history of thought (implying a process of detachment from one's convictions), the history of ideas (the analysis of systems of representations through which we see the world), and the history of mentalities (the analysis of the type of action and attitudes which give meaning to it). Thought expresses our freedom in relation to what we do, so as to establish this as an object and reflect on it as a problem. Hence Foucault referred to a 'history of problematics'. This approach does not entail deconstruction, but a critical analysis where one tries to see how different solutions to a problem have been constructed, and also how these solutions stem from a specific form of problematization. Foucault analyzed discourse as the totality, which characterizes our horizon in history. In other terms, he argued that there is no unified subject whom we might take as an agent of meaning. Rather, there is a dispersion of the (historical) subject. According to its critics, the history of discourse erases the singularity of the individual and his/her creativity (Rabinow, 1991).

The reference to Foucault serves to introduce the historical neglect of the history of ideas in the French academic world. As François Azouvi piquantly remarked more than 20 years ago, outside France to be an intellectual historian does not make one a source of national disgrace (Azouvi, 1992). A key reason for this situation has to do with the predominance of philosophy not only as an academic discipline but also as a fundamental subject in the French teaching system. The history of philosophy has been regarded as having more 'substance' than the history of ideas, as if the object of study attended to by the former had more dignity than that pursued by the latter. While for the history of mentalities and for social history, the history of ideas is too little anchored in the institutional scaffolding of a society, for the history of science and for the history of philosophy it turns out to be too little abstract. The history of ideas has also been accused of spreading some sort of idolatry of the influence and for being deprived of any normative guidance (Azouvi, 1992).

Undoubtedly, the strong role played by the *Annales* school (1950–80) has prevented the development of research in intellectual history, and in the best of cases has tied it heavily with cultural and social history (Roger Chartier). Perhaps, the tendency on the part of (many) French historians to worship archival work might have had something to do with their ignoring intellectual history. It is also important to point out that in France intellectual history has often been confused with the history of the intellectuals, a tradition which has been particularly strong from the *affaire* Dreyfus (1894–1906) and Julien Benda's *La Trahison des Clercs* (1927) up to the controversies between Camus and Sartre, and then on to Bourdieu and the philosophers of the so-called French Theory such as Althusser, Deleuze, and Derrida (Dosse, 2003). More interest has thus been poured into a sociology of the intellectuals than into genuinely critical studies of past ideas in context. This series of factors is thereby behind the still very uncertain status assigned to intellectual history, and the hostility with which it is frequently portrayed, in France.

This being said, new generations of – bi/multi-lingual – scholars are now engaging with intellectual history in refreshing fashions (Vincent, 2003; Parageau, 2010). As Sandrine Parageau has recently observed, intellectual history is criticized for its 'obsession' with methodological issues. It is accused of being parasitic with regard to traditional history and it is dismissed as "[a] cuckoo in the historical nest". After all, Foucault attacked the history of ideas for being confusing in terms of disciplinary boundaries, for its extensive borrowing from other disciplines, and for its lack of soundness. According to Foucault, it was not a discipline, but "un style d'analyse". In fact, as Parageau maintains, intellectual history moves fruitfully across disciplines in a *transdisciplinary* rather than simply *interdisciplinary* manner: while interdisciplinarity "invites exchange between disciplines", transdisciplinarity "goes further, and trenchantly proceeds to pursue a true break in the global conception of the world and of science". Therefore, for Parageau this "freedom and creativity cannot be practiced but on the fringes

of academia and its organization, that is to say there where it is still possible to get rid of disciplinary habits and dare undertake new approaches of research" (Parageau, 2010).

This view was first taken 50 years ago across the Channel in Brighton.

Sussex, Burrow, and Collini

Almost from its foundation (1961), the University of Sussex developed a congenial environment in which to pursue interdisciplinary research. With 'an independent syllabus', separate 'admissions' and 'examinations papers', intellectual history at Sussex 'was at one time the only undergraduate degree of its kind' in the UK (it was preceded by an MA program). The first lectureship in intellectual history was assigned to Larry Siedentop in 1972 and was taken up by Stefan Collini 2 years later (Collini, 2000). Thanks to courses like 'Concepts, method, and values in the social sciences' (CMV), Sussex benefited from a system of joint teaching involving literature, history, philosophy, and the history of science. The CMV's threefold methodology drew on "the history of ideas or intellectual history", attempting "to establish the filiation of particular ideas or schools of thought, their origins and lines of mutual interaction"; on "the approach of the social historian" investigating "the link between social science and the efforts of particular groups or individuals" to improve society and implement ideals into practice; and on "the social scientists approach to the history of social science" (Cuttica, 2014). Together with Siedentop and Collini, Hellmut O. Pappe, Donald Winch, John W. Burrow, Michael Moran, and others formed 'a group of like-minded scholars' who carried out research and teaching, which refused to embrace academic tribalism and departmental straightjackets. Interestingly, according to Collini, to speak of a 'Sussex School' would be 'a culpable form of exaggeration or reification' in that there never was such a programmatic adherence to one 'explicit and exclusive' method (Collini, 2000). Such an idyllic situation lasted until the axe of the government fell on small departments which were by decree incorporated within larger academic units (Cuttica, 2014).

The best way to sketch the significance of the Sussex experience is by illustrating some of the most original ideas on the *métier d'historien* elucidated by Burrow (who moved to Oxford University in 1995) and Collini (who moved to Cambridge University in 1986). The former was a prolific author of monographs devoted to unravel the complexity of many intellectual aspects of modern European culture. Averse to methodological abstraction, Burrow pursued history knowing that the latter – following Burckhardt's maxim – is "the record of what one age finds of interest in another" (Cuttica, 2014). Elegant, sharp, and erudite, Burrow's writings championed the importance of history as a narrative of the past according to its actors in contrast to readings of it imbued with presentism-tailored goals. With Collini and Winch, Burrow explained that the intellectual historian ought "to be alive to the several dimensions of the thought and feeling of the past". This means to be receptive to 'the sensibilities' of various authors; 'to the emotional and aesthetic satisfaction' past theorists 'derived from their views'; to their 'styles'; to the 'genres' through which they conveyed their principles; and to the tension between their search for 'coherence' and their 'deeply felt intuitions' (Collini et al., 1983).

Burrow viewed the practitioner as an explorer whose adventure consists in a 'genuine negotiation' with past utterances, whereby he/she both listens and speaks to them. To listen to past dialogues is a way of overcoming 'the parochialism of the present' and of enabling one's audience 'to listen with an understanding which is not superficial but intimate'. For Burrow, we need to pursue a 'two-sided' engagement with historical phenomena whereby we retain the ability to be surprised by them and their unexpected tonalities. In an equally intriguing fashion, Burrow portrayed intellectual historians as both eavesdroppers on the conversations that went on in the past and translators (Burrow, 2006). As eavesdroppers, they mingle opacity and clarity. As translators, they do not play the role of sympathetic actors reexperiencing events, but focus on transmission. They are a sort of go-between bringing to life the voices of the past. To study history means to travel back in time (and then forward into the present) equipped with concepts, which are our own and through which we attempt to make *some* sense of what took place before. To do so thoroughly is to rethink the categories we apply in our exploration of past ideas and events. All historical reconstruction has thus to be informed by a certain amount of creativity. From this reconstructive attempt emerges an intellectual historian who 'can be bothered to disturb' the 'oblivion' to which too many opinions – now unfashionable, when not disparaged as absurd – have been, and are, confined. To write intellectual history has nothing to do with what *we* think to be important at a given time, but it concerns what past actors considered relevant *then*. Therefore, its practitioner is more interested in highlighting 'intellectual dead ends', which were once fostering debate, than in concentrating on 'portentous moments, which were then not recognized as such'. To undertake this scholarly route means to descend closer to the pulse(s) of the intellectual life of past epochs (Burrow, 2000).

From nineteenth-century historical studies to analyses of the current state of affairs in our universities, Stefan Collini is an eclectic intellectual historian and a widely read author of critical essays and reviews for a multiple range of English-speaking publications. At home in different genres and styles, as well as in a broad spectrum of chronologically disparate topics, Collini writes about complex themes in a manner which, without losing in depth or elegance, is capable of reaching a public vaster than that usually addressed by the majority of academics. Collini has defined his intellectual trajectory as follows: "My formation is very much that of an historian, trained in England and the United States, working primarily on English intellectual history; after doing some work on the history of political thought and the history of sociology, my interests have moved more towards the history of cultural and literary criticism" (Collini, 1988). For Collini, "labels like 'intellectual history'" are thus no more than 'flags of convenience': they are not 'names of essences'. Hence 'those who apply the label to themselves' do not have 'any monopoly on right-minded historical practice'. Like Burrow, Collini prefers the term "intellectual history" to "the history of ideas" because the former underscores the involvement with a 'human activity' rather than with mere abstractions taken as

independent and a contextual object of study. This is why "the best work in intellectual history precisely attempts to give their proper place to…social forces, institutional frameworks, political pressures" in the life of thought and its agents.

In Collini's view, one central concern for the intellectual historian is the instability of all disciplinary boundaries. Disciplines are complex compounds of different 'practices' assembled as a unity thanks to intellectual coherence just as much as to 'historical accident or institutional convenience'. Their multifarious developments and the ways in which they change are the object of research in intellectual history. Thus, intellectual history-driven questions concentrate on why and how a certain label has been assigned to a certain scholarly activity, and consequently this leads to delineate the ambitions, interests, and convictions prevalent in a given society at a specific historical juncture. Collini calls for historians to emancipate themselves from the strictures informing 'disciplinary history' (Collini, 2001). In fact, whereas "discipline-history bores a 'vertical' hole in the past (with all the consequent dangers of 'tunnel-vision'), intellectual history attempts rather to excavate a 'lateral' site, to explore the presuppositions, ramifications, and resonances of ideas, which may often involve pursuing them into neighboring fields" (Collini, 1988).

Drawing on his critical assessment of C.P. Snow's famous and controversial 1959 Rede Lecture, *The Two Cultures*, Collini invites scholars to confront 'larger questions', which have to do with society, so as to expand a dialogue that is not confined to the frequently navel-gazing specialism of one's own profession. In this respect, he refers to 'brilliant individuals like Stephen Jay Gould, Richard Dawkins, or Stephen Hawking' as men of science *and* authors capable of engaging wide audiences without diminishing the degree of creative thinking and scientific rigor of their work. In so doing, it is possible to communicate clearly with 'a nonspecialist readership' and transmit to it the importance of highly specialized and 'technical research' (Collini, 1998).

Another major feature of Collini's writings is their de-mythologizing a series of *idées recues*, which hinder genuine historical understanding. This practice has been applied to debunk what he has called the 'absence thesis' whereby Britain is traditionally depicted as a land lacking intellectuals or one inhospitable to them compared to other countries. Collini has shown the paradoxical nature of many culturally ingrained claims about an object of study, a fashion, a tendency, or a situation. In this respect, he is a dismantler-cum-flair of all rigid assumption impoverishing the historical perspective (Collini, 2006). Allergic to the blinkered theoretical imperatives imposed on historical research, Collini has made clear that "the merits of the history written depends on qualities which no theory can adequately prescribe" (Collini, 1985). In line with this idea, and thanks to a prose sparkling with piercing irony and impressive learning, his writings embody the vocation to abandon 'the ghetto mentality' stifling academic research and to move away from specialists' entrenched 'disdain for communicating' the results of their work more broadly. Indeed, this openness is part and parcel of the type of intellectual history originally made in Sussex whose brand was the promotion of "a public language in which nonquantifiable considerations can be given their proper weight" (Collini, 1988, 1998). Perhaps, Collini's pursuits help delineate a new identity for the practitioner as critic – literary, political, cultural – whose feet are well-grounded in the sticky fields of historical narrative and whose voice reaches audiences not confined to the usual arenas of intellectual exchange.

Intellectual History and Its Future: The Practitioner

Taking the cue from Lovejoy's dictum that "the history of ideas is no subject of highly departmentalized minds" (Lovejoy, 1964), it would be professionally parochial and scholarly unhelpful to prescribe a series of criteria or, worse, rules enabling to identify *true* intellectual history from fake ones. Perhaps, less arrogantly but more usefully, one could point to some – nonexclusive nor exhaustive – traits that inform the practice of intellectual historians now working on varied topics from a plurality of perspectives. To do so might dispel some of the confusion that often shrouds all attempt at drawing intellectual history's (very mobile) borders. It might also provide an idea of future research directions likely to be taken by a large number of those who to the question opening this article would be *brave enough* to answer 'I do intellectual history'.

- Rigorous analysis of texts and critical examination of their content combined with close attention to the material conditions of their production and literary forms; their cultural and socio-political contexts; and their assumption about, e.g., gender, race, and identity.
- Interdisciplinarity and cross disciplinarity.
- Focus on the linguistic behavior of agents in the past.
- Insistence on the contextual dimension of ideas.
- Analysis of locutionary and illocutionary acts (the doings of the agents when performing linguistically), which involves a good degree of attention to the web of meanings, which gave a particular utterance or text their contemporary significance (importance of historical distance in that one deals with assumptions, opinions, references that are not his/her own).
- Equal relevance to failed ideas just as much as to successful and coherent ones. Intellectual history studies *wrong* principles and does not seek truth in the same way as it is often pursued by philosophers.
- Study of the (long-term) consequences, implications, and legacy of ideas, which, in turn, leads to concentrate on their being distorted, manipulated, or simply adjusted to new epochs and/or cultural milieus.
- Rejection of periodization as a way of fixing the boundaries between, say, 'medieval', 'modern', and 'contemporary', and consequently establishing strict areas of competence-expertise between historians.
- Interplay of diachronic–synchronic approach.
- Criticism of anachronism and prolepsis.
- Writing of intellectual biographies where the life of ideas receives as much heed as the personal sphere of the individual(s) treated.
- Interest in the materiality of things-objects charged with cultural significance (e.g., museums, houses, villas, monuments, scientific instruments, places of science).

- Expansion of the range of sources considered as valuable objects of critical exploration (e.g., poetry, fiction, paintings, nature, photographs).
- New consideration for gender and the role of women as agents of intellectual creation as well as for noncanonical authors or once-marginal topics (rejection of elitism).
- Extension of intellectual history studies to non-European, and non-Western, materials, topics, cultural regions (e.g., recent developments in Middle-Eastern, Chinese, and pre-Columbian civilization studies).

These features are intertwined with what Donald Kelley saw as the opening up of intellectual history to new avenues of research: namely, a move from thought to discourse; from the conscious to the unconscious; from creation to imitation; from intention to meaning; from authorship to readership; from the history of ideas to the social history of ideas; from tradition to canon formation; and from the sociology of knowledge to the anthropology of knowledge (Kelley, 1987). Similarly, David Harlan explained that intellectual history cannot ignore the salient part that 'postmodern literary criticism' plays in the investigation of the past and the questions it raises about the conceptual foundations on which historical work stands (Harlan, 1989). Hence intellectual historians' engagement with the issues of 'representation' and 'narrative'. By and large, these considerations confirm the central position held by language in the field: after all, as Kelley once wittily put it, "[l]anguage is an ocean in which we all swim and where we are fishes and not oceanographers" (Kelley, 2002).

Conclusion

Having delineated (by no means exhaustively) the past, the present and the future of intellectual history through the careers and works of a number of historians, a few final words might be spent in sketching a – modest – picture of the practitioner in the here and now. Opponent of the banal, iconoclast of the simplistic and promoter of the obsolete, the intellectual historian unmasks ideology and refutes the truth of a (pretended) *fact*. As critic, he/she turns into debunker of the accepted, dismantler of clichés, saboteur of the assurances of his/her readers, trespasser of academic nomenclatures. Most importantly, given that – as Brian Young has argued – intellectual history contributed 'to shape' the twentieth century just as it was molded by it, with both positive and tragic consequences, and that intellectual historians underwent on their skin the brutality of actions which derived from the ideas of historical figures as much as of fellow practitioners (Young, 2006), it is all the more decisive to reflect on their role in our contemporary society. In a(n academic) world increasingly dominated by the dictates of managerial leadership and by the disconcerting assertions of market-driven philistinism, the intellectual historian has the urgent task to alert a wide(r) public to how concepts and ideas (e.g., immigration, wealth, education, toleration) are nowadays employed in an uncomprehending and historically distorting manner by those in charge of the common good. As a stylist whose trained imagination has the liberating power of bringing together the familiar and the unfamiliar, the known and the unknown, he/she might begin by rethinking the conditions of knowledge available to twenty-first-century men and women.

See also: Absolutism, History of; Cambridge School of Intellectual History; Cultural History; Historical Thought and Historiography: Current Trends; Historicism; History and Literature; Isms and History; Linguistic Turn and Discourse Analysis in History.

Bibliography

Azouvi, F., 1992. Pour une Histoire Philosophique des Idées. Le Débat 72, 16–26.
Bevir, M., 2000. Begriffsgeschichte. History and Theory 39, 273–284.
Burrow, J.W., 2000. The Crisis of Reason. European Thought, 1848–1914. Yale University Press, New Haven and London.
Burrow, J.W., 2006. Intellectual history in English academic life: reflections on a revolution. In: Whatmore, R., Young, B. (Eds.), Palgrave Advances in Intellectual History. Palgrave, Basingstoke-New York, pp. 8–24.
Burrow, J.W., 2009. A History of Histories. Epics, Chronicles, Romances and Inquiries from Herodotus and Thucydides to the Twentieth Century. first ed., 2007. Penguin, London.
Collini, S., 1985. What is intellectual history? History Today 35, 46–54.
Collini, S., 1988. Discipline history and intellectual history: reflections on the historiography of the social sciences in Britain and France. Revue de Synthèse 109, 387–399.
Collini, S., 1998. Introduction. In: Snow, C.P. (Ed.), The Two Cultures. Cambridge University Press, Cambridge, pp. vii–lxxi.
Collini, S., 2000. General introduction. In: Collini, S., Whatmore, R., Young, B. (Eds.), History, Religion, and Culture: British Intellectual History 1750–1950. Cambridge University Press, Cambridge, pp. 1–21.
Collini, S., 2001. Postscript. Disciplines, canons, and publics: the history of the history of political thought in comparative perspective. In: Castiglione, D., Hampsher-Monk, I. (Eds.), The History of Political Thought in National Context. Cambridge University Press, Cambridge, pp. 280–302.
Collini, S., 2006. Absent Mind: Intellectuals in Britain. Oxford University Press, Oxford.
Collini, S., Winch, D., Burrow, J.W., 1983. Prologue. The governing science: things political and the intellectual historian. In: Collini, S., Winch, D., Burrow, J.W. (Eds.), That Noble Science of Politics. A Study in Nineteenth-Century Intellectual History. Cambridge University Press, Cambridge, pp. 1–21.
Conkin, P.K., 1977. Intellectual history: past, present and future. In: Delzell, C.F. (Ed.), The Future of History. Vanderbilt University Press, Nashville (Tenn.), pp. 111–133.
Cuttica, C., 2012. What type of historian? Conceptual history and the history of concepts: a complex legacy and a recent contribution. History and Theory 51, 411–422.
Cuttica, C., 2014. Eavesdropper on the past: John W. Burrow (1935–2009) and intellectual history. History of European Ideas (special issue), forthcoming.
Dosse, F., 2003. La Marche des idées. Histoire des intellectuels – histoire intellectuelle. La Découverte, Paris.
Gilbert, F., 1971. Intellectual history: its aims and methods. Daedalus 100, 80–97.
Grafton, A.T., 2006. The history of ideas: precept and practice, 1950–2000. Journal of the History of Ideas 76, 1–32.
Harlan, D., 1989. Intellectual history and the return of literature. American Historical Review 94, 581–609.
Ignatieff, M., 1998. Isaiah Berlin. A Life. Chatto and Windus, London.
Jacoby, R., 1992. A new intellectual history? The American Historical Review 97, 405–424.
Kelley, D.R., 1987. Horizons of intellectual history: retrospect, circumspect, prospect. Journal of the History of Ideas 48, 143–169.
Kelley, D.R., 2002. The Descent of Ideas. The History of Intellectual History. Ashgate, Aldershot.
Lovejoy, A.O., 1964. The Great Chain of Being: A Study in the History of an Idea. first ed., 1936. Harvard University Press, Cambridge, MA.
Mandelbaum, M., 1965. The history of ideas, intellectual history, and the history of philosophy. History and Theory 5, 33–66.
Parageau, S., 2010. L'histoire intellectuelle dans les études anglophones: un style d'analyse interdisciplinaire? In: Kaenel, A., Omhovère, C., Samin, R. (Eds.), L'Interdisciplinarité en question dans les études Anglophones. Presses Universitaires de Nancy, Nancy, pp. 15–29.

Pocock, J.G.A., 1985. Introduction: the state of the art. In: Pocock, J.G.A. (Ed.), Virtue, Commerce and History: Essays on Political Thought, Chiefly in the Eighteenth Century. Cambridge University Press, Cambridge, pp. 1–34.
Pocock, J.G.A., 1987. The concept of a language and the métier d'historien: some considerations on practice. In: Pagden, A. (Ed.), The Languages of Political Theory in Early-Modern Europe. Cambridge University Press, Cambridge, pp. 19–38.
Pocock, J.G.A., 2009. Political Thought and History: Essays on Theory and Method. Cambridge University Press, Cambridge.
Rabinow, P. (Ed.), 1991. The Foucault Reader. Penguin, Harmondsworth.
Richter, M., 1987. Begriffsgeschichte and the history of ideas. Journal of the History of Ideas 48, 247–263.
Skinner, Q., 2002. Visions of politics. In: Regarding Method, vol. I. Cambridge University Press, Cambridge.
Tully, J. (Ed.), 1988. Meaning and Context: Quentin Skinner and His Critics. Cambridge University Press, Cambridge.
Vincent, J., 2003. Concepts et Contextes de l'Histoire Intellectuelle Britannique: l'Ecole de Cambridge à l'Épreuve. Revue d'Histoire Moderne et Contemporaine 50, 187–207.
Young, B., 2006. Introduction. In: Whatmore, R., Young, B. (Eds.), Palgrave Advances in Intellectual History. Palgrave, Basingstoke-New York, pp. 1–7.

Intellectual Property Rights: Ethical Aspects

Joseph Straus, Max Planck Institute for Innovation and Competition, Munich, Germany; and NIPMO-UNISA Chair for Intellectual Property, University of South Africa (UNISA), Pretoria, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the main notions and categories of intellectual property. Some of the specific conceptual problems of intellectual property are highlighted, such as that they refer to intangible objects. Intellectual property rights are shown to have deep historical roots (e.g., in questions of book authorship), yet their role has also changed substantially due to scientific, technological, social, economic, and political developments. The article highlights how new questions of protecting genetic information and leaving material (including higher life forms) are blurring the boundaries between nonpatentable discoveries and patentable inventions. Major institutional reforms of intellectual property rights, such as those effected by the Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPS) under the World Trade Organization, are raising new questions about how to deal with intellectual property rights fairly at an international level.

'Intellectual property rights' are generally understood as an instrument of economic policy providing incentives for and rewarding a broad range of useful human creativity. Despite this prevailing economic rationale, the roots of, the justification for, and some limitations of exploitation of these rights are reasoned ethically. They will be examined in the context of the notion, economic importance, and implications deriving from scientific and international policy developments.

The Notion of Intellectual Property Rights

As a generic term, intellectual property rights stand for patents relating to inventions in all fields of human endeavor; copyright covering literary, artistic, and scientific works; neighboring rights dealing with performances of performing artists, phonograms, and broadcasts; trademarks, service marks, and commercial names concerned with signs distinguishing goods and services as well as economic subjects; industrial design protecting novel and original ornamental or aesthetic aspects of a useful article; or plant breeders' rights relating to new varieties of plants. Conceptually, intellectual property rights can be understood as rule-governed privileges that regulate the ownership and exploitation of objects which are to be characterized as intangible or incorporeal, thus abstract or intellectual. These objects are nonexclusive, since they can be in many places at once, and can be used at the same time without being consumed by their use. The possession or use of such an object by one person thus does not preclude others from possessing or using it as well. Despite their common fundamental character, i.e., being rights in intangible, abstract objects, and being territorially limited, they differ from each other in many legal details and character.

Whereas a patent also gives its owner rights against the independent creator of a similar invention, copyright confers rights against copying but does not affect the independent creation of a similar work. In other words, while patents restrict the actual usage of an idea, copyrights restrict only the copying of an expression of an idea. Necessarily, the period of protection, i.e., during which the rightful owner can prohibit unauthorized use, is tailored according to the specification of the respective rights – for instance, 20 years from filing the application in case of patents, 50 or 70 years after the death of the author in case of copyright, or renewable at will, i.e., indefinite, in case of trademarks. However, neither patents, nor any other intellectual property rights, confer on their owner a license to use the patented invention, copyrighted work, etc. Such use is prohibited whenever it would violate, for instance, laws dealing with public health, safety, environmental protection or certain ethical standards, or would infringe the rights of third parties.

Economic Importance

Due to enormous advances in science and technology, in particular in communication and information technologies, biotechnology, and life sciences in general, on the one hand, and the global economic policy developments following the conclusion of the GATT-Uruguay Round and the establishment of the World Trade Organization in 1994, on the other, intellectual property rights have gained the most remarkable economic importance. They have penetrated, or are about to penetrate, new areas of economy, such as the service sector, life sciences, and agriculture. Patents have been granted for instance, for microorganisms, plants and animals, genes, as well as computer programs, and business methods. Universities and other academic institutions, primarily aimed at generating and disseminating new knowledge, have also not remained unaffected. Consequently, the new, knowledge-based economy has been characterized as 'intellectual property economy' or 'intellectual capitalism.' This trend is reflected, for instance, by the number of patent applications filed worldwide, which doubled from approximately 1.05 million applications in 1995 to around 2.35 million applications filed by 2012 (WIPO IP Facts and Figure 2013); by the amount of royalties paid for patent licenses, which increased in the USA from US$3 billion in 1980 to over US$100 billion in 1997 (Rivette and Kline, 2000); or by the fact that the ratio of intellectual capital to total market capitalization of US companies grew

from about 17% of company value in 1975 to 80% in 2005 (Blaxill and Eckardt, 2009). Also remarkable are the numbers of US patents issued to US universities, which went up from 177 in 1974 to 5145 in 2012, the year in which US universities filed some 14 224 new patent applications and had an income of US$1.9 billion from running royalties (Straus, 2008; AUTM, 2013).

Historic Roots and Rationale

The historic roots of patents and copyrights, as the two forms of intellectual property with the longest tradition, go back to the systems of inventors' and printers' privileges at the close of the Middle Ages. Inventors' privileges formed part of the industrial policy of regional sovereigns, being aimed at promoting domestic crafts and industry, and at transferring to their own country the production of goods which had previously been imported from abroad; thus, as expressed in a grant issued by Edward III in 1331, *instruendi et informandi causa*, i.e., for the purpose of instructing and informing. The first statutory regulation, the Venetian Statute of Inventors, of 1474, with its time-limited protection of inventions against imitation, was also intended to stimulate 'deep thinkers and ingenious inventors from other regions to exercise their activity to the advantage of Venice.' Similar economic rationale is to be observed in the English Statute of Monopolies of 1624, under which the grant of patent monopolies of 14 years' duration was restricted to the 'real and first inventor,' and was conditional upon its not being contrary to law or 'mischievous to the state' in any way.

Following the invention of the press, the first printers' privileges were granted in the second half of the fifteenth century in Venice, later followed by Milan, and at the beginning of the sixteenth century by other Italian cities, France, Germany, and England. Whereas in Venice and France, some of these privileges were granted to authors, the great majority of privileges were for publishers, or, as in England, for stationers, who acquired a work from its author and organized its printing and sale. In England, where the stationers in 1534, gained protection secured by the Crown against the importation of foreign books, and in 1556, a Charter from Queen Mary, the privileges were combined with censorship over all publications, by mandatory registration with the Stationers' Company. Although the first Copyright Act, issued in 1710, in the reign of Queen Anne (the Statute of Anne), conferred the 'sole right and liberty for printing books' for 14 years to authors and their assigns, it did not change fundamentally the economic rationale of protection.

By naming booksellers and printers as falling among the author's assigns, it demonstrated that the right "stemmed from commercial exploitation rather than literary creation pure and simple" (Cornish and Llewelyn, 2003). Despite the remarkable influence which John Locke's theory – that all persons have a natural right to hold property, particularly the fruits of their own labor – played from the very beginning in justifying intellectual property, the economic policy of legislation has remained utilitarian. This is well demonstrated by Art. 1, Section 8 of the US Constitution of 1787, providing that "The Congress shall have power … (8) To promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries." This is equally true for the first modern patent laws which were passed in 1790 in the US and 1791 in France, although they strengthened the position of the inventor and awarded him a property right to his ideas. The same rationale controlled also the first US Copyright Statute of 1790.

The doctrine of intellectual property developed in eighteenth-century France, advocated primarily by D'Héricourt, Diderot, and Voltaire as a weapon against extended privileges of publishers. A first true statutory shift of the emphasis from the publisher to the author can be observed in the Decrees of the French revolutionary legislation of 1791 and 1793, which extended the term of protection to the entire life of the author and beyond; first for 5, then for 10 years after the author's death. In the course of the nineteenth century, France further extended the term of protection to cover also the life span of the author's widow, and 20 (later increased to 30) years for the benefit of the children. Eventually, in 1866, the duration of French copyright was extended generally to 50 years after the author's death. Even this shift of emphasis, however, did not add to the economic rights of the author any 'personality right' or *droit moral*, which had been first mentioned by Immanuel Kant. Kant stated that the right of the author "is not a right in an object (*in re*), i.e., in the copy (for the owner can burn this up before the author's eyes), but an innate right, inherent in his own person, which implies the possibility of opposing the attempt of another person to compel him to speak against his will." Kant characterized copyright as a supremely personal right (*jus personalissimum*) (Strömholm, 1983).

French and German legal doctrine and French case law in the second half of the nineteenth century eventually paved the way for *droit moral*: first, into the case law, and then into copyright statutes of the principal Continental countries. Although *droit moral* was introduced into the Berne Convention for the Protection of Literary and Artistic Works of 1886, as revised in 1928, in Rome, with a new Article 6 bis – which at present guarantees an author independent economic rights, and even after their assignment, "during his whole lifetime the right to claim authorship of the work and to object to any distortion, mutilation or other alterations thereof, or any other action in relation to the said work, which would be prejudicial to his honour or reputation" – the common law countries of the UK and the USA remained reserved in this respect. In general, therefore, one distinguishes between 'author's rights' of civil law countries, with their emphasis on the principles of natural rights, and the 'copyright,' rooted in the idea of utilitarianism, of common law countries. However, on balance, it was neither human rights nor copyright doctrine that decisively shaped the law, but economic interests of copyright owners, because: "In the world of copyright policymaking, a Canadian book publisher has far more in common with a Japanese book publisher than it does with a Canadian librarian who wants to make free copies for library patrons" (Goldstein, 2001).

International Legal Developments

The very nature of the objects to which intellectual property rights relate, and the development of international trade combined with the progress of transportation and

communication technologies, already by the nineteenth century revealed the need for international instruments for protecting intellectual property rights. What started in 1883, with the adoption of the Paris Convention for the Protection of Industrial Property, followed by the Berne Convention for the Protection of Literary and Artistic Works in 1886, – which established the principles of national treatment or nondiscrimination and minimum rights as binding standards for the contracting states – subsequently developed into a true network of international conventions and treaties covering the entire range of intellectual property rights under the administration of the World Intellectual Property Organization (WIPO).

However, membership in many important conventions long remained limited, and minimum rights incomplete or, as in the case of patents, nearly nonexistent. The turning point came when, during the GATT-Uruguay Round negotiations, issues of international trade, including membership in the World Trade Organization (WTO), were linked directly to an effective protection of intellectual property rights, and the International Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPS) was adopted in 1994. Under this new umbrella for international protection of intellectual property rights, all members of WTO accepted unified high standards of protection, including protection requirements, content, and limitations and duration of rights as well as their enforcement. Developing countries, countries in transition, and the least-developed countries were granted transition periods.

Old and New Ethical Aspects in Intellectual Property Rights

Traditionally, ethical aspects in intellectual property rights have been discussed in the broad context of justification of these rights. Lockean theory of natural rights to the fruits of one's labor, the doctrine of intellectual property, as well as Hegel's personality justification suggesting that the best way of progressing science and the arts is to protect scientists and artists from theft, so as to become the basis for learning by others, are such examples, and had some lasting influence on legal developments, reflected, inter alia, in broad acceptance of *droit moral*. Modern critics of intellectual property rights also put emphasis on ethical aspects when, for instance, claiming that increasing scope of patentable subject matter or legislatively creating new forms of abstract objects, such as plant variety rights, constitutes effectively the creation of capital, which has the danger that it can act as an enormous power resource for a select few. Furthermore, a global protectionist scheme for intellectual property would help to promote the formation of global factions, resulting in the danger of global rent seeking, and perhaps providing temptation to multinational elites to increase their profits through the simple stratagem of persuading a supranational body to elevate levels of protection for abstract objects already in existence (Drahos, 1996). With such reasoning, neither the ethics of the controlling doctrine nor the ethical aspect of attempts to take care of the developing world and the unprivileged in general can be overlooked. The same is true for the debate on intellectual property rights in genetic resources and benefit sharing. Although clear empirical evidence whether intellectual property rights confer a net benefit or a net loss upon society is still missing, modern lawmakers seemingly accept a limited and flexible intellectual property rights scheme as a reasonable and, as revealed by the adoption of TRIPS, internationally necessary effort to maximize benefits while minimizing costs. This attitude is demonstrated by the adoption of statutory rules or acceptance of case law, creating new objects such as DNA sequences, monoclonal antibodies, or for instance, business methods as subject matter of protection. In their overall positive attitude toward exclusive intellectual property rights modern lawmakers might have been in part influenced by the negative experience with some alternative protection found in the former socialist states such as 'investors' certificate' which did not provide for exclusivity and had no beneficial macroeconomic effects.

Special attention is paid to ethical aspects when patentability of inventions is at hand, the commercial exploitation of which would be contrary to '*ordre public*' or 'morality.' Under the TRIPS standards, members of the WTO "may exclude from patentability inventions, the prevention within their territory of the commercial exploitation of which is necessary to protect *ordre public* or *morality*, including to protect human, animal or plant life or health or to avoid serious prejudice to the environment, provided that such exclusion is not made merely because the exploitation is prohibited by their law." Thus such exclusions may only apply for ethical reasons, first, if commercial exploitation is prohibited, and, second, if they are prohibited for specifically qualified reasons, which need not necessarily be set out in statutes, however. European patent law reflects these ethical considerations by excluding from patentability the human body, at various stages of its formation and development, including germ cells, processes for cloning human beings, processes for modifying the germ line genetic identity of human beings, uses of human embryos for industrial or commercial purposes, processes for modifying the genetic identity of animals, which are likely to cause them suffering without any substantial benefit to man or animal, and in general all therapeutic methods. By declaring, explicitly in principle, that biological material, even that of human origin (provided the prior informed consent of the source person is obtained), as well as animals and plants (except varieties) is patentable, European lawmakers demonstrated clearly that they do not share respective ethical objections against patentability of life forms. This ethically reasoned approach, which differs in particular from that applied in the USA, where no such exclusions exist, however itself raises ethical concerns. On the one hand, it seems consistent that inventions, which for reasons of ethics may not be commercialized, should also not be patented, although an issue of a patent does not confirm more than that the usual patentability requirements of novelty, inventiveness (nonobviousness), and industrial applicability (utility) were met and the invention adequately disclosed. It does not contain any other value judgment. On the other hand, the invention under such circumstances cannot be commercialized in any case, as long as its exploitation is prohibited. However, if such a ban were to be lifted, anybody could use the invention, and the inventor would remain empty-handed. As history demonstrates – for instance, with contraceptives – this is not a far-fetched scenario and could easily happen again,

e.g., with regard to specific uses of stem cell technology, germ line gene therapy, etc. Empty-handed inventors, and at least to the same extent the lack of incentives for investment in potentially useful research, may be viewed as ethically doubtful results.

It should be noted, however, that under the case law of the Court of Justice of the European Union (Judgment of 18 October 2011 in Oliver Bruestle v. Greenpeace e.V.), inventions related to human embryonic pluripotent stem cells, generated from destroyed human embryos are excluded from patent protection under the aspect of ethics. This is true even though the actual working of the claimed invention does not require any further use of human embryos and the generation of stem cell lines from human embryos was fully in compliance with the applicable law. Also, the fact that products generated from human embryonic pluripotent stem cells can be commercialized in a number of EU Member States, does not matter. This raises not only questions of TRIPS compliance of the respective case law, but also of ethics in view of the fact that without patent protection free copying is allowed. Finally, mention should also be made of the case law of the US Supreme Court (Opinion of 13 June 2013 in Association for Molecular Pathology et al. v. Myriad Genetics, Inc. et al.), under which genomic DNA sequences are declared non-patentable as a naturally occurring product, even as isolated DNA segments. This indirectly reflects the ethical consideration that discoveries cannot be treated as patentable subject matter.

See also: Genomics, Ethical Issues in; Innovation and Technological Change, Economics of; Intellectual Property: Law in Context; Patenting in the Biological Sciences; Property Rights, History of; Property: Anthropological Aspects; Property: Legal Aspects; Rights: Legal Aspects; Social Work Ethics.

Bibliography

AUTM, 2013. US Licensing Activity FY 2012 Highlights. www.autm.net.

Beier, F.K., 1980. The significance of the patent system for technical, economic and social progress. International Review of Industrial Property and Copyright Law 11, 563–584.

Blaxill, M., Eckardt, R., 2009. The Invisible Edge. Portfolio, New York.

Cornish, W.R., Llewelyn, D., 2003. Intellectual Property: Patents, Copyright, Trade Marks and Allied Rights, fourth ed. Sweet & Maxwell, London.

Drahos, P., 1996. A Philosophy of Intellectual Property. Dartmouth, Aldershot, UK.

Goldstein, P., 2001. International Copyright—Principles, Law and Practice. Oxford University Press, Oxford, UK.

Grandstrand, O., 1999. The Economics and Management of Intellectual Property—Towards Intellectual Capitalism. Edward Elgar, Cheltenham, UK.

Judgment of the Court of Justice of the European Union of 18 October 2011 – Case C-34/10, Oliver Bruestle v. Greenpeace e.V.

Kitch, E.W., 1977. The nature and function of the patent system. Journal of Law and Economics 20, 265–290.

Landes, W.M., Posner, R.A., 1989. An economic analysis of copyright law. Journal of Legal Studies 18, 325–353.

Machlup, F., 1959. An Economic Review of the Patent System. U.S. Govt Printing Office Study No. 15, Subcommittee on Patents, Trademarks and Copyrights, Senate Committee on the Judiciary, 85th Congress, 2nd Session, Washington, DC.

Moore, A.D. (Ed.), 1997. Intellectual Property—Moral, Legal and International Dilemmas. Rowman & Littlefield, Lanham, MD.

Opinion of the Supreme Court of the United States of 13 June 2013 – No. 12-398 – Association for Molecular Pathology, et al., Petitioners v. Myriad Genetics, Inc. et al.

Rivette, K.G., Kline, D., 2000. Rembrandts in the Attic. Harvard Business School Press, Boston, MA.

Schricker, G. (Ed.), 2006. Urheberrecht—Kommentar, third ed. Beck, Munich, Germany.

Straus, J., 2008. Intellectual property versus academic freedom? A complex relationship within the innovation ecosystem. In: Engwall, L., Weaire, D. (Eds.), The University in the Market. Portland Press, London, UK, pp. 53–65.

Straus, J., 2013. Ordre public and morality, issues in patent eligibility. In: Takenaka, T. (Ed.), Intellectual Property in Common Law and Civil Law. Edward Elgar, Cheltenham, UK; Northampton, MA, pp. 19–49.

Strömholm, S., 1983. Droit moral—the international and comparative scene from a Scandinavian viewpoint. International Review of Industrial Property and Copyright Law 14, 1–42.

Troller, A., 1983. Immaterialgueterrecht, third ed., 2 vols. Helbing & Liechtenhahn, Basel, Switzerland.

Ulmer, E., 1980. Urheber- und Verlagsrecht, third ed. Springer-Verlag, Berlin.

WIPO, 2013. IP Facts and Figures – Economics & Statistics Series Geneva, Switzerland.

Intellectual Property: Law in Context

Christoph B Graber and Jessica C Lai, University of Lucerne, Lucerne, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Intellectual property (IP) law is often viewed as static, when it is actually chameleonlike, constantly changing to reflect technological and social developments. This article discusses how IP as we know it today – in the classical fields of copyright, patent and trademarks, and in their interaction and beyond – has developed alongside changes in technology, globalization, and society, taking into account their interrelated nature and how they spur propertization and affect the balancing of interests (both economic and noneconomic). It specifically looks at challenges relating to software and the Internet, biotechnology, and overlaps in IP, before taking a brief look at what the future may bring.

Introduction: IP Law as Contingent on the Technosocial

Analyzing intellectual property (IP) law in context invites one to view IP as a protean concept. IP is thus perceived as a set of rules that societies have been developing to regulate access to cultural and intellectual knowledge that is contingent on social and technological development. We should bear in mind that IP rights are an invention of modern society; traditional and particularly indigenous societies have developed different instruments of regulation that are often more concerned with the restriction, rather than the transfer, of rights in cultural heritage assets. Indigenous societies do not conceive of their cultural heritage in terms of exclusive rights but rather in terms of shared responsibilities, as assets to look after for future generations. Generally, they find the commodification of secret and sacred ritual objects and practices to be deeply offensive (Graber, 2009).

The technosocial contingency is obvious in the history of copyright law. There was no need for copyright law or the recognition of 'authors' before technologies were available that made the mechanical reproduction of creative artifacts possible at relatively low cost (Cornish, 2009). It was only the development of tangible commercial value in the form of a book that generated the need to identify its source (Anderson, 2009). The *Statute of Anne* (1709) was the first copyright statute to provide for an individual right to print. Before that the English Crown regulated access to printed information by bestowing selective royal grants of privileges on those craft guilds that respected the Crown's censorship standards (Drahos, 1996).

The development of modern patent law is inextricably entwined with the paradigm of industrial manufacture (Pottage and Sherman, 2010). Although statutory patents have been around since the fifteenth century, the patent rocket only started to take off at the end of the eighteenth century as a response to the industrial revolution (MacLeod, 1988; Pottage and Sherman, 2010), and started to soar in the twentieth and twenty-first centuries with the advent of mass chemical and pharmaceutical production, followed by information technology and biotechnology.

Similarly, the evolution of trademark (TM) law is a response to the development of marketing strategies, communication technologies, and mass media, which created the need to regulate the commercial use of signs (in a very broad sense, nowadays also including colors, sounds, gestures, shapes, and even smells) through a set of specific access rules. As with patents, the demand for protection of marks increased during the industrial revolution and the rise in factory production and greater transport of products. This demand increased again in the twentieth century with the growth of the popular press as a medium for advertising and with the increased ability to transport goods, intensifying competition. More recently, TM protection has swelled with the advent of globalization and technological developments, enhancing large-scale marketing efforts (Cornish et al., 2010).

Theoretical Approaches to IP Law

IP rights are rights in abstract objects (Drahos, 1996); they are rights in *corpora mystica* rather than rights in *corpora mechanica* (material objects). The question that modern society had to answer is why the law of the state should vest exclusive rights in abstract objects. Most influential has been the answer derived from a 'productive misunderstanding' of John Locke's natural law theory of property. At the end of the seventeenth century, Locke argued that "every Man has a Property in his own Person" (Locke, 1994). Therefore, according to Locke, it follows that every man also has property in his labor. Consequently, everything that the individual mixes his labor with is removed from the state of nature and becomes the individual's property (Locke, 1994). Although Locke was writing about material objects and deliberately did not extend his theory to support a perpetual right of authors, through adaptation by William Blackstone in the eighteenth century, this chain of reasoning became the philosophical justification also for property in abstract objects (Deazley, 2004).

Although Adam Smith had written about the effects of regulation on the economy as early as the eighteenth century, it was only in the 1960s that, predominantly in the United States, law and economics set out to establish itself as a new subdiscipline of legal theory (Menell, 2000). Under the growing influence of economic rationality in most areas of law and policymaking, a market-oriented foundation of IP rights gained ground as an alternative to an individualistic (natural rights-based) IP theory. From the perspective of political economy, IP

rights are monopolies vested by the law of the state in right holders to exclusively exploit the right for a certain (in the case of TMs, unlimited) time period. According to a utilitarian rationality, such monopolies are justified because they are deemed necessary incentives for creative people to invest time and money in their creations and share them with other people. Law and economics theories today strongly dominate discourse about IP law, especially as a theoretical foundation of the intersection of competition law and policy. Although some of this scholarship is very sophisticated and well aware that IP law tends to favor private over public interests, it is not able to integrate societal values – like culture, health, privacy, and so on – that compete with economic values and interests. Hence, an economic approach is capable of warding off critique that a given IP system fails to maximize utility for society, but only so long as the critique stays within the internal standpoint of classical economic rationality. Approaching utilitarianism from a moral or ethical perspective may have promise (Singer, 2011; Pogge, 2010); however, this has yet to filter into general IP theory.

Although there is a need for theoretical approaches to IP that are able to embrace noneconomic societal rationales, scholarship that is not confined to economic thinking or to an individualistic natural rights theory is scarce – particularly in the Anglo–American academic world. Influenced by legal applications of sociological system theory, however, new scholarship is emerging that views economic values of IP in a quasiconstitutional relationship with other fundamental values of society (see below, Section Constitutionalization of the Global IP Regime).

Propertization

Societies that justify IP rights with a (pseudo) Lockean theory of labor and natural right tend to overly propertize intellectual and cultural assets to the detriment of the commons (Drahos, 1996). In recent years, we have faced a paradigm shift from a presumption that 'government-granted IP rights are an exception from the general rule of free competition' to a new presumption that every new technological or cultural creation should be legally protected (Merges, 1995). This shift is evidenced by the increasing number of international agreements pertaining to IP, by the higher standards of protection advanced within the World Intellectual Property Organization (WIPO), and, most pertinently, by the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) of the World Trade Organization (WTO). With its broad membership, and by virtue of being binding and strongly enforceable through the WTO dispute resolution system, TRIPS has resulted in the propertization of intellectual assets in states that previously did not have similar IP systems and has allowed established IP states to extend the reach of their property arms into these states.

More specifically with regard to copyright, what constitutes a 'work' has extended beyond traditional literary and artistic works to include, for example, software, fictional characters, and television formats. Rights are broader, encompassing more than the rights to reproduction, communication to the public, translation and adaptation, expanding to distribution, and rental rights. Finally, although international law only requires that works be protected for the life of the author plus 50 years, many jurisdictions (such as the United States, the European Union, and Australia) have extended this posthumous protection to 70 years. Whether permitted uses of third parties (such as for teaching purposes or academic research) have concurrently been widening is contentious, raising issues of the ultimate balance between promoting creativity and allowing for future expression.

Since the early patent laws, the term of patent protection has risen from 14 to 20 years and the scope of the patentable 'invention' has grown to controversial proportions. For example, before the mid-twentieth century, it was not possible to patent chemical products per se in many states. These days, one does not blink an eye at the prospect. The development of new technologies, such as software and biotechnology, are the locus of contemporary controversies in patent law. The seeking of patents by universities and publicly funded research institutes has also become common since the World War II, whereas it had previously been viewed as contrary to the academic environment of such institutions.

Propertization in TM law is reflected in a conceptual shift from preventing fraud or consumer deception to protecting the property in signs used in trade (Bently, 2008; Lunney, 1999; Lemley, 1999). TMs originated from the courts within tort law as an unfair competition tool to prevent a trader from selling goods represented as those of another trader (Landes and Posner, 2003; Bently, 2008). The solidification of TMs as 'property' occurred in the mid-late nineteenth century (Bently, 2008). Although TMs continue to serve to minimize consumer confusion and search costs, and implicitly to indicate certain qualities of the goods on which they are used, these days, the expanding tendencies of TM protection means that, in some cases, reputation and goodwill of marks are protected even when there is no consumer confusion (Landes and Posner, 2003), controversially creating property more akin to copyright and patents (Pulos, 2005–2006). For example, the Paris Convention (Article 6bis) and TRIPS Agreement (Article 16.2–3) together require the protection of 'well-known' marks for goods and services, essentially protecting reputation/goodwill against 'dilution,' or the blurring or tarnishing of the repute of a well-known mark. This reasoning steps away from market concerns, toward protecting TMs as a thing capable of being damaged.

New drivers of propertization are complex things (more fundamental or informational) that could be perceived as being 'technologically normative,' thus adding a layer to the law's normativity. For example, patents can be obtained for genetic code, proteins (complex chemical compounds), or self-terminating seeds, which themselves produce 'technologically normative' effects irrespective of legal regulation. Genes define what chemical/biological structures a specimen can make. Proteins are catalysts that control the reactions of other simpler chemical compounds, determining the end product of a reaction. Finally, crops from self-terminating seeds can only be harvested once, the new seeds being sterile. Copyright protects Internet/software code, which inherently directs the flow of information and what can be done with content on the Internet and with hardware (Graber, 2012a). Furthermore, many copyright laws penalize the circumvention of technical

measures protecting copyrighted works that themselves decide how a work can be used irrespective of existing fair use rights. TM law broadening to protect three-dimensional shapes, possibly the shape of goods per se, has created problems when the shape is functional. These examples have attracted controversy because owners can be seen as being endowed with a combination of legal and technological normativity.

Balance

Balancing exclusive interests of individual right holders against inclusive interests of the general public has always been an important aim of IP law. In theory, a balance within IP law secures that the public benefits overall. In practice, however, the protection of IP rights has the effect of maximizing returns of private right holders at the expense of interests of the public at large. Technological development of the recent past as well as propertization has led to a considerable inflation of the private benefits that can be gained from IP monopolies to the detriment of public interests.

One of the core underlying principles of copyright law is that it protects expressions and not ideas. The ideas/expressions dichotomy is pivotal to the balance between stimulating expression and allowing for users to build off existing works, as ideas are considered to be part of the raw resources of the public, usable for further creativity. This is in part why copyright law is generally considered to be compliant with the constitutional right to freedom of expression and information, because the dichotomy acts as an internal check for the freedom (the 'internalization theory') through limiting use of only exact expression and not of ideas held within works (Torremans, 2004; Barendt, 2005). The idea/expression dichotomy is often painted as black and white, but is actually not as well-delineated as inferred by the internalization theory (Cohen, 2012) and is not on its own capable of protecting freedom of expression (Oliver, 2000; Barendt, 2005). As an illustration, the dichotomy may have little relevance in a networked environment. Take, for example, mash-ups. Users are free to use the ideas behind media works, but actually it is the snippets of expressions that they need to use in order to enjoy their freedom of expression. In fact, the ideas in the original work may be completely irrelevant to their purposes. Thus, although the internalization theory is generally questionable, it is clearly incorrect when it comes to copyright protection on the Internet.

As patent law is based on the premise that it incentivizes invention/innovation, it must be designed so as not to impinge too greatly on future invention. The incentive must be great enough to stimulate invention/innovation, but not so great as to prevent other inventors from creating based on existing inventions. As part of this balance, inventions must be novel, nonobvious (with an inventive step), and have utility (industrial application). It has also long been established that laws of nature, natural phenomena, abstract ideas, and mere discoveries are not 'inventions,' because they make up the basic tools of scientific and technological work that should be available for all to build upon, such that to grant patents over them would impede invention rather than promote it. The development of new technologies such as software, Internet business methods, as well as gene- and life-related technologies has challenged the notion of the discovery/invention dichotomy, because they are closer to (if not actually) natural phenomena or abstract ideas compared to classical inventions, such as purely mechanical machines or methods of producing chemicals or admixtures.

TMs protect balance through only protecting marks capable of distinguishing a trader's goods/services from another's, or marks that have gained such distinctiveness in the market, thus theoretically leaving the generic free in the commons. TM laws further seek balance through only protecting owners from third-party 'use in the course of trade,' which usually means that the mark is used in a TM sense to distinguish goods/services. In theory, these limitations function to respect freedom of expression and the requirement of 'use in the course of trade' particularly to restrict TM owner's rights to those required to prevent unfair competition (Helfer and Austin, 2011). For example, with a book criticizing McDonald's, the TM golden arches should be allowed to be shown on the cover, as not being 'use in the course of trade.' As with patent and copyright laws, TM law also faces challenges with the development of information technology. TMs may be needed to perform Internet searches, for example. Arguably, marks that may be distinctive in the real world may have more of a generic (or at least different) function when only used within code, not even visible to users. The nature of the technology makes it difficult to say what exactly 'use in the course of trade' means within the online environment.

Globalization

The Problem of Territoriality

Despite the increasing number of international and regional agreements, IP law is inherently territorial. The question then is how state-centered legal orders of IP are able to cope with the challenges of economic globalization. Copyright law is particularly challenged by the digital revolution and the Internet, and a territory-based legal structure is difficult to reconcile with consumers expecting easy access to digital content at all times, anywhere and through a multitude of devices. However, ambitions to get rid of copyright territoriality for the sake of creating larger content markets may easily clash with cultural policy objectives. The European Commission's effort to create a single market for copyright-protected products is a case in point uncovering collisions between economic interests of the European Union and cultural interests of member states. The Commission's conviction that Collective Management Organizations should compete in an EU-wide market for the administration of copyrights has been criticized as being one-sidedly economy oriented and ignoring important aspects of cultural diversity (Graber, 2012b).

Territorial thinking in patent law is a legacy of a time when it was difficult and sometimes even dangerous to bring an invention into a country. The trials undertaken to import an invention meant that the initial focus was on the societal benefits of the invention itself (i.e., the physical entity). With globalization and the ease with which we now move the physical, the focus has shifted to the contribution of the information behind the invention (Chisum et al., 2001), and

patent systems encourage inventors to publish the information behind their inventions that they might otherwise keep as trade secrets (Landes and Posner, 2003). That several states continue to have territorial novelty standards (as opposed to absolute standards) reflects a continuance of territorial thinking despite globalization and the shifted focus, and also despite technological developments in communication.

TM law is arguably the most territorial form of IP. This is because whether a mark is distinctive completely depends on the national market. New technologies of transport and communication and globalized business strategies challenge territoriality in TM law. What may be distinctive in one state may be generic in another in juxtaposition to the international nature of modern business. Territoriality is also highlighted by the protection of geographical indications (GIs) (e.g., Champagne) through TM-like regimes. Tension has arisen between 'old world' Europe (along with some other countries, such as India, Thailand, and Turkey) wishing to protect 'traditional' names and associated values even when there is no consumer confusion and 'new world' Americas and other settler states (such as Australia and New Zealand), which view GIs as protectionism, since they tend to be made up of multiple cultures wherein many names/symbols have become generic and are used in a descriptive manner.

There are examples of a movement away from territoriality, such as the development of the European Patent Convention and Community Patent and WIPO's global design registry.

Constitutionalization of the Global IP Regime

Recent scholarship has shown that the fragmentation of international law (UN, 2006) spurred the establishment of a global IP regime (Graber, 2014). This regime is partly formally institutionalized within WIPO, the WTO/TRIPS Agreement, and a plethora of free trade agreements (with IP chapters) and partly institutionalized outside a formal legal setting and without the involvement of states or intergovernmental organizations. Indeed, important drivers of the global IP regime are big transnational corporations of the entertainment and pharmaceutical industries in particular, whose activities may have regulatory effects and produce transnational regime law. Under the disproportionate influence of these powerful private actors, the global IP regime is developing regulatory mechanisms that are, to a certain extent, competing with the public interests of states. Within a theoretical framework that has become known as 'societal constitutionalism' (Teubner, 2012), Gunther Teubner and others have pointed out that world society is shifting from territorial differentiation to functional societal differentiation. Accordingly, the close structural coupling between state politics and the law that determines national constitutionalism does not exist at the transnational and global level, and regimes autonomously develop constitutive and limitative constitutional functions. With regard to the global IP regime, an analysis of societal constitutionalism shows that so far mainly constitutive functions have been developed and limitative elements are almost entirely missing (Graber, 2014). Public discontent with a 'neoliberal' spin on IP law and policy that is held responsible for this one-sided constitutionalization has recently become stronger and activism in social networks has been successful in mobilizing masses against projects of legislation destined to make IP enforcement more effective at the national and global level.

Specific Challenges

Software and the Internet

The question about the adequate legal protection of computer software arose only after IBM had taken the momentous decision in 1969 to unbundle software from hardware and to trade it as a separately priced asset (Bing, 2009). When the United States extended copyright protection to computer software in the 1980 amendment to the 1976 Copyright Act, this had a signaling effect on many other states to follow suit. Since the 1990s, TRIPS and the WCT have required contracting states to protect computer programs as literary works within the meaning of the Berne Convention for the Protection of Literary and Artistic Works. Scholars have criticized the copyright approach to software as mismatched and the analogy between programming language and literary language as farfetched. Whereas literary language may become the medium of a literary work due to its malleability, programming language's 'semantic' is structured as a linear relationship between orders and actions that leave no room for choice or creativity. Similarly, friction arises when distinguishing the (protected) expression in a computer program from the (unprotected) ideas and principles underlying them. Although courts in the European Union and the United States converge in the view that functional behavior of computer programs, programming language, and data formats affecting interoperability are treated as unprotected ideas and principles (Samuelson, 2012), the question whether application programming interfaces should be covered by copyright is currently a much disputed issue with far-reaching consequences for innovation in software.

Many states do not allow patents for software per se, as they are considered to be mere 'printed sheets,' instructions, or mathematical algorithms. However, software can usually be part of an invention. The exact nature of the interaction between software and hardware and whether the end result need be a 'technical effect' or can be merely informational is unclear and from the uncertainty one could say impossible to define. The United States Patent and Trademark Office has taken a very broad approach to what qualifies as an 'invention,' allowing patents that only claim steps for executing software on any computer platform and create informational results, but the courts have been inconsistent in upholding the validity of such patents. What will be accepted by the European Patent Office (EPO) is also nebulous, although it generally requires that a program directly affects the operation of hardware and results in a technical effect (Cornish et al., 2010).

In the late 1990s, 'cybersquatting' TM cases started to appear in the courts, regarding the practice of registering a well-known company domain name and trying to sell it back to the company. Courts were faced with the difficulty of finding 'use in the course of trade,' as the squatters did not actually use the webpages. The term was stretched to fit the new behavior. Concerns have also arisen with 'keyword advertising,' which is online advertising that appears as a result of keywords being entered into an Internet platform (Greene, 2009). Advertisers

buy keywords, often TMs, from the platforms so that their advertisements appear when those keywords are keyed in. For example, a computer retailer, service provider, or tech reviewer may wish to purchase the keywords 'Apple.' Indeed, Apple would also wish to purchase 'Apple,' as would perhaps its competitors. Complicated questions arise on whether there is 'use in the course of trade' (as per registered TM law), misrepresentation and associated loss or damage (for passing off), or consumer confusion (usually required to rely on consumer protection law), and the appropriateness of these terms in the Internet context.

Biotechnology

Biotech patents draw debate for several reasons. The first is that on moral or ethical grounds many do not agree with patenting life forms and gene-related technologies. The second pertains to the fact that such patents more directly affect the way that doctors treat their patients. In comparison to pharmaceuticals, which practitioners do not generally have the means to produce, many laboratories, hospitals, and academic medical centers are capable of making or using patented biotech inventions themselves (Eisenberg, 2002).

Third, biotech patents are often challenged as to whether they are 'inventions,' given their relative closeness to products of nature. What exactly constitutes a natural product or mere discovery, as opposed to an 'invention' is an open and difficult question, as all inventions in some way use and reflect natural phenomena or abstract ideas. In some situations, such as with new minerals, elements, or species found in nature, it is clear that there are only mere discoveries. The status of something found in nature, but extracted from its natural environment, is more difficult, for example, a chemical molecule extracted from a plant, wherein it is normally in a mixture with other molecules, or a genetic sequence extracted from a chromosome. Notably, both the EPO (EPO Guidelines, parts G.II.3.1 and G.III.4) and the US Supreme Court (*Association for Molecular Pathology v. Myriad Genetics, Inc.*, 2013) hold that simply isolated genes or genetic sequences are mere discoveries or natural products.

Finally, biotech patents are seen as too greatly hindering future invention, as they are often informational and fundamental in character. The shift in focus to the information behind an invention (as opposed to the physical invention itself) is problematic as it means that public access to the invention is not guaranteed. Historically, patents could often be struck from the register for nonuse; however, this is no longer the case in many jurisdictions. Patent owners are thus able to seek rent from their patents, although they themselves do not otherwise use them. This is particularly the case for more informational/fundamental inventions, which raises particular concerns because patents have in part been justified based on the costs of bringing a product to market.

Relatedly, there are serious questions about whether the incentive theory applies to all research in the biotech field, given its often fundamental nature. Yet, not all biotech inventions are fundamental and some may be costly to bring to market. Patent law need only incentivize that which would not take place free of its existence. Ultimately, we must ask at what point we should propertize in order to maximize invention/innovation, taking into account the realities of research and development.

There is a growing sentiment that patents have changed the biotech (and pharmaceutical) research arena, removing its collegial spirit. It is only in recent years that patents have been used to seek rent from public research institutes and universities. Formerly, such use was considered to fall under an informal exemption by both industry and publicly funded researchers, but the position of industry is changing. Currently, when universities and public research institutes hold patents, they usually only seek licenses or royalties from industry (Grubb and Thomsen, 2010). Empirical observation indicates that the long-standing social norm that academic researchers share their information and tools is slowly fading under the influence of IP (Eisenberg, 2009). This change may be having a negative impact on the incentive to discover. Arguably, no state has a robust experimental use exception that reflects the long-standing social norm that existed among academic researchers, although this would make more sense than depending on noncompliance and nonenforcement.

IP Overlaps

A nexus of protection exists between all three main fields of IP. The copyright–patent overlap is apparent from the discussion above relating to the protection of software by both patent and copyright laws. The overlap is even greater in states that allow for online business method patents, such as the United States. For example, a method whereby the consumer receives a copyrighted product in exchange for viewing an advertisement was allowed patent protection, meaning that claims for software with the terms 'over the Internet' or at 'an Internet Web site' are sufficient to transform an unpatentable abstract idea into a patentable process (Ultramercial v. Hulu, 2013; on appeal to the Supreme Court). Most states consider such patents to be the mere performance of mental processes and not 'inventions' (Cornish et al., 2010).

Copyright and TM protections can also overlap. Certain signs may be copyright protected if they are 'works' and 'original' (Helfer and Austin, 2011). A short phrase (such as Exxon, *Exxon Corp v. Exxon Insurance Consultants Int'l Ltd*, 1982) is unlikely to be considered a literary work, but a longer slogan could be, or a design may be an artistic work. Copyright exists without registration, such that it is possible for TM owners to spontaneously bring copyright infringement claims, creating an overlap between the two forms of IP. There are benefits because copyright protects any reproduction, not limited to use in the course of trade, or use as a TM. At the same time, there are advantages in using TMs over copyright, because TMs offer potentially perpetual protection. For example, Disney has both copyright and TMs over many of its characters. Jane Ginsburg has argued that, in some cases, an author's name can function like a TM, identifying source and indicating quality, and a sign can identify the copyright owner, such that we should be open to allowing overlap (Ginsburg, 2008).

The widening of TM protection to include three-dimensional shapes (the most famous probably being the Coca Cola bottle) has resulted in controversy because of the overlap with patent law. Shape marks are also examples of when all three areas of IP may be implicated. For example, the

'Toilet Duck' bottle shape (for a toilet bowl cleanser) was designed and patented in the mid-1980s by Walter Düring. Any designs or models for the bottle could also have design rights (if registered) or copyright. Düring also has a registered TM for the words 'Toilet Duck' and figurative marks. The patent has since lapsed. Had Düring sought a three-dimensional TM for the bottle shape, a monopoly on the shape of the bottle would still exist. Thus, TMs that cover shapes with functionality potentially give perpetual patent-like protection.

A Look to the Future

Thinking about IP as encompassing only the classical fields of copyright, patents, and TMs is, of course, narrow. New forms of IP or neighboring or related rights continue to emerge. For example, TRIPS requires TM-like protection of GIs for wines and spirits even when there would be no misled public or unfair competition, even if the true origin is indicated, and if the GI is used with expressions such as 'kind,' 'style,' 'imitation,' or 'like.' Furthermore, TRIPS requires patent-like protection of plant varieties and protection of industrial designs and integrated circuits, databases enjoy *sui generis* protection in certain jurisdictions, WIPO has treaties for the protection of performers, and producers of phonograms (the fixation of sounds other than in cinematographic or other audiovisual work), broadcasters, and pharmaceutical companies are pushing for data exclusivity protection. These reflect the move toward protecting investment and the global inertia of assigning property to anything with commercial value.

At the same time, propertization of IP is being countered by the Development Agendas of WIPO and the United Nations, and by prodevelopment nongovernmental organizations, civil liberties groups, consumer advocates, and 'pirate parties' fighting copyright, for example. A trend toward the interdisciplinarity of IP with the fields of human rights, public health, and competition law is also starting to push against the reach of IP.

Finally, recent years have shown an increasing demand by indigenous peoples for protection of their traditional knowledge, traditional cultural expressions, and genetic resources. Closely connected to this are concerns about the relationship between climate change and patent law innovation policies, and the impact of climate change on indigenous peoples' interests.

The outcome of these movements will reveal itself over time, but it is clear that IP is indeed protean.

See also: Indigenous Knowledge and Intellectual Property Rights; Patenting in the Biological Sciences; Property Rights and Urban Planning; Transaction Costs and Property Rights.

Bibliography

Anderson, J.E., 2009. Law, Knowledge, Culture. The Production of Indigenous Knowledge in Intellectual Property Law. Edward Elgar, Cheltenham, UK.

Association for Molecular Pathology v. Myriad Genetics, Inc. (2013) 569 US ___ (SC).

Barendt, E., 2005. Copyright and free speech theory. In: Griffiths, J., Suthersanen, U. (Eds.), Copyright and Free Speech. Comparative and International Analyses. Oxford University Press, New York, pp. 11–33.

Bently, L., 2008. From communication to thing: historical aspects of the conceptualisation of trade marks as property. In: Dinwoodie, G.B., Janis, M.D. (Eds.), Trademark Law and Theory: A Handbook of Contemporary Research. Edward Elgar, Cheltenham, UK, pp. 3–41.

Bing, J., 2009. Copyright protection of computer programs. In: Derclaye, E. (Ed.), Research Handbook on the Future of EU Copyright. Edward Elgar, Cheltenham, UK, pp. 401–426.

Chisum, D., Nard, C.A., Schwartz, H.F., Newman, P., Kieff, F.S., 2001. Principles of Patent Law, second ed. Foundation Press, New York.

Cohen, J.E., 2012. Configuring the Networked Self: Law, Code and the Play of Everyday Practice. Yale University Press, New Haven, MA.

Cornish, W., 2009. Conserving culture and copyright: a partial history. Edinburgh Law Review 13 (1), 8–26.

Cornish, W., Llewelyn, D., Aplin, T., 2010. Intellectual Property: Patents, Copyright, Trade Marks and Allied Rights, seventh ed. Sweet & Maxwell, London.

Deazley, R., 2004. On the Origin of the Right to Copy: Charting the Movement of Copyright Law in Eighteenth-Century Britain (1695–1775). Hart, Oxford, UK and Portland, OR.

Drahos, Peter, 1996. A Philosophy of Intellectual Property. Ashgate, Aldershot.

Eisenberg, R., 2002. Why the gene patenting controversy persists. Academic Medicine 77 (12), 1381–1387.

Eisenberg, R., 2009. Noncompliance, nonenforcement, nonproblem? Rethinking the anticommons in biomedical research. Houston Law Review 45 (4), 1059–1099.

EPO, 2012. Guidelines for Examination in the European Patent Office.

Exxon Corp v. Exxon Insurance Consultants Int'l Ltd [1982] RPC 69.

Ginsburg, J.C., 2008. Of mutant copyrights, mangled trademarks, and Barbie's beneficence: the influence of copyright on trademark law. In: Dinwoodie, G.B., Janis, M.D. (Eds.), Trademark Law and Theory: A Handbook of Contemporary Research. Edward Elgar, Cheltenham, UK, pp. 481–497.

Graber, C.B., 2009. Aboriginal self-determination vs the propertisation of traditional culture: the case of sacred Wanjina sites. Australian Indigenous Law Review 13 (2), 18–34.

Graber, C.B., 2012a. Internet creativity, communicative freedom and a constitutional rights theory response to "code is law". In: Pager, S.A., Candeub, A. (Eds.), Transnational Culture in the Internet Age. Edward Elgar, Cheltenham, UK, pp. 135–164.

Graber, C.B., 2012b. Collective rights management, competition, policy and cultural diversity: EU lawmaking at a crossroads. The WIPO Journal 4 (1), 35–43.

Greene, P., 2009. Keyword advertising, and other invisible uses of third-party trade marks in online advertising. Victoria University of Wellington Law Review 2, 1–106.

Graber, C.B. (2014) Is there potential for collective rights management at the global level? Perspectives of a new global constitutionalism in the creative sector. In: Susy Frankel, D.G. (Eds.), The Evolution and Equilibrium of Copyright in the Digital Age. Cambridge University Press, Cambridge, UK.

Grubb, P.W., Thomsen, P.R., 2010. Patents for Chemicals, Pharmaceuticals and Biotechnology: Fundamentals of Global Law, Practice and Strategy, fifth ed. Oxford University Press, Oxford.

Helfer, L.R., Austin, G.W., 2011. Human Rights and Intellectual Property. Mapping the Global Interface. Cambridge University Press, New York.

Landes, W.M., Posner, R.A., 2003. The Economic Structure of Intellectual Property Law. Harvard University Press, Cambridge, MA.

Lemley, M., 1999. The modern Lanham Act and the death of common sense. Yale Law Journal 108, 1687–1715.

Locke, J., 1994 [1689]. Two Treatises of Government. Cambridge University Press, Cambridge.

Lunney, G., 1999. Trademark monopolies. Emory Law Journal 48, 367–487.

MacLeod, 1988. Inventing the Industrial Revolution: The English Patent System, 1660–1800. Cambridge University Press, Cambridge.

Menell, P., 2000. Intellectual property: general theories. In: Bouckaert, B., de Geest, G. (Eds.), Encyclopedia of Law and Economics, The History of Methodology of Law and Economics, vol. I. Edward Elgar, Cheltenham, UK, pp. 129–188.

Merges, R.P., 1995. The economic impact of intellectual property rights: an overview and guide. Journal of Cultural Economics 19 (2), 103–117.

Oliver, J., 2000. Copyright, fair dealing, and freedom of expression. New Zealand Universities Law Review 19, 89–131.

Pogge, T., 2010. The health impact fund: better pharmaceutical innovations at much lower prices. In: Pogge, T., Rimmer, M., Rubenstein, K. (Eds.), Incentives for Global Public Health: Patent Law and Access to Essential Medicines. Cambridge University Press, Cambridge, pp. 135–154.

Pottage, A., Sherman, B., 2010. Figures of Invention: A History of Modern Patent Law. Oxford University Press, Oxford.

Pulos, M., 2005–2006. A semiotic solution to the propertization problem of trademark. UCLA Law Review 53, 833–869.

Samuelson, P., 2012. The past, present and future of software copyright interoperability rules in the EU and US. European Intellectual Property Review 34 (3), 229–236.

Singer, P., 2011. Practical Ethics, third ed. Cambridge University Press, New York.

Teubner, G., 2012. Constitutional Fragments: Societal Constitutionalism and Globalization. Oxford University Press, Oxford.

Torremans, P.L.C., 2004. Copyright as a human right. In: Torremans, P.L.C. (Ed.), Copyright and Human Rights. Freedom of Expression – Intellectual Property – Privacy. Kluwer Law International, The Hague, pp. 1–20.

Ultramercial, LLC v. Hulu, LLC (2013) No. 10-1544 (Fed. Cir.).

UN, 2006. Fragmentation of International Law: Difficulties Arising from the Diversification and Expansion of International Law. Report of the Study Group of the International Law Commission (Chaired by Martti Koskenniemi), UN Doc. A/CN.4/L.682.

WIPO, Berne Convention for the Protection of Literary and Artistic Works (1886) (adopted on 9 September 1886), as revised at Paris (24 July 1971) and as amended on 28 September 1979.

WIPO, Paris Convention for the Protection of Industrial Property, 828 UNTS 303 (adopted on 20 March 1883, entered into force 16 April 1970), as revised at Stockholm (14 July 1967).

WIPO Copyright Treaty, 36 ILM 65 (adopted on 20 December 1996, entered into force 6 March 2002).

WTO, Agreement on Trade-Related Aspects of Intellectual Property Rights, Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, 1869 UNTS 299; 33 ILM 1197 (adopted on 15 April 1994, entered into force 1 January 1995).

Intellectuals: History of the Concept

Christophe Charle, Professeur d'histoire contemporaine à l'Université de Paris I, Institut d'Histoire Moderne et contemporaine, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 11, pp. 7627–7631, © 2001, Elsevier Ltd., with updates by the Editor.

Abstract

A historical and comparative approach to the term 'intellectuals' or its equivalent in other languages shows the priority of the French use, in but also the persistent differences in its meaning even when translated in other contexts. The French model influenced other European countries, but it is only in Spain and Italy that the political mobilization of intellectuals was very similar to the French case. Whether in Russia (with the intelligentsia) or in German-speaking countries (with the Gebildete) or in the Anglo-Saxon world, the political, religious, and academic specifities remained very strong during the whole of the twentieth century.

The noun 'intellectuels' appeared in France in the 1890s and was largely diffused thanks to the polemics around the Dreyfus Affair (in particular the so-called '*Manifeste des intellectuels*,' published after the famous *J'accuse* paper by Zola issued on 13 January 1898, in *L'Aurore)*. This diffusion not only occurred in France but also, sooner or later, in the whole of Europe and even in America (North and South) (Drouin, 1994). There are two exceptions to this large diffusion of the new concept: Germany and Russia where there existed former and akin notions, *Intelligenz* and *intelligentsia*, already used as social or political denominations (Müller, 1971). Previous uses of the English translation, 'intellectuals,' are attested too before the 1890s, but they seem to have been too rare to have known a broad social circulation (Williams, 1976). In the same way, in America, 'intellectuals' became a common notion only in connection with newspaper commentaries about the Dreyfus Affair (Bender, 1987).

To understand why this new terminology received a European or even international diffusion, it is necessary to recall the issues of this historical moment. What was at stake in this crisis was not only a political problem but also the affirmation of a new group, defender of universal values against the reason of State (Charle, 1990). These values justified the fact that writers, artists, scholars, students, members of liberal professions, and so on, intervened on a collective basis in the political debate, although they were not themselves, for the main part, professional politicians. The other specificity of this moment was that, in other countries, this same cause or other similar ones favored the intervention of intellectuals, but generally in various patterns and with different contents. Since that period, social sciences have debated at a theoretical level, in order to find universal characteristics underlying this new vocabulary (Shils, 1972). As a matter of fact, the emergence of intellectuals cannot be assimilated to the apparition of a new permanent social group as some historians or sociologists suggest regularly, wrongly mixing intellectuals, intelligentsia and professions (Bell, 1973; Perkin, 1989). To know in reality what is meant by the words, intellectuals, *intellectuels*, *Intellektuelle*, *intellettuali*, *intelectuales*, and so on, it is necessary to define them within their specific cultural, social, and historical contexts.

The French Case

In France, even before the Dreyfus Affair, the neologism '*intellectuels*' was used by avant-garde circles as a social mark. The *intellectuel* was some kind of mandarin, who despised politics and wanted to distinguish himself from middle class, dominant writers and academics. *Intellectuel* was a sort of superlative of what Flaubert meant by the word '*artiste*' (Bourdieu, 1992). But, since the neologism was widely used during the Dreyfus Affair, the initial cultural meaning was replaced by a strong emphasis laid upon the political acception. In the first phase of the Dreyfus affair, the *intellectuels* were an equivalent of '*dreyfusards*' and, afterwards, when *antidreyfusard* intellectuals also intervened in a collective way, the word *intellectuels* began to define a special category of people who defended political positions based on arguments of social authority, i.e., their competence as thinkers, historians, scientists, professors, writers, or artists.

In France, the birth of *intellectuels* may be first explained by a growing inadequacy of former cultural patterns confronted to the buoyant cultural expansion of the last decades of the nineteenth century. Intellectual professions, far more numerous now, defended their social and symbolic status at once so that collective attitudes appeared, breaking with older individualistic habits. But this defense may be argued on two different grounds: with intellectual and pure values or on professional and pragmatic issues. Against avant-garde writers or academics, generally militating for the first option, professional associations appeared during the same period that were more attached to material interests. The first use of the term *intellectuels* was reserved for the first type of elitist fraternity. It was to be the germ of its political transformation as for the Russian *intelligentsia* diffused a bit earlier in the 1870s. As in Russia too, reformed universities played a major role in the process of emergence and mobilization of nonconformist intellectuals.

The second paradoxical factor of ideological and political change, was the early crisis of parliamentary democracy in France. After the stabilization of the Third Republic, apolitism predominated among intellectuals as if the end of history was already reached. A new politicization occurred with the crisis of official parties and the emergence of extremist factions, which

had a large echo among avant-garde writers (in particular anarchism and, to a lesser degree, socialism). This new trend prepared what is the specificity of the Dreyfus Affair, the invention of a new relationship with politics outside the official political scene. Intellectuals in this new terminology pretended to practice politics in a different way. This was possible because the legitimacy of Republican elite laid on the same bases as the legitimacy of intellectuals themselves, i.e., upon merit and individual talent. But in so far as these elite appeared incompetent or corrupted after different crises and scandals (Boulangism, Panama scandal, and so on), intellectuals preserved from these faults might pretend to offer an alternative elite necessary to lead any genuine democracy.

Students, avant-garde writers, and even younger generations of academics, before and during the Dreyfus Affair, expressed these new revendications of being the true representatives of the people against politicians. The State itself, with its growing intolerance towards literary innovations or extremist parties (especially anarchism) contributed to the mobilization of authors against juridical prosecutions through collective manifestos just before the Dreyfus affair.

The Dreyfus affair presented both a true continuity with preceding years and a break due to the scale of mobilization. Afterwards, the main following mobilizations of intellectuals obeyed the same collective rites and values (Ory and Sirinelli, 1986). Its founding importance was to prove that this type of mobilization might lead to real political consequences. This twofold mobilization (of *dreyfusards* and *antidreyfusards*) was new and defined, on both sides of the political scene, a general definition of intellectuals, which was not limited to leftist intellectuals. On the contrary in other countries, the equivalents of *intellectuels* are, generally speaking, confined to one side of the political spectrum.

The Peculiarities of English Intellectuals

In contrast with France, it is generally argued that no intellectuals exist at all, in the continental acception, in the UK. Since around 1980, British historians and sociologists reacted against this strong and vulgarized anti-intellectual bias. Some authors speak of an intelligentsia (Allen, 1986; Heyck, 1982), i.e., an elitist avant-garde, others of '*public moralists*,' which enhances the role of dominant and academic intellectuals (Collini, 1991) or of a '*professional class*' which assimilates intellectual professions to a new class (Perkin, 1989). A comparative approach shows that two specific factors may explain the strong difference with the French situation, in spite of the proximity of economic and political conditions in intellectual life, the persistent elitism of English academic life and the relative proximity between intellectual professions and political elite. Established elites, even if they were obliged to reform and enlarge the political system at the end of the nineteenth century, were not contested as an illegitimate power elite, to the degree they were in France. In fact, English dominant intellectuals mainly shared the same values and background as gentlemen and political leaders because they were largely issued from the same public schools and universities.

In front of these dominant intellectuals, appeared, in the last decades of the twentieth century, new types of intellectuals who presented outsider profiles. The best known were the Fabians who could not attend the best colleges and universities and had to find their way through journalism, literature, new academic institutions (e.g., the London School of Economics) or militant politics. But these avant-gardes were very different from contemporary French avant-gardes. They limited their intervention to one main field: for the Fabians, social and political questions, for aesthetes, aesthetic life, and so on. They pretended to create a voluntary structure and to indirectly influence the official sphere indirectly, not at all to destroy or affront it directly.

Even when mobilization occurred on a larger scale, as during the Boers war, English intellectuals used the official means of actions and respected legal frameworks. Finally, the main difference lay in the very different function of the State in England and in Continental Europe. In England, militant intellectuals endeavored to enlarge its role to correct social injustice, whereas, in France and even more in Germany, intellectuals tried first to weaken its authoritarian trends.

The German Case

Even if the word *Intellektuelle*, derived from French, continues almost until now to bear a derogatory nuance (Bering, 1978), the German genuine equivalents *Intelligenz*, *Gebildete*, and *Geistige*, have been used for a long time but do not imply, as the French term, the same political or social behaviors. This early appearance of the question of 'intellectuals' may be shown through the recurrent discussion about the academic proletariate (first in the *Vormärz* period, then in the 1880s, and finally in the Weimar period: Titze, 1990), with the ideological debates within the social-democratic party about the place of *Intelligenz* (Gilcher-Holthey, 1986), and also with the *Antisemitimusstreit* in 1879. This last famous polemic about the role of Jews in German society between the conservative historian Treitschke and his liberal colleague Mommsen, a former Forty-eighter (former participant to revolutionary troubles of 1848 in Germany), appears very near, in its arguments, to the debate between opposing French *dreyfusards* and *antidreyfusards*. The rights of minorities and, in particular, of individuals and Jews, lay at the center. Other affairs concerning academic freedom (Arons case or Spahn case), or freedom of creation, like the mobilization against the lex Heinze (1900), show too that the debates about intellectual autonomy were as crucial in Germany as in France and that they succeeded several times to mobilize some groups of intellectuals (Charle, 1996).

But, in all cases, mobilization was limited to specific groups and to particular issues which did not put into question the whole structure of State itself as in France. A mere political explanation (an Empire opposed to a Republic) does not suffice. What was specific and new in the Dreyfus case was the convergence of different intellectual groups about common values. In Germany, the corporatist ethos remained stronger even about general issues; free intellectuals and State intellectuals (mainly university professors) despised each other. Academics began at that time to live apart from the political sphere and preferred a general cultural function as State or *Bildung*'s defenders through different associations. Only a small

minority of free intellectuals and very few in the Academe put into question dominant elite or national causes.

The *Gebildete* assumed that they represented the true public opinion and that they were the best interpreters of general causes, but they intended to remain in their own field in order to serve their country best. This German intellectual and geographical fragmentation hindered the linkage between local or professional struggles for autonomy (Engelhardt, 1986; Hübinger and Mommsen, 1993; Ringer, 1969, 1992).

Southern Europe

In Spain as in Italy, the local equivalent of *intellectuels* seemed also to be in use in the 1890s in connection with, or even before, the Dreyfus case. The French example was very influential for the Spanish and Italian intellectuals because French cultural influence, in the two peninsulas, was already very strong since the French Revolution. Also, the inner social and political situations of Spanish and Italian intellectuals presented some analogies with the French context. As their French homologues, they thought that their countries went through a deep crisis (economic backwardness, military defeat in Spain, emigration, social riots and parliamentary corruption in Italy), which implied some sort of public intervention to find solutions. The strong anticlericalism and antimilitarism, the link between intellectuals and extreme left movements, the emergence of a new nationalism in both countries, also recalled the French debates at the turn of the twentieth century (Serrano and Salaün, 1988). Obvious differences also existed: the weaker public audience of intellectuals depending on the cultural backwardness of popular classes (high level of analphabetism) and the persistence of a large sector of opinion among Catholics, hostile to the cultural inheritance of Enlightenment, very influent in both countries, and an overproduction of *laureati* in Italy which could explain a strong commitment of academics to extremist parties (Barbagli, 1974, 1982; Michels, 1921).

Twentieth-Century Changes of Meanings

The intertwining of the sociological and the political or ethical viewpoints was perpetually renewed during the history of French intellectuals after the Dreyfus affair. A new sociologization of the word occurred with the attacks against previous *dreyfusards* in the pre-world-war period. The '*parti intellectuel*,' to use Péguy's phrase, was charged by their former allies (for example Georges Sorel) for having used the political struggle in order to conquer eminent positions of power. For their critics, this political party in fact constituted a social cluster of arrivists, a new elite of mandarins backing upon leftist politicians (Prochasson, 1992).

Between the two World Wars, there was a revival of the political and ideological emphasis on the meaning of the word during what has been called the 'Franco-french war' between the right-wing '*parti de l'intelligence*' and the left-wing '*intellectuels de gauche*.' This trend is enhanced because, at the same time, extreme left movements tried to restrict the notion to a sociological sense. The influence of Soviet Marxism may be found here in which *intellectuels* become a mere synonym for the Russian word *intelligentsia* which is nearer a sociological concept after the October revolution than the term *intellectuels* is in French. *Intellectuels* in the phraseology of the Communist Parties are assimilated to a social group in order to deny them any political autonomy and oblige them to define their political attitude within the limits of the Marxist-Leninist orthodoxy and to renounce to their own vision which is, in the French political tradition, far more linked with the French Revolution legacy and the defense of Human Rights (Benda, 1927). This intertwining of two traditions after the 1930s explains why leftist intellectuals used social concepts to attack their rightist opponents, while in contrast, rightist intellectuals laid the stress on the idealist conception of the intellectuals' role; they opposed themselves to materialism (i.e., leftist intellectuals influenced by Marxism) instead of only taking the defense of tradition as before (Sapiro, 1999).

The decline of Marxism in the *intelligentsia* and the excess of sociological terrorism in intellectual struggles during the 1950s and 1960s (Boschetti, 1985) explain why French contemporary intellectual life is dominated by some sort of revival of the primitive meaning of the word *intellectuel*. This complete historical cycle is one of the origins of the renewed interest for the study of intellectuals as political and social actors by French and foreign scholars in the 1980s and 1990s (Bourdieu, 1992; Jennings, 1993; Judt, 1992; Julliard and Winock, 1996; Sirinelli, 1990; Trebitsch and Granjon, 1998).

Anti-intellectualism, which was so frequent in England and Germany before World War I seemed to somewhat decline after the mobilization of all types of intellectuals (scholars as well as writers or journalists) during the *Union sacrée* for propaganda or practical applications devoted to National Defense. It conferred on them a new importance in all political contexts. After World War I, the democratization of politics, in both the UK and Weimar Germany, now placed intellectuals in a political context quite similar to the French one. The growing influence of left parties, of Marxist ideas, of international questions (communism, fascism, pacifism, fear for a new world war) gave birth to international debates among European intellectuals (or even American, if the case of New York intellectuals may be included: Wald, 1987). Nevertheless, this did not mean that the specific national traditions were forgotten. Even if notions like *Intellektuelle* or *intelligentsia* were present more than before in the German or English public debate, they never obtained the general influence or consensus which their equivalents enjoyed in France, Italy, or Spain (Bering, 1978; Stark, 1984). Even innovative sociologists such as Karl Mannheim continued in general to use the older lexis even to propose their new conceptions of a free-floating intelligentsia (*freischwebende Intelligenz*: Mannheim, 1929). Debates about the social decline of 'intellectual workers' in Germany or Central Europe used too old phrases such as *geistige Arbeiter* (Jarausch, 1990, Titze, 1990). The anti-intellectualism of the Nazi movement and the huge migration (after 1933) of progressist or Jewish intellectuals out of Germany and Central Europe stopped, for almost two decades, this timid convergence with the Latin tradition of intellectuals.

The decline and fall of the different fascist or communist regimes in the second half of the twentieth century gave a new

actuality to the oppositional figure of intellectuals, in particular in Central and Eastern Europe (the 'dissidents' who recalled the nineteenth century meaning of intelligentsia). In Southern Europe, in Africa, Asia, or Latin America, revolutionary intellectuals or militants for the Human Rights also rejuvenated the nineteenth century European tradition of intellectual youth mobilized for the nation or populist projects. All these convergences or political cycles explain why the historical notion of intellectuals continues to be at the center of many historical, sociological, or philosophical reflexions, trying until today to find general or transhistorical definitions of this term.

See also: Elites, Anthropology of; Elites: Sociological Aspects; Intellectuals, Sociology of; Mannheim, Karl (1893–1947); Professions and Professionalization, History of; Prosopography (Collective Biography); Social Scientists as Experts and Public Intellectuals.

Bibliography

Ahearne, Jeremy, Bennett, Oliver (Eds.), 2007. Intellectuals and Cultural Policy. Routledge, London.

Allen, P., 1986. The meanings of 'an intellectual': nineteenth- and twentieth-century English usage. University of Toronto Quarterly 55, 342–358.

Barbagli, M., 1974. Disoccupazione intellettuale e sistemo scolastico in Italia. Il Mulino, Bologna [1982 Educating for Unemployment: Politics, Labor Markets, and the School System Italy, 1859–1973. Columbia University Press, New York].

Bell, D., 1973. The Coming of Post-Industrial Society, a Venture in Social Forecasting. Basic Books, New York.

Benda, J., 1927. La trahison des clercs. Grasset, Paris [1928 The Great Betrayal. Routledge, London].

Bender, Th, 1987. New York Intellect: A History of Intellectual Life in New York City, from 1750 to the Beginnings of Our Own Time. Knopf and Johns Hopkins University Press, New York.

Bering, D., 1978. Die Intellektuellen. Geschichte eines Schimpfwortes. Klett, Stuttgart, Germany.

Boschetti, A., 1985. Sartre et les Temps modernes. Minuit, Paris [1988 The Intellectual Enterprise: Sartre and Les Temps Modernes. Northwestern University Press, Evanston, IL].

Bourdieu, P., 1992. Les rgles de l'art. Gense et structure du champ littéraire. ditions du Seuil, Paris [1996 The Rules of Art: Genesis and Structure of the Literary Field. Polity Press, Cambridge, UK].

Charle, C., 1990. Naissance des 'intellectuels' (1880–1900). ditions de Minuit, Paris.

Charle, C., 1996. Les Intellectuels en Europe au XIXme sicle, essai d'histoire comparée. ditions du Seuil, Paris.

Collini, S., 1991. Public Moralists: Political Thought and Intellectual Life in Britain 1850–1930. Clarendon Press, Oxford, UK.

Collini, S., 1993. Intellectuals in Britain and France in the twentieth century: confusions, contrasts and convergence? In: Jennings, J. (Ed.), Intellectuals in Twentieth Century France, Mandarins and Samurais. St. Martin's Press, London, New York, pp. 199–225.

Collini, Stefan, 2006. Absent Minds. Intellectuals in Britain. Oxford University Press, Oxford, New York.

Drouin, M., 1994. L'Affaire Dreyfus de A Z. Flammarion, Paris.

Engelhardt, U., 1986. Bildungsbürgertum.' Begriffs- und Dogmengeschichte eines Etiketts. Klett-Cotta, Stuttgart, Germany.

Fleck, Christian, Hess, Andreas (Eds.), 2014. Knowledge for Whom? Public Sociology in the Making. Ashgate, Farnham.

Fleck, Christian, Hess, Andreas, Lyon, E. Stina (Eds.), 2009. Intellectuals and Their Publics. Perspectives from the Social Sciences. Ashgate, Farnham.

Gilcher-Holthey, I., 1986. Das Mandat des Intellektuellen: Karl Kautsky und die Sozialdemokratie. Siedler, Berlin, Germany.

Granjon, M.-C., Racine, N., Trebitsch, M. (Eds.), 1997. Histoire comparée des intellectuels. IHTP, CNRS, Paris.

Heyck, T.W., 1982. The Transformation of Intellectual Life in Victorian England. Croom Helm, London.

Hübinger, G., Mommsen, W. (Eds.), 1993. Intellektuelle im Deutschen Kaiserreich. Fischer Taschenbuch Verlag, Frankfurt/M, Germany.

Jarausch, K.H., 1990. The Unfree Professions, German Lawyers, Teachers, and Engineers, 1900–1950. Oxford University Press, New York, Oxford.

Jennings, J. (Ed.), 1993. Intellectuals in Twentieth Century France, Mandarins and Samurais. St. Martin's Press, London.

Judt, T., 1992. Past Imperfect: French Intellectuals, 1944–1956. University of California Press, Berkeley, CA.

Julliard, J., Winock, M. (Eds.), 1996. Dictionnaire des intellectuels franais. ditions du Seuil, Paris.

Kuklick, Bruce, 2006. Blind Oracles: Intellectuals and War from Kennan to Kissinger. Princeton University Press, Princeton, NJ.

Mannheim, K., 1929. Ideologie und Utopie. F. Cohen, Bonn, Germany [1936 Ideology and Utopia, an Introduction to the Sociology of Knowledge, K. Paul, Trench, Trubner and Co., Harcourt, Brace and Co., London, New York].

Michels, R., 1921. Le Prolétariat et la bourgeoisie dans le mouvement socialiste italien. Giard, Paris.

Michels, R., 1932. Intellectuals. In: Encyclopaedia of the Social Sciences, vol. 8. Macmillan, New York, pp. 118–126.

Müller, O.W., 1971. Intelligencija. Untersuchungen zur Geschichte eines politischen Schlagwortes. Athenäum, Frankfurt/M, Germany.

Ory, P., Sirinelli, J.-F., 1986. Les intellectuels en France de l'affaire Dreyfus nos jours. Colin, Paris.

Perkin, H., 1989. The Rise of Professional Society, England since 1880. Routledge, New York, London.

Prochasson, C., 1992. Les Intellectuels, le socialisme et la guerre (1900–1938). ditions du Seuil, Paris.

Ringer, F.K., 1969. The Decline of the German Mandarins. Harvard University Press, Cambridge, MA.

Ringer, F.K., 1992. Fields of Knowledge. French Academic Culture in Comparative Perspective 1890–1920. Cambridge University Press, Cambridge, UK.

Sapiro, G., 1999. La guerre des écrivains 1940–1953. Fayard, Paris.

Serrano, C., Salaün, S. (Eds.), 1988. 1900 en Espagne, essai d'histoire culturelle. Presses Universitaires de Bordeaux, Bordeaux, France.

Shils, E., 1972. The Intellectuals and the Powers, and Other Essays. University of Chicago Press, Chicago.

Sirinelli, J.-F., 1990. Intellectuels et passions franais: Manifestes et pétitions au XXme sicle. Fayard, Paris.

Sowell, Thomas, 2012. Intellectuals and Society, Revised and Enlarged Edition. Basic Books, New York.

Stark, M. (Ed.), 1984. Deutsche Intellektuelle 1910–1933, Aufrufe, Pamphlete, Betrachtungen. Verlag Lambert Schneider, Heidelberg, Germany.

Titze, H., 1990. Der Akademikerzyklus. Vandenhoeck and Ruprecht, Gttingen, Germany.

Trebitsch, M., Granjon, M.-C. (Eds.), 1998. Pour une histoire comparée des intellectuels. Complexe, Brussels, Belgium.

Wald, A.M., 1987. The New York Intellectuals. The Rise and Decline of the Antistalinist Left from the 1930s to the 1980s. University of North Carolina Press, Chapel Hill, NC.

Williams, R., 1976. Keywords. A Vocabulary of Culture and Society. Oxford University Press, Oxford, UK.

Wolin, Richard, 2010. The Wind from the East: French Intellectuals, the Cultural Revolution and the Legacy of the 1960s. Princeton University Press, Princeton, NJ.

Intellectuals, Sociology of

Robert J Brym, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article first reviews the development of the term 'intellectual' and the social and historical circumstances that led intellectuals to become a large, moderately well-defined, and occasionally self-conscious group. Particular attention is paid to the rise of commerce and industry since the late Middle Ages, the spread of literacy, the growth of markets, and the proliferation of social contexts for intellectual discussion and debate. The article then argues that three main traditions of analysis inform the sociological study of intellectuals: class theories, theories of classlessness, and theories of shifting social networks. It is argued that the third of these traditions is the most fruitful. Evidence to support this claim is drawn from analyses of the social conditions that lead intellectuals to assume various ideological and political positions. It is concluded that, to explain intellectuals' allegiances, one must trace their paths of social mobility as they are shaped by the capacity of various social collectivities to expand the institutional milieus through which intellectuals pass. To the degree these milieus are imprinted with the interests of the social collectivities that control them, they circumscribe the interests reflected in intellectuals' ideologies and political allegiances.

Definition

Intellectuals are people who produce, evaluate, and distribute culture. Their role often involves some mixture of endorsing or criticizing the cultural objects of their attention. Following Michael Burawoy (2005), intellectual work may be said to vary along two dimensions: its intended audience (which may include academics, policy makers, various sections of the public, and the public as a whole) and the degree to which it makes instrumental versus normative claims ('what is' versus 'what ought to be'). Individual intellectuals may specialize in one or more of the cells resulting from the cross-classification of the two dimensions. Similarly, particular academic disciplines, professions, schools of thought, regions, and historical periods may highlight one or more specializations (Brym and Nakhaie, 2009).

Origins

The intellectual role has been performed in all societies. Even a nomadic tribe of fewer than 100 foragers could afford to employ a shaman to forecast the weather, predict the movement of animals, and tend to the ill. However, the Renaissance humanists (CE *c.*1300–1600) were the first influential group resembling today's intellectuals. Renaissance scholars thought they lived in a golden age of virtue and learning. To emphasize the point, they invented the term 'Middle Ages' to refer to the dark gulf separating their era from the glories of the ancient world. Promoting the study of poetry, grammar, rhetoric, history, and moral philosophy, the humanists from Petrarch in the fourteenth century to More, Erasmus, and Rabelais in the sixteenth century influenced the Reformation, the scientific revolution, and the Age of Reason.

Only since about 1600, however, have intellectuals become a large, moderately well-defined, and occasionally self-conscious group. The chief factors that helped to distinguish them were the rise of commerce and industry since the late Middle Ages, the spread of literacy, the growth of markets, and the proliferation of social contexts for intellectual discussion and debate. Flourishing commerce and industry permitted the accumulation of economic surpluses. They were used to expand state and private institutions employing intellectuals. Public education and mass literacy increased demand for intellectual services. In turn, the evolution of markets for these services allowed intellectuals greater freedom to define their political allegiances. The interests of aristocratic and church patrons no longer tightly constrained them. Finally, a host of modern social settings – ranging from universities to political movements, professional societies to coffee houses, academic journals to mass-circulation newspapers – provided contexts within which intellectuals could formulate their social and political identities.

It was not until the second half of the nineteenth century that 'intellectual' and the kindred 'intelligentsia' entered common parlance. The term 'intellectual' was first employed on a wide scale in France. In 1898, Georges Clemenceau referred to the leaders of the anticlerical and antimilitary camp that opposed the conduct of the Dreyfus trial as *les intellectuels*. Soon, the political right was deriding the self-proclaimed conscience of the French nation. Because of the resulting public debate, the term 'intellectual' stuck. The term 'intelligentsia' was popularized in Central and Eastern Europe a few decades earlier. It, too, denoted liberals, socialists, and other critics of authority. Only gradually since the end of the nineteenth century has the term 'intellectual' gained widespread acceptance and succeeded in unwrapping itself from quotation marks.

Theories

Many normative or moralistic treatments of the problem of intellectuals may be found in the literature. They criticize what intellectuals do and make a case for what they ought to do. Julien Benda (1969[1928]), Noam Chomsky (1969), and others have written important works in this tradition. In contrast, this article focuses on the analytical literature, which seeks to explain why intellectuals do what they do.

Three main analytical traditions inform the sociological study of intellectuals. Sociologists have regarded intellectuals as

(1) members of a class or as a class in their own right; (2) relatively classless; and (3) embedded in a shifting network of class and other group affiliations (Brym, 2010[1980]; Kurzman and Owens, 2002).

Class Theory

Karl Marx and Friedrich Engels (1972[1848]) regarded intellectuals as members of the *petite bourgeoisie* who would be 'precipitated into the proletariat' as capitalism developed. They believed that, under socialism, classes would disappear, so one could be "a farmer in the morning, a laborer in the afternoon, and a philosopher in the evening" (1972 [1845–6]). In contrast, anarchists – Mikhail Bakunin (1872) and Jan Wacław Machajski (Shatz, 1989) chief among them – asserted that intellectuals aspired to form a dominant class. In making this claim, they anticipated the ideas of many twentieth-century social thinkers.

By the 1950s, the middle of the class structure had expanded enormously in both capitalist and self-proclaimed socialist societies. People with university degrees filled many of the intermediate ranks. Reflecting on this change, a growing number of social thinkers argued that intellectuals constitute a class that enjoys increasing power. Alvin Gouldner (1979) held that intellectuals comprise a new 'emancipatory class.' Barbara Ehrenreich and John Ehrenreich (1979) maintained that intellectuals form part of a new 'professional-managerial class.' Irving Kristol (1972) called intellectuals the "new class of knowledge workers." Defining intellectuals broadly to include economists, technocrats, administrators, police bureaucrats, ideologists, scientists, and artists, Georg Konrad and Iván Szelényi (1979) went so far as to claim that Eastern European intellectuals had assumed class power. These and other scholars, including James Burnham (1941), Milovan Djilas (1957), and John Kenneth Galbraith (1967), disagreed about much, but they shared the view that intellectuals form a large, ideologically quite homogenous, and rising or dominant class.

Steven Brint's research exemplifies one important criticism of the view that intellectuals form a class. Using American data from a series of General Social Surveys, he demonstrated considerable ideological heterogeneity and strong cohort effects among intellectuals. Intellectuals, it turns out, are as numerous on the right as they are on the left, with the two camps tending to cluster in different occupational niches. Moreover, their ideological predispositions depend partly on the *Zeitgeist* of the historical period during which they reach late adolescence and early adulthood (Brint, 1984). Consequently, the ideological cohesiveness and presumed teleological movement of intellectuals toward class power seem greatly exaggerated.

The second main criticism of the argument that intellectuals form a rising or dominant class is that, typically, they are demonstrably subservient to other economic and/or political classes. The Communist Party of the Soviet Union always sought to ensure that universities, research institutes, and the *nomenklatura* were populated by loyalists. In Russia, the collapse of communism witnessed the ascent not of intellectuals but of people with strong connections to the state security apparatus and organized crime. Lawrence King and Iván Szelényi (2004) anticipate that, in capitalist societies, intellectuals will come to dominate an increasingly autonomous state. However, this eventuality seems increasingly unlikely in an era when giant global corporations influence governments to eliminate trade barriers and give up sovereign rights to pro-free market transnational organizations such as the World Bank and the International Monetary Fund. State autonomy has decreased in recent decades and there can be little doubt which socioeconomic strata are chiefly responsible for this development.

For a time, intellectuals were politically ascendant in some parts of the world. For instance, as Charles Kurzman (2008) has shown, between 1905 and 1911, more than a quarter of the world's population (in China, Iran, Mexico, the countries of the Ottoman Empire, Portugal, and Russia) was caught up in democratic revolutions led largely by self-identified intellectuals. Many workers, peasants, members of the bourgeoisie, and in some cases even landlords deferred to them. Within a few years, however, their ascent was checked. Russian workers wanted more than just elections and a constitution. Chinese intellectuals lost power to the army because of widespread social disorder. When their commercial interests were threatened, the great powers turned against new, democratic regimes. Among the countries listed earlier, only Portugal's democratic constitution survived, and then only until 1926.

Theories of Classlessness

Karl Mannheim (1955[1929]) held that modern intellectuals form neither a class nor part of a class. Instead, they are "members of a relatively classless stratum which is not too firmly embedded in the social order." Mannheim acknowledged that intellectuals are typically recruited from various classes. However, he also held that, because intellectuals participate in a common educational milieu, their class differences, and the variations in outlook normally associated with them, tend to be suppressed. Many scholars echoed Mannheim's conclusions. Talcott Parsons (1963) asserted that intellectuals put 'cultural considerations before social ones.' Everett Ladd and Seymour Martin Lipset (1975) wrote that the capacity of intellectuals for social criticism, creativity, innovation, and attention to facts enables them to overcome their class socialization – and, for that matter, the socializing influence of many other social collectivities to which they belong.

A recent and more nuanced variant of the classlessness thesis has been proposed by Neil Gross (2002, 2008). Gross does not regard intellectuals as entirely removed from the sway of social forces but nonetheless emphasizes that ideological and intellectual choices are largely a function of which ideas resonate with intellectuals' self-conceptions. However, this approach begs the question of how self-conceptions are rooted in social experience. Categories such as 'class,' 'ethnicity,' and 'gender' are admittedly too broad and static to capture the range of social forces that influence intellectuals' political and other choices, but acknowledging this fact does not necessitate a retreat to an explanatory framework that emphasizes coincidence or a voluntaristic phenomenology of self. Instead, it suggests the importance

of discovering more precise and dynamic social variables that influence intellectual choice.

Theories of Shifting Social Networks

Both class theories and theories of classlessness ignore crucial issues. Notably, intellectuals may be found at all points on the political spectrum. Class theories may briefly note ideological heterogeneity but tend to dismiss it as an uninteresting side issue, leaving it unexplained. Meanwhile, theories of classlessness minimize the significance of social influences on the shaping of ideas, thereby deflecting attention from the very issue that the sociology of intellectuals is mandated to analyze.

A third approach seeks to overcome both these problems by focusing on the intellectual's web of shifting group affiliations. According to Pierre Bourdieu (1984[1979]; Bourdieu and Wacquant, 1992), cultural fields tie individual and organizational actors together in various social institutions. These cultural fields are structured as social networks. Different positions in a social network are associated with different endowments of economic, social, cultural, and intellectual capital. Network position is also associated with cognitive structure. Thus, intellectuals in similar network positions share tastes, ideas, and values (Anheier et al., 1995; Lamont, 1987).

Randall Collins's magnificent analysis of philosophical schools takes this argument a step farther by showing how intellectuals gain reputations by successfully competing for attention, fame, and influence in intellectual networks. Collins demonstrates that intellectual creativity flows through chains of personal contacts. It moves by structured rivalries between intellectual networks. And it flowers when rival circles intersect in cultural metropoles. Because 'intellectual attention space' is limited, no more than six, and often as few as three, creative schools manage to propagate their ideas across generations in a given culture at a given time. Creativity dries up when a single orthodoxy dominates. Contrariwise, intellectuals create new schools of thought in response to political and economic changes that shift the distribution of material resources needed to support their work. The stars of intellectual life tend to be people who are involved in, or close to, such organizational transformations. Leading innovators are as likely to be conservatives as radicals (Collins, 1998; Gross and Frickel, 2005).

Bourdieu's and Collins's mesolevel theories are compatible with Robert Brym's approach, which, however, focuses on political allegiance rather than intellectual creativity and influence. Developing themes in Antonio Gramsci's essay on intellectuals (Gramsci, 1971[1929–35]), Brym argues that the ideologies and political allegiances of intellectuals depend on their social origins and the structure of opportunities for education, employment, and political involvement they face during their careers. These opportunity structures are, in turn, shaped by the relative power of major classes and other social collectivities. Accordingly, to explain intellectuals' ideologies and political allegiances, one must trace their paths of social mobility as they are shaped by the capacity of classes, ethnic groups, religious orders, and other social collectivities to expand the institutional milieus through which intellectuals pass. To the degree these milieus are imprinted with the interests of the collectivities that control them, they circumscribe the collective interests reflected in intellectuals' ideological and political allegiances. From this point of view, it is an oversimplification to say that intellectuals form a class, are members of a class, or are classless. They are embedded in social networks whose ties to various classes and other collectivities shift over time and help account for their ideologies and political allegiances (Brym, 1978, 2010 [[1980]]). This theme has also been taken up by Jerome Karabel (1996). The remainder of this article illustrates it.

Social Origins and Institutional Opportunities

Social Origins

In Weimar Germany, professors were largely antirepublican and right-wing. Mostly, their fathers were military officers, state bureaucrats, and academics, members or handmaidens of the aristocracy. In contrast, nonacademic intellectuals were inclined to the left. They tended to be children of successful participants in the industrial revolution, members of the upstart German bourgeoisie. Similarly, in the middle of the nineteenth century, Russia's conservative and politically moderate intellectuals were recruited overwhelmingly from the aristocracy. On the other hand, the first generation of truly radical Russian intellectuals dates from the 1860s, and it contained in its ranks a large mixture of commoners. Their fathers were merchants, peasants, and petty officials. These examples illustrate that support for the traditional social order is common among intellectuals who have been reared in upper class families that are threatened by social change. Being born into a less advantaged class has usually been associated with greater potential for intellectual radicalism.

But not always. In North America, it seems that rates of upward social mobility have historically been higher than in Europe, and class has been a less important social distinction. As a result, class origin has less bearing on the ideologies and political allegiances of North American than European intellectuals. Especially in the United States, ethnic and racial origin, educational background, field of expertise, professional and institutional status, and generational experience matter more. Accordingly, American surveys show that professors in small provincial universities are more likely to be politically conservative than are professors in prestigious Ivy League universities. African-American and Jewish intellectuals are usually more left-wing than are non-Jewish white intellectuals. Young intellectuals who work in the social sciences and cultural fields are more likely to be left-wing than are older intellectuals with business affiliations (Ladd and Lipset, 1975).

Economic Opportunities

As intellectuals reach adulthood and enter the job market, the structure of markets for intellectuals skills can modify their early political socialization. An abundance of secure jobs that

allow free expression can dampen intellectuals' early radical impulses. A lack of such jobs can exert a radicalizing effect. Some intellectuals may continue to express radical political sentiments after they have taken jobs that suit them professionally. Still, the rate and intensity of radicalism are generally greater when the number of such jobs is fewer.

For example, the Quebec educational system began to produce substantial numbers of highly educated French-speaking graduates in the 1960s. The new graduates found that an English-speaking minority controlled the larger and more efficient businesses in the private sector of the economy. Many highly educated Québécois were shut out of good jobs. This, among other factors, encouraged some of them to develop the idea that to become 'masters in their own house' Quebec must become a sovereign state. Surveys show that intellectuals became separatists in disproportionately large numbers and were a driving force leading to the election of separatist governments in 1976 and 1994.

Political Opportunities

The patterns of political opportunities intellectuals face also profoundly affect their ideas and loyalties.

At the level of social organization, political opportunities are structured by the availability of historical agents: workers, peasants, ethnic collectivities, and other groups that intellectuals may demarcate as the chief instruments of social change. Whether intellectuals become, say, socialists, populists, or nationalists is determined partly by which historical agents are mobilized for political action and which are relatively politically dormant in intellectuals' milieus at a given time.

At the level of the political system, the capacity of party organizations to absorb new talent is also important in shaping intellectuals' allegiances. For example, before the 1920s, the German Social Democratic Party had attracted most of the country's radical intellectuals. By the 1920s, however, it became (to use the then-current catchwords) 'bossified, ossified, and bourgeoisified.' Old men remained incumbents in the German Social Democratic Party for many years. Young men had little hope of rising in its ranks. Young women were even less likely to ascend to prominence, hindered as they were by the institutionalized inequality and sexism that, as Dale Spender (1982) documents, has dominated intellectual life for centuries. Lack of opportunity was one reason the young generation of radicals turned in large numbers to the Communist Party of Germany.

If historical agents are largely unmobilized and no party organization is available to sustain the intellectual's beliefs, a process of political disillusionment is likely to set in. The lack of an historical agent combined with the presence of a party organization, however, is likely to lead to intellectual elitism.

Intellectual Elitism

Ever since Plato envisaged a society ruled by 'philosopher-kings,' some intellectuals have been motivated by the conviction that they are better suited than are nonintellectuals to create and maintain ideal social arrangements. For example, in 1816, Hegel referred to Prussia as "the state of the intelligentsia." Both he and Comte characterized historical epochs as accomplishments of the human mind, which they saw as the driving force of history. In their judgment, the intellectual was the human mind's personification. Reinforcing this elitist view was the late nineteenth and early twentieth-century conception of scientists and other highly educated experts as a sort of ascetic priesthood whose dispassionate approach to their subject matter allowed them to perform extraordinary cognitive feats promoting the progress of humanity.

Lewis Feuer's studies did much to dispel the supposed ascetic roots of modern science, emphasized by Max Weber and Robert Merton, and demonstrate the antiauthoritarian, hedonistic, and individualistic mind-set that typically accompanies creative scientific work (Feuer, 1963; Merton, 1936; Weber, 1958[1905]). In recent decades, scholars working in the hybrid field known as 'social studies of science' have succeeded in bringing intellectual creativity even more firmly down to earth by demonstrating how (to quote Steven Shapin) science is "produced by people with bodies, situated in time, space, culture, and society, and struggling for credibility and authority" (Shapin, 2010; Johnson, 2008).

Turning to politics, the degree of elitism expressed by intellectuals varies independently of their position on the left–right dimension. One finds democrats and elitists on both the left and the right. Two sets of circumstances promote intellectual elitism – weak participatory demands by the masses and lack of competition from other elites.

The history of Latin America illustrates how weak mass demand for political participation encourages intellectual elitism. For example, in the 1950s, Cuban revolutionary intellectuals were faced with a small and reformist working class and a politically inert peasantry. Consequently, they took matters into their own hands, forming guerilla bands in the countryside to seize state power. The low level of popular political participation in, and mass control over, the Cuban revolutionary party diminished the likelihood of subsequent democratic development.

The history of socialism and communism in Russia reveals a similar pattern. However, the Russian case is also instructive because it illustrates the opposite tendency. Even apparently unyielding elitists can democratize if they are pressured from below. Thus, when ties between workers and Russian Marxist intellectuals were dense – during the strike wave of the 1890s, the 1905 revolution, and the period of labor militancy in 1912–14 – most intellectuals wanted their parties to operate according to democratic principles. In contrast, when working class political participation fell – due to labor quiescence caused by troughs in the business cycle, strong government and police reaction to radical activities, and enormous losses in World War I – intellectuals returned to elitist principles of party organization.

A second precondition of democratic practice among intellectuals is vigorous elite competition. Aristotle argued

that justice can be maximized not if philosopher-kings rule, but if constitutions are divided against themselves – if, in other words, there exists what is now called a separation of powers among the branches of government. Following Aristotle, scholars now commonly believe that tyranny and arbitrary rule by an elite can be prevented if the branches of government hold each other in check. One may invoke a broadly similar principle in analyzing intellectual elitism. Nonintellectual elites need to keep intellectuals in check if they are to prevent the political and social dangers that can result from intellectuals taking Plato to heart.

The most compelling case for the benefits of elite competition derives from the history of communism, the epitome of intellectual tyranny in the twentieth century. Yet the dangers of intellectual elitism have hardly subsided now that the communist era has virtually ended. In much of the world, intellectuals are making bolder claims than ever about their ability to forecast and plan social and scientific change. Governments are respecting those claims by seeking the advice of intellectuals, awarding them research contracts, and employing them by the legion. Emboldened by their growing numbers and prestige, some intellectuals, particularly in the United States and Eastern Europe, have proclaimed themselves the ruling class of the future, the real holders of power in an era when knowledge allegedly means more than capital in determining status.

One should not take intellectuals' claims about their accomplishments too seriously. For example, as Eva Etzioni-Halevy (1985) documents, economists, the 'hardest' of the social scientists, have generally failed to forecast and regulate economic trends. Natural scientists have arguably helped to create nearly as many social problems as they have solved.

On the other hand, the dangers of too much intellectual influence on political life should be carefully heeded. Individual intellectuals have proved to be exemplary political leaders, but when intellectuals are in a position to impose their blueprints on society they often do more harm than good. The 'shock therapy' advocated by the academics who formed the core of the Russian government late in 1991 caused tremendous social dislocation and political reaction that undermined reform in that country. This is only one example of what can happen when intellectuals rule. It seems reasonable to conclude that the practice of politics is much too serious a matter to be left only to the intellectuals.

See also: Enlightenment; Intellectual History; Knowledge, Sociology of; Literacy Education; Literacy and Illiteracy, History of; Renaissance; Science and Technology Studies, History of.

Bibliography

Anheier, Helmut K., Gerhards, Jürgen, Romo, Frank P., 1995. Forms of capital and social structure in cultural fields: examining Bourdieu's social topography. American Journal of Sociology 100 (4), 859–903.

Bakunin, Mikhail, 1872. On the International Workingmen's Association and Karl Marx (retrieved 03.07.12.) http://www.marxists.org/reference/archive/bakunin/works/1872/karl-marx.htm.

Benda, Julien, 1969[1928]. The Treason of the Intellectuals (Richard Aldington, Trans). Norton, New York.

Bourdieu, Pierre, 1984[1979]. Distinction: A Social Critique of the Judgment of Taste (Richard Nice, Trans). Harvard University Press, Cambridge MA.

Bourdieu, Pierre, Wacquant, Loic J.D., 1992. An Invitation to Reflexive Sociology. University of Chicago Press, Chicago.

Brint, Stephen, 1984. 'New class' and cumulative trend explanations of the liberal political attitudes of professionals. American Journal of Sociology 90 (1), 30–71.

Brym, Robert J., 1978. The Jewish Intelligentsia and Russian Marxism: A Sociological Study of Intellectual Radicalism and Ideological Divergence. Macmillan, London.

Brym, Robert J., 2010[1980]. Intellectuals and Politics. Routledge, London.

Brym, Robert J., Nakhaie, M. Reza, 2009. Professional, critical, policy, and public academics in Canada. Canadian Journal of Sociology 34 (3), 655–669.

Burawoy, Michael, 2005. For public sociology. American Sociological Review 70 (1), 4–28.

Burnham, James, 1941. The Managerial Revolution. John Day, New York.

Chomsky, Noam, 1969. American Power and the New Mandarins. Vintage, New York.

Collins, Randall, 1998. The Sociology of Philosophies: A Global Theory of Intellectual Change. Harvard University Press, Cambridge, MA.

Djilas, Milovan, 1957. The New Class: An Analysis of the Communist System. Praeger, New York.

Ehrenreich, Barbara, Ehrenreich, John, 1979. The professional-managerial class. In: Walker, Pat (Ed.), Between Labor and Capital. Black Rose Books, Montreal, pp. 5–45.

Etzioni-Halevy, Eva, 1985. The Knowledge Elite and the Failure of Prophecy. George Allen and Unwin, London.

Feuer, Lewis S., 1963. The Scientific Intellectual: The Psychological and Sociological Origins of Modern Science. Basic, New York.

Galbraith, John Kenneth, 1967. The New Industrial State. Houghton Mifflin, Boston.

Gouldner, Alvin, 1979. The Future of Intellectuals and the Rise of the New Class: A Frame of Reference, Theses, Conjectures. Seabury, New York.

Gramsci, Antonio, 1971[1929–35]. The intellectuals (Quentin Hoare and Geoffrey Nowell Smith, Trans). In: Hoare, Quentin, Smith, Geoffrey Nowell (Eds.), Selections from the Prison Notebooks. International Publishers, New York, pp. 5–23.

Gross, Neil, 2002. Becoming a pragmatist philosopher: status, self-concept, and intellectual choice. American Sociological Review 67 (1), 52–76.

Gross, Neil, Frickel, Scott, 2005. A general theory of scientific/intellectual movements. American Sociological Review 70 (2), 204–232.

Gross, Neil, 2008. Richard Rorty: The Making of an American Philosopher. University of Chicago Press, Chicago.

Johnson, Steven, 2008. The Invention of Air: A Story of Science, Faith, Revolution, and the Birth of America. Riverhead Books, New York.

Karabel, Jerome, 1996. Towards a theory of intellectuals and politics. Theory and Society 25 (2), 205–233.

King, Lawrence Peter, Szelényi, Iván, 2004. Theories of the New Class: Intellectuals and Power. University of Minnesota Press, Minneapolis.

Konrad, Georg, Szelényi, Iván, 1979. The Intellectuals on the Road to Class Power: A Sociological Study of the Role of the Intelligentsia in Socialism. Harcourt Brace Jovanovich, New York.

Kristol, Irving, 1972. About equality. Commentary 54 (5), 41–47.

Kurzman, Charles, 2008. Democracy Denied, 1905–1915: Intellectuals and the Fate of Democracy. Harvard University Press, Cambridge, MA.

Kurzman, Charles, Owens, Lynn, 2002. The sociology of intellectuals. Annual Review of Sociology 28, 63–90.

Ladd, Everett Cyrill, Lipset, Seymour Martin, 1975. The Divided Academy: Professors and Politics. McGraw-Hill, New York.

Lamont, Michelle, 1987. How to become a dominant French philosopher: the case of Jacques Derrida. American Journal of Sociology 93 (3), 584–622.

Mannheim, Karl, 1955[1929]. Ideology and Utopia (Louis Wirth, Edward Shils, Trans). Harvest, New York.

Marx, Karl, 1972[1845–6]. The German ideology: part one. In: Tucker, Robert (Ed.), The Marx-Engels Reader. Norton, New York, pp. 147–200.

Marx, Karl, Engels, Friedrich, 1972[1848]. Manifesto of the communist party. In: Tucker, Robert (Ed.), The Marx-Engels Reader. Norton, New York, pp. 331–362.

Merton, Robert, 1936. Puritanism, pietism and science. The Sociological Review 28 (1), 1–30.

Parsons, Talcott, 1963. The intellectual: a social role category. In: Rieff, Paul (Ed.), On Intellectuals. Anchor, Garden City, NJ, pp. 3–24.

Shapin, Steven, 2010. Never Pure: Historical Studies of Science as if It Was Produced by People with Bodies, Situated in Time, Space, Culture, and Society, and Struggling for Credibility and Authority. Johns Hopkins University Press, Baltimore.

Shatz, Marshall, 1989. Jan Wacław Machajski: A Radical Critic of the Russian Intelligentsia and Socialism. University of Pittsburgh Press, Pittsburgh.

Spender, Dale, 1982. Women of Ideas and What Men Have Done to Them: From Aphra Behn to Adrienne Rich. Routledge and Kegan Paul, London.

Weber, Max, 1958[1905]. The Protestant Ethic and the Spirit of Capitalism (T. Parsons, Trans). Scribner, New York.

Intelligence: Assessments of

Adam W McCrimmon and Emma A Climie, Werklund School of Education, University of Calgary, Calgary, AB, Canada
Donald H Saklofske, Social Science Centre, University of Western Ontario, London, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Intelligence is one of the most talked about, researched, and measured of the many psychological constructs describing individual differences. Intelligence tests but have undergone considerable changes over the past 100 years as our models of intelligence have become more refined and data based. This article overviews some of the more notable developments in defining and assessing intelligence, examples of currently used tests followed by a discussion of their uses.

"A rose by any other name…" Past and modern civilizations or countries have created words and descriptions that reflect individual differences in the abilities that we now have come to label as intelligence in the Western world. Intelligence, no matter the differing views on its 'structure' or 'cause,' is a major factor underlying individual differences in what and how we think, know, act, feel, learn, and solve problems. Importantly, this ability is universal and variations within and between persons and groups can be measured.

This article reviews the foundations of intelligence including definitions, types of intelligences, historical and contemporary understandings of the construct, and current models. Following a brief presentation of some of the tests that have been employed for assessing cognitive abilities, the article will highlight two of the more popular intelligence tests and focus on how the results help us to understand individual differences in intelligence. Finally, consideration of some important social and professional issues will be summarized.

What Is Intelligence?

Agreement on a description of intelligence has been challenging. Indeed, asking the seemingly simple question of 'what is intelligence?' results in varied responses. The differing descriptions of intelligence reflect the diverse nature of this concept and the ways in which it may be defined. The question may lead to several additional queries: (1) What is intelligence? (2) Are there different kinds of intelligence? (3) What causes intelligence? (4) Can we measure intelligence? and (5) How does intelligence help us to explain individual differences? Much of the psychological literature addresses these very questions (e.g., Neisser et al., 1996).

Various theories have been proposed to describe and define intelligence. Some theories define intelligence as a composite of many correlated specific abilities (e.g., Spearman, 1904), while others suggest there are different 'kinds' of intelligence (e.g., Gardner, 1983) ranging from verbal-linguistic to interpersonal intelligence. A public statement published by 52 experts (Gottfredson, 1997, p. 14) defined intelligence as a general mental capacity that includes the ability to think logically and abstractly, understand complex ideas, reason, plan, solve problems, learn quickly and from experience, communicate, and problem-solve (Gottfredson, 2008; Gottfredson and Saklofske, 2009). A further analysis of the 'knowns and unknowns' about intelligence was published 2 years later by Neisser et al. (1996) where it was recognized that both genetics and environment are determiners of intelligence, gender differences were generally small or nonsignificant in contrast to group differences, and intelligence was an important predictor of school achievement and years of education.

Brief History of Intelligence Assessment

Although there are examples of efforts to measure intelligence dating back several 1000 years, the establishment of psychology as a scientific discipline and practice provided the framework for modern intelligence assessment. In the mid- to late-1800s, the well-known Galton tests including strength of grip, memory for digits, pitch discrimination, and form boards were the standard tools for assessing intelligence (Galton, 1885, 1887). However, the first comprehensive measure of intelligence was published by Binet and Simon (1905). Initially developed for use with French school children to determine who may benefit from 'special education,' the measure consisted of 30 tasks that were sequenced in order of difficulty.

It was this initial endeavor to evaluate the mental abilities of individuals that has driven the cognitive-intelligence assessment movement. Revisions to the Binet and Simon scale in 1908 reformulated the measure so that the tasks were organized into age levels, providing a basis of comparison by which a child's performance could be evaluated against others of their same age. In addition, this revision produced a quantification of the child's performance in terms of an intelligence quotient or 'IQ' described by Stern (1912). This Binet measure gained much popularity in Europe, and was the impetus for the expansion of intelligence assessment once it was translated to English and introduced to North America by Henry Goddard. Lewis Terman (1916) developed the Stanford–Binet Intelligence Scale, which became the primary individually administered test for assessing intelligence.

A major advancement in intelligence assessment came during World War I when the United States developed the Army Alpha and Army Beta tests under the direction of Robert Yerkes (Yerkes, 1921). Designed to evaluate the cognitive intelligence of potential soldiers, the Group Examination Alpha (Army Alpha) measure was designed to assess literate

English speakers, whereas Group Examination Beta (Army Beta) evaluated illiterate or non-English speakers. The results of these group-based measures then informed decisions regarding suitability for wartime selection assignment of recruits.

Following his experiences with these early tests, David Wechsler developed the first of what is currently the most frequently used family of intelligence assessment measures by combining the verbal and nonverbal components of the two separate Army measures into a unitary assessment tool. The Wechsler–Bellevue (Wechsler, 1939) initially measured child and adult intelligence but has since been separated into adult, child, and preschool scales, each of which has undergone several revisions to update the subtests and items, revise and expand the factor structure, and update norms for each test. Many other intelligence tests have appeared during the twentieth century, some being variations on the Wechsler type tests, while others are based on different theoretical models or for other purposes (e.g., cross-cultural assessment, assessing hearing or visually impaired children).

Models for Describing Intelligence

Approaches to assessment of cognitive intelligence have typically followed from psychological theory on the broad topic of 'what constitutes human intelligence?' This topic has produced a number of long-standing debates that continue to this day. The primary discussion concerns the existence and importance of what is commonly referred to as the 'general factor of intelligence' or '*g*.' In essence, *g* refers to a statistical and theoretical convergence of performance on different measures of intelligence. As such, it is indicative of an individual's overall cognitive ability, regardless of the task by which cognitive intelligence is measured. It was initially proposed by Spearman (1904), who reported a series of significant positive intercorrelations (relationships) among various tasks that were thought to measure unrelated aspects of intelligence. Spearman proposed that all branches of intellectual activity have one fundamental function (or group of functions) in common (Spearman, 1904, p. 284). Indeed, proponents of *g* state that humans demonstrate a statistically measurable overarching level of cognitive ability, and then utilize that ability when completing individual cognitive tasks.

In contrast to this perspective, critics of *g* have cited contrary evidence of divergence in cognitive abilities (Thorndike et al., 1909) initially reflected in the description and assessment of 'primary mental abilities' by Thurstone (1938). Rather than supporting a unitary factor of intelligence, this alternative approach emphasizes the uniquely separate (or *s*) abilities such as mathematics, processing speed, or visual-spatial scanning.

Current perspectives on intelligence tend to identify the unique properties of both *g* and *s*, and highlight the interplay between an individual's unique performance on some tasks (or related tasks) and their statistically calculated overall performance, typically referred to as Full Scale Intelligence (FSIQ), which is taken to represent *g*.

While there are other viable models of intelligence including Piagetian and neo-Piagetian, neuropsychological, and chronometric descriptions, four contemporary theoretical models of intelligence will be briefly described with the aim of highlighting their core features and how this is then used in assessing intelligence. Specifically, we will review the current factor structure of the Wechsler family of cognitive abilities; the Cattell–Horn–Carroll (CHC) factor theory; the Planning, Attention Simultaneous, and Successive (PASS) theory; and Sternberg's Triarchic Theory of Intelligence.

Wechsler Factor Structure

The forthcoming *Wechsler Intelligence Scale for Children, 5th Edition* (WISC-V; Wechsler, 2014) and the current *Wechsler Adult Intelligence Scale, 4th Edition* (WAIS-IV; Wechsler, 2008) and *Wechsler Preschool and Primary Scales of Intelligence, 4th Edition* (WPPSI-IV, Wechsler, 2012) all incorporate *g* as a unitary indicator of overall global intelligence as represented by the FSIQ from each measure. In addition, the developers of the measures have painstakingly undertaken statistical analysis to determine the exact and specific underlying framework by which the individual subtests may be combined to provide an additional level of analysis between FSIQ (*g*) and the subtests (*s*). As such, psychologists are aware of the unique factor structure of each of the Wechsler measures and the resulting framework for understanding and interpreting the unique properties of cognitive intelligence demonstrated by an individual's performance on a standard administration of a measure.

Past versions of the Wechsler measures incorporated a verbal–nonverbal split as the main secondary factors subsumed by the FSIQ. Specifically, verbal abilities were represented by the Verbal IQ (VIQ) and nonverbal abilities by the Performance IQ (PIQ). These two domains were subsequently populated by subtests that were statistically shown (through factor analysis) to be strongly related to each other, and less related to subtests subsumed by the other factor. The current editions of the three measures have utilized new statistical data to support the adoption of a broader factor structure. Specifically, the WISC-V utilizes a five-factor structure with verbal abilities being represented by the Verbal Comprehension Index (VCI), visual-spatial abilities represented by the Visual-Spatial Index (VSI), fluid reasoning abilities represented by the Fluid Reasoning Index (FRI), the ability to maintain and mentally manipulate information in memory represented by the Working Memory Index (WMI), and speed of information processing by the Processing Speed Index (PSI). Conversely, the WAIS-IV uses a four-factor structure that includes the VCI, WMI, and PSI in addition to nonverbal (perceptual) reasoning abilities represented by the Perceptual Reasoning Index (PRI). The WPPSI-IV utilizes a three-factor structure (VCI, WMI, and VSI) or a five-factor structure (VCI, VSI, FRI, WMI, and PSI) depending upon the age of the child examinee.

Although some cognitive theorists argue against the factor structure of the Wechsler measures as pure indicators of human cognitive abilities (e.g., Benson et al., 2010), the Wechsler tests have received continued empirical support and are the most frequently used measures in clinical practice and research contexts. The scales have been standardized in many countries around the world and evidence suggests that the structure of the test holds up well in different countries (e.g., Georgas et al., 2003).

CHC Theory

The CHC model of human cognitive abilities is the most statistically rigorous theoretical explanation of human cognitive abilities. Similar to the Wechsler framework, CHC theory recognizes an overall level of cognitive ability that results from the overlap or correlation between the identified specific abilities. The major difference is the identification of the major broad factors that form the foundation of the Woodcock–Johnson Tests of Cognitive Abilities.

More specifically, CHC theory is a blending of two predominant theoretical frameworks. The Cattell–Horn model posits a primary division of human cognitive abilities on the basis of 'crystallized intelligence,' or the use of acquired knowledge, and 'fluid intelligence,' or abstract reasoning and problem solving. Initially, Raymond Cattell (1941, 1943) analyzed a large data collection representing numerous different tasks of intellectual abilities, the results of which indicated that tasks that tap crystallized intelligence are generally similar, whereas performance on tasks of fluid intelligence is also similar but distinct from crystallized intelligence tasks. Cattell began the tradition of utilizing specific notation to represent the different cognitive factors with a capital 'G' to represent general intelligence in combination with a lower case letter (e.g., *c* for crystallized or *f* for fluid) to represent the factors. Subsequent to this early description of G*c* and G*f*, John Horn (1965) expanded Cattell's theory to identify and include other factors such as cognitive speed (G*s*) and visual-spatial ability (G*v*). Additional research conducted over the next two decades continued to support and refine this model, with 10 unique factors being described by the 1990s (Schneider and McGrew, 2012).

Building upon this initial framework, John Carroll (1993) conducted a massive reanalysis of over 450 unique data sets on human cognitive abilities that resulted in a three-stratum model of human intelligence. In short, stratum III is the highest level, and is considered to represent *g*. Stratum II consists of the specific broad factors as initially described by Cattell and Horn. However, Carroll's reanalysis produced some modifications from Cattell and Horn's factors, with the resulting eight factors being crystallized intelligence (G*c*), fluid intelligence (G*f*), memory and learning (G*γ*), visual perception (G*v*), auditory perception (G*a*), long-term retrieval (G*lr*), cognitive speediness (G*s*), and reaction time/decision speed (G*t*). Finally, stratum I consists of several narrow abilities that are subsumed by each stratum II factor.

To date, the CHC model of human cognitive abilities is widely considered to be the most accurate representation of intelligence as measured by current cognitive tasks. It has been adapted for clinical use in the current third edition of the *Woodcock–Johnson Tests of Cognitive Abilities* (WJ-III Cog; Woodcock et al., 2001) and soon to be published fourth edition (WJ-IV Cog; Schrank, McGrew, Mather, 2014).

PASS Theory

The PASS theory of intelligence is rooted in Luria's work on brain structures and functionality (Luria, 1966, 1973a, 1973b). Luria posited that human cognition can be described as a system of three distinct yet related functional units (attention-arousal, processing and memory, and behavioral regulation) that support four psychological processes (Luria, 1973b). In turn, these four processes form a 'working constellation' of cognitive activity, which may enable different people to be successful on cognitive tasks via different cognitive processes (Luria, 1966, p. 70).

The first process in the PASS framework is Planning, which is the capacity to create, implement, and evaluate strategies to complete tasks effectively (Naglieri et al., 2012a). In effect, planning enables an individual to solve a problem and controls the other three processes. Attention involves vigilance, alertness, focus, and mental orientation to a stimulus over time and the resistance to distraction. Simultaneous Processing involves the mental integration of information into a whole and is often, though not solely, utilized in visual-spatial tasks. Conversely, Successive Processing is used to organize information into a sequential order and is more often associated with verbal information processing and reading.

The PASS theory of intelligence is operationalized by the Cognitive Assessment System–Second Edition (CAS2; Naglieri et al., 2012b) and has been found useful in the identification of children with disorders of attention or reading (Naglieri et al., 2012a). The second edition of the Kaufman Assessment Battery for Children (KABC-II; Kaufman, 2004) is another test that can be scored according to the Luria model as well as the CHC model described above.

Sternberg's Triarchic Theory

In comparison to the above models that describe cognitive abilities as factors, Robert Sternberg's conceptualization of intelligence views human cognition as a system of interrelated components that interact to allow an individual to adapt to the environment (Sternberg, 2012). Sternberg describes intelligence as internal characteristics of the individual, experiences that the individual encounters in their life, and the individual's use of their personal skills or attributes in order to adapt to and fit in with their environment. He views an individual as intelligent if they integrate and utilize their abilities to succeed within their sociocultural context. It should be noted that Sternberg's model differs significantly from the other three approaches described here in that the triarchic theory is not based on statistical approaches that provide evidence in favor of convergence of cognitive abilities (*g*) or specific and unique cognitive abilities (e.g., CHC theory). Additionally, this model has not contributed to the development of measures of intelligence but can serve as a heuristic when describing a person's intelligence observed in everyday behavior or as assessed by another intelligence test.

Models in Review

Researchers have devoted significant attention toward defining and describing human intelligence. To date, the two most widely utilized models are based on the notion of a unitary construct of general intelligence (*g*) that individuals demonstrate through varying cognitive tasks and the identification of the primary abilities that comprise the major components of intelligence. The Wechsler framework describes subcomponents of general

intelligence in terms of either four or five factors depending upon the evaluative measure being used and the age of the examinee. In contrast, CHC theory proposes that eight unique factors underlie *g* with a sizable amount of statistical evidence in support of this model. Each of these models has directly applied in the development of formal assessment measures. The PASS model describes human cognition via four interrelated processes that enable an individual to process information. Sternberg's triarchic theory differs substantially in that it emphasizes an individual's capacity to adapt to their unique sociocultural environment as a primary indicator of intelligence and as such, performance on formal measures is seen as less important than evidence of success within one's environment. In general, each model has supporters and critics, and theories of human cognition and intelligence continue to develop as a result of scientific inquiry and healthy debate.

Current Intellectual Assessment Measures

There are a very large number of intellectual assessment measures available for use by professionals (see the Buros Mental Measurements Yearbook and Tests in Print) who have obtained relevant training in standardized test administration and interpretation (i.e., psychologists, psychiatrists, physicians). These tests are designed to provide professionals with an accurate assessment of an individual's cognitive functioning across a number of domains (e.g., verbal abilities, nonverbal abilities, working memory, processing speed, cognitive fluency). Most intelligence measures provide the opportunity for both individual domain scores as well as a single intelligence score indicative of overall cognitive ability (FSIQ). It goes without saying that intelligence tests, like any other psychological measure, must demonstrate that they are valid and reliable; have been carefully standardized with comprehensive norms to aid interpretation; and have practical utility for the psychologist, the client, and any other agencies needing such information for decision-making (e.g., courts, school, military).

The following section briefly outlines two of the most commonly used batteries of cognitive intelligence for children and adults: the Wechsler family of tests and the Woodcock–Johnson Test of Cognitive Abilities. As well, a brief overview of other measures suitable for use in assessing cognitive abilities are also noted, although these measures are generally used less frequently than the Wechsler and Woodcock–Johnson tests.

Wechsler Tests

One of the most commonly used battery of intelligence measures are those initially developed by David Wechsler. His collection of tests allow for examining of intelligence from 2:6 to 90:11 years, with separate measures for preschoolers, children and adolescents, and adults. The Wechsler series also includes a brief measure of intelligence, the opportunity to measure nonverbal intelligence in language-limited populations (e.g., English language learners), and additional measures specifically examining memory and achievement.

The WPPSI-IV (Wechsler, 2012) was recently published in North America and provides a standardized measure of intelligence for preschoolers and young children from ages 2 years 6 months to 7 years 7 months. The WPPSI-IV utilizes child-friendly subscales (e.g., bug search, zoo locations) and engaging photos, and may be completed in as little as 30 min. This measure allows for a comprehensive understanding of early intellectual strengths and weaknesses, including verbal, visual-spatial, working memory, fluid reasoning, and processing speed indices (Syeda and Climie, 2014).

The most commonly used measure of intelligence for school-aged children and adolescents is the WISC-IV (Wechsler, 2003), which is now being superseded by the very recently published WISC-V (Wechsler, 2014). The WISC-V assesses the cognitive abilities of children and adolescents ages 6:0 to 16:11 years. Primary indices include verbal comprehension, visual-spatial awareness, fluid reasoning, working memory, and processing speed, as well as an overall FSIQ and various other composite scores such as the General Ability Index (GAI), incorporating verbal and visual abilities only (Prifitera et al., 2008).

The WAIS-IV (Wechsler, 2008) measures cognitive ability in adults ages 16:0 to 90:11 years. In contrast to the WISC-V, the WAIS-IV provides an understanding of cognitive abilities across four domains (verbal comprehension, perceptual reasoning, working memory, and processing speed) and two composite scores (FSIQ and GAI; Climie and Rostad, 2011; Weiss et al., 2010a).

The *Wechsler Abbreviated Scale of Intelligence – 2nd edition* (WASI-II; Wechsler, 2011) is an abbreviated measure of intelligence for individuals ages 6:0 to 90:11 years. It is appropriate for use to gain an accurate understanding of an individual's cognitive abilities but when it is not necessary or feasible to administer a full battery (e.g., WISC-V or WAIS-IV; McCrimmon and Smith, 2013). Administration time for this measure is approximately 30 min for the four-subtest form and 15 min for the two-subtest form. The WASI-II provides an understanding of verbal comprehension, perceptual reasoning, and an overall full scale score for the four-subtest (FSIQ-4) and two-subtest (FSIQ-2) administrations (McCrimmon and Smith, 2013).

Finally, the *Wechsler Nonverbal Scale of Ability* (WNV; Wechsler and Naglieri, 2006) provides a general cognitive ability assessment for individuals aged 4:0 through 21:11 years who have communication or language difficulties. It is intended for individuals who may have limited English language proficiency, developmental delays (e.g., limited language) or disorders (e.g., Autism Spectrum Disorder), deaf or hearing impaired individuals, or those with language-based learning disabilities (Massa and Rivera, 2009). The WNV provides only a full scale IQ score based on the four subtests comprising this measure.

Woodcock–Johnson Test of Cognitive Abilities

Another commonly used cognitive measure is the WJ-III (Woodcock et al., 2001) and the recently published WJ-IV Cognitive (Shrank et al., 2014). The WJ Cognitive measures can be utilized with individuals from preschool (2 years of age) to adults (over 90 years of age) and provides a strong measure of intelligence across the life span (Blackwell, 2001). Following the CHC model, the WJ-IV Cognitive provides

both standard (subtests 1 through 10) and extended (subtests 11–18) batteries, allowing for general (General Intellectual Ability, GIA; Brief Intellectual Ability, BIA; Fluid and Crystallized Composite, G*f*–G*c*) or detailed understanding of cognitive abilities in the following areas: Comprehension-Knowledge (G*c*); Long-Term Retrieval (G*lr*); Visual Processing (G*v*); Auditory Processing (G*a*); Fluid Reasoning (G*f*); Processing Speed (G*s*); Short-Term Working Memory (G*wm*); Perceptual Speed (P); and Number Facility (N). Additionally, Auditory Memory Span (MS) can be ascertained in combination with additional tasks from the WJ-IV Cognitive battery of measures.

Other Measures

Although the Wechsler and Woodcock–Johnson measures are the most frequently used intellectual assessment measures, other often used measures to examine intelligence include:

- *Stanford–Binet Intelligence Scales – 5th edition* (SB5; Roid, 2003) – the SB5 is a well-known measure of intelligence for individuals aged 2–85+ years of age with a rich history dating back to 1916.
- *Kaufman Assessment Battery for Children – 2nd edition* (Kaufman, 2004): the KABC-II is a culturally sensitive measure of ability for children aged 3–18 years of age.
- *Cognitive Assessment System – 2nd edition* (Naglieri et al., 2012b): the CAS2 is used with individuals 5:0 to 18:11 years of age and was developed following the PASS theoretical framework.
- *Differential Ability Scale – 2nd edition* (DAS-II; Elliott, 2006): the DAS-II uses profile analysis to identify cognitive strengths and weaknesses in individuals aged 2:6 through 17:11 years of age.

There are other intelligence tests available. Some are variations of the above but in shorter formats (e.g., WASI-II), while others focus on specific abilities such as the Raven's Progressive Matrices (Raven et al., 2003) or the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1997). Still other tests are intended for use when language presents a barrier to assessment, such as the WNV (Wechsler and Naglieri, 2006) or the Universal Nonverbal Intelligence Test (UNIT; Bracken and McCallum, 1998). Together, these cognitive assessment measures provide a detailed understanding of individual cognitive intelligence and allow for a reliable and accurate measure of cognitive functioning whether defined more broadly or specifically.

What Do Intelligence Tests Tell Us?

Broadly speaking, intelligence tests allow for the assessment of a person's cognitive functioning for purposes of identifying those who demonstrate average intelligence, intellectual giftedness, delays, or disability. This information may be particularly useful in the identification of intellectually gifted persons and the diagnosis, intervention planning, and remedial program development for children with learning difficulties, adolescents or adults who may require additional supported living arrangements, or persons who have are suffering intellectual decline due to head injury or dementia.

Furthermore, an examination of different score patterns using tests such as the WISC-V or WJ-IV Cog allows for strength- and deficit-based analysis of examinees' cognitive functioning, particularly in relation to distinct cognitive domains. This information may be particularly useful when creating individual learning plans for school-aged children and adolescents, and may allow individuals working with intellectually challenged adults the opportunity to build on their strengths rather than remain limited by their areas of difficulty.

Social and Professional Issues in Cognitive Assessment

As with any type of assessment, a number of social as well as ethical and professional issues must be considered when utilizing intellectual assessment measures. The following section highlights three important issues, including the necessity of appropriate training and supervision before administering cognitive assessment measures, the role of intelligence assessment with special populations (e.g., deaf or hard of hearing, intellectual disability, Autism Spectrum Disorder), and adaptations for use with culturally diverse individuals.

The first issue concerns the necessity of appropriate training and supervision for individuals who administer cognitive assessment measures. The most complex cognitive assessment measures are identified as 'level C' tests, indicating that the test is restricted to use by professionals with expert training, generally including graduate-level training in psychology (usually a doctorate degree, including specific assessment courses and supervised practice in assessment) or other related professional field, and/or registration in a state, provincial, or national organization that is related to these fields. Test publishers require proof of these credentials before an individual is allowed to purchase these measures. This check provides some control over the sale of these protected measures so as to ensure that test integrity is maintained and that only those with appropriate training and expertise in these areas are permitted to access or administer these measures. There are clear concerns with unauthorized individuals accessing these measures, including inappropriate assessment, diagnosis, or interpretation of results, and it is imperative that these assessments are conducted by appropriately trained individuals who maintain and continue to upgrade their knowledge and skills related to best assessment and ethical practices. Psychologists and others who use intelligence tests for assessment and diagnostic purposes must also adhere to the required standards for professional practice as well as the code of ethics required of them as part of their credentialing.

Second, the role of assessment in unique populations has continued to gain attention over the last few decades stemming from the earlier court challenges in the United States from some 40 years ago. Specifically, the purpose for which intelligence tests are used and the validity of these assessment measures with individuals with special learning and psychological needs has been questioned, leading to a greater emphasis on appropriate and adaptive testing measures. For example, the use of verbally loaded cognitive assessments with

language-limited individuals is clearly problematic. Individuals may present with language difficulties for a number of reasons (e.g., lack of English language knowledge, language disorder, Autism Spectrum Disorder) but may otherwise have intact cognitive abilities. As such, it is important to provide these individuals with an opportunity to demonstrate their intellectual abilities without a reliance on language that could certainly mask important abilities and result in incorrect and misleading conclusions. The WNV (Wechsler and Naglieri, 2006) provides some opportunity to measure cognitive abilities through nonverbal means. As well, other small test modifications may be necessary to ensure that all examinees receive a fair assessment. These modifications may include additional breaks between subtests (for individuals unable to concentrate for long periods of time), removal of fine motor-based subtests (for individuals with motor difficulties, such as cerebral palsy), or beginning with easier test questions for cognitively challenged individuals. These minor modifications may not affect the overall validity of the test and may provide a more accurate understanding of individual skills. A psychologist may also use a more process-oriented approach (e.g., WISC-IV Integrated; Wechsler et al., 2004) or different methods of assessing intelligence, such as dynamic assessment described by Feuerstein et al. (1979, 2002). Of course, it is key to remember that the tests themselves do not make 'decisions' about an individual, psychologists do! Thus, they must be extensively trained to understand and interpret the information gleaned from intelligence tests and other sources of information (e.g., observation, clinical history) in relation to the scores obtained from tests.

Third, the necessity of incorporating a social and cross-cultural understanding into cognitive assessment measures has gained a greater focus (Hambleton et al., 2004; Weiss et al., 2010b). Given the diverse nature of Canadian and American populations but also the very fact that most countries are becoming increasingly multicultural, the consideration of cultural implications in assessment is imperative to ensure that bias does not enter the assessment picture. Bias in assessment can occur when individuals score differently on questions or test measures as a result of individual circumstances and not as a result of a basic underlying trait or ability (van de Vijver and Tanzer, 2004). For example, asking a child 'What is the capital of Kentucky?' may yield a very different rate of accuracy in the United States compared to responses from African or European children. Indeed, even within the United States, accuracy on this question may vary based on individual geographic location (e.g., living in Kentucky versus living in Idaho) or the amount of focus on geographical capital cities within an individual classroom. Although this example is only one illustration of potential bias, on a broader scale, it is important to consider the impact of tests developed and normed in one country being used to assess an immigrant population who may not have the language capabilities or the background information to adequately demonstrate their knowledge and competencies reflective of their underlying intelligence.

Although not exhaustive, this section has noted some of the more significant concerns that must be accounted for in intellectual assessment. Together, these issues highlight some of the considerations necessary for using intellectual assessment measures with unique and/or specialized populations.

Conclusions

Overall, this article has provided a general summary of intelligence and highlighted both historical and conceptual frameworks. A description of the most commonly used cognitive assessment measures is included. Finally, some important social and professional issues are discussed with an emphasis on the ethical use of these measures. Together, it is clear that intelligence and its measurement have evolved over time and should continue to do so as our understanding of cognitive development expands.

See also: Cognitive Development During Infancy and Early Childhood across Cultures; Intellectual Functioning, Assessment of; Intelligence: Central Conceptions and Psychometric Models; Managerial and Organizational Cognition; Personal Intelligence and Competencies.

Bibliography

Benson, N., Hulac, D.M., Kranzler, J.H., 2010. Independent examination of the Wechsler adult intelligence scale – fourth edition (WAIS-IV): what does the WAIS-IV measure? Psychological Assessment 22 (1), 121.

Binet, A., Simon, T., 1905. Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. L'Annee Psychologique 11, 191–244.

Blackwell, T.L., 2001. Test review: Woodcock, R.W., McGrew, K.S., & Mather, N. (2001). Woodcock-Johnson III test. Riverside Publishing Company. Itasca, IL. Rehabilitation Counseling Bulletin 44 (4), 232–235. http://dx.doi.org/10.1177/003435520104400407.

Bracken, B.A., McCallum, S., 1998. Universal Nonverbal Intelligence Test. Riverside, Chicago.

Carroll, J.B., 1993. Human Cognitive Abilities: A Survey of Factor Analytic Studies. Cambridge University Press, New York.

Cattell, R.B., 1941. Some theoretical issues in adult intelligence testing. Psychological Bulletin 38, 592.

Cattell, R.B., 1943. The measurement of adult intelligence. Psychological Bulletin 40 (3), 153–193.

Climie, E.A., Rostad, K., 2011. Review of Wechsler adult intelligence scale – 4th edition. Journal of Psychoeducational Assessment 29 (6), 581–586. http://dx.doi.org/10.1177/0734282911408707.

Elliott, C.D., 2006. Differential Ability Scales, second ed. (DAS-II). Psychological Corporation, San Antonio, TX.

Feuerstein, R., Feuerstein, S., Falik, L., Rand, Y., 1979, 2002. Dynamic Assessments of Cognitive Modifiability. ICELP Press, Jerusalem: Israel.

Galton, F., 1885. On the anthropometric laboratory of the late international health exhibition. Journal of the Anthropological Institute of Great Britain and Ireland 14, 205–221.

Galton, F., 1887. Supplementary notes on 'prehension' in idiots. Mind 12, 79–82.

Gardner, H., 1983. Frames of Mind. Basic Books, Inc, New York, NY.

Georgas, J., Weiss, L.G., van de Vijver, F.J., Saklofske, D.H. (Eds.), 2003. Culture and Children's Intelligence: Cross-Cultural Analysis of the WISC-III. Academic Press, New York.

Gottfredson, L.S., 1997. Mainstream science on intelligence: an editorial with 52 signatories, history, and bibliography. Intelligence 24 (1), 13–23.

Gottfredson, L.S., 2008. Of what value is intelligence? In: Prifitera, A., Saklofske, D.H., Weiss, L.G. (Eds.), WISC-IV Applications for Clinical Assessment and Intervention. Elsevier, Amsterdam, pp. 545–563.

Gottfredson, L., Saklofske, D.H., 2009. Intelligence: foundations and issues in assessment. Canadian Psychology/Psychologie canadienne 50 (3), 183–195.

Hambleton, R.K., Merenda, P.F., Spielberger, C.D. (Eds.), 2004. Adapting Educational and Psychological Tests for Cross-Cultural Assessment. Lawrence Erlbaum Associates, Inc, Manwah, NJ.

Horn, J.L., 1965. Fluid and Crystallized Intelligence: A Factor Analytic Study of the Structure Among Primary Mental Abilities (unpublished Doctoral dissertation). University of Illinois at Urbana-Champaign.

Kaufman, A.S., 2004. KABC-II: Kaufman Assessment Battery for Children. AGS Pub.

Kaufman, A.S., Kaufman, N.L., 2004. Kaufman Assessment Battery for Children–II (KABC-II). American Guidance Service, Circle Pines, MN.

Luria, A.R., 1966. Human Brain and Psychological Processes. Harper & Row, New York.

Luria, A.R., 1973a. The origin and cerebral organization of man's conscious action. In: Sapir, S.G., Nitzburg, A.C. (Eds.), Children with Learning Problems. Brunner/Mazel, New York, pp. 109–130.

Luria, A.R., 1973b. The Working Brain. Basic Books, New York.

Massa, I., Rivera, V., 2009. Test review: Wechsler, D., & Naglieri, J.A. (2006). Wechsler nonverbal scale of ability. Journal of Psychoeducational Assessment 27 (5), 426–432.

McCrimmon, A.W., Smith, A.D., 2013. Review of the Wechsler abbreviated scale of intelligence, second edition (WASI-II). Journal of Psychoeducational Assessment 31 (3), 337–341. http://dx.doi.org/10.1177/0734282912467756.

Naglieri, J.A., 1997. Naglieri Nonverbal Ability Test. The Psychological Corporation, San Antonio, TX.

Naglieri, J.A., Das, J.P., Goldstein, S., 2012a. Planning, attention, simultaneous, successive: a cognitive-processing-based theory of intelligence. In: Flanagan, D.P., Harrison, P.L. (Eds.), Contemporary Intellectual Assessment: Theories, Tests, and Issues. Guilford, New York, pp. 178–194.

Naglieri, J.A., Das, J.P., Goldstein, S., 2012b. Cognitive Assessment System, second ed. PRO-ED, Austin, TX.

Neisser, U., Boodoo, G., Bouchard, T.J., Boykin, A.W., Brody, N., Ceci, S.J., Halpern, D.F., Loehlin, J.C., Perloff, R., Sternberg, R.J., Urbina, S., 1996. Intelligence: knowns and unknowns. American Psychologist 51, 77–101.

Prifitera, A., Saklofske, D.H., Weiss, L.G. (Eds.), 2008. WISC-IV Clinical Interpretation and Intervention. Academic Press, San Diego.

Raven, J., Raven, J.C., Court, J.H., 2003, updated 2004. Manual for Raven's Progressive Matrices and Vocabulary Scales. Pearson Assessment, San Antonio, TX.

Roid, G.H., 2003. Stanford-Binet Intelligence Scales, (SB5). Riverside, Rolling Meadows, IL.

Schneider, W.J., McGrew, K.S., 2012. The Cattell-Horn-Carroll model of intelligence. In: Flanagan, D.P., Harrison, P.L. (Eds.), Contemporary Intellectual Assessment: Theories, Tests, and Issues, third ed. Guilford, New York, pp. 99–144.

Shrank, F., Mather, N., McGrew, K.S., 2014. Woodcock-Johnson IV Test of Cognitive Abilities. Riverside, Rolling Meadows, IL.

Spearman, C., 1904. "General intelligence," objectively determined and measured. The American Journal of Psychology 15 (2), 201–292.

Stern, W., 1912. The Psychological Methods of Intelligence Testing (G. Whipple, Trans.). Warwick and York, Baltimore.

Sternberg, R.J., 2012. The triarchic theory of successful intelligence. In: Flanagan, D.P., Harrison, P.L. (Eds.), Contemporary Intellectual Assessment: Theories, Tests, and Issues, third ed., pp. 156–177.

Syeda, M., Climie, E.A., 2014. Wechsler preschool and primary scale of intelligence – 4th edition. Journal of Psychoeducational Assessment 32 (3), 265–272.

Terman, L.M., 1916. The Measurement of Intelligence: An Explanation of and a Complete Guide for the Use of the Stanford Revision and Extension of the Binet-Simon Intelligence Scale. Houghton Mifflin, Boston.

Thorndike, E.L., Lay, W., Dean, P.R., 1909. The relation of accuracy in sensory discrimination to general intelligence. The American Journal of Psychology 20 (3), 364–369.

Thurstone, L.L., 1938. Primary mental abilities. University of Chicago Press, Chicago.

van de Vijver, F., Tanzer, N.K., 2004. Bias and equivalence in cross-cultural assessment: an overview. Revue Européenne de Psychologie Appliquée/ European Review of Applied Psychology 54 (2), 119–135.

Wechsler, D., 2003. Wechsler intelligence scale for children, fourth ed. (WISC-IV). The Psychological Corporation, San Antonio, TX.

Wechsler, D., 2014. Wechsler Intelligence Scale for Children, fifth ed. The Psychological Corporation, San Antonio, TX.

Wechsler, D., 2012. Wechsler Preschool and Primary Scale of Intelligence, fourth ed. technical and interpretative manual. The Psychological Corporation, San Antonio, TX.

Wechsler, D., 2011. Wechsler Abbreviated Scale of Intelligence, second ed. (WASI-II). NCS Pearson, San Antonio, TX.

Wechsler, D., 2008. WAIS-IV Administration and Scoring Manual. The Psychological Corporation, San Antonio, TX.

Wechsler, D., 1939. The Measurement of Adult Intelligence. Williams & Wilkins, Baltimore.

Wechsler, D., Naglieri, J., 2006. Wechsler Nonverbal Scale of Ability (WNV). Harcourt Assessment, San Antonio, TX.

Wechsler, D., Kaplan, E., Fein, D., Morris, R., Kramer, J.H., Maerlender, A., Delis, D.C., 2004. Wechsler Intelligence Scale for Children, fourth ed. Integrated. Pearson, San Antonio, TX.

Weiss, L.G., Saklofske, D.H., Coalson, D., Raiford, S.E. (Eds.), 2010a. WAIS-IV Clinical Use and Interpretation. Academic Press, San Diego.

Weiss, L., Chen, H., Harris, J., Holdnack, J., Saklofske, D.H., 2010b. WAIS-IV use in societal context. In: Weiss, L., Saklofske, D.H., Coalson, D., Raiford, S.E. (Eds.), WAIS-IV Clinical Use and Interpretation: Scientist-Practitioner Perspectives. Academic Press, San Diego, pp. 97–140.

Woodcock, Richard W., Mather, Nancy, McGrew, Kevin S., 2001. Woodcock–Johnson III Tests of Cognitive Abilities Examiner's Manual. Riverside, Itasca.

Yerkes, R.M. (Ed.), 1921. Psychological Examining in the United States Army. Memoirs of the National Academy of Sciences, 15 (Pt. 1–3). Government Printing Office, Washington, DC.

Intelligence: Central Conceptions and Psychometric Models

Florian Schmiedek, German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany
Shu-Chen Li, TU Dresden, Dresden, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Scientific inquiries into the nature of intelligence have journeyed for more than a century and have yielded a broad collection of theoretical conceptions and measurements. This article seeks to provide an overview of research on intelligence. Central theoretical conceptions and related measurement issues are reviewed along three orientations (i.e., the product-process, the human-ecology, the lifespan developmental) as well as their relative emphases on interactions between these dimensions.

Lasting Debates in Intelligence Research: The Elusive Concept of Intelligence

Given its practical importance in individual lives and societies, the topic of 'intelligence' attracts much attention and debates both among the researchers and the lay persons alike. The research of intelligence has been pursued by psychometricians, cognitive psychologists, and developmental researchers (see Sternberg, 1994 for comprehensive reviews; see also Intelligence: Historical and Conceptual Perspectives). However, not many other terms in psychology are so elusive as 'intelligence' escaping consensual definitions over more than a century of research. Tracking chronologically the reports of three representative forums, each participated by experts in the field, sheds some light on the changing terrain of intelligence research during the twentieth century.

In 1921, a classical symposium convened by the editors of the *Journal of Educational Psychology* was held to discuss three questions: (1) What is intelligence? (2) How can it be best measured? (3) What are the most crucial next steps in research? (Thorndike, 1921). Among the 17 leading researchers who participated, 14 different answers were given! About one-fourth of the participants suggested elementary sensory and cognitive processes (i.e., perception, sensation, and attention) as primary attributes of intelligence. Another one-fourth of the experts thought that physiological mechanisms of the brain should be the determining factors, and there were still others who contented that overt adaptive behavioral responses are the key features of intelligence. Sixty-five years later, a succeeding effort with 27 experts was arranged to address exactly the same questions (Sternberg and Detterman, 1986). However, the degree of lack of consensus remains similar to what was in 1921: experts of the 1986 forum offered two dozen definitions.

Aside from the diverse conceptions among researchers, two popular books also sparked heated debates in the general public. The views range from renouncing the idea of 'measuring' individuals along the continuum of psychometrically defined intelligence (Gould, 1981) to the belief in a general intelligence factor (Herrnstein and Murray, 1994). Reacting to these controversial debates, at the dawn of approaching a new century the Board of Scientific Affairs of the American Psychological Association called for another authoritative report from a dozen experts in the field to discuss the 'knowns and unknowns' of intelligence (see the report in Neisser et al., 1996). Now we are more than a decade into the twenty-first century, a recent comprehensive update by Nisbett et al. (2012) reviews substantial research progress that has been made in the last 15 years regarding important topics like the heritability and modifiability of intelligence, group differences in intelligence, as well as its neurobiological basis. As for concepts about the definition and structure of intelligence, little new theoretical work has been added in the past decade. There remains a need for integrative frameworks to understand the diversities and controversies surrounding theoretical conceptions and the multivariate determinants of individual differences in intelligence.

Some Lasting Trends and an Attempt of Integration

In order to highlight a few continuing trends that are hidden behind the overwhelming diversities, organizers of the 1986 forum compared the main issues raised in their discussions with those that were discussed in 1921 and showed that there were some general agreements across the two efforts (Sternberg and Detterman, 1986). Specifically, conceptions including attributes such as *adaptation to the environment*, *basic cognitive processes*, and *higher-order thinking* (i.e., reasoning, problem-solving, and decision-making) were prominent in both discussions. In addition, the recurring themes of central debates seemed to surround issues about (1) whether intelligence is unitary or manifold, (2) the relative contributions of nature and nurture to individual differences in intelligence, and (3) the relative emphases on the process, content, and product aspects of intelligence.

As an attempt to compile the two dozen different views, an integrative framework (Sternberg and Detterman, 1986) with multiple levels was also proposed in the report of the 1986 forum. At the behavioral level, theoretical conceptions that originated within the psychometric tradition focus on measuring the 'products' of information processing and on identifying the structure(s) of intelligence. These theories could be compared mainly in terms of whether the structure of intelligence is, more or less, considered to be unitary or pluralistic. At the processing level, the theories attempt to integrate psychometrically defined 'products of intelligence' with various concepts of information processing. In this regard, the various conceptions differ in terms of their differential emphases either on elementary cognitive processes or higher-order cognition involving metacognition, planning, and decision-making. At the contextual level, some theories of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.25014-9

intelligence emphasize the dynamics of human–ecology interaction and the concepts of situated intelligence. Another version of the contextualist theories emphasizes the dynamic contexts of lifespan development from birth into old age. These contextualist theories evolved primarily in conjunction with a focus on the nature of intelligence as adaptive capacity.

Still there have been other theories focusing on genetic and neurobiological influences on intelligence, which are not within the purviews of this article. In the following sections, we review and compare main theoretical conceptions of intelligence with respect to their differences along the product-process, the contexts of human-ecology, and the lifespan developmental orientation, as well as their relative emphases on interactions between these dimensions.

Measuring the Products of Intelligence: From Unitary to Pluralistic and Hierarchical Views

Ever since Alfred Binet's success in devising tests to distinguish mentally retarded children from those with behavioral problems, the psychometric approach to quantify individual differences in cognitive competence using standardized intelligence scales (e.g., the Stanford–Binet test or the Wechsler test) played a dominant role in intelligence research. These scales tend to measure scholastic performance with tasks covering broad domains of functions (e.g., memory, analytical reasoning, numerical skills, verbal fluency, and practical knowledge). Thus, overall the sum scales of such tests predict school performance well. Perhaps not surprisingly because of their emphasis on school-related performances, their predictive validity drops when predicting performance outside of the classroom, such as everyday problem-solving, vocational competence (Dixon and Baltes, 1986), or other more expertiselike outcomes of real-world intellectual functioning (Ericsson and Smith, 1991).

With a more general aim to capture a broader spectrum of human cognitive abilities and their interrelations in addition to the applied orientation of predicting academic performance, a number of psychometric theories have been proposed over last century. Scholars such as Burt, Cattell, Guilford, Spearman, and Thurstone were the main proponents. At the methodological level, these theories differ with respect to the type of factor analytic methods they adopt and whether these methods are applied on the empirical-exploratory or theoretical-confirmatory basis (*see* Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis; Factor Analysis and Latent Variable Models in Personality Psychology). At the substantive level, they differ in terms of whether they orient more toward the general factor or multidimensional views of intelligence.

Two-Factor Theory

Although subtests of the intelligence scales are often designed to measure specific domains of cognitive performance (like numerical skill and verbal fluency), performances across different tests tend to be positively correlated with each other (i.e., the so-called *positive manifold*). Spearman (1904) first presented a theory of general intelligence which he later expended into what is now known as the two-factor theory. He developed the first statistical model for explaining the positive manifold among task performances. The underlying assumption of this model is that individual differences in a given test can be decomposed into two components. One variance component is common to all other cognitive tasks – a general factor of intelligence (commonly known as *g*), whereas the other component is specific to each task (Spearman, 1927). With this method, Spearman showed with several data sets that *g* accounted for all positive correlations among the test performances (see Jensen, 1998; for a review on *g*). Since then, hundreds of other studies have also found that *g* accounts for rather large portions of the common variance among a broad selection of cognitive tasks.

The Primary Mental Ability Theory

Since the discovery of *g*, a main issue in psychometric intelligence research has been whether the structure of intelligence is unitary with one general common factor or pluralistic with multiple specific factors reflecting domain-specific abilities. Among the first to question the unitary view was Thurstone, who developed multiple factor analytic techniques and conducted studies that included a broad range of paper-and-pencil items found in most other intelligence scales. In these studies, Thurstone (1938), for instance, was able to identify seven distinct factors reflecting spatial, perceptual speed, numerical, verbal, fluency, memory, and reasoning abilities. However, his assumption of these primary mental abilities being completely orthogonal to each other was not empirically supported. Instead, multiple common factor models later proposed by others showed that the positive manifold emerges again at the level of higher-order factors.

The *Gf-Gc* Theory and Related Theories

Building on the multiple primary abilities identified by Thurstone, one influential conceptual model since then has been the *Gf-Gc* theory (Cattell, 1971; Horn, 1982). Based on results from higher-order factor analyses, this theory distinguishes general fluid (*Gf*) abilities reflecting the individual's basic problem-solving, reasoning, learning, and selective attention capabilities from general crystallized (*Gc*) abilities reflecting the acquired knowledge and their utilization. Unlike Thurstone's theory, which assumes independent primary mental abilities, Cattell proposed the ontogenetic *investment theory* to specify a possible mechanism relating *Gf* to *Gc*. He postulated that the culture-, experience-, and knowledge-based crystallized abilities arise, in part, as a result of a process in which the more biology-, maturation- and process-based fluid abilities have been *invested* into suitable learning experiences and contexts as the individual develops. The *Gf* and *Gc* distinction is often compared with Hebb's (1949) differentiation between innate potentials for information processing (Intelligence A) and the acquired level of performance and comprehension (Intelligence B), and Paul B. Baltes' (1987) juxtaposition of the cognitive mechanics and the cognitive pragmatics (see more details in a later section). Besides these two central broad factors, Cattell and Horn were also able to identify several other factors reflecting visual processing ability (*Gv*), short-term memory (*Gsm*), quantitative abilities (*Gq*), associative

storage-retrieval (*Glr*), novel reasoning (*Gr*), and processing speed (*Gs*). These factors have been further verified and captured by comprehensive psychometric test batteries developed by others, such as the Stanford–Binet, Wechsler, and Woodcock–Johnson batteries.

The Structure of Intellect Model

Moving even further away from the unitary pole of *g* and working toward better delineating the process and product aspects of intelligence, Guilford (1967) proposed the structure of intellect (SOI) model, which specifies the (1) operation, (2) content, and (3) product components of intellectual abilities. The component of mental operation contains five general intellectual processes: cognitive, memory, divergent production, convergent production, and evaluation. The content component encompasses four broad areas (i.e., figural, symbolic, semantic, and behavioral) of information to which operations are applied. The product component entails six possible results (i.e., units, classes, relations, systems' transformations, and implications) of applying particular operations in particular contents. Applying factor analytic techniques in a confirmatory manner to test this model deductively, Guilford expected the structure of intellect to result in 120 single independent factors of different operation × content × product ($5 \times 4 \times 6$) combinations. However, in later modifications he acknowledged that the first-order factors could be correlated, therefore giving rise to multiple higher-order factors, but he still denied the existence of a general factor (Guilford, 1982). On the one hand, the strength of the SOI theory is its more specific and refined definition of ability structure across dimensions capturing three distinct aspects of information processing. On the other hand, the main critiques of the model targeted at the fact that this model was derived through the use of a certain factor rotation technique (i.e., the procrustean rotation) which later was criticized to be subjective to confirmatory bias and lacking empirical support for most of the factors specified.

Another theoretical model akin to Guilford's faceted conceptualization is the Berlin model of intelligence structure (BIS) by Jäger (1984). This model cross-classifies each cognitive test with respect to an (1) operation and a (2) content facet. Operation is further categorized into four ability groups (reasoning, memory, creativity, and speed), whereas content is divided into three groups (figural, numerical, and verbal). Conceptually, the 12 operation × content (4×3) combinations taken as a whole represents general intelligence. However, by factoring groups of variables that are homogeneous within one of the two facets but are heterogeneous in the other facet, one can either extract the four operational or the three content factors. Specified as such, methodologically this model implies that any given mental ability is a linear combination of the two factors, and conceptually it simultaneously specifies a mental ability with a process and a content aspect.

The Three-Stratum Theory

Reanalyzing 477 available covariance matrices of cognitive ability tasks, Carroll (1993) integrated the findings from the different theoretical schools into a hierarchical structure with three stratums. At the lowest stratum, more than 60 narrow first-order abilities could be found. Given these narrow abilities are positively correlated with each other to varying degrees, a second stratum of broad abilities, which closely resemble the primary mental abilities identified by Thurstone and by Cattell and Horn, could also be extracted. At the highest stratum, a general factor of intelligence captures the positive manifold among the broad ability factors at the second level. The hierarchy with all three stratums then describes the overall structure of intelligence. In terms of empirical breadth, this model comprehensively integrates the empirical basis of human cognitive abilities observed at the behavioral level. In terms of theoretical integration, it shows that diverging psychometric views of intelligence can be understood in terms of their differential focus on one of the stratums within the overall structure, selections of indicator tasks defining the lowest level of the structure, and adoptions of factor-analytical techniques.

Regarding the behavioral psychometric approach to intelligence research as a whole, a key issue that remains is whether factors derived by factor analysis are entities with causal status or organizational concepts capturing a system of the products of intelligence. Defining the structure of intelligence with multiple levels, going from multiple low-level specific abilities to intermediate broad abilities, and, finally, to a general factor at the highest level, has advantages for relating to other theoretical conceptions. Various conceptual connections from psychometric measures of intelligence to other theoretical conceptions that either emphasize inward cross-level integration with information-processing mechanisms and their biological correlates, outward integration with real-world everyday knowledge and problem-solving skills, or lifespan developmental integration can be more easily drawn.

Integrating Psychometric Intelligence with Cognitive Processes

Starting from late 1970s, there has been a development toward specifying the processing aspects of intelligence in hierarchical models within the psychometric tradition and the development of investigating information-processing mechanisms in cognitive science in general. With the parallel developments in both fields, researchers of intelligence started to integrate information-processing concepts into the theoretical conceptions of intelligence. Some theories focus on integrating the measurement and explanation of intelligence by looking inwardly and downwardly within the individual for elementary information-processing correlates (e.g., processing speed and working memory) of intellectual abilities (e.g., Hunt, 1980; Resnick, 1976) and their neurobiological correlates (Eysenck, 1982; see Deary, 2000 for a recent overview). Other theories focus on relating intelligence with aspects of information processing that are more closely related to real-world competence, such as complex reasoning, metacognition, planning, knowledge, vocational skills, and expertise (e.g., Ackerman, 1996; Sternberg, 1985, 1999).

Mental Speed Theory

The notion that information-processing speed may underlie individual differences in intelligence goes back all the way to

Galton, who tried to measure intelligence by simple sensory and reaction time tasks. Researchers dedicated to the mental speed approach have taken up Galton's initial idea to investigate the relation between intelligence and reaction times measured in a broad range of simple elementary cognitive tasks (see Jensen, 1998; Vernon, 1987; for reviews). Performances in these tasks tend to be perceptually based and do not require any semantic or practical knowledge. Thus, it is perhaps no surprise that the correlations between the reaction times measured with such simple tasks and intelligence have proven to be only moderate (i.e., hovering around the so called 0.3 barrier, Hunt, 1980). Proponents of the elementary information-processing theory in general and the mental-speed theory in particular admit that simple reaction time measures are inferior to more complex composite measures derived from mental tests for the measurement and prediction of intelligence. However, they argued that the moderate but consistent correlations between measures derived from simple elementary information-processing tasks and performance on mental tests may be important for unraveling the information-processing aspects of intelligence. Variations in the observed correlations across studies and their theoretical interpretations have been the issues of much discussion. The interpretations range from proposing nerve conduction velocity as the basic biological parameter for all cognitive functioning, to several alternative explanations postulating higher-level constructs like attention, motivation, personality factors, or speed-accuracy trade-offs, to be the causes for a spurious relationship between intelligence and mental speed measured with simple decontextualized tasks.

Working Memory Theory

Another aspect of cognitive function that has been deemed as a likely candidate information-processing correlate of intelligence is working memory (Baddeley, 1992). Working memory reflects people's ability to briefly maintain information, while simultaneously carrying out other mental operations on the same or other information. Empirical evidence suggests that tasks requiring greater working memory loads tend to correlate more highly with intelligence. Furthermore, it has also been found that working memory capacity correlates highly with psychometric reasoning (Kane et al., 2005; Kyllonen and Christal, 1990). Taken these findings together, proponents of this theory suggest that working memory might be a better candidate process for explaining fluid intelligence than mental speed (Engle et al., 1999; Süß et al., 2002).

The PPIK Theory

Staying with the general sentiment of relating intelligence with information processing, but orienting more toward issues on knowledge representation in artificial intelligence systems rather than the experimental cognitive approach, Ackerman (1996) proposed an integrative framework characterizing adult intellectual development as complex interactions between cognitive processes, personality traits, broad interest, and knowledge (PPIK) domains. This framework can be seen as a starting point for disentangling the complex relations between the different aspects of intellectual functioning and personality-motivational modulators (Ackerman and Heggestad, 1997). For instance, there is ample evidence indicating that personality traits, such as degrees of openness to experience and typical intellectual engagement, are substantially correlated with crystallized knowledge (von Stumm and Ackerman, 2013), but much less so with fluid ability. Such findings motivate further intelligence research to explore the roles of broad domains of occupational and avocational knowledge in adult intelligence, which has not been captured sufficiently by the traditional assessment of *Gc*.

The Triarchic Theory

Sternberg's triarchic theory of intelligence (1985) is another example of an integrative theory that seeks to combine traditional psychometric approach with information-processing perspectives at the level of higher-order cognition, rather than elementary cognitive processes. Three different kind of information-processing components are distinguished in the triarchic theory. *Metacognitive components* are higher-order control processes used for executive planning, monitoring, and evaluating task performance. *Performance components* are lower-order processes that deal with the execution of a task itself. For instance, in an inductive reasoning task, the performance components may involve encoding the test word pairs, comparing the word pairs, and retrieve from memory information about the test items. The *knowledge-acquisition components* are processes involved in learning, retaining, and integrating new information with old experiences. Sternberg (1999) had extended his original theory to encompass a broader spectrum of issues regarding successful everyday real-world competence. For instance, in everyday context, successful intelligent individuals are likely to invoke metacognitive higher-order regulatory thought processes to discern their own strengths and weaknesses, and device strategies to maximize their strengths but compensate for their weaknesses. While this theory is more integrative and ecologically valid in nature, details of many components of the theory await further specification and empirical testing.

Bioecological and Lifespan Theories of Intellectual Development

Traditionally, developmentalists have also been among the researchers of intelligence. Piaget, the pioneer of modern developmental psychology (*see* Piaget's Theory of Cognitive Development), studied the development of mental processes through which knowledge is constructed during ontogenesis (Chapman, 1988). The Piagetian view of intelligence focuses more on the processes of intelligence, rather on the products of intelligence and the individual differences therein. Specifically, central to Piaget's theory of intelligence (1960) is the view that adaptive behavior and knowledge are constructed through an intimate interplay between the child and the environment. Intellectual development results from the child generalizing her current behavior and knowledge to new situations (assimilation) and also adaptively modifying his or her current behavior

and knowledge through new experiences (accommodation). Influenced by Piaget's conceptions, two main variants of later theories of intellectual development share common endorsements of gene-environment (or biology-culture) coconstructivist and contextualist views.

The Bioecological Theory

With its main focus on understanding the complex interactions between biology and ecology in shaping child development, the bioecological theory (Bronfenbrenner and Ceci, 1994) proposes that aside from genetic effects individual differences in intellectual functioning are the result of variations in the bioecological interactions a child experiences during development. The theory has three main tenets. First, there are multiple innate abilities, as opposed to a single ability. Second, innate abilities serve as a range of possibilities that may develop, or fail to develop depending on their interactions with mechanisms and processes in the proximal developmental contexts (e.g., such as parent–child interactions and schooling). Third, motivation plays important modulatory roles in determining how much the environmental context aids the actualization of the individual's innate potential for intellectual development. The notion that interests and motivational factors play a role in intellectual development has also recently been considered in the PPIK theory (see above).

The Dual-Process Theory of Lifespan Cognitive Development

While also being coconstructive and contextual in nature, the dual-process theory of lifespan cognitive development focuses on understanding the biology–cultural interplay in the gains and losses of intellectual development across the lifespan from early childhood into old age (e.g., Paul B. Baltes et al., 1999). The lifespan developmental theory has been developed by combining both the psychometric and lifespan developmental approaches.

Combining Hebb's Intelligence A and Intelligence B distinction (see above) and the Cattell–Horn *Gf-Gc* framework with perspectives from cognitive and evolutionary psychology, the dual-process theory of lifespan cognitive development highlights two distinct but interacting aspects of intellectual functioning: the biologically driven cognitive mechanics of basic information processing, and the contextually driven cognitive pragmatics of culture-based bodies of acquired factual, procedural, and profession-related skills and knowledge. Furthermore, with an emphasis on the dynamic of lifespan development, the dual-process theory also focuses on how these two aspects of intellectual functioning interact with each other, develop, maintain, and decline throughout the lifespan. In line with Cattell's investment theory, the dual-process theory proposes differential lifespan trajectories for the cognitive mechanics and pragmatics. The biologically driven cognitive mechanics display an early growth pattern which is then invested into the acquisition of pragmatic skills and knowledge. While most abilities of the cognitive mechanics start to decline in midlife, the efficacy of culture-, experience-, and expertise-based pragmatics remain, well into old age; or, as suggested by Ackerman (1996), lifelong occupational and avocational knowledge could still be maintained or even increase up to late adulthood. However, in old age the role of biology-based cognitive mechanics in regulating the culture-based cognitive pragmatics increases. The lifespan developmental orientation to intelligence research provides the insight that the structure and dimensionality of intelligence are themselves subject to change as the individual's brain resources mature during child development, remain stable in adulthood, and decline in old age. The differentiation-dedifferentation hypothesis by Baltes et al. (1980) postulated that intellectual abilities are rather undifferentiated in childhood, undergo differentiation during maturation, leading to a multifaceted intelligence structure that remain mostly stable in adulthood; and become undifferentiated again during aging. In a lifespan sample covering the age range from 7 to 89 years, Li et al. (2004) indeed found that the dimensionality of the manifold of intellectual abilities and their component processes expands from childhood to adulthood and constricts during the course of aging.

Challenges to the Theoretical Status of Ability Factors

More than a century of research on the structure of intelligence and on correlates of cognitive ability factors in terms of their antecedents at the information-processing and biological level as well as consequences at the level of real-world outcomes was based on the investigation of differences between individuals. Yet, the scientific goal of better understanding intelligent behavior targets behaviors like problem-solving, reasoning, and learning at the level of processes within single individuals. In a seminal article, Borsboom et al. (2003) have argued that the factor-analytic approach to the structure of intelligence implies an interpretation of the latent ability factors as real entities that cause the overt behavior of performance on test items (rather than just being descriptive constructions). As the usual approach of measuring people only once on a set of test items provides no basis for demonstrating such causality at this level of individual persons, the huge available empirical evidence from hundreds of individual differences studies is actually mute about the structure and the processes underlying intelligent behavior of individuals. What is therefore required in addition to the well-established research on the structure of between-person differences is research on the structure of within-person fluctuations and changes in cognitive performance (cf Molenaar, 2004).

Borsboom et al.'s challenge of the explanatory value of ability factors as common causes of performance on different cognitive tasks resonates well with recent theoretical proposals that question the need of a general factor *g* to explain the positive manifold among cognitive tasks. van der Maas et al. (2006) proposed that this positive manifold could be explained well by mutualism, that is, beneficial interactions between different cognitive processes during development. Similarly, Dickens (2007) put forward the idea that higher levels on one or some of relatively independent skills could denote advantages for individual people, putting them into more favourable environments, which in turn foster the development of other skills, ultimately leading to a pattern of

positive correlations among performances that gives rise to a g factor. Similarly, Kan et al. (2011) recently argued that there is no need for a separate ability of *Gc* as the cause of performance on knowledge items and that the fact that different such items tend to load on a common factor should rather be interpreted as the result of the acquisition of knowledge taking place in a structured way, which is determined by, for example, the exposure to information in educational settings.

Concluding Remarks

As reviewed above, there is a continuing effort in intelligence research at evolving its subject matter in terms of which mental representation and phenotypic behavioral expressions of intelligence are part of the theoretical conceptions. A few other themes, although deviating somewhat from the more traditional conventions, have been evolving in parallel with the theoretical conceptions reviewed thus far, and merit acknowledgments. For instance, there have been growing emphases on various forms of expertise, such as professional skills or cognitive skills associated, e.g., with chess, medicine, or computer programming, as part of the core of intelligence (see Ericsson and Lehmann, 1996; Newell and Simon, 1972; for reviews). Other examples are the various lines of research on extending the predominately cognition-oriented conceptions of intelligence to social intelligence, emotional intelligence (e.g., Davies et al., 1998; Ford, 1982; Mayer and Geher, 1996; *see* Emotional Intelligence and Competencies), and other presumably separable intelligences, such as musical, spatial, and interpersonal intelligence (i.e., the Multiple Intelligence Theory, Gardner, 1993).

As it is evident from the debates and controversies since Galton and Binet, it is not fruitful for intelligence research to continue with questions such as: Is intelligence unitary or pluralistic? Is the goal of intelligence research to measure the end product of intellectual functioning and to predict performance, or is it to explain the processes of intelligence? Which plays a greater role in the development of intelligence, nature or nurture? Looking at the terrain of current intelligence research and its related fields, it is probably more fruitful to ask questions that highlight the differential utility of the various conceptions of intelligence in furthering our understanding of the complex dynamic interplay between neurobiology and sociocultural contexts in shaping intellectual development. For instance, given existing data at the different levels, which facets of intelligence maybe more biologically driven and which are more context- and culture-based? Which aspects of intelligence are primarily implemented through elementary information-processing mechanisms and their genetic and neurobiological correlates, which other aspects are open to influences from metacognitive, personality, and motivational factors? Which aspects of intelligence are of more direct relevance to real-world problem-solving? What are some aspects of intelligence that might be culture-invariant, and which other aspects are culture-dependent?

Integrating most of the central conceptions of intelligence developed over the last century reveals that intelligence as a phenomenon has multiple facets and levels, and is dynamic both in terms of information processing and in terms of the varying contexts in which it is expressed during individual ontogeny. Couched within an integrated framework, intelligence may be viewed as biologically implemented information-processing capabilities together with knowledge and expertise giving rise to adaptive behaviors in different domains of functioning that are coconstructed through the interplay between sociocultural contextual influences, and the genetic and neurobiological mechanisms across lifespan development. A growing Zeitgeist is more and more in the orientation of biocultural coconstructivism.

See also: Emotional Intelligence and Competencies; Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis; Factor Analysis and Latent Variable Models in Personality Psychology; Intelligence: Historical and Conceptual Perspectives; Piaget's Theory of Cognitive Development.

Bibliography

Ackerman, P.L., 1996. A theory of adult intellectual development: processes, personality, interests, and knowledge. Intelligence 22, 227–257.
Ackerman, P.L., Heggestad, E.D., 1997. Intelligence, personality, and interests: evidence for overlapping traits. Psychological Bulletin 121, 219–245.
Baddeley, A., 1992. Working memory. Science 255, 556–559.
Baltes, P.B., 1987. Theoretical propositions of life-span developmental psychology: on the dynamics between growth and decline. Developmental Psychology 23, 611–626.
Baltes, P.B., Cornelius, S.W., Spiro, A., Nesselroade, J.R., Willis, S.L., 1980. Integration versus differentiation of fluid/crystallized intelligence in old age. Developmental Psychology 6, 625–635.
Baltes, P.B., Staudinger, U., Lindenberger, U., 1999. Lifespan psychology: theory and application to intellectual functioning. Annual Review of Psychology 50, 471–507.
Borsboom, D., Mellenbergh, G.J., van Heerden, J., 2003. The theoretical status of latent variables. Psychological Review 110, 203–219.
Bronfenbrenner, U., Ceci, S.J., 1994. Nature-nurture reconceptualized in developmental perspective: a bioecological model. Psychological Review 101, 568–586.
Carroll, J.B., 1993. Human Cognitive Abilities: A Survey of Factor-Analytic Studies. Cambridge University Press, New York.
Cattell, R.B., 1971. Abilities: Their Structure, Growth, and Action. Houghton Mifflin, Boston.
Chapman, M., 1988. Constructive Evolution: Origins and Development of Piaget's Thought. Cambridge University Press, New York.
Davies, M., Stankov, L., Roberts, R.D., 1998. Emotional intelligence: in search of an elusive construct. Journal of Personality and Social Psychology 75, 989–1015.
Deary, I.J., 2000. Looking Down on Human Intelligence: From Psychometrics to the Brain. Oxford University Press, New York.
Dickens, W.T., 2007. What Is G? Brookings Institution, Washington, DC.
Dixon, R.A., Baltes, P.B., 1986. Toward life-span research on the functions and pragmatics of intelligence. In: Sternberg, R.J., Wanger, R.W. (Eds.), Practical Intelligence: Nature and Origins of Competence in the Everyday World. Cambridge University Press, New York, pp. 281–304.
Engle, R.W., Tuholski, S.W., Laughlin, J.E., Conway, A.R.A., 1999. Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. Journal of Experimental Psychology: General 128, 309–331.
Ericsson, K.A., Lehmann, A.C., 1996. Expert and exceptional performance: evidence of maximal adaptation to task constraints. Annual Review of Psychology 47, 273–305.
Ericsson, K.A., Smith, J. (Eds.), 1991. Toward a General Theory of Expertise: Prospects and Limits. Cambridge University Press, New York.
Eysenck, H.J., 1982. A Model for Intelligence. Springer, Berlin.
Ford, M.E., 1982. Social cognition and social competence in adolescence. Developmental Psychology 18, 323–340.
Gardner, H., 1993. Multiple Intelligences: The Theory in Practice. Basic Books, New York.
Gould, S.J., 1981. The Mismeasure of Man. Norton, New York.
Guilford, J.P., 1967. The Nature of Human Intelligence. McGraw-Hill, London.

Guilford, J.P., 1982. Cognitive psychology's ambiguities: some suggested remedies. Psychological Review 89, 48–59.
Hebb, D.O., 1949. The Organization of Behavior. Wiley, New York.
Herrnstein, R.J., Murray, C., 1994. The Bell Curve: Intelligence and Class Structure in American Life. Free Press, New York.
Horn, J.L., 1982. The theory of fluid and crystallized intelligence in relation to concepts of cognitive psychology and aging in adulthood. In: Craik, F.I.M., Trehub, S. (Eds.), Aging and Cognitive Processes. Plenum Press, New York, pp. 237–278.
Hunt, E., 1980. Intelligence as an information-processing concept. British Journal of Psychology 71, 449–474.
Jäger, A.O., 1984. Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. Psychologische Rundschau 35, 21–35.
Jensen, A.R., 1998. The G Factor: The Science of Mental Ability. Praeger, Westport, CT.
Kan, K.-J., Kievit, R.A., Dolan, C., van der Maas, H., 2011. On the interpretation of the CHC factor Gc. Intelligence 39, 292–302.
Kane, M.J., Hambrick, D.Z., Conway, A.R.A., 2005. Working memory capacity and fluid intelligence are strongly related constructs: comment on Ackerman, Beier, and Boyle (2005). Psychological Bulletin 131, 66–71.
Kyllonen, P.C., Christal, R.E., 1990. Reasoning ability is (little more than) working-memory capacity?! Intelligence 14, 389–433.
Li, S.-C., Lindenberger, U., Hommel, B., Aschersleben, B., Prinz, W., Baltes, P.B., 2004. Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. Psychological Science 15, 155–163.
van der Maas, H.L., Dolan, C.V., Grasman, R.P., Wicherts, J.M., Huizenga, H.M., Raijmakers, M.E., 2006. A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. Psychological Review 113, 842–861.
Mayer, J.D., Geher, G., 1996. Emotional intelligence and the identification of emotion. Intelligence 22, 89–113.
Molenaar, P.C.M., 2004. A manifesto on psychology as idiographic science: bringing the person back into scientific psychology, this time forever. Measurement 2, 201–218.
Neisser, U., Boudoo, G., Bouchard, T.J., Boykin, A.W., Brody, N., Ceci, S.J., Halpern, D.F., Loehlin, J.C., Perloff, R., Sternberg, R.J., Urbina, S., 1996. Intelligence: knowns and unknowns. American Psychologist 51, 77–101.
Newell, A., Simon, H.A., 1972. Human Problem Solving. Prentice-Hall, Englewood Cliffs, NJ.
Nisbett, R.E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D.F., Turkheimer, E., 2012. Intelligence: new findings and theoretical developments. American Psychologist 67, 130–159.
Piaget, J., 1960. The Psychology of Intelligence. Littlefield, Adams & Co., Paterson, NJ.
Resnick, L.B., 1976. The Nature of Intelligence. Lawrence Erlbaum Associates, Hillsdale, NJ.
Spearman, C.E., 1904. General intelligence, objectively determined and measured. American Journal of Psychology 15, 201–293.
Spearman, C.E., 1927. The Abilities of Man: Their Nature and Measurement. Macmillan, New York.
Sternberg, R.J., 1985. Beyond IQ: A Triarchic Theory of Human Intelligence. Cambridge University Press, New York.
Sternberg, R.J., 1999. The theory of successful intelligence. Review of General Psychology 3, 292–316.
Sternberg, R.J., Detterman, D.K. (Eds.), 1986. What Is Intelligence? Contemporary Viewpoints on its Nature and Definition. Ablex, Norwood, NJ.
Sternberg, R.J. (Ed.), 1994. Encyclopedia of Human Intelligence. Macmillan, New York.
von Stumm, S., Ackerman, P.L., 2013. Investment and intellect: a review and metaanalysis. Psychological Bulletin 139 (4), 841–869. http://dx.doi.org/10.1037/a0030746. Advance online publication.
Süß, H.-M., Oberauer, K., Wittman, W.W., Wilheim, O., Schulze, R., 2002. Working memory capacity explains reasoning ability and a little bit more. Intelligence 30, 261–288.
Thorndike, E.L., 1921. Intelligence and its measurement: a symposium. Journal of Educational Psychology 12, 123–147.
Thurstone, L.L., 1938. Primary Mental Abilities. University of Chicago Press, Chicago.
Vernon, P.A. (Ed.), 1987. Speed of Information-Processing and Intelligence. Ablex, Norwood, NJ.

Intelligence, Genetics of: Cognitive Abilities

Frank M Spinath and Juliana Gottschling, Saarland University, Saarbruecken, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Plomin, volume 11, pp. 7645–7651, © 2001, Elsevier Ltd.

Abstract

Behavioral genetic research has made significant contributions to the understanding of individual differences in cognitive abilities. Hardly any other behavioral dimension or disorder has been studied so extensively in behavioral genetic research as general cognitive ability. Based on the results of this research (mainly family, twin, and adoption studies), the importance of both genetic and environmental influences on human intelligence can be considered very well established. This research has also shown that genes become more important with age, drive stability while not being invariant across development, and drive the correlation among various aspects of cognition. More recent studies have further identified moderators on the heritability of general cognitive ability such as income and socioeconomic status, indicating a genotype–environment interaction (which refers to the genetic control of sensitivity or susceptibility to differences in the environment or, in other words, the environmental control of gene expression). Progress in finding specific gene variants for intelligence has been much slower, and molecular genetic studies have yet to identify variations in specific genes that are firmly associated with intelligence differences (with the exception of variation in the apolipoprotein E (APOE) gene).

One of the most prominent questions in many fields within psychology concerns the origins of observed differences among individuals. In particular, individual differences in human intelligence are in the focus of psychological as well as the public interest given their societal significance. In this article, we will address this research question within the framework of behavioral genetic research designs. Behavioral genetics is the study of the influence of an individual's genetic makeup on his or her behavior and the interaction of genes and environment insofar as they affect behavior (Plomin et al., 2013). Two methodological approaches can be distinguished within behavioral genetics: Quantitative behavioral genetics and molecular genetics.

Quantitative behavioral genetics deals with the contribution of genetic and environmental factors to the etiology of individual differences in any complex human trait. When a single trait is considered, the observed trait variation of the population is decomposed into sources of genetic and environmental variation using genetically informative designs, i.e., twin, adoption, and family designs. Additionally, genetically informative designs allow the investigation of the causes of covariation between two or more traits, as well as the causes of stability and change over time when considering multiple traits or points of measurement.

Molecular genetic research aims at identifying specific genes and gene loci responsible for interindividual differences in behavioral traits. Molecular genetic studies use candidate gene or genome-wide association techniques. Within the former approach, a gene that is polymorphic (i.e., has multiple alleles) is selected based on prior findings or on the known function of a gene's protein and it is tested whether people with different alleles tend to differ in the trait under study. While candidate gene studies are hypothesis driven, genome-wide association studies (GWAS) are hypothesis free as the association between a phenotype and a large number of genetic variants spread across the human genome is investigated. If significant genome-wide associations are identified, it is tested which of these associations are robust and replicable (Deary, 2012).

This article will provide a concise overview of behavioral genetic research results in the study of individual differences in general intelligence, not only with respect to heritability estimates and the current state of research attempting to identify specific genes, but also with a focus on the importance of genes across the life span and the environment.

General Cognitive Ability

Individuals undoubtedly differ from one another in their cognitive abilities, e.g., their verbal, spatial, and reasoning ability as well as their memory ability and processing speed. It is also well understood that these differences are predictive of a large range of important life outcomes such as school performance, academic and job success, as well as health, illness, and death (Deary, 2012; Gottfredson, 2002). Today, there is a broad consensus that cognitive abilities are hierarchically organized with a general factor of cognitive abilities (often called *g*) at the highest level (Carroll, 1993). The *g*-factor captures the common variance among the first-level (test-specific factors, i.e., the specific cognitive test that is applied) and second-level factors (broad domains of cognitive functioning, e.g., verbal ability, numeric ability, spatial ability, memory), and it has been shown in many studies that *g* accounts for around 40–50% of the variance in different psychometric intelligence tests (Jensen, 1998). In fact, the *g*-factor is one of the most replicated findings in psychology and its denotation stretches back to 1904 when Charles Spearman described the phenomenon of positively correlated cognitive test scores for the first time and interpreted this as a reflection of a single general factor of intelligence. Although there is still some ongoing debate about the exact organization and labeling of the broad domains of cognitive abilities and how they relate to *g*, it is undisputed that it is primarily *g* that captures the predictive validity as well as most of the variance in the trait intelligence (Gottfredson, 2002). Furthermore, individual differences in cognitive abilities are highly stable across the life span. A unique study conducted by Deary and colleagues was based on data of the Scottish Mental Survey and demonstrated a stability of $r_{tt} = 0.66$ between 11 and 79 years of age in a sample of more than 500 participants who

took intelligence tests in 1932 and, again, in the year 2000 (Deary et al., 2004). Indeed, most of the genetic research to date has also focused on a higher substrate of mental abilities (e.g., IQ scores).

Genetic Influence on *g*: Quantitative Behavioral Genetic Research

In quantitative behavioral genetics, nature (genetics) and nurture (environment) are studied simultaneously as two sources of interindividual differences. Differences in human traits (e.g., general cognitive ability) are attributed to relative proportions of genetic differences and differences in environmental factors that contribute to the phenotypic differences under study. Genetic variance comprises *additive* and *nonadditive* influences. The basic idea of additive genetic variance is that many genes of small effect 'add up' across alleles or loci to influence a trait (Bouchard, 2009). On the other hand, nonadditive genetic variance refers to the genetic effects that depend on interaction across alleles (i.e., dominance) or loci (i.e., epistasis). As environmental sources of variance, a distinction is made between environmental variance that specifiable groups of individuals (i.e., siblings reared together) have in common and therefore contributes to their similarity (such as socioeconomic status (SES), parenting style, family climate) and environmental variance that acts to make them different from each other (such as peers, illness, preferential treatment by one parent) (Bouchard, 2009). The former environmental effects are called *shared* or *common environmental* variance; the latter are called *nonshared* or *specific environmental* variance, which also includes measurement error. Finally, *heritability* is defined as the proportion of (measurable) variability in a phenotype that is 'accounted for' (in the usual regression sense) by variation in genotype (Nisbett et al., 2012). In view of this definition of heritability as the ratio of two variances (i.e., the genetic variance over the total phenotypic variance), it becomes clear that the heritability of a trait always depends on the relative variances of both, genotype and environment. As such, heritability is a descriptive (group) statistic that is solely applicable to the population being studied and not to the phenotype of one single person. That is, depending on the relative variability of the genetic makeup and the environment in a specific population, heritability can theoretically take any value for any trait. Based on this premise, it is apparent that the idea of a 'true' heritability for the heritability of intelligence is nonsensical.

It is also important to address a common misunderstanding regarding the interpretation of quantitative behavioral genetic results. The definition of heritability as a group statistic that is not applicable to the individual implies that chances in environmental conditions (e.g., education practice) can lead to a change in the heritability of a trait. Moreover, high heritability scores do not imply that individual intelligence scores are immutable in terms of interventions.

The relative influence of genes and environment on *g* has been studied since the beginning of psychology. In 1865, 1 year before the publication of Gregor Mendel's seminal publication on the laws of heredity, Francis Galton published two papers on the inheritance of high intelligence and other abilities, later expanded into the highly recognized book *Hereditary Genius: An Inquiry into Its Laws and Consequences* (originally published in 1869; Galton, 1869). Galton ascertained that high intelligence runs in families, an observation that lead him to conclude that high abilities are natural in origin and are inherited from one generation to another. This assumption has been confirmed by dozens of classic and recent twin, family, and adoption studies including data from a total of more than 40 000 pairs of siblings, twins, parent–offspring, and adoptive families from various countries. A first systematic review of genetic research on *g* was published in 1963 by Erlenmeyer-Kimling and Jarvik, and showed converging evidence for genetic influence. Since those early years of genetic studies, both our understanding of the underlying biological mechanism as well as the statistical methodologies have grown tremendously, but yet nothing has refuted the original conclusion that cognitive abilities are under strong genetic influence, with estimates of the effect size of genetic factors ranging from 30 to 80% with an average of 50% explained variance (e.g., see review by Nisbett et al., 2012). Moreover, results from multivariate genetic analyses, which can be used to study the causes of the covariation of specific cognitive abilities, indicate a significant overlap of genetic influences on diverse cognitive facets. This has led to the conclusion that it may be the same genetic factors that largely influence different abilities, or, in other words, that the hierarchical model of cognitive abilities is probably largely genetic in origin (Deary, 2012; Plomin et al., 2013).

At this point, it is important to mention several other factors, which need to be considered when evaluating the heritability of general cognitive abilities. One is a phenomenon that is known as *assortative mating*, which refers to nonrandom mating. With regard to intelligence, brighter women tend to mate with brighter men, reflected by a spouse correlation of about 0.40 for *g* (Plomin et al., 2013). Assortative mating results in offspring that differ more from the average than would be expected under random conditions. This way, assortative mating increases the genetic variability in a population (Plomin et al., 2013). Assortative mating also affects estimates of heritability as it increases the resemblance of first-degree relatives. In twin studies, e.g., this may result in an underestimation of the heritability of a trait because it inflates the similarity of dizygotic (DZ, 'fraternal') twins who are first-degree relatives, but not the similarity of monozygotic (MZ, 'identical') twins who share all of their segregating genes. Consequently, if assortative mating is present for a trait, the difference between MZs and DZs is smaller than under random conditions, diminishing heritability estimates, which are based on these differences.

Another issue concerns the fact that based on the majority of studies so far, most of the genetic variance in *g* appears to be additive. However, there is some evidence that nonadditive genetic effects may also be relevant, yet may well have been neglected in the classical twin design. This is because the classical twin design relies on the assumption of random mating and has to assume that either nonadditive genetic factors *or* shared environmental influences are absent since the two cannot be estimated simultaneously (see Neale and Maes, 2004 or Plomin et al., 2013). In the presence of both assortative mating and nonadditive genetic effects, the pattern of MZ and

DZ twin similarities may wrongly be interpreted as additive genetic effects, thus deflating estimates of nonadditive genetic effects. A recent study by Vinkhuyzen and colleagues utilized an *Extended Twin Family Design*. Such more complex behavioral genetic designs allow the simultaneous estimation of nonadditive genetic effects and shared environmental effects while additionally controlling for effects of assortative mating and cultural transmission (see e.g., Keller et al., 2009; Medland and Keller, 2009). This study revealed that the variance of general cognitive ability in adulthood was not only due to additive genetic factors (44%) but also to nonadditive genetic factors (27%). The remaining variance was explained by nonshared environmental factors and variants of assortative mating (11%) (Vinkhuyzen et al., 2012).

Despite those restrictions, the overall picture of behavioral genetic results for *g* appears to be rather consistent: About half of the variance is due to genetic factors, which are predominantly additive in nature. Shared environmental effects explain about 25% of the remaining variance and 25% are due to nonshared environmental effects and measurement error. It should be noted, however, that these average results are largely based on children data (Plomin et al., 2013). Furthermore, heritability estimates have been shown to vary notably across the life span as well as across different environmental conditions (e.g., socioeconomic classes).

Genetic Influence on *g* Increases during Development

It is now well established that the heritability of intelligence is not the same at different ages. Heritability estimates increase across the life span, from 0% at 3 months of age (Wilson, 1983) to about 30% at 2–6 years of age (Spinath et al., 2003) to 50% in early adulthood (e.g., Haworth et al., 2010) and to as much as 70–80% in mid to late adulthood (e.g., Reynolds et al., 2005). However, there is also some evidence for a decreasing heritability in the very old age. A review of twin studies in the elderly estimated the heritability of *g* at about 60% at 82 years of age (Lee et al., 2010), which suggests that environmental factors, such as physical health may become more important toward the end of the life span (Segal and Johnson, 2009).

Just as the genetic influence on intelligence increases across the life span, environmental influences decrease, and, interestingly, shift with regard to how they operate: While shared environmental influences appear to be more important early in life and in childhood, nonshared environmental influences become more important later in life. For example, a recent study by Haworth et al. (2010) combined raw IQ data from six large twin studies from the United States, the United Kingdom, the Netherlands, and Australia. This study provided strong statistical power and confirmed a rise in heritability from childhood (mean age of 9; 41%), to adolescence (mean age of 12; 55%), to young adulthood (mean age of 17; 66%). The nonshared environmental variance remained stable from childhood (26%) to adolescence (27%) and then dropped slightly in young adulthood (18%). The reverse trend was observed for shared environmental variance, which declined rapidly from childhood (33%) to adolescence (18%) and then remained stable at this level until young adulthood (16%).

Two likely influences have been put forward to explain this occurring pattern. First, small genetic effects present in early childhood might multiply over time, leading to larger phenotypic effects (Segal and Johnson, 2009). This explanation seems especially reasonable given the increasing complexity of cognitive tasks with age and the fact that significant structural brain developments occur from early to middle childhood (Haworth et al., 2010). Second, heritability rises and shared environmental influence declines as individuals play an increasingly more active role in selecting and shaping their environments according to their genetic makeup (Haworth et al., 2010; Johnson et al., 2011), which may also reflect active gene–environment correlation (the concept of a genetic influence on the exposure to the environment or, in other words, the fact that certain genotypes are selectively found in certain environments; Plomin et al., 2013). What is also found is that MZ twin similarity remains relatively stable over time, whereas DZ twin similarity rather tends to decrease. This could also be viewed as a stronger response to environmentally triggered change from individuals who share fewer genes (Bergen et al., 2007).

In addition to considering the variation in the heritability of *g* at different ages, longitudinal behavioral genetic studies have investigated the extent to which genetic and environmental influences contribute to stability and change in *g* (see e.g., Neale and Maes, 2004, for an introduction to multivariate behavioral genetic analyses). *Genetic change* in this respect means that genetic effects at one age differ from genetic effects at another age (Plomin et al., 2013). A genetic contribution to change is expected when intraindividual differences in *g* scores from age to age show genetic influence. Various behavioral genetic studies have been conducted to address this issue with the overall conclusion that the vast majority of genetic effects on *g* contribute to continuity from age to age rather than to change (e.g., Bartels et al., 2002). However, particularly in childhood, genetic factors also account for part of the change. In the same vein, a recent meta-analysis of longitudinal twin and adoption studies aimed at explaining the underlying mechanisms of this developmental increase in heritability (Briley and Tucker-Drob, 2013). Based on a total sample of 11 500 twin and sibling pairs (between 6 months and 18 years of age) the results indicated that the increasing genetic influence on cognitive ability in early childhood is predominantly due to innovative genetic influences. This trend, however, appears to change with age, such that by 8 years of age, the increasing heritability can largely be explained by amplified genetic influences (i.e., genetic influences on cognition are transmitted over time; Briley and Tucker-Drob, 2013). Shared environmental influences, on the other hand, appear to contribute solely to continuity from infancy to middle childhood, and nonshared environmental effects largely contribute to change in *g* across the life span (Bartels et al., 2002; Bishop et al., 2003).

Genetic Influence on *g* Appears to Differ across the Range of Socioeconomic Classes

Over the past 15 years, there is increasing interest in gene–environment interactions, i.e., whether heritability is moderated by an environmental variable such as social class, not least due to rapid methodological advances.

One of the first studies that received extensive attention was based on data from the National Longitudinal Study of

Adolescent Health. Rowe and colleagues showed that the variance in a vocabulary test was explained to different degrees by genetics and environmental influences, depending on mothers' educational level: whereas shared environmental influences accounted for most of the variance in families with poorly educated mothers, genes explained a vast amount of the variance in families with well-educated mothers (Rowe et al., 1999). In a sample of 229 pairs of twins (7 years of age) from the National Collaborative Perinatal Project, a study that is representative in terms of SES (see Nisbett et al., 2012), Turkheimer et al. (2003) examined whether the heritability of intelligence was moderated by environmental factors, i.e., the socioeconomic class. They found a large and statistically significant gene–environment interaction. At the lowest levels of SES, shared environment explained almost entirely the variation in intelligence with practically no genetic effects. By contrast, at the highest SES level, almost all of the variance was explained by genes, and almost none was accounted for by shared environmental effects. In a longitudinal study with children assessed with cognitive tests for the first time at 10 months and again at 2 years of age, there was no variation in the heritability at any level of SES initially. At 2 years of age, significant genetic variance emerged, though only at higher levels of SES (around 5% in low-SES families and 50% in high-SES families; Tucker-Drob et al., 2011). The most recent findings for children are based on data from 2-, 3-, 4-, 7-, 9-, 10-, 12-, and 14-year-old twins from the Twins Early Development Study (Hanscombe et al., 2012). The authors of this study also come to the conclusion that in low-SES families the variance of shared environmental influences is more diverse than in high-SES families, with the effect that the relative importance of genetic effects is smaller in low-SES families.

The case of genetic and environmental interactions with SES after childhood is far less clear. Harden et al. (2007), for instance, found support for a comparable trend to the childhood findings, although the SES variance in their sample of 839 pairs of 17-year-old twins was rather low. Grant et al. (2010), on the contrary, found no evidence for a moderating effect of SES in a study of 5955 male twins of 19 and 20 years of age. In an older sample of adults (36–65 years of age) also no evidence for an interaction emerged (van der Sluis et al., 2008). Based on data from 1702 adult twins (24–84 years of age), Bates and colleagues reported not only that higher SES was associated with higher mean intelligence scores but also that the total phenotypic variance in intelligence was higher in the high SES groups. The observed mean difference between the low versus high SES groups was accompanied by an absolute increase in the genetic variance whereas environmental influences were constant across the range of SES (Bates et al., 2013). From these results the authors conclude that, rather than setting lower and upper bounds on intelligence, genes may multiply environmental inputs that support intellectual growth, and thus triggering the ability to profit from environmental opportunities.

In their comprehensive review of the empirical evidence for differential heritability along the SES continuum, Nisbett et al. (2012) concluded "that the heritability of cognitive ability is attenuated among impoverished children and young adults in the United States" (p. 133) but that it is not possible to generalize this result to other countries due to the mixed results and that it is also not proofed yet whether the effect persists into adulthood. It has been suspected that these differing results among studies could be related to a cultural effect (i.e., smaller socioeconomic differences in cognitive abilities in Europe compared to the United States) and that the case of gene–environment interaction in adulthood is not as clear as in childhood due to a lack of shared environmental components overall, which make it generally difficult to detect significant interactions (Bates et al., 2013; Nisbett et al., 2012). It should also be recognized that large sample sizes are required to detect interactions (van der Sluis et al., 2012) and that an underrepresentation of low-SES families in behavioral genetic studies can cause problems for estimating the heritability of cognitive abilities (see Nisbett et al., 2012).

Genetic Influence on *g*: Molecular Genetic Studies

The quest for genes that influence intelligence has gone on for almost two decades now and many believe that the identification of specific genes will have a great impact on our understanding of the biological and neurophysiological pathways between genes and observed behavior. Advances in molecular biology and genetics have included techniques that make it possible to search the entire genome for genetic variants that may be related to psychological traits. Within the human species, we share over 99.9% of our genetic material. Still, given the size of the human genome, this leaves room for over 3 million expected genetic differences between any two unrelated individuals. These differences are called 'polymorphisms' (when they are found in more than 1% of the population). Some of these variations get translated directly into proteins that control the production of observable individual differences, such as differences in eye color, blood type, and disease susceptibility. Most variations, however, occur in sections of the DNA called 'introns,' and do not code directly for proteins, though they may have regulatory effects on other DNA segments that do. The effects of these polymorphisms on an organism's appearance or function are generally unknown, but the polymorphisms themselves are detectable at the DNA level and can be used as markers. Examples of polymorphisms include single-nucleotide polymorphisms (SNPs), in which a single nucleotide base is substituted for another at a single location in DNA. They also include 'variable number of tandem repeats,' or certain sequences of nucleotides that are repeated within a gene, but the number of repeats vary across individuals. And they include insertion/deletion polymorphisms, in which a sequence of nucleotides is present in one variant, but absent in another.

Quantitative genetic studies can provide a starting point for molecular genetic research insofar as the former type of study can indicate under what circumstances genetic influences make important contributions to the variation between individuals. For complex human traits, such as intelligence, it is reasonable to presume that they are influenced by many genes, each of which has only a relatively small effect. These genes are called quantitative trait loci (QTL), indicating that they are part of a multiple-gene system that contributes to the expression of a trait or disorder. There are two major research strategies within molecular genetics: linkage analysis and association

analysis. Traditional linkage analysis is a technique that detects associations between DNA markers and traits with the aim of mapping genes to chromosomes or to specific regions of chromosomes (see Plomin et al., 2013, for a review). Linkage analysis of large pedigrees has been effective for locating genes for single-gene disorders. For quantitative traits that involve several, up to many, genes of small effect size, however, traditional linkage analysis is not powerful enough. Association analysis in general directly investigates the correlation between a particular allele and a phenotypic trait. Association is typically more powerful than linkage analysis because allelic association has the potential to detect QTLs of small effect sizes. Association has therefore become the method of choice for many researchers over the past decade (Plomin et al., 2013). Within this approach, three different strategies can be distinguished: candidate gene studies, GWAS and genome-wide complex trait analysis (GCTA).

Candidate Gene Studies

Molecular genetics has given us considerable knowledge about many genes that produce specific proteins. Sometimes this knowledge can be used to develop ideas about genes that may be good candidates for involvement in the expression of a trait, because we also have some molecular understanding that their protein products are involved in the trait. Studies that explore this are known as 'candidate genetic association studies.' They compare frequencies of select DNA markers for groups such as individuals with a disorder (cases) versus controls or low-scoring versus high-scoring individuals on a quantitative trait (Sham et al., 2000). If association is present, a particular polymorphism or combination of polymorphisms will be seen more often in individuals carrying the trait than would be expected if it were to occur by chance. The weakness of such studies is that an association can only be detected if the genetic marker is itself the functional gene (direct association) or very close to it (indirect association or linkage disequilibrium). The candidate gene approach therefore requires strong hypotheses as to which genes are good candidates, and we often lack this knowledge (Plomin et al., 2013), since individual differences in complex traits very likely involve the influence of many genes of small effect, any single one of which may not be sufficient or necessary to associate reliably with the trait. To this date, candidate gene studies of intelligence differences have by and large been disappointing: Either no significant association was found at all or the association was not replicable (see Deary, 2012).

Genome-Wide Association Studies

GWAS have become a possible and affordable strategy to identify genes only recently (Hirschhorn and Daly, 2005) and some researchers in this area believe that it is just a matter of time and statistical power until the first robust results are found this way. Through the use of probes for large numbers of SNPs that require only little DNA, systematic scans of the entire genome (DNA) and of the entire transcriptome (RNA) are feasible. The first major GWAS was a search for type II diabetes variants published in Nature in February 2007. By 2008, nearly 100 loci for as many as 40 common diseases and traits had been identified and replicated in GWAS, many in genes not previously suspected of having a role in the diseases under study, and some in genomic regions containing no known genes (Pearson and Manolio, 2008). Since then, a large number of GWAS have been carried out, focusing on a broad spectrum of phenotypic traits, including intelligence. The GWAS strategy has resulted in the identification of reliable links between SNPs and a variety of phenotypes (Johnson and O'Donnell, 2009), including recent findings for schizophrenia (Ripke et al., 2013). In some research areas, GWAS are now analyzed by consortia (e.g., the Psychiatric Genomics Consortium; https://pgc.unc.edu/index.php) that combine forces by either working with joint data sets from different research groups or by conducting meta-analyses of genome-wide genetic data. However, to this date, GWAS findings for intelligence have been disappointing, a fact that is often referred to as the case of 'missing heritability.' For example, the largest GWAS of intelligence, which included almost 18 000 children, found no genome-wide significant association (Benyamin et al., 2014). The largest effect sizes accounted for 0.2% of the variance in intelligence scores. Another recent GWAS of a sample of 1500 children reported an association that accounted for 0.5% of the variance in intelligence scores (Desrivières et al., 2014), but this association was not present in the former study of 18 000 children (Plomin and Deary, in press). To this day, even GWAS on well-established heritable traits such as human height or weight hardly find replicable genetic associations that account for more than 1% of the population variance. Furthermore, testing 500 000 to 1 million SNPs for association, which has become common, presents an unprecedented potential for false-positive results, leading to new stringency in acceptable levels of statistical significance and requirements for replication of findings (Hunter and Kraft, 2007).

Genome-Wide Complex Trait Analysis

GCTA is a new quantitative genetic method that uses DNA to estimate overall genetic influences in large samples of unrelated individuals (Yang et al., 2011). Like other quantitative genetic designs such as the twin design, GCTA uses genetic similarity to predict phenotypic similarity. However, instead of comparing groups of individuals differing markedly in their genetic similarity (e.g., 100% for MZ twins vs 50% for DZ twins), GCTA uses genetic similarity for each pair of unrelated individuals based on that pair's overall similarity across a large number of SNPs. Each pair's genetic similarity is then used to predict their phenotypic similarity. Currently, GCTA detects only those genetic effects tagged by common SNPs and is also limited to detecting additive genetic effects (Plomin and Deary, in press). For this reason, GCTA heritability estimates are typically lower than heritability estimates from twin studies, and they represent the upper limit for detection of SNP associations in GWAS, which, like GCTA are limited to detecting additive effects of polymorphisms. Recent GCTA heritability estimates for intelligence are about half the heritability estimates from twin studies (Davies et al., 2011) suggesting that common SNPs exist which – taken together – account for a considerable portion of the variance in intelligence scores. However, it is also very likely that the heritability of intelligence is caused by thousands of DNA variants, many of which

are likely to be extremely small. It remains an empirical issue, whether and – if so – when such polymorphisms are found and replicated. Another open issue involves the scientific and practical value of such findings, which – at present – come at a high financial cost.

See also: Behavioral Genetics: Psychological Perspectives; Familial Studies: Genetic Inferences; Genetic Studies of Behavior: Methodology; Genetics of Personality; Genetics: The New Genetics; Human Genome Diversity Project: History; Human Genome Project: History and Assessment; Intelligence: Assessments of; Intelligence: Central Conceptions and Psychometric Models; Intelligence: Historical and Conceptual Perspectives; Intelligence: History of the Concept; Personality Differences and Development: Genetic and Environmental Contributions.

Bibliography

Bartels, M., Rietveld, M.J.H., Baal, G.C.M.V., Boomsma, D.I., 2002. Genetic and environmental influences on the development of intelligence. Behavior Genetics 32, 237–249.

Bates, T.C., Lewis, G.J., Weiss, A., 2013. Childhood socioeconomic status amplifies genetic effects on adult intelligence. Psychological Science 24, 2111–2116.

Benyamin, B., Pourcain, B., Davis, O.S., Davies, G., Hansell, N.K., Brion, M.-J., Visscher, P.M., 2014. Childhood intelligence is heritable, highly polygenic and associated with FNBP1L. Molecular Psychiatry 19, 253–258.

Bergen, S.E., Gardner, C.O., Kendler, K.S., 2007. Age-related changes in heritability of behavioral phenotypes over adolescence and young adulthood: a meta-analysis. Twin Research and Human Genetics 10, 423–433.

Bishop, E.G., Cherny, S.S., Corley, R., Plomin, R., DeFries, J.C., Hewitt, J.K., 2003. Development genetic analysis of general cognitive ability from 1 to 12 years in a sample of adoptees, biological siblings, and twins. Intelligence 31, 31–49.

Bouchard, T.J., 2009. Genetic influence on human intelligence (Spearman's *g*): how much? Annals of Human Biology 36, 527–544.

Briley, D.A., Tucker-Drob, E.M., 2013. Explaining the increasing heritability of cognitive ability across development: a meta-analysis of longitudinal twin and adoption studies. Psychological Science 24, 1704–1713.

Carroll, J.B., 1993. Human Cognitive Abilities: A Survey of Factor-Analytic Studies. Cambridge University Press, Cambridge, UK.

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S.E., Liewald, D., Deary, I.J., 2011. Genome-wide association studies establish that human intelligence is highly heritable and polygenic. Molecular Psychiatry 16, 996–1005.

Deary, I.J., 2012. Intelligence. Annual Review of Psychology 63, 453–482.

Deary, I.J., Whiteman, M.C., Starr, J.M., Whalley, L.J., Fox, H.C., 2004. The impact of childhood intelligence on later life: following up the Scottish mental surveys of 1932 and 1947. Journal of Personality and Social Psychology 86, 130–147.

Desrivières, S., Lourdusamy, A., Tao, C., Toro, R., Jia, T., Loth, E., Schumann, G., 2014. Single nucleotide polymorphism in the neuroplastin locus associates with cortical thickness and intellectual ability in adolescents. Molecular Psychiatry. http://dx.doi.org/10.1038/mp.2013.197.

Erlenmeyer-Kimling, L., Jarvik, L.F., 1963. Genetics and intelligence: a review. Science 142, 1477–1478.

Galton, F., 1869. Hereditary Genius: An Inquiry into Its Laws and Consequences. Macmillan, London.

Gottfredson, L.S., 2002. Where and why g matters: not a mystery. Human Performance 15, 25–46.

Grant, M.D., Kremen, W.S., Jacobson, K.C., Franz, C., Xian, H., Eisen, S.A., Lyons, M.J., 2010. Does parental education have a moderating effect on the genetic and environmental influences of general cognitive ability in early adulthood? Behavior Genetics 40, 438–446.

Hanscombe, K.B., Trzaskowski, M., Haworth, C.M.A., Davis, O.S.P., Dale, P.S., Plomin, R., 2012. Socioeconomic status (SES) and children's intelligence (IQ): In a UK-representative sample SES moderates the environmental, not genetic, effect on IQ. PLoS ONE 7, e30320.

Harden, K.P., Turkheimer, E., Loehlin, J.C., 2007. Genotype by environment interaction in adolescents' cognitive aptitude. Behavior Genetics 37, 273–283.

Haworth, C.M.A., Wright, M.J., Luciano, M., Martin, N.G., de Geus, E.J.C., van Beijsterveldt, C.E.M., Plomin, R., 2010. The heritability of general cognitive ability increases linearly from childhood to young adulthood. Molecular Psychiatry 15, 1112–1120.

Hirschhorn, J.N., Daly, M.J., 2005. Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics 6, 95–108.

Hunter, D.J., Kraft, P., 2007. Drinking from the fire hose-statistical issues in genomewide association studies. The New England Journal of Medicine 357, 436–439.

Jensen, A.R., 1998. The *g* Factor: The Science of Mental Ability. Praeger, Westport, CT.

Johnson, A.D., O'Donnell, C.J., 2009. An open access database of genome-wide association results. BMC Medical Genetics 10, 6.

Johnson, W., Penke, L., Spinath, F.M., 2011. Heritability in the era of molecular genetics: some thoughts for understanding genetic influences on behavioural traits. European Journal of Personality 25, 254–266.

Keller, M.C., Medland, S.E., Duncan, L.E., Hatemi, P.K., Neale, M.C., Maes, H.H.M., Eaves, L.J., 2009. Modeling extended twin family data I: description of the cascade model. Twin Research and Human Genetics 12, 8–18.

Lee, T., Henry, J.D., Trollor, J.N., Sachdev, P.S., 2010. Genetic influences on cognitive functions in the elderly: a selective review of twin studies. Brain Research Reviews 64, 1–13.

Medland, S.E., Keller, M.C., 2009. Modeling extended twin family data II: power associated with different family structures. Twin Research & Human Genetics 12, 19–25.

Neale, M.C., Maes, H.H.M., 2004. Methodology for Genetic Studies of Twins and Families. Kluwer Academics, Dordrecht, Netherlands.

Nisbett, R.E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D.F., Turkheimer, E., 2012. Intelligence: new findings and theoretical developments. American Psychologist 67, 130–159.

Pearson, T.A., Manolio, T.A., 2008. How to interpret a genome-wide association study. Journal of the American Medical Association 299, 1335–1344.

Plomin, R., Deary, I.J. Genetics and intelligence differences: five special findings. Molecular Psychiatry, in press.

Plomin, R., DeFries, J.C., Knopik, V.S., Neiderhiser, J.M., 2013. Behavioral Genetics, sixth ed. Worth Publishers, New York.

Reynolds, C.A., Finkel, D., McArdle, J.J., Gatz, M., Berg, S., Pedersen, N.L., 2005. Quantitative genetic analysis of latent growth curve models of cognitive abilities in adulthood. Developmental Psychology 41, 3–16.

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kähler, A.K., et al., 2013. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nature Genetics 45, 1150–1159.

Rowe, D.C., Jacobson, K.C., Van den Oord, E.J.C.G., 1999. Genetic and environmental influences on vocabulary IQ: parental education level as moderator. Child Development 70, 1151–1162.

Segal, N.L., Johnson, W., 2009. Twin studies of general mental ability. In: Kim, Y.-K. (Ed.), Handbook of Behavior Genetics. Springer, New York, pp. 81–99.

Sham, P.C., Cherny, S.S., Purcell, S., Hewitt, J.K., 2000. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. American Journal of Human Genetics 66, 1616–1630.

Spinath, F.M., Ronald, A., Harlaar, N., Price, T.S., Plomin, R., 2003. Phenotypic g early in life: on the etiology of general cognitive ability in a large population sample of twin children aged 2–4 years. Intelligence 31, 195–210.

van der Sluis, S., Posthuma, D., Dolan, C.V., 2012. A note on false positives and power in G × E modelling of twin data. Behavior Genetics 42, 170–186.

van der Sluis, S., Willemsen, G., de Geus, E.J.C., Boomsma, D.I., Posthuma, D., 2008. Gene-environment interaction in adults' IQ scores: measures of past and present environment. Behavior Genetics 38, 348–360.

Tucker-Drob, E.M., Rhemtulla, M., Harden, K.P., Turkheimer, E., Fask, D., 2011. Emergence of a gene x socioeconomic status interaction on infant mental ability between 10 months and 2 years. Psychological Science 22, 125–133.

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., Gottesman, I.I., 2003. Socioeconomic status modifies heritability of IQ in young children. Psychological Science 14, 623–628.

Vinkhuyzen, A.A.E., van der Sluis, S., Maes, H.H.M., Posthuma, D., 2012. Reconsidering the heritability of intelligence in adulthood: taking assortative mating and cultural transmission into account. Behavior Genetics 42, 187–198.

Wilson, R.S., 1983. The Louisville twin study: developmental synchronies in behavior. Child Development 54, 298–316.

Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M., 2011. GCTA: a tool for genome-wide complex trait analysis. American Journal of Human Genetics 88, 76–82.

Relevant Websites

https://www.coursera.org/course/behavioralgenetics.
https://www.genome.gov/.

Intelligence: Historical and Conceptual Perspectives

Robert J Sternberg, Office of Academic Affairs, Oklahoma State University, Stillwater, OK, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E. Hunt, volume 11, pp. 7658–7663, © 2001, Elsevier Ltd.

Abstract

This article presents historical and conceptual perspectives on human intelligence. It opens with a brief account of what intelligence is. Essentially, intelligence is viewed as the ability to implement goal-directed adaptive behavior. Then the article discusses major theories of intelligence. First, it describes implicit theories, or what also are called "folk theories" that laypeople hold regarding the nature of intelligence. Then it discusses in turn psychometric, cognitive, biological, and cognitive-contextual theories of intelligence. Next it covers the development of intelligence. And finally it reviews the measurement of intelligence.

Intelligence is the ability to implement goal-directed adaptive behavior. There have been many and diverse theories attempting to explain how intelligence is structured and how it serves adaptive purposes.

Theories of Intelligence

Implicit Theories

How do psychologists know what intelligence is? One way they find out is to ask people. For example, they might ask experts. In a 1921 symposium on the definition of intelligence, an American psychologist, Lewis M. Terman, emphasized the ability to think abstractly. Another American psychologist, Edward L. Thorndike, emphasized learning and the ability to give good responses to questions. In a similar 1986 symposium, however, psychologists generally agreed on the importance of adaptation to the environment as the key to understanding both what intelligence is and what it does. They also emphasized learning skills and understanding one's own cognitive processes.

In a set of studies, Robert Sternberg and his colleagues asked laypeople in the United States what they thought intelligence was. Three factors emerged from their responses: practical problem solving, verbal ability, and social competence. The conceptions that arise depend, however, on who is asked. In a later comparable set of studies carried out in Taiwan, four factors emerged: cognitive skills, getting along with other people, understanding oneself, knowing when to show one is smart, and knowing when not to show one is smart. In other studies conducted in Kenya, cognitive skills also were emphasized less than in the West, whereas obedience, respect, and understanding people were emphasized more.

Most theories of intelligence are explicit rather than implicit, that is, they are elicited not by asking people what they mean by intelligence, but rather, by having people perform tasks alleged to require intelligence and then analyzing the way people perform. There are several different kinds of explicit theories, especially, psychometric, cognitive, biological, and cognitive-contextual.

Psychometric Theories

Psychometric theories have generally sought to understand the structure of intelligence: What form does it take, and what are its parts, if any? Such theories have generally been based on and tested by the use of data obtained from tests of mental abilities. These tests include items such as vocabulary, completing series of numbers, seeing analogies between words, and visualizing what forms would look like if they were rotated in space.

The first of the major psychometric theories was that of the British psychologist Charles E. Spearman. In a 1904 article, Spearman argued that just two kinds of factors underlie all individual differences in test scores. The first and more important kind of factor Spearman called the "general factor," or *g*. It was said to pervade performance on all tasks requiring intelligence. The second kind of factor was specific to each test. But what, exactly, is *g*? In 1927, Spearman proposed it might be something he labeled "mental energy." Many years later, Arthur Jensen suggested it represents speed of neural conduction.

An American psychologist, Louis Thurstone, suggested instead that seven factors, or "primary mental abilities," pervade test performance: verbal comprehension (knowledge of vocabulary and in reading); verbal fluency (writing and producing words); number (solving simple arithmetical computation and reasoning problems); spatial visualization (mentally visualizing and manipulating objects); inductive reasoning (completing a number series or predicting the future based upon past experience); memory (remembering people's names or faces); and perceptual speed (rapidly proofreading to discover typographical errors in a typed text).

Other psychologists, including Raymond Cattell and John Carroll, suggested that abilities are hierarchical. At the top of the hierarchy is *g*, or general ability. But below *g* in the hierarchy are successive levels of gradually narrowing abilities, ending with Spearman's specific abilities.

Cattell suggested that general ability can be divided into two broad kinds of abilities, fluid and crystallized. Fluid abilities are the reasoning and problem-solving abilities measured by tests such as the analogies, classifications, and series completions. Crystallized abilities can be said to derive from fluid abilities and are viewed as their products, which would include

vocabulary, general information, and knowledge about specific fields. John Horn suggested that crystallized ability more or less increases over the life span, whereas fluid ability increases in the earlier years and decreases in the later ones. Philip Vernon and later Thomas Bouchard and his colleague Wendy Johnson argued instead that the distinction at the second level of the hierarchy is not between fluid and crystallized abilities, but rather between abilities that are verbal in nature and those that are more spatial or mechanical in nature.

Carroll proposed a "three-stratum" model of intelligence, which is considered by some to be the most definitive psychometric model of intelligence because it is based upon reanalyses of hundreds of data sets. According to this model, general ability is at the top of a hierarchy of abilities. At the next lower stratum are various broad abilities (including learning and memory processes and the effortless production of many ideas). At the bottom of the hierarchy are many narrow, specific abilities such as spelling ability and reasoning speed.

Cognitive Theories

In a 1957 address to the American Psychological Association, Lee J. Cronbach proposed that psychologists unite the "two disciplines of scientific psychology" – experimental and differential (the study of individual differences). His proposal led to cognitive theories of intelligence, which are derived by experimental means.

Underlying most cognitive approaches to intelligence is the assumption that intelligence comprises a set of mental representations (e.g., propositions, images) of information and a set of mental processes that can operate on the representations. A more intelligent person is assumed to mentally represent information better and also to operate more quickly on these representations than does a less intelligent person.

Earl Hunt, Nancy Frost, and Clifford Lunneborg showed in 1973 that a critical ability underlying verbal intelligence is that of rapidly retrieving lexical information, such as letter names, from memory. A few years later, Robert Sternberg identified key mental processes alleged to underlie many cognitive tasks, especially ones involving inductive reasoning. They included, among others, encoding stimuli, inferring relations between stimuli, and applying what one has learned. Arthur Jensen used a choice-reaction time apparatus to show that speed of making choices – in his view, a proxy for speed of neuronal conduction – is associated with intelligence.

Ian Deary and his colleagues have sought to understand intelligence through the study of inspection time. They have found that more intelligent individuals can discriminate the lengths of the lines with lesser stimulus duration (inspection) times than can less intelligent individuals.

Two leaders in the field of cognitive psychology, Allen Newell and Herbert A. Simon, used computers to model intelligence. The underlying idea is that computers can, in some sense, show a kind of intelligence similar to that shown by humans.

In the late 1950s and early 1960s they worked with a computer expert, Clifford Shaw, to construct a computer model of human problem-solving. Called the General Problem Solver, it could solve a wide range of fairly structured problems, such as logical proofs and mathematical word problems. In 1972, Newell and Simon proposed a general theory of problem solving, much of which was implemented on the computer. It involved heuristics such as means-ends analysis, whereby one seeks at each step of problem solving to reduce as much as possible the distance between the solution and where one is in the problem.

Marcel Just and Patricia Carpenter showed that complex intelligence-test items, such as figural matrix problems involving reasoning with geometric shapes, could be solved by a computer program at a level of accuracy comparable to that of human test-takers.

The models described above are serial-processing models, whereby the computer takes steps in sequence. David E. Rumelhart and Jay L. McClelland, proposed what they call "parallel distributed processing" models of the mind. These models postulated that many types of information processing occur at once, rather than just one at a time.

Biological Theories and Research

Biological approaches to intelligence directly examine the brain and its functioning. Intelligence as measured by IQ tests appears to be localized, in part, in the prefrontal cortex (PFC) and across the neocortex. People with higher IQs show higher levels of functioning in the superior parietal lobes, temporal, and occipital cortexes as well as in subcortical regions of the brain, especially the striatum. Integration of functioning in the parietal and frontal lobes appears to be especially important. Several different biological approaches have been used, most comparing biologically based measures to IQ.

Neural Efficiency

Complex patterns of electrical activity in the brain as prompted by specific stimuli correlate with scores on IQ tests. In particular, speed of conduction of neural impulses may correlate with intelligence as measured by IQ tests. Some investigators have suggested that this research supports a view that intelligence is based, at least in part, on neural efficiency.

Additional support for neural efficiency as a measure of intelligence can be found from studies of how the brain metabolizes glucose during mental activities. Haier and his colleagues have found that higher intelligence correlates with reduced levels of glucose metabolism during problem-solving tasks. Furthermore, Haier and colleagues found that cerebral efficiency increases as a result of learning in a relatively complex task involving visuospatial manipulations (for example, in the computer game Tetris). As a result of practice, individuals with higher IQ demonstrate lower cerebral glucose metabolism overall. But they also show more specifically localized metabolism of glucose. In most areas of their brains, persons with higher IQ show less glucose metabolism, but in selected areas of their brains (thought to be important to the task at hand), they show higher levels of glucose metabolism. Thus, people with higher IQ may have learned how to use their brains more efficiently. These results are not consistent throughout the entire literature.

Studies using electroencephalographic (EEG) methods have also noted a pattern of neural efficiency in intelligent individuals. Using EEG methods, Aljoscha Neubauer and colleagues

noted that greatest neural efficiency was observed in the brain areas associated with the individual's greatest ability.

Today, however, event-related potentials (ERPs) are used more widely than EEGs in the study of biological bases of intelligence. Research has examined the relation between intelligence test scores and P300. Quicker onset of P300 activity following stimulus presentation typically has been associated with higher intelligence test scores. The relation between P300 onset and IQ has not been consistent, however, and depends in part on the intelligence test utilized.

Brain Size and Intelligence

Some investigators have examined the relationship between brain size and intelligence. For humans, the statistical relationship is modest but significant. Obviously, the finding is only correlational: Greater brain size may cause greater intelligence, greater intelligence may cause greater brain size, or both may be dependent on some third factor. Moreover, how efficiently the brain is used is probably more important than its size. For example, on average, men have larger brains than women, but women have better connections, through the corpus callosum, between the two hemispheres. So it is not clear which sex would have, on average, an advantage – probably neither.

The relationship between brain size and intelligence does not hold across species. Rather, there seems to be a relationship between intelligence and brain size relative to the rough general size of the organism (level of encephalization).

Genetic and Heritability Studies of Intelligence

Although numerous attempts have been made to identify genes that are critical to intelligence, no single gene has been conclusively identified and it looks as though there will be no 'gene' for intelligence to be found. Up to the present, investigators have conducted at least six genomewide scans for genes contributing to intelligence and other aspects of cognition. The data from these scans vary, but there are definitely some partial overlaps. In particular, the data suggest genes related to intelligence in regions on chromosomes 2q (in four out of the six investigations), 6p (for five of the six investigations), and 14q (for three of the six investigations). IQ QTL is a research project attempting to identify quantitative trait loci (QTL) responsible for genetic variation in intelligence. The investigators have sought to identify QTLs linked to intelligence. But positive findings have generally failed to replicate or generated weak signals that have not yet been convincingly replicated in independent samples. Deary and his colleagues have found that "there is still almost no replicated evidence concerning the individual genes, which have variants that contribute to intelligence differences." G. Davies and colleagues, in a study involving 3511 unrelated adults and almost 550 000 single nucleotide polymorphisms (SNPs), have found that genetic bases of intelligence are very widely distributed across genes rather than localized. They have estimated that 40% of the variation in crystallized intelligence and 51% of the variation in fluid intelligence is accounted for by linkage disequilibrium across genotyped common SNP markers and unknown causal variants. Little is known of genetic markers for the broader aspects of intelligence discussed later in this article.

Most attempts to investigate genes underlying intelligence have been indirect rather than direct, through studies of heritability. But heritability is itself a troubled concept. Heritability (also referred to as h^2) is the ratio of genetic variation to total variation in an attribute (such as intelligence) *within* a given population. As a result, the coefficient of heritability says nothing with regard to sources of between-population variation. The coefficient of heritability further does not tell us the proportion of a trait that is genetic in absolute terms, but rather, the proportion of variation in a trait that is due to genetic variation *within a specific population*.

Variation in a trait within a given population is referred to as phenotypic variation; genetic variation in a given population is referred to as genotypic variation. Thus, heritability is a ratio of genotypic variation to phenotypic variation. Complementary to heritability is environmentality, which is a ratio of environmental variation to phenotypical variation. Both heritability and environmentality are applicable only to populations, not to individuals. There is no way of estimating heritability for a particular individual, nor is the concept of heritability even meaningful for individuals.

Heritability is typically evaluated on a 0–1 scale, with a value of 0 signifying no heritability at all (i.e., no genetic variation underlying the trait) and a value of 1 indicating complete heritability (i.e., exclusively genetic variation in the trait). Heritability and environmentality add up to 1. Thus, if IQ has a heritability of 0.50 within a certain population, then 50% of the variation in scores on the attribute within that population is due (in theory) to genetic influences. This statement is completely different from the statement that 50% of the attribute is inherited. Similarly, if a trait has a heritability equaling 0.70, it does not mean that the trait is 70% genetic for any individual, but rather that 70% of the variation across individuals is genetic.

Thus heritability is *not* tantamount to genetic influence. A trait could be highly influenced by genes and yet have low heritability (or none at all). This is because heritability depends on the existence of individual differences. If there are no individual differences, there is no meaningful heritability (because there is a 0 in the denominator of the ratio of genetic to total trait variation in a given population). As an example, being born with two eyes is 100% under genetic control (with extremely rare exceptions of malformations not discussed here). Put another way, regardless of the environment into which a person is born, the person will have two eyes. But it is not meaningful to speak of the heritability of people's having two eyes, because there are no individual differences in the trait. Heritability is not one; rather, it is meaningless (because there is a 0 in the denominator of the ratio) and cannot be sensibly calculated.

Heritability has no fixed value for a given attribute such as intelligence. Although we may read about 'the heritability of IQ' (which, according to most theories, is not exactly the same as intelligence), there is no single fixed value of heritability that represents some true, constant value for the heritability of IQ or anything else. Heritability is dependent on numerous factors, but the most important single factor is the range of environments. Because heritability represents a proportion of variation, its value will depend on the amount of variation. As Richard Herrnstein pointed out, if there were no variation at all

in the environments in which people lived, heritability would be 1, because there would be no other source of variation. If there is wide variation in environments, however, heritability is likely to decrease.

In speaking of heritability, we must remember that genes always operate within environment contexts. All genetic effects occur within a reaction range. As a result, environment will likely have differential effects on the same genetic structure. The reaction range is the range of phenotypes (observable effects of genes) that a given genotype (latent structure of genes) for any particular attribute can produce, given the interaction of environment with that genotype. For example, genotype determines a reaction range for the possible heights a particular person can attain, but other factors, such as diseases, childhood nutrition, and the like may affect the adult height that is attained. Furthermore, if different genotypes respond differently to environmental variation, heritability will differ depending on the mean and variance of relevant factors in the environment. Thus, the statistic is not a fixed, constant value. There exist no purely genetic effects on behavior, as would be demonstrated dramatically if a child were raised in a small closet with no stimulation. No genotype would allow an individual's intelligence to flourish in such an environment. Genes thus express themselves through covariation and interaction with the environment.

Because the value of a given heritability statistic is relevant only under existing circumstances, the statistic does not and cannot address the modifiability of a trait. A trait could have a high level of heritability and nevertheless be highly modifiable. The heritability statistic pertains to correlations, whereas modifiability pertains to mean effects. For example, height has a heritability of over 0.90. Yet height also is highly modifiable, as shown by the fact that average heights have risen substantially over the past several generations.

The heritability of intelligence is typically estimated as between 0.4 and 0.8. The value typically depends on the method used to estimate heritability, such as studies of degrees of relatedness (e.g., identical vs fraternal twins) or identical twins reared apart. The studies are hard to interpret, in part because their assumptions are not always met. For example, identical twins reared apart are not randomly assigned to environments so that one cannot cleanly separate genetic from environmental variation.

Matters are complicated by the fact that heritability estimates vary across populations. For example, estimates of the heritability of IQ in twin studies carried out in the former Soviet Union tended to be higher than they were in comparable studies conducted within the United States. This observation made sense in terms of our discussion above: Environmental variation in Russia under the Soviet regime was relatively constrained; most people lived in roughly comparable environments. As a result, heritability estimates were higher. Most of the IQ heritability studies up to today have been carried out in nations within the developed world. Relatively little information exists regarding the heritability of IQ in the developing world, although what evidence there is suggests moderate heritability in these nations as well.

Heritability also varies as a function of socioeconomic status (SES). Eric Turkheimer and his colleagues have found that heritability is very substantially higher in higher SES families than in lower SES families. In particular, at the lowest levels of SES, shared environment accounted for almost all of the variation in IQs whereas at the highest levels shared environment accounted for practically no variation.

In sum, heritability estimates do not explain in any meaningful sense genetic regulation of human behavior. Furthermore, they do not provide accurate estimates of the strength of the genetic regulation. Rather, as geneticist Elena Grigorenko has pointed out, genes act within the context of environments and their effects must be understood within these contexts.

Cognitive-Contextual Theories

Cognitive-contextual theories deal with the way that cognitive processes operate in various environmental contexts. Two of the major theories of this type have been proposed by Howard Gardner and Robert Sternberg.

In 1983 Howard Gardner proposed a theory of what he called "multiple intelligences." In the 2006 version of the theory, the multiple intelligences include, at minimum, linguistic, logical-mathematical, spatial, musical, bodily kinesthetic, naturalist, interpersonal, and intrapersonal intelligences. Gardner has also speculated as to the existence of an existential intelligence.

An alternative theory also taking into account both cognition and context is Robert Sternberg's 'triarchic' theory of successful intelligence. According to the initial version of Sternberg' theory, intelligence has three aspects: analytical, creative, and practical. Analytical intelligence is what is measured by conventional intelligence tests. But such tests fail to measure creative intelligence – abilities in generating new ideas that are useful – and practical intelligence – the application of intelligence to everyday life, or what may be called common sense. In more recent versions of the theory, Sternberg has suggested that wisdom – the application of one's knowledge and skills toward achieving a common good, over the long and short terms, through the infusion of ethical values, is also an important part of successful intelligence.

Peter Salovey and John Mayer have proposed the construct of emotional intelligence. It then was popularized by Daniel Goleman in a book on emotional intelligence. It is the ability to perceive accurately, appraise, and express emotion; the ability to access and/or generate feelings when they facilitate thought; the ability to understand emotion and emotional knowledge; and the ability to regulate emotions to promote emotional and intellectual growth. Several tests are now available to measure emotional intelligence. They generally show modest to moderate correlations with conventional tests of intelligence.

Development of Intelligence

There have been diverse approaches to studying the development of intelligence. Psychometric theorists, for instance, have sought to understand how intelligence develops in terms of changes in the factors of intelligence over time and changes in the amounts of the various abilities that children have. For example, the concept of mental age was popular during the first half of the twentieth century. A given mental age was believed to represent an average child's level of mental functioning for

a given chronological age. Thus, an average 10-year-old would have a mental age of 10, but an above-average 8-year-old or a below-average 12-year-old might also have a mental age of 10 years. The concept of mental age has fallen into disfavor, however, and is now used only rarely. The concept does not work well over a chronological age of roughly 16, and its assumption of perfectly smooth continuous mental development is questionable.

The Theory of Jean Piaget

Jean Piaget, a Swiss psychologist, suggested that the child explores the world, observes regularities, and makes generalizations, much as a scientist does. Two fundamental cognitive processes are apparently at work in somewhat reciprocal fashion. Assimilation involves incorporating new information into an already existing cognitive structure. Accommodation involves forming a new cognitive structure to incorporate new information. Cognitive development, according to Piaget, represents a dynamic equilibrium between these two processes of assimilation and accommodation.

Piaget also postulated that there are four major periods in intellectual development. The first, the sensorimotor period, extends from birth until roughly 2 years of age. During this period, a child learns how to modify reflexes to make them more adaptive, to coordinate actions, to retrieve hidden objects, and, eventually, to begin representing information mentally. During the second, preoperational period, from about 2–7 years of age, a child experiences the growth of language and mental imagery. He or she also learns to focus on single perceptual dimensions, such as color and shape. The third, concrete-operational period, is from about 7–12 years of age. It is the period during which a child develops an important set of skills that are referred to as conservation skills. The child will recognize that substances stay the same in amount, regardless of their form. Finally, children emerge into the fourth, formal-operational, period, which begins at about age 12 and continues throughout life. The formal-operational child develops thinking skills in all logical combinations and learns to think with abstract concepts.

Post-Piagetian Theories

Later so-called "post-Piagetian" theories of intellectual development, which have been heavily influenced by Piaget, have taken several courses. One, that of Patricia Arlin, has been to expand on Piaget's work by suggesting a possible fifth, adult, period of development, such as problem finding. A second course has been to suggest periods of development related to but different from those suggested by Piaget. A third course has been to accept that intellectual development occurs through the periods Piaget proposed, but to hold that the cognitive bases of development differ from those recognized by him. Some of these theories emphasize the importance of memory or knowledge rather than reasoning capacity.

The Theory of Lev Vygotsky

Lev Vygotsky, a Soviet psychologist, suggested that intellectual development may be largely influenced by a child's interactions with others: A child sees others thinking and acting in certain ways and then internalizes and models what is seen. Vygotsky also proposed the notion of a zone of proximal development (ZPD), which is the range of ability between a person's observable level of ability and the person's latent capacity. This latent capacity is not directly observable but may be detected by providing a context in which the latent capacity may be revealed and expressed. The ZPD is sometimes termed the zone of potential development.

Measuring Intelligence

Early Historical Background

The publication in 1859 of Charles Darwin's *The Origin of Species* had a profound effect on many lines of scientific work. Darwin suggested that the capabilities of humans are, in some sense, continuous with those of lower animals. Hence, they can be understood through scientific investigation. One person who was strongly influenced by Darwin's thinking was Sir Francis Galton, Darwin's cousin. For 7 years – between 1884 and 1890 – Galton maintained a laboratory at the South Kensington Museum in London. For a small fee, visitors could have themselves measured on a variety of psychophysical tasks. These tasks included weight discrimination and sensitivity to musical pitch. Galton believed that these tests measured more than just psychophysical abilities. He believed that psychophysical abilities are the basis of intelligence and, hence, that his tasks were measures of intelligence. Galton's intelligence test, therefore, required a person to perform simple tasks, such as deciding which of two weights was heavier or showing how forcefully he could squeeze his hand. The Galtonian tradition was taken to the United States by the psychologist James McKeen Cattell.

The IQ Test

A more influential tradition of mental testing was developed in France by Alfred Binet and his collaborator, Theodore Simon. In 1904 the minister of public instruction in Paris named a commission to study or create tests that would insure that intellectually deficient children received an adequate education. The minister was also concerned that certain children were being placed in classes for the retarded not because they were intellectually deficient but because they had behavior problems, and teachers did not want them in their classrooms. Binet and Simon proposed that tests of intelligence should measure skills such as judgment, comprehension, and reasoning – the same kinds of skills measured on most intelligence tests today. Binet's early test was taken to the United States by a Stanford University psychologist, Lewis Terman, whose version came to be called the Stanford–Binet test. This test has been revised frequently and continues to be in use.

The Stanford–Binet test, and others like it, have traditionally yielded at the very least an overall score referred to as an intelligence quotient, or IQ. In its most recent form, this test, as well as the Wechsler Adult Intelligence Scale and the Wechsler Intelligence Scale for Children, yield an overall IQ and other scores as well.

More recent tests of intelligence have expanded the range of abilities tested. For example, in 1997, J.P. Das and Jack Naglieri first produced a test, the Cognitive Assessment System, based on a theory of intelligence first proposed by a Russian psychologist, Alexander Luria. The test measures planning abilities, attentional abilities, and simultaneous and successive processing abilities.

IQ was originally computed as the ratio of mental age to chronological (physical) age, multiplied by 100. It is thus sometimes referred to as a ratio IQ. For example, if a child of 8 had a mental age of 10 (that is, performed on the test at the level of an average 10-year-old), the child was assigned an IQ of $(10/8) \times 100$, or 125. If the 8-year-old had a mental age of 6, the child's IQ would be $(6/8) \times 100$, or 75. A score of 100, whereby the mental age equals the chronological age, is average.

As discussed above, the concept of mental age has fallen into disrepute, and few tests continue to involve the computation of mental ages. Many tests still yield an IQ, but it is most often computed on the basis of statistical distributions. The scores are assigned on the basis of what percentage of people of a given group would be expected to have a certain IQ. Scores computed in this way are called deviation IQs.

The Distribution of IQ Scores

Intelligence test scores follow an approximately normal distribution, meaning that most people score near the middle of the distribution of scores. Scores drop off fairly rapidly in frequency in either direction from the center of the distribution. For example, on the IQ scale, about two-thirds of all scores fall between IQs of 85 and 115, and about 95% of scores fall between 70 and 130. Put another way, only one out of 20 scores differs from the average IQ (100) by more than 30 points.

Many psychologists now believe that IQ represents only a part of intelligence, and intelligence is only one factor in both developmental disability and intellectual giftedness. Most current definitions of mental retardation stress adaptive skills as well as IQ, and they also emphasize attributes such as creativity, motivation, and achievement in conceptions of giftedness.

See also: Cultural Intelligence and Competencies; Emotional Intelligence and Competencies; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Prior Knowledge, and Learning; Intelligence: Assessments of; Intelligence: Central Conceptions and Psychometric Models; Intelligence: History of the Concept; Personal Intelligence and Competencies; Social Intelligence and Competencies.

Bibliography

Carroll, J.B., 1993. Human Cognitive Abilities: A Survey of Factor-Analytic Studies. Cambridge University Press, New York.

Ceci, S.J., 1996. On Intelligence. Harvard University Press, Cambridge, MA.

Gardner, H., 2006. Intelligence Reframed: Multiple Intelligences for the Twenty-First Century. Basic Books, New York.

Hunt, E., 2010. Human Intelligence. Cambridge University Press, New York.

Mackintosh, N., 2011. IQ and Human Intelligence, second ed. Oxford University Press, New York.

Piaget, J., 1972. The Psychology of Intelligence. Littlefield-Adams, Totowa, NJ.

Sternberg, R.J., 2003. Wisdom, Intelligence and Creativity Synthesized. Cambridge University Press, New York.

Sternberg, R.J., Kaufman, S. (Eds.), 2011. Cambridge Handbook of Intelligence. Cambridge University Press, New York.

Vygotsky, L.S., 1978. Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA.

Intelligence: History of the Concept

John Carson, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

As a traditional division of the human soul, intelligence had long been of importance to Western philosophers and theologians. Nonetheless, before the eighteenth century intelligence, as a description of overall mental capacity, had little provenance on either side of the Atlantic. Intelligence, as a separate area of research, began to attract attention as growing curiosity about the nature of human differences meshed with the turn to scientific methods as privileged forms of explanation. This article explores that history and highlights how the Intelligent Quotient version of intelligence has come to dominate Western understandings of the concept.

Introduction

In its modern, Western guise the concept of intelligence has seemed inextricably linked to the technology for measuring it, the Intelligence Quotient (IQ) test. Developed in its modern form in 1905 by the French psychologist Alfred Binet (1857–1911) and his collaborator Théodore Simon (1872–1961), the intelligence scale and the practices of measuring intelligence it spawned spread quickly to many parts of the world, including England, Spain, Germany, Latin America, and preeminently the United States. This new technology, however, did not resolve fundamental questions about the nature of intelligence, including a precise definition of what it was, whether it was a product of heredity or environment or both, and whether it was one thing or many. To this day, these issues continue to bedevil psychologists who study intelligence and to haunt many of the public debates about education and merit that rely, either explicitly or implicitly, on particular conceptions of intelligence. In general, one can say that in most current uses of the term, 'intelligence' refers to some sort of overall mental capacity, and one that particularly highlights reasoning, problem solving, and abstract thinking. Nonetheless, research across cultures and historical periods has shown that the specific characteristics attached to the notion of overall mental ability can vary greatly, as can the importance assigned to intelligence as an individual or group-level characteristic.

The direct antecedents of the word 'intelligence' lie in the Latin *intelligentia* or *intellegentia* meaning "the action or faculty of understanding," itself derived from the Latin *intellegere* meaning 'to understand.' As one of the traditional divisions of the human soul (along with the emotions and the will), intelligence was a concept that had long been of importance to Western philosophers and theologians, especially in their attempts to differentiate human beings from other species in the animal world. Nonetheless, up to the eighteenth century it was not a word that evoked much interest on either side of the Atlantic, particular in its modern sense as a description of a person's overall mental capacity. Intelligence began to attract attention in the West as growing curiosity about the nature of human differences meshed with the turn to scientific methods as a privileged form of explanation. This article will explore some of that history and suggest how the development of the concept was intimately related to the particular cultural circumstances in which it was formulated and used.

Species, Groups, and Intelligence

Human nature was a topic of wide resonance within many intellectual fields during the eighteenth century. Not only were virtually all the emerging human sciences engaged in an intensive exploration into the fundamental characteristics of human beings, many of the most important developments in political philosophy also relied on notions about human nature as bedrock for their investigations. Nonetheless, few philosophers or other writers of the period gave much attention to intelligence per se. Rather, they tended to focus on other, somewhat related, concepts, such as 'faculties' and 'talents' and 'sensations' and, most especially, 'reason' and 'virtue.' While it is clear that there was widespread agreement that human beings differed in their mental powers, the origin, significance, and even the sheer number of those differences were all issues that elicited either a multitude of opinions, or no particular view whatsoever. In general, single all-encompassing approaches to understanding mental capacity, such as 'intelligence,' tended to be used only in limited and very particular contexts.

Interest in some sort of mental attribute characterizing overall ability to think or reason can be traced in the west at least back to the ancient Greeks and Aristotle's (384–322 BCE) definition of humans as creatures who reason. Nonetheless, intelligence began to become an object of direct scientific investigation only toward the end of the eighteenth century, when Carl Linnaeus (1707–78), the Comte de Buffon (1707–88), Johann Friedrich Blumenbach (1752–1840), and other naturalists began to compare humans systematically with other animals in the course of developing their grand taxonomies of life. Such comparisons often focused on differences in levels of mental ability, and were linked by many to the notion of the great chain of being, a linear scale that stretched from the lowest organisms through humans and up to God. The Dutch anatomist and artist Peter Camper (1722–89), for example, proposed in the 1760s that a scale of intelligence across the animal kingdom could be erected on the basis of a particular physiological feature, the facial angle, which for the first time

allowed intelligence to be converted into a type of measurable characteristic, though one operating primarily at the level of species or groups. Over the next century and a half a number of other physical features were also evaluated as stand-ins for direct determinations of degree of intelligence, with most physical anthropologists according particular significance to cranial capacity and brain weight. Elaborate programs of measuring hundreds of skulls were carried out throughout Europe and the United States, with the results frequently used to provide scientific sanction for common beliefs about the intellectual inferiority of certain groups, most especially Africans and women, on the basis of the greater proximity of their mean skull or brain sizes to those for the nonhuman primates. As more data were collected, however, extreme variations were found both within and between groups, rendering group-level rankings in the eyes of many largely meaningless. At the same time, the efforts of women in feminist movements and blacks in antislavery campaigns produced powerful denunciations of the entire project of creating hierarchies of intelligence based on physical characteristics.

Until the latter part of the nineteenth century, arguments about intelligence largely focused on the intellectual capacity of groups. Intelligence as a personal characteristic distinguishing one individual from another generated little systematic interest, save for medical concern with diagnosing the mentally deficient and legal interest in those of unsound mind. Mental philosophy and the emerging discipline of 'scientific' psychology were preoccupied with the universal characteristics of human reason and, in the case of psychology, with the process of subjecting mental features to new laboratory techniques being imported from physiology. As the evolutionary theories of Charles Darwin (1809–82) and Herbert Spencer (1820–1903) became more broadly accepted, however, interest grew in the overall mental power of individual organisms, because mental capacity was deemed to be a critical characteristic shaping the organism's adaptation to its environment. Greater pressure to understand the nature of individual differences was also developing as the result of a number of institutional and social changes occurring throughout the West. The spread of universal primary education, the increase in urban concentrations of workers, and the expansion of factory methods of production, along with the rapid growth in asylums for the mentally deficient, all provided many more opportunities to engage in systematic comparisons at an individual level, and thus helped to make intellectual differences much more visible and socially meaningful.

Intelligence and Its Tests

During the late nineteenth century, a number of investigators began to explore the meaning of individual differences, including those associated with overall level of intelligence. In England, Francis Galton (1822–1911) carried out an extensive program of quantifying human mental and physical characteristics (anthropometrics) toward the goal of identifying superior members of the human stock. His *Hereditary Genius* (Galton [1869]1972) argued that innate differences in overall superiority existed within human groups, that those differences were mental in origin, and that they were inheritable. Galton's work helped to initiate widespread enthusiasm for anthropometry, one that persisted until the end of the century. It was also a basis for his promulgation of the science of eugenics, which conceived of human progress in biological terms and advocated improvement in the quality of the breeding stock through encouraging marriages and child bearing among the most 'fit' (i.e., most intelligent) and restricting them among the least.

On the other side of the Channel, psychologist Alfred Binet, one of the pioneers in the development of *psychologie individuelle* in France, conducted a wide range of investigations at the turn of the century into the nature of human mental ability. As one of these projects, in collaboration with Théodore Simon Binet developed the initial version of what would become the first successful individual intelligence assessment instrument, the Binet–Simon Intelligence Scale (1905). The scale was comprised of a series of age-related tasks whose purpose was to reveal whether an examinee's overall intelligence was developing at a 'normal' rate. The measuring aspect of the scale was enhanced following its revisions in 1908 and 1911, when the test-giver's discretion in deciding when answers were right or wrong was severely reduced and results from the testing were reported in terms of mental age (M.A.), which facilitated conceiving of intelligence as a physical and quantifiable object.

Binet's death soon after the publication of the 1911 scale robbed the Binet–Simon of its most effective champion in France, and interest in intelligence testing there remained limited. Such was not the case elsewhere in the West, however, most especially in the United States. With the help of American psychologist Henry Herbert Goddard (1866–1957), who introduced the 1908 version of the scale to his fellow U.S. psychologists in 1910, the Binet–Simon was adopted with enthusiasm. Rival versions of the test were soon produced in great numbers and a vogue for testing developed that never completely vanished.

Stanford University psychologist Lewis M. Terman's (1877–1956) 1916 revision of the scale, his Stanford–Binet (Terman, 1916), did much to cement the place of intelligence testing within the American intellectual and cultural landscape. Thoroughly overhauled for an American population, the Stanford–Binet provided clearer age demarcations and popularized a new way of quantifying mental ability first championed by William Stern (1871–1938), the Intelligence Quotient (IQ), a ratio of mental age to chronological age (times 100) designed to remain constant over time. Terman celebrated this feature of IQ, arguing that it provided a measure of an individual's innate mental ability. He also suggested that IQ could be used to help make a variety of social decisions, including an individual's occupational choice, level of appropriate education, and possible need for institutionalization.

While not every mental tester was as certain about the implications of IQ as Terman, during the 1920s use of intelligence tests in a variety of contexts spread widely, facilitated by the creation of a group mental test (Army Alpha) for the World War I war effort that allowed 1.75 million U.S. army recruits to be assessed without needing the time-consuming one-on-one methods of the Stanford–Binet. Terman's individual scale, revised in 1937 and 1960, however, remained the gold standard for measurements of intelligence, though its preeminence began to be challenged starting in the 1940s with the creation

by David Wechsler (1896–1981) of the Wechsler Intelligence Scale for Children (WISC) and the Wechsler Adult Intelligence Scale (WAIS), instruments that generate assessments of verbal and nonverbal intelligence as well as an IQ.

The Nature of Intelligence

Concomitant with the rise of intelligence testing came a series of debates over the characteristics of the object being measured. In 1904 British psychologist Charles Spearman (1863–1945) used early test data to argue for the unitary nature of intelligence, explaining performance on mental tests in terms of general intelligence (*g*) and specific abilities (*s*). While numerous researchers – including Karl Pearson (1857–1936), Goddard, and Terman – accepted his analysis, others were skeptical, insisting instead that intelligence was composed of a number of primary independent abilities. Edward L. Thorndike (1874–1949) in the United States was among the first to articulate this position. During the 1920s, he was soon joined by two statistically sophisticated psychologists, L.L. Thurstone (1887–1955) in the U.S. and Godfrey Thomson (1881–1955) in the U.K. Both used mathematical methods similar to Spearman's (factor analysis) to show, not that there was one central factor, as had Spearman, but that there were a small number of significant and independent abilities. In the period after World War II, psychologists continued to put forward a range of interpretations of the composition of intelligence: while Hans Eysenk (1916–97) remained convinced of the reality of *g* (general intelligence), for example, Philip E. Vernon (1905–87) proposed a hierarchical model of intelligence that inter-linked specific skills, general abilities, and overall intelligence, and Joy P. Guilford (1897–1987) contended that intelligence was composed of as many as 150 independent factors. Later influential additions to these conceptualizations included Howard Gardner's (1943–present) seven discrete types of intelligence and Robert J. Sternberg's (1949–present) triarchic model of intelligence.

Overshadowing all of the arguments over intelligence, however, has been the nature–nurture question, particularly with regard to racial, ethnic, and socioeconomic groups. Figures such as Galton, Spearman, Pearson, and Terman argued early in the twentieth century strenuously for the nature position, with Galton producing studies on identical twins that have served as a model to this day for investigations into the relative weights of heredity and environment. During the 1900s to 1930s, when eugenics was at the apex of its popularity, arguments in favor of intelligence as an inheritable biological entity ran strong, and were used to justify calls for immigration restriction and for sterilization of the mentally 'unfit,' as well as for the creation of multitracked secondary schools. Claims about the 'inferiority' of non-European racial groups had been long-standing in the West, buttressed from the middle of the nineteenth century on with arguments about differences in overall levels of intelligence. During the early twentieth century these arguments continued to be promulgated using data on intelligence produced by the new intelligence tests, and indeed to be extended to less favored segments of the so-called 'white race,' including lower socioeconomic groups and peoples deriving from parts of southern and eastern Europe. There was also a particular worry about those individuals feared to be of subnormal intelligence, with professional scholars and public figures often decrying what they termed 'the menace of the feebleminded.'

At the same time, however, anthropologists were beginning to put renewed emphasis on culture as the primary determinant of human behavior, claims strengthened during the middle of the century by studies carried out especially at the Iowa Child Welfare Research Station, where IQ was found to change depending on nutrition and educational environment. Moreover, during the 1920s and 1930s a number of psychologists and sociologists in the United States, including many prominent African-American scholars, challenged much of the data on which claims about group-level differences in mental ability were based, demonstrating that the World War I testing data, for example, could be more convincingly interpreted as showing that intelligence scores were influenced markedly by the quality of education received.

During the post-World War II period the debate continued to be pressed from both sides, with increasingly sophisticated twin studies showing high IQ correlations between identical twins separated at birth, while at the same time other researchers were teasing out ever more complicated connections between intelligence development and such factors as nutrition, family child-rearing practices, socioeconomic status, and quality of education received. In the 1980s, James R. Flynn among others pointed out the significant increases in IQ test scores around the globe (now known as the 'Flynn effect'), and used those dramatic changes to argue that intelligence tests were capturing only limited aspects of the underlying entity called intelligence and that performance on these tests was strongly affected by a range of environmental factors. There was also a wholesale retreat on claims about intelligence at the level of groups, with most mainstream researchers arguing that group-level differences, even if they existed, were more than swamped by individual variation and environmental factors.

While few experts would deny the influence of both genes and the environment, the vociferous debate in the mid-1990s over *The Bell Curve* (Herrnstein and Murray, 1994), with its claims that IQ is hereditary and a prime determinant of an individual's future, indicates that broad disagreements persist about intelligence and show little likelihood of quick resolution. What they reveal as well is that the language of native intelligence has remained a powerful vehicle for discussions of a range of social issues, from the organization of an educational system to the value of affirmative action programs to the just allocation of social resources. Because of its associations with individual merit and its history of use for racial and class discrimination, intelligence continues to attract both strong proponents and vigorous critics, and remains a concept with wide cultural and political resonance.

See also: Binet, Alfred (1857–1911); Boas, Franz (1858–1942); Civil Rights Movements; Classical (Psychometric) Test Theory; Cognitive Development, Educational Theories of; Cultural Intelligence and Competencies; Darwinism: Social; Dewey, John (1859–1952); Egalitarianism; Elites, Anthropology of; Emotional Intelligence and Competencies; Eugenics, History of; Experimental Laboratories: Psychological; External Validity;

Factor Analysis and Latent Variable Models in Personality Psychology; Galton, Sir Francis (1822–1911); Governmentality; Inequality, Social; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Prior Knowledge, and Learning; Intelligence: Assessments of; Intelligence: Central Conceptions and Psychometric Models; Intelligence: Historical and Conceptual Perspectives; Meritocracy and Tokenism; Nature-Nurture Controversy, History of; Pearson, Karl (1857–1936); Personal Intelligence and Competencies; Piaget's Theory of Cognitive Development; Piaget's Theory of Human Development and Education; Piaget, Jean (1896–1980); Psychology, History of (Early Period); Psychology, History of (Twentieth Century); Psychometrics: Classical Test Theory; Psychometrics; Quetelet, Adolphe (1796–1874); Race and Racism in the Twenty-First Century; Race: Genetic Aspects; Racism, Sociology of; Social Intelligence and Competencies; Social Scientists as Experts and Public Intellectuals; Spencer, Herbert (1820–1903); Thorndike, Edward Lee (1874–1949); Vygotsky's Theory of Cognitive Development.

Bibliography

Binet, A., Simon, T., 1916. The Development of Intelligence in Children. Williams & Wilkins, Baltimore.

Blanckaert, C., 1987. 'Les vicissitudes de l'angle facial' et les débuts de la craniométrie (1765–1875). Revue de synthèse 108, 417–453.

Carroy, J., Plas, R., 1996. The origins of French experimental psychology: experiment and experimentalism. History of the Human Sciences 9, 73–84.

Carson, J., 2007. The Measure of Merit: Talents, Intelligence, and Inequality in the French and American Republics, 1750–1940. Princeton University Press, Princeton.

Cravens, H., 2002. Before Head Start: The Iowa Station and America's Children. University of North Carolina Press, Chapel Hill.

Danziger, K., 1997. Naming the Mind: How Psychology Found Its Language. Sage, London.

Degler, C.N., 1991. In: Search of Human Nature: The Decline and Revival of Darwinism in American Social Thought. Oxford University Press, New York.

Fancher, R.E., 1985. The Intelligence Men: Makers of the IQ Controversy. Norton, New York.

Galton, F., 1972[1869]. Hereditary Genius: An Inquiry into Its Laws and Consequences. Peter Smith, Gloucester.

Gould, S.J., 1981. The Mismeasure of Man. Norton, New York.

Hegarty, P., 2013. Gentleman's Disagreement: Alfred Kinsey, Lewis Terman, and the Sexual Politics of Smart Men. University of Chicago Press, Chicago.

Herrnstein, R.J., Murray, C., 1994. The Bell Curve: Intelligence and Class Structure in American Life. Free Press, New York.

Jackson, J.P., 2005. Science for Segregation: Race, Law, and the Case against Brown v Board of Education. NYU Press, New York.

Kevles, D.J., 1985. In the Name of Eugenics: Genetics and the Uses of Human Heredity. University of California Press, Berkeley.

Lemann, N., 2000. The Big Test: The Secret History of the American Meritocracy. Farrar, Straus & Giroux, New York.

Porter, T., 1995. Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton University Press, Princeton.

Richards, R.J., 1987. Darwin and the Emergence of Evolutionary Theories of Mind and Behavior. University of Chicago Press, Chicago.

Rose, N., 1990. Governing the Soul: The Shaping of the Private Self. Routledge & Kegan Paul, London.

Smith, R., 1997. The Norton History of the Human Sciences. Norton, New York.

Sokal, M.M. (Ed.), 1987. Psychological Testing and American Society. Rutgers University Press, New Brunswick, pp. 1890–1930.

Staum, M.S., 2007. Ribot, Binet, and the emergence from the anthropological shadow. Journal of the History of the Behavioral Sciences 43, 1–18.

Stern, A., 2005. Eugenic Nation: Faults and Frontiers of Better Breeding in Modern America. University of California Press, Berkeley.

Sutherland, G., 1984. Ability, Merit and Measurement. Oxford University Press, Oxford.

Terman, L.M., 1916. The Measurement of Intelligence: An Explanation of and a Complete Guide for the Use of the Stanford Revision and Extension of the Binet–Simon Intelligence Scale. Houghton Mifflin, Boston.

Wolf, T.H., 1973. Alfred Binet. University of Chicago Press, Chicago.

Wooldridge, A., 1994. Measuring the Mind: Education and Psychology in England C.1860-c. 1990. Cambridge University Press, Cambridge.

Zenderland, L., 1998. Measuring Minds: Henry Herbert Goddard and the Origins of American Mental Testing. Cambridge University Press, Cambridge.

Intelligence in Cultural, Social and Educational Context

Andreas Demetriou, University of Nicosia Research Foundation, University of Nicosia, Nicosia, Cyprus

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews recent research on human intelligence and intellectual development and the effects of culture and other social factors on both.

It also capitalizes on recent literature on how intellectual development may be strengthened by education. Contrary to a common depiction of human intelligence as a unitary trait that is invariant across time, space, and cultural context, this article argues, with Sternberg (2004), that "Intelligence cannot be fully or even meaningfully understood outside its cultural context."

Imagine persons of different ages in different cultures trying to make sense of various situations, solve problems, or make necessary decisions relevant to their age:

1. A 3-year-old toddler in New York, Athens, or Beijing trying to figure out how to make his brand new toy car move around in the room, ask for some help from the dad, or pretend that he is driving his dad's car. Alternatively, imagine a toddler in a remote small village in Africa, riding a rod pretending that it is a horse or throwing little stones to hit a small bird.
2. A 10-year-old trying to understand history or biology as presented in a school book, solve algebra problems, figure out why her best friend is angry with her and restore the relation, or convince her mother to let her meet with friends. Alternatively, imagine a child in one of the South American megacities, making a living through selling and buying fruits on the street.
3. A 16-year-old writing an essay on how principles in physics underlie the operation of mobile phones, choose a course of studies that will end up in an occupation, or make political choices based on discussions with friends, teachers, parents, and public debate (on TV or through the Internet) about governments decisions. Alternatively, imagine an adolescent trying to pass the induction tests of his tribe in the Amazon forest to become a member of the adults of the tribe or an adolescent in the Pacific islands trying to learn how to navigate from one island to the other.

However different they appear, all of these situations require from the person to specify what the problem is about and combine presently available information, previous knowledge, and thought to find a pertinent solution. In short, they require intelligence. Moreover, all of these problems are similar in some respects over all ages and cultures and different in some other respects.

To understand the human mind and intelligence, one of the mind's important manifestations, the modern science of the human mind tries to answer the following questions (Carroll, 1993; Demetriou et al., 2010; Jensen, 1998; Piaget, 1970):

1. What are the cognitive processes underlying intelligence? That is, how is information represented, connected and processed, understood, transformed, and used for the sake of learning, problem solving, and decision making?
2. What is general and what is specific in intelligence? That is, what is common across different domains of knowledge, such as mathematics, literature, or law, or different types of activity and skill, such as music, drawing, or dancing and what is particular to each of these domains?
3. What is stable and what is changing with development in intelligence?
4. Why do individuals differ in intelligence?
5. What is the contribution of the environment and culture to these differences?
6. Can intelligence be changed by education?

What Is Involved in Intelligence?

The human mind emerges from the functioning of the human brain. The human brain is an extremely complex system that evolved to enable us to know the world and cope with its demands in the sake of survival and continuation. We are connected with the world through our senses. Vision provides access to the visible aspects of the world. The space we move in, the objects we see, their color, shape, size, and number, their positions relative to each other, and their interactions, are all known through our eyes and the parts of the brain that are responsible to represent and process the information collected by the eyes. Hearing provides access to sound. All sorts of sounds produced by the natural environment and other living beings, including humans, can become known through the ears and the parts of the brain that are responsible to record and process sound. The other senses, such as smell, taste, and touch, provide other kinds of knowledge about the world. Information coming from each channel has properties that are specific to it. Through time, the human brain evolved structures and processes specializing in the processing of the various types of information provided by the senses and in the integration of different kinds of information into holistic and systematic representations enabling accurate understanding, appropriate decision making, and efficient problem solving (Jung and Haier, 2007).

The description above applies to all other organisms with a brain, to degrees varying with the complexity of their brain. It is unique to humans, however, that, they may draw knowledge and solutions to problems from stored pools of knowledge and

ideas. These are cultural products that were made initially possible because humans evolved to have a communication system that enables them to share knowledge and ideas. That is, first they acquire language, at least 100 000 years ago (Sampson et al., 2009) and then arbitrary symbol systems that stand for language, primarily writing, about 3200 years ago (Daniels and Bright, 1996), that render present knowledge and ideas virtually eternal. As a result, knowledge and problem solving becomes collective and, in principle, always available and always revisable. Obviously, the stored knowledge of different cultures reflects the history of each of them. Education is the institution that was developed through the millennia as a means for the induction of each individual into his or her own culture (Mithen, 1996).

The human mind has been the focus of three research traditions in psychology, since the late nineteenth century: the cognitive, which focused on cognitive functions and processes (Sternberg, 2011), the differential, which focused on individual differences in intellectual ability (Mackintosh, 1998), and the developmental tradition, which focused on cognitive development (Case, 1985). Because of priorities in each, research questions, methods, and theoretical constructs differed between the three traditions, especially at their early years. With time, however, the three traditions did interact with and informed each other in theory and methods. This interaction led to considerable convergences in how the human mind is analyzed and studied. According to all traditions, the human mind (1) learns actively from interactions with the world, (2) represents its interactions in mental models which are used (3) to guide further encounters, especially when they are new and unexpected. The success to deal with unexpected encounters is a measure, in all traditions, of the overall functional efficiency of the mind, gauged vis-à-vis the current situation (in the cognitive tradition), other individuals (the differential tradition), or the age of the individual (the developmental tradition). Also, all traditions are reductionist, in that they attempt to reduce models that represent the world and guide interactions with it into simpler processes aiming to capture the capacity of the mind to learn, model, choose between, and use representational systems according to current or predicted and planned goals and needs (Demetriou et al., 2010, 2011; Kazi et al., 2012).

Research and theory in these traditions converge on the assumption that the human mind is a very complex universe of processes that cooperate for the sake of understanding, problem solving, and learning. These processes are organized into systems that carry out different tasks during any understanding or problem solving task. Specifically, a set of processes interface the mind with different aspects of the environment. A second set are responsible to meet the representational and processing needs activated by the environment-oriented processes. A third set is responsible to connect and integrate information and operations vis-à-vis the goal of the moment. Finally, a fourth set of processes are responsible for the monitoring, control, and regulation of the processes activated at a given moment, if needed. That is, the four types of processes are as follows: *specialized structural systems or domains of mind*, *representational capacity*, *inference*, and *consciousness*. **Figure 1** highlights this architecture (see Demetriou et al., 2010, 2011; Hunt, 2011). These processes are involved in everyday regular functioning but also in creative mental functioning of gifted persons (Kaufman and Baer, 2006).

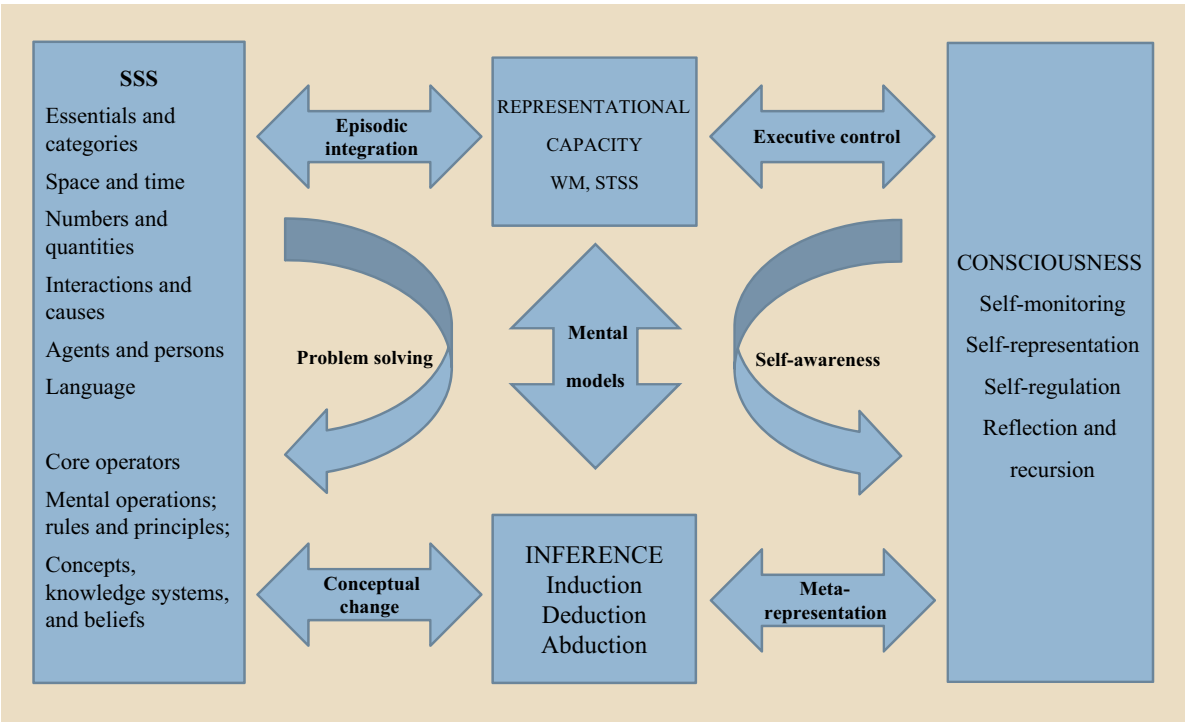

Figure 1 The general architecture of the human mind.

Domains of Intelligence

Discussion about the domains of intelligence goes on for decades. In the classical psychology of intelligence, there is general agreement about three domains of thought: language, space, and number. Command and versatility in using language, such as the comprehension and use of words, ability to mentally orient in space and grasp the relations of objects in space, such as when one must put together the pieces of a puzzle to form an integrated picture, and the ability to deal with numbers and quantitative relations, are examples of fundamental abilities in each of these three domains. Causality is recently recognized as another domain (Carey, 2009). That is, the ability to grasp how and when objects or persons cause changes in each other. In the recent years the understanding and dealing with one's own and the other persons' emotions, feelings, and thoughts is regarded as a separate domain of psychological/social intelligence. The current discussions about emotional intelligence or interpersonal and intrapersonal intelligence in Gardner's (1983) theory of multiple intelligences stand for this domain. Gardner also added kinesthetic and musical intelligence to these domains. The domain of biology may be a separate domain of thought that enables understanding of the phenomena of life and the principles underlying the behavior and condition of living beings (Carey, 1985).

Each of the domains involves three types of processes, summarized in **Table 1**. Specifically, the domains involve (1) core processes, (2) mental operations, and (3) knowledge and beliefs. Core processes are innate and geared in perception, as noted earlier. Examples are categorical perception on the basis of color, the perception of depth in the spatial system, or subitization in the quantitative system – that is, the ability of humans and other animals to automatically perceive numerosity for sets of up to three to four elements. They are present from the very first few of weeks of life if not at birth (Carey, 2009), they are the result of our evolution as a species, and they form the background for the development of the other processes in domains. With development, core processes yield mental operations that are increasingly self-guided. For example, categorization according to color in the categorical system, arithmetic operations in the quantitative system, isolation of variables in the causal system, mental rotation in the spatial system, perspective taking in the social system. Finally, the functioning of core processes and mental operations generates and updates conceptual and belief systems pertaining to the physical, the biological, psychological, and the social world: for example, money values and rules underlying everyday transactions in the quantitative domain; mental images, mental maps, locations, scenes, or layouts stored in memory in the spatial domain; representations about physical events in the causal domain or social events in the causal domain (Carey, 1985, 2009).

Domain specificity is a long-debated issue in psychology. In the domains, domain specificity varies according to level of the processes concerned. Core processes are highly domain-specific because they are constrained by perceptual

Table 1 Processes and levels in the organization of cognitive domains

Domain	*Core processes*	*Mental operations abstraction/induction*	*Knowledge and beliefs*
Categorical	Perception of similarity.	(Bayesian) inductive inferences based on similarity–difference relations. Semantic and logical relations in class hierarchies.	Conceptions and misconceptions about the world (objects, animate beings, persons), units of matter (e.g., substance, molecules, atoms, particles, etc.).
Quantitative	Subitization; pointing, counting, object arrangement (arithmetic operations).	Number alignment, induction of the mental line, proportionality.	Factual knowledge about the quantitative aspects of the world, street and supermarket mathematics, quantitative dimensions (e.g., weight, volume, area, etc.), algebraic and statistical inference rules.
Spatial/time	Location, object, size, depth, orientation perception; formation of mental images.	Mental rotation, image integration and reconstruction, location and direction tracking and reckoning.	Stored mental images, mental maps, and episodic representations and scripts about objects, locations, scenes, or occurrences.
Causal	Perception of interactions or associations in spacetime.	(Bayesian) induction of causal relations (e.g., necessary and sufficient, necessary and non-sufficient, etc.); causal search (trial and error); hypothesis formation; experimentation (isolation of variables).	Knowledge, attributions and understanding of the reasons underlying physical, biological, and social events.
Social	Recognition of conspecifics, species-valued stimuli, imitation.	Theory of mind (induction of intentions, mental states, and beliefs); decentering and taking the other's perspective.	System of social attributions about other persons, traditions, customs, and ideologies in groups, societies, and cultures.
Language and symbolic systems	Grasp of grammatical and syntactical structures of language.	Truth/falsity identification; induction of relations between "truths"; differentiation of the contextual from the formal elements; elimination of biases from inferential process; securing validity of inference.	Knowledge about grammar, syntax, and logical reasoning; metalogic about inference; awareness about knowledge production and validation processes.

channels and they are activated by specific patterns of information in the environment. Mental operations are by and large domain-specific but they are subject to training and coordination between domains, as when for instance we use numbers from the quantitative system to represent spatial relations in the spatial domain or causal relations in the causal domain. Knowledge and beliefs systems are built on the common ground of the general processes and mental operations described here but they have a large degree of specificity. For example, the disciplines of physics, biology, and psychology focus on the natural, the biological, and the human world, respectively. They all require all of the systems of mind described here to be grasped, probably to varying degrees. At the same time, each is expressed in discipline-specific semantic and rule networks that need to be mastered as such.

Abstraction and Inference

The brain is a system that encodes at one and the same time the past of an organism and its possibilities for the future. Intelligence is the functional expression of the ability to use the past in the sake of the future by enabling the individual to capitalize on his past to make decisions and find solutions for new problems. Abstraction is the fundamental process for the unification of learning and mental experiences. The ability to recognize common patterns in the flow of information, connect them into new concepts, and use them for the acquisition of new knowledge is present from birth, if not earlier in the womb (Tenenbaum et al., 2011). Relations are recognized on the basis of *similarity* between objects or concepts (e.g., color pulls together very diverse objects just because they have the same color). *Inference* carries properties from one situation or concept to another on the basis of similarity and probabilistic trends in the pattern of events. That is, missing information in a target situation is *induced* or *inferred* because the two situations are similar in some respects so that characteristics of the known situation or concept are ascribed to the base situation on the basis of their other similarities (it moves on its own so it is a living being – even without legs).

Inductive, analogical, and deductive reasoning are the main types of inference. Inductive reasoning may involve any kind of representation, such as perceptions, mental images, or propositions, and it generates relations that are only likely. Analogical reasoning is inductive reasoning applied on relations as such (e.g., London is to the United Kingdom what Paris is to France – it is the capital of). Deductive reasoning is applied only on propositions or other symbols which stand for propositions. Thus, inference in deductive reasoning transfers meaning from one set of propositions (the premises) to other propositions (the conclusion) and it always proceeds from the general (the premises) to the specific (the conclusion). In deductive reasoning if the premises are valid their conclusion is also valid and it follows necessarily from the premises (e.g., Birds fly, Guzy is a bird, it flies – even if Guzy is an Ostrich and it therefore does not fly).

Tests of intelligence always include problems addressed to these kinds of reasoning in various forms. For example, the Raven progressive matrices test (Raven and Raven, 2008), a famous test of inductive reasoning, examines the ability to find relations between visual figures and use them to specify what is missing from a pattern. Analogical and deductive reasoning is examined in various contexts, such as the specification of analogical relations between number series (e.g., 1, 5, 3, 2, 6, 4, 3, 7, ?) or concepts as in the example above about countries and their capital.

Representational Capacity

To induce relations, interrelate concepts and premises and draw conclusions, and plan solutions for problems, the relevant information and concepts must be kept in mind for as long as needed. Working memory, short-term storage, and episodic memory are terms used to refer this capacity. A common measure of working memory is the maximum amount of information and mental acts that the mind can efficiently activate simultaneously. Baddeley's (1990, 2012) model is widely regarded as a good approximation to the architecture of working memory. According to this model, working memory consists of four systems. Specifically, there are two modality-specific storage systems, namely the phonological loop and the visuospatial sketchpad, which store acoustic and visual information, respectively. There is also a central executive. This is an attention-controlling system that sustains and coordinates processing in the specialized storage systems by processes such as selection of the to be remembered information and its updating so that it is lost. These processes are effortful and conscious and they compete for the available representational resources with the specialized storage buffers. Finally, there is an episodic buffer capable of integrating information across systems into unitary episodic representations. The episodic buffer ensures the production of integrated mental products and it preserves time and the subjectivity of experience. The capacity of the storage systems and the capacity needs of the central executive set the general limits for the complexity of information and processing that can be represented. To read or follow a conversation, a minimum number of words and sentences must be represented and related to past knowledge. It is common in tests of intelligence to ask individuals to recall number digits or words in their order of presentation or in reverse order.

Attention and Executive Control

Concentrating on what one is up to is not always easy. In normal conditions, everyone is surrounded by redundant information interfering with what we are up to. Focusing on a conversation in a public place is frequently hindered by all sorts of sounds in the environment or even thoughts uncontrollably popping into one's own mind. Planning to solve a problem may go astray because irrelevant ideas spoil assembling the sequence of actions required. Therefore, to serve their purpose, representational capacity, abstraction, and inference must be protected from interference. This is the function of attention and executive control (Posner and Rothbart, 2006).

The golden trio of this function is 'focus, inhibit, shift' (Zelazo, 2011; Zelazo et al., 2005). That is, when working on a problem, keep the goal in focus, inhibit turning to irrelevant

information even if attractive or interesting or initiating irrelevant action, and shift to other information or action, if the needs of the moment justify doing so. In current research on intelligence and cognition, attention and executive control are considered as manifestations of consciousness. That is, the ability to be aware of one's own self at any moment, of the flow of experience, and the contents of representational capacity, be it information from the environment, from past knowledge, or actions made or planned. Self-regulation is to a large extent possible because of this awareness and the processes that enable one to change its focus and content.

In tests of intelligence this ability is examined by tasks requiring the manipulation of information in working memory, such as when the examinee has to recall strings of letters and number digits in the right sequence and order. Also it is examined by tasks addressed to speed of visual processing which require to focus on one aspect of the stimulus and ignore another.

What Is General and What Is Specific in Intelligence?

Performance on any task at any moment is a composite of the processes specified above. That is, it involves knowledge, skills, and mental operations specific to the domain the task belongs to, it requires some kind of inference that would integrate knowledge, skills, and operations with each other and with the current goal, it is constrained by the representational capacity available, and it needs to be protected by irrelevant information or ideas for as long as needed. Nowadays, a dominant model of how to decompose what is involved in each intellectual task is the three-stratum model proposed by Carroll (1993). According to Carroll, the first stratum includes very narrow task- or medium-specific abilities, such as word knowledge, problem solving in mathematics, memory for words, specific information about the terms used in algebra. The second stratum contains broad abilities that organize a domain, such as storage and recall processes in working memory, general mathematical knowledge and processes in quantitative reasoning, knowledge and command of the rules of language. The third stratum is general intelligence or *g* itself. Carroll posits that *g* involves complex higher order reasoning, such as inductive and analogical inference.

In terms of the model shown in **Figure 1**, representational capacity, inference, and attention and executive control are all parts of general intelligence. The web of rules, heuristics, and information in each of the domains mentioned above belongs to the second stratum. More specific knowledge belongs to the third stratum.

A distinction that is often coming in the literature in different terminology is the distinction between processes that are used to grasp and deal with the new and unknown processes used to handle familiar situations. The most famous is the distinction proposed by Cattell and Horn (Cattell, 1963) between fluid and crystallized intelligence. Fluid intelligence refers to the use of reasoning to solve unfamiliar problems. Crystallized intelligence refers to the investment of fluid intelligence into knowledge that is culturally and practically useful. In the present terms, fluid intelligence would involve inference, representational capacity, and executive control and also mental operations in the various domains. Crystallized intelligence would involve the level of knowledge and beliefs in the various domains and also the various mental operations as automated and customized for use in various professional or cultural contexts.

In Sternberg's (1997, 2004) theory, analytic intelligence refers to mental processes of representation, interrelation and transformation of information; creative intelligence refers to the processes addressed to novelty and the creation of knowledge and operation possibly useful in the future; practical intelligence generates solutions relevant and useful in a given time and environment.

What Is Stable and What Is Changing in Intelligence?

The question can be answered from three different points of view: the architecture of intelligence, individual differences in concern to it, and the kind of concepts and problems that can be dealt with at different periods of life.

Stability and Change in the Architecture of Intelligence and Intellectual Attainment

The architecture of intelligence is considered to be generally stable. The four systems described above (i.e., the relational domains, representational capacity, inference, and consciousness) are present from very early in life and they continue to be so until the end. Also, these systems are present across cultures or social groups. However, the relations between the various systems may vary according to experience, developmental phase, and culture (Demetriou et al., 2005; Kazi et al., 2012; Sternberg, 2004).

Stability and Change in Individual Differences

The stability of individual differences in IQ is debated for a long time. Many textbooks of intelligence often report that the IQ is generally stable over time, especially between 5 and 6 years of age to middle age. This position reflects the fact that the correlations between IQ scores obtained at different ages in this span are generally high (between 0.5 and 0.7). In other words, the position of about 25–50% of individuals relative to their age group remains largely stable over age.

It is now well known that the IQ of many people may both increase or decrease and that these changes can even be associated with concomitant changes in the brain (Ramsden et al., 2011). Schooling is a factor that causes increases in IQ. Many studies suggest that there is an increase of something between 1 and 4 IQ points for each extra year of schooling (Ceci, 1991; Winship and Korenman, 1997). Gustafsson (2008) found that this effect came from changes in fluid intelligence. Moreover, in the later years, after the age of 60 fluid intelligence starts to decline, especially processing efficiency and representational capacity. However, crystallized intelligence does not decline until much later (Salthouse, 2000; Schaie, 1996).

Stability and Change in Development

The quality of intelligence changes extensively with age. In differential psychology this is reflected in the construct of

mental age. That is, it is recognized that different ages are typically associated with different cognitive accomplishments, although some individuals may progress faster than others thereby excelling their age mates. Overall, with age, representational capacity expands ascending from about one chunk of information at 1–2 years of age to about five to seven chunks at about 15–16 years of age (Pascual-Leone, 1970; Kazi et al., 2012). As a result, individuals can deal with increasingly more representations and more complex relations. Also mental processing becomes faster, better focused on goal, and increasingly controllable so that it can shift across concepts or actions, if needed. As a result, self-guidance, reflection, and self-awareness increase with age. Inference, although somehow always present, becomes increasingly analytic and complex. Thus, it moves from inductions based on similarity or probabilistic judgments based on the pattern of events in time and space to deductive reasoning (Kail, 1991; Demetriou et al., 2002; Zelazo et al., 2005).

Deductive reasoning gradually evolves from an automatic, heuristic, and context-dependent use of inference embedded into one's own language, in preschool (e.g., you can play with your friends only if you finish your homework) to analytic, domain-free inference enabling processing of abstract underlying logical relations. Gradually, the focus of inference shifts from the actual content to the necessary relations between propositions, if accepted as true (e.g., "Birds fly, elephants would fly if they were birds"). This decontextualization of inference starts in the early years of primary school and continues until early adulthood.

Also, deception as contrasted to truth, the real or actual as contrasted to the phenomenal (Wellman, 1990), and symbols as tokens of realities and ideas are more easily grasped and handled with age (DeLoache, 2000). The difficulty of problems in tests of intelligence varies along these dimensions. Therefore, the higher the IQ score on the test, the higher an individual performed on these dimensions of complexity, abstraction, logicality, and flexibility.

Why Do Individuals Differ in Intelligence?

Intelligence is normally distributed in the population. According to many, this is because intelligence is like any other physical characteristic, such as height (Jensen, 1998). This interpretation is usually associated with the assumption that intelligence is genetically determined and the genes associated with physical characteristics are normally distributed in the population. Obviously, in the case of intelligence, reference is made to genes that are associated with the formation, the development, and functioning of the brain, which is the 'physical organ' of intelligence, so to speak. In fact, there is no doubt that there is a line that goes from genes to brain to intelligence. For example, genes have been discovered that relate to the overall brain volume observed in individuals and overall brain volume is related to general intelligence (Deary, 2012; Hunt, 2011). Also, the Down syndrome, which is associated with low IQ (in the range of 50–70), is caused by a specific problem in a particular position of the DNA (there are three instead of two chromosomes at the 23rd position). This problem results into specific malformations in the brain. These malformations handicap information representation and processing in the brain, resulting in low mental achievements.

Individual differences genetically determined are inheritable to a large extent. There has been extensive research of the relations between heredity and individual differences in intelligence. This research includes siblings of systematically varying genetic and environmental similarity, such as twin siblings who grow in the family of their physical parents or in different families because they have been adopted by different parents. This research makes it clear that similarity in intellectual achievements increases with genetic similarity. Overall, about 50% of variability in intelligence is due to heredity. However, this varies with age. It is about 40% in childhood, 55% in adolescence, 60% in early adulthood, and 80% later in life (Deary, 2012; Hunt, 2011). It seems that, with age, people become more selective of their environment and they even actively shape it according to their own genetic propensities, leaving little room for influences that deviate from what they feel is suitable and manageable for them. Therefore, a large part of individual differences in intelligence are indeed due to heredity but the effect of environment is also very important, particularly in early childhood.

Culture, Gender, Social Class, and Intelligence

There is a long debate about differences between cultures in intelligence. The discussion about possible differences in intelligence was polarized in the twentieth century between two completely diverging epistemological positions. The first maintains that intelligence is a primary human characteristic and that, therefore, is the same across cultures. This view takes a second step, assuming that tests of intelligence developed in the West do measure this universal human characteristic. Therefore, differences in attainment, such as differences in average IQ attained by different ethnic or cultural groups, reflect differences in how much of this characteristic is possessed by each group. Research conducted from this perspective established a rather well-documented finding. That Western nations average about 100 IQ points, Eastern nations, such as Chinese, average about 105 IQ points, and African nations average about 85 IQ points (Lynn, 2008).

The second position maintains that intelligence is the expression of the worldview of a culture. Therefore, according to this position, it is not even possible to have a common measure of intelligence across cultures with different worldviews. Thus, it is just wrong to use a system of measurement capturing a particular worldview to compare different cultures; this would miss the essence of intelligence as developed and used in all other cultures. Nisbett (2003) suggests that cultures differ along important dimensions, such as the orientation to the individual as contrasted to the social group, the orientation of a holistic as contrasted to an analytical approach to problems, and the orientation to real-world functional organization of knowledge as contrasted to an abstract principled-based organization. Differences between cultures reflect differences in their history which, in turn, reflect long-term differences in needs and activity imposed by differences in the environment. For example, the individualist stance of the

West initiated in the geographically compartmentalized world of the Ancient Greeks and their ensuing localized administrative and political structures. This is contrasted to the open large spaces of the Chinese environment and their centralized administrative and political structures. According to this position, differences in IQ, such as those noted above, do not convey any meaningful information other than IQ tests capture an expression of intelligence that is tuned to the West, it favors the East, and it disfavors the South (Nisbett, 2003).

Cross-cultural research on intelligence and cognitive development has shown that dominant worldviews do shape how problems are approached and solved. Luria (1976), for example, showed that Asian peasants refuse to take the analytical stance and apply inference to solve a simple syllogistic reasoning problem: "All bears in the wood are black; Mischa met a bear today; what color is this bear?" These peasants responded that they could not know, provided that they did not see the bear themselves!

Many other studies showed that individuals in different cultures may be very proficient in solving locally relevant problems but very inept in dealing with standard school-like or test-like problems. Cole et al. (1971) showed that children in the Kpelle tribe in Africa use locally realistic and functional criteria for classification (robin is flying) rather than taxonomic criteria (robin is a kind of bird). The use of taxonomic criteria increased with schooling. Gladwin (1970) showed that the Puluwat in the South Pacific developed a sophisticated system of navigation that integrates information about the weather, the wind, the ocean currents, and the stars. This system enables them to navigate efficiently between islands. Carraher and colleagues showed that street children in Brazil, who make their living by selling various goods on the street, were more proficient in arithmetic calculations related to their business than in the standard school-based formal context. Moreover, they used strategies involving the mental manipulation of quantities in their street calculations (e.g., they produce global quantities which they then combine) rather than the manipulation of the symbols standing for the four arithmetic operations, as it is usual at school (Carraher et al., 1985).

This pattern of findings makes meaning in the context of the framework for intelligence presented here. That is, human intelligence is the same across human cultures. However, as an interface with the environment, intelligence is affected and constrained by cultural practices and knowledge and it is expressed through them. Thus, the command of general processes, such as executive control of representational capacity and the use of analytical inference, reflect the level of abstraction attained by a group. For example, the analytical stance is secondary in development and it needs to be cultivated and practiced systematically before it can be used spontaneously. At the same time, advanced knowledge and expertise reflect how general cognitive processes and abilities interact with a cognitive domain in cultural context. The success of Brazilian street children in their street version of mathematics does not indicate that these children have a different mind than their middle-class children in school. It only reflects that their transaction needs fostered the development of general inference and quantitative reasoning in ways that were enough for the task but of course were short of the needs of formal computation skills needed to deal with abstract arithmetic tasks at school. In the same way, the sophisticated navigation system of the Puluwat indicates that general inferential mechanisms, spatial thinking, and other domains are all integrated to generate expertise in a highly valued culture-relevant domain. Obviously, this system is very remote from the abstract electronically based navigation system used to navigate an airplane. However, they both organize space and time according to a complicated system of rules and knowledge that needs extended training and practice to be mastered. The systems outlined here are involved in both cases.

Different cultures may have long established institutions that may facilitate the relations and use of key systems of intelligence making them to excel in overall intellectual performance compared with other cultures. Schooling is probably the most influential of these institutions in fostering intelligence. It is reminded that each extra year of schooling causes increases in general intelligence in each student. In fact, this may even be related to the general increase of intelligence across generations. According to the so-called Flynn effect, general intelligence increases in the general population over the years. Flynn discovered that IQ increased by about 10 IQ points every 30 years since the beginning of the twentieth century (Flynn, 1987). Flynn (2009) ascribed this increase to two main factors. The increasing pressure for abstract thought that is required by our increasingly symbolic technological environment and the increasing expansion of education.

There may be other, more specific cultural institutions that may relate to the overall intellectual achievements of different cultures. A good example here is the learning of the logographic system of the Eastern cultures, such as Chinese or Japanese. Learning this system is extremely demanding on processing efficiency and working memory because children, from early in life, must become able to recognize and produce thousands of complicated visual patterns. There is evidence showing that Chinese children are faster in recognizing visual patterns and in their working memory capacity. This enhances general cognitive fluidity due to its pressures for symbol–meaning integration, generating a general advantage in information processing and problem solving (Demetriou et al., 2005; Kazi et al., 2012; McBride-Chang et al., 2011). It might be the case that the advantage of Eastern nations in general IQ noted above is related to this advantage.

There are factors within cultures that have been associated with differences in intelligence. The role of gender has been rather controversial in the literature. Overall, it is believed, more so in the past rather than at present, that males exceed females in mathematical and spatial reasoning and females exceed makes in verbal reasoning. These differences tend to diminish over the years and they have been associated with cultural differences in education and stance toward some activities related to the genders (Hunt, 2011).

The impact of social class is equally controversial. It is well documented that children growing in less privileged environments in concern to parental education and income have lower intellectual attainments (Hunt, 2011; Mackintosh, 1998). For many, this is the result of environmental factors related to the opportunities provided to children during their developmentally

critical years (Ceci and Williams, 1997). For others, this is the result of hereditary factors associated with selection of partners for mating and environmental contexts. According to this view, more intelligent individuals, who are generally more socially and professionally successful, have children with more intelligent individuals and less intelligent individuals have children with less intelligent individuals (Herrnstein and Murray, 1994). Thus, class differences are perpetuated for hereditary rather than social reasons. An interaction between these factors may be a more balanced interpretation of the differences (Hunt, 2011).

Can Intelligence be Changed by Education?

In this section we will outline a set of principles that education would have to adopt if it would directly and systematically both raise intelligence and facilitate knowledge acquisition. These principles derive from the four-system model summarized here and they are illustrated in **Figure 2** (see Demetriou, 2014; Demetriou et al., 2011; for a more detailed discussion of these principles and their implementation and also for other relevant references).

At any phase, education must lead the student to develop and refine the following cognitive skills: focus on relevant information; scan, compare, and choose according to goal; ignore irrelevant information; represent what is chosen and associate with extant knowledge; bind into models and rehearse if necessary; evaluate models in reference to evidence; reason by deduction to evaluate truth and validity of models and conclusions; prefer solutions that are better or nicer than extant solutions; estimate consistency with beliefs, extant theories, dominant views, and so on. Encode, symbolize, and embed into the system. Being aware of epistemological issues concerning the similarities and differences between disciplines is an important part of education for critical thinking and creative thinking.

These general educational goals must be tuned to the developmental possibilities and constraints of successive developmental phases. The general rules that may govern this tuning are as follows:

1. Educational priorities must capitalize on developmental milestones to make education feel relevant, appropriate, and useful for each phase of life. For example, education in infancy should be directed to practicing core operators to enhance infants' (1) understanding of the nature, behavior, and relations of objects and (2) their own relations with objects and other persons. Education at the preschool years must provide conditions for the knowing and handling of the multiplicity of representations, points of view, and their relation to reality. In the primary school years education must focus on revealing the connections between concepts and operations and facilitate children to see their own overt or mental actions vis-à-vis the concepts and operations. Education in early adolescence should focus on developing the "if … then" stance approach to problem solving, build epistemological awareness about the characteristics, possibilities, and limitations of different knowledge domains and sharpen the use of appropriate methods for different kinds of worlds or problems.

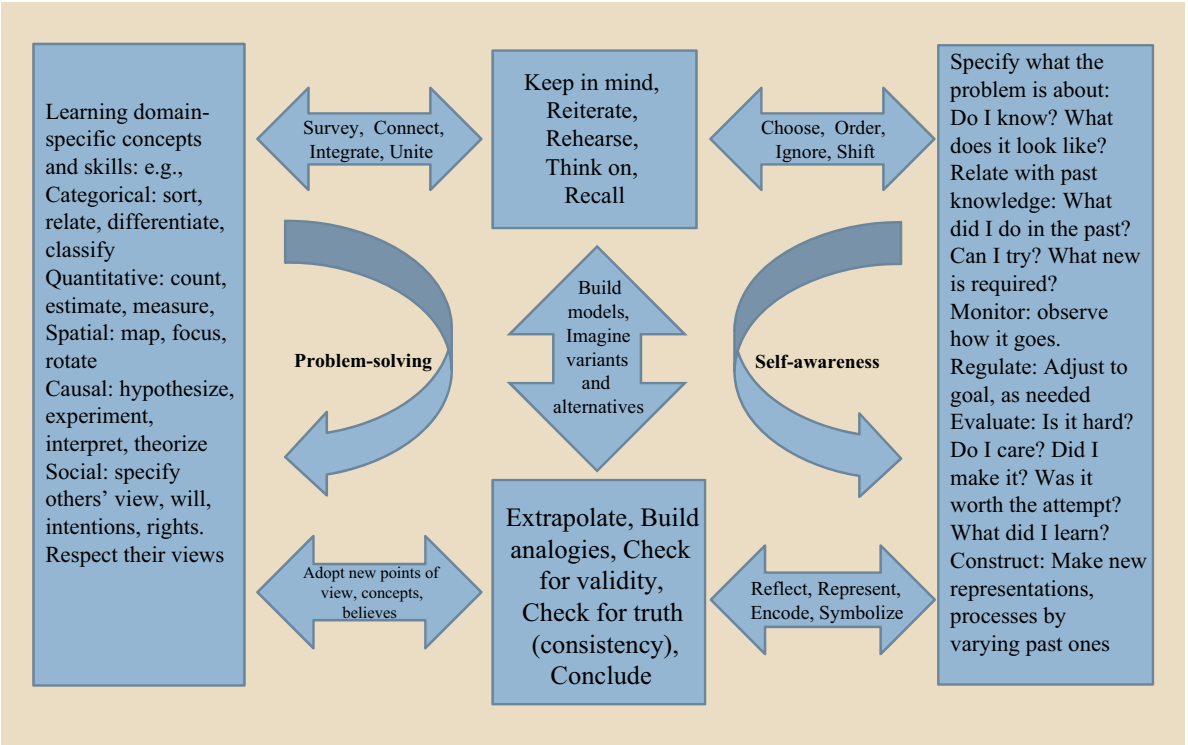

Figure 2 Function to be educated across the various systems and processes of the mind.

2. Intelligence may be improved by instruction addressed directly to the various aspects of working memory, attention, and executive control. Specifically education should aim to improve the ability to (1) monitor ongoing performance, (2) keep track of the flow of information and update it for as long as needed, (3) manage more than one task simultaneously (e.g., by alternating between blocks or types of information, if needed), (4) bind items according to type and time of presentation, and (5) inhibit irrelevant items. At the same time, organization and presentation rate of material to be learned must take into account the representational capacity and processing efficiency that is presently available to make learning possible and powerful at every phase. That is, the school must present concepts when the representational potential of individuals meets the representational demands of concepts to be taught.
3. Education must practice, consolidate, and build models for mental operations that are specific to different domains. For example, it would be useful for classificatory thought to have a template for classification and categorical relations between concepts, such as tree diagrams. Also, a model of the basic relations of causality (i.e., necessary and sufficient, necessary and not sufficient, neither necessary nor sufficient, and incompatible) would be helpful in directing the student how to manipulate and grasp cause–effect relations. Students must be induced into theme- or concept-appropriate models of the fields of interest, such as gravity, inheritance, motion. Teaching must lead the students to understand where and why these models are better than their own respective concepts or models.
4. Special attention should be given to decontextualization of inference from content and context. For example, at preschool, children may be trained to realize that the information in the premises is connected by inference. At primary school, directed comparisons across different types of arguments would enable children to differentiate form from content and understand that logic constrains inference. In adolescence education may focus on the formal aspects of different types of arguments.
5. Problem solving involves all of the processes above and it should be demonstrated and practiced as such. Students must be educated in foresight, anticipation, and formulation of alternative plans for problems. Envisaging, comparing, substantiating, debating, and choosing among alternative models, if properly directed, would make the student able to become sensitive to the role of evidence and logical and conceptual substantiation for alternative models. Moreover, with age, knowledge and context are increasingly important for successful, acceptable, and creative problem solving.
6. Finally, students must know themselves: the organization and functioning of their own mind, their own strengths and weakness, and they must be able to adjust their actions and learning accordingly. Each process has a role and each solution has a value, given the context in which it was produced and in which it will be used. For example, Zelazo (2011) suggests that mindfulness training since preschool enhances self-monitoring skills which in turn enhance executive control thereby improving learning. At primary school, education must focus on building awareness of the differences between mental functions and of their differential impact on learning. Adolescents may be guided to grasp the differences between formal disciplines, such as mathematics, and empirical disciplines, such as physics, in knowledge construction. This would enable them to evaluate the differences between the knowledge produced by these disciplines.

Conclusions

In conclusion, the human mind is a complex system of general and specialized processes. "The relations between the general and the specialized processes are complex and bi-directional. On the one hand, general processes set the limits for the construction, operation, and development of the domain-specific systems. On the other hand, specialized processes provide the frame and raw material for the functioning of general processes" (Demetriou et al., 2010: pp. 34–35). The development of each and of their relations is long and dynamic, and it occurs on several fronts at the same time, and it can be described in terms of complexity, abstraction, analytical attitude, long-term planning, and ingenuity in organizing and using knowledge for the sake of personal and collective good and development. This always takes place in particular culture at a particular time. Therefore, in Sternberg's words, "Intelligence cannot be fully or even meaningfully understood outside its cultural context" (2004: p. 325).

See also: Cognitive Development During Infancy and Early Childhood across Cultures; Cognitive Development, Educational Theories of; Cognitive Development: Child Education; Cultural Intelligence and Competencies; Intelligence, Genetics of: Cognitive Abilities; Intelligence, Prior Knowledge, and Learning; Intelligence: Assessments of; Intelligence: Central Conceptions and Psychometric Models; Intelligence: Historical and Conceptual Perspectives; Intelligence: History of the Concept; Learning to Learn; Piaget's Theory of Human Development and Education.

Bibliography

Baddeley, A.D., 1990. Human Memory: Theory and Practice. Erlbaum, Hillsdale, NJ.
Baddeley, A., 2012. Working memory: theories, models, and controversies. Annual Review of Psychology 63, 1–29.
Carraher, T.N., Carraher, D., Schliemann, A.D., 1985. Mathematics in the streets and in schools. British Journal of Developmental Psychology 3, 21–29.
Carey, S., 1985. Conceptual Change in Childhood. The MIT Press, Boston, MA.
Carey, S., 2009. The Origins of Concepts. Oxford University Press, Oxford.
Carroll, J.B., 1993. Human Cognitive Abilities: A Survey of Factoranalytic Studies. Cambridge University Press, New York.
Case, R., 1985. Intellectual Development. Birth to Adulthood. Academic Press, New York.
Cattell, R.B., 1963. Theory of fluid and crystallized intelligence: a critical experiment. Journal of Educational Psychology 54, 1–22.
Ceci, S.J., 1991. How much does schooling influence general intelligence and its cognitive components? Developmental Psychology 27, 703–722.
Ceci, S.J., Williams, W.M., 1997. Schooling, intelligence, and income. American Psychologist 52, 1051–1058.

Cole, M., Gay, J., Glick, J., Sharp, D.W., 1971. The Cultural Context of Learning and Thinking. Basic Books, New York.

Daniels, P.T., Bright, W., 1996. The World's Writing Systems. Oxford University Press, Oxford.

Deary, I.J., 2012. Intelligence. Annual Review of Psychology. http://dx.doi.org/10.1146/annurev-psych-120710-100353.

DeLoache, J.S., 2000. Dual representation and young children's use of scale models. Child Development 71, 329–338.

Demetriou, A., 2014. Learning to learn, know, and reason. In: Deakin-Crick, R., Stringer, C. (Eds.), Learning to Learn. Routledge, London, pp. 43–68.

Demetriou, A., Christou, C., Spanoudis, G., Platsidou, M., 2002. The development of mental processing: efficiency, working memory, and thinking. Monographs of the Society of Research in Child Development 67 (268).

Demetriou, A., Kui, Z.X., Spanoudis, G., Christou, C., Kyriakides, L., Platsidou, M., 2005. The architecture, dynamics, and development of mental processing: Greek, Chinese, or Universal? Intelligence 33, 109–141.

Demetriou, A., Mouyi, A., Spanoudis, G., 2010. The development of mental processing. In: Overton, W.F. (Ed.), Biology, Cognition and Methods across the Life-span. In: Lerner, R.M. (Editor-in-chief), Handbook of Life-span Development, vol. 1. Wiley, Hoboken, NJ, pp. 306–343.

Demetriou, A., Spanoudis, G., Mouyi, A., 2011. Educating the developing mind: towards an overarching paradigm. Educational Psychology Review 23 (4), 601–663.

Flynn, J.R., 1987. Massive IQ gains in 14 nations. Psychological Bulletin 101, 171–191.

Flynn, J.R., 2009. What Is Intelligence: Beyond the Flynn Effect. Cambridge University Press, Cambridge.

Gardner, H., 1983. Frames of Mind: The Theory of Multiple Intelligences. Basic Books, New York.

Gladwin, T., 1970. East Is a Big Bird. Harvard University Press, Cambridge, MA.

Gustafsson, J.-E., 2008. Schooling and intelligence: effects of track of study on level and profile of cognitiv abilities. In: Kyllonen, P.C., Roberts, R.D., Stankov, L. (Eds.), Extending Intelligence: Enhancement and New Constructs. Lawrence Erlbaum Associates, New York, pp. 37–59.

Herrnstein, R.J., Murray, C., 1994. The Bell Curve: Intelligence and Class Structure in American Life. The Free Press, New York.

Hunt, E.B., 2011. Human Intelligence. Cambridge University Press, New York.

Jensen, A.R., 1998. The G Factor. Praeger, Westport, CT.

Jung, R.E., Haier, R.J., 2007. The parieto-frontal integration theory (P-FIT) of intelligence: converging neuroimaging evidence. Behavioral and Brain Sciences 30, 135–154.

Kail, R.V., 1991. Developmental change in speed of processing during childhood and adolescence. Psychological Bulletin 109, 490–501.

Kaufman, J.C., Baer, J., 2006. Creativity and Reason in Cognitive Development. Cambridge University Press, Cambridge, UK.

Kazi, S., Demetriou, A., Spanoudis, G., Zhang, X.K., Wang, Y., 2012. Mind-culture interactions: how writing molds mental fluidity in early development. Intelligence 40, 622–637.

Lynn, R., 2008. The Global Bell Curve: Race, IQ, and Inequality Worldwide. Washington Summit Publishers, Washington, DC.

Luria, A., 1976. Cognitive development: its cultural and social foundations. Harvard University Press, Cambridge, MA.

Mackintosh, N.J., 1998. IQ and Human Intelligence. Oxford University Press, Oxford.

McBride-Chang, C., Zhou, Y., Cho, J.-R., Aram, D., Levin, I., Tolchinsky, L., 2011. Visual spatial skills: a consequence of learning to read? Journal of, Experimental Child Psychology 109, 256–262.

Mithen, S., 1996. The Prehistory of the Mind: The Cognitive Origins of Art, Religion and Science. Thames and Hudson, London.

Nisbett, R.E., 2003. The Geography of Thought: Why We Think the Way We Do. Free Press, New York.

Pascual-Leone, J., 1970. A mathematical model for the transition rule in Piaget's developmental stages. Acta Psychologica 63, 301–345.

Piaget, J., 1970. Piaget's theory. In: Mussen, P.H. (Ed.), Carmichael's Handbook of Child Development. Wiley, New York, pp. 703–732.

Posner, M.I., Rothbart, M.K., 2006. Educating the Human Brain. American Psychological Association, New York.

Ramsden, S., Richardson, F.M., Josse, G., Thomas, M.S.C., Ellis, C., Shakeshaft, C., et al., 2011. Verbal and non-verbal intelligence changes in the teenage brain. Nature 479, 113–116. http://dx.doi.org/10.1038/nature10514.

Raven, J., Raven, J. (Eds.), 2008. Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven's Quest for Non-arbitrary Metrics. Unionville. Royal Fireworks Press, New York.

Salthouse, T.A., 2000. Aging and measures of processing speed. Biological Psychology 54, 35–54.

Sampson, G., Gil, D., Trudghill, P. (Eds.), 2009. Language Complexity as an Evolving Variable. Oxford University Press, Oxford.

Sternberg, R.J., 1997. Successful Intelligence. Plume, New York.

Sternberg, R.J., 2004. Intelligence and Culture 59, 325–338. http://dx.doi.org/10.1037/0003-066X.59.5.325.

Sternberg, R.J., 2011. Cognitive Psychology. Cengage Learning, New York.

Schaie, K.W., 1996. Adulthood and old age. In: Husen, T., Postlewaithe, T.N. (Eds.), International Encyclopedia of Education. Pergamon Press, Oxford, pp. 163–168.

Tenenbaum, J.B., Kemp, C., Griffiths, T.L., Goodman, N.D., 2011. How to grow a mind: statistics, structure, and abstraction. Science 331, 1279–1285. http://dx.doi.org/10.1126/science.1192788.

Wellman, H.M., 1990. The Child's Theory of Mind. Bradford, Cambridge, MA.

Winship, C., Korenman, S., 1997. Does staying in school make you smarter? the effect of education on IQ in the Bell Curve. In: Devlin, B., Fienberg, S.E., Resnick, D.P., Roeder, K. (Eds.), Intelligence, Genes, and Success. Scientists Respond to the Bell Curve. Springer-Verlag, New York, pp. 215–234.

Zelazo, P.D., 2011. Mindfulness training in childhood. Human Development 54, 61–65.

Zelazo, P.D., Qu, L., Müller, U., 2005. Hot and cool aspects of executive function: relations with early development. In: Schneider, W., Schumann, R., Hengsteler, R., Sodian, B. (Eds.), Young Children's Cognitive Development: Interrelationships Among Executive Functioning, Working Memory, Verbal Ability, and Theory of Mind. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, USA.

Intelligence, Prior Knowledge, and Learning

Elsbeth Stern, ETH Zürich, Zürich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Intelligence test scores can account for achievement differences in many content areas to a considerable extent. An individual's intelligence quotient results from complex interactions between genes and environmental stimulation, foremost schooling. The amount of variance in intelligence to be explained by genes is the higher the more successful a society is in providing cognitively stimulating environments for everybody. Intelligence can be understood as a start-up resource of information processing which has to be invested in knowledge in order to enable competencies in a domain. Teacher's professional competencies have a major impact on how learners exploit their intelligence for learning particular subjects.

Intelligence: A Valid Predictor of Achievement Differences

Individuals with similar cultural, social, and educational backgrounds differ from one another in the time they need to process certain information, in their ability to understand complex ideas, in the efficiency with which they can deal with novel, transfer demanding tasks, and in the learning outcome that results from attending certain instructional environments. The construct of psychometric intelligence attempts to clarify what is behind such achievement variations that cannot be explained by differences in learning environments or in amount of practice.

Measuring Intelligence

About a century ago Alfred, Binet constructed problems designed to determine whether children who did not meet certain school requirements suffered from mental retardation or from behavioral disturbances. Since then many psychologists have been quite successful in developing reliable verbal and nonverbal intelligence tests for children and adults. Intelligence tests contain items composed of verbal, numerical, and pictorial material, and they require various mental operations, among them inductive and deductive reasoning, pattern recognition, and memorization. So-called speed tests contain items that are comparably easy for everybody; individual differences in the numbers of correctly solved problems only occur because of time limitations. In power tests, the items are ordered according to their difficulty, and limits in intelligence become apparent if people do not solve all problems despite having sufficient time. The distribution of achievement scores in all intelligence test scales follows the bell curve (normal distribution). This reflects the fact that the majority of people resemble each other quite a lot with respect to their cognitive capabilities, and only a few people show extraordinarily low or high competencies. Normal distribution is the statistical prerequisite for measuring intelligence on interval level by indicating deviations from the mean score. To determine the intelligence quotient (IQ), test scores are converted to a scale in which by convention the mean is 100 and the standard deviation is 15. The reliability of IQ tests, revealed either by correlation coefficients based on repeated measurement or by figuring out internal consistency, is between 0.80 and 0.90, which is higher than for most other psychometric measures. Nonetheless, a reliability lower than 1.0 only allows us to interpret a range rather than a single value. For example, if a person's tested IQ is 110, and if the reliability of the test is 0.90, the IQ of this person is between 101 and 119 with a probability of 95%. A test reliability of 0.80 reveals a range of 97–123.

Despite broad variations in the content and the form of presentation of intelligence test items, all tests have in common that they do not presuppose knowledge that can only be acquired in particular learning environments not accessible to everybody. Moreover, differences in test scores within a group only reflect differences in intelligence between these persons if all of them have had access to comparable learning environments. In other words, each member of the group must have had, in principle, the opportunity to acquire the knowledge necessary to solve the problems. Intelligence test scores can be raised considerably by practicing the respective types of items, while individual differences do not disappear but rather remain quite stable at a higher level. The Flynn effect, named for its discoverer, is the observation that modernization of a nation goes along with massive IQ gains over time. In the meantime, many highly developed countries have reached an asymptote, while nations where modernization began later still gain around three IQ points per decade. Reasons for the Flynn effect are manifold and still under debate, but it is undisputed that better schooling as well as widespread exposure to tasks which resemble intelligence test items in various media are major causes.

Individual differences in intelligence test scores between persons can only be interpreted as differences in general cognitive resources if similar amounts of practice can be presupposed. Attempts to construct so-called culture-free or culture-fair tests, which were supposed to be unaffected by prior experience, have failed because it turns out that different cultures are not prepared in the same way even for dealing with nonverbal material and mental operations that are not part of institutional schooling. Although some studies reveal ethnic differences in mean IQ, hitherto there is no convincing

evidence that these differences are genetically affected. There is rather overwhelming evidence that all ethnic groups gain IQ points as a consequence of better schooling. Within a fairly homogenous cultural context, intelligence can be considered as a personality trait mainly for two reasons. First, performance on intelligence items that are based on different contents and mental operations show significant correlations. For instance, correlations between tests on inductive reasoning with pictorial material and with verbal material are around 0.50, and correlations between tests that measure basic visual–spatial competencies such as mental rotation and tests of verbal fluency are about 0.30. In numerous studies run all over the world, multivariate statistical methods such as factor analysis have revealed that a single factor, called factor *g*, can account for 40–50% of the variance in IQ test batteries composed of various scales covering a range of content knowledge and mental operations. Second, longitudinal studies have revealed that IQ is a fairly stable measure across the life span. Long-term stability of IQ for adolescents and adults comes close to the reliability of the tests, as, among others, a once-in-a-century study from Scotland has shown (Deary, 2012). Even cognitive measures of attention gained in early childhood reveal long-term correlations are around 0.50 for both verbal and nonverbal tests. Overall, intelligence test scores predict academic performance fairly well: the correlations between IQ and grades in school and university are about 0.50. The correlations between intelligence test scores and measures of outside-school success such as income or professional status are lower but still significant. That intelligence cannot account for a larger amount of achievement variation is not at all surprising, given the importance of numerous other factors, among them social background, motivation, and effort. A longitudinal study by Duckworth et al. (2011) revealed that standardized achievement test scores were stronger related to intelligence than to motivational factors, while grades determined by teachers were determined more by factors like self-control than by intelligence.

Given that verbal and nonverbal intelligence tests are good predictors of how well an individual will succeed in school and university, they are quite helpful for making recommendations for different educational trajectories. They allow educators to identify children who cannot be expected to gain sufficiently from standard schooling and therefore need extra support adapted to their special needs, or children who might benefit from an advanced learning environment. The validity of intelligence tests can be further increased by the dynamic testing approach (Sternberg and Grigorenko, 2002), which means that all test-takers practice the items in several trials by getting feedback and thinking hints so that individual differences in familiarity with tests are compensated for. Particularly for children from disadvantageous social environments, learning tests are a more reliable and valid measures of intelligence than conventional IQ tests. Attempts to extend the construct of intelligence by including social and emotional competencies as well as striving for success are controversial because the theoretical background of these concepts is still vague, and, more importantly, because the tests designed to measure these aspects do not meet the strong diagnostic criteria that have been established in psychometric research (Neubauer and Freudenthaler, 2005).

What Is Behind Intelligence: Factor *g*, Working Memory, and Specific Abilities

What cognitive capabilities are behind factor *g* is still under debate, although in the past decade major insights could be seen to emerge. This is also due to the fact that much has been learned about the neural underpinnings of human intelligence. The substantial negative correlations between IQ and reaction times in solving simple problems that had been found in numerous studies can now be interpreted in the light of brain imaging research. A constant finding is that more intelligent individuals show less brain activation (measured by electroencephalography (EEG) or functional magnetic resonance imaging (fMRI)) when completing intelligence test items. The more intelligent a person is, the less neural resources he or she needs for solving the problem (Neubauer and Fink, 2009). Particular progress has been made in localizing the brain areas which indicate differences in intelligence: the prefrontal cortex, an area which plays a major role in dealing with complex and novel problems as well as in working memory functions. Considerable progress has been made in bringing together research on working memory and intelligence (Oberauer et al., 2008). Working memory is understood as a central cognitive function of human beings which is responsible for temporarily maintaining and manipulating knowledge during cognitive activity. Working memory functions are measured by tasks that require the goal-oriented active monitoring of incoming information or reactions under interfering and distracting conditions. Performance variations in these tasks have been found to be highly related (average correlations of 0.70 are reported) to differences in intelligence test scores. Psychologists agree that although there is more to IQ than working memory functions, the latter should be regarded as an explanatory construct for reasoning abilities.

Although factor *g* can be extracted from different intelligence tests, many correlations between intelligence subtests, though significant, are low, indicating the involvement of independent mental resources. Even between tests on inductive reasoning which are based on different forms of representation (i.e., pictures, numbers, words), only medium correlations are revealed, suggesting that cognitive processes are to a large extent guided by specific verbal, visual–spatial, or numerical abilities, among others. Since the 1960s, concerted attempts have been made to integrate general and specific abilities into broader theories of intelligence in a hierarchical model which is now known as the Cattell–Horn–Carroll theory, which has been recently discussed in light of modern IQ tests by Kaufman (2009). Psychologists now agree that the abilities which contribute to intelligence tests scores are best classified on three levels. Narrow abilities on level one either refer to speed of sensory processes or specific competencies acquired through learning (e.g., reading skills). The broad abilities on level two go on from the distinction between fluid intelligence (Gf) and crystallized intelligence (Gc). Gf mainly describes logical reasoning, measured by content-poor nonverbal tests, and Gc, which is particularly revealed in content-rich verbal tests, represents the accumulation of reading and writing skills as well as higher-order knowledge over the life span of an individual. Level three describes

a unified component of cognitive competencies which corresponds to the general factor.

The distinction between Gf and Gc is important in light of age-related changes of intelligence. While both kinds of intelligence are closely interrelated in childhood, from the age of about 25 years on, however, Gf and Gc take different developmental trajectories. While Gf starts to decrease at the age of 25, first slightly and from about 50 years on more quickly, Gc is unaffected by age and even has good chances to increase until the age of around 70. It is worth noticing that because of the large individual differences, the given age information is only a rough estimation. Developmental changes also occur with respect to the structure of intelligence. The prominence of Gf decreases during childhood while Gc as a function of repeated enlargement and rearrangement of knowledge becomes more useful.

Intelligence as the Result of Gene–Environment Interactions

The question of heritability of intelligence can still cause quite a stir in the broader public. In science, however, the controversy on whether genes have an impact on the development of intelligence is a thing of the past. To make a long story short, it is no longer the question of nature *versus* nurture but nature via nurture (Ridley, 2003). Every aspect of human behavior is embodied in genes but it is the environment which triggers gene expression. Particularly the comparison between identical and fraternal twins shed light on the impact of genes on intelligence. The correlation between the intelligence test scores of raise-together identical twins almost ties up with the reliability coefficient of the respective test. At the same time, the correlation between fraternal twins rarely exceeds the correlation found between regular siblings, although the latter group due to their age differences had been sharing less experience than twins. Other studies revealed that the IQ of young adults who had been adopted shortly after birth was more similar to their biological than to their adoptive parents.

Altogether, twin and adoption studies suggest that 50–80% of the IQ variation is due to genetic differences. This relatively large range of percentage across different studies is due to the population dependence of heritability. It is namely the case that the amount of variance in intelligence test scores explained by genes is higher as more members in a society have access to school education, health care, and sufficient nutrition. Several studies revealed lower heritability of intelligence for children raised in lower socioeconomic status (SES) families (Turkheimer et al., 2003). It turned out that lower SES fraternal twins resembled each other more than higher SES ones, indicating a stronger impact of shared environment under the former condition. Or, in other words, because of the less stimulating environment in lower SES families, the expression of genes involved in the development of intelligence is liable to be hampered.

The complex interaction between genes and environment is also founded on the fact that heritability of intelligence increases during the life span. To understand this very well-established finding, one has to realize that societies which provide access to a broad variety of cognitive activities in professional as well as in private life enable adults more than children to actively select special environments which fit their genes. People who have found their niche can perfect their competencies by deliberate learning. Although, however, it is beyond any shadow of doubt that in developed societies, genes can explain a huge amount of IQ differences, the search for the genes responsible for the expression of cognitive capabilities has not at all met with much success, despite the money and effort invested in human genome projects. Given, however, that even for height less than 20% of the variance can be traced back to already identified genes, it is far from surprising that it is almost impossible to track down the genes that are involved in intelligence (Deary, 2012). It is entirely plausible that very large numbers of genes are spread out across the entire genome and have their share, and moreover these genes seem to interact in very complicated ways with each other as well as with environmental cues. In the foreseeable future, biologists will not be able to predict a baby's cognitive capabilities from his or her DNA. Personalized education, which means that a child gets the education that fits his or her genome, is nothing but a pipe dream. What can, however, be said for sure is that in societies which provide a cognitively stimulating environment for everybody, intelligence differences will not decrease but rather increase on a high level.

Genes, however, not only indirectly guide learning and knowledge acquisition via general and specific abilities. There is rather growing evidence that during evolution, the human mental architecture has been equipped with quite specific knowledge structures, for instance, about visual or auditory patterns, numbers and magnitudes, physical objects, language use, and social situations. Such knowledge structures, typically labeled as core knowledge or privileged knowledge (Spelke and Kinzler, 2007), allow human beings appropriate cognitive and behavioral functioning from the very beginning almost without effort. Such kind of fast learning sharply contrasts with the huge difficulties human beings may encounter when it comes to learning at school. For better understanding the obstacles and often even the severe difficulties that can occur when youngsters are supposed to acquire academic skills and competencies as there are reading, writing, or mathematical and scientific reasoning, one has to take into consideration the evolution of these fields in a cultural context during the past millenniums if not centuries or even decades. A today's child, equipped with a DNA comparable to the one of a stone-age person, is expected to acquire knowledge within few years which took mankind millenniums to develop. The core knowledge human beings have been equipped with during evolution can be the starting point for building more advanced knowledge structures in different domains. In the case of mathematics, it is entirely plausible that nature has equipped human beings with intuitive number knowledge and sensitivity for magnitude. As a consequence, to the best of our knowledge, all cultures have numbers words, even illiterate ones. However, having specific symbols for numbers is even not common among all cultures with script. The Arabic place value number system, which is now common in most parts of the world, was only developed a few hundred years ago. Only after the number '0' had made its way from India via the Arabic countries to Europe that the preconditions for developing our decimal system were given. It was the Arabic number system which opened up the pathway to academic mathematics.

Cultural transformations based on invented symbol systems were the key to advanced mathematics. Central content areas in mathematics curricula of high schools, such as calculus, were only developed less than three centuries ago. What makes school a real challenge is the fact that within a few years young people have to acquire knowledge which has been developed over centuries by genius minds with tremendous effort. While intuitive quantification seems to be a universal cognitive resource almost unaffected by differences in intelligence, correlations of $r = 0.50$ and higher are found between IQ and achievement in mathematics. Discovering principles that underlie the relationships between numbers and understanding how some of these principles can be described in verbal, numerical, and graphical representations required general cognitive resources. This leads to the more general question of how domain-specific knowledge has to be represented in order to allow reasoning and drawing inferences. In the past decades, cognitive scientists have developed theories of knowledge representation which enable higher-order reasoning and academic competencies. Moreover, educational science has worked our frameworks for learning environments that can foster the acquisition of usable knowledge.

Knowledge Construction

Intelligence can be understood as a start-up cognitive resource which has to be invested in the construction of knowledge in order to enable behavioral and cognitive functioning, including problems solving and decision making. How well such activities work out particularly depends on the quality of knowledge representations in the particular content domain. Knowledge has a multifacet nature, among them knowledge about facts or abstract concepts, about how to efficiently solve routine problems, and knowledge about more generally applicable learning strategies. These different facets of knowledge all interact in contributing to a person's competence, and they can differ in their functional characteristics. Elements of knowledge can be isolated or interrelated, context-bound or context-general, abstract or concrete, implicit or conscious, inert or accessible with different degrees. Just having a high accumulation of factual knowledge in a domain is usually not at all enough for cognitive and behavioral functioning. To characterize this situation, the term 'inert knowledge' has been invented. An example of inert knowledge is vocabulary of a foreign language which has been learned and stored in memory in a dictionary style, but cannot be retrieved during natural communication.

Cognitive scientists have agreed on distinguishing between declarative (knowing that) and procedural (knowing how) knowledge, a differentiation that can be applied to a broad variety of subject areas. Declarative knowledge can be communicated because it is represented on the basis of symbol systems (language, script, mathematical, or visual–spatial representations). Declarative knowledge can be applied to concrete instances and facts, or to general and abstract knowledge of the core principles and their interrelations in a domain. It is assumed to be stored mentally in some form of relational representation, for example, schemas or semantic networks which allow for its flexible transformation through processes of inference and elaboration. Declarative knowledge is therefore not bound to specific problem types and it is, in principle, transferable to other problems.

Procedural knowledge, in contrast to declarative knowledge, is usually seen as knowledge of operators and the conditions under which they can be applied to reach certain goals. It can be automated to different degrees, depending on the extent of practice. Automated procedural knowledge can be used with minimal conscious attention and few cognitive resources. This efficiency, however, has the drawback of inflexibility. Because automated knowledge is only partly open to conscious inspection, it can hardly be verbalized or transformed by higher mental processes. As a consequence, it is often tied to specific problem types. Both kinds of knowledge are involved in reasoning and problem solving in any domain, but they are acquired in different pathways. Procedural knowledge results from practice and repetition. The more often we conduct an action in the correct way, the less attention and control will be needed.

Declarative knowledge can be subdivided into factual and conceptual knowledge. While facts can be acquired by rote-learning, conceptual knowledge, in contrast, results from conscious elaboration and reasoning with progressively focusing on defining rather than on characteristic features. This transition was studied for a great variety of concepts from different domains. Younger children, for instance, associate the concept of 'parents' with 'caring for young children,' whereas older children focus on 'having offspring.' Younger children in elementary school will agree that a pile of rice has weight, but deny that an individual grain of rice has weight as well. This seemingly implausible answer turns out to be highly plausible if one realizes that for younger children, 'weight' and 'being heavy' are still equivalent. When being asked whether a grain of rice has weight if it is put on the back of an ant, their answer is yes with deep conviction (for an overview, see Stern, 2005).

Research on learning science and mathematics has shed light on numerous differences in conceptual understanding between experts and novices. This is a particular challenge for schools in general and for teachers in particular. Particularly in subjects like science and mathematics, many teachers are disappointed about their student's scant learning gains despite the effort they put into preparing and structuring the lessons. They strictly keep to a logical sequence, and they present very clear and precise definitions, preferably based on mathematical formulas. As long as tests require students to reproduce definitions and to figure out quantitative information, performance often seems satisfactory. However, a serious lack of conceptual understanding remains, as even the most intelligent students often cannot transfer the insights they should have gained to problems that differ from those dealt with in the classroom. What has gone wrong? Often teachers hold the 'direct transmission' view of learning (Staub and Stern, 2002), according to which successful classroom practice is seen as teachers' providing of information that students memorize and retell. Such learning environments, however, may, at best, help students accumulate facts or acquire simple skills but will not support them in building up the conceptual knowledge they need to model new and complex situations, as required in science and mathematics.

The main barrier that keeps students from learning science and mathematics is not so much what they lack, but what they

have – namely, naïve scientific knowledge that often works well in everyday life but largely differs from and even contradicts scientific explanations. Thus, to support students' learning, teachers must diagnose students' initial understanding of the content at hand. Knowledge not conforming to scientific views should not be dismissed as the sad result of deficiencies in previous instruction but, rather, be recognized as an inevitable step in learning. Effective teaching requires presenting students with questions and problems that stimulate processes of knowledge reorganization and thereby help them overcome their bounded or deficient beliefs. Thanks to concerted efforts in educational research, we nowadays know very well that it is the cognitively stimulating classroom environment created by the teachers which is crucial for meaningful learning at all achievement levels. In John Hattie's book *Visible Learning*, more than 800 meta-analyses dealing with factors that might influence school-related achievement were analyzed and synthesized. Obviously accessible factors like class size, methods of instruction, or use of computers were of negligible impact. What counts for student's meaningful learning, however, are teachers who are able to transform student's errors into learning opportunities by providing tasks and tools which make the difference in light of what the students already understand and misunderstand. By presenting their students with challenging but solvable tasks and problems, competent teachers support the construction of a knowledge base where abstract concepts, facts, and procedures are integrated and enable mutual activation.

Prior Knowledge: The Best Predictor of Learning Outcome but Not a Substitute for Intelligence

When entering new learning settings, students often differ from one another. Domain-specific knowledge and intelligence have been identified as the two major sources of difference. Research on cognitive development suggests that variations in prior domain-specific knowledge can often better account for achievement differences between younger and older children than general cognitive capabilities. Moreover, longitudinal studies, such as the Munich Longitudinal Study LOGIC, revealed that within-age level achievement, differences in core elementary school subjects are to a remarkable extent determined by domain-specific prior knowledge obtained in the preschool years. Early numerical competencies could account for achievement variations in mathematics after partialing out general intelligence, and early indicators of letter identification and phonological awareness predicted later performance in reading and writing (Schneider and Bullock, 2009). The reported results are important because they show that at least in complex knowledge domains a high IQ cannot compensate for a lack of prior knowledge, and moreover, that there is no direct connection between intelligence and achievement in content domains based on rich specific knowledge. However, regression analyses based on longitudinal studies reveal that the confounded variance of prior knowledge and intelligence predicts differences in learning outcome better than each single variable. If more intelligent children are placed in stimulating learning environments, they will acquire usable knowledge which will increase their lead.

There is, however, overwhelming evidence for the pivotal role of prior knowledge for further learning and advanced performance. Studies in different areas, among the mathematics, science, and chess, have revealed much better outcomes for persons with high prior knowledge levels (experts) and somewhat lower IQ than for persons with little prior knowledge (novices) and high IQ (Grabner et al., 2007). Note, it were novices and not laypersons who were considered in these studies. In contrast to laypersons novices possess the necessary domain-specific knowledge in terms of rules and core concepts but differ from experts in their lack of practice.

For individuals who grow up in a cognitively stimulating environment, prior knowledge and intelligence are to a certain extent inextricably linked with each other. Intelligence may guide the selection of learning environments and thereby determine the acquisition of prior knowledge. A person with a low IQ will hardly follow courses on theoretical physics even if he or she is credited with extra time. Moreover, intelligence may affect the number of content areas in which a person is able to acquire a profound amount of prior knowledge. This view has been clearly confirmed by the Study of Mathematically Precocious Youth, in which individuals were identified on the basis of very high reasoning abilities before the age of 13. Thirty-five years later, these individuals achieved occupational success comparable to that of individuals attending world-class mathematics, science, and engineering graduate training programs. A remarkable result was that the ratio of very successful people in the upper quarter of this highly selected sample (percentile >99) was higher than in the lower quarter (Lubinski and Benbow, 2006).

Intelligence and Learning: Educational Implications

Comparisons between schooled and unschooled groups reveal a strong effect of education on intelligence test scores even on nonverbal tests. Only by systematic education can individuals' intelligence emerge and approach an optimum. However, given that a basic level of education has been encountered, schooling in general and special training programs in particular increase intelligence only very modestly, if at all. IQ differences remain quite stable over time in groups who have been attending stimulating learning settings. Education highlights individual differences in intelligence rather than compensating for them. Broad variance in intelligence is a challenge for designing educational environments. The question arises whether learners of different intelligence levels gain more if they are assigned to different learning environments. On a first glance, it sounds plausible that less intelligent students gain more from structured than from open instruction, while more intelligent learners show the reverse pattern. This was the hypothesis when so-called aptitude–treatment interactions were investigated. However, most studies have failed to reveal interactions between intelligence and educational treatment, some of them probably because they lacked the statistical power necessary for revealing interaction effects (Hattie, 2009). In general, when assigning learners to different learning treatments according to their intelligence, one must remember that IQ follows the normal distribution. This means that 68% of the people in a population do not differ by more than one standard deviation

in either direction from the mean – they are quite similar. Therefore, assigning an unbiased group of learners to two different learning environments by median split of IQ scores is basically problematic.

While attempts to reveal aptitude–treatment interactions have not met with much success with intelligence as indicator for aptitude, it has been shown in many contexts that learners gain from different inputs depending on their prior knowledge. For instances, for areas in science and technology, it has been shown that visual–spatial representations do help novices in understanding complex situations while they are either needless or distracting for experts (Mayer, 2009). Similar results have been shown for including means of focused processing, as there are self-explanations or metacognitive questions. While they help novices to develop a deeper understanding of a subject matter by becoming aware of their misunderstandings and wrong conclusions (Atkinson and Renkl, 2007), experts would be unnecessarily detained by such aids. Such results are not surprising because when being presented with a problem, experts and novices differ in whether they already can fall back on established procedural and conceptual knowledge structures or whether they first have to build them up. Therefore there are better reasons to assign learners to different treatments based on their prior domain-specific knowledge than based on their intelligence. Given the relationship between intelligence and efficiency in learning and information processing, a higher IQ facilitates the exploitation of learning environments, leading to the acquisition of knowledge that is broad as well as deep enough to be prepared for mastering as yet unknown demands of the future. As a consequence, the knowledge gap between more and less intelligent students will increase in the course of time. More intelligent learners who have invested their intelligence in the construction of broad and deep knowledge will be prepared for entering demanding and abstract subjects which will remain closed for those who started under less advantageous conditions. Providing different educational tracks for learners with a vocational focus and for those who are qualified for an academic path is part of the educational system of most countries. There are, however, remarkable differences concerning the age of the students at which the separation starts. While this is around 15 in many countries, in Germany and Austria, for instance, the separation starts as early as 10 years. Meta-analyses, as they are reported by Hattie (2009), revealed almost no impact of ability grouping on student's achievement growth. It rather seems to be the case that professional teachers, who are aware of their student's thinking and knowing and who provide meaningful experience in light of this knowledge, are able to boost different student's potential.

See also: Cognitive Development: Mathematics Learning and Instruction; Developmental Behavioral Genetics and Education; Education for the Gifted and Talented; Instructional Psychology; Learning Theories and Educational Paradigms; Metacognitive Development: Educational Implications; Piaget's Theory of Human Development and Education; School Learning for Transfer.

Bibliography

Atkinson, R.K., Renkl, A., 2007. Interactive example-based learning environments: using interactive elements to encourage effective processing of worked examples. Educational Psychology Review 19, 375–386.

Deary, I., 2012. Intelligence. Annual Review of Psychology 63, 453–482.

Duckworth, A., Quinn, P., Tsukayama, E., 2011. What no child left behind leaves behind: the roles of IQ and self-control in predicting standardized achievement test scores and report card grades. Journal of Education Psychology. http://dx.doi.org/10.1037/a0026280. Advance online publication.

Grabner, R., Stern, E., Neubauer, A., 2007. Individual differences in chess expertise: a psychometric investigation. Acta Psychologic 124, 398–420.

Hattie, J., 2009. Visible Learning: A Synthesis of Over 800 Meta-analyses Relating to Achievement. Routledge Chapman & Hall.

Kaufman, Alan S., 2009. IQ Testing 101. Springer Publishing, New York.

Lubinski, D., Benbow, C., 2006. Study of mathematically precocious youth after 35 years: uncovering antecedents for the development of math–science expertise. Perspectives on Psychological Science 1, 316–345.

Mayer, R.E., 2009. Multimedia Learning, second ed. Cambridge University Press, New York.

Neubauer, A., Fink, A., 2009. Intelligence and neural efficiency. Neuroscience and Biobehavioral Review 33, 1004–1023.

Neubauer, A., Freudenthaler, H., 2005. Models of emotional intelligence. In: Schulze, R., Roberts, R.D. (Eds.), Emotional Intelligence: an International Handbook. Hogrefe, Göttingen, pp. 31–50.

Oberauer, K., Sü, H.-M., Wilhelm, O., Wittmann, W.W., 2008. Which working memory functions predict intelligence? Intelligence 36, 641–652.

Ridley, M., 2003. Nature via Nurture: Genes, Experience, and What Makes Us Human. Harper Collins, New York.

Schneider, W., Bullock, M. (Eds.), 2009. Human Development from Early Childhood to Early Adulthood. Lawrence Erlbaum, Mahwah, NJ.

Spelke, E.S., Kinzler, K.D., 2007. Core knowledge. Developmental Science 10, 89–96.

Staub, F., Stern, E., 2002. The nature of teachers' pedagogical content beliefs matters for students' achievement gains: quasi-experimental evidence from elementary mathematics. Journal of Educational Psychology 93, 144–155.

Stern, E., 2005. Knowledge restructuring as a powerful mechanism of cognitive development: how to lay an early foundation for conceptual understanding in formal domains. In: Tomlinson, P.D., Dockrell, J., Winne, P. (Eds.), Pedagogy – Teaching for Learning. British Psychological Society, Leicester, pp. 153–169.

Sternberg, R., Grigorenko, E., 2002. Dynamic Testing. Cambridge, New York.

Turkheimer, E., Haley, A., D'Onofrio, B., Waldron, M., Gottesman, I.I., 2003. Socioeconomic status modifies heritability of IQ in young children. Psychological Science 14 (S), 623–628.

Intensification and Specialization, Archaeology of

Kathleen D Morrison, University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Changes in the organization of production are associated with significant historical transitions such as the development of food production, settlement aggregation and sedentism, the institutionalization of social and other inequalities, occupational specialization, urbanization, exchange, the production of surpluses, and the development of markets. Understanding production in terms of its constituent processes has thus been critical. Two of these processes, the intensification and specialization of production, lie at the heart of many of these transitions and are used to characterize the level of effort, productivity, land and labor use, and degree of skill in agricultural production and craft or artifact production.

Many of the major historical transitions studied by archaeologists – the development of agriculture, urbanism, institutionalized inequality, markets, and processes of imperial incorporation, colonialism, and political collapse – consist of, or are accompanied by, significant changes in the organization of production. Studies of intensification and specialization build on archaeological approaches to production, and must be understood in light of this construct, which differs significantly from its use in economics. Both intensification (usually examined in terms of agriculture) and specialization (generally studied in terms of craft production) are often viewed as pivotal to the development of institutionalized inequality and complex polities, although there is increasing recognition that specialized production has at times existed even in minimally stratified societies. Beyond the potential significance of intensification and specialization for the development of social complexity, both processes are also implicated in equally fundamental changes in human demography, labor organization, the development of social and gender identities, and changes in regional landscapes and floral and faunal assemblages. Recent developments in the archaeology of intensification and specialization tend to stress process over cause, highlight temporal and spatial diversity, and explore the consequences of productive organization for differently situated social actors, rather than classify productive forms on an evolutionary continuum.

The Archaeology of Production

Archaeological concepts of production are typically situated in the anthropological perspective that productive activities are embedded in larger cultural and social systems. Production – the creating, constructing, and tending of artifacts, structures, human-modified plants and animals, even entire landscapes – is a fundamental focus of the study of the archaeological record. Production, as one dimension of human economy, is both responsive to and constitutive of other domains such as processing, distribution, and consumption. This perspective is a legacy of the formalist–substantivist debate in anthropology, which centered around whether nonindustrial economies could be understood in the same terms as industrial capitalism (the formalist position), or whether the economies of nonindustrial societies typically studied by anthropologists were, in Polanyi's (1957) words, more fundamentally 'embedded and enmeshed' in social institutions. However, this picture is complicated by the existence of branches of archaeological thinking, which draw on economistic approaches or on those approaches as modified by ecology and evolution.

Research has not focused evenly on all arenas of production. Although agriculture and animal husbandry constitute important areas of research, there is a longer history of emphasis on durable goods or artifacts, objects made or modified by humans. Artifact manufacture is often termed craft production. While changes in the organization of craft production are generally conceived in terms of specialization, food production is more often viewed in terms of intensification. Part of this divergence may relate to the different kinds of information available about each arena of production. Students of craft production typically work with artifacts themselves and with evidence from production locales such as kilns, workshops, and even dumps. Food production research does involve products such as seeds and animal bones, or objects used in food preparation such as ceramics and ground stone, but these are often conceived as objects of consumption, with production studies focused primarily on locales of production such as fields.

Intensification

Archaeological studies of intensification have focused primarily on agriculture. In some sense, the historical trajectory of agricultural change can be viewed as a long, although variable, history of productive intensification, concomitant with the rise in global population, its clustered distribution on the planet, and the entrenchment of economic and social inequality.

The Concept of Intensification

The difference between intensification and simple increase is analogous to the difference between concentration and amount. Intensification of production refers to attempts to increase the concentration of production, or productive output per unit of land, labor, or some other fixed quantity.

The focus on production effort in this definition highlights the salience of input, allowing failed efforts to increase productivity to be included in studies of intensification. In archaeology, the variable held constant almost always refers to land in reference to agriculture (getting more from a given area), and labor in reference to craft production (increasing efficiency of production). Situations in which both land and labor are held constant while capital inputs are increased, as in industrialized agriculture, are less often discussed. An important exception to this is the concept of 'landesque capital' intensification. First advanced by Blaikie and Brookfield (1987), landesque capital highlights the role of farmers in creating enduring landscape modifications such as terraces that allow for more intensive forms of production over the course of many human generations without a sustained increase in effort (Widgren, 2007).

History of Approaches to Intensification

The model of intensification set forth by economist Ester Boserup (1965) has been the most influential formulation of the problem, and was widely adopted in archaeology (e.g., Cohen, 1977). Boserup's model was parsimonious, general, and comprehensive; however, like other unilineal models of cultural evolution, it greatly generalizes and simplifies this complex process. Boserup's view of population pressure as an independent variable driving intensification inverted the earlier Malthusian formulation. While Malthus saw arable land as limiting increases in production, Boserup optimistically asserted that constantly increasing population (an assumption shared with Malthus) acted as a motor, driving technological changes in land use along an extensive–intensive continuum.

The most contested aspect of Boserup's model was the causal efficacy of population pressure, a view resting on several related assumptions. First, producers are assumed to exert the minimum effort required to meet their needs (the law of least effort). Thus, the most labor-extensive regime possible will always be employed. Second, Boserup argued that there are diminishing marginal returns to labor (declining efficiency) with increasingly intensive forms of production. The disadvantages of intensive agriculture – increased labor inputs and declining efficiency of that labor – ensure, for Boserup, that more intensive forms of production will be adopted only when strictly necessary. These assumptions have been criticized on empirical grounds, as has the prime-mover status of population pressure. There is now a vast literature considering such causal factors as population density and distribution; sedentism; market-driven, prestige-driven, or politically driven demands for produce; risk; climate change; and many others (reviewed by Morrison, 1994). Contemporary approaches to intensification tend to stress multiple and cascading causes, which may be locally contextual and historically contingent (e.g., Kirch, 1994).

Consequences of Intensification

In a proximate sense, the most common consequence of intensification is increased productivity. Equally significant, however, are less proximate consequences, such as the restructuring of human labor, possibilities for differentiation of nonfood production, and the support of nonproducers. It is in this sense that intensification has been seen as underwriting social and political complexity, by which archaeologists generally mean the existence of institutionalized inequality and/or the presence of complex polities such as states or chiefdoms. Historically, intensified food production has also had dramatic ecological consequences, leading to landscape modification on a large scale, the worldwide movement of plant and animal taxa, and support for artifactual landscapes such as cities (see papers in Thurston and Fisher, 2007). Indeed, the Industrial Revolution may be seen as a case of the intensification of specialized forms of production, a process that has had significant ecological as well as social and economic consequences.

Methodological Issues in Intensification Studies

The study of intensification poses serious methodological challenges. The investigation of past agriculture, in particular, is difficult because of its large spatial scale and ephemeral material remains. For this reason, archaeologists consider numerous indicators of land use, including agricultural implements, historical documents, settlement distributions, types and distributions of agricultural facilities, paleo diet indicators, and botanical remains. Identifying intensification in the archaeological record depends on the definition adopted; views that stress output require some measures of productive success, a difficult requirement for archaeology, where agricultural productivity can rarely be quantified. Output-based models are more popular in craft production studies, where such measures may be more reasonably inferred. Production effort is usually postulated on the basis of inferences about the overall form of production, using clues such as technological forms, products, or cultigens with special productive requirements, as well as knowledge of specific environmental conditions.

The Future of Intensification Studies

Three promising trends in the archaeology of intensification may be identified. The first is increased attention to the process of intensification itself, a process incorporating a great degree of variability. Kaiser and Voytek (1983), for example, divide intensification into three components: specialization, diversification, and intensification proper. These involve changes in the amount and organization of labor and its application through technology, and require measurement of multiple variables. The second trend is increased attention to actual historical trajectories of change and the variable paths intensification may follow (e.g., Morrison, 1996). Spatial diversity in productive forms is also increasingly highlighted, in contrast to Boserup's dismissal of contemporaneous diversity as representing merely incomplete steps in a progressive development. Contributors to Widgren and Sutton (2004), for example, highlight the existence of 'islands of intensive agriculture' in East Africa, highly localized regions in which intensive forms of production have been adopted even in the absence of cities, markets, or other conventional drivers and which contrast with larger surrounding areas characterized by more extensive forms of agriculture. Finally, in accord with moves toward a more inclusive view of production, many studies now weigh

cultural meanings of food products alongside concerns of labor efficiency or energetics. This emphasis on meaning is significant in that is also tends to displace causal assumptions predicated on biological necessity.

Specialization

The origins and development of specialized production have always been major concerns in archaeology. If intensified food production is often seen as undergirding social complexity, then specialization has been regarded as its index. Students of specialization draw on Marxian traditions, which integrate concern for the means and social relations of production with anthropological attention to culture. This research tradition, however, draws equally from Adam Smith in its attention to the division of labor and focus on exchange, as well as its often embedded assumptions about the nature of economic behavior. In their concern for social and political implications of specialization, archaeologists have also drawn on Durkheimian notions of solidarity and integration.

The Concept of Specialization

Specialization may be defined as the channeling of resources, skills, and/or labor into restricted ends, a definition focusing on both the products and process of specialization. Other definitions stress the role of specialization in setting apart people, technologies, and production locales, both physically and socioeconomically. Most literature on specialization in archaeology focuses on craft production; such studies have been dominated by a concern for technological process, interest in the organization of production, and, more recently, an expanded interest in the social relations of production, including issues such as identity and meaning. Specialization implies exchange on some level and as such, is not intelligible outside larger political/economic contexts.

History of Approaches to Specialization

Most early studies of artifacts emphasized typology, using morphology and manufacturing technology to classify objects, creating the chronological and spatial systematics that still underlie disciplinary practice. Despite pioneering studies such those of V. Gordon Childe that stressed the social and political roles of craft specialists in early stratified societies, most observations on artifact variability were oriented toward defining the cultural/temporal units that were themselves the goals of research. With moves toward functionalism in the 1950s and 1960s, however, technologies of production commanded greater attention as areas of study in their own right; this focus on technology continues into recent materials science approaches.

With the 1960s and 1970s came concerted efforts to delineate the organization of production and especially the integration of specialized producers into larger political economies. Growing from this interest, much effort has been spent investigating the scale (household production, workshops, and factories) and tempo (full- vs part-time specialists) of specialized production and the material correlates of organizational forms. Also of concern are forms of control over production, strategies of labor and resource mobilization, and the disposition of manufactured goods. Several classifications have attempted to capture some of these contextual dimensions of production, distribution, and consumption (Costin, 1991; Costin and Hagstrum, 1995), even as others have pointed to the potential for significant organizational diversity and fluidity in production, making such schemes difficult to apply (e.g., Sinopoli, 2003). Studies of exchange have been particularly important, congruent with a focus on regional economies and with the development of new analytical methods for determining source areas of raw materials.

Specialization studies have been marked by a concern for the initial causes of specialization as well as its intensification, although here explanations often stress socially based demands and especially the development of prestige goods, perhaps because many artifacts are clearly not biologically necessary for survival. More recently, archaeologists have begun to stress processes of production, especially agent-oriented *chaîne opératoire* approaches, which focus on structured chains of culturally inflected decisions made by producers (Lemmonier, 1986) and on interrelations between technology and meaning. A new focus on producers, too, has followed from concern for more finely divided aspects of producer identity, including gender and ethnicity and from recognition that multiple organizational forms of production often coexist, even where the producers are the same.

Consequences of Specialization

From the time of Childe, archaeologists have explicitly linked specialized production with complex political forms and social inequality (Brumfiel and Earle, 1987). The very existence of elites is typically inferred, in part, from sumptuary objects identified by their elaboration, restriction, and sometimes exotic provenance. Curiously, specialization in craft production (sometimes cast as intensification) is generally viewed as promoting efficiency, while intensified agriculture is depicted as producing declining marginal returns. Recent work on the meaning of production, especially its ritual context, draws more fully from an anthropological tradition de-emphasizing efficiency.

Methodological Issues in Specialization Studies

Recognizing specialized production begins with the objects and by-products of manufacturing and their distribution. The definition of specialization allows for certain ambiguities, as issues of scale and context impinge. For example, the production of artifacts in one region and exchange to another may be considered regional specialization, although artifacts were manufactured by every household. Locations such as quarries are specialized places without necessary associations with producer specialization. Although the goal of archaeological research is usually to identify producer specialization, because archaeological data are fundamentally spatial, determining the organization of production is inevitably inferential. Much of the methodological development of specialization studies grew around this dilemma. Complex technological requirements or high levels of producer skill are sometimes assumed to index

specialist manufacture; another criterion advanced for identifying specialization and its scale is standardization, mass production being thought to promote standardization (for reasons of efficiency). The possible link between specialization and standardization, especially in ceramic production, has long been under active discussion (e.g., Rice, 1991; Roux, 2003). Although ethnographic and experimental research has been important in studies of specialization since the 1960s, more recently such work has expanded beyond technical guidance to address broader concerns such as the reasons for community-wide specialization (Stark, 1991), and issues of identity, gender, power, and place. Both ethnoarchaeology and experimental archaeology now have dedicated academic journals of their own.

The Future of Specialization Studies

The trends identified in intensification studies also hold for studies of specialization. Concerns for process are implicated in perspectives that draw specialized production into the larger field of the anthropology of technology; this concern is sometimes expressed in the language of 'crafting' and the ways in which crafting both draws from and creates social relationships (Costin and Wright, 1998). Recent developments in the archaeology of both intensification and specialization tend to stress process over cause and explore the consequences of productive organization for differently situated social actors rather than classify productive forms along an evolutionary continuum.

See also: Boserup, Ester (1910–99); Ceramics in Archaeology; Differentiation: Social; Labor, Division of; States and Civilizations, Archaeology of; Trade and Exchange, Archaeology of.

Bibliography

Blaikie, P.M., Brookfield, H.C., 1987. Land Degradation and Society. Methuen, London.
Boserup, E., 1965. The Conditions of Agricultural Growth. Aldine, Chicago.
Brumfiel, E., Earle, T.K. (Eds.), 1987. Specialization, Exchange, and Complex Societies. Cambridge University Press, Cambridge, UK.
Cohen, M.N., 1977. The Food Crisis in Prehistory. Yale, New Haven, CT.
Costin, C.L., 1991. Craft specialization: issues in defining, documenting, and explaining the organization of production. In: Schiffer, M. (Ed.), Archaeological Method and Theory, vol. 3. University of Arizona Press, Tucson, AZ.
Costin, C.L., Hagstrum, M.B., 1995. Standardization, labor investment, skill, and the organization of ceramic production in late Prehispanic Highland Peru. American Antiquity 60 (4), 619–639.
Costin, C.L., Wright, R.P. (Eds.), 1998. Craft and social identity. In: Archaeological Papers of the American Anthropological Association, vol. 8. American Anthropological Association, Arlington, VA.
Kaiser, T., Voytek, B., 1983. Sedentism and economic change in the Balkan Neolithic. Journal of Anthropological Research 2, 323–353.
Kirch, P.V., 1994. The Wet and the Dry: Irrigation and Agricultural Intensification in Polynesia. University of Chicago Press, Chicago.
Lemmonier, P., 1986. The study of material culture today: toward an anthropology of technological systems. Journal of Anthropological Archaeology 5, 147–186.
Malthus, T.R. Essay on the Principle of Population, as it Effects the Future Improvement of Society. Johnson, London.
Morrison, K.D., 1994. The intensification of production: archaeological approaches. Journal of Archaeological Method and Theory 1, 111–159.
Morrison, K.D., 1996. Typological schemes and agricultural change: beyond Boserup in South India. Current Anthropology 37 (4), 583–608.
Polanyi, K., 1957. The economy as an instituted process. In: Polanyi, K., Arensberg, C.M., Pearson, H.W. (Eds.), Trade and Market in the Early Empires: Economies in History and Theory. Free Press, Glencoe, NY.
Rice, P.M., 1991. Specialization, standardization and diversity: a retrospective. In: Ceramic Legacy of Anna O. Shepard. University Press of Colorado, Niwot, pp. 257–279.
Roux, V., 2003. Ceramic standardization and intensity of production: quantifying degrees of specialization. American Antiquity 68 (4), 768–782.
Sinopoli, C.M., 2003. The Political Economy of Craft Production: Crafting Empire in South India, c.1350–1650. Cambridge University Press, Cambridge.
Stark, M.T., 1991. Ceramic production and community specialization: a Kalinga ethnoarchaeological study. World Archaeology 23 (1), 64–78.
Thurston, T.L., Fisher, C.T. (Eds.), 2007. Seeking a Richer Harvest: the Archaeology of Subsistence Intensification, Innovation, and Change. Springer, New York.
Widgren, M., Sutton, J.E.G. (Eds.), 2004. Islands of Intensive Agriculture in Eastern Africa. Ohio University Press, Columbus.
Widgren, M., 2007. Precolonial landesque capital: a global perspective. In: Hornborg, A., McNeill, J.R., Alier, J.M. (Eds.), Rethinking Environmental History: World-System History and Global Environmental Change. Altamira, Walnut Creek, pp. 61–77.

Intentionality and Rationality

David Braddon-Mitchell, University of Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article discusses the central features of mental systems: intentionality and rationality. It considers rationality as the characteristic excellence of intentional states, most especially beliefs and desires, and explores the differences and similarities between conceptions of rationality in these two cases, as well as in the case of actions which arise from beliefs and desires.

Intentionality and rationality are the two essential features which, perhaps, are characteristics of a mind. Intentionality is that directedness which allows thoughts to be about other things, even about the world. Because the mind is an intentional system it can represent how things are. Rationality is the characteristic excellence of intentional states, both those that are about how it is (beliefs) and those about how one wants it to be (desires). The present article discusses the nature of intentionality, the question of what the objects of thought actually are, what relationship if any it has with consciousness, and some modern naturalistic (broadly scientific) accounts of what makes something intentional. It then considers the characteristic excellences of the intentional states of belief, desire, and of the events they result in, actions. The article concludes with a brief discussion of whether we in fact are rational, and of how to justify the normative claim that we ought to be rational.

Intentionality

Intentionality is the term given to the mind's striking capacity to have states which are about features in the world. In addition to weighing 3 g, or being composed of ectoplasmic goo, or conforming to a certain neural networking diagram, a mental state might have the property of being *about* something: perhaps *panna cotta* flavored gelato. This property – this 'aboutness' – is an intentional one. It is sometimes thought of as a kind of directedness. The mind is directed on things; your fears are directed on their objects; your beliefs likewise.

The term itself is derived from a Medieval Latin translation of an Arabic technical term. The medieval discussion was about a rather different topic than the contemporary issues that are discussed under the rubric of intentionality – perhaps slightly closer to the contemporary debate about the nature of concepts (Sorabji, 1991) (q.v. Concepts).

The World or Intermediate Objects?

A straightforward and appealing account of intentionality is to claim that it consists in some kind of relation between a mental state and the object it represents. This claim is neutral between the view that the relation is inexplicable by the natural sciences, and reductive naturalistic programs that try to explain just what scientifically mandated properties this relation is constituted by.

There are difficulties: presumably it is possible to have beliefs about things that do not exist. In some cases this can be finessed away – my belief about the gelato in the fridge, should there turn out to be no gelato in the fridge, might be redescribed as a (false) belief about the fusion of all gelato, that part of it is instantiated in the fridge. But such finessing is pointless, because there are *types* of things that do not exist, as when I have beliefs about Bunyips. So if my (false) belief about the Bunyips does not gain its intentionality in virtue of a relation between a mental state and Bunyips, what does the relation hold between?

One solution is that it holds between a mental state and a special, intermediate, intentional object. The intentional object might be there even if the mental state does not succeed in directing itself onto the world. This requires that there are such intentional objects to do the representational work. Some philosophers, including Meinong and latterly Routley and Zalta (Zalta, 1988), hold variations on the doctrine that there are such entities, but they do not have full fledged existence: the nonexistent (sometimes called Meinongian) objects have some other kind of being which is enough to do the work.

Some proponents of the Meinongian view can allow that intentionality is always relational, but that in the case of existing objects the relation is between the mental state and the existing object, and in the case of the nonexistent objects it is between the mental state and the nonexistent object. This requires them to say (in the case of a fear of a Yeti) that the nonexistent Yeti is the very same object as the thing you would be fearing if there really were Yeti.

Perhaps in favor of indirect objects is that it seems that the contents of our beliefs, desires, and so on are more finely discriminated than the things in the world. This can be noticed in the so-called referentially opaque conditions (a kind of intensional context: (q.v. Intensionality)). If Bill desires a Rosa brand Gelato, and if it so happens that Rosa brand is the worst in Sydney, most do not think that it follows straightforwardly that Bill desires the worst gelato in Sydney, even though the gelato he desires is (unbeknownst to him) the worst in Sydney. However, if some indirect intentional object is taken to be the content, then there might be two distinct intentional objects, only one of which is the object of Bill's desire, but both of which are themselves directed. One of the disturbing things about this approach is that it does not seem to solve the problem. How do the special intentional

objects do the work of ensuring that the states that are in relations to them get to be about things in the world? If it is a relation to the object in the world, then there is a regress problem. If not and there is just a mysterious power, then the mysterious power may as well be possessed by the mental state directly.

There is also difficulty for naturalists in seeing how mental states can bear relations to mysterious indirect objects. Finally, a more basic complaint is that even where there are such indirect entities, it seems that our thoughts are surely at least sometimes directly about the things in the world: it is the gelato itself Bill desires, not some intermediate object.

Intentionality and Consciousness

Some have thought that intentionality is a special feature that marks off mental phenomena from nonmental phenomena: all and only mental systems display intentionality. Something like this was intended by Brentano in the slogan claiming that intentionality is the 'mark of the mental.' This is not meant to be a nomenclatural claim which merely says that, by definition, if a system is too complex enough to be intentional then it is complex enough to be mental. Instead, it seems to proceed by some connection between consciousness and intentionality. Thus, John Searle (Searle, 1983), for example, thinks, that a physical system which could do whatever reductionists claim is enough for representation (see the following section) could none the less lack genuine intentionality, because it would not be *about* the thing it covaried with. The reasoning is that either such systems might not have what it takes for consciousness, and thus lack intentionality, or else the reverse: that since they might lack genuine intentionality, they might lack consciousness.

Reductive Approaches to Intentionality

Most common among contemporary approaches to intentionality are reductive attempts to say what scientifically respectable property of our minds it is that accounts for the difference between intentionality and nonintentionality.

The first of these approaches is to deny that the phenomenon of intentionality exists. So-called eliminative materialists (q.v. Eliminative Materialism) (e.g., Churchland, 1981) agree that we can specify scientifically respectable properties which, if they were instantiated, would guarantee that there is intentionality. But they do not think such properties are instantiated, and thus think science will discover there is no intentionality. Since their arguments depend on stringent principles for the existence of intentional entities, there are two ways for noneliminativist reductionists to argue against them: by accepting the standards and making the empirical claim that there are things of this nature, or by rejecting the standards.

Amongst those who think there is in fact intentionality in the natural world, there are, very broadly, four basic approaches in the contemporary literature: informational semantics, teleobiological theories, broadly functionalist theories, and instrumentalist theories.

Informational semantics is the view that intentionality occurs in the case of the mind only when there is covariation in between mental states and things in the world. Caveats about complexity, interrelatedness or causal, and logical dependence are often added to upgrade this condition to sufficiency. Versions of these theories can be found famously in pieces by Fred Dretske (1983) and Jerry Fodor (1990) (q.v. Content, Informational Semantics).

Objections to the view are legion, the most troubling being variations of the so-called misrepresentation problem. The problem is that in situations where there is obvious error, the theory denies that there is error. For suppose that we are trying to have a mental state about cats. It covaries with cats (i.e., it appears exactly when the agent is in the presence of cats), but occasionally – say about 1% of the time – it misfires in the dark and the state appears in the mind when a large rat is seen in the darkness. For simplicity, suppose nothing else causes this cat symbol in the head. Intuitively, it seems we should say the symbol is about cats, and occasionally we deploy that symbol to say false things of cats – as when we think that there is a cat there, but unbeknownst to us it is in fact a rat which has misled us. Unfortunately, the theory notes that there is perfect covariation: between the mental symbol on the one hand, and cats *or* rats. Thus it says the symbol is about cats or rats. And so thus we have in every case a *true* belief about of cats or rats – that they are present. Various solutions to this problem can be found in the works cited immediately above, but none uncontroversially works.

Another objection is that no purely correlational fact could account for the thought that intentional states do not just reliably indicate the state of the world, but rather such indication is what they are normatively *for*. Teleo-semantics purports to find a solution to this problem. This is the cluster of views around the work of Ruth Millikin (1984) and Karen Neander (1991), according to which evolutionary biology plays a central role. The leading idea is that states are genuinely representational because something like covariation is the *proper biological function* of the states (or is produced by states whose function is producing states of this sort). This in turn is cashed out evolutionarily: a state's proper biological function is whatever it did that caused it to be selected by natural selection. Thus a heart's function is to pump blood but not to make a noise, because while hearts have always made noises, these noises did not contribute to the fitness of those organisms with hearts, whereas pumping blood did. So, in the case of mental states, while there may be many mental states which correlate with external features of the world, only those that do so because it was advantageous during evolutionary history get to count as intentional states. Objections are of two kinds. The first kind is that the states discriminated in such ways may not turn out to be about even remotely the sorts of things that we imagine our intentional states pick out. The second kind involves worries about the historical individuation of mental states. If Creationism is true, for example, states would not be intentional – because they would have no evolutionary history. Alternately, if Creationism is true then intentionality is an entirely different thing from what it would be on the assumption that Creationism is false. This would raise questions about what is in

common between what is called 'intentionality' on the Creationist assumption, and what is called 'intentionality' on the standard assumption, that would make it right to give them the same name.

Functionalist theories say that intentionality is a phenomenon which arises when systems are complicated enough to interact with the environment in a complicated way. So what matters is not just how the things in the world affect the presence or absence of the state, but how the state interacts with other states to affect the world. A state gets to be a belief that ice-cream is around, and thus a belief, and thus intentional, not only because it is an indicator of ice-cream, but also because it potentially interacts with the world. Combined with desires for ice-cream it may cause ice-cream directed behavior. In most functionalist accounts, there are few constraints on what the internal nature of a mind has to be like to count as intentional, so long as that nature will cause the right interaction with the world (see Armstrong, 1968; Braddon-Mitchell and Jackson, 1996; Lewis, 1972).

The final major category is instrumentalism (q.v. Instrumentalism). An instrumentalist about intentionality sometimes thinks that there is no intentionality as such, but it is pragmatically useful to imagine there is, because it allows us to engage in intentional explanations of behavior, which are both quicker and more practical than the full neurophysical story. Strictly speaking, such views are versions of eliminativism which do not seek to eliminate *talk* of the properties they hold not to exist. Other theorists who are sometimes called instrumentalists hold that there are indeed intentional states, but that all it takes for a system to have intentional states is to produce effects in the world best predicted or explained by intentional talk (see e.g., Dennett, 1991) On some construals this kind of instrumentalism is close to functionalism, since like the functionalist the view is realist about intentional states, and like the functionalist it places few constraints on the intrinsic nature of the states insofar as that nature is not relevant to their causal or explanatory profile. The differences center around whether talk of states apt for prediction and explanation is a distinct way of typing the states from the functionalist's way of typing causally. The answers to these questions about the differences between functionalism and realist instrumentalism will depend in part on views about the relationships between causation and explanation.

Rationality

Introduction

If anything is common to various accounts of rationality, it is that rationality is a special excellence of the mind. Perhaps, it is to have one's higher mental processes functioning well. It must be narrowed down by excluding certain kinds of functioning – well functioning just in terms of crude computational power can be excluded. More controversially moral well functioning is often excluded (q.v. Ethical Rationalism).

It can be best thought of in terms of criticism: one can criticize various mental process as irrational. Many think that one can criticize as irrational someone who fails to draw obvious deductive inferences, one can criticize someone as irrational who has strange and bizarre patterns of change in desire (though this is controversial as we shall see later), and one can criticize as irrational someone who fails to act so as to bring about what he/she desires in the light of his/her beliefs. On the other hand it is more controversial whether one can criticize as irrational behavior or thought on the grounds that is immoral and virtually no one thinks that not being able to perform Herculean tasks of mental arithmetic is enough to count as irrationality.

This article will cover three domains: belief, desire, and practical reason, and the question of how much there is in common between what is called 'rationality' in these domains. There is also a question as to whether rationality is a feature of emotions: is it rationally required or prohibited to be in various emotional states in various circumstances (q.v. Emotions).

Rationality with respect to these domains has characteristic goals. The rational believer aims at truth, the rational desirer aims at best desires (perhaps here is a connexion with morality), and the rational actor aims to act in a way so as to satisfy his/her desire in the light of his/her beliefs. An unresolved issue is whether rationality is characterized as whatever processes bring about these goals (q.v. Externalism, Epistemology), or whether rationality can be characterized by the nature of the processes themselves, and we are lucky enough to live in a world where these processes do in fact bring about the goals on average better than other processes (q.v. Internalism, Epistemology).

Rationality of Belief

In the domain of beliefs, agents are held to be irrational if they form or revise beliefs in the wrong way. At least three kinds of belief-forming mechanisms are at issue: the source of initial beliefs, the updating of beliefs on the basis of evidence, and the deduction of the implications of these beliefs in the absence of further evidence.

There is no clear consensus on how initial beliefs should be sourced. Bayesian probability theory holds that there is no way of assessing the rightness of the initial probabilities that agents give to various claims about the world. Instead, the method for revising beliefs in the light of evidence is something that the theory specifies. So on this tradition, an agent's beliefs are rational just if they are revised in the light of evidence the way the theory specifies (see, e.g., Jeffrey, 1983, or for an accessible introduction to methods Lee, 2004 (q.v. Bayesianism)). There are competing systems for updating beliefs in the light of evidence (q.v. Inductive Logic, Belief Revision).

Some philosophers, however, grant special epistemological status to some initial beliefs (or probability assignments) depending on where they come from. Initial beliefs might be held to be held rationally insofar as they come from approved sources. These include the testimony of others (q.v. Testimony) or claims about sense impressions (q.v. Logical Positivism, Sense Data, Perception).

The rational constraints on drawing of inferences from existing beliefs unaided by further evidence are the domains of the various branches of deductive logic (q.v. Deductive Logic).

Of course, there is a richer sense of rationality in this area: the capacity to draw the *important* conclusions and ignore the irrelevant ones. Artificial intelligence researchers have long seen how irrational seeming a system will be that just sits and cranks out all the irrelevant consequences of what it believes (q.v. Frame Problem, Artificial Intelligence).

Rationality of Desire

Whilst most philosophers would allow that beliefs, or at least the process of change of belief, can be legitimately subject to criticism on grounds of rationality, there is less consensus in the case of desires. Hume's famous dictum that for all reason cares, he could prefer the scratching of his little finger to the destruction of an illustration of his idea that foundational desires are not subject to constraints of reason. An agent simply has the basic desires he/she has; the job of reason is to modify beliefs with the goal of truth, which in turn serves in the task of satisfying desires. A way of understanding the motivation for this is through considering the *point* of reason. If you think that reason has goals such as truth or the satisfaction of desire, then it is easy to see how belief formation can be irrational – it fails to track the truth – and how action can be irrational – it fails to satisfy desire in the light of belief. It is harder to see what goals basic desires have – though perhaps ease of satisfaction might accord with one of these desiderata, in which case a kind of quasi-Buddhist constraint might recommend itself, of desire minimization. It is here also that connexions between rationality and ethics might be drawn, since one constraint might be that the goal of desire is to motivate oneself toward the good life, and rational desires are those which do this (q.v. Rationalism, Ethics).

But even if there is no constraint on basic desires taken singly, there might be some constraints on combinations. In the same way that one can rationally criticize someone for inconsistent beliefs, on the grounds that not all can be true, one might criticize someone for incoherent desires, on the grounds that not all can be satisfied. But this will be a much messier business, for it is much less clear that there is anything wrong with a desire set not all of which can be satisfied. For contingently it is never the case that all our desires are satisfied, and it is hard to see why it is worse if this is elevated to a logical truth in the case where they cannot all be satisfied because two desires are logically exclusive. Indeed one can imagine that there might be some inconsistent desire sets which lead to greater flourishing than some consistent desire sets – in which case this constraint is at odds with the ethical one above. But supposing that it is a desideratum to have consistent desires, is 'rational' the right word to describe such sets? This might be thought to be an essentially terminological issue. Certainly the situation is isomorphic with the case of consistent beliefs, but whether the differences are small enough to justify using the same term may be an unsettleable piece of semantics.

There is a further question about desire evolution. How should desires change over time? Most grant that along with changes in beliefs, there are rationally mandated changes in instrumental desires. If I desire to be cured of disease, and I learn that a certain berry contains the cure, then I have a reason to acquire a desire to eat the berries in the service of my desire to be. But what of desire change over time under the impact of experience which is not grounded instrumentally? Some think that greater general consistency across time and other coherence factors are rational constraints. On this view, someone whose desires change all the time in apparently senseless ways, wishing for world peace and bananas 1 day, and world war and fruitlessness the next, is irrational. Certainly such a person in practice would have inconsistent desires (most desire happiness, and fluctuating desires is a sure-fire killer of that state). In these cases the view that costatisfiability is a requirement would be enough to claim that such a person is irrational. But even this seems like a questionable rational constraint. Inconsistent beliefs cannot both be true, so we know at least one of these beliefs is not true, and so at least one does not achieve the telos of belief. Certainly, the state of having desires which cannot be cosatisfied means that you cannot achieve the goal of having all your desires satisfied. But even that is not obviously a requirement of rationality. One might think that we ought to desire things which are, sadly, not cosatisfiable, if each of these is a proper object of desire. Perhaps we should desire both happiness and achievement. Perhaps it turns out that you cannot maximize both. Unfortunate; but not obvious evidence that you are irrational. And even if it is an irrational state, it does not follow (as it does in the belief case) that one these desires must itself not achieve the telos of desire. Only that having both does not. But what if the wildly fluctuating desires were always synchronically consistent? This would still be a very bad way to be, though the current writer thinks that it may be a borderline terminological matter whether there is enough in common between this malady and others to call it irrational rather than unfortunate.

What the question really boils down to is whether there is a theory of rational desire revision which is an analogue of the principle of rational belief revision (see Smith, 1994; Bradley, 2007).

Such a theory would require that a range of principles is used as principles of revision, such as the coherence principles discussed above. The problem, as we have seen, is in justifying any list of principles of desire revision that seem intuitively rational and showing that they are rational. In the case of principles of belief revision the justificatory role is played in showing that the principles are likely to lead to true beliefs. It is very hard to see what the analogue would be in the case of desires, barring satisfaction as the end of desire revision (sympathetic though this would be a quasi-Buddhist account of desire revision: remove the desires, so that there are very few left, and can all be easily satisfied).

One interesting approach is found in Smith (1994) and the following literature. Here the idea is that the rational principles of desire revision are the ones which lead to desiring the good, and it is success in this goal which justifies the intuitive list. The account also has the good as that which would be desired ideally, meaning that which we all would desire when revising our desires by the rational method of desire revision. The challenge for this account is persuading us that this is a nonviscous circularity, since the rational principles are understood as ones which lead to states which are themselves defined as states which are generated via the rational principles.

Rationality of Action

Rationality of action is most commonly called practical rationality. It is the kind of rationality which concerns the relations between one's beliefs and desires on the one hand, and one's actions on the other. If you desire ice-cream, if the ice-cream is in front of you, and if there are no conflicting desires or other impediments to eating it, then it is a principle of practical rationality that you should pick it up and lick. The noneater is practically irrational.

Sometimes practical rationality is characterized as reasoning *about* action. Some interpreters of Aristotle and Kant see their work in this way. Similarly, contemporary work on decision theory and game theory is often understood as what conclusions about what one *ought* to do should be drawn from premises about beliefs and desires. Another way of understanding practical rationality is the relationship between action itself and belief and desire. The difference is just that on the first way of understanding it one is practically irrational if one holds wrong *opinions* about how one should act given one's beliefs and desires, and thus it is special case of rationality about belief or theoretical reason. On the other way of understanding it, one is practically irrational if one actually acts in the wrong way given a set of beliefs and desires.

Much modern work on practical reason has focused on various calculi which either describe or prescribe how to act given one's beliefs and desires. Decision theory (see, e.g., Skyrms, 1986) (q.v. Decision Theory) is a family of theories which, on the assumption that features of the world relevant to one's decisions are themselves unaffected by those decisions, aims to give a precise account of how to choose. Game theory (q.v. Game Theory) is the calculus which has been developed to compare actions in situations where the situation itself depends on one's actions, as in many games, where one's own moves may determine the moves of others, and thus affect one's assumptions about how the world is that are relevant to choice (see, e.g., Elster, 1979).

One central philosophical worry with these theories stems from an account of what desires are. Decision and Game theory both aim, it seems, to prescribe how we should act. Thus, we ought to be able to rationally criticize someone who has acted irrationally and flouted the favored theory. Yet the most popular account of desires (or preferences as they are usually called in this literature) is the so-called revealed preference theory (q.v. Desires, Revealed Preference Theory of). According to revealed preference theory, one's desires are ascribed according to one's actions. In other words, you see what someone does, you make some assumptions about what they believe on the assumption that they are rational with respect to belief formation, and then assume that their desires are whatever it would take to produce the action. But then you can never criticize anyone's actions, for you can always attribute desires, however strange, that make sense of actions. According to revealed preference theory these are the desires that the person had.

In its most extreme version it is hard to use the theory to make recommendations about action as well. For it is hard to say what one's desires are until they are revealed in action and on strong versions they are *constituted* by the actions (among other things) so there is no fact of the matter.

So if the theories are to have prescriptive content, then we need a substantive theory of desire internal to the workings of the mind: a theory that is not tied so closely to what one actually does.

One phenomenon about which there has been a lot of philosophical discussion ever since Aristotle, and which is connected to these issues, is the so-called weakness of the will. Sometimes it seems that we act against our own desires. You might overwhelmingly desire to loose weight, resolve to avoid ice-cream, but eat it nonetheless. It might seem impossible to act against your desires if desire is constituted by actions, for there is a desire posited to explain any action. Theories of weakness of will divide between those that accept this, and posit sudden changes of desire or hidden desires to explain the behavior, and those that posit a nondesire-based mechanism that makes one sometimes behave against the all things considered desires one has at the time of the weak-willed act (q.v. Weakness of Will).

Much recent discussion of the rationality of action in decision theory is concerned with the conflict in special cases between two principles, dominance and maximizing expected utility, which intuitively seem both correct and unlikely to give clashing results (where by clashing results include cases where one principle makes no recommendations). Dominance reasoning is reasoning that proceeds by dividing the case into exhaustive possible scenarios, and showing that on each a certain action performs better, in contrast to expected utility calculations which multiply the utility of the action in each scenario by the probability of that scenario and sum these. There have for a long time been cases (q.v. Newcombe Problem) where these seem to come apart, but in cases where the implausibility of the setup has perhaps caused them to be unfairly neglected. More recently there have been shown to be well-defined games where expected utility gives no recommendation, but dominance does (see, e.g., Colyvan, 2006).

Are We Rational?

Recently, there has been much focus on the question of whether, as a matter of fact, we are rational or not. Much of this has focused on theoretical reason, or rationality about belief. Tests show that when examples of simple reasoning are given to even quite intelligent people, large proportions draw wrong inferences. Even simple modus ponens can be got wrong. Elementary statistical reason goes badly wrong, though how badly wrong it goes depends on how the information that subjects are expected to use to draw their conclusions is presented (see, e.g., Stich, 1990: Chapter 1).

There has sometimes seemed to be a debate about whether this view that irrationality is omnipresent conflicts with the views of some evolutionary psychologists who see the brain as a marvelously well engineered machine with many modules for making calculations about the environments in which it evolved. But there is surely no conflict. For what is marvelously engineered for the environments we evolved in need not be so great for our current one, and modules that calculate quickly and efficiently in demanding real-time environments are unlikely to use the same algorithms as would be best for giving considered opinions when time is not an issue (see Stich and Samuels, 2000; Kahneman et al., 1998).

Ought We be Rational?

What the question of whether we are rational or not does raise, though, is the issue of whether we ought to be. For if humans are in fact to a substantial extent not theoretically rational, then the claims that a theory of rationality has to be purely descriptive cannot be upheld. For it would be a false description. So perhaps this should be thought of as a purely normative, a theory of how we ought to think and act. And the question is why? If we lack of a foundational story about why one ought to adhere to some norms of thought, then either a skepticism about rationality or a relativism about rationality – the view that there are many sets of norms competing in the same domains, and there is no fact of the matter about which generally ought to be adopted – become attractive. The latter view has acquired some prominence in recent years (q.v. Relativism).

A leading problem for foundational accounts of the sources of the requirement to be rational is a set of considerations that arise from Carroll's Paradox (q.v. Carroll's Paradox) (Carroll, 1895). In its original form it's formulated with respect to modus ponens, the principle of belief revision which says that if you believe P and you believe $P \supset Q$, you should believe Q. Let us write the thing I might believe if I believe modus ponens as $P, P \supset Q \vdash Q$.

So now suppose I believe the following:

$$P,$$
$$P \supset Q$$
$$P, P \supset Q \vdash Q$$

I believe both P and that P implies Q, and I believe the principle that from P and $P \supset Q$, Q can be inferred.

Why should I conclude Q? It seems I could just believe all these things and not use my belief in modus ponens to infer Q. Well, the thought runs, perhaps there is another principle I need to believe. This is the principle that if you believe P, and you believe $P \supset Q$, and you believe modus ponens, you should infer Q. Let us write this as $P, P \supset Q, (P, P \supset Q \vdash Q) \vdash Q$.

But now it seems I could add that to my belief set and still not be obliged to use it to infer Q. The thought that we might add yet another principle – $P, P \supset Q, (P, P \supset Q \vdash Q), (P, P \supset Q, (P, P \supset Q \vdash Q) \vdash Q) \vdash Q$ – to our belief set leads us quickly to see that we are in the grip of a regress.

An obvious response is that I might infer Q (or perform some action in virtue of some principle of practical reason in the analogue of Carrol's Paradox for practical or normative reasons) just because I am disposed to reason with modus ponens. It's not a matter of more beliefs, it is about our dispositions. But then the paradox can be used to undermine the attempt to justify those dispositions. Perhaps I do reason with modus ponens. But why am I justified in so doing? Perhaps because I believe that I am justified in reasoning using it just if I believe it (or if it true, depending on the kind of justification in question). But what justifies that? Perhaps … and the regress reemerges.

One answer to the question of why we should adopt a set of rational norms is found in the pragmatic school: it will be somehow good for use or useful. But this is perilously subject to empirical refutation. For the evolutionary theories seem to suggest that aiming to achieve true beliefs at any cost is in fact not good for us. Perhaps quick and dirty solutions are more life preserving, even at the cost of falsehood. There is also speculation that optimism is a condition in which people systematically overestimate the likelihood of good outcomes and underestimate the likelihood of bad ones, whilst at the same time getting on better in life than the more accurate pessimists.

A more modest claim, but still unsatisfactory to many, is to make rationality itself a hypothetical imperative. Insofar as you care about truth, exhibit a theoretical reason. Insofar as you care about coherent desires, make them coherent. Insofar as you care about satisfying your desires, act so as to make them most likely to be satisfied in the light of your beliefs. Of course even the modest claim, to have any force, requires a principle very like practical reason: insofar as you care about something, follow the rule that will give it to you.

Rationality and Reasoning

The fact that rationality is often thought of as a characteristic human excellence, and so is the power of conscious thought, has sometimes led to a confusion between rationality and conscious, deliberate reasoning. In fact it is very useful to distinguish between them. Much of what we do is not the result of conscious deliberate reasoning (sometimes called ratiocination). We make decisions, pursue our goals, and act in a way which can be praised as rational or criticized as irrational, despite the fact that it may not happen in a conscious way. If it were only conscious deliberate reason that was rational, then rationality would be even harder to find than it is. Of course, in showing that a piece of cognition is rational, it sometimes helps to provide an argument which shows how the process can be *rationalized*: that is, described in a way that might look like a valid piece of deliberation or ratiocination. But this is not to say that this is what originally went on. Thus we might say of animals, even if we do not think they possess the power of conscious reasoning, that their mental processes are rational insofar as they process the information they receive about the world and respond in a rational way.

Bibliography

Armstrong, 1968. A Materialist Theory of the Mind. Routledge and Kegan Paul, London.

Braddon-Mitchell, David, Jackson, Frank, 1996. The Philosophy of Mind and Cognition. Blackwell, Oxford.

Bradley, Richard, 2007. The kinematics of belief and desire. Synthese 156 (3), 513–535.

Carroll, L., 1895. What the Tortoise said to Achilles. Mind 4 (14), 278–280.

Churchland, Paul, 1981. Eliminative materialism and propositional attitudes. The Journal of Philosophy 78.

Colyvan, M., 2006. No expectations. Mind 115 (459), 695–702.

Dennett, D., 1991. True believers. In: Rosenthal, D. (Ed.), The Nature of Mind. Oxford University Press, Oxford, pp. 339–353.

Dretske, F., 1983. Precis of knowledge and the flow of information. Behavioural and Brain Sciences 6, 55–90.

Elster, J., 1979. Ulysses and the Sirens: Studies in Rationality and Irrationality. Cambridge University Press, Cambridge.

Fodor, J., 1990. A Theory of Content and Other Essays. Bradford Books, MIT Press, Cambridge, MA.

Jeffrey, R., 1983. The Logic of Decision. McGraw-Hill, NY.

Kahneman, D., Slovic, P., Tversky, A. (Eds.), 1998. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, pp. 3–20.

Lee, P.M., 2004. Bayesian Statistics: An Introduction, third ed. Hodder Arnold, London, England.
Lewis, David, 1972. Psychophysical and theoretical identifications. Australasian Journal of Philosophy 50 (3), 249–258.
Millikan, R., 1984. Language, Thought and Other Biological Categories. MIT Press, Cambridge, MA.
Neander, Karen, 1991. Functions as selected effects: the conceptual analyst's defence. Philosophy of Science 58 (2), 168–184.
Searle, J.R., 1983. Intentionality. Cambridge University Press, Cambridge.
Skyrms, B., 1986. Choice and Chance, third ed. Wadsworth, Belmont, CA.
Smith, M., 1994. The Moral Problem. Blackwell, Cambridge, MA.
Sorabji, R., 1991. From Aristotle to Brentano: the development of the concept of intentionality. In: Blumenthal, H., Robinson, H. (Eds.), Oxford Studies in Ancient Philosophy.
Stich, S., 1990. The Fragmentation of Reason. Cambridge University Press, Cambridge, MA.
Stich, S., Samuels, R., 2000. Ending the rationality wars: how to make disputes about human rationality disappear. In: Elio, R. (Ed.), Common Sense, Reasoning and Rationality, vol. 11. Oxford University Press, New York.
Zalta, E., 1988. Intensional Logic and the Metaphysics of Intentonality. MIT Press, Cambridge, MA.

Intentionality during Infancy and Early Childhood, Development of

Dankert Vedeler, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The usual distinction between philosophical intentionality and psychological concrete intentions is here disregarded. That allows for a broader approach for studying the development of intentionality. Cognitive science, relying heavily on the concept of representation (e.g., goal representations), is compared to dynamic systems approach, which does not need to assume representations. As an alternative to a description of the development of concrete intentions in terms of goal directedness, that is, in terms of representations, dynamic systems approach opens for a perspective on the development of intentionality that provides better insight to the complexity and adaptability of human intentional behavior.

Introduction

In the late nineteenth century, the philosopher Franz Brentano (*see* Phenomenology: Philosophical Aspects), facing the emergence of psychology as a 'science of the mind,' wanted to distinguish between what was physical and what was mental about the human being. He considered mental phenomena – in contrast to physical phenomena – to have a directedness to something else than itself, and took the medieval scholastic term 'intentionality' to designate this characteristic of the mental. It may be manifest as our thinking being thinking of something. It means that the thought and what the thought is about is not the same thing. The thought is one thing, and the object of the thought is what the thought is directed toward.

However, if we go back to medieval prescholastic philosophy, we will find that Augustine was the first to use the word 'intentio' (in Latin). He was inspired by early Greek philosophy, where this directedness relationship also applies to perception and its object, to action and its object, as well as to will or desire and their object (what I want or desire) (Caston, 2008).

The Psychological Concept of Intention

Psychologists use to make a distinction between the philosophical concept of intentionality, as mental directedness, and the psychological concept of intention, meaning purpose. Many authors, however, while making the distinction, slip back into using 'intentionality' meaning 'purposefulness,' as opposed to concrete intentions linked to particular actions. This omission is actually in agreement with the above-mentioned roots of the concept in Greek philosophy. Intentions or intentional actions, as discussed by psychologists, also have their objects, for example, the glass that you grasp and the water you drink. Thus, such intentions could be subsumed under the same concept of intentionality that philosophers use. So did the philosopher Maurice Merleau-Ponty (1962), who joined the ancient Greeks by suggesting that also behavior may be intentionally directed, namely, toward what is around in the real world.

One reason for making the distinction between philosophical intentionality and psychological intentions could be to avoid ascribing mind to chimpanzees, while granting them purposes, for example, they may grasp an apple to eat it. That goes also for infants. However, in contemporary psychological literature on infant intentions, there is no mention of infant purposeful behavior. Instead of focusing on the action itself, it has been common to ask for its goal, that is, the end state to be achieved by the action (*see* Goal-Oriented Behavior among Human and Nonhuman Primates). Infant and toddler intentionality is discussed in terms of goal directedness.

This makes it necessary to consider the relationship between the concepts of intention, purpose, and goal. For example, one could argue that first you have an intention, then you act, with the purpose that corresponds to your intention, and finally you achieve the goal of your intention. This sounds like you having conceived an intention in your mind, then you have acted with your body, in order to make reality correspond to your goal, that is, you have changed reality to suit your mind. It would imply that purposeful action require mindful intentionality, in Brentano's sense. Such a conception would not grant chimps purposefulness.

What Is Intentional, and What Is Not?

If Brentano's idea is valid, then humans are intentional, and other species lack intentionality. However, if chimpanzees are not intentional and do better than small infants, we have to concede that infants come to the world as non-intentional beings. Eventually, they gain advantage over the chimps. So, they are in the unique position of becoming intentional. The development of infant intentionality is thus a privileged area of inquiry, where one needs to address the question what, actually, one should mean with the word. We will see that there are arguments for considering infants to be born as intentional beings, as well as for considering chimps, and even much simpler organisms to be intrinsically intentional.

Infant Intentionality in a Historical Perspective

William James

In the late nineteenth century, William James (*see* James, William (1842–1910)) characterized babies as a "blooming buzzing confusion" (1890: p. 488), where all sensory

impressions "fuse into a single undivided object for that mind." While using the concept of intention, in the sense of purpose and aim, he does not explain or discuss its emergence in infancy. We should notice, however, that he thinks pinpointing any particular intention will make it vanish from the ongoing stream of thought in which we, at any moment, are engaged (James, 1890: pp. 253ff). The particular intention, for example, to use a particular word in a spoken sentence, is a fugitive part of the whole that depends on the intention to convey the meaning of the sentence, which, in turn, depends on the context in which the sentence is spoken, and so on. In discussions on intentionality, this flow of consciousness, as James puts it – but we could as well say stream of thoughts, or of actions, or of behavior – is commonly overlooked, analyzing, by abstraction, this or that intention, this or that behavior. That works in the laboratory, with an appropriate experimental design. The challenge is to extract intentions from video observation of a fussy baby playing with toys, where intentions can be studied in a context.

James Mark Baldwin

Contemporary to William James, James Mark Baldwin (1892) was more outspoken on development, proposing a distinction between physiological, sensorimotor, and imitative processes. He saw the persistent, repeated imitation (*see* Imitation during Infancy and Early Childhood), to achieve perfection, as the first sign of conscious volition. What is conscious is primarily what is immediately given here and now as sensations. With persistent imitation, however, the infant shows proof of having in consciousness 'a copy' of an absent thing or event that it wants to imitate, not giving up until having obtained satisfaction with its own behavior relative to the target.

Jean Piaget

Jean Piaget (1962; *see* Piaget's Theory of Cognitive Development; Piaget, Jean (1896–1980)) also discussed imitation, however, not its role for the development of intentionality, but for the development of language and thinking. He emphasized the absence of the imitated model as the most important feature; for example, when his daughter Jacqueline imitates the temper tantrum of another child she had observed the day before, implying that she has what Baldwin would call a copy of the other child's behavior in her mind. For Piaget, such imitation, in the absence of the model, is the starting point for mentally representing absent things and events. While Baldwin does not account for but rather presupposes the mental copy in humans, for Piaget the displayed imitation may become a mental object by a process of internalization, thus laying the basis for developing a mental capacity that is particular to humans, eventually also including mental images and symbols. This allows us to 'manipulate' mental representations of objects in the mind rather than with our senses and muscles in the physical world. Eventually, some mental representations are also externalized, as drawings etc., as well as spoken or written language.

Such mental manipulation could well illustrate Brentano's conception of intentionality. That was not what Piaget meant, however, and the closest we come to the concept of intentionality is Piaget's description of the means–end relationships, that is, when the infant chooses the right means to obtain a desired end.

Lev Vygotsky

Lev Vygotsky (*see* Vygotsky's Theory of Cognitive Development), in the Soviet Union, was also preoccupied with the use of symbols and he used internalization to explain the emergence of thinking. However, instead of starting with internalization of imitations, he went directly to language, stressing its social function. Through language the child internalizes social relationships. Vygotsky thus gave thinking a social origin. From the perspective of intentionality, the influence of Vygotsky has been to give increased focus on intentional content, that is, the particular object of a thought, or the purpose of an action. By that, intentionality becomes less of a general and abstract function of human mind, and more of a world-related integration of relationships with the physical and social experiences in a person's life.

However, Vygotsky's own view on volition imposed strict requirements on children's actions for them to count as intentional. He required full consciousness (Vygotsky, 1987: p. 213), as, for example, when learning a foreign language through deliberate study of grammar and vocabulary, something quite different from the spontaneous learning of a mother tongue.

Jerome Bruner

Jerome Bruner (*see* Bruner's Theory of Cognitive Development), who helped introduce Vygotsky to Western academic psychology, was explicit on intentions: "An intention is present when an individual operates persistently toward achieving an end state, chooses among alternative means and/or routes to achieve that end state, persists in deploying means and corrects the deployment of means to get closer to the end state, and finally ceases the line of activity when specifiable features of the state are achieved. The elements of the cycle, then, comprise aim, option of means, persistence and correction, and a terminal stop order" (Bruner, 1981: p. 41). This covers both the persistence and the means–end relationship. While there is no principled difference between Piaget and Bruner on the nature of intentionality, Bruner was inspired by Vygotsky to pay more attention to the role of the child's social environment in the development of infant intentionality.

Cognitive Science

In the 1960s and 1970s, modern infant research revealed extensive skills and potentials of the human infant already at birth, something that may be contrasted with James' famous dictum mentioned here above, as well as with Piaget's view of infancy. These skills supported reactions against behaviorism concerning language and cognitive development and revived the idea of innateness, widely accepted since Chomsky's critique of Skinnerian behaviorism in the beginning 1960s. 'The cognitive revolution' that followed gave birth to a new paradigm in

psychology, cognitive science (CS), with tight interdisciplinary links to, for example, linguistics and neuroscience (Thagard, 2012). It had great impact on infant research, including on the development of intentionality.

The CS terminology and its view on the relationship between organism (person) and its environment are based on 'the computer metaphor.' This implies that information from the environment is experienced, understood, reacted to, and acted upon only after being processed by the mind, or by the brain. CS thus outlines an 'information processing' translation between the organism and the world. Information received through the senses needs to be matched with some representational content of the mental or neurophysiological apparatus that is adequate for the situation and, vice versa, from there to the body for adequate action to be executed. Eventually, CS has come far in recognizing the interactional character of such processes, but the paradigm remains the same; representations are considered as a necessary intermediary, and at the same time a dividing line between the organism and the world.

The Cognitive Science Concept of Representation

This concept of representation is thus different from Piaget's use of the same word. (There are, however, affinities between Piaget's theory and CS; the concept of 'schema' (*see* Piaget's Theory of Cognitive Development) in Piaget's theory resembles the CS concept of representation.) It first referred to something more 'mental' or abstract, such as denoted by Jerry Fodor's (1975) expression 'language of thought,' based on rule-governed relationships similar to the structure of natural languages. It then changed to be specified as 'motor' or 'neurological' representations, denoting neuromuscular or brain structures that underpin and correspond to particular behaviors, feelings, thoughts, etc. (*see* Concept Learning and Representation: Models; Knowledge Representation; Logics for Knowledge Representation; Cognitive Neuroscience; Neural Representations of Objects; Reference and Representation: Philosophical Aspects).

This is true also for the understanding of infant intentionality; Bruner (1981) stands for the earlier view, depicting intentionality in terms of action toward end states. Vittorio Gallese (2010) illustrates well the latter use of representations for the study of intentionality. He proposes a particular "intentional type of representation: a representation specifically and exclusively coded under a distinct type of motor neural activity pattern, one involving dynamic organism–object relations" (p. 205). What is represented is thus a particular type of interaction between an agent and the world he/she is acting upon. Furthermore, Gallese (p. 207) is explicit on the representational character of goals. Goals do not exist in the world other than as representations of goal states. To conclude, one may say that the development of infant intentionality from a CS perspective amounts to the development of goal representations and representations of actions to achieve goals.

Infant Intentionality as Goal Directedness

While Bruner understands goals and goal-directed action as mentally represented and Gallese has a neurophysiological understanding of the same concepts, Michael Lewis (1990) proposed a wider understanding of goal directedness. He noticed that infants, from 2 months of age, having gained control over a pleasant stimulus through its own action, did not stop trying when the triggering action failed to bring forth the stimulus, as would be predicted by a simple associative learning paradigm. Instead, the child was frustrated and increased its efforts to make the stimulus reappear. Lewis explains the results from a principled standpoint, that "all goal-directed systems are intentional" (Lewis, 1990, passim); thus these experimental results are a demonstration of infant intentionality already at 2 months and not just 'circulary reactions,' as Piaget had claimed (*see* Piaget's Theory of Cognitive Development).

Lewis relates goals to emotions, as desires. Ultimately, they serve the survival of the individual and thus have an evolutionary background. It would simply mean that survival depends on behavior having the purpose to seek pleasant emotions and to avoid unpleasant ones. Goal-directedness thus becomes the basic condition for life at any organismic level, and so would include "the intrinsic goal-directedness in the behavior of living eukaryotic cells," as proposed by Tecumseh Ficht (2008: p. 157). It would then resemble Humberto Maturana and Francisco Varela's (1980) concept of 'autopoiesis' (*see* Law, Autopoiesis in), which literally means 'self-creation,' referring to the capability of cells and organisms to maintain themselves, and, if necessary, repair themselves and change themselves through an active use of resources from the environment.

From there, Lewis conceives a developmental trajectory for intentionality, where emotion is successively complemented by cognition and, at the end, by metacognition (conscious awareness of own cognition, Lewis, 1990: p. 244, not verbatim). Five different levels of intentionality are then to be considered as levels of the self-regulation of the individual organism, from "a plant's movement toward the sun, a newborn child's movement toward a brightly colored object, and an adult's conscious goal-directed behavior" (Lewis, 1990: p. 231). Thus, we have a very broad span of different conceptions of goal-directedness, from autopoiesis, over means–ends relationships (Piaget and Bruner), to explicit conscious planning, as we saw with Vygotsky. They are all covered by Lewis' proposal on levels of intentionality. Only the highest level would satisfy Brentano's conception of intentionality.

Social Intentionality

One particularly advantageous asset of the human species is that we are social about our intentions, to an extent not found in any other species. Therefore, an interesting question is how do infants become social about own and other humans' intentions.

Theory of Mind

David Premack and Guy Woodruff (1978) introduced the concept 'theory of mind' (*see* Theory of Mind during Infancy and Early Childhood and Cultural Variation, Development of;

Mind, Theories of; Intentionality in Language and Communications), asking whether chimpanzees take into account that other beings, as well as themselves, have mental states, that is, beliefs, desires, opinions, and even pretense ('false beliefs'). This proposal gained huge success, engendering a host of research aimed at exploring what kind of theory of mind infants and possibly primates might have at different ages.

Eventually, two different conceptions of theory of mind emerged. On the one hand, 'theory' should be taken literally, that is, a kid would have a particular ('folk psychology') notion that there is something called 'mind' and that this mind to a substantial degree influences what people do and say. This interpretation was named "*theory* theory of mind." On the other hand, a bundle of what were labeled 'simulation theories' came to the fore. Research showing a parallel development in the understanding of own actions' goal-directedness and of the goal-directedness of other persons' actions supported proposals that infants associate own action to other people's actions (e.g., Sommerville and Woodward, 2010).

Mirror Neurons

Simulation theories also gained support from the discovery of 'mirror neurons' (*see* Mirror Neurons, Theory of), which fired both when a subject carried out an action and when it perceived someone else carrying out the same action (e.g., Gallese, 2010). An important point is that the activation of these neurons is not linked to the particular motor movement patterns used by the actor, but to what is considered as the goal of the action, that is, its anticipated outcome.

Thus, there has been an increased interest in the importance of social interaction for cognitive development, including the development of intentionality, spanning from sociocultural (Vygotsky) to neurophysiological approaches. We now turn to one of the strongest manifestation of this interest.

Infant Intersubjectivity

The research on infant intentionality within the CS paradigm has been conducted almost exclusively in laboratories. The point of laboratories is experimental control, such that you can be sure about the relationships between experimental variables, for example, whether awareness of own intentions goes hand in hand with awareness of other people's intentions.

Innate Intentionality

Colwyn Trevarthen had a different methodological approach. He studied the development of infant intentionality through longitudinal live video recordings of natural mother–infant interaction. He proposed that intentionality was innate, most strongly stated in a book chapter written together with Penelope Hubley: "We conceptualize this development to be by differentiation of a fundamentally coherent field of intentionality which is, however, already anatomically partitioned at birth into three modes out of which three kinds of experience and action are generated postnatally. These modes are probably three real systems of the brain that achieve functional differentiation by interaction with each other and with the environment." The three modes he conceived were "Forms of action and perceptual processing appropriate for (1) knowing and using objects (praxic mode), for (2) communicating with the human world (communicative mode), and for (3) acting in self-directed or thoughtful manner (reflective mode) appear as distinct rudiments in the newborn" (Trevarthen and Hubley, 1978: p. 213).

Other researchers have substantiated Trevarthen's proposals on the two first modes by reporting a capability for imitation (*see* Neonatal Imitation) already at birth (Kugiumutzakis, 1999) and different movement styles toward persons and toward things (Rönnqvist and Hofsten, 1994). Trevarthen used these two first modes, the praxic and the communicative mode, as elements in a three-stage model for the development of infant intentionality, where, in the first stage, the child between 2 and 4 months shows an advanced social intentionality in interaction with a parent, in what Trevarthen labeled 'primary intersubjectivity' (communicative mode). Between 4 and 9 months, as the child gains better motor control, this social interaction gives way in large part to interaction with things (praxic mode). Finally, from 9 months, the child coordinates, in 'secondary intersubjectivity,' interaction with things within social interaction, such that the social interaction itself may be about things (communicative and praxic modes combined). The difference between primary and secondary intersubjectivity is thus that first the interaction is only about mother and infant, where, as it were, the dyad is shielded from the world outside. In secondary intersubjectivity, however, the dyad is able to include the outside world, in terms of, for example, a toy, such that child and adult may share experiences (*see* Joint Attention during Infancy and Early Childhood Across Cultures) of the toy. For the younger child, then, it is either social interaction or interaction with the toy, later it is both together.

Intentionality as Organism–World Relationship

This three-stage model was not followed up in Trevarthen's later writings, which might have something to do with this understanding of intentionality being incompatible with the above-mentioned definition of the concept proposed by Bruner. Inspired by Merleau-Ponty (mentioned in the Introduction section), Vedeler (e.g., 1991) suggested that the common understanding of intentions in terms of goals could be replaced by a more general concept, in particular, concerning infant intentionality, as 'behavioral object-directedness.' Merelau-Ponty (1962: p. 138) writes: "In the action of the hand which is raised towards an object is contained a reference to the object, not as an object represented, but as that highly specific thing towards which we project ourselves" (quoted in Vedeler, 1991: p. 439). That would imply shifting focus from a cognitive to a behavioral and relational understanding of intentionality, thus avoiding a discussion about whether infants possess the mental capacity of having intentions, and instead addressing what Trevarthen observed, namely, the skills infants develop in handling both people and things, and the coordination of the two.

While never properly defining 'intention' and 'intentionality,' it is very clear that Trevarthen uses these words to describe particular relations the baby has to its environment. When it comes to secondary intersubjectivity, the important

conclusion he draws is that the coordination of person-directed and thing-directed intentions is the starting point for the development of human thinking and language. I suggest that it is also the starting point for the development in humans, both of what most philosophers mean with intentionality and what psychologists meant with intentions.

The third, reflective, mode of intentionality is not elaborated to the same degree, although Trevarthen later proposed a third, tertiary level of intersubjectivity, where child and adult are able to share meanings about things, as well as about self and others. However, this level corresponds roughly to the child's having a theory of mind, as mentioned here above.

Trevarthen uses the word 'representation' sparsely, although he seems not to oppose the CS understanding of it. The quote from him here above about the three modes of intentionality as real systems in the brain cannot be understood as anything else than hypothetical neurological representations. Also, while it has not been possible to find any statement from himself on the subject, the mirror neurons findings have generally been considered as support for his proposals on innate intersubjectivity.

Dynamic Systems Approach

Many have argued that the CS conception of representations is not necessary – and is also misleading – for describing the organism–environment relationship (cf, e.g., Costall, 2007). Several nonrepresentational approaches have been proposed. Many of them try to find a common ground with CS. Not having the space to here present several criticisms and alternative views, I choose one view, namely, dynamic systems approach (DSA) (*see* Self-Organizing Dynamical Systems, probably only about mathematical application of DSA. Certainly indispensable for understanding, but not giving full justice to the scope of possible applications). I consider DSA to have the greatest potential to offer a consistent alternative to CS.

No Need for Representations

There have been many attempts to integrate DSA into CS. In my view, however, they have not been convincing. The stumbling block is the concept of representation. While CS relies heavily on this concept, in DSA it is not needed. Things in the world and neurophysiological structures in the body (as well as the body as such) are what they are. Thus, they are on par. This means that components of a dynamic system may be both organismic and environmental, and neither have precedence over the other. If you want to calculate the product of two three-digit numbers, you may do it mentally or you may use a calculator. With the calculator, you include a component from the environment in the system you assemble for the solution of the multiplication task.

Most importantly, mental or neurophysiological structures do not represent things and events in the world. For example, brain structures consist of systems of neurons related to each other by synaptic facilitation and inhibition. That's what they are, and nothing more. Inactive, they are just a heap of neurons. They are activated if they are relevant for the solution of a task. For example, they may work together for the solution of the multiplication task. However, if, for example, the neurons in question are not well enough integrated for the task (i.e., you are not good at math), other resources might be preferred, for example, a calculator. The involvement of the brain in the task will then be quite different (e.g., pushing the digit buttons on the calculator and reading the screen).

Representations and Change

The organism is under constant transformation, as long as it is alive; for example, the world and own activity induce experiences that change synaptic links in the brain. If structures of synaptic links are to be considered as representations, it would mean that they keep their structure roughly unchanged, being activated in different situations. Anatomically, however, changes do take place, and – in all living organisms – subsequent changes depend on previous changes, such that the history of the organism becomes important. This is particularly true for humans, who mentally accumulate and retrieve experiences to an extent not found in any other species. If structures of synaptic links were to be considered as representations, the 'sameness' of some particular neuroanatomical structure would be relative, and it would be a question, at each moment, whether it is relevant or not. If not, it would no longer work as a representation.

Resources Instead of Representations

The question whether such neural structures are innate or acquired would then be beside the point. From a DS perspective, it is preferable to consider such structures as resources. Consider representations as 'hardwired': they are innate, selected by evolution, refined, developed, and adapted by experiences. While being adaptable to different situations, their function should be basically the same from time to time. Resources, on the other hand, are temporary and contingent on the situation. Are they useful or not? That depends. Sometimes some particular neuroanatomical configuration for any individual organism could be recruited to solve a particular task in a particular situation.

By habits and routines, organisms tend to assemble the same resources for similar tasks in similar situations. Thus, they help generalizing and automatizing, such that daily life becomes easier, although it may also result in malfunctioning and mistakes. However, it is always an open question which resources will be recruited in any particular situation. In DSA language, it is called 'soft assembly' (e.g., Thelen and Smith, 1994: p. 60). There are no fixed rules, no grammar, no store of solutions for solving tasks. Each situation is unique. Even if you have a strong habit for mental calculation, whether you choose to use your head or the calculator to solve your multiplication task – even for the same constellation of digits – may depend on the situation: how well you slept last night, how sure you need to be about the product, the availability of the calculator, etc.

The Impact of Dynamic Systems Approach

DSA gains increasing influence in psychology. It is not a coherent paradigm to the same degree as CS and is very

differently interpreted and used by different researches (Lewis, 2000). Shaun Gallagher (2008), as a philosopher particularly interested in neurophysiology, argues for a DSA to the analysis and explanation of intentional action. He also features the social dimension in the development of intentional action and revives Trevarthen's concepts of primary and secondary intersubjectivity mentioned here above. Ester Thelen and her group have provided the most well-known and referred to application of DSA focused on infant motor development (e.g., Thelen and Smith, 1994). In my view, the caveat is that, while to a great extent taking an antirepresentational stance, most DSA research and publications implicating intentionality – Thelen's work included – describe infant intentional actions in terms of goals and goal-directedness.

Developing through Relationships

Alan Fogel (e.g., 1993) is an exception. Based on a DSA, both methodologically and theoretically, he uses longitudinal (*see* Longitudinal Data: Event–History Analysis in Discrete Time; Longitudinal Data: Event–History Analysis in Discrete Time; Longitudinal Research: Panel Retention) video observations of live mother–infant interaction (*see* Qualitative Methodology in Developmental Psychology; Qualitative Research), where the mother and the infant are considered as one single dynamic system, and also the unit of analysis. Instead of assuming child and mother to have individual, possibly concurrent, goals, Fogel suggests that ongoing interaction organizes itself into creative achievements, where the history of the dyad in similar settings plays a central role for the further deployment of the interaction. Instead of assuming intended goals for individual actions, Fogel's approach focuses on the ongoing interaction, based on habits and routines that give the present situation its historical context and emotional color, and within which new purposes may emerge.

As an example, Vedeler (Vedeler and Garvey, 2009) used Fogel's methodology in following the transition toward secondary intersubjectivity in Trevarthen's tripartition of the development of infant intentionality. The point of the investigation was to explore the contextual resources for the coordination of social intentions with thing-oriented intentions before full-blown secondary intersubjectivity, the word 'intention' being understood as object-directed action, whether social objects or things. Weekly recordings of a recurring setting, involving a fixed set of toys, revealed the emergence over time of shared experiences of toy properties as resources for the play (cf Costall, 2007: pp. 68ff, on affordances). These new resources are, in turn, selected for further play with the same toy. Thus, situated shared experiences promote the emergence of a new kind of intentions regarding the toy, which may then become routines, later recruited for further and different tasks in the future.

Perspectives for Future Research

As to the future for research on the development of infant intentionality, I see two possibilities. Within a short perspective, most research in the area will be guided by the CS approach. In a longer perspective, however, DSA will prevail.

The CS vision of human cognitive functioning is based on laboratory findings. In DSA, reality is also, to a large extent, simplified, as one tries to find the essential components that are recruited for a particular kind of task solutions. Mathematically oriented proponents of DSA (*see* Self-Organizing Dynamical Systems) are operating with as few variables as possible. However, reality is infinitely complex. A more relaxed, 'metaphorical' use of the DSA concepts is less specific on the variables, but gains in ecological validity. This will give more importance to context and social interaction.

In this perspective, Fogel's approach might be the best candidate for giving a theoretically consistent alternative to the CS approach in studying the development of infant and early childhood intentionality. It should then not be considered as the growth of any particular mental or neurophysiological function, but as a manifestation of an activity relationship with the world, common for all living organisms, most salient, however, in the species *Homo sapiens*, having a unique skill in drawing on social resources for solution of daily life tasks.

Trevarthen was the first to draw attention to the details of early infant social propensities. With the concept of secondary intersubjectivity, he also proposed a theoretical model for human intentionality as a starting point for the development of human cognition. From a CS perspective, Trevarthen's model gains support from neurological findings, in particular on mirror neurons. He also receives recognition, in passing, from several prominent CS scholars on infant intentionality, although they rarely draw upon his model. From a DSA perspective, however, it would be wrong to assume an innate intentionality to describe primary and secondary intersubjectivity. No doubt, there are both neurological and other anatomical features of the newborn human infant (in its brain, but also, for example, the complexity of facial and finger muscles) that may be excellent resources for social interaction and manipulative skills. From there, intentions may emerge as the intricate mutuality of social interaction between mother and infant gaining particular impetus from being combined with manipulative skills.

Whether from a CS or a DSA perspective, there is by now unanimity on the importance of social interaction for human cognitive development, and thus also for intentionality in Brentano's sense. We are all developing through relationships, as the title of Fogel's (1993) book tells.

Bibliography

Baldwin, J.M., 1892. Origin of volition in childhood. Science 20, 286–287.

Bruner, J.S., 1981. Intention in the structure of action and interaction. In: Lipsitt, L.P., Rovee-Collie, C.K. (Eds.), Advances in Infancy Research, vol. I. Ablex, Norwood, NJ, pp. 41–56.

Caston, V., 2008. Intentionality in ancient philosophy. In: Zalta, Edward N. (Ed.), The Stanford Encyclopedia of Philosophy, Fall 2008 ed. http://plato.stanford.edu/archives/fall2008/entries/intentionality-ancient/

Costall, A., 2007. Bringing the body back to life: James Gibson's ecology of agency. In: Ziemke, T., Zlatev, J., Frank, R.M. (Eds.), Body, Language, and Mind, Embodiment, vol. 1. Mouton de Gruyter, Berlin, pp. 55–84.

Ficht, W.T., 2008. Nano-intentionality: a defense of intrinsic intentionality. Biology & Philosophy 23, 157–177.

Fodor, J.A., 1975. The Language of Thought. Crowell, New York.

Fogel, A., 1993. Developing through Relationships. Harvester Wheatsheaf and University of Chicago Press, London.

Gallagher, S., 2008. Are minimal representations still representations? International Journal of Philosophical Studies 16, 351–369.

Gallese, V., 2010. Of goals and intentions: a neuroscientific account of basic aspects of intersubjectivity. In: Grammont, F., Legrand, D., Livet, P. (Eds.), Naturalizing Intention in Action. MIT Press: Bradford Books, Cambridge, MA, pp. 201–225.

James, W., 1890. The Principles of Psychology, vol. 1. Holt, New York.

Kugiumutzakis, G., 1999. Genesis and development of early infant mimesis to facial and vocal models. In: Nadel, J., Butterworth, G. (Eds.), Imitation in Infancy. Cambridge University Press, Cambridge, UK, pp. 36–59.

Lewis, M., 1990. The development of intentionality and the role of consciousness, including commentaries and author's response. Psychological Inquiry 1, 231–283.

Lewis, M.D., 2000. The promise of dynamic systems approaches for an integrated account of human development. Child Development 71, 36–43.

Maturana, H.R., Varela, F.C., 1980. Autopoiesis: the organization of the living. In: Maturana, H.R., Varela, F.C. (Eds.), Autopoiesis and Cognition: The Realization of the Living. Reidel, Dortrecht, Holland, pp. 59–140.

Merleau-Ponty, M., 1962. Phenomenology of Perception (C. Smith, Trans.). Routledge & Kegan Paul, London (Original French edition published in 1945.).

Piaget, J., 1962. Play, Dreams, and Imitation in Childhood. Norton, New York (Original French edition published in 1945.).

Premack, D., Woodruff, G., 1978. Does the chimpanzee have a theory of mind? Behavior and Brain Sciences 1, 515–526.

Rönnqvist, L., von Hofsten, C., 1994. Neonatal finger and arm movements as determined by a social and an object context. Early Development and Parenting 3, 81–94.

Sommerville, J.A., Woodward, A.L., 2010. The link between action production and action processing in infancy. In: Grammont, F., Legrand, D., Livet, P. (Eds.), Naturalizing Intention in Action. MIT Press: Bradford Books, Cambridge, MA, pp. 67–89.

Thagard, P., 2012. Cognitive science. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy, Fall 2012 ed. http://plato.stanford.edu/archives/win2012/entries/cognitive-science/

Thelen, E., Smith, L.B., 1994. A Dynamic Systems Approach to the Development of Cognition and Action. MIT Press, Cambridge, MA.

Trevarthen, C., Hubley, P., 1978. Secondary intersubjectivity; confidence, confiding, and acts of meaning in the first year. In: Lock, A. (Ed.), Action, Gesture, and Symbol; The Emergence of Language. Academic Press, London, pp. 183–229.

Vedeler, D., 1991. Infant intentionality as object directedness: an alternative to representationalism. Journal for the Theory of Social Behaviour 21, 431–448.

Vedeler, D., Garvey, A., 2009. Dynamic methodology in infancy research. In: Valsiner, J., Molenaar, C.M., Lyra, C.D.P., Chaudhary, N. (Eds.), Dynamic Process Methodology in the Social and Devlopmental Sciences. Springer, New York, pp. 431–453.

Vygotsky, L.S., 1987. Thinking and speech. In: Rieber, R.W., Carton, A.S. (Eds.), The Collected Works of L. S. Vygotsky, Problems of General Psychology, vol. 1. Plenum Press, New York, pp. 39–285.

Intentionality in Language and Communication, Emergence of

Aliyah Morgenstern, Sorbonne Nouvelle University, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The attribution of intentionality is the ability to view another's actions, behaviors, gestures, or speech as having a purpose. The emergence of intentionality in infants is thought to be a foundational feature of human communication and language. We will first discuss those issues, then trace children's developmental pathway from preverbal cognitive skills such as eye contact, joint attention, intention reading and sharing to attribution of mental states (Theory of Mind), and their link to language use. We will finally focus on the role of the input and of adult scaffolding in the development of those social-cognitive abilities.

Introduction

Following Piaget (1962) and Vygotsky (1962), nonnativist scientists agree that language emerges from nonverbal cognitive and social development in infancy. The subsequent research has strived to investigate how early perceptual-cognitive capacities pave the way for language and communication. Even though children have innate biological and cognitive capacities, they need to learn conventions and formal patterns from communication and language in their environment. They gradually become fully bloomed interacting speakers building on such cognitive and social skills as the ability to follow another's gaze, to draw and maintain their attention, to imitate, to read others' intentions, to make analogies, to categorize, and to symbolize.

A wealth of studies have shown that it is the interaction and complementarity between basic perceptual, cognitive, affective processes and a favorable environment that involves caring adults or older siblings constantly adjusting their behavior to the infants' that seem to trigger and guide the emergence of shared attention and intentionality which will then lead the child into symbolic communication.

We will first present the notions of intentionality and Theory of Mind as possible foundational features for human communication and language; we will then describe the child's developmental pathway from preverbal cognitive skills to language use and finally focus on the role of the input and of adult scaffolding.

Intentionality and Theory of Mind

Definitions

The word 'intentionality' comes from the Latin 'intendere' which means to aim in a certain direction for a target. The attribution of intentionality is the ability to view another's actions, behaviors and production of gesture or speech as having a purpose. Toddlers can exhibit the understanding of intentions for example by handing over his car keys to their father who has just said 'good-bye' in the morning and is restlessly looking everywhere as he does almost every morning. The repetition of the same scenario enables the child to memorize the sequence of events and to provide the missing keys. Infants can also show that they have an intention when communicating the need to be given the ball, which is out of reach by looking and pointing at the ball as well as pulling the adult's sleeve.

According to Baron-Cohen (1991), the infant's understanding of intentions in others, a social skill found by 7–9 months of age, is a 'critical precursor' to the development of 'Theory of Mind' (ToM). ToM is the ability to attribute mental states, such as beliefs, intentions, memories, and desires to oneself and to others. It enables us to understand that mental states can be the cause of others' actions or behaviors. We only assume that others have minds based on analogy to ourselves and thanks to our participation in social interactions.

Theory of mind develops over time. It is constructed on the basis of precursor skills all the way to the understanding of mental states and their link to behaviors. Those precursor skills include joint attention (the ability to attend to or direct attention to a common object), intention reading, understanding that other people have different perspectives, use of cognitive and mental state verbal expressions, of pronoun reversals, and pretend play. Communication itself is triggered by the intention to share objects, properties, or states of affairs that are in our minds. Intentionality in communication corresponds to the encoding of messages *for* others.

A Foundational Feature of Human Communication and Language

Since the influential article by primatologists David Premack and Guy Woodruff entitled "Does the chimpanzee have a theory of mind?" (1978), much empirical research has been undertaken in investigating whether nonhuman primates can ascribe psychological states with intentionality to others and how human children develop this capacity. At the end of the twentieth century, human cognition and animal cognition were shown to differ in one important respect: our ability to understand intentions and to "take another's perspective." For Tomasello (1999), it is the very crucial biologically determined factor that distinguishes humans from primates and could also be deficient in autistic children. The emergence of intentionality is viewed as foundational because it provides the matrix to understand what others are doing.

According to Tomasello, at an important point in his/her cognitive development, between 9 and 12 months old, the child comes to experience himself/herself as an intentional agent – that is he/she views his/her behaviors as structured by goals. As he/she identifies with the adults around him/her, he/she automatically views them as intentional agents as well. Tomasello's theory, developed thanks to a wealth of

experimental studies, consists of a demonstration that all kinds of human cognitive abilities emerge from the exploitation of this basic human capacity and are transmitted culturally. Communication and language are part of those specifically human cognitive abilities that develop on the basis of this capacity. It is only because young children understand others as intentional agents that they can acquire and learn linguistic symbols: they begin to understand that their conversational partners are using these symbols to direct their actions and attention to outside entities. These symbols are organized according to conventional patterns and mutually shared rules and can only be transmitted through social learning.

Since 1999, a series of experiments has shown that though autism involves significant difficulties in understanding mental states that might be related to difficulties in processes as basic as on-line perception, which hamper their ability to respond to social stimuli such as vocal and facial expressions, autistic children and great apes understand intentions better than what was expected. They do not however engage socially with others in the way that typical children do. In more recent work, Tomasello et al. have attenuated their monolithic depiction of social cognition by giving a more gradual view (Tomasello and Rakoczy, 2003) and have suggested that an additional capacity was needed: shared intentionality (Tomasello et al., 2005). They suggest that only human beings can cooperate with each other and that it is that unique feature that allows humans to engage in the creation and use of linguistic symbols.

Children's awareness of other people's intentions, thoughts, and beliefs can only be inferred through what they do in experimental and naturalistic situations and there could be contradicting interpretations of their behaviors. Most of the evidence involving relationships between language development and theory of mind is built on correlations. It is probably premature to say that certain milestones are necessary for Theory of Mind development. But there might be an important impact of the lack of symbolic input on deficits in intentionality sharing and Theory of Mind. Indeed deaf children with no cochlear implant who do not have sign language as their first language perform very poorly as early as 18 months old in very basic tasks (Meristo et al., 2012) and at later ages perform similarly to autistic children in false belief tasks (Peterson and Siegal, 2000), whereas deaf signers perform to the same level as their hearing peers. Though the actual link between the tasks testing false beliefs and Theory of mind itself can be questioned, ToM seems to be highly dependent on specific interactive experiences during which adults behave and adjust to their infants' inferred mental states as well as they comment verbally about them. The major factor in explaining late-signing deaf children's delayed ToM development seems to be the reduced early experience of conversation and its role in perspective taking and mind sharing.

There seem to be complex interdependencies between language and Theory of Mind. It begins in infancy with joint attention and awareness of intentionality, and it continues with children's increasing use of mental state terms.

Mirror Neurons, Intentionality, and Theory of Mind

Researchers noted that certain neurons located in the prefrontal motor cortex of monkeys fired when the animals carried out an action and when they observed the same action carried out by another individual, suggesting that these neurons are important for encoding the intentions of others (Gallese et al., 1996). Since then, a number of researchers have underlined that the mirror-neuron system is an important component of the social-cognitive network. There is some evidence for impaired functioning of this system in autism. In order to understand the relationship between the processes involved in the mirror-neuron phenomena (such as imitation and empathy) and other aspects of communication and language, including theory of mind, it is now necessary to conduct longitudinal studies beginning before the onset of language and communication and of autism symptoms. Systematic neuroimaging studies should also be conducted to explore the deficits implicated in autism.

The Developmental Pathway

From Eye Contact to Shared Goals

Infants prefer their parents' faces and voices to other stimuli. They attend to their interaction partner and make eye contact almost from birth. They understand self-produced motion just a few months after birth (Bertenthal, 1996). They soon turn to look in the same direction as the adult. Woodward (1999) explains that around the age of 6 months, children make a difference between what can be expected from human beings who can reach for an object, and inanimate objects, who are not expected to reach for the same objects in similar circumstances. By 10 months, children can segment streams of continuous behavior into units (Baldwin et al., 2001). In studies involving obstacles, infants are shown to understand trying (Gergely et al., 1995). They also understand that when they succeed, adults will stop trying (Tomasello et al., 2005). Between 9 and 10 months, children engage in triadic situations involving themselves, an adult, and an object: activities such as giving and taking an object, throwing a ball to each other, putting toys away together, pointing-and-naming games. Children's gaze and actions demonstrate that they are coordinated with the other. The adult and child create a shared goal that they are going to reach together. However it is not until 12–15 months that children become involved in active joint engagement, sometimes taking the lead for they seem to understand the actual goal of the activity (Bakeman and Adamson, 1984). Not only are children able to share goals and coordinate their actions to the adults', they can also coordinate or reverse roles and take initiatives. It is around the same period that children try themselves to establish joint attention with others through pointing or other types of gestures.

Motor Control in Vocal and Gestural Productions

Children's ability to imitate adults in vocal and gestural productions shows that they make connections between their perceptions and their own actions and between their own bodies and the others'.

Before children produce their first words, they are able to systematically use prosodic cues to express a set of distinct pragmatic meanings. Thus, children at 0;9 and 0;11 are able to distinguish expressions of discontent and requests from responses and statements by means of prosody. Recent findings also report the use of adultlike intonational contours to convey

specific pragmatic functions in the one-word period (Marcos, 1987; for French; Prieto et al., 2012; for Catalan and Spanish).

Pointing gestures play an important role in the language acquisition process. They are grounded in joint attention, they trigger interaction, and they may also facilitate children's entry into word combination and syntax. Numerous studies have tackled this issue in the development of spoken language. The 'founding fathers' of the study of child development and language had great intuitions about the importance of gestures and their relation to language. In his notes on his son's development, Darwin (1877) stresses the importance of observing the transition from uncontrolled body movements to intentional gestures. Romanes (1889) compares human and animal gestures and makes fine observations on qualitative differences. He mentions the gestural language of deaf people as a sign of the universality of symbolic gestures. Stern (1924) considers pointing in particular as a precursor of intentional marking. For Werner and Kaplan (1963), pointing represents children's ability to discriminate between external objects and their own person. Communicational pointing then becomes the basis for referential behavior and reciprocity established in common activities between children and their parents (Bruner, 1975). As Tomasello et al. (2007, p. 720) underline, "pointing may thus represent a key transition, both phylogenetically and ontogenetically, from nonlinguistic to linguistic forms of human communication."

Not long after these first gestures produced to share attention with the adults, children engage in more complex communication and start using multimodal constructions including words. By doing so, they continue to share an intense collaborative activity, to enact roles now embodied as speaker and listener and to share conventional coordinating devices in conversation. Conversation is an activity in which the goal is to orient the listener's attention and to align intentions. The meanings expressed are constantly negotiated through dialog: adults scaffolding the children toward adult wording, children requiring clarification when they need it (Golinkoff, 1993). This takes place in communicative formats (Bruner, 1983). Linguistic interactions involve the exchange of communicative intentions through a multimodal symbolic system transmitted to children by their care-givers in daily activities and are built in the intertwining of language and action.

Children's cognitive and linguistic development centers on learning how to act and interact in the context of events, which serve as the basic unit of experience. The continuous and dynamic flow of sensation and experience is structured in terms of discrete events, which involve various participants and objects, temporal structure with a flow from beginning to end, and significant defining moments. It is the regularity and predictability of these events that allow children to master them as basic building blocks of experience; not only can they start recognizing typical and less typical examples of events, but they can gradually use them to make sense of much more complex sequences of events. They will eventually themselves learn to construct sophisticated mental structures and verbalize them. Indeed, when an event occurs repetitively in the child's daily experience, the predictable nature of the event structure provides a convenient entry point to language (Nelson, 2007). Joint parent–child action/interaction provides the scaffold for children's growing ability to grasp both what is happening around them, and what is being said in the situation. They learn to understand language and action together, each providing support for the other. The well-established language associated with simple events might provide the conceptual scaffold for the child to grasp more complex events.

A Complementary View: The Role of the Adult and the Input

Affect and Social Interaction

Tomasello et al.'s account (2005), for which cultural cognition is already founded at around 14 months, attributes complex cognitive skills to very young infants. Children's participation in interactions might be overinterpreted by scientists who could be underestimating the role of the input. For Hobson (2005), emotionally grounded sharing of experiences is primordial in order for interpersonal understanding and perspective shifting to develop. As described by Werner and Kaplan (1963), infants start from a primordial sharing situation and come to understand others' mental states as both similar to and distinct from their own, as they respond to and assimilate others' behaviors and attitudes.

Social interaction in infancy is dependent on the interplay between infants' affects, their neural learning processes, their perceptual and motor limitations and the structure of their social and affective environment. Social information helps infants decipher the meaning of others' language acts. Their drive to attend to the same objects helps them enter the language community.

Affectively valued outcomes are thought to trigger the infants' development of cognitive skills. The repetition of those outcomes facilitates the infant's attention shifting toward interesting social objects and people in the visual environment. Initially, infants prefer social stimuli such as the caregiver, but when the caregiver looks away at an object, it will habituate the infant to follow the same direction and get them attracted to other stimuli that might be positively endowed by the caregiver. The child will learn that it is often beneficial to follow caregivers' gaze shifts and learn to shift their own gaze to the same location, driven by pleasure (Triesch et al., 2006). Infants use their caregivers' emotional expressions toward objects to finely discriminate between them, something that so far, even sophisticated machines cannot do. Attention sharing will thus help children infer the mental states of their caregivers. They will use the information to shape their own communicational acts. Therefore, it is the interaction and complementarity between basic perceptual, cognitive, and affective processes that seem to trigger and guide the emergence of shared attention, which will then lead the child into symbolic communication.

When giving a bath a parent whose experience with water has been positive, who is safely holding the baby and preventing any kind of danger, whose knowledge about the sensation of the water with its right temperature provides confidence, expects the baby to enjoy the bath and its various elements. The parent can manipulate the meaning of the bath and its sensation for the child and set up a sequence of experiential histories that will provide a background for the subsequent bath experiences (Nelson, 2007). Verbal and gestural expressions convey a meaning that will come to be shared. Because they share their attention, to the same elements, that will minimize the differences between experiences (being in the bath itself as opposed to looking over the child). Sharing attention ensures that both participants experience some common

aspects of the event. Language during these experiences is an essential tool used by the caregiver to go beyond attention, and share meaning. The utterance "oh it feels so good" produced by Théophile's mother when he is 7 months old and giggling in his bath (*Paris Corpus*, Morgenstern & Parisse, 2012) makes sense of the child's individual experience and transforms it into a communal experience. It's through the repetition of the scene associated to the words that it becomes a script and that the words are transformed. In a similar fashion, Madeleine and her mother smell various flowers in their garden when the child is 11 months old (*Paris Corpus*). Mother and child are sharing sensations together; the mother actually takes the child's exact place in order to create some kind of community of sensations with her daughter. This is the basis of the building of transcendental intersubjectivity as Husserl would call it, the condition for a shared objective world and for some interdependence between those two specific individuals. Experience results from the encounter of the children with the environment, involving perception through smell, touch, taste and then action and interaction via communication. In those two examples, the children's sensations and experiences are entirely social because the adults spend considerable efforts to discern what they are feeling and thinking. However the child's experience still remains totally private, inaccessible in a way, since all the adults can do until the child expresses herself through conventional semiotic means such as gesture and speech is to make interpretations.

Adults' Interpretations

Kaye (1982) suggests that in order to understand intentions, infants must themselves be treated by adults as intentional beings. Adults must interpret their every move and action as having some meaning or some goal and provide various types of feedback to this effect. They therefore give meaning to the child's every gaze, gesture, facial expression, vocal production, projecting some kind of agency onto the infant. This begins very early on and creates the mental context for the emergence of the infant's intentional action. In the first months of life, parent–infant synchrony has a formative role in brain maturation. It was found to be predictive of children's self-regulation, symbolic play and more general cognitive skills (Feldman and Eidelman, 2004). Indeed, infants engage in 'protoconversations' from as young as 2 months old (Bateson, 1975; Trevarthen, 1974), which involve rhythmic attunement of adult and infant vocalizations. However the caregivers' role is not symmetrical as they produce amplified enhanced versions of their child's sounds (Papousek, 1989). Those exchanges are accompanied with smiles, mutual gaze, gestures that Stern (1985) has called "supportive emotional colored atunements(-*sic*)" and are therefore enriched with affect.

Parents might attribute intentionality to their infants but be deceiving themselves and contributing all the meaning in the exchange. When the child moves and seems to change her focus of attention, she might not be gazing at anything in particular, and the movement might not be intentional. However, the parent's illusion could be essential for the child's development as it bootstraps the child into our social and cultural world. Through repetition of similar situations and interpretations, the child can grow into a full-blown intentional agent.

Parents therefore seize and take up the sounds and movements produced by their children, in order to endow them with as much meaning as possible, and shape them into a form that could be compatible with the adult communicative system. In the following example taken from the Forrester corpus (CHILDES database, MacWhinney, 2000; Forrester, 2008), the father takes up his daughter's gesture, which could be interpreted as not being intentional and communicative at all, and transforms it into a game that serves as a transition toward meaning.

Example 1. Ella 1; 02

The father and the daughter are having breakfast
*FAT: Are you tired?
Ella whimpers and rubs her face.
*FAT: Oh a little bit.
She then makes a very unexpected gesture. Her hand goes down along her hair. She hits her head and looks at her father as she produces a short vocalization.
*ELLA: eh!
He takes up his daughter's gesture and points to her head
*FAT: baby's head.
He then points at his own head.
*FAT: daddy's head.

The father takes up what seems to be a nonintentional non-communicational gesture and transforms it by shaping it into a conventional pointing gesture, through which he can designate alternatively his own head and his daughter's head. He has changed it into a social gesture which is part of the string of routinely used pointing gestures of the various members of the family that Ella will take up and replay herself in the following sessions in the data.

The meaning of a large amount of communicative exchanges in general emerges from the collaborative process involved between the participants. This is even more obvious in the case of children and adults. A child might accidentally make a vocal or gestural production with no specific intention. Or she might endow it with a specific communicative intent, and not be understood by the adult. If the adult misinterprets the production, the child might refuse that interpretation but she might also accept this misconstrual and follow the adult's lead toward new directions. Conversation is therefore wholly collaborative and reanalysis of children's intentions can emerge and lead to new meanings.

Psychological states and interpretations expressed by parents interacting with children could facilitate their Vygotskyan internalization by children who simultaneously directly experience events and states and the adult's perspective on those same experiences. They therefore progressively acquire both a first- and a third-person perspective. There seems to be a collaborative engagement between adults and children in which multimodal communication and language do more than direct attention but serve as a mediation of children's mental lives and help them process complex representations (Clark, 1998). The Vygotskyan account underlines the importance of inner speech as the result of what is semiotically mediated in interaction with others. Dialogic exchanges are incorporated by the child and form inner speech. The multiple perspectives represented in dialogs or in multiparty conversations are thus internalized and construct the child as a dialogic being.

Cultural Variations

However, there are wide cultural variations in how adults interact with infants. In some cultures, adults do not treat infants as being fully intentional. Still, all typically developing children in all cultures develop an understanding of others as intentional agents. There are not enough cross-cultural studies to measure the correlation between adults' scaffolding and the development of intentionality sharing.

Ochs (1984, p. 338) argues that "the emphasis on personal intentions in Anglo society and scholarship is tied to a cultural ideology in which persons are viewed as individuals, i.e., coherent personalities who have control over and are responsible for their utterances and action s." The idea that we can have access to another person's mental life and beliefs is not shared across cultures. In the Samoan culture for instance, responsibility for people's actions is attributed to interpersonal relationships and situations.

Conclusion

The developmental path that goes from gaze following, joint attention, reading of intentions to cooperative communication including reversing perspectives expressed in the use of personal pronouns and of mental terms, involves complex connections between language, emergence of intentionality and theory of mind. Children are born into a social world in which they participate in daily conversations. In Western communities at least, adult conversationalists anticipate and explain behaviors based on desires, feelings, or beliefs. As studies based on analyses of primates as well as autistic, deaf, and typical children illustrate, the development of intentionality, of a theory of mind and of language skills therefore seem to be mutually supportive and intertwined processes.

See also: Conscious Control During Childhood, Development of; Empathy During Early Childhood Across Cultures, Development of; Executive Functions During Childhood, Development of; Intentionality During Infancy and Early Childhood, Development of; Intentionality and Rationality; Logical and Hypothetical Reasoning in Adolescence, Development of; Mind, Theories of; Neo-Piagetian Theories of Cognitive Development; Scientific Concepts During Childhood, Development of; Theory of Mind During Infancy and Early Childhood Across Cultures, Development of.

Bibliography

Bakeman, R., Adamson, L., 1984. Coordinating attention to people and objects in mother-infant and peer-infant interactions. Child Development 55, 1278–1289.
Baldwin, D.A., Baird, J.A., Saylor, M.M., Clark, M.A., 2001. Infants parse dynamic action. Child Development 72, 708–717.
Baron-Cohen, S., 1991. Precursors to a theory of mind: understanding attention in others. In: Whiten, A. (Ed.), Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading. Basil Blackwell, Oxford, pp. 233–251.
Bateson, M.C., 1975. Mother-infant exchanges: the epigenesis of conversational interaction. In: Aronson, D., Rieber, R.V. (Eds.), Developmental Psycholinguistics and Communication Disorders. New York Academy of Sciences, New York, pp. 101–113.
Bertenthal, B., 1996. Origins and early development of perception, action, and representation. Annual Review of Psychology 47, 431–459.
Bruner, J.,S., 1975. From communication to language: a psychological perspective. Cognition 3, 255–287.
Bruner, J., 1983. Child's Talk. Norton, New York.
Clark, A., 1998. Magic words: how language augments human cognition. In: Carruthers, P., Boucher, J. (Eds.), Language and Thought: Interdisciplinary Themes. Cambridge University Press, Cambridge, pp. 162–183.
Darwin, C., 1877. A biographical sketch of an infant. Mind 2, 285–294.
Feldman, R., Eidelman, A.I., 2004. Parent-infant synchrony and the social-emotional development of triplets. Developmental Psychology 40, 1133–1147.
Forrester, M.A., 2008. The emergence of self-repair: a case study of one child during the early preschool years. Research on Language & Social Interaction 41 (1), 99–128.
Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G., 1996. Action recognition in the premotor cortex. Brain 119, 593–609.
Gergely, G., Nadasdy, Z., Csibra, G., Biro, S., 1995. Taking the intentional stance at 12 months of age. Cognition 56, 165–193.
Golinkoff, R., 1993. When is communication a meeting of the minds? Journal of Child Language 20, 199–208.
Hobson, R.P., 2005. The interpersonal foundation of thinking. Behavioral and Brain Sciences 28, 703–704.
Kaye, K., 1982. The Mental and Social Life of Babies: How Parents Create Persons. University of Chicago Press, Chicago.
MacWhinney, B., 2000. The CHILDES Project: Tools for Analyzing Talk. In: The Database, third ed., vol. 2. Lawrence Erlbaum Associates, Mahwah, NJ.
Marcos, H., 1987. Communicative functions of pitch range and pitch direction in infants. Journal of Child Language 14, 255–268.
Meristo, M., Hjelmquist, E., Morgan, G., 2012. How access to language affects theory of mind in deaf children. In: Siegal, M., Surian, L. (Eds.), Access to Language and Cognitive Development. Oxford University Press, New-York, pp. 44–61.
Morgenstern, A., Parisse, C., 2012. The Paris corpus. French Language Studies 22 (1), 7–12.
Nelson, K., 2007. Young Minds in Social Worlds: Experience, Meaning, and Memory. Harvard University Press, Cambridge, MA.
Ochs, E., 1984. Clarification and culture. In: Schiffrin, D. (Ed.), GURT'84: Meaning, Form, and Use in Context: Linguistic Applications. Georgetown University Press, Washington, DC, pp. 325–341.
Papousek, M., 1989. Determinants of responsiveness to infant vocal expression of emotional state. Infant Behavior and Development 12, 507–524.
Peterson, C.C., Siegal, M., 2000. Insights into theory of mind from deafness and autism. Mind and Language 15, 123–145.
Piaget, J., 1962. Play, Dreams and Imitation in Childhood. Norton, New York.
Premack, D., Woodruff, G., 1978. Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences 4, 515–526.
Prieto, P., Estrella, A., Thorson, J., Vanrell, M.M., 2012. Is prosodic development correlated with grammatical and lexical development? Evidence from emerging intonation in Catalan and Spanish. Journal of Child Language 39 (2), 258–283.
Romanes, G., 1889. (French Translation 1891), L'évolution mentale chez l'homme. Origine des facultés humaines. Alcan, Paris.
Stern, W., 1924. Psychology of Early Childhood, German ed. 1914. Unwin, London.
Stern, D.N., 1985. The Interpersonal World of the Infant. Harvard University Press, Cambridge.
Tomasello, M., 1999. The Cultural Origins of Human Cognition. Harvard University Press, Cambridge, MA.
Tomasello, M., Rakoczy, H., 2003. What makes human cognition unique? From individual to shared to collective intentionality. Mind and Language 18, 121–147.
Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H., 2005. Understanding and sharing intentions: the origins of cultural cognition. Behavioral and Brain Sciences 28, 675–735.
Tomasello, M., Carpenter, M., Liszkowski, U., 2007. A new look at infant pointing. Child Development 78, 705–722.
Trevarthen, C., 1974. Conversations with a two-month-old. New Scientist 62, 230–235.
Triesch, J., Jasso, H., Deák, G.O., 2006. Emergence of mirror neurons in a model of gaze following. Adaptive Behavior 15, 149–165.
Vygotsky, L.S., 1962. Thought and Language. MIT Press, Cambridge, MA.
Werner, H., Kaplan, B., 1963. Symbol Formation. Erlbaum, Hillsdale, NJ.
Woodward, A.L., 1999. Infants' ability to distinguish between purposeful and non-purposeful behaviors. Infant Behavior and Development 22, 145–160.

Interactionism, Symbolic

Alex Dennis, University of Sheffield, Sheffield, UK
Greg Smith, School of Nursing, Midwifery, Social Work and Social Sciences, University of Salford, Salford, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D.A. Snow, volume 4, pp. 2212–2219, © 2001, Elsevier Ltd.

Abstract

Symbolic interactionism (SI) is a distinctive sociological perspective that stresses the analytic centrality of investigating the meanings people give to their activities. Originating US pragmatist philosophy and its uptake by staff and students working at the University of Chicago, it developed a strongly empirical approach to social life that significantly influenced a number of substantive sociological fields, including crime and deviance, education, health and illness, organizations and work. The article elaborates SI's theoretical and methodological implications, examines its relationship to other approaches, assesses the main lines of criticism directed toward it, and describes recent developments in the field.

Symbolic interactionism (SI) is a durable and distinctive sociological perspective that stresses the importance of the meanings people give to their activities. Humans make meaning by using symbols to interpret one another's actions, and act on the basis of these interpretations. In contrast to deterministic explanations of human conduct – psychological, biological, and social structural – SI is premised on the idea that adequate understanding requires close examination of the processes of communicative interaction operating in natural (nonexperimental) social settings. This article examines SI's origins, elaborates its theoretical and methodological implications, examines its relationship to other approaches, assesses its critics, and describes recent developments in the field.

Origins and Development

SI has a complex genealogy. Herbert Blumer (1937) coined the term 'symbolic interaction' to refer to the incorporation of pragmatist understandings of human conduct into sociological investigations, in particular, those being undertaken by staff and graduate students at the University of Chicago. Blumer drew on the works of William James and John Dewey, and, above all, Chicago philosopher and social psychologist George Herbert Mead. Other important contributors to this perspective include Charles Horton Cooley whose concepts of 'the looking-glass self' and 'sympathetic introspection' became key analytical resources. Georg Simmel's studies of forms of sociation were central to the theoretical development of the Chicago tradition.

Blumer, through both his published works and his supervision of generations of graduate students at Chicago (and then, after 1952, at the University of California, Berkeley), articulated the defining features of the SI approach to social investigation. It was not until the 1960s, however, that SI became recognized more broadly as a distinctive sociological school, partly through the publication of Blumer's (1969) collected papers. Empirical applications of Blumer's ideas were undertaken by members of the 'Second Chicago School' (Fine, 1995), who owed much to the writing and teaching of Blumer's Chicago colleague Everett C. Hughes (1971). Studies by Howard S. Becker (1963), Erving Goffman (1959, 1961), and Anselm L. Strauss (1959; Glaser and Strauss, 1965) provide key examples of the application of the perspective.

Blumer's account of SI is based on three 'premises' that define its scope and mark out its distinctive analytical stance: (1) people act toward things based on their meanings; (2) these meanings arise in social interaction; and (3) conveying and changing meanings demands that people define and interpret situations.

The first premise is a rejection of the idea that 'things' have fixed and immutable meanings intrinsic to them. What an object 'means' depends on the uses to which it can be put. Thus, for instance, a brick can be part of a built structure, a weight to hold something down, or even a weapon, and its sense derives from which of these functions it is serving at the time. This approach to meaning is derived from Dewey's (1896) critique of the stimulus-response model of human behavior, in which he demonstrated that all 'stimuli' could also be understood as 'responses,' and vice versa, depending on the context of understanding.

Meaning is not, however, an individual phenomenon, but rather arises in social interaction with others, according to Blumer's second premise. Becker's (1963) account of learning to enjoy smoking marijuana provides a useful example of this. New users must *learn* smoking techniques, how to recognize the drug's physiological effects, and how to enjoy these effects, through the cues and guidance of more experienced users. Being able to have 'the same' experience on the drug depends on social interaction between novice and experienced users.

Blumer's third premise (that people must define and interpret situations) is derived from Mead's (1934) distinction between gestures and acts. While animals can gesture at one another, as when a dog growls and raises its hackles at another, this behavior is not *symbolic*. Humans, on the other hand, use symbols (especially language) to take one another's points of view, to see things as others would see them. People therefore elicit the 'meaning' of behavior by working out *why* someone is doing what they are doing, what it means to them, and responding on the basis of these attributions. These reciprocal attributions and actions *comprise* the interpreted situation.

Theoretical and Methodological Implications

Blumer's premises lead to distinctive understandings of human actors, of society, and of how these can be studied. The human agent is a reflexive self, shaping and shaped by social situations, and society is construed in associational terms, as a complex network of interactions. These implications necessitate the rejection of many of the now-conventional aspects of social theory and methodology.

Causal analyses of human actions and explanations based on structural determinism must be abandoned. Action is not 'caused,' as it is always based on the actor's interpretation of his or her situation: the situation cannot *on its own* make the actor do anything. This can be demonstrated empirically: the 'same' phenomenon will elicit radically different responses from different people (or may elicit no response at all from others), and the 'same' behavior can be elicited by all manner of different influences. A decrease in interest rates, for example, might encourage some people to spend their savings and take out loans, thereby stimulating the economy. It could also, however, have the opposite effect on others; people could quite rationally increase their levels of saving in the hopes that economic volatility could lead to more significant price falls later on.

Blumer's (1956) key statement on mainstream sociological methodology was an attack on a leading form of causal modeling, *variable analysis*. Variable analysis is the (frequently statistical) attempt to explain social organization by demonstrating relationships between different abstract concepts, for example, between the level of unemployment and the rate of crime. Measuring the two, and correlating the results, might result in the finding that an increase in one 'goes with' an increase in the other, and public policy recommendations for reducing crime could follow. Blumer identifies three problems with this popular approach. First, 'generic' variables are ill-defined and unclear. What counts as 'social cohesion,' for instance, depends on the methods of study and topic of enquiry of particular pieces of research; cohesion in a family unit, for instance, could be measured by the family members' levels of agreement on certain important issues, while cohesion in a neighborhood could be measured by how many of his or her neighbors' names each person living in the neighborhood knows. Both measure *something*, but whether this is the same thing for all respondents is unclear. Second, variables may have little connection to the setting being investigated: they may simply be imported into studies as matters of theoretical interest. Many sociological studies, for instance, are simply examinations of the distributions and inter-relationships between age, gender, ethnicity, class, and so on, in novel settings, rather than examinations of the actual relevances of parties to those settings. Variables, here, do not emerge from settings but are imposed upon them. Third, the relationship between variables – the mechanisms whereby unemployment produces lower crime rates, for instance – remain unexplored, and no amount of additional 'intervening' variables can eliminate this absence of knowledge about what exactly causes what. No one can say clearly how variables *work*.

Following the pragmatists, therefore, SI adopted a policy of *radical empiricism*, evident in a suspicion of abstract and premature theorizing, and a privileging of investigations of the real world as the best means of answering sociological questions. Theory became a tool at the service of research rather than the other way round, and concerns about cumulativity were collapsed into comparative and formal sociological tropes (Becker, 2010). As Hughes (1951: 320) argued:

> The comparative student of man's work learns about doctors by studying plumbers; and about prostitutes by studying psychiatrists … all kinds of work belong in the same series, regardless of their places in prestige or ethical ratings. In order to learn, however, one must find a frame of reference applicable to all cases without regard to such ratings … Both the physician and the plumber do practice esoteric techniques for the benefit of people in distress. The psychiatrist and the prostitute must both take care not to become too personally involved with clients who come to them with rather intimate problems.

Although symbolic interactionists use a variety of investigative techniques (indeed, one group, the Iowa School, used primarily quantitative, questionnaire-based methods), there is a preference for gathering and examining qualitative data. This derives from the earlier investigations of the 'first' Chicago School, dominated by Robert Ezra Park's (Bulmer, 1984: 97) injunction to "go and get the seat of your pants dirty in real research."

The best way to 'get the seat of your pants dirty' is by undertaking *fieldwork*, sometimes called 'participant observation' or the ethnographic study of natural settings. Becker and Geer (1957) argued that, in contrast to interviews, the other dominant method of qualitative data collection, fieldwork allows researchers to more fully immerse themselves in the situations they are investigating, and thus, better understand the meanings and roles of participants. Because it typically takes place over a longer period of time, fieldwork also allows researchers to follow meanings as they change and develop in the course of their use.

The classic statement of these emphases is by Glaser and Strauss (1967: 1):

> Most writing on sociological method has been concerned with how accurate facts can be obtained and how theory can thereby be more rigorously tested. In this book we address ourselves to the equally important enterprise of how the discovery of theory from data – systematically obtained and analysed in social research – can be furthered. We believe that the discovery of theory from data – which we call *grounded theory* – is a major task confronting sociology today, for, as we shall try to show, such a theory fits empirical situations, and is understandable to sociologists and laymen alike. Most important, it works – provides us with relevant predictions, explanations, interpretations and applications.

Relationship to Ethnomethodology and Mainstream Sociology

SI's closest sociological relative is ethnomethodology (Garfinkel, 1967). Both are suspicious of premature theorizing and seek to minimize any theoretical framework to no more than is required to facilitate studies (Blumer, 1969;

Sharrock and Anderson, 1982). Both favor qualitative methods, in particular ethnographies (Becker and Geer, 1957; Garfinkel, 2002: 248). Both reject the idea that there can be an 'objective' description of social interaction: for ethnomethodologists the meaning of any action is reflexively tied to its context, with action, sense, and situation mutually elaborating one another *in situ* (Garfinkel, 1967: 3–4).

There are, however, also significant differences. These come to a head in the relationships between each perspective and mainstream sociology: practitioners of SI argue that theirs is a *better way of doing sociology*, while ethnomethodologists understand their work to be *an alternative to it*. The former have argued that their perspective addresses glaring errors in conventional sociological approaches to research (Blumer, 1956). The latter argued that their approach constitutes an 'asymmetrical' and 'alternate' sociology *tout court*, in which nothing is sacred, and even the most fundamental sociological concepts must be 'respecified' (Garfinkel, 2002: 114–120).

Mainstream sociological work has itself, over the last 20 years or so, increasingly engaged with themes central to the SI tradition. This has typically been unattributed, however, and the perspective itself is often misrepresented both in textbooks and in teaching. As Maines (2001: *xv*) points out, "sociologists over the years have learned a way of talking about themselves and their discipline … that has compartmentalized interactionist work and relegated it to the margins of scholarly consideration while simultaneously and unknowingly becoming more interactionist in their work." We shall now turn to the mainstream response to SI.

Criticisms

The conventional sociological view of SI is that it neglects considerations of both power and social structure. It is regarded as a perspective holding an unduly optimistic vision of the open society that is blind to the workings of various forms of institutionalized constraint and inequality. These criticisms miss the subtlety of some of the key statements in SI. Rather than neglecting the workings of power, SI studies show how power is manifested in everyday situations; rather than neglecting social structure, SI investigates social structure *in statu nascendi* as Simmel used to put it.

Research on deviance is perhaps the best-known area of interactionist work, particularly following the enormous success of Becker's (1963) *Outsiders*. Such work has led to a greater appreciation of the perspective of the deviant, the marginalized, the stigmatized, and from time to time such studies have contributed to political and institutional reforms. Yet the focus on the responses – both individual and collective – of those who are 'labeled' in various ways has all too often served to obscure the primary, and arguably more fundamental, concern of interactionist studies with the *authoritative* processes through which individuals are rendered subordinate through legally sanctioned and institutionally established procedures.

This approach can be understood to develop Durkheim's (1984) ideas about the cultural variability of laws and norms. Whereas Durkheim saw 'crime' in terms of those acts that offend the basic values or 'collective sentiments' of a society or group, SI seeks to avoid the reification of 'society' intrinsic to this kind of formulation. Its focus, instead, is on how actual rules are established, enforced, challenged and broken, often in situations where cultural consensus cannot be assumed:

> … laws and rules represent no group's values nor values of any portion of a society. Instead they are artefacts of compromise between the values of mutually opposed, but very strongly organized, associations.
>
> Lemert, 1972: 57

SI does not neglect power relations, but argues that they are the product of interaction rather than supra-individual 'structural' forces. Becker made this essential point clearly, starting with the Durkheimian theme that *social groups create deviance by making the rules whose infraction constitutes deviance* (Becker, 1963: 9, emphasis in original), then making explicit the sociological corollary that "deviance is *not* a quality of the act that the person commits, but rather a consequence of the application by others of rules and sanctions to an 'offender'" (Becker, 1963: 9, emphasis in original).

This position fully recognizes that there are patterns to such applications of rules. Becker (1963: 17–18) goes on to state:

> Who can, in fact, force others to accept their rules and what are the causes of their success? This is, of course, a question of political and economic power … Rules are made for young people by their elders … Men make the rules for women in our society … Negroes find themselves subject to rules made for them by whites. The foreign-born and those otherwise ethnically peculiar often have their rules made for them by the Protestant Anglo-Saxon minority. The middle class makes rules the lower class must obey – in the schools, the courts, and elsewhere … Differences in the ability to make rules and apply them to other people are essentially power differentials (either legal or extralegal). Those groups whose social position gives them weapons and power are best able to enforce their rules. Distinctions of age, sex, ethnicity, and class are all related to differences in power, which accounts for differences in the degree to which groups so distinguished can make rules for others.

The point is not that there are patterns of inequality and domination in society, a point with which symbolic interactionists would not disagree. Rather, the issue at hand is *through what mechanisms* are those patterns produced and reproduced. Here work such as Cicourel's (1968) examination of how police officers define the actions of black youths as 'threatening' while treating the 'same' actions are 'high spirits' when undertaken by their white counterparts is important. The discovery of a pattern of inequality animates studies of *how that pattern is generated* rather than representing the conclusion of a sociological investigation.

Developments

SI was always a tradition more varied than Blumer's guiding statements suggested. Not all symbolic interactionists subscribed to Blumer's views. An early division was identified between the Iowa School and the Chicago School of SI: the

former emphasizing structure and conventional social scientific precepts, the latter stressing process and negotiation and preferring qualitative methods. The Iowa group used self-report measures such as the Twenty Statements Test and favored an approach that did not place SI in opposition to standard scientific conceptions. Its leading proponents, including Manford Kuhn and Carl Couch, accepted more of the 'facts of social structure' (Stryker, 1980: 1) than the Chicago-oriented group. While the Chicago version of SI has predominated, there were disputes around the precise character of Mead's legacy. In particular, it was contended that Blumer's advocacy of qualitative methods at the expense of conventional social scientific procedures could not be legitimated by Mead's views. The dominance of the Chicagoan version of SI led some commentary to elide the notions of Chicago sociology, ethnographic research, and SI. Against this, it is important to remember that SI was not the only theoretical stream in Chicago's famed sociology department.

More recently SI addressed debates originating from cultural studies and postmodern and poststructuralist trends about how culture can be most adequately represented. Again, generalizing somewhat, there are differences between those who favor some form of realism, with an emphasis on fieldwork traditions of cultural description, and those who seek to pursue constructionist arguments to the point where questions about inscription – the deskwork ethnographers do to produce to their ethnographic analyses – precede all others. On both sides the question of representation, the sense that readers make of the texts that mediate culture to an audience, becomes salient (Van Maanen, 2011).

The development of the SI spawned a rich and diverse range of studies connected by an analytic commitment to articulate people's ordinary experience of the world. It long stood as a productive alternative to functionalist and structural sociologies. Interactionist ideas have significantly impacted upon a number of established sociological fields, such as crime and deviance, education, health and illness, organizations and work (see Atkinson and Housley, 2003; Fine, 1993; Reynolds and Herman-Kinney, 2003 for reviews).

SI has illuminated public issues. Interaction is central to understanding how some social condition comes to be defined as a social problem. This is an often contested matter where groups advancing competing claims (Spector and Kitsuse, 1977) about the characteristics of the problem use a range of resources – material, symbolic, and political – to persuade publics and legislators of the legitimacy of their definitions (Holstein and Miller, 2003). Here Goffman's (1974) frame metaphor has proved enormously influential in analyses of how issues and problems are defined as such.

SI played a major part in opening up new fields, notably the sociology of emotion (Hochschild, 1983) and the sociology of public places (Lofland, 1998). People's experienced emotions are intimately connected to situated interaction. Emotions are not simply biological and psychological phenomena; they are socially organized by 'feeling rules' enacted in contexts of everyday interaction. Similarly, the investigation of public places has shown that the conduct there is far from asocial. Goffman's (1983) work on the 'interaction order' demonstrated how public places are delicately ordered by local rules that regulate territoriality, civil inattention, forms of regard for the other, the expression of information about the self, and the like. W.I. Thomas's famed apothegm that if people define things as real, they are real in their consequences, justified the close study of subjectivities and catalyzed a series of interactionist-inspired studies of ordinary action. Goffman's (1974) frame analysis offered a systematic and admittedly formalistic answer to the fundamental practical problem, "what is it that is going on here?"

Ethnographic research addressed the same question more discursively and with attention to the empirical details of situated symbolism and the emergent meanings arising in natural social settings. Directly or indirectly, ethnographic studies have drawn upon the SI tradition, seeking new ways to uncover and represent the 'actor's point of view' – the experience of people as they go about their daily lives. New standards of the intensity and extensiveness of ethnographic investigations have been set, clearly marking out such work from jibes about 'tenured journalism' (Fine, 1993). Observation of people's activities over extended periods of time is one feature of interactionist-inspired ethnographies of Black American inner-city street life – Anderson (1990) was the result of 14 years' research. Anderson's (1999) subsequent book also drew on that fieldwork and another 4 years of ethnographic research in poor and well-to-do neighborhoods. Duneier's (1999) ethnography of street vendors was based on more than 3 years' fieldwork. Duneier's study was also notable for his close collaboration with his key informant and a noted photojournalist, for its pioneering audio recording of street encounters, the quasi-legalistic standards of evidence and proof applied to analytic inferences drawn, and the sharing of the financial proceeds of the book with the research participants. Similarly, Alice Goffman's (2014) study of young black parolees was based on 6 years of immersive fieldwork. Studies such as these exemplify the sustained commitment of ethnographers to depict accurately the viewpoints of research participants and to entertain seriously inconvenient facts.

Over the past quarter of a century SI has proved a tradition remarkably open to postmodern and post structuralist critiques and developments in the social sciences and humanities, including cultural studies (Becker and McCall, 1990). Several aspects of these intellectual shifts are incorporated into SI by Denzin's (2001) 'interpretive interactionism,' which explores new ways of making people's problematic experience available to the reader by artistic and social scientific methods. The examination of people's reflections about these problematic experiences ('epiphanies') was also pursued by autoethnography, a method where the ethnographer explores their own personal experience (see Ellis et al., 2010 for a review of controversies about the method). Autoethnography has proved especially effective in conveying the personal and cultural meanings encountered in challenging situations, such as facing life-threatening illnesses (Frank, 1991) or learning to become a competition-standard boxer (Wacquant, 2004).

Appraisals of SI's future (Fine, 1993; Maines, 2001) have noted its distinguished past but, prudently, have been ambivalent about its future. The SI tradition has expanded, diversified, and become internationalized. There is an important sense in which it has become mainstream and is no longer

a 'loyal opposition.' As Atkinson and Housley (2003) put it, "we are all interactionists now." Well, perhaps. SI's abiding appeal is likely to continue to be its resolutely empirical approach to the social organization of people's experiences.

Bibliography

Anderson, Elijah, 1990. Streetwise: Race, Class, and Change in an Urban Community. University of Chicago Press, Chicago.

Anderson, Elijah, 1999. Code of the Street: Decency, Violence and the Moral Life of the Inner City. Norton, New York.

Atkinson, Paul, Housley, William, 2003. Interactionism. Sage, London.

Becker, Howard S., 1963. Outsiders: Studies in the Sociology of Deviance. Free Press, New York.

Becker, Howard S., 2010. The art of comparison: lessons from the master, Everett C. Hughes. Sociologica 2010 (2), 1–12.

Becker, Howard S., Geer, Blanche, 1957. Participant observation and interviewing: a Comparison. Human Organization 16 (3), 28–32.

Becker, Howard S., McCall, Michal (Eds.), 1990. Symbolic Interaction and Cultural Studies. University of Chicago Press, Chicago.

Blumer, Herbert, 1937. Social Psychology. In: Schmidt, Emerson Peter (Ed.), Man and Society: A Substantive Introduction to the Social Sciences. Prentice-Hall, New York, pp. 144–198.

Blumer, Herbert, 1956. Sociological analysis and the "variable". American Sociological Review 21 (6), 683–690.

Blumer, Herbert, 1969. Symbolic Interactionism: Perspective and Method. Prentice-Hall, Englewood Cliffs, NJ.

Bulmer, Martin, 1984. The Chicago School of Sociology: Institutionalization, Diversity, and the Rise of Sociological Research. University of Chicago Press, Chicago.

Cicourel, Aaron V., 1968. The Social Organization of Juvenile Justice. Wiley, New York.

Duneier, Mitchell, 1999. Sidewalk. Farrar Strauss Giroux, New York.

Denzin, Norman K., 2001. Interpretive Interactionism, second ed. Sage, Thousand Oaks, CA.

Dewey, John, 1896. The reflex arc concept in psychology. Psychological Review 3 (4), 357–370.

Durkheim, Émile, 1984. The Division of Labour in Society. Macmillan, Basingstoke.

Ellis, Carolyn, Adams, Tony E., Bochner, Arthur P., 2010. Autoethnography: an overview. Forum Qualitative Sozialforschung/Forum: Qualitative Social Research 12 (1). Art. 10. http://nbn-resolving.de/urn:nbn:de:0114-fqs1101108.

Fine, Gary A., 1993. The sad demise, mysterious disappearance, and glorious triumph of symbolic interactionism. Annual Review of Sociology 19, 61–87.

Fine, Gary A. (Ed.), 1995. A Second Chicago School? The Development of a Postwar American Sociology. University of Chicago Press, Chicago.

Frank, Arthur W., 1991. At the Will of the Body: Reflections on Illness. Houghton Mifflin, Boston.

Garfinkel, Harold, 1967. Studies in Ethnomethodology. Prentice-Hall, Englewood Cliffs, NJ.

Garfinkel, Harold, 2002. Ethnomethodology's Program: Working Out Durkheim's Aphorism. Rowman and Littlefield, New York.

Glaser, Barney G., Strauss, Anselm L., 1965. Awareness of Dying. Aldine, Chicago.

Glaser, Barney G., Strauss, Anselm L., 1967. The Discovery of Grounded Theory: Strategies for Qualitative Research. Aldine, Chicago.

Goffman, Erving, 1959. The Presentation of Self in Everyday Life. Doubleday Anchor, New York.

Goffman, Erving, 1961. Asylums: Essays on the Social Situation of Mental Patients and Other Inmates. Doubleday Anchor, New York.

Goffman, Erving, 1974. Frame Analysis: An Essay on the Organization of Experience. Harvard University Press, Cambridge, MA.

Goffman, Erving, 1983. The interaction order. American Sociological Review 48 (1), 1–17.

Goffman, Alice, 2014. On the Run. University of Chicago Press, Chicago.

Hochschild, Arlie Russell, 1983. The Managed Heart: Commercialization of Human Feeling. University of California Press, Berkeley.

Holstein, James A., Miller, Gale (Eds.), 2003. Challenges and Choices: Constructionist Perspectives on Social Problems. Aldine de Gruyter, New York.

Hughes, Everett C., 1951. Mistakes at work. Canadian Journal of Economics and Political Science 17 (3), 320–327.

Hughes, Everett C., 1971. The Sociological Eye: Selected Papers. Aldine-Atherton, Chicago.

Lemert, Edwin M., 1972. Human Deviance, Social Problems, and Social Control. Prentice-Hall, Englewood Cliffs, NJ.

Lofland, Lyn H., 1998. The Public Realm: Exploring the City's Quintessential Social Territory. Walter de Gruyter, New York.

Maines, David R., 2001. The Faultline of Consciousness: A View of Interactionism in Sociology. Aldine de Gruyter, New York.

Mead, George Herbert, 1934. Mind, Self, and Society from the Standpoint of a Social Behaviorist. University of Chicago Press, Chicago.

Reynolds, Larry T., Herman-Kinney, Nancy J. (Eds.), 2003. Handbook of Symbolic Interactionism. Rowman and Littlefield, Lanham, MD.

Sharrock, W.W., Anderson, R.J., 1982. On the demise of the native: some observations on and a proposal for ethnography. Human Studies 5 (1), 119–135.

Spector, Malcolm, Kitsuse, John I., 1977. Constructing Social Problems. Benjamin/Cummins, Menlo Park, CA.

Strauss, Anselm L., 1959. Mirrors and Masks: The Search for Identity. Free Press, Glencoe, IL.

Stryker, Sheldon, 1980. Symbolic Interactionism: A Social Structural Version. Benjamin/Cummins, Menlo Park, CA.

Van Maanen, John, 2011. Tales of the Field: On Writing Ethnography, second ed. Chicago University Press, Chicago.

Wacquant, Loïc, 2004. Body and Soul: Notebooks of an Apprentice Boxer. Oxford University Press, New York.

Relevant Websites

http://dne2.ucsf.edu/public/anselmstrauss/index.html – Anselm L. Strauss resources.
http://home.earthlink.net/~hsbecker/ – Howie's Page.
https://www.brocku.ca/MeadProject/ – The Mead Project.
http://plato.stanford.edu/entries/mead/ – Mead Entry in Stanford Encyclopedia of Philosophy.
http://symbolicinteraction.org – Society for the Study of Symbolic Interaction.
http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%291533-8665 – Symbolic Interaction Journal Website.

Intercultural Communication: An Anthropological Perspective

Pál Nyíri, Vrije Universiteit, Amsterdam, The Netherlands
Joana Breidenbach, Betterplace Lab, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article is a critical review of intercultural communication (IC), an applied academic field whose goal is to understand how members of different cultural – usually national or ethnic – groups communicate, behave, and perceive the world, as well as to facilitate communication between them. The article traces the history and current state of IC, both as an academic field and as a service industry; examines the concepts of culture upon which its assumptions rest; and offers a critical perspective on its usefulness and limitations.

Intercultural communication (IC) is an applied academic field with applications particularly in management, education, and health care. It seeks to understand how members of different cultural – usually national or ethnic – groups communicate, behave, and perceive the world in order to facilitate communication between them. In academic institutions, IC is usually taught as part of courses in communication studies, management, or languages, and sometimes as a separate field of study. Outside academia, companies offer consultancy and trainings in IC, publishing houses specialize in IC texts, and business magazines run IC columns. The clients of IC consultancies include corporations, government institutions, hospitals, and militaries.

Origins

The term 'intercultural communication' was used in anthropologist Edward T. Hall's best-selling 1959 book *The Silent Language*, regarded as the 'founding document' of the field (Rogers et al., 2002: p. 11). At the request of the Foreign Service Institute, Hall developed a practical approach to sensitize future diplomats to cultural differences that obstruct communication. He divided societies into those with high vs low 'context,' those with 'monochronic' vs 'polychronic' time systems, those with higher and lower needs for personal space, and those with higher and lower speed of messages (referring to the formation of social networks). 'Low context' societies were those where people compartmentalized interaction in various personal and professional settings, and since that meant having less background knowledge about those outside their immediate circle of relatives or friends, they required more information to complete a task. (Thus, a Chinese employee could be given much simpler, verbal instructions than an American because the intense social interaction between manager and employee would create much tacit understanding between them.) Similarly, a 'monochronic' time system meant a compartmentalized management of tasks – where scheduling and precision punctuates the natural rhythm of human interaction, distinguishing work time from leisure – while 'polychronic' referred to dealing simultaneously with a range of matters and people and to a greater tolerance of delays.

Hall's views were inspired by the Culture and Personality School of Anthropology that flourished at Columbia University in the 1940s and 1950s, and included Margaret Mead and Ruth Benedict, as well as by Freudian psychoanalysis and the Sapir-Whorf school of structural linguistics (Rogers et al., 2002: p. 5). As Mitchell Hammer, an intercultural trainer, wrote, anthropology at the time described "patterns of human interaction as it exists within a particular group of people and within another group." In Hammer's view, Hall's innovation was to look at the interaction between human groups: "that's what intercultural communication was about" (Dahlén, 1997: p. 102).

Hall was the first to call the attention of nonanthropologists to the culturally constructed nature of ideas about time, privacy, and nonverbal communication. He did not offer simple, quantitative matrices to apply to the societies he described and did not explicitly divide the world into regions colored differently based on whether they were, for example, 'high or low context.' He also did not regard IC as a separate academic field. But Hall's description of culture as a 'silent language,' a code that could be cracked and taught; his use of untranslated concepts from the Japanese, which had an exoticizing effect; his general statements about 'the Americans' and 'the French'; and his insistence that 'like oil and water, the two systems' (monochronic and polychronic time) 'do not mix' lay the groundwork for the conceptual framework of the field of IC as it emerged in the 1960s and 1970s. The first IC textbook was published in the United States in 1966, and the Society for Intercultural Education, Training, and Research, which remains the leading international body of IC, was founded in 1974.

In the 1980s, IC incorporated concepts and methods taken from social psychology, a discipline that relies on the quantitative analysis of large-scale opinion surveys and on controlled experimentation in laboratory settings, and cross-cultural psychology. Cross-cultural psychology, a subfield of social psychology that acquired its own identity in the early 1970s, proceeds by identifying a number of universal dimensions of behavior, such as individualism or respect for authority, and tests them comparatively across different groups, revealing quantitative variations in reactions to particular situations through statistical methods such as factor analysis. For example, Harry Triandis, a psychologist with major influence in IC, described 'cultures' on value scales such as 'tightness,'

individualist vs collectivist, vertical vs horizontal, active vs passive, universalist vs particularist, diffuse vs specific, instrumental vs expressive, and emotionally expressive vs suppressive (Triandis, 1994). Such sets of cultural variables are sometimes derived from cross-national surveys, but very often they are simply postulated. In other words, while cross-cultural psychology generally assumes that particular cultures (i.e., nations or ethnic groups) can be distinguished by core sets of values, it is interested not in establishing what these values might be (and hence, whether they indeed form a coherent set or not) but in how behavior varies across what are assumed to be cultures driven by different values.

Key Concepts

The most-cited IC study is Geert Hofstede's survey of work-related values among 88 000 IBM employees in over 50 countries, conducted between 1967 and 1973 and published as *Culture's Consequences* in 1980. Claiming that "data obtained within a single MNC [multinational corporation] does have the power to uncover the secrets of entire national cultures" (Hofstede, 1980a: p. 44), Hofstede, trained as a social psychologist, proceeded to develop a system of 'cultural dimensions' that identifies and quantifies 'value orientations' – a concept originally developed by Kluckhohn and Strodtbeck (1961) – in different societies. In a frequently quoted definition, he saw culture as "the collective programing of the mind that distinguishes the members of one group or category of people from others" (Hofstede and Hofstede, 2004: p. 4). This 'programing' is acquired through socialization in a particular country (Hofstede, 1980b: p. 14).

The six cultural dimensions, measured on a scale from 0 to 100, are 'power distance,' collectivism–individualism, masculinity–femininity, uncertainty avoidance, long/short-term orientation, and indulgence–restraint. (Hofstede added the last two later.) To focus on one of these variables, the 'collectivism' dimension means an orientation toward the expectations and norms of the we-group (such as kin) and strict differentiation between it and outsiders. By contrast, in individualistic societies, individuals are expected to be concerned primarily about themselves and their immediate family members, and the distinction between the in- and the out-group is weak. 'Power distance' denotes acceptance of hierarchy vs preference for egalitarianism. Thus, Americans are individualists with a small power distance; Chinese, collectivists with a large power distance.

Another influential IC author and consultant, Fons Trompenaars, uses seven rather than six 'cultural dimensions' (see Table 1). Trompenaars' study, published in his book *Riding the Waves of Culture*, was based on a questionnaire administered in 1993 to 15 000 managers and administrative staff working in corporations in 50 countries. He found that Americans and Europeans are 'universalistic' while Chinese are 'particularistic.' This is because Westerners think in a logical and linear fashion and make sense of the world using abstract concepts. Chinese thinking, on the other hand, is intuitive, nonlogical and nonabstract. Dutch and Swedes are 'neutral' because they consider the control of emotions as a sign of civilization and deem it inappropriate to exhibit them in work relations, whereas Southern Europeans are 'affective' because they show their feelings openly in the workplace as well as at home.

Table 1 Trompenaars Hampden-Turner's 'seven dimensions of culture'

>>>Universalism vs Particularism
What is more important – rules or relationships?
>>>Individualism vs Communitarianism
Do we function in a group or as an individual?
>>>Specific vs Diffuse cultures
How far do we get involved?
>>>Affective vs Neutral cultures
Do we display our emotions?
>>>Achievement vs Ascription
Do we have to prove ourselves to receive status or is it given to us?
>>>Sequential vs synchronic cultures
Do we do things one at a time or several things at once?
>>>Internal vs External control
Do we control our environment or work with it?

Source: http://www.thtconsulting.com/Content/cont042.htm, accessed 17 January 2007.

Not all IC authors use Hofstede's or Trompenaars' categories, but most popular IC texts (Chaney and Martin, 2004; Gudykunst and Kim, 2002; Jandt, 2006; Lustig and Koester, 2005; Martin and Nakayama, 2003; Samovar et al., 2009; Varner and Beamer, 2005) operate with a similar set of dimensions and values. In the late 1980s, academic and corporate interest in the so-called 'Asian tigers' led to a shift in focus to Asia in the work of IC theorists. Taiwan, Hong Kong, and Singapore had all scored high in the dimensions of Collectivism and Power Distance in the Hofstede survey, but the base in both Singapore and Hong Kong had been too small for statistical analysis. A Chinese Values Survey, conducted in 23 countries by social psychologist Michael Harris Bond in consultation with Hofstede and scholars from Hong Kong and Taiwan, confirmed Hofstede's earlier findings and, additionally, came up with a set of values described as Confucian Dynamism, on the basis of which Hofstede developed the dimension he called long-term orientation. Confucian Dynamism entailed a long list of values, including persistence, industriousness, benevolent authority, loyalty to superiors, harmony with others, wealth, pragmatism, thrift, wealth, patience, respect for hierarchy and tradition, protecting 'face,' resistance to corruption, family orientation, high work morale, and patriotism (Bond, 1986: pp. 9–10). In a subsequent article, Hofstede and Bond (1988) directly linked Confucian dynamism to economic growth.

By the late 1990s, the interest of some IC theorists and practitioners began shifting away from determining group characteristics and toward measuring, and, through various trainings, expanding 'cultural' or 'intercultural competence,' a set of skills that defines an individual's ability to work with cultural difference. The Intercultural Development Inventory, developed by Milton J. Bennett and Mitchell Hammer, for example, locates individuals at one of the six stages of intercultural sensitivity, from 'denial' to 'integration' of cultural difference, on five different scales corresponding to particular cognitive mechanisms. Leading authorities on diversity management such as

Taylor Cox, Jr., tend to focus on differences in perception, influenced by the relative social status of two individuals as much as their 'core values.' Thus, Cox (1994: p. 142ff) discusses the differing evaluation of a promotion policy by black and white American employees in a company and interprets them primarily as a result of minority–majority power dynamics. Nonetheless, Cox's discussion of culture relies on 'dimensions' as developed by Hofstede, Hall, Bond, and others.

Applications

IC consultancies provide their services to both large and small businesses, as well as government institutions. Typically, these take the form of trainings or tests, often built on a 'critical incident,' a situation in which unrecognized cultural difference causes a breakdown in communication. Most trainings are specific to a particular national culture – in the context of a relocation, merger, or negotiation process – and provide clients with a compass or kit, often based on Hofstede's or Trompenaars' dimensions. Trainings that focus on cultural sensitivity place less emphasis on 'cracking cultural codes' and more on making individuals aware of the diverse factors that influence the behavior and judgment of both their business partners and themselves. These factors may include socioeconomic status, family history, or individual psychological traits.

Public health has been a major customer of 'cultural competence' trainings in North America since the 1990s, influenced not only by a concern with inequalities in health indicators and access to health care, but also by anthropological perspectives on the cultural contexts of health and illness. Medical anthropologist Arthur Kleinman (2004: p. 951) summed up the new public health paradigm as recognizing that "culture influences the experience of symptoms, the idioms used to report them, decisions about treatment, doctor–patient interactions, the likelihood of outcomes such as suicide, and the practices of professionals." This paradigm fitted in a broader trend toward 'patient-centered care.'

The U.S. Liaison Committee on Medical Education implemented a standard for cultural competence in 2000, defining it as the requirement for "faculty and students [to] demonstrate an understanding of the manner in which people of diverse cultures and belief systems perceive health and illness and respond to various symptoms, diseases, and treatments," as well as "to recognize … gender and cultural biases in health care delivery" (Association of American Medical Colleges, 2005: p. 1). In 2004, a report to the British government recommended mandatory cultural competence training for all managers and clinical staff at mental health institutions. Generally, 'cultural competence' programs and manuals endeavor to accomplish three aims: to provide information on particular 'racial' or ethnic groups, their health status and their traditions of medicine; to provide practical information on matters ranging from diet to mourning customs; and to increase the cultural sensitivity of practitioners.

A large number of trainings, especially in Europe, focus on Muslim patients. While some medical organizations warn against group stereotyping and emphasize the need for sensitivity to individual differences, most manuals are "primarily lists of characteristics for particular races and ethnic groups" (Lee and Farrell, 2006: p. 9) and treat individuals as carriers of these collective traits.

The Context of IC's Rise and Global Spread

As evidenced by the Library of Congress' holdings, publications in IC first appeared in the 1950s, and their number has since grown exponentially (Piller, 2011: p. 30). As Piller notes, the rise of IC in the United States corresponded to "the rise of a US perception of the world in terms of international competition and threat" (*ibid.*: p. 33). The Foreign Service Institute, where Edward Hall worked, grew out of a set of training programs for military personnel during World War II (*ibid.*: p. 31). In the late 1960s and 1970s, the spread of IC was mostly fueled by the emergence of Japanese companies as competitors to American ones. A Japanese translation of Edward Hall's *The Silent Language* was published as early as 1966, and universities established IC departments.

From the following decade on, the growth of IC has been increasingly linked to the emergence and spread of the contemporary multinational corporation. By the time SIETAR, the Society of Intercultural Education, Training and Research, was founded in the United States in 1974 (the first SIETAR branch outside the US was set up in Japan in 1985), an increasing number of – mostly American – businesses were operating multinationally: relying on markets or suppliers in countries other than where they are headquartered; having ethnically diverse and geographically dispersed workforces; and engaging in mergers and acquisitions abroad.

In the 1990s, with worldwide market deregulation, the global spread of consumption patterns, and the fragmentation of the production process including the outsourcing of back-office work, the globalization of corporations took on a new scale. In addition, dealing with cultural difference became part of management tasks, both in product development and marketing, and in personnel management. By the end of the decade, ethnic diversity of the workforce became both a mainstream political and often legal expectation in North America and Western Europe and a management fad. Today, large companies expect their employees to understand markets, consumer behavior, and workplace relations across borders. Meanwhile, they are confronted with the cultural pluralization of their workforce as well as their shareholders. Cultural difference, seen in the early 1990s as a technical obstacle to maximizing productivity, that good management should overcome, is nowadays often considered an asset that nonetheless should be properly managed.

Meanwhile, due to both management and policy trends that included a more explicit approach to ethnic diversity, IC took root in Europe (a SIETAR branch was established in 1991). IC chairs have been established at several European universities, and the subject of 'IC' has even become part of the German school curriculum. SIETAR branches were established in India in 2006 and in the Middle East in 2007, signaling the growing emphasis on global business in these places. Books by Hofstede and Trompenaars are now used to teach students from Chemnitz Technical University in eastern Germany to Hosei University in Japan and King Saud University in Saudi Arabia. More broadly, the continuing spread of IC is buoyed by the

currency of cultural explanations to account for a wide range of international, corporate, social, and personal conflicts in the post-Cold War world (Breidenbach and Nyíri, 2009).

Critique

A popular reader in IC now in its 12th edition defines culture as *the rules for living and functioning in society* (Samovar et al., 2009: p. 10, emphasis in original). It goes on to state that "cultural identity is derived from our 'sense of belonging to a particular cultural or ethnic group,'" that "*culture is learned*" and that "*culture is transmitted intergenerationally*" (*ibid.*: p. 11). Mainstream IC understands the world as a mosaic of distinct national societies, each of which has a set of rules that are, at their core, resistant to change. In an oft-quoted 'model,' Trompenaars compares cultures to onions (Trompenaars and Hampden-Turner, 1998: p. 6). This national cultural determinism is the main conceptual weakness of mainstream interculturalism. It reflects the inheritance of anthropology before the fundamental changes of the 1970s and the 1980s, when – as Clifford Geertz famously wrote in 1973 – it began seeing itself no longer as "an experimental science in search of law but an interpretive one in search of meaning" (Geertz, 1973: p. 5). A decade later, in a book that was hugely influential for the subsequent direction anthropology took, James Clifford turned radically against the culture concept Edward Hall had used: culture, he wrote, was "neither an 'object to be described' nor a 'unified corpus of symbols and meanings that can be definitively interpreted'" but "contested, temporal, and emergent" (Clifford, 1986: pp. 13, 19). Cultural anthropologists now broadly agree that cultures previously depicted as static and monolithic entities have in fact always incorporated outside influences, and that beneath their facades of homogeneity lie conflicts between individuals and groups who had quite different views on and uses for common cultural practices.

Similarly, linguistics "have long since moved away from structuralism with its emphasis on discrete units" (Piller, 2011: p. 32), which defined Hall's approach and continues to be influential in IC. Some cross-cultural psychologists have also been critical of national cultural dimensions failing to pay attention to individual variation, power structures, conflicts within groups, and historical change (Gjerde, 2004).

The empirical basis of mainstream IC's national typology is also problematic. Hofstede's IBM study – one of the discipline's foundational texts – was done on a relatively small sample in most of the included countries: only in six countries (five of them in Western Europe) did the number of respondents exceed 1000. Numerous other surveys have come to conclusions in direct opposition to Hofstede's making; one found that, among respondents in Japan, Korea, Taiwan, and Denmark, 'hard work' was least important for the Japanese, 'creativity' was rated highest in Korea and lowest in Denmark, and 'obedience' was more important for Danes than for any of the Asians (Helgesen, 2001). Aside from sampling problems, such inconsistence shows that questionnaires and formal interviews reflect how people see themselves and want to be seen by others, not how they actually behave. Surveys also fail to capture the influence of politics and the historical moment on *both* what people say and what they do. Cultural dimensions, postulated as an unchanging core, are ill-equipped to deal with change, whether political, social, or organizational. In the context of health care, where the focus is often on migrants, critics point out that the cultural 'trait lists' (Kleinman and Benson, 2006) rely on frequently antiquated descriptions of societies of origin, to the point where doctors are sometimes misled into looking for a cultural rationale where there is none.

IC studies that deal with differences between organizational cultures more often produce useful insights because they operate on a tangible scale of day-to-day encounters and allow to see how individuals deploy cultural difference (or similarity) strategically to achieve their aims. Yet even where the intention is to sensitize people to dealing with cultural difference (and its conscious manipulation), IC trainers often fall back on the schematic world of 'cultural dimensions.' For example, a study of US–Mexican joint ventures showed that, contrary to predictions made on the basis of Hofstede's and Trompenaars' dimensions, Mexican workers welcomed participatory management techniques introduced by the Americans, but attempts to explain these reactions within the framework of those dimensions (Rao and Teegen, 2001).

Recent Developments

Some IC scholars and practitioners are critical of the 'dimensions' approach. Alois Moosmüller, a professor of IC at the University of Munich, points out the danger of seeing culture as primordial and has devoted research to the effects of national stereotyping in German–Japanese joint ventures (Moosmüller, 2000). A more radical revisionist movement within IC has come from sociolinguistics and discourse analysis. In an IC textbook published in three editions since 1995, Scollon et al. (2012) focus on 'discourse systems' rather than 'cultures' that exist prior to discursive construction. They point out that all communication is in a sense 'intercultural,' that the 'culture' in question depends both on the multiple situatedness of individuals in various social categories and on their choice of which available discourse to deploy, and that 'avoiding miscommunication' requires an understanding of the situations from which individuals speak. In another critical text, Ingrid Piller (2011: p. 2) points out the paradox that the culturally 'pure' cast of characters IC usually operates with, does not reflect the profiles of most IC practitioners or learners. Following critics such as Brendan McSweeney (2002), Piller (2002: pp. 77–83) sees the work of Hofstede and his followers as conceptually unsound, practically useless (because of its inability to account for the internal heterogeneity of societies: an IC training on doing business with Germans would have no place for the 60 000 German businesses owned by ethnic Turks, for example) and morally reprehensible (because it reinforces stereotypes). She proposes a re-orientation of IC from its central preoccupation with group characteristics to a "material-linguistic" focus on "Who makes culture relevant to whom in which context for which purposes?" (Piller, 2011: p. 172).

Studies in this vein of IC use ethnographic or discourse analysis approaches. Some find that conflicts or

misunderstandings attributed to cultural difference stem not from ethnic fault lines but from conflicts of interest or differing central values held by subgroups within an organization, as in the case of a high-tech company operating in Germany and India (Mahadevan, 2007). Other studies focus on the consequences of language policy choices within multinational companies for relations between employees (Piller, 2011: pp. 86–87). While such studies may reflect the complex causes of conflict more accurately, it is as yet unclear whether they can take the place of Hofstedean approaches, which appear to provide easy and affordable answers to existing corporate needs.

See also: Culture: Contemporary Views; Ethnocentrism; Multiculturalism; National Character; Organizational Culture, Anthropology of; Psychological Anthropology; Values, Anthropology of.

Bibliography

Association of American Medical Colleges, 2005. Cultural competence education for medical students. Online https://www.aamc.org/download/54338/data/ (accessed 24.05.07).

Baskerville, Rachel F., 2003. Hofstede never studied culture. Áccounting, Organizations and Society 28, 1–14.

Bond, Michael Harris, 1986. The Psychology of the Chinese People. Hong Kong University Press, Hong Kong.

Breidenbach, Joana, Nyíri, Pál, 2009. Seeing Culture Everywhere… From Genocide to Consumer Habits. University of Washington Press, Seattle.

Chaney, L.H., Martin, J.S., 2004. Intercultural Business Communication, third ed. Pearson Education, London.

Clifford, James, 1986. Partial truths. In: Clifford, James, Marcus, George (Eds.), Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley, pp. 1–26.

Cox Jr., Taylor, 1994. Cultural Diversity in Organizations: Theory, Research and Practice. Berrett-Koehler, San Francisco.

Dahlén, Tommy, 1997. Among the Interculturalists: An Emergent Profession and its Packaging of Knowledge. Stockholm Studies in Social Anthropology, Stockholm no. 38.

Geertz, Clifford, 1973. The Interpretation of Cultures. Basic Books, New York.

Gjerde, Per, 2004. Culture, power, and experience: toward a person-centered cultural psychology. Human Development 47 (3), 138–157.

Gudykunst, William B., Kim, Y.Y., 2002. Communicating with Strangers: An Approach to Intercultural Communication, fourth ed. McGraw Hill, New York.

Helgesen, Geir, 2001. The 4th EPCReN workshop: perceptions of good government. IIAS Newsletter 24, 33.

Hofstede, Geert, 1980a. Motivation, leadership and organization: do American theories apply abroad? Organizational Dynamics. Summer, 42–63.

Hofstede, Geert, 1980b. Culture's Consequences: International Differences in Work-related Values. SAGE, Newbury Park, California.

Hofstede, Geert, Bond, Michael Harris, 1988. The confucius connection: from cultural roots to economic growth. Organizational Dynamics 16 (4), 4–21.

Hofstede, Geert, Hofstede, Gert Jan, 2004. Cultures and Organizations: Software of the Mind. McGraw-Hill, New York.

Jandt, F.E., 2006. An Introduction to Intercultural Communication: Identities in a Global Community, fifth ed. Sage, London.

Kleinman, Arthur, 2004. Culture and depression. New England Journal of Medicine 351 (10), 951–953.

Kleinman, Arthur, Benson, Peter, 2006. Anthropology in the clinic: the problem of cultural competency and how to fix it. PLoS Medicine 3 (10), 294.

Kluckhohn, Florence, Strodtbeck, Fred, 1961. Variations in Value Orientations. Row & Peterson, Evanston, Ill.

Lee, S. Agnes, Farrell, Michelle, 2006. Is cultural competency a backdoor to racism? Anthropology News 47 (3), 9–10.

Lustig, M.W., Koester, J., 2005. Intercultural Competence: Interpersonal Communication Across Cultures, fifth ed. Allyn and Bacon, Boston.

Mahadevan, Jasmin, 2007. Beyond national culture? Deutsch–Indische Ingenieursarbeit aus ethnologischer Sicht (German–Indian engineering work from an anthropological perspective). SIETAR Journal 1, 11–14.

Martin, J.N., Nakayama, T.K., 2003. Intercultural Communication in Contexts, third ed. McGraw Hill, New York.

McSweeney, Brendan, 2002. Hoofstede's model of national cultural differences and their consequences: a triumph of faith—a failure of analysis. Human Relations 55 (1), 89–118.

Moosmüller, Alois, 2000. Der Kulturbegriff aus ethnologischer Sicht (The culture concept from an anthropological perspective). Newsletter (SIETAR Deutschland) 1, 6–7.

Piller, Ingrid, 2011. Intercultural Communication: A Critical Introduction. Edinburgh University Press, Edinburgh.

Rao, Pramila, Teegen, Hildy, 2001. Human Resource Issues: U.S.–Mexico Joint Ventures. In: Paper Presented at the 2001 Iberoamerican Academy of Management, Mexico City, 10–12 December.

Rogers, Everett M., Hart, William B., Miike, Yoshitaka, 2002. Edward T. Hall and the history of intercultural communication: the United States and Japan. Online. Keio Communication Review 24 (accessed 24.03.12) http://www.mediacom.keio.ac.jp/publication/pdf2002/review24/2.pdf.

Samovar, Larry A., Porter, Richard E., McDaniel, Edwin R., 2009. Intercultural Communication, twelth ed. Wadsworth Cengage Learning, Boston.

Scollon, Ron, Wong Scollon, Susanne, Jones, Rodney, 2012. Intercultural Communication: A Discourse Approach, third ed. John Wiley & Sons, Chichester.

Triandis, Harry C., 1994. Culture and Social Behavior. McGraw Hill, New York.

Trompenaars, Fons, Hampden-Turner, Charles, 1998. Riding the Waves of Culture: Understanding Diversity in Global Business. McGraw Hill, Columbus.

Varner, I., Beamer, L., 2005. Intercultural Communication in the Global Workplace, third ed. McGraw Hill, New York.

Intercultural Legal Discourse: Philosophical Aspects

Otfried Höffe, Universität Tübingen, Tübingen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article focuses on the concept of intercultural legal discourses on morality, justice, and law. These discourses demand that other cultures are to be taken seriously in their own distinctiveness. Nevertheless, this does not lead into relativism. Instead, these intercultural discourses take on a universalist point of view while retaining an ethics of respect toward different cultures. To provide an example for such a discourse, the article examines the question if there is an intercultural criminal law.

The Basic Idea

The intercultural discourse on morality, justice, and law is neither limited to a particular society, nor is it committed to certain presuppositions prevalent in a single society. The practice of intercultural discourse, rather, opposes the danger of judging other cultures with the values and standards of one's own culture (ethnocentrism). This danger is especially present in the tendency of devaluing other cultures from the standpoint of European and American cultures in judging them to be deficient, primitive, barbarous, or underdeveloped (Euro or Americocentricism). Basic approaches to an ethics of respect for other cultures can be found in Herder's concept of "a people's genius" ("Genius eines Volkes") and in Hegel's theory of "people's spirits" ("Volksgeister"). Subsequently, through the course of history, different forms of culture are being developed, which are coherent and, thus far, justify themselves.

The intercultural discourse demands that other cultures be taken seriously in their own distinctiveness. This, however, does not imply relativism. There are numerous causes as to why intercultural discourses are needed. To name but a few: the challenge of pluralism, globalization in all its facets, the growing self-consciousness of various cultures, and the worldwide migratory movements. Heuristically speaking, the basic idea of the intercultural discourse can be formulated as follows: Whatever we consider to be fundamentally important, we can expect to discover in different cultures, too. This basic idea works positively – what we consider to be basic justified claims can be found in other cultures, as well – and negatively: Whatever we consider to be blameworthy, other cultures do so, too.

An intercultural discourse of morality, justice, and law takes place on several levels:

1. Within the history of moral and legal thought, the intercultural discourse tries to uncover the origin of morality and law distinctive to different cultures, and the various cross-cultural influences. Thus, the intercultural discourse points to the fact that even the oldest European faculties of law are influenced by Islamic thought. Islamic thought, in turn, is informed by ancient Greek moral, legal, and political thought. And these intercultural influences stretch even further back into history: Remember, for example, that Solon, the great legislator of ancient Greece, traveled through numerous countries; that Plato takes other city-states, for example, Sparta, as a paradigm; and that Aristotle praises Carthage's constitution, which is not Greek.
2. In comparing different legal cultures, the intercultural discourse focuses on the shared moral legacy of mankind such as the Golden Rule, basic procedural principles, universally accepted legal goods, and fundamental principles of justice. Furthermore, it recognizes that in all cultures character-traits such as integrity and righteousness, bravery, and helpfulness are highly esteemed.
3. In the justification of moral and legal principles, the intercultural discourse does not recourse to premises that are particular to a given culture, but instead uses universally valid premises that are founded on human reason and in anthropological elements. The intercultural discourse considers those arguments as valid that cannot be rejected from the viewpoint of any culture. It employs a basic distinction between universality and uniformity. Against communitarian objections, it defends universally valid principles that are justified in a formal way, thus securing openness for the distinctiveness of different cultures. It focuses on normative principles for the regulation within one society as well as on principles that should govern the interaction between societies.
4. Concerning moral and legal practice, the intercultural discourse demands that the principles be realized cautiously: Other cultures should be allowed to go through similar learning processes in due time that Europe and North America have been through with regard to human rights and international law.

An Example: Is There an Intercultural Criminal Law?

The Question

The still ongoing process of globalization reaches far beyond economical and financial issues. In the course of this process, social and cultural borders are frequently transgressed. Therefore, we need to address the question how legal orders should react toward strangers, especially toward those strangers coming from an essentially different legal culture: May legal orders use their strongest means, that is, criminal law? Do they possess a (subjective) right to use their (objective) criminal law cross-culturally? Is there an intercultural right to punish persons?

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63105-7

These questions are all the more urgent since globalization threatens an element that is indispensable for human beings: the attachment to well-defined groups. It is not just the well-being of persons that is thus endangered, but the specific kind of personal identity that relies on such fundamental legal goods as the right to one's own religion, one's own language, and one's own culture. Even where these goods are not threatened directly, an indirect danger lurks: The stranger, coming from another culture, for a long time will feel "culturally torn" between the older and the new (legal) culture. From a criminal-political point of view, this situation is highly relevant: Persons whose identities are aggrieved tend to be more criminally active than those who have a stable identity.

In a much less dramatic way, these questions arise from another perspective as well: Almost all legal cultures acknowledge that legal authority should reside with the state. Furthermore, they also share the recognition of a great amount of types of criminal offenses. And ever since the Code of Hammurabi (*c.*1770 BC) was written down, we have the information that legal goods like life and limb, property, and a good name ("honor") have been legally protected since the ancient times. Now if, first, the same acts are regarded as criminal offenses, and, second, if it can be expected by the stranger to inform himself at least with regard to criminal law about the peculiarities of his host country – especially within such realms in which peculiarities are to be expected – he cannot easily appeal to cultural variations.

Five Types of Answers

Whoever is aware of the course of history knows that the question about international criminal law is in fact not as new as might be suggested by the extent to which many contemporaries praise the connection between globalization and multiculturalism as the newest accomplishment of mankind. On the one hand, one should not forget that the problems in question already occur within a single pluralistic society: The so-called old European states have gone through a process of differentiation starting with the Reformation, affirmed by the Enlightenment, and reinforced by the Industrial Revolution. Within this process of differentiation, the relatively homogenous cultures were dissolved and multidimensional and open societies were created. In these societies exist a still-growing number of different groups with their own interests and convictions, and thus various values, manners, and life-styles. As far as it was necessary, during the course of history criminal law adapted quickly to the new circumstances via a process of liberalization: No matter whether you adhere to a different religion or whether you are an atheist, no matter whether you criticize the society or the government, no matter whether you want a divorce or whether you are homosexual – all such "deviant" behavior is no longer a criminal offense, but instead is considered irrelevant from the perspective of criminal law.

On the other hand, it is necessary to point out that accounts of multiculturalism and even globalization have a long tradition themselves: Think, for example, of what happened in the vicinities of Alexander the Great's campaigns; of the growth of the Roman Empire; of the spread of Christianity and of Islam. Last but not least, we are aware of these kinds of developments during the Age of Discovery, the Age of Colonialism and of Industrialism, and the Age of World Trade. Within all those developments, the same question arises: How to treat a stranger in cases of criminal law? From ancient times onward, certain types of answers to this question have emerged that are worth discussing even today. I will here present five such types, two of which are known to us by common sense and thus without any special legal or philosophical knowledge.

The first type of answer originates from the Old Testament: "You shall treat the stranger who sojourns with you as the native among you" (Moses 3: Leviticus 19:34). According to this principle, the stranger should be treated neither better, nor worse. The continuation of the text states the humanitarian intention of the principle: "you shall love him as yourself." The reason for this kind of equal treatment is connected with the idea of hospitality, which is grounded in one's own experience: "for you were strangers in the land of Egypt." Experience alone, of course, is not yet a justifying reason. Thus, in order to avoid the is-ought fallacy, a further, genuinely moral element is needed. This element, however, can be self-evident in a way that it need not be stated explicitly.

In case of the mentioned verse, the normative element in question can be found in the negative formulation of the Golden Rule: "Do not treat others in ways that you would not like to be treated" (for the Old Testament, cf Tobit 4:15). This rule, a fundamental principle of reciprocity, is universally valid in an empirical as well as in a justificatory way: Empirically, this rule is interculturally valid since it is present in many cultures. It can be found, for example, in Confucius, Thales (as one of the Seven Sages of Greece), in the Indian epic Mahabharata, and in the Old as well as in the New Testament. In a justificatory way, the Golden Rule is interculturally valid since it does not appeal to any presuppositions particular to a single culture. In both regards, therefore, the Golden Rule is part of the moral legacy of mankind.

This kind of humanitarian equal treatment, however, could prove to be inhuman: The stranger, bound to the laws of the host country, loses the right to follow his native custom. Since native custom is quite fundamental to a person's conduct and life-style, it might even result in the stranger rendering himself liable to prosecution. Such cases can vary from minor offenses such as in most countries of Europe, such as fishing without a permit, because such a permit is not required in a person's native country, up to graver offenses – think, for example, of the problem of sexual relationships with minors.

Whoever wants to avoid criminal prosecution needs to adapt to the custom of the host country. This, however, might result in a person estranging herself from her native custom. Here, we thus face a dilemma – the legal dilemma of intercultural encounters: A fundamental principle of justice, equal treatment, might result, if adapted to strangers, in unequal treatment, and thus in injustice: While the natives are allowed to follow their custom, strangers must adapt themselves to new custom, which often is contradictory to their own native custom. This legal dilemma of intercultural encounters thus consists in the conflict of equal treatment and subjection. Mere custom, however, rarely is relevant for criminal law. I leave aside the historical question of whether anybody has ever been legally punished for criminal offenses that were alien to him. The point is that we might find some examples – whether there are many such examples is yet another question.

In the Roman Empire, the dilemma just mentioned is basically known. As a union of states with numerous more or less autonomous states as its members, which were dependent on Rome only in virtue of a law of nations (*ius gentium*), the Roman Empire was in fact a multiethnic state with – in today's terms – a highly complex multicultural society. Within the Roman society, more types of answers to the question of international criminal law emerge, that is, a special law of nations. This Roman *ius gentium* is no law between nations, *a ius inter gentes*, but an international civil law, especially concerning trade. In contrast to Rome's autochthonous civil law, the Quirine *ius civile*, the *ius gentium* is understood to be an intercultural law known to all peoples. Philosophers later grounded this intercultural law on natural reason common to all human beings (*naturalis ratio*). Rome, thus, treated strangers (*hostes*, later: *peregrine*: foreigners) neither according to their own, nor according to Rome's own law. Rather, Rome resorted to a law known to all peoples, so that the dangers did not arise that are present in the passage from the Old Testament – estrangement and subjection.

Roman criminal law, however, is not subsumed under the *ius gentium*. According to Theodor Mommsen, our historical tradition does not link the *ius gentium* to criminal offenses only because the narrow articles of formal civil law are not applied to this realm at all (cf Mommsen, 1899: p. 119). Here, we encounter a third type of answer to our question: The difference between a citizen and a stranger is omitted, since the concept of a crime – public as well as private – does refer to a human being and not to a citizen (cf Mommsen, 1899: p. 118). Thus, we encounter a human right already in Antiquity – albeit in an objective rather than a subjective sense of right: a universal criminal law. In cases of murder, arson, theft, and material damage, it is the person who is morally to blame – regardless if she is a citizen or not.

It should be added that since the offenses just mentioned are counted as universal violations of law, some sort of a criminal law *ius gentium* is in fact drawn upon: An objective condition is employed that is tantamount to the idea of an *ius gentium*: In order to avoid any danger of hegemony, one should employ wide concepts of law, rather than locally or regionally bound concepts. Wherever this condition is not satisfied – think, for example, of the criminal offense of 'heresy' introduced under Christian emperors – there is a case of legal injustice, as well as of political mischief.

Within criminal jurisdiction, a forth type of answer emerges that need not necessarily be understood as an alternative: A further phenomenon is added to the human right of criminal law. This phenomenon reaches into the vicinity of federalism and fulfills the still-existing claim of 'no foreign judges': Especially the Eastern part of the Roman Empire, which was self-confident with regard to its legal culture, retained the right to its own criminal courts for the cities in its jurisdiction.

The civil law-variant of the *ius gentium* is necessary because of a problem that is unknown to modern states that are secular and territorial, but indispensable for a city state in Antiquity: A person became a citizen neither through subordination to a territorial state power, nor through attachment to an ethnic group, but only though active political participation. Since, furthermore, the rights of citizens heavily depended on sacral norms and norms of custom, there was a principle of territoriality rather than a principle of personality: Regardless of her residency, a person lives – a fifth type of answer – according to her law of origin. In this sense, *ius civile* literally means a civil right that is valid for the *cives*, the citizens as members of a *civitas*, a city state, and only for them. It is not equivalent to the meaning of civil law as opposed to public law.

To sum up: The first three types of answers have in common that the stranger cannot simply claim that he comes from a different legal culture. According to the Old Testament-type of law, the stranger is not allowed to appeal to a special status since he is treated equally like any native. According to the Roman civil law, the stranger is not allowed to do so because his criminal offense is to be measured on an intercultural scale. And according to the Roman criminal law the stranger is not allowed to do so since this law applies to humans as humans. While the forth, 'federal' type allows for regional differences, the fifth type – an evaluation according to the laws of a person's home – is important up to this day.

A Case Study

Within the Roman criminal law, the stranger cannot appeal to his being different. But how about issues that are of central importance and, at the same time, quite differently treated within different societies, for example, sexual offenses or matters of honor? As a case study, let me give an example that is close to a real case: A Senegalese man agrees with the parents of a Senegalese girl of minor age to take her with him to Germany and to take care of her within his household. Some time later, he has sexual intercourse with the girl who now is 16 years old.

According to German criminal law, this man is guilty of sexually abusing a child under his charge (cf StGB, § 14). If both the offender and the victim were German natives, the verdict would need to be 'guilty.' However, according to a statement by the girl which the judges considered to be trustworthy, it is part of Senegalese custom that a woman 'offers herself' to the man who takes her into his home as her protector. The girl furthermore stresses that she was not forced to any kind of action; she offered herself to him voluntarily. How is this case to be judged?

Or think of another case: An Arab man marries a second woman while living in Germany. Although this is a custom in his home country, it is forbidden in his host country as 'bigamy' (cf § 171, StGB). To make things more complicated: An Arab man is married in Europe and returns to his native country only for the sake of marrying a second woman, and then comes back to his host country to live there with both wives. Various similar cases can easily be construed. In what follows, however, I will concentrate on the first case study of the Senegal man.

Western or Modern Legal Morality: Human Rights

For a discussion of these cases to be interculturally convincing, the concept of human rights seems to impose itself – provided, however, that this concept is justifiable without recourse to premises particular to a single culture. The various prominent historical conceptions of human beings like the Greek, Jewish, Christian, Islamic, and Buddhist conception of human beings all employ culturally specific elements. Therefore, an argument

that aims at intercultural validity has to depart from these elements in order to function without any ethnocentric elements in both the descriptive and the normative sense. This is not the place to elaborate these two senses in full detail; but let me sketch the direction such an argument might take.

If one turns away from everything that is culturally specific in the descriptive sense for the purpose of an intercultural justification, one is left with the *conditio humana* and its anthropology. Here, one again encounters the danger of claiming universal validity for something that in fact is culturally specific. To overcome this danger, I propose a concept of transcendental interests (cf Höffe, 1999/2007): These are the basic human interests that are the conditions of the possibility of our everyday interests. In this sense, they are logically prior and transcendental. Even if an argumentation along these lines is new, philosophy knows its result since its beginnings. The same holds for other cultures; *de facto* they all recognize these higher-order interests.

Three levels of these culturally indifferent and thus universally valid higher-order interests can be distinguished. To recognize them nothing more than common sense is needed. Whatever interests may prevail in a given society, there exist certain kinds of interests that are the conditions of the possibility of developing interests altogether. These fundamental interests can be described as *conditions of agency*: (1) life and limb, including (material) conditions for living a complete life. Anthropology covers this level by considering the human being as a part of nature, as a living being (*zoon* or *animal*); (2) language and reason: the human being is a *zoon logon echon* viz. an *animal rationale*; and last but not least; (3) a general sociability: the human being is a social animal (*animal sociale*); and (4) a more particular capacity for being political: the human being as the *zoon politikon*, including the concepts of law (*animal juridicum*) and political society (*animal politicum*).

There maybe objections to every level. They can, however, be countered. Against the thesis of a logically prior importance of life and limb one might, for example, refer to martyrs and suicides, which in different ways relativize the interest for life. But since even these people want to decide what for, when, and how they are going to die – otherwise they would simply be killed – life and limb remain conditions for agency even for them. For securing all of these three descriptive levels, law and a morality of law are needed.

If one abstracts from everything that is culturally relative on the other, normative, side as well, one is left with the following: an idea of a regulation that rejects any despotism and violence that realizes this objection universally and thereby allows for an advantage for everyone concerned. Such a regulation can be labeled 'legal-practical reason.' It can be realized in three steps of increasing recognition. On the first step – realizing law – certain rules are introduced. On the second step – legitimizing law – these rules are strictly enforced. And on the third step – standardizing law – the rules are subjected to second-order moral rights.

So what about justification? With regard to the first level of transcendental interests – the human being as a living being with a body – the argument goes like this: Concerning his biological make-up, the human being is a possible victim as well as a possible perpetrator of violence. This fundamental situation of human beings, their *conditio humana socialis*, is answered by the morality of law: According to the concept of human rights, every human being has a subjective claim not to fall victim to violence. To put this positively: a right to life and limb.

In order to justify this right, the presupposition of an interest, even a fundamental transcendental interest, is not sufficient: From an existing interest, we can infer no claim to recognition of this interest by others. A further argument is needed. This argument starts from the observation that subjective rights are only present where others take on certain obligations; human rights are bound to respective human obligations. For example, the right to life and limb of a given person is tantamount to the obligation of all other persons to refrain from violence.

Now generally a claim to a benefit exists where this benefit is offered under the caveat that a certain similar benefit is to be expected. Claims justify themselves from reciprocity, pars pro toto: from an exchange. In order to justify human rights as subjective rights, one needs to establish a reciprocity that is part of the human being simply in virtue of being human. This situation exists when an interest of reciprocity, which is indispensable to human beings, can only be realized in reciprocity and through reciprocity. At this point, where interests are indispensable and, at the same time, bound to reciprocity, indispensability is transferred to reciprocity; the respective exchange is itself indispensable. It is not that a human right exists simply because every human being has a (higher-order) interest in life and limb. It exists, first, because this interest can only be realized in reciprocity and, second, because in the resulting 'system of reciprocity,' everybody already makes use of the benefit – the renunciation of violence by others – which is provided only by the reciprocal benefit: one's own renunciation of violence.

Principles Open for Different Cultures

The result so far is thus: Human rights are not specifically a European or Western phenomena; rather, in human rights, something common to all human beings breaks through. "Breaking through" entails a moment of evolution. Insofar as the West breaks through and other cultures do not, these other cultures might appear to be less developed. The notion of being less developed has an arrogant ring to it, which is why these cultures in question try to defend themselves against such a description. An intercultural legal discourse at this point demands caution.

In the encounter of the West with other cultures, nowadays for instance with China and Japan, a reciprocal give and take occurs. And insofar as the West demands respect for human rights from these other cultures, the West has to ask itself from time to time, if it is really human rights, simple and pure, that are being demanded, or if it is not rather a mixture, a combination of human rights with Western peculiarities, special experiences, maybe even special interests. For this reason, a *hypothesis of contamination* is sensible: The hypothesis that the alleged human rights are in fact contaminated with other elements.

Although every alleged human right can be infused with nonuniversal elements, this hypothesis does not demand total restraint. It simply demands the effort of a careful

interpretation. From the hypothesis of conglomeration, the duty of decontaminating arises. Just like the analytical chemist one needs to separate the particular from the universal in order to arrive at the real and original human rights. It is only with this effort that philosophers fulfill their duty that they have adopted ever since Greek Antiquity – the duty to ground everything solely on human reason common to all. If the philosophers fulfill this duty, they may call themselves, even if they are only modern European, with Husserl (1936/1970, 17) 'functionaries' or, even better, 'lawyers' of humanity.

It is not only the lawmaker who needs to remain culturally open when formulating human rights and criminal law in accordance with them. The judge, too, needs to remain culturally open in order to be able to judge reasonable in cases of criminal law that arise from typical conflicts of cultures.

Case Study: Further Inquiries

Let us return from these fundamental considerations back to our first case study – the Senegalese man – and let us try to weight the interests of the perpetrator and the victim. The considerations concerning their respective interests lead to certain further inquiries.

1. Perhaps the victim and the perpetrator have been in Germany only for a short time? In this case, one could consider them as people who are passing through, and whose business we need not meddle with. But let us suppose that in our case they have been in Germany for quite some time. Since this time could extend into years and maybe even decades, a political question arises: What is better for a young Senegalese woman, the – allegedly – sparse security of law in her home country, or the more encompassing security of law of the society in which she has been living for quite some time, and will possibly continue to live in even throughout her whole life?
2. Did the girl perhaps count as an adult at the age of 16 in her home country? If so, the judges might consider a mistake as to the wrongful nature of the act: One could assume that the offender had sexual intercourse with the girl only because he believed her to be an adult. Since questions about the legal age, however, are fundamental, one cannot simply assume such a mistake on behalf of the perpetrator, since he has been living in his host country for quite some time – one would at least need to consider this mistake as having been avoidable.
3. What exactly was the nature of the agreement between the man and the parents? Maybe there has already been some sort of compensation given to the perpetrator by the parents, in which case the perpetrator could be considered as having taking a double compensation.
4. If one remains at the level of custom, one could ask how voluntarily the girl's agreement in fact was. Furthermore, mere voluntariness does not in itself lead to the action not being a criminal offense. In effect, one could argue that it is hard to assume voluntariness, since the action took place within a framework of possibilities sketched by custom, which leaves but a few ways to act: Every custom and tradition limits possible ways of action. In the case at hand, this limitation goes further than in other cases: The sexual devotion in this case is a donation in return for another donation. Here, the principle of justice of exchange is called for. Taken care of by the perpetrator, the victim assumes that she owes sexual obedience to him in return. The alleged voluntariness, thus, occurs within a frame of nonvoluntariness, the feeling of owing something to somebody.
5. What is the status of the custom in question from the perspective of criminal law? The Senegalese system of law is oriented along the French system of law, and thus along a European legal tradition, so that an argument for a right for exemption from German criminal law cannot be made on solid ground – as it would be the case with a totally different legal culture. Within the new French *Code pénal*, the article in question would be 227, 27: Abuse of a person under one's charge in a case where the victim is older that 15 years, but still a minor and not emancipated though marriage. Within the old *Code pénal*, which is valid for Senegal, it is article 331-1, which in turn essentially corresponds to §174, of the German criminal law (StGB).
6. The content of the custom in question is related to a traditional, and also patriarchal society, in which – in the case of Senegal – 90% are Muslims. Since it is known what kind of value virginity has in these societies, the custom in question results in a very grave disadvantage for the girl. This situation should make us hesitant when facing the claim that this is simply a foreign custom that needs to be respected. If we take the high value into account that virginity has in these societies, an argument for a right for exemption in this case, again, can hardly be made on solid ground.
7. One could ask if the essential part of the offense – 'abuse' – could be dropped, because the relationship in question in fact is a relationship of mutual love. Against that, however, stands the recourse to a Senegalean custom, and thus to a tradition rather than to love.

An Intercultural Criminal Law

These kinds of further inquiries show that encounters between different cultures do not simply lead to situations in which the strangers can refer to their different legal cultures. Furthermore, there are experts who argue that some Africans, Indians, and others tend to use double standards with regard to Western cultures: In cases where the host culture demands a higher form of responsibility, they refer to their native legal culture, and when in their home country, they refer to Western legal culture.

In order to keep things simple, let us assume that all further inquires that we mentioned are being answered in favor of the perpetrator. Then, the key question remains: How far does the coercive power of a legal culture extend in such cases?

If we would simply need to overcome the allegedly culturally biased sentences, the case would be easy: Modern societies overcome fragments of cultural narrow-mindedness. In this sense, they have cleared their criminal law of such traces over time and deleted offenses like suicide, homosexuality, heresy, lese majesty, and criticism of the state. For the most part with criminal offenses, especially

with regard to those concerned with life and limb, there exists no cultural bias. On the contrary, these are the laws that preserve human dignity as spelled out through the human rights that are known almost universally; there is no known legal culture that does not include killing as a criminal offense.

Now, to return to our case, does sexual protection of minors fall under the charge of being culturally biased? At least prima facie, the contrary is true: Through this kind of protection, cultural biases are overcome, for example, the tendency of patriarchal societies not to consider women and children as equal.

Therefore, the question can now be answered: Insofar as the criminal offenses can be justified with reference to arguments of human rights – and this is the case with the majority of offenses, especially in liberal legal cultures – an intercultural criminal law in fact exists. The reason for it does not consist in the arrogance of a legal culture to subject strangers. It does also not consist in the idea of refusing the stranger a right of being exempted from criminal law. Finally, it does not consist in the (certainly justified) interest of a society for its own protection. The reason in question, rather, is the fact that from the perspective of criminal law the remote stranger does not exist: It is hard to come up with legal cultures that are so radically different that they do not know offenses the prohibition of which is justifiable through human rights. And even if such cultures would exist, one could demand that its members recognize human rights as soon as they travel to another society, since the offenses are justifiable through human rights. Intercultural law, therefore, reinforces the basic idea of intercultural legal discourses: Whatever we consider to be fundamentally important, we discover in different cultures, as well. And whatever we regard as fundamentally blameworthy, others do so, too.

See also: Conventions and Norms: Philosophical Aspects; Equality: Philosophical Aspects; Kantian Ethics and Politics; Law and Morality: A Continental European Perspective; Law and Morality: An Analytical Perspective; Multiculturalism; Philosophy of Law; Pluralism and Tolerance; World Republic.

Bibliography

Beccaria, C., 1764. On Crimes and Punishment (Dei delitti e delle penne) (D. Young, Trans.). Indianapolis, IN 1986.
Berner, A.F., 1876. Lehrbuch des deutschen Strafrechts (Leipzig).
Bultmann, C., 1992. Der Fremde im antiken Juda (Göttingen).
Ferri, E., 1881. I nuovi orizzonti des diritto e della procedura penale (Bologna).
Höffe, O., 1996. Vernunft und Recht. Bausteine zu einem interkulturellen Rechtsdiskurs (Frankfurt, M).
Höffe, O., 1999. Gibt es ein interkulturelles Strafrecht. Ein philosophischer Versuch (Frankfurt, M).
Höffe, O., 1999/2007. Democracy in an Age of Globalisation (Demokratie im Zeitalter der Globalisierung) (D. Haubrich, M. Ludwig, Trans.), second ed. New York, 2002.
van Houten, C., 1991. The Alien in Israelite Law. JSOT Press, Sheffield.
Husserl, E., 1936/1970. The Crisis of European Sciences and Transcendental Philosophy (Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie. Eine Einleitung in die phänomenologische Philosophie) (D. Carr, Trans.). Evanston, Den Haag, 1936/54.
Mommsen, T., 1899. Römisches Strafrecht. Leipzig.
Taylor, C., 1994. In: Gutman, A. (Ed.), Multiculturalism, Examining the Politics of Recognition. Princeton, NJ.

Interest Groups

Kay Lehman Schlozman, Boston College, Chestnut Hill, MA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by H.J. Puhle, volume 11, pp. 7703–7708, © 2001, Elsevier Ltd.

Abstract

Interest groups – for example, business and professional associations, labor unions, and environmental and consumer groups – are a feature of every functioning democracy. All interest groups perform certain functions: providing information to public officials; seeking to persuade policy makers to pursue a preferred course of action, an activity that may be difficult to distinguish in practice from the attempt to inform; and communicating with and educating members. Within the context of these widely shared functions, organized interests vary substantially both within and across democracies with respect to the difficulty they face in gaining access, the techniques they use to influence policy, and the targets of their activities. In addition, political systems differ in terms of the structure of relations between interest groups and government: some political scientists have suggested that there are two basic patterns of organized interest interaction with the state, neocorporatist and pluralist. In terms of the kind of citizen interests represented by groups, although there are both substantial differences across democracies, the interests of producers are inevitably well represented. Two recent developments affect interest-group politics on a cross-national basis: the emergence of interest representation at the European Union and the emergence of greater pluralism in democracies everywhere.

Among the requisites for democracy on a sustained basis are not only free and fair elections but also the rule of law, respect for dissent and freedom of expression, and the legitimation of the organized representation of citizen interests – often in opposition to those currently in power – through parties and interest groups. Interest groups, those myriad organizations – ranging from business and professional associations to labor unions to environmental and consumer groups to groups concerned about issues as diverse as abortion and nuclear power – that link the individual to public institutions and seek to influence public policy, are a feature of every functioning democracy. Although they may be present in nondemocratic systems as well, it is not an exaggeration to claim that democracy on a national scale is impossible without interest groups.

Problems of Definition

Differences of opinion arise in any discussion of how to categorize the various collectivities that mediate between the individual and the state in a democracy. Many of the institutions often described as interest groups – for example, corporations, foundations, or public bodies such as regional governments – are not groups at all, in that they do not have members in the ordinary sense. For this reason, some analysts prefer the term 'organized interests' to 'interest groups.' Interest groups have much in common with other social aggregates that represent citizen preferences to policy makers in a democracy. In fact, organized interests often have connections, formal or informal, with political parties or social movements. However, unlike political parties, they do not nominate candidates to campaign under their name for the purpose of running the government. They contrast with social movements in their greater degree of formal organization.

Interest Groups and the Policy Process

Wherever they are, interest groups serve as intermediaries in a two-way process of communication linking a citizen constituency with policy makers. All interest groups perform certain functions: they provide information to public officials to assist in designing policy; they seek to persuade policy makers to pursue a preferred course of action, an activity that may be difficult to distinguish in practice from the attempt to inform; and they communicate with members – keeping them apprised as to what the government is doing, educating them about the political process, and cultivating support.

Within the context of these widely shared functions, organized interests vary substantially – in ways that reflect the degree of permeability of the political system to new interests, and the nature and relative importance of formal political institutions – with respect to the difficulty they face in gaining access, the techniques they use to influence policy, and the targets of their activities. Newly emergent groups have much more difficulty penetrating interest group politics in Israel, which has a strong, centralized state, than in the USA, with its tradition of voluntarism and a public domain that, because it is divided by federalism and separation of powers, provides many points of access. Within and across polities, there is tremendous variety in the kinds of techniques groups use to realize their objectives. France, for example, has a tradition of direct action by groups not present in, say, the UK. Within any single country, organized interests differ in the extent to which they function solely as insiders utilizing traditional, low-profile means of achieving political influence or operate as outsiders, adopting more public tactics of mobilization and protest – with unions and environmental groups more likely to engage in protest activities than, say, business associations.

In choosing where to direct their energies, organized interests inevitably gravitate to power. Hence, the primary focus of interest group activity in Japan is the bureaucracy, while in Italy it is parliament. In a system as fragmented as the US political system, interest groups often have choices, and may seek to locate a political controversy in the institutional arena in which they would have the greatest probability of success.

Political systems differ not only in terms of what interest groups do and where they channel their activities but also in terms of the structure of relations between interest groups and government. Some political scientists have suggested that there are two basic patterns of organized interest interaction with the state: neocorporatist and pluralist. According to the neocorporatist model, approximated most closely in Sweden and Austria, mechanisms providing for equal representation of individuals on a geographical basis are supplemented by mechanisms providing for functional representation of organized interests – ordinarily by a relatively limited number of peak associations. Such organizations are recognized and licensed by the state and attain a regularized role in policy making through delegations of administrative power or participation on public councils or committees. Important government decisions are made only after consultation with major economic interests, most notably workers, employers, and farmers.

In contrast, according to the pluralist model, interest groups are much more numerous – less highly aggregated and organized around multiple bases of political conflict – and much more autonomous. Pluralist interest-group politics, approximated most closely by the USA, is relatively permeable to the entry of new groups, and groups arise and get involved in politics at their own initiative, rather than at the behest of the state. In a politics of pluralism, interest groups have greater freedom in choosing which political battles to fight in which political arenas. Typically, those battles involve shifting, issue-specific coalitions taking sides on issues of narrower, more limited import.

It is essential to recognize that these alternative models, which have generated considerable scholarly discussion and some controversy, are ideal types. No polity actually conforms to either model, and many countries – for example, France where a strong, centralized state has meant comparative weakness for interest groups, or Japan where regular consultation between business and government more or less excludes labor – evidence a pattern that is neither pluralist nor corporatist. Furthermore, in most countries no single pattern obtains for the making of policy in all issue areas: the formulation of foreign policy, for example, is less likely to entail regularized bargaining with interest groups than the making of agricultural policy.

Interest Group Systems: Who is Represented?

Reflecting national differences in their political traditions and political culture, in the number and depth of politically relevant social cleavages, and in the number, strength, and ideological distinctiveness of the parties, there are also, across democracies, enormous differences in the nature of interest systems, the collective set of organized interests active in politics. Polities vary widely in both the number of groups and the proportion of citizens who are members of any group at all; in the Scandinavian countries, where most employees belong to some kind of union or professional association, membership is much higher than in, for example, Italy or Greece. There are also differences in the nature and number of the axes of political cleavage encompassed in organized interest politics; the multiplicity of kinds of interests represented in the interest groups of a large and diverse country such as the USA contrasts with that in a relatively homogeneous country such as Norway.

In addition, interest group systems vary substantially in the degree to which interest representation is highly aggregated or fragmented functionally or geographically and in the extent to which interest groups have a monopoly on the organization of a particular constituency. For example, in contrast to the centralized interest group systems of Austria and Sweden are those of the Netherlands, where business associations are organized according to religious differences, or Belgium, where business associations are divided along regional lines but labor is pillarized into socialist, liberal, and Catholic national federations. Furthermore, there are marked differences in terms of the relative strength of organizations representing different kinds of interests – in particular, in the vigor of the opposition provided to traditionally powerful business organizations by organized labor and citizens' groups concerned about consumer and environmental issues.

There are, however, certain regularities in the kinds of interests represented by interest groups. It is difficult to think of a democracy with active interest group politics where producer interests are not well represented – whether by peak employer groups, sectorally based trade associations, or corporations acting individually. In contrast, with the exception of pensioners, the interests of those who benefit directly from welfare-state policies – for example, students or recipients of public assistance – rarely receive vigorous advocacy by interest groups although their causes may be championed by political parties.

Interest Groups and Parties

That the political parties in a democracy might represent interests not being championed by organized interests – and vice versa – makes clear that an overall assessment of the way in which citizen interests are represented in any particular democracy requires that its interest groups be considered in the context of its parties. Common wisdom holds that strong parties go hand in hand with weak organized interests and, conversely, weak parties with strong organized interests. In the USA, a plurality of groups broadens substantially the range of interests represented in politics by two relatively weak and ideologically moderate parties. However, strong parties do not necessarily preclude strong interest groups. Sweden has strong political parties, and interest groups are guaranteed a consultative role in corporatist policy making in the system of state commissions. And France has traditionally been characterized as having both weak parties and weak interest groups in the face of a strong state.

Democracies also differ in the extent to which interest groups have strong links to like-minded political parties. For example, at least until recently, UK trade unions have had a much closer relationship to the Labour Party than their US counterparts do to the Democrats.

The Evolution of Interest Group Politics

In spite of the substantial differences across democracies, there have been, over the past few decades, two related developments

affecting interest group politics on a cross-national basis that suggest the possibility of convergence among democracies. First is the emergence of interest representation at the European Union (EU). Although group activity in Brussels often arises within individual EU nations, many of the groups involved are either individual corporations or European-level, rather than national, organizations.

The second important trend is an advocacy explosion in democracies everywhere with – in a fragmented, pluralist fashion – more concerns being represented by more groups, and a greater reliance on professional lobbyists. This fluid style of interest advocacy is especially characteristic of the EU but has appeared in the national capitals as well. With respect to the kinds of interests represented, a growing number of citizens' groups concerned about an array of issues including nuclear proliferation, the environment, immigration, and the rights of gays, women, and ethnic minorities supplement traditional economic groups representing business, farmers, and workers. How these developments will alter interest group politics is not yet clear. Nonetheless, private associational life will surely remain essential to the functioning of democracies around the world.

See also: Issue Networks: Iron Triangles, Subgovernments, Policy Communities, Policy Networks; Pluralism; Political Representation.

Bibliography

Baumgartner, F.R., Leech, B.L., 1998. Basic Interests: The Importance of Groups in Politics and Political Science. Princeton University Press, Princeton, NJ.

Bentley, A.F., 1908. The Process of Government. University of Chicago Press, Chicago.

Crouch, C., 1993. Industrial Relations and European State Traditions. Clarendon Press, Oxford, UK.

Dalton, R.J., 1994. The Green Rainbow: Environmental Groups in Western Europe. Yale University Press, New Haven, CT.

Heinz, J.P., Laumann, E.O., Nelson, R.L., Salisbury, R.H., 1993. The Hollow Core: Private Interests in National Policy Making. Harvard University Press, Cambridge, MA.

Mazey, S., Richardson, J., 1993. Lobbying in the European Community. Oxford University Press, Oxford, UK.

Olson Jr., M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.

Schlozman, K.L., Tierney, J.T., 1986. Organized Interests and American Democracy. Harper & Row, New York.

Schmitter, P.C., Lehmbruch, G. (Eds.), 1979. Trends toward Corporatist Intermediation. Sage, Beverly Hills, CA.

Thomas, C.C. (Ed.), 1993. First World Interest Groups: A Comparative Perspective. Greenwood Press, Westport, CT.

Truman, D.B., 1951. The Governmental Process. A Knopf, New York.

Walker, J.L., 1991. Interest Groups in America. University of Michigan Press, Ann Arbor, MI.

Interest Groups, History of

Hans-Jürgen Puhle, Institut für Politikwissenschaft, Goethe-Universität Frankfurt am Main, Frankfurt am Main, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Modern interest groups can be defined as voluntary associations of individuals, firms, or smaller groups uniting in order to defend or fight for a common interest, with the intention to influence and intervene in the political process, but without ambition to form a political party. Basically, eight types can be distinguished. Their shape, functions, patterns of organization, and modes of interaction depend on the specific constellations and trajectories of a given polity, and so does their democratic potential. The systems of interest intermediation (between pluralism and corporatism) show different degrees of institutionalization. Major transformations at the end of the twentieth century have given rise to tendencies toward more fluidity and transnationalization.

Organized interests and groups, their factions or parties fighting for influence and power in a given polity or beyond, go back a long way in history. In modern political theory, the notion of 'interests' was first narrowed down by enlightened philosophers from Machiavelli to Montesquieu to material and rational interests, as opposed to and counteracting the 'passions' (Hirschman, 1977). James Madison and the Federalists considered both the passions and the different material interests, particularly of 'those who hold, and those who are without property,' to be the principal sources of factions in a society that could only be contained and controlled by building adequate institutions with a sufficient number of checks and balances (Hamilton et al., 1961; 'Federalist 10'). The concept of interest groups, although under different labels, and of interest politics in a more specific sense, began to proliferate in the Western world from the beginning of the nineteenth century when, after the American and French revolutions, the world of political parties and factions started to dissociate from the universe of interests at large and when the unequal progress of industrialization in different parts of Europe induced individuals, firms, and groups to articulate their interests in determined policies in order to improve the economic or institutional prerequisites, or cope with the social consequences of industrialization.

Therefore, the modern interest group might be defined as a voluntary association of individuals, firms, or smaller groups uniting in order to defend or fight for a common interest, with the intention of influencing and intervening in the political process, but without ambition to form a political party (exceptions notwithstanding). A number of similar terms correspond to the same definition: pressure groups, organized interests, lobbies, power groups, advocacy groups, the German 'Verbände,' the French 'syndicats,' American 'chambers,' and many 'institutes' or 'associations.' (Functional equivalents in the non-Western world are excluded here.) The more general term 'associations' also indicates proximity to and overlaps with the broader spectrum of civil society organizations (CSOs), which may also be labeled as nongovernmental organizations (NGO), voluntary-sector or nonprofit organizations (NPO), social movements, or otherwise (Della Porta et al., 1999; Maloney and Roßteutscher, 2007). All these groups usually also display functions and aspects of interest groups, so that the issue of commensurability of 'interest group aliases' has recently been raised in the literature (Halpin, 2010). As the interest groups' activities are usually directed toward the political arena asking for specific legislation, allocation of benefits and state intervention promoting their interests, the rise of state interventionism and of the modern social and welfare state since the last quarter of the nineteenth century has been closely intertwined with a secular proliferation of interest groups of different types. On the one hand, these usually reflected the characteristics of the (more or less democratic) political systems in which they originated. On the other hand, the groups' activities have often significantly contributed to modifying and modernizing the political systems and to establishing new rules and patterns of interaction. Their recognition and legitimation not only have been products of their performance but also have been dependent on the general theoretical and ideological orientations of the political consensus within a given polity, which could vary from authoritarian or Rousseauist views of single or 'particular' interests as being categorically detrimental to the common good to more pluralistic Anglo-Saxon approaches of accepting different interest articulations as expressions of a vital civil society (Tocqueville) and of taking factions for granted as vehicles of political agency (Fraenkel, 1964). In order to overcome some of the implied shortcomings of legitimation, interest groups have usually tended to develop ideological mechanisms of their own to demonstrate that their particular interests serve the common good or coincide with the interests of the community. For the great majority of interest groups, through the nineteenth and twentieth centuries, the common and still most important frame of reference has been the traditional nation state, whereas it appears that, since the 1990s, a significant process of transnationalization and 'Entgrenzung' of economic, societal, and political constellations has gained momentum, which may constitute a secular threshold in the development of the structures of politics and of the public sphere, and affect and (partly) reshape not only firms, corporations, the media, social movements, and political organizations but also interest groups. The most relevant forces behind this process are the breakthrough of the 'network society,' with all its technological and economic consequences, advancing globalization (and fragmentation) and the repercussions of regional integration (such as the European Union) and the great economic crisis.

Types

Interest groups have developed into different types. They could organize individuals, firms, associations, or public corporations (cities, states, provinces, and legislatures) at a local, translocal, national, or transnational level and develop different characteristics. With regard to the contents and quality of the interests involved it could be said that, with some simplification and not accounting for the various mixes and marginal or exceptional cases, most of the groups of the 'classical' spectrum of the nineteenth and twentieth centuries may belong to one of the following eight types: (1) professional associations; (2) business, commercial, or industrial groups; (3) labor unions; (4) agricultural organizations; (5) single (nonprofessional, nonproducer, nonpublic) interest groups; (6) ideological interest groups; (7) public interest groups; and (8) welfare associations. With regard to the more recent trends of restructuring and re-forming the public sphere, it will also make a difference (1) whether the groups are national (regional) or transnational interest groups, (2) whether they move within the consensus of what might be called an (informal) 'capitalist international' or have become 'antiglobalist,' and (3) whether they are loosely coupled ad hoc organizations (and movements) or develop structures.

(1) The most 'classical' case seems to be that of the professional associations that have taken over many of the functions of the premodern guilds, estates, brotherhoods, or clubs, particularly in regulating the requirements and mechanisms of co-optation and access to a number of key professions, be it as a privilege of professional self-government as in the United States (American Bar Association and American Medical Association) or as a task delegated by the state as in most of the German 'Kammern' (of lawyers, doctors, and artisans). The various medical associations have become particularly differentiated with the rise of compulsory health insurance or comprehensive national health services (cf. the German Kassenärztliche Vereinigungen). The organization of less professionalized groups such as teachers or employees in trade and commerce has often followed the lines of professional associations, if they did not unionize in the labor ranks as in most Latin countries.

(2), (3) The two predominant camps of interest groups of the industrial era, the associations of industry and the labor unions, have typologically developed out of earlier commercial guilds or professional associations of craftspeople and workers. The labor unions in industry have usually been able to acquire more strength and influence since they began to overcome their traditional fragmentation along the lines of professions and crafts, and started to form broader industrial unions – in Germany since the 1890s, in the United Kingdom after World War II, and in most other industrial countries during the 1920s and 1930s. In industry and labor, it also made a great difference whether there evolved a coordinating peak association or confederation, or whether there were two, as in German industry before 1919 or in Spanish labor before 1939 and after 1975, or a more fragmented variety, as in the United States. In case labor relations became institutionalized, the industrial sector and its associations occasionally created separate employers' associations with the specific objective of coordinating negotiations with the unions. Banking, insurance, and small business interests in most countries have also organized separately.

(4) Most agricultural interest groups, which came into existence from the last third of the nineteenth century, besides voicing the interests of 'agriculture' against industry and the industrial world, have reflected the structures and divisions of the various crop sectors or regions in their nation's agricultural economies, with the interests of larger producers usually dominating those of the smaller ones, a high influence of specialized production interests notwithstanding. In less developed countries or regions with a greater number of dependent small peasants and landless laborers, agricultural interests have also been organized into labor unions and social movements. In some countries (such as France) and in determined historical periods and constellations in others (such as Italy, Spain, and the United States during the Populist Era and Germany and the United States during the Great Depression) the mode of agricultural interest articulation could oscillate between interest group politics and social protest.

Two other 'classical' types are constituted by groups dedicated to voicing (5) a single nonprofessional, nonproducer, and nonpublic interest – for example, the National Rifle Association in the United States, automobile associations, groups advocating or fighting particular forms of exclusive or commercialized sports or entertainment, or the various (often 'ethnic') associations of migrants, refugees, or victims of wars and (6) by more ideologically inspired groups like those advocating imperialism, temperance, or prohibition (Anti-Saloon League), or fighting abortion or the death penalty. A number of single interest groups claim to be contributing to the public good by improving people's health – for example, most large sports associations – or by promoting human progress and empowerment of the underprivileged – for example, many organizations of minorities and of women in the United States (such as the National Association for the Advancement of Colored People, the Leadership Conference on Civil Rights, and the National Organization for Women) and in other countries. Here we also have to count a more recent type of mobilized group that has been triggered by the advancement of globalization and economic crisis since the last decade of the twentieth century and often overlaps with (increasingly transnational) social movements: the broad variety of 'antiglobalist,' and in a new way antiimperialist (and mostly anticapitalist) organizations, such as 'attac' or 'occupy.'

(7) Some of these groups might also be considered to belong, at least partly, to the type of public interest groups, the most common issues of which in the United States – following more than a century of honorable precursors such as the Anti-Slavery Society (1833) and the American Civil Liberties Union (1920) – since the 1970s have been environmental and consumer interests (Common Cause, Public Citizen, Consumer Federation of America; McFarland, 1984). In Germany and some other Central and West European countries, antinuclear and pacifist issues would have to be added. At the transnational level, groups like Greenpeace, Amnesty International, or Transparency International would belong to this category, not to forget a number of 'classical' precursors, such as the International Committee of the Red Cross (Keck and Sikking, 1998).

(8) The heterogeneous groups that might be labeled as welfare associations often combine professional interests, particularly of social workers, with those of public entities, mostly cities and counties, and the single and ideological interests of churches, religious communities, and other private agents in the welfare arena, which in parts might also be considered as serving the public interest. Many organized themselves early on into coordinating central or national associations or corporations, such as the National Conference of Charities and Correction (1874/84) in the United States (professional interests prevailing) or the Deutsche Verein für Armenpflege und Wohltätigkeit (1881) in Germany (corporate interests prevailing), long before sectoral peak associations were formed.

The various forms of intergovernmental lobbying will not be considered here because they are about influence and allocation of benefits within the public sphere only. Hence, neither the National Governors' Association and the US Conference of Mayors nor the German Kultusministerkonferenz and the Deutsche Städtetag will be considered as interest groups. The same applies to the often well-organized lobbies of foreign governments in Washington, DC, or to the representation of state or regional interests vis-à-vis the European Commission in Brussels even if they seek to increase their influence by behaving like private interests. Likewise, the numerous Political Action Committees of American interest groups, which have basically been founded for legal reasons with the specific purpose of influencing legislation, might not be considered as being a separate type of interest group, because the interests voiced by them are the same as those of their creators (Cigler and Loomis, 2012).

Alternative typologies might differentiate between status quo defenders and agents of change, or according to the constellations of the groups' organizational density and power: whether they rely more on money or on numbers, on machines and bureaucrats, or on mobilization and agitation; whether they wish to influence legislators en petit comité or think they need a broad public forum (which also depends on the political system); or whether or not they need a high level of identification or have to provide a variety of services in order to secure their members' loyalty (such as automobile associations). Among the most basic and influential differences are those between big and small interest groups and between big and small constituencies. Smaller constituencies (such as steel or farming interests) are usually much better organized and represented than broader constituencies (such as consumers), which tend to fall victim to the mechanisms of the 'free rider' effect (Downs, 1957; Olson, 1965).

Constellations and Trajectories

The shapes of interest groups and of the evolving national systems of interest intermediation have been significantly influenced by the particular economic, social, and political constellations; institutional sets; and the distinct paths and trajectories of modernization of the different countries and societies. Of particular importance here are the peculiar relations and mixes of the developments in the three most important dimensions of the modernization processes, which, simplified, might be labeled as bureaucratization, industrialization, and democratization. In the United Kingdom, industrialization has triggered democratization, and bureaucratization came late, basically in order to cope with the social consequences of industrialization. In the United States, which followed a similar pattern, the elements of democratization were stronger from the beginning, although moderated by federalism, and more checks and balances. On the European continent, bureaucratization has been the legacy of early state building under absolutism, but here the French revolution has made the great difference. In France, the bureaucratic factors combined with the different lines of democratization (representation by notables or Bonapartism, with its peculiar 'deliberative' mechanisms), whereas in most of the German states, democratization was postponed, and the characteristic patterns of modernization were dominated, down to the mid-twentieth century, by factors combined from the sets of bureaucratization and industrialization.

These different constellations have influenced the emergence and performance of interest groups with regard to their social bases and opportunity structures; their strategies and addressees, organization, modes of operation, and achievements; and their relationship with governments, parliaments, and political parties. It is important to note that of the United States and Germany – two countries that have shaped the Gestalt of modern interest groups for the twentieth century more than any others since the breakthrough of organized mass politics between 1890 and 1910 – one was not democratic. In Germany, the government was not accountable to parliament, and the decades of Bismarckian authoritarianism had weakened the political parties and given the (supposedly less political) interest groups better conditions to flourish than anywhere else.

Different Patterns of Organization

In Germany, as in many other European countries, the interest groups developed typologically in three waves:

1. Most of the associations of the first wave, some of which dated back to the first decades of the nineteenth century, were by 1900 closely cooperating with the state bureaucracy, and had become part of privileged semiautonomous corporations or 'chambers,' of agriculture, commerce, and the trades, which could no longer oppose government policies.
2. The second wave was initiated by the protectionist groups of heavy industry (Centralverband Deutscher Industrieller, 1876) and, although to a lesser degree, by heavy agriculture, in response to the great deflation of the 1870s whose principal addressees were bureaucrats and legislators. The same characteristics of interest representation vis-à-vis parliament and the executive, but not much mobilization, can be found in most groups of business and industry that were emerging during the same period in the industrial countries – usually divided into protectionists and free traders. In the United States from the 1880s, a two-tiered system was emerging with national associations like the National Association of Manufacturers (1885) and the US

Chamber of Commerce, and thousands of influential trade associations representing the various branches of production. In US agriculture, however, populist protest preceded organized lobbying (see below).

3. The groups of the third wave in Germany went beyond the previously established patterns of interest representation in that they started massive and permanent propaganda campaigns and began to influence elections, political parties, and individual politicians. They entered the stage at a time, from the 1890s onward, when the progress of industrialization induced by new technologies combined with the extension of the public sphere and the political mass market, modernization of party organization and the proliferation of social Darwinism, and imperialistic rhetoric to create an ambiente favoring a new type of vigorous, aggressive, and well-organized interest group working the public. Among these groups were the protectionist and interventionist Farmers' League (Bund der Landwirte, 1893), various petty bourgeois groups of the 'old Mittelstand' (artisans and shopkeepers), and a number of imperialistic propaganda associations like the Pan German and the Navy League. These groups were at the core of the politics of the New Right against liberalism and social democracy. They contributed much to the modernization and radicalization of conservative groups that now turned ultranationalistic and racist, and to the irreconcilable right/left polarization of the social and political forces active at the end of the monarchy that was to continue during the Republic of Weimar, which it helped to undermine (Puhle, 1975; Ullmann, 1988).

One of the models the German interest groups of the third wave decided to follow was that of 'social democracy,' meaning the party and the labor unions that soon after their emergence from clandestinity under Bismarck's Anti-Socialist Law in 1890 created efficient bureaucratic machines and initiated wide-ranging activities in expanding and consolidating the movement – which went far beyond the scope of what the earlier traditional crafts unions and associations of mutual assistance had achieved since the 1850s. The 'General Commission' of the social democratic unions, from the late 1890s, came close to what might be called a peak association of labor, despite the existence of some smaller catholic and liberal unions. The powerful industrial unions, particularly those of the metal workers and miners, could claim to be the most modern in the world. Their position was further enhanced by the fact that industry and the new employers' associations, from around 1900, began to negotiate with them on a regular basis, usually at the regional level. Besides bureaucratic traditions, the modern unions contributed to an early emergence of institutionalized labor relations in Germany, and vice versa.

In the Unites States, both the organization of industrial unions, most visible in the Congress of Industrial Organization (CIO), and institutionalized labor relations, although much more fragmented than in Germany, were not achieved before the New Deal reforms of the mid-1930s (Bok and Dunlop, 1970; Schönhoven, 1987). Most of the unions within and outside the American Federation of Labor (AFL, 1886) were organized for much longer along the lines of the traditional crafts, although not as long as the UK unions of the Trades Union Congress, which nevertheless managed to dominate the Labor Party for the first two-thirds of the twentieth century. The AFL and the CIO were reunited in 1955. In France, Southern Europe, and other parts of the world there has often been more competition between unions of different ideological and political 'colors,' which have close ties to their respective parties.

Many European interest groups of the third wave, particularly in agriculture, also saw a model in the American Populists. It is interesting to see, however, that the sequences of different modes of organization and political performance of the agricultural interests in Germany and the United States have somehow been inverse. In Germany before 1933, the prevailing patterns went from ordinary lobbying to protest and mobilization; in the United States it has been the other way, some limited revivals of protest in the Great Depression notwithstanding. The Populist revolts were followed by extensive and efficient lobbying techniques as established by the American Farm Bureau Federation (1919), now also taken up by the older National Grange (1867) and National Farmers' Union (1902). In the United Kingdom and Scandinavia, traditional sectoral lobbying has prevailed throughout. In France, one could always find several different modes of operation simultaneously – lobbying in the channels of notable politics by the big producers, mostly in the north, which organized themselves in catholic and radical groups, and protest and mobilization on the part of the smaller producers, notably the winegrowers, mostly in the south. Interest representation has hence been more fragmented, and a second tier of special product interests (associations spécialisées) has been influential. Since the 1950s, the coordinating function of the Fédération Nationale des Syndicats d'Exploitants Agricoles has become stronger. Small peasants and laborers in the French South were eventually organized in socialist unions, and in parts of Italy and Spain also in anarchist and anarchosyndicalist unions (Hervieu and Lagrave, 1992).

Systems of Interest Intermediation

The most common modes of interaction of the interest groups are still those taking place within the traditional nation states. The various interest groups interact with one another and with the legislative and executive branches of the governments, in national (occasionally also regional) systems of interest intermediation, the organization of which follows distinct patterns. These patterns depend, among other things, on the respective intensities and traditions of democratic and parliamentary government, and state interventionism. They usually oscillate between the two poles of more liberal and market-oriented pluralism, and more regulated and institutionalized corporatism, be it in the authoritarian 'state-corporatist' tradition (mostly of fascist extraction) or in the democratic variant of 'societal corporatism' or 'neocorporatism.'

Modern corporatism was originally defined by Schmitter as "a system of interest representation in which the constituent units are organized into a limited number of singular, compulsory, noncompetitive, hierarchically ordered, and functionally differentiated categories, recognized or licensed (if not created) by the state and granted a deliberate representational

monopoly within their respective categories in exchange for observing certain controls on their selection of leaders and articulation of demands and supports" (Schmitter and Lehmbruch, 1979, p. 13).

Among the essential prerequisites for the emergence of corporatist mechanisms or systems are usually counted the existence of peak associations, strong reformist labor movements, a propensity for state interventionism, and a readiness for interparty cooperation, if not consociationalism, in the political arena (Lijphart, 1984; Puhle, 1984). In reality, however, this 'pure type' of corporatism has been as rare as its opposite, the nineteenth-century model of liberal pluralism.

The systems of interest intermediation of most developed countries seem to have evolved into one of the two most frequent 'mixed' patterns. One is an organized and more or less regulated pluralism, often with strong oligarchic tendencies, which usually began with 'regulation by the regulated,' as in the United States during the Progressive Era. The other is a syndrome displaying corporatist mechanisms in a number of important sectors, not, however, adding up to a full-fledged corporatist system except in very few countries, such as Sweden, Austria, or the Netherlands in their 'classical' postwar periods. Occasionally, the two patterns even seem to converge: the 'iron triangles,' observed in certain sectors until the late 1990s, of a pluralist pattern of interest intermediation such as the American, came close to the classical 'tripartite' arrangements between industry, labor (both ideally represented by peak associations), and the state in corporatist systems (Streeck, 1994).

The degree of institutionalization of the mechanisms of interest intermediation could also vary on a sliding scale of at least four different intensities. These range from almost none (lobbying in the open political market) to codified advisory or deliberative functions of interest representation – either ad hoc, as in parliamentary hearings, or on a permanent basis, as provided by some constitutions and in the economic, social, and regional councils of the European Union – to coordinating and regulating agencies jointly organized by the interests and the state after the model of the agencies and councils established in Germany and the United States during the World War I and revitalized by the New Deal or by the (moderately neocorporatist) 'Konzertierte Aktion' of the early 1970s in Germany. A variety of corporatist arrangements in other countries would also fit into this category. The highest level of institutionalization with compulsory participation, binding decisions, and some general jurisdiction has been reached in sectoral interest representation in councils or chambers with delegated public powers, and in the different regimes of institutionalized labor relations with 'tripartite' participation, in many industrial countries. In addition, in the United States and in Europe some of the newly established legal programs for the empowerment of women or minorities, although they do not recognize interest groups, have given some of the respective groups a quasipublic monitoring function.

Secular Changes

At the end of the twentieth century, both the interest groups and the established systems of interest intermediation have been affected by two secular processes of change, first from 'more organized' to 'less organized' politics, since the 1970s, and second from an almost exclusively national frame of reference to a more transnational set, particularly since the 1990s. First, corporatism as well as the quasicorporatist functions of the 'iron triangles,' since the economic crisis of the 1970s, have been significantly weakened by the increased tendencies toward fragmentation and segmentation, of interests and politics, under the impact of decentralization and deregulation, 'new lobbyism,' and the emergence of new public interest groups. This has particularly been the case in the United States. In Europe, the most salient features of this process have been the losses in membership and the declining influence of the labor unions, and the policies and politics pursued in order to reconstruct, and often retrench, the (various types of) social and welfare states (Streeck, 2009).

A second process of change that sped up since the 1990s has significantly added to the transnationalization of interest group politics, within a broader spectrum of restructured constellations of the public sphere, under the impact, particularly, of the use of the Internet, the breakthrough of the 'network society' (Castells, 1996), with all its technological, economic, and organizational potentials, and of advanced globalization and new interdependencies on a worldwide scale. The nation state ceased to be the only (or in a number of areas the most important) addressee of interest groups, and frame of reference of the systems of interest intermediation. The impacts of 'globalization', however, went two ways. On the one hand, it has provoked resistance; triggered processes of fragmentation or regionalization; and given rise to new 'antiglobalist' interest groups, movements, and alliances (Barber, 1995), which now could be organized more easily, less formally, and less contained by national borders.

On the other hand, advancing globalization has reduced the regulatory potential of the nation states, and hence the influence of national interest groups and organizations. At the same time, transnational regulatory regimes have increasingly gained significance, such as in international criminal law, human rights, the rules of the World Bank system, World Trade Organization, fair trade, anticorruption, or labor standards (not to speak of the European Union). All these processes were further intensified by the repercussions of the great economic crisis of the first decade of the twenty-first century, although they were different in different parts of the world (e.g., crisis of the 'neoliberal' paradigm, rising protest organization, increased state interventionism coordinated transnationally). Therefore, on the whole, the agents of influence were encouraged to organize and operate more transnationally, even if the emerging transnational systems of interest intermediation so far have not yet reached the density and the degree of institutionalization of their national precursors (Schneider et al., 2006). The great exception here appears to be the unfolding of the system of interest politics of the European Union, which, after contained beginnings in the agricultural sector (the early shift of influence from the national associations of agriculture to the European interest group Comité des Organisations Professionnelles Agricoles (COPA) is a good example for 'form follows function'), from the late 1970s, has continuously developed into a full-fledged new system of interest representation and intermediation, triggering a whole new industry of social science research (e.g., Eising and Kohler-Koch, 2005;

Greenwood, 2011; Streeck and Schmitter, 1991). During the first decade of the twenty-first century, informed estimates counted around 2500 interest groups and 15 000 lobbyists in Brussels, a bit less than in Washington, DC (estimated between 13 000 and 20 000). Two-thirds of the lobbyists represented European trade federations, consultants, and firms, and only 10% national associations (Berkhout and Lowery, 2008).

Functions for Democracy

Together with the political parties, interest groups are the core agents of structured intermediacy indispensable to the functioning of a free society and a democratic polity. They contribute essentially to the aggregation, articulation, and representation of interests and serve as vehicles for identification, mobilization, and participation, and also as transmission belts for support and legitimation. On the whole, most modern interest groups during the nineteenth and twentieth centuries have contributed to the extension and preservation of liberal democracy, and to social and political changes. The existence of distinct and different organized interests has been a key prerequisite of the recognition of pluralism, for setting up channels and rules for the democratic game, and for checks and balances cf., eg., the 'classics' (Bentley, 1949; Key, 1964; Truman, 1962). Active and dynamic interest groups and 'associations' of all kinds (and under various labels, such as NGO, NPO, CSO) have contributed to consolidating the mechanisms of civil society, and their participatory and mobilizing energies have helped to revitalize political systems in cases of crisis and stagnation.

The essence of these functions has not been affected by the secular changes taking place at the end of the twentieth century, even if the interest groups, like the political parties and the mechanisms and institutions of representative government in general, have come under additional pressure by the rise of new competitors, participatory ad hoc mobilization, fragmented associationalism or deliberative bodies, and the lines between interest groups, social movements, political parties, and service firms for political communication have (again) been blurred more than before. Following the 'cultural turn' in the social sciences, recent research on interest groups has particularly emphasized their strategies for networking (and servicing their constituencies) and their management of communication (Hoffjann and Stahl, 2010; Steiner and Jarren, 2009). However, even in a more complex environment the interest groups have remained key agents of participation, pluralism, and democracy (Beyers et al., 2008; Burbank et al., 2012; Jordan and Maloney, 2007).

Interest groups, however, cannot be considered to be democratic per se in their internal structure, or to defend and promote democracy under all circumstances. In many European countries during the first half of the twentieth century, interest groups, principally of industry, agriculture, and the mobilized crisis-ridden middle classes, substantially contributed to the breakdown and destruction of democracies, and collaborated with the dictators (e.g., in Germany, Italy, and Spain), in particular by disputing the principle of parliamentary representation and the legitimacy of political parties. As, with regard to their primary political function, many tend to be somewhat ambiguous vis-à-vis the particular form of the political order, it is important that interest groups – like political parties – be held responsive and accountable to the general rules of the democratic game, that their performance in the public arena be subject to adequate control, and that, wherever necessary, efforts are made to integrate the regime of interest intermediation into the texture of democratic institutions and traditions (which may be more problematic in the case of groups acting transnationally). In dictatorial regimes, in contrast, interest groups, despite their instrumental function of organizing and occasionally mobilizing the different social sectors around the regime party or movement, in many cases by virtue of their participatory impact, have also been serving as vehicles of incremental liberalization, which eventually produces elite splits, transfers loyalties from the regime, and contributes to its demise. Research into the transitions from authoritarian rule and the consolidation of democracy has also demonstrated that the character of interest groups and the constellations within the 'partial regime' of interest intermediation have an important function for the overall process of democratic consolidation (Schmitter, 1995).

See also: Artisans and Guilds, History of; Business History; Civil Society, Concept and History of; Democracy, History of; Industrial Relations, History of; Labor History; Political Parties, History of; Professions and Professionalization, History of; Public Sphere: Nineteenth and Twentieth-Century History; Social Movements: A Social Psychological Perspective; Welfare State, History of; Work, History of.

Bibliography

Barber, Benjamin R., 1995. Jihad vs. McWorld. Ballantine Books, New York.
Bentley, Arthur F., 1949. The Process of Government: A Study of Social Pressures. Principia Press, Bloomington, IN. 1908.
Berkhout, Joost, Lowery, David, 2008. Counting organized interests in the European Union: a comparison of data sources. Journal of European Public Policy 15 (4), 489–513.
Beyers, Jan, Eising, Rainer, Maloney, William, 2008. Researching interest group politics in Europe and elsewhere: much we study, little we know? West European Politics 31 (6), 1103–1128.
Bok, Derek C., Dunlop, John T., 1970. Labor and the American Community. Simon and Schuster, New York.
Burbank, Matthew J., Hrebenar, Ronald J., Benedict, Robert C. (Eds.), 2012. Parties, Interest Groups and Political Campaigns, second ed. Paradigm, Boulder, CO.
Castells, Manuel, 1996. The Rise of the Network Society. Blackwells, Oxford.
Cigler, Allen J., Loomis, Burdett A. (Eds.), 2012. Interest Group Politics, eighth ed. Congressional Quarterly Press, Washington, DC.
Della Porta, Donatella, Kriesi, Hanspeter, Rucht, Dieter (Eds.), 1999. Social Movements in a Globalizing World. Macmillan, Basingstoke.
Downs, Anthony, 1957. An Economic Theory of Democracy. Harper and Row, New York.
Eising, Rainer, Kohler-Koch, Beate, 2005. Interessenpolitik in Europa. Nomos, Baden-Baden.
Fraenkel, Ernst, 1964. Deutschland und die westlichen Demokratien. Kohlhammer, Stuttgart.
Greenwood, Justin, 2011. Interest Representation in the European Union, third ed. Palgrave, Basingstoke.
Halpin, Darren, 2010. Groups, Representation and Democracy: Between Promise and Practice. Manchester University Press, Manchester.
Hamilton, Alexander, Madison, James, Jay, John, 1961. In: Cooke, J.E. (Ed.), The Federalist. The World Publishing Comp, Cleveland, OH.
Hervieu, Bertrand, Lagrave, Rose Marie (Eds.), 1992. Les syndicats agricoles en Europe. L'Harmattan, Paris.

Hirschman, Albert O., 1977. The Passions and the Interests. Princeton University Press, Princeton, NJ.
Hoffjann, Olaf, Stahl, Roland, 2010. Handbuch Verbandskommunikation. VS Verlag, Wiesbaden.
Jordan, Grant, Maloney, William A., 2007. Democracy and Interest Groups: Enhancing Participation? Palgrave, Basingstoke.
Keck, Margaret E., Sikking, Kathryn, 1998. Advocates beyond Borders. Advocacy Networks in International Politics. Cornell University Press, Ithaca, NY.
Key, Valdimer O., 1964. Politics, Parties, and Pressure Groups. Crowell, New York. 1942.
Lijphart, Arend, 1984. Democracies: Patterns of Majoritarian and Consensus Government in Twenty-one Countries. Yale University Press, New Haven, CT.
Maloney, William A., Roßteutscher, Sigrid (Eds.), 2007. Social Capital and Associations in European Democracies. A Comparative Analysis. Routledge, London.
McFarland, Andrew S., 1984. Common Cause: Lobbying in the Public Interest. Chatham House, Chatham, NJ.
Olson, Mancur, 1965. The Logic of Collective Action: Public Goods and the Theory of Groups. Harvard University Press, Cambridge, MA.
Puhle, Hans-Jürgen, 1975. Politische Agrarbewegungen in kapitalistischen Industriegesellschaften. Deutschland, USA und Frankreich im 20. Jahrhundert. Vandenhoeck & Ruprecht, Göttingen.
Puhle, Hans-Jürgen, 1984. Historische Konzepte des entwickelten Industriekapitalismus. 'Organisierter Kapitalismus' und 'Korporatismus'. Geschichte und Gesellschaft 10, 165–184.
Schmitter, Philippe C., 1995. Organized interests and democratic consolidation in Southern Europe. In: Gunther, Richard, Diamandouros, P. Nikiforos, Puhle, Hans-Jürgen (Eds.), The Politics of Democratic Consolidation. Southern Europe in Comparative Perspective. The Johns Hopkins University Press, Baltimore, MD, pp. 284–314.
Schmitter, Philippe C., Lehmbruch, Gerhard (Eds.), 1979. Trends Toward Corporatist Intermediation. Sage, Beverly Hills, CA.
Schneider, Volker, Streeck, Wolfgang, Grote, Jürgen, Visser, Jelle (Eds.), 2006. Governing Interests: Business Associations Facing Internationalization. Routledge, London.
Schönhoven, Klaus, 1987. Die deutschen Gewerkschaften. Suhrkamp, Frankfurt.
Steiner, Adrian, Jarren, Otfried, 2009. Intermediäre Organisationen unter Medieneinfluss? Zum Wandel der politischen Kommunikation von Parteien, Verbänden und Bewegungen. In: Marcinkowski, Frank, Pfetsch, Barbara (Eds.), Politik in der Mediendemokratie. VS Verlag, Wiesbaden.
Streeck, Wolfgang (Ed.), 1994. Staat und Verbände. Sonderheft 25, Politische Vierteljahresschrift. Westdeutscher Verlag, Opladen.
Streeck, Wolfgang, 2009. Re-forming Capitalism. Institutional Change in the German Political Economy. Oxford University Press, Oxford.
Streeck, Wolfgang, Schmitter, Philippe C., 1991. From national corporatism to transnational pluralism: organized interests in the single European market. Politics and Society 19 (2), 133–164.
Truman, David B., 1962. The Governmental Process: Political Interests and Public Opinion. Knopf, New York. 1951.
Ullmann, Hans-Peter, 1988. Interessenverbände in Deutschland. Suhrkamp, Frankfurt.

Interest, Psychology of

K Ann Renninger and Rose K Pozos-Brewer, Swarthmore College, Swarthmore, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In psychology, interest refers to both the psychological state of engagement and a relatively enduring predisposition to reengage particular content (e.g., mathematics) over time. Interest positively influences attention, goal setting, and learning strategies for people of all ages both in and out of school. This article describes characteristics of interest as well as issues of measurement, specifically the need for indicators that account for differences of context, age, and experience. A retrospective case study provides a basis for distinguishing among interest and: motivation, engagement, goals, identity, self-efficacy, self-regulation, attitude, disposition, grit, and mind-set.

Introduction

In everyday usage, interest describes attraction or preference (Valsiner, 1992). However, in psychology, interest refers to both the psychological state of engagement and the relatively enduring predisposition to return to engagement with particular content (e.g., mathematics, music, tennis) (Hidi and Renninger, 2006). Researchers' questions about interest may focus on development, emotion, task features/experience, value, or vocational interest (see Renninger and Hidi, 2011), and can refer to topics, situational interest, individual interest, or four phases in the development of interest: triggered situational, maintained situational, emerging individual, and well-developed individual interest (Hidi and Renninger, 2006; see **Table 1**).

In the following sections, we (1) survey working definitions of interest, (2) describe characteristics of interest, (3) provide a retrospective case study of interest development and use it to distinguish interest from other variables that are often used interchangeably with interest but are distinct concepts, (4) set out indicators for the measurement of interest and its development, and (5) summarize the impact of interest as a cognitive and motivational developmental variable.

Working Definitions

Presumably because of its relation to everyday experience, interest has a number of meanings, even within the field of psychology. Thus, those studying interest need to be clear about their usage and the meaning of others' usage, and avoid creating new terms if at all possible. Here, working definitions of the terms topic interest, situational interest, individual interest, and phases in the development of interest are described, with pointers to the present and/or historical context of their use.

Briefly, topic interest refers to a subtopic of a domain. For example, cells and plants are two topics in biology and a student may be interested in cells but not plants, or vice versa. Topic interest can be either a situational or an individual interest (Ainley et al., 2002).

Situational interest (also known as an earlier phase of interest development, or less-developed interest) refers to the triggering of attention that can occur in both earlier and later phases of interest development. The triggering of interest can be a response to surprise, complexity, novelty, uncertainty, or challenge, which are described by Berlyne (1960) as collative variables. Triggers for attention may also naturally occur in the environment (e.g., a thunder clap), come from the organization of instruction (e.g., hands-on activity), or be self-generated (e.g., strategies developed to stay on task). If situational interest is maintained, it will lead to reengagement over time and to the development of interest. There are two phases of situational interest: triggered situational interest and maintained situational interest (see **Table 1**).

Individual interest (sometimes called personal interest, a later phase of interest development, or more-developed interest) refers to interest that is developed beyond situational interest. It is characterized by voluntary and frequent reengagement with a discipline over time, the ability to pursue the discipline independently, and the ability to seriously engage the discipline by asking and seeking answers to curiosity questions (questions that are novel to the learner even if they are not novel to the discipline; see Renninger, 2000, 2010). There are two phases of individual interest: emerging individual interest and well-developed individual interest (see **Table 1**).

The Four-Phase Model of Interest Development describes phases in the development of interest that were identified based on a review of the empirical literature (Hidi and Renninger, 2006; see **Table 1**). The model includes two phases of situational interest and two phases of individual interest.

Phases (not stages) of interest characterize the development of interest across age levels and contexts, both in and out of school. They are described as phases because using the word 'stage' would suggest that interest develops as an invariant sequence, which is not always the case. Although interest can continue to develop, it can also go dormant if novelty and challenge are unavailable (Bergin, 1999). The continued development of interest depends on the quality of interactions with other people and of conditions in the environment (e.g., the nature of tasks, the presence of tracking, and press for achievement) (e.g., Frenzel et al., 2010).

Researchers often study either situational or individual interest as a specific phase in the development of interest. They also research earlier (situational) phases and later (individual)

Table 1 The four phases of interest development (Hidi and Renninger, 2006): definitions and learner characteristics

	Phases of interest development			
	Phase 1: Triggered situational	*Phase 2: Maintained situational*	*Phase 3: Emerging individual*	*Phase 4: Well-developed individual*
Definition	• Psychological state resulting from short-term changes in cognitive and affective processing	• Psychological state that involves focused attention and persistence over an extended period, and/or reoccurs and persists	• Psychological state *and* the beginning of relatively enduring predisposition to seek reengagement with particular classes of content	• Psychological state *and* a relatively enduring predisposition to reengage particular classes of content
Learner characteristics	• Attends to content, if only fleetingly • Needs support to engage from others and through instructional design • May experience either positive or negative feelings • May or may not be reflectively aware of the experience	• Reengages content that previously triggered attention • Is supported by others to find connections among their skills, knowledge, and prior experience • Has positive feelings • Is developing knowledge of the content • Is developing a sense of the content's value	• Is likely to independently reengage content • Has curiosity questions that leads to seek answers • Has positive feelings • Has stored knowledge and stored value • Is very focused on his or her own questions	• Independently reengages content • Has curiosity questions • Self-regulates easily to reframe questions and seek answers • Has positive feelings • Can persevere through frustration and challenge in order to meet goals • Recognizes others' contributions to the discipline • Actively seeks feedback

Reprinted with permission from Renninger, K.A., Su, S., 2012. Interest and its development. In: Ryan, R.M. (Ed.), Oxford Handbook of Motivation. Oxford University Press, New York, pp. 167–187.

phases of interest, or the development of interest across one or more phases.

Characteristics of Interest

Regardless of the type of interest on which researchers focus, there is general agreement about five of its characteristics (Renninger and Hidi, 2011). First, interest is content or object specific, meaning that interest is directed toward a particular activity, field of study, or goal. As Krapp and Prenzel (2011) explain, the learner is interested in something or someone.

Second, interest involves a relationship between learners and their environment. Interactions with others and the available resources of the environment can determine if interest continues to develop or not. The connections that the learner makes lead to deepened experience, described as magnification (Silvia, 2001) or the developing of knowledge and value components in the learner's interest development (Renninger, 1990, 2000).

Third, interest is composed of affective and cognitive components, namely affect or feelings, value, and knowledge (Hidi and Renninger, 2006). These components are present in each phase of interest, although affect may be high and knowledge and/or value may be more minimal in earlier phases of interest development. As interest develops, the levels of knowledge and value increase.

Fourth, learners may be unreflective and unaware of their interest while they engage. For example, a history student could feel excited and attentive when studying geography, but might not recognize that feeling as interest. If the student has a less-developed interest for history, he or she might not even consider it possible that interest could be triggered; whereas, a student with a more-developed interest for history might be so involved that he or she is not reflectively aware of his or her interest.

Fifth, interest has a physiological/neurological basis. Humans are hardwired to engage in seeking behaviors (Panksepp, 1998; see also Hidi, 2012). With the development of interest for particular content, they seek resources and opportunities that allow them to stretch present understanding. Characteristic behaviors include asking and seeking answers to curiosity questions (questions novel to the person who is asking them) (see Renninger, 2000, 2010).

Retrospective Case Study

In the retrospective case study that follows, we describe a student whom we call Eva. Her interest in mathematics is detailed and then analyzed in order to provide a context for considering the meaning of interest in relation to (1) its development and to (2) other terms that are distinct from interest, but often used as though they were synonymous.

Overview

During high school, Eva developed an interest in mathematics and completed seven levels of mathematics in 4 years. However, when she began high school, her only goal was to complete calculus for college applications, and she had little, if any, interest in mathematics.

Eva was the youngest child in a home where she played games that involved problem solving, such as chess and sliding-block puzzles, and engaged in a lot of open-ended LEGO construction. She liked school mathematics assignments when she was working on them but did not look forward to mathematics. She did not realize she was offtrack in learning

mathematics until she was in sixth grade and her parents and a number of the parents of other students got together and demanded that the school teach their children the mathematics they needed for middle school.

In seventh grade, the students were given sixth- and seventh-grade-level textbooks and were expected to work independently to complete them. They were to finish one section of the textbook per week, and could work on any section that they wanted at any time. Eva recalls that her teacher checked to see that students completed the sections assigned the first month and did not check the homework after that. However, Eva's mother would sit with her everyday after school and help her to work through problems in the textbooks.

Eva remembers being a little excited when her mother used colored blocks to help her learn positive and negative numbers, but she also reports that any excitement she experienced was short-lived. She was just doing what she was told to do. She did not do any extra work on problems that she did not understand, nor did she think about mathematics in her free time.

In ninth grade, Eva's counselor told her that she needed to take a course in algebra, and preferably a course in trigonometry, in order to go to college. She was initially placed in the second half of algebra, but was dropped back into the first half of the course because she struggled and did not put very much effort into learning the concepts. The first half of algebra was a good fit to Eva's mathematics abilities and the teacher was very welcoming and receptive of her timid questions. In fact, the support and encouragement that she felt led her to enroll in the second half of the course right away, even though she had not planned to take the second half of algebra until the following year.

The next term was pivotal. Eva describes this teacher as making the concepts of algebra 'come alive.' She appreciated that he provided clear explanations and she knew he liked teaching algebra. He was also the AP Calculus BC teacher, and in the course of teaching, he provided examples of how algebra was used in calculus. Eva describes herself as fascinated by algebra and excited to go to class everyday. She also reports that she began feeling confident about her mathematical abilities.

By the end of the year, Eva was determined to work her way up to AP Calculus BC. She was curious about calculus, and had come to love the challenges of mathematics. By the time Eva enrolled in trigonometry, she was voluntarily spending extra time on her mathematics homework and thinking about other ways to approach the problems assigned. Outside of school, she started reading articles about mathematics (e.g., about Fibonacci sequences in nature, such as the spirals in pinecones) and played with making circular arcs like the spirals in her artwork. Mathematics had become fun to think about. She observes that she 'sort of forgot' that she was supposed to be completing mathematics classes in order to go to college; it was only on hindsight that she realized that to someone else her interest in mathematics, and the number of courses she completed, might seem extraordinary.

Analysis

Eva's experience illustrates the range of conditions that can trigger and support interest to develop as well as differences between having a less- and a more-developed interest. It is not inconsequential that: she chose to engage games that involved problem solving in her free play as a child; her parents demanded that there be more of a mathematics program in the school; her mother would sit with her each day to support her to do her mathematics homework; she was placed in a classroom matched to her level of readiness to do mathematics; one teacher was receptive to her timid questions; and another teacher made connections between algebra and calculus, supporting her to see how topics of mathematics are related and how her studies of mathematics might unfold.

In recounting her experience, Eva does not think of her free play at home as a form of problem solving, or that its connection to mathematics was possibly dormant through middle school because of the way she was introduced to and taught mathematics. She reports being somewhat interested when she was doing mathematics in elementary school, but also says she only did her mathematics homework because she was supposed to. She did not think that she was very good at mathematics and does not think she began to understand mathematics as a discipline until she took algebra. In other words, Eva is not aware that she might have had an interest in mathematics before the second term of algebra, and even then, the fact that she enjoyed working on mathematics in and out of school is something that she had not thought about. On reflection, she remembers that her attention was triggered but not sustained by the colored blocks or the value her parents held for mathematics. However, her interest was triggered and then maintained by the support she received from her high-school teachers.

Interest, Engagement, and Motivation

When learners have interest, or are supported to develop interest, they are engaged. They are committed to learning and participating; they also are motivated in that they have identified goals that they seek to accomplish (see Järvelä and Renninger, 2014). Eva had little interest in mathematics. She could be said to have been engaged, in that she was going through with requirements to participate and self-regulate (e.g., Christenson et al., 2008). She could also be considered motivated. She had a goal to complete enough mathematics so that she could be competitive in her application to college, and she weighed the expectancies of task interest, utility, importance, and the analysis of cost in relation to her given goal (e.g., Wigfield et al., 2006).

Eva's approach to mathematics changed once her interest for mathematics was both triggered and sustained. There had been multiple potential triggers to support her to develop an interest in mathematics (e.g., her mother's help, the goal to complete enough mathematics to be competitive for college), yet Eva's sense is that it was the combination of these and the help she received from her teachers to both work with mathematics and see the tangible connections between the topics she was working with that made the difference.

Although interest can be supported to develop at any age and in any context, there are differences in the types of supports that learners need at different ages. For example, younger children are often ready to explore unfamiliar disciplines and

activity. However, beginning around 8–10 years of age, when learners begin comparing their accomplishments with those of other peers (e.g., Harter, 2006) or to their own capacities in disciplines that are already developed interests, additional and different supports to continue to persevere may be needed (Renninger, 2009; see also Durik et al., 2015).

Interest, Goals, Identity, Self-efficacy, and Self-regulation

Research has demonstrated that there are reciprocal relations between interest and other variables including goals (e.g., Harackiewicz et al., 2008), identity (e.g., Renninger, 2009), self-efficacy (e.g., Hidi et al., 2002), and self-regulation (e.g., Hidi and Ainley, 2008). When learners have a less-developed interest for a discipline, they may or may not be aware that their interest could be triggered and developed, or that their interest has been triggered. They are often unable to set goals for themselves in working with the content of potential interest or to participate productively without guidance from other people (e.g., Eva's mother sitting with her as she did her mathematics homework) or from the design of the tasks that they engage (e.g., the open-ended assignment to complete texts independently in any order that the student chose was not appropriate for Eva).

Although learners with little interest may be on task and thus appear to be engaged, their learning requires effort and they are less likely than those with more-developed interest to persevere to get through frustrations associated with completing challenging tasks. Like Eva, learners with little interest also have little self-efficacy (e.g., Bandura, 1997), or confidence about their abilities to successfully accomplish tasks in a content area that is not already a developed interest. They may be timid about asking questions, if they seek help at all. They also may or may not be in a position to self-regulate (Sansone et al., 2012).

When learners have a more-developed interest, on the other hand, they can be expected to have and/or to be developing the attention, goal setting, and learning strategies needed to seriously engage content. Like Eva following the second half of algebra, they can be expected to expend effort without it feeling effortful (Dewey, 1913), and can be expected to have developing feelings of self-efficacy about working with mathematics and growing abilities to self-regulate so that they can persist to work through challenging problem sets.

Interest, Attitude, Disposition, Grit, and Mind-set

Once Eva took the second half of the algebra course, her attitude, or belief, about mathematics was positive (e.g., Ajzen and Fishbein, 1980). Her interest was maintained and it had begun to develop. However, attitude differs from interest in several respects. First, attitude is a belief. Interest is not a belief, although a learner can have beliefs about identified interests. In the earliest phases of interest development, a person may or may not be aware that his or her interest has been triggered. Interest is a psychological state during engagement with particular content and in later phases a relatively enduring predisposition to reengage that content. Second, research on attitude distinguishes between those with positive or negative feelings. Interest, in contrast, is described as being triggered by both negative and positive emotions (see Hidi and Harackiewicz, 2000). Finally, whereas liking is an indicator of attitude, it is not considered to be sufficient as an indicator of interest, nor is it synonymous with interest because 'liking' may or may not cooccur with wanting and learning (see Berridge et al., 2009; Hidi, 2012).

Although Eva developed an interest in mathematics, the disposition (e.g., Gresalfi, 2009), or habit, to persevere to complete the tasks on which she was working was present even when her interest had not yet developed. She was able to learn, and her disposition meant that she did her homework, participated in class, and persevered (Carr and Claxton, 2002). However, the difference in the quality of her work with mathematics before and after her interest developed was marked. Before, she worked to complete mathematics assignments and did not think about them after they were completed. Once her interest developed, she was engaged in thinking about mathematics and exploring mathematics in reading and in art, as well as in her mathematics class.

Eva also had grit, or the ability to persevere (Duckworth et al., 2007). Initially, Eva persisted to complete the work that was assigned because it was expected of her. Once her interest was triggered and began to develop, she began to develop a passion for the subject and she persevered because she wanted to. As Fulmer and Frijters (2011) report, individuals who have a more-developed interest in a task are more likely to both enjoy it and persist through challenge, whereas those with less-developed interest are more likely to avoid challenge. With the shift to developing an interest for mathematics, Eva was thinking about and doing mathematics voluntarily and was continuing to develop her understanding of mathematics whether she had assignments or not.

It may be that because Eva was supported to have a growth mind-set (Dweck, 1999), a belief that she could learn, that she completed the assigned mathematics and had the disposition and grit to persevere (see Dweck et al., 2011). Unlike interest, mind-set is a belief about self and basic qualities that can affect behavior – Eva felt that she could learn and did the work that was required. Unlike mind-set, interest is not a belief; rather, in earlier phases of interest development, a learner may or may not be reflectively aware that his or her interest has been triggered and that it could be developing, and in later phases of interest, a learner can be so engaged that he or she may not be aware of the extent of this involvement (see Renninger and Hidi, 2011). Like interest, mind-set is considered to be malleable. However, interest is context dependent and mind-set is studied as a more general characteristic of the learner.

Because interest develops through interactions with context, the development of interest is idiosyncratic. Its development is dependent on the nature of potential triggers and whether these are recognized and responded to as opportunities. The development of interest has been found to reliably vary for groups based on participant characteristics: age (e.g., Neitzel et al., 2008), discipline (e.g., Denissen et al., 2007), gender (e.g., Baram-Tsabari and Yarden, 2009), goals (e.g., Harackiewicz et al., 2008), and feelings of self-efficacy (e.g., Durik et al., 2015).

Eva's situation is not the same as the next person's, and yet there are regularities in the phases of the development of her

interest. Retrospective analysis makes it appear easy to identify shifts in the phases and characteristics of interest. However, it is challenging to measure interest prospectively, especially in the earlier phases of interest development when a person may not even consider that he or she may have or could develop an interest, and cannot necessarily respond accurately on self-report measures about his or her behaviors. In the next section, we survey current approaches to the measurement of interest, and provide some detail on the nature of indicators that together have been used to assess later phases of interest development and which, in their absence, define earlier phases.

Measurement

There are no established or agreed-upon methods for measuring interest and its development, as is the case with many motivational variables (see Murphy and Alexander, 2000). Moreover, the gaps between the ways in which interest is conceptualized and the methods used to study interest are not necessarily recognized (Krapp and Prenzel, 2011). This may be because empirical reporting does not typically involve researchers in describing the relation between their conceptualization and their choice of measures (Renninger and Hidi, 2011). It may also be because the questions on which interest researchers work range from questions about interest as a dependent variable to the impact of interest as an independent variable. Also, the disciplinary contexts, learning environments, and age groups on which this research is focused vary, which necessarily affects the methods used because different groups may interpret questions and respond to methods differently.

In this section, we outline sources of data on interest, indicators of developmental interest, and give examples of items that can be used to distinguish between less- and more-developed interests. We focus on less- and more-developed phases of interest given the idiosyncratic nature of interest development. The development of interest can occur quite quickly and may also involve slow and steady change (see Renninger and Riley, 2013).

Sources of data for studying interest have included artifacts (e.g., lab notebooks), class enrollment, computer log data (e.g., log-file analysis, data collected in online modules), experience sampling, in-depth semistructured interviews, running ethnological records, self-reports, forced choice, Likert ratings, open-ended items, teacher or researcher observations of participation, think-alouds, and/or video. However, the most common source of information is the survey (e.g., Chen et al., 1999; Eccles et al., 1993; Linnenbrink-Garcia et al., 2010; Marsh et al., 2005; Rotgans and Schmidt, 2011; Schiefele et al., 1993).

Of note, although four phases in the development of interest have been identified retrospectively (e.g., Lipstein and Renninger, 2007) and there are current efforts to identify all four phases of interest prospectively in domains such as second-language learning (Cabot and Chouinard, 2012) and by generating multidimensional interest profiles (Ely et al., 2012), the idiosyncratic nature of interest mandates clarity about the indicators selected and consideration of how they apply in the settings being studied. Present understanding suggests the utility of indicators that reflect developed forms of interest in a discipline and the use of this information as a basis for assessing how developed interest is.

Four indicators of participants' developed interest include whether their engagement is frequent, deep, voluntary, and independent relative to their other engagements (Renninger and Wozniak, 1985; see the discussion in Renninger and Su, 2012). While other motivational variables such as goals, self-efficacy, and self-regulation and learner characteristics such as attitude, disposition, grit, or mind-set are related to and/or can be expected to affect interest, these variables are not the primary targets of the four indicators of developed interest.

Assessments that focus on interest development ideally address activity with particular content (e.g., mathematics) over time, but may not. They also would assess comparison among the activities of and contents engaged by the learner (rather than between-learner comparisons), learning settings in which both independent and collaborative work/play can be observed, and open-ended activity (unprompted activity). They may also need to acknowledge that depending on age and/or experience, learners may not be meta-aware of their activity, which affects their ability to provide self-reports. In completing surveys, they also may work either with outside Likert items (e.g., 1's and 5's) or inside (e.g., 2's and 4's), and/or may not interpret item stems in the way that researchers intended (see Frenzel et al., 2012).

Drawn from different age groups of participants and learning contexts, examples of the types of data that can provide information about each of the identified indicators include video and running written records of young children in free play who also completed experimental tasks (Renninger, 1990), online questionnaires, log-file analysis, artifact analysis, and follow-up interviews employed to study teachers working online with modules (e.g., Renninger, et al., 2011), and survey and think-aloud data from a study of middle-school-age students working with mathematics and reading tasks (e.g., Renninger et al., 2002).

Frequency of engagement can be assessed using running records that provide consecutive and distributed instances of participation, log files of participant work online that can be used to track the types and frequency of particular participation (e.g., in discussion, with resources), and surveys that include ratings about the particular activity being studied (e.g., mathematics) in addition to ratings of other types of activity (e.g., reading, baseball) to provide confirmation of relative frequency of engagement.

Examples of data that allow assessment of the depth of participant engagement include the exploration of a preschooler who begins trying to identify at what point a block will slide down the slope of a roof he or she is building. Measurement of work with tasks can address the presence of developed exploration, accuracy, and/or complexity.

The indicators of voluntary and independent reengagement provide a check on whether a participant is likely to reengage voluntarily or independently; it is not expected that all engagement is voluntary or independent. Rather, it is expected that engagement is not just in response to assignment or expectation (voluntary) and that it is not only social or

undertaken with others. Data that indicate that the participant is reengaging work with particular content voluntarily may be most easily assessed with behavioral data that allow multiple opportunities for the learner to reengage. During free play, this might include notes about whether engagement occurs by choice when it is neither required nor a response to social opportunity such as being invited to a birthday party. Online, collected data from artifacts or in log files allow content analysis of discussion participation, journaling, and/or problems worked as part of assigned work. On a survey, when reengagement cannot be studied directly, participants can be asked how likely they are to engage particular activity outside of school that is not assigned (e.g., How likely are you to do math problems that are not assigned? How likely are you to talk about math outside of work you need to do for class?).

Independent reengagement with particular activity can be assessed using video and running written records of free play by identifying consecutive and distributed instances of engagement with different play objects, and online, using log files that allow tracking the type of participation (e.g., discussion, use of resources, etc.). On surveys, independent reengagement can be determined by using Likert or open-ended questions that ask about the content of study and also about other activities that provide a contrast, and/or direct inquiry about whether the respondent works on the task independently.

Features of participant groups and learning environments vary and require adjustment of how indicators of interest development are employed. Thus the decision to use one or another data source and the adjustment of each of the four indicators need to account for the age of participants, the content or context of activity including interactions with educators and independent activity, multiple domains or contents, the learner's meta-awareness of interest, the learner's self-representation, the level of granularity of interest under study, and the timescale of the study.

Impact

Regardless of their particular focus, studies of interest indicate that developing interest is beneficial to motivation, engagement, and learning. They also indicate that interest can be supported to develop at all ages.

The development of interest has been repeatedly found to beneficially affect learner attention, goal setting, learning strategies, and achievement. Interest influences whether and how learners pay attention to content (e.g., McDaniel et al., 2000; see Hidi, 1995). As Durik et al. (2015) observed, learners with little interest who need to learn content are learning for reasons other than interest. At least initially, they may have difficulty attending to the task as it is intended and may need triggers for interest that guide their attention to particular features. Triggers for interest can help the learner make connections to the content and then sustain these connections long enough so that learning can take place (Ainley et al., 2002). Palmer (2009), for example, reported that there are a number of different types of situational interest associated with classroom activity, but for his high-school students, novelty was the strongest trigger for interest. As Durik et al. (2015) and Renninger (2010) also pointed out, the triggers that work for learners may differ based on the learners' phase of interest.

The presence of interest indicates the likelihood that learners have goals, or curiosity questions that they want to answer, even if these are not clearly or fully articulated. In later phases of interest, it can be expected that they will work independently to address these. As Harackiewicz et al. (2008) report, goal setting appears to be proportionally related to a person's level of interest. Learners with more-developed interest are more likely to set goals for themselves and work to realize these goals than are learners with less-developed interest. Interest also beneficially affects learning strategies. With the development of interest, the learner engages in deeper levels of processing, employs a larger range of strategies, makes decisions about when and which strategies to employ, and is able to persevere and reorganize an original plan as his or her work with a task unfolds (e.g., Schiefele and Krapp, 1996; Renninger et al., 2002).

Finally, interest and achievement are often linked, presumably because of the tailored practice that often accompanies a developed interest (e.g., Azevedo, 2013), the kind of practice that allows the learner to continue to develop his or her skills, further solidifying the basis for the ongoing development of their interest. Wigfield and Cambria (2010) called for specific consideration of age-related development in learners' interest, values, and goals in achievement outcomes, noting that it is critical to understand the relation among these variables over time. For example, in a cross-sectional study of learners between 6 and 17 years of age, Denissen et al. (2007) reported a stronger association between self-concept of ability and school grades when grades are also considered, despite findings indicating that interest is coupled with self-concept of ability. The strong relation between interest and self-concept of ability is consistent with the expectancy–value model (Eccles et al., 1983), in which the relation between expectancy and value and between self-concept of ability and interest are expected. The relation between interest and self-concept of ability is also consistent with discussions in which value has been attributed to factors other than performance. Denissen et al. (2007) also pointed out that students who perform well sometimes do not have a developed interest in the discipline (e.g., Renninger et al., 2002). However, Sansone and her colleagues (Sansone et al., 2012) have reported that grades may not predict level of learner achievement because interest can also lead a learner to focus on mastery and not necessarily on performance.

While findings to date point to the role of interest in improving learner success, how achievement is conceptualized both by the researcher and by the learner is critical, as is how achievement and interest are assessed. Data reported by Frenzel et al. (2012) suggest that learners can respond to the same item differently depending on age. Wigfield and Cambria (2010) pointed specifically to the need for conceptual clarity in items on surveys and in interviews, additional methods to complement data provided by self-reports, and person-centered methodological approaches (e.g., cluster analysis and profile analysis) in addition to variable-centered analysis.

Next steps in refining our understanding of the effects of interest involve several challenges related to measurement. Given that interest is a content-specific motivational variable, measurement needs to account for differences of interest for particular content areas, as well as for age- and experience-related development. Measurement that distinguishes between learners with less- and more-developed interests may also be useful, as there is increasing evidence that these are often distinct profiles due to the idiosyncratic nature of interest development. In this article, we describe four indicators that can be adjusted and used together with a variety of methods and data sources to assess less- and more-developed interests across age groups and contexts. Researcher agreement on indicators of interest development could help to offset the potential complications of measurement invariance introduced by differences of context, age, and experience.

Acknowledgments

We are appreciative of the thoughtful comments we received on earlier drafts of this article from Mary Elizabeth Talian and Suzanne Hidi. We thank Melissa Running for her editorial assistance. We also gratefully acknowledge support for our work from the Swarthmore College Faculty Research Fund.

See also: Academic Engagement; Affect-Regulation Motivation; Dimensional Comparisons and Their Consequences for Self-Concept, Motivation, and Emotion; Expectancy-Value-Cost Model of Motivation; Grit; Motivation and Actions, Psychology of; Motivation, Learning, and Instruction; School Achievement: Motivational Determinants and Processes; Self-Regulated Learning: Theories, Measures, and Outcomes; Wanting vs Needing.

Bibliography

Ainley, M., Hillman, K., Hidi, S., 2002. Gender and interest processes in response to literacy texts: situational and individual interest. Learning and Instruction 12, 411–428.

Ajzen, I., Fishbein, M., 1980. Understanding Attitudes and Predicting Social Behavior. Prentice-Hall, Upper Saddle River, NJ.

Azevedo, F.S., 2013. The tailored practice of hobbies and its implication for the design of interest-based learning environments. The Journal of the Learning Sciences 22 (3), 462–510.

Bandura, A., 1997. Self-efficacy: The Exercise of Control. W. H. Freeman, New York.

Baram-Tsabari, A., Yarden, A., 2009. Identifying meta-clusters of students' interest in science and their change with age. Journal of Research in Science Teaching 46 (9), 999–1022.

Bergin, D., 1999. Influences on classroom interest. Educational Psychologist 34 (2), 87–98.

Berlyne, D., 1960. Conflict, Arousal, and Curiosity. McGraw-Hill, New York, NY.

Berridge, K.C., Robinson, T.E., Aldridge, J.W., 2009. Dissecting components of reward: "liking", "wanting", and learning. Current Opinion in Pharmacology 9 (1), 65–73.

Cabot, I., Chouinard, R., April 2012. Impact of an interdisciplinary pedagogical intervention on interest, engagement and performance in French. In: International Conference of the Learning Sciences Paper Presented at the Meetings of the American Educational Research Association, Vancouver, BC.

Carr, M., Claxton, G., 2002. Tracking the development of learning dispositions. Assessment in Education: Principle, Policy & Practice 9 (1), 9–37.

Chen, A., Darst, P.W., Pangrazi, R.P., 1999. What constitutes situational interest? Validating a construct in physical education. Measurement in Physical Education and Exercise Science 3, 157–180.

Christenson, S.L., Reschly, A.L., Appleton, J.J., Berman, S., Spanjers, D., Varro, P., 2008. Best practices in fostering student engagement. In: Thomas, A., Grimes, J. (Eds.), Best Practices in School Psychology, fifth ed. National Association of School Psychologists, Bethesda, MD, pp. 1099–1119.

Denissen, J.H., Zarrett, N.R., Eccles, J.S., 2007. "I like to do it, I'm able, and I know I am": longitudinal couplings between domain-specific achievement, self-concept, and interest. Child Development 78, 430–447.

Dewey, J., 1913. Interest and Effort in Education. Riverside, Boston.

Duckworth, A.L., Peterson, C., Matthews, M.D., Kelly, D.R., 2007. Grit: perseverance and passion for long-term goals. Journal of Personality and Social Psychology 92 (6), 1087–1101.

Durik, A., Hulleman, C., Harackiewicz, J.M., 2015. One size fits some: Instructional enhancements to promote interest. In: Renninger, K.A., Nieswandt, M., Hidi, S. (Eds.), Interest and K-16 Mathematics and Science Learning, in and Out of School. American Educational Research Association, Washington, DC.

Dweck, C.S., 1999. Self-theories: Their Role in Motivation, Personality and Development. Psychology Press, Philadelphia.

Dweck, C.S., Walton, G.M., Cohen, G.L., 2011. Academic Tenacity. White paper prepared for the Gates Foundation, Seattle, WA.

Eccles, J., Adler, T.F., Futterman, R., Goff, S.B., Kaczala, C.M., Meece, J., Midgley, C., 1983. Expectancies, values and academic behaviors. In: Spence, J.T. (Ed.), Achievement and Achievement Motives: Psychological and Sociological Approaches. Freeman, San Francisco, CA, pp. 75–146.

Eccles, J., Wigfield, A., Harold, R.D., Blumenfeld, P., 1993. Age and gender differences in children's self-and task perceptions during elementary school. Child Development 64, 830–847.

Ely, R.B.W., Ainley, M.D., Pearce, J., 2012. Establishing the interests of young people, a new exploratory approach: the My Interest Now for Engagement (MINE) Project. In: Proceedings of the International Conference of the Learning Sciences 2012, 'Assessing Interests in the Service of Supporting Personalized Learning through Networked Resources'. International Conference of the Learning Sciences, Sydney, Australia.

Frenzel, A.C., Goetz, T., Pekrun, R., Watt, H.M.G., 2010. Development of mathematics interest in adolescence: influences of gender, family, and school context. Journal of Research on Adolescence 20 (2), 507–537.

Frenzel, A.C., Pekrun, R., Dicke, A.-L., Goetz, T., 2012. Beyond quantitative decline: conceptual shifts in adolescents' development of interest in mathematics. Developmental Psychology 48 (4), 1069–1082.

Fulmer, S.M., Frijters, J.C., 2011. Motivation during an excessively challenging reading task: the buffering role of relative topic interest. Journal of Experimental Education 79 (2), 185–208.

Gresalfi, M.S., 2009. Taking up opportunities to learn: constructing dispositions in mathematics classrooms. Journal of the Learning Sciences 18 (3), 327–369.

Harackiewicz, J.M., Durik, A.M., Barron, K.E., Linnenbrink-Garcia, L., Tauer, J.M., 2008. The role of achievement goals in the development of interest: reciprocal relations between achievement goals, interest, and performance. Journal of Educational Psychology 100 (1), 105–122.

Harter, S., 2006, The self. In: Damon, W., Lerner, R.M. (Series Eds.), Eisenberg, N. (Vol. Ed.), Handbook of Child Psychology; Social, Emotional, and Personality Development, vol. 3. Wiley, New York, pp. 505–570.

Hidi, S., 1995. A re-examination of the role of attention in learning from text. Educational Psychology Review 7, 323–350.

Hidi, S., June 2012. Interest and rewards. In: Paper Presented as Part of the Discussion Session, What Does Research Suggest about the Development and Deepening of Interest? What Still Needs to Be Figured Out? Jean Piaget Society, Toronto, Canada.

Hidi, S., Ainley, M., 2008. Interest and self-regulation: relationships between two variables that influence learning. In: Schunk, D.H., Zimmerman, B.J. (Eds.), Motivation and Self-regulated Learning: Theory, Research, and Application. Erlbaum, Mahwah, NJ, pp. 77–109.

Hidi, S., Harackiewicz, J.M., 2000. Motivating the academically unmotivated: a critical issue for the 21st century. Review of Educational Research 70 (2), 151–179.

Hidi, S., Renninger, K.A., 2006. The Four-Phase Model of Interest Development. Educational Psychologist 41 (2), 111–127.

Hidi, S., Berndorff, D., Ainley, M., 2002. Children's argument writing, interest and self-efficacy: an intervention study. Learning and Instruction 12, 429–446.

Hidi, S., Krapp, A., Renninger, K.A., 1992. The present state of interest research. In: Renninger, K.A., Hidi, S., Krapp, A. (Eds.), The Role of Interest in Learning and Development. Lawrence Erlbaum, Hillsdale, NJ, pp. 433–446.

Hoffmann, L., 2002. Promoting girls' learning and achievement in physics classes for beginners. Learning and Instruction 12, 447–465.

Järvelä, S., Renninger, K.A., 2014. Designing for learning: interest, motivation, and engagement. In: Sawyer, D.K. (Ed.), Cambridge Handbook of the Learning Sciences. Cambridge, New York, pp. 668–685.

Krapp, A., Prenzel, M., 2011. Research on interest in science: theories, methods, and findings. International Journal of Science Education 33, 27–50.

Linnenbrink-Garcia, L., Durik, A.M., Conley, A.M., Barron, K.E., Tauer, J.M., Karabenick, S.A., Harackiewicz, J.M., 2010. Measuring situational interest in academic domains. Educational and Psychological Measurement 70 (4), 647–671.

Lipstein, R., Renninger, K.A., 2007. "Putting things into words": 12–15-year-old students' interest for writing. In: Boscolo, P., Hidi, S. (Eds.), Motivation and Writing: Research and School Practice. Kluwer Academic/Plenum Press, New York, pp. 113–140.

Marsh, H.W., Köller, O., Trautwein, U., Lüdtke, O., Baumert, J., 2005. Academic self-concept, interest, grades, and standardized test scores: reciprocal effects models of causal ordering. Child Development 76, 397–416.

McDaniel, M.A., Waddill, P.J., Finstad, K., Bourg, T., 2000. The effects of text-based interest on attention and recall. Journal of Educational Psychology 92 (3), 492–502.

Murphy, P.K., Alexander, P.A., 2000. A motivated exploration of motivation terminology. Contemporary Educational Psychology 25, 3–53.

Neitzel, C.L., Alexander, J.M., Johnson, K.E., 2008. Children's early interest based activities in the home and subsequent information contributions and pursuits in kindergarten. Journal of Educational Psychology 100, 782–797.

Palmer, D.H., 2009. Student interest generated during an inquiry skills lesson. Journal of Research in Science Teaching 46 (2), 147–165.

Panksepp, J., 1998. Affective Neuroscience: The Foundations of Human and Animal Emotion. Oxford University Press, New York.

Renninger, K.A., 1990. Children's play interests, representation, and activity. In: Fivush, R., Hudson, J. (Eds.), Emory Cognition Series, Knowing and Remembering in Young Children, vol. III. Cambridge University Press, New York, pp. 127–165.

Renninger, K.A., 2000. Individual interest and its implications for understanding intrinsic motivation. In: Sansone, C., Harackiewicz, J.M. (Eds.), Intrinsic Motivation: Controversies and New Directions. Academic Press, San Diego, CA, pp. 373–404.

Renninger, K.A., 2009. Interest and identity development in instruction: an inductive model. Educational Psychologist 44 (2), 1–14.

Renninger, K.A., 2010. Working with and cultivating interest, self-efficacy, and self-regulation. In: Preiss, D., Sternberg, R. (Eds.), Innovations in Educational Psychology: Perspectives on Learning, Teaching and Human Development. Springer, New York, pp. 158–195.

Renninger, K.A., Hidi, S., 2011. Revisiting the conceptualization, measurement, and generation of interest. Educational Psychologist 46 (3), 168–184.

Renninger, K.A., Riley, K.R., 2013. Interest, cognition, and the case of L__ in science. In: Kreitler, S. (Ed.), Cognition and Motivation: Forging an Interdisciplinary Perspective. Cambridge University Press, New York, pp. 325–382.

Renninger, K.A., Su, S., 2012. Interest and its development. In: Ryan, R. (Ed.), The Oxford Handbook of Human Motivation. Oxford University Press, New York, pp. 167–187.

Renninger, K.A., Ewen, E., Lasher, A.K., 2002. Individual interest as context in expository text and mathematical word problems. Learning and Instruction 12, 467–491.

Renninger, K.A., Wozniak, R., 1985. Effect of interest on attentional shift, recognition, and recall in young children. Developmental Psychology 21, 624–632.

Renninger, K.A., Cai, M., Lewis, M., Adams, M., Ernst, K., 2011. Motivation and learning in an online, unmoderated, mathematics workshop for teachers. Educational Technology, Research and Development 59 (2), 229–247. Special Issue: Motivation and New Media.

Rotgans, J.I., Schmidt, H.G., 2011. Situational interest and academic achievement in the active-learning classroom. Learning and Instruction 21 (1), 58–67.

Sansone, C., Smith, J.L., Thoman, D.B., MacNamara, A., 2012. Regulating interest when learning online: potential motivation and performance trade-offs. The Internet and Higher Education 15, 141–149.

Schiefele, U., Krapp, A., 1996. Topic interest and free recall of expository text. Learning and Individual Differences 8, 141–160.

Schiefele, U., Krapp, A., Wild, K.-P., Winteler, A., 1993. Eine neue Version des Fragebogens zum Studieninteresse (FSI). Untersuchungen zur Reliabilität und Validität [A new version of the Study Interest Questionnaire (SIQ)]. Diagnostica 39, 335–351.

Silvia, P.J., 2001. Interest and interests: the psychology of constructive capriciousness. Review of General Psychology 5, 270–290.

Valsiner, J., 1992. Interest: a metatheoretical perspective. In: Renninger, K.A., Hidi, S., Krapp, A. (Eds.), The Role of Interest in Learning and Development. Erlbaum, Hillsdale, NJ, pp. 27–42.

Wigfield, A., Cambria, J., 2010. Students' achievement values, goal orientations, and interest: definitions, development, and relations to achievement outcomes. Developmental Review 30, 1–35.

Wigfield, A., Eccles, J., Schiefele, U., Roeser, R., Davis-Kean, P., 2006. Development of achievement motivation. In: Lerner, R., Damon, W. (Series Eds.), Eisenberg, N. (Vol. Ed.), Handbook of Child Psychology; Social, Emotional, and Personality Development, vol. 3. Wiley, New York, pp. 933–1002.

Interest: History of the Concept

Johan Heilbron, Centre Européen de Sociologie et de Science Politique de la Sorbonne (CESSP-CNRS-EHESS), Paris, France; and Erasmus University, Rotterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

As a central notion for understanding human behavior, the concept of interest emerged in early modern European political theory. It was part of the skeptical view of human affairs that informed the secular approach to politics and government, which arose anew in the Renaissance. In the course of the seventeenth century, the concept entered a variety of other intellectual genres as well, and obtained a new and crucial significance by becoming linked to a theory of civil exchange. To view human behavior as interest driven was not merely considered a matter of political realism, it came to be valued as socially beneficial as well. This new conception of human society was subsequently elaborated by political economists and utilitarian philosophers. From the early nineteenth century onward, the conceptual development has occurred mainly in debates accompanying the rivalries between academic disciplines. Interest was associated primarily with economic theories, which have been emulated, enlarged, corrected, as well as contested in other disciplines.

While modern social science is inconceivable without some notion of interest, the actual concept has a long and complicated history of shifting meanings. In this relatively brief article, these changes will be highlighted by following the main stages of this conceptual development. In doing so, the notion will be understood in its specific sense, abstracting from the development of similar notions (such as utility or preference), but without isolating the concept from its place in more encompassing theoretical frameworks. The most significant changes in the history of the concept are inseparable from its relations to other concepts and to the transformation of broader theoretical structures. For a proper understanding of these changes, it is important, furthermore, to realize that they are related to reconfigurations of the intellectual groups that have been the carriers of the intellectual change.

Understanding human behavior as driven mainly by (self-) interest is not in itself very noteworthy. The idea can be found in many ancient traditions of thought, both religious and secular. Most of these understandings, however, were normative rather than descriptive or explanatory. Forms of self-interest were condemned as sinful, as socially disruptive, and they were generally opposed to either religious or military virtues. Elites in agrarian societies valued religious devotion and military prowess, that is, types of behavior distinct from agricultural labor and trade or commerce. If (self-)interested behavior was not condemned as illegitimate, it was at best tolerated within very narrow social boundaries, and subsumed to more virtuous modes of behavior.

As a central term for understanding human behavior, the word interest emerged in political writings during the early modern period. It was part of the skeptical view of human affairs that informed the secular approach to politics and government that arose anew in the Renaissance. In the course of the seventeenth century, the concept entered a variety of other intellectual genres as well, and it obtained a new and strategic importance by becoming linked to a theory of civic exchange. To view human behavior as interest driven was not merely considered a matter of political realism, it now came to be valued as socially beneficial as well. This view was subsequently elaborated by political economists and utilitarian philosophers. From the early nineteenth century onward, the conceptual changes occurred mainly in debates accompanying the development of academic disciplines. Interest was primarily embedded in economic theories that have been emulated, enlarged, corrected, as well as vividly contested in other disciplines.

The Politics of Interest

The word interest, derived from the Latin *inter esse*, originally referred to procedures for compensation in Roman law. The expression *id quod interest*, meaning 'that which matters' or 'that which counts', could be applied to a variety of claims in this respect. The specific meaning of taking rent on loans was directly related to the legal notion of compensation. Interest in this particular sense came into use in many European languages during the fifteenth century, commonly as a euphemism for usury (Fuchs, 1976; Hirschman, 1992).

The more general meaning of the term emerged in the beginning of the sixteenth century. Interest then referred to the more general sense of advantage and to the human propensity for seeking benefits. It was part, not of an economic discourse, but of the skeptical anthropology and the secular approach to statecraft, appearing systematically for the first time in the work of Francesco Guicciardini (1483–1540). Guicciardini, one of the most perceptive political theorists of Renaissance Italy, frequently observes in his *Ricordi* (1512/30) that 'self-interest' prevails in nearly all human beings. This applies especially to successful leaders and others who do well in life; they always have their 'own interest in mind and measure all their actions accordingly'. Guicciardini adds, however, that 'true interest' does not necessarily reside in pecuniary advantage; it pertains, more often, to honor and to the art of knowing how to keep a good reputation.

The general meaning that Guicciardini gives to the term allows him to use it for matters of government as well. Where particular interests are in conflict with the 'public interest',

Guicciardini advocates the primacy of the latter. The conceptions of Guicciardini and his contemporary Niccolò Machiavelli (1469–1527) were developed more fully in the reason-of-state literature, which flourished in Europe around 1600. For one of the leading theorists, Giovanni Botero (1544–1617), 'reason of state' meant 'reason of interest', and state interest ought to be the supreme rule of conduct for princes and statesmen. In this view, religious and constitutional matters were to be treated as instrumental issues. For the ruler, interest of state is the only legitimate principle of action, and this interest is defined in opposition to both the interests of other states and the passions of the ruler. Control over people demands self-control; passions and other disorderly appetites need to be replaced by a rationally (ob)serving state interest.

Interests thus came to be seen as the principal motive of human behavior and as the only realistic rule of political conduct. From a predominantly critical concept, directed against ecclesiastic and humanistic virtues, it had gained a more positive meaning. While preserving its amoral connotations, interests appeared to be a more stable and more reliable motive than the passions (Hirschman, 1977). The French Huguenot leader Henri de Rohan (1579–1638) expressed this view in his influential *De l'Intérêt des Princes et des États de la Chrétienté* (1638) by stating that "princes order their people around and interest orders princes around." The prince can be deceived, his council can be corrupt, only 'interest does not lie'.

It was through the translation of Rohan's work that the term 'interest' became established in the political vocabulary in England. And it was from the political scene that the term was transferred to the market place and came to be applied to private behavior (Gunn, 1969).

In the early modern political literature, interest was not merely an analytical term. Interests were indeed perceived as the predominant motive of human behavior, but political advisors stressed that interests had to be observed in a rational and calculating way. This was not because human beings were considered to be rational decision makers, but because of the strategic advantages of rational calculation. The prudential weighing of costs and benefits was a behavioral norm for political elites rather than a form of motivational reductionism (Holmes, 1995).

Contours and Consequences of Exchange

From the political literature, written by counselors and diplomats, the notion of interest entered a variety of other intellectual genres in the course of the seventeenth century (Heilbron, 1998/2001; Lazzeri and Reynié, 1998). Among these other discourses, natural law and moral philosophy stand out; in both cases, interest-driven behavior came to be related to new conceptions of politics and society.

In modern natural law, interest was part of a foundational argument. The anthropology of the skeptics had given a seemingly universal role to self-interest and more in particular to self-preservation. Natural law theorists treated this factual observation as a natural right. On the basis of the fundamental right to assure one's self-preservation, they built a system of moral and political obligation, which was intended to overcome religious orthodoxy as well as the relativism of the skeptics.

For natural law theorists such as Thomas Hobbes (1588–1679) and Samuel Pufendorf (1632–94), people are motivated primarily by self-love. But the consequences of their view differed fundamentally. For Hobbes, a strong state was the only alternative for permanent and violent conflict between selfish individuals. Pufendorf constructed more of a historical argument. Human societies arise out of individual needs. Upon this assumption, he proposed a scheme of societal development, culminating in a form of sociability proper to 'commercial society'. Commerce at the time referred to exchange or traffic in the general meaning of these terms, not specifically to trade. The idea of a 'commercial society' designated a relatively pacified order within which various realms of exchange had been established. Contrary to Hobbes, Pufendorf thus envisioned a form of sociability based essentially on the ties of interest-based exchanges (Hont, 1987).

The very idea of a social order emerging out of the self-interested acts of individuals was more systematically elaborated in seventeenth-century French moral philosophy. In a series of remarkable essays, collected in his *Essais de morale* (1671–78), the Jansenist moral philosopher Pierre Nicole (1625–95) explained how human society could be well ordered and prosperous without assuming any religious duties or prescribing secular virtues. Following the Augustinian tradition, Jansenist theology made a strict distinction between the 'city of man' and the 'city of God'. The radically separated realms were founded on two mutually exclusive drives: self-love and love of God. Nicole and some other Jansenists and writers close to them (Pascal, La Rochefoucauld) sought somehow to adapt their uncompromising outlook to the demands of worldly life. Many of them were members of the salon of Madame de Sablé and were like Nicole searching for a *modus vivendi* for Christians who were forced to live in a corrupt world. Looking for an intermediary route between pious retreat and worldly corruption, Nicole proposed what is probably the first consistent model of commercial society.

The Hobbesian image of a war of all against all, Nicole argued, is valid only in the very beginning of humankind. Since each human being is a threat to the other and no one wants to be a victim, people unite with other people. To affirm their union, laws are established and punishment is instituted for those who violate the laws. Fear of death is therefore the first restraint on self-love. Once open violence is excluded, human beings are forced to use artificial means to satisfy their desires. These artificial means are all forms of exchange. The result of this process is a civil society, which – given the corrupt human nature after the Fall – could not be better organized, even if true religion is banned. Human traffic alone, established and regulated by self-love, had produced this result.

From the point of view of God, human society meant corruption, but from the point of view of human beings themselves, the establishment of human civility is the best possible achievement. Given the fact that very few people are prepared to withdraw from society, Nicole added that 'enlightened self-love' could be considered the most effective policy in human affairs.

For Nicole and his fellow Jansenists, human beings could not be credited with any virtue. As in the maxims of

La Rochefoucauld (1613–80), they depicted what seemed virtuous (courage, love, friendship) as being but a refined expression of self-interest. In the words of La Rochefoucauld: "Interest speaks every sort of language, and plays all sorts of roles, even that of disinterest." Since human virtue is a kind of narcissistic illusion, self-interested exchange was the most realistic alternative for the threat of civil war and a sufficient basis for an orderly society.

The argument about civil society as a process of generalized exchange on the basis of self-interest also provoked an historically important counterargument. In his critique of Jansenist anthropology, the philosopher François Fénelon (1651–1715) asserted that 'disinterestedness' was not beyond human reach. Human beings may actually experience pure and 'disinterested' love, which is close to a state of selflessness. The love of God is the highest form of this disinterested love, but Fénelon acknowledged more worldly forms as well, especially in the spontaneity of children. The term 'disinterestedness' and the argument about its possible significance were formulated in the context of a theological dispute, the *querelle du pur amour*, but it soon gained importance in other social domains, first in aesthetics, somewhat later in other areas. When Immanuel Kant at the end of the eighteenth century defined esthetical experience as disinterested pleasure, *interesseloses Wohlgefallen*, he took up the notion of disinterestedness, which British moral philosophers like Shaftesbury had taken from the French dispute over Jansenism.

Political Economy and Utilitarianism

Many of the economic and utilitarian arguments advanced in the eighteenth century were derived from these seventeenth-century writings in moral philosophy. Well before Adam Smith, Pierre de Boisguilbert (1646–1714) transformed Nicole's reasoning into the more specific idea of a self-regulating economic order. On the condition that this order could function freely, it would tend toward an equilibrium. That being the case, Boisguilbert spoke of 'laissez-aller', advocating to let this natural order run its course, thus unambiguously announcing the liberal consequences of Nicole's model for the economic realm (Faccarello, 1999).

Bernard Mandeville's tale about 'private vices' as the actual source of 'public benefits' was similarly rooted in the work of French moralists. It was not disinterestedness or self-denial that produced civil society, but vanity, avarice, and ambition. Inspired by Mandeville (1670–1733) and his predecessors, Adam Smith (1723–90) famously pointed out that it is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest. "We address ourselves, not to their humanity but to their self-love, and never talk to them of our necessities but of their advantages." Society, according to Smith, "may subsist among different men, as among merchants, from a sense of its utility, without any mutual love or affection."

The historical significance of Smith and other classical political economists was that they transformed the argument about self-interested exchange into an economic theory of self-regulating markets and the idea of harmonizing economic interests through the 'invisible hand'. But when the general argument about exchange narrowed down to a mechanism of markets, the political questions implied in the doctrine of interest reemerged. Utilitarianism was one of the attempts to rethink the political and moral issues involved.

Emerging in the eighteenth century and becoming an intellectual movement with Jeremy Bentham (1748–1832) and James Mill (1773–1836), the utilitarian system was built on the principle that every human being seeks pleasure and avoids pain. Human conduct is universally guided by the ideas and feelings that are associated with either one of these emotions. This behavioral principle was applied not only to individuals, but extended to the polity as a whole. Like individuals, governments should promote the amount of happiness and reduce the amount of pain. Moral arithmetic, based on the principle of the greatest amount of happiness of the greatest number, thus provided the means for assessing the utility of public institutions and initiating political reform.

The rise of political economy and utilitarian philosophy provoked several counter movements in the decades around 1800. In their opposition to the French revolution and the ensuing waves of democratic reform, conservatives appealed to the luster of tradition in order to counter the cold-blooded politics of interest. Romantics similarly found ways to reaffirm the power of the passions.

Disciplinary Social Science and Interdisciplinarity

When the social sciences during the nineteenth century became university disciplines, theories of interest were associated specifically with economics, which – with the exception of Marxism – was increasingly separated from the other social sciences. The broader framework of moral philosophy disintegrated and made way for a disciplinary division of labor. John Stuart Mill (1806–73) redefined the status of political economy by arguing that while different causes operate in society, some of the more powerful causes need to be studied separately. Political economy was legitimately restricted to the phenomena based on the desire of wealth and the law that a greater gain is preferred to a smaller. In Mill's view, political economists do not deny other motives, but merely abstract from them; self-interest is a methodological construct rather than a behavioral reality.

During the 'marginalist revolution' of the latter decades of the nineteenth century, the plea for scientific abstraction was combined with the postulate of maximization (Demeulenaere, 1996; Hausman, 1992). Economic behavior was modeled as rational calculation aimed at maximizing utility. This approach, which formed the heart of the neoclassical program, allowed much of the formal and technical developments that have characterized mainstream economics ever since. Self-interested behavior was now treated by means of a mathematical technique of maximization under constraint. With Vilfredo Pareto (1848–1923), the domain of economics was redefined as the study of choice or optimization, and in Lionel Robbins's definition economists study "human behavior as a relationship between ends and scarce means that have alternative uses."

While this definition is, in principle, applicable to all forms of human behavior, many economists were well aware of the limits of their approach. For Pareto, neoclassical economics had

a limited validity since many aspects of human behavior do not meet the conditions of rational calculation. In addition to the restrictions of their behavioral assumptions, many economists acknowledged the fact that the 'invisible hand' and *laissez-faire* politics were no panacea. For a variety of reasons, market outcomes can be at odds with the welfare of large parts of the population, and instead of repeating the elusive metaphor of the invisible hand, economists specified the functioning as well as the dysfunctioning of markets (Medema, 2009).

Since its inception during the first half of the nineteenth century, sociologists have generally contested purely economic and utilitarian approaches (Swedberg, 2005). Auguste Comte (1798–1857) introduced the word sociology to overcome what he saw as premature specialization and misguided generalization. Interests were an important motivational force, but so were the human affects. In his later writings, Comte coined the term 'altruism' for all nonegoistic forms of behavior, arguing for its primacy over self-interested behavior. Other sociologists differentiated human behavior in other ways, some of them elaborating richer notions of interest than merely individual, economic interest (social, collective, or public interest; vested interest; symbolic interest). Parallel to undertaking historical and empirical inquiries into the various forms of behavior and sociability, sociologists reacted critically to the abstract and exclusive nature of *homo economicus*. Emile Durkheim (1858–1917) and François Simiand (1873–1935) argued against the 'unrealistic abstractions' of economic theory, advocating a positive and more empirical approach that was not unlike that of institutional economics. Max Weber (1864–1920) proposed a broader understanding of interest and rationality and integrated these into a general typology of social action. Weber introduced 'ideal interests' aside from material interests, and distinguished instrumental rationality from value rationality. In addition to affective and traditional action, instrumental rationality refers to a means–end relationship, while value rationality pertains to choice as derived from values.

While the academic division of labor between the social sciences became much stricter during the middle part of the twentieth century, it has been increasingly contested since the 1970s. Economic theories and modeling procedures expanded into many areas, which were previously seen as the territory of other disciplines. Chicago economists like Gary Becker argued that the economic approach to human behavior, that is, the "combined assumptions of maximizing behavior, market equilibrium and stable preferences", should be used relentlessly and unflinchingly for all types of human action in all historical periods. Public choice economics and the 'law and economics' movement similarly extended notions of self-interest and maximization to the political and juridical sphere. In political science economic approaches to voting, political parties and government became mainstream, and rational choice, more generally, became the common denominator for a broad area of interdisciplinary work (Heap et al., 1992).

The expansion of economics into noneconomic domains in the latter part of the twentieth century was, be it on a more modest scale, accompanied by a movement in the opposite direction; psychologists, sociologists, and organization theorists gradually invaded the domain of economics, proposing alternative accounts for core issues of economic theory. Psychologists established a wide variety of behavioral phenomena, which contradict the assumptions of rational choice thus contributing to the establishment of 'behavioral economics'. Sociologists have repeatedly shown how cultural beliefs and institutional structures shape (economic) behavior and its outcomes. Diverse and varied as these contributions may be, they have at least demonstrated that the technical sophistication of economics depends on assumptions that can be questioned and contested on empirical as well as on theoretical grounds.

Enduring Ambiguities

Since its early uses in Renaissance political theory, the concept of interest and its terminological derivatives (self-interest, group interest, public interest, vested interest, and ideal or symbolic interest) have been part of what may be broadly considered the materialist tradition in the social sciences. But this realistic and hard-nosed orientation is itself divided over a number of fundamental issues. The development of the notion of interest uncovers at least three enduring ambiguities.

First, the term interest and its equivalents can be used for interpretative purposes in accounting for observable behavior; they can also be used in a normative way for defining what rational behavior is. In the former sense, interest refers to an explanatory and empirically testable principle; in the latter, it merely designates a logically possible and perhaps desirable course of action. The ambiguity of serving both descriptive and prescriptive functions has persisted ever since the origins of the term. Second, interest can refer to material advantages as well as to political, cultural, and symbolic benefits. If the broader understanding of interest is accepted, the ideas of rational calculation and maximization become problematic, since it is unclear how costs and benefits in various domains can be rigorously compared with one another. In the absence of an overarching metric, people have multiple interests without being able to rationally compare their relative benefits. Far from being exceptional, uncertainty is a defining characteristic of human choice. Third, actors may pursue their interests by relying on rational calculation, but they also do so on the basis of a process of socialization in which choice is not 'rational' in the economists' sense, but bound up with habitualization leading to the practical mastery of a social game (Bourdieu, 1990). In the former case, interest is part of rational choice theory; in the latter, it is not.

See also: Action, Theories of Social; Behavioral Economics, History of; Economics, History of; Interest Groups; Interest, Psychology of; Interests in Sociological Analysis; Political Economy, History of; Public Interest Law: The United States and Beyond; Public Interest; Rational Choice Explanation: Philosophical Aspects; Rational Choice Theory in Sociology; Sociology, History of.

Bibliography

Bourdieu, P., 1990. The Logic of Practice. Polity Press, Cambridge, UK.
Demeulenaere, P., 1996. Homo Oeconomicus: Enquête sur la Constitution d'un Paradigme. Presses Universitaires de France, Paris.

Faccarello, G., 1999. The Foundations of Laissez-Faire: The Economics of Pierre de Boisguilbert. Routledge, London.
Force, P., 2003. Self-Interest before Adam Smith: A Genealogy of Economic Science. Cambridge University Press, Cambridge.
Fuchs, H.-J., 1976. Interesse. In: Ritter, J., Grunder, K. (Eds.), Historisches Worterbuch der Philosophie, vol. 4. Wissenschaftliche Buchgesellschaft, Darmstadt, pp. 480–485.
Gunn, J.A.W., 1968. Interest will not lie: a seventeenth century political maxim. Journal of the History of Ideas 29, 253–276.
Gunn, J.A.W., 1969. Politics and the Public Interest in the Seventeenth Century. Routledge & Kegan Paul, London.
Hausman, D.M., 1992. The Inexact and Separate Science of Economics. Cambridge University Press, Cambridge, UK.
Heap, S.H., Hollis, M., Lyons, B., Sugden, R., Weale, A., 1992. The Theory of Choice: A Critical Guide. Blackwell, Oxford, UK.
Heilbron, J., 1998/2001. French moralists and the anthropology of the modern era: on the genesis of the notions of 'interest' and 'commercial society'. In: Heilbron, J., et al. (Eds.), The Rise of the Social Sciences and the Formation of Modernity: Conceptual Change in Context, 1750–1850. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 77–106.
Hirschman, A.O., 1977. The Passions and the Interests: Political Arguments for Capitalism before Its Triumph. Princeton University Press, Princeton, NJ.
Hirschman, A.O., 1992. The concept of interest: from euphemism to tautology. In: Rival Views of Market Society. Harvard University Press, Cambridge, MA, pp. 35–55.
Holmes, S., 1995. Passions and Constraint: On the Theory of Liberal Democracy. Chicago University Press, Chicago.
Hont, I., 1987. The language of sociability and commerce: Samuel Pufendorf and the theoretical foundations of the four-stages theory. In: Pagden, A. (Ed.), The Languages of Political Theory in Early-Modern Europe. Cambridge University Press, Cambridge, UK, pp. 253–276.
Lazzeri, C., Reynié, D. (Eds.), 1998. Politique de l'Intérêt. Annales littéraires de l'Université de Franche-Comté, Besançon.
Medema, S., 2009. The Hesitant Hand: Taming Self-interest in the History of Economic Ideas. Princeton University Press, Princeton.
Swedberg, R., 2005. Interest. Open University Press, Berkshire.

Interests in Sociological Analysis

Pierre Demeulenaere, Université de Paris-Sorbonne, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A sociological study of interests involves a definition of the concept of interests, since they can be said to vary socially. It is also concerned with the relation between interests (notably economic ones) and other human ends, for instance the values that lie behind any social order. Therefore, it implies a methodological debate about human agency: should it be considered as intrinsically selfish or unselfish, and how does one interpret any behavior that departs from the pursuit of selfish interests? This methodological debate is related to an appreciation of the different kinds of rationality that can be associated with human behavior. This article will describe mainly the way classical sociology has conceptualized the issue of interests, and the contemporary developments of those classical positions.

It is not an easy task to define the notion of interest. Throughout the tradition of sociological analysis, it is often referred to without further specification (Swedberg, 2005). In ordinary language, there are two general meanings of this particular concept: the first is that of an agent's special commitment toward some kind of object, or activity, like music. This meaning can be studied by sociology as far as it is concerned with tastes or cultural practices. Although not specifically selfish, those interests in this broad sense can nevertheless sometimes lead to conflict among persons, when the success of their realization for someone leads to an opposition to others' desires, values or interests: for instance if a street artist tags a wall in order to display what he perceives to be his artistic ability, the neighbors can feel seriously injured by such an aesthetic expression. This is a conflict of interests, although no economic or selfish dimension is directly involved. It is more a matter of aesthetic disagreement. But it is important to note that actions based on values can lead to conflict of interests.

The second and more important meaning is the one that implies some kind of selfishness in the notion of 'self-interest.' This is mainly the case with 'economic' interests, those concerned with the appropriation of scarce goods. But this idea can also be extended to noneconomic goods, like power or prestige. Wealth, power, and prestige have a common aspect: they are such that the increase of their amount for one person can often lead to a decrease of their amount for another person (what we call a zero sum game). It is not possible inside a group that prestige or power increases simultaneously for everyone, since those are relative positions. On the contrary, it is possible that wealth increases simultaneously for everyone inside a group, although the opposite can also be true: someone's wealth can increase to the detriment of another's. The two situations are very common in the social world, although people often tend to believe that only or another tends to prevail. Those selfish interests are generally described in terms of costs and benefits, or utility, with two main difficulties: first, costs and benefits cannot always be translated in monetary measures, and therefore measured at all; second various people do not necessarily have the same costs and benefits, so that the notion of costs and benefits is not something easily predictable of peoples' choices. For instance a drug consumer does not have the same costs and benefits as someone who is not interested at all in drugs. Therefore, should we consider that someone is devoted to maximize his or her utility, or maximize his or her benefits and minimize his or her costs, this would not allow us to predict his or her actual choices: costs and benefits vary according to various persons, and they vary according to social settings. Put in another way, interests can be seen as often dependent on social norms (the ones for instance that prescribe to smoke or not to smoke) or alternatively as the basis of social norms (the interest in health would lead to a norm prohibiting smoking).

Sociology and Economic Interests

The Problem of Social Order

Economics has been described as a science of human interests, when they are considered apart from any other motive, and in particular the interest for wealth. This definition is explicit in Pareto's general statement on the division of work inside social sciences. According to this description, what should the role of sociological analysis be regarding interests? Three major problems come to the forefront of the discussion.

The first one is that of social order. Any economic activity occurs within a social framework, made up of varied institutions, rules, and rights. A market, for instance, depends on property rights that allow people to exchange their goods. There are two options here. The first one is to explain this framework in terms of interests, following the tradition of Hobbes and Hume. There is a trend, especially in contemporary sociology inspired by economics, to try to revive this kind of explanation of social norms in terms of interests (Coleman, 1990). The main idea is to consider that in specific situations people have an interest in adopting social norms (for instance property rights) without which their basic interests would be threatened. Norms are based on interests and are protective of them through some limitation of their extension. This idea can be conceptualized in terms of theory of games, in a reiterated prisoners' dilemma situation that leads to the emergence and adoption of norms. In this case there are shared given interests which are the basis of the explanation of the emergence of social norms. The purpose of the norms is to reinforce the satisfaction of the same interests. However, it should be noticed that an important

part of contemporary economic reflection about property rights insists on the historical dimension of property rights that cannot be reduced to an emergence in prisoners' dilemma situations (North et al., 2009). The more important variable from a historical point of view is the relative weight of various interests, whereas in standard theory of games inequality of position regarding interests is not taken into account. Moreover, theory of games rests mainly on the postulate of the similarity of interests, whereas people, as we have seen, do not necessarily share a common conception of interests.

But the major part of sociological analysis has tried, on the contrary, to define a specific attitude toward social order that *cannot* be reduced to interests. If social order and the norms and values associated with it do not stem from individual interests, where do they come from?

Three major classical sociologists have tried to describe the problem of social order as the main area, in social life, where it seems that interests alone cannot suffice to produce the norms that unify people into a society. Moreover, social order has the role of limiting selfish interests. Comte, first, in a radical attack against what appeared to him as an 'abstract' view, namely that of social contract and economic exchange, stressed the existence of a social consensus foreign to any calculation, and prior to any economic exchange. Following Comte, Durkheim built all his sociological theory on the assumption that individual interests cannot be seen as the root of social solidarity. In a Kantian fashion, he understood morality and social norms as a repressive force against individual interests. But, in opposition to Kant, he did not find the roots of such obligations in an individual rational sense of morality, but rather derived it from a specific social level that exerts coercion upon individuals. Parsons, in his attempt to build a general theory of society, continued this tradition. He stressed the existence of common shared values on which economic interests depend. The main point that these three authors made is that selfish interests cannot always lead to cooperation. These same interests can also lead sometimes to opposition and conflict among individuals. Hence, the social primacy of cooperation should not be based only on interests. It needs a firmer basis. It is located in the functioning of the social system as such that puts pressure on individuals.

These three authors can fairly easily be grouped under the so-called holist tradition, which insists on the role of a specific social level irreducible to individual decisions, which exerts a specific constraint on individual decisions. But even in the tradition of methodological individualism, there has been a similar trend to refer interests to the values behind them that allow or restrict their development. This is notably the case of Max Weber's theorization of interests who insists on the fact that typical economic attitudes are influenced by religious beliefs. The two traditions, however, can be seen to have in common the fact that interests, related to preferences, often depend on social values, and that any interest intervenes within a social situation.

Economic Attitudes

The second classical problem is that of economic attitudes themselves. In the economic main tradition, these attitudes are generally considered from a nonhistorical point of view, as relatively constant human tendencies. For instance, in his description of economic equilibrium, Pareto makes no direct reference to any historical evolution or change of economic attitudes. For him there are only random distributions of different attitudes toward interest among different people (Pareto [1916], 1963). The *homo economicus* figure is not, in this perspective, a historical one, but a theoretical construction of typical constant attitudes in economic life (and beyond, in political life for instance when politicians try to gain more power). In most of the sociological tradition, this has been seen as irrelevant, since economic attitudes do vary historically, at least partially. Therefore, various attempts to explain the logic of these changes can be found. Such is the aim, for instance, of Mauss's ([1925]1954) *Essai sur le don*, although the gift and the potlatch rituals involve many self-interested attitudes. More generally, sociological analysis tries to relate interests to the background values that encourage or discourage them (and often constitute them). In fact, there are two aspects to be stressed from this standpoint.

The first is that the very notion of what is to be considered as an interest for an individual can vary socially. For instance, Mauss tries to show that in archaic societies there is no real interest in personal consumption, but there is a strong interest in social prestige, that is not derived from consumption, but from gifts. Therefore, the social importance of gifts can be opposed in one sense to a self-interest oriented toward personal consumption. But on the other hand, this importance of gift is linked to the interest in an increase of personal prestige. It is certainly not a purely altruistic behavior.

The second level is that, any kind of interest being admitted as such (for instance interest in wealth), there are always social values that reinforce its deployment or conversely restrict its expression. For instance, loans on interest can be authorized or forbidden in a given social community. Weber is thus famous for having, in his work on the Protestant ethic (Weber [1904–5], 1974) and on the economic ethic of world religions, stressed the influence of religious orientations on economic ones. He links the development of capitalism to a complex religious background specific to the peculiarities of western cultural tradition. Demagicalization, universalization, and moralization of relations among people, commitment to an ascetic lifestyle, and interpretation of worldly success as a sign of eschatological salvation are the key religious elements that can contribute to explaining capitalism's rise in Western societies. These religious elements are joined to other values, namely the general importance of rationality, and its effects on other institutional contexts, like the extension of rational legal systems. The specific modern capitalist attitude cuts itself from its religious roots: it becomes an acquisitive trend in a society oriented toward constant increase of goods rather than satisfaction of stable needs (Weber [1923], 1981).

There is, however, another element important in the analysis of economic attitudes: the description of tastes and preferences that lay behind interests. Parsons and Smelser (1956), for instance, in a general theorization of the relations between economics and sociology, insist upon the fact that the economic preferences that constitute interests are rooted in common social values. This point had been made earlier by Durkheim: the fact that a specific meat is allowed to be eaten

or not depends on social norms and is the basis on which individual preferences are defined. The analysis can apply also to the very constitution of economic markets and the value of profit making. Polanyi (1944) has stressed the importance of other norms than norm markets that precede them historically (norms of reciprocity and norms of redistribution), the orientation toward profit being a modern specific norm instead of a permanent human attitude.

Somehow surprisingly, this idea that economic preferences depend, at least partially, on social values and norms, can be equally found in contemporary economic analysis: beyond the idea of costs and benefits, utility maximization is derived from individual preferences. There is no utility as such, but only sets of individual choices that constitute individual utility. Therefore, individual utility can depend on social norms (this is obviously the case, for instance, for food or clothes preferences). The key issue is whether the notion of utility can be related to stable transcultural human preferences, or should be always seen as dependent on specific social settings.

The Scope of Interests

The third classical problem is that of the extension of the role of interests in social life. Economics mainly studies the production and exchange of scarce goods. But interests arise in other fields, such as political conflict, family attitudes, crime decisions, educational choices, drugs consumption, etc. The problem will be then to understand how far an analysis in terms of interests can go to explain social phenomena.

The article so far has stressed the importance of the social background that narrows or enlarges the field where interests play a role, and has also indicated that, from a sociological point of view, the very definition of interests mostly depends on a social valuation of some kind of ends.

But to this description should be added the opposite trend, the one that insists on the role of typical standard interests in several aspects of social life. This move does not only refer economic exchange to selfish interests, nor does it try to explain social values or institutions in terms of interest, but it also sometimes considers that even beliefs can be referred to interests. This can be done either with a critical stance, or only with an explanatory noncritical scope.

In the tradition of Marxist analysis, the critical attitude toward social order is related to the revelation of hidden interests not only behind oppressive institutions that serve the interests of the powerful, but also behind common moral or political beliefs and ideologies which are derived, in their proper constitution, from interests: people believe what they believe because they have an interest in doing so. The main difficulty of such a conception is that it does not take into account the specifically cognitive side of any ideology that cannot be so easily reduced to noncognitive contents and to some kind of material interests. It can be said that when people believe something, there is always a cognitive dimension of the belief that makes it acceptable for the believers given their social situation and their level of information, and the fact that it is hard to believe something that could appear to be false or completely unjustifiable. Naturally, however, beliefs can be influenced by interests and equally various unconscious biases.

But the extension of the scope of interests can also intervene without any critical intention: it is only then an effort to show that attitudes in some areas apparently removed from any economic dimension, can be interpreted in terms of interests, not only social norms, but also attitudes toward marriage, children, voting, etc. Becker's (1976) work on family or Posner's (1981) work on law and justice are representative of such ambitions. Supporters of this central role of interests in social analysis as well as its opponents have named it 'economic imperialism.' It depends on a definition of what are 'costs' and 'benefits,' against values. For instance, in fertility decisions, a child has an expected financial cost, but it has also a 'psychological' benefit; the problem is that, financial costs and psychological benefits being not equally measurable, and not easily comparable to each other, any decision can be said to be maximizing utility, since any decision is necessarily a balance among financial costs and psychological decisions. To say that it is a maximizing decision is therefore tautological, and certainly not predictive of the various choices, and explanatory of the differences among the various choices, unless we can explain why psychological benefit is not equally appreciated.

A Methodological Divide

Altruism and Selfishness

It can be said that there is a common assumption shared by the two traditions that try to explain social order in opposed fashions regarding interests: the selfishness of individuals. Individuals are considered to act according to their interests because they are selfish. This common assumption leads to two alternative results, as we have seen. The first is to consider that social order and social norms can be explained in terms of social interests, through reiterated prisoners' dilemma situations. The second is that it is not possible to do so, but, since there is a social order as well as something like a morality, which leads individuals to act morally, in an unselfish manner, it has to be referred to something different from interests.

But why should individuals be considered selfish? There are basically two different reasons for this. The first is empirical and arises from the evidence of the selfish attitudes that prevail in economic exchange, where there is a huge tendency to avoid systematically letting someone's else interest prevail over one's own.

The second element is methodological: since commitment to social values is to be explained in a way that is not tautological, it seems convenient to try to show that these norms derive from a sense of interest, that should not, for its part, be explained.

Durkheim admits straightforwardly that individuals as such are selfish when they are taken alone, but he does not justify this assumption, taken from economics. This assumption is, however, at the core of his effort to describe a specific social coercive level that should explain why individuals act in an unselfish way on certain occasions. His argument could be worded as follows: "Individuals are selfish; sometimes they devote themselves to others, but since others are just like me, I should have no reason to prefer them to myself; therefore, it is not to them that I show my respect, through my moral behavior, but to society, as a distinct entity from any

individual" (Durkheim [1906], 1996). But this statement rests on the idea that such altruistic behavior is always strange. One gives another individual what one has denied oneself, despite the fact that this other being is just another version of oneself not worthy of such an attitude.

But in fact, most often, this so-called altruistic behavior does not correspond to this description. Usually, it intervenes in three different scenarios in social life. The first relates to symmetrical attitudes of reciprocity. In this case, there is an admission of the idea that different individuals should have the same rights, and therefore, be allowed to have the same opportunity to pursue their own interests within a common framework. The respect of the right of the other is not seen as the altruistic renunciation of one's self-interest for the other's benefit. It can be derived from the acknowledgment of the fact that both individuals can be granted the same rights since they have the same interests. It is not that one favors the other at the expense of oneself. It is that everyone can understand, precisely, that since someone's interest should not systematically prevail upon the other's, the solution to this problem is the adoption of norms of reciprocity. A related analysis can be found in Harsanyi (1969). Norms of reciprocity can be based either on an utilitarian device or on the belief that one's interest should not unilaterally and systematically prevail over the other's interests. The solution to such situation is the recognition of symmetrical rights.

The second case for altruism occurs when a stronger or wealthier person helps a weaker or poorer one. But, in this case, it is obvious that if people renounce their own interests in favor of someone else, it is precisely because they are in a position to help someone else who does not have the same means. This indicates that it is not another equivalent self that is helped, but someone who is in a worse position. Such attitudes are common in social life. They can derive from the pursuit of prestige, but, more commonly, from a sense that some people need assistance, and that it is right to help them in specific situations (but not in others).

Third, it is true that sometimes there is a 'sacrifice' of the self in favor of others. Durkheim insists strongly on this kind of situation. But it could be understood, also, by considering the specific position of the person: if the captain of a ship is legally considered to be responsible for a vessel's safety because of his or her specific competence, it is therefore consistent for the captain to remain on board until everyone has evacuated. Such sacrificial attitudes are today analyzed by psychological literature (Elster, 2009).

All this means that it does not seem absolutely necessary to assume that there is some global social force behind the individual restrictions to self-interest. It is possible to understand actors' self-limitations of interest by their ability to understand the logic of a situation and the need for social rules. Therefore, three positions can be described. The first one is an individualistic and utilitarian position, that interprets most of human actions in terms of selfish interests, as in the case of Coleman's sociology.

The second is a nonindividualistic and nonutilitarian position. As we have seen, it rests upon the idea that, as individuals are selfish, and naturally committed to their own interests, the fact that they sometimes act in an altruistic manner should lead to the recognition of a specific social (not individualistic) level that is responsible for such a change of attitude and pressure on self-interest.

But since it is not obvious that individuals should definitely be considered selfish and only devoted to their own interests without the intervention of a superior social force, it seems possible to introduce a third methodological position, individualistic but nonutilitarian. Actors would have then to choose to favor their interests, or instead to avoid satisfying them, because they would be able (maybe 'rationally') to understand the appropriate course of action regarding different situations.

Rationality and Interests

It is important to introduce here the notion of rationality, which has a multisided relationship with that of interest. The classical notion of rationality that prevails in economics is an instrumental one. It designates the choice of adequate means toward ends. The idea of rationality relates to this 'correctness' of the choice of means. The important point is that ends are not integrated in this sense of rationality that concentrates on the adequacy of the choice of means.

But the rational choice model, highly influential both in economics and sociology, is not entirely clear (Demeulenaere, 1996) in this respect. On the one hand, it is officially maintained that ends, i.e., interests, are not rational by themselves, rationality being reserved to the choice of means. But on the other hand, the qualification of rational behavior in economic and sociological literature is generally applied to an attitude that does follow selfish interests, and denies respect to any kind of value or social norm that would go against them. There is also another ambiguity. On the one hand, interests are not considered to be substantial ones, as they depend on individuals' choice. But on the other hand, interests are described as substantial: money or wealth, power and prestige. Sometimes rational choice is described in terms of narrowly defined specific interests (which makes it at least predictive and falsifiable). Sometimes, it is described as a very general intentional dimension of action, which can only be opposed to unconscious attitudes.

Thus it is not clear whether this rational choice model is strongly normative or not. It would be so if it stressed that individuals are rational only if they follow their interest, in the selfish sense. But this assumption is very costly for two reasons. The first is that it is not obvious that obeying one's own interests should be the only possible rational choice. The second is that many individuals do not have the impression that they are not rational when submitting to a norm that cannot be referred to interests. On the contrary, they believe they have reasons for doing so.

Axiological Rationality

Therefore, it seems possible to introduce, in sociological analysis, a third point of view that maintains the principle of rationality, without restricting it to the pursuit of interests or to the limited choice of means toward ends. This principle of rationality should allow the actor to discriminate between interests that deserve to be followed and those that should be avoided.

How should we define this axiological rationality (Boudon, 1997)? The main point is that it would show the reasons actors have to act in a specific way, the reasons being not necessarily their interest. Boudon, for instance, shows the importance, in the voter's attitudes, of normative aspects that should not be reduced to the actor's narrow scope of interest. Those normative aspects are not arbitrary as they can be related to a system of beliefs that seems acceptable to actors regarding their general aims. Voters know that, in a large-scale election, their individual votes have no influence on the result. But they know also that other voters are aware of the same fact. If they all abstain from voting, they will lose the benefit of the democratic system. Committed to that system, they understand the importance of voting, even if it is costly in the short run and seemingly inefficient.

Such a theoretical interpretation of the actor's behavior should thus expand the concept of rationality and depart from a narrow positivistic attitude where the ends (including interests) can never be interpreted in their own rationality.

See also: Action, Theories of Social; Consequentialism (Including Utilitarianism); Decision and Choice: Bounded Rationality; Durkheim, Emile (1858–1917); Interest, Psychology of; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Macrosociology-Microsociology; Methodological Individualism in Sociology; Rationality in Society; Sociological Theory; Values, Sociology of; Weber, Max (1864–1920).

Bibliography

Becker, G.S., 1976. The Economic Approach to Human Behavior. University of Chicago Press, Chicago, IL.
Boudon, R., 1997. The moral sense. International Sociology 12 (1), 5–24.
Coleman, J., 1990. Foundations of Social Theory. Belknap Press, Chicago, IL.
Demeulenaere, P., 1996. Homo Oeconomicus. Enquête sur la constitution d'un paradigme. Presses Universitaires de France, Paris.
Durkheim, E., 1996. Détermination du fait moral. Presses Universitaires de France [1906], Paris.
Elster, Jon, 1989. The Cement of Society: A Study of Social Order. Cambridge University Press, Cambridge.
Elster, Jon, 2009. Le désintéressement. Traité critique de l'homme économique. I. Seuil, Paris.
Harsanyi, J., 1969. Rational choice models of political behavior vs. functionalist and conformist theories. World Politics 21, 4.
Mauss, M., 1954. The Gift. Cohen & West [1925], London.
North, Douglass C., Wallis et, John Joseph, Weingast, Barry R., 2009. Violence and Social Orders. Cambridge University Press, Cambridge and New York.
Pareto, 1963. The Mind and Society: A Treatise on General Sociology. Dover Publications [1916], New York.
Parsons, T., Smelser, N.J., 1956. Economy and Society. Routledge, London.
Polanyi, K., 1944. The Great Transformation: The Political and Economic Origin of Our Time. Beacon Press, Boston, MA.
Posner, R., 1981. Economics of Justice. Harvard University Press, Cambridge, MA.
Swedberg, R., 2005. Interest. Open University Press, Maidenhead, UK.
Weber, M., 1974. The Protestant Ethic and the Spirit of Capitalism. Allen and Unwin [1904–5], London.
Weber, M., 1981. General Economic History. Transaction [1923], New Brunswick, NJ.

Intergenerational Justice

Paul Kelly, London School of Economics, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Intergenerational justice poses challenges for standard theories of justice and morality because it presupposes the absence of the usual interpersonal connections that explain moral relationships. Can one use the analogy of a contract between two parties that could not interact, or where one side has an ineradicable ability to impose nonreciprocal harm? Can one have obligations to improve the welfare of those who don't exist and whose potential existence is conditional on the actions proposed? This article explores the ways standard theories such as contractarianism, utilitarianism, and communitarianism accommodate the claims of future people. It explores issues such as the nonidentity problem and various 'repugnant conclusions,' before challenging the skeptics with a threshold conception of universalism that supports environmental sustainability.

This article surveys the problem of intergenerational justice but does not address all ethical problems posed by future generations. This problem can be viewed from the perspective of how we can have obligations to those in the future and what kind of things we might owe to those in the future. The first three sections review the way traditional theories such as contractarianism and utilitarianism deal with obligations to the future. The problem of asymmetry, the nonidentity problem, and the issue of 'repugnant' conclusions are considered, by focusing on the work of contractarians such as David Gauthier and John Rawls, and consequentialists such as Derek Parfit. The final section reviews the work of Avner De-Shalit and Barry, McMahon, and Meyer who attempt to overcome these skeptical challenges and reassert the idea of obligations to the future. Avner De-Shalit uses the idea of a transtemporal constitutive community to explain the motivation for obligations to posterity. Barry, McMahon, and Meyer defend a threshold version of universalism that is indifferent to issues of temporal continuity in determining our obligations of justice. The article concludes with a review of Barry, McMahon, and Meyer's claim that our chief obligation to future generations is encompassed in the need to pursue environmentally sustainable development as part of a threshold conception of justice.

Intergenerational justice refers to the negative and positive obligations one generation owes to another. The concept of a generation is essentially indeterminate as there is no point at which a single generation stops and another begins. Within any single population a number of generations will coexist. Intergenerational justice does not refer to relationships of right and duty between different generations within one temporally continuous population. The obligations of young to old or old to young are matters of distributive justice. Intergenerational justice refers to the relationships of obligation, right, or benevolence that ought to exist between groups of people who are not temporally continuous or to that set of issues which apply to noncontinuous populations, in the same way that international justice applies to populations that do not spatially overlap.

The fact of nontemporal continuity explains the problem of intergenerational justice. Many philosophers argue that the absence of temporal continuity casts doubt on the possibility of justifying claims of obligation or right between generations. De George argues that as future people do not currently exist they cannot be the bearers of anything least of all rights or moral claims. Discussions of intergenerational justice focus on two sets of questions: first, how do we explain obligations between nontemporally continuous populations and second what obligations might we have to future generations? This article examines the bases of these doubts about extending the idea of justice to apply to the relationship between nontemporally continuous populations and then conclude with an assessment of what we might owe to posterity.

Contractualism and Asymmetry

Contemporary discussions of distributive justice are synonymous with the resurgence of contractarian thinking under the influence of philosophers such as Rawls (1971) and Gauthier (1986). The contract device is used for thinking about distributive questions in circumstances where the parties disagree about ultimate moral principles. If we cannot base our obligations to one another by appeal to a common standard such as utility maximization, we can use the idea of a rational bargain or reasonable agreement as the basis for distributive norms and obligations. The idea of a contract is a thought experiment that can be used to legitimize principles rather than an actual agreement process that groups of people have to collectively engage in. However, even as a hypothetical device the idea of a contract (in both versions of the theory) poses a major problem because of the asymmetrical relation between the participants in intergenerational cases.

The 'mutual advantage' theory argues that distributive principles can be derived from a rational bargain between participants in a cooperative venture, such as market society or political community. The issue of justice only arises when certain circumstances of justice obtain. If each individual could get everything they wanted without cooperation there would be no need for principles of justice. The circumstances of justice are threefold. First, (roughly) equal power among the parties. What one lacks in strength they can make up in guile so no one is guaranteed always getting their own way. Second, moderate

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63036-2

scarcity of goods and motivations obtains, so that social cooperation is necessary or at least in the interest of all. Finally, while all benefit from the existence of social cooperation, the participants are interested in the distribution of its benefits and burdens. As rational agents each has an interest to maximize their share of the benefits but minimize their share of the burdens of cooperation.

The problem for the 'mutual advantage' theorist is that the circumstances of justice do not unequivocally obtain between noncontinuous generations. If we use the idea of the contract within one generation we can make sense of the idea of rough equality of power and reciprocity as the key motivations for accepting the distribution of the benefits and burden of social cooperation (although we might still argue that this is a morally inappropriate way to think about justice). Yet when we extend this idea to noncontinuous generations the weakness of the method becomes apparent. In what way can we both speak of ourselves and future generations as part of the same 'cooperative' venture? If we just concentrate on the rough equality of power we can see how there is an asymmetrical relation between the generations. Take an issue such as energy policy. A present generation can unilaterally impose significant costs on future generations such as the management of toxic or nuclear wastes. I leave aside a further case where the consequences of a present decision are so egregious as to prevent there being a future generation. Does stopping potential people coming into existence constitute a harm or injustice? It is clearly possible for a present generation to take such choices without any possible negative consequences from the future generation except perhaps disapproval and poor historical reputation. The present generation is able unilaterally to affect the interests of a future generation, whereas a future generation cannot unilaterally affect the interests of the present generation, hence the asymmetry. Without rough equality of bargaining power any contractual agreement will simply reflect the advantage of the stronger party, in this case the present generation. If this is so, the contract does nothing more than allow the present generation to pursue its own interests taking account of the future only when its own reputation is at stake and when this is of higher value than any other present interest such as standard of living. Because of the asymmetrical relationship underlying an agreement it is unclear how the present generation could be motivated to assume obligations to the future given the motivational assumptions of 'mutual advantage' contractarianism. The contract not only does not generate obligations to the future, but would appear to generate reasons for ignoring such claims.

The 'impartialist' theory departs from many of the key features of the 'mutual advantage' bargain just considered. Rawls' theory introduces the idea of a hypothetical original position in which the participants deciding which principles of justice should shape the basic structure of their society as a fair system of social cooperation, are shrouded by a 'veil of ignorance.' The point of the veil of ignorance is to filter out the possibility of exploiting 'positional' advantage or inequalities of bargaining power. By denying agents behind the veil of ignorance, information about their particular circumstances and goals, and allowing them only general knowledge about society, Rawls constructs an impartial choice situation. Rational self-interest plus ignorance delivers impartiality and fairness.

Rawls' theory is a thought experiment designed to test the fairness of candidate principles of justice. The contract with the veil of ignorance is designed to model a fair choice situation. However, once again the asymmetry problem arises. The agents behind the 'veil or ignorance' are denied information that will prevent them from pursuing their own interests at the expense of each other. The agreement is effectively between members of a single generation. What is to stop the participants choosing principles in order to maximize the condition of the worst off currently in existence that will at the same time worsen the situation of those not yet born? In order to bring future generations into the agreement Rawls posits an additional motive for the participants. They are expected to reason as representatives of families who are assumed to have a concern for their descendants. Heads of families provide the intergenerational link that overcomes the asymmetrical relations between future generations. The problem with this strategy is that it merely concedes the problem of asymmetry and the inability of the contract device to generate adequate obligations to the future. At best the additional motivational assumption gives those in the present generation a reason to take account of future generations. It does not establish an obligation to future generations in the way that the contract does establish obligations within one generation. The addition of the extra motivational assumption provides a noncontractarian reason for taking account of future generations. In effect it assumes precisely what it is supposed to provide, namely a reason for taking account of the interests of future generations.

While contractarian arguments have been central to the development of current theories of distributive justice, they seem singularly ill-equipped to deal with issues of justice between generations because of the asymmetrical relationship between even hypothetical bargainers. This is a point acknowledged by Barry, one of the most significant defenders of a neo-Rawlsian contractarianism (in Dobson, 1998b). The key relationship of reciprocity and the equal ability of the partners to affect the interests of each other do not apply in the case of temporally noncontinuous generations. To overcome this, the contractarian has to appeal to noncontractarian reasons.

The Nonidentity Problem

A further ground for skepticism which is related to the asymmetry objection, but which has implications beyond contractarian arguments is the nonidentity problem. This problem has been given its most forceful statement by Parfit (1984). And it is as applicable to theories such as utilitarianism as it is to contractarian theories of justice.

The nonidentity problem draws on a paradoxical implication of the intuitive response to issues of intergenerational justice. Assume that policy makers are making decisions about technologies that deplete environmental resources or cause global warming. These decisions involve questions of intergenerational justice because they do not merely distribute the benefits and burdens of social cooperation within generations but also across generations. Any policy that unilaterally worsens the conditions of future generations

by depleting nonrenewable resources without suitable compensations for the future might be deemed to involve an intergenerational injustice as it sacrifices the interests of future generations to serve our own temporary interests. In this case the injustice consists of harming the interests of future generations by denying them opportunities or by unilaterally imposing costs upon them. Against this a skeptic might respond that our intuitive concern for the future does not withstand critical scrutiny and that we should discount the interests of future generations in determining between environmental policies.

The argument claims that only harms to assignable individuals constitute cases of injustice. An action becomes a case of injustice when it involves harm to the interests of an individual agent or groups of individuals. But if we turn to the issue of appraising environmental policies it is far from clear that they do distribute harms to assignable individuals.

Whichever environmental policy is chosen will have complex consequences for the future, many of which are indirect. One such consequence is on the procreation decisions of particular populations; that is, on who should come to mate with whom. The nonidentity problem depends upon the assumptions that any decision to procreate at one time rather than another will result in the birth of different people. The fact that I was born at the time I was and in the place I was in part made me the person I am now. If I was born either earlier or later my experiences would have been different and therefore I would have been a different person. If this claim is true then any policy which affects people's behavior will have implications for the identity of future generations. Thus, an environmental policy that forces me to use my car less and travel more on public transport may be responsible for me meeting my future partner and therefore albeit indirectly for any future persons who result from our relationship.

The important point is that whichever policy is chosen will result in different future populations and not simply different levels of welfare for the same potential population. But if different populations result from different policy choices then does it make sense to talk about obligations of intergenerational justice? We can argue that a policy of resource depletion does not involve any issue of intergenerational justice because it does not harm the interests of any assignable individual. After all whichever individuals do find themselves bearing the burdens of our choices only exist as the people they are because of choices made by us in the past. If we had chosen otherwise they would not have existed at all, but rather some other group of persons would have existed. For the contractarian the problem of showing how assignable individuals are harmed by our choices undermines the claim that we can do injustices to people in the future, and this fact renders the whole idea of intergenerational justice deeply problematic. But the problem also applies to utilitarian theories that explain the idea of injustice in terms of worsening the welfare of individuals. We cannot, for example, argue that a future population would have a higher welfare if we had adopted a policy of conservation than if we had adopted one of resource depletion for either policy would result in different populations. The issue would then become one not of whether population X had a welfare level of a or b, but whether population X should exist or whether population Y should exist. The utilitarian can respond to this claim but only at the expense of opening themselves to certain repugnant conclusions.

Utilitarianism and Repugnant Conclusions

The idea behind the nonidentity problem is that intergenerational obligations must be person regarding, that is because obligations are owed to assignable individuals. But as we have seen, as the identity of individuals in the future is itself dependent upon present actions, the idea of harming or benefiting future generations becomes problematic. Utilitarians – those philosophers who believe our most fundamental obligation is to promote, either total or average, welfare – often reject the person-regarding view of moral obligations by claiming that overall welfare is prior to questions of its distribution (see Stearns, 1972). For the utilitarian we have an overriding obligation to promote utility or welfare and this avoids the nonidentity problem because we can assess overall outcomes irrespective of the identity of the recipients of utility or welfare. This strategy raises a number of important issues when applied to intergenerational questions, but its most striking consequences arise from its application to population policy.

If we have a primary obligation to promote or maximize welfare and this obligation is prior to any obligations to assignable individuals, then we are inevitably faced with the question how many people should there be? This utilitarian obligation to promote welfare has direct implications for individual procreative decisions. Before considering whether to conceive a child or bring it into the world a couple will need to consider the impact of this decision on overall welfare. We can consider the impact of this procreative choice on either average or total welfare.

Let us first consider the case of average utility. Here we are required to consider the impact of our procreative choices on the average level of utility across the relevant population. If our society is a prosperous and happy one and the addition of further members who, although happy, would be slightly less happy than the average, then their impact would be negative.

In such a case average utilitarianism would counsel against conceiving such a child. Such a way of conceiving procreative decisions might seem suspect, but often we do apply such reasoning in thinking about population policy. Some utilitarians consider draconian population policies are justified by average utility. However, opponents of utilitarianism argue that such an approach to procreative questions is wholly inappropriate because social welfare is a morally irrelevant consideration.

Average utilitarianism does not merely have implications for whether we should allow the birth of happy children who are merely less happy than average, it also has implications for those who exist but are less happy than the average. Consider a population of 10 people two of whom are miserable while eight are happy. If we add to this consideration the fact that there is no hunger in the society so that the eight would not gain in utility if the two were removed, we could still increase the average utility by removing the two. However, we could then make similar judgments within the eight until we were left with a very small population perhaps numbering only two in which there was no difference in average utility. In existing

populations there are always groups whose elimination would, other things being equal, increase average utility.

What happens when these arguments are applied across populations, as in the case of intergenerational justice? If we accept the premise of the nonidentity argument, namely that the policies we pursue will affect the identity and size of future generations then using the average utility argument we can engage in some interesting trade-offs. If we are concerned about a policy of conservation versus one of depletion we can reason as follows. A depletion policy would have a large negative impact on a large population but a relatively low impact on a smaller future population. As long as the impact diminished the fewer people there were to be affected, we would satisfy our utilitarian obligation. This could easily be achieved by coupling the depletion policy with a draconian population policy among our existing population, such that the average utility of the small population x is greater than the average utility of population y that would have resulted from the policy of conservation. The point here is that we can make one of the aspects of the nonidentity problem a policy objective, in order to justify our depletion of resources. As long as we can manipulate the size of future populations by regulating procreative decisions we can pursue almost any policy without harming the interests of future generations.

Average utility results in some questionable conclusions, does total utility fair any better? In the case of individual procreative decisions the issue becomes one of the duties to maximize happy people. As long as each child being born is happy to some degree they contribute to the total sum of happiness. It need not be the case that they are happier than their parents only that they are happier to exist than not to exist by however small a degree. If we grant this condition, which follows from the commitment to maximizing total utility, then we are faced with an obligation to bring as many happy people into the world as possible, however small the individual level of happiness of each person is. The most striking implication for intergenerational justice is that we act justly to the future when we make the size of future populations as big as they can be irrespective of how low the average utility of each member of those future generations might be. Whereas average utility suggests the imposition of particularly draconian population policies, total utility goes to the opposite extreme by encouraging population maximization. The full impact of this repugnant conclusion can be seen when we apply the total utility strategy to the choice between policies of depletion versus conservation. We need to be able to judge the relative utility or welfare of one population against another who would have lived if an alternative policy had been pursued. As long as the total utility of population a is greater than that of population b that might have existed as a result of a different policy choice, then we are justified in pursuing policy a over b. This can help us explain and justify some trade-offs between population size and level of utility such that we might prefer a policy that resulted in a larger population with a lower utility to a smaller population with a higher utility; for example, population a with n members each having a utility of 10 as opposed to population b with $2n$ members each having a utility of 8. But this approach becomes counterintuitive and repugnant when we consider the case of a third population of $10\,000n$ each having a utility of 0.01. In this latter case we face the repugnancy of the total utilitarian approach whereby as long as the population is large enough the utility of the individual members does not matter as long as it is above zero.

By rejecting the person-regarding aspect of justice arguments, utilitarianism can overcome the criticism that we only have obligations to assignable individuals (and therefore we cannot be said to have obligations to future generations) because of the nonidentity problem. They get around the problem only by exposing themselves to different problems resulting from their indifference to issues of distribution that are also at the heart of intergenerational justice arguments. Both forms of utilitarianism considered here address the issue of intergenerational justice by the artificial means of manipulating population size. This strategy either loads the dice in favor of existing generations as long as they are prepared to manipulate the size of future generations accordingly or to reduce our obligations to the simple duty to procreate as much as possible. Neither approach seems an adequate answer to the question of what if any obligations we have to future generations. The two main candidate theories for grounding claims of justice, contractualism, and utilitarianism have serious problems in making sense of the identity of future generations and the idea that we may have obligations to them. Does this mean that the skeptic is right and the idea of intergenerational justice is flawed?

The Prospects for Intergenerational Justice

The contemporary literature on intergenerational justice is largely focused on discussions of the paradoxical consequences of applying existing theories of justice to intergenerational problems. Does this mean that the skeptic is correct in arguing that the idea of relationships of justice between generations is incoherent? Two contemporary political philosophers, De-Shalit (1995) and Barry (in Dobson, 1998b) reject this skeptical position and suggest that there are ways we can overcome the skeptical position to make sense of intergenerational obligations. However, they do so from very different perspectives, De-Shalit is a communitarian, whereas Barry is a universalist egalitarian.

Transgenerational Community

De-Shalit begins with the challenge posed by the inadequacy of contractarianism and the nonidentity problem. Faced with the problem of asymmetry in the original position, Rawls introduces an additional motivation linking the generations to justify a just savings principle as a constraint on contemporary consumption. This additional motive that requires the participants to think as heads of families is supposed to model the motive to be concerned about our posterity. De-Shalit takes this natural motive and generalizes it across the idea of constitutive communities. Why should the family be the only repository of such a motive? His point is that it is not the particular institution that is the repository of this motive but the type of relationship. The contractarian and the nonidentity argument posit a sharp break between generations that creates an unbridgeable moral gap between persons. De-Shalit challenges this individualization by arguing that our identities are shaped and constituted by membership of communities.

Membership is itself obligation creating, in the same way that Rawls assumes that acting as head of families will provide the sufficient motive to assume obligations to the future. The motivation to be concerned about posterity is for De-Shalit a natural motivation and this is missed by the narrowly 'individualistic' approaches of contractarians and utilitarians. The task for the political philosopher is then not to construct the basis of the obligation to the future. We have this natural concern because our own identity and interests are tied up with the membership and existence of a certain type of community. The task is to expand the implications of that concern for posterity and the specific obligations that can be derived from it. Two important implications arise from De-Shalit's communitarian turn. First, he broadens the notion of those goods we have a duty to bequeath to posterity. Whereas much of the literature assumes merely that our obligations to the future concern material resources and the environment, De-Shalit's emphasis on constitutive communities entails that all those definitive aspects of such communities might be 'goods' that we have an obligation to preserve for posterity. Thus, we could have obligations to preserve languages and cultures or defining institutions such as constitutions and political rights. The second implication is that De-Shalit changes the focus of intergenerational justice from a universal to a particular one. We have obligations to our posterity and not future generations as such. The British people might have obligations to preserve their constitution for future generations, or the Quebequois might have an obligation to pass on their 'distinct society' to future Quebequois, but not to future Nigerians or for the good of future Japanese. For De-Shalit it is not assignable individuals that we need to worry about but only assignable groups. Obligations only extend as far into the future as it is reasonable to think a constitutive community will retain its identity.

One possible objection to De-Shalit's approach is that it makes group membership too important. If we only have obligations to our own posterity then we only have obligations to preserve the environment for our posterity. But this raises the counterintuition that it surely cannot be right that I have a duty to protect the environment for my own posterity but that if a policy of depletion falls not on my posterity but on another people then I have no obligation or responsibility. A second objection that there is not sufficient space to explore in full is how far De-Shalit's argument creates obligations to the past to sustain community supporting practices that define the relevant political community? We might find the idea of preserving the US constitution as an obligation that present-day Americans have to the founders partly comprehensible, although Thomas Jefferson might have found this puzzling. But would we find the idea of submission to or support of the Church of England as compelling an obligation although as an institution it has had an important role in constituting the British Polity. It is because of such 'perverse' consequences that many are reluctant to abandon moral universalism, or the idea that from a moral point of view spatial and temporal location are not ultimately significant.

Universalism Defended

The key intuition underlying the threshold universalism of Barry, McMahon, and Meyer is that place and time do not provide morally relevant bases on which to discriminate between the interests of those who fall under a principle. This assumption is morally basic, as such it must be defended indirectly by challenging the challenges to it and showing how it accords with our reflectively adjusted intuitions. We might, for example, show that it overcomes the narrow parochialism of De-Shalit's communitarianism. Alternatively, we might suggest that it is the intuition that underlies Rawls' additional motivation designed to supplement contractarianism. The universalist is also likely to reject the utilitarian variant because of the repugnant conclusions that follow from abandoning a person-respecting morality. A universalist might deny that we have an obligation to bring about maximal well-being if this results in the kinds of population policies that result in repugnant conclusions.

One simple way of avoiding this utilitarian paradox is to conceive of obligations of justice not in terms of promoting the good as maximal welfare but in terms of not harming the interests of the future generations. This is nicely illustrated by an example from Woodward (1986). An individual is denied a seat on an airplane flight on the basis of race/skin color. The plane subsequently crashes with the death of all on board. In such circumstances we would be disinclined to say that the act of discrimination reduced the individual's welfare or utility but we are likely to still consider the discrimination a harm to an interest of the individual.

This negative obligation to not harm the interests of others has implications for the redistribution of resources and benefits; it does not merely involve staying one's hand and refraining from reducing welfare. In this way we can argue that we have an obligation to give weight to the interests of future generations when we make irreversible consumption decisions. We ought not merely discard future people because they are in the future.

The major objection still facing this approach is the non-identity problem. We can only have obligations to assignable individuals but the identities of these individuals are contingent on the choices we currently make such that we cannot be said to be harming them. Referring to 'future generations' does not mark out the identity of those being harmed. One way to overcome this problem is to contest the idea that only assignable individuals can be harmed. Thus, it might make sense to say that our decisions about resource depletion can harm future generations relative to the absence of that policy even though the existence of particular individuals might be dependent upon that policy because there is a threshold below which any individual, actual or possible, should not fall. The threshold conception allows us to contest the idea implicit in the non-identity problem that we cannot be said to harm the interests of members of a future generation as their interests will be a function of the circumstances in which they find themselves. While it is certainly true that any complex account of universal transtemporal interests will be hard to conceive of, there are certain things that it can be argued will be in any subsequent individual's interest. It is clearly in no future individual's interest to live in a world in which potable water is made scarce by pollution, or in which ozone damage raises levels of carbon dioxide in the atmosphere and levels of harmful radiation. Future generations are also likely to place some value on their environment and biodiversity. How much each of these things

will matter is debatable, but that they will matter is uncontestable. Where this leaves the universalist is with the idea that we have obligations to take account of the 'general' interests of future generations not to be harmed in certain ways, whoever they happen to be and whatever else they happen to want. Thus, the burdens of intergenerational justice fall on contemporaries in the form of not harming those interests of future generations by over consumption, environmental depletion, and unsustainable development without permanent advances to any future generation.

The threshold conception of universalism can then be linked to other distributive criteria such as egalitarianism and prioritarianism of sufficientariansm to identify the most plausible account of intergenerational justice. An egalitarian view might set the threshold very high as it would require us to ensure that all future generations met a threshold equal to that of present populations. On the other hand, a sufficientarian account along the lines of Barry's argument might be said to provide a minimally humane concern for whoever comes after us.

As subsequent generations extend far into the future our common intuitions suggest that our obligations become weaker in terms of just savings and the transfer of resources hence the sufficientarian approach becomes more feasible. In the case of private individuals the motive to save for offspring is strong but weakens across many generations until we reach a stage where most are indifferent to distant posterity. But the universalist approach is able to accommodate this intuition because it transforms the debate from one of just-savings and the intergenerational transfer of wealth into a present-day concern for sustainable economic and political policies that do not worsen the position of those in the future. The idea of sustainable economic development has its own broad and complex literature but the idea of sustainability gives some content to how we might think about justice between generations. In this way the idea of justice between generations converges with the reasonable environmentalists' concern with respect for the environment as a common good. If we have obligations to future generations they are to respect the common resource that is the environment which we share. Where intergenerational justice differs from environmental ethics is that it regards the value of the environment as an aspect of human interests. This of course is not the only way we might regard it, but that is merely to recognize that even in the intergenerational case there is more to morality than justice.

See also: Children, Value of; Consequentialism (Including Utilitarianism); Deserts, Philosophical Concept of; Family Theory: Economics of Intergenerational Relations; Fertility Theory: Theory of Intergenerational Wealth Flows; Generation in Anthropology; Generations, Sociology of; Justice, Conflicts, and the Justice of Conflict Resolution; Justice: Philosophical Aspects; Property: Legal Aspects of Intergenerational Transmission; Social Justice and Social Inequalities; Utilitarian Social Thought, History of.

Bibliography

Barry, B., 1989. Theories of Justice. University of California Press, Berkeley, CA.
De-George, R., 1981. The environment, rights and future generations. In: Partridge, E. (Ed.), Responsibility to Future Generations. Promethues Books, New York, pp. 157–166.
De-Shalit, A., 1995. Why Posterity Matters. Routledge, London.
Dobson, A., 1998a. Justice and the Environment. Oxford University Press, Oxford, UK.
Dobson, A. (Ed.), 1998b. Fairness and Futurity. Oxford University Press, Oxford, UK.
Gauthier, D., 1986. Morals by Agreement. Clarendon Press, Oxford, UK.
McMahon, J., 1998. Wrongful life: paradoxes in the morality of causing people to exist. In: Coleman, J. (Ed.), Rational Commitment. Clarendon Press, Oxford, pp. 208–247.
Meyer, L.H., 2003. Past and future: the case for a threshold conception of harm. In: Meyer, et al. (Eds.), Rights, Culture and Law. Oxford University Press, Oxford, pp. 143–159.
Page, E., 1999. Intergenerational justice and climate change. Political Studies 47, 53–66.
Parfit, D., 1984. Reasons and Persons. Clarendon Press, Oxford, UK.
Rawls, J., 1971. A Theory of Justice. Oxford University Press, Oxford, UK.
Stearns, J.B., 1972. Ecology and the indefinite unborn. Monist 56, 612–625.
Woodward, J., 1986. The non-identity problem. Ethics 96, 804–831.

Intergenerational Solidarity

Valeria Bordone, Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/ÖAW, WU), International Institute for Applied Systems Analysis, Laxenburg, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Intergenerational solidarity refers to any form of mutual support exchanged between generations. Focusing on the relationships between (adult) family members, Bengtson's paradigm of intergenerational solidarity distinguishes affectual, consensual, associational, structural, functional, and normative solidarity dimensions. The concept of intergenerational ambivalence additionally accounts for the coexistence of harmony and conflict in close relationships. In an attempt to understand individuals' motivation for intergenerational solidarity, several explanations are proposed and discussed in this article. Moreover, the theoretical framework introduced by Szydlik is outlined to account for the fact that intergenerational solidarity is principally a dyadic relation of individuals embedded in a family and in a social context. A discussion of the key findings on adult intergenerational solidarity is presented to illustrate patterns of (grand) parent–child exchanges.

Intergenerational Solidarity

Intergenerational solidarity refers to any form of mutual support that is exchanged between generations. While the cohort relationship is mainly a concern of social policy (i.e., the 'generational contract'), the focus of this article is on personal relations between family members of different generations, located in a system of ranked descent.

The present-day situation where members of a family belonging to different generations live many years of overlapping lives is unprecedented: in developed countries, most children have at least one parent alive until their sixth decade of life, and the grandparent–grandchild bond may hold up to four decades. Parallel with the verticalization of the age structure in the society, within the family age structure more people are occupying the top generational positions than in the past, and children are now more likely to have more grandparents than siblings or cousins (*see* Age Structure). What makes intergenerational relationships within the family so important is the fact that each person is involved in a parent–child relationship for most of their own lives. On becoming parents, individuals are likely to be involved in two types of parent–child relationships, being the child in one case and being the parent in the other.

Over the past two decades, the outstanding challenges posed by the ageing process with regard to intergenerational relationships in the family shifted the focus of research in this field more and more toward the relationships between elderly parents and adult children. In the later *empty-nest* life, the size of personal community networks tends to decrease because of the decline in resources, including income and health. A small number of geographically proximate members and long-time relationships dominate the social network of the elderly. For example, a recent study revealed that the average network size among Europeans aged 50 and above is between two and three members (Stoeckel and Litwin, 2013). 85% of them reported having a social network comprising solely (62%) or for the majority (23%) of family members. The partner and children are usually named as primary network members. However, generally having about the same age and lifestyle, the partner may fail to provide instrumental forms of help. Intergenerational relationships become, therefore, the basic opportunity for support.

Interest in the study of intergenerational solidarity has risen with the rapid aging of the population, although it dates back to the 1950s and 1960s when scholars were concerned with the weakening vertical family solidarity because of demographic trends, geographic mobility, and socio-cultural changes – the so-called *family-in-crisis hypothesis* (e.g., Sussman and Burchinal, 1962). Family studies of the time (e.g., Burgess, 1960; Parsons and Bales, 1955) emphasized the centrality of the nuclear family and the weakening of family bonds resulting in isolation of older family members. The sociologist Popenoe (1993) argued that the family has been cut back to its essentials: two generations (i.e., parents and children) and two functions (i.e., childbearing and provision of affection and companionship).

Given that "for a growing number of Europeans, living as a family today means living in longer, thinner, more often de-institutionalized (non-marital), non-co-resident families" (Hantrais, 2005, p. 4), adult intergenerational relationships within the family are shaped qualitatively and quantitatively different from the past. Yet, longstanding investigations of solidarity (*see* Family and Kinship, History of; Welfare State, History of) show that the family may have changed its structure or even its meaning, without losing its focal role of solidarity network for the family members.

This article first provides a discussion of the mostly used intergenerational solidarity paradigm; then it lays out the framework to understand family solidarity, by noting alternative theories that have been put forth by economists, sociologists, and demographers to explain the ties between generations. In addition, a discussion of the 'facts,' that is, what we know from empirical evidence on intergenerational solidarity, is included. A discussion of three-generation intergenerational solidarity concludes the article, accounting for the role of grandparents.

Solidarity, Conflict, and Ambivalence

The wide range of research on intergenerational solidarity from several fields (sociology, demography, anthropology, gerontology, and so forth) worked as an obstacle to the development of a uniform definition. Instead, many definitions were either circular or too narrow. However, "the focus on solidarity […],

whatever definition used, […] has the merit of documenting how intergenerational relations within families also maintain an important solidaristic role in increasingly individualized and welfare-state societies" (Saraceno, 2008, p. 10). The terminology of this article coincides primarily with the six domains of intergenerational solidarity which are key components of Bengtson's solidarity framework (Bengtson, 2001; Bengtson and Roberts, 1991), additionally touching on alternative schemes.

Bengtson's comprehensive scheme for describing sentiments, attitudes, values, behaviors, and structural arrangements in parent–child relationships distinguishes between affectual, consensual, associational, structural, functional, and normative solidarity. *Affectual* solidarity includes the sentiments and evaluations family members express about their relationship with other members; *consensual* solidarity refers to the agreement in opinions, values, and orientations between generations; *associational* solidarity is the type and frequency of contact between intergenerational family members; *structural* solidarity is the 'opportunity structure' for cross-generational interaction and reflects geographic proximity between family members; *functional* solidarity reflects the giving and receiving of both instrumental assets and services across generations; *normative* solidarity takes into account the expectations regarding filial and parental obligations, as well as norms about the importance of familistic values. Solidarity, along a number of dimensions, helps families function as cohesive units to fulfill member needs.

Although Bengtson's solidarity paradigm has been defined as the 'gold standard' for assessing intergenerational relationships (Silverstein et al., 2010, p. 1007), other investigators have developed broader or more detailed categories of informal helping behaviors between generations. Weiss (1974), for example, suggests a broad conception of six provisions in social relationships: attachment, social integration, opportunity for nurturance, reassurance of worth, a sense of reliable alliance, and obtaining guidance. First, individuals must experience a feeling of comfort. Next, the network should allow pooling and sharing of ideas and information. Third, individuals must have a sense of being needed by others. Fourth, others should tell or show individuals that they are important to them. Fifth, individuals have to be assured that their physical needs will be met or be supplemented significantly by others. The sixth provision is constituted of advice that individuals receive from teachers, parents, or others in authority during their life.

House (1981) adds precision to the definition of the resources that are exchanged in supportive transactions, distinguishing between *information* and advice, *appraisal support* (i.e., information relevant to self-evaluation, feedback, or help in decision making), *emotional concern*, including understanding and esteem available from others, and *instrumental aid* with needs (i.e., getting groceries, cooking, cleaning, or managing money). Informational, appraisal, and emotional support are frequently difficult to disaggregate. Krause (1987) further defines, within the instrumental dimension, the aid received (*tangible* assistance) and the support provided (*integration*).

Despite being mainly linked by mutual support, love, and happiness, family members are also likely to experience *conflicts* (Sprey, 1969). According to classical sociologists, the coexistence of closeness and conflict is inevitable in intimate relationships: parents and children may hold different interests and have power differentials. Continuing efforts to refine the solidarity paradigm, indeed, added 'conflict' as a seventh dimension to the family solidarity paradigm, generating what has come to be known as the *solidarity–conflict paradigm* (Bengtson et al., 1990).

However, assuming that affectual solidarity could capture closeness as well as conflict (being the antitheses of each other) would imply considering social interactions as being distributed over a continuum that ranges from positive to negative (i.e., from high solidarity and low conflict to low solidarity and high conflict). Yet, the relationship between parents and children may be characterized by high solidarity and high conflict or low solidarity and low conflict, both at the same time. The positive connotation carried by the term 'solidarity' should not hide the contradictions experienced in close social relationships.

The concept of intergenerational *ambivalence*, proposed initially by Lüscher and Pillemer (1998) and then expanded by Connidis and McMullin (2002), accounts for harmony and conflict in a multidimensional way. The ambivalence construct may complement the solidarity models and the conflict model, reflecting the emotional complexities of family life. Solidarity, conflict, and ambivalence are lenses through which to look at how family members deal with attachment and independence and how they negotiate them.

Especially in later life, tensions arise between the desire for autonomy and the likely need to become dependent on the adult children – to whom older parents were formerly providers (Silverstein et al., 2010). By applying the concept of ambivalence to family ties, researchers have moved away from the idea that family members maximize positive relationships and minimize negative interactions (van Gaalen and Dykstra, 2006).

Theory of Family Solidarity

The research on adult intergenerational solidarity has often been defined as 'atheoretical.' This vulnerability may be derived from the fact that family studies are mainly problem-driven and applied disciplines. A second criticism to the social science approach to intergenerational solidarity is its increasing 'microfication,' that is, giving particular attention to the psychological characteristics of the individuals.

Understanding intergenerational ties requires a theory about individuals' motivation for behavior. Indeed, several attempts have proposed explanations for intergenerational exchange. Moreover, a general theoretical framework has been outlined by Szydlik (2008) to account for the fact that intergenerational solidarity is principally a dyadic relation of individuals, embedded in a family and, beyond that, in a social context.

Starting from the individual level, the rational choice argument of the *exchange theory* regards intergenerational solidarity as a form of support with respective costs and rewards for the giver and for the receiver (i.e., the parent and the child in any of the two roles) (e.g., Emerson, 1976) (*see* Exchange: Social). Individuals will incur the costs of giving in anticipation of the (direct or delayed) rewards of receiving. In this perspective, intergenerational solidarity is based on self-interest.

In agreement with Szydlik's model (2008), the diverse conditional factors for solidarity can be classified into four groups: opportunity, need, family, and cultural–contextual structures. Costs and benefits of intergenerational solidarity are reflected in *opportunity* and *need structures* (see also Bordone, 2012 for a discussion). The first includes resources for solidarity, enabling or preventing it. For example, high levels of a child's educational attainment increase personal aspirations in the labor market (even at the cost of greater geographical distances from the parents). Education is, therefore, an opportunity that tends to prevent parent–child structural solidarity. Need structures stem for health problems or for emotional support as well as for necessity of financial nature. Solidarity is indeed *contingent* on the needs of both parents and children (Deutsch, 1975): for example, marital disruption suffered by the adult child might lead to an increase in the contact with the parent because of the child's need to get emotional comfort.

The *altruism theory* suggests that beneficence may also guide parent–child solidarity (Becker, 1974): parents and children care about each other and, therefore, provide each other various forms of support.

Within these broad perspectives there are several related models which help explain the complicated relationships observed among family members. It is plausible to assume that parents and children care about one an other and, therefore, make transfers as required to maintain or enhance the other's well-being, in concordance with altruism, exchange, and contingency hypotheses. Repeated exchanges within the family may derive from a compromise between emotional aspects and from what rationality dictates. However, the relation between parent and child also depends on *family structures*: family size, family composition, events that happened during the family life course as well as family roles and norms that reflect the entire history of socialization of the parent and the child. For example, the existence of more siblings (especially daughters) may lower the likelihood of providing functional solidarity to the parents.

At the micro-level, theoretical explanations have also been proposed concerning the tension between family members, mainly reflecting the competing pressures to maintain attachment and at the same time establish independence. Fingerman (2001), for example, refers to the competing needs of different generations as *developmental schisms*: while older parents may ask for more closeness, adult children struggle between multiple roles as partners, parents, children, and workers. Ambivalence in the relationship tends to emerge especially when the available resources do not offer many 'escape options,' that is, women, elderly, and poor will encounter more ambivalence than their privileged counterparts (for a detailed review, see Cooney and Dykstra, 2013).

Although the ideational approach widely known under the umbrella of the 'second demographic transition' (*see* Second Demographic Transition) emphasizes individual preferences in the decision-making process of individuals, intergenerational solidarity is also a feature of *cultural–contextual structures*, with roots laying far back in history. Moving upstream beyond the family and returning to a more Durkheimian orientation to social context, cultural–contextual structures should be taken into account by considering socioeconomic conditions, tax system, welfare system, labor and housing markets as well as social norms, legal duties, and policy arrangements, within which intergenerational solidarity develops. As such, they shape and reinforce intergenerational solidarity within a specific context, providing also a definition of dependence between generations.

Saraceno (2010), building on previous conceptual works, distinguishes three patterns in which legal norms and public provisions frame financial and caring obligations in families, both up and down family lines. *Familialism by default* exists when few publicly provided alternatives to family care and financial support are available. *Supported familialism* characterizes contexts where policies help families in maintaining their caring and financial responsibilities, usually through financial transfers. *De-familialization* applies when the individualization of social rights (e.g., minimum income provision, unemployment benefits, and care entitlement) limits family responsibilities and dependencies. This categorization, going beyond the public/private dichotomy of responsibilities, emphasizes that public family provisions may both motivate and lighten private family responsibilities.

The literature mainly refers to the crowding-in and crowding-out hypotheses when dealing with the offer of services from the family and the state. The state may indeed displace family services as in the case of family pensions taking over the provision of security in old age (*crowding-out*) or stimulate family support by relieving family members of time-consuming support and incentivizing them to provide voluntary services (*crowding-in*). State and family services may thus work in *complementarity*, dividing support activities according to the task-specific model (see, e.g., Motel-Klingebiel et al., 2005 for a discussion of these theses and an empirical testing).

Empirical Research: Key Findings

The intergenerational solidarity paradigm has been validated with data from many countries around the world. Although the discipline-specific language and methodology make it difficult to draw broad conclusions, empirical patterns found in previous work are discussed here with a focus on Bengtson's intergenerational solidarity framework.

Affectual and Consensual Solidarity

Emotional closeness and consensus of opinions between parents and adult offspring reflect affinity between the generations. The affective bond is central to the parent–child relationships in order to engage in solidarity exchange. Emotional ties with offspring may be of extremely high value for aging parents facing the loss of significant social ties (e.g., spouse and siblings) due to death, and the vanishing of social roles, for example, due to retirement. Indeed, children are identified by most aged parents as close confidants (Connidis and Davies, 1992). This is especially the case for mothers, who are less likely than fathers to name their respective partner as their preferred confidant. Although parent–child relationships tend to be judged positively, also from the perspective of adult offspring, they are reported of higher quality with respect to mothers than to fathers (Silverstein et al., 2010; van Gaalen and Dykstra, 2006). Moreover, adult children report less closeness

to the parents, compared to their parents' ratings of their relationship (Fingerman, 2001).

In the 1960s and 1970s, research began to give much more attention to the 'generation gap' in attitudes – i.e., the lack of consensus in values and opinions between generations – as fostered by the socio-political changes of the time. By definition, consensus is a relational variable for which an individual can provide information about the extent to which he or she perceives agreement or disagreement with another individual. Although this clearly implies that reports from a single family member are inadequate, most studies concerned with the similarity of views between generations use 'generational-level' analysis. Accordingly, self-reports from one generation have been compared with those from another. Alternatively, maintaining the family as the unit of analysis, measures of agreement provide answers to the question regarding the extent to which reports from parents and their children are similar or different.

Pruchno et al. (1994) showed that consensus between mothers and children, and between fathers and children may be best represented as two constructs (mother–child consensus and father–child consensus) rather than as one construct (i.e., parent–child consensus). Moreover, contrasting reports from mothers or fathers and their children indicate significant differences in the reports: mothers reported higher levels of agreement than did their children on how money is spent, religious matters, friends, making major decisions, and dating, while fathers reported higher levels of agreement than did their children on religious matters and dating. Children reported higher levels of agreement than did their parents on aims and goals, household tasks, career decisions, and education.

Associational and Structural Solidarity

Frequency of contact and geographic proximity between the parent and the child reflect the opportunity structure of the relationship. The associational dimension of solidarity is highly correlated with other forms of solidarity: parents and children who are more often in contact are more aware of each other's needs than those with lower frequency of contact and react to them by providing affectual solidarity and instrumental help. In particular, frequent in-person visits strongly predict functional solidarity exchange (van Gaalen and Dykstra, 2006). Both studies on American and on European data reported high levels of contact between parents and children: the majority of older parents have weekly contact with at least one child (e.g., Hank, 2007; Lawton et al., 1994). The observed decline in weekly visits to the parents in some countries may be justified by a possible counterbalance between individualistic aspirations and other compositional changes. Consistently, literature has found that parent–child contact is higher when children have own children and lower when parents are divorced (especially with the father). Greater geographical distance also predicts lower frequency of contact, not only face-to-face. The use of new technologies that facilitate communication (e.g., Internet) may increase the level of parent–child contact even when living geographically far from each other. Cross-national variation still exists within Europe, with Mediterranean countries reporting higher levels of contact than Northern European countries (e.g., Bordone, 2009; Hank, 2007).

The dimension of distance, in addition to frequency of contact, is important because close relatives are necessary to provide some other forms of solidarity. Geographical distance, however, may not be exogenous to intergenerational associational and functional solidarity: family members might choose to move closer to one another (or together) because support is required by one or both generations. Coresidence can facilitate intergenerational solidarity in numerous dimensions (e.g., emotional support and instrumental help) and can be an alternative for financial transfers both downward and upward. The person in need is usually most likely to move, and this is also why having a spouse reduces the likelihood that elderly parents reside with one of their children; while a child's divorce is a driving factor for moving in with the parents (see Cooney and Dykstra, 2013 for a more detailed discussion of related empirical evidence). Not only coresidence, but also geographical proximity may benefit both parties: living near each other allows for intergenerational solidarity as well as independence. The literature shows that fertility decisions are positively affected by having local parents, especially in countries where public childcare is limited. Moreover, having parents nearby who can provide grandparental childcare may help women balance work and family duties.

Functional Solidarity

Upward and downward exchange in terms of financial and time transfers, including all kinds of care-giving activities and instrumental support, are interrelated dimensions: assistance provided through one form of solidarity may either substitute or complement some other type. A particular form of downward time assistance is grandparenting, which we discuss subsequently. The type, direction, and amount of functional solidarity exchanged depend on three aspects, in addition to geographical distance. First, the needs and opportunities of both givers and receivers: for example, higher-income family members may substitute time-related transfers with financial transfers; second, welfare regimes: in countries where pension levels are high, the oldest old are less likely to receive money from their children and more likely to help them financially; third, the position of individuals in the life course: the oldest old tend to be net recipients and the pivot generation is more likely to engage in giving (Attias-Donfut et al., 2005).

In general, cash gifts tend to flow to the younger generations, even after children are grown and are living on their own. Using data from the Survey of Health, Ageing and Retirement in Europe (SHARE website http://www.share-project.org/), Albertini et al. (2007) reported that 21% of older parents noted providing monetary help to an adult child in the last year, with only 3% receiving such help. As shown by Attias-Donfut et al. (2005), rates of making financial transfers are higher for parents in a couple than for singles. In contrast to inter vivos transfers, often targeted to the more needy children or grandchildren, bequests are mostly divided equally among children.

Parents are also much more likely to provide time assistance (over a wide spectrum of tasks) to their children (especially to divorced or never married daughters with children) than to receive it, at least until late in life when their growing needs for help change this balance. However, functional solidarity,

understood in its informal care component (i.e., personal care, practical household help, and help with paperwork), is usually reported upward, although in modest levels in most families. A substantial minority of older American parents typically have received some routine assistance from their adult children (e.g., Davey and Eggebeen, 1998). In Europe, 16% of elders report to be recipients of instrumental support from their offspring (Albertini et al., 2007). Bonsang (2009) showed that, on average, the 65 year olds and older with at least one child in Europe receive 5.6 h of informal care per month from children not living with them. However, evidence consistently points to a significantly higher likelihood of older parents getting help from their children when needs arise.

Country differences exist in the rates and amounts of transfers; however, empirical studies in the field did not easily identify generalizable results to group countries together based on financial and time-assistance patterns (Attias-Donfut et al., 2005). The consequences of these private transfers are important for both individuals and social policies concerning the redistribution of wealth in the society (Attias-Donfut et al., 2005).

Normative Solidarity

The belief about family members' responsibilities to care for and support one another is a core aspect in theories about the family because it largely helps explaining associational and functional solidarity between parents and children. Research linking family obligation to solidarity behavior has produced inconsistent findings, often because norms for intergenerational solidarity are inadequately conceptualized and measured. Several cross-sectional investigations find that providers and receivers of support tend to report higher filial responsibility compared to those who are not involved in intergenerational exchanges. Other research finds no association between the endorsement of filial responsibility and actual support (see Cooney and Dykstra, 2013 for a deeper review of these studies).

Evidence from longitudinal studies on the effect of normative solidarity on support behavior, although mixed, tends mostly to suggest that filial obligations have a stronger motivational component for sons than for daughters. As Cooney and Dykstra (2013, p. 360) suggest: "females appear less sensitive to social prescriptions, perhaps because they take support provision for granted." Although strong feelings of family responsibility may predispose individuals to be supportive, whether assistance actually materializes depends on the specific context of need.

Inconsistencies across countries have also been found with norms being more powerful determinants of behaviors in countries with limited public support systems (see Cooney and Dykstra, 2013 for a detailed review).

Compared to the evidence on filial obligations to aging parents, less is known about how people view the responsibilities of parents in middle and late life toward adult children as studies on parental responsibility usually refer to parents' responsibility for young children. The Survey of Health, Ageing and Retirement in Europe (see SHARE website http://www.share-project.org/ for a detailed list of references), though, asks about the opinion on whether grandparents' duty is to help grandchildren's parents in looking after grandchildren or whether grandparents' duty is to contribute toward the economic security of grandchildren and their families. Descriptive evidence on these items shows that Mediterranean countries appear to be characterized by stronger family values than Central and Northern European countries.

Mutual support between generations is not always voluntary, but it is sometimes based on normative pressure or legal duties. An example is economic assistance, legally regulated, by parents to children until a certain age and later on in life by adult children to their parents (for further data on this issue, refer to the social policy indicators available on the Multilinks project website http://www.multilinks-project.eu/). National and comparative survey data, as well as qualitative studies on intergenerational support, show that the higher the space of familialism by default, and to some degree also of supported familialism, the higher the impact of gender and social class differences among the potential family carers (see Saraceno, 2010 for a review).

Ambivalence

Individuals tend to identify some problems with their adult children and with their parents, often related to communication and lifestyle. However, the majority of adult children define the ties to their parents as solely close, one-third define them as ambivalent, and only about 6% experience the relationship with the parents as solely problematic (Fingerman et al., 2004). The coexistence of high solidarity and high conflict is not a general state, and it is mainly found among young adults (van Gaalen and Dykstra, 2006).

Evidence (see Cooney and Dykstra, 2013) consistently shows that levels of ambivalence not only vary with geographical proximity (i.e., closer parent–child proximity is associated with more mixed feelings in their relationship), but also depend on the gender and marital status of the parent (ambivalent relationships are reported more with fathers than with mothers, and with currently unmarried parents compared to partnered). Empirical results also support the idea that ambivalence may result from issues of balancing attachment and independence. Parents' health is a significant predictor of ambivalence in reporting from both parents (Pillemer et al., 2007) and children (van Gaalen and Dykstra, 2006). Parents report ambivalence in their parent–child relationship also when their adult children have health problems and, especially mothers, when they have issues with the law or have chemical dependency (Pillemer et al., 2007).

Literature also shows that parents and children usually express their feelings openly. In turn, unresolved parent–child tensions may detract from affectual solidarity.

Which generation is reporting their feelings may also matter: despite providing more assistance to their offspring than receiving, mothers tend to report lower relational ambivalence than adult children. Different relational expectations may explain this evidence.

The Macro-Level

The availability of cross-country comparable surveys investigating intergenerational solidarity in its various dimensions has incentivized the proliferation of studies examining

determinants of intergenerational solidarity variation among different individuals as well as among a larger number of countries.

The first type of studies, where the micro (individual) level was the focus of the analysis, accounted for context by including country indicators to the regression models. Several comparative studies showed that both needs and opportunities of the recipients and givers exert a comparable influence on the exchange of support in different countries. However, the diversity of intergenerational relationships from country to country remains significant within Europe, highlighting a North–South divide. Therefore, a second type of studies gave attention to cultural–contextual and policy explanations at the macro-level in multilevel analysis. Although the most common discussion at the macro-level is on the distinction between weak versus strong family ties, likely linked to specific welfare systems, the nonstrict correspondence required further investigations. Political economy focused studies took into account the social context of elderly people centering around help and care from a welfare production point of view (e.g., Brandt et al., 2009). Some explorative works on the societal level have shed light on the importance of social norms in relation to family solidarity (e.g., Bordone, 2012).

Theoretical and empirical works on family solidarity tend to assume that, as the family is a reliable and nonjudgmental institution that crosses generations, it has both a direct and a mediated effect on health conditions through support between its members. Usually this strand of research shared the assumption that family solidarity (positively) influences physical and mental health status (Berkman et al., 2000). Several conceptual causal frameworks have been proposed to highlight the upstream and downstream factors that are likely to explain a causal link between solidarity and health outcomes, but this is another (although related) story, which is not discussed here.

A Three-Generation Perspective: The Role of Grandparents

The new opportunity role of grandparents has enlarged studies of intergenerational solidarity to three family generations: grandparents, parents, and (grand)children. Although in most developed countries there has been a decline in multigenerational households, grandparents still play an active and supportive role within the family, especially by taking care of grandchildren. Indeed, the proportion of elderly giving functional solidarity to offspring increases significantly when childcare is included in support measures. Taking care of grandchildren is a common activity among grandparents in Western societies, especially among women: in the USA, 50% of grandparents provide some type of childcare; similar percentages of grandfathers in Europe say they provide regular or occasional care to their grandchildren and even more grandmothers (58%) are involved in childcare (see Glaser et al., 2010 for a review).

As grandparental childcare is traditionally considered an altruistic act, research on this intergenerational exchange has focused on its effects on younger generations. By helping their children (especially daughters) with looking after grandchildren, the (unpaid) grandparents are also likely to produce welfare benefits, especially when the services offered by the market are costly and public provision of care is scarce. In some countries, the important role of grandparents as providers of childcare has already been officially recognized: in the UK, for example, grandparents who give up paid work to provide childcare can claim credits that allow them to qualify for a basic state pension (Glaser et al., 2010).

Grandparental childcare can be seen as both a downward and an upward transfer: on the one side, grandparents invest time and resources in their grandchildren; while on the other side, grandchildren are an important (emotional) resource for grandparents. Economic models have indeed hypothesized that inter vivos transfers from grandparents to grandchildren might also have an egoistic component that can help to explain why older generations transfer resources to younger generations. Downward monetary and time transfers may be motivated by the anticipation of future need, and the hope that the younger generations will be more altruistic toward their parents or grandparents in return. According to this strand of literature, grandparental childcare could be considered an investment that is expected to pay off in the future.

Focus has also been laid to consensual solidarity between grandparents and grandchildren. Giarrusso et al. (2001) found evidence for a strong relationship between grandparents and their adult grandchildren: "longer life expectancy appears to result in long-term loving relationships" (p. 473).

In sum, adult intergenerational relationships between (grand)parents and children remain a relevant aspect of aging and family studies. Reductions in mortality and decreases in fertility in most parts of the world will also reshape intergenerational solidarity patterns in economies currently in transition and in developing countries, affecting large numbers of the global population. Increasing diversity in family forms and norms will also continue to impact on the variability in family bonds, raising new research questions and requiring new theoretical explanations.

See also: Adulthood: Dependency and Autonomy; Dependency Theory; Exchange: Social; Family and Kinship, History of; Life Course: Sociological Aspects; Linked Lives; Retirement and Encore Adulthood: The New Later Life Course; Second Demographic Transition; Welfare State, History of.

Bibliography

Albertini, M., Kohli, M., Vogel, C., 2007. Intergenerational transfers of time and money in European families: common patterns – different regimes? Journal of European Social Policy 17, 319–334.

Attias-Donfut, D., Ogg, J., Wolff, F., 2005. European patterns of intergenerational financial and time transfers. European Journal of Ageing 2, 161–173.

Becker, G.S., 1974. A theory of social interactions. Journal of Political Economy 82, 1063–1093.

Bengtson, V.L., 2001. Beyond the nuclear family: the increasing importance of multigenerational bonds. Journal of Marriage and Family 63, 1–16.

Bengtson, V.L., Roberts, R.E.L., 1991. Intergenerational solidarity in aging families: an example of formal theory construction. Journal of Marriage and the Family 53, 856–870.

Bengtson, V.L., Rosenthal, C.J., Burton, L.M., 1990. Families and aging: diversity and heterogeneity. In: Binstock, R.H., George, L.K. (Eds.), Handbook of Aging and the Social Sciences, third ed. Academic Press, San Diego, CA, pp. 263–287.

Berkman, L.F., Glass, T., Brisette, I., Seeman, T.E., 2000. From social integration to health: Durkheim in the new millennium. Social Science and Medicine 51, 843–857.

Bonsang, E., 2009. Does informal care from children to their elderly parents substitute for formal care in Europe? Journal of Health Economics 28, 143–154.

Bordone, V., 2009. Contact and proximity of older people to their adult children: a comparison between Italy and Sweden. Population, Space and Place 15, 359–380.

Bordone, V., 2012. Social norms and intergenerational relationships. In: De Santis, G. (Ed.), The Family, the Market or the State? Intergenerational Support under Pressure in Ageing Societies. Springer, Netherlands, Dordrecht, pp. 159–178.

Brandt, M., Haberkern, K., Szydlik, M., 2009. Intergenerational help and care in Europe. European Sociological Review 25, 585–601.

Burgess, E.W., 1960. Aging in western culture. In: Burgess, E.W. (Ed.), Aging in Western Societies. University of Chicago Press, Chicago, pp. 3–28.

Connidis, I.A., Davies, L., 1992. Confidants and companions: choices in later life. Journal of Gerontology 47, S115–S122.

Connidis, I.A., McMullin, J.A., 2002. Sociological ambivalence and family ties: a critical perspective. Journal of Marriage and Family 64, 558–567.

Cooney, T.M., Dykstra, P.A., 2013. Theories and their empirical support in the study of intergenerational family relationships in adulthood. In: Fine, M.A., Fincham, F.D. (Eds.), Handbook of Family Theories: A Content-Based Approach. Routledge/Taylor and Francis, New York, pp. 356–378.

Davey, A., Eggebeen, D., 1998. Patterns of intergenerational exchange and mental health. Journal of Gerontology: Psychological Sciences 53B, P86–P95.

Deutsch, M., 1975. Equity, equality, and need: what determines which value will be used as the basis of distributive justice. Journal of Social Issues 31, 137–149.

Emerson, R., 1976. Social exchange theory. Annual Review of Sociology 2, 335–362.

Fingerman, K.L., 2001. Aging Mothers and Their Adult Daughters: A Study in Mixed Emotions. Springer, New York.

Fingerman, K.L., Hay, E.L., Birditt, K.S., 2004. The best of ties, the worst of ties: close, problematic, and ambivalent social relationships. Journal of Marriage and Family 66, 792–808.

Giarrusso, R., Feng, D., Silverstein, M., Bengtson, V.L., 2001. Grandparent-adult grandchild affection and consensus. Journal of Family Issues 22, 456–477.

Glaser, K., Ribé Montserrat, E., Waginger, U., Price, D., Stuchbury, R., Tinker, A., 2010. Grandparenting in Europe. Grandparents Plus, London.

van Gaalen, R.I., Dykstra, P.A., 2006. Solidarity and conflict between adult children and parents: a latent class analysis. Journal of Marriage and Family 68, 947–960.

Hank, K., 2007. Proximity and contacts between older parents and their adult children: a European comparison. Journal of Marriage and Family 69, 157–173.

Hantrais, L., 2005. Living as a family in Europe. In: Hantrais, L., Philipov, D., Billari, F.C. (Eds.), Policy Implications of Changing Family Formation. Council of Europe Publishing, Strasburg, pp. 117–181.

House, J.S., 1981. Work, Stress and Social Support. Addison Wesley, Reading, MA.

Krause, N., 1987. Life stress, social support, and self-esteem in an elderly population. Psychology and Aging 2, 349–356.

Lawton, L., Silverstein, M., Bengtson, V.L., 1994. Affection, social contact and geographic distance between adult children and their parents. Journal of Marriage and the Family 56, 57–68.

Lüscher, K., Pillemer, K., 1998. Intergenerational ambivalence: a new approach to the study of parentchild relations in later life. Journal of Marriage and the Family 60, 413–425.

Motel-Klingebiel, A., Tesch-Römer, C., von Kondratowitz, H.-J., 2005. Welfare states do not crowd out the family: evidence for mixed responsibility from comparative analyses. Ageing and Society 25, 863–882.

Parsons, T., Bales, R., 1955. Family Socialization and Interaction Process. The Free Press, New York.

Pillemer, K., Suitor, J.J., Mock, S.E., Sabir, M., Pardo, T.B., Sechrist, J., 2007. Capturing the complexity of intergenerational relations: exploring ambivalence within later-life families. Journal of Social Issues 63, 775–791.

Popenoe, D., 1993. American family decline, 1960–1990: a review and appraisal. Journal of Marriage and the Family 55, 527–542.

Pruchno, R., Burant, C., Peters, N.D., 1994. Family mental health: marital and parent–child consensus as predictors. Journal of Marriage and the Family 56, 747–758.

Saraceno, C., 2008. Families, Ageing and Social Policy. Edward Elgar, Cheltenham, UK.

Saraceno, C., 2010. Social inequalities in facing old-age dependency: a bi-generational perspective. Journal of European Social Policy 20, 32–44.

Silverstein, M., Gans, D., Lowenstein, A., Giarrusso, R., Bengtson, V.L., 2010. Old parent – child relationships in six developed nations: comparisons at the intersection of affection and conflict. Journal of Marriage and Family 72, 1006–1021.

Sprey, J., 1969. The family as a system in conflict. Journal of Marriage and the Family 31, 699–706.

Stoeckel, K.J., Litwin, H., 2013. Personal social networks in Europe: do people from different countries have different interpersonal solidarities? In: Börsch-Supan, A., et al. (Eds.), Active Ageing and Solidarity between Generations in Europe. de Gruyter, Berlin, pp. 277–287.

Sussman, M.B., Burchinal, L., 1962. Kin-family networks. Marriage and Family Living 24, 231–240.

Szydlik, M., 2008. Intergenerational solidarity and conflict. Journal of Comparative Family Studies 39, 97–114.

Weiss, R.S., 1974. The provisions of social relationships. In: Rubin, Z. (Ed.), Doing unto Others: Joining, Molding, Conforming, Helping, Loving. Prentice-Hall Spectrum Books, Englewood Cliffs, N.J, pp. 17–26.

Intergroup Relations

Jorge Vala and Rui Costa-Lopes, Institute of Social Sciences – University of Lisbon, Lisboa, Portugal

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides a succinct description of the most significant socio-psychological research and theoretical breakthroughs in the area of intergroup relations. The first section focuses on contemporary socio-cognitive theories that highlight the role played by social categorization in intergroup relations since these theories and their links with motivation and emotion massively inspire current research in this domain. The subsequent section focuses on the perspectives on intergroup relations that have emerged in the past two decades and that have reinvigorated the study of collective beliefs, ideologies, and social norms as important organizers of group behavior, specifically through the legitimation of biased intergroup attitudes.

From a socio-psychological perspective, intergroup relations are "the systematic study of relations between individuals as they are affected by group membership" (Stephan, 1985, p. 599).

The social psychology of intergroup relations has been structured in accordance with three main problems: (1) what kind of factors are responsible for the fact that our collective and individual behavior is largely determined by our membership of and identification with social groups; (2) why do intergroup relations involve a significant degree of negative aspects such as discrimination, negative attitudes, and prejudice; and (3) how can these negative aspects be reduced and intergroup relations improved? This article mainly analyzes the first two issues without overlooking the third issue.

In a historical and comprehensive piece of analysis, John Duckitt (1992) offers a highly stimulating taxonomy of the theoretical factors underlying racial prejudice. This taxonomy can also help us to understand the various socio-psychological approaches to intergroup relations and their relation to historical milestones. Accordingly, we provide a succinct description of the main phases of research on intergroup relations, starting with the period following one of the important of such milestones: the Second World War. Indeed, it was after this major conflict that research in this domain flourished, stimulated by the need to understand Nazi racial ideology and the Holocaust. During the Second World War and for a decade after it ended, negative intergroup relations were mainly conceived as an expression of a pathological need and theoretical and empirical approaches to the subject stressed personality factors and individual differences. The prototypal and most influential study carried out within this framework was Theodor Adorno and his collaborators' analysis of *Authoritarian Personality* syndrome in the 1950s.

Still in the 1950s, the focus moved from personality to social norms and sociocultural and socio-structural factors. Two works can be singled out as good examples of the foundations of this analytical perspective. First, the famous 'Robbers' Cave Experiments' carried out by Muzafer Sherif and his colleagues. These experiments highlighted the structural role played by social norms, group interests, and functional relations between groups – namely competition versus cooperation – with regard to the quality of intergroup relations. A second good example of the perspectives that emerged in this period is Tom Pettigrew's (1959) research on the relevance of socio-institutional factors to the quality of intergroup relations. Pettigrew empirically showed that, despite the importance of personality factors, sociocultural and institutional dynamics were crucial in explaining the differences between anti-black attitudes in the Southern US versus the North and between the United States and South Africa.

A new analytical shift in the analysis of intergroup relations occurred in the 1970s with the flourishing of the cognitive perspective in all domains of social psychology. Following the seminal hypotheses proposed by Gordon Allport (in his classical work *The Nature of Prejudice*) concerning the role of social categorization in the dynamics of intergroup relations, two influential works emerged. First, Doise (1976) proposed the categorical-differentiation model, which suggested that the differentiation between in-groups and out-groups generates corresponding differentiations at the evaluative and behavioral levels. Secondly, Tajfel and Turner (1979) developed Social Identity Theory (SIT), in which social comparison and social categorization constitute anchoring processes despite the fact that motivation (to achieve a positive social identity) is the core psychological mechanism behind this theory. SIT continues to inform the debate on intergroup relations based on the assumption that social categorization per se (the in-group/out-group categorization) is a sufficient condition to create in-group bias, that is, to glorify in-groups relatively to out-groups.

Besides the contribution made by the socio-cognitive approach, since the 1980s, several theories approached intergroup relations through the analysis of social norms, collective beliefs, ideologies, and implicit theories about human beings (Crandall and Eshleman, 2003; Jost and Banaji, 1994; Katz and Hass, 1988; Leyens et al., 2000; Sidanius and Pratto, 1999).

Although this encyclopedia article is framed by the taxonomic and historical perspective presented above, it addresses only the last two approaches. Thus, the first section of this article focuses on contemporary socio-cognitive theories that highlight the role played by social categorization in intergroup relations since these theories and their links to motivation and emotion massively inspire current research in this domain. The second section of this article is devoted to perspectives on intergroup relations that have emerged in the last two decades and that have reinvigorated the study of ideologies and social norms as important organizers of group behavior, specifically through the legitimation of biased intergroup attitudes. These more recent perspectives provide systematic approaches about

the content of the formal processes identified in earlier socio-cognitive approaches.

Social Categorization and Intergroup Relations

William G. Sumner carried out pioneering research into intergroup relations, introducing terms such as we-group (later transformed into in-group – the group to which we belong) and others-group (later transformed into out-group – the group to which we do not belong) and identifying ethnocentrism as a worldview in which people use the in-group as the reference by which to scale and assess the value of things. Sumner was also the first author to advance the idea that affiliation to the in-group was associated with an ongoing conflict toward the out-group. (This view was later challenged by Brewer (1999), who posited the independent existence of in-group love and out-group hate.) Several decades later, data from field studies (LeVine and Campbell, 1972) provided evidence of this tendency to differentiate social environments in accordance with in-group-out-group distinctions and to value in-group characteristics over those of relevant out-groups.

Around the same period, theoretical frameworks were being developed in the social psychology of intergroup relations in order to integrate the impact of social categorization following the paradigmatic cognitive shift mentioned above. Doise (1976) proposed the model of categorical differentiation. This model is based on the structural role played by social categorization in intergroup relations and proposes that categorical differentiation between in-groups and out-groups generates corresponding differentiations at the evaluative and behavioral levels. Simultaneously and more significantly, a work that is arguably one of the most influential theories of intergroup behavior – SIT – was being developed in response to intriguing data emerging from earlier laboratory research. In the classical minimal group studies, Tajfel and colleagues had conducted empirical research that sought to determine the minimal conditions for the emergence of in-group bias. They challenged the idea – derived from the 'Robbers' Cave Experiments' – that a conflict of group interests is a precondition for the emergence of discrimination and bias between two groups. Instead, Tajfel and colleagues believed that competition was a sufficient but not necessary condition for the emergence of such phenomena. Thus, a minimal group paradigm was developed in which the participants used matrices to allocate resources without interacting with fellow group-members and without considering personal gain. The results clearly showed that, even though a general effort to act fairly was made, the participants revealed a significant tendency to award more points to in-group than to out-group members, providing the first empirical support for the idea that the mere existence of a distinction between an in-group and an out-group could, within a specific set of circumstances, constitute the minimal condition for the emergence of in-group bias. These results were largely replicated in several independent studies (but see Jaap Rabbie and colleagues' stance on this matter).

To explain these results, Tajfel and Turner (1979) developed SIT. Unifying structural, motivational, and cognitive dynamics, this theory postulates that in-group bias (which may be elicited by categorization but is contingent on structural factors such as groups' relative status, perceptions of status legitimacy and stability, and group permeability) is a motivational response to attain a positive self-image through social comparisons (see Vignoles et al., 2000). As such, in-group bias emerges to make the in-group positively distinctive from the out-group in ways that are relevant to (and valued by) the in-group. However, according to this theory, in-group bias is just one of a set of possible strategies that an individual may use to achieve positive distinctiveness. SIT was criticized for several reasons, such as its failure to acknowledge that self-esteem was not consistently associated with in-group bias or that in-group bias does not occur for negative aspects. However, not only have most of these criticisms been debunked as misunderstandings, this theory has also influenced a highly significant amount of social psychological research on intergroup relations. After the death of H. Tajfel in 1982, Turner and colleagues sought to elaborate on the cognitive elements of SIT, leading to the creation of *Self-Categorization Theory* (SCT) (Turner et al., 1987). While SIT asserted the importance of the categorization process by invoking the value for the individual of achieving a positive distinctiveness, SCT focused on the perceptual and cognitive determinants of categorization and elaborated on the consequences of the individual perceiving him/herself as a group member for both intragroup processes and intergroup behavior.

Several other phenomena that build on the process of social categorization have been identified within social psychology and provide additional and interesting insights into the functioning of intergroup relations. One important example that combines SIT and SCT to approach both intra- and intergroup dynamics is the *black sheep effect*. – that is, when deviant in-group members are judged more negatively than comparable out-group members (Marques and Páez, 1994). The *out-group homogeneity effect* also provides an important extension of the consequences of social categorization. It consists of a tendency to see out-group members as more similar to each other than to in-group members, even though this effect is more pronounced for dominant than for dominated groups (e.g., Lorenzi-Cioldi, 1993). It is significant that this perception has been found to be associated with more negative intergroup perceptions in several studies. Another process that builds on a cognitive mechanism that is related to social categorization and is also associated with negative intergroup perceptions (namely, stereotypes) is *illusory correlation*, which can include the perception of a relationship between a particular category of membership and a particular feature when no such relationship exists (Hamilton and Sherman, 1989).

The research mentioned above consolidated the idea that categorization is the process that should be addressed in order to resolve intergroup conflict or to promote harmonious intergroup relations (see Park and Judd, 2005; for an interesting discussion on the subject). As a result, research on intergroup relations that centered on the attenuation of intergroup conflict, which began in the 1950s with the *Contact Hypothesis* with Gordon Allport (see Pettigrew and Tropp, 2006), became more focused in the 1980s on the development of theoretical models that operate (real or perceptual) transformations in group categorization.

Decategorization

Brewer and Miller (1984) developed the Decategorization Model, which built upon Allport's (1954) Contact Hypothesis. Decategorization focuses on one of Allport's conditions for successful intergroup contact: in order for contact to lead to harmonious intergroup relations, the contact interaction cannot be 'category based.' According to this model, if people do not use social categories as a mental framework by which to judge others, they will be able to recognize the uniqueness and positive characteristics of each person – in this case, the out-group member. In fact, these 'personalized' interactions are assumed to result in the disconfirmation of negative stereotypes about the out-group. A major critique of this approach relates to the supposed difficulty of making positive generalizations about other members of the same out-group precisely because individuals are not thinking within this intergroup framework.

Recategorization

In 1993, Gaertner, Dovidio and colleagues developed a different strategy within the Common In-group Identity Model (CIIM). Like Brewer and Miller (1984), the authors of this model also considered it important to reduce the salience of in-group–out-group category distinctions, albeit by using a different process. According to this model, this reduction could be achieved by transforming the cognitive representation involving the two groups (recategorization). In this case, positive intergroup attitudes would follow from the perception of a superordinate category that includes both groups (see Guerra et al., 2010). However, this strategy runs into the same problem as that associated with the decategorization strategy: ignoring the in-group–out-group categorization results in diminished generalization. Another problem with this strategy is that it seeks to eliminate group memberships that might be significant and important to individuals and which they might prefer not to abandon. One final problem with this strategy concerns the fact that the superordinate category that results from this recategorization creates an opportunity for the dominant group to assert its superiority by claiming a higher degree of prototipicality in relation to that category (Mummendey and Wenzel, 1999).

Mutual Intergroup Differentiation

While the previous two models share the assumption about the need to reduce category differentiations, the strategy proposed by Hewstone and Brown (1986) within the Mutual Intergroup Differentiation Model (MIDM) takes a different approach. Hewstone and Brown suggest that cooperative interactions between two groups should be promoted without degrading in-group–out-group categorizations. Indeed, influenced by the idea of the (positive) effects of positive goal interdependence (cooperation) depicted within Realistic Conflict Theory, these authors suggest that harmonious intergroup relations derive from cooperative interactions in which the in-group and out-group have different but complementary tasks, competencies or resources (see also Deschamps and Brown, 1983).

Dual Identity

More recently, Gaertner and Dovidio (2000) developed a refined version of the CIIM by presenting a different strategy that combines the central tenets of CIIM and MIDM, invoking an inclusive category that preserves the subgroup identities. The authors assume that this dual representation is possible since individuals can conceive of two groups (e.g., Portuguese vs Dutch) as distinct entities within a more inclusive category (i.e., European). Although this strategy partially responds to some of the problems identified with the recategorization strategy, in this approach groups that are not included in this dual categorization are targeted with even stronger negative attitudes.

A recurrent critique of the categorization approach to intergroup relations concerns the way that it neglects the affective component of intergroup attitudes. As Park and Judd (2005) put it, the intention was probably to 'add it back in' after the cognitive dynamics had been unraveled. However, aside from some notable exceptions (Mackie et al., 2000), this cognitive approach to intergroup relations continues to be unable to bring affective aspects back into the picture. If one critical aspect of the role of categorization has been its failure to establish links with emotional aspects, another issue that has appeared more recently is related to ideologies of intergroup relations. A very recent exception in this regard is the work carried out by Costa-Lopes et al. (2014), which focused on the relationship between categorization salience and bias. In this case, the discussion is not so much about whether merely dividing individuals into two groups creates in-group bias, but whether drawing attention to such a distinction per se affects intergroup attitudes. In fact, this has become the new focus of research within the general domain of categorization and intergroup relations. Following Park and Judd's (2005) idea that our understanding of this relationship could be enhanced by considering the role played by ideologies such as multiculturalism, Costa-Lopes and colleagues developed research that shows that categorization salience only leads to in-group bias when a multicultural ideology is not made salient. Although still incipient, this research constitutes an example of how intergroup-relations research has more recently been characterized by emerging perspectives that reconsider the study of ideologies and social norms as important organizers of group behavior.

Fundamental Collective Beliefs, Ideologies, and Intergroup Relations

As stated above, the central role played by social cognition in intergroup relations should not be linked only to affect, emotions, and motivation but also to fundamental beliefs and ideologies that regulate social life. It was in this vein that Henri Tajfel (1984) associated his theory of social identity and intergroup relations with social change, perceptions of fairness, and legitimacy. He recommended studying the power of social myths and conceptions of social justice in intergroup relations. This analytical perspective was followed by the approaches described below. We begin this section by presenting some of the main theories about legitimacy and intergroup relations, after which we present some theories that link justice concerns to

intergroup relations. Finally, we present some implicit theories about human nature and their impact on intergroup relations.

Intergroup Relations and Legitimacy

Social Identity Theory

The importance of legitimacy processes attributed within SIT was initially considered by Turner and Brown (1978). These authors manipulated group status (high vs low), legitimacy (legitimate vs illegitimate), and stability (stable vs unstable) of this status. The results showed that perceiving relative group status as legitimate caused the high-status group to favor the in-group and the low-status group to favor the out-group. Moreover, in the condition where group status was seen as unstable and illegitimate, members of the high-status group diminished the degree of in-group favoritism and the low-status group revealed in-group favoritism. It was only in the latter condition that the low-status group exhibited some form of in-group bias. Indeed, the issue of legitimacy is a fundamental concern in democratic societies and legitimation is a core process in intergroup relations. Social Dominance Theory (SDT; Sidanius and Pratto, 1999) and System Justification Theory (Jost and Banaji, 1994) have a prominent place among the theories that have contributed to understanding the psychological mechanisms underlying legitimation.

Social Dominance Theory

Sidanius and Pratto (1999) proposed that societies evolved as 'group-based social hierarchies' and that these social hierarchies are served by legitimizing 'social myths.' The degree of endorsement granted to group-based hierarchies constitutes the level of social-dominance orientation (SDO) and the higher this orientation, the higher the level of support for social inequalities and discrimination against socially dominated groups. Several social myths enhance (hierarchy-enhancing myths) or weaken (hierarchy-attenuating myths) the legitimacy of social hierarchies, and the prominence of a group-based hierarchy within a given society is the result of the relative salience given to these two kinds of myths. According to SDT, meritocracy is an example of a 'hierarchy-enhancing' myth that produces an illusion of justice, serves the status quo, and legitimizes established social hierarchies. In fact, Katz and Hass (1988) used unobtrusive measures to experimentally demonstrate that the salience of egalitarian values decreases prejudice whereas the salience of meritocracy increases it. However, as Son Hing and colleagues demonstrated (Son Hing et al., 2011), it is important to make a critical distinction between meritocracy as a descriptive versus a prescriptive belief. Meritocratic principles can function as a hierarchy-legitimizing factor only among those who believe that meritocracy reflects current outcome distributions in society, that is, when meritocracy is a descriptive belief.

System Justification Theory

The justification of intergroup power relations and status inequalities was also construed within an ideological approach by System Justification Theory (e.g., Jost and Banaji, 1994). The goal of this theory is to analyze the conditions that facilitate the fact that most individuals adapt, defend, and justify the status quo and tolerate social injustice and intergroup inequalities. This theory proposes that, through the process of socialization, individuals acquire a motive for justifying the status quo even when it perpetuates their own disadvantaged position. People would engage in system justification because this serves as a *palliative function* as it reduces anxiety, discomfort, and uncertainty for both advantaged and disadvantaged individuals. In this context, the theory offers an ideological explanation both for out-group favoritism, which is frequently found in both correlational and experimental research among dominated or low-status groups, and for the in-group favoritism expressed by high-status groups, thereby indicating that both such positions indicate support for the system. Inspired by cognitive dissonance theory, this theory hypothesized that disadvantaged groups will show stronger support for the social system than members of advantaged groups because the former have a stronger need to reduce ideological dissonance through system justification. However, the robustness of this effect has not been demonstrated. Be that as it may, this constitutes a mere part of the story and the question about why individuals and groups strive for social change remains pertinent. Competitive interests, identity needs, and perceived injustice constitute factors that trigger collective action and promote social change.

The Justice Motive

Perceived Justice and Relative Group Deprivation

In the context of democratic societies, legitimation is a multilevel process and relative deprivation theories offer a framework by which to understand the origins of group status (il)legitimacy and some of the conditions for intergroup antagonism, that is, the feeling of injustice generated by the perception of relative deprivation. In this vein, Walker and Pettigrew (1984) revised the concept of relative deprivation in the context of intergroup relations and linked it to SIT (see also Mummendey and Wenzel, 1999, Kawakami and Dion, 1995). Correlational research has demonstrated that feelings of illegitimacy, negative intergroup attitudes, and a predisposition to engage in social action were generated by fraternal deprivation deriving from intergroup comparisons but not by egoistic or individual deprivation deriving from interpersonal comparisons. Significantly, some theoretical approaches to justice perceptions have proposed that judgments of justice and relative deprivation are based not only on the favorability of expected outcomes but also on the perception of respect for fair procedures (e.g., Tyler and Blader, 2003). If the feeling of group deprivation could generate judgments of illegitimacy, then the feeling of group gratification could produce feelings of legitimacy in relation to group-status inequality. Guimond and Dambrun (2002) hypothesized and experimentally demonstrated that when people are assigned to a condition of relative gratification they stereotype more and evaluate minorities such as immigrants more negatively. A possible explanation for this result could be that relative gratification facilitates a motivation to justify group advantages and social inequalities. Consequently, people in that condition express a greater degree of stereotyping and more negative evaluations of underprivileged minorities. This explanation of the consequences of group gratification proposed by Guimond is inspired by Social Dominance Theory and System Justification Theory, described above.

Belief in a Just World and Intergroup Relations

The role played by people's ideas about what is fair and just in the organization of individual and collective behaviors and attitudes constitutes an important legacy of the research in social psychology. Central to the legacy of the role played by justice concerns in our societal life is the Just World Theory proposed by Lerner (1980). According to this theory, individuals believe that the world is just in the sense that, at least implicitly, people get what they deserve and deserve what they get. In this sense, injustice, particularly the suffering of innocent victims, constitutes a threat to the belief in a just world, leading individuals to engage in different strategies to reestablish this belief, such as devaluing the victims' suffering or implicitly considering that they deserve to suffer. This theory was developed in order to understand judgments of fairness at the individual or interpersonal levels. The theory was, however, extended to intergroup relations by Correia et al. (2007). The authors analyzed the hypothesis that injustices that occur to innocent victims only threaten our belief in a just world if the victims belong to 'our world' (our in-groups) but not when victims are members of out-groups. This hypothesis was experimentally examined and confirmed using unobtrusive/implicit measures; in other words, the suffering of an out-group victim was not perceived as unjust. Quoting Tajfel (1984, p. 698): "*An individual has no need to construct his own justifications of inequity or injustice, so long as acts which are oppressive, exploiting, cruel, unjust or generally 'inhuman' are committed against certain groups whose members are socially or culturally characterised as being beyond the range within which apply certain principles (whatever they may be) of interpersonal conduct.*" The concept of the scope of justice (Opotow, 1990; Staub, 1990) is particularly important in this context. This concept refers to the existence (in people's minds) of a frontier for the application of justice principles. In other words, not all human beings are perceived to be part of our moral community. Out-groups can be included in the scope of justice if they are recategorized as in-group members, a process that can follow one of the strategies described in the previous section, that is, the development of a dual identity (Gaertner and Dovidio, 2000), a strategy that could be successful in improving intergroup relations in diverse societies.

Implicit Theories about the Nature of Human Beings and Intergroup Relations

In everyday social interactions, people create and learn collective beliefs or implicit theories or social representations (Dweck, 1999; Moscovici, 1984) that concern not only the social objects that are important for the survival of individuals and their groups but also the nature of what is typically human (Leyens et al., 2000) and the central and most useful attributes in perceiving individuals and groups (Fiske et al., 2002). These implicit theories about human nature have important consequences for intergroup relations.

Infra-Humanization Theory

While several theories have addressed the process by which out-group members are dehumanized in extreme cases of social or interpersonal conflicts (e.g., Bandura, 1999; Staub, 1990), the infra-humanization theory (Leyens et al., 2001) addressed one of the most common and pervasive forms of denial of full humanity to out-group members. This process, which has deep implications for intergroup relations, could be called "banal infra-humanization." Leyens and colleagues found that people implicitly distinguish between primary emotions (e.g., anger, joy, and fear) and secondary emotions (e.g., hope, regret, and enthusiasm) and that the latter are perceived as being learned later in life and are commonly understood to constitute an essential aspect of what makes a human being. Accordingly, Leyens and colleagues (see Vaes et al., 2012; for a review) also verified that, first, individuals are more likely to describe in-groups than out-groups in terms of secondary emotions; secondly, that in-groups are considered more able to express secondary emotions than out-group members; and thirdly, that secondary emotions but not primary emotions are more readily associated with the in-group than with out-groups. Moreover, in-group members who express a secondary emotion are seen to be more human whereas out-group members expressing the same emotion are not. Research in this domain has identified several moderators of out-group infra-humanization that largely cover the basic processes related to categorization, patterns of intergroup relations, and ideologies endorsed by group members. Specifically, it has been demonstrated that the more people perceive the in-group and the out-group as sharing a common supraordinate category (Gaertner and Dovidio, 2000), the more they attribute secondary emotions to out-groups (Gaunt, 2009). Moreover, the perception that the out-group is a threat to the in-group increases infra-humanization and the less human that the out-group is seen to be, the greater the symbolic threat that it poses to the in-group. This perception is then used to justify out-group discrimination (Pereira et al., 2009). Moreover, endorsement of SDO is associated with out-group infra-humanization (Esses et al., 2005) while other studies have shown that nationalism (but not patriotism) is linked to out-group infra-humanization.

Stereotype Content Model

This model, which was developed by Fiske et al. (2002), accentuates the importance of the content of stereotypes in intergroup relations and consistently links the basic or fundamental content dimensions according to which social judgments or impressions of social groups are formed to group status, patterns of intergroup relations, and emotions. Following the insights of Asch and Kelly, the theory assumes that categorical impression formation is based on two fundamental dimensions: warmth (i.e., honest, friendly, warm, and trustworthy) and competence (i.e., competent, capable, and skilled). Indeed, these two dimensions account for 80% of the variance in individual impression formation (Wojciszke, 2005). The model hypothesizes that these two dimensions can differentiate social groups into four meaningful quadrants. The high-warmth/high-competence quadrant typically describes in-groups while the remaining quadrants describe out-groups. Moreover, several studies have shown that perceived competition between in-groups and a given out-group lowers the perceived warmth of the out-group and that the perceived high status of an out-group increases the competence attributed to it. Regarding the consequences of placing groups into one of these clusters, research has revealed that it has an impact at both the emotional and the behavioral level. High-warmth/high-competence groups,

like in-groups, elicit pride whereas low-warmth/low-competence groups produce disgust. Ambivalent groups (high-warmth/low-competence or low-warmth/high-competence) induce mixed emotions, respectively, pity and envy. At the behavioral level, in-groups and pitied groups facilitate helping behavior but low-warmth groups elicit antagonistic behavior. Significantly, Harris and Fiske (2009) used fMRI procedures to compare activity in the mPFC (the region of the brain usually involved in the perception of individuals) in a group judgment task and verified, as predicted, that this region of the brain is activated for all groups except for low-warmth/low-competence groups.

Conclusions

The social psychology literature on intergroup relations is extensive, diversified and the result of over five decades of research. This article focused on two approaches that we consider to be prominent in the analysis of intergroup relations: on the one hand, the approach of social categorization and its connection with stereotypes, prejudice, and the relational dynamics deriving from its different configurations (e.g., recategorization, decategorization, and dual identity). Next, we examined the path from the salience of the social categorization process to the emerging salience attributed to the role of ideologies and collective beliefs on intergroup attitudes and discrimination. In this path, we showed how a full understanding of intergroup relations cannot be solely based on cognitive processes, and must attend also to motivational and emotional processes. SIT continues to represent a good example of articulation of these aspects. In the second section of this article, we demonstrated how ideologies and social norms constitute fundamental triggers of intergroup relations. Indeed, ideologies and social norms lie at the root of both harmonious and conflictual intergroup relations. Until now, most literature has considered the influence of such factors as moderators of the cognitive processes. However, the analysis of the research produced in the last years in social psychology shows how the cognitive and the ideological and normative aspects have equal explanatory power over the dynamics of intergroup relations.

See also: Authoritarian Personality; Belief in a Just World; Collective Action; Collective Behavior, Social Psychology of; Dehumanization; Emotions and Intergroup Relations; Helping Intergroup Relations; Justice: Social Psychological Perspectives; Meritocracy and Tokenism; Racism: Social Psychological Perspectives; Representations, Social Psychology of; Self-Categorization Theory; Social Categorization; Social Dominance Orientation; Social Identity in Social Psychology; Social Psychology; Stereotypes in Social Psychology; Xenophobia: Social Psychological Aspects.

Bibliography

Allport, G.W., 1954. The nature of prejudice. Reading, MA: Addison-Wesley.

Bandura, A., 1999. Moral disengagement in the perpetration of inhumanities. Personality and Social Psychology Review 3 (3), 193–209.

Brewer, M.B., 1999. The psychology of prejudice: ingroup love and outgroup hate? Journal of social issues 55 (3), 429–444.

Brewer, M.B., Miller, N., 1984. Beyond the contact hypothesis: theoretical perspectives on desegregation. In: Miller, N., Brewer, M.B. (Eds.), Groups in Contact: The Psychology of Desegregation. Academic Press, Orlando, FL, pp. 281–302.

Correia, I., Vala, J., Aguiar, P., 2007. Victim's innocence, social categorization, and the threat to the belief in a just world. Journal of Experimental Social Psychology 43 (1), 31–38.

Costa-Lopes, R., Pereira, C.R., Judd, C.M., 2014. Categorisation salience and ingroup bias: the buffering role of a multicultural ideology. International Journal of Psychology 1–5. http://dx.doi.org/10.1002/ijop.12044. Advanced online publication.

Crandall, C.S., Eshleman, A., 2003. A justification-suppression model of the expression and experience of prejudice. Psychological Bulletin 129 (3), 414.

Deschamps, J.C., Brown, R.J., 1983. Superordinate goals and intergroup conflict. British Journal of Social Psychology 22, 189–195.

Doise, W., 1976. L'articulation psychosociologique et les relations entre groupes. De Boeck, Brussels.

Duckitt, J.H., 1992. Psychology and prejudice: a historical analysis and integrative framework. American Psychologist 47 (10), 1182–1193.

Dweck, C.S., 1999. Self-theories: Their Role in Motivation, Personality, and Development. Psychology Press, Philadelphia.

Esses, V.M., Jackson, L.M., Dovidio, J.F., Hodson, G., 2005. Instrumental relations among groups: group competition, conflict, and prejudice. In: Dovidio, J.F., Glick, P., Rudman, L. (Eds.), Reflecting on the Nature of Prejudice. Blackwell, Oxford, UK, pp. 227–243.

Fiske, S.T., Cuddy, A., Glick, P., Xu, J., 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology 82 (6), 878–902.

Gaertner, S.L., Dovidio, J.F., 2000. Reducing Intergroup Bias: The Common Ingroup Identity Model. Psychology Press, Philadelphia, PA.

Gaunt, R., 2009. Superordinate categorization as a moderator of mutual infrahumanization. Group Processes & Intergroup Relations 12 (6), 731–746.

Guerra, R., Rebelo, M., Monteiro, M.B., Riek, B., Mania, E., Gaertner, S.L., Dovidio, J., 2010. How should intergroup contact be structured to reduce bias among majority and minority group children? Group Processes and Intergroup Relations 13 (4), 445–460.

Guimond, S., Dambrun, M., 2002. When prosperity breeds intergroup hostility: relative deprivation and relative gratification. Personality and Social Psychology Bulletin 28, 900–912.

Hamilton, D.L., Sherman, S.J., 1989. Illusory correlations: implications for stereotype theory and research. In: Bar-Tal, D., Graumann, C.F., Kruglanski, A.W., Stroebe, W. (Eds.), Stereotyping and Prejudice: Changing Conceptions. Springer-Verlag, New York, pp. 59–82.

Harris, L.T., Fiske, S.T., 2006. Dehumanizing the lowest of the low neuroimaging responses to extreme out-groups. Psychological Science 17 (10), 847–853.

Harris, L.T, Fiske, S.T., 2009. Social neuroscience evidence for dehumanised perception. European review of social psychology 20 (1), 192–231.

Hewstone, M., Brown, R., 1986. Contact and Conflict in Intergroup Encounters. Blackwell, Oxford.

Jost, J.T., Banaji, M.R., 1994. The role of stereotyping in system – justification and the production of false consciousness. British Journal of Social Psychology 33 (1), 1–27.

Katz, I., Hass, R.G., 1988. Racial ambivalence and American value conflict: correlational and priming studies of dual cognitive structures. Journal of Personality and Social Psychology 55 (6), 893.

Kawakami, K., Dion, K.L., 1995. Social identity and affect as determinants of collective action toward an integration of relative deprivation and social identity theories. Theory & Psychology 5 (4), 551–577.

Lerner, 1980. The Belief in a Just World: A Fundamental Delusion. Plenum, New York.

LeVine, R.A., Campbell, D.T., 1972. Ethnocentrism: Theories of Conflict, Ethnic Attitudes, and Group Behavior. John Wiley & Sons, New York.

Leyens, J.P., Paladino, P.M., Rodriguez-Torres, R., Vaes, J., Demoulin, S., Rodriguez-Perez, A., Gaunt, R., 2000. The emotional side of prejudice: the attribution of secondary emotions to ingroups and outgroups. Personality and Social Psychology Review 4 (2), 186–197.

Leyens, J.P., Rodriguez-Perez, A., Rodriguez-Torres, R., Gaunt, R., Paladino, M.P., Vaes, J., Demoulin, S., 2001. Psychological essentialism and the differential attribution of uniquely human emotions to ingroups and outgroups. European Journal of Social Psychology 31 (4), 395–411.

Lorenzi-Cioldi, F., 1993. They all look alike, but so do we… sometimes: perceptions of in-group and out-group homogeneity as a function of sex and context. British Journal of Social Psychology 32 (2), 111–124.

Mackie, D.M., Devos, T., Smith, E.R., 2000. Intergroup emotions: explaining offensive action tendencies in an intergroup context. Journal of Personality and Social Psychology 79 (4), 602–616.

Marques, J.M., Páez, D., 1994. The black sheep effect: social categorisation, rejection of ingroup deviates, and perception of group variability. European Review of Social Psychology 5, 37–68.

Moscovici, S., 1984. The phenomenon of social representations. In: Farr, R., Moscovici, S. (Eds.), Social Representations. Cambridge University Press, Cambridge, pp. 3–70.

Mummendey, A., Wenzel, M., 1999. Social discrimination and the tolerance in intergroup relations: reactions to intergroup difference. Personality and Social Psychology Review 3, 158–174.

Opotow, S., 1990. Moral exclusion and injustice: an introduction. Journal of Social Issues 46 (1), 1–20.

Park, B., Judd, C.M., 2005. Rethinking the link between categorization and prejudice within the social cognition perspective. Personality and Social Psychology Review 9 (2), 108–130.

Pereira, C., Vala, J., Leyens, J.P., 2009. From infra-humanization to discrimination: the mediation of symbolic threat needs egalitarian norms. Journal of Experimental Social Psychology 45 (2), 336–344.

Pettigrew, T.F., 1959. Regional differences in anti-Negro prejudice. The Journal of Abnormal and Social Psychology 59 (1), 28–36.

Pettigrew, T.F., Tropp, L.R., 2006. A meta-analytic test of intergroup contact theory. Journal of Personality and Social Psychology 90 (5), 751–783.

Sidanius, J., Pratto, F., 1999. Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression. Cambridge University Press, New York.

Son Hing, L.S., Bobocel, D.R., Zanna, M.P., Garcia, D.M., Gee, S.S., Orazietti, K., 2011. The merit of meritocracy. Journal of Personality and Social Psychology 101 (3), 433–450.

Staub, E., 1990. Moral exclusion, personal goal theory, and extreme destructiveness. Journal of Social Issues 46 (1), 47–64.

Stephan, W.G., 1985. Intergroup relations. In: Lindzey, G., Aronson, E. (Eds.), Handbook of Social Psychology, vol. III. Addison-Wesley, New York, pp. 599–658.

Tajfel, H., 1984. Intergroup relations, social myths and social justice in social psychology. In: Tajfel, H. (Ed.), The Social Dimension: European Developments in Social Psychology. Cambridge University Press, Cambridge, UK, pp. 695–715.

Tajfel, H., Turner, J.C., 1979. An integrative theory of intergroup conflict. In: Austin, W.G., Worchel, S. (Eds.), The Social Psychology of Intergroup Relations. Brooks/Cole, Monterey, CA, pp. 33–47.

Turner, J.C., Hogg, M.A., Oakes, P.J., Reicher, S.D., Wetherell, M.S., 1987. Rediscovering the Social Group: A Self-Categorization Theory. Basil Blackwell, Oxford, UK.

Turner, J.C., Brown, R., 1978. Social status, cognitive alternatives, and intergroup relations. In: Tajfel, H. (Ed.), Differentiation between Social Groups: Studies in the Social Psychology of Intergroup Relations. Academic Press, London, pp. 201–234.

Tyler, T.R., Blader, S.L., 2003. The group engagement model: procedural justice, social identity, and cooperative behavior. Personality and Social Psychology Review 7 (4), 349–361.

Vaes, J., Leyens, J.P., Paola Paladino, M., Pires Miranda, M., 2012. We are human, they are not: driving forces behind outgroup dehumanisation and the humanisation of the ingroup. European Review of Social Psychology 23 (1), 64–106.

Vignoles, V.L., Chryssochoou, X., Breakwell, G.M., 2000. The distinctiveness principle: identity, meaning and the bounds of cultural relativity. Personality and Social Psychology Review 4 (4), 337–354.

Walker, I., Pettigrew, T.F., 1984. Relative deprivation theory: an overview and conceptual critique. British Journal of Social Psychology 23 (4), 301–310.

Wojciszke, B., 2005. Morality and competence in person-and self-perception. European Review of Social Psychology 16 (1), 155–188.

Interhemispheric Interaction in Language and Cognitive Processes

Dimitrios S Kasselimis, University of Crete, Rethymno, Greece and National and Kapodistrian University of Athens, Greece
Andreas Nidos, Department of Neurology, Metropolitan Hospital, Neo Faliro, Greece

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article's focus is on interhemispheric interaction in relation to cognitive processes. Commissural fibers facilitate interhemispheric transfer, which in turn supports several cognitive functions, such as attention, memory, and language. Two major theories have been proposed to explain the specific mechanisms of interhemispheric communication, lateralization, and hemispheric specialization: the excitatory model and the inhibitory model. Neuropsychological and neuropsychiatric manifestations of callosal syndromes are also discussed.

Introduction

In the middle of the eighteenth century, scholars working in the field of acquired language impairment provided rich evidence for cerebral lateralization, a notion that contradicted the status quo ante, i.e., the belief established by Franz-Joseph Gall, that higher cognitive functions were facilitated equally by two identical hemispheres (Tesak and Code, 2008). A 100 years later, Geschwind (1965a,b) reintroduced the idea of disconnection syndromes, according to which, several neurological disorders can be attributed to interruption of white matter tracts allowing cortico-cortical communication within a neural network supporting cognitive functions, such as language and visuospatial processing (see also Catani and ffytche, 2005). This interhemispheric interaction is possible due to the commissural fibers, the greatest of all being the corpus callosum. Studies of comparative anatomy provide useful clues regarding the phylogenetic origins of the corpus callosum and hemispheric specialization (Aboitiz and Montiel, 2003; Hopkins and Cantalupo, 2008). From an evolutionary perspective, the emergence of the corpus callosum could have played an integral role in the development of cognitive functions and hemispheric lateralization (Doron and Gazzaniga, 2008; Gazzaniga, 2000). In the present article, we will discuss issues with regard to interhemispheric communication via the commissural fibers, focusing on cognitive processes and more particularly language.

Anatomy of the Commissural Fibers

The commissural fibers are white matter tracts that interconnect the two cerebral hemispheres. The majority of these bundles interconnect homologous cortical areas, but there are also fibers participating in heterotopic connections. Homotopic commissural connections interconnect corresponding cortical areas in both hemispheres, whereas heterotopic fibers connect a cortical area of one hemisphere with nonhomologous areas in the contralateral hemisphere (Knyazeva, 2013; Nieuwenhuys et al., 2008; Zarei et al., 2006). The largest commissural fiber system is the *corpus callosum*, which forms the floor of the longitudinal fissure and the roof of the lateral ventricles. It contains more than 300 million axons and interconnects the neocortical portions of the two hemispheres. These fibers are of two different types: large-diameter fibers that mediate sensory–motor coordination and small-diameter fibers that interconnect association areas (Yazgan et al., 1995). The corpus callosum is divided into a bloated posterior part, the *splenium*; a large middle part, the *body* (or *truncus*); an anterior arched part, the *genu*; and a final tapered part, the *rostrum*, which connects the genu with the lamina terminalis. There is also an intermediate, thinned part between the truncus and the splenium, the isthmus. The axons of the corpus callosum expand into the white matter of both hemispheres, thus forming the *radiation of the corpus callosum* (Nieuwenhuys et al., 2008). The rostrum connects fronto-basal cortical regions of the two hemispheres. The genu interconnects prefrontal areas and the anterior cingulated cortex. The truncus connects motor and somatosensory regions, such as the precentral cortex, the adjacent portion of the insula, and the overlying cingulated gyrus. The caudal part of the truncus along with a portion of the splenium connects cortices from the temporal–parietal–occipital junction. The isthmus connects pre- and postcentral gyri, and the primary auditory cortex. The splenium connects posterior parietal and occipital regions, as well as medial–temporal cortices (Banich, 1995; Hellige, 1993; Raybaud, 2010). Even though callosal fibers typically project to homologous hemispheric regions (Hellige, 1993; McCulloch and Garol, 1941; Putnam et al., 2008), a small proportion projects to heterotopic regions (Clarke and Zaidel, 1994; Cook, 1984; Rakic and Yakovlev, 1968). Another commissural fiber system is the *anterior commissure*. This is a transversely oriented axon bundle and consists of two parts: a small anterior limb which interconnects olfactory cortices and a larger posterior limb which interconnects anterior areas of the middle and inferior temporal gyri. The third commissural fiber system interconnecting the two cerebral hemispheres is the *hippocampal commissure*, a part of the fornix located below the splenium that connects the two hippocampi (Nieuwenhuys et al., 2008).

Excitation versus Inhibition

Researchers have yet to establish a clear-cut pattern with regard to regulation of interhemispheric communication via the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54039-2

commissural fibers (Liederman, 1998). In terms of neurochemistry, the largest commissural fiber system, the corpus callosum can be characterized as a bundle of excitatory fibers, since the most common neurotransmitter of the callosal axons is glutamate (Conti and Manzoni, 1994). However, callosal signals can be inhibitory, due to the presence of inhibitory interneurons (Kawaguchi, 1992). Overall, the final functional effect of callosal connections can be either excitatory or inhibitory, depending on a variety of factors, including neurotransmitters, receptors, and interneurons. Due to lack of concluding evidence, the debate on a hypothesized excitatory versus inhibitory callosal function remains unresolved, focusing on two opposing models (van der Knaap and van der Ham, 2011; Bloom and Hynd, 2005).

The first model comprises theories of inhibition. According to it, the size of the corpus callosum is analogous to the degree of interhemispheric connectivity (Aboitiz et al., 1992a,b), which in turn is positively correlated with lateralization emerging due to interhemispheric inhibition (van der Knaap and van der Ham, 2011). The inhibitory model presumes that, during development, one hemisphere can inhibit the other via the corpus callosum, thus becoming dominant for a particular function (Hellige, 1993). There is reported evidence suggesting an essential role of inhibitory callosal action during ontogeny (Dennis, 1976). This phenomenon – called *metacontrol* – allows intrahemispheric processing and aids hemispheric asymmetry. It is based on the idea that transferring information between hemispheres and interhemispheric integration is energy- and time-consuming; therefore intrahemispheric processing could be more efficient with regard to simple tasks (Urgesi et al., 2005; Welcome and Chiarello, 2008). During the last four decades, many scholars have postulated several theories of inhibition. For example, Kinsbourne (1975) claimed that activation in the two hemispheres is mediated by the corpus callosum to be inversely proportional, in the context of a continuous inhibitory interaction of cortical areas connected by callosal fibers. On the other hand, Cook (1984) proposed the homotopic inhibition theory, according to which, the corpus callosum interconnects homologous areas of the two hemispheres in an inhibitory mode, and these areas also connect to adjacent cortical regions in a reciprocally inhibitory fashion. Although many variations of the inhibitory model have been proposed, all theories of inhibition state that hemispheric asymmetry is accomplished – at least in part – via callosal inhibitory mediation, which ultimately leads to hemispheric dominance for a given function, and that the degree of lateralization is proportional to the size of the corpus callosum (Hellige, 1993; Bloom and Hynd, 2005).

There are a few studies supporting the inhibitory model, by demonstrating independent intrahemispheric processing when verbal stimuli are presented to either of the two hemispheres (Hellige, 1993) or bilateral parallel processing for complex tasks (Banich and Belger, 1990). Further support is provided by dichotic listening studies, where different stimuli are presented simultaneously in each ear. According to Clarke et al. (1993), corpus callosum size is not related to laterality, but it is negatively correlated with superior ear accuracy, therefore possibly associated with the degree of functional interhemispheric inhibition. Research on the aging of the corpus callosum can also offer evidence in favor of the inhibitory view. A recent study on the possible relation between age-related degeneration of the corpus callosum and interhemispheric inhibition showed that reduced resting interhemispheric connectivity in older adults was associated with increased ipsilateral motor cortex recruitment during a simple motor task (Langan et al., 2010). The authors suggest that age-related reduced callosal connectivity reflects a deficient inhibition of the nondominant hemisphere during a simple motor task, a phenomenon resulting in greater reaction times, thus impaired performance. Putnam et al. (2008) combined functional magnetic resonance imaging and diffusion tensor imaging to investigate the relationship between callosal structure and bilateral cortical activity in healthy young adults during a verbal encoding task involving semantic judgment. While task performance was associated with increased activity in the left inferior prefrontal cortex as expected, they found that participants with low anterior callosal fractional anisotropy values demonstrated significantly greater activity in the right homologous region, compared to participants with high anterior callosal fractional anisotropy values. These results are consistent with the inhibitory model in the sense that a thinner corpus callosum would result in decreased interhemispheric inhibition indicated by increased activation of the homologous contralateral cortical area. Finally, research on developmental dyslexia suggests that the size of the corpus callosum is smaller in dyslexic children, and that splenium and genu sizes are associated with reading achievement (Hynd et al., 1995). Overall, reported data supporting the inhibitory model are scarce. It should be also noted that a study specifically designed to test Cook's homotopic inhibition theory failed to find any supporting evidence for interhemispheric inhibition mediated by the corpus callosum (Chiarello et al., 1995).

The second model is based on theories of excitation. It postulates that integration of information is enforced by interhemispheric excitation via the commissural fibers, and that the size of corpus callosum decreases as hemisphere laterality increases (Bloom and Hynd, 2005). The first evidence in favor of this view comes from early callosotomy studies (Gazzaniga and Sperry, 1967; Gazzaniga et al., 1963) showing that these patients demonstrated difficulties in integrating information received exclusively by the right or the left hemisphere. Subsequent studies confirmed that even if split-brain patients recover well a few years post surgery, being able to function successfully in everyday living activities (Sergent, 1983), they show a variety of deficits in the context of a disconnection syndrome (Banich, 1995; Crandall, 1985; Gazzaniga, 1970; Gazzaniga and Sperry, 1967; Lassonde, 1986). In sum, split-brain studies support the core idea of the excitatory model, by presenting evidence that the corpus callosum is essential for cognitive functions, through facilitating interhemispheric transfer and integration of sensory information.

One of the fundamental predictions of this model is that hemispheric asymmetry and size of the corpus callosum are inversely proportional. Studies in this direction have confirmed this hypothesis. Anatomical postmortem (Aboitiz et al., 1992a,b; Aboitiz et al., 1992) and *in vivo* (Dorion et al., 2000) studies have demonstrated an inverse association

between the size of the corpus callosum and asymmetry of homologous cortical areas. Research utilizing specialized neuropsychological tests and structural MRI has shown that performance on behavioral laterality measures is significantly negatively correlated with corpus callosum size (Clarke and Zaidel, 1994; Yazgan et al., 1995). Galaburda (1984) suggested that stronger interhemispheric connectivity leads to greater symmetry, thus indicating that emerging asymmetry during development is due to lack of interhemispheric excitatory connections. The logical consequence of this assumption is that a smaller corpus callosum would be associated with increased functional laterality, which is essential for cognitive functions, especially language – a rather lateralized function. Geschwind and Galaburda (1985) have provided anatomical evidence that the degree of hemispheric asymmetry (focusing on left language areas and their homologous areas in the right hemisphere) differs between individuals with dyslexia and normal readers. Galaburda (1995) suggests that developmental dyslexia is related to a larger right planum temporale, which interferes with the homologous dominant left area during language processing. He attributes this unexpected asymmetry to insufficient pruning. The aforementioned interference of the nondominant right planum temporale is supposed to manifest itself via the corpus callosum. This phenomenon would be aided by increased callosal connections, which could be also due to inadequate pruning during development (Galaburda et al., 1990). It should be although noted that while most researchers find abnormalities with regard to the size (Duara et al., 1991; Hynd et al., 1995; Larsen et al., 1990; von Plessen et al., 2002) and shape (Robichon et al., 2000; von Plessen et al., 2002) of the corpus callosum of individuals with developmental dyslexia, the reported data are contradicting. Some studies have found a larger corpus callosum in dyslexic readers (Duara et al., 1991; Larsen et al., 1990), others have failed to find any difference in the size of the corpus callosum between healthy controls and persons with developmental reading or language disability (Preis et al., 2000), and there are also studies providing evidence in favor of the inhibitory model (reporting smaller corpus callosum areas in dyslexics compared to normal readers) (Hynd et al., 1995; von Plessen et al., 2002). In view of such conflicting evidence, Bloom and Hynd (2005) stress the idea that the corpus callosum's size and shape may not be related to developmental dyslexia and, consequently, to reading ability.

Evidence for an excitatory role of the corpus callosum also comes from studies including patients with acquired or developmental callosal lesions. Callosotomies have been performed for many decades on epileptic patients, in order to constrain seizures. The idea behind these surgeries is that seizures from one hemisphere propagate to the other via the corpus callosum. Therefore, the effectiveness of callosotomy in preventing interhemispheric spreading of the epileptic seizures provides evidence supporting the excitatory function of the callosal fibers (but see also Roberts (1999) for a proposed inhibitory callosal function in such patients). Individuals with congenital agenesis of the corpus callosum typically demonstrate subtle linguistic and visuospatial deficits (Chiarello, 1995), similar to split-brain patients (Gazzaniga et al., 1963). Moreover, research on callosal agenesis contradicts the inhibitory view, since hemispheric lateralization develops normally and typically lateralized functions are not represented bilaterally in such patients (Lassonde, 1986). Finally, comparative anatomy research from the field of evolutionary biology provides further support to the excitatory view, by demonstrating that as laterality increases, corpus callosum size decreases across primates (Hopkins and Rilling, 2000). These data offer insight into emerging hemispheric lateralization in terms of phylogeny, and indicate that brain asymmetry evolved in Hominidae due to reduced interhemispheric connectivity.

Overall, one can find evidence for both theories in the relevant literature. However, reality with regard to interhemispheric interactions may not correspond exclusively to an inhibitory or an excitatory model. On the contrary, there may be several ways of such interactions, depending on the processing demands of a given task. Thus, these two main types of interaction are not mutually exclusive, and commissural fibers such as the corpus callosum may be inhibitory or excitatory, on the basis of specific needs for a particular cognitive function (type of stimuli, processing demands, hemispheric specialization, etc.). This proposed flexibility of the corpus callosum is suggested to increase processing efficiency in terms of time and overall performance. In some instances, when one hemisphere is superior to the other for a given task (e.g., the left hemisphere for processing linguistic information), independent operation of the hemispheres and intrahemispheric processing might be more advantageous, and therefore callosal isolation would be beneficial. On the other hand, when a particular type of stimulus can be equally processed by both hemispheres, sharing of hemispheric resources and interhemispheric cooperation might be more advantageous, and therefore callosal excitation would increase processing efficiency and performance (Bloom and Hynd, 2005; Schulte and Müller-Oehring, 2010). From a phylogenetic perspective, the emergence of anatomical and functional lateralization per se can be explained by either of the two theories. On the basis of the inhibitory claim, there could be an isolation of the contralateral hemisphere via the corpus callosum, thus leading to asymmetrical hemispheric development. In contrast, an excitatory function of the corpus callosum could have allowed unilateral mutations during interhemispheric integration. These mutations could provide a substrate for the evolutionary origins of lateralization (van der Knaap and van der Ham, 2011).

Commissural Fibers and Cognitive Functions: Interhemispheric Cooperation, Information Transfer, and Integration

There are two types of callosal fibers in terms of size and myelination. Thick, heavily myelinated fibers with a large diameter can save time by fast transmission of sensory and motor information, while thinner, less myelinated fibers with a smaller diameter demonstrate slower transmission rates between association cortical areas (Aboitiz, 1992; Aboitiz and Montiel, 2003; Bloom and Hynd, 2005; Knyazeva, 2013; Ringo et al., 1994). In any case, interhemispheric transfer and integration is a time- and energy-consuming process and

it is thus not always beneficial. However, transferring information between hemispheres can be proved to be advantageous when the cognitive load is increased. Simple tasks can be processed intrahemispherically, while highly demanding complex tasks may require bilateral activation and interhemispheric cooperation (Banich, 1998; Banich and Belger, 1990; Braver et al., 1997; Smith et al., 1996; Welcome and Chiarello, 2008). Therefore bilateral processing is preferred over processing within a single hemisphere when load-dependent benefits outweigh time and energy costs. Hemispheric lateralization for higher cognitive functions, such as the left-hemisphere dominance for language and specialization of the right hemisphere for visuospatial attention, has been argued to be a consequence of interhemispheric connectivity. It has been suggested that, in the course of the primates' evolution, intrahemispheric processing, and consequently hemispheric specialization, was induced by increased time constraints of callosal conduction delay (Hopkins and Rilling, 2000).

Callosal connections are traditionally associated with transfer and integration of sensorimotor information (Banich, 1998; Fling et al., 2011; Franz et al., 2000; Gerloff and Andres, 2002; Johansen-Berg et al., 2007; Wahl et al., 2007). However, contemporary research indicates the involvement of the corpus callosum in higher cognitive functions, and the relationship between the integrity of white matter microstructure in this commissural fiber system and performance on a variety of cognitive tasks (Roberts et al., 2013). Several studies point out that bihemispheric interaction is essential for visuospatial attention, object recognition, and processing of complex visual stimuli (Baird et al., 2005; Banich, 1998; Banich et al., 2000; Corbetta et al., 2005; Kinsbourne, 1977; Kinsbourne and Bruce, 1987; Reuter-Lorenz et al., 1990; Schulte et al., 2001; Skiba et al., 2000). Clinical research has shown that the integrity of the corpus callosum is essential for selective attention and conflict processing, in the context of interhemispheric transferring of information within a fronto-parietal attention neural network (Schulte et al., 2006). Studies on visual perception, including patients with callosal abnormalities due to normal aging (Müller-Oehring et al., 2007), commissurotomy (Robertson et al., 1993), HIV infection (Müller-Oehring et al., 2010), and alcoholism (Müller-Oehring et al., 2009), showed that compromised integrity of the corpus callosum is related to impaired local–global processing of visual stimuli. (The process of decomposing the visual world into distinct component features and the re-integrating the latter into a complex visual stimulus). Schulte and Müller-Oehring (2010) suggest that such findings indicate that lateralized whole-part perceptual integration is mediated by the corpus callosum and more specifically, "anterior callosal integrity appears to mediate local-global interference (callosal inhibition), whereas posterior callosal integrity seems to mediate local-global facilitation (callosal cooperation)" (p. 183).

Even though the corpus callosum is the greater commissural fiber system, therefore the main white matter pathway supporting interhemispheric connectivity, there are also two additional subcortical commissures, as mentioned earlier: the anterior commissure, which connects the two cerebral hemispheres at the lateral striatal ridge, and the hippocampal commissure, which connects two hippocampi. In terms of ontogeny, both of these structures develop earlier in gestation compared to the corpus callosum (Bloom and Hynd, 2005). Studies involving 'split-brain' patients (i.e., patients who have undergone callosotomy) have shown that such individuals appear to be unaffected with regard to general cognitive capacity, sense of self, and affect, and retain certain cognitive abilities, such as integrating some aspects of spatial attention and emotional information, despite the absence of callosal interhemispheric connections (Sperry et al., 1979; Gazzaniga, 1970). (It should be however noted that cognition does not remain entirely unaffected. Even though several cognitive functions, such as verbal intelligence and problem-solving capacity remain intact, specific deficits, such as in coordination, stereognosis, free recall, and short-term memory have been reported (Gazzaniga, 2000; Josheph and Bannister, 2001; Phelps, Hirst, and Gazzaniga, 1991; Zaidel and Sperry, 1974).) Moreover, acallosal patients (individuals with congenital agenesis of the corpus callosum), often show only mild cognitive deficits (Josheph and Bannister, 2001). Their difficulties remain usually undetected (in some cases callosal agenesis is revealed postmortem, after autopsy) throughout their lifespan. Normally these patients are identified only when a comorbid condition requires structural imaging (Darby and Walsh, 2005; Harris, 1995; Kasselimis and Potagas, unpublished data; Zaidel et al., 2003). Taken together, clinical data from split-brain and acallosal patients suggest that at least for some cognitive functions an intact corpus callosum does not seem to be a prerequisite, and interhemispheric transfer and integration of information may be supported by the anterior and the hippocampal commissure (but see also Gazzaniga, 2000).

Interhemispheric Interaction and Language

Since the paradigm-shifting case studies of Pierre Paul Broca (1861a,b,c, 1863), it is widely accepted that language is mainly facilitated by the left hemisphere. This specialization is also designated by functional and structural hemispheric asymmetries (Steinmann and Mulert, 2012). Rich research evidence derived from split-brain patients reveals that phonology, lexicon, and syntax are supported mainly by left-hemisphere cortices, while the opposite hemisphere appears to be rather limited regarding language and it can process only lower-level linguistic material (Baynes, 1990; Baynes and Eliassen, 1998; Baynes and Gazzaniga, 1988; Baynes et al., 1997, 1995, 1970, 1983, 2000; Gazzaniga and Miller, 1989; Gazzaniga et al., 1984, 1979; Gazzaniga and Sperry, 1967; Reuter-Lorenz and Baynes, 1992; Sidtis et al., 1981; Zaidel, 1991; Zaidel and Peters, 1981; Zeidel and Seibert, 1997).

Even though it is the left hemisphere that mainly supports language, interhemispheric cooperation has been shown to be involved in linguistic functions. Gazzaniga (2000) argues that specific callosal fibers transfer semantic representations and word information, essential for processing verbal stimuli. In terms of anatomy, there are several language-related cortical regions of the left hemisphere interconnected with their right homologues via the corpus callosum. It has been further claimed that the structural characteristics of such

connections relate to performance on speech perception tasks (for a review, see Steinmann and Mulert, 2012). Some studies argue in favor of a topographically specific functional role of the corpus callosum. For example, the posterior portion of the greater commissural fiber bundle, including the splenium and the perisplenial white matter, is shown to be crucial for phonological awareness (Dougherty et al., 2007) and the interhemispheric interplay between prosodic and syntactic information (Friederici et al., 2007). While syntax is facilitated mainly by the left hemisphere and prosodic features of speech are processed by right hemisphere cortices, the interaction of these functions relies on interhemispheric communication via the posterior corpus callosum (Friederici et al., 2007). Taking into account the accumulating data on the role of commissural fibers in language processing, Friederici (2011) corroborates the idea that the corpus callosum "builds the brain basis for the integration of local syntactic and prosodic features during auditory speech comprehension, as it connects the brain regions in which syntax and prosody are computed" (p. 1384). Finally, it has been argued that interhemispheric connections also subserve language comprehension in the context of processing semantic features of verbal information. In his recent review, Jung-Beeman (2005) suggests that current neuroimaging data support a framework for verbal semantic processing, which consists of three components: bilateral activation, integration, and selection. These components are localized bihemispherically and comprise a bilateral interactive semantic network.

Callosal Syndromes: Neuropsychological and Neuropsychiatric

Compromised integrity of commissural fibers has been related to many clinical conditions, such as perceptual deficits (Zaidel et al., 2003); dyslexia (Klingberg et al., 2000; Zaidel et al., 1998); stuttering (Chang et al., 2008; Sommer et al., 2002); neuropsychiatric features of alcoholism (Harper and Kril, 1988, 1990; Harper and Matsumoto, 2005; Pfefferbaum and Sullivan, 2002, 2005; Pfefferbaum et al., 2006; Tarnowska-Dziduszko et al., 1995); bipolar disorder (van der Knaap and van der Ham, 2011); neurodegenerative diseases, such as Alzheimer's disease, frontotemporal dementia, and amyotrophic lateral sclerosis (Murray et al., 2006); cognitive impairment and fatigue in multiple sclerosis (Yaldizli et al., 2013; van der Knaap and van der Ham, 2011); dysgraphia (Buklina, 2005); and attention-deficit hyperactivity disorder (McNally et al., 2010; Schnoebelen et al., 2010; van der Knaap and van der Ham, 2011). The detailed description of all possible impairments due to commissural fibers' abnormalities is beyond the scope of this article; therefore we will briefly discuss some of the main neuropsychological and neuropsychiatric callosal syndromes.

Pure alexia (or alexia without agraphia) is among of the renowned disconnection syndromes. It was originally described by Déjérine (1892), who reported the case of a stroke patient who had no difficulties in speech, auditory comprehension, or writing, nevertheless was unable to read. The patient's lesion was located in the left occipital lobe and included portion of the splenium. Dejerine suggested that the syndrome was due to the fact that visual information could not reach the language region in the left hemisphere. Several cases of alexia with various underlying neurological etiologies have been reported since then, the majority of which is attributed to some kind of interhemispheric disconnection. Pure alexia is usually suggested to be the result of a lesion interrupting communication between visual areas of both occipital lobes with language perisylvian areas of the left hemisphere (Coslett, 2003).

Among the psychiatric disorders that researchers have attempted to interpret under a broad theory of disconnectionism, schizophrenia has surely the first place. Wernicke was the first to propose a functional disconnection in the brain in schizophrenia (Wernicke, 1906), and this notion came to light again during the 1990s (Friston and Frith, 1995). A disconnection hypothesis of schizophrenia holds that failure in cognition arises from pathophysiology, i.e., impairment in functional integration within the brain. Friston argues that the pathophysiology in schizophrenia is due to abnormal connections within a brain with abnormal plasticity. An anomaly in connectivity is a matter of discontinuation in the synaptic flux (Friston, 1998). Schizophrenic symptoms in adults could be then evolved from dysplastic fatal brain development (Bullmore et al., 1997). The classic disconnection hypothesis though, may not account for possible changes within or between the hemispheres. Advances in imaging studies and neurostatistical techniques during the last decade offered us adequate evidence both for structural and functional connectivity for widespread disconnection in resting-state brain networks, especially in the default network (Guo et al., 2013). An investigation of abnormal interhemispheric connectivity in schizophrenia emphasizes the role of the anterior part of the corpus callosum (Crow, 1998). Also, the size of the anterior commissure as measured with diffusion tensor imaging has also been found to play a role in schizophrenia (Choi et al., 2011). Guo et al. (2014) have suggested a weakening of functional connectivity between the two brain hemispheres involving a number of cortical and limbic regions.

Autistic spectrum conditions form a diagnostic umbrella of a wide range of pervasive neurodevelopmental disorders with an onset in early childhood and a life-long symptomatology. Deficits in cognition, emotional processing, and social functionality are being studied thoroughly at the neurological level for the past decade, in many prior imaging studies (Castelli et al., 2002; Just et al., 2004; Koshino et al., 2005, 2008). Abnormal long-range functional connectivity has been studied in recent explanatory models of autism. The matter of decreased hemispheric connectivity seems a quite plausible explanation for the defective neuronal activity in autism, and is one that needs further empirical clarification. Two main models have tried to account for a decreased neuronal connectivity, especially in the frontal regions (Catarino et al., 2013): the weak central coherence model and the executive dysfunction model. As regards cognitive deficits in autism, paradigms involving task-related regions of interest have pointed out working memory, executive function, face processing, motor function, and language. Also, along with analyses from resting-state

paradigms, researchers have identified a pattern of local overconnectivity but long-distance underconnectivity in autism (Anderson et al., 2010). In imaging studies, long-range white matter pathology in autism is now evident, with the corpus callosum's volume being one of the most reliable measures: decreased callosal size is one systematically shown structural finding in autism. Another reliable measure in searching for structural abnormalities in autistic disorders is the size of the genu (Mason et al., 2008).

Clinical features of the alien hand syndrome are abundantly described in the neuropsychological literature, mainly because of its dramatic appearance. Alien hand syndrome, often paraphrased as 'the anarchic hand,' is a neurological condition in which lesions in the anterior part of the corpus callosum produce involuntary movement, while the patient fails to distinguish the affected hand from the physician's hand when these are placed in the patient's unaffected hand (Bundick and Spinella, 2000). Long before Dr Strangelove and his cinematic intermanual conflict, Dr Kurt Goldstein described the first known alien hand case, where a right-handed woman complained of her left hand being estranged and belonging to another person, while it acted on its own (Goldstein, 1908). Later in the 1960s, Norman Geschwind was the first to applaud Goldstein's efforts in stressing "the non-unity of the personality in patients with callosal section and its possible psychiatric effects," in his classic *Brain* paper (Geschwind, 1965b, p. 638). Today, a widely accepted definition for the alien hand syndrome in current neurology requires "a feeling that one limb is foreign or 'has a will of its own,' *together with* observable involuntary motor activity" (Doody and Jankovic, 1992, p. 807). Three different varieties of the alien hand syndrome have been described, with the first two sharing a frontal variant and being easily dissociable – at least theoretically (Bundick and Spinella, 2000). The first one affects the dominant hand and is found in lesions of the language medial frontal cortex and the anterior corpus callosum, while the second affects the nondominant hand and is found after lesions of the corpus callosum alone (Feinberg et al., 1992; Goldberg and Bloom, 1990). A third variety, far less common than the previous two, is reported after posterior lesions, either in corticobasal degeneration or in cerebrovascular accidents, where the protagonist is usually the nondominant hand (Gibb et al., 1989; Ventura et al., 1995).

See also: Alzheimer's Disease, Neural Basis of; Attention, Neural Basis of; Attention-Deficit and Hyperactivity Disorders, Neural Basis of; Brain Asymmetry; Brain Damage: Neuropsychological Rehabilitation; Brain Development: Ontogenetic Neurobiology of; Cerebral Cortex; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Dysexecutive Syndromes; Functional Brain Imaging of Language processes; Language Processing, Functional Magnetic Resonance Imaging of; Lateralization of Language as Demonstrated by Brain Imaging Procedures; Neural Basis of Language Development; Neural Plasticity; Neural Representations of Objects; Prefrontal Cortex Development and Development of Cognitive Function; Split Brains and Human Consciousness.

Bibliography

Aboitiz, F., 1992. Brain connections: interhemispheric fiber systems and anatomical brain asymmetries in humans. Biological Research 25 (2), 51–61.

Aboitiz, F., Montiel, J., 2003. One hundred million years of interhemispheric communication: the history of the corpus callosum. Brazilian Journal of Medical and Biological Research 36 (4), 409–420.

Aboitiz, F., Scheibel, A.B., Zaidel, E., 1992. Morphometry of the Sylvian fissure and the corpus callosum, with emphasis on sex differences. Brain 115 (5), 1521–1541.

Aboitiz, F., Scheibel, A.B., Fisher, R.S., Zaidel, E., 1992a. Fiber composition of the human corpus callosum. Brain Research 598 (1), 143–153.

Aboitiz, F., Scheibel, A.B., Fisher, R.S., Zaidel, E., 1992b. Individual differences in brain asymmetries and fiber composition in the human corpus callosum. Brain Research 598 (1), 154–161.

Anderson, J.S., Lange, N., Froehlich, A., DuBray, M.B., Druzgal, T.J., Froimowitz, M.P., Alexander, A.L., Bigler, E.D., Lainhart, J.E., 2010. Decreased left posterior insular activity during auditory language in autism. American Journal of Neuroradiology 31 (1), 131–139.

Baird, A.A., Colvin, M.K., VanHorn, J.D., Inati, S., Gazzaniga, M.S., 2005. Functional connectivity: integrating behavioral, diffusion tensor imaging, and functional magnetic resonance imaging data sets. Journal of Cognitive Neuroscience 17 (4), 687–693.

Banich, M.T., 1995. Interhemispheric processing: theoretical considerations and empirical approaches. In: Davidson, R.J., Hugdahl, K. (Eds.), Brain Asymmetry. MIT Press, Cambridge, MA, pp. 427–450.

Banich, M.T., 1998. The missing link: the role of interhemispheric interaction in attentional processing. Brain and Cognition 36 (2), 128–157.

Banich, M.T., Belger, A., 1990. Interhemispheric interaction: how do the hemispheres divide and conquer the task? Cortex 26, 77–94.

Banich, M.T., Milham, M.P., Atchley, R., Cohen, N.J., Webb, A., Wszalek, T., Kramer, A.F., Liang, Z.-P., Wright, A., Shenker, J., Magin, R., 2000. fMRI studies of Stroop tasks reveal unique roles of anterior and posterior brain systems in attentional selection. Journal of Cognitive Neuroscience 12, 988–1000.

Baynes, K., 1990. Language and reading in the right hemisphere: highways or byways of the brain? Journal of Cognitive Neuroscience 2 (3), 159–179.

Baynes, K., Eliassen, J.C., 1998. The visual Lexicon: its access and organization in commissurotomy. In: Beeman, M., Chiarello, C. (Eds.), Right Hemisphere Language Comprehension: Perspectives from Cognitive Neuroscience. Erlbaum, Hillsdale, NJ, pp. 79–104.

Baynes, K., Gazzaniga, M.S., 1988. Right hemisphere language: insights into normal language mechanisms? In: Plum, F. (Ed.), Language, Communication, and the Brain. Raven Press, New York, pp. 117–126.

Baynes, K., Kroll, N., Dronkers, N., 1997. Contributions of the corpus callosum to semantic facilitation. Journal of the International Neuropsychological Society 3, 9.

Baynes, K., Wessinger, C.M., Fendrich, R., Gazzaniga, M.S., 1995. The emergence of the capacity to name left visual field stimuli in a callosotomy patient: implications for functional plasticity. Neuropsychologia 33 (10), 1225–1242.

Bloom, J.S., Hynd, G.W., 2005. The role of the corpus callosum in interhemispheric transfer of information: excitation or inhibition? Neuropsychology Review 15 (2), 59–71.

Braver, T.S., Cohen, J.D., Nystrom, L.E., Jonides, J., Smith, E.E., Noll, D.C., 1997. A parametric study of prefrontal cortex involvement in human working memory. Neuroimage 5 (1), 49–62.

Broca, P., 1861a. Perte de la parole, ramollissement chronique de destruction partielle du lobe antırieur gauche du cerveau. Bulletins de la Societe d'Anthropologie de Paris 1861, 235–238.

Broca, P., 1861b. Remarques sur le siege de la faculte du langage articule, suivies d'une observation d'aphemie (perte de la parole). Bulletins et memoires de la Societe Anatomique de Paris 36, 330–357.

Broca, P., 1861c. Nouvelle observation d'aphemie produite par une lesion de la moitie posterieure des deuxieme et troisieme circonvolutions frontales. Bulletins et memoires de la Societe Anatomique de Paris 36, 398–407.

Broca, P., 1863. Localisation des fonctions cerebrales: siege du langage articule. Bulletins de la Societe Anthropologique de Paris 1863, 200–204.

Buklina, S.B., 2005. The corpus callosum, interhemisphere interactions, and the function of the right hemisphere of the brain. Neuroscience and Behavioral Physiology 35 (5), 473–480.

Bullmore, E.T., Frangou, S., Murray, R.M., 1997. The dysplastic net hypothesis: an integration of developmental and dysconnectivity theories of schizophrenia. Schizophrenia Research 28 (2), 143–156.

Bundick, T., Spinella, M., 2000. Subjective experience, involuntary movement, and posterior alien hand syndrome. Journal of Neurology, Neurosurgery & Psychiatry 68 (1), 83–85.
Castelli, F., Frith, C., Happé, F., Frith, U., 2002. Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. Brain 125 (8), 1839–1849.
Catani, M., ffytche, D.H., 2005. The rises and falls of disconnection syndromes. Brain 128, 2224–2239.
Catarino, A., Andrade, A., Churches, O., Wagner, A.P., Baron-Cohen, S., Ring, H., 2013. Task-related functional connectivity in autism spectrum conditions: an EEG study using wavelet transform coherence. Molecular Autism 4 (1), 1.
Chang, S.E., Erickson, K.I., Ambrose, N.G., Hasegawa-Johnson, M.A., Ludlow, C.L., 2008. Brain anatomy differences in childhood stuttering. Neuroimage 39, 1333–1344.
Chiarello, C., 1995. Does the corpus callosum play a role in the activation and suppression of ambiguous word meanings? In: Kitterle, F.L. (Ed.), Hemispheric Communication: Mechanisms and Models. Erlbaum, Hillsdale, NJ, pp. 271–300.
Chiarello, C., Maxfield, L., Kahan, T., 1995. Initial right hemisphere activation of subordinate word meanings is not due to homotopic callosal inhibition. Psychonomic Bulletin & Review 2 (3), 375–380.
Choi, H., Kubicki, M., Whitford, T.J., Alvarado, J.L., Terry, D.P., Niznikiewicz, M., McCarley, R.W., Kwon, J.S., Shenton, M.E., 2011. Diffusion tensor imaging of anterior commissural fibers in patients with schizophrenia. Schizophrenia Research 130 (1), 78–85.
Clarke, J.M., Lufkin, R.B., Zaidel, E., 1993. Corpus callosum morphometry and dichotic listening performace: individual differences in functional interhemispheric inhibition? Neuropsychologia 31 (6), 547–557.
Clarke, J.M., Zaidel, E., 1994. Anatomical-behavioral relationships: corpus callosum morphometry and hemispheric specialization. Behavioural Brain Research 64, 185–202.
Conti, F., Manzoni, T., 1994. The neurotransmitters and postsynaptic actions of callosally projecting neurons. Behavioural Brain Research 64 (1), 37–53.
Cook, N.D., 1984. Callosal inhibition: the key to the brain code. Behavioral Science 29, 98–110.
Corbetta, M., Tansy, A.P., Stanley, C.M., Astafiev, S.V., Snyder, A.Z., Shulman, G.L., 2005. A functional MRI study of preparatory signals for spatial location and objects. Neuropsychologia 43, 2041–2056.
Coslett, H.B., 2003. Acquired dyslexia. In: Heilman, K.M., Valenstein, E. (Eds.), Clinical Neuropsychology, fourth ed. Oxford University Press, New York, pp. 108–125.
Crandall, P.H., 1985. Clinical phenomenology following hemispherectomy and the syndromes of hemispheric disconnection. In: Benson, D.F., Zaidel, E. (Eds.), The Dual Brain: Hemispheric Specialization in Humans. Guilford Press, New York, pp. 277–288.
Crow, T.J., 1998. Schizophrenia as a transcallosal misconnection syndrome. Schizophrenia Research 30 (2), 111–114.
Darby, D., Walsh, K.W., 2005. Walsh's Neuropsychology: A Clinical Approach. Churchill Livingstone, London.
Déjérine, J., 1892. Contibution a l'étude anatomo-pathologique et clinique des differentes variétés de cécité-verbale. Mém Soc Biol 4, 61–90.
Dennis, M., 1976. Impaired sensory and motor differentiation with corpus callosum agenesis: a lack of callosal inhibition during ontogeny? Neuropsychologia 14 (4), 455–469.
Doody, R.S., Jankovic, J., 1992. The alien hand and related signs. Journal of Neurology, Neurosurgery & Psychiatry 55 (9), 806–810.
Dorion, A.A., Chantôme, M., Hasboun, D., Zouaoui, A., Marsault, C., Capron, C., Duyme, M., 2000. Hemispheric asymmetry and corpus callosum morphometry: a magnetic resonance imaging study. Neuroscience Research 36 (1), 9–13.
Doron, K.W., Gazzaniga, M.S., 2008. Neuroimaging techniques offer new perspectives on callosal transfer and interhemispheric communication. Cortex 44, 1023–1029.
Dougherty, R.F., Ben-Shachar, M., Deutsch, D.K., Hernandez, A., Fox, G.R., Wandell, B.A., 2007. Temporal-callosal pathway diffusivity predicts phonological skills in children. Proceedings of the National Academy of Sciences USA 104 (20), 8556–8561.
Duara, R., Kushch, A., Gross-Glenn, K., Barker, W., Jallad, B., Pascal, S., Loewenstein, D.A., Sheldon, J., Rabin, M., Levin, B., Lubs, H., 1991. Neuroanatomic differences between dyslexic and normal readers on magnetic resonance imaging scans. Archives of Neurology 48 (4), 410–416.
Feinberg, T.E., Schindler, R.J., Flanagan, N.G., Haber, L.D., 1992. Two alien hand syndromes. Neurology 42 (1), 19.
Fling, B.W., Peltier, S.J., Bo, J., Welsh, R.C., Seidler, R.D., 2011. Age differences in interhemispheric interactions: callosal structure, physiological function, and behavior. Frontiers in Neuroscience 5, 38.
Franz, E.A., Waldie, K.E., Smith, M.J., 2000. The effect of callosotomy on novel versus familiar bimanual actions: a neural dissociation between controlled and automatic processes? Psychological Science 11 (1), 82–85.
Friederici, A.D., 2011. The brain basis of language processing: from structure to function. Physiological Reviews 91 (4), 1357–1392.
Friederici, A.D., von Cramon, Y.D., Kotz, S.A., 2007. Role of the corpus callosum in speech comprehension: interfacing syntax and prosody. Neuron 53, 135–145.
Friston, K.J., 1998. The disconnection hypothesis. Schizophrenia Research 30 (2), 115–125.
Friston, K.J., Frith, C.D., 1995. Schizophrenia: a disconnection syndrome. Clinical Neuroscience 3 (2), 89–97.
Galaburda, A.M., 1984. Anatomical asymmetries. In: Galaburda, A.M., Geschwind, N. (Eds.), Cerebral Dominance. Harvard Press, Cambridge, MA, pp. 11–25.
Galaburda, A.M., 1995. Anatomic basis of cerebral dominance. In: Davidson, R.J., Hugdahl, K. (Eds.), Brain Asymmetry. MIT Press, Cambridge, MA, pp. 51–74.
Galaburda, A.M., Sherman, G.F., Rosen, G.D., Aboitiz, F., Geschwind, N., 1990. Developmental dyslexia: four consecutive patients with cortical abnormalities. Annals of Neurology 18, 222–233.
Gazzaniga, M.S., 1970. The Bisected Brain. Appleton-Century-Crofts, New York.
Gazzaniga, M.S., 1983. Right hemisphere language following brain bisection: a 20-year perspective. American Psychologist 38 (5), 525.
Gazzaniga, M.S., 2000. Cerebral specialization and interhemispheric communication does the corpus callosum enable the human condition? Brain 123 (7), 1293–1326.
Gazzaniga, M.S., Bogen, J.E., Sperry, R.W., 1963. Laterality effects in somesthesis following cerebral commissurotomy in man. Neuropsychologia 1 (3), 209–215.
Gazzaniga, M.S., Miller, G.A., 1989. The recognition of antonymy by a language-enriched right hemisphere. Journal of Cognitive Neuroscience 1 (2), 187–193.
Gazzaniga, M.S., Nass, R., Reeves, A., Roberts, D., 1984. Neurologic perspectives on right hemisphere language following surgical section of the corpus callosum. Seminars in Neurology 4 (2), 126–135.
Gazzaniga, M.S., Sperry, R.W., 1967. Language after section of the cerebral commissures. Brain 90 (1), 131–148.
Gazzaniga, M.S., Volpe, B.T., Smylie, C.S., Wilson, D.H., Le Doux, J.E., 1979. Plasticity in speech organization following commissurotomy. Brain 102 (4), 805–815.
Gerloff, C., Andres, F.G., 2002. Bimanual coordination and interhemispheric interaction. Acta Psychologica 110 (2), 161–186.
Geschwind, N., 1965a. Disconnexion syndromes in animals and man. I. Brain 88, 237–294.
Geschwind, N., 1965b. Disconnexion syndromes in animals and man. II. Brain 88, 585–644.
Geschwind, N., Galaburda, A.M., 1985. Cerebral lateralization: biological mechanisms, associations, and pathology: I. A hypothesis and a program for research. Archives of Neurology 42 (5), 428.
Gibb, W.R.G., Luthert, P.J., Marsden, C.D., 1989. Corticobasal degeneration. Brain 112 (5), 1171–1192.
Goldberg, G., Bloom, K.K., 1990. The alien hand sign: localization, lateralization and recovery. American Journal of Physical Medicine & Rehabilitation 69 (5), 228–238.
Goldstein, K., 1908. Zur Lehre von der motorischen Apraxie. Journal für Psychologie und Neurologie X (I), 169–187.
Guo, S., Kendrick, K.M., Yu, R., Wang, H.L.S., Feng, J., 2014. Key functional circuitry altered in schizophrenia involves parietal regions associated with sense of self. Human Brain Mapping 35 (1), 123–139.
Guo, W., Liu, F., Dai, Y., Jiang, M., Zhang, J., Yu, L., Long, L., Chen, H., Gao, Q., Xiao, C., 2013. Decreased interhemispheric resting-state functional connectivity in first-episode, drug-naive major depressive disorder. Progress in Neuro-Psychopharmacology and Biological Psychiatry 41, 24–29.
Harper, C.G., Kril, J.J., 1988. Corpus callosal thickness in alcoholics. British Journal of Addiction 83, 577–580.
Harper, C.G., Kril, J.J., 1990. Neuropathology of alcoholism. Alcohol and Alcoholism 25, 207–216.
Harper, C., Matsumoto, I., 2005. Ethanol and brain damage. Current Opinion in Pharmacology 5, 73–78.
Harris, L.J., 1995. The corpus callosum and hemispheric communication: an historical survey of theory and research. In: Kitterele, F.L. (Ed.), Hemispheric Communication: Mechanisms and Models. Erlbaum, Hillsdale, NJ, pp. 1–59.
Hellige, J.B., 1993. Hemispheric Asymmetry: What's Right and What's Left. Harvard University Press, Cambridge, MA.
Hellige, J.B., Taylor, A.K., Eng, T.L., 1989. Interhemispheric interaction when both hemispheres have access to the same stimulus information. Journal of Experimental Psychology: Human Perception and Performance 15 (4), 711.

Hopkins, W.D., Cantalupo, C., 2008. Theoretical speculations on the evolutionary origins of hemispheric specialization. Current Directions in Psychological Science 17 (3), 233–237.

Hopkins, W.D., Rilling, J.K., 2000. A comparative MRI study of the relationship between neuroanatomical asymmetry and interhemispheric connectivity in primates: implication for the evolution of functional asymmetries. Behavioral Neuroscience 114 (4), 739.

Hynd, G.W., Hall, J., Novey, E.S., Eliopulos, D., Black, K., Gonzalez, J.J., Edmonds, J.E., Riccio, C., Cohen, M., 1995. Dyslexia and corpus callosum morphology. Archives of Neurology 52, 32–38.

Johansen-Berg, H., Della-Maggiore, V., Behrens, T.E., Smith, S.M., Paus, T., 2007. Integrity of white matter in the corpus callosum correlates with bimanual co-ordination skills. Neuroimage 36, T16–T21.

Joseph, R.N., Bannister, C.M., 2001. Problems with interhemispheric transfer of information in complete or partial agenesis of the corpus callosum. Neurorehabilitation and Neural Repair 15 (3), 197–202.

Jung-Beeman, M., 2005. Bilateral brain processes for comprehending natural language. Trends in Cognitive Sciences 9 (11), 512–518.

Just, M.A., Cherkassky, V.L., Keller, T.A., Minshew, N.J., 2004. Cortical activation and synchronization during sentence comprehension in high-functioning autism: evidence of underconnectivity. Brain 127 (8), 1811–1821.

Kawaguchi, Y., 1992. Receptor subtypes involved in callosally-induced postsynaptic potentials in rat frontal agranular cortex in vitro. Experimental Brain Research 88 (1), 33–40.

Kinsbourne, M., 1975. The mechanism of hemispheric control of the lateral gradient of attention. In: Rabbit, P.M., Dornic, S. (Eds.), Attention and Performance. Academic Press, New York, pp. 81–97.

Kinsbourne, M., 1977. Hemi-neglect and hemisphere rivalry. Advances in Neurology 18, 41–49.

Kinsbourne, M., Bruce, R., 1987. Shift in visual laterality within blocks of trials. Acta Psychologica 66, 139–155.

Klingberg, T., Hedehus, M., Temple, E., Salz, T., Gabrieli, J.D.E., Moseley, M.E., Poldrack, R.A., 2000. Microstructure of temporo-parietal white matter as a basis for reading ability. Evidence from diffusion tensor magnetic resonance imaging. Neuron 25, 493–500.

Knyazeva, M.G., 2013. Splenium of Corpus Callosum: Patterns of Interhemispheric Interaction in Children and Adults. Neural Plasticity.

van der Knaap, L.J., van der Ham, I.J., 2011. How does the corpus callosum mediate interhemispheric transfer? A review. Behavioural Brain Research 223 (1), 211–221.

Koshino, H., Kana, R.K., Keller, T.A., Cherkassky, V.L., Minshew, N.J., Just, M.A., 2008. fMRI investigation of working memory for faces in autism: visual coding and underconnectivity with frontal areas. Cerebral Cortex 18 (2), 289–300.

Koshino, H., Carpenter, P.A., Minshew, N.J., Cherkassky, V.L., Keller, T.A., Just, M.A., 2005. Functional connectivity in an fMRI working memory task in high-functioning autism. Neuroimage 24 (3), 810–821.

Langan, J., Peltier, S.J., Bo, J., Fling, B.W., Welsh, R.C., Seidler, R.D., 2010. Functional implications of age differences in motor system connectivity. Frontiers in Systems Neuroscience 4, 17.

Larsen, J.P., Hoien, T., Lundberg, I., Odegaard, H., 1990. MRI evaluation of the size and symmetry of the planum temporale in adolescents with developmental dyslexia. Brain Language 39, 289–301.

Lassonde, M., 1986. The facilitatory influence of the corpus callosum on intra-hemispheric processing. In: Lepore, F., Ptito, M., Jasper, H.H. (Eds.), Two Hemispheres—one Brain: Functions of the Corpus Callosum. Alan Liss, New York, pp. 385–402.

Liederman, J., 1998. The dynamics of interhemispheric collaboration and hemispheric control. Brain and Cognition 36 (2), 193–208.

Mason, R.A., Williams, D.L., Kana, R.K., Minshew, N., Just, M.A., 2008. Theory of mind disruption and recruitment of the right hemisphere during narrative comprehension in autism. Neuropsychologia 46 (1), 269–280.

McCulloch, W.S., Garol, H.W., 1941. Cortical origin and distribution of corpus callosum and anterior commissure in the monkey (Macaca mulatta). Journal of Neurophysiology 4, 555–563.

McNally, M.A., Crocetti, D., Mahone, E.M., Denckla, M.B., Suskauer, S.J., Mostofsky, S.H., 2010. Corpus callosum segment circumference is associated with response control in children with attention-deficit hyperactivity disorder (ADHD). Journal of Child Neurology 25 (4), 453–462.

Müller-Oehring, E.M., Schulte, T., Fama, R., Pfefferbaum, A., Sullivan, E.V., 2009. Global-local interference is related to callosal compromise in alcoholism: a behavior-DTI association study. Alcoholism: Clinical and Experimental Research 33, 477–489.

Müller-Oehring, E.M., Schulte, T., Raassi, C., Pfefferbaum, A., Sullivan, E.V., 2007. Local-global interference is modulated by age, sex and anterior corpus callosum size. Brain Research 1142, 189–205.

Müller-Oehring, E.M., Schulte, T., Rosenbloom, M.J., Pfefferbaum, A., Sullivan, E.V., 2010. Callosal degradation in HIV-1 infection predicts hierarchical perception: a DTI study. Neuropsychologia 48, 1133–1143.

Murray, C., Viehman, A., Lippa, C.F., 2006. The corpus callosum in Pick's disease, Alzheimer's disease, and amyotrophic lateral sclerosis: Gliosis implies possible clinical consequence. American journal of Alzheimer's disease and other dementias 21 (1), 37–43.

Nieuwenhuys, R., Voogd, J., van Huijzen, C., 2008. The Human Central Nervous System: A Synopsis and Atlas, fourth ed. Springer, Berlin.

Pfefferbaum, A., Adalsteinsson, E., Sullivan, E.V., 2006. Dysmorphology and microstructural degradation of the corpus callosum: interaction of age and alcoholism. Neurobiology of Aging 27, 994–1009.

Pfefferbaum, A., Sullivan, E.V., 2002. Microstructural but not macrostructural disruption of white matter in women with chronic alcoholism. Neuroimage 15, 708–718.

Pfefferbaum, A., Sullivan, E.V., 2005. Disruption of brain white matter microstructure by excessive intracellular and extracellular fluid in alcoholism: evidence from diffusion tensor imaging. Neuropsychopharmacology 30, 423–432.

Phelps, E.A., Hirst, W., Gazzaniga, M.S., 1991. Deficits in recall following partial and complete commissurotomy. Cerebral Cortex 1 (6), 491–492.

von Plessen, K., Lundervold, A., Duta, N., Heiervang, E., Klauschen, F., Smievoll, A.I., Ersland, L., Hugdahl, K., 2002. Less developed corpus callosum in dyslexic subjects—a structural MRI study. Neuropsychologia 40 (7), 1035–1044.

Preis, S., Steinmetz, H., Knorr, U., Jäncke, L., 2000. Corpus callosum size in children with developmental language disorder. Cognitive Brain Research 10 (1), 37–44.

Putnam, M.C., Wig, G.S., Grafton, S.T., Kelley, W.M., Gazzaniga, M.S., 2008. Structural organization of the corpus callosum predicts the extent and impact of cortical activity in the nondominant hemisphere. Journal of Neuroscience 28, 2912–2918.

Rakic, P., Yakovlev, P.I., 1968. Development of the corpus callosum and cavum septi in man. Journal of Comparative Neurology 132, 45–72.

Raybaud, C., 2010. The corpus callosum, the other great forebrain commissures, and the septum pellucidum: anatomy, development, and malformation. Neuroradiology 52 (6), 447–477.

Reuter-Lorenz, P.A., Baynes, K., 1992. Modes of lexical access in the callosotomized brain. Journal of Cognitive Neuroscience 4 (2), 155–164.

Reuter-Lorenz, P.A., Kinsbourne, M., Moscovitch, M., 1990. Hemispheric control of spatial attention. Brain and Cognition 12, 240–266.

Ringo, J., Doty, R., Demeter, S., Simard, P., 1994. Time is of the essence: a conjecture that hemispheric specialization arises from interhemispheric conduction delay. Cerebral Cortex 4, 331–343.

Roberts, D.W., 1999. Corpus callosotomy in the treatment of neuronal migration disorders. In: Kotagal, P., Luders, H.O. (Eds.), The Epilepsies: Etiologies and Prevention. Academic Press, San Diego, CA, pp. 103–112.

Roberts, R.E., Anderson, E.J., Husain, M., 2013. White matter microstructure and cognitive function. The Neuroscientist 19 (1), 8–15.

Robertson, L.C., Lamb, M.R., Zaidel, E., 1993. Callosal transfer and hemisphere laterality in response to hierarchical patterns: evidence from normal and commissurotomized subjects. Neuropsychology 7, 325–342.

Robichon, F., Bouchard, P., Démonet, J.F., Habib, M., 2000. Developmental dyslexia: re-evaluation of the corpus callosum in male adults. European Neurology 43 (4), 233–237.

Schnoebelen, S., Semrud-Clikeman, M., Pliszka, S.R., 2010. Corpus callosum anatomy in chronically treated and stimulant naive ADHD. Journal of Attention Disorders 14 (3), 256–266.

Schulte, T., Müller-Oehring, E.M., 2010. Contribution of callosal connections to the interhemispheric integration of visuomotor and cognitive processes. Neuropsychology Review 20 (2), 174–190.

Schulte, T., Müller-Oehring, E.M., Pfefferbaum, A., Sullivan, E.V., 2006. Callosal involvement in a lateralized Stroop task in alcoholic and healthy subjects. Neuropsychology 20, 727–736.

Schulte, T., Müller-Oehring, E.M., Strasburger, H., Warzel, H., Sabel, B.A., 2001. Acute effects of alcohol on divided and covert attention in men. Psychopharmacology 154, 61–69.

Sergent, J., 1983. Unified response to bilateral hemispheric stimulation by a split-brain patient. Nature 305, 800–802.

Sidtis, J.J., Volpe, B.T., Wilson, D.H., Rayport, M., Gazzaniga, M.S., 1981. Variability in right hemisphere language function after callosal section: evidence for a continuum of generative capacity. The Journal of Neuroscience 1 (3), 323–331.

Skiba, M., Diekamp, B., Prior, H., Güntürkün, O., 2000. Lateralized interhemispheric transfer of color cues: evidence for dynamic coding principles of visual lateralization in pigeons. Brain and Language 73, 254–273.

Smith, E.E., Jonides, J., Koeppe, R.A., 1996. Dissociating verbal and spatial working memory using PET. Cerebral Cortex 6 (1), 11–20.

Sommer, M., Koch, M.A., Paulus, W., Weiller, C., Büchel, C., 2002. Disconnection of speech-relevant brain areas in persistent developmental stuttering. Lancet 360, 380–383.

Sperry, R.W., Zaidel, E., Zaidel, D., 1979. Self recognition and social awareness in the deconnected minor hemisphere. Neuropsychologia 17 (2), 153–166.

Steinmann, S., Mulert, C., 2012. Functional relevance of interhemispheric fiber tracts in speech processing. Journal of Neurolinguistics 25 (1), 1–12.

Tarnowska-Dziduszko, E., Bertrand, E., Szpak, G.M., 1995. Morphological changes in the corpus callosum in chronic alcoholism. Folia Neuropathologica 33, 25–29.

Tesak, J., Code, C., 2008. Milestones in the History of Aphasia: Theories and Protagonists. Psychology Press, New York.

Urgesi, C., Bricolo, E., Aglioti, S.M., 2005. Hemispheric metacontrol and cerebral dominance in healthy individuals investigated by means of chimeric faces. Cognitive Brain Research 24 (3), 513–525.

Ventura, M.G., Goldman, S., Hildebrand, J., 1995. Alien hand syndrome without a corpus callosum lesion. Journal of Neurology, Neurosurgery & Psychiatry 58 (6), 735–737.

Wahl, M., Lauterbach-Soon, B., Hattingen, E., Jung, P., Singer, O., Volz, S., Klein, J.C., Steinmetz, H., Ziemann, U., 2007. Human motor corpus callosum: topography, somatotopy, and link between microstructure and function. The Journal of Neuroscience 27 (45), 12132–12138.

Welcome, S.E., Chiarello, C., 2008. How dynamic is interhemispheric interaction? Effects of task switching on the across-hemisphere advantage. Brain and Cognition 67 (1), 69–75.

Wernicke, C., 1906. Grundrisse der Psychiatrie. Thieme, Leipzig.

Yaldizli, Ö., Penner, I.K., Frontzek, K., Naegelin, Y., Amann, M., Papadopoulou, A., Sprenger, T., Kuhle, J., Calabrese, P., Radü, E.W., Kappos, L., Gass, A., August 19, 2013. The relationship between total and regional corpus callosum atrophy, cognitive impairment and fatigue in multiple sclerosis patients. Multiple Sclerosis Journal (published online ahead of print).

Yazgan, M.Y., Wexler, B.E., Kinsbourne, M., Peterson, B., Leckman, J.F., 1995. Functional significance of individual variations in callosal area. Neuropsychologia 33 (6), 769–779.

Zaidel, E., 1991. Language functions in the two hemispheres following complete cerebral commissurotomy and hemispherectomy. In: Bollek, F., Grafman, J. (Eds.), Handbook of Neuropsychology. Elsevier, Amsterdam, pp. 115–150.

Zaidel, E., Iacoboni, M., Pacual-Leone, A.P., 1998. The Role of Human Corpus Callosum in Sensory Motor Integration: Anatomy, Physiology, and Behavior, Individual Difference and Clinical Applications. Plenum Press, New York.

Zaidel, E., Iacoboni, M., Zaidel, D.W., Bogen, J.E., 2003. In: Heilman, K.M., Valenstein, E. (Eds.), Clinical Neuropsychology, fourth ed. Oxford University Press, New York, pp. 347–403.

Zaidel, E., Peters, A.M., 1981. Phonological encoding and ideographic reading by the disconnected right hemisphere: two case studies. Brain and Language 14 (2), 205–234.

Zaidel, E., Seibert, L., October 1997. Speech in the disconnected right hemisphere. Brain and Language 60 (1), 188–192.

Zaidel, D., Sperry, R.W., 1974. Memory impairment after commissurotomy in man. Brain 97 (1), 263–272.

Zarei, M., Johansen-Berg, H., Smith, S., Ciccarelli, O., Thompson, A.J., Matthews, P.M., 2006. Functional anatomy of interhemispheric cortical connections in the human brain. Journal of Anatomy 209 (3), 311–320.

Internal/External Frame of Reference Model

Herbert W Marsh, Australian Catholic University, Strathfield, NSW, Australia; King Saud University, Riyadh, Saudi Arabia; and University of Oxford, Oxford, UK
Jens Möller, University of Kiel, Kiel, Germany
Philip Parker, Australian Catholic University, Strathfield, NSW, Australia
Man K Xu, University of Cambridge, Cambridge, UK
Benjamin Nagengast, University of Tübingen, Tübingen, Germany
Reinhard Pekrun, University of Munich, Munich, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Self-concept is an important outcome variable that facilitates attainment of other desirable outcomes. Self-concept formation is based in part on the frames of reference or standards of comparison used to evaluate one's accomplishments. The internal/external frame of reference model posits dimensional comparisons such that accomplishments in one domain are used as a basis of comparison for evaluating accomplishments in other domains, along with the accomplishments of other individuals (social comparison). Based on extensive support in educational settings that generalizes across gender, age, and nationality, it explains why high achievement in one domain (e.g., mathematics) can have negative effects on self-concepts in another domain (verbal).

A critical advance in psychology has been the growing emphasis on a positive psychology that focuses on how healthy, normal, and exceptional individuals can get the most from life (e.g., Marsh, 2007; Marsh and Craven, 2006; Seligman and Csikszentmihalyi, 2000). Consistent with this emphasis, positive self-concept and its enhancement have been the focus of research and interventions in many disciplines of psychology and the social science more generally. Self-concept is also an important mediating factor that facilitates the attainment of other desirable outcomes, such as choice behavior, planning, persistence, and subsequent accomplishments. Individuals are likely to accomplish more and perform better if they feel competent in what they do, are confident, and feel positively about themselves. However, relations between self-concept and desirable outcomes are often reciprocal, each being both a cause and an effect of the other. Programs and institutions that undermine self-concept are also likely to have negative effects on accomplishments.

There has been substantial improvement in the quality of self-concept research because of a synergy between better measurement instruments, theoretical models, quantitative methodologies, and research designs. The cornerstone of this resurgence was the classic review article by Shavelson et al. (1976). They posited self-concept as a multidimensional hierarchical construct and their structural model became the blueprint for the development of future self-concept instruments (Marsh, 2007). Shavelson et al. posited different facets of academic self-concept (ASC) to be substantially correlated and to form a single higher-order ASC factor, consistent with relations between achievements in different school subjects. However, subsequent research revealed that math self-concept (MSC) and verbal self-concept (VSC) in particular were nearly uncorrelated, leading to the Marsh/Shavelson revision (Marsh and Shavelson, 1985) that posited two higher-order ASC factors (math/academic and verbal/academic). In this model, it was proposed that underlying the verbal–math ASC factors was a continuum of ASC factors ranging from VSC at one end to MSC at the other, with an ordering of academic self-concepts in other domains along this continuum (see **Figure 1**).

The Internal/External Frame of Reference Model: Theoretical Basis

Self-concept theory emphasizes that perceptions of the self cannot be adequately understood if the role of frames of reference is ignored (Marsh, 2007). The same objective characteristics and accomplishments can lead to disparate self-concepts, depending on the frames of reference or standards of comparison that individuals use to evaluate themselves, and these self-beliefs have important implications for future choices, performance, and behaviors. The two most frequently posited frames of reference are based on social and temporal comparisons (Möller, 2005; Möller et al., 2009, 2011); self-perceptions are based in part on how current accomplishments compare with our past performances (temporal comparisons) and how our accomplishments compare with those of others within our current context (social comparisons; e.g., classmates in our school or class). Particularly in educational settings, a growing body of research based on the internal/external frame of reference (I/E) model (Marsh, 1986) demonstrates that self-perceptions may also be the result of internal comparisons, in which accomplishments in one school subject can serve as a frame of reference for another school subject – hereafter referred to as 'dimensional' comparisons (e.g., Möller and Köller, 2001a; Möller and Marsh, 2013; Möller et al., 2009; see also Möller et al., in this volume).

Dimensional comparison research has largely revolved around the I/E model, which initially was developed to explain why MSC and VSC are almost uncorrelated, even though corresponding areas of academic achievement are substantially correlated (typically 0.5–0.8, depending on how achievement

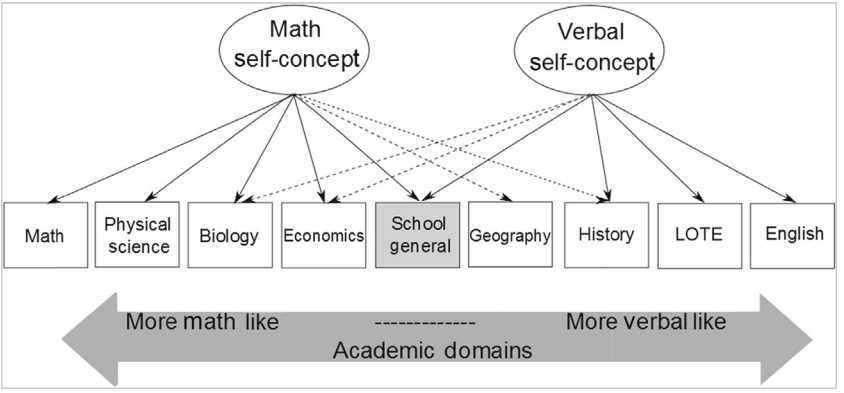

Figure 1 The Marsh/Shavelson revised academic self-concept model, which includes a wider variety of specific academic facets. Adapted with permission from Marsh, H.W., 2007. Self-concept Theory, Measurement and Research into Practice: The Role of self-concept in Educational Psychology. British Psychological Society, Leicester, UK.

is measured; for further discussion, see Marsh, 2007). People think of themselves as mostly math or mostly verbal persons, even though individuals who are good at one also tend to be good at the other (see Marsh, 1986). Why is there this stark contrast between academic self-perceptions and corresponding objective accomplishments? According to the I/E model, self-concept in a particular school subject is formed in relation to an external (social comparison) reference in which students compare their self-perceived performances in a particular school subject with the perceived performances of other students in the same school subject, and an internal (dimensional, ipsative comparison) reference in which students compare their own performances in the particular school subject with their own performances in other school subjects. Hence, students may have a favorable MSC if math is their best subject, even if they are not particularly good at math relative to other students. The joint operation of these processes, depending on the relative weight given to each, is consistent with the near-zero correlation between MSC and VSC that led to the revision of the Shavelson model (Marsh, 2007).

Tests of the I/E model typically are conducted when math and verbal achievements are regressed on MSC and VSC (see Figure 2). The external comparison process predicts that good math skills lead to higher MSCs and that good verbal skills lead to higher VSCs. According to the internal dimensional comparison process, however, good math skills lead to lower VSCs once the positive effects of good verbal skills are

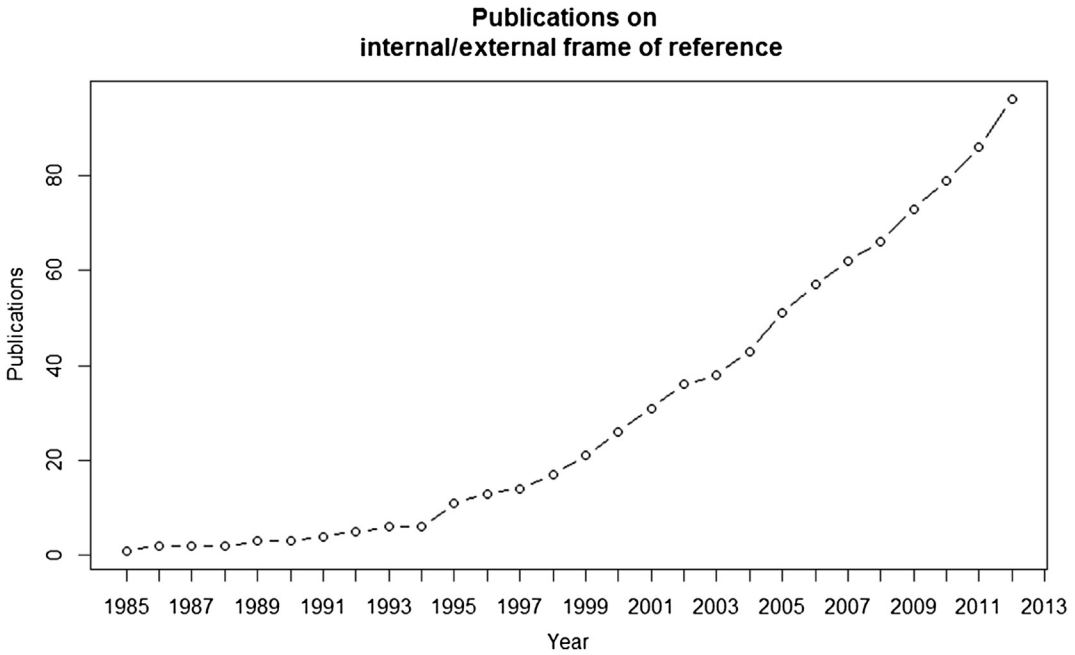

Figure 2 Cumulative references to the self-concept 'internal/external frame of reference' from Google Scholar. Only those publications with at least one other citation are included.

controlled for: The better I am at mathematics, the poorer I am at verbal subjects, relative to my good math skills. Similarly, better verbal skills lead to lower MSCs once the positive effects of good math skills are controlled for. In models used to test this prediction (see Figure 2), the horizontal paths leading from math achievement to MSC and from verbal achievement to VSC (the gray horizontal lines in Figure 2) are predicted to be substantially positive (indicated as ++ in Figure 2). However, the cross paths leading from math achievement to VSC and from verbal achievement to MSC (the dark lines in Figure 2) are predicted to be negative.

Empirical Support for the I/E Model

From the first publication in 1986, I/E research has developed into a mature and growing research area. Indeed, Google Scholar indicates that there have been more publications on this topic in the last 8 years than in all previous years combined, with 31% of all publications occurring in the last 5 years (see Figure 3). Marsh (1986) first proposed the I/E model and summarized the results of 13 studies of the relations between math and verbal achievements and self-concepts in the corresponding domains. Although there were consistently high correlations between math and verbal achievements (0.42–0.94), there were weak or even negative correlations between MSCs and VSCs (−0.10 to 0.19). In the I/E path diagram (Figure 2), the (horizontal) paths from math achievement to MSC and from verbal achievement to VSC were substantial and positive. However, the (cross) paths from math achievement to VSC and from verbal achievement to MSC were significant and negative. Subsequent research (see reviews by Marsh, 2007; Möller et al., 2009) provides strong support for the generalizability of these findings.

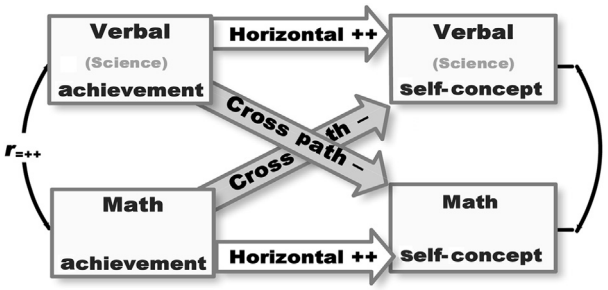

Figure 3 The 'classic' internal/external frame of reference (I/E) model relating verbal and math achievements to verbal and math self-concepts. According to predictions from the I/E model, the horizontal paths from achievement to self-concept in the matching domains (content area) are predicted to be substantial and positive (++), whereas the cross paths from achievement in one domain area to self-concept in a nonmatching domain (shaded in gray above) are predicted to be negative (−). In the present investigation, we evaluate the generalizability of these predictions to the science domain, relating science and math achievements to science and math self-concepts. Adapted with permission from Marsh, H.W., 2007. Self-concept Theory, Measurement and Research into Practice: The Role of Self-concept in Educational Psychology. British Psychological Society, Leicester, UK.

Importantly, the I/E framework has recently been extended to predict long-term educational choices. For example, Nagy et al. (2008) found that math achievement, task value, and self-concept have a positive effect on choosing math-related advanced courses but a negative effect on choosing verbal related advanced courses. Parker et al. (2012) extended this research, showing that math and verbal achievement and self-concept positively predicted final school year matriculation results and university entry. However, math achievement and self-concept positively predicted entry into university majors in math and the physical sciences but negatively predicted entry into the humanities. Such research suggests that dimensional comparison processes associated with the I/E model mean that young people use a profile of their perceived strengths and weaknesses in order to make educational decisions.

Most I/E studies (see Marsh, 2007) are based on cross-sectional or longitudinal and quasiexperimental studies. However, Möller and colleagues (e.g., Möller and Köller, 2001a; also see Möller and Marsh, 2014; Pohlmann and Möller, 2009) found support for the I/E model in true experimental studies with randomly assigned students, illustrating how experimentally manipulated feedback on achievement in one subject area had an inverse effect on self-concept in a different area. In a series of studies, Möller and Köller (2001a; Pohlmann and Möller, 2009) independently manipulated the external comparison process based on performance feedback relative to other students, and the internal comparison process based on feedback relative to performance by the same student on achievement on two subject-specific tasks. In each study, dimensional comparison information had the expected effects (e.g., positive math feedback enhanced MSC and weakened VSC). These true experimental studies provided strong support for both the dimensional and external comparison processes posited in the I/E model, and the causal interpretations of the two processes. Examining the spontaneous occurrence of dimensional comparisons with two introspective diary studies, Möller and Husemann (2006) confirmed that students spontaneously carry out dimensional comparisons in everyday life, with contrast effects from one domain to self-evaluations and emotions in the other (for more details on dimensional comparison processes, see Möller et al., this volume).

Meta-Analysis of I/E Studies

Möller et al. (2009) conducted a meta-analysis of 69 data sets ($n = 125\,308$) for studies that simultaneously evaluated the effects of math and verbal achievements on MSC and VSC. Across all studies, math and verbal achievements were highly correlated (0.67), but self-concepts were nearly uncorrelated (0.10). Path coefficients from achievement to ASC in the matching domains were positive (0.61 for math, 0.49 for verbal) but negative from math achievement to VSC (−0.21) and verbal achievement to MSC (−0.27). Support for the I/E predictions generalized across age group, gender, and country, leading the authors to conclude, "The results of our meta-analyses indicate that the relations described in the classical I/E model are not restricted to a particular

achievement or self-concept measure or to specific age groups, gender groups, or countries" (p. 1157).

A particular strength of meta-analysis is the ability to evaluate the generalizability of the results, but the strength of these tests depends upon the available data. Möller et al. found good support for the generalizability of the I/E model in relation to age, gender, and country of residence. However, there was significant study-to-study variation in the sizes of relations among I/E variables, suggesting that the results did not generalize completely. Remarkably, they found no significant differences for country of residence or age. However, most of the samples were from Western and several Asian countries (7 Asian, 12 Australian, 7 USA, and 39 European samples were included in the analysis of country as a moderator). Although the meta-analysis included responses by students from 'other nationalities,' the authors argued that the number of studies from these 'other' countries was too small to be included in analyses of country as a moderator. Similarly, it is surprising that Möller et al. found no differences as a function of age, given developmental models of how self-concept formation varies as a function of age (e.g., Marsh, 1989, 2007; Wigfield and Eccles, 2002). However, because of the paucity of available studies with young children (only 3 of 69 samples reported results for children in grade 4 or younger) the generalizability of this finding was not strong. Subsequent research focused on these apparent limitations in existing research on cross-cultural and developmental generalizability.

Cross-Cultural and Developmental Generalizability

Cross-cultural research seeks to test the cross-cultural generalizability of a theory, to establish its scope and identify cultural boundary conditions (Marsh and Hau, 2004). The I/E model has been validated in more than 35 countries including many non-English-speaking Western, Asian, and Islamic countries (Marsh et al., 2014; Marsh and Hau, 2004; Möller et al., 2009). Marsh and Hau (2004) provided persuasive support for the cross-cultural generalizability of the I/E model, based on the PISA 2000 database: nationally representative samples of 15-year-olds from 26 countries (total $N = 55\,577$). As predicted, the two horizontal paths relating math achievement to MSC (0.44) and reading achievement to VSC (0.47) were substantial and positive, whereas the two cross paths leading from verbal achievement to MSC (−0.20) and mathematics achievement to VSC (−0.26) were negative. Although the correlation between math and verbal achievement factors ($r = 0.78$) was very large, the corresponding correlation between MSC and VSC factors ($r = 0.10$) was substantially lower. In multigroup confirmatory factor analysis and structural equation modeling tests of invariance, even the highly restrictive model of total invariance (i.e., requiring every parameter to be the same in all 26 countries) provided a good fit to the data that differed only slightly from the baseline model with no invariance constraints. These results support the cross-cultural generalizability of the measures and the relations among them across these 26 countries.

As part of their meta-analysis of I/E studies, Möller et al. (2009) demonstrated that results based on the 26 PISA countries generalized well to those from the remaining samples. Noting the complementary strengths and weaknesses associated both with meta-analyses such as that of Möller et al., and with large, comprehensive primary studies, such as the Marsh and Hau (2004) study using PISA data, Möller et al. concluded "the juxtaposition of this meta-analysis with the large PISA primary study is particularly important. The similarity in the pattern of results provides important support for the I/E model" (p. 1160). Clearly, the results of this meta-analysis, in combination with previous research, provide very strong support for the I/E model. Nevertheless, because the PISA data were based on 15-year-old students from mostly Western and several Asian countries, neither the meta-analysis nor the PISA study provided strong tests of the developmental generalizability of the I/E model or even of cross-cultural generalizability beyond Western and Asian countries.

These concerns were addressed in part by the Marsh et al. (2014) study of math and science constructs, based on the Trends in International Mathematics and Science Study (TIMSS) 2007 data. This demonstrated support for the generalizability of I/E predictions in relation to math and science domains, intrinsic motivation, and self-concept. The findings, based on nationally representative matched samples of fourth and eighth grade students from 13 diverse countries, were largely generalizable over age and nationality. Averaged across all 26 (13 countries × 2 age cohorts) groups, horizontal paths from matching achievement domains were all significantly positive (.524, standard error (SE) = 0.005), while cross paths were all significantly negative (−0.235, standard error (SE) = 0.005). Consistent with a priori predictions based on developmental theory, the positive effects of achievement on self-concept and motivation in the matching domain became stronger with age. However, the critical negative effects of achievement in one academic domain on self-concept and motivation in another domain also increased across age cohorts, suggesting that with age these subject domains become more distinct. There were also substantial country differences, in which support for I/E predictions was weaker in three Middle Eastern Islamic countries compared to five Western and 4 Asian countries.

Extensions of I/E Studies: Assimilation and Contrast Effects

The majority of I/E studies have been conducted in relation to the math and verbal domains. However, the theoretical underpinning of the I/E is based on a continuum of academic domains that vary between the end points of mathematical and verbal subjects (**Figure 1**) based on a higher-order factor analysis in which lower-order ASC factors can be explained in terms of two higher-order (math and verbal) academic factors. Similarity between different academic domains in this conceptual framework is based on how close they are to each other on this underlying continuum – operationalized as correlations among the different domains that were the basis of the higher-order factor structure. There is consistent support for negative effects based on comparisons of maximally dissimilar dimensions at the end points of the academic continuum (e.g., the negative cross paths in

Figure 2), referred to as contrast effects in the broader social comparison literature (see Möller and Marsh, 2013).

The theoretical underpinning of the I/E model posits that the negative contrast effects will diminish and might even become positive assimilation effects when dimensional comparisons are based on domains that are close together on the academic continuum. Thus, for example, Möller et al. (2006) found that math achievement had a positive effect on physics self-concepts. Similarly, recent research (Parker et al., 2013) based on PISA data indicates that – consistent with the I/E model – paths leading to MSC (the only self-concept domain assessed) were positive for math achievement and negative for verbal achievement. Critically, the path from science achievement (a domain close to math on the underlying academic continuum) to MSC was also positive; science achievement was positively predictive of MSC, even after controlling for science (and verbal) achievement: an assimilation effect. However, the TIMSS I/E study (Marsh et al., 2014) tested the I/E model more fully in relation to math and science constructs (achievement, aspirations, self-concept, and positive affect in each domain); there was clear evidence of contrast rather than assimilation effects.

They had however, considered a generic science scale that ignored potentially distinct subdomains of science self-concept. Thus, for example, in their German sample, where physics was taught as a separate subject, Möller and Husemann (2006) found small positive effects of math achievement on physics self-concept and of physics achievement on MSC. We also note that contrasting contextual effects might be reinforced by the simultaneous presentation of items from only two subject domains, which could lead to an implicit demand on students to contrast themselves in relation to the two subjects. Clearly there is need for further research as to how robust the correlations are as an index of domain similarity, and the conditions under which apparently elusive assimilation effects are found.

Other researchers have focused on the generalizability of I/E effects when the focus is on two verbal domains (native language and nonnative language), typically in combination with the math domain. Although these studies (e.g., Xu et al., 2013) typically have found contrast effects (negative cross paths) from both verbal domains to MSC, the results have not been entirely consistent for the two verbal domains, suggesting that neither assimilation nor contrast effects were particularly strong.

Marsh et al. (2015) offered new predictions, theoretical insights, and methodology associated with dimensional comparison theory (DCT; Möller and Marsh, 2013) based on multiple academic domains (native language, foreign language, history, biology, physics, and math). DCT posits that the negative contrast effects based on maximally different (far) comparisons will diminish and may even become positive (assimilation) effects when dimensional comparisons are based on complementary domains that are close together on the academic continuum (near comparisons). In the first test of new DCT theoretical predictions based on such a comprehensive range of academic domains, they provided clear support for the critical prediction that paths for achievement to ASC based on near comparisons will be less negative than those based on far comparisons.

Extensions of I/E Studies: Integration with Other ASC Models

In addition to the I/E model, other theoretical models of relations between ASC and achievement have dominated recent ASC research: the reciprocal effects model (REM) and the big-fish–little-pond effect (BFLPE), described below. Although both these models have much theoretical and empirical support, they are traditionally based on responses to a single domain of academic self-concept. However, some work has been done toward integrating them with the I/E model.

Reciprocal Effects Model

Although the research reviewed here demonstrates that ASC and academic achievement are substantially correlated, a critical question is the causal ordering of these two constructs. Traditional approaches to this question took an 'either-or' approach – either achievement causes ASC (a skill-development model) or ASC causes achievement (a self-enhancement model). However, Marsh and Craven (2006) argued theoretically and empirically for a dynamic REM integrating the skill-development and self-enhancement models, in which both ASC and achievement are causes and effects of each other (see Figure 4). Support for the REM requires that both sets of paths are statistically significant, but the linkages from prior self-concept to subsequent achievement are particularly important. There now exists substantial support for the generalizability of the findings over age, nationality, different measures of self-concept and achievement (Marsh and Craven, 2006). However, of critical importance to this article, nearly all REM research is based on a single domain of ASC: typically MSC, VSC, or global ASC, while I/E studies typically are based on measures collected on a single occasion.

Based on longitudinal measures of achievement and ASC in both math and verbal domains, Möller et al. (2011) proposed the Reciprocal Internal/External frame of reference model (RI/EM), integrating the REM and I/E models. Using three waves of data, they found good support for both REM and I/E models when each was evaluated separately. Consistent with RI/EM predictions, they found positive longitudinal effects of grades and ASCs on subsequent grades and ASCs within domains, and negative effects of grades on subsequent ASCs across domains. A unique, new feature of the RI/EM is the effect on ASC in one domain on subsequent achievement in a different domain. Although these effects were also negative, the effects were small and not always statistically significant.

Big-Fish–Little-Pond Effect

Marsh (see review by Marsh et al., 2008) proposed the BFLPE to encapsulate frame of reference effects that are based on an integration of theoretical models (e.g., relative deprivation theory, social comparison theory, psychophysical judgment, social judgment). In the BFLPE model, students are hypothesized to compare their abilities with the abilities of their classmates and to use this social comparison impression as one basis for forming their own self-concept (Figure 5); individual ability is positively related to ASC (the brighter I am the higher my ASC) but school-average ability has a negative effect on ASC

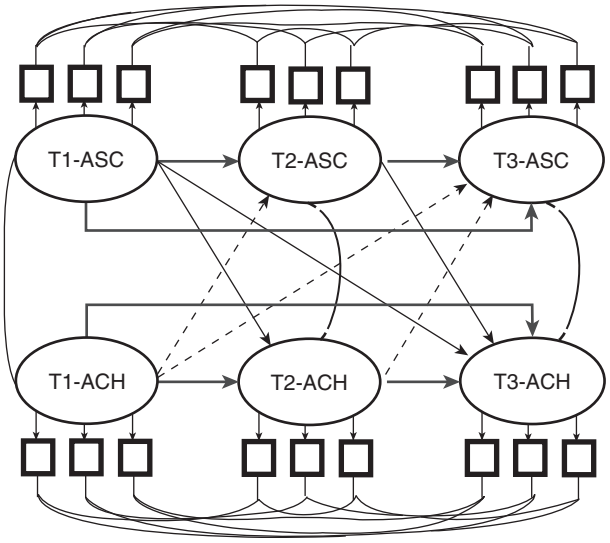

Figure 4 Prototype causal-ordering model for testing self-enhancement, skill-development, and reciprocal effects models. In this full-forward, multiwave, multivariable model, multiple indicators of academic self-concept (ASC) and achievement (ACH) are collected in three successive waves (T1, T2, and T3). Each latent construct (represented by oval shape) has paths leading to all latent constructs in subsequent waves. Within each wave, ASC and ACH are assumed to be correlated; in the first wave, this correlation is a covariance between two latent constructs, and in subsequent waves, it is a covariance between residual factors. Curved lines at the top and bottom of the figure reflect correlated uniquenesses between responses to the same measured variable (represented by boxes) collected on different occasions. Paths connecting the same variable on multiple occasions reflect stability (the solid gray paths), but these coefficients typically differ from the corresponding test–retest correlations (which do not include the effects of other variables). Dashed lines reflect effects of prior ACH on subsequent self-concept, whereas solid black lines reflect the effects of prior self-concept on subsequent ACH. Adapted with permission from Marsh, H.W., 2007. Self-concept Theory, Measurement and Research into Practice: The Role of Self-concept in Educational Psychology. British Psychological Society, Leicester, UK.

(the brighter my classmates, the lower my ASC). Extensive support for the BFLPE generalizes across student groups, subject domains, ASC instruments, and cultures (see review by Marsh, 2007; Marsh et al., 2008). However, again nearly all of this research is based on a single domain of ASC: typically MSC, VSC, or global ASC.

Indeed, integration of these two theoretical models suggests that the negative BFLPE could be accompanied by a compensatory positive effect (the BFLP-CE) when both math and verbal domains are considered simultaneously. The I/E model suggests that processes that change self-concept in one domain have a compensatory effect on self-concept in another domain, leading to the predictions that: (1) school-average ability in math has a negative effect on MSC (the BFLPE) but a positive effect on VSC (the BFLP-CE); and (2) school-average verbal ability has a negative effect on VSC (BFLPE) but a positive effect on MSC. Marsh (1990) provided initial support for this integration of I/E and BFLPE models, showing that school-average math ability negatively predicted MSC but positively predicted VSC, and vice versa for English achievement. However, compensatory effects found by Marsh (1990) were small, and finding appropriate tests of the integrated BFLP-CE model has proved difficult, because school-average math and verbal achievements typically are so highly correlated (r's > 0.95) that school contexts in the two domains cannot be easily differentiated. The following suggestions by Marsh et al. (2015) capitalized on an important feature of the German educational system; some academically selective schools select students on the basis of overall achievement, but others have a subject-specific theme and select students who excel in a particular area, rather than overall. Consistent with theoretical predictions, there were BFLPEs for both types of selective schools, but the BFLP-CE, based on the I/E model, was only evident for the themed (magnet) schools.

Theoretical Extensions of the I/E Model

Denissen et al. (2007; Eccles, 2009; Parker et al., 2012) subsequently integrated the I/E model into the traditional

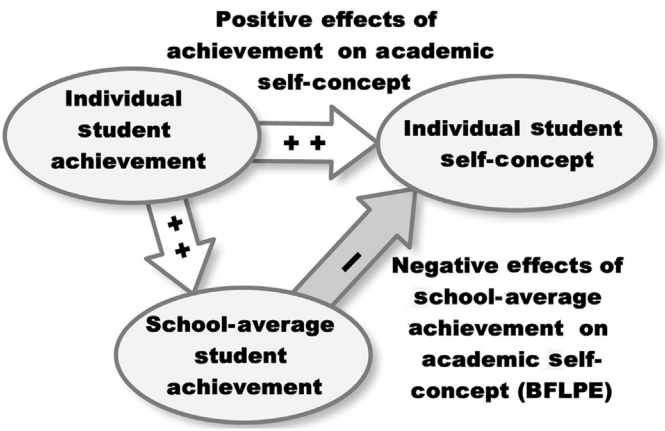

Figure 5 Path model predictions based on the big-fish–little-pond effect (BFLPE). Adapted with permission from Marsh, H.W., 2007. Self-concept Theory, Measurement and Research into Practice: The Role of Self-concept in Educational Psychology. British Psychological Society, Leicester, UK.

expectancy-value theory (EVT). EVT (Eccles, 2009; Eccles and Wigfield, 1995; Wigfield and Eccles, 2002) posits that expectancy (often operationalized as self-concept) and components of value (e.g., intrinsic, extrinsic, cost) predict a wide range of future outcomes. Although self-perceived competence was related to several different value constructs in the EVT model, the relations with intrinsic value and interest were consistently the strongest (Wigfield and Eccles, 2002). Whereas expectancy beliefs are closely associated with performance in both cross-sectional and longitudinal studies, interest and other value beliefs tend to be better predictors of choice, effort, and persistence at achievement-related activities (e.g., Marsh et al., 2005).

Emphasizing the link between EVT and the I/E model, Eccles (2009) argued "that both external and internal comparison processes are key – people assess their own skills by comparing their performances with those of other people and with their own performances across domains" (p. 82) and demonstrated how this internal dimensional comparison process has a critical influence on academic choice behavior (e.g., choice of university major; see Parker et al., 2012).

Pekrun and colleagues (Goetz et al., 2008; Pekrun, 2006) extended their control-value theory of achievement emotions to incorporate the I/E model. This was done to explain the extreme domain specificity of students' achievement emotions, based on their hypothesis that "the mechanisms addressed by Marsh's (1986) I/E model could operate for students' emotions as well" (Goetz et al., 2008, p. 7). They found that emotions were substantially more domain specific than achievement in different math and verbal subjects, and that enjoyment was the most domain specific emotion. Goetz et al. (2008) subsequently found support for the I/E model in relation to both self-concept and enjoyment, but showed that achievement/enjoyment relations were mediated by self-concept. Because enjoyment is a key component of intrinsic motivation, this research complements EVT studies showing the extreme domain specificity of intrinsic motivation.

Policy and Practice Implications

The I/E model also has implications for educational practice and for our understanding of self-concept formation. When asked to infer students' self-concepts, teachers, parents, and others imply that, consistent with corresponding measures of achievement, students who are bright in one area tend to be seen as having good ASCs in all areas, whereas students with lower ability in one area are seen as having poor ASCs in all areas. Similarly, teachers and parents may underestimate the domain specificity of affective variables that are influenced by ASC, such as students' academic emotions, interest, and motivation to learn. For example, teachers are likely to believe that students' enjoyment and anxiety generalize across school subjects. However, if teachers and others better understood the formation of self-concepts in different academic domains, then they would be able to understand their students better and provide more appropriate, credible feedback, particularly for less able students. Even bright students might have an average or below-average self-concept in their weakest school subjects; this may seem paradoxical, in relation to their good achievement (i.e., relative to other students but not relative to their own performance in other school subjects). Similarly, even poor students may have an average or above-average self-concept in their best school subject; this appears paradoxical in relation to their below-average achievement in that subject, but not when considered relative to their other school subjects.

See also: Culture and the Self: Implications for Psychological Theory; Person and Self: Philosophical Aspects; Self and Emotional Development in Adulthood and Later Life; Self and Identity Development During Adolescence across Cultures; Self in Culture: Early Development; Self-Development in Childhood and Adolescence.

Bibliography

Denissen, J.J.A., Zarrett, N.R., Eccles, J.S., 2007. I like to do it, I'm able, and I know I am: longitudinal couplings between domain-specific achievement, self-concept, and interest. Child Development 78, 430–447. http://dx.doi.org/10.1111/j.1467-8624.2007.01007.x.

Eccles, J.S., 2009. Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. Educational Psychologist 44, 78–89. http://dx.doi.org/10.1080/00461520902832368.

Eccles, J.S., Wigfield, A., 1995. In the mind of the actor: the structure of adolescents' achievement task values and expectancy-related beliefs. Personality and Social Psychology Bulletin 21 (3), 215–225. http://dx.doi.org/10.1177/0146167295213003.

Goetz, T., Frenzel, C.A., Hall, N.C., Pekrun, R., 2008. Antecedents of academic emotions: testing the internal/external frame of reference model for academic enjoyment. Contemporary Educational Psychology 33, 9–33. http://dx.doi.org/10.1016/j.cedpsych.2006.12.002.

Marsh, H.W., 1986. Verbal and math self-concepts: an internal/external frame of reference model. American Educational Research Journal 23, 129–149.

Marsh, H.W., 1989. Age and sex effects in multiple dimensions of self-concept: preadolescence to early-adulthood. Journal of Educational Psychology 81, 417–430. http://dx.doi.org/10.1037/0022-0663.81.3.417.

Marsh, H.W., 1990. The influences of internal and external frames of reference on the formation of English and math self-concepts. Journal of Educational Psychology 82, 107–116.

Marsh, H.W., 2007. Self-concept Theory, Measurement and Research into Practice: The Role of Self-concept in Educational Psychology. British Psychological Society, Leicester, UK.

Marsh, H.W., Craven, R.G., 2006. Reciprocal effects of self-concept and performance from a multidimensional perspective: beyond seductive pleasure and unidimensional perspectives. Perspectives on Psychological Science 1, 133–163. http://dx.doi.org/10.1111/j.1745-6916.2006.00010.x.

Marsh, H.W., Hau, K.T., 2004. Explaining paradoxical relations between academic self-concepts and achievements: cross-cultural generalizability of the internal-external frame of reference predictions across 26 countries. Journal of Educational Psychology 96, 56–67. http://dx.doi.org/10.1037/0022-0663.96.1.56.

Marsh, H.W., Shavelson, R., 1985. Self-concept: its multifaceted, hierarchical structure. Educational Psychologist 20, 107–125.

Marsh, H.W., Trautwein, U., Lüdtke, O., Köller, O., Baumert, J., 2005. Academic self-concept, interest, grades and standardized test scores: reciprocal effects models of causal ordering. Child Development 76, 397–416. http://dx.doi.org/10.1111/j.1467-8624.2005.00853.x.

Marsh, H.W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K.T., O'Mara, A.J., Craven, R.G., 2008. The big-fish-little-pond-effect stands up to critical scrutiny: implications for theory, methodology, and future research. Educational Psychology Review 20, 319–350. http://dx.doi.org/10.1007/s10648-008-9075-6.

Marsh, H.W., Abduljabbar, A.S., Parker, P., Morin, A.J.S., Abdelfattah, F., Nagengast, B., Möller, J., Abu-Hilal, M.M., 2014. The internal/external frame of reference model: age cohort and cross-national differences in paradoxical relations between TIMSS math and science achievement, self-concept and intrinsic motivation. American Educational Research Journal. Online version. http://dx.doi.org/10.3102/0002831214549453.

Marsh, H.W., Lüdtke, O., Nagengast, B., Trautwein, U., Abduljabbar, A.S., Abdelfattah, F., Jansen, M., 2015. Paradoxical relations between self-beliefs and achievements in multiple domains. Learning and Instruction 35, 16–32.

Möller, J., 2005. "Paradoxical" effects of praise and criticism: social, dimensional, and temporal comparisons. British Journal of Educational Psychology 75, 275–295. http://dx.doi.org/10.1348/000709904X24744.

Möller, J., Husemann, N., 2006. Internal comparisons in everyday life. Journal of Educational Psychology 98, 342–353. http://dx.doi.org/10.1037/0022-0663.98.2.342.

Möller, J., Köller, O., 2001. Dimensional comparisons: an experimental approach to the internal/external frame of reference model. Journal of Educational Psychology 93, 826–835. http://dx.doi.org/10.1037/0022-0663.93.4.826.

Möller, J., Marsh, H.W., 2013. Dimensional comparison theory. Psychological Review 120 (3), 544–560. http://dx.doi.org/10.1037/a0032459.

Möller, J., Pohlmann, B., Köller, O., Marsh, H.W., 2009. A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. Review of Educational Research 79, 1129–1167. http://dx.doi.org/10.3102/0034654309337522.

Möller, J., Retelsdorf, J., Köller, O., Marsh, H.W., 2011. The reciprocal I/E model: an integration of models of relations between academic achievement and self-concept. American Educational Research Journal 48, 1315–1346. http://dx.doi.org/10.3102/0002831211419649.

Möller, J., Streblow, L., Pohlmann, B., Köller, O., 2006. An extension to the internal/external frame of reference model to two verbal and numerical domains. European Journal of Psychology of Education 21, 467–487. http://dx.doi.org/10.1007/BF03173515.

Nagy, G., Garrett, J., Trautwein, U., Cortina, K., Baumert, J., Eccles, J., 2008. Gendered high school course selection as a precursor of gendered careers: the mediating role of self-concept and intrinsic value. In: Watt, H., Eccles, J. (Eds.), Gender and Occupational Outcomes: Longitudinal Assessments of Individual, Social, and Cultural Influences. American Psychological Association, Washington, DC, pp. 115–143. http://dx.doi.org/10.1037/11706-004.

Parker, P.D., Marsh, H.W., Lüdtke, O., Trautwein, U., 2013. Differential school contextual effects for math and English: integrating the big-fish-little-pond effect and the internal/external frame of reference. Learning and Instruction 23, 78–89.

Parker, P.D., Schoon, I., Tsai, Y., Nagy, G., Trautwein, U., Eccles, J., 2012. Achievement, agency, gender, and socioeconomic background as predictors of postschool choices: a multi-context study. Developmental Psychology 48, 1629–1642.

Pekrun, R., 2006. The control-value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. Educational Psychology Review 18, 315–341.

Pohlmann, B., Möller, J., 2009. On the benefits of dimensional comparisons. Journal of Educational Psychology 101, 248–258. http://dx.doi.org/10.1037/a0013151.

Seligman, M.E.P., Csikszentmihalyi, M., 2000. Positive psychology: an introduction. American Psychologist 55, 5–14. http://dx.doi.org/10.1037/0003-066X.55.1.5.

Shavelson, R.J., Hubner, J.J., Stanton, G.C., 1976. Self-concept: validation of construct interpretations. Review of Educational Research 46, 407–444.

Wigfield, A., Eccles, J.S., 2002. The development of competence beliefs and values from childhood through adolescence. In: Wigfield, A., Eccles, J.S. (Eds.), Development of Achievement Motivation. Academic Press, San Diego, CA, pp. 92–120. http://dx.doi.org/10.1016/B978-012750053-9/50006-1.

Xu, M.K., Marsh, H.W., Hau, K.-T., Ho, I.T., Morin, A.J.S., Abduljabbar, A.S., 2013. The internal/external frame of reference of academic self-concept: extension to a foreign language and the role of language of instruction. Journal of Educational Psychology 105 (2), 489–503. http://dx.doi.org/10.1037/a0031333.

Internal Migration: Developing Countries

Randall Kuhn, Josef Korbel School of International Studies, Denver, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D.J. Hogan and J.M. Pinto da Cunha, volume 11, pp. 7733–7737, © 2001, Elsevier Ltd.

Abstract

While internal migration is widely considered to be an essential driver of economic efficiency in More Developed Countries (MDCs), the process remains relatively understudied, undermeasured, and misunderstood in Less Developed Countries (LDCs). Existing evidence suggests that internal migration in LDCs, as in MDCs, is beneficial for development and social change, though the pathways of influence are complex. This article explores evidence on measurement, levels and trends, causes, and consequences of internal migration. The article situates internal migration within the broader frameworks of development and livelihoods, as both cause and consequence.

While internal migration is widely considered to be an essential driver of economic efficiency in More Developed Countries (MDCs), the process remains relatively understudied, undermeasured, and misunderstood in Less Developed Countries (LDCs). In recent years international migration from LDCs has gained increasing attention from researchers and policy makers, largely on the basis of the economic potential of the estimated $515 billion dollars in annual cross-national remittance flows (World Bank, 2014). While less lucrative per average episode, internal migration affects a far greater share of the population of many LDCs while in many cases carrying fewer costs or risks. Yet many LDC governments and development experts remain ambivalent about the role of internal migration in economic development and concerned about the potential burden of future migration related to climate change (De Haas, 2010; Bakewell, 2008). Existing evidence suggests that internal migration in LDCs, as in MDCs, is beneficial for development and social change, though the pathways of influence are complex.

The causes and consequences of migration can be broadly situated within processes of development and livelihoods (de Haas, 2006; Ellis, 1998). Migration scholars have recently rediscovered the mobility transition framework identified by Zelinsky (1971), which connects migration to the broader framework of demographic transition and modernization. The mobility transition implies a shift in both the extent and nature of migration, from circulating mobility over short distances (premodern), to increased and systematic migration to urban and frontier areas (early transition), to a slowdown in urban and frontier migration with increasing regional and interurban mobility as economic development deepends. These stages effectively illustrate the relationship between development and the well-documented 'migration hump' in which migration rises and falls over time (de Haas, 2006; Martin and Taylor, 1996). Migration is fed by the absolute underdevelopment of sending areas, yet processes of development also create incentives and resources necessary to migrate. Migration then feeds further processes of development (and in some cases underdevelopment), thereby creating conditions for new economic opportunities in sending areas, both via further migration and economic transformation of the sending area.

While a general story has begun to emerge, our knowledge of internal migration is incomplete in many ways, most notably due to the patchwork of data sources, causal frameworks, and micro–macro designs. Cross-national comparisons of the quantity of internal migration tend to come from exhaustive efforts to reconcile data from incomplete and incompatible data systems. By contrast, most evidence on the social, economic, cultural, and ecological drivers and impacts of migration come from high-quality data from geographically bounded surveillance sites that may not necessarily represent internal migration more broadly. Few studies have articulated across these divergent streams, and both approaches are bedeviled by the complexity of migration, beginning with the very definition of migration.

Measurement and Definitions

The definition and classification of migration depends primarily on spatial and temporal variables. Spatially, the identification of internal rather than international migration depends on clear delineation of national boundaries (King and Skeldon, 2010). Shifting boundaries challenge the classification of internal versus international moves. As just one example, were the approximately 15 million individuals displaced during the partition of India internal or international migrants, and does it matter whether a person moved 1 day before or 1 day after the actual partition on 14 August 1947? Distinctions are also confounded by the common though declining presence of frontier-free boundaries in which migrants may make little distinction between international and internal moves across the same economic, ecological, or cultural space (King and Skeldon, 2010).

Looking within the boundaries of a nation, the frequency and meaning of migration will depend greatly on the geographic boundaries across which migration is defined. The use of larger geographic units can allow for a clearer definition of migration, yet eliminates coverage of short-distance moves (Ravenstein, 1885).

Classification of migration is also temporal. In systems that allow the direct estimation of migration, including panel surveys and demographic registration systems, migration is

most often bounded by the definition of a 6-month absenc from the household. This definition may be further modified by accounting for functions and intentions. Many populations also experience high rates of seasonal migration, with significant consequences for well-being (Rogaly and Rafique, 2003). While such moves can be identified based on patterns of short-term employment, short-term movement may also relate to schooling, childbirth, provision of care to the disabled, and displacement.

Moves of long duration may also be qualified by migrant intentions, most notably by the practice of solo, temporary, or circular migration, as in South Africa and China. In South Africa, the Apartheid-era pattern of enforced circular migration among black populations persists 20 years after the end of Apartheid, with migrants returning home during episodes of unemployment, illness, or retirement (Clark et al., 2007). The Agincourt Health and Demographic Surveillance System, which collects data in a former homeland area of Mpumalanga Province, thus classifies temporary migrants as *de facto* members of the sending household, recording the number of months that the individual spends away from the household (Clark et al., 2007).

In practice, most societal estimates of internal migration are based not on prospective observation of migration events, as in a registration system, but instead on intercensal estimates of mobility. The International Migration Around the GlobE (IMAGE) project has conducted an exhaustive summary and compilation of available migration data from censuses and other sources, as summarized in **Table 1** (Bell et al., 2014; Bell and Charles-Edwards, 2013). They note that 142 of 193 nations surveyed included direct questions on migration. In terms of duration, these questions varied from the ideal but resource-intensive approach of measuring the duration of residence in the most recent location, thereby indicating the exact year of migration but excluding earlier moves. Other censuses report changes in residence over a fixed period, typically 1, 5, or 10 years, thereby capturing moves over a fixed observation period but excluding onward or circular moves. Most countries recorded only whether a respondent was living in a different location from their place of birth, meaning that lifetime circulation is ignored and rates of ever-migration will be sensitive to population aging. Whatever the temporal definition of migration, most censuses only track moves between the primary or secondary administrative units, thereby ignoring local moves. Intercensal estimates inherently exclude the behavior of nationals living overseas, except in systems that allow *de facto* residence. Finally, one-quarter of countries did not include any migration questions on their censuses, meaning the net migration flows can only be inferred indirectly by observing changes in population net of natural increase.

Bell and Charles-Edwards (2013) note that a significant number of countries, including 37 of the 51 that lack census-based estimates, have some survey-based migration data. Migration life history surveys can provide rich coverage of lifetime migration activity across a wider range of durations and distances, though such data are both cost intensive, they may still not offer cross-national comparability, and they may be subject to retrospective reporting bias (Reed, 2013) While IMAGE reports a number of migration-specific surveys, more widely available data sets carry significant limitations and biases. For instance the Demographic and Health Surveys (DHS), the most widely reproduced survey in Sub-Saharan Africa, only includes women of reproductive age and includes just two questions on migration. Some recent studies have begun to explore the use of mobile phone records and social media feeds to infer mobility rates, but these methods require further refinement, have limited potential for behavioral analysis, and raise ethical concerns over privacy (Blumenstock, 2012).

Table 1 Countries collecting internal migration data since 1995

Region	*Census*	*Register*	*Survey*	*Multiple sources*	*Total countries collecting data*	*Total countries*
Africa	43	0	38	31	50	54
Asia	37	15	24	27	41	47
Europe	31	32	32	34	41	43
Latin America and the Caribbean	32	0	12	12	32	33
Northern America	2	2	2	2	2	2
Oceania	13	1	2	3	13	14
Total	158	50	110	109	179	193

Source: IMAGE Inventory, published as Table 1 in Bell, Martin, Charles-Edwards, Elin, Kupiszewska, Dorota, Kupiszewski, Marek, Stillwell, John, Zhu, Yu, 2014. Internal migration data around the world: assessing contemporary practice. Population, Space and Place.

Levels and Trends

No explorations of levels and trends in internal migration would be complete without a discussion of urbanization measures. **Table 2** compares urbanization rates for broad geographic regions in 1970, 1990, and 2010. Most developing regions have substantially lower levels of urbanization, even after including emerging high-income states within those regions. In the past and present, Latin America remains an outlier for its high levels of urbanization. Historically, Asia and Sub-Saharan Africa have had especially low levels of urbanization. The past 20 years have seen substantial urbanization in East Asia, which rose from 33.7% in 1990 to 51.7% in 2010, resulting primarily from high levels of mobility in China. South Asia and Sub-Saharan Africa have historically had rates of urbanization that were both low and strongly correlated with levels of economic development. In more recent years Sub-Saharan Africa has seen more rapid urbanization even as economic development has proceeded more rapidly in South Asia, raising concerns about the social and economic consequences of urbanization in Sub-Saharan

Table 2 Rates of urbanization for key world regions, 1970–2010

	Percent of population in urban areas		
Major area	1970	1990	2010
South Asia	18.7	25.0	30.6
Sub-Saharan Africa	19.3	27.9	36.0
East Asia and Pacific	24.8	33.7	51.7
Middle East and North Africa	42.8	54.8	62.6
Latin America and Caribbean	57.0	70.3	78.8
Europe and Central Asia	60.6	67.4	70.0
North America	73.8	75.4	82.0

Source: World Bank World Development Prospects.

Africa. While urbanization is more readily observed than migration flows, the convolution of internal migration and urbanization is problematic for many reasons. First, a substantial share of urbanization in all regions results not from migration, but from the administrative reclassification of areas from rural to urban (Lucas, 1997). Second, rural–urban moves have historically constituted a small share of total internal migration, particularly as internal migration systems mature and already display high urbanization rates, as in Latin America (Ravenstein, 1885; Zelinsky, 1971).

The IMAGE project has attempted to reconcile disparate temporal and spatial measures of migration flows from census data (Bell and Charles-Edwards, 2013). In one summary exercise summarized in **Table 3**, they use flows of migration across the largest administrative regions in observed countries to estimate 5-year and lifetime internal migration intensity rates for each major world region. Lifetime measures of migration intensity suggest that the experience of migration is highest among countries in Latin America (18.0%) compared with Africa (12.5%) and Asia (7.2%). While lifetime migration intensity can be estimated for a wider range of censuses, it may not reflect current trends, particularly as extended duration of survival allows past migrants to live increasingly long lives. Five-year measures of migration intensity reveal greater comparability among these three regions, with the highest levels recorded in Africa (4.6%), then Latin America (4.1%), then Asia (2.9%).

The authors use these estimates to approximate the total number of internal migrants in the world today. They estimate that 763 million, 11.7% of the world's population live outside their region of birth but within their country, including 496 million in Africa, Asia, and Latin America. Both numbers are considerably higher than the estimated 200–300 million thought to be living outside their country of birth.

Finally, they use the spatial asymmetry of interzonal migration flows to estimate a Migration Effectiveness Index (MEI), the extent to which migration actually redistributes population as opposed to merely circulating populations. They observe a relatively low MEI for most Latin American Countries in the 2000 census round (e.g., 7.2% in Chile, 14.6% in Costa Rica, 23.3% in Brazil, 28.8% in Mexico), with only Ecuador showing an MEI of 63.3%. As a result, the aggregate net migration rate (the migration intensity * MEI) is actually below 2% for most Latin American countries other than Ecuador. In Asia, by contrast, MEIs are relatively high, ranging from 62.4% in China and 59.7% in Vietnam to 26.2% in the Philippines and 19.8% in Indonesia. As a result, the net impact of migration on population distribution may be much larger in Asia. For instance, they estimate that 1.7% of China's population experienced a net shift of province of residence in the 5 year preceding the 2000 census, compared with just 0.6% net movement between Brazil's states and 1.2% between states in the USA. In Africa, where fewer countries provided 5-year rates, the pattern was suggestive of high overall net migration with more minimal impact. One counterexample would be South Africa, where a 4.3% migration intensity and 33.9% MEI resulted in a fairly substantial 1.4% net movement between provinces.

Finally, contrary to popular expectation, they found that internal migration rates in most countries declined or remained

Table 3 Global estimates of the number of internal migrants in 2005, five-year and lifetime data

	Previous estimates[a]			*Updated estimates*			
Region	*PAR*[b] *(millions)*	*Migrants (millions)*	*Intensity*	*PAR (millions)*	*Migrants (millions)*	*Intensity*	*Number of countries*
			5 year				
Africa	860.7	32.9	4.0	865.6	39.7	4.6	4
Asia	3724.9	96.2	2.7	3747.7	109.8	2.9	10
Latin America and the Caribbean	531.1	18.7	3.7	529.2	21.9	4.1	16
Europe				694.2	34.8	5.0	3
North America				312.8	21.2	6.8	3
Oceania				31.9	1.8	5.7	3
More developed regions	1150.5	87.8	8.1				
Global	**5937.0**	**235.7**	**3.7**	**6181.3**	**229.2**	**3.7**	**39**
			Lifetime				
Africa[c]	860.7	104.1	12.8	911.1	113.5	12.5	13
Asia	3724.9	207.8	5.9	3945.0	282.1	7.2	14
Latin America and the Caribbean	531.1	101.6	20.2	557.0	100.2	18.0	23
Europe				730.7	166.0	22.7	10
North America				329.2	91.5		
Oceania				33.5	9.3	27.8	6
More developed regions	1150.5	326.3	29.9				
Global	**5937.0**	**739.9**	**12.5**	**6506.6**	**762.6**	**11.7**	**66**

[a]Source: Bell, Martin, Muhidin, Salut, 2009. Cross-national comparisons of internal migration. Human Development Research Paper 2009/30. UN Department of Economic and Social Affairs. Population Division, NY.
[b]PARs are calculated as 95% of the 2005 mid-year population.
[c]Zambia and Zimbabwe not included in calculation for lifetime migrants in Africa.
Source: Originally published as Table 9 in Bell, Martin, Charles-Edwards, Elin. 2013. Cross-national Comparisons of Internal Migration: An Update of Global Patterns and Trends. United Nations Population Division Technical Paper No. 2013/1.

stable between 1990 and 2000, especially in previously mobile Asian countries like Indonesia, Malaysia, and Vietnam. The notable exception was China, which saw the intensity of interprovincial migration more than double between 1990 and 2000 (1.1–2.7%), as did the MEI (from 32.2 to 62.4%), resulting in a fourfold increase in the rate of population redistribution from 0.4% to 1.7%. They note that census data for 2010, already available in Indonesia and Vietnam, show a marked rebound in migration intensities. In Latin America, migration intensities were mostly stable through the 1990 to 2000 period. None of the African censuses used provided sufficient data to track 5-year migration trends.

Causes of Migration

The decision to migrate can be situated within the neoclassical utility maximization frameworks employed for models of international migration. Arthur Lewis (1954) developed a two-sector model including an underdeveloped agricultural sector with surplus employment and a growing manufacturing sector that would draw excess labor from agricultural areas, increasing urban productivity. This emphasis on urban growth as a driver of migration led to the frequent reference to 'pull factors' in migration. Ranis and Fei (1961) refined the Lewis model to address the role of out-migration in reducing capital-labor ratios in the rural sector, thereby also improving rural productivity, which would give rise to the notion of 'push factors' driving people away from the rural area.

These sectoral models were further refined at the individual level. Sjaastad (1962) formulated the probability of migration between two locations in terms of the difference in expected wage between the two areas net of any financial and nonfinancial (psychic) costs of moving. This model opened the way for consideration of individual variations in human capital. In this framework, for instance, the tendency for migration to peak in early adulthood can be understood as a maximization of the long-term returns to migration over the life course.

One of the most frequently replicated drivers of migration is higher levels of human capital. Direct comparisons of the association of schooling to international versus rural–urban migration tend to find that rural–urban moves are more strongly associated with schooling. Even though international migration may be more lucrative, rural–urban migration may carry greater opportunities for professional advancement and skills acquired in local schooling systems may be more valuable in the domestic economy. To the extent that it has been studied, the association of schooling to short-distance and rural–urban moves is weaker than for rural–urban moves.

While the association of schooling to internal migration is well established, few studies have attempted to separate the specific effect of schooling from other forms of human capital like physical health, cognition, or psychosocial health. The international migration literature has now strongly established the 'healthy migrant effect' whereby migrants tended to be strongly selected on physical health, even in migration streams dominated by occupations that do not demand physical exertion. A few studies have just recently begun to test for the presence of health selection in internal migration, generally supporting the hypothesis (Lu, 2008).

Subsequent extensions of the neoclassical model moved beyond individual human capital to understand the role of familial financial capital, physical capital, and livelihoods in the migration process. The Harris-Todaro framework extended the standard model by arguing that the expected urban wage be taken as the average wage times the average probability of employment (Harris and Todaro, 1970). This allowed for explicit consideration of aversion to the risk of current or future unemployment, variations in the probability of employment due to migrant skills, and the role of exclusionary policies or trade unions in restricting actual access to employment. Mincer (1978) extended these models to the locus of a family of movers, arguing that the net benefit of migration to the working mover(s) must outweigh the cumulative costs experienced by all tied movers, thereby explaining the relatively low migration rates experienced by marriage people and the dynamics of tied mover migration.

Mincer's work, combined with a growing awareness of the phenomena of temporary, circular, and seasonal migration, shifted the focus of migration decision-making to the family, including both movers and nonmovers. The New Economics of Labor Migration, which has been the normative model of migration decision-making for almost two decades, posited that individual migration behavior is often the product of household or family decisions aimed at maximizing the welfare of a collective income-pooling unit (Stark, 1991). As a result, potential movers would consider a much broader range of potential outcomes than merely current income from migration, including long-term income from a combination of income streams of which migration was just one part. This also led to consideration of the unique nature of migration within the context of traditional rural agricultural production described by Lewis. Given weak capital markets in rural areas, families may see reliable, high-value cash transfers sent by migrants as a source of insurance against crop failure or other catastrophe and as a critical source of capital for investments in land and the means of production (Rosenzweig and Stark, 1989).

NELM offered two critical predictions that deviated from the standard neoclassical model. First, while migration may be a response to low current wages, it may also be part of an effort to protect or enhance existing income streams (Lucas and Stark, 1985). Households with greater land holdings may have less need for current income, but they may also have a greater interest in insuring production against shocks and in generating capital for economic expansion. Households may also be able to use land or revenue from land to finance migration. This is borne out by studies documenting a nonlinear relationship between land holdings and migration (VanWey, 2005). Second, community or contextual factors predisposing households to economic risk or limiting their ability to manage risk may increase the incentives to migrate, particularly among households with greater exposure to agricultural production (Rosenzweig and Stark, 1989).

The NELM framework was broadened through the more general linkage of migration to broader life course dynamics of production and consumption (de Haan, 1999; Ellis, 1998; de Sherbinin et al., 2008). While NELM described the conditions of an agricultural producing household with long-term interests in agricultural enterprise, it was less effective at describing

the conditions facing a range of households whose livelihoods were less deeply tied to agricultural production, but instead to a nexus of on-farm and off-farm labor activities in which land was a less direct driver of migration (Bilsborrow and DeLargy, 1990). On the other end of the socioeconomic distribution, declining family sizes and increasing socioeconomic aspirations may lead some families to view the migration of one or more family members as a pathway toward an eventual transition to the city, to the elevation of familial status in the sending area, or both.

As a result, the internal migration literature has shifted quite rapidly in the past decade to situate migration within a general livelihoods framework (Ellis, 1998). A recent review emphasized the role of human capital, financial capital, physical capital (e.g., land), natural capital (environmental resources and amenities), and social capital as drivers of migration (de Sherbinin et al., 2008). Critically, migration may be driven by a surplus or deficit of these forms of capitals, or more likely by a desire to leverage existing capital to either enhance or protect future livelihood opportunities. The livelihoods framework is useful not only for predicting migration in general, but also for predicting complex processes such as temporary versus permanent migration, family versus individual, and sequences of migration to multiple destinations (Bilsborrow and DeLargy, 1990).

One further advantage of the livelihoods framework lies in understanding that migration is as much a process of economic, social, and cultural reproduction as it is a process of disruptive innovation. If migration is deeply embedded in social structures, institutions, and livelihood strategies, then it may serve to reproduce and reinforce those structures. Numerous studies have demonstrated how the role of women in the process of migration at the family and community level reflects women's disadvantaged position within the household economy, lineage and inheritance systems, employment, and political institutions (Curran and Rivero-Fuentes, 2003; Gray, 2012; Mojola, 2014). While mobility has the potential to disrupt these well-worn patterns, migration may just as easily expose sending populations to new modes of discrimination imposed by the state, by employers, or by the globalizing forces that drive the need for migrant labor (Silvey and Elmhirst, 2003).

The State and Other Political Actors

Governments and nonstate institutions play a direct and indirect role in promoting or discouraging migration at the origin or destination. This is most notable in examples of the forced migration of ethnic, racial, or other status groups from origin areas (Raleigh, 2011). In many cases, states have also designated the eventual destination of internally displaced peoples (IDPs), most notably in the case of the system of Native American reservations in the United States, Jewish ghettoes in Europe, and the African homeland system under South African Apartheid. While direct displacement activities may affect entire populations without selectivity, evidence suggests that other forms of conflict-related migration may be subject to considerable selectivity in terms of livelihoods, vulnerability, and human capital (Bohra and Massey, 2012).

States may also play a direct role in internal displacement through development processes, as in the inundation of communities by hydroelectric dams (Hunter, 2005; Cernea, 2006). While some such developments may necessarily occur in a particular location (e.g., dams must be situated at a particular run of the river), states nonetheless have the opportunity to minimize the population impacts of such projects, to plan them in ways that are more conducive to population retention, and develop resettlements that meet the economic and cultural needs of the inhabitants. Relatively few studies have followed populations displaced by government projects, those that have been conducted suggest a high rate of onward migration among the displaced and anticipation of migration among the nondisplaced, with high selection on other factors relating to vulnerability (Scudder, 1993; Hwang et al., 2007).

States may also play a direct role in encouraging populations to migrate to specific areas. In the case of frontier migration, states may play a direct role in forest clearing; provide subsidies in the form of cash, land, or tools; and provide indirect support through legal institutions, public works, and military protection (Shrestha et al., 1993; Alston et al., 1995; Fearon and Laitin, 2011). The processes of settlement and displacement are uniquely linked in practice and often in ideology. Most areas referred to as frontiers contain indigenous human populations whose livelihoods are typically well adapted to the presettlement ecology and economy and not well adapted to the postsettlement context (Fearon and Laitin, 2011). Furthermore, settlement programs, particularly direct financial incentives, may be targeted directly toward one ethnic group and may directly or indirectly deprive other groups of access to resources, whether by intent or practice (Yiftachel and Ghanem, 2004). As a result, settlements can have adverse consequences on native and migrant populations, with frequent informal and formal conflict among the groups. The often-harsh conditions and weak institutions of frontier areas can lead to high levels of return migration and second-generation out-migration among settlers, again with significant selectivity on individual and household characteristics (Parker et al., 2008; VanWey et al., 2012).

States also exert considerable control over internal migration through legal and economic policies, through direct restrictions on mobility and through the management of policies relating to entitlements, urbanization, and land reform. A few nations have legal systems that directly control the flow of migration. The South African Apartheid labor policy was managed by a system of passes restricting migrants from settling, bringing family, or moving freely in designated white-only areas, thereby recreating the most severe disadvantages inherent in an international guest worker system in a domestic context (Wolpe, 1975). More recently, China's *hukou* system, in which an individual's *de facto* residence and access to social entitlements is assigned at birth, has similarly depressed the wages, bargaining power, and entitlement access of rural–urban migrants (Mackenzie, 2002). Restrictions on urban residency may also reduce wages by forcing solo and temporary migration, thereby shifting the costs of reproduction, retirement, and unemployment to the rural area, where smallholders may subsist through a combination of agricultural and migrant earnings. By contrast, development policies in post-WWII Latin America focused on the enclosure of lands and the promotion of high-wage, high-benefit jobs, thus encouraging permanent

migration, whether into formal employment and housing or to informal employment in informal settlements (de Janvry and Garramon, 1977). This variation may go some way to explaining the distinct patterns of urbanization in Latin America and Asia.

Public service provision at the source or destination may also alter migration flows, though in unpredictable ways. The United States and European Union have a substantial literature on the 'welfare magnet hypothesis', whereby differential provision of public transfers or services will make some areas more attractive, particularly for low-skilled populations (McKinnish, 2005). These studies have now been replicated in LDC contexts, with inconsistent results. Studies of cash transfer programs have identified attenuated or positive effects on migration, potentially due to the diversification of income streams or the availability of moving costs (Ardington et al., 2009). A study of programmatic placement of valuable health services in the Matlab area of Bangladesh, on the other hand, finds negative effects on out-migration, presumably due to the amenity value of living in the service area (Barham and Kuhn, forthcoming). A long-term study of an array of public services in Nepal found a short-term amenity effects on reducing migration and a long-term human capital effect on increased migration (Massey et al., 2009). Taken together, the evidence suggests that the role of the state in driving migration is both substantial yet mediated heavily by migration context and individual livelihoods.

Environmental Factors

More extensive research has been conducted on the role of environmental resources in determining migration (Hunter, 1998). This is understandable given the potential relationships between environmental change and modes of agricultural production that continue to support rural livelihoods (Parker et al., 2008). New patterns of environmental change at global and local scale add further urgency to such research (Bardsley and Hugo, 2010). While changes in the ecosystem services provided by a local environment can theoretically reduce incentives to migrate, most research has thus far focused on the role of the environment in driving migration, primarily as a push factor that reduces the value of the ecosystem (Gray, 2009). Conversely, most research on receiving areas has focused on the value of ecosystem services, for instance in coastal areas, as a pull to migration, and on the consequences of in-migrants for those ecosystems (Hunter, 1998; Curran, 2002).

The literature on environmental determinants of out-migration highlight the need to distinguish direct pathways of environmental risk – for instance those relating to natural hazards – from those effects that are mediated by livelihoods (Hunter, 2005). The direct pathway has been subject to considerably greater attention from policy makers and media due to the perceived risk of future climate-induced migration (Bardsley and Hugo, 2010). A number of studies have estimated that between 10 million and 40 million people are already environmental refugees, and that this number could approach 200 million by the year 2050, yet critics have questioned the definition of environmental refugee used in these studies (Gemenne, 2011).

One indication of a truly distinct process of environmental displacement would be the absence of individual- and household-level selection in the decision to migrate. If people are truly compelled to move, then there would be no place for individual agency, yet few cases of environmental migration meet this test. One of the few large-scale studies of such a disaster linked satellite data on tsunami to data on household migration behaviors in posttsunami Aceh, Indonesia (Gray et al., 2014). While the study found that individual and household selectivity in migration was much lower in more heavily damaged areas, these distinctions nevertheless existed on a continuum. Even moderately affected areas showed considerably selectivity. Even in heavily damaged areas, factors like poverty and female headship determined whether a household migrated to a displacement camp or to the home of a relative, and whether the displaced eventually returned. Hunter (2005) notes that responses to natural hazards, particularly chronic exposures like environmental contamination, tend to exist on a continuum of coping responses, including remaining in place, moving short distances, having some family members move, or total displacement, though even the displaced may return. Finally, perception of the risk of natural hazard precedes the occurrence of actual catastrophes by many years, so that far more households in risk-prone areas will use migration as part of a long-term risk management strategy to diversify livelihoods and even ensure a place of refuge in case of future disaster (Bardsley and Hugo, 2010; Hwang et al., 2007).

There is thus an emerging consensus around the notion that changing environmental conditions affect migration through a dynamic and unfolding process of livelihood depletion, livelihood risk, and perception of future catastrophic risk. While findings vary, this literature has formed a consensus around a few key findings. First, environmental factors play a more significant role in shorter distance moves than in long-distance moves, particularly overseas migration (Gray, 2009; Massey et al., 2010). Second, migration effects are more significant in the presence of measurable income shocks (for instance due to declining agricultural yields) or observable risks to livelihoods (such as drought) rather than the presence of transitory risks like flooding (Mueller et al., 2014). Third, the relationship between environmental resources and out-migration is not always negative. Environmental resources can also be framed as a source of natural capital enabling migration (Gray, 2009). Conversely, studies in the Sahel have identified conditions in which the level of environmental depletion and poverty is sufficiently acute to reduce the likelihood of migration (Findley, 1994). Finally, as with the role of the state, the environmental effects of migration are heavily mediated by community, household, and individual-level factors.

Social Capital

The practice and persistence of migration is driven in part by the accumulation of social capital, represented by knowledge, social connections, and direct opportunities to migrate through the use of social connections, often referred to as chain migration. This relationship is typically represented through the association between higher levels of past migration and

higher levels of subsequent migration at the household, community, regional, and national level (Lindstrom and Lauster, 2001; Van der Gaag and Van Wissen, 2002). Early movers tend to come from the highest strata of a community, as household wealth may buffer prospective migrants against the costs and risks associated with moving (Kuhn, 2003). As a result, some evidence suggests that migration can widen community inequalities in the short run but narrow them in the long run as opportunities diffuse more widely (Taylor and Wyatt, 1996).

Over time, some social and cultural changes engendered by migration may transcend the pathways of specific social networks and penetrate the community at large through the formation of infrastructure, institutions, values, or identities salient to the practice of migration and thus approximating an externality favoring the practice of migration (Zhao, 2003). This process of cumulative causation may include the reorientation of local markets, schooling, transport, or other infrastructure toward reducing the costs and risks of migration, changes driven in part by the expenditure of migrant remittances (Lindstrom and Lauster, 2001). As the number of individuals in a local area network having knowledge, experience, or authority in a specific destination area increases and reaches a critical density, individual social connections can be converted into a form of migration-specific social capital based on shared community or ethnic origins (Hagan and Ebaugh, 2003; Kuhn, 2003).

The notion of migration-specific social capital as a public good comes with several caveats and concerns over simplifying social capital as a public good (Portes, 1998). First, while migration opportunities may permeate some communities, the earliest opportunities may nonetheless carry the highest returns. Second, migration-specific social capital may exclude some segments of a population and reproduce existing social cleavages. A number of studies have identified the presence of separate migrant networks reflecting existing social cleavages, most notably separate networks promoting women's and men's migration (Curran and Rivero-Fuentes, 2003; Silvey and Elmhirst, 2003). Third, the distribution of a particular and potentially lucrative opportunity such as migration may carry considerable reciprocity costs that may exclude others (Burt, 2000; Winters et al., 2002). Finally, migration opportunities may often transmit through weak ties that cross community boundaries, potentially bypassing some members of a high out-migration community in favor of new entrants from other communities (Kuhn, 2003). As a result, consideration of migration impacts must take careful note of migrant selectivity and impacts on nonmigrant populations.

Impacts of Migration

Most studies consider separately the welfare of migrants themselves and those left-behind by migration. Theories of migration as a family livelihood response point to the conceptual and empirical challenge of making this distinction. Migration will typically lead to a substantial increase in the wages of the migrant, though not in all cases (Lucas, 2004). Yet a substantial proportion of migrants remit a large share of their income back to an origin household that represents a source of shared livelihood and well-being for the migrant as well. As a result, studies of migrant consumption for instance are naturally biased by the decision among many migrants to minimize consumption in the destination area specifically to maximize sending-area consumption. This decision itself has implications for the well-being and health of migrants themselves. Many internal migrants, particularly to cities, live in conditions of deprivation relative to existing urban residents, occupying low-paid and often risky jobs and living in poor living conditions, including informal settlements.

This is especially evident in the case of health, where occupational exposures, lack of access to health services, avoidance of health expenditures can interact to produce adverse health outcomes compared with urban natives and to those left behind in rural areas (Ebrahim et al., 2010). Notable concern has been expressed about the role of migration as a risk factor for sexually transmitted disease, including HIV/AIDS, though studies have lent only qualified support to the significance of this relationship among certain migration streams (e.g., short distance moves), mobility patterns (e.g., truck driving), and stages in the migration process (Deane et al., 2010). Moreover, while the tendency to reduce destination-area expenditures may carry health risks, at least one study has noted an association between the sending of remittances and increased subjective well-being of respondents in China (Knight, 2010).

The traditional focus of migration impact studies has been on the sending household and, to a lesser extent, the sending community. These studies tend to focus first on remittances from internal migrants to the sending household. There is no available global estimate of the total flow of internal migrant remittances, given the difficulty of measuring internal migration itself and the informality of many domestic remittances. Nevertheless, evidence from Asia, especially China, suggests that total domestic remittances flows are at least on the same scale as international remittances, if not as lucrative in the absolute. International migrant remittances tend to be large on a per-migrant basis, yet much of this gap is offset by the greater proportion of households that send internal migrants. Evidence from China, India, and Ghana places the total value of domestic remittances at 20–100% greater than international remittances (Castaldo et al., 2012).

Although remittances can be quite lucrative for those who receive them, the literature also highlights the selectivity of remittance flows. In addition to the notable forces of selection into migration, research also suggests selectivity in the tendency to receive remittances. Some of this selectivity is beneficial to sending households, for instance the tendency for remittances to rise during negative economic shocks (Rosenzweig and Stark, 1989). Remittances are also more likely when a migrant expects to inherit property, resulting for instance in larger remittance receipts for households with greater land holdings.

Studies of migrant and remittance impacts on sending communities generally support the presence of positive impacts, though these impacts may depend greatly on the outcome of interest, the context of migration, the social position of those impacted, migrant tenure and position, and the role of self-selection (Mendola, 2008). In general, migration and remittances result in increased consumption and the

reduction of poverty at the household level (Quisumbing and McNivel, 2010). While early studies questioned the impact of remittances on productive investments given persistent structural constraints in sending areas, some recent studies identify cases and conditions under which migration may lead to productive and forward-looking agricultural investments (Taylor et al., 2003; VanWey et al., 2012).

There is much stronger evidence to support investment of migrant remittances in human capital. A number of studies have measured positive consequences of migration for schooling investments and educational outcomes among children in migrant households and among nonmigrant children in migrant-sending communities (Quisumbing and McNivel, 2010; Kuhn, 2006).

Evidence on health outcomes among the left-behind generally point to improvements, though this depends on the outcome of interest and the social position of the beneficiary. Studies of adult self-rated health (SRH), observed physical function, mortality, and health-seeking behavior have found a positive associations with migration (Kuhn, 2005; Kuhn et al., 2011; Adhikari et al., 2011). Kuhn et al. (2011) use data from three rounds of the Indonesian Family Life Survey and a propensity matching framework to address the longitudinal effects of children's internal migration on changes in SRH, observed physical function, and mortality after accounting for self-selection into migration, though impacts were larger for high-propensity households. Some studies of adult mental health have observed negative association with migration (Adhikari et al., 2011; Arenas and Yahirun, 2011), though other studies have found a positive association (Abas et al., 2009), suggesting that the economic returns to children's migration may not fully offset the loss of personal social support.

Although there is some evidence of higher HIV/AIDS risk among migrants themselves, studies of elevated HIV/AIDS risk among the left-behind have proven far less conclusive (Deane et al., 2010). Instead research suggests that any association between mobility and HIV depends on the frequency of travel, as in the case of truck drivers or fishermen (Mojola, 2014). The association between migration and HIV/AIDS in sending communities is further confounded by the tendency for sick migrants to 'return home to die', thereby creating a spurious association between HIV/AIDS and nonmigrants living in a community (Clark et al., 2007).

Finally, a growing literature has addressed the ecological consequences of migration for rural areas, including migrant-receiving areas in cities, coastal ecosystems, and frontiers along with a few studies of migrant-sending areas. Studies of frontier areas tend to highlight the association of migration with deforestation and reduced biodiversity (Carr, 2009). Coastal ecosystem studies highlight the potential role of migration on ecosystem degradation, both through increased population and through poor stewardship on the part of migrants, but numerous case studies and systematic reviews argue that these effects depend greatly on the nature of migrant selectivity, the conditions of migration, and migrant social networks (Cassels et al., 2005; Curran, 2002; de Sherbinin et al., 2008). Studies of migrant-sending areas have identified by an agricultural deintensification effects stemming from reduced population pressure (Qin, 2010) and an agricultural intensification effect resulting from the use of remittances for mechanized agriculture, fertilizer and pesticide use, and monoculture (de Haas 2006; Gray, 2009). While the tendency for migration to increase agricultural intensity raises ecological concerns, more recent studies have noted that over the household life cycle migration may promote less ecologically demanding land uses like perennial crops and forestry (VanWey et al., 2012).

The Way Forward

Recent progress in understanding the causes and consequences of internal migration has allowed researchers to test a wide range of significant hypotheses in relation to an increasingly widely shared livelihoods framework. The inherent challenge of understanding migration as both a cause and consequence of systems of political, social, ecological and economic change requires a range of new approaches. While econometric adjustments to account for selection and community-level variation are important in establishing a causal framework for controlled impact evaluation, they also tend to obscure the truly unique aspects of migration relating to the unfolding process of reciprocal change and the variations in this process between even highly localized communities. Such reduced-form analyses must increasingly be complemented by the application of methods that explicitly address dynamic processes of change, including both qualitative narratives of change and quantitative approaches to modeling multiple interacting processes. This is particularly important in studies of migration, livelihoods, and the environment, where actual, perceived, or anticipated changes of even a relatively small scale can result in a cascade of livelihood adaptations that may be stabilizing or destabilizing (Gutmann, 2000). On the qualitative side, Mojola's (2014) recent study of local mobility, livelihoods, and gender in a wetland ecosystem in rural Kenya illustrates the possibility of weaving small but complex changes into a systemwide narrative of significant change. On the quantitative side, approaches like agent-based modeling offer the potential to quantify the consequences of migration in highly complex systems (Parker et al., 2008).

Of course complex models are only as good as their specifications, and there remain many understudied or emerging topics in migration research that require more substantial reduced-form analysis. A few topics merit special consideration. First, while numerous studies have modeled the effects of public goods on migration, few have modeled the effects of migration on the creation of public goods, particularly on the use of remittances in driving the creation of such resources. Second, most existing studies of migration and development have focused on case study areas of modest size, typically covering at most 200 000 residents. Further research must leverage novel data sources to test the hypotheses emerging at the local scale at higher levels of analysis, and perhaps conduct analysis of the impact of migration on linked ecosystems, for instance linking a rural area to an urban or frontier destination. Finally, the rise of broadband mobile communication technologies may fundamentally alter the migration process, increasing the availability to migration-specific social capital and enabling the maintenance of social relationships over long distances while also potentially reducing the need to migrate by

increasing productivity in the agricultural sector and creating opportunities to connect to global economic opportunities without physically moving. History shows that the emergence of new globalizing technologies like mobile communications will not eliminate internal migration, but merely alter its structure and function and new and previously unimaginable ways.

See also: Demographic Impact of Disasters; Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Development and Urbanization; Development: Social-Anthropological Aspects; Global Migration; Internal Migration in Industrialized Countries; International Migration and Remittances; Migration and Development; Migration and Health; Migration and Social Capital; Migration, Economics of; Migration, Theory of; Migration: Anthropological Perspectives; Migration: Motivations; Political Geography; Population Dynamics: Mathematic Models of Population, Development, and Natural Resources; Population Geography; Rural Sociology; Spatial Thinking in the Social Sciences; Urbanization.

Bibliography

Abas, Melanie A., Punpuing, Sureeporn, Jirapramukpitak, Tawanchai, Guest, Philip, Tangchonlatip, Kanchana, Leese, Morven, Prince, Martin, 2009. Rural-urban migration and depression in ageing family members left behind. British Journal of Psychiatry 195, 54–60.

Adhikari, Ramesh, Jampaklay, Aree, Chamratrithirong, Aphichat, 2011. Impact of children's migration on health and health care seeking behavior of elderly left behind. BMC Public Health 11 (143), 1–17.

Alston, Lee J., Libecap, Gary D., Schneider, Robert, 1995. Property rights and the preconditions for markets: the case of the amazon frontier. Journal of Institutional and Theoretical Economics 151 (1), 89–107.

Arenas, Erika, Yahirun, Jenjira, 2011. Left Behind: The effects of offspring's migration on parental mental health in Mexico. California Center for Population Research. PWP-CCPR-2010-060.

Ardington, Cally, Case, Anne, Hosegood, Victoria, 2009. Labor supply response to large social transfers: longitudinal evidence from South Africa. American Economic Journal: Applied Economics 1 (1), 22–48.

Bakewell, Oliver, 2008. Keeping them in their place: the ambivalent relationship between development and migration in Africa. Third World Quarterly 29, 1341–1358.

Bardsley, Douglas K., Hugo, Graeme J., 2010. Migration and climate change: examining thresholds of change to guide effective adaptation decision-making. Population and Environment 32, 238–262.

Barham, Tania, Kuhn, Randall. Staying for benefits: the effect of a health and family planning program on out-migration patterns in Bangladesh. Journal of Human Resources, forthcoming.

Bell, Martin, Charles-Edwards, Elin, 2013. Cross-national Comparisons of Internal Migration: An Update of Global Patterns and Trends. United Nations Population Division Technical Paper No. 2013/1.

Bell, Martin, Charles-Edwards, Elin, Kupiszewska, Dorota, Kupiszewski, Marek, Stillwell, John, Zhu, Yu, 2014. Internal migration data around the world: assessing contemporary practice. Population, Space and Place.

Bell, Martin, Muhidin, Salut, 2009. Cross-national comparisons of internal migration. Human Development Research Paper 2009/30. UN Department of Economic and Social Affairs. Population Division, NY.

Bilsborrow, Richard E., DeLargy, Pamela F., 1990. Land use, migration, and natural resource deterioration: the experience of Guatemala and the Sudan. Population and Development Review 16, 125–147 (Supplement: Resources, Environment, and Population: Present Knowledge, Future Options).

Blumenstock, Joshua E., 2012. Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda. Information Technology for Development 18 (2), 107–125.

Bohra, Pratikshya, Massey, Douglas S., 2011. Individual decisions to migrate during civil conflict. Demography 48 (2), 401–424.

Burt, Ronald S., 2000. The network structure of social capital. Research in Organizational Behavior 22, 345–423.

Carr, David, 2009. Population and deforestation: why rural migration matters. Progress in Human Geography 33 (3), 355–378.

Cassels, Susan, Curran, Sara, Kramer, Randall, 2005. Do migrants degrade coastal environments? Migration, natural resource extraction and poverty in North Sulawesi, Indonesia. Human Ecology 33, 329–363.

Castaldo, Adriana, Deshingkar, Priya, McKay, Andy, September 7, 2012. Internal migration, remittances and poverty: evidence from Ghana and India. In: Migrating Out of Poverty, Research Programme Consortium, Working Paper.

Cernea, Michael M., 2006. Development-induced and conflict-induced IDPs: bridging the research divide. Forced Migration Review 15, 25–27.

Clark, Samuel J., Collinson, Mark A., Kahn, Kathleen, Tollman, Stephen M., 2007. Returning home to die: urban to rural migration and mortality in rural South Africa. Scandinavian Journal of Public Health 35 (Suppl. 69), 35–44.

Curran, Sara, 2002. Migration, social capital and the environment: considering migrant selectivity and networks in relation to coastal ecosystems. In: Lutz, Wolfgang, Prskawetz, Alexia, Sanderson, Warren C. (Eds.), Population and Environment: Methods of Analysis. Population Council, New York, pp. 89–125.

Curran, Sara R., Rivero-Fuentes, Estela, 2003. Engendering migrant networks: the case of Mexican migration. Demography 40 (2), 289–307.

Deane, Kevin D., Parkhurst, Justin O., Johnston, Deborah, 2010. Linking migration, mobility, and HIV. Tropical Medicine and International Health 15 (12), 1458–1463.

Haas, De, Hein, 2010. Migration and development: a theoretical perspective. International Migration Review 44 (1), 227–264.

Ebrahim, Shah, Kinra, Sanjay, Bowen, Liza, et al., 2010. The effect of rural-to-urban migration on obesity and diabetes in India: a cross-sectional study. PLoS Med 7, e1000268.

Ellis, Frank, 1998. Household livelihood strategies and rural livelihood diversification. Journal of Development Studies 35, 1–38.

Elmhirst, Rebecca, 2002. Daughters and displacement: migration dynamics in an Indonesian transmigration area. Journal of Development Studies 38 (5), 143–166.

Fearon, James D., Laitin, David D., 2011. Sons of the soil, migrants, and Civil War. World Development 39 (2), 199–211.

Findley, Sally E., 1994. Does drought increase migration – a study of migration from rural Mali during the 1983–1985 drought. International Migration Review 28 (3), 539–553.

Gemenne, François, 2011. Why the numbers don't add up: a review of estimates and predictions of people displaced by environmental changes. Global Environmental Change 21 (S1), S41–S49.

Gray, Clark, 2009. Environment, land and rural out-migration in the southern Ecuadorian Andes. World Development 37, 457–468.

Gray, Clark, Frankenberg, Elizabeth, Gillespie, Thomas, Sumantri, Cecep, Thomas, Duncan, 2014. Studying displacement after a disaster using large-scale survey methods: Sumatra after the 2004 tsunami. Annals of the Association of American Geographers 104 (3), 594–612.

Gray, Clark, 2012. Gender, natural capital, and migration in the southern Ecuadorian Andes. Environment and Planning A 42, 678–696.

Gubert, Flore, 2000. Migration, Remittances and Moral Hazard. Evidence from the Kayes Area. CERDI. Etudes et Documents 2000-17.

Gutmann, Myron P., 2000. Scaling and demographic issues in global change research. Climatic Change 44, 377–391.

Hagan, Jacqueline, Ebaugh, Helen Rose, 2003. Calling upon the sacred: migrants' use of religion in the migration process. International Migration Review 37 (4), 1145–1162.

Harris, John R., Todaro, Michael P., 1970. Migration, unemployment and development: a two sector analysis. American Economic Review 60 (1), 126–142.

Hunter, Lori, 2005. Migration and environmental hazards. Population and Environment 26, 273–302.

Hunter, Lori M., 1998. The association between environmental risk and internal migration flows. Population and Environment 19 (3), 247–277.

Hwang, S.S., Xi, J., Cao, Y., Feng, X., Qiao, X., 2007. Anticipation of migration and psychological stress and the Three Gorges Dam project, China. Social Science and Medicine 65 (5), 1012–1024.

de Haan, Arjan, 1999. Livelihoods and poverty: the role of migration – a critical review of the migration literature. Journal of Development Studies 36, 1–47.

de Haas, H., 2006. Migration, remittances and regional development in Southern Morocco. Geoforum 37 (4), 565–580.

de Janvry, Alain, Garramon, Carlos, 1977. The dynamics of rural poverty in latin america. Journal of Peasant Studies 5 (2), 206–216.

King, Russell, Skeldon, Ronald, 2010. 'Mind the Gap!' integrating approaches to internal and international migration. Journal of Ethnic and Migration Studies 36 (10), 1619–1646. http://dx.doi.org/10.1080/1369183X.2010.489380.

Knight, John, 2010. Great expectations? The subjective well-being of rural–urban migrants in China. World Development 38 (1), 113–124.
Kuhn, Randall, 2003. Identities in motion: social exchange networks and rural-urban migration in Bangladesh. Contributions to Indian Sociology 37, 311–337.
Kuhn, Randall, 2005. In: Jatrana, S., et al. (Eds.), A Longitudinal Analysis of Health and Mortality in a Migrant-Sending Region of Bangladesh. Migration and Health in Asia. Routledge, London, pp. 318–357.
Kuhn, Randall, 2006. The effects of fathers' and siblings' migration on children's pace of schooling in a migrant-sending region of Bangladesh. Asian Population Studies 2 (1), 69–92.
Kuhn, Randall, Everett, Bethany, Silvey, Rachel, 2011. The effects of children's migration on Elderly Kin's health: a counterfactual approach. Demography 48 (1), 183–209.
Lewis, W. Arthur, 1954. Economic development with unlimited supplies of labor. The Manchester School of Economic and Social Studies 22, 139–191.
Lindstrom, David P., Lauster, Nathanael, 2001. Local economic opportunity and the competing risks of internal and US migration in Zacatecas, Mexico. International Migration Review 354, 1232–1256.
Lu, Yao, 2008. Test of the 'healthy migrant hypothesis': a longitudinal analysis of health selectivity of internal migration in Indonesia. Social Science and Medicine 67 (8), 1331–1339.
Lucas, Robert E.B., 1997. Internal migration in developing countries. In: Rosenzweig, M.R., Stark, O. (Eds.), Handbook of Population and Family Economics. Elsevier, Amsterdam.
Lucas, Robert E.B., Stark, Oded, 1985. Motivations to remit: evidence from Botswana. Journal of Political Economy 93, 901–918.
Lucas, Robert E.B., 2004. Life earnings and rural-urban migration. Journal of Political Economy 112 (S1), S29–S59.
Mackenzie, Peter, 2002. Strangers in the city: the hukou and urban citizenship in China. Journal of International Affairs 56 (1), 305–319.
Martin, Philip L., Edward Taylor, J., 1996. The anatomy of a migration hump. In: Taylor, J. Edward (Ed.), Development Strategy, Employment, and Migration: Insights from Models. OECD Development Centre, Paris, pp. 43–62.
Massey, Douglas S., Axinn, William G., Ghimire, Dirgha J., 2010. Environmental change and out-migration: evidence from Nepal. Population and Environment 32 (2–3), 109–136.
Massey, Douglas S., Williams, Nathalie, Axinn, William G., Ghimire, Dirgha J., 2009. Community services and out-migration. International Migration 48 (3), 1–41.
McKinnish, Terra, 2005. Importing the poor: welfare magnetism and cross-boarder welfare migration. Journal of Human Resources 40 (1), 55–76.
Mendola, Mariapia, 2008. Migration and technological change in rural households: complements or substitutes? Journal of Development Economics 85, 150–175.
Mincer, Jacob, 1978. Family migration decisions. Journal of Political Economy 8 (2), 749–773.
Mojola, Sanyu, 2014. Love, Money, and HIV: Becoming a Modern African Woman in the Age of AIDS. University of California Press, Berkeley, CA.
Mueller, Valerie, Gray, Clark, Kosec, Katrina, 2014. Heat stress increases long-term human migration in rural Pakistan. Nature Climate Change 4 (3), 182–185.
Parker, Dawn C., Entwisle, Barbara, Rindfuss, Ronald R., Vanwey, Leah K., Manson, Steven M., Moran, Emilio, Li, An, Deadman, Peter, Evans, Tom P., Linderman, Marc, Mussavi Rizi, S. Mohammad, Malanson, George, 2008. Case studies, cross-site comparisons, and the challenge of generalization: comparing agent-based models of land-use change in frontier regions. Journal of Land Use Science 3 (1), 41–72. http://dx.doi.org/10.1080/17474230802048151.
Portes, Alejandro, 1998. Social capital: its origins and applications in modern sociology. Annual Review of Sociology 24, 1–24.
Qin, Hua, 2010. Rural-to-urban labor migration, household livelihoods, and the rural environment in Chongqing Municipality, southwest China. Human Ecology 38, 675–690.
Quisumbing, Agnes, McNivel, Scott, 2010. Moving forward, looking back: the impact of migration and remittances on assets, consumption, and credit constraints in the rural Philippines. Journal of Development Studies 46 (1), 91–113.
Raleigh, Clionadh, 2011. The search for safety: the effects of conflict, poverty and ecological influences on migration in the developing world. Global Environmental Change 21S, S82–S93.
Ranis, Gustav, Fei, John C.H., 1961. A theory of economic development. American Economic Review 51, 533–565.
Ravenstein, Ernest G., 1885. The laws of migration. Journal of the Statistical Society of London 48 (2), 167–235.
Reed, Holly E., 2013. Moving across boundaries: migration in South Africa, 1950–2000. Demography 50, 71–95.
Rogaly, Ben, Rafique, Abdur, 2003. Struggling to save cash: seasonal migration and vulnerability in West Bengal, India. Development and Change 34 (4), 659–681.
Rosenzweig, Mark R., Stark, Oded, 1989. Consumption smoothing, migration, and marriage. Journal of Political Economy 97, 905–926.
King, Russell, Skeldon, Roland, 2010. 'Mind the gap!' Integrating approachesto internal and international migration. Journal of Ethnic and Migration Studies 36 (10), 1619–1646. http://dx.doi.org/10.1080/1369183X.2010.489380.
Scudder, Theodore, 1993. Development-induced Relocation and refugee studies: 37 years of change and continuity among Zambia's Gwenbe Tonga. Journal of Refugee Studies 6 (3), 123–152.
Shrestha, Nanda R., Velu, Raja P., Conway, Dennis, July 1993. Frontier migration and upward mobility: the case of Nepal. Economic Development and Cultural Change 41 (4), 787–816 (Published).
Silvey, Rachel, Elmhirst, Rebecca, 2003. Engendering social capital: women workers and rural-urban networks in Indonesia's crisis. World Development 31 (5), 865–879.
Sjaastad, Larry A., 1962. The costs and returns of human migration. Journal of Political Economy 70, 80–93.
Stark, Oded, 1991. The Migration of Labor. Basil Blackwell, Cambridge, MA.
de Sherbinin, Alex, et al., 2008. Rural household demographics, livelihood and the environment. Global Environmental Change 18, 38–53.
de Sherbinin, Alex, Carr, David, Cassels, Susan, Jiang, Leiwen, 2007. Population and environment. Annual Review of Environment and Resources 32, 345–373.
Taylor, J. Edward, 1986. Differential migration, networks, information, and risk. In: Stark, Oded (Ed.), Research in Human Capital and Development: Migration, Human Capital, and Development, vol. 4. JAI Press, Inc., Greenwich, CT, pp. 147–171.
Taylor, J. Edward, Wyatt, T.J., 1996. The shadow value of migrant remittances, income and inequality in a household-farm economy. Journal of Development Studies 32, 899–912.
Taylor, J. Edward, Rozelle, Scott, de Brauw, Alan, 2003. Migration and incomes in source communities: a new economics of migration perspective from China. Economic Development and Cultural Change 52 (1), 75–101.
Van der Gaag, Nicole, Van Wissen, Leo, 2002. Modelling regional immigration: using stocks to predict flows. European Journal of Population 18 (4), 387–409.
VanWey, Leah K., 2005. Land ownership as a determinant of international and internal migration in Mexico and internal migration in Thailand. International Migration Review 39 (1), 141–172.
VanWey, Leah K., Guedes, Gilvan R., D'Antona, Alvaro O., 2012. Out-migration and land-use change in agricultural frontiers: insights from Altamira settlement project. Population and Environment 34, 44–68.
Winters, Paul, Davis, Benjamin, Corral, Leonardo, 2002. Assets, activities and income generation in rural Mexico: factoring in social and public capital. Agricultural Economics 27 (2), 139–156.
Wolpe, Harold, 1975. The theory of internal colonialism: the South African case. In: Oxaal, Ivar, Barnett, Tony, Booth, David (Eds.), Beyond the Sociology of Development. Routledge, London.
World Bank, 2014. Migration and Remittances: Recent Developments and Outlook. Migration and Development Brief 22. http://siteresources.worldbank.org/INTPROSPECTS/Resources/334934-1288990760745/MigrationandDevelopmentBrief22.pdf.
Yiftachel, Oren, Ghanem, As'ad, 2004. Understanding 'ethnocratic' regimes: the politics of seizing contested territories. Political Geography 23 (6), 647–676.
Zelinsky, Wilbur, 1971. The hypothesis of the mobility transition. Geographical Review 61 (2), 219–249.
Zhao, Yaohui, 2003. The role of migrant networks in labor migration: the case of China. Contemporary Economic Policy 21 (4), 500–511.

Relevant Websites

http://perl.psc.isr.umich.edu/chitwanvalleyfamilystudy1.html – Chitwan Valley Family Study.
http://www.knomad.org/ – Global Knowledge Partnership on Migration and Development (KNOMAD).
http://www.gpem.uq.edu.au/image – Internal Migration Around the GlobE (IMAGE).
www.indepth-network.org/ – INDEPTH Network.
http://www.iom.int/ – International Organization for Migration.
http://mmp.opr.princeton.edu/ – Mexican Migration Project.
http://migratingoutofpoverty.dfid.gov.uk/ – Migrating Out of Poverty (UK Department for International Development).
http://www.migrationpolicy.org/ – Migration Policy Institute.
http://esa.un.org/unpd/wpp/Documentation/publications.htm – World Urbanization Prospects.

Internal Migration in Industrialized Countries

Michael J Greenwood, University of Colorado, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Internal migration is defined and distinguished as to its causes and consequences. Most migration research deals with causes. Migration rates are compared across industrialized countries (with Australia, Canada, New Zealand, and the United States shown to have high rates, but with US rates that have declined considerably in recent years), and various migration patterns are described, especially with respect to age and education. Migration rates are shown to fall with age after the early twenties and to rise sharply with education. Different theoretical perspectives on migration are discussed (namely, the disequilibrium and the equilibrium perspective). Moreover, various theories of migration are noted, with the human capital model singled out for special attention. Empirical findings consistent and inconsistent with this model are discussed. Several pressing questions regarding internal migration are offered, and the state of knowledge regarding the answers to these questions is discussed. The article concludes with a discussion of internal migration policies in Europe and in the United States and notes some critical issues concerning these policies.

Residential mobility and migration are pervasive facts of life in most industrialized countries, where migration is a major mechanism through which labor resources are redistributed geographically in response to changing economic and demographic forces. The migration literature may be classified into two areas of research, one dealing with the *determinants* and the other with the *consequences* of migration. Some studies treat both the determinants and consequences within the same framework. Migration research is strongly oriented toward the former.

The *determinants* of migration are the factors that affect migration, including characteristics of both places and persons and their families. This term refers to the qualitative and the quantitative importance of each factor. Place characteristics are specific to a given area, such as employment and wage opportunities, the presence of family and friends, and location-specific amenities. Personal and family characteristics help shape individual and family responses to opportunities that may exist at different locations. The *consequences* of migration refer both to the performance of migrants in their new locations relative to a benchmark (their presumed performance in their former place) and to the impacts that migrants have on others in sending and receiving areas. Traditionally, research has addressed several questions:

1. How many and who migrates?
2. Why do these people migrate?
3. Where are the migrants coming from and where are they going?
4. When do they migrate?
5. What consequences result from migration?

Many studies have aimed primarily at describing migration.

What Is Migration?

The United Nations manual, *Methods of Measuring Internal Migration* (1970), has served as the basis for several definitions of migration and migrants. It proposes the following (p. 2):

> A migration is defined as a move from one migration-defining area to another (or a move of some specified minimum distance) that was made during a given migration interval and that involved a change of residence. A migrant is a person who has changed his usual place of residence from one migration-defining area to another (or who moved some specified minimum distance) at least once during the migration interval.

Comparing Internal Migration Rates in Industrialized Countries

In most countries, people migrate from one place to another in pursuit of increased utility resulting from better employment opportunities, higher wages, and many other factors, but international comparisons of internal migration rates are difficult to make for several reasons.

Various countries define migrants differently in terms of whose movement is being considered, the type of border they must cross, and the interval over which the movement occurs. The method of measuring migration differs widely. In some countries, censuses are the main source of information concerning migration, whereas in others, population registers and other types of administrative records are the major source. The size and shape of the spatial areas between which migration is measured are not uniform either within or between countries.

Migration propensities are sensitive to national economic and demographic conditions, and consequently interregional migration propensities could differ for behavioral reasons.

The most unambiguous method of making cross-national comparisons is to focus on residential moves, but residential mobility includes more than what is commonly regarded as migration. Nevertheless, such measures suggest that rates of mobility vary widely across industrialized countries. Movement rates are high for Australia, Canada, New Zealand, and the United States. A second group of industrialized countries that is far behind these four includes France, the UK, Israel, Japan, Sweden, and Switzerland. Belgium, Ireland, and The Netherlands have the lowest rates of residential mobility.

The Determinants and Consequences of Migration – Overview

Migration Patterns

Many factors contribute to the decision to migrate. Differential characteristics of sending and receiving regions provide a potential incentive for moving, and individual and/or family traits help condition the responses to utility differentials that may arise from these different characteristics. An individual's personal characteristics – including accumulated job skills, general labor market conditions, and employment composition – will help determine the probability of gaining employment during a period of job search. Conditions in land and housing markets also may be important, and state and local taxes and the associated availability of public goods may be critical for certain potential migrants.

A number of life cycle considerations – such as marriage, divorce, completion of schooling, entry into the labor force, start of a career, birth, aging, leaving home of children, home ownership, and retirement – are critical in an individual's or a family's decision to migrate. Other personal circumstances, often related to the life cycle, are also important, such as employment status, earnings, education, accumulated skills, age, job tenure, sex, and health.

One of the most universal mobility relationships is that between age and migration. Migration propensities peak during the early to mid-20s and then decline steadily, with a slight upturn at retirement age in some countries. For example, in the United States, during the 1980s, a person aged 25–29 with 4 years of college education was over twice as likely to make an interstate move as his or her counterpart aged 35–44; over three times more likely to make such a move than someone with comparable education aged 45–64; and over three times more likely to make such a move than a peer in age who had less than 8 years of education.

Studies of the determinants of migration commonly have been formulated in the context of individual utility maximization, with some attention given to the family or the household as the decision-making unit. Models based upon such behavioral foundations have frequently been estimated with aggregate data relating both to migrants and to the determinants of migration. Before the general availability of micro- and longitudinal data, virtually all applied migration research was of necessity based on aggregate data. Although aggregate data were and are limiting in many respects, they did not prevent a boom in migration research during the 1960s and 1970s. Even in the presence of many microdata sets, aggregate data are frequently studied today. Not only are aggregate trends and tendencies of interest is their own right, but for many countries, such aggregate data are all that is available.

During the 1960s, the main thrust of migration research began to take on a more formal tone. Most of the research was not formal in a theoretical sense, but rather intuitively generated hypotheses at first were tested formally in an econometric sense with aggregate data, typically but not always with place-to-place migration data. Such models were frequently specified in the context of modified gravity models. The models are 'gravity type' in that migration is hypothesized to be directly related to the size of relevant origin and destination populations, and inversely related to distance. The models are 'modified' in the sense that the variables of the basic gravity model are given behavioral content, and additional variables that are expected to influence migration are included in the estimated relationship. These studies showed clearly that migration diminishes rapidly with increased distance and that migration is greater between more populous regions; but, migration did not always appear to be from low- to high-wage regions, nor from those with high unemployment rates to those with low rates.

Theoretical Perspectives on Migration

In almost all migration research conducted by economists prior to the late 1970s, the theoretical perspective taken was that of a disequilibrium system. The perspective is called 'disequilibrium,' because migration is assumed to be driven by the existence of a set of nonmarket clearing regional wages. Spatial variations in wages or earnings are assumed to reflect opportunities for utility gains. During more recent years, this disequilibrium perspective has been challenged by proponents of the equilibrium hypothesis, which assumes that spatial variations in wages are compensating and therefore do not reflect opportunities for utility gains.

According to the human capital approach, the potential migrant will select the locality at which the real value of the expected net benefit that accrues to him from migration is greatest. The income that the individual expects to earn at each alternative destination enters importantly into his judgment concerning the benefits associated with each location. The relevant income measure for the individual to consider is the present discounted value of his expected future stream of net pecuniary returns.

Due, in part, to the fairly consistent tendency for empirical studies based on aggregate data to fail to confirm the importance of wages or income in migration decisions, the equilibrium approach has been offered as an alternative to the traditional disequilibrium perspective described above. The equilibrium theorists begin by assuming that households and firms are in proximate equilibrium at any point in time. This assumption means that the marginal household and firm, while maximizing utility and profit, respectively, are spatially arrayed so as to receive zero consumer and producer surplus from their location. Thus, any movement from the general equilibrium configuration cannot improve utility or profit.

According to the equilibrium approach, changes in life cycle factors or generally rising real incomes continuously change the demand for consumer amenities. Because amenities are not evenly distributed spatially, migration occurs and quickly reequilibrates households. Net in-migration to amenity-rich areas tends to drive down wages and drive up the prices of locally produced goods and services and land. In amenity-poor areas, opposite patterns of change occur. Wages and local prices diverge across regions until they just compensate households for the differing amenity bundles that the various regions supply.

Family ties result in negative personal externalities that are usually internalized by the family and thus tend to discourage migration. "Tied persons" in the family are "those whose gains from migration are (in absolute value) dominated by gains (or losses) of the spouse" (Mincer, 1978: p. 753). Presuming that

their joint net returns on migrating from *i* to *j* exceed their joint net costs of migrating, a couple would presumably migrate from *i* to *j*. If, for example, the wife's expected earnings in *j* were less than in *i*, but the husband's were sufficiently greater in *j* than in *i* to offset these losses, the wife would be a 'tied mover.' On the other hand, if the husband's earnings gain in *j* were to fail to offset his wife's earnings loss in *i*, the couple would remain in *i*, and the husband would be a 'tied stayer.' Moreover, according to Mincer, such ties tend to reduce the employment and earnings of those wives who do migrate and to increase the employment and earnings of their husbands. Mincer goes on to show that increased labor force participation rates of women cause an increase in migration ties, which results in both less migration and more marital instability. Increased marital instability in turn encourages migration as well as increased women's labor force participation.

Consequences of Migration

The consequences of migration have been addressed at two levels. The first deals with the migrants themselves, where the emphasis has been on the benefits to migrating, often measured in terms of earnings gains. The second deals with migration's impact on others in the origin and the destination. Do migrants depress local wages in receiving areas and displace local residents from jobs? To the extent that migrants tend to be young and well educated, does migration deprive source regions of critically needed human capital that ensures these regions of long periods of economic stagnation?

Conclusions

Pressing Questions

Who migrates? Why do they migrate? Where do the migrants come from and where do they go? When do they migrate? What consequences does migration have for the migrants themselves and others in origin and destination regions? During the last 30 years, firm answers have been provided to many of these questions, but at the same time new, narrower, and in certain ways more refined, interesting, and certainly more challenging questions have arisen.

Migration propensities appear to vary considerably across industrialized countries for reasons that are not always obvious. Moreover, estimated parameters that address the questions raised above, while generally in qualitative agreement, are quantitatively somewhat different across industrialized countries, even when the models and data are roughly comparable. With regard to internal migration, rigorous cross-national studies are virtually nonexistent, but differences in geographic size (and hence spatial distribution of economic opportunities) and culture are likely to importantly underlie observed differences.

The Human Capital Model

Until about 20 years ago, aggregate data were used almost exclusively to study various migration phenomena. Such data embody a number of shortcomings that prevented the study of many important issues bearing on migration. The relatively recent availability of micro- and longitudinal data has had a major impact on four areas of migration research.

1. Such data have gone far toward clearing up earlier puzzles concerning the relationship between unemployment and migration.
2. They have allowed the human capital model to be tested in the migration context by allowing the estimation of migrant earnings equations.
3. They have permitted a clarification of the relationship between personal characteristics and the decision to migrate, and they have allowed a deeper understanding of the relationship between various life cycle and familial factors and migration.
4. They have permitted a detailed focus on different types of migrants, particularly primary, return, and other repeat migrants.

Researchers now know that local unemployment in the United States has a significant influence on the migration decisions of the unemployed and those who are seeking new jobs, but has little influence on individuals who are secure in their jobs. However, they do not know why this relationship fails to hold for European countries. Moreover, they do not know with great confidence why some unemployed individuals are quicker to migrate than others. To some extent, immediate migration could be discouraged if a person's spouse remains employed. What characteristics of a spouse's employment discourage migration? To what extent does it matter whether the party remaining employed is the husband or the wife? What role does the availability of unemployment insurance play in discouraging immediate migration? Institutional settings differ substantially by country, such as in the nature of unemployment insurance programs, and could cause differences in the response to personal unemployment.

The human capital model has provided a powerful analytical tool for the study of numerous important issues. It is somewhat surprising that this model has not typically provided a comparably powerful explanation of internal migration. Whether migrants are compared to otherwise comparable individuals in the origin or the destination, they appear to suffer earnings losses for at least several years after migration. This finding has been verified repeatedly for the United States and for other countries. Although several explanations have been offered for these findings, we still do not know enough about why they occur. The most likely explanation for the apparent failure of the human capital model is that the model has not failed, but rather it has just not been properly or fully implemented.

Life Cycle Impacts on Migration

The influences of life cycle changes on migration decisions have only barely been touched by researchers. The relationships between many such changes and migration, therefore, have remained undiscovered. Completion of education; birth and aging of children and the anticipation of these events; changing marital status and especially divorce; death of a spouse; acquisition of a home; retirement – these are only a few potentially important life cycle changes that are likely to importantly affect migration in industrialized countries and about which much remains to be learned.

Little is known about the interactions between migration and fertility in the contemporary United States or in other industrialized countries. A sizable literature on this topic began developing in the 1930s, but was mainly focused on rural to urban migration. A good deal of more recent work has also focused on rural to urban migrants in less developed nations, and a few studies concentrate on immigrants to the United States. However, in spite of the availability of more and better data in recent years, little has been done to study how the act of migration affects fertility in the United States or in other industrialized countries.

Migration Policies

Migration serves an equilibrating function in the economy, expediting the balancing of demand and supply forces within and across regions, and thus facilitating the operation of market economies. This facilitating role of migration has attracted the attention of policy makers. In a number of European countries, such as the UK, France, and The Netherlands, migration policy has been tied directly to regional development policy; in others, such as Sweden, migration policy has been oriented toward improving the efficiency of the labor market. In either case, a major objective has been to reduce unemployment. Subsidization of relocation expenses and employment information exchanges are the most common forms that migration policy has taken in Europe.

In the United States, policy concerns regarding migration generally are passive, although this was not the case during the nineteenth century. These concerns have been directed toward both how various national, state, and local policies have affected in- or out-migration and how migration has affected the public sector of states and localities. One policy concern that has arisen in this respect is whether the jobs created by economic development efforts go to previous residents or inward migrants.

See also: International Migration; Migrants and Migration in Europe; Migration and Fertility; Migration, Theory of.

Bibliography

Bartik, T.J., 1993. Who benefits from local job growth: migrants or the original residents? Regional Studies 27, 297–311.

Bell, M., Muhidin, S., 2009. Cross-National Comparisons of Internal Migration. United Nations Development Programme, Human Development Reports. Research Paper 2009/30.

Charney, A.H., 1993. Migration and the public sector: a survey. Regional Studies 27, 313–326.

DaVanzo, J., 1978. Does unemployment affect migration? – Evidence from micro data. Review of Economics and Statistics 60, 504–514.

Greenwood, M.J., 1975. Research on internal migration in the United States: a survey. Journal of Economic Literature 13, 397–433.

Greenwood, M.J., 1997. Internal migration in developed countries. In: Rosenzweig, M.R., Stark, O. (Eds.), Handbook of Population and Family Economics, vol. 1B. Elsevier Science, Amsterdam, pp. 647–720.

Greenwood, M.J., 1985. Human migration: theory, models, and empirical studies. Journal of Regional Science 25, 521–544.

Hunt, G.L., 1993. Equilibrium and disequilibrium in migration modelling. Regional Studies 27, 341–349.

Klaassen, L.H., Drewe, P., 1973. Migration Policy in Europe. Lexington Books, Lexington, MA.

Long, L.H., 1988. Migration and Residential Mobility in the United States. Russell Sage Foundation, New York.

Long, L.H., 1991. Residential mobility differences among developed countries. International Regional Science Review 14, 133–147.

Mincer, J., 1978. Family migration decisions. Journal of Political Economy 86, 749–773.

Molloy, R., Smith, C., Wozniak, A., 2011. Internal migration in the United States. Journal of Economic Perspectives 25, 173–196.

Nam, C.B., Serow, W.J., Sly, D.F. (Eds.), 1990. International Handbook on Internal Migration. Greenwood, New York.

Plane, D.A., 1993. Demographic influences on migration. Regional Studies 27, 375–383.

United Nations, 1970. Methods of Measuring Internal Migration. United Nations, New York. Manual VI.

Internal Reconstruction in Linguistics

Anthony Fox, University of Leeds, Leeds, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8942–8945, © 2001, Elsevier Ltd.

Abstract

After defining internal reconstruction as the method of establishing earlier, unattested forms of languages without reference to 'external,' especially comparative, evidence, and sketching the history and development of the method, the article identifies a number of related but distinct procedures which fall under this heading: historical morphophonemics, regularization of systems, and universal and typological reconstruction. Each of these is explained and exemplified in turn. Historical morphophonemics involves identifying alternations between grammatically or lexically equivalent forms, resulting from phonological changes in different contexts, and reducing them to a single form, thus reversing the changes. Regularization of systems is based on the same principle in so far as it attempts to reverse changes by eliminating anomalies, but it extends this principle to cover whole systems. Universal and typological reconstruction works backward by invoking known or assumed general processes of linguistic change. The criteria for reconstruction here may be either universal or typological, the former relying on universal characteristics of language or of language change, the latter on the observation that languages fall into a number of more or less distinct types, characterized by co-occurring features. In the latter case reconstruction is effected by restoring the assumed typological consistency of the earlier language. The justification for the different procedures is discussed, and their effectiveness and reliability evaluated.

Internal reconstruction is a method for establishing earlier, unattested forms of languages without reference to 'external,' especially comparative, evidence from other languages. In this it differs from the comparative method (*see* Comparative Method in Linguistics), which is based on such evidence, though the aims of both methods are in practice the same. To signal the difference, forms or languages reconstructed by internal reconstruction are designated 'pre-forms,' 'pre-languages,' etc., as opposed to the 'proto-forms' and 'proto-languages' reconstructed by the comparative method. The method can be applied to individual languages in cases where comparative evidence is not available, for example to isolated languages or to reconstructed proto-languages.

Internal reconstruction is an extremely powerful method, and for this reason some of its results can be both speculative and controversial, though a number of the reconstructions arrived at by this method which were earlier deemed too hypothetical have subsequently become accepted, for example the postulation of a set of 'laryngeal' consonants for Indo-European (the 'laryngeal theory'), whose existence was corroborated following the discovery of Hittite (Lindeman, 1987). As in the case of the comparative method, it evolved in practice before it was formally recognized. However, since the method rests on a quite sophisticated formal analysis of language, it did not appear until the late nineteenth century, much later than the comparative method. In a youthful work, for instance, Saussure used aspects of what would now be regarded as internal reconstruction in his study of the Indo-European vowels. However, the method was not firmly established until the 1940s, principally in the work of American structuralist linguists such as Henry Hoenigswald.

Internal reconstruction is not a single, unitary method; the term embraces a number of related but distinct procedures, which we may designate as follows:

1. historical morphophonemics;
2. regularization of systems; and
3. universal and typological reconstruction.

All of these have in common the attempt to work backward in time by reversing changes that are assumed to have taken place in the language in question, but they differ in scope and in the factors that are taken into account.

Historical Morphophonemics

This is the most limited of the methods of internal reconstruction. It applies only to the reconstruction of phonology, and works by identifying phonological alternations between grammatically or lexically related forms resulting from the differential operation of phonological changes in different linguistic contexts. By eliminating these alternations and reducing the set of alternants to a single form, it effectively reverses the changes and establishes the form that existed prior to their taking effect. In practice, it is equivalent to the synchronic process of establishing 'morpho- phonemes,' or the 'underlying forms' of classical Generative Phonology (*see* Phonology; Morphophonology) – abstract phonological entities from which the occurring forms can be derived – though it interprets these morphophonemes as historical antecedents.

Consider, for example, the Latin forms given in eqn [1].

Nominative	Genitive	
consul	consulis	('consul')
genus	generis	('kind')
corpus	corporis	('body')

[1]

As we see from *consul*, the genitive is formed by adding the affix *-is* to the stem of the word, but in the case of *genus* and *corpus* the final consonant of the stem is *-s* in the nominative and *-r* in the genitive. In synchronic phonology we might regard these two sounds as realizations of the same element, a 'morphophoneme' which appears as /s/ at the end of the word and as /r/ between vowels. But the alternation can also be

interpreted historically, if we assume that this morphophoneme was actually an original phoneme, assumed to be /s/, which was changed to /r/ between vowels. We can therefore postulate earlier forms of the genitive: **genesis* and **corposis*.

This method works, therefore, by attempting to undo the effects of the splitting of phonemes in different contexts. Here, the original phoneme /s/ split into two separate phonemes, /s/ and /r/, depending on the context in which it occurred. This split resulted in a synchronic alternation between the phonemes /s/ and /r/ in different forms of the word. By identifying such alternations, and deriving the alternating phonemes from a single source, we reconstruct the earlier form of the words.

Regularization of Systems

Like historical morphophonemics, this method attempts to reverse changes by eliminating anomalies, but it extends this principle to cover whole systems of linguistic elements.

Consider, for example, the so-called 'strong' verbs of the Germanic languages. These verbs form their principal parts not by the addition of affixes (*walk-walked*; *love-loved*, etc.), but by modifying the vowel of the stem (*sing-sang-sung*; *ride-rode-ridden*, etc.). Strong verbs fall into a number of classes, according to the pattern of alternation of the vowels; only the first three are relevant here. For these verbs, we can identify three stems, whose vowels can be reconstructed by means of the comparative method, for the Germanic languages as a whole: as in eqn [2].

ClassI	*i* :	*ai*	*i*	
ClassII	*iu*	*au*	*u*	[2]
ClassIII	*i*/*e*	*a*	*u*	

It can be seen that there is a certain regularity here, though also some anomalies. By the simple expedient of regarding the *i*: of the first stem of Class I (the colon indicates length) as *ii*, the first two classes become entirely parallel, with the vowel pattern *i–a–zero*, where the vowel is followed by *i* in Class I and *u* in Class II; in the third stem, the *i* or *u* is left as the only vowel. In Class III verbs, the vowel is followed by a nasal or liquid consonant (*r* or *l*), and the *i*/*e* alternation in the first stem depends on the following consonant: *i* occurs before a nasal and *e* before other consonants. We can therefore eliminate this alternation; the standard reconstruction assumes an *e* here, which is also generalized to the first stem of Classes I and II (*e* became *i* before the *i* and *u* of these classes as well as before the nasal of Class III). Thus, all three classes can be reconstructed with the same pattern (*e-a*) for the first two stems.

There remains the third stem, for which, as we have seen, we assume a 'zero vowel' for Classes I and II. If the pattern is the same for all three classes, Class III should have no vowel at all here, though in fact we find *u*. This *u* thus alternates with the *zero* of Classes I and II, and we can reconstruct a single source for both, which, following the first two classes, is taken to be *zero*. The nasal or liquid consonant of the stem of this class is assumed to have been 'syllabic' (i.e., it had the role of a vowel, as with the *n* or *l* of such English words as *bitten* (bitn) or *bottle* (botl)). It is presumed that these syllabic consonants later developed *u* before them. The result of these procedures is an entirely regular system, with the pattern *e–a–zero*, the differences among these three classes being in the sound which followed the vowel: *i* in Class I, *u* in Class II, and a nasal or liquid consonant in Class III. Such a regularized system can then be regarded as existing at an earlier stage of the language.

It is legitimate to object, of course, that, carried to its logical conclusion, this procedure results in the elimination of all anomalies and irregularities, and thus assumes that earlier languages had no such irregularities – an unlikely state of affairs. However, it should be borne in mind that all reconstruction, by whatever method, produces an idealization rather than an attested reality.

Universal and Typological Reconstruction

This method attempts to work backward by invoking known or assumed general attributes of language or processes of linguistic change. This is the most powerful, but also the most speculative form of internal reconstruction, since the evidence we have of such general processes of change is often of a theoretical kind.

The criteria for reconstruction here may be either universal or typological. In the former case, changes are assumed to have occurred which reflect universals, i.e., characteristics of all languages, though these may be diachronic (universal processes of change) or synchronic (universal characteristics of language states which constrain the processes of change). If we therefore encounter a particular feature in the language under investigation which is known to be the result of a certain universal process, then we may assume this process to have occurred and reinstate the form from which the feature may have derived.

Universal processes or constraints have been recognized for all components of linguistic structure: phonology, morphology, syntax, and the vocabulary. One such process, which is currently receiving considerable attention, is grammaticalization (*see* Grammaticalization), the 'downgrading' of lexical words to grammatical words and affixes (Hopper and Traugott, 1993); a similar and related process applies in the lexicon, by means of which words are progressively 'bleached' of semantic content. Recognition of such universals allows us to determine the direction of language change, and thus to reconstruct earlier stages in conformity with this direction. Thus, if we assume that all grammatical affixes are derived from original lexical words, then a reconstruction which posits grammatical morphemes as earlier stages of lexical morphemes rather than the reverse would be ruled out.

Reconstruction based on typological criteria relies on the observation that languages fall into a number of more or less distinct types, characterized by co-occurring or harmonizing features (*see* Linguistic Typology). A standard typological grouping, for example, divides languages into VO and OV types, according to whether the object (O) follows or precedes the verb (V). A number of other features of sentence structure harmonize with this: in VO languages (e.g., French) the adjective typically follows the noun, and prepositions are used;

in OV languages (e.g., Japanese) the adjective typically precedes the noun, and postpositions are used.

To apply the method, we identify the language as belonging to a particular type on the basis of some major feature, and, making the further assumption that languages are ideally typologically consistent, bring the remaining harmonizing features into line. If, for example, we find a language (such as English) which is VO but has the order adjective + noun, we will assume that at an earlier stage the language will have either been OV (which is assumed for English) or will have had the reverse order of noun and adjective. Various different typological parameters have been put forward, but their use in reconstruction generally follows the above procedure. Again there are potential weaknesses in the assumptions underlying the method, since most languages are typologically inconsistent to some degree, and it is by no means proven that typological consistency represents an ideal or original state for languages.

Conclusion

The methods of internal reconstruction are extremely powerful, but they rest on a range of assumptions about both the nature of languages and the nature and direction of language change. The accuracy and reliability of the results inevitably depend on the adequacy of these assumptions. Not surprisingly, the results of the method have often been challenged as the assumptions themselves have been called into question.

See also: Comparative Method in Linguistics.

Bibliography

Campbell, L., 1998. Historical Linguistics. An Introduction. Edinburgh University Press, Edinburgh (Chapter 8).

Fox, A., 1995. Linguistic Reconstruction. An Introduction to Theory and Method. Oxford University Press, Oxford (Chapters 7, 8).

Greenberg, J.H., 1963. Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, J.H. (Ed.), Universals of Language. MIT, Cambridge, MA, pp. 73–113.

Hock, H.H., 1986. Principles of Historical Linguistics. Mouton de Gruyter, Berlin (Chapter 17).

Hopper, P.J., Traugott, E.C., 1993. Grammaticalization. Cambridge University Press, Cambridge, UK.

Lehmann, W.P., 1992. Historical Linguistics, third ed. Routledge, London (Chapter 8).

Lindeman, F.O., 1987. Introduction to the 'Laryngeal Theory.' Norwegian University Press, Oslo, Norway.

Internal Validity

Charles S Reichardt, University of Denver, Denver, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.M. Mark and C.S. Reichardt, volume 11, pp. 7749–7752, © 2001, Elsevier Ltd.

Abstract

The effect of a treatment is assessed by comparing what happened when the treatment was present to what happened when the treatment was not present. Anything else in such a comparison that varies along with the treatment is called a treatment confound and can bias the estimate of the treatment effect. Threats to internal validity are treatment confounds. Conclusions about treatment effects are said to be internally valid if they are unbiased by threats to internal validity.

Campbell (1957) introduced the notion of internal validity by contrasting it to external validity, and both types of validity were later detailed further in Campbell and Stanley (1963, 1966). In these early works, internal validity was said to address the question of whether a treatment had an effect in a given study, while external validity was said to address the question of whether the treatment effect in a given study could be generalized to treatments, outcomes, recipients, and settings beyond those in the given study. These definitions were widely accepted as gospel across the social sciences until Cook and Campbell (1976, 1979), not being satisfied with the diptych of types of validity, introduced two new types, statistical conclusion validity and construct validity. Correspondingly, Cook and Campbell also adjusted the definitions of internal and external validity to fit the more crowded landscape. Believing the concept was often misunderstood and therefore misused, Campbell (1986) subsequently suggested that internal validity be relabeled 'local molar validity' to emphasize that internal validity concerns only the results of a specific (local) study and that, in assessing its effects, the treatment was to be taken as a single undifferentiated (molar) manipulation. However, luckily, the awkward label of 'local molar' validity never caught on.

The latest revision of the Campbellian typology is by Shadish et al. (2002), who have maintained the distinction among the four types of validity but have once again somewhat modified the definitions. According to this latest version, internal validity concerns the presence or absence of a causal relationship between the treatments as manipulated and the outcomes as measured in a specific study; construct validity concerns which higher order constructs are represented by the treatments, recipients, settings, and outcomes in the given study; and external validity concerns whether the causal relationship between the treatments as manipulated and outcomes as measured holds across variations in treatments, outcomes, recipients, and settings.

Each type of validity can suffer from unique 'threats' to validity. Threats to internal validity are treatment confounds. The next sections define treatments confounds and the types of treatment confounds that threaten internal validity and explain how to either avoid or remove threats so as to achieve internal validity.

Threats to Internal Validity

The effect of a treatment (as compared either to no treatment or to an alternative treatment) is the difference between what would have happened at time 2 if the treatment had been implemented at time 1 and what would have happened at time 2 if either no treatment or an alternative treatment had been implemented at time 1, but everything else at time 1 had been the same. Such a comparison is called the 'ideal' comparison. As just noted, the effect of a treatment can be defined in reference to a comparison either between a treatment and no treatment or between a treatment and an alternative treatment. For ease of expression, but without loss of generality, the author sometimes refers to comparisons between a treatment and no treatment and sometimes to comparisons between a treatment and an alternative treatment. The same results hold in either case.

The ideal comparison, which is a comparison between what happens after a treatment is implemented and what happens after that treatment is not implemented but everything else is the same, is impossible to obtain in practice. In any comparison that can be obtained in practice, something else must vary along with the treatment. Whatever else varies along with the treatment is called a treatment confound. Treatment confounds can introduce biases in the estimates of treatment effects. Threats to internal validity are certain types of treatment confounds.

To make the definition of a treatment confound more concrete, consider two types of comparisons often used in practice to assess treatment effects. The first type of comparison draws a contrast between outcomes obtained at different points in time. The ideal comparison requires that a treatment both be given and not be given to the same people (who will be called recipients) at the same point in time. Although such a comparison is impossible to obtain in practice, it is possible to give a treatment and not give a treatment to the same recipients at different points in time. In particular, it is possible to compare what happens before a treatment is implemented to what happens after the treatment is implemented, which is called a 'before–after' comparison. In such a comparison, differences in performances between the before and after assessments are attributed to the effect of the treatment. Anything, in addition to the treatment, that occurs between the time of the 'before' assessment and the time of the 'after' points is confounded with the treatment. Like the treatment, a confound can cause a difference between the before and after assessments. Therefore, a confound could either counteract or supplement the effect of the treatment, so the effect of the treatment is either over- or underestimated by the before–after comparison.

The second type of comparison draws a contrast between different recipients. In such a comparison, a treatment and an

alternative treatment are implemented at the same point in time, but with different recipients. That is, one set of recipients receives the treatment, while, concurrently, a different set of recipients receives an alternative treatment. Subsequent differences in performance between the two treatment groups are attributed to the effect of the treatment. Such a comparison is called a between-group comparison. Anything, in addition to the treatment, that varies across the two sets of recipients is confounded with the treatment. Like the treatment, a confound can cause a difference between the outcomes of the treatment groups. Therefore, a confound could either counteract or supplement an effect of the treatment, so the effect of the treatment is either over- or underestimated by the difference in outcomes between the treatment groups.

A variety of threats to internal validity can arise due to confounds in each of the two types of comparisons just described above. The most prominent threats to internal validity are described below.

Maturation

Due to natural maturational processes, people can change over time in any number of ways. People can become more or less knowledgeable, more or less attentive, more or less tired, more or less hungry, and so on. Campbell (1957), who first introduced the notion of internal validity, called such changes 'maturation.' If the effect of a treatment is assessed in a before–after comparison (i.e., by comparing how recipients perform before a treatment is introduced to how they perform after the treatment is introduced), the treatment is confounded with whatever maturational changes occur between the time of the before and after assessments. Maturational changes can cause a change in the after assessment compared to the before assessment, and hence maturational changes can cause the treatment to look either more or less effective than it really is. Hence maturational changes are a threat to internal validity.

History

In a before–after comparison (where the performance before a treatment is introduced is compared to performance after a treatment is introduced), the treatment is an external event imposed on the recipients. Other external events, besides the treatment, can occur between the time of the before and after assessments and impinge on the recipients. Such external events are threats to internal validity because they can cause a change in the after assessment compared to the before assessment and hence cause the treatment to look either more or less effective than it really is. The effects of external events besides the treatment are called 'history' effects. As such, history is a confound and a threat to internal validity.

Instrumentation

In a before–after comparison, the measuring instrument used to assess performance before the treatment is introduced might differ in some way from the measuring instrument that is used to assess performance after the treatment is introduced. In the extreme, two separate measuring instruments might be used for the before and after assessments. The before and after instruments might objectively be the same but might be implemented in slightly different ways. Any change in the measuring instruments or in how the measuring instruments are implemented is called a change in 'instrumentation.' A change in instrumentation is a treatment confound and could bias the estimate of the treatment effect. Hence a change in instrumentation is a threat to internal validity.

For example, consider a 10-week long intervention intended to improve children's social skills where the pretest and posttest assessments are based on parental reports of their children's behavior. After attending the program for several weeks with their children, the parents might alter their perceptions of their children's behavior. Because of the program, the parents might come to recognize, say, problems in their children's behavior that they did not recognize before. As a result, the parents might rate behavior on the 'after' assessment more negatively than they would have rated the same behavior on the before assessment. Such a change would make the program look less effective than it really is. Because the parents are the instrument by which behavior is rated and because the way in which parents rate behavior changes, the bias in the estimate of the treatment effect is due to a change in instrumentation.

Testing

In a before–after comparison, the mere act of collecting a 'before' assessment can have an effect on the 'after' assessment. Such an effect is called a 'testing' effect. A testing effect can be confounded with the treatment and thereby be a threat to internal validity. For example, in an intervention designed to improve children's social skills, a pretest assessment that asks parents to rate their children's social skills might itself have an effect on the children's social skills, quite apart from the effects of the intervention. That is, simply by rating their children on social skills during the pretest assessment, parents might be led to recognize deficiencies in their children's social skills, and this might motivate the parents to improve their children's social skills even without the intervention program.

Testing effects are often confused with instrumentation effects. A testing effect is an effect of the assessment process on subsequent outcomes, while an instrumentation effect is an effect of a change in the assessment process. When testing has an effect, a real change has occurred in the behavior being assessed, due to the process of testing or assessment. Instrumentation effects result only in a change in the recorded level of the assessed behavior but not in a real change in the level of assessed behavior.

Seasonality

The different seasons (winter, spring, summer, and fall) can induce different behaviors. The same holds for different days of the week and different times of the day. Effects due to differences in seasons, days of the week, times of day, or any other cyclical variations are collectively called 'seasonality' effects. If the before assessment, in a before–after comparison, is made during a different season, day of the week, or time of day than the after assessment, part of the difference between the two assessments could be due to seasonality effects and not just the

effects of the treatment. Hence seasonality effects can bias the estimate of treatment effects and be a threat to internal validity.

Attrition

Some recipients might fail to complete an assessment in a before–after comparison because they drop out of the treatment or for some other reason. If the before assessment in a before–after comparison is based on the scores of all the recipients, while the after assessment is based on the scores of only those recipients who complete the program or complete the after assessment, observed differences between the before and after assessments might be due not just to treatment effects but also to differences in the two different groups of recipients. If those who drop out would have performed relatively poorly on the after assessment, the resulting bias in the estimate of the treatment effect will make the treatment look more effective than it really is. Conversely, if those who drop out would have performed relatively well on the after assessment, the resulting bias in the estimate of the treatment effect will make the treatment look less effective than it really is. When recipients drop out of either the treatment or the measurement process, the loss is called 'attrition.' As just demonstrated, attrition can be a confound and a threat to internal validity.

Regression toward the Mean

Recipients in a before–after comparison are sometimes recruited into a study based on their performance on the before (or pretest) assessment. For example, students who score poorly on a pretest assessment of cognitive abilities might be recruited to receive a remedial education program. Conversely, students who score well on a pretest assessment might receive benefits and rewards, such as a scholarship. A posttest assessment could then be conducted to determine the effects either of the remedial program or of the benefits and rewards.

Selecting recipients to receive a treatment based on either high or low scores on a pretest can introduce a bias due to regression toward the mean. When recipients are selected to receive a treatment because they have relatively low scores on a pretest, regression toward the mean causes the recipients' posttest scores to be relatively higher. When recipients are selected to receive a treatment because they have relatively high scores on a pretest, regression toward the mean causes the recipients' posttest scores to be relatively lower.

Regression toward the mean occurs because some of the causes of either high or low scores on the pretest assessment are likely the result of temporary factors that tend not to be as extreme on the posttest. For example, temporary factors that can cause a pretest score to be either relatively high or low include such things as obtaining high or low scores on a pretest because of either particularly good or poor guesses on a test, feeling under the weather or feeling particularly energized at the time of the pretest, being distracted by extraneous influences or being particularly attentive because of a good breakfast or welcome news received the morning of the pretest, and so on. However, such factors are not as likely to be as high or as low on the posttest as they were on the pretest because these factors were not selected to be either high or low on the posttest as they were on the pretest. If the factors are not selected to be either high or low, they are more likely, everything else being the same, to be middling rather than either high or low. As a result, recipients selected for low scores on the pretest will tend to have higher scores on the posttest even if the treatment has no effect. Conversely, recipients selected for high scores on the pretest will tend to have lower scores on the posttest even if the treatment has no effect. As long as recipients are selected for either high or low scores on a pretest and as long as the correlation between pretest and posttest scores is not perfect, regression toward the mean will tend to make the average posttest score closer to the mean than the average pretest score, independently of any treatment effect. Hence regression toward the mean is a threat to internal validity.

Carryover Effects

Consider a comparison in which two alternative treatments are given, one after the other, to the same group of recipients. In that comparison, the performance of the recipients is assessed after each treatment is given, and the relative effects of the two treatments are assessed by contrasting these two performances. If the effects of the treatments are short lived so the effect of the first treatment has worn off before the second treatment is implemented, the effect of the first treatment will not influence the outcomes on the second assessment. However, if the effects of treatments are not short lived, the effect of the first treatment might still be present when the second treatment is implemented. As a result, a comparison of the outcomes of the two treatments can be biased by the continuing effects of the initial treatment. Such a bias is said to be due to carryover effects. Carryover effects are a threat to internal validity.

For example, consider comparing two methods of teaching children to ride a bicycle. If the first method is effective, the children will not have forgotten how to ride a bicycle by the time the second method is introduced. Therefore, even if the first method is effective and the second method is not, there will be no difference in performance following the two treatments. The success of the first method of instruction will have carried over to influence the assessment of the second method of instruction.

Selection Differences

As noted above, in a between-group comparison, one group of recipients receives a treatment, a different group of recipients receives an alternative treatment, and subsequent differences in outcomes between these two groups are attributed to differences in the effects of the treatments. No two groups of recipients are identical. One group will be smarter and one group will not be as smart, on average, even if the difference is only very slight. Similarly, one group will be taller and one group will be shorter, on average. One group will be more attentive and one group will be less attentive, on average, and so on. Whatever differences exist between the two groups of recipients before the different treatments are introduced in a between-group comparison are called selection (or baseline) differences. Selection differences are a threat to internal validity because a difference in outcomes between the treatment groups could be due either to the effects of the treatment or to the effects of selection differences.

Interactions with Selection

In a between-group comparison, the effects of selection differences can interact with the effects of other threats to internal validity. For example, history is a threat to internal validity in a before–after comparison that arises when an external event besides the treatment occurs in between the time of the before and after assessments. A bias can be introduced in a between-group comparison if different external events, besides the different treatments, are introduced in the two different treatment groups. Such a bias is said to be due to local history effects or to a selection by history interaction.

Similarly, attrition can be a threat to internal validity in a before–after comparison when some recipients drop out of the study. A bias can be introduced in a between-group comparison if different types of recipients drop out of the different treatment groups. Such a bias is said to be due to differential attrition or to a selection by attrition interaction.

Other threats to internal validity in a before–after comparison–including threats due to maturation, testing, instrumentation, seasonality, and regression toward the mean–can also introduce biases in a between-group comparison when the effects of the threats operate differentially across the groups. In such cases, the threats are said to interact with selection.

Crossover Effects

In a between-group comparison, recipients who are supposed to receive one of the two treatments might manage in some way to receive the other treatment instead. Such unauthorized reassignment of treatments is called treatment 'crossover' and is a threat to internal validity. For example, if an innovative treatment is more effective than a standard treatment, if recipients tend to cross over from the standard treatment to the innovative treatment, and if the recipients who cross over tend to be those who would perform poorly no matter which treatment they receive, crossovers will tend to make the innovative treatment appear less effective than it really is, and vice versa.

Coping with Threats to Internal Validity

To estimate treatment effects without bias, researchers must either avoid the effects of threats to internal validity so they are never present or remove them when they are present. A variety of design features and statistical procedures have been developed either to avoid or remove the effects of threats to internal validity (Reichardt, 2000; 2009). A few of the most common design features and statistical procedures are described below.

Interrupted Time-Series Designs

As noted above, a simple before–after design is susceptible to a number of threats to internal validity. One way to cope with those threats to validity is to add design features to embellish the comparison. One slightly more elaborate design includes multiple observations spread out at regular intervals both before and after the treatment is implemented. In such a design, the effect of the treatment is no longer estimated as the difference between a single pretest and a single posttest observation. Instead, the effect of the treatment is estimated by assessing the trend across time that exists in the pretest observations, projecting that trend forward in time, and comparing the projected trend to the actual trend in the posttest observations. Discrepancies are attributed to the effect of the treatment. Such a design is called an interrupted time-series (ITS) design.

Maturation is no longer a threat to validity in an ITS design because the effect of maturation, if modeled correctly by the trend in the pretest observations, is taken into account and removed. Similarly, the effects of testing are typically rendered implausible because testing effects are unlikely to continue across the multiple pretest observations that are collected before the treatment is introduced. Regression to the mean is also rendered implausible to the extent selection into the treatment is based on one of the earliest pretest observations.

An ITS design can be elaborated further by adding a comparison time series of observations. The recipients who receive the treatment provide the experimental time series of observations. Recipients who receive an alternative treatment (or no treatment) provide a comparison time series of observations. A treatment effect is estimated in each time series. The differential effect of the treatments is estimated as the difference between the estimates from the two time series. The effects of history, instrumentation, and seasonality are controlled in such an estimate to the extent these effects are the same in the two time series of observations. Further design embellishments, such as adding nonequivalent measures and treatment variations, are also possible and correspondingly can help rule out additional threats to validity (Shadish et al., 2002).

Random Assignment to Treatment Conditions

The purpose of randomly assigning recipients to treatment conditions in a between-group comparison is to cope with the biasing effects of selection differences. Contrary to what is sometimes stated, random assignment to treatments does not remove all selection differences. Rather than completely removing selection differences, random assignment makes selection differences random. The effects of these random selection differences can then be assessed using the classic statistical procedures of hypothesis tests and confidence intervals. To the extent the differences between the groups on the outcome measure are greater than chance effects due to the random selection differences, the treatment is estimated to have an effect above and beyond the effects of selection.

Another advantage of random assignment is that the effects of random selection differences can be made as small as desired simply by increasing the number of recipients (i.e., the sample sizes) in the treatment groups. As a result, random assignment can be a powerful means of addressing the threat to validity of selection differences. However, random assignment does not control all threats to internal validity. Differential attrition, local history, and crossover effects can still plague randomized experiments if researchers are not vigilant in avoiding these sources of bias.

Regression-Discontinuity Designs

Instead of assigning recipients to treatment conditions at random, recipients could be assigned using a quantitative

measure of need or merit. Those recipients whose score on the quantitative measure falls below a specified cutoff value would be assigned to one treatment group, while those recipients whose score on the quantitative measure falls above the cutoff value would be assigned to the other treatment group. The effect of the treatment is assessed by regressing the outcome scores onto the quantitative measure in each treatment group to produce a regression line in each group. A break in these regression lines at the cutoff value is attributed to the effect of the treatment. Because the analysis of the data looks for a discontinuity between the two regression lines, the design is called a regression-discontinuity (RD) design.

The RD design takes account of the effects of selection differences by modeling their effects much as an ITS design takes account of the effects of maturation (Marcantonio and Cook, 1994). An RD design can sometimes be implemented when a randomized experiment is not feasible because of administrative or ethical constraints. The disadvantage of the RD design, compared to a randomized experiment, is that the estimate of the treatment effect in the RD design requires additional assumptions and is less precise.

Matching and Propensity Scores

When recipients are assigned to two treatment conditions neither at random nor according to the specifications of an RD design, a between-group comparison becomes a nonequivalent comparison group (NECG) design. Estimates of treatment effects from an NECG design are generally less credible than estimates from either randomized experiments or RD designs because researchers cannot be as certain that the effects of selection differences have been properly modeled and removed. One of the most credible methods for taking account of the effects of selection differences in an NECG design involves matching the recipients in the different treatment groups using propensity scores (Rosenbaum and Rubin, 1983). As with ITS designs, a wide range of design and statistical embellishments can be added to between-group comparisons to take account of additional threats to validity, such as due to differential attrition, local history, and crossover effects.

Conclusion

Although firmly entrenched in textbooks on research methods across the social sciences, the Campbellian conceptualization of validity, across its evolution, has oft been challenged and criticized (Cronbach, 1982; Kruglanski and Kroy, 1975; Reichardt, 2006, 2011). One set of disputes about the Campbellian typology concerns priorities among the four types of validity. An inherent premise of the Campbellian perspective, which has been consistent over time, is that internal validity is the most important of the four types for causal inference. Without internal validity, satisfying the other three types of validity hardly advances causal inference at all. In contrast, Cronbach has argued that external validity is paramount. However, regardless of relative priorities, no serious scholar of research design doubts that internal validity is a central and fundamental building block of sound causal inference.

See also: Causal Counterfactuals in Social Science Research; Control Variables in Research; Experimenter and Subject Artifacts: Methodology; External Validity; Generalization: Conceptions in the Social Sciences; Interrupted Time Series Quasi-Experimental Designs; Nonequivalent Group Designs; Random Assignment: Implementation in Complex Field Settings; Regression Discontinuity Design; Scaling and Classification in Social Measurement.

Bibliography

Campbell, D.T., 1957. Factors relevant to the validity of experiments in social settings. Psychological Bulletin 54, 297–312.

Campbell, D.T., 1986. Relabeling internal and external validity for applied social scientists. New Directions for Program Evaluation, no. 31. In: Trochim, W.M.K. (Ed.), Advances in Quasi-experimental Design and Analysis. Jossey-Bass, San Francisco, pp. 67–77.

Campbell, D.T., Stanley, J.C., 1963. Experimental and quasi-experimental designs for research on teaching. In: Gage, N.L. (Ed.), Handbook of Research on Teaching. Rand McNally, Chicago, pp. 171–246.

Campbell, D.T., Stanley, J.C., 1966. Experimental and Quasi-experimental Designs for Research. Rand McNally, Chicago.

Cook, T.D., Campbell, D.T., 1976. The design and conduct of quasi-experiments and true experiments in field settings. In: Dunnette, M.D. (Ed.), Handbook of Industrial and Organizational Research. Rand McNally, New York, pp. 223–326.

Cook, T.D., Campbell, D.T., 1979. Quasi-experimentation: Design and Analysis Issues for Field Settings. Rand McNally, Chicago.

Cronbach, L.J., 1982. Designing Evaluations of Educational and Social Programs. Jossey-Bass, San Francisco, CA.

Kruglanski, A.W., Kroy, M., 1975. Outcome validity in experimental research: a re-conceptualization. Journal of Representative Research in Social Psychology 7, 168–178.

Marcantonio, R.J., Cook, T.D., 1994. Convincing quasi-experiments: the interrupted time series and regression-discontinuity designs. In: Wholey, J.S., Hatry, H.P., Newcomer, K.E. (Eds.), Handbook of Practical Program Evaluation. Jossey-Bass, San Francisco, pp. 133–154.

Reichardt, C.S., 2000. A typology of strategies for ruling out threats to validity. In: Bickman, L. (Ed.), Research Design: Donald Campbell's Legacy, vol. 2. Sage, Thousand Oaks, CA, pp. 89–115.

Reichardt, C.S., 2006. The principle of parallelism in the design of studies to estimate treatment effects. Psychological Methods 11, 1–18.

Reichardt, C.S., 2009. Quasi-experimental design. In: Millsap, R.E., Maydeu-Olivares, A. (Eds.), The SAGE Handbook of Quantitative Methods in Psychology. Sage, Thousand Oaks, CA, pp. 46–71.

Reichardt, C.S., 2011. Criticisms of and an alternative to the Shadish, Cook and Campbell Validity Typology. New Directions for Evaluation, no. 130. In: Chen, H.T., Donaldson, S.I., Mark, M.M. (Eds.), Advancing Validity in Outcome Evaluation: Theory and Practice, pp. 43–53.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and Quasi-experimental Designs for Generalized Causal Inference. Houghton Mifflin, Boston, MA.

International Advertising

C Samuel Craig, Stern School of Business, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

International advertising is a communication process that takes place in multiple cultures and contexts that differ in terms of values, economic development, media infrastructure, communication styles, and consumption patterns. This context shapes consumption behavior and influences how individuals respond to advertisements. International advertising is also a business activity involving advertisers who are expanding the market for their goods and services. These activities are facilitated by the advertising agencies that conduct research, create advertisements, and select media for the ads. The sum total of these activities constitutes a worldwide industry that is growing in importance and influence. International advertising is a major force that both reflects social values and helps reinforce or shape certain values worldwide.

Definition of International Advertising

International advertising entails dissemination of commercial messages to target audiences in more than one country. Its goal is to facilitate the sale of goods and services, although it may also involve advertising designed to gain acceptance of ideas and practices. One of the main challenges international advertisers face is dealing with the complexity of the diverse environments in which they advertise as well as the heterogeneity of individuals across and within countries. Advertising typically is deeply embedded in a particular culture and shaped by the context in which it takes place. Interpretation of symbols used in the advertisement, response to humorous or emotional appeals, and levels of literacy and languages spoken all influence how effective the ads are. The intended audiences differ from country to country in terms of how they perceive and respond to the advertisements. Also, consumers differ in terms of their basic needs, product preferences, and the ability to purchase products and services. The availability of media to convey messages to target audiences differs across countries. The organization of the advertising function also varies. In some cases, multinational firms centralize advertising decisions and budgets and use the same or a limited number of agencies worldwide. In other cases, budgets are decentralized and placed in the hands of local subsidiaries, resulting in greater use of local advertising agencies. These different organizational structures can also influence the types of advertisements that are created and how closely they are linked to the local culture. In situations where advertising decisions are highly centralized, the ads are often the same with minor modifications or incorporate a similar theme. When advertising decisions are decentralized, the advertisements are more likely to be highly diverse reflecting local idiosyncrasies.

At its core, international advertising is a communication process that takes place in multiple cultures and contexts that differ in terms of values, economic development, media infrastructure, communication styles, and consumption patterns. These contexts shape consumption behavior and influence how individuals respond to advertisements. International advertising is also a business activity involving advertisers who are interested in expanding the market for their goods and services. These activities are facilitated by the advertising agencies that conduct research, create advertisements, and select media to place the ads. The sum total of these activities constitutes a worldwide industry that is growing in importance and influence. International advertising is also a major force that both reflects social values and helps reinforce or shape certain values worldwide. The remainder of this article examines international advertising from these different perspectives to provide a better understanding of its role and impact on society.

The Context of International Advertising

Part of the complexity of international advertising derives from the diverse contexts in which it occurs. Advertising has long been practiced in highly industrialized countries such as the United States, countries in Western Europe, and parts of Asia. It is also a major factor influencing consumption choices being carried out in large countries whose economies are not as highly developed, but growing rapidly, such as China, Russia, and India. Furthermore, it is evident in the Middle East and Africa, but with significant variation by country. Countries differ in terms of the level of economic development, level of education, cultural values, climate, and religious traditions, among other factors. Variations in human behavior and attitudes are a set of collective and individual adaptations to contextual factors and more specifically the ecological and sociopolitical systems (Mishra et al., 1996). Individuals interact with and adapt to their physical environment in order to satisfy basic needs such as food, shelter, friendship, and achievement. Six principal contextual dimensions have been identified that influence choices and psychological values across countries: ecology, economy, education, mass media, population, and religion (Georgas et al., 2004). Mass media is the main conduit for advertising and is in turn shaped by contextual factors. Similarly, these factors impact consumer values, attitudes, and behavior patterns since these are deeply embedded in and to a large extent formed by overall social values. Three distinct categories of contextual variables relevant to behavior as consumers can be identified: the ecological context, the level of social affluence, and the religious context.

The ecological context, measured in terms of factors, such as average levels of precipitation or temperature or climatic zone, has been found to influence both values and consumption

patterns. For example, ecological factors have been found to impact value orientations and psychological variables such as involvement, power distance, and individualism (Georgas et al., 2004). Climate has been found to have an important influence on the consumption of pharmacological products such as alcohol, cocoa, coffee, tea, and tobacco (Parker, 1997).

In addition to the role of ecological factors in shaping culture, sociopolitical and economic factors are formative elements. Wealth, education, employment, governmental, and social institutions both foster and filter adaptation to the physical environment. The affluence of society is a powerful factor impacting culture and mediating its role on consumer behavior. Levels of GDP per capita have, for example, been found to be correlated with national value orientations and in particular, with Hofstede's measures of individualism and collectivism (Hofstede, 2001). Highly developed countries are more likely to have high levels of individualism and to value individualistic goals such as personal achievement and ambition. These are often important motivational factors driving the engine of economic growth and entrepreneurship. Furthermore, in poor countries, people depend on support from in-groups such as families or local communities, but as wealth increases, individuals gain access to resources which enable them to make their own personal choices and spend according to their own individual interests. Advertisers need to account for these differences in underlying value orientations in creating advertisements.

The level of education in a country measured, for example, in terms of factors such as levels of illiteracy, educational attainment, expenditure on R&D, also impacts culturally embedded consumption behavior. For example, Katz-Gerro (1999) found that individuals with high levels of education are more likely to engage in certain leisure activities. They are more likely to be ecologically concerned and to purchase in environmentally friendly stores such as The Body Shop or Whole Foods Markets and to be more concerned with fitness and health, to exercise regularly, and belong to gyms and health spas (Lee and Holden, 1999). Also they are more likely to be aware and open to ideas and information from other cultures and be willing to try unfamiliar products from other countries, even if these are relatively expensive exotic foods, such as frog's legs or truffles (Nijssen and Douglas, 2008). These tendencies and beliefs influence whether advertising campaigns for environmentally friendly products or novel products are likely to be successful in a particular locale.

Social, economic, and political institutions also play an important role in forming and perpetuating cultural patterns and behavior. In some societies, there are social hierarchies such as the social class structure typical of certain Western European countries or the caste system in India. Membership in a given social class, often defined by a combination of occupation, wealth, and birth, implies a certain social status. Each social class is bound by a system of social norms and obligations, and social interaction occurs predominantly within members of the same group, resulting in pressures for social conformity and similar consumption patterns. When creating advertising appeals, these differences need to be taken into account to maximize the effectiveness of the message.

Since wealth and possessions are a key factor defining membership in a given social class, the distribution of wealth is often closely linked to the degree of social stratification and hence variation in consumption patterns within society. Many Latin American countries, with highly skewed income distribution, are also marked by high variation in consumption patterns and price sensitivity. Stratified societies also vary in terms of the extent to which individuals are tightly enmeshed in the social structure or rather are free to develop their own lifestyles and patterns of social interaction. Conversely, in egalitarian societies, such as the Scandinavian countries, Japan, or Iceland, there is substantial social homogeneity and minimal variation in consumption patterns and behavior. In the latter case, one set of advertising appeals is likely to be sufficient, whereas in a very hierarchical society, a more segmented approach would need to be taken.

Religion is another factor with a strong influence on cultural values and consumption patterns. Some religions, e.g., Judaism, Islam, Hinduism, have very clearly prescribed rules for consumption behavior, particularly with regard to food and alcohol, which in some cases also extends to dress, particularly for women. Strong religious beliefs also are typically associated with conservative traditional values (Spika et al., 1985) with regard to social conformity and social issues such as feminism, divorce, and sexuality. These, in turn, may have either a direct or an indirect effect on the behavior of consumers and, in particular, responses to certain types of advertising appeals or product-positioning strategies.

Contextual factors play a key role in the formation and perpetuation of cultural values and behavior, as well as directly impacting attitudes and behavior as consumers. Interactions within the family unit and with friends, teachers, or colleagues at work, or members of other social and religious organizations to which an individual belongs mediate the impact of macro-environmental forces. In some cases, these interactions may reinforce the impact of contextual variables on individuals' behavior. In other cases, they may mediate their influence on values, attitudes, and behavior and act as barriers to change at a societal or an individual level. Contextual factors are thus important elements influencing consumer attitudes and behavior. Even where their influence is subtle and indirect, they need to be taken into account when developing international advertising strategy.

Research for International Advertising

Given the complexity of operating in multiple countries and contexts, a critical component of successful international advertising is research designed to develop a clearer understanding of consumer preferences, values, attitudes, and current behaviors. International advertising research is designed to facilitate the creation, refinement, and evaluation of advertising. It involves respondents and researchers from different countries and cultures and is more complex to undertake and more difficult to interpret than research carried out in a single country (see Craig and Douglas (2005) for a comprehensive treatment of international marketing research). Based on the understanding gained in the research, the advertiser is able to formulate strategies that can be turned into concepts and rough advertisements which can be tested and refined. The research also is able to help the advertiser determine whether a common

advertising campaign can be run in multiple countries or whether each country requires a separate strategy. Once the ads are run, the emphasis shifts to evaluating whether the campaign achieved the desired results.

The most critical factor in making meaningful inferences about advertisements in different countries is to be certain that the ads, themes, or potential strategy that consumers are reacting to are equivalent. To do this, the data must be comparable and equivalent from one country to another. The three main types of equivalence are *construct* equivalence, ensuring that the constructs are equivalent in all contexts and cultural settings; *measure* equivalence, ensuring that operational measures are equivalent in all research settings; and *method* equivalence, ensuring that data collection procedures are comparable and equally reliable in all contexts studied.

Construct equivalence has three distinct aspects. First, the advertiser must assess *functional* equivalence, whether a given concept or behavior serves the same function from country to country. Concepts, objects, or behaviors studied may not necessarily be functionally equivalent, that is, they may not have the same role or function in all countries studied. The second type of equivalence is *conceptual* equivalence and requires the advertiser to determine whether the same concepts or behaviors occur in different countries and whether the way in which they are expressed is similar. Functional equivalence is concerned with the role of objects and behavior in society at a macrocultural level, while conceptual equivalence is concerned with the interpretation that individuals place on objects, stimuli, or behavior, and whether these exist or are expressed in similar ways in different countries and cultures. Finally, the advertiser must assess *category* equivalence to determine whether the concept belongs to a specific class of objects or activities. The category in which objects or other stimuli are placed can vary from country to country. Variation in categories between countries creates different competitive sets and can influence advertising strategy.

Once the advertiser is confident that the constructs are equivalent, or has determined where there is divergence, the next step is to examine the equivalence of the measures used. Construct and measure equivalence are interrelated as the measure is an operational definition of the construct. Measure equivalence has three main dimensions: the calibration system used in measurement; the translation of the research instrument; and the metric equivalence of the instrument (Mullen, 1995). In developing a research instrument, equivalence has to be established with regard to the calibration system used in measurement. This includes not only equivalence with regard to monetary units and measures of weight, distance, and volume but also other perceptual cues, such as color, shape, or form, which are used to interpret visual stimuli. Advertising stimuli are complex and contain a large amount of information. They contain subtle perceptual cues such as color, pictorial form, or shape. Studies in cognitive and cross-cultural psychology suggest that a substantial degree of commonality exists with regard to the manifestations of these in different countries and cultures (Russell et al., 1997). However, ability to interpret, differentiate among, and develop gradations in these appears to differ. Often this is linked to schooling, language, and the physical environment. Translation equivalence is a central issue in the establishment of construct equivalence, since this is the stage in the research design at which the construct is defined in operational terms. The translation procedure may help pinpoint problems with regard to whether a concept can be measured by using the same or similar questions in each cultural context, and whether a question has the same meaning in different research contexts (Douglas and Craig, 2007). If different questions are used, then issues arise with regard to the minimal level of equivalence necessary for two questions to be considered the same, and what criteria for equivalence can be established.

In establishing metric equivalence, there are two main issues related to the equivalence of the scale or scoring procedure used to establish the measure, and the equivalence of response to a given measure in different countries (Mullen, 1995). The greater the emphasis placed on quantitative measurement in data interpretation, the more important the establishment of metric equivalence becomes. A second aspect of metric equivalence concerns the interpretation of a score obtained on a measure, or scalar equivalence. The question arises as to whether a score obtained in one research context has the same meaning as in another context.

Development of international advertising strategy must be guided by research that is valid and reliable and that provides adequate direction for the formulation of advertising strategy. While the issues of comparability and equivalence are the most critical, there are also other key issues. To begin with, a representative sample must be drawn of the target market or from a broader group of respondents to help define the target market. Research instruments suitable for use in all the countries or cultures have to be developed, a sampling plan developed, and appropriate administration procedures selected. Once the data have been collected, they have to be analyzed and the findings need to be presented in a form that can help facilitate development and evaluation of international advertising campaigns.

International Advertising as a Communication Process

In international markets, the process of communicating to a target audience is more complex because communication takes place across multiple cultures and contexts, which differ in terms of language, literacy, religion, and other cultural factors. Creating added complexity is the fact that media differ in their availability and in their effectiveness in conveying different appeals. A message may, therefore, not be effectively communicated to the intended audience for a variety of reasons. Structural reasons relate to the basic media infrastructure including the availability of certain media and the number of households that can be reached by a particular medium. Any advertising campaign that relies on the Internet will be severely limited in countries with low rates of Internet penetration. For example, Internet penetration rates in North America are almost 85%, making it an effective medium to deliver advertising messages to prospective customers. Furthermore, the 15% of households without an Internet connection may be of less interest to advertisers, as they may be older or have lower incomes. A very different picture is evident in Asia, where less than 32% of the households have

Internet access and in Africa less than 21% of the households have Internet access. The availability and penetration of more traditional media also varies. According to statistics compiled by Nationmaster.com, there were 803 television sets per 1000 inhabitants in the United States, 521 per 1000 in the United Kingdom, 65 per 1000 in India, and only 3.3 per 1000 inhabitants in Tanzania. Advertising campaigns relying on television as the main medium will be unable to effectively reach target audiences in countries with low household penetration. Alternative media may be found but if the advertisement requires sight and sound to communicate the message, its effectiveness will be diminished.

A particular medium may be available in a country and has sufficient household penetration to warrant its use, however, specific media vehicles to reach a target audience may be lacking. For example, *Cosmopolitan* magazine may be viewed as an appropriate media vehicle to reach women aged 18–34 and is also desirable due to editorial content that is compatible with the advertised product. However, if it is not available in that country, a substitute will have to be found. The substitute magazine may not reach the target as effectively, or it may lack the compatible editorial content. In addition to structural problems, there are more subtle issues related to people's inability to understand the advertisement, due to low levels of literacy. Advertising messages may be misinterpreted if different meanings are assigned to the words or symbols used. In instances where the message is adequately communicated, consumers may not respond to the message due to a lack of income to purchase the advertised product.

The communication process in international markets involves a number of steps. First, the advertiser in conjunction with the advertising agency determines the appropriate message for the target audience. Next, the message is encoded so that it will be clearly understood in different cultural contexts. There may be one central message that is unchanged, except for translation into multiple languages. Alternatively, the cultural contexts may be sufficiently varied that the message will have to be altered for each context. The message is then sent through media channels to the audience who decodes and reacts to the message. At each stage in the process, cultural barriers may hamper effective transmission of the message and result in miscommunication.

In encoding a verbal message, care needs to be taken in translation. Numerous examples exist of translation problems with colloquial phrases. For example, when the American Dairy Association entered Mexico with its 'Got Milk?' campaign, the Spanish translation read 'Are You Lactating?' conveying an entirely different meaning. Low levels of literacy may result in the need to incorporate visual symbols as the primary means of communicating. Here again, pitfalls can arise due to differences in color association or perception. In many tropical countries, green is associated with danger and has negative connotations. Red, on the other hand, is associated with weddings and happiness in China. Appeals to humor or sex need to be treated with considerable care as their expression and effectiveness varies from one culture to another. The dry British sense of humor does not always translate effectively even to other English-speaking countries. In many Muslim countries, extreme care must be taken on how females are depicted in advertisements as well as how male/female relations are portrayed.

In addition to encoding the message so that it attracts the attention of the target audience and is interpreted correctly, advertisers need to select media channels that reach the intended target audience. As indicated earlier, use of TV advertising may only reach a relatively select audience in certain countries. Equally, print media will not be effective where there are low levels of literacy. Certain media may also be more effective in certain cultures. For example, radio advertising has substantial appeal in South America where popular music is a key aspect of the local culture.

In addition to the influence of contextual factors covered in the 'The Context of International Advertising' section, the cultural context also impacts the effectiveness of communication. In what has been termed 'high-context' cultures, such as Japan and China, the context in which information is embedded is as important as what is said (Hall, 1976). In 'low-context' cultures, which include most Western societies, the information is contained in the verbal messages. In these cultures, audiences tend to demand explicit informational content (De Mooij, 2009). Conversely, people in high-context cultures are often more responsive to image or mood appeals, and rely on personal networks for informational content.

International Advertising as a Business Practice

When a firm develops an international advertising strategy, one important issue is whether to develop a global advertising campaign suitable for all countries, a regional advertising campaign for a specific part of the world, or to tailor communications to differences in local markets (Mueller, 2011). If the purpose of advertising is to develop a strong corporate or global image for the firm, a uniform global campaign is more likely to be used. When, on the other hand, the objective is to launch a new product or brand, or to more clearly differentiate the product or brand from other competing brands or products, local campaigns tailored to local markets are more typical.

A global campaign offers a number of advantages. In the first place, it can be an important means of building a strong and coherent global image for the firm and its products worldwide. Use of the same image in different countries builds familiarity and generates synergies across world markets. It allows utilization of good ideas and creative talent (both of which are scarce commodities) on a worldwide basis. In addition, use of a single campaign can result in substantial cost savings in copy development and production costs. Conversely, development of multiple local campaigns can lead not only to duplication of effort but also different messages in different countries. Given the increased reach of media across borders, ads run in one country may spill over into other markets. This can result in inconsistent brand images across countries and confusion in consumers' minds with regard to the benefits offered by the brand and the corporate image.

While use of uniform advertising appeals offers a number of advantages, differences in customer perceptions and response patterns across countries and cultures as well as media availability and government regulation are major barriers to use of a standardized campaign. Even though technological developments allow adaptation of advertising appeals to different languages (for example, TV can have audio channels in two

languages, Internet messages can be automatically translated), development of visual and verbal copy that works effectively in multiple countries poses major creative challenges.

Faced with this dilemma, firms may use a global umbrella campaign combined with local country- or product-specific advertising. The global umbrella campaign develops a uniform image for the company or brand worldwide, often relying on consistent visual images and the corporate logo. Product-specific or country-specific advertising builds on this image, modifying the appeal and providing information tailored to the local market. The objective of the umbrella campaign is to provide an integrating force, while local campaigns provide greater relevance to specific needs of local customers and markets.

The organizational structure of the firm often plays a key role in the choice of global versus locally adapted campaigns. If international operations are organized on a country-by-country or geographic basis and operate as local profit centers with local advertising budgets, pressures exist for use of local advertising campaigns. If, on the other hand, the company is organized by product divisions, with centralized advertising budgets at corporate or regional headquarters, use of regional or global advertising campaigns is more likely (see Douglas and Craig (2011) for more information on global strategy).

International Advertising as an Industry

The global advertising industry is characterized by a large number of small- and medium-sized advertising agencies that operate primarily in one country and by a small number of very large advertising agencies with operations in many countries. These agencies have developed extensive networks of offices throughout the world in order to coordinate the advertising process in all the countries where their clients do business. These networks often include both wholly owned subsidiaries and formal relationships with local advertising agencies to establish a presence in new markets, particularly in emerging markets.

In an effort to establish greater control over their advertising, many major advertisers are consolidating all their advertising with one agency. For some major advertisers such as IBM and Citibank, this represents annual advertising expenditures in excess of $500 million worldwide (Grein and Ducoffe, 1998). As a consequence, advertising agencies that do not have a global network are at a serious disadvantage when competing for new advertising accounts or attempting to retain existing ones that are expanding globally.

The majority of these large advertising agencies are headquartered in the United States although the ownership structure of the agency holding companies is more diverse. Of the seven largest advertising agency groups, two are headquartered in the United States, two in France, one in the United Kingdom, and two in Japan. The largest agency group, WPP, is a British agency holding company made up of four large US-based agencies as well as other firms that provide marketing communication services. WPP's revenues in 2013 were $18 billion (US dollars) with two-thirds of the revenue coming from outside the United States. WPP has 3000 offices in 110 countries and employs 179 000 people worldwide. The companies in the WPP group provide a wide range of services beyond advertising including marketing research, public relations, branding consulting, direct marketing, and digital marketing. With the exception of Dentsu and Hakuhodo, in Japan, most other agency networks generate the majority of their revenues outside their home country. Collectively, US-based advertising agencies and their subsidiaries are responsible for most of the advertising throughout the world. For example, of the approximately $60 billion in advertising placed by the top 25 agency networks in Europe during 1995, 89% of the total was placed by subsidiaries of US-based agencies. This general pattern holds in most parts of the world that do not have restrictions on foreign ownership. The major exception is Asia where the three major Japanese agencies account for 62% of the advertising placed by the top 25 agency networks. Current and comprehensive information on advertising can be obtained from *Advertising Age's* Web site, although there is a fee (www.adage.com).

Worldwide over $500 billion is spent on advertising annually. Approximately one-third of that amount is spent in the United States and the other two-thirds outside the United States. Information on advertising spending can be obtained from *Advertising Age's* Web site and from ZenithOptimedia's Web site (www.zenithoptimedia.com). The bulk of expenditures outside the United States takes place in Europe and Japan, although China, Brazil, Canada, and Australia are also important advertising markets.

The table below shows projected advertising spending in the top 10 global ad markets for 2016. The United States and Japan account for 59% of the total advertising spending in these markets and the four European markets account for an additional 20%. The concentration of spending in the United States in part explains the dominance of US-based advertising agencies. Not only do they work for US-based clients that continue to expand outside the United States, but they also accumulate knowledge, expertise, and experience in the practice of advertising that can be applied elsewhere.

Country	*Projected 2016 advertising expenditures (billions, US dollars)*	*Percent*
The United States	191.4	45
China	57.3	13
Japan	56.3	13
Germany	24.1	6
The United Kingdom	23.9	6
Brazil	20.2	5
Australia	14.3	3
Indonesia	14.2	3
South Korea	13.3	3
France	13.1	3
Total	421.3	100

Note: The expenditures reflect major media such as TV, the Internet, mobile, magazines, newspapers, radio, outdoor, and cinema. Total advertising spending for the entire world is projected to be $603 billion in 2016 and the top 10 countries account for 71% of the total.

Once the advertising message has been created, a media plan must be developed and specific media vehicles purchased to deliver the message to the target audience. This is either handled by the advertising agency or increasingly being done by specialized media buying organizations. Media differ from country to country in terms of their availability, effectiveness, and efficiency in delivering a message, and with relatively few exceptions, tend to be organized on a country-by-country basis. Notable exceptions include many Internet sites, Star TV, MTV, CNN in television, *Business Week International*, the *Asia Wall Street Journal*, the *International New York Times* in print, and selected industry and medical publication that are read worldwide. There is also a trend toward consolidation of media in order to achieve greater economies of scale and leverage content developed in one market to others. This consolidation facilitates purchase of media on a regional and global basis. In addition, the Internet is emerging as a truly global medium that does not conform to country boundaries. Web sites like Google and eBay have a global presence, although they do have local competitors. In 2014, Facebook had over 1.35 billion users and 82% were outside of the United States. Consistent with this, Facebook was available in over 70 languages (www.facebook.com). Most of the growth in advertising spending has been related to the Internet. Increased Internet penetration throughout the world is causing firms to either spend more on advertising or simply to shift expenditures from traditional media to the Internet.

International Advertising as a Social Force

From the advertiser's perspective, the primary objective of advertising is to sell products or services. In achieving this primary goal, there are often profound secondary consequences. Advertising exerts a formative influence throughout the world whose character is both pervasive and persuasive. Through the selective reinforcement of certain social roles, language, and values, it acts as an important force fashioning the cognitions and attitudes that underlie behavior not only in the market place but in all aspects of life. In an international setting, advertising has an important social influence in a number of ways. First, much international advertising is designed to promote and introduce new products that have been successful in one society into another. Often this results in radical changes in lifestyles and behavior patterns of a society, stimulating, for example, the adoption of fast food, casual attire, or specific beauty products. International advertising also encourages desire for products from other countries, as it creates expectations about 'the good life,' and establishes new models of consumption. Advertising is thus a potent force for change, while selectively reinforcing certain values, lifestyles, and role models.

Often the symbols, ideals, and mores that international advertising portray and promote are those of Western society and culture. Through global advertising, brands such as Levi's, Nike, Adidas, and McDonalds are known by and have become objects of desire for teens and young adults throughout the world. Consumers are exposed to advertisements for luxury goods such as Hermes, Rolex, Louis Vitton, and Gucci, whether or not they are available for sale in their country. Similarly, images and scenes depicted in much international advertising are either Western in origin or reflect Western consumption behavior and values. Even where adapted to local scenarios and role models, those shown often come from sectors of society, such as the upwardly mobile urban middle class, which embrace or are receptive to Western values and mores.

A criticism frequently leveled at international advertising is that it promulgates Western values and mores, notably from the United States, in other countries. This is viewed particularly negatively in societies with strong religious or moral values, which run counter to those of the West as, for example, Islamic societies in the Middle East. When Western advertising relies on sex to sell products or shows women in situations considered inappropriate or immoral, it is likely to be considered a subversive force undermining established cultural values. Equally, in some countries such as France, there is a strong negative reaction to the encroachment of US culture, values, and use of English in advertising. Promotion of tobacco products by US and UK companies in countries where there is no legislation regulating or banning cigarette advertising has also been criticized.

At the same time, international advertising acts as an integrating force across national boundaries. It disseminates messages using universal symbols and slogans, and establishes a common mode of communication among target audiences in different parts of the world. At the same time, multicultural values are reinforced by advertisers, who adopt images incorporating peoples of different nations and diverse cultural backgrounds. Coke Zero created a campaign that involved a search for dancers worldwide to star in their ads and Microsoft's television campaign for Windows-based phones that ran in 35 countries had a common theme, contained no dialog, and was tailored to various regions by using actors of difference ethnicities and local music. The impact of such campaigns is further reinforced by the growth of global media such as Star TV, CNN, MTV, or print media that target global audiences worldwide. Consequently, while international advertising can be viewed by its critics as a colonizing force propagating Western values and mores throughout the world, it is also an important force integrating societies and establishing common bonds, universal symbols, and models of communication among peoples in different parts of the globe.

See also: Advertising Agencies; Advertising: General; International Communication: History; Mass Media and Cultural Identity.

Bibliography

Craig, C.S., Douglas, S.P., 2005. International Marketing Research, third ed. John Wiley & Sons, Chichester, UK.

De Mooij, M., 2009. Global Marketing and Advertising: Understanding Cultural Paradoxes, third ed. Sage Publications, Thousand Oaks, CA.

Douglas, S.P., Craig, C.S., 2011. Convergence and divergence: developing a semi-global marketing strategy. Journal of International Marketing 19 (1), 82–101.

Douglas, S.P., Craig, C.S., 2007. Collaborative and iterative translation: an alternative approach to back translation. Journal of International Marketing 15 (1), 30–43.

Georgas, J., van de Vijver, F.J.R., Berry, J.W., 2004. The ecological framework, ecosocial indices, and psychological variables in cross-cultural research. Journal of Cross-Cultural Psychology 35, 74–96.

Grein, A., Ducoffe, R., 1998. Strategic response to market globalization among advertising agencies. International Journal of Advertising 17, 301–319.
Hall, E.T., 1976. Beyond Culture. Anchor Press, Garden City, NY.
Hofstede, G., 2001. Culture's consequences: comparing values, behaviors, institutions and organizations across cultures. Sage Publications, Thousand Oaks, CA.
Katz-Gerro, T., 1999. Cultural consumption and social stratification: leisure activities, musical tastes, and social location. Sociological Perspectives 42 (4), 627–646.
Lee, J.A., Holden, S.J.S., 1999. Understanding the determinants of environmentally conscious behavior. Psychology & Marketing 16 (5), 373–392.
Mishra, R.C., Sinha, D., Berry, J.W., 1996. Ecology, Acculturation, and Psychological Adaptation: A Study of Adivasis in Bihar. Sage Publications, Thousand Oaks, CA.
Mueller, B., 2011. Dynamics of International Advertising: Theoretical and Practical Perspectives. Peter Lang Publishing, New York.
Mullen, M.R., 1995. Diagnosing measurement equivalence in cross-national research. Journal of International Business Studies 26 (3), 573–596.
Nijssen, E.J., Douglas, S.P., 2008. Consumer worldmindeness, socialmindedness and store image. Journal of International Marketing 16 (3), 84–107.
Parker, P.M., 1997. National Cultures of the World: A Statistical Reference. Greenwood Press, Westport, CT.
Russell, P., Deregowski, J., Kinner, P., 1997. Perception and aesthetics. In: Berry, J.W., Poortinga, Y.H., Pandey, J. (Eds.), Handbook of Cross-Cultural Psychology, vol. 2. Allyn & Bacon, Boston, MA, pp. 109–143.
Spika, B., Hood, R., Gorsuch, R., 1985. The Psychology of Religion: An Empirical Approach. Prentice Hall, Englewood Cliffs, NJ.

International and Comparative Criminology

Amy Nivette, Nuffield College, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The terms 'international' and 'comparative' criminology are often used interchangeably to refer to any research concerning crime or criminal justice that either takes place outside the researcher's country of origin or involves more than one country. Each term refers to a wide body of rich comparative research that frequently intersects both theoretically and methodologically. This article gives a broad overview of comparative research, with a focus on (1) the methodological strengths and weaknesses, (2) the state of knowledge in macro- and microlevel comparative research, and (3) regional characteristics of crime.

Introduction

Crime is a multifaceted problem that faces societies of every country and every historical period (LaFree, 2006). As such, criminologists are faced with a growing body of potential explanations and solutions to the widespread, but varying, problem of crime. It is the task of comparative criminologists to explore which explanations are universal, which operate according to situational context, and which are culturally specific (Karstedt, 2001; Stamatel, 2009). Comparative researchers have already made many important contributions to the understanding of crime, violence, and criminal justice, in some cases confirming and in others rejecting common explanations.

The terms 'international' and 'comparative' criminology are often used interchangeably to refer to any research concerning crime or criminal justice that either takes place outside the researcher's country of origin or involves more than one country. Stamatel (2009) offers a more nuanced distinction between international and comparative criminology. International criminology is a general term that refers to crime or criminal justice trends in one or more countries. International criminology is not necessarily comparative since the focus of research may be on a single country; however, international comparisons are often implicit in single-country studies. Comparative criminological research is designed to investigate differences in crime and criminal justice organizations across borders. Often comparisons are made across international borders but may involve other units of analysis, such as cities, organizations, provinces, or regions. Falling under the banner of both international and comparative criminology is the more specific disciplinary term 'cross-national' criminology (Stamatel, 2009). Cross-national criminology refers to the study of crime across countries, typically with nations as the unit of analysis.

Overall, each term refers to a wide body of rich comparative research that frequently intersects both theoretically and methodologically. International comparative research covers everything from descriptive studies on the prevalence of theft in China, to multicountry school-based studies testing the universal relevance of self-control, to large-scale country-level studies testing theories of modernization. This article gives a broad overview of comparative research, with a focus on (1) the methodological strengths and weaknesses, (2) the state of knowledge in macro- and microlevel comparative research, and (3) regional characteristics of crime. It would be impossible to cover all of the many topics that fall under the label of 'international' and 'comparative' criminology, so for further reference please refer Section See also at the end of this article.

Comparing Crime across Countries

When comparing crime across countries, criminologists face two major roadblocks: the variation in legal definitions and recording, and the potential inaccuracy of reported rates. In general, it is well understood that most crimes are not comparable across borders due to fluctuations in official reporting practices and legal definitions (Neapolitan, 1997). Only homicide is comparable across countries, and this is because definitions of murder are generally historically and internationally consistent (Marshall and Block, 2004). Even so, international organizations such as the United Nations Office on Drugs and Crime (UNODC) must take special care to ensure homicide data collected from countries are comparable.

The European Sourcebook of Crime and Criminal Justice Statistics 3rd Edition (Aebi, 2006) provides detailed information on the contents of legal definitions of common crimes across 35 countries. The *Sourcebook* shows how hazardous cross-country crime comparisons are, even among countries with shared sociocultural histories. **Table 1** illustrates the differences in legal definitions of robbery and intentional homicide in seven former Soviet Republics. In order to compare robbery rates in Ukraine to those in neighboring Moldova, one must take into account inflation in Moldova due to the inclusion of muggings and pick-pocketing in the reported rates. From **Table 1** it is clear that comparing thefts across countries can be very difficult. In Armenia, burglaries are not explicitly defined in the criminal code, and so are classed as either thefts or robberies (Aebi, 2006). Moldova is the only country that includes embezzlement in the overall amount of thefts. Ukraine has the most restrictive definition of theft, including only burglaries and thefts of other items. Even considering that Lithuania, Estonia, and Russia have similar definitions, both Estonia and Lithuania have slight differences in what kinds of theft are included or excluded. Lithuania, for example, separates general theft apart

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45063-4

Table 1 Differences in legal definitions of robbery and theft in post-soviet countries

Country	Included in robbery definition
Armenia, Estonia, Georgia, Lithuania, Russia	Muggings, theft involving violence
Moldova	Muggings, theft involving violence, pick-pocketing
Ukraine	Theft involving violence
Country	**Included in theft definition**
Armenia, Georgia	Burglary, theft of motor vehicle, theft of other items, theft of small value
Lithuania, Estonia, Russia	Burglary, theft of motor vehicle, theft of other items
Moldova	Burglary, theft of other items, theft of small value, embezzlement
Ukraine	Burglary, theft of other items

Note: In Armenia, burglary is not defined in the criminal code, and so what are classed as burglaries (i.e., "gaining access to a closed part of a building or other premises by use of force with the objective to steal") are counted as robberies or thefts. (Aebi, M.F., 2006. Raw Data for the 3rd Edition (2006) of the European Sourcebook of Crime and Criminal Justice Statistics 2000–2003, p. 25. http://www.europeansourcebook.org/ (accessed 16.03.13.).). In Estonia, theft of small value is included in total thefts only when it occurs during a burglary (p. 27). In Lithuania, certain types of theft are considered separate crimes, e.g., thefts of drugs, guns, or explosives (p. 31).
Source: European Sourcebook for Crime and Criminal Justice Statistics (Aebi, M.F., 2006. Raw Data for the 3rd Edition (2006) of the European Sourcebook of Crime and Criminal Justice Statistics 2000–2003. http://www.europeansourcebook.org/ (accessed 16.03.13.).)

from thefts of guns, drugs, or explosives, and has decriminalized simple thefts up to 125 Lithuanian litas (US $47) (Aebi, 2006: p. 31).

In addition to varying legal definitions, international crime statistics may be inaccurate due to a country's weak institutional capacity to detect and record crimes, or a political interest to make the country appear safer and the state more effective. Before 1989, for example, many researchers believed crime rates in Soviet states were significantly lower than in Western Europe (see Stamatel, 2008). This was in part due to ideological propaganda by Communist political leaders, and partly due to the idea that crime was strictly and forcefully suppressed (Shelley, 1981). On the contrary, Stamatel (2008) found that homicide rates were at least as high as in Western European countries and the United States prior to 1989. The increases in crime after this period that were commonly attributed to transition and democratization were already on the rise before the fall of communism.

In order to avoid these roadblocks, researchers can draw from a range of alternative statistical sources, including victimization surveys, self-reported delinquency studies, and public health data. Victimization surveys, such as the International Crime Victims Survey, aim to capture the number of crimes committed in each country by interviewing a sample of citizens and asking if they have been the victim of a crime in the past 6–12 months (Van Dijk, 2012/2013). Self-reported delinquency studies are more commonly carried out in a single international setting, and while these studies are informative for examining universal predictors of crime, they are not always comparable across countries. However, the second International Self-Reported Delinquency study (ISRD-2) (Junger-Tas et al., 2012b) was carried out in 31 countries to create internationally comparable statistics on crime prevalence and victimization among youth. Finally, public health data are available for lethal violence across a large sample of countries from the World Health Organization (2012). The benefits of public health data are that reporting does not depend on police practices and political motives, and that the definition of lethal violence is based on the standardized International Classification of Disease Codes (ICD). The ICD-10 (10th Revision) defines homicide as "injuries inflicted by another person with intent to injure or kill, by any means" and excludes injuries due to legal intervention and war (World Health Organization, 2010).

Determinants of Crime across Countries

International and comparative investigations of criminological risk factors tend to fall into two broad categories: large-scale cross-national studies and individual-level single- or multisite studies. Both aim to test whether theories and causal risk factors of crime are universal or culturally specific.

Macrolevel Correlates of Crime

Cross-national criminological research has shown a diversity of interests including economic development, modernization, culture, political organization, human rights, punishment, and population structure. Researchers have been gradually expanding the temporal and geographical domain of comparative criminology with the emergence of more and improved international data. Recent research syntheses show that the strongest correlates of national-level homicide rates are income inequality, poverty, human development, and population heterogeneity (Nivette, 2011; Pridemore and Trent, 2010). **Table 2** shows the top 10 strongest indicators of homicide that have been tested in more than 10 studies between the years 1960 and 2010 (see Nivette, 2011). The strongest average effects, as measured by the standardized correlation coefficient (β), come from measures of income inequality ($\beta = .22$–$.48$), as well as measures of family and social structure such as the divorce rate ($\beta = .28$), female labor force participation ($\beta = .22$), and ethnic heterogeneity ($\beta = .22$).

Table 2 Top ten cross-national indicators of homicide with more than 10 contributing studies

Indicator	*No. of studies*	*Average effect size (β)*
Income inequality (ratio top/bottom wealth percentiles)	13	.42
Divorce rate	10	.28
Income inequality (Gini index)	31	.22
Female labor force participation	13	.22
Ethnic heterogeneity	13	.22
Population aged >15 years	17	.12
Development index	14	−.16
Urbanism	27	−.10
Sex ratio (males to females)	15	−.10
Social support	15	−.07

Source: Nivette, A., 2011. Cross-national predictors of crime: a meta-analysis. Homicide Studies 15, 116.

However, Nivette and Pridemore, and Trent have reservations about the conclusiveness and reliability of cross-national results. National-level risk factors are often given little theoretical and methodological attention, and countries are frequently included on the basis of available data rather than theoretical purpose.

One example is the positive link between income inequality (most often measured by the Gini index) and homicide rates. It is clear from a multitude of studies that measures of income inequality are associated with homicide, but one has yet to firmly grasp specifically why, how, and where inequality is detrimental. Indeed, a number of scholars have questioned the inequality–homicide link on both theoretical (Chamlin and Cochran, 2005) and methodological (Pridemore, 2008) grounds. Chamlin and Cochran argued that income stratification is not inherently damaging to social institutions and social relations unless it is founded on ascribed characteristics. When inequalities are ascribed, individuals perceive economic institutions as illegitimate and subsequently withdraw participation and commitment from institutions that provide social control. As such, common explanations for the effects of inequality that are based on psychological strain may be misguided, and any measures that do not distinguish between meritocratic and ascribed inequalities may have low concept validity. Pridemore (2008) argued that the inequality–homicide link is only prevalent due to model misspecification. Specifically, he questioned the absence of indicators of poverty in cross-national research, and attempted to replicate prior studies after including a proxy for poverty (the infant mortality rate). He found that in two of the three studies he reexamined, the inequality–homicide link disappeared when poverty was added to the model.

In order to rigorously test macrolevel theories and risk factors, such as the inequality–crime link, cross-national researchers require larger and more diverse samples along with clear and testable mechanisms connecting the indicator to the outcome. Despite these limitations, however, macrolevel comparative criminology has advanced our understanding of the causes and consequences of criminal behavior within and across a variety of cultural and national contexts. In Stamatel's (2009: p. 18) words,

> Because cross-national research asks fundamentally different research questions, emphasizes macro-level conditions and processes, and makes time and place central to its inquiry, it is well positioned to make unique contributions to both the breadth and depth of knowledge in criminology. Not only are contributions valuable in their own right, but they can also be applied to intra-national research questions, like crime changes in the United States, and they can inform pressing contemporary crime problems, such as terrorism and other transnational crimes.

Microlevel Correlates of Crime

Comparative microlevel studies of crime typically examine the effects of individual, family, school, or neighborhood causal risk factors on criminal behaviors. According to Murray et al. (2009: p. 3), "[r]isk factors are correlates that are shown to predict delinquency. To demonstrate something is a risk factor, a study needs to demonstrate correlations, and the variable be shown to predict the outcome." Despite a growing number of international studies searching for universal or culturally specific risk factors, there is still little conclusive evidence supporting one or the other (Eisner and Nivette, 2012).

Certain theories, such as Gottfredson and Hirschi's (1990) general theory of crime, are explicitly framed as universal theories, essentially rejecting the idea that macrolevel structural and cultural factors play a role in causing crime. A universal risk factor, like Gottfredson and Hirschi's self-control, is expected to predict crime by the same mechanism across all situations and contexts. A culturally specific risk factor, such as the concept of 'machismo' in Latin America (Neapolitan, 1994), is one that is defined by the context in which it operates, including the cultural values, situational contexts, and structural factors that determine the effect of the risk factor on crime (Karstedt, 2001).

Some risk factors have strong international support. For instance, low self-control has been shown to be a robust correlate of criminal and other problem behaviors both in the United States and cross-nationally (Rebellon et al., 2008). Rebellon et al. (2008) examined the salience of self-control on self-reported property crime and violent crime using the International Dating Violence Study (IDVS). The IDVS sampled over 20 000 university students across 32 countries, measuring a range of potential individual and social risk factors. They found that while low self-control proved to be a significant risk factor across all nations, there is evidence that individual self-control is conditioned by the macrolevel context of parental neglect.

The ISRD-2 was designed by Junger-Tas et al. (2012b) to generate comparable measures of self-reported offending and associated risk factors across 31 countries. The ISRD-2 measures a range of individual and social risk factors associated with social control, self-control, opportunity/routine activity theory, and social disorganization. So far, support for these concepts has been mixed. Generally, Junger-Tas et al. (2012a) report that certain risk factors related to the above concepts are relevant across all 31 countries, but not all. In relation to social control theory, Junger-Tas et al. (2012a) found that only measures of family bonding and parental supervision were related to delinquency in the ISRD-2 study. Gatti et al. (2011) found that associating with deviant peers is positively related to delinquency and substance use across all countries. Junger-Tas et al. (2012b) found that school-related risk factors, including school disorganization, truancy, and school bonding, were important correlates of delinquency. Additionally, Steketee et al. (2013) report that perceptions of neighborhood disorganization, individual low self-control, and positive attitudes toward violence all significantly increased the likelihood of delinquency across all countries.

Drawing from a multiethnic sample in Zurich, Switzerland (N = 1229), Ribeaud and Eisner (2010) examined a wide range of cognitive, personal, familial, social, and contextual risk factors on adolescent delinquency. The strength of the Zurich sample is in its sample composition: almost half of the children in the sample (46%) have parents born outside Switzerland. Additionally, parents immigrated from a diverse range of countries, including Germany, Croatia, Serbia, Turkey, and Sri Lanka.

Ribeaud and Eisner found that individual, parental, social, and school risk factors were significantly related to delinquency, whereas parental socioeconomic characteristics were

not. Significant cumulative risk factors include externalizing and internalizing aggressive behaviors (e.g., aggression, attention-deficit hyperactivity disorder), perinatal factors (e.g., maternal depression), internal individual factors (e.g., self-control, sensation-seeking), indoor leisure activities (e.g., extended TV exposure), family structure (e.g., broken home), parenting factors (e.g., lack of parental involvement, use of corporal punishment), school environment (e.g., school disorder), and peer status (e.g., low popularity, low trustworthiness). Sociodemographic characteristics of the parents, such as low parental education, low socioeconomic background, and whether or not the parent was a migrant, were not associated with adolescent delinquency.

Regional Characteristics of Crime

As comparative criminological research expands to every corner of the world, one is gaining a better understanding of levels and characteristics of crime across regions. Macrolevel research has identified regions that tend to share certain characteristics and trends of crime, such as high homicide rates in Latin America, low homicide rates in East Asia, or the convergence of crime rates in Western and Eastern Europe (Gruszczyńska and Heiskanen, 2011; Nivette, 2011; Ouimet, 2012). This section reviews new and unique global and regional research on trends, risk factors, and types of crime in order to demonstrate the wide range of topics concerning comparative criminologists. In addition, it is possible to highlight some potential common characteristics, risk factors, and consequences of crime across such diverse social, political, cultural, and economic contexts.

Table 3 provides a summary of total and gender-specific homicide rates (i.e., violent deaths per 100 000 persons) across eight broad geographical regions in 2008 (World Health Organization, 2012). The lowest homicide rates can be found in Europe ($M = 2.8$), Oceania ($M = 3.3$), and Central and Western Asian countries ($M = 3.8$), whereas the highest are on average found in sub-Saharan Africa ($M = 18.6$) and Central and South America ($M = 21.2$). On average, males across the globe are victimized by fatal violence more than females ($M_{\text{males}} = 15.7$; $M_{\text{females}} = 3.2$).

However, these global estimates must be considered with caution. The extent of uncertainty in certain regions is evident by the variation in the percentage of poor data quality. In many African and Asian countries, information on cause of death is unavailable, meaning the World Health Organization must estimate the number of deaths due to violence using country-specific information and death registration data from other countries in the region (WHO, 2012). For regions with widespread missing data such as Africa, this means homicide rates are based on very few available cases, and as such are likely to be unreliable. This unreliability is apparent in some countries more than others. For example, the World Health Organization estimated a homicide rate of 1.5 deaths per 100 000 persons in Somalia, a complex country that has been plagued by clan-based violence, state collapse, and invasion since 1988 (Linke and Raleigh, 2011).

In Europe, trends in homicide leading up to and following the Second World War are largely described by convergence (Eisner, 2008; Gruszczyńska and Heiskanen, 2011). Eisner (2008) reported that shortly after the Second World War, homicide levels in Western Europe were the lowest in European history (ranging from around 0.2 per 100 000 in the Netherlands in 1957, excluding infanticides, to 0.97 per 100 000 in Austria in 1966). These trends remained relatively similar as homicide increased in the late 1950s and 1960s, peaking in the 1990s and returning to hover around 1 death per 100 000 persons (Eisner, 2008; see also Marshall and Summers, 2012). The fact that the increases and subsequent declines occurred across Western European states (as well as the United States and Canada; see Zimring, 2007) suggests some similar underlying cause. Zimring (2007) argued that it was a number of factors, e.g., the decrease in number of young males, expanding welfare policy, increases in police numbers, and incarceration, that led to the 'great American crime decline' in the 1990s. However, he admitted these changes did not occur in Canada, where a similar crime trend was observed, and differences in policy and social structure are even more acute between Western Europe and the United States.

Within Europe, there are, however, several points of divergence. After 1989, crime and violence increased dramatically among postcommunist states, but more recently show decline and stability (Andresen, 2011; Marshall and Summers, 2012). Certain postcommunist states still have relatively higher levels of violence, particularly the Baltic region (i.e., Latvia, Lithuania, and Estonia), with an average of 10 deaths per 100 000 persons

Table 3 Average homicide rates (deaths by violence per 100 000 persons) in 2008 by gender across eight geographical regions

Region	*No. of countries*	*Average total rate*	*Average male rate*	*Average female rate*	*Percentage of poor data*
Europe	42	2.8	4.2	1.5	0.05%
Oceania	16	3.3	4.3	2.2	50%
Central & Western Asia	22	3.8	6.2	1.5	27%
Southern & East Asia	25	5.8	9.8	1.8	48%
Northern Africa	6	6.1	8.6	3.7	83%
North America & Caribbean	15	13.7	24.3	3.3	0%
Sub-Saharan Africa	47	18.6	30.5	7.0	94%
Central & South America	20	21.2	38.0	4.8	10%
Total	193	9.4	15.7	3.2	39%

Note: The percentage of poor data was calculated as the percentage of countries in the region where most data for cause of death were unavailable, meaning rates were estimated by the World Health Organization using country-specific information and cause of death patterns from other countries in the region.
Source: World Health Organisation, 2012. Disease and Injury Country Estimates. http://www.who.int/healthinfo/global_burden_disease/estimates_country/en/index.html (accessed 18.03.13.).

for the period 1990–2008 (Marshall and Summers, 2012: p. 59). Gruszczyńska (2004) pointed out that since independence in 1989, some postcommunist states have become 'corridors' for drug trafficking, car theft, and the expansion of organized crime rings between east and west. The problem has only intensified since Latvia, Lithuania, and Estonia gained membership to the European Union, requiring these states to more actively police their borders with Russia and Belarus (Ceccato, 2008; Gruszczyńska, 2004).

The UNODC's *Global Study on Homicide 2011* provides detailed information on the weapons used, age of the offender and the victim, and circumstances surrounding the murder where data are available. UNODC (2011: p. 40) reports that firearms are used to commit homicide more frequently in the Americas than in Europe. The percentage of homicide cases in 2008 involving a firearm ranged from about 6% in Eastern Europe to over 70% in South America. In Europe, homicides were more frequently committed using a sharp object, such as a knife.

Although data are sparse for most African countries, the *Global Burden of Armed Violence 2011* reports that in South Africa, "firearm homicide has been singled out as the leading cause of death for young men aged 15–21" (Geneva Declaration Secretariat, 2011: p. 162, citation omitted). Pierre du Toit (2001) argued that the resistance to apartheid, coupled with recent civil wars in the region, contributed to the proliferation and flow of arms in South Africa. While this may contribute to South Africa's high firearm homicide rates, the *Global Burden of Armed Violence* also implicates high levels of income inequality, unemployment, and poverty.

In sub-Saharan Africa, the rate of lethal violence and the persistence of civil conflict are often highly correlated (Geneva Declaration Secretariat, 2011). However, the context in which high rates of firearm homicides occur differs in Central and South America. UNODC (2011) reports that high homicide – in particular homicide by firearms – is not entirely attributable to conflict in Latin America and the Caribbean regions. Instead, firearm homicide is associated with organized crime and gang activity. Importantly, they are clear that organized crime/gangs are not necessarily the cause of high rates: "many countries in the region were vulnerable to crime and violence for a number of reasons, including their legacy of armed conflicts and violence, the easy availability of guns, chaotic urbanization, high income inequality, a high proportion of youth, local gang structures, as well as organized crime and drug trafficking" (UNODC, 2011: p. 49).

National-level rates can often mask gang-related intrastate variation, as in the case of Mexico's escalating drug war. Mexico's overall homicide rate has remained relatively stable from 2004 to 2009 at 11.5 deaths per 100 000 persons (Geneva Declaration Secretariat, 2011). However, in the state of Chihuahua, the homicide rate was 108 per 100 000 in 2009, which the *Global Burden of Armed Violence* attributes to an increase in drug trafficking, organized crime, and confrontations with the Mexican government. In fact, in 2010, 41% of Mexico's homicides occurred in four states: Chihuahua, Sinaloa, Guerrero, and Baja California, which together account for 11% of the total population of Mexico (UNODC, 2011). Furthermore, UNODC (2011) reports that in 2008–09, two-thirds of homicides in Chihuahua occurred in Ciudad Juárez, and three-fourths of homicides in Baja California occurred in Tijuana. This means that the nature of homicide in Mexico – particularly in areas affected by drug cartels – differs significantly from most homicides occurring in Europe, and so researchers must be careful when considering broad cross-national explanations of crime.

Comparative criminology is essential to understanding these transnational organizations and their impact on local social order and crime. International organizations and national governments are increasingly concerned about cross-border criminal activity, such as human, drug, and weapons trafficking; wildlife smuggling; the spread of organized crime syndicates; and terrorism.

For example, since the terrorist attacks in the United States on 11 September 2001, the study of terrorism has become a hot topic in comparative and international criminology. However, until recently, very little was known about the spatial patterns of terrorist activity and associated risk factors (LaFree and Ackerman, 2009; LaFree et al., 2010).

Gary LaFree and Laura Dugan (2007) have improved the comparative study of terrorism by developing a systematic database of both domestic and international terrorist events dating back to 1970. The Global Terrorism Database (GTD) uses media and other open-source databases to compile a comprehensive list of events that include information on the attacker(s), intended target(s), country of origin, country of incident, number of fatalities, number of injuries, weapons used, and intention (e.g., assassination, political protest).

LaFree et al. (2010) examined the trajectory and spatial distribution of terrorist attacks across countries using the GTD. They found that most fatal and nonfatal terrorist attacks since 1970 were clustered in four regions: Latin America (30.43% of all attacks), Western Europe (20.75%), Middle East/Persian Gulf (15.41%), and South Asia (14.45%). Terrorist attacks were least frequent in North America (1.24% of all attacks) and East/Central Asia (1.19%). LaFree et al. (2010: p. 633) point out that Latin America and Western Europe likely have more attacks because of "long-lasting and stable terrorist groups that often claim responsibility for their attacks." For example, in Colombia, the Revolutionary Armed Forces of Colombia have been operating for decades, as have the Basque Fatherland and Freedom in Spain, and the Irish Republican Army in Northern Ireland (Global Terrorism Database, 2013).

Using trajectory analyses, LaFree and his colleagues identified several countries with rapidly rising number of terrorist attacks since 1999, which they characterized as 'hot spots': Afghanistan, Bangladesh, Indonesia, Iraq, Kashmir, Nepal, Rwanda, Thailand, and West Bank/Gaza. However, LaFree et al.'s (2010) analysis of 'hot spots' only includes incidents up to 2006, meaning recent surges in attacks from Islamist group Boko Haram in Nigeria and Al-Shabaab in Somalia and Kenya are not included in LaFree et al.'s analyses (Marchal, 2009; Africa Research Bulletin, 2013). Further, the cases of Boko Haram and Al-Shabaab show the limitations of using incidents of attacks to predict future terrorist attacks. According to the GTD, since 2006, there have been four total attacks attributable to known groups in Rwanda, whereas there have been 205 in Somalia and 159 in Nigeria. (Counts only include attacks that have been attributed to a known perpetrator by the media, meet terrorism criterion 1 (political aim) and 2 (coercive intent), and exclude ambiguous and unsuccessful attacks (see Global Terrorism Database, 2013).) Predicting the success and decline of terrorist groups is understudied in comparative

criminology, but recent research suggests that the size and political/religious ideology of the group contribute to their longevity (LaFree and Ackerman, 2009).

Individual-level research on terrorism has examined a range of possible causes, including psychopathology, personality, trauma and alienation, personal status, moral neutralization, demographics, employment, education, marriage, group-level grievances, and socialization (LaFree and Ackerman, 2009). In their review of potential causes, LaFree and Ackerman find little empirical evidence to support psychopathological perspectives, which propose that terrorists disproportionally suffer from mental disorders, cognitive maladaptations, or other risk factors associated with suicide (in the case of suicide bombers). They found general support for a relationship between individual trauma or alienation and terrorist attitudes and behaviors. They argue that "while not genuinely causal in nature, trauma, perceived injustice, and alienation can be termed facilitative variables in that they may act to make individuals more antagonistic toward outgroups, more vulnerable to terrorist ideologies, or more open to mechanisms of self-justification for the use of violence" (LaFree and Ackerman, 2009: p. 350). While there is evidence that terrorists tend to be young males, other known criminological correlates of crime are not supported by evidence: employment, marriage, parenthood, and education. More generally, they find that comparative research on individual-level causes and correlates of terrorist activities is vastly underresearched.

Comparative and international research is also useful for illuminating the universal and culturally specific dangers women face across the globe. Women are more likely to be killed by a family member or spouse than men in the United States, Israel, Canada, South Africa, Australia, and across Europe (UNODC, 2011). UNODC (2011: p. 58) show that for a selection of European countries, 77% of all victims killed by their spouses or exspouses were women. This suggests that situational risk factors for female victimization are similar across countries; however, the form of victimization may vary.

The *Global Burden of Armed Violence* discusses a number of 'hidden forms' of violence against women, including 'honor killings,' 'dowry deaths,' witchcraft killings, gender-related forced suicide, and genital mutilation. Official data on violence against women likely underestimate the true levels due to underreporting and the 'hidden' nature of domestic violence. Because of this, comparative researchers often rely on secondary sources and media outlets to gain information on the characteristics and levels of 'hidden forms' of violence against women.

For example, although honor killings have existed in many different cultures for centuries, reliable data on honor killings are scarce (Goldstein, 2002). Goldstein argues that the absence of data is possibly due to an unwillingness to report crimes, a lack of legal recognition of honor crimes, and the tendency for honor killings to occur during times of war. Country-level estimates of the number of honor killings range from 400 in Yemen in 1997, to over a 1000 in Pakistan in 1999, to over 5000 worldwide each year (Goldstein, 2002: p. 31). According to the United Nations Commission on Human Rights (see Jahangir, 2000), honor killings have been recorded in a diverse range of countries, including Bangladesh, Ecuador, Great Britain, India, Israel, Jordan, and Turkey.

Sev'er and Yurdakul (2001) used a sample of honor killings reported in the media and by other academics to examine the characteristics of female honor killings in Turkey. The authors (2001: pp. 964–965, citations omitted) define honor killing as "the premeditated murder of preadolescent, adolescent, or adult women by one or more male members of the immediate or extended family." They further describe the circumstances under which honor killings tend to occur in Turkey: "These killings are often undertaken when a family council decides on the time and form of execution due to an allegation, suspicion, or proof of sexual impropriety by the victim" (p. 965). They found that most honor killings occurred in rural populations characterized by low incomes, low education, early marriages, and a 'feudal type of land ownership' that leaves "nothing to the majority of disenfranchised men (except power over women and children)" (Sev'er and Yurdakul, 2001: p. 989).

Survey-based studies of honor killing are rare. However, recently Eisner and Ghuneim (2013) conducted one of the first studies to use psychometric instruments to measure attitudes toward honor killing. Among a convenience sample of 856 ninth graders in Amman, Jordan, the authors found that 46.1% of boys agreed with items stating "It is okay for a man to kill his daughter [sister, wife] if she has dishonored her family" (p. 411). By contrast, only 22.1% of girls supported honor killings. Adolescents who supported honor killings were more likely to come from low education and 'breadwinner'-type families (i.e., mother is a house wife), espouse traditional attitudes (i.e., patriarchal and anti-Western values), believe that females should remain chaste, and rely on moral neutralization techniques to excuse violence more generally (pp. 410–412). The authors conclude that "any meaningful attempt to reduce attitudes in support of honor killings therefore probably requires a broader societal commitment to changing the underlying cultural support for patriarchal dominance, including moral justification of private violence against transgressors" (p. 415).

Conclusion

In sum, geographical and cultural regions face many different challenges to social order, but there are also seemingly common characteristics and risk factors found in different contexts. Across the globe, men are more likely to be the perpetrators and victims of lethal violence, whereas women are more likely to be victims of lethal domestic violence. Certain individual and socioeconomic risk factors of crime, such as moral neutralization, income inequality, poverty, and lack of education are likely related to a range of criminal activities, including honor killing, terrorism, organized crime, and gang violence. Finally, crimes tend to be spatially concentrated. Within Europe, crime is generally higher in Eastern Europe; however, east/west trends are likely to converge, and homicide tends to cluster in Central and South America as well as in sub-Saharan Africa. A review of violence in Mexico revealed that crimes are likely further spatially concentrated within countries. Terrorist attacks are concentrated in Latin America, Western Europe, and the Middle East.

Although this suggests criminal behaviors have some common roots across countries, more comparative and

international research is needed to explore universal and culturally specific risk factors and contexts of crime. It is important to remember that universality and cultural relativism are not inherently antagonistic to each other. In Karstedt's (2001: p. 302) words,

> we should shift our perspective systematically from differences toward universals and vice versa. Indigenization does not imply that universals or a universal theory are useless, but to the contrary: It directs attention to the culture-bound quality of those theories that we rate as universal theories, or directs the search toward hitherto neglected and potential universals. Any such tendencies toward such dichotomies are inimical to the development of the discipline.

What comparative and international criminology makes clear is that "we are a criminally challenged planet" (Adler, 1996: p. 3). While there are many methodological and theoretical problems to overcome, comparative research is nevertheless imperative to our understanding of crime.

See also: Censuses, Population: Comparative International Aspects; Comparative Studies: Method and Design; Drug Control: International Aspects; Homicide; Honor Killing: Multidimensional and Multilevel Perspectives; Inequality: Comparative Aspects; Policing in Divided Societies; Terrorism; Transnational Organized Crime.

Bibliography

Adler, F., 1996. Our american society of criminology, the world, and the state of the art – the american society of criminology 1995 presidential address. Criminology 34, 1–9.

Aebi, M.F., 2006. Raw Data for the 3rd Edition (2006) of the European Sourcebook of Crime and Criminal Justice Statistics 2000–2003. http://www.europeansourcebook.org/ (accessed 16.03.13.).

Africa Research Bulletin, March 2013. Nigeria: Boko Haram Truce remains elusive. Africa Research Bulletin: Political, Social, and Cultural Series 50, 19603C–19605C.

Andresen, M.A., 2011. The impact of accession to the european union on violent crime in Lithuania. European Sociological Review 27, 759–771.

Ceccato, V., 2008. Expressive crime in post-socialist states of Estonia, Latvia and Lithuania. Journal of Scandinavian Studies in Criminology and Crime Prevention 9, 2–30.

Chamlin, M.B., Cochran, J.K., 2005. Ascribed economic inequality and homicide among modern societies: toward the development of a cross-national theory. Homicide Studies 9, 3–29.

Cohen, D., Nisbett, R.E., 1994. Self-Protection and the culture of honor: explaining southern violence. Personality and Social Psychology Bulletin 20, 551–567.

Du Toit, P., 2001. South Africa's Brittle Peace: The Problem of Post-Settlement Violence. Palgrave, London.

Eisner, M., 2008. Modernity strikes back? A historical perspective on the latest increase in interpersonal violence (1960–1990). International Journal of Conflict and Violence 2, 288–316.

Eisner, M., Ghuneim, L., 2013. Honor killing attitudes amongst adolescents in Amman, Jordan. Aggressive Behavior 39, 405–417.

Eisner, M., Nivette, A.E., 2012. How to reduce the global homicide rate to 2 per 100,000 by 2060. In: Loeber, R., Welsh, B.C. (Eds.), The Future of Criminology. Oxford University Press, Oxford, pp. 219–228.

Gatti, U., Haymoz, S., Schadee, H.M.A., 2011. Deviant youth groups in 30 countries: results from the second international self-report delinquency study. International Criminal Justice Review 21, 208–224.

Geneva Declaration Secretariat, 2011. Global Burden of Armed Violence 2011: Lethal Encounters. Cambridge University Press, Cambridge.

Goldstein, M.A., 2002. The biological roots of heat-of-passion crimes and honor killings. Politics and the Life Sciences 21, 28–37.

Gottfredson, M.R., Hirschi, T., 1990. A General Theory of Crime. Stanford University Press, Stanford.

Gruszczyńska, B.Z., 2004. Crime in central and eastern european countries in the enlarged Europe. European Journal on Criminal Policy and Research 10, 123–136.

Gruszczyńska, B.Z., Heiskanen, M., 2011. Trends in police-recorded offences. European Journal on Criminal Policy and Research 18, 83–102.

Jahangir, A., 2000. Civil and political rights, including questions of: disappearances and summary executions: report of the special rapporteur. United Nations Commission on Human Rights. http://daccess-ods.un.org/TMP/1064358.35361481.html (accessed 27.09.13.).

Junger-Tas, J., Enzmann, D., Steketee, M., Marshall, I.H., 2012. Concluding observations: the big picture. In: Junger-Tas, J., Marshall, I.H., Enzmann, D., Killias, M., Steketee, M., Gruszczyńska, B. (Eds.), The Many Faces of Youth Crime. Contrasting Theoretical Perspectives on Juvenile Delinquency across Countries and Cultures. Springer, New York, pp. 329–353.

Junger-Tas, J., Marshall, I.H., Enzmann, D., Killias, M., Steketee, M., Gruszczyńska, B. (Eds.), 2012. The Many Faces of Youth Crime. Contrasting Theoretical Perspectives on Juvenile Delinquency across Countries and Cultures. Springer, New York.

Karstedt, S., 2001. Comparing cultures, comparing crime: challenges, prospects and problems for global criminology. Crime, Law & Social Change 36, 285–308.

LaFree, G., 2006. Expanding criminology's domain: the american society of criminology 2006 presidential address. Criminology 45, 1–31.

LaFree, G., Ackerman, G., 2009. The empirical study of terrorism: social and legal research. Annual Review of Law and Social Science 5, 347–374.

LaFree, G., Dugan, L., 2007. Introducing the global terrorism database. Terrorism and Political Violence 19, 181–204.

LaFree, G., Morris, N.A., Dugan, L., 2010. Cross-national patterns of terrorism: comparing trajectories for total, attributed and fatal attacks, 1970–2006. British Journal of Criminology 50, 622–649.

Linke, A., Raleigh, C., 2011. State and stateless violence in Somalia. African Geographical Review 30, 47–66.

Marchal, R., 2009. A tentative assessment of the Somali Harakat al-Shabaab. Journal of Eastern African Studies 3, 381–404.

Marshall, I., Block, R., 2004. Maximizing the availability of cross-national data on homicide. Homicide Studies 8, 267–310.

Marshall, I.H., Summers, D.L., 2012. Contemporary differences in rates and trends of homicide among European nations. In: Liem, M.C.A., Pridemore, W.A. (Eds.), Handbook of European Homicide Research. Springer, London, pp. 39–70.

Murray, J., Farrington, D.P., Eisner, M., 2009. Drawing conclusions about causes from systematic reviews of risk factors: the Cambridge quality checklists. Journal of Experimental Criminology 5, 1–23.

National Consortium for the Study of Terrorism and Responses to Terrorism, 2013. Global Terrorism Database. http://www.start.umd.edu/gtd/ (last accessed 14.07.14.).

Neapolitan, J.L., 1994. Cross-national variation in homicides: the case of Latin America. International Criminal Justice Review 4, 4–22.

Neapolitan, J.L., 1997. Cross-National Crime: A Research Review and Sourcebook. Greenwood Publishing Group, Westport, CT.

Nivette, A., 2011. Cross-national predictors of crime: a meta-analysis. Homicide Studies 15, 103–131.

Ouimet, M., 2012. A world of homicides: the effect of economic development, income inequality, and excess infant mortality on the homicide rate for 165 countries in 2010. Homicide Studies 16, 238–258.

Pridemore, W.A., 2008. A methodological addition to the cross-national empirical literature on social structure and homicide: a first test of the poverty-homicide thesis. Criminology 46, 133–154.

Pridemore, W.A., Trent, C.L.S., 2010. Do the invariant findings of Land, McCall, and Cohen Generalize to cross-national studies of social structure and homicide? Homicide Studies 14, 296–335.

Rebellon, C.J., Straus, M.A., Medeiros, R., 2008. Self-control in global perspective: an empirical assessment of Gottfredson and Hirschi's general theory within and across 32 national settings. European Journal of Criminology 5, 331–362.

Ribeaud, D., Eisner, M., 2010. Risk factors for aggression in pre-adolescence: risk domains, cumulative risk and gender differences – results from a prospective longitudinal study in a multi-ethnic urban sample. European Journal of Criminology 7, 460–498.

Sev'er, A., Yurdakul, G., 2001. Culture of honor, culture of change: a feminist analysis of honor killings in rural Turkey. Violence Against Women 7, 964–998.

Shelley, L.I., 1981. Crime and Modernization: The Impact of Industrialization and Urbanization on Crime. Southern Illinois University Press, Carbondale.

Steketee, M., Junger, M., Junger-Tas, J., 2013. Sex differences in the predictors of juvenile delinquency: females are more susceptible to poor environments; males are influenced more by self-control. Journal of Contemporary Criminal Justice 29, 88–105.

Stamatel, J.P., 2009. Contributions of cross-national research to criminology at the beginning of the 21st century. In: Krohn, M.D., Lizotte, A.J., Penly Hall, G. (Eds.), Handbook on Crime and Deviance. Springer, London, pp. 3–22.

Stamatel, J.P., 2008. Using mortality data to refine our understanding of homicide patterns in select postcommunist countries. Homicide Studies 12, 117–135.

United Nations Office on Drugs and Crime, 2011. Global Study on Homicide 2011. United Nations Office on Drugs and Crime, Vienna, Austria.

Van Dijk, J., 2012/2013. The international crime victims survey: latest results and prospects. Criminology in Europe 11, 24–33.

World Health Organization, 2012. Disease and Injury Country Estimates. http://www.who.int/healthinfo/global_burden_disease/estimates_country/en/index.html (accessed 18.03.13.).

World Health Organization, 2010. International Classification of Diseases and Related Health Problems 10th Revision. http://apps.who.int/classifications/icd10/browse/2010/en (accessed 16.03.13.).

Zimring, F., 2007. The Great American Crime Decline. Oxford University Press, Oxford.

Relevant Websites

http://www.acleddata.com/ – Armed Conflict Location and Event Database.

http://www.europeansourcebook.org/ – European Sourcebook of Crime and Criminal Justice Statistics.

http://www.start.umd.edu/gtd/ – Global Terrorism Database.

http://www.systemicpeace.org/inscr/inscr.htm – Integrated Network for Societal Conflict Research: Data Page.

http://www.unodc.org/ – United Nations Department on Drugs and Crime: Data and Analysis.

http://data.un.org/ – United Nations Data [General].

http://www.unicri.it/ – United Nations Interregional Crime and Justice Research Institute.

http://www.who.int/healthinfo/global_burden_disease/en/ – World Health Organization: Global Burden of Disease.

International Arbitration

Bryant G Garth, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article defines international arbitration, distinguishing between state–state arbitration, international commercial arbitration, and investment arbitration. It shows how international commercial arbitration became the default dispute resolution mechanism for transnational commercial disputes in the 1980s, and how the field has prospered since that period. The article relates international commercial arbitration to the alternative dispute resolution movement, and it concludes with a discussion of how one gets to be an arbitrator trusted to handle very high-stakes commercial disputes.

'International arbitration' refers to the private mechanisms for resolving either disputes between states or transnational business disputes. Since the 1960s, the system of international commercial arbitration has grown dramatically in importance, becoming the default dispute resolution device provided for in international commercial contracts. There is a relatively small group of international commercial arbitrators who handle most of the significant cases. The system began as a relatively informal system, close to the procedures of the European Continent. It has greatly expanded, due to the rise of trade and the increasing legitimization of this system around the world. With this expansion, the system has become much more adversarial, contested, and document intensive, due in part to the pressures of US litigators who are playing a larger role in the procedures. International commercial arbitration involves hundreds of international commercial arbitration centers competing to have the most prominent rosters of arbitrators and the largest number of high-stakes international cases, but the leading organization has long been the International Chamber of Commerce (ICC) in Paris. International arbitration can be used to resolve commercial disputes, disputes between business entities and states, or disputes between states. Arbitration procedures for disputes between states, exemplified by disputes about borders, are ad hoc and infrequently invoked. Of much more practical relevance since the 1980s is the system of private justice created to resolve international business disputes, which has become the standard system for international transactions. When two businesses with headquarters in different countries enter into a contractual arrangement, the contract will generally provide that any disputes relating to the contract will be resolved through private arbitration.

Since the 1990s, in addition, there have been increasing numbers of investment arbitrations created under treaties such as the North American Free Trade Agreement and the proliferating number of Bilateral Investment Treaties (BIT) – estimated now at close to 3000. Disputes between investing countries and host states governed by a BIT are typically referred to the International Center for Settlement of Investment Disputes (ICSID) operated under the auspices of the World Bank. The arbitrators who handle ICSID disputes are in general the same ones who staff the major international commercial arbitrations.

The number of international commercial arbitrations is impossible to determine precisely since many of the arbitrations are not administered by any formal organization, and there is no clear line between domestic and international. International commercial arbitration may be institutional, under the auspices of one of the institutions competing for the business of international commercial arbitration, or it may be ad hoc, handled solely by the parties and their counsel. It is especially difficult to count the ad hoc arbitrations.

It is clear, however, that the number of international commercial arbitrations has been increasing generally and in relation to the global financial crisis that began in 2008. The most important of the institutions is the ICC in Paris, France, and it generally has the largest caseload and the cases involving the highest stakes. In 2011, some 796 requests for arbitration were filed with the ICC, concerning 2293 parties from 139 different countries. The places of arbitration assigned to the cases involved 63 different countries on 5 continents. Arbitrators of 78 different nationalities were appointed or confirmed under the ICC Rules, and the amount in controversy exceeded US$1 million in 77% of the new cases. The ICC Court, which confirms the awards, had 508 new awards submitted in that year.

In addition to the ICC in Paris, other leading centers are the London Court of International Commercial Arbitration, the American Arbitration Association, the Chinese International Economic and Trade Arbitration Commission, the Hong Kong International Arbitration Center, the Singapore International Arbitration Center, the Korean Commercial Arbitration Board, the Japan Commercial Arbitration Association, and the Stockholm Chamber of Commerce. As noted above, ICSID is the major institution for investment arbitration. The ICC is especially useful for relatively new participants to international commercial arbitration. The ICC, while more expensive, helps to guide new parties and new arbitrators through the processes, and it also has a reputation as an institution that can ensure that its awards are respected in domestic courts.

International commercial arbitration – as arbitration generally – involves a process whereby a third party issues an opinion that is formally binding on the parties. Unlike domestic arbitration (as it has traditionally been practiced in the United States), international commercial arbitrators typically write formal opinions. These formal opinions typically are

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86039-0

neither published nor made public informally. They belong to the parties, who must therefore consent to any publication. The awards made by the arbitrators are binding and very difficult to challenge or appeal. The New York Convention of 1958, which has been adopted by well over 100 states, makes an arbitral award in practice more easily enforced within a contracting state than would be a judgment obtained through litigation. In addition, the procedures of international commercial arbitration have been converging around a set of rules that tend to detach arbitration from its site and to emphasize party autonomy. The United Nations Commission on International Trade Law model law has been the major model for this convergence.

The most prominent feature of international arbitration is that the arbitrators are private individuals without any formal connection to the state or a state judicial system. In theory, almost anyone could be selected as an arbitrator. In practice, however, there is a relatively small group that tends to dominate selection for major arbitrations. The group consists largely of European lawyers, professors, and retired judges, but there are also prominent arbitrators from other parts of the world.

The advantages of arbitration in international disputes are usually listed as follows: the ability to select elite private judges, the secrecy of the proceedings, and the ability to avoid having to submit to the judicial system of the opposing party. Not generally listed as advantages are the cost and the duration of the proceedings, since in fact arbitrations may cost as much as or more, and take as much as or more time, than litigation in a national court system.

A number of interesting sociolegal issues relate to international commercial arbitration. They include (1) how this system of private justice came to be considered the legitimate way to resolve international business disputes, (2) what this system means for issues concerned with alternative dispute resolution more generally, and (3) how people become international commercial arbitrators.

Building the International Arbitration System

International business was long a world of complex personal relations rather than formal laws and dispute resolution processes. The number of actors was relatively small in any given field, and the actors could get to know and take account of each other. There might be contracts, but the contracts were relatively unimportant documents either in the negotiation or when disputes might arise. Histories of the petroleum industry, in particular, reveal the key actors behind the 'seven sister' oil companies and how they interacted with the countries in which they did business. Lawyers were very much part of this history, but lawyers for the oil companies – including John McCloy, the famous lawyer for the Rockefeller family who was termed the 'chairman of the establishment' – relied much more on personal influence and connections than on their expertise as lawyers.

When this relatively stable set of relationships was shaken up, especially by new entrants in the 1960s and 1970s willing to offer more favorable terms to produce the oil, the lawyers for the oil companies used contract and international law to try to hold on to their privileged positions. At the same time, a small number of elite actors in the exporting countries had taken advantage of oil company paternalism to obtain education abroad – in law and engineering, for example. They and a few maverick legal advisors began to formulate legal arguments that slowly built the legal infrastructure of the industry. The major transformation of dispute resolution, however, followed the nationalization of the oil industries in most of the oil-exporting countries. The oil concession agreements typically had arbitration clauses, because arbitration had long been practiced in disputes between nations. When the oil industry was nationalized in places such as Kuwait, Libya, and Saudi Arabia, the companies invoked the arbitration clauses and began arbitration processes.

Legal accounts of this period typically celebrate the grand arbitrations that resulted, suggesting that important legal principles supporting the stability of contracts trumped the political acts of nationalization. Other accounts, including those by journalists and historians, however, do not mention the arbitrations at all. They instead emphasize the negotiations that took place between those who had long invested in the relationships, the negotiations between governmental actors at another level, and the gunboats that were exerting not-so-subtle pressure on behalf of the oil companies. Lawyers involved at the time suggest that their activity in the arbitrations was not highly valued or even much noticed by company executives.

The arbitration processes did go forward, however, and they produced raw material that equipped international business with people and principles to handle transnational disputes. The raw material was in the form of written opinions that were widely circulated, academic articles about the principles in the opinions, and individuals who gained or enhanced their reputations as potential arbitrators by virtue of their participation in the arbitrations. The petroleum arbitrations thus contributed to the production and legitimization of international commercial arbitration for transnational business disputes.

When new sets of problems arose that related to some of the themes of the petroleum arbitrations, in particular disputes involving the numerous construction projects that took place in oil-producing countries after the price of oil increased in the early 1970s, international commercial arbitration played a crucial role. The Continental academics who developed these principles into a new *lex mercatoria* also developed relationships with students from the less developed world, and the students helped to reinforce the legitimacy of these principles in their own countries. Put very simply, the process produced lawyers on both sides, legal principles, and a legitimacy that made it possible to generate a demand for what this group supplied. For the ICC, this increased demand meant that while the first 3000 requests for arbitration came between 1923 and 1976, the next 3000 came in the following 11 years.

International Commercial Arbitration and Alternative Dispute Resolution

The world of international commercial arbitration was dominated in the 1960s and 1970s by Continental academics who wrote about the *lex mercatoria*, and by arbitrators who

encouraged the parties to find a way to settle the dispute. The *lex mercatoria* could be used by the arbitrators to find a solution that fit the business relationship and left the parties satisfied with the result. Consistent with the general orientation of the leading arbitrators and lawyers associated with the pioneers of international commercial arbitration, the process was not very adversarial, was relatively inexpensive except for the arbitrators, did not tend to produce extensive documentary evidence, and did not contemplate such US practices as cross-examination of witnesses. The model was quite consistent with Continental practices, even when non-European actors participated in the processes.

When the arbitration business began to take off in the 1970s and 1980s, it coincided with the rise of business litigation in the United States. The US law firms in Paris, France, were long involved in international commercial arbitration, but they tended to play according to the Continental rules. As the practice of international commercial arbitration grew and expanded outside of a relatively small circle of people, US businesses became increasingly involved in arbitration, and developing countries found it to be in their interests to hire US law firms as well. The enhanced presence of US litigators had important consequences. They naturally sought to use the techniques of discovery, intensive production of documents, cross-examination of witnesses, and more generally the kind of aggressive adversarial behavior that was beginning to characterize litigation in the United States. They also sought to retain arbitrators who would permit the lawyers to use the techniques that they had developed. The result was that the processes of international commercial arbitration changed substantially. They did not become identical to US litigation proceedings, but they moved much closer to US adversarialism. International commercial arbitration became much more document intensive, much more adversarial, and much more expensive.

Litigation in the United States had also been transformed, and that transformation had produced the beginnings of the alternative dispute resolution movement in that country. New organizations such as the Center for Public Resources had begun to encourage corporations to practice more mediation as a way to avoid the high costs and adversarial nature of business litigation. The movement became quite influential in the United States, and some of the individuals associated with it began to suggest that it would also be useful for international commercial arbitration. Not surprisingly, the first reaction of many within the international arbitration community was that they already practiced mediation as part of the normal process. They argued that the practice of international arbitration provided an effective alternative to litigation as practiced in the United States.

As international commercial arbitration became more like US litigation, however, the US-promoted alternative began to gain more adherents within the international arbitration community. By the end of the twentieth century, the menu of US alternatives – including mediation but a range of others as well – had become much more common. This gradual transformation illustrates the way that the formal categories of dispute resolution – litigation, arbitration, and even mediation – can disguise profound transformations in the way that the underlying practices proceed. The center of gravity of international commercial arbitration moved much closer to US-style litigation, which was itself a relatively new invention, and the shift also helped to bring the US antidote, alternative dispute resolution.

Becoming an Arbitrator

The international commercial arbitration community is relatively closed, and it is often termed a 'mafia' even by insiders. One reason for this 'club' is simply that for high-stakes arbitration, the parties typically want to select arbitrators who are known quantities. Untested arbitrators are very risky from the point of view of the parties. It is also deemed important to select arbitrators who are known for their ability to 'speak the language' and be taken seriously within an arbitral tribunal. Since arbitrators often serve as counsel because of their familiarity with the processes and people, it is common for people from the relatively closed arbitral community to nominate others from the community. The key to success in the arbitration proceeding is selecting an arbitrator who will be impartial enough to persuade the chair of the tribunal but also able to understand and put forward the case of the party that appointed him or her. The selection of established names also serves a protective function in case the decision does not favor the selecting party. It is easier to avoid blame for a safe selection than for a novel one. For many reasons, including the fact that a relatively small supply serves the economic interests of the arbitrators, it is not easy to gain entrance into this community.

The easiest way to gain entrance for at least some small cases that may provide a basis to build a practice is to conform to the characteristics of those already in the field. The individuals who set the norms emphasized scholarly accomplishments, which put an emphasis on publications, cosmopolitanism, and linguistic ability. Those who seek to enter must probably also attend a number of the major conferences in order to display the appropriate characteristics and be judged more personally. However, the general point is that invitations to enter the community tend to be issued most easily to those who can offer something to the existing community – most obviously persons of high status who reflect that status back on the arbitral community, persons from areas that are not yet committed to the legitimacy of international commercial arbitration (which means they offer both credibility and a potential new caseload), or sometimes disciples of those already in the arbitral community. This process helps to explain why individuals without high status may study, write, and attend conferences, but lacking a prestigious platform, they cannot gain entry. On the other hand, individuals from developing countries who serve as judges on the World Court do typically gain entry. The process also helps explain why there are so many arbitration centers despite the fact that the caseloads are concentrated in a relatively few places. Those who create new centers gain the attention of the elite of the arbitral community and also help to provide new domains where businesses will contemplate arbitration and courts will respect the awards. As a result, the new center may promote the local leader into a position closer to the core of the arbitral community. There tend to be relatively few of these peripheral arbitrators in each relevant site.

International arbitration, especially through commercial contracts and investment treaties, is thriving today and has

been steadily gaining importance in new terrains that were once resistant to arbitration, such as Latin America. The arbitration community is still relatively small, but it has evolved and gained legitimacy through the addition of at least a small number of arbitrators from outside the major capitals of the global North.

See also: Intergenerational Justice; International Business; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; Laws, Conflict of; Lex Mercatoria; Public Goods: International.

Bibliography

Blackaby, N., Partasides, C., Redfern, A., Hunter, M., 2009. Redfern and Hunter on International Arbitration. Oxford University Press, New York/Oxford.

Craig, W., Park, W.W., Paulsson, J., 2001. International Chamber of Commerce Arbitration. Oceana Publications, Dobbs Ferry, NY.

Dezalay, Y., Garth, B., 1996. Dealing in Virtue: International Commercial Arbitration and the Construction of a Transnational Legal Order. University of Chicago Press, Chicago, IL.

Paulsson, J., Rawding, N., Reed, L., 2010. Guide to ICSID Arbitration, second rev. ed. Kluwer, Dordrecht, The Netherlands.

International Association of Schools of Social Work

Vimla V Nadkarni, International Association of Schools of Social Work, School of Social Work, Tata Institute of Social Sciences, Deonar, Mumbai, India
Tetyana Semigina, School of Social Work, National University of "Kyiv-Mohyla Academy", Kyiv, Ukraine

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Founded in 1928, the International Association of Schools of Social Work is the worldwide organization representing social work education. Comprising member schools and individuals across all continents, it works, in spite of funding and voluntary leadership challenges, to create a globally inclusive organization and promote curriculum development, research, and international exchanges.

Brief Overview

The International Association of Schools of Social Work (IASSW) is the worldwide association of schools of social work, other tertiary level social work educational programs, and social work educators. The IASSW promotes the development of social work education throughout the world, develops standards to enhance quality of social work education, encourages international exchanges, provides forums for sharing social work research and scholarship, and promotes human rights and social development through policy and advocacy activities (IASSW, 2013).

Membership in the IASSW comprises social work educational programs at the postsecondary or tertiary level and interested individuals, mostly social work educators. The Association has member schools in all parts of the world: five regional organizations in Africa, Asia and the Pacific, Europe, Latin America, and North America and the Caribbean; and three subregional associations – the Nordic subregion, the Caribbean subregion, and Eastern European subregion are affiliated with the IASSW. IASSW is governed by the Board of Directors under a Constitution approved by the biennial General Assembly. IASSW holds and, has held since 1947, consultative status with the United Nations (UN) and participates as a nongovernmental organization (NGO) in UN activities in Geneva, Vienna, and New York and other parts of the world. Through its work at the UN and with other international organizations, IASSW represents social work education at the international level.

According to Tasse (2008), some of the key features of the IASSW include: developing visions of the profession that enable it to rise to meet major challenges at critical point of history; articulating the voice of the profession in key international institutions and forums; disseminating best practices in the profession through social work education; and promoting international and comparative research.

The IASSW was founded in 1928 and is one of the largest and oldest social work professional organizations. It is a membership organization based on voluntary commitments. Over the years, IASSW has experienced turmoil caused by the expulsion of its first president by the Nazi regime in Germany, World War II, the termination of social work programs in Eastern Europe and China under Communism, internal disputes over the proper course of action in dealing with apartheid South Africa, and financial difficulties that resulted in the loss of a secretary-general and secretariat in the early 1990s. Through it all, however, under the leadership of its Presidents and Board, IASSW has remained engaged in its core purposes of promoting and representing social work education through international exchanges and advocacy. This history augurs well for the future of social work education.

IASSW Mission

The mission statement adopted in 2000 commits the IASSW as the international organization to four major purposes: (1) to develop and promote worldwide excellence in social work education, research and scholarship in order to enhance human well-being; (2) to create and maintain a dynamic community of social work educators and their programs; (3) to support and facilitate participation in mutual exchanges of information and expertise; and (4) to represent social work education at the international level.

In fulfilling its mission, members adhere to all UN Declarations and Conventions on human rights, and recognize that respect for the inalienable rights of the individual is the foundation of freedom, justice, and peace. Additionally IASSW fosters cooperation, collegiality, and interdependence among members and with others, engages in the pursuit of social justice and social development.

The IASSW carries out its mission through:

- developing global standards; and other documents including the ethics document and international definition of social work in collaboration with the International Federation of Social Workers (IFSW);
- organizing conferences for social work educators;
- organizing training programs and curriculum consultations;
- publications and usage of social media;
- representation at the UN;
- activities of Committees and Task Forces;

- funding of cross-national projects in social work education; and
- collaborating with other international social work organizations.

IASSW Current Membership

Currently, the IASSW has both school/social work program memberships and individual memberships. School membership is open to tertiary level education programs that prepare students for the profession of social work. Individual membership is open to any social work educator or other persons interested in social work education. Members of all categories must, to the best of their ability, safeguard the interests of the IASSW, pay the stipulated dues and act in accordance with the by-laws and resolutions of the IASSW which include abiding by UN Conventions on human rights, children, women, and others linked to the equal rights of minority ethnic groups, indigenous peoples, and disabled people among others.

Members of the IASSW have rights to:

- participation in ongoing projects related to social themes – poverty, macrodevelopment, peace, and human rights, ecology, women, children, AIDS and others, as well as activities of the IASSW committees and task forces;
- reduced registration fee at biennial international congresses organized by the IASSW;
- free subscription to the IASSW Listserve and *Social Dialogue* magazine;
- free subscription to *International Social Work* journal (for IASSW individual members) and discounted subscription to *International Journal of Social Welfare* (for IASSW individual members); and
- accessibility to IASSW funding regarding joint projects with other schools of social work, designed to advance IASSW Mission and to the enhancement of cooperation among schools of social work worldwide.

Each school has a vote at the General Assembly held biennially. Individual members have the same privileges as mentioned above and a prorata right to vote at the General Assembly.

IASSW has around 500 members from 62 countries. Its largest member schools are from Japan, China, and the United States. The membership across the regions is as in Figure 1, although this is subject to fluctuations.

IASSW Leadership and Governing Bodies

According to the current IASSW Constitution, the organization is governed by the Board of Directors, made up of both elected and appointed members. The nominating Committee prepares slates of candidates for the biennial elections of officers and at-large board members. Ballots are sent to all members. Regional representation is ensured through the inclusion of five regional vice presidents on the Executive Committee and Board of Directors. Currently, three subregional associations are also

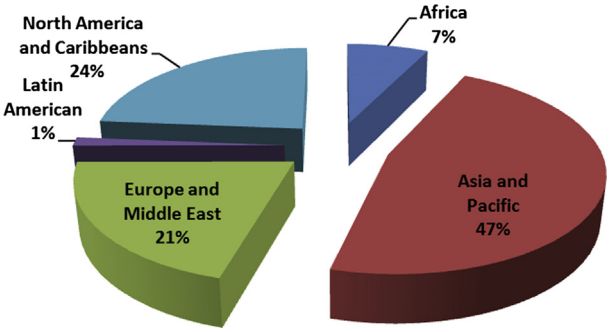

Figure 1 Distribution of 2013 IASSW membership across the regions ($N = 507$).

recognized, and each can elect a member to the Board of Directors: the Association of Caribbean Social Work Educators, the Nordic Association of Schools of Social Work, and East European Association of Schools of Social Work. National associations are entitled to a board representative if the country has at least five IASSW member schools. The Board of Directors carries responsibility for the ongoing direction of the association and development and oversight of the budget.

According to Healy (2008) the leaders of the IASSW have played major roles in shaping the organization over its 80-year history. Starting from 1928 the IASSW had 12 Presidents (see a full list in Table 1).

For more than 50 years they were from Europe, USA, and Canada. This reality is a function of the European/North Atlantic origins and the composition of the organization with the same line of interest for many years. The first President outside of Europe and North America (Abye Tasse from Ethiopia) was elected in 2004.

From 1978 to 2010 the IASSW had a position of the Honorary President. It was occupied by Katherine A. Kendall (USA) who served as the first paid Secretary General of the IASSW in the 1970s and was heavily involved in IASSW activities until her death in 2010 (CSWE, 2013). The position has remained vacant since.

Major policy changes and significant changes to the constitution require action by delegates at the General Assembly, which meets biennially during the world conference. All IASSW members participate in the Assembly, with programs

Table 1 A list of IASSW presidents (1928–2012)

Period	*Name of the president*
1928/29–1946	Alice Salomon (Germany, then – USA)
1946–1953	René Sand (Belgium)
1954–1961	Jan Floris de Jongh (the Netherlands)
1961–1968	Eileena Younghusband (United Kingdom)
1968–1976	Herman D. Stein (USA)
1976–1980	Robin Huws Jones (United Kingdom)
1980–1988	Heiner Schiller (Germany)
1988–1996	Ralph Garber (Canada)
1996–2004	Lena Dominelli (United Kingdom)
2004–2008	Abye Tasse (Ethiopia)
2008–2012	Angelina Yuen (Hong Kong)
2012–2016	Vimla Nadkarni (India)

having one full vote each and individual members having one-tenth of a vote.

Social work educators and practitioners considered professional and educational issues around thematic concerns at its biennial international congresses which were held in: 1928 Paris; 1932 Frankfurt; 1936 London; 1948 Atlantic City; 1950 Paris; 1952 Madras; 1954 Toronto; 1956 Munich; 1958 Tokyo; 1960 Rome; 1962 Brazil; 1964 Athens; 1966 Washington; 1968 Helsinki; 1970 Manila; 1972 The Hague; 1974 Nairobi; 1976 Puerto Rico; 1978 Jerusalem; 1980 Hong Kong; 1982 Brighton; 1984 Montreal; 1986 Tokyo; 1988 Vienna; 1990 Lima; 1992 Washington; 1994 Amsterdam; 1996 Hong Kong; 1998 Jerusalem; 2000 Montreal; 2002 Montpellier; 2004 Adelaide; 2006 Santiago de Chile; 2008 Durban (Seibel, 2008); 2010 Hong Kong; 2012 Stockholm; 2014 Melbourne.

During the congress held in Amsterdam (the Netherlands) in 1994, IASSW changed its constitution and structure to devolve greater powers to its regions.

The biennial international congress in Hong Kong (2010), Stockholm, Sweden (2012), and Melbourne, Australia (2014) were co-organized by IASSW, the IFSW, and the International Council on Social Welfare (ICSW). IASSW shares with these organizations the same purpose of working to increase the voice of the social work profession. Similar arrangements had been in place earlier during the 1992, 1996, and 1998 Congresses. For internal structural reasons, ICSW withdrew from the 2000 tripartite arrangements, and IASSW and IFSW held their joint Congress in Montreal.

Work of the association is furthered through several committees and task forces that are clustered under four major themes: Education, Advocacy, Administration, Budget and Finance. Since 2012, the Education Committees cluster includes capacity building, research, and international projects. The Advocacy Committee cluster covers sustainability, climate change, disaster intervention; communications and publications; human rights and the global agenda. The administration cluster comprises the Nominating Committee, Language Committee, and the Standing Committee on World Census. The IASSW has strong, active lead representatives at the UN.

IASSW Language Policy and Publications

As IASSW is open to all social work educational programs from all over the world, its aim is to practice an inclusive language policy, and refrain from becoming an elite organization dominated only by member institutions from the western world. Language policy and practice is an important factor for deciding whether this is an attractive organization for a global target group.

IASSW has four official languages: English, French, Spanish, and Japanese. These are the languages of the majority number of IASSW member schools. However, English is the dominant one due to both lack of financial and human resources for translations. At one point, IASSW's language policy included Chinese, Russian, and Arabic, but this also proved financially unfeasible and now the seven languages are reflected only in the abstracts for articles in *International Social Work*.

The IASSW is expanding knowledge of international social work through cosponsorship of the journal *International Social Work* (http://isw.sagepub.com) and publishing online magazine 'Social Dialogue' (http://www.social-dialogue.com).

In 2012, Facebook (http://www.facebook.com/IASSW.AIETS) and Twitter (@IASSW-AIETS) official accounts of the IASSW were opened.

IASSW History

The history of the IASSW was studied by Kendall (1978), Healy (2001, 2008), Kniephoff-Knebel and Seibel (2008) and others mainly within the framework of internationalization of social work. Researchers paid special attention to the periodization of the IASSW activities based on mutual respect and non-paternalistic principles and the IASSW leadership.

Seibel (2008) stressed that milestones reflected in the IASSW history were in particular the world economic crisis, the advent of fascism in 1930s, the aftermath of the World War II, the Cold War confrontation, the crisis of the welfare state, the collapse of the Soviet system, and the establishment of globalization. According to Kendall (1978) and Healy (2001, 2008), the history of the IASSW can be divided into certain stages: (1) the founding (1926–1928); (2) the early years: progress and the growing threat of depression and aggression; (3) the prewar years; (4) World War II: a crisis in leadership and the destruction of war; (5) the postwar years: recovery and opportunity; (6) the 1950s and 1960s: a time of expansion; (7) 1970s: first permanent secretariat and large scale project; (8) the 1980s: international politics and the IASSW; (9) the international politics and the IASSW – 1980s; (10) the Secretariat dissolved – 1990s; and (11) entering the twenty-first century. However, these stages might be grouped into longer periods for review of the key issues in the IASSW institutional history.

1920–1930s: The Formative Phase

The founding of the IASSW has its roots in the organizing committee established in 1926 to plan the First International Conference of Social Work in Paris (Healy, 2008). It initially comprised of 51 schools, mostly in Europe, and was known as the International Committee.

An international conference on social work education organized by Dr Rene Sand in Paris in 1928 and its section on social work education chaired by Dr Alice Salomon of Germany gave birth to the IASSW, as well as to the ICSW and to the predecessor organization of the IFSW. These major international social work organizations provided leadership in connecting international agencies and international development organizations to social work worldwide.

At the time of the First International Conference, 111 schools of social work had been identified as potential members. Of these, 46 from 10 countries agreed to join as founding members (Kendall, 1978). The minutes of this first meeting also indicate that there were representatives among the delegates from India, the Philippines, Africa, and Latin America. This was a significant indication of IASSW's commitment to be international, and impressive for its time.

The first statement of purpose showed the direction for the setting up of international exchange across schools of social work and was outlined by the delegates of the training session

at the first international conference. Alice Salomon was elected as President of the IASSW.

Religious affiliation was a source of early conflict for the organization and limited success in enrolling schools. Kendall (1978) summarizes correspondence from Salomon noting that the schools affiliated with the Catholic International Union for Social Service viewed the International Committee as too neutral on religion. On the other side of the issue, leaders from the United States indicated that they opposed any mention of religion in the organization's constitution. In her report to the International Committee meeting held in Frankfurt in 1932, Elizabeth Macadam speculated that other schools had not affiliated with the committee for financial reasons (Kniephoff-Knebel and Seibel, 2008).

Alice Salomon came from a Jewish family and was known for her commitments to pacifism, women's rights, and international connections (Healy, 2008). After 1933 the German Welfare Schools tried to dismiss from her position as president. Salomon reluctantly agreed in order to save the organization; however, the International Committee refused to follow these demands and reinstated her as President. Kuhlmann (2001) observed that it was evidence of the antiracist attitude in the IASSW that they refused to do so.

Salomon was exiled from Germany in 1937. She settled in the United States and while she retained the formal title of head of the International Committee, she was unable to provide strong leadership.

Between 1928 and 1936, exchanges between schools, comparative studies to examine similarities and differences in social work education in member countries, and collection of information on social work education were the major activities of the international committee (Kniephoff-Knebel and Seibel, 2008). The committee participated in the second World Conference of Social Work in Frankfurt in 1932 and the Third World Conference in London in 1936. The fourth conference planned for 1940 was canceled due to the Second World War.

It is notable that efforts to establish linkages with intergovernmental organizations began almost immediately after its formation. The committee established working relationships with the International Labor Organization (ILO) and with the Commission on Social Questions of the League of Nations. A particularly important achievement was the creation of a center of documentation of schools of social work in 1929 in collaboration with the ILO, the precursor to several future world surveys and censuses of social work education (Kendall, 1978). Kendall's work at this time also led to IASSW's involvement in the devising of the Universal Declaration of Human Rights and the granting of Consultative Status at the UN for IASSW in 1947.

1940–1950s: Crises and Restoration

Membership grew under Salomon's leadership, and by 1939, included 75 schools in 18 countries. During this same period, however, all the German schools withdrew from membership in protest over Salomon herself.

During the 1940s, after World War II started, international meetings and even international communications were almost impossible. Elinor Black of the United Kingdom functioned as a secretary during the war years and attempted to keep in touch with social work education to the extent possible (Healy, 2008).

According to Kendall (1978), 1945–1954 were the Restoration Years for the International Committee of Schools of Social Work. As the organization began to regroup, it immediately pursued affiliation with the newly launched UN, and in 1947 it was granted class B consultative status with the Economic and Social Council as an NGO, underscoring the importance of the mission of representing social work education in global deliberations.

By 1956, the organization had changed its name to the current one, the International Association of Schools of Social Work. The Board of Directors elected in 1956 included members from Australia, Guatemala, Japan, and India. As schools of social work developed in the newly independent countries in Asia and Africa, membership grew. In 1954, there were 217 member schools in 27 countries.

With the International Council on Social Welfare (ICSW), the IASSW launched the journal *International Social Work*. The first issue was published in 1958. IFSW joined the team of sponsors in mid-1959 (Healy, 2008).

1960–1990s: Years of Expansion

During the 1960s, the organization became a more truly international one in terms of membership and leadership. In 1968, IASSW membership included 350 schools in 46 countries (Kendall, 1978). The world conferences, too, moved to new locations. In 1962, the Congress was held in Brazil, the first IASSW Congress in Latin America. This reflected the new process of the support of social work educators in different countries. Drucker (2005: p. 6) stressed that for the nearly 30 years of the IASSW existence "International was very much American and European. It was not until the early 1960s that Western-educated social workers emerged to help create social work schools in the developing and newly independent countries."

By the late 1960s, IASSW Vice-Presidents were elected from the Philippines, Ethiopia, Panama, and Finland, further solidifying the association's global identity (Healy, 2008). The expansion in leadership of the IASSW was in step with an incredible spreading out of its educational services. Social work and social work education expanded fairly rapidly to the newly independent countries in Asia and Africa.

Another highlight of the 1970s was that the IASSW Congress was held in Africa for the first time. The 1974 conference took place in Nairobi, Kenya, and drew a larger than the usual number of delegates from Africa.

In the early 1970s, IASSW secured grant funding to carry out a major project on education for the social dimensions of family planning. The grant funded a full-time secretary-general and a secretariat. The project involved curriculum development and training seminars for faculty in 20 countries across Africa, Asia, Caribbean, and Latin America.

In 1980s, IASSW faced a serious problem of 'racism' against schools of social work in South Africa. The Nordic schools withdrew their membership as a reaction to the apartheid in South Africa. Later, after apartheid, they returned with the end of apartheid in large numbers (Healy, 2008).

Since 1983, the social work profession has sponsored Social Work Day at the UN. Initially hosted by the IFSW, IASSW became a cosponsor in 1989. While about 200 social workers participated in the early years, today the conference attracts as many as a 1000 social workers to the UN (Wetzel, 2012). The mission of Social Work Day at the UN is twofold: (1) to educate social workers and social work educators about what the UN is doing and how social work can participate, and (2) to educate the UN – its agencies, governments and missions – about the relevance of social work in achieving its shared goals. The two entities have similar human rights and social justice ethos and objectives (Nadkarni, 2013).

In 1984, to commemorate Eileen Younghusband (UK) who was an IASSW President from 1961 to 1968, the biennial Eileen Younghusband Memorial Lectures was inaugurated. A current social work educator is invited to speak at the IASSW Congress on a topic of relevance to international social work. The most recent person nominated for this honor was Professor Walter Lorenz from Italy, an established social work author and former President of the European Center for Community Education, who gave this lecture during the 2012 World Social Work Congress in Stockholm, Sweden.

In 1992, the Katherine A. Kendall (now Memorial) Award was introduced to honor the late IASSW Honorary President Dr Katherine A. Kendall (USA). The most recent recipient of this award was Professor Lena Dominelli from UK (2012).

In 1990s, IASSW faced a financial crisis due to the withdrawal of the subsidy and funds by the government of Austria to run the Vienna office as the costs continued to escalate (Healy, 2008). As IASSW was not able to keep salaried persons, it became dependent on the leadership of its Executive Committee, Board and volunteers to run the Association, emerging successfully from the crisis in 1998 (Dominelli, 2004).

2000–2010s: Internationalization of Social Work and Education

The challenges to the profession in the twenty-first century are enormous, including the impact of globalization on microlevel practice; the frequency of natural disasters; increasing armed conflicts throughout the globe; rising numbers of children and older people in poverty; social injustice; the sex trade in women and children; and organized crime. All pose fundamental issues for international organizations embedded in civil society, including IASSW, to address (Tasse, 2008). Social work education has embraced a global perspective on many fronts: curriculum transformation, study abroad, student and faculty exchange, and collaborative overseas research, engagement with the latest global issues including climate change, among others.

As the twenty-first century proceeds, the reality of global interdependency is widely appreciated (Healy, 2001). IASSW partnered with the IFSW on development and approval of four important policy statements for the profession of social work in the era of globalization: the International Definition of Social Work; the Ethical Principles; the *Global Standards for Social Work Education and Training*; and the Global Agenda for Social Work and Social Development.

The IASSW and the IFSW had worked together to formulate a Definition Statement (see http://www.iassw-aiets.org/international-definition-of-social-work) officially approved by both at their respective General Assemblies in 2000 and by both organizations jointly in 2001 in Copenhagen. It runs as follows: "The social work profession promotes social change, problem solving in human relationships and the empowerment and liberation of people to enhance well-being. Utilising theories of human behavior and social systems, social work intervenes at the points where people interact with their environments. Principles of human rights and social justice are fundamental to social work" (Hare, 2004: 418).

The new international definition of the social work profession replaced the IFSW definition adopted in 1982. It was based on the understanding that social work in the twenty-first century is dynamic and evolving, and therefore no definition should be regarded as exhaustive. Lorenz (2001) considered the ambiguities, tensions and contradictions of the social work profession, which have to be constantly negotiated and renegotiated, rather than resolved, to constitute its success and challenge. The definition and comments to it provide the parameters of broad ethical principles that cannot be refuted on an ideological level.

In 2004, the IASSW and the IFSW formally approved at the General Meetings in Adelaide (Australia) the document *Ethics in Social Work, Statement of Principles*. This document cites the definition and lists seven human rights treaties as "particularly relevant to social work practice and action." The purpose of the documents of IASSW and IFSW on ethics is "to promote ethical debate and reflection in the member organizations, among the providers of social work in member countries, as well as in the schools of social work and among social work students" (Banks et al., 2008).

A major initiative in the early twenty-first century was the development and publication of "Global Standards for Social Work Education and Training." This document was presented to the General Meetings of the IASSW and the IFSW at a joint Congress in Adelaide (October 2004) and was accepted by both organizations with implications for the transferability and portability of qualifications. This document consists of nine sets of standards: school's core purpose or mission statement; program objectives and outcomes; program curricula including fieldwork; core curricula; professional staff; student social workers; structure, administration, governance and resources; cultural diversity; and social work values and ethics (Sewpaul and Jones, 2004).

These standards are the first worldwide guidelines for enhancing social work education. At the same time neither the IASSW nor the IFSW are playing any monitoring, control, or accreditation function at the level of the any given school.

In 2012, the *Global Agenda for Social Work and Social Development: Commitment to Action* was released. This document was designed by the IASSW, the IFSW, and the ICSW to strengthen the profile of social work and to enable social workers to make a stronger contribution to policy development. The paradigm of the Global Agenda for Social Work asserts that the context in which social work is practiced is deeply influenced by globalization, particularly the political form of globalization which has seen a spread of neoliberal policies through the introduction of market forces (Jones and Truell, 2012). The Global Agenda commits the three organizations to addressing the root causes of oppression and inequality. They are dedicated to

leading a movement that will create a more socially just and equitable world.

Conclusions

IASSW's support for academic scholarship and research have borne fruit in a number of areas, including in the establishment of schools of social work where none existed before, engagement in current topical global issues, e.g., climate change, through advocacy work; and supporting the development of newly emerging fields in the profession.

See also: International Social Work; Social Work Education; Social Work Theory; United Nations and its Agencies.

Bibliography

Banks, S., Hugman, R., Healy, L., Bozalek, V., Orme, J., 2008. Global ethics for social work: problems and possibilities. Ethics and Social Welfare Journal 2 (3), 276–290.

CSWE, 2013. Katherine A. Kendall (1910–2010). http://www.cswe.org/CentersInitiatives/KAKI/AboutKAKI/KatherineKendall.aspx (accessed 10.04.13.).

Dominelli, L., 2004. Social Work: Theory and Practice for a Changing Profession. Polity Press, Cambridge.

Drucker, D., 2005. The International Association of Schools of Social Work (IASSW) and the "Pursuit of Social Justice and Social Development". Paper for International Seminar on Social Work Education in Ethiopia. http://www.aboutsweep.org/IASSW_Ethiopia_REFREV080405.pdf (accessed 26.08.13.).

Hare, I., 2004. Defining social work for the twenty first century. The International Federation of Social Workers' revised definition of social work. International Social Work 47 (3), 407–424.

Healy, L.M., 2001. International Social Work: Professional Action in an Interdependent World. Oxford University Press, New York.

Healy, L.M., 2008. Introduction: a brief journey through the 80 year history of the International Association of Schools of Social Work. Social Work and Society 6. http://www.socwork.net/sws/article/view/98 (accessed 10.04.13.).

IASSW, 2013. About IASSW. http://www.iassw-aiets.org/about-iassw (accessed 10.04.13.).

Jones, D., Truell, R., 2012. The global agenda for social work and social development: a place to link together and be effective in a globalized world. International Social Work 55 (4), 454–472.

Kendall, K.A., 1978. The IASSW 1928–1978: a journey of remembrance. In: Kendall, K. (Ed.), Reflections on Social Work Education 1950–1978. IASSW, New York, pp. 170–191.

Kniephoff-Knebel, A., Seibel, W.F., 2008. Establishing international cooperation in social work education: the first decade of the International Committee of Schools of Social Work (CSSW). International Social Work 51 (6), 790–812.

Kuhlmann, C., 2001. Alice Salomon (1872–1948) – Germany. European Journal of Social Work 4 (1), 65–75.

Lorenz, W., 2001. Social work in Europe – portrait of a diverse professional group. In: Hessle, S. (Ed.), International Standard Setting of Higher Social Work Education. Stockholm University, Stockholm Studies of Social Work.

Nadkarni, V.V., 2013. IASSW and social work day at the United Nations: a proud achievement. International Social Work 56, 252–256.

Seibel, W.F. (Ed.), 2008. Global Leaders for Social Work Education: The IASSW Presidents 1928–2008, pp. 160–161.

Sewpaul, V., Jones, D., 2004. Global Standards for Social Work Education and Training. http://www.iassw-aiets.org/global-standards-for-social-work-education-and-training (accessed 29.08.14.).

Tasse, A., 2008. International Association of Schools of Social Work (IASSW). In: Encyclopedia of Social Work, twentieth ed., vol. 2. NASW Press and Oxford University Press, pp. 472–473.

Wetzel, J.W., 2012. IASSW and the United Nations. IASSW-AIETS Newsletter. Spring Issue, 6.

Relevant Websites

www.globalsocialagenda.org – The Global Agenda For Social Work and Social Development.

http://www.facebook.com/IASSW.AIETS – IASSW on Facebook.

http://www.iassw-aiets.org/about-iassw – IASSW Web-site.

http://www.social-dialogue.com – Social Dialogue (IASSW Online Magazine).

https://twitter.com/IASSW_AIETS – Twitter.

International Business

Mauro F Guillén, The Wharton School of the University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

International business has to do with trade and direct investment across borders. Research in this area has been undertaken from economic, managerial, and sociopolitical perspectives. Globalization has introduced a new dynamic into the field, as has the rise of the emerging economies and of the emerging market multinationals.

The term 'international business' refers to economic activities that take place across the boundaries of nation-states. The most important of these are trade and foreign investment. Merchandise *trade* has grown 12-fold since the end of World War II, and trade in services has increased even faster since the early 1980s. Foreign *direct* investment (as opposed to portfolio investment not seeking the managerial control of the foreign asset) has grown at a rate twice as fast as that of trade. *Foreign direct investment* gives rise to multinational enterprises, i.e., firms with operations in more than one country (*see* Multinational Corporations). There are nearly 104 000 of these enterprises in the world today, which collectively control about 882 000 affiliates. They account for one-third of world merchandise trade and for the vast majority of trade in technology (UNCTD, 2011). While theories of international trade were first put forward during the sixteenth century and have largely remained the province of economics, the study of foreign direct investment and the multinational enterprise started only after World War II and has been approached from a variety of social and behavioral perspectives. The three most important ones – economic, managerial, and sociopolitical – are covered here (Table 1).

Modern economic thinking about foreign direct investment and the multinational enterprise properly starts with Stephen Hymer's (1960) MIT thesis, posthumously published in 1976. The key insight was to see foreign direct investment not as a capital movement but as the decision of a firm to produce some goods or service in a foreign location. Hymer noted that foreign direct investors did not seem to go abroad in search for higher interest rates – as the theory of investment would predict. Rather, he argued that "control of the foreign enterprise is desired in order to remove competition between that foreign enterprise and enterprises in other countries" (Hymer, 1976: p. 25). Thus, Hymer was the first to observe the *oligopolistic* behavior of the firm that engages in foreign direct investment. In other words, the multinational was seen as an anomaly that could only be explained by reference to imperfections in the markets for production factors and goods, economies of scale, and/or government intervention. The most complete rendering of the oligopolistic thesis is that of Knickerbocker (1973). He observed that firms in a loose-knit oligopoly in the home country match each other's investments in foreign locations following a pattern of action–reaction or move–countermove out of fear that not keeping up with rivals might erode their profitability and competitive position or endanger their sources of supply.

Hymer was also the first scholar to note that "the control [of a foreign enterprise] is desired in order to appropriate fully the returns on certain skills and abilities" (1976: p. 25), especially those having to do with product and process innovation.

Table 1 Three perspectives on international business and the multinational enterprise (MNE)

	Economic	*Managerial*	*Sociopolitical*
Name of the literature(s)	IB (International Business)	IM (International Management)	Development Studies and International Relations
Research community	AIB (Academy of International Business) and AEA (American Economic Association)	AIB and AoM (Academy of Management)	ASA (American Sociological Association) and APSA (American Political Science Association)
Research questions	Why do firms go abroad? How do they go abroad? What are the welfare implications?	Why and how do firms go abroad? How are MNEs organized?	What role do MNEs play in economic and political developments, and in international relations?
Assumptions and answers	Rationality; constrained maximization; calculation of trade-offs	Capability-based strategies; organizational structure follows strategy	Power and interest drive processes and outcomes
Ways of doing research	Deductive reasoning; formal modeling; large samples	Conceptual framing; grounded theory; case studies and large samples	Mostly inductive; grounded theory; case studies; comparative-historical method
Key scholars	Hymer (1976); Caves (1996); Knickerbocker (1973); Vernon (1966); Buckley and Casson (1976); Dunning (1979); Kogut (1983)	Perlmutter (1969); Stopford and Wells (1972); Wells (1983); Prahalad and Doz (1987); Bartlett and Ghoshal (1989)	Vernon (1971); Moran (1974); Kobrin (1982); Gilpin (1987); Haggard (1990)

 http://dx.doi.org/10.1016/B978-0-08-097086-8.73003-0

Vernon (1966) further argued that patterns of trade and foreign investment had to do with the so-called 'product cycle' and not so much with comparative advantage. Vernon saw entrepreneurs and managers as relatively myopic, i.e., most knowledgeable about, and responsive to, conditions in the home market. The features of new products reflected characteristics of the home market and affected the pattern of foreign expansion of the firm. Vernon proposed that firms would only pursue foreign market opportunities after the home market became saturated. Foreign markets relatively similar to the home market in terms of purchasing power and tastes would be entered first, and exports would probably be the preferred mode of entry. Licensing or foreign production would come later in response to host country protectionism and/or rising production costs in the home country. Vernon's insights were further developed by Swedish scholars (Johanson and Vahlne, 1977).

The economic approach to the multinational enterprise acquired a high degree of internal consistency after scholars started to draw from the insights of transaction cost economics (*see* Transaction Costs and Property Rights), which has become the backbone of the widely accepted view among economists that multi-nationals are an efficient governance response to certain market failures (Buckley and Casson, 1976; Dunning, 1979; Caves, 1996). The multinational enterprise is expected to exist if two conditions are met. First, the necessary condition is that there be differences in the productivity and cost of production factors by location (in the case of vertical expansion by multinationals) or significant transportation costs, trade protectionism, adverse changes in exchange rates, and/or needs to customize products to local demand (in the case of horizontal expansion). Transaction cost theory, however, warns that a second, sufficient condition must be met in order to guarantee that foreign production will be managerially controlled by a multinational firm: market failure due to asset specificity or uncertainty (vertical expansion) or the possession of intangible assets such as patents, brands, and other proprietary skills that cannot be protected in a contractual, arm's length fashion (horizontal expansion). A final, distinctive contribution to the economic approach was made by Kogut (1983), who proposed to see the multinational enterprise as a flexible network in which resources are arbitraged across borders, a view that revisited Vernon's insight that multi-nationals approach foreign investment decisions as a sequential process. A significant body of empirical evidence exists corroborating the major tenets of the economic approach (Caves, 1996).

While the economic approach has focused on the questions of why and how firms go abroad, the managerial approach has contributed an understanding of how multinationals are organized and managed. This second approach relies on typological conceptual frameworks and inductive reasoning rather than formal modeling. In addition, theoretical development frequently results from case studies, although testing of the main propositions has also been undertaken with large samples, as in the economic approach. Perhaps the first noteworthy typology of the multinational enterprise in the managerial literature was proposed by Perlmutter (1969), who distinguished among three types of multinational 'mindsets.' Ethnocentric multinationals are the ones assumed by Vernon's (1966) produce-cycle approach. They have simple but inflexible organizational structures. Polycentric multinationals, by contrast, adapt to local circumstances but have difficulty in coordinating. Finally, geocentric multinationals are costly but very good at gathering and integrating information from various locations. Drawing on a rich data set on US multinationals assembled under the coordination of Raymond Vernon, Stopford and Wells (1972) presented the first comprehensive treatment of the various factors at the firm, industry, and country levels affecting the choice of organizational structure and degree of ownership over the foreign subsidiaries of the multinational firm. Wells (1983) extended this kind of analysis to the case of 'Third World' multinationals, i.e., firms from developing countries that engaged in foreign direct investment.

The early insights by Perlmutter, and Stopford and Wells were further developed in their respective dissertations by Prahalad and Doz as the 'integration-responsiveness' framework (Prahalad and Doz, 1987). Multinationals are seen as organizations responding to environmental pressures for global integration (so as to reduce costs through economies of scale and scope) and to pressures for local responsiveness (so as to adapt to local tastes, institutions, or regulations). Following the strategy-structure paradigm in organizational theory, they identify different organizational structures and managerial processes that allow multinationals to operate under different configurations of global integration and local responsiveness. Bartlett and Ghoshal (1989) proposed an influential analytical framework also emphasizing the tension between global and local opportunities or demands that every multinational has to deal with. Although still in its infancy, a neoinstitutional approach to the multinational enterprise as a firm exposed to isomorphic pulls from both the home and the host country is also emphasizing the organizational problems associated with multinationality (Westney, 1993; see Institutional Theory).

A third perspective on international business has built on the insights of both the economic and managerial approaches to the study of the political and social context of multinational activity. Vernon (1971) was the first to explicitly argue that the spread of multinational corporations creates 'destructive political tensions,' and that there is a 'need to reestablish balance' between political and economic institutions.

Political scientists have looked at the impact of multinationals on international relations. For example, neorealist scholar Robert Gilpin (1987) argues that changes in the international location of activities due to foreign direct investment have reinforced the importance of domestic policies, as countries engaged in regionalization, sectoral protectionism, and mercantilistic competition during the 1980s so as to cope with globalizing trends. Other political scientists have been concerned with the relative bargaining positions of host governments and multinationals. In a classic book, Moran (1974) analyzes Chilean efforts to seize control over its vast copper industry, finding that nominal ownership and control does not guarantee success in a worldwide oligopoly. Haggard (1990) summarizes the bargaining power literature writing that the host government is relatively powerful vis-à-vis multinationals making huge fixed investments (e.g., in extractive industries), relatively

weak when dealing with export-oriented firms looking for cheap labor, and weakest when multinationals are attracted by import-substitution policies that create a political base of support among local suppliers, distributors, employees, and joint-venture partners. Kobrin (1982) offers a different, yet complementary, approach to the politics of multinational activity by examining how multinationals evaluate political risk and try to protect themselves against it.

Sociologists have also examined the effects of international business activity on economic growth and inequality in the world. One major controversy in the literature is whether foreign investment retards long-term economic growth and exacerbates income inequality in the host country or not. Capital dependency theorists have long argued in the affirmative using cross-national empirical evidence (Dixon and Boswell, 1996), while other sociologists pointed out that the data do not support such claims (Firebaugh, 1996). This protracted debate – now three decades old – does not take into account the fact that the effect of foreign investment on host-country growth and income inequality is mediated by the economic development strategy of the host government. Countries allow more or less foreign investment depending on a number of ideological, political, and economic considerations, and the effects of multinational activity differ widely between import-substitution and export-oriented contexts (Haggard, 1990). While it is easy to point to examples of the harmful effects of multinationals, and even of their outright wrongdoing, the social-science literature has been slow to recognize that they have played a largely constructive role in many countries – e.g., Singapore, Malaysia, Puerto Rico, Ireland, Portugal, and Spain – in which even organized labor and the left have welcomed them.

Over the last 20 years, the expansion of international business activity has acquired a new, qualitatively different outlook. *Globalization* at the turn of the twenty-first century is different from the internationalization of the world economy during the nineteenth century through increasing cross-border flows of trade and investment. The 'global' economy of the late twentieth century and into the twenty-first century is driven by the increasing scale of technology, the surge in cross-border collaboration between firms along the value-added chain, and the cross-border integration of information flows. While the conventional wisdom is that late-twentieth-century globalization undermines the nation-state and compels firms to converge toward 'universal best practices,' sociologists and political scientists document that governments exercise a great deal of choice in the global economy and firms continue to adopt organizational structures and patterns of behavior that are related to the institutional conditions in their home countries (Orrù et al., 1997). Thus, globalization is far from being a monolithic, uniform, and inexorable trend. Rather, it is a fragmented, contradictory, and heterogeneous process of change. More empirical and theoretical research is needed to describe the causes of globalization – including those having to do with international business activity – and to understand its effects on political, economic, and social outcomes in various parts of the world.

Perhaps the most important new development in the world of international business is the rise of the so-called emerging market multinationals, companies from emerging and developing economies that have made important investments abroad. In 2011, 29% of the 104 000 multinational firms in the world originate from emerging and developing countries, up from 21% in the year 2000. They account for about 15% of global stocks of cumulative foreign direct investment, and between 2006 and 2010 they were responsible for 41% of new foreign direct investment flows. What's more important, 120 emerging market multinationals are among the top 1000 corporate spenders on research and development, indicating that these firms are competitive in terms of low labor cost. In fact, many of them have come to dominate entire industries, like Bimbo of Mexico (bread), Embraer of Brazil (regional jets), BYD of China (rechargeable batteries), or Tata Communications of India (wholesale voice traffic). The emerging market multinationals already account for 25% of the Fortune Global 500 ranking of the largest companies in the world in terms of revenue (Guillén and Ontiveros, 2012).

The rise of the emerging market multinationals poses interesting questions about the nature and future evolution of international competition (Guillén and García-Canal, 2009). They have expanded around the world much faster on average than the multinational corporations from Europe or the United States, based on less obvious and relatively incipient competitive advantages. Many of them lacked a strong brand image before expanding abroad. They are also more likely to leverage political skills, frequently developed in the home market. They tend to use alliances, joint ventures, and mergers and acquisitions to a much greater extent than wholly owned subsidiaries and greenfield investments. Finally, most of the emerging market multinationals are companies owned and controlled by the state or families. These peculiar patterns of ownership and control raise numerous issues and have been the source of heated controversies given that governments in Europe and the United States have so far been willing to prevent some emerging market multinationals from controlling strategic assets in their economies.

As the twenty-first century roars on, the world of international business continues to undergo a profound transformation driven by the impact of information technology, social media, the growth of the emerging economies, and the fallout from the 2008 financial crisis. Research is likely to focus on the characteristics, relationships, and impact of these new actors and types of firms, and on the implications of their proliferation and growth for global competition.

See also: Multinational Corporations.

Bibliography

Bartlett, Christopher A., Ghoshal, Sumantra, 1989. Managing across Borders: The Transnational Solution. Harvard Business School Press, Boston, MA.

Buckley, Peter, Casson, Mark, 1976. The Theory of the Multinational Corporation. MacMillan, London.

Caves, Richard E., 1996. Multinational Enterprise and Economic Analysis. Cambridge University Press, New York.

Dixon, W.J., Boswell, T., September 1996. Dependency, disarticulation, and denominator effects: another look at foreign capital penetration. American Journal of Sociology 102 (2), 543–562.

Dunning, John, 1979. Explaining changing patterns of international production: in defence of the eclectic theory. Oxford Bulletin of Economics and Statistics 41, 269–295.

Firebaugh, Glenn, September 1996. Does foreign capital harm poor nations? American Journal of Sociology 102 (2), 563–575.

Gilpin, Robert, 1987. The Political Economy of International Relations. Princeton University Press, Princeton, NJ.
Guillén, Mauro F., García-Canal, Esteban, May 2009. The American model of the multinational firm and the 'new' multinationals from emerging economies. Academy of Management Perspectives 23 (2), 23–35.
Guillén, Mauro F., Ontiveros, Emilio, 2012. Global Turning Points: Understanding the Challenges for Business in the Twenty-First Century. Cambridge University Press, New York.
Haggard, Stephan, 1990. Pathways from the Periphery: The Politics of Growth in the Newly Industrializing Countries. Cornell University Press, Ithaca, NY.
Hymer, Stephen, 1976 [1960]. The International Operations of National Firms: A Study of Direct Foreign Investment. The MIT Press, Cambridge, MA.
Johanson, Jan, Vahlne, Jan-Erik, 1977. The internationalization process of the firm: a model of knowledge development and increasing foreign market commitments. Journal of International Business Studies 8 (1), 23–32.
Knickerbocker, Frederick, 1973. Oligopolistic Reaction and Multinational Enterprise. Division of Research, Harvard Business School, Boston, MA.
Kobrin, Stephen J., 1982. Managing Political Risk Assessment: Strategic Response to Environmental Change. University of California Press, Berkeley, CA.
Kogut, Bruce, 1983. Foreign direct investment as a sequential process. In: Kindleberger, C.P., Audretsch, D. (Eds.), The Multinational Corporation in the 1980s. MIT Press, Cambridge, MA.
Moran, Theodore H., 1974. Multinational Corporations and the Politics of Dependence: Copper in Chile. Princeton University Press, Princeton, NJ.
Orrù, M., Biggart, N.W., Hamilton, G.G., 1997. The Economic Organization of East Asian Capitalism. Sage, Thousand Oaks, CA.
Perlmutter, Howard, January–February 1969. The tortuous evolution of the multinational corporation. Columbia Journal of World Business, 9–18.
Prahalad, C.K., Doz, Yves L., 1987. The Multinational Mission: Balancing Local Demands and Global Vision. Free Press, New York.
Stopford, John M., Wells Jr., Louis T., 1972. Managing the Multinational Enterprise: Organization of the Firm and Ownership of the Subsidiaries. Basic Books, New York.
UNCTD (United Nations Conference on Trade and Development), 2011. World Investment Report 2011. United Nations, New York.
Vernon, Raymond, 1966. International investment and international trade in the product life cycle. Quarterly Journal of Economics 80, 190–207.
Vernon, Raymond, 1971. Sovereignty at Bay: The Multinational Spread of U.S. Enterprises. Basic Books, New York.
Wells Jr., Louis T., 1983. Third World Multinationals: The Rise of Foreign Investment from Developing Countries. The MIT Press, Cambridge, MA.
Westney, D. Eleanor, 1993. Institutionalization theory and the multinational corporation. In: Ghoshal, Sumantra, Westney, Eleanor (Eds.), Organization Theory and the Multinational Corporation. St. Martin's Press, New York, pp. 53–76.

International Business: Geographic Aspects

Sharmistha Bagchi-Sen and Torsten Schunder, State University of New York – Buffalo, Buffalo, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides an overview of the evolution of understanding multinational enterprise (MNE) activities. Between World War II and the late 1990s, there was a major focus on understanding foreign direct investment (FDI) by MNEs. In the past 20 years, recognition of commodity chains, value chains, and production networks at the global scale has prompted researchers to develop new frameworks to incorporate the complexity of MNE operations spanning many countries and industries. This article explicitly recognizes geography as a key factor (e.g., location determinants, local impact, role of policy).

Introduction

A major aspect of international business is the worldwide spread of MNEs, also known as multinational corporations (MNCs) and transnational corporations (TNCs), whose operations are organized across borders. The United States, Europe, and Japan still dominate as the main sources of MNEs. However, South Korea, Taiwan, Singapore, China (including Hong Kong), India, Brazil, and Russia are some of the countries with emerging MNEs in the twenty-first century. Foreign direct investment, an active investment in a foreign country made by either acquiring a partner in the host country or establishing a new operation, is the main mode of entry by MNEs into other countries. The theoretical foundation for studying FDI includes theories focusing on market imperfections, that is, the possession of certain advantages (e.g., knowhow, finance) by a few firms, which give them competitive advantage over others in an international market.

Historically, asset exploitation (e.g., direct sales of goods and services) has been the focus of MNEs. More recently, asset augmentation has become an integral part, that is, MNEs are now interested in gathering market intelligence to tailor their products for particular niche markets. Furthermore, firms have global networks of suppliers and distributors, as well as cooperative arrangements, like strategic alliances and joint ventures, to pool resources and reduce risks, to manage worldwide research and development (R&D), production, and sales. This article reviews the evolution of research on FDI, with a focus on the importance of the geographic location of nations – as facilitators (e.g., availability of cheap labor in the host country); as negotiators (e.g., the designation of export processing zones by the host country), and as recipients of the impact of FDI (e.g., technology transfer).

Explanations of Foreign Direct Investment

Before Hymer's (1976) groundbreaking dissertation in 1960, scholars assumed that firms invested internationally to profit from real interest rate differentials; FDI was seen as a special kind of portfolio investment with additional control functions (Hennart, 2010). Hymer challenged these assumptions by hypothesizing that firms undertake direct foreign investment to further their competitive advantages at home and abroad. By internalizing pecuniary externalities caused by market imperfections (e.g., barriers of entry, oligopolistic industry structure, government interventions), firms could achieve monopolistic profits in foreign markets based on an internal competitive advantage (Buckley and Casson, 1976; Hennart, 2010; Caves, 2007; Kedron and Bagchi-Sen, 2012).

Hymer's (1976) approach saw the motivation for FDI in the effort to preserve a position within an oligopolistic system by preventing the emergence of foreign competition in the context of reduced trade barriers; this view sees MNEs as exploitative, with governments expected to regulate their activities so that local industry and entrepreneurs are not squeezed out of the market (Hennart, 2010). Hymer's work is the foundation for the internalization theory, which explains the incorporation of processes in an organization allowing for high degrees of control and remains an important part in today's approaches to explaining FDI (Dunning, 2003). Internalization theory further explains why firms pursue FDI in lieu of exports or licensing; internalizing transaction costs in the international market to manage uncertainties is a major motivation in FDI decisions.

Dunning (1977) was the first to introduce location as an important determinant of FDI in his eclectic paradigm. This paradigm addresses the why, how, and where of FDI activities (Dunning, 2009a,b): a firm will undertake FDI if a combination of firm-specific advantages – ownership advantage (O), location-specific advantages of the host country (L), and internalization (I) of operations – create an overall benefit for the company. A company enters a foreign market via FDI if it possesses a specific competitive advantage (e.g., technology, skills), if the internalization of operations provides benefits, and if the host country provides a locational advantage (e.g., lower labor cost, resources). In the absence of internalization or locational advantages, a company would opt to export its goods/services; and licensing would be the preferred mode of entry if a company has only one particular type of ownership-specific advantage (e.g., patents) (Dunning, 1977).

FDI enables a company to maintain and protect its initial competitive advantage over competitors in a host country by internalizing operations, establishing hierarchies (e.g., vertical integration), and enhancing the security of transactions while controlling the diffusion of knowledge. The ability to internalize and thereby obtain better control of

market factors – e.g., suppliers or retailers – is another advantage. But internal markets are suitable only for specific industries or company structures whereas others fare better with external markets (e.g., outsourcing), which can eliminate inefficient parts of a production process (Hennart, 2010; Buckley and Ghauri, 2004; Buckley, 2009).

The ownership-location-internalization (OLI) paradigm has incorporated new developments as Dunning (2009b) elaborated. Generally, in the 1970 and 1980s, firms made locational choices to focus on asset exploitation (e.g., access to markets such as foreign direct investment in the United States). In recent years, asset augmentation (e.g., access to innovative ideas or codified knowledge such as FDI in Europe by Chinese and Indian firms) has been an important aspect of FDI, which includes investments by emerging market MNEs (EMMNEs) in developed economies (Ramamurti, 2010). Previously, locational determinants did not explicitly include external economies or the importance of clusters (e.g., core regions, global cities), which provide learning networks, proximity to specialized labor, and other intangible benefits not addressed by transaction cost theory. Also the importance of institutions at various geographic scales and new forms of international cooperation, ranging from a diversity of new equity-based cooperation to nonequity-based cooperation, have received increasing attention within the last decade.

In the context of emerging forms of cooperation in international business, other organizational frameworks like the global commodity chain (GCC), global value chain (GVC), and global production network (GPN) emerged. The GCC is based on networks related to a specific commodity linking households, enterprises, and states and is strongly characterized by buyer–supplier dependencies (Buckley and Ghauri, 2004). A GVC includes all value-added activities related to a product beginning with R&D, production, marketing, and distribution. Gereffi et al. (2005) identify three modes of governance (modular, relational, captive) between market relations and hierarchies, which differ in the degree of power exerted by a lead firm on its suppliers/customers without direct ownership (**Figure 1**). The GPN framework addresses the multiactor and multiscalar characteristics of transnational production systems based on power, value, and embeddedness; the GPN concept is related to the GVC approach (Coe et al., 2008).

Three main modes of governance operate within these frameworks: (1) traditional FDI is used to organize and coordinate crucial functions; (2) arm's-length transactions are buyer–seller relationships subject to market forces; and (3) nonequity modes (NEMs) are characterized by various degrees of buyer–seller dependencies (UNCTAD, 2013). Though other forms of governance exist, FDI remains an important tool to reduce uncertainty by internalizing crucial aspects of the international business process.

Because MNEs are main actors in international business, their strategic orientation and decisions are essential. Their modes of governance and internalization decisions aim to reduce economic risk by increasing reliability and control over the production process (Yeung, 1999; Guillén and García-Canal, 2009). *Vertical integration* strategies aim to internalize backward- or forward-oriented parts of the value chains, which can include resource procurement or the securing of distribution channels. *Horizontal integration* establishes operations in a new market while maintaining production in other

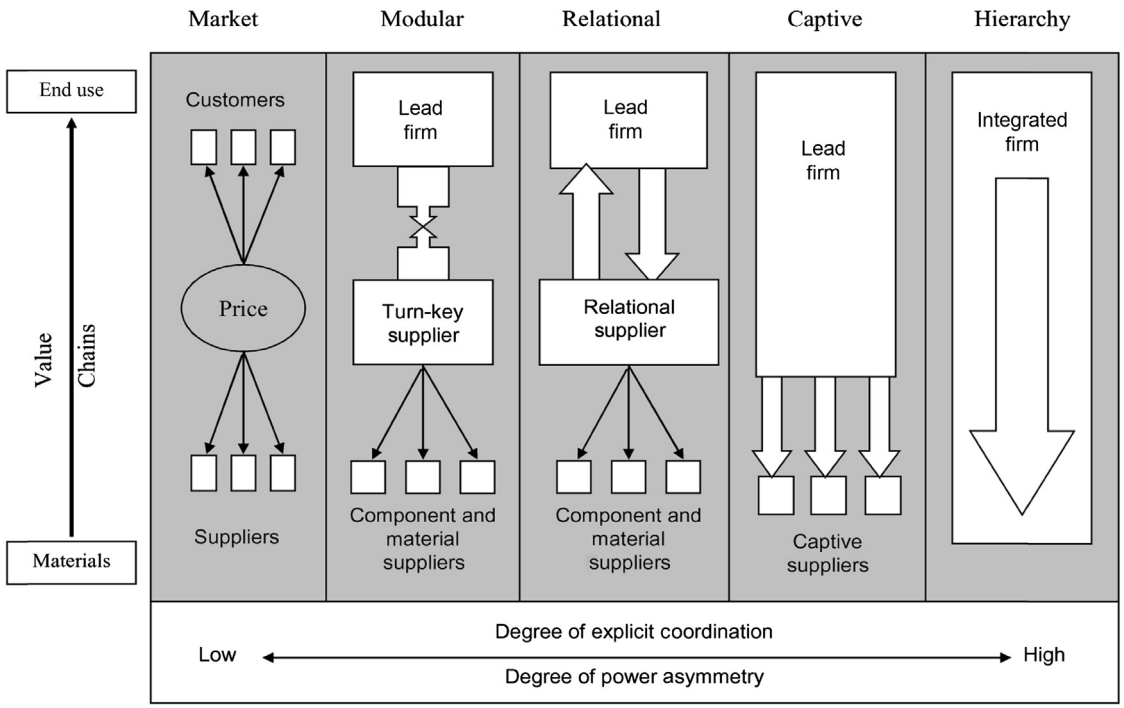

Figure 1 Five global value chain governance types. Reproduced from Gereffi, G., Humphrey, J., Sturgeon, T., 2005. Governance of global value chains. Review of International Political Economy 12, 78–104.

places. *Diversification strategies* aim to increase the product portfolio and thereby reduce a company's dependence on just one product group. *Risk reduction behavior* determines how a company enters a market (Guillén and García-Canal, 2009); for example, Chinese MNEs use joint ventures to gather market intelligence before committing to a large FDI (Yeung, 1999). Related literature discusses strategies of first movers, imitators, latecomers, and others in the international market.

Locational Determinants of Global Value Chains and Foreign Direct Investment

The recent World Investment Report (UNCTAD, 2013) provides an extensive discussion of global value chains and their locational determinants, a discussion that is also valid for understanding FDI within the GVC. In general, during all stages of the value chain, important factors are political, economic, and social stability; a sufficient institutional framework (e.g., policies, accounting and finance regulations, intellectual property rights); labor force suitability; geographic access to markets or the next value chain activity; transportation and logistics infrastructure; and the presence of local firms. All these factors lead FDI in nonextractive sectors to locate in urban or core regions.

If the globalization of R&D activities is considered, important locational factors are the national innovation system, the supply of educated labor force in science and technology, and the presence of institutions facilitating research and the translation of research for commercialization. Also important is the policy environment, including R&D policies, intellectual property laws as well as general laws and regulations governing contract research and licensing, the presence of science and technology parks, and investment incentives. Other aspects important for R&D are local consumer preference, local design capabilities, policies and institutional support to learn from local consumers (e.g., the presence of nongovernmental organizations as mediators between the multinational and the market). Relevant factors for the production aspect of a value chain are cost and quality of input materials and access to intermediate goods. Locational determinants for this stage include resources for processing, the availability of utilities, low-cost and trained labor, as well as sufficient transportation and communication infrastructure. The distribution stage requires access to a network of distribution and logistics companies.

Clearly, location determinants range from national to regional/local characteristics. With reference to operational aspects, Buckley and Ghauri (2004) note that MNE strategies are often focused on balancing the opposing pressures of global and local, that is, accommodating geographic variations in every aspect affecting business within a globalized world. Global operations provide cost advantages based on the potential of realizing economies of scale and increasing efficiency, whereas the local level requires specialization, adaptation, and product differentiation, which in turn provides a company with niche markets and high profit margins.

New locational determinants have gained importance in the context of emerging market economies, which have large markets, with large shares of low-income consumers who are highly price sensitive. People living in environments of scarcity improvise with available resources to maximize usage value, which can lead to surprisingly efficient results and an overall culture of creative makeshift innovation (e.g., 'jugaad' or frugal innovation in the Indian context). Such frugal innovation or grassroots innovation has the potential to increase overall resource efficiency and open new market opportunities at the bottom (or base) of the pyramid (BoP) if companies are able to access knowledge of the BoP and adapt to the local context (Bound and Thornton, 2012). Also within its own network, a company might be able to create competitive advantage if it is able to incorporate suppliers, customers, and global clusters of businesses cooperating with each other (Mathews, 2006). Access to such networks is critical for host country companies that want to become part of a global value chain. Furthermore, some companies originating in an emerging country have accumulated enough revenue from their own markets to venture into FDI for asset augmentation (e.g., Indian pharmaceuticals in Europe).

While manufacturing has been the traditional focus of FDI studies, more recently the R&D activities of specific industries and sector-specific analysis (e.g., information technology and other services) have attracted scholarly attention (Coughlin and Segev, 2000; Coe and Hess, 2005; Dunning and Lundan, 2009). With the increasing division of labor and new modes of organizing the labor process that enable companies to optimize the location of specific steps in the value chain (Glasmeier, 1986; Buckley and Ghauri, 2004; UNCTAD, 2013), the locational determinants of international business activities (e.g., availability of cheap educated labor) have become highly context-specific. This context specificity is illustrated in numerous studies of FDI in particular destinations such as FDI in the United States, FDI in the UK, or FDI in China. For example, the volume of FDI in the United States attracted research over the past 30 years to understand state-level patterns of FDI within the United States (Bagchi-Sen and Wheeler, 1989). In the 1980s and 1990s, Japanese FDI was a popular choice of inquiry, with focus on the variety of destinations of such investments in the automobile industry and the creation of auto alley in the United States (Poon et al., 2013). Now, the volume of FDI in emerging economies like China and India has spurred research on certain sectors and regions, such as FDI in information technology in Bangalore, India.

By using a taxonomy of market-seeking, resource-seeking, efficiency-seeking, and strategic asset-seeking FDI, work on locational determinants has mostly obscured the role of geography and place. Empirical analyses that have included geographical factors have shown how locational determinants affect business decisions at various scales ranging from the local to the national level and even global regions (Bagchi-Sen and Wheeler, 1989; Coughlin and Segev, 2000; Dunning, 2009a). Researchers have examined variations in the location determinants of FDI based on mode of entry (e.g., acquisition or greenfield), industry sector, value chain operation (e.g., type of production), structure of the multinational enterprise (e.g., centralized or decentralized governance), strategies of the MNE, and environmental factors – e.g., the 'policy environment,' such as the development of export processing zones, the level of

corruption, IP regime (see Dicken, 2011). Now, the complexity of global networks necessitates examination of specific aspects of value chain creation and their impact and a burgeoning field of study now focuses on the impact of GVC and FDI on local economies, with a focus on development outcomes (Coe et al., 2008; Dicken, 2011; UNCTAD, 2013)

Emerging Market Multinational Enterprises

Japanese FDI as well as the more recent EMMNEs from the BRIC (Brazil, Russia, India, and China) countries have required new theorization. Historically, Japan was not well integrated within the global economy until government reforms encouraged companies to invest abroad; this shift led to Japan's position alongside the United States and Europe in total volume of outward FDI, which exceeded 120 billion US dollars in 2012 and had a growth rate of 14% in 2011–2012. Japan's inward FDI was less than two billion US dollars in 2012, whereas FDI for the United States and Europe is more balanced between inward and outward flows (UNCTAD, 2013; Branstetter, 2006). Major destinations of Japanese FDI are developed countries like the United States, where Japan's goals are asset exploitation and augmentation. East and Southeast Asia, which receive a considerable amount of FDI from Japan, provide locational advantages for Japanese FDI in manufacturing, in terms of labor cost. Japanese FDI in the United States has been able to capture technology and transfer it to the home country but has also made contributions to the knowledge base of companies within the United States (Branstetter, 2006). South Korea, Taiwan, Singapore, and formerly Hong Kong are also established regions both as sources and destinations of FDI.

Recently, China and India have made great strides in integrating into the world economy based on infrastructure development, which is essential to global competition (Ramamurti, 2010). China has been a low-cost provider of unskilled labor, whereas India has been a low-cost provider of skilled labor. Ramamurti (2010) provides a taxonomy of EMMNEs and their generic internationalization strategies as well as their country- and firm-specific advantages (**Table 1**).

The geographic context in which companies operate, as well as their specific advantages and goals, strongly determine the path to internationalization. *Companies involved in natural resources* tend to integrate vertically to secure resources and downstream markets; they also enjoy privileged access to natural resources and markets in their home country. *Local optimizers* tend to operate in low-income markets; they have strong customer connections, along with an ability to optimize products and processes for their customers. This ability results in a product portfolio suitable for market entry into other emerging markets, where they might have an ownership advantage (e.g., Chinese and Indian FDI in Africa). *Companies with low-cost but proficient project management* aim to become low-cost suppliers to traditional MNEs but try to move up the value chain by resource augmentation (e.g., Indian information technology companies). Additionally, these companies diversify their supplier structure through downward FDI. *Global consolidators* are often involved in mature industries with internationally standardized products (e.g., steel), which makes it easier for EMMNEs to engage in international markets and to compete on a similar technological level globally as traditional MNEs while often having economies of scale advantages. Global consolidators tend to acquire poorly performing companies to strengthen their global position as suppliers. *Global first-movers* are different from the previous MNEs, in that they are involved in emerging technologies and industries and have no considerable disadvantages as global latecomers. While these companies have cost advantages in their home market, they are also close to the technological frontier and are in direct competition with traditional MNEs. These companies tend to acquire technological capabilities and market access in developed markets through FDI, and they use FDI to broaden their production base. China and India are both endowed with enormous human capital resources with unique characteristics that provide inherent competitive advantages, and they have a large market size. Having a large base of low-income labor, provides EMMNEs with the ability to create resource-efficient products and to implement frugal innovation more readily than their traditional counterparts from developed countries (Bound and Thornton, 2012).

A new form of internationalization is closing the gap between MNEs and EMMNEs and it has instigated a major review of traditional FDI motives and patterns. Whereas traditionally MNEs became internationally active over a long period of time, the process is accelerated for EMMNEs. EMMNEs may not have the traditional competitive advantages (e.g., proprietary knowledge) and they use a strategy of knowledge augmentation by undertaking FDI in developed markets. A large advantage over established MNEs is that EMMNEs have high organizational adaptability and are able to operate in unstable political environments. To be able to sell globally, EMMNEs normally apply a dual expansion strategy: (1) they expand into other emerging markets (through new or greenfield investments) where they might have advantages based on their own assets (e.g., specific products) in comparison to local competitors and (2) they invest in developed countries where they augment their resources through acquisitions (e.g., access to patents and brands) (Guillén and García-Canal, 2009; Kedron and Bagchi-Sen, 2012). For example, the Indian conglomerate Tata bought Jaguar and Land Rover in the UK, and the Chinese company Geely acquired Volvo Cars. These strategic moves enabled involvement in developed markets and resource augmentation for their car manufacturing-related operations.

Impact of Foreign Direct Investment and Global Value Chains

The impact of international business activities is a critical topic both for MNEs and host countries. While the economic impact, e.g., spillovers, has received wide attention, the social impact and corporate responsibility issues are increasingly visible. Research on FDI/GVC impact is limited given the dearth of data and contradictory findings from various stakeholders. Generally, the impact of international business activities is strongly determined by their specific geographic, industrial, and organizational contexts.

Table 1 Generic internationalization strategies of EMMNEs

Generic strategy	*CSAs*	*FSAs*	*Internationalization path*	*Examples*
1. Natural-resource vertical integrator	● Natural resource endowment and/or ● Large home demand for natural resources	● Privileged access to natural resources and/or ● Privileged access to home markets	● Forward integration to downstream markets and/or ● Backward integration upstream to secure natural resources	● Gazprom, Lukoil, Norilsk, Vale, Anglogold, PTT ● Petrobras, ONGC, Indian Oil, CNOOC, Chinalco
2. Local optimizer	● Low-income consumers ● Under-developed 'hard' and 'soft' infrastructures	● Ability to optimize imported products and processes to home market ● Local-customer intimacy and local embeddedness	● Target market: Other emerging markets	● HiSense, Mahindra & Mahindra, Tata Motors, Shoprite, Marcopolo
3. Low-cost partner	● Low-cost labor ● Size of skilled labor pool, including engineers/ scientists, etc.	● Process excellence ● Project management ● Ability to operate successfully in the adverse conditions of emerging markets	● Target market: Developed countries ● Up-market FDI to move up value curve ● Down-market FDI to diversify supply locations	● Wipro, Infosys, TCS, Dr. Reddy's, WEG, Sabo
4. Global consolidator	● Large and rapidly growing home market ● Price-sensitive customers	● Production and project execution excellence ● Late-mover advantages in scale, organizational processes, technology ● Strong position in home market, with strong cash flows	● Target market: Global ● Up-market FDI to acquire poorly performing companies	● Tata Steel, Hindalco, South African Breweries, Lenovo, Wanxiang, Cemex
5. Global first-mover	● Large and rapidly growing demand in a new industry ● Low-cost country for design, engineering, and production	● Close to global frontiers of technology ● Strong position in home market, including, possibly, state support	● Target market: Global ● Up-market FDI to acquire key technologies or capabilities, and customer access ● Down-market FDI to gain market access and/or to diversify production bases	● Embraer, Huawei, Suzlon Energy, Check Point, Teva

Reproduced from Ramamurti, R., 2010. What have we learned about EMNEs. In: Ramamurti, R., Singh, J.V. (Eds.), Emerging Multinationals in Emerging Markets. Cambridge University Press, Cambridge, UK, pp. 399–426.

Among other impacts, knowledge transfer and spillovers are perceived benefits of FDI; governments partially subsidize FDI to capture these spillovers (Branstetter, 2006). Empirical evidence shows contradictions: some find evidence for a bidirectional flow of knowledge between the investing firm and the host country (e.g., Branstetter, 2006), whereas others find only a unidirectional exchange benefiting the investing MNE (see de la Potterie and Lichtenberg, 2001). Specific locational conditions, the position within a global value chain, and the corporate culture, affect the nature of spillovers. In the context of developing countries, spillovers are marginal, but a major concern with FDI is its negative impact on local competition, entrepreneurship, and forced industrial consolidation. In other words, FDI often squeezes out domestic competition and future possibilities of entrepreneurship (Oetzel and Doh, 2009). If local sourcing is encouraged, some spillovers may take place, but such spillovers are generally difficult to measure. Country-specific trade and industrial policies, the negotiating power of host governments, and MNE strategy are likely to affect the extent and quality of spillovers (UNCTAD, 2013).

A question related to FDI or the activities of MNEs is how they relate to international trade. More specifically, do they substitute trade or are they supplementing or creating trade? The traditional interpretation based on market imperfections argues that FDI is an alternative way (in lieu of exports) to serve a foreign market; that is, by producing locally to replace exports from the source country (Buckley and Casson, 1981). A host country can promote local production through import substitution policies and trade barriers to restrict inward FDI and thus promote the development of indigenous industry (e.g., India prior to market liberalization in the 1990s). Other observers argue that FDI creates trade through increased intra-industry trade and the opportunity to export adapted products to similar countries through internal corporate networks (Helpman and Krugman, 1985; UNCTAD, 2013). In the context of GVCs and GPNs, companies use intraindustry trade to optimize the location of different stages in the value creation chain in different countries. However, GVC/GPN, FDI and intraindustry trade relationships are context-specific: a country with restricted inward FDI, such as Japan, will differ from a country like Canada (Pantulu and Poon, 2003).

FDI and GVCs have the potential to capture value-added activities in their host countries. Especially for developing countries involved in or planning to be involved in GVCs, it is a challenge to participate in GVCs without adjusting social and environmental conditions, often to the detriment of both. FDI

involves transfer of capital and repatriation of profits. In this context, FDI and participation in GVCs can add value and contribute to GDP growth in the host country. If goods assembled in countries include a large percentage of imports, then value-added portions are compromised for the host nation. Job creation, income generation, conditions of work, social and environmental impacts are other areas of discussion connected not only to GVCs and FDI but also to other forms of transactions (e.g., arm's-length transactions) (UNCTAD, 2013). The potential of beneficial cooperation between MNEs and nontraditional partners (e.g., nongovernmental organizations) to achieve sustainable solutions is receiving more attention in the literature on inclusive growth and development (Oetzel and Doh, 2009). Finally, upgrading and building long-term productive capabilities are other development possibilities that might be linked to international business activities, e.g., the rise of original equipment manufacturers from Taiwan, Singapore, and now China.

Concluding Remarks

Although the global value of FDI fell to $1.35 trillion in 2012, international business remains important for the world economy (UNCTAD, 2013). In 2012, developing countries led as the top destination, absorbing 52% of total FDI, while FDI from and to developed countries fell. Deeper insights into relationships between FDI and GVC/GPN/GCC are needed. With new sources and forms of MNEs, the growing importance of knowledge intensity in goods and services, the transformation of markets, and the ever-changing political environment, international business is becoming more complex; this complexity demands multiple strategies to negotiate paths of development for the MNE and the source and host economies. Future research will focus increasingly on the role of intraindustry trade in adding value to the exporting country, and this research in turn will influence studies on strategies and structures of MNEs, which are now recognized to be deeply embedded within geographical contexts.

See also: Economic Geography; Foreign Investment: Direct; Industrial Geography; International Business; International Trade: Economic Integration; Location Theory; Multinational Corporations.

Bibliography

Bagchi-Sen, S., Wheeler, J.O., 1989. A spatial and temporal model of foreign direct investment in the United States. Economic Geography 65, 113–129.

Bound, K., Thornton, I.W., 2012. Our Frugal Future: Lessons from India's Innovation System. Nesta, London.

Branstetter, L., 2006. Is foreign direct investment a channel of knowledge spillovers? Evidence from Japan's FDI in the United States. Journal of International Economics 68 (2), 325–344.

Buckley, P.J., 2009. Internalisation thinking: from the multinational enterprise to the global factory. International Business Review 18, 224–235.

Buckley, P.J., Casson, M., 1976. The Future of the Multinational Enterprise. Holmes & Meier Publishers, New York.

Buckley, P.J., Casson, M., 1981. The optimal timing of a foreign direct investment. The Economic Journal 91, 75–87.

Buckley, P.J., Ghauri, P., 2004. Globalisation, economic geography and the strategy of multinational enterprises. Journal of International Business Studies 35, 81–98.

Caves, R.E., 2007. Multinational Enterprise and Economic Analysis. Cambridge University Press, Cambridge.

Coe, N.M., Hess, M., 2005. The internationalization of retailing: implications for supply network restructuring in East Asia and Eastern Europe. Journal of Economic Geography 5 (4), 449–473.

Coe, N.M., Dicken, P., Hess, M., 2008. Introduction: global production networks-debates and challenges. Journal of Economic Geography 8, 267–269.

Coughlin, C.C., Segev, E., 2000. Location determinants of new foreign-owned manufacturing plants. Journal of Regional Science 40 (2), 323–351.

Dicken, P., 2011. Global Shift: Mapping the Changing Contours of the World Economy, sixth ed. The Guilford Press, New York.

Dunning, J.H., 1977. Trade, Location of Economic Activity and the MNE: A Search for an Eclectic Approach. Macmillan, London, England.

Dunning, J.H., 2003. Some antecedents of internalization theory. Journal of International Business Studies 34, 108–115.

Dunning, J.H., 2009a. Location and the multinational enterprise: a neglected factor? Journal of International Business Studies 40, 5–19.

Dunning, J.H., 2009b. The Key Literature on IB Activities: 1960–2006, the Oxford Handbook of International Business, second ed. Oxford Handbooks Online. http://dx.doi.org/10.1093/oxfordhb/9780199234257.003.0002.

Dunning, J.H., Lundan, S.M., 2009. The internationalization of corporate R&D: a review of the evidence and some policy implications for home countries1. Review of Policy Research 26 (1–2), 13–33.

Gereffi, G., Humphrey, J., Sturgeon, T., 2005. Governance of global value chains. Review of International Political Economy 12, 78–104.

Glasmeier, A.K., 1986. High-tech industries and the regional division of labor. Industrial Relations: A Journal of Economy and Society 25 (2), 197–211.

Guillén, M.F., García-Canal, E., 2009. The American model of the multinational firm and the "new" multinationals from emerging economies. The Academy of Management Perspectives 23 (2), 23–35.

Helpman, E., Krugman, P.R., 1985. Market Structure and Foreign Trade: Increasing Returns, Imperfect Competition, and the International Economy. MIT Press, Cambridge, MA.

Hennart, J.F., 2010. Transaction cost theory and international business. Journal of Retailing 86, 257–269.

Hymer, S.H., 1976. The International Operations of National Firms: A Study of Direct Foreign Investment. MIT Press, Cambridge, MA.

Kedron, P., Bagchi-Sen, S., 2012. Foreign direct investment in Europe by multinational pharmaceutical companies from India. Journal of Economic Geography 12 (4), 809–839.

Mathews, J.A., 2006. Dragon multinationals: new players in 21st century globalization. Asia Pacific Journal of Management 23 (1), 5–27.

Oetzel, J., Doh, J.P., 2009. MNEs and development: a review and reconceptualization. Journal of World Business 44 (2), 108–120.

Pantulu, J., Poon, J.P.H., 2003. Foreign direct investment and international trade: evidence from US and Japan. Journal of Economic Geography 3 (3), 241–259.

Poon, J., Kedron, P., Bagchi-Sen, S., 2013. Do foreign subsidiaries innovate and perform better in a cluster? A spatial analysis of Japanese subsidiaries in the US. Applied Geography 44, 33–42.

de la Potterie, B.V.P., Lichtenberg, F., 2001. Does foreign direct investment transfer technology across borders? Review of Economics and Statistics 83 (3), 490–497.

Ramamurti, R., 2010. What have we learned about EMNEs. In: Ramamurti, R., Singh, J.V. (Eds.), Emerging Multinationals in Emerging Markets. Cambridge University Press, Cambridge, UK, pp. 399–426.

UNCTAD, 2013. World Investment Report 2013. Global Value Chains: Investment and Trade for Development. UNCTAD, United Nations Publications, Switzerland.

Yeung, H.W.C., 1999. The internationalization of ethnic Chinese business firms from Southeast Asia: strategies, processes and competitive advantage. International Journal of Urban and Regional Research 23 (1), 88–102.

International Communication: History

Daniel Headrick, Roosevelt University, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since ancient times, governments, merchants, and individuals have needed to communicate over long distances. The first public postal systems date from the seventeenth century. In the late eighteenth century, France built an optical telegraph network. Starting in the 1830s, electrical telegraphy spread in Europe and North America. The 1870s saw the laying of submarine telegraph cables connecting continents and islands. The early twentieth century saw the rapid development of international electronic communications: wireless telegraphy before World War I, radio broadcasting in the 1920s and 1930s, and global telephony and the Internet in the late twentieth century.

The first long-distance communication networks were imperial posts with messengers and relay stations. Communications between polities were carried out, sporadically, by traveling embassies. Medieval merchants, students, and clergy communicated across political boundaries. In the seventeenth century, the Thurn and Taxis family provided the first public postal service throughout Europe. Since the mid-nineteenth century, government postal services have cooperated to provide worldwide mail delivery. Napoleon built a semaphore telegraph network to link France, Spain, Italy, and Germany. Electric telegraphs crossed national borders in the 1840s. By the 1870s, telegraph lines covered Europe, North America, and India, and submarine cables connected Europe with other continents. In the early twentieth century, governments supported research into wireless telegraphy to communicate with ships and with one another. Shortwave radio, invented in the 1920s, competed with international cable and airmail services. International telephony now carries voice, fax, and data. Radio has broadcast news and entertainment since the 1920s. Today, satellites permit global television broadcasting. The Internet, still in its infancy, is already replacing paper letters between countries and is about to provide international telephone, radio, and television.

International communications includes communications between the inhabitants and between the governments of different states. Such communications have increased in speed, reliability, and volume since the beginning of civilization, until they now constitute a flood.

Ancient and Medieval Communications

Messengers were the first means of conveying information over long distances. Before the Battle of Marathon in 490 BC, Philippides, 'by profession and practice a trained runner,' ran to Sparta, 120 miles away, arriving 'on the very next day.'

Inhabitants of cities and isolated communities had no choice but to rely on messengers and travelers, but kings needed more regular communications. To keep control of their far-flung domains, the Persians established a permanent postal system with relays of horses along the Royal Road that stretched from Susa in western Iran to Ephesus on the Aegean Sea (Herodotus, 1987). The Romans created a similar network of imperial messengers and relay posts along their major roads, called *cursus publicus*. The Chinese, Inca, Mongol, and other empires established similar systems. These early imperial networks were reserved for official messages; anyone else who had a message to send would have to find a traveler going in the right direction, and hope for the best.

The Middle Ages saw the decline or disappearance of the governmental apparatus of classical times and the emergence of numerous unofficial channels of communication, some of them well organized. The Benedictine monks of Cluny entrusted pilgrims with messages to other monasteries from Spain to Poland. The merchants of the Hanseatic League carried documents and messages along with merchandise. Butchers, who traveled from place to place buying and slaughtering livestock, carried messages for people. The most organized of these nongovernmental networks was the one centered on the University of Paris, whose students came from all over Europe and carried and sent messages between Paris and their homelands. Although they were supposed to serve only the university community, in fact they commonly carried messages for the public as well. These networks were slow and unreliable, however, for they did not operate on a schedule, nor did they use relays of horses or messengers.

Postal Communications

The first postal system that established permanent relays and offered its services to the general public was founded in 1489 by Franz von Taxis, postmaster to the Holy Roman Emperor Maximilian I. The Thurn and Taxis family provided a regular postal service throughout most of Europe until 1867. With up to 20 000 messengers and relays of horses on all the main routes, their postal service was reliable and, by the standards of the day, fairly swift: from Brussels to Paris in 44 h, to Innsbruck in 6 days, and to Toledo in 12 days. Service, however, was very costly.

The Habsburgs may have trusted the Thurn and Taxis family, but the rising national monarchies of France and England did not, nor did they trust any other private or nongovernmental message service. The first royal messenger service in France, established in 1464, was reserved for official messages. Private individuals could bribe the royal couriers to carry messages surreptitiously. The royal postal service was opened to the public in the early seventeenth century. It set the rates for private letters very high, to subsidize the official mail.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.95056-6

Once the royal mail service was opened to the public, the government attempted to suppress private messengers in order to protect its lucrative monopoly and permit easier surveillance of potentially subversive communications; the last nongovernmental service in France, that of the University of Paris, ended in 1719.

The evolution of the English postal system paralleled that of France. There had long been royal messengers, but the first permanent service was founded in 1512. Royal messages to and from foreign countries, however, were still hand-delivered by special messengers. Meanwhile, merchants had their own networks, such as the Merchant Adventurers' Post for English merchants engaged in foreign trade, and the Foreigners' Post for foreign merchants in England. Queen Elizabeth, suspicious of Catholic plots, ordered all mail leaving or entering the country to be carried by the royal post. This decree, though often repeated by her successors, was never fully enforced.

In 1635, the royal postal service began to serve the public; 2 years later it included foreign mails as well. As in France, 'public service' was a euphemism, for postal rates were kept extremely high as a source of income for the treasury and for the king's favorites and officials who had the right to open and read private letters. Surreptitious inspection of letters, especially those from abroad, was a common occurrence in the eighteenth and early nineteenth centuries. Most people who needed to correspond across borders preferred to send letters with a friend, a servant, a hired messenger, or a traveler. Surveillance abated in the mid-nineteenth century, as the increased volume of mail overloaded the postal bureaucracies.

The mid-nineteenth century saw a tremendous surge in the volume of mail, thanks to two innovations. One was the penny post with prepaid stamps, introduced in England by Rowland Hill in 1840 and soon imitated in other countries. The other was railroads and steamships, which made transportation cheap and fast. Letters from relatives or friends in foreign countries, a rarity until then, became a commonplace experience for middle-class families in the Western world (Headrick, 2000).

Airplanes began transporting mail right after World War I. In parts of the world, such as Africa and South America, that had poor telegraph and rail connections, airmail represented a breakthrough comparable to the effect of railways in Europe a century earlier. In places well served by traditional postal services, airmail competed with the cheapest kind of telegram, the 'night letter.' After World War II, airmail became the standard means of sending first-class postage over long distances.

The Telegraph

Many scholars believe that modern telecommunications began with the electric telegraph. From a technical perspective, the advent of electricity was certainly revolutionary. From an information point of view, however, the revolution began half a century earlier with the optical telegraph, invented by Claude Chappe in France in 1794.

Throughout the French Revolution and after, every French government expanded the optical telegraph network and supported the family that ran it. Chappe and his brothers built lines east to Strasbourg, west to Brest, and north to Dunkirk and Brussels; under Napoleon, they extended the network south to Lyon and Marseilles. As Napoleon's empire grew, lines radiated to Turin, Milan, and Venice, to Antwerp and Amsterdam, and to Mainz. After the defeat of Napoleon the French network was reduced to its domestic lines. Other European countries, as well as the USA, Egypt, and Australia, built individual lines linking their major cities to the nearest harbors. Some of these were government-owned lines reserved for official or military traffic, while others served the merchant communities. In the few instances where two systems met at a border, any message going from one country to the other had to be decoded, translated, hand-carried across the gap, and encoded for further transmission. International telecommunication had to await the electric telegraph (Wilson, 1976; Holzmann and Pehrson, 1995).

During the 1830s, several inventors came up with methods of transmitting information over long distances by electricity. The one that eventually displaced all others was invented by Samuel Morse. Morse's contribution was not the use of electricity but a code which allowed the use of only one wire, with the earth serving as a return. The cost savings of Morse's system made up for the need for trained operators.

During the next two decades, the nations of Europe and North America began installing electric telegraph lines. At first, they used incompatible systems; messages had to be written out, translated, and hand-carried across the border, then retransmitted. As traffic grew, governments strove to overcome this barrier to communication. In 1850, Austria and several German states formed the Austro-German Telegraph Union. Treaties between France and its neighbors followed. Finally in 1865, all European governments except the UK founded the International Telegraph Union, to be administered by the International Bureau of Telegraph Administrations, the first permanent international organization. It encouraged national telegraph administrations to adopt compatible technical standards and to connect their lines across border. By the 1860s, it was possible to send telegrams across Europe and between the USA and Canada (Griset, 1991; Flichy, 1991).

Meanwhile, engineers had solved the technical problems of communicating across the sea. The first successful submarine cable crossed the Channel in 1851. Cables across oceans were a greater challenge, however. The first Atlantic cable failed in 1858, as did the first cable from Egypt to India 2 years later. After many trials and errors, cable technology finally caught up with the demand. The Atlantic cable of 1866 proved reliable. In the 1870s, British companies rushed to lay cables from the UK to India and from there to China and Australia. Cables reached South America and Africa in the 1880s. By World War I, cables criss-crossed all the oceans and connected every continent and every country (Barty-King, 1979; Headrick, 1991; Winseck and Pike, 2007).

Submarine cables were phenomenally expensive, however. Only the North Atlantic had enough traffic to justify several competing cables. In other parts of the world, the high cost of cables had to be amortized over small numbers of customers, and a few firms, most of them branches of the British conglomerate Eastern and Associated, dominated the market. The high cost of sending messages over long distances led business customers and news agencies to develop elaborate code books.

Wireless Telegraphy

Although the telegraph had 'annihilated time and space' (to use a favorite expression of nineteenth-century boosters), it left many people dissatisfied, for both economic and political reasons. Economically, international telegrams were very expensive, and transoceanic ones exorbitantly so, placing them out of reach of all but the wealthiest clients and commercial enterprises. And politically, the fact that most ocean cables belonged to British firms and that most intercontinental telegrams passed through British hands, made all other governments very uncomfortable, especially after 1900, when the UK proved that it was not averse to intercepting, delaying, and censoring foreign messages during a crisis like the Boer War (1900–02).

In 1895, Guglielmo Marconi demonstrated the possibility of transmitting information through space by electromagnetic radiation. He found immediate backing from the Royal Navy and from British shipping lines, which had long wanted to communicate with their ships at sea. France, Germany, and the USA also greeted the new technology with enthusiasm, for they had fallen far behind the UK in the cable business and saw radio as a means of building international connections independent of the UK (Aitken, 1985b).

Until the late 1920s, wireless used longwaves (over 1 km, or less than 300 kHz), which required power in proportion to the distance to be covered; stations powerful enough to transmit across continents or oceans needed dozens of huge towers, miles of antennas, and alternators that used as much electricity as a small town. Within countries, wireless could not compete with the cheap and efficient telegraph networks. Across oceans it was competitive only under unusual circumstances. Before World War I, the four major powers built stations linking them to their colonies. During the war, when it was impossible to lay cables, the US expeditionary force built stations in France to supplement the overloaded transatlantic cables. After the war, to meet the booming demand for transatlantic communications, companies found it faster and cheaper to build gigantic transmitters than to lay costly cables (Aitken, 1985a; Anduaga, 2009).

Longwave wireless had two flaws to which cables were immune: they were vulnerable to atmospheric static and to interception and cryptanalysis. The British, with their heavy investment in cables and their far-flung colonies and bases, were concerned with keeping their cables viable. Thus, it came as a shock when Marconi, the man who had invented wireless, announced a new technology that threatened to make both longwave radio and submarine cables obsolete, namely shortwave radio (*c.*13–120 m or 2.5–22 MHz).

Shortwave, though more vulnerable to atmospheric conditions and less reliable than longwave, had one enormous advantage: it was extraordinarily cheap; a transmitter cost a few 100 dollars instead of hundreds of thousands. Even small inexpensive transmitters could, at certain hours of the day, reach across oceans and around the globe.

The advent of shortwave coincided with the Depression. The collapse in international trade brought the British cable companies to the brink of bankruptcy, but they were rescued at the last moment by their government for strategic reasons. Although cables served a useful function in World War II for the transmission of top-secret information, they were overshadowed by the proliferation of shortwave transmitters throughout the world. Indeed, a war of movement over enormous theaters of war like the Russian plains or the western Pacific would have been inconceivable without a multitude of radios. At the same time, the reliance of armed forces on radio meant that the outcome of battles depended on communications intelligence and cryptanalysis as much as it did on ships, tanks, and planes.

Broadcasting

Broadcasting, the transmission of radio and later television programs to many receivers at once, played a tremendous role in the cultural history of every twentieth-century nation. The first experiments in transmitting voice and music over the airwaves took place before and shortly after World War I. However, broadcasting had little impact on international relations before the 1960s for two reasons. One was that most broadcasting was in the medium wave range (150–600 m, or 500–2000 kHz), which did not carry far; only border regions could receive foreign broadcasts. The other was that programs were in the language and culture of each country (or rather, of each country's elite), with little appeal to foreigners (Douglas, 1987; Anduaga, 2009).

International shortwave programs, such as those pioneered by the BBC in the early 1930s, were specifically designed to appeal to people in other countries. Soon, every major country had its shortwave station beaming news and propaganda to the world, often in many languages. No doubt such broadcasts appealed to the converted and reminded exiles of their homelands, but whether they persuaded more than a handful of foreigners is questionable.

Mention should be made of the interesting cultural phenomenon of ham radio operators who communicated around the world, not so much to convey information as to collect lists of other distant amateurs they had contacted, much as birdwatchers collect the names of birds they have spotted.

Television uses high-frequency waves that cannot bend with the curvature of the earth. Until the advent of satellites in the 1970s, transmitters could only reach local viewers. Satellite television, however, has opened up the world to broadcasts from foreign countries; all that is needed is a dish antenna. Countries with conservative theocratic or tyrannical regimes resent the intrusion of such foreign cultural phenomena as sex, violence, and free speech. However, prohibition is hard to enforce, and people around the world are much more interested in foreign television programs than they ever were in shortwave news broadcasts.

International Telephony

Telephony is a recent phenomenon. Though invented in the 1870s, it affected few people before the early twentieth century in North America, the 1920s in western Europe, and the post–World War II period in other parts of the world; at the

beginning of the twenty-first century, there were fewer telephones in all of Africa than in Manhattan.

International telephony is an even more recent phenomenon. Due to the bandwidth requirements of the human voice, telephone messages could not use telegraph wires, but needed repeaters every few 100 miles. International calls were possible only between adjacent countries. Most telephone systems were government controlled, and calls to foreign countries had to go through an operator and were extremely costly. Calls overseas required two radio stations in each country, one to transmit and one to receive. Until the 1950s, such calls were considered newsworthy events rather than a medium of communication.

This situation began to change after the first transatlantic telephone cable was laid in 1954, and especially after the advent of satellite transmission of telephone calls in the 1970s and fiber-optic cables in the 1980s. As prices fell from dollars to pennies a minute, the number of international calls has multiplied exponentially, with no end in sight. Businesses and even many individuals now make calls to distant countries as readily as their parents once called across town.

As this is written, a new technology promises to change international telephony as dramatically as the first telephone cables: low-earth-orbit satellites that can relay calls to and from portable telephones anywhere on earth, bypassing national networks. Although governments and national companies are likely to resent this intrusion on their monopolies, telecommunications technology is too invasive to stop. Eventually, the well-to-do will be connected around the world, regardless of governments and borders, while the rest of humankind looks on in wonder (Young, 1991; Pool, 1990).

Data Communication and the Internet

The current revolution in communication is as radical as the telegraph's 'annihilation of time and space' that so amazed our ancestors a century and a half ago. By turning text, voice, music, pictures, and video images into bits, electronic digitization allows the transmission of all forms of information with equal ease and speed. Two associated technologies, microwaves and fiber-optic cables, have reduced the cost of transmission almost to zero. The results are nothing short of phenomenal. The first transatlantic telephone cable could carry a dozen or so telephone calls at once; the most recent fiber-optic cables across the Atlantic carry hundreds of thousands of simultaneous calls, or dozens of television channels, or any combination of data, voice, text, or images. Communication is not free because other costs remain, such as switching and distribution. But the price differences between local, long-distance, and international communications are shrinking fast.

The Internet began in the 1970s as a US military project to link the mainframe computers of research institutions. By the mid-1980s, when personal computers began to proliferate, engineers in Europe, North America, and Japan first created simple networks to transmit data, then linked them together into a super network called the Internet that could transmit any information to any connected computer. The demand for this new form of communication soared beyond anyone's prediction, as the public logged onto a virtual world of chat rooms, bulletin boards, electronic mail, shopping, entertainment, advertisements, and even music and video clips from around the world. And the cost is very low. For people who own a computer (even a primitive 'Web-TV' box) this amazing service costs a few dollars a month. All other costs (servers, cables, switches, software, and administration) are borne by governments or by advertisers. The Internet offers telephone and television service and competes with libraries, newspapers, and universities. And unlike the technologies that it replaces, the Internet is oblivious to borders and political jurisdictions. We can only speculate about the effect on cultures and on the politics of nation-states.

Meanwhile, computers, microwave radio transmission, and the digitization of data have revolutionized telephony. Mobile telephones, now in the billions, have reached almost everywhere on earth. Though the first ones carried only voice, newer 'smart phones' connect to the Internet and carry pictures, music, videos, and information of all sorts (Lubar, 1993).

What the public sees is only a fraction of the world of international digital data transmission. Computer networks, connected by fiber-optic cables, have revolutionized finance and commerce as well. Credit-card charges, engineering drawings, business contracts, stock market reports, late-breaking news, banking transactions, and myriad other data flit across the globe at the speed of light. The amount of money – in the trillions of dollars – that circulates at any given moment dwarfs the economies of the largest countries. The new communications media offer the potential for phenomenal economic growth and for disastrous crashes. Humans may still be bound by the laws of physics, but for information, space and time have truly been annihilated.

See also: Communication and Transportation, History of; Information Society; Postal Systems; Printing as a Medium; Radio as a Medium; Science and Technology, Social Study of: Computers and Information Technology; Telegraph; Telephone; Television: History.

Bibliography

Anduaga, A., 2009. Wireless and Empire: Geopolitics, Radio Industry, and Ionosphere in the British Empire, 1918–1939. Oxford University Press, Oxford.

Aitken, H.G.J., 1985a. The Continuous Wave: Technology and America Radio, 1900–1932. Princeton University Press, Princeton, NJ.

Aitken, H.G.J., 1985b. Syntony and Spark: The Origins of Radio. Princeton University Press, Princeton, NJ.

Barty-King, H., 1979. Girdle Round the Earth: The Story of Cable and Wireless and its Predecessors to Mark the Group's Jubilee, 1929–1979. Heinemann, London.

Douglass, S.J., 1987. Inventing American Broadcasting, 1899–1922. Johns Hopkins University Press, Baltimore, MD.

Flichy, P., 1991. Une histoire de la communication moderne. La Découverte, Paris.

Griset, P., 1991. Les révolutions de la communication, XIXe–XXe siècle. Hachette, Paris.

Headrick, D.R., 1991. The Invisible Weapon: Telecommunications and International Politics, 1851–1945. Oxford University Press, New York.

Headrick, D.R., 2000. When Information Came of Age: Technologies of Knowledge in the Age of Reason and Revolution, 1700–1850. Oxford University Press, New York.
Herodotus, 1987. In: Grene, D. (Ed.), The History Book VI. University of Chicago Press, Chicago, pp. 105–106 (Trans.).
Holzmann, G.J., Pehrson, B., 1995. The Early History of Data Networks. IEEE Computer Society Press, Los Alamitos, CA.
Lubar, S., 1993. InfoCulture: The Smithsonian Book of Information Age Inventions. Houghton Mifflin, Boston.
Pool, I.D.S., 1990. Technologies without Boundaries: On Telecommunications in a Global Age. Harvard University Press, Cambridge, MA.
Wilson, G., 1976. The Old Telegraphs. Phillimore, London.
Winseck, D., Pike, R.M., 2007. Communication and Empire: Media, Markets, and Globalization, 1860–1930. Duke University Press, Durham, NC.
Young, P., 1991. Person to Person: The International Impact of the Telephone. Granta, Cambridge, UK.

International Council on Social Welfare

Sergei Zelenev, International Council on Social Welfare, Bronx, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

As the International Council on Social Welfare (ICSW) looks back over its past 85 years of history, closely linked to the history of social work and social practice, unfolding across geopolitical borders and aimed at improving the human condition and well-being in the face of socioeconomic uncertainties and risks generated by market-driven societies, the organization recognizes not only its achievements, but also the multiple challenges. For the ICSW, retaining its influence and finding a niche in the new international environment requires much professionalism and dedication and a constant quest of the meaningful, innovative, and cost-effective policies.

The History of the International Council on Social Welfare in Brief

Established in 1928, the International Council on Social Welfare (ICSW) is one of the oldest international nongovernmental organizations (NGOs) aimed at promoting social development, social justice, and social welfare everywhere in the world. The ICSW's history is closely linked to the history of social work and social practice, unfolding across geopolitical borders and aimed at improving the human condition and well-being in the face of socioeconomic uncertainties and risks generated by market-driven societies. While the emphasis placed by the ICSW on international social work and social development, nationally and internationally, has been shifting, the balance between these has evolved. However, a two-pronged approach has been a hallmark since its inception.

The immediate predecessor of the ICSW – the first International Conference on Social Work – was convened in Paris in July 1928. That ground-breaking event became possible following the dedication and vision of several extraordinary personalities willing and capable of assuming leadership in the preparation of the Conference (Among these, the name of the Belgian doctor, scientist, and educator, René Sand (1877–1953) stands out. In 1921, he became Secretary-General of the 'League of Red Cross Societies,' which was created in 1919. Sand endorsed the idea of a large-scale world forum on social welfare in 1923, thereby reviving a proposal floated earlier by Dr Clotilde Mulon at the 46th US Conference for Social Work. Julia Lathrop, the President of the National Conference for Social Work, invited René Sand to address the 50th American Conference on the issue of international cooperation, where he was able to generate a positive reaction to the idea.) in the aftermath of World War I, which brought devastation and suffering to many nations. Preparatory work for the International Conference began in earnest with several meetings of the Preparatory Committee being held over a 3-year period. These culminated in a meeting in Prague in 1927, where the guidelines for the organization of the conference were elaborated and adopted (Le Conseil international d'action sociale, 2008). The stated goals of the conference were fostering discussion and idea sharing between social workers and social welfare organizations from around the world, assistance with dissemination of ideas and professional information, and establishing personal contacts. These efforts promoted international cooperation within the social work profession and beyond, bringing on board numerous activists, experts, and practitioners, and working apart from state-sponsored structures for improving the human condition. The largest intergovernmental organizations of that time – the League of Nations and the International Labour Organisation – provided technical support, with financial backing from several national and regional foundations.

Mapping and identifying the field of proposed activities, the first International Conference in Paris underscored that social work could not be constrained to individual cases but required broad-based interdisciplinary efforts to relieve suffering resulting from want (i.e., providing palliative assistance), to restore people to 'normal' conditions of life (through curative assistance), to prevent social afflictions (preventive assistance), and to improve social conditions and raise standards of living (constructive assistance) (Macdonald, 1975). The importance of individual practice was acknowledged, along with the need to promote legislative and administrative measures targeting a group or community and coupled with research in the fields of social well-being and social welfare. Attempting to delineate a field even further, the participants at the Paris Conference underscored a link between the need to prevent or address social hardships and the availability of resources for social purposes. (Premiere Conference international du service social, Paris, 1928, tome I, 1929, Statuts, pp. 25, 172.) There was agreement on the need for professional qualifications and skills, but opinions among more than 2500 participants, who represented a diverse group of social workers, health care providers, welfare officers, and other activists, differed as to the respective roles of the state and private institutions. The variety of views was also reflected in the submissions of national committees and in publications of the executive committee, even though there was an agreement about the basic values of international social work practice and the general directions as outlined above.

Since the first conference, the importance of the thematic approach was recognized; the first priority theme was 'Social Work and Industry.' About two-thirds of the participants at the Paris Conference were women. The Conference elected a woman – Alice Masarykova, the President of the Czechoslovakian Red Cross – as the first President of the ICSW for 4 years.

Doctor René Sand became the first Secretary-General of the International Conference on Social Work, a post he held from 1928 to 1932, after the Provisional Preparatory Group of the Paris Conference transformed itself in 1928 into a permanent organization under the name "International Conference on Social Work" (Eilers, 2003).

The aim of this move was to capitalize on the positive momentum and translate the wealth of ideas that had emerged at the first conference into practice so as to continue its activities without any interruption. René Sand became President in 1936 and held that position until the outbreak of World War II (WWII). Having become a permanent organization, the International Conference on Social Work represented national councils on social work and social welfare, promoting humanistic values and advocating for disadvantaged population groups, and also assisting its members with information and consultative work. At the same time, recognition of the substantial differences in approaches taken at the national level and counterproductive responses to encouraging the same strategy for everyone led to acceptance of the substantial autonomy of each organization (Le Conseil international d'action sociale, 2008, p. 16).

Before WWII, two international conferences were convened with a 4-year interval – in 1932 in Frankfurt, Germany, on the topic 'Social Work and the Family,' and in 1936 in London, UK, on 'Social Work and the Community.' The ICSW global conferences resumed in 1948 after the interruption caused by WWII. Since that time, the global conferences of the ICSW have been held periodically, initially every 4 years, and since 2010 every 2 years. Global conferences on social work and social development were organized through close collaboration among the ICSW, International Association of Schools of Social Work (IASSW), and International Federation of Social Workers (IFSW) – the three sister organizations as they have been termed. (In 1929, an International Committee of Schools of Social Work was established to give prominence to the specific needs of social work education, explore best practices, and facilitate international cooperation; the first conference of IASSW was in 1929 in Berlin. The first conference of IFSW was held in 1956 in Munich, at the same time as the ICSW conference (see Le Conseil international d'action sociale, 2008, p. 50).)

The international conferences regularly convened by the ICSW either separately or together with IASSW and IFSW have dealt with recurring issues and new subjects that reflected political realities and deep changes on the global scene. During the 1950s and 1960s, the emergence as independent political actors of a large group of developing countries facing entirely new social and economic challenges caused the ICSW to recast its agenda on socioeconomic development to encompass social rights, social welfare, and well-being, including dominant policy discourses around the improvement of the lives of those living in poverty, participation, social cohesion, community development, and growing significance of social services. These have become a constant feature of the global discourses on social issues, gradually expanding the scope and objectives of social welfare goals, analyzing interaction of key actors, and presenting analysis of policy instruments (Gough, 2004). The membership of the ICSW expanded, as national councils from many developing countries in Asia, Africa, and Latin America became members.

Organizational adjustments followed, and in 1966, the ICSW – which until then was known as the 'International Conference on Social Work' – changed its constitution and name to the 'International Council on Social Welfare,' without having to change its English acronym, ICSW. Commenting on that change, Kate Katzki, the Secretary-General (In 1996, the position of the 'Secretary-General' was renamed the 'Executive Director.') of the ICSW from 1967 to 1988, underscored the fact that the ICSW could no longer focus on only one single goal, namely, an international forum for discussion among social workers. (Kate Katzki, 60 years of ICSW, New York, 1988, p. 16 quoted in Le Conseil international d'action sociale, 2008, p. 32.) The name change reflected the profound structural changes in the organization, the growing diversification in its activities, and the higher political profile sought by the ICSW leadership.

Structure and Organization of Activities

According to the Constitution of the ICSW, (The most recent constitution was adopted by the Committee of Representatives on 29 December 2007, and amended on 14 June 2010. The Committee of Representatives itself was given a new name of 'General Assembly,' retaining the same duties and prerogatives.) "the Council" is a "global, non-governmental, non-political, non-sectarian and non-profit organization which represents a wide range of national and international member organizations that seek to advance social justice, social welfare, social work and social development" (The ICSW Constitution, clause 1).

The primary objectives of the Council are as follows: Promote forms of social and economic development that aim to reduce poverty, hardship, and vulnerability, especially among disadvantaged people; strive for recognition and realization of fundamental rights to employment, income, food, shelter, education, health care, and security; promote equality of opportunity, freedom of expression, freedom of association, participation, and access to human services and oppose discrimination; promote policies and programs that strike an appropriate balance between social and economic goals and that respect cultural diversity; strengthen civil society throughout the world to achieve these objectives; seek implementation of these objectives by governments, international organizations, and nongovernmental agencies in cooperation with its network of member organizations. (The ICSW Constitution, clause 1.)

The *General Assembly*, which adopts the 4-year global program and the 2-year budget framework, is the supreme governing body of the Council, electing the President, Vice President, Treasurer, and other members of the *Management Committee*, responsible for all matters not specifically allocated under the Constitution to the General Assembly or to the *Supervisory and Advisory Board*, which comprises the Regional Presidents, the President, the Vice President, and the Treasurer. It is entrusted with supervision and monitoring the *Management Committee*, creating and amending the bylaws, and establishing and restructuring regions. (The ICSW Constitution, clause 9, 10, and 11.)

While the ICSW is a worldwide nongovernmental organization with regional structures, it is not a mass organization, and only one national council from each country can be accepted into membership. The regional organizations are indigenous to the regions. The regions play an active role, but membership is assumed directly with the international headquarters, and not via regions. The ICSW is a unified organization that could be compared to an amalgam rather than a mosaic as far as its membership is concerned. (ICSW membership is classified into several different categories. Category A covers national committees having a broadly based membership in their respective countries and seeking to enhance social justice, social welfare, social work, and social development in that country. Category B is reserved for international nongovernmental organizations trying to achieve the same objectives. Categories C and D are intended for civil society organizations that cannot claim to work on a full range of issues and be the key party representing social welfare issues in the country, but seeking to enhance social justice, social welfare, social work, and social development anyway. (The ICSW Constitution, Clauses 5–7).)

The regional structure is established under the bylaws, and regions have their own decision-making bodies, as well as their own programs of activities. Each member, with the exception of International Member Organizations, is allocated to a region by the Board for the purposes of the allocation of membership fees, and each member is entitled to the same rights in the region as globally. At the same time, there is substantial organizational flexibility – any member organization is able to cooperate with any member organization in other regions, as well as, by agreement, participation in other regional meetings, activities, and ICSW structures. Regional General Assemblies and other regional organs have been established under global and regional by laws. (The ICSW Constitution, clause 8.) Some regional entities have their own legal status separate from the ICSW.

The President of the ICSW is the highest elected official and representative of the Council and is elected to a single 4-year term, without the possibility of a second consecutive term. In his capacity of the strategic leader of the organization, the President sets overall direction and proposes innovative policy solutions, maintaining interaction with other partners and stakeholders at the regional and global levels.

The ICSW Executive Director is responsible for implementing the decisions of the General Assembly, Board, and Management Committee and for keeping the Management Committee informed on issues that are important to the proper functioning of the organization. The Executive Director combines supervision of the day-to-day operations of the Global Office with substantive outreach activities, editing the Global Cooperation Newsletter published by the ICSW in English, French, and Spanish, and maintaining links with various partners, including other NGOs and intergovernmental organizations.

The ICSW has consultative status with the United Nations Economic and Social Council (UN ECOSOC), International Labour Organization (ILO), and other UN agencies. This status enables the ICSW to address UN bodies through broad-based statements and advocacy, participate in the intergovernmental discussions, and present recommendations to intergovernmental bodies. The ICSW's Global Programme is a strategic framework of the organization designed to operate at the global, regional, and national levels, setting the goals and umbrella policy for the Council operations for 4 years.

Expanding the Scope of Work at the Global Level

Through the years, the priorities of the ICSW have been changing, reflecting the political changes in the world, the development and refinement of a global perspective in its work, as well as the growing awareness of the need to reconcile social and economic policy. In the face of the progressing integration and liberalization of the world economy and the new intensity of international competition – with its profound, and not necessarily positive, effect on social spending and delivery of social services – the welfare situation has varied significantly from one region to another, with growing uncertainty at the local, national, and regional levels and widespread setbacks in living standards.

The development discourse has also not been static. Human development has emerged as one of the central concepts in international development since the early 1990s, supported not only by civil society but also by many governments and international organizations. The need to view people as the ends of development as well as its means, focus on both poor and the poorest people, and prioritize capability enhancing services (such as food security, education, and health), and the multidimensional conceptualization of human well being and poverty have shifted the focus of many economists and policy makers away from the earlier emphasis on per capita economic growth as the sole indicator of socioeconomic progress. The 1990s marked a watershed in the evolution of ideas about poverty reduction (Hulme, 2007).

At the onset of the 1980s, the ICSW had three core priorities: convening conferences and ensuring space for discussions, providing support to national committees, and building stronger relations with the UN. The situation changed in the late 1980s and early 1990s, when a more comprehensive agenda emerged both at the ICSW headquarters and in the regions, covering such themes as access to the means required for economic independence, strengthening social infrastructures, the issue of violence and social development, and the social dimensions of technological change. Suggestions were made to go further and address on a consistent basis topics such as the significance of aging, human immunodeficiency virus/acquired immunodeficiency syndrome pandemics, and North–South relations. In the 1980s, a gradual shift in the philosophy of the organization became apparent, including its main thrust from the idea of 'social work' to 'social welfare' and then to 'social development' (Le Conseil international d'action social, 2008, p. 45).

A deliberate and consistent move was made in the 1970s and 1980s to go beyond social work and enhance social development dimension in its global activities and broaden the scope of ICSW's activities. The interdisciplinary and holistic nature of the ICSW approach to development thinking formed after WWII and was vividly demonstrated at the ICSW global conferences. The ICSW has been invited to take part in numerous conferences, expert group meetings, research

seminars, and other events of international significance, to share its experiences and provide substantial inputs. Its global image was strongly reinforced during the preparation of the World Summit for Social Development (WSSD) convened in Copenhagen in 1995.

The preparatory process of the WSSD provided the ICSW with an opportunity to present its analysis and vision of the world social situation to influential global audiences. First of all, the ICSW actively participated in the work of the preparatory committees for the summit, promoting in accordance with its Constitution forms of social and economic development that aimed to reduce poverty, hardship, and vulnerability throughout the world, and making specific suggestions regarding ways to ensure the progressive realization of all human rights, including economic and social rights. Apart from discussions at intergovernmental meetings, deliberate efforts were made to link closely the activities of the ICSW with the preparatory process for the Summit. For example, the Tampere Manifesto adopted by the 1994 ICSW global conference was conceived with the Summit agenda in mind; the document advocated "an integrated and integrative approach to social development," promoted rights to equal opportunities, supported grassroots initiatives, and strongly condemned violence, intolerance, and xenophobia. It also addressed the priority themes of the WSSD: poverty eradication, employment promotion, and social integration, and was submitted to the Preparatory Committee. (Le Conseil international d'action sociale, 2008, pp. 70–71. Summing up the ICSW's contributions and addressing the Plenary Session of the World Social Summit in Copenhagen on 7 March, 1995, ICSW President Dirk Jarré indicated that the Council "has been actively and constructively involved in every phase of the preparatory process leading to the Summit. We have mobilized our global resources—in personal, intellectual, organizational and financial terms—to inform NGOs about the Summit and to motivate them to participate in the endeavour, to contribute ourselves directly to the Summit's content and to give support to the process itself" (Address to the Plenary Session of the WSSD by Dirk Jarré, 1996. In: Social Priorities of Civil Society. Speeches by Non-Governmental Organizations at the World Summit for Social Development, NGLS, New York, p. 60.).

The WSSD represented an important political threshold for the ICSW, not only in terms of opportunities to contribute to the outcome documents but also by the fact that it became an important benchmark for ICSW-affiliated organizations during the implementation phase; in many ways ICSW activities in the post-Summit period became aligned to the Ten Commitments made in Copenhagen and the recommendations of the Plan of Action adopted at the Summit. The principles of social justice, equity, and equitable societies, high on the WSSD agenda, are shared by the ICSW, which has been striving for many years to put these principles into practice on the ground and promote them in high-level international forums. Creating an environment to enable achievement of social development, eradicating poverty through decisive national action and international cooperation, promoting full employment as a basic priority, and fostering stable, safe and just societies to promote social integration – all those commitments that emerged at the Summit have been taken on board as guiding principles by the ICSW in its own activities. The global consensus reached at the Summit that poverty eradication should remain a priority goal for development was a crucial element that gave a forward momentum to several new post-Summit initiatives, including the elaboration and adoption of the Millennium Declaration approved on 8 September 2000, and the elaboration of the Millennium Development Goals (MDGs) a year later based on the Millennium Declaration where the eradication of extreme poverty and hunger were put on the top of the list.

The ICSW also actively participated in the preparation of the twenty-fourth special session of the General Assembly convened in Geneva in June 2000, where the commitments of the WSSD were reviewed with respect to further action and initiatives. As a follow-up to the WSSD, the ICSW has participated in all of the annual sessions of the Commission for Social Development (CSocD) – one of the functional Commissions of ECOSOC, which reviews the outcome of the Summit every year. The written and verbal statements made by the ICSW representatives highlight the efforts of the ICSW-affiliated organizations in the context of the implementation of the Summit commitments and increased priority that social development has assumed vis-à-vis other policy objectives.

Some key topics and crosscutting themes raised in the statements could be mentioned. Depending on the context of the discussions and the priority theme of the CSocD and in line with ICSW objectives, the ICSW statements addressed issues such as the urgent need of an integrated approach to social development, the missing links between strategy and policy, called for more focused policies aimed at the equitable distribution of resources and opportunities. Reflecting the spirit of the World Social Summit, the statements made by ICSW representatives sought to highlight the goal of poverty eradication and draw attention to the role of employment as the nexus between economic and social policy and the plight of the working poor and the need to address growing inequality within and between nations. The ICSW consistently speaks for social inclusion and promoting 'a society for all' – a goal identified by the WSSD as one of the core priorities.

Promoting social protection as one of the pillars of a forward-looking social policy has become one of the key priorities and recurrent themes in ICSW advocacy efforts. The shift away from universal, publicly provided social protection schemes to needs-tested targeted assistance, which became evident with spread of neoliberal economic policies, encouraged by the international financial institutions has been extremely painful for many societies. ICSW efforts have sought to draw attention to the existence of positive alternatives to targeting, and the detrimental dimensions for well-being of the 'race to the bottom' resulting from cuts in social spending; and advocate consistently for the principles of the universal public provision of basic social services and basic income. Protecting people against contingencies and risks, along with developmental and social justice functions of social protection are well known (Department of Economic and Social Affairs. Report on the World Social Situation 2001. United Nations, New York, 2001, p. 217.) and have always been underscored by the ICSW in its policy statements.

The new impetus to these efforts was provided when Recommendation 202 on national floors of social protection was adopted by the International Labour Conference at its 101st session in 2012. There is increasing recognition of the fact

that the implementation of nationally defined social protection floors (SPFs) may be regarded as an important tool in the fight against poverty and inequality. A rights-based approach aimed at protecting women, men, and children from risks and vulnerabilities helps to sustain their well-being throughout the life cycle, guaranteeing access to health care and other basic services, as well as at least a basic level of income security. The ICSW has supported that initiative since its inception by the UN in 2009. (See the ICSW Statement on Poverty eradication, Commission for Social Development, Fiftieth session.)

Joining forces with more than 60 other NGOs, the ICSW formed a Global Coalition for Social Protection Floors, which has an increasingly important role to play in shaping the debate, raising awareness, achieving a national consensus on policy priorities and monitoring progress of SPFs. But the coalition with NGO partners is not the only joint effort undertaken by the ICSW in connection with social protection promotion. Among a limited number of other international NGOs, the ICSW has been invited to take part in the work of the Social Protection Inter-Agency Cooperation Board chaired jointly by the ILO and the World Bank. The Board has been instrumental in considering social protection in the context of the deliberations on the post-2015 development agenda when the existing MDGs expire, interagency work at the country level, cooperation in the field of social protection statistics, and other key issues linked to social protection promotion.

Another important partnership was formed by the ICSW at the UN with civil society organizations working to promote the rights of older persons. The ICSW has been closely collaborating with the Global Alliance for the Rights of Older People, voicing its support for the elaboration of a new legal instrument to protect the rights of older persons. Despite its undeniable importance and wide-ranging socioeconomic consequences, population aging is one of the issues that has not been addressed in the framework of the MDGs. In that light, the role of the Open-Ended Working Group on Ageing (The working sessions of the OEWG were convened twice in 2011, and once in 2012 and in 2013.) established by the UN General Assembly in December 2010 to consider the existing international framework for the human rights of older persons and to identify possible gaps and how best to address them, including by considering the feasibility of further instruments and measures, is the most important. Its significance as a new and influential international forum on aging and development should be recognized, particularly given the fact that aging presents a challenge not only for the developed countries but also for the developing world. Since the inception of the Working Group, the ICSW has actively participated in the discussions, which have covered numerous political and socioeconomic aspects of population aging at the national, regional, and international levels, highlighting both the emerging issues and the opportunities presented by aging, but always dovetailing around its original declared objective – how to protect better the rights and interests of older persons.

Participating in the deliberations, the ICSW has not only submitted official statements (See ICSW statements to the Open Ended Working Group at the third and fourth sessions in 2012 and 2013, respectively: http://social.un.org/ageing-working-group/.) but also usually provided ad hoc comments during the discussions, making specific suggestions on how to move the process forward for the benefit of older persons. The importance of clarifying the scope and content of human rights norms has been stressed, given the fact that non-implementation of the existing standards is rampant, even though in many cases and particularly in developing countries those standards are themselves quite low, remaining inadequate for the effective protection of older persons. (See, for example, OHCHR. Normative standards in international human rights law in relation to older persons. Analytical outcome paper. August 2012.)

Apart from ties built within the UN environment, the ICSW has also built a strong collaboration and partnerships with sister organizations that have the same historical roots, namely, with the IASSW and IFSW. The three organizations work together to produce the *International Social Work* journal, and collaborate in organizing joint world conferences on social work and social development issues every 2 years. The conferences provide ample opportunities for the stakeholders to discuss issues that the social work and social welfare sectors face on everyday basis and to ensure a closer link between evidence-based practice, policy objectives, and social development goals. Furthermore, the joint global conferences have revealed how actions geared on accomplishing the priority goals of the joint Global Agenda adopted by the ICSW together with IASSW and IFSW might impact on the conditions of people and demonstrate how actions by the social work and social development community can contribute to an enduring physical environment, promote compliance with a range of climate-related policies and measures, and help raise awareness and build capacity among practitioners regarding the imperatives of sustainable development.

Coming together in Paris in mid-November 2012 to consider specific actions in the context of the implementation of the Global Agenda, the three partner organizations decided to establish a Global Agenda Observatory to gather evidence about the activities of social workers, educators, and social development practitioners who support the implementation of the Global Agenda and to give visibility and credibility to their contributions and promote further action. This new entity is structured around the four themes of the Global Agenda commencing with the theme of promoting social and economic equalities. The themes to be covered in subsequent years include promoting the dignity and worth of peoples, working toward environmental sustainability, and strengthening recognition of the importance of human relationships. A focus on ensuring an appropriate environment for practice and education will be maintained throughout. (The Global Agenda Observatory is conceived as a mechanism for monitoring and reporting upon the implementation of the Agenda Commitments and will consist of networks or consortia composed of institutions of higher education and professional/practice-based organizations that jointly conduct research, analyze, synthesize, and report on Agenda activities. It also involves a process that will evolve over time. It aspires to establish a strong, credible monitoring mechanism over the next 10 years. The Global Agenda Observatory will be established at the local, national, regional, and international levels, working through the existing structures and work plans, taking into consideration the existing priorities of each of our organizations.)

Regional and National Activities

At the regional level, the ICSW seeks to increase the consideration of social development by regional groupings of governments; at the same time, at the national level, the ICSW works to strengthen the ability of national councils to influence social policy and programs in their respective countries, taking into account local conditions and priorities of national socioeconomic development.

Activities covered by the national councils are wide-ranging, but within the remit of the ICSW Global Programme. For example, in 2012–2013, regional activities of the ICSW-affiliated organizations in Latin America concentrated on the affordability of social protection and the feasibility of the SPF initiative, exploring the new tools and policy measures. In Asia, activities covered issues of applicability and affordability of social protection schemes to the specific regions (South Asia), highlighted the importance of partnerships and collaboration among governments and NGOs on social welfare matters (Southeast Asia and the Pacific), and explored challenges of aging and long-term care (Northeast Asia).

ICSW Europe has convened seminars to explore the situation of the working poor, people who have been pushed into the lower end of the job market. Their plight represents a growing concern in countries at all levels of development, including economies in transition and developed-market economies.

In June 2013, the ICSW in the Middle East and North Africa Region organized a high-level forum on the importance of national civil dialogue, where issues of social inclusion and participation were discussed, along with finding ways and means to ensure that the 'voices' of various groups in society are heard, giving real substance to the concept of citizenship and democracy. The event was extensively covered by the regional press and other mass media (Le Matin, Rabat, 13 June 2013). In the African region, ICSW activities focused on capacity building, promoting partnerships, and addressing resource mobilization issues. (One of the most innovative projects in the context of promoting North–South partnerships is the joint project between the Tanzanian Council for Social Development and the Finnish Federation for Social Welfare and Health. (recently renamed SOSTE), which aims at regional capacity-building programs to strengthen national councils to be effective advocates for social development and poverty reduction.)

The ICSW training program to strengthen national councils is an integral part of its ongoing work, representing a combination of bottom-up and top-down efforts. Apart from capacity development, training sessions organized by the ICSW provide opportunities to strengthen existing links with the academic community and to build new knowledge-based networks.

Conclusion

As the ICSW looks back over its past 85 years of history, the organization recognizes not only its achievements, but also the multiple challenges. The number of NGOs dealing with social development issues has mushroomed during the last 20 years. For the ICSW, retaining its influence and finding a niche in the new international environment requires much professionalism and dedication, and a constant quest of the meaningful, innovative, and cost-effective policies.

See also: International Association of Schools of Social Work (IASSW).

Bibliography

Eilers, Kerstin, 2003. Social policy and social work in 1928. The First International Conference of Social Work in Paris Takes Stock. In: Hering, Sabine, Waaldijk, Berteke (Eds.), History of Social Work in Europe (1900–1960): Female Pioneers and Their Influence on the Development of International Social Organizations. Leske + Budrich, Opladen, p. 120.

Gough, Ian, 2004. Welfare regimes in development context. In: Gough, Ian, Wood, Geof, Barrientos, Armando, Bevan, Philippa, Davis, Peter, Room, Graham (Eds.), Insecurity and Welfare Regimes in Asia, Africa and Latin America. Social Policy in Development Contexts. Cambridge University Press, p. 22.

Hulme, David, December 2007. The Making of the Millennium Development Goals: Human Development Meets Results-Based Management in an Imperfect World. The University of Manchester, Brook World Poverty Institute, p. 2.

Le Conseil international d'action sociale (ICSW), 2008. Quatre-vingts ans D'histoire (1928–2008). Presses de l'Ecole des hautes Etudes en santé publique, Rennes, p. 14.

Macdonald, John, 1975. The International Council on Social Welfare. Yesterday, Today and Tomorrow, a Short History. ICSW, New York, p. 7.

International Issues in Child Protection

Jonathan Parker, Bournemouth University, Bournemouth, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines single shared assessments in social work practice, promoted as a means of reducing duplication, cutting bureaucracy, providing a single lead point across services for assessment, and addressing the needs of people who use human services. The development of the single shared assessment process and its location with social work assessment, policy, and legislation in the United Kingdom is described and the technology is critiqued.

Introduction

This article explores the developing concept and practice of 'single shared assessment (SSA)' or 'Single Assessment Process (SAP)' as it is emerging within social work, particularly in the United Kingdom, but akin to other models of integrated service delivery such as that in Quebec, Canada (Hébert and Veil, 2004). The SSA primarily relates to assessments that are shared across various agencies. Social work processes are adaptive, moldable constructs, and the assessment and interventive practices that are currently employed not only differ within the separate UK administrations, as well as internationally, but also are dynamic and changing as sociopolitical developments and social work, as it is locally embodied and enacted, interact to build received practices.

The article will provide an overview of assessment in social work practice, locating it as a central element globally of the social work process. It will identify some of the types and meanings associated with assessment and explore the political rationale that drives much of it, and of which social workers should be aware if they are to act alongside, with and on behalf of people who use their services.

Social work's central role in assessment will be highlighted but this will be problematized as government policy seeks to create human profession isomorphs in conducting assessment that may leave out the unique contribution of social work to enhancing people's life chances and choices. However, these conflations have credence within policy and guidance and the evolution of SSAs and their deployment will be explored and critiqued as both recognition of the centrality of the person-in-environment as expert and the isomorphic tendencies of marketized human services bureaucracies.

Social Work Assessment in Context

The notion that assessment is a key task in social work practice has long been recognized (Hollis, 1964; Pincus and Minahan, 1973) and yet there has been very few attempts to construct or provide a systematic conceptual and theoretical base for assessment (Clifford, 1998) with an increasing reliance on instrumental and functional approaches (Martin, 2010).

There may be agreement that assessment is central to social work; there is dissension on what assessment actually means, how it should be conducted, who should do what with it, and with whom it should be done. These reflect social, political, organizational, and practice-based questions that have, over time, assumed a more overt focus on 'risk' and its identification, aversion, acceptance, or mitigation. Perhaps this, in part, results from the desire to move away from the traditional, essentialist approach promoted by Mary Richmond who borrowed scientific language and method in her *Social Diagnosis* (1917). Also, the rejection of positivist thought in social sciences in favor of fluidity and uncertainty (Clifford, 1998) has continued the debate. The purposes of assessment are as varied as the methods designed for undertaking them. While a systematic theory of assessment has not been produced, it could be argued that the dynamic and fragmented interactions of human lives are better served by a theoretical orientation that allows difference and change rather than an attempt at rigorous definition (Milner and O'Byrne, 2009).

Risk has always been part of social work assessment as it has been part of the social work process itself; whether assessing the risk of the 'undeserving' gaining charitable alms, the risk of significant harm occurring to a child, or the risk a person with mental health problems presents of injuring themselves or others. However, risk has become 'concretized' as an entity that is believed to be subject, potentially, to accurate calculation and, thereby, mitigation. Risk assessments have come to represent regulatory technologies of practice, working to protect the professionals rather than service users (Dominelli, 2004). It is something with which social workers must grapple, while understanding its temporally fluid, socially constructed status, and recognizing some of the tensions arising from a risk focus that may work against the wishes of those with whom social workers engage from child protection to adult social work (Littlechild, 2008; Parker, 2013). A paradox of injustice may, indeed, arise from uncritical application of risk assessment in social work (Reamer, 2001).

Many models are linear, prescriptive, and stylized in a way that failed to meet the needs of contemporary social workers (Milner and O'Byrne, 2009). Examples abound within government frameworks often sporting large number of questions and areas to undertake an assessment. The lack of theoretical base may be construed as a weakness in social work highlighting the focus on instrumentality and

a disjuncture between social workers' understanding of complexity and social policy and guidance promoting normativity. Focusing on an holistic assessment of the person-in-environment challenges such instrumentality allowing concerns for social justice to emerge (Healy, 2005), something that Dominelli (2012) has extended.

Part of this definitional debate concerns the perceived dichotomy between social work as an 'art' or as a 'science.' The debate is well rehearsed in the literature (England, 1986; Parker and Bradley, 2010). A measured approach would suggest that social work is both an art and a science since it integrates skills and 'professional' knowledge in its application.

Assessment is popularly understood in terms of judging or valuing the worth of something. This is indicative of an objective activity. It implies the use of standards against which something can be appraised. This certainly appears to be the case in many social work assessments. However, it leaves out the interactive nature and consideration of the person assessed, which also feature in social work assessments (Wilson et al., 2011).

Assessment in social work is also firmly rooted in the context of social work values calling for an explicit acknowledgment of their influence. Social work assessments, therefore, combine the judgment of something with the explicit acknowledgment of the importance of values, diversity, and the views of others. One way of examining assessment in social work is to separate the 'types' of assessment used.

Social work assessments can be separated into two basic types. First, as an ongoing, fluid, and dynamic process (Coulshed and Orme, 2006; Parker and Bradley, 2010). This develops the 'what and how' that needs to be done. It is something that continues over time and aligns with social work as an integrated process (Compton and Gallaway, 1975; Dominelli, 2004). Assessments may also be evaluative and reflexive continuing through the process of social work (Hardiker and Everitt, 1996) and specifically when working with sexually abusive young people (Hackett, 2004). Bartlett and Saunders (1970) consider the assessment process to represent an intellectual exercise that, aligned with professional judgment, can promote holism and social justice (Healy, 2005). However, social workers often do not continue the assessment throughout a planned piece of work. In fact, UK community care legislation and practice may preclude such a continuous process because of the separation of assessment as a discrete activity. Assessments in child protection may also be undertaken as a time-specific formulation that, when employed instrumentally, focuses on function rather than on individuals and acts to protect social workers rather than working on behalf of service users (Lloyd and Taylor, 1995; White et al., 2010). This has ramifications for the ways in which SSAs are both construed and conducted.

This latter kind of assessment suggests norms and standards. It offers a way of examining a situation, event, or individuals at a specific period in their lives and making predictions or identifying needs, goals, and ways of achieving these. Time-specific assessments, at times, offer a baseline, from which change can be continually monitored. It does not, however, necessarily give the assessor the depth of information they need, and it may reflect the values of those making or commissioning the assessment rather than those being assessed.

The levels at which assessments are conducted may be broad-based and holistic or focused on a particular issue serving to inform a particular intervention. These levels do not preclude one another. It may be appropriate in many circumstances to undertake a broad social assessment prior to focusing on an agreed goal and target for intervention that may demand a much more specific focus and activity. In practice, social work assessments often not only reflect different levels throughout the intervention but contain elements of both types identified above. Being explicit about the type used and the implications of this for the clients or service users is important (Fook, 2012).

It is important to acknowledge the differential and varied impact that legislation, policy and procedure, and the theoretical orientation, beliefs, and values of practitioners have on the practice of social work assessments (Milner and O'Byrne, 2009; Parker and Bradley, 2010). It is also imperative that workers be aware of and check their individual bias and how this may well affect the assessment undertaken. A degree of reflexivity, reflecting on how one reflects as a social worker, allows social workers to question practice and outcome in assessment, allowing challenges to received wisdom (Taylor and White, 2001).

Social work assessment serves to reach initial conclusions that describe, explain, predict, evaluate, and prescribe interventive methods. Thus, it has a purposive element. Traditionally, assessments have followed resource-led pathways rather than the broader needs-led approach. The needs-led approach separates out ends and means and is not meant to be dependent upon available resources. However, Coulshed and Orme (2006) indicate that needs-led principles cannot work unless there are plenty of resources and that in social work there are likely to be situations where demand exceeds supply.

A social work assessment is a focused collation, analysis, and synthesis of relevant collected data pertaining to the presenting problem, identified needs, and a pertinent to agency function both prior to and during planned and purposeful intervention (Parker and Bradley, 2010). All social work assessments, therefore, run the risk of inducing normative behavior and need to be 'troubled.' It is important that assessment is not seen simply as an activity, skill, or practice that can be undertaken in a linear fashion without recognition of theoretical and philosophical underpinnings and the importance of working together with those who use services (Hepworth et al., 2009). Grouping the ways in which assessment can be understood can help social workers to develop a questioning, critically reflective appreciation of the meanings and possible impacts assessment can have on individuals. The following model groups assessment around prescribed and political approaches, 'tribal' allegiances fostered by theoretical ideologies, and processes or rituals involved in the 'dance' or interrelational conduct of assessment.

1. Frameworks and regulatory frameworks – (political assessment), for example, Common Assessment Frameworks, the SSA process (see Turney et al., 2011)

2. Theoretically influenced assessment – (tribal assessments), for example (Payne, 2005; Milner and O'Byrne, 2009),
   a. Evidence-based practice
   b. Psychodynamic assessments
   c. Cognitive-behavioral assessments
   d. Person-centered assessments
   e. Community assessments
3. Procedural assessment – (ritual – liturgical or prescriptive)

Understanding and critiquing the assessments employed by agencies and organizations and the ways in which these are conducted (Smale et al., 1993; Preston-Shoot, 2012)

4. Interpersonal assessment – (ritual – relational)

Critical reflection on the interpersonal processes involved in assessment (Ruch, 2013) – data choice and gathering; data analysis and interpretation; the impacts of relationship.

The SSA model can align to all the above types. While politically determined, prescribed, and driven, it is possible for individual practitioners to engage with individuals, to recognize theoretical and personal biases and to engage with others rather than 'apply' an assessment.

What People Who Use Social Work Services Want and Say

User voice and involvement in the process of assessment has influenced the development of new approaches that recognize the importance of the participant, moving them from a passive subject status to an active participative one (Doel and Best, 2008). This, in part, has led to the development and promotion of SSA processes.

Early work concerning the user voice highlights two particular concerns; first, the quality of the relationship with the practitioner and, second, the effectiveness of the services received (Mayer and Timms, 1970). To achieve these requires planning. Planning involves making a comprehensive and careful assessment formulating goals and developing a sequential program aimed at achieving these goals while choosing methods and establishing criteria against which the outcome can be evaluated. However, in the task-oriented approach to inclusion that this may foster, it is central to maintain a clear emphasis on relationship (Ruch, 2013).

With variations in emphasis, the consumerist approaches to involvement from the political Right, to the activist approaches of the political Left, there has been a focus on aspects of clarity, honesty, openness, accountability, and efficacy in the relationship. It is the latter, however, which permeates the debate. This is taken up by Ruch (2013) as an important aspect of any assessment. If the SAP or SSA is to work, it is crucial that a constructive working relationship is formed, and important that this human interaction in the complex arena of assessment is allowed to mitigate the normative effects of instrumentality and standardization in the process.

There is an element of control involved in assessments especially when these are carried out by statutory social services agencies. Not only do they control access to services but also tend to control, exercising power over the disadvantaged sections of society (Ney et al., 2013). It is essential, therefore, to be accurate in factual content, and to recognize subjectivities as these may conceal normative values applied through the assessment process. In this regard, systematic and multiple source data collection methods help. It may be construed that SSAs could assist here, as they will triangulate interpretation from different disciplinary perspectives.

Smale et al.'s (1993) tripartite model of assessment offers a perspective on the ways in which SSA may be undertaken: as a 'questioning model' that emphasizes the expertise of the social worker; a 'procedural model' that emphasizes agency function and may employ checklists that demand little professional judgment or critical thinking; or an 'exchange model' that recognizes that people are experts on their own problems. The emphasis is on an exchange of information and mobilizing individual strengths to meet agreed goals.

Multidisciplinary assessments were developed as a potential safeguard in community care, the purpose of which is to create individual and flexible packages of care from a range of service providers, agencies, and groups. It is partly these laudable aims that led to the concept of SSAs across many countries. This process should be undertaken in conjunction with clients, seeing them as 'experts' in respect of their own situations (Preston-Shoot, 2007).

The Single/Shared Assessment Process

It is important to understand the context in which social work assessment has developed as it provides the sites in which SSA has developed, such as the search for inclusion of the user voice, multidisciplinarity, a reduction of the burden of telling the story, and recognition of the complexity of human situations. It also provides a significant role for social workers as lead professionals who coordinate assessment acting as the central point of contact for service users and professionals alike. However, the concept(s) of SSAs are not quite as they seem and require a degree of critique, something recognized early by Glasby (2004). In this article, examples from the United Kingdom will be used to illustrate potential learning for social work globally.

The New Labour Government in the United Kingdom made partnership working a key priority from 1997 onward, despite the undeveloped evidence base in its favor (Miller and Cameron, 2011). The development of an integrated comprehensive assessment; a SSA; or Single Assessment Process/Tool has been part of the overall vision for health and social care services in the four administrations of the United Kingdom since 2001 at least. Challis et al. (2010), however, trace concern for comprehensive, holistic assessments to the heady age of the birth of the Welfare State in Britain (Warren, 1946). Terminology remains contested and to an extent confused and differentiated, reflecting the different levels adopted in pursuit of the process. In this article, the various terms will be used interchangeably reflecting country preferences, but there will be inclination toward SSA and SAP.

A key aim of the SAP, in the United Kingdom, was to standardize the assessment processes for older people among professionals working in health and social care. An underlying principle was also to promote independence for older people – a principle that was pronounced in key policy documents (Department of Health, 2000, 2001). This approach has been challenged to an extent by older people themselves whose experiences of the SAP indicate a need for interdependence recognizing the limits to self-responsibility (Powell et al., 2007), and hinting, no doubt, at the unspoken politicoeconomic agendas underpinning pursuit of independence. There have been delays in operationalizing the process and the SAP, although designed to be implemented by April 2004, has been integrated in a patchy and nonstandardized way nationally (Tucker et al., 2007). There also remains division between health and social care agencies based on cultural barriers and different funding streams (Miller and Cameron, 2011).

Assessment in adult social care in the United Kingdom is located in a system of care management. Concerns for escalating costs and belief in the efficiency and appropriateness of care within the person's own home led, following the Griffiths Report (1988) and White Paper Caring for People (Department of Health, 1989), to legislation in England, the National Health Service and Community Care Act 1990, and subsequent guidance which, in turn, heralded the care management system.

In part, the introduction of the SAP, was a government response to earlier criticisms of multidisciplinary working within the context of care management (Manthorpe et al., 1996). The aim of the SAP was to "promote better care services and better outcomes for older people, and more effective use of professional resources." Particularly, the SAP "should ensure that the scale and depth of assessment is kept in proportion to older people's needs, agencies do not duplicate each other's assessments, and professionals contribute to assessments in the most effective way" (Department of Health, 2002: p. 1). SSA models have been introduced elsewhere, for instance, the Australian Community Care Needs Assessment of 2007. However, there is greater emphasis on the importance of multidisciplinarity in the Australian model. While pioneers in favor of the development of a comprehensive assessment process promoted multidisciplinary and integrated assessments as faster and less costly (Øvretveit, 1993), the ongoing divisions between health and social care and chauvinistic protection of budgets derailed the smooth implementation of the process.

The original intentions indicated that local stakeholders, including older people, service users, carers, providers, other local statutory agencies such as housing, and local voluntary organizations should be involved in drawing plans for the local implementation of the SAP.

The Scottish Executive (2001) sought to implement SSA in Scotland by April 2002, echoing aims expressed by the English administration and building on guidance from the Scottish Joint Futures Group and community care guidance (SWSG 11/91, 1992; SWSG 10/98, 1998). The aim to reduce bureaucracy and to develop and work with shared protocols was again explicit. Assessment should be shared and led by one single professional, drafting in specialist involvement as appropriate. New legislation was not envisaged in Scotland for this process and the duty to assess under s. 12a of the Social Work (Scotland) Act 1968 remained.

The explicit principles on which the Scottish model was developed also demonstrated alignment with human service values. The model reflects an ideal and demands recognition of needs that set some of these principles in tension such as the development of a shared and agreed assessment tool that necessarily channels and limits the process; an implementation process facilitating joint working across health, social work, and housing; the development of minimum expectations and standards, which again are not overtly recognized as having sociopolitical, economic, and cultural elements attached. It is unlikely that a process could ever be developed that did not have such underpinnings but the necessity of bringing these to the surface and making them visible is something that was not undertaken, thus reducing the impact and effectiveness of reach of the SSA, and potentially leading to instrumental completion of assessments rather than the comprehensiveness desired.

When producing initial guidance on the SAP, the Department of Health in England was not envisaging radical change. Much of the required infrastructure was already in place within care management. In preparation for the implementation of the SAP in April 2004, nominal guidance, therefore, was issued (Department of Health, 2004) that principally reinforced and upheld earlier guidance providing a framework for a successful implementation. It was also emphasized that a key to a consistent approach was the way in which assessment information, subject to consent of the service user, was stored and shared, using the Single Assessment Summary or local equivalent as described in Annex I (Department of Health, 2002). This Annex was modified in November 2003. The need for appropriate training of professionals to enable them to carry out person-centered assessments and care planning (see Annex J, Department of Health, 2002) was reinforced and examples of good practice in the early implementation of the SAP were cited in a range of local authorities (Department of Health, 2004).

The UK government appeared confident that work on the SAP had been well developed and that initiatives were underway to build on this and develop a Common Assessment Framework 'to ensure less duplication across different agencies and allow people to self assess where possible' (Department of Health, 2006: p. 115). However, in practice, the SAP is still 'bedding in' and there has been a level of drift in its implementation that may be a reflection on the priority given to this process at national and local levels. Indeed, it was the government's intention that the implementation of SAP would be cost neutral and this has been met with a healthy skepticism (McNally et al., 2003). Anecdotal information indicates that practitioners are finding that the contact and overview assessments have a tendency to promote 'social work by numbers.' The forms are seen as extensive and health dominated and focus on a 'need to know' basis rather than a holistic assessment. The first professional person who visits the service user in his or her home is normally required to fill in the appropriate sections, but this does not always happen. Practitioners are wary of disclosing sensitive information on the service user (for example, financial information) that is usually left in a brightly colored

folder in a conspicuous place in the home. There is also the challenge created by professional differences and the meanings given to different terminology, for example, that of care planning (Parker and Bradley, 2010). Progress has also been hampered by the fact that not all organizations have a shared database and/or compatible electronic systems. On the other hand, local authority care managers can see the benefit of sharing assessments with allied professionals; in the old system the task normally fell to them and they were not always the best person to do this work.

Reflections on the Single Shared Assessment Process

From early days, research has exemplified continuing problems with implementing an SSA process (Glasby, 2004). Clare and Cox (2003) opine that the emphasis on inclusion, while laudably reflecting the increasing presence of the voices of those using human services may not extend as far as people with complex needs, who may well appear too difficult to involve meaningfully and therefore may remain excluded. Taylor (2003) further suggested that a shortfall in both nursing and social work practitioners limited the efficacy of the SSA/SAP in providing appropriate community care services that meet needs at all levels of complexity. It is perhaps to be expected that ideologically motivated models often driven by political aims do not work as well as assumed in complex human situations. Social workers must remain aware of the hidden agendas of neoliberalism that concern cost-effectiveness and efficiency but work through such assumed social goods as standardization, multidisciplinary approaches, joined-up foci, and reduced bureaucracy. A critically reflexive stance is important.

Challis et al. (2010), writing about England, reported on a 2005–06 survey concerning comprehensive assessment patterns and their prevalence, including multidisciplinary models. The survey suggested that comprehensive assessments relating to older people with complex needs are hampered by a lack of multidisciplinary experts, and that the intended outcomes of the SAP in England – to develop systematic multidisciplinary services – was not working, and, indeed, would not become effective without further policy directives.

Related to Challis et al. (2010) and Abendstern et al. (2011) explored the levels and nature of integrated or shared assessments across health and social care in England after the implementation of the SAP. She and her colleagues considered the premise that the provision of timely and effective services required health and social care agencies to work closely together in conducting assessments. Of those responding to the survey, a broad commitment to integrated assessment was found, but involvement in the SAP and its implementation was patchy. The survey of lead officers for the SAP boasted an 82% response rate ($n = 122$), with almost 72% being local authority personnel. While promoted as integrated multidisciplinary assessment processes, it was social workers who most frequently engaged as lead professional when conducting SSAs, followed by district nurses and occupational therapists. Key findings indicated that the majority of areas surveyed had agreements in place to support integrated assessments, but the use of SAP data appeared more useful for determining intermediate care and care management as opposed to access to health funded continuing care or mental health services. Research by Powell et al. (2007), concerning the views of older people experiencing the SAP, confirms the potential for identifying 'low-level' need, but not as effective where formal, complex care may be needed. The 26 older people interviewed by Powell et al. reported that the process does not assist when there are resource constraints, illustrating the tensions between identified needs and limited resources. They also highlighted the centrality of social relationships in respect of health and well-being, which extends to relationships between professionals.

A search of Web-based local government policies on single shared assessments shows the drive toward standardization as an unquestioned good; a desire for mimetic practices that suggest normative social work best meets people's needs. The policies developing do not necessarily reflect research-led and evidence-informed approaches but desire for legitimacy, not seeming out-of-step, or being open to challenge. The dangers inherent in this for social workers include being funneled into practices predicated on assumed 'evidence' and coercive structures (the policy guidance and legislative underpinnings) that require critique if social workers are to maintain their stand together with the marginalized and disenfranchised within and outside of society. From December 2008 to March 2009, Miller and Cameron (2011) undertook a comprehensive, although not systematic, review of interagency assessment, including 23 papers. Their review included earlier systematic reviews of evidence and literature reviews (Abenderstern et al., 2008; Challis et al., 2008). They found similar policy drivers and the introduction of eligibility criteria for services in the four administrations of the United Kingdom. They also found common difficulties in devising assessment documentation and balancing the need for free text with tick-box approaches, and a resulting lack of standardization in the process that makes comparison and evaluation problematic to undertake. However, their review also identified positive features of the SSA process in managing risks, reducing duplication, improving standards in practice and recording, joint working and communication, and involving service users.

Undertaking any social work assessment demands two core skills:

1. administration, and
2. human relationship or interpersonal skills (Coulshed and Orme, 2006).

Alongside demands for an ability to organize, systematize, and rationalize assessment knowledge, there is a concomitant need reflected within the SSA process for humane sensitivity and a valuing of the uniqueness of each individual involved. Whether this person-centered approach is realized depends on the integrity and approach of the social work, rather than the standardized approaches determined, although the levels of support and concern for social workers performing the assessments are also important to the outcome. SSA or the SAP represents an attempt to humanize social work and health care services but this cannot be achieved by and of

itself. Social workers have a role in developing a reflexive and questioning approach that itself aligns well with the assessment task, and at times must challenge the assumed yet unsubstantiated positives associated with SSA if state, organization, or professional needs appear to be privileged above the needs of those individuals and communities whom social workers serve.

See also: Child Neglect; Child Protection; Elder Abuse; Social Work Theory.

Bibliography

Abenderstern, M., Clarkson, P., Challis, D., Hughes, J., Sutcliffe, C., 2008. Implementing the Single Assessment Process for older people in England: lessons from the literature. Research, Policy and Planning 26 (1), 15–32.

Abendstern, M., Hughes, J., Clarkson, P., Sutcliffe, C., Challis, C., 2011. The pursuit of integration in the assessment of older people with health and social care needs. British Journal of Social Work 41 (3), 467–485.

Bartlett, H., Saunders, B., 1970. The Common Base of Social Work Practice. National Association of Social Workers, Washington.

Challis, D., Clarkson, P., Hughes, J., Abendstern, M., Sutcliffe, C., 2008. A systematic evaluation of the development and impact of the single assessment process in England. Research and Policy Update 2. Manchester University: Personal Social Services Research Unit.

Challis, D., Abendstern, M., Clarkson, P., Hughes, J., Sutcliffe, C., 2010. Comprehensive assessment of older people with complex care needs: the multidisciplinarity of the Single Assessment Process in England. Ageing and Society 30 (7), 1115–1134.

Clare, L., Cox, S., 2003. Improving service approaches and outcomes for people with complex needs through consultation and involvement. Disability and Society 18 (7), 935–953.

Clifford, D., 1998. Social Assessment Theory and Practice: A Multidisciplinary Framework. Ashgate, Aldershot.

Compton, B., Gallaway, B., 1975. Social Work Processes. Dorsey Press, Homewood, IL.

Coulshed, V., Orme, J., 2006. Social Work Practice: An Introduction, fourth ed. Palgrave MacMillan, Basingstoke.

Department of Health, 1989. Caring for People: Community Care in the Next Decade and beyond, (Cm 849). HMSO, London.

Department of Health, 2000. The NHS Plan: A Plan for Investment. A Plan for Reform. Department of Health, London.

Department of Health, 2001. National Service Framework for Older People. The Stationery Office, London.

Department of Health, 2002. The Single Assessment Process Guidance for Local Implementation and Annexes (HSC 2002/001; LAC2002/1). www.dh.gov.uk/PolicyandGuidance/HealthandSocialCareTopics (accessed 01.12.12.).

Department of Health, 2004. Single Assessment Process for Older People. Requirements for April 2004. Department of Health, London.

Department of Health and Ageing, 2007. The Australian Community Care Needs Assessment: Synopsis of Field Test Report and Eligibility and Assessment Working Group Response. Australian Government Department of Health and Ageing, Canberra.

Doel, M., Best, L., 2008. Experiencing Social Work: Learning from Service Users. Sage, London.

Dominelli, L., 2004. Social Work: Theory and Practice for a Changing Profession. Polity Press, Cambridge.

Dominelli, L., 2012. Green Social Work: From Environmental Crises to Environmental Justice. Polity Press, Cambridge.

England, H., 1986. Social Work as Art: Making Sense for Good Practice. Allen and Unwin, London.

Everitt, A., Hardiker, P., 1996. Evaluating for Good Practice. Palgrave Macmillan, Basingstoke.

Fook, J., 2012. Social Work: A Critical Approach to Practice, second ed. Sage, London.

Glasby, J., 2004. Social services and the Single Assessment Process: early warning signs? Journal of Interprofessional Care 18 (2), 129–139.

Griffiths, R., 1988. Community Care: Agenda for Action, (The Griffiths Report). HMSO, London.

Hackett, S., 2004. What Works for Children and Young People with Harmful Sexual Behaviours. Barnado's, Barkingside.

Healy, K., 2005. Social Work Theories in Context: Creating Frameworks for Practice. Palgrave Macmillan, Basingstoke.

Hébert, R., Veil, A., 2004. Monitoring the degree of implementation of an integrated delivery system. International Journal of Integrated Care 4 (3), 1–11.

Hepworth, D., Rooney, R., Rooney, G., Strom-Gottfried, K., Larsen, J., 2009. Direct Social Work Practice: Theory and Skills, eighth ed. Brooks Cole, Belmont, CA.

Hollis, F., 1964. Casework: A Psychosocial Theory. Random House, New York.

Littlechild, B., 2008. Child protection social work: risks of fears and fears of risks – impossible tasks for impossible goals. Social Policy and Administration 42 (6), 662–675.

Lloyd, M., Taylor, C., 1995. From Hollis to the Orange Book: developing a holistic model of social work assessment in the 1990s. British Journal of Social Work 25 (6), 691–710.

Manthorpe, J., Stanley, N., Bradley, G., Alaszewski, A., 1996. Working together effectively? Assessing older people for community care services. Health Care in Later Life 1 (2), 143–155.

Martin, R., 2010. Social Work Assessment. Learning Matters/Sage, London.

Mayer, J., Timms, N., 1970. The Client Speaks: Working Class Impressions of Casework. Routledge and Kegan Paul, London.

McNally, D., Cornes, M., Clough, R., 2003. Implementing the single assessment process: driving change or expecting the impossible? Journal of Integrated Care 11 (2), 18–29.

Miller, E., Cameron, K., 2011. Challenges and benefits in implementing shared interagency assessment across the UK: a literature review. Journal of Interprofessional Care 25 (1), 39–45.

Milner, J., O'Byrne, P., 2009. Assessment in Social Work, third ed. Palgrave/Macmillan, Basingstoke.

Ney, T., Stoltz, J., Maloney, M., 2013. Voice, power and discourse: experiences of participants of family group conferences n the context of child protection. Journal of Social Work 13 (2), 184–202.

Øvretveit, J., 1993. Co-ordinating Community Care: Multidisciplinary Teams and Care Management. Open University Press, Buckingham.

Parker, J., 2013. Assessment, intervention and review. In: Davies, M. (Ed.), The Blackwell Companion to Social Work, fourth ed. Blackwell, Oxford, pp. 311–320.

Parker, J., Bradley, G., 2010. Social Work Practice: Assessment, Planning, Intervention and Review, third ed. Learning Matters/Sage, London.

Payne, M., 2005. Modern Social Work Theory, third ed. Macmillan, Basingstoke.

Pincus, A., Minahan, A., 1973. Social Work Practice: Model and Method. F.E.Peacock Publishers, Itasca, IL.

Powell, J., Robison, J., Roberts, H., Thomas, G., 2007. The Single Assessment Process in primary care: older people's accounts of the process. British Journal of Social Work 37 (6), 1043–1058.

Preston-Shoot, M., 2007. Whose lives and whose learning? Whose narratives and whose writing? Taking the next research and literature steps with experts by experience. Evidence and Policy 3 (3), 359.

Preston-Shoot, M., 2012. The secret curriculum. Ethics and Social Welfare 6 (1), 18–36.

Reamer, F., 2001. The Social Work Ethics Audit: A Risk Management Tool. National Association of Social Workers, Washington.

Ruch, G., 2013. Understanding contemporary social work: we need to talk about relationships. In: Parker, J., Doel, M. (Eds.), Professional Social Work. Learning Matters/Sage, London, pp. 46–59.

Smale, G., Tuson, G., Biehal, N., Marsh, P., 1993. Empowerment, Assessment, Care Management and the Skilled Worker. Macmillan, Basingstoke.

Taylor, A., August 21, 2003. Single assessment process rendered ineffective by lack of social workers. Community Care 1486, 6.

Taylor, C., White, S., 2001. Knowledge, truth and reflexivity: the problem of judgement in social work. Journal of Social Work 1 (1), 37–59.

The Scottish Executive, 2001. Guidance in Single Shared Assessment of Community Care Needs. http://www.sehd.scot.nhs.uk/publications/DC20011129CCD8single.pdf (accessed 01.12.12.).

Tucker, S., Baldwin, R., Hughes, J., Benbow, S., Barker, A., Burns, A., Challis, D., 2007. Old age mental health services in England: implementing the National Service Framework for Older People. International Journal of Geriatric Psychiatry 22 (3), 211–217.

Turney, D., Platt, D., Selwyn, J., Farmer, E., 2011. Social Work Assessment of Children in Need: What Do We Know? Messages from Research. Department for Education, London.

Warren, M., 1946. Care of the chronic aged sick. The Lancet 247 (6406), 841–843.

Wilson, K., Ruch, G., Lymbery, M., Cooper, A., 2011. Social Work: An Introduction to Contemporary Practice. Pearson Education, Harlow.

White, S., Wastell, D., Broadhurst, K., Hall, C., 2010. When policy o'erleaps itself: the 'tragic tale' of the Integrated Children's System. Critical Social Policy 30 (3), 405–421.

International Law and Treaties

Mark W Janis, University of Connecticut, Hartford, CT, USA; and University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

International law is the law generated by more than one state to regulate the international transactions of both states and private parties. It includes rules drawn from treaties, customary international law, general principles of law, natural law, and equity. It is applied in international diplomacy, international courts and arbitrations, and domestic courts and arbitral tribunals. Its fields include the laws of war and peace, the law of the sea, the laws of diplomatic practice, the law of treaties, the law of recognition and succession of states and governments, international trade and commerce, international human rights law, the law of international organizations, international environmental law, international business law, and international criminal law.

For thousands of years what we now call *international law and treaties* have provided legal rules and processes, generally generated by more than one state, to regulate the international transactions of both states and private parties. Modern *international law and treaties* are grounded in the Roman *jus gentium* as transformed by seventeenth-century Dutch theorist Hugo Grotius, who argued that the law of nations consisted both of positive law (in the explicit form of international agreement and the implicit form of international custom) and of natural law as drawn from reason, religion, and necessity. The field of *international law and treaties* is vast and growing. It includes traditional topics such as the laws of war, the law of the sea, the law of ambassadors and diplomacy, the law of treaties, the law of recognition and succession of states, and the law of trade and merchants. New topics include the law of international organizations, international human rights law, international business law, international tax law, international environmental law, and international criminal law. International legal rules are enforced not only by international courts and arbitration, but by domestic courts, and in diplomatic practice.

Definition

International law and treaties is a term that in English has at least three modern meanings: (1) legal rules and legal processes generated by more than one state; (2) legal rules and legal processes that concern interstate relations; and (3) legal rules and legal processes that regulate the international transactions of states and private parties.

Nature and History

Traces of what we now term *international law and treaties* are as old as written history. Evidence of compacts between states and rulers of states dates back at least 4000 years; they chronicle agreements about, inter alia, war and peace, foreign merchants, and ambassadors, all still lively topics of the discipline. It is safe to say that international law and treaties figure as necessary features of human society, regardless of any juristic definition, description, or argument.

The modern intellectual tradition relating to international law and treaties owes much to Roman law. The Roman term *jus gentium*, the law of nations, described legal rules pertaining generally to foreigners when specific foreign rules were unknown or in conflict. These were rules thought to be so basic that they were shared by all nations. Hence, like natural law, the law of nations was presumed to exist without positive expression in formal legislation. In Roman law, *jus gentium* ordinarily was employed in legal matters concerning private transactions, such as marriage, wills, and commerce.

Despite its origins in Roman private law, the law of nations was adopted by European jurists in early modern times to describe certain public legal relations. For example, in the sixteenth century, Spanish writers such as Suarez and Vitoria employed principles of *jus gentium* to argue that the Spanish Crown owed legal duties of basic humanitarian conduct to all peoples, even the indigenous peoples conquered by Spain in the Americas.

The often alleged 'father' of international law is Hugo Grotius (1583–1645), a Dutch jurist exiled to Paris following the suppression of the liberal party in the Netherlands. Already an accomplished diplomat, jurist, and theologian, Grotius devoted the remainder of his life to the cause of peace. In the midst of Europe's disastrous 30 Years War (1618–48), Grotius published the law of nations' most famous book, De Jure Belli Ac Pacis, to demonstrate that legal rules and processes defined and limited the sovereign rights and aggressions of otherwise bitterly divided Catholic and Protestant states and that such legal rules and processes even applied to relations between Christian and non-Christian states.

Grotius elaborated a legal theory that still serves as the foundation of modern international law. He argued that sovereign states were defined and bound by two kinds of legal rules drawn from the law of nations: positive law and natural law. The positive law of nations, expressed in treaties and custom, was made by the explicit and implicit agreement of sovereign states. The natural law of nations, expressed in the natural laws of reason and religion, was the result of necessary rules binding everyone, even rulers of states.

The first part of Grotius' theory of *jus gentium* was quickly married to emerging seventeenth-century political theory, like that of Hobbes, positing the idea of *sovereign* states. Grotius'

consensual theory of the law of nations helps explain why treaties, such as the Peace of Westphalia that terminated the 30 Years War in 1648, are legally binding: sovereign states have authority both to legislate rules internally for their own territory and to make rules externally for themselves through interstate agreement. Grotius' legal theory also explains why sovereign states should not interfere in the internal affairs of other sovereign states: a nation's sovereign authority must be limited so that the sovereign authority of other states is protected. Such international legal principles that weave sovereign states together and protect them from each other continue to guide international relations today.

In his age, Grotius' *jus gentium* excited many rulers, including Gustavus Adolphus of Sweden (who is said to have used Grotius' book as a pillow while campaigning on his many wars), as well as many universities. In Germany, The Netherlands, and England, professorial chairs in the new discipline of the law of nations were created, commencing an outpouring of scholarly works further describing and developing the discipline.

Not all students, of course, were pleased by their professors; one of the earliest pupils of the noted English law professor, William Blackstone, rebelled against what he was taught about the law of nations. In 1789, Jeremy Bentham (1748–1832) rejected the term the *law of nations* as professed by Blackstone; Bentham argued in *An Introduction to the Principles of Morals and Legislation* that the discipline should be renamed *international law*. Bentham's invented word, *international*, successfully passed from international law to denominate many other disciplines, for example, *international relations*, *international politics*, and *international economics*.

More or less inadvertently, Bentham changed not only a term but the very definition of the discipline. Grotius' and Blackstone's classical definition of the *law of nations* focuses on legal rules and processes generated by more than one state. Bentham's definition of *international law*, however, focuses on the subjects of the discipline; these Bentham said were states and states alone. The classic *jus gentium* regards individuals, as well as states, as its subjects, possessing legal rights and duties and accessing legal process. For the sake of theory, Bentham rejected individuals as proper subjects of *international law*.

A third definition of the discipline, along with a new term to describe it, was provided by the US jurist Philip Jessup, who in 1956 proposed replacing both the term the *law of nations* and the term *international law* with a new term, *transnational law*. Jessup defined the discipline of *transnational law* as encompassing any legal rule or process, municipal or international, that has to do with international transactions of any state, entity, or individual.

It may or may not make a difference how one defines the discipline. For example, a rule about delimitation of the continental shelf not only emerges from a multistate source, a treaty, *à la* Grotius' *law of nations*, but it also relates to interstate relations *à la* Bentham's *international law*, and deals with international transactions *à la* Jessup's *transnational law*. Hence, such a rule easily fits within the discipline as described by all three definitions. However, other times the different definitions lead to conflict. For example, if one believes, along with Bentham, that only states may be subjects of *international law*, then one cannot believe that individuals have international legal rights and duties. This means that they cannot be protected by international human rights law against abusive state action or be subject to international criminal law. A contrary position is usually taken by those who subscribe to the ordinary definitions of the *law of nations* and *transnational law*. Nowadays in practice, although the term *international law* is very popular, some of its exclusivities have been softened both by the classical tradition of the *law of nations* and by the modern sentiment expressed in *transnational law*.

Domain

The domain of *international law and treaties* is vast and growing. The discipline's classical fields, many hundreds of years old, include the *laws of war* (both about going to war, *jus ad bellum*, and about the proper conduct of war, *jus in bello*), the *law of the sea*, the *law of ambassadors and diplomacy*, the *law of treaties* (how treaties are made, interpreted, and implemented), the *law of the recognition and succession of states*, and the *law of merchants* (traditionally, *lex mercatoria* was part of the law of nations). More recently, new fields in the discipline have emerged, including the *law of international organizations and institutions*, *international human rights law*, *international trade law*, *international financial law*, the *international law of Antarctica*, the *law of outer space*, *international environmental law*, *international refugee and immigration law*, *international criminal law*, the *law of international arbitration*, *international tax law*, and the *international law of state responsibility*. The proliferation of fields in the discipline of international law is due in large measure to the increasing numbers and kinds of international relationships among states, peoples, corporations, and institutions, as well as to the inability of national law and regulation to adequately cope with such international transactions.

The Legislative and Enforcement Questions of International Law

To comprehend the fundamentals of modern international law, it is important to understand that two questions have always beset the discipline. First, how is it possible to authoritatively generate international legal rules? And second how is it possible to realistically enforce international law? In short, without an effective international political system, it seems at first glance that there is neither an authoritative way to make international law nor a realistic process to enforce it. These questions are answered in practice by remarkably effective systems of international legal rules and international legal process.

International Legal Rules

How are international legal rules authoritatively made? International lawyers often turn to Article 38 of the Statute of the International Court of Justice (the 'ICJ') to help describe the formation of international legal rules. Article 38 sets forth the rules available to the ICJ, the court of the United Nations, in deciding cases before it, albeit there are other kinds of

international legal rules as well. The three principal sorts of international legal rules available to the ICJ are rules drawn from explicit international agreements (mostly treaties), rules drawn from international custom, and rules drawn from general principles of law. These are sometimes described as formal sources of international law in that they are the sorts of rule-making that are acceptable in form as legitimate in international law (cf the *material sources* of international law which are the evidences that show the substance of a specific rule).

Rules drawn from international agreements include international legal rules set out in treaties, compacts, conventions, protocols, or whatever may be a written agreement between or among states governed by international law. Nowadays, there are tens of thousands of such instruments; their norms constitute the greatest number of international legal rules. These explicit international rules are comparable not only to statutes in municipal law (in that they establish rules for general conduct), but to domestic contracts (in that they can make ad hoc agreements based on party consent) and to constitutions (in that they can constitute international organizations). They can also be more or less aspirational. Treaty rules are the international legal rules that tend to be both least ambiguous and most clearly based on state consent. All other kinds of international legal rules are more troubled by problems both of ambiguity and of authority.

A second type of international legal rule is *customary international law*. Ordinarily, customary international law is perceived as based on the implicit agreement of states and is composed of two elements. First is a consistent pattern of state practice; this defines the rule. Second is a belief, an *opinio juris*, that compliant state behavior is based on a sense of binding legal obligation. In practice, the determination of a consistent pattern of state practice is almost always controversial and involves a subjective weighing of evidences from state practice proving and disproving the rule. *Opinio juris* is even more controversial: states rarely indicate that they are behaving in international relations in a certain way because they feel compelled to do so by a sense of international legal obligation. Rather, *opinio juris* is usually an evaluation added by a judge or a commentator as a description of state practice.

The third kind of legal rule prescribed by Article 38 of the ICJ Statute is *general principles of law*. These are legal principles generally shared by most or all legal systems. This concept is rather similar to the original Roman notion of *jus gentium*, that is, rules so basic that they are to be found everywhere. In modern practice, general principles of law are usually developed by jurists through a comparative law search. The method proceeds by examining diverse municipal legal systems. If a rule is found to be common to most or all of the municipal legal systems examined, then the rule is also presumed to exist as a general principle of law in international law.

Of the three formal sources of international law provided for in the ICJ Statute, treaty and custom are, more or less, consensual in nature. Even general principles of law can be seen to have a consensual aspect since it can be argued that if states already recognize a legal rule internally, they may be presumed to accept that self-same rule externally. Beyond the ICJ Statute, however, other formal sources of international legal rules exist without displaying such real or alleged consensual foundations.

Such nonconsensual rules may be seen as being no longer in the camp of positive law, but belonging to that of natural law. Indeed, natural law itself has long been an important formal source of international law. However, the term *natural law*, once so common, is nowadays rarely used in discussions of international law. Instead, newer equivalent terms have emerged. Key among these are *jus cogens* and *fundamental norm*, both expressing the natural law idea that some legal rules are so fundamental that they can be found everywhere and the idea that they trump any conflicting rule of positive law. *Jus cogens* or fundamental norms are often thought to include both the basic principles underpinning the international system, for example, *pacta sunt servanda* (agreements are binding) and the very notion of 'sovereignty,' as well as those basic principles protecting key human rights, for example, legal norms protecting individuals against genocide, slavery, and torture.

Another nonconsensual form of international law is *equity*. Sometimes viewed as a corrective of the strict application of a legal rule (equity *contra legem*), equity can also involve rule creation (equity *praeter legem*) and rule interpretation (equity *infra legem*). Equitable rules or principles can be based on nonlegal criteria, such as distributive justice and basic notions of fairness.

International Legal Process

How are international legal rules enforced in practice? Most times when international legal rules are recognized, they are applied by the nation states themselves. So, for example, all kinds of international legal rules, whether they be drawn from treaties, custom, general principles of law, *jus cogens*, or equity, are applied from time to time by municipal courts. Approaches to this process usually are thought to be either *dualistic* or *monistic*. *Dualism*, the prevalent view, holds that municipal legal systems and international legal systems are distinct from each other (dualist) and that there need to be rules of *incorporation* providing why, when, and how a rule of international law is to be applied by a municipal court. *Monism*, the less usual view, sees both international law and municipal law as parts of a unified system of law, where incorporation of international legal rules into municipal law is to be presumed.

Different municipal legal systems have different answers to questions of the incorporation of international law. Moreover, a municipal legal system's rules about incorporation may vary depending on the type of international legal rule concerned. So, for example, though treaty rules are sometimes deemed to be *self-executing* in the United States, that is, requiring no domestic legislation to put them into internal effect, in Great Britain treaty rules are never self-executing: all require an act of Parliament before they can be employed as rules of decisions by municipal courts. However, respecting customary international law, the United States and Great Britain share a similar approach. Both municipal legal systems deem some sorts of customary international law to be part of the *law of the land*, available to municipal courts without legislative act.

Besides providing rules of decision for municipal courts which then enforce the international rules alongside municipal law, international law is also enforced by procedures in the international legal system. There are, for example, an increasing

number of international courts. Oldest among these is the International Court of Justice at The Hague, which dates its antecedents to the Permanent Court of Arbitration established by The Hague Peace Conference of 1899, and to the Permanent Court of International Justice, established under the auspices of the League of Nations in 1921. The International Court of Justice was established alongside the United Nations, its parent institution, in 1945. The Court is composed of 15 judges elected by the UN General Assembly. There are usually ICJ judges from China, France, Russia, the United Kingdom, and the United States.

Reference has already been made to the ICJ Statute, the treaty that authorized the creation of the Court. The ICJ Statute also provides for two kinds of jurisdiction for the Court. First is its *contentious jurisdiction*; here the Court, with the prior consent of the parties to the dispute, actually decides cases between states. However, neither international organizations nor individuals may be party to the ICJ's contentious jurisdiction. The second sort of ICJ jurisdiction is its advisory jurisdiction: at the request of some international organizations (though neither states nor individuals), the ICJ may render a legally nonbinding advisory opinion on a legal matter. Since 1921, the practice of the Permanent Court of International Justice and the International Court of Justice has remained roughly steady: about three or four judgments or advisory opinions rendered each year.

Since 1950, a great many new international courts have been created. Each of these courts is limited either geographically or by its subject matter, but some have become a great deal busier than the older and more universal ICJ. Europe's two regional courts – the European Court of Human Rights in Strasbourg (established by the European Convention of Human Rights in 1953 and comprising 47 member countries) and the European Court of Justice in Luxembourg (the highest court of the European Union) – nowadays decide hundreds of cases each year. Less busy but still important are international courts and international dispute-settlement procedures such as the Law of the Sea Tribunal in Hamburg, the World Bank's International Center for the Settlement of Investment Disputes in Washington, the Inter-American Court of Human Rights in Costa Rica, the World Trade Organizations' dispute-settlement procedures in Geneva, the international criminal tribunals for the former Yugoslavia in The Hague and for Rwanda in Kenya, and the International Criminal Court in The Hague. Unlike the states-only provisions limiting the ICJ, all these newer international tribunals permit individuals in some circumstances to be parties to litigation.

In addition to the increasing number of international courts, there are a great many international arbitral procedures available to provide ad hoc jurisdiction to settle international legal disputes, public and private. Many of these international arbitrations are organized under the auspices of one of the international arbitral institutes. Two of the busiest are the International Chamber of Commerce in Paris and the American Arbitration Association in New York.

Finally, international law is often enforced simply by the diplomatic practice of governments. Reciprocity and good faith, outside any formal dispute settlement procedures, account for the lion's share of cases in which international law is effectively employed. It would be difficult for much of the world's business, public and private, to get done unless governments followed similar practices set out in treaties, custom, and the other kinds of international law.

Prognosis

It is likely that international law and treaties, already employing tens of thousands of international lawyers worldwide, will continue to prosper and expand. Modern trade, communications, and transportation have all been globalized much more thoroughly than modern politics, which remains largely state-centered. It is likely that without a political or social impetus for internationalization of political decision making, the sovereign state will remain the principal vehicle of government. Accordingly, international law and treaties will shoulder the burden of harmonizing and facilitating the growing number of transactions among the world's nations.

See also: International Arbitration; International Business; International Trade: Commercial Policy and Trade Negotiations; Lex Mercatoria; Public Goods: International.

Bibliography

Bentham, J., 1970. In: Burns, J.H., Hart, H.L.A. (Eds.), An Introduction to the Principles of Morals and Legislation. Athena Press, London.
Brierly, J.L., 1949. The Law of Nations, fourth ed. Oxford University Press, Oxford, UK.
Brownlie, I., 2008. Principles of Public International Law, seventh ed. Oxford University Press, Oxford, UK.
Franck, T.M., 1990. The Power of Legitimacy Among Nations. Oxford University Press, Oxford, UK.
Grotius, H., 1925. In: Scott, J.B. (Ed.), The Law of War and Peace. Carnegie Endowment for International Peace, Washington, DC.
Higgins, R., 1994. Problems & Process: International Law and How We Use It. Oxford University Press, Oxford, UK.
Janis, M.W., 2010. America and the Law of Nations 1776–1939. Oxford University Press, Oxford/New York.
Janis, M.W., 2012. International Law, sixth ed. Aspen, New York.
Janis, M.W., Noyes, J.E., 2011. International Law Cases and Commentary, fourth ed. West, St Paul, MN.
Jennings, R., Watts, A., 1992. Oppenheim's International Law, ninth ed. Longman, Harlow, UK.
Jessup, P.C., 1956. Transnational Law. Yale University Press, New Haven.
McNair, A.D., 1938. The Law of Treaties. Columbia University Press, New York.
Rousseau, C., 1970. Droit International Public. Sirey, Paris.
Wheaton, H., 1836. Elements of International Law. Carey, Lea & Blanchard, Philadelphia, PA.

International Migration

Giovanni Facchini, University of Nottingham, Nottingham, UK; and University of Milan, CEPR, CES-Ifo, CReAM, IZA and LdA, Milan, Italy
Tommaso Frattini, University of Milan, CReAM, IZA and LdA, Milan, Italy
Anna M Mayda, Georgetown University, CEPR, CReAM, IZA and LdA, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Immigration is on the rise in recent decades. In this article, we first provide an overview of the recent evolution of global migration flows. We then analyze the main drivers of international migration, and how they have changed over time. Finally, we review the theoretical and empirical findings of the growing literature that analyzes the economic consequences of immigration in destination countries.

Introduction

Migration is an old phenomenon. Population movements have characterized the history of mankind but the size of the international migrant stock, both in absolute terms and relative to the world population, has never been higher. In 2013, 231 million people (3.2% of the world population) live in a country, which is different from where they were born, up from 75 million (2.3%) in 1965. These numbers indicate that migration is becoming increasingly relevant, although it has been influentially argued (see for instance Pritchett, 2006) that labor mobility is still 'too low' compared to the mobility of capital and goods.

In this article, we first provide a systematic assessment of the main facts and trends that characterize recent international migration flows. We then examine the most important migration drivers identified in the literature, and discuss their evolution over time. Finally, we review the main findings about the economic consequences of migration in receiving countries.

The Changing Geography of Migration

The period after World War II has seen significant changes in the main group of destination countries of international migration flows. **Figure 1** shows the worldwide distribution of the stock of immigrants in 1960. Darker shaded areas indicate locations that have been more intensively exposed to the phenomenon. As is immediately clear, in 1960, immigration was particularly important for countries in Oceania like Australia and New Zealand, and for North American destinations like Canada. Argentina also was characterized by a large number of foreign-born, representing 12.6% of the total population. On the other hand, during this period, the stock of foreign-born in the United States reached a historic low, as a result of the restrictive policies introduced in the aftermath of World War I (see Goldin, 1994), which were removed only in 1965. Importantly, immigration was sizable also in a variety of other destinations, particularly in the developing world.

During the boom years of the 1960s, immigration took off on a global scale. In the United States, the 1965 reform led to a radical change in the channels of entry of new immigrants, with family reunification taking a leading role. The result was large inflows from Latin American countries, and Mexico in particular, that substantially reshaped the ethnic composition of the US population. During this period, the other major North American destination country, Canada, continued to pursue a relatively open immigration policy stance but, in 1967, it introduced the so-called 'point-based system,' a pioneering approach to select immigrants based on observable individual characteristics. Australia and New Zealand continued their efforts to populate their vast territories, promoting in particular the arrival of 'white' immigrants. As for Europe, traditional destinations continued to actively recruit foreign workers throughout the 1960s and the early 1970s. The oil crisis of 1973 represented a critical turning point, since most countries suspended or withdrew from the bilateral workers recruitment agreements they had signed during the boom years. While this policy change had a clear impact in the short run, leading to a substantial decline in new inflows, it made also apparent the limitations of the temporary immigration programs. In fact, only very few of the temporary migrants who in the 1960s arrived in Northern European countries ended up returning home in the 1970s.

Throughout the 1970s and the 1980s, immigration flows started to pick up in the United States and Canada (average annual inflows to the United States during the 1980s reached almost 650 000 and to Canada 126 000; source: OECD International Migration Database) and in Australia and New Zealand. At the same time, the number of foreigners residing in European countries was broadly stable (Dustmann and Frattini, 2013) and started to increase only during the late 1980s, as a result of the better conditions prevailing in the local labor markets.

Figure 2 captures the distribution of international migrants in 2013. At the global level, several important changes have occurred compared to 1960. First, Australia, Canada, and New Zealand have become increasingly reliant upon foreign workers throughout this period and in fact, in 2013, well over 20% of their population was made up by foreign-born individuals. This development has been the result of immigration policies that have allowed large inflows of foreign workers and

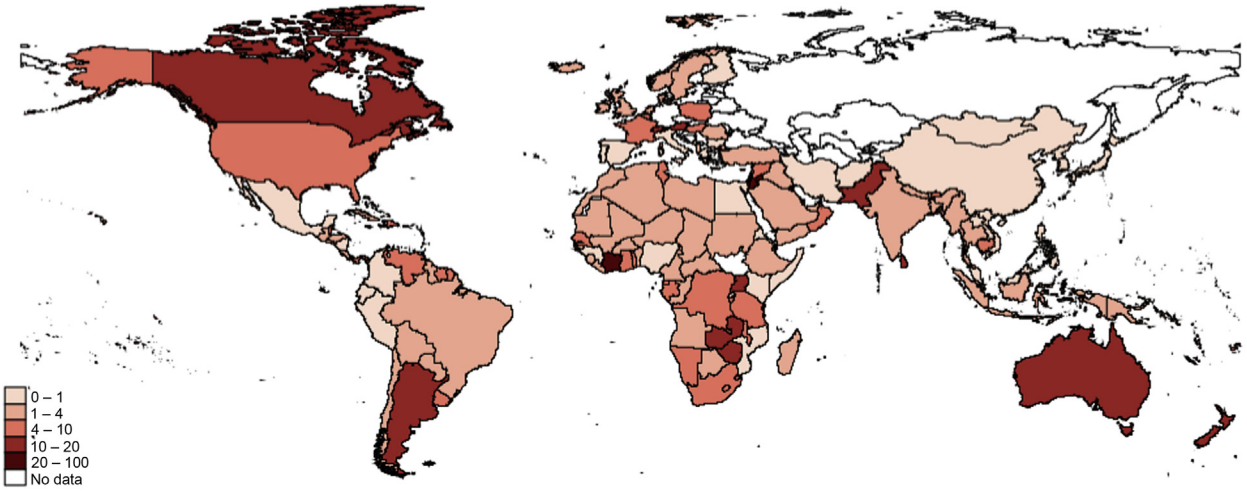

Figure 1 Migrant stock as % of total population, 1960. Source: Our elaboration on United Nations Population Division, Trends in Total Migrant Stock: 2008 Revision.

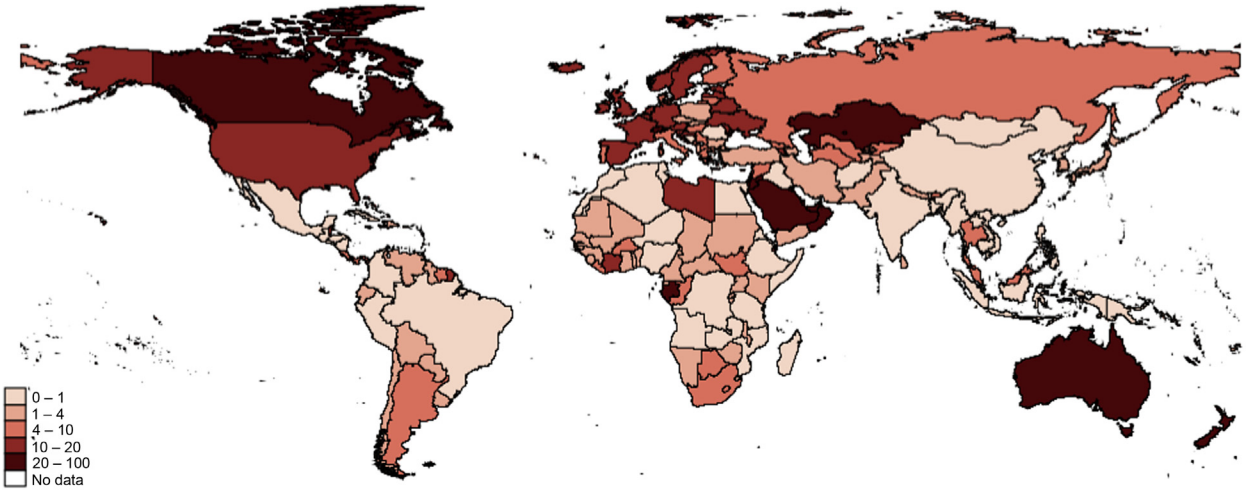

Figure 2 Migrant stock as % of total population, 2013. Source: Our elaboration on United Nations, Department of Economic and Social Affairs (2013). Trends in International Migrant Stock: The 2013 revision.

increasingly focused on skilled migrants, in particular after 1985. At the same time, large inflows of immigrants throughout the 1990s and the following decade have made the foreign-born population in the United States reach levels not seen since the end of the great migration of the early twentieth century. In fact, by 2013, slightly more than 14% of the US population was born abroad (i.e., a total of 45 million individuals).

Second, oil-rich countries in the Persian Gulf area, but also elsewhere like Gabon and Libya, have emerged as the host of very large numbers of foreign-born individuals. While these inflows were the result of skill shortages – of engineers and oil-field workers as well as of financial professionals and service-sector workers – the integration of foreigners in the local societies has been the source of much concern both among source country governments and international organizations (see Ruhs, 2013). Compared to 1960, it is also important to point out how some areas of the world have become less attractive destinations for foreign workers. This is, in particular, the case of Latin America, where over the period 1960–90, the migration inflow came to a complete stop. Some Latin American countries actually turned into positive net *emigration* countries.

The fall of the Berlin wall and a rapidly growing economy during the 1990s have also contributed to substantially change the geography of migration in Europe. By 2013, traditional Northern European destinations harbored large numbers of foreigners, often representing well in excess of 10% of the total population. At the same time, Southern European countries, which in the 1950s and 1960s had been the source of large outflows of workers, have now turned into the destination of large migration inflows. In 2013, 13.8% of the Spanish population is foreign-born and a large immigrant population is also harbored by Italy.

The Drivers of International Migration

A growing literature (see e.g., Mayda, 2010; Ortega and Peri, 2013) has investigated – using aggregate data – the determinants of international migration flows, identifying a series of factors that play a key role. These can be summarized as follows:

- Average income differentials between the home and destination countries
- The distribution of income in the source and the destination countries
- Geographic and cultural distance
- Immigration networks
- Immigration policies

In what follows, we will review the role of each of these determinants.

Average Income Differentials

At least since Sjaastad's (1962) pioneering contribution, economists have modeled the decision to migrate as the result of an income maximization choice, where different locations offer different opportunities. Earlier contributions have focused on bilateral differences, while more recent analyses (Bertoli and Fernandez Huertas Moraga, 2013) have explicitly considered the choice among several alternative locations. Estimates in the literature show that international migration flows are highly responsive to per capita GDP in the destination country. This is consistent with **Figure 3**, where the share of international migrants in the host country's population is positively and significantly correlated with destination country's per capita GDP. The evidence in the literature on the impact of origin country's per capita GDP is instead more mixed.

The Distribution of Income in Source and Destination Countries

While average income differentials are key drivers of the migration decision, workers differ along multiple dimensions and the literature has emphasized the role that individual characteristics are likely to play. An important framework to understand the selection of immigrants along observable features – with respect to both the population in the origin and destination countries – has been proposed by Borjas (1987). The basic idea of his model is that – under specific assumptions on the structure of the migration costs – the relative dispersion of income in the source and the destination countries is a key determinant of the characteristics of who will migrate. Two scenarios are particularly relevant. In the first one, known in the literature as 'positive hierarchical sorting,' migrants are positively selected both at the source and the destination. This is likely to occur when the return to skills in the destination country is higher than in the country of origin and the correlation between the return to skill in the two countries is sufficiently high. This is a desirable situation from the point of view of the host country, as immigrants represent the 'best and the brightest.' An alternative scenario, known as 'negative hierarchical sorting' implies instead a situation in which immigrants are negatively selected both at the source and the destination. This scenario is likely to occur when the return to skills is higher in the source than in the destination country. This is a less desirable outcome from the point of view of the destination, as in this case migrants would tend to be overwhelmingly low-skilled workers.

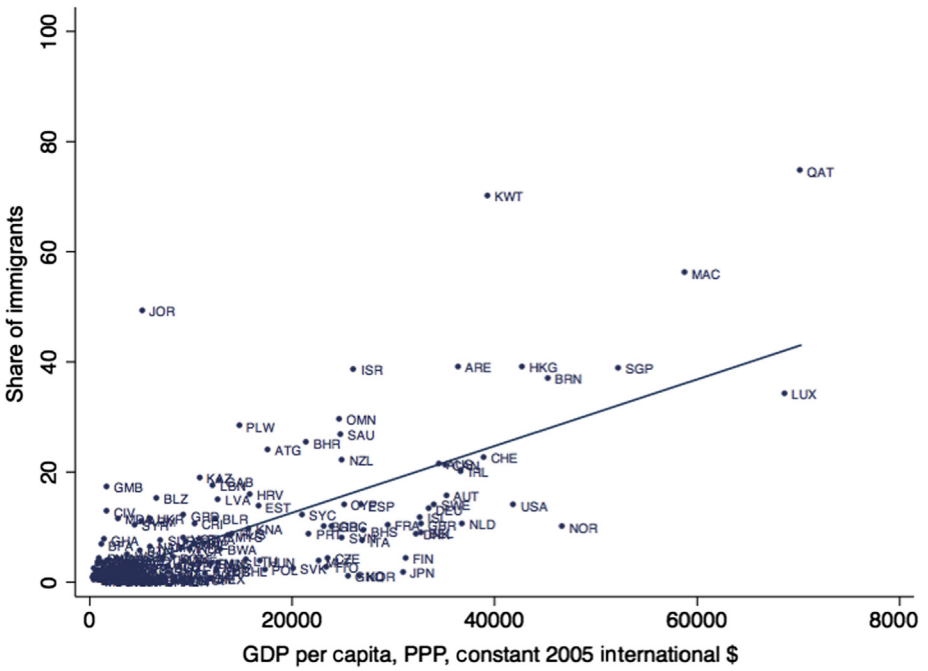

Figure 3 GDP per capita and immigrant share. Source: Our elaboration based on data from the World Bank.

Several macro papers have pointed out the relevance of inequality in the source and the destination countries in explaining immigration flows. A recent historical study (Abramitzky et al., 2012), based on a particularly rich data set of Norwegian families at the end of the nineteenth century, provides instead detailed microlevel evidence in an environment where the predictions of the model can be tested more systematically. In particular, in this article, the authors are able to compare brothers who emigrated to the United Sates, with brothers who instead remained in Norway, supplying very clean evidence in favor of Borjas' (1987) model.

Geographic and Cultural Distance

The third channel that has been emphasized by the literature concerns immigration costs and how they shape the economic calculus of foreign workers when deciding where to settle. We start by looking at the cost of moving abroad – e.g., transportation costs – and we then turn to proxies for other migration costs, e.g., the cultural distance between the source and the destination country's society.

Figure 4 uses data from the UK Department of Transportation to calculate the real price of an average international airfare during the post-1970 period. As we can see, there has been a rapid and sustained decline in the costs of moving from one country to another, on average involving a reduction of 49.7% in the case of business travel, and an even more remarkable decline of 69.1% when it comes to leisure flights.

Measuring the evolution of cultural distance across countries is a much more complex endeavor, but we can consider at least two proxies that provide useful information to that end. First, Figure 5 reports information on the dynamics of communication costs for a group of six European countries.

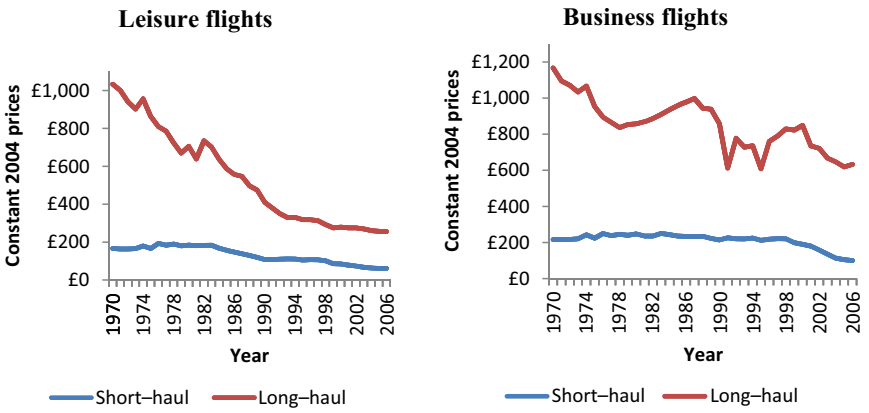

Figure 4 Average fares on international flights. Source: Our calculations based on data from the UK Department of Transportation and the National Archives.

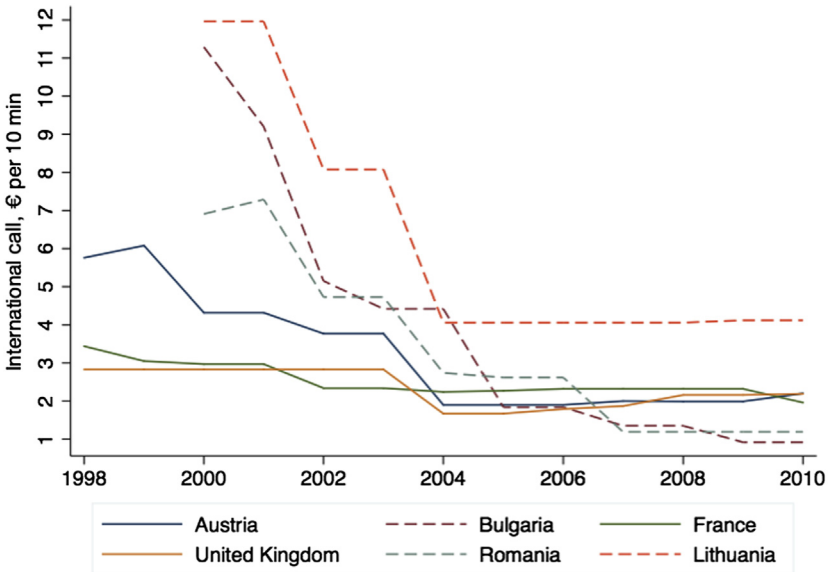

Figure 5 International phone call rates. Source: Our elaboration on Eurostat data.

The data, taken from Eurostat, provide a picture of the rapid decline in the cost of international phone calls, both in immigrant destination and source countries.

Our second proxy for the evolution of cultural distance concerns the diffusion of the Internet. The idea is that as a larger share of the population gains access to the World Wide Web, individuals become more exposed to the value systems prevailing in other countries, and are thus better informed concerning the characteristics of possible destinations before relocating there. Figures 6 and 7, comparing the situation in 1995 and 2012, show that, while a clear digital divide continues to persist between the South and the North of the world, the Internet access has grown exponentially in many developing countries. This has led to a dramatic reduction in the uncertainty faced by individuals in their decision. Economic theory suggests that, everything else equal, this will increase migrants' probability to move.

Immigrant Networks

The fourth determinant of immigration flows that has been emphasized by the literature is the so-called network channel. The basic stylized fact is that migrants (especially from the same origin country) tend to cluster in specific areas (see e.g., Munshi, 2003). This is particularly true for less-skilled individuals and for new arrivals. This last point is nicely illustrated for the United States in a paper by Scott et al. (2005), who establish a strong positive relationship between the existing stock of immigrants in a metropolitan area in 1990 and the immigration flow between 1990 and 1998. The particularly striking result is that six US cities account for approximately 70% of the total settlement of new immigrants, suggesting that preexisting immigrant communities do play a very important role in attracting new arrivals.

But why do migrants tend to cluster together? Several explanations have been put forward in the literature, and the

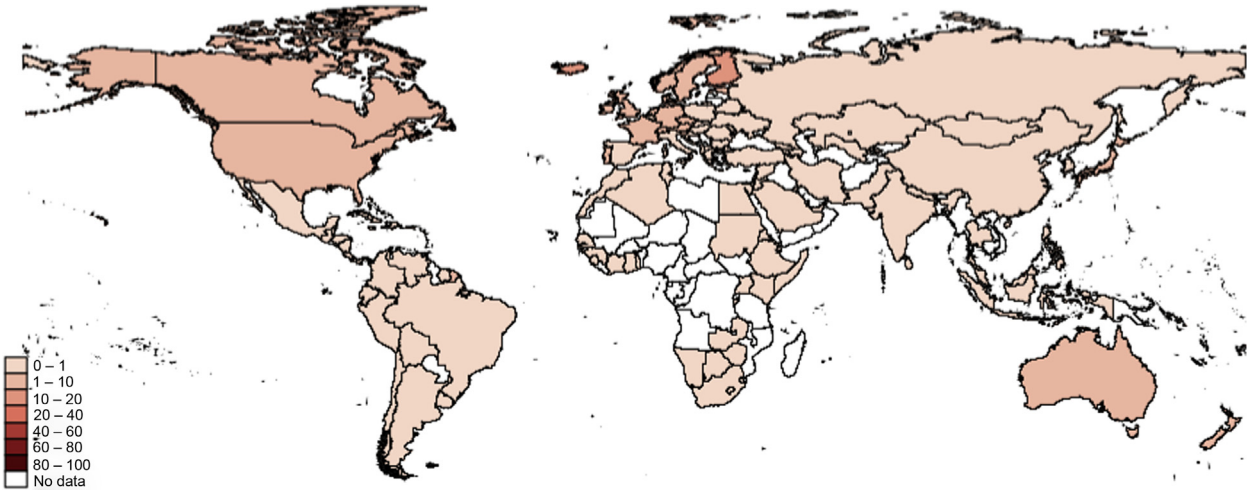

Figure 6 Internet users per 100 inhabitants, 1995. Source: Our elaboration on World Bank, World Development Indicators data.

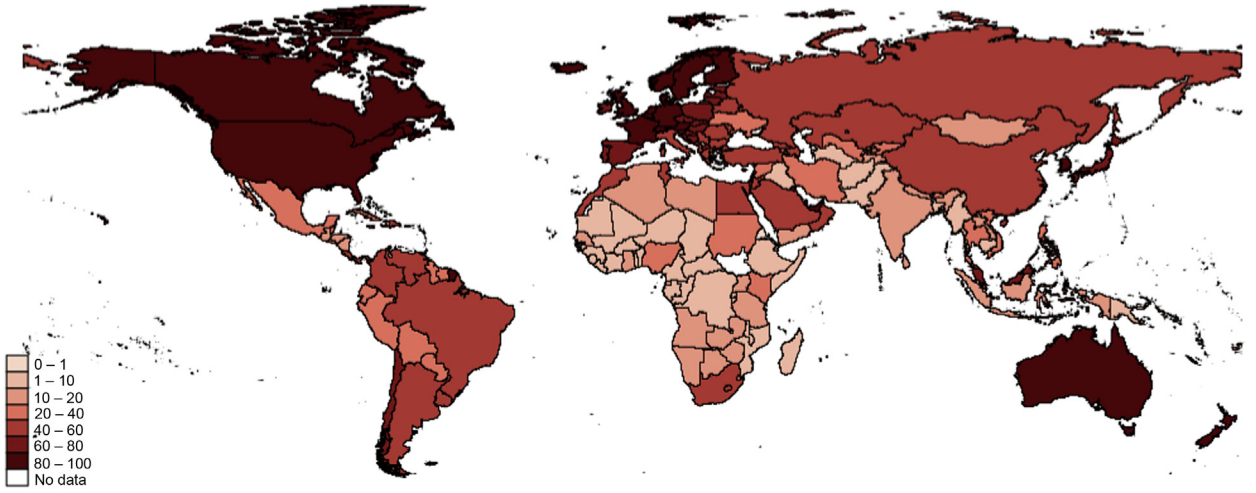

Figure 7 Internet users per 100 inhabitants, 2012. Source: Our elaboration on World Bank, World Development Indicators data.

basic idea is that immigrant networks tend to reduce the cost of migrating (Munshi, 2003).

Immigration Policies

As argued by several observers, the gap in living conditions between the North and the South of the world is so large that the unencumbered flow of workers would allow the global economy to reap huge efficiency gains. Restrictions have made these gains hard to obtain and, in particular, a key role has been played by the policies adopted by destination countries. A few studies have investigated the views of the different actors that play a key role in shaping immigration policy in the destination countries. In particular, Facchini and Mayda (2008) have looked at individual voters' attitudes toward immigration, whereas Facchini and Mayda (2010) have studied the views of key officials in the government. Both suggest that – at least for the near future – it is not realistic to expect large changes in the policy stance, as both public opinion and government officials are not in favor of substantially liberalizing existing policies.

The Impact of International Migration Flows on Destination Countries

International migration affects the destination country's economy through a variety of channels. In this section, we review the main ones, first highlighting the theoretical mechanisms at work and then briefly discussing the most recent empirical findings.

Labor-Market Effects

Assume that skilled and unskilled labors are combined to produce a single good according to a constant returns to scale production function. In this model, immigration inflows increase the labor supply in the destination economy, thus generating a *surplus* with respect to the premigration situation. Importantly, this surplus will not be evenly distributed among the residents of the host country. The income-distribution effects of migration depend on the skill composition of migrants relative to natives. If immigrants are, on average, less skilled than natives, skilled workers will benefit from migration, while unskilled workers will be on the losing end. The reason is that the arrival of immigrants will give rise to a decrease in the unskilled wage and an increase in the skilled wage. If, on the other hand, immigrants are, on average, more skilled than natives, the income-distribution effects of migration are reversed, i.e., immigrants will hurt skilled natives and benefit unskilled ones, as their arrival will induce a decrease in the skilled wage and an increase in the unskilled wage (see Borjas, 1995).

Of course, these predictions rely on the assumption that similarly skilled natives and immigrants are perfect substitutes in production. This assumption might not be very realistic: migrants might lack some country-specific dimension of human capital (like language) and thus cannot compete with native workers with the same level of formal skills. Two recent studies have analyzed the substitutability of immigrants and natives in the UK (Manacorda et al., 2012) and the US (Ottaviano and Peri, 2012). Both studies find that immigrants are not perfect substitutes for natives, even within narrowly defined skill groups, but partly complement their skills. In particular, these studies show that once imperfect substitutability within skill groups is taken into account, the group most negatively affected by new immigrant inflows is that of previous immigrants. One of the reasons of the imperfect substitutability between native and immigrant workers is that these two types of workers tend to specialize in different tasks. In fact, it has been shown (Peri and Sparber, 2009) that unskilled immigrants to the United States have specialized in occupations requiring manual and physical skills, while unskilled natives are employed in language-intensive occupations. Interestingly, a similar pattern of specialization is at work among skilled workers (Peri and Sparber, 2011). To conclude, while these considerations suggest caution in defining skill levels based on educational qualifications, the basic idea underlining the models of perfectly competitive labor markets still remains valid, once care is taken in appropriately defining skill groups (see for instance Dustmann et al. (2013), where skills are defined in terms of position in the wage distribution).

Note that, if immigration affects factor prices, it will also affect the production cost of goods and services. Assuming that output markets are sufficiently competitive, changes in production costs should be reflected in output prices. This channel will operate especially for nontradable goods and services, whose prices are set locally, whereas it is less likely to be at work in the case of tradable goods, as their prices are set on international markets. Does this change in production costs pass through to consumers? Some studies suggest that this is indeed the case (e.g., Cortes, 2008). In particular, immigration has been shown to reduce price growth in low-wage and immigrant-intensive service sectors, with the strongest effects being concentrated in those sectors that use more intensively low-wage workers.

Effects on Output Mix and on Production Techniques

In recent years, researchers are paying increasing attention to other channels – beyond changes in factor and output prices – through which the economy absorbs migration flows. One widely studied mechanism is represented by changes in the output mix. Consider a stylized economy characterized by two factors, skilled and unskilled labor, and two sectors which are, respectively, skilled- and unskilled-labor intensive. In this case, as long as the destination country is a small open economy, it is diversified in production and the size of the migration stock is small, an inflow of unskilled migrants will increase (decrease) the production of the unskilled-labor-intensive (skilled-labor-intensive) good while leaving wages unaffected.

A further adjustment might take place through changes in the production techniques used by firms. In other words, if, for example, an inflow of unskilled migrants takes place, firms can adjust to the changes in relative labor supplies by adopting technologies that use more intensively unskilled labor. There is evidence that immigration has affected the choice of techniques of production in the United States. Even within the same narrowly defined industrial sector, firms in metropolitan areas that have received more unskilled immigrants have decreased their use of automation equipment and have

consequently decreased their equipment-output ratio (Lewis, 2011). Similarly, it has been convincingly shown that in areas with a relatively large immigration-induced increase in the skilled workforce, the adoption of personal computers between 1990 and 2000 took place more rapidly than in areas receiving less-skilled immigrants (Doms and Lewis, 2006). What about changes in output mix? Hanson and Slaughter (2002) find evidence of immigration-related changes in output mix in US states. Some studies have explicitly analyzed the relative importance of changes in techniques of production as compared to changes in output mix in explaining how economies adjust to immigration (see Lewis (2013) for a comprehensive survey of this literature). The results are remarkably robust and suggest that economies adjust to immigration-induced changes in skill mix mostly via within-industry changes in skill intensity; conversely, changes in output composition account for only a small fraction (typically around 10%) of the adjustment.

Productivity and Innovation Effects

One related important issue, which has only recently been studied, is the contribution of immigrants to productivity growth and innovation in the host country. The effects on total factor productivity (TFP) may arise for several reasons. For example, immigration gives rise to changes in the skill composition of the workforce, which affect the probability of innovation. In particular, high-skilled immigrants might benefit the host country via 'human capital spillovers' (see e.g., Moretti, 2004). In addition, immigrants display systematic differences in cultural background, risk attitudes, and language skills compared to the destination country population, etc.

It has been shown, for instance, that labor productivity increased more slowly in the US states that attracted a larger share of immigrants during the 1980s, both in low- and high-skill industries, which suggests that the effect is not due to the low-skill levels of immigrants, but rather to their imperfect short-run assimilation in the US economy (see Quispe-Agnoli and Zavodny, 2002). However, the negative effect on productivity growth may disappear as immigrants acquire language skills and familiarity with the labor market institutions of the host country. Indeed, analyses of the long-run effects of immigration on TFP have shown that, over the period 1960–2006, immigration exerted a strong positive effect on TFP in the United States (see Peri, 2012).

Finally, Hunt and Gauthier-Loiselle (2010) investigate the impact of immigration on innovation and find a positive relationship between the number of patents per capita and skilled immigration at the state level. Kerr and Lincoln (2010) show evidence of a positive effect of the number of H1B visas granted on the direct contribution of ethnic innovators. These papers find that skilled immigrants in the US patent at twice the native rate. In addition, the higher patenting rate of immigrants has not crowded out natives' patenting. On the contrary, there are positive spillovers of immigration on a state's innovation activity: a 1% point increase in the share of immigrant college graduates in the population increases patents per capita by 9–18%. Immigration seems to have boosted innovation in Europe, too. A recent study on 20 European countries (Bosetti et al., 2012) has demonstrated that skilled immigration has led to higher levels of knowledge creation, measured either by the number of patents or by the number of citations to published articles.

Welfare-State Effects

The main OECD destination countries of immigrant flows are characterized by large welfare states (Boeri et al., 2002), through which the public sector redistributes a substantial fraction of national income across individuals. In these contexts, immigration has a nonnegligible impact on public finances, since foreign workers both contribute to and benefit from the welfare state. The aggregate net effect of immigration on the welfare state is either positive or negative, depending on the socioeconomic characteristics of immigrants relative to natives. If immigration is skilled, natives are likely to benefit from the presence of foreign workers, due to a positive welfare spillover. On the other hand, if migration is unskilled, natives may be negatively affected, through a welfare-state leakage effect. Besides this aggregate effect, through the welfare-state channel, immigrants also give rise to income-distribution effects (see Facchini and Mayda, 2009). Consider a simple redistributive system, in which all income sources are taxed at the same rate and all individuals in the economy, i.e., both natives and immigrants, are entitled to an equal lump sum per capita benefit. By construction, this simple welfare system redistributes resources from high- to low-income individuals. Thus, the extent to which natives are affected by migration through the welfare-state channel differs according to each voter's income level. That is, there will be income-distribution effects. These effects depend on the adjustment mechanism of the welfare state to migration. On the one hand, migration can bring about changes in the tax rate, while per capita benefits are kept constant (*tax adjustment model*). On the other, the per capita benefits can adjust, while tax rates are unchanged (*benefit adjustment model*).

Under the tax adjustment model, if immigration is skilled compared to the native population, all individuals benefit from the inflow of foreign workers, since the government will be able to lower the tax rate (while keeping the level of per capita benefits unchanged). However, the reduction in the tax rate will benefit rich individuals to a greater extent than poor ones, since tax payments represent a larger fraction of rich voters' net income. If immigration is on the other hand unskilled, all individuals are worse off, since the tax rate needs to increase to keep the level of per capita benefits unchanged. Richer individuals are those more negatively affected, since the burden of financing the welfare state falls disproportionately more on richer members of society. To summarize, under the tax adjustment model, any change in tax rates (either negative or positive, depending on the skill mix of migrants) will have a bigger impact on high-income individuals.

As in the tax adjustment model, also under the benefit adjustment model, an inflow of skilled immigrants benefits all individuals, while an inflow of unskilled immigrants worsens the position of all natives. The difference is that, under the benefit adjustment model, an inflow of immigrants will have a more pronounced effect (either positive or negative, depending on the skill mix of migrants) on those individuals who are at the receiving end of the welfare system, i.e., those

with a low income. The reason is that the per capita benefit represents a larger fraction of a poor individual's net income, thus changes in per capita benefits affect this group more strongly. In particular, if a country receives skilled migrants, the per capita benefit will increase, and all natives will be made better off, but poor ones more so than rich ones. On the other hand, a country receiving unskilled migrants will be forced to reduce per capita benefits, thus hurting everybody, but again low-income voters to a greater extent than high-income ones.

Conclusions

In this article, we have systematically assessed the main facts and trends that characterize recent international migration flows and have provided a short review of the research on the causes and consequences of migration. Space constraints have forced us to focus only on the effects of immigration on destination countries. However, it is important to keep in mind that immigration significantly affects source countries too, especially through large remittance flows, and through its impact on the availability of human capital.

It is not an overstatement to say that international migration could radically change the future of the world economy in the next decades (as has happened with trade in goods and services and capital flows). The drop in transport costs and cultural distances in modern economies, coupled with persistent income differentials across countries, have led to an unprecedented increase in migration pressures. A clear understanding of the determinants and impacts of this phenomenon is the starting point of sound economic policies, both in destination and origin countries.

See also: Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Global Migration; Immigration Policy; Immigration, Applied Research on; Immigration, Citizenship, and Integration: Social Work Connections; Immigration, Emigration, and Citizenship; Immigration: Health and Mortality; Immigration: Political Aspects; Migration, Economics of; Migration, Ethnicity, Aging and Social Work Practice; Migration: Motivations; Migrations, Colonizations, and Diasporas in Archaeology.

Bibliography

Abramitzky, R., Boustan, L., Eriksson, K., 2012. Europe's tired, poor, huddled masses: self-selection and economic outcomes in the age of mass migration. American Economic Review 102, 1832–1856.

Bertoli, S., Fernandez Huertas Moraga, J., 2013. Multilateral resistance to migration. Journal of Development Economics 102, 79–100.

Boeri, T., Hanson, G.H., McCormick, B., 2002. Immigration Policy and the Welfare System. Oxford University Press, Oxford.

Borjas, G.J., 1987. Self-selection and the earnings of immigrants. American Economic Review 77 (4), 531–553.

Borjas, G.J., 1995. The economic benefits from immigration. Journal of Economic Perspectives 9 (2), 3–22.

Bosetti, V., Cattaneo, C., Verdolini, E., 2012. Migration, cultural diversity and innovation: a European perspective. IGIER Working Paper n.469. Bocconi University, Milan, Italy.

Cortes, P., 2008. The effect of low-skilled immigration on U.S. Prices: evidence from CPI data. Journal of Political Economy 116 (3), 381–422.

Doms, M., Lewis, E.G., 2006. Labor Supply and Personal Computer Adoption. Federal Reserve Bank of San Francisco Working Paper No. 2006-18.

Dustmann, C., Frattini, T., Preston, I., 2013. The effect of immigration along the distribution of wages. Review of Economic Studies 80 (1), 145–173.

Dustmann, C., Frattini, T., 2013. Immigration: the European experience. In: Card, D., Raphael, S. (Eds.), Immigration, Poverty, and Socioeconomic Inequality. Russell Stage Foundation, New York.

Facchini, G., Mayda, A.M., 2008. From individual attitudes to immigration policy: theory and evidence. Economic Policy 56, 651–713.

Facchini, G., Mayda, A.M., 2009. Does the welfare state affect individual attitudes toward immigrants? Evidence across countries. Review of Economics and Statistics 91 (2), 295–314.

Facchini, G., Mayda, A.M., 2010. What drives immigration policy? Evidence from a survey of government officials. In: Gang, I., Epstein, G.S. (Eds.), Culture and Migration (Frontiers of Economics and Globalization), vol. X. World Scientific, Singapore, pp. 605–648.

Goldin, C., 1994. The political economy of immigration restriction in the United States, 1890 to 1921. In: Goldin, C., Libecap, G. (Eds.), The Regulated Economy: An Historical Analysis of Political Economy. University of Chicago Press, Chicago.

Hanson, G.H., Slaughter, M.J., 2002. Labor-market adjustment in open economies: evidence from US states. Journal of International Economics 57 (1), 3–29.

Hunt, J., Gauthier-Loiselle, M., 2010. How much does immigration boost innovation? American Economic Journal: Macroeconomics 2 (2), 31–56.

Kerr, W., Lincoln, W.F., 2010. The supply side of innovation: H-1B visa reforms and U.S. ethnic invention. Journal of Labor Economics 28 (3), 473–508.

Lewis, E., 2011. Immigration, skill mix, and capital-skill complementarity. Quarterly Journal of Economics 126 (2), 1029–1069.

Lewis, E., 2013. Immigration and production technology. Annual Review of Economics 5, 165–191.

Manacorda, M., Manning, A., Wadsworth, J., 2012. The impact of immigration on the structure of wages: theory and evidence from Britain. Journal of the European Economic Association 10 (1), 120–151.

Mayda, A.M., 2010. International migration: a panel data analysis of the determinants of bilateral flows. Journal of Population Economics 23, 1249–1274.

Moretti, E., 2004. Workers' education, spillovers and productivity: evidence from plant-level production functions. American Economic Review 94 (3), 656–690.

Munshi, K., 2003. Networks in the modern economy: Mexican migrants in the U.S. labor market. Quarterly Journal of Economics 118, 549–599.

Ortega, F., Peri, G., 2013. The role of income and immigration policies in attracting international migrants. International Migration Studies 1, 1–28.

Ottaviano, G.I.P., Peri, G., 2012. Rethinking the effect of immigration on wages. Journal of the European Economic Association 10 (1), 152–197.

Peri, G., 2012. The effect of immigration on productivity: evidence from U.S. states. Review of Economics and Statistics 94 (1), 348–358.

Peri, G., Sparber, C.S., 2009. Task specialization, immigration, and wages. American Economic Journal: Applied Economics 1 (3), 135–169.

Peri, G., Sparber, C.S., 2011. Highly-educated immigrants and native occupational choice. Industrial Relations 50 (3), 385–411.

Pritchett, L., 2006. Let Their People Come: Breaking the Gridlock on Global Labor Mobility. Center For Global Development, Washington, DC.

Quispe-Agnoli, M., Zavodny, M., 2002. The effect of immigration on output mix, capital and productivity. Federal Reserve Bank of Atlanta Economic Review 87 (1), 17–27.

Ruhs, M., 2013. The Price of Rights: Regulating International Labor Migration. Princeton University Press, Princeton, NJ.

Scott, D.M., Coomes, P.A., Izyumov, A.I., 2005. The location choice of employment-based immigrants among U.S. metro areas. Journal of Regional Science 45 (1), 113–145.

Sjaastad, L., 1962. The costs and returns of human migration. Journal of Political Economy 70, 80–93.

International Migration and Remittances

Dean Yang, Department of Economics, Gerald R. Ford School of Public Policy, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The remittances that these migrants send to origin countries are one of the largest types of international financial flows to developing countries. This article reviews the relationships between remittances and the economic development of the world's poorer nations. It first reviews the definition, magnitude, and some basic characteristics of remittance flows. Then, it treats the motivations for remittances, the impacts of remittances on development outcomes, and their role in helping respond to adverse shocks. It concludes with thoughts about the immediate future of remittances and promising areas for future research on the topic.

Between 1965 and 2000, individuals living outside their countries of birth grew from 2.2 to 3.1% of world population, reaching a total of 214 million people in 2010 (United Nations, 2011). The remittances that these migrants send to origin countries are one of the largest types of international financial flows to developing countries.

This article is on remittances and their relationship with the economic development of the world's poorer nations. It will first review the definition, magnitude, and some basic characteristics of remittance flows. Then, it will treat the motivations for remittances, the impacts of remittances on development outcomes, and their role in helping origin countries and – households respond to adverse shocks. It will conclude with thoughts about the immediate future of remittances and promising areas for future research on the topic.

Definition, Magnitude, and Basic Characteristics

Remittances are household income received from overseas, primarily initiated by international migrants sending funds to home countries. Remittances may be sent as cash or in kind, and may flow through a variety of formal or informal channels. Formal channels include dedicated money transfer operators (e.g., Western Union, MoneyGram), banks, and credit unions. Informal channels include systems – such as *hawala* and *hundi* in South Asia and *padala* in the Philippines – operated by nonfinancial firms or brokers. In international balance of payments data, remittances are measured as the sum of two categories of transactions: 'personal transfers' (all current transfers in cash or in kind made or received by resident households to or from nonresident households) and 'compensation of employees' (earnings of temporary workers who are not resident in their host countries) (International Monetary Fund, 2013).

Since the late 1990s, remittances sent home by international migrants have exceeded official development assistance, and in several years have approached the magnitudes of foreign direct investment (FDI) flows. Figure 1 graphs these four categories of financial flows to developing countries from 1991 to 2011. In 2011, migrant remittances sent to developing countries amounted to US$353 billion. Developing country receipts of

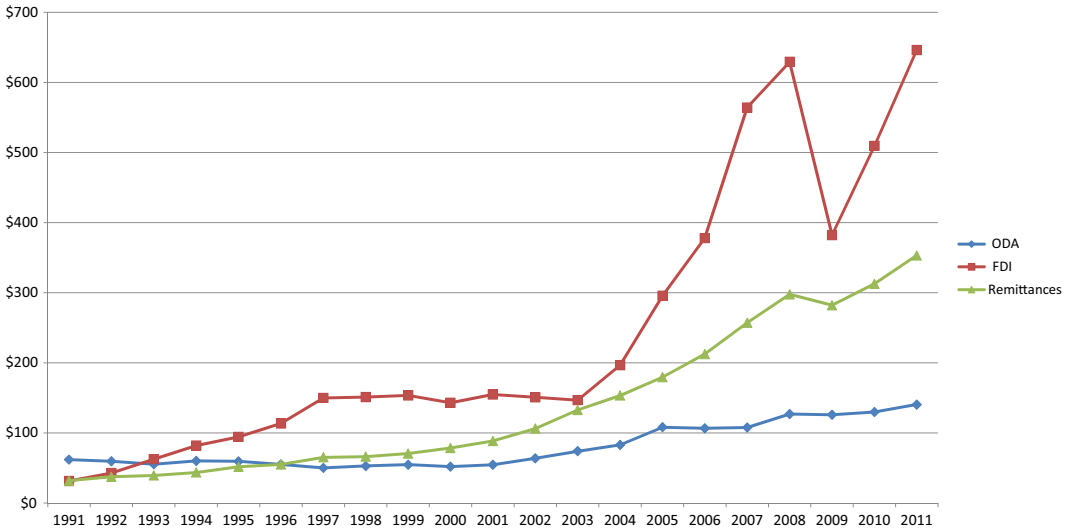

Figure 1 Remittances vs other international financial flows to developing countries (1990–2011). Note: Data are from the World Bank's World Development Indicators (WDI) 2013. Data are in billions of constant (2005) US$, in total across developing countries (low and middle income as classified by World Bank). Variables displayed are 'net official development assistance and official aid received (current US$),' 'foreign direct investment, net inflows (BoP, current US$),' and 'personal remittances, received (current US$).'

Table 1 Top remittance recipient countries

	Remittance receipts (USD billions), 2013 estimate		*Remittances received as % of GDP, 2012*
India	71	Tajikistan	48
China	60	Kyrgyz Republic	31
Philippines	26	Nepal	25
Mexico	22	Lesotho	25
Nigeria	21	Moldova	24
Egypt	20	Armenia	21
Bangladesh	15	Haiti	21
Pakistan	15	Samoa	21
Vietnam	11	Liberia	20
Ukraine	9	Lebanon	17

Note: Data on the dollar value of remittances received are estimated for 2013, and data on remittances received as share of GDP are from 2012. For estimation methodology, see Ratha, Dilip, Eigen-Zucchi, Christian, Plaza, Sonia, Wyss, Hanspeter, Yi, Soonhwa, 2 October 2013. Migration and Remittance Flows: Recent Trends and Outlook, 2013–2016. Migration and Development Brief 21. Development Prospects Group, World Bank.

FDI (the largest type of international financial flow to the developing world) were less than double that figure ($646 billion). Receipts of official development assistance (foreign aid) came in a poor third to remittances and FDI in 2011, amounting to just $141 billion. (Data are from the World Bank's World Development Indicators 2013. The 2008–09 financial crisis had a substantial negative impact on FDI flows, while remittances and ODA were by contrast relatively stable. In 2009, FDI, remittance, and ODA flows to developing countries were $382, $282, and $126 billion, respectively.) The World Bank estimates that remittances to developing countries will amount to $414 billion in 2013 (Ratha et al., 2013). Motivated by their large magnitudes, international financial institutions and developing country governments are keenly interested in finding policies that can stimulate remittances and enhance their development impacts.

Table 1 displays remittance data for the 10 largest remittance-receiving countries, ranked by dollar value (column 1) and by share of GDP (column 2). The largest remittance-receiving countries in 2013 by (estimated) dollar value are India and China, which received $71 billion and $60 billion, respectively. The Philippines and Mexico are nearly tied for the next two places, receiving $26 billion and $22 billion, respectively. When it comes to remittances as a share of GDP, on the other hand, rankings (based on 2012 data) are very different. Countries with small populations but large migrant flows end up at the top of this list, led by Tajikistan (where remittances amount to 48% of GDP) and followed by the Kyrgyz Republic (31%), Nepal (25%), Lesotho (25%), and Moldova (24%).

Remittances are not only large in aggregate magnitude. They are also prominent in the context of individual financial lives, making up a substantial fraction of migrant workers' earnings. Table 2 presents data on remittances sent as a fraction of earnings from several surveys of various migrant populations. For many migrant populations, the share of earnings remitted is substantial. Mexican migrants (surveyed upon return to Mexico) remit 31.1% of earnings, while migrants from El Salvador surveyed in the Washington, DC, area in 2007–2008 report remitting 37.7% of their United States earnings. Senegalese in Spain remit 49.9% of earnings, and Ghanaians in Italy 23.3%. In other surveyed migrant populations, the share of remittances out of earnings is more modest: Moroccan immigrants in France remit 10.4% of earnings, Turks in Germany 2.1%, Chinese in Australia 6.1%, Filipinos in the United States 5.8%, and Cubans in the United States 2.3%.

Remittances differ from the other large types of international financial flows to developing countries – such as FDI and official development flows – in that they are sent at relatively high frequencies and in relatively small magnitudes. This pattern has been found in a diverse variety of migrant populations.

For example, Yang (2011) analyzes transaction-level data from a money transmitter in Washington, DC, and finds that

Table 2 Remittance activity in selected migrant origin-destination country pairs

Origin country	*Migrant destination country*	*Average remittances as a percentage of earnings (%)*	*Average annual remittances ($ value)*	*Data source*	*N*
China	Australia	6.09	552	Australia LSIA	65
Morocco	France	10.37	1283	France 2MO	128
Turkey	Germany	2.14	512	Germany SOEP	334
Ghana	Italy	23.28	2528	Italy NIDI	497
Senegal	Spain	49.91	3304	Spain NIDI	399
Mexico	United States	31.12	4125	MMP	1268
El Savador	United States	37.72	5314	ESSMF	877
Philippines	United States	5.84	958	US NIS	344
Cuba	United States	2.32	398	US PEW	111

Note: Data are from the following sources: China–Australia: 1997 Longitudinal Survey of Immigrants to Australia (Australia LSIA), http://www.immi.gov.au/media/research/lsia/; Morocco–France: Survey of Households' Transfer of Funds to their Countries of Origin (France 2MO), Miotti, Luis, Mouhoud, El Mouhoub, Oudinet, Joel, 2009. Migrations and Determinants of Remittances to Southern Mediterranean Countries: When History Matters. Paper Presented at the 2nd Migration and Development Conference, Washington, DC, September 10–11; Turkey–Germany: 2000 German Socio-Economic Panel (Germany SOEP), http://www.diw.de/english/soep_overview/33899.html; Senegal–Spain: Netherlands Interdisciplinary Demographic Institute International Migration Survey (Spain NIDI), Groenewold, George, Bilsborrow, Richard, 2004. Design of Samples for International Migration Surveys: Methodological Considerations, Practical Constraints and Lessons Learned from a Multi-country Study in Africa and Europe. Population Association of America (General Conference); Mexico–United States: Mexican Migration Project (MMP), http://mmp.opr.princeton.edu/; Philippines–United States: New Immigrant Survey (US NIS), http://nis.princeton.edu/; El Salvador–United States: El Salvador Survey of Migrant Families (ESSMF), Ashraf, Nava, Diego Aycinena, Claudia Martinez A., Dean Yang, forthcoming. Savings in transnational households: a field experiment among migrants from El Salvador. Review of Economics and Statistics. Cuba–United States: Pew National Survey of Latinos (US Pew), http://pewhispanic.org/datasets/signup.php?DatasetID=7.

migrants from El Salvador sent an average of 16.9 remittances over a 12-month period in 2006–2007, with more than a fifth of migrants undertaking 26 or more remittance transactions (at least biweekly) over the period. The average amount sent per remittance was $299.21, and the median was $200. In 88.6% of transactions, the amount sent was $500 or less. These patterns are consistent with other surveys of Hispanic immigrants in the United States. Orozco and Fedewa (2006) report that 81% of individual remittance transactions sent to a major bank in Guatemala were equal to or less than $300. Bendixen and Associates (2008), in a survey of 5000 Hispanics in the United States, found that 50% sent remittances on a regular basis, remitters sent on average 15 remittances per year, and the average amount sent per remittance was $325. Other survey-based studies finding similar results include Menjivar et al. (1998), DeSipio (2000), Clark and Drinkwater (2001), Bendixen and Associates (2001, 2004a,b), and Amuedo-Dorantes et al. (2005).

Motivations Behind Remittances

The motivations for remittances are varied, and often overlap with motivations behind the original decision to migrate. Migration is frequently well-described as an investment on the part of the migrant and his or her family, where the returns to the investment take the form of higher earnings in a more developed migration destination. In this context, remittances often constitute the portion of the investment return from migration that accrues to individuals in the migrant origin country (Clemens and Ogden, 2013; De Haas, 2010). Docquier and Rapoport (2006) provide a model of the remittance-sending decision that incorporates a variety of motives, including altruism, exchange (compensation for services rendered to the migrant by recipients), insurance, loan repayment, and investment, some or all of which could operate simultaneously.

Altruistically motivated migrants could send remittances to increase recipient consumption levels (Stark, 1995), and may also be responsive to shocks experienced by recipients and thus play an insurance function (Cox et al., 1998; Gubert, 2002). Remittances may be intended for physical or human capital investments by recipients. Remittances could also have self-interested motivations, such as repayment of debts incurred for the migrant's education in the home country or the initial fixed costs of migration (Poirine, 1997; Ilahi and Jafarey, 1999). Remittances could also be intended for migrants' future investments in the home country, or for monitoring or administration of investment assets. Remittances may be sent to secure a future inheritance from elders being supported in the home country (Hoddinott, 1994; de la Briere et al., 2002; Osili, 2004).

Impacts of Remittances

Remittances have a variety of impacts on recipient countries and households. Aggregate analyses of the relationship between remittances and economic performance at the country level are inconclusive. Some studies find a positive relationship between remittances and economic growth (Faini, 2006; World Bank, 2006; Barajas et al., 2009), while others find a zero or negative relationship (Chami et al., 2003; IMF, 2005; Giuliano and Ruiz-Arranz, 2005).

Studies using country-level data face inherent challenges in conclusively establishing causal impacts of remittances on national-level outcomes. Studies using data at the micro (household or individual) level can often delineate causal pathways more convincingly, and can also shed light on remittance impacts with greater detail and nuance.

Many papers argue that remittances rarely fund productive investments, and instead mainly allow higher consumption: for example, see Brown and Ahlburg (1999), and references cited in Durand et al. (1996). However, others have found that migration and remittance receipts are positively correlated with various types of household investments in developing countries. Examples include Brown (1994), Massey and Parrado (1998), McCormick and Wahba (2001), Dustmann and Kirchkamp (2002), Woodruff and Zenteno (2007), and Mesnard (2004) on entrepreneurship and small business investment in a variety of countries; Adams (1998) on agricultural land in Pakistan; Taylor et al. (2003) on agricultural investment in China; Cox-Edwards and Ureta (2003) and Adams (2005) on schooling investments in El Salvador and Guatemala, respectively; and others. (Of course, neither consumption nor investment should be assumed a priori to be a 'better' use of remittances. It could be optimal for households to use remittances mainly on consumption, particularly if they are starting from very low consumption levels.)

A key methodological concern with existing work that attempts to understand effects on households is that remittances are not randomly allocated across households. In general, therefore, any observed relationship between migration or remittances and household outcomes may simply be due to unobserved third factors, or reverse causation. A more recent research literature on the impacts of migration (and the associated remittance flows) seeks to go beyond establishing correlations (as interesting as they may be), and pays close attention to establishing causal impacts on outcomes of interest.

An experimental approach to establishing causal impacts of migration and remittances would randomly assign migration opportunities (or fluctuations in the size of remittances), and examine how households or localities in the migrant-source country are affected. While there are substantial practical obstacles to conducting such experiments, a number of studies have identified situations in the real world – or 'natural experiments' – that in key respects approximate the randomized experiment just described.

One type of natural experimental study involves taking advantage of lotteries that migration destination countries have implemented to decide which of many applicants will receive a limited number of immigration visas. Lottery winners can then be compared with lottery losers, and any differences between the groups can be interpreted as due to differences in the opportunity to migrate. A number of papers have measured impacts of winning lotteries for immigration visas to New Zealand from Pacific island nations. Gibson et al. (2010) finds that Tongans moving to New Zealand due to winning the lottery increase their income 263% within the first year of moving, and also show improved mental health (Stillman et al., 2009), compared with lottery losers.

Gibson et al. (2011) then examine the impact of migrating due to visa lottery winning on household members remaining behind in Tonga. They find that per capita income and consumption of remaining household members are actually lower for left-behind household members of lottery winners, compared with similar individuals in lottery-loser households. This appears to be due to the fact that migrants send home lower amounts in remittances than they were earning domestically prior to migration. This negative finding does not appear universal, however. Gibson et al. (2013) conduct a similar analysis among Samoan households with and without winners in the New Zealand immigration visa lottery, finding that migration due to winning the visa lottery led to lower poverty rates and higher household incomes among left-behind household members.

Another natural experimental approach to establishing the impact of migration and remittances was implemented by Clemens and Tiongson (2013). Their focus is on Filipinos applying to migrate to Korea for temporary contract work. Their analysis exploits the fact that eligibility for this migration opportunity was conditional on receiving a certain minimum score on a Korean language test. Applicants scoring at the score cutoff or above were deemed eligible, and those scoring below were not. Their analysis involves comparing applicants who scored at or just above the Korean test score cutoff, and those scoring just below, on the basis that scoring just above or just below is as good as randomly determined. They find that eligibility for this migration opportunity on the basis of the Korean test score leads to large increases in the likelihood of international labor migration. The origin households of successful applicants in turn receive higher remittances, have higher expenditures, and invest more in children's human capital 3–5 years later.

Another revealing type of natural experimental analysis takes advantage of shocks to the economic conditions of migrants, and examines how origin households are affected. Yang and Martinez (2005) and Yang (2008b) take advantage of the fact that a nonnegligible fraction of households in the Philippines have one or more members working overseas at any one time. These overseas Filipinos work in dozens of foreign countries, many of which experienced sudden changes in exchange rates due to the 1997 Asian financial crisis. These exchange rate changes were unexpected and varied in magnitude across overseas Filipinos' locations. The net result was large variation in the size of the exchange rate shock experienced by migrants across source households. Taking advantage of this variation in the size of migrant exchange rate shocks, these papers examine the impact of the shocks on changes in outcomes in migrants' origin households. Yang (2008b) shows that more positive exchange rate shocks (from the standpoint of the migrant, appreciations of the overseas currencies in which they earn vs the Philippine currency) lead to increases in remittances sent home. Households whose migrants experience more favorable exchange rate shocks raise their nonconsumption disbursements in several areas likely to be investment related – in particular, educational expenditures – and show enhanced participation in entrepreneurial activities. Households raise hours worked in self-employment, and become more likely to start relatively capital-intensive household enterprises like transportation/communication services and manufacturing. By contrast, there is no identifiable effect of the exchange rate shocks on current household consumption. Yang and Martinez (2005) shows that these positive migrant exchange rate shocks also lead to lower poverty rates among migrant origin households.

One question that is often raised is whether remittance lowers recipients' incentive to work (sometimes referred to as a 'dependency' effect of remittances). Across studies, there is little evidence of such an effect. Gibson et al. (2011) find no statistically significant effect of migration due to visa lottery winning on labor supply of left-behind household members. Yang (2008b) find no effect of migrants' exchange rate shocks on wage labor in origin households, and actually finds a positive effect on hours worked in self-employment. The Clemens and Tiongson (2013) study also finds no effect of labor migration to Korea on labor supply of nonmigrant origin household members.

Insurance Role of Remittances

Remittances may also bring benefits to origin households of migrants by serving an insurance role. Households in developing countries are exposed to a variety of types of risk, such as from shocks to weather or other determinants of agricultural productivity, economic fluctuations of other sorts, or adverse health events within the household. Overseas migrants could respond to adverse events in origin households by sending assistance in the form of increased remittances. Such insurance provided by migrants could help origin households maintain human capital investments, encourage them to adopt new (and riskier) production technologies or entrepreneurial activities.

Remittance flows in the aggregate do appear to rise in response to negative shocks at the country level. Mishra (2005) examines aggregate remittances in 13 Caribbean countries from 1980 to 2002 and finds that every 1% decrease in GDP is associated with a 3% increase in remittances 2 years later. Yang (2008a) examines the impact of hurricanes on international financial flows using country-level data, and finds that for poorer developing countries, remittances rise after countries are hit by hurricanes, and amounting to about one-fifth of the value of hurricane damages (a response about one-quarter as large as the response of foreign aid).

Household-level analyses come to similar conclusions. For example, Yang and Choi (2007) ask whether migrant remittances serve as insurance by focusing on responses to income fluctuations driven by weather in the Philippines. Among households with migrant members, they find that remittances rise when incomes fall due to negative weather shocks. The amount of insurance provided is large, with about three-fifths of declines in income replaced by new remittances sent by migrants.

Concluding Thoughts

Remittances will likely remain one of the most important types of international financial flows to developing countries. Increasing competition in the money transfer industry, leading to declines in remittance transaction costs, will drive continued strong growth of remittances. Innovations such as cell phone-based remittance services will make sending remittances more

and more convenient for migrants. The existing empirical evidence suggests that migrants are likely to respond to the reductions in transaction costs and increases in convenience by sending substantially more remittances in dollar terms (Aycinena et al., 2010.) As a result, remittances are projected to continue to grow in the coming years, reaching $540 billion in 2016 (Ratha et al., 2013).

There are attractive prospects for continued remittance-related innovation on the part of money transmitters and financial institutions, and collaborations with academics and the nonprofit and government sectors. One direction for research is on the extent to which migrants value control over how recipients use remittances, for which there is emerging new evidence, for example in the area of control over savings in the home country (Ashraf et al., forthcoming). It should be useful to explore the demand for and impact of improving migrant control over other remittance uses, such as human capital investments or investments in small enterprises. Such innovations may lead to enhanced impacts on economic development outcomes in home areas, and may also encourage migrants to send more remittances.

Another area where future investigations might focus is on reasons why remittances tend to be sent in low-value transactions, at high frequencies. This pattern is a puzzle, because of substantial per-transaction remittance fees. For example, remittances examined in Yang's (2011) sample, remittances to El Salvador typically incurred a flat fee of $9 or $10 per transaction. An open question is why remitters fail to minimize fees paid by sending large amounts relatively rarely. There are a number of possible explanations for low-value, high-frequency transactions, such as liquidity constraints on the part of senders, self-control problems on the part of senders or recipients, or risk of theft or loss on either the sender or recipient side of the transaction. Valuable future work could seek to shed light on the likely causes of this very prominent pattern in remittance sending behavior.

Notwithstanding recent advances in our knowledge on migration, remittances, and their impacts, much more remains to be learned. We are likely to see major advances on remittance-related questions in the coming years. It remains a fruitful area for future research and experimentation.

See also: Internal Migration: Developing Countries; International Migration by Ethnic Chinese; International Migration by Ethnic South and Southeast Asians; International Migration; Migration and Fertility; Migration to and from Latin America; Migration, Theory of.

Bibliography

Adams Jr., Richard H., 1998. Remittances, investment, and rural asset accumulation in Pakistan. Economic Development and Cultural Change 17 (1), 155–173.

Adams Jr., Richard H., 2005. Remittances, household expenditure and investment in Guatemala. In: Ozden, Caglar, Schiff, Maurice (Eds.), International Migration, Remittances, and the Brain Drain. World Bank, Washington, DC.

Amuedo-Dorantes, Catalina, Bansak, Cynthia, Pozo, Susan, 2005. On the remitting patterns of immigrants: evidence from Mexican survey data. Economic Review 90 (1), 37–58.

Ashraf, Nava, Diego Aycinena, Claudia Martinez A., Dean Yang, forthcoming. Savings in transnational households: a field experiment among migrants from El Salvador. Review of Economics and Statistics.

Aycinena, Diego, Claudia Martinez, A., Yang, Dean, 2010. "The Impact of Remittance Fees on Remittance Flows: Evidence from a Field Experiment Among Salvadoran Migrants," Working Paper. University of Michigan.

Barajas, A., Chami, R., Fullenkamp, C., Gapen, M., Montiel, P., 2009. Do Workers' Remittances Promote Economic Growth? IMF Working Paper WP 09/153.

Bendixen and Associates, 2001. Survey of Remittance Senders: U.S. To Latin America. Bendixen and Associates, Miami, FL. http://www.bendixenandassociates.com/studies/IDB%20Remesas%202001.pdf.

Bendixen and Associates, 2004a. Remittances and the Dominican Republic: Survey of Recipients in the Dominican Republic and of Senders in the United States. Bendixen Associates, Miami, FL. http://www.bendixenandassociates.com/studies/IDB%20Dominican%20Republic%20Presentation%20FINAL%202004.pdf.

Bendixen and Associates, 2004b. State by State Survey of Remittance Senders: U.S. To Latin America. Bendixen and Associates, Miami, FL. http://www.bendixenandassociates.com/studies/IDB%20National%20Survey%20of%20Remittance%20Senders%202004.pdf.

Bendixen and Associates, 2008. Survey of Latin American Immigrants in the United States. Bendixen and Associates, Miami, FL. http://www.bendixenandassociates.com/studies/IDB_2008_National_Survey_Presentation.ppt.

Brown, R.P.C., 1994. Migrants' remittances, savings, and investment in the South Pacific. International Labour Review 133 (3), 347–367.

Brown, R.P.C., Ahlburg, D., 1999. Remittances in the south Pacific. International Journal of Social Economics 26 (1/2/3), 325–344.

de la Brière, Bénédicte, Sadoulet, Elisabeth, Janvry, Alain de, Lambert, Sylvie, 2002. The roles of destination, gender, and household composition in explaining remittances: an analysis for the Dominican Sierra. Journal of Development Economics 68 (2), 309–328.

Chami, Ralph, Fullenkamp, Connel, Jajah, Samir, 2003. Are Immigrant Remittance Flows a Source of Capital for Development? IMF Working Paper 03/189. International Monetary Fund, Washington, DC.

Clark, Kenneth, Drinkwater, Stephen, 2001. An Investigation of Household Remittance Behavior. University of Manchester Discussion Paper no. 0114.

Clemens, Michael, Ogden, Timothy, 2013. Migration as a Strategy for Household Finance: A Research Agenda on Remittances, Payments, and Development. Financial Access Initiative Working Paper, October 2013.

Clemens, Michael, Tiongson, Erwin, 2013. Split Decisions: Family Finance when a Policy Discontinuity Allocates Overseas Work. CGD Working Paper 324. Center for Global Development, Washington, DC.

Cox, Donald, Eser, Zekeriya, Jimenez, Emmanuel, 1998. Motives for private transfers over the life cycle: an analytical framework and evidence for Peru. Journal of Development Economics 55 (1), 57–80.

Cox-Edwards, Alejandra, Ureta, Manuelita, 2003. International migration, remittances, and schooling: evidence from El Salvador. Journal of Development Economics 72 (2), 429–461.

De Haas, H., 2010. Migration and development: a theoretical perspective. International Migration Review 44 (1), 227–264.

DeSipio, Louis, 2000. Sending Money Home…For Now: Remittances and Immigrant Adaptation in the United States. Tomas Rivera Policy Institute and Inter-American Dialogue, Washington, DC.

Docquier, Frédéric, Rapoport, Hillel, 2006. The economics of migrants' remittances. Handbook of the Economics of Giving, Altruism and Reciprocity 2, 1135–1198.

Durand, Jorge, Kandel, Willia, Parrado, Emilio A., Massey, Douglas S., 1996. International migration and development in Mexican communities. Demography 33 (2), 249–264.

Dustmann, Christian, Kirchkamp, Oliver, 2002. The optimal migration duration and activity choice after re-migration. Journal of Development Economics 67 (2), 351–372.

Faini, Ricardo, 2006. Migration and Remittances: The Impact on the Countries of Origin, Mimeo, University of Rome. Available at: http://www.eudnet.net/download/Faini.pdf.

Gibson, J., Mckenzie, D., Stillman, S., 2010. How important is Selection? experimental vs. non-experimental measures of the income gains from migration. Journal of the European Economic Association 8 (4), 913–945.

Gibson, J., Mckenzie, D., Stillman, S., 2011. The impacts of international migration on remaining household members: omnibus results from a migration lottery program. Review of Economics and Statistics 93 (4), 1297–1317.

Gibson, J., McKenzie, D., Stillman, S., 2013. Selectivity and the estimated impact of emigration on incomes and poverty in sending areas: evidence from the Samoan quota migration lottery. Economic Development and Cultural Change 61 (2), 248–280.

Giuliano, Paola, Ruiz-Arranz, Marta, 2005. Remittances, Financial Development, and Growth. IMF Working Paper 05/234. International Monetary Fund, Washington, DC.

Groenewold, George, Bilsborrow, Richard, 2004. Design of Samples for International Migration Surveys: Methodological Considerations, Practical Constraints and Lessons Learned from a Multi-country Study in Africa and Europe. Population Association of America (General Conference).

Gubert, Flore, 2002. Do migrants insure those who stay behind? Evidence from the Kayes area. Oxford Development Studies 30 (3), 267–287.

Hoddinott, John, 1994. A model of migration and remittances applied to western Kenya. Oxford Economic Papers 46 (3), 450–475.

Ilahi, Nadeem, Jafarey, Saqib, 1999. Guestworker migration, remittances and the extended family: evidence from Pakistan. Journal of Development Economics 58 (2), 485–512.

International Monetary Fund, 2013. Balance of Payments and International Investment Position Compilation Guide (BPM6 CG). Washington, DC.

Massey, Douglas, Parrado, Emilio A., 1998. International migration and business formation in Mexico. Social Science Quarterly 79 (1), 1–19.

McCormick, Barry, Wahba, Jackline, 2001. "Overseas work experience, savings and entrepreneurship amongst return migrants to LDCs. Scottish Journal of Political Economy 48 (2), 164–178.

Menjivar, Cecilia, DaVanzo, Julie, Greenwell, Lisa, Burciaga Valdez, R., 1998. Remittance behavior among Salvadoran and Filipino immigrants in Los Angeles. International Migration Review 32 (1), 97–126.

Mesnard, Alice, 2004. Temporary migration and capital market imperfections. Oxford Economic Papers 56 (2), 242–262.

Miotti, Luis, Mouhoud, El Mouhoub, Oudinet, Joel, 2009. Migrations and Determinants of Remittances to Southern Mediterranean Countries: When History Matters. Paper presented at the 2nd Migration and Development Conference, Washington, DC, September 10–11.

Mishra, Prachi, 2005. Macroeconomic impact of remittances in the caribbean. International Monetary Fund Working Paper.

Osili, Una Okonkwo, 2004. Migrants and housing investment: theory and evidence from Nigeria. Economic Development and Cultural Change 52 (4), 821–850.

Orozco, Manuel, Fedewa, Rachel, December 2006. Leveraging Efforts on Remittances and Financial Intermediation. Inter-American Development Bank working paper 24.

Poirine, Bernard, 1997. A theory of remittances as an implicit family loan arrangement. World Development 25 (5), 589–611.

Ratha, Dilip, Eigen-Zucchi, Christian, Plaza, Sonia, Wyss, Hanspeter, Yi, Soonhwa, October 2, 2013. Migration and Remittance Flows: Recent Trends and Outlook, 2013–2016. Migration and Development Brief 21, Development Prospects Group, World Bank.

Stark, Oded, 1995. Altruism and beyond. Basil Blackwell, Oxford and Cambridge.

Stillman, Steven, McKenzie, David, Gibson, John, 2009. Migration and mental health: evidence from a natural experiment. Journal of Health Economics 28 (3), 677–687.

Taylor, J. Edward, Rozelle, Scott, de Brauw, Alan, 2003. Migration and incomes in source communities: a new economics of migration perspective from China. Economic Development and Cultural Change 52 (1), 75–101.

United Nations, 2011. Trends in International Migrant Stock: Migrants by Age and Sex (United Nations database, POP/DB/MIG/Stock/Rev.2011). Department of Economic and Social Affairs, Population Division.

Woodruff, Christopher, Zenteno, Rene, 2007. Migrant networks and microenterprises in Mexico. Journal of Development Economics 82 (2), 509–528.

World Bank, 2006. The Development Impact of Workers' Remittances in Latin America, vol. 2. World Bank, Washington. Detailed Findings, Report No. 37026.

Yang, Dean, 2008a. Coping with disaster: the impact of hurricanes on international financial flows, 1970-2002. B.E. Journal of Economic Analysis and Policy 8 (1) (Advances), Article 13.

Yang, Dean, Claudia Martinez, A., 2005. Remittances and poverty in migrants' Home areas: evidence from the Philippines. In: Ozden, Caglar, Schiff, Maurice (Eds.), International Migration, Remittances, and the Brain Drain. World Bank, pp. 81–122.

Yang, Dean, April 2008b. International migration, remittances, and household investment: evidence from Philippine migrants' exchange rate shocks. Economic Journal 118, 591–630.

Yang, Dean, Summer 2011. Migrant remittances. Journal of Economic Perspectives 25 (3), 129–152.

Yang, Dean, Choi, HwaJung, 2007. Are remittances insurance? Evidence from rainfall shocks in the Philippines. World Bank Economic Review 21 (2), 219–248.

International Migration by Ethnic Chinese

Emmanuel Ma Mung, CNRS, University of Poitiers, Poitiers, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The international migration of ethnic Chinese started to be numerically significant more than a century ago. It continued into contemporary times, mainly from continental China, but also from numerous countries in Southeast Asia where the first migrants settled. It is, therefore, not easy to assess the extent of this migration. However, their number can be estimated to 40–45 million. Two morphological characteristics lead to the definition of this migration as a 'diaspora': first, the multipolarity of the migration (migration toward several countries of destination) and second, the interpolarity of relations, that is, the fact that there are migratory, economic, and cultural ties between individuals who have settled in different countries. The diaspora-like character of this migration has led to the development of migratory, economic, and financial networks, which play a growing role worldwide, particularly in China and Southeast Asia.

The mass international Chinese migration began in the middle of the nineteenth century and today makes up 40–45 million people. It has evolved toward a diasporic configuration characterized by a multipolarity of migration and an interpolarity of relations in which networks play an important role and explain the economic efficiency of this diaspora. The revival of migrations from continental China has reinforced the diaspora and modified its composition and role.

The Migration from China

For centuries, or even millenniums, there has been a migration from China to Southeast Asia. These migrations mainly comprise an economic, political, or military urban elite. The Chinese migration really reached its peak in the middle of the nineteenth century when the combination of several factors made people leave. It was the result of the two Opium Wars, the Taiping Rebellion, the abolition of slavery, and the colonial enhancement of Southeast Asia by the English, the Dutch, and the French.

The two Opium Wars (1840–42 and 1856–60) and the Taiping Rebellion (1860–65) provoked tremendous social change and terrible famines, especially in the southern provinces, which are principal centers of departure even today. The progressive abolition of slavery during the first half of the nineteenth century brought about a demand for cheap labor in order to replace slave work in former slavery countries. But above all, the colonization of Southeast Asia and the development of new English, French, and Dutch colonies (Indochina, Indonesia, Malaysia, Singapore, and so on) also created a demand for workers in plantations (hevea, pepper plant, sugarcane, tobacco, and others) and in mines, a demand that was no longer filled by slavers or by local labor.

From then on, the geography of international Chinese migration was maintained for a long time, almost exclusively affecting the southern provinces of continental China and mainly spreading into Southeast Asia and into former slavery areas. This geography still plays a fundamental role in the general distribution of migration today, although there have also been substantial modifications. As a result of the treaties of the Opium Wars, the creation of the 'coolie trade' allowed colonial powers to hire workers in order to develop their dependent territories. The coolie trade produced flows of departure toward the Caribbean, the Indian Ocean islands, and Polynesia, but especially toward Southeast Asia. These workers were, for the most part, landless peasants. They were among the poorest and saw in these contracts, however miserable, the possibility for them and for their families to survive. The large majority of workers came from the southern provinces of China (Guangdong, Fujian).

Within 25 years, from 1876 to 1901, 4 850 000 emigrants, a considerable figure, are assumed to have passed in transit through ports in southern China. During that time, the number of people belonging to the Chinese diaspora was estimated at 8 million, most of whom were found in the Nan Yang areas (South China Sea) of Asia. Moreover, several dozens of thousands of people were scattered over the insular constellations of the Indian and Pacific Oceans, some hundreds of thousands went to the eastern rim of the Pacific Ocean, from Canada to Chile, and also to the Caribbean islands, in particular to Cuba and Jamaica.

Chinese International Migration Today

The ancientness of the migration, interracial marriages, the constant flow of migration, and the fact that the ethnic Chinese usually take the nationality of their receiving countries make an accurate measure of the extent of the diaspora difficult, because it is hard to determine what an 'ethnic Chinese' is. What differentiates ethnic identity from other social identities (professional, religious, and so on) is, as Max Weber states, the shared belief in a common origin, whether it is real or imagined. Ethnic identity is, therefore, a social construction not a natural one. It is still a valid criterion for defining a population because this belief strongly configures the relationships among individuals themselves and thus determines specific social behaviors with important effects on economic, political, and cultural patterns.

Bearing these considerations in mind, one can estimate the world distribution of ethnic Chinese from various sources (national censuses; Institute of Chinese Studies of Ohio State

University, Records of Overseas Compatriot Population Distribution of the Republic of China, press articles, research articles). From these numerous and diverse estimates, one can consider the number of ethnic Chinese in the world to be around 40–45 millions (see Table 1).

Table 1 Distribution of ethnic Chinese in 2010 (estimated in thousands, countries of more than 10 000 ethnic Chinese)

Asia	30 000–35 000
Thailand	9400
Indonesia	8800
Malaysia	6960
Singapore	2800
Vietnam	1300
Philippines	1150
Myanmar	1100
South Korea	690
Japan	660
Cambodia	340
Kazakhstan	300
Laos	190
Brunei	70
Kyrgyzstan	60
Israel	15
America	6000–6500
United States	3380
Canada	1610
Peru	500
Brazil	200
Argentina	120
Panama	50
Venezuela	50
Jamaica	30
Mexico	23
Costa Rica	15
Guatemala	15
Cuba	10
Europe	2300–2800
Federation of Russia	990
France	400
United Kingdom	400
Italy	180
The Netherlands	140
Spain	145
Germany	79
Serbia	30
Austria	30
Belgium	30
Portugal	20
Ireland	16
Hungary	15
Switzerland	13
Oceania	900–1000
Australia	670
New Zealand	150
French Polynesia	15
Papua New Guinea	10
Samoa	10
Africa	400–600
South Africa	250

(Continued)

Table 1 Distribution of ethnic Chinese in 2010 (estimated in thousands, countries of more than 10 000 ethnic Chinese)—cont'd

Angola	70
Nigeria	50
Madagascar	50
Algeria	30
Mauritius	27
Réunion	25
Zimbabwe	10
Sudan	10
Total	40 000–45 000

Sources: Various, including national censuses; Institute of Chinese Studies, Ohio State, University; Records of Overseas Compatriot Population Distribution, Republic of China; press articles; research articles.

Ethnic Chinese are present in nearly every part of the world. One can note the presence of more than 10 000 individuals in more than 55 countries, 100 000 in nearly 24 countries, and a million individuals in at least 9 countries. The majority (75%) lives in Asia, but their presence there is proportionally less important than it was 50 years ago (96%). There is now a redeployment of the diaspora, notably toward the nonAsian Pacific (North America, Australia), Europe, and Latin America and Africa.

Ethnic Chinese were estimated to be around 20 million at the beginning of the 1980s, the numbers of the Chinese diaspora would thus have doubled in about 30 years. This impressive increase is due to the natural growth of this population but especially by the revival of migrations from mainland China.

Revival of Migrations from Mainland China

The economic opening of China was accompanied with liberalization of conditions to emigrate and in 2006 more than 20 millions people has a passport. The annual number of Chinese migrants according to the census figures in Beijing increased rapidly from 57 000 in 1982 to 757 000 in 2000. The annual flow of entries of Chinese citizens in the countries of OECD rose from 140 000 in 1995 to 335 000 in 2000, and 539 000 in 2008. These flows of new migrants have affected the already settled groups.

Southern Chinese provinces – Guangdong, Fujian, and Zhejiang – are still the main source of migrants as evident from the large numbers of clandestine migrants going to Europe from these. But since the 1990s, migrants have been arriving from new places of origin, such as great urban areas like Beijing, Tianjin, and Shanghai, but above all, the Dongbei region comprising the northeast provinces of Heilongjiang and Liaoning, and Hubei in central China. One characteristic of this migration flow is the increasing number of female migrants.

A new migratory phenomenon linked to the increase in the level of education in the country is the emigration of skilled and high-skilled people especially of students. The Chinese authorities encouraged during the 1990s students to study abroad. In 2001, around 150 000 of them went abroad, mainly

to the Western countries. And at the beginning of the 2000s, they were 460 000, half of them were in the United States and 20% in Europe.

Migrations from China or from other settling countries of the Chinese diaspora are oriented more and more to the developed countries. According to national censuses, from the beginning of the 1990s to the mid-2000s the number of Chinese in overseas increases of 1.7 million (from 1.6 million to 3.3 million) in the United States and of 1 million in Canada. In Japan, it is multiplied by four to reach 610 000, and in Western Europe (France, United Kingdom, Italy, Spain) the number increase from 361 000 to 783 000, and Australia now has 670 000. New settling areas have appeared, such as the Central and Eastern European countries. One can find there several tens of thousands of Chinese migrants, and in Eastern Siberia several hundreds of thousands of them, including farmers who cultivate land deserted by their Soviet predecessors. Other new destinations are Central Asia, and more recently Africa and Latin America.

Since the 1990s, temporary and contractual migrations developed in relation with Chinese companies working abroad and involved in the exploitation of raw materials (oil, ores, wood, etc.) and the construction of infrastructures for that purpose. In addition, they are engaged in large scale projects involving the building of equipments in public works and telecommunications and the distribution of electricity in particular in Africa. These new type of migrations are also carried out by Chinese agencies, which sale this labor force to companies in different countries through bilateral agreements in different parts of the world. In 2004, there were 597 000 such migrants, the large majority of them were in Asia and Africa.

Over the last decade the migration of merchants has developed rapidly worldwide, reflecting the development in the production of consumer goods in China. They distribute products 'made in China' consisting of textile, bicycles, motor cycles, tools, electric household appliances, electronic goods, data-processing equipments, computer, etc. One can find these new Chinese merchants in Western Europe (mainly in France, Italy, and Spain), in Central and Eastern Europe (notably Romania, Bulgaria, Hungary, and the Czech Republic), etc. In Africa, the phenomenon is highly visible.

Mostly, this merchant migration is resulted by the opening of wholesale centers to supply retailers, whether Chinese or native. In Aubervilliers a town near Paris, several hundreds of Chinese wholesalers are settled in a dedicated area. The same is true of Italy (Naples) and in Spain (Madrid). The same scene is repeated in South Africa, where one of the largest wholesale centers is located in Johannesburg, Nigeria, Ghana, Cameroon, Togo, Senegal. Apart from these wholesalers, one also finds many retailers established in the main cities. Some have also settled in cities of intermediate sizes or in small towns.

The Chinese Diaspora

The existence and the development of migratory, economic, informational, but also affective and emotional relations between the various poles of settlement, at the local and global scale, circumscribe a social entity of which the unit is given by the subjective belief of the people in a common origin – China – which is the base of a collective identity. It is in this sense that one can speak of Chinese diaspora. The Chinese overseas constitute a diaspora, which is characterized – morphologically – by the multipolarization of the migration (installation in several countries) and the interpolarity of the relations; i.e., the existence of relations between the various places of settlement.

This diaspora is 'entrepreneurial' in the sense that enterprises ensure its reproduction as social body. Companies play a crucial role with regard to employment. The majority of the ethnic Chinese work in these companies and these companies employ mainly ethnic Chinese.

Chinese migrations gave rise in most settling countries to an economic organization made up by small firms articulated between them on the local scale in the fields of supplying, financing, or labor. They form kinds of local economic clusters, often related ones to others at international scale.

This economic organization as 'entrepreneurial diaspora' has important effects on Chinese migrations, because these movements meet more the demand for labor of the ethnic Chinese companies than the demand of the general labor market. The multiplication of these companies encourages an immigration that contributes to their development. In return, it reinforces the labor demand, and consequently migrations.

Multipolarity of the Migration

Since the nineteenth century, Chinese migration has had a pronounced multipolar feature because of its direction toward many territories in Southeast Asia, the American continent, Polynesia, the Indian Ocean, the Caribbean, South Africa and, in a more modest way, in Europe. Throughout the twentieth century, a change in the destinations and a reinforcement of the Chinese presence was noticeable. As previously stated, the presence of more than 10 000 overseas Chinese in each of more than 55 countries is recorded, with more than 1 million people in at least nine countries. One of the main features of this distribution is therefore a strong multipolarity of the migration. The other feature of the international Chinese migration is a strong interpolarity of relations.

Interpolarity of Relations

Interpolar migrations are often linked to major geographical trouble. The wars in Vietnam, Laos, and Cambodia, and the political takeover by totalitarian regimes that followed them, caused the departure of more than 3 million people, most of whom where first or later generations of Chinese origin. For the most part, they joined already established communities in Europe, North America, Asia, or Australasia by means of diverse itineraries, sometimes passing through various receiving countries. During the 1990s in Indonesia, where more than 8 million ethnic Chinese are recorded, demonstrations against the government and President Suharto have at times degenerated into anti-Chinese pogroms. From 100 000 to 125 000 ethnic Chinese are thought to have left Indonesia within 1 year, for the most part heading toward Singapore, Malaysia, Australia, or even Hong Kong. Fortunately, all migration is not due to

dramatic events, but is rather a search for better opportunities elsewhere. This is the case for the Chinese who settled in Italy or in Spain, among whom a good part came from France in the 1980s and 1990s because the French government immigration laws became ever stricter, and because of the relatively greater freedom in the new receiving countries at this moment. It is worth noting that these Chinese also came to join small communities that had already been settled for several decades. This also happened between France and the Netherlands during the period between the two world wars.

The diversity of the national origins of the Chinese within the receiving countries is a good indicator of the interpolarity of the migrations. In France, and today also in Spain and in Great Britain, more than 20 different national origins are recorded among the ethnic Chinese, with even a greater number in North America: nearly 30 in Montreal alone, for instance, and 40 in the United States.

These interpolar migrations are the basis of relations between the different settling poles. They represent the networks of movement of persons, materials, and nonmaterial values (goods, money, information, social norms and codes, cultural models). The classic example is the establishment of trading networks between China and the different receiving countries of the diaspora. These networks have been strengthened considerably since the opening of the Chinese economy. On a more subtle scale, such migration represents an extension of distributional networks in the various areas of settlement in a given country, as long as the migration spreads. The most interesting point is the growth of trading networks among the different receiving countries. The way a diaspora works, it induces a diversification of the geographical areas of supply and distribution. In grocery stores, to take a simple example, specialized products came from several dozen different countries in addition to continental China, namely from Taiwan, Singapore or Hong Kong, Hawaii, Thailand, Brazil, and even Spain and California. A study made out in an Asian supermarket in Paris revealed that the products came from 37 different countries that had an ethnic Chinese population. A similar study from a Chinese supermarket in San Diego (California) revealed more than 40 different countries of origin.

The existence of interpolar relations can allow people engaged in business to move from one country to another when working conditions worsen. This explains why Sino–Indonesian capital was particularly withdrawn toward Singapore and Thailand in the 1960s and 1970s. This is also the reason why today there is a shift in capital from Southeast Asia to the American west coast in order to take advantage of economic opportunities (financial flow from Taiwan or Singapore) and vice versa.

Decisive Economic Effects

Southeast Asia makes up 75% of the number of individuals in the Chinese diaspora. The presence of these immigrants has asserted itself in the nineteenth century already, and it has been reinforced constantly ever since. At first, this immigration included labor force for mines and plantations, but it shifted quickly to small shopkeepers, craftsmen, and sometimes to manufacturers and businessmen. Already at the beginning of the twentieth century, the rice business and commercial distribution were operated by these immigrants or their descendants even in the most remote villages. Since then, their presence in economic activities has constantly been reinforced. In some Southeast Asian countries like Indonesia, Philippines, Thailand, Malaysia, and especially Singapore, the ethnic Chinese control is an important part of the private economy, the country trade, and the financial investments.

Since the opening of China, the Chinese diaspora has increasingly had important economical relations with this country. This has contributed in a very decisive way to the spectacular development of this country since the mid-1980s. Practically disappeared from the 1950s, the remittances of emigrants started again in the 1980s and increased in an impressive way: $US175 millions in 1990, $US5.2 billions in 2000, and $US51.3 billions in 2010, according to the World Bank. In these remittances are added an important part of the foreign direct investments (FDI) in China realized by companies and financial groups belonging to ethnic Chinese. In 2001, the diaspora would have contributed for 63% of FDI ($US30 billions), and in 2004, for 60% of FDI ($US36 billion). The economic takeoff of China is thus widely connected to these transfers and the most economically developed provinces are also the ones of strong migratory tradition.

Conclusion

The international migration of ethnic Chinese, which started in the mid-nineteenth century, has evolved into a diaspora-type configuration. Today, this migration constitutes the most important population in a diaspora (40–45 million people). Networks play an essential role in such a diaspora. Even if today it seems too powerful and rich, one should remember that originally it was composed of labor migration, and that this remains an element to this day. Current migrants are still mainly proletarians.

This migration, like other migrations, form new social groups – the diasporas – settled on transnational territory. A diaspora is based on an identity of ethnic features. This ethnic consideration does not exclude other elements. Studies of the Chinese diaspora constantly remind the investigator of the singularities of the individuals who have different national and/or geographical origins and speak many different languages. In everyday life, ethnic Chinese meet fellowmen who are nevertheless different. Therefore, an 'internal' cosmopolitanism is at work in the diaspora, which guarantees multiple allegiances across the different countries where the diaspora has been established, as well as with China.

See also: Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Global Migration; Immigration Policy; Immigration and Crime in the United States; Immigration, Applied Research on; Immigration, Citizenship, and Integration: Social Work Connections; Immigration, Emigration, and Citizenship; Immigration: Health and Mortality; Immigration: Political Aspects; Immigration: Social Psychological Aspects; Migration and Development; Migration and Health; Migration and Social Capital; Migration, Economics of; Migration, Ethnicity, Aging and Social Work Practice; Migration:

Anthropological Perspectives; Migration: Cultural Aspects; Migration: Motivations; Migrations, Colonizations, and Diasporas in Archaeology.

Bibliography

Chen, D., 1923. Chinese Migrations, with Special Reference to Labor Conditions. Bulletin No. 340. United States Bureau of Labor Statistics, Washington, DC.

Guerassimoff, C., 2012. La Chine et sa nouvelle diaspora. La mobilité au service de la puissance. Ellipses, Paris.

Kuah-Pearce, K.E., Davidson, P. (Eds.), 2008. At Home in the Chinese Diaspora: Memories, Identities and Belongings. Palgrave Macmillan, Richmond, UK.

Kuhn, P.A., 2008. Chinese Among Others. Emigration in Modern Times. National University of Singapore, Singapore.

Ma Mung, E., 2000. La diaspora chinoise: géographie d'une migration. Ophrys, Paris.

Nyiri, P., Saveliev, I.R., 2002. Globalizing Chinese Migration: Trends in Europe and Asia. Ashgate, Aldershot, Hampshire, England; Burlington, VT.

Ong, A., 1999. Flexible Citizenship: The Cultural Logics of Transnationality. Duke University Press, Durham, NC.

Pan, L. (Ed.), 1999. The Encyclopedia of Chinese Overseas. Curzon Press, Richmond, UK.

Pieke, F., Nyiri, P., Thuno, M., 2004. Transnational Chinese: Fujianese Migrants in Europe. Stanford University Press, Stanford.

Poston, D.L., Yu, M.Y., 1990. The distribution of the overseas Chinese in the contemporary world. International Migration Review 24 (3), 480–508.

Suryadinata, L. (Ed.), 2011. Migration, Indigenization and Interaction: Chinese Overseas and Globalization. World Scientific Publishing Company, Singapore.

Tan, C.B. (Ed.), 2013. Routledge Handbook of the Chinese Diaspora. Routledge, New York.

Thuno, M. (Ed.), 2007. Beyond ChinaTown: New Chinese Migration and the Global Expansion of China. Nordic Institute of Asian Studies, Copenhagen, p. 280.

Wang, G., 1992a. Community and Nation. Allen & Unwin, St Leonards, UK.

Wang, G., 1992b. China and the Overseas Chinese. Times Academic Press, Singapore.

Zhou, M., 2009. Contemporary Chinese America: Immigration, Ethnicity, and Community Transformation. Temple University Press, Philadelphia.

International Migration by Ethnic South and Southeast Asians

Graeme Hugo, The University of Adelaide, Adelaide, SA, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

South and Southeast Asia are currently home to one-third of the world's population and as such are one of the main global reservoirs of potential international migrants in the contemporary context of rapid globalization. Moreover, within the region the last two decades have seen international migration become part of the calculus of choice of an increasing proportion of residents in the region due to education, internationalization of labor markets, and transformation of transport and communication systems. However, while the region contains several of the world's largest 'labor surplus' nations and hence potentially (as well as actually) several of the globes' major source areas of international migrants, the patterns of international movement influencing the region are more complex than a simple 'south–north' flow. Although the latter is significant, it has to be recognized not only that south–north migration takes various forms and is a complex two-way process, but there are also important migration flows within the Asian region. The types of migration examined include permanent resettlement, refugee movement, labor migrations, student migrations, and undocumented migrations. The article summarizes these patterns and then proceeds to examine some of the characteristics of the migrants themselves. Here, attention is focused on the increasing levels of involvement of women in the movement and their selectivity in terms of skill and education. It is pointed out that while there are important flows of skilled people, there are also major flows of unskilled migrants (especially migrant workers) out of the labor surplus nations of the region. The relevance of neoclassical economic theories of migration is assessed and such a theory is seen to be only partial in its explanation. Other elements considered to be important are the proliferation of social networks and the development of an immigration industry in the region. Government in both sending and receiving nations are also becoming increasingly influential in shaping international migration. A number of issues relating to international migration in the region are then addressed. These include the scale and impact of remittances sent home by migrants from the region who are abroad, issues relating to the protection and vulnerability to exploitation of migrants, the role of intermediaries in the migration process, the level of trafficking, people smuggling, and undocumented migration in the region, the brain drain, the relationship between migration and economic development, and the formation of ethnic enclaves in destination countries.

Introduction

South (including the countries of India, Pakistan, Bangladesh, and Sri Lanka) and Southeast (the countries of Indonesia, the Philippines, Thailand, Vietnam, Singapore, Malaysia, Brunei, Cambodia, Laos, and Myanmar) Asia were home to 33.1% of the world's population in 2013 (2367.8 million persons) (United Nations, 2013a). Moreover, the region has experienced massive population change over recent decades (Williams and Guest, 2012; Veron, 2008). The rapid increase in population movement between countries that has accompanied the process of globalization has affected all of the countries of the region. There has been a major increase in the levels, complexity, and impact of international migration. While two decades ago mobility was limited to a selected small proportion of residents in the region, it is now within the calculus of choice of a major part of the population as they weigh up their life chances. With the increase in movement has come an increased awareness of it, and intervention to influence it, by governments in the region. This article seeks to summarize the evolving patterns of movement and address some of the major issues relating to it.

The increasing significance of international migration of South and Southeast Asians is evident in Table 1. This indicates that while the world's population currently live in South and Southeast Asian countries – 2367.8 million people – a further 55 million people born in these countries are currently living in different countries. The largest diaspora (14.17 million) is of Indians, while the countries where the diaspora is largest in relation to their resident population are Laos, Afghanistan, Brunei, and Bhutan. These measurements of diaspora are, however, partial because

- they exclude second and later generations born in destination countries,
- they only include countries of destination, which collect information on the foreign-born in their census.

Hence, estimates of the actual size of South and Southeast Asian diaspora are generally substantially larger than those in Table 1.

Patterns

The data available to examine international migration in South and Southeast Asia are limited. The increase in significance and intensity of movement has outpaced the development of data systems to measure it. Moreover, undocumented migration in the region is arguably as great in scale as legal migration, and by definition, the information relating to that movement is very limited. While there is a great deal of complexity in the spatial patterning of the movement, the degree of permanency, the characteristics of movers and its legality, it is possible to identify three main subsystems of movement in the region. These include a so-called south–north, more or less permanent, flow to developed countries, a predominantly temporary labor flow

Table 1 South and Southeast Asia: migration stocks in other countries, 2013

Country of origin	*Migrant stocks in other countries*	*Residents in origin*	*Migrants as % of residents*	*Migrants as % of global total*
Afghanistan	5 109 921	30 552 000	16.73	2.21
Bangladesh	7 757 315	156 595 000	4.95	3.35
Bhutan	90 116	754 000	11.95	0.04
Brunei Darussalam	50 894	418 000	12.18	0.02
Cambodia	1 115 567	15 135 000	7.37	0.48
India	14 166 558	1 252 140 000	1.13	6.12
Indonesia	2 992 550	249 866 000	1.20	1.29
Iran (Islamic republic of)	1 061 005	77 447 000	1.37	0.46
Lao people's democratic republic	1 293 447	6 770 000	19.11	0.56
Malaysia	1 445 890	29 717 000	4.87	0.62
Maldives	1327	345 000	0.38	0.00
Myanmar	2 647 960	53 259 000	4.97	1.14
Nepal	1 044 707	27 797 000	3.76	0.45
Pakistan	5 687 563	182 143 000	3.12	2.46
The Philippines	5 486 902	98 394 000	5.58	2.37
Singapore	303 544	5 412 000	5.61	0.13
Sri Lanka	1 252 482	21 273 000	5.89	0.54
Thailand	894 002	67 011 000	1.33	0.39
Timor-Leste	39 438	1 133 000	3.48	0.02
Vietnam	2 604 992	91 680 000	2.84	1.13
Total South and Southeast Asian stocks	55 046 180	2 367 841 000	2.32	23.78

Sources: United Nations, 2013b. Trends in International Migrant Stock: Migrants by Destination and Origin. United Nations database, POP/DB/MIG/Stock/Rev.2013; UNESCAP, 2013. 2013 ESCAP Population Data Sheet. UNESCAP.

to the Middle East, and a third set of 'south–south' interactions within the region. Each subsystem involves different balances of forced and nonforced movement, males and females, permanent and nonpermanent, documented and undocumented, labor related and other movement, etc.

South–north movement has recently become the most dominant global international migration flows (United Nations, 2013c), and South and Southeast Asia is one of the dominant origin areas of such flows. In the largest nation in the region, the Indian government in May 2012 estimated the size of its diaspora at 21.9 million. (Source: Ministry of Overseas Indian Affairs, http://moia.gov.in/writereaddata/pdf/NRISPIOS-Data(15-06012) new pdf (accessed 08.11.13.).) However, other countries in the region also have large diasporas and many of them are in OECD countries as a result of so-called south–north migration. **Table 2** presents estimates of the numbers of South and Southeast Asian migrants in more developed countries. These are underestimates of the impact of immigration because the data are obviously incomplete; undocumented migrants are excluded and the children born overseas to migrants and later generations are not included. Nevertheless, the scale of the movement can be appreciated.

However, more or less permanent settlement in North countries is not the only type of movement out of South and Southeast Asia. International labor migration whereby workers from the region move to other countries to work, usually under contract, for a set period (usually around 2 years) before returning to their home country has expanded in the last two decades. One of the major streams of this type out of the region is directed toward the Middle East. There has been a long tradition of South Asians working in that region, which extends back to colonial times; but the steep increase in oil prices after 1973 saw massive flows of foreign exchange into the region that fostered massive infrastructure construction, which in turn saw the expansion of recruitment of foreign workers initially

Table 2 Reported numbers of Southeast- and South Asia-born populations in 'north countries'

Origin	*USA (2011)*	*Canada (2011)*	*Australia (2011)*	*Japan (2010)*	*New Zealand (2006)*	*Europe (2013)*
Southeast Asia						
Myanmar	95 726	8255	21 760		843	20 828
Thailand	239 942	14 985	45 464	41 279	6159	262 864
Laos	183 768	14 540	9932		897	47 026
Cambodia	160 212	24 275	28 329	2683	5853	72 210
Vietnam	1 259 317	168 420	185 039	41 781	4875	431 586
The Philippines	1 813 597	502 295	171 233	21 081	15 285	457 750
Malaysia	58 268	23 710	116 196	8364	14 547	78 651
Singapore	30 608	12 190	48 647	2512	4857	54 030
Indonesia	97 244	14 870	63 159	24 895	4614	185 512
Total (inc. Other)	3 939 341	787 600	701 790	331 695	58 257	1 612 280
South Asia						
India	1 856 777	572 435	295 363	22 497	43 344	1 155 737
Pakistan	303 915	161 380	30 221	10 299	2211	788 484
Bangladesh	184 469	47 180	27 808	10 175	1437	382 415
Sri Lanka	45 762	135 400	86 412	9097	7257	383 554
Nepal	73 528	9050	24 636		579	63 976
Total	2 464 451	925 445	464 440	52 068	47 571	2 774 166

Sources: 2011 US Census Bureau American Community Surveys; New Zealand 2006 Census; ABS 2011 Census; Statistics Canada 2011 Census; Statistics Japan, Ministry of Internal Affairs and Communication; United Nations, 2013b. Trends in International Migrant Stock: Migrants by Destination and Origin. United Nations database, POP/DB/MIG/Stock/Rev.2013.

from South Asia but later from Southeast Asia. In the late 1980s, this movement changed to include more women and service workers. Again, data are poor since many Middle Eastern countries do not release information on migrant workers due to the sensitivity attached to heavy reliance on foreign labor in several countries where these workers in some cases outnumber local workers. Nevertheless, Table 3 presents some information for the main labor-exporting countries. The bulk of migrant workers from Southeast Asia are unskilled, while South Asia has sent both skilled and unskilled workers. Female domestic servants have become more significant in the movement over time, especially those from Indonesia, the Philippines, and Sri Lanka.

Table 3 Estimates of migrant workers from South and Southeast Asia deployed to the Middle East

Country of origin	*Saudi Arabia*	*Other*	*Total*	*Year*	*Source*
Southeast Asia					
Thailand			25 054	2011	ADBI and OECD, 2013
The Philippines	293 049	368 426	661 475	2010	Baruah, 2013
Indonesia	289 297	313 862	603 159	2011	Baruah, 2013
South Asia					
India	137 643	72 238	209 881	2011	ADBI and OECD, 2013
Pakistan	138 495	283 163	421 658	2008	ADBI and OECD, 2013
Bangladesh	15 039	445 140	460 179	2011	ADBI and OECD, 2013
Sri Lanka	70 896	172 870	243 766	2010	ADBI and OECD, 2013
Nepal	71 116	169 706	240 822	2011	Baruah, 2013

In the 1980s and 1990s, there was a growing tempo of movement of international labor migrants to Asian destinations, including other countries in the Southeast Asian region. This is evident in Table 4, which consolidates estimates of the stocks of overseas migrant workers from the various nations of South and Southeast Asia. These indicate that for many the main destination is another Asian nation. In fact, several of the East and Southeast Asian nations, which began their transitions to low fertility in the 1980s and 1970s, have suffered significant labor shortages in recent years. This has been a function of both rapid economic expansion and the declining numbers of new entrants into their workforces due to low fertility. This has been accompanied by a distinct segmentation of their labor forces whereby local workers have eschewed entry into low-income, low-status jobs, which have created niches for overseas migrant workers in these areas. Hence particular industries in particular Asian countries have become quite dependent on foreign labor. These include some activities in countries like Thailand and Malaysia, which themselves are supplying migrant workers to Japan and Taiwan. Hence in Thailand the rice milling and fishing industries are heavily reliant on labor from Myanmar, while in Malaysia the plantation and forestry

Table 4 South and Southeast Asian countries: Estimates of stocks of migrant workers in other countries

Origin countries	*Number*	*Main destinations*	*Source*	*Year*
Southeast Asia				
Brunei	24 300	Australia, Canada, the UK, the US, Germany, Ireland, Malaysia, the Netherlands, and New Zealand	Scalabrini Migration Center, 2012	2010
Myanmar	3 000 000	Thailand, the US, India, Malaysia, Australia, the UK, Japan, Canada, Korea, and Germany	*Migration News*, April 2013	2013
Thailand	810 000	Saudi Arabia, Taiwan, Myanmar, Singapore, Brunei, Malaysia, Cambodia, the US, Germany, Japan, Korea, Australia, and the UK	Scalabrini Migration Center, 2012	2010
Laos	366 600	Thailand, the US, France, Australia, Canada, Japan, Germany, Belgium, Cambodia, New Zealand	Scalabrini Migration Center, 2012	2010
Cambodia	350 400	Thailand, the US, France, Australia, Canada, and New Zealand	Scalabrini Migration Center, 2012	2010
Vietnam	2 200 000	Korea, Japan, Malaysia, the US, Australia, Canada, Cambodia, Germany, France, Thailand, and the UK	Scalabrini Migration Center, 2012	2010
The Philippines	10 000 000	Middle East, Malaysia, Japan, the US, Canada, Australia, Italy, and the UK	*Migration News*, July 2012	2012
Malaysia	1 480 000	Singapore, Australia, Brunei, the UK, the US, Canada, NZ, India, Japan, and Germany	Scalabrini Migration Center, 2012	2010
Indonesia	6 500 000	Malaysia, Saudi Arabia, Taiwan, Singapore, South Korea, and the United Arab Emirates	*Migration News*, April 2013	2013
Total	24 731 300			
South Asia				
India	7 500 000	Middle East	*Migration News*, October 2011	2011
Pakistan	3 973 000	Middle East, Malaysia	Arif, 2009	2004
Bangladesh	8 000 000	Saudi Arabia, Malaysia	*Migration News*, April 2012	2012
Sri Lanka	1 700 000	Middle East, Italy, Singapore, South Korea	*Migration News*, July 2013	2013
Nepal	3 500 000	Middle East, India, Malaysia	*Migration News*, January 2012	2011
Total	24 673 000			

industries rely upon gaining workers from Indonesia, the Philippines, and Bangladesh.

It is apparent then that one can recognize three distinct subsystems in South and Southeast Asian migration. One has involved movement to the so-called northern nations; another links the region to the Middle East; while a third involves intra-Asia movement. While, permanent settlement in the case of the first subsystem and labor migration in the case of the second, two are dominant, there are a number of other forms of international movement involved in the subsystems. Dealing first with the south–north movement, a most important factor was the abolition of racially biased immigrant selection policies in the main immigration countries (United States, Canada, Australia, etc.) in the 1960s and 1970s. Immigration policies were developed in which selection was based upon skill/education and/or family linkages as well as acceptance of refugee/humanitarian migrants. Hitherto migrants from South and Southeast Asia could not compete on an equal basis for migrant places with people from Europe. In addition, migration flowed along linkages established by colonial rule (e.g., Indians and Pakistanis to the UK, Indonesians to the Netherlands, etc.) or by military involvement (Koreans to the United States, etc.).

An important element in the south–north movement, as well as migration within Asia, has been that of the refugees. Table 5 shows the current stock of refugees in countries in the region. While there have been several such flows over the last half century, two have been on a particularly large scale and have differed substantially in type. The outflow of refugees from Indochina following reunification of Vietnam in 1975 was predominantly initially to other Asian (especially Southeast Asian) nations. This involved more than 2 million people the vast majority of whom have been settled in third countries (predominantly in the 'north') or repatriated back to their home countries. These movements have resulted in substantial subsequent migration since these refugee settlers functioned as anchors in destinations to bring other family members to these destinations, often using the family reunification components of migration programs.

The second major refugee migration occurred from Afghanistan, especially into the neighboring nations of Pakistan and Iran. This involved several million people most of whom were not settled in third countries or repatriated but stayed for long periods in their countries of temporary asylum. There have been many smaller but nevertheless significant movements, mostly involving movement into neighboring countries – Myanmar into Thailand, Tibet into India, Bangladesh into India, the Philippines into Malaysia, etc.

Another strand in the south–north migration has been the acceleration in the migration of students to undertake secondary but especially tertiary education. Table 6, for example, provides data on the percentage of foreign students from Asia studying in tertiary institutions in some northern countries. In Australia, for example, over the 1983–2013 period the number of overseas students increased from 9098 to 462 411 with almost half of the latter being from South and Southeast Asia. There is a strong connection between such movement and eventual settlement in the destination nation. Indeed, gaining an educational qualification in the 'north' country can facilitate entry as a permanent settler.

A related element in the intra-Asian subsystem involves the movement of trainees from labor surplus nations like Indonesia, Thailand, and the Philippines to countries with shortages such as Japan and Korea. The movement ostensibly is for workers of multinational companies in the former countries to spend a period in the home country of the companies to gain training. There has been discussion of the extent to which companies are using this system to import unskilled or semiskilled workers when government policy prevents such movement.

While the net effect of the south–north movement is emphatically in favor of the north, the flow in the opposite direction is not to be ignored. This in part involves former emigrants returning to their home country, but in the case of some Southeast Asian nations also involves natives of northern countries to meet the demands for skilled workers in areas like accountancy, engineering, management, etc., which cannot be met locally. Most of the latter will remain only temporarily in the Asian country. The fact that there is return migration into Asian countries from the north, however, is often neglected. It

Table 5 South and Southeast Asia: Refugees by country of asylum/origin, end 2012

	Number ('000)	
	Destination	*Origin*
Southeast Asia		
Cambodia	–	14.0
East Timor	–	–
Indonesia	1.8	10.1
Laos	–	8.0
Malaysia	89.2	0.5
Burma/Myanmar	–	215.3
The Philippines	0.1	0.1
Thailand	84.5	0.4
Vietnam	–	336.9
South Asia		
Afghanistan		2585.6
Bangladesh	30.7	10.2
Bhutan	–	41.6
India	185.7	14.3
Nepal	56.3	7.6
Pakistan	1638.5	33.6
Sri Lanka	0.1	132.8

Source: United Nations High Commissioner for Refugees (UNHCR), 2013. Displacement: The New 21st Century Challenge – UNHCR Global Trends 2012. UNHCR.

Table 6 Percentage of foreign students from Asia and the Pacific studying in tertiary institutions in some north countries

Country	*Year*	*Total number of foreign students*	*Percentage from Asia and the Pacific*
USA	2010	684 714	60.8
Australia	2010	271 231	78.0
Japan	2010	141 599	46.2
New Zealand	2010	37 878	79.0

Source: United Nations Educational, Scientific and Cultural Organization (UNESCO), 2013. Global Education Digest 2012. UNESCO, Paris.

often can take the form of a 'reverse brain drain' involving highly qualified people who return and take up important positions in the economy or government of their home country.

Obtaining estimates of international migration in the South and Southeast Asian regions is very difficult. Data collection systems are not generally well developed and there is a high level of undocumented migration. The highly complex nature of the movement needs to be stressed as does the fact that there are complex linkages between the different types of movement.

The Migrants

Migration is always a selective process in that the groups of people who move differ from those who are left behind or those which they join at the destination. Each of the three main subsystems briefly described above is distinctive not just in terms of the countries involved, but also the types of people who move. The south–north migration can generally be characterized as being selective of the better-educated and skilled populations since entry criteria in the destination countries usually require this, although 'business' migration programs involving the investment of minimum levels of personal wealth at the destination are increasingly common. Under family reunification and refugee/humanitarian programs and the United States Quota Program, there are some, but limited, opportunities for unskilled people to move but the overall selectivity remains. One characteristic of the south–north movement is the increasing involvement of women such that in several northern countries women immigrants from the region (especially Southeast Asia) outnumber their male counterparts.

The composition of the flows in the Asian and Middle East subsystems is quite different. While there are certainly exceptions (these include some of the flows out of South Asia and the Philippines), the out-movements are dominated by unskilled workers. These are directed at particular labor market niches at the destination, which due to low wages or low status are unable to attract local workers. Several of these flows are selective of women so that in some sending countries (e.g., Sri Lanka, Indonesia, and the Philippines) women outnumber men among the outgoing migrant workers. These women are especially directed to a number of specific occupations, which sometimes make them vulnerable to exploitation. These especially include work not only as domestics in private homes in destination countries but also in some cases are associated with the so-called entertainment industry. In some countries (e.g., the Philippines), it has been found that migrant workers sometimes take up jobs in the destination country, which are at a level below their qualifications.

The Process

There can be no doubt that the level of international population movement in South and Southeast Asia has increased substantially over the last two decades. While in the 1970s the possibility of international migration was only within the reach of relatively few people in the region, it is now within the calculus of choice of a much larger number of people and a wider range of socioeconomic, cultural, and gender groups. Clearly the whole process of globalization, especially as it has impinged upon labor markets, has been an important factor. This has greatly increased the linkages between nations involving flows of goods, finance, ideas, etc., and it is inevitable that movements of people also have increased although the barriers erected by nation states to flows of people have been reduced.

Neoclassical economic explanations have relevance for migration in the region. South and Southeast Asia contain four of the 10 largest labor surplus countries in the world. The demographic and economic gradients between richer and poorer countries both globally and within Asia have increased over the last two decades so that there are steep differences not only in income and wages between nations, but also in the rate of increase of the working age populations. Labor importing economies are undergoing economic transformation from manufacturing to service-based economies and this is creating jobs for migrant workers. Labor shortages in areas like information technology, management, and a number of skilled areas not only are encouraging the development of international labor markets, but also the labor market segmentation referred to earlier is creating opportunities for migrant workers in low skilled, low paid, and low status areas. The South and Southeast Asian region is able to supply workers both in the skilled areas (from South Asia, the Philippines) and the unskilled occupations.

Nevertheless, the widening gaps between nations in population and economic growth, and income and wage differentials are not sufficient explanation of the burgeoning of migration in the region although they are an important component of any comprehensive accounting for it. There are a number of other elements, which are also involved. Among the most important is the development of social networks between origin and destination nations. Empirical research indicates that the majority of movers from the region migrate to join family and friends who migrated earlier to the destination. A large proportion of South and Southeast Asians now have social capital in countries other than their own in the form of a family member or friend who lives and/or works in those countries. Those linkages form important conduits along which international migration, both documented and undocumented, takes place. The development of ever cheaper and better transport and communication systems enhances the operation of these networks in initiating migration.

Another element involves the so-called migration industry, which has a long history in the region but has proliferated in recent times. This involves a plethora of immigration agents, recruiters, travel agents, travel providers, immigration officials, labor brokers, and other intermediaries who are involved in initiating and facilitating international migration. Part of this infrastructure are the people smugglers and traffickers who transport migrants outside of government regulations but it also involves many who operate within the law.

Another element is the greater involvement of government in international migration in the region. While governments in destination countries are predominantly concerned with controlling the numbers and characteristics of immigrants, those in sending countries are increasingly becoming involved

in the emigration process. Realization that international migration can relieve under- and unemployment at home and enhance foreign exchange earnings has seen governments in several nations intervene to encourage emigration and to protect the rights of their nationals living and/or working in other countries. Some countries have been active in encouraging former migrants with needed skills to return, while others have policies and programs to encourage their nationals in other countries to invest their earnings in their home countries.

Issues

The proliferation of international migration in South and Southeast Asia has seen the emergence of a number of significant issues both in the countries of origin and destination as well as among the migrants themselves and their families. One of the most debated of these relates to remittances – those money earned by migrants abroad that are sent to or brought back to the country of origin. There is debate as to the scale and impact of these financial flows, but the data relating to them are largely incomplete. This is because much of the flow does not pass through official channels. Nevertheless, it is clear that the impact of remittances can be considerable. Table 7, for example, shows the significance of reported remittances in relation to foreign exchange earnings from trade in three Southeast Asian countries. While global figures such as those in Table 7 understate the scale and impact of remittances nationally, there can be no doubt that their impact within regions of South and Southeast Asian countries is much greater. This is because migrants tend to be drawn from particular *regions* rather than randomly across the nation. Hence remittances tend to be focused on particular parts of countries such as Kerala in India or Flores in Indonesia. Moreover, these regions are often among the poorest in the origin countries and the remittances represent one of the few sources of investment in those regions. For example, a December 1999 report from the State Planning Committee of Laos indicates that remittances from abroad represented 28% of all household earnings in the Vientiane Valley compared with 25% from agriculture, 22% from wages, and 18% from business (Lintner, 2000: 48). However, remittances have also been seen to increase costs and widen income distribution in origin areas as well. There have been several interventions by governments in the region to encourage nationals overseas to put their earnings in national banks and to invest in their home countries as well as to encourage the sending of remittances to the home country.

Another important issue relates to the exploitation of migrants from the region. While it is incorrect to depict all such migrants as victims subject to exploitation, there can be no doubt that some groups are vulnerable to exploitation of a number of kinds. This is particularly the case among some groups of female migrants, especially those who work in occupations where protection is difficult like as domestic workers in private homes and in the entertainment industry. Those who are undocumented are especially vulnerable to exploitation, but lack of knowledge and local languages and customs can also increase vulnerability of immigrants. Governments of sending countries vary considerably in the efforts they have made to protect their nationals abroad. The Philippines, for example, leads the way in this area but even so significant exploitation still occurs. The issue of protection of

Table 7 Main Southeast Asian labor exporting countries: Workers' remittances relative to exports and imports in US$ million, 1980–2012

Country	*Year*	*Workers' remittances*	*Total merchandise*		$\frac{R}{X}$	$\frac{R}{M}$
			Exports (X)	*Imports (M)*		
Indonesia	1980	33	21 908	10 834	0.2	0.3
	1992	264	33 815	27 280	0.8	1.0
	2012	7212	186 146	190 225	3.9	3.8
The Philippines	1980	421	5744	8295	7.3	5.1
	1992	2538	9790	15 465	25.9	16.4
	2012	24 641	51 995	65 360	47.4	37.7
Thailand	1979	189	5240	7158	3.6	2.6
	1992	1500	32 473	40 466	4.6	3.7
	2012	4713	229 519	247 590	2.1	1.9
Bangladesh	1980	339	885	2545	38.3	13.3
	1992	912	1903	2527	47.9	36.1
	2012	14 085	25 113	34 132	56.1	41.3
Pakistan	1980	2048	2958	5709	69.2	35.9
	1992	1574	7264	9360	21.7	16.8
	2012	14 007	24 596	44 157	56.9	31.7
India	1980	2757	11 265	17 378	24.4	15.9
	1992	2897	19 795	22 530	14.6	12.8
	2012	67 258	293 214	489 394	22.9	13.7
Sri Lanka	1980	152	1293	2197	11.8	6.9
	1992	548	2487	3470	22.0	15.8
	2012	6001	9480	19 087	63.3	31.4

Sources: Hugo, G.J., 1995. Labour export from Indonesia: an overview. ASEAN Economic Bulletin 12 (2), 275–298; World Bank World Development Report, various issues; World Bank Remittances dataset, October 2013.

nationals overseas is most pressing, but governments find it difficult to interfere in the affairs of other countries. Accordingly, some of the most effective protection comes from the activities of nongovernment organizations and through informal networks established by the migrants. Exploitation of the migrants can occur in the process of movement when they may be forced to pay exorbitant amounts of money or suffer physical abuse. There have been several documented cases of undocumented migrants from the region losing their lives due to suffocation or drowning in the process of migration. At destinations, they can be subject to exploitation and discrimination especially if they are undocumented migrants. Such migrants are usually depicted as criminals rather than as victims of the groups that exploit them in the process of migration or at the destination.

The pressing issue of protection of migrants and migrant workers from the region is related to that of controlling the intermediaries in the migration process. While all such intermediaries do not operate outside of the law, a substantial number do. Indeed the involvement of crime syndicates in people smuggling and trafficking has increased exponentially and has become an important issue in international crime. The multinational nature of these groups makes them difficult to control and their influence is widespread, even involving officials involved in immigration. The thriving of such organizations is clearly related to the failure of documented migration systems and enhanced policing of migration is unlikely to solve the problem of undocumented migration. Improvement of official migration systems must be part of the increasing international effort to combat the people smuggling and trafficking, which is especially rife in South and Southeast Asia. In countries of origin there is a need for information programs, which provide potential migrants with detailed, accurate information about the costs of the migration process and the nature of the experience at the destination. Studies in several countries indicate that migrants are often poorly informed on both these counts and this can lead to them paying excessive amounts to move and having false expectations of the destination. The latter, for example, results in up to one-half of Indonesian women going to the Middle East to work as domestics, returning to Indonesia before their contract time has expired. There needs to be a crackdown on the activities of unscrupulous intermediaries in several countries.

The issue of south–north migration representing a 'brain drain' of skill and talent needed for development in the home countries is a longstanding one in the region. In several origin countries, however, the occupational profile of migrant workers is largely an unskilled and semiskilled one, so that movement out of a country like Indonesia certainly cannot be characterized as a brain drain. On the other hand, much of the movement out of countries like India, Pakistan, the Philippines, and Singapore is of skilled and highly educated groups. In those cases, however, the movement is not necessarily a net economic loss to the nation. Economic studies have shown that in some cases individuals have a net positive economic impact if they go overseas and remit money back to their home country rather than remain behind. In countries like India and the Philippines, there appears to be an over supply of some skilled groups in relation to the opportunities available locally. Nevertheless, there can be no doubt that the emigration of some skilled groups can be detrimental to local economies in some cases. With the globalization of labor markets, such mobility is inevitable. Some countries in the region (e.g., Pakistan) have put in place programs to encourage former emigrants with certain skills to return to their home country. The growth of India's information technology industry, for example, has lured some former emigrants to return.

With the increasing recognition of the significance of international migration's influence on economic development, the last decade has seen a growing interest of governments in adopting explicit policies to influence international migration. Table 8 shows that there has been an increase in the number of countries with explicit immigration policies and an increase in the number seeking to increase immigration levels. Moreover, the realization that in a favorable policy context emigration can enable development in origin areas has seen an increase in the number of countries with diaspora policies (Table 9).

The interrelationship between international migration and development in the region is much argued. There can be no doubt that international migration can have both positive and negative effects on the well-being of communities left behind. This presents a challenge to policy makers to put in place institutional structures, which maximize these benefits and minimize the negative effects. All empirical evidence points to the fact that there is little to be gained from policies and programs designed to stop population movement when there are economic imperatives behind that movement. The challenge is in creating policies and structures, which enhance the benefits of the movement.

There are a range of issues that relate to the experience of migrants from South and Southeast Asia in destination countries. Several groups settling in northern countries show a tendency to concentrate in particular communities or develop

Table 8 Asia: views and policies of governments regarding immigration, 1976–2011

	Views on level of immigration				*Goal of policies on immigration*			
	Too low	*Satisfactory*	*Too high*	*Total number of countries*	*Raise*	*Maintain or no intervention*	*Lower*	*Total number of countries*
1976	4	32	1	37	4	32	1	37
1986	1	30	7	38	1	30	7	38
1996	1	35	10	46	2	23	21	46
2005	4	30	13	47	4	26	17	47
2011	2	33	12	47	7	27	13	47

Source: United Nations, 2013d. World Population Policies 2011. United Nations, New York.

Table 9 Asia: Views and policies of governments regarding emigration, 1976–2011

	Views on level of emigration				*Goal of policies on emigration*			
	Too low	*Satisfactory*	*Too high*	*Total number of countries*	*Raise*	*Maintain or no intervention*	*Lower*	*Total number of countries*
1976	4	31	2	37	4	31	2	37
1986	3	28	7	38	5	25	8	38
1996	2	31	13	46	3	32	11	46
2005	7	25	15	47	9	24	14	47
2011	7	28	12	47	12	26	9	47

Source: United Nations, 2013d. World Population Policies 2011. United Nations, New York.

ethnic enclaves, especially in major metropolitan areas. Debate on these ranges between those who see such areas as barriers to the adjustment of migrants and those who see them as having a cushioning effect in allowing migrants to adjust to life in the new country in a gradual way and with the support of community networks. Other issues relate to discrimination experienced by some groups in housing and job markets as well as nonrecognition of qualifications leading to occupational skidding and deskilling.

Conclusion

The South and Southeast Asian region represents one of the world's major reservoirs of potential international migrants accounting for more than one-quarter of the global population and several of its largest labor surplus economies. Until recently, the impact of international migration in the region has been limited, but it is now of significance in all countries in the region. Although fertility decline has been significant, the labor force in the region continues to grow rapidly. Economies in several countries in the region are also growing rapidly but all of the pressures, which have created greater movement in the last two decades than previously will continue to operate. Hence the twenty-first century will see an increase in the scale and complexity of movement out of, within, and into the region.

See also: Assimilation of Immigrants; Demographic Impact of Disasters; Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Dependency Theory; Discrimination: Racial; Economic Sociology; Ethnic, Racial, and Nationalist Movements; Immigration: Health and Mortality; International Migration and Remittances; International Migration by Ethnic Chinese; International Migration; Migration and Development; Migration and Fertility; Migration and Social Capital; Migration, Theory of; Migration: Cultural Aspects; Migration: Motivations; Social Networks.

Bibliography

Arif, G.M., July 2009. Recruitment of Pakistani Workers for Overseas Employment: Mechanisms, Exploitation and Vulnerabilities. International Labour Office, Geneva.

Asian Development Bank Institute (ADBI) and Organisation for Economic Cooperation and Development (OECD), 2013. Managing Migration to Support Inclusive and Sustainable Growth. ADBI and OECD.

Baruah, N., 23–25 January 2013. Trends and outlook for labour migration in Asia. In: Presentation for the 3rd ADBI-OECD-ILO Roundtable on Labour Migration in Asia: Assessing Labour Market Requirements for Foreign Workers and Developing Policies for Regional Skills Mobility, Thailand.

Hugo, G.J., 1995. Labour export from Indonesia: an overview. ASEAN Economic Bulletin 12 (2), 275–298.

Lintner, B., 27 July 2000. Living on the edge. Far Eastern Economic Review 48.

Migration News, 2013, Volume 20, Number 2, April, https://migration.ucdavis.edu/mn/more.php?id=3832_0_3_0, (accessed 06.01.04.).

Scalabrini Migration Centre, 2012. Asian Migration Outlook 2011. Scalabrini Migration Centre, Quezon City, Philippines.

UNESCAP, 2013. 2013 ESCAP Population Data Sheet. UNESCAP.

United Nations, 2013a. World Population Prospects: The 2012 Revision. United Nations, New York.

United Nations, 2013b. Trends in International Migrant Stock: Migrants by Destination and Origin. United Nations Database. POP/DB/MIG/Stock/Rev.2013.

United Nations, September 2013c. International Migration 2013: Migrants by Origin and Destination Population Facts. No. 2013/3. Department of Economic and Social Affairs, Population Division.

United Nations, 2013d. World Population Policies 2011. United Nations, New York.

United Nations Educational, Scientific and Cultural Organization (UNESCO), 2013. Global Education Digest 2012. UNESCO, Paris.

United Nations High Commissioner for Refugees (UNHCR), 2013. Displacement: The New 21st Century Challenge – UNHCR Global Trends 2012. UNHCR.

Veron, J., 2008. The demography of South Asia from the 1950s to the 2000s: a summary of changes and a statistical Assessment. Population-E 63 (1), 9–90.

Williams, L., Guest, M.P. (Eds.), 2012. Demographic Change in Southeast Asia: Recent Histories and Future Directions. Southeast Asia Program Publications. Cornell University, New York.

World Bank, 2000. World Development Report 1999/2000. Various Issues. Oxford University Press, New York.

International Organisations and Education

Pradeep A Dhillon, University of Illinois, Urbana-Champaign, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Explicit research interest in the consanguinity between international organizations and education as a field of inquiry is a recent development. It has grown with the demise of formal colonialism and the rise of the economic liberalization that emerged, and advanced rapidly, in the wake of the Cold War. The development of new communication technologies that shrink both time and space have played a crucial role in these more recent efforts. They shape the demands that globalization, and the knowledge economy, places on international organizations as agents of global change.

Introduction

International organizations and education have often been intimately linked. Thus, for example, we can note the relationship between the Catholic Church and its rich tradition of scholarship and teaching in various parts of the world. We could also bring to mind the East India Company and the folding of education – particularly education in English – into its mercantilist involvement around the globe. Nevertheless, explicit research interest in the consanguinity between international organizations and education, particularly higher education, as a field of inquiry in itself is a relatively recent development (Altbach, 2010). It has grown with the demise of formal colonialism and the rise of the economic liberalization that emerged and advanced rapidly in the wake of the Cold War. The development of new communication technologies that shrank both time and space have played a crucial role in these more recent efforts and have come to shape the demands that our contemporary understanding of globalization, and the knowledge economy, place on education. It is the connection between these different, albeit related, strands of global, economic, and epistemological change that has compelled international organizations to play an increasingly active role first in education more broadly and now increasingly in the domain of higher education (Marginson, 2013).

Theoretical Approaches

Many methods have been invoked to understand the development and growth of education by international organizations tied primarily to the United Nations, but also to other organizations that bear a strong commitment to international social, cultural, scientific, and economic development. Following World War II, the efforts of international organizations were directed toward enabling previously colonial countries to modernize and establish capacities for full and equal participation in an international system. Thus, for example, UNESCO adopted in 1960 the Convention against Discrimination in Education. This convention called for a prohibition on all kinds of discrimination in access to education based on the right to education laid out as a fundamental human right in the Universal Declaration of Human Rights (Byani et al., 1994). Various theories were advanced both to promote this agenda and to provide a critique of it. Thus, if modernist theories sought to improve education as part of a global modernization project, neo-Marxist theories were critical of these efforts and saw them as promoting a new kind of colonialism and imperialism spurred by the interests of either countries wielding global power or dominant sections of society within nations themselves (Carnoy, 1974).

Furthermore, there was an effort to establish educational institutions as a component of the nation-building process within newly decolonized states; these were seen to be were part of the legitimizing process as these newly minted nation-states sought to develop in the mode of Western states. That said, however, the international organizations involved with education were often viewed with suspicion by the very nation-states that were seeking to develop though a Western conception of modernity and seen as tools in the hands of capitalist and parochial political power brokers rather than as agents of democratic change. This suspicion, for example, led India in the context of the Indo-Pakistan conflict around the war for Bangladeshi independence to end the functioning of the Peace Corps within its borders. In addition, most of the efforts of the international organizations involved with education were directed toward education at the primary and secondary levels with the goal of providing universal education to members of all the segments of society both within and across countries. Above all, at this stage, education was largely seen as a local interest, and international organizations saw themselves primarily as facilitators for providing universal access to primary and secondary education.

International Organizations as Educational Agents

The supportive educational role that international organizations played has changed with globalization, liberalization, and the knowledge economy that developed in the wake of the Cold War and the rapid growth of technological innovation. Emphasis has now shifted away from primary and secondary education to higher education, and international organizations – rather than being facilitators for national efforts at education – have become agents in educational reforms and the building and transformation of educational institutions around the globe. The relationship between universities and the world economy had been often noted, as by Manuel Castells (1994), but the focus

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92100-7

today is on preparing young men and women to work within a global knowledge economy.

In other words, the previous emphasis on educating for citizenship within a national ideological context and even the world economy in the context of the Cold War has shifted in the post-Cold War period toward educating for a global economy that relies heavily on the development and use of new information technologies (Torney-Purta et al., 2001). Thus, for example, the World Bank recently released a report titled *Skilling Up Vietnam: Preparing the Workforce for a Modern Market Economy*, which stresses that the nature of work in a modern market economy will need employees who display a mix of higher-quality cognitive, behavioral, and technical ability. Furthermore, the new and emerging technologies vital to the modern market need internationally shared standards of competence if they are to be successful in driving the global economy. Hence, experts argue that it is time to consider the active role that international organizations are now playing, and being increasingly called on to play, in building a world educational culture, one in which all can fully and equally participate (Resnik, 2006).

Technological Education and International Organizations

The rhetoric promoting such an approach to education might sound like the previous call for education for all, but there are deep differences. The earlier call was driven by a normative sense of egalitarianism; the more recent call is, to a large extent, driven by economic considerations. Thus, for example, it is tempting to see current efforts at providing open source access to systems of knowledge as a move toward more equitable distribution of research and innovation opportunities. Nevertheless, such efforts could easily prove to be tied to more narrow national and international economic interests and competition, as we can see from the educational goals articulated by Vietnam and also in other countries around the globe. In India, for example, there is a rapidly growing system of Indian Institutes of Information Technology originally funded by the Ministry of Human Resources and Education and now increasingly through private and public partnerships. Twenty such institutes have already been established, and several more are envisioned.

Regardless of the efforts of individual countries, as already noted, international organizations promoting and supporting education in this globalized, neoliberal knowledge economy have come to operate as intergovernmental organizations, and become central actors in a globalizing world with demands for shared educational standards and results as well as governance. In addition, since they first arose, each international organization has developed its own history, organizational structure, and culture. Importantly, operating as independent organizations, they have come to develop and pursue their own interests and agendas and often stand accused of having lost sight of their original shared vision of creating global equity of opportunity through the development of social, cultural, scientific, and economic capacities in harmony with each other under the umbrella of large global organizations like the United Nations.

Furthermore, these initial normative commitments are often obscured by more recent demands placed on international organizations by the shifts in the global economic landscape, even though the placement of education predominantly in the service of economic development is not a new concern (Vaizey, 1967). Critics have often noted the growing participation of economists of education and international organizations in the diffusion and consolidation of education around the world. That said, what is different in today's neoliberal economy is the explicit justification that entrepreneurship, product development, and managerial efficiency with a focus on educational outcomes have come to provide for educational research and practice. Innovation, in addition to diffusion and consolidation, has become vital to the business of developing a world culture of education. Typically, it is international organizations that place this kind of demand on educational institutions at all levels of what is an increasingly overt international capitalist system. It is not surprising, then, to see fast-growing emphasis on educating for scientific and technological excellence at the secondary levels, and for business and technology at the university levels. Even the humanities are only considered valuable to the extent that they are in the service of technological advancement. Human values – humanism in the old-fashioned sense of the term – are almost completely ignored in the vision of a posthuman world that now drives much of contemporary global educational policy. In the introduction to a volume on higher education in South Africa, the novelist J.M. Coetzee (2013) cautions against the elision of the humanities specifically, and the retreat of nations as educational policy is increasingly driven by international organizations, like the World Bank, in the service of the global economy: "The response of the political class to the university's claim to a special status in relation to the polity has been crude but effectual: if the university, which, when the chips are down, is simply one among many players competing for public funds, really believes in the lofty ideals it proclaims, then it must show it is prepared to starve for its beliefs. I know of no case in which a university has taken up the challenge."

Moreover, in addition to social and critical literacy, this push for technological innovation and development ignores the educational charge of fostering social cohesion and concern for the disenfranchised while respecting individuality and cultural differences. Thus, the World Bank is often accused of promoting a clear agenda for higher education in the service of the global economy, leaving other organizations within the global educational structure to pay attention to heretofore overlooked global concerns like the need for deepening and promoting human rights education. Anna Russell (2010) paints a particularly poignant picture of the manner in which international organizations have only just started to engage with the human rights community. Using the right to water as a case study, she demonstrates the ways in which international organizations either conceived of development in complete isolation of human rights and human rights education – it was not seen as relevant to practice – or addressed these issues under vague and unenforceable terms like *nondiscrimination* and *tolerance*. Russell calls for international organizations to develop robust institutional structures that would allow for

cooperation among organizations in development, placing human rights at its heart.

Human rights education is only one of the normative demands placed on the international organizations responsible for developing a world educational culture. Educating for social cohesion and citizenship, both within nation-states and at the international level, is an equally important task that international organizations must take responsibility for as they play their role as agents of change through education in the international system. It was Frederic of Prussia who noted that "there are few lands in which all citizens have the same religion, and the question arises: is such unity to be forced or can one permit everyone to think according to his own views?"(cited in Heynman, 2003). The turn to tolerance suggested by the Enlightenment philosophers as a solution to be achieved through education is that much more important today as we seek to improve tolerance, cooperation, and harmony between nation-states and subnational populations within a dense and increasingly integrated and equitable international system. Such universalism, however, is not to be thought without the specificity of the histories and cultures of the nation-states in question and also those of the international organizations themselves. It is to be kept in mind that until quite recently education has for the most part been viewed traditionally as relevant to local culture and therefore subject to local control. Certainly, the educational landscape is changing as international organizations become agents in the service of neoliberal global policies and the knowledge economy. Educators and policy makers are increasingly being called on to establish a general research agenda "on the formulation and transmission of world-level educational ideals and prescriptions to individual countries" (McNeely, 1995).

Conclusion

In sum, international organizations are now considered important in developing a worldwide pattern of educational institutionalization. Mass education – hence the calls for diversity and inclusion in all educational institutions, but especially those of higher education – has now become the imaginary that drives much of educational policy in the discourses of international organizations as well as national and regional educational institutions. For educators it is important to emphasize that this not merely an economic imaginary but also a normative one and as such demands education as a mark of citizenship. It is fair to say, though, that the policies developed and promoted by international organizations are often prescribed without sufficient awareness of local cultures and national political demands that leave the policies to be adopted in name alone, without effective local implementation. Thus, for example, the new technologies allow teaching to large numbers across the globe – thus meeting the demands of equal access to knowledge – but could really only be effective for the most culturally thin content. Even if we are committed to building global capacity for the knowledge economy, we would still need to attend to local social, cultural, and historical conditions for such learning to be effective and the development and use of appropriately relevant innovations. It is important to note, however, that organizations like UNESCO are indeed actively involved in supporting local efforts at preserving not only concrete heritage sites around the world but also cultural sites of intangible heritage (Ruggles and Silverman, 2006). Above all, given growing democratic demands across the international system and despite economic instability, postcolonial conditions, and the declining significance of public education, international organizations will have to play an even more active role as vast populations of the planet still stand in need of education policies that would meet the challenge of ameliorating the conditions of persistent global poverty and addressing the looming threat of increasing climatic disasters.

See also: Cross-National Comparisons in Education: Findings from PISA; Environmental Education and Education for Sustainable Development; Social Inequality and Schooling.

Bibliography

Altbach, P.G., 2010. Higher education: an emerging field of research and policy. In: Bassett, R.M., Moldanado, A. (Eds.), International Organizations and Higher Education Policy: Thinking Globally, Acting Locally? Routledge, London, pp. 9–26.

Bayley, C., 2004. The Birth of the Modern World 1780-1914: Global Connections and Comparisons. Blackwell, Oxford.

Boli, J., Thomas, G.M. (Eds.), 1999. Constructing World Culture: International Nongovernmental Organizations since 1875. Stanford University Press, Stanford, CA.

Brown, P., Lauder, H., Ashton, D., 2011. The Global Auction: The Broken Promises of Education, Jobs, and Prosperity. Oxford University Press, New York.

Byani, C., Graham-Brown, S., Gundara, J., Jones, C., Thornberry, P., Warner, R., 1994. Education Rights of Minorities. London.

Carnoy, M., 1974. Education as Cultural Imperialism. McKay, New York.

Castells, M., 1994. The university system: engine of development in the new world economy. In: Salmi, J., Vespoor, A. (Eds.), Revitalizing Higher Education. Pergamon, Oxford, pp. 14–40.

Chabott, C., 2009. Constructing Education for Development: International Organizations and Education for All. Routledge, London.

Coetze, J.M., 2013. Introduction in: Higgins, J. Academic Freedom in a Democratic South Africa Johannesburg. University of Witwatersrand, South Africa.

Enders, J., 2004. Higher Education, internationalization, and the nation-state: recent developments and challenges for governance theory. Higher Education 47 (3), 361–382.

Evans, P., Rueschemeyer, D., Skopol, T. (Eds.), 1985. Bringing the State Back in. Cambridge University Press, New York.

Heynman, S., 2003. Education, social cohesion, and the future role of international organizations. Peabody Journal of Education 78 (3), 25–38.

Marginson, S., 2013. Imagining the global. In: King, R., Marginson, S., Naidoo, R. (Eds.), Handbook of Globalization and Higher Education. Edward Elgar Publishers, Northhampton, MA, pp. 10–39.

McNeely, C.L., 1995. Constructing the Nation-State: International Organization and Prescriptive Action. Praeger.

Peters, M., Marginson, S., Murphy, P., 2009. Creativity and the Global Knowledge Economy. Peter Lang, New York.

Resnik, J., 2006. International organizations, the "education-economic growth" black box, and the development of world education culture in. Comparative Education review 50 (2), 173–195.

Rothkopf, D., 2008. Superclass: The Global Power Elite and the World They Are Making. Little Brown, New York.

Ruggles, D.F., Silverman, H. (Eds.), 2006. Cultural Heritage and Human Rights. Springer, New York.

Russell, A.F.S., 2010. International organizations and human rights: realizing, resisting and prepackaging the right to water? in. Journal of Human Rights 9 (1), 1–23.

Scott, P., 2013. The university as a global institution. In: King, R., Marginson, S., Naidoo, R. (Eds.), Handbook of Globalization and Higher Education. Edward Elgar Publishers, Northhampton, MA, pp. 59–75.

Shin, J.C., Teichler, U. (Eds.), 2013. The Future of the Post-massified University at the Crossroads: Restructuring Systems and Functions. Springer, New York.
Torney-Purta, J., Lehmann, R., Oswald, H., Schultz, W., 2001. Citizenship and Education in 28 Countries: Civic Knowledge and Engagement at Age 14. International Association for the Evaluation of Educational Achievement, The Hague, the Netherlands.
UNESCO, 2010. Education for All Global Monitoring Report 2010: Reaching the Marginalized. Oxford University Press, Oxford.
Vaizey, J.E., 1967. Education in the Modern World. McGraw-Hill, New York.
World Bank, 29 November 2013. Skilling up Vietnam: Preparing the Workforce for a Modern Market Economy.
Young, M., 2007. Bringing Knowledge Back in. Routledge/Falmer, London.

International Social Problems: A Systems Perspective

Heather K Manion, Royal Roads University, Victoria, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

There is no shortage of international social problems facing us today. How problems are defined and addressed depends on their conceptualization and interaction with other parameters within a wider system. This article suggests a set of domains that are needed to help individuals, communities, and societies move from surviving to thriving and suggests that international social problems emerge when these elements are out of equilibrium. The article argues that new, more holistic solutions that harness and enhance innate strengths or assets can address complex international social problems.

What constitutes an international social problem depends on what is meant by 'international' and 'social problem,' who is defining it, and how they are theoretically or paradigmatically positioned. Global conversations are influencing how international social problems are understood and addressed. Given contemporary global interconnectedness, the growing population, and new communication tools available, there are more influential conversations than ever. This behooves social workers and other social change agents to recognize the inherent complexity of international social problems so that they can better utilize key international partners and hidden assets and design more impactful solutions with more sustainable change.

International social problems encompass a huge range of issues and this article can only touch on the breadth of topics. It begins with definitions, explores the various global drivers of social change, including globalization, looks at the biases through which international social problems are conceptualized and illustrates key examples. The article then investigates several contemporary attempts to solve international social problems and examines the international players involved in defining and addressing social problems, including social workers, and concludes with a glimpse of future possibilities.

Definitions and Key Concepts

International social problems are dynamic and relative. A social problem is an issue that arises from a conflictual interaction between individuals and their environment (social or natural, domestic or cross-boarder). In 2000, Khan and Dominelli (2000) noted how social workers were experiencing the impact of globalization and an internationalization of some social problems. Many social problems are first felt locally, but expand in size or across nations to become international issues or concerns. For example, people displaced by war and forced to migrate beyond their borders, face a new set of policies, social and cultural practices, and laws as they attempt to make new lives for themselves.

Whether a social problem is deemed domestic or international does not significantly change its definition. I argue that given the intensely interconnected way people, goods, services, and ideas interact across borders, all social problems have become internationalized to some degree, either in how globalization shapes the way they emerge at a local level or in how they are understood and addressed across time and space. For example, George and Page (2004) illustrate how a domestic social problem, i.e., family violence, has been impacted by globalization in several ways: increasing poverty and unemployment leading to a rise in family violence; facilitate processes that allow people to share information about family violence and mobilize action that impacts upon international and national policy agendas. Untangling how issues are connected internationally can be complex, as can mobilizing social action across cultures and jurisdictions. However, the ability to mobilize action across the globe offers the possibility of garnering broader momentum of support, more diverse information and expertise, and more innovative approaches (Leadbeater, 2008).

Globalization and Unequal Distribution of Wealth

No examination of international social problems would be complete without recognizing globalization as a dominant global force. The impact of globalization has meant that the professionals who work on preventing or tackling international social problems are increasingly interrelated and influenced by global events and conversations. "… The internationalization of social problems blurs boundaries between nation-states. Problems experienced by one individual or group in one country can require solutions that involve professionals everywhere" (Dominelli, 2010: p. 57). Dominelli proposes that just as the global village shrinks, multiple communities of practice are developing across regions, nations, communities coalescing around similar issues.

While arguably there has been an increase in consumer-based wealth and well-being in the world, the gap between the richest and the poorest has increased significantly since the 1970s (Schuerkens, 2010; Collier, 2007). This gap breeds social problems and civil unrest (including increased relative poverty, unemployment, crime, lower average rates of health, more disparate access to property and public services, and more fragile democracy). The global distribution of money is

well documented. Forty percent of the world's population controls only 5% of the world's wealth, while 20% of the world's richest individuals accumulate three quarters of the world's income (Muhr, 2011). Muhr suggests those in the middle feed the global economic beast and those at the end are irrelevant. Inequity breeds disadvantage. This is at the heart of most international social problems (George and Page, 2004).

Ways of Viewing International Social Problems

People build lenses as a way to make sense of the complex world. International social problems can be seen through various lenses: sociological theories, practice approaches, and political–philosophical stances. Lens-specific approaches can mark cross-disciplinary or cross-cultural work by impeding dialogue. Analyzing the lenses through which an individual or group understands an international social problem can help to identify and ameliorate understandings leading to better analysis of problems and potential solutions (see for instance Foucault, 1969). Sociology has been responsible for creating lenses and making them more transparent (e.g., Kendall, 2007). For example, an examination of international social problems could focus on either the individual's relationship to the problem or conversely the underlying structure that shapes the problem.

Mazur (2007) suggests social problems are built on unique histories and conflicts; rooted in individual contexts. Glyn et al. (1996: p. 3) defines social problems as "situations, policies, or trends that are (1) distressing or threatening to large numbers of people, (2) contrary to the mores (essential practices) or moral beliefs of the society, and (3) partly or wholly correctable through the actions of social groups." Alternatively Kendall (2007: p. 23) defines social problems as forging individual and social conditions or patterns of behavior that necessitate public concern or collective action to make change.

Bronfenbenner's ecological systems theory (1979) envisioned a series of nested systems. Although his theory was originally focused on children's development, within the context of international social problems I use it to refer to the microsystem (the system most close to the individual, e.g., the family, community, and services used); the mesosystem (the interplay between microsystems); and the macrosystem (societal norms and the systems that hold these).

A systems approach broadens one's perspective by helping to conceptualize complex interrelations across paradigms. Systems approaches were built from multiple disciplines and have been criticized for lacking a cohesive framework. However, they provide an opportunity for conceptualizing problems within the context of their history, setting and take into account inherent benefits, detriments, strengths, and assets. This holistic transdisciplinary approach promises a better understanding of how different components of a system interact, and helps develop more sophisticated solutions where both intended and unintended consequences can be better predicted.

Analytical lenses, cultural 'common sense' principles, and assumptions are weaknesses that need to be recognized and when necessary, counterbalanced or at least questioned when addressing international social problems. That many social problems persist is testament in part to the limitations of our analytical prowess.

Re-Envisioning International Social Problems

Globally there are a plethora of complex, interlinking, and overlapping social and environmental problems. As a result, people are increasingly working on clusters of factors influencing social problems rather than focusing on single issues. For example, providing access to health care for a refugee lowers their vulnerability across a range of factors unrelated to health. Individuals and groups with inherent or imposed vulnerabilities feel the impacts of international social problems more profoundly and in greater frequency than others do. Similarly social problems are manifested differently over the life span and across micro- and macro-levels.

International social problems can be clustered in many ways. For instance, Mkandawire (2006) suggests these global issues: poverty and underdevelopment, democratic transition and consolidation, and social protection and inclusion. Similarly, they could be grouped according to the 30 articles outlined in the United Nations Universal Declaration on Human Rights (1948). Alternatively, an individual approach could be used, such as Maslow's (1954) hierarchy of needs (i.e., physiological needs, safety needs, love and belonging, esteem, and self-actualization) or a human development approach, as proposed by Nussbaum (2006) (i.e., life bodily health, bodily integrity, senses, imagination and thought, emotions, practical reason, affiliation, other species, play and control over one's environment).

Any clustering has advantages and disadvantages. This section provides a taxonomy for international social problems using seven domains. Drawing on elements from the above, these seven interdependent domains include (1) prosperity and equality, (2) sustainable livelihoods, (3) people and power, (4) interpersonal relationships, (5) people and the environment, (6) health and well-being, and (7) people and knowledge. This taxonomy attempts to challenge traditional typologies of international social problems by articulating a strengths-based approach. Each domain is prefaced with a vision of an ideal before outlining how international social problems materialize when these domains are out of balance.

Prosperity and Equality: From Survival to Thriving

Vision: a society where all individuals and communities have equal access to items required to survive (i.e., basic needs) and thrive.

To survive, people need access to shelter and protection, sanitation, nutritional food and safe water, social inclusion, and basic health care. In order for people to thrive and live sustainably with dignity they additionally need education and knowledge, sustainable livelihoods, communication technologies, community support, and self-determination (see Maslow, 1954 or Nussbaum, 2006).

Internationally, human beings are not thriving, and nearly 1 billion people are barely surviving (Collier, 2007). Extreme poverty exists both in rural and urban areas. Economic or material poverty and inequality are the most commonly cited international social problems. George and Page (2004) argue that the processes of globalization have exasperated poverty and unemployment in many parts of the world. Gaps between the richest and the poorest can precipitate issues like civil unrest locally, nationally, and cross-nationally. Collier (2007) suggests that while most of the world has benefited from global economic prosperity, 1 billion of the world's poorest are extremely marginalized. They are localized in 58, mainly African, nations. Collier argues that they share the experience of civil war or ethnic violence, failed or corrupt states, lack of access to seaports and dependence on natural resource exports that cannot provide economic prosperity or stability.

Poverty can also be conceptualized as an 'access' problem. In this way poverty is more than an issue of economics. It encompasses access to environment, social interaction, healthy housing, and justice. Who has access is culturally variable, complex, and often illogical. As an example, natural disasters and war heighten vulnerability and highlight 'access' issues, not least because individuals are separated from their normal support mechanisms and infrastructure. Resulting problems may include food insecurity, famine (and mass migration), starvation, and long-term homelessness. Unequal access to resources is inextricably tied to many social problems outlined in the following domains.

Sustainable Livelihoods

Vision: sustainable livelihoods provide access to basic needs, the opportunity to thrive, alongside a sense of identity and belonging.

Sustainable livelihoods can provide sustenance, economic viability, a sense of identity, social status, and the means to produce and distribute resources and services. While livelihoods provide benefits to some, the societal system is rife with social problems, including lack of access to employment or sufficient employment, and exploitation. These issues are often felt intimately at the domestic level, and become international when workers need to move across nations and regions to survive and thrive.

Focusing on global employment, Schuerkens (2010) notes that a global decrease in union power has led to more instability for the workforce and greater chances of worker exploitation. Underemployment is also on the increase with more people in part-time and casual positions. Although for some, this may be a lifestyle choice for many it is not. Discrimination in employment remains entrenched. Women still earn less than men in most nations (Schuerkens, 2010).

The search for employment can spur international migration. For some it is a matter of survival for themselves and their families through remittances (Lyons et al., 2006). International mobility can be seen, for example, with Pasifika workers doing seasonal work in Australia and New Zealand, Latin Americans traveling up the coast of North and Central America following the harvests, domestic laborers, and nannies brought into the Middle East or Western Europe, or the miners and retail workers from China brought into Canada. Exploitation is a risk for some global migrants who lack full legal privileges, access to education and skills training, or a social safety net.

More extreme problems include people trafficking, for instance, people trafficked from Sri Lanka to Australia or from Western Africa to Western Europe or young girls trafficked for sexual exploitation from Eastern Europe to North America and Western Europe (Lyons et al., 2006). The risk of exploitation can come from both traffickers and employers when workers have no recourse to legal or social support.

People migrate through choice, necessity, or displacement. Risks of exploitation and insecurity are increased further by the circumstances causing migration, especially in the case of famine or war. International social problems incurred from mass migration impact both those migrating and those left behind. Where mass exoduses have occurred economic collapse often follows, for instance in the US cities where big industry have relocated internationally or in the environmental devastation left behind with the abandonment of massive refugee camps in West Africa. Conversely, population explosion in many global cities has meant a growth in the number and size of urban slums. These often have their own social problems linked to crumbling infrastructure and inadequate transportation (Mazur, 2007).

The choice (or lack thereof) to participate in the labor market is particularly pertinent at either end of the life span, and for those with impediments that curtail their participation. For children, the participation in labor is contentious. Western concepts of 'childhood' are not universal (Rock et al., 2012). Yet, Myers and Bourdillon (2012) suggest that efforts to eradicate child labor internationally have imposed Western ideals without understanding local contextual rationales for child labor. Some families, communities, and even children such as those living on the streets are reliant on children's labor input at home, in the field, or on the street. However, children lack a political voice making them more easily exploited, for example, in commercial child sexual exploitation, as child soldiers or land mine removers. More holistic systems analysis of child labor has sidestepped purely right/wrong arguments, and advocated for solutions that better understand local context and support whole communities through more flexible education approaches that utilize mentoring and apprenticeship models that combine education with humane treatment and skills development (Myers and Bourdillon, 2012).

People and Power

Vision: individuals and communities recognize their power, and use it for the betterment of all.

Kahane (2010), building on Paul Tillich's theories, suggests the fundamental dichotomous drives underlying all complex social problems are power (expressed as self-actualization) and love (expressed as drive to unity). He argues that international social problems are left unsolved because one drive is encouraged over the other; either love over power or power over love.

Terms like governance, leadership, and transparency permeate international development literature. They purportedly keep

power in check, especially expressed as 'power over.' These debates are designed to codify 'access to' principles, for example, in the forms of rights for woman, fertility, indigenous cultures, cultural minorities, children, migrant workers, and disabled people (e.g., Lyons et al., 2006; Hölsher, 2012).

Discordant power creates an environment for corruption and corruption hampers social development everywhere. Through transparency and accountability, citizen participation in decision making is sometimes seen as an antidote to power imbalance. However, "participation is still circumscribed to the spaces authorized by the powerful" (Mkandawire, 2006: p. 400). More inclusive ways to balance power at both local and international levels have to be found.

Interpersonal Relationships

Vision: an empathic society where people live in harmony with each other, respecting each other's similarities and differences.

Conflict can be related to access, equality, human rights, and discordant societal values and norms. Social problems related to conflict and discordance include interpersonal violence, family violence, child abuse, crime, mass violent crime, social injustice, strife, weapon trafficking, deadly combat, child soldiers, sexual violence, dehumanizing acts, and war (George and Page, 2004; Kendall, 2007; Schuerkens, 2010; Mazur, 2007; Glyn et al., 1996; Collier, 2007; Hölsher, 2012).

Neuroscience has illustrated that people quickly and unconsciously group others into ingroups (i.e., friends) and outgroups (i.e., foes). Rock (2009) suggests that the conceptual circuits used to interpret what is being said from a 'foe' leads to a greater chance of misunderstanding and misinterpretation and less empathy. This coalescing around group identity has caused many international ethnic conflicts (Mazur, 2007) where little empathetic response exists. Calloway-Thomas (2010: p. 7) suggests "[e]mpathy is the moral glue that holds civil society together; unless humans have robust habits of mind and reciprocal behaviour that lead to empathy, society as we know it will crumble."

Finding ways to ameliorate the ingroup/outgroup divide and foster empathic societies could help to reinvigorate community practice to tackle international social problems. That people are not continuously at war with their neighbors points to an innate human asset, namely that people generally appreciate each other's company.

People and the Environment

Vision: policies and practices that support people to live symbiotically with their environment.

Growing recognition exists that human activity is changing the planet at a systems level. People are now bombarded with a litany of global environmental impacts causing social problems, including pollution, natural resource degradation, loss of entire species of plants and animals, climate change, extreme weather events, effects from extractive industries, artificial disasters, famine, food insecurity, rising sea levels, and others. These profoundly impact social well-being and threaten to increase (Steffen, 2011).

People's relationship with the environment may increasingly impact a multitude of international social problems. As an example, food security illustrates a tangible global threat. For many, 'vulnerability' is manifested by living on marginal unproductive or unsafe land (whether rural or urban). Urban areas house the majority of the global population and the ability to ensure food is both produced and distributed to those areas is tenuous (George and Page, 2004).

Health and Well-Being

Vision: social policies and practices put people's physical, emotional, spiritual and cognitive health, and well-being at the center.

Numerous social problems are affiliated with health and well-being, including, communicable and noncommunicable diseases, mental illness, health care provisions, access to medicines, pandemics, hygiene, immunization, protection, happiness, and identity. The interplay between ill-health and poverty is complex and well documented (e.g., Collier, 2007; George and Page, 2004; Glyn et al., 1996; Mills and Shillcutt, 2004; Schuerkens, 2010).

International efforts have gone into tackling international health problems. Consequently, infant and child mortality rates and the burden of communicable diseases including HIV/AIDS and malaria have been lowered and life expectancy has drastically increased (Horton, 2012). However more can be done with concerted international effort. The most common causes of death include heart disease, stroke, and respiratory illnesses and the largest global risk factors include high blood pressure, tobacco, alcohol, and poor diet (Horton, 2012). The most commonly cited health issues within international development are tuberculosis, malaria, and AIDS (Mills and Shillcutt, 2004). Mills and Shillcutt suggest the significant direct and indirect impact of communicable diseases on families and societies globally through mortality, morbidity, fertility, and decreased intellectual capacity is significant.

Well-being is trickier to quantify, but attempts to introduce a happiness index and human development index have begun. These emphasize a broader set of indicators to measure well-being, such as a strong identity, access to and participation in art and culture, civic engagement, a sense of purpose and belonging, and a home. People's access to land rights, culture and heritage, individual and collective memories, and language is integral to their sense of identity and well-being (Berry, 2008). People have complex relationships with place (e.g., home) and their sense of belonging. Not having self-determination and sovereignty can undermine personal and collective identity and lead to significant international social problems, including armed conflict and civil war (Hölsher, 2012).

When well-being is negatively expressed it can lead to social problems including self-harm, suicide, substance abuse, gambling, and other addictions, with wider socio-environmental conditions compounding issues. These traditionally individual problems are increasingly international in

nature with the double drivers of globalization and propagation of social and traditional media that can promote or encourage risky behaviors (George and Page, 2004). Supporting initiatives, attitudes, and movements that foster a sense of belonging and community is fundamental to helping communities thrive.

People and Knowledge

Vision: social policies and practice support full access to, and recognition of, diverse cultural knowledge and a freely available education system that helps people flourish.

Codified as a basic right within the UN Universal Declaration of Human Rights, education has been thought of as a fundamental means of extricating people from poverty. Utilization of knowledge has changed with the development of the Internet and social media, making education more accessible to more people (for those with access to this technology).

Ownership and control of this information is increasingly discussed. Movement toward democratization of information is taking the form of shared software development, Wiki pages, knowledge banks, and increased access to school at all levels, subjects and language, including freely accessible courses from the world's best universities. While a digital divide still exists, effort has gone into making information communication technology more accessible and affordable, including in poorer nations.

This opens opportunities for dialogue on international social problems, as well as improved knowledge sharing and education across a range of subjects, such as literacy, numeracy, health promotion, and agricultural advancements (e.g., Leadbeater, 2008; Kahn, 2012; UN, 2012; Schuerkens, 2010). New avenues for redressing some international social issues may include social workers in southern Africa using online repositories like the Australian Domestic and Family Violence Clearinghouse to access examples of how to work with perpetrators of violence; or community workers in rural El Salvador opening an online dialogue with people at Kew Gardens in the UK about building seed banks; or social entrepreneurs in Chennai sharing resources with their counterparts in Toronto through Changemakers; or social work educators in Ethiopia sharing classroom discussions with students in Japan on the efficacy of antioppressive practice through LinkedIn (see Section Relevant Websites).

Access to knowledge is related to the concept of ownership and monetization of knowledge. Issues surrounding intellectual property rights still produce major social inequity (Manion, 2005). Food security, indigenous knowledge of medical plants, and artistic license are just some of the areas regularly debated. Hindering an individual's access to their own or indigenous knowledge, heritage, and language has profound implications on their well-being (Berry, 2008). Many international social innovations are emanating from ancient indigenous practices. This requires an improved way to honor, understand, and respect knowledge from a broader range of sources and epistemologies, and open the door to new possibilities (see Blackstock, 2010; Dominelli, 2012).

Reframing Questions and Seeking Solutions

Grappling with international social problems often begins with identifying problems, formulating possible interventions, testing, redrafting and implementing programs, and evaluating impact. Most international projects continue to rely on time-limited, project-based interventions that struggle to create enduring social change (Jaques and Vanbalkom, 2012: p. 413). Preventative measures are costly, but after-the-fact measures are more so (Kendall, 2007). Similarly micro-level solutions or individual solutions can benefit individuals with personal problems, but they ignore systemic issues. Mid-range attempts to solve social problems, including groups that help individuals cope with problems and grassroots groups that work for community-based change, often lack sustainability and struggle for larger scale change. Macro-solutions are particularly at risk of political interference and often lack acceptance or acknowledgment of individual circumstance, wisdom, or choice (Kendall, 2007).

At this unique point in history, where globalization has both exacerbated some international social problems and created new avenues for tackling them beyond the constraints of national and regional borders, there are both new challenges and opportunities for social workers at the micro-, meso-, and macro-levels (Dominelli, 2012). An alternative way for social workers to tackle these problems is to focus on what is needed to succeed and thrive within populations across national or regional borders before being hampered by the problem itself. For example, a maternity clinic was set up for young expectant mothers in Kenya (Acumen, 2013). Women had had poor experiences of the clinic and were not using it. They decided to focus on what the clinic would look like and how it would function, if it were to successfully attract patients. A success-focused approach allowed the health provider to explore a broader range of solutions. Design played an important role. Field trips with expectant mothers included visits to luxury hotel lobbies to distill what creates a welcoming environment. As a result of changes made to the clinic, women started using the service, thereby reducing infant mortality rates. Acumen's Human Centered Design model has been used successfully in many countries. This approach is akin to an asset or strength-based approach that allows participation and a wider array of possibilities because it focuses on solutions, not problems.

Strengths-based practice sidesteps fragmented problem solving and focuses on holistic approaches, which assume self-determination using strengths and assets (Saleebey, 1992). Saleebey (1992) proposed asking the miracle question, akin to asking what success looks like or what is needed to help individuals and societies thrive. I propose a macro-scale strengths-based miracle question. Outside the realm of social work, several multidisciplinary community workers are beginning to look at ways to solve entrenched (international) social and environmental problems. Examples of innovative and collaborative ways to tackle old problems are available on these Web sites: Charles Leadbeater, Acumen, Changemakers, WorldChanging Web sites.

The Actors

The direction of change is driven by those who define what international problems, or constellations of problems are important and which will be tackled. Recent history has shown that actors driving international development debates are predominantly located in the more economically developed countries. They include international agencies such as UN agencies, International Non-Government Organizations, World Bank, International Monetary Foundation, World Trade Organization, government foreign policy and foreign aid programmes, private foundations, and some consortia and university scholars. The constriction of power vested in a small number of bodies can itself be a social problem, requiring an international and institutional response. As an example Collier (2007) cynically argues that the decisions to fund or focus on one issue or region over another is less about choosing the issues where assistance is most needed, or where it potentially will make the biggest impact, but more spuriously by international consultants and organizations who act in their own self-interests or by public pressure based on the whims of concerned celebrities.

Macro-level attempts to solve social problems include special interest groups for political change and national and international social movements, including the UN as a major player. Although the UN plays a number of roles in tackling international social problems, its four main purposes are keeping peace; developing friendly relations among nations; helping nations work together to improve the lives of poor people, conquer hunger, disease, and illiteracy, and encourage rights and freedoms; and supporting nations to achieve these goals (http://www.un.org/en/aboutun/).

Two notable examples of macro-level global attempts at social policy include the Copenhagen Consensus and Millennium Development Goals (MDGs) (**Table 1**). The Copenhagen Consensus involved global economists prioritizing the global challenges that held the most promising opportunities in 2004. Their list included, in order: climate change, communicable diseases, conflicts and arms proliferation, access to education, financial insecurity, governance and corruption, malnutrition and hunger, migration, sanitation and access to clean water, subsidies, and trade barriers (Lomborg, 2004).

The MDGs are social and environmental aspirations set for 2015. Some targets have been met ahead of time, some are on track and many are unlikely to be met (see UN, 2012). MDGs represent the closest attempt the world has had to a universal set of ideals and an intentional means of actualizing change across a range of international social problems. The following Web site outlines the eight goals and tracks their achievement about progress in each region from 2012 (http://www.un.org/millenniumgoals/pdf/2012_Progress_E.pdf).

Role of Social Workers

In recent years social work has grown significantly and shifted into increasingly diverse areas of practice. Dominelli (2010: p. 4) suggests that social work has been shaped by globalization through globalization of the profession and rising demand for services, fiscal crisis between state and market, technological revolution, increase in disasters, commitment to human rights and social justice, influence of international social work organizations, and persistent optimism. The growing internationalization of social problems, the increase in social workers working in international organizations, and the increasing interrelatedness of localized practice, all lend themselves to supporting initiatives, processes, and activities that promote innovative social change. The International Federation of Social Workers (IFSW), the International Association of Schools of Social Work (IASSW), and the International Council on Social Welfare (ICSW) illustrated a commitment to this with the Global Agenda for Social Work and Social Development: Commitment to Action (2012). Specifically, between 2012 and 2022, these institutions are committed to promoting social and economic equalities, the dignity and worth of peoples, working toward environmental sustainability, and strengthening recognition of the importance of human relationships.

Looking to the Future

Old thinking will not produce new solutions. The way forward is to ask better questions and open dialogue with a diverse range of stakeholders with a vested interest in facilitating positive change. Block (2008) and Kahane (2010) both provide different, but powerful examples of how this has been done at local grassroots level with community members (Block) and at the geopolitical level between multinational actors looking to solve issues of food security (Kahane). The new approaches needed to solve long-standing issues will require citizens, social professionals, and power brokers working at the micro-, meso-, and macro-levels. Current conversations attest to the appetite for new approaches to complex social problems. Approaches, interventions, practices, and policies that promote positive change are varied, including movements toward global governance, convergence of priorities and sectors, investments in private–public partnerships, the spread of sharing networks, rise of social enterprises and entrepreneurs, increased calls for meaningful dialogue to solve tough problems, and increased peer-to-peer advocacy and action.

Social justice is at the heart of organizations and professionals mandated to help, but ensuring this value is integrated into international social policy is a mammoth task. Social problems require a systemic approach that combines sociological, economic, political, educational, and public health analysis and considerations. Scalable innovative solutions to complex social problems will emerge where there is a confluence of these aspects. A more balanced approach that integrates best practices, historical and local contexts and integrates the wisdom and ability to act inherent in micro- and macro-approaches could provide more impactful solutions. Fostering actions, interventions, policies, and practices that enhance people's ability to thrive in their environment and flourish in communities will provide a better foundation for an international society that bolsters human dignity and provides a place for everyone.

Table 1 Millennium development goals

Goal	*Target*	*Progress (according to UN, 2012)*
1. Eradicate extreme poverty and hunger	Reduce extreme poverty by half	"Extreme poverty falls in every region" (p. 6)
	Productive and decent employment	"Eastern Asia sets the pace for worldwide progress on working poverty" (p. 8)
	Reduce hunger by half	"The numbers of malnourished have stabilized since 1990" (p. 11)
2. Achieve universal primary education	Universal primary schooling	"Progress on primary school enrollment has slowed since 2004, even as countries with the toughest challenges have made large strides" (p. 16)
3. Promote gender equality and empower women	Equal girls' enrollment in primary school	"Parity is achieved in developing world primary schools, even though some regions lag behind" (p. 20)
	Women's share of paid employment	"Equal access to job opportunities remains a distant target for women in some regions … Women, more often than men, turn to the informal economy" (pp. 22–23)
	Women's equal representation in national parliaments	"Women continue to gain representation in parliaments, but the pace is slow" (p. 24)
4. Reduce child mortality	Reduce mortality of under – five-year olds by two-thirds	"Child mortality falls by more than one third, but progress is still too slow to reach the target" (p. 26)
5. Improve maternal health	Reduce maternal mortality by three quarters	"Maternal mortality has nearly halved since 1990, but levels are far removed from the 2015 target" (p. 30)
	Access to reproductive health	"Fewer teens are having children in most regions, but progress has slowed" (p. 34)
		"The unmet need for family planning remains persistently high in regions with low levels of contraceptive use" (p. 36)
		"Aid to reproductive health care and family planning remains low" (p. 37)
6. Combat HIV/AIDS, malaria and other diseases	Halt and begin to reverse the spread of HIV/AIDS	"New HIV infections continue to decline in the hardest-hit regions" (p. 38)
	Have halted by 2015 and begun to reverse the incidence of malaria and other major diseases	"Significant and durable progress against malaria is underway" (p. 42)
		"The tide has been turning in the spread of tuberculosis" (p. 44)
7. Ensure environmental sustainability	Integrate the principles of sustainable development into country policies and programmes and reverse the loss of environmental resources	"Forest area increase in Asia is helping to slow, but not reverse, global losses worldwide" (p. 46)
	Reduce biodiversity loss, achieving, by 2010, a significant reduction in the rate of loss	"Biodiversity is still being lost, even as more areas of the earth's surface are protected" (p. 49)
	Halve proportion of population without improved drinking water	"Halve proportion of population without improved drinking water" (p. 52)
	Halve proportion of population without sanitation	"Despite improvement in most of the developing regions, the MDG sanitation target is still out of reach" (p. 54)
	Improve the lives of slum dwellers	"Improvements in the lives of 200 million slum dwellers bring achievement of the MDG target, even as rapid urbanization swells the ranks of the urban poor" (p. 56)
8. Develop a global partnership for development		"Core development aid falls in real terms for the first time in more than a decade, as donor countries face fiscal constraints" (p. 58)
	Address the special needs of the least developed countries, landlocked countries, and small island developing states	"Aid dips as a share of total donor income, as well as for LDCs and sub-Saharan Africa" (p. 59)
	Develop further an open, rule-based, predictable, nondiscriminatory trading and financial system	"Trade rebounds after the crisis, with market access largely unimpaired" (p. 60)
	Deal comprehensively with developing countries' debt	"Interrupted in 2009, the downward trend of developing countries' debt-service ratios resumes" (p. 62)
	In cooperation with the private sector, make available the benefits of new technologies, especially information and communications	"The extraordinary rise in mobile cellular subscriptions continues in the developing world" (p. 63)

Adapted from United Nations, 2012. Millennium Development Goals Report 2012 United Nations, New York.

See also: Globalization and Health; Globalization and Human Trafficking; Globalization and World Culture; Globalization, Anthropology of; Globalization: Geographical Aspects; Globalization: Welfare and Social Work; Income Inequality; Inequality, Social; Inequality: Comparative Aspects; Near Middle East/North Africa Studies: Society and History; Social Inequality in History (Stratification and Classes); Social Work Theory; Subjective Wellbeing and Culture.

Bibliography

Acumen, 2013. Human Centred Design Workshop Workbook. Acumen, New York.

Berry, J., 2008. Globalisation and acculturation. International Journal of Intercultural Relations 32 (4), 328–336.

Blackstock, C., 2010. Getting into Trouble for Doing the Right Thing (Keynote Speech). SNAICC Conference. Alice Springs, Australia.

Block, P., 2008. Community: The Structure of Belonging. Berett-Koehler Publishers, San Francisco.

Bronfenbrenner, U., 1979. The Ecology of Human Development: Experiments by Nature and Design. Harvard University Press, Cambridge, MA.

Collier, P., 2007. The Bottom Billion: Why the Poorest Countries Are Failing and what Can Be Done about it. Oxford University Press, Oxford.

Dominelli, L., 2010. Social Work in Globalizing World. Polity, Cambridge.

Dominelli, L., 2012. Globalisation and indigenisation: reconciling the irreconcilable in social work. In: Lyons, K.H., Hokenstad, T., Pawar, M., Huegler, N., Hall, N. (Eds.), Sage Handbook on International Social Work. Sage, London, pp. 39–55.

Foucault, M., 1969. The Archaeology of Knowledge. Routledge, London.

George, V., Page, R.M. (Eds.), 2004. Global Social Problems. Polity, Cambridge.

Glyn, J.A., Hohm, C.F., Stewart, E.W., 1996. Global Social Problems. HarperCollins, New York.

Hölsher, D., 2012. Social Justice. In: Healy, L.M., Link, R.L. (Eds.), Handbook of International Social Work: Human Rights, Development, and the Global Profession. Oxford University Press, New York, pp. 44–54.

Horton, R., 2012. GBD 2010: understanding disease, injury, and risk. The Lancet 380 (9859), 2053–2054.

IASSW, ICSW, IFSW, 2012. Global Agenda for Social Work and Social Development: Commitment to Action. http://ifsw.org/get-involved/agenda-for-social-work/ (accessed 18.06.13.).

Jaques, L., Vanbalkom, D., 2012. International social development projects. In: Healy, L.M., Link, R.L. (Eds.), Handbook of International Social Work: Human Rights, Development, and the Global Profession. Oxford University Press, New York, pp. 411–416.

Kahane, A., 2010. Power and Love: A Theory and Practice of Social Change. Berrett-Koehler Publishers, San Francisco.

Kendall, K., 2007. Social Problems in a Diverse Society, fourth ed. Pearson A & B, Boston.

Khan, P., Dominelli, L., 2000. The impact of globalization on social work practice in the UK. European Journal of Social Work 3 (2), 95–108.

Khan, S., 2012. The One World School House. Twelve Hachette Book Group, New York.

Leadbeater, C., 2008. We-think: Mass Innovation, Not Mass Production: The Power of Mass Creativity. Profile Books, London.

Lomborg, B. (Ed.), 2004. Global Crisis Global Solutions. Cambridge University Press, Cambridge.

Lyons, K., Manion, K., Carlsen, M., 2006. International Perspectives on Social Work. Palgrave MacMillan, London.

Manion, H.K., 2005. A global perspective on intellectual property rights. International Social Work 48 (1), 77–78.

Maslow, A., 1954. Motivation and Personality, vol. 48 (1). Harper, New York pp. 77–87.

Mazur, A., 2007. Global Social Problems. Rowman & Littlefield Publishers, New York.

Mills, A., Shillcutt, S., 2004. Communicable diseases. In: Lomborg, B. (Ed.), Global Crisis Global Solutions. Cambridge University Press, Cambridge, pp. 62–111.

Mkandawire, T., 2006. Social development policies: new challenges for the social sciences. International Social Science Journal 58 (189), 395–404.

Muhr, T., 2011. Constructing the Bolivian Alliance for the peoples of our America. In: Schuerkens, U. (Ed.), Globalization and Transformations of Social Inequality. Routledge, New York, pp. 115–134.

Myers, W., Bourdillon, M., 2012. Introduction: development, children, and protection. Development in Practice 22 (4), 437–447.

Nussbaum, M., 2006. Frontiers of Justice: Disability, Nationality, and Species Membership. Harvard University Press, Cambridge, MA.

Rock, D., 2009. Your Brain at Work: Strategies for Overcoming Distraction, Regaining Focus, and Working Smarter All Day Long. Collins, New York.

Rock, L., Karabanow, J., Manion, H.K., 2012. Childhood and youth in international context: life-course. In: Lyons, K.H., Hokenstad, T., Pawar, M., Huegler, N., Hall, N. (Eds.), Sage Handbook on International Social Work. Sage, London, pp. 39–55.

Saleebey, D. (Ed.), 1992. The Strengths Based Perspective in Social Work. Longman, White Plains, New York.

Schuerkens, U. (Ed.), 2010. Globalization and Transformations of Social Inequality. Routledge, New York.

Steffen, A. (Ed.), 2011. WorldChanging: A Survivor's Guide for the 21st Century. Abrams, New York.

United Nations, 2012. The Millennium Development Goals Report 2012. United Nations, New York.

Relevant Websites

http://acumen.org/ – Acumen.

www.austdvclearinghouse.unsw.edu.au – Australian Domestic and Family Violence Clearinghouse

www.changemakers.com – Changemakers.

www.charlesleadbeater.net – Charles Leadbeater.

www.kew.org/science-conservation/save-seed-prosper/millennium-seed-bank – Kew's Milleniun Seed Bank Partnership.

www.linkedin.com – LinkedIn.

http://www.worldchanging.com/ – Worldchanging.

International Social Survey Programme

Tom W Smith, NORC at the University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The International Social Survey Program (ISSP) studies societies in comparative and temporal perspective. Since its founding in 1984 by Australia, Germany, Great Britain, and the United States, the ISSP has grown to cover 49 countries. Substantive topics and specific measures are repeated at regular intervals and include such themes as the role of government, national identity, work orientation, and social inequality.

The International Social Survey Program (ISSP) is a continuing, annual program of cross-national collaboration. It brings together preexisting, social science projects and coordinates research goals, thereby adding a cross-national perspective to the separate, national studies. The comparative perspective is important because in a globalizing world more and more social processes are international and interconnected and by placing countries in cross-national perspective one can differentiate between country-specific patterns and more universal phenomena.

ISSP evolved from a bilateral collaboration between the Allgemeinen Bevolkerungsumfragen der Sozialwissenschaften (ALLBUS) of the Zentrum fuer Umfragen, Methoden, und Analysen (ZUMA), West Germany, and the General Social Survey (GSS) of the National Opinion Research Center (NORC), University of Chicago, in the United States. (ZUMA is now part of Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen (GESIS) – Leibniz Institute for the Social Sciences.) Both the ALLBUS and GSS are replicating, time-series studies. The ALLBUS has been conducted biennially since 1980 and the GSS 29 times since 1972. In 1982, GESIS and the NORC devoted a segment of the ALLBUS and GSS to shared questions on job values, important areas of life, abortion, and feminism. Again in 1984, collaboration was carried out, this time on class differences, equality, and the welfare state.

Meanwhile, in late 1983, the National Centre for Social Research (NCSR)(then known as Social and Community Planning Research), London, which was starting a social indicators series called the British Social Attitudes Survey (BSA) similar to the ALLBUS and GSS, secured funds from the Nuffield Foundation to further international collaboration. Representatives from GESIS, NORC, NCSR, and the Research School of Social Sciences, Australian National University, organized ISSP in June 1984 and agreed to (1) jointly develop topics dealing with important areas of social science, (2) field the studies as a 15-min supplement to the regular national surveys (or a special survey if necessary), (3) include an extensive common core of background variables, and (4) make the data available to the social science community as soon as possible.

To govern ISSP, Working Principles were adopted (www.issp.org). These expressed ISSP's basic goals and established the structure for achieving them. ISSP is a collegial and democratic organization. The core institution is the annual plenary meeting of member countries where all members have equal standing and decisions are made by either consensus or majority rule. ISSP is coordinated by a Secretariat. A country is elected as the Secretariat for a 3-year term and may serve two consecutive terms. Great Britain (1985–91) served first followed by Germany (1991–97), the United States (1997–2003), Norway (2003–09), and Israel (2009–15). A Standing Committee consisting of the Secretariat and five other countries directs ISSP between the plenary meetings. The other permanent committee is the Methodology Committee, which oversees and reviews all technical matters relating to the design and execution of ISSP surveys.

The Working Principles also outline the basic procedures for drafting studies such as the role of drafting groups and the general meeting and the timeline for developing topics and final questionnaires. Moreover, it indicates ISSP's commitment to high data quality and functional equivalence across countries and stipulates standards for conducting and documenting surveys. This is shown in particular by the establishment of study monitoring reports, which are compiled for each round and detail the methodology used in each survey. The initial team of ISSP researchers was small and involved people running similar programs in their respective countries and could easily conduct cross-national research without a formal compact. By creating the Working Principles and setting down the rules and procedures, the team established a mechanism for the expansion of ISSP from beyond a collegial coterie into a large, diverse, international collaboration of social science institutes.

Each research organization funds all of its own costs. There are no central funds. The merging of the data into a cross-national data set is performed by the GESIS Data Archive for the Social Sciences, University of Cologne, in collaboration with the Analisis Sociologicos, Economicos y Politicos in Madrid.

Since 1984, ISSP has grown to 49 nations, the founding four – Germany, the United States, Great Britain, and Australia – plus Austria, Ireland, Hungary, the Netherlands, Israel, Norway, the Philippines, New Zealand, Russia, Japan, Bulgaria, Canada, the Czech Republic, Slovenia, Poland, Sweden, Spain, Cyprus, France, Portugal, Slovakia, Latvia, Chile, Denmark, South Africa, Switzerland, Venezuela, Brazil, Belgium, Finland, Mexico, Taiwan, Korea, Uruguay, the Dominican Republic, Croatia, Turkey, China, Italy, Ukraine, Argentina, Estonia, Iceland, Palestine, and India. In addition, East Germany was added to the German sample upon reunification. Past participants not currently active include Bangladesh and Brazil. The affiliated organizations are listed in Table 1. Other nations have replicated particular studies

 http://dx.doi.org/10.1016/B978-0-08-097086-8.44080-8

Table 1 ISSP

		ISSP studies									
Members	*Countries*	*Role of government*	*Social support*	*Social equality*	*Family/gender*	*Work*	*Role of government II*	*Religion*	*Social equality II*	*Environment*	*Family/gender II*
RSSS	Australia	D86	D87	D88	–	–	D90	D93	D93	D93	–
IS	Austria	D86	D86	D88	D88	D89	–	D93	D93	–	D95
BUP	Bangladesh	–	–	–	–	–	–	–	–	–	–
IUPERJ	Brazil	–	–	–	–	–	–	–	–	–	–
ASA	Bulgaria	–	–	–	–	–	–	–	D93	D94	D95
SC	Canada	–	–	–	–	–	–	–	D92	D93	D94
CES	Chile	–	–	–	–	–	–	–	–	–	–
CAR	Cyprus	–	–	–	–	–	–	–	–	–	–
ISCAS	Czech Republic[a]	–	–	–	–	–	–	–	D92	D93	D94
DEPPA	Denmark	–	–	–	–	–	–	–	–	–	–
FSD	Finland	–	–	–	–	–	–	–	–	–	–
CDA	Flanders	–	–	–	–	–	–	–	–	–	–
France-ISSP	France	–	–	–	–	–	–	–	–	–	–
GESIS	Germany[b]	D85	D86	D87	D88	D89	D90	D91	D92	D93	D94
NCSR	Great Britain[c]	D85	D86	D87	D88	D89	D90	D91	D92	D93	D94
Tarkai	Hungary	–	D86	D87	D88	D89	D90	D91	D92	D93	D94
SSRC	Ireland	–	D89	D89	D89	D89	D91	D91	–	D93	D94
TAU	Israel	–	–	–	–	D89	D91	D91	–	D93	D94
Eurisko/CENSIS/UEP	Italy[d]	D85	D87	D87	D88	D89	D91	D91	D92	D93	D94
BCRI	Japan	–	–	–	–	–	–	–	–	D93	D94
LAS/LSRC	Latvia	–	–	–	–	–	–	–	–	–	–
CEO/IMO	Mexico	–	–	–	–	–	–	–	–	–	–
SCP	The Netherlands	–	–	D87	D89	D89	–	D91	–	D93	D94
MU	New Zealand	–	–	–	–	–	–	D91	D92	D93	D94
NSD	Norway	–	–	–	–	D89	D90	D91	D92	D93	D94
SWS	The Philippines	–	–	–	–	–	–	D91	D92	D93	D94
ISS	Poland	–	–	–	–	–	–	D91	D92	D93	D94
ICS	Portugal	–	–	–	–	–	–	–	–	–	–
VCIOM	Russia	–	–	–	–	–	–	D91	D92	D93	D94
IS-SAS	Slovakia	–	–	–	–	–	–	–	–	–	–
POMCRC	Slovenia[e]	–	–	–	–	–	–	D91	D92	D93	D93
HSRC	South Africa	–	–	–	–	–	–	–	–	–	–
CIS/ASEP	Spain	–	–	–	–	–	–	–	–	D94	D94
UU	Sweden	–	–	–	–	–	–	–	D91	–	D94
SIDOS	Switzerland[f]	–	–	D87	–	–	–	–	–	D93	–
AS	Taiwan	–	–	–	–	–	–	–	–	–	–
NORC	USA	D85	D86	D87	D88	D89	D90	D91	D92	D93	D94
LACSO	Venezuela	–	–	–	–	–	–	–	–	–	–
Others:											
IFS	Poland	–	–	D87	–	–	–	–	–	–	–
BS	Lithuania	–	–	–	–	–	–	–	–	–	D94

Table 1 ISSP—*Cont'd*

		National ID	*Role of Government III*	*Work II*	*Religion II*	*Social equality III*	*Environment II*	*Social support*	*Family/gender III*	*National ID II*	*Citizenship*
RSSS	Australia	D96	D98	–	D98	D00	–	D01	D02	D03	D05
IS	Austria	D95	–	–	D99	D00	D01	D01	D02	D04	D04
BUP	Bangladesh	–	–	D97	–	–	–	–	–	–	–
IUPERJ	Brazil	–	–	–	–	–	–	D01	D03	–	D06
ASA	Bulgaria	D95	D97	D97	D99	D99	D00	D01	D02	D03	D05
SC	Canada	D95	D96	D99	D99	D00	D00	D02	–	D04	D04
CEP	Chile	–	–	–	D98	D00	D00	D01	D03	D03	D05
CAR	Cyprus	–	D96	D97	D98	D99	–	D01	D02	–	D04
ISCAS	Czech Republic[a]	D95	D96	D97	D99	D99	D00	D01	D02	D03	D04
DEPPA	Denmark	–	–	D97	D98	D01	D01	–	D03	D04	D05
FSD	Finland	–	–	–	–	–	D01	D01	D02	D03	D04
CDA	Flanders	–	–	–	–	–	–	–	D02	–	D04
FRANCE-ISSP	France	–	D97	D98	D98	D99	–	D01	D02	D03	D05
GESIS	Germany	D95	D96	D97	D98	D00	D00	D02	D02	D03	D03
NCSR	Great Britain[b]	D95	D96	D97	D98	D99	D00	D01	D02	D03	D04
Tarkai	Hungary	D95	D96	D97	D98	D98	–	D01	D02	D03	D04
SSRC	Ireland	D95	D96	–	D98	–	D02	–	D02	D03	D03
TAU	Israel	–	D96	D97	D98	D99	D00	D01	D02	D04	D05
Eurisko/CENSIS/UEP	Italy[d]	D95	D96	D97	D99	–	–	D02	–	–	–
BCRI	Japan	D95	D96	D97	D98	D99	D00	D01	D02	D03	D04
SRC	Korea	–	–	–	–	–	–	–	–	D03	D04
LAS/LSRC	Latvia	D95	D96	–	D98	D99	D00	D01	D03	D03	D04
CEO/IMO	Mexico	–	–	–	–	–	D01	–	D03	–	D06
SCP	The Netherlands	D96	–	D98	D98	–	D00	–	D03	D05	D05
MU	New Zealand	D96	D97	D97	D98	D99	D00	D01	D02	D03	D04
NSD	Norway	D95	D96	D97	D98	D99	D00	D01	D02	D03	D04
SWS	Philippines	D95	D96	D97	D98	D99	D00	D01	D02	D03	D04
ISS	Poland	D95	D97	D97	D99	D99	–	D01	D02	D05	D05
ICS	Portugal	–	–	D97	D99	D99	D00	–	D03	D04	D04
VCIOM	Russia	D96	D97	D98	D98	D99	D00	D01	D02	D03	D05
IS-SAS	Slovakia	D96	–	–	D98	D01	–	–	D02	D04	D05
POMCRC	Slovenia	D95	D95	D97	D98	D98	D00	D01	D03	D03	D03
HSRC	South Africa	–	–	–	–	–	–	D01	–	D03	D04
CIS/ASEP	Spain	D95	D96	D97	D98	D99	D00	D01	D03	D03	D04
UU	Sweden	D95	D96	D97	D98	D99	D01	–	D02	D03	D04
SIDOS	Switzerland[f]	–	D98	D98	D99	–	D02	D01	D03	D03	D04
AS	Taiwan	–	–	–	–	–	–	–	D02	D03	D04
NORC	USA	D96	D96	D98	D98	D00	D00	D02	D02	D04	D04
DE/SI	Uruguay	–	–	–	–	–	–	–	–	D04	D04
LACSO	Venezuela	–	–	–	–	–	–	–	–	D04	D04

(*Continued*)

Table 1 ISSP—*Cont'd*

ISSP members	*Countries*	*Work orientation III*	*Role of government IV*	*Leisure*	*Religion III*
CEOP	Argentina	–	–	D08	–
RSSS	Australia	D05	D06	D07	D08
IS	Austria	–	P06	D08	D08
BUP	Bangladesh	–	–	–	–
IUPERJ	Brazil	–	–	–	–
ASA	Bulgaria	D05	–	D09	–
SC	Canada	D06	D06	–	D08
CEP	Chile	–	D06	D07	D08
ISR	Croatia	–	D06	D06	D09
CAR	Cyprus	D05	–	D07	D08
ISCAS	Czech Republic[a]	D05	D06	D07	D08
DEPPA	Denmark	D06	D08	–	D09
FUNGLODE	Dominican Republic	D05	D06	D07	D08
FSD	Finland	D05	–	D07	D08
CDA	Flanders	D05	–	D07	D10
FRANCE-ISSP	France	D05	D06	D07	D08
GESIS	Germany	D05	D06	D08	D08
NCSR	Great Britain[b]	D05	D06	D07	D08
Tarkai	Hungary	D05	D06	D07	D08
SSRC	Ireland	D06	D06	D08	D08
TAU	Israel	D05	D05	D07	D08
Eurisko/CENSIS/UEP	Italy[d]	–	–	–	D08
BCRI	Japan	D05	D06	D07	D07
SRC	Korea	D05	D06	D07	D08
LAS/LSRC	Latvia	D	D07	D07	D09
CEO/IMO	Mexico	D06	–	D08	D09
SCP	The Netherlands	D	D06	D08	D08
MU	New Zealand	D05	D06	D07	D08
NSD	Norway	D05	D06	D07	D08
SWS	The Philippines	D05	D06	D08	D08
ISS	Poland	–	D08	D08	D08
ICS	Portugal	D07	D07	–	D09
VCIOM	Russia	D06	D07	D07	D09
IS-SAS	Slovakia	–	–	D08	D08
POMCRC	Slovenia	D05	D06	D06	D08
HSRC	South Africa	D05	D06	D07	D08
CIS/ASEP	Spain	D05	D07	–	D08
UU	Sweden	D05	D06	D07	D08
SIDOS	Switzerland[f]	D05	D07	D07	D08
AS	Taiwan	D05	D06	D07	D09
IPC	Turkey	–	–	–	D08
KIIS	Ukraine	–	–	–	D08
NORC	United States	D06	D06	D08	D08
DE/SI	Uruguay	–	D06	D08	D08
LACSO	Venezuela	–	D	–	D08

D = Done P = Planned.
AS = Academia Sinica (Taipei).
ASA = Agency for Social Analyses (Sofia).
BCRI = Broadcasting Culture Research Institute, NHK (Tokyo).
BS = Baltic Surveys, Lithuania (Vilnius).
BUP = Bangladesh Unnayan Parishad (Dhaka).
CAR = Center for Applied Research, Cyprus College (Nicosia).
CEO = Centro de Estudios Opinion, University of Guadalajara (Guadalajara).
CEOP = Centro de Estudios de Opinion Publica, University of Buenos Aires (Buenos Aires).
CDA = Centrum voor Dataverzameling en-Analyse, Katholieke Universiteit Leuven (Leuven).
CENSIS = Centro Studi Investimenti Sociali (Rome).
CES = Centro de Estudios Publicos (Santiago).
CIS/ASEP = Centro de Investigaciones Sociologicas and Analisis Sociologicos, Economicos y Politicos (Madrid).
DEPPA = Department for Economic, Politics, and Public Administration, Aalborg University (Aalborg).
DE/SI = Department of Economic/Statistical Institute, University of the Republic (Montevideo) Eurisko, in collaboration with the University of Milan (Milan).
FSD = Finnish Social Science Data Archive, University of Tampere (Tampere).
FRANCE-ISSP = consortium of Centre de Recherche en Economie et Statistique, Centre d' Informatisation des Donnees Socio-Politiques, Observatoire Francais des Conjonctures Economiques, and Laboratoire d' Analyse Secondaire et de Methodes Appliquees en Sociologie (Paris).
FUNGLODE = Fundacion Global Democracia y Desarrollo (Santo Domingo).
GESIS = Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen, formerly Zentrum fuer Umfragen Methoden und Analysen (Mannheim).
HSRC = Human Sciences Research Council (Pretoria).
ICS = Instituto de Ciencias Sociais, University of Lisbon (Lisbon).
IFS = Instytut Fiozofuu i Socjologii, University of Warsaw (Warsaw).
IMO = Institute of Marketing and Opinions (Mexico City).
IPC = Istanbul Policy Centre, Sabanci University (Istanbul).
IS = Instituet fuer Soziologie, University of Graz (Graz).
ISCAS = Institute of Sociology, Czech Academy of Sciences (Prague).
ISR = Institute for Social Research (Zagreb).
ISS = Institute of Social Studies, University of Warsaw (Warsaw).
IS-SAS = Institute of Sociology, Slovak Academy of Sciences (Bratislava).
IUPERJ = Instituto Universitario de Pesquisas do Rio de Janerio (Rio de Janerio).
KIIS = Kyiv International Institute of Sociology (Kiev).
LACSO = Laboratorio de Ciencias Sociales (Caracas).
LAS/LSRC = Latvian Academy of Science and Latvia Social Research Centre (Riga).
NCSR = National Center for Social Research; formerly Social and Community Planning Research (London).
NORC = National Opinion Research Center, University of Chicago (Chicago).
NSD = Norsk Samfunnsvitemskapelig Datajeneste, University of Bergen (Bergen).
MU = Massey University (Palmerston).
POMCRC = Public Opinion and Mass Communication Research Center, University of Ljubljana (Ljubljana).
RSSS = Research School of the Social Sciences, Australian National University (Canberra).
SC = Survey Center, Carleton University (Ottawa).
SCP = Sociaal en Cultureel Planbureau (Rijkswijk).
SI = Soziologisches Institut, University of Zuerich (Zurich).
SIDOS = Swiss Information and Data Archive Service for the Social Sciences (Neuchatel).
SRC = Survey Research Center, Sungkyunkwan University (Seoul).
SSRC = Social Science Research Center, University College (Dublin).
SWS = Social Weather Stations (Quezon City).
Tarki = Tarsadalomkutatasi Informatika Tarsula (Budapest).
TAU = Tel Aviv University (Tel Aviv).
UEP = University of Eastern Piedmont (Alessandria).
UU = University of Umea (Umea).
VCIOM = Soviet Center for Public Opinion and Market Research (Moscow).
[a]Includes Slovakia in 1992.
[b]Includes East Germany starting in 1990.
[c]Includes Northern Ireland 1989–91, 1993, 1994, 1998–2002, and 2008.
[d]CENSIS replaced Eurisko in 2001.
[e]Partial version of 1986 social support study.
[f]The 1987 and 1993 studies were fielded by the Soziologisches Institut, University of Zuerich.

without joining ISSP: East Timor, Indonesia, Kenya, Singapore, Sri Lanka, and Tanzania.

The annual topics for ISSP are developed over several years by a drafting group and pretested in various countries. The annual plenary meeting of ISSP then adopts the final questionnaire. Methodologically, ISSP researchers especially concentrate on developing questions that (1) are meaningful and relevant across countries and (2) can be expressed in an equivalent manner in all relevant languages. The questionnaire is originally drafted in English and then translated to other languages using rigorous translation and comparability checking procedures.

The themes covered in ISSP studies and the nations collecting data are listed in Table 1. The first theme on the role of government covered attitudes toward (1) civil liberties, (2) education and parenting, (3) welfare and social equality, and (4) the economy. The second theme was on social networks and support system. It contained detailed behavioral reports on contacts with various friends and relatives and then a series of questions about where one would turn for help when faced with various situations such as financial need, minor illness, career advice, and emotional distress. The third study, on social equality, concerned beliefs about what factors affect one's chances for social mobility (e.g., parental status, education, contacts, race, etc.), explanations for inequality, assessments of social conflicts, and related questions. It also asked people to estimate the average earnings of various occupations (e.g., farm laborer and doctor) and what the average earnings of these occupations should be. The fourth study covered the impact on the family of the changing labor force participation of women. It included attitudes on marriage and cohabitation, divorce, children, and child care and special demographics on labor force status, child care, and earnings of husband and wife. The fifth study on orientations toward work dealt with motivations to work, desired characteristics of a job, problems relating to unemployment, satisfaction with one's own job (if employed), and working conditions (if employed).

The sixth study in 1990 repeats the role of government theme. By replicating substantial parts of earlier rounds (at least two-thirds), ISSP has not only a cross-national perspective, but also an over time perspective. One is not only able to compare nations and test whether similar social science models operate across societies, but also able to see if there are similar international trends and whether parallel models of societal change operate across nations.

The seventh study covers the impact of religious beliefs and behaviors on social, political, and moral attitudes. It includes questions on religious upbringing, current religious activities, traditional Christian beliefs, and existential beliefs. The nonreligious items concern topics such as personal morality, sex roles, crime and punishment, and abortion. The eighth study in 1992 replicates and extends the 1987 social equality study. The ninth study in 1993 is on the environment. It includes an environmental knowledge scale along with attitudinal and behavioral measures.

The 10th study in 1994 repeats the 1988 study on gender, family, and work. It also adds items on household division of labor, sexual harassment, and public policy regarding the family. The 11th study in 1995 was on national identity. It assesses nationalism and patriotism, localism and globalism, and diversity and immigration. The 12th study in 1996 was the second replication of role of government. The 13th study in 1997 was the first replication of the 1989 study on work orientations. The 14th study in 1998 was the first replication of the 1991 religion study. The 15th study in 1999 was the second replication of the social inequality study fielded in 1987 and 1992. The 16th study in 2000 was the first replication of the 1993 environment study. The 17th study in 2001 was related to, but not a strict replication of, the 1986 social relations and social support study. In 2002, the 18th study was the third fielding of the gender, family, and work study. In 2003, the 19th study was the first replication of the 1995 national identity study. In 2004, the 20th study was a new study on citizenship. In 2005, the 21st study was the second replication of the work orientation study. In 2006, the 22nd study was the third replication of the role of government study. In 2007, the 23rd study was a new study on leisure-time activities. In 2008, the 24th study was the second replication of the religion study. In 2009, the 25th study was the third replication of the social inequality study. In 2010, the 26th study was the second replication of the environment study. In 2011, the 27th study was on the new topic of health policy and in 2012, the 28th study was the fourth fielding of the gender, family, and work study. In 2013, the 29th study will be the third fielding of national identity; in 2014, the first replication of citizenship; and in 2015, the fourth fielding of work orientation.

As of 2012, data from the first 26 studies are available from the GESIS Archive and various national archives such as the UK Data Archive at the University of Essex and Interuniversity Consortium for Political and Social Research at the University of Michigan. The 2011 study on health policy and the other rounds will be released in 2013 and later rounds will be available as soon as the data can be processed.

Publications based on ISSP are listed in a bibliography available at www.issp.org. The 2012 edition lists over 5200 publications. In recent years, over 400 publications have been added annually.

There are 12 collections of ISSP research: (1) Roger Jowell, Sharon Witherspoon, and Lindsay Brook, eds., *British Social Attitudes: Special International Report*. (Aldershot: Gower, 1989); (2) J.W. Becker, James A. Davis, Peter Ester, and Peter P. Mohler, eds., *Attitudes to Inequality and the Role of Government*. (Rijswijk, The Netherlands: Sociaal en Cultureel Planbureau, 1990); (3) Roger Jowell, Lindsay Brook, and Lizanne Dowds, eds., *International Social Attitudes: The 10th BSA Report*. Aldershot: Dartmouth Publishing, 1993; (4) Alan Frizzell and Jon H. Pammett, eds., *Social Inequality in Canada*. Ottawa: Carleton University Press, 1996; (5) Alan Frizzell and Jon H. Pammett, eds., *Shades of Green*. Ottawa: Carleton University Press, 1997; (6) Roger Jowell, John Curtice, Alison Park, Lindsay Brook, Katrina Thomson, and Caroline Bryson, eds., *British – and European – Social Attitudes: The 15th Report*. Aldershot: Ashgate, 1998; (7) Niko Tos, Peter Ph. Mohler, and Brina Malnar, eds., *Modern Society and Values: A Comparative Analysis Based on ISSP Project*. Ljubljana: University of Ljubljana and ZUMA, 2000; (8) the special issue on Pro-democratic Values in Cross-National Perspective of the *International Journal of Sociology*, 32 (Spring 2002), 1–108, (9) the special issue of the *International Journal of*

Migration Studies, 9 (No. 2, 2005), 82–156, (10) Betty Hilliard and Maire Nic Ghiolla Phadraig, eds., *Changing Ireland in International Comparison*. Dublin: Liffey Press, 2007; (11) Max Haller, Roger Jowell, and Tom W. Smith, eds., *Charting the Globe: The International Social Survey Programme, 1984–2009*. London: Routledge, 2009; and (12) Annamari Konttinen, ed., *Civic Mind and Good Citizenship: Comparative Perspectives*. Tampere: Tampere University Press, 2009.

For further information there are two Web sites that can be contacted:

1. GESIS Archive, University of Cologne: http://www.gesis.org/en/issp/
2. ISSP Secretariat: http://www.issp.org/

ISSP marks several important developments in the area of cross-national research. First, since the collaboration between organizations is not special or intermittent, but routine and continuing, this has resulted in an enormous amount of data being collected. From 1985 through 2010, ISSP conducted 644 national studies with 870 275 respondents. With more and more countries joining ISSP, this mass of data will continue to grow. Second, ISSP makes cross-national research a basic part of the national research agenda of each participating country. This means that cross-national research is a central and integral part of each country's research program and not a separate and occasional appendage. Third, by combining a cross-time with a cross-national perspective, two powerful research designs are utilized to study societal processes. This provides a solid basis for the scientific study of human society and the rigorous testing of theories concerning societal processes and development. Fourth, ISSP epitomizes the perspective that cross-national research should be collaborative. It is the opposite of the safari research tradition in which researchers in a single country or a small group selected a topic, designed an instrument, and then exported it to other countries. ISSP studies are developed and approved by the assembly of all member countries. As such, they reflect a collective, international point of view. ISSP follows the principles of inclusiveness and equality and also improves cross-national research by being more sensitive and responsive to all the societies being studied. Fifth, ISSP is democratic. ISSP is not run by a small cadre of principal investigators. Decisions are made by the collective membership at general meetings either via consensus or by majority rule following the one-country, one-vote principle. One might suppose that 40-some countries seated around a table and having serious scientific debates over the merits of topics and how to measure them would lead to chaos and gridlock, but in fact ISSP has operated smoothly and successfully produced annual studies for over a quarter century. Sixth, analysis is egalitarian and open to all. Once a merged data set is created, it is released to all and non-ISSP members have the same access as ISSP members do to the data. This greatly increases usage and advances social science research. Finally, ISSP has inspired and helped establish additional cross-national research such as the East Asian Social Survey and the European Social Survey. The later in particular has advanced cross-national research notably and helped establish new standards of excellence in comparative research (Jowell et al., 2007; Smith, 2007, 2010).

Cross-national survey research is central to advancing social science knowledge. As Durkheim (1938, p. 139) noted in 1895, "Comparative sociology is not a particular branch of sociology; it is sociology itself, in so far as it ceases to be purely descriptive and aspires to account for facts." And this of course applies to the social sciences as a whole. First, it recognizes that the cultures and countries of the world are increasingly interdependent and that many processes are essentially global rather than merely national. As the Working Group on the Outlook for Comparative International Social Science Research has noted, "A range of research previously conceived as 'domestic,' clearly needs to be reconceptualized in light of recent comparative/international findings." Unless a comparative perspective is adopted, "models and theories will continue to be 'domestic' while the phenomena being explained are clearly not" (Luce et al., 1989). Second, comparing countries is very useful since placing societies in perspective illuminates their social location and knowing a country's relative position is as important as its absolute attributes. As Butz and Torrey (2006) have observed, "international replication of national findings is critical for distinguishing between local and universal phenomena. The recent development of internationally harmonized, cross-sectional data sets has already demonstrated their potential." Finally, even when one's interest is purely monocultural, comparative data can be very valuable. For example, if one follows a societal trend in a single country, one is prone to seek country-specific explanations for the trend. But if one has cross-national data and finds that the trend is multinational, then one needs to seek a different and broader set of explanatory factors beyond events and developments in a single country.

ISSP assists the expansion of social-scientific knowledge by gathering data on important societal topics in comparative perspective. As House et al. (2004) note, "The globalization of survey research creates the opportunity for comparative and multilevel research on the grand scale of nations or societies. Adding comparative survey-based data to the analysis can yield more precise and reliable answers to questions of great scientific and social importance."

Bibliography

Butz, William P., Torrey, Barbara Boyle, June 30, 2006. Some frontiers in social science. Science 312, 1898–1899.

Durkheim, Emile, 1938. The Rules of Sociological Method. The Free Press, Glencoe, IL.

House, James S., et al., 2004. Survey research and social science: retrospect and prospect. In: House, James A., et al. (Eds.), A Telescope on Society: Survey Research and Social Science at the University of Michigan and Beyond. University of Michigan Press, Ann Arbor, MI.

Jowell, Roger, et al., 2007. Measuring Attitudes Cross-nationally: Lessons from the European Social Survey. Sage, Los Angeles.

Luce, R. Duncan, Smelser, Neil J., Gerstein, Dean R. (Eds.), 1989. Leading Edges in Social and Behavioral Science. Russell Sage, New York.

Smith, Tom W., 2007. Codes of ethics and standards in survey research. In: Donsbach, Wolfgang, Traugott, Michael (Eds.), Handbook of Public Opinion Research. Sage, London.

Smith, Tom W., 2010. The globalization of survey research. In: Harkness, Janet, Edwards, Brad, Johnson, Timothy, Mohler, Peter, Smith, Tom W., Braun, Michael, Pennell, Beth-Ellen, Lyberg, Lars (Eds.), Multinational, Multicultural, and Multiregional Survey Methods. Wiley & Sons, New York.

International Social Work

Bala R Nikku, School of Social Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia; and Nepal School of Social Work, Kathmandu, Nepal

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents an overview of definitions of international social work (ISW), current debates, and future course of ISW profession and teaching. After a brief introduction, Section What Is ISW?: The Debate presents the debate on what is ISW. Scholars have argued that social work is a contextual profession and hence international and global social work debates are vane. Section Internationalization of Social Work: Profession and Education further elaborates the debate on internationalizing social work profession and education. The following Sections International or Universal Social Work? and International but Not Global? present the discussions on quest for universalism and on global social work. Section Conclusion is a concluding one.

Introduction

The term international social work (ISW) was first used by George Warren in 1943, serving on the American President Truman's Presidential Advisory Committee on Political Refugees. Warren must have used this term to describe social work practice in agencies that were engaged in organized international efforts (Xu, 2006). During this time, the UN Relief and Rehabilitation Administration's (UNRRA) programs became a hub of ISW activities. The UNRRA's programs contributed to the spread of ISW, as well as ISW education (Healy, 2001). The UNRRA was an international relief agency, largely dominated by the United States but representing 44 nations. Founded in 1943, UNRRA became part of the United Nations in 1945, was especially active in 1945 and 1946, and largely shut down operations in 1947 and its functions were transferred to several UN agencies, including the International Refugee Organization and the World Health Organization. Much prior to this initiative one can note the historical events at the end of the nineteenth century like the Jane Addams's adaptation of practices from the British Settlement Movement and Toynbee Hall in London to create Hull House for poor people in Chicago, and during the same time the Charity Organization Societies and their growth both in the UK and the USA as the humble beginnings of the ISW.

What Is ISW?: The Debate

Periodic conceptual articles on ISW reveal a growing scholarly understanding of the concept. ISW is a complex concept that has been defined by scholars both from global north and south drawing on their both local and international work (Cox and Pawar, 2006; Lyons et al., 2006; Healy, 2001). Nevertheless, some authors express their disagreement and questioned whether social work can be truly international in nature given its diversity, irrelevance of western models, and its primary commitment to local issues (Gray, 2005; Drucker, 2003; Webb, 2003; George, 1999; Harris, 1990). These scholars have argued that social workers are asked to perform vastly different tasks in different countries and cultures. The methods and values of social workers in different countries or even communities in the same country (like India) can be quite diverse, complicated, and socially incongruent, considering or recognising the different set of social arrangements and values given the different globalization impacts on the local communities in an ever changing country situations. Gray and Webb critique that "Social work is no different now than it was in the last century as it spread from the west to the rest with its colonizing civilizing mission replacing local, Indigenous healing practices and communitarian values with psychologizing individualistic treatment regimes. Culturally relevant social work practice is by its very nature localized and 'indigenized.'" They also further argue that "Internationalizing processes tread on the toes of indigenization or the adaptation of western social work to local cultural contexts" (www.socmag.net/?p=56).

During the 1930s, entries titled 'International Social Work' appeared regularly in the *Social Work Yearbook* published in the USA by the Russell Sage Foundation. These early definitions were quite specific and comprehensive, and included practice across borders, international assistance, work through international bodies such as the International Labour Organization and the League of Nations, and the international exchange facilitated by world conferences in social work (Warren, 1939, cited from Healy and Thomas, 2007: p. 584).

Hokenstad et al. (1992) defined ISW from a broad perspective as a professional practice that crosses national boundaries, and relies on contacts and exchanges between countries. Over time, the term ISW also came to encompass domestic social work practice with immigrants and refugees (Sanders, 1984).

Lyons (1999), writing from England, agreed that ISW "is a nebulous concept, with elements of cross-national comparison and application of international perspectives to local practice, as well as participation in policy and practice activities which are more overtly cross-national or supra-national in character" (1999: p. 12). Lyons also suggests a distinctive field of ISW that includes cross-border practice and efforts to influence policy and practice at a supranational level. Similarly, Lyons et al. refer to "the idea of international social work as a perspective or specialist activity" (2006: p. 5).

Barker's (1999) "international social work is about using social work methods or personnel internationally; about social work cooperation between countries; and/or about transfer between countries of methods or knowledge about social work" (p. 250).

Healy (2001) defined ISW as "international professional action and the capacity for international action by the social work profession and its members"; arenas for action include international professional exchange, international practice and policy, and those aspects of local and domestic practice that 'transcend national boundaries' (p. 7). She does not view ISW practice as a field, but uses the term international action and makes the case that due to increased global interdependence, both local social workers and domestic human service agencies have greater opportunities and even a responsibility to engage in international activities (see also Healy, 1995).

Haug's (2005) critique of the Western-dominated discourse on ISW called for emphasis on human rights and social justice, 'guided by a clear vision of a more equitable global society.' Haug believes that "international social work includes any social work activity anywhere in the world, directed toward global social justice and human rights, in which local practice is dialectically linked to the global context" (2005: p. 133).

Mohan (2008) asserts that "international social work may be defined as a discursive discipline that employs the knowledge and tenets of social work practice in a diverse, dynamic and interdependent world" (2008: p. 13). He further argues that "international social work should be redefined as a professional discipline that promotes trans-national knowledge, studies and experiences to foster equality and justice as vehicles of international understanding, collaboration and collective human social development" (p. 19).

Lyngstad (2013) observes two possible reasons for the controversy or the debates in ISW: "The first has to do with the fact that the raison d'etre of social work is contextual; it is hard to imagine that good social work is possible without focusing on the local context. The second reason is related to the way we are using the concepts. Concepts such as ISW, comparative social work, contextual social work and internationalizing social work are often used in a helter-skelter manner, thus the discourse related to these issues is somewhat blurred since the concepts are being used in a disorderly way" (2013: p. 401).

It is evident from the above discussion that there is a 'plethora' of definitions about ISW. It is also clear that the interpretations of what is ISW, ranged from placements and work in other countries, to work in one's own country with service users originating from different parts of the world, to collaboration between countries and international organizations as well as a consideration of the impact of global developments on local situations (Hugman, 2010). Concepts such as cross-national, global, and ISW are often used interchangeably (Healy, 2008). Thus, the concept is used in a wide sense "as a broad, umbrella term referring to any aspect of social work involving two or more nations" (Healy, 1995: p. 442).

The notion/concept of ISW has evolved over a period of time from a simple activity to a transnational discipline. What is common in all of these definitions by various scholars is the influence of globalization, human rights violations, migration, poverty and disasters, and the urgent need for cross-cultural, cross-boarder, interprofessional action. These are some of the concepts central to ISW. This debate is very important to inward looking, self-criticizing social work profession to further grow and develop as a full profession in different parts of the world. This debate also reminds us that whether ISW as a concept still evolving or only revolving?

Internationalization of Social Work: Profession and Education

Internationalization is again a very broad term. Some relate to internationalization of the curriculum, student exchanges, faculty mobility programs, the student market, and employability in a globalized context. Internationalization of education especially higher education is a global trend and social work is not isolated to this process. Higher education today globally is a competitive industry and a common resource. Many countries see it as an engine of economic prosperity and an artifact of intellectual property and for some a subtle way of neocolonialism. So what is internationalization of social work? How is it happening? And what are the impacts?

In recent years, there has been an increased emphasis on the internationalization of the social work profession (Caragata and Sanchez, 2002; Johnson, 2004; Heron, 2005), even though the notion of internationalization and knowledge exchange has existed for decades (Lalayants et al., 2013). Nagy and Falk (2000) have stressed the interconnected nature of problems across the globe. There is an urgent need for thorough understanding of the consequences and solutions for addressing these global issues. Given this complexity, only competent, trained social workers with an international orientation can work effectively. Sanders and Pedersen (1984) further argued that an additional benefit of international education is creating opportunities to promote collaborative relationships in social work education, practice, and research.

Some scholars viewed the internationalization of social work is a positive and dynamic aspect of contemporary practice. However few others have argued that it is another form of Western dominance. Askeland and Payne (2006) have argued "internationalizing social work education creates a Western hegemony in disseminating knowledge." Dominelli (2012, 2009) states that internationalization refers to objectives, processes, practices, policies, challenges, and strategies that link the local with the global and vice versa. Internationalization may be beneficial or exploitative; its outcomes vary according to the values and aims of different institutions practicing it. She argues that internationalizing practices cross-national borders and may or may not respect locality-specific values, cultures, and perspectives.

Midgley (1997) calls internationalism as an ideology that focuses on processes of international collaboration to reduce ethnic tensions and share the commonalities of a global humanity. Hugman et al. (2010) explored varied ideas describing the concept of ISW. They looked at ISW from the idea of globalization, and discussed how ISW is figuratively embodied in migrations. To cite an example, Germany is the fifth largest international contributor to the UK social work workforce after Australia, USA, South Africa, and India (Hussein et al., 2011).

Akimoto (2008) states that the following are not necessarily ISW. They are: (1) practice or research on or in other countries, (2) joint work with practitioners or professors of other countries, (3) international comparative research and analysis, (4) assistance to the second and or third World and research on South–North relations, (5) cross-cultural social work, (6) devotion to ISW organizations, and (7) participation in an international conference. He concludes that the 'internationalization of social work' is *not* identical with ISW. He further presents that ISW is irreconcilable with egocentrism, ethnocentrism, and xenophobia, and unable to do without 'compound eyes' or eyes from the outside.

The Council on Social Work Education (CSWE) in the USA is very supportive of social work educators' activities in international policy development and advocacy and professional exchange. The CSWE standards call for the insertion of global and cross-cultural content in social work curricula (CSWE, 2008). This has placed a responsibility on social work educators who are now challenged to find ways to deliver such knowledge to their social work students through internationalizing their curricula and introducing relevant changes in coursework and the field practicum requirements (Wehbi, 2011). Similar emphasis can also be seen in the UK and other countries.

As a response to increased demand for social workers who are competent in cross-cultural practice, some universities/schools of social work (e.g., Umea University in Sweden, University of Michigan School of Social Work offers a certificate in Global Social Work) have started offering courses in Internationalization and Social Work or some others titled as ISW as part of their social work programs. This trend suggests that some schools of social work have recognized the need for internationalizing their curriculums, teaching methods with an objective that their graduates will be able to work both in home country and other countries. In turn, this trend has raised increased pedagogical demands for the social work educators in these institutions. Pedagogical tools like 'reverse mission,' cultural immersion, and use of self-reflective journaling and reflective writing, group dialogue are being used (see Furman et al., 2008; Kreitzer et al., 2012; Lalayants et al., 2013).

The past and ongoing social, political, economic, and cultural processes in the global northern and southern states have resulted into a diversity of social welfare systems and governance structures that lead to differences in how social workers are taught, trained, employed, organized, and regulated. There is also a need for further research and publication into the student choices, preferred learning styles when it comes to ISW content. The ongoing social work innovations and scholarship building across many schools of social work provide pathways to further strengthen internationalizing social work teaching, learning, and ultimately the profession (Nikku and Hatta, 2014).

International or Universal Social Work?

From the early humble beginnings of social work in the nineteenth century, war and developments in one country or area influenced developments in others. The first international activities being organized in the nineteenth century when representatives of public and private charities began to organize conferences to share their experiences and exchange information on successful methods of practice. Despite this history, ISW has been described as the youngest branch of social work (Friedlander and Apte, 1968).

As early as 1960s, Benjamin Goldman (1962) published in International Social Work journal on ISW as a professional function and called for the development of ISW as a fourth method, to complement community organization, group work, and casework. Xu (2006) comments that "it is noteworthy that almost 80 years after social work first 'went global,' international social work is still struggling to become a well recognized and permanent field of social work" (2006: p. 682).

Similarly, the quest for a 'universal social work' is seen in the efforts of global organizations like IASSW, ICSW, and IFSW. These organizations are in the forefront in organizing joint global social work and social work development conferences beginning 2010 in Hong Kong, 2012 in Stockholm, 2014 in Melbourne, and 2016 in South Korea. The quest was also evident from the revisions of Global Social Work definitions in 2000 and 2014. The endorsement of Global Standards of Social Work Education (first presented at the international social work conference held in Montpelier in 2002, they were adopted by the social work community in Adelaide in 2004, see Sewpaul and Jones, 2004) is another hallmark of this quest for universalism in social work (see Gray and Fook, 2004).

International but Not Global?

The phrase 'Think Globally and Act Locally' has also been used by many social workers in their day to day practice. Some scholars have argued that social work may be international but not global. Some have argued it is neither international nor global. Yet some other asked: Is there a global common core to social work?

Weiss (2005) carried out a cross-national comparative study that examined the commonalities and differences in professional ideology among social work graduates in 10 countries by studying their attitudes toward poverty and the goals of social work. The major finding is the substantial similarity in the students' professional ideology despite the different contexts of their professional socialization process. Students in all the countries attributed poverty to structural causes and supported the extension of state social welfare as a way of dealing with poverty. They expressed a large degree of support for individual well-being and social justice as major goals of social work. The similarities and differences reflect the interplay of centrifugal and centripetal forces that fashion the characteristics of global social work.

Influential ideas and scholarship on critical social work and human rights, which I believe are part of ISW, came from Asia especially from Australia (see Allan et al., 2003; Ife, 1997; Napier and Fook, 2000; Pease and Fook, 1999). Australia and Aotearoa-New Zealand remain the important centers of social work in Asia and the Pacific region, through ISW activities emanating from these centers, they have had an impact on

social work in the surrounding Asia-Pacific region and internationally (Gray and Agllias, 2013).

Criticizing international or global social work, Stephen Webb (2003) suggested that "…social work has at best a minimal role to play with any new global order, should such an order exist" (p. 191). He argues that the concept 'global social work' is problematic. His thesis, that the role of the nation state remains significant and that social work is still essentially a locally based activity, is one with which, in many respects, we concur. Webb's suggestion that extension of social work to the global level and talk of 'ISW' is essentially a thinly disguised form of intellectual or linguistic imperialism among the social professions is one to be taken seriously (Braye et al., 2004: p. 4).

Conclusion

To conclude, we come back to the question that is been asked in the beginning: What is ISW? Is 'ISW' the same thing as 'universal social work' or similar to global social work? The definition and nature of ISW, however, remains contested. The notion or the idea of 'ISW' is a concept that is defined by various scholars and practitioners depending on their location and ideology. Hence it focuses on social work which crosses national boundaries and the practice is evolving for sure. With the advancement of global technologies, increased mobility and use of Information and Communications Technologies (ICTs), and academic mobility and student international placements to name a few, ISW has the potential to become a legitimate method of social work both in the global south and north. To become so it warrants cross-cultural understanding and mutual agreement to address some of the old and emerging global issues. It actually has the power to balance the western hegemony or professional imperialism if any by recognizing the other partners are also equal and crucial to practice ISW.

See also: Child Protection: International Issues; Culturally Appropriate Interventions in Social Work; Global Agenda for Social Work and Social Development; International Association of Schools of Social Work; International Social Problems: A Systems Perspective; International Social Work Student Exchanges; Neocolonialism and Social Work.

Bibliography

Akimoto, T., 2008. What Is International Social Work? Its Contribution to Social Work in a Global Society. Presentation at the centenary of the Alice Salomon Hochschule on 23 October 2008 in the Pestalozzi-Fröbel-Haus, Berlin.

Askeland, G.A., Payne, M., 2006. Social work education's cultural hegemony. International Social Work 49, 731–743.

Allan, J., Pease, B., Briskman, L. (Eds.), 2003. Critical Social Work: An Introduction to Theories and Practices. Allen & Unwin, Sydney.

Barker, R.L., 1999. The Social Work Dictionary, fourth ed. NASW Press, Washington, DC.

Braye, S., Lyons, Karen, Preston-Shoot, Michael, 2004. Editorial. European Journal of Social Work 7 (1), 3–5.

Caragata, I., Sanchez, M., 2002. Globalization and global need: the new imperatives for expanding international social work education in North America. International Social Work 45 (2), 217–238.

Chamberlain, E.R., 1991. The Beijing seminar: social work education in Asia and the Pacific. International Social Work 34 (1), 27–35.

Chow, Nelson C.W., 1996. Social work education – East and West. Asia Pacific Journal of Social Work 6 (2), 5–15.

Council on Social Work Education, 2008. Handbook of Accreditation Standards and Procedures, fifth ed. Author, Alexandria, VA.

Cox, D., Pawar, M. (Eds.), 2006. International Social Work: Issues, Strategies, and Programs. Sage, Thousand Oaks, CA.

Dominelli, L., 2009. International comparisons in social work. In: Pearce, R., Weinstein, J. (Eds.), Innovative Education and Training for Care Professionals: A Providers' Guide. Jessica Kingsley, London, pp. 25–42.

Dominelli, L., 2012. Internationalisation and indigenisation. In: Lyons, K., Hokenstad, T., Pawar, M., Huegler, N., Hall, N. (Eds.), Sage Handbook of International Social Work. Sage, London, pp. 39–55.

Drucker, D., 2003. Whither international social work? A reflection. International Social Work 46 (1), 53–81.

Estes, Richard J., 1992a. Internationalizing Social Work Education: A Guide to Resources for a New Century. University of Pennsylvania School of Social Work, Philadelphia.

Estes, Richard J., 1992b. Internationalizing Social Work Education: A Guide to Resources for a New Century. University of Pennsylvania School of Social Work, Philadelphia.

Fook, J., 2002. Social Work: Critical Theory and Practice. Sage, London.

Friedlander, W.A., Apte, R.Z., 1968. Introduction to Social Welfare. Prentice-Hall, Englewood, NJ.

Furman, R., Coyne, A., Negi, N.J., 2008. An international experience for social work students: self-reflection through poetry and journal writing exercises. Journal of Teaching in Social Work 28 (1/2), 71–85.

Goldman, B.W., 1962. International social work as a professional function. International Social Work 5 (3), 1–8.

George, J., 1999. Conceptual muddle, practical dilemma: human rights, social development and social work education. International Social Work 42 (1), 15–26.

Gray, M., 2005. Dilemmas of international social work: paradoxical processes in indigenisation, imperialism and universalism. International Journal of Social Welfare 14 (2), 230–237.

Gray, M., Fook, J., 2004. The quest for a universal social work: some issues and implications. Social Work Education 23 (5), 625–644.

Gray, Mel, Agllias, Kylie, 2013. International Social Work and Social Welfare: Australia and Pacific Islands, Encyclopedia of Social Work. National Association of Social Workers Press and Oxford University Press.

Harris, R., 1990. Beyond rhetoric: a challenge for international social work. International Social Work 33, 203–212.

Healy, K., 2000. Social Work Practices: Contemporary Perspectives on Change. Sage, London.

Healy, L.M., Thomas, R.L., 2007. International social work. A retrospective in the 50th year. International Social Work 50 (5), 581–596.

Healy, L.M., 1995. Comparative and international overview. In: Watts, T.D., Elliott, D., Mayadas, N.S. (Eds.), International Handbook on Social Work Education. Greenwood Press, Westport, CT, pp. 421–440.

Healy, L.M., 2001. International Social Work: Professional Action in an Interdependent World. Oxford University Press, New York.

Healy, L.M., 2008. International Social Work: Professional Action in an Interdependent World, second edn. Oxford University Press, New York.

Healy, L.M., 1988. Curriculum building in international social work: toward preparing professionals for the global age. Journal of Education for Social Work 24 (3), 221–228.

Heron, B., 2005. 'Changes and Challenges: Preparing Social Work Students for Practicums in Today's Sub-Saharan African Context'. International Social Work 48 (6), 782–793.

Hokenstad, M.C., Midgley, James (Eds.), 1997. Issues in International Social Work: Global Challenges for a New Century. NASW Press, Washington, DC.

Hugman, R., 2010. Understanding International Social Work. Palgrave Macmillan, Basingstoke.

Hugman, R., Moosa-Mitha, M., Moyo, O., 2010. Towards a borderless social work: reconsidering notions of international social work. International Social Work 53 (5), 629–643.

Haug, E., 2005. Critical reflections on the emerging discourse of international social work. International Social Work 48 (2), 126–135.

Hokenstad, M.C., Khinduka, S.K., Midgley, J., 1992. Profiles in International Social Work. NASW Press, Washington, DC.

Hussein, S., Stevens, M., Manthorpe, J., Moriarty, J., 2011. Change and continuity: a quantitative investigation of trends and characteristics of international social workers in England. British Journal of Social Work 41, 1140–1157.

Ife, J., 1997. Rethinking Social Work: Towards Critical Practice. Longman, Melbourne, Australia.
Johnson, A.K., 2004. Increasing internationalization in social work programs Healy's continuum as a strategic planning guide. International Social Work 47 (1), 7–23.
Kendall, K.A., 1987. International Social Work Education, Encyclopedia of Social Work, eighteenth ed. National Association of Social Workers, Silver Springs, MD. 987–996.
Kreitzer, L., Barlow, C., Schwartz, K., Lacroix, M., Macdonald, L., 2012. Canadian and EU social work students in a cross-cultural program: what they learned from the experience. International Social Work 55, 245–267.
Lalayants, M., Doel, M., Kachkachishvili, I., 2013. Students' perceptions of international social work: A comparative study in the USA, UK, and Georgia. International Social Work. http://dx.doi.org/10.1177/0020872812473140.
Lusk, M.W., Stoesz, D., 1994. International social work in a global economy. Journal of Multicultural Social Work 3 (2), 101–113.
Lyons, K., 1999. International Social Work: Themes and Perspectives. Ashgate, Aldershot.
Lyons, K., Manion, K., Carlsen, M., 2006. International Perspectives on Social Work: Global Conditions and Local Practice. Palgrave Macmillan, Houndmills, Basingstoke.
Lyngstad, R., 2013. Contextual social work and internationalizing social work education: two sides of the same story? Journal of Social Work 13 (4), 400–418.
Marina, Lalayantsm, Doel, Mark, Kachkachishvili, Iago, 2013. Pedagogy of international social work: a comparative study in the USA, UK, and Georgia. European Journal of Social Work. Published online: 04 July 2013.
Midgley, J., 1990. International social work: learning from the third world. Social Work 35 (4), 295–301.
Midgley, J., 1997. Social work and international social development: promoting a developmental perspective in the profession. In: Hokenstad, M.C., Midgley, J. (Eds.), Issues in International Social Work. NASW Press, Washington, DC, pp. 11–26.
Midgley, J., 2001. Issues in international social work: resolving critical debates in the profession. Journal of Social Work 1 (1), 21–35.
Mohan, B., 2008. Rethinking international social work. International Social Work 51 (1), 11–24.
Nagy, G., Falk, D., 2000. Dilemmas in international and cross-cultural social work. International Social Work 43 (1), 49–60.
Napier, L., Fook, J. (Eds.), 2000. Breakthroughs in Practice: Theorising Critical Moments in Social Work. Whiting and Birch, London.
Nikku, B.R., Hatta, Z.A. (Eds.), 2014. Social Work Education and Practice: Scholarship and Innovations in the Asia Pacific. Primrose Hall Publishing Group, Australia.
Pease, B., Fook, J., 1999. Transforming Social Work Practice: Postmodern Critical Perspectives. Routledge, London.
Sanders, D., 1984. Developing a graduate social work curriculum with an international cross-cultural perspective. Journal of Education for Social Work 13 (3), 76–83.
Sanders, D., Pedersen, P., 1984. Introduction. In: Sanders, D., Pedersen, P. (Eds.), Education for International Social Welfare. University of Hawaii Press, Honolulu, pp. Xi–xxvi.
Sewpaul, V., Jones, D., 2004. Global standards for social work education and training. Social Work Education 23 (5), 493–514.
Warren, G., 1939. International Social Work. In: Kurtz, R. (Ed.), Social Work Yearbook. Russell Sage Foundation, New York, pp. 192–196.
Webb, S.A., 2003. Local orders and global chaos in social work. European Journal of Social Work 6 (2), 191–204.
Wehbi, S., 2011. Key theoretical concepts for teaching international social work. Revista de Asistenta Sociala 4, 23–29.
Weiss, I., 2005. Is there a global common core to social work? A cross-national comparative study of BSW graduate students. Social Work 50 (2), 101–110.
Xu, Q., 2006. Defining international social work. A social service agency perspective. International Social Work 49 (6), 679–692.

International Social Work Student Exchanges

Rosemary J Link, Simpson College, Iowa, IA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article identifies the goals, orientation, planning, and outcomes of international social work student exchanges. It argues that cross-cultural communication is a key skill developed by an exchange. The value base proposed includes belief in global interdependence, a willingness to scrutinize the history of student national and cultural identity, and commitment to evaluation. The knowledge base includes human rights, the history of social and economic development, and models of assessment and social justice.

Introduction

In his address to the Council on Social Work Education (CSWE) in Washington, D.C. in 2012, Tiong Tan from Singapore, asked his listeners to realize that we have to'break the mold' (Tan, 2012). Tan referred to the history of the social work profession in the Western World as still focusing on micro-systems, casework and the client as victim or person caught up in pathological relationships. Although many of the audience were radical social workers and community organizers, it is difficult to refute this imagery of the West when so much agency and therapeutic work is based on individual interventions financed through health insurance or in some countries, departments of child protection or education.

In contrast to this individualized and as experienced in some countries, 'colonial' model, Dr Tan characterized Asian social work as focused on the wider community, where the individual sees herself/himself as part of a greater good and the individual is deemphasized. Part of the challenge lies in funding sources for the profession. The exchange of information across countries can alert us to opportunities and strategies developed by other systems to separate the principles and implementation of social work practice from purely fiscal control. One example of fiscal control in the US is the influence of insurance company administrators on which claims for healthcare are appropriate. Fiscal responsibility is a professional expectation. When caught up in familiar systems, professionals may miss the wider array of practice that comes with a global view. Student exchange in particular, is a way to escape narrow vision. The focus of this article is the learning that can happen when students are encouraged into exchanges and conversations with social work students, educators and professionals in countries and cultures other than their own. In addition to conceptual arguments, reference is made to the increase in efforts to assess the impact of programs through research and evaluation (Van Hoof and Verbeeten, 2005; Center for Global Education, 2011).

The following introduction gives the background to the goals of student international exchanges in social work. Then the article develops in four sections focusing first on theory, followed by practical aspects of exchange and analysis:

- Introduction of Concepts and Theoretical Frameworks for International Exchange
- The Practical Framework and Types of Exchange
- The Future and Assessment of Exchanges
- Summary and overview of the goals and results of exchanges.

This article recognizes that there is a place for both of the approaches to professional practice that Tan refers to, plus some combination of medical/therapeutic and community/social and economic development change models. The premise of international exchange is to broaden thinking and recognize that limiting intervention to one model or another is to underserve communities. The tension becomes how to change traditions and find compatible language to share plans for effective future international and global collaboration in a fast globalizing world. Compounding this East–West divide is the current demonizing, especially in the US but also in Europe, of China's fast industrial growth and dominant market position (Editor, Economist, December 2012). For example, it was part of the Republican platform in the US during their 2012 presidential election, to cast China in the negative light of cheap labor and exploitation. An alternative view sees a vast country with much work to do in terms of fair working conditions, catching up on the same cycles of industrial and postindustrial growth that countries of the West have passed through (Midgley, 2012). India and China, are two vast countries caught up in rates of growth that strain every aspect of life (Editor, Economist, December 2012).

Letting go of ingrained stereotyping is one of the key opportunities and challenges of student exchange. Thus the first goal of student exchange is to reach beyond the familiar and to be curious about new cultural norms and the second is to realize that appreciating the cultural traditions of another does not lessen one's personal identity. As a result of a formal exchange in China, student Cali Breese illustrates this goal:

> I was invited into Peace Corps service … I thought I was ready…but each day I felt challenged to understand the dialects, the culture, strange sounds and smells, different social rules …
>
> Cali Breese, excerpt from Link and Ramanathan, Rowman and Littlefield, 2011: p. 199

Communicating effectively across cultures rests on the exchange of ideas with humility and respect, where students are eager to learn new ideas, in the context of an explicit (third) goal to reach intercultural understanding. Beyond the frank interpretation of meaning – what is trash to one, is rubbish to another is treasure to yet others – is the task to be truly curious, without feelings of superiority. It means realizing that our sense of reality is just one perspective on a complex gathering of cultures that make up the kaleidoscope of the world. Carter and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28049-5

Qureshi (1995) define culture in a way that reflects this notion of variety and equal worth: "culture is a learned system of meaning and behaviors that is passed from one generation to the next" in a geographical setting. Similarly Razack (2002: p. 251) puts culture at the center of student exchange, where colonial and oppressive history is looked in the eye and students challenge themselves to trace their generational identity.

It follows, that the theory of student exchange also focuses on challenging the 'hindsight' and superior interpretations of social and economic development. First, second or third world concepts are replaced with the (fourth) goal of recognizing that every country is part of a development cycle and all have elements of learning and human insight to share and to benefit from. A fifth goal, integration of this recognition of our universal generational experience of social and economic development, leads to more open attitudes to learning, to future career choices that include international opportunities and relationships (Bess and Link, 2012). Without this internalized recognition, social workers are at risk of ignoring valuable international policy instruments and worse, exposing themselves to unethical or illegal practice, such as the inadequate documentation of international adoptions (Rotabi, 2012; Shukovsky, 2003). These goals build the theory of student exchange, which is still relatively new, but the practice of exchange goes back much further (Lager and Mathiesen, 2012).

Thus the goals of student international exchange are about more than acquiring new knowledge and skill – they depend on a willingness to study cultural efficacy and to reflect upon values awareness. Researcher Terry Cross (1986) proposes a values continuum in relation to prejudice and racism, which offers a useful theoretical framework in organizing curriculum for an exchange. This theory is applied in the following section.

Introduction of Concepts and Theoretical Frameworks for International Exchange

International exchange in the social work profession has been flourishing for at least 100 years:

> … from the days of Toynbee Hall, the Settlement House in London visited in 1904 by the US community organizer and peace activist Jane Addams, to the early days of the Tata Institute in Mumbai and its center for international visitors. A leap forward occurred at the 1928 International Social Work conference in Paris, when social work leaders, including Addams and Jebb called for increased professional opportunities to exchange ideas and practice, with a primary goal being to promote peace in the world
>
> Cacinovic in Healy & Link, 2012: p. 343

Peace in the world was an imperative following the devastation of the nineteenth century battles in Europe and South Africa including the Boer War which ushered in the twentieth century; and the First World War (1914–17) which provided the catalyst for international exchange. Then, after the Great Depression and the eruption of the Second World War (in Europe 1939–45), there was rapid expansion in international organizations, pulling East and West into dialogue. Most central to the new forums for exchange during the 1930s was the League of Nations, and after 1948 the United Nations. Another step was the Commonwealth Youth Exchange Council, established in Nairobi and London in 1965. However, the domination of the norms of colonial power slowed true exchange. This is illustrated in the context of language that perpetuated attitudes of, First, Second, and Third worlds, first used in economic and development sociology (Midgley, 2012).

In 2013, the numbering of worlds in terms of economic and political status has been abandoned and the scene may be set for more genuine exploration of cultural meaning and exchange. Instead of a WalMart US style organization expecting its staff to smile broadly at strangers and stay open on Sundays in Europe, corporations are learning the hard way how to be successful – by listening to local custom. The transnational corporation 'Dairy Queen (DQ)' set up store in Ljubljana in Slovenia, in a country that has the best ice-cream and gelato outside Italy. It should be no surprise that DQ soon left, having given scant attention to effective forms of cultural exchange and underappreciating what was already in existence (Link and Cacinovic, 2012). This failure of a company to respect local norms is an important lesson for social work students reflecting on their own cultural awareness and communication.

The theoretical frameworks that apply to this exploration of international exchanges are those of Norton and the 'Dual Perspective,' Cross's 'Cross Cultural Continuum' and Bosch's 'cultural efficacy.' In Cross's continuum, students share their ideas about culture and come to realize that 'I respect everyone, I treat everyone the same' translates to the limiting 'same as me' (Cross, T. 1986: p. 6). Norton prompts initial awareness of how a person has the advantage when they join a workplace or community that reflects many of the same cultural norms and patterns that someone is raised with and in particular seems to put newcomers, migrant workers or exchange students at a disadvantage (Norton, 1997). Assignments and orientation to student exchange can be organized in ways that infuse these theoretical frameworks and focus on values and styles of communication.

In organizing student exchanges it is widely recognized that for some students this is their first foray into new lands and cultures (Mathiesen and Lager, 2007). There is a natural ignorance which can be compounded by lack of realization of the layers of norms that we build over a lifetime (Cross, 1986: p. 6). Norton speaks of students who settle easily into higher education because the context is familiar. A student at the University of Ljubljana, who speaks Slovenian, has ties with Italy settles into class in Ljubljana more easily than the student from Croatia or Bosnia who speaks different languages or dialects, experiences life on the outside of the European Union and is still caught up in the tensions of war.

Henna John-Fisk is researching the resilience of children of prostitutes through an exchange in Mumbai, India (Henna Fisk, 2012: p. 147). Part of John-Fisk's data highlights the preference of the children interviewed, to stay with their mothers. The research reports that the children understood what was pressuring their parent into prostitution and wanted to stay with them. For a Western reader, ignorant of the cultural ties in Mumbai that take into account the community of prostitutes, the oppression of the control of prostitutes by male handlers, it is easy to be shocked at the idea of a child staying with her mother in such circumstances. This is the cultural norm and professional response, even though the women

involved are often controlled by exploitive and violent men. It can be argued that there is a secondary oppression by society's attitude to their capacity for mothering.

It is realized that student exchange is multilayered and the theoretical sources referenced here become key design elements for educators. Lager and Mathiesen have identified 10 'facilitating factors' which are a combination of theoretical and practical approaches to student exchange in the context of international field placements. Five of these are:

1. A communication feedback loop with supervisor, field liaison, and student with mutual evaluation and debriefing.
2. Contact should be maintained at the micro, mezzo, and macro with the host country.
3. The common ethical framework established by the principles established by the IFSW is adopted by all participants.
4. A reciprocal strengths-based approach for all parties is assumed.
5. Mutual goals (Lager and Mathiesen, 2012: p. 338).

These concepts for building international exchange lead directly into the practical planning, orientation, implementation, evaluation, synthesis, cycles of exchange. The concepts also emphasize the important of ongoing learning and the message to students that this is a journey to be continued when they return from their exchange with their new perspectives and learning.

Practical Frameworks and Types of Exchange

An underlying theme of this article is: how can international exchange transform later professional behavior. What are some of the goals and challenges that need to be addressed, what are the best practices in international student exchange and what variety of forms do they take. The increase in commitment to exchanges is facilitated by the fact that the core professional international organizations for social workers endorse exchange, including the International Association of Schools of Social Work.

In the second decade of the twenty-first century examples of international exchanges are expanding fast and include: the Singapore Institute of Management which has extensive ties and professional exchange with social work programs in China (Tan, 2012); Harvard University in the US, which expects all students to have passports and to travel. Most of the universities of the European Union participate in former European Community Action Scheme for the Mobility of Univesity Students (ERASMUS) and now Socrates and Tempus (since the end of the Soviet Union) exchange programs:

Celebrating its 25th anniversary in 2012, ERASMUS is the most successful student exchange program in the world. Each year, more than 230 000 students study abroad thanks to the Erasmus program. It also offers the opportunity for student placements in enterprises, university staff teaching and training, and it funds cooperation projects between higher education institutions across Europe (www.ec.europa.eu/education/lifelong-learning-programme/erasmus).

Schools in India, including the Tata Institute in Mumbai, have long-standing relationships with employers in the Western World, and participate in student field placement exchange and international conferences. The National University of Singapore has extensive ties with India and the International Association of Schools of Social Work has just published a report on the 'IASSW and Indonesian Social Work: Building Capacity Together' (www.iassw-aiets.org, 24.01.13). An outstanding feature of this growth is the recognition that international student exchange contributes to curriculum development for faculty as well as learning amongst students.

Table 1 demonstrates the widening goals of exchange.

International exchanges occur through a variety of formats including:

- physical two-way exchange between host and sending organizations (including field placements)
- interactive video exchange between classrooms in different countries
- instructional exchange online.

Asamoah (2003) speaks of the unbalanced approach to exchanges in the past, with schools in the Western World designing the goals and expecting those visited to grasp their curriculum design and expectations:

> Many of the early exchange models are most accurately described as one-way consultations without a true exchange component, and some models lacked cultural relevance and sensitivity…
> Asamoah 2003: p. 3

The first step in planning then is to review the goals of the exchange with potential partners and to design the curriculum

Table 1 Goals for exchange

- Identifying a variety of solutions to common problems, such as child abuse.
- Encouraging social work students to question their own sense of cultural identity and 'norms' so that they become more accepting and welcoming of others.
- Expanding the knowledge of a variety of countries amongst students so that they understand the original impetus for immigration.
- Developing a sense of the strength of the profession worldwide to foster collaboration, especially at borders and in questions of human rights such as international adoptions.
- Building a curriculum that is relevant beyond the local situation which recognizes issues that are common to social workers worldwide, such as access to health, healthy pregnancy, access to safe environments for vulnerable people including older people, people with disabilities and children.
- Sharing and expanding theory and practice, for example the work in Ljubljana to 'cocreate' solutions for families. This work builds on the idea of a 'fourth wave' in social work practice – the first being community work, the second clinical practice and the medical model, the third being psychosocial work with an understanding of the person in their environment and the fourth being the expectation of working alongside rather than ahead or above service users (Link and Cacinovic, 2012, in OUP Handbook).

in outcomes language, for example: as a result of this exchange students will be able to explain the reason for professional adherence worldwide to the Convention on the Rights of the Child. The most successful physical and practical exchanges seem to be those where the emphasis in on institutionalizing the planning so that it endures changes in personnel, while also emphasizing the importance of building true friendly relationships. Hokenstad (2003) describes the ongoing exchange between the Department of Social Work at Eotvos Lorand University (ELTE) in Budapest and the Mandel School of Applied Social Sciences at Case Western Reserve University (CWRU) in Cleveland. He identifies advanced planning, regular contact and faculty sharing of curriculum innovation. Similarly the University of Ljubljana, Slovenia has developed multiple exchanges with partners in Europe, including Barcelona and Finland as well as long-standing exchanges with North America. Such enduring exchanges require the following planning, orientation, implementation and evaluation phases:

- in depth planning for each stage;
- visits between leaders through interactive video or preferably in person, including to experience ahead of the students where possible, accommodations, food, access to water and travel arrangements;
- shared goals for orientation in advance of travel;
- readings recommended by host setting for each leg of the exchange;
- frank discussion of cultural expectations and norms;
- norms for travel (partners; avoiding clothing with national slogans; attention to nonverbal behavior; adherence of local laws, for example, in relation to alcohol use, gum chewing, spitting, sneezing); and
- review of evaluations for implementation of changes and improvements each time.

Such detailed planning also includes crisis management and strategies for student misadventure or ill health. During an exchange in Slovenia, which included brief field agency experiences, the group leader (Link) was telephoned with the news that 'Victoria has blown up.' While taking the fast train to investigate, the group leader was reviewing her conflict negotiation skills. On arrival, she found that language had intervened to send the wrong message: the student 'threw up.' Language, particularly when participants are under the impression that they share it, such as English, can lead to much misinterpretation and deserves attention (Moss, 1994).

Planning leads to the detailed 'orientation' and preparation for learning prior to travel. The nuts and bolts of successful exchanges are characterized with well-documented organization such as demonstrated the Center for Global Education. In their 25th anniversary book, they address the elements of their orientation that have evolved over the years:

> As educators, we see our role as one that engages students and participants in the world, facilitating critical analysis and reflection that leads to action. We believe that intercultural dialogue and collaboration with decision makers and historically disadvantaged urban and rural communities are a way of developing greater understanding of the power relations in the world.
>
> McBride, 2007: p. 2

Orientation takes many forms but essentially the goal is to engage students in the exchange prior to travel. Key elements include identifying the theoretical framework, such as a human rights perspective for a social work exchange between Slovenia and the US, or as above, a feminist perspective for a program of exchange with Mexico and Nicaragua. As discussed above, cultural awareness is both a goal and an activity for students as they begin to apply their theoretical learning. Lyons and Ramanathan (1999: pp.175–176) identify activities during orientation including:

- Beginning a journal.
- Small group discussion of advantages and disadvantages of undertaking an international field placement exchange.
- Write a paper for presentation about the country you will be going to. What are its main demographic, geographic, historic, economic, cultural and political characteristics?
- Using these discussions draw up a checklist and action plan for departure and arrival in a new country.

Clearly orientation also includes all the practical matters of organizing a group and these include: a sense of geography is essential so maps have to be studied and local conditions including weather and terrain; travel arrangements are made well in advance so that students have necessary visas and passports, health preparation and immunizations are scheduled, students practice social norms, such as greetings, supplies are planned including water purification tablets, warm clothing, travel, health and loss insurance, finance planning and access to necessary currency. Guidance on international exchange can be found in many locations including the US State Department country information, Lonely Planet Gravel Guides and the World Health Organization country updates (www.who.org).

This short article on international exchange will not dwell on the details of implementation other than to identify key elements as listed above and references and resources. It is etched on this writer's memory to instill travelers with a knowledge of their surroundings and the importance of knowing where they are on the globe. Flying has robbed students of the sense of geography and borders.

Similarly, regular processing of events as an exchange unfolds assists students to see the relevance of their learning to their own practice.

In a program with the Center for Global Education in Mexico, a group called 'Luce Y Libertad' agreed to publication of a case study as long as it was in both Spanish and English so that they could share it with their community. This exchange between students and the women's self-esteem project led to direct learning for the students. They participated in classes on organizing themselves as a food cooperative and took their learning back to the MidWest, US (Link and Ramanathan, 2011).

In addition to physical, face-to-face exchanges, there is increasing attention to interactive video exchange between different countries, sometimes in a dyad and sometimes expanding to a trio of relationships. In 2004, the University of Singapore agreed to participate in a video exchange with Augsburg College, Minneapolis in the US, sponsored by the State Department. The project was one of 10 interactive video programs offered during 2004 and 2005 and is available both

through US State Department records and narratives by the faculty and staff involved (Link/Bill, 2005).

The interactive video classroom involves similar cultural awareness, practical planning and thoughtful exchange of ideas concerning curriculum to the planning for a physical exchange. The most notable differences lie in the protocols that the students themselves develop that include ways of nonverbally taking turns to converse; greetings; use of language; attention to process (Link and Bill, 2005).

An extension of the use of technology in multiple forms, such as Skype, Facebook, Twitter, and all other social media, is the opportunity for students to connect with one another outside the virtual classroom, to study, to become friends and to share insights about their practice. Technology is racing ahead and the expectations of the twenty-first century workplace mean that students require experience and practical assistance with online learning. Many universities have courses completely online so that it is a natural form of exchange for students to find themselves in virtual classrooms with students across the globe. The distinction here is that the online course is content driven and disciplinary based, and does not address the issues of cultural exchange expected in the true formal international exchange courses.

Also, many schools are now involved in a US initiative begun at universities such as the Massachusetts Institutes of Technology and Harvard, termed 'Massively Open Online Courses' or MOOCs. These courses are offered free of charge, enroll hundreds of students and are taught by professors who offer certificates of completion. There is lively debate as to the disruption caused by MOOCs and the positive stimulus for administrators in Higher Education to think more creatively about open access for students of all income levels. It follows that as courses become available online worldwide, that education will become more widely accessed in a variety of countries. Although MOOCs have yet to spread to social work, they are on their way as a new form of exchange.

The Future and Assessment of Exchanges

While every exchange will have its own itinerary and level of involvement with local social workers and human service professionals, there are some universal aspects to endings and evaluation. As described by the Center for Global Education (CGE) all of their courses include oral and written evaluations and follow-up meetings at 're-entry' to home country (CGE, 2007). Similarly, one of the enduring features of the Slovenia-Midwest US exchange has been the thorough review of feedback so that learning is built-in to the next iteration of the exchange. For example, 1 year costs were reduced via less expensive accommodations in a hostel used by many travelers, but the location, noise levels, and general stress of the accommodation led to re-instating an earlier location.

Van Hoof and Verbeeten undertook a detailed research project, including but not focused on human service, surveying 1487 undergraduate students from a variety of countries who were incoming or US students outgoing on courses abroad. They achieved a 23.74% response rate and although a limitation lies in the focus on one large, public US university, the data is useful. For example, the reasons for international exchange and studying 'abroad' were:

1. it is a good opportunity to live in another culture;
2. it is a good opportunity to travel;
3. I liked the country my exchange program was located in…
4. and when it came to determining what it had meant to their personal development, the respondents were most enthusiastic: 67.7% considered it extremely relevant (Van Hoof and Verbeeten, 2005: p. 42).

One of the most vivid quotes occurs in the introduction to Van Hoof and Verbeeten's evaluation research. A student, Penny Barend, asked for water at the first supper with her 'host' family in Italy, only to be told: "Wine is for drinking, water is for washing" (Van Hoof and Verbeeten, 2005: p. 44).

In a study of exchanges, Vestal (1994) spoke of the 'promise of today' and the expectations of the next generation of students. It is nearly 20 years since Vestal foresaw the expansion in student interest and recognition of global interdependence (Vestal, 1994). Students are recognizing the ideas of Dower and Williams research, that national borders are now false limits on knowledge and research (Dower and Williams, 2002). The diseases such as severe acute respiratory syndrome (SARS) and acquired immunodeficiency syndrome (AIDS) cross all boundaries and natural disasters unite everyone on the plant in the need for emergency preparedness and collaboration. At this time the United Nations is the focus for international learning and the work of the UN in establishing policy frameworks for practice are bearing fruit. An example of this is a simple one: staff at the adolescent refuge in Celje, Slovenia, say that it is a natural part of their work to use the Convention on the Rights of the Child. This contrasts with students in the US who live in one of the last countries in the world that is yet to ratify the Convention on the Rights of the Child.

Conclusions

As stated earlier, an underlying theme of this entry is, how can international exchange transform later professional behavior? The examples cut across all forms of social work intervention. Some of the impact of exchange dialogue is highlighted in the example of the use of the UN Convention on the on the Rights of the Child. Another is the impact of the University of Mona, West Indies on practice with elders in Hartford, following the longstanding exchange with the University of Connecticut, School of Social Work Center for International Social Work. The social workers at the Seattle Adoption Agency may wish they had been more involved in formal social work exchange to avoid their malpractice (Shukovsky, 2003).

This narrative has identified at least four key goals for international exchanges:

- to reach beyond the familiar and find new ways of doing familiar things;
- to be curious about new cultural norms and in doing so to realize that appreciating the cultural traditions of another does not lessen one's personal identity;
- to communicate effectively across cultures and make new friends in distant places of the planet; and
- to expand innovations, effective strategies and implementation of international policy instruments in social work practice.

These are just examples of goals highlighting the need for expanded dialogue across cultures and regions of the world. The goals undergird the hard and detailed work of making practical arrangements, implementing and assessing courses.

Apparently there is much thinking to do if we are to understand one another better; thinking and talking and walking together. Student international exchanges represent a key step toward this talking with one another. If social work exchanges continue to build, the next generation will take the new 'Global Agenda' seriously. The Global Agenda combined the work of the International Federation of Social Workers (IFSW), the International Association of Schools of Social Work (IASSW), and the International Council on Social Welfare (ICSW) and was agreed at a conference in Hong Kong in March 2012 (IFSW, 2012). The agenda includes 'a set of objectives to meet our joint aspirations for social justice and social development' (IFSW, 2012). The Global Agenda expects students and educators, through their exchanges, to join in communicating and organizing in solidarity with social work practitioners across the world. To be effective as a profession in a globalizing world, we have to seize the moment and build international exchanges.

See also: Afrocentric Approaches to Social Work.

Bibliography

Asamoah, Y., 2003. International collaboration in social work education: overview (Chapter 1). In: Healy, L., Asamoah, Y., Hokenstad, M.C. (Eds.), Models of International Collaboration in Socoial Work Education. CSWE, Alexandria, VA, p. 3.

Bess, A., Link, R.J., 2012. International Careers in Social Work (Chapter 71). In: Healy, L.M., Link, R.J. (Eds.), Handbook of International Social Work. OUP, New York.

Braley, M., Dybvig, N., 2012. Full-time volunteers, Christian Peacemaker teams, based in Barrancabermeja, Colombia. Chapter 71, international careers in social work. In: Bess, A. (Ed.), Handbook of International Social Work. OUP, New York, p. 482.

Breese, C., 2011. Peace corps volunteer: China. In: Link, R.J., Ramanathan, C.S. (Eds.), Case Study in Human Behavior in a Just World. Rowman & Littlefield, New York, p. 199.

Carter, R.T., Qureshi, A., 1995. A typology of philosophical assumptions in multicultural counseling and training. In: Pontevotta, J., Suzuki, L.A., Alexander, C. (Eds.), Handbook of Multicultural Counseling. Sage, Thousand Oaks, CA, pp. 239–262.

Center for Global Education, 2007. 25th Anniversary Issue. Augsburg Press, Minneapolis.

Cross, T., 1986. Cultural Competence Continuum. University of Washington Press, Seattle.

Dominelli, L., 2010. Social Work in a Globalizing World. Polity Press, Cambridge.

Dower, N., Williams, J., 2002. Global Citizenship: A Critical Introduction. Routledge, New York.

Economist, December 2012a. India's Bridge of Dreams: Halfway to Paradise, p. 119.

Economist, December 2012b. China and India's Bridge to Nowhere.

Hokenstad, M.C., 2003. Collaboration in an era of social transition: the Eotvos Lorand University, Hungary-Case Western Reserve University Connection. Chapter 5 (Chapter 1). In: Healy, L., Asamoah, Y., Hokenstad, M.C. (Eds.), Models of International Collaboration in Social Work Education. CSWE, Alexandria, VA, p. 3.

IFSW, 2012. The Global Agenda for Social Work and Social Development. http://ifsw.org (retrieved 02.01.13).

International Association of Schools of Social Work: IASSW and Indonesian social work: building capacity together. www.iassw-aiets.org.

John-Fisk, H., 2012. Voices of Children of Prostitutes in Mumbai. Presentation and abstracts to the Council on Social Work Education Annual Program meeting. CSWE, Aleandria, p. 147.

Lager, P.B., Mathiesen, S.G., 2012. International field placement (Chapter 51). In: Healy, L.M., Link, R.J. (Eds.), Handbook of International Social Work. OUP, New York, pp. 337–342.

Link, R.J., Bill, R., 2005. Lessons without Leaving. Presentation to the Annual Program Meeting, Council on Social Work Education, Nashville.

Link, R.J., Cacinovic, V.G., 2012. Models of international exchange (Chapter 52). In: Healy, L.M., Link, R.J. (Eds.), International Handbook of Social Work. OUP, New York.

Link, R.J., Ramanathan, C.S., 2011. Leaving footsteps. Textbox 2.2 example of reality analysis in chapter 2, common human condition. In: Human Behavior in a Just World. Rowman & Littlefield, New York, pp. 32–35.

Lyons, K., Ramanathan, C., 1999. Models of field practice in global settings (Chapter 10). In: Ramanathan, C., Link, R. (Eds.), All Our Futures. Brooks/Cole, Pacific Grove.

Mathiesen, S.G., Lager, P.B., 2007. A model for developing international student exchanges. Social Work Education 26 (3), 280–291.

McBride, K., 2007. Center for Global Education: Building a Just and Sustainable World: Educating for Transformation. CGE publication, Minneapolis, p. 2.

McIntosh, P., 1988. White Privilege and Male Privilege: A Personal Account of Coming to See Correspondences through Work in Women's Studies. Wellesley College, Center for Research on Women, Wellesley, MA.

Midgley, J., 2012. Development (Chapter 4). In: Healy, L.M., Link, R.J. (Eds.), Handbook of International Social Work. OUP, New York.

Moss, N., 1994. British/American Language Dictionary for More Effective Communication. Passport Books, Chicago.

Norton, D., 1997. Dual perspective. In: Lortie, et al. (Eds.), Human Behavior in the Social Environment. Brooks/Col/Thomson, Pacific Grove, CA, pp. 4–6.

Razack, N., 2002. A critical examination of international student exchange. International Social Work 45 (2), 251–265.

Rotabi, K., 2012. International Adoption (Chapter 11). In: Healy, L.M., Link, R.J. (Eds.), Handbook of International Social Work. OUP, New York.

Shukovsky, P., 2003. Feds Claim Adopted Orphans had Parents, U.S. Agents Break up Local Agency Dealing in Cambodia. Seattle Post-Intelligencer, 17 December, p. A1.

Tan, T.N., November 2012. East West Divide in Social Work. Plenary presentation to the Annual Program Meeting of the Council on Social Work Education, Washington, D. C., USA.

Van Hoof, H.B., Verbeeten, M.J., Spring 2005. Wine is for drinking, water is for washing: student opinions about international exchange programs. Journal of Studies in International Education 9 (1), 42–61.

Vestal, T.M., 1994. International Education: Its History and Promise for Today. Praeger, Westport, CT.

Relevant Websites

www.who.org – World Health Organization.
www.lonelyplanet.com – Lonely Planet

International Trade: Commercial Policy and Trade Negotiations

Andreas Dür, University of Salzburg, Salzburg, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Commercial policy has a strong negotiated component, with many international trade negotiations aiming at the liberalization of trade. Why do we see these trade negotiations, that is, why do countries even aim to liberalize trade? What explains whether countries engage in multilateral or bilateral and plurilateral negotiations? And how do countries negotiate over trade, that is, which role does power play and what explains the choice of bargaining strategy in commercial negotiations?

Commercial Policy

Most countries today are more open to international trade than ever before in their history. While trade flows were severely curtailed in the interwar period, since the end of World War II trade has grown substantively faster than world output. In 2011, therefore, world merchandise trade amounted to no less than US$18 trillion (World Trade Organization, 2013: 5). International trade no longer only concerns trade in goods; services trade, for example, the cross-border provision of financial services, also plays a crucial role in the expansion of trade. Moreover, countries trade an increasing amount of intermediary goods, leading to the establishment of global value chains and global production networks (Elms and Low, 2013). As a consequence, the world economy has become highly dependent on world trade.

The rapid growth in cross-border trade has been driven by technological innovations that have reduced the costs of trade. Not only have shipping and air travel become less expensive, but also the cross-border provision of services has become much easier as a result of the Digital Revolution. Next to technological changes, however, also the commercial policies adopted by countries across the globe have facilitated the growth in trade. Commercial policy is the part of government policy that directly aims at influencing the amount of trade that crosses borders. This can mean boosting trade, for example, by reducing barriers that hinder imports. Trade barriers include both tariffs and behind-the-border measures, such as discriminatory standards, that limit foreign imports. Boosting trade can also be achieved by facilitating or even subsidizing exports. Commercial policy can also be protectionist, that is, be aimed at reducing trade, by increasing trade barriers both at and behind the border.

Commercial policy involves unilateral trade policy measures, such as unilateral increases or reductions in tariffs. In the United States, for example, the Smoot Hawley Act (1932) unilaterally increased tariffs across most imports, leading to a substantial decline in foreign imports. More recently, many emerging economies unilaterally reduced their tariffs, with Chile a particularly prominent example (Edwards and Lederman, 1998). Some trade policy measures are also inherently unilateral. This is the case for antidumping actions and countervailing measures. The former are duties imposed when a country suspects that another country sells a product below the price of production. Countervailing measures are a response to subsidies given by another country. Because commercial policies do not only affect the own country but also trading partners, however, commercial policy also has a strong negotiated component. In fact, ever since countries engage in trade, they have signed agreements that facilitate or restrict trade between two or more countries.

Commercial policy measures can be applied preferentially or nonpreferentially. A tariff cut, for example, can be extended to only a few countries (for example, in form of a preferential trade agreement (PTA)) or to all countries benefitting from 'most-favored-nation' treatment. Most-favored-nation treatment means that a country receives at least the same if not more trade advantages than any other country. In the words of Jackson (1997: 160–161), it establishes "an obligation to activities of a particular foreign country or its citizens at least as favorably as it treats the activities of any other country." Antidumping duties or countervailing measures create preferential treatment, as they target specific suppliers or specific countries.

Overall, then, commercial policy measures can liberalize or protect a market; can be imposed unilaterally or be the result of negotiations; and can be applied preferentially or in a nonpreferential manner. **Table 1** summarizes these three dimensions that can be used to characterize commercial policy. It offers examples for seven of the eight possible types of commercial policy measures. The eighth – a negotiated, nonpreferential protectionist measure cannot be observed in practice.

Trade Negotiations as an Instrument of Commercial Policy

Trade negotiations are a key instrument of commercial policy. Most trade negotiations aim at the liberalization of trade, that is, the removal of trade barriers that hinder the flow of goods or services across borders. After World War II, 23 countries signed the General Agreement on Tariffs and Trade (GATT), which foresaw the stepwise liberalization of trade following the principle of reciprocity. Reciprocity means that countries exchange tariff concessions; one country agrees to reduce the tariff on a product in exchange for another country reducing the tariff on the same or another product. Establishing reciprocity, especially among many countries, required trade negotiations.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71040-3

Table 1 A typology of trade policy measures

	Unilateral		*Negotiated*	
	Preferential	*Nonpreferential*	*Preferential*	*Nonpreferential*
Liberalizing	The unilateral removal of a specific trade barrier that only benefits one or a few countries	A tariff cut that is applied to all countries enjoying most-favored-nation treatment	A preferential trade agreement that liberalizes trade among two or a few countries	A trade agreement that leads to liberalization on a most-favored-nation basis
Protectionist	The imposition of antidumping duties that only affect one or a few countries	A tariff increase that is applied to all countries	A voluntary export restraint agreement	–

Note: the cell entries provide examples for the various measures. They do not represent a comprehensive list of trade policy measures.

Such trade rounds in the framework of the GATT took place in 1947, 1949, 1950–51, 1956, 1960–61, 1964–67, 1973–79, and 1986–93 (see **Table 2**).

The most important of these trade rounds, the Kennedy Round in the 1960s, the Tokyo Round in the 1970s, and the Uruguay Round in the 1980s and 1990s resulted in major tariff cuts and also the reduction of other barriers to trade (Cohn, 2002; Brown, 2003). The Uruguay Round also led to the creation of the World Trade Organization (WTO), which replaced the GATT in 1995. In the framework of the WTO, a major trade round – known as Doha Development Agenda – was started in 2001, with the negotiations still ongoing.

Trade negotiations also take place outside the GATT/WTO. Especially, the last two decades have seen the rapid spread of PTAs that are the result of negotiations between two or among a few countries (Mansfield and Milner, 1999; Dür et al., 2014). Examples are the North American Free Trade Agreement (signed in 1992), the agreement between Japan and Switzerland (signed in 2009), and the Australia-New Zealand Closer Economic Relations Trade Agreement (signed in 1983). At the time of writing, Australia, Chile, Japan, Mexico, the United States, and a series of other countries engage in negotiations for a Trans Pacific Partnership agreement and the United States and the European Union envisage the conclusion of a Transatlantic Trade and Investment Partnership.

Not all trade negotiations aim at liberalizing trade. Especially in the 1970s and 1980s, countries negotiated 'voluntary' export restraint agreements that restrict international trade (Bhagwati, 1988: 44). In these agreements, an exporting country agrees to reduce exports of a specific product to avoid the imposition of trade barriers by the importing country. A prominent example for a voluntary export restraint is the 1981 agreement by Japan to reduce exports of cars to the United States.

Why Do Countries Engage in Trade Negotiations to Liberalize Trade?

Trade negotiations thus mainly aim at the liberalization of international trade. But why are countries even interested in trade liberalization? It is generally agreed that free trade maximizes the welfare of all countries because it allows for goods and services to be produced in the most efficient locations. Free trade allows countries to specialize on producing the goods or providing the services for which they have a comparative advantage, meaning that they can produce or provide them at a lower opportunity cost than in another country. Countries thus do not need to have an absolute advantage in producing some goods or providing some services in order to benefit from trade; as long as countries differ in their relative efficiencies in producing different goods, all countries benefit from trade.

Trade, however, also has distributional effects (Alt and Gilligan, 1994). An expansion in trade redistributes wealth in a society, that is, it benefits some and hurts others. Two theorems offer contradictory predictions about who will win and who will lose from an expansion in trade. The Stolper–Samuelson theorem predicts that the relative price of the factor that is used most intensively in the production of

Table 2 Major multilateral trade negotiations

Name	*Time*	*Achievements*
Geneva	1947	Creation of the GATT and tariff reductions
Annecy	1949	Tariff reductions
Torquay	1950–51	Tariff reductions
Geneva II	1956	Tariff reductions
Dillon	1960–61	Tariff reductions
Kennedy	1964–67	Tariff reductions and an antidumping code
Tokyo	1973–79	Tariff reductions and several codes for nontariff barriers
Uruguay	1986–93	Tariff reductions, nontariff barriers, service liberalization, foreign direct investment protection, intellectual property rights protection. Creation of the WTO
Doha Development Agenda	2001–ongoing	The Bali package agreed upon in 2013 covers trade facilitation, food security, and preferential rules for least developed countries

a good will increase as production of the good expands. By contrast, the owners of the factor of production that is used less intensively in the production of this good will lose from an expansion of production. In a country exporting a labor-intensive good such as textiles, consequently, trade liberalization will result in an increase in wages. At the same time, the returns on capital will decline as the country increases its exports of textiles. This theorem builds on the assumption of perfect factor mobility, that is, it assumes that a factor of production can be moved without costs from one sector of production to another.

The Ricardo–Viner model, by contrast, assumes that factors of production are specific to a sector. This may be because the capital used in one sector is different from the capital used in another sector. For example, a cotton spinning machine cannot be used in the production of cars. Following this model, an expansion in exports of a good will increase the price of capital and labor employed in the specific sector; and decrease the price of capital and labor in other sectors. The expectation then is for capital owners and labor in exporting industries to favor trade liberalization, and capital owners and labor in import-competing industries to favor protectionist trade policies.

Several studies have tried to find out which of these two models better accounts for trade preferences (Magee et al., 1989; Rogowski, 1989; Hiscox, 2002). Magee et al. (1989) found support for the Ricardo–Viner model. At the same time, the Stolper–Samuelson theorem has been used to show that increases or decreases in trade, for example, those resulting from cheaper transportation, can have major political consequences (Rogowski, 1989). Socialism may have been a response to the expansion of trade in a backward economy with a low land-labor ration. West European Fascism may have resulted from a decline in trade in an advanced economy with a low land-labor ratio. Hiscox (2002), finally, showed that factor mobility varies over time; sometimes, then, the Stolper–Samuelson theorem applies and at other times the Ricardo–Viner model.

For many years, these distributional effects of trade have been used to explain countries' choices between protectionism and trade liberalization (Alt et al., 1996). The expectation has been for the losers from trade to push for protectionist trade policies; and the winners to support trade liberalization. Both the losers and the winners may create pressure groups that push for their respective stances by donating to decision makers, providing information or mobilizing the public. The political economy of trade then expects a government to take into account the relative strength of the two sides when designing its commercial policy. If factor mobility is high, the demands for protection or trade liberalization may also be reflected in party programs. In either case, the expectation is that whenever the protectionist side is stronger, the country will implement a protectionist trade policy; whenever the free trade side is stronger, the country will opt for trade liberalization.

Early on, however, scholars realized that the two sides in the trade policy debate may not be equally capable of getting organized and influencing government policy. In fact, traditionally the proponents of protectionist trade policies have been seen as having an advantage in mobilizing for political activity. The highly protectionist Smoot–Hawley Act that US Congress passed in 1932, for example, may have been the result of the highly effective mobilization of protectionist interests, whereas the free trade interests were all but absent from the political scene (Schattschneider, 1935). In this view, the benefits of protection are concentrated on a few producers, while the costs of protection are distributed among many consumers. Collective action problems (Olson, 1965) hence are seen as benefitting protectionist interests.

What then explains the move from protectionist trade policies to trade liberalization that we can observe just before and after World War II? Much research has focused on the passage of the Reciprocal Trade Agreements Act (RTAA, 1934) in the United States as the initial impetus for trade liberalization. The RTAA empowered the President to negotiate reciprocal trade agreements with trading partners. In a series of negotiations, the President used these powers to open foreign markets and bring down US trade barriers. For many, this piece of legislation, by shifting the responsibility for trade policy from Congress to the executive, allowed for the subsequent liberalization of trade (Lohmann and O'Halloran, 1994; Bailey et al., 1997; Gilligan, 1997). The institutional change may have been important because the President has a larger constituency than members of Congress and thus needs to take into account the negative externalities of protectionist policies for the welfare of the country as a whole (Lohmann and O'Halloran, 1994). The trade agreements that the president then negotiated with third countries locked in lower tariffs and made it more difficult for the at that time predominantly protectionist Republican Party to raise tariffs once back in control of Congress (Bailey et al., 1997). By concentrating the benefits of reciprocity on particular exporters, moreover, reciprocity created a powerful lobby for free trade (Gilligan, 1997).

These arguments, however, have not gone unchallenged. Referring to the RTAA as a 'magic bullet,' Hiscox (1999) argued that this account may explain what happened after the RTAA was passed, but cannot account for the passage of this piece of legislation. He argues that the act only passed because the back then free trade oriented Democratic Party controlled Congress at that time. The RTAA also did not lead to a stark reversal in the strength of protectionist and free trade interests in the United States in the following decades. Rather, for exogenous reasons, namely a shift in the geographic distribution of production in the United States, the Republican Party became increasingly divided with respect to trade policy. This allowed the RTAA to survive in the 1940s and 1950s even after the Republicans regained a majority in Congress.

The explanation provided in Dür (2010) also runs counter the institutionalist explanation. Here the argument is that twentieth-century trade liberalization was a response to the discrimination created first by the system of imperial preferences implemented by the United Kingdom and its dominions and then the process of European integration. Congress passed the RTAA only 2 years after the Ottawa agreements established the system of imperial preferences. Both societal interests and decision-makers supportive of the RTAA clearly alluded to the negative consequences of the Ottawa agreements for American exporters in making their case for the RTAA. Later, the creation of the European Economic Community in the late 1950s reinvigorated the American interest in trade liberalization, leading to the far-reaching trade liberalization agreed upon in the Kennedy Round. Most recently, following this account, the

discrimination created by the spread of PTAs propelled further trade liberalization.

Trade liberalization may also result from a change of ideas (Goldstein, 1993; Schonhardt-Bailey, 2006). In the nineteenth century, economic ideas may have played a major role in the British unilateral shift from protectionism to free trade, as exemplified by the repeal of the highly protectionist Corn Laws (Schonhardt-Bailey, 2006). More recently, the perception that the protectionist policies of the interwar period contributed to the Great Depression may have made decision-makers favor free trade policies. What is not completely clear from this explanation, however, is why most trade liberalization was achieved in international trade negotiations, rather than implemented unilaterally.

When Do Countries Engage in Multilateral and When in Bilateral or Plurilateral Trade Negotiations?

Commercial negotiations comprise only either two or a few countries (bilateral or plurilateral negotiations) or many countries (multilateral negotiations). The results achieved in trade negotiations, moreover, can be applied preferentially or on a most-favored nation basis. For many years, countries have been granting most-favored nation status to other countries (Horn and Mavroidis, 2001). Most-favored nation treatment also became a key element of the trading system that emerged after World War II. A key principle of the multilateral trade rounds taking place in the framework of the GATT/WTO is that any concessions offered by a country to another are extended to all members of the GATT/WTO.

At the same time, many bilateral or plurilateral trade negotiations aim at the establishment of PTAs that liberalize trade between two or among a few countries without extending these concessions to third countries. From a welfare perspective, under many circumstances PTAs are inefficient, as they may lead to the substitution of more efficient imports by less efficient production within the preferential trading zone. This effect is known as trade diversion (Viner, 1950). PTAs also create trade if inefficient domestic production is substituted by more efficient production from within the preferential trading area.

Despite their ambiguous welfare effects, PTAs have been mushrooming since around 1990. In the period from 1945 to 1990, countries around the world signed 208 PTAs (Dür et al., 2014). After 1991, this number substantially increased, with countries signing 525 new agreements between 1991 and 2009. Many of these new agreements not only reduce tariffs on goods, but also contain provisions regarding the liberalization of services trade, the protection of foreign direct investments, the harmonization of standards, and many other behind-the-border trade barriers.

Several explanations exist for why countries pursue preferential rather than multilateral trade agreements. For one, countries may find it easier to negotiate far-reaching agreements between two or among a few countries (Krugman, 1993). As of 2014, the WTO has 159 members. Agreeing on a multilateral agreement among so many countries has proven difficult. The Doha Development Agenda has been ongoing since 2001, with only limited results. In this view, PTAs simply satisfy countries' needs for negotiated trade liberalization in the absence of a multilateral agreement.

PTAs may not only be a second-best policy, however. Producers of goods that benefit from economies of scale may favor PTA because they allow them to increase their production and hence reduce unit costs without facing world-wide competition. A PTA then can serve as a springboard for these producers to achieve global competitiveness. Realizing this, these producers may lobby governments to sign preferential rather than or in advance of multilateral trade agreements (Chase, 2005).

Countries may also pursue PTAs for geopolitical reasons. A PTA may allow a dominant country to increase other countries' dependence on the own market, which in turn should increase the dominant country's power over these other countries (Hirschman, 1945). Similarly, a PTA may be an instrument to increase the military power of a set of countries. PTAs increase trade among the member countries; this trade creation enhances the national income of the participating countries, allowing them to invest more in defense (Gowa, 1994).

A PTA may also be a response to the PTAs signed by other countries. By liberalizing trade only among a few countries, PTAs discriminate against third countries. Exporters from these third countries suddenly find themselves in a situation in which they are no longer competitive in the PTA market, because they face higher trade barriers than firms inside the PTA. Third country exporters may then lobby their governments to also sign a PTA with the members of a PTA or to form a competitive PTA, leading to a domino effect that makes more and more countries sign PTAs (Baldwin, 1993; Dür, 2010; Baccini and Dür, 2012).

PTAs may also be a signal to voters in democracies that the government is pursuing efficient economic policies (Mansfield and Milner, 2012). The argument here is that voters often do not know whether an economic downturn is a result of the economic policies pursued by the government or exogenous circumstances. In doubt, they use elections to oust the government. Knowing this, the government can sign PTAs that signal to voters that it is pursuing an efficient economic policy, allowing it to stay in power even during economic downturns.

Finally, especially for governments in developing countries, PTAs may also serve to lock in domestic economic reforms (Whalley, 1998). Economic reforms can be painful; once implemented, they may not survive a change in government. By linking these reforms to better foreign market access in PTAs, governments can ensure the durability of these reforms.

How Do Countries Negotiate over Trade?

Early trade negotiations in the GATT were dominated by a few countries. Developing countries mainly were passive recipients of the concessions exchanged by the key trading nations. It was not before the Uruguay Round that developing countries became increasingly vocal. The increasing number of active players in multilateral negotiations has made the analysis of these negotiations substantially more demanding. Both multilateral and preferential trade negotiations have also become increasingly complex because of the growing number of topics covered in trade negotiations. Whereas originally the

negotiations mainly tackled tariff cuts, contemporary trade negotiations have a scope that goes far beyond tariffs and encompasses behind-the-border barriers to trade.

One important consequence of the increasing complexity of multilateral trade negotiations has been the emergence of coalitions that explicitly coordinate their position. In the Uruguay Round, for example, several coalitions of developing countries played an important role (Drahos, 2003; Narlikar, 2003). Bargaining together allowed them partly to balance the superior power of the developed countries. As preferential trade negotiations mainly do not allow for such coalition building, weaker countries may be more open to exploitation in these negotiations than in the GATT/WTO system.

An analysis of international trade negotiations thus always begs the question of who has power in these negotiations. There are several sources of power in trade negotiations (Drahos, 2003; Dür, 2010). A country's market size is key, as countries are more eager to get access to larger than to smaller markets. Access to the own market, then, can be 'sold' in exchange for concessions from the other countries. A country's domestic political institutions may also matter for the country's bargaining power in international trade negotiations. The literature on international negotiations shows that countries with high ratification requirements tend to be more powerful in international negotiations (Putnam, 1988). Finally, negotiation strategies may matter for a country's bargaining power. Choosing the right tactics at the right moment (for example, building a coalition) can give a country an advantage in the negotiations.

Existing research on the power of countries in international trade negotiations indeed provides evidence for power imbalances. Gruber (2000) showed that countries that are able to conclude trade negotiations without waiting for more protectionist countries possess what he calls 'go-it-alone-power.' This go-it-alone power allows them to impose certain trade policy choices on excluded countries. Kim (2010) argues that developed countries used their power to shape the multilateral trade institutions following their preferences. By doing so, they managed to affect trade flows in a way that benefit them to the detriment of developing countries. Dür (2010) argues that countries that discriminate against other countries manage to gain power in international trade negotiations. The creation of the European Economic Community, for example, enabled the European countries to extract concessions from the United States.

Some research has also looked at the strategies that countries use in international trade negotiations. Strategies are sets of tactics that countries adopt in pursuit of their objectives (Odell, 2006: 15). They vary between distributive and integrative bargaining (Walton and McKersie, 1965) or between hard and soft bargaining (Dür and Mateo, 2010). Distributive bargaining refers to a strategy that is aimed at claiming value from others. Integrative bargaining, by contrast, aims at creating value, that is, increasing the size of the pie that can be distributed in a negotiation. Hard bargaining refers to the use of conflictual or aggressive tactics; and soft bargaining to the use of cooperative or friendly ones.

A few studies report empirical evidence on the choice of strategies in international trade negotiations. Davis (2003) analyzed how countries use issue linkages to reach compromises in agricultural trade negotiations. Elms (2006) studied US strategies in negotiations with Japan and South Korea in the 1990s. She shows that countries adopt a distributive or value-claiming strategy when they face losses in trade negotiations. They do so even though a distributive strategy is likely to result in a suboptimal outcome.

Outlook

The increasing scope of international trade negotiations has made them increasingly salient to the public. The Seattle WTO Ministerial Meeting of 1999, which was supposed to start the multilateral trade negotiations later known as Doha Development Agenda, witnessed violent protests (Smith, 2001). The Anti-Counterfeiting Trade Agreement signed in 2011 and 2012 was not ratified by the European Union because of a massive mobilization of citizens opposing the transnational protection of intellectual property rights (Dür and Mateo, 2014). The negotiations for a Transatlantic Trade and Investment Partnership have also experienced a bumpy start after the plans for investor-state dispute settlement – namely the ability for private investors to sue states before an international tribunal – were publicly debated. While earlier trade negotiations were nearly exclusively decided by a few negotiators, this greater public salience will have a major effect on current and future trade negotiations. It will influence the speed of liberalization; countries' choice of multilateral or bilateral negotiations; and power and strategies in negotiations.

See also: Area and International Studies: Political Economy; Globalization: Geographical Aspects; International Trade: Economic Integration; Trade Policy.

Bibliography

Alt, James E., Gilligan, Michael, 1994. Survey article: the political economy of trading states: factor specificity, collective action problems and domestic political institutions. The Journal of Political Philosophy 2 (2), 165–192.

Alt, James E., Frieden, Jeffry, Gilligan, Michael J., Rodrik, Dani, Rogowski, Ronald, 1996. The political economy of international trade: enduring puzzles and an agenda for inquiry. Comparative Political Studies 29 (6), 689–717.

Baccini, Leonardo, Dür, Andreas, 2012. The new regionalism and policy interdependence. British Journal of Political Science 42 (1), 57–79.

Bailey, Michael A., Goldstein, Judith, Weingast, Barry R., 1997. The institutional roots of American trade policy: politics, coalitions, and international trade. World Politics 49 (3), 309–338.

Baldwin, Richard E., 1993. A Domino Theory of Regionalism. NBER Working Paper No. 4465.

Bhagwati, Jagdish N., 1988. Protectionism. MIT Press, Cambridge.

Brown, Andrew G., 2003. Reluctant Partners: A History of Multilateral Trade Cooperation, 1850–2000. University of Michigan Press, Ann Arbor.

Chase, Kerry A., 2005. Trading Blocs: States, Firms, and Regions in the World Economy. University of Michigan Press, Ann Arbor.

Cohn, Theodore H., 2002. Governing Global Trade: International Institutions in Conflict and Convergence. Ashgate, Aldershot.

Davis, Christina L., 2003. Food Fights over Free Trade: How International Institutions Promote Agricultural Trade Liberalization. Princeton University Press, Princeton.

Drahos, Peter, 2003. When the weak bargain with the strong: negotiations in the World Trade Organization. International Negotiation 8 (1), 79–109.

Dür, Andreas, 2010. Protection for Exporters: Power and Discrimination in Transatlantic Trade Relations, 1930–2010. Cornell University Press, Ithaca.

Dür, Andreas, Baccini, Leonardo, Elsig, Manfred, 2014. The design of international trade agreements: introducing a new database. Review of International Organizations. http://dx.doi.org/10.1007/s11558-013-9179-8.

Dür, Andreas, Mateo, Gemma, 2010. Bargaining power and negotiation tactics: the negotiations on the EU's financial perspective, 2007–2013. Journal of Common Market Studies 48 (3), 557–578.
Dür, Andreas, Mateo, Gemma, 2014. Public opinion and interest group influence: how citizen groups derailed the anti-counterfeiting trade agreement. Journal of European Public Policy. http://dx.doi.org/10.1080/13501763.2014.900893.
Edwards, Sebastian, Lederman, Daniel, 1998. The political economy of unilateral trade liberalization: the case of Chile. National Bureau of Economic Research. Working Paper No. 6510.
Elms, Deborah, 2006. How bargaining alters outcomes: bilateral trade negotiations and bargaining strategies. International Negotiation 11 (3), 399–429.
Elms, Deborah K., Low, Patrick (Eds.), 2013. Global Value Chains in a Changing World. World Trade Organization, Geneva.
Gilligan, Michael J., 1997. Empowering Exporters: Reciprocity, Delegation, and Collective Action in American Trade Policy. The University of Michigan Press, Ann Arbor.
Goldstein, Judith, 1993. Ideas, Interests, and American Trade Policy. Cornell University Press, Ithaca.
Gowa, Joanne, 1994. Allies, Adversaries, International Trade. Princeton University Press, Princeton.
Gruber, Lloyd, 2000. Ruling the World: Power Politics and the Rise of Supranational Institutions. Princeton University Press, Princeton.
Hirschman, Albert O., 1945. National Power and the Structure of Foreign Trade. University of California Press, Berkeley.
Hiscox, Michael J., 1999. The magic bullet? the RTAA, institutional reform, and trade liberalization. International Organization 53 (4), 669–698.
Hiscox, Michael J., 2002. International Trade and Political Conflict: Commerce, Coalitions, and Mobility. Princeton University Press, Princeton.
Horn, Henrik, Mavroidis, Petros C., 2001. Economic and legal aspects of the most-favored-nation clause. European Journal of Political Economy 17 (2), 233–279.
Jackson, John H. (Ed.), 1997. The World Trading System: Law and Policy of International Economic Relations, second ed. Cambridge University Press, Cambridge.
Kim, Soo Y., 2010. Power and the Governance of Global Trade: From the GATT to the WTO. Cornell University Press, Ithaca.
Krugman, Paul R., 1993. Regionalism versus multilateralism: analytical notes. In: De Melo, Jaime, Panagariya, Arvind (Eds.), New Dimensions in Regional Integration. Cambridge University Press, New York, pp. 58–78.
Lohmann, Susanne, O'Halloran, Sharyn, 1994. Divided government and US trade policy: theory and evidence. International Organization 48 (4), 595–632.
Magee, Stephen P., Brock, William A., Young, Leslie, 1989. Black Hole Tariffs and Endogenous Policy Theory: Political Economy in General Equilibrium. Cambridge University Press, Cambridge.
Mansfield, Edward D., Milner, Helen V., 1999. The new wave of regionalism. International Organization 53 (3), 589–627.
Mansfield, Edward D., Milner, Helen V., 2012. Votes, Vetoes, and the Political Economy of International Trade Agreements. Princeton University Press, Princeton.
Narlikar, Amrita, 2003. International Trade and Developing Countries: Bargaining and Coalitions in the GATT and WTO. Routledge, London.
Odell, John S. (Ed.), 2006. Introduction. In: Negotiating Trade: Developing Countries in the WTO and NAFTA. Cambridge University Press, Cambridge, p. 1.
Olson, Mancur, 1965. The Logic of Collective Action. Harvard University Press, Cambridge.
Pahre, Robert, 2008. Politics and Trade Cooperation in the Nineteenth Century: The "Agreeable Customs" of 1815–1914. Cambridge University Press, Cambridge.
Putnam, Robert D., 1988. Diplomacy and domestic politics: the logic of two-level games. International Organization 42 (3), 427–460.
Rogowski, Ronald, 1989. Commerce and Coalitions: How Trade Affects Domestic Political Alignments. Princeton University Press, Princeton.
Schattschneider, E.E., 1935. Politics, Pressures, and the Tariff: A Study of Free Private Enterprise in Pressure Politics, as Shown in the 1929–1930 Revision of the Tariff. Prentice Hall, New York.
Schonhardt-Bailey, Cheryl, 2006. From the Corn Laws to Free Trade: Interests, Ideas and Institutions in Historical Perspective. MIT Press, Cambridge.
Smith, Jackie, 2001. Globalizing resistance: the battle of seattle and the future of social movements. Mobilization: An International Quarterly 6 (1), 1–19.
Viner, Jacob, 1950. The Customs Union Issue. Carnegie Endowment for International Peace, New York.
Walton, Richard E., McKersie, Robert B., 1965. A Behavioral Theory of Labor Negotiations: An Analysis of a Social Interaction System. ILR Press, Ithaca.
Whalley, John, 1998. Why do countries seek regional trade agreements? In: Frankel, Jeffrey A. (Ed.), The Regionalization of the World Economy. University of Chicago Press, Chicago, pp. 63–89.
World Trade Organization, 2013. World Trade Report 2013: Factors Shaping the Future of World Trade. WTO, Geneva.

Relevant Websites

www.designoftradeagreements.org – Design of Trade Agreements (DESTA) Database
http://ec.europa.eu/trade/ – European Commission, Trade.
http://www.sice.oas.org/ – SICE. The Organization of American States Foreign Trade Information System.
http://www.un.org/en/development/progareas/trade.shtml – United Nations, International Trade.

International Trade: Economic Integration

Anthony Venables, London School of Economics, London, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7843–7848, © 2001, Elsevier Ltd.

Abstract

Regional economic integration occurs when countries come together to form free trade areas or customs unions, offering members preferential trade access to each others' markets. The article reviews the economic effects of such agreements on member countries and on the world trading system. Effects on member countries include the benefits and costs of trade creation and trade diversion, as well as gains from increased scale and competition. 'Deeper' integration can be pursued by going beyond abolition of import tariffs and quotas, to further measures to remove market segmentation and promote integration. Effects on the world trading system are not clear-cut. There is little evidence that regionalism has retarded multilateral liberalization, but neither is there support for the view that continuing expansion of regional agreements will obviate the need for multilateral liberalization efforts.

Regional integration agreements (RIAs) are groupings of countries formed with the objective of reducing barriers to trade between members. They have long historical antecedents, sometimes being a stage to nation building (most famously in the Zollverein of nineteenth-century Germany), and sometimes as part of colonial trading arrangements. In the postwar period, developments were led by the European Union (EU, originally the European Economic Community (EEC), founded in 1958), and in the 1960s and 1970s there was a number of rather inward looking (and largely unsuccessful) RIAs between developing countries. Since the mid-1980s there has been a dramatic increase in regional integration activity. Of the 194 RIAs notified to the General Agreement on Trade and Tariffs/World Trade Organization (GATT/WTO) at the beginning of 1999, 87 were notifications since 1990. Now almost all countries are members of at least one RIA, and more than one third of world trade takes place within such agreements. The new developments include the expansion and deepening of the EU; the construction of new and more open RIAs between developing countries; and the advent of RIAs in which both high-income and developing countries are equal partners, lead by the North American Free Trade Area (NAFTA) which, in 1994, extended the Canadian–USA free trade agreement to Mexico.

RIAs come in many shapes and sizes. They vary in income levels, in openness to trade, and in the share of trade that takes place within the RIA, (60% for the 15 members of the EU, but just 10% for the eight members of the West African Economic and Monetary Union). Structures vary, from the loose agreements to facilitate trade of the African Cross-Border Initiative and Asia Pacific Economic Cooperation (APEC) forum, through to the deep integration of the EU, involving the construction of shared executive, judicial, and legislative institutions. The focus of this article, in common with most of the economics literature, is on RIAs as mechanisms for trade liberalization, not addressing the integration of factor markets, monetary systems, or political institutions. Even with this focus, RIAs differ widely in structure. 'Free trade areas' remove internal tariff and non-tariff barriers, but permit members to retain independent external tariff policies. 'Customs unions' go further, in fully harmonizing external trade policy. And it is increasingly recognized that effective integration of markets requires more than reducing tariffs and quotas. Many other barriers – such as differing national product standards and a host of minor border frictions – impede trade and support market segmentation, and some RIAs are now pursuing 'deep integration' policies to eliminate these barriers.

This article is organized under four main headings. Section Economic Effects on Member Countries outlines the main effects of RIAs on member countries, Section Policy Making discusses some of the wider policy choices faced in RIA formation, and Section Regional Integration Agreements and the World Trading System looks at the implications of RIAs for the world trading system as a whole. Section Conclusions concludes.

Economic Effects on Member Countries

Trade Creation and Diversion

The modern analysis of RIAs dates from Viner (1950), who drew the distinction between the trade creating and trade diverting effects of RIA formation. The classical source of gains from trade is that global free trade allows consumers and firms to purchase from the cheapest source of supply, hence ensuring that production is located according to comparative advantage. In contrast, trade barriers discriminate against foreign supply, inducing domestic import-competing producers to expand even though they have higher costs than imports. This in turn starves domestic export sectors of resources and causes them to be smaller than they otherwise would be. Since a RIA liberalizes trade, reducing at least some of the barriers, doesn't it follow that it too will generate gains from trade? Viner's contribution was to show that the answer is; not necessarily. The gains from trade argument applies if all trade barriers are reduced, but need not apply to a partial – and discriminatory – reduction in barriers, as in a RIA. This is because discrimination between sources of supply is not eliminated, it is just shifted. If partner country production displaces higher-cost domestic production then there will be gains – trade creation. But it is possible that partner country production may displace lower cost imports from the rest of the world, and this is welfare reducing trade diversion.

The analysis of trade creation and trade diversion constitutes one of the first formal analyses of the more general

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71039-7

problem of 'second-best welfare economics.' Given that distortions remain in place in some activities in the economy, it is not necessarily the case that removing just some of the distortions (e.g., eliminating trade barriers on partner countries and leaving them in place on external countries) is welfare improving. In the literature on regional integration the response to the fundamental ambiguity created by the second best took three main forms.

First, authors established circumstances under which there is no interaction between formation of the RIA and external trade flows, so no possibility of trade diversion. Meade (1955) pointed out that if trade barriers with nonmembers take the form of fixed quantitative restrictions, then a RIA must raise the total welfare of member countries since there is no possibility that imports from the rest of the world are displaced. Ohyama (1972) and Kemp and Wan (1976) showed how, when external trade barriers take the form of tariffs, it is possible to adjust these to hold external trade volumes constant, so preventing trade diversion from occurring.

Second, researchers identified conditions, in terms of changes in endogenous variables, for welfare gain. For example, welfare increases if the initial-tariff weighted change in trade volume is positive (Meade, 1955). If internal tariffs are close to zero, then reducing them to zero raises welfare if it increases tariff revenues earned on external trade (Ethier and Horn, 1984).

The third approach is to identify features of economies (in terms of their underlying exogenous characteristics) under which they are more or less likely to gain or lose from RIA membership. Lipsey (1957) argued that joining with countries that are already one's largest trading partners is unlikely to lead to diversion, since the fact that the countries were originally the largest trading partners suggests that they are the lowest cost source of supply. Similar reasoning, including transport costs in the costs of supply, leads to the 'natural trading bloc' argument (Wonnacott and Lutz, 1989; Summers, 1991). Venables (2000) shows that those members of an RIA with comparative advantage most different from the world average are most likely to lose from trade diversion, as their trade is diverted to partner countries with comparative costs between theirs and the world average.

Empirical work on trade creation and trade diversion has taken two main forms; econometric studies of changes in trade flows, and computer simulation studies of the full general equilibrium effects of RIA membership.

Econometric studies seek to quantify the changes in trade flows attributable to membership of a RIA, and thereby identify trade creation and diversion. A variety of different econometric models have been developed, the most common being based on the gravity model which estimates bilateral trade between countries as a function of their GDPs, populations, the distance between them, and physical factors such as sharing a land border, and being landlocked or an island. Dummy variables capture whether or not countries are in a particular RIA, their estimated effect indicating whether countries in a RIA trade more or less than would otherwise be expected. Using this technique, Bayoumi and Eichengreen (1997) find that the formation of the EEC reduced the annual growth of member trade with other industrial countries by 1.7% points, with the major attenuation occurring over 1959–61, just as trade preferences were phased in. Soloaga and Winters (1999) look at a wide range of RIAs, producing a mixed picture with little evidence of widespread trade diversion. Overall, there appears to be weak evidence that external trade is smaller than it otherwise might have been in at least some of the blocs that have been researched, but the picture is sufficiently mixed that it is not possible to conclude that trade diversion has been a major problem. Furthermore, it cannot be inferred that trade diversion has been economically damaging without information on relative costs and tariff structures, variables that are not revealed in this sort of aggregate exercise.

The second empirical approach is based on computable equilibrium modeling. This involves construction of a full computer model of the economies under study and use of the model to simulate the effects of the policy changes associated with the RIA. Such a model typically contains a great deal of microeconomic detail, so can be used to predict changes in production in each sector, and changes in factor prices and real incomes. In models that assume a perfectly competitive environment, the combined effects of trade diversion and trade creation typically give very small welfare gains – just a fraction of 1% of GDP (see Baldwin and Venables, 1997 for a survey). The strength of these models is that they have sufficient microeconomic structure for the effects of a policy change to be traced out in detail, and its real income effects to be calculated. They are also often used for prediction – to estimate the likely effects of a policy change before it is implemented. But they have the major weakness that they are not usually fitted to data as carefully, or subject to the same statistical testing, as econometric models. The cost of the microeconomic detail is a complexity that makes rigorous econometric estimation impossible.

Although the focus of the trade creation and diversion literature has been on the changes in trade flows induced by regional integration, two consequent effects are important. First, changes in trade flows may change world prices, possibly improving the terms of trade of member countries, although this gain arises at the expense of outside countries. For example, if trade diversion occurs then RIA imports from outside countries are reduced, and any reduction in import prices that this causes is a terms of trade gain. Empirical work on this issue by Winters and Chang (2000) shows that Brazil's membership in Mercosur has been accompanied by a significant decline in the relative prices of imports from nonmember countries.

Second, changes in tariffs and trade volumes will lead to loss of government tariff revenue. This can occur directly (as intra-RIA tariffs are cut) and as a consequence of trade diversion (as imports are diverted away from external, tariff inclusive, sources of supply). Its cost depends on the social cost of raising funds by alternative means, and can be severe in some developing countries. For example, in the South African Development Community, where some countries are quite heavily dependent on trade with South Africa, substantial amounts of revenue are involved, amounting to perhaps 5.6 and 9.8% of government revenue for Zambia and Zimbabwe respectively. Cambodia derived 56% of its total tax revenues from customs duties prior to its entry into the Association of South East Asian Nations (ASEAN), and Fukase and Martin (1999) argue that

entry into ASEAN provided a powerful stimulus for the introduction of a value-added tax.

Scale and Competition Effects

A second mechanism through which member countries are affected by RIA membership derives from the fact that countries may be too small to support, separately, activities that are subject to large economies of scale. Regional cooperation offers a route to overcome the disadvantages of smallness, by pooling resources or combining markets. These scale benefits can arise in public projects (see World Bank, 2000) and also at the level of the private firm, where they typically interact with imperfectly competitive market structures. These considerations are absent from the trade creation and trade diversion approach outlined above, which is based on the perfect competition and constant returns to scale paradigm of traditional trade theory. It was only in the 1970s and 1980s that formal analysis of the interaction between trade, economies of scale, and imperfect competition began with the 'new trade theory,' and these techniques have now been extensively applied to regional integration.

The basic argument is that there is a trade-off between the extent to which firms can achieve economies of scale, and the intensity of competition in the market. For a given size market, larger firms means fewer firms and hence more monopolistic outcomes. If regional integration combines markets, then it shifts this trade-off, potentially allowing firms to be bigger and markets to be more competitive (Smith and Venables, 1988). For example, there might be an initial situation in which two economies each have two firms in a particular industry, and these firms exploit their 'duopoly' power, setting prices well above marginal cost. After formation of the RIA this becomes four firms in one combined RIA market. This increases the intensity of competition, and possibly induces merger (or bankruptcy), perhaps leaving only the three most efficient firms. The net effect is increased competition, increased firm scale, and lower costs. 'Triopoly' competition is likely to be more intense than the original duopolies, and surviving firms are larger and more efficient, so can better exploit economies of scale. A further source of gains comes from possible reductions in internal inefficiencies that firms are induced to make. If the RIA increases the intensity of competition, it may induce firms to eliminate internal inefficiencies (X-inefficiency) and raise productivity. Since competition raises the probability of bankruptcy and hence layoffs, it also generates stronger incentives for workers to improve productivity, and increases labor turnover across firms within sectors. Although these are significant potential sources of gain, they have been difficult to achieve in many RIAs. This is addressed further in Section The Depth of Integration.

Winners and Losers

A continuing concern is with the distribution of the costs and benefits of regional integration between member countries. Do central regions gain at the expense of peripheral ones, and do poor countries tend to catch up or get left behind? The evidence is, broadly, that RIAs composed of developed countries tend to show convergence (for example, the narrowing of per capita income differentials observed in the EU, see Ben-David, 1993). However, the picture for RIAs composed of developing countries is more mixed, with some examples of divergent performance (World Bank, 2000).

The analytical literature on these questions is quite sparse, but provides several clues why this might occur. First, as mentioned above, trade diversion is more likely for countries with 'extreme' comparative advantage, suggesting that in a RIA among developing countries it might be the lowest income countries that experience diversion. For example, their imports of manufactures might be diverted from nonmember countries to a partner that has a comparative advantage in manufactures within the RIA, but not relative to the world at large. Second, industries might tend to cluster in locations that have relatively good market access, or that are well supplied with business services or provision of other intermediate goods. This is more likely to occur in developing countries than in developed ones, partly because of their sparser provision of business infrastructure, and partly because the small size of their manufacturing sectors means that clustering is less likely to run into congestion and other sources of diminishing returns. The clustering may lead to wages being bid up in one member country at the expense of others.

Policy Making

The Depth of Integration

The simplest form of regional integration is the elimination of tariffs (or quotas) between member countries. Beyond this there is a wide range of policy options open to countries considering integration, many of which turn on the 'depth' of integration sought by member countries – ranging from modest trade liberalization, through full economic integration, to the formation of shared institutions.

The distinction between a free trade area and a customs union, in which a common external tariff is set, has already been made. The latter involves greater sharing of sovereignty and requires establishing procedures for revenue sharing, but in return can yield much greater market integration. In a free trade area where countries set different external tariffs the free internal circulation of goods is impossible; border formalities have to be maintained to ensure that external imports do not all enter through the member with the lowest external tariff, for re-export to other member countries. Since these imports include intermediate goods that are further processed in member countries, in practice this involves enforcing complicated 'rules of origin' governing trade flows within the RIA. (The EU's agreement with Poland has 81 pages of small print in its rules of origin section, and NAFTA some 200, Krueger, 1997.) These rules are not necessary in a customs union, enabling simplification – or elimination – of internal border formalities.

It is increasingly recognized that tariffs and quotas alone may be just a small part of the overall barriers to trade created by an international border. Rules of origin create frictions, and so too does contingent protection (such as antidumping rules, known to have a 'trade chilling' effect even when not used), duplicative customs procedures, differing national product standards, and simple border red tape. The cost of border formalities on intra-EU trade in the early 1990s has been

estimated to be more than 1% of the gross value of internal trade, despite implementation of procedures to cut these costs; in many RIAs the costs of border formalities are much larger.

Furthermore, relatively minor border frictions may constitute large real trade barriers because it can be in the interests of firms to try and limit effective cross-border competition in order to maintain collusive market-sharing arrangements. This will prevent the 'scale and competition' effects discussed in Section Scale and Competition Effects from being achieved.

Recognition of the importance of these barriers – and of the failure of more than 25 years tariff-free trade to truly integrate markets – was a motivation behind the 'deep integration' of the EU's 1992 Single Market Program (Flam, 1992). The program involved adoption of almost 300 measures falling into the following five main types.

1. Simplification and in some cases abolition of intra-EC border controls, involving, e.g., replacing border paperwork by an EU-wide system of administering value added tax on cross-border transactions.
2. Adoption of the 'mutual recognition' principle for product standards, under which a product that can be legally sold in any EU country can be legally sold in all, thereby removing the need for expensive re-testing and re-certification of products.
3. Progress towards deregulation of the transport sectors of EU countries, including measures to reduce restrictions on truckers from one country accepting loads in another.
4. Opening of public procurement in EU countries to effective competition from suppliers in all EU countries.
5. Deregulation of service sector activities, including opening financial services to competition and giving service providers and professionals the right of establishment in other EU countries.

Estimates of the gains from these measures range up to as much as 5% of EU GPD, although these are based largely on computable equilibrium studies rather than an actual survey of outcomes (see Baldwin and Venables, 1997).

Policy Reform and Commitment

Trade policy reforms – and other sorts of reform – are often hampered by the expectation that they may be reversed. Adjusting to reform typically involves investments, but these investments will not be made unless investors are confident that the reform will persist. These problems are mitigated if a country has a 'commitment mechanism' guaranteeing that the reform will be durable, and membership of a RIA can, in some circumstances, provide such a mechanism (Fernandez and Portes, 1998).

The commitment mechanism operates most obviously for trade policy – membership requires that tariffs with member countries be cut, and reneging on agreed internal liberalization is likely to bring swift retaliation by partner countries. However, it has been argued that RIAs are valuable as commitment mechanisms for a much wider range of measures. Although NAFTA was ostensibly about trade policy, an important part of its motivation was the desire on the part of both the Mexican and US governments to lock in the broad range of economic reforms that the Mexican government had undertaken in the preceding years. The EU Articles of Agreement with eastern European accession candidates are explicit in promoting 'full integration into the community of democratic nations.' And the intervention of other Mercosur countries is credited with having averted a military coup in Paraguay in 1996 (Survey on Mercosur, *The Economist*, 12 October 1996). Paradoxically, it is even suggested that the value of a RIA as a commitment mechanism is greatest in areas other than trade policy, because there is already a way committing to tariff reductions – the tariff bindings of the GATT/WTO.

Regional Integration Agreements and the World Trading System

The effect of regionalism on the world trading system as a whole has been the subject of extensive recent debate. Is the growth of regionalism part of a process towards global free trade, or is it a substitute for it, damaging to multilateral negotiations and likely to lead to a situation of protected trading blocs? It is noteworthy that RIAs go against the fundamental principle of the GATT/WTO, the 'most favored nation' principle under which trade policy must be non-discriminatory. They are permitted only via Article XXIV of the GATT, which requires that they cover 'substantially all trade' of member countries and lead to external tariffs that 'shall not on the whole be higher or more restrictive.'

There are essentially three issues. First, does a given structure of RIAs create a force for more or less liberal external trade policy? Second, is there an inherent dynamic of RIA formation which will lead to ever larger RIAs – perhaps even a RIA of the whole world, giving global free trade? And third, is there evidence that the presence of RIAs has assisted or retarded progress in the multilateral trade system?

An insight into the effect of the number of RIAs on incentives to set external tariffs was provided by Krugman (1993) who noted that if tariffs are set to improve member countries' terms of trade, then they will be lowest – and consequently world income greatest – in two opposite circumstances. One is when there is a single world trade bloc containing all countries, i.e., global free trade. The other is when trade policy is set by many small independent jurisdictions, each so small as to have no market power and no reason to deviate from free trade. Between these extremes each trading bloc has an incentive to use external tariffs to try and improve its terms of trade, suggesting that a situation of relatively few large trading blocs might be the worst possible outcome.

This line of argument has been criticized from a number of different angles. From the theory side, it should be recognized that tariff setting is a repeated game, so the incentives to cooperate, or to deviate from cooperation, need to be taken into account (Bond and Syropoulos, 1996; Winters, 1998). From the institutional side, it has been pointed out that article XXIV of the GATT expressly forbids RIAs from raising external tariffs. And from the empirical side, no evidence has been found to suggest that RIAs are in fact prone to set higher external tariffs than are separate countries (Foroutan, 1998).

What of the dynamics of RIA formation? Does adding a member country to a RIA increase the incentives for further countries to join, and will existing members be willing to permit

unrestricted entry? A number of researchers have argued that the incentives to join increase as RIAs become larger – a phenomenon termed 'domino regionalism' by Baldwin (1995). This may be because of perception of growing benefits of membership, or because of increasing costs of being outside. These costs can arise as outsiders' terms of trade may decline (the other side of the members improving terms of trade referred to in Section Trade Creation and Diversion). Perhaps more importantly, countries fear that firms may relocate, in search of the benefits of a larger market (evidence of foreign direct investment flows in Europe at the time of the Single Market Program suggests that this occurred, Baldwin et al., 1996). Another source of loss from nonmembership of RIAs is the risk of being isolated if a trade war occurs (Whalley, 1998).

Although these arguments suggest a growing demand for membership of RIAs, it is less clear that this will be matched by a willingness to accommodate new members, at least within existing RIAs. The most important tradeoff here is between the benefits of 'deeper' integration, and the difficulty of achieving this with larger memberships. Thus, while the EU has continued to enlarge, worries about the difficulty of accommodating new entrants impedes progress. An important contribution to the ideas on regionalism was made when the Asia Pacific Economic Cooperation (APEC) forum announced the principle of 'open regionalism.' Although this term has been given a number of quite different meanings, the key idea is that of open access, whereby the RIA announces that any country willing to abide by its rules may join. However, at time of writing, the idea remains to be put to the test, since APEC itself has yet to develop as a RIA. Overall, then, it seems that regional integration does create its own dynamic, attracting further countries to want to join. But there is no reason to believe that all such requests should be successful, or that the end of the process is more likely to be a single world free trade area than it is to be a number of competing trade blocs.

Finally, are there any reasons to believe that the presence of RIAs has facilitated or impeded progress in multilateral trade negotiations? The argument for impedance hinges mainly on finite government capacity. Investment of time and effort in regionalism reduces the capacity of governments to invest in multilateral negotiation. On the other side, it is argued that regionalism can help, by reducing the number of separate negotiators (since members of a customs union negotiate as a single body) – an argument that begs the question of how members formulate their common position.

Also, it is suggested that regionalism has proved valuable by providing 'laboratories' for trying new approaches to issues of trade reform, particularly on more difficult issues. As for the track record of RIAs and multilateral negotiations, views are again divided, although there is a view that fear of fragmentation into regional blocs provided a spur for successful completion of the Uruguay round of trade negotiations (WTO, 1995).

Conclusions

Joining a RIA is a major 'one-off' event for a country. It affects all aspects of the economy – goods prices, the structure of production, and income flows to different economic agents. Generalizing about the implications of such occasional and large changes is inevitably difficult. Nevertheless, research has established the circumstances under which gains or losses are more or less likely, and drawn out the implications of this for the design of agreements (World Bank, 2000). Future research needs to pay greater attention to strategies for developing countries, and the role of RIAs in assisting them to participate effectively in the world trading system.

See also: Commodity Chains; Economic Geography; Globalization: Geographical Aspects; Globalization: Legal Aspects; International Business; International Trade: Commercial Policy and Trade Negotiations.

Bibliography

Baldwin, R.E., 1995. A domino theory of regionalism. In: Baldwin, R.E., Haaparanta, P., Kiander, J. (Eds.), Expanding Membership in the European Union. Cambridge University Press, Cambridge, UK.

Baldwin, R., Forslid, R., Haaland, J.I., 1996. Investment creation and diversion in Europe. The World Economy 19, 635–659.

Baldwin, R.E., Venables, A.J., 1997. International economic integration. In: Grossman, G.M., Rogoff, K. (Eds.), Handbook of International Economics, vol. 3. North-Holland, Amsterdam.

Bayoumi, T., Eichengreen, B., 1997. Is regionalism simply a diversion: Evidence from the evolution of the EC and EFTA. In: Ito, T., Krueger, A. (Eds.), Regionalism versus Multilateral Trade Arrangements. University of Chicago Press, Chicago, pp. 141–168.

Ben-David, D., 1993. Equalizing exchange: Trade liberalization and income convergence. The Quarterly Journal of Economics 108, 653–679.

Bhagwati, J., 1993. Regionalism and multilateralism: An overview. In: De Melo, J., Panagariya, A. (Eds.), New Dimensions in Regional Integration. CEPR, Cambridge University Press, London, pp. 22–51.

Bond, E.W., Syropoulos, C., 1996. The size of trading blocs: Market power and world welfare effects. Journal of International Economics 40, 411–437.

Ethier, W., Horn, H., 1984. A new look at economic integration. In: Kierszkowski, H. (Ed.), Monopolistic Competition and International Trade. Oxford University Press, Oxford, UK, pp. 204–229.

Fernandez, R., Portes, J., 1998. Returns to regionalism: An analysis of nontraditional gains from regional trade agreements. World Bank Economic Review 12, 197–220.

Flam, H., 1992. Product markets and 1992: full integration, large gains? Journal of Economic Perspectives 6, 7–30.

Foroutan, F., 1998. Does membership in a regional preferential trade arrangement make a country more or less protectionist? The World Economy 21, 305–335.

Frankel, J.A., 1997. Regional Trading Blocs in the World Economic System. Institute for International Economics, Washington, DC.

Fukase, E., Martin, W., 1999. Evaluating the Implications of Cambodia's Accession to the ASEAN Free Trade Area: A General Equilibrium Model (CGE) Approach. World Bank, Washington, DC.

Kemp, M., Wan, H.Y., 1976. An elementary proposition concerning the formation of customs unions. Journal of International Economics 6, 95–97.

Krueger, A.O., 1997. Free trade agreements versus customs unions. Journal of Development Economics 54, 169–187.

Krugman, P., 1993. Regionalism versus multilateralism: Analytical notes. In: De Melo, J., Panagariya, A. (Eds.), New Dimensions in Regional Integration. Cambridge University Press, Cambridge, UK.

Lawrence, R.Z., 1996. Regionalism, Multilateralism and Deeper Integration. Brookings Institution, Washington, DC.

Lipsey, R.G., 1957. The theory of customs unions: Trade diversion and welfare. Economica 24, 40–46.

Meade, J.E., 1955. The Theory of Customs Unions. North-Holland, Amsterdam.

Ohyama, M., 1972. Trade and welfare in general equilibrium. Keio Economic Studies 9, 37–73.

Smith, A., Venables, A.J., 1988. Completing the internal market in the European Community: Some industry simulations. European Economic Review 32, 1501–1525.

Soloaga, I., & Winters, L. A. (1999.). Regionalism in the 1990s; what effect on trade?. London: Centre for Economic Policy Research Discussion paper no.2183.

Summers, L. 1991. Regionalism and the world trading system. In Policy Implications of Trade and Currency Zones. A symposium sponsored by the Federal Reserve Bank of Kansas, KS.

Venables, A. J. 2000. Winners and Losers From Regional Integration Agreements. London: CEPR discussion paper.

Viner, J., 1950. The Customs Union Issue. Carnegie Endowment for International Peace, New York.

Whalley, J., 1998. Why do countries seek regional trade agreements? In: Frankel, J. (Ed.), The Regionalization of the World Economy. University of Chicago Press, Chicago, pp. 63–83.

Winters, L.A., 1998. Regionalism versus multilateralism. In: Baldwin, R., Cohen, D., Sapir, A., Venables, A. (Eds.), Market Integration, Regionalism and the Global Economy. Cambridge University Press, Cambridge, UK.

Winters, L.A., Chang, W., 2000. Regional integration and the prices of imports: An empirical investigation. Journal of International Economics 51, 363–377.

Wonnacott, R.I., Lutz, M., 1989. Is there a case for Free Trade Areas. In: Schott, J.J. (Ed.), Free Trade Areas and US trade Policy. Institute for International Economics, Washington, DC, pp. 59–84.

World Bank, 2000. Trade Blocs. World Bank, Washington, DC.

World Trade Organization, 1995. Regionalism and the World Trading System. WTO Secretariat, Geneva, Switzerland.

Internationalization: Patterns in Business Firms

Ivo Zander, Uppsala University, Uppsala, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews historical and contemporary patterns of how business firms extend their operations outside their country of origin. Addressing the 'why,' 'when,' 'where,' and 'how' of business firms' internationalization, it illustrates how over time traditional and incremental approaches to internationalization have evolved toward increasingly heterogeneous and rapid internationalization processes. Especially the emergence and internationalization of multinational corporations from the developing countries suggest ongoing and potentially far-reaching changes to the international business environment.

Introduction

Patterns are about regular and intelligible forms or sequences discernible in the way in which something happens or is done (http://oxforddictionaries.com). This article will be concerned with such patterns, specifically patterns in how business firms have been observed to internationalize, that is, extend their operations outside their country of origin and into foreign countries.

In a brief introduction to the article, it is useful to also consider the meaning of the rather amorphous concept of internationalization. While a firm's country of origin may be defined as the country in which it was incorporated and in most cases also carried out its first commercial activities, internationalization or the extension of operations into foreign countries covers a range of activities. These activities include the foreign traveling of firm founders and managers, the borrowing of foreign capital, the importing and exporting of products or services, but also the development of more extensive operations such as marketing, manufacturing, and research and development, which then typically involve various degrees of ownership of assets in foreign countries.

In other words, internationalization is a broad concept, which can only be partly covered in this article. The main emphasis here will be on the internationalization of business firms in terms of their efforts to either source, sell, or develop products and services outside their country of origin, and particularly but not exclusively on investments that involve some degree of ownership and control over foreign operations. Whenever the latter has occurred, it will be referred to as foreign direct investments and business firms referred to as multinational corporations (MNCs). Yet, as will be shown presently, ownership and control over foreign resources may not be a core or defining characteristic of internationalization, especially when it comes to the international expansion of contemporary business firms.

As an introductory roadmap, the presentation will focus on a set of questions which include the 'why,' 'when,' 'where,' and 'how' of business firms' internationalization (cf Welsh and Luostarinen, 1988). It will focus primarily on patterns as they have been observed and evolved over time, but also to a more limited extent explore the underlying drivers, consequences, or tangential issues as discussed in the vast international business literature (e.g., Jones, 1996; Dunning and Lundan, 2008; Dunning, 2009). Needless to say, the presentation can cover only a small proportion of the literature on each of the questions, and to every pattern of internationalization that is proposed there are surely contrasting and contradictory examples to be found. The main ambition has been to give an overview of the phenomenon, offering some select references that may serve as entry points into the specialized literatures as they have evolved over the past decades.

'Why'

The answers to this question may seem obvious, but over the past century and a half there have been some interesting shifts in the ways by which firms have decided to invest in foreign markets. As documented by Dunning (1983) and Wilkins (1988), early European and North American multinationals were looking to either secure the supply of inputs and raw materials necessary for their operations, or to expand the size of the market for their products. Securing the supply of inputs and raw materials appears to have been the dominant motive when firms in the mining, oil, rubber, agricultural, and food industries internationalized in the late nineteenth century, whereas market-seeking motives or the ambition to expand sales volumes came to the fore after World War II.

Both resource- and market-seeking motives reflect the desire to protect or enhance the profitability of the company, sometimes complemented by a host of other but in many cases secondary motives. Such motives include the desire to improve overall performance by locating and coordinating select business activities in foreign countries, which is sometimes referred to as efficiency-seeking behavior, or to monitor and interact with competitors in international oligopolistic industries. More peripheral motives include founders' and managers' desire to create international business empires, which may coincide with other ambitions that do not necessarily prioritize shareholder wealth and dividends.

Especially in the case of market-seeking internationalization, firm-specific advantages and unique products and services have been instrumental for developing the desire and confidence to go abroad, and also for overcoming the disadvantages or liabilities of foreignness that affect firms when they first enter specific foreign markets (Zaheer, 1995). Such firm-specific

advantages may be proprietary technology and knowledge about products or services, or any set of factors that allows the firm to offer products or services that are unique or particularly appealing in terms of their cost–performance relationship. In many cases, such firm-specific advantages and unique products or services reflect the structure and characteristics of local industry, so that internationalizing firms may serve as indicators of industries in which the home country has proven particularly successful and competitive. There is also much that speaks for the gradual development of firm-specific advantages that derive from the firm's presence in several countries and markets (Rugman and Verbeke, 2009), but these types of advantages will not be examined in detail in the current account.

Emerging New Motives

While the desire to exploit existing firm-specific advantages has been a consistent theme in firms' internationalization, the past decades have nevertheless seen what appears to be a shift toward internationalization aimed at uncovering new or additional firm-specific advantages. What have been referred to as asset-seeking investments (cf Dunning and Lundan, 2008) imply that firms engage in business activities in foreign countries not to exploit their already existing advantages, but instead to seek complementary knowledge, build new advantages, or compensate for competitive weaknesses by establishing operations in foreign locations. While some firms may be looking for expertise in select functions or parts of their value-adding activities (e.g., firms may establish design centers in locations known for being the sources of new consumer trends), others may buy their way into technological capabilities and products that offer different and better growth prospects than those associated with the company's current business portfolio.

In yet other cases, the establishment of foreign operations can be a way to learn about foreign market demand and more efficient management practices. Such motives have been observed particularly among emerging MNCs from developing countries, which by linking into the global value chains created by established MNCs, leveraging their resources in terms of technology, knowledge, and brand names, and engaging in concerted learning efforts have created increasingly broad-based advantages to compete internationally (Mathews, 2002a; Buckley, 2009). Ultimately, such processes may result in acquisitions of firms in developed economies to get first-hand access to technology, management, and marketing expertise, or to capture established brand names and networks of suppliers and distributors (Deng, 2007). In yet other cases, investments in developed countries have been motivated by the need to alleviate or circumvent institutional constraints in the home market (Luo and Tung, 2007). From this perspective, and while the exploitation of firm-specific advantages probably has remained the main explanation for firms' internationalization, internationalization may capture both exploitative and explorative ambitions of the expanding business firm.

'When'

It is clear that internationalization of various forms has been going on for centuries, but in the context of multinational firms that own, control, and actively manage their international investments broad-based internationalization efforts coincided with industrialization and started taking off in the latter half of the nineteenth century. The foreign direct investment activities of emerging and already established MNCs increased steadily from the late nineteenth century and onward, especially among firms from North America and Europe (Dunning, 1983). Internationalization activity then decelerated during the interwar period, only to pick up again cautiously in the two decades after World War II and then more forcefully over the decades leading up to the present (**Figure 1**).

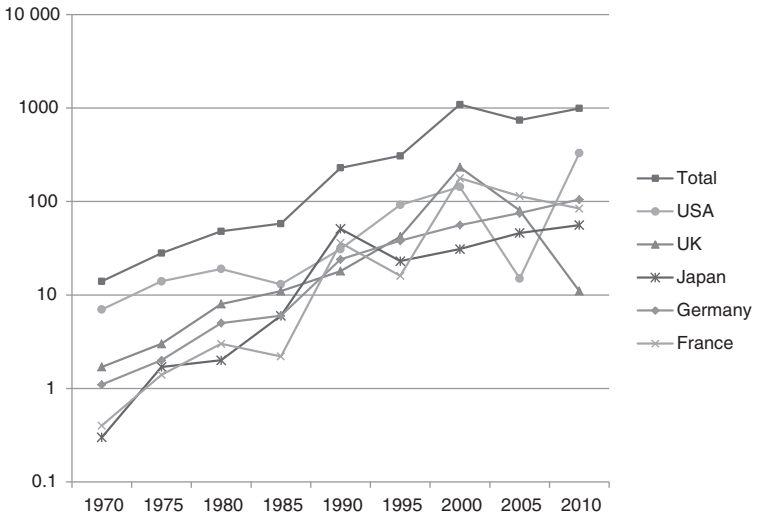

Figure 1 Growth in outward foreign direct investment, select developed countries ($ billion). UNCTAD, 2013. http://unctadstat.unctad.org/TableViewer/tableView.aspx?ReportId=88.

Patterns of foreign direct investment activity have been intimately connected to general conditions in the international business environment (Jones, 1996). During the interwar period, political unrest and spreading nationalism gradually curbed firms' internationalization ambitions and in some cases resulted in either divestments or nationalizations. In contrast, the postwar integration of world markets and more liberal attitudes toward foreign direct investments resulted in an environment much more benign toward firms' internationalization efforts (Dicken, 2007). Reduced tariffs, improved means of communication, and lowered transportation costs have all contributed to MNCs' willingness and ability to establish more extensive and coordinated networks of geographically dispersed units.

It is also clear that business cycles of shorter duration affect firms' willingness to engage in internationalization activities. Generally, it has been found that peaks in the global economy come with increased international merger and acquisition activity, whereas periods of slowdown or decline have the opposite effect (**Figure 2**). Such changes in the global economy may also affect the relative attractiveness of different modes of foreign operation, for example, greenfield investments and foreign acquisitions (UNCTAD, 2012).

The Firm Perspective

An alternative approach to the 'when' question is to take the firm's perspective, asking at what time in its development it decides to become engaged in international operations. There is not so much systematic information about this issue, partly because it is difficult to establish a common definition of the start of international activities. There are, however, indications that the timing of firms' internationalization efforts has seen some shifts over time.

For many and probably most business firms, the search for requisite resources and especially markets for their products and services starts in their home country (Wilkins, 1986). This is where founding entrepreneurs discover and develop their business ideas, and also where business models are first tested out and perfected. This also implies that at the time of the first internationalization efforts most firms have reached a certain age and size, although there may be quite some variation across the two parameters. Especially among firms with small home markets and which operate in scale-intensive industries, international activities and sales can be expected to start comparatively early on.

One of the most significant developments over the past decades is the emergence of a new type of firm, which appears to internationalize from very early on in its existence. These are firms that have been given a variety of names, including born globals (Knight and Cavusgil, 2004) and international new ventures (Oviatt and McDougall, 1994). Drawing upon an increasing number of internationally experienced individuals and often found in the high-tech industries, they already from inception combine and coordinate international operations, including the raising of capital, manufacturing and sales, and research and development activities. As resources tend to be scarce in such newly established and rapidly expanding firms, internationalization also tends to employ alternatives to foreign direct investment and ownership of assets. These alternatives include technology licensing and alliances, or various forms of contractual arrangements for securing access to critical components or distribution of products and services.

'Where'

The 'where' question divides into two related parts, one that concerns the sources of internationalization activity – or countries from which internationalizing firms originate – and one that concerns the target countries.

For a long period of time, the greater part of internationalization activity was connected to firms originating from relatively advanced economies. Wilkins (1988: p. 11) summarizes the conditions that applied to the growth of MNCs at the turn of the nineteenth century: "The more industrialized, the more economically advanced the country, *ceteris paribus*, the more likely for it to be the headquarters for business abroad." Throughout the first half of the twentieth century and beyond, more than 95% of all foreign direct investments were accounted for by MNCs from the developed countries, including the United States, the United Kingdom, Germany, and France (Dunning, 1983). From the 1960s and onward, firms originating from Japan came to play an increasingly

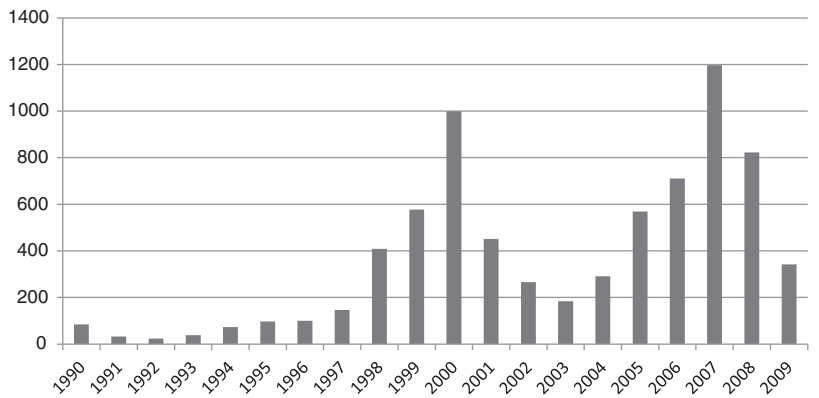

Figure 2 Value of cross-border mergers and acquisitions (valued at over $ 1 billion). UNCTAD, 2009. World Investment Report – Transnational Corporations, Agricultural Production and Development. United Nations, New York, and Geneva.

prominent role in total foreign direct investment activities (Dunning, 2009).

From about the 1990s and onward, however, the sources of foreign direct investments have increasingly been found among firms from developing countries. While developing country MNCs have existed for a long time, historically their foreign operations have been associated primarily with low-cost and undifferentiated products, target markets consisting primarily of other developing countries, and only marginal relative proportions of worldwide foreign direct investment activity. Much of this has changed during the past two decades, both in terms of modes of operation and overall impact on worldwide foreign direct investments (**Figure 3**). These are firms that will be returned to presently when discussing the 'how' question, as their entry into the world economy seems to have been made possible by strategic and organizational innovations well adapted to the international business environment at the opening of the twenty-first century.

The second part of the 'where' question concerns the countries that are selected as targets for internationalization and foreign direct investment activities. Generally speaking, while outward flows of foreign direct investments have traditionally been concentrated to a limited number of developed economies, the targets for internationalization activities and investments have been more diverse, with a significantly greater proportion accounted for by developing economies (Dunning and Lundan, 2008).

Patterns of Incrementalism

From the individual firm's perspective, it appears that the extension of the geographical scope of operations typically proceeds incrementally. In the early days of internationalization, starting in the late nineteenth century, internationalizing firms thus extended their activities into nearby or otherwise related countries (Dunning, 1983). Geographical and cultural proximity, or ties from periods of colonization played an important role in determining the location of the first investments of expanding MNCs. Such tendencies have been traced to the concept of 'psychic distance' (O'Grady and Lane, 1996), broadly defined as anything that prevents the flow of information between countries and creates barriers to learning about business conditions in a specific foreign market. Some of its main components include geographical distance and differences in culture, language, and political systems.

Geographical and other types of proximity indeed appear to underlie a consistent pattern in which MNCs prefer to trade and invest selectively in what may be referred to as 'close' countries (Dicken, 2007; Dunning and Lundan, 2008). For example, US multinationals have had a much stronger propensity to invest in Mexico and Latin America than their European and Japanese counterparts, and a major part of Latin American and Asian foreign direct investments have been made elsewhere in those regions. This, of course, does not preclude other investments in what would be more 'remote' countries, and there remains quite some variation in preferences across countries and individual firms, particularly in the case of born global firms or international new ventures. Nevertheless, the overall pattern indicates that the idea of truly global corporations in terms of operations and foreign investment activities is limited to a small number of cases (Rugman, 2001).

'How'

One major assumption about how firms internationalize has been that they commit resources to foreign markets incrementally, and only as uncertainty about local business environments is reduced and the investing firm gains access to wider opportunities in the local environment. The typical development of activities in any given country would go from initial exports or independent agent relationships to more committed forms such as wholly owned sales subsidiaries, and ultimately to wholly owned subsidiaries that involve more advanced functions such as manufacturing and research and development (Johanson and Vahlne, 1977).

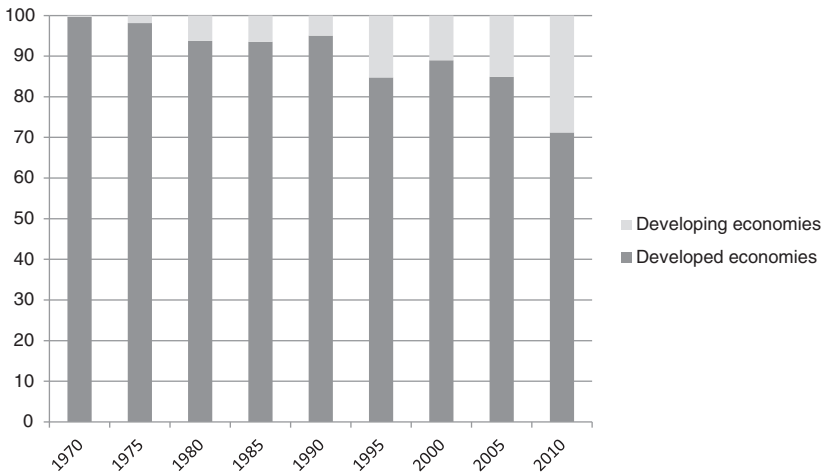

Figure 3 Distribution of outflows of foreign direct investment. UNCTAD, 2013. http://unctadstat.unctad.org/TableViewer/tableView.aspx?ReportId=88.

While the tendency to invest incrementally and cumulatively has been documented across a number of country settings (Welsh and Luostarinen, 1988; Johanson and Vahlne, 1990), developments over the past decades may have reduced its general validity. Some of the main influencing factors include more generally available information about the international business environment and more internationally experienced individuals, managers, and firms. For example, improved information and a sense of greater familiarity with foreign markets may partly explain why the proportion of greenfield ventures relative to acquired units declined from the turn of the nineteenth century and until the 1960s, and why since then mergers and acquisitions have become a favored but fluctuating means of foreign expansion, especially among MNCs from the developing countries (UNCTAD, 2012).

Other factors that may cause a departure from incremental and cumulative investment processes include industry-specific conditions and any firm-specific strategies for international expansion. It could also be added that through government intervention some countries may not be open to all types of investments by foreign firms, making it impossible for internationalizing firms to utilize otherwise preferred modes of local operations (e.g., China upheld strict restrictions on wholly owned foreign direct investments until the early 2000s). While many companies could thus be expected to commit resources to foreign markets in an incremental and cumulative way, individual cases may display significant deviations from the general pattern.

Internationalization at the Functional Level

Some investigations have looked at patterns in how firms internationalize various parts of their value-adding activities, including manufacturing and research and development activities. Some early attempts to describe and explain systematic changes in the nature of involvement in foreign countries emphasized the local origins of new product ideas and how throughout the product cycle product standardization causes a shift of manufacturing operations from the country of origin to foreign countries (Vernon, 1966). While the formation of well-established MNCs and increasing homogenization of the world's economies have altered some of the model's validity, the tendency to locate standardized manufacturing and other activities to low-cost countries has remained a significant feature of the well-established MNCs.

Starting in the late 1970s, the search for lowered costs of standardized manufacturing and other activities gave rise to a pattern by which well-established MNCs offshored and sometimes also outsourced select parts of their value-adding chains. Enabled by the rapid developments in information technology, much of these offshoring initiatives emanated from the United States and found their home in India, although over time patterns tended to become more geographically diverse. Offshoring has also come to include more advanced operations such as engineering and research and development (Lewin et al., 2009), perhaps reminiscent of the incremental and cumulative processes identified in firms' overall internationalization processes.

A second pattern in the internationalization of individual value-adding activities is the gradual increase in the amount and proportion of research and development carried out in foreign locations (UNCTAD, 2005). While asset-seeking investments are part of the explanation, the observed increase is probably in large part explained by independent and cumulative growth of foreign subsidiaries (Blomkvist et al., 2010). This growth appears to follow a pattern by which over time individual foreign units become more embedded in their local business environments – that is, forge closer contacts with existing suppliers, customers, and other local actors – and ultimately develop the capability to create and commercialize products and services that are sold locally and in some cases internationally as well. Such developments would then reflect a gradual deepening and extension of business activities carried out in individual foreign locations, although there remains quite some variation across types and levels of sophistication of foreign establishments.

New Currents

The general expectation about incremental internationalization and expansion of foreign operations disguises a more diverse pattern of internationalization among MNCs of more recent origin, as first documented in the context of firms from the Asia-Pacific region (Mathews, 2002b). Such firms may undergo accelerated internationalization by simply accessing and leveraging already existing resources in the global business environment while not necessarily becoming engaged in extensive ownership of these resources (Mathews, 2006; Mathews and Zander, 2007). They may not even start out with a distinctive technological edge or firm-specific advantage, but through technology acquisitions, alliances, and concerted learning efforts catch up with or leapfrog established competitors. Seen from a different perspective, such firms utilize already existing resources in the global business environment to identify and rapidly pursue business opportunities that have been either overlooked or on account of negligence and organizational heritage left unattended by incumbent firms.

Conclusions and Future Developments

Various forms of internal business activities have been carried out for centuries, and it is clear that in the long-term perspective there have been discernible patterns of internationalization, which fall outside of the scope of the current account. The patterns of internationalization that have been outlined here pertain to developments from the late nineteenth century and onward, and concern the operations of what could be described as modern MNCs, or those that engage in some form of ownership and coordination of geographically dispersed resources and operations.

At the fundamental level, patterns of internationalization reflect a set of processes that have remained more or less unchanged over time. As individuals discover new business opportunities they seek to exploit those opportunities in the best possible way, which includes attempts to protect and secure the new business and extend sales to markets as large as possible.

The exploitation of new business opportunities is accompanied by imitation and emerging competition, necessitating concerted efforts to control costs and to continuously upgrade existing products and services. In some cases, the continuation of profitable business may also require the development and entry into significantly new product and service offerings. In all of these undertakings, the international business environment offers potential advantages for expanding business firms, and it is thus not surprising that as a group they have embraced these possibilities and in many cases reacted in fairly similar ways.

In the expansion of international operations, whether upstream or downstream in the value chain, firms should be expected to favor locations where business contacts have been established, and also those locations that can be accessed, controlled, and coordinated with relative ease. From that perspective, and considering significant limitations to the extent individuals may move about in time and geographical space, internationalization patterns that embrace and build on the familiar and accessible are likely to persist also in the foreseeable future. Accordingly, there will also remain a fundamental tendency to approach and incorporate locations that are perceived of as more distant only in the later stages of the expansion of international operations. As firms face resource constraints in the short run, it seems equally likely that by and large investments in foreign locations are made incrementally and cumulatively.

This said, since the mid-twentieth century there have no doubt been considerable and perhaps accelerating changes in the preconditions for initiating and carrying out international business operations. From early on, through traveling, international education and later on involvement and careers in internationalized business firms, the individual can now increasingly establish and maintain a relatively broad network of international contacts. These contacts can then be used not only for the discovery of new business opportunities, but also for the expansion of new businesses in the international domain. Other enabling factors that come into play are easier access to knowledge and technology (e.g., through the Internet, commercially available databases, and markets for technology), improved and more rapid means of communication and transportation, and the harmonization of institutional conditions and legal systems to make different countries and business environments appear increasingly similar and 'familiar.'

There are thus good reasons to expect locationally determined and incremental patterns of internationalization to give way to increasingly diverse ways of creating and exploiting new business opportunities. This then involves expectations about a growing proportion of various types of born global firms or international new ventures, and also about firms originating from developing economies that find new and in many cases innovative ways of responding to previously unexplored business opportunities. It could be added that such firms have probably existed longer than their relatively recent discovery in the academic literature suggests, but as of yet there are few sources that can firmly document how the relative proportions of different types or 'species' of firms have evolved over time.

Old and New Patterns of Internationalization

It may indeed be useful to make a distinction between patterns of internationalization that apply to the old and by now well-established MNCs, and those that apply to multinationals of more recent origin. The now well-established MNCs typically came about during times when knowledge and the individual's experiences were location-bound, giving rise to patterns of internationalization characterized by a gradual expansion from a relatively well-defined point of departure. In contrast, MNCs of more recent origin would be less tied to local conditions and experiences in the identification of new business opportunities, and through processes of linkage, leverage, and learning also be able to rapidly access and utilize a larger set of geographically dispersed resources (Mathews, 2002a). The opening up of previously protected and regulated markets such as China,

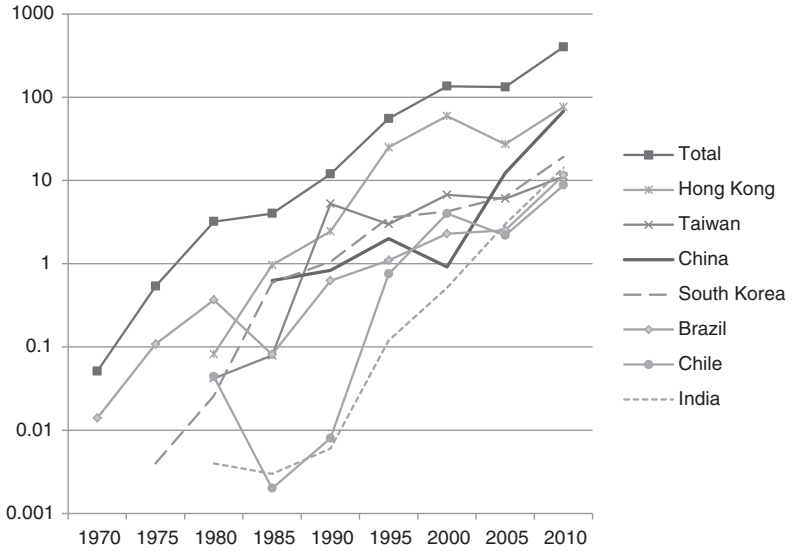

Figure 4 Growth in outward foreign direct investment, select developing countries ($ billion). UNCTAD, 2013. http://unctadstat.unctad.org/TableViewer/tableView.aspx?ReportId=88.

India, and other developing economies offers additional opportunities for new firms to enter into markets that have traditionally been neglected by many old-time MNCs. The overall conclusion would be that apart from assuming a more prominent position in the population of different types or species of MNC, recently established MNCs will display internationalization patterns that are increasingly diverse in terms of origins and characteristics and also unfold more rapidly than before.

What seems to be beyond dispute is that new firms originating from the developing economies have now started to account for an increasing proportion of all outward-bound foreign direct investment activity, and at a pace that surpasses that of MNCs from the developed economies (Figure 4).

The pattern which has been observed over the past three decades may signal a significant reshaping of the competitive landscape, and also the advent of some still unexplored patterns in the interaction and international activities of established and newcomer firms.

See also: Business Models; Commodity Chains; Cyclical Industrial Dynamics; Diversification, Relatedness, and the New Logic of Co-creation; International Business; Interorganizational Relationships and Networks; Multinational Corporations; Strategic Adaptation; Strategic Management; Supply Chain Management.

Bibliography

Blomkvist, K., Kappen, P., Zander, I., 2010. Quo vadis? The entry into new technologies by foreign subsidiaries of the multinational corporation. Journal of International Business Studies 41 (9), 1525–1549.

Buckley, P.J., 2009. The impact of the global factory on economic development. Journal of World Business 44, 131–143.

Deng, P., 2007. Investing for strategic resources and its rationale: the case of outward FDI from Chinese companies. Business Horizons 50, 71–81.

Dicken, P., 2007. Global Shift – Mapping the Changing Contours of the World Economy, fifth ed. Sage Publications, London.

Dunning, J.H., 1983. Changes in the level and structure of international production: the last one hundred years. In: Casson, M. (Ed.), The Growth of International Business. George Allen & Unwin, London.

Dunning, J.H., 2009. The key literature on IB activities: 1960–2006. In: Rugman, A.M. (Ed.), The Oxford Handbook of International Business, second ed. Oxford University Press, Oxford.

Dunning, J.H., Lundan, S.M., 2008. Multinational Enterprises and the Global Economy, second ed. Edward Elgar, Cheltenham.

Johanson, J., Vahlne, J.-E., 1977. The internationalization process of the firm—a model of knowledge development and increasing foreign market commitments. Journal of International Business Studies 8 (1), 23–32.

Johanson, J., Vahlne, J.-E., 1990. The mechanisms of internationalization. International Marketing Review 7 (4), 11–24.

Jones, G., 1996. The Evolution of International Business: An Introduction. Routledge, London.

Knight, G.A., Cavusgil, S.T., 2004. Innovation, organizational capabilities, and the born-global firm. Journal of International Business Studies 35 (2), 124–141.

Lewin, A.Y., Massini, S., Peeters, C., 2009. Why are companies offshoring innovation? The emerging global race for talent. Journal of International Business Studies 40, 901–925.

Luo, Y., Tung, R.L., 2007. International expansion of emerging market enterprises: a springboard perspective. Journal of International Business Studies 38 (4), 481–498.

Mathews, J.A., 2002a. Dragon Multinational: Towards a New Model of Global Growth. Oxford University Press, New York.

Mathews, J.A., 2002b. Comparative advantages of the latecomer firm: a resource-based account of industrial catch-up strategies. Asia Pacific Journal of Management 19, 467–488.

Mathews, J.A., 2006. Dragon multinationals: new players in 21st century globalization. Asia Pacific Journal of Management 23, 5–27.

Mathews, J.A., Zander, I., 2007. The international entrepreneurial dynamics of accelerated internationalization. Journal of International Business Studies 38 (3), 1–17.

O'Grady, S., Lane, H., 1996. The psychic distance paradox. Journal of International Business Studies 27 (2), 309–333.

Oviatt, B.M., McDougall, P.P., 1994. Towards a theory of international new ventures. Journal of International Business Studies 25 (1), 45–64.

Rugman, A.M., 2001. The End of Globalization – Why Global Strategy is a Myth & How to Profit from the Realities of Regional Markets. AMACOM, New York.

Rugman, A.M., Verbeke, A., 2009. Location, competitiveness, and the multinational enterprise. In: Rugman, A.M. (Ed.), The Oxford Handbook of International Business, second ed. Oxford University Press, Oxford.

UNCTAD, 2005. World Investment Report – Transnational Corporations and the Internationalization of R&D. United Nations, New York and Geneva.

UNCTAD, 2009. World Investment Report – Transnational Corporations, Agricultural Production and Development. United Nations, New York and Geneva.

UNCTAD, 2012. World Investment Report – Towards a New Generation of Investment Policies. United Nations, New York and Geneva.

UNCTAD 2013. http://unctadstat.unctad.org/TableViewer/tableView.aspx?ReportId=88.

Vernon, R., 1966. International investment and international trade in the product cycle. The Quarterly Journal of Economics 80 (2), 190–207.

Welsh, L.S., Luostarinen, R., 1988. Internationalization: evolution of a concept. Journal of General Management 14 (2), 34–64.

Wilkins, M., 1986. The history of European multinationals: a new look. The Journal of European Economic History 6, 483–510.

Wilkins, M., 1988. European and North American multinationals, 1870–1914: comparisons and contrasts. Business History 30, 8–45.

Zaheer, S., 1995. Overcoming the liability of foreignness. Academy of Management Journal 38 (2), 341–363.

Internet and Culture

Marc Verboord and Susanne Janssen, Erasmus University Rotterdam, Rotterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The increased use of the Internet since the late 1990s has had a large impact on how arts and culture products such as films, books, cds, museum exhibitions, etc., are produced, distributed, acquired, and received. This article aims to discuss some of the most relevant developments in the study of the Internet's relation to arts and culture from a social-scientific perspective. It focuses on (1) processes of producing and distributing culture (material production); (2) ways in which cultural products are classified and legitimated (symbolic production); and (3) behaviors and perceptions of cultural audiences.

Introduction

The increased use of the Internet since the late 1990s has had a large impact on how arts and culture products such as films, books, cds, museum exhibitions, etc., are produced, distributed, acquired, and received. The World Wide Web has shortened communication lines between companies and between persons, enabling global contacts to strengthen and cultural exchanges to increase. How increasing usage of Internet and other new media continues to shape our social life has been among the fastest growing topics in communication research in the past decade. The consequences for the domain of arts and culture are in some respects much debated (e.g., the controversies of illegal downloading), yet in other respects rather taken for granted (e.g., the decline of cultural authority).

This article aims to discuss some of the most relevant developments in the study of the Internet's relation to arts and culture from a social-scientific perspective. Although the focus will mainly be on those cultural artifacts and events which are classified as forms of 'art' – be it highbrow, middlebrow, or popular – by legitimate experts in society, one of the conclusions of this article will be that on the Internet also new, less classifiable cultural genres emerge. Similarly, the boundaries of the Internet are not always clear, as its increased usage is accompanied by many related new media and *information and communications technologies* (ICTs) (e.g., hardware devices such as cell phones and iPod and software) which can also be objects of study. This overview aims to discuss three aspects of cultural fields in which the influence of Internet is observed: processes of producing and distributing culture (material production); ways in which cultural products are classified and legitimated (symbolic production); and behaviors and perceptions of cultural audiences. In doing so, we draw upon various research perspectives, which will first be introduced in Section Research Perspectives. This article cannot claim to be exhaustive, but at the same time we point out that large parts of the territory are still unexplored. Not only does much of the current literature have a much broader take on 'culture,' most of the work that does address arts and culture in a more strict sense remains largely theoretical or practical, but yet lacks empirical backup.

Research Perspectives

Internet has revolutionized (re)production, distribution, and consumption of media materials in at least three important ways. First, its networked structure allows Internet users – regardless of their status or profession – to connect directly to almost all other Internet users and exchange media content without the intervention of professional gatekeepers. Second, Internet has incorporated technology that is both available and accessible for many Internet users, which can be applied to create, edit, upload, and download all types of text-based and audiovisual content. Thirdly, Internet is increasingly seen – particularly by the most recent birth cohorts – as the principal medium to gather information on or access to culture and entertainment at the expense of print media, radio, and television. Losing the monopoly to mass production and distribution of cultural goods and weakened control over information flows have both economic and symbolic implications for cultural producers as well as consumers. At the core we find the hierarchically ordered 'cultural field' in which professionals engage in cultural production through highly institutionalized practices and logics (Bourdieu, 1993) being challenged by alternative practices. Scholars have begun to theorize and, to a lesser extent, empirically study these developments in various ways.

Within communication research, we find three important perspectives. Digital inequality studies are concerned with audiences and users of ICT and Internet – more particularly, who have access to new media and who are most likely to actually use such media for social participations (e.g., in the realm of culture). Increasingly, however, notions of 'audiences,' 'users,' and 'consumers' are contested for dismissing the blurring of boundaries between consumers and producers. In particular, work on peer production – or user-generated content – details the work of 'prosumers': nonprofessionals who create cultural content online which sometimes invokes semi-institutionalized fields on its own, such as the blogosphere (Bruns, 2008). These studies also showed that despite the wide reception and application of this user-generated content, only a small segment of the population is involved in its active creation (Van Dijck, 2009). A third and related scholarship analyses computer-mediated communication and

its effects on behaviors and attitudes. While often not particularly concerned with culture, it has produced relevant insights in how more or less institutionalized forms of communication (e.g., online reviews of amateurs vs professionals) influence recipients' perceptions. The significance of expertise and authority is among the focal points – particularly in studies on how 'collective knowledge' is produced on the Internet (e.g., online encyclopedia Wikipedia).

Internet's economic implications with regards to arts and culture tend to diverge in two main directions: consequences for business models and for marketing strategies. Media economists (but also other publicists) have started to analyze how the cultural industries have (not) adapted to changes in the way cultural products are acquired through online outlets (e.g., illegal downloading, web stores).

At the same time, proposals for new operative business strategies are being launched which explicitly incorporate characteristics of online consumption modes, such as 'collective knowledge' production (e.g., Wikipedia) or 'crowdsourcing' (e.g., www.sellaband.com). Marketing specialists have recognized one particularly salient trait of Internet communication: the possibility to create 'word-of-mouth' effects in an online context, thereby generating a certain 'buzz' which is both perceived as authentic valorization by peers and yet can have the scale of mass advertising. This has been labeled Electronic Word-of-Mouth (eWOM). Noninstitutionalized platforms for 'user-generated content' such as cultural blogs can consequently be targeted to have marketing messages distributed through bottom-up processes.

A third strand of research addresses the symbolic production of culture on the Internet. One emancipatory manifestation of such work – emphasizing how fans and other 'common' (prod)users – negotiate the meaning attributed to popular culture – is found within the paradigm of cultural studies. Most of this work concerns ethnographic accounts that elucidate the place of audiences in the 'convergence culture' (Jenkins, 2006), a constellation in which media are understood to fulfill increasingly similar functions for ever more indistinct social groups. As such, often social inequalities in terms of education, age, or gender are argued to be no longer significant markers for taste and behavior, and notions of cultural authority are challenged. However, studies in cultural sociology argue that Bourdieu's (1984) concept of 'distinction' may have weakened due to the ascendance of more omnivorous cultural consumption patterns – that is combining highbrow and popular culture – but cannot be completely abandoned. Popular culture has also become more institutionalized over time, and many of such off-line configurations tend to be reproduced in online contexts.

Material Production and Distribution

Most discussion on how Internet affects the world of arts and culture apparently concerns the way cultural products are produced and distributed. Illegal downloading is a major concern in the cultural industry: all cultural goods that can be electronically copied (music, film, photographs, books) are increasingly being disseminated without permission of the original cultural producers who consequently suffer loss of revenue (Levine, 2012). Yet also legal ways in which products find their way to audiences (via web stores or intermediary sites such as YouTube or Flickr) affect how the cultural industry as mediators are being sidetracked. Although the disruption of traditional business models in the cultural industries has prompted many observers to signal the start of a completely new era – for the worse or for the better – in many ways there is also continuation of old discussions.

In mid-twentieth century, the introduction of 'reproduction' of recorded images (film, television) and sounds (music), aroused strong debates on what constitutes 'art' and to what extent cultural products could be 'commodified' in an almost industrial way (e.g., Frankfurter Schule). Despite reproducible products becoming key ingredients in the diets of many cultural audiences in the second half of the century – and thus losing most of their controversy – criticism never completely disappeared. While many sociological and media/cultural economic studies pursue objective analysis of production practices and their implications, normative perceptions on cultural value or the functioning of the industry are still widely spread. Examples of such expressions are for instance found in critiques on how book publishing threatens standards of literary culture, or in the political economist's concerns that the corporations involved in cultural production reduce the variety of opinions and tastes through their shaping of news, entertainment, and other media expressions.

These discussions resonate in theoretical publications on cultural production in the digital age, which tend to take two competing perspectives. Both of them, however, appear to draw on three underlying argumentative dimensions: economic, artistic, and moral considerations. Among Internet's criticasters, a first group emphasizes the destructive effects that Internet-related practices have on the existing cultural industries. Their main argument is that the 'free riding' behavior, which Internet provokes will decrease incomes such that it will ultimately drive cultural producers out of business (e.g., Levine, 2012). This view is countered, however, by various theorists who argue that Internet also offers new opportunities for business as producers can take advantage of 'collective wisdom' production and user-generated content to gain more insight in consumer demands (e.g., Tapscott and Williams, 2008).

A second type of arguments concerns the 'quality' of the output. Antagonists mainly lament the decline of artistic – but also journalistic – quality and diversity in content that Internet allegedly induces. In their view, 'user-generated content' does not meet the standards of institutionally produced content, due to the absence of professional gatekeepers, which would degrade the available supply (e.g., Keen, 2007). Again, opposite viewpoints are taken. Particularly in the domain of news production, some argue that 'user-generated' content increases diversity of voice, and as such contributes to a multiform society. In addition, Chris Anderson (2006) in his book *The Long Tail* points at increasing diversity at the aggregate level. Centralizing the storage and searchability of cultural products in an online context (e.g., a web store like Amazon) enables consumers to discover products out of the mainstream, and, consequently, producers to exploit niche markets, which were not profitable before the digital age. He argues that this may very well lead to the end of the 'blockbuster economy.'

Finally, there is a third group of scholars that stresses more moral and ethical aspects of Internet, particularly the issue of copyright and ownership of cultural content. Violation of copyright is one of the main concerns of the industry with regards to Internet distribution of culture – and this obviously includes also the creators of culture themselves (the artists). Besides the loss of economic revenue, infinite reproduction and unofficial distribution via new media also invokes a moral contemplation of the value of 'creation.' Artists who take considerable pains in creating new works of art see themselves confronted with audiences who increasingly expect free and instant satisfaction of their individual cultural wants – disregarding the artists' investments. The advent of 'user-generated content' fueled this debate on copyrights, as we find also theorists who argue that taking existing content to recreate and remix cultural expressions is not very different from traditional artists finding inspiration in folk tales or using samples and quotes (Lessig, 2008). On the most fundamental level, however, copyright discussions address the question whom the current system truly benefits most – and whether that is changing: the media conglomerates that do most investments, the artists who are the source, and/or the art that seeks innovation.

Many of these issues regarding production and distribution currently remain unresolved, as empirical evidence is often lacking or hard to find (Curran et al., 2012: 69ff). For instance, there is, as yet, no clear evidence that The Long Tail has indeed become such profitable market niche that it inflicts death on the blockbuster. And while illegal downloading in all likelihood affects the cultural industries, these industries have also been slow to innovate: record companies have been retailing their back catalog for years (via the cd); film companies have increasingly centered their production around formulaistic high-spectacle blockbusters. Meanwhile many new business models – such as crowdsourcing models or 'value-based pricing models' – are still too young for definite conclusions (Wikström, 2009). And lastly the copyright wars are not yet decided – despite new initiatives such as 'blanket protection' in which the rights to distribute products are granted to Internet service providers in exchange for payment – being underway (Levine, 2012).

Meanwhile new media are increasingly applied within the marketing of cultural products. In the film industry, trailers of new movies are increasingly disseminated via the larger film blogs that in many ways have taken over the dominance of trade journals in breaking news and reporting on the movie scene. For the book industry, Thompson (2010: 243ff) shows how author tours and print advertising are increasingly being replaced by online advertisements and building word-of-mouth via the blogosphere. Getting positive reviews are not the aim. Instead, publishers target a relatively small group of influential bloggers in order to create a certain 'buzz,' which can translate in off-line demand.

Symbolic Production

Consensus on what cultural manifestations are artistically valuable or important – in other words, (are) legitimate – has traditionally been reached through highly hierarchical systems. Within thematically organized domains or fields, institutions operated through agreements on how production and distribution practices should be performed (Bourdieu, 1993). Both institutions and individual agents within these institutions nevertheless compete for cultural authority through securing and expanding their influence in the field, as higher levels of assigned expertise translate in more status. At the same time, the agency of audiences – when contributing to the process of legitimation – has been more of a receptive nature than a productive one.

Yet value attribution to cultural products is not hierarchical per definition. Legitimation should be seen as a social process, since institutional agents rely heavily on having an audience that accepts their selections and evaluations, or in other words puts 'belief' in them (Bourdieu, 1993). In the traditional system, print media such as newspapers, opinion magazines, and academic journals served as outlets for institutional agents, stressing the top-down process of legitimation by its unidirectional character from producer to consumer. The reason for this was twofold: (1) the selection of persons who could enter the field was confined to those showing the ideas already dominant in the field and (2) the ideas of these agents working in the media field on what makes legitimate culture resembled the ideas of their audiences since they were largely recruited from the same social ranks (Bourdieu, 1993).

The arrival of the Internet has led to less hierarchical systems of classifying information – among which evaluations of culture – due to its specific characteristics, as is theorized by Verboord (2010, 2013). The open nature of the Internet has enabled cultural audiences without institutional ties to engage in discussions of quality or value beyond merely reacting to professional critics. Not only do audiences have the means to be faster to evaluate (as publications of traditional print media are often still slowed down by editorial checks and publishers' time to process), they also publish outside traditional institutional settings. Particularly within popular culture, discussion forums and consumer review sites circumvent professional critics and authorities. Often, evaluating culture is also more explicitly linked to the purchasing of cultural goods (e.g., web stores such as Amazon) and is rescaled from a few extensive, in-depth evaluations to a large number of short, sometimes ill-written, biased, or even invalid comments (cf David and Pinch, 2006). Symbolic production in the cultural field has thus been restructured in the sense that dispersion of audience attention, larger diversity of voice, uneven quality in writing and argumentation has made it harder for cultural critics to solidify their authority. In this fast changing media landscape, media outlets, and the critics therein, have more and more difficulty to stay connected to specific audience groups. This declining homology between media and audiences is particularly salient on the Internet, which puts considerable pressure on the legitimacy of critics and related institutional agents. This does not mean expertise is no longer important, however. The empirical results of Verboord (2010) demonstrate that expert critics are still valued higher than online critics. Similarly, studies in computer-mediated communication show that textual properties and communication tactics are highly important to be considered as prominent and 'useful' reviews (e.g., Otterbacher, 2011). Traits associated with professionalism such as correct spelling and high levels of argument

density are thus used to filter out credible information in the enormous online supply. Still, the ascendance of peer-produced content challenges the hierarchical model of cultural evaluation, though it remains in use, but particularly stimulates the use of popular aesthetics as the standard form of evaluative discourse (Verboord, 2013).

It remains to be seen if symbolic production has become more democratic and multiform in the digital age. Traditionally, cultural genres with more social status tend to show more inequalities in terms of gender, ethnicity, and social class. An exploratory analysis of the fiction book output in 1 month in the United States showed that female authors generally were more often reviewed on Amazon and social network Goodreads than male authors, whereas print media showed no significant difference (Verboord, 2011). This result may, however, be typical for the book market in which women form a large part of the readership.

Audiences

Cultural Participation

Much of the reduction in control that cultural producers used to have in producing and disseminating content has accrued to audiences. Provided that one has the access and the digital skills to go online, it is often easier, faster, and cheaper to purchase products and tickets, find information, contact like-minded persons, share opinions and emotions on experiences, etc. in this digital environment. Representative studies into the way cultural participations are affected by online mediation are rare, however. One of the few attempts to map cultural participation through new and traditional media is conducted in 2009 at the Netherlands Institute for Social Research for a representative sample of the Dutch population, aged 18 years and older (Sonck and De Haan, 2012). Contrasting popular culture (pop music, film) and traditional culture (theater, visual arts, cultural heritage, classical music) in both receptive and more active modes of participation, they show huge discrepancies across activities and cultural forms. Overall, old media are still used more often than new media. Even among the youngest age group – 18–34 – new media have a complementary rather than a substitutional role.

In absolute terms, Internet is most often used for finding information: more than 40% of the population has ever looked for information on pop music, film, or theater. For the other art forms, this percentage lies between 10 and 20%. About a quarter of the population buys tickets and/or recordings of pop music or film via the Internet. Still, only with regards to pop concert tickets, online buying has actually surpassed off-line buying. Purchasing highbrow culture via the Internet is mainly restricted to buying theater tickets (34% vs 36% offline). Actual (mediatized and receptive) cultural participation – viewing and/or listening to objects and recordings – is predominantly done via audiovisual media. Online participation amounts to almost 20% for all cultural forms, except classical music (only 5%) and pop music (almost 30%). Despite web2.0 tools becoming increasingly available, active forms of participation are not widely spread. Creating and sharing cultural content were established for a third of the Dutch, but this includes photographs and diary fragments which may be uploaded just for friends. Only 5% reported to ever have posted messages on arts and culture.

These online participation trends – though perhaps partly superseded by advancing ICTs and media logics – reveal the contextual imperative of the social shaping of technology. Online participation variably relates to particular issues such as scarcity (ticket purchase for performances), searchability (information), replicability (music files), or archivability (cultural heritage, visual arts). Contrary to what some Internet gospellers state, simply having the opportunity to use the Internet for a particular purpose is clearly not enough to propel such usage. Dependent on their skills, attitudes, and incentives, audiences pursue applications that enable them to efficiently find desired gratifications and affordances. Which online services come into vogue is difficult to predict. Successful online services seem to originate more often from outside traditional cultural production and benefit particularly popular culture. Although producers of highbrow products similarly need to relate to their clientele, this audience group is generally smaller, older, and more aligned to institutionally rather than peer-produced expertise. As a result, many of these producers still take the tradition as the point of reference, such as virtual museums, which are principally created to make collections and associated documentations more accessible, and secondly want to provide an entertaining and educational experience (Styliani et al., 2009).

Whereas such highbrow endeavors to reach out to audiences still command high levels of control, more popular culture has generated mostly interactive platforms where audiences are also enabled to become creators themselves. Video Web site YouTube is perhaps the most used and most famous form of such participatory culture: about half of its contents appears to be user-generated (Burgess and Green, 2009). The Web site exemplifies the notion of 'convergence culture' through its integration of different forms of media (television, film, music, photographs), but just as much in how culture is represented. Next to the wide circulation of original content produced in cultural industries' formats, many alternative versions of such materials are available in various degrees of editing, mixing-up, animation, imitation, etc. (Burgess and Green, 2009: pp. 42–43). Such media work of (semi-)amateurs consequently generates new, at times unclassifiable, genres outside of institutionalized settings. It is here where the cultural participation of audiences comes closest to reordering the status quo of cultural fields.

Cultural Information Regimes: Between Mass and Niche

The experiential nature of arts and culture together with the infinite variety of products implies that information on cultural products comprises a double-edged uncertainty. Since audiences can only evaluate products after consuming and experiencing them, they do not know whether they choose the 'right' product before acquisition. Consequently, producers do not know what products will become successful among audiences. Several information regimes are applied to reduce uncertainty. The most traditional, highbrow-oriented regime of cultural classification was already discussed in the previous section. On the Internet, market information

regimes, which inform producers and consumers on popularity, appear more important, however. It is almost standard practice on the web to inform Web site visitors about the popularity of the Web site itself, particular texts on that site (e.g., by letting visitors assess the usability, or give a Facebook like), or products on the site (e.g., book sales on Amazon.com; number of previous downloads).

The potential power of online-disseminated market information was demonstrated in a web-based experiment by Salganik and Watts (2008) that created parallel online universes in which the same (relatively unknown) pop songs were attributed different popularity scores by the researchers. Once consumers could start to listen to the song corpus and download their preferred tracks, the demand curve developed consistently in favor of those songs which had the largest (manipulated) download scores from the start. Yet if the same songs in a different online universe had been flagged as not been downloaded a lot, demand did not grow hard. This result underscores the more general observation that although Internet has increased the supply of media content to almost infinite proportions, demand appears to be highly concentrated because information overload is an impetus to use filters such as popularity markers.

Despite these results, empirical studies suggest that interpersonal ties still prevail above Internet sources when looking for cultural information (Kayahara and Wellman, 2007). While Internet is frequently used to find detailed information on particular products or events, recommendations on which artists or products to look for in the first place often still comes from the personal off-line network. What is more, despite technology use to find new music being highly diversified, it does not seem to help more people to become opinion leaders in their social circles nor online (Tepper and Hargittai, 2009). In a similar vein, participation in social network sites mainly follows from off-line contacts or interests, making it easier to connect to like-minded persons, but not necessarily leading to new friendships (Baym and Ledbetter, 2009). Still, there is increasing evidence that social media are reshaping cultural consumption practices (Dewan and Ramaprasad, 2012).

Discussing Culture: The Voice of the Consumers

Besides transforming systems at the societal level – such as access to cultural content and currency of information regimes – user-generated content is also strongly associated with new ways in which content is discussed and deliberated at the level of individual consumers of (popular) culture. Again, partly this follows from increased network opportunities. Internet has provided fans of popular culture with the means to (virtually) meet like-minded persons and share thoughts, interpretations, and evaluations on specific web forums. To be sure, the rise of fan culture has been underway since the breakthrough of popular culture in the 1960s, but from the start Internet proved to be an important facilitator of communication. Having started out as a domain for 'geek culture' in which mostly 'tech-savy' males propagated genres such as science fiction and alternative rock, it has widened substantially. Moreover, online fan culture has developed a voice, which exceeds in-crowd discussions on trivial content details. Particularly, Jenkins (2006) provides an account – within the broader theoretical framework of media convergence – of how new media enable fans to influence practices of corporate producers. Fans contribute to the producers' deeper understanding of how their work is interpreted and evaluated, and, ultimately, how program loyalty can be achieved. Interactive television formats in which viewers can for instance vote in talent shows accommodate this connection to the fan base. At the same time, some content fans create online may harm the interests of producers. Examples concern new stories on Harry Potter, which violate copyright laws or the emergence of 'spoiler' sites on which fans try to disclose television content before it is officially broadcasted. Although originally many of such practices are the work of small groups situated in relatively obscure online communities, grassroots convergence has empowered fans through collective intelligence. Interactivity first stirs the depths and intensity of communication within such communities. Subsequently, the content at stake – if it gains sufficient momentum – gets picked up by outlets higher in the media hierarchy, after which it may continue to be circulated until it reaches mainstream audiences. One important remark needs to be made: while such bottom-up processes do allow new products or creators to rise to stardom (e.g., rock band Arctic Monkeys via MySpace), most participations reported in these fan studies still concern popular culture produced in the cultural industries. Moreover, this is in line with results from other studies into participatory culture such as YouTube: user-created content augments diversity yet audience attention is most drawn by what originated from traditional media (Burgess and Green, 2009).

Prospects

Despite the rapidly growing literature on new media and Internet, research which focuses on their impact on arts and culture remains relatively limited and scattered across disciplines. Internet's novelty, its omnipresence and the fast pace at which it penetrated Western society not only spurred a great variety of perspectives, but also stimulated theory to regularly develop without empirical falsification.

More empirical evidence for theorized effects is needed. Part of such empirical underpinning would entail employment of multiple methods and cross-disciplinary perspectives. For instance, studies into online fan culture remain persistently qualitative in nature – thereby preventing generalization of findings. At the same time, research into effects of online reviews largely relies on quantitative measures, often based on student experiments.

Connecting different research perspectives thus appears to be one of the major challenges for the study of arts and culture in the digital age. Such interconnections help to better understand some of the paradoxes that emerge from the current literature. For instance, how the creation of user-generated content has reversed many cultural flows from top-down to bottom-up, but at the same time most cultural content which is circulated online still springs from the cultural industries (either in original or remixed versions) and only a limited number of Internet persons can actually be called 'produsers.'

Or the observation that the Internet facilitates the convergence of media, cultural genres, and even consumption and production practices, yet what is actually consumed and produced is subject to increasingly larger fragmentation and individualization. Also, the role of social network sites (e.g., Facebook) in taste formation, the use of technology to recommend consumers mainly cultural products very similar to what they purchased before, and the future of physical cultural outlets in an increasingly digitalizing world are important themes for future research.

See also: Computer Mediated Communication; Cultural Mediators and Gatekeepers; Cultural Production in Networks; Culture, Production of: Prospects for the Twenty-First Century; Markets: Artistic and Cultural; New Media and the Digital Divide; Symbolic Boundaries; Video Cultures: Television.

Bibliography

Anderson, C., 2006. The Long Tail. Why the Future of Business is Selling Less of More. Hyperion, New York.

Baym, N.K., Ledbetter, A., 2009. Tunes that bind? Predicting friendship strength in a music-based social network. Information, Communication & Society 12, 408–427.

Bourdieu, P., 1984. Distinction. A Social Critique of the Judgment of Taste. Routledge & Kegan and Paul Ltd, London.

Bourdieu, P., 1993. The Field of Cultural Production. Polity Press, Cambridge.

Bruns, A., 2008. Blogs, Wikipedia, Second Life and beyond: From Production to Produsage. Peter Lang, New York.

Burgess, J., Green, J., 2009. YouTube. Online Video and Participatory Culture. Polity Press, Cambridge.

Curran, J., Fenton, N., Freedman, D., 2012. Misunderstanding the Internet. Routledge, London.

David, S., Pinch, T., 2006. Six degrees of reputation: the use and abuse of online review and recommendation systems. First Monday 11 (3). http://www.firstmonday.org/issues/issue11_3/david (accessed 15.11.07.).

Dewan, S., Ramaprasad, J., 2012. Music blogging, online sampling, and the Long Tail. Information Systems Research. First published online 13 February 2012, as. http://dx.doi.org/10.1287/isre.1110.0405.

Jenkins, H., 2006. Convergence Culture. Where Old and New Media Collide. New York University Press, New York.

Kayahara, J., Wellman, B., 2007. Searching for culture – high and low. Journal of Computer-Mediated Communication 12, 824–845.

Keen, A., 2007. The Cult of the Amateur. How Today's Internet is Killing Our Culture. Doubleday/Currency, New York.

Lessig, L., 2008. Remix. Making Art and Commerce Thrive in the Hybrid Economy. Penguin, London.

Levine, R., 2012. Free Ride. How the Internet is Destroying the Culture Business and How it Can Fight Back. Vintage, London.

Otterbacher, J., 2011. Being heard in review communities: communication tactics and review prominence. Journal of Computer-Mediated Communication 16, 424–444.

Salganik, M.J., Watts, D.J., 2008. Leading the herd astray: an experimental study of self-fulfilling prophecies in an artificial cultural market. Social Psychology Quarterly 71, 338–355.

Sonck, N., De Haan, J., 2012. De Virtuele Kunstkar. SCP, The Hague.

Styliani, S., Fotis, L., Kostas, K., Petros, P., 2009. Virtual museums, a survey and some issues for consideration. Journal of Cultural Heritage 10, 520–528.

Tapscott, D., Williams, A., 2008. Wikinomics: How Mass Collaboration Changes Everything. Atlantic Books, London.

Tepper, S.J., Hargittai, E., 2009. Pathways to music exploration in a digital age. Poetics 37, 227–249.

Thompson, J.B., 2010. Merchants of Culture. The Publishing Business in the Twenty-First Century. Polity Press, Cambridge.

Van Dijck, J., 2009. Users like you? Theorizing agency in user-generated content. Media. Culture & Society 31, 41–58.

Verboord, M., 2010. The legitimacy of book critics in the age of the Internet and omnivorousness. European Sociological Review 26, 623–637.

Verboord, M., 2011. Cultural products go online. Comparing the Internet and print media on distributions of gender, genre and commercial success. Communications 36, 441–462.

Verboord, M., 2013. The impact of peer-produced criticism on cultural evaluation. A multi-level analysis of discourse employment in online and offline film reviews. New Media & Society. First published on 3 July 2013 as. http://dx.doi.org/10.1177/1461444813495164.

Wikström, P., 2009. The Music Industry: Music in the Cloud. Polity Press, Cambridge.

Internet and Privacy

Andreas Busch, Georg-August-Universität Göttingen, Göttingen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The issue of informational privacy has grown in political prominence and importance as use of the Internet has spread around the globe. Economic interests base business models on users' indirect payment with personal data; criminals engage in fraudulous 'identity theft'; and private and public sector surveillance is conducted in order to increase security. But attempts at protecting privacy through regulation on both national and international levels face substantial difficulties which relate to both technical and political causes.

Introduction

The Internet has become an indispensable feature of everyday life in the first decade of the third millennium. No longer confined in use to experts but user-friendly enough to require little training, the Internet is now available not only to citizens in developed countries (where around 80% now have access), but will be to almost 3 billion people around the world by the end of 2014 (40% globally; 32% in developing countries). The contribution to economic growth and dynamism undoubtedly counts among the beneficial consequences of this development. But worries about the negative effects that may accompany this have also emerged, for in a world where users are continuously 'online,' privacy can be endangered if their every move and request can be traced, stored, searched, and potentially forever replicated. Threats to privacy can come from various directions: well- and ill-intentioned governments, eager to learn about their citizens' lives, or focused on protecting them from terrorist threats count among them, but so do commercial enterprises interested in turning personal data into revenue streams or wanting to maximize service to their customers. Whether firms like *Facebook* or *Google* – whose customers number in the hundreds of millions – use their accumulated knowledge in the right way has thus become a subject of debate in recent years. Most discussion about threats to privacy today has to do with the connectedness and linkage of personal or person-related data through the Internet. Both public and private sector actors have responded to these worries with attempts at (self-)regulation, on the national and on the international level. However, much of this field remains contested and little has been achieved in terms of encompassing goals and standards for such regulation. Since June 2013, privacy has also come on the political agenda in many countries after revelations about systematic snooping by security agencies such as the National Security Agency (NSA) in the United States, Government Communications Headquarters (GCHQ) in the United Kingdom, and Bundesnachrichtendienst (BND) in Germany emerged.

Privacy and Technology

While privacy is a basic human need and is constitutive for the individual, it remains difficult to describe and define. Across societies, the desire and need for privacy has varied historically (Moore, 1984), and in the process of modernization, demand for privacy has tended to grow in order to balance societal intrusions resulting from developments such as greater population density. In the theoretical debate, however, no generally accepted definition has so far emerged. More than 45 years ago, Alan Westin (1967, p. 7) wrote in his seminal study on the subject: "Few values so fundamental to society as privacy have been left so undefined in social theory or have been the subject of such vague and confused writing by social scientists." A recent comprehensive work in the area makes basically the same point when it states: "Privacy, however, is a concept in disarray. Nobody can articulate what it means." (Solove, 2008, p. 1). The main reason for this lies in the fact that privacy is defined negatively as an absence of intrusion; such intrusions, however, can come from many directions and vary in type and subjectively perceived intensity, hence leading to different priorities for their avoidance.

Discussions about privacy and its intrusions have almost always been linked to technological developments. Already in the late nineteenth century, the emergence of 'instantaneous photographs' was seen as invading privacy, prompting calls for a right "to be let alone" (Warren/Brandeis, 1890). In the 1920s, the tapping of telephone lines for purposes of law enforcement led to debates resulting in stringent regulation and reporting requirements for such actions. In the 1960s and 1970s, plans for comprehensive 'data banks' and 'national data centers' containing electronic records of every citizen prompted debates about the relative merits of optimizing the state's social planning versus protecting privacy (Miller, 1971) and stopped most of these developments. From the 1990s onward, the proliferation of CCTV systems led to discussions about a 'surveillance society,' in which there was "no place to hide" (O'Harrow, 2006), as every move could be watched and recorded. After the terrorist attacks in the United States on 11 September 2001, plans emerged to put technology to the service of antiterrorism, but technocratic dreams of 'Total Information Awareness' through 'data mining' of all available public and private sector information sources turned out to be neither feasible nor desirable (National Research Council, 2008).

In the context of technology, privacy almost always means informational privacy and thus focuses on a specific part of the debate about privacy (Rössler, 2005). Analyzing the relationship between the Internet and privacy thus requires looking at the competing interests over person-related data as well as at the

attempts to set rules about their creation, proliferation, storage, and use.

Internet Use and Personal Data

Development of Internet Usage

Usage of the Internet has made the transition from expert domain to everyday use in the last two decades. Originally developed by military and academic experts to provide a safe means of communications even in the event of a nuclear war, and used first to link up leading research centers, the 'Internet' is actually a system of various connected computer networks that function by adhering to common technical protocols. Hiding the underlying complexity from the user and making a vast network of billions of nodes easily maneuverable through naming systems and graphical interfaces has been a condition for its exploding usage as well as for its commercial success, with the latter providing the funds for the Internet's ever-increasing capacity.

Internet use has increased dramatically since the turn of the century. In 2000, the percentage of individuals using the Internet was between 20% and 40% in developed countries (France: 14%; UK: 27%; Germany and Japan: 30%; US: 43%), and the global number was at a mere 6.5%. By 2013, numbers had shot up to around 85% in the first group (US: 84%; France: 82%; Japan: 86%; UK: 90%; Germany: 84%), while the global number had increased sixfold to 40% ('ITU database' http://www.itu.int/). This massive growth was unexpected even by experts who did not foresee the explosive development of mobile Internet access at broadband speeds which helped in particular to connect citizens of developing countries to the Internet in only a few years. As a result, an estimated 2.7 billion people around the globe had access to and used the Internet in 2013 (International Telecommunication Union, 2013, p. 1).

While these impressive growth figures have contributed substantially to development goals in developing countries (where they have, e.g., helped to introduce mobile banking and to spread public health and education information), it is in the developed countries that commercial activity initiated by and through the Internet is strongest. Electronic commerce has been a growth engine in many developed economies, contributing further to a shift toward a knowledge and information-based economy that has been under way at least since the 1960s (Machlup, 1962). But the Internet has also triggered the emergence and meteoric rise of new corporate stars such as Google (search engine) and Facebook (social media). These firms (which are used here as illustration for many others that are similar) build their business models on the use and processing of person-related data, for they charge no fees for their services from their customers. Instead, customers pay indirectly with their personal data which firms – through sophisticated mechanisms – monitor, store, and then use to create profiles. These profiles in turn allow them to sell advertisements which are highly targeted to the user's sociodemographic characteristics, thus increasing their success rate and extracting premium prices from advertisers.

How much successful firms like Google and Facebook depend on this business model is evident from business data submitted in companies' quarterly reports to the US Securities and Exchange Commission (SEC): in the first half of 2012, advertising revenues made up 96% of Google's overall revenues; and Facebook (which reported 955 million 'monthly active users' and 552 million 'daily active users') earned 84% of its total revenue from advertising ('SEC,' http://www.sec.gov/). In short, these firms would not be commercially viable without advertising revenues from personal data.

User Demand for Internet Privacy

In their corporate mission statements firms like Google and Facebook neither mention the commercial use they make of person-related data nor their dependency on them. They describe their mission as, respectively, "to organize the world's information and make it universally accessible and useful" (Google) and "to make the world more open and connected" (Facebook).

Being more open about their actual business models might cause worries among the hundreds of millions of customers who use their services on a daily basis. Google, for example, answers between four and five billion search questions per day; it logs and keeps all search queries as well as users' IP addresses, in addition to 'cookies' (small data files that track a user's Web browsing), and (where applicable) location information ('Google,' https://www.google.com/policies/privacy/). Thus the firm obtains a lot of person-related data which it may combine with information that users turn over voluntarily when registering for a user account with the firm, such as name, e-mail address as well as telephone and credit card numbers or device identifiers. Given that the company's market share is 95% of the search engine market in some European countries, this means that users may have few alternatives to using these services.

According to surveys, users value their privacy highly and are concerned about the protection of their personal data on the Internet. General worries about privacy and data protection predate the advent of the Internet, but the availability of online electronic commerce means most consumers were only confronted concretely with the issue of being asked to enter personal data online in the late 1990s. In an internationally comparative survey commissioned by IBM, 80% of respondents in the US, 68% in the UK, and 79% in Germany agreed with the statement "Consumers have lost all control over how personal information is collected and used by companies" (Louis Harris & Associates, 1999, p. 22). Data from the Eurobarometer survey show that the level of 'concern' about data privacy has changed little since the early 1990s, with 68% expressing it in 2008 and 66% in 1991 – the worries are thus not primarily triggered by the Internet. Underlying this stable trend, however, are substantial differences in attitudes between countries: Austria and Germany top the field with a rising trend and 86% expressing concern, while at the bottom of the list are Finland and the Netherlands with 35% and 32% respectively and (in the latter case) a declining trend. With respect to the Internet specifically, worries tend to be more pronounced: 82% of EU respondents thought that data transmission over the Internet was not sufficiently secure, with some variation across countries but little across sociodemographic differences (Gallup Europe/European Commission, 2008). On the other hand, the disclosure of personal information is seen by many

as a part of modern life. They do disclose such information for the purposes of online services and online shopping, but 70% of Europeans are concerned that their personal data held by companies may be used for a purpose other than for which it was collected (European Commission/TNS Opinion and Social, 2011).

In sum, a strong preference for privacy, and worries about safety of personal data submitted online do not prevent citizens from using the Internet for online activity, but survey evidence indicates that they use it in a guarded way.

Problems of Crime, Security, and Surveillance on the Internet

The Internet was originally designed without central control in order to maximize the autonomy of heterogeneous networks working together. This not only conformed with military priorities of resilience in the case of a nuclear attack, it was also appropriate for the relatively small academic community in which users trusted each other; in addition it resonated with the libertarian impulses of the founding generation of engineers many of whom had links to the counterculture (Markoff, 2006). The result has been interpreted as a 'techno-utopian' attempt at self-regulation (Hofmann, 2012), a space almost completely free from formal boundaries and rules, in which the eventual emergence of problems with cybercrime is primarily an unforeseen consequence of past design decisions.

Cybercrime and Cybersecurity

What is now labeled 'cybercrime' has been emerging from the mid-1990s onward – when the originally wholly state-funded backbones were privatized and commercial use of the Internet took off – and taken various forms. The sending of unsolicited mass messages to initially many thousands, later millions of users advertising goods and services ('spam') was a first (and more annoying than dangerous) way of using the Internet for unintended activities. Initially done through e-mail (making use of the open nature of Internet e-mail servers), this activity was more recently primarily conducted through 'botnets' of PCs that had been taken over through the installation of malware. Such malware (viruses and 'trojans') can also be used to obtain passwords by logging and sending out users' keystrokes or to scan hard drives for files containing confidential information. Individual servers on the Internet, but also whole networks of firms or governments can be attacked through flooding them with IP packets (often from botnets), thus effectively shutting them down ('distributed denial of service'); and such attacks can be combined with attempts at extortion. Other forms of cybercrime include the sending of e-mail which claims to be from a user's bank or business partner, but which contain links redirecting responses to forged Web sites ('phishing') and the systematic use of personal data to impersonate other people for fraudulent purposes ('identity theft').

It is very difficult to obtain reliable information about the scale of the problem of cybercrime. Figures about its costs often make headlines – the British *Financial Services Authority*, e.g., estimated the cost of identity fraud to the UK economy at £1.7 billion per annum. This estimate, however, includes substantial amounts for lost or stolen credit cards, VAT losses, and even costs to state agencies for antifraud procedures. Furthermore, crime statistics do not routinely count whether IT was instrumental in committing a crime. Without more precisely targeted attempts to identify crimes committed online, a UK public inquiry concluded, "it is simply impossible to assess the scale of the problem" (House of Lords. Select Committee on the Constitution, 2009, vol. 1, pp. 15–16). As regards the inflated headline figures, it is clear that both in the public sector (law enforcement) and in the private sector (IT security firms) some actors clearly have an interest in pushing publicity of the issue.

Internet users, as far as survey evidence shows, have taken the dangers to heart and modified their behavior accordingly by trying to protect their privacy. 40% of Internet users in the EU are concerned about someone stealing or misusing their personal data; and 38% worry about the security of their online payments. Many of them have therefore changed their online behavior, and most of them try to avoid disclosing personal information online. When using the Internet for online banking or shopping, the two most common concerns are about someone taking or misusing personal data (mentioned by 40% of Internet users in the EU) and security of online payments (38%) (European Commission/TNS Opinion & Social, 2012, p. 4).

Both private and public sector actors have reacted to the threat and to the worries of users. IT security firms developed sophisticated tools to protect their customers' PCs from attacks through viruses and other malware and have built a substantial business around this; states have initiated international agreements (like the 2001 *Convention on Cybercrime* by the Council of Europe which contains common definitions and procedures on the subject as well as rules for international cooperation), passed national legislation, and set up competence centers to develop appropriate defenses.

Surveillance on the Internet

Security firms and state agencies thus work to protect users' privacy against illegitimate and foreign intrusion. But in order to do that, they collect lots of personal data which means we can look at the very same actors also with reference to the debate about surveillance. That debate focuses on how users on the Internet can be protected from the actions of legitimate firms and government agencies, and has been the subject of considerable social activism (Bennett, 2008). Firms' commercial interests were already mentioned above; government agencies, in turn, have an interest in using the Internet for the collection of person-related data for purposes of state and policy administration, social planning, law enforcement, national security, etc.

The concept of surveillance has been prominent in the sociological literature on personal and person-related information in recent years and has given rise to the academic field of *surveillance studies* (for an overview see Lyon, 2007; Ball et al., 2012). In this debate, the term surveillance has been defined very broadly as "the focused, systematic and routine attention to personal details for purposes of influence, management, protection or direction" (Lyon, 2007, p. 14), and as a consequence, much has been described and discussed in this literature. Most of this is not or only indirectly linked to the Internet,

such as CCTV cameras in public space, workplace surveillance practices of employers, DNA databases for policing or the tagging of consumption patterns through radio frequency identification chips or 'smart meters.' With the growing use and commercialization of the Internet from the mid-1990s onward, surveillance has also taken place through the Internet (e.g., through easy availability of background checks, and the massive collection of data to monitor and profile Internet users). Increasing transparency of individuals' actions is the result of these developments – often without their awareness. However, literature from the *surveillance studies* tradition has been criticized for being somewhat one-sided and pessimistic in its assessment of surveillance on the Internet and furthermore poor on empirical and technical detail (Bennett/Parsons, 2013).

While surveillance from the private sector is (at least in principle) subject to attempts at regulation on the national and supranational level, even if their extent is contested (see below), this is much less the case for surveillance from the public sector. As studies conducted by the OpenNet Initiative have demonstrated, Internet censorship and surveillance are becoming more technically sophisticated, and both are increasingly prevalent not only in authoritarian regimes, but also in liberal democracies (Deibert et al., 2010). As a result, the initially almost universally optimistic assessment of the emancipatory societal and political consequences of further spread of Internet use – prominently visible, e.g., in the role played by social media during the 'Arab Spring' of 2011 – has been called into question. Rather than bringing empowerment to the individual, fostering transparency and democracy, and undermining authoritarian rule, it has been pointed out, the surveillance potential of Internet technology can also be used by dictatorships to spy on dissidents, uncover their identities, and censor access to information (Morozov, 2011). The revelations of routine and widespread snooping by national security agencies that surfaced in the summer of 2013 and are discussed in a separate section below have contributed to increased sensitivity in this area in many countries.

Regulating Privacy on the Internet

Problems to protect privacy against intrusions are in principle amenable to solutions through regulation. Like in most other policy domains, attempts to impose such regulations meet resistance from interested parties and political forces opposed to regulation. But there are two domain-specific aspects that further complicate regulation of privacy on the Internet: on the one hand, due to its purposefully decentralized technical nature, there is not one single point from which the whole of the Internet can be affected by regulation; and on the other hand, attempts at regulation only started after the privatization of the Internet backbones and the opening up of the Internet for commercial use, i.e., in the mid-1990s. At that point in time, however, the Internet had already undergone a long period without any regulation, while privacy regulation had been under way for some time, both on the national and international level, without taking into account the specific problems the Internet created in that area. Attempts at regulation thus continue to take place in a highly fragmented situation, with a multitude of actors (many of whom focus not only on regulation on the Internet) and strongly diverging interests. Outcomes are therefore often characterized by conflict and blockade.

National Level Regulation

Due to the technical aspects of the Internet and the sequence of historical developments mentioned above, regulation of privacy on the national level is not primarily focused on the Internet. Privacy became a topic of political regulation in many Western democracies in the 1970s, when countries (starting with Sweden and West Germany) began to discuss and pass comprehensive privacy legislation and set up institutions charged with protecting personal information, often labeled 'privacy commissioners' or 'data protection agencies.' While reacting to common problems, national solutions varied considerably due to varying political preferences, administrative traditions, and institutional interests; but processes of diffusion and learning spread privacy legislation across the globe in the decades since (Newman, 2008). The United States, however, followed a different trajectory with a focus on performance-based regulation centering on 'fair information principles' largely developed by the administration, and supplemented by area-specific legislation instead of comprehensive legislation.

National regulators and privacy commissioners have little effective influence on the whole Internet. Their work thus mainly focuses on national level implementation of rules and questions that are amenable to national regulation. Under the heading 'Internet,' a recent report (published in 2011) of the German Data Protection Commissioner, e.g., discussed such topics as the collection of location data, WLAN access, Google Street View, and the setting up of a 'Common Internet Center' by German security authorities to fight terrorism. Such reports can influence national political debates, serve as an antagonist to security agency and law enforcement interests in the political process, and have educational effects on the general public which uses the Internet, but it is unlikely to achieve much beyond national borders. Borrowing a term from environmental policy, it can thus primarily be described as 'end-of-pipe regulation.'

International Level Regulation

Potentially more effective regulation can be achieved on the supranational and international level. Indeed, looking at privacy regulation, one can say that the international level has been leading the way here, with bodies such as the Council of Europe and the OECD being early initiators of conventions and guidelines from the 1970s onward. In 1985, the OECD first adopted a 'Declaration on Transborder Data Flows' which pledged support for international exchange of data and information and (given its focus on economic policy) framed obstacles to it as protectionist; in 1998, a 'Declaration on the Protection of Privacy on Global Networks' followed which confirmed the importance of effective privacy protection also for the future success of e-commerce. In 1995, the European Union passed its 'Directive on the Protection of Personal Data' in order to harmonize regulation across its then 15 member

states and restrict the transfer of personal data to third countries without adequate protection of personal data. Since EU directives are legally binding to member states, this had more direct impact than the previous documents which largely had the status of recommendations.

However, none of these supranational and international organizations are in a position to exercise direct regulatory power over the Internet. Indeed, who controls the Internet is both unclear and contested (Goldsmith/Wu, 2006); and given the principal decisions in favor of technological openness that lie at the heart of the Internet (see above), the fact that a single set of standards cannot be enforced universally on the whole Internet will not change. Furthermore, conceptions and priorities regarding privacy vary on the national level: European countries perceive privacy as a fundamental human right which is a precondition for an individual's autonomy; it can therefore not be given up individually, and society has a duty to protect it. In the United States, in contrast, privacy is seen as a property right that an individual can trade away if he or she chooses to, and markets are thus seen as a more effective way of protecting it than state regulation (Busch, 2011).

With divergent preferences between Europe and the US, no international regime governing privacy on the Internet could be established; instead, rival standards are the most likely outcome in such a situation, and that has indeed been the case (Drezner, 2007). As a consequence, conflicts have arisen over data protection on the Internet between the EU and the US in the past. While a compromise could be reached in the economically important area of e-commerce with the 'Safe Harbor Agreement' of 2000, outcomes have been more conflictuous and characterized by unilateral imposition of US preferences in areas related to security and antiterrorism policy, such as the questions of transferring personal data of aviation passengers (PNR) or international financial transactions data (SWIFT) (Busch, 2013).

The Internet and National Security

Information technology has been playing an ever more important role in security policy over the last decades due to the massive improvements in information and communications technology (ICT). This development was further speeded up significantly after the terrorist attacks of 11 September 2001, leading to a transformation of state activities which saw a blurring of the boundaries between foreign and domestic security policy, the blending of police, secret service, and military tasks, and a central role for ICT as a national security state tool (Busch, 2015).

The extent to which national security agencies have interfered with communication on the Internet in order to obtain information became clear after a whistleblower, the former NSA contractor Edward Snowden, published classified information from the summer of 2013 onward. It revealed comprehensive operations by intelligence agencies such as the US NSA, the UK's GCHQ, the German BND, and various others more to intercept information from the Internet and the telephone system through both indiscriminate bulk data collection as well as targeted spying under programs with names like PRISM, Quantum, Tempora, XKeyscore, or MUSCULAR. The material also made public a previously unknown level of cooperation and data sharing between these agencies which could also serve to circumvent constitutional barriers to their operations, such as a ban on spying against their own country's citizens. Apparently the NSA also had direct access to the servers of technology firms like Google, Facebook, Microsoft, and Apple (Harding, 2014).

The revelations ruffled feathers diplomatically as it became clear that the US had not only collected data from close allies such as Germany and France and bugged a fax machine of the European Commission, but had even targeted Chancellor Merkel's personal cell phone. As these transgressions were criticized in political discourse, they made clear that a number of state security agencies seemed to have overstepped their brief. A principal-agent problem had evidently emerged in the oversight of security services, and legislatures in some countries moved to tighten the reins. In the United States, for example, President Obama appointed a *Review Group on Intelligence and Communications Technologies* that reported in December 2013. Besides its 46 concrete recommendations the report emphasizes the need to "protect, at once, two different forms of security: national security and personal privacy" (President's Review Group, 2014, p. xv). It remains to be seen to what extent the Review Group's recommendations will result in changes to US policy. In any case a continuation of past practices could severely harm US business interests if customers from other countries decided not to entrust their electronic data to servers located in the US any longer.

Internet and Privacy: Outlook

Growing use of the Internet is only the latest in a string of technological developments that has imperiled personal privacy. While opinion polls show a strong preference of Internet users to protect their personal data, a number of interests combine to make use of such data, mainly for the purposes of economic growth, commercial profit, and the protection of individual and national security. While in principle regulation could help strike a balance between these competing interests, various factors have been working against this. Thus regulation of the Internet is principally difficult due to its decentralized nature on the technical level, its global connectedness and spread across many jurisdictions, and private ownership of large parts of its infrastructure. In addition, the view that states could not regulate the Internet (and that states are generally on the way out as globalization spreads) dominated for a long time, resulting in few attempts at regulation. When – parallel to the growing importance of the Internet – state attempts at regulation increased, they encountered collective action problems emanating from different preferences and regulatory path dependencies.

Whether competition between rival regulatory standards on the international level, mainly those of the US and the EU, will lead to the emergence of common solutions remains to be seen. While such regulation can affect privacy interests of Internet users, the resistance against that may in the future be less stringent than it was in the past, for attitudes in this area have changed, as a leading advocate in that field has stated: "When I first wrote the book, two ideas seemed to

dominate debate about the Net: first, that the government could never regulate the Net, and second, that this was a good thing. Today, attitudes are different. There is still the commonplace that government can't regulate, but in a world drowning in spam, computer viruses, identity theft, copyright 'piracy', and the sexual exploitation of children, the resolve against regulation has weakened. We all love the Net. But if some government could really deliver on the promise to erase all the bads of this space, most of us would gladly sign up" (Lessig, 2006, p. 27).

In principle, a third approach to the problem of protecting privacy on the Internet – besides national and international level regulation – is available, namely a technical and organizational solution applied at the level of the information system ('privacy by design,' 'privacy enhancing technologies'). While this has both been discussed and encouraged (e.g., by information commissioners) for some time, implementation has been reluctant as substantial interests exist that push, e.g., against data minimization. Whether that balance of interests will shift as the number of Internet users rises further and discussions about the "virtue of forgetting in the digital age" (Mayer-Schönberger, 2009) become more prominent, remains open to speculation. The debate following the European Court of Justice ruling of April 2014 about the "right to be forgotten" (C131/12, *Google Spain SL, Google Inc. v Agencia Española de Protección de Datos*) has demonstrated the difficulty of balancing central values such as free speech and the protection of privacy. But above all it has been known for some time that who prevails in collective action is not primarily determined by the sheer number of people involved on both sides or the quality of the argument, but rather by their relative incentives and capabilities to organize and gain influence on the political decision-making process (Olson, 1965).

Botnet	A group of computers that are connected to the Internet and under the control of a malicious party, typically after malware (such as a virus program) has been executed under legitimate use.
Data mining	A process of attempting to discover meaningful patterns in (often very large) sets of already collected data without prior forming of hypotheses.
Privacy enhancing technologies (PET)	Technological, procedural, or organizational tools to enhance user control over their personal data.
Radio frequency identification (RFID)	A wireless technology that uses electromagnetic fields to transfer data, mostly without battery, from a tag attached to an object to identify its properties.
Total information awareness (TIA)	Program run by the US Department of Defense after 11 September 2001 to integrate information from diverse sources. After allegations of privacy violations funding was withdrawn in 2004.
Ubiquitous computing	A paradigm of computer use in which many, often mobile and networked computers replace the one traditional desktop computer

See also: Cyber Crime: Identity Theft; Cyberbullying; Internet and Culture; Internet and Social Media: Anthropological Aspects; Oversharing: The Eclipse of Privacy in the Internet Age; Privacy: Theoretical and Legal Issues; Surveillance Studies; Surveillance and Privacy, Geography of.

Bibliography

Ball, Kirstie, Haggerty, Kevin, Lyon, David (Eds.), 2012. Routledge Handbook of Surveillance Studies. Routledge, London.

Bennett, Colin J., 2008. The Privacy Advocates. Resisting the Spread of Surveillance. MIT Press, Cambridge MA.

Bennett, Colin J., Parsons, Christopher, 2013. Privacy and surveillance: the multidisciplinary literature on the capture, use, and disclosure of personal information in cyberspace. In: Dutton, William H. (Ed.), The Oxford Handbook Internet Studies. Oxford Univ. Press, Oxford, pp. 486–508.

Busch, Andreas, 2011. The regulation of privacy. In: Levi-Faur, David (Ed.), Handbook on the Politics of Regulation. Edward Elgar, Cheltenham UK, Northampton MA, pp. 227–240.

Busch, Andreas, 2013. The regulation of transborder data traffic: disputes across the Atlantic. Security and Human Rights 24, 119–138.

Busch, Andreas, 2015. The changing architecture of the National Security State. In: Leibfried, Stephan, Huber, Evelyne, Lange, Matthew, Levy, Jonah, D., Nullmeier, Frank, Stephens, John (Eds.), The Oxford Handbook of Transformations of the State. Oxford University Press, Oxford, pp. 536–553.

Deibert, Ronald, Palfrey, John, Rohozinski, Rafal, Zittrain, Jonathan (Eds.), 2010. Access Controlled. The Shaping of Power, Rights, and Rule in Cyberspace. MIT Press, Cambridge Mass.

Drezner, Daniel W., 2007. All Politics Is Global. Explaining International Regulatory Regimes. Princeton Univ. Press, Princeton.

European Commission and TNS Opinion & Social, 2011. Attitudes on Data Protection and Electronic Identity in the European Union. Special Eurobarometer 359. Brussels.

European Commission and TNS Opinion & Social, 2012. Cyber Security. Special Eurobarometer 390. Brussels.

Gallup Europe and European Commission, 2008. Data Protection in the European Union: Citizens' Perceptions. Analytical Report. Brussels.

Goldsmith, Jack L., Wu, Tim, 2006. Who Controls the Internet? Illusions of a Borderless World. Oxford University Press, Oxford.

Harding, Luke, 2014. The Snowden Files. The Inside Story of the World's Most Wanted Man. Guardian Books, London.

Hofmann, Jeanette, 2012. Et in Arcadia Ego: from Techno-Utopia to cybercrime. In: Margetts, Helen, 6, Perri, Hood, Christopher (Eds.), Paradoxes of Modernization. Unintended Consequences of Public Policy Reform. Oxford Univ. Press, Oxford, pp. 81–100.

House of Lords. Select Committee on the Constitution (Ed.), 2009. Surveillance: Citizens and the State. 2nd Report of Session 2008–09. The Stationery Office, London.

International Telecommunication Union, 2013. Measuring the Information Society 2013. International Telecommunication Union, Geneva.

Lessig, Lawrence, 2006. Code: Version 2.0. Basic Books, New York.

Louis Harris & Associates, Inc, 1999. IBM Multi-national Consumer Privacy Survey. A Comprehensive and Comparative Look at Consumers in the United States, Germany, and United Kingdom and Their Attitudes toward Privacy in Everyday Business Transactions. IBM Global Services, New York.

Lyon, David, 2007. Surveillance Studies. An Overview. Polity, Cambridge.

Machlup, Fritz, 1962. The Production and Distribution of Knowledge in the United States. Princeton Univ. Press, Princeton, NJ.

Markoff, John, 2006. What the Dormouse Said. How the Sixties Counterculture Shaped the Personal Computer Industry. Penguin, New York.

Mayer-Schönberger, Viktor, 2009. Delete. The Virtue of Forgetting in the Digital Age. Princeton Univ. Press, Princeton, NJ.

Miller, Arthur R., 1971. The Assault on Privacy. Computers, Data Banks, and Dossiers. University of Michigan Press, Ann Arbor.

Moore Jr., Barrington, 1984. Privacy. Studies in Social and Cultural History. M.E. Sharpe, Armonk, N.Y.; London.

Morozov, Evgeny, 2011. The Net Delusion. The Dark Side of Internet Freedom. Public Affairs, New York.

National Research Council (Ed.), 2008. Protecting Individual Privacy in the Struggle against Terrorists. A Framework for Program Assessment. National Academies Press, Washington DC.
Newman, Abraham L., 2008. Protectors of Privacy. Regulating Personal Data in the Global Economy. Cornell University Press, Ithaca ua.
O'Harrow, Robert, 2006. No Place to Hide. Behind the Scenes of of Our Emerging Surveillance Society. Free Press, New York, NY.
Olson, Mancur, 1965. The Logic of Collective Action. Public Goods and the Theory of Groups. Harvard University Press, Cambridge, Mass.
President's Review Group on Intelligence and Communications Technologies, 2014. The NSA Report. Liberty and Security in a Changing World. Princeton University Press, Princeton (NJ), Oxford.
Rössler, Beate, 2005. The Value of Privacy. Polity, Cambridge (UK), Malden (MA).
Solove, Daniel J., 2008. Understanding Privacy. Harvard University Press, Cambridge, Mass.
Sommer, Peter, Brown, Ian, 2011. Reducing Systemic Cybersecurity Risk. OECD/IFP Project on "Future Global Shocks". 14th January 2011. Paris.
Warren, Samuel D., Brandeis, Louis D., 1890. The right to privacy. Harvard Law Review IV, 193–220.
Westin, Alan F., 1967. Privacy and Freedom. Atheneum, New York.

Relevant Websites

https://www.google.com/policies/privacy/ – Google.
http://www.itu.int/ITU-D/ict/statistics/explorer/index.html – ITU database.
http://www.sec.gov/ – U.S. Securities and Exchange Commission.

Internet and Social Media: Anthropological Aspects

Paula Uimonen, Stockholm University, Stockholm, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Generally defined as a network of networks, the Internet has had a profound impact on the social organization and cultural meaning of modern society. Since it entered the public domain in the early 1990s, the Internet has grown exponentially and is now used by one-third of the world population. Anthropologists have studied the Internet from its early social history, especially in non-Western countries. Over time, this research has evolved into the subdiscipline of digital anthropology, which studies the development and use of digital media in different cultural contexts.

Cyberculture and Digital Anthropology

Generally defined as a network of networks, the Internet has had a profound impact on the social organization and cultural meaning of modern society. It is a media technology that has grown at an exponential rate: from being used by approximately 2% of the world population in 1997 to 35% or 2.45 billion people in 2011 (estimates by International Telecommunication Union). The growth in social media use, also known as Web 2.0 technologies, has been equally staggering. Since it was launched to the public in 2006, Facebook grew from 100 million users in 2008 to 1 billion in 2012. Meanwhile, Wikipedia, Twitter, and YouTube have become commonplace sources of information, communication, and entertainment, while new touchscreen interfaces have facilitated the interactive use of digital media, across generations.

The Internet was originally developed by and for researchers, to enable them to share costly computer resources. The first manifestation of the Internet was the Advanced Research Projects Agency Network (ARPANET) in the United States. From the outset, the Internet built on the idea of 'open-architecture networking' or what came to be known as Transfer Control Protocol/Internet Protocol, allowing interconnections or 'internetting' between different computers and computer networks (see *Brief History of the Internet* published online by the Internet Society). From the late 1960s and onward, the Internet grew steadily in research communities around the world, first in the United States, but as of late 1980s in other parts of the world as well. It was the usefulness of being able to share computer resources combined with the popularity of e-mail, which was introduced in 1971 that spread the Internet around the world. The Internet was originally funded by the US Department of Defense, but its purpose was to strengthen research. It was only in the early to mid-1990s that the Internet became public and commercial. As the Internet entered the public domain, it was suggested that it was originally built by the military to withstand nuclear war. While this interpretation of the origins of the Internet have been vehemently denied by the scientists who were involved in its design, it offers an interesting 'origin myth' of what was initially known to the public as an unregulated, even anarchistic, global media technology, the development of which was inspired by the countercultures of the 1960s (Castells, 1996; Uimonen, 2001).

Already in 1994, Arturo Escobar encouraged anthropologists to engage in a field of inquiry that he referred to as the 'anthropology of cyberculture. Underlining that all technologies are cultural interventions and that science and technology constitute 'crucial arenas for the creation of culture in today's world, Escobar (1994: p. 211) encouraged anthropologists to 'venture into this world in order to renew their interest in the understanding and politics of cultural change and cultural diversity. His reading of cyberculture did not necessarily refer to the Internet, which was mentioned in a footnote, but focused on computer, information, and biological technologies, primarily from the perspective of science and technology studies (STS). Even so, his interest in cyberculture was clearly inspired by emerging work on cyberspace, a term he used more widely than Internet in the paper. At the time, some anthropologists were also inspired by the concept cyborgs, which was introduced by feminist scholar Donna Haraway (1991) to describe fabricated hybrids of machine and organism. Sociologist Sherry Turkle (1995) was another influential scholar, and her work on human–computer interaction inspired many anthropologists for years to come (Garsten and Wulff, 2003). In the 1990s, some anthropologists referred to their Internet research in terms of cyberanthropology, but the term fell out of use. Alternative ways of classifying Internet research included digital ethnography, Internet ethnography, virtual ethnography, virtual anthropology, and eventually digital anthropology, which has evolved into the most common way of describing this research field.

In August 2010, the European Association of Social Anthropologists (EASA) organized a workshop in Digital Anthropology at the biennial EASA conference in Maynooth, Ireland, a milestone in the evolution of this subdiscipline. The first of its kind at EASA, the workshop was convened by Daniel Miller and Heather Horst, two of the leading researchers in this field, with presentations by senior and junior Internet anthropologists. In 2012, a Center for Digital Anthropology was established at the University College of London (UCL), directed by Daniel Miller. Over the years, digital anthropology has also been taught at different universities in Europe and United States. In August 2008, the Department of Social Anthropology at Stockholm University initiated a graduate course in digital anthropology led by Paula Uimonen, in 2009 the UCL launched an MSc program in digital anthropology

directed by Daniel Miller, and in 2012 the first textbook in digital anthropology was published (Horst and Miller, 2012). In the United States, Michael Wesch at Kansas State University has been leading a Digital Ethnography program, Christopher Kelty has been teaching courses about the Internet at University of California, Los Angeles, and Tom Boellstorff has been teaching courses in virtual anthropology at University of California, Irvine, to name some contemporary digital anthropologists.

Digital anthropology is closely related to media and transnational anthropology, as well as STS and visual anthropology. Topics related to digital anthropology are often discussed in the *Media-Anthro* mailing list of EASA's media anthropology network as well as *Savage Minds*, the most popular blog for anthropologists. There is also a digital anthropology group in the *Open Anthropology Cooperative* (OAC), a social networking site where many disciplinary specializations in anthropology coalesce. The OAC was initiated in 2009 by Keith Hart, one of the first anthropologists to write about Internet, social transformation, and world society (2000).

Research topics in digital anthropology combine distinguishing features of digital media with typical areas of interest for anthropologists, such as cultural identity, power structures, and social networks. Many researchers in digital anthropology are experienced in other research areas, bringing broader insights into their digital work. For example, Daniel Miller is well known for his work on material culture and consumption, while Tom Boellstorff has done a lot of work on globalization and sexuality. As younger scholars now turn to digital anthropology there will probably be a greater specialization in this research field, but hopefully future generations of scholars will also broaden their research interest to other anthropological domains. After all, digital anthropology is no different from other subdisciplines within anthropology in that it builds upon as well as contributes to developments in mainstream anthropology. Meanwhile, a growing number of anthropologists are paying attention to digital media in their studies in other domains, not least in the fields of transnational migration, diaspora, material culture, popular culture, and social movements. This body of research is also contributing to anthropological analyses of digital media. For an overview, see Coleman (2010).

According to a recent definition, 'digital anthropology focuses on the development and use of digital media and communication technologies in different social and cultural contexts' (Uimonen, 2012: p. 36). It is still an emerging subdiscipline, which does not have a theoretical canon, but draws upon the accumulated body of anthropological knowledge in general, and related research fields or subdisciplines within anthropology in particular. Digital anthropologists carry out fieldwork online as well as offline, often combining traditional ethnographic research methods with more experimental methods such as multisited and online ethnography (Kelty, 2008), ethnographic fieldwork in virtual worlds (Boellstorff et al., 2012) or participatory digital video (Uimonen, 2012). Like many other research fields in anthropology, digital anthropology originates from scholars based in Europe and the United States, but their fieldwork is often carried out in non-Western contexts, thus focusing on cultures that tend to be underinvestigated by digital media scholars from other disciplines. With time, the research domain will hopefully be more globalized, attracting anthropologists from all parts of the world, especially Africa, Asia, and Latin America.

Hyperbole and Social Change

Ever since it entered the public domain, the Internet has been cast in hyperbolic predictions of radical social change. Embedded in discourses on the information society or knowledge society, the Internet has been thought to usher in a new era in human development, the information age. Utopian champions have envisioned new premises for democracy and openness, while dystopian critics have warned of increased surveillance and government control. Similarly, optimistic predictions of enhanced individual freedoms have been countered with gloomy reports of the online circulation of legally prohibited and morally offensive content. Some of the most influential predictions have related to changes in the economic system, the virtual economy offering new premises for growth and value creation in global capitalism. These utopian forecasts have even influenced stock markets as exemplified in the 'dot com' boom, when newly established Internet companies of various kinds were estimated to be worth considerable amounts of money in their initial public offerings. The ensuing dot com crash revealed the false premises of many of these speculative estimates, although some of the dot coms managed to survive in the long term. It was perhaps no coincidence that the dot com bubble took place around 1997–2000, concomitantly with the millennium shift, utopian expectations of a new virtual economy thus coinciding with an epochal shift in human history. At the time, dystopian predictions of disaster were also quite influential, as exemplified by the Y2K (year 2000) debate. As the calendar shifted to year 2000, technical date standards were considered to be problematic and unless remedied, societies could face technical failures of unknown proportions. The Y2K problem caught the attention of governments and even the United Nations, and substantial resources were invested into fixing the 'millennium bug. To date, it remains unclear whether the Y2K problem was exaggerated or real, but billions of US dollars were spent fixing it.

Historical changes coinciding with the millennium shift have also been anticipated by social scientists, the most influential one being the sociologist Manuel Castells and his theories of the 'network society. In the years leading up to the millennium shift, Castells published his trilogy *The Information Age: Economy, Society and Culture*, comprising the volumes *The Rise of the Network Society* (1996), *The Power of Identity* (1997), and *End of Millennium* (1998). Leading sociologist Anthony Giddens referred to Castells as the 'new Weber, thus equaling his analysis of the information age with Max Weber's theories of the industrial age. A basic tenet in Castells' work is the notion that networks have become the dominant social structure in the age of informational capitalism, characterized by 'space of flows, 'timeless time, and 'culture of real virtuality. His understanding of the network society is heavily influenced by the Internet, the

technological configuration of which he draws upon when describing this new social structure, but he extrapolates social changes to all domains, cultural, economic, and political. Castells has been criticized for technological determinism, but remains one of the world's most well-cited scholars on matters related to Internet, power, communication, and globalization.

From an anthropological perspective, notions of technology-induced social change can be interpreted as forming part of the narratives of modernity framing Internet development. By approaching technology from the perspective of 'social constructivism, the STS school has been able to problematize this common understanding of progress, while deepening our understanding of the role of technological determinism in notions of modernity (Escobar, 1994). Bruno Latour, sociologist of science and self-proclaimed anthropologist of modernity, is an influential scholar in STS, and one of the key theoreticians of Actor–Network Theory. Latour (2005) challenges the modernist distinction between nature and culture, while focusing on social transformation through mediators, which can be human or nonhuman agents, and their acts of translation or mediation in networks of relations.

More recently, social media have been the subject of popular and scholarly claims of radical social change, as exemplified by discourses on the Twitter and Facebook revolutions in North Africa and the Middle East in 2011. Although social scientists as well as social activists have vehemently challenged the suggestion that social media played a critical role in the popular uprisings of the 'Arab spring, mainstream Western media have perpetuated this view, not least by retransmitting information from various online sources. Media professionals have joined politicians, business representatives, and civil society advocates in praising the revolutionary power of social media, while techsavvy hackers have helped local activists bypass Internet censorship introduced by nervous governments.

This second wave of hyperbole has been more focused on media than technology, reflecting the ubiquity and convergence of technological interfaces that has taken place over the last decade. Some scholars still distinguish between 'new' media and 'old' media, thus accentuating differences between digital media and other media. The emphasis on 'social' media is a good example of this logic. All media are in fact social; they are 'machineries of meaning' or 'cultural technologies' (Hannerz, 1996). Each medium has its own characteristics, whether print or broadcast, but they all convey shared meanings between people. Social media are supposed to be different from other media in that the producer is also the consumer, there is a greater degree of interaction, and user-generated content can be created, exchanged, and circulated with considerable ease. With the advent of 'smart phones, social media have become even more accessible, and in addition to mobile interfaces of popular sites, new interactive mobile applications are constantly developed. Although these decentralized and interactive features have characterized the Internet all along, new technological interfaces have made them more accessible, thus contributing to their popularity as well as the hyperbole framing their use.

Anthropologists tend to offer a more empirically grounded appraisal of digital media and social change. Rather than taking radical claims at face value, anthropologists use their knowledge and methods to determine what changes are *de facto* taking place in a given cultural context, and how these changes are influenced by and in turn influence the development and use of digital media. In so doing, anthropologists can build upon one of the main theorems of digital anthropology, namely the intricate interlinkages between online and offline social worlds. From the outset, anthropologists have insisted that the Internet should not be viewed as something that stands apart from society, affecting it from the outside, but rather something that is socially embedded, the Internet both influencing and being influenced by society (Miller and Slater, 2000; Uimonen, 2001). More recently, this argument has been made in relation to virtual worlds, Boellstorff (2008) underlining that the virtual is not opposed to the real but to the actual. The argument has also been made and creatively illustrated in relation to social media. Michael Wesch's early analysis of social media and hypertext underlines the human essence of the Internet. His introductory YouTube video *Web 2.0… The Machine is Using Us* has proven quite popular and has been viewed by over 11 million people since it was posted in January 2007. In his recent Facebook ethnography, Miller describes the profound impact of this online platform for the management of social relationships, while explaining that individuals have always been social networking sites (Miller, 2011).

This insistence on the interconnectedness of the virtual and the real, online and offline social worlds, digitally mediated and nonmediated interactions enable anthropologists to grasp momentary as well as lasting changes brought about by digital media, transformations that build upon as well as alter existing social realities.

A complementary way of determining social changes introduced by digital media is through Internet ethnography that is, ethnographic studies of technological development. Since the Internet, like all technologies, is a social construct, it is possible to ethnographically study the people who are developing the technology, along with the cultural values and meanings that influence their work. In other words, the very stuff of social embeddedness – the people who develop the technology – can also be ethnographically investigated. In his detailed study of computer geeks in different places in Asia, Europe, and the United States, Christopher Kelty (2008) is able to identify the cultural significance of free and open source software. The cultural meanings and social practices of software developers offer an ethnographic entry point into a much broader appraisal of how technological standards interact with and influence societal trends. Thus Kelty is able to show how the cultural practices of 'recursive publics' are impacting other social domains, well beyond software development. In an earlier Internet ethnography, Uimonen (2001) could show how the activities of Internet pioneers were influenced by and in turn helped influence the cultural meaning of the Internet in various developing countries, thus spreading not only Internet technology but also the 'culture of networking' to different parts of the world.

Internet and Globalization

Inseparable from the broader process of globalization, the Internet has contributed to the interconnectedness of the world, as evidenced in the intensification of transnational flows of people, ideas, goods, and services over the last few decades. The Internet has also connected people around the world in a myriad of decentralized networks of multimodal communication and interaction. Now that the public Internet has been around for two decades, especially in the well-connected rich economies of the world, it would be difficult to imagine a world without global interactivity. But it was not long ago that a fax machine was the cheapest and fastest way to communicate with people in distant places, the facsimile being a precursor to e-mail, in terms of speed and reach even more so than ordinary mail – or what came to be known as snail mail when e-mail took over written communication. Similarly, it was quite recently that satellite TV opened up the world to us, and before that television. Within years of becoming public, the Internet has brought us interactive media, making users consumers, producers and distributors of multimedia content. The change has indeed been remarkable, but perhaps not quite as radical as is often assumed, since the uneven spread of digital media has also highlighted the continuation of historically shaped inequalities around the world, especially the perpetuation of center–periphery relations.

Anthropological analyses of globalization help identify the cultural significance of the global spread of the Internet. While media scholar Marshall McLuhan used 'global village' primarily in relation to television in the early 1960s, the term was reinvigorated through the rapid growth of Internet use. Unfortunately, the concept global village portrays a somewhat illusory notion of social harmony and is therefore of limited explanatory value (Hannerz, 1996). By contrast, anthropologists have theorized the social and cultural significance of global interconnectedness, including more conflictual aspects like asymmetric center–periphery relations. Anthropological theory building on globalization pays close attention to the role of media in transnational flows and imaginaries, as exemplified in the work of Ulf Hannerz (1996) and Arjun Appadurai (1996). Indeed, some of the most advanced anthropological theories of globalization have preceded the Internet, yet help explain it. In a paper published in 1992, Ulf Hannerz referred to world society, or the 'global ecumene' as he called it, in terms of a 'network of networks. This was before the Internet, which is generally defined as the network of networks, had entered the public domain. Hannerz' elaborations on social networks built on his earlier work in urban anthropology (1980), where he had already described the city in terms of a network of networks. In his later work, he introduced the concept 'network of perspectives' to capture the social organization of cultural meaning in an increasingly interconnected world. Since they precede Castells' elaborations on the network society, anthropological network theories offer a complementary perspective on the global network society.

The global spread of the Internet has been accompanied by discourses on the 'digital divide, a concept that refers to uneven patterns of access and use, especially the gaps between the global North and the global South. Digital divides are also found within societies, such as gaps between rural–urban, old–young, educated–illiterate users and non-users. These gaps reflect the uneven distribution of infrastructure, not least electricity, as well as skills and abilities. Digital discrepancies are also evident in the production of digital technology, the availability of digital content, and the disposal of digital artifacts. Labor-intensive hardware is produced in low-cost countries, very little online content is produced by and for peripheral cultures, and computers discarded in Western countries are shipped to developing countries for redistribution. The availability of digital technology tends to follow patterns of overall development, the richest countries in the world being the most well-connected. Inspired by the binary logic of digital technology, Castells discusses the digital divide in the global network society in terms of inclusion and exclusion, while referring to the least connected and thus the most excluded continent in the world, Africa, in terms of the 'black hole of informational capitalism' (Castells, 1998: pp. 161–165). This binary view is, however, somewhat limited.

Anthropological analyses offer a more nuanced understanding of digital stratification, as exemplified by the concept 'partial inclusion' (Uimonen, 2012). Instead of looking at social inequalities in binary terms, an emphasis on patterns of digital stratification allows for a more varied appraisal of access and use in different parts of the world. From this perspective, it becomes clear that access and use in the peripheries can be a matter of partial inclusion in the global network society. In many parts of the world, access is shared and available in designated areas only (computer labs, Internet cafes, offices), while digital media skills can be quite rudimentary. Even so, these islands of connectivity can stimulate perceptions of global interconnectedness, thus corresponding to local aspirations for membership in modern world society (Ferguson, 2006). Limited as it may be, the state of partial inclusion is indicative of much broader social and cultural patterns that are of great interest to anthropologists in their quest to understand and explain the current state of the world.

The ways in which digital media are used in different parts of the world, not least in the peripheries, can also shed light on the social and cultural significance of transnationalism and translocality. From early on, anthropological studies have shown that far from erasing local identities, the Internet tends to reinforce a localized sense of belonging. If anything, the Internet helps people explore their cultural identities, a process that has been described in terms of 'expansive realization' and 'expansive potential' (Miller and Slater, 2000). At the same time, the Internet can also stimulate a more cosmopolitan identification with the world at large (Miller and Slater, 2000; Hart, 2000; Uimonen, 2001, 2012). If anything, the Internet clarifies the extent to which people have become connected to many more places than the place where they may physically live or reside. Through the Internet, people can manage their translocal and transnational social relations, thus maintaining both local and global cultural identities. The Internet is also stimulating the expression of more hybrid and creolized identities, especially through visual interfaces like Facebook (Uimonen, 2012).

Online Knowledge Production

The Internet opens up new possibilities for anthropological knowledge production, while raising new challenges. The Internet enables online fieldwork, which does not require the physical relocation of the researcher. Compared to travel to distant localities for an extended period of time, this is a less costly and in many ways more convenient way of carrying out fieldwork. It is still rare, however, that anthropologists carry out their research solely online. Boellstorff (2008, 2012) is a notable exception with his ground breaking fieldwork in the virtual world Second Life. It is more common that anthropologists combine on- and offline fieldwork, often feeling more comfortable having face-to-face encounters with informants in their offline environments as well as online. Online fieldwork also has the benefit that large amounts of field data can easily be gathered and stored. For instance, social media like blogs and Facebook retain all user data for many years. In the case of blogs, the material is even publicly available, while Facebook requires a relation with the page owner. This can of course easily lead to data overload since the amounts of information can be overwhelming. Neither is it always certain that websites remain online, making online content rather precarious. Another challenge is that the Internet develops at a rapid pace, making it rather stressful to keep abreast of technological developments. Many anthropologists are challenged by technology and shun away from the Internet, keeping their own use at a minimum. In general, digital anthropologists are drawn from anthropologists with a keen interest in digital media, i.e., more geeky anthropologists, although this is by far not a prerequisite. Younger generations of anthropologists tend to be more techsavvy and thus less hesitant or shy about doing research through and about digital media.

Online research raises new ethical issues, from informed consent to the protection of identities. Similarly to offline public places, it is quite impossible to get informed consent from users of an online forum, especially over a longer period of time. It can also be difficult to disguise the real identities of informants when research results are published in audiovisual formats online. And while anthropologists may be used to doing research about peoples who will not read their work, online research is often understood and in many cases challenged by informants. There are no clear cut answers to these issues, but anthropologists need to face them as they venture into brave new online worlds.

When it comes to data gathering, anthropologists can look forward to exciting opportunities. Over the last decade, digital technology has become more accessible and affordable. Digital recording devices like cameras, video cameras, and sound recorders are now readily available, at reasonable costs. Technological convergence makes these tools even more easy to use. With a smart phone, the researcher can record images, film, and sound, and upload these online, while a new generation of digital cameras enable the user to upload images to the Internet straight from the camera. As the devices become more easy to use, new opportunities are also opening up for participatory interactive research, since informants can be equipped with digital cameras or videos and create their own digital stories. Technological advancements are paving the way for more visual and sensory ethnography, as the researcher is better able to record field data gathered through all senses. In this regard, digital anthropology can benefit from the work carried out in the fields of visual anthropology and the anthropology of the senses.

Online publication of anthropological knowledge is still in its infancy, but interesting explorations have already begun. Although some of the earliest monographs were accompanied by websites (Miller and Slater, 2000; Uimonen, 2001), these sites were mainly textual and after a few years they ceased to exist. Kelty's *Two Bits* (2008) was a ground breaking example of online publication. True to his subject matter of free and open source software, Kelty's entire monograph is available online, for download and/or modulation. The blog where the book is posted, http://twobits.net/, lists comments, and points users to other sites where the book is reviewed and debated.

More recently, Uimonen's (2012) monograph on digital media in Tanzania has been published with an accompanying website with photos and videos from fieldwork as well as a hyperlinked ebook. Published in Routledge's new book series *Innovative Ethnographies*, the monograph *Digital Drama* explores innovative forms of ethnographic representation. The book and the series website weave together visual and audiovisual narratives in artistic formats, *Digital Drama* is the first ethnography of digital media in an African context and the ebook version is hyperlinked with the book's website, at http://www.innovativeethnographies.net/digitaldrama.

Anthropological perspectives on Internet and social media can clearly advance our understanding of the social and cultural impact of digital media in our interconnected world. As Escobar noted in the early days of digital anthropology, 'cyberculture, moreover, offers a chance for anthropology to renew itself without again reaching, as in the anthropology of this century, premature closure around the figures of the other and the same. These questions, and cyberculture generally, concern what anthropology is about: the story of life as it has been lived and is being lived at this very moment' (Escobar, 1994: p. 223).

See also: Center–Periphery Relationships; Cosmopolitanism; Creolization: Sociocultural Aspects; Infrastructures of Society, Anthropology of; Modernity; Popular Culture.

Bibliography

Appadurai, Arjun, 1996. Modernity at Large: Cultural Dimensions of Globalization. University of Minnesota Press, Minneapolis, Minn.

Boellstorff, Tom, 2008. Coming of Age in Second Life. An Anthropologist Explores the Virtually Human. Princeton University Press, Princeton, New Jersey.

Boellstorff, Tom, Nardi, Bonnie, Pearce, Celia, Taylor, T.L., 2012. Ethnography and Virtual Worlds: A Handbook of Method. Princeton University Press, Princeton, NJ.

Castells, Manuel, 1996. The Rise of the Network Society. In: The Information Age: Economy, Society and Culture, vol. 1. Blackwell, Malden, MA.

Castells, Manuel, 1997. The power of identity. In: The Information Age: Economy, Society and Culture, vol. 2. Blackwell, Malden, MA.

Castells, Manuel, 1998. End of millennium. In: The Information Age: Economy, Society and Culture, vol. 3. Blackwell, Malden, MA.

Coleman, Gabriella, 2010. Ethnographic approaches to digital media. Annual Review of Anthropology 39, 487–505.

Escobar, Arturo, 1994. Welcome to Cyberia. Notes on the anthropology of cyberculture. Current Anthropology 35 (3), 211–231.
Ferguson, James, 2006. Global Shadows. Africa in the Neoliberal World Order. Duke University Press, London.
Garsten, Christina, Wulff, Helena (Eds.), 2003. New Technologies at Work: People, Screens and Social Virtuality. Berg, Oxford.
Hannerz, Ulf, 1980. Exploring the City: Inquiries toward an Urban Anthropology. Columbia University Press, New York.
Hannerz, Ulf, 1992. The global ecumene as a network of networks. In: Kuper, Adam (Ed.), Conceptualizing Society. Routledge, London, pp. 34–56.
Hannerz, Ulf, 1996. Transnational Connections: Culture, People, Places. Routledge, London.
Haraway, Donna, 1991. Simians, Cyborgs and Women: The Reinvention of Nature. Routledge, New York.
Hart, Keith, 2000. The Memory Bank. Money in an Unequal World. Profile Books, London.
Horst, Heather, Miller, Daniel (Eds.), 2012. Digital Anthropology. Berg Publishers, Oxford, UK.
Kelty, Christopher, 2008. Two Bits: The Cultural Significance of Free Software. Duke University Press. Online at: http://twobits.net.
Latour, Bruno, 2005. Reassembling the Social: An Introduction to Actor-network Theory. Oxford University Press, Oxford.
Miller, Daniel, 2011. Tales from Facebook. Polity Press, Cambridge.
Miller, Daniel, Slater, Don, 2000. The Internet: An Ethnographic Approach. Berg, Oxford.
Turkle, Sherry, 1995. Life on the Screen: Identity in the Age of the Internet. Simon & Schuster, New York.
Uimonen, Paula, 2001. Internet, modernization and globalization. In: Stockholm Studies in Social Anthropology, vol. 49. Almqvist & Wiksell, Stockholm. Transnational.dynamics@development.net. Internet, modernization and globalization.
Uimonen, Paula, 2012. Digital Drama. Teaching and Learning Art and Media in Tanzania. Routledge, New York.

Relevant Websites

http://www.innovativeethnographies.net/digitaldrama – Digital Drama.
http://www.itu.int – International Telecommunication Union (ITU).
http://www.silva-sandbox.ucl.ac.uk/digital-anthropology/centre-for-digital-anthropology – Centre for Digital Anthropology.
http://www.ucl.ac.uk/anthropology/digital-anthropology/ – MSc in Digital Anthropology.
http://mediatedcultures.net/ – Digital Ethnography.
http://www.media-anthropology.net – Media Anthropology.
http://openanthcoop.ning.com/ – Open Anthropology Cooperative (OAC).
http://www.isoc.org – Internet Society (ISOC).
http://twobits.net/ – Two Bits.
http://www.youtube.com/watch?v=6gmP4nk0EOE – Web 2.0: The Machine is Us/ing Us.
http://savageminds.org/ – Savage Minds.

Internet Research: Ethical Concerns

Annette N Markham, Aarhus University, Aarhus N, Denmark
Elizabeth A Buchanan, Center for Applied Ethics, University of Wisconsin-Stout, Menomonie, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Across the globe, societies are experiencing significant flux and transformation in the way digital information and networked technologies are woven into and influence everyday life, workplace practices, social norms, and institutional structures. Likewise, all academic disciplines are challenged to make sense of these transformations in light of ethical research practice. Actively confronting these challenges means not only being aware of them, but thoughtfully considering how we might better address these challenges. This article discusses various ethical challenges in applied research involving digital media or Internet-related contexts.

Introduction

Ethical considerations are a constant companion for any researcher. Whether one is in the hard sciences, social sciences, humanities, or some inter- or intradisciplinary space, researchers face a continuum of choices and decision points around benefits, harms, and risks in their work. Ethical research explores what is possible, just, and fair in a given context. Internet research, broadly conceived, provides numerous avenues for researchers across the disciplines: New methods, newly constituted forms of data, novel forms of data collection, and once unimaginable mixing and reuse of data have emerged in the rich and fertile Internet landscape of social media applications, big data streams, and the conglomeration of technologies into discrete data streams. This article will discuss various ethical challenges in applied research involving digital media or Internet-related contexts.

Ethical issues are more obvious in certain situations than others, such as drug testing on human subjects. Here, potential harm is anticipated and analyzed in advance; the participant in the study is physically present; and the vulnerability of the participant is more readily recognized. In other research situations, the dilemmas might not be so clear. Blood samples from a previous study might become a convenient sample for a researcher looking at a certain population's propensity for schizophrenia. Any harm to individuals who gave the blood anonymously for a diabetes study – four decades ago – might seem minimal. Still, if a study of this blood shows that a particular indigenous tribe located in Arizona was not really indigenous, plus they have a tendency toward schizophrenia, this knowledge can cause significant psychological harm to the tribal community. This actually happened, and the spark occurred when a tribal member attending a medical conference in Washington, DC suddenly discovered this information about her tribe, the Havasupai, while listening to a professional colleague's report of these findings. (This case involved blood and DNA samples taken from the Havasupai Indian tribe in Arizona. See http://www.whoownsyourbody.org/havasupai.html.) This case is a prime example of individual and group harms that can occur as a result of research activities, even those that rely on 'anonymous' data or specimens.

In research situations involving the Internet and digital materials, ethical questions have focused for many years on such areas as representation, privacy, the nature of public data, and identities. These questions include: What counts as a human? Can a person's avatar in *Second Life* or another virtual world qualify for privacy protection? What about one's LinkedIn or Facebook profiles, should these be considered authored texts or representatives (avatars) of individuals? Do platform developers, marketers, journalists, or researchers have the right to examine our online activities, simply because they are publicly available? Do companies have the right to conduct mass experiments on users of their technological platforms? Given current capacities to track, collect, archive, and analyze our digital traces, is it even possible to entirely protect or anonymize our personal data? What happens when our participants want to be given credit for their participation? How do we balance a requirement to protect privacy through anonymity with a competing request to ensure publicity through nonanonymity?

These questions have no definitive answers from a technological, social, or legal point of view, but different disciplinary approaches have been offered, as have recommendations from various federal research agencies, such as Canada's Tri-Council, Norway's National Committee for Research Ethics in the Social Sciences and Humanities, and the U.S. Secretary's Advisory Committee to the Office for Human Research Protections.

Ethical practice is essential to responsible research, yet the parameters to guide such practice are neither universally accepted nor straightforward. This article will summarize persistent contemporary dilemmas of Internet-related research and outline some of the ways these dilemmas have been or might be better addressed by legislators, research communities, and scientific research oversight committees. Because the topic of Internet research ethics has undergone significant changes in recent decades, this article strives to present the most current articulations within these changing conversations. Hence, it presents strong ethical decision making as a case-based, inductive process, rather than a practice of narrow or strict interpretations of legal or institutional guidelines. The reader is urged to consider this stance within the legal, institutional, regional, and cultural parameters of the specific research project or situation. (This may become an issue of critical concern when doing research that involves information or persons outside one's own legal and cultural boundaries. Beyond knowledge of basic philosophical principles and extant

guidelines, researchers must be aware of legal requirements in the countries involved. This can create important and potentially unresolvable tensions that should be acknowledged (Markham and Buchanan, 2012).) To note, this discussion is not intended as a set of codes or guidelines, but as an overview of how ethical discussions have shifted in the United States because of the complications brought about by technologies and capacities associated with digital materials and the Internet.

The material in this article is largely based on ethical discussions within the general research arena of Internet Studies. More specifically, it reflects the stance developed by the Association of Internet Researchers (AOIR). Some of the material in this article is derived directly from the 2012 report coauthored by Annette Markham and Elizabeth Buchanan with feedback from the AOIR Ethics Working Committee (Markham and Buchanan, 2012). Readers should check the AOIR Web site for the most recent updates about these issues.

Origin of Ethical Guidelines for Scientific Research

Grave ethical missteps in research environments involving people prompted a worldwide response in the mid-late twentieth century. Most ethics guidelines were created with the specific goal to protect persons. Several documents now serve as a foundation for the ethical treatment of persons in research studies or as they are often known, 'human subjects,' including the UN Declaration of Human Rights (1948), the Nuremberg Code (1949), the Declaration of Helsinki (1964), and the Belmont Report (1979). As Markham and Buchanan (2012) note, "the basic tenets shared by these policies include the fundamental rights of human dignity, autonomy, protection, safety, maximization of benefits and minimization of harm, or, in the most recent accepted phrasing, respect for persons, justice, and beneficence" (p. 4).

Not exclusively, but especially in the United States, these three principles – respect for persons, justice, and beneficence – have been operationalized in specific ways. This is very briefly reviewed below.

Respect for Persons

In principle, this means that individuals should be treated as autonomous beings, capable of making their own decisions. In addition, it is recognized that certain persons should receive additional protections if they have less autonomy. This category of vulnerability includes, for example, young, disabled, pregnant, or incarcerated persons. In practice, or as described within many regulatory documents, this principle has been translated to informed consent. Persons must be given adequate information, clearly understand the risks and benefits of a study, and be able to make a voluntary decision to participate. They must also be allowed to choose to end their participation at any time.

Justice

The initial concept of justice as envisioned by the Belmont Report focuses attention on fairness of distribution of scarce benefits. This essentially means that equal people should be treated equally, or that people should not be denied the benefits of research without good reason. In practice, this concept has been operationalized to mean that the selection of research subjects should be fair so that not only certain select groups gain from the benefits or suffer the risks of such research. This concept has been extended to include fair access, so that persons have an equal opportunity to participate in studies that might benefit them, as in the case of innovative methods, procedures, or therapies in biomedical clinical trials.

Beneficence

This principle holds that humans should not be harmed in the course of study. In practice, at least in the United States, harm is assessed in relation to a risk–benefit analysis. If the benefit is seen to be great, this may outweigh the risks or costs. This utilitarian approach differs significantly from other cultural contexts that promote a more deontological approach. For example:

> [At] least on an ideal level, European approaches tend to emphasize more deontological approaches – i.e., approaches that take basic human rights (self-determination, privacy, informed consent, etc.) as so foundational that virtually no set of possible benefits that might be gained from violating these ethically justifies that violation.
>
> Ess and The AOIR Working Committee, 2002

Ethics of Research in the Digital and Internet Age

In line with the basic principles outlined in the Belmont Report and other international documents, the AOIR finds the following principles to be fundamental to an ethical approach to Internet research:

The greater the vulnerability of the community/author/participant, the greater the obligation of the researcher to protect the community/author/participant.

Because 'harm' is defined contextually, ethical principles are more likely to be understood inductively rather than applied universally. That is, rather than one-size-fits-all pronouncements, ethical decision making is best approached through the application of practical judgment attentive to the specific context.

Because all digital information at some point involves individual persons, consideration of principles related to research on humans (versus data) might be necessary even if it is not immediately apparent how and where persons are involved in the research data.

When making ethical decisions, researchers must balance the rights of subjects (as authors, as research participants, as people) with the social benefits of research and researchers' rights to conduct research. In different contexts the rights of subjects may outweigh the benefits of research.

Ethical issues may arise and need to be addressed during all steps of the research process, from planning to publication and dissemination.

> Ethical decision-making is a deliberative process, and researchers should consult as many people and resources as possible in this process, including fellow researchers, people participating in or familiar with contexts/sites being studied, research review boards, ethics guidelines, published scholarship (within one's discipline but also in other disciplines), and, where applicable, legal precedent (Markham and Buchanan, 2012, footnotes removed).

Even with these starting points, ethical conundrums arise and are rarely easy to resolve when encountered during a research project, due to multiple competing or conflicting values and norms at play in any situation. Likewise, different ethical issues arise at different critical junctures of a project, when the researcher is making logistical or methodological decisions about how to collect, manage, analyze, and interpret what might be called data, or how to present the results of one's research.

Persistent Dilemmas in Conceptualizing Ethical Considerations

We have found that Internet or digitally related contexts present sticky or persistent ethical dilemmas for social researchers. These could be discussed in many productive ways, through the lenses of philosophy, research methods, design, everyday use, work practices, and so forth. Here, four complications most directly related to the core tenets of ethical research guidelines are discussed: human subjects, privacy, personally identifiable information, and informed consent. These four concepts bring with them irresolvable tensions, as Markham and Buchanan (2012) note, since there are many ideas, values, and stakeholder interests competing in how these concepts get defined or actualized, whether in a general sense or in the specific context of a study. Rather than delimiting debate around these issues, we take the stance that they are best understood and managed in the most ethically sensible way, according to the needs determined by the specific research case. (For more philosophical discussion of phronesis as applied to digital research ethics and decision making, see the first version of the AOIR ethics report (Ess and the AOIR Ethics Working Committee, 2002) or Ess, 2009.) While general research ethics principles guide the work of ethics review boards as well as researchers, each and every case and context is unique and presents nuances for consideration.

Human Subjects

As written in the AOIR ethics guidelines (2012): "The concept of 'human subjects' is a persistent and contentious example of a dialectical tension between disciplinary/regulatory models and context-specific sensibilities" (p. 6). Emerging directly from biomedical research practices, the term 'human subject' is used broadly across all research disciplines in the United States as a baseline determinant of whether or not the research project should undergo ethical supervision. Yet this term "has been long critiqued for being ill-suited for models of inquiry that follow non-biomedical procedures for interacting with people" (p. 6), such as ethnographic observation or interviews. This ill fit becomes more apparent when deciding whether or not digital information qualifies for human subject protections and if so, how. As Markham and Buchanan clarify, the answer is not easily determined by relying on traditional or regulatory frameworks:

> Manipulation and close study of information generated by social media networks certainly constitutes a different research environment than sticking a needle into a volunteering person in a medical laboratory. On the other hand, entire communities have felt harm from use of their DNA data more than a decade after it was collected and anonymously aggregated. Interacting with a person in a face-to-face interview seems to raise different considerations than interviewing a person in an anonymous chatroom. Yet, anonymous members of online communities have felt harm after researchers published reports about what they perceived to be private community activities. Blogs are often considered public, published texts. On the other hand, users have described their blogs as a part of their identity, not to be treated as simply publicly accessible data.
>
> 2012: p. 6, including footnote text

Thus, the term 'human subject' may not be suitable as the baseline prompt for considering potential ethical questions. As AOIR concluded in the 2012 guidelines:

> [W]hen considered outside a regulatory framework, the concept of 'human subject' may not be as relevant as other terms such as harm, vulnerability, personally identifiable information, and so forth. We encourage researchers to continue vigorous and critical discussion of the concept of 'human subject,' both as it might be further specified in internet-related research or as it might be supplanted by terms that more appropriately define the boundaries for what constitutes inquiry that might be ethically challenging.
>
> Markham and Buchanan, 2012: p. 6

Privacy Concepts

Not only is our personal privacy a crumbling myth, the concept itself is contested and ambiguous. Researchers struggling with the practical and methodological challenges surrounding the issue of privacy find themselves on constantly shifting conceptual terrain. The term 'privacy' is almost always discussed as a binary that contrasts private with public. The effort on one level is to determine what constitutes public versus private, which has been seen as a useful way to guide one's actions in relation to this determination. Researchers typically use this distinction to help decide whether the material is a published text or a person, or to determine whether and how privacy needs to be protected.

Distinguishing privacy in relation to notions of the public is problematic on many levels. Commonly, people use publicly accessible information spaces but maintain strong expectations of privacy. Ikonomidis Svedmark and Nyberg (2009) interviewed dozens of Swedish youth (ages 15–25) who posted revealing pictures of themselves in public online spaces, such as blogs. Their responses revealed a complex and ambiguous understanding of privacy. As one Swedish girl put it: "I have this blog online because it's the only place where I feel safe and be sure my parents won't find it. If it had it in a diary in my room, they would find it."

This sort of experience of perceived, rather than actual, privacy has guided much of our current thinking about privacy and ethical social Internet research. Sveningsson (2003) argues that privacy is not only a continuum, it is

also a concept understood in relation to whether the information/content/context is considered sensitive or nonsensitive. She offers the following grid, which has been useful in broadening the conceptualization of privacy in digital contexts (Figure 1):

While this model is a useful progression from earlier notions of public and private, it remains problematic because the definitions remain embedded in binaries that do not adequately encompass experience or attitudes about privacy, sensitivity, and vulnerability. What is 'private' remains predominantly a location or state, opposed to or separate from 'public.' Questions about privacy in research design and process, then, also center on this binary, and the main question remains: "What is public/private?" This is the wrong question in contexts where one's information flow is constant and public, where people are always-connected, or where cutting, pasting, forwarding, reposting, and other mashup practices remix our personal information in globally distributed, complex networks.

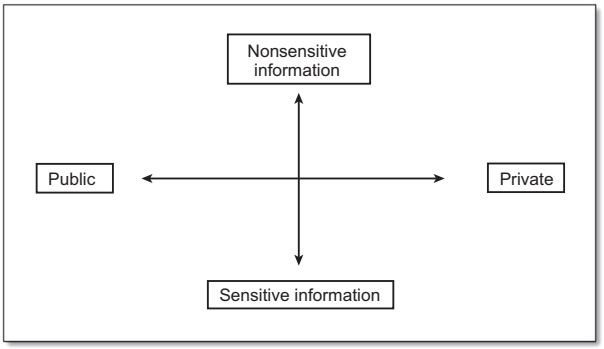

Figure 1 Sveningsson's model of Internet privacy, 2004.

McKee and Porter (2009) add other dimensions to Sveningsson's model to make more fine-grained distinctions (Figure 2).

This model adds more dimensions to how we might assess a research context, along various continua. They argue that in making decisions about whether or not a researcher needs to gain consent, the researcher should consider the extent to which the context might be public or private, how likely it is that the material or data might be connected to an individual or not, how much interaction the researcher has with a person, the level of sensitivity of the topic, and the vulnerability of the person involved. As with Sveningsson's model, this graphic retains a binary distinction between public and private. It presents a useful tool to think with, but can create other sets of binaries that become problematic in practice. When we try to use this model to explain tough cases, its utility breaks down.

It is quite common that people acknowledge that the substance of their communication is public, but are shocked when the information is taken out of context or used in a way that is unexpected or unwanted. For example, youth interviewed by Danah Boyd and Alice Marwick (2011) felt angry that their Facebook information was used by teachers and administrators in a school-wide public presentation to illustrate the dangers of posting private information in public spaces. They knew the information was relatively public. It was the manner in which it was used that made them feel violated (p. 6). Context, more than type of information, was the critical consideration. A similar attitude is expressed by the Swedish youth mentioned above (Ikonomidis Svedmark and Nyberg, 2009).

Thus, while for some users, privacy is a disintegrating condition or state, seemingly impossible to protect, for many other users, privacy – in any traditional sense – has never been a perceived right or concern. In mediatized contexts, as Nissenbaum points out, "what people care most about is not

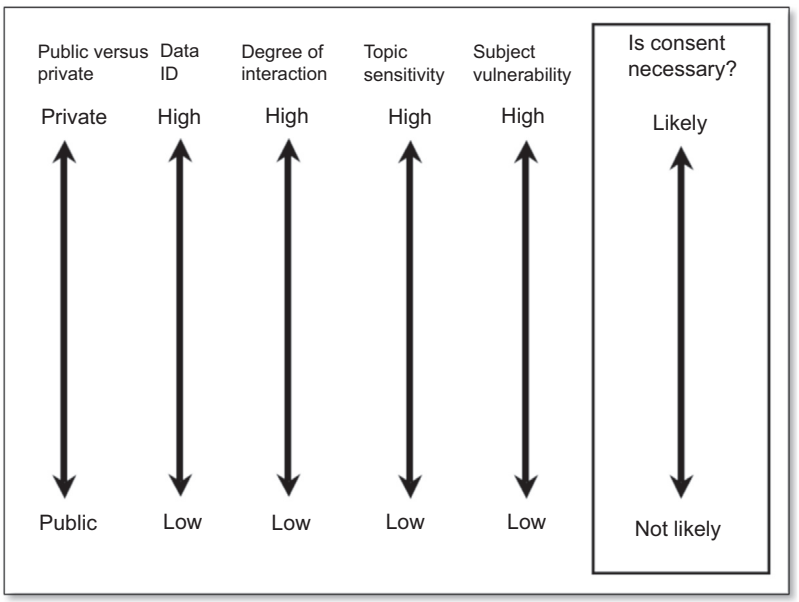

Figure 2 McKee and Porter's dimensions for informed consent, 2009.

simply restricting the flow of information but ensuring that it flows appropriately" (2010: Intro, para. 2). This is also supported by Boyd's discussions of youth and the emerging concept of private publics (2014). So to consider how we might protect participants, researchers may need to consider a different set of questions, such as: "What is a person's relationship to their information in this context?" or more specifically, even: "How does a person feel about his or her relationship to his or her information?" (Markham, 2012). These questions can be extended, of course, to larger groups or to discussions of data flow more broadly, as the case of the Havasupai tribe at the beginning of this article so vividly illustrates.

Informed Consent

This term poses many challenges in Internet-related research contexts. Emerging historically from ethical dilemmas of biomedical research, such as intravenous drug testing, the process of obtaining informed consent is designed to ensure that persons understand and accept the potential risks of research and willingly participate. This formulation is based on certain premises: first, it assumes that the researcher can literally see the potential participant to verify certain demographic information, such as age, socioeconomic status, or race that helps to determine whether or not the participant falls into a known category of vulnerability. Second, it assumes that procedures, purposes, and potential risks and benefits can be delimited and explained adequately. Both assumptions are faulty, but this history sustains a strong bias toward visual verification of 'informed,' 'consent,' and 'vulnerability.'

In Internet or digitally related research contexts, the objective of informed consent may be well intended but problematic. Researchers have sought and used robust and innovative approaches to online consent, and while no form of consent is perfect, consent in Internet research can raise particular issues. For example, is consent needed, and if so, How can we gain consent from anonymous chatroom or online forum participants? How can we ensure our participants do not fall into a vulnerable category? Where are waivers of consent appropriate in Internet research? What sort of protection should we afford information in large-scale digital databases, or, where are the human subjects in big data? These questions have compelled reconsideration of the goals and even the concept of informed consent. Many scholars such as Ikonomidis Svedmark and Nyberg (2009) provide good reasoning for why informed consent should not be an a priori (regulatory) requirement, but a (researcher) decision made based on the contingencies of the specific situation. Case studies illustrate how obtaining informed consent can cause harm, how informed consent is sometimes impossible, and how informed consent can be counterproductive. Svedmark's more recent work (2012) poignantly demonstrates that harm is – or should be – a more central consideration than informed consent.

In light of this, the AOIR ethics committee (Markham and Buchanan, 2012) suggests that researchers and ethical review boards can make stronger decisions about whether and how to obtain informed consent by returning to the rationale or premises for informed consent rather than taking for granted the regulatory norms for obtaining informed consent in a priori fashion. This stance would enable a more proactive role in determining how best – on a case-by-case basis – to enact beneficence, justice, and respect for persons. Notably, regulatory bodies are shifting more and more toward this stance. The National Science Foundation (2008) has extended guidelines to recognize and facilitate emergent, case-by-case considerations of informed consent. The U.S. Research Council is in the process of recommending significant changes to the way informed consent is conceptualized within the complexities of twenty-first-century information and data contexts (National Research Council, 2014). Additionally, research ethics boards are finding the flexibilities in the extant regulations to better match the fluid and synchronous mode of Internet communications, and also seeing more varied online consent models, including portals, virtual rooms, survey-type consents, and iterative consent.

The topic of informed consent continues to grow more complex as technologies and human social experience entangle with profit-minded stakeholders. Take, for example, the 2014 case of Facebook conducting mass manipulation of users' news feeds to conduct experiments on emotions. (A URL search for the Facebook emotional contagion study will yield a wide range of positions and reports on the topic and controversy. To view the original report, see Kramer et al. (2014).) Informed consent was not obtained, but was also not considered a relevant consideration. Users agreed to terms of service that allow for such manipulation. Significant public outcry is a powerful reminder that use of big data to conduct mass experiments on the digitally saturated public comes with ethical if not legal responsibility.

(Protection of) Personally Identifiable Information

Much of the discussion of privacy and ethics in the first decade of Internet research focused on protecting the privacy and anonymity of individuals and groups in online contexts. Now, at the end of the first decade of the twenty-first century, we face daunting challenges associated with data protection. As noted by the U.S. National Research Council (2014), "There is so much information that is freely and openly available about individuals that informational risk is ubiquitous in society. …. The issue of data protection spans the spectrum of methods and modes of inquiry in the social and behavioral sciences, whether qualitative or quantitative, longitudinal or experimental, observational or questionnaire-based, or micro- or macro-level or large-scale" (p. 83). As it gets easier to collect data in large-scale and automated ways, data and personally identifiable information become increasingly complicated categories. Archiving everyday deliberate activities, such as surfing patterns or demographic information, is commonplace. But it is also possible to collect nondeliberate activities such as backspacing and deleting. Programs can archive information passively or incidentally, as when a restaurant recommendation site not only tags but also stores one's geolocation, or when a self-tracking exercise device logs, statistically manages, and then archives one's physical data over time (sleep cycles, weight gain or loss, heart rate, running times, blood pressure, medication schedules, and so forth).

Snowden's (2013) revelation that the National Security Agency was secretly collecting personally identifiable information increased public awareness of the crumbling myth of data privacy. For research communities, too, this issue is

a growing ethical concern. While there is no consensus on how to handle and effectively protect personally identifiable information, the problem is recognized as a universal priority.

It is important to note that ethical issues continue to grow more complicated as new technologies and capacities develop. Although the above terms point to persistent concerns, others may emerge. Thus, it is important that ethical decision making remain active and case-based, rather than passive acceptance of or adherence to existing norms and parameters of regulatory bodies. This is an important shift in how regulators, policymakers, and research communities are talking about how ethical guidelines are best applied in actual research situations.

Provocations for Further Consideration

This article provides less direct guidance than questions or provocations to help guide thoughtful reflections. In this sense, this article aligns with a case-by-case or process-based approach to ethical decision making. This means more questions than answers.

In addition to suggesting that readers consult the ever-growing list of best practice guides, the following two principles may be useful: Ethical dilemmas are less visible from a distance, and harm or vulnerability can never be fully understood in advance. These two provocations confront Internet-specific considerations and serve as a rough summary of the past 20 years of shifting emphasis in the area of research ethics, based on the rise of computational power, storage capacity, and the ubiquitous characteristics of mobile and networked digital technologies.

Ethical Dilemmas Are Less Visible from a Distance

Some years ago, in an early draft of the AOIR ethics document, we suggested:

> As knowledge production becomes more dependent on information derived from some sort of Internet research, we recognize that the objects of scientific research are more person-based than not. This has become clear to our group over the past twenty years of Internet development. For instance, a researcher cannot talk about discursive or physical participation in the online information sphere without acknowledging that this necessarily involves a person somewhere in the process – a person who is thinking and behaving within his or her own cultural and moral stances. The Internet complicates the fundamental research ethics question of *personhood* (Is an avatar a person?) and *harm* (does the connection between data and persons or between an online identity and a real-life person enable psychological, economic, physical, harm?). Due to the complexity of Internet contexts, harm may not be immediately visible, but may emerge at any point in the research process.

One useful way of exploring the extent to which Internet texts or data sets might connect to persons is through what we might call the 'distance principle.' When the conceptual or experiential distance between the researcher and author/participant is close, we are more likely to naturally identify and frame the research scenario as one that involves 'humans' and therefore warrants ethical considerations of risks and potential harms. Arguably, as we work with aggregate data sets, separate texts from their original context, or use automated means of collecting, sorting, managing, and analyzing digital materials, the experiential or perceptual distance grows. As this happens, researchers may be more likely to define the research scenario as one that does not involve 'human,' simply because persons are no longer obviously present. This is most crudely illustrated when we compare an online interview through instant messaging with scraping data archives through an application programming interface and stripping this data of personally identifiable information.

In the latter, there is a greater possibility for forgetting that there was ever a person somewhere in the process that could be directly or indirectly impacted by the research. The distance principle can help one examine the role between the researcher and the researched. As the AOIR committee emphasizes, this sort of principle is not intended to establish the nature or extent of relevant ethical issues involved, but to simply help demonstrate that ethical issues are possibly present even if they are difficult to identify at that moment in time, or based on the technologies in use at that time. That is, ethical issues may be 'downstream' and only rise to the surface due to a change in Internet architecture, Internet norms, or even legal changes.

Harm or Vulnerability Can Never Be Understood in Advance

Although ethical guidelines were developed as a result of direct physical or psychological harm to persons participating in research studies, harm is discussed far less than those proposed methods for preventing it: informed consent and protecting privacy. While these latter terms are important, their regulatory construction often overshadows considerations of harm itself. This becomes evident when we look at counterintuitive examples of harm and vulnerability.

For example, Ikonomidis Svedmark and Nyberg (2009) studying blogging practices in Sweden wanted to publish excerpts from certain blogs for a publication. Based on typical regulatory interpretations of harm and vulnerability, they determined that a mother blogging intensively about the year prior to her daughter's suicide would be a vulnerable participant. They determined that an interior designer blogging about daily life and posting pictures of the family would not be considered a vulnerable person. The decision to base vulnerability on blog content is a sensible and common practice. Yet in this case, it mischaracterized both bloggers' feelings about their blogging activities. The mother of the suicide victim encouraged the researchers to post excerpts, name her directly, and point readers to her actual site. In other words, she did not want her confidentiality protected and she did not consider herself to be vulnerable. On the other hand, the interior designer was shocked and dismayed by the researchers' request and insisted that they not post any excerpts, photos, or other personally identifiable information about her blog. Potential vulnerability in this case could not be determined by using typical means of categorizing participants.

As our lives become more informational, potential harm can be more closely tied with considerations of how our personal information or personally identifiable information is being used outside the original context. As noted elsewhere (Markham, 2012):

> What may seem ephemeral or innocuous at one point in time might shift rapidly into something that causes real or perceived harm. Following information as it flows across multiple contexts, for example, may yield different assessments of privacy in each. For example, although I might willingly broadcast all my tweets to the public world at large, and would not mind my tweets being used in visual mapping of trends, I would feel violated to find all of my tweets – sent out in a sporadic fashion, in an ephemeral sense, or over a long period of time – combined together and analyzed by a researcher. Despite the fact that users recognize the overtly public nature of their presentation of self via digital media, this has no universally agreed upon or a priori correspondence with the harm that might eventually be felt.
>
> p. 4

As the terrains of research continue to shift along with new technologies, it behooves researchers to consider harm from a grounded, case-by-case perspective. This does not mean abandoning regulatory advice, but rather, paying attention to the intentions of these regulations. This involves at the very least interrogating the common (and often misleading) assumptions that harm is only connected to known categories of vulnerability, harm will be mitigated with informed consent, or that protecting privacy through confidentiality can prevent harm.

Shifting from Regulation-Based to Processes-Based Models

Across the globe, societies are experiencing significant flux and transformation in the way digital information is woven into and influences everyday life, workplace practices, social norms, and institutional structures. Likewise, all academic disciplines are challenged to make sense of these transformations in light of research practice. Actively confronting these challenges means not only being aware of them, but thoughtfully considering how we might better address these challenges.

This article has addressed the issue from a somewhat conservative perspective, focusing on the rights of participants and emphasizing the risks of current research practices. This is deliberately skewed, to urge researchers to assume a somewhat worst-case scenario approach (following the advice of pioneering Internet researcher Storm King in 1996). Certainly over the past 20 years, the landscape for what we might consider within the realm of ethical practice has shifted radically. We encourage a methodologically and disciplinarily sound approach for protecting research participants while engaging in the most productive of research endeavors.

Recommended Resources

For more information on ethics and research guidelines within the U.S. context, visit the Department of Health and Human Services Web site. This Web site offers an international compilation of human research standards (hhs.gov/ohrp/international/intlcompilation/intlcompilation.html). Notably, these conversations continue to change and grow. The information and data sharing capacities of the Internet have broadened discussions among a range of scientific and regulatory communities. The AOIR documents some of these discussions on their ethics wiki at ethics.aoir.org.

For more information on philosophical discussions of ethics within technological contexts or computing, see Buchanan's edited collection on virtual ethics (2004); Ess's work on *Digital Media Ethics* (2009); Moor's discussions of *Cyberethics* (2007); or Tavani's edited volume on *Ethics and Technology* (2013).

For more information on process-based models of ethical decision-making practices in digital contexts, see the AOIR ethics documents (2002, 2012); McKee and Porter's work on casuistic or case-based ethics (2009); Markham's discussions of ethics as method (2006); Wyatt's work on ethics in e-research (2011); or Nissenbaum's discussions of contextual integrity (2012).

For general overviews of Internet research ethics, see the AOIR ethics wiki, which links to the AOIR Ethics Committee guidelines as well as other resources. For an extensive review of key issues, see the entry by the same name in Stanford University's *Encyclopedia of Philosophy*: http://plato.stanford.edu/entries/ethics-internet-research/.

Intellectual property, copyright, and fair use are important ethical topics that were not addressed in this article. Copyright bridges legal and ethical realms. Conceptualizations, regulations, and laws vary by country. This is an important consideration in Internet research. Copyright and intellectual property may be a more relevant concern than the issue of human subjects in certain disciplines – i.e., linguistics. To get more information about these debates and best practices, see American University's Center for Media and Social Impact Web site, which contains an extensive section on fair use of copyrighted materials. For more information on copyright law by country, see http://en.wikipedia.org/wiki/Category:Copyright_law_by_country.

See also: Archiving: Ethical Aspects; Common Rule: US Regulatory Protections for Human Research Participants; Ethical Practices: Institutional Oversight and Enforcement; Informed Consent of Research Subjects; Media Ethics; Privacy of Individuals in Social Research: Confidentiality; Research Ethics, Cross-Cultural Dimensions of; Research Ethics: Research.

Bibliography

Boyd, D., 2014. It's Complicated: The Social Lives of Networked Teens. Yale University Press.

Boyd, D., Marwick, A., 2011. How Teens Understand Privacy, unpublished manuscript. Available online: http://www.danah.org/papers/2011/SocialPrivacyPLSC-Draft.pdf.

Ess, C., 2009. Digital Media Ethics. Polity, Cambridge.

Ess, C., The AOIR Ethics Working Committee, 2002. Ethical Decision-making and Internet Research: Recommendations from the AOIR Ethics Working Committee. Available from: http://www.aoir.org/reports/ethics.pdf.

Ikonomidis Svedmark, E., Nyberg, A., 2009. Om det privata i publika och digitala rum. In: Semig: Unga Om Sex Och Internet. Davidsons Tryckeri AB, Stockholm, pp. 354–383.

King, S.A., 1996. Researching Internet communities: proposed ethical guidelines for the reporting of results. The Information Soeicty 12 (2), 119–128.

Kramer, A.D.I., Guillory, J.E., Hancock, J.T., 2014. Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences of the United States of America 111 (24), 8788–8790. http://dx.doi.org/10.1073/pnas.1320040111.

Markham, A., 2006. Method as ethic, ethic as method. Journal of Information Ethics 15 (2), 37–55.

Markham, A., 2012. Fabrication as ethical practice: qualitative inquiry in ambiguous internet contexts. Information, Communication, & Society. http://dx.doi.org/10.1080/1369118X.2011.641993.

Markham, A., Buchanan, E., 2012. Ethical Decision-Making and Internet Research: Recommendations from the AOIR Ethics Working Committee (Version 2.0). Available from: http://www.aoir.org/reports/ethics2.pdf.

Markham, A., Buchanan, E., 2011. Ethical Decision-Making and Internet Research: Recommendations from the AOIR Ethics Working Committee (Draft), unpublished manuscript.

McKee, H.A., Porter, J.E., 2009. The Ethics of Internet Research: A Rhetorical, Case-Based Process. Peter Lang Publishing, New York.

National Research Council, 2014. Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences. Committee on Revisions to the Common Rule for the Protection of Human Subjects in Research in the Behavioral and Social Sciences. Board on Behavioral, Cognitive, and Sensory Sciences, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, Board on Health Sciences Policy, Institute of Medicine. The National Academies Press, Washington, DC.

National Science Foundation, 2008. Interpreting the Common Rule for the Protection of Human Subjects for Behavioral and Social Science Research. Available: http://www.nsf.gov/bfa/dias/policy/hsfaqs.jsp.

Nissenbaum, H., 2010. Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford University Press, Stanford.

Nuremberg Code, 1949. Trials of War Criminals before the Nuremberg Military Tribunals under Control Council Law No. 10, vol. 2. U.S. Government Printing Office, Washington, DC, pp. 181–182.

Snowden, E., 2013. The whistleblower behind the NSA surveillance revelations. The Guardian, 9 June 2013, p. 2. Online version posted 11 June 2013 at http://www.theguardian.com/world/2013/jun/09/edward-snowden-nsa-whistleblower-surveillance.

Svedmark, E., 2012. Att skydda individen från skada. En forkninsetusj balansakt. In: Kalman, H., Lövgren, V. (Eds.), Etiska dilemman. Forskningsdeltagande samtycke och utsatthet. Gleerups, Malmö, Sweden, pp. 106–107.

Sveningsson, M., 2003. Ethics in Internet ethnography. In: Buchanan, E. (Ed.), Readings in Virtual Research Ethics: Issues and Controversies. Hershey: Idea Group, pp. 45–61.

Tavani, H., 2011. Ethics and Technology: Controversies, Questions, and Strategies for Ethical Computing, third ed. Wiley Press, Hoboken, NJ.

United Nations Commission on Human Rights, 1948. The Universal Declaration of Human Rights. Available from: www.un.org/en/documents/udhr/index.shtml.

United States Department of Health, Education, and Welfare, 1979. The Belmont Report. Available from: http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html.

World Medical Association, 1964. WMA Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. Available from: www.wma.net/en/30publications/10policies/b3/.

Wyatt, S., 2011. Ethics of e-research in social sciences and humanities. In: Heider, D., Massanari, A. (Eds.), Digital Ethics, Research and Practice. Peter Lang, New York, pp. 5–20.

Interoception

André Schulz, University of Luxembourg, Walferdange, Luxembourg

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article summarizes contemporary knowledge on the psychology of interoception, i.e., the perception of bodily processes, and the underlying physiological mechanisms. First, theoretical models of interoception are presented and brain regions are discussed, which may play a role for the processing of interoceptive neural signals. Second, the most relevant methods to assess interoceptive accuracy are described. Third, the relevance of interoception for disciplines of general psychology is outlined. Fourth, the role of interoceptive accuracy for the development and maintenance of symptoms in clinical mental disorders are presented. Finally, some future challenges in interoceptive research are discussed.

Introduction

Sensory information from the body originates from different sensory receptor types, which can be distinguished (Dworkin, 2007) into (1) exteroceptors (e.g., mechanoreceptors in the skin), (2) proprioceptors (e.g., receptors in the spindles of skeletal muscles), and (3) interoceptors (e.g., mechano-, chemo-, thermo-, or metaboreceptors in visceral organs). Exteroceptors process external sensory information, such as visual, auditory, or tactile stimuli. A limited number of body-internal macroevents, such as heartbeats, may also partially be processed via exteroceptive, i.e., somatosensory receptors. Information on body position, muscular movement or tension are relayed by proprioceptors. Sensory information indicating the state of visceral organs is considered to constitute interoception. There is a variety of receptor subtypes that are responsible for interoceptive signal processing. Every organ system has its own population of interoceptors. They can be categorized using the following characteristics: (1) their functionality; receptors may be sensitive to physical movement or pressure (mechano- or stretch receptors), to a chemical substance or a specific concentration of this substance (chemoreceptors), to an under- or overrun of a threshold of a certain temperature (thermoreceptors), or to the registration of a metabolic process (metaboreceptor). (2) Their sensory threshold; e.g., one population of baroreceptors in arterial blood vessels respond to blood pressure changes within a range up to 100 mmHg, while another one responds to changes over 100 mmHg. (3) Their adaptability; the timeframe, within which interoceptors adapt to continuous stimulation may vary in subpopulations (fast- vs slow-adapting). Sensory information from interoceptors is continuously transmitted to the central nervous system. The majority is relayed over reflex circuits in the spinal cord or the brainstem without any conscious perception of these processes. However, this interoceptive information can reach consciousness if a predefined range of 'normal' functioning is exceeded. For instance, when body temperature will reach 38 °C, it is perceived as fever, if the filling of the stomach is exceeded, it is perceived as nausea, or if heart rate and blood pressure dramatically increase in a stressful situation, it is perceived as aversive. Interoception research does not only focus on conscious perception of bodily processes, but it also aims to understand interoceptive signal processing below consciousness level. One objective of this article is to summarize psychological disciplines, in which the conscious or unconscious processing of interoceptive signals play a decisive role.

Neurophysiology and Neuropsychology of Interoception

The majority of afferent signals from interoceptors are transmitted over the Nervus vagus and the Nervus glossopharyngeus and then relayed over the nucleus tractus solitarius (NTS), the main sensory brainstem center for visceral-afferent information. Reflex circuits for the maintenance of homeostasis in physiological processes, such as blood pressure (i.e., arterial baroreflex), oxygen saturation, or body temperature, involve synapses within the NTS. The NTS is in close exchange with other medullary structures, the caudal (CVLM) and rostral ventromedial medulla (RVLM), and the dorsal motor nucleus of the vagus (DMNV), which all play a role in these reflex circuits. From the NTS afferent signals are projected onto midbrain areas (parabrachial nucleus, PBN; periaquaeductal gray, PAG). The PBN and PAG are connected with higher structures, involving the limbic system, hypothalamus and thalamus, as well as cortical regions. The conscious perception of bodily signals certainly requires the involvement of the thalamus as important relay for sensory information. Cortical areas that are responsible for the processing of interoceptive neural signals are the anterior cingulate cortex (ACC), the anterior insula (AI) in the temporal lobe, the somatomotor cortex (SMC), the somatosensory cortex (SSC), and the prefrontal cortex (PFC). Neuroimaging studies have shown that a heartbeat detection task activates the right AI, the ACC, and the SMC, whereas the gray matter volume in the AI correlates with the accuracy in such tasks (Critchley et al., 2004). The processing of interoceptive signals, such as heartbeats, may result in the generation of specific electrocortical potentials (e.g., heartbeat-evoked potentials). These potentials are located in the ACC, the right IA, the SSC, and the PFC (Pollatos et al., 2005), which play a role in interoceptive signal processing. In addition to the transmission of afferent neural signals over the cranial nerves, another pathway over the spinal cord may also play an important role in interoception, the lamina1-spinothalamocortical pathway (Craig, 2002). Projections over

 http://dx.doi.org/10.1016/B978-0-08-097086-8.14145-5

this pathway also reach the NTS and thalamic structures, but it also integrates other afferent information from the body, such as information on nociception and tissue damage.

Concepts of Interoception

There is an ongoing debate, which mechanisms in afferent signal processing from the body can be called 'interoception.' Given the importance of interoception for psychology, there is also a shortness on theoretical models to include different levels of interoceptive signal processing. Vaitl (1996) views interoception as a general term to be subdivided into somatosensory, proprioceptive, and visceroceptive subsystems. In his model of interoception, sensory information from all three subsystems is transmitted to the CNS, where their representations are processed. CNS representation may reach awareness if attention is focused on visceral sensations. Learning processes are assumed to be the consequence of conscious CNS representations and may influence subjective reports, behavior or evaluation and could further affect the previous levels of interoceptive signal processing. The assumption of this regulatory circuit is congruent with feed-forward models of somatosensory amplification. The integration of somatosensory, proprioceptive and visceroceptive afferent signals is well in line with models on interoception and emotion, which propose that the experience of emotions relies on afferent signals from different sources (Wiens, 2005). Each of the methods to assess interoception described below involve different components of interoception in Vaitl's model. Heartbeat-evoked potentials, for example, require the stimulation of interoceptors *and* CNS representation, while awareness, report, or learning are not necessary, while they would be for performance in heartbeat detection tasks. From the current point of view, afferent signal transmission represents an independent factor, as it has been repeatedly shown that interoceptive signal processing is deficient in individuals with degeneration of afferent autonomic nerves, while interoceptors are still intact (Leopold and Schandry, 2001; Schulz et al., 2009). Furthermore, CNS representations can be differentiated into brainstem representations and higher (cortical) representations.

The denomination of interoceptive measures has been extensively discussed in literature. Common terms are 'interoceptive awareness,' 'interoceptive sensitivity,' or 'interoceptive accuracy,' although they all describe the same kind of measure (here: accuracy in a heartbeat detection task). Garfinkel and Critchley (2013) have addressed this issue and postulated a hierarchical model of interoception. They understand 'interoceptive sensibility' as dispositional tendency to be internally focused and thus as a prerequisite for the successful completion of interoceptive tasks. In their words 'interoceptive sensitivity' should represent the objective accuracy in these tasks, whereas 'interoceptive awareness' could be understood as metacognitive measure that quantifies individuals' explicit knowledge of their interoceptive accuracy. Although this differentiation is plausible, the measure indexed by interoceptive tasks should be called 'interoceptive accuracy' instead, since this term is rather neutral. It does not suggest that a specific score reflects 'sensitivity' or 'awareness,' since some individuals may be hypersensitive, which leads to misinterpretation of signals and thus to lower accuracy, or others may focus a maximal grade of awareness toward interoceptive stimuli that may still be below the sensory threshold.

Methods to Assess Interoception

Heartbeat Detection Tasks

The majority of research is based on experimental paradigms, in which participants are requested to detect their heartbeats. They are further divided into heartbeat counting tasks and discrimination tasks. Heartbeat counting tasks were developed by Schandry (1981), who named the task 'mental tracking test.' The task consists of time intervals of different duration (original version: 35, 45, and 55 s), during which the participants are instructed to silently count the number of their heartbeats. This number is later compared with the actual number of heartbeats in this interval. The absolute value of the difference between both is divided by the actual number of heartbeats and this 'inaccuracy' index is then subtracted from 1, which results in the accuracy score of the heartbeat counting task. It has been repeatedly demonstrated that the accuracy score depends on the wording of the instruction given. The comparison between a standard instruction ("count all heartbeats you feel in the body") and a strict instruction ("count only those heartbeats about which you are sure") suggests that an individual's knowledge on their heart rate and capacity to accurately estimate the duration of time intervals may be important factors for the accuracy in heartbeat counting tasks (Ehlers and Breuer, 1996). Heartbeat discrimination tasks were originally developed by Brener and Jones (1974) and further elaborated by others. In these tasks participants are asked to judge whether a set of consecutive exteroceptive stimuli (e.g., lights, tones, tactile stimuli) appear simultaneously with their own heartbeats (S+ trials) or not (S− trials). The available variants of this task mainly differ in the setup of S− trials, since some present exteroceptive stimuli with a fixed delay to heartbeats, while others simulate a set of artificial stimuli without any relation to the actual heartbeats. In a systematic investigation on possible time delays between heartbeats and the elicitation of an exteroceptive signal it was demonstrated that the optimal delay for the concurrent perception of both is about 230 ms after an R-wave (Brener and Ring, 1995). Heartbeat detection is reduced in individuals with a degeneration of afferent autonomic nerves due to diabetes mellitus.

Heartbeat-Evoked Potentials

Heartbeat-evoked brain potentials (HEPs) represent electrocortical potentials, which are related to the perception of cardiac signals, such as heartbeats. HEPs can be measured 250–600 ms after a cardiac R-wave and have their largest amplitude over the right hemisphere at frontal and central sites. HEP amplitudes have been demonstrated to reflect an individuals' performance in heartbeat detection tasks, motivation to perform in those tasks, and attentional focus on heartbeats. Nevertheless, the conscious perception of heartbeats is not required to produce HEPs, since afferent signals from the cardiovascular system continuously reach higher cortical areas. As a consequence, HEPs are measured during heartbeat

detection tasks (Leopold and Schandry, 2001), during rest (Schulz et al., 2013b) or during another foreground task (Gray et al., 2007). HEPs are calculated as event-related potentials with the R-wave in the ECG as 'event.' Since it was demonstrated that up to a time delay of 450 ms the electrocardiac field of the ECG and the HEP are partially overlapping, both effects have to be disentangled (Gray et al., 2007). Again, HEPs are reduced in individuals with a degeneration of afferent autonomic nerves (Leopold and Schandry, 2001).

Visceral Modulation of Startle

Afferent signals from visceral organs have the potential to modulate higher central information processing. Initially, this relationship was investigated in the cardiovascular system. The additive stimulation or unloading of arterial baroreceptors by the cardiac cycle and an external pressure/suction device affects the human startle response to auditory stimuli, a defensive reflex that is relayed over the brainstem. Even without external stimulation startle responses are lower during the early cardiac cycle phase (R-wave + 230 ms) than during the late cardiac cycle phase (R + 530 ms), as are subjective intensity ratings of the startle stimuli. The difference of startle amplitudes during early and late cardiac cycle phase has been called 'Cardiac Modulation of Startle' (CMS) effect. This effect is also largely diminished in individuals with diabetic neuropathy (Schulz et al., 2009). The CMS effect, therefore, reflects baroafferent neural traffic from the cardiovascular system. In contrast to the other methods described, the CMS may serve as a methodology to estimate afferent neural signals relayed at the level of the brainstem, since only brainstem processes are involved (arterial baroreflex circuit, acoustic startle circuit). Furthermore, afferent signals from cardiopulmonary baroreceptors, the gastrointestinal system and the respiratory system may also influence the human startle response. Thus, the concept of CMS could eventually be extended to 'Visceral Modulation of Startle.' HEPs and CMS are considered psychophysiological indicators of interoceptive signal processing.

Factors Determining Accuracy in Interoception Tasks

Differential Aspects of Interoceptive Accuracy

A number of sociodemographic factors play a role in the accuracy in interoceptive tasks. In general, results suggest that men are more accurate in interoceptive tasks related to sensations of the cardiovascular, the respiratory or the gastrointestinal system than women. Possible explanations for these effects could be sex differences in vascular musculature or hemispheric lateralization, which also play a role in interoception. The higher proportion of body fat in women than in men could also account for these differences, as it has been shown that individuals with lower body fat perform better in heartbeat perception tasks, and accuracy in these tasks is negatively correlated with BMI (Herbert et al., 2013). Furthermore, Khalsa et al. (2009) have demonstrated that accuracy in heartbeat perception tasks and age show a negative association ($r = -.30$). This correlation could be explained by reduced sympathetic tone, decreased sensitivity of baroreceptors, or reduced arterial stiffness with increasing age. Contrary to common wisdom, experienced meditators do not show higher accuracy in interoceptive tasks (Khalsa et al., 2008), than normal controls, although they report subjectively increased awareness to bodily states.

Physiological Determinants

Arterial baroreceptors represent one important type of interoceptors, which are responsible for cardiac interoception. In response to head-up tilt or isometric exercise, accuracy in heartbeat detection tasks is highly correlated with stroke volume (SV; $r = .59$) and momentum ($r = .67$), which represents the product of SV and velocity of blood outflow (Schandry et al., 1993). Furthermore, accuracy in heartbeat detection is associated with the change in the pre-ejection period (PEP) and the cardiac output in response to a mental arithmetic stressor. It could thus be concluded that cardiodynamic parameters, which are considered reflecting inotropic activation (contractility), are positively associated with accuracy in heartbeat detection. Nevertheless, this relationship has not been consistently reported in the literature. Heartbeat perception accuracy is positively related to systolic blood pressure (SBP), but negatively related to heart rate. Typically, there is a negative relationship between heart rate (HR) and SV, except for conditions of extensive exercise or stress. Increased SV, decreased PEP, and increased SBP may generally induce higher stimulation of arterial baroreceptors. Psychophysiological parameters that imply increased stimulation of baroreceptors per single heartbeat are, therefore, positively correlated with accuracy in heartbeat detection tasks (SV↑, PEP↓, SBP↑), while parameters that imply decreased stimulation per heartbeat (HR) show the opposite pattern. Individuals with a high degree of physical training tend to have superior accuracy in heartbeat detection compared with nontrained individuals.

Interoception and Emotions

Early Concepts

In the original James-Lange theory of emotion, all emotional processes are understood as epiphenomena of visceral activation, which takes place when an individual is confronted with an emotionally relevant stimulus. Nevertheless, this theory has not been empirically supported. According to the modified theory by Cannon and Bard, visceral responses and the genesis of emotions are concurrently elicited through the perception of an emotionally relevant stimulus. Individuals with interrupted transmission of visceral-afferent signals, for example, still experience emotions, although visceral activation is too slow to match the genesis of emotions. In the seminal study by Schachter and Singer (1962) individuals received epinephrine, which drives the sympathetic nervous system, or a placebo substance. One subgroup of participants was informed about the 'adverse effects' of the substance, another subgroup was naïve. The authors observed that the informed group attributed their subjective level of arousal to the adverse effects of the drug, while the naïve participants reported a higher intensity of anger after being exposed to anger provocation by an experimenter. These results suggest that both the perception of bodily states and cognitive appraisal are important for the experience of an emotion.

Current Point of View

Contemporary theories do not assume interoceptive signals as crucial for the experience of emotions anymore. Nevertheless, individuals with spinal cord injury have been found to report reduced intensity of emotions, which suggests that afferent information of bodily states is important, but not a prerequisite for emotion genesis. Current concepts assume that interoceptive neural traffic reaches 'first order' central structures, such as the dorsal pontine areas, the AI, and the somatosensory cortex, which may be associated with a phenomenological level of emotional experience. In a 'second order' of central representation of interoceptive signals, structures such as the anterior cingulum integrate these genuine interoceptive signals and other sources of bodily information, e.g., somatosensory feedback and attention. Self-reported emotional experience may thus be the consequence of integrating interoceptive signals and other information of bodily states, as well as paying attention to these processes (Wiens, 2005). A large body of evidence indicates that a more intense experience of emotions is positively associated with accurate heartbeat perception. Neuroimaging studies suggest that the processing of interoceptive signals and the experience of emotion share neural structures, such as the AI, the ventromedial prefrontal cortex (VMPFC), and the ACC (Pollatos et al., 2007).

Interocepion and Decision Making

Damasio's Somatic Marker Hypothesis

The influential somatic marker hypothesis by Damasio (1994) postulates that interoceptive signals are regularly integrated into processes of decision making. In every context, in which decisions are expected to be made, the outcomes of possible action alternatives are anticipated. This anticipation produces a specific visceral response to be integrated in affective responses to the expected outcome. These responses, either transmitted over somatosensory or visceral-afferent circuits, are called 'somatic markers.' Damasio hypothesized that these 'somatic markers' shape decision making. The neural structure that is thought to be responsible for the integration and evaluation of somatic markers is the VMPFC. There is a large body of evidence supporting the somatic marker theory. In experimental research, human decision making is assessed through specific paradigms, such as the Iowa Gambling Task. Performance in this task is positively associated with sympathetic responses in anticipation of punishment or reward, putatively reflecting the availability of somatic markers, as well as with accuracy in interoceptive tasks. Patients with neurological damage in the VMPFC do not show comparable electrodermal responses to anticipated punishment or reward (Bechara et al., 1996). Dunn et al. (2006) provide a comprehensive review of the somatic marker hypothesis.

Interoception and Stress

Interoceptive Accuracy during and after Acute Laboratory Stressors

While interoception describes the transmission of ascending signals on the brain–body axis, stress represents a prominent example for descending information on the brain–body axis. Stress and activation drive the two physiological stress systems, the sympathetic–adrenomedullary (SAM) axis including the sympathetic nervous system, and the hypothalamus–pituitary–adrenocortical (HPA) axis with cortisol as final product. Neural and endocrine signaling resulting from the activation of both axes may change the activity of visceral organs (e.g., inotropic and chronotropic effects on the myocardium) and typical brain regions, which are responsible for their perception (e.g., thalamus, insular cortex). Thus it can be argued that interoception and stress are closely linked. Laboratory stress tests are designed to provide a stressful environment that guarantees for a maximal degree of standardization, ecological validity (i.e., cognitive, affective, and behavioral aspects of stress), and effectiveness. However, the direction of the effect of a laboratory stressor on interoceptive accuracy largely depends on the methods used to investigate interoception. Comparing both heartbeat detection tasks, these opposite effects were observed with a strong laboratory stressor, the Socially Evaluated Cold Pressor Test (SECPT) (Schulz et al., 2013a): Stress seems to increase interoceptive accuracy if attention is solely focused on visceral sensations (i.e., heartbeat counting), while it decreases if attention is additionally focused on exteroceptive stimuli (i.e., heartbeat discrimination). One possible conclusion is that stress effects on interoception depend on attentional processes.

Evidence on the impact of stress on psychophysiological indicators of interoceptive signal processing is sparse. Acute stress induced by an SECPT was found to change the pattern of the cardiac modulation of startle (Schulz et al., 2011). Two studies investigated the impact of acute laboratory stressors on heartbeat-evoked potentials. A moderate arithmetic stressor could not modulate the HEP amplitude per se, but the HEP was correlated with cardiac responses to the stressor (Gray et al., 2007). Acute pain during a mild version of the cold pressor test attenuated HEP amplitudes (Shao et al., 2011), but possible later effects or relationships to both stress axes were not reported.

Stress Hormones

Certain stress hormones are considered to play an important role in the processing of interoceptive signals. First, the activation of the autonomic nervous system may result in a release of catecholamines, such as epinephrine. Although catecholamines do not cross the blood–brain barrier and, therefore, cannot modulate the processing of interoceptive signals at the level of the CNS, epinephrine binds to beta1-adrenergic receptors at the myocardium. The binding of epinephrine to beta1-adrenoceptors causes positive inotropic (increased contractibility), chronotropic (increased heart rate), and dromotropic effects (faster depolarization), which lead to an increase of SV and a decrease of the PEP (Schachinger et al., 2001). It was repeatedly shown that cardiodynamic parameters, such as PEP or SV, are associated with the accuracy in heartbeat detection tasks (Schandry et al., 1993). It is likely that the increased stimulation of arterial baroreceptors caused by increased SV is responsible for this effect. Furthermore, peripheral sympathetic activation induced via beta1-adrenergic

agents and central sympathetic activity elicited by alpha1--adrenergic agents increases heartbeat perception accuracy (Moor et al., 2005). Second, cortisol, the final product of HPA axis activation, may also have the potential to modulate interoceptive signal processing. A dose of 4 mg of intravenously administered cortisol may induce higher heartbeat-evoked potentials in a state of high alertness compared with a state of low alertness (Schulz et al., 2013b). This effect may feed into a vicious circle of increased attentional focus on physical symptoms, increased anxiety and higher levels of cortisol, which may represent possible psychoneuroendocrine mechanism underlying psychological feed-forward models of somatosensory amplification. Interestingly, cortisol does not affect the cardiac modulation of startle, suggesting that this effect is limited to the cortical representation of interoceptive signals, but does not alter the brainstem processing of afferent bodily signals (Schulz et al., submitted). In contrast to catecholamines, cortisol crosses the blood–brain barrier and affects cells in the entire body, including the CNS. The described results can be attributed to activity in the thalamus, which represents a major relay center for exteroceptive and interoceptive sensory information, and was demonstrated to be affected by cortisol administration. The importance of the HPA axis for altered perception of physical sensations is further underlined by the fact that dysregulation of the HPA axis can be observed in many body-related mental disorders, such as somatoform disorders (Rief and Barsky, 2005) or depression (Terhaar et al., 2012).

Chronic Stress

There is very little research on chronic stress and interoception. This is surprising in view of the widely accepted notion that chronic stress exposure causes dysregulation of the SAM and the HPA axes, which in turn are of particular relevance for the above mentioned mental disorders. Clinical data suggests that these dysregulations contribute to the generation of physical symptoms and may thus disrupt the adequate processing of interoceptive signals (Rief and Barsky, 2005). Nevertheless, an association between self-reported chronic stress and interoceptive accuracy in laboratory paradigms has not been demonstrated, as yet (gastric stimulation: Rosenberger et al., 2009; heartbeat detection task: Schulz et al., 2013a). Both studies, however, assessed chronic stress in healthy participants. It remains for future studies to investigate possible abnormalities in interoceptive accuracy in chronically stressed individuals, which may already show SAM and HPA axes dysregulation.

Interoception in Learning and Memory

The model of stress and memory consolidation by McGaugh (2000) postulates that two physiological processes are crucial for the enhancement of memory consolidation after stress: (1) glucocorticoids (rodents: corticosterone; humans: cortisol) that bind on glucocorticoid receptors within the NTS, the locus coeruleus (LC), and the basolateral amygdala (BLA) and (2) peripherally circulating epinephrine, which cannot cross the blood–brain barrier, but activates visceral organs, whose afferent signals are transmitted to the NTS and LC. From these structures, over noradrenergic pathways, beta-adrenergic synapses in the BLA are activated, which induces the release of cyclic adenosine monophosphate (cAMP) and cAMP-dependent protein kinase. Both substances can enhance memory consolidation. Postsynaptic efficacy of beta-adrenergic synapses is increased by glucocorticoids. Results from animal models show that both processes, glucocorticoid secretion and afferent signal transmission from visceral organs, are necessary for memory consolidation. In humans, the increase of baroafferent neural feedback induced by norepinephrine-infusion improved long-term memory (Moor et al., 2005). Interoceptive neural signals, therefore, play an important role for the stress-induced improvement of memory consolidation in animals and humans.

Interoception in Mental Disorders

Panic Disorder

The most frequently investigated disorder in relation to interoception is panic disorder. In cognitive models of panic disorder it is assumed that panic patients have enhanced selective attention for threat-related stimuli in general and body-related cues in particular. This attentional bias may result in increased interoceptive accuracy. In concordance with this hypothesis, many empirical papers/articles found superior accuracy in heartbeat counting tasks in patients with panic disorder as compared with healthy individuals or other mental disorders (Ehlers and Breuer, 1996). A positive relationship between panic symptoms and interoceptive accuracy has also been observed in children (Eley et al., 2004). Cognitive-behavioral therapy of panic disorder does not change performance in heartbeat perception, but heartbeat perception negatively predicts the outcome of therapeutic intervention (Ehlers and Breuer, 1996). Nevertheless, there are also negative reports, and Willem Van Der Does et al. (2000) conclude from a re-analysis of the data of many of these studies that the majority of patients with panic disorder do not have generally increased accuracy, but only a minority. Moreover, higher accuracy in heartbeat counting tasks was only found when using the standard instruction ("count all heartbeats you feel in the body") and not in the strict instruction ("count only those heartbeats about which you are sure") (Ehlers and Breuer, 1996). This highlights the importance of knowledge about one's own heart rate in panic patients for performing this task, which may be more precise in panic patients, while healthy individuals tend to underestimate their heart rate in this task.

Depression

Affective disorders are often accompanied by autonomic dysregulations, which are also reflected in specific symptoms, such as the disruption of the sleep–wake cycle, dysregulation of appetite, sexual dysfunction, or psychomotor inhibition or agitation. It is plausible that the dysregulation of autonomic processes also induces alterations in the perception of phenomena that are autonomically controlled. Dunn et al. (2007) have reported that individuals with a mild depression may have reduced interoceptive accuracy, while individuals with a severe depression may not differ in their accuracy

from healthy controls. However, in this investigation it is unclear whether the difference between both depression groups is due to the severity of disease or to their use of medication, since the number and dosage of antidepressants was notably higher in the severe depression group. A reduction of interoceptive accuracy in major depression is also reflected in a decrease of heartbeat-evoked potentials as compared with healthy control individuals (Terhaar et al., 2012), which may be associated with altered activity in the insular cortex, which could be observed in depressed individuals. Since in a clinical setting there is a substantial overlap of depression and anxiety symptoms, it is not clear to what degree both components may differentially affect interoception. In a student sample it was demonstrated that there is a negative relationship between depression and interoceptive accuracy only in highly anxious individuals (Pollatos et al., 2009), which suggests the more important role of anxiety for interoception. Furthermore, it was reported that anxiety and anhedonia in a subclinical sample interact in that there is a positive relationship between anxiety and interoception only in individuals with low anhedonia (Dunn et al., 2010). Taken together, depressive symptomatology is associated with a tendency for decreased interoceptive accuracy, which may additionally be moderated by anxiety level.

Somatoform Disorders

The perception-filter model of somatoform disorders (SD) postulates that hyperactivation of the autonomic nervous system and the HPA axis contribute to increased perception of physical sensations in SD (Rief and Barsky, 2005). Study findings, however, are very mixed. This could be due to the heterogeneity of the conditions classified as SD, some of which emphasize the misperception of bodily signals (e.g., medically unexplained symptoms), while others focus more on medically unestablished health-oriented anxiety (e.g., hypochondria). Even within the field of medically unexplained symptoms (MUS) it is questionable whether individuals with complaints in different organ systems should be classified under the same condition (e.g., cardiovascular symptoms vs. chronic back pain). Furthermore, there is a debate in the literature on the expected direction of the difference between somatoform patients and controls. As summarized by Schaefer et al. (2012), some models assume an increased focus of attention on bodily processes in SD, which may lead to increased accuracy in interoception, while opposing models propose deficient processing of interoceptive signals in SD. Empirical data are in partial support of both hypotheses. On the one hand, Pollatos et al. (2011) found that a mixed group of somatoform patients showed lower accuracy in a heartbeat counting task than a healthy control group. On the other hand, high symptom reporters in the normal population as well as patients with MUS both show less accurate perception of CO_2 ratio in the air than healthy individuals (Bogaerts et al., 2010). Schaefer et al. (2012) reported that a mixed somatoform outpatient group and a matched healthy control group did not differ in the accuracy of the heartbeat counting or discrimination task. However, a negative correlation between the number of symptoms and the accuracy in the heartbeat counting task was observed. It can be summarized that there is a tendency to observe lower accuracy in interoceptive tasks in patients with MUS. Nevertheless, SD in general and MUS in particular may be characterized by an inappropriateness of processing afferent information from the body. It is questionable whether this deficiency is solely reflected by a quantitative reduction in interoceptive paradigms or if this alteration should not rather be seen as a qualitative change. Moreover, the role of possible dysregulated stress responses in SD, and the role of depression for altered interoception is still neglected in the current literature.

Eating Behavior and Eating Disorders

Patients with anorexia nervosa (AN), bulimia nervosa (BN), and binge eating disorder (BED) all report increased subjective awareness to bodily processes. In contrast, patients with AN and BN show lower interoceptive accuracy than healthy controls (Herbert and Pollatos, 2014; Pollatos et al., 2008), while there is no data on patients with BED. One challenge in interoception research with eating disorders is to determine which components of a putatively altered interoception may be attributed to the mental disorder and which may be a physiological and metabolic result of the dysregulated food intake. For instance, the increased interoceptive accuracy in obesity could be a result of an increase in arterial blood pressure. Furthermore, with a quasiexperimental approach of clinical group comparisons, it is impossible to disentangle cause and effect. There is empirical evidence supporting both alternative hypotheses. On the one hand, eating behavior has been experimentally manipulated to assess possible effects on interoception. Short-term fasting was found to increase accuracy in heartbeat counting tasks (Herbert et al., 2012), and the amplitude of heartbeat-evoked potentials (Schulz et al., 2014). While short-term food deprivation induces an increase in sympathetic tone, these effects are inverted in long-term food deprivation. It is, therefore, plausible that long-term fasting, such as observed in AN, may be a way of regulating the perception of bodily sensations. This assumption is also in line with the finding that AN patients show reduced interoceptive accuracy (Pollatos et al., 2008). On the other hand, interoceptive accuracy is positively associated with intuitive eating (Herbert et al., 2013), suggesting that high interoceptive accuracy contributes to healthy eating behavior.

Future Directions

In the majority of the studies described, interoception was operationalized using interoceptive tasks, e.g., heartbeat perception tasks, or psychophysiological indicators of interoceptive processing, e.g., heartbeat-evoked potentials. These studies demonstrate linear relationships between accuracy in interoceptive tasks or group comparisons in terms of 'higher than' or 'lower than.' A careful revision of the literature on stress and interoception, as well as interoception in mental disorders as compared with healthy controls demonstrates that the assumption of linear relationships does not hold. Future research should investigate more complex relationships between interoceptive measures and mental disorders, and also focus on qualitative instead of only quantitative changes in

interoceptive processing. Another challenge for future research lies in the assessment of different interoception modalities at a time with the same technique (horizontal integration), as well as different levels of the same interoceptive modality (e.g., signals at the level of the brainstem, the cortex, and subjective reports; vertical integration).

See also: Chronic Pain: Models and Treatment Approaches; Co-morbidity of Mental and Physical Conditions; Explanatory Style and Health; Health Risk Perception; Illness Behavior and Care-Seeking; Stress and Cardiac Response; Stress, Coping and Health; Symptom Perception, Awareness and Interpretation.

Bibliography

Bechara, A., Tranel, D., Damasio, H., Damasio, A.R., 1996. Failure to respond autonomically to anticipated future outcomes following damage to prefrontal cortex. Cerebral Cortex 6, 215–225.

Bogaerts, K., Van Eylen, L., Li, W., Bresseleers, J., Van Diest, I., De Peuter, S., Stans, L., Decramer, M., Van Den Bergh, O., 2010. Distorted symptom perception in patients with medically unexplained symptoms. Journal of Abnormal Psychology 119, 226–234.

Brener, J., Jones, J.M., 1974. Interoceptive discrimination in intact humans: detection of cardiac activity. Physiology & Behavior 13, 763–767.

Brener, J., Ring, C., 1995. Sensory and perceptual factors in heart beat detection. In: Vaitl, D., Schandry, R. (Eds.), From the Heart to the Brain. The Psychophysiology of Circulation – Brain Interaction. Peter Lang, Frankfurt, pp. 193–221.

Craig, A.D., 2002. How do you feel? Interoception: the sense of the physiological condition of the body. Nature Reviews Neuroscience 3, 655–666.

Critchley, H.D., Wiens, S., Rotshtein, P., Ohman, A., Dolan, R.J., 2004. Neural systems supporting interoceptive awareness. Nature Neuroscience 7, 189–195.

Damasio, A., 1994. Descartes' Error: Emotion, Reason and the Human Brain. Grosset/Putnam, New York.

Dunn, B.D., Dalgleish, T., Lawrence, A.D., 2006. The somatic marker hypothesis: a critical evaluation. Neuroscience & Biobehavioral Reviews 30, 239–271.

Dunn, B.D., Dalgleish, T., Ogilvie, A.D., Lawrence, A.D., 2007. Heartbeat perception in depression. Behaviour Research and Therapy 45, 1921–1930.

Dunn, B.D., Stefanovitch, I., Evans, D., Oliver, C., Hawkins, A., Dalgleish, T., 2010. Can you feel the beat? Interoceptive awareness is an interactive function of anxiety- and depression-specific symptom dimensions. Behaviour Research and Therapy 48, 1133–1138.

Dworkin, B.R., 2007. Interoception. In: Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (Eds.), Handbook of Psychophysiology, third ed. Cambrigde University Press, Cambridge.

Ehlers, A., Breuer, P., 1996. How good are patients with panic disorder at perceiving their heartbeats? Biological Psychology 42, 165–182.

Eley, T.C., Stirling, L., Ehlers, A., Gregory, A.M., Clark, D.M., 2004. Heart-beat perception, panic/somatic symptoms and anxiety sensitivity in children. Behaviour Research and Therapy 42, 439–448.

Garfinkel, S.N., Critchley, H.D., 2013. Interoception, emotion and brain: new insights link internal physiology to social behaviour. Commentary on: "Anterior insular cortex mediates bodily sensibility and social anxiety" by Terasawa et al. (2012). Social Cognitive and Affective Neuroscience 8, 231–234.

Gray, M.A., Taggart, P., Sutton, P.M., Groves, D., Holdright, D.R., Bradbury, D., Brull, D., Critchley, H.D., 2007. A cortical potential reflecting cardiac function. Proceedings of the National Academy of Sciences of the United States of America 104, 6818–6823.

Herbert, B.M., Blechert, J., Hautzinger, M., Matthias, E., Herbert, C., 2013. Intuitive eating is associated with interoceptive sensitivity. Effects on body mass index. Appetite 70C, 22–30.

Herbert, B.M., Herbert, C., Pollatos, O., Weimer, K., Enck, P., Sauer, H., Zipfel, S., 2012. Effects of short-term food deprivation on interoceptive awareness, feelings and autonomic cardiac activity. Biological Psychology 89, 71–79.

Herbert, B.M., Pollatos, O., 2014. Attenuated interoceptive sensitivity in overweight and obese individuals. Eating Behaviors, 445–448.

Khalsa, S.S., Rudrauf, D., Damasio, A.R., Davidson, R.J., Lutz, A., Tranel, D., 2008. Interoceptive awareness in experienced meditators. Psychophysiology 45, 671–677.

Khalsa, S.S., Rudrauf, D., Tranel, D., 2009. Interoceptive awareness declines with age. Psychophysiology 46, 1130–1136.

Leopold, C., Schandry, R., 2001. The heartbeat-evoked brain potential in patients suffering from diabetic neuropathy and in healthy control persons. Clinical Neurophysiology 112, 674–682.

McGaugh, J.L., 2000. Memory–a century of consolidation. Science 287, 248–251.

Moor, T., Mundorff, L., Bohringer, A., Philippsen, C., Langewitz, W., Reino, S.T., Schachinger, H., 2005. Evidence that baroreflex feedback influences long-term incidental visual memory in men. Neurobiology of Learning and Memory 84, 168–174.

Pollatos, O., Gramann, K., Schandry, R., 2007. Neural systems connecting interoceptive awareness and feelings. Human Brain Mapping 28, 9–18.

Pollatos, O., Herbert, B.M., Wankner, S., Dietel, A., Wachsmuth, C., Henningsen, P., Sack, M., 2011. Autonomic imbalance is associated with reduced facial recognition in somatoform disorders. Journal of Psychosomatic Research 71, 232–239.

Pollatos, O., Kirsch, W., Schandry, R., 2005. Brain structures involved in interoceptive awareness and cardioafferent signal processing: a dipole source localization study. Human Brain Mapping 26, 54–64.

Pollatos, O., Kurz, A.L., Albrecht, J., Schreder, T., Kleemann, A.M., Schopf, V., Kopietz, R., Wiesmann, M., Schandry, R., 2008. Reduced perception of bodily signals in anorexia nervosa. Eating Behaviors 9, 381–388.

Pollatos, O., Traut-Mattausch, E., Schandry, R., 2009. Differential effects of anxiety and depression on interoceptive accuracy. Depress Anxiety 26, 167–173.

Rief, W., Barsky, A.J., 2005. Psychobiological perspectives on somatoform disorders. Psychoneuroendocrinology 30, 996–1002.

Rosenberger, C., Elsenbruch, S., Scholle, A., De Greiff, A., Schedlowski, M., Forsting, M., Gizewski, E.R., 2009. Effects of psychological stress on the cerebral processing of visceral stimuli in healthy women. Neurogastroenterology and Motility 21, 740-e45.

Schachinger, H., Weinbacher, M., Kiss, A., Ritz, R., Langewitz, W., 2001. Cardiovascular indices of peripheral and central sympathetic activation. Psychosomatic Medicine 63, 788–796.

Schachter, S., Singer, J.E., 1962. Cognitive, social, and physiological determinants of emotional state. Psychological Review 69, 379–399.

Schaefer, M., Egloff, B., Witthoft, M., 2012. Is interoceptive awareness really altered in somatoform disorders? Testing competing theories with two paradigms of heartbeat perception. Journal of Abnormal Psychology 121, 719–724.

Schandry, R., 1981. Heart beat perception and emotional experience. Psychophysiology 18, 483–488.

Schandry, R., Bestler, M., Montoya, P., 1993. On the relation between cardiodynamics and heartbeat perception. Psychophysiology 30, 467–474.

Schulz, A., Ferreira De Sá, D.S., Dierolf, A.M., Lutz, A., Van Dyck, Z., Vögele, C., Schächinger, H., 2014. Short-term Food Deprivation Increases Amplitudes of Heartbeat-evoked Potentials. Psychophysiology. Accepted for publication.

Schulz, A., Lass-Hennemann, J., Nees, F., Blumenthal, T.D., Berger, W., Schachinger, H., 2009. Cardiac modulation of startle eye blink. Psychophysiology 46, 234–240.

Schulz, A., Lass-Hennemann, J., Sutterlin, S., Schachinger, H., Vogele, C., 2013a. Cold pressor stress induces opposite effects on cardioceptive accuracy dependent on assessment paradigm. Biological Psychology 93, 167–174.

Schulz, A., Plein, D.E., Richter, S., Blumenthal, T.D., Schachinger, H., 2011. Cold pressor stress affects cardiac attenuation of startle. International Journal of Psychophysiology 79, 385–391.

Schulz, A., Richter, S., Ferreira De Sá, D.S., Vögele, C., Schächinger, H. Cardiac Modulation of Startle Is Not Affected by Intravenously Administered Cortisol, submitted for publication.

Schulz, A., Strelzyk, F., Ferreira De Sa, D.S., Naumann, E., Vogele, C., Schachinger, H., 2013b. Cortisol rapidly affects amplitudes of heartbeat-evoked brain potentials-implications for the contribution of stress to an altered perception of physical sensations? Psychoneuroendocrinology 38, 2686–2693.

Shao, S., Shen, K., Wilder-Smith, E.P., Li, X., 2011. Effect of pain perception on the heartbeat evoked potential. Clinical Neurophysiology 122, 1838–1845.

Terhaar, J., Viola, F.C., Bar, K.J., Debener, S., 2012. Heartbeat evoked potentials mirror altered body perception in depressed patients. Clinical Neurophysiology 123, 1950–1957.

Vaitl, D., 1996. Interoception. Biological Psychology 42, 1–27.

Wiens, S., 2005. Interoception in emotional experience. Current Opinion in Neurology 18, 442–447.

Willem Van Der Does, A.J., Antony, M.M., Ehlers, A., Barsky, A.J., 2000. Heartbeat perception in panic disorder: a reanalysis. Behaviour Research and Therapy 38, 47–62.

Interorganizational Relationships and Networks

Mark Ebers, Faculty of Management, Economics and Social Sciences, University of Cologne, Koeln, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Research on interorganizational relationships and networks studies overall properties of relations between and among individuals, groups, or organizations, their antecedents, development processes, and outcomes. Focusing mainly on the organizational level, this article highlights why organizations forge different forms of relationships and what consequences result from the interorganizational links and the positions of organizations within their networks of relationships.

Interorganizational relationships refer to longer-term relationships between and among organizations (e.g., suppliers, customers, competitors, trade associations, and public sector organizations) that are pursuing a mutual interest while also remaining independent and autonomous, thus retaining separate interests. They can take a variety of forms, such as long-term sourcing agreements, partnerships, consortia, joint ventures, or alliances. Research on interorganizational relationships is concerned with properties and overall patterns of relations between and among organizations, their origins, the processes through which they are established, maintained, changed, and dissolved as well as with the consequences associated with such relationships (see Cropper et al., 2008). Interorganizational networks among individuals, groups, or organizations represent the overall patterns that evolve from, constitute, and foster such relationships (Knoke, 2012). All of the approaches in research on interorganizational relations and networks basically create their distinctive frames of enquiry from two core building blocks: a set of attributes describing the relating organizations and a set of attributes describing the nature of the relationships through which they are linked.

This article highlights why organizations forge different forms of relationships and what consequences result from the interorganizational links and the positions of organizations within their networks of relationships. Since it primarily concentrates on organization-level phenomena, it mainly highlights research that originates in, or has been adopted by, organization science. However, it should be noted that other disciplines have made significant contributions to our understanding of interorganizational relationships and networks, for instance, political science, economics, social psychology, economic geography, and legal studies. This article will highlight three major research traditions and their most important findings: first, the research tradition grounded in strategic management and organization science studying mainly bilateral relationships between organizations; second, the two research traditions devoted to interorganizational networks. Social network analysis is mainly founded in social psychology and sociology. The study of clusters organizations, i.e., local agglomerations of cospecialized firms and other institutions, primarily is based in economics, economic geography, and sociology. While the boundaries separating these research traditions lately seem to have become more permeable, as scholars mutually recognize and adopt concepts, methods, and research findings, the following sections will nevertheless discuss the three noted research traditions separately, as sufficient differences remain between them.

Interorganizational Relationships

The phenomenon of interorganizational collaboration is not new. Since the 1980s, however, one could witness in many industries a fast proliferation of joint ventures, alliances, consortia, and other forms of collaboration that has triggered a burgeoning literature. Among the prominent research questions addressed by this research are the following: Why do organizations engage in interorganizational cooperation, and when? With whom are organizations likely to ally, and why? How do organizations govern and manage their cooperative relationships? What are the outcomes and implications of particular forms of interorganizational relationships for the participating organizations and for third parties?

Motives

Actors' motives for engaging in interorganizational relationships can be quite varied (see Glaister and Buckley, 1996). Some firms cooperate in order to collude against common rivals or reduce competition by binding competitors as allies. Others use cooperation in an attempt to access needed resources and/or capabilities in order to enhance their competitiveness, e.g., in terms of improved products, better or wider market access, or faster market entry. Moreover, the formation of interorganizational relationships is seen as a means of fast, effective, and efficient learning, thus short-circuiting the process of acquiring and appropriating knowledge. In addition, cost savings in various ways can motivate the formation of interorganizational relationships. Through joint research, procurement, production, or marketing, firms strive to realize economies of scale and/or scope. Furthermore, firms can also cooperate in order to achieve specialization benefits, when they decide not to engage in the full spectrum of a particular product's value chain but to specialize in particular stages of production, and then to cooperate in the provision of the final product. A better information flow and improved coordination of resource flows among cooperating firms can also lead to cost and time savings. Finally, reduction of risk and uncertainty represent motives for forming interorganizational relationships when organizations seek to spread financial or other risks, for

example, when shouldering (mostly large) innovations or other projects (Nohria and Eccles, 1992).

Forms of Interorganizational Relationships

There exists considerable variation in the formal and informal structure of interorganizational relationships. Research has suggested numerous and diverging conceptualizations and measurements in order to capture this variety. Perhaps the most widely used conceptualization refers to the contractual and organizational arrangements by which collaborating firms govern their relationship. Specifically, different forms of collaboration are distinguished on the basis of the distribution of ownership- and decision-rights between the parties, and the coordination mechanisms employed in the relationship. In this vein, research distinguishes more hierarchical, hybrid and more marketlike forms of governance (Ebers, 1997; *see* Hierarchies and Markets). Other conceptualizations focus on the collaborators' motives, the types of resource interdependencies that exist between the collaborating organizations, or on the norms, rules, and processes that characterize the collaboration.

Consequences of Interorganizational Relationships

Depending on the specific rationale and design of a relationship, interorganizational collaboration may have implications for a number of organizational outcomes. Most prominently, research has studied how through their interorganizational relationships organizations enhance their innovativeness through mutual learning, augment their resource base and capabilities through privileged access to their partners' knowledge and resources, enhance their reputation through association with reputed partners, or improve their cost position through better coordination of interdependencies and production cost savings (Oliver, 2009). However, as a consequence of collaborating, firms may also experience an erosion of their competitive position by imitating or faster-learning partners. Moreover, lock-in effects of collaboration may impede necessary adjustments due to partner-specific investments and limited access to new information. In addition, a number of studies have addressed indirectly some consequences of interorganizational relationships by studying the stability and termination of interorganizational relationships (Doz and Hamel, 1998).

Theoretical Approaches and Methods

As in the field of organization studies as a whole, research examining when, why, how, and to what effects organizations engage in interorganizational relationships has employed a number of different approaches and theories. Most prominently, these include resource dependence theory, organizational economics, evolutionary theory, industrial marketing and purchasing, strategic management, neo-institutional theory, critical perspectives, and the management perspective theory.

Resource dependence theory views interorganizational relations as a means for securing needed resources and of enhancing control over resource supply. Power relations and the effective management of resource dependencies are at the center of researchers' interest. Organizational economics approaches, mostly informed by agency or transaction cost theory (*see* Transaction Costs and Property Rights), offer comparative assessments of various forms of interorganizational relationships. These stress the economizing effects of particular arrangements that result from an alignment of particular transaction and actor attributes with contractual and organizational features of the relationship. Evolutionary theory approaches seek to explain processes of boundary emergence, maintenance, shift, and erosion around interorganizational relationships. It analyzes interorganizational relationships through the generic conceptual scheme of variation, selection, and retention (Carroll and Hannan, 2000). Industrial marketing and purchasing approaches, often with regard to technological collaboration, study how the resource, informational and social interdependencies that exist between the collaborating organizations impact on their behavior (Håkansson and Snehota, 1995). Scholars from the field of strategic management focus on the exchange and allocation of resources and information within interorganizational relationships, and on how interorganizational relationships can be utilized for improving a firm's conduct and profitability (Jarillo, 1993). Neo-institutional theory, by contrast, studies how legitimacy pressures, social institutions, and institutionalized behaviors shape the forms and functioning of interorganizational relationships (Strang and Sine, 2002). Critical perspectives on interorganizational relations highlight how divergent interests and asymmetrical power relations among organizations, including the political use of language, shape interorganizational relations and influence their outcomes (Hardy and Phillips, 1998). The management perspective on interorganizational relationships focuses on the series of processes undertaken by a team of individuals, with various skills and capabilities, that shape both the direction to be taken by the collaborating organizations and the allocation and implementation of resources toward those ends (Huxham and Vangen, 2005).

With regard to methods employed, case study and survey research dominate the study of interorganizational relationships. Longitudinal studies are still rather rare, as are experiments and simulations.

Interorganizational Networks: Social Network Analysis

Social network analysis views interorganizational networks as a set of linkages (e.g., resource, friendship, informational ties) among a set of actors (individuals, groups, or organizations). Social network analysis offers various algorithms designed to describe and analyze positions of nodes within, and structural properties of, interorganizational networks. The basic aim of this research tradition is to understand how the overall structure of an interorganizational network, and an actor's position within that network, provides opportunities and constraints for actors' behavior (*see* Network Analysis). Compared to research on interorganizational relationships, this strand of research thus tends to pay more attention to the overall embeddedness of organizational and individual actors in networks of relationships than to individual attributes of the actors and of their individual relations.

Social network analysis suggests two broad ways in which overall structural features of social networks influence actors' behavior. First, particular network structures and positions have implications for the information that is available to actors and thus shape their decision-making and behavior. Second, particular network structures and positions in various ways invest actors with, or deprive them of, power to exert control over other actors. Such power may stem, for example, from utilizing informational advantages, privileged access to other actors, invoking obligations associated with particular relations, or from being able to mobilize sanctions by third parties (see Burt, 1992).

Principal Methods of Network Analysis

Simulation and experiment represent important methods within social network analysis. However, more and more social network analysts have begun to complement these traditional methods with survey research and sometimes ethnographic analysis. Important concepts that have been explored include network and actor centrality (i.e., measures of the connectedness of actors within a network); measures concerning structural properties of networks such as the density of a network (indicating the number of established ties in relation to the possible ties among the actors in a network); and the concept of clique (denoting a subset of actors within a network all, or most, of whom are directly tied to one another); or similarity measures: for instance, structural equivalence (two actors occupy structurally equivalent positions if they entertain identical linkages with the same third parties in the network) (see Wasserman and Faust, 1994).

Major Research Subjects

Prior to the 1980s, the bulk of research in social network analysis concentrated on developing and analyzing mathematical models of social structures and ensuing behaviors. Since then, however, there has been a proliferation of studies that apply these models to empirical phenomena (see Grabher and Powell, 2004). Two issues in particular have received considerable attention: interlocking directorates and social capital.

Interlocking Directorates

This research studies how the composition of firms' boards of directors influences firm behavior. If firms' directors sit on the boards of other firms, the directors of which in turn sit on the boards of other firms, this creates interlocking directorates. There exist several competing views as to the role that such interlocks play. However, many scholars suggest that interlocks represent an indicator of social relations among organizations, and of the extent to which firms are integrated into particular networks. Research has shown that interlocking directorates influence, for example, the diffusion of practices, executive pay, the use of 'greenmail' and of 'golden parachutes' policies for top management, the prevalence of structural defenses against takeover attempts, participation in mergers and acquisitions, and other elements of firm strategy such as the extent of research and development and the form of firm financing (see Mizruchi, 1996).

Social Capital

The notion of social capital originates from community studies that highlighted the role of strong, intense personal relationships among the members of a local community that invested them with mutual trust and the power for collective action. Social capital endows actors with a credential that entitles them to credit, and with a resource that can be utilized in action. At least two competing conceptualizations of social capital can be distinguished. A structural view sees social capital as a function of the information and control advantages that an actor enjoys who is in a broker position between otherwise unconnected actors; this position exploits a 'structural hole' in a network (Burt, 1997). A relational view sees social capital as a function of the intensity and extensiveness of personal bonds among actors. In this perspective, strong ties among actors, intense relationships characterized by mutual trust, a sense of obligation, common norms, and expectations, create social capital on which the respective actors can draw (Coleman, 1990).

In examining the consequences of social capital, research has identified two main effects. First, social capital can be utilized for increasing the efficiency of action through increasing the efficacy of information diffusion. Strong ties and intense relationships among actors that share many characteristics make information exchange easier and less ambiguous; likewise, exploiting the position of a structural hole enables a more efficient diffusion of information, because redundant contacts are minimized. Second, social capital can be utilized as a productive resource that enhances actors' goal achievement. This can be realized, for instance, by exploiting the position of a structural hole for learning about, and exercising control over, more rewarding opportunities, for instance, with regard to career advancement or firm profit. Likewise, in densely connected networks characterized by strong ties among actors, information about actors' behavior is disseminated quickly and evaluated on the basis of similar norms; thus, behavior can readily be socially sanctioned, cooperation is enabled, and opportunism constrained. Interorganizational networks whose members command significant relational social capital were thus found to thrive and prosper.

Interorganizational Networks: Clusters of Organizations

It has long been a well-known phenomenon that firms are not evenly distributed in geographic space but to a greater or lesser extent cluster locally. From the streets and quarters occupied by specialized trades in medieval towns, to the diamond-cutting industry of Antwerp and Tel Aviv; the coal and steel industry in the Ruhr, Lothringia, and Silesia regions; the cutlery industry in Sheffield and Solingen; the financial services industry in London and New York City; to the software industry in California's Silicon Valley, firms during different periods of time have formed regional or local clusters. While regional interorganizational networks thus by no means are a novel

phenomenon, research interest in regional interorganizational networks since its inception in the late 1800s (Marshall, 1920) has experienced a revival beginning in the 1980s.

The Formation of Clusters

The literature identifies two broad reasons why firms, together with some supporting organizations, cluster regionally or locally. First, through clustering firms can realize agglomeration economies. These result from both urbanization advantages and localization advantages. Urbanization advantages include economies of scale with regard to a shared infrastructure (e.g., for communication or transport), economies of scope (e.g., technological spillovers and informational externalities), specialization benefits (leading to technological or skill advantages), and transaction cost savings (e.g., reduced search costs for buyers and suppliers) (Krugman, 1995). Localization benefits comprise most importantly easy access to locally concentrated resource inputs (e.g., raw materials, specialized skills, specific attitudes of a local workforce, or colocated suppliers). A related line of reasoning stresses that clustering among firms is enabled and supported by regionally embedded institutions such as chambers of commerce, employers' unions, banks, science parks, universities, and training centers. These institutions often provide important resources, e.g., capital or access to distribution channels and potential customers. Moreover, they often act as informational brokers that support the exchange of information among firms, encourage and facilitate mutual learning, and can thus foster the responsiveness, adaptability, and innovativeness of networked firms.

Second, relational advantages emanating from the business and social relationships among geographically concentrated firms and individuals can also foster clustering. Vigorous competition from nearby competitors and face-to-face contact with demanding local customers can put pressure on the competitiveness and innovativeness of firms. Frequent contact with employees from rival firms and their products can further mutual learning. Family and friendship ties among local business people, common membership in local trade associations, sports clubs, and political institutions can be a source of knowledge and learning and provide better access to resources, e.g., capital and political influence that are required for successful business development. Moreover, intense social ties, behavioral norms, and routines may smooth cooperation, enable the exchange of tacit knowledge, encourage a ready and more trustworthy exchange of information among network members, and may thus support mutual learning and innovation (Porter, 1990).

Outcomes of Local Clusters

In line with these driving forces, research has identified a number of consequences associated with regional clusters of firms. Perhaps the most widely debated outcome concerns the increased competitiveness of regional cluster firms. Two major driving forces can be identified. On the one hand, clusters enhance competition through intense competitive pressure from nearby rivals and customers; this puts high pressure on firms and their employees to compete. On the other hand, localization and relational benefits provide competitive advantage over nonclustered rivals.

A second theme widely discussed in the literature concerns learning and innovation advantages. Local proximity and shared norms and traditions are said to allow for better cooperation even among rival firms, and to promote technology and informational spillovers. Together with competitive pressure, this can contribute to enhanced incentives as well as opportunities for learning and innovation. In addition, it is noted that clusters of small cospecialized firms that can flexibly recombine their resources and capabilities can be lighter on their feet than larger, more integrated competitors. Collectively, these firms are therefore said to be able to adapt more quickly to changes in demand, both quantitatively and qualitatively (Piore and Sabel, 1984).

Both postulated effects make regional clusters a favored theme in discourses on regional economic development. Endogenous growth and development of regions as well as sustained regional competitiveness are among the promises that have been associated with regional clusters (Staber et al., 1996).

However, there is evidence from research that the above-noted effects of regional clusters can, but need not necessarily, arise. There exist numerous examples of clusters that thrived at one point in time, only to fail later. Many coal and steel regions in the USA and Western Europe have shared this fate, as has the once admired minicomputer cluster along Route 128 in Massachusetts, USA. Several factors may contribute to the decline of regional clusters. Some of these are inherent to the functioning of regional clusters, some associated with external developments. They include the following: failing demand for a cluster's product, obsolescence of particular modes of production, competition from similar clusters in other regions, loss of the ability to coordinate cospecialized activities, and inertia due to a loss of developmental abilities. As this debate about the consequences of regional clusters for the participating firms and their respective regions shows, performance evaluations of business networks and regional districts are extremely difficult and problematic.

Methods and Approaches

Most research to date builds on case study evidence. A large number of conceptual studies engage in transferring more general insights from particular disciplines, for instance, economics or sociology, to the field of regional clusters.

At the end of the nineteenth century, economics was among the first disciplines to analyze the phenomenon of regional clusters. Paradoxically, the increasing globalization in the 1980s revived interest in internationally competitive, regional clusters, both in economics and economic geography (Porter, 1990; Krugman, 1995). Evolutionary economics developed a particular interest for the learning and innovation characteristics of regional clusters (Nooteboom, 2000) that is also shared by the innovative milieu approach (Maillat, 1991). Industrial sociology analyzed regional networks employing the labels of industrial districts and new industrial spaces (Storper, 1997). Institutional theory sociologists emphasized the social and institutional structures underlying regional clusters (Herrigel, 1996).

Conclusion

Over the last decades, interorganizational relationships and networks have become considerably more prominent in both practice and research. Pertinent studies have mainly emphasized the benefits that organizations experience when participating in various forms of supraorganizational structures, yet also noted a few risks. As interorganizational relationships and networks are scrutinized by many disciplines focusing on different phenomena, applying different theories and methods, this field of study has been characterized as a being fragmented into different silos of research. While silo'ing offers benefits of specialization, it makes it difficult to accumulate findings across silos. However, as some research in the field is beginning to demonstrate, there exist potentially fruitful opportunities for enhanced dialogue across extant research silos, be they constituted by different disciplines, theoretical lenses, or study objects (Cropper et al., 2011). The resulting cross-fertilization could contribute to an enrichment of our understanding, as well as a consolidation of empirical findings and theoretical conclusions, concerning why, when, and how different kinds of relationships and network structures within and among organizations have which effects on organizational and higher-level outcomes.

See also: Actor-Network Theory; Boundaries and New Organization Forms; Conflict: Organizational; Cultural Production in Networks; Economic Geography; Network Analysis, History of; Network Analysis; Network Paradigm: Applications in Organizational Science; Networks, Geography of; Policy Networks; Social Capital; Social Network Algorithms and Software; Social Network Analysis; Social Network Models: Statistical; Social Networks.

Bibliography

Burt, R.S., 1992. Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge, MA.

Burt, R.S., 1997. The contingent value of social capital. Administrative Science Quarterly 42, 339–365.

Carroll, G.R., Hannan, M.T., 2000. The Demography of Corporations and Industries. Princeton University Press, Princeton.

Coleman, J.S., 1990. Foundations of Social Theory. Belknap Press of Harvard University Press, Cambridge, MA.

Cropper, S., Ebers, M., Huxham, C., Ring, P.S., 2011. Packing more punch? Developing the field of inter-organisational relations. International Journal of Strategic Business Alliances 2, 153–170.

Cropper, S., Ebers, M., Huxham, C., Ring, P.S. (Eds.), 2008. The Oxford Handbook of Inter-organizational Relations. Oxford University Press, Oxford.

Doz, Y.L., Hamel, G.R., 1998. Alliance advantage: The Art of Creating Value through Partnering. Harvard Business School Press, Boston, MA.

Ebers, M., 1997. Explaining interorganizational network formation. In: Ebers, M. (Ed.), The Formation of Interorganizational Networks. Oxford University Press, New York, pp. 3–40.

Fisher, M.M., Suarez-Villa, L., Steiner, M. (Eds.), 1999. Innovation, Networks, and Localities. Springer-Verlag, Berlin.

Glaister, K.W., Buckley, P.J., 1996. Strategic motives for international alliance formation. Journal of Management Studies 33, 301–332.

Grabher, G., Powell, W.W. (Eds.), 2004. Networks. Edward Elgar, Cheltenham.

Håkansson, H., Snehota, I. (Eds.), 1995. Developing Relationships in Business Networks. Routledge, London.

Hardy, C., Phillips, N., 1998. Strategies of engagement: lessons from the critical examination of collaboration and conflict in an interorganizational domain. Organization Science 9, 217–230.

Herrigel, G.B., 1996. Industrial Constructions: The Sources of German Industrial Power. Cambridge University Press, New York.

Huxham, C., Vangen, S., 2005. Managing to Collaborate: The Theory and Practice of Collaborative Advantage. Routledge, London.

Jarillo, J.C., 1993. Strategic Networks: Creating the Borderless Organization. Butterworth-Heinemann, Oxford, UK.

Knoke, D., 2012. Economic Networks. Polity Press, Cambridge, UK.

Krugman, P.R., 1995. Development, Geography, and Economic Theory. MIT Press, Cambridge, MA.

Lundvall, B.-A. (Ed.), 1992. National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning. Pinter Publishers, London.

Maillat, D., 1991. The innovation process and the role of milieu. In: Bergman, E., Maier, G., Tödtling, F. (Eds.), Regions Reconsidered: Economic Networks, Innovation and Local Development in Industrialized Countries. Cassel, London, pp. 103–117.

Marshall, A.E., 1920. Principles of Economics, an Introductory Volume, eighth ed. Macmillan, London.

Mizruchi, M.S., 1996. What do interlocks do? An analysis, critique and assessment of research on interlocking directorates. Annual Review of Sociology 22, 271–298.

Nohria, N., Eccles, R.G. (Eds.), 1992. Networks and Organizations: Structure, Form, and Action. Harvard Business School Press, Boston, MA.

Nooteboom, B., 2000. Learning and Innovation in Organizations and Economies. Oxford University Press, Oxford, UK.

Oliver, A., 2009. Networks for Learning and Knowledge-creation in Biotechnology. Cambridge University Press, Cambridge, UK.

Oliver, A., Ebers, M., 1998. Networking network studies: an analysis of conceptual configurations in the study of interorganizational relationships. Organization Studies 19, 549–583.

Piore, M.J., Sabel, C.F., 1984. The Second Industrial Divide: Possibilities for Prosperity. Basic Books, New York.

Porter, M.E., 1990. The Competitive Advantage of Nations. Free Press, New York.

Staber, U.H., Schaefer, N.V., Sharma, B. (Eds.), 1996. Business Networks: Prospects for Regional Development. Walter de Gruyter, Berlin.

Storper, M., 1997. The Regional World: Territorial Development in a Global Economy. Guilford Press, New York.

Strang, D., Sine, W.D., 2002. Interorganizational institutions. In: Baum, J.A.C. (Ed.), Companion to Organizations. Blackwell Publishing, Oxford, pp. 497–519.

Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications. Cambridge University Press, New York.

Interpersonal Attraction, Psychology of

Shelly Zhou, Stony Brook University, NY, USA; and University of Toronto, Toronto, ON, Canada
Marie B Chelberg and Arthur Aron, Stony Brook University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This encyclopedia article reviews the major predictors of interpersonal attraction, the desire to be around and form a friendship or romantic relationship with another person. The main predictors include physical attractiveness, desirable personality traits, similarity, familiarity, and reciprocal liking. Other factors that facilitate or limit these effects, as well as situational and extradyadic influences, are also discussed. Finally, findings related to each of these predictors are tied together under the overarching frameworks of social exchange theory and sexual strategies (evolutionary) theory.

Interpersonal attraction refers to the desire to be around and form a friendship or romantic relationship with another. This encyclopedia article reviews the literature on interpersonal attraction, discussing main findings and underlying processes, as well as facilitating and limiting conditions of its major determinants.

Physical Attractiveness

Not surprisingly, people like others who are good-looking. Research on heterosexual romantic attraction shows that people like, talk more to, are friendlier toward, and request more dates from those who are physically attractive (Walster et al., 1966). This preference exists across cultures (Buss and Schmitt, 1993) and for both sexes. Men express a greater emphasis on the physical attractiveness of potential mates than women do, but interestingly, women's implicit (i.e., not consciously recognized) preferences for physical attractiveness (as assessed with response time and date requests) are just as strong as that of men (Eastwick et al., 2011). Research indicates that looks are also important in nonromantic attraction, finding that people are more interested in potential friends who are attractive (Black, 1974). So, the liking for attractive others is quite robust, transcending gender, culture, and relationship type.

From an evolutionary perspective, people like good-looking mates because they contribute to the viability of one's offspring. One reason for this is that physically attractive people are healthier in some ways, suggesting that the preference for good-looking mates represents an attempt to find high-quality genes with which to mix one's own. For example, facial and bodily symmetry is attractive and is also associated with good health (Perilloux et al., 2010). Second, the physical features that men and women find attractive in the opposite sex correspond with the differential sexual strategies that each sex uses to augment the viability of their genes in future generations. Men, who have low minimal costs in reproduction, can have many offspring and invest little in each; they therefore prefer physical indicators of fertility in women such as youthful features and a 0.7 waist-to-hip ratio (Furnham et al., 2005; Jones, 1995). Women, who have higher minimal costs in reproduction, invest more in the survival and success of a few offspring and are therefore tuned to physical indicators (signs of status, such as a large jaw and prominent forehead; Rhodes, 2006) that a man can provide resources for the offspring.

In addition to augmenting this evolutionary explanation for heterosexual romantic attraction, a social perspective can also explain homosexual romantic and general/friendship attraction. First, being physically attractive carries social benefits, not only for oneself, but also for one's romantic partner and friends. For example, men received more favorable evaluations when they were with an attractive versus unattractive woman (Sigall and Landy, 1973), and both sexes were seen as more attractive and credible when they had attractive friends (Walther et al., 2008).

Additionally, people tend to assume a whole host of good qualities about attractive people, including that they are nicer, more successful, have better marriages, and are happier in general. This "what is beautiful is good" bias (Dion et al., 1972) may be true to some small extent; for example, when participants had a phone conversation with a target (who they could not see), they rated the physically attractive targets as more likable and socially skillful (Goldman and Lewis, 1977). The question, then, is whether these desirable attributes are inherent in physically attractive people or result from receiving preferential treatment based solely on looks. One line of research supports the latter explanation. Specifically, when judges believed that they were talking with an attractive (versus unattractive) opposite sex target, both conversation partners were warmer and more animated regardless of the target's actual attractiveness (Synder et al., 1977; Andersen and Bem, 1981). Therefore, the "what is beautiful is good" bias seems to be at least partially a self-fulfilling prophecy.

Desirable Personality

Although not as extensively studied as physical attractiveness, personality is also important for attraction. One way in which personality affects attraction is through physical attractiveness: when one is told that a target has desirable personality traits, one views them as more physically attractive, but the opposite occurs when the target has undesirable personality traits (Lewandowski et al., 2007). Personality also influences attraction directly: across genders, everyone prefers potential mates

and friends who are kind, intelligent, honest, and dependable (Anderson, 1968; Buss and Barnes, 1986).

Similarity, Differences, and Ideal Self-Attraction

Similarity

Consistently, a large body of research has supported the notion that "birds of a feather flock together," showing that people are attracted to those whom they see as having similar attitudes to oneself (e.g., Byrne et al., 1970). Most of this work used the 'bogus other' technique, in which participants read responses to an attitude questionnaire (researchers gave certain responses to vary the degree of similarity to participant responses) and rated their liking for the respondent. Using another research paradigm, people also like partners with whom they are matched on maximal similarity after in-person interactions (Black, 1974; Byrne et al., 1970). Established couples are also more similar to each other in various ways (not just attitudes) than random strangers. This preference for similar others is not restricted to romantic attraction and is just as strong in friendship attraction (Black, 1974).

There are reasons that similarity is important for attraction. On the individual level, similar others validate and reinforce one's values and beliefs, which make the person feel good about these aspects of the self and provide a sense of stability (Byrne and Clore, 1970). On the dyadic level, people get along better with others who share their attitudes, values, and background (Davis, 1981), possibly because shared worldviews ease communication and diminish conflict. Additionally, people expect similar others to like them more than nonsimilar ones (Condon and Crano, 1988), which encourages them to reciprocate this perceived attraction (we discuss 'reciprocal liking' in more detail, later in this article). Taking these together, similarity may ultimately serve to increase the probability that initial attraction will lead to successful relationship formation (Aron et al., 2006).

Another line of research has examined factors that affect the strength of the similarity-attraction effect. A robust finding is that similarity is most important for values, beliefs, attitudes, and interests; findings for a similarity effect of personality have been much weaker and less consistent (Simpson and Harris, 1994). Additionally, central aspects of the self are more important in the similarity-attraction effect than peripheral self-aspects (Clore and Baldridge, 1968). Finally, this effect disappears when the other shares an aspect of the self that one dislikes (Novak and Lerner, 1968) – this makes sense considering that this kind of interpersonal similarity would reinforce and remind people of their unwanted aspects.

It is important to note that it is perceptions of, rather than actual, similarity that drives this effect. Previous studies that matched real people based on similarity confounded perceived and actual similarity, since partners knew about this matching criterion. Studies separating these variables have found that perceived similarity indeed predicts attraction better than actual similarity (Tidwell et al., 2013; Montoya et al., 2008). These variables are distinguishable because there are motivational processes and social schema that affect perceptions of similarity, especially at this nascent stage of relationship formation when people are still becoming familiar with each other. For example, people tend to assume that others are similar to themselves until they have reasons to believe otherwise (Rosenbaum, 1986) and they perceive physically attractive others as more similar to themselves than unattractive ones (Marks et al., 1981).

Differences and Self-Expansion

According to Aron et al. (2013), people are motivated to increase their potential efficacy by acquiring new content (i.e., resources, perspectives, and identities) for the self and thereby expand the sense of who they are. One important source of this self-expansion is close relationships. Specifically, people include close others into their own sense of self and thereby behave, feel, and think, to some extent, as though the other's self-content were actually their own. When the close other differs from oneself, one can incorporate new resources, perspectives, and identities that he or she did not previously possess. So, the self-expansion model predicts that people like different others who can provide self-expansion, which contradicts the robust findings that interpersonal similarity (in the attitude and background domains) is desirable. Resolving this contradiction, recall that similarity's function is to aid in relationship formation; thus, if another factor performs this function, would one prefer differences to capitalize on its self-expanding benefits? Indeed, when told that a reliable matching algorithm indicated that they would get along well, participants preferred a bogus other with different (rather than similar) interests (Aron et al., 2006).

Ideal Self

Another, albeit small, line of research has looked at the idea that people like others who resemble their own ideal self (the person one wants to become). Studies that have separated the effects of the actual versus ideal self have favored the latter. These effects were found for romantic (Klohnen and Luo, 2003) and general attraction (Wetzel and Insko, 1982) toward bogus others as well as in established friendships (LaPrelle et al., 1990; Thompson and Nishimura, 1952). One explanation for this ideal-attraction effect comes from the Michelangelo phenomenon (Rusbult et al., 2009), an effect whereby close partners support each other's goal pursuits (by modeling, reinforcing, and prompting desired behaviors) and thereby move each other toward their respective ideal selves. Given that partners resembling one's ideal self perform better in this role (Rusbult et al., 2009), liking these individuals represents an attempt to best engage the Michelangelo phenomenon and enjoy its self and dyadic benefits.

An extension of the self-expansion model attempts to consolidate the roles of ideal-similarity with those of actual similarity and differences in attraction. According to Zhou et al. (2008), people wish to expand toward their ideal self, acquiring resources, perspectives, and identities that are ideal-consistent. So, where the actual and ideal selves are disparate, one would prefer different others because they offer coveted self-aspects that the individual lacks. However, on self-aspects where the actual and ideal selves are close, one would prefer similar others to reinforce this desired proximity. Again, research on the ideal self's role in attraction is nascent, and

should be taken as speculative and a promising direction for future research.

Reciprocal Liking (Liking People Who Like You)

Another influential idea is that people like others who like them. Research on reciprocal liking has indeed found that people like others more when told that the other likes them, versus when no given information or when told that the other dislikes them (Backman and Secord, 1959). This effect exists for romantic as well as nonromantic attraction and also extends to intergroup attraction (Curtis and Miller, 1986; Lowe and Goldstein, 1970).

There are several reasons that people reciprocate others' liking for them. First, it decreases the probability of rejection and increases the belief that this other will be a source of pleasant interactions. Although these factors are themselves motivating, they also (like interpersonal similarity) increase the probability of successful relationship formation. Another explanation for reciprocal liking comes from balance theory (Heider, 1958), which states that one is motivated to maintain consistency among one's attitudes toward the self, other, and an object. In the case of reciprocal liking, the self is also the object, so if both one and the other like the self, one is motivated to also like the other to create or maintain the desired consistency.

Most research into the facilitators and limitations of reciprocal liking has been in the romantic domain. First, speed-dating studies suggest that reciprocal liking is restricted to people who like oneself uniquely, as opposed to those who like everyone (Eastwick et al., 2007). The strength of reciprocal liking also increases with the target's physical attractiveness (Greitemeyer, 2010). Interestingly, some evidence suggests that reciprocal liking peaks at moderate levels of other-liking and wanes beyond that. This may be because moderate other-liking adequately staves off rejection concerns and simultaneously creates some uncertainty that encourages thinking about the target (Whitchurch et al., 2011). Another relevant idea is gain–loss theory (Aronson and Linder, 1965), which proposes that reciprocal liking is generally stronger if the other went from disliking to liking oneself, rather than if the other liked oneself all along. The appeal of the former case may come from interpreting that the attitude change represents a successful attempt at 'winning the other over.'

Familiarity

Another major determinant of attraction is familiarity, the idea that one likes others to the extent that one has spent time with and knows them. This increase in attraction, interestingly, occurs even when interactions are superficial or passive and in the absence of any information about the target. For example, this effect emerged when participants' only exposure to the target was that she sat in their class (without talking to anyone); students liked her more when she attended class more frequently (Moreland and Beach, 1992). This effect also occurs when repeatedly viewing a stranger's photograph, even when it is presented subliminally (Bornstein et al., 1987; Zajonc, 1968). The familiarity-attraction effect is a specific application of Zajonc's mere exposure effect, the idea that simple exposure to any stimulus can increase liking for it.

Related to the familiarity effect is the propinquity effect, the idea that people like those who are physically close to them. Whether this proximity refers to how close someone sits to them in class or how close someone lives to them, the effect is robust and can be explained by the increased frequency at which people who are physically proximate see each other (i.e., repeated exposure).

Given that familiarity is an important factor in interpersonal attraction, a surprisingly little amount of work has gone into exploring its intricacies. (Much more work has been done on the overarching idea of mere exposure, but that work has generally not used social stimuli, and so will not be discussed here.) The most likely reason for this effect is that, in the absence of other information, familiarity signals that a person is safe. This is likely because, having lived in small groups during our ancestral past, humans were in close proximity to and frequently saw other people with whom they were interdependent. Therefore, throughout evolutionary history, familiarity and propinquity may have developed as signals of the trust and camaraderie associated with the in-group. Of course, it is likely that subsets of group members disliked and merely tolerated each other, making repeated exposure increasingly more discomfiting rather than pleasant. Indeed, one limit to this effect is that exposure needs to be at least neutral, and repeated exposure in unpleasant contexts can decrease liking (Perlman and Oskamp, 1971).

Situational Influences on Interpersonal Attraction

So far, the discussion has centered on determinants that are inherent in the dyad. However, there are also extradyadic and situational factors that can increase one's attraction to a target. For example, after seeing pictures of highly attractive women, men rated other women and even their own romantic partners as less physically attractive than did men who viewed images of average-looking women (Kenrick et al., 1989). A similar effect exists for women: after seeing highly dominant men, women are less satisfied with their own romantic relationships (Kenrick et al., 1994).

Another situational influence is the 'closing time effect,' which refers to the tendency for singles to perceive opposite-sex bar and nightclub patrons as more attractive as closing time nears (Pennebaker et al., 1979; Madey et al., 1996). The explanation of this effect comes from Commodity Theory (Brock, 1968), which states that the value of a resource increases with its scarcity. Indeed, as the night draws to a close and patrons are about to leave, potential partners become more scarce and thus more attractive.

Another factor is the influence of arousal on attraction, whereby romantic or sexual attraction can come from physiological arousal caused by an external source (Foster et al., 1998). This idea is based on Schachter and Singer's (1962) notion that emotions result from interpretations of physiological arousal. Specifically, when one becomes aroused, he or she searches the environment to locate the cause of this arousal and then uses this information to identify the particular emotion they are feeling. For example, arousal can be

experienced as fear (if a lion is running toward you) or as attraction (if an appropriate romantic target is nearby). In applying this idea, Dutton and Aron (1974) proposed specifically that arousal from one source may be misinterpreted as romantic/sexual attraction if (1) an appropriate target is close by and (2) the true source of the arousal is not immediately obvious. Indeed, their study found that men were more attracted to a female research assistant while they crossed a shaky bridge over a high canyon (which was presumably anxiety-inducing and physiologically arousing) rather than a stable bridge over a low rivulet.

Another idea is that the opinions of same-sex peers (or competitors) influence one's own perceptions of opposite-sex targets. When female participants see another woman smiling at a man, they find him more attractive than if the woman looked at him with a neutral expression (Jones et al., 2007). Furthermore, in a speed-dating study, both men and women were more attracted to potential partners if they saw that other people were romantically interested in these targets (Place et al., 2010). The influence of peers also extends to negative evaluations: knowing that other women found a man unattractive caused female participants to decrease their attraction toward this target (Graziano et al., 1993).

Conclusions

This encyclopedia article has reviewed several determinants of interpersonal attraction, including physical attractiveness, desirable personality traits, similarity, differences, ideal self, reciprocity, familiarity, and several extradyadic factors. As summarized here, many of these predictors of attraction, each with unique moderators and underlying motivations, tie together nicely into the overarching frameworks of social exchange and evolutionary theory.

Social exchange theory, the idea that people view relationships in economic terms of exchanging rewards and costs (Thibaut and Kelley, 1959), neatly ties together the major predictors of attraction. This is because, as previously discussed, selecting relationship partners using these criteria proffers certain psychological and social benefits. In sum, one likes similar others because they validate and reinforce the self, whereas one likes different others because they offer opportunities for self-expansion. People also like others who are ideal-consistent because they help move the self toward its ideal and familiar others because they signal safety. People also like others who like them because it boosts self-esteem and achieves a desired balance in attitudes. It is also important to remember that relationships are costly, requiring (material and psychological) resources and putting one at risk for rejection. Therefore, according to social exchange theory, interpersonal attraction is a function of the net gains versus losses that one perceives the target is likely to provide.

The evolutionary perspective of sexual strategies theory (Buss and Schmitt, 1993) also explains many of these findings. Physical attractiveness strongly predicts interpersonal attraction because it signals the ability to produce healthy offspring. Kindness and loyalty are important because people with these traits make good spouses and parents for one's future offspring. Throughout much of human history, people lived together in small tribal groups and knew one another; thus, similarity and familiarity may indicate that the person 'belongs to our group' and can provide safety and security. Ultimately, all romantic attraction serves the evolutionary purpose of obtaining the best mate possible for reproduction and parenting of offspring that will survive and successfully carry on one's genes.

See also: Hooking up and Casual Sex; Human Mate Choice, Evolution of; Human Mate Choice; Love and Intimacy, Psychology of; Online Dating; Social Relationships in Adulthood.

Bibliography

Andersen, S.M., Bem, S.L., 1981. Sex typing and androgyny in dyadic interactions: individual differences in responsiveness to physical attractiveness. Journal of Personality and Social Psychology 41, 74–86.

Anderson, N.H., 1968. Likableness ratings of 555 personality-trait words. Journal of Personality and Social Psychology 9, 272–279.

Aron, A., Lewandowski, G.W., Mashek, D., Aron, E.N., 2013. The self-expansion model of motivation and cognition in close relationships. In: Simpson, J.A., Campbell, L. (Eds.), Oxford Handbook of Close Relationships. Oxford University Press, New York.

Aron, A., Steele, J.L., Kashdan, T.B., Perez, M., 2006. When similars do not attract: tests of a prediction from the self-expansion model. Personal Relationships 13 (4), 387–396.

Aronson, E., Linder, D., 1965. Gain and loss of esteem as determinants of interpersonal attractiveness. Journal of Experimental Social Psychology 1 (2), 156–171.

Backman, C.W., Secord, P.F., 1959. The effect of perceived liking on interpersonal attraction. Human Relations 12, 379–384.

Black, H.K., 1974. Physical attractiveness and similarity of attitude in interpersonal attraction. Psychological Reports 35 (1), 403–406.

Bornstein, R.F., Leone, D.R., Galley, D.J., 1987. The generalizability of subliminal mere exposure effects: influence of stimuli perceived without awareness on social behavior. Journal of Personality and Social Psychology 53 (6), 1070–1079.

Brock, T.C., 1968. Implications of Commodity Theory for value change. In: Greenwald, A.G., Brock, T.C., Ostrom, T.M. (Eds.), Psychological foundations of attitudes. Academic Press, New York, pp. 243–275.

Buss, D.M., Barnes, M., 1986. Preferences in human mate selection. Journal of Personality and Social Psychology 50 (3), 559–570.

Buss, D.M., Schmitt, D.P., 1993. Sexual strategies theory: an evolutionary perspective on human mating. Psychological Review 100 (2), 204–232.

Byrne, D., Clore, G.L., 1970. A reinforcement model of evaluative responses. Personality. An International Journal 1 (2), 103–128.

Byrne, D., Ervin, C.R., Lamberth, J., 1970. Continuity between the experimental study of attraction and real-life computer dating. Journal of Personality and Social Psychology 16 (1), 157–165.

Clore, G.L., Baldridge, B., 1968. Interpersonal attraction: the role of agreement and topic interest. Journal of Personality and Social Psychology 9 (4), 340–346.

Condon, J.W., Crano, W.D., 1988. Inferred evaluation and the relation between attitude similarity and interpersonal attraction. Journal of Personality and Social Psychology 54 (5), 789–797.

Curtis, R.C., Miller, K., 1986. Believing another likes or dislikes you: behaviors making the beliefs come true. Journal of Personality and Social Psychology 51 (2), 284–290.

Davis, D., 1981. Implications for interaction versus effectance as mediators of the similarity-attraction relationship. Journal of Experimental Social Psychology 17 (1), 96–117.

Dion, K., Berscheid, E., Walster, E., 1972. What is beautiful is good. Journal of Personality and Social Psychology 24 (3), 285–290.

Dutton, D.G., Aron, A.P., 1974. Some evidence for heightened sexual attraction under conditions of high anxiety. Journal of Personality and Social Psychology 30 (4), 510–517.

Eastwick, P.W., Eagly, A.H., Finkel, E.J., Johnson, S.E., 2011. Implicit and explicit preferences for physical attractiveness in a romantic partner: a double dissociation in predictive validity. Journal of Personality and Social Psychology 101 (5), 993–1011.

Eastwick, P.W., Finkel, E.J., Mochon, D., Ariely, D., 2007. Selective versus unselective romantic desire: not all reciprocity is created equal. Psychological Science 18 (4), 317–319.

Foster, C.A., Witcher, B.S., Campbell, W.K., Green, J.D., 1998. Arousal and attraction: evidence for automatic and controlled processes. Journal of Personality and Social Psychology 74 (1), 86–101.

Furnham, A., Petrides, K.V., Constantinides, A., 2005. The effects of body mass index and waist-to-hip ratio on ratings of female attractiveness, fecundity, and health. Personality and Individual Differences 38 (8), 1823–1834.

Goldman, W., Lewis, P., 1977. Beautiful is good: evidence that the physically attractive are more socially skillful. Journal of Experimental Social Psychology 13 (2), 125–130.

Graziano, W.G., Jensen-Campbell, L.A., Shebilske, L.J., Lundgren, S.R., 1993. Social influence, sex differences, and judgments of beauty: putting the interpersonal back in interpersonal attraction. Journal of Personality and Social Psychology 65 (3), 522–531.

Greitemeyer, T., 2010. Effects of reciprocity on attraction: the role of a partner's physical attractiveness. Personal Relationships 17 (2), 317–330.

Heider, F., 1958. The Psychology of Interpersonal Relations. John Wiley & Sons, Hoboken, NJ.

Jones, B.C., DeBruine, L.M., Little, A.C., Burriss, R.P., Feinberg, D.R., 2007. Social transmission of face preferences among humans. Proceedings of the Royal Society Biological Sciences 274 (1611), 899–903.

Jones, D., 1995. Sexual selection, physical attractiveness, and facial neoteny: cross-cultural evidence and implications. Current Anthropology 36 (5), 723–748.

Kenrick, D.T., Gutierres, S.E., Goldberg, L.L., 1989. Influence of popular erotica on judgments of strangers and mates. Journal of Experimental Social Psychology 25 (2), 159–167.

Kenrick, D.T., Neuberg, S.L., Zierk, K.L., Krones, J.M., 1994. Evolution and social cognition: contrast effects as a function of sex, dominance, and physical attractiveness. Personality and Social Psychology Bulletin 20 (2), 210–217.

Klohnen, E.C., Luo, S., 2003. Interpersonal attraction and personality: what is attractive - self similarity, ideal similarity, complementarity, or attachment security? Journal of Personality and Social Psychology 85 (4), 709–722.

LaPrelle, J., Hoyle, R.H., Insko, C.A., Bernthal, P., 1990. Interpersonal attraction and descriptions of the traits of others: ideal similarity, self similarity, and liking. Journal of Research in Personality 24 (2), 216–240.

Lewandowski Jr., G.W., Aron, A., Gee, J., 2007. Personality goes a long way: the malleability of opposite-sex physical attractiveness. Personal Relationships 14 (4), 571–585.

Lowe, C.A., Goldstein, J.W., 1970. Reciprocal liking and attributions of ability: mediating effects of perceived intent and personal involvement. Journal of and Social Psychology 16 (2), 291–297.

Marks, G., Miller, N., Maruyama, G., 1981. Effect of targets' physical attractiveness on assumptions of similarity. Journal of Personality and Social Psychology 41 (1), 198–206.

Madey, S.F., Simo, M., Dillworth, D., Kemper, D., Toczynski, A., Perella, A., 1996. They do get more attractive at closing time, but only when you are not in a relationship. Basic and Applied Social Psychology 18 (4), 387–393.

Montoya, R.M., Horton, R.S., Kirchner, J., 2008. Is actual similarity necessary for attraction? A meta-analysis of actual and perceived similarity. Journal of Social and Personal Relationships 25 (6), 889–922.

Moreland, R.L., Beach, S.R., 1992. Exposure effects in the classroom: the development of affinity among students. Journal of Experimental Social Psychology 28, 255–276.

Novak, D.W., Lerner, M.J., 1968. Rejection as a consequence of perceived similarity. Journal of Personality and Social Psychology 9 (2), 147–152.

Pennebaker, J.W., Dyer, M.A., Caulkins, R.S., Litowitz, D.L., Ackreman, P.L., Anderson, D.B., McGraw, K.M., 1979. Don't the girls get prettier at closing time: a country and western application to psychology. Personality and Social Psychology Bulletin 5 (1), 122–125.

Perilloux, H.K., Webster, G.D., Gaulin, S.J.C., 2010. Signals of genetic quality and maternal investment capacity: the dynamic effects of fluctuating asymmetry and waist-to-hip ratio on men's ratings of women's attractiveness. Social Psychological and Personality Science 1 (1), 34–42.

Perlman, D., Oskamp, S., 1971. The effects of picture content and exposure frequency on evaluations of Negroes and Whites. Journal of Experimental Social Psychology 7 (5), 503–514.

Place, S.S., Todd, P.M., Penke, L., Asendorpf, J.B., 2010. Humans show mate copying after observing real mate choices. Evolution and Human Behavior 31, 320–325.

Rhodes, G., 2006. The evolutionary psychology of facial beauty. In: Fiske, S.T., Kazdin, A.E., Schacter, D.L. (Eds.), Annual Review of Psychology, Vol. 57. Annual Reviews, Palo Alto, CA, pp. 199–226.

Rosenbaum, M.E., 1986. The repulsion hypothesis: on the nondevelopment of relationships. Journal of Personality and Social Psychology 51 (6), 1156–1166.

Rusbult, C.E., Finkel, E.J., Kumashiro, M., 2009. The Michelangelo phenomenon. Current Directions in Psychological Science 18 (6), 305–309.

Rusbult, C.E., Kumashiro, M., Kubacka, K.E., Finkel, E.J., 2009. The part of me that you bring out: ideal similarity and the Michelangelo phenomenon. Journal of Personality and Social Psychology 96 (1), 61–82.

Schachter, S., Singer, J., 1962. Cognitive, social, and physiological determinants of emotional state. Psychological Review 69 (5), 379–399.

Sigall, H., Landy, D., 1973. Radiating beauty: effects of having a physically attractive partner on person perception. Journal of Personality and Social Psychology 28 (2), 218–224.

Simpson, J.A., Harris, B.A., 1994. Interpersonal attraction. In: Weber, A.L., Harvey, J.H. (Eds.), Perspectives on close relationships. Allyn & Bacon, Needham Heights, MA, pp. 45–66.

Snyder, M., Tanke, E.D., Berscheid, E., 1977. Social perception and interpersonal behavior: on the self-fulfilling nature of social stereotypes. Journal of Personality and Social Psychology 35 (9), 656–666.

Thibaut, J.W., Kelley, H.H., 1959. The Social Psychology of Groups. Wiley, New York.

Thompson, W.R., Nishimura, R., 1952. Some determinants of friendship. Journal of Personality 20, 305–314.

Tidwell, N.D., Eastwick, P.W., Finkel, E.J., 2013. Perceived, not actual, similarity predicts initial attraction in a live romantic context: evidence from the speed-dating paradigm. Personal Relationships 20 (2), 199–215.

Walster, E., Aronson, V., Abrahams, D., 1966. Importance of physical attractiveness in dating behavior. Journal of Personality and Social Psychology 4 (5), 508–516.

Walther, J.B., Van Der Heide, B., Kim, S.Y., Westerman, D., Tong, S.T., 2008. The role of friends' appearance and behavior on evaluations of individuals on Facebook: are we known by the company we keep? Human Communication Research 34 (1), 28–49.

Wetzel, C.G., Insko, C.A., 1982. The similarity-attraction relationship: is there an ideal one? Journal of Experimental Social Psychology 18 (3), 253–276.

Whitchurch, E.R., Wilson, T.D., Gilbert, D.T., 2011. He loves me, he loves me not...: uncertainty can increase romantic attraction. Psychological Science 22 (2), 172–175.

Zajonc, R.B., 1968. Attitudinal effects of mere exposure. Journal of Personality and Social Psychology 9 (2), 1–27.

Zhou, S.C., Wright, S.C., Moretti, M.M., 2008 (April). Directed self-expansion: ideal self and individual differences in self-expansion motives. Paper presented at the 88th annual meeting of the Western Psychological Association, Irvine, CA.

Interpersonal Psychotherapy

Denise E Wilfley and Allison L Shore, School of Medicine, Washington University, St. Louis, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Interpersonal psychotherapy (IPT) is a brief, time-limited treatment that was originally developed for the treatment of major depression. IPT is derived from theories in which interpersonal function is recognized as a critical factor in psychological adjustment and well-being. IPT is also based on empirical research connecting change in the social environment to the onset and maintenance of depression. Although the initial goal of IPT is to reduce symptoms of depression, the major goal is to improve the quality of the patient's current interpersonal relations and social functioning. IPT moves through three defined phases, each of which is associated with specific strategies and tasks for the therapist and patient. Its well-defined and unique treatment strategies are aimed at resolving problems within four social domains: grief, interpersonal role disputes, role transitions, and interpersonal deficits. IPT has been found to be efficacious for major depression and has also been successfully adapted to treat other types of mood and nonmood disorders (e.g., eating disorders and anxiety disorders). In this article, IPT's development, treatment phases and strategies, therapeutic stance, and outcome research are described. In addition, areas in need of further research are delineated.

Introduction

Interpersonal psychotherapy (IPT) is a brief, time-limited therapy that was developed in the late 1960s by Klerman and colleagues, and then, in 1984, was indicated for the treatment of depression in a manual by Klerman, Weissman, Rounsaville, and Chevron (Weissman et al., 2000). IPT was derived from theories in which interpersonal functioning is recognized as a critical factor in psychological adjustment and well-being. IPT is also based on empirical research connecting change in the social environment to the onset and maintenance of depression (Paykel et al., 1969). The objective of IPT is to improve the quality of patients' current interpersonal relations and social functioning, which in turn reduces symptoms of depression. IPT was initially formulated not as a novel therapy, but as a descriptive attempt to represent the current practice of psychotherapy for depression (Klerman and Weissman, 1993). IPT makes no assumptions about the causes of depression but instead takes into consideration any genetic and biological, psychological, and social factors that predispose, precipitate, and perpetuate the disorder (Ravitz et al., 2013b). Moreover, IPT assumes that the development of clinical depression occurs in a social and interpersonal context and that the onset, response to treatment, and outcomes are influenced by the interpersonal relations between the depressed patient and significant others.

IPT moves through three defined phases, each of which is associated with specific strategies and tasks for the therapist and patient. Its well-defined and unique treatment strategies are aimed at resolving problems within four social domains: grief, interpersonal role disputes, role transitions, and interpersonal deficits.

IPT has been found to be efficacious for major depression and has also been successfully adapted to treat other types of mental disorders. In this article, IPT's development, treatment phases and strategies, therapeutic stance, and outcome research are described. In addition, areas in need of further research are delineated. The extensive empirical background and theoretical foundation, including the strategies and techniques of IPT, are fully described in a comprehensive book (Weissman et al., 2000).

Development

Adolf Meyer's (1957) theory that psychopathology is a result of maladaptive adjustment to the social environment lays the groundwork for IPT. Harry Stack Sullivan (1953) stands as the theorist who most clearly articulated the interpersonal paradigm and popularized the term 'interpersonal' as a balance to the then-dominant intrapsychic approach. Sullivan believed that psychiatry involves the scientific study of people and the processes that go on among them; accordingly, the unit of clinical study is the patient's interpersonal relationships, as people cannot be understood in isolation from them. In his theory, Sullivan (1953: 13) posited that people have "relatively enduring patterns of recurrent interpersonal situations," which can either foster self-esteem or result in hopelessness, anxiety, and psychopathology.

Sullivan developed a comprehensive theory of the connections between psychiatric disorders and interpersonal relations for the developing child in the family and for the adult in the multiple transactions of life. The roles of major interest to IPT occur within the nuclear family (as parent, child, sibling, partner); the extended family; the friendship group; the work situation (as supervisor, supervisee, or peer); and the neighborhood or community. The interpersonal approach views the relationship between social roles and psychopathology as occurring in two ways: disturbances in social roles can serve as antecedents for psychopathology; and mental illness can produce impairments in the individual's capacity to perform social roles. IPT is also associated with the work of John Bowlby (1982), the originator of attachment theory, who acknowledged the importance of early attachment to subsequent interpersonal relationships and psychopathology. In sum, IPT is derived from theories in which interpersonal functioning is

recognized as a critical component of psychological adjustment and well-being.

IPT is based not only on theory, but also on empirical research linking change in the social environment to the onset and maintenance of depression. Considerable research supports the key interpersonal problem areas. For example, people become depressed in the circumstances of grief, marital disputes, and the life changes associated with interpersonal role transitions, particularly in the absence of social ties. Moreover, longitudinal studies show that not only can life events trigger depression, but indeed depression can also trigger adverse life events (for a review see Frank and Spanier, 1995; Weissman et al., 2000). In combination, these data provide strong support for use of an interpersonal approach for the understanding and treatment of depression.

Depression is conceptualized as having three component processes: symptom function, social and interpersonal relations, and personality and character problems. Symptom function entails the development of depressive affect and the neurovegetative signs and symptoms (e.g., low energy, sleep and appetite disturbance), both of which are believed to be precipitated by biological and psychological factors. Social and interpersonal relations entail interactions in social roles with other persons derived from learning based on childhood experiences, concurrent social reinforcement, and personal mastery and competence. Personality and character problems refer to enduring traits such as inhibited expression of anger or guilt, poor psychological communication with significant others, and difficulty with self-esteem. These traits determine a person's reactions to interpersonal experience. IPT intervenes in symptom formation and social dysfunction associated with depression, rather than enduring aspects of personality, because of its relatively short duration and low level of psychotherapeutic intensity. There is little expectation that IPT will have a pronounced effect upon personality structure; however, many IPT patients do acquire new social abilities that may help offset personality difficulties. An important component of IPT for depression is the deliberate avoidance of issues related to personality functioning during the treatment of the acute symptomatic episode (Weissman et al., 2000).

The methods and techniques of many schools of psychotherapy share common ground (Frank, 1973), including attempts to help patients gain a sense of mastery and reduce social isolation. A major difference among the therapies lies in their conceptualization of the causes of the patient's problems, and in the typical length and focus of treatment. For instance, IPT differs from traditional psychodynamic approaches in that it is time-limited, focused, examines current rather than past relationships, and recognizes but does not focus on intrapsychic defense mechanisms and internal conflicts. Furthermore, IPT differs from cognitive and behavioral approaches in that maladaptive thoughts and behaviors are addressed only as they apply to problematic interpersonal relationships. The goal is to change the relationship pattern rather than associated depressive cognitions, which are acknowledged in IPT as depressive symptoms. Although the therapeutic style is less directive in IPT than in cognitive and behavioral approaches (i.e., less didactic/prescriptive), the therapies are similar in terms of their active current focus on specified target areas. In contradistinction to other psychotherapies, IPT recognizes but does not directly focus on the patient's personality characteristics.

Treatment Phases, Goals, and Strategies

The initial phase, ordinarily the first one to three sessions, includes diagnostic assessment and psychiatric history and establishes the context for the treatment. After a standardized diagnostic symptom review is conducted, the patient is diagnosed as depressed and assigned the 'sick role.' Patients are given the 'sick role' to exempt them from additional social pressures, increase their awareness of the need for help, and elicit their cooperation in the process of recovery. A detailed examination of the patient's interpersonal history (the interpersonal inventory) is conducted, which includes a review of the patient's current social functioning and current close relationships, their patterns and mutual expectations. During this review, changes in relationships are illuminated that were proximal to the onset of symptoms (e.g., the death of a significant other, changing to a new job, increasing marital discord, or disconnection from a friend). The interpersonal inventory provides a structure for elucidating the social and interpersonal context of the onset and maintenance of depressive symptoms and delineates the focus of treatment.

An evaluation of need for medication is based on the severity of symptoms, previous experience with medications, and patient preference. The therapist and patient then discuss the diagnosis and what the patient might expect from treatment. Symptom relief starts with helping the patient understand that the vague and uncomfortable symptoms are a part of a known syndrome, which responds to several treatments and has good prognosis. The therapist then links the depressive syndrome to one of four interpersonal problem areas: grief, role transitions, interpersonal role disputes, or interpersonal deficits. After the major interpersonal problem area is identified, the therapist makes a specific treatment plan with the patient to work on this problem area.

During the intermediate phase of treatment, typically the middle 7–11 sessions, the therapist implements treatment strategies that are specific to the identified problem area. *Grief* is identified as the problem area when the onset of the patient's symptoms is associated with the death of a loved one, either recent or past. The goals for treating grief include facilitating mourning and helping the patient to find new activities and relationships to substitute for the loss. *Role transition* includes any difficulties resulting from a change in life status (e.g., divorce, retirement or change in one's work role, moving, leaving home, diagnosis of medical illness). The patient is helped to deal with the change by recognizing positive and negative aspects of the new role they are assuming, and pros and cons of the old role this replaces. *Interpersonal role disputes* are conflicts with a significant other (e.g., a partner, other family member, coworker, or close friend), which emerge from differences in expectations about the relationship. The therapist assists the patient to identify the nature of the dispute and generates options to resolve it. If resolution is impossible, the therapist assists the patient in dissolving the relationship and in mourning its loss. *Interpersonal deficits* include patients who are socially isolated or who are in chronically unfulfilling

relationships. The goal is to reduce the patient's social isolation by helping to enhance the quality of existing relationships and encouraging the formation of new relationships. The problem area may change during the course of treatment. If the patient has several related problem areas, the therapist selects the problem area that is most prominent or likely to result in change.

In the termination phase of treatment, typically the last three sessions, the therapist assists the patient in evaluating and consolidating gains, acknowledging the feelings associated with termination, detailing plans for maintaining improvements in the identified interpersonal problem area(s), and outlining remaining work. Patients are also encouraged to identify early warning signs (e.g., low energy) and to identify plans of action at signs of relapse.

Throughout IPT, focus stays on the interpersonal context of a patient's life. For example, if a patient is having depressive symptoms, the therapist will focus on relationship difficulties that exacerbate the symptoms. Meetings should not pass without reference to the interpersonal goals, as they are unique and require specific and directive interventions by the therapist. Indeed, the therapist's ability to maintain focus on interpersonal themes has been associated with better outcomes (Frank et al., 1991). In session, unfocused discussions are redirected to the key interpersonal issues, and abstract and general discussions are minimized in order to preserve focus. Therapists refrain from making inquiries that evoke vague or passive responses, such as general questions about the patient's week. Rather, sessions begin with questions such as 'What would you like to work on today?' and 'How have things been since we last met?' These questions provide more direction for patients and focus them on recent interpersonal events and recent mood, which the therapist attempts to connect. Patients who describe problem areas will be asked about recent mood and other depressive symptoms in relation to interpersonal events; alternatively, if the patient focuses on symptoms, the therapist asks about recent life events and interactions (Weissman et al., 2000).

Therapeutic Stance and Techniques

Each model of psychotherapy can be characterized in part by the positioning of the therapist in terms of the nature of the relationship with the patient and the level of therapeutic activity. The IPT therapeutic stance is one of warmth, support, and empathy. The IPT therapist is active and advocates for the patient rather than remaining neutral. Moreover, the therapeutic relationship is not viewed as a manifestation of transference, nor is it to be treated as a friendship. Some of the most frequently used techniques include exploratory techniques, encouragement of affect, clarification, and communication analysis; other techniques include use of therapeutic relationship, behavior change techniques, and adjunctive techniques.

In addition to the traditional individual format, IPT can also be delivered in groups, known as interpersonal psychotherapy for group (IPT-G) (Wilfley et al., 2000). There is one 2 hour pretreatment meeting at the start of treatment; during this session, the interpersonal inventory is administered, problem areas are identified, and treatment goals are established. In addition, there is a mid-treatment individual meeting to review treatment goals and a posttreatment individual meeting to review changes made and strategize about ways to maintain treatment gains. The rationale for utilizing a group is that it serves as an 'interpersonal laboratory' in which there are multiple opportunities to practice one's interpersonal skills so that skills can be readily applied to outside relationships. In addition, the group setting breaks patterns of social isolation and stigma, while providing a supportive environment. Moreover, IPT administered in a group setting is a more cost-effective treatment modality than individual sessions.

The key elements of IPT-G are similar to traditional IPT in that it is time-limited, semistructured, and focused on changing interpersonal relationships. Groups typically include 7–9 members who are homogeneous with regard to diagnosis and heterogeneous with regard to interactional style.

Outcome Research

Research has demonstrated outcome and efficacy for IPT in the following domains: mood disorders (i.e., recurrent depression (Weissman et al., 2000), bipolar mood disorder (Frank et al., 2005), and major depression (Weissman et al., 2000)); eating disorders (i.e., bulimia nervosa (Fairburn et al., 1995; Nevonen and Broberg, 2006), binge eating disorder (Wilson et al., 2010), anorexia nervosa (Carter et al., 2011), and subclinical eating disorders (Nevonen and Broberg, 2005)); anxiety disorders (i.e., post traumatic stress disorder (PTSD; Markowitz et al., 2009), body dysmorphic disorder (BDD; Weissman et al., 2000), and social anxiety disorder (SAD; Bohn et al., 2013)); as well as other disorders (i.e., multisomatoform disorders (Sattel et al., 2012) and borderline personality disorder (Bellino et al., 2010)).

Interpersonal Psychotherapy for Group

IPT-G has been shown to be efficacious in adults for a number of disorders, including major depressive disorder (Bolton et al., 2013), comorbid depression and substance use disorder among female prisoners (Johnson and Zlotnick, 2008), PTSD (Krupnick et al., 2008), perinatal depression (Reay, 2013), bulimia nervosa (Nevonen and Broberg, 2006), binge eating disorder (Wilfley et al., 2002), and eating disorders not otherwise specified (Nevonen and Broberg, 2005).

Adaptations of IPT

IPT has been adapted for various populations, including depressed older adults (Miller and Reynolds, 2007), Hispanic populations (Markowitz et al., 2009), adolescents with mood disorders (IPT Adolescent Skills Training; IPT-AST; Young et al., 2010), adolescents with depression who engage in nonsuicidal self-injury (Jacobson and Mufson, 2012), couples (Weissman et al., 2000), women who are pregnant (Grote et al., 2009) or women who are experiencing postpartum depression (O'Hara et al., 2000), and patients with comorbid medical conditions (e.g., patients infected with human immunodeficiency virus; Ransom et al., 2008; Weissman et al., 2000).

New adaptations to IPT currently under pilot study are partner-assisted IPT for the treatment of perinatal depression (Brandon et al., 2012), IPT for adolescents with learning disabilities (adapted from IPT-A; Brunstein-Klomek et al., 2013), clinician-managed IPT (CM-IPT; O'Hara and Stuart 2013), and interpersonal and social rhythm therapy (IPSRT) used as monotherapy for the acute treatment of bipolar II depression (Swartz et al., 2009). IPT has also been adapted for women with depression and chronic pain (Poleshuck et al., 2010). IPT for the prevention of excessive weight gain has been piloted in adolescent girls (Tanofsky-Kraff et al., 2010) and is currently being evaluated in a large-scale randomized controlled trial.

IPT has been modified for brief delivery (Swartz et al., 2008), long-term delivery (Weissman et al., 2000), as a maintenance intervention following weekly delivery (Frank et al., 2007), and for use as adjunctive self-help (Weissman et al., 2000). IPT has been modified for delivery via telephone (Ransom et al., 2008) and in primary care settings (Peeters et al., 2013).

IPT has been translated into several languages (Markowitz et al., 2009; Weissman et al., 2000). Together, these findings indicate that IPT is not only an effective and promising treatment for a range of psychiatric disorders, but it also has the potential for widespread dissemination (Markowitz and Weissman, 2004).

Future Directions

Treatment Efficacy

Though IPT has been shown to be efficacious for several mental disorders, no study has determined the optimal duration or dosing of treatment delivery. A critical next step is to determine the optimal treatment length and dosing of IPT for particular disorders. Cuijpers et al. (2013) suggest longer treatment lengths do not produce better outcomes and may be associated with negative treatment outcome. The authors propose using a greater number of treatment sessions per week over a shorter treatment length (Cuijpers et al., 2013). In contrast, Weissman and Verdeli (2012) propose a model of three sessions IPT followed by a triaging component (e.g., monthly phone calls) based on patients' symptom severity. In cases where patients have improved but still endorse moderate symptoms and impairment, additional treatment sessions can be conducted. Therefore, continued research to investigate optimal timing, duration, and frequency of treatment delivery is needed.

Mediators and Moderators

Although a number of outcome studies in IPT document its efficacy, little is known about the mechanisms by which IPT exerts its effects (Frank and Spanier, 1995). A greater understanding of the mechanisms by which change occurs in IPT would assist in further refinements of the treatment and may yield insights about the nature of the psychiatric syndrome under investigation. Continued research to evaluate the efficacy of IPT with specific patient subgroups is needed. For example, Chui et al. (2007) found an ethnicity by treatment effect for the reduction of objective binge eating episodes among patients with bulimia nervosa. Specifically, African-American participants, compared to other racial/ethnic groups, showed greater reductions in binge eating episodes when treated with IPT than cognitive behavioral therapy (CBT) (Chui et al., 2007). Wilson et al. (2010) demonstrated specificity of treatment effects in that individuals with low self-esteem and greater eating psychopathology performed best in IPT at 2-year follow-up compared to those receiving behavioral weight loss and CBT-guided self-help. Thus, considerations to adapt IPT for particular patient subgroups may need to be employed to increase the overall effectiveness and applicability of IPT.

Widespread Dissemination of IPT

IPT was developed as a research intervention but still has not been well disseminated among clinicians (Weissman et al., 2000). Evidence-based psychotherapies are being deployed in primary care worldwide, and there is a continued need to integrate mental health services into primary care in the United States (Weissman, 2013). Substantial evidence exists that clients are not receiving evidence-based treatments, such as IPT, in routine clinical care (Wilfley et al., 2002; Wilson et al., 2010). Furthermore, even when clients do receive IPT, it is often not delivered competently (Weissman et al., 2006). Additionally, a limited number of clinicians have been proficiently trained to use IPT (Weissman et al., 2006).

Practical, cost-effective methods of training clinicians to implement IPT are urgently needed. For example, increasing training during residency programs may be beneficial, because, over the course of training, residents became more comfortable with using evidence-based treatments as first-line interventions; this approach has been used for the success implementation of IPT in psychiatry resident programs (Lichtmacher et al., 2006). To increase the dissemination and sustainability of training, alternative strategies that extend traditional training techniques (e.g., in-person workshops, which require substantial time and resources) may be of benefit. The 'train-the-trainer' approach may be a more sustainable training strategy than traditional training techniques. In this approach, an internal champion is trained in evidence-based care delivery, and then trains other clinicians in their setting (Fairburn and Cooper, 2011). Continuing education in evidence-based psychotherapies is an underused strategy for addressing gaps in training in underserviced communities. A program of four peer-led weekly courses on the fundamentals of evidence-based care demonstrated that participants favored the peer-led group format over self-directed learning, which may facilitate increased treatment implementation (Ravitz et al., 2013a).

It may also be helpful to explore the use of Internet-based methods for disseminating IPT. Online programs for clinical training are more cost-effective and scalable than in-person approaches (Fairburn and Cooper, 2011). Online training can be offered to geographically dispersed trainees using minimal person-based resources. Trainees can view material on their own time, increasing the acceptability of this approach. Additionally, trainees are able to go back and review material already learned, which can reinforce session content. Training can be readily updated to efficiently disseminate new information and materials. Consequently, online training overcomes many of the barriers to in-person training and thus has

strong potential to increase the number of therapists trained to deliver evidence-based interventions.

Conclusion

IPT is a specialty, time-limited treatment that has demonstrated efficacy for the treatment of many mental disorders in individual and group formats. The IPT model assumes that social problems produce an environment in which disorders are created and sustained; thus, IPT targets current interpersonal problems in the areas of interpersonal deficits, interpersonal role disputes, role transitions, and grief, in order to reduce disorder symptomatology. The widespread dissemination of this highly acceptable, evidence-based treatment is greatly needed to increase access to care.

See also: Behavior Therapy: Evidence-Based Practice, Clinical Applications, Unresolved Issues; Cognitive Behavioral Therapy; Eating Disorders; Evidence-Based Psychotherapy; Group Psychotherapy, Clinical Psychology of.

Bibliography

Bellino, S., Rinaldi, C., Bogetto, F., 2010. Adaptation of interpersonal psychotherapy to borderline personality disorder: a comparison of combined therapy and single pharmacotherapy. Canadian Journal of Psychiatry 55 (2), 74–81.

Bohn, C., Aderka, I.M., Schreiber, F., Stangier, U., Hofmann, S.G., 2013. Sudden gains in cognitive therapy and interpersonal therapy for social anxiety disorder. Journal of Consulting and Clinical Psychology 81 (1), 177–182.

Bolton, P., Bass, J., Verdeli, H., Clougherty, K.F., Ndogoni, L., 2013. Group interpersonal psychotherapy for depression in rural Uganda: 6-month outcomes. Journal of the American Medical Association 289 (23), 3117–3124.

Bowlby, J., 1982. In: Attachment and Loss: Retrospect and Prospect, vol 52 (4). Basic Books, New York, pp. 664–678.

Brandon, A.R., Ceccotti, N., Hynan, L.S., Shivakumar, G., Johnson, N., Jarrett, R.B., 2012. Proof of concept: partner-assisted interpersonal psychotherapy for perinatal depression. Archives of Women's Mental Health 15 (6), 469–480.

Brunstein-Klomek, A., Kopleman-Rubin, D., Al-Yargon, M., Mufson, L., Apter, A., Mikulciner, M. (Eds.), June 10–15, 2013. Psychological Intervention Based on IPT-A for Adolescents Diagnosed with Learning Disorders. Fifth International ISIPT Conference: Mechanisms, Modalities and Moving Ahead, Iowa City, Iowa.

Carter, F.A., Jordan, J., McIntosh, V.V., Luty, S.E., McKenzie, J.M., Frampton, C.M., Bulik, C.M., Joyce, P.R., 2011. The long-term efficacy of three psychotherapies for anorexia nervosa: a randomized, controlled trial. International Journal of Eating Disorders 44 (7), 647–654.

Chui, W., Safer, D.L., Bryson, S.W., Agras, W.S., Wilson, G.T., 2007. A comparison of ethnic groups in the treatment of bulimia nervosa. Eating behaviors 8 (4), 485–491.

Cuijpers, P., Huibers, M., Ebert, D.D., Koole, S.L., Andersson, G., 2013. How much psychotherapy is needed to treat depression? A meta-regression analysis. Journal of Affective Disorders 149 (1–3), 1–13.

Fairburn, C.G., Norman, P.A., Welch, S.L., O'Connor, M.E., Doll, H.A., Peveler, R.C., 1995. A prospective study of outcome in bulimia nervosa and the long-term effects of three psychological treatments. Archives of General Psychiatry 52 (4), 304–312.

Fairburn, C.G., Cooper, Z., 2011. Therapist competence, therapy quality, and therapist training. Behaviour Research and Therapy 49 (6–7), 373–378.

Frank, E., Kupfer, D.J., Buysse, D.J., Swartz, H.A., Pilkonis, P.A., Houck, P.R., Stapf, D.M., 2007. Randomized trial of weekly, twice-monthly, and monthly interpersonal psychotherapy as maintenance treatment for women with recurrent depression. American Journal of Psychiatry 164 (5), 761–767.

Frank, E., Kupfer, D.J., Thase, M.E., Mallinger, A.G., Swartz, H.A., Fagiolini, A.M., Monk, T., 2005. Two-year outcomes for interpersonal and social rhythm therapy in individuals with bipolar I disorder. Archives of General Psychiatry 62 (9), 996–1004.

Frank, E., Kupfer, D.J., Wagner, E.F., McEachran, A.B., Cornes, C., 1991. Efficacy of interpersonal psychotherapy as a maintenance treatment of recurrent depression. Archives of General Psychiatry 48, 1053–1059.

Frank, E., Spanier, C., 1995. Interpersonal psychotherapy for depression: overview, clinical efficacy, and future directions. Clinical Psychology: Science and Practice 2, 349–369.

Frank, J.D., 1973. Persuasion and Healing: A Comparative Study of Psychotherapy, rev ed. Schocken, Oxford, England.

Grote, N.K., Swartz, H.A., Geibel, S.L., Zuckoff, A., Houck, P.R., Frank, E., 2009. A randomized controlled trial of culturally relevant, brief interpersonal psychotherapy for perinatal depression. Psychiatric Services 60 (3), 313–321.

Jacobson, C.M., Mufson, L., 2012. Interpersonal psychotherapy for depressed adolescents adapted for self-injury (IPT-ASI): rationale, overview, and case summary. American Journal of Psychotherapy 66 (4), 349–374.

Johnson, J.E., Zlotnick, C., 2008. A pilot study of group interpersonal psychotherapy for depression in substance-abusing female prisoners. Journal of substance abuse treatment 34 (4), 371–377.

Klerman, G.L., Weissman, M.M., 1993. Interpersonal psychotherapy for depression: background and concepts. In: Klerman, G.L., Weissman, M.M. (Eds.), New Applications of Interpersonal Psychotherapy, 1. American Psychiatric Press, Washington, DC, pp. 3–26.

Krupnick, J.L., Green, B.L., Stockton, P., Miranda, J., Krause, E., Mete, M., 2008. Group interpersonal psychotherapy for low-income women with posttraumatic stress disorder. Psychotherapy Research 18 (5), 497–507.

Lichtmacher, J., Eisendrath, S.J., Haller, E., 2006. Implementing interpersonal psychotherapy in a psychiatry residency training program. Academic Psychiatry 30 (5), 385–391.

Markowitz, J.C., Patel, S.R., Balan, I.C., Bell, M.A., Blanco, C., Yellow Horse Brave Heart, M., Lewis-Fernandez, R., 2009. Toward an adaptation of interpersonal psychotherapy for Hispanic patients with DSM-IV major depressive disorder. Journal of Clinical Psychiatry 70 (2), 214–222.

Markowitz, J.C., Weissman, M.M., 2004. Interpersonal psychotherapy: principles and applications. World Psychiatry 3 (3), 136–139.

Meyer, A., 1957. Psychobiology: A Science of Man. Thomas, Springfield, IL.

Miller, M.D., Reynolds 3rd, C.F., 2007. Expanding the usefulness of interpersonal psychotherapy (IPT) for depressed elders with co-morbid cognitive impairment. International Journal of Geriatric Psychiatry 22 (2), 101–105.

Nevonen, L., Broberg, A.G., 2005. A comparison of sequenced individual and group psychotherapy for eating disorder not otherwise specified. European Eating Disorder Review 13 (1), 29–37.

Nevonen, L., Broberg, A.G., 2006. A comparison of sequenced individual and group psychotherapy for patients with bulimia nervosa. The International Journal of Eating Disorders 39 (2), 117–127.

O'Hara, M.W., Stuart, S., Gorman, L.L., Wenzel, A., 2000. Efficacy of interpersonal psychotherapy for postpartum depression. Archives of General Psychiatry 57 (11), 1039–1045.

O'Hara, M., Stuart, S. (Eds.), 2013. Fifth International ISIPT Conference: Mechanisms, Modalities and Moving Ahead June 10–15 in Iowa City Iowa.

Paykel, E., Myers, J., Kienelt, M., Klerman, G., Lindethal, J., Pepper, M., 1969. Life events and depression: a controlled study. Archives of General Psychiatry 21 (6), 753–760.

Peeters, F., Huibers, M., Roelofs, J., van Breukelen, G., Hollon, S.D., Markowitz, J.C., Arntz, A., 2013. The clinical effectiveness of evidence-based interventions for depression: a pragmatic trial in routine practice. Journal of Affective Disorders 145 (3), 349–355.

Poleshuck, E.L., Talbot, N.E., Zlotnick, C., Gamble, S.A., Liu, X., Tu, X., Giles, D.E., 2010. Interpersonal psychotherapy for women with comorbid depression and chronic pain. Journal of Nervous and Mental Disease 198 (8), 597–600.

Ransom, D., Heckman, T.G., Anderson, T., Garske, J., Holroyd, K., Basta, T., 2008. Telephone-delivered, interpersonal psychotherapy for HIV-infected rural persons with depression: a pilot trial. Psychiatric Services 59 (8), 871–877.

Ravitz, P., Cooke, R.G., Mitchell, S., Reeves, S., Teshima, J., Lokuge, B., Zaretsky, A., 2013a. Continuing education to go: capacity building in psychotherapies for front-line mental health workers in underserviced communities. Canadian Journal of Psychiatry 58 (6), 335–343.

Ravitz, P., Watson, P., Grigoriadis, S., 2013b. Interpersonal psychotherapy for depression. In: Ravitz, P., Maunder, R. (Eds.), Psychotherapy Essentials to Go. W.W. Norton & Company, Inc., New York.

Reay, R. (Ed.), June 10–15, 2013. "Group Interpersonal Psychotherapy (IPT-g) for Perinatal Depression." Fifth International ISIPT Conference: Mechanisms, Modalities and Moving Ahead. Iowa, Iowa City.

Sattel, H., Lahmann, C., Gündel, H., Guthrie, E., Kruse, J., Noll-Hussong, M., Ohmann, C., Henningsen, P., 2012. Brief psychodynamic interpersonal psychotherapy for patients with multisomatoform disorder: randomised controlled trial. The British Journal of Psychiatry 200 (1), 60–67.

Sullivan, H.S., 1953. The Interpersonal Theory of Psychiatry, first ed. Norton, New York.

Swartz, H.A., Frank, E., Frankel, D.R., Novick, D., Houck, P., 2009. Psychotherapy as monotherapy for the treatment of bipolar II depression: a proof of concept study. Bipolar Disorders 11 (1), 89–94.

Swartz, H.A., Frank, E., Zuckoff, A., Cyranowski, J.M., Houck, P.R., Cheng, Y., Shear, M.K., 2008. Brief interpersonal psychotherapy for depressed mothers whose children are receiving psychiatric treatment. American Journal of Psychiatry 165 (9), 1155–1162.

Tanofsky-Kraff, M., Wilfley, D.E., Young, J.F., Mufson, L., Yanovski, S.Z., Glasofer, D.R., Salaita, C.G., Schvey, N.A., 2010. A pilot study of interpersonal psychotherapy for preventing excess weight gain in adolescent girls at-risk for obesity. International Journal of Eating Disorder 43 (8), 701–706.

Weissman, M.M., Markowitz, J., Klerman, G.L., 2000. Comprehensive Guide to Interpersonal Psychotherapy. Basic Behavioral Science Books, New York.

Weissman, M.M., Verdeli, H., Gameroff, M.J., Bledsoe, S.E., Betts, K., Mufson, L., Fitterling, H., Wickramaratne, P., 2006. National survey of psychotherapy training in psychiatry, psychology, and social work. Archives of General Psychiatry 63 (8), 925–934.

Weissman, M.M., Verdeli, H., 2012. Interpersonal psychotherapy: evaluation, support, triage. Clinical Psychology and Psychotherapy 19 (2), 106–112.

Weissman, M.M., 2013. Psychotherapy: Increasing in poor countries, decreasing in the United States. Huffpost Healthy Living. Retrieved from: http://www.huffingtonpost.com/myrna-m-weissman-phd/psychotherapy-low-income-countries_b_3991151.html.

Wilfley, D.E., MacKenzie, K.R., Welch, R.R., Ayres, V.E., Weissman, M.M., 2000. Interpersonal Psychotherapy for Group. Basic Books, New York.

Wilfley, D.E., Welch, R.R., Stein, R.I., Spurrell, E.B., Cohen, L.R., Saelens, B.E., Dounchis, J.Z., Frank, M.A., Wiseman, C.V., Matt, G.E., 2002. A randomized comparison of group cognitive-behavioral therapy and group interpersonal psychotherapy for the treatment of overweight individuals with binge-eating disorder. Archives of General Psychiatry 59 (8), 713–721.

Wilson, G., Wilfley, D.E., Agras, W., Bryson, S.W., 2010. Psychological treatments of binge eating disorder. Archives of General Psychiatry 67 (1), 94–101.

Young, J.F., Mufson, L., Gallop, R., 2010. Preventing depression: a randomized trial of interpersonal psychotherapy-adolescent skills training. Depress Anxiety 27 (5), 426–433.

Interpersonal Trust across the Lifespan

Ken J Rotenberg, School of Psychology, Keele University, Staffordshire, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by K.J. Rotenberg, volume 11, pp. 7866–7868, © 2001, Elsevier Ltd.

Abstract

The article outlines the development of trust across the life span. The development comprises: (1) honesty-trusting behavior toward caregivers during infancy; (2) honesty and reliability cognitive-based trust during early childhood; (3) honesty, reliability, and emotional trust distinctly in peers during middle childhood; (4) moral/abstract trust including social networks during adolescence; (5) romantic relationship and workplace trust during early adulthood; and (6) care provision and inheritance concerns trust during late adulthood and old age, respectively. Research shows that (1) trust beliefs during childhood longitudinally predict psychosocial functioning and (2) children who hold very high trust beliefs show elevated psychosocial problems.

The importance of trust and the development of trust are recognized in criminology, sociology, psychology, political science, and industry/organization among other disciplines (see Rotenberg, 2010). An EBSCO/PSYINFO search using the combination of psychology and sociology with the term 'trust' in the abstract yields a total of 112 345 publications across history. Of those, 2714 are publications within the last 6 months of July 2014, which shows that there are approximately 452 publications on trust per month. The current volume of research examining trust is indeed 'staggering.'

I provided a developmental outline of changes in interpersonal trust in the first edition, which was guided by the bases × domains × target dimensions trust framework (BDT; Rotenberg, 1994). This article will expand upon that proposal drawing upon current research. In addition, this article will deal with the related issues of: (1) whether or not childhood trust has 'consequences' for their psychosocial functioning across the life span, and (2) whether or not children who hold very high trust beliefs are at risk for psychosocial problems.

A Theoretical Framework

The BDT framework (Rotenberg, 2010) is now composed of three bases (reliability, honesty, and emotional) × three domains (cognitive/affective, behavior-dependant, and behavior-enacting) × two target dimensions (specificity and familiarity). The three bases of trust are: (1) reliability, which refers to the fulfillment of word or promise; (2) emotional trust, which refers to the reliance on others to refrain from causing emotional harm, such as being receptive to disclosures, maintaining confidentiality of them, refraining from criticism, and avoiding acts that elicit embarrassment; and (3) honesty, which refers to telling the truth and engaging in behaviors that are guided by benign rather than malicious intent and by genuine rather than manipulative strategies.

The three domains of the BDT comprise: (1) cognitive/affect, which entails individuals' beliefs and feelings that others demonstrate the three bases of trust (e.g., believes that others keeps promises); (2) behavior-dependent trust, which entails individuals behaviorally relying on others to act in a trusting fashion as per the three bases of trust (e.g., depends on another to keep a promise); and (3) behavior-enacting trust, which entails individuals behaviorally engaging in the three bases of trust (e.g., keeps promises). The term 'trust beliefs' will be used to describe the cognitive-affect domain and term 'trustworthiness' will be use to describe the behavior-enacting domain for ease of expression. The BDT framework includes the specificity dimension of the target of trust (ranging from general category versus a specific person) and familiarity of the target of trust (ranging from slightly familiar to highly familiar). According to the BDT framework, trust is the result of reciprocal processes whereby individuals demonstrate matching patterns of trust and establish a common social history.

The following developmental pattern was proposed. During infancy, trust is primarily composed of behavior-dependent honesty that is directed toward caregivers and strangers. During early childhood, children establish honesty and reliability cognitive trust that is manifested in behavior, which are directed toward a growing network of peers and adults. During middle childhood, the honesty, reliability, and emotional bases of trust are evidenced in interactions with adults, in peer relationships, and academic achievement that is linked to their reasoning about violations of social expectations and moral principles. During adolescence, all bases and domains of trust are extended to social-political events, moral issues in society, and sexual matters. Trust during early and middle adulthood becomes focused on romantic partners, and superiors and coworkers in the workplace. Finally, trust during later adulthood and very old age is focused on vulnerability to criminal acts, whether others will provide care, family members' inheritance claims, and whether the trust placed in family members or mankind will be confirmed after death. The proposed developmental pattern is formulated as a description of the *principal* basis and focus of trust at different age periods across the life span.

This article will address two additional issues: First, it addresses the extent to which psychosocial functioning across the life span is a consequence of childhood trust. Second, it addresses the question regarding whether or not children who hold very high trust beliefs are at risk for psychosocial problems?

Trust Across the Life Span

Infancy (0–2 Years)

During this period, individuals (1) form a sense of trust versus mistrust (Erikson, 1963); (2) establish different qualities of attachment with caregivers that include a sense of trust in them as represented in an internal working model (IWM) (Cohn, 1990; Lieberman, 1977; Dwyer, 2005; Sakai, 2010); (3) establish an IWM that affects subsequent social relationships; and (4) demonstrate shyness, which entails a low trusting orientation toward unfamiliar others. The primary form of trust during this period is behavior-dependent honesty, which is the tendency for infants to behavioral rely on care providers and others predicated on the *feeling* (i.e., affect) that their actions are guided by benign intentions.

Early Childhood (2–6 Years)

Trust during early childhood entails honesty as cognitive/affective basis of trust, behavior-dependent trust, and trustworthiness. This proposition has received support by the research by Harris and his colleagues (Chen et al., 2013; Corriveau and Harris, 2010; Harris, 2012) as well as others (e.g., Robinson et al., 2011). In the research by Harris and his colleagues, 4- and 5-year-old children are typically confronted with figures depicting hybrids of animals (i.e., they had features of two animals) and mothers or others named the animals. It has been found that young children select: (1) the name of the animal provided by their mothers when the object was a complete hybrid; (2) the accurate name of the animal when it had more distinguishing features (i.e., those distinctly of one animal) despite the mothers' selection of names; (3) the name of the animal provided by an adult's source with a history of accuracy rather than inaccuracy; (4) the name of the animal provided by others when expressed as a consensus, and (5) the name of the animal when the race of informants is the same as the child's (see Chen et al., 2013; Corriveau and Harris, 2010). These findings may be interpreted as showing that even very young children hold beliefs (and corresponding affect) regarding whether or not sources of communication are honest (i.e., truthful) and depend on those sources as a guide to their behavior.

The preceding line of investigation identifies the factors that affect children's reliance on others as a source of information and thus the children's acquisition of knowledge regarding science and religion (see Harris, 2012). These findings may be interpreted as showing young children engage in behavior-dependent trust as it pertains to honesty (Corriveau and Harris, 2010). This interpretation is uncertain because it is not clear that the children's behavior is a result of their sensitivity to the intentions of others (e.g., benign as opposed to malevolent) as required for honesty according to the BDT.

Children's Theory of Mind (ToM) comprises their ability to infer, and utilize in judgment, the premise that others have cognitive and emotional states such as intentions, desires, beliefs, knowledge, and emotion (see Peskin and Ardino, 2003). Research has demonstrated that young children's ToM is associated with their honesty trustworthiness, notably telling the truth rather than deceiving/lying (see Polak and Harris, 1999), and with their emotional trust beliefs (Rotenberg et al., in press).

Middle Childhood (7–12 Years)

During this period, parents continue to affect children's social cognition, affect, and behavior. The children become increasingly involved, however, in school, peer friendships, and peer groups (primarily same sex) and the neighborhood. As a consequence, children's honesty, reliability, and emotional trust both as beliefs and behavior are (1) affected by parents' trust beliefs and behavior (Rotenberg, 1995); (2) an integral part of same-sex peer friendships (Rotenberg, 1986); (3) part of peer-group interactions (Rotenberg and Boulton, 2013); (4) linked to their reasoning regarding social expectations and moral principles (Kahn and Turiel, 1988); (5) linked to the trustworthiness of their neighborhoods (Rotenberg et al., 2012); and (6) associated with academic achievement (Imber, 1973). Because of the increasing exchange of intimacies (secrets) between peer friends (see Rotenberg and Boulton, 2013), reciprocal patterns of emotional trust is shown during middle childhood.

Adolescence (13–19 Years)

Adolescents' sexual development, moral development, increased ability to comprehend abstract concepts, and their involvement in broader social networks and social events result in further changes in trust. As a consequence, adolescents' trust extends to sexual matters such as birth control and the contraction of sexually transmitted diseases including AIDS (see Lear, 1997). Also, adolescents begin to prepare for their role in democracy from their school involvement and to display concern with the moral issues pertaining to wider groups within society that bear on trust (see Flanagan and Stout, 2010; Settle et al., 2011). Adolescents' trust in institutions is lower than those found in older adults (Quintelier, 2007), and younger generations demonstrate lower generalized trust in others than do older generations (see Flanagan and Stout, 2010).

Early and Middle Adulthood (20–29 Years and 30–55 Years, Respectively)

During this period, individuals most commonly become involved in extended romantic relationships and enter the workplace. Accordingly, trust during this period is highlighted by trust within the context of romantic relationships. Trust in romantic partners is: (1) a defining aspect of their romantic attachment style notably security (Mikulincer, 1998), (2) undermined by reciprocal self-concealment (Uysal et al., 2012), and (3) linked to relationship-promoting memory biases regarding transgressions of partners (Luchies et al., 2013). Individuals' trust in marriage partners has been found to be associated with their love of their partners and the belief that partners are concerned about their (the individuals') welfare (Rempel et al., 1985). Also, there is evidence that individuals' show an impulsive form of trust, which increases their caring for their romantic partners (Murray et al., 2011). This period is also marked by trust issues concerning superiors and coworkers in the workplace, which affect psychosocial functioning and productivity (see Knoll and Gill, 2011). All bases, domains, targets, and

dimensions of trust in Rotenberg's (1994) framework are viable throughout adulthood.

Later Adulthood (55–75 Years)

This period is marked by individuals' concerns over quality of life and physical safety. Older adults' trust centers on the extent to which their offspring (adult children), social networks, and social agencies will provide care and help to establish or sustain a satisfactory quality of life (Bierhoff, 1992). Weak trust beliefs in family and friends or neighbors are associated with a subjective sense of insecurity during this age period (Fagerström et al., 2011). During this period, individuals are concerned about their vulnerability to being victims of crime and thus their trust centers on their physical safety (Rotenberg, 1990).

Very Old Age (75+ Years)

During very old age, individuals confront issues of distribution of inheritance and, as a consequence, reevaluate their trust in family members. They deal with the issue of their immortality and specifically whether the trust they place in family members and perhaps in society or mankind to construct a better world will be confirmed after death. It is important to note that research shows that the holding low generalized trust in others by individuals is predictive of their mortality, notably in males (Barefoot et al., 1998; Nummela et al., 2012). Among other interpretations, this longitudinal relation may be attributed to trust beliefs serving as a buffer to stress or to distrust per se as a stressor that reduces health and, in turn, results in mortality.

Psychosocial 'Consequences' of Childhood Trust

Erikson (1963) proposed that trust versus mistrust was the first of the eight stages of psychosocial development and evidenced during infancy. The resolution of that stage was believed to affect the healthy to unhealthy psychosocial resolution of subsequent stages during the course of development. Although there is dearth of research directly examines that hypothesis, researchers have investigated whether or not later psychosocial functioning across the life span is a consequence of childhood trust. Malti et al. (2013) carried out a four-wave longitudinal investigation of 1028 children that ranged from 8 to 11 years of age. Latent growth-curve analyses demonstrated that low trustworthiness (i.e., low promise-keeping behavior) was predictive of stable aggressive behavior as reported by teachers. In a related study, Rotenberg et al. (2004) examined trustworthiness, as assessed by peer-reported promise-keeping and secret-keeping, and the number of friendships for children/early adolescents from Canada and the United Kingdom. It was found that trustworthiness longitudinally predicted the number of friendships across an 8-month or 1-year period.

As further investigation of the potential consequences of childhood trust, Qualter et al. (2013) examined the relation between trust shown by 7-year-olds and their psychosocial and health functioning at 2-year intervals until they were 17 years of age. Qualter et al. (2013) found that children's reliability-based trust beliefs in others and behavior-dependent trust on others longitudinally predicted stable low loneliness across time. In a related study, Rotenberg et al. (2010) carried a short-term longitudinal examination of the relations between trust beliefs (reliability and emotional) during early childhood, middle childhood, and early adulthood. It was found that trust beliefs longitudinally predicted low loneliness in each age group. The findings show that psychosocial functioning such as low loneliness and low aggression is a probable consequence of children's trust beliefs and trustworthiness. It should be emphasized that the findings suggest that those are probable consequences because they represent associations across time, albeit longitudinal in nature. Also, the research has not yet addressed whether those consequences persist across the entire life span.

Are Children with Very High Trust Beliefs at Risk for Psychosocial Problems?

There is a growing body of research showing that children with very high trust beliefs are at a risk for psychosocial problems. Rotenberg et al. (2005) found curvilinear U-shape (quadratic) relations between 9- and 11-year-old children's trust beliefs in the peer group/peer friends and psychosocial functioning. Children with very high and those with very low trust beliefs were more rejected by peers, more excluded by peers, and held lower perceptions of self-perceived social acceptance, than were children with midrange trust beliefs. Children's trust beliefs linearly predicted changes in psychological maladjustment (depression, loneliness, and anxiety) but that was qualified by a quadratic pattern. Children with very high and those with very low trust beliefs showed elevated levels of changes in psychological maladjustment compared to a linear relation. Also, quadratic relations showed that, compared to children within the middle range of trust beliefs, children with very high trust beliefs and those with very low trust beliefs in peers were (1) lower in self-perceived social acceptance, (2) more excluded by peers, and (3) more rejected as opposed to accepted by peers. Similarly Betts et al. (2009) found quadratic patterns in 5- to 7-year-old children. Children with very low and those with very high generalized trust beliefs in other and demonstrated greater loneliness and fewer friendships than did children with the middle range of generalized trust beliefs. The findings attest to the principle that being 'too trusting' can have detrimental effects. Research has not yet shown that that pattern is evidence during other periods in development.

Summary

The proposed developmental pattern redresses the lack of such a proposal in the field. It is hoped that the currently proposed pattern, when linked to research, will both shed light on individuals' experience of trust during the life span and promote relevant research. Although, the growth of research on trust is unprecedented, the body of knowledge is not growing accordingly because of the diversity of conceptualizations and measures of that construct. The BDT framework and the proposed developmental offers hope for integrating the expanding work on the topic of trust.

See also: Trust, Political; Trust, Sociology of; Trust: Philosophical Aspects.

Bibliography

Barefoot, J.C., Maynard, K.E., Beckham, J.C., Brammett, B.H., Hooker, K., Siegler, I.C., 1998. Trust, health and longevity. Journal of Behavioral Medicine 21, 517–526.

Betts, L.R., Rotenberg, K.J., Trueman, M., 2009. The early childhood generalised trust belief scale. Early Childhood Research Quarterly 25, 175–185.

Bierhoff, H.W., 1992. Trust and trustworthiness. In: Montada, L., Filipp, S.H., Lerner, M.J. (Eds.), Life Crisis and Experience of Loss in Adulthood. Erlbaum, Hillsdale, NJ, pp. 411–433.

Cohn, D.A., 1990. Child-mother attachment of six-year-olds and social competence at school. Child Development 61, 152–162.

Corriveau, K.H., Harris, P.L., 2010. Young children's trust in what other people say. In: Rotenberg, Ken (Ed.), Interpersonal Trust during Childhood and Adolescence. Cambridge University Press, Cambridge, UK, pp. 87–109.

Chen, E.E., Corriveau, K.H., Harris, P.L., 2013. Children trust a consensus composed of outgroup members – but do not retain the trust. Child Development 84, 269–282.

Dwyer, K.M., 2005. The meaning and measurement of attachment in middle and late childhood. Human Development 48, 155–182.

Erikson, E.H., 1963. Childhood and Society, second ed. Norton, New York.

Fagerström, L., Gustafson, Y., Jakobsson, G., Johansson, S., Vartiainen, P., 2011. Sense of security among people aged 65 and 75: external and inner sources of security. Journal of Advanced Nursing 67 (6), 1305–1316.

Flanagan, C.A., Stout, M., 2010. Developmental patterns of social trust between early and late adolescence: age and school climate effects. Journal of Research on Adolescence 20 (3), 748–773.

Harris, P.L., 2012. Trusting What You're Told: How Children Learn from Others. Belknap Press of Harvard University Press, Cambridge, MA, USA.

Imber, S.C., 1973. Relationship of trust to academic performance. Journal of Personality and Social Psychology 28, 145–150.

Kahn Jr., P.H., Turiel, E., 1988. Children's conceptions of trust in the context of social expectations. Merrill-Palmer Quarterly 34, 403–419.

Knoll, D.L., Gill, H., 2011. Antecedents of trust in supervisors, subordinates, and peers. Journal of Managerial Psychology 26 (4), 313–330.

Lear, D., 1997. Sex and Sexuality: Risk and Relationships in the Age of AIDS. Sage, Thousand Oaks, CA.

Lieberman, A.F., 1977. Preschoolers' competence with a peer: relations with attachment and peer experience. Child Development 48, 1277–1287.

Luchies, L.B., Wieselquist, J., Rusbult, C.E., Kumashiro, M., Eastwick, P.W., Coolsen, M.K., Finlel, E.J., 2013. Trust and biased memory of transgressions in romantic relationships. Journal of Personality & Social Psychology 104 (4), 673–694.

Malti, T., Averdijk, M., Ribeaud, D., Eisner, M.P., Rotenberg, K.J., 2013. Children's trust and developmental trajectories of aggressive behavior in an ethnically diverse sample. Journal of Abnormal Child Psychology 41, 445–456.

Mikulincer, M., 1998. Attachment working models and the sense of trust: an exploration of interaction goals and affect regulation. Journal of Personality and Social Psychology 74, 1209–1224.

Murray, S.L., Pinkus, R.T., Holmes, J.G., Harris, B., Gomillion, S., Aloni, M., Derrick, J.L., Leder, S., 2011. Signaling when (and when not) to be cautious and self-protective: impulsive and reflective trust in close relationships. Journal of Personality & Social Psychology 101 (3), 485–502.

Nummela, O., Raivio, R., Uutela, A., 2012. Trust, self-rated health and mortality: a longitudinal study among ageing people in Southern Finland. Social Science & Medicine 74 (10), 1639–1643.

Peskin, J., Ardino, V., 2003. Representing the mental world in children's social behavior: playing hide-and-seek and keeping a secret. Social Development 12, 496–512.

Polak, A., Harris, P.L., 1999. Deception by young children following noncompliance. Developmental Psychology 35, 561–568.

Qualter, P., Brown, S.L., Rotenberg, K.J., Vanhalst, J., Harris, R.A., Goossens, L., Bangee, M., Munn, P., 2013. Trajectories of loneliness during childhood and adolescence: Predictors and health outcomes. Journal of Adolescence 41, 325–338.

Quintelier, E., 2007. Differences in political participation between young and old people. Contemporary Politics 13 (2), 165–180.

Rempel, J.K., Holmes, J.G., Zanna, M.P., 1985. Trust in close relationships. Journal of Personality and Social Psychology 49, 95–112.

Robinson, E., Butterfill, S.A., Nurmsoo, E., 2011. Gaining knowledge via other minds: children's flexible trust in others as sources of information. British Journal of Developmental Psychology 29 (4), 961–980.

Rotenberg, K.J., 1986. Same-sex patterns and sex differences in the trust basis of children's friendship. Sex Roles 15, 613–626.

Rotenberg, K.J., 1990. A measure of the trust beliefs of elderly individuals. International Journal of Aging and Human Development 30, 141–152.

Rotenberg, K.J., 1994. Loneliness and interpersonal trust. Journal of Social and Clinical Psychology 13, 152–173.

Rotenberg, K.J., 1995. The socialization of trust: parents' and their children's interpersonal trust. International Journal of Behavioural Development 18, 713–726.

Rotenberg, K.J. (Ed.), 2010. Interpersonal Trust during Childhood and Adolescence. Cambridge University Press, New York.

Rotenberg, K.J., Addis, N., Betts, L.R., Fox, C., Hobson, Z., Rennison, S., Trueman, M., Boulton, M.J., 2010. The relation between trust beliefs and loneliness during early childhood, middle childhood and adulthood. Personality and Social Psychology Bulletin 36, 1086–1100.

Rotenberg, K.J., Betts, L.R., Eisner, M., Ribeaud, D., 2012. Social antecedents of children's trustworthiness. Infant and Child Development 21, 310–322.

Rotenberg, K.J., Boulton, M.J., 2013. The relations between interpersonal trust consistency and quality of peer relationships during childhood. Social Development 22, 225–241.

Rotenberg, K.J., Boulton, M.J., Fox, C., 2005. Cross-sectional and longitudinal relations among trust beliefs, psychological maladjustment, and social relationships, during childhood: are very high as well as very low trusting children at risk? Journal of Abnormal Child Psychology 33, 595–610.

Rotenberg, K.J., McDougall, P., Boulton, M.J., Vaillancourt, T., Fox, C., Hymel, S., 2004. Cross-sectional and longitudinal relations among relational trustworthiness, social relationships, and psychological adjustment during childhood and adolescence in the UK and Canada. Journal of Experimental Child Psychology 88, 46–67.

Rotenberg, K.J., Petrocchi, S., Lecciso, F., Marchetti, A., in press. The relation between children's trust beliefs and theory of mind abilities. Infant and Child Development.

Sakai, A., 2010. Children's sense of trust in significant others: genetic versus environmental contributions and buffer to life stressors. In: Rotenberg, Ken J. (Ed.), Interpersonal Trust during Childhood and Adolescence. Cambridge University Press, New York, pp. 56–84.

Settle, J.E., Bond, R., Levitt, J., 2011. The social origins of adult political behavior. American Politics Research 39 (2), 239–263.

Uysal, A., Lin, H.L., Bush, A.L., 2012. The reciprocal cycle of self-concealment and trust in romantic relationships. European Journal of Social Psychology 42 (7), 844–851.

Interpretation in Anthropology

Johannes Fabian, University of Amsterdam, Amsterdam, The Netherlands

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7874–7878, © 2001, Elsevier Ltd.

Abstract

Interpretation, often paired with or opposed to explanation, signifies practices (including methods of selecting and preparing ethnographic material, ways of defining questions and arguing positions, and styles of writing) by which anthropologists produce and represent knowledge. The idea goes back to a tradition of text-oriented critical philology and history that, together with the project of a natural history of mankind, emerged as the discipline of anthropology during the Age of Enlightenment. Approaches to interpretation in anthropology that were developed subsequently varied depending on how the discipline defined its object of study (for instance, as a historical record, as the result of processes of evolution, or as symbol-based systems orienting action). Throughout this period, interpretation retained its general meaning, until in the 1970s, pleas for an interpretive or hermeneutic anthropology were formulated by those who argued that cultural anthropology could not be practiced either as a (narrowly defined) natural science or as a purely semiotic discipline. Assigning central importance to the concept of text (and context) was a common feature of this critical movement; its strategies varied greatly between promoting text as a metaphor for culture, regarding ethnographic data as texts, promoting performative views of culture as texts-in-action, and calling for attention to the political aspects of investing texts with authority.

Since the beginnings of the discipline, interpretation has occupied and affected everyone who practices anthropology. Being all-pervasive, a concept or practice can become invisible such that in most standard works on the theory and history of anthropology 'interpretation' does not rate an entry in the index, and that goes, curiously, also for the book that more than others put interpretation on the map of debates in recent anthropology (Geertz, 1973). Half a decade later, anthropology figured prominently among the social sciences that were said to have turned to an 'interpretive or hermeneutic approach' (Rabinow and Sullivan, 1979, but see Bauman, 1978 for the long and deep history of the challenge of hermeneutics). Significantly, this claim was made in a Reader, a collection of authoritative essays. As a genre of scholarly books, we expect Readers to appear when a theoretical concept and its attendant practices of research and writing are either programmatically announced or have already congealed as a distinctive (sub)discipline. The latter, as far as we can tell, has not happened. Since a clearly defined hermeneutic anthropology does not exist, a statement about 'interpretation in anthropology' is a peculiar task, and one which calls for narrative rather than systematic presentation.

The Burden of History: Hermeneutics

A way to get a grip on elusive 'interpretation' in anthropology is to trace the concept to the discipline's modern beginnings. Most historians locate them in the Age of Enlightenment, and this proves helpful provided that the scope is not narrowed unduly (and anachronistically) to a search for eighteenth-century predecessors of anthropology as a natural or social science. Thinkers whom we recognize as proto-anthropologists formulated their ideas in a context of theological, philosophical, and (often overlooked) philological debates. These debates were carried out in a field of tension that developed between the Enlightenment and, largely contemporary, Romantic thought (a complex story still best covered in Gusdorf, 1973). From this broader view of our genealogy one can derive a heuristic model whereby modern conceptions of interpretation are seen to go back to moves between two poles. One of them is a long tradition of the exegesis of authoritative texts such as the Scriptures, the Classics but also corpora of Law. Typically, interpreters of such texts partook in the authority of their sources and had an interest in promoting and defending a transhistorical status for their sources. At the other pole, we find a tradition of secular critique. It emerged from moves to historicize authoritative texts by submitting them, often building on methods and techniques developed by theological and philological exegetes, to literary, social, and political analysis and interpretation (a conceptual pair that shows up also in most modern writings). Biblical criticism and what was called the Homeric question had demonstrated that these texts of quasi-transcendent religious and aesthetic authority had a mundane history of creation, collection, and editing under multiple individual and collective authorship. Because they had a past they were relevant to a critical understanding of the present. In a line leading from Hegel to Marx and Weber, to reveal the historicity of human affairs became a foremost task of social science.

Anthropology owes its openness to ideals of rational inquiry into, and appreciation of, the creations of other traditions ('cultures') at least as much to this secular tradition of critical interpretation as to the principles and habits of natural science. Given this double inheritance, distinctions between interpretive hermeneutic and explanatory scientific positions that keep the two facing each other in binary opposition become stultifying. The same applies when exegesis, interpretation working close to the material surface of texts, is contrasted with hermeneutics concerned with deep meanings. Least of all do such divisions correspond to that between a subjective quest for authority and power and the objective pursuit of scientific truth. Those who interpret and those who explain have about the same potential for turning into authoritative pundits. In sum, to acknowledge the double genealogy of anthropology in a hermeneutic and a scientific

lineage maps a conceptual space necessary to discuss interpretation in anthropology; distinctions between them do not validate or invalidate, or even classify approaches.

While unproductive oppositions should be rejected, it is important to keep in mind that, in its modern usage, interpretation has been a critical and often militant concept. To cite one of Bauman's conclusions: "*The method of sociological hermeneutics,* like that of empirical science, *can serve the practice of communication only in its negative capacity, as the method of criticism*" (1978: p. 241, emphasis in the original).

Interpretation and the Object of Anthropology

Anthropology may have owed much to a tradition of critical interpretation, and J. G. Herder, a thinker steeped in that tradition, can be credited with having bequeathed to anthropology essential elements that went into its guiding idea, culture. Yet interpretation or hermeneutics did not assume a prominent position in debates about the production of anthropological knowledge until the late 1960s. Of the two paradigms under which anthropology gained admission as an academic discipline toward the end of the nineteenth century, evolutionism (as represented by E. B. Tylor) emerged from the Scottish Enlightenment and embraced principles and taxonomic methods of natural history. It derived its methods above all from disciplines such as geology, biology, and comparative anatomy. Diffusionism can be traced to the Romantic project of a universal cultural history but it practiced a hermeneutic that had been narrowed to a positivist methodology of interpreting historical sources, as exemplified in F. Graebner's (1911) adaptation of the work of E. Bernheim (1970, orig. 1889), a medieval historian. In both paradigms, notwithstanding the literary ambitions of some proponents, such as James Frazer, and the influence of literary critic Matthew Arnold on E. B. Tylor, anthropology seemed to lack an object for critical interpretation comparable to the great texts, the Scriptures, or the Classics. This did not change with the succeeding paradigms of British functionalism, French structuralism, or even (and at first) American culturalism. Primitive, nonwestern societies were not to be interpreted; they called for functional explanation, structural decoding, or structural-functional systemic analysis, until the study of cultural 'meaning' took the center stage.

This last development was tied to the concept of symbol that, unbeknownst to many who wielded it (leaning on Freud, Cassirer, Kenneth Burke, or Susanne Langer), always threatened to take anthropology back to precritical hermetic traditions of interpretation. Most daringly stated by Roy Wagner (1975), a 'symbolic anthropology' staked its claims – also with a Reader (Dolgin et al., 1977) – at about the time when calls for an 'interpretive and hermeneutic' approach were first heard (possibly prepared by Palmer's excellent general introduction to hermeneutics in English published in 1969). Clifford Geertz was invoked as an authority by the symbolic as well as the interpretive movement. He remained somewhat aloof of both and continued to call his approach semiotic (i.e., sign-centered) although he went further in the direction of hermeneutics than any of his predecessors when he suggested that anthropology's object of study, culture, might be approached as (or like) a text (an 'ensemble of texts' in his words). Geertz referred to an essay by Paul Ricoeur (1971, reprinted in Rabinow and Sullivan, 1979: pp. 73–101). Ricoeur had advanced the 'hypothesis' that the human sciences, by adopting a hermeneutic approach, could proceed from 'Explanation to Understanding' (for a broader statement of the relevance of hermeneutics and interpretation to the social sciences, see Ricoeur's influential study of Freud, 1970). Earlier influences from Dilthey and Weber had prepared the ground for giving to 'understanding' this new and sharp focus.

Interpretation and Method

Thus, by the mid-1970s, one of its most influential proponents seemed to have steered anthropology back to the Romantic tradition of interpretation. However, important as this move may have been at the level of theory (i.e., in redefining the object of anthropology and in clarifying the status of the discipline as an interpretive rather than explanatory science), its practical consequences were limited. The model of the text did not generate a new distinctive method; thick description and a knack for discerning 'webs of signification' continued to work as the ethnographer's 'magic.' For those who embraced the hermeneutic stance, text remained essentially a metaphor and few among these writers were compelled by this figure to direct their ethnographic research and writing to actual texts and their interpretation. The latter happened in a movement that was roughly contemporary with Geertzian hermeneutics and had intellectual roots that, acknowledged or not, go back to W. von Humboldt. It took its departure from a critique of formalist structural linguistics, of purely classificatory approaches in folklore, and of positivist views of ethnography. It was in this field of innovative endeavors where critical thought in language-centered anthropology (Hymes, 1974) and the philosophy of social science (Habermas, 1988, orig. 1967) converged that we began to see how interpretation might develop from a theoretical figure to a practical guideline, a method organizing the production of anthropological knowledge from field research to writing.

Whether or not certain choices, steps, procedures, and strategies devised for interpretation can count as an anthropological method is an epistemological question: how must we conceive the nature of anthropological inquiry, how the nature of its object? From a hermeneutic position the answer would roughly be this: anthropology, though aware of, and in some of its branches concerned with, neural, etiological, and certain statistically demonstrable determinations of human behavior, starts with the assumption that human experience as well as the study of that experience are always mediated. Mediated means that experiences (and our understandings) are made, shared, and transmitted by means (lit. things in the middle, media) that include language (in the broadest sense of the term), practices of communication and representation, and material objects. This is an epistemological, not an ontological view: When we conceive of an ensemble of mediations as culture this is not to posit culture as an entity. Culture is a discourse on mediations and practices. In this view, mediations are objectifications of human action somewhat but (as we shall argue) not quite in the sense envisioned by Ricoeur when he stated

that "action itself, action as meaningful, may become an object of science, without losing its character of meaningfulness, through a kind of objectification similar to the fixation which occurs in writing. By this objectification action … constitutes a delineated pattern which has to be interpreted according to its inner connections" (1971: p. 538). The analogy with fixation in writing (the reason to introduce the model of the text in the first place) "prepares the detachment of the *meaning* of the action from the *event* of the action" (1971: p. 538). In the passages from which these quotations are drawn Ricoeur deploys the vocabulary of Weberian–Parsonian culturalism (pattern, meaning, meaningful action, interaction) and thereby advocates interpretation as an operation on elements of culture that, though objectified, are disembodied – a point often made in the critique of American culturalism and echoed in Bourdieu's (1977: p. 96) qualifying the hermeneutic stance as 'objectivism.'

In contrast, approaches to text and interpretation in anthropology that, explicitly or not, focus on mediation and on culture as praxis are concerned with maintaining or restoring the event-character of action. Interpretation then becomes a method of translating the texts and documents, that is, those 'fixations in writing' that are the 'data' produced by interactive, communicative ethnography (notes, recordings, collected texts, including some that made their way into historical or administrative archives), into events and eventually into processes. Not the model of 'the' text but models of text making are required for such interpretive work, much of which has been inspired by Dell Hymes's (1974: orig. 1964) seminal observations on components of speech events. His ideas, sometimes called 'sociolinguistic,' were further developed by the contributors to Bauman and Sherzer (1974). Linking the notion of speech event to that of genre, Dan Ben Amos has shown methodological consequences for folklore studies (1976), and a classic in this vein is Dennis Tedlock's collection of essays (1983). Fabian has stated the philosophical (anti-positivist) point of departure, offered a model of text production, and a warning against tendencies to grammaticalize Hymes's pragmatic approach to texts (1991, Chaps. 1, 3, 5).

Interpretation and Performance

The move away from semiological to pragmatic concerns reflected and promoted yet another understanding of interpretation. 'Interpretation' used to be a discipline and skill taught in speech and drama departments; today, many of them have become part of a multifaceted field called Performance Studies. Performance, as a theoretical concept usually paired with 'competence,' was adopted by Hymes from generative linguistics and applied to his ethnography of speaking but it undoubtedly gained its wider currency in anthropology in the more specific, theatrical understanding promoted, above all, by Victor Turner (stated most comprehensively in 1986). Actually, as his work shows, attention to performance in the interpretation of culture has a complex intellectual history. Turner began by giving focus and life to drama, a pervasive but mostly dead theatrical metaphor in social theory (for instance, in Parsonian functionalism–structuralism) that was at the origin of such crucial sociological terms as role, actor, and enactment. He saw drama above all in ritual, and came close (as did Max Gluckman) to postulating a distinctive type of premodern society whose social dynamic was not determined by linear accumulation of change and innovation but by cyclical enactment of a dramatic script. Undoubtedly, this presented a gain over static views of primitive or traditional societies. However, not unlike the model of culture as a text, the drama metaphor had its limits when it came to understand its practical consequences for the production of ethnographic knowledge. What was needed was an epistemological conception of performance as both a way in which a culture works and a method by which it may be studied and represented. Taking an increasingly literal approach to performance as theatrical performance (including experimental collaboration with Richard Schechner, a producer and theorist in the field of theater), Turner ended in an impasse that eventually led him to advocate a biological (neurological) position not unlike that earlier taken by Lévi-Strauss (see Turner, 1986: pp. 156–177). Nevertheless, his work helped to make anthropologists aware that large parts, perhaps most expressions, of cultural knowledge are performative in the sense that what they say is also constituted by how contents are enacted. An emerging 'ethnography of reading' (Boyarin, 1993) applies this to texts fixated in writing and opens to anthropology a wide field of interpretation no longer predicated on categorical distinctions between orality and literacy; for methodological consequences and procedures, see Finnegan (1992) and Tonkin (1992). What changed, and was gained, by including pragmatic (sociolinguistic) and performance aspects in methods of interpreting oral sources is best illustrated by comparing Vansina's classic work on oral tradition (1961) with its radical revision (1985).

Finally, it should be mentioned that a pragmatic and critical stance toward the interpretation of texts also leads to a re-evaluation of 'context,' a concept that has had a long history in anthropology. It was applied mainly in the validation of ethnography and often, much like text, in a metaphorical manner in admonitions to place specific facts or data in their historical, political, economic, religious, etc. 'context.' In-as-much as this suggested that, though relevant to an interpretation of findings, context was itself not subject to interpretation (at least not at the moment when it was invoked), this has become an untenable proposition. Context poses problems of interpretation, it does not solve them.

The Politics of Interpretation

A review of theoretical and practical views of interpretation in anthropology would not be complete without developments that emerged in the 1980s. They consisted of moves giving anthropology a 'literary turn,' an apt designation if it is understood as promoting attention to writing in the production of ethnography and anthropological knowledge. Not metaphorical invocations of text as a model, but calls for critical interpretation of ethnographies as texts were formulated by a group resolved to lift the veil under which modern, scientifically minded anthropology had kept ethnographic authorship (Marcus and Cushman, 1982, Clifford and Marcus, 1986, Clifford, 1988). Postmodernism became a convenient but doubtful label for this movement – doubtful because it obscures

its roots in decidedly modern advancements in the historiography of our discipline, in the theoretical confrontation with its colonial and postcolonial political role that had occupied critical anthropology since the 1960s, and, though less obviously, in feminist critique of anthropological discourse.

As regards interpretation, the subtitle of the Clifford and Marcus volume, 'The Poetics and Politics of Ethnography,' pointed to the directions further work could, and did, take. Along the lines of a hermeneutical understanding of the term, interpretation now applied to the canon of anthropological writing the tools developed by literary theory and criticism (exemplified by Geertz, 1988). Interpretation in the performative sense encouraged, beyond literary deconstruction, 'experimentation with genre' which resulted, if not quite in its demise, then in relativizing the status of the monograph as the authorized form of ethnography: narrative, dialogue, poetry, collage, and autobiographic inclusion of the ethnographer in his or her account have begun to shape much of anthropological writing. Perhaps it is not always easy to draw a line between hard-won freedom from scientism and gratuitous literary posing; postmodern ethnography is at its best when it is able to justify its choice of representational forms on epistemological grounds.

How an author translates practices of research into writing has become a foremost task of critical interpretation in anthropology. To be able to discern narrative plots and strategies, identify tropes and other rhetorical devices, and to recognize intertextual connections is now required when we discuss the validity of ethnographies. Conversely, though it would be foolish to cast away all our prior habits and techniques, field research can no longer rely largely on preconceived methods that once were thought to yield data without being themselves subject to constant interpretation (which would have made methods useless in a classical positivist research setup). Attempts to formulate these insights have reached the level of textbooks and handbooks that steer a difficult course between opening a back door to flat-footed methodology and scaring anthropological novices with hermeneutic hypersensitivity.

Distinguishable but never separable from attention to poetics is the concern with the politics of ethnography, that is, with relations of power that determine production of anthropological knowledge in all its phases from research to writing. Therefore the question of interpretation in anthropology cannot be raised without (taking a lesson from Michel Foucault) questioning the devices of control and domination deployed in the discourse anthropology pronounces on others. Interpretation may be 'just interpretation,' but neither as a hermeneutic alternative to scientism nor as a pragmatic method should it make anthropology complacent about its achievements.

Bibliography

Bauman, R., Sherzer, J. (Eds.), 1974. Explorations in the Ethnography of Speaking. Cambridge University Press, Cambridge, UK.

Bauman, Z., 1978. Hermeneutics and Social Science: Approaches to Understanding. Hutchinson, London.

Bernheim, E., 1970. Lehrbuch der historischen Methode und der Geschichtsphilosophie. Burt Franklin, New York (Reprint, original edition 1889).

Ben Amos, D. (Ed.), 1976. Folklore Genres. University of Texas Press, Austin, TX.

Bourdieu, P., 1977. Outline of a Theory of Practice. Cambridge University Press, Cambridge, UK.

Boyarin, J. (Ed.), 1993. The Ethnography of Reading. University of California Press, Berkeley, CA.

Clifford, J., 1988. The Predicament of Culture: Twentieth Century Ethnography, Literature, and Art. Harvard University Press, Cambridge, MA.

Clifford, J., Marcus, G.E. (Eds.), 1986. Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley, CA.

Dolgin, J.L., Kemnitzer, D.S., Schneider, D.M. (Eds.), 1977. Symbolic Anthropology: A Reader in the Study of Symbols and Meanings. Columbia University Press, New York.

Dundes, A., 1980. Interpreting Folklore. Indiana University Press, Bloomington, IN.

Fabian, J., 1991. Time and the Work of Anthropology: Critical Essays 1971–1991. Harwood Academic Publishers, Chur, Switzerland.

Finnegan, R., 1992. Oral Traditions and the Verbal Arts: A Guide to Research Practices. Routledge, London.

Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.

Geertz, C., 1988. Works and Lives: The Anthropologist as Author. Stanford University Press, Stanford, CA.

Gusdorf, G., 1973. L'Avenement des sciences humaines au siècle des lumières. Payot, Paris.

Habermas, J., 1988. On the Logic of the Social Sciences (S.W. Nicholsen, J.A. Stark, Trans.). Polity, UK.

Hymes, D., 1974. Foundations in Sociolinguistics: An Ethnographic Approach. University of Pennsylvania Press, Philadelphia, PA.

Marcus, G.E., Cushman, D., 1982. Ethnographies as texts. Annual Review of Anthropology 11, 25–69.

Palmer, R.E., 1969. Hermeneutics: Interpretation Theory in Schleiermacher, Dilthey, Heidegger, and Gadamer. Northwestern University Press, Evanston.

Ricoeur, P., 1970. Freud and Philosophy: An Essay in Interpretation (S. Savage, Trans.). Yale University Press, New Haven, CT.

Ricoeur, P., 1984. The model of the text: Meaningful action considered as a text. Social Research 51 (1–2), 185–218.

Tedlock, D., 1983. The Spoken Word and the Work of Interpretation. University of Pennsylvania Press, Philadelphia, PA.

Tonkin, E., 1992. Narrating Our Pasts: The Social Construction of Oral History. Cambridge University Press, Cambridge, UK.

Turner, V.W., 1986. The Anthropology of Performance. Performing Arts Journal Publications, New York.

Vansina, J., 1961. De la tradition orale. Essai de méthode historique. Musée Royal de l'Afrique Centrale, Tervuren.

Vansina, J., 1985. Oral Tradition as History. University of Wisconsin Press, Madison.

Interpretive Methods: Macromethods

R Stryker, University of Minnesota, Minneapolis, MI, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7878–7881, © 2001, Elsevier Ltd.

Abstract

This article uses Weber's work to elucidate key features of macrointerpretive methods. Although relying on a logic different from that of frequentist statistical methods, macrointerpretive methods are based on coherent methodological principles and procedures. The goal of all macrointerpretive method is to render large-scale social actions and processes intelligible. Imputation of chains of meaning, motivation, and consequent likely courses of action are one hallmark. Macrointerpretive work is profoundly historical, and is well suited to reconciling agency and structure. It exhibits a continuous interplay of ideas and evidence, and combines inductive with deductive reasoning to solve anomalies, explain historically specific event sequences, and seek limited historical generalizations that are value-relevant and culturally significant. Enriched by multiple analytic techniques including diverse formal and nonformal narrative and comparative methods, macrointerpretive scholarship embodies tensions between the prominence accorded to narrative vs. comparisons, the concrete and particular vs. the abstract and general, conceptual understanding vs. causal analysis, and subjectivism vs. objectivism. Although built around a diversity of techniques, macrointerpretive scholarship of all sorts is well suited to forming and empirically examining concepts, hypotheses, and theories that help us understand how social structures work, how they reproduce themselves, and how they change.

The goal of all interpretive methods, both macro and micro, is to make sense of some slice of reality – to render it understandable or intelligible. Since the work of Max Weber (1949, 1958, 1978) has been central to articulating macro interpretive methods, this essay uses Weber's work to elucidate key features of these methods and to help outline their key tensions.

Weber's Protestant Ethic *as an Illustrative Case*

Weber's (1958) *The Protestant Ethic* exemplifies the meaning-oriented (verstehende) and explanatory goals and strategies of macrointerpretive work. Noting a historical association between the Protestant reformation and the rise of capitalism in Western Europe, Weber outlines a puzzle. How could a religion leading to more *conformist* behavior than did Catholicism become associated with radically *innovative* economic behavior? Unraveling this historical and theoretical puzzle requires understanding the cultural beliefs occasioned by the Reformation and the dilemmas they created for the Protestant believer. Detailing exactly how the new beliefs provided for social control allows Weber to make sense of a culturally significant and historically specific pattern of institutional change.

To do this, Weber constructs the ideal types 'capitalist ethos' and 'Protestant ethic.' *Ideal types* are analytic tools for interpretive, explanatory work. The social scientist creates ideal types by abstracting elements from concrete historical situations, then combining the elements into a meaningful conceptual whole. The ideal type does not exist in empirical reality, but it is constructed inductively from that reality to illuminate something of cultural significance (Weber, 1958, p. 47,1949, pp. 91–111,1978, pp. 18–22).

According to Weber (1958: 51–68), the spirit of capitalism combines two elements: an absolute ethical duty to accumulate capital and to do so in a rigorously calculating rational way. The ideal-typical capitalist must increase his or her capital as an end in itself, methodically identifying and following courses of action that lead to profit-making, while strictly avoiding any enjoyment of life. Weber argues that any account of the rise of modern capitalism must explain how this ethos, which then helped transform economic institutions, evolved.

The ideal-type Protestant ethic analytically combines three elements (Weber, 1958: 79–154). First is Luther's idea of the calling – that people have a moral duty to their work. Second is ascetic mastery of the world, the idea that at each minute of every day, the individual must bring God's glory to fruition in the world, rather than by withdrawing to a monastery or nunnery. Third is the Calvinist doctrine of predestination, arguing that by the decree of God for the manifestation of God's glory, some people are predestined for grace and others for damnation. Only God knows which people are which, and nothing people do in their lives can change their preordained fate.

Once Weber has theoretically constructed his object of inquiry – the capitalist ethos – and his explanatory concept – the Protestant ethic – he creates the interpretive links tying the two together. He asks and answers the question, what does predestination *mean* and how is it likely to orient behavior for the people who believe in it, given that they *also* believe in the calling and ascetic mastery of the world? He then argues as follows.

> In its extreme inhumanity, this doctrine must above all have had one consequence for the life of a generation [that] surrendered to its magnificent consistency. This was a feeling of unprecedented inner loneliness of the single individual. In what was for man of the age of the Reformation the most important thing in life, his eternal salvation, he was forced to follow his path alone to meet a destiny which had been decreed for him from eternity … the complete elimination of salvation through the Church and the sacraments was what formed the absolutely decisive difference from Catholicism.
>
> (Weber, 1958: 104)

Attaining salvation is so central that believers will question constantly whether they are saved. Since there is no way of

knowing, most people will feel intolerable psychological strain. Strain is amplified because at the same time that participation in Protestant churches and rituals cannot guarantee salvation, believers are told by religious leaders that it is "an absolute duty to consider oneself chosen … lack of self confidence is the result of insufficient faith, hence of imperfect grace" (Weber, 1958: 111). It becomes imperative to resolve this 'Catch-22' situation and justify certainty in one's salvation.

What will the believer do? Faced with a preordained outcome, some people might become fatalistic and others might commit suicide. To explain why believers do not take these paths, Weber situates predestination in its broader belief context. Because of the conjunction of belief in predestination with beliefs in the calling and ascetic mastery of the world, the Protestant believer is likely to try and relieve his or her incredible psychological insecurity by bringing God's glory to fruition through efficacious action in a calling in the world. 'Good works' cannot bring salvation, but they *do* provide an effective way to get rid of the fear of damnation, assuming that the believer lives an entire "life of good works combined into a unified system" (Weber, 1958: 115–117).

In short, predestination and the need for proof of salvation provide a psychological basis for constant self-regulation and the rational ordering of moral life. The other elements of the Protestant ethic explain how and why the solution to the need for proof will be found in the systematic rationalization of moral life in the world rather than being found elsewhere. As a meaningfully interpreted whole, the Protestant ethic is powerful brew. The interpretive link to the capitalist spirit now is clear: purposive-rational orientation to capital accumulation as an end in itself is a strategy to provide the believer with a clear, continuous and unambiguous sign of his or her own salvation. As a collective attribution of meaning and action orientation, the strategy – and the macro religious structure that created it – unintentionally reorders economic behavior and institutions.

Key Features and Tensions in Macrointerpretive Work

As illustrated by Weber, imputation of chains of meaning, motivation, and consequent likely courses of action are a hallmark of macrointerpretive explanation. Because actors are front and center, with social institutions providing the cultural scripts that orient actors' attributions of meaning and their behavior, macrointerpretive work is ideally suited to reconciling agency and structure (Sewell, 1996). Similarly, macrointerpretive work is profoundly historical, although it is not necessarily historicist in the narrowest meaning of that term.

Because most macrointerpretive work is concerned with imputing actors' meanings, culture and ideology loom large as objects of inquiry or as explanatory factors (Wuthnow, 1989; Mahoney, 1999). However, macrointerpretive methods also signal the interrelated construction of historical narratives and social science concepts to interpret historical wholes, processes or particularities, so that they are rendered meaningful to us as social scientists (Ragin, 1987). These two analytically distinct aspects of macrointerpretive work tend to be found together, as in Weber's *Protestant Ethic*, but they sometimes appear separately (Mahoney, 1999).

Macrointerpretive scholarship is enriched by multiple analytic techniques including diverse formal and nonformal narrative and comparative methods (see, e.g., Ragin, 1987, 2000; Griffin, 1993; Stryker, 1996; Mahoney, 1999). Some scholars view strategies such as narratives stressing sequence, path dependence, and 'eventful temporality' and comparisons emphasizing the divergent structural and cultural contexts of events as alternative approaches to doing macrointerpretive work (Griffin, 1992; Sewell, 1996). However, others argue these strategies are complementary (Quadagno and Knapp, 1992; Stryker, 1996; Mahoney, 1999).

Like Weber's own work, macrointerpretive scholarship embodies a tension between the concrete and particular and the abstract and general (compare Bendix, (1978) with Moore, (1966)). Often, macrointerpretivists resolve the tension between specificity and generality, as did Weber, by using the concepts they create to resolve puzzles and explain historically specific events and event sequences, while also seeking limited historical, or conditional, generalizations (Quadagno and Knapp, 1992; Stryker, 1996; Mahoney, 1999; Paige, 1999). The underlying assumption about the relationship between history and social science theory guiding macrointerpretive work is that history is theorized at the same time as theory is historicized. In other words, the two are 'mutually constructed' and 'mutually adjusted' so that they are 'logically interdependent processes' rather than 'analytically independent things' (Stryker, 1996: 312).

Macrointerpretive work often is portrayed as wholly inductive. But as Weber's *Protestant Ethic* demonstrates, it cannot be characterized fairly in this way. Neither can its practitioners be characterized as formalistically applying Mill's method of agreement, indirect method of difference or some combination of the two. Notwithstanding much literature describing comparative variants of macrointerpretive methods in terms of Mill's canons, Goldstone (1997: 108) shows why 'Mill's methods cannot be used and are not used by comparative case-study analysis.' Instead, Goldstone, (1997) likens case-oriented comparative methods to sophisticated detective work. Like all macrointerpretive analysis, these methods involve a constant interplay of ideas and evidence to solve anomalies, they carefully combine inductive with deductive reasoning, and they involve attending to the complexities of multiple and conjunctural causation (Stryker, 1996; Goldstone, 1997; Ragin, 1987). A final hallmark of macrointerpretive scholarship is selection of research topics according to their 'value-relevance' and 'cultural significance' for the scholar and intended audience (Weber, 1949: 76–77; Sewell, 1996; Mahoney, 1999).

This run-down of multiple currents and tensions within macrointerpretive work is not exhaustive, but it should suffice to show that macrointerpretivists debate (a) whether or not interpretation must involve causal explanation as well as imputing meaning and (b) the prominence that should be accorded to narrative as opposed to comparative techniques. Much, but not all, macrointerpretive work is explicitly or implicitly causal-analytic (Skocpol and Somers, 1980; Ragin, 1987; Goldstone, 1997). Similarly, much macrointerpretive work is explicitly or implicitly comparative, even as it relies heavily on narrative strategies (Skocpol and Somers, 1980; Mahoney, 1999). For a good introduction to diverse causal-analytic strategies and to the multiplicity of comparative

techniques employed by macrointerpretivists, see Skocpol and Somers (1980), Skocpol (1984), Ragin (1987), Rueschemeyer et al. (1992), Goldstone (1997), and Mahoney (1999). For a good introduction to the diversity of narrative techniques, including very different concepts of narrativity and its role in macrointerpretive social science, see Stone (1979), Griffin (1993), Sewell (1996), Stryker (1996), and Somers (1997).

A final tension involves questions of subjectivity and objectivity in social science. Some contemporary scholars – notably feminist standpoint theorists such as Smith (1990) and Collins (1990) – suggest that it is neither possible nor desirable to transcend the lived experience of the actors one studies. Meanings should be imputed through dialogic techniques practiced with real subjects or, if this is not possible, with historical texts, so that the actors studied would recognize and verify these meanings as their own.

Deconstructionist critiques such as these have been useful to elucidate how class, race, gender, and other categories of domination and subordination are implicated in the production of social science knowledge. However, most macrointerpretive scholars follow Weber (1949) in assuming that *interpretive objectivity* is both possible and desirable. Although there is substantial disagreement over which concepts and procedures are best, most macrointerpretive scholars presume that if the analyst is clear about his or her concepts and how they are operationalized and if he or she employs an explicit set of procedures to establish the interpretations he or she makes, he or she can arrive at interpretive objectivity. Here Weber's (1949, 1978) discussion of the use of ideal types, counterfactuals, and principles of objective possibility to arrive at explanations that are causally adequate and adequate on the level of meaning is seminal. Contemporary promotion of techniques of 'strategic narrative,' (Stryker, 1996), of formal qualitative comparative analysis and its extension into fuzzy set logic (Ragin, 1987, 2000), and of event structure analysis (Griffin, 1993), can be seen as updates on Weber's (1949) rules for achieving objectivity in macrointerpretive social science. In these extensions, issues of validity, reliability, and replicability are emphasized.

Conclusion

In sum, there is no one macrointerpretive method. Macrointerpretive scholarship is built around a diversity of techniques. Both old methodological stalwarts and new developments are playing their role in generating the culturally significant findings of macrointerpretive research. In the twenty-first century, there is an enormous body of writing illuminating the rich variety found within macrointerpretive methods. What all varieties share with Weber's *Protestant Ethic* is an emphasis on imputing meaning to render largescale social actions and processes intelligible. As comparisons of such diverse studies as Wallerstein's (1974) on the world system and Somers (1997) on English working-class formation suggest, macrointerpretive scholarship of all sorts is well suited to forming and empirically examining concepts, hypotheses, and theories that help us understand how social structures work, how they reproduce themselves and how they change. Although macrointerpretive scholarship relies on analytic logic completely different from that of frequentist statistical methods, it proceeds based on coherent methodological principles and procedures.

See also: Archival Methods; Case Study: Methods and Analysis; Case-Oriented Research; Causal Counterfactuals in Social Science Research; Classification, Understandings of in the Social Sciences; Comparative Studies: Method and Design; Content Analysis; Explanation: Conceptions in the Social Sciences; Generalization: Conceptions in the Social Sciences; Grounded Theory: Methodology and Theory Construction; Idealization, Abstraction, and Ideal Types; Interpretive Methods: Micromethods; Narrative, Sociology of; Weber, Max (1864–1920).

Bibliography

Bendix, R., 1978. Kings or People? Power and the Mandate to Rule. University of California Press, Berkeley, CA.
Collins, P.H., 1990. Black Feminist Thought: Knowledge, Consciousness and Empowerment. Unwin-Hyman, Boston.
Goldstone, J., 1997. Methodological issues in comparative macrosociology. Comparative Social Research 16, 107–120.
Griffin, L.J., 1992. Temporality, events and explanation in historical sociology: An introduction. Sociological Methods and Research 20, 404–427.
Griffin, L.J., 1993. Narrative, event-structure and causal interpretation in historical sociology. American Journal of Sociology 98, 1094–1133.
Mahoney, J., 1999. Nominal, ordinal, and narrative appraisal in macrocausal analysis. American Journal of Sociology 104, 1154–1196.
Moore, B., 1966. The Social Origins of Dictatorship and Democracy. Beacon Press, Boston.
Paige, J.M., 1999. Conjuncture, comparison, and conditional theory in macrosocial inquiry. American Journal of Sociology 105, 781–800.
Quadagno, J., Knapp, S., 1992. Have historical sociologists forsaken theory? Thoughts on the history/theory relationship. Sociological Methods and Research 20, 481–507.
Ragin, C., 1987. The Comparative Method. University of California Press, Berkeley, CA.
Ragin, C., 2000. Fuzzy-set Social Science. University of Chicago Press, Chicago.
Rueschemeyer, D., Huber Stephens, E., Stephens, J.D., 1992. Capitalist Development and Democracy. University of Chicago Press, Chicago.
Sewell Jr., W., 1996. Three temporalities: Toward an eventful sociology. In: McDonald, T.J. (Ed.), The Historic Turn in the Human Sciences. University of Michigan Press, Ann Arbor, MI, pp. 245–280.
Skocpol, T. (Ed.), 1984. Vision and Method in Historical Sociology. Cambridge University Press, New York.
Skocpol, T., Somers, M., 1980. The uses of comparative history in macrosocial inquiry. Comparative Study of Society and History 22, 174–197.
Somers, M., 1997. Deconstructing and reconstructing class formation theory: Narrativity, relational analysis, and social theory. In: Hall, J.R. (Ed.), Reworking Class. Cornell University Press, Ithaca, NY, pp. 73–105.
Smith, D.E., 1990. The Conceptual Practices of Power: A Feminist Sociology of Knowledge. Northeastern University Press, Boston.
Stone, L., 1979. The revival of narrative: Reflections on a new old history. Past and Present 95, 3–25.
Stryker, R., 1996. Beyond history versus theory: Strategic narrative and sociological explanation. Sociological Methods and Research 24, 304–352.
Wallerstein, I., 1974. The Modern World System. Academic Press, New York.
Weber, M., 1949. The Methodology of the Social Sciences. Free Press, New York [1903–17].
Weber M, 1958. The Protestant Ethic and the Spirit of Capitalism. Scribner, New York [1904].
Weber, M., 1978. In: Roth, G. (Ed.), Economy and Society. University of California Press, Berkeley, CA [1921] Vol. 1.
Wuthnow, R., 1989. Communities of Discourse: Ideology and Social Structure in the Reformation, the Enlightenment and European Socialism. Harvard University Press, Cambridge, MA.

Interpretive Methods: Micromethods

Norman K Denzin, University of Illinois at Urbana-Champaign, IL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7881–7885, © 2001, Elsevier Ltd.

Abstract

Recent developments in qualitative research are analyzed. Qualitative inquiry is located within an interdisciplinary and historical framework. The uses and forms of qualitative inquiry are reviewed, including those forms associated with participative, action research. Speculations on the future of interpretive, mirco-methods are offered.

In this article, recent developments in the uses and representations of interpretive, research methods are analyzed. Micromethods are focused on those interpretive approaches which yield close readings of texts and social situations. This will require serious attention to forms of narrative analysis, and the uses to which narrative material can be put by a society and its members, including social scientists. Accordingly, my intentions are fourfold. First, to analyze the problems and issues confronting the use of micro-qualitative methods and research evidence. Second, to define and locate this sprawling, interdisciplinary field of qualitative inquiry within the history of interpretive methods in the United States in the twentieth century. Third, to discuss the multiple uses and forms of qualitative inquiry, including its uses in participatory, applied action research settings. Speculations on where interpretive narrative inquiry will go in the twenty-first century form the conclusion.

The Field of Interpretive Research and its History

Interpretive research methods produce descriptions and accounts about the ways of life of the writer and those written about. There are multiple forms, or strategies of qualitative research (see below), including ethnography, participant observation, and case study analysis. The use of these strategies is connected to specific methodological practices, ranging from interviewing, to participant observation, to the use of visual, narrative, and personal experience methods. These methods generate specific types of empirical materials. These materials come in multiple forms, ranging from transcribed interviews and field observations, to life stories, cultural texts (stories, music, cinema, soap operas, TV news, televised sporting events), the artifacts and symbols of material culture (building and burial sites, pottery, tools, clothing), photographs, archival records, biographical, and historical documents. Traditionally, these materials are seen as producing the evidence that qualitative researchers use when they formulate interpretations of social and cultural phenomenon. Three interconnected activities define the interpretive research process. They go by a variety of labels or names, including theory, method, and analysis. Behind these terms stands the researcher who has a gendered biography, and who speaks from a particular class, race, cultural, and ethnic perspective. This researcher confronts the world with a set of ideas, or basic beliefs. These beliefs can be called a paradigm (Guba, 1990). A paradigm consists of an epistemology, an ontology, and a set of research strategies. Specific types of methodological practice are specified by a paradigm's epistemology. Ontology raises basic questions about reality and the nature of human nature. Methodology addresses the issue of how we gain knowledge about the world. Today in qualitative inquiry multiple paradigms, each with its particular ontology, epistemology, and methodology, compete for attention. These paradigms include logical postivism, postpositivism, constructionism and feminism, Marxism and critical theory, cultural studies and ethnic models. (Logical positivists believe in an external reality that can be apprehended by the methods of objective science. In contrast, postpositivists assert that there may be an external reality, but it cannot be known perfectly.)

The positivist paradigm, within interpretive methodology, presumes an objective world that can be studied with objective, micro-research methods. The constructivist paradigm presumes multiple realities, and contends that the researcher and those studied create understandings about the world and its meanings. Terms like trustworthiness, credibility, transferability, dependability, and confirmability replace the traditional positivist criteria of internal and external reliability and validity. Feminist, ethnic, critical theory, and cultural studies models privilege a real world that makes a material difference in the lives of men, women, and children. Interpretive methods are used, and researchers experiment with new ways of writing and presenting empirical evidence.

These paradigms, which structure qualitative research, operate in a complex historical field which cross-cuts five historical moments. These five moments simultaneously operate in the present. They are defined as the *traditional* (1900–50), the *modernist*, or golden age (1950–70), *blurred genres* (1970–86), the *crisis of representation* (1986–90) and *postmodern*, or present moments (1990–present). The present moment is defined by the refusal to privilege any method or theory, or methods claim to validity.

Successive waves of critical, epistemological discourse move across these five moments. The *traditional period* is associated with the logical positivist, quantitative research paradigm. In this period qualitative researchers attempted to justify their research methods in terms of traditional criteria of reliability and validity. Researchers sought to maintain a detached, objective stance toward the world and their objects of study. Research methods were the means by which evidence about the

world was collected. The other who was studied was alien, foreign, and strange.

The field worker, during this period, was lionized, made into a larger-than-life figure who went into and then returned from the field with stories about strange people. These accounts were structured by the norms of classical ethnography. This sacred bundle of terms (Rosaldo, 1989: p. 31) organized ethnographic texts in terms of four beliefs and commitments: a commitment to objectivism, a complicity with imperialism, a belief in monumentalism (the ethnography would create a museum-like picture of the culture studied), and a belief in timelessness (what was studied never changed). This model of the researcher, who could also write complex, dense theories about what was studied, holds to the present day.

The *modernist* and *blurred genres* moments are connected to the appearance of postpositivist arguments. The logic of multiple methods (and triangulation) was adopted, embodied in the development of empirically grounded theory. During this period, researchers started to move away from the detached observer stance of the traditional moment. At the same time a variety of new interpretive, qualitative perspectives made their presence felt, including hermeneutics, structuralism, semiotics, phenomenology, cultural studies, and feminism.

The modernist phase or second moment builds on the canonical works from the traditional period. Social realism, naturalism and slice-of-life ethnographies are still valued. This was the golden age of rigorous qualitative analysis.

By the beginning of the third stage (1970–86), *'Blurred Genres,'* qualitative researchers had a full complement of paradigms, methods, and strategies to employ in their research. Geertz's two books, *Interpretation of Cultures* (1973) and *Local Knowledge* (1983), defined the beginning and end of this moment. In these two works he argued that the old functional, positivist, behavioral, totalizing approaches to the human disciplines was giving way to a more pluralistic, interpretive, open-ended perspective. This new perspective took cultural representations and their meanings as its point of departure. Calling for 'thick descriptions' of particular events, rituals, and customs, Geertz suggested that all qualitative writings were interpretations of interpretations. The observer had no privileged voice in the interpretations that were written. The central task of theory was to make sense out of a local situation.

In the blurred genre phase the humanities became central resources for critical, interpretive theory, and the interpretive research project broadly conceived. The blurred genres phase produced the next stage, the crisis of representation. Here researchers struggled with how to locate themselves, their empirical materials, and their subjects in reflexive texts. It became increasingly clear that research methods were not neutral tools that produced value-free observations. Rather all observations were understood to be theory-laden; facts were not independent of the observer's values; and observers could not sustain an objective, detached view of themselves, the world, and the research process.

The *fifth moment* is the present. It is characterized by a sensibility which doubts all previous paradigms and historical moments. Theories are now read in narrative terms, as tales from the field (Geertz, 1988). Preoccupations with the representation of the 'Other' remain. New epistemologies from previously silenced groups emerge to offer solutions to this problem. The concept of the aloof researcher has been abandoned. Action, activist-oriented research is on the horizon, as are more social criticism and social critique. The search for grand narratives is being replaced by more local, small-scale theories fitted to specific problems and specific situations.

Reading History

The following conclusions can be drawn from this brief history. Each of the earlier historical moments continues to operate in the present, either as legacy, or as a set of practices that researchers follow or argue against. The multiple and fractured histories of qualitative research make it possible for any given researcher to attach a project to a canonical text from any of the previous historical moments. Multiple criteria of evaluation now compete for attention. There have never been so many paradigms, strategies of inquiry, or methods of analysis to draw upon and utilize. Researchers are in a moment of discovery and rediscovery, as new ways of looking, interpreting, arguing, and writing are debated and discussed. The qualitative research act can no longer be viewed from within a neutral or objective perspective. Class, race, gender, and ethnicity shape the process of inquiry, making research a multicultural process.

Interpretive Inquiry as a Site of Multiple Methodological Research Practices

Any definition of micro-interpretive research and the evidence generated by qualitative methods must work within this complex historical field. Qualitative research and qualitative evidence mean different things in each of these moments. Nonetheless, an initial, generic definition can be offered.

Interpretive inquiry is multimethod in focus, involving a naturalistic approach to its subject matter. This means interpretive researchers study things in their natural settings, attempting to make sense of these things in terms of the meanings people bring to them. *Qualitative research* involves the studied use and collection of case study, personal experience, introspective, life story, interview, observational, historical, interactional, and visual texts which describe routine and problematic moments and meanings in individual's lives.

Qualitative research, as a set of interpretive practices, privileges no single methodology over another. It has no theory, or paradigm that is distinctly its own. Multiple theoretical paradigms, as argued above, claim use of qualitative research methods and strategies, from constructivism, to cultural studies, feminism, Marxism, and ethnic models of study. Qualitative research does not belong to a single discipline.

These separate and multiple uses and meanings of the methods of qualitative research make it difficult to agree on any essential definition of the field, for it is never just one thing. Still a definition must be made. I borrow from, and paraphrase Nelson, Treichler, and Grossberg's attempt to define cultural studies (1992: p. 4).

> Qualitative research is an interdisciplinary, counter-disciplinary field. It cross-cuts the humanities, the social and the physical sciences. Qualitative research is many things at the same time. It is multi-paradigmatic in focus. Its practitioners are sensitive to the value of the multi-method approach. They are committed to the

naturalistic perspective; and to the interpretive understanding of human experience. At the same time the field is inherently political and shaped by multiple ethical and political positions. Qualitative research embraces two tensions at the same time. On the one hand it is drawn to a broad, interpretive, critical sensibility. On the other hand it is drawn to more narrowly defined positivist, postpositivist, humanistic and naturalistic conceptions of human experience and its analysis.

This rather awkward statement means that qualitative research, as a set of interpretive practices, embraces within its own history, constant tensions, and contradictions over the project itself, including its methods, and the forms its findings and interpretations take.

The Crisis of Interpretation

Today qualitative researchers confront a double crisis of representation and legitimation. Embedded in the discourses of poststructuralism and postmodernism (Denzin, 1997: p. 6; Lather, 1993), these two crises are, as Lather (1993) notes, coded in multiple terms, variously called and associated with the narrative, 'interpretive, linguistic, and rhetorical turns' in the social sciences. This linguistic turn makes problematic two key assumptions of social theory and interpretive research. The first assumption, as suggested above, presumes that theorists and researchers can no longer directly capture lived experience; such experience, it is argued, is created in the social text written by the researcher. This is the representational crisis. It confronts the inescapable problem of representation, but does so within a framework that makes the direct link between experience and text problematic.

The second assumption makes the traditional criteria for evaluating and interpretive theory and research problematic. This is the legitimation crisis. It involves a serious rethinking of such terms as validity, generalizability, and reliability, terms already re-theorized in postpositivist, constructionist-naturalistic, feminist, and interpretive discourses. This crisis, asks, 'How are interpretive, ethnographic studies to be evaluated in the poststructural moment?' Clearly these two crises blur together.

The Representational Crisis

A single but complex issue defines the representational crisis. It involves the assumption that much if not all social science and ethnographic writing is a narrative production, structured by a logic that separates writer, text, and subject matter. Any social text can, accordingly, be analyzed in terms of its treatment of four paired terms: (1) the 'real' and its representation in the text, (2) the text and the author, (3) lived experience and its textual representations, and (4) the subject and his or her intentional meanings. The text presumes a world out there (the real), that can be captured by a 'knowing' author through the careful transcription (and analysis) of field materials (interviews, notes, etc.). The author becomes the mirror to the world under analysis. This reflected world then re-presents the subject's experiences through a complex textual apparatus which typically mingles and mixes multiple versions of the subject. The subject is always a textual construction, for the 'real' flesh-blood person is always translated into either an analytic subject as a social type, or a textual subject who speaks from the author's pages.

Qualitative researchers have historically assumed that their methods probe and reveal lived experience. They have also assumed that the subject's word is always final, and that talk directly reflects subjective-lived experience. The literal translation of talk thus equals lived experience and its representation.

Poststructuralism challenges these assumptions. Language and speech do not mirror experience, they create it and in the process of creation constantly transform and defer that which is being described. The meanings of a subject's statements are, therefore, always in motion. There can never be a final, accurate representation of what was meant or said, only different textual representations of different experiences. As Lather (1993: p. 3) observes, these arguments do not put an end to representation, they signal instead the end of pure presence (Lather 1993: p. 3). The task at hand is to understand what textually constructed presence means, since there is only ever the text. This leads to the question of a text's authority.

The Legitimation Crisis

Many contemporary qualitative researchers challenge post-positivist arguments concerning the text and its validity. Lather (1993) and Scheurich (1997: p. 84) argue that validity is a text's call to authority and truth. Lather calls this version of validity epistemological. That is, a text's authority is established through recourse to a set of rules concerning knowledge, its production, and representation. These rules, as Scheurich (1997) notes, if properly followed, establish validity. Without validity there is no truth, and without truth there can be no trust in a text's claims to validity. With validity comes power, and validity becomes a boundary line 'which divides good research from bad, separates acceptable (to a particular research community) research from unacceptable research … it is the name for inclusion and exclusion' (Scheurich, 1997: p. 84).

Poststructuralism reads the discussions of logical, construct, internal, ethnographic and external validity, text-based data, triangulation, trustworthiness, credibility, grounding, naturalistic indicators, fit, coherence, comprehensiveness, plausibility, truth, relevance, as attempts to reauthorize a text's authority in the postpositivist moment. Such moves still hold (all constructionist disclaimers aside) to the conception of a 'world-out-there' that is truthfully and accurately captured by the researcher's methods and written text.

These words, and the methodological strategies that lie behind them, represent attempts to thicken and contextualize a work's grounding in the external empirical world. They represent efforts to develop a set of transcendent rules and procedures that lie outside any specific research project. These rules, if successfully followed, allow a text to bear witness to its own validity. Hence a text is valid if it is sufficiently grounded, triangulated, based on naturalistic indicators, carefully fitted to a theory (and its concepts), comprehensive in scope, credible in terms of member checks, logical, truthful in terms of its reflection of the phenomenon in question. The text's author then announces these validity claims to the reader. Such claims

now become the text's warrant to its own authoritative representation of the experience and social world under inspection.

Resistances to Qualitative Studies

The academic and disciplinary resistances to interpretive, micro-research further illustrate the politics embedded in this field of discourse. The challenges to qualitative research are many. Qualitative researchers are called journalists or soft scientists. Their work is termed unscientific, or only exploratory, or entirely personal and full of bias. It is called criticism not theory, or it is interpreted politically, as a disguised version of Marxism, or Humanism.

These resistances to qualitative inquiry reflect an uneasy awareness that its traditions commit one to a critique of the positivist project. But the positivist resistance to qualitative research goes beyond the 'ever-present desire to maintain a distinction between hard science and soft scholarship' (Carey, 1989: p. 99). The positive sciences (physics, economics, and psychology) are often seen as the crowning achievements of Western civilization, and in their practices it is assumed that truth can transcend opinion and personal bias (Carey, 1989: p. 99). Qualitative inquiry is seen as an assault on this tradition, leading some positivists to retreat into a 'value-free objectivist science' (Carey, 1989: p. 104) model to defend their position. But these critics seldom attempt to make explicit, and critique the 'moral and political commitments in their own contingent work' (Carey, 1989: p. 104).

The opposition to positive science by the postpositivists and the poststructuralists is seen as an attack on reason and truth. At the same time, the positive science attack on qualitative research is regarded as an attempt to legislate one version of truth over another.

Action Research

In the contemporary period greater use is being made of qualitative evidence in participatory and applied action programs. Life stories, case histories, and personal narratives are often used as key documentary evidence to establish the need for social change. Such materials are also used as proof that an applied action program works. Indeed qualitative (and narrative) evidence lends itself especially well to those applied action programs which stress the subjective, reflexive dimensions of social experience. They further imply that such materials are critical to the implementation of participatory action research as a practical, collaborative, participatory, emancipatory, and reflexive project. First-person narrative texts allow Third World and indigenous persons to share in the ownership of the research endeavor. Such texts anchor research in the context of ongoing community life. They help community members develop a shared orientation toward collective action aimed at addressing social injustice.

Participatory ethno-dramas, forum, and political theater can be used to mobilize community consciousness around an instance of perceived injustice or repression. For example, Brecht's Epic or Dialectical Theatre is deliberately disruptive and political, and thus anticipates the more radical forms of contemporary postmodern, political theater that has connections back to Artaud's concepts of a pure theatre (Birringer, 1993: p. 217). Third World popular theater has been used by 'oppressed Third World people to achieve justice and development for themselves' (Etherton, 1988: p. 991). This form of dramatic representation extends what theater and narrative can do politically.

In Conclusion

The legacies of the present move in several directions at the same time. They all turn on social texts and narratives, their uses and how they are read. The narrative turn in the social sciences presumes that social texts, including recorded or visual texts, interviews, fieldwork notes, transcribed conversations, speeches, film, music, advertisements, personal, electronic, and archival documents, can be rigorously and scientifically analyzed through the methods of narrative analysis.

But Trinh (1989: p. 141) warns us that traditional, empiricist narrative methods represent an approach to storytelling that must be avoided. They turn the story told into a story analyzed. In so doing they sacrifice meaning for analytic rigor. They only hear and read the story from within a set of predetermined structural categories. They do not hear the story as it was told. The goal is to recover these lost stories.

See also: Interpretive Methods: Macromethods; Thick Description: Methodology.

Bibliography

Birringer, J., 1993. Theatre, Theory, Postmodernism. Indiana University Press, Bloomington, IN.
Carey, J.W., 1989. Culture as Communication. Unwin Hyman, Boston.
Denzin, N.K., Lincoln, Y.S., 1994. Introduction. In: Denzin, N.K., Lincoln, Y.S. (Eds.), Handbook of Qualitative Research. Sage, Thousand Oaks, CA, pp. 1–17.
Denzin, N.K., 1997. Interpretive Ethnography: Ethnographic Practices for the 21st Century. Sage, Thousand Oaks, CA.
Etherton, M., 1988. Third World popular theatre. In: Banham, M. (Ed.), The Cambridge Guide to Theatre. Cambridge University Press, Cambridge, UK, pp. 991–992.
Geertz, C., 1973. Interpretation of Cultures: Selected Essays. Basic Books, New York.
Geertz, C., 1983. Local Knowledge. Basic Books, New York.
Geertz, C., 1988. Works and Lives. Stanford University Press, Stanford, CA.
Guba, E.G., 1990. The alternative paradigm dialog. In: Guba, E.C. (Ed.), The Paradigm Dialog. Sage, Newbury Park, CA, pp. 17–30.
Lather, P., 1993. Fertile obsession: validity after poststructuralism. Sociological Quarterly 34, 673–693.
Nelson, C., Treichler, P.A., Grossberg, L., 1992. Cultural studies. In: Grossberg, L., Nelson, C., Treichler, P.A. (Eds.), Cultural Studies. Routledge, New York, pp. 1–16.
Rosaldo, R., 1989. Culture and Truth. Beacon, Boston.
Scheurich, J.J., 1997. Research Method in the Postmodern. Falmer Press, Washington, DC.
Trinh, T Minh-ha, 1989. Woman, Native, Other: Writing Postcoloniality and Feminism. Indiana University Press, Bloomington, IN.

Interrupted Time Series Quasi-Experimental Designs

Richard McCleary, University of California, Irvine, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The time series quasi-experiment was proposed initially by Campbell (1963) as a means of assessing the causal effect of a discrete intervention on a social science time series. Although the design had been widely used in the physical sciences, social science applications presented idiosyncratic inferential problems. Addressing these problems as threats to internal and external validity, Campbell (1963) concluded that the time series design was the strongest of the common quasi-experimental designs. Addressing threats to statistical conclusion and construct validities as well as internal and external validities, Cook and Campbell (1979) subsequently concurred with the optimistic validity assessment for the design.

In the simple time series quasi-experiment described by Campbell (1963), a binary variable X_t breaks the time series Y_t into pre- and post-intervention segments of N_{pre} and N_{post} observations. Thus

$$Y_t = \mu + \omega X_t + u_t \qquad [1]$$

The u_t term is a strictly exogenous time series of 'white noise' innovations:

$$u_t \sim \text{iid } N\left(0, \sigma^2\right) \qquad [1a]$$

Scalar parameters μ and ω are interpreted, respectively, as the pre-intervention intercept and the post-intervention change in intercept. Rejecting the null hypothesis

$$H_0 : \omega = 0 \qquad [1b]$$

leads to the inference that X_t affected a change in the intercept of Y_t. The comment label for all quasi-experimental models of this sort is 'interrupted time series model.'

Since most social science time series are nonstationary, and since the effects of social interventions are often dynamic, time series quasi-experiments were not widely used until Box and Jenkins (1970; Box and Tiao, 1975) generalized the simple model [1] to account for nonstationary time series and dynamic effects. The Box–Jenkins–Tiao model can be written as

$$Y_t = f(u_t) + g(X_t) \qquad [2]$$

Here, $f(u_t)$ is an empirically specified *Auto*Regressive *I*ntegrated *M*oving *A*verage (ARIMA) model; $g(X_t)$ is a theoretically specified dynamic transfer model that distributes the scalar effect parameter ω over the post-intervention segment. Component models f and g are constructed by rules that guarantee the existence of inverses f^{-1} and g^{-1}.

The Box–Jenkins–Tiao model [2] controls not only the common threats to statistical conclusion validity, but also some of the most plausible threats to internal validity. Threats to construct validity are controlled by the dynamic transfer component, $g(X_t)$. The common threats to external validity are controlled by replication across settings. Threats to the four validities and the models and methods used to control these threats are illustrated with examples drawn from the applied literature.

Statistical Conclusion Validity

The nine common threats to statistical conclusion validity described by Shadish et al. (2002, p. 45) are spurious inferences associated with misstatements of the Type I and Type II error rates. Type I errors occur when H_0 is true (i.e., when $\omega = 0$) but H_0 is mistakenly rejected. In practice, the Type I error rate is fixed a priori at a conventional level, usually $\alpha = 0.05$ corresponding to 0.95 confidence.

The Type I threats to statistical conclusion validity accrue a consequence of stochastic time series properties, particularly trend or drift and autocorrelation, that are represented explicitly in the $f(u_t)$ component of [2]. For the lag operator $L^k Y_t = Y_{t-k}$, and the difference operator $\nabla^d = 1 - L^d$, $f(u_t)$ can be written as an ARIMA model.

$$f(u_t) = \varphi^{-1}(L)\nabla^{-d}\theta(L)u_t \qquad [2a]$$

Although the $\varphi(L)$ and $\theta(L)$ structures of [2a] were available in the 1920s (e.g., Slutsky, 1937; Yule, 1927), the more general Box–Jenkins–Tiao model was developed in the 1970s. Practical applications required computing resources, of course. Box and Jenkins (1970) described a modeling algorithm which was implemented shortly thereafter. Modern time series software packages included in virtually all-commercial statistical analysis systems evolved from this original modeling algorithm.

If the parameters of the autoregressive and moving average polynomials $\varphi(L)$ and $\theta(L)$ are constrained to the stationary-invertible segment [2a] is solved for the innovation as

$$u_t = \theta^{-1}(L)\varphi(L)\nabla^{-d}Y_t \qquad [2b]$$

If the innovation series estimated from [2b] has the properties [1a], the ARIMA model renders all of the Type I threats to statistical conclusion validity implausible.

The Type II errors occur when H_0 is false (i.e., when $\omega \neq 0$), but H_0 is mistakenly *not* rejected (or even *accepted*). Given a fixed a priori α-error rate, the Type II error rate is optimized by design to be less than a conventional threshold level. For the conventional $\alpha = 0.05$ Type I error rate, the Type II error rate is usually optimized at $\beta < 0.2$ corresponding to statistical power of at least 0.8 (Kendall and Stuart, 1979, Chapter 22).

Compared to the Type I threats, the Type II threats to statistical conclusion validity have received little attention. There are

two reasons for this. First, the Type II threats assume some variation of the Neyman–Pearson hypothesis test (Neyman and Pearson, 1933). In variations of the traditional Fisher significance test (Fisher, 1925, 1955), the Type II threats are undefined. But for the second, assuming a Neyman–Pearson test, most of the Type II threats are related to the length of the time series. Because estimation algorithms rely on asymptotic properties, analyses of short time series may fall below the conventional 0.8 statistical power threshold. Given a long time series, on the other hand, say $N \geq 50$ observations, in practice, the ARIMA model controls both Type I and Type II threats to statistical conclusion validity.

Otherwise, not withstanding the magnitude of the scalar parameter ω – small effects are difficult to detect in even the best circumstances – the statistical power of a time series quasi-experiment is determined by series length, balance, and model adequacy. For constant N and α, the statistical power of a time series quasi-experiment is maximized when $N_{pre} = N_{post}$. Since statistical power is proportional to the absolute fit of the ARIMA model, model adequacy does guarantee that β will fall below the conventional threshold. Nevertheless, spurious inferences attributed to Type II threats to statistical conclusion validity appear rare.

Internal Validity

Although in principle, all nine of the common threats to internal validity identified by Shadish et al. (2002, p. 55) are plausible in the time series quasi-experiment, four of the nine – history, maturation, instrumentation, and regression – account for virtually all of the spurious inferences found in the applied literature. The threats of history and maturation are often made plausible by the relatively long time frame of this design and to the nature of long time series. The threats of instrumentation and regression are often made plausible by the nature of planned interventions. For unplanned interventions, a special case that Campbell (1969) calls the 'natural experiment,' neither threat is necessarily plausible.

When available, a suitable control time series renders the threat of history implausible. Hennigan et al. (1982) uses the time series in **Figure 1** to document an increase in property crime following the introduction of commercial television. In 34 'early' and 34 'late' cities, television broadcasting begins in 1950 and 1954, respectively. Considered separately, history is a plausible threat to the internal validity of both effects. The 'late' effect may be due to an economic recession that began in 1955, for example, and the 'early' effect may be an artifact of the Korean War mobilization. As men were called away to military service and women joined the workforce, homes were left undefended against property criminals. Although history is a plausible threat to the internal validity of each effect, it is implausible for the joint effect of both interventions.

The other three common threats to internal validity are less robust to standard design features. **Figure 2** shows monthly burglaries for Tucson before, during, and after an intervention (McCleary et al., 1982). Prior to 1979, burglaries were investigated by uniformed police officers. When this task was taken over by detectives, burglaries dropped abruptly. Two years later, when the task was transferred back to uniformed officers,

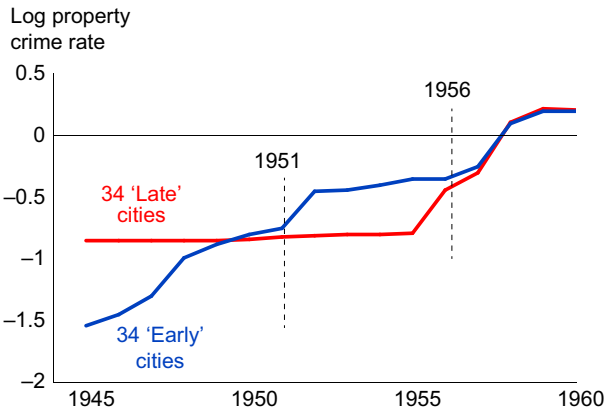

Figure 1 Annual property crimes for cities with commercial TV broadcast service in 1951 and for cities with TV broadcast service in 1956. Source: Hennigan, K., Del Rosario, M.L., Heath, L., Cook, T.D., Wharton, J.D., Colder, B.J., 1982. Impact of the introduction of television on crime in the United States: empirical findings and theoretical implications. Journal of Personality and Social Psychology 42 (3), 461–477.

burglaries returned abruptly to pre-1979 levels. While the on–off component of this design rules out the threat of history, the other threats are still plausible. In fact, this effect is entirely explained by differences in the way detectives and uniformed officers keep records; it is an instrumentation artifact. Because interventions often affect the way in which a target phenomenon is measured, instrumentation is always a plausible threat to the evaluation of planned interventions.

Regression threats are also associated with planned interventions but for different reasons. Following a 1955 traffic 'crackdown' in Connecticut, highway fatalities dropped significantly. Examining a long time series, however, Campbell and Ross (1968) show that the reduction was a regression artifact. The 'crackdown' was implemented in 1955 because traffic fatalities in that year were abnormally high; regression to the mean was virtually guaranteed.

The threat of regression is plausible in the Connecticut example because the hypothesis test relied on only $N_{post} = 1$

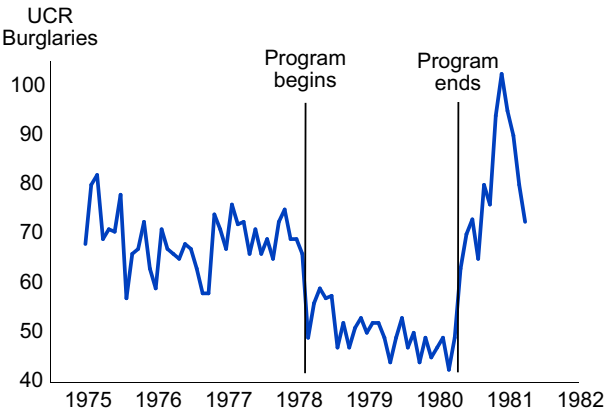

Figure 2 Monthly burglaries for Tucson, 1975–81. Source: McCleary, R., Nienstedt, B.C., Erven, J.M., 1982. Uniform crime reports and organizational outcomes: three time series quasi-experiments. Social Problems 29, 361–372.

annual observation. Given enough post-intervention observations to model the temporary disturbance, the regression threat is implausible. The Connecticut example aside, no spurious inferences attributed to this threat are found in the applied literature.

The time series quasi-experiment is distinguished from other common quasi-experimental designs by the use of a statistical model to control common threats to internal validity, especially the most plausible threat. Maturation as a threat to internal validity is plausible whenever a time series trends deterministically or drifts stochastically. Because most of the long social science time series exhibit trend and/or drift, maturation is a plausible threat to the internal validity of most time series quasi-experiments.

Figure 3 shows a weekly productivity series from the Hawthorne experiment. From 1927 to 1932, researchers manipulated daily rest breaks and, though not conducting a formal analysis, claimed that rest breaks affected a large, positive increase in productivity. However, formal analyses of these data challenge the original finding. Regressing step functions coded for three 'punishments' regimes, Franke and Kaal (1978) demonstrate that the productivity increases were motivated by fear. The fundamental problem with this interpretation, as well as with the original interpretation, is that the productivity time series exhibits a systematic trend throughout its history. To the extent that the bank wiring room machine operators were becoming more productive over time, maturation is a plausible threat to the internal validity of both interpretations.

To illustrate the problem, the punishment effects estimated by Franke and Kaule are plotted against the productivity series in Figure 3. The effects are collinear with the natural trend of this series. To the extent that a difference equation accounts for this trend, the ARIMA model [2a] renders this particular maturation threat implausible. Reanalysis of this time series quasi-experiment with a difference equation finds a modest but significant effect for rest breaks and no effect for punishment (McCleary, 2000). In general, maturation threats to internal validity are implausible when Box–Jenkins–Tiao models are used to estimate intervention effects.

Construct Validity

Of the 14 common threats to construct validity identified by Shadish et al. (2002, 73), the threats associated with placebo artifacts – experimenter expectancies and novelty and disruption – are plausible in many time series quasi-experiment applications. In true experiments, these threats are controlled by blinding or masking the intervention to subjects and experimenters. In a time series quasi-experiment, the threats are controlled by modeling the effect in a dynamic transfer model for the $g(X_t)$ component of [2]. One popular version of $g(X_t)$ uses a single lagged operator to distribute the effect over the post-intervention segment.

$$g(X_t) = \delta(L)\omega\nabla^d X_t \qquad [2c]$$

If the parameter δ is constrained to $0 \leq \delta \leq 1$, and if d is either zero or one, [2c] includes a subset of models with meaningful and useful interpretations (McCleary and Hay, 1980; McDowall et al., 1980).

Figure 4 shows 50 daily self-injurious behavior counts for an institutionalized patient (McCleary et al., 1999). When the patient is treated with an opiate-blocker on the 26th day, the incidence of self-injurious behavior drops abruptly and stays at the new, lower level for the remaining 25 days of observation. This effect is modeled by setting $\delta = \nabla = 0$ in [2c]. Figure 5 shows the daily counts for a patient who received a placebo on the 26th day instead of the opiate-blocker. This patient's self-injurious behavior also drops abruptly but then returns to the pre-intervention level within a few days. This effect is modeled by setting $\nabla = 1$ in [2c] and estimating a nonzero value of δ.

The visually apparent decaying 'spike' placebo effect in Figure 5 is commonly found in social science applications.

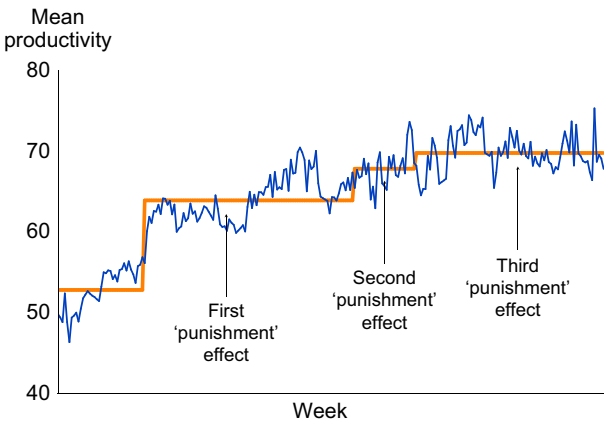

Figure 3 Weekly productivity measures for five machine operators in the first bank wiring room (Hawthorne) experiment. Estimated step function effects are plotted against the series. Source: Franke, R.H., Kaul, J.D., 1978. The Hawthorne experiments: first statistical interpretation. American Sociology Review 43, 623–643.

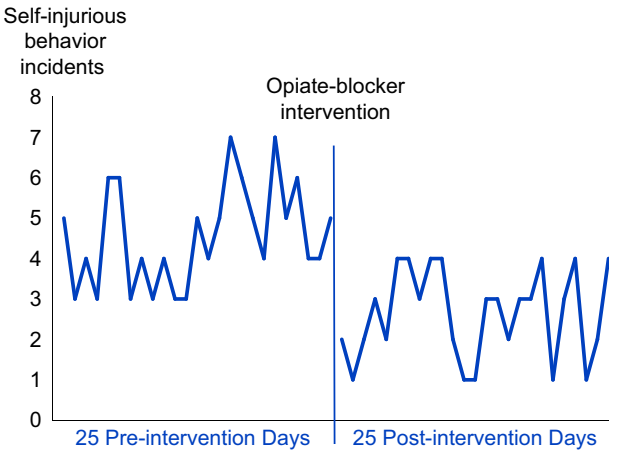

Figure 4 Daily self-injurious behavior counts for a single patient before and after an opiate-blocker. Source: McCleary, R., Touchette, P., Taylor, D.V., Barron, J.L., 1999. Contagious models for self-injurious behavior. In: Poster Presentation, 32nd Annual Gatlinburg Conference on Research and Theory in Mental Retardation.

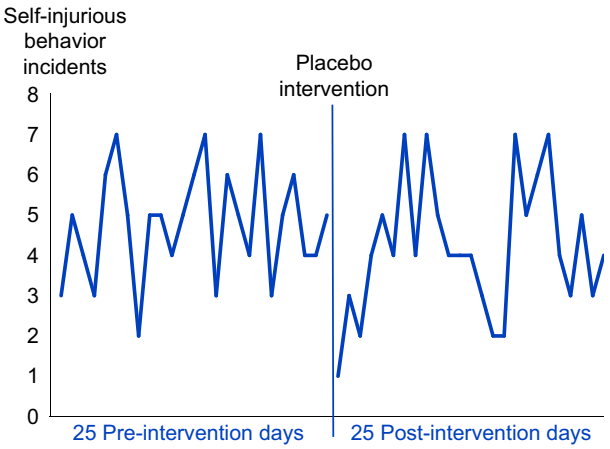

Figure 5 Daily self-injurious behavior counts for a single patient before and after a placebo. Source: McCleary, R., Touchette, P., Taylor, D.V., Barron, J.L., 1999. Contagious models for self-injurious behavior. In: Poster Presentation, 32nd Annual Gatlinburg Conference on Research and Theory in Mental Retardation.

For example, McCleary and Riggs (1982) demonstrate that new laws, whether ultimately successful or not, often produce 'decaying' spikes that have much in common with placebo effects. The ability to model these temporary effects illustrates a salient strength of the time series quasi-experiment. Other common quasi-experimental designs have a difficult time accommodating dynamic effects that are not static and permanent.

External Validity

When Campbell (1957) initially proposed a dichotomous validity system, the common threats to external validity were defined as any that were not obvious threats to internal validity. When Cook and Campbell (1979) expanded the dichotomy to the current four-validity system, the threats to external validity were again defined as the residual set left over after the threats to construct validity had been defined. Shadish et al. (2002, p. 87) take the same position, noting that external validity "concerns inferences about the extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes." These concerns are not always well described in terms of discrete threats, however, especially as they relate to the external validity of time series quasi-experiments.

As software became more widely available after 1970, applications of the time series quasi-experiment began to accumulate. In evaluation research, effect estimates of replicated social programs often-supported discrepant substantive inferences about the program's effectiveness. Literature evaluating the effectiveness of gun-control legislation is typical in that the published estimates include positive, negative, and null effect estimates (Reiss and Roth, 1993, pp. 255–287). Whereas a small proportion of randomly distributed null effects can be ignored, a broader range of quasi-experimental effect estimates suggests that the inter-replication variance is more important than the general effect itself.

The external validity of several time series quasi-experiments can be analyzed in either of two ways. First, a set of k time series can be defined as vector time series, Y_t

$$Y_t' = \{Y_{1,t}, Y_{2,t}, \ldots, Y_{k,t}\} \quad [3]$$

The vector Y_t' can then be analyzed as a vector ARIMA process (e.g., McGaw and McCleary, 1985). Second, variance in the set of k reported effect estimates

$$\omega' = \{\omega_1, \omega_2, \ldots, \omega_k\} \quad [3a]$$

can be decomposed into components associated with variance in 'persons, settings, treatments, and outcomes.' Since the second approach is a special, restricted case of fitting the set of k related quasi-experiments into a vector ARIMA model; the first approach is preferred. In practice, unfortunately, few time series satisfy the rigid assumptions required for a vector ARIMA model.

Although less general, the second approach is more practical and flexible in most situations. The statistical models used to analyze the set of k related effect estimates, which were originally developed in meta-analysis or research synthesis, decompose the effect variance into components that can be attributed to setting (e.g., place, historical period, and so on), properties of the intervention (e.g., resources devoted to the intervention, mode of implementation, and so on), methodology (e.g., time series properties, dependencies, and so on), and other potential threats to external validity. McDowall et al. (1980) demonstrate this method for combining effect estimates from evaluations of gun control laws; McCleary (2000) uses a similar model for combining effect estimates from the Hawthorne experiment plotted in **Figure 3**.

Future Directions

At this point in its development, solutions to the questions of internal validity have been proposed, tested, and accepted. Solutions to the questions of construct validity for the time series quasi-experiment are resolved by the general Box–Jenkins–Tiao model [2]. Although solutions to the questions of external validity are still too novel to enjoy the same level of consensus, it seems that the currently available solutions will become more widely used and accepted in the near future. All of the unresolved issues relate to statistical conclusion validity. Problems of missing data and short time series (e.g., $N < 30$ observations) pose particular needs.

Otherwise, the consensus view of hypothesis testing is undergoing rapid change in the social sciences. All of the common threats to statistical conclusion validity identified by Shadish et al. (2002, p. 45) are artifacts of what Gigerenzer et al. (1989, p. 106) call the 'hybrid' null hypothesis significance test. As conventions and methods in this area change, fundamental thinking about the threats to statistical conclusion validity is likely to change as well.

See also: External Validity; Internal Validity; Quasi-Experimental Designs; Time Series: ARIMA Methods; Time Series: Advanced Methods; Time Series: General.

Bibliography

Box, G.E.P., Jenkins, G.M., 1970. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco, CA.

Box, G.E.P., Tiao, G.C., 1975. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association 70, 70–79.

Campbell, D.T., 1957. Factors relevant to the validity of experiments in social settings. Psychological Bulletin 54, 297–312.

Campbell, D.T., 1963. From description to experimentation: interpreting trends as quasi-experiments. In: Harris, C.W. (Ed.), Problems in Measuring Change. University of Wisconsin Press, Madison, WI.

Campbell, D.T., 1969. Reforms as experiments. American Psychologist 24, 409–429.

Campbell, D.T., Ross, H.L., 1968. The Connecticut crackdown on speeding: time series data in quasi-experimental analysis. Law and Society Review 3, 33–53.

Campbell, D.T., Stanley, J.C., 1966. Experimental and Quasi-experimental Designs for Research. Rand-McNally, Skokie, IL.

Cook, T.D., Campbell, D.T., 1979. Quasi-experimentation: Design and Analysis Issues for Field Settings. Houghton-Mifflin, Boston, MA.

Fisher, R.A., 1925. Statistical Methods for Research Workers. Oliver and Boyd, Edinborough.

Fisher, R.A., 1955. Statistical methods and scientific induction. Journal of the Royal Statistical Society. Series B 17, 69–78.

Franke, R.H., Kaal, J.D., 1978. The Hawthorne experiments: first statistical interpretation. American Sociology Review 43, 623–643.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Kruger, L., 1989. The Empire of Chance: How Probability Changed Science and Everyday Life. Cambridge University Press.

Hennigan, K., Del Rosario, M.L., Heath, L., Cook, T.D., Wharton, J.D., Colder, B.J., 1982. Impact of the introduction of television on crime in the United States: empirical findings and theoretical implications. Journal of Personality and Social Psychology 42 (3), 461–477.

Kendall, M., Stuart, A., 1979. The Advanced Theory of Statistics, fourth ed., vol. 2. Charles Griffin, London.

McCleary, R., 2000. Evolution of the time series experiment. In: Bickman, L. (Ed.), Research Design: Donald Campbell's Legacy. Sage, Thousand Oaks, CA, pp. 215–234.

McCleary, R., Hay Jr., R.A., 1980. Applied Time Series Analysis for the Social Sciences. Sage, Beverly Hills, CA.

McCleary, R., Riggs, J.E., 1982. The 1975 Australian Family Law Act: a model for assessing legal impacts. New Directions for Program Evaluation 1982 (16), 7–18.

McCleary, R., Nienstedt, B.C., Erven, J.M., 1982. Uniform crime reports and organizational outcomes: three time series quasi-experiments. Social Problems 29, 361–372.

McCleary, R., Touchette, P., Taylor, D.V., Barron, J.L., 1999. Contagious models for self-injurious behavior. In: Poster Presentation, 32nd Annual Gatlinburg Conference on Research and Theory in Mental Retardation.

McDowall, D., Loftin, C., Wiersema, B., 1992. A comparative study of the preventive effects of mandatory sentencing laws for gun crimes. Journal of Criminal Law and Criminology 83, 378–394.

McDowall, D., McCleary, R., Hay Jr., R.A., Medinger, E.E., 1980. Interrupted Time Series Analysis. In: University Papers Series: Quantitative Applications in the Social Sciences, vol. 21. Sage, Beverly Hills, CA, and London.

McGaw, D.B., McCleary, R., 1985. PAC spending, electioneering, and lobbying: a vector ARIMA time series analysis. Polity 17, 574–585.

McSweeney, A.J., 1978. The effects of response cost on the behavior of a million persons: charging for directory assistance in Cincinnati. Journal of Applied Behavioral Analysis 11, 47–51.

Neyman, J., Pearson, E.S., 1933. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 231, 289–337.

Richardson, J., Kroeber, A.L., 1940. Three centuries of women's dress fashions: a quantitative analysis. Anthropological Records 5 (2), 111–154.

Reiss, A.J., Roth, J.A., 1993. Understanding and Preventing Violence. National Academy Press, Washington, DC.

Ross, H.L., Campbell, D.T., Glass, G.V., 1970. Determining the effects of a legal reform: the British 'breathalyzer' crackdown of 1967. American Behavioral Scientist 13, 493–509.

Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and Quasi-experimental Designs for Generalized Causal Inference. Houghton-Mifflin, Boston.

Shlaifer, R., 1980. The relay assembly test room: an alternative statistical interpretation (Comment on Franke and Kaul, ASR. October 1978). American Sociological Review 45, 995–1005.

Slutsky, E., 1937. The summation of random causes as the source of cyclic processes. Problems of Economic Conditions 3, 1.

Webb, E.J., Campbell, D.T., Schwartz, R.D., Sechrest, L., 1966. Unobtrusive Measures. Rand-McNally, Skokie, IL.

Yule, G.U., 1927. On a method of investigating periodicity in disturbed series with special reference to Wolfer's sunspot numbers. Philosophical Transactions A226, 267.

Intersectionality

Stephanie G Guittar and Nicholas A Guittar, Valdosta State University, Valdosta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Intersectionality is a methodology and theoretical framework that is designed around drawing scholarly attention to variations in people's lived experiences. Our experiences are driven by our social identification with different dimensions, such as race, class, gender, and sexuality among others, and intersectionality is aimed at making sense of these differences. Intersectionality was born out of a critique of feminist theories as lacking inclusivity, and it has morphed into a variety of approaches designed to investigate previously understudied elements of identity and their connection to disparate social outcomes.

Intersectionality refers to the study of intersecting social categories – such as race, gender, and social class – with which an individual identifies. Intersectionality uniquely addresses a holistic understanding of the lived experiences of an individual within a society. It is the recognition that social outcomes cannot be properly explained by investigating independent social categories and treating them as stand-alone variables. Women are not a monolithic category, and neither are African-Americans. People are affected simultaneously by the social treatment of a multitude of characteristics. Consider the intersection of race and gender – the most storied interaction detailed in early intersectionality literature. This single intersection creates categories of people whose experiences are oftentimes vastly different from groups who share only one dimension (e.g., African-American women and White women) or perhaps none of these dimensions (e.g., African-American women and White men).

The multiple dimensions of our social identities are not easily teased apart. It is essential then to emphasize the synergy of two or more dimensions in order to understand their compound effects on individuals, groups, even social institutions. It is difficult to determine which social category has the most influence over our interactions and experiences because each category is intimately tied to the next – our social demographics operate simultaneously in everyday interactions. We are all the sum of our parts; thus, according to intersectionality, it does not make sense to study each part separately as it would not provide an accurate depiction of our social realities. Studying race, class, and gender as intersecting categories allows us to understand how each of these categories manifests itself within a given social context and in relation to each other.

Intersectionality, as a theoretical framework and/or methodology has been positioned as particularly useful for understanding and explaining the experiences of social categories of people who experience compound disadvantage (e.g., African-American women). Historically, these two categories – race and gender – have been treated as being exclusive of one another. Thus, research was commonly focused on 'women' or 'African-Americans' but rarely on 'African-American women'. Add a third dimension into the mix, such as social class, as you begin to see how complicated it can be to investigate social phenomena without trivializing the experiences of any particular populations. Herein lies the power of intersectionality.

> Fundamentally, race, class and gender are intersecting categories of experience that affect all aspects of human life; thus, they simultaneously structure the experiences of all people in this society. At any moment, race, class or gender may feel more salient or meaningful in a given person's life, but they are overlapping and cumulative in their effects.
>
> (Andersen and Collins, 2012: 4)

Intersectionality was largely born out of a critique of feminist theories as being ill fit to depict or explain the experiences of African-American women. As a theory, intersectionality has continued to be employed most often to investigate the intersection of categories, which result in compound disadvantage; thus enabling researchers to better investigate underserved or marginalized populations. But intersectionality is also being increasingly used to investigate categories of people who experience compound privilege (e.g., White men), as well as categories of people who experience both oppression and privilege on differing dimensions (e.g., gay men).

Intersectional approaches to studying social life are necessary since individuals and groups do not have equal access to scarce resources in contemporary societies. The stratification of society among multiple dimensions has created a social reality where people have such different experiences based upon their ascribed and achieved statuses. At the center of determining our access to valued resources are the polarizing concepts of privilege and oppression. On any social dimension there are groups who enjoy privilege (i.e., unearned advantage) and many other groups who find themselves being pressed or formed into 'lesser' positions in society (see Marilyn Frye's, 1983 essay 'Oppression' for further insight defining oppression).

Without a focus on intersectionality, we would be left with the false sense that all members of a particular category of people (all women, all Hispanics, etc.) are affected equally by various social forces. This is particularly troublesome considering how, in the contemporary United States, members of privileged categories (Whites, men, heterosexuals, people with class privilege, etc.) oftentimes fail to recognize how their privileged position improves their social outcomes. It is as if their privilege is invisible, and this invisibility threatens the ability of privileged groups to recognize how oppressed groups are adversely affected. Conversely, members of oppressed groups firmly recognize and understand that the characteristics by which they are oppressed are intimately tied to their social

outcomes. This disconnect – between those who are privileged and those who are oppressed – can lead privileged people to erroneously believe that they have similar lived experiences to members of disadvantaged groups. This reality was characterized very clearly by Michael Kimmel and Ferber (2010: xiv–xv):

> I first confronted this invisibility in the early 1980s, when I participated in a small discussion group on feminism. In one meeting, a white woman and a black woman were discussing whether all women were, by definition, "sisters" because they all had essentially the same experiences and because all women faced a common oppression by men. The white woman asserted that the fact that they were both women bonded them, in spite of racial differences. The black woman disagreed.

> "When you wake up in the morning and look in the mirror, what do you see?" she asked. "I see a woman" replied the white woman. "That's precisely the problem," responded the black woman. "I see a *black* woman. To me race is visible every day, because race is how I am *not* privileged in our culture. Race is invisible to you because it's how you are privileged. It's why there will always be differences in our experience."

This single interaction between two women characterizes the importance of intersectionality. Two individuals who share a single dimension (in this case, gender) do not experience a unified reality. Their experiences are fractured along the lines of countless other dimensions, most immediately of which is race. Thus, it becomes essential that researchers consider the intersectional nature of our identities.

Hegemonic ideologies further complicate things by erroneously assuming that sex, gender, sexuality, and race are binary terms. One is male or female, a man or a woman, gay or straight, White or Black. Intersectionality challenges this ideology and highlights the variance within each category. Without careful attention to the diversity found within each category researchers will (inadvertently or not) continue to reify false dichotomies. This practice has contributed to the invisibility of intersexuality, transgender, and bisexuality (Lucal, 2008), among other social categories. In fact, many social scientists emphasize that there is greater diversity within a single category of people (e.g., women) than there is between the 'average, typical' members to two different categories (e.g., men and women). Considering this within-group diversity, it is plain to see that there is not just one monolithic lived experience for any social category. Intersectionality allows researchers to better investigate and learn about the unique experience of people in less privileged positions. Before taking a look at contemporary examples of intersectionality 'in action', it is important to delve deeper into the theoretical roots and early formation of intersectionality.

Background and Early Roots

Intersectional analysis stems from a phenomenological and feminist perspective. By employing phenomenological methodology the researcher seeks to understand the lived experiences of individuals through the subjective reality of their experiences in the world. The basic proposition of phenomenology is that reality is socially constructed; that all of our ideas, norms, and beliefs are taken for granted by members of society and hence, not questioned. Phenomenology, as pioneered by Alfred Schutz (1967) is rooted heavily in Weber's concept of verstehen. Verstehen (which is German for understanding) stresses that you must put yourself in another person's shoes if you wish to truly understand their social reality. Phenomenology, then, translates perfectly to intersectionality: in order to understand a person's subjective reality one must simultaneously take into consideration all elements of their social identity (race, class, gender, and the like).

Phenomenological researchers challenge our culturally learned ideas and urge us to question the taken-for-granted beliefs and behaviors that we observe. This last element of phenomenology has obvious methodological implications, and it has become a cornerstone of intersectionality. Consider, for example, that the average person in the United States views sexuality as a 'real' thing – something that can be identified and categorized. Thus, sexuality researchers who are not mindful of their assumptions could enter into a study believing that everyone identifies with a concrete sexual identity. Sexual identities seem like commonplace units for categorizing sexuality, but only recently have sexual identities been used as such. Prior to the 1940s, sexuality was framed almost exclusively as mere behavior and nothing more. Our current practice of categorizing sexual identities emerged in the mid-1940s as the word 'heterosexual' began being used as a way to identify people who are attracted to members of the other sex (Katz, 2007). Fast forward to the present day, and we find that many teens and young adults are bucking these categories, opting out of sexual politics, and stating that they simply like who they like (Yon-Leau and Munoz-Laboy, 2010). Shedding assumptions and taken-for-granted beliefs enables researchers to focus on people's lived experiences and uncover these sorts of revolutionary findings. Phenomenology then encourages the researcher to take the role of a stranger and view the world through someone else's standpoint to understand their subjective reality. This factor positions phenomenology as particularly helpful when researching intersections that do not align with one's own categorical characteristics.

Feminist standpoint theory elaborated on phenomenology by bringing the experiences of women to the forefront. Dorothy Smith (1979) developed feminist standpoint theory, which emphasizes not only that women's perspectives are unique, but also that women live their lives according to a bifurcated consciousness. This concept, similar in form to W.E.B. Du Bois' double consciousness (Du Bois and Edwards, 2007), involves the duality of functioning within a society in which one is viewed as the 'other', while also maintaining one's own lived experience. DuBois' double consciousness is centered on the reality that African-Americans live and operate in two Americas – one dictated by dominant White middle-class ideologies, and another, which is the more authentic experience of African-Americans. Smith similarly emphasized that women's lived experiences are produced along a 'line of fault' within their personal, lived, and reflected-on experience. For women the 'line of fault' is the juncture between what they know, their reality, and the external expectations and culture that exists around them (the patriarchal rules). Hence,

standpoint theory emerged as a way to bring to light the unique experiences of women to a world that had focused purely on men's experiences.

Although the aforementioned works by Smith and DuBois predate the contemporary conception of intersectionality, their focus on the unique experiences of historically underrepresented and consequently oppressed groups is clearly a root of intersectionality. By looking macroscopically and collectively at both scholars' works, we begin to see not only the unique standpoints of women and African-Americans, but the unique experiences of intersectional categories, such as African-American women, African-American men, and White women – all of whom have lived experiences markedly different from the traditionally privileged category of White men. Standpoint theory is not without its critics. The principal criticism related to intersectionality is that standpoint theory is not a holistic approach because it utilizes gender as the main standpoint for women's experience. Intersectional feminists emphasize that race, class, sexual orientation, nationality, and other social categories create different 'standpoints' for women – none of which should be predetermined as more essential than the next. Present-day intersectionality is viewed as a much more holistic and inclusive approach to studying the experiences of individuals in less privileged positions.

The Advent of Intersectionality

Historically, social inquiry has recognized some forms of stratification (namely, gender and race), although the discussion of intersecting layers of stratification did not widely enter the social sciences until the late 1980s and early 1990s. Intersectionality gained recognition as a stand-alone theory and methodology in the late 1980s. Kimberlé Crenshaw (1989) is credited with coining the term 'intersectionality', but it is important to note that many women before her had written about how women of color were marginalized within feminist discourse (Davis, 2008). As early as the mid-nineteenth century, women like Sojourner Truth wrote about the omission of Black women in civil rights movements. Feminist theorists, such as Patricia Hill Collins, Barbara Smith, bell hooks, Lynn Weber, and Audre Lorde, among others have contributed to the development of intersectionality to include the unique experiences of minority women among feminist thought.

Intersectionality arose from a criticism of feminist theories for omitting women of color in feminist research or including "hand picked Black women's voices to avoid criticisms that they are racist" (Collins, 2000: 6). Intersectional theorists argued that feminist perspectives inaccurately depicted the experiences of *all* women as a collective, shared experience. While feminist theorists brought awareness to gender inequality, they typically overlooked other forms of oppression (e.g., racism, classism) that are equally central to the experiences of minority women. In a similar vein, minority women were historically left out of academic discussions of race. Thus women of color largely existed on the fringes of research on two of the central most elements of their identity (i.e., gender and race). The experiences of women of color were insufficiently represented by feminist and racial discourses, and intersectionality emerged as a way to counter this relative invisibility. A 1993 article by Bernice Barnett serves as an excellent example of how early intersectional research worked to mend minority women's invisibility on both fronts. Her work details the invisibility of Black women's contributions to the civil rights movement, and it showcases the power of employing an intersectional methodology in order to highlight the unique experiences of a particular category of people.

Intersectionality gained steam throughout the 1990s as a key perspective for unlocking the interaction between social categories, such as gender, race, and social class (along with other dimensions) that permeate people's lives. More specifically, intersectionality began to emphasize "the outcomes of these interactions in terms of power" (Davis, 2008). Power, in a Weberian sense, is the probability that one actor within a social relationship will be in a position to carry out his or her own will despite resistance. Through intersectional analyses, one is able to see how social categories work simultaneously in conjunction with one another to produce a compound influence that can limit or enhance an individual's ability to carry out his or her own will. Thus, intersectionality is essential to understanding how power operates in a multicultural society. Grouping all women into one category, as prior feminist writings had done, is problematic. It does not enable researchers to conceptualize the power struggles that adversely affect women, racial minorities, sexual minorities, and the underclass, among others.

Patricia Hill Collins is largely credited as the chief expert on intersectionality – beginning with her seminal work *Black Feminist Thought* (2000). Although Collins – whose work garners much more attention below – wrote most heavily about intersectionality as it relates to Black women, intersectionality research has expanded far beyond the dimensions of race and gender. Intersectionality is clearly applicable to every person or group in society. A poor, White, gay woman may be privileged because of her race but she simultaneously experiences oppression as a result of her social class, sexuality, and gender. Similarly, a Black man may be privileged because of his gender but he experiences disadvantages due to his race. Therefore, using single categories to describe an individual's experience neglects all possible forms of oppression or privilege that one can experience. Intersectionality can even be employed to critique 'majority' groups and emphasize different dimensions of privilege.

Some scholars speak of intersectionality as a theory, others as a concept, and others as an analytical approach (Cole, 2008; Davis, 2008). As a theory, it is vague in that there is not a clear-cut methodology to properly test it. There is not a consensus on the social categories that should be analyzed (beyond race, gender, and social class) or where the line should or could be drawn. On the other hand, its vagueness and the vast number of categories that can be intersected reconcile the gaps between critical feminist theory and postmodern feminist theory. Intersectionality challenges the complexity, the distinctiveness, and the entirety of the wide range of social categories. No social category is binary, and intersectionality as an analytical tool allows researchers to study the effects of these fluid and complex categories on all lived experiences. Such is a major reason why intersectionality has become so central to contemporary applications of queer theory, feminist theory, and critical race theory.

The Matrix of Domination

> Oppression and resistance remain intricately linked such that the shape of one influences that of the other. At the same time, this relationship is far more complex than a simple model of permanent oppressors and perpetual victims. Another way of approaching power views it not as something that groups possess, but as an intangible entity that circulates within a particular matrix of domination and to which individuals stand in varying relationships.
> (Collins, 2000: 274)

Collins conceptualizes intersectionality as a matrix of domination. She uses her intersectional framework to illustrate the ways in which each demographic category intersects with structures of power and privilege to form unique lived experiences for various group. In the matrix of domination, there are multiple interlocking levels of privilege and oppression that stem from our stratified positions within various elements of US society. Collins initially focused almost exclusively on the empowerment of African-American women through an understanding of the multiple oppressive forces of gender and race – which work together to undermine individual and collective effort. She has since broadened her theoretical use of the matrix of domination to encompass all groups formed by the intersections of race, class, gender, nationality, and sexuality. One is not simply oppressed or an oppressor; privileged or disadvantaged. Oppression and privilege do not play out in simple binary terms (of either having it or not), they are instead forces to which all individuals have differing relations. The social structuring of race, class, gender, and sexuality "affects individual consciousness, group interaction and group access to institutional power and privileges" (Andersen and Collins, 2012: 5). Through the perspective of a matrix of domination, one social category is not more important than another. One cannot be more important than another at all times. As mentioned earlier, one can be more salient than others at one time but it is the combination and the cumulative effect of all of our privileges and disadvantages that shape our experiences. We are never viewed as just one 'something'. A woman is never viewed as *just* a woman; there are other social categories, such as her race, class, sexuality, nationality, education that will affect interactions, and her access to power and privilege.

Intersectionality: Different than Studying Difference, Diversity or Multiculturalism

Patricia Hill Collins makes a great distinction between intersectionality and the study of difference, diversity or multiculturalism. At first, intersectionality may seem intuitive – as in 'how is this a new perspective?' Have we not researched race, class, and gender for decades? While race, class, and gender have been used as 'variables' in quantitative studies, and plenty of feminists have studied the experiences of women, intersectionality is not as simple as it sounds. To study the intersections of race, gender, and class, one must surpass the simple task of comparing experiences of difference. When individuals think of the experiences of those who are not like them, they may simply engage in a comparative analysis. For example, a White male listening to the experiences of a Latino male may think to compare the experiences of Latino men to other Latino men from the Caribbean, Central, or South America. In doing so, the White male has grouped the Latino male experience as one based strictly on his ethnicity and assumed it to be comparable to others of the same ethnicity. Engaging in this type of comparison overlooks how an individual's experience can be shaped by nationality, racial variations within an ethnic group, and social class (among other categories) that represent a position in a system of power, where a person can be privileged in one category and oppressed by others. Though learning about different groups is a step in the right direction toward enhancing knowledge of others' experiences, it is important to recognize that there is more to the experience than simply the influence of one social category; in this case, ethnicity.

Comparative analysis can also lead one to compare how one group's oppression is similar to other oppressed groups. For example, some have compared the marriage equality movement to the civil rights movement. While racial minorities and sexual minorities have both experienced oppression in our society, comparing their oppression is not an appropriate approach because it overlooks the various power mechanisms at play for each member of these groups. Not to mention the civil rights movement of the 1960s was not focused solely on race, but on equality as a whole. Comparing groups also lends itself to setting up a hierarchy where one group's oppression is ranked against the others as if being the most oppressed somehow earns more attention or empathy. It can also lead to a comparison to the dominant group, the 'norm' that automatically classifies those who differ as 'abnormal' and creates stigma. Engaging in comparative analysis of the experiences of oppressed groups does not accurately describe or analyze the experiences in terms of the interlocking mechanisms of privilege and oppression.

Collins also points out that intersectionality is different from studying diversity. Diversity has become a popular word in our society, and increasing diversity and inclusion has become a primary goal for many institutions. Diversity initiatives aim to increase awareness about different cultures, races, and ethnicities. But in doing so the underlying message is that, while we may look different, beneath it all we are all the same. Thus pure diversity-based approaches risk completely overlooking the inequalities that various minority groups experience as a result of their position in society. When the idea of 'beneath it all, we are all the same' is promoted, race, class, and gender are understood as strictly aesthetic differences between us rather than vehicles of opportunities and constraints within the structures of power in our society. Collins states:

> The very term diversity implies that understanding race, class, and gender is simply a matter of recognizing plurality of views and experiences in society—as if race, class, and gender were benign categories that foster diverse experiences instead of systems of power that produce social inequalities.
> (Andersen and Collins, 2012: 8)

Intersectionality analyzes experiences based on the interconnection of ethnicity, race, class, gender, nationality, religion, sexuality, and any other social categories that situate one's experience in relation to power in society. This focus on the differential influences of power and privilege set

intersectionality apart from studies of diversity and difference. From Collins' perspective, the matrix of domination is the most accurate approach to analyzing intersectionality "because of its focus on structural systems of power and inequality" (Andersen and Collins, 2012: 9).

Intersectional Methods and Analysis

By now, the importance of researching intersectionality should be obvious. While many scholars agree that intersectional research is needed, few scholars have outlined a clear methodology and analytical strategy to conduct intersectional research. Such is a major criticism of intersectionality. There seem to be a number of common themes for intersectionality scholarship that make it difficult to narrow down exact methods to conduct intersectional research. In this section, some of the approaches that have been discussed in the literature are summarized. McCall (2005) outlines three approaches to studying intersectionality: anticategorical, intracategorical, and intercategorical (although she emphasizes that not all intersectional research fits nicely into one of these approaches). These are not new ways of studying intersectionality but place the current intersectional methodologies into three categories.

The anticategorical complexity approach, as the name suggests, rejects the labeling of any particular category. The meaning of any social category, such as race, gender, or sexuality is highly contextual and socially constructed. Weber (1998: 16) states, "Although they [race, class, gender, and sexuality] persist throughout history, race, class, gender and sexuality hierarchies are never static and fixed, but constantly undergo change as part of new economic, political, and ideological processes, trends, and events." Since these categories are never static, any category can be challenged and eventually become insufficient. For example, it is no longer sufficient to categorize individuals as 'gay' or 'straight'. There are numerous sexual identities, some of which are depicted as falling along a sexual continuum (e.g., bisexual), and others, which defy traditional gender dichotomies (e.g., pansexual). Further, some contemporary youth reject any sexual identity labels. The same way that White and Black are not sufficient racial categories, the same can be said of categories for other forms of stratification. Categories are social constructs that develop and evolve over time. Therefore, the anticategorical complexity approach focuses on the avoidance of placing individuals into preformed categories. This approach is similar in vain to a poststructuralist perspective, which opposes binary labels as they suggest a hierarchy or opposition between categories. Poststructuralist researchers are encouraged to deconstruct the labels to understand the realities of individuals who may not ascribe to such labels. Considering the proliferation of progressive sexualities that defy simply categorization, many practitioners of queer theory have opted to utilize anticategorical complexity approaches in contemporary scholarship.

The intracategorical complexity approach, also referred to as the microreductionism approach (Walby, 2007), has managed the complexity of intersectionality with the use of case studies, along with ethnographic and narrative methods, to analyze the intersections of multiple social categories. Many researchers have utilized case studies to identify 'invisible' populations not commonly represented in the literature. In doing so, researchers shed light on these 'new'" groups, but by utilizing case studies of individuals in these groups they uncover the differences *within* the group (hence the *intra*). This approach can yield valuable qualitative insight on understudied group, but this approach can be problematic because it has "become a strategy of seeking out ever finer units for analysis, in pursuit of a pure intersecting category. But, there are no pure groups; there are always more forms of difference" (Walby, 2007: 457).

In the continued pursuit of carving out new intersectional methodologies, many scholars find themselves researching 'storied' social categories that seem to have been reified as concrete entities by prior research. For example, categories such as African-American women and Latinas have been extensively researched and therefore serve as excellent starting points in terms of studying intersectional dynamics. This approach to intersectionality, dubbed intercategorical complexity, "requires that scholars provisionally adopt existing analytical categories to document relationships of inequality among social groups and changing configurations of inequality along multiple and conflicting dimensions" (McCall, 2005: 1773). Some scholars view this approach as contradicting the design of other forms of intersectional research. However, McCall emphasizes that this approach is not contradictory – it simply involves analyzing relationships of inequality between two or more social groups instead of analyzing complexities within a single social group. In other words, this approach is helpful in ascertaining the closing or widening of gaps in social outcomes that may exist between groups. For example, an intercategorical analysis of race and ethnicity would start with broad categories like Black and White, Hispanic and non-Hispanic. Analyses could then investigate intergroup differences along more detailed social groups that lie at race/ethnicity intersections (e.g., White Hispanics, Black Hispanics, White non-Hispanics, and Black non-Hispanics). Other dimensions, such as social class or sexuality can be added, which result in increased intercategory comparisons. These comparisons may appear rather reductionist when viewed as individual comparisons, but their combined analyses yield detailed intersectional findings that are quite holistic once pieced together.

Intercategorical complexity has been employed as a helpful approach for intersectional inquiry in various quantitative studies. Considering the structural reliance on the use of social categories in most quantitative research, this approach enables finite scrutiny of intersectional differences between group outcomes. Rather than using social categories (race, gender, class) as stand-alone independent variables, the use of 'interaction terms' or 'interaction effects' enables researchers to engage in intercategorical analyses to learn more about the combined effects of various social attributes. An example of this can be seen in a recent article by Guittar and Pals (2014), where the researchers investigated the combined effects of gender, race, and religiosity on attitudes toward homosexuality. One could argue that such a research design treats social categories as essentially different. But the use of interaction terms in ordinary least squares (OLS) regression analyses (which include the use of various control variables) can isolate extraneous influences, which oftentimes result in the appearance of 'real' differences rooted in race or gender. In the Guittar and Pals study they found that purported intersectional differences between race/gender categories (Black men, White

women, etc.) were not about race or gender itself, but about religiosity and gender role beliefs among these intersectional categories of people.

Intersectionality is still a young, evolving methodology and theory for studying unique experiences that lie at the intersection of various social categories. From an original focus on the intersection of race and gender to a contemporary emphasis on broader topics, such as the transcendence of categorization, intersectionality aims to draw scholarly attention to the variation in people's lived experiences. People have layered identities rooted in privilege and oppression, and these identities contribute to disparate social outcomes and unequal access to valued resources. Intersectionality empowers some scholars to emphasize within-group variation while enabling others to emphasize across-group differences in social outcomes. While the approaches to intersectionality are quite varied, they all hold consistent the notion that our identities are complex, and that we must be ever mindful of our relationships to numerous structures of power.

See also: Inequality, Social; Minorities; Phenomenology in Sociology; Postmodernism in Sociology; Power in Society; Prejudice in Society: Sociological Perspectives; Race Identity; Racial Relations; Racism, Sociology of; Social Classification; Social Identity in Sociology; Social Stratification; Sociological Theory.

Bibliography

Andersen, M.L., Collins, P.H., 2012. Race, Class, & Gender: An Anthology, eighth ed. Thomson Wadsworth, Belmont, CA.

Barnett, B., 1993. Invisible southern black women leaders in the civil rights movement: the triple constraints of gender, race, and class. Gender & Society 7 (2), 162–182.

Cole, E.R., 2008. Coalitions as a model for intersectionality: from practice to theory. Sex Roles 59, 443–453.

Collins, P.H., 2000. Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment. Routledge, New York.

Crenshaw, K., 1989. Mapping the margins: intersectionality, identity politics, and violence against women of color. Stanford Law Review 43, 1241–1299.

Davis, K., 2008. Intersectionality as buzzword: a sociology of science perspective on what makes a feminist theory successful. Feminist Theory 9 (2), 67–85.

Du Bois, W.E.B., Edwards, B.H., 2007. The Souls of Black Folk. Oxford University Press, Oxford.

Frye, M., 1983. The Politics of Reality. The Crossing Press, Trumansburg, NY.

Guittar, N.A., Pals, H., 2014. Intersecting gender with race and religiosity: do unique social categories explain attitudes toward homosexuality? Current Sociology 62, 41–62.

Katz, J., 2007. The Invention of Heterosexuality (With a New Preface). University of Chicago Press, Chicago, IL.

Kimmel, M.S., Ferber, A.L., 2010. Privilege: A Reader. Westview Press, Boulder, CO.

Lucal, B., 2008. Building boxes and policing boundaries: (De)Constructing intersexuality, transgender, and bisexuality. Sociology Compass 10, 519–536.

McCall, L., 2005. The complexity of intersectionality. Signs 30, 1771–1800.

Schutz, A., 1967. The Phenomenology of the Social World. Northwestern University Press, Evanston, IL.

Smith, D., 1979. Sociology for women. In: Sherman, J.A., Beck, E.T. (Eds.), The Prism of Sex: Essays in the Sociology of Knowledge. University of Wisconsin Press, Madison, WI, pp. 135–187.

Walby, S., 2007. Complexity theory, systems theory, and multiple intersecting social inequalities. Philosophy of the Social Sciences 37, 449–470.

Weber, L., 1998. A conceptual framework for understanding race, class, gender, and sexuality. Psychology of Women Quarterly 22, 13–32.

Yon-Leau, C., Munoz-Laboy, M., 2010. 'I Don't like to say that I'm anything': sexuality politics and cultural critique among sexual-minority latino youth. Sexuality Research and Social Policy 7, 105–117.

Intersectionality and Crime

Brenda S Blackwell, Georgia Southern University, Statesboro, GA, USA
Jennifer Cruze, Georgia State University, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Intersectionality, coined by Crenshaw (1989), is a framework utilized across a broad range of social sciences to explore how individuals within multiple systems of oppression have unique experiences and responses to those experiences. Within the study of crime, intersectionality has primarily been utilized by feminist criminologists to identify how and why women with different social statuses or from different racial or ethnic backgrounds have different levels of criminal involvement and experiences with victimization. This article briefly reviews the state of research utilizing intersectionality frameworks to examine crime and victimization.

Introduction

The concept of 'intersectionality' as relevant to the study of power differentials between groups and individuals was first raised in the 1980s. Researchers interested in a variety of status defining variables, such as race, gender, and class, recognized that certain groups of individuals crossed lines, or categories, of oppression – experiencing oppression based on more than one social category or socially benefiting because of statuses attained across more than one category. Crenshaw (1989) first coined the term 'intersectionality' in her efforts to portray how laws may actually omit particular groups experiencing discrimination (or oppression). Since this time, the use of intersectionality to formulate research questions and to theorize about relational oppression has expanded. Studies employing intersectional frameworks can be found across a wide range of disciplines, such as political science, social work, public policy, public health, gerontology, psychology, and sociology (Best et al., 2011; Daly, 1997). While a number of scholars have recognized the impacts of specific structural positions denoted by membership in specific categories (such as race, sex, and class) in the field of criminology, it is only recently that scholars specifically and deliberately have begun to apply intersectional frameworks to develop and explore questions within the fields of criminology and criminal justice.

Its relatively new stature within criminology and criminal justice yields a framework that has been inconsistently utilized, and which, when utilized, is associated with a variety of terms theoretical questions, and methodological approaches. This article explores the use of 'intersectionality' in criminology and criminal justice. We start with a description of intersectionality, examining its roots, key methodological discussions, and key issues. We then review criminal justice and criminological research that has utilized intersectional approaches, both overtly and through less deliberate tests of intersectionality.

Background

Conceptually, the ideas behind intersectionality grew out of criticisms predominately offered by black feminists critical of mainstream feminism for assuming that all women had shared experiences. Such critiques declared that the focuses of feminism generally reflected the desires and needs of middle-class women, signally a lack of understanding that black women, historically and today, have significantly different experiences when compared to their white counterparts (Sokoloff and Burgess-Proctor, 2011). Movements within black feminism and throughout women's studies increasingly revealed that not all women are the same. Hence, predicating research, policies, and programs on the assumption that women stand as a single analytical category was argued to be a flawed approach (McCall, 2005). Importantly, it was recognized that simple additive models that included gender and race missed the unique and complex experiences of black women, poor black women in particular, who occupy a unique social position generated by overlapping systems of oppressions (Burgess-Proctor, 2006; Collins, 2007). As Bernard (2013: pp. 3–4) noted, positions held by marginalized women who occupy multiple subjugated statuses yield "vulnerabilities that … can be difficult to measure." While race was highlighted as a crucial factor distinguishing women within early works (Crenshaw, 1989; but see also Collins, 2007), additional structuring factors, such as class or age, were soon recognized as influential in shaping experiences. To date, however, the majority of intersectionality research still primarily considers gender, race, and class, which are treated as the key master statuses comprising the axis of oppression.

While intersectionality generated a great deal of discussion and debate within gender studies and legal feminist discussions, intersectionality works in criminology lagged. This is not to say that criminological research has not examined the multiplicative effects yielded for those oppressed by race, sex, and class, but rather that claims of theory building through intersectionality or research driven by intersectional methods have not been the norm. Instead, while early criminological studies produced by researchers such as Daly (1989), Messerschmidt (1993), and Miller (1998) clearly reflected intersectionality as foundational in questions, approaches, and methods, explicit reference to an 'intersectional framework' per se was not always highlighted. Recent years have seen a rise in intersectionality references in research on crime and victimization.

As with research in other fields, in criminology, the primary use of intersectionality as a framework has been to provide a tool that ensures greater understanding of the systems in place and processes that work together to create inequality for certain groups and subgroups in society. While research in

mainstream criminology does at times reflect this understanding, more overt discussions of unique statuses created through the intersections of multiple oppressed categories is most likely to be found by researchers interested in the study of women and crime. As with other fields, the primary master statuses considered are sex, race, and class; the assumption is that the intersections of inequalities generated by these socially structured systems yield multiplicative, not simply additive, results that predict not only offending but also victimization.

Given broader discussions of intersectionality, and the debate about its usefulness for theory building, as well as for the development of research strategies, it is useful to consider how intersectionality is linked to theoretical developments in the study of crime, as well as how it has been employed to design research that is aimed at understanding the multiple positions of women. Notably, authors across numerous fields have described intersectionality as a framework, or paradigm, useful for theory building and testing rather than as a theory itself (Davis, 2008).

Theoretically, intersectionality is used to stress the need for criminological research to reconsider the broad categories frequently utilized to understand criminality within extant theories. As well, discussions of identity and systems that externally structure development of identities within specific groups or systems have emerged (Messerschmidt, 1993). Further, the need to develop multilevel explanations for crime and responses to crime are underscored by intersectional research, as systems of oppression do not work independently, but rather interactively, to produce various planes of oppression (Bernard, 2013; Collins, 2007). Finally, this perspective focuses attention on the role of relationships between individuals; hence, theories that recognize the role of relationships in generating and reflecting power differentials and resultant influences on crime and victimization are prime candidates for expansion through intersectional frameworks (Sokoloff and Burgess-Proctor, 2011; Bernard, 2013).

Methodological Approaches and Intersectionality

Because the foundation of intersectionality is in the study of women and the law, use of this framework by legal scholars is unsurprising. Legal research has explored the utilization of law to understand and promote human rights for women, immigrants, children, and other groups with minority statuses, marriage and other rights for same-sex couples, and to meet the needs of adults and children with disabilities (Best et al., 2011). Ultimately, legal scholars use intersectionality to examine how the law, by ignoring intersectional statuses, may actually exclude groups who exist in multiple oppressions because such groups are not represented within many legal protections (Conaghan, 2009). The effective portrayal of the problems facing black women by Crenshaw (1989) since has been applied to others who experience multiple oppressions to examine how intersecting oppressive statuses influence treatment under the law and how the law can be utilized to redress inequities.

With acceptance of intersectionality extending outside of legal scholarship came questions of how this framework could be utilized in research to address additional theoretical and policy questions. Understanding the broader picture of how intersectionality is linked to research methodologies is important for any review or utilization of this framework. McCall (2005) provided a comprehensive analysis of how using intersectionality requires careful consideration in methodological approaches due to the complexity of the social phenomenon that it seeks to address. Acknowledging that a broad array of methodologies are utilized in intersectional research, McCall classifies these methodologies based on forms that intersectionality may take, which she references as anticategorical complexity, intercategorical complexity, and intracatagorical complexity.

According to McCall (2005), intersectional research was initiated within an intracategorical framework, which recognizes, but is critical of, stable, enduring social categories such as race, sex, and class. Intracategorical research primarily is concerned with identifying groups that emerge at specific intersections (such as sex and race or race and class) and highlighting the importance of shared points of oppression. Research conducted in the intracategorical approach, while conceding that there are such social categories, ultimately questions the validity of the processes through which social boundaries are created and defined. To accomplish this, researchers may select a broad category – such as women – then seek to identify differences between women that occur on the basis of other social categories (class or race, for instance). Intracategorial research frequently takes form through narrative essays, case studies, and studies of single groups, as well as through ethnographic research.

Anticategorical research is more keen to recognize the complexity of social life and, more specifically, socially identified categories that yield or reflect inequities. Researchers utilizing the anticategorical approach, according to McCall, question the viability of categories, and ultimately deconstruct categories, because they recognize that individuals are not only members of multiple oppressed categories but also that such membership is both fluid and relative. Such is the case, for example, for scholars interested in research on transgendered individuals, who argue that gender contains many more categories than just male and female or masculine and feminine. McCall (2005) links methodologies such as geneology and new ethnography to this category of intersectional study.

Finally, intercategorical complexity falls on the more conservative end of the spectrum. Researchers employing an intercategorical approach recognize and accept extent social groups and categories associated with oppressive experiences. Researchers from this frame are interested in the differences between categories and the relationships present between the oppressed and the powerful. They are particularly interested in how interactions between socially situated groups are created, maintained, and managed by both those in power positions, as well as those experiencing inequality. Scholarship from this vein is the most likely to employ quantitative methods alongside qualitative methods; however, with the addition of complexity added by the intersectional approach, researchers must employ more sophisticated methodologies. More specifically, quantitative research from an intercategorical approach is punctuated by interaction effects (rather than simple additive models), multilevel modeling, ecological models, or contextual approaches because only by jointly

examining structural categories can the effects of intersections be adequately examined (McCall, 2005).

McCall (2005) concludes that these organizing categories are not the only way such research can be classified; rather, McCall recognizes her scheme may be incomplete, arguing that researchers should not accept these categories within her classification scheme as immutable, but rather that research can, and likely should, reach across these different frameworks to more fully examine questions and yield more comprehensive conclusions.

While others have discussed the methodological approaches to the study of intersectionality, McCall's (2005) classification schemes have garnered substantial acceptance across literatures, and her caveats are well attended. Indeed, most research addressing intersectionality seems to employ a mixed methods approach with qualitative methods considered key to understanding intersectional experiences identified by quantitative approaches.

Intersectionality and Criminality: Theoretical Links and Developments

As previously noted, within criminology, the use of intersectionality as a framework is relatively recent, and primarily is utilized within feminist criminology to enhance understanding of female criminality and victimization. Since early critiques that social scientists, and criminologists in particular, neglected women in the study of deviance, the study of women and crime has significantly increased. Daly and Chesney-Lind (1988) called for researchers interested in gender and crime to address a number of key questions, two of which have served as the foundation of most research in this arena. First, researchers have begun to explore the viability of traditional criminological theories for enhancing our understanding of female criminality, known as the 'generalizability question.' To a lesser degree, researchers have examined the 'gender ratio' question – searching for explanations of the gender gap in criminality (Daly and Chesney-Lind, 1988). More recently, however, criminologists have specifically begun to explore differences in criminality that exist 'within' sex categories, as well as between, yielding work that identifies gendered pathways into offending (Daly, 1989; Messerschmidt, 1993; Miller and Mullins, 2006), as well as work more directly nested in an intersectional paradigm (see discussion in Burgess-Proctor, 2006).

Today, intersectionality is inherently embedded in the feminist criminology, which Miller and Mullins (2006: p. 221) suggest is primarily linked to providing 'narrower, more focused explanations of contexts and situations,' or middle-range theorizing. As such, when a broad review of the literature is taken, one can identify research that reflects intersectional analyses, even though such research may not explicitly reference intersectionality at the outset. Researchers have drawn from intersectionality to reexamine traditional theories and to hone concepts such as class within such theories. As well, research on 'doing identity' frequently utilizes intersectionality. Moreover, intersectionality exists at the foundation of multilevel explanations for crime, as well as research aimed at uncovering how relationships between individuals affect criminality. Finally, more recent research drawing from intersectional approaches has pointed out the need to understand the intersections of crime–victim statuses in generating unique offender subgroups.

Traditional Theories of Crime and Intersectionality

Intersectional research that has led to the reconsideration of broad categories within traditional criminological theories has yielded conclusions that this framework is useful for identifying weaknesses in such perspectives, as well as for suggesting improvements. For example, using quantitative analyses, Simpson and Gibbs (2006) explored interactions between race, gender, and class within traditional criminological models to determine if an intersectionality framework would identify differences in effects of theoretical constructs not yet identified. Their findings indicate that while traditional theories are useful, dissecting constructs using this framework identified differences that should be further explored.

Likewise, Bernard (2013) delved into Merton's strain theory drawing on an intersectionality framework to demonstrate how concepts from this perspective are improved with consideration of differences within subgroups. Bernard noted that Merton, in his version of strain theory, suggested that crime is the result of strains produced by feelings of frustration yielded for particular segments of society who accept the goal of success but do not have available legitimate means to attain these goals. Bernard extended Merton's discussion by arguing that women, in contrast to men, may be more likely to become involved in criminality as a result of the constraints they experience rather than strains; she noted that women who exist under multiple vulnerable statuses respond to oppressions according to the social contexts they occupy. She concluded by highlighting that 'doing identity' reflects how groups who occupy intersectional statuses adapt to their unique positions.

A broader exploration of multiple criminological theories generated the conclusion that while traditional criminological theories are useful for understanding broad differences in offending across boys and girls, important differences exist within these categories for which such theories cannot account. For example, Katz (2000) found differences in correlates of delinquency between black and white females. Specifically, among females, black youths were more delinquent when they reported feelings of alienation, while white delinquent girls were more likely to report feelings of depression.

Notably, and perhaps not surprisingly, most feminist criminologists' applications of intersectionality to improving theories of crime have focused on gender and race. Bell (2013), however, is critical of the categories of race that have been the focus of most of this research, as it is largely limited to comparing white and black offenders. She draws on intersectionality to examine the life course perspective, with a focus on determining whether the intersections of multiple oppressions are more or less salient in generating delinquency and crime depending on life course stages. In contrast to other research, she considers the status of Hispanics, as well as whites and blacks. Her results yielded support for the application of intersectionality, with the intersection of race and gender playing a key role for young adults, but not for juveniles. Specifically, among juveniles race and gender had direct effect, with boys more likely than girls to commit delinquent offenses

and blacks more likely than whites and Hispanics to be involved in delinquency. However, no race × gender interactions yielded significant results. Among young adults, in contrast, Bell (2013) reported that race and gender did significantly interact, and moreover, the race effect was enhanced by broadening the categories considered. Specifically, young adult males were more likely than not to be offenders, and young Latino men were more likely to be involved in property crimes than to be criminally uninvolved (Bell, 2013: p. 114). All females were less likely than males to be offenders when they were young adults.

In sum, research generated from an intersectionality perspective supports the need to consider how individuals positioned in multiple oppressive statuses may not fit the models generated by traditional criminological theories. Furthermore, this emergent literature yields the conclusion that patterns identified through quantitative analyses could be further understood by exploring differences using qualitative research. This approach, as suggested by McCall (2005), would be useful for improving our knowledge and further honing our theories.

Considering Intersectionality in 'Doing' Identity through Crime

Another branch of intersectionality research recognizes that managing multiple social oppressions leads some individuals into criminal behavior as a mechanism for achieving success. This branch of intersectionality is primarily associated with the works of Pat Carlen (1988) and James Messerschmidt (1993), each of whom associates intersections of gender and class with criminality, with the former focusing on criminality of women, and the latter on male participation in crime. Each scholar recognizes that one's position in larger social structures leads to different forms of relations at an individual level. These interactions are the basis in which individuals 'do identity' (Bernard, 2013), and it is the context in which these interactions occur that dictates whether an individual's behaviors are deemed as appropriate or deviant. The concern is with how individuals 'do gender,' or act in ways considered appropriate in any given situation based on one's sex category.

Carlen (1988) draws on the association between class and gender to examine women's participation in crime by linking failure to achieve 'success' in one of these two domains to criminal participation. She describes gender norms for women as striving to have happy and strong families with men who are the primary breadwinners. According to Carlen, women's participation in crime results when either the woman must generate their own income (violating the 'class deal') or when she is unable to develop or maintain a happy family life and/or relationship (violating the 'gender deal'). Notably, in her work, Carlen recognizes the importance of multiple statuses and does not position either class or gender as more important or relevant than the other, rather it is the process of navigating the experiences generated by unique positioning of multiple oppressions that is of interest.

While Messerschmidt (1993) also is interested in class and gender, in contrast to Carlen his position is that crime is one way marginalized youths have to 'do' masculinity when other, legitimate, routes for demonstrating masculinity are unavailable. Messerschmidt argues that masculinity may be achieved in different ways depending on the social context in which an interaction takes place. Researchers continue to build on Messerschmidt's work in efforts to deepen our understanding of factors that influence behaviors linked to 'doing masculinity.' For example, Shade et al. (2011) explore the role of fatherhood in 'doing masculinity' for adolescent males involved in the juvenile justice system, concluding that understanding life course stages and experiences within intersectional contexts is important for the development of policies and programs to reduce teen parenthood.

Miller (1998) extends 'doing identity' work extensively across numerous studies. Like Messerschmidt and Carlen, Miller asserts that gender is achieved situationally, as males and females engage in behaviors that establish their femininity and masculinity. For example, while Miller (1998) notes that males and females cite similar reasons behind their involvement in robbery, such as thrill seeking and as a way to obtain money or other material goods, the methods they utilize distinctly reflect 'doing gender.' Specifically, men's use of physical force and targeting of male victims in the commission of robbery allows them to demonstrate their masculinity. Female robbers, in contrast, varied their methods by their victim; women are likely to target other women as victims – citing less need for force or use sex if they are robbing a male victim. Likewise involvement and activities for boys and girls involved in gangs reflect situationally defined ways of 'doing gender' (Miller, 2008).

Reviewing criminological work on 'doing identities' highlights the usefulness of intersectionality for examining not only microlevel intersections but also intersections that occur at the macrolevel and generate unique settings, contexts, and groups relevant to understanding crime rates as well as criminal behaviors. As described by Collins (2007), macrolevel influences also intersect, or interlock, generating group level positions associated with multiple factors that structure experiences, as well as available choices and social and personal capital for those who inhabit those spaces.

Intersectionality and Crime at the Macrolevel

Less criminological research has utilized intersectionality to examine macrolevel differences in crime. This is surprising given the theoretical frameworks that call attention to macrolevel phenomenon. For example, a number of criminological theories consider class (anomic, strain, social disorganization) and race (social disorganization, racial threat) at the community level. However, it is notable that macrolevel theories rarely consider the impacts of gender. Bell (2013) is critical of this omission, citing research that depicts sex differences in criminality within communities; in marginalized communities, it is noted that while black women may have higher offending rates than white women, their rates of offending are lower than those of black men. Hence, even at the macrolevel, within-sex differences exist.

Haynie and Armstrong (2006) and Parker and Hefner (2013) utilized intersectionality to examine rates of domestic violence and homicide, paying particular attention to the victims of these offenses. Each study concluded that

drawing on such a framework yielded findings of interest. Specifically, in reviewing homicide rates of cities, Haynie and Armstrong (2006) determined that not only are gender and race separately strong predictors, unique findings are uncovered when race and gender are considered together. For example, while structural disadvantage was positively associated with black and white females' homicide rates, its association with the former was stronger. However, when comparing the effect of structural disadvantage within race, its association with male homicide rates was stronger than for females among whites but stronger for females' than males' among blacks. Likewise Parker and Hefner (2013) examined intersectionality at the macrolevel, specifically exploring the effects of economic disadvantages, family structure, and changes in incarceration and domestic violence policies with a focus on those who experience multiple disadvantaged statuses – specifically race and gender. They also determined that the intersectional framework yielded a more elaborate understanding of the unique positions and disadvantages held by black women in our society, as economic deprivation was more salient for understanding the homicide rates of black women than of white women.

Using Intersectionality to Understand System Involvement

The links between race, class, sex, and offending have long been explored, and the recent shifts to examining their intersecting effects for those experiencing multiple oppressive statuses clearly point to the need to further utilize an intersectionality framework to enhance our understanding of offending. Just as these and other variables have been identified as salient to understand offending, it also has been recognized that they impact how an offenders and victims experience the criminal justice system. Steffensmeier et al. (1998) outlined the focal concerns model for enhancing our understanding of judicial sentencing decisions and the ways that race, gender, and age both separately and together impact such decisions. According to the focal concerns perspective, it is those who reside at the intersection of multiple statuses of inequality, namely race, age, and gender, who are perceived in the justice system as the most likely to put society at risk. This model purports that judges make decisions based on three focal concerns: offender blameworthiness and amount of harm experienced by victim, community protection, and individual and organizational practical implications of sentence outcomes. Because judges frequently have little contact or time to observe offenders as a result of the plea-bargaining process, they may draw on additional information from experiences and stereotypes to aid in their decision-making process. This yields the targeting of young, black, male offenders as those who experience more severe sanctions. Their research (Steffensmeier et al., 1998) supports the focal concerns model; the effects of race were dependent on the gender and age of the offender, the effect of gender was dependent on race and age, and the effect of age was influenced by gender and race. Qualitative data reflecting judicial perceptions of offenders based on race, gender, and age further explicated findings, supporting the focal concerns assumption that young black male offenders were perceived as more dangerous than other groups, as well as the belief that incarcerating women and older offenders, who are seen as less blameworthy, yields higher social costs. Extensions of this framework to examine bail (Demuth and Steffensmeier, 2004) and parole decisions (Huebner and Bynum, 2006), along with additional studies of sentencing, provide further support that these variables, and specifically different combinations of them, yield interactive rather than simple additive effects.

The focal concerns perspective reflects the intersectionality framework, with findings that not only do multiple statuses interact to create unique positions affecting identities and resultant experiences, but also to influence others' reactions to them. Indeed, support for the focal concerns perspective indicates that system officials look not simply at singular characteristics when making decisions but rather simultaneously consider multiple characteristics that reflect such statuses. These findings have far-reaching implications, particularly when policies are developed that reflect past/current practices in decision making or when they target particular groups. Indeed, poor women and men of color are disproportionately impacted by a myriad of policies, ranging from drugs to welfare and foster care to housing. Notably, it is not only offenders who are impacted by such policies but also victims. Furthermore, research on victims reveals that those with multiple oppressive statuses consider such policies as they determine whether to rely on the system for help.

Victims of Crime: Views from an Intersectional Lens

Sokoloff and Burgess-Proctor (2011) identify intersectionality as a key approach in the study of crime and victimization because of its ability to identify how systematized elements of inequality intersect to provide groups who are more (or less) likely to become engaged in criminal behavior or to be victimized. In this vein, scholars interested in victims of crime have begun to utilize intersectionality to expand understanding of not only who victims are but also whether programs and resources are appropriate across victims who present with different statuses.

Work on the portrayal of victims from an intersectional framework demonstrates that not all victims receive the same attention. Indeed, not only are victims with certain characteristics more (or less) likely to garner media attention, but the types of media and the portrayals of victims in the media vary according to social statuses. For example, white, middle, and upper class women who are missing or are victimized are not only more likely to be deemed newsworthy but also are less likely to be portrayed as holding responsibility for their victimization than counterparts who are poorer, lower status, and/or, are minorities (Gilchrist, 2010); those who hold multiple statuses linked to oppression are the least likely to receive any or positive media attention.

As with offending, application of an intersectionality framework to victimization is believed to enhance our understanding of victims' experiences within a social justice framework. Domestic violence, in particular, has been viewed with an intersectional lens, with the goals of improving theories, understanding justice responses, and improving services provided to victims and offenders involved in battering

relationships. While widely recognized that women are more likely than men to be injured in domestic violence incidents, less is known about how marginalized women experience domestic violence. Early intersectional research drew attention to the fact that black and white women's lives yield very different experiences with and responses to domestic violence. Sokoloff and Dupont (2005) recently reaffirmed that this framework is an important tool to utilize to enhance our understanding of the unique characteristics of intimate partner violence in marginalized communities, specifically minority, lower-class communities. Research in the past decade has further drawn on this perspective to explore the distinct experiences of other groups, such as immigrants, disabled individuals, or same-sex couples, in domestic violence experiences.

Using this framework, researchers have elucidated how women's responses to domestic violence are impacted by the multiple statuses that they hold, concluding that the statuses held by victims yield different reactions or different groups of victims may have different reasons far behind similar responses to such events (Josephson, 2002). For example, assumptions are made that black women who are victimized by black men are discouraged from reporting black men to police in domestic violence incidents, and that, as a result, they are less likely to report when they are assaulted in these communities or by their partners. Moreover, it is posited that these victims may fear that they will lose custody of their children if they report abuse. This is particularly the case when these victims reside in communities that are segregated not only racially but also by class (Cramer and Plummer, 2009). Findings, however, are mixed; some evidence supports these assumptions and other studies indicate that poor minority women are more, not less, likely to seek prosecution of their abusers (see discussion in Josephson, 2002). More research clearly is needed to further identify commonalities and differences across domestic violence victims.

Immigrant status adds an additional layer, particularly for those who are in the country illegally; these women are isolated not only within their communities but also from services available to broader categories of those who experience domestic violence. Erez et al. (2008) report that these victims also are threatened by their abusers with their immigration status, further reducing the likelihood that they will report assault incidents to police. Domestic violence within same-sex relationships also has been reviewed using an intersectional lens, with researchers noting that societal level homophobia impacts how the justice system responds to abuse within such couples; moreover, individuals in these relationships may be less likely to report or seek services because of their fears about how society views them (Sokoloff and Dupont, 2005).

Studies examining how system officials and service providers respond to victims of domestic violence reinforce the need for intersectional framing. For some time, concerns have been raised that services for victims of interpersonal violence were developed with a goal of responding to 'average' victims, but that the 'average' victim is conceptualized from a white majority framework. Josephson (2002) asserts that this has resulted in programs and services that do not recognize the experiences of not only minority women, but more importantly, women who present with a complex array of linked issues, including minority status, poverty, and residence in communities without resources or strong social networks. She argues that provision of services is hampered by the context of the multiple marginalizations that such women experience.

For example, many already utilize social services and therefore have developed relationships with government agencies associated with such services. Hence, they have formulated opinions about government agencies in general and may be less likely to seek out assistance when they are victimized because of prior negative experiences with agency personnel. In addition, because women in marginalized communities have less social capital and resources to draw on, they recognize that they are potentially more at risk when they seek assistance.

Moreover, frequently agencies in marginalized communities focus on addressing a single issue and as a result fail to recognize and address the myriad of factors that impact those who have multiple layers of oppression that need to be addressed in order to provide an adequate response to victimization. For example, groups aimed at addressing racism or poverty may fail to recognize or address compounding factors generated for those who experience DV or other victimizations (Sokoloff and Dupont, 2005; Josephson, 2002). This is further hampered by policies that, developed with an eye to one issue, ultimately constrain victims from making decisions that would help them leave abusive situations. Josephson (2002) points specifically to the requirements of AFDC/TANF that affect payments and potentially even contacts with abusers in order to ensure establishment of paternity and/or child support payments as potentially harmful to victims who desire to leave or reduce contact with abusers.

Finally, intersectionality likewise points to the need to explore resources available for individuals who do not fit into even these categories. Cramer and Plummer (2009) call attention to the fact that women who are disabled are quite isolated, frequently relying on their abuser for mobility; moreover, abuse of such individuals may not be physically violent but instead involve threats to or withholding of medicine or treatments. These victims, as well as elderly victims, are underexplored within the criminological literature and it is a lens of intersectionality that is bringing attention to them as generally unrecognized and underserved.

Conclusion

In sum, utilizing intersectionality as a lens to review traditional criminological theories, as well as to examine crime policies and justice practices, has begun to highlight populations that have to date been underexplored. While critics of intersectionality frequently raise concerns that use of this framework runs the risk of deconstruction, researchers studying crime have effectively employed intersectionality to raise awareness of how individuals who hold multiple oppressive statuses are further victimized, not only as victims of crime but even as offenders in a system that is underprepared in its identification of sources of criminality and responses to victims and offenders. As intersectionality work continues to evolve in the study of crime, we can anticipate that additional populations who are underserved will be identified, greater consideration of contradicting policies and programs will occur, and resources will be better utilized to reduce crime propensity and serve victims. Furthermore, it is

likely that such research will also clarify how experiencing these complex identities impacts not only offending or victimization but also how these two sides of criminal events share experiences within some shared settings. The intersectional lens holds significant promise for future research in the study of crime.

See also: Communities and Crime; Crime and Class; Crime and Gender; Crime, Race and Ethnicity; Feminist Criminology and Gender Studies; Intersectionality; Queer Theory and Intersectionality.

Bibliography

Bell, K.E., 2013. Young adult offending: intersectionality of gender and race. Critical Criminology 21, 103–121.

Bernard, A., 2013. The intersectional alternative: explaining female criminality. Feminist Criminology 8 (1), 3–19.

Best, R.K., Krieger, L.H., Edelman, L.B., Eliason, S.R., 2011. Multiple disadvantages: an empirical test of intersectionality theory in EEO litigation. Law and Society Review 45, 991–1025.

Burgess-Proctor, A., 2006. Intersections of race, class, gender, and crime: future directions for feminist criminology. Feminist Criminology 1 (1), 27–47.

Carlen, P., 1988. Women, Crime, and Poverty. Open University Press, Bristol, PA.

Collins, P.H., 2007. Pushing the boundaries or business as usual? Race, class, and gender studies and sociological inquiry. In: Calhoun, C. (Ed.), Sociology in America: A History. University of Chicago Press, Chicago, pp. 572–604.

Conaghan, J., 2009. Intersectionality and the feminist project in law. In: Grabham, E., Cooper, D., Krishnadas, J., Herman, D. (Eds.), Intersectionality and beyond: Law, Power and the Politics of Location. Routledge, London, pp. 21–48.

Cramer, E.P., Plummer, S., 2009. People of color with disabilities: intersectionality as a framework for analyzing intimate partner violence in social, historical, and political contexts. Journal of Aggression, Maltreatment, and Trauma 18 (2), 162–181.

Crenshaw, K., 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory, and Antiracist Politics, vol. 14. University of Chicago Legal Forum, pp. 538–554.

Daly, K., Chesney-Lind, M., 1988. Feminism and criminology. Justice Quarterly 5 (4), 497–538.

Daly, M., 1989. Women and Poverty. Attic Press, Dublin.

Daly, K., 1997. Different ways of conceptualizing sex/gender in feminist theory and their implications for criminology. Theoretical Criminology 1 (1), 25–51.

Davis, K., 2008. Intersectionality as buzzword: a sociology of science perspective on what makes a feminist theory successful. Feminist Theory 9 (1), 67–85.

Demuth, S., Steffensmeier, D., 2004. The impact of gender and race-ethnicity in the pretrial release process. Social Problems 51, 222–242.

Erez, E., Adelman, M., Gregory, C., 2008. Intersections of immigration and domestic violence: voices of battered immigrant women. Feminist Criminology 4 (1), 32–56.

Gilchrist, K., 2010. "Newsworthy" victims? Exploring differences in Canadian local press coverage of missing/murdered aboriginal and white women. Feminist Media Studies 10 (4), 373–390.

Haynie, D.L., Armstrong, D.P., 2006. Race and gender-disaggregated homicide offending rates: differences and similarities by victim-offender relations across cities. Homicide Studies 10 (1), 3–32.

Huebner, B., Bynum, T.S., 2006. An analysis of parole decision making using a sample of sex offenders: a focal concerns perspective. Criminology 44 (4), 961–991.

Josephson, J., 2002. The intersectionality of domestic violence and welfare in the lives of poor women. Journal of Poverty 6 (1), 1–20.

Katz, R.S., 2000. Explaining girls' and women's crime and desistance in the context of their victimization experiences. Violence Against Women 6, 633–660.

McCall, L., 2005. The complexity of intersectionality. Signs 30 (3), 1771–1800.

Messerschmidt, James, 1993. Masculinities and Crime: Critique and Reconceptualization of Theory. Rowman and Littlefield, Lanham, MD.

Miller, J., 1998. Up it up: gender and the accomplishment of street robbery. Criminology 36 (1), 37–66.

Miller, J., 2008. Getting Played: African American Girls, Urban Inequality, and Gendered Violence. New York University Press, New York.

Miller, J., Mullins, C.W., 2006. The status of feminist theories in criminology. In: Cullen, F.T., Wright, J.P., Blevins, K.R. (Eds.), Taking Stock: The Status of Criminological Theory, Advances in Criminological Theory, vol. 16. Transaction, New Brunswick, NJ, pp. 216–249.

Parker, K.F., Hefner, M.K., 2013. Intersections of race, gender, disadvantage, and violence: applying intersectionality to the macro-level study of female homicide. Justice Quarterly. http://dx.doi.org/10.1080/07418824.2012.761719.

Shade, K., Kools, S., Weiss, S.J., Pinderhughs, H., 2011. A conceptual model of incarcerated adolescent fatherhood: adolescent identity development and the concept of intersectionality. Journal of Child and Adolescent Psychiatric Nursing 24, 98–104.

Simpson, S.S., Gibbs, C., 2006. Making sense of intersections. In: Heimer, K., Kruttschnitt, C. (Eds.), Gender and Crime: Patterns in Victimization and Offending. New York University Press, NY, pp. 269–302 (NCJ-214516).

Sokoloff, N.J., Dupont, I., 2005. Domestic violence at the intersections of race, class and gender: challenges and contributions to understanding violence against marginalized women in diverse communities. Violence Against Women 11 (1), 38–64.

Sokoloff, N.J., Burgess-Proctor, A., 2011. Remembering criminology's 'forgotten theme': seeking justice in U.S. crime policy using an intersectional approach. In: B, M., Hoyle, C. (Eds.), What Is Criminology? Oxford University Press, Oxford, UK, pp. 235–248.

Steffensmeier, D., Ulmer, J., Kramer, J., 1998. The interaction of race, gender, and age in criminal sentencing: the punishment cost of being young, black, and male. Criminology 36 (4), 763–798.

Interviewing, History of

Raymond M Lee, Royal Holloway University of London, Surrey, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes the historical development of interviewing in the social sciences. It begins by noting the transition toward the end of the nineteenth century from the interviewing of informants about the lives of other people toward the elicitation of direct accounts from individuals themselves. The history of survey interviewing in the modern sense is traced from the 1920s to the period just after World War II, and from the 1950s to the present day. The emergence of less structured styles of interviewing in the early part of the twentieth century is described. Attention is directed to the variety of critical approaches to traditional conceptions of interview that became important from the 1960s onward. Briefly discussed are changes in the work roles of interviewers, the interviewing of 'special populations,' and shifts in the social acceptability of the interview itself. Finally, the implications for the interview of recent developments in new communication technologies are outlined.

Introduction

As the oral historian Paul Thompson (2000) has shown, the direct elicitation of information through question and answer has a long and venerable history. The word 'interview' seems to have originated in the sixteenth century to describe a meeting of two people of equal rank and status to discuss a matter of some importance to them. In the later part of the nineteenth century, the practice of interviewing prominent individuals in order to obtain from them an authoritative view concerning some pressing topic of the day became a common practice in journalism (Schudson, 1994). Slightly later on, the term began to take on something more like its modern sense of a device for eliciting information about some aspect of people's lives. It was not, however, until the early decades of the twentieth century that interviewing began to be seen in a methodologically self-conscious way by social scientists interested in respondents' activities, circumstances, and subjective states. By the middle of the twentieth century, the interview – in its various guises – had some claim to have become a central, indeed one might say dominant, research method in the social sciences.

The Historiography of the Interview

No detailed full-length monographic history of the interview exists, although Jennifer Platt has provided a concise, and characteristically incisive, overview of the topic (2001). The early attempt by Riesman and Benney (1956) to sketch the history of interviewing must now be considered speculative. The development of interview practice in the social sciences has been characterized by many (often unattributed) borrowings and cross-fertilizations between anthropology, political science, psychology, consumer research, social work, and sociology. Moreover, although the history of social science interviewing is treated here largely in isolation, it cannot strictly speaking be divorced from related developments such as the emergence of sampling methods, or shifts in recording technologies (Lee, 2004). It should also be noted that, as Platt (2001) points out, the methods current at a particular time often now have to be recovered from prescriptive or programmatic accounts, those used for training, for example, which might present a partial or idealized view of actual practice.

Nowadays, interviews come in many different forms. They can be highly structured with standardized questions or loosely organized around a set of topics. They can be conducted face to face, by telephone, or online. While interviews are usually conducted with single individuals, they can also be used with small groups. The diversity of interview forms, the extent of variation in interview practice, and the theoretical and methodological traffic between and within disciplines all combine to make it difficult to produce a simple linear narrative of how the social science interview developed over time. It is convenient, however, to begin with the history of survey interviewing.

The History of the Survey Interview I

During the nineteenth century, concerns arose in a number of countries about the social conditions faced particularly by the urban poor and the means by which those conditions might be ameliorated. Shaped by such concerns, a variety of groups and individuals began to undertake social investigations to identify the nature and magnitude of social problems and to provide a basis for action (Bulmer et al., 1991). Often referred to as 'surveys' but distinct from the modern conception of the term, investigations of this kind relied on the systematic collection of information about families and households usually by means of an exhaustive enumeration of local areas. Although there were those like Henry Mayhew in England, Wilhelm Riehl in Germany, and Frédéric Le Play in France, Belgium, and elsewhere who interviewed workers, peasants, and others directly, and who were not unaware of the methodological issues involved in doing so, for the most part nineteenth century, social investigators tended to use officials or functionaries as proxy informants about social conditions rather than interviewing respondents directly. Mutual distrust and suspicion between the classes, underlying conceptions of social pathology, and practical problems of access encouraged researchers to devalue direct accounts

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03222-0

from respondents in favor of indirect methods for gathering information (Marsh, 1985).

Marsh (1985: pp. 214–216) points to a number of factors that, in England at least, encouraged the transition from respondent to informant interviewing. These include the use of directly employed field-workers rather than volunteers, improvements to working hours and employment conditions that increased accessibility to working class respondents, a shift away from pathologizing discourses that blamed the poor for their own misfortunes, and a growing interest in the subjective sphere of beliefs, attitudes, and values. That parallel shifts occurred in other countries is suggested by Sklar's (1991) account of Florence Kelley's work at the Hull House in Chicago, and Lazarsfeld and Obserschall's (1965) discussion of Max Weber's involvement in a succession of empirical research projects in Germany. Bulmer et al. (1991) note that in Britain the reformist social survey had largely run its course by 1914 and the onset of World War I. The tradition remained strong in the United States through the 1920s before eventually falling victim to the Great Depression (Converse, 1987). Economic austerity reduced the foundation funding, which was an important source of support for local surveys. Research on social conditions began increasingly to be conducted under state or federal auspices, and with an enhanced role for academic social scientists who brought with them newer ideas about sampling and measurement techniques. What Converse (1987: p. 39) calls the 'new attention to measurement' encouraged the collection of individual-level data, although the emphasis remained on behavioral rather than subjective measures. Together with the increased understanding and acceptability of probability sampling, the outlines of the modern sample survey began to emerge.

The growth of newspaper advertising, radio listening, and a desire by businesses, faced with shrinking profit margins as the economic climate worsened, to understand consumer attitudes and behavior all contributed to the development of a burgeoning market research industry in the United States in the 1920s and the 1930s (Converse, 1987). The exploration of consumer attitudes soon expanded to cover wider social and political issues, a trend buoyed by pollsters' successful prediction of the 1936 US Presidential election in contrast to the failure of the magazine *Literary Digest* to predict the election result with a huge but fatally unrepresentative mail-in ballot of its subscribers.

As the interpenetrated worlds of market research and opinion polling expanded, debates began to emerge on topics such as sampling, questionnaire design, question wording, and the role of interviewers. Disagreement – partly shaped by differences between academic and polling practice – existed about the extent to which the interview should be flexible or conversational in tone or follow a more tightly structured format based on standardized questions (Converse, 1987: pp. 95–98; pp. 137–138). The amount of freedom interviewers had to interpret questions and record responses could depend on quite specific matters such as the perceived sensitivity of particular questions, but practice also reflected institutional culture and organizational history. Those bodies that gave interviewers latitude in how they conducted the interview tended, by and large, to have smaller field forces, to recruit well-educated interviewers, and to provide them with detailed training. Different interviewing styles placed different demands on interviewers. These in turn had implications for data quality – including the possibility of interviewer malpractice – interview recruitment, training, and payment. The possible role of 'interviewer effects,' the extent to which the status or demeanor of the interviewer might affect the responses obtained in the interview, was a particular concern. Converse suggests that pollsters retained relatively cordial links with sociologists and political scientists through the 1930s and into the 1940s. Disagreements with social psychologists remained, however, and intensified when the advent of war brought with it increased competition for funds, prestige, and institutional advantage.

By the 1930s, governments in many countries had begun to regard morale as an important element in the planning and prosecution of war. In the United States, when war came, a number of government agencies were set up to study military and civilian morale. Many leading figures from the world of survey research – Cantril, Lazarsfeld, Likert, Stouffer, and others – contributed to this effort. The exigencies of wartime work left little space for methodological development. What did emerge out of the wartime experience was a degree of institutional crystallization in the postwar period that increasingly saw survey research located within university-affiliated, externally funded research centers and institutes. Bodies such as the Bureau of Applied Social Research at Columbia, the National Opinion Research Center at Chicago, and the Survey Research Center at Michigan all pursued work on methodological issues, which was funded by research contracts focused on substantive topics usually of an applied nature. The extent to which experimental variations in interview practice could be built into survey designs equipped researchers with a powerful tool for studying their own methods. In the later part of the 1940s and the early 1950s, researchers at the National Opinion Research Center (NORC) conducted a series of experiments that overturned much of the conventional wisdom of the time concerning sources of interviewer bias and the role of 'rapport' in the interview (Hyman, 1954), implicitly encouraging a focus on the instrumental and task-oriented elements of the interviewer's role.

After the war, research on the use of surveys increased in a number of social science disciplines. This shift was particularly marked in sociology, but can also be seen in economics and political science, if rather less so in social psychology (Presser, 1984). Except for economics, as Presser shows, this general trend reflects the growing importance American social scientists placed on self-report measures of attitudes, beliefs, values, and other subjective phenomena. As the survey became dominant, attention turned to topics such as sampling, scaling, and questionnaire design (Platt, 2001). Discussion of survey-oriented topics such as interviewer effects and the selection, training, and supervision of interviewers also became noticeable at this time.

Unstructured Interviewing

Known early on as the 'personal' interview or 'verbatim' interview; later as the 'depth,' 'in-depth,' or 'intensive' interview; and thereafter as the 'unstructured' or 'semistructured' interview, an open, exploratory style of interviewing has often been seen as

an alternative to the more heavily standardized survey interview. In sociology, interviewing of this kind has often been associated with the work of the Chicago tradition. Initially, however, the interview was somewhat disfavored by Chicago sociologists, who preferred instead the written life history and 'undesigned materials' (Thomas, 1912), letters, diaries, memoirs, and the like. Sociologists in this period did use information from interviews conducted by social workers (Platt, 1996). Given that social work was regarded at the time as a feminized, somewhat low-status activity, it is possible that direct use of the method by sociologists was somewhat inhibited as a result. When Chicago-trained sociologists did interview, they tended to base their understanding of the procedure on literary genres derived from the life history. In consequence, they often relied on rather mechanistic strategies for invoking responses from the interviewee rather than focusing on the interactional dynamics of the interview (Lee, 2008). Ironically, it was social workers rather than sociologists who produced in the 1920s a relatively comprehensive and sophisticated approach to interviewing heavily informed by contemporary progressivist assumptions about the value of professional expertise and social scientific knowledge. This work, now largely forgotten, explored in a relatively systematic way issues related to interview dynamics, the functions of interviews, the role of the interviewer, and methods for recording and analysis (see, e.g., Queen, 1928).

In the interwar period, Carl Rogers, partly influenced by the Freudian Otto Rank, drew on his experience of working in child guidance to develop what he described as a 'nondirective' approach to counseling. Applied to research interviewing, nondirective interviewing allowed the respondent, rather than the interviewer, to assign significance to the topics covered and used a variety of strategies to signal a nonjudgmental understanding of the respondent's thinking, and the need as the interview proceeded for further elaboration or self-reflection. A similar style of interviewing emerged out of the Hawthorne studies, based this time on principles derived from the Australian psychologist Elton Mayo's reinterpretation of Jean Piaget's work (Lee, 2011).

Nondirective interviewing became widely used in the social sciences. The Hawthorne studies were central to the newly burgeoning field of human relations, and to related fields like consumer research, and organizational studies. In the late 1930s, the social psychologist Rensis Likert, who directed the Division of Program Surveys, a research unit in the US Department of Agriculture strongly oriented to the use of an open, conversational style of interviewing, invited Rogers to train the division's interviewers in nondirective methods, thus helping to widen the visibility of the method. In time, though, the popularity of nondirective interviewing waned. The method drew criticism from those who saw it as overly bland or who felt that interviewers should be free to home in on topics of specific interest, or to probe silence on the part of the respondent more aggressively. In sociology in particular, as the force of these criticisms began to be felt, attention shifted, it can be argued, toward other methods, such as participant observation, at the expense of interviewing methods (Lee, 2011).

Soon after the United States entered the war in 1941, the sociologist Robert Merton became involved in a study in which groups of soldiers were interviewed about the effects of an enemy propaganda film that they had just watched. 'Focused interviewing' (Merton and Kendall, 1946), the method that evolved out of this study, diffused soon after into marketing and consumer research. Subsequently known as the 'focus group,' the method became popular in the 1960s and the 1970s. Its relative cost-effectiveness encouraged take-up of the focus group as a method of consumer research as did, somewhat surprisingly, the invention of sales tracking technologies such as the barcode, which encouraged companies to shift from monitoring market share using unreliable recall-based surveys to a methodology that allowed exploration of the reasons behind consumer decisions (Lee, 2010). The use of focus groups subsequently diffused back into sociology and became very widely used across the social sciences in many different countries.

The History of the Survey Interview II

There have been three key postwar trends affecting interview practice in survey research: the internationalization of the survey method, the proliferation of survey modes, and the turn toward cognitive methods. The internationalization of survey research began in the late 1930s with George Gallup using franchising arrangements to set up Gallup affiliates in a number of countries. After the war, polls and surveys came into common use in many Western industrialized countries. Survey methods diffused through personal contacts and academic exchanges, the growing availability of conferences, summer schools and courses, commercial arrangements, and the encouragement of international organizations (Bulmer, 1998). Receptivity to survey-based methods depended often on cultural and political factors. For example, in France, the survey interview was not always well regarded, with some (particularly elite) social groups seeing it as a rather alien cultural form (Riesman, 1956). In the countries of the then Eastern bloc and other parts of the Communist world, survey research met with official disapproval, although this could be eased during periods of political uncertainty or change when party officials wished to monitor popular opinion. In these contexts, though, individuals could be anxious about agreeing to participate in an interview and about how to respond to politically sensitive questions (Henn, 1998).

The variety of means and media through which survey responses are elicited has expanded greatly over the years. The traditional methods of face-to-face interviewing and self-completion questionnaires were joined in the 1970s by telephone interviewing. Subsequently, a variety of computer- and Internet-mediated methods have come to the fore, and the use of mixed-mode surveys has increased (Couper, 2011). Modes differ in their implications for the role and purpose of interviewers. How these will change in the future as technologies develop remains unclear.

Beginning in the late 1970s, collaboration began to emerge between cognitive psychologists and survey researchers. In some ways, a reinvigoration of the historic links between survey research and social psychology, much of this work focused initially on question design (Jobe and Mingay, 1991) Somewhat later, something of a symbiotic relationship also developed between survey researchers and sociologists working in

the ethnomethodological tradition (see, e.g., Maynard and Schaeffer, 2000). The focus of this work – how in question–answer exchanges meaning emerges out of interaction – has obvious implications for understanding the interview.

Interviewers

If interviewing was originally a gentlemanly pursuit, by the middle of the twentieth century, survey interviewing had become a largely feminized occupation. Although historically, survey organizations differed in the extent to which they recruited by age and educational level (Converse, 1987), they retained a preference for women as interviewers, which rested on the then conventional gendered assumptions about levels of pay, household roles, and emotional labor. Research based on samples of interviewers working on the multicountry European Social Survey suggests that today, with some exceptions, men are well represented on interviewing field forces (Blom et al., 2011). Work conditions have implications for how workers perform their role, whether one sees this as 'cheating' or as an accommodative response to the demands of the job. How interviewers conceive of their role, their attitudes toward it, and the operational routines and practices they adopt have begun to be incorporated in a relatively systematic way into work on interviewer effects (see, e.g., Durrant et al., 2010). There still remains, however, a dearth of work that looks at interviewing as an occupational role, and training contexts aside (Converse and Schuman, 1974), how interviewers themselves experience their role has received relatively little attention, a reflection perhaps of their relatively low status (Peneff, 1988).

Special Populations

Interviewing, particularly survey interviewing, has tended to be directed toward members of the general population. Particular kinds of respondent – children, the powerful, members of stigmatized groups, and so on – often pose challenges that lie beyond the standard repertoire of interviewing approaches. Work in these fields has tended to produce a set of specialized literatures. Although, obviously, specific groups throw up specific issues, there are a number of key concerns that crosscut these literatures whatever the character of the group involved. Among these concerns one can note (1) the relative social positions of the interviewer and the interviewee and how these might reflect disparities of power and status, (2) the respective identities of the interviewer and the interviewee, (3) the importance of avoiding stereotypical conceptions as well as an alertness to possible unconscious assumptions, and (4) the possibility of cross-cultural misunderstandings. The (possibly problematic) role in the interview of self-disclosure by the interviewer has also been a concern with some groups.

Critical Voices

Voices were raised in the 1960s against the dominance of the interview and other self-report methods in the social sciences (see, e.g., Webb et al., 1966). More broadly, in many disciplines, radical and critical positions of various kinds emerged to challenge what was seen as a dominant orthodoxy characterized as positivistic, scientistic, and politically quiescent. In all of this the interview was not abandoned. How as a method it was to be used, however, was seen to require a fundamental rethinking of the dynamics of interviewing and a radical reconfiguration of interview practice. Some writers argued for a focus on powerful actors and institutions and advocated a robust, interrogative style of interviewing taking investigative journalism as a model (on the development of 'adversarial' research strategies, see Lee, 1993). In France, Touraine developed an approach, morphologically similar to the focus group, but embedded within politically engaged processes for social intervention and consciousness raising (Brincker and Gunderlach, 2005). More commonly, the interview was to be seen not as a more or less fleeting instrumental and extractive encounter governed by the researcher's need for information but as a shared interactional accomplishment open to the biographical, contextual, and relational underpinnings of people's lives, and in which the voice of the interviewee was to be afforded clear and unhindered expression. In work of this kind, the communicative aspects of the interview came under detailed, indeed minute, scrutiny with the focus on how, through the course the interview, both interviewer and interviewee formulate and frame questions and responses out of their reciprocal understandings of the meanings situationally produced within the interview itself. This opened to view discursive features of the interview such as the role within it of narrative, paralinguistic communication or the grounds on which cultural misunderstandings emerge, topics usually treated as invisible in conventional treatments or relegated to the status of technical 'troubles.'

Feminist writers, in particular, produced a widely influential strand of critical writing on the interview. The aim here was to overcome the traditional invisibility of women in social research, and to explicate how gendered patterns of power, domination, and disadvantage were implicated in mainstream interviewing practice. Traditional interviewing styles with their closed questions and rigid prespecification of topics tightly framed by the researcher's interests were seen as objectifying female experience and muting women's self-expression. Instead, feminist researchers emphasized as an ethical commitment the importance of establishing nonhierarchical forms of relationship between the researcher and the researched and of giving voice to those they studied. Both these aims were seen to be met more readily by qualitative interviewing methods than by other approaches. Despite broadly shared principles and commitments, debates have emerged within feminist interviewing practice about, for example, the place of self-disclosure by the interviewer, boundaries appropriate to the interview relationship and its aftermath, and the effect on the interview of the relative social positions of the interviewer and the interviewee (Reinharz and Chase, 2001).

A decidedly marginal interest until recently, the use of 'biographical methods' – a term covering a wide range of approaches, including life history and narrative methods, biography and autobiography, and oral history methods – has blossomed in the early part of the twenty-first century. All these approaches have an emphasis on the understanding of individual lives, lived out within biographical, historical, and

social contexts. Some work in the field has been driven by an explicit commitment to interdisciplinarity (Bornat, 2008). For example, in a number of European countries, a degree of openness to psychoanalytic thinking has been a feature of research focused on the interpretive understanding of individual biographies, often in relation to policy-relevant areas like migration and health care (see, e.g., Chamberlayne et al., 2000). At the same time, a degree of compartmentalization, often based on an adherence to disciplinary allegiances, can also be seen within the field (Bornat, 2008).

The biographical turn might be seen more broadly as part of a cultural trend emerging after World War II in which 'testimony' and 'witnessing' were seen to have positive value. This trend, Thomson (2007) suggests, emerged in part because the recounting of suffering, indignity, and loss provided a way for victims to respond individually and collectively to the violent traumas and upheavals of the twentieth century. Witnessing came to be regarded, too, as providing a means by which those silenced by dominant social and cultural arrangements could seek repair and redress. Since that silencing is seen to inhere in the assumptions and practices associated with dominant research traditions within Western colonialism and imperialism, an important development has been the emergence of 'indigenous methodologies' (Smith, 2012). This notion emphasizes the necessity for research to advance the needs, aspirations, and cultural integrity of colonized peoples, as defined and articulated by those peoples themselves. Typically, quantitative methods have been rejected in favor of qualitative or participative approaches, but even here, conventional methods such as interviewing have been rethought, for example, by aligning information gathering to indigenous discursive practices such as storytelling or deliberative discussion (see, e.g., Bishop, 2012).

In recent years, a rather sophisticated view of the interview has emerged from writers interested in recasting the method in the light of postmodernism. Postmodernist treatments of the interview draw on and build upon existing critiques of interview practice found in feminism, discourse analysis, sociolinguistics, and the like, but go beyond them in various ways. There has been a degree of interest, for example, in blurring the conventional boundaries between genres, with interview transcripts becoming vehicles for representational experiment through the use of poetry or performance art. Writers have also experimented with 'autoethnography,' an approach that involves a highly collaborative interviewing style with considerable attention paid to the emotions evoked for the interviewer and interviewee by the interview process itself (for a summary, see Fontana, 2001).

The Social Acceptability of the Interview

It is difficult to disentangle popular attitudes to the interview as a social encounter between the interviewer and the interviewee from wider public understandings of polling, sampling, and social science research in general. If the interview could seem, particularly to the middle classes, like an intrusion or an imposition in the 1920s, it came to be regarded in subsequent decades with varying degrees of curiosity, acceptance, or resistance (Igo, 2008). Survey participation seems to have peaked in the 1960s and researchers have fretted about declining response rates ever since. What component of this decline is due to changing cultural understandings of the interview, as opposed to other factors, is difficult to assess. There are signs, too, that more qualitatively oriented approaches have sometimes faced resistance from disadvantaged, marginalized, and exploited groups. While this clearly prompted a reorientation of interview practice across a number of fields, it might also have encouraged a shift in research focus toward groups amenable to the interview as a method (Manning, 2006). The decline in survey response rates has encouraged research on why people participate in survey interviews (see, e.g., Groves and Couper, 2012). However, there remains comparatively little work on participation in research using qualitative interviews or on how those who do participate regard the experience of being interviewed. This issue has relevance particularly in countries that require prior ethical review of research projects. Review boards often consider intensive interviews on 'sensitive' topics to be ethically problematic (Buckle et al., 2010). The emergent character of such interviews can be seen as complicating the process of obtaining informed consent, especially so where research participants are regarded as being 'vulnerable' in some way. Furthermore, the possibility that interviews might cause emotional distress to interviewees has also been a concern. As well as describing strategies that researchers can use to address such concerns, Buckle et al. (2010) point to evidence suggesting that, in fact, research interviews, even on difficult and emotionally charged topics, are often regarded positively by interviewees.

Conclusion

Atkinson and Silverman have spoken of the extent to which people today live within an 'interview society.' In such a society, the self becomes 'a proper object of narration' (1997: p. 315). Everyone is seen as having discernible views, understandings, and feelings about their own experiences. It is assumed, on the one hand, that individuals can articulate these views, understandings, and feelings, and, on the other, that it is reasonable for information about them to be elicited through direct questioning. In consequence, the interview has become ubiquitous as a method for obtaining personal information especially within professional–client relationships, while the interview also provides the mass media with a vehicle for blurring of the boundaries between public and private. In this context, it is tempting to see the interview remaining both as a pervasive aspect of life in modern societies and as a central tool of social research. One might also suppose that the methodological future of interviewing as a method resides in incremental technical change and/or innovation that remains local to particular specialisms or disciplines. The major challenge to this scenario lies, of course, in the potential impact on interview practice of the revolution in new communication technologies.

The development of the interview has always been entwined with technological change. Originally, researchers had relied on reconstructing interviews from memory, unobtrusively jotted notes, or the use of standard or private shorthands to record interviews. In some contexts, the use of stenographers was common (Lee, 2004). Carl Rogers pioneered the phonographic

recording of unstructured interviews during the 1940s. After World War II, the availability of relatively portable magnetic recorders encouraged the use of audio recording in field settings. Technological developments, such as transistorization and miniaturization (and much more recently digital recording), have allowed improvements both in audio quality and the portability of recording devices to the extent that the recording of unstructured interviews became ubiquitous in the social sciences (Lee, 2004). Although detailed case studies of adoption are rare, and the pace of technologically induced change has not always been uniform (Schober and Conrad, 2008), the development of the survey interview has also plainly been shaped, directly and indirectly, by the various affordances inherent in data processing technologies, including Hollerith machines, mainframe computers, and statistical software packages, and data acquisition technologies such as the telephone and, more recently, portable digital devices.

New communication technologies pose questions about how the dynamics of the interview itself might change in technologically mediated contexts. Advances in video-based technology raise, for example, the question of what it means to talk of 'rapport' when the respondent is virtually, but not physically, copresent (Fielding, 2010). How might the long-standing preference for standardized interview questions on surveys be modified by automated systems that use intelligent, agent-driven, interviewing techniques (Schober and Conrad, 2008)? As Schober and Conrad point out, these and other questions challenge researchers to find ways of harnessing the possibilities produced by new communication technologies without allowing those possibilities to impose themselves on the interview process in unwelcome ways. Beyond this, contemporary shifts in the use of social media, the growing ubiquity of geotagged information, the development of data fusion techniques, and the availability of large-scale commercially generated data sets all potentially have implications for the place of the interview itself within the repertoire of social research (for a provocative view of what these implications might be, see, e.g., Savage and Burrows, 2007). The interview is unlikely to disappear as a basic method in the social sciences. It is highly likely, however, that its nature, scope, use, and relationship to other methods will be radically transformed in the years to come.

See also: Attitude: History of Concept; Empirical Social Research, History of; Sample Survey Methodology, History of.

Bibliography

Atkinson, P., Silverman, D., 1997. Kundera's immortality: the interview society and the invention of the self. Qualitative Inquiry 3, 304–325.

Bishop, R., 2012. Indigenous methods in qualitative educational research. In: Delamont, S. (Ed.), Handbook of Qualitative Research in Education. Edward Elgar Publishing, Cheltenham, pp. 126–140.

Blom, A.G., de Leeuw, E.D., Hox, J.J., 2011. Interviewer effects on nonresponse in the European Social Survey. Journal of Official Statistics 27, 359–377.

Bornat, J., 2008. Biographical methods. In: Alasuutari, P., Bickman, L., Brannen, J. (Eds.), The Sage Handbook of Social Research Methods. Sage Publications Ltd., London, pp. 344–356.

Brincker, B., Gundelach, P., 2005. Sociologists in action: a critical exploration of the intervention method. Acta Sociologica 48, 365–375.

Buckle, J.L., Dwyer, S.C., Jackson, M., 2010. Qualitative bereavement research: incongruity between the perspectives of participants and research ethics boards. International Journal of Social Research Methodology 13, 111–125.

Bulmer, M., 1998. Introduction: the problem of exporting social survey research. American Behavioral Scientist 42, 153–167.

Bulmer, M., Bales, K., Sklar, K.K., 1991. The social survey in historical perspective: 1880–1940. In: Bulmer, M., Bales, K., Sklar, K.K. (Eds.), The Social Survey in Historical Perspective: 1880–1940. Cambridge University Press, Cambridge, MA, pp. 1–48.

Chamberlayne, P., Bornat, J., Wengraf, T., 2000. Introduction: the biographical turn. In: The Turn to Biographical Methods in Social Science: Comparative Issues and Examples, pp. 1–30.

Converse, J.M., 1987. Survey Research in the United States: Roots and Emergence 1890–1960. University of California Press, Berkeley, CA.

Converse, J.M., Schuman, H., 1974. Conversations at Random: Survey Research as Interviewers See It. Wiley, New York.

Couper, M.P., 2011. The future of modes of data collection. Public Opinion Quarterly 75, 889–908.

Durrant, G.B., Groves, R.M., Staetsky, L., Steele, F., 2010. Effects of interviewer attitudes and behaviors on refusal in household surveys. Public Opinion Quarterly 74, 1–36.

Fielding, N., 2010. Virtual fieldwork using access grid. Field Methods 22, 195–216.

Fontana, A., 2001. Postmodern trends in interviewing. In: Gubrium, J.F., Holstein, J.A. (Eds.), Handbook of Interview Research: Context and Method. Sage Publications Inc., Thousand Oaks, CA, pp. 161–175.

Groves, R.M., Couper, M.P., 2012. Nonresponse in Household Interview Surveys. John Wiley & Sons, New York.

Henn, M., 1998. Opinion polling in central and eastern Europe under communism. Journal of Contemporary History 33, 229–240.

Hyman, H.H., 1954. Interviewing in Social Research. University of Chicago Press, Chicago.

Igo, S.E., 2008. The Averaged American: Surveys, Citizens, and the Making of a Mass Public. Harvard University Press, Cambridge, MA.

Jobe, J.B., Mingay, D.J., 1991. Cognition and survey measurement: history and overview. Applied Cognitive Psychology 5, 175–192.

Lazarsfeld, P.F., Oberschall, A.R., 1965. Max Weber and empirical social research. American Sociological Review 30, 185–199.

Lee, R.M., 1993. Doing Research on Sensitive Topics. Sage Publications Ltd, London.

Lee, R.M., 2004. Recording technologies and the interview in sociology, 1920–2000. Sociology: Journal of the British Sociological Association 38, 869–889.

Lee, R.M., 2008. Emory Bogardus and the new social research. Current Sociology 56, 307.

Lee, R.M., 2010. The secret life of focus groups: Robert Merton and the diffusion of a research method. The American Sociologist 41, 115–141.

Lee, R.M., 2011. "The most important technique…": Carl Rogers, Hawthorne, and the rise and fall of nondirective interviewing in sociology. Journal of the History of the Behavioral Sciences 47, 123–146.

Manning, P.K., 2006. Fieldwork in transition. In: Hobbs, D., Wright, R. (Eds.), The Sage Handbook of Fieldwork. Sage Publications Ltd., London, pp. 361–376.

Marsh, C., 1985. Informants, respondents and citizens. In: Bulmer, M. (Ed.), Essays on the History of British Sociological Research. Cambridge University Press, Cambridge, pp. 206–227.

Maynard, D.W., Schaeffer, N.C., 2000. Toward a sociology of social scientific knowledge: survey research and ethnomethodology's asymmetric alternates. Social Studies of Science 30, 323–370.

Merton, R.K., Kendall, P.L., 1946. The focused interview. American Journal of Sociology 51, 541–557.

Peneff, J., 1988. The observers observed: French survey researchers at work. Social Problems 35, 520–535.

Platt, J., 1996. A History of Sociological Research Methods in America, 1920–1960. Cambridge University Press, Cambridge.

Platt, J., 2001. The history of the interview. In: Gubrium, J.F., Holstein, J.A. (Eds.), Handbook of Interview Research: Context and Methods. Sage, Thousand Oaks, CA, pp. 33–54.

Presser, S., 1984. The use of survey data. In: Turner, C.F., Martin, E. (Eds.), Surveying Subjective Phenomena, vol. 2. Russell Sage Foundation, pp. 93–114.

Queen, S.A., 1928. Social interaction in the interview. Social Forces 6, 545–557.

Reinharz, S., Chase, S.E., 2001. Interviewing women. In: Gubrium, J.F., Holstein, J.A. (Eds.), Handbook of Interview Research: Context and Methods. Sage, Thousand Oaks, CA, pp. 220–239.

Riesman, D., 1956. Orbits of tolerance, interviewers, and elites. The Public Opinion Quarterly 20, 49–73.

Riesman, D., Benney, M., 1956. The sociology of the interview. Midwestern Sociologist 18, 3–15.
Savage, M., Burrows, R., 2007. The coming crisis of empirical sociology. Sociology 41, 885–899.
Schober, M.F., Conrad, F.G., 2008. Survey interviews and new communication technologies. In: Conrad, F.G., Schober, M.F. (Eds.), Envisioning the Survey Interview of the Future. John Wiley & Sons, Inc., Hoboken, NJ.
Schudson, M., 1994. Question authority: a history of the news interview in American journalism, 1860s–1930s. Media Culture and Society 16, 565–587.
Sklar, K.K., 1991. Hull-House maps and papers: social science as women's work in the 1890s. In: Bulmer, M., Bales, K., Sklar, K.K. (Eds.), The Social Survey in Historical Perspective: 1880–1940. Cambridge University Press, Cambridge, MA, pp. 111–147.
Smith, L.T., 2012. Decolonizing Methodologies: Research and Indigenous Peoples, second ed. Zed Books, London.
Thomas, W.I., 1912. Race psychology: standpoint and questionnaire, with particular reference to the immigrant and the Negro. American Journal of Sociology 17, 725–775.
Thompson, P., 2000. In: Voice of the Past: Oral History, third ed. Oxford University Press, Oxford.
Thomson, A., 2007. Four paradigm transformations in oral history. The Oral History Review 34, 49–70.
Webb, E.J., Campbell, D.T., Schwartz, R.D., Sechrest, L., 1966. Unobtrusive Measures: Nonreactive Research in the Social Sciences. Rand McNally, Chicago, IL.

Intimate Partner Abuse, Applied Research on

Jordana N Navarro, Cookeville, Tennessee Technological University, TN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although recognition of intimate partner abuse (IPA) did not begin until the 1970s, the gravity of the social problem is undisputed. In contrast to other fields, studying family violence has always had an implicit applied component, because research is often conducted to improve prevention and treatment programs. As a result of this applied focus, programs have evolved from an emphasis on the individual to recognition of systemic factors of abuse. This article reviews applied research on IPA that has contributed to the development of evidence-based programs combatting the problem.

Introduction

Research continues to indicate that intimate partner abuse (IPA) is a serious *global* social problem (Bacchus et al., 2010; Gadomski et al., 2001; García-Moreno et al., 2005; Gover et al., 2003; Rolling and Brosi, 2010). (This article focuses exclusively on research addressing IPA. The term 'intimate partner abuse' is used in recognition of the fact that batterers use a variety of tactics to maintain power and coercive control on victim/survivors. However, when referring to the research field in general the term 'family violence' is used.) Approximately 25% of women have experienced IPA at some point in their lifetime (Saltzman and Houry, 2009). (The debate regarding gender symmetry in IPA is ongoing and outside the scope of this article. Thus, for the purposes of this article, information will focus on male-to-female perpetrated IPA unless specifically stated otherwise.) In contrast to men, women are at greater risk of being beaten, killed, raped, or stalked by an intimate partner compared to an unknown person (Mahoney et al., 2001; Tjaden and Thoennes, 2000). As research into IPA has grown throughout the years, the definition of 'abuse' has broadened beyond physical, emotional, and sexual abuse (El-Mouelhy, 2004; Nicolaidis and Paranjape, 2009; Saltzman and Houry, 2009) to include cyberstalking (U.S. Department of Justice, 2001) and economic abuse (Corso, 2009). These advancements in the understanding of IPA have subsequently led to stronger prevention and intervention programs.

In contrast to other fields of study, research on family violence has always had an inherent applied mandate as scholars typically search for knowledge to contribute to prevention and intervention efforts combatting IPA (Gelles, 1985). The goal of these efforts is to promote evidence-based violence prevention and intervention practices. In this article, the current state of knowledge – stemming from applied research on IPA – will be reviewed. In addition, important terminology, guiding theoretical frameworks, and future research directions are discussed. To begin, a review of the definitional differences between the terms 'prevention' and 'intervention' is warranted.

Definitions

Although the terms 'prevention' and 'intervention' are often used interchangeably by laypersons, the fundamental difference between the two is important to discuss. According to Hamby (1998), prevention programs are designed to prevent violence and typically are classified into three main categories: primary, secondary, and tertiary prevention. Primary prevention programs are designed to prevent abuse before onset (Andrews, 1994; Hamby, 1998; Navarro, 2013; Smithey and Straus, 2004). In order to be successful, primary prevention programs typically attempt to foster a concurrent change in environment and increase education (Andrews, 1994; Smithey and Straus, 2004).

Secondary prevention, in contrast to primary prevention, are efforts targeted at a specific group of individuals at risk to experience the problem (Hamby, 1998). However, primary and secondary preventions are similar in that both are designed to prevent violence before onset (Hamby, 1998). Finally, tertiary prevention is very similar to intervention: services are targeted at those already experiencing the problem with the goal of preventing subsequent occurrences (Hamby, 1998). Although some programs fall neatly within one category, the lines between these programs are not impermeable (Hamby, 1998). In addition to understanding the various types of prevention and intervention models, it is equally important to understand the theoretical perspectives that often drive program development.

Theoretical Perspectives

Scholars have advocated that effective programs are those grounded in scientific theories of how problems develop and persist (Harvey et al., 2007). Utilizing theoretical perspectives in the development of prevention and intervention programs not only provides a framework for effectively planning and efficiently delivering services to target populations, but also assists scholars conducting program evaluations to discern impact. Although several theories are applicable in the development of prevention and intervention programs (see Smithey and Straus, 2004), this article will focus on two of the most widely used perspectives: the socioecological framework and deterrence theory.

According to Andrews (1994), a substantial amount of family violence prevention and intervention programs are 'person-centered' or focused on the victim/survivor or perpetrator. However, as research has called attention to the structural factors leading to IPA, the utilization of multidimensional perspectives has increased. One of the most widely referenced multidimensional frameworks utilized in the development of

prevention programs – including programs developed by the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO) – is Bronfenbrenner's (1979) socioecological perspective.

As Bronfenbrenner (1979) theorized, individuals are constantly acted upon and shaped by their surrounding environments, which is comprised of the microsystem (e.g., the individual), the mesosystem (e.g., the family), the exosystem (e.g., the community), and the macrosystem (e.g., larger culture). Since its conception, the socioecological perspective has been utilized to guide prevention programming combatting IPA both within the United States (Navarro, 2013) and internationally (Chan et al., 2009). Typically, programs guided by the socioecological perspective engage in some variation of the following activities: (1) at the individual level, programs may target the effects of witnessing marital violence and stimulant use; (2) at the family level, programs may target male dominance and economic dependence on men; (3) at the community level, programs may target increasing the social support in the surrounding community; and (4) at the societal level, programs may target changing cultural norms that support marital violence (Harvey et al., 2007; Heise, 1998; Mitchell and Vanya, 2009).

As illustrated, the socioecological framework is especially useful, because it considers the interplay among personal, situational, and sociocultural factors (Heise, 1998). As a result, the socioecological perspective incorporates the tenets of several theoretical frameworks into one holistic model. For example, by acknowledging the importance of witnessing marital violence in the understanding of IPA perpetration, the socioecological model incorporates tenets from social learning theory. Moreover, by acknowledging the importance of cultural norms supportive of male privilege in the understanding of IPA perpetration, the socioecological model incorporates tenets from patriarchal theory. Thus, the socioecological framework draws on the strengths of many perspectives to provide a broad understanding of IPA. Aside from this framework, another widely used theoretical perspective – which also serves as the backbone of the criminal justice system – is deterrence theory.

The premise of deterrence theory can be traced back to Durkheim, who assumed that the negative consequences of punishment curbed individuals from engaging in criminal behavior – especially in instances where punishment was certain, swift, and severe (Sherman and Berk, 1984; Pogarsky, 2009). The first type of deterrence, referred to as specific deterrence, describes the process of an offender avoiding engagement in subsequent illegal activities (e.g., IPA) due to the perceived costs of such behaviors (Smithey and Straus, 2004). Because the population experiencing this includes offenders, or those experiencing the problem, programs developed using the tenets of specific deterrence are most akin to tertiary prevention and intervention. In contrast, general deterrence describes the process in which individuals (in general) abstain from engaging in crime (e.g., IPA) in order to avoid receiving sanctions (Smithey and Straus, 2004). Because the population experiencing this includes the entire community, general deterrence is most akin to primary prevention. Perhaps the most famous study applying the ideas of deterrence theory to the prevention and intervention of IPA is the Minneapolis arrest field experiment (Maxwell et al., 2002; Sherman and Berk, 1984) discussed in the following section.

Current State of Knowledge

As interest in family violence has grown throughout the years, scholars have highlighted the multiple causes of IPA – each of which provides the opportunity for prevention and intervention (Smithey and Straus, 2004). This section reviews the current state of knowledge regarding the many advancements in prevention and intervention that have occurred throughout the years, which have been empirically supported through applied research. In order to structure this section, information is organized by target populations if initiatives focused on a specific group. Following that review, the last section notes findings from comprehensive initiatives that include many components usually targeting whole communities.

Abusers

Perhaps the most famous study evaluating an intervention designed to reduce IPA is the 1984 Minneapolis arrest experiment (Sherman and Berk, 1984). In this seminal work, Sherman and Berk (1984) evaluated the impact of three law enforcement responses on batterers' recidivism (e.g., arrest, informal counseling, and ordering the batterer to vacate the premises for 8 h). At the conclusion of their study, data indicated that arrested offenders recidivated less than offenders ordered to leave the premises (Sherman and Berk, 1984). As a result, sweeping reforms occurred that led to mandatory arrest policies, despite that other scholars encountered considerable difficulty in replicating these results or replicated studies produced contradictory findings (Maxwell et al., 2002).

Acknowledging these discrepancies, Maxwell et al. (2002) recently synthesized data generated through various replication attempts in order to readdress whether arrest acted as a specific deterrent for IPA offenders. In their study, Maxwell et al. (2002) found a consistent reduction in offenders' subsequent violence following arrest. As a result of these data, the scholars noted the original findings of the Minneapolis arrest experiment were supported despite concerns from other scholars that mandatory arrest may not always be in the best interest of the victim/survivor. Aside from the debate regarding mandatory arrest, another area of inconsistent research is the impact of batterer intervention programs (BIPs).

According to Hamilton et al. (2013), three types of BIP models dominate today: the cognitive-behavioral, the psychodynamic, and the profeminist treatment (i.e., the Duluth model). Although BIP curriculum slightly varies, all focus on the behavior of the abuser and encouraging the batterer to accept responsibility for that behavior (Gondolf, 2011). Moreover, BIPs tend to focus on promoting healthy relationships, while also addressing underlying beliefs that relate to IPA (Gondolf, 2011). Unfortunately, while the aforementioned objectives are promising, only a handful of evaluations have been conducted on these programs to date – all of which indicate there is little to no lasting effect in curbing subsequent IPA (Babcock et al., 2004; Gondolf, 2011). In recognition of

this lack of impact, Gondolf (2011) evaluated several proposed ideas of how to improve these programs (e.g., treat attachment issues associated with battering, diversifying programs to account for different batterer types, including culturally sensitive material, etc.). However, after reviewing available evidence – including program evaluations – Gondolf (2011) notes there was little evidence to support these alternatives. As a result, an effective intervention program that successfully reduces the likelihood of subsequent violence remains an elusive goal.

Businesses

Since the mid-1990s, family violence scholars have increasingly drawn attention to the important role businesses have in the prevention and intervention of IPA (Murray and Powell, 2008; Urban and Bennett, 1999). The role of businesses in preventing IPA, or intervening when confronted with signs of IPA in workplace, is vital for a few reasons. First (and most importantly), victim/survivors spend considerable time at work; as a result, problems occurring at home are brought into the workplace. For example, according to a CDC report (2003), costs stemming from IPA exceeded $5 billion in the United States, which includes approximately $1.8 billion in the indirect costs of lost productivity. Second, aside from the monetary costs, an estimated 13 000 acts of violence occur in workplaces within the United States each year (Bachman, 1994) obviously placing victim/survivors at risk as well as fellow employees. Similar consequences have been documented internationally. For example, Henderson (cited in Murray and Powell, 2008) notes businesses in Australia lose more than $1.5 billion annually as a result of IPA. Due to these consequences, many businesses have now recognized their role in the prevention and intervention of IPA, including large employers such as Polaroid, Liz Claiborne Inc., and Marshalls, Inc. (Isaac, 1997).

The three main types of IPA prevention and intervention models utilized in workplaces include employer-led programs, brokerage partnerships, and union-based initiatives – many of which have undergone formative evaluations (Clarke, 1999 cited in Murray and Powell, 2008). In employer-led programs, IPA is addressed by incorporating an awareness of the problem into existing human resources structures and organizational processes (Murray and Powell, 2008). The aforementioned includes training fellow employees and managers about the dynamics of abuse in order to increase awareness of, and sensitivity to, signs of IPA in the workplace (Murray and Powell, 2008). In brokerage partnerships, employers partner with IPA social service agencies and these organizations engage in activities to raise awareness and promote understanding within the workplace (Murray and Powell, 2008). Finally, in union-based initiatives, group leaders organize to promote positive change for workers; for example, developing IPA resources to assist in prevention efforts (Murray and Powell, 2008).

According to Urban and Bennett (1999), although the goals of employees and employers sometimes differ, both groups share the common objective of preventing and reducing IPA in the workplace. In recognition of this common goal, Urban and Bennett (1999) guided the development, implementation, and evaluation of the Domestic Abuse Program (DAP) (not the actual name of the program (see Urban and Bennett, 1999)). Implemented in a plant in a rural area of Tennessee, initial program evaluation findings of DAP indicate trainings provided as part of the initiative were successful (Urban and Bennett, 1999). There were statistically significant differences among DAP trainees in three out of seven IPA concepts from pretest to a year after training (Urban and Bennett, 1999). In addition, there was a marked decrease in the projected utilization of negative and controlling behaviors among trainees following DAP (Urban and Bennett, 1999). As a result of these initial findings, Urban and Bennett (1999) note that DAP created an environment in which victim/survivors could be supported by their colleagues.

Although the aforementioned describes and illustrates the types of prevention and intervention programs currently utilized in the workplace, more work is necessary as scholars have found that employers sometimes react to IPA by punishing victim/survivors with poor evaluations or engaging in other tactics to justify their terminations (Park, 2003). These negative reactions are unfortunate, because research has found that promoting an open and supportive work environment benefits employees and employers alike. For example, Swanberg and Logan (2005) found that when employees experiencing IPA disclosed abuse to supportive employers, both benefitted. Employed victim/survivors were grateful, became increasingly loyal, and demonstrated increased motivation to excel in their duties (Swanberg and Logan, 2005).

Despite the noted advancements, a dearth of information exists on program evaluations of IPA prevention and intervention initiatives within the workplace (Urban and Bennett, 1999). Moreover, if information about workplace prevention and intervention strategies do exist, documents discussing these initiatives are rarely made available (Murray and Powell, 2008). As a result, scholars should endeavor to address this gap in knowledge moving forward by emphasizing the important benefits of program evaluations – chief of which is to ensure that initiatives were implemented properly (Murray and Powell, 2008). However, the advancements that have been made thus far further underscore the point that IPA is not a private matter between individuals, but rather a social problem necessitating the involvement of multiple levels of society – including communities.

Communities

In order to successfully combat IPA, community mobilization must occur (Harvey et al., 2007). Involving communities in prevention and intervention programs is important as these efforts can stop IPA by increasing awareness and education among residents regarding the complexities of abuse (Smithey and Straus, 2004). In addition, engaging communities in the prevention and intervention of IPA is critical as these environments can assist families throughout the lifespan and, as a result, reduce related social problems: exploitation, neglect, and violence (Andrews, 1994). While community involvement is essential to the prevention and intervention of IPA, it is particularly important in rural communities where increased isolation and limited services further complicate the problem (Gadomski et al., 2001).

According to Gadomski et al. (2001), there is minimal information on whether public education campaigns increase

awareness of IPA or promote changes in attitudes and behaviors regarding abuse – particularly among rural communities. To address this gap in knowledge, the scholars conducted a 7-month campaign and utilized the following: mailings to libraries and clergy, posters, printed advertisements and media articles, and radio advertisements (Gadomski et al., 2001). At the conclusion of the campaign, Gadomski et al. (2001) found statistically significant increases in favorable behavioral intention (e.g., assisting the victim, consulting with others about IPA) among residents of the targeted community – particularly among men. Similar public information and awareness campaigns have also been utilized internationally.

Although Harvey et al. (2007) note that initiatives based solely on public awareness campaigns cannot typically foster *sustained* change, these programs can be successful in reaching large numbers of people. For example, the United Nations' (UN) Interagency Campaign to End Violence Against Women consisted of a regional mass media campaign held in Africa, Asia, Latin America and the Caribbean, and the Pacific (Harvey et al., 2007). The types of media utilized by the UN to increase awareness of IPA included the following: badges, community events, public service announcements, posters, theme songs, and T-shirts (Harvey et al., 2007). Based on the outcomes of the campaign, the initiative appeared to be successful at fostering change at a macrolevel in several countries: legislative reviews were conducted, new laws were implemented, and funding was allocated to IPA agencies (Harvey et al., 2007). Although programs may utilize public awareness and education campaigns to reach specific audiences, other initiatives – particularly found in the international community – engage community residents directly in the effort to combat IPA.

One example of such an initiative exists in Ecuador, where close friends and relatives (referred to as compardres) are partnered with newlywed couples for a period of time in order to assist in instances of marital conflict (Smithey and Straus, 2004). A similar program exists in Marshall Islanders of Oceania, where elders supervise and intervene in episodes of IPA (Counts et al., 1992 cited in Smithey and Straus, 2004). In other programs, not only did community members become directly engaged with combatting IPA (through attending educational trainings), but passed this knowledge on by training fellow residents as well (Kugel et al., 2009). As a result, the *Familias con Voz* (Families with a Voice) prevention program provides a sustainable method to combat IPA – especially among migrant communities (Kugel et al., 2009). Indeed, utilizing community leaders as 'promoters' (referred to as *promotoras*) of change has achieved great success in other fields (e.g., health services), and is now recognized as a viable method to address blatant inequities (Ingram et al., 2008). Overall, researchers have found that community-based approaches are most successful when residents take ownership of the issue as well as participate in program efforts to address the problem, which includes engaging youth as well (Harvey et al., 2007).

Educational Institutions

School-based education programs are vital components of prevention and intervention, because targeted youth learn nonviolent conflict resolution and interpersonal skills necessary to stop the cycle of violence (Smithey and Straus, 2004). Indeed, recognizing that research has supported the relationship between exposure to violence and a higher likelihood of engagement in IPA – the involvement of youth in prevention and intervention programs is important (Harvey et al., 2007). The aforementioned is also particularly critical, because evaluation research has largely shown that youth are less accepting of IPA after exposure to prevention and intervention programming (Jaffe et al., 1992). In reviewing the various forms of school-based IPA prevention and intervention, scholars have identified four different levels of programming, which become more involved and detailed as children grow (Powell et al., 1996).

The first and most commonly used level entails programming directed at students, such as teaching sessions (Powell et al., 1996). The second level typically involves parents and peers in the program in order to reinforce nonviolent messages inside the classroom through peer groups and outside the classroom through parents (Powell et al., 1996). The third level of prevention requires working with social groups, such as neighborhoods, to promote nonviolent behaviors inside school and in the surrounding community (Powell et al., 1996). Finally, the fourth level of prevention targets economic opportunities, firearm availability, media exposure, and similar policy/legislative matters in order to promote broader social change that addresses violence in general (Powell et al., 1996). Aside from the level utilized, it is also important to consider the age of the target population in program development and implementation as effects are greatest when the activities are age appropriate (Harvey et al., 2007).

Prekindergarten (Pre-K) and elementary school programs typically include fostering a general understanding of interpersonal abuse rather than focusing on specific types of violence (Allen, 2009; O'Brien, 2001). The rationale for generalizing the topic is that children within these age groups are not old enough to fully understand the complexities of IPA; however, they still need strategies of how to react when confronted with violence (Allen, 2009; O'Brien, 2001). Fortunately, research has identified several evidenced-based practices utilized in early violence programs: bibliotherapy, group discussions, games, and role playing (Jaffe et al., 1986 cited in Thompson and Trice-Black, 2012). Although many early programs focus on older children (e.g., elementary school and older) evidence suggests Pre-K youth can also benefit from prevention activities (Allen, 2009). For example, evidence indicates that the Creating a New Generation of Peacemakers (hereafter Peacemakers) program successfully impacted Pre-K children (Allen, 2009).

During a 5-week period, the Peacemakers program promoted conflict avoidance tactics, conflict resolution skills, and encouraged respect for others among Pre-K children (Allen, 2009). In order to accomplish the aforementioned, five one hour interactive educational sessions were hosted over a 5 week period, which consisted of various activities: designing a square for a Peacemakers quilt, creating a puppet to illustrate feelings, drawing a safe place in the home or neighborhood (Allen, 2009). At the conclusion of the program, youth involved in the Peacemakers initiative demonstrated statistically significant improvements in desirable skills: conflict avoidance, conflict resolution, and respect for others (assessed by collecting

evaluations of children from teachers) (Allen, 2009). In addition to Peacemakers, several other similarly focused early programs have been successful at reducing the impact of violence on young children: Childreach; Second Step: A Violence Prevention Curriculum; Resolving Conflict Creatively Program; PeaceBuildiers; and The Incredible Years (Guttman et al., 2006). In contrast to young children, programs targeting middle and high school youth are normally more direct.

Programs targeted at middle school and high school youth are more detailed and usually target dating behavior (O'Brien, 2001). Moreover, by the time youth reach high school age, programs typically comprise of multiple components: an integrated curricula, peer groups, and violence-awareness events (O'Brien, 2001). One example of a school-based dating violence prevention program targeting older youth (13–15 years of age) is the Safe Dates program in North Carolina (Harvey et al., 2007). In the Safe Dates program, youth attended 10 educational sessions as well as participated in a poster contest and theater production (Harvey et al., 2007). At the conclusion, youth involved in the program reported less engagement in negative behaviors: psychological, moderate physical, and sexual dating violence perpetration (Harvey et al., 2007). Moreover, there were statistically significant declines in moderate physical victimization (Harvey et al., 2007). Aside from targeting youth at particular grade levels, initiatives have also been developed for student organizations (e.g., student athletes) within schools.

According to Miller et al. (2012), the Coaching Boys into Men (CBIM) program utilizes the relationship between male high school student athletes and male coaches to promote positive development among youth. In the CBIM program, coaches serve as positive role models who lead student athletes in weekly discussions about respect and dating violence prevention (Miller et al., 2012). In order to lead discussions, coaches themselves undergo a 60-min training session led by a violence prevention advocate at which they are introduced to a Coaches Kit to help them facilitate the program (Miller et al., 2012). In total, the program encompasses three main components: (1) raising awareness of negative behavior and encouraging positive alternatives, (2) promoting gender equality and respect, and (3) developing bystander intervention skills through the modeling of coaches (Miller et al., 2012).

Results indicate the CBIM is an effective school athletics-based program to reduce dating violence (Miller et al., 2012). According to Miller et al. (2012), CBIM athletes reported greater intentions to intervene and more positive bystander behavior. Similarly, the Mentors in Violence Prevention (MVP) program is another effective method to engage student athletes and leaders in order to raise awareness of IPA, challenge norms supportive of violence, and promote bystander intervention (Harvey et al., 2007). Results from the MVP initiative indicate trainees improved on knowledge and attitudes of IPA following the program (Harvey et al., 2007). Overall, school-based violence prevention and intervention initiatives are essential to reducing IPA and dating violence, but more evaluation is necessary to understand long-term impacts of these efforts (Harvey et al., 2007; O'Brien, 2001). Another target population that is critical to combatting IPA involves those individuals interacting with victim/survivors on a daily basis – first responders and the general medical community.

First Responders and Medical Personnel

As knowledge of IPA has grown, the importance of first responders and medical personnel as 'first lines of defense' in the effort to combat abuse has also increased. For example, according to Smithey and Straus (2004), information about IPA is now part of medical school curricula and research is routinely funded to study the problem. Additionally, the American Medical Association suggests all providers screen for IPA, a practice that was subsequently endorsed by other professional organizations: American College of Emergency Physicians and the American College of Obstetricians and Gynecologists (Alpert, 2007). Finally, the Joint Commission on the Accreditation of Healthcare Organizations mandates hospitals have IPA screening and referral protocols in place to assist victim/survivors (Alpert, 2007). The importance of first responders and medical personnel cannot be understated as these individuals are also usually the first individuals to interact with victim/survivors in rural communities where there is limited access to other social services – thereby further emphasizing the significance of their role in the prevention and intervention of IPA (Elliott, 1993 cited in Hyman et al., 2000).

The convergence of victim/survivors and first responders/medical personnel provides important prevention and intervention opportunities. Scholars have found approximately 17% of IPA victim/survivors seeking emergency care disclose abuse to personnel (Larkin and Parks, 2009). However, despite this knowledge, research indicates a majority of medical personnel do not query patients about IPA during consultations (Rodríguez et al., 2001). This lack of questioning is likely due to several disclosure barriers found within the medical setting: patients' belief that clinicians lack the time to discuss abuse, patients' fear of involving law enforcement, patients' perceptions that medical personnel lack interest in discussing IPA, and patients' concerns about confidentiality (Rodríguez et al., 2001). In recognition of the aforementioned barriers, programs have focused on engaging various personnel to increase intervention within the medical setting. For example, after attending training on IPA, medical students reported greater comfort with responding to signs of abuse (Moskovic et al., 2008). Research conducted on programs with a similar focus produced comparable outcomes: medical students reported greater awareness and understanding of IPA (Wielichowski et al., 1999 cited in Davidson et al., 2001; Jonassen et al., 1999). In addition to medical students, efforts are targeted at practitioners directly.

In a review of programs targeted at health care providers, Davidson et al. (2001) note that relatively few published studies have assessed long-term impacts among practitioners. However, one study evaluated the impact of the RADAR (Routine screening, Ask direct questions, Document your findings, Assess patient safety, Review patient options and referrals (Harwell et al., 1998)) program on health care providers located within various community health centers (Harwell et al., 1998). Per Alpert (2007), the RADAR program has been utilized extensively in various health care settings since 1992 and is especially suited to encouraging intervention among medical personnel. Composed of three parts, the RADAR program includes (1) training to all health care staff, (2) theoretical context to understand impacts of abuse on

various parties involved, and (3) follow-up support to health care staff (Harwell et al., 1998).

According to Harwell et al. (1998), after phase one of the RADAR program was completed, trainees reported greater comfort and knowledge regarding IPA. Moreover, most trainees expressed satisfaction with the training component of RADAR, and that the training met their expectations as far as information presented (Harwell et al., 1998). Although some of these immediate impacts of the training declined by 3 months and there was little improvement in personnel actually documenting abuse, Harwell et al. (1998) note the training was effective in providing the necessary knowledge to providers in order to assist IPA victim/survivors in the medical setting. While victim/survivors typically seek out first responders and medical personnel to help 'physically' heal them, many seek out religious leaders to assist in 'spiritual' healing. Therefore, religious leaders are critically important to prevention and intervention efforts.

Religious Leaders

Research indicates that victim/survivors of IPA often turn to religious leaders after victimization to obtain guidance and acquire context for understanding abuse (Fortune, 2001; Strickland et al., 1998). Hence, religious leaders can have a pivotal role in the prevention and intervention of IPA, which can either be a positive or negative experience for the victim/survivor (Fortune, 2001; Strickland et al., 1998). Unfortunately, in cases of the latter, research has documented victim/survivors experiencing the following: abandonment by their faith communities, 'shaming' and blaming for *their* abuse, and otherwise being held responsible for their own victimization (Fortune, 2001). As a result, engagement of religious leaders in the prevention and intervention of IPA is especially important.

In one noteworthy study, Jones and Fowler (2009) describe the implementation and evaluation of a Faith Leader Training Program, designed to train religious leaders about how to recognize and respond appropriately to IPA. Developed as a six hour training module, the basic components included providing information regarding: an overview of IPA, effective and appropriate responses to abuse, unique challenges faced by faith leaders (e.g., misinterpretation of sacred texts), and local IPA services for victim/survivor referral (Jones and Fowler, 2009). In addition, a quarterly newspaper was mailed to all program participants that kept the topic visible (Jones and Fowler, 2009). According to Jones and Fowler (2009), pre- and posttraining surveys indicate statistically significant changes in trainees' attitudes, beliefs, and knowledge of IPA as well as issues concerning victim/survivor safety. Finally, aside from training results, other indications suggest that these efforts had a substantial impact. For example, a church involved in the program developed a spiritual spa for victim/survivors residing at a local women's shelter (Jones and Fowler, 2009).

Putting It All Together: Coordinated Community Action Models

The Coordinated Community Action Model (CCAM) developed initially from the work of advocates in the 1980s who were combatting IPA from a holistic perspective (Barner and Carney, 2011; Hamby, 1998; Pence and Paymar, 1993; Shepard et al., 2002). Referred to as the Duluth Domestic Abuse Intervention Project (DAIP or (later) the Duluth model), the following nine agencies collaborated to develop standard policies and procedures for responding to IPA: the county court, the county jail, three mental health agencies, the police department, the probation department, the prosecutor's office, and the shelter (Pence and Paymar, 1993). Overall, the seminal work of Pence and Paymar (1993) was critically important to the family violence field for several reasons: (1) DAIP shifted focus away from victim/survivor responses to emphasizing punitive responses to abusers' violent episodes, (2) DAIP shifted focus away from individual-level factors to recognition of violence rooted in patriarchy, and (3) DAIP established the important role of collaboration in intervening and preventing IPA. Throughout the progression of the family violence field, CCAM like DAIP have continued to be utilized as methods of prevention and intervention of IPA (Barner and Carney, 2011; Hamby, 1998; Navarro, 2013; Shepard et al., 2002) including internationally (Chan et al., 2009). Moreover, although relatively few evaluations exist on CCAM, Hamby (1998) reports several available evaluations have produced positive results indicating the viability of these programs.

Although CCAM vary in scope, most models include a collaboration with counselors, judges, police, probation officers, prosecutors, and shelter advocates (Shepard et al., 2002). Through this collaboration, most CCAM continue the traditional focus of Duluth's DAIP: emphasize offender accountability (e.g., pro-arrest or mandatory arrest policies) and supportive victim/survivors (Shepard et al., 2002). (It is important to acknowledge the ongoing debate regarding whether mandatory arrest policies are always in the best interest of victim/survivors.) In one noteworthy CCAM evaluation, Shepard et al. (2002) reviewed whether enhancements made to the existing DAIP model impacted offender recidivism. DAIP enhancements, referred to as EDAIP, included the use of danger assessment tools, a probation-sentencing matrix, and a computerized monitoring system (Shepard et al., 2002).

Although there were several challenges encountered during the course of the evaluation, Shepard et al. (2002) note their data show EDAIP offenders had lower rates of recidivism compared to those in DAIP. As a result, while Shepard et al. (2002) state that their study was unable to evaluate the effectiveness of any one component, they claim their data suggest that risk of recidivism can be reduced with enhanced coordinated responses. According to Chan et al. (2009), although CCAM were originally implemented more along the lines of community justice approaches, the concept of coordinating community responses to social problems has shown merit in prevention as well.

Program evaluation findings from a CCAM utilized in a small community in Orlando, Florida, support the claims made by Chan et al. (2009). In contrast to the previous study noted, this CCAM (referred to as Project Courage) focused on the primary prevention of IPA by engaging several stakeholders within the target population: businesses, community residents, educators and youth, first responders, and religious leaders (Navarro, 2013). In addition to involving multiple groups in the prevention effort, Project Courage included multiple activities as part of the overall initiative: a training to empower individuals to 'Recognize, Respond, and Refer' (R3) for IPA,

a business initiative referred to as 'Purple Key,' multiple youth programs, and religious leader roundtables (Navarro, 2013). Although there were several challenges encountered during the implementation and progression of the program, results from the first year of the initiative indicate that Project Courage was successful at fostering an immediate change in the recognition and response to IPA. Thus, evidence suggests CCAM continue to be a novel and effective approach to the prevention and intervention of IPA.

Methodological Limitations and Future Directions

This article provided an overview of the current state of knowledge regarding evidence-based IPA prevention and intervention efforts – stemming from applied research on the topic. While the aforementioned studies highlight the important advancements made after four decades of effort, there are several limitations that should be noted. First, many of the studies noted in this article were limited in their ability to assess long-term impact. As a result, additional research assessing long-term impacts (e.g., >1 year) is warranted before firm conclusions are drawn.

Second, despite the important role that certain individuals have in the prevention and intervention of IPA (e.g., religious leaders and employers), there was a lack of information regarding initiatives targeting these groups specifically. Thus, there was limited applied research to report. Moreover, few studies focused on marginalized populations or special populations (e.g., military). For example, in one cutting-edge study, Dichter and Marcus (2013) note relatively little is known about the prevalence of IPA among female veterans. As a result, future research should endeavor to address these persistent gaps in knowledge.

Third, applied research often presents unique challenges to scholars, which may impact the desired rigor of studies. While scholars attempt to overcome these challenges and utilize the most rigorous methods (e.g., randomized designs, comparison groups, sufficient follow-up, replication, etc.; Whitaker et al., 2013), this is not always possible. As a result, the quality of studies presented in this article is not consistent. This lack of consistency is well recognized in the literature, and several scholars have discussed the challenges and barriers faced in applied research (see Navarro, 2013; Riger, 2002; Whetstone and Tewksbury, 2000). Taking all the aforementioned into account, this article should serve as a starting point to conduct a more in-depth review of the applied research presented here.

Acknowledgments

Information included in this article is part of a dissertation entitled, 'Promoting Courage: An Evaluation of Harbor House of Central Florida's Domestic Violence Primary Prevention Initiative Project Courage.'

See also: Applied Social Research, History of; Applied Sociology; Collective Efficacy and Crime; Communities and Crime; Crime and Gender; Crime in the Rural Context; Deterrence Theory: Crime; Domestic Violence: Sociological Perspectives; Gender and Women's Studies, Applied Research on; Intersectionality and Crime; Intimate Partner Violence; Prevention of Crime and Delinquency; Rape and Sexual Coercion; Rural Sociology; Sexism; Social Learning Theory and Crime; Workplace Safety and Health.

Bibliography

Allen, S.F., 2009. A study of a violence prevention program in prekindergarten classrooms. Children & Schools 31, 177–187.

Alpert, E.J., 2007. Healthcare professionals' roles in domestic violence. In: Jackson, N.A. (Ed.), Encyclopedia of Domestic Violence, first ed. Routledge, New York, NY.

Andrews, A.B., 1994. Developing community systems for the primary prevention of family violence. Family & Community Health 16, 1–9.

Bacchus, L.J., Bewley, S., Vitolas, C.T., Aston, G., Jordan, P., Murray, S.F., 2010. Evaluation of a domestic violence intervention in the maternity and sexual health services of a UK hospital. Reproductive Health Matters 18, 147–157.

Bachman, R., 1994. Violence and Theft in the Workplace: National Crime Victimization Survey. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Barner, J.R., Carney, M.M., 2011. Interventions for intimate partner violence: a historical review. Journal of Family Violence 26, 235–244.

Babcock, J.C., Green, C.E., Robie, C., 2004. Does batterers' treatment work? A meta-analytic review of domestic violence treatment. Clinical Psychology Review 23 (8), 1023–1053.

Bronfenbrenner, U., 1979. The Ecology of Human Development: Experiments by Design and Nature. Harvard University Press, Cambridge, MA.

Chan, Y.-C., Lam, G.L., Cheng, H.C., 2009. Community capacity building as a strategy of family violence prevention in a problem-stricken community: a theoretical formulation. Journal of Family Violence 24, 559–568.

Control, N. C. F. I. P. A, 2003. Costs of Intimate Partner Violence against Women in the United States. National Center for Injury Prevention and Control, Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, Atlanta, GA.

Corso, P., 2009. Economic analysis and the prevention of intimate partner violence. In: Mitchell, C., Anglin, D. (Eds.), Intimate Partner Violence: A Health-based Perspective. Oxford University Press, New York, pp. 53–60.

Davidson, L.L., Grisso, J.A., Garcia-Moreno, C., Garcia, J., King, V.J., Marchant, S., 2001. Training programs for healthcare professionals in domestic violence. Journal of Women's Health & Gender-Based Medicine 10, 953–969.

Dichter, M.E., Marcus, S.C., 2013. Intimate partner violence victimization among women veterans: health, health care service use, and opportunities for intervention. Military Behavioral Health 1, 107–113.

El-Mouelhy, M., 2004. Violence against women: a public health problem. The Journal of Primary Prevention 25, 289–303.

Fortune, M.M., 2001. Religious Issues and Violence against Women. In: Renzetti, C.M., Edleson, J.L., Bergen, R.K. (Eds.), Sourcebook on Violence Against Women. Sage Publications, Inc., Thousand Oaks, CA.

Gadomski, A.M., Tripp, M., Wolff, D.A., Lewis, C., Jenkins, P., 2001. Impact of a rural domestic violence prevention campaign. The Journal of Rural Health 17, 266–277.

García-Moreno, C., Jansen, H., Ellsberg, M., Heise, L., Watts, C., 2005. WHO Multi-country Study on Women's Health and Domestic Violence against Women. World Health Organization, Geneva.

Gelles, R.J., 1985. Family violence. Annual Review of Sociology 11, 347–367.

Gondolf, E.W., 2011. The weak evidence for batterer program alternatives. Aggression and Violent Behavior 16 (4), 347–353.

Gover, A.R., Macdonald, J.M., Alpert, G.P., 2003. Combating domestic violence: findings from an evaluation of a local domestic violence court. Criminology & Public Policy 3, 109–131.

Guttman, M., Mowder, B., Yasik, A., 2006. Early violence prevention programs: implications for violence prevention against girls and women. Annals of the New York Academy of Sciences 1087, 90–102.

Hamby, S., 1998. Partner violence: prevention and intervention. In: Jasinski, J.L., Williams, L.M. (Eds.), Partner Violence: A Comprehensive Review of 20 Years of Research. Sage, Thousand Oaks, CA.

Hamilton, L., Koehler, J.A., Lösel, F.A., 2013. Domestic violence perpetrator programs in Europe, part I: a survey of current practice. International Journal of Offender Therapy & Comparative Criminology 57 (10), 1189–1205. http://dx.doi.org/10.1177/0306624X12469506.

Harvey, A., Garcia-Moreno, C., Butchart, A., 2007. Primary Prevention of Intimate-Partner Violence and Sexual Violence: Background Paper for WHO Expert Meeting. World Health Organization, 1–37.

Harwell, T.S., Casten, R.J., Armstrong, K.A., Dempsey, S., Coons, H.L., Davis, M., 1998. Results of a domestic violence training program offered to the staff of urban community health centers. American Journal of Preventive Medicine 15, 235–242.

Heise, L.L., 1998. Violence Against Women: An Integrated, Ecological Framework. Violence against Women 4, 262–290.

Hyman, I., Guruge, S., Stewart, D.E., Ahmad, F., 2000. Primary prevention of violence against women. Women's Health Issues 10, 288–293.

Ingram, M., Sabo, S., Rothers, J., Wennerstrom, A., De Zapien, J.G., 2008. Community health workers and community advocacy: addressing health disparities. Journal of Community Health 33, 417–424.

Isaac, N.E., 1997. Corporate Sector Response to Domestic Violence. Harvard University School of Public Health, Cambridge, MA.

Jaffe, P.G., Sudermann, M., Reitzel, D., Killip, S.M., 1992. An evaluation of a secondary school primary prevention program on violence in intimate relationships. Violence and Victims 7, 129–146.

Jonassen, J.A., Pugnaire, M.P., Mazor, K., Regan, M.B., Jacobson, E.W., Gammon, W., Doepel, D.G., Cohen, A.J., 1999. The effect of a domestic violence interclerkship on the knowledge, attitudes, and skills of third-year medical students. Academic Medicine 74, 821–828.

Jones, A.S., Fowler, T.S., 2009. A faith community-domestic violence partnership. Social Work and Christianity 36, 415–429.

Kugel, C., Retzlaff, C., Hopfer, S., Lawson, D.M., Daley, E., Drewes, C., Freedman, S., 2009. Familias con voz: community survey results from an intimate partner violence (ipv) prevention project with migrant workers. Journal of Family Violence 24, 649–660.

Larkin, G.L., Parks, J., 2009. Acute intervention for intimate partner violence in the medical setting. In: Mitchell, C., Anglin, D. (Eds.), Intimate Partner Violence: A Health Based Perspective. Oxford University Press, New York, NY.

Mahoney, P., Williams, L.M., West, C.M., 2001. Violence against women by intimate relationship partners. In: Renzetti, C.M., Edleson, J.L., Bergen, R.K. (Eds.), Sourcebook of Violence Against Women, Second ed. Sage Publications, Thousand Oaks, CA.

Maxwell, C.D., Garner, J.H., Fagan, J.A., 2002. The preventive effects of arrest on intimate partner violence: research, policy and theory. Criminology & Public Policy 2, 51–80.

Miller, E., Tancredi, D.J., Mccauley, H.L., Decker, M.R., Virata, M.C.D., Anderson, H.A., Stetkevich, N., Brown, E.W., Moideen, F., Silverman, J.G., 2012. "Coaching boys into men": a cluster-randomized controlled trial of a dating violence prevention program. Journal of Adolescent Health 51, 431–438.

Mitchell, C., Vanya, M., 2009. Explanatory Frameworks of Intimate Partner Violence. In: Mitchell, C., Anglin, D. (Eds.), Intimate Partner Violence: A Health Based Perspective. Oxford University Press, New York, NY.

Moskovic, C.S., Guiton, G., Chirra, A., Núñez, A.E., Bigby, J., Stahl, C., Robertson, C., Sims, A., Thul, E.C., Miller, E., Sims, A., Sachs, C.J., & Pregler, J.P., 2008. Impact of participation in a community-based intimate partner violence prevention program on medical students: a multi-center study. Journal of General Internal Medicine 23, 1043–1047.

Murray, S., Powell, A., 2008. Working it Out: Domestic Violence Issues and the Workplace. Australian Domestic & Family Violence Clearinghouse, UNSW.

Navarro, J., 2013. Promoting Courage an Evaluation of Harbor House of Central Florida's Domestic Violence Primary Prevention Initiative Project Courage. University of Central Florida, Orlando, FL.

Nicolaidis, C., Paranjape, A., 2009. Defining intimate partner violence: controversies and implications. In: Mitchell, C., Anglin, D. (Eds.), Intimate Partner Violence: A Health Based Perspective. Oxford University Press, New York, NY.

O'Brien, M.K., 2001. School-based education and prevention programs. In: Renzetti, C.M., Edleson, J.L., Bergen, R.K. (Eds.), Sourcebook on Violence Against Women. Sage Publications, Inc., Thousand Oaks, CA.

Park, S., 2003. Working towards freedom from abuse: recognizing a 'public policy' exception to employment-at-will for domestic violence victims. New York University Annual Survey of American Law 59, 121.

Pence, E., Paymar, M., 1993. Education Groups for Men Who Batter: The Duluth Model. Springer Publishing Company, New York, NY.

Pogarsky, G., 2009. Deterrence and decision making: research questions and theoretical refinements. In: Handbook on Crime and Deviance. Springer, New York, NY, pp. 24–258.

Powell, K.E., Dahlberg, L.L., Friday, J., Mercy, J.A., Thornton, T., Crawford, S., 1996. Prevention of youth violence: rationale and characteristics of 15 evaluation projects. American Journal of Preventive Medicine 12, 3–12.

Riger, S., Bennett, L., Wasco, S.M.M., Schewe, P.A., Frohmann, L., Camacho, J.M., Campbell, R., 2002. Evaluating services for survivors of domestic violence and sexual assault. Sage Publications, Thousand Oaks, CA.

Rodríguez, M.A., Sheldon, W.R., Bauer, H.M., Pérez-Stable, E.J., 2001. The factors associated with disclosure of intimate partner abuse to clinicians. Journal of Family Practice 50, 338–344.

Rolling, E., Brosi, M., 2010. A multi-leveled and integrated approach to assessment and intervention of intimate partner violence. Journal of Family Violence 25, 229–236.

Saltzman, L.E., Houry, D., 2009. Prevalence of nonfatal and fatal intimate partner violence. In: Mitchell, C., Anglin, D. (Eds.), Intimate Partner Violence: A Health Based Perspective. Oxford University Press, New York, NY.

Shepard, M.F., Falk, D.R., Elliott, B.A., 2002. Enhancing coordinated community responses to reduce recidivism in cases of domestic violence. Journal of Interpersonal Violence 17, 551–569.

Sherman, L.W., Berk, R.A., 1984. The specific deterrent effects of arrest for domestic assault. American Sociological Review 49, 261–272.

Smithey, M., Straus, M.A., 2004. Primary prevention of intimate partner violence. In: Kury, H., Obergfell-Fuchs, J. (Eds.), Crime Prevention–New Approaches. Weisser Ring Gemeinnutzige Verlags-GmbH, Mainz.

Strickland, G.A., Welshimer, K.J., Sarvela, P.D., 1998. Clergy perspectives and practices regarding intimate violence: a rural view. The Journal of Rural Health 14, 305–311.

Swanberg, J.E., Logan, T., 2005. Domestic violence and employment: a qualitative study. Journal of Occupational Health Psychology 10, 3.

Thompson, E.H., Trice-Black, S., 2012. School-based group interventions for children exposed to domestic violence. Journal of Family Violence 27, 233–241.

Tjaden, P., Thoennes, N., 2000. Full Report of the Prevalence, Incidence, and Consequences of Violence against Women Series: Research report. Atlanta, GA: Centers for Disease Control and Prevention.

U.S. Department of Justice, Office of Justice Programs, 2001. Stalking and domestic violence: Report to Congress. Report No. 186157.

Urban, B.Y., Bennett, L.W., 1999. When the community punches a time clock evaluating a collaborative workplace domestic abuse prevention program. Violence against Women 5, 1178–1193.

Whetstone, T.S., Tewksbury, R., 2000. Desperately seeking significance: problems and difficulties associated with evaluating the efficacy of domestic violence response programs. Social Pathology 6, 102–120.

Whitaker, D.J., Morrison, S., Lindquist, C., Hawkins, S.R., O'Neil, J.A., Nesius, A.M., Mathew, A., Reese, L.R., 2006. A critical review of interventions for the primary prevention of perpetration of partner violence. Aggression and Violent Behavior 11, 151–166.

Intimate Partner Violence

Jennifer T Buntin, Lewis University, Romeoville, IL, USA; and One University Parkway, Romeoville, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sexual and intimate relationships are often plagued with violent acts. This article focuses on two key challenges that researchers, as well as groups and organizations seeking to provide support, face in interpreting and understanding peoples' real experiences of violence within their relationships and the consequences that these interpretations have for understanding and providing support to victims. The first challenge revolves around distinguishing between different types of intimate partner violence, as categorized by Michael P. Johnson (1995, 2008). The second challenge considers the role that gender, sexual orientation, race, and ethnicity play in the interpretations and explanations for intimate partner violence.

Sexual and intimate relationships are often plagued with violent acts. Data from the National Intimate Partner and Sexual Violence Survey show that more than one in three women (35.6%) and one in four men (28.5%) in the United States report rape, physical violence, and/or stalking by an intimate partner in their lifetime (Black et al., 2011). These experiences of violence within their relationships impact the individual's life in a variety of ways, such as post-traumatic stress disorder, injury recovery, sexually transmitted diseases, long-term anxiety and stress, depression, emotional detachment, sleep disturbances, suicidal behavior, missed work or school, homelessness, substance abuse, and more (Aguilar and Nightengale, 1994; Breiding et al., 2008; Crofford, 2007; Heise and Moreno, 2002; Johnson and Leone, 2005).

Given the prevalence of violence within intimate relationships, much research, including those studies listed above, details the patterns of violent behavior in intimate relationships and builds an understanding of how and why such violence occurs (for overviews, see Hattery, 2009; Johnson and Ferraro, 2000). This article focuses on two key challenges that researchers, as well as groups and organizations seeking to provide support, face in interpreting and understanding people's real experiences of violence within their relationships and the consequences that these interpretations have for understanding and providing support to victims. The first challenge revolves around distinguishing between different types of intimate partner violence, as categorized by Michael P. Johnson (1995, 2008). The second challenge considers the role that gender, sexual orientation, race, and ethnicity play in the interpretations and explanations for intimate partner violence.

Johnson's Categories of Intimate Partner Violence

Early sociological research on intimate partner violence generally followed two perspectives: the family-violence perspective, attributed originally to Gelles (1974) and Strauss (1979); and the feminist perspective, exemplified by Dobash and Dobash (1979), Walker (1984), and Kurz (1989). The two perspectives utilized different types of data and came to very different conclusions about the prevalence and causes of intimate partner violence. The family-violence perspective relies primarily on large-scale quantitative data analysis and concentrates on discrete, violent acts occurring during family conflict. Studies using this perspective perceive violence as a relatively common element of intimate and family relationships, which can be exacerbated by stress. In contrast, scholars working within the feminist perspective are more likely to use qualitative data, such as interviews of women living in battered women's shelters. The feminist perspective views intimate partner violence as a consequence of patriarchal relations, constructions of gender roles, and structural constraints that limit exit opportunities for women experiencing intense, systematic beatings (Buntin et al., 2004).

The differences in conceptualizing violence in relationships, and the reliance on quite different sources of data resulted in very different conclusions about the prevalence and nature of intimate partner violence from these two perspectives. The findings have been so different that researchers from both perspectives have long argued that the other perspective misunderstands the nature of family violence (Johnson, 1995). Scholars in the family-violence paradigm see violence as a natural, relatively common, and sometimes mild aspect of intimate relationships, whereas feminist researchers perceive violence in intimate relationships as more intense, more brutal, and the result of male attempts to physically and psychologically control women.

Michael P. Johnson (1995, 2008), however, argues that rather than one perspective being wrong and the other right, what really creates the confusion and conflict between the family-violence perspective and the feminist perspective is a homogeneous conceptualization of intimate partner violence. In his groundbreaking analysis, Johnson (1995) argues that both perspectives are accurate, and that they are simply examining different phenomena. He originally distinguishes between two types of intimate partner violence: patriarchal terrorism and common couple violence. Due to the large-scale survey sources utilized, the family-violence perspective primarily observes 'common couple violence' or violence that is generally more moderate in nature, rarely escalating to extreme levels, and more gender balanced, i.e., both men and women participate in the violence at similar rates (Johnson, 1995). In other words, this is the 'everyday' sort of violence that occurs within approximately half of intimate relationships (Olson, 2002). In contrast, the

feminist perspective's dependence on case studies from battered women's shelters yields data on 'patriarchal terrorism,' the more intense, extreme cases of violence in which the male partner physically and psychologically controls and abuses his female partner. This form of violence and control is deeply embedded in "patriarchal traditions of men's right to control 'their' women…a form of terroristic control of wives by their husbands that involved the systematic use of not only violence, but economic subordination, threats, isolation, and other control tactics" (Johnson, 1995: 284). This form of violence is less frequent; therefore it shows up less in large-scale survey data. However, its higher level of intensity means that the consequences for the victims and their children are more extreme, thus leading some women to seek help at battered women's shelters where they come to the attention of the researchers working from the feminist perspective.

Johnson's differentiation between different types of violence helped make sense of the conflicting findings within the intimate partner violence literature. He has since revised and developed his framework, testing it on additional data (Johnson, 2001, 2008; Johnson and Leone, 2005). Other scholars have also tested his typology on different data sets and found support for this approach to domestic violence (Graham-Kevan and Archer, 2003; Leone et al., 2004; Olson, 2002). Johnson's revised typology (2008) includes four categories of intimate partner violence rather than two: intimate terrorism, violent resistance, situational couple violence, and mutual violent resistance. The first, intimate terrorism, is what Johnson previously referred to as patriarchal terrorism. It is violence committed with the intent to gain general control over one's partner. This is the violence that has been the focus of feminist scholars and activists for decades. It is also the type of violence that receives the most media attention and is generally what most people mean when they think of 'domestic violence.' Not all violence perpetrated against intimate partners can be categorized as such. In addition to 'situational couple violence,' Johnson's revised term for the common couple violence, in which there is violence without the intent to fully control the partner, his more recent framework also includes two more categories. 'Violent resistance' refers to the violence perpetrated by partners, generally women, who are the victims of intimate terrorism. In this category, the violence erupts as a way to resist or escape the violence and control of the intimate terrorist. This category includes legally defined self-defense, but also acts of violence that would not be considered self-defense in a courtroom, but stem from the desperation of an individual who has been terrorized and isolated for extended periods. Finally, Johnson defines a much smaller category of intimate partner violence, that of 'mutual violent control' (Johnson, 2008). In these cases, both partners engage in violence and controlling behaviors against each other. In other words, the violence is an active battle for control and domination within the relationship (Johnson, 2008).

Johnson's differentiation of intimate partner violence and the subsequent theoretical development and empirical testing by him and other intimate partner violence researchers has added significantly to our understanding of the role of violence within intimate relationships, as well as the diverse psychological, social, and even economic consequences it can have for the participants and their families.

The Influence of Gender, Sexual Orientation, and Race/Ethnicity on Understandings of Intimate Partner Violence

Gender

Not surprisingly, intimate partner violence is interpreted quite differently by the public and by help providers as a consequence of the race, class, gender, and sexual orientation of the couple. Johnson's work, discussed above, highlights the role that gender plays in shaping understandings of what constitutes intimate partner violence, as well as the likelihood of an individual to experience it. In Johnson's (1995) earlier conceptualization, patriarchal terrorism, the gendered nature of the category is more apparent. Even translated to intimate terrorism, however, research shows that this combination of violence and control-seeking behaviors is primarily perpetrated by men against women (Graham-Kevan and Archer, 2003; Johnson, 2001). Situational couple violence, in contrast, is more gender balanced (Johnson, 1995, 2008).

When it comes to seeking help from organizations and institutions providing assistance to victims of intimate partner violence, the feminist perspective leads us to expect that victims of intimate partner violence are female, potentially making it difficult for male victims to seek help and find assistance. In addition, the expected gendered nature of intimate partner violence may obscure situational couple violence and result in fewer resources for couples for whom the violence is not part of a systematic attempt by the male partner to control the female partner. For these couples, seeking assistance from an institution that perceives all intimate partner violence as intimate terrorism perpetrated by men on women will be useless at best.

In addition, qualitative data collected in four Chicago neighborhoods suggest that race, class, and sexual orientation can also create a miscommunication between victims and help providers. The Chicago Health and Social Life Survey (CHSLS) (Laumann et al., 2004) included qualitative interviews with community leaders and service providers within these neighborhoods that vary by race, ethnicity, and sexual orientation. These data provide a window into the ways that these social characteristics may shape the understandings of violence by the help providers, as well as the likelihood that victims will seek their assistance and the information that the victims may feel comfortable sharing.

Sexual Orientation

Intimate partner violence within same-sex relationships has only recently begun to receive research and media attention. The assumption that domestic violence is primarily perpetrated by men against women meant that it was perceived as a phenomenon found almost exclusively in heterosexual relationships. Chicago's 'Shoreland' neighborhood (a pseudonym) is home to a significant portion of the city's gay community.

> Annie, a leader in the gay and lesbian community in Shoreland, describes intimate partner violence as 'a huge issues in the queer community. Most women don't believe it, though, so it is hard to bring pressure.' She is concerned about the lack of services available to gay and lesbian victims of domestic violence. She feels that homophobia in women's shelters prevents lesbians from seeking and receiving help there. She also points out that there are no places for male victims to go at all.
>
> Buntin et al., 2004: 250–251

Interviews with police officers in the neighborhood illustrate confusion and discomfort regarding same-sex intimate partner violence. Same-sex domestic violence calls were perceived as more violent than heterosexual calls, and therefore more dangerous to the officers. In addition, the common assumption that intimate partner violence is violence perpetrated by men on their female partners creates confusion when officers are confronted with same-sex partners. They can no longer depend on this mental framework in trying to discern the perpetrator from the victim. "A male officer in Shoreland described gay domestic violence calls as disorienting and unsettling…" (Buntin et al., 2004: 251). In general, the officers also appeared to offer little in the way of referrals to service organizations, perhaps also as a result of the uncertainty of where to send gay and lesbian victims of intimate partner violence for assistance.

Although more research is being done in this area, the literature on intimate partner violence within same-sex relationships is still limited. Adam M. Messinger (2011) finds that victims with a history of same-sex relationships are more likely to experience verbal, controlling, physical, and sexual intimate partner violence, even from opposite sex partners. Other researchers have also begun to explore the role that violence plays in same-sex relationships (see, for example, Cruz and Firestone, 1998; Halpern et al., 2004; Merrill and Wolfe, 2000; Renzetti, 1992).

Race/Ethnicity

In the two Latino neighborhoods profiled in the CHSLS, cultural explanations of intimate partner violence were common among service providers. Many perceived Mexican or Puerto Rican culture as the source of intimate partner violence in these neighborhoods, focusing on cultural concepts of male machismo and female subservience. One Catholic church worker stated that "domestic violence is so much a part of the culture that people lack a vocabulary to identify it" (Buntin et al., 2004: 254).

These cultural explanations, however, serve to obscure the structural and economic factors at play as well. For example, a Catholic priest interviewed described violence as "the dark side of Mexican culture" (Buntin et al., 2004: 254). He went on, however, to describe a scenario for violence that is filled with structural and economic influences rather than Mexican culture.

> When a man comes up to the United States, he is suddenly much richer in absolute terms than he was in Mexico, but much poorer in relation to the wider population. He feels inadequate, belittled. Often he will require his wife to stay at home and not work. But even if this is not the case, the only person he can push around is his wife, or the kids when he is in his castle. So there is a very high incidence of domestic violence in Westside.
>
> Buntin et al., 2004: 254

Thus, while he perceived Mexican culture as the source of the violence, the example he gives points toward other factors, such as the migration experience, socioeconomic inequality in the US, and feelings of powerlessness in society by members of marginalized groups. Similarly, service providers interpret Latinas' refusal to leave a violent situation in terms of the cultural value of familia or family. While they may be accurate in part, these cultural explanations are blind to the powerful role that can be played by economic dependence on the perpetrator, lack of family support in the area, and perhaps fear of seeking assistance because of undocumented immigrant status.

In the African-American neighborhood referred to as Southtown, a different type of explanation for intimate partner violence dominated. While service providers in the Latino neighborhoods perceived that their clients willingly accepted violence in their relationships for cultural reasons, providers and community leaders in Southtown often refused to recognize that intimate partner violence was a significant problem in their neighborhood at all. Since the CHSLS data show rates of intimate partner violence in Southtown that are higher than any of the other neighborhoods, such statements suggest a denial or ignorance of an important aspect of residents' family life. The lack of recognition of intimate partner violence in the neighborhood likely stems from a lack of willingness by residents to seek help from service providers in the neighborhood (Buntin et al., 2004). CHSLS researchers also perceived a lack of willingness on the part of the African-American community to fully participate in research conducted by outside organizations. "Although the data are limited, there does seem to be a clear pattern – especially among those affiliated with religious organizations – of violence being interpreted primarily in terms of 'moral choice' and only secondarily in terms of such structural and economic factors as poverty, racism, and lack of education" (Buntin et al., 2004: 259). These moral perspectives may serve to discourage people from seeking help in Southtown. It is also unclear how an organization emphasizing moral choice can intervene effectively except by continuing to preach the virtues of moral strength. Within the context of a structural reality of little economic opportunity, financial instability, racism, etc., individuals may be less inclined to seek assistance from organizations focusing on a moral approach.

These findings from the Chicago neighborhoods included in the CHSLS (Laumann et al., 2004) suggest that gender, sexual orientation, race, and ethnicity play a crucial role in shaping opportunities for assistance available to victims of intimate partner violence. Gendered, cultural, and religious beliefs on the part of the service providers can limit their ability to effectively provide assistance to victims in their

neighborhoods. The findings of the CHSLS suggest that more emphasis on structural and economic factors may improve the ability of service providers to positively impact the lives and circumstances of families dealing with intimate partner violence. For example, service providers working with immigrant victims of domestic violence could gain insight into their clients' experiences and choices by increasing their understanding of immigration policy and the social process of migration. For example, a woman whose legal status in the United States is dependent on her relationship with her husband may feel that she must accept abuse from him rather than risk her legal status. Her acceptance of abuse, in this case, stems not from a cultural belief about women's roles in the family, but on her political and economic position within the US. Consequently, assisting her escape from this situation requires different tactics on the part of the organization.

Similarly, to effectively address violence within African-American and gay/lesbian neighborhoods, those individuals and organizations seeking to assist victims of domestic violence must recognize the ways that cultural understandings about domestic violence (i.e., that 'it doesn't happen in our church' or the implicit expectation that victims are female and perpetrators are male) serve to mask the very real instances of violence and abuse found in predominantly African-American or same-sex neighborhoods.

See also: Crime, Immigration, and Ethnicity; Heterosexuality; Masculinity, Scripts of; Partnership Formation and Dissolution in Western Societies; Sexuality: Cultural Aspects.

Bibliography

Aguilar, R., Nightengale, N., 1994. The impact of specific battering experiences on the self- esteem of abused women. Journal of Family Violence 9, 35–46.

Black, M., Basile, K., Breiding, M., et al., 2011. The National Intimate Partner Violence Survey (NISVS): 2010 Summary Report. National Center for Injury Prevention and Control, Centers for Disease Control and Prevention, Atlanta, GA.

Breiding, M., Black, M., Ryan, G., 2008. Chronic disease and health risk behaviors associated with intimate partner violence—18 U.S. states/territories, 2005. Annual Epidemiology 18, 538–544.

Buntin, J., Lechtman, Z., Laumann, E., 2004. Violence and sexuality: examining intimate-partner violence and forced sexual activity. In: Laumann, E., Ellingson, S., Mahay, J., et al. (Eds.), The Sexual Organization of the City. University of Chicago Press, Chicago, pp. 226–263.

Crofford, L., 2007. Violence, stress, and somatic syndromes. Trauma Violence Abuse 8, 299–313.

Cruz, J., Firestone, J., 1998. Exploring violence and abuse in gay male relationships. Violence and Victims 13 (2), 159–173.

Dobash, R., Dobash, R., 1979. Violence against Wives. Free Press, New York.

Gelles, R., 1974. The Violent Home: A Study of Physical Aggression between Husbands and Wives. Sage, Newbury Park, CA.

Graham-Kevan, N., Archer, J., 2003. Intimate terrorism and common couple violence: a test of Johnson's predictions in four British samples. Journal of Interpersonal Violence 18 (11), 1247–1270.

Halpern, C., Young, M., Waller, M., et al., 2004. Prevalence of partner violence in same-sex romantic and sexual relationships in a national sample of adolescents. Journal of Adolescent Health 35 (2), 124–131.

Hattery, A., 2009. Intimate Partner Violence. Rowman & Littlefield Publishers, Lanham, MD.

Heise, L., Garcia-Moreno, C., 2002. Violence by intimate partners. In: Krug, E., Dahlberg, L., Mercy, J., et al. (Eds.), World Report on Violence and Health. World Health Organization, Geneva (Switzerland), pp. 87–121.

Johnson, M., 1995. Patriarchal terrorism and common couple violence: two forms of violence against women. Journal of Marriage and the Family 57, 283–294.

Johnson, M., 2001. Conflict and control: symmetry and asymmetry in domestic violence. In: Booth, A., Crouter, C., Clements, M. (Eds.), Couples in Conflict. Lawrence Erlbaum, Mahwah, NJ, pp. 95–104.

Johnson, M., 2008. A Typology of Domestic Violence: Intimate Terrorism, Violent Resistance and Situational Couple Violence. Northeastern University Press, Boston.

Johnson, M., Ferraro, K., 2000. Research on domestic violence in the 1990s: making distinctions. Journal of Marriage and the Family 62 (4), 948–963.

Johnson, M., Leone, J., 2005. The differential effects of intimate terrorism and situational couple violence. Journal of Family Issues 26 (3), 322–349.

Kurz, D., 1989. Social science perspectives on wife abuse: current debates and future directions. Gender & Society 3, 489–505.

Leone, J., Johnson, M., Cohan, C., Lloyd, S., May 2004. Consequences of male partner violence for low-income, ethnic women. Journal of Marriage and Family 66, 471–489.

Merrill, G., Wolfe, V., 2000. Battered gay men: an exploration of abuse, help seeking and why they stay. Journal of Homosexuality 39 (2), 1–30.

Messinger, A., 2011. Invisible victims: same-sex IPV in the national violence against women survey. Journal of Interpersonal Violence 26 (11), 2228–2243.

Olson, L., 2002. Exploring "common couple violence" in heterosexual romantic relationships. Western Journal of Communication 66 (1), 104–128.

Renzetti, C., 1992. Violent Betrayal: Partner Abuse in Lesbian Relationships. Sage, Thousand Oaks, CA.

Strauss, M., 1979. Measuring intrafamily conflict and violence: the Conflict Tactics (CT) scales. Journal of Marriage and the Family 41, 75–88.

Walker, L., 1984. The Battered Woman Syndrome. Springer, New York.

Intracarotid Sodium Amytal Test and Language

Caleb M Pearson, Barrow Neurological Institute, Phoenix, AZ, USA
Alexander I Tröster, Barrow Neurological Institute, Phoenix, AZ, USA; and Barrow Center for Neuromodulation, Phoenix, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Many individuals suffering from epilepsy are refractory to medical management. Uncontrolled seizures are psychologically and socially disruptive. Brain surgery affords many individuals the opportunity to lead normal lives by dramatically reducing, or eliminating seizure activity. Surgery, however, carries the risk of postoperative cognitive decline. One way to minimize this risk is to undergo intracarotid sodium amytal testing (IAT). This article familiarizes the reader with the IAT and discusses the strengths and weaknesses of this particular approach to the presurgical examination of the epilepsy patient. Contributions of the IAT to our understanding of the neuroanatomical substrates of language will also be discussed.

Introduction

Approximately 20–40% of individuals with epilepsy do not respond adequately to treatment with anticonvulsant medications (Aicardi and Shorvon, 1997). Uncontrolled seizures are not only psychologically and socially troublesome to individuals with epilepsy, but are also economically disruptive, as many individuals with epilepsy are either underemployed or unemployed. Neurosurgical intervention affords patients with pharmacologically intractable epilepsy the opportunity to eliminate, or at least reduce seizure activity. While epilepsy surgery has become increasingly safe and effective, it is not completely without risk as structures important to both language and memory are commonly epileptogenic and are the target of neurosurgical intervention. Thus, in order to avoid significant insult to language (i.e., aphasia) or memory (i.e., amnesia) the treatment team must ascertain whether the tissue to be resected is functional, and if so, must determine the likelihood and degree of possible postoperative cognitive impairment. It is, therefore, important that both language and memory be evaluated presurgically to assess the risk of impairment postoperatively. The most widely used method for determining language dominance in potential epilepsy surgery cases is known as the intracarotid amobarbital test (IAT), a method employing selective anesthetization of brain tissue. Not only has this procedure been clinically useful, it has also led to important insights into the neuronal correlates of language – particularly our understanding of inter- and intra-hemispheric differences in language organization. This article will detail the basic elements of the IAT, the strengths and weaknesses of the procedure, and the lessons it has taught us regarding the cortical organization of language. Although the Wada test is also considered highly useful in predicting postoperative memory outcome (Milner et al., 1962; Chelune, 1995), we will restrict our focus in this article to the use of the Wada test as it applies to language functioning.

Historical Overview

Gardner (1941) was the first to apply an anesthetic directly to the brain in an attempt to identify eloquent language cortex. He injected procaine directly into the cortex of two left-handed patients who were being evaluated for neurosurgical intervention. The first patient was a 35-year-old man undergoing neurosurgical intervention to remove a tumor located in the right frontal premotor area, while the second was a 27-year-old woman with a history of left-frontal craniotomy and excision of a cerebral cyst who subsequently developed epilepsy. This second patient was to undergo a second left-frontal surgical intervention in order to excise the presumed epileptogenic tissue. Gardner understood that the patients' left-handedness increased the likelihood of having right-hemisphere language dominance. Additionally, he postulated that he might be able to do a more complete resection if language dominance was located in the hemisphere contralateral to the one being operated upon. In the case of the first patient, a trephine involving the right frontotemporal region was completed and a hypodermic needle was inserted into the underlying cortex. Twenty-seven cc of 0.75% procaine hydrochloride was then injected causing a mild left-facial droop without aphasia. The right cerebral hemisphere was subsequently removed successfully without language complications. To evaluate language dominance in the second patient, the left-facial motor area was located using electrical stimulation. A solution of 10 cc of 1.0% procaine was injected just anterior to this, without subsequent aphasia. This patient subsequently underwent a large left-frontal resection with no postoperative aphasia and marked improvement in seizure frequency.

Following Gardner, Juhn Wada, a Japanese-Canadian neurologist, published his paper outlining the IAT (Wada, 1949). Wada originally developed the IAT, known to many as the 'Wada test,' to study the spread of epileptiform discharges between the left and right halves of the brain in patients undergoing unilateral electroconvulsive therapy. Wada noted a disruption of language took place when the presumed language-dominant hemisphere was injected with amobarbital and anesthetized. Based on this observation, he reasoned that this technique may be useful in determining hemispheric language dominance in neurosurgical candidates, and thus reduce the cognitive complications following neurosurgical intervention. Working as a fellow with Theodore Rasmussen at the Montréal Neurological Institute, Wada then deployed the IAT as a part of the preoperative evaluation of persons with

medically refractory epilepsy, and in 1960 the first English publication describing the IAT appeared (Wada and Rasmussen, 1960).

Current Use of the IAT

Currently, the IAT generally has three purposes in epilepsy surgery centers. First, the IAT helps to determine the language-dominant hemisphere of those being considered for neurosurgical intervention for retractable epilepsy. Language lateralization not only helps to reduce postoperative language impairment, but can also help predict risk for postoperative memory impairment (Chelune et al., 1993). The second purpose of the IAT is to corroborate the presumed hemispheric seizure onset determined by other diagnostic methods (e.g., electroencephalography (EEG), video EEG telemetry and ictal semiology, structural and functional radiography, etc.). Finally, the IAT helps to assess the functional memory capacity of the hemisphere contralateral to the one injected (Chelune, 1995).

Although the IAT is an invasive procedure and carries a small degree of risk (e.g., stroke, embolism, etc.), it remains an integral component of the comprehensive presurgical evaluation of patients with intractable epilepsy. It is not without its detractors, however, and some have suggested the use of less invasive procedures such as functional magnetic resonance imaging (fMRI) for language localization is preferable. Others have suggested that the use of baseline neuropsychological testing in conjunction with other test results (e.g., ictal EEG and structural MRI findings) can satisfactorily predict postoperative memory decline (Stroup et al., 2003; Baxendale et al., 2006). Indeed, routine use of the IAT has declined over the last decade, even as the number of epilepsy surgeries has increased. In one survey, only 12% of centers reported that every surgical candidate underwent IAT (see **Figure 1**). By contrast, 85% of centers used the IAT on every patient in 1993 (Rausch et al., 1993). Despite criticisms and a decline in its routine use, the IAT remains the gold standard for the lateralization of language dominance. A recent world-wide survey of 92 epilepsy surgery centers from 31 countries showed that nearly half of all epilepsy surgery centers routinely utilize the IAT during presurgical workups (Baxendale et al., 2008). Furthermore, almost half of all epilepsy surgery centers in the United States conduct the IAT on 75% or more of their patients. Moreover, in the United States the capacity to perform the IAT is one of the certification criteria required of epilepsy centers under the current National Association of Epilepsy Centers (NAEC, 2001) guidelines.

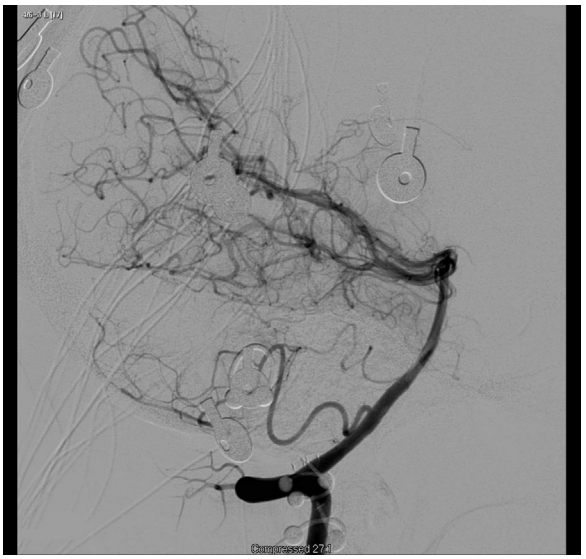

Figure 1 Cerebral angiogram of the left vertebral artery showing the posterior circulation of the brain. The posterior cerebral arteries can be seen toward the middle and upper portion of the figure. These arteries are typically the target of selective Wada testing.

Rationale and Description of the IAT

The basic concept behind the use of the IAT is to determine how well one hemisphere of the brain can support essential cognitive functions such as language and memory, while the other hemisphere is anesthetized. Because the IAT anesthetizes only part of one cerebral hemisphere, it allows one to assess the cognitive functions served by each hemisphere; that is, it is assumed that disruptions of language and memory during the IAT are a consequence of the temporary 'lesioning' of the injected hemisphere, and that the IAT mimics the effects of resective surgery involving that hemisphere. Moreover, because there is a greater incidence of atypical speech development (i.e., right hemisphere or bilateral speech representation) in those with epilepsy, it is important to evaluate rather than presume language lateralization in individual patients.

Although there are other study techniques that can be used to help determine the presumed cognitive functioning of specific brain regions (e.g., fMRI, magnetoencephalogram (MEG)), the IAT is relatively unique in that it uses deactivation as opposed to activation of functional tissue to assess presumed cognitive functioning. In order to do this, the IAT makes use of an agent to anesthetize a portion of one hemisphere of the brain – typically those areas of the brain that receive their blood supply from the internal carotid artery (ICA; commonly known as the anterior circulation, which comprises the anterior cerebral artery, middle cerebral artery, and the anterior choroidal artery). The areas of the brain subserved by these vessels include the lateral frontal and temporal lobes, as well as a portion of the mesial temporal structures such as the anterior one-third of the hippocampus. Alternative selective Wada tests are increasingly used. These procedures make use of microcatheters to anesthetize more circumscribed portions of the brain such as those subserved by the posterior cerebral or anterior choroidal arteries. Potential advantages of selective Wada tests are that they allow assessment of a specific cognitive function such as memory, without disrupting other cognitive abilities such as language. This selectivity has led some to conclude that there may be fewer false positive memory test results, meaning fewer memory test failures due to aphasia rather than amnesia, compared to the more traditional IAT (Der-Jen et al., 2006). Nonetheless, many centers are reluctant to routinely perform selective studies due to possibly higher complication rates and the increased technical expertise demanded. At our center, for example, selective Wada testing

is attempted only when the results of the initial IAT are of questionable validity (e.g., because the patient was overly sedated during the procedure) or when results conflict with data from other tests.

The most widely used anesthetic agent is sodium amobarbital, which crosses the blood–brain barrier (a functional obstruction in the brain's circulatory system, which usually keeps harmful substances out of brain tissue) easily, allowing for a rapid anesthetic effect. Other agents, such as methohexital, pentobarbital, etomidate, and propofol, are also used at some centers. Although the exact methodology varies from center to center, the IAT typically begins with a cerebral angiogram. Angiography is performed with the patient lying in a supine position on a table. Access to the ICA is gained via a catheter inserted through an incision or puncture near the groin that allows access to the femoral artery. The catheter is then typically directed by a neuroradiologist or interventional neurologist until it is placed in the ICA just distal to the common carotid bifurcation contralateral to the cerebral hemisphere to be studied. Cerebral angiography is then performed to examine the amount of cross-flow from one hemisphere to another and to detect atypical vascular organizations that may lead to complications or atypical results (see Figure 2). For instance, a fetal origin to the posterior communicating artery may lead to excess amobarbital being dispersed into the posterior cerebral artery, while the presence of a fetal origin persistent trigeminal artery may lead to amobarbital being delivered to critical respiratory and cardiac centers in the brainstem – an event that could result in death.

The IAT usually begins with the patient holding his arms into the air with his palms upward and counting aloud. Although the amount of amobarbital varies by center, typically anywhere between 50 and 150 mg is injected into the catheter over the course of 4–6 s. The drug is then flushed through the catheter with saline. Immediately following the injection, the patient will often begin to display the telltale signs of neurological dysfunction associated with the anterior cerebral circulation such as contralateral hemiparesis or hemiplegia (weakness or paralysis of the limbs), dysarthria, and at times the patient's eyes will deviate ipsilaterally. If injection of the language-dominant hemisphere was undertaken, abnormalities such as speech arrest and paraphasic errors will often be observed. Speech arrest alone does not necessarily signify the disruption of speech centers within the brain as it is often seen following nondominant hemisphere injections as well, presumably due to frontal lobe dysfunction causing an akinetic mutism-like effect. Thus, speech must be assessed carefully to ensure proper identification of language dominance. Throughout the procedure the patient is monitored via EEG to assess the degree of ipsilateral slowing.

Figure 2 Cerebral angiogram via the right internal carotid artery. Notice the mild degree of cross-flow of contrast to the left anterior cerebral artery via the anterior communicating artery. This degree of cross-flow is typically not concerning as amobarbital is injected at a much slower rate than the contrast used for angiogram, thus leading to a lesser degree of cross-flow during the actual intracarotid sodium amytal testing examination.

After an assessment of basic neurological functions, the language and memory portion of the IAT is typically performed. Most centers use a combination of spoken words and objects presented to the patient while under the effects of amobarbital. At our center, the neuropsychologist performing the assessment uses a tablet computer secured to the hand via a strap to present both written words and common objects to the patient. The patient is asked to read the word or name the object in order to assess their language ability during the procedure. If the patient is unable to produce the correct response, the word is read or the object is named three times to the patient to ensure encoding. Following the presentation of stimuli, which typically takes between 2 and 3 min, the patient's language abilities are assessed. Baseline language functions typically return within 8–10 min, and afterward the patient is asked to recall test stimuli. Most often, the patient is unable to spontaneously recall many of the test stimuli, in which case the patient is presented with a recognition task and asked to select the previously learned stimuli from a set of three similar words/objects.

At many centers, including our own, a second injection is performed on the opposite hemisphere. This allows for a more accurate language assessment as it gives the examiner the opportunity to rule in or out the possibility of bilateral language representation. A second injection also allows for the assessment of functional capacity in the portion of the brain to be resected, which gives the examiners further insight into potential postoperative cognitive decline. The second internal carotid artery is typically injected anywhere between 30 min and several hours after the first injection to ensure that the effects of the amobarbital from the first injection have completely resolved. The test procedure is then repeated, typically using different, but equivalent stimuli in order to reduce difficulties interpreting test results due to practice effects.

Problems and Considerations

Although the rationale underlying the IAT seems straightforward, the performance and interpretation of the procedure are not. Both the determination of speech dominance and that of memory representation are subjective, and the protocols used in

the IAT vary, sometimes greatly, between centers. Additionally, given the invasiveness of the procedure, very little is known about how individuals absent of neurological abnormalities might perform. As a result, there is no data with which to make comparisons between those being considered for surgery and those who are neurologically normal. The following discussion examines some of the procedural difficulties as well as the complications involving the interpretation of results.

Although many centers perform the IAT, there is a great deal of argument regarding whether or not it is necessary to evaluate both cerebral hemispheres. Some centers perform only a unilateral IAT to either determine language lateralization or to assess whether or not the hemisphere contralateral to the planned surgical resection can support cognitive functioning (a concept known as functional reserve; Chelune, 1995). Wellmer et al. (2005) argue that a bilateral IAT is often redundant. In their retrospective review of 107 cases having undergone bilateral IAT, results from the unilateral IATs performed on the side of intended surgery would have accurately lateralized language in over 80% of cases. Of course that means the 20% of patients who were incorrectly classified would be at a much greater risk of postoperative language impairment following a standard resection. Loddenkemper et al. (2004) reported a case of aphasia following a left-sided hemispherectomy in a 55-year-old man with early childhood onset epilepsy as a result of a motor vehicle accident. Preoperatively the patient was assessed using unilateral IAT of only the left hemisphere. The patient was fully capable of speech and was able to follow commands during the IAT and it therefore was assumed that he was right-hemisphere language dominant. However, the patient's postoperative aphasia would argue that his language representation was bilateral, which may have been detected had he undergone a bilateral IAT (in which case he would presumably be fully able to speak and follow commands following a right hemisphere injection as well). Transient language impairment can be seen in more selective resections as well. We reported on a case in which a 46-year-old patient showed transient moderate to severe naming deficits (2 months, but not 18 months) after a right subtemporal selective amygdalohippocampectomy (Pearson and Packwood, 2012). The patient underwent a unilateral right-sided IAT in which no apparent language deficits were observed. Given the patients postoperative deficits, it was assumed the patient actually had bilateral language representation and that a bilateral IAT would have helped us understand the patient's postoperative risks more fully.

Various drug parameters, for example, amobarbital dosages, concentrations, volume of amobarbital and saline mixture, rate of delivery (steady or incremental), the use of alternative anesthetic agents, and method of delivery (hand or automated injection) also vary widely between epilepsy centers (Rausch et al., 1993). These variations in drug delivery may lead to discrepant findings that make comparison of studies difficult. Drug parameters affect the extent and duration of anesthesia. For instance, a faster rate or larger volume of injection will typically perfuse a more extensive volume of brain, thus compromising more domains of function, whereas smaller volume or slower rate injections will lead to a greater concentration of drug in a more circumscribed area of the brain, possibly leading to more intense or prolonged drug effect.

The desired level of anesthetization is a fine balance that avoids patient obtundation compromising patient test participation on the one hand, while avoiding inadequate hemisphere anesthetization on the other. Ideally, the drug should be dosed and delivered in such a way that it models as closely as possible the effects of surgery. Finally, the drug effects should be enduring enough to allow sufficient time to administer the IAT completely.

Interpretation of the IAT is predicated on the assumption that brain regions supplied by the anterior and middle cerebral arteries are inactivated during anesthesia. However, a neuroimaging study by Hart et al. (1993) suggests that this assumption is not necessarily true. Using single-photon emission computed tomography, they found that there is a great deal of interindividual variability in the regions actually perfused by amobarbital after ICA injection. Further, varying degrees of cross-flow into the contralateral hemisphere or posterior circulation is present depending on the individual vasculature of each patient, which leaves some degree of uncertainty about validity of the results obtained (Hart et al., 1993). Although these individual characteristics can be evaluated to some degree before the IAT using the cerebral angiogram, it is impossible to predict all factors that may complicate interpretation of results. Further, the correlation between contrast medium and amobarbital distribution is limited by differences in methods of injection (e.g., Rausch et al., 1993) and the physical properties of the material injected.

Procedural differences are not the only factors that complicate the interpretation IAT results. The definition of what constitutes language dominance also varies somewhat between epilepsy centers. Benbadis et al. (1998) described two popular ways of determining hemispheric dominance (often referred to as the 'laterality index') during IAT. One method compares hemispheric differences in duration of speech arrest to produce an index varying from 1 (strong left-hemisphere dominance) to −1 (strong right-hemisphere dominance). The second method uses language performance on a specific battery of language tests obtained during anesthetization of each hemisphere to produce a ratio ranging from −100 (strong right-hemisphere dominance) to 100 (strong left-hemisphere dominance). Although these two methodologies would appear to measure similar constructs, this does not seem to be the case. Indeed, in a study of 21 patients by Benbadis et al. (1998), there was no significant correlation between the laterality indices based on speech arrest and performance on a language battery. Many centers have their own 'in-house' methodologies to evaluate language. These tasks range from simple naming, repetition, and comprehension paradigms to more robust assessments that evaluate the full gamut of language abilities. Despite these differences in protocols, a multicenter study of seven epilepsy centers found a relatively high degree of interrater reliability for language lateralization using the IAT (Haut et al., 2002). Additionally, the validity of the IAT to establish hemispheric language lateralization is well established (Dodrill, 1993).

Contribution of the IAT to the Study of Language

The IAT has a number of advantages as a research tool relative to other methodologies. Compared to lesional studies, the IAT

makes it possible to obtain baseline data in the individual to be studied. Additionally, it has allowed researchers to gather large amounts of group data where the 'lesion'-to-assessment interval is relatively homogeneous and thus minimizes confounding variables related to point of recovery. The IAT also allows for the production of a fully reversible lesion, which is useful when it is necessary to study the bilateral functioning of a single subject. Despite these advantages, there are a number of disadvantages the most notable being the brevity of the drug effect, and the possibly limited generalizability of the findings. When gathering clinically relevant information is of paramount importance, there is, within the time constraints of the IAT, little opportunity to manipulate the nature of the stimuli. Furthermore, because the IAT is carried out in individuals with typically chronic cerebral dysfunction (epilepsy), it is unclear if the findings obtained pertain to normal populations, or even to groups with other neurological conditions. Nonetheless, the IAT has allowed researchers to explore not only the intra- and interhemispheric differences in language representation and their correlates, but also, by looking at individuals with brain injury, the functional resiliency and plasticity of brain structures.

Early Studies of Language Lateralization

The first large-scale studies of language using the IAT were carried out at the Montreal Neurological Institute in the mid-1960s. Branch et al. (1964) reported on 119 mostly left handed or ambidextrious patients who underwent either bilateral (97 patients) or unilateral IAT. Of the 71 left-handed individuals assessed, 34 patients (48%) were found to be left-hemisphere language dominant, 27 patients (38%) showed evidence of right-hemisphere language dominance, and 10 patients (14%) were determined to have bilateral language representation. Of the 48 right-handed individuals assessed, 43 (90%) were determined to be left-hemisphere language dominant. The remaining 5 (10%) individuals demonstrated right-hemisphere language dominance, while none of the right-handed subjects showed bilateral language representation. Also assessed was the degree to which early cerebral injury contributed to language representation. Of 27 left-handed patients with left hemisphere cerebral injury before the age of 5 years, 18 (67%) were right-hemisphere language dominant, and 3 (11%) were found to have bilateral language. Those individuals who had not sustained an early cerebral injury were much less likely to have right-hemisphere language dominance. Indeed, of these 44 subjects who had undergone IAT, 28 (64%) were found to be left-hemisphere language dominant, while 9 (20%) and 7 (16%) were found to be right-hemisphere dominant or to have bilateral language representation respectively.

Milner et al. (1966) further described the 18 subjects from the previous series with presumed bilateral language representation. Seventeen (94%) of these individuals were left handed. In six of these cases, no speech arrest was noted following either the right or left hemisphere injections. The remaining patients showed speech arrest and language impairment after both injections.

In a larger study of 371 individuals, Milner (1975) examined the issue of language lateralization and handedness even further. Of 140 right-handed individuals without any evidence of cerebral injury before the age of 5, 134 (96%) were left-hemisphere language dominant. None of these individuals showed bilateral language representation. One-hundred-and-twenty-two left-handed patients were also reported upon, of which 82 (67%) were left-hemisphere dominant, and two groups of 18 (15%) individuals were found to be right-hemisphere language dominant and to lack clear dominance respectively. Of the 78 left-handed patients who had sustained an early cerebral insult, 40 (51%) were right-hemisphere language dominant, 23 (30%) were left-hemisphere language dominant, and 15 (19%) had bilateral language representation. Much less variability was seen in the 31 right-handed individuals who had sustained an injury before the age of 5, as 25 (81%) showed left-hemisphere dominance, and 4 (13%) and 2 (6%) showed right and bilateral dominance, respectively. Thus, left handedness appears to play a role in the likelihood of atypical speech dominance following early cerebral injury.

These early studies on language lateralization provide an excellent framework for understanding language dominance; however, they are not without flaws. First, at the time of publishing, the IAT was in its infancy. Angiography was not performed on any of the patients reported. Thus, the amount of cross-flow and potentially abnormal vasculature was undetermined. Second, a portion of the subjects used in all studies underwent only unilateral injections (although a majority of subjects also underwent direct cortical stimulation and language mapping in an attempt to confirm language lateralization).

The Medical College of Georgia has one of the most comprehensive databases of patients to have undergone the IAT. Lee (2010) recently reported on the IAT results from presurgical candidates evaluated at the Medical College of Georgia and found that approximately 81% of dextral individuals were found to be left-hemisphere language dominant, while 18 and 1% were found to be mixed and right-hemisphere language dominant, respectively. With regards to left-handers, 50% showed left-hemisphere language dominance, while 42 and 8% were mixed and right-hemisphere language dominant, respectively. Although similar in many regards, this particular study shows that bilateral language representation in left hemisphere patients undergoing a presurgical workup for retractable epilepsy may be more common than previously thought.

Predictors of Atypical Language Representation

Atypical language representation has been linked to a variety of patient characteristics. As revealed by the studies of Milner and colleagues, age of injury to the left hemisphere plays an important role in language dominance. Since the resiliency of the brain is maximal during the early years of life, injury at a very early age might induce either the intact or damaged hemisphere to reorganize (i.e., accommodate speech and language zones) so as to compensate for functional losses. Intrahemispheric language reorganization has been found to be more likely than interhemispheric reorganization after 6 years of age. In an early study conducted by Rasmussen and Milner (1977) examining left-handed patients who had sustained a left cerebral hemisphere insult before 6 years of age, 49

(53%) patients displayed right cerebral language dominance and 17 (19%) patients showed evidence of bilateral language representation. Persistence of neurological deficits, most notably hemiparesis, following early brain injury also appears to increase the likelihood of atypical language dominance. Indeed, several studies have suggested that left hemisphere injuries sufficient to cause hemiparesis have been found to be associated with atypical language dominance (Ajersch and Milner, 1983; Woods et al., 1988; Rey et al., 1988).

The Crowding Hypothesis

Interhemispheric language reorganization is more likely to occur at earlier ages; however, there is some evidence that reorganization comes at a cost to cognitive abilities usually subserved by the right hemisphere. This phenomenon is often referred to as the 'crowding' effect (Teuber, 1974). Evaluation of verbal and nonverbal functions during the IAT in epileptic patients with early onset of left hemisphere dysfunction has revealed that patients with atypical (right hemisphere) speech representation perform as well as patients with left hemisphere speech on most measures of language functioning (Strauss et al., 1990). However, patients with atypical language representation often perform more poorly than left-hemisphere language-dominant patients on nonverbal tasks. Lansdell (1969) has reported lower than expected nonverbal IQ scores in individuals with right-hemisphere language dominance. Novelly and Naugle (1985) have also shown that males tend to have lower PIQ than VIQ scores in patients with right-hemisphere language dominance. Other studies have found no differences between groups of individuals with left- and right-hemisphere language dominance (Mateer and Dodrill, 1983; Rausch et al., 1991).

Patterns of Hemispheric Speech Representation

Amobarbital studies of patients with epilepsy have also challenged the notion that hemispheric language dominance falls into discrete categories. Rather, it is more likely that language representation falls on a continuum from strongly left-hemisphere dominant to strongly right-hemisphere dominant, just as handedness seems to fall on a continuum from strongly dextral to strongly sinistral. However, other data suggest that cases might not fall along all points of such a continuum. Specifically, there are four major patterns of hemispheric speech representation. Toward the left-dominant end of the spectrum, there are indeed several clusters representing degrees of left speech dominance; however, individuals with bilateral or right speech dominance tended to center at these respective points, without the gradation observed for left hemisphere representation (Kurthen et al., 1994).

Language Lateralization in Bilingual and Deaf Individuals

A limited number of amobarbital studies have addressed two other important issues, namely the intra- and interhemispheric organization of bilinguals, and the organization of sign language. With respect to studies of bilinguals, several reports have consistently shown that the interhemispheric organization of both languages is complementary (i.e., hemispheric dominance for the two languages is similar). Intrahemispheric organization, however, appears to differ by language. Gomez-Tortosa et al. (1995) report on a bilingual patient who underwent left common carotid amobarbital injection with subsequent language evaluation in both languages. Speech impairment was noted in both languages; however, following the removal of a left parasylvian arteriovenous malformation, the patient developed deficits speaking her native tongue, without any language impairment of her second language.

Because of the potential for deaf individuals to have experienced some degree of cortical reorganization, reports of IAT results in this population are of particular significance. Such studies provide tentative data about whether findings from normal-hearing individuals concerning the organization of sign language apply to deaf individuals. One such study conducted by Damasio et al. (1986) showed evidence of complete left hemisphere dominance in a right-handed individual in both signed English and finger spelling. Thus, evidence of bilateral language representation for sign language was not found.

Alternative Methodologies in the Study of Language and Future Directions

Even though the IAT continues to be the standard presurgical means of assessing lateralization of language, its invasiveness, and the degree of inherent risk involved with the procedure make alternative methodologies very attractive. Some centers have begun using less-invasive strategies such as fMRI or MEG to help determine the functional organization of language in the workup of patients with intractable epilepsy. However, the reliability and predictive value of these tools have only begun to be assessed clinically. Although data suggest a high concordance rate between these less-invasive techniques and IAT in right-handers, that does not seem to be the case with regards to mixed and left-handed individuals (Dym et al., 2011). Additionally, techniques such as fMRI require a great deal of technical expertise and dedicated time in order to make work. Despite this, it is almost a certainty that these various noninvasive strategies will begin to take the place of the IAT language assessment across major epilepsy centers around the world.

See also: Bilingual Aphasia; Epilepsy; Functional Magnetic Resonance Imagery, Analysis of; Handedness and Language, Relationship between; Interhemispheric Interaction in Language and Cognitive Processes; Language Processing, Functional Magnetic Resonance Imaging of; Magnetic Resonance and Molecular Imaging in Psychiatry; Transient Epileptic Amnesia.

Bibliography

Aicardi, J., Shorvon, S.D., 1997. Intractable epilepsy. In: Engels, J., Pedley, T.A. (Eds.), Epilepsy: A Comprehensive Textbook. Lippincott-Raven, Philadelphia, PA, pp. 1325–1331.

Ajersch, M.K., Milner, B., 1983. Handwriting posture as related to cerebral speech lateralization, sex, and writing hand. Human Neurobiology 2, 143–145.

Baxendale, S., Thompson, P., Harkness, W., Duncan, J.S., 2006. Predicting memory decline following epilepsy surgery: a multivariate approach. Epilepsia 47, 1887–1894.

Baxendale, S., Thompson, P., Duncan, J.S., 2008. The role of the Wada test in the surgical treatment of temporal lobe epilepsy: an international survey. Epilepsia 49, 715–719.

Benbadis, S.R., Binder, J.R., Swanson, S.J., Fischer, M., Hammeke, T.A., Morris, G.L., et al., 1998. Is speech arrest during Wada testing a valid method for determining hemispheric representation of language? Brain and Language 65, 441–446.

Branch, C., Milner, B., Rasmussen, T., 1964. Intracarotid sodium amytal for the lateralization of cerebral speech dominance: observations in 123 patients. Journal of Neurosurgery 21, 399–405.

Chelune, G.J., Naugle, R.L., Lüders, H., Sedlak, J., Awad, I.A., 1993. Individual change after epilepsy surgery: practice effects and base rate information. Neuropsychology 1, 41–52.

Chelune, G.J., 1995. Hippocampal adequacy versus functional reserve: predicting memory functions following temporal lobectomy. Archives of Clinical Neuropsychology 10, 413–432.

Damasio, A., Bellugi, E., Damasio, H., Poizner, H., Van Gilder, J., 1986. Sign language aphasia during left-hemisphere amytal injection. Nature 322, 363–365.

Der-Jen, Y., Jiing-Feng, L., Yang-Hsin, S., Ian-Kai, S., Tung-Ping, S., Chien, C., Hsiang-Yu, Y., Chun-Hing, Y., 2006. Selective posterior cerebral artery amobarbital test in patients with temporal lobe epilepsy for surgical treatment. Seizure 15, 117–124.

Dodrill, C.B., 1993. Preoperative criteria for identifying eloquent brain: intracarotid amytal for language and memory testing. Neurosurgery Clinics of North America 4, 211–216.

Dym, R.J., Burns, J., Freeman, K., Lipton, M.L., 2011. Is functional MR imaging assessment of hemispheric language dominance as good as the Wada test?: a meta-analysis. Radiology 261, 446–455.

Gardner, W.J., 1941. Injection of procaine into the brain to locate speech area in left-handed persons. Archives of Neurology and Psychiatry 46, 1035–1038.

Gomez-Tortosa, E., Martin, E., Gaviria, M., Charbel, F., Ausman, J., 1995. Selective deficit of one language in a bilingual patient following surgery in the left parasylvian area. Brain and Language 48, 320–325.

Hart, J., Lewis, P.J., Lesser, R.P., Fisher, R.S., Monsein, L.H., Schwerdt, P., et al., 1993. Anatomic correlates of memory from intracarotid amobarbital injections with technetium Tc 99mm hexamethylpropylene-amine oxime SPECT. Archives of Neurology 50, 745–750.

Haut, S.R., Berg, A.T., Shinnar, S., Cohen, H.W., Bazil, C.W., Sperling, M.R., et al., 2002. Interrater reliability among epilepsy centers: multicenter study of epilepsy surgery. Epilepsia 43, 1396–1401.

Kurthen, M., Helmstaedter, C., Linke, D.B., Hufnagel, A., Elger, C.E., Schramm, J., 1994. Quantitative and qualitative evaluation of patterns of cerebral language dominance. Brain and Language 46, 536–564.

Lansdell, H., 1969. Verbal and nonverbal factors in right-hemisphere speech: relation to early neurological history. Journal of Comparative and Physiological Psychology 69, 734–738.

Lee, G.P., 2010. Neuropsychology of Epilepsy and Epilepsy Surgery. Oxford, New York.

Loddenkemper, T., Dinner, D.S., Kubu, C., Prayson, R., Bingaman, W., Dagirmanjian, A., Wyllie, E., 2004. Aphasia after hemispherectomy in an adult with early onset epilepsy and hemiplegia. Journal of Neurology Neurosurgery and Psychiatry 75, 149–151.

Mateer, C.A., Dodrill, C.B., 1983. Neuropsychological and linguistic correlates of atypical language lateralization: evidence from sodium amytal studies. Human Neurobiology 2, 135–142.

Milner, B., Branch, C., Rasmussen, T., 1962. Study of short-term memory after intracarotid injection of sodium amytal. Transactions of the American Neurological Association 82, 224–226.

Milner, B., Branch, C., Rasmussen, T., 1966. Evidence for bilateral speech representation in some non-right-handers. Transactions of the American Neurological Association 91, 306–308.

Milner, B., 1975. Psychological aspects of focal epilepsy and its neurosurgical management. In: Purpura, D., Penry, J., Walter, R. (Eds.), Advances in Neurology, Neurosurgical Management of the Epilepsies, vol. 8. Raven Press, New York, pp. 299–321.

National Association of Epilepsy Centers: Guidelines for Essential Services, Personnel, and Facilities in Specialized Epilepsy Centers in the United States, 2001. Epilepsia 31 (suppl 1), S1–S12.

Novelly, R.A., Naugle, R.I., 1985. Gender specific effects on VIQ versus PIQ in acquired right-hemisphere speech. Journal of Clinical and Experimental Neuropsychology 7, 627.

Pearson, C.M., Packwood, S., 2012. Naming deficits following right subtemporal selective amygdalohippocampectomy in an individual with left-hemisphere language dominance. The Clinical Neuropsychologist 26, 425–426.

Rasmussen, T., Milner, B., 1977. The role of early left-brain injury in determining lateralization of cerebral speech functions. Annals of the New York Academy of Science 299, 355–369.

Rausch, R., Boone, K., Ary, C.M., 1991. Right-hemisphere language dominance in temporal lobe epilepsy: clinical and neuropsychological correlates. Journal of Clinical and Experimental Neuropsychology 13, 217–231.

Rausch, R., Silfevenius, H., Wieser, H., Dodrill, C., Meador, K., Jones-Gotman, M., 1993. Intraarterial amobarbital procedures. In: Engel, J. (Ed.), Surgical Treatment of the Epilepsies, second ed. Raven Press, New York, pp. 341–357.

Rey, M., Dellatolas, G., Bancaud, J., Talairach, J., 1988. Hemispheric lateralization of motor and speech functions after early brain lesions: study of 73 epileptic patients with intracarotid amytal test. Neuropsychologia 26, 167–172.

Strauss, E., Satz, P., Wada, J., 1990. An examination of the crowding hypothesis in epileptic patients who have undergone the carotid amytal test. Neuropsychologia 28, 1221–1227.

Stroup, E., Langfitt, J., Berg, M., McDermott, M., Pilcher, W., Como, P., 2003. Predicting verbal memory decline following anterior temporal lobectomy (ATL). Neurology 60, 1266–1273.

Teuber, H.L., 1974. Why two brains? In: Schmitt, F.O., Worden, F.G. (Eds.), The Neurosciences: Third Study Program. MIT Press, Cambridge, pp. 71–74.

Wada, J., 1949. A new method for the determination of the side of the cerebral speech dominance: a preliminary report on the intracarotid injection of sodium amytal in man. Igaku Seibutsugaku 14, 221–222.

Wada, J., Rasmussen, T., 1960. Intracarotid injection of sodium amytal for the lateralization of cerebral speech dominance: experimental and clinical observations. Journal of Neurosurgery 17, 266–282.

Wellmer, J., Fernandez, G., Linke, D.B., Urbach, H., Elger, C.E., Kurthen, M., 2005. Unilateral intracarotid amobarbital procedure for language lateralization. Epilepsia 46, 1764–1772.

Woods, R.P., Dodrill, C.B., Ojemann, G.A., 1988. Brain injury, handedness, and speech lateralization in a series of amobarbital studies. Annals of Neurology 23, 510–518.

Intrinsic Motivation: A Cultural Perspective

Qian Wang, The Chinese University of Hong Kong, Hong Kong SAR, China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Western-originated research on intrinsic motivation has well documented its implications for optimal psychological functioning, and the role of autonomy-supportive environmental factors in cultivating intrinsic motivation. There exist theoretical analyses suggesting either a culture-specific or a universal view about intrinsic motivation across cultures. This article reviews these theoretical analyses with reference to the extant empirical research on intrinsic motivation among non-Westerners. It concludes with some reflections to highlight the value of, as well as a warranted caution in, taking a cultural perspective to study intrinsic motivation.

Intrinsic motivation is defined commonly as driving forces of behavior that individuals deem inherent to the activities undertaken – namely, the activities themselves being interesting, enjoyable, and meaningful – in contrast to non-intrinsic forces in terms of incentives or pressures (e.g., rewards or deadlines) that are external to the activities. There has been much research among Westerners documenting (1) how intrinsic motivation may be related to optimal cognitive and socioemotional functioning, and (2) what factors may foster or hinder intrinsic motivation (see Deci and Ryan, 2001). This Western psychology of intrinsic motivation has been recently called into question regarding its universality (vs culture specificity), like many other areas of research in psychology (e.g., Arnett, 2008; Henrich et al., 2010). This article thus takes a cultural perspective to summarize and evaluate theoretical views and empirical evidence regarding intrinsic motivation among non-Westerners (mostly East Asians and particularly Chinese, whom the majority of the extant research concerns).

Original Research on Intrinsic Motivation in Western Cultures

As reviewed comprehensively elsewhere (e.g., Deci and Ryan, 2001; Lepper et al., 2008; Ryan and Deci, 2000; Sansone and Harackiewicz, 2000; Wigfield et al., 2006), intrinsic motivation has been found to be related to an array of optimal psychological functioning among Westerners, ranging from enhanced self-regulation, persistence, and achievement in various domains including academics, sports, and health behavior, to enhanced well-being in terms of positive emotional experience and self-esteem. It has also been evidenced among Westerners that environmental influences such as autonomy-supportive parenting or teaching may foster intrinsic motivation, whereas controlling parenting or teaching may hinder intrinsic motivation. Specifically, autonomy-supportive parents or teachers respect children's or students' sense of autonomy, and thus encourage them to freely express their thoughts and feelings and also give them opportunities to make their own decisions, whereas controlling parents or teachers allow children or students little freedom to express themselves or make decisions on their own, but pressure them to follow through with agendas and meet up standards and expectations set by parents or teachers.

Theoretical Views Regarding Intrinsic Motivation in Non-Western Cultures

How may the aforementioned findings among Westerners with respect to (1) the implications of intrinsic motivation for psychological functioning, and (2) the environmental influences that affect intrinsic motivation, may be generalized to non-Western cultures? There are theoretical views suggesting either culture specificity or universality.

Under the rubrics of individualism vs collectivism (see Oyserman et al., 2002) or independent vs interdependent orientations (see Markus and Kitayama, 1991), Western cultures (prototypically the European American culture in the United States) have been contrasted with Eastern cultures (prototypically the Chinese-originated Confucian culture in East Asia). In particular, it has been contended that the collectivistic or interdependent cultural orientation in East Asian societies such as China, Japan, and Korea prioritizes social relations over personal attributes, and hence, intrinsic motivation (which entails individuals personally finding the activities undertaken interesting, enjoyable, and meaningful) may not play a significant role in East Asians' psychological functioning (see Hau and Ho, 2010; Wigfield et al., 2006). Another theoretical analysis compares the so-called Socratic approach to learning in Western cultures and the so-called Confucian approach to learning in Eastern cultures (Tweed and Lehman, 2002). This analysis reveals a focus on pragmatic outcomes of learning (e.g., a good job, rather than intrinsic motivation for the sake of learning itself) in Eastern cultures, also suggesting the possibility that intrinsic motivation may be of less import in Eastern (vs Western) cultures.

As to the issue concerning environmental influences, it has been argued that autonomy-supportive vs controlling parenting or teaching may have different ramifications for the development of intrinsic motivation among East Asian (vs Western such as European American) children and students (see Iyengar and DeVoe, 2003; Markus and Kitayama, 2003). This argument is also based on the notion that the collectivistic or interdependent cultural orientation discussed earlier

predominates in East Asian societies. Taking controlling parenting or teaching as an example, Eastern Asian (vs Western such as European American) children or students who highly value their relationships with their parents or teachers may be more likely to find the agendas, standards, and expectations set by their parents or teachers meaningful and take them on willingly. Moreover, the so-called Confucian approach to learning requires respect and obedience to authority figures such as parents and teachers and prefers learning to be structured by authority figures rather than self-directed by children or students (Tweed and Lehman, 2002), which also suggests possibly different effects of autonomy-supportive vs controlling parenting or teaching on children's or students' intrinsic motivation in Eastern (vs Western) cultures.

Contrary to the foregoing analyses suggesting culture specificity, there are both Western-originated and indigenous Eastern views suggesting universality in how intrinsic motivation may affect psychological functioning and how environmental factors may influence intrinsic motivation. American-originated self-determination theory (SDT) makes the case that there is a basic psychological need for autonomy and satisfaction of this need promotes optimal human functioning universally (see Deci and Ryan, 2000). Intrinsic motivation is one of the conditions that satisfy the universal need for autonomy by ensuring individuals acting out of their own determination (i.e., out of personally recognized interest, enjoyableness, and meaning of the activities undertaken) and thus may be conducive to psychological functioning regardless of culture. Vice versa, satisfaction of the need for autonomy also cultivates intrinsic motivation by affording individuals the freedom to pursue activities that they personally find interesting, enjoyable, and meaningful. According to SDT, given the universality of the need for autonomy, the role of environmental influences such as autonomy-supportive parenting or teaching in satisfying this need and fostering intrinsic motivation (vs the role of environmental influences such as controlling parenting or teaching in thwarting the need for autonomy and hindering intrinsic motivation) may well be similar across cultures (see Deci and Ryan, 2001).

In analyzing indigenous Chinese notions, Lee (1996) identified intrinsic motivation of learning as an inherent element in Confucianism, which entails "learning for self-realization" (i.e., learning for the purpose to "cultivate oneself as an intelligent, creative, independent, and autonomous" human being). Of particular note is that Lee (1996) observed: "While much of the discussion on Eastern cultures points to collectivism, there is a neglect of 'individualism' or individuality in the Eastern tradition". Wang and Li (2003) echoed: "The Chinese self is often viewed as consisting of a predominant orientation toward interdependence and relatedness… it is sometimes taken to an extreme to mean that the Chinese lack a sense of individual self…" Specifically, Wang and Li (2003) highlighted individualistic goals for learning including loving/enjoying learning, and autonomous regulation of learning including a desire to be free (e.g., "I do not like anyone to watch over my schoolwork") expressed by Chinese students when probed to describe their beliefs about learning (see also Li, 2006). Taken together, these analyses suggest that the beneficial effects of intrinsic motivation on psychological functioning and the role of autonomy-supportive parenting or teaching in promoting intrinsic motivation (vs the role of controlling parenting or teaching in impeding intrinsic motivation) documented among Westerners may also be evident among non-Westerners such as Chinese.

Notably, some pioneering cross-cultural research has also hinted that intrinsic motivation as a key element of the individualistic or independent orientation is not exclusively absent in collectivistic or interdependence-oriented cultures, such that value of autonomy indeed is endorsed universally across cultures (Schwartz, 1992). Moreover, socioeconomic developments that have been happening in many societies around the world are moving traditionally collectivistic or interdependence-oriented cultures toward "increased emphasis on individual freedom-focused values" including intrinsic motivation (Inglehart and Oyserman, 2004).

Empirical Evidence Regarding Intrinsic Motivation in Non-Western Cultures

There is a relatively small body of empirical work investigating intrinsic motivation in non-Western cultures that may help to appraise the culture-specific vs universal views.

Implications of Intrinsic Motivation

Several studies utilized international data sets to examine the relations from intrinsic motivation to secondary school students' academic achievement in Western and Eastern cultures. Chiu and colleagues (Chiu and Chow, 2010; Chiu and Zeng, 2008) analyzed the data set of the Organization for Economic Cooperation and Development's Program for International Student Assessment (OECD-PISA, 2000), and showed that 15-year-old students' intrinsic motivation in reading and mathematics (e.g., "When I read/do mathematics, I sometimes get totally absorbed.") was positively associated with their actual performance in reading and mathematics in 41 countries around the world including Western societies such as the United States of America, the United Kingdom, Germany, and France, and Eastern societies such as Hong Kong, Indonesia, Korea, and Thailand. Zhu and Leung (2011) analyzed the data set of the Trends in International Mathematics and Science Studies (TIMSS 2003), and showed that intrinsic motivation (e.g., "I enjoy learning mathematics.") was positively associated with actual performance in mathematics among East Asian 8th graders in Hong Kong, Japan, Korea, Singapore, and Taiwan and their Western counterparts in Australia, England, the Netherlands, and the United States of America.

Perhaps due in part to Chinese students' outstanding academic achievement (e.g., OECD, 2002), most existing studies concerning intrinsic motivation in non-Western cultures focused on Chinese students' academic motivation. Consistent with the findings from the above mentioned studies based on international data sets, investigations specifically into Hong Kong Chinese primary- and secondary-school students' reading motivation and achievement revealed positive relations from intrinsic motivation (e.g., "I like to read what interests me.") to actual performance (e.g., reading comprehension; Law, 2009; Lin et al., 2012) and engagement

(e.g., amount of time spent on reading or use of reading strategies; Lau, 2009; Law, 2009). When it comes to motivation, engagement, and achievement generally across school subjects, intrinsic motivation has also been found to be positively associated with use of deep-processing learning strategies and self-reported rank of school performance among Hong Kong Chinese secondary school students (Lai et al., 2006). Moreover, positive relations from intrinsic motivation to use of cognitive and metacognitive learning strategies, participation in class, persistence at schoolwork, and actual performance in terms of grades have been evidenced among Taiwan Chinese primary- and secondary school students (d'Ailly, 2003; Shih, 2005, 2008), as well as Korean high school students (Jang et al., 2009). Examining a relative autonomy index, which is a weighted indicator of intrinsic (e.g., "I study the course because it's fun.") relative to nonintrinsic motivation (e.g., "I study the course because I'll get in trouble if I do not."), Vansteenkiste et al. (2005) found that the greater intrinsic relative to nonintrinsic motivation Chinese postsecondary students harbored, the more adaptive their learning experience (e.g., more positive attitude toward the course and better time management) and emotional adjustment (e.g., greater life satisfaction and more positive affect), and the better their exam performance. Also investigating the relative autonomy index, Wang and Pomerantz (2009) showed that the greater intrinsic relative to nonintrinsic motivation Chinese 7th graders harbored upon their entry into junior high school, the better their actual performance in terms of grades one and half years later at the end of 8th grade.

Notably, one study investigated children's intrinsic motivation in a laboratory setting (d'Ailly, 2004). It revealed among Taiwan Chinese 5th and 6th graders nonsignificant relations from interest in a foreign language learning computer game to effort spent on playing the game (as indicated by the computer-recorded amount of time of children's engagement in the game and the computer-recorded number of mouse clicks made by children when playing the game). To summarize, almost all extant empirical research on intrinsic motivation in non-Western cultures focuses on the academic arena and mostly involves East Asian (particularly Chinese) participants. This body of work seems quite consistent (except d'Ailly, 2004) in showing that intrinsic motivation in academic activities is associated with enhanced academic engagement and achievement as well as enhanced emotional well-being among non-Western learners, similar to what has been found among Westerners and lending support to the universal view.

Environmental Influences on Intrinsic Motivation

As to the issue of how environmental factors may affect intrinsic motivation in non-Western cultures, one line of work looks into autonomy-supportive vs controlling parenting or teaching. d'Ailly (2003) revealed a positive relation from mothers' autonomy support (e.g., encouraging children to solve problems in schoolwork on their own, rather than imposing directives upon children) to Taiwan Chinese primary school students' intrinsic motivation. Vansteenkiste et al. (2005) showed that the greater perceived parental autonomy support (e.g., "My parents allow me to choose what to do whenever possible.") vs psychological control (e.g., "My parents are less friendly to me if I do not see things like they do."), the greater Chinese postsecondary students' intrinsic relative to nonintrinsic motivation in learning (i.e., the relative autonomy index). Such a positive relation from perceived parental autonomy support to Chinese students' intrinsic relative to nonintrinsic motivation was evident throughout the first 3 years of college (Pan and Gauvain, 2012). There is also evidence that non-Western students' intrinsic motivation benefits similarly from teachers' autonomy support. Positive associations between perceived autonomy support by teachers (e.g., "My teacher encourages me to ask questions." "My teacher provides me with choices and options.") and students' intrinsic motivation in learning have been found among Taiwan Chinese primary- and secondary school students (d'Ailly, 2003; Shih, 2008), rural Chinese primary school students (Zhou et al., 2009), Korean (Jang et al., 2009), and Russian high school students (Chirkov and Ryan, 2001).

Another line of work scrutinizes particularly the role of personal choice in affecting intrinsic motivation among non-Westerners. In Study 1 conducted by Iyengar and Lepper (1999), 4th graders were let freely decide how much time to spend solving anagrams after a session of working on anagrams in which they were either allowed to make their own choice (i.e., the personal choice condition) or led to believe that their mothers had made the choice for them as to which type of anagrams to do (i.e., the mom choice condition). It was found that Asian American 4th graders spent more time solving anagrams during the free-play period (arguably as an indicator of intrinsic motivation) in the mom (vs personal) choice condition, whereas Anglo American 4th graders spent more time solving anagrams during the free-play period in the personal (vs mom) choice condition. In Study 2 conducted by Iyengar and Lepper (1999), 5th graders were asked to report their liking of a mathematics-related computer game after a session of playing the game in which they were either allowed to make their own choices (i.e., the personal choice condition) or led to believe that their classmates had made the choices for them concerning certain features of the game (i.e., the in-group choice condition). It was found that Asian American 5th graders reported greater liking of the game (arguably as an indicator of intrinsic motivation) in the in-group (vs personal) choice condition, whereas Anglo American 5th graders reported greater liking of the game in the personal (vs in-group) choice condition. These findings have turned out to be perhaps the best known and the most cited support for the culture-specific view on intrinsic motivation – that is, the role of autonomy support (in terms of granting personal choice) in cultivating intrinsic motivation may not be universal.

Bao and Lam (2008) further examined the issue concerning personal choice by studying Hong Kong Chinese primary school students. Their findings depict a refined picture of how personal choice may affect Chinese students' intrinsic motivation: When students feel closely related to their mothers or teachers, there are no differences in their intrinsic motivation in personally vs mother or teacher chosen activities; while when students do not feel closely related to their mothers or teachers,

they tend to have greater intrinsic motivation in personally vs mother or teacher chosen activities. In sum, the majority of the aforementioned studies (except Iyengar and Lepper, 1999) seem to be in line with the universal view that autonomy support from their surrounding environment may help children and students in non-Western cultures to develop intrinsic motivation, similar to what has been found among their Western counterparts.

Concluding Remarks

As reviewed above, the relatively small body of extant empirical research on intrinsic motivation in non-Western cultures may be quite limited in that the majority of it (1) focuses on the academic arena (rather than examining diverse activities in diverse domains of life like in sports and health behavior and in diverse settings like in businesses as research among Westerners has concerned), and (2) involves Eastern Asian (particularly Chinese) participants only (rather than comparing participants from multiple cultures as the few studies based on international data sets did). Hence, even though most of the existing studies seem to support the universal (vs culture-specific) view regarding the implications of intrinsic motivation for psychological functioning and the environmental influences that may affect intrinsic motivation, it is avoided consciously to draw any definite and blanket conclusion here. Instead, several reflections are made below about taking a cultural perspective to understand intrinsic motivation.

First of all, it is critical that cultural analyses are made on the basis of unambiguous conceptualization and operationalization of intrinsic motivation and related constructs, and with full consideration of alternative explanations. As mentioned earlier, in Study 1 conducted by Iyengar and Lepper (1999), time spent on doing the task during the free-play period was taken as an indicator of intrinsic motivation, but this is dubious because it is quite likely that Asian American children would decide to spend more time working on the task in the mom (vs personal) choice condition for nonintrinsic reasons such as to please their mothers. Moreover, while d'Ailly (2004) found nonsignificant relations from Taiwan Chinese children's interest to their effort when they were instructed to play a computer game in the laboratory, it was also found in the study that when explicitly told that there would be no test, Taiwan Chinese children's interest in the game increased significantly. The latter finding suggests a plausible explanation for the former finding such that the laboratory setting may be perceived by Taiwan Chinese children as a test setting in which they tend to exert much effort regardless of interest, while it remains an open question as to the relations from interest to effort among Taiwan Chinese children when they do not feel the pressure of being tested. The foregoing closer look at the studies conducted by Iyengar and Lepper (1999) and d'Ailly (2004) calls for rigorous research across cultures that entails efforts to understand adequately the cultural meanings of the assessments and the procedures, as well as to ensure between-culture measurement equivalence (for reviews on measurement equivalence, see Davidov and Depner, 2011; Vandernberg and Lance, 2000). Also of note is that caution is warranted against mistaking likely universal boundary conditions (e.g., interest plays less of a role in effort expenditure when individuals are under the pressure of being tested) for 'dichotomous' cultural differences in fundamental psychological processes (e.g., intrinsic motivation matters for Western but not Chinese children).

Second, it is essential to theorize and research intrinsic vs nonintrinsic motivation in an inclusive and integrative, rather than exclusive manner – an approach which is inherent in Western-originated theories (see Deci and Ryan, 2001) but often missing in cultural analyses. For instance, the aforementioned findings by d'Ailly (2004) of nonsignificant relations from interest to effort among Taiwan Chinese children are often interpreted as evidence for the culture specificity of intrinsic motivation (i.e., intrinsic motivation may not be relevant in Chinese culture, unlike in Western cultures; see Hau and Ho, 2010), but not as evidence for the universality of the role of nonintrinsic factors (e.g., perceived pressure of being tested) in undermining the functional relevance of intrinsic motivation (see Deci and Ryan, 2001). Also, the findings from the studies conducted by Iyengar and Lepper (1999) that Asian-American children demonstrated greater intrinsic motivation in the mom (vs personal) choice condition (Study 1) or in the in-group (vs personal) choice condition (Study 2) are often highlighted over the findings from the same studies that Asian American children in fact demonstrated greater intrinsic motivation in the personal (vs experimenter) choice condition (Study 1) or in the personal (vs out-group) choice condition (Study 2), thereby downplaying the relevance of personal choice for Asian-American children's intrinsic motivation.

Indeed, the salience of nonintrinsic factors (e.g., pressure of being tested; d'Ailly, 2004) and socially oriented motivation (e.g., great engagement demonstrated by Asian American children in the mom choice or the in-group choice condition; Iyengar and Lepper, 1999) in non-Western cultures has been a major argument for the culture-specific view on intrinsic motivation. However, the functional relevance of intrinsic motivation does not exist exclusively with the functional relevance of nonintrinsic factors (see Covington, 2000) or socially oriented motivation (see Pomerantz et al., 2012). Rather, the social context may facilitate a process termed internalization (see Deci and Ryan, 2001) through which individuals come to find interest, enjoyableness and meaning in activities that do not carry such features initially (see also Sansone and Harackiewicz, 2000). For example, positive relations from attitudes toward family obligations to intrinsic value of schoolwork have been found among adolescents of European, Asian, and Latin American background in the United States (Fuligni and Tseng, 1999).

While stimulated by the apparent contradiction between the theoretical analyses suggesting culture-specificity vs universality regarding intrinsic motivation, this article concludes with the observation that those theoretical analyses (i.e., individualism vs collectivism, independent vs interdependent orientations, or the so-called Socratic vs Confucian approaches to learning, in contrast to Self-Determination Theory or indigenous Chinese notions of learning) may well be compatible when neither culture nor motivation is 'dichotomized'. As pointed

out by other researchers as well, non-Western cultures such as Eastern Asian cultures exemplified by Chinese-originated Confucianism are not monolithic entities of collectivism, interdependent orientation or the so-called Confucian approach to learning, to the exclusion of individualism, independent orientation or the so-called Socratic approach to learning (Gurung, 2002; Lee, 1996; Wang and Li, 2003). Moreover, although clearly differentiable in their antecedents and consequences, intrinsic motivation and nonintrinsic driving forces of behavior coexist in human reality and weigh out dynamically in co-shaping human strivings, depending on individual and situational as well as cultural characteristics (Covington, 2000; Deci and Ryan, 2001; Gurung, 2002; Wigfield et al., 2006). The value of a cultural perspective lies in enriching and refining the understanding of the full humanity by paying due attention to possibly different availability, functionality, accessibility and applicability of various mental attributes and mechanisms across as well as within cultures (Norenzayan and Heine, 2005). Such a value shall not be undermined by obscuring perhaps unwittingly yet quite misleadingly the universality of certain fundamental psychological processes, which quite likely include intrinsic motivation.

Acknowledgment

Preparation of this article was supported by General Research Fund 451809 to Qian Wang from Hong Kong Research Grants Council.

See also: Cross-Cultural Study of Education; Cross-National Comparisons in Education: Findings from Pisa; Interest, Psychology of.

Bibliography

Arnett, J.J., 2008. The neglected 95%: why American psychology needs to become less American. American Psychologist 63, 602–614.

d'Ailly, H., 2003. Children's autonomy and perceived control in learning: a model of motivation and achievement in Taiwan. Journal of Educational Psychology 95, 84–96.

d'Ailly, H., 2004. The role of choice in children's learning: a distinctive cultural and gender difference in efficacy, interest, and effort. Canadian Journal of Behavioral Science 36, 17–29.

Bao, X.-H., Lam, S.-F., 2008. Who makes the choice? Rethinking the role of autonomy and relatedness in Chinese children's motivation. Child Development 79, 269–283.

Chirkov, V.I., Ryan, R.M., 2001. Parent and teacher autonomy-support in Russian and U.S. adolescents: common effects on well-being and academic motivation. Journal of Cross-Cultural Psychology 32, 618–635.

Chiu, M.M., Chow, B.W.Y., 2010. Culture, motivation, and reading achievement: high school students in 41 countries. Learning and Individual Differences 20, 579–592.

Chiu, M.M., Zeng, X., 2008. Family and motivation effects on mathematics achievement: analyses of students in 41 countries. Learning and Instruction 18, 321–336.

Covington, M.V., 2000. Intrinsic versus extrinsic motivation in schools: a reconciliation. Current Directions in Psychological Science 9, 22–25.

Davidov, E., Depner, F., 2011. Testing for measurement equivalence of human values across online and paper-and-pencil surveys. Quality & Quantity 45, 375–390.

Deci, E.L., Ryan, R.M., 2000. The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. Psychological Inquiry 11, 227–268.

Deci, E.L., Ryan, R.M., 2001. Psychology of intrinsic motivation. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Elsevier, Oxford, pp. 7886–7888.

Fuligni, A.J., Tseng, V., 1999. Family obligation and the academic motivation of adolescents from immigrant and American-born families. In: Urdan, T.C., Maehr, M.L., Pintrich, P.R. (Eds.), Advances in Motivation and Achievement, vol. 11. JAI Press, Stamford, CT, pp. 159–183.

Gurung, R.A.R., 2002. Comparing cultural and individual learning tendencies. American Psychologist 57, 145–146.

Hau, K.T., Ho, I.T., 2010. Chinese students' motivation and achievement. In: Bond, M.H. (Ed.), The Oxford Handbook of Chinese Psychology. Oxford University Press, New York, pp. 187–204.

Henrich, J., Heine, S.J., Norenzayan, A., 2010. The weirdest people in the world? Behavioral and Brain Sciences 33, 61–135.

Inglehart, R., Oyserman, D., 2004. Individualism, autonomy, and self-expression: the human development syndrome. In: Vinken, H., Soeters, J., Ester, P. (Eds.), Comparing Cultures, Dimensions of Culture in a Comparative Perspective. Brill, Leiden, the Netherlands, pp. 74–96.

Iyengar, S.S., DeVoe, S.E., 2003. Rethinking the value of choice: considering cultural mediators of intrinsic motivation. In: Murphy-Berman, V., Berman, J.J. (Eds.), Nebraska Symposium on Motivation, Cross-Cultural Differences in Perspectives on the Self, vol. 49. University of Nebraska Press, Lincoln & London, pp. 1–57.

Iyengar, S.S., Lepper, M.R., 1999. Rethinking the value of choice: a cultural perspective on intrinsic motivation. Journal of Personality and Social Psychology 76, 349–366.

Jang, H., Reeve, J., Ryan, R.M., Kim, A., 2009. Can self-determination theory explain what underlies the productive, satisfying learning experiences of collectivistically oriented Korean students? Journal of Educational Psychology 101, 644–661.

Lai, P.Y., Chan, K.W., Wong, K.Y.A., 2006. A Study of Intrinsic Motivation, Achievement Goals, and Study Strategies of Hong Kong Chinese Secondary Students. http://publications.aare.edu.au/06pap/lai06321.pdf (accessed 04.11.13.).

Lau, K.-L., 2009. Reading motivation, perceptions of reading instruction and reading amount: a comparison of junior and senior secondary students in Hong Kong. Journal of Research in Reading 32, 366–382.

Law, Y.-K., 2009. The role of attribution beliefs, motivation and strategy use in Chinese fifth- graders' reading comprehension. Educational Research 51, 77–95.

Lee, W.O., 1996. The cultural context for Chinese learners: conceptions of learning in the confucian tradition. In: Watkins, D.A., Biggs, J.B. (Eds.), The Chinese Learner: Cultural, Psychological, and Contextual Influences. Comparative Education Research Center, Hong Kong, pp. 25–41.

Lepper, M.R., Master, A., Yow, W.Q., 2008. Intrinsic motivation in education. In: Maehr, M.L., Karabenick, S.A., Urdan, T.C. (Eds.), Advances in Motivation and Achievement: Social Psychological Perspectives, vol. 15. Emerald Group Publishing Limited, Bingley, UK, pp. 521–555.

Li, J., 2006. Self in learning: Chinese adolescents' goals and sense of agency. Child Development 77, 482–501.

Lin, D., Wong, K.K., McBride-Chang, C., 2012. Reading motivation and reading comprehension in Chinese and english among bilingual students. Reading and Writing 25, 717–737.

Markus, H.R., Kitayama, S.K., 1991. Culture and the self: implications for cognition, emotion, and motivation. Psychological Review 98, 224–253.

Markus, H.R., Kitayama, S.K., 2003. Models of agency: sociocultural diversity in the construction of action. In: Murphy-Berman, V., Berman, J.J. (Eds.), Nebraska Symposium on Motivation, Cross-cultural Differences in Perspectives on the Self, vol. 49. University of Nebraska Press, Lincoln & London, pp. 1–57.

Norenzayan, A., Heine, S.J., 2005. Psychological universals: what are they and how can we know. Psychological Bulletin 131, 763–784.

OECD, 2002. PISA 2000 Technical Report. OECD, Paris.

Oyserman, D., Coon, H.M., Kemmelmeier, M., 2002. Rethinking individualism and collectivism: evaluation of theoretical assumptions and meta-analyses. Psychological Bulletin 128, 3–72.

Pan, Y., Gauvain, M., 2012. The continuity of college students' autonomous learning motivation and its predictors: a three-year longitudinal study. Learning and Individual Differences 22, 92–99.

Pomerantz, E.M., Cheung, C.S.-S., Qin, L., 2012. Relatedness between children and parents: implications for motivation. In: Ryan, R.M. (Ed.), The Oxford Handbook of Human Motivation. Oxford University Press, New York, pp. 335–349.

Ryan, R.M., Deci, E.L., 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist 55, 68–78.

Sansone, C., Harackiewicz, J.M., 2000. Intrinsic and Extrinsic Motivation: The Search for Optimal Motivation and Performance. Academic, New York, pp. 443–453.

Schwartz, S.H., 1992. Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. Advances in Experimental Social Psychology 25, 1–65.

Shih, S.-S., 2005. Taiwanese sixth graders' achievement goals and their motivation, strategy use, and grades: an examination of the multiple goal perspective. The Elementary School Journal 106, 39–58.

Shih, S.-S., 2008. The relation of self-determination and achievement goals to Taiwanese eighth graders' behavioral and emotional engagement in schoolwork. The Elementary School Journal 108, 313–334.

Tweed, R.G., Lehman, D.R., 2002. Learning considered within a cultural context: confucian and socratic approaches. American Psychologist 57, 89–99.

Vandernberg, R.J., Lance, C.E., 2000. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. Organizational Research Methods 3, 4–69.

Vansteenkiste, M., Zhou, M., Lens, W., Soenens, B., 2005. Experiences of autonomy and control among Chinese learners: vitalizing or immobilizing? Journal of Educational Psychology 97, 468–483.

Wang, Q., Li, J., 2003. Chinese children's self-concepts in the domains of learning and social relations. Psychology in the Schools 40, 85–101.

Wang, Q., Pomerantz, E.M., 2009. The motivational landscape of early adolescence in the United States and China: a longitudinal investigation. Child Development 80, 1272–1287.

Wigfield, A., Eccles, J.S., Schiefele, U., Roeser, R.W., Davis-Kean, P., 2006. Development of achievement motivation. In: Eisenberg, N., Damon, W., Lerner, R. (Eds.), Handbook of Child Psychology, Social, Emotional, and Personality Development, sixth ed., vol. 3. Wiley, New York, pp. 933–1002.

Zhou, M., Ma, W.J., Deci, E.L., 2009. The importance of autonomy for rural Chinese children's motivation for learning. Learning and Individual Differences 19, 492–498.

Zhu, Y., Leung, F.K.S., 2011. Motivation and achievement: is there an East Asian model? International Journal of Science and Mathematics Education 9, 1189–1212.

Introspection: History of the Concept

Kurt Danziger, York University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of introspection, in the sense of a dispassionate observation of one's own mental states, first became prominent in the context of eighteenth-century philosophical empiricism. Although its practice was widely rejected as a basis for building a mental science, some late nineteenth century and early twentieth century psychologists advocated its limited use under laboratory conditions. Behaviorists rejected its use under any circumstances, though admitting verbal reports as sources of scientific data. More recent interest in the experimental study of consciousness and its neural basis has revived questions regarding verbal reports on experienced mental states.

The term 'introspection' refers to the self-observation of mental events. Historically, it is a relatively recent term that achieved prominence in the course of debates about the status and the subject matter of psychology as a science. Critics stressed the incompatibility of self-observation and objective, scientific, observation; defenders proposed versions of introspection that could lead to reliable results.

The Philosophical Background

Origins

There is no documented occurrence of the term 'introspection' before the late seventeenth century. Serious discussion of the phenomenon described by this term began a century later and only gathered momentum in the course of the nineteenth century. Positivists like Auguste Comte now began to contrast the reliable, objective methods of science with the speculations of philosophers supposedly based on self-observation. Introspection, thus defined, would be as old as philosophy, a claim that is hardly supported by an analysis of the relevant early texts.

The origins of the modern concept of introspection are closely tied to the rise of an empiricist philosophy at the end of the seventeenth century. For John Locke and others (1690) all knowledge was based on experience, but there were two sources of experience: the external senses and "the perception of our own mind within us," which he likened to an "internal sense." The modern concept of introspection emerged out of the empiricist philosophers' belief that self-knowledge was based on acts of internal observation that were closely analogous to the acts of observation by which we gain knowledge of the external world.

The Classical Controversy

Objections to the empiricist concept of introspection were launched from two entirely different directions, though there was some convergence on specific criticisms. One set of objections came from the German Idealist philosophers, the other from Scientific Positivists like Comte. Among the first group a beginning was made by Immanuel Kant, who gave several reasons why an empirical study of the psyche could never amount to a natural science. Among these reasons was the consideration that one could not observe one's mental states without thereby altering them. This position was based on a rejection of the empiricist analogy between knowledge of the natural world and self-knowledge. At best, self-observation would yield information about the self as it appears to itself, not as it really is. Idealism led to the development of methods of investigation in the human sciences that were neither introspective nor 'natural–scientific' in character.

'Scientific Positivism,' on the other hand, rejected introspection along with any knowledge gained by methods other than those of the natural sciences. These methods demanded a strict separation of the observer and the objects observed; a separation that, according to Comte, no one could achieve in regard to his own mind. Introspection did not merely modify what was observed, it was actually impossible.

Safeguarding the foundations of philosophical empiricism, John Stuart Mill replied to Comte by offering what came to be the classical defense of introspection: Mental events do not disappear in a flash – they are remembered. What is actually observed during introspection are the memories of recently past mental events, not the events as they occur. In other words, introspection is really retrospection, the examination of memory images. This defense was taken over by William James (1890) and other empiricists, though by the closing years of the century their advocacy of introspection was qualified by warnings about possible errors of memory. Some conceptual clarification was introduced by Franz Brentano (1973/1874), who distinguished between the immediate awareness of ongoing mental events, which he called "internal perception," and "self-observation" (the retrospective attending to past mental events). It can be confusing to refer to both as "introspection."

Introspection and Experiment

The turn to experimental methods of psychological investigation was closely linked to the controversial nature of introspection. Wilhelm Wundt, who played a major role in the launching of experimental psychology, believed that the psychological laboratory provided the only conditions under which reliable introspective data could be gathered (Danziger, 1990). He accepted the distinction between inner perception

and self-observation but unlike his predecessors, totally rejected the latter. Instead, he advocated a control of the conditions of internal perception so as to approximate the conditions of external perception. This could be accomplished in the laboratory by such measures as the repetitive presentation of relatively simple stimuli and the requirement that reactions be immediate and automatic. However, adherence to these conditions would necessarily limit the scope of experimental psychology (and of valid introspection) to the investigation of 'elementary' psychological events, mainly in the areas of sensation and reaction times.

At the beginning of the twentieth century, these restrictions were rejected by a new generation of experimental psychologists who greatly extended the scope of experimental introspection so as to enable the investigation of more complex psychological processes like thinking and problem solving. Reliability of results was now thought to depend on the adoption of certain attitudes while introspecting. But there was no unanimity about the nature of these attitudes. E.B. Titchener (*see* Structuralism), a major exponent of experimental introspection, demanded that introspective descriptions should be in terms of simple, irreducible units and should abstract from any meaning that the stimulus might have. In Europe, members of the Wurzburg School (see Humphrey, 1951) emphasized the importance of clearing one's mind of preconceptions while introspecting (Titchener's requirements were prime examples of such preconceptions). Later, the *Gestalt* psychologists extended this phenomenon-based approach by insisting that perceptions should be described exactly as they appear in daily life, i.e., in terms of interrelated, meaningful patterns.

The investigations of the Wurzburg School raised a new kind of question: What is the relationship between the conscious experience of the introspector and the verbal report he or she makes of this experience? In the empiricist tradition it had always been assumed that the relationship was one of description, not different in principle from the description of objects experienced as located outside oneself. However, it now appeared that some of the verbal reports in introspection experiments could not be regarded as descriptions of inner states, they were much more like 'expressions' of these states, e.g., when someone reports feeling: "Oh no, not that again!" The German term for this is *Kundgabe* which has connotations of broadcasting and proclaiming. This distinction pointed toward a recognition of the communicative features of introspective reports. It also suggested that the result of treating introspective reports purely as descriptions of mental events was likely to be a misleadingly intellectual account of such events.

Behaviorism and Its Aftermath

Divergent attitudes to the practice of experimental introspection led to divergent results among different laboratories. Titchener claimed that sensory images could always be discovered in any thought activity but this was disputed by the Wurzburg School, by Binet in France, and by others. This became known as the "imageless thought controversy." Situations in which the results from one laboratory are at variance with those from other laboratories are far from unknown in the history of science. However, the 'imageless thought controversy' acquired a special significance because its existence was used as evidence for the uselessness of introspection by a new movement in American psychology, namely, 'Behaviorism.'

Early behaviorists like J.B. Watson considered introspection to be entirely bogus, while others, like K.S. Lashley, conceded that it might provide "cues to physiological problems." Certainly, no scientific hypothesis could ever be verified by means of introspection. But in retrospect it seems that the behaviorists drew unwarranted conclusions from the practical difficulties of introspection. (Mackenzie, 1977; Howe, 1991). Both pragmatic and metaphysical reasons were operative here. Pragmatically, it seemed advisable to dispense with a practice that would never yield the kind of knowledge that the science of psychology increasingly aspired to, i.e., knowledge that would be useful in dealing with practical psychological problems on a large scale. Metaphysically, the behaviorists were committed to the elimination of mind from the category of things that could have effects in the real world. Therefore, the study of mental events was pointless.

Behaviorism was more successful in denigrating introspection than in offering an alternative explanation of what actually happened when someone claimed to be introspecting. Its attempts to do so generally involved assumptions about the existence of subtle physical processes, like subliminal movements of the vocal cords and other as yet undetectable examples of "covert" behavior. It was these, rather than mental states, that supposedly gave rise to introspective reports. Behaviorism effectively accepted the original philosophical analogy between internal and external observation. It differed from those it called 'introspectionists' only in regard to the kind of stuff that was the target of internal observation.

There was a strong rhetorical element in the behaviorists' critique of introspection, which led to the construction of historical legends. One such legend was that of a category of "introspective psychology" that comprised prebehaviorist and antibehaviorist psychologies in general. This blurred the significant distinctions that existed among those who were not prepared to outlaw all forms of introspection. None of them had characterized themselves as "introspectionists," and all had often been critical of many forms of introspection.

Debates about the nature of introspection reached their height during the first quarter of the twentieth century and were followed by a period during which the topic was usually regarded as a closed chapter in the history of psychology. Subjects in psychological experiments could still be questioned about the reasons for their responses but their replies were treated as "verbal reports," a form of overt behavior, not a description of internal mental events. Theoretical discussions about the nature of introspection were largely restricted to professional philosophers, beginning with Russell, Wittgenstein, and Ryle, and had little or no influence on scientific psychology. One exception was the latter day behaviorist B.F. Skinner (1953) who continued to regard introspective reports as an effect of various forms of covert behavior, but also added the possibility that at times they might be no more than a claim by a person to have accomplished a particular discriminative task. Disappointing evidence in the area of "covert behavior," as well as the unscientific multiplication of highly speculative hypotheses, eventually robbed the behaviorist position of its early plausibility.

Subsequently, the question of introspection resurfaced in the context of computer models of mind and consciousness (Dennett, 1978 *see* Computers: Impact on the Social Sciences). Here, introspective reports were seen as the output of some kind of language processing system that receives input from various sources, notably, one or more memory systems, an attentional–perceptual system, and perhaps a central 'control' system. The nature of the output will depend on the "information" that is in these systems and on their functional interrelationships. In some versions of the model the final subsystem only has access to the results of previous information processing, not the process itself. This accounts for certain inadequacies of introspective evidence; problems of memory account for others (see Ericsson and Simon, 1993).

Some authors pointed to the fact that all experimental introspection and much everyday introspection occurs in a social context and that introspective reports are always limited by the linguistic resources available to the reporter (Lyons, 1986). The content of introspective reports is therefore circumscribed by social and cultural factors.

In the early twenty-first century, a revived interest in the experimental study of consciousness led to a reopening of questions about introspection (see Overgaard, 2006). Attempts at finding the neural correlates of consciousness had been propelled by major methodological advances on the neural side but were hardly matched by greater conceptual clarity on the other side of the correlation.

See also: Structuralism.

Bibliography

Brentano, F., 1973/1874. Psychology from an Empirical Standpoint. Routledge, London.

Danziger, K., 1990. Constructing the Subject: Historical Origins of Psychological Research. Cambridge University Press, Cambridge.

Dennett, D.C., 1978. Brainstorms. MIT Press, Cambridge, MA.

Ericsson, K.A., Simon, H.A., 1993. Protocol Analysis: Verbal Reports as Data. MIT Press, Cambridge, MA.

Howe, R.B.K., 1991. Introspection: a reassessment. New Ideas in Psychology 9, 25–44.

Humphrey, G., 1951. Thinking. Methuen, London.

James, W., 1890. Principles of Psychology. Holt, New York.

Lyons, W., 1986. The Disappearance of Introspection. MIT Press, Cambridge, MA.

Mackenzie, B.D., 1977. Behaviourism and the Limits of Scientific Method. Routledge, London.

Overgaard, M., 2006. Introspection in science. Consciousness and Cognition 15, 629–633.

Skinner, B.F., 1953. Science and Human Behavior. Collier Macmillan, New York.

Invariance in Statistics

Morris L Eaton, University of Minnesota, Minneapolis, MN, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 7893–7897, © 2001, Elsevier Ltd.

Abstract

Statistical models, typically consisting of a collection of probability distributions, are used to describe patterns of variability that random variables or data may display. Describing the invariance of such models is often done via group theory. Although the mathematical notion of a group is relatively simple, the ideas of group theory provide a very convenient way to describe how statistical models change when random variables are transformed. In particular, statistical models that are closed under a class of transformations are called invariant. Often this invariance has important implications for inferential problems associated with the model. The principle of invariance asserts that whenever a problem is invariant under a group of transformations, then the solution to the problem should also be invariant. Applications of this principle occur in both estimation and hypothesis-testing problems. For example, maximum-likelihood estimators and likelihood-ratio tests are invariant solutions to inference problems when the model is invariant.

Invariance in Statistics

The use of the word invariance in a statistical context most often refers to situations where a statistical model transforms in a suitable way when potential observations are transformed. Describing this context requires the specification of two ingredients. The first is the collection of probability distributions which specify the statistical model. The second is the group of transformations which describes what type of invariance is under consideration. Using the invariance of a statistical model to suggest inferential schemes, such as estimators or tests of hypothesis, is one of the most important applications of invariance reasoning. Such uses of invariance, via concrete examples, form the connective tissue in this article. These examples provide applications of the *principle of invariance* which asserts that invariant problems should have invariant solutions.

The ideas of invariance provide not only the motivation for many statistical procedures but supply the insight and techniques for establishing both distributional and optimality properties of modes of inference. This is most pronounced in the area of linear statistical models – univariate as well as multivariate. For example, in standard regression and analysis of variance models, the affine invariance is reflected in maximum likelihood estimation, in classical F-testing procedures as given by ANOVA tables, and in the structural equation approach to linear models (Fraser, 1968) (*see* Nonparametric Regression). In addition, nonparametric procedures based on ranks and inference based on permutation distributions receive their basic statistical justification from invariance considerations. A full explication of the role of invariance in statistics requires no mean amount of mathematical and statistical theory. However, much of the motivation and pioneering work can be appreciated without all the formal details. Readers wishing just to get the main ideas are advised to skip Section 4 and focus on the concrete examples given in Sections 2, 5, 7, and 8 of what follows.

An Example Due to E. J. G. Pitman

One of the first full fledged invariance arguments to appear in the statistical literature of estimation is due to (Pitman, 1939). Here is a special case of his argument which provided motivation for later work on invariant estimation.

Example 1. Suppose f is probability density function on the real line R^1. For each real number $\theta \in R^1$,

$$f(u-\theta), \quad u \in R^1 \qquad [1]$$

is a density with a *translation parameter* θ. The density in (1) serves as a *translation model* in some very simple experimental situations. Now, suppose that $X_1,\ldots,X_n$ are independent and identically distributed according to (1). To describe the statistical model for the data $X_1,\ldots,X_n$, first form the $n \times 1$ column vector X with coordinates $X_1,\ldots,X_n$. On Euclidean n-space R^n, X has a density

$$p(x|\theta) = \prod_{i=1}^{n} f(x_i - \theta) \qquad [2]$$

where x is an element of R^n with coordinates $X_1,\ldots,X_n$. eqn [2] gives the likelihood of the data $x = (x_1, \ldots, x_n)'$ where the prime denotes vector transpose. The density p in (2) together with $\theta \in R^1$ specifies the *statistical model* for the vector X.

Now, let c be a real number and consider the transformed random variables and parameter

$$\tilde{x}_t = X_i + c, \quad \tilde{\theta} = \theta + c \quad \text{for } i = 1, \ldots, n \qquad [3]$$

If X has a distribution corresponding to θ, then $-\overline{X} = (\tilde{x}_1, \ldots, \tilde{x}_n)'$ has a distribution corresponding to $\tilde{\theta}$. In other words, if X_i are all translated by c, then the model parameter is translated by c. For this reason the model for X is said to be *invariant under translation*.

Next suppose the statistical problem is to estimate θ. Given an estimator

$$T = t(X_1, \ldots, X_n) \in R^1 \qquad [4]$$

let

$$r_T(\theta) = \mathscr{E}_\theta (T-\theta)^2 \qquad [5]$$

be the mean squared error of T. Here, $\mathscr{E}$ denotes expectation under the distribution of X.

In essence Pitman now argued as follows: *Consider an estimate* $t(x_1,\ldots,x_n)$ *of* θ. *From the invariance of the model,* $t(x_1+c,\ldots,x_n+c)$ *is an estimate of* $\theta+c$. *Thus,* $t(x_1+c,\ldots,x_n+c)-c$ *is also an estimate of* θ. *Equating these two estimates yields*

$$t(x_1+c,\ldots,x_n+c) = t(x_1,\ldots,x_n)+c \qquad [6]$$

for all $X_1,\ldots,X_n$ and all c.

An estimate satisfying (6) is called an *invariant estimate*. Pitman asserted that the invariance of the model strongly suggested that only invariant estimates of θ should be considered. He then showed that, within the class of invariant estimators, there is one which is *best* in the sense that it has the smallest mean squared error (see (5)) for every θ. This best invariant estimate is given by the formula

$$t_0(x) = \int_{-\infty}^{\infty} \theta p(x|\theta)\mathrm{d}\theta \Big/ \int_{-\infty}^{\infty} p(x|\theta)d\theta, \quad X \in R^n \qquad [7]$$

To say that the *Pitman estimator* $T_o = t_o(X)$ is *best invariant* means that if T is any invariant estimator, then $r_{To}(\theta) \le r_T(\theta)$ for all $\theta \in R^1$.

The argument leading to the best invariant estimator T_0 is based on two observations. First, the model given by (2) for the random vector X transforms in a very simple way when the data are translated (see (3)). Second, the problem of estimating θ also transforms in a simple way and rather naturally leads to estimates satisfying (6) – namely, the invariant estimates. Generalizing Pitman's argument to cover other inferential problems is much of what invariance is about, although the earlier work of Hotelling (1936) had already introduced invariance into hypothesis testing. The next sections describe in mathematical terms what is meant by an invariant statistical model. It is helpful to keep the Pitman example in mind.

Statistical Models

Consider an experimental situation which will give rise to a vector $X = (X_1,\ldots,X_n)'$ of observations. In essence, a statistical model for this situation is a collection of probability distributions for X, one of which is supposed to be an adequate description of the potential variability that might be expected in X. Because X is an n-dimensional column vector, the distributions for X are defined on R^n. The collection of distributions is often written $\{P(\cdot\,|\theta)|\theta \in \Theta\}$ and the index θ for the collection is a *parameter*. The set Θ is called the *parameter space*. The collection $\{P(\cdot\,|\theta)|\theta \in \Theta\}$ is the *statistical model* for X. Statistical models are typically based on a detailed knowledge of the science underlying the generation of the vector X, the type of data similar experiments have yielded, plus a variety of other considerations including analytic tractability.

In Example 1, the sample space is R^n, the parameter space is $\Theta = R^1$, and $P(\cdot\,|\theta)$ is given by

$$P(A|\theta) = \int_A p(x|\theta)\mathrm{d}x \qquad [8]$$

where A is a subset of R^n. The density $p\,(\cdot|\theta)$ is given in (2). Therefore $\{P(\cdot\,|\theta)|\theta \in \Theta\}$ is the model for the Pitman example when $P(\cdot\,|\theta)$ is given by (8).

Groups of Transformations

A simple but very useful type of algebraic system is a group. In essence, a *group* is a set G together with a binary operation o so that the following hold:

1. if g_1 and g_2 are in G, then $g_1 \circ g_2$ is in G. Thus G is *closed* under the binary operation;
2. there is an element e in G called the *identity* so that $e \circ g = g \circ e = g$ for all $g \in G$;
3. each element g has an *inverse*, g^{-1}, which satisfies $g \circ g^{-1} = g^{-1} \circ g = e$ and
4. for g_1, g_2 and g_3 in G, $g_1 \circ (g_2 \circ g_3) = (g_1 \circ g_2) \circ g_3$. Thus the binary operation is *associative*.

It is common notational practice to delete the binary operation o and write g_1g_2 for $g_1 \circ g_2$. This convention is followed here. A standard example of a group is the set R^n where the binary operation is addition. A more interesting example is the set Gl_n of all $n \times n$ nonsingular matrices where the binary operation is matrix multiplication. For further examples see MacLane and Birkhoff (1967).

The notion of a group of transformations arises when each element g of a group G can also be thought of as defining a bijective (one-to-one and onto) transformation on a set $\mathscr{X}$. Here are two familiar examples.

Example 2. Take G to be the set of real numbers and the binary operation to be addition. Let $\mathscr{X}$ be Euclidean n-space R^n and let x_0 be a fixed vector in R^n. Each $g \in G$ defines a bijection of R^n to R^n given by $x \to x + gx_0$. Here gx_0 is the real number g times the vector x_0. The above notation means that the function defined by g maps x onto $x + gx_0$. For example, if x_0 consists of the vector of all ones, then $x + gx_0$ is just the vector x with g added to each coordinate. This case arose in Example 1 – see (6).

Example 3. Take G to be all $n \times n$ nonsingular matrices with the binary operation of matrix multiplication. With $\mathscr{X} = R^n$, each g defines a bijective transformation $x \to gx$ where gx is the matrix g times the vector $x \in R^n$.

Although not completely obvious in the above two examples, the group operation 'fits together' with function composition when the group element is thought of as a bijective function. In Example 3, what this means is that

$$A_1(A_2x) = (A_1A_2)x \qquad [9]$$

The left hand side of (9) is the 'function' A_1 evaluated at A_2x. Hence $A_1(A_2x)$ is the composition of the two functions A_1 and A_2 in that order. The right hand side of (9) is the 'function' A_1A_2 evaluated at x where A_1A_2 is the group element formed by composing the two group elements A_1 and A_2. This 'fitting together' is what defines a *group of transformations*.

Let $\mathscr{X}$ be a set, G be a group, and assume each $g \in G$ defines a bijective transformation from $\mathscr{X}$ to $\mathscr{X}$. The value of the transformation g at x is written gx. For notational convenience, the action of the function g on $\mathscr{X}$ is written

$$x \to gx \qquad [10]$$

That is, x is mapped onto gx by the function g.

The group G is said to *act on* $\mathscr{X}$ if

a. $g_1(g_2x) = (g_1g_2)x$ *for* $x \in X$ and $g_1, g_2 \in G$.
b. $ex = x$ *for* $x \in \mathscr{X}$ where e is the identity in G.

Conditions (a) and (b) imply that the group inverse of g is the function inverse of g when g is thought of as a bijective transformation. When G acts on $\mathscr{X}$, G is a *group of transformations*.

In statistical applications, it is common for a group G to act on more than one set in the same problem. This happens in Example 1 where G acts on the sample space R^n as noted in Example 2 and G acts on the parameter space $\Theta = R^1$ (3). More complicated examples are given in the next section.

When the group G acts on $\mathscr{X}$, the set $\{gx|g \in G\}$ is called the *orbit* of x. The orbits decompose $\mathscr{X}$ into disjoint sets. Further, G defines an equivalence relation on $\mathscr{X}$, say $\cong$, where $x_1 \cong x_2$ if and only if $x_1 = gx_2$ for some $g \in G$. The partition of $\mathscr{X}$ induced by $\cong$ consists exactly of the orbits (MacLane and Birkhoff, 1967). A function h defined on $\mathscr{X}$ is called *invariant* if $h(x) = h(gx)$ for all $x \in \mathscr{X}$ and $g \in \mathscr{X}$. In other words, h is invariant if and only if h is constant on orbits. Functions h which are invariant and which take different values on different orbits are called *maximal invariants* because such functions identify the orbits in $\mathscr{X}$. Maximal invariants are important objects in the study of invariant hypothesis testing problems, which are discussed later in this article. Here is a simple example of a maximal invariant.

Example 4. Take $\mathscr{X} = R^n$ and $G = O_n$, the group of $n \times n$ orthogonal matrices. Then G acts on R^n by matrix multiplication of vectors

$$x \to gx, \quad x \in R^n \quad \text{and} \quad g \in O_n$$

The orbits in R^n are $S_r = \{x|x'\ x = r^2\}$ where $r \geq 0$. The orbit S_r is just the rim of the sphere of radius r in R^n. Then $h(x) = x'x$ is a maximal invariant.

Invariant Models

Coupling the idea of a statistical model with that of a group acting on a set leads naturally to the notion of an invariant statistical model. Consider a random vector X with values in a sample space $\mathscr{X}$ and a statistical model $\{P(\cdot|\theta)|\theta \in \Theta\}$ for the distribution of X. The parameter θ may also be a vector. Assume a group G acts on both the sample space and the parameter space, although the action of G on the two spaces may be quite different.

Definition. The statistical model for X is *invariant under* G if for all $g \in G$, when X has distribution $P\cdot|\theta)$, then gX has distribution $P(\cdot|g\theta)$, for all $\theta \in \Theta$.

Example 5. Suppose the real valued random variable X is normal with mean μ and variance σ^2. With $\theta = (\mu, \sigma^2)$, X is $N(\mu, \sigma^2)$ when the parameter value is θ. In this example, the sample space $\mathscr{X}$ is R^1, while Θ is

$$\Theta = \left\{(\mu, \sigma^2)\middle|\mu \in R^1, \sigma > 0\right\}$$

The group G is

$$G = \left\{(a, b)|a > 0, b \in R^1\right\}$$

and the binary group operation is

$$(a_1, b_1)(a_2, b_2) = (a_1a_2, a_1b_2, + b_1)$$

The action of G on $\mathscr{X}$ when $g = (a, b)$ is

$$x \to gx = ax + b$$

while the action of G on Θ is

$$\theta = (\mu, \sigma^2) \to g\theta = (a\mu + b, a^2\sigma^2) \qquad [11]$$

Of course, when X is $N(\mu, \sigma^2)$, then $aX + b$ is $N(a\mu + b, a^2\sigma^2)$. This motivates the group action in (11) and shows that the normal model with arbitrary mean and variance is invariant under the action of G.

The above example is far too simple to be of much interest in inferential problems. A natural extension of Example 5 follows and is of direct relevance in both estimation and hypothesis testing – topics to be considered in sections to follow.

Example 6. The notation established in Example 5 is used here. Suppose $X_1, \ldots, X_n$ are independent and identically distributed as $N(\mu, \sigma^2)$. The random column vector $X = (X_1, \ldots, X_n)'$ takes values in R^n and the distribution of X is the n-dimensional normal distribution $N_n(\mu 1, \sigma^2 I_n)$ where I_n is the $n \times n$ identity matrix and $\mathbf{1}$ is the vector of ones in R^n. The parameter space for this model is Θ given in Example 5. Take G to be the group in Example 5, so G acts on Θ as in (11).

For $g = (a,b) \in G$, the transformation on R^n defined by this g is

$$x = (x_1, \ldots, x_n)' \to ax + b1$$

The model for X is $\{P(\cdot|\theta)|\theta = (\mu, \sigma^2) \in \Theta, P(\cdot|\theta) = N_n(\mu\mathbf{1}, \sigma^2 I_n)\}$. This is an invariant model for X when G acts on R^n and Θ in the manner indicated.

The Principle of Invariance

When a statistical model $\{P(\cdot|\theta)|\theta \in \Theta\}$ on a sample space $\mathscr{X}$ is invariant under a group G, it is common statistical practice to search for good inferential procedures, such as estimation schemes or hypothesis tests, which are also invariant. The normative form of this practice leads to the principle of invariance: If a statistical model is invariant under a group, then invariant inference problems should have invariant solutions.

The argument in Example 1 which led to invariant estimators (Eqn. [6]) is a direct application of this principle applied to an estimation problem. Although a precise definition of what an invariant inference problem is would take us far afield, the estimation and testing examples of the following sections will well illustrate the principle. For a thorough and rigorous treatment of invariant inference problems, see Ferguson (1967), Eaton (1983, 1989), Berger (1985), Lehmann (1986), Lehmann and Casella (1998).

Invariant Estimation

Given an observable random vector X whose distribution belongs to an invariant statistical model $\{P(\cdot|\theta)|\theta \in \Theta\}$, the search for a good invariant estimation of the parameter θ is a common inferential problem. Formally an estimator is a function t defined on the sample space $\mathscr{X}$ and taking value in

the parameter space Θ. The abstract version of the argument used by Pitman is the following:

> The model for X is invariant under the group G so if X has distribution $P(\cdot|\theta)$, then gX has distribution $P(\cdot|g\theta)$. If $X = x$ is observed then $t(x)$ is the estimated value of θ. But if $gX = gx$ had been seen rather than x, then $t(gx)$ should be the estimated value of $g\theta$ since gX has distribution $P(\cdot|g\theta)$. Therefore $g^{-1}t(gx)$ should estimate θ. Equating these two point estimates of θ leads to $t(x)g^{-1}t(gx))$, or equivalently

$$t(gx) = gt(x) \quad \text{for } x \in \mathscr{X}, g \in G \qquad [12]$$

Estimators which satisfy (12) are called *invariant* by some authors and *equivariant* by others. It should be emphasized that in (12), the expression gx means the transformation determined by g evaluated at $x \in \mathscr{X}$ where as $gt(x)$ means g evaluated at the point $t(x) \in \Theta$. eqn [12] involves the action of G on both $\mathscr{X}$ and Θ.

Example 7. The notation and assumptions of Example 6 hold here. The maximum likelihood estimators (MLE) for μ and σ^2 are

$$\widehat{\mu}(x) = \frac{1}{n}\sum_{1}^{n} x_i = \bar{x}$$

$$\widehat{\sigma}^2(x) = \frac{1}{n}\sum_{1}^{n}(x_i - \bar{x})^2$$

Then, written in vector form, the MLE for $\theta = (\mu, \sigma^2)$ is

$$t_0(x) = \left(\widehat{\mu}(x), \widehat{\sigma}^2(x)\right), \quad x \in R^n$$

Thus t_0 is an estimator defined on R^n taking values in Θ. This estimator satisfies eqn [12] so that the MLE is invariant.

Example 7 is a special case of a general phenomenon – namely, when statistical models are invariant under a group, then the MLE is an invariant estimator. This is not to say that the MLE is necessarily a good (or the best) invariant estimator, but just that finding MLEs respects the principle of invariance. For example, the Pitman estimator (7) is usually not the MLE.

Invariant Hypothesis Testing

In multiparameter hypothesis testing problems, it is rather unusual to find a testing procedure with optimality properties when compared with all other tests. However, certain tests in multivariate statistical analyses (Muirhead (1982), Eaton (1983), Anderson (1984)) do have optimality properties within the class of invariant tests when the testing problem is invariant.

Again suppose an invariant model $\{P(\cdot|\theta)|\theta \in \Theta\}$, for $X \in \mathscr{X}$ is given where the group G acts on both $\mathscr{X}$ and Θ. Consider a hypothesis testing problem

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta - \Theta_0$$

such that H_0 is the null hypothesis and H_1 is the alternative. This testing problem is called *invariant* if the model under $H_0, \{P(\cdot|\theta)|\theta \in \Theta_0\}$ is also invariant under G. A testing procedure is invariant if it can be expressed in terms of a maximal invariant (see Sect. 4).

Example 8. Let $X_1,\ldots,X_n$ be a random sample from a $N_1(\mu, \sigma^2)$ population and consider the classical Student-t problem of testing H_0: $\mu = 0$ vs. H_1: $\mu \neq 0$. With $\bar{x} = n^{-1}\sum x_i$ and $s^2 = \sum(x_i - \bar{x})^2$, the Student-$t$ test rejects H_0 for large values of $(\bar{x}/s)^2$. This testing problem is invariant under the group $G = \{a | a \in R^1, a \neq 0\}$ where the group operation is multiplication. The action of G on the sample space is

$$x_i \to ax_i, \quad \mathrm{i} = 1, \ldots, n$$

and the action of G on the parameter space is

$$\mu \to a\mu, \quad \sigma^2 \to a^2\sigma^2$$

The Student-t test is not uniformly most powerful (UMP) within the class of all tests but it is UMP within the class of invariant tests (Lehmann, 1986). It is also true that the Student-t test is UMP within the class of unbiased tests – a fact which does not extend to the multivariate case. However, the invariance argument does extend as the following example shows.

Example 9. Let $X_1,\ldots,X_n$ be a random sample from a p-variate normal distribution $N_p(\mu, \Sigma)$. The mean vector μ is in R^p and the $p \times p$ covariance matrix Σ is assumed to be nonsingular. The multivariate analog of the Student-t problem is the hypothesis testing problem H_0: $\mu = 0$ *vs.* H_1: $\mu \neq 0$. Set $\overline{X} = n^{-1}\sum X_i$ and $S = \sum(X_i - \overline{X})(X_i - \overline{X})'$ The likelihood ratio test (LRT) of H_0 *vs.* H_1 rejects H_0 for large values of Hotelling's T^2 statistic (Anderson, 1984) given by up to constants)

$$T^2 = \overline{X}' S^{-1} \overline{X}$$

This is a natural multivariate extension of the Student-t statistic. The above hypothesis testing problem is invariant under the group $G = Glp$ of all $p \times p$ non-singular matrices with matrix multiplication as the group composition. The action of G on the sample space is for $A \in G$

$$X_i \to AX_i, \quad i = 1, \ldots, n$$

and on the parameter space is

$$\mu \to A\mu, \quad \sum \to A \sum A'$$

Hotelling's T^2 test is invariant and within the class of invariant tests, it is UMP.

Example 9 contains elements that are common to a variety of invariant testing problems. First, the LRT is always an invariant test for any testing problem which is invariant. Therefore restricting attention to invariant tests never excludes the LRT. Second, within the class of invariant tests, there is one test that is optimal. This happenstance occurs in some, but certainly not all, invariant testing problems in multivariate analysis (see Eaton, 1983).

Concluding Remarks

The examples given above reflect only the simplest invariance arguments. In fact invariance has found many applications outside the hypothesis testing and estimation contexts

described here. For example, Lehmann (1986) illustrates such arguments in non-parametric problems while other applications can be found in Eaton (1983, 1989), Berger (1985), and Wijsman (1990).

A review of the statistical literature shows that many of the best known statistical procedures have attractive invariance properties. These properties not only appeal to our aesthetic sense of symmetry, but also provide powerful theoretical tools for the study and development of new inferential methods. Although the theory of invariance was well developed by the 1970s, it continues to be an active area of research in mathematical statistics. For example, some recent work in statistical prediction theory has been heavily influenced by invariance considerations (see Eaton and Sudderth, 1999).

See also: Likelihood in Statistics; Multivariate Analysis: Discrete Variables (Overview); Probability Theory: Formal; Probability Theory: Interpretations.

Bibliography

Anderson, T.W., 1984. An Introduction to Multivariate Statistical Analysis, second ed. Wiley, New York.
Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis, second ed. Springer-Verlag, New York.
Eaton, M.L., 1983. Multivariate Statistics: A Vector Space Approach. Wiley, New York.
Eaton, M.L., 1989. Group Invariance Applications in Statistics. The Institute of Mathematical Statistics and The American Statistical Association, Hayward, CA.
Eaton, M.L., Sudderth, W.D., 1999. Consistency and strong inconsistency of group-invariant predictive inferences. Bernoulli 5, 833–854.
Ferguson, T.S., 1967. Mathematical Statistics: A Decision Theoretic Approach. Academic Press, New York.
Fraser, D.A.S., 1968. The Structure of Inference. Wiley, New York.
Hotelling, H., 1936. Relations between two sets of variates. Biometrika 28, 321–377.
Lehmann, E.L., 1986. Testing Statistical Hypotheses, second ed. Springer, New York.
Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation, second ed. Springer, New York.
MacLane, S., Birkhoff, G., 1967. Algebra. Macmillan, New York.
Muirhead, R.J., 1982. Aspects of Multivariate Statistical Theory. Wiley, New York.
Pitman, E.J.G., 1939. The estimation of location and scale parameters of a continuous population of any given form. Biometrika 30, 391–421.
Wijsman, R.A., 1990. Invariant Measures on Groups and their use in Statistics. In: Lecture Notes – Monograph Series, vol. 14. Institute of Mathematical Statistics, Hayward, CA.

Ion Channels and Molecular Events in Neuronal Activity

Bryan Kolb, Department of Neuroscience, University of Lethbridge, Lethbridge, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 12, pp. 7897–7903, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The nervous system's basic mechanism for communicating information is the opening and closing of ion channels. Ion channels are found in the membranes of cells and different channels allow the passage of different ions such as Ca^{2+}, Na^{+}, K^{+}, and Cl^{-}. In neurons, the Ca^{2+} channel mediates neurotransmitter release, which is the basis for interneuronal communication. The amount of Ca^{2+} entering a channel controls the activity level of neurons, with increased Ca^{2+} leading to increased activity. Excitatory synaptic input increases intracellular Ca^{2+} directly, by activating ionotropic and metabotropic receptors. Ca^{2+} influx also stimulates differentiation and directs the course of development in neurons. Large and sustained increase in intracellular Ca^{2+} on the postsynaptic side of the synapse is the primary event for inducing long-term forms of synaptic plasticity. Excess Ca^{2+} can trigger a cascade of events leading to neural death.

Introduction

The most fundamental mechanism for imparting information in the nervous system is the opening of ion channels. This responsibility begins early in life, where opening of ion channels in progenitor cells induces neuronal differentiation. Indeed, ion channels mold the growth and behavior of neurons throughout each phase of the life span. While there are many different classes of ion channels involved in regulating nervous system activity, Ca^{2+} entry serves as the internal ionic message of neural experience. Developmentally, calcium triggers differentiation, guides the growth of neuronal processes, and contributes to the selective elimination of neurons. Ca^{2+} mediates neurotransmitter release, the simplest form of interneuron communication, and its entry is the primary intracellular event for encoding the frequency and number action potentials generated pre- and postsynaptically. Activity- or use-dependent increases in Ca^{2+} serve as an intracellular message for initiating neuronal plasticity, which may be involved in circuit modifications required for learning a memory. Finally, completing the cycle of neuronal ontogeny, Ca^{2+} in excess triggers a cascade of mechanisms leading to neuronal degeneration during senescence and age-related disease.

Activity-Dependent Increases in Neuronal Ca^{2+}

Internal Ca^{2+} is maintained at extremely low concentrations creating an enormous driving force for its movement into the cytoplasm, from internal stores or the extracellular space. Ca^{2+} levels are estimated to be at about 100 nM in neurons at their resting hyperpolarized state, while extracellular Ca^{2+} is in the 2–3 mM range and 10 μM to 1 mM inside the endoplasmic reticulum (Mattson et al., 2000). The gating mechanisms for Ca^{2+} entry into the cell's cytoplasm include plasma membrane voltage-dependent Ca^{2+} channels (VDCC), plasma membrane neurotransmitter or ligand-activated channels, and Ca^{2+} release from internal stores. All of these mechanisms are recruited by neuronal activity; they experience Ca^{2+}-mediated feedback regulation, and are modulated by neurotransmitter activation of G proteins (Lee et al., 1999; Zhang et al., 1998; Zamponi and Snutch, 1998; Zuhlke et al., 1999). Each mechanism of Ca^{2+} entry is compartmentalized to produce localized increases in intracellular Ca^{2+} and site-specific activation of Ca^{2+}-sensitive pathways (Magee et al., 1998; Neher, 1998; Yuste et al., 2000).

Adaptive thresholds for Ca^{2+} entry ensure its action as a second messenger will reflect high levels of neuronal activity. Voltage-gated and neurotransmitter-gated channels interact during periods of neuronal activity to determine the degree of Ca^{2+} entry and thus the direction of Ca^{2+}-mediated adaptation. The classic sequence for initiating Ca^{2+} entry begins with presynaptic release of neurotransmitter acting postsynaptically to depolarize the cell beyond the threshold for triggering an action potential (Colquhoun and Sakmann, 1998). The pioneering work of Hodgkin, Huxley, and Katz demonstrated action potentials were initiated by activating voltage-dependent Na^{+} channels and rapidly terminated by the combined action of Na^{+} channel inactivation and the opposing activation of voltage-dependent K^{+} channels (Hodgkin et al., 1952). Activation of VDCCs is facilitated by the rapid depolarization initiated by Na^{+} channel activation, but it is not an absolute requirement. High- and low-threshold Ca^{2+}-mediated action potentials can be triggered by depolarization independent of Na^{+} channel activation (Magee et al., 1998; Huguenard, 1996). The duration and peak depolarization of the action potential, and consequently the extent of Ca^{2+} entry, is limited by activation of voltage-dependent K^{+} channels and intrinsic kinetics of Ca^{2+} channel inactivation.

Unique patterns of Ca^{2+} can be achieved by the selective expression of ion channels in different populations of nerve cells. For example, the expression of inwardly rectifying K^{+} (K_{ir}) channels raises the threshold for action potential generation by hyperpolarizing the resting membrane potential of striatal neurons, while the expression and interaction of low-threshold Ca^{2+} channels (T-type), Ca^{2+}-activated K^{+} channels, and hyperpolarization-activated K^{+} channels contribute to regular, intrinsic firing in neurons of the substantia nigra and thalamus (Huguenard, 1996; Santoro et al., 2000; Wilson and Callaway, 2000). Regional differences in ion channel distribution can also create microenvironments of Ca^{2+} control within the cell.

Clustering of Na^+, K^+, and Ca^{2+} channels in dendrites create local sites for integrating synaptic input (Magee et al., 1998; Safronov 1999), and direct association of P/Q- and N-type Ca^{2+} channels with vesicle fusion proteins mediates activity-dependent release of neurotransmitter in synaptic terminals (Goda and Sudhof, 1997; Neher, 1998).

The molecular diversity of ion channels present in neurons creates opportunity for exquisite control of Ca^{2+} influx through differences in gating mechanisms and kinetics of activation and inactivation (Catterall, 1995; Ghosh and Greenberg, 1995). Calcium channels are heteroligomeric complexes containing α_1, $\alpha_2\delta$, and β subunits. Ca^{2+} channels contain a single α_1 subunit, which serves as a pore and voltage sensor. Nine α_1 subunit genes have been identified, corresponding to each type of VDCC channel previously categorized by pharmacology. α_{1a} encodes for ω-Aga IVA-sensitive P-type; α_{1B} encodes for ω-Ctx GVIA-sensitive N-type; α_{1C}, α_{1D}, and α_{1S} all encode for dihydropyridine-sensitiveL-type; and α_{1G} and α_{1H} encode for Ni^{2+}-sensitive T-type calcium channels (Zamponi and Snutch, 1998). K^+ channels represent a much larger and more diverse collection of ion channels. At present, nine families of voltage-activated K^+ channels are described (Kaczorowski and Garcia, 1999). Many are derived from homologues of Drosophila channel genes Shaker, Shah, Shaw, and Shal, referred to as K_V1, K_V2, K_V3, and K_V4, respecively, in mammals. They usually form homotetrameric channels, but association with β-subunits can alter channel kinetics (Kaczorowski and Garcia, 1999). A separate class of K^+ channel fundamental in regulating neuronal excitability is Ca^{2+}-activated K^+ channels. They are classified by single channel conductance of 2–25 pS (small or SK channel), 25–100 pS (intermediate or IK channel), and 100–300 pS (large or BK channel) (Kaczorowski and Garcia, 1999). Ca^{2+}-activated K^+ channels serve as an activity-dependent feedback mechanism for reducing neuron-firing frequency and returning the membrane potential back to its resting state.

Excitatory synaptic input increases intracellular Ca^{2+} directly, by activating ionotropic and metabotropic glutamate receptors (mGluRs), and secondarily by driving the membrane potential beyond the voltage-dependent threshold for activating VDCCs. The cooperation of these mechanisms ensures robust Ca^{2+} signals in response to strong synaptic drive. Ionotropic glutamate receptors (AMPA, kainate, and NMDA receptors) are the postsynaptic mediators of fast excitatory synaptic transmission. Functional ionotropic glutamate receptors contain four or five structurally related subunits. The combination of subunits GluR1–4 makes functional AMPA receptors; GluR5–7 and KA1–2 make kainate receptors; and NR1 and NR2A–D make NMDA receptors (Bigge, 1999). Flip or flop splice variants of AMPA receptor subunits influence gating kinetics. AMPA receptors generally have low Ca^{2+} permeability, due to the Q/R site of GluR2 subunits.

GluR2 is the most common subunit found in the brain, but examples of receptors without GluR2 exist and are implicated in Ca^{2+}-dependent changes in neuronal physiology (Conti and Weinberg, 1999; Bigge, 1999). NMDA receptors have greater glutamate affinity than AMPA receptors, slower activation and deactivation kinetics, and they are highly permeable to Ca^{2+}. The Ca^{2+} conductance/Na^+ conductance is >5 for NMDA receptors, 2 for AMPA receptors lacking GluR2, and 0.1 for AMPA receptors containing GluR2 (Conti and Weinberg, 1999). The combined action of greater Ca^{2+} permeability and slower channel inactivation results in dramatically more Ca^{2+} influx for NMDA receptors. NMDA receptors are made of NR1 and NR2 heterodimer complexes and generally composed of four to five subunits. NR1 is essential for channel function and NR2 confers the voltage-dependent Mg^{2+} block and glycine modulation of NMDA receptors (Conti and Weinberg, 1999). The voltage-dependent Mg^{2+} block acts as an activity sensor, allowing large increases in intracellular Ca^{2+} only when the depolarization threshold for releasing the Mg^{2+} block is exceeded. Kainate receptors, like AMPA receptors, lack voltage dependency. Compared to AMPA receptors, kainate receptors contribute little to the peak postsynaptic response, but their slower decay kinetics results in comparable charge transfer (Ferking and Nicoll, 2000). Kainate receptors also mediate presynaptic inhibition of excitatory and inhibitory synapses in the hippocampus.

Slow excitatory synaptic events mediated by activation of mGluRs is another important mechanism for elevating intracellular Ca^{2+}. Group-I mGluRs stimulate phospholipase C via GTP-binding proteins. Phospholipase C cleaves phosphoinositol-(4,5)-bisphosphate (PIP_2) into diacylglycerol (DAG) and inositol 1,4,5-triphosphate (IP_3). DAG activates protein kinase C (PKC) and IP_3 activates IP_3 receptors located on the surface of the smooth endoplasmic reticulum (Fagni et al., 2000; Mattson et al., 2000). IP_3 receptors trigger the release of Ca^{2+} stored in the endoplasmic reticulum at concentrations of 10–100 μM (Mattson et al., 2000). Other neurotransmitters use the same mechanism for releasing Ca^{2+} from intracellular stores, but glutamate is by far the most ubiquitous.

Transient receptor potential (TRP) channels are a recently described type of channel that was first identified as membrane proteins mediating phototransduction in fruit flies (Hardie and Minke, 1992). In mammals, 28 different TRP channels have been identified and grouped into six subfamilies (TRPA, TRPC, TRPM, TRPML, TRPP, and TRPV) (Wu et al., 2010). TRP channels are unique in that they act as cellular sensors for a wide range of physical and chemical stimuli (Clapham, 2003). TRP channels exhibit a nonselective cation permeability and a weak voltage sensitivity. They are abundant on nerve endings in the skin and mouth as well as throughout the brain in both neurons and astrocytes (Scimemi, 2013). They are sensitive to a wide range of stimuli including Ca^{2+}, cyclic nucleotides, phosphorylation potential, and temperature as well as environmental inputs that can be either beneficial or harmful (Zheng, 2013).

Activation of TRP channels changes the membrane potential and translates signaling ions across the cell membrane. A key feature of TRP channels is their sensitivity to multiple types of distinct stimuli. For example, the TRPV1 channel is sensitive to capsaicin, extracellular pH, and heat, whereas TRPM8 is sensitive to menthol and cold (Zheng, 2013). Understanding how each TRP channel responds to its peculiar set of activation energy is poorly understood and a subject of much research.

Activity-Dependent Control of Neuronal Development

Ca^{2+} influx stimulates differentiation and directs the course of development in excitable cells (Spitzer et al., 1994). VDCCs,

ionotropic glutamate receptors, and mGluRs all participate in regulating intracellular Ca^{2+} at key steps in neuronal development. Glutamate activation of NMDA and AMPA receptors induces proliferation of progenitor cells during neurogenesis and their migration to appropriate cell layers (Haydar et al., 2000). Evidence also exists for neurogenesis mediated by Ca^{2+} entry through voltage-dependent L-type Ca^{2+} channels and Ca^{2+} waves mediated by release of Ca^{2+} from intracellular stores. In postmitotic cells, Ca^{2+} continues to direct neuronal differentiation by influencing ion channel expression, neurotransmitter phenotype, and growth factor expression; and neurite VDCCs (L- and N-types) stimulate cell adhesion molecules (CAMs) (NCAM, cadherin, and L1) to promote neurite outgrowth via CAM-mediated cell-to-cell interactions and the binding of CAM to actin-based cytoskeleton (Doherty and Walsh, 1994). Ca^{2+} may initiate neurite extension by binding to calmodulin (CaM) and phosphorylating microtubule associated proteins (MAP2) through the activation of Ca^{2+}/CaM-dependent protein kinase II and IV (Kuhn et al., 1998). A parallel mechanism implicated in MAP phosphorylation is the mitogen-activated protein kinase (MAPK) pathway stimulated by mGluR activation (Quinlan and Halpain, 1996). It is clear that Ca^{2+}-mediated remodeling is based upon a Ca^{2+} set point, since differences in responses exist depending on the Ca^{2+} level achieved and the neuronal population examined. Some neurons experience growth cone collapse with increases in Ca^{2+}, due to severing of microfilaments by Ca^{2+}-activated gelsolin (Neely and Nichols, 1995). Support for this mechanism comes from the observation of downregulated VDCCs coincident with neurite extension and increases in neuronal activity producing decreases in CAM expression.

Protein synthesis is also critical to activity-dependent plasticity and growth, a process Ca^{2+} mediates through interacting steps of protein phosphorylation (Ghosh and Greenberg, 1995). Activity-dependent increases in Ca^{2+} stimulate the expression of immediate early genes like c-fos and c-jun, as well as delayed 'effector' genes for neurotrophins and, in a mechanism of positive feedback, genes for VDCCs, NMDA receptors, and AMPA receptors (Ghosh and Greenberg, 1995). The pattern of gene expression is dependent on the pattern of neuronal activity and Ca^{2+} accumulation. Intracellular pathways appear to act in parallel to alter gene expression, but they may respond differentially to changes in Ca^{2+} concentration. A final point of convergence for Ca^{2+}-stimulated pathways is phosphorylation of cAMP-responsive element-binding protein (CREB), which activates RNA polymerase by forming a complex with CREB-binding protein. The independent pathways proposed to trigger this mechanism include Ca^{2+} activation of Ras/MAPK pathway, Ca^{2+} activation of adenylate cyclase and subsequently protein kinase A (PKA) and Ca^{2+} binding to CaM, and activation of Ca^{2+}/CaM kinase II and IV (Ghosh and Greenberg, 1995). A clear division does not exist between the actions of Ca^{2+} and neurotrophins in driving activity-dependent change in neurons. Activity not only increases intracellular Ca^{2+}, but it also causes the release of neurotrophins, and neurotrophins and Ca^{2+} produce similar responses in nerve cells (McAllister et al., 1999).

Ca^{2+} and neurotrophins both trigger the Ras/MAPK pathway to change gene expression, neurotrophins by binding to membrane-bound tyrosine kinase (trk) receptors, and Ca^{2+} by activating Ras directly (Heumann, 1994). Neurotrophins have both direct rapid and slow modulatory effects on ion channels involved in excitation (Benedikt and Poo, 1996). Neurotrophins cause rapid depolarization by activating Na^{+} channels linked to trk receptors and they modulate glutamate-activated channels and intrinsic voltage-activated channels to increase intracellular Ca^{2+} (Jarvis et al., 1997; Kafitz et al., 1999). Neurotrophins also increase neuronal excitability and thus Ca^{2+} influx through acute inhibition of $K_{ir}3.1$, $K_{ir}3.4$, and $K_{ir}2.1$ (Wischmeyer et al., 1998). Neurotrophins also act presynaptically to increase neurotransmitter release (McAllister et al., 1999). Testament to the intimate relationship between neurotrophins and Ca^{2+}, neurotrophins produce long-term changes in Ca^{2+} entry by increasing the expression of VDCCs, NMDA receptors, and AMPA receptors, and activity-dependent increases in Ca^{2+} produce a reciprocal increase in neurotrophin expression (McAllister et al., 1999).

The gearing of neurons toward excitation during development is clearly illustrated by the unique behavior displayed by neurotransmitter-activated channels early in development. Embryonic synaptic AMPA receptors are highly Ca^{2+} permeable and miniature excitatory postsynaptic current (mEPSC) amplitude can be more than twofold larger than at mature synapses (Rohrbough and Spitzer, 1999). Postnatal neurons preferentially express flip variants of AMPA receptors, which increase the depolarization-mediated Ca^{2+} signal by slowing desensitization (Conti and Weinberg, 1999). It should be noted, however, that many synapses show only functional NMDA receptors and only with activity to 'silent' AMPA receptors appear (Petralia et al., 1999). Early postnatal development also shows a shift in the molecular structure and physiological properties of NMDA receptors. NR2B subunits are preferentially expressed early in development and they show less voltage-dependent block by Mg^{2+} and slower kinetics, properties which lower the threshold for Ca^{2+} entry during synaptic excitation. With maturity, there is a shift to NR2A and NR2C subunits, which increase the Mg^{2+} block and speed up the NMDA-activated response. This postnatal disposition toward synaptic excitation is augmented by a developmental shift in the property of $GABA_A$ receptor-mediated inhibition (Ben-Ari et al., 1997). Rather than inhibiting action potential generation and VDCC activation, as is seen in adults, immature $GABA_A$ synapses cause slow depolarizations triggering bursts of action potentials and Ca^{2+} influx through VDCCs (Ben-Ari et al., 1997). Later in development, the Cl^{-} gradient changes resulting in hyperpolarizing $GABA_A$ responses and an inhibition of action potential generation and Ca^{2+} channel activation (Ben-Ari et al., 1997).

Activity and Synaptic Function

Ca^{2+} entering presynaptic terminals during action potentials stimulates the fusion of synaptic vesicles with the presynaptic membrane and the release of neurotransmitter into the synaptic cleft. P/Q- and N-type Ca^{2+} channels are the primary VDCCs subserving this function in the central nervous system (Goda and Sudhof, 1997). Entrance of Ca^{2+} through individual channels is proposed to form microdomains, where Ca^{2+} levels may reach as high as 100 μM (Neher, 1998). It is

proposed that 'core complex' proteins of the synaptic fusion machinery maximize their proximity to these domains by binding syntaxin to Ca^{2+} channels (Goda and Sudhof, 1997). Prior to and independent of Ca^{2+} influx, synaptic vesicles dock to the presynaptic membrane in active zones and undergo priming to make them ready for releasing neurotransmitter with Ca^{2+} influx (Goda and Sudhof, 1997). The docking and priming steps are mediated by an interaction of a 'core complex' of NSF attachment protein receptors (SNAREs) located on the vesicle (v-SNARE) and attachment SNAREs located on the target membrane (t-SNARE). The core complex or SNARE is made up of syntaxin and synaptosome-associated membrane protein of 25 kDa (SNAP-25), which are localized in the synaptic plasma membrane, and the vesicular protein synaptobrevin (also known as vesicle associated membrane protein) (Goda and Sudhof, 1997). The process of vesicle fusion and neurotransmitter release is believed to be mediated by the interaction of the synaptotagmin Ca^{2+} sensor with the 'core complex.' Synaptotagmin I and II are the leading candidates as Ca^{2+} sensors, although other presynaptic proteins with Ca^{2+} C_2-binding domains are proposed to serve the same function (Goda and Sudhof, 1997).

The mechanism for fusion is proposed to be a rapid electrostatic or conformational change in the fusion complex mediated by Ca^{2+} binding to synaptotagmin (Goda and Sudhof, 1997). The vesicle collapses into the membrane and is retrieved away from the active zone by a clathrin-dependent mechanism involving the GTPase dynamin I (Cremona and De Camilli, 1997).

Invasion of the presynaptic terminal by multiple action potentials results in a short-term change in the probability of vesicle fusion and the release of neurotransmitter.

Depending on the number and timing of the action potentials, and thus the degree and duration of Ca^{2+} increase, the processes of facilitation (time constant less than 1 s), augmentation (time constant of several seconds), and potentiation (time constant of a few minutes) are initiated (Zucker, 1999). One mechanism proposed for mediating short-term synaptic enhancement is the binding of Ca^{2+} to synapsin I. Synapsin I is a Ca^{2+}-regulated ATPase, which binds to actin and may facilitate vesicle transport via the actin matrix (Zucker, 1999). Other mechanisms proposed to participate in short-term synaptic plasticity include facilitating actions of synaptotagmin and Ca^{2+}-dependent facilitation of presynaptic Ca^{2+} currents (Zucker, 1999). An explanation for the longest form of short-term plasticity, PTP, is based on Ca^{2+} dumping by mitochondria following mitochondrial sequestration of Ca^{2+} during spike trains (Zucker, 1999). New studies indicate another key role for Ca^{2+} in mediating short-term synaptic plasticity lies in a clathrin-independent form of rapid endocytosis of synaptic vesicles. The proposed mechanism suggests that Ca^{2+} binds to synaptophysin during repeated excitation and Ca^{2+}-stimulated synaptophysin binds to dynamin I to produce rapid endocytosis (Daley et al., 2000).

Large and sustained increases in intracellular Ca^{2+} on the postsynaptic side of the synapse is the primary event for inducing long-term forms of synaptic plasticity or long-term depression (LTD) and long-term potentiation (LTP). Critical to determining whether LTP or LTD occurs are the extent and kinetics of the Ca^{2+} signal and the molecular mechanism underlying the Ca^{2+} increase (Zucker, 1999). These factors and other mechanisms vary considerably between brain areas. Based on cortical and cerebellar studies, a few general themes emerge. Ca^{2+} entry through NMDA receptors is a critical event in the induction of LTP and the Mg^{2+} block of these receptors sets a stimulus and depolarization threshold. Weaker stimuli, which activate VDCCs, often lead to the induction of LTD (Zucker, 1999). LTD also requires the activation of group I mGluRs at many synapses and the subsequent activation of nitric oxide synthetase (Daniel et al., 1998). VDCCs can augment the NMDA Ca^{2+} signal if back-propagating action potentials precede activation of NMDA receptors (Zucker, 1999). This receptor-based scheme depends on the parameters of the Ca^{2+} signal, and thus examples of NMDA receptor-dependent LTD and mGluR-dependent LTP also exist (Nicoll et al., 1998).

The mechanisms used to induce synaptic plasticity in adult animals often follow the same pathways utilized for development and growth in the nervous system. Phosphorylation plays a critical role in LTP and it is triggered by Ca^{2+} influx (Soderling and Derkach, 2000). Ca^{2+} activation of Ca^{2+}/CaM kinase and Ca^{2+} activation of the Ras/MAPK pathways play distinct roles in the maintenance phase of LTP.

Autophosphorylation of Ca^{2+}/CaM kinase makes it Ca^{2+} independent and creates a mechanism for maintaining long-term phosphorylation. Active Ca^{2+}/CaM kinase increases AMPA receptor responses and may be involved in the unmasking of silent receptors and the trafficking of new receptors to the extracellular membrane (Malinow et al., 2000; Soderling and Derkach, 2000). Keeping Ca^{2+}/CaM kinase phosphorylated and active is opposed by the activity of protein phosphatase I. Herein lies another critical step in LTP: inhibiting protein phosphates. Phosphatase inhibition is mediated by Ca^{2+} activation of adenylate cyclase, activation of PKA, and phosphorylation of inhibitor 1 (I1). I1 shuts off protein phosphatase I and Ca^{2+}/CaM kinase thus stays phosphorylated and active (Soderling and Derkach, 2000). Another critical component of LTP expression is protein production. Ca^{2+}/CaM kinase is also implicated in gene expression as well as Ca^{2+}-activation of Ras/MAPK and, just like observed developmentally, phosphorylation of CREB (Platenik et al., 2000). Another similarity to development is the clear involvement of neurotrophins in LTP expression. Neurotrophins directly modulate Ca^{2+} physiology and they activate the Ras/MAPK pathway (McAllister et al., 1999; Klintsova and Greenough, 1999).

Ca^{2+} in Aging and Degeneration

The role of Ca^{2+} in development and plasticity requires flawless control over its intracellular concentration. Aging in the brain, and in many other tissues, occurs in part through a disruption in the ability to maintain Ca^{2+} homeostasis (Thibault et al., 1998). A system level theory for explaining aging poses that chronic exposure to glucocorticoid stress hormones creates a disruption of Ca^{2+} homeostasis (Thibault et al., 1998). In the hippocampus, stress hormones and/or aging increases the expression of L-type Ca^{2+} channels (Thibault et al., 1998). The increased Ca^{2+} signal suppresses hippocampal neuron activity by stimulating Ca^{2+}-activated K^+ channels and, consequently, the hippocampus becomes less effective in its role of

terminating the stress response. Circulating levels of stress hormones therefore progressively increase over an animal's lifetime and consequently the deleterious effects of Ca^{2+} overload also become progressively worse. Elevated Ca^{2+} in aged neurons is proposed to create dysfunctional plasticity and to reduce the threshold for Ca^{2+}-mediated cell death.

Disruption of Ca^{2+} homeostasis is also proposed to be a critical event in apoptosis and necrosis produced in many age-related neurodegenerative disorders (Mattson, 1998; Mattson et al., 2000; Thibault et al., 1998). Periods of excessive excitation, as might occur during stroke or in response to β-amyloid stimulation, cause abnormally high levels of Ca^{2+} ions to enter the cell. The candidate mechanisms for the pathological increase in Ca^{2+} influx include VDCCs, NMDA receptor-gated channels, and mGluR activation and downstream release of Ca^{2+} from internal stores. The consequent Ca^{2+} overload stimulates Ca^{2+}-sensitive proteases and it plays a significant role in the generation of reactive oxygen species (ROS) (Mattson et al., 2000). High levels of cytoplasmic Ca^{2+} cause mitochondrial Ca^{2+} loading, which stimulates mitochondrial ROS production and release of apoptotic factors (Mattson, 1998).

The flood of Ca^{2+} triggers abnormal activation of calpain and free radicals activate caspases, both of which are viewed to create excessive proteolysis. In addition, Ca^{2+} sequestration by mitochondria during excessive excitation reduces the production of ATP with a critical cost of shutting down ATPases responsible for maintaining the cell's membrane potential (Albin and Greenamyre, 1992). Plasma membrane Na^{+}–K^{+}-ATPases require huge amounts of energy and their reduced activity leads to cellular depolarization, activation of VDCCs, and the removal of Mg^{2+} voltage-dependent block from NMDA-gated channels.

Ca^{2+} also triggers cytoplasmic production of ROS through the activation of Ca^{2+}-dependent enzymes. Ca^{2+}-activated proteases convert xanthine dehydrogenase into xanthine oxidase, which produces superoxide anions and H_2O_2. High levels of cytoplasmic Ca^{2+} triggered by NMDA-receptor activation also activate nitric oxide synthetase (NOS). NO produced by the activation of NOS is a free radical, as well as a ubiquitous mediator of biological processes (Dawson and Dawson, 1998). NO turns into a much more potent oxidant, however, after it reacts with the superoxide anion under conditions of oxidative stress to produce highly reactive peroxynitrite. The ROS generated by these mechanisms cause lipid, protein, and nucleic acid oxidation and ultimately cell death.

See also: Neural Plasticity; Neurotransmitters.

Bibliography

Albin, R.L., Greenamyre, J.T., 1992. Alternative excitotoxic hypotheses. Neurology 42, 733–738.

Ben-Ari, Y., Khazipov, R., Leinekugel, X., Caillard, O., Gaiarsa, J.L., 1997. $GABA_A$, NMDA and AMPA receptors: a developmentally regulated 'menage a trois'. Trends in Neurosciences 20, 523–529.

Benedikt, B., Poo, M-m, 1996. Fast actions of neurotrophic factors. Current Opinion in Neurobiology 6, 324–330.

Bigge, C.F., 1999. Ionotropic glutamate receptors. Current Opinion in Chemical Biology 3, 441–447.

Catterall, W.A., 1995. Structure and function of voltage-gated ion channels. Annual Reviews in Biochemistry 64, 493–531.

Clapham, D.E., 2003. TRP channels as cellular sensors. Nature 426, 517–524.

Colquhoun, D., Sakmann, B., 1998. From muscle endplate to brain synapses: a short history of synapses and agonist-activated ion channels. Neuron 20, 381–387.

Conti, F., Weinberg, R.J., 1999. Shaping excitation at glutamatergic synapses. Trends in Neurosciences 22, 451–458.

Cremona, O., De Camilli, P.D., 1997. Synaptic vesicle endocytosis. Current Opinion in Neurobiology 7, 323–330.

Daniel, H., Levenes, C., Crepel, F., 1998. Cellular mechanisms of cerebellar LTD. Trends in Neurosciences 21, 401–407.

Daley, C., Sugimori, M., Moreira, J.E., Ziff, E.B., Llinas, R., 2000. Synaptophysin regulates clathrin-independent endocytosis of synaptic vesicles. Proceedings of the National Academy of Sciences USA 97, 6120–6125.

Dawson, V.L., Dawson, T.M., 1998. Nitric oxide in neurodegeneration. Progress in Brain Research 118, 215–229.

Doherty, P., Walsh, F.S., 1994. Signal transduction events underlying neurite outgrowth stimulated by cell adhesion molecules. Current Opinion in Neurobiology 4, 49–55.

Fagni, L., Chavis, P., Ango, F., Bockaert, J., 2000. Complex interactions between mGluRs, intracellular Ca^{+} stores and ion channels in neurons. Trends in Neurosciences 23, 80–88.

Ferking, M., Nicoll, R.A., 2000. Synaptic kainate receptors. Current Opinion in Neurobiology 10, 342–351.

Ghosh, A., Greenberg, M.E., 1995. Calcium signaling in neurons: molecular mechanisms and cellular consequences. Science 268, 239–246.

Goda, Y., Sudhof, T.C., 1997. Calcium regulation of neurotransmitter release: reliably unreliable? Current Opinion in Cell Biology 9, 513–518.

Hardie, R.C., Minke, B., 1992. The trp gene is essential for a light-activated Ca^{2+} channel in Drosophila photoreceptors. Neuron 8, 643–651.

Haydar, T.F., Wang, R., Schwartz, M.L., Rakic, P., 2000. Differential modulation of proliferation in the neocortical ventricular and subventricular zones. Journal of Neuroscience 20, 5764–5774.

Heumann, R., 1994. Neurotrophin signalling. Current Opinion in Neurobiology 4, 668–679.

Hodgkin, A.L., Huxley, A.F., Katz, B., 1952. Measurements of current–voltage relations in the membrane of the giant axon of Loligo. Journal of Physiology (London) 116, 424–448.

Huguenard, J.R., 1996. Low-threshold calcium currents in central nervous system neurons. Annual Reviews in Physiology 58, 329–348.

Jarvis, C.R., Xiong, Z.G., Plant, J.R., Churchill, D., Lu, W.Y., MacVicar, B.A., MacDonald, J.F., 1997. Neurotrophin modulation of NMDA receptors in cultured murine and isolated rat neurons. Journal of Neurophysiology 78, 2363–2371.

Kaczorowski, G.J., Garcia, M.L., 1999. Pharmacology of voltage gated and calcium-activated potassium channels. Current Opinion in Chemical Biology 3, 448–458.

Kafitz, K.W., Rose, C.R., Thoenen, H., Konnerth, A., 1999. Neurotrophin-evoked rapid excitation through TrkB receptors. Nature 401, 918–921.

Klintsova, A.Y., Greenough, W.T., 1999. Synaptic plasticity in cortical systems. Current Opinion in Neurobiology 9, 203–208.

Kuhn, T.B., Williams, C.V., Dou, P., Kater, S.B., 1998. Laminin directs growth cone navigation via two temporally and functionally distinct calcium signals. Journal of Neuroscience 18, 184–194.

Lee, A., Wong, S.T., Gallagher, D., Li, B., Storm, D.R., Scheuer, T., Catterall, W.A., 1999. Ca^{2+}/calmodulin binds to and modulates P/Q calcium channels. Nature 399, 155–159.

Magee, J., Hoffman, D., Colbert, C., Johnston, D., 1998. Electrical and calcium signaling in dendrites of hippocampal pyramidal neurons. Annual Review of Physiology 60, 327–346.

Malinow, R., Mainen, Z.F., Hayashi, Y., 2000. LTP mechanisms: from silence to four-lane traffic. Current Opinion in Neurobiology 10, 352–357.

Mattson, M.P., 1998. Free radicals, calcium, and the synaptic plasticity-cell death continuum: emerging roles of the transcription factor NF-B. International Review of Neurobiology 42, 103–168.

Mattson, M.P., LaFerla, F.M., Chan, S.L., Leissring, M.A., Shepel, P.N., Geiger, J.D., 2000. Calcium signaling in the ER: its role in neuronal plasticity and neurodegenerative disorders. Trends in Neurosciences 23, 222–229.

McAllister, A.K., Katz, L.C., Lo, D.C., 1999. Neurotrophins and synaptic plasticity. Annual Review of Neuroscience 22, 295–318.

Neely, M.D., Nichols, J.G., 1995. Electrical activity, growth cone motility and the cytoskeleton. Journal of Experimental Biology 198, 1433–1446.

Neher, E., 1998. Vesicle pools and Ca^{+} microdomains: new tools for understanding their roles in neurotransmitter release. Neuron 20, 389–399.

Nicoll, R.A., Oliet, S.H., Malenka, R.C., 1998. NMDA receptor-dependent and metabotropic glutamate receptor-dependent forms of long-term depression coexist in CA1 hippocampal pyramidal cells. Neurobiology of Learning and Memory 70, 62–72.

Petralia, R.S., Esteban, J.A., Wang, Y.X., Partridge, J.G., Zhao, H.M., Wenthold, R.J., Malinow, R., 1999. Selective acquisition of AMPA receptors over postnatal development suggests a molecular basis for silent synapses. Nature Neuroscience 2, 31–36.

Platenik, J., Kuramoto, N., Yoneda, Y., 2000. Molecular mechanisms associated with long-term consolidation of the NMDA signals. Life Sciences 67, 335–364.

Quinlan, E.M., Halpain, S., 1996. Emergence of activity-dependent, bidirectional control of microtubule-associated protein MAP2 phosphorylation during postnatal development. Journal of Neuroscience 16, 7627–7637.

Rohrbough, J., Spitzer, N.C., 1999. Ca^{2+}-permeable AMPA receptors and spontaneous presynaptic transmitter release at developing excitatory spinal synapses. Journal of Neuroscience 19, 8528–8541.

Safronov, B.V., 1999. Spatial distribution of Na^{+} and K^{+} channels in spinal dorsal horn neurons: role of the soma, axon and dendrites in spike generation. Progress in Neurobiology 59, 217–241.

Santoro, B., Chen, S., Luthi, A., Pavlidis, P., Shumyatsky, G.P., Tibbs, G.R., Siegelbaum, S.A., 2000. Molecular and functional heterogeneity of hyperpolarization-activated pacemaker channels in the mouse CNS. Journal of Neuroscience 20, 5264–5275.

Scimemi, A., 2013. A TRP among the astrocytes. Journal of Physiology 591 (Pt 1), 9–15.

Soderling, T.R., Derkach, V.A., 2000. Postsynaptic protein phosphorylation and LTP. Trends in Neurosciences 23, 75–80.

Spitzer, N.C., Gu, X., Olson, E., 1994. Action potentials, calcium transients and the control of differentiation of excitable cells. Current Opinion in Neurobiology 4, 70–77.

Thibault, O., Porter, N.M., Chen, K.C., Blalock, E.M., Kaminker, P.G., Clodfelter, G.V., Brewer, L.D., Landfield, P.W., 1998. Calcium dysregulation in neuronal aging and Alzheimer's disease: history and new directions. Cell Calcium 24, 417–433.

Wilson, C.J., Callaway, J.C., 2000. Coupled oscillator model of the dopaminergic neuron of the substantia nigra. Journal of Neurophysiology 83, 3084–3100.

Wischmeyer, E., Doring, F., Karschin, A., 1998. Acute suppression of inwardly rectifying $K_{ir}2.1$ channels by direct tyrosine kinase phosphorylation. Journal of Biological Chemistry 273, 34063–34068.

Wu, L.J., Sweet, T.B., Clapham, D.E., 2010. International union of basic and clinical pharmacology. LXXVI. Current progress in the mammalian TRP ion channel family. Pharmacological Review 62, 381–404.

Yuste, R., Majewska, A., Holthoff, K., 2000. From form to function: calcium compartmentalization in dendritic spines. Nature Neuroscience 3, 653–659.

Zamponi, G.W., Snutch, T.P., 1998. Modulation of voltage dependent calcium channels by G proteins. Current Opinion in Neurobiology 8, 351–356.

Zheng, J., 2013. Molecular mechanism of TRP channels. Comprehensive Physiology 3, 221–242.

Zhang, S., Ehlers, M.D., Bernhardt, J.P., Su, C.-T., Huganir, R.L., 1998. Calmodulin mediates calcium-dependent inactivation of N methyl-D-aspartate receptors. Neuron 21, 443–453.

Zucker, R.S., 1999. Calcium- and activity-dependent synaptic plasticity. Current Opinion in Neurobiology 9, 305–313.

Zuhlke, R.D., Pitt, G.S., Deisseroth, K., Tsien, R.W., Reuter, H., 1999. Calmodulin supports both inactivation and facilitation of type calcium channels. Nature 399, 159–162.

IPAT (Impact, Population, Affluence, and Technology)

Geoffrey McNicoll, Population Council, New York, NY, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 7903–7906, © 2001, Elsevier Ltd.

Abstract

For a given population group, the impact of human activity on the natural environment can be decomposed into the per capita impact × population size. Per capita impact, in turn, can be decomposed into the per capita level of consumption or 'affluence' × the environmental impact per unit of consumption under the existing technology. These statements can be combined as the identity $I = PAT$ (Impact = Population × Affluence × Technology). In this form the equation can be attributed to Ehrlich and Holdren (1971, 1972), and is sometimes called the Ehrlich–Holdren equation. IPAT refers to the equation and to the broader view of population–environment relations that it connotes.

'Impact' here indicates a persisting rather than a transitory environmental effect. In Commoner's (1971) definition, it is an external intrusion into an ecosystem which tends to degrade the system's natural capacity for self-adjustment. For population–environment interactions the impacts of most concern are those that reduce the system's capacity to provide environmental services. An example of an environmental impact is a country's carbon dioxide emissions, which degrade the environmental service provided by the atmosphere in regulating heat radiation from the earth's surface. The decomposition in that case would be population × per capita GDP × the 'carbon intensity' of the economy. At a given level of affluence and carbon intensity, emissions rise in proportion to population.

IPAT as a Causal Relationship

The IPAT equation is more widely employed not as an identity, but as a causal relationship, intended to point to the environmental damage that results from one or more of population size, a high-consumption lifestyle, and environmentally destructive technology, each amplifying the others. Implicitly, it asserts that these factors can together be seen as the main human causes of degradation. That assertion is debatable: in particular, social organizational and behavioral factors would often warrant separate scrutiny as causes of degradation rather than being subsumed within A and T.

This causal use is mostly confined to elementary or rhetorical discourse about the environment. A familiar instance is in contrasting the environmental impacts of population growth in rich and poor regions. If per capita energy use is a proxy for $A \times T$, the environmental impact of an additional person in America or Europe consuming at the average level for those societies would be an order of magnitude greater than that of an average person in Asia or Africa – a disparity cited in many popular writings that seek to assign responsibility for degradation or resource depletion. The relationship is also used simply to organize a discussion, sometimes expressed in a more general functional form: $I = f(PAT)$.

If P, A, and T were independent of each other, the multiplicative relationship would be equivalent to an additive relationship among growth rates. In the case of carbon dioxide, the growth rate of emissions would equal the sum of the growth rates of the three components. However, P, A, and T are not independent of each other. For any defined population and environment, they are variables in a complex economic, demographic, and sociocultural system. Each also has major distributional dimensions and is a function of time. Consumption – or any other measure of human welfare – is an output of this system; environmental effects, both intended and unintended, are outputs as well. And even at the global level the system is not autonomous: it is influenced by 'natural' changes in the environment and by environmental feedbacks from human activity.

One kind of dependency among P, A, and T follows from the common situation of diminishing returns: successive increments to population may do more than proportional environmental damage if they require recourse to ever more powerful technologies to maintain consumption levels. The IPAT equation might then be seen as illustrating a best-case scenario. But other kinds of dependency would tend to lessen the impact. For example, as affluence increases, so typically does a society's demand for a clean environment and wilderness preservation – that is, there is a high-income elasticity of demand for environmental quality. This would support a policy response emphasizing conservation or restoration, and the development and adoption of 'clean' technologies. Similarly, population growth rates generally fall, even to negative values, with rising incomes. If such interactions are acknowledged, the PAT decomposition retains heuristic and pedagogical value but does not bear much empirical scrutiny. In particular, it cannot resolve disputes on the relative contributions of factors responsible for environmental degradation.

For this latter task, Preston (1994) has proposed looking at the variances and covariances of the growth rates of I, P, A, and T over different regions or countries. In an illustrative analysis of carbon emission data for major world regions over 1980–90, he finds that population growth makes only a minor contribution to the total emission variance; the major contributors are the growth of A and T, with a substantial offsetting effect from the interaction of A and T. Given the 50% or so increase in global population projected for this century, however, the future role of population growth in carbon emissions is nonetheless important. Detailed studies of this relationship include Bongaarts (1992) and O'Neill et al. (2001, 2010). Important too, of course, are the demographic consequences of any

resulting climate change, such as those working through shifts in food production, disease patterns, and sea levels.

A more sophisticated version of the IPAT formula, more suited to statistical estimation, includes parameters describing the elasticities of impact with respect to the component factors:

$$I = aP^bA^cT^de$$

where a, b, c, and d are to be estimated and e is an error term. This has been termed the STIRPAT equation (York et al., 2003). It has been used in a number of studies in preference to the original formulation, although the fitting process often encounters problems of multicollinearity. At the country level there is some evidence that the population elasticity b (interpreted as the percentage change in impact from a 1% increase in population) is close to 1 for carbon emissions, but may be higher for some other pollutants (see Cole and Neumayer, 2004).

Environmental Spillovers

With a strong management regime or effective community norms and sanctions, population growth in a region need not adversely affect the local environment. Access to a resource can be rationed, behavior strictly governed, or the growth can be diverted elsewhere, usually to urban centers. Ultimately, limits on the environmental services that can be drawn on would be reached and thereafter degradation would probably ensue. Well short of that limit, however, conditions of rapid economic or political change can undermine a management regime or erode norms and sanctions, with adverse environmental results. Excessive deforestation can often be traced to such institutional breakdowns (or to ill-considered efforts at institutional reform) rather than to population growth. In other cases, a resource may have been so abundant that no management or sanctions were needed: this is a setting where the familiar 'tragedy of the commons' may unfold as the number of claimants to the resource or their exploitative abilities increase (see Hardin, 1968).

Environmental effects of human activities may extend well beyond the region of origin, as with the case of acid rain or watershed destruction. In addition to such literal spillovers, trade and other economic relationships among societies may allow adverse effects to be 'exported' to another region: a poorer society may be more willing to incur environmental damage in return for economic gain, or be less able to prevent it. The concept of a community's 'ecological footprint' offers one means of accounting for such displaced effects in terms of the total area required to sustain the community's population and level of consumption (see Wackernagel and Rees, 1996; *see* Carrying Capacity of the Environment). Where the environmental effects involve degradation of a global commons, as with carbon emissions into the atmosphere, an effective response requires devising an international regulatory regime that takes account of economic level, demographic scale, and technological capacity.

At the global level, a calculation of human environmental impact, in the spirit of IPAT, is that of the extent of human domination of various components of the global ecosystem. Examples would include the proportion of the land surface transformed by human activity, the anthropogenic share of atmospheric carbon dioxide, and the proportion of marine fisheries that are fully exploited or depleted (Vitousek et al., 1997). A summary impact measure proposed by Vitousek et al. (1986) is the proportion of total organic material produced through photosynthesis that is used directly by humans or used in human-dominated ecosystems, which they calculated at the time to be around 30% of the terrestrial total and 2% of the marine total. For a critical discussion of the bases of such calculations, however, see Smil (2013: Chapter 11).

Nonlinearities in Environmental Systems

An evident difficulty for impact assessment is the nonlinearity of environmental systems. The Millennium Ecosystem Assessment (2005), a major global study, foresees a future of accelerating, possibly abrupt, and potentially irreversible changes in ecosystems, with important consequences for human well-being. Ecosystems may have multiple equilibria; they may exhibit cyclical or chaotic behavior. Small human populations and primitive technology do not ensure a small environmental impact: in prehistoric times minuscule populations of hunter-gatherers may have been responsible for some mammalian extinctions and possibly for large-scale ecological change following the introduction of new fire regimes (see Martin and Klein, 1984). Ecosystems may be resilient under the pressure of human activity until a point is reached at which there is sharp discontinuous change (see Holling, 1986). Kasperson et al. (1995) identify a series of thresholds in nature–society trajectories as human activity in a region intensifies beyond sustainability: first a threshold of impoverishment, then endangerment, and finally criticality – the stage at which human wealth and well-being enter an irreparable decline.

The thresholds may not be evident in advance. Under some greenhouse scenarios, for example, gradual ocean warming leads to an abrupt shift in the course of ocean currents – possibly over a period as short as a few decades – with major effects on weather patterns and the climates of coastal states. As systems models of the environment, such as general circulation models of global climate, expand in size and sophistication their value in prediction is expected to increase. However, they may always be defeated in this task by the inherent uncertainties created by nonlinearities in nature itself. Application of the precautionary principle, aimed at keeping human environmental impact below the threshold of sustainability, may lower the likelihood of a surprise outcome but cannot eliminate it.

Subjectivity of Impact Assessments

The transformation of natural environments by human activity is typically identified with damage or degradation. In the case of agricultural practices leading to erosion, salination, laterization, etc., that assessment is clearly justified. The spread of crop and forest monocultures is seen as ecologically damaging through the consequent loss of biodiversity and increase in vulnerability. Some human interventions, however, can be seen as positive in ecological terms – and not only in cases of restoration of earlier damage. It is also important to note that scientific judgment in characterizing environmental impact does not necessarily accord

with public sentiment. Radically transformed environments – say, English hedged fields or Balinese rice terraces – may be regarded as aesthetically pleasing, provided their ecological stability (at a new level) is maintained. Indeed, Dubos (1980) notes that some of today's most admired landscapes are products or by-products of human activity, representing 'areas deforested, swamps drained, hillsides gouged of their stones and sand.' Prior to the Romantic age, 'wilderness' tended to evoke feelings of horror more than of beauty.

Thus, it is necessary to distinguish between an environmental impact of human activity as construed by ecologists (summarized in stylized fashion by the *I* of the IPAT equation) and the subjective assessment of that effect. In common usage the environment is nature viewed through human eyes, valued by human preferences. Preferences differ among people and over a person's life; they may be socially constructed; they tend to adapt to or track realities. An environmental impact that is seen by ecologists as an effectively irreversible change in a natural system – for example, extinction of a species or loss of a tropical rainforest, or the much earlier loss of temperate forests – may offer apparently compensating gains at the time and in the locality; the full costs and accompanying regret may only be felt much later or by groups far removed geographically. Moreover, regret wanes as people's expectations adjust to the new environmental circumstances. An active and timely public response in limiting adverse impacts calls for heightened ecological awareness at the various different levels of society and fuller realization that ecological stability and environmental services cannot be taken for granted.

See also: Sustainable Development.

Bibliography

Bongaarts, J., 1992. Population growth and global warming. Population and Development Review 18, 299–319.

Cole, M.A., Neumayer, E., 2004. Examining the impact of demographic factors on air pollution. Population and Environment 26, 5–21.

Commoner, B., 1971. The Closing Circle. Knopf, New York.

Dubos, R.J., 1980. The Wooing of Earth. Athlone, London.

Ehrlich, P., Holdren, J., 1971. The impact of population growth. Science 171, 1212–1217.

Ehrlich, P., Holdren, J., 1972. One-dimensional ecology. Bulletin of the Atomic Scientists 28, 16–27.

Hardin, G., 1968. The tragedy of the commons. Science 162, 1243–1248.

Holling, C.S., 1986. The resilience of ecosystems: local surprise and global change. In: Clark, W.C., Munn, R.E. (Eds.), Sustainable Development of the Biosphere. Cambridge University Press, Cambridge, UK.

Kasperson, J.X., Kasperson, R.E., Turner II, B.L. (Eds.), 1995. Regions at Risk: Comparisons of Threatened Environments. United Nations University Press, Tokyo.

Martin, P.S., Klein, R.G. (Eds.), 1984. Quaternary Extinctions: A Prehistoric Revolution. University of Arizona Press, Tucson, AZ.

Millennium Ecosystem Assessment, 2005. Ecosystems and Human Well-Being: Synthesis. Island Press, New York.

O'Neill, B.C., MacKellar, F.L., Lutz, W., 2001. Population and Climate Change. Cambridge University Press, Cambridge, UK.

O'Neill, B.C., et al., 2010. Global demographic trends and future carbon emissions. Proceedings of the National Academy of Sciences 107, 17521–17526.

Preston, S.H., 1994. Population and the Environment. International Union for the Scientific Study of Population. Belgium, Liege.

Smil, V., 2013. Harvesting the Biosphere: What We Have Taken from Nature. MIT Press, Cambridge, MA.

Vitousek, P.M., Ehrlich, P.R., Ehrlich, A.H., Matson, P.A., 1986. Human appropriation of the products of photosynthesis. BioScience 36, 368–373.

Vitousek, P.M., Mooney, H.A., Lubchenco, J., Melillo, J.M., 1997. Human domination of Earth's ecosystems. Science 277, 494–499.

Wackernagel, M., Rees, W., 1996. Our Ecological Footprint: Reducing Human Impact on the Earth. New Society Publishers, Philadelphia.

York, R., Rosa, E.A., Dietz, T., 2003. STIRPAT, IPAT and ImPACT: analytic tools for unpacking the driving forces of environmental impact. Ecological Economics 46, 351–365.

Irrationality: Philosophical Aspects

Richard Samuels, The Ohio State University, Columbus, OH, USA
Stephen P Stich, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Over the past four decades, human reason and rationality has been among the most intensely investigated topics in psychology, cognitive science, and economics. At the heart of this debate is a view of human rationality, often associated with the *Heuristics and Biases* tradition, on which much of our reasoning and decision making is normatively problematic because it relies on heuristics and biases rather than rational principles. In this article we describe briefly some of the evidence that has been invoked in support of this contention and consider some of the more prominent critical responses, especially from evolutionary psychology and dual-process theorists.

The Aristotelian Tradition vs the Heuristics and Biases Program

There is a philosophical tradition, reaching at least as far back as Aristotle, which maintains that man is a rational animal. According to this Aristotelian traditional, human beings are *intrinsically* rational, though of course even the most ardent defender of this view would admit that people's decisions and thought processes are sometimes very irrational indeed. When people are tired, or drunk, or in the grip of rage, they can think and act in ways that no account of rationality would condone. To accommodate this obvious fact, contemporary defenders of human rationality often appeal to a distinction between *competence* and *performance* similar to the one invoked by contemporary linguists (Chomsky, 1965; Samuels et al., 1999). The claim that human beings are rational is interpreted to mean that normal humans posses a rational reasoning competence – a tacit knowledge of normatively appropriate principles for reasoning and decision making. Under ordinary circumstances these principles are reflected in people's thinking and their actions, though sometimes, for various reasons, our performance can fail to reflect this underlying rational competence.

In the early 1970s, however, Amos Tversky, Daniel Kahneman, and a number of other psychologists began reporting findings suggesting that the Aristotelian view of human rationality is seriously mistaken. What these studies demonstrated is that even under quite ordinary circumstances where fatigue, drugs, and strong emotions are not factors, people reason and make decisions in ways that appear systematically to violate familiar canons of rationality on a broad array of problems. Those first surprising studies sparked the growth of an enormously influential research program whose impact has been felt in a wide range of disciplines including psychology, economics, political theory, and medicine. The expression *Heuristics and Biases* is often used as a label for this tradition, since it holds that much of our reasoning and decision making relies on Heuristics and Biases rather than rational principles.

In the sections to follow we will offer some examples of the sort of experimental findings that many have interpreted as showing that the reasoning and decision-making competence of normal humans is *far* from rational. (For more comprehensive reviews of the literature see Dawes, 1988; Baron, 2008; Gilovich et al., 2002.) But first we need to say what standards of rationality are being applied by those who think the experimental findings demonstrate widespread irrationality. Though writers in the Heuristics and Biases tradition rarely offer an explicit and general theory of rationality, the most plausible reading of their work is that they are assuming some version of what Edward Stein (1996) calls the *Standard Picture*. On this view, to be rational is to reason and make decisions in accordance with principles that are based on the rules of logic, probability theory, and decision theory. Sometimes, of course, it is far from clear how these theories are to be applied – a problem that we will return to in Section Do These Studies Have 'Bleak Implications' for the Extent of Human Rationality?. For the moment, however, we follow the lead of those who work in the Heuristics and Biases tradition and assume that the irrationality manifested in the experiments we describe is intuitively obvious.

Irrationality in Reasoning

The Conjunction Fallacy

In what has become perhaps the most famous experiment in the Heuristics and Biases tradition, Tversky and Kahneman (1982) presented people with the following task.

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Please rank the following statements by their probability, using 1 for the most probable and 8 for the least probable.

(1) Linda is a teacher in elementary school.
(2) Linda works in a bookstore and takes Yoga classes.
(3) Linda is active in the feminist movement.
(4) Linda is a psychiatric social worker.
(5) Linda is a member of the League of Women Voters.
(6) Linda is a bank teller.
(7) Linda is an insurance sales person.
(8) Linda is a bank teller and is active in the feminist movement.

In a group of naive subjects with no background in probability and statistics, 89% judged that statement (8) was more

probable than statement (6) despite the obvious fact that one cannot be a *feminist* bank teller unless one is a *bank teller*. When the same question was presented to statistically sophisticated subjects – graduate students in the decision science program of the Stanford Business School – 85% made the same judgment! Results of this sort, in which subjects judge that a compound event or state of affairs is more probable than one of the components of the compound, have been found repeatedly since Tversky and Kahneman's pioneering studies, and they are remarkably robust. This pattern of reasoning has been labeled *the conjunction fallacy*.

Base-Rate Neglect

On the familiar Bayesian account, the probability of a hypothesis on a given body of evidence depends, in part, on the prior probability of the hypothesis. However, in a series of experiments, Kahneman and Tversky (1973) showed that subjects often seriously undervalue the importance of prior probabilities. One of these experiments presented half of the subjects with the following 'cover story'.

> A panel of psychologists have interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. You will find on your forms five descriptions, chosen at random from the 100 available descriptions. For each description, please indicate your probability that the person described is an engineer, on a scale from 0 to 100.

The other half of the subjects were presented with the same text, except the 'base rates' were reversed. They were told that the personality tests had been administered to 70 engineers and 30 lawyers. Some of the descriptions that were provided were designed to be compatible with the subjects' stereotypes of engineers, though not with their stereotypes of lawyers. Others were designed to fit the lawyer stereotype, but not the engineer stereotype. And one was intended to be quite neutral, giving subjects no information at all that would be of use in making their decision. Here are two examples, the first intended to sound like an engineer, the second intended to sound neutral:

> Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles.
>
> Dick is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues.

As expected, subjects in both groups thought that the probability that Jack is an engineer is quite high. Moreover, in what seems to be a clear violation of Bayesian principles, the difference in cover stories between the two groups of subjects had almost no effect at all. The neglect of base-rate information was even more striking in the case of Dick. That description was constructed to be totally uninformative with regard to Dick's profession. Thus the *only* useful information that subjects had was the base-rate information provided in the cover story. But that information was entirely ignored. The median probability estimate in both groups of subjects was 50%.

Before leaving the topic of base-rate neglect, we want to offer one further example illustrating the way in which the phenomenon might well have serious practical consequences. Here is a problem that Casscells et al. (1978) presented to a group of faculty, staff, and fourth-year students at Harvard Medical School.

> If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs? ____%

Under the most plausible interpretation of the problem, the correct Bayesian answer is 2%. But only 18% of the Harvard audience gave an answer close to 2%. Forty-five percent of this distinguished group completely ignored the base-rate information and said that the answer was 95%.

Irrationality in Decision Making

Framing

In a study that is widely believed to illustrate a deeply irrational feature of human decision making, Tversky and Kahneman (1981) presented a group of subjects with the following problem:

> Imagine that the US is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:
>
> If Program A is adopted, 200 people will be saved.
>
> If Program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved.

A second group of subjects was given an identical problem, except that the programs were described as follows:

> If Program C is adopted, 400 people will die.
>
> If Program D is adopted, there is a 1/3 probability that nobody will die and a 2/3 probability that 600 people will die.

On the first version of the problem most subjects chose Program A. But on the second version most chose Program D, despite the fact that the outcome described in A is identical to the one described in C.

The Ultimatum Game

Consider a game that works as follows: There are two players, an 'Allocator' and a 'Recipient' whose identities are unknown to each other. At the beginning of the game the Allocator is given a sum of money (say $10) and asked to divide it between himself and the other player. The only restriction is that the Allocator must offer *something*. So, if the smallest unit of money available is $1, then the Allocator must offer at least this much to the Recipient. Once the Allocator has divided the money, the

Recipient gets to decide whether to accept or reject the amount offered. If the Recipient accepts, each player gets to keep the sum specified by the Allocator. If the Recipient rejects the offer, then both players receive *nothing*. The game is not repeated, and the players know from the outset that they will play only once.

On one widely accepted analysis of the problem, it is never rational for a Recipient to reject an offer, since it is better to get *some* money rather than none. If this is right, then it is always rational for the Allocator to offer the Recipient the smallest amount of money possible, since rational Recipients will always accept the offer, and this will maximize the amount of money the Allocator receives. However, in many studies in both developed and developing countries, subjects very rarely make these choices. Indeed, in these studies it is often the case that the modal offer is to divide the money equally. Moreover, Recipients frequently reject offers of less than 25% of the total, preferring instead to go home empty-handed. (For an overview of the literature see Thaler, 1992.)

The Public Goods Game and Costly Punishment

The Public Goods Game is another game that has been widely used to study economic decision making. In one version of this game, four subjects are each given a sum of money (say $10) and are offered the opportunity of contributing some or all of this money to a group fund. Decisions are made without discussion or negotiation, and subjects are told that they will play the game only once with this set of subjects. After all the subjects have had an opportunity to contribute to the fund, the money in the fund is multiplied by a fixed amount (50% for example) and then distributed equally among all the players, regardless of their contribution. On one analysis, a rational agent concerned to maximize his economic gain should contribute nothing to the group fund, since that will enable him to keep his initial $10 and to get 25% of whatever the other subjects have contributed to the fund. However, in many studies it has been shown that this is rarely what subjects do. The majority of subjects contribute something to the group fund, and many subjects contribute more than 50% of the sum they are initially given.

In one frequently used alternative version of the Public Goods Game, subjects are offered the opportunity to 'punish' other players whose contribution to the group fund displeases them. Specifically, after the game described above is completed, subjects are told that they can spend some portion of their game money to take money from another player, even though they, themselves, will not be allowed to keep the money. They might, for example, be permitted to spend $1 to take $3 from another player, under the proviso that the $4 will be kept by the experimenter. Obviously, under such circumstances, both players are worse off financially. Moreover, the game is designed to minimize the collateral benefits to punishers: decisions about punishment are made privately, and subjects are aware that they will not play the game again with the same set of people. Nevertheless, in a manner that runs quite contrary to standard economic analyses, subjects often chose to spend their money to punish other players who have made smaller than average contributions to the group fund (Yamagishi, 1986).

Do These Studies Have 'Bleak Implications' for the Extent of Human Rationality?

The results we have sketched are just a tiny sample of an enormous and growing literature on human reasoning and decision making. But what conclusions about human rationality – and irrationality – should we draw from these findings? On this question opinions are sharply divided. In one often quoted comment, two leading investigators in the Heuristics and Biases tradition claimed that the experimental results have 'bleak implications' for human rationality (Nisbett and Borgida, 1975). Other researchers have concluded that "individuals are generally affected by systematic deviations from rationality" (Bazerman and Neale, 1986). And still others have suggested that the fault may be in our genes, or at least in our evolutionary history – "that people lack the correct programs for many important judgmental tasks…[because we]… have not had the opportunity to evolve an intellect capable of dealing conceptually with uncertainty" (Slovic et al., 1976). But many other theorists have argued that conclusions like these are unwarranted. In what follows we discuss four of the more prominent responses to the 'bleak implications' view and the evidence adduced in its support. (For an extended review of the debate see Samuels et al., 2002.)

Problems with Specific Experiments

There is a substantial literature arguing that one or another experiment in the Heuristics and Biases tradition is flawed and thus does not really demonstrate that the subjects' performance is irrational. Some of the putative flaws are methodological. One charge that obviously has some merit is that it is often less than clear that subjects understand the problem in the way that experimenters want them to. For example, principles of conversational implicature might lead subjects in the 'feminist bank teller' experiment to assume that choice (6) (Linda is a bank teller) means that she is a bank teller who is *not* active in the feminist movement (Schwarz, 1996). Other putative flaws turn on the way that the experimenters propose to apply the normative principles to particular problems. For instance, Gigerenzer (2000) argues that there are typically several different ways in which the principles of statistics and probability can be applied to a given problem and that these different analyses of the problem lead to different answers – or in some cases to no answer at all. If this is correct, then obviously we cannot conclude that subjects are being irrational simply because they do not give the answer that the experimenters prefer.

Darwinian Responses

A second prominent response to the claim that the human mind is prone to 'systematic deviations from rationality' comes from work by evolutionary psychologists. These theorists argue that if we are adequately to assess the extent of human rationality, we need to pay due attention to issues about the evolved structure and function of our cognitive systems. Further, they maintain that attention to such matters provides grounds for rejecting the 'bleak implications' view.

Informational Format

One especially prominent instance of this sort of critique arises in research regarding our ability to reason about probabilities. Evolutionary psychologists proceed by noting that the effective use of probabilistic information would have been highly advantageous to our hominid forebears, and therefore, they maintain, it would be surprising if we had not evolved 'an intellect capable of dealing conceptually with uncertainty'. But, they continue, we should expect this capacity to manifest itself most clearly when probabilistic information is presented in a format that would have been common in ancestral environments. In these environments, probabilistic information would almost invariably have been presented in the form of information about frequencies as opposed to single events. Thus, evolutionary psychologists argue, we should expect people to do much better on probabilistic reasoning tasks if the problems are presented in a frequency format. And it appears that they do. In the 'feminist bank teller' problem, for example, if the description of Linda is followed by a question like this one:

There are 100 people who fit the description above. How many of them are

…

(6) bank tellers?

…

(8) bank tellers and active in the feminist movement?

…

the number of subjects who commit the conjunction fallacy drops from over 90% to only about 10%.

Further evidence comes from Cosmides and Tooby's (1996) systematic exploration of the 'Harvard Medical School problem' discussed in Section Base-Rate Neglect. They showed that subjects find it much easier to use base-rate information rationally on 'frequentist' versions of the problem, like the one that follows.

1 out of every 1000 Americans has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease.

Imagine that we have assembled a random sample of 1000 Americans. They were selected by lottery. Those who conducted the lottery had no information about the health status of any of these people.

Given the information above:

on average,

How many people who test positive for the disease will *actually* have the disease? _____ out of _____.

On this version of the problem, 76% of subjects – as opposed to the mere 18% in Casscells et al.'s original study – gave the correct Bayesian answer.

Preference Structure

Evolutionary psychologists have also proposed a related response to the claim that subjects who make substantial offers in the Ultimatum Game contribute to the group fund in the Public Goods Game or pay to punish low contributors in the Public Good Game are being irrational. They concede that it may be irrational to make these decisions if one's only goal in these games is to maximize one's financial gain. But, they note, money is not the only thing that people value. Many people also value fairness and thus they assign positive utility both to being fair in their own offers and contributions, and to punishing others whom they judge to have behaved unfairly in these games. Further, since people in many cultures make fair offers in the Ultimatum Game and pay to punish below average contributions in Public Goods Games, some evolutionary psychologists have suggested that this preference for fairness may be innate and pan-cultural (Baumard et al., 2013).

The suggestion that there is a pan-cultural preference for fairness has not gone unchallenged, however. For example, Henrich et al. (2004) conducted Ultimatum Game experiments in 15 small-scale societies, and found that while people in some of these groups made offers that were similar to those made by subjects in previously studied groups, people in other groups exhibited a wide range of behaviors (for instance, among the Machiguenga, a group in the Peruvian Amazon, the modal Allocator offers only 15% of the money provided by the experimenter, and low offers are almost always accepted). Even more surprising are the results of Gächter and colleagues (Herrmann et al., 2008) who found that in a number of societies scattered around the globe subjects in Public Goods Games with Costly Punishment would pay to punish other players who had contributed *more* than average to the group fund. Ongoing research is aimed at explaining these decisions, which appear to be blatantly irrational and to pose a prima facie challenge to the hypothesis that humans have an innate preference for fairness.

Ecological Rationality

In light of findings such as those just mentioned, some evolutionary psychologists have maintained that our normative standards of rationality require modification. Most influentially, Gigerenzer and collaborators have proposed that for many purposes a conception of rationality that they call *ecological rationality* may be more appropriate as a criterion for normative assessment than the Standard Picture typically assumed by Heuristics and Biases researchers (Gigerenzer et al., 1999). On this view, an inferential or judgmental strategy is ecologically rational if it is accurate and efficient on the sorts of tasks that were important in the environments in which we evolved. If we interpret Aristotle's thesis as the claim that *man is an ecologically rational animal*, evolutionary psychologists maintain that there is a growing body of evidence indicating that he was right (Gigerenzer and Gaissmaier, 2011).

Dual Process Accounts of Human Reasoning

Another response to the 'bleak implications' view that has become quite influential in recent years – even among early advocates of the Heuristics and Biases program – comes from dual-process theories of reasoning (Evans, 2008; Kahneman, 2011; Sloman, 1996; Stanovich, 1999). Dual-process theorists accept that people's reasoning and decision making often deviate systematically from appropriate norms of rationality; but they also maintain that there are many contexts in which people perform quite well. More significantly, they advocate a kind of explanation of this pattern that appears to chart a 'middle way' between those who maintain that normatively problematic heuristics are the *only* reasoning resources people possess, and

those evolutionary psychologists who suggest that *all* reasoning is subserved by normatively unproblematic cognitive mechanisms.

Though dual-process theories differ from one another in many details (Evans and Frankish, 2009), they all propose that reasoning and decision making are subserved by two quite different sorts of system. One kind of system is typically characterized as relatively fast, holistic, automatic, largely unconscious, and requires relatively little cognitive capacity. The other is relatively slow, rule based, more readily controlled, and requires significantly more cognitive capacity. Further, systems of the former kind are widely assumed to have emerged relatively early in human evolution and, as evolutionary psychologists suggest, to have been shaped by natural selection to do a good job on the problems that would have been important to our hominid forebears (Stanovich, 1999, 2004). The other kind of system, by contrast, is assumed to have evolved more recently, is more heavily influenced by culture and formal education, and is often more adept at dealing with many of the problems posed by a modern, technologically advanced and highly bureaucratized society (Stanovich, 1999, 2004).

Since the new system requires more cognitive capacity, is more influenced by culture and education, and does not get used automatically, Stanovich hypothesized that there might be significant individual differences in people's ability and inclination to use it. More specifically, he reasoned, people with higher cognitive capacity, as measured by instruments like the Scholastic Test (SAT), and with a cognitive style that emphasizes 'epistemic self-regulation' should do better on tasks that the old system was not designed to handle. Stanovich agrees with evolutionary psychologists that many of the tasks studied in the Heuristics and Biases tradition fall into this category. In extensive studies of these tasks he has shown that, while the average performance on these tasks is indeed quite poor, there are *some* subjects who give the answer that the Standard Picture suggests is normatively correct on *many* of these problems, and these subjects typically have significantly higher SAT scores and score higher on tests designed to detect cognitive styles that include epistemic self-regulation.

If Stanovich and other dual-process theorists are on the right track, then the unbridled optimism sometimes suggested by evolutionary psychologists is unwarranted, since most untutored people do indeed lack the capacity to deal with a wide range of problems that are important in a technological society. But the glum pessimism often associated with the Heuristics and Biases tradition is not warranted either. Since the fast, automatic, and evolutionarily older system requires little cognitive capacity, everyone has the capacity to deal rationally with many reasoning and decision-making problems that were important in the environment in which we evolved. Moreover, since the new, slow, rule-based system can be significantly affected by education, there is reason to hope that better educational strategies will improve people's performance on those problems that the old system was not designed to deal with.

See also: Bayesian Theory, History, Applications, and Contemporary Directions; Decision Making, Psychology of; Decision Making: Nonrational Theories; Decision and Choice: Bounded Rationality; Decision and Choice: Economic Psychology; Decision and Choice: Heuristics; Moral Reasoning in Psychology; Practical Reasoning: Philosophical Aspects; Probability Theory: Formal; Probability Theory: Interpretations.

Bibliography

Baron, J., 2008. Thinking and Deciding. Cambridge University Press, Cambridge.
Baumard, N., André, J., Sperber, D., 2013. A mutualistic approach to morality: the evolution of fairness by partner choice. Behavioral and Brain Sciences 36, 59–122.
Bazerman, M., Neale, M., 1986. Heuristics in negotiation. In: Arkes, H., Hammond, K. (Eds.), Judgment and Decision Making: An Interdisciplinary Reader. Cambridge University Press, Cambridge.
Casscells, W., Schoenberger, A., Grayboys, T., 1978. Interpretation by physicians of clinical laboratory results. New England Journal of Medicine 299, 999–1000.
Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.
Cosmides, L., Tooby, J., 1996. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. Cognition 58, 1–73.
Dawes, R., 1988. Rational Choice in an Uncertain World. Harcourt Brace Jovanovich, Orlando, FL.
Evans, J.St.B.T., 2008. Dual-processing accounts of reasoning, judgment and social cognition. Annual Review of Psychology 59, 255–278.
Evans, J.St.B.T., Frankish, K., 2009. In Two Minds: Dual Processes and Beyond. Oxford University Press, Oxford, UK.
Gigerenzer, G., 2000. Adaptive Thinking: Rationality in the Real World. Oxford University Press, New York.
Gigerenzer, G., Gaissmaier, W., 2011. Heuristic decision making. Annual Review of Psychology 62, 451–482.
Gigerenzer, G., Todd, P., The ABC Research Group, 1999. Simple Heuristics that Make Us Smart. Oxford University Press, New York.
Gilovich, T., Griffin, D., Kahneman, D., 2002. Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge University Press, Cambridge, UK.
Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., 2004. Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies. Oxford University Press, Oxford, UK.
Herrmann, B., Thöni, C., Gächter, S., 2008. Antisocial punishment across societies. Science 319, 1362–1367.
Kahneman, D., 2011. Thinking, Fast and Slow. Farrar, Straus and Giroux, New York, NY.
Kahneman, D., Tversky, A., 1973. On the psychology of prediction. Psychological Review 80, 237–251.
Nisbett, R., Borgida, E., 1975. Attribution and the social psychology of prediction. Journal of Personality and Social Psychology 32, 932–943.
Samuels, R., Stich, S., Tremoulet, P., 1999. Rethinking rationality: from bleak implications to Darwinian modules. In: LePore, E., Pylyshyn, Z. (Eds.), What Is Cognitive Science. Blackwell, Oxford.
Samuels, R., Stich, S., Bishop, M., 2002. Ending the rationality wars: how to make disputes about human rationality disappear. In: Elio, R. (Ed.), Common Sense, Reasoning and Rationality. Oxford University Press, Oxford, UK.
Schwarz, N., 1996. Cognition and Communication: Judgmental Biases, Research Methods and the Logic of Conversation. Erlbaum, Hillsdale, NJ.
Sloman, S.A., 1996. The empirical case for two systems of reasoning. Psychological Bulletin 119, 3–22.
Slovic, P., Fischhoff, B., Lichtenstein, S., 1976. Cognitive processes and societal risk taking. In: Carol, J.S., Payne, J.W. (Eds.), Cognition and Social Behavior. Erlbaum, Hillsdale, NJ.
Stanovich, K.E., 1999. Who is Rational? Studies of Individual Differences in Reasoning. Erlbaum, Mahwah, NJ.
Stanovich, K.E., 2004. The Robot's Rebellion: Finding Meaning the Age of Darwin. University of Chicago Press, Chicago, IL.
Stein, E., 1996. Without Good Reason. Clarendon Press, Oxford.
Thaler, R., 1992. The Winner's Curse: Paradoxes and Anomalies of Economic Life. Princeton University Press, Princeton, NJ.
Tversky, A., Kahneman, D., 1981. The framing of decisions and the psychology of choice. Science 211, 453–458.
Tversky, A., Kahneman, D., 1982. In: Kahneman, Slovic, Tversky (Eds.), Judgments of and by Representativeness.
Yamagishi, T., 1986. The provision of a sanctioning system as a public good. Journal of Personality and Social Psychology 51 (1), 110–116.

Irrigation Societies

John S Lansing, School of Anthropology, University of Arizona, Tucson, AZ, USA; Santa Fe Institute, NM, USA; and Stockholm Resilience Center, Stockholm, Sweden
Florence Durney, School of Anthropology, University of Arizona, Tucson, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the social sciences, the concept of 'irrigation societies' derives from the view that the construction and operation of large-scale irrigation societies leads to a concentration of power in the state, as argued by Marx and Wittfogel. Thus the control of water is linked to the appearance of many archaic states. Small-scale irrigation systems, on the other hand, often promote egalitarian systems of governance. Contemporary research on irrigation societies consists of case studies; comparative cross-cultural investigations of the relationship between irrigation and power; analysis of the 'design principles' of successful irrigation systems; and studies of the ecological impact of irrigation.

Introduction

The notion of 'irrigation societies' derives from one of the most durable ideas in social science: that the construction and operation of large-scale irrigation systems leads to a concentration of power in the state. An early formulation of this proposition is found in Marx's writings on 'Asiatic societies': "This prime necessity of an economical and common use of water – necessitated in the Orient – the centralizing power of Government" (Marx, 1853, p. 7). In the 1950s, Karl A. Wittfogel (1957) elaborated Marx's thesis, correlating irrigation scale and political hierarchy. Beginning in the 1960s, archaeologists incorporated these ideas into theories of state formation in the Old and New Worlds. The control of water was linked to the appearance of civilizations in Mesopotamia, Peru, the Valley of Mexico, the Indus Valley, China; and to chiefdoms in the prehistoric American Southwest, Polynesia, and elsewhere (Wright, 1986). In the 1970s, sociocultural anthropologists began to take issue with the Wittfogel thesis, noting that in some cases small-scale irrigation systems actually appear to promote egalitarian social systems, not despotic states. Contemporary research on irrigation societies consists of case studies; comparative cross-cultural investigations of the relationship between irrigation and power; analysis of the 'design principles' (Ostrom, 1992) of successful irrigation systems; and studies of the ecological impact of irrigation.

Overview of Irrigation Systems

It is estimated that humanity now uses 26% of total terrestrial evapotranspiration and 54% of runoff that is geographically and temporally accessible (Postel, 2000). Most of this water is used for irrigation. Groundwater use in irrigation is expanding quickly, and almost 40% of the irrigated area is now reliant upon groundwater as either a primary source, or in conjunction with surface water (Siebert et al., 2010). The global total of irrigated land is now estimated to be between 299 and 310 million hectares, with just over 100 million coming from groundwater sources (Siebert et al., 2010). In the past 50 years, there has been an enormous expansion of new irrigation systems. The FAO estimates that the amount of irrigated land roughly doubled from 1961 to 2009 (2011). According to the World Bank, an average of 885 large dams (>15 m high) were constructed per year between the 1950s and the 1980s. Recently, the rate of construction of large dams has begun to decline, ranging between 160 and 310 per year, and this number is expected to decline further because of rising economic, social, and environmental costs (Postel et al., 1996, 2000, World Commission on Dams, 2000, p. 10). But while the rate of increase in large dams is decreasing, the amount of irrigated land is still expected to increase globally by a factor of 1.9 by 2050 (Lobell et al., 2008).

Meanwhile, small-scale, hand-built, gravity-flow canal systems managed by small groups of farmers still account for about 85% of the world's irrigated cropland, according to World Bank estimates. Small-scale systems operated by local irrigators are still used to grow most of the crops in Asia. For example, more than 15 000 systems of less than 50 ha are managed by farmer's associations in Korea (Oh, 1978). In southern India (Andhra Pradesh, Karnataka, Kerala, and Tamil Nadu), more than 125 000 community tanks, many of which are centuries old, provide water for 30% of the irrigated croplands (Sakthivadivel and Shanmugham, 1989; Shah, 2008). More than six million small tanks and reservoirs are managed at the village level in China (Yu and Buckwell, 1991). Similarly, in Morocco, 830 000 ha are irrigated by locally managed systems utilizing perennial flows and seasonal floods, which range in size from 50 to 3000 ha (Abdellaoui, 1989). About a million hectares are cultivated in Nigeria using traditional irrigation or flood-recession systems, while an estimated quarter million hectares of pre-Hispanic Peruvian terraces are still irrigated in the Andean highlands (Mabry, 1996).

Small-scale irrigation systems are also found in the developed world. Fifty thousand local irrigation systems exist in Japan, some in continuous operation since the early seventeenth century, and together constitute about a quarter of Japan's total irrigated area (Takeuchi, 1980). Centuries-old irrigation systems also thrive on the Mediterranean coast of Spain (Glick, 1970), in the Swiss Alps (Netting, 1981), and in the Southwestern USA (Mabry, 1996).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12096-3

Environmental Constraints on Irrigation

The physical environment constrains the kinds of irrigation systems that can be constructed. Wittfogel's thesis was based on the labor demands imposed by large-scale, premodern canal irrigation systems, fed by large rivers. Such systems, typical of arid and semiarid regions such as those adjoining the Tigris, Euphrates, and Indus rivers, required the coordinated labor of thousands to be constructed and maintained. Tropical, semitropical, island, and desert environments impose quite different constraints, leading to different technical and managerial solutions. Scarborough has proposed an 'accretional' model for the modification of tropical and semitropical environments, where there is great species diversity but little abundance of one specific species in any single niche or patch (1998). Archaic civilizations evolving in these settings adapted to this dispersed resource base in a manner markedly different from that of early states located in semiarid settings with access to large rivers. According to Scarborough, the political economies of tropical archaic states took advantage of the interconnectivity of dispersed resources, discovering and refining the ecological pathways that made useable resources accessible to humans. For example, the civilizations of the archaic Lowland Maya, and in many regions of South and Southeast Asia developed irrigation systems based on tanks designed to capture wet-season flows and release them in the dry season. Such tanks are still in use in southern India and Sri Lanka.

Oasis civilizations of Central Asia, China, and Arabia developed tunnels called *qanats* to capture the snowmelt from the Himalayas and other mountain chains, transporting the flow for great distances under the deserts. In parts of North Africa, including the Nile Valley and the American Southwest, flood-recession systems were invented to capture and control scarce seasonal rainfall. Mandelbrot and Wallis showed that the annual flood of the Nile River exhibits chaotic behavior (1968). Using this historic data from the Nile as well as modern data from the Senegal river basin, Park has argued that farmers adapted to the risks of too much or too little water by creating institutions to reallocate land among cultivators annually, fostering control by the elite and class stratification long before population pressure reached significant levels (1992).

Management of Irrigation

The relationship between the technical and social aspects of irrigation is an ongoing topic of research. There are two major schools of thought. One tradition, emanating from Wittfogel and pursued primarily by archaeologists, is concerned with the power-centralizing effects of large-scale irrigation systems. A second tradition approaches small-scale irrigation systems from the perspective of common property management (or 'common-pool resources'). A key issue for both research traditions is the relationship between the scale of irrigation systems and the structure and complexity of management institutions. Hunt distinguishes between 'irrigation communities' and 'water–user associations': the former are autonomous, self-contained organizations that are supported by consensual, mutually beneficial institutional arrangements while the latter are units within larger bureaucratic structures that require communication and coordination across organizational levels (1989).

It appears that many small-scale systems of irrigation management are self-governing. Hunt compared 15 irrigation systems in industrialized and developing countries, ranging from 700 to 30 000 ha in extent, and found no strong relationship between system size and the structure of authority. But he predicted that any system larger than about 100 ha has a 'very high probability' of being managed by a 'unified administrative authority structure' (Hunt, 1988, p. 347). Netting argued that water scarcity can be a stimulus for the centralization of decision making, when it threatens widespread conflict that will seriously reduce the efficiency of the system (1981). Kappel identified 5000 as the upper limit of the number of people who can be supported by local self-governing systems, based on a study of 17 historic and contemporary cases. However, Uphoff found that irrigation systems with command areas of up to 40 ha tend to be managed by the entire group of irrigators; those between 40 and 400 ha are managed by a central official, either elected by the irrigators or appointed by the state, and systems between 400 and 4000 ha often have three layers of management (Mabry, 1996).

Turning to the dynamics of management in small-scale irrigation systems, Ostrom notes that 'a self-governing irrigation system is a prime example of a public enterprise in which a segment of society governs itself for itself' (1992, p. vii). A frequent starting-point for the analysis of irrigation communities is Hardin's 'tragedy of the commons' thesis, which holds that rational self-interest will motivate individuals to take as much of a common resource as possible (1998). This is linked to the 'tail-ender' problem in irrigation systems: irrigators closer to the source can control the flow to their downstream neighbors, so the farmer at the tail end is always vulnerable. How then are sustainable cooperative systems of water sharing possible?

One answer is suggested by game theory, specifically the iterated Prisoner's dilemma, which shows how it can be in one's interest to cooperate with other users if there is an expectation that such cooperation will result in continuing access to a valuable, shared resource. Irrigation systems not only provide benefits, they typically entail high labor costs, and this can also be a motive for cooperation. Ostrom found that in negotiations over the rules of water use between the members of a local water–user association, the bargaining power of tail-enders is greater if their labor is needed to maintain the system (1992). Bardhan also found that caste homogeneity increased the likelihood of cooperative behavior in irrigation systems in Tamil Nadu (2000).

Ecological factors may also influence patterns of cooperation. Rice farmers on the Indonesian island of Bali must cope with a suite of rice pests, including rats, insects, and bacterial and viral diseases. The traditional method for controlling these pests is to plant large blocks of rice terraces at the same time. After harvest the fields are flooded, depriving the pests of their habitat, and in some cases killing them outright. Thus upstream farmers, who control the flow of water, may benefit by cooperating with their downstream neighbors. In this way, the scheduling of irrigation can be used to manage pests (Lansing, 1991, 2006).

Irrigation and Agroecology

Salinization resulting from overirrigation and poor drainage has significantly reduced yields on more than 60 million hectares, about a quarter of the total worldwide, and 25 million hectares have been abandoned because of salt accumulation (Mabry, 1996, p. 229). The amount of salinized irrigated land varies greatly from area to area, ranging between 10 and 50% across nations (Goudie, 2003). Annually, for every additional hectare brought into production by new irrigation projects, another goes out of production because of salinization (Umali, 1993). In ancient Mesopotamia, irrigators may have developed successful strategies to slow the rate of salinization (Altaweel and Watanabe, 2012).

Flowing water can pick up and transport mineral nutrients needed by plants, such as nitrogen, phosphate, and potassium. Traditional Asian rice paddies use irrigation systems to bring these nutrients, as well as water, to the plants. For example, volcanic ash from Indonesian volcanoes contains about 1% potassium, and relatively high levels of phosphate. Until the advent of the 'Green Revolution' in rice in the 1970s (the spread of high-yielding crops designed to rely on chemical fertilizers), most of the nutrients needed in Asian rice paddies were supplied by this process of mineral leaching. In Western Europe and the USA, the spread of 'Green Revolution' farming techniques has created an overabundance of nutrients: too much of a good thing. Worldwide, between 1960 and 1980 the application of nitrogen fertilizers increased more than fivefold, and in the decade that followed, more synthetic fertilizer was spread on land than had been applied in the entire previous history of agriculture. Rivers like the Mississippi and the Colorado acquire heavy loads of nutrients like phosphate, and generate enormous pollution problems downstream. For example, in 2009, 21 756 square kilometers of the deep waters of the Gulf of Mexico became a 'dead zone' (Nixon, 1998). European countries like the Netherlands restrict fertilizer use in order to protect freshwater ecosystems, but meanwhile pollution problems have scaled upward from lakes and streams to the Baltic and North Sea. Worldwide, virtually all coastal zones are experiencing rapidly increasing pollution from agricultural runoff and domestic waste (Diaz and Rosenberg, 2008). Archaeologists generally agree that environmental problems associated with the expansion of irrigation systems led to the decline of many ancient states. Today, the spread of irrigated agriculture on which the continuing expansion of the human population depends poses a 'tragedy of the commons' problem on a global scale (Lansing and Kremer, 2011).

See also: Civilizations; Land Tenure; Resilience.

Bibliography

Abdellaoui, R.M., 1989. Small-scale irrigation systems in Morocco. In: Public Intervention in Farmer-managed Irrigation Systems. International Irrigation Management Institute, Digana, Sri Lanka, pp. 165–173.

Altaweel, M., Watanabe, C., 2012. Assessing the resilience of irrigation agriculture: applying a social–ecological model for understanding the mitigation of salinization. Journal of Archaeological Science 39 (4), 1160–1171.

Bardhan, P., 2000. Irrigation and cooperation: an empirical analysis of 48 irrigation communities in South India. Economic Development and Cultural Change 48 (4), 847–865.

Diaz, R.J., Rosenberg, R., 2008. Spreading dead zones and consequences for marine ecosystems. Science 321 (5891), 926–929.

FAO, 2011. The State of the World's Land and Water Resources for Food and Agriculture: Managing Systems at Risk. Earthscan, Milton Park, Abingdon.

Glick, T., 1970. Irrigation and Society in Medieval Valencia. Harvard University Press, Cambridge, MA.

Goudie, A.S., 2003. Enhanced salinisation. In: Alsharhan, A.S., Wood, Warren W. (Eds.), Water Resources Perspectives Evaluation, Management and Policy. Elsevier, Amsterdam, pp. 287–293.

Hardin, G., 1998. Extensions of the tragedy of the commons. Science 280, 682–683.

Hunt, R.C., 1988. Size and structure of authority in canal irrigation. Journal of Anthropological Research 44 (4), 335–355.

Hunt, R.C., 1989. Appropriate social organization? Water user's associations in bureaucratic canal irrigation systems. Human Organization 48, 79–90.

Lansing, J.S., 1991. Priests and Programmers: Technologies of Power in the Engineered Landscape of Bali. Princeton University Press, Princeton, NJ.

Lansing, John Stephen, 2006. Perfect Order: Recognizing Complexity in Bali. Princeton University Press, Princeton, N.J.

Lansing, J. Stephen, Kremer, James N., 2011. Rice, fish and the planet. Proceedings of the National Academy of Science United States of America 108 (50), 19841–19842.

Lobell, D.B., Burke, M.B., Falcon, W.P., Naylor, R.L., Tebaldi, C., Mastrandrea, M.D., 2008. Prioritizing climate change adaptation needs for food security in 2030. Science 319 (5863), 607–610.

Mabry, J.B. (Ed.), 1996. Canals and Communities: Small-scale Irrigation Systems. University of Arizona Press, Tucson, AZ.

Mandelbrot, B.B., Wallis, J.R., 1968. Noah, Joseph and operational hydrology. Water Resources Research 4 (5), 909–918.

Marx, K., 1968 [1853]. The British rule in India. In: Avineri, S. (Ed.), Karl Marx on Colonialism and Modernization. Doubleday, Garden City, New York.

Netting, R.M., 1981. Balancing on an Alp. Cambridge University Press, Cambridge, UK.

Nixon, S., 1998. Enriching the sea to death. Scientific American Quarterly 9 (3), 48–63.

Oh, H.-S., 1978. Customary rules of water management for small irrigation reservoirs in Korea. Journal of Rural Development 1, 96–110.

Ostrom, E., 1992. Crafting Institutions for Self-governing Irrigation Systems. ICS Press, San Francisco, CA.

Park, T.K., 1992. Early trends toward class stratification: chaos, common property and flood recession agriculture. American Anthropologist 94 (1), 90–117.

Postel, S.L., Daily, G.C., Ehrlich, P.R., February 1996. Human appropriation of renewable fresh water. Science 271, 785–787.

Postel, Sandra L., 2000. Entering an era of water scarcity: the challenges ahead. Ecological Applications 10 (4), 941–948.

Sakthivadivel, R., Shanmugham, C.R., 1989. Issues relating to interventions in farmer-managed irrigation: rehabilitation of a tank irrigation system. In: Public Intervention in Farmer-managed Irrigation Systems. International Irrigation Management Institute, Digana, Sri Lanka, pp. 113–130.

Scarborough, V.L., 1998. Ecology and ritual: water management and the Maya. Latin American Antiquity 9 (2), 135–159.

Shah, Esha, 2008. Telling otherwise: a historical anthropology of tank irrigation technology in South India. Technology and Culture 49, 652–674.

Siebert, S., Burke, J., Faures, J.M., Frenken, K., Hoogeveen, J., Doll, P., Portmann, F.T., 2010. Groundwater use for irrigation – a global inventory. Hydrology and Earth System Sciences 14 (10), 1863–1880.

Takeuchi, S., 1980. Japan. In: Takeuchi, H. (Ed.), Farm-level Water Management in Selected Asian Countries. Asian Productivity Organization, Tokyo, pp. 76–90.

Umali, D.L., 1993. Irrigation-induced Salinity. Technical Paper No. 215. World Bank, Washington, DC.

Wittfogel, K.A., 1957. Oriental Despotism: A Comparative Study of Total Power. Yale University Press, New Haven, CT.

World Commission on Dams, 2000. Dams and Development: A New Framework for Decision-making: the Report of the World Commission on Dams. Earthscan, London.

Wright, H.T., 1986. The evolution of civilizations. In: Meltzer, D., Fowler, D., Sabloff, J.A. (Eds.), American Archaeology, Past and Future. Smithsonian Institution Press, Washington, DC, pp. 323–368.

Yu, C.L., Buckwell, A., 1991. Chinese Grain Economy and Policy. CAB International, Wallingford, UK.

Islam: Asia

Mark R Woodward, Arizona State University, Tempe, AZ, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 7916–7920, © 2001, Elsevier Ltd.

Abstract

The majority of the world's Muslim population is located in South, East, and Southeast Asia. Muslim control of overland and ocean trade routes linking the Mediterranean basin China and the Spice Islands of eastern Indonesia brought Islam to East and Southeast Asia. Afghan conquests were largely responsible for the establish of Islamic political authority and the spread of Islam among the populations of South Asia. Relationships between religion, including Islam, and ethnicity in the region are complex and highly variable. In societies and states with Muslim majorities ethnic distinctions among Muslims are socially and politically significant. In those with Muslim minorities, particularly China, India, Burma, and the Philippines, 'Muslim' is often regarded as an ethnic as well as religious category. Everywhere there is a dynamic tension between universal Muslim values and legal norms encoded in Arabic and other religious texts and those of local cultures. This tension is expressed in many cultural domains ranging from ritual performance to concepts of political authority. Contemporary 'Islamist' or 'fundamentalist' movements are attempts to shift the balance between local culture and Islamic universalism in the latter direction.

Islam is often understood as a Middle Eastern and, even more narrowly, an Arab religion. While the origins and core scriptures of Islam are indeed Arabic, the majority of the world's Muslim population is located in South, Southeast, and East Asia. This article concerns the Islamicate civilizations and societies of these regions. It examines theoretical and methodological approaches to the study of Islam in the region, the historical processes through which Islam came to be an Asian religion, relationships between ethnicity and religion in Asian societies, popular religion and vernacular literatures in Islamic Asia, the diverse roles played by Islam and Muslims in public life in contemporary Asian States, and the significance of Islamist or fundamentalist movements.

'Islam' and 'Muslim' Definitional and Theoretical Concerns

In both western and Muslim writing the terms 'Islam' and 'Muslim' are used in various and often contradictory ways. Conflicts concerning their use and definition are apparent in, and have done much to shape, scholarly discourse about religion and culture in Muslim Asia. In a theological sense the term Islam means 'submission to god.' A Muslim is a person or *jinn* (spirit) who practices submission. Debates concerning the meaning of submission are a central feature of Islamic discourse. Muslim jurists define Islam as an all encompassing behavioral code and do not distinguish between religious and other modes of social behavior. Mystics (*Sufis*) define Islam as an emotional or spiritual state and are less concerned with personal and social behavior. Proponents of both positions often describe their opponents as *kafir* (non-Muslim).

The line between social scientific and confessional scholarship in and about the Muslim world is a fine one. Many social scientific studies share the Muslim juristic assumption that Islam is first and foremost a legal tradition and that there is a universal and normative core against which the 'Islamicness' of particular cultures can be evaluated. The result has been that societies in which the influence of Muslim mysticism predominates have often been described as nominally and even trivially Muslim (see, for example, Snouck Hurgronje, 1906; Geertz, 1960). Since the early twentieth century many of the basic reference works on Islam in European languages has been written by Muslim apologists whose definitions of 'Islam' are theological motivated (for a contemporary example see Rahman, 1980). For the most part their works assume a legalistic understanding of Islam. To the extent that the social science literature has been shaped by these works, it employs Muslim theological categories. Muslim scholars have, in turn, begun to use social scientific works by Western scholars with whom they are inclined to agree, as authoritative references in overtly apologetic treatises.

Recent scholarship has challenged this approach and the related view that Arabs are the guardians of Muslim orthodoxy. Increasingly, social scientists and textualists have been concerned with the dynamic tension between textual and lived Islam in diverse cultural and historical settings. The textual tradition is now understood as the product of discursive practice. Recent scholars have examined the ways in which this tradition describes, and argues for, a variety of possible 'Islams.' This perspective attempts to capture both the variety of Islamic texts and cultures and historical and doctrine linkages among them. The term 'Islamicate' is now used in reference to societies and civilizations whose development has been shaped by the interaction of Islam, as a religion, and a variety of local cultures. The use of this term circumvents concerns with orthodoxy inherent in early scholarship.

The chief difficulty with this approach is that by privileging the local character of Muslim discourse, it runs the risk of accepting the local as normative. Progress in understanding the local Islams of Asia, and elsewhere, will require additional attention to the ways in which elements of the pan-Islamic textual tradition are interpreted, understood, and put into practice, and alternatively the ways in which specific histories and cultures constrain and limit the implementation of the social ideals described in universally recognized Muslim texts. There is also a need for comparative studies of the various Asian Islams.

Historical Background

Political, economic, and military factors all contributed to the establishment of Islam as an Asian religion and the foundation

of Asian Islamicate civilizations. In the seventh and eighth centuries Arab Muslim armies carried Islam into central Asia and western India. Muslim military traditions resulted in the conquest of most of South Asia. Here, Muslim expansion was motivated as much by the religious conviction that Muslim rulers have an obligation to expand the Muslim world (*dar al Islam*) as by economic considerations. Commerce spread Islam further. Arab control of the western terminus of overland and maritime trade routes linking the Mediterranean basin and China led to the expansion and consolidation of Islam in Central, East, and Southeast Asia.

South Asia

There has been a prominent Muslim presence in what are now Afghanistan and Sind (western Pakistan) since the early eighth century. It was not, however, until approximately 1000 CE that the conquest of the plains of what is now India began in earnest. A long series of raids and invasions based in Afghanistan led to the establishment of Islamicate empires centered in the Gangetic plain. The expansion of Muslim polities continued until the establishment of British power early in the eighteenth century. The Central Asian origins of South Asian Islam account for Turkish and Persian influences on South Asian Muslim literature and popular religion and for the presence of a significant *Shia* minority not found elsewhere in Asia. British conquest of Burma in the nineteenth century (1825–85) led many South Asian Muslims to emigrate to this Southeast Asian nation. Arab Muslim traders reached the southwest coast of India and Sri Lanka in the eighth century. Here, trade, rather than conquest, was the primary factor facilitating the expansion of Islam. The Islamicate cultures of this region resemble those of Southeast Asia and southern China more closely than those of northern portions of the Indian subcontinent.

Southeast Asia

Islam came to Southeast Asia in the thirteenth century. It was brought to the Malay world by Arab and South Asian Muslim traders who dominated the Indian Ocean trade during that period. Muslim states were established on both sides of the straits of Malacca and later spread southward to Java and as far east as the southern Philippines. Largely owing to the power of coastal Sultanates it rapidly became the dominant religion in the Malay-speaking world. Muslim communities in Vietnam, Cambodia, southern Thailand, and Burma originated in similar ways, but never displaced the political dominance of Indianized and, in the case of Vietnam, Sinicized states. Islam came to Burma from two additional sources in the precolonial period, from Bengal in the west and from Southern China. During the period of British rule large numbers of south Asian Muslims migrated to Rangoon, Mandalay, and other Burmese urban centers. There are also scattered Chinese Muslim populations in northern regions of Thailand, Laos, and Vietnam.

China

Arab Muslim armies reached China's western border in AD 714. Muslim control of the silk road stretching across central Asia to western China was the avenue through which Islam came to be the religion of the various Turkic-speaking communities of central Asia, and substantial portions of northwestern China. The Mongol conquest of China and the Islamic lands of central Asia promoted closer contacts between China and the Middle East. The Mongol conquest of Persia and the sack of Baghdad in 1258 CE was a devastating blow to Islamicate cultures. However, the subsequent conversion of the Mongol elite to Islam enhanced the position of Chinese Muslim communities and resulted in the Islamicization of most of central Asia.

Legends preserved by Muslims in Southeastern China speak of Muslim emissaries and traders arriving during the lifetime of the Prophet Muhammad. While the veracity of these accounts is difficult to determine, Chinese records and archaeological evidence point to a significant Muslim presence in Southern China from the seventh century onwards. Muslims played a major role in Chinese international and internal commerce and politics from the eighth century until the establishment of the Ching dynasty in 1644. Muslims were a sizable and 'convenient' minority. Islam enabled them to establish and maintain relationships with the outside world that was both feared and belittled by the Chinese elite. At the same time Islam precluded its adherents from competition for political power within China. Javanese and other Southeast Asia chronicles mention China, as well as Arabia and India, as being among the sources of Islam. This speaks of the importance of the premodern Asiatic commercial system in the expansion of Islam in Asia.

Islam and Ethnicity in Asian Societies

Sociologically, religion and ethnicity are closely intertwined in the Asian Muslim world, though in complex and varied ways. In some instances Islam transcends ethnic distinctions, in others Muslim communities are divided along ethnic lines, despite the normative doctrine and political ideology of 'Muslim brotherhood.'

In nation states with Muslim minorities 'Muslim' has become an ethnic, and often political category, even where Muslims themselves would not recognize it as such. In China, the Philippines, and the Buddhist societies of mainland Southeast Asia, majority populations and governments define Islam as an ethnic category, despite linguistic and religious differences within the Muslim community. Burma provides a clear example. As the result of the inclusion of Burma in the British Indian empire, the Muslim community is extremely diverse. In addition to communities with historical ties to many regions of India, there are Chinese and ethnic Burmese Muslim communities. Typically each of these communities has its own mosque and religious officials. Burmese Buddhists and the national government ignore these distinctions and categorize all of these diverse communities as simply 'Muslim.' In nation-states with Muslim majorities Islam is generally understood as a unifying force that transcends ethnic differences.

In colonial and postcolonial societies Islam has often been used as a unifying ideology. The use of Islam as an organizing principle contributed to the rise of nationalist movements in British India and Malaysia and the Dutch Indies (Indonesia). The effectiveness ofIslam as a political principle has proved to be uneven. It was largely responsible for the division of British India into the states of India and Pakistan in 1949. It provided the basis for Malay nationalism in Malaysia, but also contributed to

Singapore's succession and for ongoing tension between Malay Muslim, South Asian Hindu, and Chinese communities.

South Asia

Religion and ethnicity have long been salient features of social and political discourse in South Asia. Pakistan was founded on the basis of a common religious identity, obscuring ethnic and cultural relationships between Hindus and Muslims in both the Northwest and Eastern regions as well as ethnic and linguist differences among Muslims. Pakistan subsequently split into Pakistan and Bangladesh, a division that reflects, geographic, ethnic, and linguistic differences. Bangladesh is an ethnic as well as a religious state. Pakistan is a multiethnic state that uses Islamic ideology to transcend ethnic divisions. The rise of militant Hindu nationalism as a potent political force in India has led to the identification of Islam and Muslims as an ethnically distinct and 'foreign' element of Indian society. Contemporary Indian Muslims are inclined to disregard sectarian and ethnic differences in the face of a mounting tide of Hindu religious nationalism.

Southeast Asia

Southeast Asia is characterized by extreme ethnic and religious diversity. Islam is the dominant religion in Indonesia, Brunei, and Malaysia. There are important Muslim minorities in the Buddhist countries (Burma, Thailand, Laos, and Cambodia), Sinicized Vietnam, and the Christian Philippines.

The role of Islam in the political systems of the majority Muslim nations of the region, Indonesia, Malaysia and Brunei differs fundamentally. Malaysia is officially an Islamic state, although little more than 50 percent of its citizens are Muslims. Political parties compete to appear 'more Islamic than each other.' Despite the fact that it is the world's most populous Islamic nation, Indonesia defines itself as being neither a religious or a secular state, but rather as one with the state ideology *Panca Sila* (the Five Principles) which comprise: belief in the one true God, a just and civilized humanity, national unity, democracy based on consensus of the people, and wise leadership and social justice for all of the Indonesian people. This ideology seeks to transcend religious differences in a Muslim state in which there are small, but important, Christian and Hindu minorities. Islamic political parties, several of which advocated the creation of an Islamic state, play significant roles in Indonesian politics. These groups understand Islam as an ideology that transcends ethnic distinctions. Other, more liberal, Muslim organizations have attempted to establish an Islamic theology of tolerance promoting ethnic and religious pluralism. Brunei is by far the smallest of the Muslim states of Southeast Asia. It is officially defined as a Malay Islamic Monarchy. Power is located with the Sultan and members of his immediate family. The current Sultan is known as a pragmatist and has sought to limit the influence of Muslim scholars and fundamentalists.

Other Southeast Asian nations have significant Muslim minorities. There is, however, considerable variation in the social and economic status of Muslim minority populations. There is a basic distinction between immigrant and native Muslim communities, that is, between those who define themselves in terms of ethnic categories indigenous to the region and those of Arab, Persian, and South Asian origin. The largest communities of native Muslims are found in the Southern Philippines and Southern Thailand. While Muslims make up less than 10 percent of the Philippine population, they are concentrated in the Sulu archipelago and the southern Island of Mindanao. Phillippino Muslims are culturally and linguistically related to those of eastern Indonesia and the Indonesian/Malaysian island of Kalimantan (Borneo). They fiercely resisted both Spanish and US colonialism and have never been fully integrated into the modern Philippine state. The conquest of these territories was not completed until 1914. The modern state has continued the anti-Islamic policies of Spanish and US colonialists, and encouraged Christian immigration to Muslim majority areas. The result has been continued armed struggle between the army and the Moro National Liberation front. Muslims make up approximately 10 percent of the population of Thailand and are of diverse ethnic origin. Most of the Muslim population of the Southern provinces are Malay, although there are also significant Muslims of the Thai origin. Much of this area was not part of the Thai state until the mid-nineteenth century. There are still periodic outbreaks of resistance to Thai rule. In the vicinity of Bangkok there are diverse Muslim communities of Malay, Thai, Persian, and Cham (Cambodian) origin. In the northern region there is a substantial community of Chinese Muslims originating in Yunan. In general the Thai Muslim community is not well organized as it faces continuous pressure to assimilate Thai cultural norms and Theravada Buddhism. Prior to the Khmer Rouge persecutions there were approximately 500 000 Cambodian Muslims Most are the descendants of Muslim refugees from the kingdom of Champa which was destroyed by the Vietnamese in the fifteenth century. They retain distinct ethnic and linguistic identity. Very few survive.

China

Chinese Muslims (*Hui*) are recognized by the People's Republic as an official nationality, despite considerable cultural and religious differences within the community. Some southern Chinese communities that trace their descent to Arab Muslim traders have sought to be officially recognized as *Hui* to obtain economic benefits granted to minority nationalities. Descent and refusal to eat pork are the primary factors distinguishing them from the larger *Han* Chinese community. In northwest China Turkic Muslim communities employ a combination of ethnic and religious criteria to distinguish themselves from the *Han* majority. The collapse of the Soviet Union led to an Islamic revival among the Turkic people of central Asia. This has led to an increased concern with ethnic and Islamic identity among the Muslim peoples of northwest China. The Chinese Nationalist government of Taiwan views relationships between religion and ethnicity differently than the People's Republic. It does not officially recognize the distinctiveness of Muslim Chinese. Nor does it limit or regulate expressions of Muslim religiosity.

Popular Religion and Vernacular Literatures

Arabic religious texts are known throughout Islamicate Asia. The *Qur'an Hadith* (traditions attributed to the Prophet

Muhammad), legal, mystical, and theological works are studied in theological academies throughout the region. There are also significant vernacular literatures in Urdu, Bengali, Tamil, Javanese, and Malay. Vernacular literatures are often based on a combination of pan-Islamic and pre-Islamic components. Many recount tales of the spread of Islam and the exploits of local Muslim saints, kings, and warriors. Throughout the twentieth century there have been concerted efforts to translate Arabic theological texts into Asian languages. This has contributed to an increased awareness and appreciation of Islamic theology and law in almost every Asian Muslim society.

The ritual complex mandated by Islamic law – the five daily prayers, alms, fasting during the month of *Ramadan,* and the pilgrimage to Mecca – are known and practiced throughout the region. Other kinds of popular religion reflect the significant role played by Muslim mystics *(Sufis)* in the conversion process. These include pilgrimage to the tombs of saints, ritual recitation of the *Qur'an,* healing rites, and the celebration of the *maulid* or birthday of the Prophet Muhammad. Sufi orders continue to play significant parts in social, religious, and political life in most Asian Islamicate societies. For many Asian Muslims Sufi devotionalism is more important than that required by Muslim law.

Islamist Movements

Islamist movements advocate the 'purification' of Muslim religious life. In general they seek to restrict devotional acts to those prescribed by Islamic law. They strongly oppose *Sufism* and rituals rooted in or associated with pre-Islamic traditions. They also seek to establish the Islamic law as the basis for social and political life. They are significant voices in intellectual, social, and political discourse throughout Islamicate Asia. Most have intellectual roots in the theology of the nineteenth century Egyptian modernist Muhammad Abduh. They emphasize a combination of modern education and religious zealotry. Many believe that the realization of the goals can be facilitated by the establishment of Islamic states. Support for these movements is found primarily among urban, middle-class movements. They are strongly opposed by traditional religious elites and in most instances ignored by rural populations.

See also: International Migration by Ethnic South and Southeast Asians; Islam: Middle East; Islam: Sub-Saharan Africa; Islamic Fundamentalism; Islamic Social Work: Hope and the Commitment to Change; Near Middle East/North Africa Studies: Religion; Religion and Economic Life; Religion: Evolution and Development; Religions of India.

Bibliography

Bowen, J.R., 1993. Muslims through Discourse: Religion and Ritual in Gayo Society. Princeton University Press, Princeton, NJ.

Dhofier, Z., 1999. The Pesantren Tradition. Program for Southeast Asian Studies Monograph Series. Arizona State University Press, Tempe, AZ.

Esposito, J., 1987. Islam in Asia. Religion, Politics and Society. Oxford University Press, New York.

Geertz, C., 1960. The Religion of Java. Free Press, Glencoe, IL.

Gladney, D.C., 1991. Muslim Chinese. Ethnic Nationalism in the People's Republic. Harvard University Press, Cambridge, MA.

Israeli, R., Johns, A. (Eds.), 1984. Islam in Asia. Volume II. South and East Asia. Jerusalem University Press, Jerusalem.

Lapidus, I.M., 1988. A History of Islamic Societies. Cambridge University Press, Cambridge, UK.

McNeill, W.S., Waldman, M. (Eds.), 1973. The Islamic World. The University of Chicago Press, Chicago.

Rahman, F., 1980. Major Themes of the Qur'an. Bibliotheca Islamica, Minneapolis, MN.

Hurgronje, C.S., 1906. The Acehnese, Vol. 2 (O'Sullivan AWS trans.). Brill, Leyden, The Netherlands.

Woodward, M.R., 1989. Islam in Java: Normative Piety and Mysticism in the Sultanate of Yogyakarta. University of Arizona Press, Tuscon, AZ.

Islam: Middle East

Charles Lindholm, Boston University, Boston, MA, USA.

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 7920–7927, © 2001, Elsevier Ltd.

Abstract

More than any other religion, Islam understands itself in relationship to its past, as believers seek to emulate the lives of the Prophet and his associates. This means that the language and culture of Arabia is of critical spiritual import for all Muslims. This article will outline the historical experience of Islam in Arabia, with attention paid to the social circumstances surrounding the rise of Muhammad's prophecy, and to the development of hadith scholarship, jurisprudence, and sectarianism over time within the Arabic Muslim community. The problematic relationship of Islam with the state will be briefly considered in historical perspective, as will the rise of charismatic and millenaristic movements. But what will be stressed is the pragmatic and personal moral message of Islam, and its effect on ordinary Muslims in Arabia.

Islam in Historical Context

Islam claims to be the direct word of the almighty and omniscient God (Allah), as directly revealed to His Messenger, Muhammad. Written down in the Quran, Allah's commands provide Muslims (those who submit) with the final and authoritative pathway to salvation, as well as a code of proper conduct covering all aspects of ordinary life. Islam is arguably the fastest growing of all religions, daily continuing its rapid spread throughout every corner of the world. Yet, despite its cosmopolitan character, Islam remains inextricably linked in language and in character to its past in the desolate heart of Saudi Arabia, where it arose in the beginning of the seventh century AD (for standard histories, see Hodgson, 1974; Lapidus, 1988).

This arid place was an extraordinarily unlikely spot for a new world religion to bloom. In fact, Arabia had avoided conquest by the stronger surrounding powers primarily because of its geographical isolation and physical harshness. The marginalization of Arabia meant that none of the major religions of the surrounding world had gained ascendance there, although the Arabs traced their own genealogical charter back to Ishmael, the son of Abraham and Hagar, and therefore had a historical connection with Judaism and Christianity. But in practice they worshipped a plethora of local gods (who remain powerful today under the guise of *jinns*). These divinities offered their devotees protection from the unpredictable environment and acted as guardians during holy months when clan feuds were set aside and markets were organized on sanctified ground. One of these markets, according to many scholars, was to be found in the central Arabian crossroads of Mecca, site of the kaaba, the holy black stone that is now the goal of the *hajj*, the annual Muslim pilgrimage (for a contrary view, see Crone, 1987).

There is great controversy among Western scholars over the accuracy of Islamic records concerning the early years of the religion (c.f. Crone and Cook, 1977; Humphreys, 1991; Wainsbrough, 1977; Watt, 1973, 1988), but most accept that Muhammad, the Prophet ofIslam, was born in Mecca in approximately the year 570 to a member of the Quraysh clan, the kin group that controlled the market and its gods. As a posthumous child he had to rely on the guardianship of his paternal uncle, Abu Talib, and in his later revelations Muhammad was especially concerned with the protection of widows and orphans. With the help of his first wife, the successful and much older tradeswoman Khadija, he eventually became a well-to-do merchant, much respected for his integrity, good judgment, and fairness by his fellows, who called him al-Amin, the trustworthy.

At the age of 40, during a meditative retirement on Mount Hira, Muhammad was gripped by the first of the uncanny revelations that would eventually be collected into the Quran. In his vision, the angel Gabriel spoke to him, commanding him to 'Proclaim! In the name of thy Lord and Cherisher, who created – created man out of a leech-like clot.' Later revelations occurred at intervals throughout his life. One hundred and fourteen of these revelations, presented in rhymed verse of surpassing beauty and novelty, were definitively collected and collated in sections or *suras* of descending order of length (not chronology) shortly after the Prophet's death. These became the Quran, the central holy text of Islam. For the Arabs, whose major art form is poetry, the extraordinary aesthetic power of the Quranic verses was a convincing proof of their divine source. Oral recitation from the Quran remains today as the most important way for the faithful to experience the sacred (Nelson, 1986).

According to Muslim hagiography, Muhammad's revelations originally had little influence in Mecca. His first converts were Khadija and some of his closest relatives, including Abu Talib's son Ali, who was later to marry Khadija's daughter Fatima. But most of the early believers were those who were poor, disenfranchised, and humble. They were drawn to the Quran's condemnation of excessive riches, to its advocacy of generous donations to care for the disadvantaged, and to its repudiation of the arrogance and selfishness of the wealthy. Yet the Quran also had a place for the interests of the upper strata. According to Muhammad's revelation, wealth and power were not bad in themselves – only greed, arrogance and an absence of community responsibility.

Nonetheless, most of the Meccan populace were unwilling to follow the Islamic demand that they pay *zakat* (a tax to be

redistributed to the community of believers), cease to hoard, participate in community charity, and refuse interest on loans. According to legend, many were especially angered at Islam's absolute denial of the might of their clan gods in favor of the complete domination of Allah – a threat to the basic structure of their patriarchal society. Also galling for them was the Quranic demand that the great show submission by prostrating themselves in prayer to the Almighty God who would judge them on the Last Day.

After the death of his defender and uncle, Muhammad and his followers suffered increasing persecution in Mecca and looked for refuge elsewhere. Eventually, in 622 the Aws and Khazraj tribes of the agricultural oasis of Yathrib (later called Medina), anxious for a leader who could mediate their interminable feuds, begged Muhammad to migrate there. Fearing for his life in Mecca and hopeful of increasing his following, Muhammad accepted this invitation, and undertook the *hijra* to Medina with his followers – an event marked as the beginning of the Muslim calendar. It was there that he built the foundations of his holy empire, attracting converts from the surrounding tribes as he sought to conquer the Meccan confederation that had repudiated him. Following his inspiration, he rallied his following to struggle against the odds, defeating a much larger Meccan army at the battle of Badr in 624, successfully raiding Meccan caravans, and threatening Meccan commercial dominance. Finally in 630 the Muslim army occupied Mecca itself, as Muhammad's Qurayshian cousins were obliged to convert to his cause. When Muhammad died in 632, his followers had conquered the whole Arab peninsula, and had begun their great and ultimately successful war against the surrounding Sasanid and Byzantine empires. All of these events remain in the forefront of Muslim imagination to this day.

The Muslim explosion which burst from the hinterlands of Arabia was historically unparalleled; it united the Middle East into one open trade area, paving the way for the rapid creation of new prosperity. At the same time, the radical Muslim message of the equality of all believers struck a cord with the common people of the empires, who, theoretically at least, were liberated from their inferior status by the simple act of conversion. The rise of Islam was thus both an economic and social revolution, offering new wealth and freedom to the dominions it assimilated under the banner of a universal brotherhood guided by the message of the Prophet of Allah. Of course, in practice the situation was more complex and ambiguous, as Arabs – proud of their early support of Muhammad – claimed superiority over the more recently converted Persians and Berbers, who in response demanded equal rights for themselves under the *sharia* or Muslim law. There was also rivalry between the old Qurayashian elite and the upstarts who had gone into exile with Muhammad, and now claimed special precedence for themselves. Nonetheless, Islam expanded rapidly, and soon predominated throughout the Middle East, from Morrocco to Afghanistan (for accounts, see Bulliet, 1994; Gibb, 1962; Shaban, 1971; Wellhausen, 1927).

In the Muslim imagination this was a period when the 'City of God' was actually realized in historical reality, under the authority of the Prophet himself and the four pious Caliphs after him: Abu Bakr (d. 634), Umar (d. 644), Uthman (d. 656), and his patrilateral parallel cousin Ali (d. 661) – who led Islam in its holy war against the unbelievers. To compare the Muslim experience with Christianity, it is as if Jesus had commanded an army that subdued Jerusalem, and Peter had become Emperor of Rome. Muslim thought is saturated with longing for a return of this idealized era when ordinary men and women worked together selflessly to realize the will of Allah under the leadership of just and divinely guided leaders, and were rewarded for their efforts by vast power and wealth.

Recalling their millennial past, the Muslim devout, unlike their Christian cousins, have never inwardly consented to the disjuncture between the religious experience of the community of believers (equal before God, led by the Prophet and his deputies) and the later reality of power-seeking secular rulers given to political intrigue and physical coercion. The remembered unity of politics and religion during the first years of Islam has always remained available as an inspiration to Muslim reformers and as a rebuke to politicians. That unity is evoked today by Muslim reformers, who inevitably decry the corruption of the times and call for a return to the pure faith and practice of the past (for an example, see Mitchell, 1969).

To a great degree, the extraordinary victories achieved by early Islam were a result of the manner in which Muhammad's message transformed and expanded the worldview of the tribesmen who became fighters for the faith. As in a tribe, within the Muslim community – the *umma* – co-responsibility was required, and all believers were strictly prohibited from exploiting, injuring, or enslaving one another; they were obliged as well to share their resources through charitable contributions redistributed to the community at large. They were also expected to unite together in holy war (*jihad*) against the enemies of the faith. Joining the *umma* was a voluntary matter, but once the faith was accepted then a Muslim was forever committed, just as a tribesman was forever linked to his kin.

Also ratified and transformed by Islam were the mercantile values Muhammad had grown up with in Mecca. The Quran does not repudiate commerce, nor did the growth of Islam destroy trade and markets. On the contrary, Muhammad's message favored entrepreneurship, and integrated the ethics of the merchant into Islamic virtue. Profit, if fairly made and appropriately spent in the service of God and to help the deprived, is praised in the Quran, while the absolute right of individuals to own and dispose of personal property is assumed, as is the right to make contracts. In fact, the Quran is permeated with the imagery of the marketplace, which must have appealed directly to the individualistic mercantile values of the people of the time, both urban and rural, who were all so dependent on trade for their livelihoods (Rodinson, 1973; Torry, 1892).

Most importantly, the Quran emphasized the spiritual equivalence of all persons, who are held to be responsible for their own destinies and who will be judged as individuals by Allah on the Last Day. In principle, Islam is the least hierarchical of all religions: there are no priests, no church, no monks, no difference between clergy and laymen. Everyone is alike in the eyes of the transcendent and absolute God to whom all must submit. This aspect of Islam applied even to the Prophet, who was at pains to deny any sacred or superhuman status to himself. When doubters asked Muhammad to verify his revelation by miracles, he replied that the miracle of the

Quran was sufficient, and that he was but an ordinary man, who had been singled out by God as the messenger of His word. In its egalitarianism and individualism, Islam accurately reproduced the dominant mores of its tribal and mercantile Arabian environment (Lindholm, 1996).

Yet Muhammad's community, though it reflected the surrounding culture, was also something completely different than had existed previously. In their own imagination of themselves, the early Muslims were not yet another tribal grouping gathered around a sacred marketplace; they belonged to the 'party of God' – opponents were infidels. Nor was Muhammad one judge among many; he was the ultimate and final authority over a millenaristic nativist movement that aimed at nothing less than transformation of the world. Despite his protestations, for his followers Muhammad was not like everyone else; he was the perfect man who had actually seen the face of God (an event celebrated throughout the Muslim world).

In its prophecy, Islam followed in the tradition of the Middle Eastern exemplary religions already proclaimed by Moses, Jesus, and Zoroaster, who had all called for believers to submit to the moral codes imposed by an all-powerful yet personalized God: for them, faith and deeds, not heredity and power, were the true measure of spiritual enlightenment; the ethical action of individuals in this world was a prerequisite for passage into paradise (see Weber, 1978 on the nature of exemplary prophecy). Far from wishing to repudiate his predecessors, Muhammad believed his mission as 'the seal of the Prophets' was to correct and complete these earlier revelations, which had been misinterpreted and garbled. Because Muslims believe that the Quran preserves God's direct words in a pristine state, most are extremely reluctant to apply modern Western methods of textual criticism to the holy book.

The Rise of Shi' ism

While Muhammad lived, the *umma* was united under his charismatic leadership, but after his death, it inevitably began to lose its moral cohesion. Rebellions and internal schism rent the community, and after the death of Ali in 661 the realm of Islam was plunged into a cycle of tyrannical rulers and foreign invasions, punctuated by intermittent religiously inspired insurrections. Ali's own son, Husain, led one such revolt, but was pitilessly slaughtered in the battle of Karbela in 680 – an event re-enacted and mourned today in the great self-lacerating celebrations of Muharram held among Shi' ites – the 'partisans' of Ali – who are the largest sectarian group in the Middle East (Chelkowski, 1979). Today Shi' ites of different types compose nearly all the population of Iran, over 50 percent of the population of Iraq, 70 percent of Bahrain, and one-third of the populace of Lebanon. Altogether, the Shi'ites are estimated to make up 11 percent of all Muslims.

Ever since Husain's martyrdom, uprisings against the corruption of the faith have periodically been undertaken in the name of Ali and his descendants, who were believed by their partisans to be the legitimate inheritors of Muhammad's sacred authority. By restoring Ali's line to the throne, the schismatics hoped to re-establish the realm of God on earth. Usually, these movements fell victim to repression by powerful rulers, and they also suffered from internal disorganization, as there was disagreement over which of the possible claimants deserved the title of the Imam – the divinely appointed ruler. Even when successful, Ali's partisans suffered from problems of maintaining a sacred polity over several generations. For example, the great Fatimid dynasty of Egypt (969–1171) disintegrated over issues of succession, as one Imam denied his own divinity, while a later one had no sons (for a short history of Shi'ite, see Richard, 1995).

The most successful Shi'ite strategy was followed by the so-called twelvers, who presently dominate in Iran. After their eleventh Imam died without issue, it was argued that the next Imam was actually living in occlusion, and would not return until the final judgement, though he could sometimes be contacted in dreams and visions. This sect then developed a complex and rigorous educational system that aimed to produce religious scholars capable of both interpreting sacred texts and of attaining mystical communion with the hidden Imam. Such men, acclaimed by their fellows for their special spiritual gifts, became exemplary figures in themselves, and were recognized as *ayatollahs*, 'miraculous signs of God.' Every *ayatollah* had his own loyal disciples, as well as access to endowments and to donations from followers; most importantly, each retained the right to independent judgment. Only with the political ascendance of Ayatollah Khomeini in the 1979 revolution was this pluralistic system overthrown, and a new centralized spiritual authority inaugurated by the state, which proclaimed itself to be the final arbiter of truth. The future of this experiment remains to be seen, but it must be noted that it flies in the face of twelver Shi'ite tradition (Arjomand, 1988).

Sufism: The Mystical Path

A second reaction to the disintegration of the Caliphate was Sufism: so called, it is said, because practitioners wore rough woolen robes to indicate their unworldliness. Like the Shi'ites, Sufis also hoped for a renewal of the faith, but for them this could only be accomplished by a spiritual return to what they believed were the original practices ofIslam. The more ardent of them withdrew into the wilderness where they practiced severe austerities in pursuit of an immediate experience of the divine. Many devoted themselves to repetition of devotional phrases (*dhikr*) which were combined with special breathing patterns and rhythmic movement to induce spiritual illumination (Arberry, 1950; Schimmel, 1975).

By the twelfth century, the Sufi pursuit of enlightenment had become institutionalized, as saintly figures (often called *shaikhs*) founded various lodges *(tariqas)* where they taught followers their own mystical practices. Students *(murids)* were immersed in a community of brothers detached from the ordinary world, obedient to the will of their spiritual guide, who alone could provide the blue patched 'frock of blessing' and dervish cap to pupils judged to have achieved a measure of divine grace. These newly enlightened individuals could then act as spiritual guides to their own pupils, adding to the *silsila*, or chain of knowledge, reaching back to the original teacher. While the Shi'ites argued for a hidden Imam, the Sufis claimed there was a *qutb*, a pivot of the universe, a perfect man veiled from ordinary men, to be discerned only by the spiritual elite, who themselves were organized in a secret hierarchical order of

the 'friends of God,' manifested on the local level in the *tariqas* (Trimingham, 1971).

Sufism had great influence throughout the premodern Middle East, and throughout the Muslim world, as wandering mystics furnished a cohesive vision of the cosmos in the face of a splintering polity while simultaneously providing new converts with an experiential religion suitable for an unsophisticated audience. Establishing their lodges in the hinterlands around the tombs of saints, Sufis regularly served as the mediators in local tribal disputes and as guardians of the peace, gaining disciples and considerable power – so much so that they could occasionally become leaders in revolts, uniting tribes against the state (Gellner, 1981). The greatest historical example were the Safavids, who originated as followers of a Sufi saint, and who ruled Persia from 1501 to 1722. Sufism also found a receptive audience in the cities, where the ranked organizations of worker's guilds and local clubs paralleled the hierarchical order of the *tariqas*. Military groups too found the Sufi message to their taste, while intellectuals and scholars were attracted to the controlled 'sober' forms of Sufism, so much so that it was quite common for an orthodox Muslim cleric to be a Sufi adept as well – the best known example being al-Ghazali (d.1111) (Watt, 1953). In contrast, radical 'drunken' forms of Sufi worship appealed to the impoverished and sensation-seeking urban masses, who could lose themselves in ecstatic performances of *dhikr* (see Massignon, 1982).

Sufism spread with Islam to the rest of the Muslim world, where it still is a powerful spiritual force. However, in the modern Middle East, Sufism has more or less lost its significance. It only remains powerful in Turkey, where the complete assimilation of orthodox Islam into the Ottoman state meant a repudiation of orthodoxy when that state lost legitimacy, leaving Sufism as the major alternative to secularism. Elsewhere, the oppositional role traditionally played by Sufism has been taken over by various fundamentalist groups which loudly accuse Sufis of being anti-Islamic saint worshippers. Yet, like Sufi *tariqas*, these groups too are founded by zealous individuals who believe they have direct access to a purer faith, and who demand the absolute and self-abnegating loyalty of a cadre of followers. The Sufic social structure thus lives on, even though its mystical content has been denied (Roy, 1994).

Islam in Theory and Practice

While there has always been a powerful impulse in Middle Eastern Islam toward charismatic leadership, for the vast majority of Muslims the Quran remains the primal source of spiritual knowledge. Ideally there should be no need at all for human leadership when God has sent down rules to regulate and evaluate all human actions; rules which can be discovered by anyone who can read the holy text. But the Quran, being finite, does not and cannot explicitly cover all circumstances. There are many areas of behavior left out, some questions are not answered, other precepts are ambiguous, and sometimes there are contradictions or apparent impossibilities in the text.

In the face of these difficulties, some early Islamic thinkers (Mutazilites), influenced by Greek thought, attempted to force the Quran to conform to reason, but this effort was defeated in the ninth century by the 'traditionalists' led by the great theologian Ibn Hanbal (d. 855), who argued that the omnipotence of God meant that any constraints on Him and on His word – even the constraints of reason – were a form of idolatry. This did not mean that reason was banished – in fact, the use of analytic argument flourished in Muslim scholarship – but reason was firmly limited to explication of the legal implications of the sacred text; it could not be extended to put limits on God.

In place of an explanatory theology that attempted to unite logic and religion, Islam concerned itself with the obligations and prohibitions commanded by Allah: individual devotional practice and acceptance of the Word are what is crucial, not reflection or rationalization. In its emphasis on application, Islam has been appropriately described as a religion not so much of orthodoxy as of orthopraxy, reliant on ritual actions to reveal and inculcate faith (Smith, 1957). The very word 'Islam' signifies an act – the initiatory act of surrender and submission to God. The essential practices required of all Muslims are five: the public profession of the faith (God is one, and Muhammad is his Prophet – recitation of this is sufficient for conversion), five daily prayers, the pilgrimage to Mecca for those who are able, donation of *zakat*, and the month-long fast of Ramadan. Some Muslims now say holy war should also be seen as compulsory, but this is a matter of controversy, and in any case can be interpreted as a struggle against one's own base instincts.

In all instances, these shared practices are simple, direct, and spiritually compelling. The aim of each is to infuse the faithful with the necessity of a willed subordination of the self to the dictates of the overarching Muslim community while offering in return both the promise of reward in the afterlife and immediate compensation in the here and now. The human rhythm of Muslim practice is clearest in the fast of Ramadan, which is the ninth month of the Islamic lunar calendar. During this period it is strictly forbidden to eat, drink, smoke, or have sexual contact. However, fasting is not regarded by Muslims as a way of atoning for sins through suffering, as it is in the Judeo-Christian tradition. Rather, all restrictions are lifted as soon as the sun sets, so that a greater quantity of food is eaten during the evenings of Ramadan than during the rest of the year, and the meals are of higher quality. Ramadan feasts also involve celebratory gatherings of friends and family, and a public redistribution of alms and food to the needy.

The rhythm of constraint and release, solitude and community, culminates in the great Muslim festival of *Eid al-fatr* – literally the lesser festival, but in actuality the major celebration of the year (*Eid al-attar*, the 'greater festival,' is a re-enactment of Abrahamic sacrifice in which a sheep is slaughtered and shared out with family, dependents and the poor). *Eid al-fatr* is very much an affirmation of participation and renewal, as Muslims everywhere put on new clothes, visit, embrace, share food and drink, and gather for prayer. Forgiveness of all insults and enmity is mandated; debts are canceled and old enemies embrace to show their renewed brotherhood. The music and celebrations that marked the whole month are intensified, and the tombs of the ancestors are visited, reaffirming again the continuity of the community with its sacred past. High and low are leveled through the giving of alms; sharing in hunger gives the wealthy a sense of the contingency of their plenty and brotherhood with the

impoverished, who, like all other Muslims, take part in both the fast and the feast. The egalitarian and individualist message of Islam is reaffirmed through action, as salvation is achieved by the deeds of individuals in this world, who unite with their fellows in the performance of socially and spiritually rewarding acts of self-sacrifice.

Within this theology of practice there is no original sin. Mankind has been placed on the pathway and given directions, he only needs to heed and obey. The Quran is a light to guide travelers as they work their way through the darkness of earthly life toward the climactic moment of the final judgment. Redemption, atonement, an irredeemable inner sense of evil – these aspects of Christian dogma are left aside in Islam where the punishment for turning away from the path is simple and terrible – God withdraws Himself, and leaves the individual alone to wander without a guide in the moral wilderness, destined never to reach paradise.

Yet despite the absence of original sin and the emphasis on practice, Islam is not a religion of positive thinking. Election to the ranks of the saved cannot be assumed to follow from proper practice or good works; salvation is solely a consequence of God's ineffable and unpredictable grace. Fear of the utter power and unknowability of God is mitigated only by reliance on His positive qualities. In the Quran Allah has continually described Himself as merciful and compassionate, and Muslims hope for His forgiveness. Notwithstanding God's compassion, the impossibility of pleasing an utterly transcendent, unpredictable, and indeterminable Deity necessarily leaves Muslims in a state of trepidation and perplexity as to the state of their souls.

The Primacy of Memory

In their effort to please God, and fearing their own all-too-human fallibility and hypocrisy, Muslims have naturally sought other means to secure salvation beyond the study of the Quran and following its precepts. In particular, they have tried to emulate the actual practices of Muhammad and his followers (*Sunna* – the 'trodden path'), reasoning that these must serve as the most reliable template for behavior approved by Allah. As a result, Islam, more than any other religion, has always been preoccupied with discovering, preserving, and imitating its own past. Such historical memories serve, with the Quran, as the major source for Muslim moral consciousness.

So it is that where other religions interpret their holy texts through the decisions of authoritative legislative bodies, or through deduction from general principles, Muslims instead discovered principles and authority through the collection of historical records of the words and deeds of the founder and his companions. This was the task of the preservers of reported history or, as they are conventionally known, the *ahl al-hadith*, people of 'tradition,' who at first were simply individuals who had lived in the original *umma*, and could report on its practices.

Over the generations, reciters became professionals who concentrated on developing their memories and eliminating personal biases so they could repeat, verbatim, what they had heard from those before them, without interpretation, comment, elaboration, or condensation. They also had to be moral exemplars, since *hadith* were judged reliable not only on the basis of historical criterion, but also on the basis of the integrity of the reciter. Those approved could then enter into the *isnad*, or 'chain' of reliable transmitters that was recited whenever the tradition was spoken. As Islam spread, the traditions were codified, validated, and published in six so-called 'canonical' collections, which served, with the Quran, as the basis for legal decisions.

Legalism and the Problem of Interpretation

The jurists (*faqih*) who interpreted the sacred texts at first overlapped almost completely with the *ahl al-hadith*. Neither had any formal education, and both were devoted to the study of the Quran, the *sunna*, and the traditions. The difference was that the lawyers developed regulations for matters left unclear by scripture, such as the exact nature of a property transfer. They also had to clarify the weight of any particular rule: was a practice obligatory, merely recommended, or neutral; if obligatory, should failure be punished by men, or left to the discretion of Allah?

However, from the point of view of the traditionalists, the jurists' use of reason and analogy to fill in lacunae or resolve apparent contradictions in the holy texts was frivolity at best, idolatry at worst. The potentially disastrous dispute between literalists and interpretivists was classically resolved by the theologian al-Shafi (d. 820), who argued that *ijma*, the agreement of knowledgeable scholars, ratified by the community itself, decides the manner in which ambiguities in sacred texts are to be understood. In other words, Islamic theological disputes ought to be decided democratically, by consensus – albeit a consensus arrived at over generations.

This consensus is never final, since an authoritative dissent can, in principle, cause a reopening of the debate if that dissent manages to gain scholarly sanction and public endorsement. Shafi's compromise retained the egalitarian and populist character of Islam while simultaneously recognizing the importance of expert legal knowledge as the arbiter of tradition. Continuity through collective consensus still permitted the theoretical possibility for new interpretation – a possibility that continues to be the source of intense debate among contemporary Muslims.

Two centuries after Shafi's reforms the legal schools or *madhhab* (literally a 'chosen way') were whittled down to four, each named after its supposed founder: al-Shafi, Ibn-Hanbal, Malik b. Anas (d. 795), and Abu-Hanifah (d. 767), and each teaching its version of the law in specialized colleges (*madrasas*), and producing scholar-judges (*ulema* – people of knowledge). After some violent wrangling, it was eventually accepted that the schools should be considered equally accurate in their interpretations of divine law, since 'only God knows' the truth (Makdisi, 1981). Though ideally one could choose any school to follow, most Muslims accepted the *madhhab* prevalent in their area. The central Middle East remains Hanafite or Shafite, the Malikites prevail in North Africa, while the Hanbalite school remains fairly small, but is intellectually influential. Shi'ites too have their own legal schools, slightly variant from those of the Sunnis – in fact, the first college in the world, al-Azhar in Cairo, was founded in 970 by the Shi'ite Fatimids.

Even today, despite public efforts at rationalization, Islamic law has remained remarkably personalized and open to interpretation, since there is still no central ecclesiastical organization enforcing rules, nor any standard procedure for appeal to

a higher ecclesiastical court (indeed, there is no hierarchy of courts whatsoever). For these reasons, different juristconsults can issue *fatwas* (legal opinions) at odds with one another – much to the confusion of Western observers (Coulson, 1964; Goldziher, 1981).

Muslim Resistance to the State

Efforts by Middle Eastern states to incorporate, rationalize, and secularize Islamic law have always met with ambiguous results, as the majority of Muslims remain suspicious of the motives of secular authorities. For a recent example, when al-Azhar was completely integrated into the Egyptian state in 1961, its graduates were increasingly seen as minions of the government. In response, many self-taught Muslims have opened their own local schools for the training of religious scholars (Gaffney, 1994). Similar patterns have occurred elsewhere in the Middle East, where popular mistrust of the official *ulema* has periodically led laymen and reformers to try to find a pathway to a purer and more egalitarian form of Islam.

This is so much the case today that, except in Iran, those offering an 'Islamist' critique of the state are usually autodidacts self-consciously distancing themselves from the class of the traditionally religiously educated, who are tainted by their graduation from state schools. But although ordinary Muslims are suspicious of the state, they are equally distrustful of those zealots who say they have privileged access to the truth. The best course, they believe, is to avoid conflict with the secular authorities, submit to the precepts of the Quran, *sunna,* and *hadith,* and rely on Allah's infinite mercy.

See also: Historical Thought and Historiography: Islamic Tradition; Islamic Fundamentalism; Middle East, North Africa: Sociocultural Overview; Near Middle East/North Africa Studies: Religion; Prophetism; Religion, Family, and Kinship; Religion: Evolution and Development; Religion: Morality and Social Control; Religion: Nationalism and Identity; Religious Fundamentalism: Cultural Concerns; Religious Nationalism and the Secular State.

Bibliography

Arberry, A.J., 1950. Sufism: The Religious Attitude and Life in Islam. University of Chicago Press, Chicago.
Arjomand, S.A., 1988. The Turban and the Crown: The Islamic Revolution in Iran. Oxford University Press, New York.
Bulliet, R., 1994. Islam: The View from the Edge. Columbia University Press, New York.
Chelkowski, P. (Ed.), 1979. Ta'ziyeh: Ritual and Drama in Iran. New York University Press, New York.
Coulson, N., 1964. A History of Islamic Law. Edinburgh University Press, Edinburgh, UK.
Crone, P., 1987. Meccan Trade and the Rise of Islam. Princeton University Press, Princeton, NJ.
Crone, P., Cook, M., 1977. Hagarism: The Making of the Islamic World. Cambridge University Press, Cambridge, UK.
Gaffney, P., 1994. The Prophet's Pulpit: Islamic Preaching in Contemporary Egypt. University of California Press, Berkeley, CA.
Gellner, E., 1981. Flux and Reflux in the Faith of Men. In: Gellner, E. (Ed.), Muslim Society. Cambridge University Press, Cambridge, UK.
Gibb, H.A.R., 1962. Studies on the Civilization of Islam. Princeton University Press, Princeton, NJ.
Goldziher, I., 1981. Introduction to Islamic Theology and Law. Princeton University Press, Princeton, NJ.
Hodgson, M., 1974. The Venture of Islam: Conscience and History in a World Civilization. University of Chicago Press, Chicago. Vols 3.
Humphreys, R.S., 1991. Islamic History: A Framework for Inquiry. Princeton University Press, Princeton, NJ.
Lapidus, I., 1988. A History of Islamic Societies. Cambridge University Press, New York.
Lindholm, C., 1996. The Islamic Middle East: An Historical Anthropology. Blackwell Publishers, Oxford, UK.
Makdisi, G., 1981. The Rise of Colleges: Institutions of Learning in Islam and the West. Edinburgh University Press, Edinburgh, UK.
Massignon, L., 1982. The Passion ofal-Hallaj: Mystic and Martyr of Islam. Princeton University Press, Princeton, NJ. Vols 4.
Mitchell, R.P., 1969. The Society of the Muslim Brothers. Oxford University Press, Oxford, UK.
Nelson, K., 1986. The Art of Reciting the Quran. University of Texas Press, Austin, TX.
Richard, Y., 1995. Shi'ite Islam: Polity, Ideology and Creed. Blackwell Publishers, Oxford, UK.
Rodinson, M., 1973. Islam and Capitalism. Pantheon, New York.
Roy, O., 1994. The Failure of Political Islam. Harvard University Press, Cambridge, MA.
Schimmel, A., 1975. Mystical Dimensions of Islam. University of North Carolina Press, Chapel Hill, NC.
Shaban, M.A., 1971. Islamic History A.D. 600–750 A.H. 132. A New Interpretation. Cambridge University Press, Cambridge, UK.
Smith, W.C., 1957. Islam in Modern History. Princeton University Press, Princeton, NJ.
Torry, C., 1892. Commercial-Theological Terms in the Koran. E. J. Brill, Leiden, The Netherlands.
Trimingham, J., 1971. The Sufi Orders of Islam. Oxford University Press, Oxford, UK.
Wainsbrough, J., 1977. Quranic Studies: Sources and Methods of Scriptural Interpretation. Oxford University Press, Oxford, UK.
Watt, W.M., 1953. The Faith and Practice of al-Ghazali. Allen and Unwin, London.
Watt, W.M., 1973. The Formative Period of Islamic Thought. Edinburgh University Press, Edinburgh, UK.
Watt, W.M., 1988. Muhammad's Mecca: History in the Qur'an. Edinburgh University Press, Edinburgh, UK.
Weber, M., 1978. Economy and Society. University of California Press, Berkeley, CA.
Wellhausen, J., 1927. The Arab Kingdom and its Fall. Curzon Press, London.

Islam: Sub-Saharan Africa

Peter B Clarke, King's College, London, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 7927–7931, © 2001, Elsevier Ltd.

Abstract

Nigeria's Muslim population, generally estimated to be over 50 million, is the largest in this part of Africa, and the focus in this article is on the growth and impact of Islam there, and in East Africa, especially in Tanzania. Broadly speaking, there have been three stages in the development of Islam in tropical Africa: quarantine, mixing, and reform. In the first of these stages Muslims, mainly merchants from North Africa, tended to live in closely knit communities and apart from the indigenous population. The second stage was characterized by the mixing of Muslim and traditional beliefs and rituals, a process which gave rise to the third stage, that of reform. The article is particularly concerned with the role of the scholar-mystic as reformer and nationalist and the impact of colonialism and Islam in the context of the post-independence nation-state.

An estimated 140 million Muslims live in tropical Africa, also referred to as sub-Saharan and Black Africa. The majority are concentrated in the Sahelian and Savannah regions that lie to the south of the Sahara proper, between the Atlantic to the west and the Red Sea to the east, in the Horn of Africa and along the East African coast from Somalia to the northern region of Mozambique. Nigeria's Muslim population, generally estimated to be over 50 million, is the largest in this part of Africa, and the focus in this article will be on the growth and impact of Islam there, and in East Africa, especially in Tanzania where, as in Nigeria, there is also a heavy concentration of Christians, a situation that affects the character, policies, and objectives of both religions in both countries. Elsewhere either Muslims or Christians are in a clear majority.

The advance of Islam and Christianity throughout tropical Africa has not undermined all traditional beliefs and rituals. On the contrary, as modernization has brought new challenges to community and personal life, people have not rejected the traditional belief systems for the 'newer' religions but have tended to juxtapose – the term syncretize is misleading – them, and to have recourse to either one or the other depending on the concerns that require a spiritual solution.

Broadly speaking, there have been three stages in the development of Islam in tropical Africa: quarantine, mixing, and reform. In the first of these stages Muslims, mainly merchants from North Africa, tended to live in closely knit communities and apart from the indigenous population. The second stage was characterized by the mixing of Muslim and traditional beliefs and rituals, a process which gave rise to the third stage, that of reform.

West Africa Until the Nineteenth Century

It was through commerce that, over 1000 years ago, Islam began to spread to West Africa and indeed to most of tropical Africa. During the eighth century, Muslim merchants of both Arab and Berber origins began to work their way across the trans-Saharan trade routes to the Ancient kingdoms of Ghana and Kanem Borno situated within the frontiers of what is today western Africa.

After gradually establishing itself as the religion of court and commerce, Islam came to be seen as a progressive, modernizing force in Tropical Africa. The traveler Ibn Battuta (1304-68/77) left an account of the healthy condition of Islam in Mali in the middle years of the fourteenth century. He was impressed by the meticulous observance of the Friday prayers, the importance attached to the learning of the *Qur'an* by heart and the fidelity with which *zakat* or almsgiving was practiced. Ibn Battuta also pointed to what has been a feature of Islam in West Africa since its beginnings there: mixing. He noted the retention in Muslim circles of traditional religious customs, including divination, prostration before a ruler, and the matrilineal system of inheritance, considered to be contrary to Islam, and given by reformers as the principal reason for *jihad*.

Early on Islam encountered Christianity in tropical Africa. By the middle of the seventh century, Arab Muslim traders had come into contact with the Christian kingdom of Nubia which, when unified, stretched from the second to the sixth cataract of the Nile, although there was little by way of Islamization until the arrival of the Mamluk dynasty in Egypt in the second half of the thirteenth century. A combination of military pressure from Egypt and royal marriages between the Egyptian and Nubian dynasties resulted in the establishment of Muslim rule in the capital of the Nubian state of Muqurra in the first half of the fourteenth century.

The Scholar-Mystic as Reformer and Nationalist

While Islam in tropical Africa has been spread largely by peaceful means and has used the same means to reform itself, the launching of the Almoravid movement in the southern sector of the Sahara in 1054 by Abdullah b. Yasin initiated a tradition of militant reform that was to influence the course of West African Islam down the centuries, inspiring future *jihads* or holy wars.

In the fifteenth century, this tradition was given intellectual and doctrinal weight by the Muslim scholar Al-Maghili (1425–1502) from Tlemcen in southern Algeria, who spent many years teaching Islam and counseling rulers in Hausaland (present day northern Nigeria) and the central Sudanic kingdom of Songhay (present day Niger). His radical teachings on the Shari'a (Islamic law), Muslim government and education, gender and taxation were communicated in reform manifestos to scholars and rulers across the western Sudan.

The seventeenth century saw the beginnings of a series of *jihads* in western Africa culminating in the Sokoto *jihad* which was to change the cultural and religious landscape of that region during the first half of the nineteenth century. It was launched by the scholar and mystic from the hinterland of Hausaland in northern Nigerian, Shaykh Uthman dan Fodio (1754–1817) early on in the thirteenth century of the Muslim *hijra* (1785–1882) – the century which, it was widely believed, would witness the advent of the Mahdi (God-guided one) who would ensure the triumph of Islam everywhere. The principal target of this jihad, one of the most thoroughgoing in the history of Islam, were the nominally Muslim urban based government administrators and the jurists who were accused of every kind of corruption and oppression. Oppression of fellow Muslims, Shaykh Uthman preached, was tantamount to infidelity on the ground that a believer was duty bound to befriend a fellow believer.

The Madhi, many were convinced, had arrived in the eastern Sudan in 1881 in the person of Muhammad Ahmad. A member of the puritannical *Sammaniyya* Sufi order, Muhammad Ahmad, in a brief period, overthrew Egyptian rule in the Sudan and established an Islamic state there which lasted until 1898. Like that of Shaykh Uthman dan Fodio, the Sudanese Mahdi's religious nationalist movement was both rural based and directed against the corrupt urban based elites, mostly Turco-Egyptian officials. The careers of both men highlight the crucial role of the mystic as reformer in African Islam.

After the Madhi the Sudan was to become, in theory, a condominium under joint British and Egyptian rule until 1952. The Sudan has suffered from a civil war between the largely Muslim North and the Christian, traditionalist South since 1955. Since that time several administrations have been engaged in extending Islamic law and education across the South.

Islam on the East African Coast Until the Nineteenth Century

As in West Africa, Islam's early development on the East Coast was slow. It spread more easily and evenly among the nomadic and seminomadic populations of Somalia and the Horn generally, while further south it remained confined to many of the island populations including those of Lamu, Pate, Pemba, Comoros, Mombasa, and Zanzibar, and to a few settlements on the coast itself. The best documented of the early coastal settlements is Kilwa, founded by Shirazi immigrants from the Arabian peninsula in either the tenth century or in the twelfth century. Though early Islam was largely confined to the coast Arab immigrants had established a Muslim state far inland in Shoa in Ethiopia by the ninth century which was later absorbed by Ifat, one of seven Muslim kingdoms all tributaries of Christian Ethiopia.

By the first half of the fourteenth century a well-developed Muslim society had been established on parts of the coast of East Africa. Once again Ibn Battuta, the chronicler and traveler, who visited Mogadishu in Somalia in 1331, provides a picture of the state of Islam at the royal palace at this time. He describes a deeply Islamized court around the Arabic speaking ruler and/or shaykh who was assisted by a Muslim judge or *qadi* from Egypt. Furthermore, East African students had already begun to travel to study in the Islamic university of Al-Azhar in Egypt, a tradition begun in West Africa much earlier.

Though historically less militant than western African or Sudanese Islam, the fusion of Islam with local coastal customs and traditions created the context for reformers to engage in *jihad* or holy war. The reformer Ahmad b. Ibrahim (1506–43), better known as Ahmad Gran, waged several *jihads* from the Muslim state of Adal, situated in the interior, into the Christian kingdom of Ethiopia which involved not only the purification of Islam but also the conversion of Christians. Paradoxically, the indirect consequence of those *jihads* was the disintegration of the largely Muslim region of Adal, the main area of relatively settled Islamic life in the interior, although Harar, to the south, remained for Somalis a center of Muslim orthodoxy. In the second half of the sixteenth century the Galla, originally from south central Ethiopia, began to expand and by the nineteenth century a number of ethnic groups further north including the Wallo were largely Muslim.

Though Islam made little further progress in the East African interior until the nineteenth century, Arab Muslim culture had begun to interact with the indigenous Bantu cultures through marriage and concubinage to produce the new and distinctive culture known as Swahili, an Arabized-Bantu lingua franca spoken in Kenya, Tanzania, and in the eastern sector of the Democratic Republic of the Congo. The Muslim presence on the coast, on the other hand, was strengthened by the arrival of the Ottomans who occupied Masawa in 1557, and came into conflict with the Portuguese who had burnt down Zayla' in 1517.

East African Islam Turns Inland

The Yao people, situated midway between Lake Nyasa and the Indian Ocean and part of a trading network that included slave trading stretching from the Indian Ocean west to Katanga, were the only major ethnic group south of Somalia to turn Muslim prior to the nineteenth century. During the first half of that century increasing economic expansion inland began to effect a profound reorientation in outlook, attitude, and priorities among the Muslims of the coastal settlements. Muslim traders and agriculturalists began to move into the interior to Tabora, Ujiji, Nyangwe, Buganda, Lake Nyasa, Katanga, and elsewhere, and while not actively proselytizing they were, nonetheless, spreading Islamic culture by imitation.

The first Muslims arrived in Buganda, part of present day Uganda, from Zanzibar in the mid-nineteenth century during the reign of Kabaka Suna. Arab influence increased in the 1860s under Suna's successor Mutesa, not a full Muslim himself (he was not, for example, circumcised) and this was manifest in Buganda's adoption of the Islamic calendar, the observance of Ramadan, and the wearing of Arab dress which had become fashionable.

Tension created by the uneasy relationship between the demands of Islam and local custom and culture was never far below the surface and erupted in violence as protests by strict, orthodox Muslim visitors to Buganda against Mutesa's leading of the Friday prayers led to a backlash in which 100 Muslims were massacred. The situation was further complicated by the arrival of Christian missionaries in Buganda in 1877. Though in 1888 Christians and Muslims united to expel Mutesa's successor Mwanga, too much a pagan for either, unity soon turned to

hostility as the Muslims expelled the Christians and installed Kalema as Kabaka. A civil war followed and by 1890 the Christians had gained the upper hand and were to receive the backing of the colonial regime. By the treatises of 1890 and 1892, Captain, later Lord, Lugard, acting in the name of the Imperial British East African Company, allocated Bugandan landed chieftainships along religious lines in favor of Christians and this was confirmed by the Ugandan Agreement of 1900. Christianity's dominance in Uganda had been institutionalized.

Colonialism and Islam

Muslim resistance to colonialism has been mentioned in the case of the Sudan and Muslim opposition, both passive and militant, emerged in many other parts of tropical Africa including the Senegambia, Somalia, and Northern Nigeria.

While profoundly disturbing culturally, psychologically, and spiritually everywhere in tropical Africa, the indirect consequences of colonial rule on the actual spread and institutional development of Islam were uneven both in time and place and depending on colonial aims, interests, rivalries, and manpower requirements. At first the German use of Muslims for administrative purposes in East Africa contributed to Islam's progress in the interior, while later this same practice worked against its advance by dividing government personnel from the indigenous peoples opposed to foreign domination. The strongest opposition to German colonial rule came in the form of the Maji-Maji (*maji* in Swahili means water) rebellion (1905–07), organized by Kinjikitile, once again a Sufi or Muslim mystic.

British colonialism which replaced the German system after World War I tended to slow the progress of Islam in Tanzania and elsewhere in East Africa. The system of Indirect Rule at the level of local government tended to protect the power of the local chiefs from being undermined by wealthy Muslim merchants from the coast. Moreover, Christian schools became the main supply of local administrators, teachers, police, and army personnel. The perception was formed of Christianity as the religion of progress, while Islam, as it was gradually sidelined, came to be seen as an obstacle to advancement in the modern sector of the economy and politics.

By contrast with East Africa, British colonial rule in Northern Nigeria relied heavily on Muslim personnel to implement the system of Indirect Rule and even extended Muslim authority to cover the so called 'pagan areas' such as the Jos Plateau. Muslim courts and schools were improved and, as in the Sudan, Colleges of Islamic Law were opened. What mostly assisted the spread of Islam during the colonial era was the improvement in communications, while what mostly disadvantaged Islam and assisted Christianity was the growing importance of western education.

Postindependence Islam in the Context of the Nation State

The independence struggle and the regaining of independence itself provided Muslims, previously under represented in the modern sector of political life and in the civil service in many parts of Tropical Africa, with the opportunity to insert themselves in a more positive and sometimes aggressive manner into the political affairs of the new nation state.

Since independence in 1963, Nigerian Muslims have created a nationwide network of Islamic institutions and have supported the territorial integrity of the federation when threatened by seccessionist movements which continue to exist and the most serious of which, in recent times, came from the eastern region and resulted in the Biafran Civil War (1967–70). On the other hand, in the late 1970s as the country prepared to return to civilian rule after more than a decade of military rule, Muslim members of the Constituent Assembly appointed to approve the Constitution of the Second Republic became embroiled in a serious controversy over the establishment of a Supreme Shari'a Court at the Federal level that threatened the unity of the nation. Many Nigerians not only resented and felt threatened by the Muslim demand for a Federal Shari'a Court but also by the Muslim call for the removal from the preamble to the 1978 draft of the Constitution of the Second Republic the term 'secular state' on the grounds that it promoted secularism, atheism, and immorality. Those skeptical of northern Muslim intentions and more favorable, if not to seccession, then to greater regional autonomy, claim that Muslim enthusiasm for a strong central government reflects a concern not with harmony and unity but over the resources of the country, principally oil, which are located in the South.

Evident in tropical Africa since the 1970s is not only the greater role of Islam in the affairs of the nation state but also the steady increase in Muslims in what were traditionally non-Muslim areas such as Meru country in Central Kenya. The precise number of Muslims in Kenya is unknown but the best estimate would seem to be somewhere between 10 and 20 percent of the population. Islam is also expanding among the black population in South Africa where it was first introduced by slaves and political prisoners from the Dutch East Indies, Madagascar, and West Africa in the late seventeenth to early eighteenth century. Numbers were boosted by the influx of Muslim traders, many of them from Bombay, who distinguished themselves from the indentured laborers from India, a majority of whom were Hindu, who began arriving from 1860.

In the 1980s Muslims joined with Christians of all denominations, Jews, and Hindus in South Africa against the enforced pluralism of apartheid. Today most of South Africa's major cities have a sizeable Muslim presence, and increasing attention is given to education, community integration, and links with the wider Muslim world in Africa and beyond from which it was largely cut off during the apartheid era.

Conclusions

It is usual to stress the fit between Muslim culture and the African condition, and while there was, and is, undoubtedly much common ground between the two, Islam, like Christianity, has been both a destructive and creative force. It has produced many distinguished scholars, mystics, and philanthropists, among them a Muslim William Wilberforce in the person of Shaykh Uthman dan Fodio (1754–1817) of Hausaland in present day northern Nigeria whose writings and program of reform continue to influence and motivate contemporary Muslims not only in Africa but also in the United States and Europe.

Islam has been a major agent of globalization in tropical Africa. Wherever it has settled it has brought numeracy and literacy in Arabic and turned what were regional and/or national trading and commercial networks into international ones. It has also introduced new forms of religious community and spirituality in the form of the *turuq* or Sufi brotherhoods that link believers in tropical Africa to those in the wider Muslim world.

Whether Islam, or for that matter Christianity, has been an innovator or simply a catalyst, where the introduction of such core beliefs as belief in the Supreme Being as totally unique and independent is concerned, is both a fascinating and much debated point among social scientists and historians of religion, but one which space prevents us from discussing here.

See also: Historical Thought and Historiography: Islamic Tradition; Islam: Asia; Islam: Middle East; Islamic Fundamentalism; Middle East, North Africa: Sociocultural Overview; Near Middle East/North Africa Studies: Religion; Religion, Sociology of.

Bibliography

Clarke, P.B., 1990. Islam in Tropical Africa in the Twentieth Century. In: Clarke, P.B. (Ed.), Islam. Routledge, London, pp. 180–192.

Hunwick, J.O., 1900. Islam in Tropical Africa to c. 1990. In: Clarke, P.B. (Ed.), Islam. Routledge, London, pp. 164–179.

Lohdi, A.Y., Westerlund, D., 1999. Tanzania. In: Westerlund, D., Svanberg, I. (Eds.), Islam Outside the Arab World. Curzon Press, Richmond, UK, pp. 97–110.

Nimtz, A.H., 1980. Islam and Politics in East Africa: The Sufi Order in Tanzania. University of Minnesota Press, Minneapolis, MN.

Pouwels, R.L., 1987. Horn and Crescent: Cultural Change and Traditional Islam on the East African Coast. Cambridge University Press, Cambridge, UK.

Rasmussen, L., 1993. Christian-Muslim Relations in Africa: The Cases of Northern Nigeria and Tanzania Compared. Academic Press, London.

Trimingham, J.S., 1964. Islam in East Africa. Books for Libraries, New York.

Westerlund, D., 1980. Ujamaa na Dini: A Study of Some Aspects of Society and Religion in Tanzania. Almqvist and Wiksell International, Stockholm.

Islamic Fundamentalism

Youssef M Choueiri, School of Arts, Languages and Cultures, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Islamic fundamentalism came into being as a sociopolitical movement in the twentieth century. It was largely a reactive ideology that was directed against what it deemed to be the complete secularization of states and societies. Such an outcome was related to a variety of systems and worldviews, ranging from democracy and nationalism to socialism and liberalism. Consequently, it perceived Islam as the main victim of this relentless onslaught, whereby its principles, regulations, and strict code of Shari'a law were no longer observed or adhered to, even in Muslim societies. Such a state of affairs called therefore for the establishment of a new revolutionary organization, espousing Jihad as its method of struggle and fully committed to rebuilding a totalitarian Islamist state.

Islamic fundamentalism is a political and ideological phenomenon of the twentieth century. Born in the 1920s, it gathered momentum under the impact of social, political, cultural, and economic changes in Muslim lands in the last quarter of the twentieth century. Taking its cue from movements of Muslim Reform at the end of the nineteenth century, Islamic fundamentalism dented and finally rejected the apologetic modernism of Muslim Reform, while appealing to a return to Muslim origins as announced in the Koran.

Islamic fundamentalism projected an imagined golden age as a social project to be copied in the present, married this to populism typical of the twentieth century, and propagated an antimodernist social and cultural program for resocialization. These included separating men and women in public spaces, turning the veil in its extreme form into a religious duty, and forming vigilante-armed groups to enforce Islamic rituals and obligations. It rested in its moderate forms on social conservatism, but its radical tendencies borrowed ideas of revolution and total societal transformation. It is a movement of varying incidence and depth in Muslim lands and among Muslim Diasporas, with particular appeal to the subaltern intelligentsia and to structurally marginalized sectors of the population, both of whom regard the socially expansive definition of religion a solution to a variety of crises and historical impasses.

Islamic fundamentalism grew among sections of Muslim peoples worldwide in the twentieth century. It is not a unitary and coherent movement but represents a constellation of political, social, ideological, and intellectual trends with varying orientations and national settings.

General Features

Islamic fundamentalism is too much a product of modern history to admit to being regarded, as is often the case, as *sui generis*. It is much too severed from Muslim learned and popular traditions over centuries to admit to being considered, as is habitually presumed, an authentic, albeit condensed, resultant of these traditions. Insofar as it represents the political expression of a particular attitude toward scriptural material, it is a phenomenon generically belonging to the history of religions, with distinctive affinities to the other monotheistic fundamentalisms, Judaism, and Christianity, as well as to other fundamentalisms that are recasting themselves along monotheistic patterns, namely Buddhism, especially in Sri Lanka, and to a lesser extent Hinduism. Insofar as this attitude toward scripturalist religious origins is made to converge with the modern political ideologies that power modern political movements, Muslim fundamentalism subtends conceptions of history, society, state, and of political action that quite naturally freely draw on the universal modules of populism, vitalism, and the entire repertoire of standard anti-Enlightenment motifs in the history of European and non-European peoples.

Finally, insofar as Islamic fundamentalism, in its wide variety of national settings, ideological and sectarian inflections, and political orientations, was an ideology of increasingly wide appeal in the closing decades of the twentieth century, it shares the common socioeconomic characteristics of contemporary subalternist populist movements everywhere, both as diffuse sentiment and as centrally organized, party political movements of contestation and of revolution, mobilizing crowds as well as building up organizational structures and, in some extreme cases in Egypt, Algeria, Afghanistan, and some Western countries, resocializing their members completely and constructing countersocieties.

Intellectual Origins

Muslim Reformism

The ultimate intellectual, as distinct from social, origins of Islamic fundamentalism go back to the latter part of the nineteenth century. Much like right-wing European nationalist and royalist ideologies on the ascendant at the time, and in line with the long tradition of sentiment and polemic against the Enlightenment and the French Revolution associated with, among others, Johann Gottfried von Herder 1744–1803, Edmund Burke 1729–97, and Joseph de Maistre 1753–1821, these were premised upon a vitalist conception of society and history. At the heart of this conception lay the contention that societies and histories constituted natural organic wholes sundered and rendered into a condition of morbid distemper by the historical processes of modernity in its violent

revolutionary changes…. While these processes in Europe were identified as French Republicanism, socialism, and anticlericalism, in Muslim domains, and particularly in Ottoman lands, which were at the forefront of rapid change among Muslim peoples, they were identified with the state-led reforms in the domains of law, educational and cultural advancement, and modernist social engineering – all of which were galvanized by a positivist notion of historical evolution correlative with the assumption by the Ottoman state of direct control on the Bonapartist pattern, which saw cultural and legal hegemony as a primary task and prerogative.

Just as Herder and other conservative Germans derided the court of Frederick the Great as effeminate and 'inauthentic,' so the subaltern rungs of the emergent Ottoman bureaucracy and modern, nonclerical intelligentsia saw the reformist Ottoman state and the upper rungs of the state administration as being captive to European domination. These groups likewise saw this domination personified in westernized lifestyles and progressive ideas. Expressed initially by the Young Ottomans, these sentiments were vigorously expressed and most consequentially by the Persian Jamal al-Din al-Asadabadi, alias al-Afghani (1839–97) and his then disciple, the Egyptian Muhammad 'Abduh (1849–1905), through the medium of the short-lived, but highly suggestive, and influential review, *al-'Urwa al-Wuthqa*, *The Indissoluble Bond* (a Koranic expression), published in Paris in 1882–83.

It was suggested in this review that nations are corporate entities which cohere around moral drives, a notion which bears close comparison with Herder's notion of *Kräfte*. Once adulterated by foreign influences, these drives tend to atrophy, and to weaken the body-national which, in this social Darwinist world of nations, is perforce made subject to stronger parties. Thus history is a cyclical process, an alternance of purity and adulteration, corresponding to phases of power and powerlessness: in the history of Islam, *al-'Urwa al-Wuthqa* traced the causes of adulteration of the early Muslim empire to non-Arab newcomers to the faith, and to heterodox sects, and the adulteration of the Ottoman polity to the infiltration of modernist ideas and institutions of European provenance. Afghani and 'Abduh discounted race and ethnicity from their definition of a nation, and viewed religion as 'the sure bond' of Muslim nations. It was the Muslim religion, very broadly defined in culturalist and utilitarian rather than in devotional or dogmatic terms, which caused the organic unity of the nation to cohere, to acquire direction, and to gather force. Political action was thus conceived as an educational enterprise for the broad mass of people captive to superstition, and a conspiratorial political task directed at winning over prominent statesmen who might provide the authoritarian armature of this historical project.

Over subsequent decades, and until the late 1920s, the two prongs of this enterprise bifurcated, with Afghani, a charismatic and highly complex arch-conspirator but a superficial thinker, failing to build a Muslim internationalist coalition. 'Abduh, who became Grand Mufti of Egypt, independently charted the intellectual and cultural highlights of a lasting religious reform, explicitly inspired by the example of Luther, whose main thrust was to combat what he regarded as superstitious and irrational accretions that overcame what he took for the original Koranic foundations of the Muslim religion: folk religiosity, traditionalism, clericalism, and virtually all interpretations of religion that were contrary to modern science (including Darwinism) and modernist forms of political, legal, and social organization. At 'Abduh's hands, Islam became highly latitudinarian and episodically regarded almost as a natural religion corresponding to the social evolution of humanity. Reason and individual judgment *(ijtihad*) were highlighted as the true characteristics of Islam. Hence the modernist and reformist labels attached to this movement.

From Reformism to Fundamentalism

It was with 'Abduh's star pupil, the Syrian Muhammad Rashid Rida (d. 1935), initially as liberal as his teacher, that what might be termed political Afghanism became ascendant. In his monthly review *al-Manar* (1898–1934), published from Cairo and prodigiously influential all the way from Morocco to Java, he laid the foundations of an alliance between social conservatism and a cultural–political activism in the name of a religious reform shorn of its previous latitudinarianism, retained as a slogan voided of practical consequence, and with a pronounced antimodernist polemical tone that sought to expand the definition of religion to encompass the entire surface of social life, rendering politics, in the words of a prominent follower, 'social devotions.' The emphasis on the strict observance of ritual, the emblematic use of politicocultural spectacle such as extreme forms of punishment, the vituperative strictures against equally emblematic manifestations of modernity, particularly secularism and the social liberation of women, vigorous opposition to nearly all but literalist interpretations of the Koran and to miraculous and salvational views of Islamic history, and the hostility to modern legal reforms in the name of fidelity to Koranic and classical norms, propounding the 'restoration' of the *Shari'a* as a fundamental political slogan, and the revival of the notion of apostasy. These became leitmotifs of all subsequent Islamic fundamentalisms in the Arab World and beyond.

This expansive definition of religion was influenced by the example of the literalist revivalism of the Wahhabi movement which led to the foundation of the Kingdom of Saudi Arabia in 1932. This movement, born in a deprived and isolated territory during the era of the French Revolution, represented a ritualistic Puritanism which referred directly to the Muslim scriptures and certain fundamentalist figures in Muslim traditions, at the same time impoverishing these considerably, and reducing them to formulism. It is in this context that one begins to see that revivalism preceded reformism as a political movement in the Muslim world. It denotes Islamic movements that emerged in the eighteenth and nineteenth centuries. It was mainly confined to peripheral areas lying beyond the reach of central authorities. Its social base had a predominantly tribal formation. It articulated its outlook in the form of an internal Islamic dialogue. In the 1740s Central Arabia witnessed the formation of such a movement under the religious guidance of Muhammad b. 'Abd al-Wahhab and the political leadership of Muhammad b. Al-Sa'ud.

The Mahdist uprising in the Northern Sudan in 1881–85 was one of the last revivalist movements. This state survived

until 1898 following Lord Kitchener's reconquest. On the other hand, Islamic Reformism was an urban movement that came into being in the nineteenth century, and lasted well into the twentieth. Its leaders were state officials, intellectuals, or 'ulama' (religious leaders) fiercely opposed to traditional interpretations of religion. It conducted an open dialogue with European culture in an attempt to stem the tide of decline. By the middle of the twentieth century, both revivalism and reformism were superseded by the emergence of sovereign nation-states throughout the Muslim world.

These states carried out varied programmes of development and relegated Islam to a personal sphere (Turkey), or co-opted Islamic reformist ideas into state structures. Islam was for the first time dissected and revaluated under the watchful eyes of advancing expansionist Western European norms and concepts were inevitably borrowed or recast in this comparative endeavour. This was succinctly expressed by the Ottoman reform movement known as the Tanzimat 1839–1876 or 1800–1905, for the whole period of reforms. Reformists embarked on a search of viable cultural models. European progress was deemed as the direct outcome of intellectual developments leading to political liberty and economic progress. Martin Luther (1483–1546) was often considered a latter-day Muslim anxious to combat superstitions and restore religion to its original progressive nature.

It was in this historical juncture that the fundamentalism of Rashid Rida was widened to include an entire programme of political and social rejuvenation.

Culture into Politics

To this expansive definition of the religious sphere was added an organized political moment in the 1920s and 1930s with the emergence of the Society of Muslim Brothers, first in Egypt, later spawning similar movements elsewhere.

Incipient Organization

Like subsequent moments in the history of Islamic fundamentalism, this political emergence followed the rhythms and patterns of international history which, in the interwar period, was characterized by the anti-Enlightenment and antirationalist vitalist motifs that provided ideological sustenance to racism, hypernationalism, xenophobia, and cognate trends, both organized and socially diffuse. Thus, the Society of Muslim Brothers and its Muslim and Hindu homologs in India, the Jamaat-i Islami and the Rashtriya Swayamsevak Sangh, for instance, conceived themselves as sporting and paramilitary youth organizations, ostensibly beyond party and class politics in the name of the unity and homogeneity of the nation, conceiving of the state as a natural and organic emergence from a homogenous society, fully corresponding to its essence, momentarily in diremption and in the course of restoration to an Arcadian past prior to its violation and corruption by inauthentic modernity.

This rhetoric of history underlying a political project of hegemonism and of the seizure of state power by means both gradual (emphasizing reeducation and resocialization) and insurrectional, was given added ideological and systematic force and relevance in the period of the Cold War, with large-scale borrowing from state-led nationalism in Muslim countries and from populist conceptions of popular will, interpreted as fundamentally and abidingly religious, whereby Muslim peoples were thought to be, by natural instinct, Islamist in political orientation. This explains to a large degree the extraordinary degree of social, symbolic, and physical violence needed for nurturing these movements in their quest for power. Correlatively, these movements derived concrete logistical, institutional, and organizational sustenance from the United States and from Saudi Arabia, and may be said to have constituted the cultural plank of the Truman Doctrine, to which military and political aid was added, most dramatically in the Malay Peninsula in the 1950s, in Indonesia in 1965, and in Afghanistan in the 1980s.

Radical Islamism and Society

The main ideological innovation in this third phase of Islamic fundamentalism resides in the notion of *takfir*, the notion that Muslim societies are not only aberrant and unfaithful to their natural history, but that they are territories of degeneration and godlessness in need of a total recasting along models derived from the prophetic period in Islamic history, now transformed into a golden age. First propounded by Abul-A'la Mawdudi (1903–79), the founder of the Jamaat-i Islami in India and later its leading light in Pakistan, it was taken over by the Egyptian Sayyid Qutb (1903–66), crossed with the notion of degeneration current in right-wing European political thought during the *fin-de-siècle* and the interwar periods (prominent here was the influence of Alexis Carrel), and provided the cornerstone of the extremist fundamentalism of the 1990s: among certain groups in Egypt, Algeria, and Afghanistan, and more episodically elsewhere. Most specifically, the civil war in Afghanistan spawned a class of itinerant professional Islamist revolutionaries of various nationalities who took up arms in Algeria, Egypt, the United States, Bosnia, Kashmir, and elsewhere.

At the close of the twentieth century, further elements were added to the ideological and activist repertoire of these movements, some of which had taken over state power (as in Iran, Afghanistan, and Sudan) or became potent players on the national scene (as in Pakistan), where they have been directing operations of social engineering aimed at recasting laws and public mores. The most important of these elements, again reflecting international flavors of the time, is the hypernationalist use of the vitalist notions of 'identity,' of cultural difference, and their cognates in a situation marked by the waning of projects of development and their cultural and social correlates, the weakening of states in the developing world under conditions of global deregulation, including cultural and social deregulation. These derive sustenance from armies of the slum-dwellers, the structurally unemployed, small businessmen belonging to a previous age, and subaltern intellectuals including members of the clerisy, most potently, and successfully in Iran, where the breakup of the modernist state was most spectacularly displayed in the revolution of 1979 which has highly specific ideological and organizational characteristics.

An expansion of religious thought, in the name of culturally specific protocols of knowledge, led to an auxiliary and the most important element that sustains fundamentalist bids for intellectual hegemony, namely the institutional and

conceptual development of the 'Islamization of knowledge': Islamic social science, Islamic economics, Islamic democracy, the renewed hostility to Darwinism, and Islamic feminism. These currents of thought exercised collateral influence on sectors of political thought in many Muslim countries, some of them irreligious and hostile to political Islamism, and certain of these vocabularies seeped into the mainstream of political discourse, quite apart from state-sponsored religion which had hitherto been marked by a latitudinarian reformism, now feeling the need to protect its flanks and pushed into a more conservative direction.

Radical Islamism was thus a reactive movement that aspired to reinvent Islam according to its interpretation of the trajectory of world history, pronounced to be utterly secular, and consequently atheistic in its ultimate configurations. Hence, all societies, including Muslim countries, were deemed to have entered a new age of ignorance (*Jahiliyya*). This tragic calamity called therefore for radical solutions that had to cast aside the idea of reform. These solutions embraced the notion of Jihad as the most potent instrument prescribed for all believers in their endeavor to reinstate the true religion of God. Jihad became henceforth the sixth pillar of Islam (the five original pillars were belief in the oneness of God and the prophethood of Muhammad, Zakat or lams-giving, praying, fasting, and undertaking the pilgrimage to Mecca). Furthermore, Jihad was to be launched by a well-organized revolutionary vanguard steeped in the original tenets of Islam. This ideology received its most coherent elaboration at the hands of its Egyptian ideologue, Sayyid Qutb. His program of action was taken up by various Islamic groups with the gradual consolidation of Muslim nation-states and their modernist policies.

It was now thought that all previous attempts at building an Islamic state in the modern era had failed as a result of a number of interrelated factors: lack of a coherent ideology and a corresponding plan of action, coupled with the mistaken belief in the possibility of reforming existing institutions in a gradual and piecemeal manner. Such attempts at reform had failed, it was argued, because Islam could not be considered as a partial or disjointed collection of rituals and social practices. Islam is to be restored as a complete system of thought and practices, a weltanschauung, not unlike, fascism, Marxism, or liberalism. Each of these ideologies possesses its own repertoire of concepts, plans of action, constituencies, and political organization. Islam should in this sense differentiate itself clearly and emphatically in all its labels, slogans, tactics, and strategies. All societies had reverted to a new type of ignorance, irrespective of their political or cultural designations. Consequently, two mutually opposed camps should face each other across well-demarcated fields: those who uphold belief in the only authentic message of God, and all others, irrespective of their varied designations – democratic, communist, socialist, nationalist etc. – particularly those who deceitfully pretend, in spite of all evidence to the contrary, to be applying the precepts of Islam. In other words, Jihad launched by a revolutionary vanguard becomes a religious obligation that is incumbent on all those who look forward to the foundation of an Islamic state. It is, moreover, a worldwide movement that recognizes no borders, national territories, or international conventions. However, it was initially supposed that this religious duty, whose highest form was armed struggle, or indiscriminate violence, was to be carried out against the immediate enemy, or the national state. This was the case in the 1970s, 1980s and early 1990s, with Egypt, Syria, and Algeria figuring as the most conspicuous cases. As hopes of a speedy victory began to fade owing to repressive measures and counterterrorist policies pursued by expanded intelligence and security services of these national states, a new approach began to be formulated. This was especially the case of al-Qa'ida organization that was originally set up in the 1980s to combat the Soviet military presence in the country in close cooperation with the United States, Saudi Arabia, and Pakistan. With the Soviet withdrawal from Afghanistan in 1989 and the subsequent downfall of its communist-led government in 1992, al-Qa'ida, under the leadership Osama Bin Laden, a Saudi millionaire, tried to diversify its activities by launching suicide and other operations in Muslim countries, with no long-lasting results. Shunned by Saudi Arabia and the United States in the wake of their victory in Afghanistan, and faced with efficient repressive regimes across the Muslim world, Osama Bin Laden and his Arab recruits and disciples, along with other Jihadist groups, took the crucial decision of transforming their struggle into an internationalist encounter pitting the West and other secular states against an Islamist organization that perceived Jihad as a relentless series of violent acts designed to sow terror, intimidate the enemy, and lead to the final collapse of an unjust and ungodly world system. It was in this context that the attack on the twin towers of the World Trade Center and the Pentagon took place on 9 September 2001. Between 2001 and 2010 the West coined its own term, war on terror, to describe or justify its military, political, and surveillance operations across the world. Beginning with the invasion of Afghanistan in 2001, and Iraq in 2003, and ending with illegal practices ranging from the irregular rendition of suspected persons, or the extrajudicial transfer of a suspect from a Western state to non-Western one whereby torture could be inflicted with impunity, and setting up in the United States a surveillance system of phone tapping, email, and private correspondence interception. In this sense, Jihadist indiscriminate violent operations were often overshadowed by indiscriminate response and illegal violations of human rights, not to mention the national sovereignty of independent states.

The eruption of the Arab Spring at the end of 2010 marked a new watershed in the fortunes and future prospects of Islamic fundamentalism. This was most clearly pronounced in the bifurcation of the movement into two branches, with one espousing the pursuit of a peaceful pathway to political power, and the other accentuating its rhetoric of continuing its armed struggle until the final Day of Judgment. Moreover, the internationalist phase seems to have been unceremoniously discarded without official burial notices. This act received its recent confirmation in the eruption of a peaceful uprising in Syria, in line with other Arab Spring countries, only to be transformed within 2 years into a violent civil war, with Jihadists pouring into Syria from all over the World. Such a state of affairs indicates the emergence of a new Islamic fundamentalist paradigm which combines national and extra-territorial elements in an effort to stem the tide of secular change deemed harmful to the intrinsic nature of a specific and narrow version of Islam. It is, however, doubtful whether such a strategy will yield permanent results in the face of increasing

numbers of Muslim individuals, organizations and states embracing and applying democracy as a preferred system of government and means of resolving social and political differences and diversities. This is the case in Indonesia, Turkey, and Tunisia, to mention only a few. The example of the Muslim Brothers in Egypt is another case in point, whereby they lost power within a year of winning parliamentary and presidential elections in 2013. This indicates that citizens become almost immediately disillusioned with a movement that uses a religiously saturated call to action without a corresponding ability to deliver some of its widely trumpeted promises, albeit in an ambiguous language that most strategists find difficult to translate into concrete achievements. Lacking a clear social and economic program, these movements seem to excel under certain conditions in gaining the support of some sections of their populations, but once in power this support rapidly evaporates, leading to their ignominious downfall, as was the case in Egypt. Furthermore, Western governments are revising their traditional policies, adopting in the process more nuanced stances that may in the long-term eventuate in more balanced strategies and political decisions ranging from the feasibility of military confrontations to targeted economic aid.

Given these new political and social developments, Islamic fundamentalism no longer seems as the wave of the future. Nevertheless, such a diagnosis does not preclude the persistence of certain movements and individuals to swim against the tide or commit suicide in the name of a distorted version of their own religion.

See also: Islam: Middle East; Middle East, North Africa: Sociocultural Overview; Near Middle East/North African Culture Private; Near Middle East/North African Studies: Politics; Near Middle East/North African Studies: Religion; Religion, Sociology of; Religious Fundamentalism: Cultural Concerns.

Bibliography

Al-Azmeh, A., 1996. Islams and Modernities, second ed. Verso, London.
Appleby, R.S., Marty, M.E. (Eds.), 1991. The Fundamentalism Project, 5 vols. Chicago University Press, Chicago.
Ayubi, N., 1991. Political Islam. Routledge, London and New York.
Choueiri, Y., 2010. Islamic Fundamentalism. Continuum, London and New York.
Dabashi, H., 1993. Theology of Discontent. The Ideological Foundations of the Islamic Revolution in Iran. New York University Press, New York.
Enayat, H., 1982. Modern Islamic Political Thought. Macmillan, London.
Gerges, Fawaz A. (Ed.), 2014. The New Middle East. Protest and Revolution in the Arab World. Cambridge University Press.
Kerr, M., 1966. Islamic Reform. University of California Press, Berkeley and Los Angeles.
Roy, O., 1994. The Failure of Political Islam. Harvard University Press, Cambridge, MA.

Islamic Law

Lena Salaymeh, University of California (UC) Berkeley School of Law, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Islamic law' refers to the diverse legal systems that have been and continue to be produced with the objective of being in line with the Islamic faith. This article provides an overview of Islamic legal history and a basic explanation of orthodox Islamic jurisprudential methodology that demonstrates the diversity, flexibility, and breadth of Islamic legal systems.

Introduction

'Islamic law' refers to the diverse legal systems that have been and continue to be produced with the objective of being in accord with the Islamic faith. Islamic legal systems operate in multiple and sometimes discontinuous ways. Usage of the singular term 'Islamic law' should not be understood as suggesting the absence of legal polycentricity (multiple groups and institutions generate Islamic law) or legal pluralism (within Islamic societies, since Islamic and non-Islamic legal systems coexist). There are significant historical and substantive distinctions between 'Islamic law' and 'Muslim legalities' (the legal systems in use by Muslims). 'Islamic law' refers to juristic interpretations (*fiqh*) of divine law (*sharīʿah*); 'Muslim legalities' refers to either state law (where Muslims are the majority or minority) or the legal practices of non-state Muslim communities. The key distinction between these two overlapping categories is that Islamic jurisprudence is generated by an interpretive process anchored in canonical Islamic texts; in comparison, Muslim legalities are generated by an interpretive process anchored in a state or other legal system that may or may not be Islamic and with a population that may or may not have a Muslim majority (Salaymeh, 2014). The historical section on Islamic law in this article focuses on Arabia and other parts of the Near East, but Islamic law spread internationally by the beginning of the modern era. This article provides a basic overview of Islamic legal history and key aspects of Islamic jurisprudence.

Common Terms and Concepts	
Qurʾān	Islam's holy book, revealed to Muḥammad (d. 632)
sunnah	Normative precedents of the earliest Muslim community
ḥadīth	Reports of the sayings and acts of Muḥammad and some of his early companions; compiled in canonical collections
sharīʿah	Divine law
fiqh	Human understanding or interpretation of divine law
faqīh	Jurist, one who interprets divine law
uṣūl al-fiqh	Principles for interpreting divine law

(*Continued*)

—cont'd

Common Terms and Concepts	
ijmāʿ	Legal consensus of Companions or of Muslim community
qiyās	Analogical reasoning
ijtihād	Independent legal judgment
taqlīd	Following established legal precedents
fatwá (pl. fatāwá)	A nonbinding legal opinion issued by a jurisconsult (a specific kind of jurist qualified to give legal advice), sometimes in response to a question
muftī	A jurisconsult, meaning a jurist qualified to formulate Islamic legal opinions
qāḍī	Judge
madhhab (pl. madhāhib)	Around the eighth century, referred to a group of students, jurists, and judges who adopted the doctrine of a particular leading jurist; later referred to the following of a particular legal school

Islamic Legal History

The periodization presented here identifies approximate moments in Islamic legal history and is not intended to convey developmental or evolutionary progress.

Legal Circles and Networks (*c.* CE 610–750): The Beginnings of Islamic Law

Islamic legal history begins with Prophet Muḥammad's judgments over the first Muslim community in both Mecca (*c.*610–622) and Medina (*c.*622–632). Some of these laws were recorded in the Qurʾān and others were orally transmitted as ḥadīth. To derive legal rulings, the Prophet's Companions and the first generation of Muslims after them (known as the Successors) interpreted the Qurʾān and sunnah and expressed personal opinions (*rāʾy*) or interpretations of Islamic law; in generating Islamic law, they both innovated new laws and followed legal precedents (of the Prophet, Arabian customs, and broader Near Eastern legal practices). As was the case for the general transmission of knowledge in the region, legal education occurred primarily in study circles

around particular scholars (Companions or Successors), who sometimes served as judges in local courts. Jurists with a reputation for piety or learning attracted larger circles of students.

With the Muslim conquests and the establishment of the Umayyad Empire (CE 661–750), Muslims created a burgeoning empire and integrated the region's diverse religious and ethnic communities – as well as some of their legal practices. While legal education initially took place in legal circles (local schools), they evolved into legal networks (geographic or personal schools, following a reputed jurist), and later consolidated into legal schools (*madhāhib*, groups of legal professionals following a particular jurisprudential methodology). Legal circles were informal groups of students who sat around a teacher (in a circle) to memorize and discuss wide-ranging Islamic legal and historical traditions; these local legal circles often met in mosques or private homes. Legal networks were larger, more formal groups of students who studied with a teacher to memorize and discuss a limited set of legal and historical traditions; meeting in mosques, homes, and informal academies, these networks comprised both students who were local to the teacher's community and those who traveled from afar in order to study with the teacher. Many of these teachers were also jurists who served the legal needs of their communities. Traditional Islamic sources identify seven jurists in Medina as having been influential: Abū Bakr ibn ʿAbd al-Raḥmān al-Makhzūmī (d. 712/3), Khārija ibn Zayd ibn Thābit (d. 717–9), ʿUrwa ibn al-Zubayr (d. *c.*715), Sulaymān ibn Yasār al-Ḥilālī (d. 718/9), ʿUbayd Allāh (d. 716/7), Saʿīd ibn al-Musayyab (d. 712/3), and al-Qāsim ibn Muḥammad (d. 728). Jurists outside of Medina also gained fame; for example, famous jurists in Mecca included ʿAṭāʾ ibn Abī Rabāḥ (d. 723) and Mujāhid ibn Jabr (d. *c.*720) and in Kufa included Ibrāhīm al-Nakhaʿī (d. 714) and Ḥammād ibn Abī Sulaymān (d. 737) (Hallaq, 2005: p. 65).

There was a spectrum of Muslim juristic approaches, with one end comprising rationalists (*ahl al-rāʾy*, primarily in Iraq) and the other end consisting of traditionists (*ahl al-ḥadīth*, primarily in Ḥijāz). Rationalists anchored jurisprudence in logical thinking, whereas traditionists anchored jurisprudence in sacred and textual sources (Hallaq, 2009: pp. 55–60). Muslim jurists of this period cannot easily be categorized as belonging to one camp or the other because most jurists likely combined approaches, by both relying on textual authorities and engaging in logical reasoning in formulating legal opinions. Shīʿī sectarianism led to the development of another approach, based on the legal authority of Caliph/Imām ʿAlī and his Imāmī successors. The largest Shīʿī legal school is the Imāmī one – sometimes identified as the Jaʿfarī school, after the fifth and sixth Imāms who lived in the late seventh through mid-eighth centuries; it was initially strongest in Medina and Kufa.

Professionalization of Legal Schools (*c.* CE 750–950)

After the ʿAbbāsids (r. 750–1258) established Baghdad as the political-intellectual capital (in CE 762) of the Islamic empire, jurists shifted from associating with legal circles or networks to identifying with professional legal schools (*madhāhib*). Parallel to the ʿAbbāsid Empire's development of an increasingly more complex bureaucratic system, Muslim jurists became professionalized – as evidenced in their literature, education, and organizational structures. Islamic legal literature initially integrated historical narratives and recorded legal judgments, but Muslim scholars gradually differentiated genres (historiography, Qurʾānic interpretation, ḥadīth studies, and law) as the juristic profession became more specialized. Muslim jurists debated the role of oral law – the orally transmitted traditions about the legal practices of the earliest generations of Muslims – in Islamic jurisprudence; challenges to the authority of oral reports led to the increasing importance of recorded (written) ḥadīth. Consequently, a broad synthesis of the rationalist and traditionist approaches to jurisprudence took place in this period.

Similarly, Islamic legal education became more professionalized. Islamic law was initially taught in mosques and in private homes; but the demands of students and scholars traveling in search of religious knowledge led to the establishment of hostels and the organization of more dedicated spaces of learning. The college of legal studies (*madrasah*) gradually emerged in Iran and Iraq to provide spaces for instruction and libraries, in addition to housing for professors and students. But these buildings served additional communal functions, as they often included tombs or public fountains. Law students studied with many scholars in many places, not being limited to one legal school; in turn, their travels resulted in the diffusion of legal methodologies to new areas.

Over the course of the first four centuries of Islam, hundreds of legal circles and networks emerged and disappeared. The legal reasoning and doctrines of these now extinct legal circles and networks are difficult to reconstruct from the available historical sources. They were incrementally superseded by legal schools, for which we have more historical evidence – particularly those that survive; this situation of the historical evidence can result in overemphasizing the results of a long historical process, rather than the sociopolitical dynamics that produced the consolidation of legal schools. Among the extinct legal schools, most scholarly attention has been focused on the schools of three jurists: al-Awzāʿī (d. 773; Syria), al-Ẓāhirī (d. 883; Iraq), and al-Ṭabarī (d. 922; Iraq).

The ʿAbbāsid state's demand for professional jurists and sectarian resistance encouraged the consolidation of Islamic legal schools from hundreds to several. The eponyms of the surviving legal schools flourished in this era: Abū Ḥanīfah (d. 767; Kufa, Iraq), Mālik ibn Ānas (d. 795; Medina), al-Shāfiʿī (d. 820; Egypt), Aḥmad ibn Ḥanbal (d. 855; Baghdad, Iraq). Each school is known for a particular characteristic: Ḥanafīs for their reliance on personal opinion, analogical reasoning (*qiyās*), and juristic preference (*istiḥsān*); Mālikīs for their prioritization of Medinan practice (*ʿamal*) and consensus (*ijmāʿ*); Shāfiʿīs for their synthesis of rationalist and traditionist methodologies; Ḥanbalīs for their avoidance of reasoning in favor of textual sources. Still, it would be a mistake to reduce the complexity of these legal schools to these common motifs or assume that these jurists and their students represent all Islamic legal activity of this period. It was not until the beginning of the tenth century, at the earliest, that these legal schools became clearly identified, especially in terms of a particular jurisprudential methodology (Hallaq, 2009: pp. 60–71). In addition, legal schools sometimes played sociopolitical roles.

For instance, throughout the ninth and tenth centuries, Ḥanbalī preachers occasionally raised popular demonstrations for or against Caliphal policy, but the school is known for generally avoiding political controversies (Lapidus and Salaymeh, 2012b).

Coinciding with the rise of formal institutions of learning and the consolidation of the schools of law, there was an increase in the importance of the jurisconsult (*muftī*) – a jurist qualified to formulate Islamic legal opinions. A legal decree consisted of both the question sent in writing to the jurisconsult and his answer. In addition, judges often consulted with jurisconsults for their expertise in complicated cases, as evidenced by the inclusion of a jurisconsult's legal decree within the court register. The earliest surviving collections of a jurisconsult's opinions (plural, *fatāwá*) date to the tenth century. The general public consulted a jurisconsult in ways similar to how we now consult attorneys for legal advice; some individuals 'shopped' for a legal decree (singular, *fatwá*), asking several jurisconsults until they obtained a desired outcome.

A judge (*qāḍī*) was responsible for the administration of the courts and selected both his own court recorders and witness certifiers (responsible for evaluating the reliability of witnesses). Judges were also involved in managing charitable endowments and, sometimes, managing the assets of orphans. While court recorders wrote down case proceedings, these documents were not effectively preserved and Islamic court records are largely lost for periods before the sixteenth century. The ʿAbbāsids appointed the first chief judge (*qāḍī al-quḍāh*) – Abū Yūsuf (d. 798), a Ḥanafī scholar – in the late eighth century.

In addition to the formal court system, there were other judicial or quasi-judicial institutions. As early as the ʿAbbāsid Caliph al-Mahdī (r. 775–785), if not earlier, Muslim rulers heard complaints or petitions in courts of equity (*maẓālim*). As under pre-Islamic rule, market-inspectors acted on behalf of the state to ensure fair dealings in the marketplace, which evolved into enforcing morally proper behavior. Reflecting the professionalization of legal knowledge, jurists produced a genre of legal literature (*ḥisbah*) that delineated the duties and limitations of a market inspector.

Consolidation and Formalization (*c.* CE 950–1250)

Professionalization of jurists and institutionalization of learning contributed to the growing dominance of a religious orthodoxy. Diversity of opinions on a variety of legal and theological matters gradually became narrower and, after the tenth century, a religious canon becomes discernible. By the eleventh century, the majority of Muslim jurists appear to have identified with one of the still extant Sunnī legal schools (Ḥanafī, Shāfiʿī, Ḥanbalī, or Mālikī). (Other extant schools include Imāmī Shīʿīs and Ismāʿīlīs, as well as Ẓāhirīs (literalists), Zaydīs, and Khārijīs.) Legal schools succeeded or failed based on popular support, the sociopolitical influence of the school's jurists, and the compatibility between local, customary practices, and the school's legal opinions. A legal orthodoxy emerged through the consolidation of the main legal schools and the gradual marginalization of peripheral or minority legal schools. The four surviving Sunnī legal schools grew in terms of number, power, and influence for a variety of both internal (i.e., legal education and popular appeal) and external (i.e., state support) reasons. In eleventh and twelfth century Nishapur, Ḥanafīs and Shāfiʿīs struggled for local political power over the judiciary, academies, and charitable endowments; these struggles affected the influence of each legal school in the region. Yet, an individual Muslim's identification with a 'legal school' did not necessarily indicate a commitment to the legal school's jurisprudence; common people were loyal to local leaders and not necessarily to the abstract principles of the legal school.

Within this consolidation process, identification with a particular legal school was not necessarily exclusive. Law students studied with scholars outside their own legal school – particularly ḥadīth transmission – and freely switched legal school affiliation. Since caliphs appointed a chief judge who, in turn, oversaw the appointment of judges throughout the empire, jurists increasingly tended to associate with the larger, state-backed legal schools. Some jurists shifted their legal school affiliation in order to procure administrative or judicial appointments. Indeed, as legal education became more systematized and salaries of judges increased, the legal profession most likely attracted students drawn to the pragmatic – rather than spiritual or intellectual – aspects of the vocation of law. In terms of producing the most influential jurisprudential texts, the most important jurists were unaffiliated with the state.

The formalization and professionalization of the legal schools are reflected in changing genres of legal literature. In later periods, the legal opinions of groups of jurists who belonged to the same legal school were compiled, as opposed to the opinion of a single jurist. Muslim jurists wrote commentaries on canonized or codified legal texts of groups of predecessor jurists (who had been identified as authoritative) and created condensed legal manuals (plural, *mukhtaṣarāt*). *Mukhtaṣar* (singular) texts abbreviated earlier legal texts to offer straightforward legal answers based on each school's most authoritative or widely accepted legal opinions; these abridgements became the main texts of study for students and became increasingly prevalent after the eleventh century. By selectively choosing which legal opinions would be considered authoritative, these compendia created an official resource for later jurists to perpetuate and disseminate the rulings and the methodology of their school. Another genre of legal literature that intensified in this period is texts on the etiquette of judging (*adab al-qāḍī*). Based on earlier court practices, these texts outlined the duties and proper behavior of a judge and general court procedures, thereby further formalizing the legal profession.

Beginning in the tenth century, legal education became increasingly formalized as professional law colleges (*madāris*; singular, *madrasah*) were endowed with permanent sources of income, such as land or rent-bearing urban property, set aside in perpetuity. These endowed trusts (*awqāf*) funded faculty salaries, student stipends, religious clerical staff, administrative staff, and maintenance staff (janitors, gate keepers, etc.). Donors created these endowed trusts in order to prevent their property from being divided in accordance with inheritance laws or from being confiscated by the government. These trusts also allowed donors to identify individuals (especially daughters) who could receive a regular income from the endowment. Endowing colleges gave the donor prestige.

Niẓām al-Mulk (d. 1092), a Seljuq minister, made it the state's policy to support Islamic educational institutions and endowed both Ḥanafī and Shāfiʿī colleges in all of the empire's major cities. While the colleges were initially dedicated to only one legal school, the Mustanṣiriyya college (founded in Baghdad in 1233) was devoted to several legal schools (Lapidus and Salaymeh, 2012b).

Jurist-professors designed the curriculum of these colleges, which varied according to local practice and the legal school affiliation. In previous centuries, jurists typically had earned their living through trade or craftwork, but with the institutionalization of education in the form of endowed colleges, religious scholars became professionals, administrators, and salaried employees. Through state patronage, salaries, and the administration of endowments, scholars became a professional class. Islamic legal education was basically informal, flexible, and built on individual teacher and student relationships, rather than institutional ties; students did not earn grades or receive degrees. Learning revolved around the reputation of a respected scholar, who in his individual capacity could give students a specific license to teach, to transmit ḥadīth, to judge, or to participate in scholarly debates – depending on the course of study. In addition to law, many colleges taught Qurʾān, ḥadīth, Arabic grammar, theology, Sufism, medicine, and philosophy. These colleges consolidated informal legal instruction into an official institution with a physical center and permanently endowed funds, thereby facilitating professional full-time study of Islamic law and the expanded training of cadres of Islamic legal teachers and administrators (Lapidus and Salaymeh, 2012b).

Still, a typical student 'heard' (i.e., listened to and memorized oral lectures) his subjects from a variety of teachers on a variety of topics in a variety of places and study circles. Education occurred beyond colleges and the function of colleges extended beyond education (into religious propaganda and political action). In addition to endowed colleges, the transmission of religious knowledge took place in informal study circles (*ḥalaqāt*; s. *ḥalaqah*), in public mosques, and in private homes; students could come to a college without being a resident or receiving a stipend. Women not only studied with relatives, but also received certificates from outside scholars, including women scholars; they most commonly transmitted ḥadīth. College professors, Qurʾān and ḥadīth reciters, and storytellers all had audiences with the general public. Both formal colleges and informal learning circles encouraged the consolidation of the surviving legal schools by training students to further propagate the law school's methodology (Lapidus and Salaymeh, 2012b).

Technocratization (*c.* CE 1250–1450)

By the thirteenth century, Muslim jurists were increasingly becoming technocrats, specialized in knowing the doctrinal rulings of a particular school – rather than in applying the Islamic jurisprudential methodology of a particular school. In this era, Muslim jurists composed more abridged versions of canonical legal commentaries and the locus of jurisprudential activity was shifting away from the Iraqi sphere of influence (both geographically and intellectually). This was likely a result of the Mongol destruction of Baghdad (capital of the ʿAbbāsid Caliphate) and its libraries in 1258. The city's role as an intellectual center came to an end as regional centers of intellectual production arose to replace it. In provincial centers of legal activity, legal manuals and their commentaries were produced; in particular, one type of abridged Islamic legal text – sometimes identified as a 'common law *mukhtaṣar*' (Fadel, 1996) – became prevalent.

The Mamluk Sultanate (r. 1250–1517) of Egypt appointed a chief judge for each of the four main Sunnī legal schools, solidifying the position of the largest schools and confirming the demise of the once numerous smaller schools. A potential consequence of these juristic appointments was that the Mamluk Sultan used these four legal options to his advantage, by selectively seeking legal decrees from the legal school most likely to meet his objectives. The daily operations of Mamluk courts are preserved in the fourteenth century court documents archived in Jerusalem's Islamic museum. While generally jurisconsults were not affiliated with or regulated by the state, the Mamluks did appoint some jurisconsults to courts of equity. In the eleventh century, jurists had defined a jurisconsult as being a *mujtahid* (i.e., one who has the ability to independently reason; the highest rank of a jurist). By the middle of the thirteenth century, however, it appears that the prerequisites were lowered and a jurisconsult was expected by most scholars to be a *muqallid* (i.e., able to articulate a legal opinion based on the precedents and methodology of a particular legal school; a lower rank than *mujtahid*). This reflects a far reaching phenomenon in this era: many Muslim jurists preferred to follow the legal opinions of their predecessors in orthodox legal schools (*taqlīd*), rather than develop legal opinions based on the jurisprudential methodology of an orthodox legal school (*ijtihād fī al-madhhab*).

Bureaucratization (*c.* CE 1450–1800)

The Ottoman Empire (r. 1299–1923) bureaucratized Islamic legal practices of preceding periods by making law an essential component of the state's administrative apparatus. After the fifteenth century, Ottoman courts became the primary venue for litigation throughout the Ottoman Empire. Ottoman courts applied Islamic legal precedents (specifically, *fiqh*) in a judicial system characterized by its systematization. Following the precedent of the Mamluks in Egypt, the Ottomans officially appointed a jurisconsult (*muftī*) to provide formal legal opinions, thereby integrating these formerly independent legal thinkers into the state bureaucracy. Although the training of legal professionals had previously been controlled primarily by jurists with minimal state intervention, the Ottoman state became intimately involved in every aspect of judicial activity – from training jurists and court personnel, to selecting judges and applying legal codes. The Ottomans established schools in the cities/towns they conquered and made the Ḥanafī legal school the state's official legal school. As a result, law students began to migrate away from other legal schools to seek training in the Ḥanafī school as a step toward a career in the government (Lapidus and Salaymeh, 2012a).

The Ottomans systematized law throughout the empire by promulgating legal codes, such as the *qānūnnāme* (code of regulations) of 1519 and the criminal code of 1540. Having been composed in both Ottoman Turkish and Arabic, the *qānūnnāme*'s language indicates that Arabs under Ottoman rule

were expected to conform to Ottoman laws, which were based on both medieval Islamic jurisprudence and customary laws (Ergene, 2001). The Ottoman criminal code changed a variety of criminal penalties into fixed fines (proportional to the offender's economic means) and permitted some circumstantial evidence in determining adultery, but otherwise largely continued earlier Islamic legal practices (Imber, 1996). Legal codes were enforced by Ottoman judges in official, newly established Ottoman courthouses, which were more numerous when compared to the pre-Ottoman period. Ottoman court fees for litigating cases and for official documents (deeds, marriage certificates, etc.) financed the court's operational costs and even contributed to the Ottoman treasury (Ergene, 2002).

Despite the proliferation of Ottoman courts and laws throughout the empire, there was some local resistance to Ottoman legal hegemony. Some Ottoman subjects objected to some Ottoman laws on the grounds that they were not based on 'Islamic' legal principles. By way of example, in the eighteenth century, a local, provincial judge objected to two types of taxes (*avarïz* and *nüzül*) because he believed they were impermissible according to *sharīʿah*. (Ergene, 2001: p. 83) This tension was part of an ongoing debate about the relationship between Ottoman laws and the legal precedents established by the medieval Islamic legal schools. Ottoman jurists invoked Islamic jurisprudential terminology – such as *ḍarūra* (necessity) and *maṣlaḥah* (public interest) – to legitimate setting aside certain precedents (recorded in the jurisprudential texts of the orthodox legal schools). Similarly, local customs could prevail over the legal opinions recorded in jurisprudential texts. Thus, the Ottoman acceptance of usury should be understood as an Ottoman interpretation of Islamic scripture that differed from the medieval orthodox position. What makes the Ottoman period distinctive in Islamic legal history is the state's participation in legal activity, not the legal outcomes of a normative jurisprudential process.

The pinnacle of the Ottoman state's control over legal matters was the Ottoman *şeyhü'lislām (shaykh al-Islām)*, who acted as a leader of Ottoman religious scholars and chief jurist of the Ottoman state. He issued legal decrees and intervened in political matters by issuing legal opinions related to the state's foreign policies. Molla Hüsrev (d. 1480) was a fifteenth century *şeyhü'lislām* who wrote about Islamic law in Arabic. Likewise, Ibrahim b. Muhammed al-Halebi (d. 1549) wrote in Arabic (since he was from Egypt) and his work was translated into Turkish in the seventeenth century. Ebu's-su'ud (d. 1574) was one of the most well-known Ottoman jurists to hold the position of *şeyhü'lislām*. Ebu's-su'ud gave the Ottoman sultan the title of Caliph, justified interest-loans, and modified tax practices (Imber, 1997). Collections of legal decrees – such as *Behçetü'l-fetāvā* by Yenişehirli 'Abdullāh, who was *şeyhü'lislām* from 1718 to 1730 – provide scholars with important information about daily life and legal practices of the Ottomans.

Law in any society is not merely the product of the state or legal elites, but also is informed by local groups and interests. Guild law, for instance, was independently developed and Ottoman courts honored their customs in adjudicating disputes between guild members. Similarly, Ottoman judges (like judges in any legal system) sometimes deferred to customary law, thereby legitimating local customs as a source of law. By the way of example, the Ottoman *qānūnnāme* introduced a local custom by fining a man who was aware of his wife's infidelity, but did not act upon that knowledge (Imber, 1996: p. 185). In response to local customs, some Ottoman legislation was promulgated to deal with customary realities that do not appear to have been dealt with by medieval Muslim jurists. Records of complaints against Ottoman officials – such as the *Şikayet Deferi* of 1675 – indicate that lower class Ottoman citizens had recourse to the courts against government elites. This suggests that the Ottoman courts were not merely tools used by the elite to dominate society (Lapidus and Salaymeh, 2012a).

Some powerful interest groups were able to acquire exceptions to Ottoman laws that resulted in economic advantages. In the seventeenth and eighteenth centuries, the Ottoman Empire granted legal rights (*ahdnames*) to foreign merchants who were enabled to conduct commercial dealings according to foreign, rather than Ottoman, laws. The agreements dictated everything from customs rates, to taxation, to inheritance laws that would apply to both foreign merchants and non-Muslim Ottomans associated with them (effectively changing the nationality of the latter). These Ottoman capitulations eventually became a pretext for European intervention into internal Ottoman affairs (Lapidus and Salaymeh, 2012a).

A common Western stereotype about the nature of the Ottoman state as despotic has been refuted by numerous studies demonstrating that power was decentered through the Ottoman bureaucracy and that there were legal limits to the Ottoman sultan's power. A combination of tradition and bureaucracy prevented a sultan from indiscriminately using his power. Sultan Süleyman (d. 1520) was known as the lawgiver and acquired a reputation for justice by punishing Ottoman administrators who had abused their power. Süleyman promulgated legislation (known as *kanun*) and compiled the judgments of predecessor Ottoman sultans in his legal code, the *kanun-i Osmani*, which was the law of the Ottoman Empire for several centuries. Other sultans passed regulations that introduced or limited a variety of legal practices. For example, after 1544, Ottoman judges were prohibited – by sultanic order – from conducting marriages without the permission of a woman's guardian (a legal option permitted by Ḥanafī *fiqh*) (Lapidus and Salaymeh, 2012a).

Another common stereotype about Ottoman judges is embodied in the term coined by Weber, *kadi justice*, which refers to a judge who has unlimited personal discretion. Ottoman court records, however, illustrate that Ottoman judges actually followed well-known procedural and substantive rules. Indeed, later Ottoman judges were bureaucrats and their method of adjudication was as systematized as other Ottoman administrative acts. While multiple, sometimes conflicting, legal entities operated in the fourteenth and fifteenth centuries, these were subsequently centralized by the Ottoman state (Lapidus and Salaymeh, 2012a).

Islamic law also existed outside the Ottoman Empire and other empires played similar roles in codifying Islamic law. Awrangzīb, Mughal Emperor of India, ordered the compilation of an Islamic law compendium based on the most authoritative legal opinions of the Ḥanafī school; the result is known as *al-Fatāwá al-ʿĀlamgīrīyah* (compiled between CE 1664 and

1672) and the text was an important reference work throughout South Asia and beyond for generations.

Colonialism and Reformation, Liberalization and Authoritarianism (*c.* CE 1800–2010)

Whether through direct occupation or indirect political and economic pressure, Western colonial interventions in the Muslim world provoked a variety of legal reforms and legal innovations. During and after colonization, many Muslim-majority states merged civil or common-law traditions from the West with Islamic legal traditions. Some of these changes were prompted by the needs of modern nation-states and modern administrative practices, while other changes were triggered by socioeconomic and political pressures. These processes generated the distinctions between 'Islamic law' and 'Muslim legalities.' The 'secularization' of law that began with Western colonialism led to the emergence of these overlapping spheres of legal activity. Muslim jurists continued to teach and implement Islamic legal traditions, while secular legal professionals practiced law in states that happened to have Muslim-majority populations. For more detailed information on how these modern distinctions between 'Islamic law' and 'law in Muslim majority states' (or 'Muslim legalities') play out in some parts of the Muslim world, please *see* Middle Eastern and North African Legal Traditions.

Western commercial and penal codes were gradually adopted throughout the Muslim world in the nineteenth and twentieth centuries. South Asia, for example, was incorporated into the British common law system. The Ottoman Empire pursued a legal reform movement (*Tanẓīmāt*) that led, in 1847, to mixed courts with European and Ottoman judges and, in 1850, to a commercial code in line with the global market economy. In addition, a series of 'land reform' laws were passed during the mid-nineteenth century – such as an Ottoman law (1858) and a series of French laws in Algeria (1856–73) – that resulted in the privatization of land. This, in turn, facilitated colonial purchases of land throughout the Middle East and North Africa; in some areas, colonialists owned between 20 and 25% of land. One of the more widely known works of Islamic legal reformation is the *Mecelle* (1869–76), which was issued by the Ottoman government as an attempt to codify Ḥanafī laws in the areas of transactions and some court procedures. Ottoman capitulations permitted Europeans to operate their own courts in the Ottoman Empire, thereby bypassing Ottoman legal sovereignty; this resulted in significant economic advantages for Europeans living in Ottoman territories.

Islamic constitutionalism appears to have begun in the late nineteenth century, with the drafting of constitutions in Tunisia (1861) and in the Ottoman Empire (1876), but intensified with the Iranian constitutional revolution (1906–11) (Arjomand 116–7). While Islamic law was initially understood as a limit on governance, it eventually became identified in various constitutions as a source of law. By 2000, approximately 24 Muslim-majority states identified Islamic law as a (or 'the') source of law (Arjomand 123). Some countries abolished Islamic courts, while other countries established and maintained two parallel court systems (one civil and the other Islamic).

Islamic reform (*iṣlāḥ*) movements, with a special focus on legal issues, emerged in response to colonialism and broad regional legal reforms. There were two common responses to this situation: first, a rejection of the orthodoxy of legal schools and a return to the primary textual sources of Islamic law (advocated by different streams of Salafīs); second, an attempt to combine the legal opinions of both the orthodox and extinct legal schools, relying on the objectives and legal necessity of Islamic legal theory (advocated by the Muslim Brotherhood). Muḥammad ibn ʿAbd al-Wahhāb (d. 1787) preached that Muslims should establish laws based on the Qurʾān and sunnah, instead of relying on the orthodox legal schools; his ideas remain influential today. Many Muslim intellectuals and activists who were involved in the socio-political changes of the nineteenth and twentieth centuries were legal thinkers who advocated certain positions vis-à-vis law. Rifāʿah al-Ṭaḥṭāwī (d. 1873) studied in France and returned to his home country of Egypt where he advocated for modifying Islamic legal traditions based on social need. Among the more influential Sunnīs were the Egyptian grand muftī Muḥammad ʿAbduh (d. 1905) and his student, the Syrian Muḥammad Rashīd Riḍā (d. 1935). Both scholars proposed Islam – and adherence to Islamic law in particular – as essential to colonial resistance. ʿAbduh (d. 1905) issued numerous legal decrees and advocated for choosing (*takhayyur*) the most suitable legal opinion from across the orthodox Sunnī legal schools or combining (*talfīq*) legal opinions from different schools. Riḍā (d. 1935) articulated Islamic reform as a return to the ways of the earliest generations of Muslims, the *salaf* (from which is derived the term *salafī*). Hassan al-Bannā (d. 1949) was the founder of the Muslim Brotherhood, a Muslim social and political party in Egypt. A number of academies and institutions were established or strengthened with the intention of 'revitalizing' Islamic law throughout the Muslim world during the twentieth century. Yet legal education was increasingly secularized and legal actors with training in Islamic law primarily engaged the private sphere or served the public in family law.

Recent debates about the authority of Islamic legal traditions are one aspect of an Islamic revival movement that began in the 1970s as a reaction against the legacy of European colonialism in the region. The reliance on Western laws in the relatively new states of the region is perceived by some as lingering foreign intervention. This is why some states (including Afghanistan, Iran, Pakistan, Sudan, and Saudi Arabia) reinstated Islamic criminal punishments and why many Muslim scholars advocate Islamizing modern nation-states. Sayyid Quṭb (d. 1966) was a prominent Muslim Brotherhood intellectual who advocated for Islamic governance and was executed by the Egyptian government. Pakistan's highly influential thinker, Abū al-ʿAlá Mawdūdī (d. 1979) argued for the necessity of an Islamic state. Iraqi Shīʿī jurist Muḥammad Bāqir al-Ṣadr (d. 1980) expounded on the role of Muslim jurists in Islamic states, which later influenced the Iranian constitution and popular ideas about Shīʿī governance. Al-Ṣadr was the author of an important work on Islamic economics (*Iqtiṣādunā*), whose proposals for an Islamic alternative to capitalism and communism extended throughout the Arab world. His political writings influenced the Iranian revolution and he was active in forming an Islamist party in Iraq. The 1978–79 Iranian revolution provoked renewed attention to the

role of Muslim jurists in modern states. Ayatollah Khumayni (d. 1989) elaborated, advocated, and implemented a theory (*wilāyat al-faqīh*) that gave Muslim jurists authority over political and governmental life. In contrast to this official role for jurists, many other Muslim-majority states include a reference to Islamic law as a source of law in their constitutions, without thoroughly defining who will implement Islamic law.

The primary focus of Islamic legal change in the twentieth century had been family law; this arena is the principal focus of Western/colonial attention to and criticism of Islamic law. Turkey issued a family law code in 1917 that was based on the four orthodox Sunnī schools of law, as well as some opinions from extinct or minority legal schools. Other states followed the Turkish model, with Egypt issuing its personal status laws in 1929, Syria in 1953, and other countries soon thereafter. These family law codes not only selected legal rulings from across several legal schools, they also introduced reforms – particularly in limiting polygyny or the divorce prerogatives of men. One of the more well-known legal reformers is ʿAbd al-Razzāq al-Sanhūrī (d. 1971), who led the drafting of the Egyptian Civil Code (1948/9) and the Iraqi code (1951) and participated in the drafting of codes for neighboring states (including Iraq, Syria, Jordan, Kuwait, and Libya). Sanhūrī earned two doctorates from the University of Lyon, was the dean of both an Egyptian and an Iraqi law school, and held several government positions. Sanhūrī sought to modernize Islamic law and develop a modern legal system that would benefit from comparative analysis, but emphasize Islamic values.

In the modern era, the traditional legal schools became less influential than they once were. The Ḥanafī school is prevalent in Turkey, the Levant, and parts of South Asia; Mālikīs are present in North Africa and small parts of the Gulf; the Shāfiʿī school is dispersed in Indonesia, Malaysia, southern Egypt, the Gulf, East Africa, and the Levant; Ḥanbalīs are concentrated in the Gulf. (The Ḥanbalī school spread in the eighteenth century as a result of the Wahhabī movement in the Arabian Peninsula.) Contemporary Islamic legal changes are most evident in the scholarly writings and legal decrees of reputed jurists, as well as the judicial decisions and statutory legislation of some Muslim-majority states. The sociopolitical role of the jurisconsult (i.e., *muftī*) has broadened in the modern era, as a result of the institutional structure (*dār al-iftāʾ*) that supports the issuing of legal decrees, as well as the emergence of print and other media to disseminate legal opinions. More recently, Sunnī jurists have attracted followers through their issuing of legal decrees, including on Internet sites such as Islam Online, www.islamonline.net. Several contemporary Muslim intellectuals who speak publicly about Islamic law are either exiled or living under complicated circumstances in their home countries; examples include the Iranian ʿAbd al-Karīm Surūsh, who lives in exile, and the formerly exiled Tunisian Rāshid al-Ghannūshī, cofounder of al-Nahḍah, the largest political party in post-uprising Tunisia. Still, some Muslim jurists have strong alliances with their governments.

Legal reforms of the nineteenth and twentieth centuries were incremental, as modern nation-states created courts that gradually extended their jurisdiction into areas of the law that had been administered by Muslim jurists. Just as the jurisdiction of Islamic courts narrowed to particular doctrinal areas (particularly family law), so too did popular understandings of 'sharīʿah' become synonymous with certain legal norms; rather than a tradition of professional juristic training and a set of jurisprudential methodologies, Islamic law came to mean a set of laws. Indeed, it is an irony of the contemporary Islamic revival movement that the state's cooptation of the previously independent judiciary is not the focus of resistance.

Growing communities of Muslims who are minorities in non-Muslim-majority states has generated interest in 'minority Islamic jurisprudence' (*fiqh al-aqallīyāt*). The Fiqh Council of North America has issued legal decrees encouraging the assimilation of Muslims in US civil and political society. The European Council for Fatwa and Research is similarly active in publishing legal rulings that meet the diverse and sometimes complex needs of Muslims living in non-Muslim societies. An example of these kinds of rulings is legal decrees permitting Muslims to serve in the armies of Western states.

Resistance (CE 2010–present)

While the colonial period formally ended in the mid-twentieth century, Western interventions and neoimperialism continue to shape the everyday practice of Islamic law. In the midst of uprisings throughout the Arab world, Muslim jurists played roles either in negating the validity of demonstrations (as in the case of state-appointed muftī Dr. ʿAlī Jumʿah's defense of the Egyptian regime in 2011) or in encouraging demonstrations (as in the case of "independent," Qatar-based jurist Yūsuf al-Qaraḍāwī's articulation of the legitimacy of regime change in 2011). A new genre of legal literature devoted to revolutions (*fiqh al-thawrah*) is emerging. Questions about Islamic constitutionalism or the role of Islamic law in public society continue to be debated.

Methodology of Orthodox Islamic Jurisprudence

Orthodox Islamic jurisprudence is *uṣūl al-fiqh*, which translates to "the principles by which Islamic law is derived" (from a set of normative Islamic legal sources). The two main textual sources of Islamic law are the Qurʾān and *sunnah*. Whereas *uṣūl al-fiqh* refers to the roots of Islamic law (or jurisprudence), *furūʿ al-fiqh* refers to its branches (or doctrinal law, both substantive and procedural). Jurisprudential texts discuss hermeneutic guidelines for interpreting and classifying the legal content of Qurʾān and *sunnah*, as well as the identification of consensus (*ijmāʿ*) and methods for analogical reasoning (*qiyās*). Legal reasoning (*ijtihād*) comprises a variety of distinct techniques for deriving legal rulings (Kamali, 1991). Most jurists classified legal acts into five basic categories – obligatory, recommended, permitted, abhorred, and prohibited. A significant distinction between Sunnī and Shīʿī jurisprudence is that the latter tradition considers Imāmī legal opinions as infallible legal sources. The differences of legal opinions between and within legal schools are produced as a result of distinct hermeneutic approaches and of dissimilar reliance on textual and nontextual legal sources.

In Islamic legal literature, *furūʿ al-fiqh* is where one finds legal opinions on a wide array of topics, commonly divided into ritual matters (such as purity, prayer, charity taxation,

fasting, and pilgrimage) and social relations (marriage, divorce, inheritance, sales, loans, partnerships, crimes, torts, and court procedures). Two types of *furūʿ* texts coexisted – the *mukhtaṣar* (abridged legal compendium) and the *mabsūṭ* (comprehensive commentary and explanatory legal compendium).

Conclusion

Beginning with the founding of a Muslim community in seventh-century Arabia, Islamic law has undergone significant and frequent changes. Actual legal practice has been infinitely varied, according to locality, legal school, and many other situational factors. Custom or socioeconomic circumstances often decide whether or not aspects of the law are applied. The application of the law to practical situations and the procedure of consulting scholars for legal opinions generate legal changes. The relationship between Muslim jurists and political structures varied according to time and place. Islamic law is both historically and normatively flexible.

See also: Constitutionalism, Comparative; Islam: Middle East; Law: Change and Evolution; Legal Pluralism; Legal Systems, Classification of; Middle Eastern and North African Legal Traditions; Postcolonial Law; Religion: Nationalism and Identity; Secularization; Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns.

Bibliography

Arjomand, S.A., 2007. Islamic constitutionalism. Annual Review of Law and Social Science 3, 115–140.

Ergene, B.A., 2001. On Ottoman justice: interpretations in conflict (1600–1800). Islamic Law and Society 8, 52–87.

Ergene, B.A., 2002. Costs of court usage in seventeenth- and eighteenth-century Ottoman Anatolia: court fees as recorded in estate inventories. Journal of the Economic and Social History of the Orient 45, 20–39.

Fadel, M., 1996. The social logic of *taqlīd* and the rise of the mukhtaṣar. Islamic Law and Society 3, 193–233.

Hallaq, W.B., 2005. The Origins and Evolution of Islamic Law. Cambridge University Press, Cambridge.

Hallaq, W.B., 2009. Sharīʿa: Theory, Practice, Transformations. Cambridge University Press, Cambridge.

Imber, C., 1996. Studies in Ottoman History and Law. Isis Press, Istanbul.

Imber, C., 1997. Ebu's-su'ud: The Islamic legal tradition. Stanford University Press, Stanford.

Kamali, M.H., 1991. Principles of Islamic Jurisprudence. Islamic Texts Society, Cambridge [England].

Lapidus, I.M., Salaymeh, L., 2012a. Ottoman Law. Islamic Societies to the Nineteenth Century: A Global History. Cambridge University Press, Cambridge.

Lapidus, I.M., Salaymeh, L., 2012b. Scripturalism: Quran, Hadith, and Law. Islamic Societies to the Nineteenth Century: A Global History. Cambridge University Press, Cambridge.

Salaymeh, L., 2014. Commodifying 'Islamic law' in the U.S. legal academy. Journal of Legal Education 63 (4), 640–646.

Islamic Social Work: Hope and the Commitment to Change

Nada Eltaiba, The University of Western Australia, Perth, WA, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

There is a growing interest in religious and spiritual healing among social workers. Religion and spirituality have been the subject of discussions by many scholars and are increasingly considered an essential source of understanding about human beings and their ways of coping with psychological and social problems. This article addresses the concept of hope in relation to religious and spiritual healing in the Islamic texts, the Qur'an and the Hadith. The concepts of religion and spirituality overlap in the Islamic context and are used interchangeably in this article.

Introduction

There is a growing interest in religion and spirituality among social workers. This is reflected in the significant growth in the number of publications and the breadth of research on spirituality and religion in social work (Al-Krenawi and Graham, 2003; Eward R. Canda and Furman, 2010; El-Amin, 2010; Eltaiba, 2007; Furman et al., 2011; Graham and Shier, 2011; Seinfeld, 2012). An understanding of religion and spirituality is increasingly considered an essential source of insight into human beings and their ways of coping with challenges and difficulties. In social work, the interest in religion and spirituality is related to the increased attention to enhancing cultural competence including in the spiritual and religious realms. This is considered an ethical obligation for social work practice (AASW, 2000).

Religion and spirituality have found to enhance optimism, confidence, self-worth, and respect. Studies show that people become more religious when they encounter challenging experiences or are affected by psychological problems (Al-Krenawi and Graham; Edward R. Canda and Furman, 2010; Chen and Koenig, 2006; Graham and Shier, 2011; Kvarfordt and Sheridan, 2007; Seinfeld, 2012). In most cases, religion is significant in their lives at the personal and the social level. People tend to cope better when they turn to religion to make sense of turmoil or to find hopeful insights.

The concepts of religion and spirituality are the subject of great discussion and are considered by many to be controversial. Some researchers, practitioners, and individuals experience difficulties differentiating between the two terms (Edward R. Canda and Furman, 2010; George et al., 2000). According to Canda, "Religions are institutionalised and organised patterns of belief, morals, rituals, and social support systems that have spirituality as their central concern." He further defines spirituality thus: spirituality involves the search for a sense of life purpose, meaning, and morally fulfilling relationships between oneself, other people, the universe, and the ultimate ground of reality (Canda, 1999: p. 12). Both religion and spirituality are personal and share sacred core.

For Muslims, religion and spirituality are considered, in essence to be the same; the concepts do not occur distinctly from each other (Hyman and Handal, 2006). Both religion and spirituality must be considered within a collective context and framework. In the following discussion, the term religion is used to include spirituality and religiosity. The concept of religion is considered to incorporate spiritual aspects, beliefs, values, acts, and symbols that comprise Islam (Eltaiba, 2007).

Religious values and practices for Muslims incorporate various social, political, and spiritual activities. For the majority of Muslims, religious beliefs have a fundamental place in their interpretation of and coping with different problems. For example, research has found that there is a strong inclination among members of Muslim societies to express themselves using religious discourse and toward combining spiritual healing with other approaches when dealing with difficulties. This includes referring to the Islamic texts, the Qur'an and the Hadith (Al-Krenawi and Graham, 1999; Eltaiba, 2007).

Islamic texts are part of the daily life of most Muslim people. The Qur'an for Muslims is considered the ultimate book of the Islamic religion (Horrie and Chippindale, 2003). The Qur'an has organized values for all aspects of human life, including health and healing values. The Qur'an provides Muslims with guidance on what to do and what not to do in order to be close to Allah. Consequently, individuals may be kept from some problems or be enabled to deal with these problems more constructively. There is a belief in Islam and among Muslims that the Qur'an has healing powers (Al-Krenawi and Graham, 1999; Eltaiba, 2007). The power of healing of the Qur'an is stated as follows: "And We send down of the Qur'an, that which is a healing and a mercy to the believer" (Qur'an 17:82).

The Hadith, books of 'prophet Muhammad's traditions and sayings,' is considered another main source that explains how the Qur'an can be applied to Muslim lives in order to lead the best life possible and helps people to understand problems according to the Islamic faith.

Islamic texts are recognized as providing a rich context within which to interpret difficulties. Understanding the options for coping processes that these texts provide is valuable and necessary for social workers practicing in Muslim communities. Some of the central concepts with which social workers need to be familiar are hope and the obligation to change.

Hope in Islam

Defining hope is challenging. It has predominantly been linked to expectations and the achievement of goals (Whitley, 2010). Hope comprises four components: the affective (e.g., trust, confidence, humor, and positive emotions); the cognitive (e.g., reflecting on past experiences, goal setting, planning, and

assessing the likelihood of success); the behavioral (e.g., motivation and personal activity); and the environmental (e.g., availability of resources, health care, and relationships) (Schrank et al., 2008).

The promotion of hope in association with religion is seen as important in the journey of coping with difficulties. Hope, for example, is identified in mental health literature as a significant aspect of the recovery process and it is strongly manifested in the recovery discourse (Anthony, 2003; Bonney and Stickley, 2008; Whitley, 2010). Recovery from mental health problems is linked to hope, which provides a strong foundation to recover. In addition, hope is seen to speed the recovery process and aid in sustaining it.

Likewise, hope is an important theme in the Qur'an and the Hadith. It is presented as a feeling of relief and comfort despite facing difficulties or crises. The believer expects that the situation will ease and things will be better; they will never give up and will keep going. In that sense, hope will keep motivating the person to work, be determined to achieve, and to be in a better place in their life.

In the Islamic context, hope means to have gratitude that there will always be a possibility of overcoming problems and difficulties. Hope helps people to tolerate and overcome life's challenges and catastrophes. Without such hope, things would be much harder for those suffering. In the Qur'an and the Hadith, the difficulties in life are seen as being followed by periods of relief. The Qur'an states

> So, verily, with every difficulty, there is relief: Verily, with every difficulty there is relief.
> Qur'an 94:4, 5

This also was confirmed by the teaching of the prophet as the Hadith states: "Prosperity comes after agony and sorrow; and every ordeal is followed by easy circumstances" (Nawawi and Abbasi, 1980), endurance of suffering is rewarded. Human beings are expected to strive to do their best in order to cope with problems. Facing challenges is part of the journey of human life. These challenges are a trial from Allah and human beings are expected to suffer through the process. The Qur'an says:

> Or do you think that ye shall enter the Garden of (bliss) without such (trials) as came to those who passed away before you? They encountered suffering and adversity and were so shaken in spirit that even the Apostle and those of faith who were with him cried: 'When (will come) the help of God?' Ah! Verily, the help of God is (always) near.
> Qur'an 2:214

As hope is part of being a believer, losing hope is forbidden in the Islamic tradition. Not having hope and not being motivated to work and change is considered to be against the teaching of the Qur'an and the Hadith. The Qur'an says

> And never give up hope of God's soothing Mercy: truly no one despairs of God's soothing Mercy, except those who have no faith.
> Qur'an 12:87

There is an acknowledgment in Islamic tradition that regret and feelings of guilt may lead to a psychological state in which people become affected by some psychological problems. These feelings might create some difficulties for the person, which prevent them from looking into the issues of their life in a positive manner. This will also cause the person to be less hopeful, less accepting, or less motivated. Learning about the forgiveness of Allah, as it is represented in the Qur'an and the Hadith, is considered a way of coping with these feelings. This is especially so in cases where individuals feel guilt associated about committing actions that go against religious commands or if they feel they are not deserving of good. In the Qur'an, believers are ordered not to be demanding on themselves, as there is a space where these misdeeds are forgiven:

> O my servants who have transgressed against their souls! Despair not of the Mercy of Allah' for Allah forgives all sins for He is oft-forgiving, Most Merciful.
> Qur'an 39:53

On the other hand, the feeling of hope is a sign of a strong relationship with Allah and it is based on the belief that Allah will respond to people when they ask for help. This is stated in the Qur'an in the following verse:

> And never give up hope of God's soothing Mercy: truly no one despairs of God's soothing Mercy, except those who have no faith.
> Qur'an 12:87

One of the important features of the loss of hope as forbidden in Islam is that of harming of the self. Self-harm is not allowed in the Islamic tradition; Suicide is considered a sin in Islamic teaching. If a Muslim does harm themselves, he or she will not be granted heaven after death. The Qur'an says:

> Nor kill (or destroy) yourselves: for verily God hath been to you Most Merciful!
> Qur'an 4:29, 30

A wish for death is also considered forbidden in Islamic tradition, as the Prophet said, "None of you should wish for death because of a calamity befalling him" (Sahih Bukhari: vol 8, p. 202). While self-harm or wishing for death is not allowed, the individual, if in pain that is too hard to endure, can pray by saying, as the prophet instructed "O Allah! keep me alive as long as life is better for me, and let me die if death is better for me" (Sahih Bukhari: vol. 8, p. 202).

Hope in Islam is associated with an important theme, which is the *Tawakul*, or the reliance on Allah, and the commitment to change. *Tawakul* is important in Islam and means sincere trust in Allah in all aspects of life. It is also about the belief that everything refers to Allah (Ibn Kathir, 1966). *Tawakul* means the recognition of the limitations of human beings and acknowledging that illness and healing are in the hands of Allah. The belief that Allah is the source of everything supposedly will assist the person with mental health problems to have trust and depend on God in lifting the burdens of their

illness. The Qur'an states, "And on Allah let all men of faith put their trust," and "Sufficient is God for me! in him trust those who put their trust" (Qur'an 14:11; 39:38). *Tawakul* is associated with holding on to the hope that Allah will definitely make things easier for individuals who put their trust in Him.

Trust in Allah or *Tawakul* requires safeguarding Allah's commands or what is called in Islam, *Taqwá*. These concepts are associated with hope and prosperity. Trust in Allah and the safeguarding of his commands will consequently result in achieving goals and relieving distress. The Qur'an says:

> And for those who fear Allah, He (ever) prepares a way out, And will provide for him from (a quarter) whence he hath no expectation.
> Qur'an 65:2, 3

In the Muslim relationship with God, Allah is the creator and the human being is the creation of Allah. Oneness is the basic element in the connection with Allah. The role of the human being is to submit to the power of God and to worship him alone. The importance of this relationship is frequently emphasized in the Qur'an and the Hadith. An example of this concept is identified in the following verse, "He is Allah the One and Only" (Qur'an 112:1).

The belief that Allah is the source of everything and anything, and that Allah's power and knowledge is the ultimate power, forms the central understanding of this relationship. Everything is from Allah, things happen only with his permission, and if he wills them. This notion is frequently identified in the Qur'an and the Hadith. Distress, illness, poor health, and any other difficulty in life are all the result of Allah's will. In this relationship, God is the one who has the ultimate supremacy. Allah's power is emphasized, as the following verses states:

> To Him is due The Primal origin of the heavens and the earth When He decreeth a matter He saith to it: 'Be'. And it is.
> Qur'an 2:117

In this sense the truth about human beings, including their social or psychological problems, are all within Allah's knowledge. For example, human interpretations of what might be the source of their mental health problems are only endeavors. They are only possible interpretations, which might be right or wrong. Humanity's knowledge is limited within the knowledge of Allah. This concept that is frequently mentioned in the Qur'an, strongly indicates the vulnerability of humans in their relationship with Allah the creator and the one who has ultimate knowledge: "It is Allah Who knows, and ye who know not" (Qur'an 3:66).

Hope and Change: Transferring the Self

Hope is connected to dedication to making efforts to change. There is a strong correlation between hope and change. Hope will increase the motivation of a person to look for change and, furthermore, when they achieve any change, they will have more hope.

To achieve change in human life a person must begin this change within themselves. Change within requires the person to commit to modifications in their emotional and cognitive process. It is about defining goals, organizing a plan for the change, being positive, and becoming an active participant in the change process. The Qur'an states,

> Verily never will Allah change the condition of a people until they change it themselves (with their own souls).
> Qur'an 13:11

It is also acknowledged that this change requires building knowledge and commitment to transformation of the self.

Change requires the search for knowledge, and searching for knowledge is seen as a continuing process of exploration. Humans are required to be flexible, to change their interpretations, and to accept new information. All of this is within the oneness of their relationship with Allah. The Qur'an says, "Verily in that are Signs for those who reflect" (Qur'an 30:21). Individuals are responsible for seeking knowledge about the possible causes of their disorders, for ways of coping, and for help. Knowledge is used to make sense of a problem and to find practical solutions. People with psychological problems, for example, are encouraged to consult with experts and practitioners in different professions to increase their awareness of the problem and to develop strategies to alleviate their problems.

Strengthening the Relationship with Allah

One of the main ways to enhance hope and change is for the person to commit themself to rebuilding a strong relationship with Allah. Change within requires the person to strengthen their relationship with Allah. This is achieved by increasing their submission to Allah, their repentance to Allah, their reliance on Allah, and by being patient. It also requires looking for practical steps to overcome problems. Building their relationship with Allah and strengthening their spirituality will motivate the person to change and to have hope for the positive outcomes related to working on the self, seeking help, and changing one's actions.

In terms of strengthening one's relationship with Allah, it is considered that Allah's will is the essence of understanding the causes and the help that is being sought. The notion of building one's relationship with Allah is rated highly in Islamic tradition: the Qur'an says

> O ye who believe! Endure, out-do all others in endurance, be ready, and observe your duty to Allah, in order that ye may succeed.
> Qur'an 3:200

Seeking Allah should not only to take place during the easy times but also during the difficult and distressful times, during the misfortunes in life and the hardships encountered. The notion is that the difficulties in life are trials or blessings to prevent worse things from happening.

A person's relationship with Allah can also be strengthened by praying, fasting, giving to charities, *Dua* (prayers), and reciting the Qur'an. All these activities are supposedly ways to enhance hope and increase the change within people with mental health problems. While looking for help, the person is expected to be patient and tolerant. This is emphasized in the Qur'an:

> O ye who believe! Seek help in steadfastness and prayers. Lo Allah is with the steadfast.
> Qur'an 2:15

Saber, as well as being patient and tolerant, is important for succeeding in the journey of recovery. Making *dua* or asking Allah for help is another important way to foster this relationship with Allah as the Qur'an says,

> Call on Me. I will answer your prayer.
> Qur'an 40:60

Looking for practical help, whether it is from professionals or family, friends, and mentors, is encouraged in Islamic tradition as one of the responsibilities of the person who encounters difficulties. It is a method of making changes and acquiring hope. The person will seek help through these avenues, while at the same time having hope and trusting in Allah. Two of these avenues might be social networking or professional help. In relation to social networking, the person is encouraged to maintain their connections with family, neighbors, and the community. The Prophet emphasized that by saying "… And Allah will aid His slave so long as he aids his brother…" (Nawawi and Abbasi, 1980). Members of these groups are obliged to provide support and they will be rewarded for doing so. The duty of the family and community is to provide information and resources that the people in need can access (Ahmed, 1998; Kinzie, 2000). For example, looking for scientific treatment for problems related to health is part of the responsibility of Muslims. One of the basic notions in Islam is that there is a cause for illnesses and Muslims are commanded to look for treatments for their illness. They are also encouraged to integrate medical and physical healing as well as psychological and spiritual healing (Kinzie, 2000). There is always hope that medical and/or scientific developments will lead to treatment options being available in the future. The seeking of medical treatment is encouraged by Islamic tradition and this treatment should be sought from anyone who possesses scientific knowledge or who can provide support, regardless of their own religious or spiritual beliefs. The Prophet Muhammad explained to his followers in the Hadith,"There is no harm of any treatment as long as it is beneficial to human being" (Sahih Bukhari: vol. 7, p. 362).

Keeping Balance

In their journey to change within and on the practical level, people are encouraged to keep balance in their lives. In the Islamic faith, the concept of keeping balance in life is a crucial aspect of coping. This concept means that the person needs to pay attention to all aspects of their life, so as to not to emphasize one aspect at the expense of another. Balance means that the person pays attention to their physical, spiritual, social, and intellectual needs. The importance of this concept in the Islamic tradition is emphasized by its mention in the Hadith: "There is share for your eyes, share for your own self, share for your family" (Sahih Muslim: p. 564).

It is important to take care of the human and everyday needs of the lives of individuals, but also not to omit or overlook spiritual needs. Ansari notes that the concept of *nafs* (self, or living entity) is considered, in Islamic philosophy, to be a mixture of the biological, the psychological, and the spiritual (Ansari, 2002). All these parts need to be nurtured. Examples of doing this might include regular exercise (such as weekly), visits to family, relaxing with friends, and being industrious in some kind of work (whether paid or unpaid). This concept is stressed in many occasions in the Qur'an and the Hadith, where there is an emphasis on maintaining balance in order to reach the balance of the *nafs*. In the Hadith, Muslims are encouraged to balance the needs of the self and the needs of others around them, such as family members. The believer needs to revise his life frequently in order to make balance in everyday activities, family, work, and relationships.

Implication for Social Work Practice

Social workers for both Muslims and non-Muslims can utilize these concepts of hope and change when working with clients from Muslim backgrounds. It is important that they are aware of hope and change within the context of Islamic philosophy. This is because these concepts can play a major role in motivating clients to work actively on problem solving, help seeking, and coping with various life difficulties.

Before talking with clients about these the concepts of hope and change, social workers need to first explore with the client their level of religiosity, spirituality, and willingness to talk about these concepts. This provides a safe environment for clients to speak. Social workers can then explore these concepts, asking what hope and change mean for them, how these concepts are related to their faith, and exploring to what extent the clients relate these concepts to their problems, social workers can also help clients to create a plan to strengthen their spirituality, to help them better cope with difficulties. Social workers and clients can together look into the relationship between hope and change.

In summary, the concepts of hope and the changing of the self are important values of Islamic religion and spirituality. They are crucial parts of the journey of individuals to accessing the help of social workers. Hope and change help Islamic individuals to make sense of disorder, and motivate them to look for ways of coping with their problems. As a result, social workers can use these concepts to assist Islamic clients to enhance their lives, to be motivated to strengthen themselves internally, and to look for resources to help them to deal with their problems – professionally and within the community.

Bibliography

ASW, 2000. Barton. AAW Code of Ethics. Australian Association of Social Workers.

Ahmed, M., 1998. Health and Healing in the Qur'an. Triumph Pub. Co. Ltd., Sa'adu Zungur, Kano.

Al-Krenawi, A., Graham, J.R., 1999. Social work and Koranic mental health healers. International Social Work 42 (1), 53–65.

Al-Krenawi, A., Graham, J.R., 2000a. Islamic theology and prayer: relevance for social work practice. International Social Work [H.W. Wilson - SSA] 43 (3), 289.

Al-Krenawi, A., Graham, J., 2000b. Culturally sensitive social work practice with Arab clients in mental health settings. Health and Social Work 25 (1), 9–22.

Al-Krenawi, A., Graham, J.R., 2003. Principles of social work practice in the Muslim Arab world. Arab Studies Quarterly 25 (4), 75.

Ansari, Z., 2002. Islamic psychology. In: Olson, R.P. (Ed.), Religious Theories of Personality and Psychotherapy: East Meets West. Haworth Press, Inc., New York, NY, pp. 325–357.

Anthony, W., 2003. Expanding the evidence base in an era of recovery. Psychiatric Rehabilitation Journal 27 (1), 1–2. http://dx.doi.org/10.2975/27.2003.1.2.

Bonney, S., Stickley, T., 2008. Recovery and mental health: a review of the British literature. Journal of Psychiatric and Mental Health Nursing 15 (2), 140–153.

Canda, E., 1999. Spiritually sensitive social work: key concepts and ideals. Journal of Social Work Theory and Practice 1 (1), 1–15.

Canda, E.R., Furman, 2010. Spiritual Diversity in Social Work Practice: The Heart of Helping, second ed. Oxford University Press, New York, NY, USA; Oxford.

Chen, Y.Y., Koenig, H.G., 2006. Do people turn to religion in times of stress? an examination of change in religiousness among elderly, medically ill patients. Journal of Nervous and Mental Disease 194 (2), 114–120.

El-Amin, C.W., 2010. Personal and professional spirituality: Muslim social workers' perspectives. Dissertation Abstracts International Section a: Humanities and Social Sciences 70 (7-A), 2727.

Eltaiba, N., 2007. Perceptions of Mental Health Problems (Islamic Context). VDM Publisher.

Furman, L.D., Benson, P.W., Canda, E.R., 2011. Christian social workers' attitudes on the role of religion and spirituality in U.S. social work practice and education: 1997–2008. Social Work & Christianity 38 (2), 175–200.

George, L.K., Larson, D.B., Koenig, H.G., McCullough, M.E., 2000. Spirituality and health: what we know, what we need to know. Journal of Social & Clinical Psychology 19 (1), 102–116.

Gomi, S., Starnino, V.R., Canda, E.R., November 2013. Spiritual assessment in mental health recovery. Community Mental Health Journal n.p.

Graham, J.R., Shier, M.L., 2011. Making sense of their world: aspects of spirituality and subjective well-being of practicing social workers. Journal of Religion & Spirituality in Social Work: Social Thought 30 (3), 253–271.

Hall, R.E., Livingston, J.N., Brown, C.J., Mohabir, J.A., June 2011. Islam and Asia Pacific Muslims: the implications of spirituality for social work practice. Journal of Social Work Practice 25 (2), 205–215.

Horrie, C., Chippindale, P., 2003. What Is Islam? A Comprehensive Introduction (revised and updated ed.). Virgin, London.

Hyman, Corine, Handal, Paul J., 2006. Definitions and evaluation of religion and spirituality items by religious professionals: a pilot study. Journal of Religion and Health 45 (2), 264.

Ibn Kathir, I. i. i. U. c, 1966. Tafsir al-Quran al-'azim ([al-Tab'ah 1]). In: Bayrut (Ed.), Dar al-Andalus lil-Tiba'ah wa-al-Nashr.

Kinzie, J., 2000. The historical relationship between psychiatry and the major religions. In: Boehnlein, J.K. (Ed.), Psychiatry and Religion: The Convergence of Mind and Spirit Issues in Psychiatry. American Psychiatric Publishing, Inc., Washington, DC, pp. 3–26.

Kvarfordt, C.L., Sheridan, M.J., 2007. The role of religion and spirituality in working with children and adolescents: results of a national survey. Journal of Religion & Spirituality in Social Work: Social Thought 26 (3), 1–23. http://dx.doi.org/10.1300/J377v26n03_01.

Nawawi, Abbasi, S.M.M., 1980. Riyad al-salihin = Riyadh-us-saleheen: Arabic-English. Dar el Arabia, Beirut.

Schrank, B., Stanghellini, G., Slade, M., 2008. Hope in psychiatry: a review of the literature. Acta Psychiatrica Scandinavica 118 (6), 421–433. http://dx.doi.org/10.1111/j.1600-0447.2008.01271.x.

Seinfeld, J., 2012. Spirituality in social work practice. Clinical Social Work Journal 40 (2), 240–244. http://dx.doi.org/10.1007/s10615-012-0386-1.

Whitley, R., 2010. Rediscovering hope. Psychiatric Rehabilitation Journal 33 (3), 239–241. http://dx.doi.org/10.2975/33.3.2010.239.241.

Isms and History

Cesare Cuttica, Université Paris 8-Vincennes, Saint-Denis, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the role of *ism*-categories in history and in history writing. Very often casually used, *isms* permeate not just scholarly work but also everyday parlance – and yet neither outside of nor within academia do people often ask exactly why and how we use them. What do they mean? What types of *things* are they supposed to encapsulate? The article presents an overview of some of these issues through references to a chronologically and thematically varied spectrum of sources; unveils some significant early modern instances of *isms*; and reflects on how history is written *now*. Departing from a widespread epistemological approach to the study of history, it is here suggested that to employ *isms* can be useful in the attempt to understand the past, its languages, ideas, people, and systems of beliefs. Indeed, the following pages show how a less essentialist use of *isms* might provide new meaningful readings of past thoughts and events. Finally, by rethinking the role of the (intellectual) historian in dealing with *isms*, some methodological considerations on the practice of history writing as an act of observation as much as one of creation will also be advanced.

Introduction

The *Oxford English Dictionary* (OED) defines an *ism* as "[a] form of doctrine, theory, or practice having, or claiming to have, a distinctive character or relation". It also points out that *isms* are "chiefly used disparagingly" (*OED*, "ism, *n.*"). Above all, from the *OED* we learn that *isms* refer to the following units of meaning:

1. nouns of action where the suffix 'ismòs' derived from verbs in 'izein' indicates a thing done which is 'rarely concrete' (e.g., 'agonism', 'baptism', 'criticism');
2. words that express 'action or conduct of a class' of people (e.g., 'heroism', 'patriotism', 'despotism');
3. nouns applied to 'the condition of a person or thing' (e.g., 'barbarism', 'anomalism', 'parallelism');
4. names of 'a system of theory or practice' – often named after a person, subject, or object: e.g., Cartesianism, Kantism, Poujadism, Thatcherism – related to the 'religious', 'ecclesiastical', 'political', 'philosophical', 'social' spheres (e.g., 'Arminianism', 'Buddhism', 'Calvinism', 'Catholicism', 'Chartism', 'Liberalism', 'Machiavellism', 'Epicureanism', 'Positivism', 'Radicalism');
5. 'class-names' or terms describing 'doctrines or principles' (e.g., 'altruism', 'feminism', 'universalism'); and
6. terms 'denoting' linguistic characteristics (e.g., 'Americanism', 'Latinism', 'Orientalism', but also, e.g., 'archaism', 'colloquialism', 'witticism') (*OED*, "-ism, *suffix*").

This dry taxonomy is enlivened by some – mainly negative – historical examples that detail how *isms* and their users have been regarded since the eighteenth century. A telling sample of various social mind-sets and literary conventions comes from a letter dated 4 November 1789 where Horace Walpole reproached his interlocutor for his tendency to "squabble about Socianism, or some of these isms" (*OED*, "ism, *n.*"). Twenty-two years later, in a display of filial disaffection Percy Bysshe Shelley contemptuously proclaimed: "He [his own father] is nothing – no 'ist', professes no 'ism', but superbism and irrationalism" (Hogg, 1858). More cheerfully, in October 1820 in *A Letter to M. Allen* Thomas Carlyle stated: "I expect much pleasure from talking over old bygone things, from discussing Spürzheimism [Spürzheim was the founder of phrenology], Whiggism, Church of Englandism, and all other imaginable 'isms'." On a harsher note, the issue of *Fraser's Magazine for Town and Country* from June 1840 – targeting socialism and Owenism – unsympathetically declared that "[a]ll the untidy *isms* of the day shall be dissipated". In general, in nineteenth-century Britain it was believed that *isms* had spread from France as a result of the French Revolution (Cuttica, 2013). As for French views of *isms*, in 1921–22 Marcel Proust sarcastically dismissed the famous French diplomat Talleyrand as "one of the inventors of dilettantism, of I couldn't-care-lessism ['je m'enfichisme'], of many words in 'ism' now in fashion with our little snobs" (Proust, 1954). Another French literary doyen, Philippe Sollers, had one of his fictional characters mull over "[a]narchism, cubism, surrealism, communism" and scathingly remark that "'isms' after all give the impression of having been fabricated in order to hide the dawn of new names" (Sollers, 1983).

If *isms* flourished after 1789, in the early modern period it was the adjectival forms 'ist' and 'ian' and their derivative nouns 'ists' and 'ians' to be more commonly employed. *Isms* were applied as convenient shorthands like in the case of 'Arminianism' in seventeenth-century English political and religious discourse or when describing the thought and/or activities of one person from whom a movement or group of ideas took their origins (e.g., 'Hobbism'). Nowadays, it is often assumed that *isms*, which abound not only in political jargon but also in sociological, juridical, and journalistic language, constitute technical words, which the specialist is supposed to decode for us. Implied in this view is the distinction between common language as it is utilized by people in their daily interaction with the world and a more intellectually subtle conceptual use.

Searching for examples of the first kind, that is of how *isms* are present in everyday parlance, we find that on 9 August 2010 in an article published in the Italian daily *La Stampa*, Alberto Guareschi, son of the famous Italian journalist, cartoonist, and humorist Giovannino Guareschi, said that his father "was

against all 'isms', fascism, communism […] and he made use of the pen as a weapon, writing and drawing. He never took himself too seriously and his self-mockery made him closer to his readers" (Ferrari, 2010). This statement is significant in that it reveals a die-hard conviction that *isms* are not only the opposite of irony, but that they encapsulate the ideological poison whose lethal effects caused the tragic events, which occurred in the twentieth century. To confirm this view stands none other than Joseph Goebbels who referred to communism, fascism, Catholicism, and liberalism as "lies big enough to be believed" (Newbury, 2010). However, it can be said with reasonable confidence that positive (or, at least, neutral) connotations are generally attached to artistic movements such as Expressionism, Symbolism, Cubism, and Surrealism.

Other *isms* might be perceived as less dangerous, but they are still capable of arousing vehement reactions. Among these is intellectualism, which is identified with the practice of turning simple things into complicated matters. Intellectualism is despised as the opposite of good and matter-of-fact (hence healthy) conduct. Confirming this diagnosis is the case of the philosopher Edmund Husserl who in 1935 predicted that he would be accused of rescuing "an intellectualism which loses itself in theories alienated from the world, with its necessary evil consequences of a superficial lust for erudition and an intellectualistic snobbism" (Husserl, 1970). Husserl's observation proved tragically accurate: the Nazis made of 'antirationalism' and 'antiintellectualism' two explicit touchstones of their propaganda.

To be a source of heated disputes does not prevent *isms* from being successful: barbarism – to be found in Immanuel Kant's *What Is Enlightenment?* (1784) – and Orientalism – rocketed to academic stardom thanks to the greatly influential as well as highly divisive book on Western attitudes to the East written by Edward W. Said (1978) – are only two examples among a plethora of terms frequently used by the literati. It is also important to consider that there are *isms*, which are marked by the *anti* prefix. In this regard, the central question is whether it is plausible to adopt a polemical term such as anti-Americanism in a scholarly context. Obviously, this type of *ism* has derogatory connotations, which might impede a critical analysis of a phenomenon or set of ideas, and thus exclude debate. However, at other times the category can be legitimately used by scholars studying certain mind-sets, views, or policies (e.g., anti-Americanism in the context of the Cold War).

Scholars and Isms: Antipathy at First Sight

Despite their recurrence both in common parlance and in a good range of intellectual fields, there are very few studies that attend to *isms* (Williams, 1983; Quinion, 2002). One exception is A. London Fell's work on the major *isms* of early modernity and modernity (London Fell, 1996). Interestingly, London Fell started off by pointing to the numerous "forms of ideology expressed as 'isms'" thriving in the US during the fiery political debates of the 1990s between Republicans and Democrats. The then President Bill Clinton accused his opponent Newt Gingrich of unduly inciting 'rivalry' between conservatism and liberalism. For London Fell such a public interest in *isms* showed that what in the nineteenth century had been 'endemic' had by then become 'epidemic'. Myriad terms ending in *ism* could be found not just in the public discourse and the media, but also in a wide spectrum of academic disciplines from the history of art to philosophy and legal-political thought. More worryingly, historians' misuse of *isms*, according to London Fell, had led to sundry anachronistic interpretations of the past. For this reason, he adopted a – loosely conceived – deconstructionist technique to these "problematical 'constructs'", so as "to exploit" and "to explode" them (London Fell, 1996). Behind his anti'modernistic' deconstructionist project lay what London Fell called an attempt at "de-Hegelianizing" the welter of "anachronisms [sic] and ideological 'isms'" that have dominated scholarship on the modern state in the last two centuries. Viewing them as 'artificial academic devices' disconnected from the historical texts they purported to analyze, he astringently declared that 'to apply the suffix 'ism'" implies a large degree of scholarly blindness in that it obfuscates one's 'perception of a problem'. In consequence, this process 'needs to be *dis*pelled'.

In summary, if London Fell was right to argue that to study history through *isms* is open to pitfalls, it is also true that to use *isms* does not always correspond to invoking a quasiHegelian mystique of development. Rather, it means to advance a more fragmented reading of ideas in a given *historical* context with the platitudinous but fundamental goal of communicating knowledge to modern readers. It also means to accept that the historian is inevitably caught in the whirlwind of his/her time's intellectual activity, scholarly fashions, and cultural interests, which shape the direction of his/her research. If warning against anachronisms superimposed on historical sources and presentist interpretations of them is fruitfully pursued by people like London Fell, it has nonetheless to be observed that his type of approach presupposes that original texts and meanings exist, and that as such they can be recovered thanks to the *oracular* voice of the historian.

Harro Höpfl is the only theorist to have dealt systematically with the practice of using *isms* in historical and philosophical scholarship, and with the ensuing methodological problems (Höpfl, 1983). Focusing on etymology, Höpfl argued that *ism*-formations did not express the content of a doctrine or body of ideas. In their original Greek form 'ismòs', *isms* served as derogatory terms with which to define heretical, unorthodox, or false beliefs held by one's adversaries, and to characterize their conduct as much as their way of being. Subsequently, *isms* played a role in 'the construction of developmental histories' to define 'forces and agencies' behind social progress or change as well as to explain such histories in a more accessible fashion within a polemical context. Therefore, they came to represent more than simple labels for people's creeds: they expressed 'agents and causes'. In light of this evolutionary trend, Höpfl provided a series of 'genealogical' units with which to classify *isms*:

- those denoting the conduct or practices of a group;
- those generated in doctrinal and politico-religious contexts of conflict;
- those born out of politico-intellectual disputes since the nineteenth century;
- those adapted to scholarly use from their original polemical use;

- those specifically created for scholarly purposes; and
- those first employed by scholars but afterward taken up in common parlance.

Less neutrally (and a little too reductively), Höpfl asserted that *isms* serve to create 'a subject-matter for oneself'. Their popularity stems from the fact that a precise definition of what they mean is not required. Instead of being seen as a problem of synthesis, they should lead their users to 'a pause of reflection' in order to specify what one is referring to by them. This method is easier to pursue, according to Höpfl, when the referent is a set of doctrines instead of a phenomenon (e.g., 'fascism, communism, colonialism, leftism, pacifism, individualism, collectivism'). This is allegedly so because in the former case the category entails a process of self-identification on the part of the actors involved, while in the latter instance one has to confront the impossibility to grasp historical phenomena by dint of a simple *ism*. To remedy this last problem, Höpfl concluded, it is not sufficient for scholars to declare what is it they mean by their *ism* since this is tantamount to highlighting one aspect to the detriment of a larger range of components pertaining to the occurrence or circumstances under scrutiny (Höpfl, 1983).

Much more trenchantly, the founding father of the history of ideas, Arthur O. Lovejoy, had warned that *isms* are "trouble-breeding and usually thought-obscuring terms, which one sometimes wishes to see expunged from the vocabulary of the philosopher and the historian altogether" (Lovejoy, 1936). By the same token, Nicholas Henshall proclaimed that to study early modern political history implies to dispense with "a series of early nineteenth-century 'isms', which still obscure the differences between early modern consciousness and our own". According to Henshall, terms like "'liberalism', 'socialism', 'communism', 'nationalism' and 'absolutism' met the needs of political and social polemic in the 1820s and 1830s", but they do not help historical recovery of the past. Henshall disparagingly argued that "[t]he beauty of 'isms' is that everything can be made to fit them": they are generalizations which are used as historical tools (Henshall, 1996).

Pace Lovejoy and Henshall (and to some extent Höpfl too), many abhorred *isms* are nevertheless endowed with sound explanatory power and, therefore, can be used to describe portions of political thinking, social life, and economic trends expressed in various mediums in the past across the world. In this respect, Höpfl's approach whereby there are "three sorts of -isms which, at least in the hands of careful users, occasion no particular difficulty" – that is, *isms* accepted by those grouped under a category; *isms* referring to stances in a specific debate; and *isms* created 'ad hoc' to speak of "some determinate doctrine already identified independently of –isms" – is more incisive and deserves our attention (Höpfl, 1983).

As for the first remark, it is worth pointing out that the criterion of self-identification (on the part of a group of presumed adherents) as determinant in establishing the validity of an *ism* can hinder the work of the historian in that it assumes that to use an interpretative category to describe something in the past, that same category had to exist then. Such a view restrains the creative element informing history writing. In fact, if applied with due caution and awareness of their conventionality, *isms* can serve as interpretative torches with which to throw light on a more or less coherent plurality of ideas formulated at a certain time in a specific historical milieu within a more or less wide array of texts. Thus, one can reasonably adopt 'radicalism' (a nineteenth-century coinage) to refer to "radical moments of actions devoid of a programmatic radicalism, that is the development of radically pragmatic responses under the sheer pressure of events and circumstances", to describe the situation in which many early seventeenth-century English Puritans thought and acted (Davis, 2007).

With regard to Höpfl's second point on *isms* to be found in intellectual debates as objects of controversy or weapons with which to attack one's adversaries, it is necessary to consider some historical examples detailing the first appearances of some of the categories most frequently mentioned in the scholarship. This way of proceeding will provide a unique illustration of how *isms* played a decisive role in early modern polemics. It will also show that they established the meanings of concepts or currents of thought which were to rise to prominence in the nineteenth and twentieth centuries. Thus, searching for early instances of *isms* (notably, in the English context) we find that Machiavellism, Hobbism, Jesuitism, fanaticism, and a few others were fairly common both as ideological markers in political and religious controversies, and as labels with which to identify sets of ideas, principles, and patterns of behavior.

Isms: Historical Explorations

During the Elizabethan and Stuart eras, 'Machiavellism' appeared as an abusive epithet in the works of authors such as Thomas Nash (1593), George Downame (1603), Thomas Morton (1606), Andrew Willet (1607), Robert Bolton (1631), and William Burghley (1675). This might not come as a surprise, given the troubled pedigree of Machiavelli's doctrines and their impact on European political theory, and practice. And yet knowledge of this fact is a far cry from providing a thorough account of its fortunes. In fact, as Enzo Baldini has underlined, Machiavellism is an 'ambiguous and difficult-to-seize' term. By and large, it indicates 'theoretical arguments and political conducts' which throughout the centuries have directly referred to the Florentine's works. However, it has also been identified with all kinds of unscrupulous behavior and has often been equated to the paradigm of 'reason of State'. As a result, we can say that while 'Machiavellian' has to do with the theories elaborated by Machiavelli himself, 'Machiavellism' denotes a rearticulation of these theories in another historical context. This process includes translations and interpretations of his writings pursued in a more or less faithful way. In addition to this distinction, we have the terms 'Machiavellic' and 'Machiavellist': the former is employed in common parlance to refer to cunning and untrustworthy conduct (with an emphasis on plotting and dissimulation); the latter points to a disciple of Machiavelli, that is one who puts into practice his teachings without any moral scruple but with the exclusive aim of achieving personal gain. Finally, from the usage of the Florentine's works and ideas it has been possible to define Machiavellism as a 'political principle', whereby for a state it is licit to act in any way which is

conducive to its own benefit, especially if this is done in contrast to other states (Baldini, 2009).

Equally controversial albeit less enduring than Machiavellism was 'Hobbism'. If John Collinges argued that 'the world' was 'drunk with *Hobbism* & *Parkerism*' (Collinges, 1679), which he saw as nefarious political models, sundry thinkers associated Hobbism primarily with 'Atheism'. It appeared in this vein in the works of, among others, Anne Conway (late 1670s), Zachary Cawdrey (1681), Edward Stillingfleet (1681), William Sherlock (1687), Jeremy Collier (1689), and Richard Baxter (1689). Making people who were normally at odds with one another agree, 'Jesuitismus' – the bogeyman of sixteenth- and seventeenth-century English public polemics – was used by Laurence Humphrey as early as 1582, and was then to be found in the tracts of Francis Hastings (1600), David Owen (1610), and William Prynne (1655). 'Iesuitisme' appeared soon after (Georg Sohn, 1592; Christopher Bagshaw, 1601; Antoine Arnauld, 1602). It was vehemently attacked as the embodiment of scheming politics, ideological danger, and reprehensible moral conduct. Equally imbued with negative connotations and thrown around as a term of insult to describe doctrinally extreme positions typical of zealous and enthusiastic people was 'fanaticism', which can be found in the works of John Gaule (1657), Samuel Fisher (1660), John Corbet (1661), and Jeremy Taylor (1663).

Besides these fairly straightforward categories, one can point to others where flexibility of meaning and contextual changes engendered opposite definitions of the very same term. This was the case of 'patriotism', an instance of which featured in Thomas Urquhart's *Logopandecteision* (1653). In fact, whereas in the Tudor era the appeal to patriotism indicated support for the monarch, in the 1620s claiming to be on the patriotic side meant to manifest one's loyalty to the nation in contrast to the king. Patriotism served – sometimes simultaneously – opposite political projects and theoretical discourses (Condren, 2006). Furthermore, while 'Anglicanism' was not employed before the 1830s and 1840s, 'Anglianisme' had already been used in 1616 in Thomas Hanab's *Tessaradelphus* (Davies, 1992). Earlier still, attacking the jailed separatist lawyer Henry Barrow, the Episcopal archenemy of Puritans Richard Bancroft had complained about the 'new frenzy of Barrowisme' (1593) (Winship, 2009). In 1605 the nonconformist pamphleteer William Bradshaw had issued the laconically titled *English Puritanisme*, providing evidence that this other scholarly contentious *ism* existed in the Jacobean period (Bradshaw, 1605). Going back in time we discover that 'Lutheranysme' and 'Lutheranisme' as well as 'Caluinisme' were also part and parcel of early modern religious parlance, especially to signify two forms of 'schism' (Cuttica, 2013).

Having so far addressed two of the three points raised by Höpfl, it is now time to turn to the last and most relevant one: this has to do with *isms* created 'ad hoc' to speak of "some determinate doctrine already identified independently of –isms". To analyze this complex issue entails to do two things: first, to propose an inevitably limited but hopefully insightful classification of legitimate *ism*-categories which might help us to spot opinions, ideas, and moods present in a particular context; secondly, to examine what kinds of tools are deployed when interpreting the past, and, more specifically, to reflect on the practice of history writing from the methodological perspective of intellectual history. The next and final paragraphs deal with these two themes.

Rethinking Isms

If Lovejoy's claim that the historian's tool kit should be purged of *isms* is exceedingly harsh, his observation that *isms* "are names of complexes, not of simples" (Lovejoy, 1936) serves as a cornerstone of the present article. While one needs to be aware of the intricate network of theoretical avenues crossing history and of the difficulty of undoing them, it is nonetheless important to sketch a variegated multiplicity of language-patterns to be found on the canvas of this very history where they have been woven in different ways by men and women. This means that as neologisms signal a change in the perception of the world within a community, so the historian's invention of new *isms* helps to shed light on phenomena hitherto left unearthed and to dig out novel ideas sedimented in the past.

This is precisely the case of a recent article by Alexander Schmidt where the category of 'irenic patriotism' is employed to analyze 'different phases of irenic debate in the Empire' in the late sixteenth and early seventeenth centuries. By carving out 'irenic patriotism' as distinct from 'reason of State' or '*politique*' patriotism, Schmidt's research prompts important methodological questions on how to address paradigms of thought in a given historical context and on how to unfold features of past-discourses, which have not previously been taken into account (Schmidt, 2010; Cuttica, 2013). Schmidt's piece shows how *isms* display new objects of study, foster "new projects of knowledge" (Said, 1978) and unmask hermeneutical deception. Against arguments that they inevitably produce essentialist universalism, Schmidt's article confirms the didactically useful potential carried by *isms* in uncovering intellectual complexity. It also encourages us to break the unitary epistemological and interpretative model whereby the past is made of polar opposites (often expressed via misconceived and reductive *isms*).

Thus, the foregoing paragraphs enable us to formulate a different (less evolutionary than Höpfl's) proposal on *isms*, their *mode d'emploi* and their value. Six nonmutually exclusive categories emerge:

1. Cultural *isms*: they are employed in everyday conversation; they bring attention to certain phenomena, ways of speaking, habits, and mind-sets. In so doing, they often denounce, mock, and scorn. Hence they can hold an ethical dimension.
2. Descriptive *isms* (quasi-factual): they indicate specific governmental policies or strategies; programs implemented by groups with a particular intent; and individual behaviors informed by one main feature contained in the term composing the *ism*.
3. Derogatory *isms*: they are created to target people, groups, (e.g., sects, heretical fringes, rival parties, iconoclasts) and attitudes reputed to be abnormal or not in line with majority-thinking or dismissed as insignificant or disregarded as inferior.
4. Didactic *isms* (disciplinary-academic dimension): their purpose is to explain and, to a certain extent, to clarify

movements and ideas, historical periods and events, for an audience in education or part of the so-called educated-public.

5. Ideological *isms*: they refer mainly to the political, philosophical, and sociological spheres in that they denote grand narratives of thought and principles characterizing units of ideas interwoven with historical developments, especially in the nineteenth and twentieth centuries.
6. Technical *isms*: they are employed in scholarly work as a creative key with which to unlock hidden and/or neglected objects of study.

This way of proceeding is inspired by Ludwig Wittgenstein's idea of discourse as social practice and language in action. It applies to *isms* a sort of Wittgensteinian-linguistic cartography, an anthropology of language usage (Wittgenstein, 1963). This is not some kind of Borgesian utopia/dystopia whose ambition is to catalog every application of *isms*. Rather, it is an attempt to provide a picture of some central *ism*-uses. Despite being a work *ad infinitum*, this strategy might lead historians to recognize some cardinal points on their intellectual horizon and on that of the men and women they study. As a geographical map does not exactly correspond to the surface of the earth, so a schematic grid of *ism*-applications does not exhaust the entirety of all possible uses, but makes us more aware of the territory on which we move as investigators of the past. Exploring the ways in which *isms* might be adopted as well as how they fit with past terminology can be illuminating for the historian at pains to understand different theories, convictions, and prejudices.

In summary, what has been attempted above is no more than an outline of some important debates and tendencies concerning the use of *isms* in both public discourse and the scholarship. Having provided a broad picture of how various *isms* first emerged, of how they function in nonstrictly academic circles and of how they have been – and still are – treated by scholars, the article has then concentrated on the only work expressly dedicated to the study of *ism*-categories. Relying on Harro Höpfl's pioneering but now dated article, it has advanced some suggestions on how to make a more accurate and thorough use of *isms*. Now, these suggestions need to be connected to the activity of history writing (notably, from the viewpoint of intellectual history). To do so might further elucidate important aspects of the *ism*-question.

Isms and History Writing

If we take heed of several voices heard in this piece, we should get rid of *isms* altogether. This is due – we are told – to their reductionist configuration of the past of which they are supposed to give account. And yet the problem remains: what do we replace (many of the) *isms* with? What are the alternatives? And, if we identify them, why might these be in any lasting sense less problematic terms to become part and parcel of our scholarly vocabulary and methodological toolbox? To attempt an answer to these questions implies setting forth a few considerations on the *métier d'historien*. Practitioners of different disciplines within the larger framework of history writing disagree about *the right thing to do* with regard to method and often dismiss ways of proceeding which are not in tune with theirs as unfounded and sloppy. Instead of perpetuating this less than sympathetic trend and prescribe recipes for *the good scholar*, it might be more valuable to explore the field of research in which one pursues his/her work. In brief, the issue of whether *isms* should or should not be used might depend on how we look at our practice.

More specifically, it might have to do with what one thinks of the conventional devices historians deploy in order to put forward new readings of history. As terms like 'early modern', 'modern', and 'postmodern' help to divide into smaller units of time the chronological vastness opened before us, so the employment of *isms* helps to better grasp portions of the theoretical panorama of the past. Equally, classifications of past traces according to whether they are written or oral, documents or texts, public/official or private, erudite or popular, in a foreign language or our own, translated and, if so, by whom and under what circumstances, constitute indispensable lenses through which we bring into focus historical experience(s). All of these *tricks* – *isms* included – are the fruit of conventions, which can facilitate the study of history and strengthen our inevitably limited approach to it.

To address the complex issue of *isms* is also an attempt to rethink the conditions of knowledge available to the intellectual historian. In this sense, it is crucial to ask whether the (intellectual) historian is an antiquarian, a retriever of the past or a shaper of its cultural codes for consumption in the here and now. Is studying the past a reproductive endeavor or does it imply a creative effort? Should historical research be guided by the ideal goal of telling what *really* happened? Or, more humbly, should it consist in a quest for understanding through the fallible instruments at our disposal, the meanings of textual traces left behind on that horizon? If the task of the intellectual historian is to become aware of concepts and ideas nowadays accepted in an often uncomprehending manner, then to examine how *isms* are employed is not only justified but necessary. As an archaeologist excavating the past and bringing to life remnants of it, historians should not refrain from using languages, categories, and models which, while being respectful of the rhetoric of past authors, can communicate to modern readers a sense of what it *might have been like* back then (Cuttica, 2012).

This project entails an inventive effort on the part of the historian. It presupposes a practitioner willing to adopt a different pose before past-fragments left in, e.g., texts, documents, paintings: one that applies imagination and irony to the art of unraveling *and* narrating what went on in both the foreground and the background of historical experience. Moved by the aim of molding new ideas, images, or concepts so as to better capture events and actions, words and theories scattered on the often alien canvas of history, the intellectual historian carries out his/her activity more like an *artist* than like a scientist. As Brian Young put it, he/she is active and persistent 'in cultivating the artist's penetrating eye'. Shaped by "the idea of history as literature, and literature as history", intellectual historians are "uniquely sensitive masters of rhetoric (and entrepreneurs of the rhetorical repertoire rather than philosophers)" (Young, 2002). This profile suggests that, "historians,

like novelists, are makers of order" (Worden, 2001). It also reminds us that to use *isms* is integral to the enterprise of translating past utterances for twenty-first-century readers. In other words, to convey the past to an audience requires transmitting images in an intelligible way, and this is precisely what *isms* do by helping readers to see connections that might otherwise have remained obscure.

Conclusion

Might the (intellectual) historian thus be described as an *ismist*, as one cultivating the art of *ismism*, which may well, from time to time, involve not just the radical redefinition of existing *isms* but even the invention of new *isms*? If so, to operate within the methodological framework of intellectual history does not mean to a-priori proclaim whether *isms* should be used or strictly avoided, but to ask ourselves – as has been done in this article – how *history* has employed them; how we can understand them; and what use of them might be productive of new knowledge. When handling *isms* the intellectual historian acts, on the one hand, as somebody who traces their developments through successive epochs and multifarious contexts and, on the other, as somebody who carefully clarifies how he/she is managing them in his/her effort to illuminate and narrate the past.

This critical inquiry should conduce us to realize that *isms* are both the elaboration and the outcome of choices made at different times in history. This approach might also encourage us to think about the future of *isms*, notably of those that, faced with the incumbent global economic recession, have failed to provide solid answers to it (e.g., capitalism, neoliberalism, socialism, libertarianism). By doing so, we would keep alive the controversial but engaging history of *isms* and their practice.

Acknowledgment

An earlier and longer version of this article was originally published as an article: C. Cuttica, '*To Use or Not to Use*… The Intellectual Historian and the Isms: A Survey and a Proposal', *Études Épistémè*, 23 (Spring 2013), G. Buchanan: Textes et traductions, Varia. URL: http://revue.etudes-episteme.org/?to-use-or-not-to-use-the. The author thanks the editors of the journal for their permission to reuse some of the original material here.

See also: Absolutism, History of; Cambridge School of Intellectual History; Cultural History; Historical Thought and Historiography: Current Trends; Historicism; Intellectual History; Linguistic Turn and Discourse Analysis in History; Political History: History of Politics.

Bibliography

Baldini, A.E., 2009. Machiavellismo e machiavellismi: progetto di ricerca e messa a punto di un concetto. In: Arienzo, A., Borrelli, G. (Eds.), Anglo-american Faces of Machiavelli. Polimetrica, Monza, Italy, pp. 23–48.
Bradshaw, W., 1605. English Puritanisme (London).
Collinges, J., 1679. A Reasonable Account Why Some Pious, Nonconforming Ministers in England Judge It Sinful for Them to Perform Their Ministerial Acts […]. London.
Condren, C., 2006. Argument and Authority in Early Modern England. The Presupposition of Oaths and Offices. Cambridge University Press, Cambridge.
Cuttica, C., 2012. What type of historian? Conceptual history and the history of concepts: a complex legacy and a recent contribution. History and Theory 51, 411–422.
Cuttica, C., 2013. To use or not to use… The intellectual historian and the isms: a survey and a proposal. Etudes Epistémè 23 (Spring 2013), G. Buchanan: Textes et traductions, Varia. URL: http://revue.etudes-episteme.org/?to-use-or-not-to-use-the.
Davies, J., 1992. The Caroline Captivity of the Church. Charles I and the Remoulding of Anglicanism 1625–1641. Clarendon Press, Oxford.
Davis, J.C., 2007. Afterword: reassessing radicalism in a traditional society: two questions. In: Burgess, G., Festenstein, M. (Eds.), English Radicalism. Cambridge University Press, Cambridge, pp. 338–372.
Ferrari, M.B., 2010. Incontri sul Grande Fiume 5. Ma cosa fai Guareschi in bicicletta? Nel paese del parmense dove vive il figlio: Pedalava per raccogliere storie e dimagrire. La Stampa, 9 August, p. 21.
Henshall, N., 1996. Early modern absolutism 1550–1700: political reality or propaganda? In: Asch, R.G., Duchhardt, H. (Eds.), Der Absolutismus – ein Mythos? Strukturwandel monarchischer Herrschaft in West- und Mitteleuropa (ca. 1550–1700). Bölau, Köln, pp. 25–53.
Hogg, T.J., 1858. The Life of Percy Bysshe Shelley, 4 vols. (i). E. Moxon, London.
Höpfl, H., 1983. Isms. British Journal of Political Science 13, 1–17.
Husserl, E., 1970. The venna lecture. Appendix 1: philosophy and the crisis of European humanity. In: Husserl, E. (Ed.), The Crisis of European Sciences and Transcendental Phenomenology (Translated, with an Introduction, by D. Carr). Northwestern University Press, Evanston, pp. 269–299.
London Fell, A., 1996. Origins of Legislative Sovereignty and the Legislative State: Volume 5, Modern Origins, Developments, and Perspectives against the Background of Machiavellism. Book II: Modern Major Isms (17th–18th Centuries). Praeger, Westport, Conn., and London.
Lovejoy, A.O., 1936. The Great Chain of Being: A Study of the History of an Idea. Harvard University Press, Cambridge, Mass.
Newbury, R., 2010. Graham Greene: Spiare, vizietto di famiglia. La Stampa, 31 December, p. 35.
Oxford English Dictionary. Ism, n. http://www.oed.com (accessed 30.01.14.).
Oxford English Dictionary. -Ism, suffix http://www.oed.com (accessed 30.01.14.).
Proust, M., 1954. Sodome et Gomorrhe. In: À la Recherche du Temps Perdu. Bibliothèque de la Pléiade. Gallimard, Paris (tome ii, ii partie, ch. iii).
Quinion, M., 2002. Ologies and Isms. A Dictionary of Word Beginnings and Endings. Oxford University Press, Oxford.
Said, E.W., 1978. Orientalism: Western Conceptions of the Orient. Routledge, London.
Schmidt, A., 2010. Irenic patriotism in sixteenth- and seventeenth-century German political discourse. The Historical Journal 53, 243–269.
Sollers, P., 1983. Femmes. Gallimard, Paris.
Williams, R., 1983. Isms. In: Williams, R. (Ed.), Keywords. A Vocabulary of Culture and Society, first ed. 1976. Flamingo, London, pp. 173–174.
Winship, M.P., 2009. Freeborn (Puritan) englishmen and slavish subjection: popish tyranny and puritan constitutionalism, c. 1570–1606. English Historical Review 124, 1050–1074.
Wittgenstein, L., 1963. Philosophical Investigations (E. Anscombe, Trans.). Blackwell, Oxford.
Worden, B., 2001. Roundhead Reputations: The English Civil Wars and the Passions of Posterity. Penguin, Harmondsworth.
Young, B., 2002. The tyranny of the definite article: some thoughts on the art of intellectual history. History of European Ideas 28, 101–117.

Issue Evolution

Edward G Carmines, Indiana University, Bloomington, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The standard model of partisan change emphasizes the role of critical elections in bringing about dramatic and quickly realized changes in the party system. In contrast, the model of issue evolution focuses on more gradual and incremental change in the party system – change that is driven by issue competition among partisan elites, which eventually leads to a transformation in mass party alignments along the issue dimension. Thus, there are two characteristics that distinguish this model of partisan change. First, issue evolution presumes that most partisan change occurs gradually over a lengthy span of time, unfolding over several elections and perhaps decades instead of a single election. Second, the model emphasizes the role of partisan elites in initiating and leading political change. The crucial role of the mass electorate is to respond – or, more frequently, fail to respond – to the issue cues of partisan elites. The process of issue evolution, then, involves elite reorientation on the issue that eventually leads to mass issue realignment.

Issue evolution refers to a model of partisan change in which issue competition among political elites, often spurred on by political activists, gradually leads to a transformation in mass party alignments along the issue dimension. There are two characteristics that distinguish this model of partisan change. The first is the critical role played by elites. Unlike most models of elite–mass linkage political elites are not seen as merely responding to the demands of the mass electorate but as shaping and establishing the reputation of political parties on a political issue. The masses are seen as eventually reacting to these elite cues (MacDonald and Rabinowitz, 1987; Zaller, 1992). However, since it takes time for parties to establish a reputation on a political issue and for citizens to perceive and act on this difference between the parties, mass partisan change typically occurs gradually over a lengthy span of time. Ultimately, the transformation of mass party alignments occurs when the issue moves current voters to change their partisan identifications and biases the recruitment of new partisans.

The Standard Model of Partisan Change: The Critical Election Perspective

The standard model of partisan change was provided by V.O. Key in the 1950s (1955). His model of critical elections postulates that dramatic events can quickly transform the fortunes of political parties, leading to a realignment of the party system during the course of a single election. His account seems to fit almost perfectly the New Deal realignment. The Republicans, dominant in national politics since the 1860s, were held responsible for the Great Depression and, as a result, millions of voters deserted the 'Grand Old Party' and became Roosevelt Democrats during the 1932 presidential election. The New Deal realignment thus seems to be a case of partisan change that was both dramatic and quickly realized, resulting in the rapid replacement of an old majority party by a new one.

Criticisms of the Standard Model

Key's model of critical election realignments, especially as amplified in Burnham's *Critical Elections and the Mainsprings of American Politics* (1970), became the standard framework for understanding changes in the US party system. Despite its popularity and influence, however, several major criticisms have been directed at the critical election realignment perspective (Mayhew, 2002). In the first place, it has been impossible to specify exactly what constitutes a critical election. This vagueness has led to claims of realignment for virtually every recent presidential election (Bass, 1991). The generalizability of the theory has also been questioned since it is difficult to build a general theory based on a single historical case (Carmines and Stimson, 1989; Silbey, 1991). Finally, it has been argued that the critical election perspective overemphasizes short-term partisan change, while neglecting more gradual changes that extend over a long period of time (Carmines and Stimson, 1989).

The Issue Evolution Model

Carmines and Stimson (1981, 1986, 1989; also see Carmines and Wagner, 2006) have developed an alternative model of partisan dynamics which they refer to as issue evolution. Unlike the classic realignment model, it presumes that most partisan change occurs gradually and incrementally over a lengthy span of time, unfolding over several elections and perhaps decades instead of a single election. At no single point in time is partisan change necessarily striking but its culmination through time can leave an indelible imprint on the party system.

Most issues, according to Carmines and Stimson, lack the capacity to generate significant partisan change. This is because most issues have one or more of the following limitations: they reinforce the existing party alignment, they never capture public attention or they have effects that are entirely short term. It is only those relatively rare issues that transcend these limitations that have the capacity to lead to fundamental and permanent change in the party system. In other words, an issue evolution only occurs when a new issue emerges that cuts across the existing party–issue alignment, captures the public's attention for an extended period of time, and cannot be resolved in the normal policy-making process.

Carmines and Stimson (1980) put particular emphasis on the amount of cognitive processing required to deal meaningfully with a political issue as a critical dividing line between those issues that can and cannot lead to issue evolutions. Issues that are complex, technical, and obscure are not good candidates for issue evolution. Instead, if issues are to transform mass party alignments, they must be able to be understood at a 'gut' level, engaging the raw emotions of the citizenry.

Two basic aspects of an issue evolution are its structure and sequence.

The Structure of Issue Evolution

There are two main actors in issue evolution: partisan elites and the mass electorate. Partisan elites include presidents, members of Congress and state legislatures, and candidates for major political offices. A party's position and reputation on a political issue is established through the public acts and pronouncements of their elites. When a president sets out his position on an issue in the State of the Union Address or when senators express their opinions on an issue through their roll-call votes, they are helping to build the position of their party on a given issue. When this process is repeated over time and in different political arenas, the party establishes a reputation on the issue.

A party's reputation on the issue, in turn, serves as a partisan cue to the mass public.

But why do elites of different parties adopt contrasting positions on at least some political issues and how does this process begin? Riker (1982) suggests a plausible answer. A political party, finding itself temporarily in the minority, always has an incentive to defeat the majority winning party and, according to Riker, political issues are a natural vehicle for this purpose. New issues – if they can split the coalition of the majority party and are sufficiently attractive to the electorate – can turn old losers into new winners (and old winners into new losers) (Carmines, 1991). Thus, for Riker, losing minority parties always have an electoral incentive to adopt distinctive positions on new issues in their never-ending quest to upset the current status quo. And for the same reason, minority parties are particularly attentive to the appeals of grass roots political activists who are looking for parties to advance their political causes.

Riker's logic can also be applied to minority factions within majority party coalitions. For example, the new left operating within the majority Democratic coalition appealed to anti-Vietnam War forces and those on the left on the new moral and cultural issues of abortion and gay and women rights to help it take control of the party in 1972 (Layman, 2001). The point is that it makes sense for losing and disadvantaged political elites to develop and advertise new issue proposals in an effort to improve their public standing and gain political leverage.

The other major group in the issue evolution process is the mass electorate. The political masses are not only less active in the political process than elites they also have no particular incentive or capacity to create new issue alternatives themselves. Their key role in issue evolution is to respond – or, more frequently, fail to respond – to the issue cues of party elites. Moreover, even when the citizenry does respond to the issue cues of partisan elites, it does not do so instantaneously. Instead, the response is gradual, delayed, and protracted.

The Sequence of Issue Evolution

According to Carmines and Stimson's model of issue evolution, changes in elite partisan behavior do not lead directly to mass partisan response. Rather, two intervening steps are necessary to link elite policy shift to mass issue realignment. First, the mass electorate must perceive a difference in the issue positions of the parties. But awareness of party differences on an issue, by itself, is not likely to lead to mass issue realignment. Citizens must care about the issue. The changed perceptions of the parties on the issue must carry with them a strong dose of affection and disaffection for the parties if they are to overcome the stubborn inertia of existing partisan identifications (Campbell et al., 1960). Only when these two intervening conditions are met – clarified mass cognitive images of the parties and polarized affection toward them – will issue reorientation among partisan elites lead to changes in mass issue alignment.

Time ordering is critical. That is, the issue evolution process should follow a specific temporal sequence with changes in the cognitive and affective images of the parties occurring after elite reorientation on the issue but before mass partisan response. According to Carmines and Stimson's model of issue evolution, party activists play a pivotal role in this process, linking elite initiatives to mass responses.

Illustrations

What partisan changes have been illuminated by the issue evolution model? Let us consider three cases of issue/partisan change in contemporary US politics that have been examined through the lens of the of issue evolution perspective. For other examples, see Burns (1997) on taxes, Fordham (2007) on defense, and Shoch (2001) on trade.

Racial Issues

Carmines and Stimson (1989) offer a detailed account of racial politics in mid-twentieth century America as an example of issue evolution. Historically, Republicans – the party of Lincoln – had a more progressive racial tradition than did the Democrats with their support first of slavery and later of segregation. But during the mid-1960s their positions on race became reversed. The Democrats became the party of racial liberalism while the Republicans became the party of racial conservatism.

Carmines and Stimson show that this partisan transformation on race closely followed the model of issue evolution. In the first place, the change was gradual and incremental, unfolding gradually over many decades instead of within a single critical election. Indeed, according to evidence collected by Schickler (2013) the parties' altered positions on race have their origins in the late 1930s when a series of developments began to indicate that the Democratic Party rather than the Republican Party would eventually become more hospitable to blacks' civil rights. Through an analysis of

presidential speeches, party platforms, and congressional roll-call votes, Carmines and Stimson demonstrate that partisan elites altered their racial positions years before mass partisans reversed their positions. There was strong evidence, in other words, that elites rather than the general public initiated the issue evolution on race. Finally, the racial issue case follows the sequence of changes postulated in the issue evolution model. The process began with reorientation among partisan elites and ended with mass issue realignment but bracketed between the beginning and ending of this process was the electorate's increasingly accurate perceptions of the parties' racial positions and polarized affections based on them. Party activists played a crucial role in clarifying for the public the evolving racial positions of the parties through their highly visible activities undertaken during election campaigns. All in all, the partisan transformation of racial issues fits the issue evolution model extremely well.

Abortion

Changes in the abortion issue during the 1980s and 1990s provide another example of an issue that has undergone a partisan evolution, according to a study by Adams (1997). Examining roll-call votes in the House and Senate, Adams shows that Democrats and Republicans were only moderately divided over abortion throughout the 1970s. But, beginning in the early 1980s, the difference between these partisan elites grew steadily during the next 15 years. By the mid-1990s, congressional Democrats clearly had adopted the pro-choice position just as Republicans had become overwhelmingly pro-life. This change unfolded gradually and incrementally over a two-decade period, a pattern predicted by the issue evolution model.

Mass partisans have also become more polarized on abortion during this period, as Adams demonstrates through an analysis of public opinion polls. Throughout the 1970s, and up until the mid-1980s, Democrats, in the aggregate, were actually more pro-life than Republicans. But during the 1990s a partisan shift occurred, resulting in Democrats becoming significantly more pro-choice than Republicans.

Adams shows, finally, that there is likely a causal connection between the elite and mass changes, with elite-level changes producing mass level responses. Not only did the shifts in abortion attitudes occur earlier among partisan elites than among ordinary partisans but the degree of issue polarization is sharper and more pronounced in the former than the latter. Further, other research demonstrates that party activists played a key role in the partisan transformation of the abortion issue (Carmines and Woods, 2002).

In sum, in every major regard, abortion is consistent with the issue evolution model. Issue polarization gradually develops among both partisan elites and masses and as the issue cues of partisan elites become clearer and more distinct over time the partisan masses respond more readily to them. The process culminates in issue realignment among mass partisans.

The New Religious Divide and Moral and Cultural Issues

In *The Great Divide: Religious and Cultural Conflict in American Party Politics* (2001), Layman provides a comprehensive account of how the new traditionalist–modernist cleavage in American religion has come to shape the contemporary American political party system, the extent to which the religious and cultural divide between the Democratic and Republican parties has grown over time, and the process through which this growth has occurred. Similar to Carmines and Stimson's discussion of the types of issues capable of creating issue evolutions, Layman contends that the traditionalist–modernist religious cleavage came to affect party politics because it is associated with political issues that are easily understood, arouse strong emotions in a large number of people, reflect a deep-seated social conflict, cut across the lines of party conflict that existed in the 1970s, and appealed to the strategic incentives of certain groups of politicians within both parties.

He draws on the issue evolution framework to develop a model of the process through which polarized stands by party leaders on cultural and moral issues such as abortion, gay rights, and the role of women have been translated into changes in the religious composition of the parties' coalitions. Party activists have played a significant role in making this connection by sustaining and extending elite polarization and by providing cues to an inattentive public about the nature of issue change at the elite levels of party politics. The mass electorate gradually responded to elite- and activist-level change by first adjusting its perceptions of and affect toward the parties, and ultimately by reshaping its partisan ties along traditionalist–modernist religious lines.

Layman's research shows that the process of partisan religious change generally has followed the patterns predicted by the issue evolution model. It began among party elites and activists and activists also provided the link between elite-level issue change and the response by a relatively inattentive public. Over time, the mass public has come increasingly to see differences in the positions of the parties on cultural and moral issues and to care about these differences. These changes in mass perceptions of the parties and affect toward them have led ultimately to a substantial shift in the religious and cultural composition of the parties' electoral coalitions.

See also: Law: Change and Evolution; Political Communication.

Bibliography

Adams, G.D., 1997. Abortion: evidence of an issue evolution. American Journal of Political Science 41, 718–737.

Bass Jr., H.F., 1991. A reader's guide and bibliography. In: Shaffer, B.E. (Ed.), The End of Realignment? University of Wisconsin Press, Madison WI, pp. 141–178.

Burnham, W.D., 1970. Critical Elections and the Mainsprings of American Politics, first ed. Norton, New York.

Burns, J., 1997. Party policy change: the case of the democrats and taxes 1956–1968. Party Politics 3, 513–532.

Campbell, A., Converse, P., Miller, W., Stokes, D., 1960. The American Voter. Wiley, New York.

Carmines, E.G., 1991. The logic of party alignments. Journal of Theoretical Politics 3, 65–80.

Carmines, E.G., Stimson, J.A., 1980. The two faces of issue voting. American Political Science Review 74, 78–91.

Carmines, E.G., Stimson, J.A., 1981. Issue evolution, population replacement, and normal partisan change. American Political Science Review 75, 107–118.

Carmines, E.G., Stimson, J.A., 1986. On the structure and sequence of issue evolution. American Political Science Review 80, 901–920.
Carmines, E.G., Stimson, J.A., 1989. Issue Evolution: Race and the Transformation of American Politics. Princeton University Press, Princeton, NJ.
Carmines, E.G., Wagner, M.W., 2006. Political issues and party alignments: assessing the issue evolution perspective. Annual Review of Political Science 9, 67–81.
Carmines, E.G., Woods, J., 2002. The role of party activists in the evolution of the abortion issue. Political Behavior 24, 361–377.
Fordham, B., 2007. The evolution of republican and democratic positions on cold war military spending. Social Science History 31, 603–634.
Key Jr., V.O., 1955. A theory of critical elections. Journal of Politics 17, 3–18.
Layman, G., 2001. The Great Divide: Religions and Cultural Conflict in American Party Politics. Columbia University Press, New York.
MacDonald, S.E., Rabinowitz, G., 1987. The dynamics of structural realignment. American Political Science Review 81, 775–796.
Mayhew, D.R., 2002. Electoral Realignments: A Critique of an American Genre. Yale University Press, New Haven, CT.
Miller, G., Schofield, N., 2003. Activists and partisan realignment in the United States. American Political Science Review 97, 245–260.
Riker, W.H., 1982. Liberalism against Populism. W H Freeman, San Francisco, CA.
Sanbonamatsu, K., 2002. Democrats, Republicans, and the Politics of Women's Place. University of Michigan Press, Ann Arbor, MI.
Schickler, E., 2013. New deal liberalism and racial liberalism in the mass public, 1937–1968. Perspectives on Politics 11, 75–98.
Shoch, J., 2001. Trading Blows: Party Competition and U.S. Trade Policy in a Globalizing Era. University of North Carolina Press, Chappel Hill, NC.
Silbey, J.H., 1991. Beyond realignment and realignment theory: American political eras, 1789–1989. In: Shafer, B.E. (Ed.), The End of Realignment? University of Wisconsin Press, Madison, WI.
Wolbrecht, C., 2000. The Politics of Women's Rights: Parties, Positions, and Change. Princeton University Press, Princeton, NJ.
Zaller, J.R., 1992. The Nature and Origins of Mass Opinion. Cambridge University Press, Cambridge, UK.

Issue Networks: Iron Triangles, Subgovernments, Policy Communities, Policy Networks

Mark Thatcher and Juergen Braunstein, London School of Economics, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

As policy making has increasingly been recognized to involve interest groups, analytical frameworks have been developed of state–interest group relations. Initially, there were limited and simple categories, such as issue networks, iron triangles, and policy communities. However, more complex policy network 'models' have been created that offer typologies of state–interest group relations and then seek to explain the forms of such relations. Most recently interorganizational approaches to policy networks have been established that draw on institutional and collective action frameworks that examine policy making by studying how the formation of ties among actors influences network output and the interaction among actors. Network frameworks allow inclusion of public and private actors and represent a move away from seeing policy making as an activity confined to public actors arranged in a simple hierarchy of relations.

Studies of issue networks and other forms of networks examine the relationships between interest groups and the state. As political scientists analyzed policy making, they found that simple models of representative government in which only elected officials and political parties took decisions and had a clear, coherent overview of policy across government were inadequate. Instead, interest groups were active participants in decision-making and implementation, building up close relationships with elected and unelected officials. Policy making was fragmented, with differing participants and relationships in each segmented policy field.

Early Development

Initially, in the 1960s and 1970s, there were two distinct views of the relationships between interest groups and officials. 'Iron triangles' and 'policy communities' suggested that very closed 'subgovernments' of stable participants in long-term relationships effectively made policy in well-delineated policy fields such as agriculture or defense. In the United States, 'iron triangles' were composed of the interest groups active in a particular policy field, members of Congress (especially those sitting in the relevant committees), and officials (particularly unelected) in the relevant parts of the bureaucracy with powers in the field. These actors knew each other, were in constant interaction and exchanged concessions, advantages, and information. They were a closed group, able to exclude outsiders. The concept of 'policy community' was used and developed in studies of Britain from the 1970s onward (Jordan, 1990). Policy formation was argued to be segmented into subsystems composed of policy communities of civil servants from government departments (or units within them) and interest groups. They were built on exchange relationships and a sense of 'community,' and possessed a set of dominant values (Richardson and Jordan, 1979). Both iron triangles and policy communities paid close attention to interest groups with powerful and continuing material interests, the basis of exchange relationships.

The concept of 'issue networks' was developed by Hugh Heclo in opposition to the 'iron triangles' literature in the United States, which he argued monopolized attention on closure and subgovernments, but ignored forces for fragmentation and the development of new forms of interest groups (Heclo, 1978). An issue network consists of a large number of issue-skilled 'policy activists' drawn from conventional interest groups and sections of the government, together with academia and certain professions but also comprising expert individuals regardless of formal training. Participants are constantly changing, and their degree of mutual commitment and interdependence varies, although any direct material interest is often secondary to emotional or intellectual commitment. No one is in control of the policies and issues covered by a network and it is very difficult to identify clearly the dominant actors.

Neither policy communities nor issue networks offered a free-standing, developed model of policy making (Thatcher, 1998). Rather, they represented enriching metaphors and useful concepts that were empirically applicable, and provided counterbalances to previous well-established views of policy making. They encouraged disaggregated analysis and drew attention to the effects of the fragmentation of government into subsystems, the interdependence of government and interest groups, and increased specialization in policy making. However, issue network, iron triangle, and policy community literature suffered from several limitations: they covered only specific categories of relations between groups and officials; they omitted key aspects of policy making, such as ideas, the distribution of power among actors, and change; the factors explaining the emergence of different types of an issue network or policy community remained general.

Network Typologies

In the face of these limitations, more general network approaches were developed. One path involved creating typologies of networks. The concept of 'policy network' or sometimes 'policy community' became generic, covering different forms of relationship between interest groups and officials arising from variations in networks characteristics. Rhodes and Marsh (1992) offer a six category typology

covering a continuum from issue networks to policy communities and based on differences in the degree of integration, membership, and distribution of resources among members. Jordan and Schubert (1992) claim that different terms for state/group relations in fact refer to variants of networks and hence they place terms such as iron triangles, clientelism, subgovernment, policy community, and various types of corporatism and pluralism in relation to each other using three 'dimensions,' namely number of members, whether the network is sectoral or transectoral, and the level of its stability.

In extending the scope of networks, the typologists pursue more ambitious theoretical aims. They seek to relate existing different terms and categories to each other within one overarching typology in order to avoid a jumble of overlapping and confusing terms. They claim to assist in ordering and categorizing information, permitting policy subsystems to be described, classified, and compared. Explanatory aims have ranged from specifying the key characteristics of policy networks in order to allow the systematic ordering of material, linking the 'structural' features of state and society and policy making (Atkinson and Coleman, 1989) and offering a 'model' of the role of interest group intermediation based on different forms of power dependence embodied in the types of network (Rhodes and Marsh, 1992).

Yet network typologies have suffered from considerable difficulties (Dowding, 1995). Terminological disputes over whether the generic term should be policy network or policy community have created confusion. Placing differing forms of state–interest group relations under the heading of 'policy network' has not led to a common terminological currency. As the concept of policy network/community became generic, definitions have become loose and frequently contradictory. The empirical applicability of network typologies has encountered important obstacles. Several of the 'dimensions' that form the basis of categories are difficult to operationalize, or are unlikely to generate agreement between different researchers; examples include the 'degree of integration,' 'preeminence,' 'serving the interests of producers' (Rhodes and Marsh, 1992). Categories are poorly defined and identifying a type of network in practice is difficult, since categories are not mutually exclusive and different types of policy network or aspects of several network types can exist in the same policy area.

Interorganizational Analyses

An alternative generalizing strategy has been to develop interorganizational analyses of networks. Drawing on social network analysis and work on interorganizational relations, they view policy making as an example of decision-making by interdependent actors who engage in exchange relations (Knoke et al., 1996). Policy is seen as the product of complex interactions among actors drawn from both 'public' and 'private' sectors. Such actors are partly interdependent but also partly autonomous. They are linked horizontally without being part of a single organizational hierarchy; their relations are based on exchange, thereby producing policy networks, but nevertheless combine elements of conflict with those of cooperation (Marin and Mayntz, 1991). Analysis focuses on the interaction of actors; the unit of analysis consists of the relations among network participants.

Interorganizational network analyses 'map' network structures, using qualitative and quantitative methods to reveal linkages between actors. Frequently mathematical techniques derived from formal network analysis are used to establish the network position of actors and relationships among them, applying indicators/measures such as intensity of communication, reputation among network participants, or resources. Mapping networks to provide a systematic picture of linkages serves several purposes. It draws attention to the interrelations and interdependence between public and private actors. Quantification enables complex patterns to be revealed that would be difficult to find because the events studied are too numerous, vast, and/or scattered and allows comparison of networks across sectors, countries, and time periods. The most important ambition of interorganizational approaches is to offer a 'structural analysis' by looking at patterns of network relations. It is insufficient merely to analyze the attributes of individual actors, for the influence of such attributes on behavior depends on the network of relations of each actor (Emirbayer and Goodin, 1994). Instead, relations among actors must be studied as a whole. A structural analysis of private and public actor configurations in policy formation identifies the position of actors and the ways in which networks influence actors' capacity to use their resources. It therefore permits explicit consideration of the ways in which different structures create different mechanisms or logics in policy making.

The scope of interorganizational approaches is very broad. Thus, for example, policy network refers to 'structures' located 'beyond or between policy markets and policy hierarchies' that are composed of 'relatively autonomous action units' (Marin and Mayntz, 1991). From an 'organizational state' perspective, Knoke et al. have conceived policy networks as relations linking organized interests with public officials, transforming heterogeneous communities of policy actors into webs of common benefit-seeking actors (Knoke et al., 1996). To operationalize and apply in empirical studies, important choices must be made to narrow down the scope of interorganizational networks. Similarly, applying quantitative techniques requires the choices over matters such as questions posed, categories created, the relationships to be analyzed, and variables to be controlled for. Social network analysis may embody a particular 'theoretical orientation' toward the structure of the social world, notably the need to study behavior as a set of relationships, but it provides a particular set of concepts and methods and a framework for testing arguments/models rather than a substantive set of hypotheses or a specific body of substantive theory. Social network analysis is good at explaining stability and continuity of networks and policy outcomes by drawing attention to networks as forms of interest intermediation or as new forms of governance, which substitute government. However, these approaches poorly explain changes in policy networks and changes in policy outcomes (for a summary of the argument see Richardson, 2000).

Analyzing Networks and Change

Addressing this criticism, more recent network research focuses on explaining changes in policy networks as well as changes in policy outcomes over time by adding additional explanatory

factors. These include ideas, processes and mechanisms, and contextual macrolevel factors (Daugbjerg, 1999; Peterson, 2003). Networks are perceived as intermediating and opportunity structures through which ideas, individual and group preferences operate to influence policy outcomes. As such these studies argue that the mere existence of ideas and preferences fails to explain or predict the specific design of policy. Therefore, they suggest looking at networks as intervening variables in order to understand national differences in the implementation of ideas into policies (Daugbjerg and Pederson, 2004). Furthermore they propose to integrate macrolevel factors such as the impact of national and supranational institutions with veto points at the mesolevel network dynamics.

Analyses outline the linkages between networks and policy outcomes, as well as the relationship between broader macrolevel institutions and subsystems, often by using case studies to illustrate connections. They identify differences in policy outcomes and relate them to variations in network structures in increasingly systematic cross-sectoral and cross-national comparisons (Howlett, 2002). In addition, research draws attention to the need to go beyond correlation toward more causal analyses. Studies increasingly try to outline a set of causal mechanisms connecting a set of explanatory variables to the explananda, which can either be policies or the network itself. They seek to identify the processes through which policies or networks came about and to reconstruct them through historical narratives. Thus far only few attempts have been made to integrate mechanisms from different disciplines, such as cognitive psychology (see Hindmoor, 2009), into social network analysis. Thus for instance, Hindmoor (2009) looks at how policy makers' preferences are formed in contexts of uncertainty. By drawing on cognitive psychology literature he identifies 'priming' (convincing people to believe something through releases, articles, interviews) as a mechanism of persuasion whereby social networks operate to influence policy decisions. Empirically, he uses the framework to examine how different policy network structures influenced different responses to foot and mouth disease.

Increasing the Theoretical Leverage of Social Network Analysis

Another stream of research aims to deepen the theoretical and explanatory leverage by linking up network frames with other analytical approaches (see Thatcher, 1998). It suggests that combining network analysis with other established analytical frames means complementing subsystem/mesolevel analysis, with insights from micro- and macrolevel analysis. Drawing on various forms of new institutional as well as rational choice and constructivist analyses social network research attempts to further develop and test hypotheses.

Recent positivistic accounts go beyond earlier interorganizational network analysis by linking statistical-descriptive and inferential analysis to theories of institutional analysis and development (see Lubell et al., 2012; Ward et al., 2011) and collective action (Sandström and Carlsson, 2008). Analyses linking up with collective action frameworks investigate how network structure relates to how actors are connected within a network and how this can influence policy. Social networks are perceived as opportunity structures, which can impose constraints on collective action. As such location and connection of actors within the network have important implications (for example, the size and diversity of a group influences collective action). Studies look at factors that influence the formation of linkages in networks and investigate the impact of networks ties on individual behavior and attitude.

They apply multivariate statistics to test hypotheses about the influence of networks on a variety of explananda, such as network performance. To capture interdependency effects among different actors within social networks research draws on different estimation techniques, including quadratic assignment procedure, exponential random graph models, and stochastic actor-oriented models (for an overview see Lubell et al., 2012; Ward et al., 2011). They utilize them in connection with longitudinal and cross-sectional data for analyzing different aspects of network formation. For example, stochastic models are frequently applied to longitudinal datasets for measuring interdependency effects and change over time (Snijders et al., 2010). Changes are analyzed with reference to earlier observations and studies investigate how individual actors shape social networks through the formation of ties among actors. While stochastic models assume that actors choose ties that maximize their utility, exponential random graph models are frequently combined with other theoretical frameworks. For example, Lee et al. (2012) draw on an institutional collective action framework (see Feiock and Scholz, 2010; Scharpf, 1997), which guides their expectations about the conditions under which actors are expected to cooperate and structures are expected to evolve. In contrast to stochastic models, exponential random graph models rely specifically on cross-sectional data, which offer a snapshot of a network at a particular moment in time. This has the advantage of capturing the whole network dynamics (e.g., network centrality) at a certain point and testing it against hypotheses. Together these inferential models can make predictions about tie formation among actors and the impact of these on other network performances and how network structures mediate actor relationships.

One critical aspect of positivistic analyses is their reliance on complete information. Because social networks are analyzed as structures, researchers need to have a totality of data, which reflects accurately the totality of connections among actors (Ward et al., 2011). Incomplete data affects the reliability of statistical measurement. To address the missing data researchers increasingly integrate so-called 'latent space approaches,' which are based on Bayesian data augmentation to account for unobserved data (Koskinen et al., 2010; Hoff et al., 2002). The calculated likelihood is related to observed patterns. As long as the missing data is random, the application of inferential statistics allows making accurate predictions. Consequently, unobserved systematic patterns have a critical impact on results. Despite recent advancements, even proponents of latent space approaches, most notably Lubell et al. (2012), acknowledge the need for further refinement, specifically with regard to large populations with unknown nodes.

Dialectical approaches were developed as a critique of positivistic social network analysis. In an attempt to go beyond

'simple causal models' and highlight the complexity of the relationship between network and agency Marsh and Smith denote the notion of a dialectical policy network analysis. They specify dialectical as the "interactive relationship between two variables in which each affects the other in a continuing iterative process" (Marsh and Smith, 2000: 5). This relationship refers to structure of network and agents, network and context, and network and policy outcome. Based on that they formulate a set of logic and broad statements – for instance, arguing that external factors as well as agents' decisions affect network formation, that policy outcomes are the results of interaction between agents and structure, that policy outcomes influence networks as well as being influenced by, and that network change is a product of the interaction between the context and the network. However, these statements are not specific enough to provide testable hypotheses and they are also not explicating mechanisms. Dialectical studies use qualitative methods, often focusing on sectoral comparisons within countries across time. Studies in this tradition share a critical realist ontology and maintain that policy making takes place in a multilayered self-organizing subsystem (Marsh and Smith, 2000; Evans, 2001; Toke and Marsh, 2003). Networks are in a constant mode of change because actors within the network interpret context and policy feeds back through network structures. To fully understand policy outcomes one needs to examine the interactive relationships. These models are complex and difficult to test because they attempt to explain the emergence as well as the change of networks as well as the content of policy outcomes.

Decentered analyses go further by adopting an interpretative perspective. By drawing on an antifoundational perspective (Bevir, 2003) decentered approaches apply a constructivist anthropological perspective, which emphasizes the importance of discourse, beliefs, tradition, and meanings (Bevir and Richards, 2009). Their primary focus is on network creation, by investigating how networks are created and altered through individual action. These actions are not conditioned by institutional norms but by beliefs, which individuals adopt against the background of traditions and in response to dilemmas. Decentered approaches maintain that in order to understand practices and institutions the researcher needs to understand meaning, beliefs, and preferences of peoples involved (Bevir and Rhodes, 2010; Bevir and Richards, 2009). Decentered approaches aim to investigate people's beliefs and practices by qualitative methods. They are using historical analysis to understand traditions and using ethnographic analysis and tools, such as studying individual behavior in context, unstructured approach, observations, textual analysis, oral histories, biographies, to investigate beliefs and practices. The ambition of decentered approaches is to "[turn] current approaches to networks on their head" (Bevir and Richards, 2009: 8)

Located between the extreme position of positivistic inferential and interpretative-decentered approaches, the advocacy coalition framework is an alternative partner for social network analysis (Thatcher, 1998). According to the advocacy coalition framework bounded rational actors are expected to rely on policy core beliefs and form coalitions with other actors to pursue common objectives. Policy outcomes and policy changes reflect belief systems of actors within subsystems. The unit of analysis is the subsystem, which can overlap with other subsystems and can be nested in other macro systems. Subsystems can be linked to policy change through policy learning (i.e., changes in beliefs), external subsystem events; internal events; cross-coalition learning. Advocacy coalition frames focusing on the role of ideas to explain policy outcomes and drawing on concept of advocacy coalitions. Based on their belief systems, actors can be aggregated into coalitions. Similar belief systems reduce coordination and transaction costs among actors.

Advocacy coalition frameworks apply quantitative and qualitative methods in order to explain internal logic of subsystems and policy change/content. Studies analyzing policy change and content apply qualitative methods to outline the mechanisms and processes whereby network characteristics affect policies (see Nohrstedt, 2011; Albright, 2011). They explain how and why advocacy coalitions affect policy changes through processes that link external events to coalition responses and policy outcomes (Albright, 2011; Nohrstedt, 2011). By drawing on cross-temporal comparisons, these studies refine existing advocacy coalition frames with regard to causal mechanisms.

Analyses investigating the internal logic of social networks and explaining advocacy coalitions are primarily quantitative (Matti and Sandström, 2011; Ingold, 2011). They focus on how individual beliefs shape policy networks and how external events are mediated through advocacy coalitions. Similar to other positivistic approaches, quantitative advocacy coalition analyses rely on statistical tools. They identify coalitions in policy areas by using block models (Imgold, 2011), and linking coordination and beliefs within policy areas by utilizing quadratic assignment procedure (Matti and Sandström, 2011). Responding to earlier criticism, most notably Schlager (1995), recent advocacy coalition accounts integrate aspects of collective action and testing whether similar beliefs lead to similar coordinated action (Weible and Sabatier, 2005).

Conclusion: Constant Expansion

Thus analyses of networks have become increasingly broad: from modest origins as concepts of issue networks and iron triangles, they have expanded to overarching frameworks of relations between public officials and private interests.

Differences in epistemological and ontological traditions among diverse theoretical lines of inquiry have regularly led to differences and sharp debates (see Marsh and Smith, 2000; Dowding, 1995, 2001). While there are those, most notably Marsh and Smith (2001) and Bevir and Richards (2009), who call for more constructivist perspectives in social network analysis by using anthropological and historical approaches, others, such as Dowding (1995, 2001) and van Bueren et al. (2003), suggest a positivistic empiricist approach by developing network analysis with rational choice approaches, such as game theory or actor-centered institutional analysis. Hence recent developments in network analysis offer a powerful example of the split between highly positivist analyses using quantitative and formal techniques, and interpretivist and constructivist studies that increasingly promote ethnographic studies.

See also: Interest Groups, History of; Interorganizational Relationships and Networks; Network Analysis, History of; Network Analysis; Network Centrality, Measures of; Policy Networks; Social Network Analysis.

Bibliography

Albright, E.A., 2011. Policy change and learning in response to extreme flood events in Hungary: an advocacy coalition approach. The Policy Studies Journal 39, 485–511.

Atkinson, M.C., Coleman, W.D., 1989. Strong states and weak states. British Journal of Political Science 19, 46–67.

Bevir, M., 2003. A Decentred Theory of Governance. In: Bang, H. (Ed.), Governance as Social and Political Communication. Manchester University Press, Manchester.

Bevir, M., Richards, D., 2009. Decentring policy networks: a theoretical agenda. Public Administration 87, 1–14.

Bevir, M., Rhodes, R., 2010. The State As Cultural Practice. Oxford University Press, Oxford.

Bueren, E.M., et al., 2003. Dealing with wicked problems in networks: analyzing an environmental debate from a network perspective. Journal of Public Administration Research and Theory 13, 193–212.

Daugbjerg, C., 1999. Reforming the CAP: policy networks and broader institutional structures. Journal of Common Market Studies 37, 407–428.

Daugbjerg, C., Pedersen, A.B., 2004. New policy ideas and old policy networks: implementing green taxation in Scandinavia. Journal of Public Policy 24, 219–249.

Dowding, K., 1995. Model or metaphor? A critical review of the policy network approach. Political Studies XLIII, 136–158.

Dowding, K., 2001. There Must Be End to Confusion: Policy Networks, Intellectual Fatigue, and the Need for Political Science Methods Courses in British Universities. Political Studies 49, 89–105.

Emirbayer, M., Goodwin, J., 1994. Network analysis, culture, and the problem of agency. American Journal of Sociology 99, 1411–1454.

Evans, M., 2001. Understanding dialectics in policy network analysis. Political Studies 49, 542–550.

Fawcett, P., Daugbjerg, C., 2012. Explaining governance outcomes: epistemology network governance and policy network analysis. Political Studies Review 10, 195–207.

Feiock, R., Scholz, J., 2010. Self-Organizing Federalism: Collaborative Mechanisms to Mitigate Institutional Collective Action Dilemmas. Cambridge University Press, New York.

Heclo, H., 1978. Issue networks and the executive establishment. In: King, A. (Ed.), The New American Political System. AEI, Washington.

Hindmoor, A., 2009. Explaining networks through mechanisms: vaccination, priming and the 2001 foot and mouth disease crisis. Political Studies 57, 75–94.

Hoff, P., et al., 2002. Latent space approaches to social network analysis. Journal of the American Statistical Association 97, 1090–1098.

Howlett, M., 2002. Do networks matter? Linking policy network structure to policy outcomes: evidence from four Canadian policy sectors 1990–2000. Canadian Journal of Political Science 35, 235–267.

Ingold, K., 2011. Network structures within policy processes: coalitions, power, and brokerage in Swiss climate policy. The Policy Studies Journal 39, 435–459.

Jordan, G., 1990. Sub-governments, policy communities and networks: refilling the old bottles? Journal Theoretical Politics 2, 319–338.

Jordan, G., Schubert, K., 1992. A preliminary ordering of policy network labels. European Journal Political Research 21, 7–27.

Knoke, D., Pappi, F.U., Broadbent, J., Tsujinaka, Y., 1996. Comparing Policy Networks: Labour Politics in the U.S., Germany, and Japan. Cambridge University Press, Cambridge.

Koskinen, J., et al., 2010. Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. Statistical Methodology 7, 366–384.

Lee, Y., Lee, W.I., Feiock, R.C., 2012. Interorganizational collaboration networks in economic development policy: an exponential random graph model analysis. The Policy Studies Journal 40, 547–573.

Lubell, M., et al., 2012. Testing policy theory with statistical models of networks. The Policy Studies Journal 40, 351–374.

Marin, B., Mayntz, R., 1991. Policy Networks: Empirical Evidence and Theoretical Considerations. Campus Verlag, Frankfurt.

Marsh, D., Smith, M., 2000. Understanding policy networks: towards a dialectical approach. Political Studies 48, 4–21.

Marsh, D., Smith, M., 2001. There is More than One Way to Do Political Science: on Different Ways to Study Policy Networks. Political Studies 49, 528–541.

Matti, S., Sandström, A., 2011. The rationale determining advocacy coalitions: examining coordination networks and corresponding beliefs. The Policy Studies Journal 39, 385–410.

Nohrstedt, D., Weible, M.C., 2010. The logic of policy change after crisis: proximity and subsystem interaction. Policy Studies Organization 1, 1–31.

Nohrstedt, D., 2011. Shifting resources and venues producing policy change in contested subsystems: a case study of Swedish signals intelligence policy. The Policy Studies Journal 39, 461–484.

Peterson, J., 2003. Policy networks. Institute for Advanced Studies. Political Science Series 90, 1–29.

Richardson, J., Jordan, G., 1979. Governing under Pressure: The Policy Process in a Post-Parliamentary Democracy. Martin Robertson, Oxford.

Richardson, J., 2000. Government, interest groups and policy change. Political Studies 48, 1006–1025.

Rhodes, R.A.W., Marsh, D., 1992. Policy networks in British politics: a critique of existing approaches. In: Marsh, D., Rhodes, R.A.W. (Eds.), Policy Networks in British Government. Clarendon Press, Oxford.

Sandström, A., Carlsson, L., 2008. The performance of policy networks: the relation between network structure and network performance. The Policy Studies Journal 36, 497–524.

Scharpf, F., 1997. Games Real Actors Play: Actor-Centered Institutionalism in Policy Research. Westview Press, Boulder.

Schlager, E., 1995. Policy making and collective action: defining coalitions within the advocacy coalition framework. Policy Sciences 28, 243–270.

Scholz, J.T., et al., 2008. Do networks solve collective action problems? Credibility, search, and collaboration. The Journal of Politics 70, 393–406.

Snijders, G., et al., 2010. Introduction to stochastic actor-based models for network dynamics. Social Networks 32, 44–66.

Thatcher, M., 1998. The development of policy network analyses. Journal of Theoretical Politics 10, 389–416.

Toke, D., Marsh, D., 2003. Policy networks and the GM crops issue: assessing the utility of a dialectical model of policy networks. Public Administration 81, 229–251.

Ward, M., et al., 2011. Network analysis and political science. Annual Review Political Science 14, 245–264.

Weible, M.C., et al., 2011. A quarter century of the advocacy coalition framework: an introduction to the special issue. The Policy Studies Journal 39, 349–360.

Weible, C., Sabatier, P., 2005. Comparing policy networks: marine protected areas in California. The Policy Studies Journal 33, 181–202.

J

Jackson, John Hughlings (1835–1911)

W Ian McDonald[†], Royal College of Physicians, London, UK
Hans-Joachim Freund, Heinrich Heine University, Dusseldorf, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 7943–7945, © 2001, Elsevier Ltd.

Abstract

John Hughlings Jackson (1835–1911) was arguably the most influential thinker in the history of neurology. He elaborated the localisation of function in the human brain on the basis of clinical observations in patients with brain lesions and epileptic seizures. His work induced the first experimental studies that further established the structure–function relationships in the brain. He also laid the ground for our understanding of the nature of epileptic discharges.

John Hughlings Jackson was born in Providence Green, Yorkshire, England on 4 April in 1835, of English antecedants on his father's side and Welsh on his mother's (Critchley and Critchley, 1998). His father was a gentleman farmer and a brewer who, as a surviving letter shows, was a wise and kindly man. Jackson's mother died aged 29 when he was one year old. He was educated at an undistinguished boarding school which he left at the age of 15, and did not go to university. In later years 'he held that by exemption from over-teaching his mind had retained more of freedom and energy than might otherwise have been the case' (Hutchinson, 1925).

When Jackson left school he was at once apprenticed to a general practitioner in York. Two years later, in 1852, he enrolled at the York Medical School and in 1855 he proceeded to St. Bartholomew's Hospital in London. He qualified in 1856 and returned to the York Dispensary for three years as a resident medical officer. In 1860 he submitted a thesis (the title of which is unrecorded) to the University of St. Andrews and was awarded the MD.

By this time Jackson had returned to London. He had become increasingly interested in philosophy and seriously considered abandoning medicine for this field. He was dissuaded by his friend Jonathan Hutchinson (who later became one of the leading surgeons and ophthalmologists of his day). About this time (Hutchinson records) Jackson came under the influence of C E Brown-Sequard who 'told him strongly, that it was foolish to waste his efforts in wide observation of disease in general, and that if he wished to attain anything he must keep to the nervous system' (Hutchinson, 1925). As a result, Jackson was appointed in 1862 to the staff of Brown-Sequard's hospital, the National Hospital for the Paralysed and Epileptic in Queen Square. In the same year he became a clinical assistant at the Royal London Ophthalmic Hospital (now known as the Morefields Eye Hospital) in London and in the following year he was appointed to the London Hospital. These three institutions provided him with the experience on the basis of which he made his fundamental contributions to our understanding of the organization of nervous function.

Jackson was a shy, modest, self-effacing individual, much loved by his students and colleagues. His wife died after only eleven years of marriage. A place at the lunch table was always set for him at the home of his colleague Dr. Thomas Buzzard. Jackson came to lunch irregularly and without notice. As Buzzard's son recalled, '... he would pull a chair up to the table, exchange a few remarks, perhaps tell a humorous story with only a twinkle of the eye disturbing the solemnity of his features, and after five, perhaps ten minutes he would mumble something about remembering an engagement and leave the room' (Buzzard, 1934). In his last years Jackson lived an even more solitary, though serene life. He retained an interest in the lives of his pupils and their families to the end. He died on the 7 October 1911. His contribution was widely acknowledged by the academic world during his lifetime, but to the surprise of all he received no national honor.

Jackson's contemporaries and pupils recorded that he was a meticulous observer of clinical phenomena, and from his earliest days in medicine he was committed to the necessity for pathological verification of clinical inferences. His understanding of the nervous system based on this approach was molded by two main influences. The first derived from his teacher at medical school, Dr. Thomas Laycock (later Professor of Medicine in Edinburgh) who was interested in psychology and in nervous function and dysfunction. The other influence was the writings of the philosopher Herbert Spencer with whom Jackson corresponded up to the beginning of the twentieth century. It was from Spencer that Jackson derived the notion of the evolution of nervous function, which led to his idea of a hierarchical arrangement of functions, and that disease (neurological and psychiatric) led to a successive regression from the 'least automatic' and 'least deeply organised' to the 'most automatic' and 'most deeply organised' levels. As Walshe (1961) has pointed out, these ideas derived from Jackson's lifelong study of five kinds of clinical phenomena: focal seizures, chorea, hemiplegia, speech disorders, and mental disorders.

A remarkable consequence of Jackson's work was his prediction, on the basis of unaided clinical observation, of

[†]Deceased.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61058-9

a number of physiological principles which were later confirmed experimentally. One of them is the concept of localized cerebral discharges generating focal seizures, and that such focal seizures are caused by local brain disease. The comparison between the clinical symptomatology and subsequent neuropathology showed that the distribution of focal motor and sensory signs depended on the site of the lesion. The topographical structure–function relationships laid the ground for the topological functional organization of the cortex long before cytoarchitectonic anatomical maps were established. The general idea of the functional specialization of the cortex had already been proposed by Gall (1825). His phrenology was derived from the attribution of certain personality features to the individual 'microstructure' of the scalp as inferred from palpation. Gall's postulate that cerebral function must be localized was correct in principle but based on inadequate evidence. The first real evidence came from Jackson, who published his first account of epilepsy in 1861, almost 10 years before his clinically based concept was confirmed by using electrical stimulation of the cortex (Fritsch and Hitzig, 1870; Ferrier, 1876).

The first pioneers of cortical neurophysiology indeed recognized the pivotal role of Jackson in the development of the concept of epileptogenic discharges in distinct areas generating focal seizures as the following two quotations show: '... I was precisely the first one, now some thirty or more years ago, to confirm by *physiological experiment* and to define more closely that which Hughlings Jackson had concluded from clinical facts' (Hitzig, 1900); '... Dr. Hughlings Jackson who from a clinical and pathological standpoint anticipated many of the more important results of recent experimental investigation into the functions of the cerebral hemispheres. (Ferrier, 1886).

Another concept put forward by Jackson related to the effect of lesions on function. In order to understand the pathophysiology of the effects of lesions he introduced the term 'negative and positive elements.' The negative elements were derived from the more global effects of epileptic seizures as apparent from dreamy states and the loss of consciousness. This was regarded as a dissolution 'of so many nervous elements of the highest centres' by the epileptic discharge, so that the 'most special and complex sensorimotor arrangements' would affect the adjustment of the organism as a whole to the environment. Convulsions were recognized as due to excessive neuronal discharge. From his clinical observations Jackson deduced that 'there are two ways in which nerve tissue suffers. It may be destroyed, and then there is loss of function. It may be unstable, and then there is disorder of function – discharge' (Jackson, 1931, vol. 1, p. 29).

The 'positive element' was exemplified by speech disturbances, where the inability to express oneself demonstrates the negative effect of a brain lesion, whereas the paraphasia – the insertion of wrong words or inadequate sentence formations – would show the effect of the damaged part on the function of the remaining neural circuitry. These views later inspired Sigmund Freud's contributions to aphasia, in particular his early insight into the emotional determinants of speech disruptions. The positive and negative elements also had an impact on subsequent hypotheses on psychotic disorders.

Jackson's explanation of the clinical deficits associated with lesions was based on his view of the three levels in the evolution of the nervous system. Each element of each level was seen to contain a complete representation of the next lower level. Consequently, each element of the middle and highest levels would contain a representation of the entire body. If the nervous system is damaged the clinical impairment is not due just to the damaged part, but represents the compensatory action of the lower level circuitry that would then partially compensate for the function of the destroyed tissue.

The concept of evolution of the nervous system was therefore instrumental to his view that dissolution of the nervous system due to the effects of lesions results from the function of the remaining, nondamaged parts being liberated from control by the damaged part. This approach to understanding the pathophysiology and nature of the neurological signs and symptoms was later replaced by 'localizationists' who considered these to be deficits in a mechanical framework, i.e., saw them as loss of function of the destroyed tissue. It soon became clear that phenomena like spasticity, tremor etc., could not be explained in this way. With the advent of new, experimentally-based conceptions of the brain as a complex, densely interconnected and reentrant network, the Jacksonian views were basically confirmed.

Jackson's contribution was not limited to the conceptual advances we have described. His skill in precise clinical observation, together with his analytical and deductive powers, led to advances of practical value in clinical neurology. For example, he was one of those who pioneered the introduction of the Helmholtz ophthalmoscope to medicine, his first paper on the topic appearing in 1863. His papers of 1871 and 1873 on 'optic neuritis' from intracranial disease, and the diagnosis of brain tumours, represented a major advance. He made contributions to neuro-ophthalmology, particularly in relation to eye movement, and to the effects of syphilis on the nervous system.

The writings of Jackson have a reputation for being difficult to read, being filled as they are with qualifications in the text and in extensive footnotes, as he strove to get as close as possible to the truth about inherently complex problems; they are still illuminating. Many appeared in *Brain* (of which he was a co-founder) but many others appeared in journals which are now difficult to obtain. Fortunately a comprehensive selection was published in 1931 and reprinted in 1958.

See also: Brain, Evolution of; Cerebral Cortex; Epilepsy; Long-Term Potentiation and Long-Term Depression.

Bibliography

Buzzard, E.F., 1934. Hughlings Jackson and his influence on neurology. Lancet 2, 909–913.

Critchley, M., Critchley, E.A., 1998. John Hughlings Jackson. Father of English Neurology. Oxford University Press, New York, p. 228.

Ferrier, D., 1876. The Function of the Brain. Smith, Elder, London.

Ferrier, D., 1886. The Function of the Brain, second ed. Putnam, New York, p. 498.

Fritsch, G.T., Hitzig, E., 1870. Uber die elektrische Erregbarkeit des Großhirns. Archiv fur Anatomie und Physiologie, 300–332. (Engl. Trans. Wilkins H 1963 Neurosurgical classics. *Journal Neurosurgery* 20: 904–16).

Gall, F.J., 1825. Fonction du Cerveau. Balliere, Paris.

Hitzig, E., 1900. Hughlings Jackson and the cortical motor centres in the light of physiological research. Brain 23, 545–581.

Hutchinson, J., 1925. Recollections of a lifelong friendship. In: Taylor, J. (Ed.), Neurological Fragments. Oxford University Press, Oxford, UK.

Jackson, J.H., 1925. Neurological Fragments. In: Taylor, J. (Ed.). Oxford University Press, Oxford, UK.

Jackson, J.H., 1931. Selected Writings of John Hughlings Jackson. Hodder and Stoughton, London.

Spurzheim, G., 1825. Phrenology in Connection with the Study of Physiogonomy. Part I. Characters. Treuttel, Wurtzand Richter, London.

Walshe, F.M.R., 1961. Contributions of John Hughlings Jackson to neurology. Archives of Neurology 5, 119–131.

Jacobs, Jane (1916–2006)

Mary P Corcoran, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Jane Jacobs' work has been influential in a broad and disparate range of disciplines including economics, sociology, psychiatry, journalism, playwriting, social activism, politics, urban planning, biology, and architecture. She is primarily remembered as a public intellectual, an activist, a champion of all things urban, and an irascible opponent of urban renewal models. She offered a particular diagnosis of what was wrong with the modern city, and a blueprint for urban recovery. While she has been criticized for her somewhat utopian portraiture of how city life might be, many of her practical recommendations for reimagining city space are integral to urban policy and planning practice today.

Introduction

Jane Jacobs' extraordinary intellectual legacy has lead to her being variously described as a visionary, urban philosopher, urban sociologist, urban economist, and urban designer (Hirt and Zahm, 2012). Her work has been influential in a broad and disparate range of disciplines including economics, sociology, psychiatry, journalism, playwriting, social activism, politics, urban planning, biology, and architecture. She is primarily remembered as a public intellectual, an activist, a champion of all things urban, and an irascible opponent of urban renewal models. She offered a particular diagnosis of what was wrong with cities, and a blueprint for their recovery. Many of her urban 'tactics' are an accepted part of urban policy and planning practice today. In sum, Jacobs was "an astute chronicler of urban life, she squarely addressed issues, eloquently showcased what worked, and advanced proposals for improving our cities" (Ellin, 2010: p. 45).

An autodidact, Jane Jacobs' incisive exploration of the modern urban form was largely based on her own self-directed research and learning. Her *modus operandi* might be termed serendipitous exploration. In correspondence with Richard Carroll Keeley in the late 1980s, she wrote about her approach to her work: "I read as omnivorously as I can manage, in anything that interests me. Seek patterns, explore them and look out for contradictions … messy, muddly, work which I am inclined to think that thinking maybe usually is" (Keeley, 1989a: p. 35). Jacobs practised the discipline of dialectical thinking and questioning, but her frame of reference was never abstract or abstruse. Rather, her life's work as a student of the city was grounded in the materiality of everyday urban life. She had a unique ability "to move from the particularities of the local interaction of incident and morphology to the larger patterns of urban life" (Sorokin, 2010: p. xvi). Inevitably, her excoriation of twentieth-century urban planning particularly in North America but also beyond, drew criticism from establishment figures within academe and architectural practice. Jacobs was dismissed as lacking both systematic disciplinary training and a grounding in theoretical knowledge. In fact, Jacobs' approach to her metier was to marry rigor and intuition, cool observation and empathy, historicity and an abiding and unerring sense of the human subject.

Biographical Facts

Jane Jacobs was born in 1916 in Scranton, Pennsylvania, and died in Toronto in 2006. Her father was a physician and her mother a former school teacher and nurse. She was raised a Presbyterian. After completing High School where she was a self-confessed 'trouble-maker,' Jacobs worked for a short time in her home city before moving with her sister to New York. Here Jacobs worked in a variety of secretarial and writing jobs which afforded her the opportunity to savor the geographical and sociological variety of the city. These experiences, she claims, "… gave me more of a notion of what was going on in the city and what business was like, what work was like" (www.pps.org/reference/jjacobs-2). She met a young architect Robert Jacobs during this period and he became her husband and lifelong comrade in arms. He died in 1996.

Jacobs began to develop a critical perspective on city planning and urban renewal during her period of tenure as associate editor of *Architectural Forum*, where she worked throughout the 1950s. After hearing her speak publicly about the threats posed by current urban planning policies to urban living, the distinguished urbanist and author William H. Whyte invited her to write an article in *Fortune* magazine, titled 'Downtown is for People.' On foot of this article Jacobs won a Rockefeller foundation grant to embark on her first and most well-known book, *The Death and Life of Great American Cities* (1961), "a prophetic text" (Suarez, 2010: p. 19) which would change her life, and influence generations of urban theorists and urban practitioners.

In the 1960s Jacobs took a very public role in opposing the urban renewal policies in New York City, favored by the city's Parks Commissioner Robert Moses. Her contribution to campaigns to stop building highways through historical neighborhoods and urban villages contributed to the defeat of the Lower Manhattan Expressway plan and the eventual fall from power of Moses. Furthermore, "Jacobs' harsh criticism of 'slum-clearing' and high-rise housing projects was also instrumental in discrediting these once universally supported planning practices" (www.pps.org/reference/jjacobs-2).

The Jacobs family left New York City in 1968 because of their opposition to the Vietnam War. Jacobs found the culture of Toronto very different to that of New York City. She believed that immigrants had more capacity to do things in the city than

she had observed in her native New York. She recalled that in the late 1960s there was little evidence that city precincts had been subjected to redlining (as had been the case in New York) and that ethnic populations were not ghettoized. Nevertheless in her adopted home she remained a critic of top–down city planning, and blockbuster developments which threatened the integrity of the city's neighborhoods.

Contemporaries recall that while she was best known in New York City for the struggles which she fronted in the 1960s, for Torontonians she became a significant and accessible mentor for those concerned about the development of the city (Greenberg, 2010). She took Canadian citizenship in 1974 and lived in Toronto until her death. Her qualities as an activist were legend. The New York Times obituary on her death referred to the 'steeliness beneath her folksiness' and quoted the Canadian weekly magazine, MacLean's which had feted her as 'a lioness in winter' (*The New York Times*, 26 April 2006). Jacobs remained a social thinker and activist with her feet very firmly planted on the ground. As Greenberg has observed "she refused to be deified or reified: she never wanted to be a guru and strenuously resisted the temptation to freeze her thinking into cult or dogma … and resolutely rejected being franchised" (2010: p. 117). Nonetheless, streets, walks, awards, medals, and fellowships are named after her.

Description of the Major Works and Publications

Jane Jacobs brought depth and complexity to how we view the city. She explored the city as a distinctive social space and economic space, and tracked the myriad ways in which the built environment, economic relations, and social viability of the city are interpolated. Hers is a powerful articulation of the interactive order of everyday city life that is frequently eclipsed by the focus on the competitive entrepreneurial city and the proliferation of civic boosterism. Moreover, Jacobs was an early critic of low-density, single use settlement patterns that have a deleterious impact on the environment (Kidder, 2012). Her approach followed the template of grounded theory, using empirical observation in the real world to generate sensitizing concepts that could be deployed as starting points for new research (Glaser and Strauss, 1967). Jacobs opposed abstraction and wrote in an accessible style. Her virtuosity resided in her ability to marry historical and comparative perspectives with case study analysis and personal observation.

Jacobs's first and most feted book *The Death and Life of Great American Cities* (1961) made an enormous impact on its publication which resonates to the present day. This work is the basis of her core reputation as a defender of cities, and of city streets. Indeed, it has been described as "a long paean to the humanizing, secure, child-friendly, and economically coherent nature of the urban street" (Suarez, 2010: p. 16). The book literally shook the planning and architectural establishment to its core, and it remains prescribed reading for each new generation of architecture, planning, and social science students.

For Jacobs, cities are an ongoing work accomplished by those who live, work, and socialize in them. Diversity, intensity, and creativity are the sources of their economic potential. In the tradition of the Chicago School and the ecological explorations pioneered by Robert Park in the 1920s, Jacobs approached the city as a laboratory of organized complexity to be explored literally from the ground up. She utilized a social processes framework by paying attention simultaneously to the economic, social, cultural, political, and regulatory forces that constitute the urban.

In *The Death and Life of Great American Cities* Jacobs advanced the view that a gradualist, evolutionary approach to urban development was preferable in terms of maintaining urban vitality to an intensive, deterministic approach then favored by urban planners. Jacobs vigorously opposed 'one size fits all' planning approaches. Rather she emphasized the benefit of "allowing places to evolve gradually through natural selection and infill development, with minimal government intervention and in careful consultation with local residents" (King, 2013: p. 2408). This view of the city as the outcome of ecological forces has been heavily critiqued from a political economy perspective, and from a planning perspective can be read as almost heretical. Urban planning in Western cities today proceeds through large-scale, public–private redevelopments, neighborhood-level regeneration programmes, and cultural projects designed primarily with the interests of capital accumulation in mind. They seek to appeal not to the local citizenry but to prospective gentrifiers and tourists. Yet, it is arguable that in two respects Jacobs' creed has been incorporated into planning practice: there is nowadays a greater respect for the urban vernacular and an acknowledgment that planners should take account of the lived experience of urbanites in planning and urban design.

The Death and Life of Great American Cities focused on four key dimensions of urban life which are summarized here: (1) The city is the natural progenitor of interactive human life. The nature of cities is explored by examining its public realm, sidewalks, parks, and neighborhoods, all of which Jacobs viewed as crucial sites of civic interface. Among her key theses that remains highly pertinent is the notion that collective security can be maintained through having low-level collective surveillance (eyes on the street) in vibrant public spaces. (2) The well-being and liveability of a city are connected to level of diversity. Jacobs understood diversity not so much in relation to population make up, but more in terms of different kinds of economic functions, the mixing of diverse cultural groupings and tolerance of diverse cultural practices. This is a perennial argument made about cities and has been repopularized in more recent times by writers such as Ray Oldenburg and Richard Florida. Her recipe for diversity included multifunctionality districts or neighborhoods, shorter city blocks that are more people friendly, architectural mingling of old and new, and greater density of urban populations. Diversity, Jacobs believed, affords opportunity for encounters with difference and with strangeness, and such encounters incite question, wonder, and discovery (Keeley, 1989b). Multifunctionality in neighborhoods creates the basis for diversity in urban life. (3) Cities are involved in processes of creative destruction through the forces of decline and regeneration, or what she termed 'slumming and unslumming.' She presciently argued that too much homogeneity or – as she termed it duplication of function – militates against diversity. Homogenization results in a loss of appeal, which in turn contributes to a spiral of decline resulting in urban stagnation

and decay. Interestingly, Jacobs references The City area of Central London as an example of dediversification, and predicted (rather accurately as it turns out) that eventually the City area and the adjacent West End "will become one dreary sea of office blocks" (1992: p. 149). (4) The final section of the book deals with tactics for restoring cities. Most of these tactics are now commonly deployed in urban planning and project development but they were highly innovative in the early 1960s. These include providing integrated social housing, limiting car usage in the city, revisualizing the city and city space through the application of design principles, opening up and preserving the public realm, and refashioning city governance.

The book is replete with examples and instances of the 'common good' which Jacobs sees in the lively neighborhood which acts as both a focus for place attachment and a space for civil interactions where people of the neighborhood, and those visiting or passing through can easily comingle. Jacobs believed that in cities that lacked a natural and casual public life, people would become socially isolated, they would "lack public acquaintanceship … and no practice or ease in applying the most ordinary techniques of city public life at lowly levels" (1992: p. 65). She contended that the public realm of the city must be open, accessible, and traversed by all. She recognized the significant role played by public characters, people who act as lynchpins within informal community networks, ensuring the transmission of information within those networks. Jacobs went on to identify another important constituency – the hop-skip people – whose role is to make links across neighborhoods, creating a summative district neighborhood. Though she does not use the terminology, Jacobs inadvertently subscribed to the now commonplace view that loose ties with others can forge bridging social capital, which complements the denser forms of social capital associated with communities where networks are intimately enmeshed. Hop-skip people are the civic minded, who move across and between networks and are the facilitators of genuine self-governance (Byrne, 1989). Such self-governance in the cause of achieving the common good can only emerge if there is sufficient economic prosperity in the city and the complexion is sufficiently diverse. In the book she paints an imagined urban community (utopia) constituted through diversity and dynamic interdependence. Moreover, she makes a key distinction between the goals of efficiency and creativity in the urban context. According to Jacobs, the two are not mutually exclusive but complementary: giving full rein to creativity heightens efficiency in the long run (Cichello, 1989).

Jacobs' work, though highly original, has many antecedents from within the social sciences. There is a clear debt to the Chicago School and their ethnographic observations of everyday life in the city. Her work also grappled with the modernist conundrum presented by the city, a theme that featured heavily in Georg Simmel's oeuvre. The latter was struck by the creative imperatives borne of the intensity, diversity and tempo of city living, and the struggle to retain subjectivity in the modern metropolis.

Like many classics, *The Death and Life of Great American Cities* generated controversy when first published but over time this has given way to an acknowledgment of the timeless insights which it contains. In the twenty-first century many of Jacobs's observations and prescriptions have become integral to how cities see themselves, and seek to be seen by others. Jacobs argued that grand planning schemes intending to redevelop large swaths of a city according to a central theoretical framework fail because planners do not understand that healthy cities are organic, spontaneous, messy, complex systems that result from evolutionary processes. The less master planning to which a city is subject, the more potential for an organic outcome, she believed. As King observes her framework of analysis suggests that neighborhoods that have experienced gradual rather than intensive redevelopment would experience better allocation of land uses, perhaps resulting in more walking, less vacant area, and a greater diversity of uses (King, 2013: p. 2409). Cities concerned with liveability nowadays are turning increasingly toward Jacobian principles focusing on creating more walkways and cycle paths through city streets, reducing the impacts of car culture, promoting 'living over the shop' schemes, developing the public realm of city centers and renewing high streets.

After the resounding success of *The Death and Life of Great American Cities*, Jane Jacobs took a more philosophical turn broadening her agenda to include discussions of the city in its economic context, and to explore the moral basis of social and economic relations. Subsequent books include *The Economy of Cities* (1970); *The Question of Separatism* (1980), an analysis of the case for Quebec sovereignty; *Cities and the Wealth of Nations* (1984), a prescient study of the urban city region as a counterpoint to the nation-state; *Systems of Survival* (1993), an exploration of the morality of commerce and politics; *The Nature of Economies* (2000), which employs an ecological frame of analysis to interpret the economic system, and *Dark Age Ahead* (2004), a work which explored how suburban sprawl might be tackled through more creative and intensive usage of land. Once again, Jacobs proved to be ahead of the curve given her concerns with making suburbia more sustainable. She extended her urban brief to the suburbs envisioning roadways being converted into urban streets and arterial routes transformed into boulevards that offered equal access to transit and cycle lanes. However, she suggested that such changes would only be possible on foot of a catalystic event such as an economic crisis (Greenberg, 2010). It is arguable that in Western cities at least, the effects of climate change coupled with economic crisis are leading toward a more pronounced focus on creating more sustainable cities and suburbs.

Jacobs' later works represented a sustained attempt to understand cities within the broader context of the political economy of the region and the state. With considerable forethought she anticipated the potential of cities as economic engines or incubators. In *The Economy of Cities* she analyses how cities come into being, and what they must do economically to flourish. The city she observes is "a settlement that consistently generates its own economic growth from its own local economy" (Jacobs, 1970: p. 258). Growth is envisioned in terms of human innovation and creativity through a process of discovery of new forms of work, through application of new knowledge, or recombinations of knowledge. The rise of the now ubiquitous creative quarters, digital hubs, and silicon alleys all bear powerful witness to her central thesis. Nevertheless, these economic growth engines remain niche rather than mainstream, and cities – even in the most advanced

societies – also continue to house sweet shops, to sustain informal economies, and to generate minimum wage service jobs.

Cities, Jacobs argued, operate through a reciprocating economic system which serves to catalyze relationships between cities. [In fact cities more often compete with each other for inward investment on an increasingly globalized stage.] Jacobs advocated a system of import substitution in cities as well as export development. From her perspective, export generation is the process by which cities increase the amount and range of goods and services that they trade (Cichello, 1989). This she believed would create a virtuous cycle or multiplier effect. Import replacement, on the other hand, can occur when the dynamic growth and innovation in the city make it possible to start producing locally what previously had to be imported. Once again, processes are central to her mode of analysis. The creatively minded entrepreneur is the instigator of this virtuous economic cycle, which she contended could create a sustainable growth model. From today's standpoint this reads as a particularly naive and nostalgic version of benign capitalism, a capitalism that could be locally tamed. But even as she was writing at the end of the 1960s, this version of capitalism was on the way out, and a far more rapacious form of neo-liberal capitalism was being ushered in.

The Economy of Cities explores the economic implications as opposed to the social implications of her liberal democratic model of the city (Kidder, 2012). She emphasized freedom and autonomy in relation to the economic activity of cities. Here she anticipated the growing strength of the urban regime, which to some degree challenges the nation-state nowadays as a site of economic activity through its own civic boosterism. She anticipated the emergence of the city and city regions as constituting new sites of power and political practice at subnational level.

A moral distinction is made in *The Economy of Cities* between commercial values (dominated by relations of commercial exchange) on the one hand and guardian values (adherence to a collective responsibility and orientation toward a shared common good) on the other. Her innate feel for sociology is evidenced here in her Weberian juxtapositioning of an ethic of rationality oriented toward means and growing out of a calculative frame of mind, with an ethic of ends which is much more concerned with the value systems that underpin our actions and engagements with others.

Cities and the Wealth of Nations (1984) has been described as her most ambitious project but also most pessimistic book (Keeley, 1989b). In this book, Jacobs built upon the microeconomic base elaborated in her earlier works, to develop a macroeconomic analysis that frames a theory of wealth, trade, growth, and decline of cities both individually and collectively (Cichello, 1989). She critiqued macroeconomic theory taking issue with the many anomalies that are thrown up by conventional economic theory predicated on a very narrow conceptualization of human behavior. (Once again, her critique was prescient. More than 50 years later the rise of behavioral economics points to a central concern of her life's work, namely, that context heavily affects the way we make decisions and take action.) The book which provides a critical analysis of economic development theory and offers a populist reading of global political economy develops further themes implicit in *The Economy of Cities*. The city is placed in the context of its city region or hinterland. Jacobs traced the stultifying effect on the economy of the displacement of the innovator by the bureaucratic corporation. Against this backdrop, she assessed the viability of the city-state. According to Jacobs, tradition has bound us to a system of nation-states, and the primacy of this economic unit has systematically blocked the innovative and dynamic potential of subnational economic units, cities. In this she anticipates the rise of the modern city-state which directly competes with nation-states for investment (see, e.g., the rise of metropolitan corridors such as the Rhine–Rhur metropolitan region and silicon valley in California). It is a fact today that a rising share of the economic action is urban. Even in the technology sectors, themselves traditionally suburban, capital is following talent to the city (Luce, 2014).

Jacobs as ever was interested in the lived experiences of urban resilience, local generation of wealth, microfinancing, and entrepreneurial development, particularly in the global South. Her writings take on an inevitable sepia hue when viewed from the purview of a digitally driven, hyper globalized world. Still, Jacobs did not shy away from confronting the failures of the economic system of national states, in the process challenging the foundations and premises upon which that system is predicated (Cichello, 1989).

Assessment

Jacobs' work lends itself to reading and rereading. Many of her predictions have been borne out and her analysis remains *au courant* for students of the city, and for those who seek urban transformation. Her ideas remain perennially popular. For instance, her contention that a mix of housing types to include older as well as newer buildings has a positive impact on social relations and the climate of sociability in a neighborhood continues to garner attention. A recent study tested Jacobs' hypothesis that mixed age buildings could foment enhanced social relations. Drawing on data in Chicago on housing stock age and data on rates of social cohesion, control, intergenerational closure, and reciprocal exchange King demonstrated significant links between housing age diversity (historical development pace) and measures of neighborly social relations, even when controlling for other neighborhood housing features, such as social composition and individual socio demographics (King, 2013: p. 2421). Although one cannot infer causality where a correlation holds, King suggests that these cross-sectional results do suggest the validity of Jacobs's argument.

In terms of her impact outside of the academy, her ideas have been taken up both by New Urbanism and Integral Urbanism movements. Integral Urbanism is a movement within urban design that aims "to heal the wounds of the modern and postmodern era by engaging in a range of restorative efforts incorporating five qualities: hybridity, connectivity, porosity, authenticity and vulnerability" (Ellin, 2010: p. 48). At its core this reorientation in urban design places value on process as well as product, something which Jacobs championed throughout her writing and activist career. Jacobs' influence on design thinking is detectable in the preoccupations with sustainable urbanism and the focus on place-making

among academics, policy-makers, and politicians today. Jacobs spent a lifetime defending public space against privatization, both in her scholarly works and in her political activism. Community participation in planning, community gain clauses in development plans, environmental assessment all exist today but are traceable directly or indirectly to the incisive and forensic account of a defective planning system promulgated in *The Death and Life of Great American Cities*.

There is a strong ethical dimension to her work, and particularly so in *The Death and Life of Great American Cities*, *The Economy of Cities*, and *Cities and Wealth of Nations*, which provided powerful arguments about the ethical underpinnings of liberal democracy (Kidder, 2012). She consistently endorsed the twin principles of individual and collective determination in cities. In this her work harks back to that of Simmel who saw cities as both arenas for intense social interaction and for the elaboration of subjectivity in the face of the encroachment of objective culture. Jacobs was concerned with freedom, autonomy, and diversity in the city all of which Simmel had seen as constitutive of the modern metropolis. In her life and her work she sought a balance between the expression of individual freedom and the ordered, rationalized world of the money economy where all is reduced to the question of 'how much?' Such questions intrigued Jane Jacobs and continue to exercise social scientists, architects, planners, and economists today. How do we achieve an equitable balance between "the freedom of the individual and the order of collective responsibility, between large scale organizations and small ones, between public life (we) and private life (I)" (Hamdi, 2010: p. 37).

Her belief that diversity and sociability are essential requirements of cities is reflected directly or indirectly in the work of many contemporary urbanists: the 'strength of weak ties' (Granovetter, 1973); the concept of the 'third place' (Oldenburg, 1989); the creative class thesis (Florida, 2002); the concept of 'civil integration' (Vertovec, 2007); the notion that 'civicism' is the key to identity in the modern world (Bell and de-Shalit, 2011); and the idea that cooperative relations are central to human success (Sennett, 2012), to name just some.

Critics have described her work as representing a Kuhnian paradigm shift in how we think about and theorize cities (Cichello, 1989). In particular, *The Death and Life of Great American Cities* published more than 50 years ago, emphatically threw down the gauntlet to planners, architects, and urban theorists wedded to modernism and urban renewal. Jacobs attempted to reframe the way in which the city should be understood (Keeley, 1989b). As Sorokin astutely observes she had a unique "way of looking that saw urbanity not in the reduction of its order to the regularity of clockwork but in its vivid spontaneity and expansive, ever-transforming difference" (2010: p. xv). Much as John Berger's, *Ways of Seeing* (1972) challenged the hegemony of Western cultural aesthetics by revealing the hidden ideologies of visual images, Jacobs' work challenged the taken-for-granted assumptions about city growth and urban dynamics, revealing the latent properties that make cities lovable and livable. She offered a powerful and compelling critique of the high modernist notion of the city as a machine for living promulgated by Le Corbusier. By contrast, for Jacobs "the living city can achieve a level of complexity, and promote an array of individual ends, that far surpasses any designed construction, whether physical or social, precisely because cities are the emergent, collective expression of many individual minds interacting day after day" (Ikeda, 2010: p. 27). The city for her is fundamentally a set of social processes, which themselves are always in a state of flux.

Jacobs has been criticized for the utopianism latent in her own vision of the city. In particular, her work veers toward an excessively nostalgic vision of diverse and intimate neighborhoods peopled by self-employed artisans, petit bourgeoisie, and urban entrepreneurs. Her vision of communal diversity animated at local level through ongoing interactions between neighbors involved in low-level surveillance of their mutually shared space, can if brought to its logical conclusion lead to a more dystopian version of urban reality-gated communities such as exclusionary communitarian movements which serve to socially polarize the contemporary city (Harvey, 2000). Moreover, Jacobs displays no real interest in, and presents no sustained analysis of issues of power, of class, of gender, of racial and ethnic cleavages all of which are germane to the constitution of city and urban conflict.

Jacobs' contribution lies in her capacity to cut through quantitative metrics and take into account the whole complex range of variables and processes that made up the city and everyday city life. This was no less than a call for a humanistic–interpretive approach to urban analysis as a substitute for the more positivist approach that dominated in the twentieth century, and in particular in the years after World War II. Hers was a clarion cry to the American public and to publics everywhere, an encouragement to citizens to think about processes; to work inductively rather than deductively and to seek the particulars (fragmentary evidence) which can provide more insight into the generalities (statistics and metadata sets) of how things operate (Jacobs, 1992[1961]: p. 440). She proved to be a formidable leader in charting this course, in the process creating a legacy that persists to the present day.

See also: Just City; Place in Geography; Planning Ethics; Right to the City: A Liberal-Democratic Perspective; Social Ecology; Spatial Thinking in the Social Sciences; Urban Anthropology; Urban Design and Public Space; Urban Geography; Urban Planning (Western), History of; Urban Sociology; Urbanization.

Bibliography

Bell, D.A., de-Shalit, A., 2011. The Spirit of Cities: Why the Identity of a City Matters in a Global Age. Princeton University Press, Princeton, NJ.

Byrne, P.H., 1989. Jane jacobs and the common good. In: Lawrence, F. (Ed.), Ethics in Making a Living: The Jane Jacobs Conference. Scholars Press, Atlanta, GA, pp. 169–190.

Cichello, A., 1989. In defense of Jane Jacobs: an appreciative overview. In: Lawrence, F. (Ed.), Ethics in Making a Living: The Jane Jacobs Conference. Scholars Press, Atlanta, GA, pp. 99–161.

Ellin, N., 2010. The tao of urbanism. In: Goldsmith, S., Elizabeth, L. (Eds.), What We See: Advancing the Observations of Jane Jacobs. New Village Press, Oakland, CA, pp. 44–56.

Florida, R., 2002. The Rise of the Creative Class. Basic Books, New York.

Glaser, B.G., Strauss, A.L., 1967. The Discovery of Grounded Theory: Strategies for Qualitative Research. Aldine Publishing Company, Chicago.

Goldsmith, S., Elizabeth, L., 2010. What We See: Advancing the Observations of Jane Jacobs. New Village Press, Oakland, CA.
Granovetter, M.S., May 1973. The strength of weak ties. American Journal of Sociology 78 (6), 1360–1380.
Greenberg, K., 2010. The interconnectedness of things. In: Goldsmith, S., Elizabeth, L. (Eds.), What We See: Advancing the Observations of Jane Jacobs. New Village Press, Oakland, CA, pp. 116–124.
Hamdi, N., 2010. The intelligence of informality. In: Goldsmith, S., Elizabeth, L. (Eds.), What We See: Advancing the Observations of Jane Jacobs. New Village Press, Oakland, CA, pp. 34–43.
Harvey, D., 2000. Spaces of Hope. Edinburgh University Press, UK.
Hirt, S., Zahm, D. (Eds.), 2012. The Urban Wisdom of Jane Jacobs. Routledge, Oxfordshire, UK.
Ikeda, S., 2010. The mirage of the efficient city. In: Goldsmith, S., Elizabeth, L. (Eds.), What We See: Advancing the Observations of Jane Jacobs. New Village Press, Oakland, CA, pp. 24–33.
Jacobs, J., 1992[1961]. The Death and Life of Great American Cities. Vintage, New York.
Jacobs, J., 1970. The Economy of Cities. Vintage, New York.
Jacobs, J., 1985. Cities and the Wealth of Nations. Vintage, New York.
Keeley, R.C., 1989a. Some paths through Jane Jacob's thought. In: Lawrence, F. (Ed.), Ethics in Making a Living: The Jane Jacobs Conference. Scholars Press, Atlanta, GA, pp. 29–37.
Keeley, R.C., 1989b. The vision of Jane Jacobs: an overview and interpretation. In: Lawrence, F. (Ed.), Ethics in Making a Living: The Jane Jacobs Conference. Scholars Press, Atlanta, GA, pp. 39–98.
Kidder, P., 2012. The right and the good in jane jacobs' urbanism. In: Hirt, S., Zahm, D. (Eds.), The Urban Wisdom of Jane Jacobs. Routledge, Oxfordshire, UK, pp. 9–24.
King, K., 2013. Jane Jacobs and the need for 'aged' buildings: neighbourhood historical development pace and community relations in. Urban Studies 50 (12), 2407–2424.
Lawrence, F. (Ed.), 1989. Ethics in Making a Living: The Jane Jacobs Conference. Scholars Press, Atlanta, GA.
Luce, E., 20 January 2014. Talent deserts gridlocked Washington for the cities. Financial Times, p. 7.
Martin, D., 25 April 2006. Jane jacobs, urban activist is dead at 89. New York Times. http://www.nytimes.com/2006/04/26/books/26jacobs.html?pagewanted=all (accessed 02.07.2014).
Oldenburg, R., 1989. The Great Good Place. Paragon House, New York.
Sennett, R., 2012. Together: The Rituals, Pleasures and Politics of Cooperation. Penguin, UK.
Simmel, G., 1971. The metropolis and mental life. In: Levine, D. (Ed.), Georg Simmel on Individuality and Social Forms. University of Chicago Press, Chicago, pp. 321–340.
Sorokin, M., 2010. Jane's spectacles. In: Goldsmith, S., Elizabeth, L. (Eds.), What We See: Advancing the Observations of Jane Jacobs. New Village Press, Oakland, CA, pp. xv–xvii.
Suarez, R., 2010. Jane Jacobs and the 'battle for the street'. In: Goldsmith, S., Elizabeth, L. (Eds.), What We See: Advancing the Observations of Jane Jacobs. New Village Press, Oakland, CA, pp. 15–23.
Vertovec, S., 2007. New Complexities of Cohesion in Britain: Super-diversity, Transnationalism and Civil Integration. Commission of Integration & Cohesion, West Yorkshire: Communities and Local Government Publications. www.pps.org/reference/jjacobs-2 (accessed 13.01.14.).

Jakobson, Roman (1896–1982)

Linda R Waugh, University of Arizona, Tucson, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Roman Jakobson was one of the major linguists, literary theorists, and semioticians of the twentieth century, and one of its most creative minds. He contributed to linguistics major fundamental concepts and showed the importance for linguistics of such diverse issues as child language acquisition, poetry, the act of communication, and typology and universals. After a brief overview of Jakobson's life, major themes of his work are discussed in detail: structural phonology, child language acquisition, and aphasia; distinctive features and functional view of sound; invariance and relational structure; functions of language; metaphor/metonymy (similarity/contiguity); grammar; and Jakobson's legacy in linguistics and other fields.

Roman Jakobson (1896–1982) was one of the major linguists, poeticians, and semioticians of the twentieth century. He contributed to linguistics such concepts as (distinctive) feature, binary opposition, markedness, redundancy, structuralism, and universals. He also showed the importance for linguistics of child language acquisition, aphasia, poetry, metaphor and metonymy (similarity and contiguity), the act of communication (speech event), the meaning of grammatical categories, typology and universals, and the systematicity of language change. And he pointed out the relationship of linguistics to semiotics, cultural anthropology, literary studies, studies of the brain, neurology, and so forth. Jakobson helped to found Russian Formalism, was a cofounder of the Prague Linguistic Circle, and was a dominant force in American linguistics from the early 1940s onward. He was a leading proponent of Prague functional structuralism and pointed out the importance of all the functions of language (including the poetic function) for linguistic analysis. His work in Prague and then in the US aided the development of functionalist approaches, more generally, pragmatics, sociolinguistics and ethnography of communication, discourse and text analysis, psycholinguistics, neurolinguistics, aphasia studies, structuralist approaches to language change, and geographical linguistics. He also influenced literary analysis, poetics, semiotics, social anthropology, mythology, and folklore. (Most of Jakobson's work is reprinted in his *Selected Writings*. For a one-volume introduction to his work in linguistics, see his *On Language*. For more detail and further bibliography for this article, see Waugh and Monville-Burston, 1990, 2002; see also Bradford, 1994.)

Brief Overview of Jakobson's Life

Jakobson was born in Moscow on 11 October 1896 and died in Cambridge, Massachusetts, on 18 July 1982. While still a student in his teenage years he helped to found the two groups known now as Russian Formalism (a combination of the Moscow Linguistic Circle and the Society for the Study of Poetic Language in St. Petersburg), both of which brought together linguists, literary scholars, and writers (especially, aestheticians and poets of the avant-garde). They insisted on the autonomy of literary studies and called for an imminent analysis of literary works, with a focus on the properties that distinguish literary material from any other kind. In 1920, Jakobson moved to Prague, Czechoslovakia and in 1926 helped to found the Prague Linguistic Circle, the cradle of (functional) structuralism. For the Prague School, language serves for communication and thus its internal structure has to be studied from the standpoint of the tasks it performs. Jakobson first coined the term structuralism in 1929 as a cover term for this viewpoint, and was a leading proponent of this highly influential twentieth-century trend in linguistics, and in many allied humanistic, social, and behavioral fields.

In 1939, Jakobson fled the Nazis and in 1941 went to the USA. He taught at the *Ecole Libre des Hautes Etudes*, New York, from 1942 to 1946 and in 1943 he began teaching at Columbia University. In 1949, he obtained a Professorial Chair at Harvard University. From 1957 onward, he was also an Institute Professor at MIT. He was President of the Linguistic Society of America in 1956 and received the International Prize for Philology and Linguistics in 1980 and the Hegel Prize in 1982.

Structural Phonology, Child Language, and Aphasia

During his Prague period, Jakobson and fellow Russian linguist (Prince) Nikolai Trubetzkoy argued that phonological systems are structural wholes, which have to be studied in terms of their relational properties, and that language change has to be seen in terms of systems. They investigated types of phonological systems and established the study of linguistic affinities between languages – how languages that are geographically and culturally related may acquire linguistic traits in common (see Jakobson, 1971a).

During this time Jakobson also turned his attention to the acquisition of phonology by children and its dissolution in aphasics. His monograph – *Child Language, Aphasia and Phonological Universals* (Jakobson, 1941) – demonstrated that the order of acquisition follows the laws of typology: from universal to nonuniversal, unmarked to marked, simple (optimal) structures to more complex ones. He also showed that aphasic loss follows a reverse order.

Theory of Distinctive Features and Functional View of Sound

In his American period and with the aid of advances in acoustic phonetics, Jakobson developed his famous theory that

phonemes such as /t/ and /d/ in *ten* and *den* and /b/ and /m/ in *bat* and *mat* are not the ultimate (smallest) constituents of language. Rather, the distinctive features are the smaller properties of sound by which phonemes are differentiated from each other. For example, /t/ is differentiated from /d/ as unvoiced (voiceless) versus voiced in many languages, and /b/ from /m/ as oral (nonnasal) versus nasal. Jakobson also defined a small set of such (acoustic-perceptual) features that underlie the phonological systems of all the languages of the world (Jakobson et al., 1952; Jakobson and Halle, 1956; Jakobson and Waugh, 1979). He also showed the relevance of these notions for grammatical meaning, which he decomposed into semantic features (see Jakobson, 1984).

Influenced by communication (information) theory and cybernetics, Jakobson explored more deeply the function of language for communication. He insisted, in particular, that all facets of the speech sound have a function (Jakobson and Waugh, 1979): while distinctive features differentiate between words that are different in meaning, redundant features serve to support and enhance the distinctive features. Other features may show the boundaries of words (configurative), inform about the attitude of the speaker (emotive), or indicate the age, sex, social class, etc. of the speaker (physiognomic).

Since distinctive and redundant features have no positive content, but all other types of signs in language do (including words and sentences), there are two types of signs in language; and language and the genetic code are the only systems to have this kind of structure. This differentiation led to a new dichotomy for language (see Jakobson and Waugh, 1979; Jakobson, 1980): mediacy (elements with an indirect relation to meaning such as distinctive features) versus immediacy (direct relation between sound and meaning: phenomena such as sound symbolism; synesthesia; glossolalia; and mythic, poetic, and magical uses of sound). In many cases of mediacy, there is an iconic (nonarbitrary) relation between sound and meaning (see also Jakobson, 1966). This dichotomy is also related to the hemispheres of the brain (left vs right, respectively) in Jakobson (1980).

All of the feature types together represent the entire makeup of a sound and there is nothing that is pure sound (Jakobson and Waugh, 1979). Consequently, the oppositions traditionally made between linguistic and nonlinguistic, abstract and concrete, functional and nonfunctional, (phon)emic and (phon)etic are inappropriate.

Invariance and Relational Structure

Invariance was, for Jakobson, the dominant topic and methodological device underlying his research: any sign is defined by invariant properties that differentiate it from the other elements in the system of which it is a part. Across languages too, he also insisted on the importance of universal invariants (language universals), establishing the equivalence of diverse signs. In addition, he correlated invariance with contextual variation: any sign evidences variation as it enters specific contexts. Contextual variants are not equal, but hierarchized: some are more basic, others more marginal. Invariance and variation apply to the *signans* (e.g., sound, as in variation across phonemes) and the *signatum* (e.g., the semantic properties of the case and verbal systems of Russian (see Jakobson, 1984)). In semantics, by incorporating invariance (context-independent meaning, general meaning) and variation (context-specific meaning, polysemy, reference), he gave the basis for a rigorous approach to grammatical and lexical meaning.

Invariance for Jakobson was always invariance in relationship (equivalence); and the primary type of relation in language is that of binary opposition, phonological (e.g., nasal vs oral), grammatical (e.g., singular vs plural), or lexical (e.g., 'near' vs 'far'), etc. (see Jakobson, 1971a,b, 1987; Jakobson and Waugh, 1979). Such binarism is based on an asymmetry in the relationship between a marked (focused or weighted) and an unmarked (neutral) term, for example, marked nasal, plural, 'near' vs unmarked oral, singular, 'far.' For Jakobson, markedness is a fundamental way by which we organize not just language but cultural systems in general.

Functions of Language

For Jakobson, the *raison d'être* for language is communication and language is suited to various communicative goals, which in turn are correlated with the act of communication in which language is used. In his famous article, *Linguistics and Poetics*, he defined the six primary factors of any speech event: (1) speaker (addresser), (2) addressee, (3) context (including the thing referred to), (4) message (*parole*, language use), (5) code (*langue*, linguistic system), and (6) contact (medium) of communication. In conjunction with these, he then defined the functions of language in terms of a focus in the message on one of the factors: (1) emotive function (focus on the speaker), for example, intonation showing anger; (2) conative function (focus on the addressee), for example, imperatives and vocatives; (3) referential function (focus on the context), for example, talk about the real world; (4) poetic function (focus on the message), for example, poetry; (5) metalingual function (focus on the code), for example, definitions of words; and (6) phatic function (focus on the contact), for example, 'hello, do you hear me?' These functions may be dominant in a message or may be subsidiary; for example, a referential message may carry expressive information about the opinion of the speaker.

Metaphor and Metonymy: Similarity/Contiguity

In 1956, Jakobson published his famous *Two Aspects of Language* where he analyzed the relation between communicative processes and properties of linguistic structure. First, he distinguished two fundamental operations: selection (substitution) for encoding (production) and combination for decoding (comprehension). Then, he contrasted two types of relations: similarity (all types of equivalence) and contiguity (temporal and spatial neighborhood, causal relations, part–whole), and said that they underlie the two operations. Put simply, the elements in a selection set are normally associated by similarity and those in combination by contiguity. Since in aphasia, the normal patterning of speech is disturbed, Jakobson proposed that these factors be used as a way of categorizing types of aphasic disturbances.

For Jakobson similarity/contiguity is a fundamental polarity of language, texts, culture, and human thought (cognition) in

general. He used it for much of his work in semiotics of film and music. He also used it to refine his definition of the poetic function (1960): in poetry, where focus on the message is dominant, equivalence (similarity) relations are used to build the combination rather than only to underpin the elements of the selection set. Thus, rhyme, alliteration, parallelisms, metaphor, etc. help to structure the poetic text, whereas in prose contiguity is the essential constructional principle.

Grammar

Jakobson also focused on the function of grammatical categories. His famous article on *Shifters* (1957) focuses on those elements whose general meaning in the code can only be specified by taking into account their use in messages, because the codal meaning includes information about particular elements of the speech event. For example, deictic pronouns designate speaker (I) and addressee (you). Thus, language encodes pragmatic factors of the context of utterance and linguistics necessarily includes syntax, semantics, and pragmatics.

Grammatical categories (both morphological and syntactic), for Jakobson (citing Franz Boas), are defined as those which are obligatorily present in the construction of acceptable messages (1959a,b), whereas particular lexical categories (e.g., words referring to space) are optional. Through this view of grammar, Jakobson provided a semantic and operational approach to the relation between language and cognition: grammatical categorizations provide the necessary patterns of thought.

Jakobson was also inspired by Charles Sanders Peirce's notion that the essence of a sign is its interpretation, that is, its translation by some further sign. Henceforth, he defined the *signatum* as that which is interpretable or, better, translatable (Jakobson, 1959b) by a potentially unlimited series of signs. He characterized the Peircian approach as the only sound basis for a strictly linguistic semantics.

Jakobson's Legacy in Linguistics and Other Fields

Jakobson's influence on twentieth-century thought was considerable and many of his contributions have become a permanent part of American and European views of language. He has had such a towering role in linguistics that his work has defined the field itself. Many of Jakobson's concepts and discoveries are now so deeply ingrained in the theoretical basis of modern linguistics that they are thought to be commonplace or self-evident. This article will consider, in particular, his influence on linguistics in the last 30 years or so and later offer some brief notes on other disciplines. (For more details of Jakobson's influence, see Waugh and Monville-Burston, 1990, 2002.)

Influence on Linguistics

Jakobson's view of language as a relational whole and a communicative tool with many functions opened the door to functionalist approaches more generally, and abetted the development of linguistic studies that look at discourses and texts and consider their wider linguistic and 'real-world' environment. His work on *Shifters* showed that grammar, semantics, and pragmatics are but different sides of the same phenomenon and aided present-day interest in pragmatics. Work in pragmatics has led linguists to explore the boundaries of their discipline with the neighboring fields of anthropology, sociology, mythology, philology, and philosophy in a truly Jakobsonian interdisciplinary spirit. His writings on grammatical meaning have inspired others to investigate the interdependence of invariance and variation, the semantics of case systems, and the interrelation of tense and aspect with syntax, discourse, and the lexicon.

Jakobson's insistence that language is a complex and dynamic whole with a variety of subcodes (speech styles) helped to pave the way for the advent of sociolinguistics. The ethnography of communication has also taken its inspiration from Jakobson's multifunctional perspective on communication. His work on the functions of language and their relation to the speech (communication) act remains central in the investigation of speech in interaction. Anthropological linguists – who adopt a functional and pragmatic perspective with an interest in folklore, poetics (oral traditions), and discourse analysis – draw on and develop the ideas of the Prague School and Jakobson in particular.

Jakobson stressed the need to search for general laws governing all linguistic systems and thus helped to foster the current interest in generalizations. He saw a definite equivalence between the search for explanatory adequacy in generative grammar and the elicitation and interpretation of the entire network aimed at by Prague structuralism. His monograph on child language and aphasia, generally regarded as one of his most original and influential works, has been signaled as the point of departure of a new era in linguistics and has been lauded by many because it showed the relation between child language and phonology, launched psycho- and neurolinguistics, and ushered in a new focus on universal and law-governed, general (explanatory) properties of language.

The widespread movement that presently investigates universal properties of language, both universal grammar and language universals, is deeply indebted to Jakobson. When the linguist Noam Chomsky came in contact with Jakobson and worked with Jakobson's student and coauthor Morris Halle, he fully realized the explanatory power of universals and their relation to language acquisition; thereafter, he set out to determine universal constraints on the form of grammars. The other stream of research in universals, practiced by Joseph Greenberg and his associates, examines concrete materials from a great variety of languages in order to establish typological classifications and establish empirically the general laws that underlie the structure of human language.

Many scholars agree with Jakobson that typology is important for the study of language change, areal linguistics, and historical reconstruction. His work in historical, geographical, and areal linguistics continues to be reprinted in readers in these areas. The seminal idea that changes must always be treated in view of the system that undergoes them is the essential premise for textbooks in historical linguistics. And Jakobson's view of the teleological basis for language change has been instrumental in inspiring further work.

By revealing the close relationship between developmental studies and vital questions of linguistic theory, Jakobson's *Child Language* also launched psycholinguistics. Many psycholinguists now recognize that the broad outlines of this bold synthesis continue to be confirmed. Following Jakobson's example, scholars interested in language pathology have studied aphasia from a linguistic point of view and, for many, Jakobson's distinction between similarity disorders and contiguity disorders is regarded as fundamental, and firmly supported by contemporary ideas of the functional organization of the human brain. Jakobson was the first modern linguist to investigate in depth the concurrent constituents of linguistic elements. Many linguists have claimed that the distinctive feature was Jakobson's greatest insight and, after the phoneme, the most significant step forward in modern phonology. Generative phonology recognizes Jakobson as one of its founding fathers, in particular, through Halle. Markedness and the correlated notion of binary opposition are now widely used in linguistics and other disciplines. Componential analysis – analyzing an item into smaller properties and representing them as a combination – was generalized as a methodological model for semantics and lexical representation, and has been adopted in other approaches as a means of describing the structure of a variety of semantic fields.

Influence on Other Fields

Jakobson's influence on contemporary thought has also spread to disciplines such as literature, poetics, semiotics, social anthropology, mythology, and folklore. His theoretical papers *Linguistics and Poetics* and *Poetry of Grammar and Grammar of Poetry* and his analytical studies of poems by Baudelaire and Shakespeare (as well as many other poets in many other languages) established him as an important theoretician of literature (see Bradford, 1994). He is the source of not only structuralist but also other contemporary trends in literary theory. He also helped to develop semiotics as a discipline in its own right and has been called the major catalyst in the contemporary semiotic reaction (see Jakobson, 1973). And he is credited with introducing the American semiotician Charles Sanders Peirce to Americans and with showing them the importance of Peirce's insights. His work in semiotics has also had repercussions in fields as diverse as economics, mathematics (especially catastrophe theory), sociology, and philosophy.

Jakobson's influence on anthropology is considerable. In American anthropology, for example, three convergent Jakobsonian ideas have proven to be particularly fruitful: the structure of the poetic text, the closeness of ties between poetry and mythology, and the relationship between discourse and cultural context. His greatest impact, however, has been on structural anthropology, especially through Claude Lévi-Strauss, who considers himself a disciple of Jakobson and has repeatedly acknowledged his indebtedness to him. He used many Jakobsonian ideas in his work on kinship systems, cooking, folktales, myths, and primitive thought.

With the breadth and depth of his knowledge, his brilliant imagination, his prodigious output, the originality of his accomplishments, and his international influence on a variety of fields, Roman Jakobson has been recognized worldwide as one of the major creative minds of the twentieth century.

See also: Anthropology: Overview; Aphasia; Areal Linguistics; Avant-Garde Art and Artists; Boas, Franz (1858–1942); Brain Asymmetry; Ethnography in Applied Social Research; First Language Acquisition, Linguistic Theory of; Functional Approaches to Grammar; Functionalism in Anthropology; Generative Grammar; Iconicity in Linguistics; Information Theory; Interpretation in Anthropology; Language Acquisition; Lexical Semantics; Linguistic Typology; Linguistics: Overview; Pragmatics, Linguistic; Psycholinguistics: Overview; Semiotics; Structuralism.

Bibliography

Bradford, R., 1994. Roman Jakobson: Life, Language, Art. Routledge, London; New York.

Jakobson, R., 1941. Kindersprache, Aphasie und allgemeine Lautgesetze. Almqvist and Wiksells, Uppsala, Sweden (1968. Child Language, Aphasia, and Phonological Universals. Mouton, The Hague; Paris).

Jakobson, R., 1956. Two aspects of language and two types of aphasic disturbances. In: Jakobson, R., Halle, H. (Eds.), Fundamentals of Language. Mouton, The Hague.

Jakobson, R., 1957. Shifters, verbal categories, and the Russian verb. In: Jakobson, R. (Ed.), Selected Writings II: Word and Language. Mouton, The Hague, pp. 130–147.

Jakobson, R., 1959a. On linguistic aspects of translation. In: Brower, R. (Ed.), On Translation. Harvard University Press, Cambridge, MA, pp. 232–239 (1966. Second ed. Oxford University Press, New York).

Jakobson, R., 1959b. Boas' view of grammatical meaning. In: Goldschmidt, W. (Ed.), The Anthropology of Franz Boas. Essays on the Centennial of His Birth.

Jakobson, R., 1960. Linguistics and poetics. In: Sebeok, T. (Ed.), Style in Language. MIT, Cambridge, MA.

Jakobson, R., 1966. Quest for the essence of language. Diogenes 51, 21–37.

Jakobson, R., 1971a. Selected Writings I: Phonological Studies, second ed. Mouton, The Hague; Paris.

Jakobson, R., 1971b. Selected Writings II: Word and Language. Mouton, The Hague; Paris.

Jakobson, R., 1973. Main Trends in the Science of Language. George Allen and Unwin, London.

Jakobson, R., 1980. Brain and Language: Cerebral Hemispheres and Linguistic Structure in Mutual Light. Slavica, Columbus, OH.

Jakobson, R., 1981. Selected Writings III: Poetry of Grammar and Grammar of Poetry. Mouton, The Hague; Paris; and New York.

Jakobson, R., 1984. Russian and Slavic Grammar: Studies, 1931–1981. Mouton, Berlin; New York.

Jakobson, R., 1987. Selected Writings VIII: Major Works, 1976–1980. Mouton, Berlin.

Jakobson, R., 1990. On Language. Harvard University Press, Cambridge, MA.

Jakobson, R., Fant, C.G.M., Halle, M., 1952. Preliminaries to Speech Analysis. MIT, Cambridge, MA.

Jakobson, R., Halle, M., 1956. Fundamentals of Language. Mouton, The Hague.

Jakobson, R., Jones, L.G., 1970. Shakespeare's Verbal Art in Th'Expence of Spirit. Mouton, The Hague; Paris.

Jakobson, R., Lévi-Strauss, C., 1962. Les chats de Charles Baudelaire. L'Homme 2, 5–21.

Jakobson, R., Waugh, L.R., 1979. The Sound Shape of Language. Indiana University Press, Bloomington, IN, 1987. Second, augmented edition. Berlin: Mouton de Gruyter. 1987. Second edition reprinted in Selected Writings, Volume VIII, by Jakobson, R., Rudy, S. (Eds.), Berlin: Mouton.1–315. 2002. Third, augmented edition. Berlin: Mouton de Gruyter. 2002. Third edition reprinted in Selected Writings, Volume VIII, by Jakobson, R., Rudy, S. (Eds.), Berlin: Mouton.

Jakobson, R., 1990. In: Waugh, L.R., Monville-Burston, M. (Eds.), On Language. Harvard University Press, Cambridge, MA. Updated in 2002. Introduction to Roman Jakobson: Selected Writings. On Language: The Life, Work and Influence of Roman Jakobson. In Roman Jakobson, Selected Writings, I, third edition, v–xcviii.

James, William (1842–1910)

M Brewster Smith, University of California, Santa Cruz, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 7949–7954, © 2001, Elsevier Ltd.

Abstract

William James, American philosopher and psychologist, remains a vivid presence in psychology because of his two classics, *Principles of Psychology* (1890/1981) and *Varieties of Religious Experience* (1902/1985). His work at once embodies the culmination of nineteenth century psychology, in his critical review of the older philosophical psychology and of the new experimental psychology, and the launching of twentieth century psychology, especially the functionalist tradition. He is remembered in philosophy as one of the last philosophers in the grand style who also sought to communicate his insights to a broad audience, an audience attracted by the informal felicity of his writing, public lecturing, and teaching. Distinctively American in his promotion of a pragmatism that viewed abstract concepts critically in terms of their 'cash value,' he participated actively in the European intellectual scene and, like his eminent novelist brother Henry James, advanced the linkage of the American and European intellectual and cultural worlds.

William James, American philosopher and psychologist, remains a vivid presence in psychology because of his two classics, *Principles of Psychology* (1890/1981) and *Varieties of Religious Experience* (1902/1985). His work at once embodies the culmination of nineteenth century psychology, in his critical review of the older philosophical psychology and of the new experimental psychology, and the launching of twentieth century psychology, especially the functionalist tradition. He is remembered in philosophy as one of the last philosophers in the grand style who also sought to communicate his insights to a broad audience, an audience attracted by the informal felicity of his writing, public lecturing, and teaching. Distinctively American in his promotion of a pragmatism that viewed abstract concepts critically in terms of their 'cash value,' he participated actively in the European intellectual scene and like his eminent novelist brother Henry James advanced the linkage of the American and European intellectual and cultural worlds.

His enduring significance at the end of the twentieth Century is reflected in the publication of definitive scholarly editions of his complete writings in 17 volumes (1975b–88) and of his correspondence (1992) of which 12 volumes were planned. As members of an unusually creative family, he and his brother Henry and sister Alice have been the subject of extensive biographical attention (e.g., Lewis, 1991; Myers, 1986; Perry, 1935).

Family Background and Early Life

The James family was economically privileged. William's father, Henry James Sr. (1811–82), who had inherited a comfortable living, was a restless, questing intellectual who found inspiration in the writings of the Swedish mystic Emanuel Swedenborg. He attracted visits from the New England transcendentalists Emerson and Thoreau while William was an infant in New York City, where he was born 11 January 1842. William's education was irregular, including periods of schooling or tutoring in England, France, Germany, and Switzerland. The intense moral and intellectual climate of the family seemingly entailed severe costs in physical and mental health for William and his gifted siblings. Teasing out the sources of creativity and of pathology in the James family has challenged its biographers.

When William was 18, he embarked on the study of painting in the Newport studio of William Hunt, a prominent American artist. But he gave up painting after about 6 months, turning next to science. He pursued studies in chemistry and comparative anatomy at Harvard, and, in a smaller shift in direction, entered Harvard Medical School when he was 22. These were years of increasing doubt, depression, and psychosomatic discomfort. James interrupted his medical training for a year to take part in an expedition to the Amazon with the Harvard biologist Louis Agassiz, and interrupted it once more to spend a year and a half in Germany taking therapeutic baths and reading widely in German literature, philosophy, and science. Returning to complete his M.D. degree in 1869, he spent several years of mental distress and near collapse living with his parents.

It was during this period that, by his own account, he drew on philosophy in a turn toward recovery, as he recorded in a subsequently famous passage in his diary of 30 April 1870:

> I think that yesterday was a crisis in my life. I finished the first part of (Charles) Renouvier's second *Essais* and see no reason why his definition of free will – 'the sustaining of a thought because I chose to when I might have other thoughts' – need be the definition of an illusion. At any rate, I will assume, for the present – until next year – that it is no illusion. My first act of free will shall be to believe in free will (Perry, 1935, vol.1, p. 323).

This passage heralds the theme of voluntarism that pervades his later writings, as well as suggesting roots of his pragmatic philosophy.

James's productive life began in 1873 when he was appointed instructor in anatomy and physiology at Harvard. By 1875 he was teaching his first course in psychology, in connection with which he established a teaching laboratory that, as American historians of psychology like to note, preceded Wilhelm Wundt's laboratory at Leipzig in 1879,

a date often enshrined as the founding of scientific psychology. By 1878, when he was 36, he had contracted to write the textbook on psychology that was not to see publication until 1890. It was also in 1878 that he married Alice Gibbens, a teacher who shared his intellectual and spiritual interests and had spent 5 years in Heidelberg with her widowed mother. This event reflected the stabilization of his personal life and also contributed greatly to it. In the following year he offered his first course in philosophy. He was effectively launched in his dual career as psychologist and philosopher at Harvard where he remained for the rest of his life. He died on 26 August 1910, at his family retreat at Chocorua, New Hampshire.

Major Themes

Three themes pervade James's contributions to psychology and philosophy. One is his commitment to voluntarism, the primacy of human agency, which led him to reject mechanistic doctrines. A second is his continual struggle to find a place for religious spirituality as exemplified by his father in the world revealed by Darwinian science. The third, an essential ingredient in his handling of the first two, is his commitment to pluralism and openness in theorizing, as shown in his distaste for abstract systemization and in his opposition to what he saw as the absolutism of both scientific materialism and Hegelian idealism. These themes characterize James's thought throughout his career as psychologist and philosopher, but in the spirit of the third theme, paradoxical inconsistencies recur in his psychology and philosophy.

James's Psychology

In his great book, *The Principles of Psychology* (1890/1981), James presented an account of mental life in the positivistic spirit of natural science. That meant for him that speculative entities beyond experience like the soul or psychological faculties were to be avoided. He took a commonsense dualism of mind and body for granted, though he was explicit that fundamental metaphysical issues would have to be faced eventually. The principal psychological method for him was introspection, by which James mostly meant the informal and avowedly fallible self-observation that he did so well, not the disciplined and quantified introspection of the new German experimental laboratories, which he respected while noting wryly that the method 'could hardly have arisen in a country whose natives could be *bored*.' (p. 192).

James's Functionalism

To apply a term that emerged with his immediate American successors, James was presenting a functional psychology, in the senses subsequently intended by the Chicago functionalists. His focus was on mental processes (functioning), not on the analysis of mental contents into constituent sensory elements. Further, he emphasized the adaptive functions of psychological processes, in line with evolutionary theory. Consciousness, he proposed, emerged to organize and direct neural processes in the brain that had become too complex to be guided by a mechanistic concatenation of reflexes. His vision of scientific psychology was committed to a mind–body interactionism troublesome to both scientific materialists and philosophers.

James viewed the nervous system and the brain as organs of mind, and devoted early chapters of the *Principles* to the functions of the brain. There follows a chapter on habit, foreshadowing the later preoccupations of J.B. Watson's behaviorism. James's account of habit was primarily physiological. Currents in the cerebral cortex originating in the sense organs 'leave their traces in the paths which they take. The only thing they can do … is to deepen old paths or to make new ones. … (A) simple habit … is, mechanically, nothing but a reflex discharge; and its anatomical substratum must be a path in the system' (p. 112). From this physiological base James proceeded with recommendations for the cultivation of good habits, putting neural mechanism to voluntary use for moral purposes, a characteristically Jamesian manoeuver.

The Stream of Thought

If James's treatment of the brain and of habit pointed toward later developments in neuropsychology and behaviorism, his classic treatment of the stream of thought provided a sensitive and persuasive alternative to the prevalent approaches of introspective psychology that sought to decompose ideas and images into sensory elements. For James, the proper starting point in psychology is not sensations or feelings, as in prevalent elementaristic psychologies, but the sheer fact that thought goes on. 'Every thought tends to be part of a personal consciousness,' he wrote (p. 220). Thought seems continuous but is always changing. It deals actively and selectively with objects independent of itself. The idea that thought is actively selective underlies James's later discussion of attention, which in turn provides the rationale for his treatment of will, in terms of holding a thought of action in focus long enough for the action to occur.

No two states of consciousness are ever the same. 'A permanently existing "idea"... which makes its appearance before the footlights of consciousness at periodical intervals is as mythological an entity as the Jack of Spades' (p. 230). Relationships and transitional states are given in the primary data of introspection just as directly as the more substantive states on which consciousness may momentarily 'perch.' James's views here anticipated those of the Würzburg school regarding imageless thought, and foreshadowed Gestalt psychology in their rejection of sensory elements. James's evocative term, 'the stream of consciousness,' found a place in the context of world literature.

The Consciousness of Self

Since the stream of thought is always a personal consciousness in which the objects of awareness are divided into 'me' and 'not me,' James followed his treatment of the stream of thought with a chapter on 'The Consciousness of Self' that was very influential on later scholarship. If 'I' is taken to label the person as agent and introspective observer, 'me' is the self as object of introspective observation. For the purposes of his psychology, the 'I' is simply the passing thought in the personal stream,

which appropriates the person's memories of the past and anticipations of the future. The 'me' comprises a hierachy of three selves. First is the material self – one's body, clothes, and the material possessions that one regards as 'mine.' The social self, or more properly, selves, consists of the views of oneself held by others that one cares about, as one understands them. The spiritual self is 'a man's inner or subjective being, his psychic faculties and dispositions, taken concretely … the most enduring and intimate part of the self' (p. 283). These three aspects of 'me' are the focus of motives of self-seeking and self-preservation, and of self-feelings of self-complacency and self-dissatisfaction. James provided a famous 'formula' for self-esteem as the ratio of Success to Pretensions, anticipating much later research and theory about Level of Aspiration.

The distinctively human attribute of reflective self-awareness highlighted in James's treatment of selfhood was subsequently developed mainly by social psychologists in the sociological tradition, especially C. H. Cooley and G. H. Mead, but returned as a central concern of psychologist students of personality and social behavior in the latter half of the twentieth century.

Emotion

Of all the special psychological topics that James discussed, his theory of emotion evoked more notice and controversy than any other. It was known as the James–Lange theory because in the *Principles* James linked his ideas as previously published in the 1880s with similar views of the Danish physiologist Carl Georg Lange published at about the same time. James's version held that the stronger or 'coarser' emotions of grief, fear, rage and love arise from feeling bodily changes that are aroused directly by an 'exciting fact.' In contrast with folk belief, we do not weep because we feel sad; we feel sad because we are weeping. We feel fear because we are set to run away. We feel because we act.

During the first half of the twentieth century, the James–Lange theory did not fare well. In one important critique, the physiologist Walter Cannon showed in animal research that centrally mediated emergency emotional responses do not depend on feedback from the viscera, a finding that bore more critically on Lange's version of the theory than on James's. More importantly, in a period dominated by behaviorism, James's central interest in emotion as conscious feeling seemed scientifically irrelevant. With the subsequent decline of behaviorism, each of several competing treatments of emotion looked back to James as anticipating their conclusions. Paul Ekman and Carroll Izard focused on the role of sensory feedback from the facial musculature in emotional expression as crucial to the distinctiveness of felt emotions, whereas Stanley Schachter launched a cognitively focused line of research emphasizing the interpretive context in which arousing stimulation is encountered. As cognitive perspectives on human psychology developed one-sidedly, James's emphasis on the centrality of feeling became more attractive.

Other Aspects

As embodied in the *Principles*. James's psychology dealt with many other topics, some familiar in later textbooks, others less so. At first encounter, James's account of the will seems foreign to present students of psychology, who are accustomed instead to nonconsensual chapters on motivation. For James, voluntary movements are a secondary development from the primary organic functions involved in reflex action, instincts, and emotional movements (p. 1099). The link between thought and voluntary action is provided by ideo-motor action, the concept that thoughts of a movement lead directly to making the movement, unless contrary inhibiting thoughts intervene. With the resurgence of cognitive and neuropsychology, James's focus on the selectivity of attention and on the relation between thought and action returned as central to the discipline. His century-old formulations are no longer useful, but his identification of significant problems for psychology earns renewed respect.

Other topics broached in the *Principles*, particularly hypnosis and divided consciousness, anticipated concerns beyond the boundaries of positivistic science that continued to preoccupy James during his mainly philosophic period that followed its publication (see Taylor, 1983). James was much interested in the work of the French psychopathologists Charcot, Janet, and Liébeault, who preceded psychoanalysis in giving attention to nonrational processes outside the scope of normal consciousness.

James as a Philosopher

In spite of James's expressed intent in the *Principles* to avoid metaphysics, the book was infused with his philosophic thinking, particularly in his critique of earlier philosophic psychology, and contains the germ of most of the philosophic positions on which he elaborated subsequently. His specific impact on philosophy is identified with pragmatism and radical empiricism. Pragmatism, launched by James, became a vigorous philosophic movement which enjoyed a renaissance late in the twentieth century. Radical empiricism was James's attempt at the fundamental metaphysical construction that he postponed in the *Principles*. The work was incomplete at the time of his death, and the book presenting it (James, 1912/1976) was assembled posthumously from journal articles.

Pragmatism

When James published *Pragmatism* (1907/1975a), a popular book that was much reprinted and translated, he drew on his earlier writing and thinking over three decades. Here, he offered the pragmatic method as a way of settling metaphysical disputes that might otherwise be interminable, such as those between tenderminded religious and toughminded materialistic world views: 'The pragmatic method … is to try to interpret each notion by tracing its respective practical consequences. What difference would it make practically to anyone if this notion rather than that notion were true?' (1907/1975a, p. 28) The pragmatic method entails a perspective on meaning. As he writes, 'to develop a thought's meaning, we need only determine what conduct it is fitted to produce: that conduct is for us its sole significance' (p. 29). And with pragmatic meaning comes a pragmatic conception of truth, as characterizing beliefs

that receive confirmation when acted upon. 'The truth of an idea is not a stagnant property inherent in it. Truth *happens* to an idea. It *becomes* true, is *made* true by events. Its verity *is* in fact an event, a process …' (p. 97). With characteristic generosity, James attributed his initial ideas about pragmatism to his friend the semiotic philosopher C. S. Peirce, who did not agree with James's developed approach.

The extensive philosophical controversy about pragmatism in the first decade of the twentieth century appears to reflect considerable mutual misunderstanding on the part of both James and his critics, as well as some problems with his mostly informal exposition (see Myers, 1986). He can be interpreted as treating claims of pragmatic truth as similar to the provisional truth claims of empirical science, in the spirit of Peirce and, later, John Dewey, or alternatively, as rationalizing wishful thinking (religious faith has good personal and social results, so it may be accepted legitimately as true by its believers). The idea that philosophical abstractions and assertions should be judged in terms of their practical consequences came to seem typically American, to both his supporters and his critics.

Late in the twentieth century, interest in pragmatism revived, with similar diversity of interpretation. One strand of thought prominently represented by the philosopher Rorty (1979) became part of the 'postmodernist' attack on the scientific world view to which some disaffected psychologists subscribed. Knowledge, in this version, is a human construction for human purposes and cannot be taken as a 'mirror of nature,' a view with roots in James. Some other humanistically inclined applied psychologists called for a pragmatic psychology focused on useful contextually limited truths rather than general laws. This version of pragmatism retains a commitment to evidential empiricism like that of the early James.

Radical Empiricism

As James entered his final decade increasingly troubled by his failing heart, he turned to the fundamental metaphysical issues that he had tabled when he wrote the *Principles*. In pragmatism, he had crystalized a theory of knowledge developed over his full career. Radical empiricism, in contrast, proposed an ontology that continued his attack on Hegelian absolute idealism but broke sharply with the commonsense mind–body dualism of his scientific psychology. To transcend this Cartesian dualism, and the cognitive dualism of thought and its object or consciousness and its contents, he now proposed pure experience as the fundamental reality. Pure experience prior to analysis or conceptualization does not distinguish between thing and thought. The commonsense categories, he believed, could be recovered from the contexts in which ingredients of pure experience are embedded and their various relationships. His apparent move toward philosophical idealism was qualified by his continuing commitment to pluralism: '… there is no *general* stuff of which experience at large is made. If you ask what any one bit of pure experience is made of, the answer is always the same: "It is made of *that*, of just what appears, of space, of intensity, of flatness, brownness, heaviness, or what not"' (1912/1976, p. 14–15).

Religion and Psychic Research

James continually struggled to find a place for the moral and hopeful view of humanity's place in the universe that religions can provide. Pragmatism was one route through which he sought to reconcile the world views of religion and science. Another was psychic research. For two years he was president of the Society for Psychical Research, to which he belonged from 1884 until his death. He was ever hopeful of finding supportive evidence for spiritual realities, and never satisfied that the evidence actually obtained was adequate (cf James, 1986).

James was already esteemed as a leading philosopher and psychologist when he delivered invited lectures on 'natural religion' at the University of Edinburgh in 1901–02. Their publication as *Varieties of Religious Experience* (1902/1985) became his second enduring classic, one that, a century later, is not only cited and respected but also read. Displaying the meaning of religion in the religious experience of individuals, it draws richly on accounts by saints, mystics, and others who provide evidence of their relation to what they regard as the Divine. The conceptual distinctions around which he organized his description and analysis have entered our common culture: the religion of healthy-mindedness vs. that of the sick soul; the 'divided self' and 'once-born' and 'twice-born' characters.

Although his sympathetic feeling for the many forms of religious belief is clear throughout, his presentation of religious experience is rightly esteemed by believers and unbelievers alike. His way of deploying 'case materials' to display the nature of a human phenomenon is a model for contemporary qualitative research to emulate.

James regarded the subconscious mental states revealed by the French psychopathologists as important in religious conversion. Speculatively, he declared his conviction that subliminal consciousness overlaps other realms of consciousness in mystical experience. By 'being faithful in my poor measure to this over-belief, I seem to myself to keep more sane and true' (1902/1985, p. 408). Here, James made radical and pragmatic use of his developing 'radical empiricism.' In the late twentieth century, this mystical side of James appealed to disaffected psychologists identified with a 'transpersonal' psychology with religious undertones.

On Public Issues: Imperialism and War

As a public intellectual, James was involved in the social issues of his day. A political as well as a philosophic liberal, he was opposed vehemently to emerging American imperialism in the wake of the Spanish American War. It was as an advisor to the peace movement that he wrote *The Moral Equivalent of War* (1910/1982), one of his last essays published in his lifetime, which remains a classic document in the psychology of war and peace.

After an evolutionary account of human warlike propensities, he noted that the militarists of his day held warfare to be the mainstay of admirable, martial virtues, and that as a 'pleasure economy' supplants the earlier 'pain and fear economy,' these virtues are likely to degenerate. He agreed that war fosters 'manliness' and surrender of private interest, but, given the

monstrosity of modern war, he proposed planning for alternative sources of energizing challenge. Specifically, he suggested 'instead of military conscription a conscription of the whole youthful population to form for a certain number of years a part of an army enlisted against *nature* ...' (p. 171). That he focused on 'manliness' and cast nature in the role of enemy shows James as a creature of his time, but his idea of the moral equivalent of war remains attractive. The equivalence is not in regard to the release of aggression, however, but to self-testing for the common good against difficult challenge.

William James Today

James's vivid and unconventional writings assure his continued influence, as they won him acclaim but evoked controversy in his lifetime. In North American psychology, James has become such a revered ancestor that his name gets called upon in many divergent contexts. His pragmatism is still very much alive. At one point, James considered humanism as an alternative term to pragmatism. Certainly, the influence of James in psychology and philosophy has favored humanistic responsiveness to the richness and variety of human experience, in contrast with the abstractness of absolute idealism in his own day and the dehumanized barrenness of behaviorism, logical positivism, and analytic philosophy that followed. His sensitive phenomenological treatment of thought and self remains unsurpassed, though it has been built on and superseded in various ways. His openmindedness and pluralism set a valuable counter-ideal to the dogmatism that is always a potential liability of mainstream doctrine. This counter-ideal inspired mid-century psychologists of personality such as Gordon Allport, Gardner Murphy, and Henry Murray. At the turn of the millenium, it further inspires transpersonal psychologists who find in James respectable company for their own tendencies toward mysticism. The controversy surrounding James continues.

See also: Darwin, Charles Robert (1809–82); Dewey, John (1859–1952); Freud, Sigmund (1856–1939); Functional Explanation: Philosophical Aspects; Functionalism in Anthropology; Intellectual History; Pragmatism: Philosophical Aspects; Pragmatist Social Thought, History of; Religion and Sciences; Religion, Psychology of; Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

James, W., 1907/1975a. Pragmatism: A New Name for Some Old Ways of Thinking. Harvard University Press, Cambridge, MA.
James, W., 1975b. The Works of William James, vol 17. Harvard University Press, Cambridge, MA and London.
James, W., 1912/1976. Essays in Radical Empiricism. Harvard University Press, Cambridge, MA.
James, W., 1890/1981. The Principles of Psychology. Harvard University Press, Cambridge, MA.
James, W., 1910/1982. The moral equivalent of war. In: James, W. (Ed.), Essays in Religion and Morality. Harvard University Press, Cambridge, MA, 162–173.
James, W., 1902/1985. The Varieties of Religious Experience: A Study in Human Nature. Harvard University Press, Cambridge, MA.
James, W., 1986. Essays in Psychical Research. Harvard University Press, Cambridge, MA.
James, W., 1992. The Correspondence of William James, vol 12. University Press of Virginia, Charlottesville, VA and London.
Lewis, R.W.B., 1991. The Jameses: A Family Narrative. Farrar, Straus and Giroux, New York.
Myers, G.E., 1986. William James: His Life and Thought. Yale University Press, New Haven, CT.
Perry, R.B., 1935. The Thought and Character of William James. Little, Brown, Boston.
Rorty, R., 1979. Philosophy and the Mirror of Nature. Princeton University Press, Princeton, NJ.
Taylor, E., 1983. William James on Exceptional Mental States: The 1896 Lowell Lectures. Scribners, New York.

Janet, Pierre (1859–1947)

John I Brooks, III, Fayetteville State University, Fayetteville, NC, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 7954–7956, © 2001, Elsevier Ltd.

Abstract

Pierre Janet, French psychologist and psychiatrist, was among the first to develop a dynamic psychology using the concepts of subconscious ideas and psychological force. Janet contributed to the understanding of hysteria and of obsessive-compulsive disorders through his theories of dissociation and psychasthenia. His etiology of mental illness stressed both constitutional weakness and the role of traumatic memories. His method, 'psychological analysis,' moved beyond an early reliance on hypnotherapy to include intensive case study and the recovery of dissociated traumatic memories. In many ways, Janet can be ranked with Sigmund Freud as one of the great systematic thinkers in the history of psychology.

Pierre Janet, pioneering French psychologist and psychiatrist, was born 30 May 1859 in Paris. He studied philosophy at the elite *Ecole Normale Supérieure* and began his career teaching philosophy in aprovincial *lycée*. While teaching at Le Havre from 1883–89, he worked with local physicians to conduct experiments on hypnosis with hysterical patients. These experiments led to several publications, including the thesis for his doctorate in philosophy, *L'automatisme psychologique* (1889). In the same year, Janet obtained a position teaching in a Paris *lycée* and began formal training in medicine. He worked at the Salpêtrière hospital under the direction of the foremost authority of the day on hysteria, the neurologist JeanMartin Charcot. Janet received his medical degree in 1893 and continued to work at the Salpêtrière until 1910. The cases he studied there formed the foundation for some of his most important works. In 1902, Janet was named to the chair of experimental psychology at the Collège de France, where he taught until 1935. He also co-founded and edited the *Journal de psychologie normale et pathologique*. He died on 24 February 1947 in Paris.

Janet's thought can be separated into an early and a late period, with 1910 as a rough dividing line. In the early period, which was by far the more influential, Janet focused on extending the domain of mental life and on elaborating a psychology based on a distinction between automatic and synthetic mental activity. As a result of his investigations into hypnotism and post-hypnotic suggestion, Janet rejected the then-common assumption in scientific and medical circles that unconscious mental functioning was physiological in nature. He insisted that it depended on the intervention of ideas, just as conscious mental functioning does. These ideas simply are not connected with the consciousness of the self. Because the word unconscious implied 'neurological' for so many at the time, Janet coined the term 'subconscious' to characterize ideas separated from consciousness.

The distinction between automatic and synthetic mental activity separated Janet from British associationist psychology. Janet held that association could only recall connections previously established by the mind; in other words, association explained automatic functions such as habit and conditioned response. Associations initially are established by mental synthesis, which combines the disparate elements of perception into related wholes. Synthesis is the basis for the integration and interpretation of experience. Beginning at the simplest level of combining sensory data into perceived objects, synthesis works on its own products to build ever higher levels of object, culminating in the concepts of the self and the world. By conceiving the self as the product rather than the origin of synthetic activity, Janet rejected the traditional Cartesian notion of the self as simple and primary; rather, the self is one of the most complex mental phenomena, and it appears only at an advanced stage of mental development. In addition, Janet dismissed the philosophical tendency to characterize abstract reason as the pinnacle of human mental functioning. Influenced by evolutionary theory, Janet insisted that adaptive behavior in the real world required the greatest mental effort and that, in the real world, the social world posed more challenges than the natural world.

Synthesis requires energy, which Janet characterized through a concept of psychological force. An individual's psychological force determines the quantity of psychological stimuli they can hold together in synthetic unity. To describe the quality or level as well as the quantity of synthetic activity, Janet developed the idea of psychological tension. When individuals lose psychological force, they can no longer synthesize all of the stimuli impinging upon them. The inability to integrate perception leads to various kinds of mental illness, which Janet grouped into two major classes. One was hysteria, in which sufferers respond by abandoning or dissociating whole classes of perception. They may lose consciousness and control of parts of their body (producing hysterical paralysis), memory (amnesia), or behavior patterns (tics and fugues). In extreme cases, synthesis continues to operate on the dissociated portions of mental life beyond the awareness or control of the self, producing what was then known as double consciousness and what today would be called multiple personality or dissociative identity disorder.

To describe the other major class of mental illnesses, Janet coined the term 'psychasthenia,' which he used instead of the existing term 'neurasthenia,' to emphasize the psychological rather than biological nature of the condition. Psychasthenics retain consciousness of their uncontrollable thoughts and actions even as they are powerless to stop them. Such individuals suffer from obsessive-compulsive disorders and from feelings of inadequacy related to their inability to control themselves.

Janet attributed the lack of psychological force to two main causes – constitutional deficiency and traumatic emotions. Following in a long hereditarian tradition, Janet held that some

people are born with less psychological force than others. Such individuals are more prone to mental illness, because they are more likely to be overwhelmed by the external world. In addition, traumatic incidents may produce emotions that sap energy and disrupt the normal process of synthesis. Individuals lose the ability and the desire to integrate into consciousness painful memories, which continue to play a disruptive role as dissociated subconscious ideas.

Janet's treatment of patients began with what he called 'psychological analysis,' during which the therapist attempts to separate symptom complexes into their elements and to uncover the origins and development of the ideas behind these elements. Janet used a variety of techniques in psychological analysis. He began his career relying heavily on hypnotism to reveal subconscious memories and dissociated personalities. He supplemented hypnotism with automatic writing and other techniques designed to elicit responses from the patient's subconscious. Such methods had limitations, however. Janet became sensitive to issues of suggestion in the production of hysterical symptoms. Indeed, he was one of the first historians of hypnotism, tracing the origin of the classic hysterical symptoms described by Charcot to the effects of earlier magnetizers on the patients Charcot examined. Janet also found that psychasthenics tended to be less susceptible to hypnotism. He came to rely more on repeated observations of and conversations with the patient over long periods. Through this interaction, Janet was able both to discover the origins of neurotic symptoms and to establish a relationship with the patient that would aid in the therapeutic process.

Once the origins and elements of the symptoms were discovered, Janet attempted to eliminate their pathogenic effects by a variety of methods. By enabling patients to reintegrate dissociated memories into consciousness, Janet was able to reduce their disruptive power. He also tried to dissolve such memories through suggestion and substitution. One case he reported was that of Justine, who had hysterical crises provoked by an intense fear of cholera. Through analysis, Janet discovered that as a child Justine had developed a fear of death by helping her mother, who was a nurse, care for dying patients. This fear was further focused by an incident in which Justine saw the putrefying corpses of two cholera victims. Janet recovered these memories, and then, through an elaborate and gradual process of suggestion, substituted more acceptable images – a Chinese general with the name Cho-le-ra for a diseased corpse. Eventually, the word cholera lost its power, and the symptoms disappeared.

Janet also tried various measures to adapt a patient's psychological force to the demands of her or his environment in order to prevent a recurrence of symptoms and to avoid the development of new illnesses. He used mental exercises to increase the patient's psychological force and recommended reductions in the stresses imposed by the patient's occupation and social life. Though powerless in most cases to change the sufferer's situation, Janet recognized the significance of environment in contributing to the onset of mental illness.

After about 1910, Janet began to redescribe his psychology in terms of conduct rather than ideas. He did this because he came to view the proper goal of psychology as the explanation of the adaptation of the organism to its environment. Ideas are instruments that lead to adaptive action. Janet also expanded and systematized his conception of the hierarchy of psychological levels. In addition, he became increasingly interested in the psychology of religious belief. These new directions came together in his most important late work, *De l'angoisse à l'extase* (1926), based largely on a case study of a highly religious patient with hysterical symptoms.

Janet's influence was considerable at the end of the nineteenth and beginning of the twentieth centuries. His concepts of dissociation and of subconscious ideas found adherents from Carl Jung to William James. Alfred Adler acknowledged his indebtedness to Janet's analysis of feelings of inadequacy. Jean Piaget cited Janet's work on the levels of psychological activity as inspiration for his own developmental theory. In 1906, Janet gave a series of lectures at Harvard Medical School, which made him better known in the US.

However, Janet's influence diminished rather than grew with time. Janet failed to train many followers. He also suffered in competition with Sigmund Freud. Not only was Freud more adept at forming a devoted following, he was more successful at defending the issues over which he differed with Janet, such as the sexual etiology of neuroses and the origin of repressed memories in fantasy. The rise of behaviorism, which despite its focus on action owed little directly to Janet, also eclipsed Janet's own psychology of conduct. In psychiatry, neurological and pharmacological approaches became dominant where Freudian psychoanalysis did not prevail, and in psychology, experimental methods replaced Janet's reliance on psychopathology.

In the last two decades of the twentieth century, a number of psychologists and psychiatrists began to study Janet anew. As Freud's abandonment of the seduction theory was challenged, some found a corrective in Janet's emphasis on actual traumatic events. Theorists who questioned Freud's insistence on the sexual origins of neuroses turned to Janet for a more inclusive view of psychological trauma. The resurgence of dissociative identity disorder as a diagnostic category also led researchers to seek guidance in one of the early students of that condition. Renewed interest in these and other areas to which Janet made substantial contributions assure him an important place in the history of psychology and psychiatry.

See also: Freud, Sigmund (1856–1939); James, William (1842–1910); Jung, Carl Gustav (1875–1961).

Bibliography

Brooks III, J.I., 1998. The Eclectic Legacy: Academic Philosophy and the Human Sciences in Nineteenth-century France. University of Delaware Press, Newark, DE.

Ellenberger, H.F., 1970. The Discovery of the Unconscious: The History and Evolution of Dynamic Psychiatry. Penguin, London.

Janet, P., 1889. L'automatisme psychologique: Essai de psychologie expérimentale sur les formes inférieures de l'activité humaine. Alcan, Paris.

Janet, P., 1893. L'état mental des hysteriques, vol. 2. Rueff, Paris (1977 *The Mental State of Hystericals*, vol. 2. University Publications of America, Washington, DC).
Janet, P., 1898. Névroses et idées fixes, vol. 2. Alcan, Paris.
Janet, P., 1903. Les obsessions et la psychasthénic, vol. 2. Alcan, Paris.
Janet, P., 1907. The Major Symptoms of Hysteria: Fifteen Lectures Given in the Medical School of Harvard University. Macmillan, New York.
Janet, P., 1919. Les médications psychologiques, vol. 3. Alcan, Paris (1925 *Psychological Healing*, vol. 2. Macmillan, New York).
Janet, P., 1926. De l'angoisse á l'extase: Études sur les croyances et les sentiments, vol. 2. Alcan, Paris.
Prevost, C.-M., 1973. La psycho-philosophie de Pierre Janet: Économies mentales et progrés humain. Payot, Paris.
Sjövall, B., 1967. Psychology of Tension. An Analysis of Pierre Janet's Concept of 'tension psychologique.' Together with an Historical Aspect. Svenska Bokförlaget, Stockholm.

Japan: Sociocultural Aspects

Emiko Ohnuki-Tierney, University of Wisconsin, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

With a primary emphasis on the nonlocalized nature of Japanese culture, this article introduces how almost all features of what we call Japanese culture, including shifting self-identities, were born at the interface of local and global forces. It highlights the internal complexities, comprised of the 'Agrarian Japanese' vs the minorities, those along the lines of gender and age, and overseas Japanese and returnees. The Japanese educational system, religions, health care, youth culture, and sports have all undergone transformations at the confluence of the local and global, showing the dynamic nature of Japanese culture and society.

Shifting political boundaries of the nation-state called Japan, the multiethnic composition of its population, and transnational/global flows involving the Japanese, together with recent epistemological and theoretical questions in various fields, have made anthropological studies of Japan highly complex. Anthropology of Japan, practiced by an increasing number of scholars worldwide, is an enormously active field now that anthropologists are no longer specialists of nonliterate peoples. Neither is it a field for Westerners alone; other Asians are actively engaged in it. After introducing the two dominant schools of anthropology in Japan and a history of anthropology of Japan in the United States, the article points to major developments in contemporary anthropology of Japan including: emphases on minorities and other social groups that had been underrepresented in the past studies; studies on the Japanese not bound to the spatial unit of Japan, including the overseas Japanese, various types of returnees to Japan; studies explicating how almost all features of Japanese culture/history were born at the confluence of local and global forces at present (e.g., youth culture) and in the past (influence of the Chinese and Western civilizations).

A sociocultural overview of Japan appears straightforward. Yet, shifting political boundaries of the nation-state called Japan, its multiethnic composition of the population, and transnational/global flows involving the Japanese, together with recent epistemological and theoretical questions in various fields, have reminded us that such an undertaking is a highly complex enterprise. The anthropology of Japan is a testing ground for new approaches to anthropology, including various attempts to historicize anthropology (Ohnuki-Tierney, 1987, 1993). Following is a brief summary of the field identifying important contributions but focusing primarily on scholarship based in Japan, the United States, and the United Kingdom (see also Hendry, 1992; Eisenstadt, 1996).

Anthropology of Japan: In the Beginning

Anthropology *in* Japan may be represented by two major approaches, both called *minzokugaku*, each having a different character for *zoku*, one referring to folklore or folk traditions, and the other, 'people.' 'Folklore' is equivalent to 'ethnography,' rather than 'folklore' as practiced in the United States. The basic approach was founded by Yanagita Kunio (1875–1962), who emphasized the fieldwork aspect of the discipline. Its emphasis on rural and other marginalized populations was later extended to urban populations under Miyata Noboru (1936–2000). Orikuchi Shinobu (1887–1953), another pillar of this school, made extensive use of classic literature as ethnographic texts, thus introducing history to ethnography.

The other major approach is sociocultural anthropology, which began with the influence of the Viennese school, then turned to British social anthropology, and most recently to American cultural anthropology. Most Japanese scholars in Japan in this tradition specialize in non-Japanese cultures. American anthropology in Japan is largely a product of World War II, when the American government commissioned scholars to decipher its enemy, with a most celebrated, and often criticized, case being *The Chrysanthemum and the Sword* (1946) by Ruth Benedict. Immediately after World War II, scholars focused on 'village Japan,' where they sought a more 'traditional' Japan than modernized city life.

Making the Majority and Marginalizing the Internal 'Others'

Wet-rice agriculture, introduced from the Asian continent to Japan around 400 BCE, gradually supplanted the previous culture, called Jōmon, based on a foraging subsistence economy (Kobayashi, 2003); it provided the political, economic, and symbolic foundation for the Yamato state and the imperial system, which lost its political power after the twelfth century, until it was nominally 'restored' through the Meiji Restoration in 1868.

The rice-based agrarian political economy and cosmology provided a lasting impact on the social structure and cosmological tradition, which became *the* Japanese cosmology *sui generis* for all the Japanese, including non-agrarian peoples. This development contributed to the representation of Japan as 'Agrarian Japan,' writing off many social groups from Japan's history and society. We now see a burgeoning interest in 'minorities' and counter discourses within Japanese society. Since the cultural, linguistic, and biological identities of the

Japanese, Koreans, Ainu, and Gilyaks (Nivkh) remain uncertain, and anthropology no longer recognizes 'race' as such, the question of Ainu identities remains controversial. The Ainu engaged in a foraging economy without a writing system but with a highly developed oral tradition (Ohnuki-Tierney, 1981). Subscribing to the scheme of social evolution, the Japanese as well as Westerners were eager to identify the Ainu as 'primitive hunter-gatherers' or even 'archaic Caucasoids.' Their way of life and physical characteristics were exoticized. As they were colonized by the Japanese (Takakura, 1960), they were forced to abandon their way of life and became a minority, the 'other' within the Japanese nation state. Previously occupying the Kuriles, southern Sakhalin, Hokkaido, and northeastern Honshu, the Ainu are now confined mostly to Hokkaido. Their population peaked at 80 000 in the eighteenth century and then dropped to between 25 000 and 50 000 at the end of the twentieth century. Since the 1960s, various Ainu political groups have made great strides in asserting their indigenous rights.

The so-called *hisabetsu-burakumin* is another by-product of the development of 'Agrarian Japan' (Ohnuki-Tierney, 1987). Until the late medieval period, they constituted loosely defined social groups in non-agrarian occupations dealing with culturally defined 'impurity,' including animal and human deaths, and the visual and performing arts. In 1582, they were codified as 'outcaste,' becoming increasingly marginalized in later periods. Despite legal emancipation in 1870, social discrimination has continued into the twenty-first century. The 1987 government census figures are 4603 districts with a total population of 1 166 733, while the Buraku Liberation League suggests at least 6000 districts with a population of 3 000 000. The Okinawans also constitute a minority whose 'distinctiveness,' if there is any, is based primarily on regional differences. Okinawa's past isolation from the main islands and its location as the gateway of influence from outside created some cultural differences, and their dialect was once regarded as non-Japanese.

There are various 'foreign' residents, of which 85% (700 000) are Koreans. Others include Chinese, Westerners, and more recent immigrants from Southeast Asia and the Middle East. In other words, Japan, although stereotyped as demographically homogeneous, has been a multiethnic nation throughout its history (*see* East Asian Studies: Culture).

Although scholarly work on women had previously been available, studies undertaken at the end of the twentieth century have included feminist and other new perspectives, foregrounding their voices, their life experiences, and women as social agents who actively carve out or change courses of their lives (Hara Chieko, Iwao Sumiko, Karen Kelsky, Takie Lebra, and Ueno Chizuko).

'Culture' – Nonlocalized

Another burgeoning genre of scholarly work derives from an emphasis on how 'culture,' or a particular people such as the Japanese, is not localized. Our awareness of the nonisomorphic nature of culture, locality, and people and an unprecedented flow of population across borders of nation states has led to a number of studies of 'transnationalism' and transnationals in the past and in the contemporary context of globalization. These have included studies of overseas Japanese communities, young Japanese who have returned to Japan after being raised overseas due to their parents' occupation (Goodman, 1990), Japanese businessmen overseas (E. Ben-Ari), and emigrants who have returned to Japan, as well as transnational economic activities (Bestor, 2004). Studies of Japanese colonization have become another powerful lens to understand Japanese culture and society, along with Japanese cultural and political nationalisms (K. Yoshino).

'Culture/History' at the Confluence of the Local and the Global

If the political economy and cosmology of rice agriculture were born as a result of the interaction between the global flow and the local development, so there were subsequent major historical watersheds of Japanese history. Japan's 'modern' era was ushered in because of the threat of Western colonialism all over Asia. In 1868, Meiji oligarchs used the 'return of the imperial system,' whose political power had been virtually nonexistent since the end of the twelfth century, as a means of overthrowing the shogunal government, which had proved incapable of coping with Western threats. The Constitution of Imperial Japan, drafted primarily by German and French scholars, with the Prussian constitution as its model, transformed the emperor from an apolitical deity in the Japanese pantheon to a sovereign, Almighty God, and Commander of the military without, however, practical military power. With Japan's defeat in World War II, a new constitution was written by the American occupying forces, creating a 'symbolic' emperor and introducing equal rights for women, coeducation in public schools, and land reform that drastically leveled economic inequality.

Changes in the educational system too are the result of the local–global interactions. With a high literacy rate, less than 4% of students leave school at the age of 15, while almost all the rest complete senior high school education. Some 500 four-year universities and 580 two-year colleges in the 1990s contrast with the prewar system, which consisted of the National Higher Schools and the Imperial Universities, whose graduates, all male, dominated the key posts in the government, business, and intellectual fields. Today, in addition to the University of Tokyo, University of Kyoto, other major national universities, and the top private universities, such as Keiō and Waseda, there are alternative routes to success, for example, the International Christian University or Jōchi University, run by Jesuits, and Chūō University for law, among others. Holders of doctorates from universities in the United States and United Kingdom receive increasing recognition. A major factor in potential curriculum changes in preuniversity education and also in teacher training is the call for 'globalization' or 'internationalization.' The latest push in this direction is the so-called Global 30, a 5-year program for the internationalization of Japanese universities, begun in 2009 and financed by the government; the program increases the number of foreign students, foreign faculty members, and courses given in English. The initiative was adopted by 13 universities, mostly the top universities, although it has encountered difficulties and opposition, and its future remains

uncertain. Despite democratization of the educational system in some sense, the recent affluence has produced more economically based inequality, with wealthy parents sending their offspring to costly 'cram schools' to increase their chances of passing the entrance examination to prestigious universities (Hendry, 1992).

Late twentieth-century debates over the enforcement of the national anthem and the national flag, textbook revisions, and curriculum changes form an ongoing debate about Japan's military past. After World War II, the American occupation forces enforced the well-known 'blacking out' of passages about Japan's aggression from school textbooks. Half a century later, the Japanese face the problem of how to reintroduce the Pacific war experiences, including atrocities committed by the Japanese military, such as the Nanking Massacre, 'comfort women,' and chemical experiments. In this soul-searching process, the Japanese attempt to locate themselves more solidly among other Asians, rather than upholding Westerners as their mirror, as has been the case in the past.

Culture as a product of local–global interactions is well articulated in Japanese religions. Nominally, the Japanese religions consist of native Shintoism and two religions of foreign origin: Buddhism and Confucianism. In addition, the mountain ascetic practice (*shugendō*) continues to be practiced, while numerous 'new religions' came into being often during periods of unrest. At the end of the twentieth century, several new religions enjoy large memberships, including Sōka Gakkai, which also has formed a successful political party. Mahayana Buddhism came from India via China in the late sixth century; it was first embraced by the elite but later became influential among common people. With the rise of political and military nationalism in the early Meiji period, it was targeted as a 'foreign' religion, and many Buddhist temples and works of art were destroyed. In contrast, Shintoism was 'elevated' to be the exclusive national religion (Hardacre, 1989). These changes at the institutional levels, however, had little to do with the common folk who subscribed simultaneously to Shintoism and Buddhism, both of which had become a part of their daily lives. Most rites of passage celebrating life are Shinto rites, while most mortuary rituals are Buddhist. In Japanese folk religiosity, humans and superhumans constitute a continuum – exceptional humans are assigned power over humans who nonetheless manipulate and cajole the superhumans to their advantage. Japanese religions are decisively 'this worldly,' rather than focusing on salvation after life. The major 'functions' of deities and Buddhas are to provide business prosperity, traffic safety, recovery from illness, good health, success in one's undertaking, including high school and university entrance examinations, and other worldly gains (Kitagawa, 1990(1966)).

Christianity was introduced to Japan when Jesuit missionaries arrived in 1549. The converts were persecuted during the sixteenth and seventeenth centuries, when the rulers saw Christianity as potentially subversive of their authority, and again later during the nationalistic period. At the beginning of Meiji period, American Protestantism was introduced and made a profound influence. Some prominent political and intellectual leaders were converted to Christianity and were highly respected by a broad spectrum of people as pillars of moral conscience under the increasingly totalitarian government. Today, Christianity has penetrated into the daily fabric of the Japanese, with 217 'mission schools' at various levels of education.

Numerous practices that fall under the rubric of 'magic' and 'superstitions' are found in profusion in urban centers and are embraced by a large number of Japanese, including the young and the educated, who carry amulets in their cars and hang them on children's rucksacks and hospital beds. Children and high school students hang votive plaques on which they write a prayer, for example, for success in examinations and dating. Young and old alike practice divination of all kinds when choosing a marriage partner, a date for a wedding, a date for opening of a store, and so forth. Although worldly concerns occupy a central place in Japanese religiosity, at its core is a belief in the soul, especially that of the deceased. Various practices associated with souls of the deceased constitute the most important rituals at the end of the twentieth century. These include funerals and later rituals connected with them, as well as the daily offering of water, tea, freshly cooked rice, and favorite foods of the deceased on their ancestral alcove. Municipalities own large cemeteries where the family tombs of the residents are located. During the Bon festival, when the ancestors are supposed to return to this world, and at other prescribed times, families in cars form caravans to visit these cemeteries. This practice continues in intensity despite the almost complete absence of a vision of the afterworld (Smith, 1974). The belief in the soul – the most fundamental of Japanese religiosity – also explains a long history of an enormous importance of the goblins, ghosts, and related figures in Japanese imagination, paving a way to the more contemporary anime, in which the soul of a doll represents the fluid boundaries the Japanese recognize between humans, nonhumans, and inanimate objects (Komatsu, 2010).

With a population of about 123 million, Japan's demographic pattern is characterized by a large proportion of people over 65 (about 20% at the end of the twentieth century), the world's lowest infant mortality rate, and a very low birth rate. The household, consisting of an extended family with its emphasis on generational continuation and the primogeniture system, has been considered the basic social unit, although it was codified only in the Meiji period and outlawed in 1947. This image is misleading, since the 'continuation' does not mean a continuation through blood lines – a high incidence of adoption characterizes Japanese kinship, and the system was most important only for families with substantial wealth and/or important occupation. For most, the nuclear family has been far more important. Since the mid-nineteenth century, industrialization accelerated the migration of nuclear families to urban centers where well over three-quarters of the population is concentrated today, leaving the rural areas primarily in the hands of the elderly. Since the end of World War II, social structure/hierarchy has undergone significant changes (Ishida and Slater, 2010).

Abortion, legalized in 1948 although practiced with high frequency even earlier, and infanticide were formerly the primary means for birth control. At the end of the twentieth century, the condom (80% of contraceptive users) and the rhythm method are used, with the oral contraceptives legalized only in 1999. In spite of the lowest infant mortality rate in the

world, the Japanese population growth is alarmingly low. The physical and psychological involvement of fathers in child rearing has increased in recent years, so they now spend more time with their families, especially on weekends.

While studies of workplaces by anthropologists are by no means neglected, popular culture now receives much attention (A. Allison; J. Callas; I. Condry; M. Creighton; K. Kelsky; P. Noguchi). 'Globalization' is most visible in 'youth culture,' a product especially of American influence. Whether or not they are in the mainstream of society, today's youth flourish in their youth culture, listening to the latest pop music, following the latest youth fashions, and eating out with their friends at popular restaurants. Many color their hair, sometimes blond. They have adopted body piercing after the fashion of American youth, in contrast to the centuries-old Japanese refusal to imitate Chinese ear piercing, which was seen as a form of impurity. Before the last decades of the twentieth century, even married couples did not show their affection in public. At the beginning of the twenty-first century, public places are crowded with young dating couples who openly display their feelings.

Twenty-first century Japanese, young and old, spend a great deal of money and time on leisure activities, including 'traditional' nature viewing – viewing of cherry blossoms, chrysanthemums, and colored leaves. Television and games constitute the major sources of leisure activity. Sports have become primarily spectator sports watched on TV, with baseball and sumo the most popular. Both professional and high school baseball are enormously popular (Kelly, 2009). Sumo, the oldest professional sport in Japan, received a great deal of attention when it broke with tradition by giving the highest ranking of *yokozuna* to foreign-born wrestlers. There are now sumo organizations throughout the world (Tierney, n.d.). One of the most popular participant sports is golf. Domestic and foreign tourism has become a major leisure activity for Japanese of all ages. The young sometimes combine tourism with their passion for consumerism, going to France, Italy, and the United States to purchase 'brand name' merchandise. A 'unique' leisure activity is the reading of comic books, with some comic books having a weekly circulation of over 4 million – the only arena of recreation competing against the electronic media.

The health care system, too, has developed as a result of Japan's interaction with various others. After the introduction of Chinese medicine, Dutch medicine was introduced during the early eighteenth century, followed by the German system. Today most Japanese go to a biomedical doctor for major acute diseases and surgery, while using Chinese medicine for day-to-day health maintenance and chronic illnesses, in addition to going to temples and shrines to pray for speedy recovery (Ohnuki-Tierney, 1984). In 1959–60, all Japanese became legally eligible for pension coverage and health insurance, provided either by one's employer or the government. During the affluent mid-1970s, a national system of free medical care for the aged was introduced, and there was a massive increase in the pension benefits. When the 'bubble' burst, the national scheme of free care for the elderly was discontinued in 1983, and various other contractions were enforced. At the beginning of the twenty-first century, state-of-the-art biomedical technology is available in Japan, although brain death was legally recognized only in 1999, and the Japanese have shown considerable resistance toward organ transplantations in general.

Japan enjoys one of the lowest crime rates in the industrialized world. However, organized crime is a serious problem without any realistic solution in sight. The targets of organized crime are usually highly specific, without threatening the general public. However, periodically, Japan is shaken by outbursts of crime such as that perpetrated by the Aum Shinrikyō cult, which in 1995 spread poisonous gasses in a Tokyo underground station, causing 5000 deaths and injuries.

Conclusions

The critiques of 'totalizing' and 'essentializing' representations of culture/society/people have led the anthropology of Japan in several productive directions, foregrounding diversity, and multiple voices, whereas the awareness of the nonlocalized nature of culture/society/people has stimulated anthropologists to come to terms with various new ways of looking at what indeed is the 'anthropology of Japan.' These new perspectives place Japanese culture and society in the global historical flow and geopolitics, rather than treating them as a hermetically sealed entity.

See also: Meiji Revolution, The.

Bibliography

Benedict, R., 1946. The Chrysanthemum and the Sword. Houghton Mifflin, Boston, MA.
Bestor, Theodore C., 2004. Tsukiji: The Fish Market at the Center of the World. University of California Press, Berkeley, CA.
Eisenstadt, S.N., 1996. Japanese Civilization. University of Chicago Press, Chicago, IL.
Goodman, R., 1990. Japan's International Youth. Clarendon Press, Oxford, UK.
Hardacre, H., 1989. Shintō and the State, 1868–1988. Princeton University Press, Princeton, NJ.
Hendry, J., 1992. Understanding Japanese Society. Routledge, London.
Ishida, H., Slater, D.H. (Eds.), 2009. Social Class in Contemporary Japan: Structures, Sorting and Strategies. Routledge.
Kelly, W., 2009. Samurai baseball: the vicissitudes of a national sporting style. International Journal of the History of Sport 26 (3), 429–441.
Kitagawa, Joseph M., 1990(1966). Religion in Japanese History. Columbia University Press, New York.
Kobayashi, Tatsuo, 2003. Jomon Reflections: Forager Life and Culture in the Prehistoric Japanese Archipelago. Sainsbury Institute, Oxbow Books, London.
Komatsu, K. (Ed.), 2010. Kaiyō Bunka No Dentō to Sōzō (The Tradition and Creation of the Culture of the Goblins). Serika Shobō, Tokyo.
Ohnuki-Tierney, E., 1981. Illness and Healing among the Sakhalin Ainu. Cambridge University Press, Cambridge, UK.
Ohnuki-Tierney, E., 1984. Illness and Culture in Contemporary Japan. Cambridge University Press, Cambridge, UK.
Ohnuki-Tierney, E., 1987. The Monkey as Mirror. Princeton University Press, Princeton, NJ.
Ohnuki-Tierney, E., 1993. Rice as Self. Princeton University Press, Princeton, NJ.
Ohnuki-Tierney, E., 2002. Kamikaze, Cherry Blossoms and Nationalisms: The Militarization of Aesthetics in Japanese History. University of Chicago Press, Chicago.
Smith, R., 1974. Ancestor Worship in Contemporary Japan. Stanford University Press, Stanford, CA.
Takakura, S., 1960. The Ainu of Northern Japan. American Philosophical Society, Philadelphia, PA.
Tierney, R.K., n.d. Wrestling with Tradition: The Spaces and Histories of Sumo.

Jeffreys, Sir Harold (1891–1989)

Arnold Zellner, University of Chicago, Chicago, IL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 7960–7963, © 2001, Elsevier Ltd.

Abstract

An overview of Sir Harold's life, work, awards, honors, and many achievements is presented. Emphasis is placed on his fundamental work in creating and applying an operational axiom system for probability theory that is helpful to scientists and applied workers in their efforts to learn from their data and experience and to create and use models to explain the past and predict, fundamental objectives of science. His reasonable degree of belief definition of probability is operational and permits probabilities to be associated with degrees of belief in alternative hypotheses and models. These degrees of belief can be updated logically by use of Bayes's Theorem to reflect information in new data and used to perform significance tests that Jeffreys developed which are very useful in choosing between or among alternative hypotheses and models. That his estimation, testing, prediction, and other techniques work well using nonexperimental data from geophysics and astronomy, as well as data generated from controlled experiments is of great interest to many, including many social scientists who work mainly with nonexperimental data.

Sir Harold Jeffreys, F. R. S. was born on 22 April 1891 in Fatfield, County Durham, England, and died on 18 March 1989 in Cambridge, England. He was educated at Armstrong College, Newcastle-upon-Tyne and at St. John's College, Cambridge University. After brilliant scholastic performance, he became a Fellow of St. John's College in 1914. From 1946 to 1958, he was Plumian Professor of Astronomy and Experimental Philosophy at Cambridge University. In 1925, he was elected a Fellow of the Royal Society, later became its Bakerian Lecturer and Royal Medalist and received the Copley Medal in 1960. He was knighted in 1953, President of the Royal Astronomical Society, 1955–57, and President of the International Association of Seismology and Physics of the Earth's Interior, 1957–60. In 1962, the Harold Jeffreys Lectureship was created by the Royal Astronomical Society and Jeffreys presented the inaugural lecture in 1963.

The obituary in *The Times* of London, 23 March 1989, noted that Jeffreys's 'famous treatise, *The Earth: Its Origin, History and Physical Constitution*, which with re-editing and improvements, went through seven editions between 1924, when it first appeared, and 1976, did perhaps more than any other book to make the geophysics of the solid earth a coherent study. His researches in seismology, in particular, established his reputation as one of the leading authorities on that subject.'

His early research was centered on astronomical and meteorological dynamics. An accomplished mathematician, he wrote two monographs, entitled *Cartesian Tensorsand Operational Methods in Mathematical Physics*. He and his wife, Dr. Bertha Jeffreys, a very able mathematician, wrote *Methods of Mathematical Physics*, which was published in several editions by Cambridge University Press and became a highly respected and widely cited classic. In a review of the work in *Nature*, it is stated that 'The authors distinguished for their knowledge of mathematical physics and cosmogony, and by their researches in these fields, possess also a remarkably comprehensive and deep understanding of pure mathematics. Their purpose has been to synthesize the mathematical knowledge for the service of physics ... The work is a benefaction to the cause of progress in natural philosophy.'

The Collected Papers of Sir Harold Jeffreys on Geophysics and Other Sciences was published by Gordon and Breach, London, in six volumes. Volume 6 is entitled, *Mathematics, Probability and Miscellaneous Other Sciences*. Work reported in this volume and in his books, *Scientific Inference* (1931 and later editions) and *Theory of Probability* (1939 and later editions) reflected Jeffreys's concerns with the foundations of science and, in particular, scientific inductive procedures. His book, *Theory of Probability* presents an axiom system for inductive inference applicable in all the sciences and operational procedures for description, measurement, modeling, data analysis, estimation, testing, and prediction. Many applications of these procedures to important statistical problems encountered in many sciences are provided in his books and papers.

Since he began research on probability theory after World War I in a series of papers written jointly with Dr. Dorothy Wrinch, his contributions have been of fundamental importance to the philosophy of science, scientific method, and theoretical and applied statistics. The general nature of this research has been stated in the preface of the second edition of *Scientific Inference* as follows: 'The general standpoint that scientific method can be understood if and only if a theory of epistemological probability is provided, remains unaltered. Consequently, I maintain that much that passes for theory of scientific method is either obscure, useless or actually misleading.'

In his book, *Theory of Probability*, Jeffreys presents a theory of epistemological probability, discusses its relevance for work in all areas of science, and applies it to solve many practical, important estimation, testing, prediction, and other statistical problems including those considered by R.A. Fisher, Joey Neyman, Karl Pearson, and others. Irving J. Good, an eminent statistician who reviewed Jeffreys's 1961 edition of *Theory of Probability* wrote that Jeffreys's book '... is of greater importance for the philosophy of science, and obviously of greater immediate practical importance, than nearly all the books on probability written by professional philosophers lumped together.'

As regards to Jeffreys's work on scientific method, he along with Pearson (1938) stressed the 'unity of science principle,' namely, that any area of study, for example, economics,

physics, psychology, etc., can be scientific if scientific methods are employed in analyzing data and reaching conclusions. Or as Pearson (1938, p. 16), cited by Jeffreys states, 'The unity of all science consists alone in its method, not in its material.' With respect to scientific method, Jeffreys, in his book *Theory of Probability*, provides an axiom system for probability theory which is useful to scientists in their efforts to learn from their data, explain past experience, and make predictions regarding as yet unobserved data, fundamental objectives of science, that are applicable in the natural, biological, and social sciences. He states that scientific induction involves observation and measurement and the production of generalizations or models to explain past data and predict new data, a view of induction that is fundamentally different from those of Mach and Popper. Further in induction, deduction plays a role but alone is inadequate for scientific work, mainly because deduction provides just three attitudes toward a proposition, namely, proof, disproof, or complete ignorance. As Hume and others have stressed, there is always the possibility of an exception to a scientific proposition. Thus, Jeffreys saw a need for statements less extreme than those of deductive logic, namely proof, disproof, or complete ignorance. His inductive logic provides such statements that involve reasonable degrees of belief in scientific propositions.

Jeffreys considers a probability to be a measure of the reasonable degree of confidence that an individual has in a proposition given the individual's available background information and data. This individualistic, subjective definition is compared to other definitions, including the axiomatic, long-run frequency, and hypothetical infinite population definitions in Chapter 7 of *The Theory of Probability*. He points out that the first definition is widely employed, for example, in the work of J. Neyman, the second is the Venn limit and strongly advocated by R. Mises, and the third is usually associated with R.A. Fisher although it had appeared earlier in statistical mechanics in the work of Gibbs and others. His criticisms of these concepts of probability are ingenious and devastating and have not been refuted in the literature, perhaps because they are irrefutable. In addition to specific detailed examples illustrating the inadequacies of these three definitions, Jeffreys comments, 'No probability has ever been assessed in practice, or ever will be, by counting an infinite number of trials or finding the limit of a ratio in an infinite series ... A definite value is got on them only by making a hypothesis about what the result would be. The proof even of the existence is impossible. On the limit definition, ... there might be no limit at all ... the necessary existence of the limit denies the possibility of complete randomness, which would permit the ratio in an infinite series to tend to no limit.'

After further detailed criticisms of the above three definitions of probability, he writes, 'The most serious drawback of these definitions, however, is the deliberate omission to give any meaning to the probability of a hypothesis' (1998, p. 377). Also, he remarks, '... in practice no statistician ever uses a frequency definition, but that all use the notion of degree of reasonable belief, usually without even noticing that they are using it and that by using it they are contradicting the principles they have laid down at the outset' (1998, p. 369). See also Jeffreys's (1998 pp. 30–3) discussion of the 'personalistic' or 'moral expectation' views of probability.

Having provided a definition of probability, Jeffreys's axiom system provides procedures for using such probabilities in a coherent way. That is, he provides basic analysis that yields the addition and product rules and other results, including Bayes's Theorem or the principle of inverse probability that Jeffreys states (1998, p. 28) was '... first given by Bayes (1763). It is the chief rule involved in the process of learning from experience.' It is invaluable in using data to revise our initial beliefs in propositions regarding the values of parameters, alternative hypotheses and models, future values of variables, etc. With this analysis, Jeffreys presents a logically consistent framework for learning from data to explain the past and to make predictions about future outcomes, primary objectives of science, as mentioned above. He demonstrates, through analyses of many applied problems using data, that his estimation, prediction, and testing procedures work well in practice and makes many comparisons with results provided by the maximum likelihood approach of Fisher and others, Neyman-Pearson testing procedures, Fisher's p-value approach to testing, etc.

Central in Jeffreys's work on statistical estimation, prediction, and testing problems is Bayes's Theorem. Whatever the estimation problem, use of Bayes's Theorem enables an investigator to combine his prior information regarding possible values of the parameters, represented by a 'prior density' with the likelihood function that incorporates the information in the data to obtain a 'posterior density' for the parameters that incorporates all the prior and sample information. That is, using Bayes's theorem, the posterior density is proportional to the prior density times the likelihood function with the factor of proportionality a numerical constant such that the integral of the posterior density over its range is equal to one. Thus for whatever problem, given the inputs, a prior density and a likelihood function, a posterior density for the parameters is available. This posterior density can be employed to compute the probability that a parameter's value lies between any two given values, say between 0.4 and 0.6, a solution to the famous inverse problem posed by Bayes. Optimal posterior point estimates for parameters are obtained easily from a given posterior density, say a posterior mean, median, or modal value that are optimal in terms of minimizing posterior expected loss *vis à vis* squared error, absolute error, and zero-one loss functions, respectively. Further, in large samples, Jeffreys (1998, p. 193) shows that under general conditions, the posterior density assumes a normal shape centered at the maximum likelihood estimate with a covariance matrix equal to the inverse of the estimated Fisher information matrix (1998, p. 193) and remarks that R.A. Fisher's maximum likelihood methods yields '... results which are indistinguishable from those given by the principle of inverse probability (in large samples), which supplies a justification for it' (1998, p. 194).

Since a suitable prior density is needed to utilize Bayes's Theorem and has been a bone of contention for many years, Jeffreys devoted considerable attention to the problem of formulating appropriate prior densities. In particular, he put forward a reproducible way of formulating invariant prior densities appropriate for use in situations in which an investigator is ignorant, or wishes to proceed as if he were ignorant, about the possible values of parameters. His famous recommendation

in such circumstances is to take the prior density for the parameter(s) proportional to the square root of the determinant of the Fisher information matrix (see pp. 179ff for material on invariant priors). Use of this 'ignorance' or 'diffuse' or 'noninformative' prior density provides invariant posterior probability statements for all continuous one-to-one transformations of the parameters, a remarkable and highly original solution to the choice of prior problem. Jeffreys derived many invariant priors and used them in analyses of applied problems. He is usually given credit for being a pioneer in considering and solving the problem of formulating invariant priors.

With respect to prediction, in general a probability density function for future, as yet unobserved, observations is formulated conditional on parameters with unknown values. If we have, for example, a proper prior or posterior density for the parameters, the product of it and the conditional predictive density can be formed and the parameters integrated out to yield an operational predictive density. It can be used to make predictive probability statements, for example, to compute the probability that a future, as yet unobserved, observation will lie between, for example, 3.4 and 4.7. Also, optimal point predictions can be derived for given predictive loss functions, for example, a mean for quadratic loss functions, etc. For more on uses of predictive densities for forecasting and control, see the general references cited above.

As regards to the evaluation of alternative models or hypotheses put forward to explain empirical phenomena, Jeffreys recommended workers to start with simple models and complicate if necessary. He and Dorothy Wrinch put forward the Simplicity Postulate, namely a statement that simpler models have higher probabilities of performing well. Jeffreys considered all possible models ordered with respect to degree of complexity and assigned probabilities to each, with simpler models receiving higher probabilities in an infinite sequence that is assumed to converge.

Within this broad modeling framework, on considering two alternative hypotheses, Jeffreys used Bayes's Theorem to derive an expression for the posterior odds on these two hypotheses that is equal to the prior odds on the two hypotheses times the Bayes's factor, the ratio of the predictive densities for the observations under the two alternative hypotheses. Jeffreys used posterior odds in his pathbreaking work on significance tests and compared his results for fundamental testing problems with those provided by use of Fisher's p-values and Neyman-Pearson testing procedures. Since Jeffreys's pioneering work on significance testing, many have employed and extended his methods to deal with a wide range of testing and model selection techniques, see, for example, the review paper by Kass and Raftery (1995). Given Jeffreys's brilliant formulation, interpretations, and uses of significance tests based on posterior odds, much improved methods for evaluating and/or combining hypotheses and models evolved and are now in widespread use.

Since Jeffreys's initial work in the 1920s many have come to appreciate his outstanding contributions to the philosophy and practice of science. It is noteworthy that he analyzed much nonexperimental astronomical and geophysical data. Yet his analysis of randomization in experimental design in his book *Theory of Probability* is extremely penetrating and valuable. In addition, he has made important contributions in the areas of distribution theory, asymptotic theory and expansions, robust estimation, invariance theory via differential geometry, information measures for distributions, mixture distributions, contingency tables, a key nonparametric problem, and in modifying the Pierre Simon Laplace Rule of Succession. And last, that Jeffreys's methods work not only in solving physical science problems but also in solving problems in biology, the social sciences, business, law, and other areas is noteworthy. Finally, while many orthodox statisticians were critical of Jeffreys's work years ago, it is the case that his approach and results are currently widely employed and valued.

See also: Bayesian Statistics; Geographic Thought, History of; Probability Theory: Interpretations; Statistics, History of.

Bibliography

Bayes, T., 1763. An essay toward solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London 53, 370–418.
Berger, J.O., 1985. Statistical Decision Theory, second ed. Springer, New York.
Bernardo, J.M., Smith, A.F.M., 1994. Bayesian Theory. Wiley, New York.
Berry, D.A., Chaloner, K.M., Geweke, J.K., 1996. Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner. Wiley, New York.
Box, G.E.P., Tiao, G.C., 1993. Bayesian Inference in Statistical Analysis. Wiley Classics Library, New York.
Cook, A.H., 1990. Sir Harold Jeffreys biographical memoirs. Fellows of the Royal Society of London 36, 303–333.
Jaynes, E.T., 1983. E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics. Reidel, Dordrecht, The Netherlands.
Jeffreys, H., 1931. Scientific Inference. Cambridge University Press, Cambridge, UK. (Reissued with addenda, 1937, second ed. 1957, third ed. 1973).
Jeffreys, H., 1939. Theory of Probability. Oxford University Press, Oxford, UK (second ed. 1948, third ed. 1961, third rev. ed. 1967, Reprinted 1988 and 1998).
Jeffreys, H., 1971. Collected Papers of Sir Harold Jeffreys on Geophysical and other Sciences, vol. 1–6. Gordon and Breach, London.
Jeffreys, H., 1998. Theory of Probability, third rev. ed. Oxford University Press, Oxford, UK.
Kass R.E. (ed.) Tribute to Sir Harold Jeffreys. Chance 4 (2), 1991.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90 (430), 773–795.
Kass, R.E., Wasserman, L., 1996. The selection of prior distributions by formal rules. Journal of the American Statistical Association 91, 1343–1370.
Pearson, K., 1938. The Grammar of Science. Everyman, London.
Zellner, A. (Ed.), 1980. Bayesian Analysis in Econometrics and Statistics: Essays In Honor of Harold Jeffreys. North-Holland, Amsterdam.
Zellner, A., 1996. An Introduction to Bayesian Inference in Econometrics. Wiley Classics Library, New York.
Zellner, A., 1997. Past and Recent Results on Maximal Data Information Prior Distributions in Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers. Edward Elgar, Cheltenham, UK.

Jewish Diaspora

Sergio DellaPergola, The A. Harman Institute of Contemporary Jewry, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Demography of world Jewry demonstrates the unique interplay that exists in the long run between the general concept of Jewish population and the normative concept of Jewish peoplehood. This article outlines the evolution of Jewish population worldwide between 1700 and 2013 covering population growth (besides the dramatic collapse caused by the Shoah), geographical distribution (between and within continents), life cycle events (births, deaths, and marriage), migrations (international and urbanization), conversions and assimilation, and the emerging age composition. The evolution of Jewish population globally reflects cultural, socioeconomic, and legal determinants and constraints experienced under minority and majority status.

Database and Conceptualization

The Jewish Diaspora is widely scattered and typically represents a small minority of the total population. The very definition of Jewishness may be subject to different interpretations. Therefore studying the demographic profile and trends of the Jews is subject to lack of uniformity of available sources. A singular overlap emerged all along Jewish history between the general concept of *Jewish population* and the normative concept of *Jewish peoplehood*. In the former sense, size and characteristics of Jewish population were significantly affected over time by the willingness, ability, and sometimes feasibility of individuals to identify as part of the given group. In the latter sense, the Jews' collective and shared belonging was often deeply affected by trends and events involving the demographic sphere, such as sudden quantitative decline because of major epidemics, massive murder, or enforced conversions; more rarely increase because of large-scale affiliation or high birthrates; or cultural–contextual change associated with large-scale migration movements.

Jewish demographic trends cannot be assessed without first addressing the empirical definition of *who is a Jew*. Figures reported below mostly relate to the concept of *core Jewish population*, i.e., *all those who, when asked, identify themselves as Jews; or, if the respondent is a different person in the same household, are identified by him/her as Jews*. This is an intentionally comprehensive approach, reflecting both subjective feelings and community norms and bonds. Definitions are admittedly looser in the Diaspora than in Israel where personal status is subject to ruling by the Ministry of Internal Affairs. The *core Jewish population* broadly overlaps but does not necessarily coincide with the *halakha* (Rabbinic) definition of a Jew as someone who *is the child of a Jewish mother or converted by appropriate religious and legal procedure*. Inclusion in *core* Jewish population statistics does *not* depend on any measure of a person's Jewish commitment or behavior in terms of religiosity, beliefs, knowledge, community affiliation, or otherwise. The *core* Jewish population includes all those who converted to Judaism, or decided to join the Jewish group and declare themselves Jewish. It excludes those of Jewish descent who have formally adopted another religion, as well as other individuals who did not convert out but currently refuse to recognize their Jewishness (DellaPergola, 2013).

Concurrently, the concept of *enlarged Jewish population* includes the sum of (1) the *core* Jewish population, (2) all other persons, Jewish by birth or parentage who do *not* currently identify as Jews, and (3) all the respective *non-Jewish* household members (spouses, children, etc.). The *enlarged* Jewish population is by definition significantly larger than the *core* population. The *Law of Return* – Israel's distinctive legal framework for the eligibility and absorption of new immigrants – further extends its provisions to all current Jews, their Jewish or non-Jewish spouses, children, and grandchildren, and the respective spouses. As a result of its three-generation time perspective and extended family definition, the Law of Return applies to a much wider population than *core* and *enlarged* Jewish populations alike. The Law of Return, per se, does not affect a person's Jewish status. While it defines clear-cut normative rules for the attribution of certain rights and prerogatives, the initiative for being entitled to its provisions stems from the people's individual awareness or willingness to belong to the Jewish collective, or to have some Jewish ancestry. Since the second half of the twentieth century, cultural identities have tended to become an increasingly important factor in determining the boundaries, size, and demographic characteristics of Jewish population.

Two main interacting drivers crucially shaped Jewish population size and characteristics in the longer run: external forces out of the control of Jewish communities and mostly operating in the global and regional political sphere; and internal forces stemming from the peculiar culture and social structure of the Jewish collective itself (DellaPergola, 2002). A further significant role was played by the changing modes of interaction between Jewish minorities and the majority of society, under rules and constraints usually dictated by the latter. At least five more specific determinants deeply affected the different desirability, feasibility, and availability of socio-demographic events among Jews, their frequency compared to others, and trends over time at the community level:

- unique *traditional culture and community organization*, or the minority's religious and social norms relevant to individual

demographic and social events, as well as collective frameworks and institutions apt to implement or facilitate those norms;

- legal provisions or – more relevant to contemporary history – subjective perceptions concerning a minority's *dominance/dependence* status vis-à-vis the majority of society or other minorities within it;
- *social class stratification*, and the consequent inter- and intra-group differences in perceived interests and access to resources;
- available *knowledge* on the given socio-demographic process, whether acquired through formal education or other channels;
- *biological constraints* of genetic or other nature, in relation to inherited properties affecting exposure to health risks or chances for survival. In this respect, it is essential to stress the difference between contemporary understanding of the human genome, scientifically grounded and prejudice-free, and anti-Jewish racialist approaches of a previous generation.

Demographic trends also very much draw on nondemographic factors such as the general state of world and regional economies, global cultural change, the role of religion, available technology, and political and military affairs. The combination of these factors and their interactions with demography determined the size and characteristics of Jewish populations in specific locations – hence the broader Jewish profile globally.

Jewish Population Growth 1700–1939

Until the late seventeenth century, poor health among populations generally, and inferior access to civil rights among Jews particularly, rendered population growth slow and unstable. When these negative constraints began to weaken, thanks especially to improved nutrition and to early emancipatory steps, world Jewry started to grow. Compared to a rough estimate of 1.1 million Jews in 1700, an estimated total of 2.5 million was attained around 1800. While until the seventeenth century overall Jewish population growth had been slower than that of world population, during the eighteenth and nineteenth centuries, world Jewry increased at a comparatively faster pace. Jewish population reached 4¾ million in 1850, and between the 1880s and 1914 yearly growth rates of 1.6–1.8% were attained by Jews globally. In 1900, world Jewish population was estimated at 10.6 million, a fourfold increase over 1800, and in 1914 it reached 13.5 million. Close to World War I Jewish growth rates began to slow down, but the absolute increase in Jewish population continued to be substantial. By the eve of World War II, the Jewish population had reached a historical maximum of 16.5 million. The annual rate of growth diminished to about 0.8%, while the world's total grew from 1920 to 1940 by 1% annually. Whereas in the nineteenth century the relative growth of the Jews had largely exceeded that of the peoples of Europe, North America, and the less-developed regions, this was no longer true in the period between the two World Wars.

The demographic expansion just outlined was not evenly spread across the different regions of Jewish settlement in the Diaspora, but mostly affected Jewish communities in Eastern Europe (see Figure 1). Jewish population growth rates of over 2% a year prevailed in Eastern Europe during the last quarter of the nineteenth century, surpassing Europe's faster growth rates at that time (1.5–1.7% annually for England, Poland, and Russia between 1880 and 1910). The rapid emergence of Eastern Europe's communities as the largest centers of Jewish life can be appreciated by noting that whereas around 1650 they possibly consisted of less than 350 000 people, by 1900 – also counting Jewish emigrants from Eastern Europe to Western countries and their descendants – they had grown to an estimated 7.5 million.

Jewish population growth outside Eastern Europe was slower. A moderate pace of internal growth characterized most

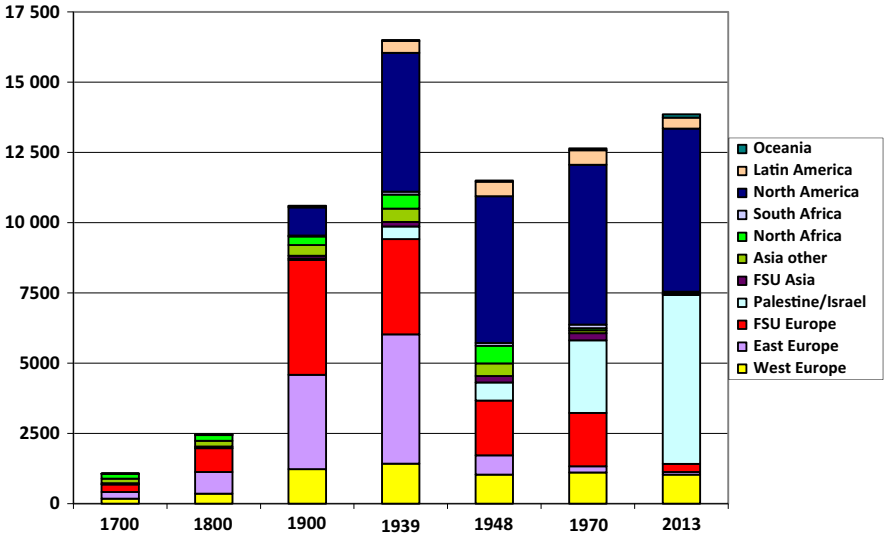

Figure 1 Jewish population by major regions, 1700–2013.

Jewish communities in Western and Central Europe, where early low growth rates had been induced during the time of Jewish seclusion into ghettos. In North Africa and the Middle East, the slower development of societal modernization and higher death rates hindered population growth, and initial demographic expansion of Jewish communities did not take place until the last decades of the nineteenth century or the early decades of the twentieth century.

One major consequence of these regional gags was a radical change in the geographical distribution of the Jews. During the late Middle Ages, the Jewish presence had been more or less equally divided between the areas of influence of Christianity and the Islam. The demographic rise of the Jewish population in Eastern Europe during the eighteenth and nineteenth centuries gave it overwhelming predominance in the context of world Jewry. Around 1880 – before the beginning of major intercontinental migrations – Jews in Eastern Europe (including the Balkans) represented about 75% of the world total, up from an estimated 52% around 1700. Jewish communities in North Africa and Asia (including Palestine's small *yishuv* = Jewish community), while stagnating or slowly growing in absolute numbers, declined in relative share from about 35% of world Jewry in 1700 to 8% in 1880. Western and Central European Jewry kept their share of the total (14% in 1880 vs 13% in 1700), while new Jewish communities were emerging in America (3% of world Jewry in 1880 vs less than 1%, mostly in the Caribbean area, in 1700).

As a result of large-scale international migrations, many veteran Diaspora communities in Eastern and Central Europe, the Middle East, and North Africa shrank and became small and precarious, or virtually disappeared. Instead, two major demographic centers arose: the United States, where the final consolidation of the Jewish community occurred mainly in the period 1881–1924, and Palestine, which absorbed significant immigration between World Wars I and II but whose definitive demographic expansion followed the establishment of the state of Israel in 1948. Several secondary Jewish population centers emerged: in Western Europe, especially in France and England; in Canada, alongside the major Jewish Diaspora population in the United States; in South America, especially in Argentina; in South Africa and Australia. The more spectacular change in the geographical distribution of world Jewry was the increase in the number of Jews in America, from about 250 000 in 1880 to 1 175 000 in 1900 and nearly 3 500 000 in 1914, and from 3 to 11% and, by 1914, 25% of world Jewry. This rapid expansion of American Jewry reflected the migration of about 2 400 000 Jews between 1881 and 1914, of whom 2 150 000 went to the United States and Canada.

By converse, the relative share of Europe's Jews among total world Jewry continued to decline from about two-thirds in 1914 to 58% in 1939. On the other hand, the proportion of American Jews among world Jewry grew from about a quarter in 1914 to a third in 1939. But because of the U.S. immigration quotas which very strongly affected the Jews, after 1924 relative Jewish population growth became much greater in Latin America and South Africa, although the initial numbers were much smaller. On the other hand, there was a rise in the relative share of Jews in North Africa and, especially, in Asia. The main reasons were considerable Jewish migration to Palestine/Eretz Israel, to the Asian territories of the USSR, and also to Egypt, and an upward swing in the natural increase of the local Jewish population thanks to the eventual onset of mortality decline.

Jewish Population Trends 1939–2013

World Jewish population was estimated at 16.5 million in 1939, of which 9.5 million (58%) was in Europe, and was reduced to 11 million in 1945, of which 3.8 million (35%) was in Europe. The total loss of 5.7 million Jews in Europe reflects a somewhat higher number of direct and indirect victims of Shoah victims, minus some natural increase that occurred in countries not subject to enemy's persecution and annihilation. If the Shoah had not occurred, Jewish population was expected to grow by at least another 6 million by the end of the twentieth century (DellaPergola, 2006c). But instead, between 1939 and 1945, Jewish population diminished by about 400 000 in Western and Central Europe, and by about 5.3 million in Eastern Europe and the Balkans (including Turkey, with most of its Jewish population located on the European side of the Bosporus). Jewish population between 1939 and 1945 globally declined by 33%, in Europe by 60% – 30% in Western Europe and 65% in Eastern Europe.

The countries where Jews suffered the most during the war period were Poland (97%), followed by Greece (89%), Austria and Czechoslovakia (−88%), Yugoslavia (−84%), Germany and the Netherlands (−77%), the Baltic countries (−74%), Belgium (−66%), Byelorussia (−61%), Hungary (−55%), and Ukraine with Moldova (−51%). These figures point to the exceptionally tragic incidence of Shoah in countries all across occupied Europe's in Eastern Europe, the Balkans, Central and Western Europe. The Jewish population in several other European countries, including the major Soviet republic of Russia and other Western European countries suffered somewhat to quite lower losses. A few Western areas like the United Kingdom, the Scandinavian countries, the Iberian Peninsula, and Switzerland added some Jews through absorption of refugees or internal growth. Moreover in 1945 there remained a significant bulk of over 200 000 displaced persons, nearly double the amount on the eve of the war.

The process of reconstruction of European Jewry that started immediately after World War II brought about gradual normalization of the more acute cases of displacement, return to normal family life, and a temporary substantial increase in the low Jewish birthrate that had prevailed since the 1930s. Much broader Jewish population redistribution involved the emigration to Israel of many Shoah survivors and the massive transfer thereof large numbers of Jews from Islamic countries prompted by violence that accompanied Israel's war of Independence. Subsequently, Israel continued to constitute the main country of access for Jewish migrants from Eastern Europe – that in the meantime had mostly fallen under the control of the Soviet Union. In turn, with the advancement of decolonization especially in North Africa, large-scale Jewish migration impacted the geography of European Jewry. The net consequence during the years 1945–70 was a further reduction of European Jewry by nearly 600 000 (−15%), but while Western Europe gained on balance 168 000 (+18%), Eastern Europe lost 736 000 (−26%). The European country that

benefited the most was France, with a nearly threefold growth (Bensimon and DellaPergola, 1984), with moderate gains of 10–15% in the United Kingdom, Austria, Italy, and Spain. On the other hand Jewish population diminished in Eastern Europe by as much as −91% in Poland, −84% in Romania and Bulgaria, −67% in Czechoslovakia, −61% in Hungary, −42% in Yugoslavia, and −33% in Germany. The Soviet Republics in Europe were more stable when some of those displaced during the war resettled while emigration to the West and Israel was only at its incipient stages.

Still regarding Europe, the more recent period (1970–2013) was dominated by significant emigration from the Soviet Union after the June 1967 war, the collapse of the USSR, and the massive exodus from what now was the Former Soviet Union (FSU). The Jewish population in Europe diminished by 1.8 million (−56%), from over 3.2 million in 1970 to just above 1.4 million in 2013 (10% of world Jewry). Decline was about 100 000 in Western Europe (−7%) and over 1.7 million in Eastern Europe (−82%). Looking at individual countries, the most affected were Byelorussia and Ukraine with Moldova (−92%), Romania (−87%), the Baltics (−83%), Russia (−76%), Bulgaria (−71%), Poland (−64%), Turkey (−56%), Czechoslovakia (−54%), Yugoslavia (−46%), Hungary and Greece (−31%). Among Western countries, the United Kingdom and Ireland underwent a decline of −26%, and several other countries (France, Italy, Switzerland) lost around or more than 10%. The main beneficiary of population transfers was Germany whose Jewish population grew nearly fourfold, with smaller growth in Spain and Portugal, Austria, and Scandinavia.

Of the 11 million Jews left in the world in 1945, about 564 000 lived in Palestine; in May 1948, at Israel's independence, there were about 650 000. The state of Israel became the main outlet of Jewish international migration. After the initial inflow of 690 000 immigrants in 1948–51, hence nearly doubling the initial Jewish population, the intensity of immigration waves that reached the country tended to decline over time, until the major upsurge of 1990–91 when 376 000 immigrants mostly from the FSU arrived. Israel's Jewish population grew rapidly, reaching 1 million in 1949, 2 million in 1962, 3 million in 1976, 4 million in 1991, 5 million in 2001, and 6 million in 2013. Mass immigration from Europe, Asia, and Africa was the major determinant of growth until the early 1960s. In later years, with the reduction of Jewish international migration after the mass exodus of Soviet Jewry, natural increase became the predominant factor of Israel's Jewish population increase.

Table 1 summarizes the main changes in Jewish population size and distribution between 1948 and 2013, when the world's Jewish population was assessed at 13 855 000, of which 6 014 000 were in Israel (43% of world Jewry). The fast growth in Israel's population was matched by massive decline in the Middle East, North Africa, the FSU, and other countries in Eastern Europe. Significant decline occurred also in Latin America and South Africa. Oceania is the region out of Israel that experienced the fastest Jewish population growth, while North America experienced lesser relative growth but larger absolute population increase. Western Europe experienced initial growth followed by minor decline.

Reflecting continuing Jewish population increase in Israel and stagnation or continuing decline among Jewish communities in the rest of the world, over 90% of world Jewry lived in five countries in 2013, of which 82.6% in the two largest of Israel and the United States, and 98.7% lived in 22 countries with at least 15 000 Jews (Table 2). Excluding Israel from the count, over 90% of Diaspora Jews lived in seven countries, of which 69.2% in the United States alone and 97.8% lived in 21 countries. Besides the two major Jewish populations (Israel and the United States), comprising over 5 million persons, another seven countries each had more than 100 000 Jews each – three in Western Europe (France, the United Kingdom, and Germany), one in Eastern Europe (the Russian Federation), one in North America (Canada), one in South America (Argentina), and one in Oceania (Australia). The dominance of Western countries in global Jewish population distribution is a relatively recent phenomenon and reflects the West's relatively more hospitable socioeconomic and political circumstances vis-à-vis the Jewish presence. The growth, or at least

Table 1 Jewish population by major regions: number, percent distribution, and percent change, 1948–2013

	Number (thousands)			*Percent*			*Percent change*		
Region	*1948*	*1970*	*2013*	*1948*	*1970*	*2013*	*1948–1970*	*1970–2013*	*1948–2013*
Israel	650	2582	6014	5.7	20.4	43.4	297	133	825
West Europe	1035	1113	1054	9.0	8.8	7.6	8	−5	2
East Europe	715	222	91	6.2	1.8	0.7	−69	−59	−87
FSU Europe	2000	1897	270	17.4	15.0	1.9	−5	−86	−87
FSU Asia	225	254	20	2.0	2.0	0.1	13	−92	−91
Other Asia	365	100	20	3.2	0.8	0.1	−73	−80	−95
North Africa	630	83	4	5.5	0.7	0.0	−87	−95	−99
South Africa	105	124	71	0.9	1.0	0.5	18	−43	−32
North America	5215	5686	5805	45.3	45.0	41.9	9	2	11
Latin America	520	514	385	4.5	4.1	2.8	−1	−25	−26
Oceania	40	70	120	0.3	0.6	0.9	75	71	200

FSU, Former Soviet Union.
Source: Adjusted from DellaPergola, S., 2009. International migration of Jews. In: Ben-Rafael, E., Sternberg, Y. (Eds.), Transnationalism: Diasporas and the Advent of a New (Dis)order. Brill, Leiden, Boston, pp. 213–236.

Table 2 Countries with largest core Jewish populations, 1 January 2013

			Percent of total Jewish population			
			In the world		In the Diaspora	
Rank	Country	Jewish population	%	Cumulative %	%	Cumulative %
1	Israel[a]	6 014 300	43.4	43.4	[b]	[b]
2	The United States	5 425 000	39.2	82.6	69.2	69.2
3	France	478 000	3.5	86.0	6.1	75.3
4	Canada	380 000	2.7	88.8	4.8	80.1
5	The United Kingdom	290 000	2.1	90.9	3.7	83.8
6	Russian Federation	190 000	1.4	92.2	2.4	86.3
7	Argentina	181 500	1.3	93.5	2.3	88.6
8	Germany	118 000	0.9	94.4	1.5	90.1
9	Australia	112 500	0.8	95.2	1.4	91.5
10	Brazil	95 200	0.7	95.9	1.2	92.7
11	South Africa	70 000	0.5	96.4	0.9	93.6
12	Ukraine	65 000	0.5	96.9	0.8	94.4
13	Hungary	48 000	0.3	97.2	0.6	95.1
14	Mexico	40 000	0.3	97.5	0.5	95.6
15	Belgium	30 000	0.2	97.7	0.4	96.0
16	The Netherlands	29 900	0.2	97.9	0.4	96.3
17	Italy	28 100	0.2	98.1	0.4	96.7
18	Chile	18 500	0.1	98.3	0.2	96.9
19	Switzerland	17 400	0.1	98.4	0.2	97.2
20	Turkey	17 300	0.1	98.5	0.2	97.4
21	Uruguay	17 200	0.1	98.6	0.2	97.6
22	Sweden	15 000	0.1	98.7	0.2	97.8

[a]Includes Jewish residents in East Jerusalem, the West Bank, and the Golan Heights.
[b]Not applicable.
Source: Adjusted from DellaPergola, S., 2013. World Jewish population 2013, In: Dashefsky, A., Sheskin, I. (Eds.), American Jewish Year Book. Springer, Dordrecht.

slower decrease, of Jewish population in the more developed Western countries is matched by a higher share of Jews in a country's total population. The share of Jews in a country's total population tends to be related to the country's level of development, with most Jews concentrated in large, highly developed, central metropolitan areas. In 2013, more than half of the world Jews lived in five metropolitan areas, in descending order: Tel Aviv, New York, Jerusalem, Haifa, and Los Angeles. By adding another 10 areas (South Florida, Be'er Sheva, San Francisco, Washington/Baltimore, Boston, Chicago, Paris, Philadelphia, London, and Toronto), 75% of the world Jewish population were covered.

Regarding core Jewish populations in 2013, the share of Jews out of the total population in Israel (including Jews in East Jerusalem, the West Bank, and the Golan Heights, but excluding Palestinians in the West Bank and Gaza) was 75.4%. If also including the Palestinians in Gaza and the West Bank, the extant Jewish majority shrank to 50.5% of total inhabitants. Israel obviously is a special case, but it has become also quite a developed country (sixteenth by the UN Index of Human Development (2013)). Elsewhere Jews represented 17.3 per 1000 of total population in the United States; 3.9 per 1000 on average in the other seven countries with over 100 000 Jews; 0.8 per 1000 on average in the other nine countries with over 18 000 Jews; and virtually nil in the remaining countries which comprise the overwhelming majority of world population. In 2013, 95 countries had at least 100 Jews, but only 17 had more than 25 000. In 29 countries out of the 95, there was at least one Jew per 1000 of total population. The relationship between Jewish population size and its share of a country's total population is summarized in Table 3.

Table 3 World Jewish population distribution, by number and proportion (per 1000 total population), 1 January 2013

Number of Jews in country	Jews per 1000 total population			
	Total	0.0–0.9	1.0–4.9	5.0+
Jewish population distribution				
Total number[a]	13 854 800	296 200	1 130 500	12 427 600
Total percent[a]	100.0	2.1	8.2	89.7
100–4999	0.5	0.5	0.0	–
5000–24 999	1.3	0.5	0.7	0.1
25 000–99 999	3.0	1.2	1.8	–
100 000 or more	95.2	–	5.6	89.8
Number of countries				
Total	95	66	22	7
100–4999	61	56	4	1
5000–24 999	17	7	9	1
25 000–99 999	8	3	5	–
100 000 or more	9	–	4	5

[a]Grand total includes countries with fewer than 100 Jews, for a total of 500 Jews. Minor discrepancies are due to rounding. Israel includes Jewish residents in East Jerusalem, the West Bank, and the Golan Heights.
Source: Adjusted from DellaPergola, S., 2013. World Jewish population 2013, In: Dashefsky, A., Sheskin, I. (Eds.), American Jewish Year Book. Springer, Dordrecht.

International Migration

Among the mechanisms of Jewish population change, international migration crucially shaped the geographical distribution of the Jewish Diaspora across history and especially since the second half of the nineteenth century. About 9.2 million Jews moved between continents or across different geo-cultural areas between 1880 and 2012 (see summary in Table 4). Most Jewish migrants came from Eastern Europe, North Africa, and the Middle East. Over the 133 years estimated here, 59% went to Western countries, primarily the United States, and 41% to Palestine/Israel.

The continuous wavelike pattern in Jewish international migration (see Figure 2) demonstrates recurrent crises affecting the position of Jewish communities in different parts of the world and the consequent need for prompt relocation, which in turn suggests dependency of the Jewish Diaspora on changing political, socioeconomic, and cultural circumstances in the respective countries. Some of the stimulating factors of large-scale emigration operated in recent years essentially like they did at the turn of the twentieth century in spite of the deep intervening sociopolitical changes across world society and Jewish communities alike.

Three major Jewish migration waves were related to: (1) the pre–World War I mass migration from Eastern Europe to North

Table 4 Estimated Jewish international migration, by major areas of origin and destination – absolute numbers and percent distribution, 1881–2012

Areas of origin and destination	*1880–1918*	*1919–1948*[a]	*1948*[a]*–1968*	*1969–2012*	*Total*
Total (thousands)	2400	1615	1881	3277	9173
Yearly average (thousands)	63	55	91	76	69
Total (percent)	100	100	100	100	100
From Eastern Europe[b]	97	86	33	52	65
To Western countries	95	65	6	21	45
To Palestine/Israel	2	21	27	31	21
From Asia–Africa[b]	3	5	52	15	18
To Western countries	2	2	15	3	5
To Palestine/Israel	1	3	37	12	13
From Palestine/Israel[c]	0[d]	3	10	17	8
From Western countries[e]	0[d]	6	5	16	8
Regional subtotals:					
To Western countries	97	70	31	41	59
To Palestine/Israel	3	30	69	59	41

[a]May 15 (Israel's independence day).
[b]Since 1990, Asian regions of FSU included in Asia–Africa.
[c]To Western countries.
[d]Less than 1%.
[e]To Palestine/Israel.
Source: Adjusted from DellaPergola, S., 2009. International migration of Jews. In: Ben-Rafael, E., Sternberg, Y. (Eds.), Transnationalism: Diasporas and the Advent of a New (Dis)order. Brill, Leiden, Boston, pp. 213–236.

America, peaking in fiscal years 1905–06 (154 000 Jews admitted), 1906–07 (149 000), and 1913–14 (138 000); (b) the massive transfer to Israel of displaced Shoah survivors and of sizable Middle Eastern and North African Jewish communities between 1948 (peaking with 239 000 in 1949) and 1951; and (c) the more recent exodus from the FSU (228 000 in 1990 and 195 000 in 1991). These major moments of crisis or postcrisis adaptation occurred at comparable intervals of about 40–45 years, much in response to major events in the global polity (pre–World War I and post–World War II adjustments, the end of the Soviet Union). While no mechanical or deterministic assumption is suggested about the alignment of global geo-political forces, economists and demographers have discussed the possible cyclical unfolding of economic change and its impact on national political systems (Kuznets, 1958). Specific consequences for Jewish ethno-religious or social groups tended to stimulate mass migration. In between the major migration waves, several other significant waves appear, also quite regularly spread over the time axis, although each wave had its specific determinants and geographical origins.

The volume of the main Jewish migration flows is summarized in **Table 4** according to a fourfold geographical typology: Eastern Europe; Muslim countries in Asia and Africa; Western countries including America and Western Europe; Palestine/Israel, under Ottoman rule until 1918, under British mandate until 1948, and since 1948 the state of Israel. Until 1918, Eastern Europe was the main sending region and the United States the main recipient. During the interwar period, Eastern Europe continued to be the main area of origin but migration was sparser and Palestine emerged as a major country of destination. Since 1948, Israel became the dominant country of destination, especially between 1948 and 1968 (69% of total migrants) when the plurality of Jewish migrants came from Muslim countries in Asia and Africa. Since 1969 Eastern Europe again became the supplier of the majority of Jewish migrants, and Israel received 59% of the world total. In the balance of migration flows between Western countries and Israel, the former ended up with a moderate net surplus.

Jewish migration to Israel, usually called by the value-oriented Hebrew term of *aliyah* (*ascent*) consistently followed a wavelike pattern (Bachi, 1977; see **Figure 3**). Historiography

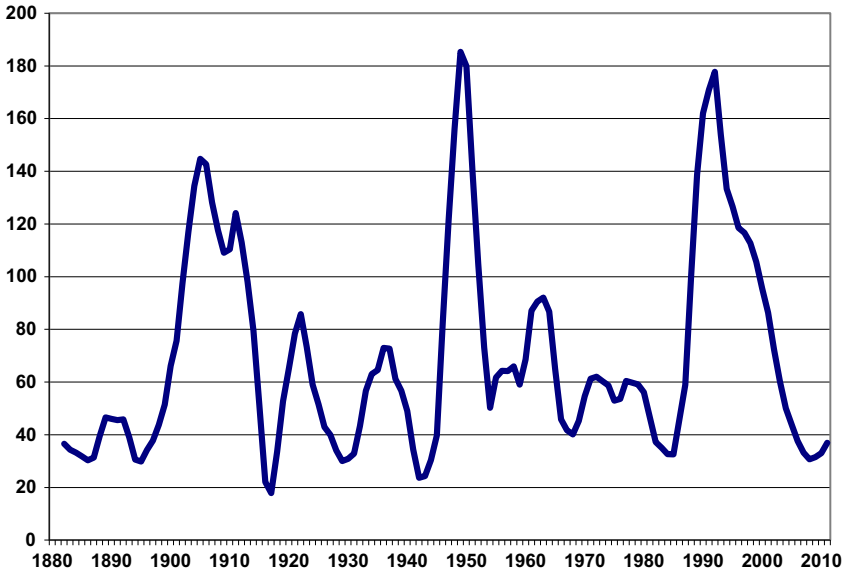

Figure 2 International migration of Jews, 1880–2011 – 5-year moving averages. Source: adjusted from DellaPergola, S., 2009. International migration of Jews. In: Ben-Rafael, E., Sternberg, Y. (Eds.), Transnationalism: Diasporas and the Advent of a New (Dis)order. Brill, Leiden, Boston, pp. 213–236.

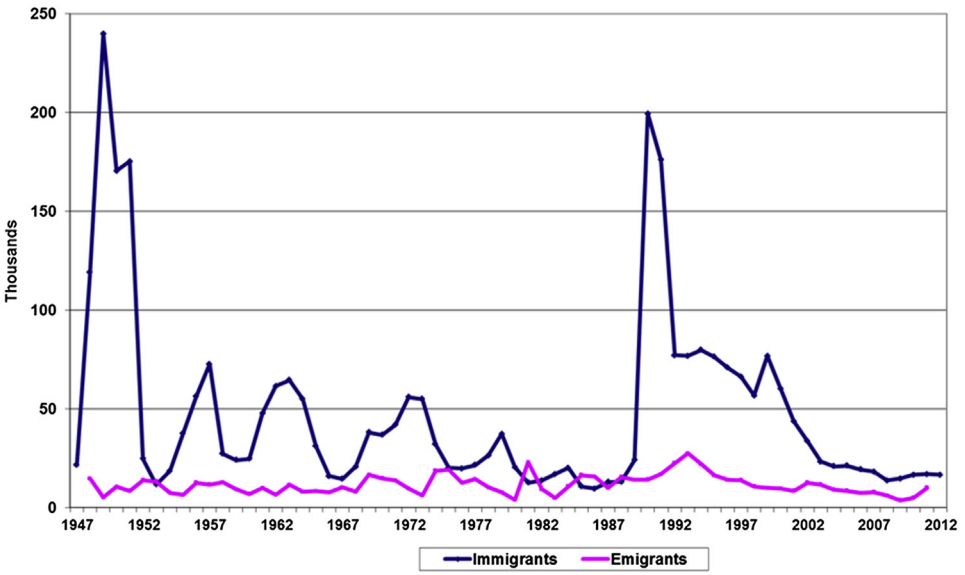

Figure 3 New immigrants and emigrants, Israel 1947–2012, thousands. Source: Israel Central Bureau of Statistics, annual.

has numbered such migration waves serially, starting with the 'first' (1881–1903) through the 'fifth *aliyah*' (1932–38). Similar cyclical patterns actually prevailed throughout the nineteenth century, though with relatively small numbers of migrants, and clearly continued throughout and beyond Israel's first 50 years of independence. In absolute numbers, the greatest single wave occurred in 1948–51, followed by others of declining size until the major surge of the early 1990s. However, relative to a rapidly expanding receiving population, the highest and quite unique impact of immigration on the receiving society was recorded during the 'fourth *aliyah*' (283 immigrants per 1000 Jewish residents in 1925), followed by Israel's 'great *aliyah*' (236 per 1000 in 1948 and 26 per 1000 in 1949) and the 'fifth *aliyah*' (201 per 1000 in 1935). The major wave of 1990–91 only generated 50–60 per 1000 yearly increases to a much larger receiving Israeli population.

Jewish migration was stimulated by a combination of political, socioeconomic, and cultural factors. Negative circumstances in the countries of origin determined the observed wavelike patterns. The volume and direction of Jewish migration largely reflected policies in the countries of origin and destination. Against limiting quotas enforced by the United States and other countries in the period following World War I and until the end of World War II, Israel since 1948 offered Jews nearly unrestricted immigration, socioeconomic incentives, and citizenship rights. Consistently with a comparatively less amount of return migration, the demographic and socioeconomic characteristics of Jewish migrants were generally less selective than among total international migrants. Despite its heterogeneity and initial low status, migration led to the growth and development of contemporary Israeli society and of the major Jewish communities worldwide and stimulated Jewish social mobility (Metzer, 1998; Schmelz et al., 1991).

Assistance provided by Jewish international and Israeli agencies generally enhanced migration. Various bodies promoted, directed, supported, and subsidized Jewish migrants, such as the Jewish Colonization Association, the World Zionist Organization and Jewish Agency for Israel, the American Jewish Joint Distribution Committee, the Hebrew Immigrant Aid Society, and Nativ, the liaison bureau of Israel's Government in the FSU.

Throughout the major part of the twentieth century, the overwhelming majority of world Jewish population was located in countries with high emigration propensities, but emigration significantly emptied those reservoirs. With the recent predominance of a Jewish presence in developed and free countries with low emigration propensities, short of significant changes in the current global and regional circumstances, the volume of Jewish international migration was bound to decline.

Vital Statistics and Population Composition

The historical transition from high to low birthrates and death rates mostly happened earlier among Jews than among surrounding total populations. Jews in spite of their obviously lower social status were among the forerunners of demographic change in modern societies, along with other small minorities of European society like the aristocracy and members of the urban *bourgeoisie*. The peculiarities of the nineteenth-century Jewish demography (lower death rates coupled with lower birthrates) were noted by scientists and writers who were developing their theories about *human races*. But instead of correctly interpreting these differentials in the light of cultural, social, and economic differences at the community level, these differences were attributed to alleged genetic characteristics. The myth of a special and dangerous *Jewish race*, able to survive under any climate and prone to multiply much faster than other populations, would lead to a quest to limit or altogether delete the Jewish presence, with catastrophic consequences for the Jews (Schmelz, 1971; DellaPergola, 2006a).

Lower death rates reflected significant differences in the incidence of various causes of death among Jews and non-Jews

in the same countries. According to the nineteenth- to early twentieth-century data, Jews suffered comparatively less from smallpox, typhus, cholera, tuberculosis, venereal disease, and excessive use of alcohol. They were reportedly more affected by diabetes, respiratory diseases, genetically transmitted disabilities, and some types of mental illness. Early declines in infant mortality contributed to higher population growth but also induced earlier fertility reductions among the Jews.

During the nineteenth and early twentieth centuries, higher birthrates, inasmuch as they existed, reflected nearly universal marriage and comparatively high marital fertility levels. With no exception, however, Jews anticipated by many years the respective surrounding populations in adopting more stringent control over reproduction, through postponing marriage and more effective birth control, and sometimes also through applying Jewish restriction about sex relations. This occurred true in a variety of environments in Western, Central, and Eastern Europe, in North Africa, as well as in communities generated by intercontinental migration, such as in the United Kingdom, North and Latin America. Conversions and other passages to and from Jewry, on the other hand, generally caused net losses to the Jewish population – hence a significant factor of attrition in long-term Jewish population growth.

After World War II, a 'baby boom' occurred especially in the United States, followed by renewed decline in Jewish birth. Jewish populations in Europe reached a state of demographic stagnation and decline, with deaths consistently outnumbering births (Table 5). This negative gap was strikingly high in the FSU European republics. In the Russian Federation in 2000, there were only 600 recorded Jewish births compared to over 8200 recorded Jewish deaths – a net loss of 7600. Such striking deficit reflects extreme population aging (see below), in part the consequence of the intensive emigration of younger Jewish adults and nuclear families with the consequence that large numbers of elderly remained behind in the FSU. In the United States, Jewish vital statistics are not directly available. However Jewish population projections based on available age composition and cautious assumptions about the age-specific frequency of motherhood and deaths suggest that the core Jewish population generates annually about 50 000 births and 58 000 deaths. The likely deficit of about 8000 is being compensated by a positive Jewish immigration balance.

Israel is the only exception to these recessive demographic trends. Steady immigration produced a doubling of Israel's

Table 5 Jewish vital statistics in selected countries, 1988–2012

Country and year	*Births*	*Deaths*	*Difference*
Russian Federation			
1988	3710[a]	13 826	−10 116
2000	613[b]	8218	−7605
The United Kingdom			
1991	3200	4500	−1300
2000	2786	3660	−874
2007	3313	2878	+435
2010	–	2734	–
Germany			
1990	109	431	−322
2002	151	1000	−849
2010	168	1081	−913
2012	199	1282	−1083
The United States[c]			
2011	50 000	58 000	−8000
Israel			
1990	73 851	25 759	+48 092
2000	91 936	33 421	+58 515
2010	120 763	33 948	+86 815
2012	125 409	35 919	+89 490

[a]Births to Jewish mothers, of which 2148 are to non-Jewish fathers. Assuming as many births to Jewish fathers and non-Jewish mothers, the total births would be 5858.
[b]Births to Jewish mothers, of which 444 are to non-Jewish fathers. Assuming as many births to Jewish fathers and non-Jewish mothers, the total births would be 1057.
[c]Estimated from age composition and age-specific survivorship rates.
Source: Tolts, M., 2004. The Post-Soviet Jewish population in Russia and the world. Jews in Russia and Eastern Europe 52, 37–63; Vulkan, D., Graham, D., 2008. Population Trends among Britain's Strictly Orthodox. Board of Deputies of British Jews, London; Zentralwohlfahrtsstelle der Juden in Deutschland, 2013. Mitgliederstatistik der jüdischen Gemeinde und Landesverbände in Deutschland für das Jahr 2012. Frankfurt a.M; Israel Central Bureau of Statistics. Annual. Statistical Abstract of Israel. Jerusalem and author's projections.

Table 6 World Jewish population, by country and recent intermarriage frequency, 2010

Current rate of Jews marrying non-Jews (%)[a]	*Country*[b]	*Jewish population*	
		Thousands	*Percentage*
	World	13 578	100.0
0–0.9	West Bank[4]	290	2.1
1–4.9	Israel[1], Yemen[4]	5414	39.9
5–14.9	Mexico[1], Gibraltar[4], China[4], Iran[4], Syria[4], North Africa[4]	56	0.4
15–24.9	Bahamas[4], Costa Rica[4], Guatemala[2], Panama[4], Venezuela[1], India[3], Japan[4], Singapore[4], South Africa[3]	101	0.7
25–34.9	Canada[1], Chile[2], Central and South America not elsewhere reported[4], Turkey[2], Asia not elsewhere reported[4], Africa not elsewhere reported[4], Australia[1], New Zealand[3]	540	4.0
35–44.9	Argentina[3], Brazil[2], Uruguay[2], France[1], the United Kingdom[1], Western Europe not elsewhere reported[3]	1144	8.4
45–54.9	The United States[1], Italy[2], the Netherlands[1], Switzerland[1], FSU in Asia[3]	5520	40.7
55–74.9	Austria[1], Germany[1], Eastern Europe (excluding FSU)[3]	216	1.6
75+	FSU in Europe[2], Cuba[3]	298	2.2
Average rate			
48.1	Diaspora		
29.1	World		

[a]Not Jewish at time of marriage. Intermarriage estimates are national or regional estimates and ignore geographic variations within countries. Demographers refer to this intermarriage rate as the 'individual intermarriage rate.' Rates are for marriages in the past 5–10 years.
[b]Data quality rated as follows: 1: Recent and reliable data; 2: Partial or less recent data of sufficient quality; 3: Rather outdated or incomplete data; 4: Conjectural.
Source: Adjusted from DellaPergola, S., 2009b. Jewish out-marriage: a global perspective. In: Reinharz, S., DellaPergola, S. (Eds.), Jewish Intermarriage around the World (pp. 13–39). London, New Brunswick: Transaction.

Jewish population between 1970 and 2004, which was reinforced by a significant Jewish natural increase. In 1990, 73 900 Jewish births and 25 800 Jewish deaths produced a natural increase of 48 100. In 2012, 125 400 Jewish births and 35 900 Jewish deaths produced a net increase of 89 500. Demand for children continues to be strong among both the religious and the secular populations, rooted partly in Jewish communal identity and partly in a broader sense of economic optimism and life satisfaction, and resulting in significantly larger families in Israel than among Jews in other countries.

Low Jewish birthrates and population aging in the Diaspora are significantly impacted by high and continually increasing rates of intermarriage (Table 6). Overall, the rate of intermarriage has been increasing among Jews, but significant differences persist by country. In recent years, in the Russian Federation about 70% of recently married Jewish women and 80% of recently married Jewish men married non-Jews. In the United States, and in several medium-size European Jewish communities, the intermarriage rate was over 50%; in France and the United Kingdom, over 40%; in Canada and Australia, over 30%; and in South Africa and Venezuela, over 15%. Of the major Jewish communities, probably only Mexico had an intermarriage rate lower than 15%. The incidence of intermarriage significantly depends on the ethno-religious composition of parents: most of the total increase in intermarriage occurs among Jewish adults who are themselves the children of intermarried parents (Phillips, 2013). In Israel, the rate of intermarriage is less than 5%, low but not negligible, reflecting the growing number of non-Jews immigrated under the Law of Return, particularly from the FSU. Many of these intermarriages are performed in Cyprus (Dvorin, 2006). On average, based on 2010 global Jewish population distribution and recent intermarriage rates in different countries, about 29% of all recently married Jews worldwide, and 48% of all recently married Jews in the Diaspora, started a new family with a non-Jewish partner. Scattered data on cohabitation among young Jewish adults point to much higher rates of intermarried couples (DellaPergola, 2009b).

A further factor is the Jewish identity of the children of intermarriages. The percentage of the children of intermarriage raised as Jews during the early 1990s was about 20% in both the

Table 7 Selected Jewish populations, by main age-groups, 1970–2011

Country[a] *and type*	*Year*	*Total*	*0–14*	*15–29*	*30–44*	*45–64*	*65+*	*Median age*[b]
Traditional								
Ethiopia	1991	100	**51**	20	13	11	5	14.7
Transitional								
Israel	1970	100	**30**	27	16	20	7	26.1
Iran	1976	100	**30**	28	19	17	6	25.7
Mexico	1991	100	24	**27**	20	22	7	29.3
Israel	2011	100	**26**	22	19	21	12	31.5
Moderate aging								
The United States	1970	100	22	24	17	**26**	11	35.5
Canada	1971	100	21	25	15	**27**	12	34.2
Greater Paris	1975	100	21	25	18	**25**	11	34.1
The United Kingdom	1990	100	19	19	**26**	19	17	37.6
South Africa	1991	100	19	**22**	22	21	16	36.1
Venezuela	1998	100	24	19	21	**24**	12	36.0
Mexico	2000	100	20	24	**25**	24	7	33.6
Advanced aging								
Russian Federation	1979	100	8	15	21	**31**	25	49.2
Switzerland	1980	100	18	18	19	22	**23**	41.1
Lithuania	1993	100	14	15	18	**27**	26	47.7
Hungary	1995	100	14	18	19	23	**26**	44.4
Brazil	2000	100	16	19	22	**25**	18	40.3
The United States	2001	100	16	20	19	**26**	19	41.5
The United Kingdom	2001	100	16	17	19	**26**	22	44.2
Canada	2001	100	18	19	18	**27**	18	40.7
Turkey	2002	100	10	16	22	**34**	18	47.2
Greater Buenos Aires	2004	100	15	21	16	**28**	20	43.3
Australia	2006	100	17	18	18	**28**	19	42.6
Italy	2009	100	15	16	19	**27**	23	44.9
New York, 8 counties	2011	100	18	18	15	**27**	22	45.0
Terminal								
Romania	1979	100	5	11	10	34	**40**	59.1
Ukraine	2001	100	6	10	14	35	**35**	56.4
Russian Federation	2002	100	5	11	14	33	**37**	57.5
Germany	2009	100	7	14	15	29	**35**	54.3

[a]Countries sorted by date within type. The largest age-group in each population is in bold font.
[b]The median is the age in reference to which one-half of the population is younger and one-half is older. Some of the median ages were computed from more detailed age distributions than those shown here.
Adjusted from DellaPergola, S., 2012. World Jewish population 2012. In: Dashefsky, A., Sheskin, I. (Eds.), American Jewish Year Book. Springer, Dordrecht, pp. 213–283.

United States and the Russian Federation. In 2001, this percentage had increased in the United States to more than one-third, still far from the 50% that would be required so as not to contribute to a decrease in the number of Jews. Nonidentification with Judaism of many children of intermarriages, added to low fertility levels, produces an even lower 'effective Jewish birthrate.'

Age composition plays a crucial role in population change (Table 7). Data for 1975–2012 exemplify extreme variations in age composition following the transition from higher to lower birthrates and death rates. Jewish populations can be classified into five types, gradually moving from traditional to transitional, moderately aging, advanced aging, and terminal. Traditional Jewish populations, characterized by very high percentages of children, have virtually disappeared. Jews in Ethiopia before their mass immigration to Israel in 1991 were the last surviving example. The transitional type occurs as fertility is controlled and mortality declines. Such populations still feature relatively high percentages of children, an increasing share of adults, and a median age around 30 years or under. Israel in 2012 provided the only persisting example of a Jewish population where the percentages in each major age-group regularly decreased from the younger age-groups to the older age-groups. In moderately aging communities, the center of gravity moves to age 45–64, but children under 15 years of age are still more numerous than adults aging 65 and over. This type, with a median age about 35–40 years, predominated during the 1970s and through the 1990s in the United States, and still persisted in some communities in Central and South America like Mexico. More recently, Jewish communities in the United States – namely in New York, in Canada, in Western and Central European countries, in Central and South American countries like Argentina and Brazil, as well as in Australia and Turkey, joined the advanced aging type. In these populations, persons aged 65 and over outnumber children under the age of 15 years, and median ages mostly range between 40 and 45 but also tend to approach 50. The terminal age composition pattern is typical of the Russian Federation, other FSU republics, Germany, and several other Eastern European countries. It comprises percentages of elders that are double or more the percentage of children, with a median age of 50 or higher, eventually tending toward 60 years and over. Such populations are bound to continuous decline.

See also: Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Ethnicity and Migration in Europe; Global Migration; Immigration Policy; Immigration, Citizenship, and Integration: Social Work Connections; Immigration, Emigration, and Citizenship; Immigration: Political Aspects; Migration History; Migration and Development; Migration and Social Capital; Migration, Ethnicity, Aging and Social Work Practice; Migration: Anthropological Perspectives; Migration: Cultural Aspects; Migration: Motivations; Migrations, Colonizations, and Diasporas in Archaeology.

Bibliography

Avni, H., 1991. Argentina and the Jews: A History of Jewish Immigration. The University of Alabama Press, Tuscaloosa, AL.

Bachi, R., 1977. The Population of Israel. The Hebrew University, and Paris: CICRED, Jerusalem.

Bensimon, D., DellaPergola, S., 1984. La Population Juive de France: Socio-démographie et Identité. The Hebrew University, and Paris: Centre National de la Recherche Scientifique, Jerusalem.

DellaPergola, S., 1998. The global context of migration to Israel. In: Leshem, E., Shuval, J.T. (Eds.), Immigration to Israel: Sociological Perspectives. Transaction, New Brunswick, pp. 51–92.

DellaPergola, S., 2002. Demography. In: Goodman, M. (Ed.), The Oxford Handbook of Jewish Studies. Oxford University Press, Oxford, pp. 797–823.

DellaPergola, S., 2006a Demography, Encyclopedia Judaica, second ed., vol. 5 Jerusalem, pp. 553–572.

DellaPergola, S., 2006b Migration, Encyclopedia Judaica, second ed., vol. 14 Jerusalem, pp. 207–219.

DellaPergola, S., 2006c. Between science and fiction: notes on the demography of the holocaust. Holocaust and Genocide Studies 10 (1), 34–51.

DellaPergola, S., 2009. International migration of Jews. In: Ben-Rafael, E., Sternberg, Y. (Eds.), Transnationalism: Diasporas and the Advent of a New (Dis) order. Brill, Leiden, Boston, pp. 213–236.

DellaPergola, S., 2009b. Jewish out-marriage: a global perspective. In: Reinharz, S., DellaPergola, S. (Eds.), Jewish Intermarriage around the World. Transaction, London, New Brunswick, pp. 13–39.

DellaPergola, S., 2009c. Fertility Prospects in Israel: Ever below Replacement Level? United Nations Expert Group Meeting on Recent and Future Trends in Fertility. United Nations Secretariat, Department of Economic and Social Affairs, Population Division, New York.

DellaPergola, S., 2010. Jews in europe: demographic trends, contexts, outlooks. In: Schoeps, J., Ben-Rafael, E. (Eds.), A Road to Nowhere? Jewish Experiences in Unifying Europe? Brill, Leiden-Boston, pp. 3–34.

DellaPergola, S., 2011. Jewish Demographic Policies: Population Trends and Options in Israel and in the Diaspora. The Jewish People Policy Institute, Jerusalem.

DellaPergola, S., 2012. World Jewish population 2012. In: Dashefsky, A., Sheskin, I. (Eds.), American Jewish Year Book. Springer, Dordrecht, pp. 213–283.

DellaPergola, S., 2013. World Jewish population 2013. In: Dashefsky, A., Sheskin, I. (Eds.), American Jewish Year Book. Springer, Dordrecht.

DellaPergola, S., Rebhun, U., Tolts, M., 2000. Prospecting the Jewish future: population projections 2000–2080. American Jewish Year Book 100, 103–146.

Dvorin, N., 2006. Marriages of Israelis abroad and the role of former Soviet Union immigrants. Megamot 44 (3), 477–506.

Friedlander, D., Goldscheider, C., 1979. The Population of Israel. Columbia University Press, New York.

Hersch, L., 1931. International migration of the Jews. In: Wilcox, W. (Ed.), International Migration, vol. 2. National Bureau of Economic Research, New York, pp. 471–520.

Israel Central Bureau of Statistics. Annual. Statistical Abstract of Israel. Jerusalem.

Kuznets, S., 1958. Long swings in the growth of population and in related economic variables. Proceedings of the American Philosophical Society 102, 25–52.

Lestschinsky, J., 1960. Jewish migrations, 1840–1956. In: Finkelstein, L. (Ed.), The Jews: Their History, Culture and Religion, vol. 2. Harper, New York, pp. 1536–1596.

Metzer, J., 1998. The Divided Economy of Mandatory Palestine. Cambridge University Press, Cambridge, UK.

Phillips, B.A., 2013. New demographic perspectives on studying intermarriage in the United States. Contemporary Jewry 33 (1–2), 103–120.

Schmelz, U.O., 1971. Migrations. In: Encyclopaedia Judaica, vol. 16. Keter, Jerusalem, pp. 1518–1529.

Schmelz, U.O., DellaPergola, S., Avner, U., 1991. Ethnic Differences among Israeli Jews: A New Look. The Hebrew University and American Jewish Committee, Jerusalem.

Sicron, M., 1957. Immigration to Israel, 1948–1953. Falk Project for Economic Research in Israel and Israel Central Bureau of Statistics, Jerusalem.

Tolts, M., 2004. The Post-Soviet Jewish population in Russia and the world. Jews in Russia and Eastern Europe 52, 37–63.

United Nations, Human Development Programme, 2013. Human Development Report 2013: The Rise of the South: Human Progress in a Diverse World. New York.

Vulkan, D., Graham, D., 2008. Population Trends among Britain's Strictly Orthodox. Board of Deputies of British Jews, London.

Wischnitzer, M., 1948. To Dwell in Safety: The Story of Jewish Migration since 1800. Jewish Publication Society, Philadelphia.

Zentralwohlfahrtsstelle der Juden in Deutschland, 2013. Mitgliederstatistik der jüdischen Gemeinde und Landesverbände in Deutschland für das Jahr 2012. Frankfurt a.M.

Job Analysis and Work Roles, Psychology of

Heiner Dunckel, International Institute of Management, University of Flensburg, Flensburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Job analysis is essential for understanding and changing working conditions and the workers themselves. Analyzing jobs is a crucial element in (psychological) work analysis because jobs and the tasks constituting them are the basic (design) units of an organization and the key interface between human beings and organization. This article briefly outlines job analysis methods and techniques, distinguishing them according to the criteria: unit of analysis, intended applications, theoretical foundation, method users and methods of data collection, methodological standards, and results. Emphasis is given – particularly in view of practical requirements – to the need for methods and techniques to be scientifically proven in terms of their reliability and validity. Finally, the future importance of job analysis and the necessary further developments are discussed.

Introduction

Job and work analysis is essential if we are to describe, understand, and change the way people work and their work environment. It is important because it is the basis for the solution of different human resource problems (Brannick et al., 2007). Job and work analysis also plays an important role in the European Union, as the European Framework Directive 89/391/EEC issued in 1989 legally commits every employer to conduct regular measures to encourage improvements with regard to the safety and health of workers at their workplaces. Generally speaking: If we wish to solve work-related problems, work analysis is not normally a matter of choice; we are either obliged to perform it explicitly or at least base our decisions on implicit work analysis information.

Job analysis is part of work analysis. It is necessary (see also Ghorpade, 1988) because

- jobs and their constituent tasks are the basic (design) elements of an organization, forming the basis for the constitution of workgroups, departments, and other larger divisions;
- jobs are consciously designed and changed by people (e.g., industrial engineers) who do not normally perform them;
- modern organizations are constantly changing (job analysis allows a continuous adjustment of job-related requirements to changing organizational goals).

There is a large number of jobs (e.g., in the production, office/administration, and service sectors) and elements of the work situation (e.g., organizational structures, agreements on working hours, work objects, and means of work), and a wide variety of ways in which job-related information can be used, that practically preclude the possibility of adopting a universal approach or analysis method here. Accordingly, there is no one job analysis method, but rather a number of systems, techniques, and methods that can be used to analyze jobs (Dunckel, 1999; Gael, 1988).

Nor are there any universally valid definitions, terms such as job, position, task, and element being used differently in different countries and approaches. This fact should be taken into account by the reader, for this article, too, uses terms in a way that would certainly not be endorsed by all job analysts.

Job analysis is the systematic, empirically oriented, problem-based collection of information on the interaction between humans and the work tasks and conditions that constitute a job or are connected with it "… by dividing it [the job] into smaller units, where the process results in one or more written products with the goal of describing what is done in the job or what capabilities are needed to effectively perform the job" (Brannick et al., 2007: p. 8).

Depending on the respective problem or goal (e.g., organizational planning and design, human resource management), different approaches and different techniques and methods are normally required (Dunckel, 1999; Gael, 1988; Ghorpade, 1988). A distinction should be drawn between job analysis and job evaluation: job evaluation in the narrower sense aims to determine compensation levels for a job (Landy, 1993); in a broader sense, job evaluation is also the evaluation of jobs according to given criteria, e.g., the extent to which jobs are conducive to health or personality development.

Jobs are composed of tasks, i.e., the same jobs are characterized by largely identical tasks. Task analysis is therefore a key element in job analysis (Dunckel and Resch, 2010; Landy, 1993). Task is generally taken to mean the purpose of a work system or the goal to be achieved under given conditions. Tasks specify goals to be attained by the worker, taking into account different overall conditions. Elements of jobs are therefore defined as work tasks if they relate in each case to different goals (more detailed treatment in Oesterreich and Volpert, 1987; Ghorpade, 1988). Tasks are the points in a job at which the organization meets concrete individuals and makes concrete demands on the activities of these individuals. The task is thus the 'interface' between the organization and the individual and an essential element in psychological job analysis.

Psychological job analysis systematically captures and evaluates information on a working individual's activity. Here, it is concerned not only with the externally visible activity and the conditions surrounding it, but also with the psychological processes (e.g., thinking, motivation) and structures (e.g., memory) determining and regulating this activity. Psychological job analysis thus focuses on the concrete work activity as a psychologically regulated activity – i.e., the activity of a working individual.

Unlike 'classical' work studies or technical (job) analyses, psychological job analysis is inconceivable without the involvement of the concrete working individual concerned. Supplying the affected workers with comprehensive information on the aims and objectives of an analysis, taking into account the fears and anxieties caused by it, and providing feedback to the workers on the results are (or should be) essential elements of any psychological job analysis.

Classification Criteria

Job analysis methods and techniques can be distinguished on the basis of various criteria (also McCormick, 1976): unit of analysis, intended applications, theoretical foundation, method users and methods of data collection, methodological standards, and results.

Unit of Analysis

The job analysis unit is the job itself and the tasks that comprise it. For this reason, many job analysis systems relate directly to the tasks and the immediate work conditions. More comprehensive job analysis systems (Dunckel, 1999; Gael, 1988) take into account the fact that essential features (e.g., potential for defining goals self-reliantly) and results of the tasks only become comprehensible if the overall organizational conditions and the knowledge, skills, and attitudes of the workers are taken into consideration. They therefore extend the analysis to include both the organizational conditions (e.g., degree of responsibility, codetermination potential, values and norms, leadership climate) and the working individuals with their respective knowledge, skills, and attitudes.

Methods can be classified, based on widely accepted criteria, according to whether they are more job/task-oriented or worker/person-oriented (Brannick et al., 2007). Job-oriented analysis is concerned with analyzing tasks, job conditions, or job features, without regard for the concrete individuals involved and their different knowledge, skills, abilities, and attitudes. Person-oriented analysis, on the other hand, is centered on the person and specifically concerned with differences among individual workers in the perception, interpretation, and performance of the job. Typical job-oriented approaches are: functional job analysis, task inventories, health services mobility approach (Gael, 1988; Ghorpade, 1988); typical person-oriented approaches are: position analysis questionnaire, critical incident technique, ability requirement scales, cognitive task analysis (Gael, 1988; Ghorpade, 1988).

The distinction between job-oriented and person-oriented analysis is of conceptual and practical significance. In stress research, for example, it is conceptually important to begin by determining stress factors independently of the person involved (and their individual perceptions and coping behavior) in a job-oriented manner, then going on to examine how objectively identical stress factors are differently perceived and dealt with by different individuals and how they affect different people. However, the (person-oriented) analysis of these interindividual differences means first determining objective stress factors because we are concerned here with interindividual differences in relation to objectively identical stress factors. This distinction is of practical significance when planning future workplaces for which there are not yet any workers. Here, the job-oriented approach is the only way of obtaining job analysis information.

Intended Applications

Techniques and methods have been and are being developed for different goals and potential applications. Job analysis literature contains numerous different lists of goals and proposed applications for job analysis information (Ash, 1988; Lees and Cordery, 2000). Basically speaking, two complexes of goals can be distinguished: work/job and organization design, and personnel development. More concrete aims of job analyses are, for instance, comparing work activities, changing and planning the work situation and organization, determining skill requirements and factors defining aptitude requirements, technology assessment, maintenance of industrial health and safety standards, and job evaluation.

Job analysis methods and techniques play an important part in human resources management and personnel selection (Algera and Greuter, 1998; Brannick et al., 2007). They are used to determine more precisely the requirements a person must meet to perform his or her work tasks. These requirements may be specified as tasks to be accomplished, behavioral requirements (e.g., required behavior or behavioral repertoire), eligibility requirements (e.g., knowledge and skills), or trait requirements (e.g., abilities and interests). For each of these requirement types, there are a number of techniques available (Gael, 1988).

If the results of such analyses are combined, job analysis can also be used to describe (work) roles as defined by Katz and Kahn (1978) (see also Dierdorff and Morgeson, 2007). According to these authors, roles are standardized behavior patterns demanded of all persons involved in a given functional relationship. Here, job analysis can also help to identify when and on what conditions role expectations (of the different organization members with respect to individual workplace occupants) can lead to conflicts and job stress.

The intended applications of job analysis are not necessarily independent of one another. If, for example, the goal is to change the work organization, it will often be necessary to analyze both the technological implications and the consequences for the workers involved.

Besides the application purpose, the application area also plays an important role. Job analysis techniques and methods consider different levels and units of an organization. A distinction must be made between

- sectors (e.g., industrial, administrative, service);
- levels of an organization (e.g., enterprise as a whole, business division, department, workplace group, workplace, work task);
- professional groups (e.g., executives, specialized professional groups); and
- activity classes (e.g., assembly, control and monitoring activities, administrative activities, service activities).

Depending on the specific concern, the emphasis will be on different information. Job analysis, too, faces the problem of breadth versus depth. The more detailed the information, the

more limited the application purpose. It may therefore be a good idea to proceed in several steps, starting with rough analyses to determine the key analysis areas, which are then analyzed in greater detail.

Theoretical Foundation

The theoretical foundation largely determines which information is captured at which level or with which analysis unit.

Work studies as defined by Taylor or Gilbreth (see Ghorpade, 1988) are based on an additive *movement* (or motion) model. Techniques working on this basis thus attempt to define elementary movement units (e.g., grasping with the hand), combining these additively in order to then determine, say, the standard time required to perform the work.

Approaches rooted in behaviorism also attempt to identify elementary units (of behavior) (e.g., processing materials, recognizing optical differences). Their units are bigger and hence they are therefore not movement-oriented but *behavior*-oriented.

These approaches, however, fail to take into account the fact that movements merge to form 'wholes,' are integrated in complex webs of activity and the regulating mental processes and representations, and are codetermined by these. This is why many developments, especially in Germany, are based on the '*action regulation* theory' (Frese and Sabini, 1985; Oesterreich and Volpert, 1987), thus giving priority to questions relating to the psychological regulation of action, the level of psychological regulation, the completeness of actions, the degrees of freedom (Hacker, 1998), or the scope for action or decision. The guiding idea here is that of humane work, i.e., work geared to human strengths and enabling individual workers to perform their job under permanently tolerable conditions, without impairment of their well-being and in a manner conducive to their personal development.

In the emphasis they place on characteristics such as scope for action, variability, identity, and importance of the task, action theory approaches are in keeping with the traditions of industrial sociology, e.g., the work of Turner and Lawrence (1965) and the work of Hackman and Oldham (1975) that build on this. Since the latter approaches are specifically concerned with questions relating to the 'motivation potential of work,' they can also be classified as oriented to *motivation theory* (*see* Work Motivation).

Besides drawing on approaches based on behavior, action, and motivation theory, job analysis systems also have recourse to concepts of stress theory (*see* Workplace Stress), ergonomics, and human engineering (*see* Human Factors and Ergonomics); in addition, worker-oriented approaches draw on concepts of personality theory.

Method Users

The main users of job analysis methods are the workers themselves, first-level supervisors, higher-level supervisors, job analysts, technical experts and other company experts, but also works and staff councils.

Whether and to what extent a method can be used depend, among other things, on the application requirements that must be met by the users. Important factors here, besides formal qualifications, are the amount of experience the users need in analysis techniques, whether they can teach themselves how to use such techniques or whether special training is required.

Methods of Data Collection

There are also fundamental differences between techniques in terms of the data collection methods used, for example:

- interview methods (e.g., individual and group interviews, technical conferences with experts, more or less structured questionnaires and checklists);
- observational methods (e.g., direct and indirect observation, continuous observation, work sampling);
- analyses of company data (e.g., working hours lost, accident statistics, workplace descriptions);
- analyses of documents (e.g., file analyses, form analyses); and
- work activities performed by job analysts.

Each of the methods has its advantages and disadvantages. It is therefore a good idea to combine several methods (Brannick et al., 2007; Gael, 1988). These advantages and disadvantages can be highlighted by comparing interview and observation methods.

Interview Methods

It is a good idea to interview workers, for they are the ones who know their own work activities best. Furthermore, such interviews are indispensable when the workers' subjective assessment of the work is needed or psychological processes are to be evaluated because these can only be accessed directly by introspection. In addition, interviews, especially questionnaires, are frequently the method of choice because they are relatively easy to develop and use. Interview methods are the most frequently used job analysis technique.

This should not, however, obstruct our view to the fact that interview methods have a number of weaknesses. Some typical problems are:

- comprehension problems of workers who are not so accustomed to dealing with the written language (e.g., in the case of questionnaires);
- the ambiguity of everyday language;
- difficulties in translating scientific terms into everyday language; and
- the problem of putting into words many aspects of psychological regulation processes (Hacker, 1998).

Observation Methods

Observation methods are generally used in cases where it is important to avoid the sort of errors that can occur in interview methods or 'bias' as a result of evaluation and interpretation processes on the part of the workers, or when, in future workplace design, no workers are yet available for the planned jobs.

Observation methods are often seen as a way of getting around the problems inherent in interview methods and obtaining 'more objective' data. In reality, they are subject to the same sort of problems as interview methods, in some cases giving rise to additional problems:

- The quality of job observations deteriorates for complex work activities.

- Certain temporally dynamic aspects of the work activity (e.g., pressure of time) are harder to observe.
- Infrequent events, which are nevertheless of significance for the job, (e.g., starting and stopping machines, annual accounts) are often not included.
- Observers, too, are subject to evaluation, interpretation, and 'biasing' processes. For instance, observers tend to rate workplaces as uniformly good or bad.

These typical advantages and disadvantages mean that proper job analysis involves considering precisely the methods which are suitable. This also means that users must be aware of the problems inherent in these methods, carrying out, where necessary, appropriate training measures to reduce them.

It is also a good idea – whenever this is feasible – to combine different methods, e.g., questionnaires, interview, and observation methods. For this reason, many techniques also include the *observational interview* as a proven data collection method, based on structured observation of the work processes and related interviews with the workers involved at their workplace.

This is also reasonable and required from a methodological point of view as a number of surveys show that data provided by incumbents have a significantly lesser reliability than data collected by analysts and experts (Dierdorff and Wilson, 2003; Voskuijl and Sliedregt, 2002; see also Dunckel and Resch, 2010).

Methodological Standards

Job analysis methods differ, among other things, in their degree of standardization – ranging from nonstandardized 'free' descriptions to semistandardized interviews to observations and interviews following exactly prescribed rules. They differ in the amount of time they take (from 30 min to several hours), their psychometric quality, the number of dimensions captured, etc. Which method to choose cannot be decided in a general manner; this depends largely on the intended application and on available knowledge and theory building with respect to the problems under investigation.

Scientifically based methods should be reliable and valid:

- The *reliability* of a method shows whether and to what extent a method can be used to obtain 'stable,' 'reliable,' or 'replicable' results. Ideally, repeated measurements of the same object should show as little deviation as possible (Oesterreich and Bortz, 1994).
- The *validity* of a method indicates the extent to which it actually measures what it is supposed to measure.

Examining these quality criteria is not only of scientific interest, but also of utmost practical importance. If the results of the analysis are to have practical consequences, e.g., for job design, decision makers must be able to depend on the fact that the results are reliable and valid.

Results

Job analysis methods provide quantitative and qualitative results. Qualitative results are mostly verbal, narrative descriptions of the job or the tasks; quantitative results are presented in numerical – and in some cases graphical – form. Complex methods generally present results in both qualitative and quantitative forms.

Outlook

Job analysis will continue to constitute a major basis for organization and personnel development also in the future. Given the numerous potential applications, there will be no one universal method covering all application purposes, even though the latest techniques allow more effective forms of information analysis and evaluation.

Job analysis methods were developed principally for the industrial and administrative sectors. They exhibit shortcomings, for example, when analyzing activities in the service sector, domestic and family work, and the self-employed sector. Furthermore, job analysis, too, will be obliged to respond to changes in the working world. New forms of work, e.g., jobs in virtual and networked organizations, teamwork within virtual structures, telecommuting and electronic business, dynamic and complex jobs with vague boundaries, also present new challenges for job analysis techniques and methods (Brannick et al., 2007; *see* Teamwork and Team Performance Measurement).

In addition, in the area of human resources management, job analysis is still needed (Ash, 1988) to help reduce gender discrimination with respect to pay by showing that the still existing differences in income between men and women on the same jobs cannot normally be justified by differing job requirements and conditions.

See also: Job Satisfaction; Organizational Citizenship Behavior; Organizational Climate in the Work Setting; Performance Appraisal and Evaluation; Teamwork and Team Performance Measurement; Work Motivation.

Bibliography

Algera, J.A., Greuter, A.M., 1998. Job analysis. In: Drenth, P.J.D., Thierry, H., Wolff de, C.J. (Eds.), Handbook of Work and Organizational Psychology, second ed., vol. 3. Psychology Press, Hove.

Ash, R.A., 1988. Job analysis in the world of work. In: Gael, S. (Ed.), The job analysis handbook for business, industry and government, vols. I and II. Wiley, New York.

Brannick, M.T., Levine, E.L., Morgeson, F.P., 2007. Job and Work Analysis, second rev. ed. Sage, Los Angeles.

Dierdorff, E.C., Morgeson, F.P., 2007. Consensus in work role requirements: the influence of discrete occupational context on role expectations. Journal of Applied Psychology 92 (5), 1228–1241.

Dierdorff, E.C., Wilson, M.A., 2003. A meta-analysis of job analysis reliability. Journal of Applied Psychology 88 (4), 635–646.

Dunckel, H. (Ed.), 1999. Handbuch psychologischer Arbeitsanalyseverfahren (Handbook of Psychological Job Analysis Techniques). vdf, Zürich.

Dunckel, H., Resch, M.G., 2010. Arbeitsanalyse (Work analysis). In: Kleinbeck, U., Schmidt, K.-H. (Eds.), Enzyklopädie der Psychologie, Band Arbeitspsychologie. Hogrefe, Göttingen.

Flanagan, J.C., 1954. The critical incident technique. Psychological Bulletin 51, 327–358.

Frese, M., Sabini, J. (Eds.), 1985. Goal Directed Behavior: The Concept of Action in Psychology. Lawrence Erlbaum, Hillsdale, NJ.

Gael, S. (Ed.), 1988. The Job Analysis Handbook for Business, Industry, and Government, vols. I and II. Wiley, New York.

Ghorpade, J., 1988. Job Analysis. A Handbook for the Human Resource Director. Prentice Hall, Englewood Cliffs, NJ.

Hacker, W., 1998. Allgemeine Arbeitspsychologie (General Industrial Psychology). Huber, Bern.

Hackman, J.R., 1970. Tasks and task performance in research on stress. In: McGrath, J.E. (Ed.), Social and Psychological Factors in Stress. Holt, Rinehart & Winston, New York, pp. 202–237.

Hackman, J.R., Oldham, G.R., 1975. Development of the job diagnostic survey. Journal of Applied Psychology 60, 159–170.

Katz, D., Kahn, R.L., 1978. The Social Psychology of Organizations, second rev. ed. Wiley, New York.

Landy, F., 1993. Job analysis and job evaluation: the respondent's perspective. In: Schuler, H., Farr, J.L., Smith, M. (Eds.), Personnel Selection and Assessment. Lawrence Erlbaum, Hillsdale, NJ, pp. 75–90.

Lees, C.D., Cordery, L., 2000. Job analysis and design. In: Chmiel, N. (Ed.), Introduction to Work and Organizational Psychology. A European Perspective. Blackwell, Oxford, pp. 45–68.

McCormick, E.J., 1976. Job and task analysis. In: Dunnette, M.D. (Ed.), Handbook of Industrial and Organizational Psychology. Rand McNally, Chicago, pp. 651–697.

Oesterreich, R., Bortz, J., 1994. Zur Ermittlung der testtheoretischen Güte von Arbeitsanalyseverfahren (On determining the test-theoretical quality of job analysis techniques). ABO Aktuell 3, 2–8.

Oesterreich, R., Volpert, W., 1987. Handlungstheoretisch orientierte Arbeitsanalyse (Action-theory-oriented job analysis). In: Kleinbeck, U., Rutenfranz, J. (Eds.), Arbeitspsychologie (Enzyklopädie der Psychologie, Themenbereich D, Serie 111, Band 1). Hogrefe, Göttingen.

Turner, A.N., Lawrence, P.R., 1965. Industrial Jobs and the Worker. Harvard University, Boston.

Voskuijl, O.F., von Sliedregt, T., 2002. Determinants of interrater reliability of job analysis. A meta-analysis. European Journal of Psychological Assessment 18 (1), 52–62.

Job Design and Evaluation: Organizational Aspects

Alan J Arthurs, University of Bath, Bath, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 12, pp. 7977–7980, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editors.

Abstract

Job evaluation is a method that is used to produce a hierarchy of jobs in an organization as the basis for determining relative pay levels. It seeks to measure the relative value of jobs, not that of the job holders. The main aim of job evaluation is to provide an acceptable rationale for determining the pay of existing hierarchies of jobs and for slotting in new ones. It may be implemented unilaterally by management or with varying degrees of participation by the workforce.

Job evaluation is a method that is used to produce a hierarchy of jobs in an organization as the basis for determining relative pay levels. It seeks to measure the relative value of jobs, not that of the job holders. The main aim of job evaluation is to provide an acceptable rationale for determining the pay of existing hierarchies of jobs and for slotting in new ones. It may be implemented unilaterally by management or with varying degrees of participation by the workforce. Acceptability, consensus, and the maintenance of traditional hierarchical structures are normally the goal of such schemes. Job evaluation schemes do not directly determine rates of pay. They are concerned with relationships, not absolutes. The rate for the job or the salary for a job grade is influenced by a number of factors outside the scope of most schemes. Normally pay is linked to external market rates, the relative bargaining strengths of the negotiating bodies, and traditional patterns of pay differentials between jobs. In recent years equal pay legislation has exposed job evaluation to sometimes rigorous analysis in the courtroom. Out of this are emerging principles which seek to ensure that job evaluation schemes are constructed and implemented transparently and without obvious bias (Acker, 2006).

Job evaluation is a method used to produce a hierarchy of jobs in an organization as the basis for determining relative pay levels. It seeks to measure the relative value of jobs, not that of the jobholders. Ideally, the performance of the individual should not enter into job evaluation although, in practice, it may be difficult to dissociate individuals from their jobs. Job evaluation is based on the assumption that components of work can be isolated and that these can be decomposed into elements considered worthy of remuneration.

The main aim of job evaluation is to provide an acceptable rationale for determining the pay of existing hierarchies of jobs and for slotting in new ones. It may be implemented unilaterally or with varying degrees of participation by the workforce. Acceptability, consensus, and the maintenance of traditional hierarchical structures are normally the goals of such schemes.

Job evaluation schemes do not directly determine rates of pay. The rate for the job or the salary market for a job grade is influenced by a number of factors outside the scope of most schemes. Often, the pay determinants and indeed hierarchies are linked to external market pay rates, the relative bargaining strengths of the negotiating bodies, and traditional patterns of pay differentials between jobs. Job evaluation is concerned with relationships, not absolutes.

The Rise of Job Evaluation

Job evaluation grew up in the first half of the twentieth century, in an era of large, bureaucratic organizations and increasing unionization. There was a shift toward more 'rational' and 'scientific' methods of determining pay differentials (Ruona and Gibson, 2004). A major concern was to justify pay levels since many industrial disputes were concerned with wages. Collective bargaining, unilateral wage determination, or individual negotiations were seen as unsystematic compared to job evaluation. The rise of mass production and sometimes-militant industrial unionism was also important. As a result of new production methods introduced at the end of the 1930s and during World War Two, the number of varied yet unskilled occupations increased sharply. The clear line between mass production and traditional craft jobs had become blurred. Pay on the basis of job content, not personal attributes, complemented this system of industrial relations.

Main Emphases in Job Evaluation Research

For many years the main interest in job evaluation has come from scholars in the fields of industrial relations, labor economics, and industrial psychology.

During World War II some academic economists had been members of, or attached to, the US National War Labor Board, a body which encouraged the use of job evaluation. Subsequently, some of these scholars, notably Clark Kerr, wrote about the industrial relations aspects of the subject (Kerr and Fisher, 1977). One of his conclusions was that the design of a scheme is of less practical importance than the context into which it is introduced and the way in which it is administered. Harold Northrup, another scholar with links to National War Labor Board, concluded that job evaluation was used by management 'partly to deter or prevent unionization, partly to rationalize its wage scales prior to unionization … and partly to stabilize the wage structure and eliminate continuous bargaining over particular rates after unionization.' (quoted in Hutner, 1986: p. 25).

In the 1940s a very different strain of research was started by industrial psychologists interested in the design of rational, equitable pay systems and psychometric issues of reliability and validity. This early research generally concluded that trained individuals, familiar with the organization's jobs and given detailed instructions on how to analyze them, could reach similar conclusions on overall job worth (see, for example, Ash, 1947). Interest in sex discrimination in pay systems stimulated further research on the validity of different job evaluation methods in the 1980s. This produced results which were much less clear (Arvey, 1986). Madigan (1985) concluded that, while the reliability of various methods tested was in the range from .70 to .90, the accuracy with which the predicted pay grade matched the actual grade was much lower.

The validity of job evaluation has been investigated by looking at the results of different methods, using predictive validity procedures, factor analysis techniques, and testing for potential sources of bias. Results have been contradictory, since some researchers have concluded that the results will be similar whatever the method used (for example, Robinson et al., 1974), whereas others have found significant differences (see Madigan and Hoover, 1986). There is some consensus that the skills and knowledge dimension, present in nearly all methods, is an overriding predictor of job evaluation results (Rogers, 1946; Madigan, 1985). It has been argued that this means that job evaluation mainly measures abilities that a person brings to the job, rather than the job content (Rogers, 1946). Another research finding is that where the evaluators know the job title this significantly influences the results (Smith et al., 1989). A review of the psychological literature has proposed, unsurprisingly, that 'research models need to be expanded beyond focusing on the psychometric properties of job evaluation' (Gerhart and Milkovich, 1992).

Recent research has drawn upon organization theory to study how individual and group power may influence the outcomes of job evaluation. The coalition model of organizations suggests that negotiation, bargaining power, and politics are key determinants of decision behavior, particularly resource allocation decisions (Pfeffer and Salancik, 1974). Using this approach Welbourne and Trevor (2000) examined whether power in a university affected job evaluation outcomes. They concluded that both departmental and individual positional power helped predict the results. Departmental power predicted the number of job grade allocations and reallocations achieved and positional power predicted the probability of favorable reallocation decisions. Powerful departments were more likely to approach the Human Resources Department with marginally acceptable requests and indulge in more extensive negotiations about the appropriateness of job descriptions and the proposed grade. The school's dean was likely to intervene prior to an official submission, particularly for higher graded positions. Welbourne and Trevor conclude that a rational model of behavior to explain outcomes is inadequate and that job evaluation captures the dynamics of the job's environment as well as its worth. Writing over 60 years earlier Viteles was less measured, claiming that 'beneath the superficial orderliness of job evaluation techniques and findings, there is much that smacks of chaos' (Viteles, 1941: p. 165).

Equal Pay and Job Evaluation

In the 1960s equal pay legislation was introduced in many western economies. The concept of equal pay for equal work has been widened in many jurisdictions to include equal pay for work of equal value. In the European Union, where a dispute arises as to whether work is of equal value, a country is required to provide a process by which an assessment of value can be made. Such a process involves some form of comparison based on job evaluation or classification. Where job evaluation is used it must be shown not to discriminate against women. In some countries, particularly the US, equal pay for work of equal value, also known as comparable worth, has proved to be a highly contentious issue. Economists, labor lawyers, historians, and sociologists have joined the debates. The role of job evaluation in bringing about equal pay for work of equal value has been scrutinized and vigorously disputed.

Among economists there are many that dismiss the use of job evaluation as a remedy for pay discrimination against women. They believe that to attempt to compare the value of different jobs using job evaluation is like 'comparing apples with oranges.'

The power dimension in job evaluation has been made more transparent by historical analyses of the way in which job evaluation has been used to maintain traditional male differentials, reflecting the prevailing values or relative power in society. Job evaluation, despite its rhetoric of fairness, does not necessarily lead to pay equity for women. During World War Two a US National War Labor Board case by the United Electrical Workers against General Electric (GE) and Westinghouse sought equal pay for work of equal value (Milkman, 1987). The job evaluation system used for all GE jobs gave different pay rates for men and women. The company responded that these gender-based rate differentials were longstanding, accepted by the union during bargaining rounds, and in line with industry practices. They argued that to pay women more would put the firms at a competitive disadvantage and that the board should consider 'sociological factors,' including women's different social and economic roles. Although the board decided in the union's favor, the companies ignored the decision (Milkman, 1987: pp. 80–1). Although explicit references to male and female rates disappeared gendered concepts, such as the distinction between 'light' and 'heavy' work, became the basis for unequal pay.

Until recently most job evaluation schemes continued to incorporate the gender bias of the early twentieth century (Treiman and Hartmann, 1981). Common practices were: the separation of men and women into different job evaluation schemes; greater weight being given to the attributes of male-dominated occupations; and using different scales for men and women when converting job evaluated points to pay rates (Baker and True, 1947).

The Law and Job Evaluation

Equal pay legislation has exposed job evaluation to sometimes rigorous analysis in the courtroom. Out of this are emerging principles that seek to ensure that job evaluation schemes are constructed and implemented transparently and without obvious bias. A distinction accepted in the European Court is

between analytical and non-analytical methods. In analytical schemes jobs are broken down into their component elements for the purposes of comparison. In non-analytical schemes the relative worth of the jobs may be based on a whole job comparison. More formal methods, particularly analytical schemes, have been judged to be more transparent and less susceptible to bias. However, it has been accepted that no scheme can ever be fully objective since the whole process is based on a series of judgments made about the facts presented to the evaluators. These judgments reflect each evaluator's background, experience, and attitudes.

Future Directions

By the middle of the 1980s job evaluation was being used by about 75% of US firms (Beer et al., 1985). However, for the past two decades increased individualization of the employment relationship has been developing and unions in most western economies have been on the retreat. Employers have been exhorted to be more flexible and willing to consider fundamental changes in employment practices. Increasingly job evaluation has been portrayed as bureaucratic, inflexible, reinforcing hierarchy, and not in tune with the needs of the times. Lawler has argued that in a rapidly changing, highly competitive environment, particularly where companies depend on advanced technology and the knowledge of their employees, pay systems should emphasize growth, development, and performance (Lawler, 1986, 1994). He claims that this is inhibited by job evaluation, which depersonalizes individuals by equating them with a set of duties, rather than concentrating on who they are or what they can do.

These views have been influential and there has developed a greater emphasis on paying people for their skills, knowledge, and performance. However, job evaluation, after becoming less fashionable in the 1980s and early 1990s, is in many countries more popular than ever. One reason for this is that the case against job evaluation has been overstated. Many of the problems associated with it can be overcome by redesigning schemes to fit the needs and values of modern organizations. A second reason is that there exists a flourishing job evaluation consultancy industry which has been quick to respond to some of the criticisms by modifying existing schemes and introducing new ones which incorporate more emphasis on skills, knowledge, competency, and performance. Broadbanding has introduced fewer, wider pay bands, thereby reducing hierarchical levels and the need for constant reevaluations.

A further influence has been that in many countries organizations have introduced job evaluation in order to be seen to be responding to legislative demands for pay equity between men and women. Job evaluation, carried out analytically, transparently, and with the participation of those concerned, normally provides a defence against potentially very expensive legal action.

Job evaluation will continue to change, but is unlikely to disappear. Employees expect equitable treatment. Even in an age of flexibility and individualism, it remains the most reliable means of achieving a pay structure that is perceived as both rational and fair by employees and their managers (Sholihin and Pike, 2009).

See also: Corporate Culture; Ethical Codes, Professional: Business Codes; Human Resource Management, Psychology of; Organizational Behavior, Psychology of; Organizational Climate; Organizations: Authority and Power; People in Organizations; Public Administration: Organizational Aspects; Stress at Work; Workplace Autonomy; Workplace Safety and Health.

Bibliography

Acker, J., 2006. Inequality regimes: Gender, class, and race in organizations. Gender & Society 20 (4), 441–464.
Arvey, R.D., 1986. Sex bias in job evaluation procedures. Personnel Psychology 39, 315–335.
Ash, P., 1947. The reliability of job evaluation procedures. Journal of Applied Psychology 32, 313–320.
Baker, H., True, J.M., 1947. The Operation of Job Evaluation Plans: A Survey of Experience. Princeton University, Industrial Relations Section, Princeton, NJ.
Beer, M., Spector, B., Lawrence, P.R., Mills, D.Q., 1985. Human Resource Management. Collier Macmillan, London.
Gerhart, B., Milkovich, G., 1992. Employee compensation: research and practice. In: Hough, L.M., Dunnette, M.G. (Eds.), Handbook of Industrial and Organizational Psychology, vol. 1. Consulting Psychologists Press, Palo Alto, CA, pp. 481–569.
Heneman, R.L., 2003. Job and work evaluation: A literature review. Public Personnel Management 32 (1), 47–71.
Hutner, F.C., 1986. Equal Pay for Comparable Worth. Praeger, New York.
Kerr, C., Fisher, L.H., 1977. Effects of environment and administration on job evaluation. In: Kerr, C. (Ed.), Labor Markets and Wage Determination. California University Press, Berkeley, CA, pp. 53–86.
Lawler, E.E., 1986. What's wrong with point-factor job evaluation. Compensation and Benefits Review 18, 20–28.
Lawler, E.E., 1994. From job-based to competence-based organizations. Journal of Organizational Behavior 15, 3–15.
Madigan, R.M., 1985. Comparable worth judgments: a measurement properties analysis. Journal of Applied Psychology 70 (1), 137–147.
Madigan, R.M., Hoover, D.J., 1986. Effects of alternative job evaluation methods on decisions involving pay equity. Academy of Management Journal 29, 84–100.
Milkman, R., 1987. Gender at Work: The Dynamics of Job Segregation by Sex during World War II. University of Illinois Press, Urbana, IL.
Pfeffer, J., Salancik, G.R., 1974. Organization decision making as a political process: The case of a university budget. Administrative Science Quarterly 19, 135–151.
Robinson, D.D., Wahlstrom, O.W., Mecham, R., 1974. Comparison of job evaluation methods. Journal of Applied Psychology 59, 633–637.
Rogers, R.C., 1946. Analysis of two point rating job evaluation plans. Journal of Applied Psychology 30, 579–585.
Ruona, W.E.A., Gibson, S.K., 2004. The making of twenty-first-century hr: An analysis of the convergence of HRM, HRD, and OD. Human Resource Management 43 (1), 49–66.
Sholihin, M., Pike, R., 2009. Fairness in performance evaluation and its behavioural consequences. Accounting & Business Research (Wolters Kluwer UK) 39 (4), 397–413.
Smith, B.N., Hornsby, J.S., Benson, P.G., Wesolowski, M., 1989. What is in a name: The impact of job titles on job evaluation results. Journal of Business and Psychology 3, 341–351.
Treiman, D.J., Hartmann, H.I. (Eds.), 1981. Women, Work, and Wages: Equal Pay for Jobs of Equal Value. National Academy Press, Washington, DC.
Viteles, M.S., 1941. A psychologist looks at job evaluation. Personnel 17, 165–176.
Welbourne, T.M., Trevor, C.O., 2000. The roles of department and position power in job evaluation. Academy of Management Journal 43, 761–771.

Job Displacement Insurance

Donald O Parsons, George Washington University and IZA Bonn, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Earnings losses from permanent job separations are a serious threat to the financial security of long-tenured workers. Job displacement insurance is presumably designed to offset these losses, but evidence suggests that consumption smoothing among the long-tenured displaced workers is seriously incomplete, at least in lightly regulated labor markets. Unemployment and reemployment wage insurance could fully cover these losses, but are costly to provide. Severance pay has emerged as a supplemental, if much criticized, instrument. Moral hazard limitations on unemployment insurance (UI) generosity mean that severance pay functions as scheduled (partial) UI and scheduled wage insurance. Consumption smoothing over time through savings and borrowing is less efficient than ideal insurance, but may be preferred in second-best situations. Long-tenured separated workers are older on average, which introduces special problems, but also additional policy options, including early access to retirement accounts.

Introduction

Earnings form the bulk of worker wealth in industrial economies and job displacement is a serious financial threat. Job displacement losses vary with the individual, the structure and flexibility of the labor market, and the business cycle, but are typically large and quite persistent among long-tenured workers. Although the economic risk of 'job displacement from a long-tenured job' is a real one to all workers, the fraction that experiences the event is not large. (For his sample of displaced workers in their 50s in the US, Couch reports a 2-year job displacement rate of 4.7% (Couch, 1998: p. 10).) Many workers will never become long-tenured, and long-tenured workers may remain employed until disability or retirement ends their work lives. Structurally, the long-tenured displaced worker is likely to have held a job with relatively stable demand, punctuated by infrequent but sharp permanent contractions.

Events that involve a small probability of a large loss are ideally insurance events, not saving or borrowing events. Earnings losses are usefully decomposed for insurance purposes into (1) those incurred through spells of unemployment and (2) those incurred through reemployment wage (rate) losses. Reemployment wage losses are apparently costly to insure, perhaps for administrative reasons, while unemployment losses are difficult to insure fully because of incentive problems deriving from the greater leisure available to the unemployed. If the utility value of benefits and the unemployed leisure is higher than that of working, incentives to seek and accept job offers fail (search moral hazard). (The resulting distortions can be dramatic if unemployment insurance (UI) rules are poorly written (Riddell and Kuhn, 2010).) Other distortions, often labeled 'firing cost' effects, arise if the employer bears substantial expected separation costs.

Standard commercial insurance is rare for this risk, but employers, trade unions, and governments offer various consumption smoothing mechanisms. Each has its own limitations, and supply elements play an important role in the composite insurance that emerges to moderate these risks. (A detailed discussion of the job displacement insurance policy alternatives can be found in Parsons (2014a).) Employers can offer simple separation packages to displaced workers and have one distinct advantage – they can avoid job displacement altogether by maintaining the worker on the job in low-demand times (labor hoarding). (Excessive labor hoarding is the focus of its own literature, carrying the label firing costs.) Trade unions can negotiate guaranteed employment agreements or separation benefits and even provide benefits directly. However, employers and labor unions have one critical disadvantage, an inability to monitor worker activities following displacement, which is required for actual-loss-based unemployment and reemployment wage insurance. Governments can reasonably monitor reemployment, at least if labor markets are largely formal, but struggle with subtler information like the intensity of job search. As a consequence, most separation packages involve only partial UI. (Card et al. (2007) and Chetty (2008) raise provocative questions about the welfare implications of UI-induced increases in unemployment duration. UI effects may include income (liquidity) effects as well as the usual incentive effect that concern most policy designers.) Severance pay, fixed-sum payments at the time of separation, serves as both scheduled wage insurance and (partial) scheduled UI in this second-best world. (Scheduled benefits mean fixed benefits linked to expected, rather than actual, losses.)

The important function of severance pay as scheduled wage insurance is often neglected in theoretical studies. These studies fail to distinguish between temporary separations, which normally involve no wage rate losses, and permanent separations, which do. (A detailed review of job displacement loss studies is provided below.) For example, Azariadis (1975), Rosen (1985), and Blanchard and Tirole (2008) focus on contracting issues and simplify by ignoring the reemployment market, essentially treating unemployment as an absorbing state. Shavell and Weiss (1979) and Shimer and Werning (2008) consider the behavior of unemployed workers only, so the models are silent on prior wages and therefore wage loss. Flemming (1978), Hansen and Imrohoroglu (1992), Hopenhayn and Nicolini (1997), and Pissarides (2010) introduce a reemployment market in a contracting framework

but simplify by assuming that wages are the same for workers pre- and postlayoff, thus permitting them to focus on unemployment.

Governments are quite active in this insurance market, often mandating that firms provide severance pay and, less frequently, directly supplying unemployment insurance; a summary of these efforts is provided below. The success of these measures in smoothing consumption following job displacement of long-tenured workers is more problematic. Cochrane (1991), using a general consumption approach proposed by Mace (1991), screened the U.S. for specific adverse events and identified involuntary job loss and, to a lesser extent, long-term illnesses (greater than 100 days) as being especially poorly 'insured.' A number of consumption studies, surveyed below, confirm Cochrane's finding. At least in US-style economies (US, UK, and Canada), long-tenured displaced workers experience large average declines in consumption.

One demographic group is of special interest. Long-tenured displaced workers are disproportionately older and may be eligible for disability and retirement benefits, with consequences for labor force participation and consumption smoothing. The decline in health status and marketable skills that often comes with age, when combined with the private and public income support systems meant to ease the impact of these declines on consumption, makes these workers an especially attractive policy target. Like private firms, governments often reduce the aggregate workforce by encouraging 'early retirement,' especially when unemployment is high.

Job Displacement Losses

There are a variety of reasons for job separation, with quite different consequences for earnings. Layoffs (employer-initiated job separations without cause) are the main focus here and must be distinguished from quits (employee-initiated job separation), discharges (employer-initiated job separations with cause), and other separations, including most prominently retirement. (These are U.S. Bureau of Labor Statistics (BLS) definitions. Elsewhere, BLS layoffs are often called redundancies.) Quits (voluntary separations from the perspective of the worker) are often advantageous to the worker, while discharges for cause are not, but for obvious reasons employers are averse to subsidizing such separations. Retirement carries the expectation that the worker will leave the labor force and often involves employer, union, or government financial support independent of job displacement insurance. Within separation categories, distinctions are also important, including that between permanent and temporary layoffs. Permanent layoffs are synonymous with job displacements. Temporary layoffs, common in the United States and Canada, are layoffs expected to result in rehire, often after a known time period.

Job displacement losses are often substantial. Consider the recent experience of job displacements of long-tenured workers (3 or more years of tenure at displacement) in the United States, derived from the BLS's Displaced Workers Survey. Of the long-tenured displaced workers from 2009 to 2011, only 56% were employed in January 2012, with another 26.7% unemployed, and the remainder (17.4%) out of the labor force altogether (BLS, 2012). The fraction out of the labor force jumps dramatically among workers 55 years of age and older (**Figure 1(a)**: Panel A). Among displaced workers from full-time wage and salary jobs who were employed in January 2012, 26.6% suffered wage losses of 20% or more (**Figure 1(b)**: Panel B).

Economists have devoted considerable attention to the structure of these losses; a few key studies are summarized in **Table 1**. The paradigm study is that of Jacobson et al. (1993a), who used administrative data to capture important elements of the dynamics of job displacement loss. (Jacobson et al. (1993b), Fallick (1996), Kletzer (1998), and Kuhn (2002) provide surveys of the literature.) Long-tenured workers (6 years of tenure or more) suffered earnings losses that begin well before displacement, with a sharp earnings drop at displacement, followed by a lengthy, incomplete recovery. As they reported, "We find that high-tenured workers separating from distressed firms suffer long-term losses [6 years following displacement] averaging 25% per year." (Jacobson et al., 1993a: p. 685). (Losses are relative to a control group of retained workers.) Short-term earnings losses were much higher, 45% in the first quarter after displacement.

The Jacobson, LaLonde, and Sullivan data represent a single state with a large heavy-manufacturing sector (Pennsylvania) and may be atypical. Couch and Placzek (2010) replicated the Jacobson, LaLonde, and Sullivan study on administrative data for Connecticut in a more robust, service-oriented labor market and find that earnings losses are significantly lower. Couch and Placzek (2010: p. 587) reported earnings losses of 32–33% in the quarter after displacement and 13–15% 6 years later. The lower loss estimates are to be expected given the countercyclical property of job displacement earnings losses (Davis and Von Wachter, 2011). The basic time profile of earnings losses was quite stable.

Studies based on survey data provide additional details of this process, including the partition of earnings losses into unemployment losses and wage losses. Ruhm (1991), using the Panel Study of Income Dynamics (PSID), reported persistence earnings losses following job displacement, on the order of 14.7% 4 years after job loss. With the survey data, Ruhm was able to break down the profile of losses by source. The high unemployment experienced following job loss had largely (85%) dissipated 4 years later, but earnings losses linked to wage rate losses remained large, with reemployment wages 14–15% below preseparation earnings.

Stevens (1997), also using the PSID, provided details on the nature of this loss process – not only are the displaced workers often forced to accept entry-level wages but also the new job will almost, by definition, have no seniority rights, making workers vulnerable to additional spells of unemployment. She reported relative earnings losses among displaced workers with 3 or more years of tenure of 26% immediately following separation. Earnings losses of 11% remained 6 or more years after displacement, almost all wage (rate) losses (Stevens, 1997: p. 178).

The U.S. Job Displacement Surveys provide ample historical evidence of earnings losses (up to 3 years following displacement), both on average and by job tenure (Farber, 2004). Average losses among the displaced workers included

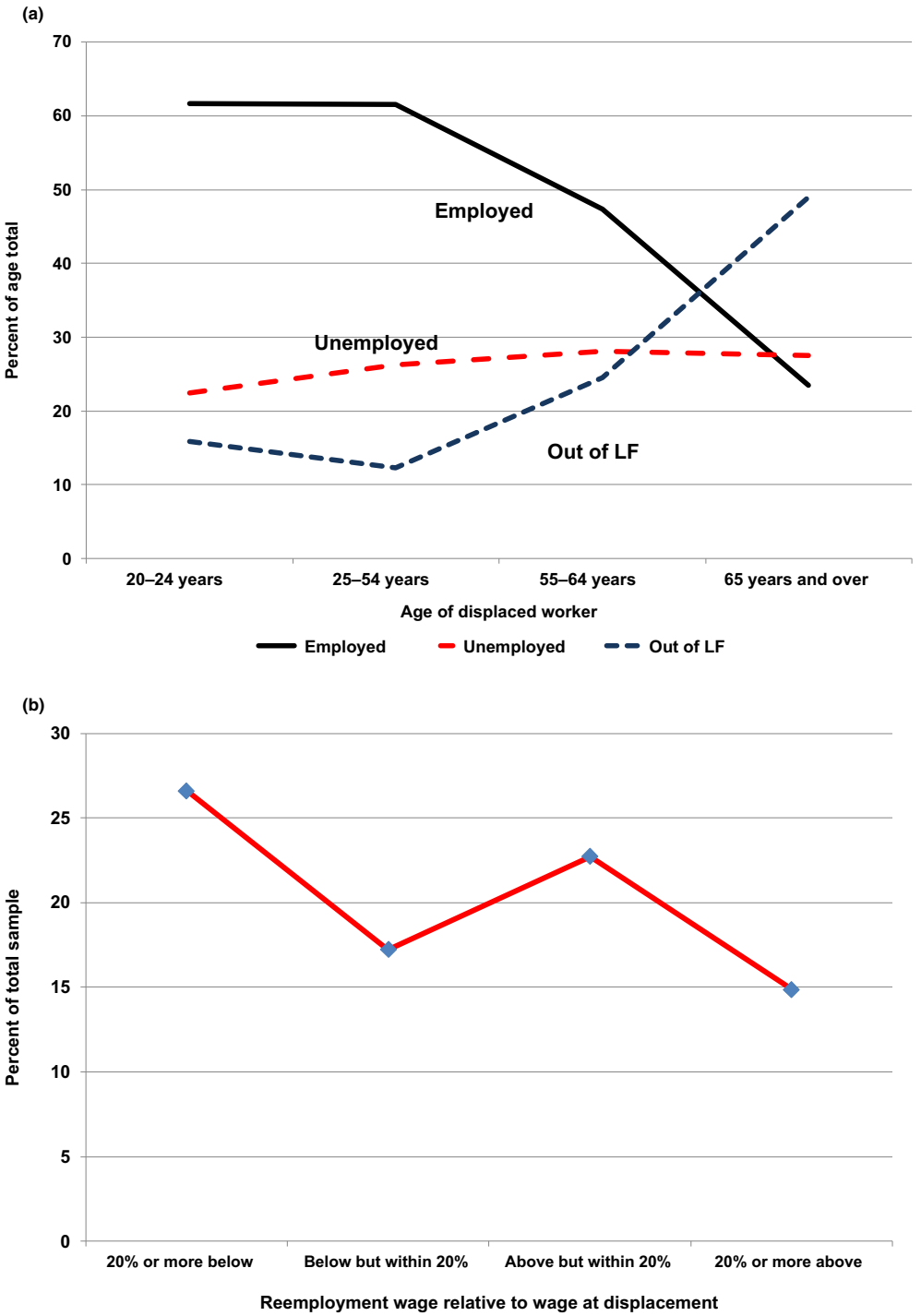

Figure 1 (a) Labor force status by age January 2012 displaced workers with 3 or more years tenure 2009–11. (b) Distribution of reemployment wage loss in January 2012 of long-tenured (full-time) workers displaced in 2009–11. LF, labor force. Bureau of Labor Statistics (U.S. Department of Labor), 2012. Worker Displacement: 2009–2011. BLS news release USDL-12-1719.

employment losses of 32%, while wage losses among the full-time to full-time reemployed average 8% (Farber, 2004: pp. 113, 115). Earnings losses varied widely across individuals, industries, and economic conditions. Indeed, many individuals report wage gains, especially in good times (Helwig, 2001).

Expected losses were systematically larger for long-tenured workers. Averaged over the first ten Displaced Worker Surveys (1982–84 through 1999–2001), losses among full-time workers who secure full-time reemployment were as follows: 3.3% among workers with 1–3 years of tenure (relative to those

Table 1 Job displacement losses among long-tenured workers

Study	*Data source*	*Sample restrictions*	*Short-term earnings loss*	*Long-term earnings loss*
Jacobson et al. (1993a)	Administrative data, Pennsylvania	Tenure 6+ years mass layoffs	45% (Q1)	25% ($t+6$)
Couch and Placzek (2010)	Administrative data, Connecticut	Tenure 6+ years mass layoffs	32–33% (Q1)	13–15% ($t+6$)
Ruhm (1991)	PSID	Plant closings, permanent layoff	17.5% ($t+1$)	14.7% ($t+4$)
Stevens (1997)	PSID	Tenure 3+ years plant closings, permanent layoff	26% ($t+1$)	11% (6+ years)
Farber (2004)	DWS	Slack work, plant closing, position/shift abolished	32.4% employment loss 7.8% wage loss, FT to FT reemployed (3 years or less)	NA

DWS, Displaced Worker Survey; FT, full-time.

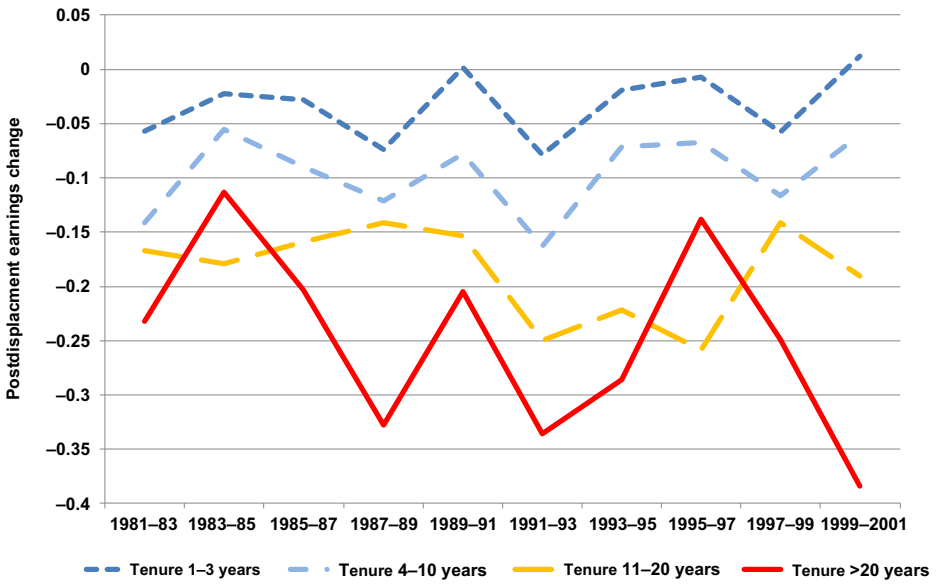

Figure 2 Earnings change of displaced workers by tenure full-time employment to full-time reemployment first ten Displaced Worker Surveys. Farber, Henry S., 2004. Job loss in the United States, 1981–2001 accounting for worker well-being. Research in Labor Economics 23, 69–117.

with less than 1 year), 9.6% among those with 4–10 years, 18.6% for those with 11–20 years, and 24.7% among those with more than 20 years of tenure. (Author's calculations from Farber (2004: Table 3, pp. 96–97.) These are partial effects, controlling for gender, race, age, and years since displacement, for workers 20 years and older.) Loss *differentials* by tenure appear quite stable over the business cycle (Figure 2).

Job Displacement Insurance across Economies

Job separation packages vary widely across economies (2009), but not without logic (Table 2). (Holzmann et al. (2012) provide a detailed empirical account of variations in the size and composition of separation packages worldwide.) The administrative convenience of severance pay – the government need only mandate that private employers provide the

Table 2 Job displacement insurance coverage worldwide (countries with populations exceeding 1 million in 2009)

	Countries
Only statutory severance	69 (46.3%)
Both statutory severance pay and UI/UA	61 (40.9%)
Only UI/UA	13 (8.7%)
With significant collectively bargained severance	12
With no significant collectively bargained severance	1
None	6 (4.0%)
Total	149

UI, unemployed insurance; UA, unemployed assistance.
Source: Author's calculation from amended Holzmann, Robert, Pouget, Yann, Weber, Michael, Vodopivec, Milan, 2012. Severance pay programs around the world: history, rationale, status, and reforms. In: Holzmann, Robert, Vodopivec, Milan (Eds.), Reforming Severance Pay: An International Perspective. World Bank, Washington, DC, Annex B.

instrument – makes it especially popular. (Severance pay has its own potential limitations, the firing cost problem, discussed in Insurance Design below. Parsons (2013) provides a recent review of severance pay theory and empirics.) In a sample of 149 countries with 1 million or more inhabitants, 130 (87%) had some form of mandated severance pay, including 27 that had both statutory severance and collectively bargained severance pay in a 'significant manner' (Holzmann et al., 2012: Annex B). (Author's calculation from Holzmann et al. (2012: Annex B).) Another 13 (9%) had only significant collective bargaining severance. (This group includes the US, where collective bargaining agreements do often include severance pay provisions (Pita, 1996), but where collective bargaining is limited. Voluntary plans are however prevalent (Bishow and Parsons, 2004).) Only six (4%) had neither.

UI is more difficult to provide, and, not surprisingly, much less common. Of the 130 countries in the sample with statutory severance pay, only 61 (47%) provided UI, while 69 did not. Another eight (6%) with statutory severance pay offered means-tested unemployment assistance. Of the 13 countries with only collectively bargained severance, 11 countries provided UI and another unemployment assistance. None of the six countries that had neither statutory nor collectively bargained severance (the Democratic Republic of the Congo, Iraq, Haiti, Lebanon, Gambia, and Timor-Leste) provided UI.

The various combinations of UI/assistance and mandated severance pay are not random, but are closely linked to country income and presumably government competencies. GDP per capita in the six countries that had neither mandated severance nor some form of UI was only $1999. Of the 69 countries that had only mandated severance pay and no UI, average GDP per capita was $5810. Of the 61 countries with both mandated severance and some form of unemployment benefit, the average GDP per capita (in 2009 U.S. dollars) was $14 418.

Insurance Adequacy

On average, displaced workers with long tenure suffer serious long-term earnings losses, which, in the absence of consumption smoothing instruments, will have adverse effects on household consumption. There are many such devices, private and public, and it is natural to ask how effective these efforts are at smoothing the consumption of displaced workers. A number of consumption studies of displaced workers provide evidence on that question. Unfortunately, these studies are limited to 'Anglo-Saxon' countries (the UK, Canada, and the US), which have broadly similar labor markets and social insurance systems. In all three countries, labor markets are relatively 'flexible,' UI replacement rates are moderate, and severance pay is modest. (Severance pay is voluntary in the U.S. and, if offered, typically provides 1–2 weeks pay per year of service. Canada mandates 1 week pay per year of service, the U.K approximately one half-week per year of service (Holzmann et al., 2012: Annex B).)

Assessment studies focus on the consumption, not the utility, effects of displacement, no doubt in the belief that wage changes upon reemployment will have only modest effects on work hours. The U.S. study by Cochrane (1991) flagged job displacement as an especially serious threat to consumption stability. "Households that experienced an involuntary job loss had consumption growth 24–27% points lower than households that did not …" (Cochrane, 1991: p. 969). Other studies, directly focused on the negative consumption effects of job displacement, confirmed that they are substantial. Stephens, for example, tracked out the impact of job loss ('plant closed/employer moved' and 'laid off/fired') on earnings and on consumption over time in the U.S. (PSID). He found a long-term decline (6 years after displacement) in the earnings of the household head of 21.9%, with a decline in family food consumption of half that amount, or 11% (Stephens, 2001: p. 33). Among older workers (age 50 to early 60s), the relative change (loss) in annual food consumption within 1 year of job loss was 15% (PSID)–17.5% (Health and Retirement Study (HRS)) (Stephens, 2004: p. 268) (**Table 3**).

Consumption studies using Canadian and U.K. data revealed similar regularities. Browning and Crossley (2008), using Canadian data, assess the impact on consumption of *permanently* displaced workers (plant closure and other mass layoffs). They reported consumption losses of about 6% relative to consumption growth for those laid off with known recall data. Losses were higher for workers with tenure greater than 10 years or age greater than 40 years (10%). Bloemen and Stancanelli (2005) estimated a model of food consumption growth (from prior to layoff to 3 months after) in the United Kingdom, and report a decline in consumption of 17%.

Researchers have also estimated the impact of variations in job displacement insurance, especially UI generosity, on consumption of the unemployed. Presumably, consumption will vary (positively) more closely with benefits the more critical the benefits are as consumption support when earnings drop. Browning and Crossley (2001), using the same Canadian data set discussed above, estimated the sensitivity of consumption to changes in the key unemployment benefit parameter, the replacement ratio. They reported almost no impact of UI generosity on consumption – an elasticity of 0.05 (a 10% decrease in the replacement rate will induce a half-a-percent decrease in consumption). (Gruber (1997) reported extremely large UI replacement rate effects on consumption in the U.S. (consumption elasticity with respect to UI benefits of 0.25 for the whole population), but this estimate is inconsistent with later estimates and should probably be treated as an aberration.) The effect was somewhat higher for vulnerable subgroups, 0.17, for example, for sole breadwinner respondents who report no family assets.

Similarly, Bloemen and Stancanelli (2005) estimated the impact of more generous job displacement benefits on food consumption growth (from prior to layoff to 3 months after) in the United Kingdom. Severance pay ('end-of-job payment') had a positive and significant effect on consumption growth as well as debt reduction following job loss. The UI replacement rate was not significant if entered without interactions, but was significant if interacted with the importance of the respondent's income to the household. "… for a household where the head's earnings are the only source of income, a drop in the replacement rate of 10% leads to a 2% fall in food expenditure" (Bloemen and Stancanelli, 2005: p. 443).

Table 3 Consumption losses among involuntarily displaced workers

Study	*Country/data*	*Sample restrictions*	*Consumption losses*	*Remarks*
A. Consumption losses following displacement				
Cochrane (1991)	U.S./PSID	Involuntary job loss	24–27%	
Stephens (2001)	U.S./PSID	Plant closings and permanent layoffs	11.0% (6+ years) 9.8% (all years)	Earnings decline 21.9% (6+ years)
Stephens (2004)	U.S./PSID and HRS	Older workers (age 50 to early 60s) Business closed/moved or laid off	15.0% (PSID) 17.7% (HRS)	Within 1 year of displacement
Browning and Crossley (2008)	Canada/COEP	Plant closure and mass layoff Tenure greater than 10 years Age greater than 40 years	6.4% 10.4% 9.3%	Canadian Out-of-Employment Panel (COEP) total expenditures
Bloemen and Stancanelli (2005)	U.K./LSUS	Unemployed for at least 3 months	17% (simple average)	Living Standards during Unemployment Survey
B. Responsiveness of consumption to job displacement insurance				
Browning and Crossley (2001)	Canada/COEP	Permanent layoff At least 6 months unemployment	0.05 Consumption/UI RR elasticity 0.17 Consumption/UI RR elasticity if no assets and respondent sole breadwinner	Canadian Out-of-Employment Panel (COEP) total expenditure
Bloemen and Stancanelli (2005)	U.K./LSUS	Unemployed for at least 3 months	UI RR effect insignificant except 0.2 Consumption/UI RR elasticity if zero assets, no severance and sole breadwinner	Severance pay has significant, positive impact on consumption
C. Responsiveness of consumption to permanent/transitory earnings shocks				
Blundell et al. (2008)	U.S./PSID and CEX	Distinguish transitory and permanent shocks	0.64 Elasticity of consumption change to permanent income change 0.05 Elasticity of consumption change to transitory income change	Heterogeneity important

CEX, Consumer Expenditure Survey; HRS, Health and Retirement Survey; PSID, Panel Study of Income Dynamics; RR, replacement rate; UI, unemployment insurance.

Blundell et al. (2008) extended this literature (PSID) in a general way, noting that:

> We find some partial insurance of permanent shocks, especially for the college educated and those near retirement. We find full insurance of transitory shocks except among poor households.
> Blundell et al., 2008, p. 1887

These results are consistent with the belief that the temporarily unemployed workers and short-tenured displaced workers find ways to cushion the effects of earnings shocks on household consumption, but that the long-term displaced do not, with the (partial) exception of the college educated and those near retirement.

Apparently, the mixture of private and public severance pay and UI offered in the countries studied does not guarantee consumption stability in the face of job loss among long-tenured workers. Severance pay and, perhaps, unemployment benefit levels do appear to contribute to consumption smoothing. It is, of course, simple to design a system that would insure uninterrupted consumption following job displacement – offer unemployment benefits of unlimited duration equal to the original wage – because the worker would have no incentive to accept a lower offer. Such a plan, however, would insure that the long-tenured displaced workers do not return to work, except in robust labor markets. The policy challenge is to design systems that will better insure the long-tenured displaced workers without wasting their productive capacities. In the remainder of this article, alternative job displacement insurance systems are considered within a standard demand and supply framework.

The Insurance Market

Insurance Demand

The capitalized value of expected job displacement losses of long-tenured workers will vary with discount rates and potential years of remaining work life, but is large for all but those close to retirement. The literature focuses on employer and government efforts to provide consumption smoothing devices, but individual efforts are nontrivial. At the elementary level, multiearner households pool resources, though the pools are generally quite small. Households also adjust labor market behaviors, with other family members entering the workforce to offset the earnings losses of those displaced.

For frequent, small-loss events – temporary layoffs and displacement of short-tenured workers – savings and credit instruments may be quite satisfactory. Saving instruments are abundant in industrialized economies, but third-party borrowing opportunities are naturally limited by the likelihood that the loan will be paid back. (See Gruber (2001) for a description of financial assets upon entering unemployment and Sullivan (2008) for a discussion of access to unsecured credit (credit cards) among the unemployed.) As instruments provided by employers, unions, and especially government have become dominant, studies of these informal approaches have evolved into the question: Do public efforts crowd out informal efforts or do they contribute on net to worker financial security (Cullen and Gruber, 2000; Engen and Gruber, 2001)? Apparently, crowd-out effects are substantial.

Because layoff in the first instance is an employment event, it is natural to imagine the issue within an employment contracting framework. (Reviews of the employment contracting literature include Parsons (1986), Malcomson (1999), and Salanié (2005).) In a simple single-period model, the worker and the employing firm negotiate a contract that specifies (1) the conditions under which the worker will be retained in the firm with corresponding wages and (2) the separation benefits to be paid if the worker is laid off (Parsons, 2014b). Unemployment is an important aspect of this model, and worker utility is correspondingly multidimensional (consumption and leisure). The worker is assumed to be risk averse and the firm to be risk neutral. (In simple cases, two-dimensional (consumption, leisure) 'risk averse' can be defined by an indirect utility function that is concave in income (Stiglitz, 1969), in this case full income – nonearned income plus earnings if the worker devotes all available hours to work.)

A key issue is the severity of the demand shock that will induce the firm to lay off workers. Layoff behavior, or conversely job retention, has been much studied, in part because of the link with firing cost concerns (see Section Insurance Design below). Azariadis (1975), in his seminal work on employment contracting, derived the conditions under which the firm will lay off workers in an economy in which (1) there is no possibility of reemployment and (2) job displacement insurance is not available. His model highlights the insurance value of labor hoarding: workers will negotiate full-employment contracts over a larger domain of shocks.

If demand contractions are sufficiently large, however, layoffs will occur and the displaced worker will find himself or herself unemployed or reemployed elsewhere, often at a lower wage. A complete separation package will then be a combination of (1) wage insurance and (2) UI.

The first-best magnitude of each of these is dependent on the structure of the state utility functions, sometimes labeled von Neumann–Morgenstern utility functions. Insurance redistributes consumption across states, but the utility value of consumption is in general not independent of the amount of leisure available to the worker, and the displaced worker by definition faces 'excess' leisure in the unemployed state. Many design studies use a special case for this utility function, additively separable concave functions of consumption and of leisure. (Early examples include Azariadis (1975), Shavell and Weiss (1979), and Hopenhayn and Nicolini (1997).) This model has the convenient property that, if insurance can be costlessly provided, *consumption will be equal across states in the ideal contract*. If layoffs are justified, the ideal consumption plan can be implemented by

1. *wage insurance benefits* equal to the difference between the earnings of retained workers and reemployed workers, and
2. *UI benefits* equal to reemployment earnings.

This special utility structure insures that search moral hazard problems will arise – consumption is equal across states of nature, while leisure is greater in the unemployed state. (Perhaps to emphasize nontraditional concerns, Blanchard and Tirole (2008) employ a utility model with a constant cash

equivalent value of leisure that induces neither search moral hazard nor firing cost problems (see Section Insurance Design below).)

Insurance Supply

The usual third-party suppliers of insurance products do not exist for job displacement insurance. Instead, three types of insurers are prominent: labor unions, employers, and the government. European trade unions were early pioneers in the provision of UI, but were never able to maintain reserves sufficient to finance benefits when times were hard. (Willoughby (1897) provides an interesting contemporaneous account of these efforts.) These failed efforts induced government intervention, the Ghent system, in which governments essentially financed union benefit programs. (See ILO (1955) for a broader historical review of early public UI programs.) In the US, unions have negotiated guaranteed employment contracts as well as severance pay and supplements to government-supplied UI benefits (supplemental unemployment benefits or SUB plans) (Pita, 1996). That said, today employers and governments are the main suppliers of job separation insurance.

The efficient provision of full job displacement insurance requires information on a variety of processes:

1. Was the worker involuntarily separated and therefore eligible for benefits?
2. Is the worker, while on benefits, actively seeking the next job and willing to accept reasonable job offers he or she uncovers?
3. Has the worker become reemployed, so that benefits can stop?
4. Is the drop in earnings following reemployment the consequence of a decline in wage rate rather than reduced work hours?

The first question is most easily answered by the employer, though the employer's response may be distorted by self-interest, depending on the financing system. For example, if benefits are limited to involuntary separations and partly or wholly paid by the employer, the employer has an incentive to categorize a separation as voluntary. If the layoffs are large scale, involuntary separations are hard to disguise, but that is not the case in small-numbers situations. Indeed, in the U.S. UI system, employers bear a substantial share of the cost of benefits paid out to their own laid-off workers (experience rating), and significant resources are required to adjudicate disputes over the nature of the separation. Conversely, if benefits are wholly paid by the government, the employer may be tempted to report any separation as involuntary, treating generously its own past workers at the expense of other taxpayers, typically firms.

UI requires substantial information following separation. The second and third questions seriously strain employer capabilities, and perhaps even those of governments. Even the conceptual simple task of determining whether the displaced worker has taken another job faces practical difficulties. In more developed economies, for example, undetectable jobs arise in the shadow economy of cash transactions. In less-developed economies, tracking workers across jobs of any sort may be difficult. (For recent estimates of the pervasiveness of the shadow economy, see Buehn and Schneider (2012).) Governments may be efficient at monitoring employment if the shadow economy is small and UI program coverage is broad. In that case, the worker's reemployment would be entered into the system at the time of the next hire, assuming of course unique individual identifiers.

Problem (2) is not easily monitored by any agent and has been the focus of much program design effort. In the simplest case, unemployment involves a release from work, and actual loss UI benefits must be limited in order to induce appropriate search and job acceptance behavior, with the reduced UI benefits partly and imperfectly offset by scheduled benefits (severance pay) (Baily, 1977; Rogerson and Schindler, 2002; Parsons, 2014b).

The provision of wage insurance pivots critically around the last question (4), the answer to which is typically unavailable to the displacing firm and often to the government. (Full reemployment wage insurance would require pooling of the earnings of the winners and the losers, but this has not been seriously proposed. Most policy designers propose a negative option payment – for those who make losses, some fraction of the losses will be paid through premiums from all (Baily et al., 1993). For additional discussions, see Parsons (2000), Kling (2006), and LaLonde (2007).) The U.S. UI system, for example, relies largely on earnings records for eligibility and benefit calculations, and not on work hours. But wage insurance is really wage rate insurance and assumes a return to full-time work. Otherwise it becomes a variant of UI with similar problems. The separation of total earnings into hours and wage rates is not a serious problem in large enterprises, which must themselves accurately track worker activities, but can be more difficult to verify in smaller enterprises. In this environment, wage insurance benefits are likely to be 'scheduled,' that is based on expectation of loss, rather than actual loss.

In broad outline, then, information problems are likely to shift actual loss unemployment provision to the government, and transform the insurance package from UI and wage insurance to (incomplete) UI and severance pay. Severance pay functions as scheduled (partial) UI and scheduled wage insurance.

Insurance Market Problems

Job Search and Offer Acceptance Incentives

The concern that generous UI benefits might discourage return to work has almost certainly limited unemployment benefits in countries that offer unemployment benefits. As a consequence, a great deal of thought has gone into the consideration of alternative mechanisms for inducing appropriate job search and offer acceptance by the unemployed. Two approaches have received attention – sanctions and active labor market policies. Sanctions typically involve monetary penalties, often in the form of reduced benefits, while active labor market policies involve time penalties.

Sanctions

An obvious approach is to impose penalties or sanctions on those who do not behave appropriately. Sanctions can be in the form of fines or, more commonly, reductions in future benefits. (Grubb (2000) provides a review of UI sanctions in OECD countries.) Denying or delaying benefits to workers who quit or are fired from their last jobs before entering unemployment is one such sanction. The withholding of benefits if workers do not undertake, or at least report, reasonable job search effort is another. The penalty may be the same for workers who quit their jobs and those sanctioned for lack of search effort (Grubb, 2000).

The U.S. UI system has search requirements, which vary by state, with the loss of future benefits as the sanction. A plan might require the unemployed recipient to make three employer search contacts in 1 week. Such requirements are not realistically enforceable in the US – it is apparently not possible to confirm or deny a large portion of claimed employer contacts (St. Louis et al., 1986). At a deeper level, it is a rare job that does not require some evidence of enthusiasm for the job by the applicant, but this is something easily withheld and impossible to monitor. Somewhat oddly, demonstration projects in the U.S. suggest that dropping the requirements *entirely* leads to large increases in unemployment duration, but requiring more than one search has no incremental impact on duration (Meyer, 1995).

Abbring et al. (2005) find, for the Netherlands, that the imposition of a sanction does significantly hasten the return to work, and moreover does so even after the current penalty has been lifted. They interpret this result to suggest that sanctioned workers perceive themselves as under extra scrutiny. One might imagine that well-publicized sanctions may have similar threat effects on others.

Active Labor Market Policies

A historically important technique for smoothing earnings shocks in limited-information situations is public works, with the government providing employment to workers in the absence of private demand. Public works, used on a large scale in the US and elsewhere during the Great Depression and often by colonial powers in times of crop failure, remain a potential tool in less-developed countries (Ravallion, 1991). Even if the work yields little product, it assures the provider that the beneficiary is willing to work and is not employed elsewhere. (Pissarides (2010) specifies conditions under which a private employer might undertake the equivalent of a public works project.)

The same approach has been extended to UI beneficiaries under the label 'active labor market policies,' although the work activity might instead involve mandatory training or search activities (see Kreiner and Tranaes (2005) for a discussion and analysis). Search moral hazard arises because workers value leisure as well as consumption, and unemployment by definition includes excess leisure. If the government wants to direct income to the unemployed without discouraging work effort, it can do so by imposing a countervailing work (activity) requirement on the beneficiary. Such requirements are labeled 'ordeals' by Nichols and Zeckhauser (1982).

As Besley and Coate (1992) nicely illustrate, ordeals work best, and perhaps only, if the insurer cares solely about income and consumption and not about worker utility, which would include leisure. For that reason, sanctions and ordeals appear in government-mandated programs, not voluntary contracts in which utility is paramount. Modern applications of this public works approach often soften the edges of the policy by labeling the work activity 'job training,' although the evidence on positive earnings effects of government job training programs is decidedly mixed. The signature of an ordeal is that the activity be mandatory, not voluntary, although gradations exist here. One might conjecture that the more resources the government devotes to the activity, whether job training or counseling, the more effective the program and the less onerous the ordeal.

There is perhaps a natural limit to the ordeal of sanctions and active labor-market policies. Imposing penalties on those one is trying to assist may be preferred to across-the-board benefit limitations, but is otherwise unappealing. In the best of circumstances, the threat of benefit sanctions or mandatory work requirements would induce ideal behavior without the threat ever being acted upon. (See, for example, Black et al. (2003). Such good fortune is perhaps rare.)

Firing Cost Issues

The importance of severance pay in the job displacement package is perhaps reflected in the pervasiveness of severance pay mandates and the voluntary provision of severance where it is not mandated. Economists, however, have regularly expressed concerns about mandated severance pay. (See, for example, Emerson (1988) and Lazear (1990) and the reviews in Heckman and Pagés (2004), Holzmann et al. (2012), and Parsons (2012a).) If employers must self-finance job displacement insurance, they might be tempted to retain workers in low-demand states who would, on efficiency grounds, be released to secure more productive matches elsewhere. Labor hoarding may be excessive.

Firing costs are often associated exclusively with severance pay, but this is a historical accident. In a model in which employers supply all separation benefits and cannot reinsure those benefits in third-party markets, firing costs would be the sum of severance pay and expected unemployment benefits. The concern about severance pay distortions arose with the practice, common in much of Europe, of mandating that firms must provide severance pay, leaving firms to fully fund the mandate. The European unemployment system, in contrast, has no user cost element; firms can lay off workers without any concern that higher benefit payouts to their workers will lead to higher taxes or surcharges. In that situation, firing costs are limited to severance pay.

In the US, severance pay is voluntary and is fully financed by the employer, so the same firing cost concerns arise. UI benefits are fully funded by taxes on payrolls, but the tax rate varies with the UI benefits distributed to the firm's laid-off workers ('experience rating') and will therefore enter firing cost calculations. Although severance pay is less generous in the US than in Europe, the experience rating of unemployment benefits raises the possibility that firing cost distortions may be higher in the US.

As it happens, severance pay alone appears to have only modest effects on labor hoarding, although it is important to recognize that large benefit mandates are found only in

countries in which labor laws are poorly enforced (Parsons, 2012a). Neither do firms engage in active avoidance of severance mandates, a possibility proposed by Lazear (1990). Apparently, they do not. That said, firing cost distortions may be a problem in substantially more generous benefit systems, an important caveat if one believes current levels are inadequate.

Insurance Design

A multiperiod framework permits more complex insurance schemes as well as introducing savings and borrowing strategies. Embedding the insurance contract in a multiperiod context introduces direct effects on severance pay schedules and, more subtly, on UI schedules. We consider insurance schemes in this section and saving and borrowing plans in the next.

Service-Conditioned Severance Benefits

The fact that reemployment wage losses increase systematically with job tenure (**Figure 2**) has the unsurprising consequence that severance benefits are almost always increasing in 'service' with the employer, essentially tenure. In voluntary plans in the US, a benefit rule of 1 or 2 weeks of pay per year of service is common (Parsons, 2005). The same basic weeks-pay-per-year-of-service algorithm, though with greater benefit levels, is common in mandated severance worldwide. See the scatter diagram of mandated benefits at 20 years relative to 5 years (**Figure 3**) for the sample of mandated plans in medium and large countries (1 million or more inhabitants, Holzmann et al., 2012: Annex B). The typical mandate is, however, well in excess of 1 week per year of service. Of the 118 countries with severance mandates and valid nonzero benefit information at 20 years of service, 90 (76%) mandate more than 20-weeks pay (Holzmann et al., 2012: Annex B). The median mandated benefit at 20 years of service in this sample of countries is 33.5-weeks pay.

UI Benefits over the Duration of Unemployment

Of (much) greater design subtlety, multiperiod models introduce interesting new UI benefit strategies. In a multiperiod setting, the benefit algorithm must specify a level of benefits for each week of unemployment. In practice, most government benefit schemes are quite simple. The U.S. benefit schedule has two steps, with zero benefits in the second step. The behavior of unemployed workers at or near the point at which benefits are 'exhausted' has been the focus of considerable empirical study. (Boone and van Ours (2012) provide a summary of this literature as well as their own contribution. See Mortensen (1977) and Meyer (1990) for theoretical and empirical analyses, respectively, of finite duration benefit schemes.) In a number of European countries, benefits are reduced more gently through a transition to social assistance, often means-tested welfare programs (Atkinson and Micklewright, 1991). (In the Holzmann et al. (2012) data, 16 of the 149 medium and large countries had both UI and unemployment assistance (author's calculations from Annex B). Another six had only unemployment assistance.)

Economists have devoted a great deal of attention to the ideal benefit payout stream as a function of unemployment

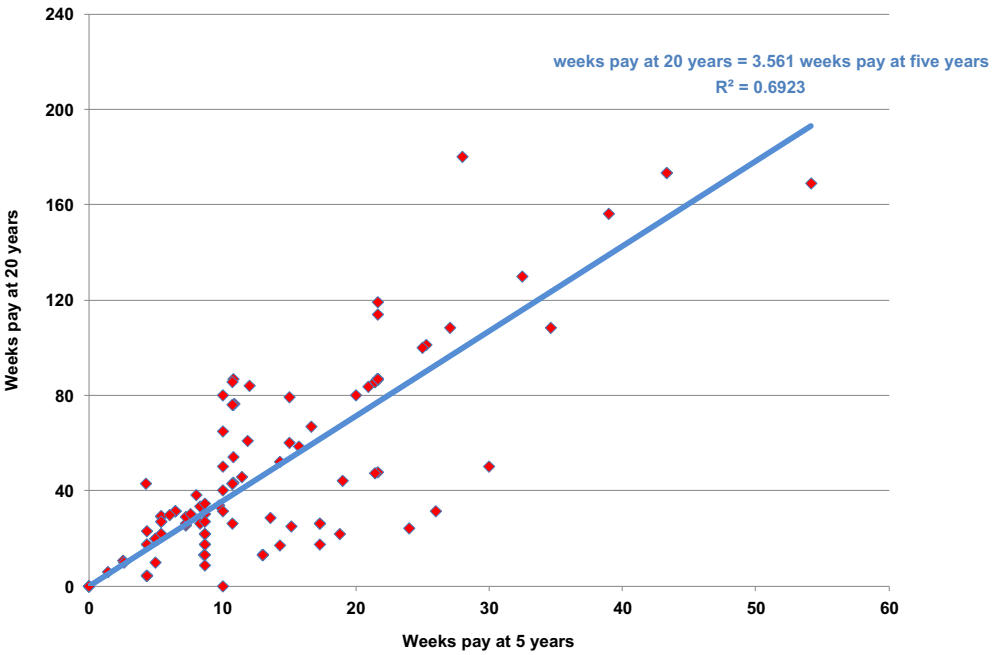

Figure 3 Mandated severance pay benefits (weeks pay) at 5 years and 20 years of service, 124 countries. Holzmann, Robert, Pouget, Yann, Weber, Michael, Vodopivec, Milan, 2012. Severance pay programs around the world: history, rationale, status, and reforms. In: Holzmann, Robert, Vodopivec, Milan (Eds.), Reforming Severance Pay: An International Perspective. World Bank, Washington, DC, pp. 17–120, Annex B.

spell duration. (Paradigm theoretical studies of optimal UI include Azariadis (1975), Baily (1977), Mortensen (1977), Shavell and Weiss (1979), and Hopenhayn and Nicolini (1997). Holmlund (1998), Karni (1999), and Fredriksson and Holmlund (2006) provide excellent reviews.) A key early finding was that UI benefits should optimally decline with the duration of unemployment (Shavell and Weiss, 1979). The intuition is straightforward. By throwing benefits forward in time in the unemployment spell, workers have less incentive to delay return to work despite currently generous benefits. Reduced future benefits are a threat that may not be realized. Early theoretical contributions, including Shavell and Weiss, assumed away the possibility that the worker could transfer resources across time except through the program under study. The need to repress the possibility of private saving is obvious. The worker, post-separation, has no desire to be squeezed in this way and would likely save early UI benefits in order to supplement low benefits in later periods. (Pavoni (2007) expresses concern that declining benefit schemes in the saving-constrained case might lead to unacceptably low levels of benefits for longer unemployment spells. This problem is less serious if the worker is permitted to save.) This no-saving assumption becomes formally awkward in a job displacement setting in which severance pay is almost certainly part of the optimal vector of benefits. (The fixed-sum benefits may be paid out as a lump-sum or as periodic payments, but without reference to unemployment status.)

Hopenhayn and Nicolini (1997) analyzed an environment in the spirit of Shavell and Weiss, with no private saving or borrowing, but introduced the possibility that the insurer, presumably the government, can pay out additional benefits early and finance the expense by taxing future worker earnings. The optimal benefit scheme will still involve declining benefits, but, at least in the example developed, at a rate much lower than in the Shavell and Weiss model. The penalty for engaging in longer unemployment spells is spread over both lower expected future unemployment benefits and lower future after-tax earnings. This model also unravels if the worker can privately reallocate resources across periods, and perhaps has assets to start – as the worker would, for example, with severance payments.

In what may now be considered an early attempt to apply behavioral economics, great effort has also gone into assessing empirically a quite different, discrete payment plan – reemployment bonuses – which offers a lump-sum reward if a worker takes another job in a specified time interval. An eye-catching bonus to successful early job finders may accelerate job search and acceptance sufficiently that the bonuses pay for themselves. One early demonstration project to test this conjecture, Illinois, raised hope that the approach might be of practical value (Woodbury and Spiegelman, 1987), and led to an ambitious series of demonstration projects that generated much valuable insight into the behavior of the unemployed. (Meyer (1995) provides a thorough review of the demonstration projects and their outcomes.) Alas, the demonstrations provided no support for the validity of the core proposition. Lump-sum bonuses for early job acceptance do not usually pay for themselves.

Saving and Borrowing Strategies

Mandated saving and borrowing options have received considerable attention as *alternative* mechanisms for helping unemployed workers smooth consumption in response to earnings shocks (Feldstein and Altman, 2007; Stiglitz and Yun, 2005). The theoretical advantages of unemployment savings accounts and severance savings accounts are that they should induce neither (1) search moral hazard nor (2) firing cost concerns. The reason of course is that the worker already owns the payout, which means the worker cannot capture additional benefit payouts by (inefficiently) remaining unemployed, and the employer cannot avoid benefit payouts by (inefficiently) retaining the worker on the payroll (Parsons, 2012b).

Savings accounts may be an excellent substitute for earnings insurance *for temporary layoffs* in which the worker returns to the contract firm, presumably at the original wage. UI benefits paid to workers on temporary layoffs amount to externally financed vacation pay for a firm's employees, and are likely to be heavily used (Feldstein, 1976). The puzzle that remains is why workers and governments are so reluctant to embrace the savings approach – even for temporary layoffs. Savings are, however, a relatively poor tool for smoothing consumption among long-tenured, displaced workers. That said, if otherwise uninsured, even large shocks are eased by access to savings and credit markets.

As noted in the last section, the worker's ability to save privately fundamentally changes what the government can accomplish. Penalizing extended unemployment through declining consumption, as in Shavell and Weiss (1979) and Hopenhayn and Nicolini (1997), may not be feasible. If the government administrator can observe assets, the declining unemployment benefit plan can be salvaged if the government is willing to tax away positive assets as they accumulate (insuring that none will have any), though that seems unlikely as a public policy.

Insurance plans and saving and borrowing plans are not mutually exclusive, and can be offered jointly. Worker access to savings instruments changes the profile of optimal UI benefits. In a model with no moral hazard issues, Shavell and Weiss (1979) demonstrated that positive UI benefits should be offered only after assets are depleted, and then at a fixed benefit level. (Unless assets are sufficiently small that they can effectively be ignored.) First exhausting assets and then accessing insurance makes sense in this model because the worker has no alternative uses for the savings (no retirement or bequest, no risk of future layoffs or wage rate declines). Shavell and Weiss (1979) did not analyze the (search-effort) information-constrained case, but provide a two-period model that demonstrates that optimal benefits could either increase or decrease with longer unemployment spells, depending on circumstance such as initial wealth position. Lentz and Tranaes (2005) developed a model of job search and asset choice with fixed unemployment benefits of unlimited duration that generates the plausible prediction that savings will be used to support consumption during an unemployment spell, with search effort increasing as savings dwindle. It is plausible that workers would demand increasing UI benefits across unemployment spells in this situation, and more generally negotiate

benefit plans that are conditioned on asset levels (Rendahl, 2012).

If the displaced worker is able to 'hide' assets, the nature of the optimal benefit stream is apparently highly model-specific. Kocherlakota (2004: p. 541) reported being able to "provide a complete characterization of the optimal contract" if search effort is linear in the probability of securing a job offer (and auxiliary assumptions to deal with the linearity problem are met). "The agent's consumption is constant while he is unemployed, and jumps to a higher constant and history-independent level of consumption when he finds a job." Shimer and Werning (2008)constructed a model with a simple search algorithm and a constant absolute risk aversion utility function, which has the convenient property that the solution, in this case of constant benefits throughout the unemployment spell, does not depend on worker wealth – which makes moot the question of whether assets are observed or hidden. Mitchell and Zhang (2010), in a model in many ways similar to Shimer and Werning (2008), concluded that the optimal time profile of benefits should be increasing over the unemployment spell. As a practical matter, approximately 50% of workers surveyed in the Displaced Workers Survey exhaust their unemployment benefits (GAO, 2012), so the exact shape of the benefit algorithm across unemployment spells may not be as important for permanent layoffs as it could be for temporary layoffs. UI is a fixed sum for UI exhaustees – that is to say, severance pay.

Aging and the Long-Tenured Displaced Worker

Long-tenured displaced workers are, by the very nature of the event, older, and issues of aging and of the job displacement of long-tenured workers overlap. At the simplest level, there is the concern that older workers may be disadvantaged in the labor market. The measured impact of job displacement on older workers who return to work is large and justifies concerns about the welfare of these individuals (Couch, 1998), although perhaps no more than younger workers. Chan and Stevens (1999, 2001) analyzed the consequence of job loss on older workers (HRS). Among workers in their 50s or beyond, reemployment wage losses of displacement workers were on the order of 24% (relative to the not displaced) with a mean duration of unemployment of 12 months (Chan and Stevens, 1999: pp. 212–213).

Measurement of job displacement effects is difficult because a greater fraction of older workers than others never return to work. That may be because jobs are scarce or because older workers have access to a variety of alternative support systems, including disability insurance and private and public retirement plans, which ease withdrawal from the labor force. Borrowing against pensions and other postwork income streams can also smooth consumption during unemployment periods if feasible (Stiglitz and Yun, 2005). Hairault et al. (2012) argue this freedom is critical if the worker is close to retirement.

Recognizing this demographic reality, firms often offer early retirement buyouts, under the belief, apparently true, that workers relatively close to retirement are willing to separate voluntarily with modest inducements (the shadow value of returning to work is low). Public systems also treat older workers more generously, reflecting both (1) concern for older individuals and (2) the possibility that encouraging older workers to leave the labor force will open up jobs for the young (the fixed jobs hypothesis). Early retirement programs, expanded unemployment duration eligibility, reduced active search requirements – all contribute to earlier withdrawal from the workforce (OECD, 2006).

Older displaced workers do apparently respond robustly to these program incentives and in the expected direction. As Tatsiramos (2010: p. 157) summarizes his cross-country findings:

> … in countries with relatively more generous unemployment insurance provisions for the older unemployed, which offer a pathway to early retirement, older displaced workers exhibit lower re-employment and higher retirement rates compared to the non-displaced.

The welfare implications of these incentives and responses are less obvious, and may depend on the macroeconomic model felt to be relevant in the economy.

See also: Earnings Inequality; Insurance; Retirement and Encore Adulthood: The New Later Life Course; Retirement and Health; Retirement, Economics of; Social Insurance: Legal Aspects.

Bibliography

Abbring, Jaap H., van den Berg, Gerard J., van Ours, Jan C., July 2005. The effect of unemployment insurance sanctions on the transition rate from unemployment to employment. Economic Journal 115 (505), 602–630.

Atkinson, Anthony B., Micklewright, John, December 1991. Unemployment compensation and labor market transitions: a critical review. Journal of Economic Literature 29, 1679–1727.

Azariadis, Costas, December 1975. Implicit contracts and underemployment equilibria. Journal of Political Economy 83, 1183–1202.

Baily, Martin N., July 1977. Unemployment insurance as insurance for workers. Industrial and Labor Relations Review 30, 495–504.

Baily, Martin N., Burtless, Gary, Litan, Robert E., 1993. Growth with Equity: Economic Policymaking for the Next Century. Washington, DC, 110–141.

Besley, Timothy, Coate, Stephen, 1992. Workfare versus welfare: incentive arguments for work requirements in poverty-alleviation programs. American Economic Review 82, 249–261.

Bishow, John, Parsons, Donald O., May 2004. Trends in Severance Pay Coverage in the United States. 1980–2001. Available at SSRN: http://ssrn.com/abstract=878144.

Black, Dan A., Smith, Jeffrey A., Berger, Mark C., Noel, Brett J., November 2003. Is the threat of reemployment services more effective than the services themselves? Evidence from random assignment in the UI system. American Economic Review 93 (4), 1313–1327.

Blanchard, Olivier J., Tirole, Jean, March 2008. The joint design of unemployment insurance and employment protection: a first pass. Journal of the European Economic Association 6 (1), 45–77.

Bloemen, Hans G., Stancanelli, Elena G.F., 2005. Financial wealth, consumption smoothing and income shocks arising from job loss. Economica 72, 431–452.

Blundell, Richard, Pistaferri, Luigi, Preston, Ian, December 2008. Consumption inequality and partial insurance. American Economic Review 98 (5), 1887–1921.

Boone, Jan, van Ours, Jan C., December 2012. Why is there a spike in the job finding rate at benefit exhaustion? De Economist 160 (4), 413–438.

Browning, Martin, Crossley, Thomas F., April 2001. Unemployment insurance benefit levels and consumption changes. Journal of Public Economics 80 (1), 1–23.
Browning, Martin, Crossley, Thomas F., July 2008. The long-run cost of job loss as measured by consumption changes. Journal of Econometrics 145 (1–2), 109–120.
Buehn, Andreas, Schneider, Friedrich, February 2012. Shadow economies around the world: novel insights, accepted knowledge, and new estimates. International Tax and Public Finance Springer, 19 (1), 139–171.
Bureau of Labor Statistics (U.S. Department of Labor), 2012. Worker Displacement: 2009–2011. BLS news release USDL-12-1719.
Card, David, Chetty, Raj, Weber, Andrea, November 2007. Cash-on-hand and competing models of intertemporal behavior: new evidence from the labor market. Quarterly Journal of Economics 122 (4), 1511–1560.
Chan, Sewin, Stevens, Ann H., May 1999. Employment and retirement following a late-career job loss. American Economic Review 89 (32), 211–216.
Chan, Sewin, Stevens, Ann H., April 2001. Job loss and employment patterns of older workers. Journal of Labor Economics 19 (2), 484–521.
Chetty, Raj, April 2008. Moral hazard versus liquidity and optimal unemployment insurance. Journal of Political Economy 116 (2), 173–234.
Cochrane, John H., October 1991. A simple test of consumption insurance. Journal of Political Economy 99 (5), 957–976.
Couch, Kenneth A., February 1998. Late life job displacement. Gerontologist 38 (1), 7–17.
Couch, Kenneth A., Placzek, Dana W., 2010. Earnings losses of displaced workers revisited. American Economic Review 100 (1), 572–589.
Cullen, Julie B., Gruber, Jonathan, July 2000. Does unemployment insurance crowd out spousal labor supply? Journal of Labor Economics 18 (3), 546–572.
Davis, Steven, Von Wachter, Till, 2011. Recessions and the costs of job loss. Brookings Papers on Economic Activity (Fall), 1–55.
Emerson, Michael, 1988. Regulation or deregulation of the labour market: policy regimes for the recruitment and dismissal of employees in the industrialized countries. European Economic Review 32, 775–817.
Engen, Eric M., Gruber, Jonathan, June 2001. Unemployment insurance and precautionary saving. Journal of Monetary Economics 47 (3), 545–579.
Fallick, Bruce C., October 1996. A review of the recent empirical literature on displaced workers. Industrial and Labor Relations Review 50 (1), 5–16.
Farber, Henry S., 2004. Job loss in the United States, 1981–2001 accounting for worker well-being. Research in Labor Economics 23, 69–117.
Feldstein, Martin, October 1976. Temporary layoffs in the theory of unemployment. Journal of Political Economy 84, 937–958.
Feldstein, Martin, Altman, Daniel, 2007. Unemployment insurance savings accounts. Tax Policy and the Economy 21, 35–63.
Flemming, J.S., 1978. Aspects of optimal unemployment insurance. Journal of Public Economics 10, 403–425.
Fredriksson, Peter, Holmlund, Bertil, 2006. Improving incentives in unemployment insurance: a review of recent research. Journal of Economic Surveys 20 (3), 357–386.
Government Accountability Office (U.S), February 2012. Unemployment Insurance: Economic Circumstances of Individuals Who Exhausted Benefits, Report to the Chair, Committee on Finance, U.S. Senate. GAO, 12–408.
Grubb, David, 2000. Eligibility criteria for unemployment benefits. OECD Economic Studies (31), 147–184.
Gruber, Jonathan, March 1997. The consumption smoothing benefits of unemployment insurance. American Economic Review 87 (11), 192–205.
Gruber, Jonathan, October 2001. The wealth of the unemployed. Industrial and Labor Relations Review 55 (1), 79–94.
Hairault, Jean-Olivier, Langot, Francois, Menard, Sebastien, Sopraseuth, Thepthida, June 2012. Optimal unemployment insurance for older workers. Journal of Public Economics 96 (5–6), 509–519.
Hansen, Gary D., Imrohoroglu, Ayse, February 1992. The role of unemployment insurance in an economy with liquidity constraints and moral hazard. Journal of Political Economy 100 (1), 118–142.
Heckman, James J., Pages, Carmen, 2004. Introduction. In: Heckman, Pages (Eds.), Law and Employment: Lessons from Latin America and the Caribbean. The University of Chicago, Chicago, pp. 1–107.
Helwig, Ryan T., June 2001. Worker displacement in a strong labor market. Monthly Labor Review 124 (6), 13–28.
Holmlund, Bertil, 1998. Unemployment insurance in theory and practice. Scandinavian Journal of Economics 100 (1), 113–141.
Holzmann, Robert, Pouget, Yann, Weber, Michael, Vodopivec, Milan, 2012. Severance pay programs around the world: history, rationale, status, and reforms. In: Holzmann, Robert, Vodopivec, Milan (Eds.), Reforming Severance Pay: An International Perspective. World Bank, Washington, DC, pp. 17–120.
Hopenhayn, Hugo A., Nicolini, Juan P., April 1997. Optimal unemployment insurance. Journal of Political Economy 105 (2), 412–438.
International Labour Office (ILO), 1955. New Series, No. Unemployment Insurance Schemes. Studies and Reports, 42. Geneva International Labour Office.
Jacobson, Louis S., LaLonde, Robert J., Sullivan, Daniel G., September 1993a. Earnings losses of displaced workers. American Economic Review 83 (4), 685–709.
Jacobson, Louis S., LaLonde, Robert J., Sullivan, Daniel G., November/December 1993b. Long-term earnings losses of high-seniority displaced workers. Economic Perspectives 17, 2–20.
Karni, Edi, October 1999/2006. Optimal unemployment insurance: a survey. Southern Economic Journal 66, 442–465.
Kletzer, Lori G., 1998. Job displacement. Journal of Economic Perspectives 12 (n1) (Winter), 115–136.
Kling, Jeffrey R., 2006. Fundamental Restructuring of Unemployment Insurance: Wage-Loss Insurance and Temporary Earnings Replacement Accounts. The Hamilton Project, Brookings Institution. Discussion Paper 5 September 2006.
Kocherlakota, Narayana R., July 2004. Figuring out the impact of hidden savings on optimal unemployment insurance. Review of Economic Dynamics 7 (3), 541–554.
Kreiner, Claus T., Tranaes, Torben, September 2005. Optimal workfare with voluntary and involuntary unemployment. Scandinavian Journal of Economics 107 (3), 459–474.
Kuhn, Peter J., 2002. Summary and synthesis. In: Losing Work, Moving on: International Perspectives on Worker Displacement. W.E. Upjohn Institute for Employment Research, Kalamazoo, Michigan, pp. 1–103.
LaLonde, Robert J., 2007. The Case for Wage Insurance. Maurice R. Greenberg Center for Geoeconomic Studies at the Council on Foreign Relations. http://www.cfr.org/publication/13661/case_for_wage_insurance.html.
Lazear, Edward P., August 1990. Job security provisions and employment. Quarterly Journal of Economics 105, 699–726.
Lentz, Rasmus, Tranæs, Torben, July 2005. Job search and savings: wealth effects and duration dependence. Journal of Labor Economics 23 (3), 467–489.
Mace, Barbara J., October 1991. Full insurance in the presence of aggregate uncertainty. Journal of Political Economy 99 (5), 928–956.
Malcomson, J., 1999. Individual employment contracts. In: Ashenfelter, Orley, Card, David (Eds.), Handbook of Labor Economics, vol. IIIB. North Holland Press, pp. 2291–2372 (Chapter 35).
Mitchell, Matthew, Zhang, Yuzhe, November 2010. Unemployment insurance with hidden savings. Journal of Economic Theory 145 (6), 2078–2107.
Mortensen, Dale T., July 1977. Unemployment insurance and the job search decision. Industrial and Labor Relations Review 30, 505–517.
Meyer, Bruce D., July 1990. Unemployment insurance and unemployment spells. Econometrica 58 (4), 757–782.
Meyer, Bruce D., March 1995. Lessons from the U.S. Unemployment insurance experiments. Journal of Economic Literature 33 (1), 91–131.
Nichols, Albert L., Zeckhauser, Richard J., May 1982. Targeting transfers through restrictions on recipients. American Economic Review 72 (2), 372–377.
OECD, 2006. Employment Outlook 2006.
Parsons, Donald O., 1986. The employment relationship: job attachment, work effort, and the nature of contracts. In: Ashenfelter, Orley, Layard, Richard (Eds.), Handbook of Labor Economics, vol. II. North Holland Press, pp. 789–848.
Parsons, Donald O., 2000. Wage insurance: a policy review. Research in Employment Policy 2, 119–140.
Parsons, Donald O., December 2005. Benefit Generosity in Voluntary Severance Plans: The U.S. Experience. http://ssrn.com/abstract=877903.
Parsons, Donald O., 2012a. Mandated severance pay and firing cost distortions: a critical review of the evidence. In: Holzmann, Robert, Vodopivec, Milan (Eds.), Reforming Severance Pay: An International Perspective. World Bank, Washington, DC, pp. 121–157 (Chapter 3).
Parsons, Donald O., 2012b. The firing cost implications of alternative severance pay designs. In: Holzmann, Robert, Vodopivec, Milan (Eds.), Reforming Severance Pay: An International Perspective. World Bank, Washington, DC, pp. 159–176 (Chapter 4).
Parsons, Donald O., September–December, 2013. Understanding severance pay. Cuadernos de Economía (Spanish Journal of Economics and Finance) Elsevier 36 (106), 155–165.
Parsons, Donald O., 2014a. Job Displacement Insurance: A Policy Taxonomy. Mimeo, Economics Department, George Washington University.
Parsons, Donald O., 2014b. The Simple Analytics of Job Displacement Insurance. Mimeo, George Washington University.
Pavoni, Nicola, September 2007. On optimal unemployment compensation. Journal of Monetary Economics 54 (6), 1612–1630.

Pissarides, Christopher A., 2010. Why do firms offer 'employment protection'? Economica 77, 613–636.
Pita, Cristina, July 1996. Advance notice and severance pay provisions in contracts. Monthly Labor Review 119.
Ravallion, Martin, July 1991. Reaching the rural poor through public employment: arguments, evidence, and lessons from South Asia. World Bank Research Observer 6 (2), 153–175.
Rendahl, Pontus, August 2012. Asset-based unemployment insurance. International Economic Review 53 (3), 743–770.
Riddell, Chris, Kuhn, Peter J., January 2010. The long-term effects of unemployment insurance: evidence from New Brunswick and Maine, 1940–1991. Industrial and Labor Relations Review 63 (2), 183–204.
Rogerson, Richard, Schindler, Martin, September 2002. The welfare costs of worker displacement. Journal of Monetary Economics 49 (6), 1213–1234.
Rosen, Sherwin, September 1985. Implicit contracts: a survey. Journal of Economic Literature 23, 1144–1175.
Ruhm, Christopher J., March 1991. Are workers permanently scarred by job displacements? American Economic Review 81 (1), 319–324.
St. Louis, Robert D., Burgess, Paul L., Kingston, Jerry L., 1986. Reported vs. Actual job search by unemployment insurance claimants. Journal of Human Resources 21 (1), 92–117 (Winter).
Salanié, Bernard, 2005. The Economics of Contracts: A Primer, second ed. MIT Press.
Shavell, Steven, Weiss, Laurence, December 1979. The optimal payment of unemployment insurance benefits over time. Journal of Political Economy 87, 1347–1362.
Shimer, Robert, Werning, Ivan, December 2008. Liquidity and insurance for the unemployed. American Economic Review 98 (5), 1922–1942.
Stephens Jr., Melvin, February 2001. The long-run consumption effects of earnings shocks. Review of Economics and Statistics 83 (1), 28–36.
Stephens Jr., Melvin, February 2004. Job loss expectations, Realizations, and household consumption behavior. Review of Economics and Statistics 86 (1), 253–269.
Stevens, Ann H., January 1997. Persistent effects of job displacement: the importance of multiple job losses. Journal of Labor Economics 15 (1, Part 1), 165–188.
Stiglitz, Joseph E., October 1969. Behavior towards risk with many commodities. Econometrica 37 (4), 660–667.
Stiglitz, Joseph E., Yun, Jungyoll, December 2005. Integration of unemployment insurance with retirement insurance. Journal of Public Economics 89 (11–12), 2037–2067.
Sullivan, James X., 2008. Borrowing during unemployment: unsecured debt as a safety net. Journal of Human Resources 43 (2), 383–412 (Spring).
Tatsiramos, Konstantinos, May 2010. Job displacement and the transitions to re-employment and early retirement for non-employed older workers. European Economic Review 54 (4), 517–535.
Willoughby, William F., September 1897. Insurance against unemployment. Political Science Quarterly 12 (3), 476–489.
Woodbury, Stephen A., Spiegelman, Robert G., September 1987. Bonuses to workers and employers to reduce unemployment: randomized trials in Illinois. American Economic Review 77 (4), 513–530.

Job Satisfaction

Howard M Weiss and Kelsey L Merlo, Georgia Institute of Technology, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Research on job satisfaction, or the positive or negative evaluation one makes about one's job, began in the early 1930s and has been a productive area of research for decades. In this article, the concept is defined, the history of the research on job satisfaction is described, and major theoretical positions and empirical findings are presented.

Definition

Among the various concepts that can describe how people think about and relate to their work, psychologists have found it useful to include an overall evaluative judgment about one's job. This evaluation is labeled job satisfaction and, thus, job satisfaction can be defined as *a positive or negative evaluation one makes about one's job or job situation.* Conceived this way, job satisfaction is an attitude. Therefore thinking about the measurement of satisfaction or developing theories of its causes and consequences are best approached by recognizing that the concept of job satisfaction falls under the broad umbrella of general attitude theory. This has not always been understood and as a consequence, unnecessary confusion about the nature of satisfaction and its relation to related but different concepts like affect has existed. However, more conceptual clarity has developed in the last 20 years, and the definition of satisfaction an attitude or an evaluative judgment has been broadly accepted.

Traditionally, psychologists have made a distinction between *overall satisfaction*, satisfaction with one's job as a whole, and *facet satisfactions*, expressed satisfactions with the aspects of one's job (e.g., pay, supervision, the work one does, etc.). These differences simply relate to the differences in the object of evaluation. In the past, the language surrounding these concepts has created confusion by implying some part–whole relationship among the various attitudes, as if the overall attitude was composed of the facet attitudes. Although beliefs and attitudes about various work features may influence overall evaluations about jobs, they are only one such influence. Attitudes are always evaluations of something, be they concrete objects like coworkers or abstract concepts like jobs or career opportunities, and any element of work experience that is at all discriminable can be an object of evaluation. Understood this way, when psychologists focus on the evaluation of the totality of a job, they are referring to overall job satisfaction. When they focus on the evaluation of a specific aspect of job experience, they are referring to some facet satisfaction.

Psychologists have also developed a number of other related concepts to describe the aspects of work experience, such concepts are commitment, engagement, work subjective well-being, and so on. In some cases, the differences from satisfaction are subtle but important. In other cases, the differences are more apparent than real, as evaluation turns out to be the disguised but essential defining feature. Only one distinction is important to raise in this article and that is the difference between affect (moods and emotions) and satisfaction, as satisfaction has sometimes been erroneously defined as an affective or emotional reaction to one's job. Satisfaction, as an attitude, is an evaluative judgment. Moods and emotions are affective states with discernible beginnings and endings and the subjective experience of particular feelings. Such affective experiences can influence attitudes but they are not equivalent to the attitudes.

Measuring Job Satisfaction

As is typical with attitudes generally, job satisfaction is most frequently measured using self-reports in questionnaire format. Satisfaction is assessed both in research and organizational practice contexts (e.g., organization-specific morale surveys, climate surveys, etc.), and while researchers tend to use a few well-developed, widely available measures, organizations tend to use proprietary measures either developed in-house or by consulting firms.

Measures used in the research community tend to be of two types: overall satisfaction measures or facet-based measures. Overall satisfaction measures ask for overall evaluations of one's job directly. That is, they specifically reference the totality of one's job as the object of evaluation. One popular example is the 'faces' scale (Kunin, 1955), which asks respondents to use a series of scaled, smiling faces to indicate their overall evaluation. The faces response format can be used to measure evaluations of particular facets too, but it is frequently used as a measure of overall satisfaction. Another popular measure is the Job in General Scale (Ironson et al., 1989) patterned after the facet-focused Job Descriptive Index (see below).

Facet measures assess evaluations of specific aspects of job experience (e.g., pay, supervision, coworkers, etc.). Generally, overall satisfaction is derived as a sum of the individual facet evaluations. The two most popular facet satisfaction measures are the Minnesota Satisfaction Questionnaire (MSQ; Lofquist and Dawis, 1969), which assesses 20 job satisfaction facets and is available in both long and short forms, and the Job Descriptive Index (Smith et al., 1969), which measures five facets. These measures are convenient because they assess both facet evaluations and overall job evaluations and have long histories of use. However, the assessment of overall satisfaction through facet summation is problematic because it assumes that people's overall satisfaction is derived from some mental calculation involving the facet evaluations. There is good reason to suppose this not to be true. Indeed, the Job in General Scale for measuring overall satisfaction directly was developed as a response to this concern. In addition to these rigorously developed overall and

facet measures, there are a number of single- and multiple-item assessments of facet and overall job satisfaction that have been developed in an ad hoc manner and appear in the literature.

History of Research on Job Satisfaction

Research on job satisfaction began in earnest in the early 1930s and was heavily influenced both by the economic and employment crises of the depression and by the new developments in attitude measurement (e.g., work by Thurstone and Likert, in particular.) The economic situation provided an important reason for studying employee 'morale,' while the advances in attitude scaling provided an approach to the task, even if that approach was not often followed rigorously.

Interestingly, research published in the 1930s showed an interest in more varied aspects of work experience than what is captured by the construct of job satisfaction. For example, Fisher and Hannah examined the relationship between 'emotional maladjustment' and job satisfaction in their clinical studies reported in their book *The Dissatisfied Worker* (1931). Rexford Hersey (1932), using what would today be called mood checklists, studied the daily rhythms of worker affect and related those rhythms to daily changes in productivity, physiological states, and off-work life events. Hoppock (1935), interested in the far-reaching consequences of the depression, focused his attention on the social implications as opposed to the organizational implications of job dissatisfaction.

Later, interest in worker experience narrowed to emphasize evaluative judgments of the aspects of work, i.e., job satisfaction and the causes and consequences of those attitudes. Perhaps the research most influential for the development of the dominant paradigm was conducted by Kornhauser and Sharp (1932). In their research, conducted at Kimberley Clark, Kornhauser and Sharp laid the foundation for what was to come by using questionnaires to ask about facets of job satisfaction and relating those facets to aspects of performance of interest to management. The examination of work attitudes became the primary vehicle for studying work experience, questionnaires became the primary method, facet measurement became the primary content, and organizational effectiveness became the primary reason for the research. By the end of the decade, an eclectic stew of thinking about worker experience had boiled down to a puree composed of worker attitudes or job satisfaction. Satisfaction became the predominant way of capturing worker experiences and the research after that mostly concentrated on the predictors and performance correlates of satisfaction, how best to measure satisfaction, and the development of theories of satisfaction formation.

The 1940s and 1950s saw mostly empirical efforts to catalog the various work environment features and personal characteristics that predict differing levels of work satisfaction. There were also continued empirical efforts to demonstrate the ability of satisfaction to predict job performance in its various forms. Although even Kornhauser and Sharp (1932) in their landmark study reported an inability of satisfaction to predict performance, there existed a generally accepted belief that workers would perform better and be more motivated in their work, if they found satisfaction in their jobs. The reasons for this assumption were based more on intuition than sound conceptual analysis. Then in 1955, Brayfield and Crockett published an influential review of studies correlating satisfaction and performance and found the relationship to be negligible. Subsequent reviews by Vroom (1964) and Iaffaldano and Muchinsky (1985) came to the same conclusion. Although later reviews (for example, Judge et al., 2001) may be more optimistic about the relationship between satisfaction and performance, by the late 1950s it became clear that more theoretical development of the concept was needed, including better understanding of the development of work attitudes and a clearer picture of the many ways in which employees contribute to organizational effectiveness might be predicted by levels of satisfaction.

In this regard, the 1960s might be considered the 'golden age' of satisfaction research as psychologists mostly abandoned the empirical efforts to catalog the correlates of satisfaction and began to develop influential theoretical positions. In 1959, Herzberg, Mausner, and Snyderman introduced the 'two-factor' theory, which dominated research and practice for over a decade. Important theoretical positions were offered by Porter (1961), Adams (1963, 1965), Katzell (1964), and Locke (1969) in a tradition that Lawler (1973) labeled 'discrepancy theories' to illustrate that they all involved workers comparing assessments of working conditions to some personal standard. Vroom (1964) developed a theory of satisfaction out of his broader expectancy theory of motivation. The 1960s also saw the development of the two most important theoretically driven measurement systems, the Job Descriptive Index (Smith et al., 1969) and the MSQ (Lofquist and Dawis, 1969). Important theoretical developments continued in the 1970s, particularly with the introduction of social information processing theory by Salancik and Pfeffer (1977, 1978), which emphasized the importance of the social context of satisfaction judgments, and the job characteristics model of Hackman and Oldham (1975), which emphasized the importance of features of the tasks that people work on.

Finally, after a period of theoretical and empirical quiescence, the 1990s saw psychology's renewed interest in affective states thereby making its way to the study of work experience. In some ways this helped work psychology come full circle to the initial interests of Hersey (1932) or Fisher and Hanna (1931), broadening discussions of work experience beyond the confines of satisfaction. More importantly for satisfaction research, the interest in affect helped bring clarity to the construct of satisfaction, positioning it as an attitude or evaluative judgment, influenced by but not equivalent to affective states. Important overarching discussions of this were presented by Weiss and Cropanzano (1996) with the presentation of Affective Events Theory (AET) and then in more detail by Weiss (2002). Now, as indicated earlier, job satisfaction is one of a number of psychological constructs used to understand the subjective experiences of workers. A thorough history of the research on job satisfaction can be found in Weiss and Brief (2001).

Consequences of Differing Levels of Job Satisfaction

An overriding impetus for the study of job satisfaction has been a belief that it is an important influence on job performance.

Consequently, the nature of that relationship continues to be an important direction for the application of empirical and conceptual efforts. As employees can contribute to organizational effectiveness in many ways, through their task performance, commitment to the organization, reliability of attendance, provision of support and mentoring, and so on, no single statement can be made about satisfaction and performance. Instead, the importance of work attitudes to performance appears to be a function of what aspect of performance one is referring to.

Initial interest focused on what might be called immediate task performance and by 1955 enough empirical studies of that relationship had been done to justify a review by Brayfield and Crockett. Their conclusion was rather surprising for the time. "It appears that there is little evidence in the available literature that employee attitudes of the type usually measured in morale surveys bear any simple – or for that matter, appreciable – relationship to performance on the job." (1955: p. 408). Later Vroom (1964) reviewing essentially the same literature came to the same conclusion as did an early *meta*-analysis by Iaffaldano and Muchinsky (1985). More recently, a *meta*-analysis by Judge et al. (2001) reported a more moderate satisfaction–performance relationship ($r = 0.30$, corrected). They suggest that the difference is due primarily to methodological differences (Judge et al., 2001). A *meta*-analysis by Harter et al. (2002) found a moderate relationship between employee satisfaction and unit-level performance ($r = 0.22$). Regardless, it is important to note that even the most optimistic results find satisfaction explaining less than 10% of the variance in task performance. Further, most of the data points in these *meta*-analyses are from correlational studies, leaving causal direction ambiguous. Finally, Bowling (2007) suggests that the satisfaction–performance relationship is largely spurious. He found that the relationship between satisfaction and performance dropped to 0.09 after controlling for domain-specific personality traits (i.e., organization-based self-esteem), suggesting that satisfaction may not be causally related to performance. Overall, and contrary to intuition, these results suggest modest to negligible relationships between job satisfaction and task performance.

Employee contributions to organizational effectiveness go beyond the contributions made by immediate task performance and therefore research has also looked at the ability of satisfaction to predict other criteria like job withdrawal (turnover, absenteeism, etc.) and extra-role behaviors (organizational citizenship behavior (OCB) and counterproductive work behavior (CWB)). *Meta*-analyses of studies examining satisfaction–turnover relationships generally produce findings similar to those for task performance, moderate but not substantial. For example, Carsten and Spector (1987) found that turnover was related to satisfaction ($r = -0.26$, corrected) and Tett and Meyer (1993) found that turnover intentions and withdrawal intentions were highly correlated with job satisfaction ($r = -0.58$, corrected), but observed turnover less so ($r = -0.24$). A number of other *meta*-analytic studies found relationships of a similar magnitude (Cotton and Tuttle, 1986; Fried et al., 2008; Griffeth et al., 2000; Hom et al., 1992; Podsakoff et al., 2009). Overall, satisfaction is most strongly related to turnover intention, not actual turnover. The attenuated relationship between intentions and behavior may be due to a myriad of other factors that are outside of the individual's control (i.e., ability to get another job).

Similar to the satisfaction–turnover relationship, satisfaction has been weakly related to absence duration ($r = 0.09$) and absence frequency ($r = -0.18$; Scott and Taylor, 1985). The restricted relationship between absenteeism and satisfaction may be due to the employee's ability to express such behaviors (i.e., strict absence policies).

In the 1980s, interest shifted to behaviors that were outside of the formal job requirements, such as OCB (see Bateman and Organ, 1983; Smith et al., 1983). OCBs, broadly defined as organizationally useful behaviors that are not formally prescribed by formal job requirements, appear to be moderately correlated with satisfaction ($r > 0.16$; Dalal, 2005; LePine et al., 2002; Organ and Ryan, 1995). CWBs, or behaviors intended to harm coworkers or the organization as a whole, can be considered the opposite of OCBs. Mount et al. (2006) found significant path coefficients between satisfaction and organizational CWBs ($r = 0.41$) and interpersonal CWBs ($r = 0.40$), and Dalal (2005) found a moderate relationship between satisfaction and overall CWBs ($r = -0.37$, corrected). In comparison to in-role behaviors, more substantive effect sizes for OCBs and CWBs may be due to the greater level of control that the employees have over these behaviors. OCBs and CWBs are not strongly regulated behaviors, and thus allow employees more latitude in expression.

Harrison et al. (2006) have argued that this behavior-by-behavior approach to satisfaction consequences is inconsistent with basic attitude theory. Working from the compatibility principle of attitude theory, they argue that overall job satisfaction, as a general attitude, should be a better predictor of an aggregated index of employee contributions than of any single behavior. In a very creative *meta*-analysis, where they combine results for both withdrawal and extra-role behaviors, they conclude that the expected relationship of overall satisfaction with an integrated index of contribution approached 0.60. Their work is important because it shows a utility for job satisfaction that has been heretofore hard to find and because it demonstrates the usefulness of incorporating attitude theory into the study of job satisfaction.

Finally, interest in satisfaction is not driven entirely by organizational considerations as there is reason to believe that job dissatisfaction has negative effects on mental health. For example, a *meta*-analysis by Lee and Ashforth (1996) found significant relationship with satisfaction and Maslach's (1982) three dimensions of burnout: emotional exhaustion ($r = -0.31$, corrected), depersonalization ($r = -0.44$, corrected), and personal accomplishment ($r = 0.29$, corrected). Further, a *meta*-analysis by Faragher et al. (2005) found that satisfaction is related to anxiety ($r = 0.42$), depression ($r = 0.42$), and self-esteem ($r = 0.43$), suggesting that satisfaction at work can be meaningfully related to employee's mental health.

Predictors of Job Satisfaction

Organizational research has had a long tradition of examining the predictive correlates of job satisfaction, with interest in predictors across populations and for specific populations (e.g., nurses, firefighters, teachers, etc.). The result is a literature

of thousands of studies demonstrating that this work feature or that personal characteristic correlates with job satisfaction.

Judge and Kammeyer-Mueller (2012) found it useful to organize the most studied work environment predictors of job satisfaction into four broad categories: the nature of the work itself, the support from coworkers, leader quality, and organizational practices. Evidence exists for the importance of experiences in each of these categories.

Of primary importance among work-related factors is the nature of the work itself. Since the introduction of the Job Characteristics Model (Hackman and Oldham, 1976) of satisfaction and motivation, empirical studies consistently show that when people work on jobs, they report as being more complex, requiring more skills, providing more intrinsic feedback, and so on, they also report higher levels of job satisfaction. In addition, interventions that change jobs to enhance these characteristics generally enhance reports of satisfaction as well. Related to the structure of work, but not the specific task per se, is research on role conflict and role ambiguity. Both role conflict, the extent of incompatible job demands, and role ambiguity, uncertainty concerning job responsibilities, are consistently correlated with reports of job dissatisfaction. Importantly, conflict between work and nonwork roles is also a consistent predictor of dissatisfaction. Kossek and Ozeki (1998) found a negative relationship between all types of work–family conflict and satisfaction ($r = -0.31$), suggesting that problematic work and nonwork relationships are detrimental to satisfaction.

Social interactions in the workplace, both with coworkers and with supervisors, are well-known predictors of job satisfaction. Not surprisingly workers report that the degree of support they get from their coworkers is an important influence on the extent to which they like their jobs as is the degree to which their supervisors show concern, consideration, and respect for them (Judge and Kammeyer-Mueller, 2012).

Organizational practices and policies also influence satisfaction. For example, Baltes and colleagues investigated the differential effect of nontraditional work schedules on satisfaction (Baltes et al., 1999). Flextime schedules allow individuals to decide when to arrive and leave from work. Compressed workweeks condense the workweek to less than 5 days by increasing the number of hours per day. A compressed schedule was more strongly related to satisfaction ($r = 0.28$) than a flextime schedule ($r = 0.07$). Much research has shown that being treated in a just and fair manner correlates with reported job satisfaction. Interestingly enough, while most people would think that pay levels are the single most important influence on satisfaction, *meta*-analytic results suggest that pay may only be weakly related to job satisfaction (Judge et al., 2010).

Personal antecedents, such as personality, also predict satisfaction. Bruk-Lee et al. (2009) found that job satisfaction was positively but weakly related to the Big Five personality traits (Goldberg, 1990) of conscientiousness, extraversion, and agreeableness ($r \geq 0.12$), negatively related to neuroticism ($r = -0.25$), and not related to openness to experience. Satisfaction was also related to trait positive affectivity ($r = 0.41$) and negative affectivity ($r = -0.25$) (Bruk-Lee et al., 2009). Another *meta*-analysis (Connolly and Viswesvaran, 2000) found similar relationships between positive and negative affectivity and satisfaction.

The ability to review the thousands of studies on the predictors of job satisfaction is clearly beyond the scope of this article, but there is no compelling necessity to do so. The purely empirical database, the observation that this factor or that variable predicts satisfaction, might have some practical value in a particular context but it provides little organizing explanatory insight into the nature of work attitudes or their consequences. This conclusion has not gone unnoticed by job satisfaction researchers, and as a result a rich history of theoretical activity has developed side by side with the purely empirical database.

Theories of Job Satisfaction

Two general points can be made about the broad array of theoretical positions developed regarding job satisfaction. First, although job satisfaction is an attitude held about one's job (or the aspects of one's job when referring to facet satisfactions), very few of the major theoretical positions are informed by basic attitude theory (but see work by Schleicher et al., 2011; Harrison et al., 2006; as a notable exceptions). Second, most theoretical positions focus on developing a causal structure for satisfaction, instead of trying to place satisfaction within a broad nomological network of causes, consequences, moderators, and so on. Interestingly, all of the theoretical approaches have received some empirical support, partly because they are mostly complementary rather than antagonistic, partly because they each touch on some element of the truth, and partly because they are infrequently stated or tested in ways conducive to disconfirmation.

Theories of job satisfaction can be placed within four broad categories: cognitive judgment theories, social influence theories, dispositional theories, and affect theories. Historically, cognitive judgment theories are the oldest positions and affect theories are found more recently. These theoretical positions are best understood as presenting differences of emphases, not incompatibility. For example, affect theories like AET generally include a role for personality.

Cognitive judgment approaches dominated the theoretical landscape in the 1960s; and although they are more rarely researched today, they still form an implicit framework for many discussions of satisfaction. Locke's (1976) value-based theory is perhaps the most influential, but theories with a similar structure were offered by Porter (1961) and Katzell (1964), among others. Although each of the theories has unique elements, a general framework can be described. In this framework, the work environment is perceptually organized as a set of discrete concrete or abstract features (e.g., coworkers, pay, supervisors, career opportunities, etc.). Workers have a corresponding set of standards that they use to evaluate the features directly or the outcomes derived from the features (i.e., the pay I receive compared to the pay I desire or expect). Finally, perceptions are matched to standards to yield facet satisfactions. Overall satisfaction is derived from the combination of facet satisfactions.

More recently, a variant of the discrepancy approach can be found in fit theories of satisfaction. Fit can be defined as "the compatibility between an individual and a work environment that occurs when their characteristics are well matched"

(Kristof-Brown et al., 2005). In a *meta*-analysis of research on fit, Kristof-Brown et al. (2005) found support for the contention that the degree to which job and organizational demands match personal skills, need, and values predicts job satisfaction. Fit theories have a 'matching' structure similar to those found in cognitive judgment theories, although judgment theories postulate specific cognitive processes while fit generally does not.

Social influence theories, most notably represented by social information processing theory (Salancik and Pfeffer, 1977, 1978), give emphases to the role that the social environment has in influencing judgments about work. That is, they postulate that attitudes of coworkers can influence perceptions and evaluations of particular features of one's job as well as overall evaluations of the job as a whole. In this sense, social influence theories are not alternatives to cognitive judgment theories but rather help understand how the perceptions that are critical to these judgments are influenced by the social environment. Not surprisingly, research has consistently supported these social influences on work perceptions and evaluations.

Dispositional approaches focus on personality characteristics as influences on satisfaction. As described by Weiss and Cropanzano (1996), the basic idea is that "to some degree, a person's job satisfaction reflects his or her general tendency to feel good or bad about all aspects of life and this general tendency is independent of the specific nature of the job" (p. 7). Evidence for the position is found in tendencies for satisfaction to show some stability across time and jobs and for satisfaction to correlate with established personality constructs. The most notable theoretical position in this area is core self-evaluations (CSE) theory (Judge et al., 1997). CSE represent fundamental evaluations of the self and include personality characteristics of self-esteem, self-efficacy, locus of control, and neuroticism. Research supports the relationship between the CSE composite and job satisfaction.

In the 1990s, the popularity of research on job satisfaction gave way to a new interest in affective states like moods and emotions. However, the new-found interest in true affect did not lead to the disappearance of thinking about job satisfaction. Instead, it helped make that thinking more precise. An important theoretical position in this domain is AET (Weiss and Cropanzano, 1996) that attempts to clarify key concepts in satisfaction and affect research. In particular, it explicates differences between satisfaction, as a relatively stable attitudinal evaluation of one's job, and moods and emotions as transient affective experiences (see also Weiss, 2002), and it discusses how discrete affective experiences influence overall judgments like satisfaction. It also draws a distinction between affect-driven behaviors (criteria driven by immediate affective states like moods and emotions) and judgment-driven behaviors (criteria driven by evaluative judgments like satisfaction). Now, these distinctions between affect and attitude are well accepted and their interrelationships represent a productive area of research.

Summary

Jobs satisfaction, the overall evaluative judgment one has about one's job, has a long and rich history in organizational research and practice. In some ways it is seen as the enduring 'person side' of such research, in contrast to performance interest of much organizational research. Even this characteristic is flawed because, historically, much satisfaction research was driven by supposed performance implications. Nevertheless, over the long history of work psychology, satisfaction has generally carried the load for research on work experience. Today work experienced is captured by a plethora of related constructs like emotions, moods, engagement, commitment, and so on. Still, as part of this umbrella of work experience concepts, satisfaction remains an important way to describe worker–employer relationships of interest to researchers and management.

See also: Happiness and Work; Industrial–Organizational Psychology: Science and Practice; Job Analysis and Work Roles, Psychology of; Organizational Citizenship Behavior; Work Motivation; Workplace Stress.

Bibliography

Adams, J.S., 1963. Toward and understanding of inequity. Journal of Abnormal Psychology 67, 422–436.

Adams, J.S., 1965. Inequity in social exchange. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 2. Academic Press, New York, pp. 267–299.

Baltes, B.B., Briggs, T.E., Huff, J.W., Wright, J.A., Neuman, G.A., 1999. Flexible and compressed workweek schedules: a meta-analysis of their effects on work-related criteria. Journal of Applied Psychology 84 (4), 496–513.

Bateman, T.S., Organ, D.W., 1983. Job satisfaction and the good soldier: the relationship between affect and employee "citizenship". Academy of Management Journal 26 (4), 587–595.

Bowling, N.A., 2007. Is the job satisfaction–job performance relationship spurious? A meta-analytic examination. Journal of Vocational Behavior 71 (2), 167–185. http://dx.doi.org/10.1016/j.jvb.2007.04.007.

Brayfield, A.H., Crockett, W.H., 1955. Employee attitudes and employee performance. Psycholigical Bulletin 52, 396–424.

Bruk-Lee, V., Khoury, H.A., Nixon, A.E., Goh, A., Spector, P.E., 2009. Replicating and extending past personality/job satisfaction meta-analyses. Human Performance 22 (2), 156–189. http://dx.doi.org/10.1080/08959280902743709.

Carsten, J.M., Spector, P.E., 1987. Unemployment, job satisfaction, and employee turnover: a meta-analytic test of the Muchinsky model. Journal of Applied Psychology 72 (3), 374–381.

Connolly, J.J., Viswesvaran, C., 2000. The role of affectivity in job satisfaction: a meta-analysis. Personality and Individual Differences 29, 265–281.

Cotton, J.L., Tuttle, J.M., 1986. A meta-analysis and review with implications for research. The Academy of Management Review 11 (1), 55–70.

Dalal, R.S., 2005. A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. Journal of Applied Psychology 90 (6), 1241–1255. http://dx.doi.org/10.1037/0021-9010.90.6.1241.

Faragher, E.B., Cass, M., Cooper, C.L., 2005. The relationship between job satisfaction and health: a meta-analysis. Occupational and Environmental Medicine 62 (2), 105–112. http://dx.doi.org/10.1136/oem.2002.006734.

Fisher, V.E., Hanna, J.V., 1931. The Dissatisfied Work. Macmillan, New York.

Fried, Y., Shirom, A., Gilboa, S., Cooper, C.L., 2008. The mediating effects of job satisfaction and propensity to leave on role stress-job performance relationships: combining meta-analysis and structural equation modeling. International Journal of Stress Management 15 (4), 305–328. http://dx.doi.org/10.1037/a0013932.

Goldberg, L.R., 1990. An alternative "description of personality": the big-five factor structure. Journal of Personality and Social Psychology 59 (6), 1216–1229.

Griffeth, R.W., Hom, P.W., Gaertner, S., 2000. A meta-analysis of antecedents and correlates of employee turnover: update, moderator tests, and research implications for the next millennium. Journal of Management 26 (3), 463–488.

Hackman, J.R., Oldham, G.R., 1975. Development of the job diagnostic survey. Journal of Applied Psychology 60, 159–170.

Hackman, J.R., Oldham, G.R., 1976. Motivation through the design of work: test of a theory. Organizational Behavior and Human Performance 16, 250–279.

Harrison, D.A., Newman, D.A., Roth, P.L., 2006. How important are job attitudes? Meta-analytic comparisons of integrative behavioral outcomes and time sequences. The Academy of Management Journal 49 (2), 305–325.

Harter, J.K., Schmidt, F.L., Hayes, T.L., 2002. Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: a meta-analysis. Journal of Applied Psychology 87 (2), 268–279.

Hersey, R.B., 1932. Workers' Emotions in Shop and Home: A Study of Individual Workers from the Psychological and Physiological Standpoint. University of Pennsylvania Press, Philadelphia.

Herzberg, F., Mausner, B., Snyderman, B., 1959. The Motivation to Work. Wiley, New York.

Hom, P.W., Caranikas-Walker, F., Prussia, G.E., 1992. A meta-analytical structural equations analysis of a model of employee turnover. Journal of Applied Psychology 77 (6), 890–909.

Hoppock, R., 1935. Job Satisfaction. Harper, New York.

Iaffaldano, M.T., Muchinsky, P.M., 1985. Job satisfaction and job performance: a meta-analysis. Psycholigical Bulletine 97 (2), 251–273.

Ironson, G.H., Smith, P.C., Brannick, M.T., Gibson, W.M., Paul, K.B., 1989. Construction of a job in general scale: a comparison of global, composite, and specific measures. Journal of Applied Psychology 74 (2), 193–200.

Judge, T.A., Locke, E.A., Durham, C.C., 1997. The dispositional causes of job satisfaction: a core evaluations approach. Research in Organizational Behavior 19, 151–188.

Judge, T.A., Locke, E.A., Durham, C.C., 1997. The dispositional causes of job satisfaction: a core evaluations approach. In: Staw, L.L.C.B.M. (Ed.), Research in Organizational Behavior. JAI, Greenwich, CT.

Judge, T.A., Piccolo, R.F., Podsakoff, N.P., Shaw, J.C., Rich, B.L., 2010. The relationship between pay and job satisfaction: a meta-analysis of the literature. Journal of Vocational Behavior 77 (2), 157–167. http://dx.doi.org/10.1016/j.jvb.2010.04.002.

Judge, T.A., Thoresen, C.J., Bono, J.E., Patton, G.K., 2001. The job satisfaction - job performance relationship: a qualitative and quantitative review. Psychological Bulletin 127 (3), 376–407.

Katzell, R.A., 1964. Personal values, job satisfaction and job behavior. In: Borow, H. (Ed.), Man in a World of Work. Houghton-Mifflin, Boston, pp. 341–366.

Kornhauser, A.W., Sharp, A.A., 1932. Employee attitudes; suggestions from a study in a factory. Personnel Journal 10, 393–404.

Kossek, E.E., Ozeki, C., 1998. Work-family conflict, policies, and the job-life satisfaction relationship: a review and directions for organizational behavior-human resources research. Journal of Applied Psychology 83 (2), 139–149.

Kristof-Brown, A.L., Zimmerman, R.D., Johnson, E.C., 2005. Consequences of individuals' fit at work: a meta-analysis of person-job, person-organization, person-group, and person-supervisor fit. Personnel Psychology 58, 281–342.

Kunin, T., 1955. The construction of a new type of attitude measure. Personnel Psychology 8, 65–78.

Lawler, E.E., 1973. Motivation in Work Organizations. Brooks/Cole Publishing Company, Monterey, CA.

Lee, R.T., Ashforth, B.E., 1996. A meta-analytic examination of the correlates of the three dimensions of job burnout. Journal of Applied Psychology 81 (2), 123–133.

LePine, J.A., Erez, A., Johnson, D.E., 2002. The nature and dimensionality of organizational citizenship behavior: a critical review and meta-analysis. Journal of Applied Psychology 87 (1), 52–65. http://dx.doi.org/10.1037//0021-9010.87.1.52.

Locke, E.A., 1969. What is job satisfaction? Organizational Behavior and Human Decision Processes 4 (4), 309–336.

Locke, E.A., 1976. The nature and causes of job satisfaction. In: Dunette, M.D. (Ed.), Handbook of Industrial and Organizational Psychology. Rand McNally, Chicago, pp. 1297–1349.

Lofquist, L.H., Dawis, R.V., 1969. Adjustment to Work. Appleton-Century-Crofts, New York.

Maslach, C., 1982. Burnout: The Cost of Caring. Prentice Hall, Englewood Cliffs, NJ.

Mount, M., Ilies, R., Johnson, E., 2006. Relationship of personality traits and counterproductive work behaviors: the mediating effects of job satisfaction. Personnel Psychology 59, 591–622.

Organ, D.W., Ryan, K., 1995. A meta-analytic review of attitudinal and dispositional predictors of organizational citizenship behavior. Personnel Psychology 48, 775–802.

Podsakoff, N.P., Whiting, S.W., Podsakoff, P.M., Blume, B.D., 2009. Individual- and organizational-level consequences of organizational citizenship behaviors: a meta-analysis. Journal of Applied Psychology 94 (1), 122–141. http://dx.doi.org/10.1037/a0013079.

Porter, L.W., 1961. A study of perceived need satisfactions in bottom and middle management jobs. Journal of Applied Psychology 45, 1–10.

Salancik, G.R., Pfeffer, J., 1977. An examination of need-satisfaction models of job attitudes. Administrative Science Quarterly 22, 427–456.

Salancik, G.R., Pfeffer, J., 1978. A social information processing approach to job attitudes and task design. Administrative Science Quarterly 23, 224–253.

Schleicher, D.J., Hansen, S., Fox, K.E., 2011. Job Attitudes and Work Values, vol. 3. American Psychological Association, Washington, DC.

Scott, K.D., Taylor, G.S., 1985. An examination of conflicting findings on the relationship between job satisfaction and absenteeism: a meta-analysis. Academy of Management Journal 28 (3), 599–612.

Smith, C.A., Organ, D.W., Near, J.P., 1983. Organizational citizenship behavior: its nature and antecedents. Journal of Applied Psychology 68 (4), 653–663.

Smith, P.C., Kendall, L.M., Hulin, C.L., 1969. The Measurement of Satisfaction in Work and Retirement: A Strategy for the Study of Attitudes. Rand McNally, Chicago.

Tett, R.P., Meyer, J.P., 1993. Job satisfaction, organizational commitment, turnover intention, and turnover: path analyses based on meta-analytic findings. Personnel Psychology 46, 259–293.

Vroom, V.H., 1964. Work and Motivation. Wiley, New York.

Weiss, H.M., 2002. Deconstructing job satisfaction: separating evaluations, beliefs, and affective experiences. Human Resource Management Review 12, 173–194.

Weiss, H.M., Brief, A.P., 2001. Affect at work: a historical perspective. In: Payne, R.L., Cooper, C.L. (Eds.), Emotions at Work: Theory, Research and Applications for Management, Wiley, New York, pp. 133–171.

Weiss, H.M., Cropanzano, R., 1996. Affective events theory: a theoretical discussion of the structure, causes and consequences of affective experiences at work. Research in Organizational Behavior 18, 1–74.

Joint Action

Margaret Gilbert, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

For the purposes of this article, a joint action is the action of two or more people. Thus, two or more people may hold a conversation, go hiking together, and so on. Joint actions are central to everyday life, but it is not easy to say precisely what they amount to. A variety of accounts have been envisaged by theorists. This article reviews some of the best-known accounts, with special reference to their capacity to explain the nature and source of the entitlements and obligations associated with joint actions. Some of these proposals see the core of any joint action as an array of personal intentions in favor of the joint action, where personal intentions are expressible by the individuals they characterize in sentences of the form 'I intend….' Others have argued that something else lies at the core of any joint action. Thus, John Searle contends that each participant must manifest what he refers to as "collective intentionality," which is a matter of an individual 'we'-intending, as opposed to his or her having a personal intention. Margaret Gilbert argues rather that joint actions are grounded in an appropriate "joint commitment" of the parties. An appeal to joint commitment provides an explanation of the entitlements and obligations of joint action. Some, including Michael Bratman, appeal rather to a complex moral principle to provide an explanation of these entitlements and obligations.

Some Initial Considerations on Joint Action

Joint Action Need Be Cooperative Only in a Broad Sense

Joint action is not necessarily a matter of amicable cooperation. We may be quarreling with one another or engaged in a fight. Quarreling and fighting are species of joint action. Though it need not be amicable, there is clearly a broad sense in which joint action is always cooperative or collaborative. I cannot quarrel with you unless you quarrel with me; I cannot fight with you unless you fight with me. If you refuse to take me seriously we cannot quarrel. If you lay down your arms, we are no longer fighting.

Obligations and Entitlements in the Context of Joint Action

Informal observation of the way people behave when doing things together makes it clear that the parties often rebuke one another for not acting accordingly. Such rebukes are commonly understood to be in order. For example, Jill and John are going for a walk, John begins to lag behind, and Jill rebukes John for doing so. John understands that, in the context of their walk, Jill has the standing to rebuke him.

Less punitive responses are also commonplace, and routinely understood to be in order. For example, Jill asks John why he is lagging behind. John may not like being questioned like this, but he does not dispute that Jill is in a position to question him. Or, Jill tells John to hurry up, in order to spur him into appropriate action. John understands that she is in a position to do so. He does not say 'Who are you to tell me what to do?'

Underlying these diverse responses is the understanding that participation in a joint action somehow entitles each of the parties to relevant actions from the other or others. In related terms, each party is obligated to the others to act accordingly. In other terms again, he or she owes them appropriate actions. If one party does not produce the relevant actions, therefore, the others have the standing to question that party, and so on, as they see fit.

A further point is this. Suppose John and Jill have set out together on a 5-mile hike. At some point along the way, John gets tired. He announces to Jill that he is turning back. If he does this, then, failing some special background understandings, something is amiss. It would be different if he asked for Jill's concurrence with his turning back. In general terms, absent special background understandings, no one party to a joint action has the power to determine that the joint action is at an end, or to change its established course. If one party acts as if he or she has that power, something is amiss.

Inferences from Statements about What We Are Doing

From 'We are on a 5-mile hike' and 'In order for us to hike 5 miles, I must hike 5 miles,' it seems that John can infer that he has reason to hike 5 miles. Notably, it is not obvious that there need be any interposition in this reasoning of a premise about John himself. For instance, it is not necessary for John to add that he wants to hike for 5 miles with Jill.

Two Central Questions

A satisfying account of joint action will incorporate an explanation of the observations just given. Two related questions deserve emphasis.

What States of the Participants Are Involved in Joint Action?

When two people are involved in a joint action, what must be true of each of them? Must they, for instance, personally intend to act as is necessary for completion of the joint action? This may seem to be an obvious assumption, and many theorists have assumed something like it. Often they put things initially in terms of an underlying joint intention, then parse this in terms of personal intentions of the parties. As we shall see, serious problems attend this approach.

Of What Kind Are the Obligations and Entitlements Associated with Joint Intention?

The parties to a joint action understand their participation to entitle each of them to actions in conformity with the relevant joint intention. What kind of entitlement is at issue? One might assume that it is moral entitlement founded on moral principle. On the other hand, it may seem that joint action is too basic and primitive a phenomenon to require moral understanding in its participants. It is certainly not obvious that it does require such understanding. Indeed, it seems possible that those with little or no moral perspicacity are both capable of joint action and of understanding that they have the entitlements in question when engaged in such action.

A related point is that those who engage in morally wrongful joint actions with the understanding that their actions are morally wrong may take themselves to have the standings discussed earlier. Thus, one bank robber may rebuke another for holding back at some point, and be understood by his partner to have the standing to do so, though morally speaking his holding back may be the right thing for him to do, as he knows.

Accounts of Joint Action in terms of Meshing Personal Intentions

Michael Bratman sees the core of a joint action as a joint (or, in his terms, 'shared') intention which, he argues, is constituted by personal intentions of a particular kind (Bratman, 1999 contains some central articles). A somewhat simplified version of Bratman's account of joint intention runs as follows. We jointly intend to go for a walk if and only if (a) (i) I intend that we go for a walk, (ii) you intend that we go for a walk; (b) I intend that we go for a walk in accordance with and because of (a) (i) and (ii), and you intend likewise; and (c) (a) and (b) are common knowledge between us.

As to the 'common knowledge' in condition (c), what is intended is, roughly, that (in the example) the fact that conditions (a) and (b) are fulfilled is completely out in the open as far as the two people involved are concerned. (For the first published discussion of 'common knowledge,' see Lewis, 1969.)

There are several problems with an account of joint action in terms of joint intention on Bratman's account. Two will be detailed here. Analogous problems arise for accounts of joint action that are broadly in the spirit of Bratman's proposals.

One problem has to do with the entitlements associated with joint action as such. It does not appear that anyone is entitled to relevant actions from anyone else just by virtue of the satisfaction of the conditions laid down in Bratman's account of joint intention. Thus my intending that we do something, and intending that this occur by virtue of my so intending and you so intending, does not seem to entitle you to any particular action from me, or vice versa. Nor does the fact that my various intentions are common knowledge between us seem sufficient to bring any such entitlement into existence. The fact that you have similar intentions, as is common knowledge between us, does not seem to change things as far as anyone's entitlements go.

If you know that I intend that we go for a walk, and have reason to think that I will not change my mind, then there is a sense in which you are 'entitled to expect' that I am going to make the appropriate contribution to our walking together. But the quoted phrase just means that you have a good basis for predicting that I am going to act thus. It does not mean that you are entitled to my acting thus in such a way that rebuking me, for instance, is in order.

In related discussion, Bratman has invoked a complex moral principle to argue for the presence of the relevant entitlements in many situations in which his conditions on joint intention hold. The principle in question is the 'principle of fidelity' proposed in Scanlon (1990) and elsewhere to explain the obligations of those who make promises. This principle runs as follows (Scanlon, 1990: p. 208): If (a) A voluntarily and intentionally leads B to expect that A will do x (unless B consents to A not doing x); (b) A knows that B wants to be assured of this; (c) A acts with the aim of providing this assurance, and has good reason to believe that he or she has done so; (d) B knows that A has the beliefs and intentions just described; and (e) A intends for B to know this, and knows that B does know it; and B knows that A has this knowledge and intent; then, in the absence of some special justification, A must do x unless B consents to his not doing so.

As Bratman allows, Scanlon's principle of fidelity is not applicable in all situations where there are joint intentions on his account of them. The principle is applicable to the parties, Bratman assumes, if each makes a prior promise to the others; but this is not always the case when there is joint intention and action. Given that joint action by its nature entitles the participants to conforming actions, then appeal to Scanlon's principle will not rescue an account of joint action in terms of joint intention in Bratman's sense.

Another problem with an appeal to Scanlon's principle in this context is that even when it does apply, it is by no means clear that the parties owe one another actions appropriate to the joint action in question, in the sense of 'owe' that is in question here. See Gilbert (2004). This concern also applies to Alonso's appeal to a moral principle of reliance in Alonso (2009).

A further problem with an account of joint action in terms of joint intention in Bratman's sense is that it does not explain the need for the concurrence of all parties in order to terminate the joint action. One with a personal intention is in a position to change it unilaterally. It follows that an individual who is party to a joint intention in Bratman's sense is in a position unilaterally to do away with one of the foundational personal intentions at any time. There is no need for the permission of the other party or parties.

Appeals to a Set of Personal 'We-Intentions'

John Searle has claimed that at the core of joint action lies irreducible states of mind of each participant expressible by utterances of the form 'We intend to do such-and-such' (see, for instance, Searle, 1995; also Sellars, 1963). These 'we-intentions' are irreducible in the sense that they cannot be reduced or broken down into thoughts about what 'I intend' and what 'you intend.' Searle's central concern is to describe the inner or subjective side of joint action with respect to each participant. He does not focus on the question: when is a given 'we-intention' appropriate? What he writes at one point

suggests that, in his view, one person's 'we-intention' is correct as long as the other person or persons in question have corresponding 'we-intentions' of their own. This particular proposal does not seem to be right, if 'We intend to do such-and-such' is understood in accordance with everyday understandings about joint intention.

Suppose that Ben is currently thinking, with respect to himself and Elaine, 'We intend to get married.' Elaine is in a similar position. According to the proposal under discussion, it follows that each one is correct in thinking 'We intend to get married.' Yet according to everyday understandings it would not be plausible to claim that, solely by virtue of the facts described, Ben and Elaine jointly intend to get married. If one learned that they had never discussed the matter, for instance, one would judge that their thoughts were not properly grounded.

In addition to such concerns, it is not clear that those with corresponding we-intentions owe each other corresponding actions. Nor is it clear that the concurrence of all parties is necessary in order to terminate a joint action, on this account.

A Role for Joint Commitment

Joint Commitment

Margaret Gilbert has proposed that the core constituent of a joint action is a pertinent joint commitment (see Gilbert, 2000). The nature of joint commitment as Gilbert understands this will now be explained.

To be committed to do A in the relevant sense is to have sufficient reason to do A. This means that, all else being equal, one will act in error should one not do A. According to Gilbert, if Jane decides to go shopping this afternoon, she is thereby subject to a commitment to go shopping then. She is not subject to such a commitment if she is merely inclined to go shopping or hopes to do so.

Jane's decision creates what Gilbert calls a commitment 'of the will.' It is a 'personal' commitment of the will, the will in question in this case being that of Jane herself. A personal commitment of the will is the creation of the person in question, who is in a position unilaterally to rescind it.

According to Gilbert, there are both personal and joint commitments of the will. A joint commitment of the will is not a conjunction of corresponding personal commitments of each party. Rather, it is a commitment of all the parties by all the parties. A joint commitment requires the participation of each party in its creation in one way or another.

Without prior authorization, no one can unilaterally rescind or dissolve a joint commitment: all must rescind it together. Otherwise it stands as a constraint on the parties. A joint commitment can be violated by one party alone, but this is precisely a matter of violation as opposed to dissolution. Unlike those who rescind or dissolve it, the violator acts against a standing joint commitment.

There are both basic and derived forms of joint commitment. In order to create a joint commitment of the basic type, each party must express to the rest his or her readiness to create with them a particular joint commitment. When this has been done in conditions of common knowledge, the relevant joint commitment is in place. Derived forms of joint commitment will be discussed shortly.

Gilbert has argued that a pertinent joint commitment helps constitute not only joint goals and intentions but also joint or collective beliefs, social rules, and many other central social phenomena (see Gilbert, 1989, 1996, 2000). With respect to the constitution of joint action, she initially invoked a joint goal, and then a joint intention. More recently she has again invoked a joint goal (see Gilbert, 2006: Chapters 6 and 7). The choice in question is not crucial for present purposes. It will be assumed in what follows that a joint intention is at issue.

On Gilbert's account it is necessary and sufficient for two or more people jointly to intend to do such-and-such that they are jointly committed to intend as a body to do such-and-such. Spelling this out: they are jointly committed together to emulate, in relevant ways, one who intends to do such-and-such, by virtue of the actions of each. Actions in this context include utterances. In Gilbert's technical terminology a joint commitment to intend as a body to do such-and-such is necessary and sufficient, by definition, to create a "plural subject" of the intention in question. Thus we intend to do such-and-such, if and only if we constitute a plural subject of intention.

How do people express to one another their readiness to be jointly committed to intend as a body to do something? There are many such ways. Among these are such humdrum verbal exchanges as 'Shall we go for a walk this afternoon?' with a following 'Sure, let's.'

Joint commitments are not necessarily set up through verbal exchanges. Two men who jump into a rowing boat in hot pursuit of a fugitive may signal by their gestures and utterances that, in effect, each is ready to be jointly committed with the other to intend as a body to row to where the fugitive is hiding. As each moves the oars in light of their ensuing commitment, they engage in the joint action of rowing the boat.

In principle, it may take some time for a particular joint commitment to be established. Each party's readiness to be party to the joint commitment may be made clear only gradually and over the course of many interactions with the other people in question.

Derived Joint Commitments

The foregoing examples of the formation of a joint commitment involve joint commitments of the basic type. Here is an example involving a derived joint commitment. Don asks his wife Sally 'Where are we going on vacation this summer?' He presumes that though he is unaware of any vacation plans, there may already be one, and Sally will know what it is. This presumption is based on the existence of a background joint commitment of the basic type. Don and Sally are jointly committed to accept as a body that Sally may make vacation plans for the two of them. With this background commitment in place, Sally is in a position unilaterally to create joint vacation plans for the two of them. Thus, by virtue of the background joint commitment and Sally's unilateral decision, Don and Sally may take on a derived joint commitment such as the joint commitment to intend as a body to go to Spain this summer. Another example: Don has been authorized by a basic joint commitment of the two to decide on the end point of his and Sally's hike. Sally asks him: "Where are we heading?" She asks, in effect, for the content of a derived joint commitment

whose content is so far unknown to her, a joint commitment that may well have been established already.

Large Populations

Members of large populations, such as the many inhabitants of a large island, are unlikely to know one another personally, or even to know of one another as individuals. If there are population-wide joint commitments in such populations, this will be through expressions of readiness to be jointly committed with all others with a certain common feature, such as 'those living on this island.'

By virtue of background joint commitments of the kind discussed in the Section Derived Joint Commitments, the members of a large population can have a particular joint intention in Gilbert's sense, though many of the members do not know of the intention in question. This accords with everyday understandings according to which it may be said that, for instance, a particular country intends a particular action, without it being assumed that everyone in that country knows of the intention, or indeed of the ensuing action.

Those who form intentions for the whole population may be authorized to ensure that these intentions are implemented. Their implementation need not involve all members of the population, in order that the action in question be considered an action of the population as a whole.

Obligations and Entitlements of Joint Commitment

Gilbert has argued at some length that those who are jointly committed in some manner are obligated to one another to act accordingly, and correspondingly entitled to such action from the others (see Gilbert, 2000; also Gilbert, 2006: Chapter 7). A brief argument for her position might note that if one violates a commitment that is joint, it is evident that one is answerable to the other parties to the commitment, who can observe that one has violated 'our' commitment as opposed to a commitment of one's own. They have a clear basis for questioning the violator and for administering rebukes on their own behalf. This implies that the parties to a joint commitment are entitled to conformity from the other parties. These others have corresponding obligations to conform: they are obligated to all parties to the commitment, their obligation being to conform to the commitment.

This and related arguments are arguments from the nature and structure of joint commitment. It may thus be best not to think of the relevant entitlements and obligations as moral entitlements and obligations. All that one needs to understand, in order to understand that they exist, is the nature of a joint commitment. One need not appeal to the value of anything such as, in the case of Scanlon's principle of fidelity, the value of what Scanlon refers to as 'assurance.'

Gilbert's account of joint action offers an explanation of how those who act together have associated entitlements to conformity from one another. These entitlements come from the constitutive joint commitment. There are associated entitlements to ask questions, rebuke one another, and so on, should deviation from the constitutive joint intention be mooted. Gilbert can allow that in some situations where there are joint intentions it is morally permissible not to act in accordance with them. Such situations may include those where one party coerced another into participating in the relevant joint commitment, and those where the joint commitment is to intend something immoral. All things considered, then, it may be morally permissible not to fulfill the relevant obligations of joint commitment. This does not entail that those obligations no longer exist. They exist as long as the sustaining joint commitment exists (see Gilbert, 1996).

Explaining Inferences from 'We Are Doing…'

The plural subject account of joint action provides an explanation of how from 'We are rowing to the other side' and 'My rowing faster will best facilitate our rowing to the other side,' I can infer 'I should row faster' without the interposition of some premise about myself, such as a premise stating that I personally want to row to the other side. My being party to the relevant joint commitment gives me reason to row faster if this will best facilitate our rowing to the other side. I need have no personal desire, inclination, or attitude in favor of our doing this.

Joint Action and Related Phenomena

Gilbert has argued that those who participate in joint actions on her account constitute a social group in a central sense of the term, though the group may be small and transient (Gilbert, 1989). Indeed, she argues, any plural subject in her sense constitutes a social group. List and Pettit (2011) discuss constraints that are necessary, in their view, for a persisting group to constitute a genuine agent.

On Gilbert's joint commitment account of social groups, it can be argued that obligations of a special kind are part and parcel of group membership. This makes it possible to bring Gilbert's plural subject theory of joint action and related phenomena to bear on a classic question of political philosophy, the question of political obligation. Is there an obligation to obey the laws of one's own country in particular, and if so, what is the source of this obligation? Evidently, one possible source is participation in a joint commitment (see Gilbert, 2006).

The topic of joint action is relevant to the topic of collective moral responsibility (see the essays in French and Wettstein, 2006). In this connection Gilbert has argued that, given her joint commitment account, it may sometimes be reasonable to ascribe moral responsibility for a joint act to the group whose act it is. The moral responsibility of particular group members is then a separate matter. This perspective on joint action allows for the intelligibility of a common phenomenon: the experience of pride or guilt over the actions of one's group, as opposed to one's own actions (see Gilbert, 1996).

The Nature and Role of Agreements

Another theorist who has written extensively on joint intention and action is Raimo Tuomela. In Tuomela (1984), he put what he called 'we-intentions' at the center of an account of joint action. Each party to a joint action must have a relevant we-intention. Tuomela's complex definition of a we-intention runs roughly as follows: a member Ai of a collective G

we-intends to do X, where X is some act of G, if and only if (a) Ai intends to do his own part of X, given that every member of G will do his own part of X; (b) Ai believes that every member of G will do his own part of X; and (c) there is a mutual belief in G to the effect that (a) and (b).

Tuomela (1995) introduces the "underlying assumption that we-intentions are formed in a process involving the making of an explicit or implicit agreement (or at least believed agreement)" (p. 127). Assumed agreements determine what joint action is afoot, if any; we-intentions ensue.

Clearly, agreements are a paradigmatic context for the production of joint intentions and actions. However, as the example of the rowers pursuing a fugitive suggests, agreements as these are ordinarily understood and are not essential. If one says that there is an implicit agreement in such cases, the question arises as to what precisely this claim amounts to, since an implicit agreement does not seem to be an agreement proper.

One might do well to start with an account of paradigmatic agreements. There has been less concentrated analytical focus on such agreements than might have been expected, given the undoubted importance of agreements within the scheme of everyday social concepts. A standard claim, perhaps echoing statements in contract law, has been that an agreement is an exchange of promises. Gilbert has argued at length against this view, on one interpretation of it. She has proposed that typical everyday agreements are, in effect, joint decisions, where the core of a joint decision is a joint commitment to endorse as a body the action plan in question (see Gilbert, 1996: Chapter 13; also Gilbert, 2006: Chapter 10).

Suppose that Phyllis and Joe jointly commit to endorse as a body the decision to get married next year. Presumably, it is reasonable then to regard them as jointly committed to intend as a body to get married next year. Given Gilbert's account of agreements and joint intentions, then, agreements give rise to joint intentions, providing a basis for joint action. Agreements are not necessary for joint action, however, insofar as a joint commitment to intend as a body can be brought into being in the absence of a joint commitment to endorse as a body a certain decision. This accords with the way the relationship of agreements to joint actions is commonly understood.

Individualism, Holism, and Joint Action

Annette Baier has objected to the "individualistic bias" she finds in recent work on joint intention and action (see Baier, 1997). She objects in particular to the assumption that multiperson actions are constructed out of single persons, their abilities, and intentions. Baier proposes that doing things with others is more basic than doing things alone. To say this, however, is not yet to explain what doing things with others amounts to.

Clearly, the range of existing accounts of joint intention and action is wide. They include accounts in terms of personal intentions, with appeal to moral principles such as Scanlon's principle of fidelity to account for the associated entitlements and obligations. (Both Bratman and Tuomela appeal to Scanlon's principle.) They also include Gilbert's account, which derives the associated entitlements and obligations from a joint commitment. Gilbert's account goes beyond what she refers to as singularism, insofar as a joint commitment in her sense is not simply a conjunction of personal commitments (see Gilbert, 1989).

Accounts that make personal intentions central accord with the approach to social phenomena of sociologist Max Weber, who put at the center of his fundamental sociological concepts the concept of an action that is oriented to the actions of others (called by Weber (1922) a "social action"). They harmonize well, also, with the classical ways of characterizing social processes in terms of the mathematical theory of games. Gilbert's plural subject account of social phenomena accords rather with the approaches of Emile Durkheim (1982) and Georg Simmel (1971), among sociologists, and of Jean-Jacques Rousseau (1983) and other political philosophers who invoke an actual contract at the base of a human society. (Gilbert (1989), proposes an interpretation of Durkheim that suggests a commonality between Gilbert's approach and that of Durkheim.) These thinkers all emphasized in one way or another that people engaged in joint action were unified in a significant sense. Economists have begun to move beyond individual choice to investigate what has been called 'team reasoning' (see Bacharach, 2006).

See also: Cooperation: Sociological Aspects; Group Processes, Social Psychology of; Individualism versus Collectivism: Philosophical Aspects; Interactionism, Symbolic.

Bibliography

Alonso, F., 2009. Shared intention, reliance, and interpersonal obligations. Ethics 119, 444–475.
Bacharach, M., 2006. Beyond Individual Choice: Teams and Frames in Game Theory. Princeton University Press, Princeton, NJ.
Baier, A.C., 1997. The Commons of the Mind. Open Court, Chicago, IL.
Bratman, M.E., 1999. Faces of Intention: Selected Essays on Intention and Agency. Cambridge University Press, Cambridge, UK.
Durkheim, E., 1982. The Rules of Sociological Method, first American ed. Free Press, New York.
French, P., Wettstein, H. (Eds.), 2006. Midwest Studies in Philosophy: Shared Intention and Collective Responsibility. Blackwell Publishing, Malden, MA.
Gilbert, M., 1989. On Social Facts. Routledge, London.
Gilbert, M., 1996. Living Together: Rationality, Sociality, and Obligation. Rowman and Littlefield, Lanham, MD.
Gilbert, M., 2000. Sociality and Responsibility: New Essays in Plural Subject Theory. Rowman and Littlefield, Lanham, MD.
Gilbert, M., 2004. Scanlon on promissory obligation: the problem of promisees' rights. Journal of Philosophy 101, 83–109.
Gilbert, M., 2006. A Theory of Political Obligation: Membership, Commitment, and the Bonds of Society. Oxford University Press, Oxford.
Lewis, D.K., 1969. Convention: A Philosophical Study. Harvard University Press, Cambridge, MA.
List, C., Pettit, P., 2011. Group Agency: The Possibility, Design, and Status of Corporate Agents. Oxford University Press, Oxford.
Rousseau, J.-J., 1983. On the Social Contract. Hackett, Indianapolis, IN.
Scanlon, T., 1990. Promises and practices. Philosophy and Public Affairs 19, 199–226.
Searle, J.R., 1995. The Construction of Social Reality. Free Press, New York.
Sellars, W., 1963. Imperatives, intentions, and the logic of 'ought'. In: Castaneda, H.-N., Nakhnikian, G. (Eds.), Morality and the Language of Conduct. Wayne State University Press, Detroit, MI.
Simmel, G., 1971. Georg Simmel: On Individuality and Social Forms. University of Chicago Press, Chicago, IL.
Tuomela, R., 1984. A Theory of Social Action. Dordrecht, Boston, MA.
Tuomela, R., 1995. The Importance of Us. Stanford University Press, Stanford, CA.
Weber, M., 1922. Economy and Society, vol. 1. University of California Press, Berkeley, CA.

Joint Attention during Infancy and Early Childhood Across Cultures

Jessica L Kinard and Linda R Watson, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides an overview of joint attention development. Typical development of joint attention is discussed, followed by joint attention in special populations. Finally, the influence of diverse cultures on joint attention is explored.

Introduction

"Look!" says Ellie, pointing to a seagull. She glances to her mother and then back at the bird.

"Wow, a seagull," says Ellie's mother. "And look! A kite!" She points to the sky and Ellie looks in the same direction, laughing as the kite swoops overhead.

In this interaction, Ellie is demonstrating an important skill, known as 'joint attention.' Joint attention is defined as coordinating attention between a referent and another person, in order to share interest in the referent (Bakeman and Adamson, 1984; Crais et al., 2004). To engage in joint attention, a child needs to either (1) direct someone's attention to a referent (i.e., 'initiation of joint attention') (Crais et al., 2004) or (2) respond when another person directs his/her attention to a referent (i.e., 'response to joint attention') (Mundy et al., 2007). In other words, joint attention interactions include three main components: (1) person A (e.g., the child), (2) a referent (e.g., an object, person, or event), and (3) person B (e.g., an adult or another child). Because joint attention involves these three components, it is sometimes referred to as a triadic interaction (Shin, 2012).

Initiating and responding to joint attention is a complex skill, crucial for infancy and early child development. In this article, joint attention is discussed in the following contexts: (1) typical development, (2) development in special populations, and (3) the influence of diverse cultures on joint attention.

Joint Attention in Typically Developing Children

This section discusses the milestones for joint attention development, types of communicative behaviors associated with joint attention, and relationships between joint attention and other developmental domains.

The nature of joint attention can be clarified by comparing it to other functions of communication. In early development, infants typically communicate for three main reasons, or 'communicative functions': (1) to influence another person's behavior in order to request or protest (i.e., behavioral regulation); (2) to draw attention to self for social purposes, such as during social games (i.e., social interaction); and (3) to draw attention to a referent, in order to share interest in the referent (i.e., joint attention) (Bruner, 1981; Crais et al., 2004). Behavioral regulation may occur as a dyadic (i.e., involving the child and another person) or triadic (i.e., involving the child, another person, and a third referent) interaction. In contrast, joint attention always involves three components; the child needs to establish a joint focus of attention on a referent outside of the dyad. Infants first communicate for behavioral regulation and social interaction; afterward, they begin to communicate for joint attention (Crais, 2011; Crais et al., 2004). Infants begin to develop joint attention around 7–10 months of age and have full joint attention skills between 12 and 18 months of age (Crais et al., 2004; Salley and Dixon, 2007).

Infants communicate to establish joint attention through a variety of communicative behaviors, or 'communicative means,' which could include gestures, eye gaze, vocalizations, and words. When developing ways of communicating, infants first use eye gaze and/or eye gaze combined with vocalizations, followed later by gestures and eventually words. Gestures tend to emerge without vocalizations, and then are combined with vocalizations shortly afterward (Crais et al., 2004). Some gestures are used exclusively for joint attention (e.g., showing an object to someone to share interest), whereas other gestures may be used to communicate for joint attention, as well as other purposes (e.g., pointing at an object to share interest, or pointing at an object to request).

Response to and initiation of joint attention both uniquely contribute to later language development, suggesting that both facets of joint attention play important roles in children's learning (Mundy et al., 2007). In the first 2 years of life, pointing and following another person's point, in particular, are important joint attention milestones (Colonnesi et al., 2010; Vallotton and Ayoub, 2010). Pointing is not the only way to initiate joint attention, nor is pointing restricted to the function of joint attention; however, this gesture is frequently used by young children and, thus, the literature on pointing may be especially relevant to the topic of joint attention.

Pointing and understanding others' pointing is strongly associated with the following skills, particularly when the pointing is for joint attention: (1) current language skills (i.e., in the first 2 years of life) and later developing language skills, including vocabulary related to cognition (e.g., 'think'), and (2) the ability to interpret another person's intentions later on (Colonnesi et al., 2010; Kristen et al., 2011; Mundy et al., 2007). Children can also use joint-attention-pointing as a tool for learning social emotional concepts, even after they learn how to talk (Vallotton and Ayoub, 2010). This literature suggests that both the function and form of joint attention play important roles in early child development.

Since pointing is an important gesture for joint attention development, a discussion of its development is warranted.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23172-3

Researchers suggest that pointing develops gradually, first with an open-hand point and followed by an index-finger point (Liszkowski and Tomasello, 2011). Children who point with an index finger often demonstrate more communicative abilities than children who point with their whole hand. When compared to whole-hand pointers, children who point with an index finger tend to display the following behaviors: (1) pointing at referents more often, (2) combining vocalizations with their pointing more often, and (3) responding more appropriately to another person's point. In contrast, children who only point with their whole hand may look to where another person is pointing, but display less understanding of why the person is pointing to the referent, suggesting that children who point with an index finger have reached a higher stage of social cognitive development (Liszkowski and Tomasello, 2011).

Brain development research provides support for the relationship between pointing and cognition. Responding to joint attention may be associated with posterior attention systems of the brain, allowing children to reflexively respond to another person's bid for joint attention, whereas initiating joint attention may be associated with frontal attention systems in the brain, such as the anterior cingulate cortex, allowing children to understand their own and others' reasons for communicating (Hopkins and Taglialatela, 2013; Mundy et al., 2007). The posterior, reflexive systems are established early on, whereas the anterior, intentional systems are established later in infancy (Mundy et al., 2007). Based on this information, it seems reasonable that how children use pointing for joint attention could provide clues to their brain development.

Despite these linguistic and cognitive benefits, joint attention is not always associated with development of later skills. After children are about 2 years old, joint attention is not as predictive of how well children will learn vocabulary and grammar, perhaps because most children have gained joint attention by that time, and so joint attention is less able to differentiate between children who go on to develop higher or lower language skills (Salley and Dixon, 2007). Also, unlike the ability to follow a point, following another person's gaze is not associated with the ability to interpret that person's intentions; this difference may be because pointing is more salient to infants as a form of communication (Kristen et al., 2011).

In the first 2 years of life, however, joint attention predicts aspects of later development (Colonnesi et al., 2010; Kristen et al., 2011; Mundy et al., 2007; Vallotton and Ayoub, 2010). The child's use of more pointing may result in more joint attention with an adult, and thus, more verbal input to the child about a topic (Colonnesi et al., 2010; Goldin-Meadow et al., 2007; Vallotton and Ayoub, 2010). As adults respond to the child's focus, these interactions may help the child develop language (Goldin-Meadow et al., 2007) and social emotional understanding (Vallotton and Ayoub, 2010).

Although joint attention may promote language acquisition, the reverse may also be true: language skills could help children develop their joint attention. The diversity of a child's vocabulary may influence how well the child engages in joint attention with others, perhaps because a greater vocabulary can help children share more experiences (Vallotton and Ayoub, 2010). Children begin to develop their expressive vocabulary around 12–15 months of age (Mundy et al., 2007). Between 14 and 36 months of age, children continue to solidify their verbal and gestural behaviors, and become more adept at using them to interact with their parents (Vallotton and Ayoub, 2010). Children between 15 and 20 months, in particular, begin combining pointing and language, as their vocabulary increases and they begin putting words together (Colonnesi et al., 2010). Thus, although pointing is an important gesture prior to 1 year of age, it develops even stronger associations with language in the second year of life.

Children follow a general pattern of joint attention development, but their development may vary according to their communicative environment. The following variables could influence development of joint attention: (1) child temperament, (2) child gender, (3) maternal behaviors, and (4) the people involved in the interaction (Benigno et al., 2007; Crais et al., 2004; Mundy et al., 2007; Salley and Dixon, 2007; Shin, 2012).

In terms of temperament, toddlers who display more frequent negative feelings, such as fear and discomfort, tend to have lower language skills, point less often, and initiate joint attention less often; in contrast, toddlers who display more frequent positive affect, such as pleasure, tend to have higher language abilities. Possibly toddlers experiencing negative emotions focus on regulating their feelings more than paying attention to their environment, and thus have more difficulty developing joint attention and language skills (Salley and Dixon, 2007).

Regarding gender, 9-month-old girls initiate joint attention with eye contact more often than their male peers. This gender difference is not specific to joint attention, however, as girls from 9 to 24 months of age also initiate behavioral regulation more often than boys of similar ages (Mundy et al., 2007). These differences in early prelinguistic communication of joint attention and behavior regulation may give girls an advantage in developing language skills at an earlier age than boys.

Finally, the people involved in the interaction may influence the child's joint attention development. When parents engage their children socially and emotionally, these interactions foster imitation, joint attention, and conversational skills. Researchers argue that this social emotional engagement supports the later development of joint attention skills, allowing children to practice paying attention to another person before paying attention to both their play partner and a shared referent (Farrant et al., 2011).

More joint play tends to occur when toddlers are engaged in a one-to-one interaction with a parent, as opposed to a parent–child–sibling interaction. However, when an older sibling joins the play, the toddler and parent demonstrate more joint attention with each other than in the one-to-one situation, suggesting that toddlers are capable of establishing joint attention even when both, they and their sibling, are vying for their parent's attention. Furthermore, toddlers demonstrate larger vocabularies when they establish joint attention in a parent–child–sibling interaction, as opposed to one-to-one situations with a parent. In parent–child–sibling contexts, the siblings establish joint attention more often with their parent than with each other (Benigno et al., 2007).

Some researchers speculate that infants do not actually point for joint attention, but for behavioral regulation (Southgate et al., 2007), an idea that would be supported by

findings that infants do not use pointing with other children, only with adults who are more capable of meeting the infants' behavioral regulation demands. However, Shin (2012) examined social communication skills in 9- to 23-month-old children in a child care setting, and determined that, in addition to regulating each other's behavior, children also communicated with each other for purely joint attention purposes. Therefore, children do appear to communicate with each other simply to share interest in referents. Joint attention between infants and toddlers appears to be an important way for children to develop social cognition and play skills (Shin, 2012).

Joint Attention in Children with Special Needs

Some disorders change the developmental trajectory of joint attention. In this section, joint attention in the following conditions will be discussed: autism spectrum disorder (ASD), Down syndrome (DS), Williams syndrome (WS), blindness, and deafness.

Autism Spectrum Disorder

Joint attention deficits are key impairments in children with ASD. According to the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders, ASD is characterized by deficits in two main areas: (1) social communication and social interaction and (2) restricted, repetitive patterns of behavior, interests, or activities (Paul and Norbury, 2012). Deficits in social communication and social interaction could impact the following areas: back-and-forth social interactions, understanding and using nonverbal communication, and/or starting and keeping social relationships. The category of restricted and repetitive behaviors includes the following areas: repetitive speech, movements, or interactions with objects; insistence on following routines and resistance to change; obsessive interests or fixations; and unusual reactions to sensory stimuli in the environment.

Based on these criteria, it seems logical that children with ASD would struggle with joint attention. First, joint attention is highly dependent on understanding and producing nonverbal communication, particularly when infants are first developing joint attention skills (Crais et al., 2004); however, nonverbal communication is often a deficit in children with ASD (Mundy et al., 1986; Paul and Norbury, 2012). Second, joint attention involves sharing interest in a referent with another person (Crais et al., 2004); however, children with ASD struggle with social interactions (Adamson et al., 2009; Paul and Norbury, 2012). Poor social motivation to share interests with another person may contribute, at least partially, to difficulties with joint attention skills in children with ASD.

Finally, restricted and repetitive behaviors could also disrupt joint attention development. If children with ASD fixate on obsessive interests and have difficulty with change, it will likely be difficult for them to switch their attention between a referent and another person, whether to initiate or respond to joint attention (Adamson et al., 2009; Mundy et al., 2007; Paul and Norbury, 2012). Unusual sensory reactions may also hinder joint attention development; for example, if children are undersensitive to stimuli, then they may not respond easily when someone tries to gain their attention (Paul and Norbury, 2012). It is interesting to note that, although children with ASD have relatively strong behavioral regulation skills, they tend to struggle with 'triadic' behavioral regulation (e.g., switching attention between a referent and another person to request help) more than with 'dyadic' behavioral regulation (e.g., paying attention to only one entity to request) (Phillips et al., 1995; Watson et al., 2013). These difficulties with triadic behavioral regulation suggest that, in addition to social motivation, the triadic nature of joint attention could be a factor that makes joint attention difficult for children with ASD.

Supporting these ideas, a wealth of research exists on the development of social communication in children with ASD. The minimal use of joint attention gestures is what consistently distinguishes children with ASD from other children, both those with typical development and those with other developmental disabilities (Adamson et al., 2009; Paparella et al., 2011; Watson et al., 2013). Infants and young children with ASD are less likely to communicate for joint attention purposes when compared to other children, even when children with ASD have a large expressive vocabulary (Adamson et al., 2009; Watson et al., 2013). For example, typically developing children and children with non-ASD disabilities markedly increase their use of joint attention gestures over the 9- to 18-month age range; conversely, children with ASD have a relatively flat trajectory of joint attention development (Watson et al., 2013) and continue to have difficulties with joint attention at preschool age (Adamson et al., 2009). Gestures associated with joint attention seem to be difficult for children with ASD to learn, such as following another person's gaze, showing referents, and pointing (Paparella et al., 2011). It is important to note that, even though children with ASD clearly struggle with joint attention skills, some infants later diagnosed with ASD communicate for joint attention purposes, even at very young ages (Paparella et al., 2011).

As with typically developing children, joint attention skills of children with ASD may vary according to their background. Caregivers can play a large role in how well children with ASD learn joint attention skills. Parents of children with ASD do not appear to differ from parents of other children, in terms of how much they follow their child's lead or direct their child's behavior (Siller and Sigman, 2002; Watson, 1998); however, a highly responsive caregiving style (i.e., following the child's lead, rather than requiring that the child switch his/her attention) is even more strongly associated with later social communication gains in children with ASD (Siller and Sigman, 2002, 2008) than in typically developing children (Landry et al., 2001; Tamis-LeMonda et al., 2001). This type of parenting style may be particularly important for children with ASD, given their difficulties with joint attention.

Down Syndrome

DS is a genetic condition, which, in most cases, results from an extra copy of chromosome 21. Children with DS exhibit characteristic physical traits, such as low muscle tone, large tongue, fold of skin on the eyelid, and small height. Children with DS also demonstrate delays in many areas of development, such as cognition, motor skills, and receptive and expressive language (Paul and Norbury, 2012).

Despite difficulties with language, social communication appears to be a strength for children with DS (Adamson et al., 2009). Children with DS tend to develop joint attention more slowly than typically developing children, but in a similar developmental progression (John and Mervis, 2010). Compared to 18- and 30-month-old typically developing children, 30-month-old children with DS can often maintain joint attention at similar levels (Adamson et al., 2009). Preschoolers with DS can also frequently tell the difference between communicative gestures and actions that look like gestures, but are noncommunicative (e.g., pointing with an index finger versus holding out a finger, but not actively pointing at anything) (John and Mervis, 2010). Children with DS can often find hidden objects when an adult gestures toward the object, indicating an ability to understand others' gestures, even when the children with DS have low cognitive and language skills (John and Mervis, 2010). As with typically developing children and children with ASD, children with DS tend to have difficulty following another person's gaze and understanding what the gaze means; pointing combined with eye gaze appears to be an easier gesture to understand (John and Mervis, 2010; Kristen et al., 2011; Paparella et al., 2011). Despite their overall abilities to engage in joint attention, children with DS have more difficulty than children with ASD at producing expressive language, even after 30 months of age (Adamson et al., 2009).

Williams Syndrome

WS is a genetic condition that occurs when about 25 genes are deleted from chromosome 7q11.23. Children with WS demonstrate characteristic physical features, such as a small chin and upturned nose, and characteristic behaviors, such as being overly friendly and anxious. Children with WS also demonstrate delays in many domains, such as visual-spatial difficulties, nonverbal delays, and intellectual disabilities; however, some children with WS have intellectual abilities in the low normal range (Paul and Norbury, 2012).

Similar to children with DS, children with WS can often tell the difference between a communicative and noncommunicative gesture. Unlike children with DS, most children with WS have difficulty finding hidden objects when an adult gestures toward the object, suggesting that understanding these gestures is a challenge. Not surprisingly, children with WS struggle to interpret another person's eye gaze and have an easier time following pointing. Overall, children with WS demonstrate relatively strong language skills and motivation for social interaction; however, they struggle with social communication skills, like understanding and producing joint attention gestures. Therefore, children with WS and DS each have unique profiles: children with WS demonstrate strengths in language and difficulties with joint attention, whereas children with DS struggle with language, but have relatively strong joint attention (John and Mervis, 2010).

Blindness

Children who are blind from birth may have early difficulties learning language, but typically catch up by the time they enter school (Paul and Norbury, 2012). Joint attention skills, on the other hand, are more challenging for these youngsters. For many congenitally blind children, early social communication may appear similar to children with ASD, although there are many differences between the two conditions (Bigelow, 2003; Brown et al., 1997; Hobson et al., 1999). Children who are blind tend to demonstrate more typical social affect and more instances of pretend play than sighted children with ASD, even if the children who are blind demonstrate other characteristics of ASD (Hobson et al., 1999).

It is easy to imagine why joint attention would be difficult for children who have visual impairments. Typically developing children largely rely on vision to coordinate their attention between a person and a referent, both to see what another person is referring to and to direct another person's attention (Crais et al., 2004). Children who are blind cannot rely on vision for joint attention, but instead must rely on other means, such as touch, hearing, and memory of spatial orientations (Bigelow, 2003). These alternate means may be difficult for some blind children to learn, such as children with cognitive impairments; thus, blind children with cognitive impairments may display more ASD-like characteristics than blind children without additional deficits (Brown et al., 1997).

Cognitive skills are highly involved in developing non-visually based joint attention. For children who are blind, a stepping stone to joint attention is the development of object permanence. These infants must first understand that objects exist and can be physically manipulated, before they can learn to coordinate their attention between objects and another person. Infants may demonstrate understanding of object permanence by actively searching for objects and by using another person as a tool for finding those objects. For example, an infant might feel along a person's arm to locate the object in his/her hand, but not actively communicate with the other person to share interest in the object. Infants who are blind might also engage in self-stimulating behaviors, such as head shaking, which parents can help shape into communicative behaviors, like shaking head for 'no' at around 18 months. These self-stimulating behaviors tend to decrease as children are better able to control their surroundings (Bigelow, 2003).

As children who are blind learn words, they start to use those words to label objects and follow verbal instructions, indicating a growing understanding of objects in their surroundings (Bigelow, 2003). Early response to joint attention can be seen when a child and a parent are playing with an object together, but the child fusses when the parent stops playing. However, even once object permanence and social interactions are established, initiating joint attention is often difficult for children who are blind, since they cannot see what is available in their environment. Once they begin to establish a vocabulary, children who are blind can more easily respond to and initiate joint attention; however, they continue to have difficulty initiating joint attention about distant objects and abstract concepts. Communication partners can provide context clues to help children who are blind perceive their environment and focus their attention (Bigelow, 2003), perhaps helping to reduce ASD-like symptoms (Hobson et al., 1999). Therefore, although joint attention is important for developing language skills, the reverse is also true: language

skills are important for developing joint attention, particularly for children who are blind (Bigelow, 2003).

Deafness

Children with severe to profound hearing impairments demonstrate a range of language and social communicative abilities. The language environment in which children grow up plays a large role in how they learn to communicate, such as whether they are born to parents who are deaf or with typical hearing. The communicative styles of these parents are highly variable, which could impact how children learn to engage in joint attention (Gale and Schick, 2009; Paul and Norbury, 2012).

Regarding both oral and signed communication, deaf 2-year-olds with deaf parents tend to engage in less joint attention using words than hearing 2-year-olds with hearing parents. This difference may be because, when using sign language, deaf children need to split their attention between looking at another person's signs and looking at the object, whereas hearing children can simultaneously look at and communicate about the object. Also, deaf children may need more time to visually explore objects before they can communicate about them. Among deaf 2-year-olds of hearing parents, toddlers tend to use more words (including both oral and signed vocabulary) during joint attention than deaf children of deaf parents, but fewer words than hearing children of hearing parents (Gale and Schick, 2009).

Despite these differences in word use during joint attention, deaf 2-year-olds with deaf parents tend to have similar expressive vocabularies as hearing children with hearing parents, when both sign and oral vocabularies are included in the count. This suggests that the deaf toddlers of deaf parents are not necessarily at a disadvantage for learning language within joint attention routines, simply because they use less words than hearing children in these routines; rather, deaf toddlers may simply need more time to look back and forth between the referent and adult when communicating. On the other hand, deaf 2-year-olds with hearing parents tend to have much lower expressive vocabularies (including both oral and signed vocabulary) than hearing children with hearing parents (Gale and Schick, 2009).

A possible explanation for these differences is that deaf parents may be more sensitive to their deaf child's need to visually explore objects, giving them more time to process information, whereas hearing parents tend to engage their children in several, short interactions about an object, rather than maintaining attention on one topic (Gale and Schick, 2009). Deaf parents also tend to use tactile cues, such as tapping the child to gain his/her attention, before signing words (Akhtar and Gernsbacher, 2008). Another possible explanation for these vocabulary differences relates to the parents' linguistic fluency. Deaf parents are likely to be fluent in sign language, and so could provide rich vocabulary models for their children; on the other hand, hearing parents may be beginning learners of sign language, and so might not provide as many vocabulary models for their children (Paul and Norbury, 2012). Furthermore, although hearing parents may have rich oral vocabularies, their child's hearing impairment prevents the child from fully accessing this information. These styles of interaction may influence how many words a toddler is able to process and learn within joint attention routines.

Joint Attention in Diverse Cultures

In this section, the influence of diverse cultural backgrounds on joint attention will be considered, both in relation to typical development and development of children with special needs.

In general, infants from diverse cultures demonstrate similar social communication skills as infants in the United States (Blake et al., 2005; Callaghan et al., 2011; Childers et al., 2007; Salomo and Liszkowski, 2013). For example, as Japanese and Italian-Canadian infants approach 14 months, they tend to increase their joint attention gestures (Blake et al., 2003, 2005). In Nigeria, infants also tend to increase the complexity of their joint attention from 1 to 2 years of age (Childers et al., 2007). These joint attention interactions seem to be important for later receptive language development in both Japanese and Italian-Canadian infants (Blake et al., 2003) and expressive language development in Nigerian toddlers (Childers et al., 2007), similar to studies of children in the United States (Mundy et al., 2007). Infants from the Netherlands, China, and Mexico also demonstrate joint attention gestures at 11 months of age, such as pointing, showing, and offering objects to establish joint attention (Salomo and Liszkowski, 2013). Somewhat differently from these groups, French infants tend to increase their joint attention up to about 11.5 months of age, and then reduce their frequency of joint attention, while infants who are English-Canadian tend to have high levels of joint attention gestures throughout the 9- to 14-month age range (Blake et al., 2005).

Regarding response to joint attention, infants from Canada, Peru, and India are able to respond to joint attention bids at a similar rate, including both (1) following a point to an interesting picture from 8 to 13 months of age and (2) following gaze to find hidden objects at 12 months of age. However, infants from Canada show somewhat more success at finding hidden objects than the infants from Peru and India, perhaps because the Canadian infants have more opportunities to move around independently (Callaghan et al., 2011).

Despite these similarities, differences in joint attention and related aspects of social communication also exist across cultures. Some of these differences may relate to the amount of time that infants spend in social interaction. For example, infants in China may have more opportunities for interaction than Dutch infants in the Netherlands, and so Chinese infants tend to point more often; along similar lines, Dutch infants may have more opportunities for interaction than Mayan infants in Mexico, and so, in turn, Dutch infants tend to point more often than Mayan infants (Salomo and Liszkowski, 2013). Parent education may play a role in how children from different cultures develop gestures. In a study of 12-month-olds from India, Canada, and Peru, Indian infants pointed for joint attention significantly less often than Canadian infants, while Peruvian infants pointed at a rate somewhere in the middle, not significantly different from either the Canadian or Indian infants (Callaghan et al., 2011). The authors proposed these differences may relate to

maternal education, rather than cultural differences per se, since the mothers from India had lower education than the other two groups.

The use of eye gaze, so important in Western cultures in establishing joint attention, is another behavior that varies across cultures (Akhtar and Gernsbacher, 2008). For example, from 9 to 14 months, Japanese infants may not increase how frequently they coordinate eye contact with other gestures for communication, indicating that, despite the importance of eye contact in English-speaking cultures, it may not be as significant in Japanese social communication (Blake et al., 2003). Additionally, mothers in Kenya tend use less eye contact with their infants than mothers in the United States, instead focusing on interactions that involve touch and holding the child close. Families in Papua New Guinea and in Palestine tend to hold their infants facing outward, rather than inward, making eye contact more difficult to establish and changes in head motions more difficult to notice. Therefore, tactile and verbal means of communication may be more relevant in these cultures than visual communicative means, and could potentially function in the same role as eye gaze (Akhtar and Gernsbacher, 2008). The use of touch is particularly important for developing joint attention of infants who are blind, as well as infants who are deaf, so it seems logical that tactile cues could also be important for joint attention in other cultures (Akhtar and Gernsbacher, 2008).

In summary, infants from diverse cultures demonstrate both similarities and differences in joint attention development. An interesting consideration is whether diverse culture plays a role in the joint attention of children with disabilities. Research on this topic is rare, but a study was recently completed in the United States, comparing the communication of monolingual and bilingual toddlers with ASD (Valicenti-McDermott et al., 2013). Although the bilingual English-Spanish-speaking toddlers demonstrated more cooing, gesturing for behavior regulation, and simple pretend play than the monolingual English-speaking toddlers, the groups were similar in their use of social interaction gestures (e.g., waving goodbye) and joint attention gestures (e.g., showing objects). Based on these findings, it appears that learning two languages does not impair the joint attention development of children with ASD, and may be associated with benefits in other aspects of communication development.

Conclusion

Joint attention is a social communication ability that develops early in infancy and progresses throughout early childhood, allowing children to coordinate their gaze between a referent and another person, in order to share interest in that referent. In early development, communicating for the purposes of behavioral regulation and social interaction generally precedes communicating to establish joint attention, and the acquisition of these two communicative functions may help form the foundation for joint attention. Although development varies based on individual differences, children progress through a typical hierarchy of social communication development, allowing them to solidify their joint attention skills through use of gestures, eye gaze, vocalizations, and words. In turn, these joint attention skills are related to concurrent and longitudinal gains across many developmental areas, such as language, social cognition, social emotional understanding, and joint play.

Developmental and physical disabilities may disrupt this typical progression, making it difficult for children to acquire joint attention. Overall, infants and toddlers from diverse cultures demonstrate similar joint attention skills; however, some cultural differences in joint attention do exist. Cultural differences may also influence how social communication symptoms are presented in children with disabilities.

See also: Autism Spectrum Disorders; Cognitive Development During Infancy and Early Childhood across Cultures; Culture as a Determinant of Mental Health; Infancy and Human Development; Infant Gaze Behaviors; Intentionality During Infancy and Early Childhood, Development of; Language Development: Influence of Socio-Economic Status; Mother-Child Communication: Cultural Differences; Neurophysiological Markers of Language Development; Parenting Attitudes and Beliefs across Cultures; Play During Infancy and Early Childhood: Cultural Similarities and Variations; Social Referencing During Infancy and Early Childhood across Cultures; Social and Emotional Development in the Context of the Family; Temperament and Human Development; Theory of Mind During Infancy and Early Childhood Across Cultures, Development of.

Bibliography

Adamson, L.B., Bakeman, R., Deckner, D.F., Romski, M., 2009. Joint engagement and the emergence of language in children with autism and Down syndrome. Journal of Autism and Developmental Disorders 39, 84–96.

Akhtar, N., Gernsbacher, M.A., 2008. On privileging the role of gaze in infant social cognition. Child Development Perspectives 2 (2), 59–65.

Bakeman, R., Adamson, L.B., 1984. Coordinating attention to people and objects in mother-infant and peer-infant interaction. Child Development 55 (4), 1278–1289.

Benigno, J.P., Clark, L., Farrar, M.J., 2007. Three is not always a crowd: contexts of joint attention and language. Journal of Child Language 34 (1), 175–187.

Bigelow, A.E., 2003. The development of joint attention in blind infants. Development and Psychopathology 15, 259–275.

Blake, J., Osborne, P., Cabral, M., Gluck, P., 2003. The development of communicative gestures in Japanese infants. First Language 23 (1), 3–20.

Blake, J., Vitale, G., Osborne, P., Olshanskey, E., 2005. A cross-cultural comparison of communicative gestures in human infants during the transition to language. Gesture 5, 201–217.

Brown, R., Hobson, R.P., Lee, A., 1997. Are there "autistic-like" features in congenitally blind children? Journal of Child Psychology and Pyschiatry 38 (6), 693–703.

Bruner, J., 1981. The social context of language acquisition. Language and Communication 1 (2/3), 155–178.

Callaghan, T., Moll, H., Rakoczy, H., Warneken, F., Liszkowski, U., Behne, T., Tomasell, M., 2011. Early social cognition in three cultural contexts. Monographs of the Society for Research in Child Development 76 (2), 1–142.

Childers, J.B., Vaughan, J., Burquest, D.A., 2007. Joint attention and word learning in Ngas-speaking toddlers in Nigeria. Journal of Child Language 34 (2), 199–225.

Colonnesi, C., Stams, G.J.J.M., Koster, I., Noom, M.J., 2010. The relation between pointing and language development: a meta-analysis. Developmental Review 30, 352–366.

Crais, E., Douglas, D.D., Campbell, C.C., 2004. The intersection of the development of gestures and intentionality. Journal of Speech, Language, and Hearing Research 47, 678–694.

Crais, E.R., 2011. Testing and beyond: strategies and tools for evaluating and assessing toddlers. Language, Speech, and Hearing Services in Schools 42, 341–364.

Farrant, B.M., Maybery, M.T., Fletcher, J., 2011. Socio-emotional engagement, joint attention, imitation, and conversation skill: analysis in typical development and specific language impairment. First Language 31 (1), 23–46.

Gale, E., Schick, B., 2009. Symbol-infused joint attention and language use in mothers with deaf and hearing toddlers. American Annals of the Deaf 153 (5), 484–503.

Goldin-Meadow, S., Goodrich, W., Sauer, E., Iverson, J., 2007. Young children use their hands to tell their mothers what to say. Developmental Science 10 (6), 778–785.

Hobson, R.P., Lee, A., Brown, R., 1999. Autism and congenital blindness. Journal of Autism and Developmental Disorders 29 (1), 45–56.

Hopkins, W.D., Taglialatela, J.P., 2013. Initiation of joint attention is associated with morphometric variation in the anterior cingulate cortex of chimpanzees (*Pan troglodytes*). American Journal of Primatology 75, 441–449.

John, A.E., Mervis, C.B., 2010. Comprehension of the communicative intent behind pointing and gazing gestures by young children with Williams syndrome or Down syndrome. Journal of Speech, Language, and Hearing Research 53, 950–960.

Kristen, S., Sodian, B., Thoermer, C., Perst, H., 2011. Infants' joint attention skills predict toddlers' emerging mental state language. Developmental Psychology 47 (5), 1207–1219.

Landry, S.H., Smith, K.E., Sank, P.R., Assel, M.A., Vellet, S., 2001. Does early responsive parenting have a special importance for children's development or is consistency across early childhood necessary? Developmental Psychology 37, 387–403.

Liszkowski, U., Tomasello, M., 2011. Individual differences in social, cognitive, and morphological aspects of infant pointing. Cognitive Development 26, 16–29.

Mundy, P., Block, J., Delgado, C., Pomares, Y., Vaughan Van Hecke, A., Venezia Parlade, M., 2007. Individual differences and the development of joint attention in infancy. Child Development 78 (3), 938–854.

Mundy, P., Sigman, M., Ungerer, J., Shermon, T., 1986. Defining the social deficits of autism: the contribution of non-verbal communication measures. Journal of Child Psychology and Psychiatry 27 (5), 657–669.

Paul, R., Norbury, C., 2012. Language Disorders from Infancy through Adolescence: Listening, speaking, reading, writing, and communicating, fourth ed. Elsevier Mosby, St. Louis, MO.

Phillips, W., Gómez, J.C., Baron-Cohen, S., Laá, V., Riviére, A., 1995. Treating people as objects, agents, or "subjects": How young children with and without autism make requests. Journal of Child Psychology and Psychiatry 8, 1383–1398.

Paparella, T., Stickles, G.K., Freeman, S., Kasari, C., 2011. The emergence of nonverbal joint attention and requesting skills in young children with autism. Journal of Communication Disorders 44, 569–583.

Salley, B.J., Dixon, W.E., 2007. Temperamental and joint attentional predictors of language development. Merrill-Palmer Quarterly 53 (1), 131–154.

Salomo, D., Liszkowski, U., 2013. Sociocultural settings influence the emergence of prelinguistic deictic gesture. Child Development 84 (4), 1296–1307.

Shin, M., 2012. The role of joint attention in social communication and play among infants. Journal of Early Childhood Research 10 (3), 309–317.

Siller, M., Sigman, M., 2002. The behaviors of parents of children with autism predict the subsequent development of their children's communication. Journal of Autism and Developmental Disorders 32 (2), 77–89.

Siller, M., Sigman, M., 2008. Modeling longitudinal change in the language abilities of children with autism: parent behaviors and child characteristics as predictors of change. Developmental Psychology 44, 1691–1704.

Southgate, V., van Maanen, C., Csibra, G., 2007. Infant pointing: communication to cooperate or communication to learn? Child Development 78 (3), 735–740.

Tamis-LeMonda, C.S., Bornstein, M.H., Baumwell, L., 2001. Maternal responsiveness and children's achievement of language milestones. Child Development 72 (3), 748–767.

Valicenti-McDermott, M., Tarshis, N., Schouls, M., Galdston, M., Hottinger, K., Seijo, R., Shulman, L., Sinner, S., 2013. Language differences between monolingual English and bilingual English-Spanish young children with autism spectrum disorders. Journal of Child Neurology 28 (7), 945–948.

Vallotton, C.D., Ayoub, C., 2010. Symbols build communication and thought: the role of gestures and words in the development of engagement skills and social-emotional concepts during toddlerhood. Social Development 19 (3), 601–626.

Watson, L.R., 1998. Following the child's lead: mothers' interactions with children with autism. Journal of Autism and Developmental Disorders 28, 51–59.

Watson, L.R., Crais, E.R., Baranek, G.T., Dykstra, J.R., Wilson, K.P., 2013. Communicative gesture use in infants with and without autism: a retrospective home video study. American Journal of Speech-language Pathology 22, 25–39.

Journalism

Daniel C Hallin, University of California, San Diego, La Jolla, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Journalism refers to the activity of gathering and presenting news for public dissemination. It has played a central role in modern society as a political institution, a business, and a form of popular culture. It has taken many forms in different historical periods and political and economic contexts, varying, for example, in its relation to political institutions, its status as a profit-making business, and the degree to which it is 'professionalized'.

Journalism is generally defined as the activity of gathering and presenting information and commentary about current events for public distribution. From a social science point of view, it can be conceived in a number of ways: as a distinct occupation with its own patterns of recruitment and socialization and occupational subculture; as a distinct form of social activity involving the gathering and processing of information; as a social institution with a set of relations with other institutions like the state; or as a cultural form or form of discourse.

Disputes about the boundaries of journalism have been common, among both scholars and practitioners. Chalaby (1998), for example, has argued that journalism properly so-called emerged in Britain and the United States with the development of commercial mass media; the production of news as a commodity, in his view, is radically distinct from the earlier practice of using the press to publicize political ideas. Journalists' organizations, particularly in countries with state-owned broadcasting systems, have sometimes excluded those who worked in broadcast news, on the grounds that they did not practice journalism independently, but merely transmitted the views of others. In many countries today there is debate about whether independent Internet 'bloggers' or those who produce new forms of 'infotainment' like talk-show hosts can legitimately be called journalists. There is little to be gained conceptually by trying to adjudicate these disputes and establish the 'true boundaries' of journalism. These disputes are enlightening precisely because multiple forms of journalism have always coexisted, and because journalism has never been rigidly differentiated from other social activities, but has overlapped with and existed in relations of mutual interdependence with a variety of political, economic, social, and cultural institutions and practices.

Historical Evolution

The term 'journalist' emerged in the seventeenth century; originally, 'journalism' referred simply to the activity of producing a printed periodical. One can trace many of the social functions involved in journalism to earlier eras, but it seems best to say that journalism emerged with the development of the newspaper and other forms of print culture. The emergence of journalism was connected not only with the development of print technology, but also with an interrelated set of social and cultural changes. First, it was connected with the development of capitalism and the market. This was true in two senses. The market, first of all, increased the demand for specialized sources of information. Those involved in market relationships had an intensified need to know about occurrences beyond their own world of face-to-face relationships. Merchants were among the first producers and consumers of newspapers, and as increasing parts of the population were drawn into the market economy, newspaper readerships grew. Second, newspapers were themselves distributed through the market, and in many countries newspapers were among the first mass-produced consumer commodities. This is not true in all cases, however. Party newspapers in Europe, for example, were often distributed through party-related organizations, and newspapers in Communist systems also distributed through organizations and in workplaces.

Second, the emergence of journalism was connected with the rise of the nation-state and political citizenship. It is connected of course with the emergence of democratic politics; early newspapers were often primarily vehicles for public political participation. Expansions of newspaper readership were often closely connected with expansions of the electoral franchise. In the United States, for example, the rise of the 'penny papers', which greatly expanded the audience for news, came on the heels of the Jacksonian revolution and the elimination of property qualifications for voting. In Uruguay, one of the few Latin American countries in which a mass circulation press has developed, newspaper circulations increased from 1 newspaper for every 11 inhabitants in 1870 to 1 for every 4 in 1916, a period that also saw the rise of mass-based political parties.

The notion of citizenship is broader than that of democracy, however, and the connection between journalism and the rise of the nation-state goes beyond the development of democratic political institutions. Another way in which the modern nation-state draws the mass public into political life is through mobilization for war; newspaper circulations have typically expanded in times of war as well as in times of expanded political participation. As Anderson (1991) has pointed out, the culture of nationalism involves a sense of belonging to an imagined community of compatriots; journalism is one of the key cultural forms in the creation of this sense of national collectivity, and it can play this role with or without political democracy.

Finally, journalism is connected with the rise of realism as a cultural form; it emerges simultaneously with a number of

related realist genres, including the novel and scientific writing, and indeed has in certain historical periods overlapped with these, as novels and scientific information were published in newspapers and novelists often wrote for the press. All three of these historical changes are connected with the rise of the 'public sphere'.

Max Weber (1946) once characterized the journalist as a 'type of professional politician'. And indeed in its early days journalism was substantially – though not exclusively – a part of the world of politics. Journalists were often advocates for political causes; Benjamin Franklin, for example, who like many others in his age was simultaneously a journalist and a politician, described journalism as essentially a form of oratory and thus part of a tradition that could be traced back to the political assemblies of Greece and Rome. In democratic countries, newspapers typically became associated with political parties, as these began to emerge, or factions of parties, and often were supported financially by parties or individual politicians. In authoritarian countries, newspapers were often political mouthpieces of either the state or insurgent movements. In Africa, for example, many leaders of the independence movements of the twentieth century were involved in journalism.

The identification between journalism and politics was substantially changed, however, by rise of a commercial newspaper industry. This began in the Unites States in the 1830s, with the development of the 'penny papers', which expanded newspaper circulations by an order of magnitude and became profitable cultural industries by selling the attention of a mass audience to advertisers (Schudson, 1978). In Britain, commercialization of the press was delayed by restrictive taxes, but got under way in the 1850s. In Japan, similarly, commercial papers began to compete with the earlier, politically oriented papers at the end of the 1870s. Commercialization did not push politics aside altogether. Most newspapers in the United States, for example, continued to have strong partisan loyalties throughout the nineteenth century, and their owners and editors to be active in politics. They were, however, financially independent of parties, and their owners were independent political powers in a way their predecessors often had not been. In this sense, it could be said that commercialization increased the power of journalism as an institution. Commercialization also changed the agenda and style of journalism, diminishing political advocacy in favor of the competition to provide news in a timely and sometimes sensational fashion.

A sharp debate has emerged in media studies over the implications of commercialization of the news media for democratic politics. One view holds that commercialization gives the press the independence it needs to serve as a 'watchdog' of the state and to provide the public as a whole with information 'without fear or favor'. The contrary view holds that commercialization tends to drive political content out of the press, replacing it with sensational human interest material, and also to consolidate control of the press in the hands of the upper class, leading to unequal political competition between competing social interests (Curran and Seaton, 1997). This argument has been developed especially from the case of Britain, where a highly politicized working-class press flourished in the period when taxes inhibited the development of commercial media. Weber (1946) also observed that the commercial papers in Germany "have been regularly and typically the breeders of political indifference".

Commercial newspapers were the first to employ hired reporters, editors, and eventually artists, photographers, and other specialized journalistic personnel. Throughout the nineteenth century, journalism developed as a distinct occupation with specialized practices and a sense of identity. Many of the conventions for gathering and presenting information, which we now think of as distinctively journalistic, emerged in these years, including the summary lead, which synthesizes the 'most important' information in a news story for the reader, and the practice of interviewing public figures, which was important in part because it gave journalists a more active role in the creation of news, and in some cases increased the public profile of individual journalists.

In the twentieth century, the notion of journalism as a 'profession' began to emerge, most distinctively, perhaps, in the United States, where the idea of neutral expertise became a particularly strong element of political culture. Professionalization meant, above all, the development of the idea that the journalist serves the public as a whole, rather than particular political tendencies, owners, or other interests. It is associated with the shift, noted by Siebert et al. (1956), from the older 'libertarian' conception of press freedom to the 'social responsibility' model, which sees the press as the holder of a public trust. Professionalization was also associated with a shift toward balance and 'objectivity' as primary journalistic values (though more strongly in the United States, than in many other countries), and with increased autonomy of journalists, as against owners or managers, within news organizations. Journalistic autonomy has nevertheless always been limited: despite aspirations in some countries for direct journalistic control of news organizations (perhaps most notably France after World War II; the journalists at *Le Monde* once controlled and still participate in the election of the director), ultimate authority in virtually all news organizations remains outside the hands of journalists.

Several decades of scholarship have painted a complex picture of the implications of the professionalization of journalism. Professional norms can be seen, for example, both as a mechanism of social control, constraining journalists and often serving to exclude nonconforming content from the news, and as a basis for journalists' autonomy, legitimizing their right to a space of action not constrained by their owner's political views, the pressures of advertisers, etc. The concept of professional routines has become central to the analysis of the social role of journalism. Routines are the standard practices that enable news organizations to operate efficiently and that legitimate the many choices that must be made in the production of news, allowing them to be treated as matters of shared professional judgment rather than of political debate. Where professionalism is strong, these routines are powerful enough that particular individuals, including both journalists and owners, have limited ability most of the time to shape news content. These routines, however, often have social biases built into them, and most scholars consider the biases built into journalistic routines to be the most important explanation for the politics of news (Schudson, 2003). Among the most important routines is the use of certain kinds of sources of

information that are considered 'authoritative' – principally government officials and other elite members of society (Gans, 1979; Hall et al., 1978).

There are many views about what kinds of social 'biases' prevail in contemporary journalism. Some, for example, have argued that journalism generally undermines the authority of established social institutions (this view began to grow in the 1970s, which probably saw the high point of journalistic autonomy in most Western countries, and also a decline in the legitimacy of many social institutions). Some have argued that they serve more or less consistently to support established social interests, and others that they are essentially neutral, and represent the full range of contending interests. The most common view among scholars is probably that in liberal societies media routines tend to reproduce the existing structure of power, but not in an entirely consistent way; under certain circumstances, for example, the dominant routines can open the media to penetration by new social movements or lead to periods of contestation over the boundaries of political debate.

Differentiation

Structural functionalism holds that human societies tend to evolve toward increased differentiation, in which institutions become increasingly specialized in the functions they perform. Some scholars – most notably Alexander (1981) – have argued that the development of journalism should be seen in this light. This does make considerable sense in societies where commercialization and professionalization are relatively advanced. Journalism in these societies has increasingly become separated from party politics, and has developed norms and routines of its own, a 'media logic' distinct from the logic of politics. In other ways, however, the notion that journalism tends to evolve toward differentiation seems doubtful. Most importantly, such a view neglects the tendency for journalism to become integrated into market mechanisms. Professionalization, to be sure, limits this to some degree, but the 'media logic' that has emerged in contemporary societies is very much a joint product of professional norms and the pressures of the market, which many argue are growing increasingly strong. Other scholars thus give a very different account of the history of journalism than the structural functionalists. Habermas, for example, decries the 'colonization' of the public sphere by the cultural industries, and Bourdieu argues that the journalistic field once heavily influenced by the political field is now heavily influenced by that of the market.

Journalism and the State

In liberal societies, journalism is generally thought of as standing apart from the state, serving to check its power as a watchdog or the fourth estate. This is accurate to a significant degree, as reflected in incidents like the publication by *The New York Times* and other papers of the 'Pentagon Papers' (a secret government study of US decision making on the Vietnam War 'leaked' to the press by a dissident former official). But close ties between the media and the state also exist in liberal societies. State subsidies and restraints do exist in such societies, as do clientelist relationships between politicians and media owners. But more important is the close relationship between journalists and government officials as sources and subjects of news, respectively, a relationship that has proved mutually beneficial to officials seeking publicity and control over the news agenda, and to journalists seeking news. This relationship is highly developed, with the news media organizing much of the newsgathering process around the institutions of the state, and the state devoting considerable resources to accommodating them. In many cases, the relationship is formally organized, as in the lobby system in the British parliament or the press clubs of Japan. The closeness of the relationship between contemporary journalists and state officials means that journalists have access to considerably more information about the inner workings of government than they would have as outsiders. It also means that they are often in an important sense part of the community of political elites, and reflect the prevailing viewpoints within that community. Here, again, doubts can be raised about the differentiation theory of journalism history: it could be argued that journalism is more entangled with the state than it was a century or two ago, when journalists had no privileged access to the corridors of power.

In many political systems, the news media are much more directly integrated into the state, because either they are directly owned by state and party apparatuses, as in China, or few sources of capital exist to support news media outside of the state, as in some developing countries. In other cases, they will be tied more indirectly, through dependence on state advertising, for example, payments to journalists, or personal ties between owners and politicians.

Varieties of Journalism

Journalism has always been plural; indeed one of the distinctive characteristics of the modern newspaper is the fact that it mixes widely disparate types of content, not only in terms of subject matter, but also in terms of the functions it fulfills for readers. Journalism provides both information and entertainment. It also provides moral and ideological orientation. Even in countries where objectivity became a dominant professional value, there continue to be many occasions when journalists communicate value judgments to their readers, particularly when the society's consensus or hegemonic values are considered to be at stake, in covering international conflict, for example, or in investigative reporting. Some scholars have also argued that journalists perform a ritual function, promoting a sense of community solidarity, particularly in coverage of ceremonial events and crises.

In most societies, there is also some differentiation of journalism into distinct styles, often serving different parts of the community. One clear example is the differentiation of the British press between the 'quality' or broadsheet press, serving middle- and upper-class readers and emphasizing informational journalism, and the mass or tabloid press, serving primarily working-class readers with a much more entertainment-oriented form of journalism. In France, a distinction has often been drawn among 'rationalist', 'sensationalist', and 'ideological' newspapers, the latter including Catholic and Communist

organs. In many countries, local and regional papers are more commercially oriented and lack distinct political orientations, while national (or capital city) papers reflect distinct political tendencies. In the United States, ethnic minority media have played an important role, and often practice a form of journalism that was intended less to provide neutral information than to defend the interests of a particular part of the community.

Diverse social and political systems have also given rise to diverse forms of journalism. In Western Europe, newspapers tied directly or indirectly to political parties and other organized social groups like trade unions or churches remained important for most of the twentieth century, and the functions of journalism included much more emphasis on representing ideas and social interests in the public sphere as well as participating in the process of bargaining among groups, parties, and politicians, alongside the information and entertainment functions associated with commercial media. The party press has declined dramatically since the 1970s in Europe, but most newspapers in Southern Europe still have partisan identities and alliances, newspapers across the region continue to have more clearly differentiated political orientations than those in the United States, and European journalists are more likely to see commentary and to some degree advocacy as integral to the journalist's role than their US counterparts (Hallin and Mancini, 2004).

In many parts of the world, journalistic partisanship in one form or another is more the norm than the exception. Its form is often different from the party press of Western Europe, as the strongly institutionalized political party system of twentieth-century Europe is distinct from the political structures found elsewhere. In much of Asia, for example, the relation between media and politics is characterized by a fluid kind of partisanship based on shifting personal alliances, in which journalists both serve and participate in complex processes of maneuvering among political and economic power brokers. Rather than stable partisan alliances and identities, as one often finds with European news media, many Asian newspapers are characterized by what McCargo (2012) calls 'partisan polyvalence,' with different journalists at the same paper expressing the views of different actors with which they have ties. In many systems, journalism is very much an instrument by which particular actors – politicians, business people, leaders of ethnic groups or clans – pursue their economic and political interests; in this context, of course, the analysis of 'professional routines' referenced above, which is derived from the Western experience, is much less relevant to understanding the behavior of journalists or the content and role of the news media.

Journalism in Communist societies also evolved into distinct models that remain strong in countries like China and Vietnam. This model had its roots in the Western European tradition of party journalism, oriented toward political mobilization. But with the fusing of party and state, journalists became official figures and distinctive functions evolved. Journalism is an important institution in a Communist system, and news media are typically well supported by the state, trade unions, and other official institutions to which they are attached. Their functions are varied and complex, and include mobilizing people around the policy goals of the party-state, gathering social information for confidential dissemination within the political elite, serving as vehicles for debate among factions and ideological tendencies within the party and society, as well as serving a kind of ombudsman function, as an institution to which ordinary people could complain, hoping to be heard by those in power. In the Chinese case, the introduction of market mechanisms into parts of the Chinese media has introduced further complexity of different forms and functions of journalism, without changing its essential integration into the party-state (Zhao, 2012).

Current Developments

A number of powerful forces are currently transforming journalism and its social role.

Commercialization

In many countries, the news media have become increasingly commercialized and journalism more strongly influenced by market factors. In the West, the news media have been substantially commercialized since the nineteenth century. But for much of this century the effects of market forces on the news media have been partly counterbalanced by professionalization, public service broadcasting (which had a monopoly in Europe until the 1980s), and press subsidies, which many European countries expanded in the 1970s to keep politically oriented papers alive and prevent market forces from eliminating competing voices. These counterbalancing forces have been weakened in recent years, though public service broadcasting remains an important institution in much of Europe and some other parts of the world.

The culture of journalistic professionalism has been undermined in important ways by the shift to commercialism, and may also have been undercut by a general decline in the prestige of political life. A key element of that culture was the idea that journalists should make independent judgments about what citizens needed to know about public affairs. Today, it is increasingly assumed that it is the journalist's job to provide whatever content interests the media consumer. The news agenda has shifted away from public affairs in the traditional sense, and toward human interest and 'service' journalism useful to ordinary people in everyday life (investment advice, health reporting) (Hallin, 2000). Journalistic autonomy has been undercut in important ways, as the business and editorial sides of news organizations are increasingly integrated.

Some argue that these changes have made journalism less elitist, more responsive to the concerns of the majority of the population, and others that they undermine the integrity of the system of public communication. Journalism clearly has important functions for the society as a whole that most private industries do not have. It is central to political life, and to the process of opinion formation more generally; some have argued that it is increasingly central, as other institutions, political parties and trade unions, for example, have weakened. It remains one of the key questions of media studies whether market mechanisms can provide the fulfillment of these functions, and if not, what other social mechanisms can be devised.

The Economic 'Crisis of News'

In the United States and some other parts of the world, the economic model that underpinned the most important news media has been under strain for some time, and reached a point of crisis during the first global economic downturn of the twenty-first century. The number of newspapers has generally been declining in Western countries since the 1950s, and since the 1980s, many news organizations have cut back drastically on their staffs. The US television networks, for example, which once had foreign correspondents stationed around the world, now have virtually none, and most newspapers have cut the size of their reporting staffs dramatically, eliminating many specialized beats (arts and culture, for example) and much of their daily coverage of government, business, and other social institutions.

Emergence of New Forms of Partisanship and the Decline of 'Objectivity'

The party press, as noted above, declined in Western countries in the mid- to late twentieth century, this occurred particularly early and strongly in the United States. That trend has been partially reversed, however, as partisan media have been revived in the United States, and a shift has taken place from factual reporting of information to an emphasis on opinion, commentary, and satire, from forms of journalistic discourse that emphasize 'objectivity' to more 'subjective' ones (Baym, 2010). This change is associated with cultural shifts (from 'modernism' to 'postmodernism') and with an economic and technological shift from a highly centralized media environment where the monopoly newspaper and the government-regulated television networks were central to the multichannel media environments of radio, cable television, and the Internet. In the latter, many media target niche audiences distinguished by demographic category, lifestyle, and worldview, and partisanship is often an effective marketing device.

Democratization

In many developing and newly industrialized countries, there has been a move toward liberal democracy in recent years, with many authoritarian regimes giving way to multiparty systems and greater political openness. An increased emergence of independent news media has been associated with these changes, probably as both cause and effect. Opposition parties have increasingly gained access to the media in many countries, political scandals have proliferated (Waisbord, 2000), and elites have lost much of the tight control over public discourse they once held. Often commercialization of the media has accompanied democratization, and this has led to a dramatic reorientation of journalism, in many cases, away from serving government officials and toward the mass audience, sometimes resulting in rapid shifts from 'protocol news' to sensational crime reporting. In many cases, however, early visions of the emergence of independent journalism oriented toward public service have been undercut, sometimes by violence against journalists, often connected with the rise of drug cartels (Mexico, Honduras); sometimes by political polarization (Venezuela, Ecuador); and sometimes by reconsolidation of control by a new power center (Russia).

Internationalization

News media have developed primarily within the context of national political systems. There has always been some degree of international influence and exchange, most notably, starting in the late nineteenth century, through the role of the global wire services, which were an important basis for Western dominance of the flow of world news. But most journalists worked within a primarily national context. Most still do, but the internationalization of journalism has clearly accelerated. The advent of CNN as a global news service was an important step in this direction. The most important contemporary example is Al Jazeera, which, along with some other forms of pan-Arab television, has created an increasingly significant transnational public for news within the Arab world. The Internet has also made news increasingly accessible across borders.

New Technologies

The Internet and other forms of digital technology are changing journalism in many ways. They have undercut the separation of technology-based forms of journalism, particularly television and print, as Web-based journalism combines both. They have changed the timescale of journalism from the periodic cycle of daily news to a more continuous flow. They have provided new means for dissemination of news, and increased the audience for many news organizations, though often it has been difficult for news organizations to 'monetize' their online audiences, and the Internet has contributed to the collapse of the business model of newspapers. And they have displaced professional journalists in important ways from the central role they once played as 'gatekeepers' of news, as other actors, including politicians, political institutions, interest groups, and individual citizens, can disseminate news and information directly to the public. Much of the information circulated on the Internet is still produced by journalists for major news organizations, and in that sense the continuing importance of journalists in the flow of information should not be underestimated; often bloggers and other citizens recirculate and comment on what appears in the 'mainstream news media' rather than gathering information as professional journalists do. But there are many examples where these nonprofessional actors have major impacts, where video shot by citizens with cell phones becomes central to reporting, for example, or where a public is built around some political issue first in the 'blogosphere' and later is taken up by professional journalists (Williams and Delli Carpini, 2011).

See also: Agenda Setting, Media Effects On; Citizen Journalism; Freedom of the Press; Journalism and Journalists; Mass Media and Politics; Media Events; New Media, News Production and Consumption; News Interview; News: General; Political Communication; Public Sphere and the Media.

Bibliography

Alexander, Jeffrey, 1981. The mass news media in systemic, comparative and historical perspective. In: Katz, E., Szecscö, T. (Eds.), Mass Media and Social Change. Sage, Beverly Hills, pp. 17–52.

Baym, Geoffrey, 2010. From Cronkite to Colbert: The Evolution of Broadcast News. Paradigm Press, Boulder, CO.
Chalaby, Jean, 1998. The Invention of Journalism. Macmillan, London.
Curran, James, Seaton, Jean, 1997. Power without Responsibility: The Press and Broadcasting in Britain. Routledge, London.
Gans, Herbert J., 1979. Deciding What's News: A Study of CBS Evening News, NBC Nightly News, Newsweek and Time. Pantheon, New York.
Hall, S., Crichter, S., Jefferson, T., Clarke, J., Roberts, B., 1978. Policing the Crisis. Macmillan, London.
Hallin, Daniel C., 2000. Commercialism and professionalism in the American news media. In: Curran, James, Gurevitch, Michael (Eds.), Mass Media and Society, third ed. Arnold, London, pp. 218–237.
Hallin, Daniel C., Mancini, Paolo, 2004. Comparing Media Systems: Three Models of Media and Politics. Cambridge University Press, New York.
McCargo, Duncan, 2012. Partisan polyvalence: characterizing the political role of Asian media. In: Hallin, D.C., Mancini, P. (Eds.), Comparing Media Systems beyond the Western World. Cambridge University Press, New York.
Nichols, John, McChesney, Robert W., 6 April 2009. The death and life of great American newspapers. The Nation, 1, 11–20.
Schudson, Michael, 1978. Discovering the News: A Social History of American Newspapers. Basic Books, New York.
Schudson, Michael, 2003. The Sociology of News. W.W. Norton, New York.
Siebert, Fred S., Patterson, Theodore, Schramm, Wilbur, 1956. Four Theories of the Press. University of Illinois Press, Urbana, IL.
Waisbord, Silvio, 2000. Watchdog Journalism in South America: News, Accountability and Democracy. Columbia University Press, New York.
Weber, Max, 1946. Politics as a Vocation. In: Gerth, H.H., Wright Mills, C. (Eds.), From Max Weber: Essays in Sociology. Oxford University Press, New York.
Williams, Bruce A., Delli Carpini, Michael X., 2011. After Broadcast News: Media Regimes, Democracy and the New Information Environment. Cambridge University Press, New York.
Zhao, Yuezhi, 2012. Understanding China's media system in a world historical context. In: Hallin, D.C., Mancini, P. (Eds.), Comparing Media Systems beyond the Western World. Cambridge University Press, New York.

Journalism and Journalists

Slavko Splichal, University of Ljubljana, Ljubljana, Slovenia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Journalism was born in the seventeenth century as a supplementary, part-time occupation of printers, postmen, or tradesmen; with industrial and bourgeois revolutions, it became fully institutionalized in the mass media. In earlier periods, journalists largely belonged to self-sustained free professions of the middle class; with industrialization they joined the class of workers employed by the press publishers. In the age of media digitalization and convergence, an increasing number of journalists belong to the precariat class or – according to the 'ideology of nonprofessionalism,' which suggests that 'everyone can be a journalist' in the digital age – they are not (considered) part of the professional journalistic workforce at all. Journalists are commonly associated nationally into trade unions and professional associations. The largest international association of journalists is the International Federation of Journalists.

The journalist is the key employee of news organizations who, in the definition adopted by the International Labor Organization, "researches, investigates, interprets and communicates news and public affairs through newspapers, television, radio and other media"; main tasks of journalists include "collecting, reporting and commenting on news and current affairs for publication in newspapers and periodicals, or for broadcasting by radio, television or webcast media." Due to national peculiarities in labor statistics and recent developments in 'online journalism,' a global estimate of the number of journalists remains quite unreliable. International Federation of Journalists (IFJ) claims to represent around 600 000 members in more than 100 countries (2012). Most journalists are employed in the United States, Japan, Germany, and the United Kingdom.

Journalistic work is primarily based on the *symbolic manipulation* of the content (sense making) in contrast to those, like typographers, responsible for the *mechanical* production of the newspaper or other news medium. In large and complex electronic media systems, journalism is contextualized and codetermined by other media output and occupations. Several groups of employees, who are occupied with symbolic manipulation or involved in administrative and organizational tasks, are conventionally not considered journalists. The development of electronic and, in particular, digital production and distribution of media contents, and the growth of new formats and genres of journalism (e.g., tabloids in the newspaper industry; talk shows on TV; blogging, social media in the Internet; hybrid genres such as 'infotainment' and 'advertorials') brought about numerous occupational groups (technicians and other employees in broadcasting, documentary production, audiovisual workers, computer programmers, designers) whose work cannot be classified exclusively into either side, i.e., they participate in both symbolic and technical (re)production of contents. The boundaries between journalism and nonjournalism, and between full-time and part-time journalists become even more problematic in the cyberspace dominated by a constant flux of information produced, reproduced, and distributed by millions of Internet 'visitors.'

There is no uniform 'journalistic culture' or concerted form of professional operation of journalists that would lead to a single concept of journalistic work and norms. In contrast to the rather technical meaning of journalism in English and most other languages of Western Civilization referring to the workforce in institutionalized forms of (mass) communication, the Greek concept of journalism, *δημοσιογραφία* (*demosiografía*), means 'writing about/for the people' and is focused on the social substance of journalism by emphasizing people rather than a particular class, or nature. Since *demos* means more than population but *citizenship*, journalism implies political relevance, which indeed was its most significant common trait until very recently. A critical labor perspective in the 1960s led to the concept of *newsworkers*, suggesting a distinction between news and entertainment or fiction without solving the problem of a clear sociological referent of the concept. In the Muslim world, journalists are required, similar to those in the West, not to 'mix truth with falsehood' and to verify news, but reporting is not allowed to be conceived on suspicion, to ruin people's reputation, or to mock them, and most importantly, news must have a 'use value' in that it must provide benefit (*Nafa*') to the people. The latter idea was central to 'development journalism' emerging in the 1960s in developing countries, emphasizing that reporting should contribute to national development of the country concerned. Similarly, 'public' or 'civic journalism' movements in the West in the 1990s suggest that journalists should not merely report, but discover a community's 'relevance' by meeting the people and experiencing their lives.

People have always passed on information from one to the next, but journalism as a distinctive occupational task did not begin until the seventeenth century, when it became a supplementary, part-time occupation of otherwise professionally committed individuals in the social communication process – e.g., printers, postmen, or tradesmen. Modern journalism, as a result of industrial and bourgeois revolutions, became fully institutionalized with the mass media. During that period, the press was largely liberated from censorship and party subservience. R. Park noted that the rise of the reporter had been one of the most important events in the 'age of news.' While in earlier centuries, journalists largely belonged to self-sustained free professions of the middle class, with industrialization they joined the class of workers employed by the press publishers. The separation between the occupations of publisher and journalist was followed by the establishment

of national and international professional organizations and trade unions of journalists.

During the nineteenth century, newspapers were generally still closely linked to political parties and journalists subject to the political credo of newspapers. The idea of objectivity represented an alternative to partisan journalism and a way to legitimate the professional status of journalists. The quest for objectivity was related to the need to establish a sort of professional, scientifically validated method for journalists to ensure their professional autonomy based on a 'standard' university education, also typical of traditional professions. The idea of journalistic objectivity as a professional norm was not widely recognized until the end of the World War I. In the early 1900s, the demand for objective reporting was embedded in the broader cultural movement of 'scientific naturalism,' which stressed the importance of methods in giving evidence of facts.

The conflict between subjective storytelling (the blending of facts and opinions, real events and trivial fictional material, news and entertainment into a tightly packaged story) and 'objective news' (the classic function of journalism to sort out factual and reliable accounts of daily, particularly politically relevant events) was never solved. The growing number of journalistic *codes of ethics* since the beginning of the twentieth century to mark professional standards in journalism has never resulted in a single code that would universally define the standard of 'objectivity,' including the professional standards adopted by IFJ. There are two main reasons: (1) The two tendencies of storytelling and objective reporting are not mutually exclusive but intermingling extreme strategies in reporting events; and (2) The competition for consumers causes competition for the latest news and makes journalists vulnerable to politics; political authorities enter journalism either directly through interviews and press conferences, or indirectly through 'information subsidies' (making information available to journalists on a *quid-pro-quo* basis), which influence the journalistic agenda setting, if not attitudes and actions, and eventually make journalism subsidiary to *public relations*. For example, in Japan, news sections of newspapers are produced in close cooperation with powerful news sources in government, industry, and the civil service, organized through the 'press club' system.

Journalism lacks the objective criteria – stressing the importance of universal norms, service to clients or altruism, intellectual activity and scientifically based expertise, autonomy and responsibility, and ethics – that would place it in the same social position as 'true' professions such as medicine or law. In many countries journalists have no specific professional standards to meet, which became even truer with the most recent development of 'digital journalism' in the Internet. The rise of the Internet and other technologies has radically changed the production and diffusion of news; an array of new authors, web publishers, and journalists, professional and nonprofessional, were brought in to create and distribute news online.

Although national codes of ethics share some 'universal' norms such as prohibiting editorial advertising and misleading audiences, protecting individuals against invasion of privacy, defamation and damaging publicity in general, and social groups against hate speech, it is increasingly difficult to enforce them in the online environment. Specific forms of (self-)regulation such as courts of honor, press or media councils, or press or media ombudsmen exist in a great number of countries to supervise the implementation of ethical standards by professional journalists and the media. They may be composed of journalists alone (as in the case of courts of honor within journalist associations) or representatives of the general public and media organizations.

There are invariably controversies between scholars, media workers, and media industries about the structure of journalism studies and the type of education needed for journalists. Early courses in journalism were mostly practical or vocational and connected with on-the-job training. In the 1920s and later, journalism curricula expanded into social sciences and included courses such as communication theories, public opinion, history of journalism, and ethics. The number of educational institutions in journalism has been growing throughout the twentieth century, starting with the University of Missouri in 1908 in the United States (which at present has 350 schools of journalism; one-third of them accredited at the university level) and soon followed by other regions in the world, where educational programs and materials were largely based on those from the USA and Europe. One of the first journalism schools linked to universities was set up in 1918 at the National Peking University in China. In Latin America, the number of journalism faculties and schools increased from 13 in 1950 to almost 300 by the end of the century. There is no 'typical' journalism curriculum, but it usually combines journalistic and nonjournalistic courses, with the prevalence of the latter (e.g., in the USA in a ratio 25–75). Many schools offer specialization in print, radio and television journalism, public relations, and advertising; fewer have more specialized programs, e.g., in photojournalism, magazine journalism, or television documentary. Generally, two seemingly contradictory trends may be found in the development of journalism programs: more and more specialized and even fragmented curricula in different fields of journalism on the one hand, and a more 'generic' approach to journalism stressing the need for instruction in communication skills cutting across specific job demands and combined with nonjournalistic specialization, such as arts, agriculture, economics, law, or politics. Both trends are restricted to larger and developed countries where many universities and schools can compete to 'supply' journalists to widely differentiated media.

In the developed world, a large proportion (often exceeding 50%) of newcomers to the profession graduate in journalism at a university or its equivalent. An international comparative study indicates that on average, students choose journalism as senior teenagers, primarily because they believe that the work of journalists is interesting and useful for society, and that they personally are talented for this job. However, many of those who graduate in journalism never enter a journalistic career but rather other communication-related or even noncommunication occupations. Specific education is not a necessary condition to enter a journalistic occupation. The lack of entry conditions such as agreed upon professional qualifications and licensing may be related to the liberal idea that freedom of expression and press freedom should not be restricted in order not to violate fundamental citizens' freedoms. Almost all countries outside the United States have a body of press or media law but journalists have no special legal status other

than defined by the constitutional rights and freedoms of citizens and, in some countries, some minor state legislation, e.g., the protection of information sources. Nevertheless, media owners and editors may define specific qualities and conditions that candidates for journalism must fulfill, and journalists may use mechanisms such as codes of ethics and forms of self-regulation to protect not only group autonomy but also monopoly in the labor market.

The lack of formal education in terms of theoretically grounded knowledge within a particular field contributes to the controversy about professionalism in journalism. Acquisition of such knowledge is central to the selection processes (search for the 'right' person for a job), and instrumental (skills) and expressive socialization (norms, ethics), by which entry into a profession may be controlled. The latter dimension of socialization gives primacy to the intellectual component of professionalism and to the production of expert knowledge, resulting from research. In 1922, W. Lippmann complained that no research had been conducted in the field of journalism despite its immense impact on democratic life. Although German *Zeitungswissenschaft* studied systematically the press beginning in the nineteenth century (e.g., journalistic genres, economics of newspapers), M. Weber was the first to suggest in 1910 to pay attention to journalism as a profession. He explained that a sociology of newspapers ought to include the study of journalists, part-time journalists, and other newspaper employees; 'collectivism and individualism in the creation of the content of the newspaper'; 'the anonymity of newspaper' and its effects upon journalists; cultural specificities influencing journalism and the social functions of the press; and 'the fate and the situation of journalism as a profession' (particularly the questions of education and professional demands upon the journalist). Weber's ambitious research program was not systematically carried out until the 1950s when during the rise of 'mass communication as a scientific discipline' in the United States, journalists became the subject of empirical research. Similarly to university education, however, studies in media occupations and professionals were not always well received by working journalists.

Before the rise of online journalism, in most countries journalists were closer to lawyers and politicians than teachers or public servants in terms of prestige and salaries. They were part of local and national élites although they were on average much younger if compared to other professions. Particularly, journalists in developing countries began with rather high social status. Generally, however, the journalists' opportunity for internal promotion within profession was limited even then. Comparatively low pay, inadequate work conditions, routine work, and limited opportunities for vertical mobility are key factors of horizontal mobility of journalists. They often move to other media fields such as public relations or advertising, and less frequently they leave the field of mass media, e.g., by moving to politics.

Since the 1990s, the rise of the Internet and digitalization of the traditional mainstream media led to the decline of the traditional media (particularly newspapers), stimulated the development of new online genres, and changed the nature and social (professional) status of journalism. These processes also brought about practices once incompatible with professional journalism, such as free Internet publication of newspapers' content to attract more advertisers; the flood of free media financed by advertisements; free 'public relations articles' replacing news stories and thus undermining newspaper credibility; advertising on editorial pages – undermining newspaper credibility even further; and particularly, reduction of editorial staff and newsroom budgets. In addition, strong professional mobility between the worlds of journalism and politics, which was particularly characteristic of the former socialist countries, and a too close relationship of journalists with advertising, entertainment, and public relations may severely endanger the autonomy of journalists and mute their critical voices.

However, when journalists are critical to political and commercial events and actors, as in 'investigative journalism,' they are often exposed to intimidation and harassment by the elites, and even assassinations are used to silence unwanted journalistic voices not only in wartime, or against war correspondents. In 21 years since 1990, when the IFJ has been issuing annual reports on the killing of journalists and media staff, 2271 journalists and media staff were killed or murdered in the course of their work, and many other disappearances were listed as under investigation by IFJ and the International Press Institute. The killings are only the tip of an iceberg of physical assaults and jailings, which affect journalists every year. However, the most widespread and epidemic peril of independent journalism in many countries is an increased number of lawsuits against journalists based on the legal provisions that the media should not harm the rights to privacy and reputation of individuals (including public officials); libel actions may be a disguised way of preventing journalists from performing a 'watchdog function' in society. Civil libel compensations became a sword of Damocles particularly in Eastern Europe and Russia in the 1990s, but prosecutions of journalists and high compensations threaten the media with bankruptcy and generally dampen investigative journalism in many other countries with a longer democratic tradition.

In terms of social and demographic characteristics, journalists do not represent the population of citizens who rely on their reporting and interpretations, but are expected – like other professions – to carry out their work in the interests of the audiences rather than their own interests. Such expectations may be disputed by the facts that professional journalists disproportionately originate from the families in which professional and white-collar occupations prevail, live in elite neighborhoods, are disproportionately male (although this disparity is lessening during last decades), and their views of the world commonly depart from those of an 'average citizen,' usually leaning toward more liberal and democratic political views, in contrast to technical staff in media organizations who are usually more conservative. For example, in a 2000-*Los Angeles Times* survey, 82% of US journalists were found to back abortion compared to only 49% of the general population. This is not to argue that journalists are a politically coherent population. In some countries, such as Italy, journalists are clearly 'divided' among different political affiliations, a tendency advocated by some contemporary regulatory measures supporting (in the case of press also subsidizing, e.g., in Norway and Sweden) diversity and political plurality of the media.

The composition of the journalistic workforce is generally becoming more diversified in terms of gender, but men still

represent the majority of the journalistic workforce. The increase of the number of female journalists indicates that journalism is no more a male preserve, but it also reflects a tendency of the intimization that permeates journalism. Yet the proportion of female journalists is still lower than the proportion of female students of journalism; female journalists are concentrated in the junior posts and in the nonnews sector (e.g., magazines), but the number of women producers, executives, chief editors, publishers, and other decision-making positions remains very low (in some countries even below 10%) and they tend to be paid less than men despite the fact that they are, on average, better educated than their male counterparts. The differences may be partly a consequence of the fact that female journalists relatively recently entered the profession, but they also may have to do with discrimination against women in journalism.

The rise of digital journalism and global recession that started in 2007 radically transformed the publishing world, and the nature and structure of journalist jobs. More and more newspapers, magazines, radio, and TV stations cut down their workforce in order to lower expenses and become more competitive on the market, or they had to close down altogether for the loss of audiences or advertising revenues. As all media got digitalized to attract particularly young audiences, which largely are not reading conventional news any more, and the Internet became the mode of carriage for all traditional press and broadcast media, journalists are increasingly supposed to master digital communication tools from traditional e-mail services to robust interactive Web sites. They are expected to be able to do everything in the digital environment with converging platforms: interviewing and writing, audio and video recording, photographing, and producing graphics. By repurposing the existing content in different (offline and online) platforms, fewer journalists can produce more. In consequence, traditional news media are often believed to have decreased professional standards of journalism to the detriment of objective reporting on current events and in-depth analysis of controversial national and international affairs, in order to remain competitive on the global market. An increasing number of (particularly young) journalists are not part of the professional journalistic workforce or they belong to the precariat class with part-time jobs and no job security.

These changes also spurred an 'ideology of non-professionalism' suggesting that 'everyone can be a journalist' in the digital age. Individual citizens – 'citizen journalists' – may present news through their blogs or online media services without being 'professional' journalists. The traditional editor or journalist is no longer a gatekeeper who selects news. Citizens are journalists and editors by themselves and for themselves, up- and downloading what they want with their personal computer devices. Due to the low cost of modern communication technology, the difference between digitalized traditional mainstream media (such as online newspapers) and new digital nonmainstream media (such as individual blogs) in terms of audience size and production costs is becoming less indicative than it used to be in the traditional divide between the mainstream and alternative media.

The competitiveness of commercially driven 'citizen journalism' where citizens – in addition to advertisers – play an active role by collecting, reporting, analyzing, and distributing news is best exemplified by the South Korean 'citizen journalism organization,' *OhmyNews* (www.ohmynews.com) with the motto "Every Citizen is a Reporter." The online newspaper, which was founded in 2000, now employs 70 full-time staff members; it has 62 000 citizen contributors sending 150 stories each day for no or (exceptionally) very low fee, and more than 2 million daily users. In addition to the Korean language site, it also has an English (OhmyNews International) and Japanese site.

The same hybrid development in 'global participatory citizen journalism' was followed by many other commercial news providers, including *CNN iReport* (http://ireport.cnn.com/) launched in 2006. Stories provided by members of the CNN audience are published on *iReport* Web site without being edited unless they are made part of CNN's 'professional' news coverage on CNN television and CNN.com. iReport claims to 'receive a submission from every country on the planet with an average of 2.4 million unique users each month.' Whereas this hybrid model of journalism is no doubt profitable for the owners, it represents a brutal devaluation of the labor of newsworkers. Changing patterns of journalism and the murky future of journalism as a vocation and full-paid occupation threatens to bring journalism close (back) to what Weber (1919) suggested a century ago: "The journalist belongs to a sort of pariah caste, which is always estimated by 'society' in terms of its ethically lowest representative."

The least ambiguous mark of 'professionalism' in journalism is the establishment of professional organizations. Journalists are commonly associated nationally either into trade unions or professional associations; frequently, their organizations combine both these functions. Yet a long-term tension exists between professional associations, which stress the professional status of journalists (and usually also include employers and managers), and trade unions, concerned primarily with collective negotiations on behalf of working journalists (newsworkers) with employers. Journalists often tend to define themselves as experts claiming to be above politics, but in some countries, journalistic associations also pursue political goals. The criteria for membership of journalists' trade unions and professional associations are fluid, particularly in relation to part-time journalists and 'boundary occupations' such as public relations. Trade unions of journalists negotiate for salaries, working conditions, and fringe benefits, and in some cases they also protect journalists' independence, e.g., by financial support to journalists penalized for refusing to do work incompatible with the professional standards (National Union of Journalists in the UK) or, as in France, by effectuating the same benefits to journalists who would leave jobs because of the substantial political changes in the editorial policy as to those who lose their jobs involuntarily. In many situations unions also serve as 'tools of the employer' because an imbalance in power exists between unions and employers due to the lack of unity among unions and guilds, their weak economic position and goals narrowed to individuals' rewards.

The largest international associations of journalists and media staff are the International Federation of Journalists (IFJ) and the Union Network International (UNI) Global Union, replacing Media and Entertainment International (MEI). First established in 1926 and relaunched in 1946 and 1952, IFJ is

the largest international journalistic organization with 600 000 members and 140 member unions in 125 countries on all continents (May 2012). The aims of IFJ include the promotion of international actions to defend press freedom through autonomous national trade unions or (as in the USA) guilds of journalists and other 'authors,' representation of journalists in the international trade union movement, and assistance to journalists and their unions in fighting for their professional and industrial rights. Journalists from socialist and some other countries established in 1946, the International Organization of Journalists (IOJ) as the counterpart to the IFJ. The IOJ, which also claimed the IFJ of 1926 as its predecessor, was dissolved in 1991 after the collapse of socialism. Members of 'technical' occupations in the arts, audiovisual, cultural, entertainment, and mass media sector who formerly represented the bulk of the membership of the MEI, merged in 2000 with three other international union bodies into the UNI, which represents 15.5 million members in 900 unions from more than 140 countries (May 2012).

See also: Agenda Settting, Public Policy In; Attitudes, Political and Public Opinion; Audience Measurement; Citizen Journalism; Information Society; Internet and Social Media: Anthropological Aspects; Journalism; Mass Communication: Normative Frameworks; Media Ethics; News Interview; News: General; Professions, Sociology of; Public Opinion: Social Attitudes; Television: History.

Bibliography

Bourdieu, P., 1998. On Television–The Power of Journalism. The New Press, New York.

de Beer, A.S., Merrill, J.C., 2008. Global Journalism: Topical Issues and Media Systems, fifth ed. Allyn & Bacon, Boston.

Dahlgren, P., Sparks, C. (Eds.), 1991. Communication and Citizenship: Journalism and the Public Sphere in the New Media Age. Routledge, London.

Hardt, H., Brennen, B. (Eds.), 1995. Newsworkers: Toward a History of the Rank and File. University of Minnesota Press, Minneapolis.

Javnost-The Public, 1996. Special issue "journalism at the crossroads" (3, 3).

Journal of Communication, 1992. Special issue on "journalism in crisis and change" (42, 3).

Kunczik, M., 1988. Concepts of Journalism North and South. Friedrich-Ebert-Stiftung, Bonn.

Lippmann, W., [1922]1960. Public Opinion. Macmillan, New York.

Merrill, J.C., 1983. Global Journalism. Longman, New York.

OECD, 2010. The Evolution of News and the Internet. http://www.oecd.org/dataoecd/30/24/45559596.pdf.

Park, R.E., 1967. On Social Control and Collective Behavior. University of Chicago Press, Chicago.

Robert, W., McChesney, R.W., Pickard, V. (Eds.), 2011. Will the Last Reporter Please Turn Out the Lights: The Collapse of Journalism and what Can Be Done to Fix It. The New Press, New York.

Schudson, M., 1978. Discovering the News: A Social History of American Newspapers. Basic Books, New York.

Splichal, S., Sparks, C., 1994. Journalists for the 21st Century. Ablex, Norwod, NJ.

Weber, M., [1910]1976. Toward a Sociology of the Press. Journal of Communication 26 (3), 96–101.

Journals, Scholarly

Mark Ware, Mark Ware Consulting, Bristol, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The journal plays a central role in scholarly communication. Its core functions of registration, certification, dissemination, archiving, and rewarding have remained intact so far across the online transition and the exploration of new publishing models like open access. Peer review is critical to the journal, and remains widely supported despite some robust criticism. New technologies have the potential radically to reimagine the online journal, but conservative scholarly communication behaviors suggest that future innovation will be evolutionary rather than disruptive.

Journals and Scholarly Communication

The journal – born in, and emblematic of, the Enlightenment – remains central to most fields of scholarly endeavor and of growing importance even in those fields where the monograph has previously held sway. In many ways, its transition to the online world has been smoother than for other media, preserving its core functions, its visible appearance, and its business model. At the same time, its details have never been more contested.

To understand the journal, we should consider it as an integral part of the wider system of scholarly communication, and bear in mind that the exponential growth in journals and articles is not a creation of the publishing industry, but is a consequence of the growth of investment in research.

Scholarly communication takes place across a spectrum from informal to formal. At the most casual are communications between individual researchers: e-mails, telephone calls, and conversations at conferences. Blog posts, still rare in most fields, are rather more formal (they can have quasi-persistent uniform resource locators, and editors increasingly allow their citation). Preprints and working papers are situated significantly further along the spectrum, followed by conference presentations and papers, and at the most formal, the journal article and scholarly monograph.

Defining the Journal

History of the Journal

The first journals were established in the 1660s; the first were *Philosophical Transactions* and the *Journal des Scavans*, established, respectively, by Henry Oldenburg on behalf of the Royal Society and Denis de Sallo in Paris. They took advantage of then new technologies – the printing press and the postal service – to distribute the findings of learned society members, and (crucially) to establish priority over claims.

The Five Core Functions

Within the scholarly communication system, the journal's role is substantially wider and more important than just the dissemination (or 'publication') of results. Its functions are usually seen as follows:

- *Registration*: third-party recording of the author's precedence and ownership of an idea
- *Certification*: providing quality control and establishing the validity of a scholarly claim through peer review
- *Dissemination* or *Awareness*: communicating the findings to the intended audience
- *Archiving*: the preservation of the scholarly record over time
- *Reward*ing: participants (primarily authors) are rewarded for their performance based on various metrics derived from the scholarly communication system.

Journals play an important navigational role in filtering readers' finite attention on relevant articles. There are a large number of journals, and even within individual specialist disciplines there can be many more journals and articles published than an individual researcher could ever hope to read. Although the 'article economy' is increasingly important, the filtering of publications by journals focused on particular topics, serving a particular community, and stratified by quality or significance of findings remains important.

Publishing in journals is thus an integral part of being a researcher in most scholarly fields. It connects them to other researchers in their field and keeps them abreast of developments. It allows (at least in principle) work to be repeated and the results tested. It provides a permanent record, sometimes referred to as the 'minutes of science,' on which future work can reliably be built. Researchers benefit from the promotion and dissemination of their work and gain recognition, credit, and funding in the process.

In the print world, these canonical functions (registration, certification, etc.) were bound together in the physical medium of the journal. Many commentators have pointed out, however, that this need not be the case in the online world: the journal could be disaggregated, and in particular the dissemination function could be separated from that of certification. We will return to this idea under 'Peer Review.'

The Journals Publishing Cycle

The movement of information between the different participants in the journal publishing process is called 'the publishing

 http://dx.doi.org/10.1016/B978-0-08-097086-8.41037-8

cycle.' The author's article passes through peer review and (if accepted) on to the publisher, thence to subscribing institutional libraries – typically via a subscription agent – before ending up back in the hands of the reader. In the world of electronic publishing, of course, readers also obtain journal articles directly from the publisher. This is a 'cycle' because scholars then build on the ideas and information in published articles to formulate new research questions and hypotheses.

In this traditional model, there is a clear-cut distinction between the informal and the formal, the unpublished and the published.

An emerging (though minority) model adopts a more open perspective: early drafts (with closed dissemination) are followed by preprints (open dissemination), open prepublication peer review, formal publication (open distribution), and then discussion and reuse through which various kinds of post-publication review and evaluation take place.

Peer Review

Central to all this is the role of peer review, which both endorses the quality of the researcher's work – its validity, significance, and originality – and brings about improvements to the quality of published articles.

Types of Peer Review

The main difference between approaches to peer review has been in the extent to which the authors' and reviewers' names are known to each other or hidden ('blinded').

The most common has been single-blind review, in which the authors' names are known to the reviewers, but the reviewers' names are withheld from the authors.

Double-blind review is the next most common approach after single-blind; here neither the authors' nor the reviewers' identities are disclosed to the other.

Open peer review is a poorly defined term, with the 'openness' applying variously to the actors' identities, to the submitted manuscript and/or the reviewers' reports, and to the set of people able to comment on submitted manuscripts. In most variants, the authors' and reviewers' names will typically be known to each other, though journals then vary on whether the reviewers' names are subsequently made public. Some journals, however, use open review to describe the publication of the reviewer reports and editorial correspondence even if this remains anonymous.

Critiques

For all its totemic status, peer review is far from immune to criticism. The comprehensive set of charges leveled at it concerns its validity, effectiveness, efficiency, and fairness.

The charge that there is an absence of reliable empirical evidence that peer review leads to a higher quality of research (as opposed to research publications) seems damning. The absence of evidence is not the same thing as evidence of absence, though, and the critics accept that this is both a high hurdle and a question of considerable difficulty to test in the real world.

Various studies have shown peer review indeed lacks effectiveness in some areas. For instance, it has been shown to be unreliable (e.g., interreviewer reliability is generally low), and not very good at detecting errors. It also adds delay to the publication process, arguably slowing down the dissemination and hence progress of science (Ware, 2011).

Peer review can be seen as a 'burden' on researchers, in the sense that it represents real work to be done, but this is generally accepted to be an appropriate part of the professional duties of academics.

The evidence for a 'crisis' in peer review, with growth in submissions overwhelming the numbers of available researchers, is weak. Research outputs are correlated directly with the number of active researchers, so more research outputs means more potential reviewers. There may, however, be transitional regional imbalances.

Peer review is also criticized for its potential unfairness. Single-blinded review clearly allows scope for reviewer bias or misconduct. Journals are exploring the increased use of both double-blinded review and of more open review to mitigate these flaws.

New Developments in Peer Review

The foundational role of the peer-reviewed journal in the scholarly record, combined with its importance in evaluating research, is reflected in a conservative approach to peer review. Despite this, there has been a significant degree of innovation during this century, aided by the potential of the Web to support new approaches. These can be divided into attempts to improve the prepublication review process, and those that seek to add a layer of review after publication.

Starting with prepublication changes, the change with the biggest impact to date is probably the 'soundness not significance' approach. Reviewers are asked to assess the technical soundness of submitted papers, but not to judge them by the allegedly more subjective criteria of impact or significance. The approach is frequently associated with open access (OA) journals.

'Cascade' review allows the reviewers' comments (with the author's consent) to follow a manuscript when it is rejected by one journal and submitted to another. It plays some part within publishers but attempts to implement it across publishers (such as the Neuroscience Peer Review Consortium) have had limited impact to date, though new initiatives continue to emerge.

A more radical change would be to separate peer review from individual journals entirely, with the review being carried out separately and reviewer reports being attached to manuscripts for the benefit of any journals to which the author chose to allow access. At least two organizations had begun to offer such services as of mid-2013.

The other new approach to peer review is to conduct assessment of research articles after publication rather than before.

In its most radical form, a system of postpublication review would see papers published without prior review (beyond minimal screening), with various kinds of appraisal, rating, and review, both formal and informal, taking place subsequently, and combined with metrics reflecting usage and impact.

Proponents argue that filtering prior to publication is ineffective, and is costly and time consuming.

A more pragmatic approach to postpublication review sees it complementing conventional review, rather than replacing it.

Publishing Ethics

While misconduct in research and scholarly communication remains unusual, it does happen and may be more common that most would like to believe. We can distinguish between misconduct in research and in publishing, though there are some overlaps. Research misconduct at its most extreme includes fabrication and falsification of data, as well as omitting or discarding data that does not support the authors' research, overoking use of flawed data, or not maintaining adequate records.

Publishing misconduct includes plagiarism, not citing others' work, not presenting work that contradicts their own, publishing the same data in multiple articles or multiple submissions, giving inappropriate author credits ('guest authorship'), not crediting an author ('ghost authorship'), withholding details of methodology, and undeclared conflicts of interest.

Growth in Article Retractions

The sharp rise in article retractions over the past decade is well documented. Data from the PubMed biomedical literature database show that retractions have grown both in absolute terms and as a proportion of articles published.

It cannot be determined from these data whether incidence of misconduct is increasing, whether the misconduct is now more likely to be detected, whether ethical standards have risen, or some combination of the three. The time between publication and retraction has fallen, which is consistent with improved detection, but does not rule out increased misconduct.

Committee on Publication Ethics

A number of organizations provide support and resources to journals and publishers in dealing with ethical issues, notably the Committee on Publication Ethics. The World Association of Medical Editors and International Committee of Medical Journal Editors also provide codes of conduct and other guidance.

The CrossCheck service brings a more technological approach to bear. It allows journals to screen submissions for potential plagiarism by using software to compare their content to a database of scholarly articles.

Using the Journal

Authors and Readers

It used to be said that the average scholarly article was only read by a very small number of people. This appears to have been a misunderstanding (King et al., 2006), but whatever the truth in the past, the usage data of electronic publishing have given the lie to this perception.

Some fields of study are of course small and specialized, but the overall audience for scholarly work can be large. There are estimated to be 6.5–8.5 million researchers worldwide (Elsevier, 2011; Royal Society, 2011). Their numbers are determined primarily by levels of investment in research and development, and are growing by some 4–5% per year; much of this growth arises in the emerging economies.

There is a distinction to be made between the core active researcher segment and the wider journal-reading community, which is likely to be much larger. We can think of a central core, where most readers are also active researchers, surrounded by an intermediate layer where users are regular readers but only occasional authors, in turn wrapped in an outer layer of occasional readers.

This readership profile will vary by discipline: in fields like pure mathematics or theoretical physics there may be little readership outside the core of active researchers, but in a field such as nursing or medicine, practitioner readers will be many times more numerous than authors.

Many of these additional readers may also be more peripheral and infrequent readers, including postgraduate and undergraduate students. There appears to be no robust evidence sizing this wider journal reader community but publishing industry estimates suggest a total global journal readership of around 10–15 million.

Scientific journal articles are written primarily by academics. For instance, Tenopir and King report that although only 10–20% of the scientists in the United States are employed in universities, they account for about 75% of articles published (King and Tenopir, 2004).

Reading Patterns

We know that journals are widely used within academia. Industry estimates put global annual downloads of scholarly articles at well over 1 billion (Ware and Mabe, 2012), and one study found that each UK registered academic library user downloaded a total of 47 articles per year (RIN, 2011b). The distribution of downloads roughly follows a power law, with a few highly accessed articles and a 'long tail' of infrequently accessed work.

Of course article downloads measure something different from actual reading. There is a long history of reading studies, however, and from these we know, for example, that the average number of articles read annually by academics increased from around 150 in the 1970s to over 270 in the late 2000s, while the average time spent reading each has declined (Tenopir et al., 2011). These pressures have led to what has been dubbed 'strategic reading' (Renear and Palmer, 2009), and to speculation that new software tools may in the future undertake at least part of this work on behalf of (human) readers.

Citations and the Impact Factor

The most tangible evidence for most authors of article readership and impact is a citation elsewhere in the peer-reviewed literature. Citations are central to scholarly discourse, helping the author build their arguments by reference to earlier work, and helping readers by pointing them to related work.

Citations have grown faster than the number of articles. The factors explaining this include the cumulative growth in the literature, more permissive journal policies allowing longer reference lists, and growth in coauthorship. There are big differences in average citation rates between disciplines, with most of the variation explained by differences in coauthorship rates.

Of course, citations play a central role in bibliometrics, notably in the Journal Impact Factor (JIF). The JIF, a proprietary figure produced by Thomson Reuters, measures the average number of times an article in a journal has been cited in a particular period. It was originally introduced to aid librarians in the selection of journals but has subsequently acquired a greater significance as a measure of the 'quality' of a journal. This focus on a single metric has inevitably led to 'gaming' strategies by editors and publishers.

The use of the JIF as a proxy for the quality of research outputs of individuals or departments is a particularly egregious misuse of statistics, since the distribution of citations is highly skewed, and JIFs depend on the research field and on technical factors unrelated to journal quality (Seglen, 1997).

Criticism of the overuse and misuse of the impact factor has led to a growth in the number of available metrics (for example, the *h*-index) and in the field of 'altmetrics.' The latter seeks to obtain a broader measure of article impact by combining multiple metrics and usage indicators.

Authors' Behavior, Perceptions, and Attitudes

While academics may be radical in thought, they are conservative in deed, at least when it comes to scholarly communication. Harley et al. (2010) found researchers remained focused on conventional formal publication, and cautious about new models of Web-based scholarly communication, a picture confirmed in other studies (e.g., RIN, 2010).

Despite a widespread transition to use of electronic journals over the past two decades, the core motivations of authors appear remarkably constant, in terms of a need for attribution and recognition, quality control (including peer review), and visibility. Securing funding and furthering their careers remain two key motivators for authors, albeit generally given as secondary to disseminating their findings. The factors used by authors to select journals for publication are also stable: perceived quality, relevance, and speed of publication remaining the most important, and support for peer review remains widespread.

We should also be cautious of the idea of the 'digital native,' a younger generation of scholars that grew up with Facebook and is alleged to lead the way in adoption of new digital media. Younger scholars showed similar patterns of behavior in these studies in their professional activities as their elders, whatever their nonwork behaviors, and in fact, well-established researchers appeared the most likely to lead adoption of new tools and media.

Disciplinary Differences

Journals vary by discipline. One obvious difference is size, with science and technology journals in the Thomson Reuters index more than three times the size of those in the humanities and social sciences. These differences reflect those in the underlying disciplines: researchers in science, technology, engineering, and mathematics (STEM) disciplines produce the most articles, followed by those in social sciences, and then in the humanities. Joint authorship also varies, with the highest levels again in STEM disciplines, especially biomedicine. The relative importance given to journal articles compared to other types of research output such as monographs is also highest in STEM fields, while a few fields favor (peer-reviewed) conference proceedings. There are also clear differences in attitudes to peer review, in reading and usage patterns, in citation behaviors, in the half-lives of cited and downloaded materials, in the importance of the certification function (as reflected in the willingness to use and formally cite un-peer-reviewed materials), and more recently in attitudes toward OA.

Long-Term Preservation

The continued use of today's journal content into the future depends on its long-term preservation. History has shown that acid-free paper, stored under the right conditions, can be preserved for centuries or longer. We lack the same experience with electronic media, but confidence in their longevity is important for the acceptance of electronic-only journals. New infrastructure is currently under development for the preservation of electronic media, including distributed storage arrangements that mimic the multiple print copies held by libraries, and digital archives maintained by national libraries. Challenges remain, however, in developing some aspects of the necessary technologies, and in the funding requirements.

The Journal in the Global Marketplace

Journal and Articles Numbers and Trends

There are a lot of journals: Ulrich's Directory records well over 28 000 peer-reviewed academic journals, collectively publishing some 1.8–1.9 million articles each year. The archive is correspondingly large and increasingly available online: at the end of 2012, the CrossRef database contained over 46 million digital object identifiers (DOIs) referencing journal articles.

An important subset is the 10 800 journals included in Thomson Reuter's Journal Citation Reports, of which 8400 are in the Science Edition and 3000 in the Social Sciences Edition: these collectively publish about 1 million articles annually. These journals are on average substantially larger than those not included, but their importance primarily arises because they are the most cited. The other major database, Scopus, is intentionally broader in scope, and covers about 20 000 peer-reviewed journals.

The number of peer-reviewed journals published annually has been growing at a steady rate of about 3.5% per year for over three centuries, with some acceleration in the postwar period 1944–78. The number of articles has also been growing by about 3% per year over similar timescales. This growth in the literature is directly correlated with (and likely driven by) the number of researchers in the world, which in turn is driven by the levels of research and development (R&D) funding.

Global Trends in Scientific Output

Global spending on R&D has consistently outpaced global gross domestic product. This trend seems likely to continue as governments seek competitive advantage from their skills and knowledge bases, and see R&D as critical to innovation and economic growth.

Consequently, we should not be surprised to see the fastest growth in journal article output in the large emerging economies that have prioritized R&D investment. China has overtaken the United Kingdom to publish the second largest annual number of research papers, with its share now at over 10%, and is set to overtake the United States well before 2020.

Journal Market Size and Economics

Journals are not just a core part of scholarly communication; they also represent a substantial market valued at about $9–9.5 billion, with books worth a further $3.8 billion. Journal revenues derive mainly from library subscriptions and license fees (68–75% of the total revenue), followed by corporate subscriptions (15–17%), advertising (4%), membership fees and personal subscriptions (3%), and various author-side payments (3%) (Ware and Mabe, 2012).

On the cost side, one of the more reliable estimates of global costs was made by Cambridge Economic Policy Associates (RIN, 2008; updated for RIN, 2011c), giving the average 2010 journal article cost of production (print + electronic) at £3095. This was made up as follows: first copy costs (the costs incurred regardless of the number of copies distributed, e.g., peer-review management, copy-editing, typesetting, and origination): £1261; variable costs (printing, paper, distribution): £581; indirect costs (staff and overheads): £666; and surplus: £586.

Researchers' Access to Journals

The extent to which researchers have adequate access to journal content has been a contested area. There is no doubt that the 'Big Deal' and similar bundling models (in which consortia of libraries purchase online access to additional, previously unsubscribed journals from a publisher at heavily discounted rates) have led to a large expansion of access. This of course came at a cost, including squeezing smaller publishers' share of library budgets, but has documented benefits in terms of sharply increased availability, usage, and decreased cost per download (RIN, 2011b). Many groups that could benefit from access have nevertheless found themselves excluded from such arrangements, leaving access gaps at corporations, small and medium enterprises, independent professionals such as engineers, clinicians not attached to research/teaching hospitals, and others.

Paradoxically, measures of researcher satisfaction and dissatisfaction are both high. For instance, in a 2011 survey, 93% of researchers described their access as easy or very easy, and yet 39% of the same respondents said their top priority was to improve access (RIN, 2011a). The likely reasons for this are that 'easy access' is no longer good enough; researchers can be frustrated by knowing that essentially all articles are available online somewhere; and the improvements in search engines and other discovery tools means that relevant research is ever more visible.

One special case is access in developing countries. Research-4Life, a collaboration between publishers and various United Nations agencies, provides access either free of charge (for the poorest countries) or at very low rates.

Open Access

OA refers to the making available of content (especially, though not exclusively, journal research articles) in online digital copies, free of charge, and free of most copyright and licensing restrictions, and free of technical or other barriers to access (such as digital rights management or requirements to register to access).

The combination of electronic delivery, bundling models, and philanthropic schemes, has significantly improved access to scholarly content. This has by no means reduced pressures for OA, however, and OA journals are by far the fastest growing section of the market.

Gold and Green Open Access

There are two broad approaches for journal articles: 'Gold' OA is a publishing model in which journals make the final published versions of articles freely available immediately on publication. By contrast, 'Green' OA is a sharing or distribution model in which a version of the article is archived in, and made available from, an open repository, either immediately on publication or following an embargo period.

The most common business model (by numbers of articles) for Gold OA publishing is the article processing charge (APC). Green OA does not have a self-sustaining business model itself, but assumes journals' business models will remain unaffected by the free parallel distribution (or will flip to OA).

Other Open Access Variants

In addition to these primary forms, there are two notable variations. First, some journals make their full content freely available on the publisher's platform after an embargo period. Second, most subscription-based journals now offer authors optional OA at the article level (on payment of an APC).

System-Wide Perspectives

If the goal of wider access is widely supported, the details of implementation and the relative benefits of particular approaches and public policies have been heavily contested.

The main drivers of OA have been the interventions and policies of research funders and policy makers, together with the growth and maturing of the OA publishing sector itself.

Research funder initiatives have probably been the most important, including in the UK the Finch Group report and subsequent changes to public funding policies, and the policies and initiatives of the Wellcome Trust (which collectively have tended to give greater weight to the Gold model), and in the United States the 'public access' policies of the National Institutes of Health and (more lately) the White House Office of

Science & Technology Policy (which are more inclined toward the Green route).

Several economic studies have considered the direct and wider social benefits of OA. Whether moving to a Gold OA model in place of (or more likely, in parallel with) subscriptions would result in direct financial savings for academia will depend on the average APC. The wider economic benefits to society of OA to research outputs depend on long-term economic growth models; the approach may be valid but the outputs are sensitive to assumptions that are not easily tested (Houghton et al., 2009; RIN, 2011c; Finch Working Group, 2012).

Open Access Impacts on Use and Citation

Another contested area has been whether making research articles OA leads to increased citations. All work has confirmed a marked increase in downloads of OA materials, but studies differ on whether or not this is accompanied by increased citations (e.g., see Davis and Walters, 2011). Although the common-sense hypothesis suggests more views would lead to more citations, the best available evidence at this point suggests that OA articles do not receive more lifetime citations, but they do get them sooner.

Copyright Issues

A growing focus is on the rights to reuse content, not just to access it. This has led to a debate over the most appropriate copyright licensing regime for OA articles: the most liberal, the Creative Commons Attribution license (CC-BY) is widely endorsed by OA advocates, especially in the sciences, but many publishers and indeed authors and editors prefer more restrictive Creative Commons licenses.

New Developments in Scholarly Communication

Journal Publishing Platforms

There is no doubt that technology has had an impact on the journal and has the potential radically to reimagine it in the future (for instance, by separating its core functions), but to date the electronic journal remains remarkably familiar to users of print.

Apart from the increased efficiency of dissemination and of searching (and the large behind-the-scenes efficiencies in production processes), one could argue that one of the most important early advantages of e-journal platforms was the hyperlinked citation and 'cited-by' links, made possible by the development of the DOI and the infrastructure that sits behind it. Looking forward, we see journal platforms becoming more interoperable in ways that extend this simple idea, for instance, using semantic technologies and open Application Programming Interfaces (APIs).

New Forms of Electronic Articles

There is interest in exploring new forms of journal article. These range from attempts to improve the on-screen usability and functionality to more radical changes such as the recent introduction of data papers, and more radically the proposed nanopublications.

Data-Intensive Science

One catalyst for change will surely be the growth of data-intensive science. The explosion of data presents challenges, but also creates opportunities to increase the efficiency of research and its reproducibility through 'intelligent' data sharing (Royal Society, 2012). Journals are thus now developing policies with respect to data sharing. A new article type, the 'data paper,' is also emerging with the launch of data journals.

Semantic Web and Linked Data

Closely associated with data publication is the notion of the semantic Web. Present journals (at least in science, technology, and medicine) make considerable use of semantic technologies to enrich content, but the formal concept of the semantic Web remains complex, expensive, and elusive.

Semantic enrichment is the tagging of content to add value by identifying and linking terms and concepts of interest to a particular domain. The benefits include improved search and discovery, including matching content to user interests and behaviors; enabling new products and services such as semantically grouped collections, and facilitating text and data mining; and productivity gains in journal production and peer review.

Text and Data Mining

Text and data mining, which draw on natural language processing and information extraction to identify patterns and find new knowledge from collections of textual content, offer the potential to transform the way researchers make use of the literature (e.g., Nature, 2012). This potential has yet to be realized, however, and there are some substantial challenges still to be addressed, including access, licensing issues, data normalization, and others.

Social Media

One of the more obvious changes in the way the general public uses the Web has been the dramatic growth of social media. This has not been paralleled in scholarly communication, however, where uptake remains low and researchers stubbornly conservative in their preference for conventional channels. For example, a recent study estimated that only 1 scholar in 40 used Twitter (Priem et al., 2012) and other studies (noted above) put all forms of social media use at about 15% or lower. Use in some areas (including Twitter) is growing fast, however, and it seems inevitable that integration with the journal will gradually become more important. One emerging idea is that of open annotation.

Mobile Access and Apps

Similarly, despite the widespread adoption of smartphone, tablets, and other mobile devices, accesses by mobile devices

currently represent well under 10% of traffic at most electronic journal platforms. On current trends, increasing mobile access seems unlikely to change the nature of the journal much (in contrast to some of the other trends discussed in this section), but an important impact may be to accelerate a shift of personal reading from print to screen.

Identity and Trust

Two related concepts central to the value of the journal are those of identity and trust. In peer review, reputation is critical to the editor selecting the reviewers and evaluating their reports, and the same is as true for social media recommendations or open annotation.

An important problem has therefore been the disambiguation of researchers with similar or identical names, or with inconsistent transliteration of names from non-Roman alphabets, and the tracking of researchers across name changes. While there are proprietary approaches, a better system is now emerging in the form of Open Researcher and Contributor ID. This initiative provides a central registry (which permits researchers to create and manage their own records), and a Web platform to support system-to-system communication and authentication.

In Conclusion

The journal has thus proved a remarkably resilient format, little changed in its core functions since its introduction in the seventeenth century. Despite the professionalization of peer review in response to the rapid expansion of science following World War II, the commercialization of the sector in the decades that followed, and more recently the transition to the Web, the core functions of the journal – registration, certification, dissemination, archiving, and rewarding – remain intact, for now at least. Its central importance to scholarly communication has probably never been higher, and its role in research assessment and reward is still growing.

See also: E-Communications Platforms and E-Learning; Libraries; Publishing: Academic and University Presses; Research Funding: Ethical Aspects; Scientometrics.

Bibliography

Davis, P.M., Walters, W.H., 2011. The impact of free access to the scientific literature: a review of recent research. Journal of the Medical Library Association 99 (3), 208–217. http://dx.doi.org/10.3163/1536-5050.99.3.008.

Elsevier, 2011. International Comparative Performance of the UK Research Base, Report for UK Department of Business, Innovation and Skills. Retrieved from: http://www.bis.gov.uk/assets/biscore/science/docs/i/11-p123-international-comparative-performance-uk-research-base-2011.pdf.

Finch Working Group, 2012. Accessibility, Sustainability, Excellence: How to Expand Access to Research Publication. Retrieved from: http://www.researchinfonet.org/wp-content/uploads/2012/06/Finch-Group-report-FINAL-VERSION.pdf.

Harley, D., et al., 2010. Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines. Retrieved from escholarship.org: http://escholarship.org/uc/item/15x7385g.

Houghton, J., et al., 2009. Economic Implications of Alternative Scholarly Publishing Models. JISC. Retrieved from: http://www.jisc.ac.uk/publications/reports/2009/economicpublishingmodelsfinalreport.aspx.

King, D.W., Tenopir, C., 2004. An evidence based assessment of author pays. Nature Focus on Access to the Literature. Retrieved from: http://www.nature.com/nature/focus/accessdebate/26.html.

King, D.W., Tenopir, C., Clarke, M., 2006. Measuring total reading of journal articles. D-lib Magazine 12 (10). http://dx.doi.org/10.1045/october2006-king.

Nature, 2012. Gold in the text? (Editorial). Nature 483 (7388), 124. http://dx.doi.org/10.1038/483124a.

Priem, J., Costell, K., Dzuba, T., 2012. Prevalence and Use of Twitter among Scholars. Retrieved from: http://figshare.com/articles/Prevalence_and_use_of_Twitter_among_scholars/104629.

Renear, A.H., Palmer, C.L., 2009. Strategic reading, ontologies, and the future of scientific publishing. Science 325 (5942), 828. http://dx.doi.org/10.1126/science.1157784.

RIN, 2008. Activities, Costs and Funding Flows in the Scholarly Communications System in the UK. Research Information Network. Retrieved from: http://rinarchive.jisc-collections.ac.uk/our-work/communicating-and-disseminating-research/activities-costs-and-funding-flows-scholarly-commu.

RIN, 2010. If You Build It, Will They Come? How Researchers Perceive and Use Web 2.0. Research Information Network, London. Retrieved from: http://rinarchive.jisc-collections.ac.uk/our-work/communicating-and-disseminating-research/use-and-relevance-web-20-researchers.

RIN, 2011a. Access to Scholarly Content: Gaps and Barriers. Retrieved from: http://rinarchive.jisc-collections.ac.uk/node/1172.

RIN, 2011b. E-Journals: Their Use, Value and Impact – Final Report. Retrieved from: http://rinarchive.jisc-collections.ac.uk/our-work/communicating-and-disseminating-research/e-journals-their-use-value-and-impact.

RIN, 2011c. Heading for the Open Road: Costs and Benefits of Transitions in Scholarly Communications. Retrieved from: http://www.rin.ac.uk/system/files/attachments/Dynamics_of_transition_report_for_screen.pdf.

Royal Society, 2011. Knowledge, Networks and Nations: Global Scientific Collaboration in the 21st Century. The Royal Society.

Royal Society, 2012. Science as an Open Enterprise. Retrieved from: http://royalsociety.org/policy/projects/science-public-enterprise/report/.

Seglen, P.O., 1997. Why the impact factor of journals should not be used for evaluating research. British Medical Journal (Clinical Research Ed.) 314 (7079), 497. http://dx.doi.org/10.1136/bmj.314.7079.497.

Tenopir, C., Mays, R., Wu, L., 2011. Journal article growth and reading patterns. New Review of Information Networking 16 (1), 4–22. http://dx.doi.org/10.1080/13614576.2011.566796.

Ware, M., 2011. Peer review: recent experience and future directions. New Review of Information Networking 16 (1), 23–53. http://dx.doi.org/10.1080/13614576.2011.566812.

Ware, M., Mabe, M., 2012. The STM Report: An Overview of Scientific and Scholarly Journal Publishing, third ed. International Association of Scientific, Technical and Medical Publishers. Retrieved from: http://www.stm-assoc.org/2012_12_11_STM_Report_2012.pdf.

Relevant Websites

http://www.alpsp.org/Ebusiness/Home.aspx – Association of Learned and Professional Society Publishers.
http://publicationethics.org – Committee on Publication Ethics.
http://www.doaj.org – Directory of Open Access Journals.
http://www.stm-assoc.org – International Association of Scientific, Technical & Medical Publishers.
http://www.ingentaconnect.com/content/alpsp/lp – Learned Publishing (Journal).
http://www.mdpi.com/journal/publications – Publications (Journal).
http://www.scopus.com – Sciverse Scopus.
http://www.sspnet.org – Society for Scholarly Publishing.
http://wokinfo.com – Web of Knowledge.

Judaism

Stephen Sharot, Ben-Gurion University of the Negev, Beer-Sheva, Israel

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by J. Neusner, volume 12, pp. 8004–8010, © 2001, Elsevier Ltd.

Abstract

From a religion that focused on sacrifices in the Temple of Jerusalem, Judaism emerged as a religion that required the implementation of religious laws by all Jews in their daily life. New patterns of Jewish identity in the nineteenth century involved a decline in religious observance and modifications in the synagogue services that represented accommodations to non-Jewish society. The patterns of religious observance today in the two largest Jewish communities, those of the United States and Israel, are broadly similar, but they differ in their religious affiliations and identities. Ultra-Orthodox Jews, whose numbers have increased considerably in recent decades, present a strong contrast with the secularized pattern of the majority.

From Israelite Religion to Rabbinical Judaism

Max Weber's *Ancient Judaism* (1917–19/1952) was the first comprehensive sociological analysis of religious changes in Ancient Israel. Weber recognized both division and continuity between the 'Israelite religion' of the period prior to the destruction of the First Temple in 586 BCE and the 'Judaism' that emerged from the period of the Babylonian Exile, the return of some of the exiles after 538 BCE, and the period of the Second Temple (516 BCE–CE 70). A genuine monotheism emerged only in the later period, but Weber traced its origins to a period when the Israelites constituted a confederacy of seminomadic stock breeders and warriors. The tribes and kin groups that made up the confederacy had, at first, their own separate gods, but as the confederacy became more powerful it became increasingly focused on one god. Weber emphasized the importance of the Covenant in this process as it both signified and promoted the uniqueness of the relation of Israel to its God. It was constituted of mutual promises: the people of Israel promised to obey God's law, and in return, Yahweh, who had already liberated them from slavery in Egypt, would grant them domination over the Promised Land and protect them from their enemies.

The gradual crystallization of the monotheistic idea continued during the period of centralized monarchy from the eleventh century and was advanced considerably by the new type of prophecy that emerged by the end of the sixth century BCE. Weber argued that the appeal of the prophets can be understood, in part, against the background of the development of the monarchy into despotic states, the demilitarization of the peasants, and mounting external threats to the Israelite kingdoms. The tradition preserved those prophecies that appeared to have come true, and it was during the Babylonian captivity, when the predictions of doom were seen to have come to pass, that the prestige of prophetic religion reached its height. A broad urban stratum whose lifestyle was not conducive to magic was responsive to ethical prophecy and became the carrier of religious, ethical action. Weber argued that changes in Ancient Judaism represented a significant development in ethical rationalization, but the influence of postexilic Judaism on Western society was limited because it remained a religiosity of law, centered on concrete, discrete norms rather than on abstract principles.

Weber's work influenced biblical scholars, but although a number of sociologists have attempted to systematize Weber's somewhat fragmented analysis, particularly with respect to the development of monotheism and rationalization, very few sociologists have attempted to expand upon Weber's analysis on the basis of subsequent findings of biblical scholars and archaeologists. Controversy remains among biblical scholars with respect to the development of monotheism. Some argue that there was little continuity between the preexilic and postexilic periods, claiming that monotheism emerged totally in the exilic (Babylonian) and postexilic eras, and they minimize or discount any significant preexilic contributions to its emergence. In these accounts, there was little to distinguish preexilic Israelite religion from so-called Canaanite polytheistic religions, and it was the religious elites of Yahwism in exile and in the postexilic period who fictionalized what became the biblical narratives for the preexilic era. Other scholars attribute greater importance to preexilic developments, such as the early shift, possibly from the late tenth century, to Yahwistic 'monolatry' (exclusive worship of a god enforced by prohibitions of allegiance to other gods without denying their existence), but they are also inclined to recognize the Babylonian exile as having an overwhelming impact that ensured the permanency of mono-Yahwism.

The assumption that early Israelite society was a seminomadic, tightly knit tribal league has been challenged by scholars who have proposed that it was primarily an agrarian society, composed mostly of peasants indigenous to Canaan who worshiped a number of deities in addition to Yahweh. The centralization of sacrificial worship in Jerusalem was a slow process before it gained ritual hegemony in the seventh century. The Covenant is now seen by many as established no earlier than the first half of the seventh century BCE, and one view is that it was originally conceived as between Yahweh and the king and later reformulated as between Yahweh and the people. In place of the idea that the roots of monotheism emerged in nomadic or seminomadic societies, scholars have emphasized the royal shrines at Jerusalem and the urban centers of Mesopotamia during the Babylonian exile as the major sites of the emergence of monotheism. There is little doubt that in all of these periods, different strata and groups of the peoples of Israel and Judea practiced a variety of religions (family, local, royal, or elite),

and that a uniform orthodoxy was not achieved (Gnuse, 1997; Gottwald, 2002).

In the most general terms, the transformation of Israelite religion into Judaism can be described as a change from a religion that had come to be focused on sacrifices and other ceremonies performed by the priests in the Temple in Jerusalem to a religion that emphasized the reading and teaching of the Torah, forms of prayer worship in the synagogue, and the implementation of the religious laws by all Jews. Elements that had been part of Temple worship, particularly with respect to food, were accommodated into domestic life, both on a daily basis and on festivals such as Passover. Some of the seeds of this transformation can be traced to the period of the Second Temple, particularly among the Pharisees, who represented one of a variety of 'Judaisms' alongside the Sadducees, Essenes, Christians, Samaritans, and other groups in the first century CE.

The changes in the religion and the identity of the people associated with the religion have prompted a debate among scholars of antiquity over the issue of when it becomes appropriate to use the terms 'Jew' (*Yehudi* in Hebrew) and 'Judaism' (*Yahadut* in Hebrew). At first, *Yehudi* referred specifically to the tribe of Judah (*Yehuda*), as distinguished from the other tribes of Israel, but the name came to designate anyone resident in or originating from the kingdom of Judah. S.J.D. Cohen (1999) argues that a semantic shift from 'Judean' to 'Jew' is justified from the latter part of the second century BCE, when the Greek term for *Yehudim*, *Ioudaioi* (singular, *Ioudaios*), came to be applied to people who were not necessarily geographic or ethnic Judeans but had come to believe in the God of the Judeans. A non-Judean could become a *Ioudaios* by joining the Judeans in worshiping and venerating the God whose temple was in Jerusalem. Critics of Cohen have argued that the semantic change is justified only at a later date. As a translation of the Greek *Ioudaismos*, 'Judaism' in antiquity has been understood to denote the religion and/or culture of Judeans or Jews, but there are very few uses of the term in Second Temple sources. 'Judaism' came to be commonly used in gentile Christian literature, but among Jews its use remained rare until modern times.

The destruction of the Temple in CE 70 and the crushing of the revolt in CE 132–135 proved to be decisive events in the transformation of the religion; the institutional base of the priesthood was gone, and Torah scholars emerged as the dominant religious elite. It was the sages of the religious academies of Palestine and Babylonia who produced, over the first six centuries of the Common Era, a huge corpus of religious literature that came to be known as the Talmud. Oral traditions were eventually edited around the year CE 200 into a large corpus known as the Mishna, and the subsequent voluminous commentaries on this literature came to be known as the Gemara. The Palestinian Gemara was completed in about CE 450, but it was the Babylonian Talmud (the Mishna plus the Babylonian Gemara), completed in about CE 600, that came to be accepted by most Jewish communities as authoritative (Cohen, 1987).

The term 'rabbi,' meaning 'my master,' came to be used to designate those scholars who created Talmudic Judaism, which became the dominant form of Judaism from around the turn of the first millennium until the nineteenth century. It took some time for the rabbis to be accepted as religious leaders, and during most of the period of the production of the Talmud, from the second to the seventh centuries CE, the rabbis had little impact on the vast majority of Jews. The rabbis appear to have recruited disciples who were literate, relatively prosperous Jewish males, but little is known of the social processes by which rabbinical Judaism came to be accepted among the communities of the expanding diaspora (Lapin, 2010). One social process that took place among the Jews during those centuries when Rabbinic Judaism was taking hold was their occupational transition from agriculture to urban occupations. It has been proposed that the transformation of Judaism into a religion centered on reading and learning provided the Jews with a competitive advantage in the new urban centers of the Muslim empire, especially during the Abbasid rule from the mid-eighth to the early ninth centuries, and later in the towns of Europe. The shrinkage of the Jewish population, far greater than that of the general population, from the first century until the early eighth century is explained by voluntary conversion, mainly to Christianity, of Jewish illiterate farmers who did not adopt Rabbinic Judaism (Botticini and Eckstein, 2005).

Judaism and Jewish Identity

An explicit distinction between the religious and ethnic elements in Jewish identity was made only in the modern era, but it is possible to trace how these elements came together and received canonical authority in Rabbinic Judaism. The rabbinic texts defined a Jew as a person who belongs to a particular descent group, 'the children of Israel/Jacob,' and the common ancestors were linked to a religious event. The 'people of Israel' had entered into a unique relationship with Yahweh through the covenants with Abraham, Isaac, and Jacob, and that relationship was more sharply defined through the Covenant at Sinai, an event that, according to the Rabbinic tradition, defines Israel (Porton, 1999). The rabbis of the second century CE standardized the conversion process by requiring that all converts accept the commandments of the Torah, that male converts be circumcised, and that all converts should undergo ritual immersion. This standardization provided a group identity with permeable boundaries. It was also in the second century CE that rabbinic law established the matrilineal principle: only the offspring of a Jewish mother were recognized as Jews. This ruling, which was surprising in the context of the patriarchialism of Israelite and Judean laws and culture, together with the prohibition on intermarriage, expressed the rabbinic concern for genealogical purity (Cohen, 1999).

The fusion of religion and ethnicity formulated by the rabbis came apart in the modern era when rabbinic hegemony came to an end and Jews were freed from the segregation that had been imposed upon them by non-Jews. Toward the end of the eighteenth century and in the nineteenth century, new patterns of Jewish identity developed in Europe and North America in response to secularization and the construction of new collective identities of modernizing nations. The construction of boundaries of citizenship and identity involved formulating terms for the Jews' inclusion or exclusion. Where secularization was accompanied by willingness to

include the Jews – often made conditional upon the Jews' adoption of the dominant culture and national identity – many reinterpreted their Jewish identity as a purely religious one and eschewed any Jewish national, or what came later to be known as Zionist, identity. Proclaiming that they differed from their non-Jewish compatriots only in terms of religious persuasion, the English people, French people, Germans, and Americans of the Jewish or Mosaic faith developed forms of Judaism (Reform, 'modern' or 'decorous' Orthodox, and later, in the United States, Conservative Judaism) that they believed would allow them both to participate in the wider society and to retain a religious identity as Jews. The religious observance of a large proportion of Western and Central European and American Jews in the nineteenth century came to be limited to the rites of passage, infrequent visits to the synagogue, and a few yearly observances in the family setting.

The forms and degrees of changes made within synagogue services varied among the secularized Western communities in the nineteenth century; Reform Judaism became important in German and American communities, whereas modified forms of Orthodoxy were the most popular in the communities of England, France, and Italy. These differences were related to the influences of the different forms of Christianity in the environment of acculturated Western Jews (Catholicism in France and Italy, Anglicanism in England, and liberal Protestantism in parts of Germany and the United States), and the extent to which economically mobile Jews were socially accepted by non-Jews in the middle and upper classes. One aspect of religious accommodation was common to all the secularized communities: the concern with decorum. Acculturated Jews disliked the outward emotionalism and demonstrativeness of the traditional service – the loud chanting of prayers, the body movements, the coming and going during the service, and the inappropriate business of selling religious honors, with congregants making material offerings when called upon to recite benedictions before and after readings of the Torah. The decorous Orthodoxy in, for example, the English communities provided rigidly patterned services with a passive audience of worshipers listening to the cantor, a choir, and regular sermons in the vernacular by the rabbi or 'minister.' The Reform movement in Central Europe and North America made far more extensive changes in the synagogue or 'temple' services; in addition to mixed choirs accompanied by an organ, mixed seating, and sermons in the vernacular, the liturgy was considerably abbreviated and many prayers were read in the vernacular. In the United States, a Conservative movement seeking a middle path between Orthodoxy and Reform emerged at the end of the nineteenth century. The Conservative movement attracted the children of immigrants from Eastern Europe and grew rapidly after World War II; its services had more Hebrew and were more traditional than Reform ones, but, like Reform and in contrast to Orthodoxy, mixed seating was allowed.

In Eastern Europe, where the construction of national and ethnic identities involved marking the Jews as unassimilable outsiders, the inroads of secularism and the breakup of traditional communities toward the end of the nineteenth century led many Jews to adopt a nonreligious national, ethnic, or cultural Jewish identity. This type was, in some respects, the converse of the type that had become common in Central and Western Europe and the United States. A greater proportion of secularized Eastern European Jews believed that their problems required radical solutions (a socialist revolution or an autonomous Jewish state), and religion was seen as an irrelevance or obstacle in the attainment of such solutions (Sharot, 1976; Goldscheider and Zuckerman, 1984).

The ideologically secularist Jewish movements attracted only a minority of Eastern European Jews, but a number of the secularists joined the mass migration to the West beginning from the 1880s and transplanted their socialist, nationalist, and Yiddish cultural movements to those communities. In the United States, the secularist movements of these Eastern European immigrants continued to find support among the second generation in the years between the two world wars; secularist ethnic identities remained the preference of a minority, but antireligious Zionists, secularist Yiddish culturists, and a number of anarchist, socialist, and communist Jewish organizations could still be found (Glazer, 1957).

Secularist nationalists, especially socialist Zionists from Eastern Europe, were prominent in the first modern waves of Jewish migration to Palestine, and the construction of the national identity of 'New Hebrews' or 'Israelis' involved attempts to make a clear differentiation between nationality and religion. The emphasis on nationality rather than religion was signaled by substituting the 'people of Israel' for 'God' when drawing upon biblical and traditional literature. The dissociation from Judaism in formulations such as 'Canaanites' was intended to contrast between the 'Hebrew' and the 'Jew,' the latter being associated with the negatively evaluated condition of exile.

Both the religious antinational or nonnational Jewish identity espoused by Reform Judaism, and the national, antireligious, or nonreligious Jewish identity of secularist Zionism, were expressions of an acute preoccupation with the boundaries of identity and were cast in highly ideological forms. They were examples of modernist movements in their rejection of the traditionalist religious societies, their conceptions of progress, and their reinterpretations of the Jewish past to legitimize their endeavors and visions of the future. Reform Jews found their values of social justice in the teachings of the ancient Jewish prophets, and secularist Zionists pointed to the Hebrew kingdoms of antiquity as expressions of the primary ties between the people and the land, and the importance of political autonomy. Finding an appropriate link with a distant past, and unburdening themselves of the deadweight of tradition, both Reform Judaism and secularist Zionism presented themselves as carriers of a mission that would bring further enlightenment and freedom in the future (Sharot, 1982).

A number of events and social processes have resulted in the decline of the modernist, ideological forms of the denationalized religious Jewish identity and the dereligionized national Jewish identity. The exclusively religious identity that was given ideological and ritual expression in classical Reform Judaism lost its attraction in the face of vehement racial anti-Semitism, the Holocaust, the achievements of the Zionist movement, the establishment of the state of Israel, and the wars between Israel and Arab countries. The second generation of American Jews tended to be drawn to Conservative Judaism,

which combined religion and ethnicity and provided, in its ritual forms, an attractive middle road between the old-society Judaism of their parents and the Reform Judaism of the more established American Jews. The third and subsequent generations have shifted more to Reform – but this brand of Reform no longer attempts to dispute the national component of Jewish identity, and places greater emphasis on Israel and Zion (Sklare, 1955; Sharot, 1976, 2011).

The Six-Day War accelerated the decline of both the nonnational religious identity among Jews in the Diaspora and the antireligious national identity in Israel. The American Council for Judaism, a Reform offshoot that represented the anti-Zionist position of classical Reform Judaism, virtually collapsed, and in Israel ideological secularism was marginalized by the interpenetration of politics and religion that followed the war (Liebman and Don-Yehiya, 1983). In addition to the Six-Day War and its repercussions, two longer term processes contributed to the decline of a secularist or antireligious Jewish ethnicity, both in the Diaspora and in Israel. One was secularization. Secularist Jewishness had its high point at a time when the religious worldview was beginning to lose its shared, taken-for-granted qualities, but religion still retained a strong influence over a large part of the community. Once the pervasive influence of religion came to be confined to a minority enclave (or a number of enclaves) that rarely impinged on the daily lives of the majority of Jews, secularist Jewishness lost its appeal and was replaced by minimal religiosity. Ideological secularism may appear to have retained some importance in Israel, but it tends to take the form of opposition to the religious establishment and to the political power of the religious sectors, especially the ultra-Orthodox or *haredim*, rather than to religion per se.

Another development that contributed to the decline of the antireligious or nonreligious Jewish identity, especially in the Diaspora, was the decline of ethnic cultural elements (especially the Yiddish language), which in the past provided a cultural basis for nonreligious ethnicity, and the decline of communal and social structures, such as ethnic neighborhoods and industries where Jews provided the majority of both employers and employees. The destruction of Yiddish culture in Eastern Europe and the decline of its remnants in the West undercut the foundations of a Jewish ethnic identity that could draw upon rich cultural resources with little regard to, or in opposition to, religious heritage. Where an all-encompassing religion has lost its strength, and where nonreligious ethnic cultural sources are no longer available, instead of rejecting religious symbols, secularized Jews tend to reinterpret them, extracting their supernaturalist meanings and making them suitable for familial and ethnic celebrations (Sharot, 2011).

Secularization and Fundamentalism

In comparison with the traditional community, in which the great majority observed a comprehensive system of *mitzvot* (religious commandments) that guided their everyday life, Jewish religious observance today is highly selective and fragmented. Rituals that were observed by the great majority in the past, such as following the many dietary laws or the donning of *tephillin* (phylacteries) by males during weekday morning prayers, are observed today only by small minorities. With respect to religious beliefs, a 2008 survey of the American population found that the proportion of Jews who affirm such beliefs as absolute certainty in the existence of God and life after death is considerably lower than that of Christians. There is little difference between American Jews and Christians with respect to joining religious congregations and their social participation in those congregations. However, the proportion of Jews who attend religious services frequently is far lower than that of Christians. About one-half of American Jews attend synagogue on a few occasions, on the high holidays (Yom Kippur and Rosh Hashanah) or for weddings and bar or bat mitzvahs.

The historical identity of Judaism and the Jewish people has meant that a predominantly ethnic Jewish identity can be expressed only through religious symbols and practices, but the symbols and practices that were chosen by secularized Jews are those that relate the individual and family to the themes of the continuation of the Jewish people and its historical emergence and struggles for survival. The rites of passage – circumcision, bar or bat mitzvah, and religious marriage and burial – can be interpreted in this way, and they are observed by the great majority who also signal their Jewish identity by placing the mezuzah (a piece of parchment inscribed with a specific set of verses from Deuteronomy, covered by a protective case) on the doorposts of their homes. The two most popular ritual occasions in all Western communities and Israel are the Seder meal at Pesach (Passover) and the lighting of candles during Hanukkah. The Passover meal is a family occasion, often of the extended family, that celebrates the deliverance from Egypt. More than two-thirds of American Jews and 9 out of 10 Israeli Jews participate in the Passover Seder. Hanukkah commemorates the successful struggle against Syrian-Greek rule and the rededication of the desecrated Second Temple. More than 80% of Jews in both the United States and Israel reported that they lit candles on the nights of Hanukkah. Whereas in the United States Hanukkah, which is close to Christmas, has become the major children's festival, in Israel other ritual occasions (Purim and Lag b'Omer) have become major celebrations that focus on children (with fancy costumes and bonfires, respectively).

Another yearly ritual, fasting on Yom Kippur, is observed by more than one-half of the American Jewish population and by more than two-thirds of the Israeli Jewish population. It is somewhat misleading to compare the American and Israeli populations because Israeli Jews include a large proportion of immigrants and their children from the more traditional communities of North Africa and Asia. A more appropriate comparison is with Israeli-born Jews of European origins, of whom about one-half report that they fast on Yom Kippur. With respect to less popular practices, such as Sabbath observance and separate dishes for milk and meat, the levels of observance are higher in Israel, even when the comparison is confined to Israelis of European origins.

The lower levels of religious observance among American Jews are related to their greater emphasis on voluntarism or choice in the observance of religious customs. American Jews live in a society of religious voluntarism and have adopted the

market system of denominations; among their religious options, they are free to choose to affiliate (or not). The emphasis in Israel is on the protection of Jewish interests by centralized institutions, which include an Orthodox establishment, and the consequence of the almost exclusive state support for the Orthodox is that the Reform and Conservative movements are disadvantaged in competing for the allegiance of consumers of religious services.

American and Israeli Jews attribute different meanings to the term 'religious Jew.' For Israeli Jews, the meaning of *dati* (pl. *datiim*, 'religious') refers to a Jew whom American Jews call Orthodox: a Jews who makes the *Halacha*, the system of religious laws, the basis of his or her way of life. Many American Jews believe that there are different ways of being a religious Jew, and the term can be applied to Reform, Conservative, or even nonaffiliated Jews who may observe only a few religious customs. In recent years in the United States, there has been a growth of those who identify with Reform Judaism or who express no denominational preference, on the one hand, and a growth of those who identify with Orthodoxy, on the other. The growth of the Reform and Orthodox sectors has been at the expense of Conservative Judaism, but the differences among the denominations are viewed as artificial and irrelevant by many lay American Jews who seek a Jewish spirituality that transcends the denominations (Liebman and Cohen, 1990; Cohen and Eisen, 2000; United Jewish Communities, 2003; Levy et al., 2002).

In comparison with surveys of the religious preferences of American Jews, which give denominational options and a 'Just Jewish' option, surveys of Israeli Jews provide the options of *haredi* (ultra-Orthodox), *dati* (religious), *mesoriti* (traditional), and *hiloni* (secular). *Haredi* (pl. *haredim*) is a Hebrew epithet meaning a God-fearing devotee. The term was used in the past to denote any Jew who was punctilious about religion, but in the last 50 years it has been increasingly used to designate those Jews who go beyond Orthodoxy or, a more appropriate term in the context of Judaism, Orthopraxy. *Mesoritim* (pl. of *mesoriti*) accept the Orthodox interpretation of Jewish law as the authentic Judaism, but they admit that they do not observe it fully and are flexible in their observance. They observe those parts of the religious law that they believe are essential to their religious identity. For example, certain *Shabbat* (Sabbath) and *kashrut* (dietary) observances, largely focused on the family, are viewed as essential, whereas regular synagogue attendance is not (Yadgar, 2011).

The most recent comprehensive survey of the religiosity of Israeli Jews, conducted in 2009, found that 8% identified as *haredi* (up from 5% in the 1999 survey), 15% as *dati* (11% in 1999), 33% as *mesoriti* (32% in 1999), and 46% as *hiloni* (down from 52% in 1999). These identities correlate with reported levels of religious observance: *haredim* and *datiim* report that they observe the religious law 'meticulously' or 'to a great extent,' while the *mesoritim* are divided between those who observe 'to a great extent' and those who observe 'to some extent.' Most of the secular who say that they are not antireligious (43% of the total sample) observe the religious tradition 'to some extent,' while most of the secular antireligious (6% of the total sample) say that they 'do not observe the tradition at all.' Far more *Mizrachim*, Israeli Jews of North African and Asian descent, define themselves as *mesoriti* (44% in the 2009 survey) compared with Ashkenazim, Jews of European descent (18%), or those of mixed *Mizrachi*-Ashkenazi parentage (23%). Ashkenazim and those of mixed parentage tend to be more polarized in their religious identifications as either *haredi* and religious or secular (Arian and Keissar-Sugarmen, 2012).

The pattern of a minimal religiosity, largely focused on a few yearly festivals, among a large proportion of contemporary Jews has occurred at the same time as an increase in the ultra-Orthoprax or, as they are called in Israel, the *haredim*. Following World War II, the future of ultra-Orthoprax Judaism had seemed bleak; most of the traditionalist communities of Eastern Europe were destroyed by the Holocaust, and it was expected that the defections to secular society would accelerate in the open societies of the West and Israel. Instead, the last decades of the twentieth century saw a considerable strengthening and growth of the ultra-Orthoprax communities, both in the West and in Israel. By building an enveloping system of socialization and education, separated from the wider society, the *haredim* succeeded in retaining the vast majority of those born within the communities, and, as a consequence of high birth rates, they have grown in numbers. The *haredim* in the United States now represent about one-third of Orthoprax Jewry, and they have spread in the New York area from their established neighborhoods in Brooklyn to *haredi* 'villages' in the suburbs (Heilman, 2006). In Israel, they now constitute about 8% of the Israeli Jewish population and have spread from their centers in Jerusalem and Tel Aviv to an increasing number of neighborhoods and towns.

After the Holocaust, when only a few thousand children were in the *haredi* educational system, the *yeshivot* (higher Talmud learning institutions for males; singular, *yeshiva*) were seen by *haredim* as the principal means by which a Torah-based society could be rebuilt and the youth kept apart from the secular world. Unlike in the former Eastern European communities, the traditional way of life could no longer be taken for granted and had to be constructed on the basis of voluntary communities in which there was a more overt commitment to a strictly religious way of life. Within an encompassing framework, *haredi* males are expected to devote long years to Talmud learning, a way of life that is presented as requiring great effort and devotion and as absolutely essential for the survival of the Jewish people. Removed from the concerns of earning a living and often sheltered from the pressures of families, *yeshiva* students stress stringent interpretations of the religious law. The tradition of written codes has gained in strength at the expense of folk traditions anchored in families and local communities, and younger *haredim* often supersede their fathers in their religious knowledge and strict conformity to the religious laws (Friedman, 1987).

Long years in the *yeshivot* have limited the participation of *haredim* in the workforce, and this is particularly the case in Israel, where government subsidies for the *haredi* way of life are extensive. The traditional rabbinical emphasis on hard work and activism in the world was replaced by interpretations that stressed the role of miraculous events in economic maintenance, the need to abandon profane work in order to worship God, and the devaluation of work as a distraction from the higher calling of study and as an obstacle to salvation. Some *yeshiva* students are now challenging this attitude

toward work and complain that the withdrawal from labor markets has resulted in poverty. This challenge is part of the changes that have been evident in the *haredi* community in Israel since the 1990s. The use of modern technology, especially computers, cellular phones, and DVDs, has given access to a wider knowledge, and various mystical and New Age groups have influenced the community. Some *yeshiva* students wish to be more involved in family life, obtain professional training, join the military, and participate in sociopolitical debates. The religiosity of many *yeshiva* students may be shifting from an otherworldly to a this-worldly piety, but they find it difficult to propose solutions that combine both their religious piety and their material and emotional needs (Stadler, 2009).

The case of Judaism in modern society suggests that the classical thesis of secularization should be modified by attention to polarization between a relatively secularized majority and a neotraditionalist minority. Among the majority of Jews, an often steep decline in religiosity followed the breakup of cohesive local communities, exposure to scientific education and secular beliefs, and an involvement in legal-rational structures. At the same time, the processes of modernization encouraged the countertrend of fundamentalism or neotraditionalism. This was not simply a matter of a rejection of modernity, or important aspects of it, by conservatives or fundamentalists who felt threatened by change. The processes of modernity have themselves provided congenial conditions for the growth and strength of neotraditionalism. In Judaism, it was the reorganization of ultra-Orthopraxy on the basis of voluntaristic communities and reorganized institutions that produced growth, greater militancy, and a radical alternative to the secularized pattern of the majority.

See also: Historical Thought and Historiography: Jewish Tradition; Judaism and Gender; Religion, Sociology of; Religion: Nationalism and Identity; Religiosity; Religious Fundamentalism: Cultural Concerns; Secularization.

Bibliography

Arian, Asher, Keissar-Sugarmen, Ayala, 2012. A Portrait of Israeli Jews: Beliefs, Observance, and Values of Israeli Jews, 2009. Israel Democracy Institute, Jerusalem.

Botticini, Maristella, Eckstein, Zvi, 2005. Jewish occupational selection: education, restrictions, or minorities? The Journal of Economic History 65 (4), 922–948.

Cohen, Shaye J.D., 1987. From the Maccabees to the Mishnah. Westminster John Knox Press, Louisville, Kentucky.

Cohen, Shaye J.D., 1999. The Beginnings of Jewishness: Boundaries, Varieties, Uncertainties. University of California Press, Berkeley.

Cohen, Steven M., Eisen, Arnold M., 2000. The Jew within: Self, Family, and Community in America. Indiana University Press, Bloomington, IN.

Friedman, Menachem, 1987. Life tradition and book tradition in the development of ultraorthodox Judaism. In: Goldberg, Harvey E. (Ed.), Judaism Viewed from Within and from Without. State University of New York Press, Albany.

Glazer, Nathan, 1957. American Judaism. University of Chicago Press, Chicago, IL.

Gnuse, Robert Karl, 1997. No Other Gods: Emergent Monotheism in Israel. Sheffield Academic Press, Sheffield.

Goldscheider, Calvin, Zuckerman, Alan S., 1984. The Transformation of the Jews. University of Chicago Press, Chicago, IL.

Gottwald, Norman K., 2002. Rethinking the origins of ancient Israel. In: Gunn, David M., McNutt, Paula M. (Eds.), 'Imagining' Biblical Worlds: Studies in Spatial, Social and Historical Constructs in Honor of James W. Flanagan. Sheffield Academic Press, London, pp. 190–201.

Heilman, Samuel C., 2006. Sliding to the Right: The Contest for the Future of American Jewish Orthodoxy. University of California Press, Berkeley.

Lapin, Hayim, 2010. The rabbinic movement. In: Baskin, Judith R., Seeskin, Kenneth (Eds.), The Cambridge Guide to Jewish History, Religion, and Culture. Cambridge University Press, New York, pp. 58–84.

Levy, Shlomi, Levinsohn, Hanna, Katz, Elihu, 2002. Beliefs, Maintaining Tradition, and Values of Jews in Israel 2000. Guttman Center, Israeli Institute for Democracy, and Avi Chai, (in Hebrew) Jerusalem.

Liebman, Charles S., Cohen, Steven M., 1990. Two Worlds of Judaism: The Israeli and American Experiences. Yale University Press, New Haven, CT.

Liebman, Charles S., Don-Yehiya, Eliezer, 1983. Civil Religion in Israel: Traditional Judaism and Political Culture in the Jewish State. University of California Press, Berkeley.

Porton, Gary G., 1999. Who was a Jew? In: Neusner, Jacob, Avery-Peck, Alan Jeffery (Eds.), Judaism in Late Antiquity: Where We Stand – Issues and Debates in Ancient Judaism, vol. 2, Brill, Leiden, pp. 197–218.

Sharot, Stephen, 1976. Judaism: A Sociology. David & Charles, Newton Abbot.

Sharot, Stephen, 1982. Messianism, Mysticism, and Magic: A Sociological Analysis of Jewish Religious Movements. University of North Carolina Press, Chapel Hill.

Sharot, Stephen, 2011. Comparative Perspectives on Judaisms and Jewish Identities. Wayne State University Press, Detroit.

Sklare, Marshall, 1955. Conservative Judaism: An American Religious Movement. Free Press, New York.

Stadler, Nurit, 2009. Yeshiva Fundamentalism: Piety, Gender, and Resistance in the Ultra-Orthodox World. New York University Press, New York.

United Jewish Communities, 2003. National Jewish Population Survey 2000–01. United Jewish Communities, New York.

Weber, Max, 1952. Ancient Judaism. Free Press, New York.

Yadgar, Yaacov, 2011. Secularism and Religion in Jewish-Israeli Politics: Traditionalists and Modernity. Routledge, New York.

Judaism and Gender

Laura S Levitt, Temple University, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article offers an overview of contemporary research on Jews, Judaism, and gender. It provides a review of scholarly, religious, cultural, and communal efforts to transform Jewish life in light of feminist insights beginning with the second wave of American feminist activism and scholarship. This article addresses the range and depth of these efforts. It also shows how these efforts have forged for a broader vision Jewish community to include others who have been marginalized especially lesbian, gay, bisexual, transgender, and queer Jews in the twenty-first century, and across various Jewish domains.

Building on Virginia Woolf's suggestion that we add a supplement to history as a way of remedying the inequities between men and women, feminist historian Joan Scott calls our attention to the ambiguities within what she has termed "the logic of the supplement." As Scott explains, "women are both added to history and they occasion its rewriting, they provide something extra and they are necessary for completion, they are superfluous and indispensable" (Scott, 1993: p. 241). This contradictory logic is at the heart of what it means to think about gender, Jews, Jewishness, and Judaism well into the twenty-first century. In the context of Jewish communal, religious, and scholarly life, a consideration of women, much less the larger categories of gender and sexuality, has resulted in a radical rethinking of the basic assumptions that have defined every dimension of contemporary Jewish life. On the one hand, the inclusion of women within Jewish life constitutes an expansion of existing narratives. On the other hand, it has become clear that any overview of Jewish life without a consideration of gender and sexuality would be both inaccurate and incomplete.

Late twentieth- and early twenty-first-century Jewish feminist scholarship and activism continue to transform notions of Jews, Jewishness, and Judaism. Key changes in the world of Jewish scholarship are charted by sources noted in the bibliography, including works by Adler, Alpert, Boyarin, Heschel; Hyman, Hyman and Moore, Kaplan; Peskowitz and Levitt, Plaskow, and Prell. Similarly, important discussions of Jewish feminist activism – some of which overlap with the scholarly discussion – include works by Balka and Rose, Beck, Brettschneider, Klepfisz and Kaye/Kantrowitz, Ruttenberg, and Shneer and Aviv. Jewish culture, religion, and history have been profoundly transformed through these efforts, as the simple logic of inclusion has given way to more radical reconfigurations of the Jewish past, present, and future. Jewish feminist activists and scholars have been a part of the broader feminist movement, not only in what is known as the second wave of US feminist activism but also in earlier efforts. Often they have served as leaders in their movements (Klepper, 2013; Hyman, 1995; Kaplan, 1991).

Some Historical Background

As feminist scholars have argued (Kaplan and Moore, 2010; Prell, 1999; Hyman, 1995; Levitt, 1997), the call for inclusion has its origins in the legacy of Jewish emancipation in the West where individual Jewish men – mainly bourgeois – were granted rights as citizens of liberal nation-states. They were emancipated as individuals, not as Jews. By giving up various forms of Jewish communal authority, they came to see themselves reflected in the universal democratic vision of the Age of Revolution.

Although the revolutions of the eighteenth century excluded many people, their vision of universalism offered the promise of emancipation to those who remained disenfranchised within Western culture, including women of the affluent middle classes. Jewish women, whose husbands, brothers, uncles, and fathers had become citizens, began to ask questions about their own status. Using the logic of these liberal revolutions, by the late nineteenth century, Jewish feminists became active in larger women's movements. Together with other women they fought for the right to own property and the right to vote (Klepper, 2013; Hyman, 1995). Such struggles also inspired them to challenge various inequities within Jewish communal life. One result was the dramatic increase in the number of American Jewish women attending synagogues for public worship and the abandonment of gender segregation in seating. By the 1970s, Jewish women used notions of liberal justice and equality to demand their inclusion in all aspects of communal worship. By arguing that specific Jewish communal practices either excluded women or treated them as inferior to men, they demanded their rightful place on the *bema* (the platform with the reading desk at the front of the synagogue), reading torah and leading prayer services. In these ways, they were able to critique the masculinist bias of traditional worship. This liberal critique has expanded to include Jewish social and communal practices, from religious observance to the various fields of Jewish Studies scholarship, although the glass ceiling and powerful inequities in pay and advancement continue to mark organized Jewish communal life and its various institutions (Bronznick et al., 2008).

The growing presence of Jewish women and attention to issues of gender and, increasingly, sexuality have occasioned the rewriting of Jewish history and the rereading of classical Jewish texts in the various subfields of Jewish Studies (religion, literature, history, sociology, political science, and language study) as well as a resurgence in Jewish cultural production including visual art, poetry, fiction, film, music, and dance.

Scholars have used gender analysis to rethink, for example, the gender-segregated spaces of Eastern European Jewish life,

bringing into focus the ingenuity of Jewish women (Parush, 2004). Increasingly, scholars and writers are also addressing the rich gendered histories of Mizrahi – Middle Eastern and North African – Jewish life, and their diasporas (Khazzoom, 2003).

Jewish feminist politics and scholarship, echoing larger scholarly and political trends, has opened up space for serious examination of lesbian, gay, bisexual, transsexual, and queer (LGBTQ) issues in Jewish life, in Jewish texts, and in Jewish ritual and communal practice – perhaps most overtly in the ordination and investiture of LGBTQ people as rabbis and cantors and in the performance of Jewish wedding and commitment ceremonies.

Jewish Religious and Communal Life

In Jewish worship, these changes have been palpable, including a transformation in the language of prayer. Not only is there now gender-inclusive language but also whole new liturgies take account of the experiences of Jewish women (Falk, 1996; Adler, 1998). By the same token, the life experiences of LGBTQ people in the Jewish community have received increasing attention (Drinkwater et al., 2009). Since the 1970s, in increasing numbers of Jewish communities, women are engaging in once restricted religious practices. They are chanting sacred texts, participating in worship services, and receiving honors once granted only to Jewish men. They are leading communal worship as rabbis and cantors (Nadell, 1998). They are also becoming presidents of congregations and Jewish religious organizations. Although many of these efforts remain contested, especially among various Orthodox Jewish communities, among liberal Jews, in Reform, Conservative, Reconstructionist, and Renewal communities, the inclusion of women in all of these realms, including their role as rabbis and cantors, are now very much established practices, especially in the United States. Even among certain Modern Orthodox communities, these roles and practices are changing. Likewise, the growth of Orthodox Jewish feminist prayer groups, study centers, conferences, and the creation of women rabbinic interns throughout the 1990s all attest to the growth of an Orthodox Jewish feminist movement (Greenberg, 1981). Even within ultraorthodox communities, women are reclaiming the rituals of family purity, for example, as attractive Jewish practices for women (Fader, 2009; El-Or, 1994).

In the twenty-first century, formerly marginalized Jews especially gender and sexual minorities and LGBTQ communities, but also disabled Jews, have been granted increasing inclusion. Jewish women and more open LGBTQ Jews are also a part of passing on Jewish religious traditions through the work of Jewish education. Since the nineteenth century, Jewish women have played an ever-increasing role in the education of Jewish children in what had been an exclusively male domain. Women are more often running Jewish day schools sponsored by the various religious movements and training day school teachers and administrators in graduate programs in Jewish education. Among liberal Jews, women are playing an important role in the education of future rabbis and cantors. The number of women on the faculties of the Conservative, Reform, and Reconstructionist rabbinical colleges is growing. Jewish women are also taking on leadership roles as administrators at these institutions; many have come out as LGBTQ people as they serve in these capacities. In all of these realms, there is a growing commitment to include women and LGBTQ Jews in the vital life of Jewish religious practices and Jewish communal life more broadly understood (Prell, 1999). Jewish women are taking on a variety of leadership roles both in North America and in Israel, running Jewish agencies, social service organizations, and federations, although they continue to earn less than their male peers (Bronznick et al., 2008).

In the late twentieth century, feminists made substantial contributions as rabbis, lay leaders, activists, and scholars (Alpert, 1997; Alpert et al., 2001; Shneer and Aviv, 2002; Plaskow, 1990; Beck, 1982/1984/1989; Balka and Rose, 1989). Notions of family, community, and religiosity were powerfully challenged by their strong presence within and outside of organized Jewish communities, leading to more recent and more radical changes. They also challenged the criteria for membership in Jewish congregations and communal organizations, constructing new rituals of recognition such as marriage and commitment ceremonies and events to celebrate "coming out" in a Jewish context. They also joined gay men in demanding that the Jewish community address homophobia and the impact of the AIDS epidemic on Jews, especially as the crisis peaked in North America in the late twentieth century (Balka and Rose, 1989; Shneer and Aviv, 2002).

Jewish Studies

The various fields of scholarship in Jewish Studies have also witnessed the growing presence of women as a precursor to more critical transformations of scholarly practice, particularly in the fields of text study, religion, history, literature, and the various social sciences. Although Jewish women were already active in the academic study of Judaism beginning in the late nineteenth century, the most powerful and dramatic changes came with the second wave of US feminist activism. Feminist scholars have not only added women to the study of ancient Jewish texts, Jewish philosophy, theology, history, literature, and social scientific studies, they have raised questions about both the content of Jewish Studies courses and the ways they are being taught in colleges and universities (Prell, 2007; Diner et al., 2010; Peskowitz and Levitt, 1997; Davidman and Tenenbaum, 1994).

Work on classical Jewish texts, the Hebrew Bible (Levkovitz, 2010; Pardes, 1992), and especially rabbinic literature have changed how these sources are read (Ilan et al., 2007; Labovitz, 2009; Baker, 2002; Peskowitz, 1997; Boyarin, 1997; Hauptman, 1998; Adler, 1998; Biale, 1984). Feminist scholars have not only looked at how Jewish women figure in these texts, they are increasingly raising questions about how gender is produced in these texts and how these visions of gender have become normative. Here reading practices are crucial. As Miriam Peskowitz (1997) notes, such normative texts can and should be read as legal fantasies, following the lead of recent scholarship in critical legal theory; these texts tell us about the desires of those who produced them.

Jewish feminist scholars have also taken on the traditions of Jewish theology, ethics, and philosophy to offer new approaches to these various fields of Jewish religious studies

scholarship (Tirosh Samuelson, 2004; Shapiro, 2003; Adler, 1998; Plaskow, 2005, 1990; Alpert, 1997; Levitt, 1997; Falk, 1996; Weissler, 1998). These scholars not only question the gender of God but the subject of ethical discourse. Who is the person making moral choices, and from what standpoint are these issues being addressed? They also challenge the mind–body dualisms that mark medieval as well as modern Jewish thought and their implications for Jewish women. In Jewish history, feminist scholars have used the incorporation of women to their readings of the Jewish past to rethink some of the grand narratives that have shaped Jewish historical study. By demonstrating the role women played in emancipation and enlightenment, modern Jewish historians have shown us how vital Jewish women were to these transformations in Jewish life (Hyman, 1995; Kaplan, 1991). Jewish feminist historians have also called our attention to the role of Jewish women in assimilation into American culture (Diner et al., 2010; Kaplan and Moore, 2010; Prell, 1999, 2007).

Broader works on twentieth-century Jewish history have also been transformed by accounts of gender. Included in this scholarship are powerful accounts of Zionism as a masculinist national movement (Boyarin, 1997; D. Biale, 1992) as well as a growing body of literature about the role of women in the devastation of the Holocaust (Ofer and Weitzman, 1998; Rittner and Roth, 1993). As Daniel Boyarin (1997) points out, Zionism was a way of remasculinizing the Jewish people. On the other hand, feminist scholars have begun to examine how the experience of genocide was gendered. They point to how and in what ways pregnancy, menarche, menstruation, and motherhood marked Jewish women in the Holocaust. Feminist scholars have also addressed the particular devastation of sexual assault in this context, and spearheaded work on Holocaust memory and its representations (Hirsch, 2012; Levitt, 2007).

Finally, women's historians have called attention to the roles of Jewish women in the ancient past. Not only were they wives, daughters, and mothers, they were also workers. They were dyers and weavers, spinners and artisans alongside men, and they were also communal leaders (Baker, 2002; Brooten, 1982; Peskowitz, 1997; Meyer, 1988). The work of feminist scholars has offered us new evidence of the lives of Jewish women in all these eras of the past. In the social sciences, feminist scholars have called our attention to the role of Jewish women in Jewish communal and political life, including their role in the process of Jewish assimilation. How did they dress themselves and their families or decorate their homes? What did they serve their families for dinner? These are all crucial questions for feminist ethnographers who are re-examining American Jewish assimilation (Diner et al., 2010; Prell, 1999; Kaplan and Moore, 2010; Davidman and Tenenbaum, 1994).

The work in Jewish literature is massive and complex. Feminist scholars have not only discovered heretofore forgotten or unknown texts by Jewish women in various languages including Hebrew, Ladino, Yiddish, Russian, and English (Balin, 2000; Zierler, 2004), they have also offered new accounts of more well-known texts (Fuchs, 1987; Seidman, 1997). Scholars like Seidman (1997) have also asked us to reconsider the relationship between Jewish languages and how gender is used to naturalize the privileging of Hebrew over Yiddish. As she explains, like the relationship between husband and wife, Yiddish, the language of daily life, was subservient to Hebrew, the masculine language of prayer and traditional Jewish study.

Jewish Cultural Production

Recent works of feminist art and literature have not only built on the work of Jewish scholars in a variety of fields, they have also used popular and material culture to offer critical perspectives on Jewishness as a cultural and ethnic as well as a religious identity. Some works highlight the absence of Jewish women from certain forms of religious or cultural production, like the reading and writing of sacred texts (Aylon, 2012). Others have used these artifacts including texts to create new theatrical or cinematic productions on Jewish women. Since the 1970s, there has been a growing body of Jewish feminist literature especially by Jewish lesbian feminists (Alpert, 1997; Alpert et al., 2001; Beck, 1982/1984/1989; Brettschneider, 2006; Klepfisz and Kaye/Kantrowitz, 1986/1989; Ruttenberg, 2001). These writers often focus on Jewishness as a cultural and political and not a religious practice. More recent efforts have included the reclaiming of klezmer music and the resources of Yiddish film for new cultural production.

In the arts, the inclusion of women and queers has expanded the realm of Jewish cultural production in the United States as well as Israel. Increasingly, Jewish women and LGBTQ Jews are playing an active role in these efforts by creating new works, sponsoring exhibits, and participating in various forums to present feminist and queer artistic works to a broader audience (Aylon, 2012; Antler, 1998; Prell, 1999). Here Jewish museums have become a crucial site for new forms of Jewish feminist and queer cultural production (Shneer and Aviv, 2002, 2005).

Future Prospects

Although Jewish women are a part of Jewish communal life, now more than ever, there is much more work to be done. Efforts to make explicit the inequities at the heart of many Jewish communal practices, much less the exclusion of women from sacred texts and the problematic roles they do have when they are included in these texts, all demand ongoing critical examination. Such efforts have bolstered, challenged, and enhanced a growing awareness of the many other forms of exclusion that haunt Jewish life, not only in terms of gender and sexuality but also disability. Among Jewish feminists, the idea of giving up these traditions is too simple. Although they are critical of these texts, they continue to claim them and try to find ways of forging new and more powerful relationships with these legacies. Jewish women and LGBTQ Jews are carrying out this labor as university scholars, day school teachers and administrators, cantors and rabbis, and poets, artists, musicians and filmmakers, and as bloggers.

See also: Antisemitism; Catholicism and Gender; Historical Thought and Historiography: Jewish Tradition; Holocaust, The; Judaism; Nazi Law; Religion, Sociology of; Women in Buddhism; Women's Religiosity.

Bibliography

Adler, R., 1998. Engendering Judaism: An Inclusive Theology and Ethics. Jewish Publication Society, Philadelphia, PA.

Alpert, R., 1997. Like Bread on the Seder Plate: Jewish Lesbians and the Transformation of Tradition. Columbia University Press, New York.

Alpert, R., Levi Elwell, S., Idelson, S. (Eds.), 2001. Lesbian Rabbis: The First Generation. Rutgers University Press, New Brunswick, NJ.

Antler, J. (Ed.), 1998. Talking Back: Images of Jewish Women in American Popular Culture. Brandeis University Press, Lebanon, NH.

Aylon, H., 2012. Whatever is Contained Must Be Released: My Jewish Orthodox Girlhood, My Life as a Feminist Artist. The Feminist Press, New York.

Baker, C., 2002. Rebuilding the House of Israel: Architectures of Gender in Jewish Antiquity. Stanford University Press, Redwood City, CA.

Balin, C., 2000. To Reveal Our Hearts: Jewish Women Writers in Tsarist Russia. Hebrew Union College Press, Cincinnati.

Balka, C., Rose, A. (Eds.), 1989. Twice Blessed: On Being Lesbian, Gay and Jewish. Beacon Press, Boston, MA.

Baskin, J. (Ed.), 1991/1999. Jewish Women in Historical Perspective. Wayne State University Press, Detroit, MI.

Baskin, J., 2002. Midrashic Women: Formations of the Feminine in Rabbinic Literature. Brandeis University Press, Lebanon, NH.

Beck, E.T., 1982/1984/1989. Nice Jewish Girls: A Lesbian Anthology. Beacon Press, Boston, MA.

Biale, D., 1992. Eros and the Jews: From Biblical Israel to Contemporary America. University of California Press, Berkeley, CA.

Biale, R., 1984. Women and Jewish Law: An Exploration of Women's Issues in Halakhic Sources. Schocken Books, New York.

Boyarin, D., 1997. Unheroic Conduct: The Rise of Heterosexuality and the Invention of the Jewish Man. University of California Press, Berkeley, CA.

Boyarin, D., Itzkovitz, D., Pellegrini, A. (Eds.), 2003. Queer Theory and the Jewish Question. Columbia University Press, New York.

Brettschneider, M., 2006. The Family Flamboyant: Race Politics, Queer Families, Jewish Lives. State University of New York Press, Albany, NY.

Bronznick, S., Goldenhar, D., Linsky, M., 2008. Leveling the Playing Field: Advancing Women in Jewish Organizational Life. Advancing Women Professionals and the Jewish Community (AWP), Cambridge Leadership Associates, Cambridge, MA.

Brooten, B.J., 1982. Women Leaders in the Ancient Synagogue: Inscriptional Evidence and Background Issues. Scholars Press, Chico, CA.

Davidman, L., Tenenbaum, S. (Eds.), 1994. Feminist Perspectives on Jewish Studies. Yale University Press, New Haven, CT.

Diner, H., Kohn, S., Kranson, R. (Eds.), 2010. A Jewish Feminine Mystique: Jewish Women in Postwar America. Rutgers University Press, New Brunswick, NJ.

Drinkwater, G., Lesser, J., Shneer, D., Plaskow, J. (Eds.), 2009. Torah Queeries: Weekly Commentaries on the Hebrew Bible. New York University Press, New York.

El-Or, T., 1994. Educated and Ignorant: Ultraorthodox Jewish Women and Their World. Lynn Reiner, Boulder, CO.

Fader, A., 2009. Mitzvah Girls: Bringing up the Next Generation of Hasidic Jews in Brooklyn. Princeton University Press, Princeton, NJ.

Falk, M., 1996. The Book of Blessings: New Jewish Prayers for Daily Life, the Sabbath, and the New Moon Festival. Harper Collins, San Francisco, CA.

Fonrobert, C., 2000. Menstrual Purity: Rabbinic and Christian Reconstructions of Biblical Gender. Stanford University Press, Redwood City, CA.

Fuchs, E., 1987. Israeli Mythogynies: Women in Contemporary Hebrew Fiction. State University of New York Press, Albany, NY.

Fuchs, E. (Ed.), 2005. Israeli Women's Studies: A Reader. Rutgers University Press, New Brunswick, NJ.

Greenberg, B., 1981. On Women and Judaism: A View from Tradition. Jewish Publication Society, Philadelphia, PA.

Hauptman, J., 1998. Rereading the Rabbis: A Woman's Voice. Westview Press, Boulder, CO.

Heschel, S. (Ed.), 1983/1995. On Being a Jewish Feminist: A Reader. Schocken Books, New York.

Hirsch, M., 2012. The Generation of Postmemory: Writing and Visual Culture after the Holocaust. Columbia University Press, New York.

Hyman, P.E., 1995. Gender and Assimilation in Modern Jewish History: The Roles and Representations of Women. University of Washington Press, Seattle, WA.

Hyman, P.E., Moore, D.D. (Eds.), 1997. Jewish Women in America: An Historical Encyclopedia, 2 vols. Routledge, New York.

Ilan, T., Or, T., Salzer, D.M., Steuer, C., Wandrey, I. (Eds.), 2007. A Feminist Commentary on the Babylonian Talmud: Introduction and Studies. Mohr Siebeck, Tübingen.

Irshai, R., 2012. Fertility and Jewish Law: Feminist Perspectives on Orthodox Responsa Literature. Brandeis University Press, Lebanon, NH.

Kaplan, M.A., 1991. The Making of the Jewish Middle Class: Women, Family, and Identity in Imperial Germany. Oxford University Press, New York.

Kaplan, M.A., Moore, D.D. (Eds.), 2010. Gender and Jewish History. Indiana University Press, Bloomington.

Kaufman, D.R., 1991. Rachel's Daughters: Newly Orthodox Jewish Women. Rutgers University Press, New Brunswick, NJ.

Khazzoom, L. (Ed.), 2003. The Flying Camel: Essays on Identity by Women of North African and Middle Eastern Jewish Heritage. Seal Press, New York.

Klapper, M., 2013. Ballots, Babies, and Banners of Peace: American Jewish Women's Activism, 1890–1940. New York University Press, New York.

Klepfisz, I., Kaye/Kantrowitz, M. (Eds.), 1986/1989. The Tribe of Dina: A Jewish Women's Anthology. Beacon Press, Boston, MA.

Labovitz, G., 2009. Marriage and Metaphor: Constructions of Gender in Rabbinic Literature. Lexington Books, Lanham, MD.

Levitt, L., 1997. Jews and Feminism: The Ambivalent Search for Home. Routledge, New York.

Levitt, L., 2007. American Jewish Loss after the Holocaust. New York University Press, New York.

Levkovitz, L., 2010. In Scripture: The First Stories of Jewish Sexual Identities. Rowman and Littlefield Publishers, New York.

Meyer, C.L., 1988. Discovering Eve: Ancient Israelite Women in Context. Oxford University Press, New York.

Nadell, P., 1998. Women Who Would Be Rabbis: A History of Women's Ordination, 1889–1985. Beacon Press, Boston, MA.

Nadell, P. (Ed.), 2003. American Jewish Women's History: A Reader. New York University Press, New York.

Ochs, V., 2007. Inventing Jewish Ritual. The Jewish Publication Society, Philadelphia, PA.

Ofer, D., Weitzman, L.J. (Eds.), 1998. Women in the Holocaust. Yale University Press, New Haven, CT.

Pardes, I., 1992. Countertraditions in the Bible: A Feminist Approach. Harvard University Press, Cambridge, MA.

Parush, I., 2004. Reading Women: Marginality in Nineteenth Century Eastern European Jewish Society. Brandeis University Press, Lebanon, NH.

Peskowitz, M., 1997. Spinning Fantasies: Rabbis, Gender, and History. University of California Press, Berkeley, CA.

Peskowitz, M., Levitt, L. (Eds.), 1997. Judaism since Gender. Routledge, New York.

Pinsky, D., 2009. Jewish Feminists: Complex Identities and Activist Lives. University of Illinois Press, Champaign.

Plaskow, J., 1990. Standing Again at Sinai: Judaism from a Feminist Perspective. Harper Collins, San Francisco, CA.

Plaskow, J., 2005. The Coming of Lilith: Essays on Feminism, Judaism, and Sexual Ethics, 1972–2003. Beacon Press, Boston, MA.

Prell, R., 1999. Fighting to Become Americans: Jews, Gender and the Anxiety of Assimilation. Beacon Press, Boston.

Prell, R. (Ed.), 2007. Women Remaking Judaism. Wayne State University Press, Detroit.

Rittner, C., Roth, J.K. (Eds.), 1993. Different Voices: Women and the Holocaust. Paragon House, New York.

Ruttenberg, D. (Ed.), 2001. Yentl's Revenge: The Next Wave of Jewish Feminism. Seal Press, Berkeley.

Scott, J.W., 1993. Women's history. In: Kauffman, L.S. (Ed.), American Feminist Thought at Century's End: A Reader. Blackwell, Cambridge, UK, pp. 234–257.

Seidman, N., 1997. A Marriage Made in Heaven: The Sexual Politics of Hebrew and Yiddish. University of California Press, Berkeley.

Shapiro, S.E., 2003. 'And God created woman': reading the Bible otherwise. In: Levinas and Biblical Studies. Brill, Leiden, pp. 159–195.

Shneer, D., Aviv, C. (Eds.), 2002. Queer Jews. Routledge, New York.

Shneer, D., Aviv, C., 2005. New Jews: The End of the Jewish Diaspora. New York University Press, New York.

Tirosh Samuelson, H. (Ed.), 2004. Women and Gender in Jewish Philosophy. Indiana University Press, Bloomington.

Umansky, E.M., Ashton, D. (Eds.), 2008/1992. Four Centuries of Jewish Women's Spirituality: A Sourcebook. Brandeis University Press, Lebanon, NH (the original [1992] edition was published by Beacon Press).

Weissler, C., 1998. Voices of the Matriarchs: Listening to the Prayers of Early Modern Jewish Women. Beacon Press, Boston, MA.

Zierler, W., 2004. And Rachel Stole the Idols: The Emergence of Modern Hebrew Women's Writing. Wayne State University Press, Detroit, MI.

Judges

José J Toharia, Metroscopia, Madrid, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

As the official specifically charged with the administration of the law, the figure of the judge has adopted a varied number of shapes in different times and societies. However, the multiplicity of historical experiences can be reduced, for analytical purposes, to a few basic typologies (or 'ideal types,' in Weberian terminology), which are briefly discussed. Particular and more detailed attention is given to the two specific configurations of the judicial role (i.e., the Civil Law and the Common Law judge), which clearly predominate in most contemporary societies.

A judge is a public officer whose role is to administer the law in a Court of Justice impartially and independently. Judicial independence is instrumental to judicial impartiality, since only an independent judge can be expected to be fully impartial: hence the configuration of the role as that of an officer within some governmental structure, a measure originally meant to provide protection against pressures from the confronting parties (Shapiro, 1981). However, while the official status of the judicial function may provide protection against the parties it may also facilitate pressures and influences from those to whom the judge owes the office. In some highly infrequent cases (such as early Roman Law) the judge is freely elected by the parties who commit themselves to abide by his decisions. However, in most historical instances the judge's authority and binding capacity is independent from the will of the involved parties and derives from the backing granted to judicial actions by the political system. This creates a direct link between the legitimacy of judicial institutions and the type of legitimacy of the prevailing form of political government. Judicial impartiality can thus be expected to be reasonably guaranteed only under strict democratic rule, one of the basic requisites of a democratic polity being the existence of an effectively independent judiciary.

Officially empowered to determine in a binding way with which of the contending parties lies the legal right, the judge constitutes a key, central figure in all legal systems, past or present. However, the judge is but one among other possible conflict-solving institutions. Rather than as a discrete, fully differentiated category, he or she can be better understood as one end of a continuum that would run: *go-between*, *mediator*, *arbitrator*, and *judge* (Shapiro, 1981: p. 3). All those roles require impartial implementation but differ in the degree of formalization and depersonalization in presiding over the relation with the contending parties. The *go-between*, who operates on a pure consent basis, represents the more informal, face-to-face and equity-oriented end of the continuum; the judge, professionally trained, vested with official authority and required to lack all personal acquaintance with the parties, occupies the opposite end.

As societies become more complex and differentiated, and as the state-controlled legal system increases its centrality in social life, the role of the judge gains predominance. However, the other conflict-solving roles do not disappear; instead, they subsist as alternative – and usually very active – instances. The continuum can be thus perceived as a cumulative process, allowing for the emergence of progressively more complex and formalized agencies of conflict resolution that coexist with the preceding ones (*see* Dispute Resolutions in Economics).

Types of Judges

A basic typology of judges can be attempted from the combination of three factors that shape judicial roles. First, *the type of legal system* within which the judge operates: a basic distinction can be made between Western legal systems (comprising both Common Law and Civil Law systems), and all other types of systems, such as *Far Eastern* (Chinese and Japanese), *Religious* (Islamic or Hindu), or *Socialist* (see Zweigert and Kötz, 1998). The second factor is the type of political legitimacy underlying the body of laws to be applied, and consequently underlying the political system, which upholds those laws (Weber, 1947). Finally, the basic function that is attributed to the judge: a basic distinction can be made between authoritative adjudication and mediating arbitration. The following classificatory scheme results.

Democratic Judge

A type of judge who adjudicates cases applying rules of general validity, enacted by duly elected representatives of the citizenry. This type of judicial system corresponds to representative democracies (legal-rational type of domination in Weberian terms). There are two possible formats for the *democratic* type of judges: Common Law and Civil Law.

Traditional or Predemocratic Judge

A predominantly arbitrating, conciliatory type of judge who applies norms embodying a traditional legacy, contained in religious, moral or ethical codes, or in local customs, *mores*, or folkways. This type of judge would correspond to the Weberian category of 'traditional' political domination. Several classical types of judicial roles can be included within this category: the historical Islamic *Qadi* Judge (who attempts conciliation – if and when the compromise does not permit what is forbidden or does not forbid what is permitted by the religious law – and who adjudicates on a case-by-case basis according to equity); the traditional Far Eastern Judge (Chinese or Japanese), more concerned with mending and preserving social harmony and

equilibrium than with legal arguments or questions of individual rights; and the African or Native American type of customary judge.

Nondemocratic Judge

A type of judge corresponding to autocratic political situations where the norms to be applied may be claimed to express the will of the people, *but as interpreted and elaborated by an autocratic ruler* (or ruling body), self-appointed as exclusive, natural spokesman for the collective will, and as such above any need for elective confirmation. This third situation could be associated with Weber's concept of *charismatic* rule.

Increasing Convergence

This threefold typology provides basic global models or *ideal types* but is not meant as a description of real contemporary judicial systems. The judicial systems of the more economically developed and more democratically stable countries would generally fit well into the first ('*democratic judge*') category. Most other countries striving in many cases to consolidate modernized and democratic polities, and as such immersed in transitional situations, represent mixed, intermediate stages. However, the generalization can be made that as processes of economic, social, and cultural globalization take place at a world-scale, and as political democratization and the rule of law tend to become increasingly undisputed, universal goals, the tendency can be detected in most developing countries toward the adoption of judicial systems parallel to the two models of justice prevailing in stable democracies (predominantly, the Civil Law type of judicial system). In quite a few cases this has implied some degree of *denaturalization* of the imported model, since the modern Civil Law (as well as the Common Law) type of judge is designed to function under democratic conditions. Certainly this was not always the case: both systems (Civil and Common Law) had emerged and consolidated well before the modern idea of the rule of law had even begun to develop. But their current configuration results from the substantial redefinitions they both experienced in the wake of the democratic revolutions of the eighteenth century.

Judicial Systems in Developing Countries

Purely customary judicial systems have become a rarity of mainly anthropological or historical interest. By 1990, all Francophone African countries, with the exception of Cameroon, had suppressed all types of customary courts, establishing a state system of courts of the Civil Law type (Mangin, 1990). In Anglophone Africa, customary courts are still allowed to exist in most countries but as marginal institutions, their decisions in any case being appealable to state courts, mostly modeled along Common Law lines (Coldham, 1990). One of the most conspicuous instances of Far Eastern traditional-type of judge, the Japanese, is now extinct and has given way to a Civil Law type system of justice, with some elements of Common Law (Oda, 1992).

Apart from China, one of the few remnants of the 'socialist' category of judicial system, most traditional judicial systems are experiencing assimilation to the Western models. In India, where a process of *hybridization* gave birth to 'Anglo-Muhammadan' and 'Anglo-Hindu' law (Zweigert and Kötz, 1998: 310 and 318), these bodies of law are administered by state judges in Common Law style.

In the case of the Islamic world a distinction has to be made between: (1) countries, which in the wake of Islamic fundamentalist revolutions reorganized as Islamic Republics and which would still fit into the traditional/religious category of judicial systems; and (2) countries which though experiencing, in different degrees, a trend toward the re-islamization of social, legal, and political institutions still can be considered to maintain a predominantly secular state-courts system created in most cases in the 1950s and 1960s, along the Civil Law model (Botiveau, 1994).

Civil and Common Law Judges

The two types of judicial systems increasingly predominant in the contemporary world (Civil and Common Law) share a basic common, albeit distant, root: Roman Law, Justinianean and Classic, respectively (van Caenegem, 1988). Their historical development led to differing legal and judicial traditions: Common Law is "judge-made law, that is to say (consisting of) rules to be collected from the judgments of the courts" (Dicey, 1961/1885), whereas codification stands out as the basic identifying feature of Civil Law (Merryman, 1969). But they both embody what seems to be the final stage of the centuries-long process of evolution of the Western legal tradition. As pointed out by Berman (1983) the emergence of Canon Law (which gave rise to the first full legal system), the Protestant Revolution (which secularized the legal world), and the English Revolution (which established the supremacy of Parliament) paved the way for the two versions of democratic judicial systems predominant today and brought about by the American (1776) and French (1789) Revolutions. As democratically reborn judicial systems, both the Civil Law (as reshaped by the French Revolution) and Common Law (as amended on American soil) share a similar contextual prerequisite for their adequate operation: the existence of the rule of law within a pluralistic, representative democracy.

However, both systems diverge in the way in which they attempt to keep the organization of justice in line with the commonly shared dogma of popular sovereignty. The variant of the Civil Law, which originated in the French Revolution was basically intended to prevent judges from interfering with the popular will as expressed in the statutes enacted by duly elected representative bodies – a reaction against the preeminence of *Parlements* under the *Ancien Régime*. In the new, democratic situation the judge had to be just 'the mouth of the law,' in Montesquieu's words. Only from an unswerving application of democratic laws would the judge acquire a democratic character. Contrarily, the main purpose of the judicial system established by the 1787 American Constitution was to make the judge an essential element of the tripartite system of checks and balances that the new democratic polity was intended to be – Congress, the President, and the Judiciary. To this effect,

the judge had to apply the new statutory law democratically enacted (which should be given priority in cases of collision with judge-made law); at the same time the judge was empowered to insure that statutes passed by ephemeral, temporary majorities would not contravene the basic covenant contained in the 'supreme law of the land' (i.e., the Constitution) (*see* Judicial Review).

As a consequence, two clearly differentiated models of 'democratic justice' crystallized. On the one hand, the Common Law judge, as watchperson of the Law, that is, of the "patchwork quilt of legislative provisions held together by thread woven out of general principles and sewn by the judges" (Bell et al., 1998: 5), performs his functions with a basic policy-implementing orientation (Damaska, 1986) predominantly within an adversarial system. On the other hand, the Civil Law judge, as a specialized anonymous civil servant, decides the cases brought to him applying a body of written legal norms – assumed to provide a complete and coherent repertoire of solutions; his functional orientation is inquisitorial and conflict solving. Intended to be a system of justice rigidly respectful of the popular will, the Civil Law system turned out in practice to be a machinery of justice administration insulated from politics. Although this sounds favorable, what it means is that the Civil Law model of justice, devised for a democratic setting, proved equally functional when the overall political system ceased to be democratic. The system had been designed to provide a dutiful application of the laws in force, whatever their nature, not to scrutinize their legitimacy or interpret their contents. A change from democracy to autocracy meant no difference for the internal logic of the system, which could continue operating practically unchanged. The cases of Italy, Germany, and Spain in the twentieth century, to mention but some particularly conspicuous cases where breakdown of democracy, autocratic rule, and return to democracy did not entail any proportionally correlative reorganization of the judicial system and even just marginal changes in the composition of the judiciary provide telling examples (Toharia, 1975). This unintended political *floating* ability and adaptability of the Civil Law model explains the successive attempts to amend its original late-eighteenth-century design in order to reinforce its original democratic structure. The increasing importance given to *jurisprudence* (or binding judicial interpretation of norms), a word which Robespierre wanted to see removed from the French vocabulary; the gradual introduction of varying forms of judicial review (in most cases in the form of Constitutional Courts) largely under Hans Kelsen's influence in the 1920s and 1930s; the practical disappearance of the inquisitorial principle (which in its most classic and pure form has become now a rarity likely to disappear soon, as is the case of Spain); and new forms of judicial activism (Guarnieri and Pederzoli, 1996) are clear examples of the reorganization that the original structure has been experiencing. In a sense, these changes reflect an increasing pattern of convergence between Civil and Common Law models (Cappelletti, 1989), a convergence that goes both ways: a telling example is the impact that the legal system of the European Union is having on the legal and judicial practices of the United Kingdom, which has meant substantial and gradual redefinitions of some basic Common Law features (White, 1999).

The Expanding Power of Judges

Most democratic countries have experienced in recent years a process of *judicialization* of their political life, with "the infusion of judicial decision-making and of court-like procedures into political arenas where they did not previously reside" (Vallinder, 1995). This has produced an unprecedented expansion of the power and influence of judges, especially in Civil Law countries where judges have traditionally kept a very low profile in public affairs. Basically this process has resulted from the combined effect of two major trends: on the one hand, an increased judicial activism in areas such as the control of governmental decisions and the protection and assertion of individual rights; on the other hand, the tendency among political decision makers to defer to the Courts of Justice the resolution of issues with a highly divisive and conflictive potential. This second pattern has been a special cause for concern as it may imply a weakening of the legitimacy of representative institutions. If judicial adjudication and political decision making are perceived not as a dichotomy but as the two ends of a scale (that would run from total majoritarianism to total government by judiciary) what can be concluded is that most democratic countries seem to have experienced over the last two decades a relative swing toward the government by judiciary end of the scale (Vallinder, 1995).

The Selection of Judges

A key element in the configuration of any judicial system is the way in which judges are selected. An ideal system of judicial recruitment should allow for: (1) the selection of competent and qualified persons; (2) through mechanisms unlikely to impair their impartiality and independence while in office; (3) making room for mechanisms of discipline, evaluation, and accountability of the judicial personnel; and (4) aiming to create a judiciary as representative as possible of the social body (in terms of ideology, social class, gender, race, religion, or other relevant social attributes). These four ideal requirements are obviously difficult to meet and most existing systems of judicial recruitment attempt to achieve them through a wide variety of arrangements – which probably explains why "no other governmental official is selected by so many different methods" as the judge (Stumpf, 1998: 135).

A basic, preliminary distinction can be made between *professional* and *bureaucratic* systems of selection (Guarnieri and Pederzoli, 1996), corresponding respectively to Common Law and Civil Law judicial systems. The *professionalized* selection looks for judges among already experienced, practicing lawyers, which produces an older body of judges who have relatively short tenure as judges. (On mechanisms of selection see *Judicial Selection*). By contrast, the *bureaucratic* system of selection prevailing in Civil Law systems allow for much longer judicial careers: from an average entrance age barely around or below 30, to compulsory retirement around age 70 or even 75. The system is not meant to select mature, experienced lawyers and make them judges but rather to select young law graduates, without any previous legal practice, and prepare them for a lifelong career in the judiciary. The sense of achievement within the legal profession and the social aura and moral

auctoritas that characteristically go with the image of the Common Law judge, is lacking in the Civil Law counterpart.

Judges' Social Representativeness

As professionals called to perform a highly specialized task implying extensive legal training, judges have tended historically to be recruited among members of educated, well-to-do sectors of society. This has resulted historically in an almost universal double structural bias within all Judiciaries: a *class bias* (middle- to lower-class people, as well as people from minority groups, were likely to be strongly underrepresented) and a *gender bias* (women were particularly likely to be strongly underrepresented, as a result of the late recognition of their rights to equal education and equal access to professional life). A third, additional bias might even be said to have characterized the judiciary traditionally: a *conservative bias* (as suggested originally by Max Weber, who pointed to the inherent conservative implications of legal reasoning and thence its more likely appeal for conservative-minded persons). As a result women, members of minorities, persons of lower-middle or working-class origin, and people with ideologically progressive orientations were only rarely found within judiciaries.

To a large extent, all these biases seem to be gradually fading, as educational systems have become more inclusive and affirmative action programs have promoted higher levels of social equality. As for the social origin of their members, Civil Law judiciaries have been traditionally less elitist than their Common Law counterparts: in countries such as France, Italy, or Spain most judges tended to come from rural and middle- to lower-middle class families whereas in the US or in the United Kingdom urban upper or upper-middle class origin usually prevailed. But in both cases the presence of judges with lower or working-class origin was practically nil. This situation is changing substantially in both systems as judiciaries gradually become more representative of the prevailing social structure (Toharia, 1987). The data in the table (see Inset) reflect the gradual increase of social pluralism within most judiciaries over the last three decades. The change is particularly perceptible with respect to the increased presence of women, especially in Civil Law systems. The universalistic, competitive nature of the *bureaucratic* process of selection (generally based purely on personal performance in public entrance examinations, which leaves less room for blatant discrimination), and the fact that it takes place at a much earlier point in professional life seems to account for this seemingly more advanced degree of *feminization* of most Civil Law judiciaries. The presence of women within the judiciary may be said in any case to have well passed beyond the 'tokenization' phase, and to be becoming a fully institutionalized pattern. Even in Muslim countries such as Tunisia, in spite of the controversy reopened by the religious fundamentalists' questioning of the eligibility of women for judicial office, in 1994 one out of every four judges was a woman (Botiveau, 1994). To a somewhat lesser degree, the presence of racial and religious minority groups within the judiciary has also increased. As for the *ideological bias* the issue seems to have lost much of its former intensity. In Common Law countries, appointment or direct election of judges on the basis of their ideological orientation is perceived as a legitimate way of creating a bench that 'looks like the country.' In Civil Law countries, where judicial systems are supposed to be person-proof, the scant data available indicate nonetheless that the ideological composition of the judiciary tends to reflect quite closely the ideological pluralism prevailing in the global society (Toharia, 1987). The pattern of affiliation to the ideologically oriented unions of judges existing in countries such as Italy, France, or Spain provides an additional evidence of such pluralism.

How socially representative is the Judiciary?

Women and other social groups within the Judiciary. Some selected historical landmarks

Date	*Country*	
1835	USA	First Roman Catholic appointed to the Supreme Court (Roger B. Taney)
1916	USA	First Jewish-American appointed to the Supreme Court (Louis Brandeis)
1961	USA	African-American judges represent 0.8% of all state and federal judges
1962	UK	First woman appointed as County Judge (Elizabeth Lane)
1965	UK	First woman appointed to the High Court Bench (Elizabeth Lane)
1967	USA	First African-American appointed to the Supreme Court (Thurgood Marshall)
1967	Tunisia	First woman admitted into the Judiciary
1977–80	USA	President Carter's positive action program results in the appointment of 100 women, African-Americans and Hispanic-Americans to the Federal Bench.
1979	UK	England's first Sikh judge appointed
1979	California, USA	Governor Brown appoints America's first openly homosexual judge
1980–81	USA	3.1% of all US State Supreme Court justices are women
1981	USA	First woman appointed to the Supreme Court (Sandra Day O'Connor)
1981	California, USA	Governor Brown appoints first openly lesbian judge

(Continued)

How socially representative is the Judiciary?
Women and other social groups within the Judiciary. Some selected historical landmarks

1986	USA	First Italian-American appointed to the Supreme Court (Antonin Scalia)
1989	Germany	18% of all judges are women (vs 3% in 1960)
1990	The Netherlands	28% of all judges are women (vs 11% in 1973)
1992	Georgia, USA	First African-American woman appointed to the Supreme Court of a former Confederate State
1992	Costa Rica	42% of all judges are women
1992	Buenos Aires, Argentina	26% of all judges are women (vs 11% in 1985).
1993	USA	10% of all federal judges are African-Americans
1993	USA	3.3% of judges in all levels of the US Judiciary are African-Americans
1994	Tunisia	23% of all judges are women
1994	Venezuela	First woman ever appointed as Chief Justice of a country's Supreme Court (Ms. Cecilia Sosa)
1994	UK	Percentage of women in the different levels of the British Judiciary: Heads of Division: none; Lords Justices of Appeal: one out of 32; High Court Judges: 6%; Circuit Judges: 6%; District Judges: 10%.
1995	Italy	41% of all judges are women
1996	France	47% of all judges are women
1996	USA	Of all judges within the Federal Judiciary, 9% are African-American, 4% Hispanic-American, 0.6% Asian-American and 0.2% Native American. These minority groups add up to 13.8%, vs 4.9% in 1976.
1996	USA	Within the Federal Judiciary, 17% are women (vs 1% In 1976 and 9% in 1989). During President Clinton's first administration (1992–96) 31% of all appointments to the Federal Bench were women (as compared with 0.4% under Nixon, 16% under Carter, 8% under Reagan and 19% under Bush), and 33% to members of minority groups.
1999	Spain	34% of all judges are women, vs 12% in 1987. Among the last class of new judges selected, 58% are women.
2010	USA	Upon the confirmation of Associate Justice Elena Kagan, the US Supreme Court counts three women among its nine members for the first time in history.
2012	USA	33% of all active district (or trial) court judges are women
2012	USA	President Obama establishes a record, adding 72 women judges to the Federal Bench, the most ever appointed by a President in a single term. President Clinton had a larger number of women confirmed as federal judges during his two-term presidency (111).
2012	European Union	34% of all supreme court judges in EU member states are women. This percentage varies very substantially between the different countries the differences in this proportion
2012	USA	32% of the 164 active judges sitting on the 13 federal courts of appeal are women
2013	European Union	18 of the 47 judges (38%) of the European Court of Human Rights are women

See also: Constitutional Courts; Courts and Adjudication; East Asian Legal Traditions; Islamic Law; Judicial Behavior; Judicial Independence; Judicial Review; Law and Courts in Authoritarian States; Legal Traditions of Southeast Asia; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Middle Eastern and North African Legal Traditions; Sub-Saharan African Legal Traditions.

Bibliography

Bell, J., Boyron, S., Whittaker, S., 1998. Principles of French Law. Oxford University Press.
Berman, Harold J., 1983. Law and Revolution. The Formation of the Western Legal Tradition. Harvard University Press, Cambridge, MA.
Botiveau, B., 1994. La loi islamique et jugement moderne. Droit et Cultures 28, 25–45.
van Caenegem, R.C., 1988. The Birth of the English Common Law, second ed. Cambridge University Press.
Cappelletti, M., 1989. The Judicial Process in Comparative Perspective. Clarendon Press, Oxford.
Coldham, S., 1990. Les systèmes judiciaires en Afrique anglophone. Afrique Contemporaine 156, 21–26.
Dakolias, M., 1995. A strategy for judicial reform: the experience in Latin America. Virginia Journal of International Law 36, 166–231.
Damaska, M.R., 1986. Two Faces of Justice and State Authority. Yale University Press, New Haven.
Dicey, A.V., 1961. An Introduction to the Study of the Law of the Constitution. Macmillan, London (Originally published 1885).
Guarnieri, C., Pederzoli, P., 1996. La puissance de juger. Michalon, Paris.
Lee, H.P., 2011. Judiciaries in Comparative Perspective. Cambridge University Press.
Mangin, G., 1990. Quelques points de repère dans l'histoire de la Justice en Afrique francophone. Afrique Contemporaine 156, 21–26.

Merryman, J.H., 1969. The Civil Law Tradition. Stanford University Press, Stanford.
Oda, H., 1992. Japanese Law. Oxford University Press, Oxford.
Royer, J.-P., 1995. La Justice en Europe. Pouvoirs 74, 139–152.
Shapiro, M., 1981. Courts. A Comparative and Political Analysis. The University of Chicago Press, Chicago.
Shetreet, S., Forsyth, C. (Eds.), 2011. The Culture of Judicial Independence. Brill/Martinus Nijhoff.
Stumpf, H.P., 1998. American Judicial Politics. Prentice-Hall, Englewood Cliffs, NJ.
Toharia, J.J., 1975. El juez español. Un análisis sociológico. Tecnos, Madrid.
Toharia, J.J., 1987. Pleitos tengas. Introducción a la cultura legal española. CIS/Siglo XXI, Madrid.
Vallinder, T., 1995. When the courts go marching in. In: Tate, N.C., Vallinder, T. (Eds.), The Global Expansion of Judicial Power. New York University Press, New York, p. 13.
Weber, M., 1947. The Theory of Social and Economic Organization. Oxford University Press, New York (Originally published in German in 1925 as the first part of Economy and Society).
White, R.C.A., 1999. The English Legal System in Action. The Administration of Justice, third ed. Oxford University Press.
Zweigert, K., Kötz, H., 1998. An Introduction to Comparative Law. Clarendon Press, Oxford.

Judicial Behavior

Michael J Nelson, The Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

How do judges decide cases? This article reviews the myriad factors that may influence judicial behavior, including law, ideology, and strategic behavior, and argues that no single factor can explain judicial decision-making; instead, the choices justices make are influenced by a combination of factors. The article concludes with some suggestions for future research that acknowledges individual-level variation in the effects of these factors.

What factors influence a judge's decision? To what extent are judges influenced by law, ideology, and politics as they write their opinions and decide cases? Refining the answers to these questions has been the major project of judicial scholars for decades. For many years, scholars leaned upon separate models of judicial decision-making that privileged singular explanations, such as legal or attitudinal factors, as the dominant force underlying judicial decisions. Now, however, scholars have come to the consensus that judicial decision-making is influenced by a variety of legal and extralegal factors, a view summarized well by Gibson (1983): "judges' decisions are a function of what they prefer to do, tempered by what they think they ought to do, but constrained by what they perceive is feasible to do" (9).

This article reviews the political science literature on judicial behavior, with a particular attention to the varying factors that research indicates affect judicial behavior. In particular, the article discusses the relationships between a number of factors, including law, personal preferences, other judges in the legal system, Congress, the executive branch, organized interests, and the public, and judicial behavior. The article concludes with some suggestions for future research that acknowledge individual-level variation in the effects of these factors.

Law

As any elementary school student learns, the task of judges is to interpret the law and to apply prior rulings to the present case. Levi (1948) describes this classic model of legal reasoning and judicial decision-making: "similarity is seen between cases; next the rule of law inherent in the first case is announced; then the rule of law is made applicable to the second case" (2). In other words, judges decide cases through a process of analogical reasoning; they wade through the body of precedent, find the cases most similar to the present one, and then apply the law from those past and similar cases to the facts of the case at hand. But, in practice, to what extent do legal factors affect the outcomes of cases? As such, a natural starting point for understanding judicial behavior is the law.

Conceptually, one can view law both as a constraint and an opportunity for judges (Hansford and Spriggs, 2006). On the one hand, clear precedent bounds judicial behavior by forcing judges to decide a given case in a particular way while also prohibiting judges from deciding cases in other ways. A trial court judge, for example, lacks the discretion to sentence a jaywalker to death. On the other hand, law provides an opportunity for justices. If lower court judges are bound by the decisions of higher court judges, as they are in common law systems, then high court judges have an opportunity to write their policy preferences into law by crafting law that must be followed by lower court judges. Thus, law provides an opportunity for high court judges to chisel their policy preferences into law that binds the decisions of future judges.

One of the earliest attempts to quantify the influence of law on the decisions of judges is Segal's (1984) study of U.S. Supreme Court search and seizure cases 1962–81. Segal coded a variety of facts for each case (e.g., the presence of a warrant) and then used those facts to predict whether the Court would find the search to be a reasonable one. His results indicate that one can predict the Supreme Court's finding about the reasonableness of a search, given only information about legal factors; this provides strong evidence that law plays some role in structuring the decisions of judges.

Of course, not every case fact is equally important; oftentimes, legal rulings give prominence to particular legal factors. For example, the U.S. Supreme Court's ruling in *Adarand Constructors, Inc. v. Peña* (1995) held that the constitutionality of race-based classifications (a case fact) must be decided using strict scrutiny (a legal standard). In other words, if a judge is deciding a case in which a classification that treats people differently based on their race is at issue, then that judge is bound to apply a particular legal standard (strict scrutiny) in his/her decision. If legal considerations are at the forefront in judicial decision-making, then case facts or standards of review should have differential effects before and after a Court ruling that codifies their importance. This is the theory that Richards and Kritzer (2002) test with regard to content-neutrality doctrine in free expression law. They present persuasive evidence that jurisprudential regimes condition judicial decision-making by changing the set of facts that are deemed important for judges to consider when they decide a case in a given area of law while implicitly giving less weight to other case. Still, other scholars have cast doubt upon these findings, suggesting that jurisprudential regimes play only a minor role in structuring the choices justices make. However, the weight of the evidence about legal factors on judicial decision-making indicates that law does play an important, albeit partial, role in structuring the choices that judges make.

Ideology

Strictly legal explanations have a major drawback: they struggle to explain nonunanimous decisions. If legal factors are the predominant base of judicial decision-making, then nonunanimous decisions are 'mistakes' brought about by judges who misinterpret the law. Yet, while unanimous decisions represent the plurality of the decisions of the U.S. Supreme Court (a fact that is often overlooked), a significant number of Supreme Court decisions invite the dissent of one or more judges, particularly in the Court's most political and salient cases. To that end, attitudinal explanations of judicial behavior provide a second factor that can explain the dissenting behavior of judges.

The clearest statement of ideological influences in judicial decision-making comes from Segal and Spaeth (2002), who formalized what they called the attitudinal model of judicial decision-making. They write: "[t]his model holds that the Supreme Court decides disputes in light of the facts of the case vis-à-vis the ideological attitudes and values of the justices. Simply put, Rehnquist votes the way he does because he is extremely conservative; Marshall voted the way he did because he is extremely liberal" (65). In other words, these attitudinalists argue that the personal political values of judges affect their legal decisions.

Segal and Spaeth present a variety of statistical evidence to demonstrate that attitudes have an independent effect on the decisions of judges. First, they return to Segal's (1984) finding that legal factors can predict the Court's eventual finding of the reasonableness of a search, but they go a step further and determine whether, once one considers attitudinal explanations, legal factors still can predict the reasonableness of a search. They find that judicial ideology has an independent and unique effect on judicial decision-making even after the effects of case facts on judicial outcomes are considered. Second, they examine the effects of the norm of *stare decisis*, the legal principle that states that past decisions must be followed when deciding a current case, on the decisions of justices. If law, and not ideology, is the predominant force behind judicial decision-making, then judges who were in dissent in a previous case should be constrained by the precedent set in that past case and vote differently when the issue comes before the court again. Yet, Segal and Spaeth present a variety of evidence suggesting that judges tend to support their initial position, rather than the precedential position, when an issue comes before the court a second time. This finding further bolsters their theory.

This evidence suggesting that judges' personal preferences influence their decisions creates a problem: how does one know what the personal preferences of judges are? After all, the ability of scholars to parse the effects of ideology from the effects of law is limited by their ability to construct a measure of ideology that is both valid and reliable. As a result, a major strand of research over the past 30 years has sought to measure the ideology of judges.

In one of the first attempts to quantify judicial ideology, Pritchett (1948) sought to determine whether Roosevelt's appointees voted differently than those justices who were on the Supreme Court when Roosevelt took office. To this end, Pritchett took a simple, but powerful, approach: he simply counted the number of times that Justice X cast a vote that agreed with Justice Y and then calculated the percentage of cases in which Justice X agreed with Justice Y. His analysis of the Roosevelt court indicated that Roosevelt's appointees were more likely to vote with each other than they were to vote with other president's appointees.

While Pritchett's method was ingenious because it relied upon no assumptions about the ideological direction of the Court's decision, scholars have also measured ideology by aggregating the votes of justices and assigning each vote a liberal or conservative direction. Scholars then use the justice's aggregated liberal or conservative voting in one or more past terms to predict their votes in future cases. However, reliance on justices' votes as an explanatory factor in a model that seeks to predict justices' votes presents a circularity problem: scholars are trying to predict a quantity of interest by using that same quantity of interest as a predictor.

With the circularity problem in mind, scholars have searched for a measure of judicial ideology is valid and does not rely on justices' votes to produce the measure of ideology. With this in mind, a wealth of literature has relied (and, in some cases, still relies) on the party of the appointing president as a measure of judicial ideology. The idea behind the measure is a simple one: if one knows who sought to place a judge on the bench, one can guess accurately about the personal policy preferences of that judge. This measure relies on a major assumption, namely that presidents attempt to cement their legacy by appointing judges who share their views to the U.S. Supreme Court. Still, party identification has been widely used (and is still commonly used today) as a measure of judicial ideology, particularly for lower federal courts in the United States.

Of course, even presidents from the same party vary in their personal views, and the fact that Senatorial confirmation is required of the U.S. federal judges means that any presidential nominee to the bench must be approved by the Senate suggests that presidents may not always be able to place their ideal nominee on the bench. If this is the case, then presidential party is an overly simplistic measure of judicial ideology. To this end, Segal and Cover (1989) relied upon newspaper editorials discussing the views of nominees to the U.S. Supreme Court to create an indicator of judicial ideology that is not dependent upon the eventual votes of the nominees once they take their seat on the Supreme Court. However, because these editorials come at the time of a justice's nomination to the Court, the Segal–Cover scores measure a justice's perceived ideology at the time they were nominated to the Court, and these scores therefore cannot account for any ideological drift that occurs once a justice begins his/her tenure on the Court.

Another common measure of judicial ideology relies upon the votes of judges. Martin and Quinn (2002) updated Pritchett's methodology, using Markov Chain Monte Carlo methods to fit a dynamic item response model on the votes of justices. These Martin–Quinn Scores have become the most widely used measure of judicial ideology for scholars seeking to quantify the observed ideology of U.S. Supreme Court justices. Each justice receives a new score every term, allowing each justice's ideology to change over time, and providing a distinct advantage over measures, like the party of the appointing president, that remain static over a justice's entire tenure on the Court.

The measurement of judicial ideology is still an active research frontier. Traditionally, scholars have assumed that the

decisions of U.S. Supreme Court justices can be explained by knowing the judge's position on a single, liberal-conservative continuum. However, recent research has challenged this assumption. Lauderdale and Clark (2012) estimated a complex statistical model that allows justices' ideologies to vary by issue area, and they present persuasive evidence that the traditional one-dimensional view of judicial ideology is too simplistic, and that justices vary on their ideological location by area of law. Thus, for example, a justice could be fairly liberal on equal protection issues but conservative on criminal procedure cases. Determining the extent to which issue-level variation in ideology affects judicial decision-making is a major open question for scholars to investigate.

Other Justices

Strictly attitudinal explanations suggest that judges always vote sincerely; they only vote in ways that are congruent with their personal policy preferences. But these attitudinal explanations also pose a problem: empirical scholarship has demonstrated that justices sometimes vote strategically rather than sincerely. In other words, justices do not always vote in line with their personal preferences. As Epstein and Knight (1998) put it:

> Justices may be primarily seekers of legal policy, but they are not unconstrained actors who make decisions based only on their own ideological attitudes. Rather, justices are strategic actors who realize that their ability to achieve their goals depends on a consideration of the preferences of other actors, and the choices they expect others to make, and the institutional context in which they act (10).

Thus, under strategic explanations of judicial behavior, justices' beliefs about the consequences of their decisions are involved. Whereas strictly legal explanations posit single-minded judges, who are unconcerned about external reaction, and the attitudinal explanations expect judges to vote sincerely on every case without regard to sound public policy, the strategic explanations depend heavily on the expected interactions between and among actors within the political system.

Moreover, these strategic explanations of judicial behavior focus both on the legal rule adopted in the case rather than simply whether the case is decided in a liberal or conservative manner. In other words, this class of explanations acknowledges that judges can resolve a case in a 'liberal' manner in any number of ways: for example, they can issue a number of decisions or few decisions and they can rely on some precedents to the exclusion of others. Indeed, Maltzman et al. (2000) argue that the U.S. Supreme Court justices may occasionally vote against their personal preferences in order to achieve a final policy outcome that is more preferable to them.

Internally, this type of behavior is seen most clearly through the bargaining process that coincides with opinion writing. Epstein and Knight (1998) argue that individual justices often try to predict the votes of their colleagues in a case in order to act in a sophisticated manner. Therefore, an analysis that simply centers on the justices' votes may miss evidence of strategic behavior; analysis of judicial behavior under these strategic explanations must look beyond the vote on the merits and incorporate the construction of legal rules.

Strategic behavior happens throughout the process of writing opinions. Examining what they term 'the collegial game,' Maltzman et al. (2000) provides a variety of evidence that such sophisticated behavior happens on the Court. In particular, they show that justices are strategic throughout the entire process of drafting their opinions; decisions about how to respond to draft opinions and when to accommodate other justices' views are colored deeply by justices' strategic behavior.

If judges care about policy, then colleagues on their own court are not the only jurists whose actions they need to consider. While the Supreme Court lies at the apex of the federal judiciary, its decisions set precedent for the lower federal courts. Of course, lower federal judges are subject to the same life tenure as Supreme Court justices. Thus, it is difficult to force them to follow any particular Supreme Court precedent; they can theoretically ignore a decision of the Supreme Court without fear of retaliation. As a result, a variety of studies (e.g., Songer et al., 1994) have found that implementation of the decisions of Supreme Court by lower court is neither automatic nor apolitical. Thus, if justices are the policy-minded sophisticated actors posited by strategic explanations, one may expect to find that their voting decisions attempt to maximize the probability that lower federal courts will adopt and cite their legal policies. Murphy (1959) explored the relationship between the Supreme Court and lower federal courts, and argued that "[t]he Supreme Court must take into account the reaction of inferior judges" when making its decisions (1031). However, while many studies have investigated lower court compliance with Supreme Court precedent, there has been little research on the extent to which lower court behavior constrains the actions of the U.S. Supreme Court. Thus, while one may suspect that justices may adjust their voting to increase the probability of lower court implementation, there is no strong empirical evidence to support that assertion.

Courts and Legislatures

The U.S. Congress has the final say over the membership of the federal judiciary, and decides the Court's budget, the size of the federal judiciary, and even its jurisdiction. With Congress's broad powers with respect to the judiciary in mind, strategic explanations of judicial behavior suggest that Court may tailor its opinions in response to the preferences of other actors in an effort to increase the likelihood that the legislature will respond favorably to the decision (Epstein and Knight, 1998). Since other actors control the implementation of their decisions as well as their budget and salary, strategic explanations of judicial behavior suggest that external actors could constrain the Court through their control over various aspects of the judiciary.

Scholars have studied the relationship between congressional preferences and the Court's decision to hear a case. Of course, the Supreme Court has almost complete agenda-setting power; indeed, with few exceptions, the Court is able to choose between cases it would decide and cases it would decline to hear (Perry, 1991). With such wide discretion, there is reason to believe that the Court may act strategically to select certain cases and reject other, less desirable cases. Harvey and Friedman (2009) find evidence of a relationship between congressional preferences and the decision to hear a case in

their analysis of statutes passed between 1987 and 2000; using a statute-based data set and examining whether or not a particular statute is reviewed in a given year, they argue that the Court's docket consists primarily of cases in which the Court will face little constraint from the other branches of government as the justices reach their preferred ruling.

On the other hand, Owens (2010) uses data from the personal papers of the justices to compare attitudinal and strategic accounts of decision-making at this stage in the judicial process. Owens examines justices' votes to hear a case between 1953 and 1993, and his analysis considers alternate explanations about the agenda-setting process absent in Harvey and Friedman's study. Still, Owens finds no support for strategic behavior in any of his models. Given Owens's results and the weakness of Harvey and Friedman's findings, current research implies that any external constraint that does exist in the agenda-setting stage is weak, at best.

Political scientists have devoted substantially more attention to the content of the Court's final opinions and their relationship with the preferences of the other branches of government; again, the evidence is mixed. On the one hand, Segal (1997) and Sala and Spriggs (2004) found only negligible evidence of strategic voting behavior. In his study, Segal predicts justice's voting positions based on their preferences and then looks for evidence of strategic behavior as the political environment changes. Thus, Segal's approach allows him to compare directly attitudinal and strategic explanations; Segal finds that attitudinal explanations are superior in their ability to explain a justice's final voting behavior. Additionally, Segal and Spaeth (2002) update and refine Segal's work and reach a similar conclusion; they argue that justices' policy goals lead them to vote sincerely in the vast majority of cases. Furthermore, Spriggs and Hansford (2001), examining the factors that lead to the overruling of precedent by the Court, find no support for a separation of powers constraint; thus, there is strong evidence that the Court is not influenced by the ideology of the other branches of government.

Still, not all of the empirical evidence suggests that the Court is completely unconstrained by congressional preferences. First, Spiller and Gely's (1992) analysis of the Supreme Court's decisions in the National Labor Relations Board cases found that the Court was significantly constrained by congressional preferences. Second, Bergara et al. (2003) reanalyze Segal's data regarding congressional and judicial preferences using Spiller and Gely's methodology and find that justices act strategically in response to congressional and presidential preferences and that the level of constraint varies widely over time. Finally, Clark (2009) demonstrates that, over time, the Court's use of judicial review has been directly proportional to the number of court-curbing bills (such as legislation to reduce the Court's jurisdiction) recently considered by the U.S. Congress. As Congress becomes more hostile toward the Court, the Court becomes less likely to use its power of judicial review.

Courts and Executives

But what of the executive branch? The Supreme Court often invites the Solicitor General – the federal government's lawyer – to share his views with the Court in important cases, and his views are so important that he is often called the 'Tenth Justice.' The Solicitor General serves a number of important roles for the Supreme Court, including providing credible information about the case as well as a cue about the views of the executive branch. By providing additional legal argumentation as well as the official position of the executive branch on a given case, the Solicitor General plays a vital role in the Court's decision-making process.

Regardless of the specific role it serves, the Solicitor General is incredibly successful when he appears before the Court. Indeed, studies show that the Court is much more likely to heed the recommendations of the Solicitor General in either the merits or *cert* stages of its decision-making process than any other attorney or actor. Indeed, the Solicitor General is the winningest lawyer to consistently appear before the Court.

Organized Interests

Beyond mass public opinion, the justices may also heed the views of organized interests as they shape their docket and rule on the merits. The most common way for organized interests to participate in the decision-making process is by filing *amicus curiae* (friend of the court) briefs at the merits and/or *cert* stages of the Court's decision-making process. These *amicus* briefs provide the Court with additional legal arguments as well as cues about what sorts of groups support various positions in the case and how important the case is to the public.

Research suggests that the Court responds strongly to these briefs. Caldeira and Wright (1988) demonstrate that, as the number of *amicus* briefs filed at the *cert* stage increases (regardless of whether the briefs suggest the Court should hear the case), the Court becomes more likely to hear a case. Similarly, a wealth of other research suggests that the Court incorporates arguments made by interest groups in their *amicus* briefs into the final opinions, providing evidence that these briefs are an efficacious way for organized interests to shape judicial policy on issues they care about.

The Public

Beyond listening to the other branches of government, justices may also heed the will of the people. There are many reasons to expect that the Court is considerate of the public will as it decides cases. First, if the justices are strategic actors concerned about the implementation of their policies, they should pay attention to public opinion; Rosenberg (2008) notes that the Court's opinions only reach their full potential when they are in line with public opinion. Second, the Court may act strategically out of an attempt to increase its legitimacy. Since courts have neither the power of the purse nor the sword, maintaining institutional legitimacy is of utmost importance to Supreme Court justices (Gibson and Nelson, forthcoming).

While one typically does not think of the Court as a representative institution, Marshall (2008) argues that the Court's decisions have been consistent with public opinion in between three-fifths to two-thirds of cases since the 1930s. Furthermore, Marshall cites reasons to believe that the Court may pay attention to public opinion; for example, the Rehnquist court

referred to public opinion in about six of its opinions per year. If the Court references public opinion directly in the text of its decisions, then there exists a strong theoretical argument that the Court both is aware of and may be influenced by public opinion. In a study of state trial court judges, Nelson (2014) presents results that suggest that, as judges have access to better information about public opinion, they adjust their decisions to fall in line with citizen preferences.

However, evidence to support the claim that public opinion has a direct effect on the Court's voting behavior has been mixed. Epstein and Martin (2010) summarize the state of the literature in the title of their article: "Does Public Opinion Influence the Supreme Court? Possibly Yes (But We're Not Sure Why)." Their empirical examination finds, in line with other studies, that the Court tends to issue more liberal decisions as the U.S. public becomes more liberal and more conservative decisions as the public mood trends in that direction. Yet, there are two distinctive mechanisms through which this ideological change could occur. First, justices might simply respond to public opinion. In other words, public opinion has a causal effect on judicial decisions, and Supreme Court justices respond directly to public opinion. Second, justices are people, too. For that reason, the same social forces that lead the mass public to become more liberal (conservative) might also act on the justices, causing them to become more liberal (conservative). Thus, public opinion may only have an indirect effect on judicial decision-making. The two explanations are difficult to parse because they are often observationally equivalent.

Still, scholars have done their best to parse the two explanations. Casillas et al. (2011) develop a model that controls social factors that cause macro-level opinion change, allowing them to separate the effect of direct and indirect public opinion. Their results suggest that public opinion has a causal effect on judicial decision-making, particularly in nonsalient cases. Additionally, Uribe (2014) targets hypotheses to find specific evidence that justices directly follow public opinion to maintain high levels of public support. She finds no evidence that justices behave strategically in this way. Instead, her results suggest that justices vote based on their sincere ideological preferences and that the similarity between a justice's ideology and public mood affects whether justices vote in line with, irrespective of, or against public opinion.

Conclusion

The factors that affect judicial decisions are multifaceted and varied, and include legal, attitudinal, and strategic factors. Still, one must remember that each individual explanation is not mutually exclusive; understanding how judges decide cases is rarely simple as understanding the effect of a singular factor. Rather, judges' decisions result from a rich set of considerations that encompass both law and politics. Indeed, George and Epstein (1992) demonstrate that the predictive power of models that include just legal or just attitudinal factors is surpassed by a model that includes both of those factors. Likewise, the effects of one explanation may be condition on the effects of another. For example, Bartels (2009) shows that legal considerations condition the effects of ideology; the predictive power of ideology is enhanced in cases that grant judges broad discretion and is mitigated in cases where judges are constrained by law.

Additionally, one must remember that all of the explanations discussed here have focused upon average behavior on the U.S. Supreme Court. Surely, judges on different courts may be more or less affected by one of these factors than judges on a different court. For example, Zorn and Bowie (2010) show that the influence of ideology on judicial decision-making increases as one moves up the judicial hierarchy. Moreover, the effects of these explanations may vary across judges, even those who sit on the same court. Indeed, some judges may have different views of the judicial role that leads to individual-level variation in the predictive power of these explanations. This is an important frontier for future research.

See also: Judicial Behavior; Judicial Independence.

Bibliography

Bartels, Brandon L., 2009. The constraining capacity of legal doctrine on the U.S. Supreme Court. American Political Science Review 103 (3), 474–495.

Bergara, Mario, Richman, Barak, Spiller, Pablo T., 2003. Modeling Supreme Court strategic decision making: the congressional constraint. Legislative Studies Quarterly 28, 247–280.

Caldeira, Gregory A., Wright, John R., 1988. Organized interests and agenda setting in the U.S. Supreme Court. American Political Science Review 82, 1109–1127.

Casillas, Christopher J., Enns, Peter K., Wohlfarth, Patrick C., 2011. How public opinion constrains the U.S. Supreme Court. American Journal of Political Science 55, 74–88.

Clark, Tom S., 2009. The separation of powers, court curbing, and judicial legitimacy. American Journal of Political Science 53, 971–989.

Epstein, Lee, Martin, Andrew D., 2010. Does public opinion influence the supreme court? Possibly yes (But we're not sure why). University of Pennsylvania Journal of Constitutional Law 13, 263–281.

Epstein, Lee, Knight, Jack, 1998. The Choices Justices Make. CQ Press.

George, Tracey E., Epstein, Lee, 1992. On the nature of Supreme Court decision making. American Political Science Review 86, 323–337.

Gibson, James L., 1983. From simplicity to complexity: the development of theory in the study of judicial behavior. Political Behavior 5 (1), 7–49.

2014 Gibson, James L., Nelson, Michael J. Is the U.S. Supreme Court's legitimacy grounded in performance satisfaction and ideology? American Journal of Political Science, forthcoming.

Hansford, Thomas G., Spriggs II, James F., 2006. The Politics of Precedent. Princeton University Press.

Harvey, Anna, Friedman, Barry, 2009. Ducking trouble: congressionally induced selection bias in the supreme court's agenda. Journal of Politics 71, 574–592.

Lauderdale, Benjamin E., Clark, Tom S., 2012. The supreme court's many median justices. American Political Science Review 106 (4), 847–866.

Levi, Edward H., 1948. An Introduction to Legal Reasoning. University of Chicago Press.

Maltzman, Forrest, Spriggs II, James F., Wahlbeck, Paul J., 2000. Crafting Law on the Supreme Court: The Collegial Game. Cambridge University Press.

Marshall, Thomas R., 2008. Public Opinion and the Rehnquist Court. State University of New York Press.

Martin, Andrew D., Quinn, Kevin M., 2002. Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. Supreme Court, 1953–1999. Political Analysis 10, 134–153.

Murphy, Walter F., 1959. Lower court checks on supreme court power. American Political Science Review 53, 1017–1031.

Nelson, Michael J., 2014. Representative justice? Public opinion and the criminal justice system. Journal of Law and Courts 2, 117–152.

Owens, Ryan J., 2010. The separation of powers, judicial independence, and strategic agenda setting. American Journal of Political Science 54, 412–427.

Perry Jr., H.W., 1991. Deciding to Decide: Agenda Setting in the United States Supreme Court. Harvard University Press.

Pritchett, C. Herman, 1948. The Roosevelt Court: A Study in Judicial Politics and Values 1937–1947. The Macmillan Co, New York.

Richards, Mark J., Kritzer, Herbert M., 2002. Jurisprudential regimes in supreme court decision making. American Political Science Review 96, 305–320.
Rosenberg, Gerald N., 2008. The Hollow Hope: Can Courts Bring about Social Change?, second ed. University of Chicago Press.
Sala, Brian R., Spriggs II, James F., 2004. Designing tests of the supreme court and the separation of powers. Political Research Quarterly 57, 197–208.
Segal, Jeffrey A., 1984. Predicting supreme court cases probabilistically: the search and seizure cases, 1962–1981. American Political Science Review 78, 891–900.
Segal, Jeffrey A., 1997. Separation-of-powers games in the positive theory of congress and courts. American Political Science Review 91, 28–44.
Segal, Jeffrey A., Cover, Albert D., 1989. Ideological values and the votes of U.S. Supreme Court justices. American Political Science Review 83, 557–565.
Segal, Jeffrey A., Spaeth, Harold J., 2002. The Supreme Court and the Attitudinal Model Revisited. Cambridge University Press.
Songer, Donald R., Segal, Jeffrey A., Cameron, Charles M., 1994. The Hierarchy of Justice: Testing a Principal-Agent Model of Supreme Court-Circuit Court Interactions. American Journal of Political Science 38, 673–696.
Spiller, Pablo T., Gely, Rafael, 1992. Congressional control or judicial independence: the determinants of U.S. Supreme Court labor-relations decisions, 1949–1988. The RAND Journal of Economics 23, 463–492.
Spriggs II, James F., Hansford, Thomas G., 2001. Explaining the overruling of U.S. Supreme Court precedent. The Journal of Politics 63 (4), 1091–1111.
Uribe, Alicia, 2014. Chasing the Public? An Analysis of the Relationship between Public Opinion and Supreme Court Votes. Working Paper. University in St. Louis, Washington.
Zorn, Christopher, Bowie, Jennifer, 2010. Ideological influence in the federal judicial hierarchy: an empirical assessment. Journal of Politics 72 (4), 1212–1221.

Relevant Websites

http://artsandsciences.sc.edu/poli/juri/ – The Judicial Research Initiative (JuRI) at the University of South Carolina.
http://complaw.wustl.edu/ – Comparative Law.
http://epstein.usc.edu/blackmun.php – Lee Epstein, Washington University in St. Louis.
http://scdb.wustl.edu – The Supreme Court Database.

Judicial Independence

Michael J Nelson, The Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the literature on the measurement and consequences of judicial independence worldwide. First, the article discusses the difficulties inherent in defining judicial independence. Next, the article describes various measures of *de facto* and *de jure* judicial independence and their properties. Finally, the article describes the consequences of judicial independence for politics and economic development worldwide.

Scholars have long been concerned about the ability of judges to make decisions without the fear of reprisal by external political actors. Because judges generally lack the legitimacy bestowed upon presidents and legislators by the electoral process, scholars have concerned themselves with the ability of judges to render legitimate decisions without undue external political influence. In other words, if judges are to fulfill a role as effective checks on the arbitrary and capricious use of power by the executive and legislative branches of government, they must be able to issue decisions that effectively limit the powers of those branches of government with the confidence that those decisions will be respected.

Given the normative importance of judicial independence, scholars have investigated both the determinants of judicial independence as well as its effects on the relationships between that judicial institutions and other actors within the political system and the effects of judicial independence on a variety of economic and political outcomes. Research indicates that the level of judicial independence enjoyed by a court affects both how judges make decisions as well as the economic prosperity of countries.

Still, while widely invoked in the literature, scholars have had widely divergent ideas about what it means for a court (or a judge) to be independent, as well as how one might know if a judge or court is actually independent. This article reviews the literature on judicial independence. It begins by defining judicial independence and exploring the two major facets of the concept – *de facto* judicial independence and *de jure* judicial independence – as well as the myriad ways scholars have attempted to measure both types of judicial independence. Next, the article discusses the effects of judicial independence worldwide with particular attention to the U.S. Supreme Court and the courts of last resort in the U.S. states. The article concludes with some comments about directions for future research on judicial independence worldwide.

Defining Judicial Independence

Scholars generally argue that an independent judiciary is an important feature of a country's political system because it enables the judiciary to serve as a strong check against arbitrary or excessive uses of power by the executive or legislative branches of government. However, scholars disagree about how one would know if a court is actually independent. As a result, any conceptualization of judicial independence must grapple with questions involving the types of protections to which judicial independence refers, from whom a court is independent, and the appropriate unit of analysis suggested by the notion of an independent court. These three questions highlight the difficulties inherent in conceptualizing judicial independence.

First, does judicial independence refer to formal protections from outside influence or more on-the-ground, practical considerations? This question is perhaps the major question that has vexed judicial scholars interrogating the concept of judicial independence. On the one hand, formal constitutional provisions provide a clear way to demarcate judiciaries with clear independent bases of power from those who do not have those protections. For example, we might believe that judges, like those who serve on the U.S. Supreme Court, who are granted life tenure at the time of their appointment are automatically more independent than those, like most in the U.S. States, who must rely on period reelection or retention in order to keep their seats on the bench. The logic is this: life tenured judges have no formal reason to mind the preferences of the legislature or the public in order to keep their position on the bench whereas judges who are subject to reelection or reappointment may be more likely to defer to the whims of the body who will eventually choose whether they should remain in office.

On the other hand, not all judicial powers are formally written in the constitution; indeed, the U.S. Supreme Court's power of judicial review appears nowhere in the text of that country's constitution. As a result, one might also measure judicial independence based on the extent to which a court is independent in practice, rather than according to formal institutional rules. Implementation provides a key example of the importance of considering judicial independence from a practical perspective. Even given life tenure, a judge must rely on some external actor, such as a bureaucracy or a legislature to translate her opinions into practice. To this end, even a judge who need not worry about her continued tenure on the bench may still, in practice, make deferential decisions out of a concern for implementation.

To this end, empirical scholars have tended to acknowledge that judicial independence can be partitioned into two categories: *de jure* and *de facto* judicial independence (Feld and Voigt, 2003). The former category assesses the separation of the judicial branch from the other branches of government

given the formal institutional rules and procedures that govern political actors; the concept of *de jure* judicial independence involves close attention to what Rios-Figueroa and Staton (2014) call "institutions of judicial insulation" (4). The latter category – *de facto* judicial independence – assesses the extent to which those rules and constraints actually function to affect the behavior of judicial actors. *De facto* judicial independence is present when a judge's (or court's) decisions represent her (or its) own preferences without regard for the preferences of other actors (Rios-Figueroa and Staton, 2014).

The second major question facing scholars concerns from whom a court is independent. Judges may be dependent upon many actors: bureaucrats who implement their decisions, lower court judges who interpret the decisions of high court judges, the public whose legitimacy courts need in order to make countermajoritarian decisions without reprisal, and/or the other institutions of government whose law and/or actions the Court must review. Any attempt to measure or discuss judicial independence requires scholars to determine both which of these individuals and institutions an independent court is liberated from and how to classify those courts who are independent from some of these actors but not other actors. Put differently, is a court only 'independent' if it can act free from the influence of the executive branch, the legislative branch, *and* the public, or might a court be independent if it is reliant on some of those actors but not others? To this end (and discussed in more detail below), scholarship addressing the measurement of judicial independence has tended to treat the concept as a continuum ranging from courts that are completely dependent on other political actors to, at the other extreme, courts whose actions are not reliant upon the will of those actors.

A final question facing scholars when defining judicial independence concerns the appropriate unit of analysis. In other words, is judicial independence a characteristic of countries, courts, or judges? Put differently, are judicial independence something that is constant across all courts within a country? Indeed, it seems increasingly likely (especially in countries or states where judges face highly political reappointment procedures) that a preference for judicial independence is best ascribed at the level of the individual judge. Such a theory is bolstered by findings, like those of Ramseyer and Rasmusen's (2001), that judges vary individually in their deference to other branches of government. Ramseyer and Rasmusen (2001) investigate the propensity of Japanese judges to rule against the government, and they find that career-minded Japanese judges are less likely to rule against the government in politically important cases for fear that ruling against the government will lead them to receiving undesirable career placements in the future. However, most traditional measures of judicial independence treat the concept as one which is a characteristic of a country.

Measuring Judicial Independence

Scholars have grappled with these three practical concerns discussed above when creating measures of judicial independence. Given the difficulties inherent in defining the concept, scholars have generated a variety of different ways to measure the extent to which a country's courts are independent. Given the distinction between *de facto* and *de jure* judicial independence, scholars have tended to develop measures of each concept separately. In this section, I review existing measures of judicial independence before discussing the empirical evidence about the relationship between *de jure* and *de facto* judicial independence. For a comprehensive discussion of each measure, see Rios-Figueroa and Staton (2014).

Measuring *De Jure* Judicial Independence

As discussed above, *de jure* judicial independence refers to the extent to which a court is formally freed from the influence of other political actors. Given that *de jure* independence is based upon formal, written considerations, scholars have relied heavily upon content analysis of constitutions when constructing measures of the concept. This section outlines the components underlying three prominent measures of *de jure* judicial independence.

Feld and Voigt's (2003) measure of *de jure* independence relies upon the presence of formal institutional characteristics of national judiciaries. The measure assesses the presence or absence of 23 characteristics in 12 categories, including constitutional amendment difficulty, judicial selection and removal procedures, salary protections, judicial review, and whether or not the highest court is required in the constitution. Feld and Voigt create an index that ranges from 0 to 12 and, due to a lack of data from some countries, divide the total number of indicators present within a country by the total number of indicators for which there is available data. The resulting measure is a continuous measure of *de jure* independence.

Keith (2002) dissolves the concept of *de jure* independence further into separate indicators of the concept; Keith's measure also relies upon content analysis of national constitutions. The resulting measure is composed of seven separate trichotomous measures, each corresponding to separate dimensions of *de jure* judicial independence. The scales include measures for financial autonomy, a separation of powers system, whether or not the constitution insures the finality of judicial decisions, and whether or not judges have exclusive authority over their jurisdiction.

A final measure of *de jure* judicial independence was derived by La Porta et al. (2004). Again relying on content analysis of national constitutions, La Porta et al. examine judicial tenure, judicial policy-making authority, and the extent of constitutional review to create a measure of "judicial checks and balances." La Porta et al. average each of these characteristics together and normalize the resulting score to create a continuous measure of judicial independence.

Since these three measures of judicial independence purport to measure the same concept, one natural question to ask is the extent to which the three measures suggest similar results for a given country. Rios-Figueroa and Staton (2014) conducted an analysis of the association between empirical indicators of *de facto* judicial independence and found disheartening results. Their results indicate that these different indicators tend to be correlated with each other at very low levels, suggesting that they are actually measuring different types of judicial independence. More importantly, the low correlation between the

separate indicators of *de facto* independence indicates that these measures are unstable measures of the underlying concept.

Measuring *De Facto* Judicial Independence

De facto judicial independence moves away from a reliance on formal guarantees of judicial independence toward more practical considerations regarding the extent to which courts appear to be independent in practice. To this end, measures of *de facto* judicial independence rely on data from two sources: expert judgments about the political climate in a given country and reports published by experts, such as the United States State Department. While the sources for the measures differ, each of the following measures attempts to quantify the degree to which the decisions of judicial actors are, in practice, constrained by the preferences of other political actors within their country.

Feld and Voigt's (2003) measure of *de facto* judicial independence draws upon a survey of experts and political elites. The measure includes length of tenure, the number of judges removed from office, judicial salaries, and compliance by other political actors in their implementation of judicial opinions. As with their measure of *de jure* judicial independence, the measure begins with a count of the number of characteristics present in a particular country, and Feld and Voigt divide that count by the total number of characteristics for which they were able to collect data for a particular country.

The Bertelsmann Transformation Index (2010) also relies on expert judgments to create another measure that captures *de jure* judicial independence. The scale ranges from 1 to 10, and it incorporates expert judgments on a number of factors including the extent of corruption, the presence of judicial review, the professionalization of the judiciary, and the institutional differentiation of the judiciary.

The Political Risk Services (2010) 'Law & Order' measure integrates both political and social attributes of the efficacy of the rule of law within a country. Both 'law' (based on "the strength and impartiality of the legal system") and 'order' (capturing the "popular observance of law") are measured on a separate, 0–3 ordinal scale, and the two scales are combined for the final measure (5). Thus, "a country can enjoy a high rating – 3 – in terms of its judicial system, but a low rating – 1 – if it suffers from a very high crime rate of if the law is routinely ignored without effective sanction (for example, widespread illegal strikes)" (5).

Yet another measure of *de facto* judicial independence comes in POLITY's measure of constraints on the executive (XCONST). The seven-point scale is used to score the relative level of restraint placed on the executive branch's ability to act. These constraints are based on the efficacy of' 'accountability groups' like "the ruling party in a one-party state; councils of nobles or powerful; advisors in monarchies [and] the military in coup-prone polities" in their attempts to constrain the executive (Marshall, 2010: 24). Along with these accountability groups, the independence of the judiciary is included as one factor that comprises this variable.

Henisz (2000) combines the POLITY measure with the Political Risk Services measure to form a dichotomous measure of judicial independence; in his measure, a country scoring at or above the median score on both scales is labeled independent while those below the median on one or both measures are labeled not independent.

Other measures of *de facto* judicial independence rely upon reports compiled by the U.S. State Department. In one such attempt, Howard and Carey (2004) code a judiciary as fully independent if it, in practice, functions independently of both the executive branch and the legislature, provides basic due process to criminal defendants, and is basically free of corruption and bribery. Their resulting measure is a three-category ordinal scale.

Other scholars have also relied upon the U.S. State Department's reports. Cingranelli and Richards (2008) construct a trichotomous measure of judicial independence using these reports; by their measure, a country has a fully independent judiciary if the court has the power of judicial review, judges on the national supreme court have a tenure of at least 7 years, court hearings are open to the public, judgeships are open to the public, political elites have limited removal power, and if actions of the political branches of the government may be challenged in the judiciary.

Tate and Keith (2009) construct a similar trichotomous measure using U.S. State Department reports. By their measure, a judiciary is independent if it operates with no mention of corruption or outside influence or interference from the executive or other branches of government.

Rios-Figueroa and Staton (2014) also analyze the extent to which various measures of *de facto* judicial independence seem to measure the same underlying concept. Their examination indicates that measures of *de facto* judicial independence tend to correlate with each other at higher levels than do measures of *de jure* judicial independence, providing some evidence of the reliability of these measures.

The Relationship between *De Facto* and *De Jure* Independence

Although scholars generally have developed separate measures of *de facto* and *de jure* judicial independence, a natural question to ask is the extent to which the two types of judicial independence are related to one another. To put the question differently, how well can we tell if a judiciary is independent in practice if we know how independent they are on paper? To answer this question, Rios-Figueroa and Staton (2014) conduct an empirical analysis of many of the measures discussed above. Their empirical analyses suggest that the two types of judicial independence are quite different from one another; measures of *de facto* and *de jure* judicial independence capture different underlying concepts. Moreover, given information about *de jure* or *de facto* judicial independence in one country does little to predict levels of the other type of judicial independence; measures of each concept are not highly correlated with each other.

The Effects of Judicial Independence

In addition to the major research programs attempting to quantify and conceptualize judicial independence, another

robust literature on the topic addresses the effects of judicial independence on politics and policy within a country and across countries. This section reviews the effects of judicial independence on countries worldwide, giving special attention to judicial independence at the state and federal levels in the United States.

Judicial Independence in Comparative Perspective

One major question regarding judicial independence concerns the cross-sectional effects of judicial independence on economic and political features of a country. Put differently, scholars have sought to determine the extent to which variations in countries' levels of judicial independence, both over time and across countries, affect government, politics, and economics in those countries.

One prominent concern in the literature relates to the relationship between judicial independence and a country's system of government. There is widespread evidence that judicial independence promotes democratic governance by lowering the stakes of being in power; by limiting the benefits leaders can capture by holding power, independent courts promote stable democratic governance (North et al., 2000). Likewise, Reenock et al. (2013) demonstrate that independent courts encourage the survival of democratic regimes, particularly in situations were compromise between classes is promoted by economic conditions.

Though high levels of judicial independence tend to promote democratic, rather than authoritarian, rule, some authoritarian countries do have independent judiciaries. Why do authoritarian governments permit independent courts? After all, authoritarian governments centralize power in the hands of a few people, and independent judiciaries are empowered to take action to limit the use of that power. Thus, it might seem irrational for dictators to allow independent courts to flourish. Moustafa (2007) investigates this question thoroughly, documenting the introduction of an independent judiciary in Egypt. His research demonstrates that authoritarian regimes allow independent judiciaries in an attempt to credibly commit to other countries that they will permit the rule of law, encouraging those countries to form economic linkages with their country.

Beyond studies seeking to explain the relationship between systems of government and levels of judicial independence, scholars have also sought to examine the relationship between levels of judicial independence and respect for rights. One set of studies has documented the importance of an independent judiciary for the protection of rights. North and Weingast (1989) argue that independent judiciaries are valuable because they make credible the promises of states to respect the rights of individual citizens. Empirical studies have tended to support this assertion, demonstrating that countries with greater levels of judicial independence tend to have higher levels of human rights protections (Keith, 2002; Powell and Staton, 2009).

Finally, another set of studies has documented the positive effects of judicial independence on a country's economic outcomes; by protecting property rights, independent courts are able to encourage economic development (North, 1990). For example, Barro (2000) documents the relationship between respect for the rule of law (*de facto* independence) and finds that economic growth is greater in countries with higher levels of respect for the rule of law, that investment is higher in countries with greater respect for the rule of law, and that countries with more respect for the rule of law have lower levels of income inequality.

Judicial Independence and the U.S. Supreme Court

The framers of the U.S. Constitution took a variety of measures to limit overt political influence on the decisions of the U.S. Supreme Court. The most obvious institutional shield against external influence on the decision making of the U.S. Supreme Court is the fact that all Justices are appointed for a period of 'good behavior' which, in practice, has come to mean life tenure. Justices who retire are still given a salary, a law clerk, and the opportunity to hear and decide cases. By eliminating the need to worry about either reappointment or a postjudicial career, the framers took strong steps to ensure the formal independence of the U.S. judiciary.

A variety of studies confirm the fact that, for the most part, there is little evidence that the overt preferences of other branches of government affect the decisions of the U.S. Supreme Court. In the most notable study, Segal (1997) finds no evidence that the U.S. Supreme Court is responsive to changing congressional preferences in its decision making. Likewise, Owens (2010) finds a similar effect as the Court selects its docket.

Still, other studies provide some evidence that the Court's *de facto* independence is not unshakable. Clark (2009), for example, documents an inverse relationship between the number of court-curbing bills (e.g., bills that attempt to limit the court's jurisdiction or to change the size of the court in an attempt to make it more amenable to the legislature's policy preferences) introduced in Congress in 1 year and the Court's propensity to use its power of judicial review to strike down a federal law in the following year. Likewise, the executive branch also plays an important role in the Court's work. The U.S. Supreme Court frequently asks for the formal opinion of the executive branches on the cases it is reviewing by inviting the Solicitor General to file an *amicus* (friend of the court) brief in a pending case or by inviting the Solicitor General to participate in oral argument. When the executive branch participates in a case, it is very successful; the Court is more likely to hear a given case after the executive branch has recommended review, and the Solicitor General is the winningest attorney that appears before the Court. Thus, there is some evidence that suggest limits to the Court's independence.

Judicial Independence and Judicial Elections

The literature on courts in the American states is replete with concerns about judicial independence. The American states are unique in that their judges, unlike every other judicial system in the world except the Bolivian national courts, are selected and retained through popular elections. The American states began to adopt judicial elections in the mid-1800s as a response to legislative overspending; judicial elections, the drafters of state constitutions believed, would reduce legislative deference to the legislative branch, empower courts to use the power of

judicial review to curb legislative overspending, and increase levels of judicial independence (Shugerman, 2012). Similar arguments were made in Bolivia with the ratification of their 2009 constitution; judicial elections were initially adopted in both circumstances as a way to empower courts and increase judicial independence from the traditionally elected branches of government.

In the United States today, though, judicial elections are seen as a mechanism to promote judicial accountability (Nelson, 2011). Because one major function of popular elections in a democratic political system is to promote accountability by encouraging elected officials to heed public will and suggesting that voters should incorporate retrospective evaluations of candidate performance as one element of their voting calculus, opponents of judicial elections argue that these institutions weaken the independence of the judiciary by making judges over-reliant on public opinion. Supporters of judicial elections would counter that argument with one made by some of the leading researchers on the subject: "at the state level, law should represent public preferences and political culture, as long as the mandates of the United States Constitution and other federal law is [sic] observed" (Bonneau and Hall, 2009: 15).

The empirical literature on the effects of judicial elections on the independence of the judiciary suggests that all judges – regardless of how they are selected and retained – respond to some sort of external influence, but the method that is used to select and retain a judge changes the external influences to which a judge responds. Scholars have determined that judges who are popularly elected tend to follow public opinion more closely in their decisions on high-profile issues (such as the death penalty) than judges who will never stand for election (Brace and Boyea, 2008). Additionally, research suggests that the type of election that is used conditions the extent to which judges are congruent with public opinion in their decision making. For example, judges who run in elections that do not provide information on their party affiliation to voters on the ballot tend to follow public opinion more closely than judges who run in partisan elections (Caldarone et al., 2009). Still, the literature provides little reason to believe that appointment systems are a panacea for judicial independence. Indeed, Johnson (2014) shows that judges who never face election, but are appointed by governors or legislators, are less likely to rule against executive power when given the chance. Thus, it seems that judicial selection is less a question about judicial independence from all external influences but rather a question about from which influences one wishes judges to be independent.

Conclusion

Judicial independence is an important concept in an array of work on comparative and domestic judicial politics, and scholars have developed a number of different measures of the concept. Yet, while each measure purports to measure one part of the same concept, there are significant empirical and theoretical limitations among the measures; some measures capture concepts other than judicial independence, many of the measures make strong assumptions about the nature of judicial independence, and many measures of seemingly the same concept are correlated with each other at very low levels. Thus, future research should continue to unpack the concept of judicial independence at a theoretical level and work to create new, theoretically grounded, measures of the concept.

Future work on judicial independence will continue the work of Rios-Figueroa and Staton (2014) to determine the extent to which empirical indicators of judicial independence actually capture the concept they purport to measure and to examine the relationships between various empirical indicators of judicial independence. Rios-Figueroa and Staton's findings that measures of *de facto* and *de jure* judicial independence are quite different from one another should give scholars pause about conflating the two concepts. Future work on the topic should examine the conditions under which *de facto* and *de jure* judicial independence are closely related within a country and the conditions under which they are not.

See also: Judicial Behavior.

Bibliography

Acemoglu, Daron, Johnson, Simon, Robinson, James A., 1991. The colonial origins of comparative development: an empirical investigation. American Economic Review 91, 1369–1401.

Barro, Robert J., 2000. Inequality and growth in a panel of countries. Journal of Economic Growth 5, 5–32.

Bonneau, Chris W., Hall, Melinda Gann, 2009. In Defense of Judicial Elections. Routledge.

Brace, Paul, Boyea, Brent D., 2008. State public opinion, the death penalty, and the practice of electing judges. American Journal of Political Science 52, 360–372.

Caldarone, Richard P., Canes-Wrone, Brandice, Clark, Tom S., 2009. Partisan labels and democratic accountability: an analysis of state supreme court abortion decisions. Journal of Politics 71, 560–573.

Cingranelli, David L., Richards, David L., 2008. The Cingranelli-Richards (CIRI) Human Rights Dataset. http://www.humanrightsdata.org.

Clark, Tom S., 2009. The separation of powers, court curbing, and judicial legitimacy. American Journal of Political Science 53, 971–989.

Feld, Lars P., Voigt, Stefan, 2003. Economic growth and judicial independence: cross-country evidence using a new set of indicators. European Journal of Political Economy 19, 497–527.

Henisz, Witold J., 2000. The institutional environment for economic growth. Economics and Politics 12, 1–31.

Howard, Robert M., Carey, Henry F., 2004. Is an independent judiciary necessary for democracy? Judicature 87, 284–290.

Johnson, Gbemende, 2014. Judicial deference and executive control over administrative agencies. State Politics and Policy Quarterly 14, 142–164.

Keith, Linda Camp, 2002. Constitutional provisions for individual human rights: are they more than mere window dressing? Political Research Quarterly 55, 111–143.

La Porta, Rafael, Lopez de Silanes, Florencio, Pop-Eleches, Christian, Shleifer, Andrei, 2004. Judicial checks and balances. Journal of Political Economy 112, 445–470.

Marshall, Monty G., 2010. POLITY IV Project: Dataset Users' Manual. Center for Systemic Peace.

Moustafa, Tamir, 2007. The Struggle for Constitutional Power. Cambridge University Press, New York.

Nelson, Michael J., 2011. Uncontested and unaccountable? Contestation rates in trial court elections. Judicature 10, 208–217.

North, Douglass C., 1990. Institutions, Institutional Change and Economic Performance. Cambridge University Press, New York.

North, Douglass, Weingast, Barry, 1989. Constitutions and commitment: the evolution of institutions governing public choice in 17th century England. Journal of Economic History 49, 803–832.

North, Douglass, Summerhill, William, Weingast, Barry, 2000. Order, disorder, and economic change: Latin America vs. North America. In: de Mesquita, Bruce Bueno (Ed.), Governing for Prosperity. Yale University Press, New Haven, CT.

Owens, Ryan J., 2010. The separation of powers, judicial independence, and strategic agenda setting. American Journal of Political Science 54, 412–427.
Political Risk Services, 2010. International Country Risk Guide Methodology. The PRS Group, Inc.
Powell, Emilia J., Staton, Jeffrey K., 2009. Domestic judicial institutions and human rights treaty violation. International Studies Quarterly 53, 149–174.
Ramseyer, J. Mark, Rasmusen, Eric B., 2001. Why are Japanese judges so conservative in politically charged cases? American Political Science Review 95, 331–344.
Reenock, Christopher, Staton, Jeffrey K., Radean, Marius, 2013. Legal institutions and democratic survival. Journal of Politics 75 (2), 491–505.
Rios-Figueroa, Julio, Staton, Jeffrey K., 2014. An evaluation of cross-national measures of judicial independence. Journal of Law, Economics & Organization 30 (1), 104–137.
Stiftung, Bertelsmann, 2010. Transformation Index of the Bertelsmann Stiftung 2010. Center for Applied Policy Research.
Segal, Jeffrey A., 1997. Separation-of-powers games in the positive theory of congress and courts. American Political Science Review 91, 28–44.
Shugerman, Jed, 2012. The People's Courts. Harvard University Press, Cambridge.
Tate, C. Neal, Keith, Linda Camp, 2009. Conceptualizing and Operationalizing Judicial Independence. Working Paper.

Relevant Websites

http://www.humanrightsdata.org.
http://polisci.emory.edu/faculty/jkstato/page3/index.html – Judicial Independence Data by Jeffrey K. Staton.
http://www.prsgroup.com – The PRS Group.
http://www.systemicpeace.org/polity/polity4.htm – Polity IV Project.

Judicial Review

Jesse Merriam and Joel B Grossman, Johns Hopkins University, Baltimore, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.B. Grossman, volume 12, pp. 8020–8024, © 2001, Elsevier Ltd.

Abstract

Judicial review may be defined as the institutional capacity and authority of courts of law to assess the validity of actions taken by both lower courts and coordinate branches of government. The following article traces the doctrine's historical origins, examines its sometimes tense relationship with theories of democratic governance, explores its recent globalization beyond the United States, and concludes with discussion of contemporary debates concerning the judicialization of politics.

Judicial review may be defined as the institutional capacity and authority of courts of law to assess the validity of actions taken by both lower courts and coordinate branches of government. In the United States, judicial review has been associated historically with Supreme Court decisions regarding the allocation of powers among the branches of the national government, and between the national and state governments. In the last half of the twentieth century, the federal judiciary also used the power of judicial review to play a prominent role in protecting individual rights. Most modern nation-states now incorporate some variant of this power into their laws and constitutions.

Historical Origins

Judicial review is an outgrowth of efforts to limit the scope of power held by those authorized to interpret both sacred and secular texts. Societies, particularly (but not exclusively) those with complex governing structures, have created systems of law that constrain or influence official and social behavior through expressed codes or instructions. These, in turn, are overseen by special, trained officials. The fear that written laws and those empowered to enforce them will subvert the basic structures and values of a society form a prominent chapter in the history of self-government. Enabling government while protecting against its abuses is both an age-old and universal concern.

In the Judeo-Christian tradition, for example, one finds such concerns in the books of Exodus, Leviticus, and Kings I and II. There, following the example of the primary lawgiver, Moses, the prophets often contested the interpretations of both sacred and civil law provided by the kings and other secular rulers. Likewise, intellectuals in ancient China engaged in an extensive debate over whether *li* (ritual) or *fa* (law) should govern human behavior. Mo Tsu, one such intellectual, argued that government is necessary to establish and govern meanings, including those pertaining to human conduct. Other ancient Chinese intellectuals countered that the attempt to codify essential rituals or forms of conduct would lead to sophistry in political life. The prince's advisors, it was reasoned, would confuse the prince with various interpretations of essential political terms (Pocock, 1971).

Tension with Democracy

In modern democratic polities, the dilemma presented by judicial review is often referred to as the 'countermajoritarian difficulty' (a phrase famously coined by Alexander Bickel). This term recognizes the tension between democracy and the rule of law. In many legal systems, judges are not elected through the political process, but are rather appointed by elected officials (often for life), thereby giving the judiciary some, though by no means an absolute, independence from political influences. Because courts are thus somewhat insulated from political pressures, many commentators have seen judges as ideal guardians for protecting minorities from the tyranny of the majority and for preserving a political system's fundamental principles and values. On this ground, British jurist A.V. Dicey famously contended that an impartial and independent judiciary is central to the rule of law. Yet because courts are often not majoritarian democratic institutions, their decisions and interpretations of laws and constitutions may be at variance with popular values. There is always the fear that, in the guise of interpretation, a judge will substitute his or her personal values for those of a democratic majority or the preferences of a previous ruling party.

The emergence and practice of judicial review in the United States provide a textbook example of this continuing tension between democracy and the rule of law. Although the US Constitution does not explicitly mention judicial review, the founders were quite familiar with the practice. The doctrine had precedents in English law, the most famous of which is *Bonham v. College of Physicians* (1610), also known simply as *Dr Bonham's Case*. That case arose after the Royal College of Physicians sought to punish Dr Thomas Bonham for practicing medicine without a license issued by the college. The Court of Common Pleas ruled that even though an Act of Parliament had authorized the college to act in this way, Dr Bonham still could prevail in his suit challenging the legality of the punishment, because, in the words of the court's chief justice, Sir Edward Coke, "when an Act of Parliament is against common right and reason, or repugnant, or impossible to be performed, the common law will control it, and adjudge such an Act to be void."

Drawing from this principle, the US Supreme Court held almost 200 years later in *Marbury v. Madison* (1803) that "it was the province and duty of the judicial department to say what the

law is," and that, under a written constitution of enumerated powers, the Court could not enforce a law repugnant to it. The facts of the *Marbury* case, springing from the acrimonious 1800 presidential election and the resulting showdown between the outgoing Federalists and the incoming Democrat-Republicans, underscore the political dimensions of judicial review.

After Thomas Jefferson defeated John Adams in the 1800 election, Adams spent his last days in office seeking to fill as many governmental positions as possible with Federalists. To this end, Congress created and Adams appointed many justices of the peace in Washington, DC. But before the last of these commissions could be delivered, Jefferson's tenure as president began, and Jefferson promptly ordered his secretary of state, James Madison, not to deliver any of the remaining commissions. William Marbury, one of the appointees yet to receive his commission, brought a suit in the Federalist-dominated US Supreme Court to order Madison to deliver the papers. Chief Justice John Marshall, a leading Federalist whom Adams had appointed to the Court in his final months as president, began the Court's opinion by noting that Marbury was indeed entitled to his commission. Nevertheless, Marshall continued, the Court did not have the authority to mandate its delivery, because the Constitution did not give the Court original jurisdiction over the matter. Therefore, Marshall concluded, the federal statute that authorized Marbury to bring the suit to the Supreme Court was invalid.

Before the *Marbury* decision, state courts had exercised this power of judicial review, but until that point no court had declared the power so forcefully as Marshall did; he held that the US Constitution is supreme over ordinary statutes and, moreover, that the Constitution specifically makes the judiciary the ultimate arbiter of its meaning. The *Marbury* decision thereby carved the path for the federal judiciary to invalidate any governmental action inconsistent with its understanding of the Constitution. In locating this judicial power in the US Constitution, Marshall pointed to many sources, such as the general structure of the document, specific provisions dealing with the judiciary's duty to uphold the Constitution as the supreme law of the land, and Alexander Hamilton's claim in Federalist 78 that the judiciary must be able to use its power of persuasion to counter the executive's sword and the legislature's purse.

Many scholars have challenged Marshall's reasoning as vastly oversimplifying whether the Constitution does indeed contemplate a power of judicial review, let alone an exclusive power setting the court as the ultimate arbiter of the Constitution's meaning. But almost all scholars recognize the political acumen that Marshall displayed, refusing to enlarge the Federalists' power in that particular instance while at the same time significantly broadening the Supreme Court's power overall to invalidate legislative enactments it deems unlawful. By not mandating the commission, Marshall made it appear as though he was exercising judicial restraint and not furthering the Federalists' agenda. But in so doing, he expanded the authority of the only branch of the federal government, the judiciary, still dominated by Federalists. Jefferson saw through Marshall's plan, writing that judicial review was "a very dangerous doctrine indeed," but he had little choice but to accept the ruling, being that it gave him what he had wanted, the denial of Marbury's commission.

Given this danger of the judiciary invalidating an action taken by a coordinate branch of government, there is always the possibility that the other branches will use their powers over the sword and purse to ignore a judicial decision, and for this reason the Supreme Court has often exercised great caution in exercising judicial review. Indeed, the Court invoked the power sparingly in the nineteenth century, invalidating only a few laws as unconstitutional during that time. But in the twentieth century, as the doctrine gained acceptance among the citizenry, the scope and incidence of judicial review expanded rapidly. Today, it is a centerpiece of a 'judicialized' American political system. Nevertheless, although there is general agreement at the theoretical level about the utility of judicial review, there is still disagreement about its application in particular cases and the alleged perils of an 'imperial judiciary.' Indeed, a handful of scholars, most notably Alexander Bickel, Sandy Levinson, Mark Tushnet, and Larry Kramer, have contended that the nation would be better served by eliminating or at least vastly narrowing the power of judicial review, because, according to these thinkers, judicial review undermines the democratic process and this harm outweighs any benefits that might accrue from the doctrine (Bickel, 1986; Levinson, 1989; Tushnet, 2000; Kramer, 2004). One could say, then, that the battle between Marshall and Jefferson continues into the twenty-first century.

Globalization of Judicial Review

As Americans continue to debate the use and consequences of judicial review, it has taken root elsewhere. Many postwar constitutions in Europe, as well as India, South Africa, and Japan, incorporated some variant of it. And it now has been adopted, at least formally, by many of the nations of the former Soviet empire. Many of these nations had, in fact, once deliberately rejected it. Post-Napoleonic Europe, for example, firmly placed the judiciary under the control of the positivistic codes and enactments of the parliaments. And many of the Asian countries struggled throughout the nineteenth and early twentieth centuries to develop independent judiciaries that would be 'acceptable' to the United States and the other Western powers (Grovogui, 1996; Millard, 1931). The rise of authoritarian and totalitarian regimes with little regard for basic human rights in the twentieth century led to an 'Americanization' of the governing regimes in Turkey, the Philippines, Japan, France, and Germany after World War II. All of these nations have joined many others in adopting some measure of constitutional and liberal reform in their political arrangements, reforms that include an expressed power of judicial review. In Africa, judicial review and 'constitutional courts' have taken root in places like Benin, South Africa, Ghana, and many other countries where, previously, the judiciary was almost completely subservient to the executive or legislature (Shapiro and Stone, 1994; Prempeh, 1999).

In the United Kingdom, which has no written constitution, no bill of rights, a strong tradition of parliamentary supremacy, and no formal mechanism for judicial review, in the same period that the United States was undergoing a transition toward a more 'activist' judiciary with respect to civil rights, British judges increasingly policed the administrative realm for violations of governing statutes. Some scholars have looked to

internal factors, such as intergovernmental support and rights advocacy organizations, as well as the influence of the increasingly rights-oriented and litigious culture of the United States, to explain the development and persistence of what might be called clandestine review (Shapiro and Stone, 1994; Epp, 1998).

Nevertheless, courts in Europe, Africa, and Asia have retained much of their traditional character and have not adopted a full-fledged version of American judicial review. This is particularly true of countries with civil code rather than common law systems. Most European constitutional courts, for example, are special 'constitutional' courts that hear cases principally brought by the government, although this is not the case for all European nations; Germany, for example, has a special procedure to hear lawsuits brought by private litigants. In France, a civil code country, the Constitutional Council may review legislation only in the immediate time period after passage and prior to promulgation. Moreover, the fact that members of both the Assembly and Senate can, with great ease, bring a constitutional challenge to legislation has further blended the edges separating legislating from adjudicating. The legislator proposes and debates a law 'as if' a judge will hear the same; and judges write decisions 'as if' the audience will be a legislator. Whether the parliament is the creature of the judge or the judge is the creature of the parliament is in all respects an important question (Shapiro and Stone, 1994: pp. 400–402).

Characteristics such as those retained by the French point to important departures from American cultural and jurisprudential conventions. The US Supreme Court has never viewed itself as a court of 'instruction.' Indeed, the Court has consistently held that it lacks the power to issue merely advisory opinions; pursuant to Article III of the US Constitution, federal courts may adjudicate cases only in the context of an adversarial 'case or controversy' proceeding. Organizing principles of government such as federalism and the separation of powers also vary widely in their acceptance, meaning, and application, thereby insuring many variants of judicial review. Recent research shows that over the past decade, foreign courts have been citing American courts less and less frequently, and many nations have been moving away from the American structure of governance. Some scholars attribute both phenomena to the United States' declining influence and status as a hegemonic power. If the United States' power abroad further declines, as many predict it will, the variations of judicial review will likely increase, making the massive power employed by the US Supreme Court through judicial review even more of an anomaly.

Judicialization of Politics?

Although few nations have adopted the robust form of judicial review currently practiced by courts in the United States, it is clear that judicial review has become a global phenomenon, leading some social scientists to speak of a 'judicialization' or 'juridicalization' of politics. At its root, the thesis suggests that political institutions such as legislatures and administrative agencies have come to expect and even welcome courts as arbiters or overseers of their actions. They tailor their actions to accord with an expected response by the judiciary (Stone Sweet, 1999: p. 164; Shapiro, 1994). The electorate itself has come to have similar expectations. In the American political context, for example, judicialization has had the paradoxical effect of legitimizing, *de facto*, the role of judicial review in a democratic process. While most Americans lack a deep knowledge of American history, politics, or law, the judicialization of politics has enabled them to speak the language of rights. Many Americans can, with some measure of accuracy, defend or dispute the validity of the legal holdings involved in such famous cases as *Brown v. Board of Education* (1954), *Miranda v. Arizona* (1966), and *Roe v. Wade* (1973) (Friedman, 1990; Grossman and Epp, 1991). The judicialization of politics has also led many Americans to believe that the US Supreme Court is and should be the final arbiter of legal disputes and that this does not contravene any basic democratic values, despite the fact that federal judges are not elected. Judges, in this view, are the lesser of many evils in a world where state and national institutions have come to be mistrusted. What citizens oppose most are decisions with which they disagree!

Outside of the United States, the tension between judicial review and democracy has been even less of an issue in the public eye. This is due, without doubt, to the association of judicial review with postauthoritarian reforms and constitutions that establish the primacy of fundamental human rights principles and entrust their protection to constitutional courts. The rise of supranational courts, such as the European Court of Justice, has not only created a second level of review that transcends the courts and laws of individual countries, but also added considerable force to the argument that the enforceability of basic rights is critical to the maintenance of democracy and a civilized world order.

The legal philosopher H.L.A. Hart pointed out years ago that judicial review, or some variant of it, is going to be a problem to any political community in which those who are authorized to make or interpret the laws governing the community are institutionally distinct from the community at large. In such cases, the conflict of law is often a matter of making both first-order laws, such as constitutions, electoral systems, and branches of government, and secondary laws that guide and constrain rule making (Hart, 1961). In less theoretical terms, the nature of the legal system becomes itself part of the political vocabulary of the system. This, after all, is what the 'judicialization' of politics is about: political actors adopting legal and constitutional discourse to express their political views. Governing structures and particular social and political conflicts are continually readjusted to each other as officials and some segment of the public alike discuss what is taking place and whether it conforms to basic political and legal arrangements.

Contemporary Debate

While the American literature on judicial review might seem to be of immediate and obvious interest to legal reformers in Europe, Asia, and Africa, it remains to be seen whether the American experience (or the 'countermajoritarian difficulty,' for that matter) will be of deep interest in those places. Rather, as judicial review has become less a species of American exceptionalism and more of a global practice, current scholarship has turned increasingly to a hybrid of the American debate. At issue is

the empirical question of *why* judicial review develops in some times and places and whether, normatively speaking, judicial review may be better employed in some systems than in others.

Some scholars have argued that for judicial review to take root, there must be some level of commitment to the rule of law, and to government that is both democratic and limited. In polities where pure majoritarianism prevails, there is, of course, no necessary limit to what the legislature can do in the electorate's name. Where the government is limited by some overarching set of principles, the ability of courts of law to intervene and veto legislation in apparent conflict with such principles is more pronounced (Shapiro, 1998). In a twist on the American experience, then, the inversion of the relationship of rights-to-the-state in regimes throughout Europe after World War II has enshrined their constitutional courts as defenders of preexisting and constitutive rights held by the citizenry. Many of these rights find their sources not merely in written constitutions (such as in the German Basic Law's protections of 'human dignity') but also in sources both internal and external to the political system (e.g., treaties and conventions) (Shapiro and Stone, 1994). Conversely, in some African countries like Ghana, the long-standing tradition of judicial deference to executive authority has led judges to use their new powers to adopt a jurisprudence seemingly at odds with the more recent, liberal written constitutions. In these nations, commentators have suggested, new 'constitutional courts' with judges learned in foreign jurisprudence and committed to constitutional reform will have to be substituted for a judiciary that is, in large part, a holdover from the *ancien régime* (Prempeh, 1999).

Other oft-cited factors influencing the development of judicial review include the existence of commerce and internal geographic and cultural divisions, such as federalism. Commercial actors conventionally use legal devices to secure both formal agreements and sanctions should the terms of those agreements be violated. Written law and the search for a neutral third party who is an expert in such law have long been important parts of trade. Social scientists have argued that commerce was of great importance in accounting for the historical emergence of judge-made law. The origins of the professional lawyer, Max Weber argued, are to be found in the *cavere*, the Roman practice of drafting contracts, wills, and other legal devices. The lawyers, seeking to make commercial relations more predictable for their clients, would change the meanings of extant legal terms or invent new legal devices and then seek out judges who would institutionalize the changes as authoritative (Weber, 1954: p. 72).

The existence of federative or federal government is also cited as a factor in how judicial review develops in a given country. Horizontal divisions of power among disparate electoral, legislative, and executive bodies, as well as vertical divisions of power, invariably occasion conflict that is itself caused by, and therefore cannot be corrected easily by, 'normal' political procedures. In countries experiencing such conflict, the judge might be seen as the best means of clarifying or mediating the existing division of power among these governmental entities, with the judge taking on the role of a theorist of the state and its basic arrangements (Dworkin, 1986).

Supranational courts, such as the European Court of Justice, have added yet another layer of review. That Court, for example, has claimed a sweeping power of judicial review for itself by converting the various treaties it oversees into a sort of overarching constitution binding on both member nations and the courts and citizens of those nations. Such a practice undermines the close connection between judicial review and national sovereignty, and it may also contravene basic principles of international law, at least in theory. Yet the European Court has been successful in part because it provides stability and some manner of coherence, while also balancing national sovereignty interests with the goal of nation-state integration.

Scholars also cite more temporally specific factors as relevant to the development of judicial review. In the United States, commercial elites have dominated constitutional litigation for most of US history. But this changed substantially in the postwar era, when nontraditional litigants were able to pool time, monetary resources, and available legal talent to press for important constitutional guarantees (Epp, 1998). And the US Supreme Court, over time, 'adjusted' the Constitution to facilitate the development and legitimacy of the American welfare state. In Europe, somewhat differently, the devolution of legislative power in parliamentary systems and the concomitant rise in cabinet- and party-enforced government have occasioned a role for courts, particularly in those countries (e.g., Germany) where even a governing majority cannot prevent legislation from being referred to the constitutional court (Shapiro and Stone, 1994: p. 405).

French social-historian and philosopher Michel Foucault argued that, since the early modern period, juridical institutions and discourses have been displaced gradually by administrative institutions and their discourses. The world of sovereigns and their landed jurisdictions had given way to a world of populations and institutions to 'normalize' these populations (Foucault, 1978). If we take both the American experience and the recent proliferation of legal and constitutional norms in Europe and elsewhere as our guide, the thesis would have to be revised to account for the appearance of administrators who come garbed as judges. The influence of legal and constitutional norms, vocabularies, and institutions on the socioeconomic and political structure of much of the planet has been considerable; the effects of this process are, however, to date poorly understood and in need of further study. In this respect, 'postmodern' arguments have to be seriously qualified. Whether justified in fact and theory, the expectation that judges and legal norms can ground individual rights and provide a constitution for transnational and intercultural contact is shared increasingly today. The very American *Marbury*, as it were, has found a place for itself in the global environment.

See also: Balance of Power, History of; Constitutionalism, Comparative; Courts and Adjudication; Democratic Theory; Power; Supreme Courts.

Bibliography

Bickel, A., 1986. The Least Dangerous Branch: The Supreme Court at the Bar of Politics. Yale University Press, New Haven, CT.

Dworkin, R., 1986. Law's Empire. Belknap Press of Harvard University Press, Cambridge, MA.

Epp, C.R., 1998. The Rights Revolution: Lawyers, Activists, and Supreme Courts in Comparative Perspective. University of Chicago Press, Chicago, IL.
Foucault, M., 1978. The History of Sexuality. In: An Introduction, vol. 1. Pantheon Books, New York.
Friedman, L., 1990. The Republic of Choice. Harvard University Press, Cambridge, MA.
Grossman, J.B., Epp, C., 1991. The reality of rights in an 'atolerant' society. This Constitution, 20–28.
Grovogui, S.N.Z., 1996. Sovereigns, Quasi-Sovereigns, and Africans: Race and Self-Determination in International Law. University of Minnesota Press, Minneapolis, MN.
Hart, H.L.A., 1961. The Concept of Law. Clarendon Press, Oxford, UK.
Kramer, L., 2004. The People Themselves: Popular Constitutionalism and Judicial Review. Oxford University Press, New York, NY.
Levinson, S., 1989. Constitutional Faith. Princeton University Press, Princeton, NJ.
Millard, T.F., 1931. The End of Extraterritoriality in China. The A.B.C. Press, Shanghai, China.
Pocock, J.G.A., 1971. Politics, Language, and Time. Atheneum, New York.
Prempeh, H.K., 1999. New jurisprudence for Africa. Journal of Democracy 10 (3), 135–149.
Shapiro, M., 1994. Juridicalization of politics in the United States. International Political Science Review 15 (2), 101–112.
Shapiro, M., 1998. The European court of justice: of institutions and democracy. Israel Law Review 32 (1), 3–50.
Shapiro, M., Stone, A., 1994. The new constitutional politics of Europe. Comparative Political Studies 26, 397–420.
Stone Sweet, A., 1999. Judicialization and the construction of governance. Comparative Political Studies 32, 147–184.
Tushnet, M., 2000. Taking the Constitution away from the Courts. Princeton University Press, Princeton, NJ.
Weber, M., 1954. In: Rheinstein, M. (Ed.), On Law in Economy and Society. Harvard University Press, Cambridge, MA.

Jung, Carl Gustav (1875–1961)

Angela Graf-Nold, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Carl Gustav Jung (1875–1961) was one of the main founders of psychoanalysis and general depth psychology, beneath Sigmund Freud and Alfred Adler. He was copy editor of the first academic psychoanalytic periodical *Jahrbuch für psychopathologische und psychoanalytische Forschungen* (ed. by Bleuler and Freud) and founder and president of the International Psychoanalytic Association (1910–13). With an academic background from the Zurich school of psychiatry (Burghölzli clinic, University of Zurich), encounters with Freud, and in confrontation with his own unconscious processes, Jung developed his own psychological approach under the term *analytical psychology*. His research was centered around the nature and the typical constellations of the psychic functions (reasoning, feeling, sensation, intuition) and their relations to the individuation process, as well as the nature of the unconscious and its relation to consciousness (archetypes, cross-cultural patterns of the individuation process, synchronicity).

Jung was born in Switzerland at Kesswil, near Lake Constance, on 26 July 1875, son of the parson Paul Jung and his wife, Emilie B. Preiswerk; in 1902 he married Emma Rauschenbach (1882–1956), daughter of a wealthy manufacturer, with whom he had five children, one son and four daughters. He died at the age of 85 on 6 June 1961, in his house in Küsnacht, Lake Zürich.

In 1895, Jung began his medical studies at the University of Basel. The sudden death of his father in 1886 caused some financial troubles, and left him to take care of his mother and sister Gertrud, who was 9 years younger than he was. His poverty contrasted with the continued reputation of both of his late, well-known grandfathers, both fathers of 13 children and both professors at Basel University. Carl Gustav Jung, Senior (1794–1864), a political refugee from Mannheim, Germany, had been a professor of medicine and president of Basel University. Samuel Preiswerk (1799–1871), C.G. Jung's maternal grandfather, professor of theology and ancient Hebrew at the university, had served as the president of the reformed ministers in the Basel Canton.

After his exams in December 1900, Jung decided to leave Basel and follow a professional career as a psychiatrist, beginning as a second assistant at the psychiatric hospital Burghölzli in Zurich. At the time, the Burghölzli asylum, southwest of the city of Zurich, was under the direction of Eugen Bleuler; it was then the most 'modern' institution for mentally ill persons, was affiliated with the University of Zurich, and its director was at the same time also professor of psychiatry at the university.

The context of the work at Burghölzli asylum is visible in Jung's later work. The reputation of Burghölzli asylum was the merit of Bleuler's predecessor Auguste Forel, who during his 20 years directorship from 1879 to 1898 had established the professional standards of the daily routine in the clinic and had set legal standards for admission, treatment, and dismissal of the patients; with regard to the discipline of psychiatry, he outlined and represented an unbiased empirical approach to mental phenomena, focusing as well on the neurological, psychological, and social aspects of mental disorders (bio-psycho-social model of mental illness/Zurich school of psychiatry). At the beginning an ardent pioneer in neurology, the everyday misery of his patients soon led him to abandon neurological research; instead, Forel (a Swiss francophone) introduced hypnosis (as practiced by the 'French School of Nancy' of H. Bernheim) as a first psychotherapeutic treatment for inpatients. In hypnosis, Forel found the proof for the existence of subconscious processes that were equivalent to externally observable neurological phenomena (scientific monism). His textbook on hypnotism became very famous and is still available in print (Forel). In 1898, Forel quit his job in order to dedicate himself further to the prevention of mental illness; he wrote an international bestseller about sexual education, *The Sexual Question*, 1905, and internationally got involved in promoting alcohol abstinence. Still very active as an organizer, author, lecturer, and psychiatric expert, he stayed a dominant figure in Swiss psychiatry, and was a critical observer of the activities at the Burghölzli clinic where Jung soon began to play a dominant role.

Bleuler, who was director of the clinic when Jung started, wholeheartedly shared Forel's conception of mental illness. As Forel's assistant, he had supported him in his pioneering neurological research, then he had participated very actively in the exciting introduction of hypnosis at the Burghölzli 1885/86. In 1886, he took over the directorship of the remote asylum Rheinau, an institution for about 400 chronically, mentally ill patients; there he lived as a bachelor with only one medical colleague and the housekeeper's family together with the patients.

As director of the Burghölzli clinic, Bleuler obviously appreciated the end results of his scientific and personal isolation. He immediately established a scientific debating club for his medical staff and other colleagues, which met weekly, alternating in the Burghölzli and in the private clinic of his colleague, the Russian–Swiss neurologist C. von Monakow. This circle constituted an important part of the 'creative scientific community' (Loewenberg, 1995) of the Burghölzli staff; the debates about Freud and psychoanalysis began here, and finally psychoanalytic interpretations were applied not only to patients' behavior and dreams but also mutually to colleagues and friends.

It was this environment that Jung experienced at Burghölzli hospital; it was quite a shock and like an entrance in a "monastery of the world." Four medical doctors, beneath Bleuler, as well as a secondary doctor and one ordinary assistant, were in charge for about 400 mentally ill patients; he and his colleagues lived as interns in single rooms; the director and the secondary doctor had apartments, and about 80 waiters slept on foldaway beds in the patients' rooms or in the corridors and could go home only once every two weeks.

"My inability to understand gave me such a sense of inferiority, that I couldn't bear to go outside the hospital" (Jung, 1925/1989). He read through the available literature and dedicated himself to the study of his cases. Soon he began his dissertation, 'On the psychology and pathology of so-called occult phenomena.'

1902

In this doctoral thesis, Jung studied his spiritualistic experiments with his cousin Hélène Preiswerk, which he had carried out with her at the end of his medical studies, and treated her case in the context of (hysterical) patients of the Burghölzli hospital. In a quite minor context, he referred to Freud and Breuer's *Studies on Hysteria* and to Freud's *Interpretation of Dreams*, but mainly he relied on the new studies of Pierre Janet and Théodore Flournoy. Flournoy, professor of psychology in Geneva, had just published a scientific best-seller *Les Indes à la Planète Mars* (1900), a long-term study of a spiritualistic medium. Jung got in touch with Flournoy and showed special appreciation for his 'objective-distanced' approach to mental phenomena (which was influenced by his personal friend William James) and his having arranged the new findings in a bigger context. Jung's untroubled friendship with Flournoy outlasted the much more famous, but never as untroubled, friendship with Freud, which began several years later.

In October 1902, Jung quit his strenuous job at Burghölzli. He went to Paris to study at the Collège de France with Pierre Janet and also learned English at the Berlitz School. In 1903, he married 21-year-old Emma Rauschenbach, the daughter of a factory owner, and settled in Zurich. At the end of 1903, he returned to Burghölzli hospital, initially as an external representative for the resident doctor who had fallen ill. In December 1904, he finally took over the regular post of resident doctor, and moved into the doctor's family apartment at the hospital. Three of his five children were born there and grew up together with Bleuler's family. Bleuler had also recently married the philologist and writer Hedwig Waser, one of the first academic women in Switzerland. Both women shared the interests of their husbands and took part in their scientific discussions; doubtless, there also existed a common role expectation for both wives to support their husbands in their professional duties.

This 'participation of everybody' was due to Bleuler's ideal of an 'active community' (*tätige Gemeinschaft*). Bleuler's previous experience as director of the remote Rheinau asylum, with the close contact with mostly 'incurable' chronically mentally ill patients, had led him to emphasize the obvious therapeutic value of an 'active community,' in which everybody worked according to their abilities in meaningful activities. For his special scientific research interest in this field – the dynamics of psychotic behavior – he favored a hypothesis of neurological conditions that caused a specific discontinuity in the forms of association.

Therefore, it was with some enthusiasm that, in 1904, Jung and Bleuler set up a well-equipped psychological laboratory for association experiments. These played a prominent part in psychological research at the time, especially in the school of Wilhelm Wundt at the University of Leipzig (Germany). Their findings soon led to passionate debates in the debating club. In giving accounts not only of the regularities of the test persons' responses to the stimulus words, as was usual in the Wundtian approach, but also to the 'intruders in the mind,' the irregularities (delays of reaction times, repetitions, and other disturbances), Jung and his coworkers (initially Franz Riklin, Sr., Jung's wife, Ludwig Binswanger, and several volunteers) identified the 'emotional-toned complexes' (*gefühlsbetonte Komplexe*) of the test persons as autonomous expressions of their repressed or dissociated unconscious conflicts. These results were particularly decisive for Jung. Behind completely 'mad' behavior he discovered quite normal hopes and desires. Paranoiac ideas and hallucinations revealed their hidden sense in the context of the patient's life stories. "At the bottom we discover nothing new and unknown in the mentally ill; rather, we encounter the substratum of our own nature," he stated. "That was at that time a very moving experience" (Jung and Jaffé, 1963).

When Jung was preparing a paper for his application as lecturer at Zurich University in 1905, he realized that his findings threw a new light on Freud's theory of repression. He decided to stand up for him, despite Freud's bad reputation in academic psychiatry. In 1906, Jung published a first volume of Association Studies, which he sent to Freud in Vienna. In 1907, he published *On the Psychology of Dementia Praecox*, which included a detailed case study with association experiments of an old female patient of the clinic, who since Forel's time had served as a demonstration case for students. These two publications inaugurated a remarkably passionate personal relationship with Freud, which is documented in hundreds of letters (Freud and Jung, 1974). Bleuler shared a special enthusiasm for Freud, although, like Jung, he was always somewhat critical about Freud's recently elaborated sexual theory and especially criticized Freud's dogmatic, sectarian tendencies as an impropriety toward the scientific community (Freud and Bleuler, 2013). In October 1907, Jung – against some opposition from Bleuler's side – founded an informal 'Freud Club' at the Burghölzli, which met independently from the other regular scientific staff meetings. In April 1908, Jung organized a first Psychoanalytical Congress (private meeting) in Salzburg. On this occasion, a first joint publication was founded – the *Jahrbuchfür psychopathologische und psychoanalytische Forschungen*, edited by Bleuler and Freud, and compiled by C.G. Jung. Thus Jung, with the decisive support of Bleuler, inaugurated the reception of Freudian psychoanalysis in academic psychiatry and the beginning of the so-called 'psychoanalytic movement.'

The culmination of – as well as the turning point in – the relationship of Freud and Jung was their joint trip in September 1909 to the 20th Anniversary Conference of Clark University, Worcester, Massachusetts. Both had to lecture and were honored with their first honorary degrees. Jung had been invited independently, because of traditional connections

between the Burghölzli hospital and Clark University through Forel's connections with his former pupil and friend Adolf Meyer, a Swiss immigrant to the United States, which played a dominant role in the formation of American Psychiatry. In the 6 weeks together with Freud during the trip to the United States, some strange personal behavior and some critical points in mutual dream interpretations inevitably revealed to Jung the limits of a friendly and cooperative relationship with Freud. Nevertheless, he continued to act for Freud as organizer and founder of the International Psychoanalytical Association, a role that he later remarked on as *collaboration à toux prix*. In 1910, he was elected president of the International Psychoanalytical Association. Shortly after Freud's provoked break with Alfred Adler in 1912, Freud challenged Jung's leadership with the organization of a 'secret committee,' consisting only of very 'confidential' followers – excluding Jung. Jung subsequently resigned as the editor of *Jahrbuch* in October 1913, and then as president of the International Psychoanalytic Association in April 1914.

In March 1909, some months before his trip to the United States with Freud, Jung quit his job at the Burghölzli clinic and opened a private practice as psychiatrist in Küsnacht, Lake Zurich. In the 1920s and 1930s, his popularity began to grow. After 1929 he was engaged in the General Medical Society for Psychotherapy (*Allg. Aerztl. Gesellschaft f. Psychotherapie*) and was an author and coeditor of the affiliated *Zentralblatt für Psychotherapie* (Berlin); after the resignation of the society's president Kretschmer in 1933 after Hitler's rise to power, Jung accepted the presidency of this organization and editorship of the affiliated *Zentralblatt für Psychotherapie*. He defended this compromising collaboration with German psychotherapists with his aim to reorganize the society as an international institution in order to help some colleagues maintain their functions, and to prevent a general abolition of psychotherapy as a medical discipline in Germany. After initial and partial success with these goals, Jung finally resigned in 1940 (Giovanni Sorge, 2012).

In 1933, Jung resumed his academic lecturing, as a lecturer for 'modern psychology' within the General Department of the Swiss Federal Institute of Technology (ETH), Zurich, he lectured courses in 'Modern Psychology' (psychology of the unconscious/the process of individuation). In 1934, he was awarded the title of professor; he resigned because of health reasons in summer 1941. From 1944 to 1945 he held a professorship in medical psychology at the University of Basel. In 1948, the Jung Institute in Zurich (later Küsnacht) was established by an international committee of some of his associates as a research and training center for Jungian analysts.

Analytical Psychology: Contribution to Science

In 1909, immediately after Jung returned from the United States, he immersed himself in archeological and mythological literature in order to differentiate his own view of the unconscious from Freud's. The result was his major work *Wandlungen und Symbole der Libido* (1912), later published in English as *The Psychology of the Unconscious* (1916). In contrast with Freud's personalistic view of the psyche – i.e., his attempt to reduce the unconscious to a hiding place of the personal, sexual conflicts of the dreamer – for Jung, sexuality was not the sole determining factor; rather, the main point was that the unconscious had an autonomous life, which reflected not only the personal, but also the common, collective history of the mind. The manifestations of the unconscious stood not only for a causal, reductive purpose, but also had a final, i.e., forward-looking, meaning. Relying on some products of automatic writing (two hymns and the dramatic myth of an Indo-American hero) produced by a young American, Miss Frank Miller (first published in 1905 in the *Archives de Psychologie*, ed. by Flournoy), Jung presented his own conception of the libido as a general 'psychic energy' with a dual nature, split into positive and negative poles with a steady flux between construction and de(con)struction. He also outlined an 'anatomy and embryology of mental development,' which followed common patterns of changing imaginary processes in the history of mankind, as well as in the personal history of each individual person. Wavering between an analysis of Miss Miller's fantasies, some hidden self-analysis, a detailed analysis of religious hero myths of humankind from Mithraism to Christianity, and an analysis of classical and romantic philosophy and literature, he developed his theory of the structure and energetic functioning of the psyche without yet coining a decisive terminology of his own.

In the years between 1912 and 1918, Jung experienced a time of personal insecurity and scientific isolation. He resigned from his post as lecturer at the university and then dedicated himself to his international practice as a psychotherapist, concentrating on his private studies and on his confrontation with his own unconscious. He documented these studies in black notebooks, elaborating on them with calligraphy and careful illustrations in a red leather–bound representative volume; in an epilogue from 1959 he stated that it was necessary to absorb the overwhelming power of his original experience until they found their way in reality with his acquaintance with alchemy, which allowed him to integrate them into a whole. This Red Book was recently published (Jung 2010, ed. by Sonu Shamdasani, 2010).

In 1921, Jung published his major book, *Psychological Types*, in which he linked his historical research with the findings of his self-analytical process. He outlined the key concepts that later became characteristic of analytical psychology such as 'types,' 'collective unconscious,' and 'individuation process.'

Jung had become interested in the question of psychological types because of the differences he perceived between Adler and Freud. With critical reference to historical material, Jung elaborated a model of typology, with eight psychological types – two main attitude types of extroverts and introverts, further subdivided by the predominance of one of four psychological functions: thinking, feeling, sensation, and intuition. While sensation conveys concrete reality through the senses, intuition conveys the hidden possibilities in the background. And while thinking leads to a judgment about the meaning of concrete reality, feeling leads to a judgment of its worth. Normally one attitude or function prevails (superior function) and the opposite (inferior) attitude or function remains undeveloped, more unconscious, and undifferentiated. The value of this model lies less in its contribution to differential psychology than in the theory of knowledge and the theory of the therapeutic process as a quest for wholeness.

From this experience of the relativity of every standpoint with respect to psychic reality, Jung highlighted the significance of the 'personal equation' as a prior condition of any psychological statements – i.e., the position that what a psychologist claimed to be true of human psychology as a whole was first and foremost a subjective confession. Additionally, he developed a conception of the therapeutic process, which involved not only becoming conscious of repressed contents of the unconscious but also integrating the psyche and developing a much more complete approach to inner and outer realities. This process of individuation involved withdrawing projections that are structured according to certain autonomous images of the unconscious (archetypes); these reflect the common history of mankind (archetypes of the collective unconscious) as well as the repressed images of one's own personal history (personal complexes). Common and ubiquitous archetypes are the shadow, the unconscious negative side of the personality; the hero, the idealized picture of the person; the anima, the animus; and the repressed female/male soul picture in men and women. This process provokes an ethical conflict in each patient; they have to take responsibility for their mental attitudes and develop a fuller approach to reality not only in ideas, but also in feelings, sensations, and intuition. As a result, the center of a person moves from the ego (the center of consciousness) toward the self (the center of the whole, individual, conscious and unconscious psyche). As therapeutic methods for this process, Jung developed active imagination (as a method of assimilating unconscious contents by several forms of self-expression) and amplification (circumambulation) of the symbols (instead of free association in Freudian analysis).

The foundation of the dynamic concept that became known as analytical psychology and the institutionalization of Jungian analysis were connected with the founding of the Psychological Club in 1916. This was made possible by the generous financial gift of Jung's American patient, Edith Rockefeller McCormick. After Jung's separation from the Burghölzli clinic and the Freudian circle, the Club reestablished an important debating forum in which he could communicate his ideas to interested patients, colleagues, and scientists. Here many of his former patients – among them his life-long special assistant and friend Toni Wolff, his secretary and the compiler of his biography, the psychologist Aniela Jaffé, and the philologist in ancient languages Marie Louise von Franz – were encouraged to take up public lecturing. In the Club he also met outstanding academics in his main subjects of interest, such as mythology, religious studies, philosophy, psychology, and archeology. Club lecturers included Karl Kerényi, pioneer in Greek mythology, the young Mircea Eliade, pioneer in the science of religions, Gershom Scholem, the German-born Israeli philosopher and historian, and other representatives of different Christian, Indian, Jewish, and Eastern religious traditions. Richard Wilhelm, missionary to China and sinologist, had a special impact on Jung. Wilhelm shared with him the old mystical Taoist text *The Secret of the Golden Flower*. It was in this text that Jung first discovered the correspondence of his individuation concept with Eastern philosophy and worked out his mandala symbolism for psychic wholeness or the self.

From 1933 on, the Club lectures and Jung's public lectures at ETH were supplemented by the Eranos Conferences, which were organized by Olga Fröbe, an English-Dutch spiritualist and scholar, every summer for a fortnight at her house at Lake Maggiore in Moscia, near Ascona, Switzerland. In the atmosphere of these meetings with contemporary pioneers and authorities in all fields of modern science and spirituality, such as the Jewish philosopher Martin Buber, the Indologist Heinrich Zimmer, the biologist Adolf Portmann, and many others, Jung presented and discussed his view of traditional Christianity and medieval alchemy, the archetypes of the collective unconscious, and the principle of synchronicity (Hakl, 2013).

An encounter with death in 1944 marked the beginning of the period of Jung's final elaboration of ideas. His main work on traditional Christianity, *Answer to Job* (1952), and his major work on alchemy, *Mysterium Conjunctionis* (Mystery of Conjunction; 1955) appeared at this time. The common principles of the reconciliation of the opposites of the dual nature of the psyche, which he had worked out in this major book, led to the final formulation of his main pioneering concept, *Synchronicity as a Principle of Acausal Relations* (1952), which he published in collaboration with Wolfgang Pauli, professor of physics at ETH. The idea of an ultimate psychoid archetype represents the connecting principle of psyche and matter, which are two aspects of the same thing and can realize themselves in both dimensions 'synchronistically.' The relative autonomy and unpredictability of the synchronistic phenomena correspond with the new findings of subatomic and cosmic physics and hint toward a metaphysical order to which both the psyche of the perceiving person as well as nature itself submit.

Contemporary Influence

In contrast with Freud, who gave the impression that psychoanalysis "may be likened to Athena's springing forth fully armed from the head of Zeus" (Jones, 1953–57), Jung always tried to trace back his ideas to older, universal traditions. His pioneering work was the rediscovery of an uninterrupted chain of psychic manifestations and spiritual traditions from antiquity, gnosticism over alchemy to the present. His early claim for a constructive, forward-looking meaning in unconscious manifestations led to some changes in early psychoanalytic treatment that are now common in psychotherapy, for instance, the symmetrical setting (instead of the patient lying down), focus on present conditions as responsible for the outbreak of neurosis, and focus on transference and countertransference as the crucial point for the change of mental attitude. In addition, Jung's conception of the spiritual nature of psychic life influenced the later transpersonal and humanistic schools of psychotherapy. His concern with spiritual mystical traditions often led to controversial claims that he himself was a mystic, selling false dreams of spiritual redemption. The alleged evidence of the most recriminatory voice (Noll, 1997) has been traced back to an obviously misidentified and misinterpreted document and shown to be fallacious (Shamdasani, 1998). Jung emphasized that he had no 'doctrine' to offer, but rather some methods of looking at facts, phenomena, and natural laws. His approach to these facts is empirical in the sense of a critical reflection on the nature of all

experience that relates to self-experience and the nature of the perceiving psyche.

This view, which valuates the psyche as 'the mother of all science and arts,' reveals its special importance in the approach to religious phenomena. The unconscious psyche with its autonomous operators (collective archetypes) reflects the numinous, awful, frightening, and timeless spirit in changing images that correspond to changing religious images from ancient times to the present. In differentiating between religious images and religion itself, Jung pointed to the inner experiences open to everyone independently of every external belief system.

Jung received numerous honorary degrees: in 1909 the first honorary degree, from Clark University Worcester, Massachusetts, USA; in 1936 from Harvard University, Cambridge, Massachusetts, USA; in 1937 from Hindu-University Benares and the universities of Allahabad und Calcutta/India; in 1938 from Oxford University, UK; in 1945 from the University of Geneva, Switzerland; and in 1955 from the Swiss Federal Institute (ETH), Zurich.

See also: Bleuler, Eugen (1857–1939); Eliade, Mircea (1907–86); Freud, Sigmund (1856–1939); Janet, Pierre (1859–1947); Psychoanalysis, History of; Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Bair, Deirdre, 2003. Jung: A Biography. Little, Brown, Boston.
Bleuler, Eugen, 1923. Textbook of Psychiatry. Allen and Unwin, London.
Ellenberger, Henry F., 1970. The Discovery of the Unconscious. The History and Evolution of Dynamic Psychiatry. Penguin Books, London.
Flournoy, Théodore, 1900/1994. Les Indes à la Planète Mars. Atar, Genève, 1994. From India to the Planet Mars. A Case of Multiple Personality with Imaginary Languages. Ed. Shamdasani, S., Trans. Vermilye, D., Princeton Univeristy Press, Princeton.
Fordham, Frieda, 1953. An Introduction to Jung's Psychology. Penguin Books, Melbourne, Vic.
Forel, Auguste, Forel, A., 1889–2013. Hypnotismus. Engl.: Hypnotism 1907–2009, Rebmann. BiblioLife, New York/Charleston, SC.
Forel, Auguste, 1908–1933. (105) Die Sexuelle Frage. Engl. Adap.:The sexual question, by C.F. Marshall.
Freud, Sigmund, Bleuler, Eugen, 2013. Ich bin zuversichtlich, wir erobern bald die Psychiatrie. Birefwechsel 1904–1937. Schwabe Verlag, Basel.
Freud, Sigmund, Jung, Carl Gustav, 1974 (Mannheim, R, Hull, R.F.C., Trans.). In: McGuire, W. (Ed.), Letters. Princeton University Press, Princeton, NJ.
Graf-Nold, Angela, 1991. History of psychiatry in Switzerland. History of Psychiatry 2, 321–328.
Graf-Nold, Angela, 01/2007. Jung's position at the Swiss Federal Institute of Technology (ETH) Zurich. Jung History: a, Semiannual Publication of the Philemon Foundation 2 (2), 14–17.
Hakl, Hans Thomas, 2013. Eranos/an Alternative Intellectual History of the 20th Century. Equinox, Sheffield.
Jones, Ernest, 1953-57. Sigmund Freud: life and work. Hogarth Press, London.
Jung, Carl Gustav, 1969. In: Read, H., Fordham, M., Adler, G., McGuire, W. (Eds.), Collected Works, 20 Vol. Princeton University Press, Princeton, NJ.
Jung, Carl Gustav, 1989. In: McGuire, W. (Ed.), Analytical Psychology, Notes on a Seminar, Given 1925. Princeton University Press, Princeton, NJ.
Jung, Carl Gustav, Jaffé, Aniela, 1963. Memories, Dreams, Reflections. Pantheon, New York.
Jung, Carl Gustav, Pauli, Wolfgang, 1952. Naturerkläung und Psyche. Rascher Verlag, Zürich, 1955. The Interpretation of Nature and the Psyche. Pantheon, New York.
Jung, Carl Gustav, 2010. In: Shamdasani, S. (Ed.), The Red Book. Norton, New York.
Jahrbücher, Eranos, 1934–1972. Bd. 1–38. Rhein-Verlag, Zurich.
Loewenberg, Peter, 1995. The creation of a scientific community, the Burghölzli 1902–1914. In: Fantasy and Reality in History. New York, pp. 46–89.
Noll, Richard, 1997. The Aryan Christ: The Secret Life of Carl Jung. Random House, New York.
Shamdasani, Sonu, 1998. Cult Fictions. C.G. Jung and the Founding of Analytical Psychology. Routledge and Kegan Paul, London.
Sorge, Giovanni, Fall 2012. Jung's presidency of the international general medical society of psychotherapy. New insights. Jung Journal: Culture and Psyche 6 (4), 31–53.

Relevant Websites

www.aras.org – The Archive for Research in Archetpyal Symbolism.
www.iaap.org – IAAP - International Association for Analytical Psychology.
http://www.cgjunggesellschaften.eu.
www.cgjungpage.org – Reflections on Psychology, Culture and Life. The Jung page.

Film

Face to Face: BBC -Interview with C.G. Jung, John Freeman, October 1959.
Matter of Heart (1986), 107 min. Carl Gustav Jung, featuring interviews with those who knew him and archive footage of Jung. (Withney, M., Wagner, S.).

Juries

Shari S Diamond, Northwestern University School of Law, Chicago, IL, USA; and American Bar Foundation, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes the jury, including changes over time and variations in the use of lay juries in the United States and internationally. It considers the jury not only as a decision maker, but also as a political institution. The topics covered include: methodological issues in studying the jury, how jurors evaluate evidence, jury composition, jury size and decision rule, individual differences, effects of the media, jury competence, jury instructions, deliberations, a comparison of the jury with other potential triers of fact, and reactions to jury trial and the reform movement.

The jury is a group of laypersons recruited by the legal system to resolve a dispute. In contrast to professional judges, jury members are amateur and nonrecurring decision makers. They are selected from the community at large rather than for their legal expertise, and they adjudicate disputes only on occasion rather than on a routine basis. The jury is both a cultural icon and a popular scapegoat, praised as a champion of democracy and denounced as incompetent or biased when it reaches an unpopular verdict. Empirical research offers a more measured view of the jury, revealing a motivated set of decision makers, often admirable in the way they handle the complex and contentious disputes they are asked to resolve, but always human in their struggles to deal with limited resources and imperfect information and to balance competing values in reconstructing events and assessing mental states.

The ideal jury has many potential roles: a thorough and impartial judge of the evidence presented at trial, a faithful representative of the community from which it is drawn, an obedient follower of legal instructions, a protector of citizens against the power of the state, and a repository of common-sense justice that tempers the harsh hand of the law. Critics charge that juries fail to fulfill one or more of these roles. Behavioral scientists who study the jury find some variation from this set of ideal roles, in part because the roles can be inconsistent with one another, in part because the jury's cognitive and motivational characteristics influence its behavior.

Jury Variations

The jury has gone through numerous incarnations over time and exists internationally in a variety of forms. The medieval jury was limited to male property owners, who came to court with extensive knowledge about the case. They acted as witnesses, investigators, and tribunal, determining both the facts and the law. The modern American jury is drawn from the citizenry at large, with the parties entitled to eliminate some prospective jurors if they have preconceived notions about what the verdict should be before the trial begins. Jurors generally are expected to absorb and evaluate the evidence presented at trial, apply the law as it is described in the judge's instructions, deliberate until they reach a verdict or an impasse, and disperse without personal consequences at the end of the case, except for the occasional inquiries of the American press in high-profile trials. The right of citizens to have both criminal and civil controversies resolved by groups of ordinary fellow citizens is guaranteed by the US Constitution and by every state constitution.

Although the majority of criminal trials decided by a jury occur in the United States, juries are also part of the criminal justice system in the countries of the British Commonwealth. Moreover, a new and renewed interest in juries is taking place internationally (e.g., in Argentina, Spain, Russia, and South Korea), reflecting the perception that citizen adjudication in the justice system is a mark of a democratic government. In civil cases, jury trials occur almost exclusively in the United States and Canada. England eliminated juries in most civil cases by the 1930s, retaining the civil jury primarily in defamation cases. Most modern juries are small, compared to the ancient Greek jury of several hundred that tried Socrates. For example, some American jurisdictions have juries of six. All states with the death penalty, however, use a jury of 12 for capital cases. The South Korean jury, introduced in 2008, has five, seven, or nine jurors depending on the seriousness of the offense.

Although the jury is prominent in discussions about the American legal system, legal disputes are infrequently resolved by the verdict of a jury, even in the United States. Most criminal cases in the United States end in dismissals or guilty pleas; most civil cases are dismissed or settled. An estimated 150 000 jury trials occur annually in state courts and an additional 5000 in federal courts (Mize et al., 2007). The influence of the jury extends far beyond the trials it actually decides, however, as litigants and attorneys anticipate what a jury would do if asked to resolve the dispute. Inaccurate expectations are common, based on impressions gleaned from high profile but idiosyncratic trials and from trials resulting in unusually high awards that attract media attention. The absence of direct and reliable feedback on what would have happened if a jury had actually decided a case that was resolved without trial provides little opportunity to correct misimpressions. The value of the jury as a signal of likely trial outcomes has been threatened in recent years by the drop in the percentage of both criminal and civil cases that culminate in trials.

Political Roles of the Jury

In addition to adjudicating disputes, the jury is a potential source of legitimacy for the legal system. To the extent that the

jury is viewed as representing a fair cross section of the community, its verdict is likely to be seen as the product of fair consideration and can carry a degree of legitimacy that the decision of the judge, as an employee of the state, may lack. Even when its verdict is unpopular and even if observers believe that the jury does not fairly represent the community, the jury acts as a lightning rod, insulating the judge and other parts of the state legal system from criticism. American opinion polls consistently show that a majority of citizens support the jury system.

The jury also can act as a conduit for community standards. For example, in evaluating a claim of self-defense, the jury must determine what a reasonable person would be expected to believe, as well as what the particular defendant did believe. Although the jury is charged with applying to the facts the law it receives from the judge, the self-defense example illustrates the fuzziness of the division between law and facts. The jury must often inject its understanding of appropriate standards into its fact finding even while scrupulously following the instructions that the judge provides.

An additional political role for the jury is its educative function, identified by De Tocqueville as the jury's great strength. Surveys suggest that more than half of American adults have had some personal involvement with the courts, and of this subgroup, half have served on a jury. Citizens also receive information and misinformation on the courts from other sources, including the media. The extent to which jury service provides additional or corrective information is unclear, but jury experience tends to make jurors feel more positive about the jury system.

Finally, experience as a juror has a wider effect on the citizens who serve. Recent evidence has shown that it promotes civic engagement and political participation (e.g., prior jurors showed an increased likelihood of voting, Gastil et al., 2010).

Methodological Challenges in Studying the Jury

Study of the jury is a particularly thorny project because the jury speaks publicly only to deliver its verdict: deliberations occur outside the presence of any nonjuror. Although not all jurisdictions have legislation that bars access to jury deliberations, there have been only a few instances in which deliberations have been observed or recorded, and only one systematic study of real jury deliberations (Diamond et al., 2003). Researchers have used a variety of indirect approaches to study juries, each with strengths and limitations: (1) archival studies of jury case characteristics and verdict patterns, which provide information on large samples of cases, but can only reflect the limited variables that have been recorded and may, as in the case of commercial jury verdict reporters, have systematically biased samples, favoring, for example, the publication of cases with large awards (Lee and Waters, 2011); (2) posttrial surveys or interviews that question jurors about the actual process of reaching a verdict, but must depend on what jurors noticed, accurately recorded, remember, and are willing and able to report; (3) surveys of other trial participants, such as judges and attorneys, who are informed court observers, but have only indirect information on jury behavior; (4) field experiments, which can offer strong evidence on the impact of a legal reform like juror note taking, but provide little information on process and require unusual cooperation to implement correctly; and (5) mock jury experiments, the most widely used method for studying the jury. A key strength of mock jury experiments is that they combine control with the opportunity to examine the process of decision making. The extent to which behaviors observed in these experiments mirrors the behavior of real juries is hard to assess. Some evidence, however, shows substantial correspondence between verdict results and simulations that vary in realism, suggesting that not all differences will affect the results (Bornstein, 1999).

How Jurors Evaluate Evidence

The traditional legal model of the jury trial portrays jurors as passive recipients of the evidence and legal instructions. Yet even legal instructions send a double message: decide the case based solely on the evidence presented in court, yet use commonsense and experience in reaching a decision. Empirical studies of jury behavior find jurors to be active processors of incoming information. Jurors enter the courtroom with preconceptions and expectations that affect the way they react to evidence. They actively search for causal explanations to make sense of what they are hearing and seeing, and they fill in gaps and interpret ambiguities in ways that influence their decisions. Although it is possible to imagine a method of decision making in which a decision maker gives each piece of evidence a value and a weight, and then combines the results algebraically to reach a verdict, this mechanical approach does not describe the decision-making process of jurors. Nor have more elaborate mathematical modeling efforts (e.g., Bayesian accounts) captured what jurors do.

A more cognitively grounded, and more widely accepted, model of individual juror decision making is the story model. According to this model, jurors employ an explanation-based judgment process. They construct a summary, mental picture of what decision-relevant events occurred and the causal relations among them (Pennington and Hastie, 1986). The construction of a cognitively coherent explanation is a central focus of related decision models as well (e.g., Simon, 2004). Jurors select, organize, elaborate, and interpret the evidence, drawing inferences to connect the pieces to develop an explanation-based narrative. Each decision maker utilizes a general comprehension strategy for understanding human action that includes some features that legal doctrine may not consider relevant, like the defendant's motive (Nadler, 2012).

Some models of jury behavior generated in the laboratory assume that jurors begin deliberations with clear verdict preferences. Yet when jurors begin their deliberations, many hold only tentative views about the most plausible version of what happened and a rough understanding of the legally relevant criteria for reaching a verdict. As a result, the deliberation process often involves an early consideration of the content of the jury instructions (Diamond et al., 2012). Overall, the jurors attempt to develop the most plausible reconstruction of the events by pooling their assessments of the incomplete and conflicting stories that the witnesses tell. They are sensitive to apparently inconsistent behavior and testimony, within and across witnesses.

Jurors in a minority of cases begin their deliberations in agreement; in the majority of cases, there is more than one way to view the evidence. Deliberations may begin with a vote "to see where we stand" or, more often, with an attempt to construct a group understanding of the evidence. Despite the common image that the jury nearly always takes a vote at the beginning of deliberations, early calls for a vote are frequently derailed by discussion about the evidence or the instructions. Moreover, the longer the trial, the longer the prevote discussion, suggesting that a heavier evidentiary load requires more processing time before jurors attempt to reach a verdict.

Jury Composition

The modern American jury is far more heterogeneous and representative of the citizenry at large than was the early English jury or even the jury in the early twentieth century. Nonetheless, the jury is not a random sample of citizens. It is the product of a multistage selection process that typically begins with a list of potentially eligible jurors drawn from voter registration lists and often supplemented by individuals holding drivers' licenses in the general geographic area where the court sits. Prospective jurors may be excused from jury service on the basis of hardship, but losses also arise from geographic mobility, a failure to update the lists, and nonresponse by prospective jurors to court summons. The loss of prospective jurors in the qualification and summons process results in a systematic underrepresentation of minorities, younger individuals, and those at lower income levels.

The final stage in jury selection occurs when prospective jurors are brought into the courtroom and questioned to determine whether they will serve in the particular case. Those who clearly express preconceived notions about what the verdict in the case should be or those with clear conflicts of interest are excused by the judge, known as a challenge for cause (Rose and Diamond, 2008). In addition, the parties can excuse a limited number of prospective jurors without giving a reason (the peremptory challenge). The US Supreme Court has ruled that peremptory challenges based on race or gender are constitutionally prohibited, but evidence suggests that the prohibition fails to eliminate racially and gender motivated challenges (Sommers and Norton, 2008). Courts generally do not require much justification from the party initiating the peremptory challenge when the opposing party charges that the challenge was improperly motivated. Also, the small number of challenged jurors in the typical trial means that an attorney can generally identify a unique and nondiscriminatory reason for each challenge.

The result of this variety of shaping and sometimes crosscutting forces is that juries tend to be somewhat more educated, wealthier, and older, and less likely to include a representative number of minorities than the distribution of these groups in the adult population. The selection process that occurs before the jurors enter the courtroom is primarily responsible for the difference between the composition of the American jury and the adult population as a whole. Although the difference is likely to persist, the American jury today is more representative than it has ever been. Moreover, jury participation is extensive. Surveys indicate that 25% of American citizens are likely to serve on a jury trial at some point in their lives. That percentage may drop with the decreasing number of jury trials, but because most courts have reduced terms of jury service, often limiting the term to one day of service or one trial, the distribution of lifetime service among citizens may not change.

Jury Size and Decision Rule

Traditional English and American juries consisted of 12 members who were required to reach a unanimous verdict. Some American jurisdictions now permit juries with as few as six members and nonunanimous verdicts of 9:3 or 10:2; England permits a 10:2 verdict if the jury has been unable to reach unanimity within a specified period of time. Reducing jury size decreases the representation of minority perspectives on the jury and increases the likelihood of an aberrant jury verdict. A majority decision rule tends to reduce the rate of hung juries and to shorten deliberations because jurors in the majority do not need to gain the support of all jury members, but hung juries are most likely to occur when the jury has more than one or two holdouts. Jury deliberations under a majority decision rule can be less thorough because the deliberators can arrive at a verdict without considering the reasons why there is disagreement among the members of the jury. Reflecting this evidence, the 2005 American Bar Association Principles on Juries and Jury Trials endorsed twelve-member juries and a unanimous decision rule.

Individual Differences

Attempts to predict juror verdict preferences from background characteristics have generally been able to explain only a modest amount of variance in outcomes (Lieberman and Sales, 2007). Demographic characteristics like gender, race, and age generally account for very little of the variation in response. Attitudinal characteristics can be more powerful, albeit also modest, predictors. Expectations, beliefs, and values affect the way that jurors react to evidence. In that respect, jurors are no different from any other decision makers, because people ordinarily scrutinize more carefully and are more likely to reject information that is inconsistent with their beliefs and expectations, because it is generally easier for people to remember theory-consistent information than theory-inconsistent information, and because ambiguous information tends to be interpreted as theory consistent.

Some types of cases are more likely than others to implicate strongly held beliefs or values. A primary example involves murder punishable by the death penalty: even among jurors willing to impose the death penalty in some cases, the strength of their support for the death penalty can strongly influence the likelihood that they will vote for a death sentence. Moreover, those who support the death penalty are more likely to find a capital defendant guilty of murder.

Juries and the Media

The American constitutional right to a free press has always permitted the media to provide the public with pretrial

information, putting the constitutional rights to a free press and to a fair trial potentially at odds. Traditionally, this media exposure did not occur in countries like Canada and Great Britain where the press is prohibited from writing about impending trials. The Internet has changed all of that. Now that nontraditional media sources, including individual bloggers, can use the Internet to swiftly and widely disseminate both accurate and inaccurate information about cases and parties, both interested and casual users of the Internet may receive these communications. If these sources include potentially prejudicial information about issues like a defendant's alleged confession, the danger is that this exposure will affect the expectations and beliefs of prospective jurors. Although some research on traditional pretrial publicity suggests that the effects are generally small, there is currently little systematic research on the effects of new media on the jury.

Attorneys too, when they have access to the names of prospective and sitting jurors, can use the Internet to conduct background research on the jurors. The line between due diligence and intrusive prying is unclear at this point, as is the impact of gathering this information. In the twenty-first century world of cable and the Internet, limited exposure may be a thing of the past. Moreover, the new technology has spawned legitimate concerns about the potential interactions between jurors and media sources on the Internet, both during and before trial, as well as juror uses social media to communicate with sources of potential influence outside the trial setting. Courts have responded with increasingly specific judicial admonitions and warnings, but it is not yet clear how effective these efforts are likely to be and, to the extent that they are not effective, the consequences for trials. Use of the Internet has grown exponentially, but increasing sophistication about the potentially misleading content available on the Internet may reduce its impact. Nonetheless, the Internet has reduced the effectiveness of moving the location of a high-profile trial as a way to counteract potential effects of pretrial publicity.

Competence, Including Reactions to Experts

Both criminal and civil trials increasingly involve the testimony of experts who present technical and scientific evidence (e.g., DNA analysis, econometric modeling, epidemiological and toxicological research; Vidmar and Diamond, 2000–2001). Such complex evidence presents a challenge for the experts who must communicate with a lay audience, for the judge who acts either as a gatekeeper in ruling on which experts can testify and what they can say or as the trier of fact, and for the lay jury.

Although jurors typically exert considerable effort in trying to understand the content of expert testimony, questions arise about the ability of jurors (and judges) to evaluate technical and scientific evidence. Jurors (and judges) can be confused by expert testimony, but that confusion can often be reduced or avoided if the expert is also a good teacher. Moreover, jurors frequently express skepticism about the trustworthiness of experts presented by parties within the adversary system and, although cross examination may not be as powerful a corrective mechanism as traditional legal doctrine presumes (Levett and Kovera, 2009), jurors adopt reasonable strategies to evaluate expert credibility. In addition, juries can and do draw on the expertise of their most competent members, some of whom may have scientific or technical backgrounds, to assess the strength of the evidence, providing a resource that is unavailable to the trial court judge.

Reactions to the Law, Including Nullification

Jurors are instructed to apply the law as the judge describes it to the facts as the jury finds them. Juries may not always follow the instructions, but the primary cause of that failure is the inadequacy of the instructions rather than a willful disregard by the jury. Jurors may fail to follow the judge's instructions on the law if they are either unmotivated or unable to apply the instructions, but they typically see themselves as obligated to apply the law and spend a significant portion of their time during deliberations discussing the law. Nonetheless, legal instructions often fail to provide jurors with helpful legal guidance. The major goal in writing jury instructions on the law is to produce language that will accurately reflect the relevant law so that appellate courts will find the instruction acceptable. Insufficient attention is generally given to ensuring that the meaning of the instruction is clearly communicated. In addition to legal jargon, poor communication may result not only from legal jargon. It may also occur if omissions leave jurors without guidance on some legal issues or if structural problems arising from the piecemeal construction of jury instructions leave jurors confused about how the pieces fit together (Diamond et al., 2012). Comprehension of instructions can be significantly improved by using plain language, simplifying sentence construction, providing definitions for technical terms, and specifically correcting jurors' inaccurate expectations and misconceptions.

Although the most common source of jury deviations from legal standards is a failure of the legal instructions to convey clearly what the appropriate legal standard is, jurors also may deviate from the path outlined in the instructions due to cognitive or motivational biases. For example, jurors admonished to disregard particular information may find it difficult to do so. Other legal instructions may ask the jurors to engage in mental gymnastics – for example, to use a defendant's criminal record only to assist in evaluating his credibility, but not as evidence of bad character; to forget that they learned about damaging evidence that the judge ruled inadmissible – and jurors may be unwilling or unable to perform the required cognitive adjustments. It is unclear whether these failures are significantly less likely when the trier of fact is a judge. Some recent studies of judges suggest instead that cognitive biases represent heuristic patterns of using information that neither juries nor judges can be expected entirely to overcome.

Finally, jurors may depart from the judge's legal instructions when the application of the legal standard to the particular case so substantially violates the jurors' sense of justice that they are persuaded to temper the letter of the law in that case. Although such deviations are rare, explicit jury nullification plays a central role in conceptions of the appropriate role of the jury and has been a source of extensive debate. Courts have long recognized the power (as opposed to the right) of the jury to nullify, but both courts and some empirical evidence raise concerns about whether instructing juries about that power invites unwarranted rejection of legal standards (Horowitz et al., 2006). When jurors

are explicitly told that they have the right not to apply the law as the judge describes it, they are more willing to reach verdicts that avoid the literal application of the law.

Deliberations

Deliberations resulting in a group verdict distinguish the jury from its chief alternative, the trial court judge, in two ways. By requiring agreement from multiple jurors, the jury verdict in principle reduces the likelihood that the decision will represent an idiosyncratic view of a single deviant decision maker. Moreover, in theory, deliberations give the jury an opportunity to profit from the resources of its multiple members, to pool its knowledge and sensibilities, and to resolve differences, ultimately producing a verdict that reflects more than what could be achieved either by a single decision maker or by mechanically combining or averaging the preferences of the individual members.

The extent to which deliberations actually do affect jury verdicts, however, is in dispute. Early scholars (e.g., Kalven and Zeisel, 1966) suggested that jury verdicts simply reflect the position of the majority before deliberations begin. The picture they presented was consistent with the verdict-driven jury that takes an immediate vote to see where each juror stands and then focuses its attention on persuading the minority to join the position initially held by a majority of the jurors. When a vote is immediate, it is likely to reflect predeliberation preferences. When a discussion of the evidence precedes a vote (the so-called evidence-driven jury), that vote will be affected by any changes that have occurred as a result of the discussion. Although faction size is a strong predictor of verdict (Devine, 2012) and most juries probably do end up with a verdict that reflects the majority that is visible by the time most of the jurors have expressed a verdict preference in deliberations, recent research suggests a more prominent role for deliberations in determining trial outcomes (e.g., Hans et al., 2003). Moreover, the outcome of deliberations is less predictable when a majority position is less likely to emerge because there are multiple verdict choices or in civil cases where damages are at issue.

Civil Liability and Damages

Civil juries are sometimes viewed as proplaintiff when it comes to liability, although there is little evidence to support the claim that they are more likely than judges to find the defendant liable (Greene and Bornstein, 2003). Although jurors do hold corporations to a high standard of behavior, the standard primarily arises from an expectation that corporations are better able to control their behavior than an individual might be (Hans, 2000). On the issue of damages, the pattern of jury behavior is less clear. Compensatory damage awards generally vary with the severity of the injury (i.e., vertical equity), but there is some evidence that plaintiffs with small losses tend to be overcompensated and those with large losses tend to be undercompensated. Compared to special damages (e.g., lost wages, medical expenses), the difficult-to-measure general damages (e.g., pain and suffering), show greater variation in the amounts awarded for injuries at the same level of severity (i.e., horizontal inequity). In the absence of a standard for the 'correct' amount that will compensate for pain and suffering, some scholars have suggested that juries be supplied with information on the pattern of pain and suffering awards in previous jury trials involving similar injuries. Although the problem of defining 'similar injuries' remains, this additional information may reduce some unwarranted disparity across cases.

Civil juries are sometimes instructed that they can award damages for punishment and deterrence (i.e., punitive or exemplary damages) as well as damages for compensation. Punitive awards most frequently occur in cases involving intentional harms and financial rather than physical harms. They are rare, but they are the subject of great controversy, in part because there is little agreement on what award levels are appropriate to suitably punish, as well as deter, these harmful behaviors.

The Jury vs Alternative Triers of Fact

Discourse on the performance of the jury frequently ignores a basic reality of legal systems: in evaluating the jury, the relevant comparison is not some hypothetical ideal decision maker, whatever qualities such a model decision maker would have. Rather, the appropriate comparison is the human alternative or set of alternatives that might be used in the jury's stead. Alternatives to the jury include professional, legally trained judges, and mixed tribunals composed of laypersons and legally qualified professionals. Studies that have compared the reactions of laypersons and legal professionals to the same criminal trials indicate, first, high levels of agreement on verdict (74–90%), and, second, a greater inclination to acquit by laypersons when they disagree with the professionals (e.g., Eisenberg et al., 2005; Kim et al., 2013). In addition, agreement rates do not vary with the complexity of the trial, suggesting that disagreement is not a product of an inability of the laypersons to understand the evidence.

The few studies that compare the process of decision making by judges and jurors show a surprising correspondence even in response to case characteristics that might be expected to influence laypersons more than professionals (Robbenolt, 2005). The assumption of the legal system that judges can 'compartmentalize' and discount legally inadmissible information, and avoid being influenced by it, appears to be an inflated expectation. Both laypersons and professional judges fail to disregard inadmissible evidence, are influenced by outcome information in their judgments about the probability of that outcome (the hindsight bias), and are reluctant to find for the plaintiff when presented with statistically based evidence. If judges are influenced by inadmissible evidence, the jury trial may provide an effective way to avoid burdening the trier of fact with much of the potentially biasing information that judges must regularly confront. In addition, the trial judge in a jury trial retains the power to override a conviction by a jury if the judge determines that the conviction is unwarranted by the evidence, providing a rarely used but available extra protection for the defendant in a criminal jury trial.

Reactions to Jury Trial and the Reform Movement

As a lightning rod for popular debate, the jury is regularly the subject of calls for reform. Some reforms, such as those permitting jurors to submit questions for witnesses during the trial, do not assume that juries are failing to perform adequately but rather are designed to assist them in reaching well-considered judgments, to improve the comfort of the conscripted citizens who serve as jurors, and generally to optimize jury performance and juror satisfaction (Dann, 1993; Diamond et al., 2006). Other proposed reforms aim at reducing the right to jury trial or limiting the power of the jury, such as calls for a complexity exception to the right to jury trial in civil cases and for legislative caps that place a ceiling on jury awards. High-profile trials resulting in unpopular verdicts have stimulated some of the proposed changes in the jury system as have the concerted lobbying efforts of particular interest groups (e.g., the insurance industry). It is likely that calls for reform of all kinds will continue to follow the jury. The real question is whether the increasing scholarly literature on jury behavior will inform popular and political discussions about the jury.

See also: Courts and Adjudication; Criminal Procedure: Adversarial and Inquisitorial Legal Systems; Decision Making, Psychology of; Insurance and the Law; Torts in Law.

Bibliography

American Bar Association, 2005. Principles for Juries and Jury Trials. Thomson West, Eagan, MN.

Bornstein, B.H., 1999. The ecological validity of jury simulations: is the jury still out? Law & Human Behavior 23, 75–91.

Dann, B.M., 1993. "Learning lessons" and "speaking rights": creating educated and democratic juries. Indiana Law Journal 68, 1229–1279.

Devine, D.J., 2012. Jury Decision Making: The State of the Science. New York University Press, New York.

Diamond, S.S., Vidmar, N., Rose, M., Ellis, L., Murphy, B., 2003. Juror discussions during civil trials: studying an Arizona innovation. University of Arizona Law Review 45, 1–81.

Diamond, S.S., Rose, M.R., Murphy, B., Smith, S., 2006. Juror questions during trial: a window into juror thinking. Vanderbilt Law Review 59, 1927–1972.

Diamond, S.S., Murphy, B., Rose, M.R., 2012. The "kettleful of law" in real jury deliberations: successes, failures and next steps. Northwestern University Law Review 106, 1537–1608.

Eisenberg, T., Hannaford-Agor, P.L., Hans, V.P., Waters, N.L., Munsterman, G.T., Schwab, S.J., Wells, M.T., 2005. Judge-jury agreement in criminal cases: a partial replication of Kalven and Zeisel's The American Jury. Journal of Empirical Legal Studies 2, 171–206.

Gastil, J., Deess, E.P., Weiser, P.J., Simmons, C., 2010. The Jury and Democracy: How Jury Deliberation Promotes Civic Engagement and Political Participation. Oxford University Press, New York.

Greene, E., Bornstein, B.H., 2003. Determining Damages: The Psychology of Jury Awards. American Psychological Association, Washington, DC.

Hans, V.P., 2000. Business on Trial: The Civil Jury and Corporate Responsibility. Yale University Press, New Haven, CT.

Hans, V.P., Hannaford-Agor, P., Mott, N., Munsterman, G.T., 2003. The hung jury: The American Jury's insights and contemporary understanding. Criminal Law Bulletin 39, 33–50.

Horowitz, I.A., Kerr, N.L., Park, E.S., Gockel, C., 2006. Chaos in the courtroom reconsidered: emotional bias and juror nullification. Law and Human Behavior 30, 163–181.

Kalven Jr., H., Zeisel, H., 1966. The American Jury. University of Chicago Press, Chicago.

Kim, S., Park, J., Park, K., Eom, J., 2013. Judge-jury agreement in criminal cases: the first three years of the Korean jury system. Journal of Empirical Legal Studies 10, 35–53.

Lee, C.G., Waters, N.L., 2011. A Verdict on the Reporters: The Representativeness and Accuracy of Commercially Published Jury Verdict Reports, paper presented at the Conference on Empirical Legal Studies, November 2011.

Levett, L.M., Kovera, M.B., 2009. Psychological mediators of the effects of opposing expert testimony on juror decisions. Psychology, Public Policy & Law 15, 124–148.

Lieberman, J.D., Sales, B.D., 2007. Scientific Jury Selection. American Psychological Association, Washington, DC.

Mize, G.E., Hannaford-Agor, P., Waters, N.L., 2007. The State-of-the-States Survey of Jury Improvement Efforts: A Compendium Report. National Center for State Courts, Williamsburg, VA.

Nadler, J., 2012. Blaming as a social process: the influence of character and moral emotion on blame. Law and Contemporary Problems 75, 1–31.

Pennington, N., Hastie, R., 1986. Evidence evaluation in complex decision making. Journal of Personality & Social Psychology 51, 242–258.

Robbenolt, J.K., 2005. Jury decision making: evaluating juries by comparison to judges: a benchmark for judging? Florida State University Law Review 32, 469–509.

Rose, M.R., Diamond, S.S., 2008. Judging bias: juror confidence and judicial rulings on challenges for cause. Law & Society Review 41, 513–549.

Simon, D., 2004. A third view of the black box: cognitive coherence in legal decision making. University of Chicago Law Review 71, 415–586.

Sommers, S.R., Norton, M.I., 2008. Race and jury selection: psychological perspectives on the peremptory challenge debate. American Psychologist 63, 527–539.

Vidmar, N., Diamond, S.S., 2000–2001. Jurors and expert evidence. Brooklyn Law Review 66, 1121–1180.

Just City

Susan S Fainstein, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term 'the just city' refers to a body of work that develops a theory of urban justice and derives criteria from it to apply to urban policy. Until recently, critical urban studies identified injustices within the urban realm but, with the exception of the work by David Harvey and Henri Lefebvre, did not explicitly create a normative theory to depict what constitutes justice. Many of the studies with an urban political economy perspective focused on property development and urban social movements, primarily within wealthy Western countries; the criteria on which they based their critique of urban governance and flows of capital remained implicit. In the 1990s and thereafter, urban scholars began to address the topic of justice explicitly. Particularly influential was Iris Marion Young's book, *Justice and the Politics of Difference*, which defines justice as the absence of forms of domination and which moves beyond Marxist-inspired political economy to identify nonmaterial forms of injustice arising from the marginalization of groups. Building on her argument, many scholars named diversity as a chief component of justice. One important strand of thought, based on the writings of Jürgen Habermas, emphasized communicative rationality. Susan S. Fainstein, in her book *The Just City* considers that there are three major components of urban justice: democracy, diversity, and equity, with priority to be given to the latter.

The Just City is a book written by Susan S. Fainstein that was published in 2010. It develops a theory of urban justice and a set of criteria to apply to urban policy derived from these principles. The effort builds on a body of work in geography, planning, politics, and sociology that counters the dominant approach in social science by which individual decisions and market processes constitute the primary explanatory and normative framework for social–spatial description.

Development of Concept within Urban Studies

Although the subject of justice has always been a major topic within political philosophy, contemporary social scientists largely avoided normative statements, adhering to a positivist methodology that aimed at scientific rigor. In the 1960s and 1970s, however, dissenters contended that the ostensibly factual basis of empirical analysis disguised ideological support for an unjust status quo, even while not explicitly developing a concept of justice (Sayer and Storper, 1997). Those attacking injustice within cities sought to demonstrate how spatial manipulation underlay the disadvantages suffered by residents of poor neighborhoods. The theoretical basis for much of the discussion about urban justice was developed by the French theorist Henri Lefebvre (inter alia, 1968, 1970, 1996, 2003), who defined space as being constituted by social relations rather than, as had been the case until the 1960s, by its territorial, physical, and demographic characteristics. Under this reading, space ceased to be a container of buildings, population, and production but instead became a constituent of the relations of production and reproduction and a contributing source of inequality and by implication of injustice.

During the 1970s, a group of theorists who were influenced intellectually by Lefebvre and practically by the uprisings that occurred in European and American cities during the preceding decade broke away from the types of analysis constituting the mainstreams of their disciplines and applied a political--economic epistemology to the study of urban development. Many of these thinkers were based in the Research Committee on the Sociology of Urban and Regional Development of the International Sociological Association (RC21), which was established in 1970. Although the auspices of the research committee were within the discipline of sociology, membership was highly interdisciplinary, and many of the founders of the research committee were neo-Marxists who aspired to inject a spatial element into Marxist theory. Among the founders of this group were Manuel Castells, Michael Harloe, David Harvey, Peter Marcuse, Enzo Mingione, Chris Pickvance, and Edmond Preteceille. Hailing from a number of different, primarily European countries, these scholars developed a critical approach to the relatively recent field of urban studies and started a journal, the *International Journal of Urban and Regional Research*, edited for many years by Michael Harloe, which published articles reflecting their viewpoint.

The two books that proved most influential in setting the course of later work were Manuel Castells's *La Question Urbaine* (1972, 1977) and David Harvey's *Social Justice and the City* (1973). Castells identifies the city as the site of social reproduction and thereby moves the emphasis in Marxist analysis of cities toward questions of distribution, in particular to the role of the local state in providing the means of subsistence or collective consumption (i.e., goods and services provided outside the market, usually under the aegis of state bureaucracies). Within this framework, the important societal agents became state actors and urban social movements demanding benefits from bureaucratic institutions rather than mobilized workers seeking to take over the means of production. In *Social Justice and the City*, Harvey, who is generally considered to be the most prominent geographer of the latter half of the twentieth century, describes how he moved away from more conventional analysis to a focus on the relationship between urban space and the social situation of residents. He argues that space, social justice, and urbanism must be understood in relation to each other and that "it is, of course, the power of Marx's analysis that it promotes such a reconciliation among disparate topics" (1973, p. 17). For both Castells and Harvey, the crucial issue for study

is how power relations (as determined by the interaction between state authority, economic ownership, and urban residents) affect urban outcomes and, in particular, how spatial relations reinforce injustice. Unlike Castells, who limited himself to analysis rather than advocacy, Harvey openly stated that his project was achieving a more just society (even while maintaining that justice could not be defined as an abstract principle). He did not, however, identify an agent of change beyond a general reference to the role of the working class. Rather, Harvey focused on the logic of capital, especially as it circulated through the financial and real estate sectors in cities, regions, and the globe. He pointed to contradictions within the capitalist economic framework and saw these contradictions as undermining the system. Harvey uses the term the 'spatial fix' to describe how capitalists shift investment across spatial boundaries so as to circumvent crises of profitability in any particular place. Castells (1983), who regarded social movements rather than political parties or classes for themselves as the principal force for change, in a later work developed a theory of grassroots movements.

These two foci – property development and community struggle – developed by Castells and Harvey became the preoccupations of later works by a variety of authors concerned with urban justice. Many of them were inspired by insurgent movements against allegedly value-free social science within the traditional academic disciplines. Although these later scholars continued to work within the framework of a critical political economy, they did not usually define themselves as Marxists. Even those who, like Castells, began as Marxist theoreticians, with the exception of Harvey, typically moved away from Marxist terminology and became theoretically eclectic. At the same time, many geographers retained an allegiance to Lefebvre, especially to his concept of the right to the city (see, inter alia, Purcell, 2008; Smith, 2003; Soja, 1989), which became equated with a transcendent ideal of justice that went beyond distributional issues to the right to participate in the creation of the city (Marcuse, 2009).

Foundations for Normative Criteria in Empirical Research

The discussion around urban justice in the terms described above arose primarily among researchers within wealthy, Western countries, who in particular were reacting against the depredations of state-sponsored programs for highway construction and urban redevelopment (or, in British terminology, regeneration) that had the effect of destroying the housing and communities of low-income communities. Among the many works describing such involuntary displacement are Herbert Gans's *The Urban Villagers* (1962), which chronicles the demolition of Boston's predominantly low-income Italian West End and its replacement by luxury apartment buildings; Jane Jacob's *Death and Life of Great American Cities* (1961), which calls redevelopment programs the rape of cities; Alan Altshuler's *The City Planning Process* (1965), which depicts the annihilation of an African-American neighborhood to allow for an expressway; and Parkinson, Foley, and Judd's *Regenerating the Cities* (1988), which presented a group of case studies drawn from UK and US urban renewal experiences. Initial concerns over the razing of neighborhoods and their replacement by high-end residential or commercial structures shifted to the identification of economic development approaches subsidizing wealthy property developers (Fainstein, 2001; Le Galès, 1991) and to the transformation of neighborhoods achieved through gentrification (Smith and Williams, 1986; Hamnett and Randolph, 1988; Lees et al., 2008).

Another thread of research picked up on the question of agents of change. A number of works traced the effects of urban social movements (Piven and Cloward, 1971, 1977; Fainstein and Fainstein, 1974; Mayer, 2010). Paul Davidoff (1965) argued that planners should be advocates for disadvantaged groups, and Norman Krumholz (Krumholz and Forester, 1990) contended that even planners employed by governments could engage in 'equity planning' for the benefit of poor and minority communities. Urban social movements made demands for citizen participation and 'community control' (Altshuler, 1970), contesting the notion that experts were the best judges of the public interest.

Critical analysis of urban change was, as noted above, carried out in an interdisciplinary framework. Nevertheless, scholars from different disciplines had different emphases. Geographers used the term 'uneven development' to depict the way in which capital flowed in and out of different parts of the city, region, and nation, continually creating relative wealth and relative poverty (Smith, 1991). Sociologists particularly noted the persistence of segregation by race and class (Massey and Denton, 1993) and the social basis of power within cities (Logan and Molotch, 1987). Political scientists analyzed power differentials and identified progrowth coalitions of business elites as the force behind the destruction of low-income communities and the subsidization of commercial interests (Mollenkopf, 1983; Stone, 1989). Planners were especially concerned with the exclusion of low-income communities from the planning process (Arnstein, 1969). As will be discussed below, many of these works assumed a normative standard of justice against which public actions were measured but did not usually make this standard explicit.

Spatial Justice in the Developing World

The literature so far discussed critically examined the relationship between spatial phenomena and social justice within the context of cities with large economic surpluses. Concern with the distributional effects of policy evolved from the perception that in prosperous metropolitan areas there was more than enough wealth to allow everyone a decent standard of living. Within poor countries, the issue of spatial justice takes on a different form. Dependency theory attributes continued widespread poverty to the domination by core countries of peripheral areas. Theories of uneven development, originally referring to flows of capital among and within cities in the developed world, are applied to the relations among nations. While framed also within an understanding that spatial relations and the real estate industry contribute to immiseration of low-income communities, the focus of scholarship is on the region and the nation rather than the city (see, e.g., Wallerstein, 1974; Amin, 1976; Cardoso and Faletto, 1979).

Making the Implicit Explicit

Until the 1990s, much urban literature has as a subtext a critique of injustice and a normative standard on which the critique was based, but only David Harvey actually uses the word 'justice' in developing his analysis, and he manifests considerable ambivalence toward the term. On the one hand, he declares, "We cannot do without the concept of justice for the simple reason that the sense of injustice has historically been one of the most potent seedbeds of all to animate the quest for social change" (Harvey and Potter, 2009, p. 41). On the other, he sees all concepts of justice as embedded in power relations and rejects "idealist specifications of rights and of justice" (Harvey and Potter, 2009). It is not clear, however, from where oppositional concepts derive if not from idealist specifications, albeit filtered through existing social practices.

During the 1990s, urban scholars finally began to address the topic of justice explicitly. Iris Marion Young, although a philosopher by discipline, taught in an urban planning department at the University of Pittsburgh. Her book, *Justice and the Politics of Difference* (1990), specifically addresses the issue of justice within the city. Her identification in that book of injustice with types of oppression led her to define justice as the absence of forms of domination (exploitation, marginalization, powerlessness, cultural imperialism, and violence), an approach adopted in a number of subsequent works. Harvey's concern with social justice and the city prompted a conference and then the publication of a book (Merrifield and Swyngedouw, 1996) on the occasion of the twentieth anniversary of his *Social Justice and the City*. The intent of this initiative was to bring to the fore the issue of social justice which, the editors opined, had been "relegated to the hinterlands of academic inquiry" (1996, p. 2).

Considerable effort has been made since the beginning of the new millennium to define justice and prescribe approaches to realizing it within the urban context. Where earlier scholars had not specified the value system underlying their critiques or offered prescriptions for more just policies, they began to hold conferences, produce books, and start journals that named justice as their objective (e.g., Marcuse et al., 2009; Hayward and Swanstrom 2011; Brenner et al., 2012; the French journal *Justice Spatiale/Spatial Justice*) In doing so, they looked to political philosophy for guidance since philosophers, unlike social scientists, have long discussed the lineaments of a just society. Mostly, however, philosophers avoided making practical recommendations for its realization.

The most important twentieth-century book developing a theory of justice is by John Rawls. Famously, Rawls argued that the allocation of goods in a society should be governed by the 'difference principle,' whereby policies should only improve the situation of those better off when "doing so is to the advantage of those less fortunate" (Rawls, 1971, p. 75). Rawls posits that in the "original position," behind a "veil of ignorance" that prevents individuals from knowing what will be their ultimate position in society, they will opt for a more egalitarian distribution of goods. By this logic, egalitarianism is a rational choice for individuals. Although Rawls does not address justice within the city, his argument for a just society defined by greater equity stimulated much later work within political philosophy that in turn filtered into urban social science. In particular, the writings of Jürgen Habermas as well as Henri Lefebvre and Iris Marion Young influenced urban theorists.

Choosing justice as the norm for urban policy represents a reaction to the growing inequality and social exclusion arising from the use of neoliberalism as the template for public policy (Brenner and Theodore, 2002). 'Neoliberalism' refers to the doctrine that market processes produce the most efficient allocation of resources, provide incentives that stimulate innovation and economic growth, reward merit, and consequently are conducive to the greatest good of the greatest number. For the market to work, state action that distorts prices and interferes with rewards to investors must be minimized; rather the local state needs to offer incentives to investors if it is to compete within the world system of cities. Within this ideological framework, policies to reduce inequality and provide advantages to minority groups hinder the workings of the market's invisible hand and cause the economy to perform at a suboptimal level. Under this governing principle, efficiency becomes the single criterion for evaluating public policy, and cost-benefit analysis becomes the tool for its realization.

The use of cost-benefit analysis to assess policy derives from philosophical utilitarianism. Its governing principle of maximizing the greatest good for the greatest number underlies the types of calculations generally used for choosing among policy alternatives. Despite its prevalence it has been strongly criticized for failing to take into account the impacts of policies on minorities and its blindness to questions of distribution (Rawls, 1971; Campbell and Marshall, 2002). Thus, to take the example of urban redevelopment programs, displacement of residents is justifiable if the majority benefits even marginally, regardless of the serious costs to those displaced. Further, it usually involves a quantification of benefits that is biased in favor of the already well-off. For instance, investment in commuter rail, if it adds to time savings by high-income individuals, will have a greater payoff than expenditures on buses for low-income people whose time has less monetary value. The counterargument is that justice as the yardstick for measuring public policy effectiveness does not negate efficiency as a goal but instead requires the policy maker to ask to what end efficiency applies. If a policy serves the goal of assisting the most disadvantaged without wasting resources, in line with Rawls's difference principle, then the policy is efficient even if it does not maximize an aggregate cost-benefit ratio.

Democracy, Diversity, Equity

Three main approaches to the issue of justice in the city characterize recent writing on the subject; the differences among them are primarily a matter of emphasis. Most prominent during the latter part of the twentieth century within planning theory is communicative rationality (Forester, 1993; Innes, 1995), sometimes referred to as the collaborative approach (Healey, 2006). Derived from Habermas's concept of communicative rationality, the emphasis here is on democratic decision making and represents a reaction to the high-handed technocratic methods underlying urban renewal programs. It is

assumed that the stronger the role of disadvantaged groups in policy decisions, the more redistributional will be the outcomes; thus, broad participation and deliberation should produce more just outcomes.

Communicative rationality starts with an abstract proposition. According to Patsy Healey (1996, p. 239):

> A communicative conception of rationality … replaces[s] that of the self-conscious autonomous subject using principles of logic and scientifically formulated empirical knowledge to guide actions. This new conception of reasoning is arrived at by an intersubjective effort at mutual understanding. This refocuses the practices of planning to enable purposes to be communicatively discovered.

Within communicative theory, the planner's primary function is to listen to people's stories and assist in forging a consensus among differing viewpoints. Rather than providing technocratic leadership, the planner is an experiential learner, at most providing information to participants but primarily being sensitive to points of convergence. Leadership consists not in bringing stakeholders around to a particular planning content but in getting people to agree and in assuring that, whatever the position of participants within the social–economic hierarchy, no group's interest will dominate. Among political scientists, this viewpoint is generally captured in the term deliberative democracy and assumes that through interaction participants will modify their ideas and reach a consensus.

Critics of this approach contend that initial conditions of inequality cause any consensual outcome that results in to reflect the interests of the most powerful (see Huxley and Yiftachel, 2000; Campbell and Marshall, 2006; Flyvbjerg, 1998). They consider that bringing about changes favoring relatively disadvantaged groups necessarily involves conflict (Purcell, 2009). In relation to the role of the planner, Fainstein (2010) asserts that planners whose aim is justice need to intervene in the planning process, calling for policies that favor low-income and minority groups. Further, she worries that democracy can lead to exclusionary practices, since a situation in which all social classes are proportionally represented will rarely occur. Even when representatives of the poor do participate, they may be co-opted or manipulated. She concedes, however, that planners have little independent power and will only succeed in improving the situation of the less powerful when they have political backing from either grassroots movements or progressive politicians.

The debate over how much importance to attach to democratic decision making as opposed to the differential impacts of policy boils down to a dispute over the importance of process versus outcome. For those who regard a process-centered approach as inadequate for obtaining justice, the problem is that it fails to take into account the unequal positions of the participants. In Nancy Fraser's words (1997, p. 79): "Liberal political theory assumes that it is possible to organize a democratic form of political life on the basis of socioeconomic and sociosexual structures that generate systemic inequalities."

Diversity

Among philosophers, Rawls has been criticized for his individualistic approach to society. The critique centers on his seeming obliviousness to group identity (Young, 1990) and his insensitivity to nonmaterial forms of discrimination (Honneth, 2003). According to this viewpoint, discrimination against groups defined by their color, cultural inheritance, or sexual preferences cannot simply be collapsed into a manifestation of material inequality. Young (1990) argues that social justice requires recognition of and respect for group differences. She calls for the rejection of an assimilationist model under which difference is eradicated but rather the assertion of a positive sense of group difference wherein the group defines itself rather than being defined by others.

Toward the end of the twentieth century, the concern among philosophers with recognition of the other converged with an emphasis among urbanists on diversity. As used by urbanists the term has a variety of meanings (Fainstein, 2010). Among urban designers it refers to mixing building types; among planners it denotes both mixed uses of land and social heterogeneity of residential buildings, neighborhoods, and public spaces. Jane Jacobs (1961) was particularly influential in her call for a physical heterogeneity that would promote a corresponding social mixing. She maintains that such diversity will prove beneficial not only for producing attractive cities but for making cities economically productive. Her reasoning anticipates the widely publicized argument of Richard Florida (2002) that diversity stimulates creativity and in turn economic growth.

A number of urban scholars regard the merit of diversity not in its promotion of economic growth but rather in enhancing the right to the city. Leonie Sandercock (1998) terms her ideal city 'Cosmopolis' and depicts a metropolis that gives people from a variety of ethnic and racial backgrounds equal rights to city space and "the pleasures of anonymity." Richard Sennett (1970, 1990), who long ago championed "*The Uses of Disorder*," similarly praises the benefits of difference and even of conflict. New urbanists, who both write about and actually design communities, aim to achieve social diversity through a variety of building types, while land-use planners call for inclusionary zoning that requires new development to provide for low-income housing.

The achievement of diversity, however, may militate against other values. Moving people against their will in order to produce diversity, as has occurred in American programs to replace public housing with mixed-income projects (Goetz, 2003), is contrary to democratic norms of self-determination. Furthermore, there is a tension between heterogeneity and community. According to Robert Putnam (2007, p. 137), "immigration and ethnic diversity tend to reduce social solidarity and social capital. New evidence from the US suggests that in ethnically diverse neighbourhoods residents of all races tend to 'hunker down'. Trust (even of one's own race) is lower, altruism and community cooperation rarer, friends fewer." Iris Young supports neighborhood homogeneity within a metropolitan context of porous borders and multiple groups because of the need to recognize difference and to understand that living among those like oneself provides existential security (Young, 2000).

Equity

Edward Soja (2010) has highlighted the concept of uneven development in his investigation of spatial justice. Like Iris

Marion Young and David Harvey he begins with a depiction of injustice and considers that geography is "a significant causal force in explaining [inequitable] social relations and societal development" (2010, p. 63). He argues that the pursuit of justice requires gaining control over the processes producing unjust urban geographies. He does not identify specific programs to reduce spatial injustice but rather looks to coalitions of groups demanding the right to the city as the vehicles for achieving both greater material equity and also greater respect for marginalized populations. He describes such coalition building in Los Angeles and also the teaching of spatial justice practice at the University of California at Los Angeles as strategies for reaching a more just city.

Fainstein (2010) points out that the values of equity, diversity, and democracy may be in conflict, and she gives equity priority among them. She is particularly concerned with planning and redevelopment decisions made at the city level, and she examines them through an investigation of the cases of New York, London, and Amsterdam. She concludes that Amsterdam is the most equitable and democratic of the three and that New York has been most successful in absorbing immigrants. London is an intermediate case in regard to equity. Johannes Novy and Margit Mayer (2009, p. 104), however, criticize Fainstein for holding up the social democratic cities of Europe as a model, contending that it "distracts from and, worse, naturalizes the inequalities and injustices which mark them … ."

The ultimate intent of Fainstein's book is to specify policies that would benefit relatively disadvantaged social groups, as defined by income or marginality. To this end, she lists programs that enhance equity, stressing housing development for low-income households, preventing involuntary displacement, giving priority in economic development programs to the interests of employees and small businesses, and lowering intraurban transit fares. In furtherance of diversity, she calls for ending discriminatory zoning, insuring that boundaries between districts remain porous, providing widely accessible and varied public space, and mixing land uses. In regard to democracy, she calls for the use of advocates to represent groups that do not participate directly in decision making, consultation of target populations in areas to be redeveloped, and broad consultation for areas that are not yet developed but are under development pressure. She notes, however, that, if the aim is justice, the purpose of inclusion in decision making is to have interests fairly represented, not to value participation in and of itself. This further value may well underlie a vision of the good city, but it is not necessary for her definition of justice.

An important difference between her and David Harvey is that she is willing to embrace reform through existing political–economic processes rather than viewing greater justice as unattainable under capitalism. In Harvey's view, her approach "is thus constrained to mitigating the worst outcomes at the margins of an unjust system" (Harvey and Potter, 2009, p. 46). He calls capitalist liberal democracy a régime of rights and asserts: "My objection to this régime of rights is quite simple: to accept it is to accept that we have no alternative except to live under a régime of endless capital accumulation and economic growth no matter what the social, ecological, or political consequences" (Harvey and Potter, 2009, p. 43).

At the root of their disagreement is the question of whether or not the capitalist state within a political democracy can be an instrument for redistribution. Harvey, in line with Marx's dismissal of the capitalist state as the executive committee of the bourgeoisie, expects little from it. Furthermore, consonant with Marx's emphasis on production rather than distribution as the basis for inequality, he looks to restructuring the relations of production and the processes of spatial development rather than to enlarging the state's role in providing a social wage as key to creating "the city of one's heart's desire" (2009, p. 40). For Harvey, to begin with a prescription for policy outcomes rather than with a strategy for transforming power relations is a losing proposition.

It is not within the power of municipal governments to achieve transformational change. Only the nation state has this kind of leverage, and the more egalitarian cities of Europe admired by Fainstein are in fact underpinned by strong national welfare programs. At the same time, local policies make life better or worse for people. There are many decisions, especially involving housing and transport, made at the local level that differentially affect people's quality of life. The extent to which these decisions have responded to the needs of the relatively disadvantaged has been the result of local political pressures and the ways in which problems are framed. The hope underlying the discussion of the just city is that it can change the rhetoric around urban policy from a single-minded focus on competitiveness to a discourse about justice.

See also: Advocacy and Equity Planning; Citizen Participation; Gentrification; The Right to the City; Urban Neoliberalism.

Bibliography

Altshuler, Alan A., 1965. The City Planning Process. Cornell University Press, Ithaca.
Amin, Samir, 1976. Unequal Development. Trans. by BrianPearce. Monthly Review Press, New York.
Altshuler, Alan A., 1970. Community Control. Pagasus, New York.
Arnstein, Sherry R., 1969. A ladder of citizen participation. Journal of the American Planning Association 35 (4), 216–224.
Brenner, Neil, Marcuse, Peter, Mayer, Margit, 2012. Cities for People, Not for Profit. Routledge, New York.
Brenner, Neil, Theodore, Nik (Eds.), 2002. Spaces of Neoliberalism. Blackwell, Oxford.
Campbell, Heather, Marshall, Robert, 2002. Utilitarianism's bad breath? A re-evaluation of the public interest justification for planning. Planning Theory 1 (2), 163–187.
Campbell, Heather, Marshall, Robert, 2006. Towards justice in planning. European Planning Studies 14 (2), 239–252.
Cardoso, Fernando Henrique, Faletto, Enzo, 1979. Dependency and Development in Latin America. Trans. by Marjory Mattingly Urquidi. University of California Press, Berkeley.
Castells, Manuel, 1972. La question urbaine. F. Maspero, Paris.
Castells, Manuel, 1977. The Urban Question: A Marxist Approach. Trans. by Alan Sheridan of *La question urbaine*. MIT Press, Cambridge.
Castells, Manuel, 1983. The City and the Grass Roots. University of California Press, Berkeley.
Davidoff, Paul, 1965. Advocacy and pluralism in planning. Journal of the American Institute of Planners 31 (4), 544–555.
Fainstein, Susan S., 2001. The City Builders. University Press of Kansas, Lawrence, KS.
Fainstein, Susan S., 2010. The Just City. Cornell University Press, Ithaca, NY.
Fainstein, Susan S., Fainstein, Norman I., 1974. Urban Political Movements. Prentice-Hall, Englewood Cliffs, NJ.
Florida, Richard, 2002. The Rise of the Creative Class. Basic Books, New York.

Flyvbjerg, Bent, 1998. Rationality and Power. Trans. by Steven Sampson. University of Chicago Press, Chicago.
Forester, John, 1993. Critical Theory, Public Policy, and Planning Practice. State University of New York Press, Albany.
Gans, Herbert J., 1962. The Urban Villagers. Free Press of Glencoe, New York.
Goetz, Edward, 2003. Clearing the Way. Urban Institute Press, Washington, DC.
Hamnett, Chris, Randolph, Bill, 1988. Cities, Housing and Profits. Hutchinson, London.
Harvey, David, 1973. Social Justice and the City. Johns Hopkins University Press, Baltimore.
Harvey, David, Potter, Cuz, 2009. The right to the just city. In: Marcuse, Peter, Connolly, James, Novy, Johannes, Olivo, Ingrid, Potter, Cuz, Steil, Justin (Eds.), Searching for the Just City. Routledge, New York, pp. 40–51.
Hayward, Clarissa Rile, Swanstrom, Todd (Eds.), 2011. Justice and the American Metropolis. University of Minnesota Press, Minneapolis.
Healey, Patsy, 1996. Planning through debate: the communicative turn in planning theory. In: Campbell, Scott, Fainstein, Susan (Eds.), Readings in Planning Theory. Blackwell, Oxford, UK, pp. 234–257.
Healey, Patsy, 2006. Collaborative Planning, second ed. Palgrave Macmillan, New York.
Honneth, Axel, 2003. Redistribution as recognition: a response to Nancy Fraser. Trans. by Joel Golb, James Ingram and Wilke, Chritiane. In: Fraser, Nancy, Honneth, Axel (Eds.), Redistribution or Recognition? Verso, London, pp. 110–197.
Huxley, Margo, Yiftachel, Oren, 2000. New paradigm or old myopia? Unsettling the communicative turn in planning theory. Journal of Planning Education and Research 19 (4), 333–342.
Innes, Judith E., 1995. Planning theory's emerging paradigm: communicative action and interactive practice. Journal of Planning Education and Research 14 (3), 183–189.
Jacobs, Jane, 1961. Death and Life of Great American Cities. Vintage, New York.
Krumholz, Norman, Forester, John, 1990. Making Equity Planning Work. Temple University Press, Philadelphia.
Lees, Loretta, Slater, Tom, Wyly, Elvin, 2008. Gentrification. Routledge, New York.
Lefebvre, Henri, 1968. La droit à la ville. Éditions Anthropos, Paris.
Lefebvre, Henri, 1970. La révolution urbaine. Gallimard, Paris.
Lefebvre, Henri, 1996. Writings on Cities/Henri Lefebvre; Selected. Translated, and Introduced by Kofman, Eleonore and Lebas, Elizabeth. Blackwell, Oxford.
Lefebvre, Henri, 2003. The Urban Revolution. Trans. by Bononno, Robert. University of Minnesota Press, Minneapolis.
Le Galès, Patrick, 1991. L'invention d'une politique de développement économique local en France et en Grande-Bretagne. A.N.R.T. Université de Lille III, Lille.
Logan, John R., Molotch, Harvey, 1987. Urban Fortunes. University of California Press, Berkeley.
Marcuse, Peter, Connolly, James, Novy, Johannes, Olivo, Ingrid, Potter, Cuz, Steil, Justin (Eds.), 2009. Searching for the Just City. Routledge, New York.
Marcuse, Peter, 2009. Postscript. In: Marcuse, Peter, Connolly, James, Novy, Johannes, Olivo, Ingrid, Potter, Cuz, Steil, Justin (Eds.), Searching for the Just City. Routledge, New York, pp. 240–254.
Massey, Douglas S., Denton, Nancy A., 1993. American Apartheid. Harvard University Press, Cambridge.
Mayer, Margit, 2010. Social Movements in the (Post-) Neoliberal City. Bedford Press, London.
Merrifield, Andy, Swyngedouw, Erik, 1997. Social justice and the urban experience: an introduction. In: Andy Merrifield, Andy, Swyngedouw, Erik (Eds.), The Urbanization of Injustice. New York University Press, New York, pp. 1–17.
Mollenkopf, John H., 1983. The Contested City. Princeton University Press, Princeton.
Novy, Johannes, Mayer, Margit, 2009. As 'just' as it gets? In: Marcuse, Peter, Connolly, James, Novy, Johannes, Olivo, Ingrid, Potter, Cuz, Steil, Justin (Eds.), Searching for the Just City. Routledge, New York, pp. 103–119.
Parkinson, Michael, Bernard, Foley, Dennis, R. Judd (Eds.), 1988. Regenerating the Cities. Manchester University Press, Manchester.
Piven, Frances, Cloward, Richard, 1971. Regulating the Poor. Pantheon, New York.
Piven, Frances, Cloward, Richard, 1977. Poor People's Movements. Pantheon, New York.
Purcell, Mark, 2008. Recapturing Democracy. Routledge, New York.
Purcell, Mark, 2009. Resisting neoliberalization: communicative planning or counter-hegemonic movements. Planning Theory 8 (2), 140–165.
Putnam, Robert D., 2007. E pluribus unum: diversity and community in the twenty-first century. Scandinavian Political Studies 30 (2), 137–174.
Rawls, John, 1971. A Theory of Justice. Harvard University Press, Cambridge.
Sandercock, Leonie, 1998. Towards Cosmopolis: Planning for Multicultural Cities. John Wiley, New York.
Sayer, Andrew, Storper, Michael, 1997. Ethics unbound: for a normative turn in social theory. Environment and Planning D: Society and Space 15 (1), 1–18.
Sennett, Richard, 1970. The Uses of Disorder. Vintage, New York.
Sennett, Richard, 1990. The Conscience of the Eye. Knopf, New York.
Smith, Neil, 1991. Uneven Development. Blackwell, Oxford.
Smith, Neil, 2003. Foreword to Henri Lefebvre, The Urban Revolution. University of Minnesota Press, Minneapolis.
Smith, Neil, Williams, Peter, 1986. Gentrification of the City. Allen and Unwin, Boston.
Soja, Edward W., 2010. Seeking Spatial Justice. University of Minnesota Press, Minneapolis.
Soja, Edward W., 1989. Postmodern Geographies. Verso, New York.
Stone, Clarence, 1989. Regime Politics. University Press of Kansas, Lawrence.
Wallerstein, Immanuel M., 1974. The Modern World System. Academic Press, New York.
Young, Iris Marion, 1990. Justice and the Politics of Difference. Princeton University Press, Princeton.
Young, Iris Marion, 2000. Inclusion and Democracy. Oxford University Press, Oxford, UK.

Justice, Access to

Earl Johnson, California Court of Appeal, CA, USA; and University of Southern California Law School, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Equal access to justice for people no matter their economic means is a fundamental value in organized societies, especially in democracies. Yet nations vary greatly in their commitment to providing that access to their lower income citizens. Some guarantee access to justice as a statutory or constitutional right and invest substantial public resources to supply subsidized lawyers to poor citizens, but many do not. Beyond the provision of subsidized counsel, other approaches for improving access to justice include contingent or conditional fees allowing lawyers to earn a percentage of their clients' recoveries while not charging if they lose; simplified procedures and/or laws that lessen the need for lawyers; substitution of lay advocates for lawyers in some forums; providing out-of-court assistance to allow disputants to represent themselves; and technologies that enhance access to justice in various ways.

History and political theory both underscore the fundamental importance of providing equal access to justice for all citizens no matter what their economic and social status. It was promised to the subjects of emperors and kings, and sometime delivered, long before those subjects began to even dream of the possibility of governing themselves. Then democratic theory, especially the social contract theory so influential with the founders of democracies in Europe and North America, emphasized that no man would surrender his natural right to resolve disputes through force unless government provided a forum to resolve those disputes where all citizens, rich and poor, had an equal chance of prevailing if in the right (Johnson, 2000, 2010). Thus, it is no surprise that constitutions and basic laws of nearly every democracy guarantee their citizens 'due process' or 'fair hearings' and 'equal protection of the laws' or 'equality before the law' (Johnson, 2000, 2010), as do multi-nation 'fundamental rights' agreements (European Convention on Human Rights and Fundamental Freedoms (1950), European Charter of Fundamental Rights (2009), and International Covenant on Civil and Political Rights of the United Nations (1966).)

Despite these solemn guarantees, however, for most democratic nations it was well into the twentieth century before governments became serious about offering their lower income citizens equal access to justice. Progress has been uneven, with some countries forging far ahead of others. This has become a field of study for political scientists, sociologists, historians, and legal scholars – seeking to analyze the problems, to devise solutions, to evaluate experimental approaches, to comprehend the political opposition to this goal, and to measure the progress.

Social scientists ponder how to measure the extent of need for legal services among low- and sometimes moderate-income people. Many countries have conducted survey research querying respondents about whether they have experienced common problems that experts have identified as being 'legal' or 'justiciable.' Although time frames for the studies varied, the percentage of people reporting at least one eligible problem ranged from a low of 19.5% in Japan to a high of 67% in the Netherlands, with Scotland, New Zealand, the UK, Northern Ireland, and the US in between (Currie, 2007). All these studies reflect that many respondents experienced more than one such 'legal' or 'justiciable' problem during the relevant time frame – some of them several such problems (see, e.g., American Bar Association, 1994; Currie, 2007).

These studies appear to document a vast need for legal services among low- and in some instances moderate-income populations. But they may not be an accurate measure of the 'effective demand' for lawyers in those communities, that is, how many of these problems poor people would take to lawyers if those services were free and readily available to them. Legal needs studies in nations where legal aid is a matter of right and fully funded reflect that far more poor people have justiciable problems than ask lawyers for their help each year, even though that help is plentiful and free. The continued existence of such a gap between 'legal needs' and 'effective demand' for legal assistance may be the result of (1) an irrational aversion to lawyers among some poor people, (2) a rational evaluation that their position lacks enough merit for a lawyer or anyone else to cure the problem, or (3) a feeling that some of the problems they have can be handled just as well by other services in the community – or a combination of these factors. Nevertheless, effective demand for lawyers remains substantial among low-income people in every country, while nations differ dramatically in their willingness to satisfy that demand.

An Overview of Access to Justice Strategies

The study of access to justice issues in modern societies begins with an appreciation of the essential role lawyers play in the regular civil courts in nearly every nation advanced enough to have a professional court system. To varying degrees, these societies place much of the burden on the competing parties in civil cases to investigate the facts, find the applicable law, and present the case to neutral, often passive decision-makers, whether professional judges or lay juries, or a mix. In many nations, those neutral decision-makers often are the only elements of the process the government supplies. Civil-law

countries on the European continent and elsewhere may use an 'inquisitorial system,' with judges guiding the discovery process and asking most of the questions at any hearings. Yet even then at least some responsibility for procuring the relevant factual evidence still often falls on the parties and requires the expertise and thus the expense of lawyers. Common-law countries such as Britain and its former colonies, and others using an 'adversarial system,' impose a much heavier burden on the competing parties, expecting them to find the relevant evidence, organize the presentation, and ask most of the questions at any evidentiary hearings and even locate and argue which legal principles should govern the resolution of the case. Depending on the division of responsibility between the government-funded courts and the private litigants, the latter can bear anything from a significant proportion to nearly all of the cost of the dispute resolution function as carried out in modern nations.

This article analyzes how modern nations seek to supply counsel to those unable to afford their own, and also how some nations have sought to restructure the dispute resolution function to reduce the need for lawyers. Three basic strategies are covered: first, providing government-funded lawyers to disputants who are unable to afford their own legal counsel (Section Increasing Access to Justice by Providing Government-Paid Lawyers); second, providing incentives to encourage private lawyers to represent people unable to afford to pay legal fees but without using government funds (Section Arrangements that Encourage or Require Private Attorneys to Represent Lower Income People without a Government Subsidy); and third, providing dispute resolution forums where disputants can participate fairly without the assistance of a lawyer (Section Increasing Access to Justice without Supplying Free Lawyers, Government Paid or Otherwise).

Increasing Access to Justice by Providing Government-Paid Lawyers

Empirical studies conducted in the USA support the hypothesis that litigants who lack lawyers are severely disadvantaged in the courts, at least as those courts still operate in the early years of the twenty-first century (Galanter, 1974; Sandefur, 2010; Greiner et al., 2013). Debtors sued in collection cases who were represented by counsel had a six times better chance of winning than those who defended themselves (Caplovitz, 1974). A megaanalysis examining scores of prior empirical studies on the impact of counsel in court proceedings found that representation by counsel increased the chances a party would prevail by anywhere from 17 to 1380% (Sandefur, 2010). Almost none of these studies used randomized assignment and thus are open to the possibility that the lawyers picked the more winnable cases. But a recent study employing this research strategy reported that tenants represented by lawyers still fared three times better than those who appeared in court without counsel (Greiner et al., 2013).

Most governments did not require empirical research to reach the conclusion that their lower income citizens required free counsel if they were to enjoy effective access to justice in the regular courts. During the second half of the twentieth century, nearly all industrial democracies began increasing their public investments in civil legal aid, thus increasing access to justice for their low-income citizens. But the level of commitment to equal justice for all citizens and the willingness to fund the lawyers necessary to that goal vary greatly from nation to nation. The World Justice Project's 'Rule of Law Index' ranked 124 nations on, among other things, 'access to civil justice.' Among the 23 'high-income' nations, the Netherlands ranked first on that scale, United Kingdom tenth, the United States twentieth, and Poland last (World Justice Project, 2011). This result may be explained in large part by the relative investments of these countries in civil legal aid as a percentage of their national gross domestic product (GDP) – the best measure of their comparative commitment to equal access to justice. As might be expected, no nation invested a substantial percentage of its GDP in civil legal aid. Nonetheless, by this measure, in 2012, Germany (a civil-law country) spent 2.5 times as much on civil legal aid as the US did, even though much less is expected of the parties and their lawyers in the inquisitorial process used in that country. Ireland and Hong Kong spent 3 times, Ontario (Canada) 4.6 times, Norway and Scotland 6.5 times, the Netherlands 8 times, and England and Wales 11 times more than the US did (Johnson, 2014; tables on pp. 920–921).

The economic downturn in Europe during recent years has prompted some of those nations to reduce their financial commitment to legal aid, along with other programs benefiting their low-income populations. The United Kingdom, for instance, is seeking to reduce its 2013–14 legal aid expenditures by 7.7% from the prior fiscal year (Legal Aid Agency, 2013–14). Yet this still would mean that nation would be spending more than 10 times as much of its GDP on civil legal aid as the US does.

The legal status of a citizen's right to counsel in the regular civil courts also varies greatly among nations and helps account for the 'access to justice' rankings on the 'Rule of Law' Index. While nearly all European and some Commonwealth nations have enacted a broad statutory right to counsel in civil as well as criminal cases, neither the federal nor any state government has done so in the USA. (In the US, the right to counsel currently is limited to criminal proceedings.) England first enacted a statutory right to counsel for poor litigants in civil cases in 1495, France did the same in 1852, Germany in 1877, the rest of Northern Europe in the late nineteenth or early twentieth century, Italy in 1923, Brazil in 1934, and several Canadian provinces and Commonwealth countries in the 1960s and early 1970s (Cappelletti et al., 1975; Johnson, 2000, 2010, 2014). In these countries, statutes guarantee low-income people that the government will supply them with a lawyer in any case where they qualify financially for such assistance, although they typically require the litigant's position to have some merit and the stakes to be significant (Johnson, 2010, 2014).

When most countries initially created a right to counsel they did not compensate the lawyers appointed to provide that representation but expected them to serve on a pro bono basis (Cappelletti et al., 1975). But over the course of the twentieth century more and more industrial democracies began to pay the lawyers serving poor people with government funds, until by the end of the century a majority of European nations and all of the Commonwealth countries chose to do so (Johnson, 2000). Italy, the only Western European democracy to rank

below the USA on the 'access to civil justice' scale on the 'Rule of Law' Index, is the only one on the list that still fails to pay the lawyers it assigns to fulfill the right to counsel (World Justice Project, 2011; Cappelletti et al., 1975; Johnson, 2014).

Europe and the USA also have taken different routes when considering whether to recognize a 'constitutional' or 'quasi-constitutional' right to counsel for indigent (poor) litigants. Despite interpreting nearly identical constitutional language, the U.S. Supreme Court and the European Court of Human Rights have reached opposite conclusions on the right to counsel in civil cases. In 1979, the European Court of Human Rights ruled that the 'fair hearing' guarantee found in the European Convention on Human Rights and Fundamental Freedoms required member governments to supply free counsel to indigent civil litigants in the regular courts (Airey v. Ireland, 1979). In a later decision, that court confirmed this right and extended the rationale to require that indigent litigants be given roughly 'equal arms' when contending with an opponent who has a lawyer (*Steel v. United Kingdom* 2005). In contrast, the U.S. Supreme Court has held the constitutional guarantee of 'due process' does not require the provision of free counsel to civil litigants, although it might in some cases (Lassiter v. North Carolina, 1981).

Some countries not bound by the European Convention have adopted the European, rather than the American approach, to this issue. For example, in 1937 the Swiss Supreme Court held that nation's constitutional guarantee that "All Swiss are equal before the law" required cantonal governments to provide free lawyers to those Swiss citizens who were too poor to hire their own in civil cases (*Judgment of 18 October 1937*; O'Brien, 1967). In 1999, the Canadian Supreme Court found the 'fair hearing' guarantee in that nation's new Charter of Fundamental Rights required governments to provide free counsel, at least in the type of case before the court, when authorities sought to retain custody of a parent's children (*New Brunswick v. G.J.* 1999). In 2001, the South African Land Reform Court held that indigent parties had a constitutional right to counsel in cases before that tribunal (*Nkuzi v. South Africa and the Legal Aid Board* 2001). Brazil has had a civil right to counsel provision in its national constitution since 1934, but only a few states comprehensively implement the right (Alves, ILAG Brazil 2005, 2007, 2009).

Lacking either a constitutional or statutory guarantee, poor people in many other countries around the world, including the USA, some Canadian provinces, and the Peoples Republic of China, receive legal assistance with their civil legal problems only if they can obtain help from a limited number of lawyers supplied through private or legislative charity (Zemans, 1979; ILAG China 2011, 2013, ILAG Canada 2013).

Nations also differ in how they deliver their government-subsidized legal services. There are three common models: compensated private counsel, salaried staff attorneys, and mixed systems. (These models are described in Johnson (2010) based on the authorities cited therein.) Under the compensated private counsel model, private lawyers represent lower income people just as they would paying clients. But when serving those low-income clients the lawyers submit their bills to the government rather than the client. The low-income clients, in turn, can usually choose any lawyer who has expressed a willingness to participate in the government program. This was the predominate delivery system in most of the largest nations of continental Europe, including the Commonwealth countries, when they first started providing public-funded legal services to the poor. This compensated private counsel model remains the sole approach in several of those nations, including France and Germany (Johnson, 2010, and sources cited therein).

Under the salaried staff model the government establishes separate legal aid offices, sometimes at the courts, but more often in 'clinics' or 'neighborhood law offices' in close proximity to concentrations of low-income people. The offices are staffed with salaried lawyers, usually employed full-time. When poor people have legal problems they bring them to one of these offices and are represented by a salaried lawyer. The USA, the People's Republic of China, South Africa, Ireland, a few Canadian provinces, and some Brazilian states, along with a few other jurisdictions, rely principally on salaried legal aid lawyers (Johnson, 2010; ILAG national reports cited therein.)

Several nations have moved to a mixed system that attempts to combine the virtues of the other two. Some of these mixed systems, such as the one in Quebec, Canada, adopt a 'client option' approach, allowing clients to choose between salaried staff lawyers and compensated private attorneys to represent them in a given case. Others, such as in Ontario, Canada, divide legal problems between the two systems on a 'subject matter' basis – typically with clients required to take all 'poverty law' cases (e.g., housing, consumer, government benefits, etc.) to salaried attorney offices and the more traditional legal problems such as divorce to compensated private counsel. Still others have developed a 'functional' divide between salaried staff and compensated private counsel. In the Netherlands, salaried lawyers and supervised paraprofessionals handle all preliminary legal advice and short-term legal assistance, but private lawyers provide all the representation in litigated cases. Finland follows that same pattern, except it offers clients involved in litigation the choice between salaried lawyers or compensated private lawyers (Johnson, 2010).

Some countries using the compensated private counsel model have experienced ever-expanding legal aid budgets, often caused by higher legal fees. These budgetary considerations have led some governments, most notably in England, to begin exploring modifications calculated to reduce costs. One such approach is 'franchising,' where the government signs contracts with certain private lawyers and law firms to provide services at reduced fees; another is 'capitation,' where lawyers agree to handle all the legal problems generated by the entire low-income population in a defined geographic area for a fixed per capita sum (ILAG England 2007, 2009; Moorhead, 1998).

Arrangements that Encourage or Require Private Attorneys to Represent Lower Income People without a Government Subsidy

In some nations, government-subsidized lawyers coexist with fee arrangements that encourage private attorneys to represent lower income clients without government funding in certain categories of cases. The most common of these incentive arrangements are the following: (1) contingent fees (Section Contingent Fees), (2) conditional fees (Section Conditional

Fees), and (3) court-awarded fees, usually in cases that further some important government policy (Section Court-Awarded Fees). In addition to these economic incentives, in many countries public policies encourage or require private lawyers to devote some of their time to the representation of poor people on a pro bono (uncompensated) basis (Section Voluntary and Mandatory Pro Bono Legal Services).

Contingent Fees

The USA is almost unique in allowing lawyers to represent clients on a 'contingent fee' basis. Under fee contracts allowing them a percentage (typically 33–40%, but sometimes lower for early settlements and higher for trials) of the client's recovery, lawyers agree to assume the risk of failure, not charging any fee if they lose. Thousands of the poorest citizens are represented each year in cases where they are suing for a substantial sum of money and have a good chance of winning. At the same time, the contingent fee system offers no help to low-income people who are defendants, or who are plaintiffs in the many categories of legal disputes that do not involve monetary stakes or where the potential financial recovery is too modest to cover the risk of loss (Regan et al., 1999).

The contingent fee system is not without its critics. Some contend that the fee percentages are too generous to lawyers, leaving their clients with less than full compensation for their injuries. Others, however, respond that without the bonus of substantial fees when they win, contingent fee lawyers could not afford to take riskier or smaller cases and thus would contribute less to increasing access for the low-income population (Regan et al., 1999). Some jurisdictions have legislated 'caps' on injured parties' damage recoveries and on the percentages of those recoveries lawyers can earn – at least in certain types of cases (Regan et al., 1999). Typical are caps on recoveries and legal fees in medical malpractice cases, which can make it difficult for poor people to find lawyers unless they suffer a catastrophic injury (California Medical Injury Compensation Reform Act).

Conditional Fees

Most countries, other than the USA, use a 'loser pays' rule in allocating legal fees. That is, losing litigants not only pay their own lawyers but also remit at least a 'reasonable fee' to the winners' lawyers. A few nations have altered this rule in a way that may allow lawyers to serve low-income clients without a government subsidy in a variety of cases.

In 1995, the United Kingdom introduced a modified form of the 'conditional fee' model, one that applies to nearly all civil litigation involving monetary stakes no matter the economic means of the parties. Under this model lawyers do not charge their clients a fee if they lose, just as is true in the 'contingent fee' system. But lawyers who win can charge an 'uplift' (usually called a 'success fee') of up to 100%, in effect doubling their fee. Until 2013, the losing litigant had to pay this 'success fee' as well as the regular fee, to the winner. But as a result of a change enacted that year, the 'success fee' is paid out of the clients' monetary recovery. However, the total amount of the winning lawyers' fee (including the 'success fee') cannot exceed 25% of that recovery after deducting any provisions for future ongoing care or future lost income. Should they lose, clients are not relieved of their obligation to pay the other side's fees. But, if they can afford it, they can purchase insurance to cover this risk (Ministry of Justice, 2013).

The U.K. government expects the 'conditional fee' system to supplant legal aid for the claimant's side of most cases offering the prospect of a monetary recovery. Critics point out the prospects of success must be unusually favorable before lawyers will be willing to provide representation under a 'conditional fee' system. They will not receive a fee from either side if they lose and the 'reasonable fee' they receive as an 'uplift' from their clients' recoveries will be far less generous than the 33–50% of the judgment that contingent fee lawyers typically earn when they win (Moorhead, 2000).

Court-Awarded Fees

Most jurisdictions in the USA depart from the general 'American' rule which expects all litigants, winners, and losers to bear their own legal fees and costs (Luebsdorf, 1984) and instead shift fees to the losing party in certain categories of cases in order to increase access to justice in those cases. These are

1. Statutory fee-shifting – in lawsuits under statutes specifically authorizing fee-shifting in favor of private parties whose cases enforce the public policies embodied in those statutes.
2. Public interest fee-shifting – in lawsuits where the court, rather than the legislature, determines that the public interest will be promoted by awarding winning litigants the bonus of requiring the loser to pay their legal fees.

Both statutory fee-shifting and public interest fee-shifting offer an alternative source of funding for those unable to afford their own counsel. They are especially important in major litigation not offering the possibility of a substantial monetary recovery, such as those ordering the government or some other institutional defendant to change its policies or practices in the future. The court is typically not limited by the size of the winning party's economic recovery, which often is nothing, but can award winning lawyers their regular full hourly fees for the time they spent on the case – at least if the request is for a 'reasonable' amount of time and an appropriate hourly fee in the local community (Murr, 2000). Beyond that, in some jurisdictions, lawyers can apply a 'multiplier,' often doubling or tripling the lawyers' fee in order to recognize the riskiness, complexity, or significance of a particular piece of litigation (Klaber, 2006).

Voluntary and Mandatory Pro Bono Legal Services

As mentioned earlier, most European countries originally implemented the statutory right to counsel by appointing lawyers to provide free legal representation to poor people without compensation – a form of *mandatory* pro bono services. During the twentieth century most of these nations found this approach unsatisfactory and shifted to a system that compensated the lawyers representing the poor (Cappelletti et al., 1975; Huls, 1994; Johnsen, 1994), although a few, like Italy, still depend on drafting lawyers to provide legal aid (Cappelletti et al., 1975).

Voluntary pro bono services depend entirely on the charitable impulses of the private legal profession. Faced with a shortage of government-funded legal aid the USA in the last decades of the twentieth century and the first decade of the twenty-first century, legal aid supporters have attempted to encourage more private lawyers to volunteer time to represent the poor. In many states and cities, organized pro bono programs have been established. Some require private lawyers to report annually on the hours devoted to pro bono services (see, e.g., Florida Bar Association, 2014). A national program issued a 'pro bono challenge' to the 500 largest law firms, asking them to commit 3–5% of the firm's billable hours to pro bono representation of the poor. As of 2012, this effort had succeeded in signing up over 100 firms, and corporate law firms donated over two-and-a half-million hours of legal services to the low-income population that year (AmLaw Pro Bono Report, 2012). Some other nations – Sweden and Australia, among them – also have begun to look to pro bono services to fill some of the gaps created by reductions in government legal aid budgets (Regan et al., 1999).

There is a serious question, however, whether societies can expect pro bono legal services to play more than a limited, supplementary role in the delivery of legal representation to low-income people. It seems there is no more reason to expect lawyers to donate enough services on a pro bono basis to meet the low-income population's total need for legal services than to expect medical doctors and hospitals to supply enough of their services pro bono to replace the present government-funded health care systems for the poor – e.g., Medicare and Medicaid in the USA, the National Health Service in the UK, etc.

Increasing Access to Justice without Supplying Free Lawyers, Government Paid, or Otherwise

Professional lawyers appear essential to fair proceedings in the regular courts of most modern countries, as the European Court on Human Rights held in Airey v. Ireland (discussed in Section Increasing Access to Justice by Providing Government-Paid Lawyers). The legal rules governing the decisions are complex, the procedures intricate, and investigatory techniques difficult to master. For all these reasons, unrepresented parties seldom have a chance for a 'fair hearing' in the regular courts – especially if the opposing party has a lawyer.

Yet the European Court of Human Rights also recognized it may be possible for governments to create forums – for some cases at least – where the assistance of lawyers is not crucial and thus effective access to justice can be achieved without providing free counsel to lower income disputants (Airey v. Ireland, 1979). It is possible to identify five approaches that various countries have tried, alone or in combination, in an attempt to deprofessionalize dispute resolution: (1) simplification of procedures (or diversion to simpler forums) (Section Redesigned Forums – Simplification of Procedures and/or Redistribution of Functions and Responsibilities); (2) simplification of the substantive law (Section Simplification of Substantive Legal Rules Governing Disputes); (3) the use of less expensive, nonlawyer advocates (Section Use of Less Expensive, Nonlawyer Representatives), (4) out-of-court assistance to prepare disputants to represent themselves (Section Out-of-Court Assistance to Unrepresented Litigants), and (5) technological substitutes for legal expertise (Section Technological Substitutes for Legal Expertise).

Redesigned Forums – Simplification of Procedures and/or Redistribution of Functions and Responsibilities

For simpler disputes it might be feasible to construct simpler dispute resolution processes, and for any dispute it might be possible to shift most of the responsibility for discovering and presenting the facts from the private parties to the judges or others the government pays to operate the forum. But there are competing values as well as questions about the efficacy of these approaches that have not yet been resolved through empirical study or extensive experience.

Many complex procedures have been created not to enrich the legal profession by making lawyers essential to operate the system but for the purpose of allowing the courts or other forums to function in an efficient and timely way, and to produce accurate and just results. For instance, typically the parties must go through several proceedings before reaching a full-blown trial in order to weed out meritless claims and defenses as soon as possible. Another set of procedural rules limits the forms of evidence parties may introduce in order to ensure disputes are decided on the basis of reason and not emotion and prejudice. But if that procedural complexity is important if not essential to the efficiency of the courts and the validity of their decisions, when if ever is it desirable to strip away that complexity and thus presumably make the system more accessible to unrepresented parties?

There are categories of cases that seem possible candidates for a simplified process – in essence where all the usual steps can be collapsed into a single court appearance. During that session, the judge can hear and evaluate the evidence from both sides and render a decision, without worrying whether the case could have been resolved earlier and with less expenditure of judicial time. But this brand of one-shot consolidated proceeding appears practical only for relatively simple cases, where there is little advantage in weeding out weaker cases without a full hearing at earlier stages in the process.

Early in the twentieth century, several countries started 'small claims courts' to hear disputes involving minor stakes. Using simple procedures and expecting the judges to take a more active role in developing the facts, these courts were designed to function well without lawyers (Turner and McGee, 2000). Some criticize small claims courts for pitting unrepresented individual litigants against lawyers hired by institutions such as credit companies, landlords, and businesses (see, e.g., Elwell et al., 1990; Sterling et al., 1990). A few jurisdictions, such as California, responded by banning lawyers entirely from these courts. This did not always succeed in leveling the playing field in those courts, however, as institutional litigants began deploying well-educated nonlawyer employees to specialize in handling their cases in the small claims courtrooms (Moulton, 1969). Overall, studies of small claims courts have arrived at mixed findings about their fairness to individual litigants, especially in cases involving institutional opponents who were

represented by lawyers or nonlawyer advocates (Yngvesson and Hennessey, 1975).

Some jurisdictions also have created specialized forums roughly analogous to the small claims court model to handle certain narrow categories of cases (housing courts, labor courts, social welfare courts, and the like) with the expectation that their simplified processes would allow many disputants to represent themselves. Once again, the results are mixed (Cappelletti et al., 1978–79).

In addition to simplified courts within the court system itself, some jurisdictions have created administrative bodies empowered to resolve certain categories of disputes. For example, in the United Kingdom, 'tribunals' composed of part-time volunteers have been used to decide disputes arising out of public housing and other government benefit programs (Cappelletti et al., 1978–79). In some Canadian provinces, an agency called the 'Office of the Rentalsman' was set up to resolve most landlord–tenant disputes (Rentalsman). In many jurisdictions administrative hearings decide whether a worker is entitled to compensation for job-related injuries or whether a discharged employee will receive unemployment compensation. Once again, the procedures are designed to be simple compared to the regular courts and typically disputants are encouraged to appear without lawyers. A recent study of one such unemployment compensation hearing system reported that unrepresented employees fared as well as those who were represented by supervised Harvard law students (Greiner et al., 2012).

In recent years, private mediation and arbitration (often called Alternative Dispute Resolution) have emerged as major alternatives to the regular civil courts, especially in the USA. The primary motive behind the growth of these alternative forums, however, arises not from a desire to increase access to justice for lower income individuals, but to help more affluent parties, especially businesses, reduce cost, delay, and the risk of large damage awards (Abel, 1982; Johnson, 1993). While private mediation and arbitration have not proved a significant strategy for increasing access to justice when the disputes pit individuals against institutions, they have offered a useful alternative to litigation in disputes between evenly matched individuals (Menkel-Meadow, 1999).

Simplification of Substantive Legal Rules Governing Disputes

One of the major reasons people must employ lawyers to represent them is the complexity of the legal rules the judge or jury will apply to decide who wins. Some suggest we can reduce the need for lawyers and thus improve access for lower income disputants by simplifying those rules. What is called 'simplifying the law' seems to embrace three quite different meanings of the term.

For many, it merely means expressing the legal rules in 'plain language' that is, converting multisyllable technical legal terms into language understandable to people with a secondary or sometimes elementary school education. To the extent this can be done, this form of law simplification would seem to be helpful to unrepresented litigants. But many legal rules are based on concepts that remain complex and difficult to fully comprehend even when expressed in plain language. Furthermore, complexity also often arises because of the sheer number of rules that may interact to determine the proper resolution of a given dispute – the several elements of each possible cause of action, the several elements of each possible defense, and the like. No matter how simple the language and even the concepts that may be involved, it often takes expertise to work through the maze of rules. For these reasons, simplifying the language of law has its limits as a way of increasing access for unrepresented litigants.

A second form of law simplification reduces the number of rules that determine the outcome of a given class of cases. An example of this is the substitution of 'no-fault' divorce for a 'fault-based' system under which those wanting to terminate their marriage had to satisfy one or the other of several alternative grounds – cruelty, desertion, adultery, etc. In some jurisdictions, legislatures not only enacted no-fault divorce laws but authorized couples to obtain 'summary divorces' if they had no children, limited property, and short-term marriages (see, e.g., California Family Code). Japan has long had a system of 'mutual consent divorce,' where parties who have agreed to divorce and settled the terms of the divorce can merely register that event at the same government office where they obtained their marriage license, a process that accounts for about 90% of divorces in that country (Yokochi, 2009). In an ever more complex society, however, it is questionable whether there will be that many opportunities to reduce the number of legal elements required to decide cases – especially because most of the complexity is the product of interest groups and their advocacy for new laws or new interpretations of old laws they deem favorable to them.

Third, law simplification can mean the substitution of a single generalized standard in place of a sequence of precise rules the judge is expected to apply in deciding certain disputes. This grants the judge broad discretion in reaching what he or she perceives to be the correct resolution – to do justice as he or she sees it. It also means unrepresented parties do not have to attempt the daunting task of mastering a whole set of detailed legal standards that will determine the outcome of their disputes, lessening the disadvantage when contending with parties who have lawyers (Zorza, 2014). The potential downside of reliance on a general standard is its subjective nature. It offers no check on any biases the judge may harbor in favor or against certain parties or for a personally preferred outcome, making the process closer to a system 'of men, not laws.'

Use of Less Expensive, Nonlawyer Representatives

Some specialized lower cost forums permit nonlawyers to represent parties appearing before them. Some even rely exclusively on nonlawyer advocates for this role. Presumably the use of these less-educated and less-expensive personnel enhances access by lowering the cost of representation to the parties or to the government employing the advocates. A 1998 study in the United States evaluated the difference between representation by lawyers and representation by lay advocates in certain administrative forums, and found only marginal differences in the quality of representation or the results achieved (Kritzer, 1998b). Unresolved, however, is whether those findings would hold up in the regular courts that decide a broader range of disputes or where one litigant was

represented by a lawyer and the other a lay advocate. Another issue is whether nonlawyer advocates are necessarily less expensive than lawyers and thus their use can increase access to justice. The United Kingdom allows nonlawyers to represent litigants in certain of its tribunals. However, it is unlikely this lowers the cost and thus enhances access to justice because the fees are a percentage of the recovery and remain the same irrespective of whether the claimant's representative is a lawyer or a nonlawyer (Ministry of Justice, 2013).

Out-of-Court Assistance to Unrepresented Litigants

Several countries, especially those with only limited lawyer resources committed to legal aid, have seen unrepresented litigants engulf their courts. Some U.S. jurisdictions report 70–80% of litigants in family law court appearing without counsel, and as many as 90% of tenants are unrepresented in eviction cases while 90% of landlords are (Engler, 2010). Prompted as much by the judges as by advocates for the poor, and unwilling or unable to provide lawyers to the overwhelming numbers appearing in their courts, many of those jurisdictions have begun providing comparatively low-cost services designed to help those litigants prepare and present their own cases.

This assistance can take several forms, sometimes more than one available in a single jurisdiction. In some places, unrepresented litigants are offered classes that teach the participants how to draft pleadings and file them in court, along with an explanation of what they can expect when they appear in court and how litigants are expected to behave in those forums. At a more individualized level of assistance, what are often called 'self-help' centers offer one-on-one guidance from lawyers or, more frequently, supervised paralegals. At these centers, often located in courthouses, staff members help litigants fill out the pleadings or paperwork the courts require and show them how to conduct themselves in presenting their particular case to the judge. Some show videos to help litigants visualize what it will be like when they interact with a judicial officer. None of these centers, of course, provide representation in the courtroom – it is purely pretrial assistance they offer (Zorza, 2002).

Interviews and anecdotal evidence suggest judges find litigants who have received self-help assistance present the court with more coherent pleadings and other written documents and are somewhat better prepared to present their cases. In that sense, these services make the judge's job easier (Henschen, 2002; Graecen, 2002). Self-help assistance also usually is adequate – as well as far less expensive – in uncontested cases such as mutual consent divorces with no disputes over property, child custody, etc., and where the only task is properly preparing paperwork and navigating the process to an agreed-upon conclusion (Cavanaugh, et al., 1976). What is less clear is whether self-help assistance is sufficient to increase substantially the litigant's ability to win contested cases – with one study showing litigants receiving self-help assistance faring no better than those lacking that help (Steinberg, 2011) and another finding those represented by lawyers were three times more likely to win than those limited to out-of-court assistance (Greiner et al., 2013).

Even the most committed advocates for self-help assistance recognize unrepresented litigants – with or without pretrial assistance – find it difficult if not impossible to succeed in the extreme adversarial world typical of common law courts. Thus, these advocates have urged judges to relax the adversarial model when unrepresented litigants are involved in the proceedings and to assume more responsibility for eliciting the evidence needed to decide the issues in the case (Albrecht et al., 2003; Zorza, 2002). In response, some jurisdictions have enacted rules expressly authorizing judges to give assistance to unrepresented litigants and to assume a more inquisitorial role in such cases (California Judicial Council, 2007). Whether most judges will be willing to shift to that inquisitorial model and whether judicial budgets will allow them the time to do so remains a critical unresolved issue (Johnson, 2014).

Technological Substitutes for Legal Expertise

Several technologies have been deployed to give some level of legal assistance to low- and moderate-income people – often similar to the classes or the individualized help otherwise provided in person at self-help centers as described in the previous Section Out-of-Court Assistance to Unrepresented Litigants.

Many court systems or legal services programs in the United States have developed Web sites filled with forms, plain language explanations of the law, information about how to represent oneself in court, and other relevant information, often including video clips illustrating what to do and how to do it (Staudt, 2005).

Nearly every state now has at least one telephone hotline providing legal information to callers, many of which provide one-on-one advice and limited assistance as well as referrals for further legal help (Houseman, 2005).

In a more sophisticated form of technological support, some court systems and legal services providers have developed interactive computer programs that help the litigants produce appropriate pleadings to submit to court. Sitting down at a kiosk or a computer linked to the Web, an unrepresented litigant responds to a series of questions appearing on the screen. Once those questions have been answered, the kiosk or a printer attached to the computer the litigant is using spews out the relevant court documents and instructions about where and how to file them. That way the language of the street is turned into the language of the law. Typically, users also can click on videos that orient them to the tasks ahead when they take their cases to court (iCAN! Legal; Staudt, 2005).

There is survey evidence suggesting courts find the resulting documents a substantial improvement over what these litigants prepare on their own (Greasen, 2002). But as of yet there has been no assessment of outcomes to determine whether and how much these computer-generated litigation documents help litigants to prevail in court if they lack legal representation when they appear before the judge.

The Beginnings of Comprehensive Fully Integrated Systems Seeking to Provide Cost-Effective Justice to Low- and Moderate-Income People

The Netherlands appears to be leading the way in its efforts to integrate technology, nonlawyer legal services, self-help

assistance, and full representation into a comprehensive system for giving lower income people access to justice in the most cost-effective way. It has recently introduced a sophisticated interactive Web site called the Rechtwijzer (Roadmap to Justice). This Web site uses a series of questions based on a decision tree that encourages and seeks to allow prospective litigants to analyze and solve common legal problems – family, housing, consumer, etc. – on their own. Yet when users reach an impasse the site refers them to outside experts and organizations (Legal Aid Board). This same Web site includes a still more sophisticated interactive program for 'divorce and parenting' issues available at rechtscheidingsplan.nl (Legal Aid Board, Rechtwijzer 2.0, www.openlab.com/2014/01/16rechtwijzer-justice-platform). It guides the parties through the entire process – from the decision whether to divorce through the issues, such as child custody and support that must be confronted and resolved. Along the way the program encourages the parties to settle their differences through online negotiation or online mediation rather than moving ahead with litigation (Legal Aid Board; New Law Journal, 18 April 2013).

When the Rechtwijzer Web site proves insufficient or not attractive to citizens, they turn to the next layer of assistance – 'law counters' that provide in person advice and short-term assistance. These 'law counters' are staffed by lawyers and supervised paralegals and their services are not means-tested. When advice and short-term assistance are not enough to solve the problem or where the users have already been sued, those who are financially eligible are referred to private lawyers compensated by the government who will provide them the necessary service including representation in court (ILAG Netherlands National Report, 2011, 2013).

Technology, computer technology in particular, may have only begun to transform the dispute resolution process in ways that could improve access for lower income litigants. Video technology combined with smart phones, laptops, and tablets may make it possible to avoid long trips to courthouses and costly days off in order to hold court hearings and resolve disputes. Voice-recognition software combined with translation software may soon enable people who do not speak the language of the court or other forum to communicate easily with those serving on that court or forum. The words the litigants speak in their language will be heard by the decision-makers in theirs, and vice versa. Moreover, advanced artificial intelligence could lower the cost of litigation by quickly analyzing the law and the facts of a dispute and producing a suggested decision.

Future Issues in the Movement toward Equal Access to Justice

The major issues for government-funded legal aid and equal access to justice in the twenty-first century include (1) whether a 'globalization of constitutional values' will lead all nations to recognize a constitutional right to equal justice (and to counsel when needed) in civil cases; (2) whether those nations that already have such a right will develop delivery mechanisms and other dispute resolution reforms allowing them to guarantee effective access for lower income citizens at politically sustainable levels of public investment; (3) whether the vital importance of equal access to justice for democratic government and political stability will cause developing countries and former communist nations to institute significant legal aid programs; and (4) whether alternatives to counsel – nonlawyer advocates, simplified procedures and substantive law, self-help assistance, technology, and inquisitorial judges – will be able to afford low-income disputants fair and equal justice at less cost and in which circumstances that will be possible. As it has in the past, social science will have much to contribute not only to the design of these initiatives, but more importantly to their evaluation. The insistence on empirical proof will offer the most effective guarantee that what emerges indeed grants every citizen, irrespective of means, the right to equal justice that democratic societies promise to all.

See also: Due Process in Law; Equality and Inequality: Legal Aspects; Equality of Opportunity; Equality: Philosophical Aspects; Inequality: Comparative Aspects; Justice and Law; Justice, Conflicts, and the Justice of Conflict Resolution; Justice: Philosophical Aspects; Justice: Social Psychological Perspectives; Law and Democracy; Law, Mobilization of; Lawyers, Regulation of; Lawyers: Social Organization of the Profession; Legal Culture and Legal Consciousness; Litigation; Poverty Law: United States; Rights: Legal Aspects; Social Justice and Social Inequalities.

Bibliography

Abel, R. (Ed.), 1982. The politics of informal justice: the American experience, vol. 1. Academic Press, New York. Comparative Studies, vol. 2.
Airey v. Ireland, 1979. Eur. Court H.R. (Ser. A) Judgment of Oct.
Albrecht, R., Greacen, J., Hough, B., Zorza, 2003. Judicial techniques for cases involving self-represented litigants. Judges' Journal 42, 16.
American Bar Association, 1994. Legal Needs and Civil Justice: A Survey of Americans, Major Findings from the Comprehensive Legal Needs Study. American Bar Association, Chicago.
AmLaw, 2012. Pro Bono Report (2012). American Lawyer. http://www.americanlawyer.com/PubArticleTAL.jsp? id=1202560034999&2012_Pro_Bono_Survey&slreturn=20130126003728.
Buckhannon v., 2001. West Virginia Department of Health and Human Resources, 532 U.S. 598.
California Applicant Attorneys Association. Web site at: www.caaa.org.
California Civil Code. Section 1717.
California Family Code. Sections 2400–2406.
California Judicial Council., Administrative Office of the Courts, 2007. Handling Cases Involving Self-Represented Litigants: A Benchguide for Judicial Officers. California Judicial Council, San Francisco section 2–2.
Caplovitz, D., 1974. Consumers in Trouble: A Study of Debtors 8–9, 215–221.
Cappelletti, M., Gordley, J., Johnson Jr., E., 1975, 1981. Toward Equal Justice: A Comparative Study of Legal Aid in Modern Societies. Oceana, New York.
Cavanaugh, R., Rhode, D., 1976. The unauthorized practice of law and pro Se divorce. An Empirical Analysis, Yale Law Journal 86 (104), 130–141.
Cappelletti, M., Bryant, G. (Eds.), 1978–79. Access to Justice, 3 vols. Giuffre, Milan.
Currie, A., 2000. Legal aid delivery models in Canada: past experience and future developments. University of British Columbia Law Review 33, 285.
Currie, A., 2007. The Legal Problems of Everyday Life: The Nature, Extent and Consequences of Justiciable Problems Experienced by Canadians. Department of Justice-Canada, Ottawa.
iCAN! Legal, Orange County (California) Legal Aid Society. at: www.iCAN!Legal.
Elwell, S., Carlson, C., 1990. The Iowa small claims court: an empirical analysis. Iowa Law Review 75 (422), 511.

Engler, R., 2010. Connecting self-representation to civil Gideon: what existing data reveal about when counsel is most needed. Urban Law Journal 37 (37), 66–72.
Farmer, A., Tiefenthaler, J., 2003. Explaining the recent decline in domestic violence. Contemporary Economic Policy 21, 158.
Florida Bar Association, 2014. Rule 4–6.1(d), Pro Bono Public Service, Rules Regulating the Florida Bar.
Galanter, M., 1974. Why the 'haves' come out ahead. Law and Society Review 9, 95.
Graham v. Daimler Chrysler, 2004. 34 Cal. 4th 553.
Greacen, J., 2002. Self-represented Litigants and Court and Legal Services Responses to Their Needs: What We Know. www.courtinfo.ca.gov/programs/cfcc/pdffiles/SRLwhatweknow.pdf.
Greiner, J., Pattanayak, C., 2012. Randomized evaluation in legal assistance: what difference does representation make. Yale Law Journal 121, 2118.
Greiner, J., Parranayak, C., Hennessey, J., 2013. The limits of unbundled legal assistance: a randomized study in a Massachusetts district court and prospects for the future. Harvard Law Review 126 (901), 907–910, 920–32.
Gunn, S., 1995. Eviction defense for poor tenants: costly compassion or justice served. Yale Law and policy Review 13 (385), 411.
Hall, R., 1991. Eviction Prevention as Homelessness Prevention; the Need for Access to Legal Representation for Low-Income Tenants.
Henschen, S., 2002. Lessons from the Country: Serving Self Represented Litigants in Rural Jurisdictions.
Houseman, A., 2005. The Future of Legal Aid in the United Sates. www.clasp.org/publications/future_legal_aid.pdf. at 9.
Huls, N., 1994. From pro Deo practice to a subsidized welfare state provision: twenty-five years of providing legal services to the poor in the Netherlands. Maryland Journal of Contemporary Legal Issues 4, 326.
ILAG (International Legal Aid Group), 2005/2007/2009/ 2011/2013. National Reports for Nations Cited in Article, All Found. http://ilagnet.org/about.php.
Johnsen, J., 1994. Nordic legal aid. Maryland Journal of Contemporary Legal Issues 4, 301.
Johnson Jr., E., 1978. Justice and Reform: The Formative Years of the American Legal Services Program. Transaction Press, New Brunswick, NJ.
Johnson Jr., E., 1981. Lawyers' choice: a theoretical appraisal of litigation investment decisions. Law and Society Review 15, 567.
Johnson Jr., E., 1993. National report on American law: preserving justice in an era of global economic competition. In: The International Symposium on Civil Justice in the Era of Globalization, vol. 211. Shinzan, Tokyo, pp. 218–219.
Johnson Jr., E., 1994. Toward equal justice: where the United States stands two decades later. Maryland Journal of Contemporary Legal Issues 5199.
Johnson Jr., E., 2000. Equal access to justice: comparing the United States and other industrial democracies. Fordham International Law Journal 24, 83.
Johnson Jr., E., 2008. Justice for America's poor in the year 2020: some possibilities based on experiences here and abroad. DePaul Law Review 58, 393.
Johnson Jr., E., 2010. Equality before the law and the social contract: when will the United States finally guarantee its people the equality before the law the social contract demands. Fordham Urban Law Journal 37, 157.
Johnson Jr., E., 2014. Santa Barbara. In: To Establish Justice for All: The Past and Future of Civil Legal Aid in the United States, 3 vols. Praeger.
Judicial Council of California, 2005. Model Self-Help Pilot Program: A Report to the Legislature.
Klaber, M., 2006. A uniform fee-setting system for calculating court-awarded attorneys' fees: combining ex ante rates with a multifactor lodestar method and a performance-based mathematical model, 6. Maryland Law Review 66, 228.
Kritzer, H., 1998a. Contingent-fee lawyers and their clients: settlement expectations and settlement realities. Law and Social Inquiry 23, 795.
Kritzer, H., 1998b. Legal Advocacy: Lawyers and Nonlawyers at Work. University of Michigan, Ann Arbor, MI.
Lassiter v. North Carolina. 1981 452 US 18.
Lawrence, S., 1990. The Poor in Court: The Legal Services Program and Supreme Court Decision-Making. Princeton, Lawrenceville, NJ.
Legal Aid Agency, 2013–14. Business Plan 2013–14. UK Ministry of Justice.
Leubsdorf, J., 1984. Toward a history of the American rule on attorney fee recovery. Law & Contemporary Legal Problems 4, 9.
Lu, A., 2012. Do You Need a Lawyer for a Workers' Comp Case?. http:/blogs.findlaw.com/injured/2012/12/do-you-need-a-lawyer-for-a-workers-comp-case.htm.
Menkel-Meadow, C., 1999. Do the 'haves' come out ahead in alternative judicial systems? 'Repeat players' in ADR. Ohio State Journal on Dispute Resolution 15, 19.
Ministry of Justice (UK), 2013. No Win No Fee Agreements. www.justice.gov.uk/civil.../main-changes.
Moorhead, R., 1998. Legal aid in the eye of the storm: rationing, contracting and a new institutionalism. Journal of Law and Society 25, 365.
Moorhead, R., 2000. Conditional fee agreements, legal aid, and access to justice. University of British Columbia Law Review 23, 471.
Moulton, B., 1969. The persecution and intimidation of the low-income litigant as performed by the small claims court in California. Stanford Law Review 21, 1657.
Murr, G., 2000. Analysis of the valuation of attorney work product according to the market for claims: reformulating the lodestar method. Loyola University-Chicago Law Journal 31, 599.
New Law Journal, 18 April 2013. The New Model. www.newlawjournal.co.uk/nlj/content/new-model.
O'Brien, W., 1967. Why not appoint counsel in civil cases: the Swiss approach. Ohio State Law Review 28 (1).
Regan, F., Paterson, A., Goriely, T., Fleming, D. (Eds.), 1999. The Transformation of Legal Aid: Comparative and Historical Studies. Oxford University Press, Oxford, UK.
Rentalsman, Office of, New Brunswick. http://www.snb.ca/e/1000/1000-2/e/1000-2_003_e.asp#1.
Reynolds, S., Harris, C., Peoples, R., 2007. Back to the future: an empirical study of child custody outcomes. North Carolina Law Review 85, 1619.
Sandfur, R., 2010. Impact of counsel: an analysis of empirical evidence. Seattle Journal of Social Justice 9, 51.
Schefergegen Appenzell A.R.H. Regierungsrat, Judgment of 8 October 1937, BGE 63 la 209(1937).
Schmertz, T., 1967. The indigent civil plaintiff in the District of Columbia: facts and commentary. Federal Bar Association Journal 27, 235–243.
Schwartz, M., Mitchell, D., 1970. An economic analysis of the contingent fee in personal-injury litigation. Stanford Law Review 27, 1125.
Seron, C., et al., 2001. The impact of legal counsel on outcomes for poor tenants in New York City's housing court: results of a randomized experiment. Law & Society Review 35, 419.
Singer, L., 1979. Nonjudicial dispute resolution mechanisms: the effects on justice for the poor. Clearinghouse Review 13, 569.
Staudt, R., 2005. Technology for justice customers: bridging the digital divide facing self-represented litigants. University of Missouri Law Journal on Race, Religion, Gender & Class 5 (71), 71–76.
Steinberg, J., 2011. In pursuit of justice? Case outcomes and the delivery of unbundled legal services. Georgetown Journal on Poverty Law and Policy 18 (453), 482–495.
Sterling, H., Schrag, P., 1990. Default judgments against consumers: has the system failed? Denver University Law Review 67, 357.
Turner, J.C, McGee, J.A., 2000. Small Claims Reform: A Means of Expanding Access to the American Civil Justice System District of Columbia Social Law Review 5, 177.
United Nations, 1966. International Covenant on Civil and Political Rights. Article 14, section 1, 99 U.N.T.S. 171.
World Justice Project, 2011. The Rule of Law Index 2011 Report.
Yngvesson, B., Hennessey, P., 1975. Small claims, complex disputes: a review of the small claims literature. Law and Society Review 9, 219.
Yokochi, M., 2009. Civil Law Issues Related to Psychiatric Treatment and Divorces Not Included in the Statistics: Filing for Uncontested Divorce without the Spouse's Consent. www.repository.oai.yamaguchi-u.ac.jp.
Zemans, F., 1979. Perspectives on Legal Aid. Frances Pinter, London.
Zorza, R., 2002. The Self-Help Friendly Court: Designed from the Ground Up to Work for People without Lawyers. National Center for State Courts, Charlottesville.
Zorza, R., 2014. Some first thoughts on court simplification: the key to civil access and justice transformation. Drake Law Review 845, 863–864.

Justice and Health

Johannes Siegrist, Heinrich-Heine-University of Duesseldorf, Duesseldorf, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This contribution explores adverse effects of experienced injustice on health. At the macrosocial level, income inequality is one such source of experienced injustice, which is associated with population health in epidemiological studies. At the mesosocial level, perceived unfairness in organizations (organizational injustice) has been associated with poor health of employees. Similarly, violations of the principle of justice of exchange at work elicit stress reactions with adverse long-term effects on health, as documented by research based on the effort–reward imbalance model. Unjust social exchange is an important psychosocial determinant of health.

Background

The association of justice with health is commonly interpreted in the frame of unequal access to health care. In fact, being deprived from health care provision in the presence of an urgent health problem violates a basic human right. In order to tackle this basic injustice all health care systems have developed procedures, which, to different extents, regulate people's access to health care. Yet, given the universal dilemma between demand and supply, inequalities in the provision of health care are widespread in a majority of health care systems worldwide. A vast literature in the fields of health economics, medical ethics, and medical sociology deals with core questions of injustice in this context (*see* Socioeconomic Status and Health Care).

The topic of this contribution differs from this main line of reasoning. Here, we address the question whether the experience of injustice has an impact on human health and well-being and, if so, what mechanisms are responsible for this association. How can we define and measure experienced injustice and what is the empirical evidence for its association with health? In other words, we maintain that experienced injustice acts as a distinct psychosocial stressor that generates strong negative emotions and is associated with psychobiological stress responses with adverse long-term consequences on health.

In most instances, injustice cannot be observed directly. Apart from the extreme situation where injustice manifests itself as direct physical violence there is a broad spectrum of conflicting social interactions or violations of social norms, which evoke experiences of injustice. It is crucially important to identify those socioenvironmental constellations that trigger experiences of injustice. Yet, it is equally important to assess the subjective appraisals of persons exposed to these encounters as these appraisals affect the individual consequences.

Against this background different notions of 'justice' are mentioned, and a leading psychological theory of perceived justice is described. In the following sections three main scientific approaches toward studying health effects of perceived justice or injustice and their main findings are summarized, dealing with income inequality, organizational injustice, and injustice of exchange at work.

Notions of Justice

Justice is a broad concept composed of different dimensions. One dimension addresses the preferences and behaviors of individuals who are prepared to respect others, to treat them fairly, and not to harm them. In this sense, being just is an individual attribute or disposition. Another dimension of justice concerns its corrective function, the criminal law. It regulates the compensation of all types of criminal offense by applying sanctions and punishments. Procedural justice is a third dimension. Here, the process of how established rules are observed and how decisions are reached matters most. Core aspects include the equality of opportunity, the transparency of information, the avoidance of biased judgments, and the justification of decisions. Nevertheless, in everyday life, two further dimensions of justice are particularly important, distributive justice and justice of exchange. These notions relate to all types of social cooperation and conflict between people. Distributive justice deals with the principles of how scarce and desired resources are distributed among individuals, whereas the principle of justice of exchange maintains that benefits received from a partner should be reciprocated by equitable returns. Some form of inequality is inherent in many cases of distributive justice, while equity is the leading principle of justice of exchange.

There are several other notions of justice (e.g., interpersonal, informational, intergenerational, environmental justice), but distributive justice and justice of exchange have eminent significance for human health because they mainly concern the things 'worth driving for' in social life, such as appreciation, love, trust, material rewards, security, control, engagement, and belonging.

Substantial progress in analyzing the psychological processes generating experiences of injustice was achieved by J.S. Adams (1965). In his equity theory he maintained that a person (A)'s costs invested in an activity that is useful for another person (B) are expected to be compensated by comparable investment from that person (B) toward person (A). The perception of justice or injustice is contingent (1) on the notion of balance between 'cost' (transferred from A to B) and 'gain' (transferred from B to A), and (2) on the fit of this balance or ratio with a comparable balance or ratio of 'cost' and 'gain' experienced by a reference person (C) or a reference

group. Reference persons or a reference group are usually defined by horizontal comparison, where persons living in similar socioeconomic or sociocultural circumstances are chosen as a standard of evaluation. According to Adams, two types of perceived distributive injustice are to be distinguished, overfitting and underfitting. In the first case, the ratio of 'gain' relative to 'cost' in person (A) is higher than the corresponding ratio in reference person (C). A feeling of unease due to unjustified advantage may result from this experience. In case of underfitting, the ratio of 'gain' relative to 'cost' in person (A) is lower than the corresponding ratio in reference person (C), giving rise to negative emotions of anger and irritation. Equity theory posits that both states of injustice trigger some kind of inequity aversion that motivates people to reduce these discrepancies by behavioral or cognitive changes with regard to their 'costs' or 'gains.' A stronger motivation of change is expected in case of underfitting as compared to overfitting, given the powerful impact of loss experience.

In Adams' concept, 'inequity distress' was introduced as a hypothetical construct which was not measured in empirical tests of the theory, but which was useful in explaining subsequent behavioral or cognitive changes (Greenberg, 2010). Adams' main interest was to understand adaptive behavioral changes following perceived inequity rather than to explain individual health effects. This latter interest received primary importance in the frame of a more recent development, the effort–reward imbalance model that, in part, builds on Adams' equity theory (see Section Health Effects of Injustice: The Model of Effort–Reward Imbalance).

The following sections summarize core findings of three main scientific approaches toward studying health effects of perceived justice or injustice. The first approach represents a macrosocial perspective where inequity refers to the distribution of income across society. In the second approach, the concept of organizational justice developed in organizational psychology and management sciences is applied to epidemiological studies testing associations of organizational procedures with employees' health. In the third approach, injustice of exchange is analyzed in the frame of employment conditions where efforts spent by working people are not reciprocated by appropriate rewards. Adverse health effects are examined with reference to psychobiological stress models.

Health Effects of Injustice: Income Inequality

One of the most robust findings of social epidemiology documents a social gradient of morbidity and mortality in modern societies: the lower peoples' positions in the social hierarchy, the worse their health (Marmot and Wilkinson, 2006). Income is an important determinant of socioeconomic position, demonstrating strong effects on health and life expectancy (*see* Socioeconomic Status and Health). However, if correlations of household income with mortality, or correlations of gross national product with life expectancy, are compared across countries this correlation almost disappears in countries with the highest income levels. Here, it appears that the distribution of income influences average life expectancy of a society more so than the absolute level (Kawachi, 2000). In countries with a high degree of income inequality population health seems to be worse than in countries with a more egalitarian distribution. While high income inequality is interpreted as a condition amenable to experiences of distributive injustice, several questions need to be answered: Is the correlation of income inequality with population health restricted to cross-country comparisons, or does it equally exist within countries? Does the correlation persist after taking major confounding factors into account? And how can these associations be explained?

There are several ways of measuring the distribution of income (Kawachi, 2000). Among these, the Gini coefficient is most often used and best known. It theoretically ranges from 0.0 (perfect equality) to 1.0 (perfect inequality). In a majority of cross-national studies strong correlations between national Gini coefficient and mortality rate (e.g., age-specific, all-cause mortality) were observed. However, given a restricted comparability of respective data, a more convincing test may result from within-country analyses. Several such analyses are now available. They confirm the reported association, most impressively for the USA, where income inequality and poverty together accounted for about one-fourth of the state variations in total mortality (Kawachi, 2000).

When interpreting these associations one should notice that they represent ecological correlations, thus being subject to ecological fallacy. An association observed at the level of aggregate data cannot readily be translated to the level of individual data. Several confounding factors (e.g., individual household income, household size, education) have to be taken into account in order to disentangle compositional effects (e.g., a high proportion of poor people with elevated mortality risks living in areas with high income inequality) from contextual effects (the direct impact of income inequality in a certain area on individual health of its inhabitants). Although methodological concerns still prevail in this research there is some evidence of contextual effects of wide income distribution on poor population health (Wilkinson and Pickett, 2011). Therefore, it is important to know how this type of distributive injustice may affect health.

One explanation maintains that societies with a high level of income inequality are less likely to invest into public goods and services, thus weakening the quality of infrastructure and the social capital or social cohesion among residents of respective areas. Material and psychosocial adversity in turn may exert negative effects on population health. Another explanation focuses on negative emotions and stress reactions following exposure to relative deprivation. High income inequality evokes feelings of being treated unfairly and socially excluded among those who are at a clear distance from the top. Perceived distributive injustice is associated with frustrated expectations, feelings of inferiority, disrespect, and even shame experiences with adverse effects on health (Wilkinson and Pickett, 2011). Although plausible these associations have not yet been examined with convincing experimental and epidemiological evidence (see Section Health Effects of Injustice: The Model of Effort–Reward Imbalance). Therefore, more research on the association of income inequality with population health is needed, given the high sociopolitical and economic importance of this topic.

Health Effects of Injustice: The Concept of Organizational Justice

Whereas research on income inequality addresses the macrosocial level the concept of organizational justice relates to the mesosocial level of institutions and organizations. The notion of organizational justice was elaborated in organizational psychology and management sciences several decades ago, and was applied to health research only a couple of years ago in the frame of investigating psychosocial determinants of work-related disorders in epidemiological cohort studies (Elovainio et al., 2002). Three dimensions are distinguished in this concept: distributive justice, the perceived fairness of an organization's distribution of relevant resources; procedural justice, the perceived fairness of observing rules and applying agreed-upon standards in decision making; and interactional justice, the perceived fairness of interaction and communication in organizations, in particular between managers and employees (e.g., being treated with respect; relational justice; being informed in an accurate and timely way (informational justice) (Adams, 1965; Greenberg, 2010)).

It is obvious that each dimension contributes to a positive climate in organizations, strengthening job performance, and organizational commitment among employees. What is the evidence that perceived justice in organizations is associated with mental or physical health? To answer this question different scales measuring perceived justice were applied in epidemiological studies. As the most frequently used instrument focuses on procedural and interactional justice (Moorman, 1991), fewer results are available so far on the effects of distributive justice on health (see Section Health Effects of Injustice: The Model of Effort–Reward Imbalance). In a recent review on mental health, it was concluded that 10 prospective studies found higher mental health risks among employees reporting poor relational justice, with increases in the outcome measures ranging from 1.2 to 1.6. Similarly, low procedural justice was associated with higher risk of future depressive symptoms and medically certified depression in six prospective studies, with relative effects ranging from 1.4 to 1.9 (Ndjaboué et al., 2012). Additional findings confirmed elevated risks of sickness absence and of coronary heart disease in association with low organizational justice (Kivimäki et al., 2005; Greenberg, 2010).

In conclusion, epidemiological evidence points to a robust association of some aspects of perceived organizational injustice with poorer mental health. Additional health risks were also explored in a few studies, but the pathways linking organizational injustice with health outcomes are still poorly understood. At the level of worksite health promotion programs these results support investments in leadership training and improved interactional skills in organizations.

Health Effects of Injustice: The Model of Effort–Reward Imbalance

'Effort–reward imbalance' was proposed as a stress-theoretical model of a health-adverse psychosocial work environment that is based on the notion of justice of exchange (Siegrist, 1996). Social reciprocity, a fundamental principle of transactions that are characterized by some form of utility, lies at the core of the work contract, which defines distinct obligations in exchange for rewards (money, appreciation, career opportunities (promotion, job security)). The model asserts that failed reciprocity, in terms of high effort and low reward, generates strong negative emotions and associated stress reactions with adverse long-term health consequences. Failed reciprocity occurs frequently if the workers have no alternative choice in the labor market or if they are exposed to highly competitive jobs. Moreover, people with excessive work-related motivations and attitudes are at risk of experiencing this imbalance. Although there is conceptual overlap with the analysis of distributive justice in equity theory (Adams, 1965), the model has its original focus in sociological role theory, specifically the work role, emphasizing stressful effects of threats to social status and related rewards. In addition, the model combines extrinsic (situational) and intrinsic (personal) conditions of continued (chronic) imbalance, and it provides a new explanation of the occurrence of stress-related disorders by linking the experience of failed reciprocity in core social roles to the brain reward system and related psychobiological processes (Siegrist, 2005). The model also differs from the concept of organizational justice as it is applied to social exchange beyond work-related organizational contexts; in addition it explicitly deals with interpersonal comparisons beyond the relevant 'give' and 'take' relationship.

Effort–reward imbalance at work, as measured by a standardized, psychometrically validated questionnaire (Siegrist et al., 2004), was shown to be associated with significantly elevated risks for several chronic conditions in epidemiological studies, in particular depression (Rugulies et al., 2013) and coronary heart disease, including some of its main risk factors (Greenberg, 2010; Steptoe and Kivimäki, 2012; Xu et al., 2009). Importantly, these findings result from studies in different regions of the world, thus pointing to a rather basic type of stressful human experience. This latter notion is supported by evidence demonstrating associations of failed reciprocity in other social roles (e.g., volunteering, caring, homework, close social relationships) with several health indicators (Chandola et al., 2007) (*see* Stress at Work).

In addition to epidemiological findings this model underwent careful testing in experimental and 'naturalistic' investigations. Single or combined scales measuring effort–reward imbalance at work were related to dysregulated hormonal profiles, altered cardiovascular responses, and markers of reduced immune system function (Chandola et al., 2010). Moreover, first intervention studies revealed that mental health improves following a theory-based work stress reduction program (Bourbonnais et al., 2011).

Taken together, the available evidence of research on violations of the strong principle of justice of exchange in core social roles, and specifically in the work role, documents elevated subsequent risks for several disorders, such as depression or cardiovascular disease. As social reward frustration is generally more frequent among people with lower socioeconomic status, and as these stress-related disorders follow a similar social gradient, interventions should target socioeconomically deprived group with priority. Such interventions are not restricted to the individual, group, or organizational level, but may include national labor and social policies that aim at improving justice of exchange in core areas of everyday life.

Conclusion

Experiencing justice or injustice in social life matters for human health and well-being. Although these experiences reflect people's subjective appraisals they usually evolve from unequal social exchange and social deprivation, triggered by disregard of notions of fairness and by violation of underlying social norms. Fairness as a social preference seems to play a relevant role in human evolution (Fehr and Fischbacher, 2003), and its disregard is likely to evoke strong negative emotions and stress reactions. In three areas of unequal social distribution, we demonstrated elevated risks of stress-related physical and mental disorders: income inequality, organizational injustice, and inequity of exchange in core social roles, such as the work role.

There is much room for extending this line of research in several directions. One direction concerns the comparative cross-country analysis of distributive injustice, for example, addressing access to basic resources including employment and good quality of work, social security, or environmental health. In another direction, the temporal dimension in terms of intergenerational justice of exchange may be explored. Furthermore, additional experimental evidence on the effects of microsocial inequity on brain functioning and behavioral responses is of interest. In terms of policy implications intervention studies are of primary importance as they demonstrate the potential benefit of reduced injustice in social life for human health and well-being.

See also: Occupational Health; Social Integration, Social Networks, and Health; Socioeconomic Status and Health Care; Socioeconomic Status and Health; Stress at Work.

Bibliography

Adams, J.S., 1965. Inequity in social exchange. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 2. Academic Press, New York, pp. 267–299.

Bourbonnais, R., Brisson, C., Vézina, M., 2011. Long-term effects of an intervention on psychosocial work factors among health care professionals in a hospital setting. Occupational and Environmental Medicine 68, 479–486.

Chandola, T., Marmot, M., Siegrist, J., 2007. Failed reciprocity in close social relationships and health: findings from the Whitehall II study. Journal of Psychosomatic Research 63, 403–411.

Chandola, T., Heraclides, A., Kumari, M., 2010. Psychophysiological biomarkers of workplace stressors. Neuroscience and Biobehavioral Reviews 35, 51–57.

Elovainio, M., Kivimäki, M., Vahtera, J., 2002. Organizational justice: evidence of a new psychosocial predictor of health. American Journal of Public Health 92, 105–108.

Fehr, E., Fischbacher, U., 2003. The nature of human altruism. Nature 425, 785–791.

Greenberg, J., 2010. Organizational injustice as an occupational health risk. The Academy of Management Annals 4, 205–243.

Kawachi, I., 2000. Income inequality and health. In: Berkman, L.F., Kawachi, I. (Eds.), Social Epidemiology. Oxford University Press, Oxford, pp. 76–94.

Kivimäki, M., Ferrie, J.E., Brunner, E., Head, J., Shipley, M.J., Vahtera, J., Marmot, M., 2005. Justice at work and reduced risk of coronary heart disease among employees. Archives of Internal Medicine 165, 319–326.

Marmot, M., Wilkinson, R.G. (Eds.), 2006. Social Determinants of Health. Oxford University Press, Oxford.

Moorman, R.H., 1991. Relationship between organizational justice and organizational citizenship behaviors: do fairness perceptions influence employee citizenship? Journal of Applied Psychology 76, 845–855.

Ndjaboué, R., Brisson, C., Vézina, M., 2012. Organisational justice and mental health: a systematic review of prospective studies. Occupational and Environmental Medicine 69, 694–700.

Rugulies, R., Aust, B., Madsen, I.E.H., Burr, H., Siegrist, J., Bültmann, U., 2013. Adverse psychosocial working conditions and risk of severe depressive symptoms. Do effects differ by occupational grade? European Journal of Public Health 23, 415–420.

Siegrist, J., 1996. Adverse health effects of high effort-low reward conditions at work. Journal of Occupational Health Psychology 1, 27–43.

Siegrist, J., Starke, d, Chandola, T., Godin, I., Marmot, M., Niedhammer, I., Peter, R., 2004. The measurement of effort-reward imbalance at work: European comparisons. Social Science & Medicine 58, 1483–1499.

Siegrist, J., 2005. Social reciprocity and health: new scientific evidence and policy implications. Psychoneuroendocrinology 30, 1033–1038.

Steptoe, A., Kivimäki, M., 2012. Stress and cardiovascular disease. Nature Reviews in Cardiology 9, 360–370.

Wilkinson, R.G., Pickett, K., 2011. The Spirit Level: Why Greater Equality Makes Societies Stronger. Bloomsbury Press, New York.

Xu, W., Zhao, Y., Guo, L., Guo, Y., Gao, W., 2009. Job stress and coronary heart disease: a case-control study using a Chinese population. Journal of Occupational Health 51, 107–113.

Justice and Law

David Luban, Georgetown University Law Center, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article proposes that the commonsense conception of legal justice is simply treating people as the law provides. Law, however, can be flawed from the standpoint of extralegal justice. In that case, the term 'legal justice' sounds like a misnomer – a wrongful appropriation of the terminology of justice viewed as a moral virtue. The article examines two avenues of response to this problem: first, theories that narrow the gap between legal and extralegal justice, and second, theories according to which the central virtue of legal systems is something other than justice. Among the former, the article discusses natural law theories, moralized theories of interpretation, equity, and certain forms of alternative dispute resolution. Among the latter, it discusses economic, pragmatic, and feminist alternatives to a justice-centered account of law's virtues.

Justice has been called the fundamental moral virtue of social institutions (Rawls, 1971). In similar fashion, 'legal justice' is the name commonly given to the fundamental moral virtue of legal institutions. Theorists through the centuries have differed substantially over what this virtue is, and what relation legal justice bears to other forms of justice. In its commonsense meaning, according to which legal justice means treating people as the law provides, legal justice helps ensure that people are treated consistently and predictably. At the same time, however, it seems unlikely that treating people as the law provides invariably means treating them justly in a moral sense, because the law is at best an imperfect instrument of corrective justice, distributive justice, or even elementary decency. Some theorists have responded to this concern by offering alternative accounts of legal justice. Others, of more skeptical bent, have doubted that legal justice has much to do with moral justice, either because they deny that legal institutions have moral aims, or more commonly because they believe that other virtues, such as economic efficiency, pragmatic problem solving, or compassion, are more fundamental aims of legal institutions.

The Commonsense Conception of Legal Justice

In its primary everyday sense, legal justice means treating people as the law provides. This view may be referred to as the commonsense conception of legal justice. Under the commonsense conception, law, rightly interpreted and correctly applied to facts determined rationally, yields legal justice. Indeed, 'legal justice' turns out to be nothing more than the name for the state of affairs in which law is rightly interpreted and correctly applied to facts determined rationally; the phrase has no legitimate meaning other than this. Under the commonsense conception, legal justice is roughly synonymous with legality itself. It is what most people mean when they speak of the 'rule of law,' an abbreviation for 'the rule of law, not of men.' Where the rule of men subjects people to the will of officials, who may be partial or capricious, the rule of law impartially subjects people only to the law itself (*see* Rule of Law (and Rechtsstaat)).

In Western philosophy, it was Plato (428–347 BC) who in his late dialogue *The Laws* first introduced the idea that human rulers should be 'slaves of the law.' Aristotle (384–322 BC) also endorsed the rule of law idea, in his *Politics* (Bk. 3, ch. 16); it is also Aristotle who, in *Nichomachean Ethics* (Bk. 5), provided the first systematic analysis of the forms of justice. Notably, however, Plato did not entirely endorse the rule of law. Drawing on the analogy of physicians who change their course of treatment when they find it is killing their patients (Statesman 295c–297c), he argued that officials should depart from legal rules when the good of the city requires it. The fear that rigid legality may lead to injustice haunts the commonsense conception of legal justice.

The commonsense conception raises large jurisprudential questions. How is law interpreted rightly? How is it applied properly? How are facts determined rationally? Competing philosophical schools offer various theories of legal interpretation; proceduralists disagree about how to adjudicate properly; and evidence law balances epistemological accuracy against investigative cost and the protection of personal rights. As a result, the commonsense view that legal justice means treating people as the law provides is not as straightforward as it seems. It is fraught with ambiguity – an essentially contested concept if ever there was one. The commonsense conception of legal justice turns out to be an entire family of conceptions.

Perhaps the most fundamental divide among theories of legal interpretation and law application exists between views that try to bind decision makers through nonmoral general criteria and views that require decision makers to exercise moral judgment. Leading examples of the former are textualism – the idea that laws must be interpreted according to the exact wording of their texts – and intentionalism, the idea that laws must be interpreted according to the intention of those who enacted them, ascertained through objective historical evidence. An example of the latter is the theory of the distinguished philosopher Ronald Dworkin (1931–2013), who believes that judges are required to interpret law in the morally best light it can sustain (Dworkin, 1977, 1986). (For further discussion on this point, *see* Legal Reasoning and Argumentation.)

Justice and Legal Justice

Regardless of how its ambiguities are resolved, the commonsense conception implies that legal justice will never be any better than the law itself. If the law itself is unjust, then treating people as the law provides will produce injustice. US Supreme Court Justice Oliver Wendell Holmes wrote in a letter, "I have

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86043-2

said to my brethren many times that I hate justice, which means that I know if a man begins to talk about that, for one reason or another he is shirking thinking in legal terms" (Holmes, 1929). Legal justice, understood as legality, differs essentially from what might be called extralegal justice. If Holmes is right that legal institutions do not deliver extralegal justice, then two rather disquieting conclusions appear to follow. First, the term 'legal justice' is fundamentally misleading: the word 'justice' in it really does not mean justice in the moral sense that is its home terrain. Second, there may be no reason to think of legal justice as a moral virtue at all. A good murderer is not good, and legal justice may be unjust.

Moreover, even when a law is just, inflexible application may yield outcomes in particular cases that are inequitable – unjust, in another everyday sense of the term – because of the peculiar facts of the case. Laws are general in scope, but, as Aristotle observes, even a good general rule may yield an unjust outcome when the case does not fit the rule. Thus, alongside the problem of unjust laws there is a problem of unjust outcomes. Both problems generate a tension between justice and legal justice.

The Natural Law Alternative to the Commonsense Conception of Legal Justice

The most ambitious approach to the problem of unjust law simply denies the possibility that law can be unjust. At the first blush, this sounds absurd, but the proposition that unjust law is not really law at all (*lex injusta non est lex*) has had powerful defenders in the so-called natural law tradition, among them St Augustine and St Thomas Aquinas (*see* Natural Law). The natural law argument may be put as follows. Inherent in the idea of law is the claim that the law creates moral obligations over and above whatever penalties attach to it. That is why a tax law is different in kind from a gunman's command to hand over your wallet, and lawful authority is not merely a 'gunman writ large.' The difference is that one does nothing blameworthy by evading the gunman, but it is commonly believed that tax evasion is wrong. If the law is unjust, however, the moral obligation vanishes, and with it the status of law (Soper, 1984).

The leading natural law theorist John Finnis (b. 1940) puts the argument somewhat differently. Legal systems are instituted for a purpose, namely, to promote the common good. The common good requires that there be lawmakers and law enforcers, who solve the various coordination problems that arise in organized society. Their role in fulfilling the common good is the sole reason that lawmakers and law enforcers have authority over others, and why others ought to obey them. The law-making authority therefore has an implicit limit: if legislators are not aiming at the common good, they have exceeded their authority, and the laws they enact are laws in form alone (Finnis, 1980). This was the same conclusion reached by the German theorist Gustav Radbruch in a famous argument that racist Nazi laws were no laws at all (Radbruch, 1946/2006).

The intuitive basis to the natural law argument is that every recognizable legal system is built around a core set of rules, such as the proscriptions of violence, theft, and fraud. These are moral requirements recognizably directed toward the common good; they flow from the nature of human society, and ultimately from human nature itself. They are 'natural laws' in the sense that reasoning from human nature seems to dictate that they belong in every legal system. To natural lawyers, the idea that any arbitrary set of rules could constitute a legal system is unrealistic, because it denies that reason itself reveals the core of natural law at the heart of legal systems.

The natural law position can be understood as an amendment to the commonsense conception of legal justice. Where the latter understands legal justice to mean treating people as the law provides, whatever that may be, the natural law conception understands the very character of law to be such that it cannot treat people in a manner inconsistent with natural law.

To the natural law argument there are at least two responses. The most prominent is the argument of legal positivists (*see* Legal Positivism). Positivists hold that law consists at bottom of nothing other than posits of the sovereign authority. Any valid enactment by the sovereign authority is law, whether or not it is just. From the positivist point of view, natural lawyers are merely playing definitional games when they insist that a duly enacted statute should not be honored with the title 'law' unless it is just. Taking away the label 'law' does not take away the problem of unjust law, because officials remain duty bound to treat people as the positive law provides. The commonsense conception of legal justice cannot be merely defined away.

Furthermore, even if the natural law argument is right that not just any capricious set of rules counts as law, and that only a system devoted to the common good will be recognizable as a legal system, it does not follow that every single law in a legal system is just: a system devoted to the common good can contain flaws and blind spots. The natural law argument applies most plausibly to entire legal systems, not individual laws. Perhaps a largely unjust system is not really law, but an unjust enactment of a statute in a largely just system is presumptively law. In that case, the problem of unjust law reappears within the natural law conception of legal justice.

The American philosopher Lon Fuller proposed a distinctive version of the natural law argument. According to Fuller, a legal system must satisfy eight procedural requirements, which he called 'the inner morality of law' and 'procedural natural law.' Enactments must be general in scope, public, prospective rather than retroactive, clear, consistent, relatively constant over time, capable of being obeyed, and enforced as written. Fuller believed that law serves as a baseline around which citizens organize their activities, and citizens must be able to know, understand, and comply with the baseline. The eight procedural requirements ensure that law satisfies these conditions. For Fuller, then, legal justice has more specific content than the commonsense conception – law must satisfy the eight procedural requirements – but less content than the traditional natural law view.

Fuller's eight requirements have been widely accepted as an analysis of the rule of law. Fuller also believed that any legal system satisfying the eight procedural requirements would be substantively just, but his efforts to demonstrate this ambitious conclusion are unpersuasive. American slave law satisfied the eight procedural requirements; so did many legal regimes that treated women as mere property of their fathers or husbands. Such legal regimes were procedurally just but substantively unjust, and most theorists agree that Fuller's theory of

'procedural natural law' fares no better than traditional natural law theories in eliminating the problem of unjust law.

Building Justice-Enhancing Devices into the Law

An alternative approach to the problem of unjust law concedes that unjust law is possible, but tries to narrow the gap between justice and legal justice by building into legal systems devices for bringing law into line with justice. Legal language can be interpreted flexibly rather than rigidly, or interpreted in its morally best light, as Dworkin proposes. Many legal systems permit statutes to be invalidated when they conflict with constitutional principles, which are often framed in broadly moralistic language. In European law, the principal human rights treaty must be interpreted in the way that most effectively protects human rights, even if that is not the most literal interpretation; that is because of the European Convention on Human Rights' "special character as a treaty for the collective enforcement of human rights and fundamental freedoms" (*Soering v. United Kingdom*, 1989, para 87). In the United States, juries have the power to nullify unjust law by acquitting a criminal defendant in defiance of the law. Officials, too, may decline to enforce unjust laws, nullifying them through a kind of benign neglect. Arguably, nullification defeats legal justice in the name of extralegal justice; however, some theorists have pointed out that the discretion to nullify law is itself sanctioned by law, and therefore may be regarded as an escape clause built into the concept of legal justice (e.g., Kadish and Kadish, 1973; Butler, 1995). Law must be administered, and because laws are stated in general form, officials will always have some discretion to determine when laws apply; thus, the power to nullify will always exist.

Devices like these, which are internal to the law itself, may well lessen the tension between justice and legal justice, but few theorists believe that they eliminate it. No method of interpretation is flexible enough to yield any desirable result, nor is every unjust law automatically unconstitutional. Furthermore, nullification is a device better suited to the exceptional case than to the rule itself.

Law and Equity

What about situations in which a law is just, but applying it strictly in a particular case yields an unjust outcome? Aristotle defines the word equity as "a correction of law where it is defective owing to its universality," and argues that justice requires that law should yield to equity. This, arguably, is what nullifiers are up to. In Aristotle's memorable image, rules of law should be like the soft lead rulers that builders use to measure irregularly shaped molding: "the rule adapts itself to the shape of the stone, and so too the decree is adapted to the facts" (*Nichomachean Ethics*, Bk. 5, ch. 10). (For further discussion, *see* Equity in Law.)

But while the utility of soft rulers is clear, soft rules are more problematic. In contemporary terms, Aristotle was arguing that law should consist of standards rather than rules. A law that says 'drive no faster than 100 km per hour' is a rule; a law that says 'drive at a reasonably safe speed' is a standard. Standards, unlike rules, are adaptable to facts in the case-by-case way that Aristotle recommends.

There are two problems with replacing rules with standards, however. The first is that standards are useful only when there is widespread social agreement on how they apply across a broad range of cases. Where no such agreement exists, law must resort to rules if it is not to become fatally vague. For that reason, mature legal systems always contain both rules and standards.

The second problem is that the shift from rules to standards will not eliminate the need to supplement law with equity. Standards, like rules, are general, and can therefore yield injustice in particular cases. For example, a driver rushing a stroke victim to the hospital may be driving faster than 100 km per hour and driving at an unsafe speed. If it seems unjust to punish the driver, that will not be because the speed limit takes the form of a rule rather than a standard. Rather, it is unjust to punish the driver because the emergency situation does not fit any universal speed limit, either a rule or a standard.

Perhaps, then, the universality of laws should be abandoned. The problem with this approach is that the universality of legal rules and standards is the hallmark of laws, distinguishing them from orders and directives. Furthermore, the virtues of legal justice lie precisely in the universality of law. Because law is general, legal justice treats like cases alike, and in that way it is fair and impartial; in addition, generality makes the law predictable, which allows people to plan their affairs around the law. Treating people as the law provides honors the values of fairness and predictability, whereas rule by decree undermines both these values.

Alternative Dispute Resolution

One way to reduce the tension between law and equity is to divert a dispute from the formal legal system to mediation, arbitration, out-of-court settlement, or other forms of alternative dispute resolution (ADR). Proponents argue not merely that ADR can resolve disputes more quickly, cheaply, and cooperatively than formal adjudication, but also that by particularizing outcomes to the unique facts of each case ADR creates higher quality justice – it creates something that looks more like equity than law.

Here as elsewhere the argument is very ancient. For Greek authors of the Homeric period, the word 'justice' (*diké*) was synonymous with a mutually agreeable settlement. In *The Laws*, Plato distinguished between the best judge, who destroys the wicked, the second best judge, who subordinates the wicked to the good, and the third best judge, who reconciles the contending parties with each other. Plato argued that in a real, nonutopian city the third best judge, the one who settles cases through compromise, is better than the other two (627e–628a).

However, the virtue of ADR and the settlement process – that ADR makes an end run around legal formalities – raises questions about how to evaluate the justice of settlements. Negotiations, which often turn on nothing but naked bargaining power, can yield outcomes fully as inequitable as trials. Furthermore, critics of ADR suspect that informal processes may reinforce unequal power relationships, and that the lack of formal protections will leave vulnerable groups even more vulnerable. Should the justice of a mediated settlement be evaluated against the standard of an ideally fair trial (the 'shadow verdict'), or an ideally fair

negotiation (the 'shadow bargain'), or something entirely *sui generis*? How are any of these standards to be determined? Does compromise promote extralegal justice, or does it merely compromise legal justice? Such questions yield no easy answers. (For further discussion of this point, *see* Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Negotiation and Bargaining: Role of Lawyers.)

Origins of the Conflict between Justice and Legal Justice

How grave the tension between extralegal and legal justice is depends on what extralegal justice consists of. Philosophical theories of justice differ dramatically from each other (*see* Justice: Philosophical Aspects), and none commands anything resembling a consensus. When Holmes wrote 'I hate justice,' his skepticism can be read as a complaint about the vagueness and debatability – hence, the contentlessness – of the concept of extralegal justice.

Yet despite the philosophical controversies, there is a recognizable Western concept of justice of which the philosophical theories may be said to offer alternative interpretations or conceptions. This concept draws on the two chief sources of Western moral thought, Greek philosophy and the Bible.

For Plato and Aristotle, and likewise for the biblical authors and redactors, justice is primarily a virtue of persons rather than of institutions. Thus, Plato's *Republic* investigates justice in the city only because Plato believes that the city is a plausible analogue to the individual soul, his real topic of interest; so too, Aristotle places his account of justice in his treatise on ethics rather than in his treatise on politics. In the Bible, a just community is simply a community filled with just people. In this respect, the classical and biblical sources differ profoundly from contemporary theorists, who treat justice as a property of institutions rather than of persons.

It was Aristotle who first observed that the words 'unjust' and 'just' are ambiguous – the former refers both to 'the lawless man and the grasping or unfair man,' and similarly the latter refers both to the law-abiding and the fair man (*Nicomachean Ethics*, Bk. 5, ch. 1). One form that the virtue of justice assumes, then, is obedience to the law. Related to this form are special virtues of officials – to punish lawbreaking, and to judge impartially (which is a requirement found both in the Hebrew Bible and in Aristotle).

What about extralegal justice? Both the Jewish and Christian traditions locate the core ideal of justice in the Golden Rule and the injunction "Love your neighbor as yourself" (Leviticus 19:18; Matthew 22:39), which commentators often regard as alternative formulations of the same duty. In addition, both traditions lay special emphasis on the injustice of neglecting or oppressing the poor – a recurrent theme among the biblical prophets. The Hebrew word for almsgiving, *tzedakah*, derives from '*tzedek*,' the word for justice. *Tzedakah* is a duty of the just individual, and a community will be unjust to the degree that it neglects the needy. In contemporary discussions, people often use the term 'social justice' in precisely this way, to express a social duty of justice to the poor. Similar themes appear in non-Western religious traditions as well (including a version of the Golden Rule in the Analects of Confucius).

This sense of extralegal justice as solidarity with the least fortunate will inevitably generate conflicts with legal justice. Even within Leviticus 19, the duty to care for the needy lies in tension with the duty of judges not to favor the poor. Furthermore, as functionalists argue, laws aim to maintain and preserve a social order; law is therefore an innately conservative institution, whereas the prophetic conception of social justice is innately utopian and radical. One cannot really expect that laws in any society will be systematically formulated to enforce wealth transfers from the prosperous and powerful to the poor and powerless.

There is another reason as well to anticipate divergence between legal and extralegal justice. Since Aristotle, it has been traditional to distinguish distributive justice from corrective justice. The former concerns fundamental questions of distribution in society; the latter deals with the rectification of injuries, both civil and criminal. Legal institutions such as courts deal principally with issues of corrective justice, and have very little to do with distributive justice (which is not to deny that legal decisions often have distributive effects). As writers from Aristotle onward have observed, corrective justice may run contrary both to moral virtue and distributive justice. For, if a good person accidentally injures a wicked one, or a deserving poor person injures an unjustly wealthy one, the injurer will owe compensation, even though in the former case the good person may be ruined for the sake of the bad, and in the latter case the distribution of wealth becomes even more lopsided than it was before. Legal institutions take established distributive principles and patterns as a given, regardless of whether they are just. In this sense, then, even when courts achieve corrective justice, they may well reinforce distributively unjust practices.

Alternatives to Legal Justice

So far this article has surveyed four conceptions of legal justice: the commonsense conception, according to which legal justice secures the goods of impartial consistency and predictability in the law; the traditional natural law conception, according to which legal justice secures the common good; Fuller's 'procedural natural law,' according to which legal justice conforming to the eight canons of lawmaking provides an intelligible baseline against which people may order their own affairs; and the Aristotelian idea that genuine legal justice combines universal legal standards with equity. None of these is able to relieve the tension between legal and extralegal justice. The question remains whether other alternatives do better. Several alternatives have become prominent in contemporary debates.

Law and Economics

One response is simply to deny that achieving extralegal justice is an important aim of legal systems. A particularly influential contemporary version has been offered by law and economics theorists. According to Richard Posner (b. 1939), a founder of law and economics, the aim of legal systems is to maximize aggregate social wealth. If legal rules systematically fail to maximize social wealth, disputing parties could realize shared gains by bargaining around the law; inevitably, then, the law

gravitates toward wealth-maximizing rules. Furthermore, wealth maximization is a normatively desirable aim, better than utilitarianism because it cultivates productive virtues rather than pleasure seeking (Posner, 1983). Critics, however, have found both the descriptive and normative claims of law and economics exaggerated. Inefficient legal rules sometimes linger for centuries, and the argument that a law prohibiting murder or rape can be justified only if it can be shown to maximize social wealth seems unsupportable (Dworkin, 1980; Kronman, 1980). (For further discussion of this point, *see* Law and Economics: Empirical Dimensions.)

Pragmatism

Another prominent alternative to the commonsense conception of legal justice is pragmatism. As one proponent defines pragmatic adjudication, "a pragmatist judge always tries to do the best he can for the present and the future, unchecked by any felt duty to secure consistency in principle with what other officials have done in the past" (Posner, 1998: p. 237). To the objection that this is an invitation to lawlessness, pragmatists respond that doing the best for the present and the future will in some cases require the judge to apply law and precedent in a rigidly formalistic manner; in other cases, however, pragmatism requires judges to look past formalities.

Pragmatists collapse legal justice into extralegal justice, where the latter is understood as achieving the best consequences overall. In one respect, pragmatism descends from the legal realists (*see* Legal Realism), who argued that legal formality is merely a disguise for implicit policy choices, which should be made explicit and be guided by concern for consequences. In another sense, pragmatism represents a radical version of Aristotle's plea for context-specific equitable standards. The attractions of pragmatism are plain, but the general objection to broad standards applies in an acute form to pragmatism: there will seldom be agreement on what 'doing what is best for the present and the future' requires in any controversial case; thus, the virtues of predictability, transparency, and treating like cases alike will erode from the law.

Feminism

Feminists influenced by the work of psychologist Carol Gilligan have argued that the ethic of impartial justice is a quintessentially male construct – one that is in no way superior to the feminine 'ethic of care.' To illustrate the ethic of care, some feminist legal scholars have fastened on the image of Shakespeare's Portia, who pleads that the 'quality of mercy' should replace rigid legal justice. The ethic of care focuses on repairing relationships and is directed toward individual needs rather than rules, policies, or the best overall consequences; unsurprisingly, this leads some proponents of the ethic of care to favor ADR over formal adjudication. On the other hand, other feminists caution that the ethic of care is part of legal justice rather than an alternative to it; conversely, care itself becomes corrupted without the virtues of consistency, integrity, and evenhandedness – precisely the virtues promoted by legal justice in its commonsense conception (West, 1997). (For further discussion on this point, *see* Feminist Legal Theory.)

Conclusion

It is often unclear whether law and economics, pragmatism, and feminism are intended as alternatives *to* legal justice or as alternative analyses of legal justice. In a sense, all the theories surveyed here are responses to two fundamental tensions. The first is the tension between the rule of law, which subordinates particular cases to general rules, and the moral demand to treat particular cases according to their individualized merits. The second is the tension between the values of consistency, predictability, and stability that legal justice serves and the more Utopian values implicit in broader notions of justice. It may well be that no intellectual approach can eliminate this tension, or reconcile the competing senses of the word 'justice.'

See also: Common Sense, Anthropology of; Equity in Law; Feminist Legal Theory; Injustice: Legal Aspects; Justice, Conflicts, and the Justice of Conflict Resolution; Justice: Philosophical Aspects; Justice: Social Psychological Perspectives; Law and Economics; Legal Formalism; Legal Realism; Legal Reasoning and Argumentation; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Nazi Law; Pragmatism: Philosophical Aspects; Social Justice and Social Inequalities.

Bibliography

Aristotle. Politics (many editions).
Aristotle. Nicomachean Ethics (many editions); 1998. Dove Publications, Mineola, NY.
Butler, P., 1995. Race-based jury nullification: black power in the criminal justice system. Yale Law Journal 105, 677–725.
Dworkin, R.M., 1977. Taking Rights Seriously, rev ed. Harvard University Press, Cambridge, MA.
Dworkin, R.M., 1980. Is wealth a value? Journal of Legal Studies 9, 191–226.
Dworkin, R.M., 1986. Law's Empire. Harvard University Press, Cambridge MA.
Finnis, J., 1980. Natural Law and Natural Rights. Clarendon Press, Oxford.
Fuller, L.L., 1969. The Morality of Law, rev ed. Yale University Press, New Haven, CT.
Gagarin, M., 1986. Early Greek Law. University of California Press, Berkeley, CA.
Gilligan, C. (Ed.), 1982. A Different Voice: Psychological Theory and Women's Development. Harvard University Press, Cambridge, MA.
Holmes, O.W., 1929. Letter to John C H Wu, 1 July 1929. In: Justice Holmes to Dr. Wu: An Intimate Correspondence 1921–32 [1947]. Central Books, New York.
Kadish, M.R., Kadish, S.H., 1973. Discretion to Disobey: A Study of Lawful Departures from Legal Rules. Stanford University Press, Stanford, CA.
Kronman, A.T., 1980. Wealth maximization as a normative principle. Journal of Legal Studies 9, 227–242.
Menkel-Meadow, C., 1985. Portia in a different voice: speculations on a women's lawyering process. Berkeley Women's Law Journal 1, 39–63.
Plato. The Laws (many editions).
Plato. The Statesman (many editions).
Posner, R.A., 1983. The Economics of Justice. Harvard University Press, Cambridge, MA.
Posner, R.A., 1998. Pragmatic adjudication. In: Dickstein, M. (Ed.), The Revival of Pragmatism: New Essays on Social Thought, Law, and Culture. Duke University Press, Durham, NC.
Rawls, J., 1971. A Theory of Justice. Belknap Press of Harvard University Press, Cambridge, MA.
Radbruch, G., 1946/2006. Statutory lawlessness and superstatutory law (B.L. Paulson, S.L. Paulson Trans.). Oxford Journal of Legal Studies 26, 1–11.
Soering v. United Kingdom 1989, 161 Eur. Ct. H.R. 87 (1989), reprinted in 28 I.L.M. 1066 (1989).
Soper, P., 1984. A Theory of Law. Harvard University Press, Cambridge, MA.
West, R., 1997. Caring for Justice. New York University Press, New York.

Justice, Conflicts, and the Justice of Conflict Resolution

Leo Montada, University of Trier, Mozartstr, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Whereas concerns for justice are universal, views of what is just and what is unjust are not universally shared. Divergent views of justice often cause social conflicts. Frequently, the parties involved in conflicts are convinced that their own view is the solely valid one. As we have no access to an objective truth about justice, conflicts may be reconciled by the judgement of an authority accepted by all parties or by a negotiated agreement between the parties. Agreements are just when the parties are equally free in their decision and equally informed about all relevant facts and possible outcomes.

Justice – A Universal Concern of Humans

Claims for justice and protests against injustice are ubiquitous in social life. Political movements, revolutions, and wars are inflamed by experienced or perceived injustices. Justice is debated in all fields of politics. The courts are swamped with lawsuits aimed to enforce supposedly just entitlements. Resentment of injustice is at the core of conflicts in all contexts of life from the domestic to the public domain. It is hard to cope with a fate that is perceived as unjust.

The scope of concern for justice becomes obvious by the spectrum of subjects that are evaluated as just or unjust: *distributions* of wealth and poverty, opportunities and risks, rights and duties, etc.; *social interactions and relations* where, for example, goods and services, love and loyalty, or hostilities and retributions are exchanged; *retributions for offenses*; *valuations of achievements and failures*; *constitutions and laws of states, the bans and rules of religions*; *social roles, rules, and conventions in cultures and communities*; and *procedures in elections and in decision making*.

The Justice Motive

Appraisals of justice emotionalize and have a motivational impetus. Humans have a justice motive; they are averse to injustice (Ross and Miller, 2002). This assumption does not fit to the conception of man as homo oeconomicus implying the assumption that self-interest is the cardinal human motive. The following examples lay out some of the differences between these two conceptions of man.

No doubt, the 'flag of justice' often serves to mask self-interest (Montada, 2003). Politicians may promise to fight against some injustice hoping to thereby raise their own popularity. Employers may care about the productivity of their employees, which declines when these feel treated unjustly. However, a merely strategic concern for justice proves that the actors believe that justice is an important concern of other people, which they try to abuse in a 'parasitic' way for their selfish goals (Elster, 1989).

Rich empirical evidence is available for the fact that people care about justice also when this does not serve their self-interest. Resentment is not only evidenced in cases of self-experienced injustices, but also in cases of injustice suffered by foreign people all over the world, e.g., by victims of natural disasters, poverty, violence, negligence, or despotism. Feelings of existential guilt were observed with people living in fortunate conditions when they became aware of the bad fate, the harm, and hardships suffered by foreign people. Both resentment and existential guilt stimulate prosocial commitments aiming at reducing these injustices (Montada, 2003).

Using experimental approaches, such as the Ultimatum or Public Goods or Resource Dilemma games, innumerable studies in various cultures have shown that a majority of people surrender personal advantages and take considerable costs only to prevent or to retaliate unjust selfish advantage taking of other actors. These findings can only be understood by assuming justice motives (Fehr and Gächter, 2002; Fehr and Schwarz, 2003). The justice motive cannot be reduced to self-interest, it is a primordial motive.

Divergent Views about Justice Cause Social Conflicts

While the justice motive is universal, the views of justice are not at all universally shared. Whatever the target of valuation may be – distributions, social exchanges, evaluations of deeds and misdeeds, achievements and failures, policies, laws, and other normative standards, procedures in elections and decision making – we find divergent views. The convictions vary between cultures, collectives, social milieus, and individuals when specific cases are appraised and also when general maxims of justice are at stake, e.g., the various principles for just distributions (Montada, 2012).

Conflicts may be carried out constructively, e.g., by rational discourses about diverging convictions and positions. Quite often they are carried out destructively. By far not every action powered by justice motives is commendable. Under the flag of justice, dire wrong is done. Frequently, actors 'justify' their violence, crimes, selfishness as retaliation of injustice or as prevention of expected injustice (Montada, 2007). Since Plato, justice was valued as one of the cardinal virtues in the antique and medieval philosophy. The justice motive, however, cannot generally be valued as a virtue. To be a virtue it needs cultivation, namely, the readiness and the competences to reflect on the validity of one's own convictions and those of others.

We are speaking of justice in the singular – *the* justice – as if one single view of justice would be the valid one in general or

in a concrete case and had to be acknowledged and shared by everybody else (Rüthers, 1991). And we tend to prefer mates, friends, and peers who share our own convictions. Many people have normative convictions that they believe to be 'indisputably true' or even 'sacred' and 'not negotiable.' They may be shared by more or less large collectives, they are, however, not shared universally.

The belief of justice as universal truth is to be found in religions – most of the bans and rules of religions can be understood as norms of justice. However, different religions hold different bans and rules. It is also to be found in political philosophies. These, too, hold different maxims of justice, e.g., egalitarianism, liberalism, social welfarism, or utilitarianism. No one is universally acknowledged. Consequently, the question of what is just and what is unjust is debatable and may cause social conflicts. Within states, communities, and further social units, conflicts are less frequent when views about justice are widely shared and when the justice of institutions as parliaments and courts is widely acknowledged within the population (Würtenberger, 1984). In pluralistic populations, divergent views of justice and divergent attitudes toward authorities and institutions are the rule.

The Spectrum of Justice Principles in Various Subject Fields

Equality is widely considered as the basic principle of justice (Gosepath, 2004). What does equality precisely mean? Equal freedom for all, equal opportunities or equal outcomes (cf Barry, 1989; Kolm, 1996; Schmidt, 2000; Sen, 2010)? Human beings are different in many respects. Is everybody to be treated equally, does everybody have equal rights and duties or are only equals to be treated equally and have equal rights and duties as Aristotle has postulated? Many debates about justice start with the question which inequalities are relevant and to be considered in the evaluation of justice and injustice.

Distributive Justice

For instance, which inequalities can be seen as relevant concerning the *just distribution* of goods, burdens, rights, duties, etc.; gender, age, social status, kinship, nationality, merits, skills and expertise, needs? Should various relevant inequalities be accounted for in combination? How relevant are responsibilities for needs, losses, damages, injuries, diseases, etc., with respect to distributive justice? These questions are debated by philosophers, in political arenas and in many public and private contexts.

Walzer (1983) proposed the concept of *complex equality*, meaning that distributions in different spheres of justice (e.g., material wealth, social recognition, political power, education, kinship and love, hobbies, etc.) are not perfectly correlated. Thus, a lower rank in one sphere may be compensated by a higher rank in another. Taking into account that the subjective importance of these spheres varies between people, the subjectively experienced and perceived overall inequality in a population may be much lesser than 'egalitarians' presume and criticize as unjust.

Justice in Social Exchange Relationships

The equality criterion for appraising justice can also be found in some modified forms in other domains of social and societal life. In exchange relationships, reciprocity and equity are prominent principles. Social exchanges have been regarded to be just if reciprocity is established (Gouldner, 1960). This is true for positive exchanges, where equal mutual advantage is the normative standard, as well as for negative exchanges, which are often described by the metaphor of 'an eye for an eye and a tooth for a tooth.'

Same as in the domain of distributive justice, the actors in exchange relationships may not be equal, for instance, with regard to their needs, abilities, wealth, burdens, social roles, opportunities, etc. These and further inequalities have to be considered when the justice of exchanges is evaluated. In many relationships, equal mutual advantages are neither possible nor expected. Nevertheless, the exchange relation may be evaluated as just, especially when the *equity principle* is applied or when social role relationships are assessed.

Equity means equal ratios of investments/costs and outcomes/benefits of all parties involved in exchange relationships. Actually, the subjective appraisals of these ratios, not the objective balances are crucial (cf Tyler et al., 1997). Subjectively, different exchanges (goods, services, love, respect, trust, loyalty, harm, negligence, hostilities, etc.) may be focused on and balanced. Indirect and chained exchanges may be included in the subjective balances. In many close relationships, subjectively all provided benefits and services may be experienced as compensated by gratitude or by love. In social role relations, too, equity does not mean equality of reciprocal exchanges. For example, teachers will be motivated to give their best and feel justly gratified by their pupils' progress. The kind of relationship is crucial. On the marketplace, the exchange relations are not the same as in employment contracts, in families or in friendships. Consequently, the justice standards are not the same.

Retributive Justice

The preliminary question is: What is considered an offense, a crime, a sin? The answers vary a lot in national or religious codes and in the normative standards of cultures and communities. Consequently, the answers are debatable in terms of justice. Further questions are: What should be done in case these norms are violated? Are the offenders to be punished or to be educated? Who is to blame for the offense: the offenders, their families, or the social system? How is the punishment of an offense to justify: as retaliation, as deterrence, as rehabilitation? These questions, too, are debatable in terms of justice.

What are just assessments of offenses (Darley, 2002)? The main principle of *retributive justice* is specified as proportionality of guilt for misdeeds and their punishment. Again, inequalities between the actors and the specifics of the action context are to be considered. The guilt of offenders may be qualified by their life history, their living contexts and by situational specifics, which may qualify their responsibility, e.g., provocations by the victims (Auhagen and Bierhoff, 2001). Impartiality of verdicts means that the same standards are applied to similar offenders and similar offenses.

The focus of *restorative justice* is not retribution of the offense but support for the victims and reintegration of the offenders into society (Rössner, 2004). This approach frequently brings offenders and victims together. The offenders shall become aware of the effects that their offense had on the victims. They get an opportunity to apologize and to make amends. In cases of an interactions history between offenders and victims as in cases of mobbing, the victims may learn whether and by which behavior they may have provoked the offenders.

Procedural Justice, Just Decision Making

In view of the broad spectrum of options for appraising and constructing justice in various fields of social life, conflicts about justice can be expected. Depending on the kind of conflicts and the social system, various ways for their settlement are available: lawsuits, arbitration, negotiation, mediation, decision making in parliaments, organizations, and communities. In all these contexts, procedural justice gains eminent importance (Ackermann, 1980). A prerequisite of accepting a decision as just is the perceived legitimacy of decision-making authorities or bodies.

Moreover, those who are affected by decisions value the impartiality and the objectivity of authorities and decision making bodies as criteria for justice (Leventhal, 1980). Above all, they want to have 'voice,' meaning the opportunity to present their views and claims in the process of decision making (Thibaut and Walker, 1975). Well confirmed by empirical research is the fair-procedure effect (cf Tyler et al., 1997), meaning that perceived procedural fairness helps the parties to accept decisions or outcomes that are less favorable than they had expected and hoped for. Decent and polite treatment by authorities may add to this effect (Lind and Tyler, 1988).

The Justice of Constitutions and Codified Laws and Ethics

Justice convictions are also crucial in appraisals of constitutions, legal codes, religious bans, and rules. These and further bodies of normative maxims and rules exhibit large divergencies that give reason for disputing about their justice. Every law, every normative maxim can be criticized on the basis of some justice principle. Legal codes may violate human rights or the bans and rules of some religion. The charter of human rights contravenes the normative traditions of various cultures. For instance, it is open to question whether the dignity of human beings may be determined universally or whether it may be specified according to culturally shared values or according to individual values formulated in every social interaction up to living wills? Different cultures have diverging conceptions of justice, rights and duties, standards for respectful behavior, honor, etc. Cultures are definable by their specific morals and manners – codified or not.

Efforts to Remove Injustice May Create New Injustices

It is easier to criticize some injustice than to establish justice within social and societal systems. Efforts to correct one injustice may result in new injustices as illustrated by a few examples. (1) Affirmative action policies are justified as just compensation for the unquestionable historical disadvantages of women in the labor market. Surely, there are good reasons for removing these disadvantages. However, it is disputable whether it is just to privilege young women who have personally not been disadvantaged so far in competition with young men who have personally not been privileged. (2) Justice for the defendants may mean injustice for the victims of crimes. When the 'benefit of doubt' is granted to the defendant – doubtlessly a significant progress in the history of criminal law – the victim's claim for justice may be violated, assuming that the victim has no doubt that the defendant is guilty (Orth, 2000). (3) Sentences may be deserved and just retributions for offenses, but they bear the risk of unjust social discrimination of people close to the offenders, for instance their children. Looking at the branched systemic effects of every intervention, it is not easy to avoid new injustices. Efforts to establish justice have to be diligently checked for the risks of new injustices.

Injustice and Conflicts

At the core, all social conflicts are justice conflicts. People expect that others respect the normative standards, entitlements, and claims that they believe to be just. Conflicts result from violated normative expectations on how other people, agencies, etc., have to behave, from violated or disregarded claims, whatever their normative basis may be: human rights, a legal code, the codex of a religion, a formal or an implicit contract, or good manners that are valid in a culture. Crucial is the fact that actors consider their own convictions as legitimate and as rightfully to be claimed.

Perceived injustices, not incompatibilities instigate social conflicts. Incompatible goals, plans, opinions, etc., may bear the risk of conflict, but only in case that actors violate the normative expectations of others. Diverging views may be appreciated as worthy of consideration or else as disrespectful, as withdrawal from the community. Diverging goals may be considered as legitimate or as an offense. Diverging religious beliefs may be tolerated as a matter of personal freedom or condemned as a sin or a sacrilege. Fair competition is not a social conflict. If all actors in competitive markets and sports are considered to behave legitimately, frustrations and losses are possible but they do not evoke resentment. There are winners and losers but not victimizers and victims. The losers have nothing to reproach the winners for.

Perceived injustice instigates resentment, reproaches, wishes for retaliation, or claims for compensation. In conflicts, people claim justice for themselves or for others for whom they feel sympathy or responsibility. They react with resentment when they view the perpetrating actors as responsible, and when they do not see a convincing justification for the norm violation. Conflicts will become manifest when the actors are blamed and do neither confess nor apologize for their misdemeanor nor excuse or justify their behavior convincingly. In manifest conflicts, the parties justify their own behavior and reproach the behavior of the others.

Conflicts and Justice

Looking at all the divergent views of justice, we have to conclude that conflicts are unavoidable. When conflicts are not reconciled, but carried out destructively, they may be costly for one or for all parties. They may be a mortgage for future exchanges and end in lasting hostility. Conflicts may also be an opportunity to recognize and remove injustices, to establish a sustainable peace, and to create productive social relationships. Using the right procedures, conflicts may trigger productive developments in social and societal life.

Justice by the Settlement of Conflicts

If the assumption is valid that social conflicts basically result from incompatible normative expectations or claims, making conflicting convictions converge is the self-evident strategy for settling conflicts. This point can be illustrated by the fact that sincere apologies by an offender have a pacifying effect. As Goffman (1971) has emphasized, by apologies the perpetrators express that they fully share the victims' normative view of the case. They affirm the violated norm as valid, they admit having offended the norm, to be responsible for their action or neglect, and to be blameworthy because their offense was not justified. Sincere apologies reconcile victims (as well as judges and observers) and reduce their resentment and desire for retribution (Ohbuchi et al., 1989). Thereby, a complete consent about the origin of the conflict is expressed, and the conflict is settled.

Convergence of normative convictions can be attempted by *normative discourses*. Discourses must not have the objective to find 'the' normative truth. In the mediation of conflicts they merely serve to convey the insight that good reasons can be put forward not only for one's own views and claims but also for the opponent's views and claims, and also for further stances that could be taken in the case. Normative dilemmata can be disclosed. In discourses, the conflict parties get motivated to seek not only arguments in favor of their own position but to deliberate also on counterarguments and arguments in favor of the position held by the opponents (Montada, 2012).

Settling conflicts is made easier when the opponents acknowledge that divergent norms or principles of justice may be advocated with good reasons, and, consequently, that neither party is solely right or wrong. If the opponents recognize that their conflict reflects a normative dilemma, they no longer view the position of the other side as completely illegitimate and their own position as the solely legitimate one. This is the reason why sacred, not negotiable normative positions are a big problem for the mediation of conflicts.

For example: Which employees should be dismissed first when business is running low? Many criteria may be deemed just: seniority, acquired merits, abilities, current performance level, need of the job for the support of a family age, wage level, etc.? Or should nobody be dismissed and the wages of all employees cut back? For each of these criteria plausible arguments and counterarguments may be pleaded.

In normative discourses, the conflicting positions or claims for justice are both supported and criticized. When good arguments are put forward, the opponents may qualify their views and claims. They recognize that for none of the conflicting claims exclusive validity can be claimed. It is the wisdom of institutions to consider and integrate various principles of justice in their regulations and decisions (Elster, 1992). The social market economy, for instance, is an attempt to harmonize the right of all citizens to free economic activities with the maxims of the social welfare state. Rawls' 'Maximin principle' is also a suggestion to combine the freedom to economic activities – which produces common wealth – with the entitlement of every citizen to participate in the common prosperity (Rawls, 1971).

Settlement of Conflicts by Just Contracts

Only a just peace is a sustainable peace – an unjust settlement of a conflict would cause new conflicts. However, not 'objective justice' is required for peace, but 'intersubjective justice,' meaning that all parties consider the settlement of their conflict to be just. Sentences and decisions by authorities carry the risk of producing winners and losers. The losers may believe to be treated unjustly. Settling conflicts by an agreement, by a contract, is the preferable option.

Just and Unjust Contracts

Contracts are regarded to be just when the parties are equally free to consent or not to consent and when the parties are equally well informed about all relevant facts and risks. This does not mean objective justice but intersubjective justice. A contract that the parties have agreed upon freely and well informed cannot be unjust to one of the parties. Justice would be at risk if relevant information had been withheld from one of the contracting parties, if pressure had been exerted, if one party had not been free to refuse the contract on account of some predicament, e.g., neediness. Therefore, unequal power, unequal neediness, unequal abilities to grasp complex information may jeopardize the justice of agreements.

Because contracts are of eminent importance for peace in social life, many legal norms have been established which specify the obligations of the contracting parties. In many states, legal rules have been established to protect the supposedly less powerful or less informed parties in business contracts. These risks can be reduced if the agreements are negotiated under the guidance of mediators who have the duty to actively balance such inequalities. Mediators also have to take care of the justice of contracts with regard to affected third parties.

Since Hobbes 1648/1970, equal freedom and equal information have also been the core idea in *contract theories of the state* (cf Kersting, 1994): Free citizens agree upon the constitution and the legal code of a state. This agreement constitutes the justice of the constitution and the legal code. According to this model, majority decisions in parliaments bear the risk that the losing minority and many citizens consider it as an unjust imposition.

In the settlement of conflicts, the opponents are considered autonomous to negotiate an agreement. However, appraising the justice of contracts would be incomplete without having a look at their *impact on third parties*. Adverse effects on third

parties would raise new justice problems. Contracts may be fair with respect to the exchanges between the contracting parties but may be seriously unfair with respect to third parties or a community as, for instance, cartel agreements that may be fair for the contracting parties at the expense of others. Therefore, it is prudent to expand the view from the contracting parties to the impact of an agreement on third parties and the public. Contracts at the cost of third parties are externally unjust and a cause for new conflicts. In conflict mediation, it is a wise rule to think of all third parties who may be affected by an agreement and to reflect their legitimate concerns.

Just contracts are binding, they are to be observed. This is the normative premise for making a contract. Typically, contracts are concluded on the base of assumptions about the future. Sometimes, these outcomes turn out to be wrong. If one of the parties has deceived the other one by false information the contract may be contested. What about erroneous assumptions that do not result from a deception? The parties to a contract are well advised to make explicit their shared assumptions as a base of the contract and to include the clause to adjust the contract in case that these turn out to be wrong. To ensure that contracts are just in the long run it is opportune to plan an evaluation of the feasibility and productivity of a contract and include the option to adjust and optimize the agreement in view of new facts, new insights, and new views of justice. *Adjustable contracts* are preferable to strictly and for all times specified ones because they allow to avoid injustices that have not been foreseen. This is also true for the laws and the institutions of a state, which have to be adjusted to new facts, to changed needs and preferences within the population, to changed views of justice, etc.

In the mediation of many *conflicts* (Deutsch, 2006; Montada and Kals, 2013), adjustable contracts are recommendable. The opponents should be made aware of the fact that their agreement is based on assumptions and expectations, which may turn out to be wrong or incomplete, and of the fact that they may have not considered all relevant aspects and questions and not all options for reconciling their conflict by an agreement. Therefore, the option may be considered to evaluate the current agreement after an appropriate period and to adjust it if this would be opportune with the same sense of fairness and the same climate of cooperation established in the actual mediation.

Summary

Humans universally care about justice, but their views of what is just are divergent. People tend to view their own convictions as the true or valid ones. We neither have cognition of an objective truth about justice nor authorities whose opinions would be acknowledged universally. Divergent convictions are causes for conflicts. Basically, all social conflicts are justice conflicts triggered by violations of subjectively just claims and expectations of how others have to behave.

If the parties in a conflict have trust in the justice of an authority or an institution they may appeal to them for settling the case. Conflicts may also be settled by an agreement or contract. An agreement is just between the parties when they have equal freedom of decision and equal information about relevant facts and expected outcomes. Whether an agreement is just or unjust with respect to affected third parties has to be examined case by case. Whether the agreement will come up to the expectations of the parties cannot be taken for sure. Including an evaluation of the agreement after an adequate period of experience offers the option to adjust it consensually.

See also: Business Ethics; Economics and Ethics; Equality: Philosophical Aspects; Ethics and Values; Female Quotas in Parliament: Do They Make a Difference?; Intercultural Legal Discourse: Philosophical Aspects; Justice: Philosophical Aspects; Morality 'East' and 'West': Cultural Concerns; Pluralism and Tolerance; Rational Choice Explanation: Philosophical Aspects; Religious Nationalism and the Secular State; Value Pluralism; Virtue Ethics; Wisdom: Philosophical Aspects.

Bibliography

Ackermann, B., 1980. Justice in the Liberal State. Yale University Press, New Haven.
Auhagen, A.E., Bierhoff, H.W. (Eds.), 2001. Responsibility – The Many Faces of a Social Phenomenon. Routledge, London.
Barry, B., 1989. Theories of Justice. University of California Press, Berkeley.
Darley, J., 2002. Just punishments: research on retributional justice. In: Ross, M., Miller, D.T. (Eds.), The Justice Motive in Everyday Life. Cambridge University Press, New York, pp. 314–333.
Deutsch, M., 2006. Justice and conflict. In: Deutsch, M., Coleman, P.T., Marcus, E.C. (Eds.), The Handbook of Conflict Resolution. Jossey – Bass, San Francisco, pp. 43–68.
Elster, J., 1989. The Cement of Society. Cambridge University Press, Cambridge.
Elster, J., 1992. Local Justice. Russel Sage Foundation, New York.
Fehr, E., Gächter, G., 2002. Altruistic punishments in humans. Nature 415, 137–140.
Fehr, E., Schwarz, G. (Eds.), 2003. Psychologische Grundlagen der Ökonomie: Über Vernunft und Eigennutz hinaus. Verlag Neue Züricher Zeitung, Zürich.
Gerhard, C., 2007. Gemeinwohl vs. Eigennutz. VDM Verlag Dr. Müller, Saarbrücken.
Goffman, E., 1971. Relations in Public: Microstudies of the Public Order. Penguin, Harmondsworth.
Gosepath, S., 2004. Gleiche Gerechtigkeit. Grundlagen eines liberalen Egalitarismus. Suhrkamp, Frankfurt/M.
Gouldner, A., 1960. The norm of reciprocity. American Sociological Review 25, 161–178.
Hobbes, T., 1648/1970. Leviathan. Penguin, Harmondsworth.
Kersting, W., 1994. Die politische Philosophie des Gesellschaftsvertrags. WBG, Darmstadt.
Kolm, S.C., 1996. Modern Theories of Justice. MIT Press, Cambridge, MA.
Leventhal, G.S., 1980. What should be done with equity theory? New approaches to the study of fairness in social relationships. In: Gergen, K.J., Greenberg, M.S., Willis, R.H. (Eds.), Social Exchange: Advances in Theory and Research. Plenum Press, New York.
Lind, E.A., Tyler, T.R., 1988. The Social Psychology of Procedural Justice. Plenum Press, New York.
Montada, L., 2003. Justice, equity, and fairness in human relations. In: Weiner, J. (Ed.), Handbook of Psychology, vol. 5. Wiley, Hoboken, NJ, pp. 537–568.
Montada, L., 2007. Justice conflicts and the justice of conflict resolution. In: Törnblom, K., Vermunt, R. (Eds.), Distributive and Procedural Justice. Research and Applications. Ashgate/Glower, Burlington, pp. 255–268.
Montada, L., 2012. The normative impact of empirical justice research. In: Kals, E., Maes, J. (Eds.), Justice and Conflicts. Springer, Berlin/Heidelberg, pp. 3–19.
Montada, L., Kals, E., 2013. Mediation: Psychologische Grundlagen und Perspektiven, third ed. Beltz, Weinheim.
Nozick, R., 1974. Anarchy, State and Utopia. Basic Books, New York.
Ohbuchi, K., Agarie, N., Kameda, M., 1989. Apology as aggression control. Journal of Personality and Social Psychology 56, 219–227.

Orth, U., 2000. Strafgerechtigkeit und Bewältigung krimineller Viktimisierung. Weisser Ring Verlag, Mainz.
Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge.
Rössner, D., 2004. Mediation im Strafrecht. In: Maelicke, B., Sonnen, B.R. (Eds.), Handbuch der Resozialisierung. Nomos, Baden-Baden, pp. 203–231.
Ross, M., Miller, D.T., 2002. The Justice Motive in Everyday Life. Cambridge University Press, New York.
Rüthers, B., 1991. Das Ungerechte an der Gerechtigkeit. Defizite eines Begriffs. Edition Interfrom, Zürich.
Schmidt, V., 2000. Bedingte Gerechtigkeit. Soziologische Analysen und philosophische Theorien. Campus, Frankfurt/M.
Sen, A., 2009. The Idea of Justice. Penguin Books, London.
Sen, A., 2010. Die Idee der Gerechtigkeit. C.H. Beck, München.
Tyler, T.R., Boeckmann, R.J., Smith, H.J., Huo, Y.J., 1997. Social Justice in a Diverse Society. Westview, Boulder.
Thibaut, J., Walker, L., 1975. Procedural Justice: A Psychological Analysis. Wiley, New York.
Walzer, M., 1983. Spheres of Justice: A Defense of Pluralism and Equality. Basic Books, New York.
Würtenberger, T., 1984. Legitimität und Gesetz. In: Rüthers, B., Stern, K. (Eds.), Freiheit und Verantwortung im Verfassungsstaat. C.H. Beck, München, pp. 533–550.

Justice: Philosophical Aspects

Thomas Pogge, Yale University, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Justice is predicated mainly of three kinds of *iudicanda*: *subjects* (persons and collectives), the *conduct* of subjects, and *social rules*. Recent discussions have focused on *social justice*, concerned with social rules. Here John Rawls's seminal theory has biased the discussion toward purely recipient-oriented theorizing, considering solely how human beings fare under various social rules. An alternative approach would also consider how we are treating one another, making morally significant our specific roles in imposing social rules and the causal pathways on which such rules affect people.

Introduction

Justice is a central moral notion, associated with fair and impartial decision procedures (*procedural justice*) and with persons and groups being treated even-handedly (*formal justice*, treating like cases alike) and in a morally fitting way (*material* or *substantive justice*). The demands of justice are frequently presented as uncompromising and strictly required, in contrast to the more flexible – and sometimes opposing – demands of mercy, charity, humanity, decency, beneficence, kindness, and equity, which may be merely supererogatory.

Justice is predicated mainly of three kinds of *iudicanda*, namely, *subjects* (persons and collectives), the *conduct* of subjects (actions and omissions, such as accusations, refusals, policies, and wars), and *social rules* (including social institutions, laws, norms, practices and procedures, economic regulations, and educational and health care systems). In recent decades, discussion has focused on *social justice*, which is concerned with the third *iudicandum* of social rules.

Justice is often differentiated by theme. Thus *distributive* justice governs the assignment of benefits and burdens (such as rights, duties, property, income, and taxes), *commutative* justice governs trades and exchanges, and *corrective* and *retributive* justice govern, respectively, compensating the victims and punishing the perpetrators of prior wrongdoing. *International* justice is typically viewed as a separate area, in which the four mentioned themes are overshadowed by reflections on the just use of force. Just war theories are generally organized around two topics: just cause in going to war (*ius ad bellum*) and just ways of fighting a war (*ius in bello*). Kant adds a third topic: the just way of concluding a war so as to lay the foundation for a stable peace (*ius post bellum*, Kant, 1996, §58).

The manifest complexity of the notion of justice makes it doubtful that any unified account of justice can be adequate across *iudicanda* and themes. In response to such doubts, it is tempting, as numerous historical examples attest, to simplify matters through reductionist moves. One such move, prominent among deontological conceptions of ethics, is to reduce judgments of character to judgments of conduct, perhaps by holding that a just subject is one stably disposed toward just conduct, as independently specified. (One may object that the justice of subjects is in part a function also of their inner states: of the feelings and attitudes of individuals and the internal mode of organization of organized collectives. Thus, the city of Athens may be called unjust not merely on account of its conduct toward Socrates and the Melians, but also on account of the severe disadvantages its laws inflicted upon women, metics, and slaves.) Following Aristotle (1985), some conceptions of ethics have proposed the reverse: to reduce judgments of conduct to judgments of character. Just conduct is, then, conduct that a just agent (independently specified) would engage in. Locke (1960) and Nozick (1974) have proposed a third kind of reduction: social rules are just if and only if they were, or at least could have been, established through just conduct. Conversely, Rawls has suggested that persons and conduct are just when they support and obey applicable just social rules or help establish just social rules insofar as such are lacking (Rawls, 1971, 115, 334).

Other reductionist attempts seek to reduce judgments about various kinds of *iudicanda* to judgments about states of affairs. One consequentialist reduction of this sort holds that subjects, conduct, and social rules are just in proportion to their tendency to promote the overall goodness of the world or whatever part thereof they are specially responsible for. To be discussed in this article, John Rawls offers a sophisticated such reduction of judgments about social institutions to judgments about distributions.

The plausibility of such reductions depends on the overall plausibility of the theory of justice in which they are embedded, which in turn depends on whether the evaluative and normative judgments derivable from this theory are morally sound. So far no detailed theory of justice, reductionist or otherwise, has managed to cover several themes and kinds of *iudicanda* in a plausible way. Contemporary writers have therefore sought to develop more limited conceptions of justice. The paradigmatic effort in this vein, that of Rawls (1971), is focused narrowly on how distributive justice might be instantiated in the most important and pervasive rules, constituting the *basic structure*, of a self-contained society.

The Theory of John Rawls

A criterion of justice should entail plausible absolute and comparative judgments about feasible basic structure designs, actual or proposed. Rawls seeks a criterion that, in addition, facilitates real-world agreement among the citizens accepting it on how to maintain or reform their shared social institutions.

This second desideratum speaks against any criterion that, like the principle of utility, is too difficult to apply and hence too easily abused. Rawls seeks the morally best *public* criterion of social justice.

His central idea is to answer this moral question through a prudential one: morally best is that public criterion of justice that prospective citizens would choose on prudential grounds. The criterion and any basic structure designs based on it should be as good as possible for the persons living under them; nothing else need be considered. This central thesis is, of course, itself a moral claim – to be developed and examined in what follows.

Citizen's interests and preferences vary and are shaped by the actual basic structure design of their society. For his prudential question to have a determinate answer, Rawls therefore needs a vantage point that attains some independence from the status quo and also ensures agreement. This vantage point is the *original position* – an imagined deliberation among *parties* who, acting in the interest of prospective citizens, are to agree on a public criterion of justice without knowing, due to a *veil of ignorance*, the distinguishing features (endowments, tastes, needs, desires, creeds, and values) of their clients or even the natural and historical context of their clients' society.

To have grounds for deliberation, the parties need some information about their clients' interests. Rawls makes them assume that all clients have three interests, which he portrays as closely connected to their role as citizens in a democratic society and hence as not biased toward any particular religious, philosophical, or ethical worldview or way of life. Rawls calls these the *three higher-order interests*, suggesting that they are interests in the content and fulfillment of other interests (like second-order desires are desires about desires) and also are deep, stable, and normally decisive. These interests are (1) to develop and exercise a sense of justice, which involves a sincere allegiance to some conception of justice (whose content remains unspecified so as not to prejudge the parties' decision); (2) to develop and exercise the capacity to form, to pursue, and rationally to revise a conception of the good; and (3) to protect and advance one's particular conception of the good (Rawls, 1993, 19, 74). The main question in the original position is, then, which *social order* – consisting of a public criterion of justice and of the basic structure designs this criterion would actually bring forth in various alternative contexts – affords the best prospects for citizens' higher-order interest fulfillment?

Rawls makes the dramatic claim that the parties would, because of their severely limited information, make their choice pursuant to the *maximin rule*, seeking to optimize the worst social position that the public acceptance of their criterion might generate. They would examine candidate criteria of justice, asking of each how badly off, in terms of the three higher-order interests, persons might come to be in societies in which it is the accepted public criterion of justice. And they would prefer the criterion that promises the higher floor (i.e., the *max*imum *min*imorum), even if it would also lead to a much lower ceiling and average.

Rawls argues that the parties would favor a criterion of justice formulated in terms of certain *social primary goods*, such as rights, liberties, income, and wealth. Access to these all-purpose means is not only a good proxy for higher-order interest fulfillment, but also much easier to measure than the latter. Measurability is important because the criterion is supposed to be used for public justification: to foster agreement on institutional design decision. Rawls claims that the parties would choose these two principles as the public criterion of social justice (Rawls, 1993, 5f).

> *First Principle:* Each person has an equal claim to a fully adequate scheme of equal basic rights and liberties, which scheme is compatible with the same scheme for all; and in this scheme the equal political liberties, and only those liberties, are to be guaranteed their fair value.
>
> *Second Principle:* Social and economic inequalities are to satisfy two conditions: first, they are to be attached to positions and offices open to all under conditions of fair equality of opportunity [the opportunity principle]; and second, they are to be to the greatest benefit of the least advantaged members of society [the difference principle].

The first principle is to have lexical priority over the second, and, within the second, the opportunity principle is to have lexical priority over the difference principle (Rawls, 1971, 302f). 'Lexical' is short for 'lexicographical' and thus invokes how words are arranged alphabetically, giving priority to the first letter over the second, to the second over the third, and so on. Rawls conceives priorities among principles accordingly: in the comparison of alternative feasible basic structure designs, superiority in terms of a lexically prior principle overrides any inferiority in terms of lexically subordinate ones.

The proposed public criterion of justice requires, then, that a society's basic structure be arranged to secure above all the basic liberties. These include the political liberties, the rights and freedoms included in the rule of law, freedom of thought and liberty of conscience, freedom of association, and the freedom and integrity of the person including freedom of movement and the freedom to hold personal property. Except in the case of the political liberties, the first principle does not require that citizens have sufficient means, such as income and wealth, to take full advantage of their basic liberties; the distribution of such means is governed by the second principle. Yet, the first principle does require more than legal recognition: citizens must have secure access to their freedoms, and attempts to disturb their enjoyment or exercise must be effectively prevented and deterred.

The second principle addresses the social and economic aspects of a basic structure. The opportunity principle requires that persons have roughly equal chances of success irrespective of their initial (childhood) socioeconomic position. The difference principle requires that rules affecting socioeconomic inequalities (such as the tax system and the social safety net) be designed to optimize the lowest socioeconomic position. Existing poverty is justified if, and only if, no institutional reform decreasing or increasing inequality would raise the lowest socioeconomic position.

The Bias toward Purely Recipient-Oriented Theorizing

While Rawls's theory has received much criticism, it has also in the last 40 years shaped the debate about justice and, most

importantly, the dominant concept of justice. There is little awareness now of the distinctive features of, or alternatives to, the specific concept of justice employed by Rawls and his many critics and defenders.

While the word 'just' can, and often does, function as a one-place predicate, the concept of justice, like the concept of mother, essentially involves a second place: to be unjust, a *iudicandum* A must be unjust *to someone*: to certain *recipients* B.

The idea of recipients, of those who have justice claims on a particular *iudicandum*, invokes the question of what their claims on the *iudicandum* are, the question of what these recipients have and do not have, and what they ought to have or not to have. The concept of justice essentially involves, then, also a third place for relevant *goods and ills*, broadly conceived: A is just or unjust to B by affecting whether and to what extent B has C. (As the category of conduct includes omissions, so the category of goods and ills includes the null case of receiving nothing, being denied a good or spared an ill.) For B to suffer an injustice, there must be goods or ills C that B either has-but-should-not-have or lacks-but-should-have.

The concept of justice as explicated thus far – fundamentally a three-place concept, with frequent but derivative one-place and two-place employments – frames current discussions of justice. According to it, justice is a feature of certain *iudicanda*, which depends on their effects on the distribution of certain goods and ills to or among their recipients. One might call this the *passive* concept of justice, because it directs moral attention to those by whom (in)justice is suffered. Employing this passive concept, one will focus on a *iudicandum*'s effects on its recipients – specifically, on its impact on what these recipients are able to have or to be relative to what they ought to be able to have or to be. One will consider a *iudicandum* just in proportion to its impact on the moral quality of the distribution of relevant goods and ills among its recipients: to be just is to promote good distributions. The passive concept thus leads easily to a conception of justice as *distributional*, and much of the current justice literature is indeed about how to code and rank distributions.

This framing of the debate may well be a remnant of long-dominant consequentialist, especially utilitarian, theorizing (Bentham, 1982; Sidgwick, 1981), which Rawls has superficially done so much to displace. The central consequentialist idea is that subjects, conduct, or social rules should be assessed solely by their effects on how well the world goes. Most current theorizing about justice involves the closely related thought that the justice of subjects, conduct, or social rules should be assessed by their effects on the moral quality of the overall distribution of goods and ills among the *iudicandum*'s recipients. In both cases, the interest in outcomes is indirect. How well the world goes is of little practical interest in itself, but can be of great practical interest if one judges the moral quality of other *iudicanda* by their impact on how well the world goes. Similarly, the moral quality of distributions is of little practical interest in itself, but can be of great practical interest if one judges the justice of other *iudicanda* by their impact on the moral quality of distributions.

Does it make sense to take the assessment of distributions as basic and then to judge other *iudicanda* by their distributional effects? With respect to subjects and their conduct, this mode of assessment is widely rejected. Among the timeworn staples of ethics classes are gruesome organ redistribution stories designed to elicit the response that it is often unjust for subjects and conduct to aim at or produce the best overall distributions. Abstractly considered, a situation in which everyone has at least one eye and one kidney is surely morally better than an otherwise similar situation in which some, through no fault of their own, have no functioning eye or kidney while many others have two. But redistributive conduct and subjects aiming at such an abstractly better distribution are nonetheless considered gravely unjust.

With respect to subjects and their conduct, cases of this sort may be used to draw the conclusion that we ought to distinguish between *treating recipients justly* and *promoting a good distribution among them*. The moral significance of this distinction can be illustrated with matched doing/allowing scenarios such as the following. In Scenario One, Art has two functioning kidneys and Sue has none; we must decide whether to give one of Art's kidneys to Sue. In Scenario Two, Art has one functioning kidney and Sue has none; we must decide whether to give a donated kidney to Art or to Sue. (Other things are presumed equal across scenarios, with both Art and Sue urgently wanting the kidney in question.) In terms of their distributional impact, the two decisions are on a par: in each scenario, we decide whether the final distribution of kidneys will be 2:0 or 1:1. A decision in Sue's favor is then in both scenarios equally promoting of a good distribution. Intuitively, however, the scenarios differ dramatically: justice requires a decision for Art in Scenario One and a decision for Sue in Scenario Two.

With respect to social rules, a similar distinction seems called for, and for similar reasons. Just social rules for allocating donated kidneys favor those who, through no fault of their own, have no functioning kidney over those who have one, and thereby promote a better distribution of kidneys. But just social rules do not mandate the forced redistribution of kidneys from those who have two to those who have none, though doing so would also optimize the distribution of kidneys. Once again, treating recipients justly does not boil down to promoting the best distribution. What matters is how social rules *treat* their recipients, not merely how they affect the quality of the distribution among these recipients.

This distinction has been remarkably neglected in current work on social justice. It is not surprising that it plays no role in consequentialist theorizing, which is committed to the view that social rules should be judged by their effects on the overall outcome, irrespective of how they produce these effects. Remarkable is that Rawls's theory, billed as a deontological alternative, likewise makes the justice of social rules depend exclusively on the overall distribution they tend to produce. Rawls assigns no moral relevance to *how* a basic structure design affects the overall distribution of goods and ills. Most of his critics follow him in this.

At least this decision is clearly built into the centerpiece of Rawls's theory: "the persons in the original position try to acknowledge principles which advance their systems of ends as far as possible. They do this by attempting to win for themselves the highest index of primary social goods, since this enables them to promote their conception of the good most effectively whatever it turns out to be. The parties … strive for as high an absolute score as possible" (Rawls, 1971, 144). By

judging alternative social orders in this *purely recipient-oriented* way, Rawls ensures from the start that they are judged exclusively by their 'output' in terms of what overall distributions they generate among their participants – without regard to how they may bring it about that particular social goods or ills end up with particular persons. While Rawls differs from utilitarians in regard to how distributions should be coded and assessed – fulfillment of higher-order interests rather than happiness, and maximin rather than sum ranking or averaging – he shares with them the purely recipient-oriented mode of theorizing about justice.

This approach should lead Rawls to hold that, to be just, social rules must compensate for natural handicaps. It is evident that the extent to which a person's higher-order interests, especially the third, are fulfilled depends not merely on her share of social goods but also on her natural endowments. Someone genetically predisposed toward good health, cheerfulness, intelligence, and good looks is better able to advance her conception of the good than someone genetically predisposed toward sickliness, melancholy, low intelligence, and homely looks. How, if at all, would the parties in the original position take account of this fact? Leaving genetic modifications aside, the parties can formulate a criterion of justice that takes account of the natural distribution by preferring basic structure designs that favor the naturally disfavored in the distribution of social goods (ensuring that no one is very badly off in terms of both natural and social goods) insofar as doing so raises the floor of higher-order interest fulfillment.

Rawls claims nonetheless that his parties would adopt a criterion of justice that assesses each basic structure design solely on the basis of the distribution of *social* primary goods it produces: "a hypothetical initial arrangement in which all the social primary goods are equally distributed … provides a benchmark for judging improvements" (Rawls, 1971, 62). The two principles do not take into account natural primary goods, such as "health and vigor, intelligence and imagination" (*ibid.*). "The natural distribution is neither just nor unjust" (Rawls, 1971, 102).

How might one react to Rawls's apparent inconsistency? First, one might reject and revise his exclusive emphasis on the distribution of *social* goods and ills. In this vein, Amartya Sen (1982, 2009) has argued that, since natural differences among persons evidently influence their quality of life, we should focus on the distribution not of social primary goods but of capabilities. How badly off someone is depends on her access to being well nourished, adequately clothed and sheltered, and so on, which, in turn, depends not merely on her income but also on her physiology as well as the climate and culture she confronts. To this critique, Rawls has responded that he had, for purposes of a first approximation, left severe natural handicaps and disabilities aside and that the remaining natural disadvantages – Sen had mentioned differential caloric and textile needs – are so minor that their accommodation within a criterion of justice is not worth the cost in terms of increased complexity and consequent divisiveness. But this response is unconvincing: even ordinary natural handicaps – such as shortness, poor posture, ugliness, or a tendency toward melancholy and depression – can be quite devastating. And even if severe natural handicaps of mind and body can be left aside in a first approximation, they cannot be ignored in the final version of a theory that real-world citizens are to use as a shared basis for shaping and reforming their society's basic social institutions.

Second, one might defend Rawls by arguing that natural inequalities are of little importance to a person's *true* quality of life: her ability to fulfill her higher-order interests. This response is also unpromising. Innate mental disabilities, resulting in low intelligence or an inability to concentrate, can surely undermine a person's capacity to understand and apply a conception of justice as well as her capacity to form, revise, and rationally pursue a conception of the good. Such disabilities, and genetically based homeliness and melancholy as well, surely also affect profoundly a person's success in advancing her particular conception of the good.

Third, following Rawls's proposed criterion of justice at the expense of his purely recipient-oriented thought experiment for deriving this criterion, one might argue that the distribution of natural goods and ills, however important to persons, is irrelevant to the comparative assessment of basic structure designs, which should be based exclusively on the distribution of *social* goods and ills that each generates. The task of a scheme of social rules is to do justice to, or to treat justly, all those whose lives are regulated by this scheme. This requires a just allocation of the benefits and burdens of social cooperation, not promotion of the best attainable distribution of all social and natural goods and ills that may be relevant to higher-order interest fulfillment or quality of life. Natural inequalities precede social rules, so to speak, as persons already have their genetic endowments and handicaps when they enter society. A justly designed basic structure should not, then, make the naturally favored compensate the naturally disfavored any more than a just parent, administrator, or civil judge should use benefits and burdens under her control to even out natural inequalities. This view faces the difficulty of drawing precisely enough the distinction between social and natural inequalities to which it gives so much moral significance. It is not clear how one can, without considerable arbitrariness, measure the extent to which a given disease or inequality in earning power is due to genetic and to social factors. Taken strictly, this view would moreover lead to the controversial conclusion that there is nothing unjust about a society that has a just economic system but then leaves the allocation of medical care (as well as of the means of mitigating melancholy, homeliness, low native intelligence, etc.) entirely to the market, where those with special disabilities are bound to pay more for insurance.

Toward a More Adequate Theory

A fourth option is to develop a conception of social justice that neither (like Rawls's two principles) ignores natural goods and ills nor (as the original position would suggest) takes account of them on a par with social goods and ills. The key idea is that a criterion of social justice must be sensitive not merely to the intrinsic features of goods and ills (their kind, magnitude, and frequency), which make them relevant to the recipients' quality of life, but also to an extrinsic feature, namely, the *causal pathway* by which social rules affect their distribution. Such a criterion must take account of – and its application thus requires information about – the particular relation between

social rules and human quality of life, which may affect whether some institutionally avoidable shortfall is an injustice at all and, if so, how great an injustice it is.

The intuitive moral significance of this relational feature can be shown by distinguishing, in a preliminary way, six ways in which a social order may affect the goods and ills of its recipients. The following illustration uses six different scenarios, arranged in order of their intuitive moral significance, in which, due to social rules, certain innocent persons are avoidably deprived of some vital nutrients V (e.g., the vitamins contained in fresh fruit). First-class shortfalls are *officially mandated*, paradigmatically by the law: legal restrictions prevent certain persons from buying foodstuffs containing V. Second-class shortfalls arise from *legally authorized* conduct of private subjects: sellers of foodstuffs containing V lawfully refuse to sell to certain persons. Third-class shortfalls are ones that social institutions do not specifically require or authorize but rather *foreseeably and avoidably engender* through conduct they stimulate – as when certain persons, excluded from employment within an ill-conceived economic order, cannot afford to buy foodstuffs containing V. Fourth-class shortfalls arise from private conduct that is *legally prohibited but barely deterred* – as when, emboldened by mild penalties and lax enforcement, sellers of foodstuffs containing V illegally refuse to sell to certain persons. Fifth-class shortfalls arise from *natural factors whose effects social rules avoidably leave unmitigated* – as when certain persons, unable to metabolize V due to a treatable genetic defect, lack access to the treatment that would correct their handicap. Sixth-class shortfalls, finally, arise from *self-caused factors whose effects social rules avoidably leave unmitigated* – as when certain persons, unable to metabolize V due to a treatable self-caused disease (brought on, perhaps, by their maintaining a long-term smoking habit in full knowledge of the medical dangers associated therewith), lack access to the treatment that would correct their ailment.

Even without all necessary qualifications and refinements, this list makes clear that it is highly relevant to the justice of social rules *through what causal pathway* they affect the distribution of goods and ills. It is no truer of social rules than of subjects and conduct that they are just if and insofar as they promote a good distribution.

How can the moral significance of this relational feature be explained? Social rules are not merely imposed *on* people, their *recipients*, but also imposed *by* people. Purely recipient-oriented theorizing ignores these imposers of social rules, and the passive concept of justice has no place for them. Once the assessment of social rules takes account of both recipients and imposers, the sought explanation is at hand: social rules that cause harms by mandating them are more unjust – because they reflect a different attitude on the part of their imposers – than social rules that engender equally frequent and severe harms through overly mild penalties and lax enforcement. The first injustice is close to the negative-duty end of the spectrum (bringing harm about), while the second injustice is closer to the positive-duty end (letting harm happen).

This hypothesis also explains why justice does not require social rules that mandate beatings whenever the total number of beatings is thereby reduced – and why it may (*pace* Rawls, 1971, ch. 4) oppose official restrictions of basic liberties that, by making such liberties generally more secure, enhance citizens' freedom overall. Rawls must count official and private threats to basic liberties on a par: from the prudential perspective of prospective citizens, it is no worse for one's basic liberties to be threatened by laws or government agents than for them to be equally threatened by criminals, terrorists, or drunk drivers (Rawls, 1971, 242). Here Rawls collides with the widely held belief that official violations are morally more significant, so that, for example, a society should not allow its government to restrict or violate the basic liberties of some or all citizens whenever doing so will prevent greater harms to basic liberties by private actors. The mere fact that some 'reform' enhances the safety of the representative citizen on balance is not enough to justify weakening constraints on policing (during a 'war' on drugs or terror) or introducing lower standards of evidence or disproportionate punishments (such as the death penalty for recidivist drunk drivers).

This diagnosis suggests that a genuine alternative to purely recipient-oriented theorizing can be developed by deploying a more active concept of justice, one that includes a fourth essential place for what we might call the *agents* of justice, those who bear or share moral responsibility for the (in)justice of the *iudicandum*. Pursuant to this alternative concept, the injustice of any *iudicandum* presupposes some agent(s), D, responsible for its injustice or for making it more just. Calling this the active concept of justice alludes to its diverting some attention from those who suffer (in)justice to those who produce it.

While the passive concept easily leads to a conception of justice as *distributional*, the active concept suggests a conception of justice as *relational*, involving a relation between particular agents and recipients mediated through a *iudicandum*. By drawing attention to what we owe one another (Scanlon, 1999), to each agent's specific responsibilities for particular *iudicanda* and, through them, to their particular recipients, the active concept leads away from the idea that a concern for the justice of a *iudicandum* is simply a concern for its promoting a good distribution among its recipients. Rather than tie justice to an outcome and thus view it as something that *obtains* to a greater or lesser extent (what would justice be like, and what ought recipients be able to have or to be?), the active concept presents justice as something that *is done* to a greater or lesser extent: by particular agents, through a *iudicandum*, to particular recipients.

To conclude, unjust subjects, conduct, and social rules involve victims unjustly harmed by them. But it does not follow that the magnitude of the injustice is proportional to the magnitude of these harms. The magnitude of the injustice may also vary with how these harms are produced. The active concept of justice invites sensitivity to such relational factors. In regard to social rules, such sensitivity can take the form of a scalar version of the familiar distinction between negative and positive duties. Such a scale is needed to weight the harms that rules produce by how they produce them (mandate, authorize, engender, etc.). A similar such scale is needed to gauge the degree to which different agents contribute to imposing unjust rules. Moral responsibility for the justice of some *iudicanda* may vary from agent to agent. So there may be competition among *iudicanda* – one may have responsibilities with regard to achieving and maintaining the justice of several *iudicanda*, and one may then have to decide how much of an effort one ought to make with regard to each. These issues concerning

responsibilities and their prioritization are crucial for giving justice a determinate role in the real world. And they tend to be overlooked from the start – or grossly oversimplified through a generic duty to promote just *iudicanda* – when the topic is approached in terms of the passive concept of justice.

See also: Consequentialism (Including Utilitarianism); Contractarianism; Equality: Philosophical Aspects; Intergenerational Justice; Kantian Ethics and Politics; Rawls, John (1921–2002); Reflective Equilibrium.

Bibliography

Ackerman, Bruce, 1980. Social Justice and the Liberal State. Yale University Press, New Haven.
Aristotle, 1985 [ca. 330 BCE?]. Nicomachean Ethics. trans.T. Irwin. Hackett, Indianapolis.
Barry, Brian, 1989. Theories of Justice. University of California Press, Berkeley.
Bentham, Jeremy, 1982 [1789]. In: Burns, J.H., Hart, H.L.A. (Eds.), An Introduction to the Principles of Morals and Legislation. Methuen, London and New York.
Cohen, G.A., 2008. Rescuing Justice and Equality. Harvard University Press, Cambridge, MA.
Dworkin, Ronald, 1981. What is equality? Philosophy and Public Affairs 10 185–246, 283–345.
Fleischacker, Samuel, 2005. A Short History of Distributive Justice. Harvard University Press, Cambridge, MA.
Gauthier, David, 1987. Morals by Agreement. Cambridge University Press, Cambridge.
Hume, David, 1983 [1751]. In: Schneewind, J.B. (Ed.), An Enquiry Concerning the Principles of Morals. Hackett, Indianapolis.
Kant, Immanuel, 1996 [1797]. The Metaphysics of Morals. transl. Mary Gregor. Cambridge University Press, Cambridge.
Locke, John, 1960 [1690]. In: Laslett, Peter (Ed.), Two Treatises of Government. Cambridge University Press, Cambridge.
Nozick, Robert, 1974. Anarchy, State, and Utopia. Basic Books, New York.
Okin, Susan, 1991. Justice, Gender and the Family. Basic Books, New York.
Plato, 1974 [ca. 380 BCE]. The Republic. transl. G.M.A. Grube. Hackett, Indianapolis.
Rawls, John, 1971. A Theory of Justice. Harvard University Press, Cambridge MA.
Rawls, John, 1993. Political Liberalism. Columbia University Press, New York.
Sandel, Michael, 2012. Justice: What is the Right Thing to Do? Farrar, Straus and Giroux, New York.
Scanlon, Thomas M, 1999. What We Owe To Each Other. Harvard University Press, Cambridge, MA.
Sen, Amartya, 1982. Equality of what? In: Sen, A. (Ed.), Choice, Welfare and Measurement. Cambridge University Press, Cambridge.
Sen, Amartya, 2009. The Idea of Justice. Harvard University Press, Cambridge, MA.
Sidgwick, Henry, 1981 [1874]. The Methods of Ethics. Hackett, Indianapolis.
Smith, Adam, 1976 [1759]. In: Raphael, D.D., Macfie, A.L. (Eds.), The Theory of Moral Sentiments. Oxford University Press, Oxford.
Walzer, Michael, 1984. Spheres of Justice. Basic Books, New York.

Justice: Social Psychological Perspectives

Ali Kazemi, University of Skövde, Skövde, Sweden
Kjell Törnblom, ETH Zurich, Zurich, Switzerland
Gerold Mikula, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G. Mikula, volume 12, pp. 8063–8067, © 2001, Elsevier Ltd.

Abstract

This article discusses the concept of justice from a social psychological perspective. The authors begin by briefly outlining some theoretical precursors of current social psychological research on justice. Another part is devoted to an overview of theories and empirical research in the areas of distributive and procedural justice. The article furthermore addresses interactional justice, why people care about justice, and reactions to injustice in the context of which individual differences in conceptions of justice are briefly discussed. The article ends by discussing some current trends, social applications, and some future challenges in social psychological inquiries of social justice.

The Social Psychological Discussion of Justice

In general terms, justice is assumed to prevail when people get what they are entitled to or deserve on the basis of who they are and/or what they have done. Although relatively clear in abstract terms, to be practically useful this definition needs to be more specific about exactly why, what, and how much particular people are entitled to receive with regard to benefits as well as burdens. Similarly, the basic rule of distributive justice, that people have the right to be equally treated as others who are alike them, leaves unsettled the criteria according to which the likeness of people should be assessed and the kind of treatment that the people should get. Justice can be unambiguously defined only on an abstract level. The abstract definition is, however, open to multiple translations into concrete terms. For that reason, different people or groups frequently disagree in their justice judgments of given conditions or circumstances. Even if views of justice are socially shared, this does not change the basically subjective nature of justice judgments.

In contrast to studies of justice within other scientific disciplines, social psychologists take a descriptive rather than a normative (prescriptive) approach. Their aim is not to stipulate how justice *should* be defined and achieved. Instead, they focus on the subjective sense of (in)justice and its impact on human action and judgment. Thus, social psychologists study what people regard as just and unjust under given circumstances, how people deal with the concept of justice, how they react to situations that they regard as unjust, and under which circumstances, and why, people care about justice.

The subjective nature of justice judgments makes it plausible that people will differ in their assessments of a given situation. Divergent evaluations and interpretations are particularly likely among people who view the situation from different perspectives because they occupy different social positions (Mikula et al., 1998a; Tyler et al., 1997; Walster et al., 1978). The following three perspectives are particularly relevant: the victims (i.e., those who suffer from injustice), the perpetrators (i.e., those who caused injustice), and the observers (i.e., those who witness injustice without being personally affected). Perpetrators typically tend to regard their own behavior as appropriate and just, and they frequently attribute any negative behavior on their part to external circumstances. Victims and observers are more likely to regard the same behavior as unjust and attribute responsibility and blame to the actor (cf Mikula, 1994).

Origins and Early Development of the Field

The social psychological discussion of justice originated from relative deprivation research (Runciman, 1966; Stouffer et al., 1949) and Homans' (1961) propositions about distributive justice. According to Homans (1982) another intellectual root of social justice research is the application of the frustration–aggression principle in social exchange theory. Although in slightly different terms, all these writings proposed that people evaluate outcomes they receive relative to some kind of comparison standard rather than in absolute terms. The comparison standards define the quality of outcome people expect to receive based on the outcomes that other relevant people receive. Thus, the evaluation of a given outcome may vary depending upon the particular comparison standard used. The outcome will be experienced as gratifying and/or just if it is equal to the one received by the comparison standard and unfavorably or favorably unjust if it is lower or higher, respectively. Crosby (1976) suggested that deprivation is related to feelings of resentment, and to feel resentment, five conditions have to be satisfied. The individual who lacks something (X) must (1) perceive that another individual possesses X; (2) have a desire for possessing X; (3) feel entitled to possess X; (4) consider coming into possession of X as feasible; and (5) not feel responsible for lacking X.

Adams' (1965) elaborations (to be described under the next section) refined and extended Homans work and were influenced by Festinger's social comparison (1954) and cognitive dissonance (1957) theories. It should be noted that formulas similar to Adams' had already been proposed by Sayles (1958) and Patchen (1961). These models were initially employed to explain industrial workers' reactions to under- and overpayment at work. Later, Walster et al. (1978) developed equity

theory as a general theory of social justice and applied it to various kinds of social interactions and relationships.

Another important impetus was based on the *just world hypothesis* (Lerner, 1980) according to which people have a basic desire to believe in a just world where people get what they deserve. This desire prompts people to act or engage in cognitive distortion to maintain or reestablish the belief that the world around them is just.

Distributive Justice

Early social psychological conceptualizations of justice focused on one particular form, namely *distributive justice* (i.e., the justice of the distribution of outcomes or resources), as well as on mainly one particular principle of distributive justice, the equity or contribution principle (although equality was intermittently and unsystematically discussed).

An important step in the development of the study of distributive justice was the research and theorizing by Adams (1965) on affective, cognitive, and behavioral consequences of perceived inequity. Like Homans, Adams suggested that people evaluate what they contribute to an interaction or relationship (i.e., their inputs) and what they receive in return (i.e., their outcomes) by comparing the ratio of their own outcomes and inputs with that of others and expect the ratios to be equal. If the outcome–input ratios of a person and the comparison other do not match, people experience inequity. Inequity is an aversive state that motivates people to restore equity either 'actually' by behavioral means or 'psychologically' by cognitive redefinition of the situation. Interestingly, this is assumed to hold equally true for people who are worse off as well as for those who are better off than the comparison other.

The narrow focus of early theory and research on the equity principle was later extended in several respects. Alternative conceptualizations, the multiprinciple approaches, emphasized the existence of several justice norms, in addition to the equity or contribution principle. The equality (to each the same, regardless of inputs) and the need principles (to each according to actual or perceived needs) are two of the most commonly discussed. These approaches assume that different distribution rules are regarded as just and serve as standards of action and evaluation in different circumstances (Deutsch, 1975; Mikula, 1980). For example, the equality principle is assumed to predominate in groups for which solidarity is important, the equity/contribution principle in groups that strive to maximize production, while the need principle is favored by groups that are concerned about the care and nurturing of its members. Empirical studies in this tradition were designed to identify situational and personal conditions that determine which distribution principles are regarded as just and used as a criterion for resource distribution, as well as what positive or negative consequences (regarding work satisfaction, task performance, personal relationships, etc.) will follow from the application or violation of the various principles (Törnblom, 1992).

The various 'main' distributive justice principles may be further specified (or operationalized) into several 'subrules.' For example, the equity/contribution principle may be conceptualized as contributions of effort, accomplishment, or ability; the need principle may be conceived in terms of biological, basic, or functional needs; and equality may be understood as equality of opportunity, treatment, or results. Thus, a specification of a justice principle into the particular subrule that is meant to represent it promotes more conceptual clarity and precision in research, hinders arbitrary practical applications, as well as minimizes the occurrence of communication conflicts (e.g., Kazemi and Törnblom, 2011; Törnblom and Jonsson, 1985).

Procedural Justice

Further developments in the field widened the focus beyond distributive justice and considered additional forms of justice. Justice judgments are frequently made about the decision-making procedures by which the distributions are arrived at as well, that is, procedural justice. The social psychological study of procedural justice was stimulated by the work of Thibaut and Walker (1975). These authors studied people's responses to different third-party procedures of conflict resolution. They found that people regard procedures that provide them with control of the decision-making process (process control or voice) and/or the decision itself (decision control) to be fairer than when either type of control is lacking. In addition, the perceived fairness of the procedures turned out to affect people's satisfaction with their outcomes. Subsequent research studied particular features of the procedures that seemed to contribute to the perception of procedural (in) justice and the consequences of these perceptions (Lind and Tyler, 1988; Tyler et al., 1997).

Justice researchers identified a variety of procedural elements that contribute to the perceived fairness of procedures. For instance, Leventhal (1980) proposed that procedures will be regarded just if they ensure consistent treatment across persons and over time, utilization of accurate information, suppression of personal biases of decision makers, existence of appeal mechanisms by which wrong decisions can be corrected, representation of the affected parties in the decision-making process, and compatibility with fundamental moral and ethical values.

Empirical evidence suggests that people consider procedures that allow them to participate in the decision-making process and express their opinions and views (voice) to be fairer. Interestingly, this seems to occur even when the expressed opinion does not influence the final decision, as long as the respective parties feel that their views have been seriously considered (e.g., Lind et al., 1990).

The perception that the procedures are just has a variety of positive consequences. Procedural justice promotes people's acceptance of decisions, their long-term adherence to agreements, and their willingness to cooperate with groups they belong to. Perceived procedural justice also enhances the perceived legitimacy of civil authorities, institutions and rules, and people's willingness to defer to authorities and their decisions. Finally, perceived procedural justice improves people's satisfaction with the outcomes they receive. Strikingly, these positive consequences result even if the decision procedures lead to unfavorable outcomes for the persons concerned (Tyler et al., 1997).

Interactional Justice

Distributive and procedural justice has received the most attention in the literature. But since the mid-1980s, theory and research have also emphasized the importance of the social and informal aspects of justice in daily encounters (Bies and Moag, 1986; Greenberg, 1993; Mikula et al., 1990). Greenberg (1993) pointed out the importance of considering interactional justice in justice theorizing and distinguished between interpersonal and informational justice as two forms of interactional justice. This has subsequently been supported by Colquitt et al. (2001; see also Colquitt, 2001). There is however some controversy concerning whether or not interactional justice should be subsumed under the broader concept of procedural justice (Bies, 2005; Bobocel and Holmvall, 2001). Procedural and interactional justice may constitute aspects of 'fairness of treatment' (Van den Bos, 2005).

While procedural justice is related to the formal and structural aspects (i.e., rules and policies), interpersonal justice concerns the quality of the relationship or the exchange and social processes taking place between people. Just as distributive and procedural justice is defined in terms of a number of criteria or principles, interpersonal justice is defined in terms of four main criteria: respect, truthfulness, justification, and propriety (Bies and Moag, 1986). Information justice refers to how allocation decisions or enacted procedures are explained to those concerned – the type of information given or how it is given to the individuals. This gives the impression of neutrality and increases people's trust in the authorities. Research on information justice and its effects are very scarce (see Kazemi and Törnblom, 2009).

Why Do People Care About Justice?

The question of why people care about justice received different answers in different social psychological theories. One group of theorists assumes that people are concerned about justice because it serves their interest in maximizing their own outcomes. For instance, equity theorists (Walster et al., 1978) proposed that social systems develop rules of justice, and induce members to follow these rules, because this can maximize collective outcomes. People want to conform to these rules and expect others to do so as long as it serves their self-interest. Lerner and Clayton (2011) recently analyzed existing research relevant to the two motivational forces of justice and self-interest and proposed a model that seeks to integrate them into a comprehensive theory of behavior.

Justice is "conceived as a norm, a value, a virtue, a standard of evaluation of almost any aspect of life and living together, and as a human motive affecting thoughts, emotions, and actions" (Kals and Maes, 2012: p. v). The importance of justice in human social life is related to the fact that justice fulfills four different types of identity maintenance needs: (1) *instrumental* (material identity needs; Thibaut and Walker, 1975; Skitka, 2003), (2) *relational* (social identity needs; Lind and Tyler, 1988; Tyler and Lind, 1992), (3) *protective* (certainty or predictability needs; Lind and Van den Bos, 2002), and (4) *moral* (personal identity needs; Skitka and Mullen, 2002). How justice satisfies each need is briefly described below.

Early theorizing on procedural justice (Thibaut and Walker, 1975) posits that people care about fair procedures, and regard procedures that give them control over outcomes to be fairer, because they help them to maximize their outcomes in the long run. The focus is on the expected and actual outcomes that people receive (i.e., the instrumental perspective). The relational or group-value model of procedural justice (Lind and Tyler, 1988; Tyler and Lind, 1992) provides an alternative perspective. According to this theory, people are concerned about their position in groups. They use previous experiences of how they were treated by authorities as a source of information about their standing or status in the group. Evidence of just treatment is taken as an indicator that they are regarded as valuable and worthy group members that, in turn, has positive implications for their self-esteem and self-worth. Justice judgments are relational in nature and the issues of neutrality, trustworthiness, and status recognition are central to understanding justice. Thus, the relational account holds that people not only have material identity needs, but also need to be treated in a way that makes them feel as respected citizens with certain rights. While providing different explanations of why people care about justice, that is, because of instrumental concerns about outcomes or of relational concerns about group status, both groups of theorists view justice as a means rather than as an end in itself. It seems possible to understand the instrumental and relational models as two sides of the same coin rather than two different coins, that is, the relational model is 'instrumental' as well, as it focuses on obtaining status in the group, while the instrumental model has control as an end. Thus, from that point of view they merely differ in terms of the social resource they see as outcome – instrumentally motivated people seek control and relationally motivated people seek status. Other authors criticize that these views do not capture the essential meaning of justice as a moral ought. Without denying that justice may be used instrumentally as a means to acquire some end or outcome, they argue that justice is inherently an end in itself. People are motivated intrinsically to behave justly, and sometimes act in ways that are irrelevant to self-interest motives and maximization of own outcomes.

According to the protective account, fairness is important as it is conducive to reducing the uncertainty that citizens may experience in their relationships with authorities and fellow citizens. Justice is a norm, and as such functions to coordinate behavior. If I believe in fair play, and expect others to do so, a kind of consensus effect results. Uncertainty management theory (Lind and Van den Bos, 2002) posits that activation of fairness processes is an indication of fairness judgments being utilized to settle some important social or psychological issue. Lind and Van den Bos (2002) suggest that fairness serves people in managing uncertain situations (e.g., are everyone else going to pay his/her taxes?) by giving them "confidence that they will ultimately receive good outcomes and because it makes the possibility of loss less anxiety-provoking" (p. 195).

Finally, the moral account stresses the importance of personal moral standards as an important determinant of how people reason about fairness. For instance, if the individual strongly believes that a mismatch between norms of input (e.g., equitable tax payment) and norms of outcome (e.g., equal or need based access to and utilization of

Medicare and educational services) is immoral and wrong, then s/he tends to perceive the system as unfair regardless of procedural or interactional justice considerations (Skitka and Mullen, 2002).

Reactions to Injustice

Social psychological theory and research distinguished three major types of responses to injustice: (1) cognitive (psychological) restoration of justice by changing the interpretation of the situation, (2) actual restoration of justice by behavioral means, and (3) nonaction or resignation (Adams, 1965; Tyler et al., 1997).

People may prefer cognitive over behavioral reactions to injustice for various reasons. People who caused injustice, and people who are unfairly advantaged, may be inclined to justify the given situation to avoid the need for compensation or restitution and to protect a positive self-image. People who are unfairly disadvantaged may be motivated to deny being victims of injustice to protect their self-esteem, or they may conclude that they are responsible for their own fate. They may also view their negative situation as a result of a legitimate process and thus not regard it as unjust. One's 'moral community' might be redefined, excluding some victims from the rights to be treated justly. Furthermore, moving the focus from one's own situation to that of others,' observers frequently distort others suffered injustices as a result of their own behavior ('blaming the victim,' 'the just world hypothesis').

Behavioral attempts to restore justice include seeking compensation from or retaliating against the responsible agent (sometimes a third party). Examples are disobedience, theft, sabotage, violence, murder, and sick leave. Leaving the unjust relationship is another alternative, even though this option does not restore justice. The likelihood of behavioral reactions of the former type increases when people see a chance of improving the situation, and when they assume that the situation would not change 'by itself.' In addition, anticipated personal benefits and costs may influence the likelihood of behavioral reactions. However, the intensity of experienced injustice seems to be the most important factor. People with strong feelings of injustice are more likely to react behaviorally to injustice regardless of the costs involved and/or the likelihood of success. Behavioral reactions may occur as individual-level responses, collective-level responses or both. Collective reactions are more likely when the injustice is perceived as a collective rather than an individual issue and may be expressed in the form of revolts, strikes, terrorism, war, and the like.

An implicit assumption of equity theory was that a person's inequitable treatment in one relationship will not affect her behavior in other second-partner relationships. In such a case, an additional new inequity would be created. However, studies by Austin and Walster (1974, 1975) suggest that people sometimes try to do so; they focus on equity with the world (i.e., equity across several relationships) rather than person-specific equity. They proposed that individuals are sometimes interested in the amount of net equity they receive across relationships, while at other occasions they may prefer both person-specific equity and equity with the world. Furthermore, Moschetti and Kues (1978) found that transrelational redress by inequitably treated subjects took place in third-partner relationships as well. Some shootings at schools and work places are tragic examples of transrelational equity or equity with the world.

Although fundamental and highly relevant, the daily observation that people sometimes do not react to injustice has not been systematically investigated by researchers in the field (see Törnblom and Kazemi, in press). People may not react to injustice (cognitively or behaviorally) because they feel powerless and have resigned to their fate. Another reason is that people may be afraid of retaliation. This is particularly true in asymmetrical relations (e.g., employer–employee relationship). Finally, they may perceive their situation as normal and appropriate because they compare themselves predominantly with similar others who are also disadvantaged. Sometimes people are even inclined to regard 'what is' as 'what ought to be' (Moore Jr., 1978).

We have already discussed the cognitive and behavioral consequences of injustice but only hinted at subsequent emotional reactions, as broadly conceived by theorists like Homans (1961) and Adams (1965). These theorists proposed that people are likely to experience distributive injustice when individuals in an exchange relationship receive rewards that are seen as disproportional to their investments. People were assumed to react with anger to unfavorable and with guilt to favorable disproportionality. Adams (1965) used the general terms distress and dissatisfaction to indicate that perceived injustice causes aversive affective reactions. Specific examples are stress, dissatisfaction, depression, self-depreciation, feeling of guilt, loss of motivation, and apathy. The connection between emotion and justice is complex and, until recently, a neglected area of theory and research. As noted by several justice researchers (e.g., Hegtvedt, 2006; Mullen, 2007) as emotions are capable of influencing cognitions and evaluations (and behavior), negative affect may very likely precede and result in perceptions of injustice (in addition to the other way around) which, in turn, will lead to additional affective responses (for studies on emotions as reaction to perceived injustice see Clayton, 1992; Mikula et al., 1998b; Sprecher, 1992). Justice judgments (which are cognitive assessments) are assumed to be more time consuming than the less conscious types of emotional reactions which are more immediate and spontaneous.

During the past couple of decades some justice researchers have focused their attention on the importance of *individual* differences in the context of tolerance for or sensitivity to violations of moral norms and injustices (Huseman et al., 1987; Lovas and Wolt, 2002). Two such individual difference constructs related to justice reasoning and reactions to injustice are *belief in a just world* and *justice sensitivity*. Lerner (1980) assumed that people generally believe that the world is a fair place. When this idea is threatened, people usually employ various protective strategies. The conception of the world as a fair (and predictable) place seems more firmly rooted in some people than in others and suggests that people get what they deserve and deserve whatever positive or negative treatments they get. Various measures have been used to tap the construct (see Dalbert, 2001; Lipkus, 1991). Injustices are a major contradiction for individuals who score high on such a scale (i.e., perceive the world as a very fair place). They

are more aversive to injustices, see them as more threatening, and react more strongly than those who see the world as less fair. Justice sensitivity (Schmitt et al., 1995, 2005) is conceived as a dispositional sensitivity to befallen injustice, comprised by four components: frequency (of experiencing yourself as a victim of injustice), intensity of anger (subsequent to unfair distribution or treatment), intrusiveness of thoughts (moral mental load, how much the person dwells on an unfair event), and punitivity (to what extent and in what way you think that an agent of injustice should be forced to restore justice).

Some Areas of Current Interest and Social Applications

Social justice research has expanded to include analyses of several different areas of social life: work organizations, legal institutions, educational settings, civil institutions, social policy and economics, gender issues, other than human species, environmental issues, international conflicts, family issues, close personal relationships, and coping with life crises and victimization. People's perceptions of justice and their reactions to injustice in work organizations have been one of the main areas of social justice research from the very beginnings of equity theory. The rapid growth of this line of research has led some authors to use the generic term 'organizational justice' (Cropanzano and Greenberg, 1997). Empirical evidence shows that the perception of distributive justice affects, for example, work performance, absenteeism, turnover, and people's satisfaction with outcomes they receive. The perception of procedural justice influences workers' organizational citizenship behaviors and reactions to a variety of organizational measures such as personnel selection procedures, performance evaluations, layoff decisions, strategic organizational planning and change, and workplace drug screening.

A full description of an exchange or distribution situation must include information about the object of distribution. What kinds of goods or bads are being distributed, and does the particular type of transferred item affects the process and outcome of the distribution? These types of questions are addressed by *resource theory of social exchange* (Foa, 1971; Foa and Foa, 1974). It features a classification of resources (i.e., any object which is transmitted via interpersonal behavior) into six types: the value of love, status, and services (particularistic resources) is derived from the identity of the provider and his/her relationship to the recipient, which is generally not the case for money, goods, and information (universalistic resources). The exchangeability of these resources (some concrete and some abstract) varies with the personal/social/cultural context within which the resource transaction takes place. The various properties of the resources yield a number of new exchange rules. Resource theory has implications for justice theory in that inappropriate exchanges may generate dissatisfaction and feelings of injustice. Receiving a friendly pat on the back from a customer in return for repairing his car is likely to result in dissatisfaction and arouse feelings of injustice. Furthermore, a given allocation principle may be viewed as just for the distribution of some resources but not for others. Workers usually prefer wages to be paid according to contributions rather than equally to all. The resource features that have most frequently been focused in justice research are their nature, quantity, and valence (i.e., whether the resource is a good or a bad, an asset or a liability, a gain or a loss). Resource theory has stimulated an impressive amount of research – theoretical extensions and integrations (e.g., Markovsky and Berigan, 2012; Törnblom and Vermunt, 2007) as well as empirical research and social applications (for more details see Törnblom and Kazemi, 2012a).

An additional resource issue with implications for justice considerations is the way resources are produced and acquired. Social psychological discussions of distributive justice have focused on resources of consumption to the neglect of the connection between justice and the production of those resources and the access to their means of production. Whether or not a fur coat is genuine or an ornament is made of real or imitation ivory (i.e., produced by murdering animals) will have implications for justice evaluations of exchanges and distributions. The complex linkages between justice and the production, possession, and provision of transacted social resources generate new research avenues (Törnblom and Kazemi, 2007).

Due to the subjective nature of justice, misunderstandings, disagreements, and conflicts between involved parties will easily occur. As Montada (2007: p. 256) points out, "All social conflicts may be interpreted as justice conflicts," and "The aim of conflict settlement is a just peace" (Montada, 2012: p. 15). Relatively scant theoretical as well as empirical attention has been given to this important area of inquiry, perhaps partly due to the imprecise meaning of the umbrella concept of conflict (see Törnblom and Kazemi, 2012b) – for exceptions see Kals and Maes (2012) and Mikula and Wenzel (2000).

Other Areas of Interest and Future Challenges

Frequently individuals and collectives (as well as animals and the environment) are excluded from our moral community. Thus, justice is considered irrelevant and not applicable to some entities in some situations and contexts which, in turn, may lend legitimacy to morally unacceptable treatment and victimization of humans and other animals. Several prominent justice researchers have contributed valuable insights concerning the 'scope of justice' (notably Opotow, 1990, 1993, 1995). A thorough analysis of empirical research within this area of investigation was conducted by Hafer and Olson (2003). As moral behavior and justice are also relevant guidelines in the context of plants and nonliving entities, the threats from global warming, natural resource depletion, and pollution attract much needed research within the area designated as environmental justice (see Clayton and Myers, 2009; Clayton and Opotow, 2003).

Additional areas deserving research attention are intergroup justice, justice versus other competing motives guiding behavior, the proactive usage of justice to promote well being, and theory integration. Further inquiries into the processes in the connection between emotion and justice is another challenging task that would take justice research to the next level (see Cropanzano et al., 2011). These lines of research merely

exemplify the diversity of topics awaiting further theoretical and empirical research (Kazemi and Törnblom, 2008).

Existing social psychological inquiries of justice are to a great extent experimentally oriented; research hypotheses are investigated under controlled conditions. Even though some qualitative studies have been conducted lending ecological validity to experimental results (e.g., Clayton, 1992; Mikula, 1986; Randall and Mueller, 1995), additional qualitative justice research are needed to gain insights into the phenomenology of justice. Also, due to the overrepresentation of research concerned with self-reported behaviors and behavioral intentions, research featuring actual behavior observations is warranted. Furthermore, with the relative rapid accumulation of empirical evidence in the field of justice studies, the need is increasing for theoretical integration, propositional inventories, summaries of research evidence, and meta-analyses. Meta-analyses provide us with effect sizes and assist in identifying boundary conditions and directions in which new research are needed.

Concluding Comments

Some of today's most important global problems are gender inequality, immigration and minority issues, allocation of resources in health care, discrimination, housing segregation, poverty, financial aid, the prevention and resolution of conflict and war, to mention a few. The core of these problems generally boils down to problems of justice. As social psychological research has convincingly shown that people care about justice, and that justice norms and conceptions influence human conduct and thinking, this research may contribute to our understanding and the amelioration of social problems on different levels – the interpersonal, intergroup, organizational, societal, and cultural.

The field of social justice research is very vital, attracting interest of an increasing number of researchers from different disciplines. The dominant social psychological line of research is published in highly ranked international journals such as the *Journal of Personality and Social Psychology*, *Personality and Social Psychology Bulletin*, *European Journal of Social Psychology*, *Personality and Social Psychology Review*, *Journal of Vocational Behavior*, *Journal of Applied Psychology*, *Social Issues, and Social Forces*. *Social Justice Research* (founded in 1987 by Melvin Lerner) is the official journal of the International Society for Justice Research.

See also: Applied Criminology and Criminal Justice; Belief in a Just World; Conflict and Conflict Resolution, Social Psychology of; Crime and Criminal Justice: Applied Research from Routine Monitoring to Evidence-Based Practices; Criminal Justice and Social Work Practice; Criminal Justice, Sociology of; Environmental Justice in the United States; Genetics and Social Justice; Intergenerational Justice; Intergroup Relations; Justice and Health; Justice, Access to; Justice, Conflicts, and the Justice of Conflict Resolution; Juvenile Justice: International Law and Children's Rights; Organizational Justice; Social Action and Social Justice; Social Exclusion, Social Deprivation and Health; Social Justice and Social Inequalities; Social Psychology; Women and Criminal Justice.

Bibliography

Adams, S.J., 1965. Inequity in social exchange. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology. Academic Press, San Diego, pp. 267–299.

Austin, W., Walster, E., 1974. Reactions to "Equity with the world". Journal of Experimental Social Psychology 10, 528–548.

Austin, W., Walster, E., 1975. Equity with the world: the trans-relational effects of equity and inequity. Sociometry 38, 474–496.

Bies, R.J., 2005. Are procedural justice and interactional justice conceptually distinct? In: Greenberg, J., Colquitt, J.A. (Eds.), Handbook of Organizational Justice. Erlbaum, Mahwah, NJ, pp. 85–112.

Bies, R.J., Moag, J.S., 1986. Interactional justice: communication criteria of fairness. In: Lewicki, R.J., Sheppard, B.H., Bazerman, M.H. (Eds.), Research on Negotiation in Organizations. JAI Press, Greenwich, CT, pp. 43–55.

Bobocel, D.R., Holmvall, C.M., 2001. Are interactional justice and procedural justice different? Framing the debate. In: Gilliland, S., Steiner, D., Skarlicki, D. (Eds.), Research in Social Issues in Management: Theoretical and Cultural Perspectives on Organizational Justice, vol. 1. Information Age Publishing, Greenwich, CT, pp. 85–108.

Clayton, S.D., 1992. The experience of injustice: some characteristics and correlates. Social Justice Research 5, 71–91.

Clayton, S., Myers, G., 2009. Conservation Psychology: Understanding and Promoting Human Care for Nature. Blackwell, Oxford, UK.

Clayton, S., Opotow, S. (Eds.), 2003. Identity and the Natural Environment: The Psychological Significance of Nature. MIT Press, Cambridge, MA.

Colquitt, J.A., 2001. On the dimensionality of organizational justice: a construct validation of a measure. Journal of Applied Psychology 86, 386–400.

Colquitt, J.A., Conlon, D.E., Wesson, M.J., Porter, C.O.L.H., Ng, K.Y., 2001. Justice at the millennium: a meta-analytic review of 25 years of organizational justice research. Journal of Applied Psychology 86, 425–445.

Cropanzano, R., Greenberg, J., 1997. Progress in organizational justice: tunneling through the maze. International Review of Industrial and Organizational Psychology 12, 317–372.

Cropanzano, R., Stein, J.H., Nadisic, T., 2011. Social Justice and the Experience of Human Emotion. Taylor and Francis, New York.

Crosby, F., 1976. A model of egoistical relative deprivation. Psychological Review 83, 85–113.

Dalbert, C., 2001. The Justice Motive as a Personal Resource. Springer, New York.

Deutsch, M., 1975. Equity, equality, and need: what determines which value will be used as the basis of distributive justice? Journal of Social Issues 31, 137–149.

Festinger, L., 1954. A theory of social comparison processes. Human Relations 7, 117–140.

Festinger, L., 1957. A Theory of Cognitive Dissonance. Row and Peterson, Evanston, IL.

Foa, U.G., 1971. Interpersonal and economic resources. Science 171, 345–351.

Foa, U.G., Foa, E.B., 1974. Societal Structures of the Mind. Charles C. Thomas, Springfield, Illinois.

Greenberg, J., 1993. The social side of fairness: interpersonal and informational classes of organizational justice. In: Cropanzano, R. (Ed.), Justice in the Workplace: Approaching Fairness in Human Resource Management. Erlbaum, Hillsdale, NJ, pp. 79–103.

Hafer, C.L., Olson, J.M., 2003. An analysis of empirical research on the scope of justice. Personality and Social Psychology Review 7, 311–323.

Hegtvedt, K.A., 2006. Justice frameworks. In: Burke, P. (Ed.), Contemporary Social Psychological Theories. Stanford University Press, Stanford, CA, pp. 46–69.

Homans, G.C., 1961. Social Behavior: Its Elementary Forms. Harcourt, Brace and World, New York.

Homans, G.C., 1982. Foreword. In: Greenberg, J., Cohen, R.L. (Eds.), Equity and Justice in Social Behavior. Academic Press, New York, pp. xi–xviii.

Huseman, R.C., Hatfield, J.D., Miles, E.W., 1987. A new perspective on equity theory: the equity sensitivity construct. Academy of Management Review 12, 222–234.

Kals, E., Maes, J. (Eds.), 2012. Justice and Conflicts: Theoretical and Empirical Contributions. Springer, New York.

Kazemi, A., Törnblom, K.Y., 2008. Social psychology of justice: origins, central issues, recent developments, and future directions. Nordic Psychology 60, 209–234.

Kazemi, A., Törnblom, K., 2011. Perceived fairness of six allocation decision rules in the context of hiring and firing. In: Jacobsson, C., Ricciardi, M.R. (Eds.), The Individual and the Group: Future Challenges. Gothenburg University Press, Gothenburg, pp. 44–51.

Kazemi, A., Törnblom, K., 2009. Rättvist ledarskap (Fair leadership). In: Kazemi, A. (Ed.), Välbefinnande i arbetslivet - socialpsykologiska perspektiv (Social Psychological Perspectives on Well-being in Working Life). Studentlitteratur, Lund, pp. 37–63.

Lerner, M.J., 1980. The Belief in a Just World: A Fundamental Delusion. Plenum, New York.
Lerner, M.J., Clayton, S., 2011. Justice and Self-interest. Two Fundamental Motives. Cambridge University Press, Cambridge.
Leventhal, G.S., 1980. What should be done with equity theory? New approaches to the study of fairness in social relationships. In: Gergen, K.J., Greenberg, M.S., Willis, R.H. (Eds.), Social Exchange. Advances in Theory and Research. Plenum, New York, pp. 27–55.
Lind, E.A., Kanfer, R., Earley, P.C., 1990. Voice, control and procedural justice: instrumental and noninstrumental concerns in fairness judgments. Journal of Personality and Social Psychology 59, 952–959.
Lind, E.A., Tyler, T.R., 1988. The Social Psychology of Procedural Justice. Plenum, New York.
Lind, E.A., Van den Bos, K., 2002. When fairness works: toward a general theory of uncertainty management. In: Staw, B.M., Kramer, R.M. (Eds.), Research in Organizational Behavior, vol. 24. JAI Press, Greenwich, CT, pp. 181–224.
Lipkus, I., 1991. The construction and preliminary validation of a global belief in a just world scale and the exploratory analysis of the multidimensional belief in a just world scale. Personality and Individual Differences 12, 1171–1178.
Lovas, L., Wolt, R., 2002. Sensitivity to injustice in the context of some personality traits. Studia Psychologica 44, 125–131.
Markovsky, B., Berigan, N., 2012. Resource types and fairness perceptions in social dilemmas. In: Törnblom, K.Y., Kazemi, A. (Eds.), Handbook of Social Resource Theory: Theoretical Extensions, Empirical Insights, and Social Application. Springer, New York.
Mikula, G., 1980. On the role of justice in allocation decisions. In: Mikula, G. (Ed.), Justice and Social Interaction: Experimental and Theoretical Contributions from Psychological Research. Springer, New York, pp. 127–166.
Mikula, G., 1986. The experience of injustice: toward a better understanding of its phenomenology. In: Bierhoff, H.W., Cohen, R.L., Greenberg, J. (Eds.), Justice in Social Relations. Plenum, New York, pp. 103–123.
Mikula, G., 1994. Perspective-related differences in interpretations of injustice by victims and victimizers: a test with close relationships. In: Lerner, M.J., Mikula, G. (Eds.), Entitlement and the Affectional Bond. Plenum, New York, pp. 175–203.
Mikula, G., Athenstaedt, U., Heschgl, S., Heimgartner, A., 1998a. Does it only depend on the point of view? Perspective-related differences in justice evaluations of negative incidents in personal relationships. European Journal of Social Psychology 28, 931–962.
Mikula, G., Petri, B., Tanzer, N., 1990. What people regard as unjust: types and structures of everyday experiences of injustice. European Journal of Social Psychology 20, 133–149.
Mikula, G., Scherer, K.L., Athenstaedt, U., 1998b. The role of injustice in the elicitation of differential emotional reactions. Personality & Social Psychology Bulletin 24, 769–783.
Mikula, G., Wenzel, M., 2000. Justice and social conflict. International Journal of Social Psychology 35, 126–135.
Montada, L., 2007. Justice conflicts and the justice of conflict resolution. In: Törnblom, K., Vermunt, R. (Eds.), Distributive and Procedural Justice: Research and Social Applications. Ashgate Publishing Company, Burlington, VT.
Montada, L., 2012. The normative impact of empirical justice research. In: Kals, E., Maes, J. (Eds.), Justice and Conflicts: Theoretical and Empirical Contributions. Springer, New York, pp. 3–19.
Moore Jr., B., 1978. Injustice. The Social Bases of Obedience and Revolt. Macmillan, London.
Moschetti, G.J., Kues, J.R., 1978. Transrelational equity comparisons: extensions to the third partner relationship and a decision-making model. Journal of Personality and Social Psychology 36, 1107–1117.
Mullen, E., 2007. The reciprocal relationship between affect and perceptions of fairness. In: Törnblom, K., Vermunt, R. (Eds.), Distributive and Procedural Justice: Research and Social Applications. Ashgate Publishing Company, Hampshire, England, pp. 15–37.
Opotow, S., 1990. Moral exclusion and injustice: an introduction. Journal of Social Issues 46, 1–20.
Opotow, S., 1993. Animals and the scope of justice. Journal of Social Issues 49, 71–85.
Opotow, S., 1995. Drawing the line: social categorization, moral exclusion, and the scope of justice. In: Bunker, B.B., Rubin, J.Z. (Eds.), Conflict, Cooperation, and Justice. Jossey-Bass, San Francisco.
Patchen, M., 1961. The Choice of Wage Comparisons. Prentice Hall, Englewood Cliffs, NJ.
Randall, C.S., Mueller, C.W., 1995. Extensions of justice theory: evaluations and employees' reactions in a natural setting. Social Psychology Quarterly 58, 178–194.
Runciman, W.G., 1966. Relative Deprivation and Social Justice: A Study of Attitudes to Social Inequality in Twentieth-Century England. Routledge & Kegan Paul, London.
Sayles, L.R., 1958. Behavior of Industrial Work Groups: Prediction and Control. Wiley, New York.
Schmitt, M., Gollwitzer, M., Maes, J., Arbach, D., 2005. Justice sensitivity: assessment and location in the personality space. European Journal of Psychological Assessment 21, 202–211.
Schmitt, M., Neumann, R., Montada, L., 1995. Dispositional sensitivity to befallen injustice. Social Justice Research 8, 385–407.
Skitka, L.J., 2003. Of different minds: an accessible identity model of justice reasoning. Personality and Social Psychology Review 7, 286–297.
Skitka, L.J., Mullen, E., 2002. Understanding judgments of fairness in a real-world political context: a test of the value protection model of justice reasoning. Personality and Social Psychology Bulletin 28, 1419–1429.
Sprecher, S., 1992. How men and women expect to feel and behave in response to inequity in close relationships. Social Psychology Quarterly 55, 57–69.
Stouffer, S.A., Suchman, E.A., De Vinney, L.C., Star, S.A., Williams Jr., R.A., 1949. The American Soldier: Adjustments during Army Life. Princeton University Press, Princeton, NJ.
Thibaut, J., Walker, L., 1975. Procedural Justice: A Psychological Analysis. Erlbaum, Hillsdale, NJ.
Törnblom, K.Y., 1992. The social psychology of distributive justice. In: Scherer, K.R. (Ed.), Justice: Interdisciplinary Perspectives. Cambridge University Press, New York, pp. 177–236.
Törnblom, K.Y., Jonsson, D.R., 1985. Subrules of the equality and contribution principles: their perceived fairness in distribution and retribution. Social Psychology Quarterly 48, 249–261.
Törnblom, K.Y., Kazemi, A., 2007. A resource production approach to distributive justice. In: Törnblom, K.Y., Vermunt, R. (Eds.), Distributive and Procedural Justice: Research and Social Applications. Ashgate Publishing Ltd, London, pp. 39–66.
Törnblom, K.Y., Kazemi, A. (Eds.), 2012a. Handbook of Social Resource Theory: Theoretical Extensions, Empirical Insights, and Social Applications. Springer, New York.
Törnblom, K.Y., Kazemi, A., 2012b. Advances in justice conflict conceptualization: a new integrative framework. In: Kals, E., Maes, J. (Eds.), Justice and Conflicts: Theoretical and Empirical Contributions. Springer, New York, pp. 21–51.
Törnblom, K.Y., Vermunt, R., 2007. Towards an integration of distributive justice, procedural justice, and social resource theories. Social Justice Research 20, 312–335.
Törnblom, K., Kazemi, A., in press. Distributive justice: revisiting past statements and reflecting on future prospects. In: Cropanzano, R., Ambrose, M. (Eds.), The Oxford Handbook of Organizational Justice. Oxford University Press, New York.
Tyler, T.R., Lind, E.A., 1992. A relational model of authority in groups. In: Zanna, M. (Ed.), Advances in Experimental Social Psychology. Plenum, New York, pp. 115–191.
Tyler, T.R., Smith, H.J., 1998. Social justice and social movement. In: Gilbert, D.T., Fiske, S.T., Lindzey, G. (Eds.), Handbook of Social Psychology. McGraw-Hill, Boston, pp. 585–629.
Tyler, T.R., Boeckmann, R.J., Smith, H.J., Huo, Y.J., 1997. Social Justice in a Diverse Society. Westview Press, Boulder, CO.
Van den Bos, K., 2005. What is responsible for the fair process effect. In: Greenberg, J., Colquitt, J.A. (Eds.), Handbook of Organizational Justice. Erlbaum, Mahwah, NJ, pp. 273–300.
Walster, E., Walster, G.W., Berscheid, E., 1978. Equity. Theory and Research. Allyn and Bacon, Boston.

Juvenile Justice: International Law and Children's Rights

Barry Goldson, The University of Liverpool, Liverpool, UK
John Muncie, The Open University, Milton Keynes, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a replacement of the previous edition article by G. van Bueren, volume 12, pp. 8067–8069, © 2001, Elsevier Ltd.

Abstract

Since the early twentieth century, most young offender legislation worldwide has been formulated on the basis that children and young people who find themselves in conflict with the law should be dealt with separately from adults and protected from the full weight of the criminal law. These young offender laws are based on an assumption that young people under a certain age are *doli incapax* (deemed incapable of forming the intent to commit a crime) and, accordingly, cannot be held fully responsible for their actions. A commitment to act in a 'child's best interests' and to protect and promote children's human rights are foundational principles for all systems of juvenile justice. A key driver of this phenomenon is the United Nations Convention on the Rights of the Child (UNCRC), formally adopted by the United Nations General Assembly in 1989 and subsequently ratified by 193 countries. Similarly, representative are the authoritative rights-informed guidelines on 'child-friendly justice' that have been ratified by the Council of Europe. A key question, however, is whether and to what extent individual nation-states actually comply with their 'child-friendly justice' obligations. Indeed, there remains a clear disjuncture between the promise of international rights discourse on the one hand and the limitations of territorial jurisdictional implementation on the other.

Universal Human Rights Instruments and Children

The formalization of universal human rights was consolidated with the creation of the United Nations and the adoption of the Universal Declaration of Human Rights in 1948. Subsequently, the United Nations General Assembly adopted five further pivotal human rights treaties: the International Convention on the Elimination of All Forms of Racial Discrimination (1965); the International Covenant on Economic, Social and Cultural Rights (1966); the International Covenant on Civil and Political Rights (1966); the Convention on the Elimination of All Forms of Discrimination against Women (1979); and the Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (1984). Each and all of these treaties apply to children, young people, and adults, but it was not until 1989 when the United Nations Convention on the Rights of the Child (UNCRC) was adopted by the United Nations General Assembly, and not until 1990 when the Convention came into force, that a universal human rights instrument focused exclusively and comprehensively on protecting and promoting children's particular interests.

Article 1 of the UNCRC provides that the term 'child' refers to 'every human being below the age of 18 years.' The Convention comprises 54 articles bringing together children's economic, social, cultural, civil, and political rights. General measures incorporated within the UNCRC include a fundamental obligation on governments (referred to as 'States Parties') to develop and sustain a children's human rights infrastructure within their jurisdictional spheres comprising, for example, the right to nondiscrimination (Article 2); the primacy of the child's best interests (Article 3); the right to life and maximum development (Article 6); and the right of children and young people to have their views given due weight in all matters affecting them (Article 12). The UNCRC also provides a range of 'civil rights' including the child's right to freedom of expression and association; the right to receive information; and the right to protection from all forms of violence, abuse, neglect, and mistreatment. The Convention further provides for every child's right to an adequate standard of living and the right to the best possible health care and educational services.

The Universal Declaration of Human Rights (1948), the UNCRC (1989), and each of the intervening human rights instruments have a bearing on juvenile justice law, policy, and practice, be it direct or indirect. Economic, social, and cultural rights; civil and political rights; the elimination of all forms of discrimination; safeguards against torture and other cruel, inhuman or degrading treatment or punishment; protection from violence, abuse, neglect, and mistreatment; a recognition of the 'special status' of childhood; 'best interest' principles; the right to life and maximum development; the right to be informed and the right to be heard, all have salience regarding the treatment of children in conflict with the law. Furthermore, specific and identifiable provisions of these instruments, together with an additional range of more 'specialist' global and European human rights standards, treaties, rules, conventions, and guidelines, relate more explicitly still to juvenile justice.

Human Rights Instruments and Juvenile Justice

International human rights provisions with regard to juvenile justice were initially formulated via three key instruments.

First, the United Nations Standard Minimum Rules for the Administration of Juvenile Justice (the 'Beijing Rules') were adopted by the United Nations General Assembly in 1985. The 'Rules' provide guidance for the protection of children's human rights in the development of separate and specialist juvenile justice systems. They were a direct response to a call made by the 'Sixth United Nations Congress on the Prevention of Crime and the Treatment of Offenders' that convened in 1980. Rule 4.1 provides: "juvenile justice shall be conceived as an integral part of the national development process of each country, within a comprehensive framework of social justice for all juveniles" (United Nations General Assembly, 1985).

Second, the United Nations Guidelines on the Prevention of Delinquency (the 'Riyadh Guidelines') were adopted by the United Nations General Assembly in 1990. The Guidelines are underpinned by diversionary and nonpunitive imperatives: "the successful prevention of juvenile delinquency requires efforts on the part of the entire society to ensure the harmonious development of adolescents" (para. 2); "formal agencies of social control should only be utilized as a means of last resort" (para. 5) and "no child or young person should be subjected to harsh or degrading correction or punishment measures at home, in schools or in any other institutions" (para. 54) (United Nations General Assembly, 1990a).

Third, the United Nations Rules for the Protection of Juveniles Deprived of their Liberty (the 'Havana Rules') were adopted by the United Nations General Assembly in 1990. The 'Rules' center a number of core principles including: deprivation of liberty should be a disposition of 'last resort' and used only 'for the minimum necessary period' and, in cases where children are deprived of their liberty, the principles, procedures, and safeguards provided by international human rights standards, treaties, rules, and conventions must be seen to apply (United Nations General Assembly, 1990b).

Furthermore, the core provisions contained within the 'Beijing Rules,' the 'Riyadh Guidelines,' and the 'Havana Rules' were bolstered, in 1990, when the UNCRC – the most widely ratified human rights instrument in history – came into force. The 'Articles' of the Convention that have most direct bearing on juvenile justice include the following:

- In all actions concerning children … the best interests of the child shall be a primary consideration (Article 3)
- States Parties (governments) should recognize the rights of the child to freedom of association and to freedom of peaceful assembly (Article 15)
- No child shall be subjected to arbitrary or unlawful interference with his or her privacy, family, home, or correspondence (Article 16)
- No child shall be subjected to torture or other cruel, inhuman, or degrading treatment or punishment (Article 37a)
- No child shall be deprived of his or her liberty unlawfully or arbitrarily. The arrest, detention, or imprisonment of a child shall be in conformity with the law and shall be used only as a measure of last resort and for the shortest appropriate period of time (Article 37b)
- Every child deprived of liberty shall be treated with humanity and respect for the inherent dignity of the human person and in a manner that takes into account the needs of persons of his or her age. In particular, every child deprived of liberty shall be separated from adults unless it is considered in the child's best interest not to do so (Article 37c)
- Every child deprived of his or her liberty shall have the right to prompt access to legal and other appropriate assistance, as well as the right to challenge the legality of the deprivation of his or her liberty before a court or other competent, independent and impartial authority, and to a prompt decision on any such action (Article 37d)
- States Parties recognize the right of every child alleged as, accused of, or recognized as having infringed the penal law to be treated in a manner consistent with the promotion of the child's sense of dignity and worth, which reinforces the child's respect for the human rights and fundamental freedoms of others and which takes into account the child's age and the desirability of promoting the child's reintegration and the child's assuming a constructive role in society (Article 40(1))
- States Parties shall seek to promote the establishment of laws, procedures, authorities, and institutions specifically applicable to children alleged as, accused of, or recognized as having infringed the penal law and in particular: (1) The establishment of a minimum age below which children shall be presumed not to have the capacity to infringe the penal law; (2) Whenever appropriate and desirable, measures for dealing with such children without resorting to judicial proceedings, providing that human rights and legal safeguards are fully respected (Article 40(3)) (United Nations General Assembly, 1989).

The United Nations human rights instruments have been further reiterated internationally by the Vienna Guidelines (Office of the United Nations High Commissioner for Human Rights, 1997) and, within the European context, by a movement toward 'child-friendly justice' driven by the Council of Europe. By extending the human rights principles that inform the 'European Rules for Juvenile Offenders Subject to Sanctions or Measures' (Council of Europe, 2009), the Committee of Ministers has formally adopted specific 'Guidelines for Child Friendly Justice' (Council of Europe, 2010). The 'guidelines' echo the UNCRC in stating that "a 'child' means any person under the age of 18 years" (Council of Europe, 2010: Section II(A)) and they apply "to all ways in which children are likely to be, for whatever reason and in whatever capacity, brought into contact with … bodies and services involved in implementing criminal, civil or administrative law" (Council of Europe, 2010: Section I: para. 2). The Council of Europe emphasizes the 'unifying' human rights objective of the 'guidelines' by explaining that they are intended to

> guarantee the respect and the effective implementation of all children's rights at the highest attainable level … giving due consideration to the child's level of maturity and understanding and the circumstances of the case … [Child friendly] justice is accessible, age appropriate, speedy, diligent, adapted to and focused on the needs and rights of the child, respecting the rights of the child.
> (Council of Europe, 2010: Section II(C))

It follows that the European 'guidelines for child-friendly justice' serve to underline many of the provisions of global human rights instruments including: "the minimum age of criminal responsibility should not be too low and should be determined by law"; "alternatives to judicial proceedings … should be encouraged whenever these may best serve the child's best interests"; "respect for children's rights as described in these guidelines and in all relevant legal instruments on the rights of the child should be guaranteed to the same extent in both in-court and out-of-court proceedings" (Council of Europe, 2010: Section IV(B): paras 23–26); and "any form of deprivation of liberty of children should be a measure of last resort and be for the shortest appropriate period of time" (Council of Europe, 2010: Section IV(A): para. 19).

Collectively, therefore, the United Nations and the Council of Europe human rights standards, treaties, rules, conventions,

and guidelines provide what is now a well-established 'unifying framework' for encouraging domestic legislative reform, modeling juvenile justice statutes, formulating policy, and developing practice in all nation-states to which they apply (Goldson and Hughes, 2010; Hamilton, 2011). As such – at face value at least – it might legitimately be argued that the same instruments provide the basis for 'globalized' human rights–compliant and 'child-friendly' juvenile justice. After all, with the notable exceptions of the United States and Somalia, each of the United Nations member states – 193 'States Parties' in total – have formally ratified the UNCRC and committed themselves to its implementation.

Restorative Justice

Key elements in 'acting in a child's best interest' and in the implementation of 'alternatives to judicial processing' have been encapsulated within the notion of restorative justice. Advocates maintain that various forms of conferencing (particularly 'family group conferencing'), sentencing circles, victim–offender mediation, reparation, and restitution provide greater potential to reintegrate the 'offender' back into the community. Both the Council of Europe and the United Nations have offered support for restorative justice and, in the juvenile justice sphere, located it within the context of 'best interests' principles and human rights discourses.

In 2002, the United Nation's Economic and Social Council formulated a number of foundational universal principles relating to restorative justice including noncoercive 'offender' and 'victim' participation, voluntarism, and confidentiality. These principles were integrated within human rights–based procedural safeguards (United Nations, 2002: paras 6–11).

Three years later, the Eleventh United Nations Congress on Crime Prevention and Criminal Justice (United Nations, 2005: p. 3) undertook to 'enhance restorative justice' and urged Member States to recognize the importance of further developing restorative justice policies, procedures, and programs. Similarly, the Council of Europe has recommended that restorative justice approaches should be made available to cover all stages of the juvenile justice process. It further recommends that such approaches should form autonomous mechanisms for conflict resolution and operate independently from conventionally formal means of judicial processing. The Committee of Ministers has stated that:

> Alternatives to judicial proceedings such as mediation, diversion (of judicial mechanisms) and alternative dispute resolution should be encouraged whenever these may best serve the child's best interests. The preliminary use of such alternatives should not be used as an obstacle to the child's access to justice.
> (Council of Europe, 2010: para. 24, our emphasis)

It follows that restorative justice has proliferated globally. Although such proliferation has inevitably given rise to multiple and competing definitions – alongside a diverse breadth of operational practices – the notions of 'reconciliation,' 'redress,' and 'restoration' remain the primary legitimizing logics.

A critical reading of restorative justice raises concerns that its free-ranging and discretionary nature may be at odds with human rights imperatives in the juvenile/youth justice sphere, including principles of proportionality, due process, 'best interests,' and the protection of children in conflict with the law from excessive incursion and unaccountable intervention. Rarely, if ever, do restorative justice approaches replace the conventional apparatus of formal adversarial 'justice'; rather, it is not uncommon for 'restorative' interventions to be reserved for low-level child 'offenders' and/or the readily compliant, while more retributive punitivism is visited upon the 'heavy-enders' and/or the recalcitrant. Rather than serving to replace conventional retributive applications of juvenile justice with 'child friendly' and human rights–compliant variants, restorative justice may serve to facilitate net widening, system expansion, disproportionality, and diversified (but interdependent) technologies of criminalization. In these ways, ostensibly 'informal' and 'restorative' processes might do more to relegitimize rather than challenge formal adversarial systems of juvenile justice (Goldson and Muncie, 2012). It is possible that the 'inclusionary' mechanisms of restoration may allow 'States Parties' to claim adherence to 'best interests' principles and compliance with the UNCRC, while simultaneously accommodating coercive, responsibilizing, and unaccountable interventions on the one hand and failing to disturb the hegemony of retribution and punishment, on the other. This critical reading of the potential of restorative justice suggests its limits in the juvenile justice sphere.

Monitoring Compliance

In order to monitor the translation of the UNCRC into the realm of law, policy, and practice within the national borders of respective States Parties, an international monitoring body has been established. The United Nations Committee on the Rights of the Child comprises 18 democratically elected members drawn from the 193 States Parties. The Committee normally meets three times a year in Geneva, Switzerland, and each of its members is required to act in an independent capacity as distinct from being a 'representative' of their specific country. The Committee has two principal functions: first, to issue 'General Comments' in order to elaborate the means by which the provisions and requirements of the UNCRC should be applied within specific subject domains; second, to periodically investigate the degree to which each State Party is implementing the Convention – and retaining compliance with it – within its corpus of law, policy, and practice. On both counts, the Committee has identified institutionalized obstructions to the implementation of the UNCRC in general and, more specifically, serious breaches and violations of the human rights of children within particular juvenile justice systems.

The 'General Comment' in respect of juvenile justice (United Nations Committee on the Rights of the Child, 2007) emphasizes the principle that jurisdictional systems must seek to safeguard the child's dignity at every point of the justice process and stresses that recourse to judicial proceedings (and especially custodial detention) should only ever be operationalized as measures of last resort. In short, the 'Comment' provides that the primary aims of human rights–compliant

juvenile justice are to meet the child's needs and to protect their best interests. In practice, however, the Committee concludes that implementation of the UNCRC is often piecemeal and that the human rights obligations of States Parties frequently appear as little more than afterthoughts within juvenile justice reform:

> …many States parties still have a long way to go in achieving full compliance with CRC, e.g., in the areas of procedural rights, the development and implementation of measures for dealing with children in conflict with the law without resorting to judicial proceedings and the use of deprivation of liberty only as a measure of last resort … The Committee is equally concerned about the lack of information on the measures that States parties have taken to prevent children from coming into conflict with the law. This may be the result of a lack of a comprehensive policy for the field of juvenile justice. This may also explain why many States parties are providing only very limited statistical data on the treatment of children in conflict with the law.
> (United Nations Committee on the Rights of the Child, 2007: para. 1)

More recently and more significantly, the Committee has detected regressive movement in the form of:

> …a growing tendency in both developed and developing regions of the world to consider — or even reform earlier progressive legislation on juvenile criminal justice — by lowering the age of criminal responsibility and increasing penalties for children found guilty, in a misguided effort to reduce increasing public insecurity and, as a result, weakening the realization of children's rights.
> (United Nations General Assembly, 2012: para. 38)

Concerns regarding piecemeal application or, worse still, regression in the implementation of international human rights standards, stem – at least in part – from the fact that the UNCRC is ultimately permissive and breach attracts no formal sanction. In this sense, Abramson (2006) finds that it may be the most ratified of all international human rights instruments but it also appears to be the most violated, particularly with regard to juvenile justice and that, moreover, such violations occur within a context of relative impunity. While presenting an otherwise positive assessment of the practical impact of the UNCRC in 'transforming the world of children and adolescents,' Abramson (2006) argues that juvenile justice is essentially peripheralized and/or unduly disregarded, even to the point of being 'unwanted.'

Jurisdiction-Specific Noncompliance

Two years after a State Party ratifies the UNCRC, it is obliged to submit an initial report to the United Nations Committee on the Rights of the Child – outlining how it is applying the UNCRC. Following the submission of the initial report, each State Party is required subsequently to provide periodic reports at 5-year intervals. The Committee also considers written evidence submitted by discreet government departments, nongovernmental organizations, national independent human rights institutions (such as Children's Commissioners and Ombudspersons), and children and young people themselves. The principal purpose of such inquiry is to ascertain the extent to which law, policy, and practice within individual nation-states is UNCRC compliant and the degree to which children are treated in accordance with the spirit, if not the word, of the Convention (Kilkelly, 2008; Goldson and Kilkelly, 2013).

With regard to the 'administration of juvenile justice,' the United Nations Committee on the Rights of the Child has repeatedly reported violations of children's human rights. The recurring concerns raised by the Committee particularly relate to Articles 37, 39, and 40 of the UNCRC (see above). Indeed, despite having had over 20 years to move toward full implementation, most states appear to have failed to integrate and embed the Convention within juvenile justice law, policy, and practice. (The Committee's reports on individual countries can be accessed on http://www2.ohchr.org/english/bodies/crc/.) For example, in Western Europe, Austria, Finland, Ireland, Germany, Portugal, Switzerland, and the United Kingdom have each been specifically criticized for failing to separate children from adults in custody and/or for facilitating easier movement between adult and juvenile systems owing to diminishing distinctions between the two (as is characteristic of controversial processes of 'juvenile transfer' or 'juvenile waiver' that allow for the prosecution of children in adult courts in the United States). Similarly, the United Nations Committee report of 2004 on Germany condemned the increasing number of children placed in detention – especially children of foreign origin – including the custodial detention of children with persons up to the age of 25 years. The report on the Netherlands (2004) expressed concern that custody was no longer being used as a last resort, whereas in its report of the same year on France the Committee expressed concern over legislation and practice that tends to favor repressive over educational measures, increases in the numbers of children in prison and the resulting worsening of conditions. The Committee has long been critical of the low age of criminal responsibility adopted in the three UK jurisdictions. It also lists a number of additional concerns with regard to UK jurisdictions, including the inappropriate use of punitive preemptive interventions (such as 'antisocial behavior orders') and the excessive use of custody (Goldson and Muncie, 2006, 2012; Muncie, 2008).

Racialized (In)justice

The persistent recurrence of human rights violations – over time and across space – is compounded by a manifest racialization of juvenile justice practice. Muncie (2008) reported that of 18 Western European jurisdictions studied, 15 were explicitly exposed to critique by the United Nations Committee for negatively discriminating against children from minority ethnic communities and migrant children seeking asylum. The overrepresentation of such children is particularly conspicuous at the polar ends of the system – arrest and penal detention. This especially appears to be the case for the Roma and traveler communities in England and Wales, Finland, France, Germany, Greece, Ireland, Italy, Northern Ireland, Portugal, Scotland, Spain, and Switzerland; for Moroccan and Surinamese children in the Netherlands; and for other North African children in Belgium and Denmark.

Trends across Europe, as elsewhere in the world, appear to indicate that the regulation and governance of 'urban marginality' and poverty is being increasingly prized away from the social welfare apparatus and redefined as 'crime problems' within burgeoning 'penal states' (Wacquant, 2009). Moreover, within such shifts, minority ethnic communities and immigrant groups are bearing the brunt of a 'punitive upsurge' (Wacquant, 2009). In particular, 8 million Roma – the largest minority ethnic group in the European Union – are widely reported as enduring systematic discrimination, harassment, ghettoization, forced eviction, expulsion, and detention. By collating data drawn from 22 country-specific reports, Gauci (2009: p. 6) notes: "most … reports identify the Roma … as being particularly vulnerable to racism and discrimination … in virtually all areas of life." Increased ghettoization of 'foreigner' and Roma communities in various European countries, whether as a result of institutional decisions or practical realities such as chronic unemployment, is serving to consolidate structural exclusion and systematic marginalization: "the creation of spatial segregation and socially excluded localities where communities are effectively denied access to basic services such as water and electricity" (Gauci, 2009: p. 10).

Of course, such phenomena are not restricted to Europe. Cunneen and White (2006) note similar processes of racialized justice in Australia, for example, in particular a persistent overrepresentation of indigenous young people held in juvenile justice detention, despite an apparent favoring of diversion and community-based sentences for other offenders. In the United States, African-American and Hispanic children and young people are markedly overpenalized (Miller, 1996; Acoca, 1999). Indeed, some of the most punitive elements of juvenile justice worldwide appear to be increasingly reserved for, and applied to, children from minority ethnic and/or immigrant populations.

Violence against Children

In February 2003, Paulo Sergio Pinheiro was appointed, at Assistant Secretary-General level, to direct the 'United Nations Secretary-General's Study on Violence Against Children' (the Study). The Study was undertaken in collaboration with the Office of the High Commissioner for Human Rights, the United Nations Children's Fund, and the World Health Organization. It comprises the most wide-ranging and detailed analysis of its type in history. The final report – published and presented to the United Nations General Assembly in November 2006 (Pinheiro, 2006) – concluded that

> Millions of children, particularly boys, spend substantial periods of their lives under the control and supervision of care authorities or justice systems and in institutions such as orphanages, children's homes, care homes, police lock-ups, prisons, juvenile detention facilities and reform schools. These children are at risk of violence from staff and officials responsible for their well-being. Corporal punishment in institutions is not explicitly prohibited in a majority of countries. Overcrowding and squalid conditions, societal stigmatization and discrimination and poorly trained staff, heighten the risk of violence. Effective complaints, monitoring and inspection mechanisms, and adequate government regulation and oversight are frequently absent. Not all perpetrators are held accountable, creating a culture of impunity and tolerance of violence against children.
>
> (Pinheiro, 2006: p. 16)

More specifically, throughout the course of the United Nations Study, a series of 'thematic consultations' – involving leading international experts – convened in order to provide periodic subject-specific reports. In the 'Violence Against Children in Conflict with the Law' report, the Director of the Study observed that "children in conflict with the law … are one of the *most vulnerable* groups to the worst forms of violence" (Pinheiro, 2006: pp. 17–18, emphases added). The juxtaposition of universal human rights discourse and international recognition of state responsibilities to safeguard children on the one hand, and the pervasive violation of children by the States Parties themselves – particularly those in conflict with the law and, more still, child prisoners – on the other is, to say the least, anomalous (Goldson, 2009). Such anomaly echoes Abramson's (2000) earlier analysis of the implementation of the UNCRC within juvenile justice systems in 141 countries that revealed widespread absence of 'sympathetic understanding.' He argued that a complete overhaul of juvenile justice is required in many countries where a range of violations are evident including inadequate or nonexistent training of judges, police, and prison authorities; no (or limited) access to legal assistance, advocacy and/or representation; delayed trials; disproportionate sentences; insufficient respect for the rule of law; incidence of police brutality; and improper use of the juvenile justice system to address other social problems. Particular concerns center on penal detention including failure to develop alternatives to incarceration; overcrowding and poor conditions in custodial facilities; limited prospects of rehabilitation; infrequent contact between child prisoners and their families; lack of separation between child and adult prisoners; inhumane treatment; and, at the extremes, torture (Goldson and Kilkelly, 2013). Beyond this is the question of child execution.

Age of Criminal Responsibility: Corporal Punishment, Life Imprisonment, and Execution

Although there is no categorical international standard regarding the age at which criminal responsibility can reasonably be imputed on a child, the provisions of a number of human rights instruments are pertinent. For example, Article 4(1) of the United Nations Standard Minimum Rules for the Administration of Juvenile Justice (the Beijing Rules) provides: "in those legal systems recognising the concept of the age of criminal responsibility for juveniles, the beginning of that age shall not be fixed at too low an age level." Furthermore, the official commentary on this provision states: "in general there is a close relationship between the notion of responsibility for delinquent or criminal behavior and other social rights and responsibilities (such as marital status, civil majority, etc)." In this way, the Beijing Rules advocate consistency and integrity

within and across the wider corpus of criminal and civil law providing for 'social rights and responsibilities' within a given jurisdiction.

However, the age of criminal responsibility differs widely around the world. In the United States, the minimum age is set at just 6 years (in North Carolina) but most states have set no minimum age at all (Cipriani, 2009). In some Islamic societies, such as Iran, the age of criminal responsibility is linked to the age of maturity or puberty that, according to Sharia law, is 9 years for girls and 15 years for boys. Across the United Kingdom, the minimum ages of criminal responsibility – among the lowest in the European Union – vary from 8 years in Scotland (although from 2010, children under the age of 12 years cannot be prosecuted in a criminal court), to 10 years in England, Wales, and Northern Ireland. Even greater variance is evident across Europe where children up to the age of 14 years (Germany, Italy, and Spain), 15 years (Denmark, Sweden, Norway, and Finland), or 18 years (Belgium and Luxembourg) are deemed to lack full criminal responsibility and, as a result, tend to be processed in civil tribunals rather than criminal courts (Muncie, 2009). Such differences reveal both the extent to which compliance with international standards – regarding the criminalization of children – fluctuates between nation-states and, more problematically, the manner in which some jurisdictions flagrantly flout their human rights obligations. Indeed, from the time of its establishment until February 2009, the Committee on the Rights of the Child has had to recommend, on no less than 65 occasions, that a State Party either increase the minimum age of criminal responsibility or institute a minimum age of criminal responsibility where one did not exist at all (Hamilton, 2011).

The low ages of criminal responsibility that are widely applied in many countries of the world raise serious implications especially with regard to sentences of life imprisonment. In many states in the United States, for example, some offences committed by children have attracted a mandatory minimum sentence of life imprisonment without the possibility of parole. In 2012, the effect of the Supreme Court's ruling that such practices violated the Constitution's prohibition of 'cruel and unusual punishment' was restricted to making the life without parole disposition optional, rather than mandatory (*Miller v. Alabama*, 2012). Furthermore, in 'reviewing the laws of the Commonwealth States with regards to life imprisonment of children' – including 'life imprisonment without parole,' 'life imprisonment with the possibility of parole,' 'detention at the pleasure of the executive or the courts,' and/or 'indefinite detention sentences' – Ratledge (2012) reports that "45 out of 54 Commonwealth States provide for one or more of the types of life imprisonment." Moreover, in at least 40 countries of the world – almost all of which have ratified the UNCRC – children can still be sentenced to whipping, flogging, caning, or amputation and, in at least seven countries, child offenders can lawfully be sentenced to death by lethal injection, hanging, shooting, or stoning (Child Rights Information Network, 2012). At the most extreme end of the continuum of human rights violations, between 1990 and 2011 Amnesty International (2012: n.p.) has documented 87 executions of child offenders in nine countries observing that:

> The use of the death penalty for crimes committed by people younger than 18 is prohibited under international human rights law, yet some countries still execute child offenders. Such executions are few compared to the total number of executions in the world. Their significance goes beyond their number and calls into question the commitment of the executing states to respect international law.

Children's Rights: Contingent and Conditional

The practical translation of universal children's human rights standards and principles is, at least in part, culturally relative. It is also contingent upon social and economic materiality. A persistent problem is how human rights might be equally distributed within a world that is profoundly divided and polarized by social and economic inequalities. The United Nations Children's Fund (UNICEF, 2010) has revealed that even in rich nations, identifiable groups of children are unnecessarily 'left behind,' subjected to poverty, denied access to 'well-being,' and exposed to inequality. Worldwide, more than 1 billion children lack proper nutrition, safe drinking water, decent sanitation, health care services, shelter, and/or education and every day, 28 000 children die from poverty-related causes (Goldson and Muncie, 2006). According to Save the Children UK (2007), children are treated as commodities across all continents in myriad ways including child trafficking, child prostitution, bonded child labor, child slavery, and child soldiers. Whether international bodies and standards of juvenile justice can help in the efforts to bring about much needed changes in conditions, attitudes, and practices remains to be seen.

See also: Children and Law; Civil Law; Civil Liberties and Human Rights; Crime Policy in the United States; Critical Legal Studies; Death Penalty; Discrimination and the Law; Due Process in Law; Equality and Inequality: Legal Aspects; Globalization: Legal Aspects; Human Rights: Legal Aspects; Indigenous Rights; Injustice: Legal Aspects; Justice and Law; Justice, Access To; Law, Sociology of; Law: The Socio-Legal Perspective; Legal Systems, Classification of; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Punishment: Social and Legal Aspects; Race and the Law; Rights: Legal Aspects.

Bibliography

Abramson, B., 2000. Juvenile Justice: The 'Unwanted Child' of State Responsibilities. An Analysis of the Concluding Observations of the UN Committee on the Rights of the Child, in Regard to Juvenile Justice from 1993 to 2000. International Network on Juvenile Justice/Defence for Children International, Brussels.

Abramson, B., 2006. Juvenile justice: the unwanted child. In: Jensen, E., Jepsen, J. (Eds.), Juvenile Law Violators, Human Rights and the Development of New Juvenile Justice Systems. Hart, Oxford.

Acoca, L., 1999. Investing in girls: a 21st century strategy. Journal of the Office of Juvenile Justice and Delinquency Prevention VI (1), 3–13.

Amnesty International, 2012. Executions of Juveniles since 1990. Available at: http://www.amnesty.org/en/death-penalty/executions-of-child-offenders-since-1990.

Cipriani, D., 2009. Children's Rights and the Minimum Age of Criminal Responsibility: A Global Perspective. Ashgate, Farnham.

Child Rights Information Network, 2012. Ending Inhuman Sentencing of Children. Available at: http://www.crin.org/violence/campaigns/sentencing/.

Council of Europe, 2009. European Rules for Juvenile Offenders Subject to Sanctions or Measures. Council of Europe Publishing, Strasbourg.

Council of Europe, 2010. Guidelines of the Committee of Ministers of the Council of Europe on Child Friendly Justice (Adopted by the Committee of Ministers on 17 November 2010 at the 1098th meeting of the Ministers' Deputies). Council of Europe, Strasbourg.

Cunneen, C., White, R., 2006. Australia: containment or empowerment? In: Muncie, J., Goldson, B. (Eds.), Comparative Youth Justice. Sage, London.

Gauci, J.-P., 2009. Racism in Europe. European Network Against Racism (ENAR), Brussels.

Goldson, B., 2009. Child incarceration: institutional abuse, the violent state and the politics of impunity. In: Scraton, P., McCulloch, J. (Eds.), The Violence of Incarceration. Routledge, London.

Goldson, B., Hughes, G., 2010. Sociological criminology and youth justice: comparative policy analysis and academic intervention. Criminology and Criminal Justice 10 (2), 211–230.

Goldson, B., Kilkelly, U., 2013. International human rights standards and child imprisonment: potentialities and limitations. International Journal of Children's Rights 21 (2), 345–371.

Goldson, B., Muncie, J., 2006. Rethinking youth justice: comparative analysis, international human rights and research evidence. Youth Justice: An International Journal 6 (2), 91–106.

Goldson, B., Muncie, J., 2012. Towards a global 'child friendly' juvenile justice? International Journal of Law, Crime and Justice 40 (1), 47–64.

Hamilton, C., 2011. Guidance for Legislative Reform on Juvenile Justice. Children's Legal Centre and UNICEF, New York.

Kilkelly, U., 2008. Youth justice and children's rights: measuring compliance with international standards. Youth Justice: An International Journal 8 (3), 187–192.

Miller v. Alabama, 132 S. Ct. 2455 (2012). Available at: http://www.supremecourt.gov/opinions/11pdf/10-9646.pdf.

Miller, J., 1996. Search and Destroy: African-American Males in the Criminal Justice System. Cambridge University Press, Cambridge.

Muncie, J., 2008. The punitive turn in juvenile justice: cultures of control and rights compliance in western Europe and the USA. Youth Justice: An International Journal 8 (2), 107–121.

Muncie, J., 2009. Youth and Crime, third ed. Sage, London.

Office of the United Nations High Commissioner for Human Rights, 1997. Guidelines for Action on Children in the Criminal Justice System (The 'Vienna Guidelines') Recommended by Economic and Social Council Resolution 1997/30 of 21 July 1997. United Nations, New York.

Pinheiro, P.S., 2006. World Report on Violence against Children. United Nations, Geneva.

Ratledge, L., 2012. Inhuman Sentencing: Life Imprisonment of Children in the Commonwealth. Child Rights Information Network – Available at: http://www.crin.org/docs/Inhuman_Sentencing.pdf.

Save the Children UK, 2007. The Small Hands of Slavery. Save the Children UK, London.

UNICEF, 2010. The Children Left behind: A League Table of Inequality in Child Well-being in the World's Rich Countries. Innocenti Report Card 9. Unicef Innocenti Research Centre, Florence.

United Nations, 2002. ECOSOC Resolution 2002/12: Basic Principles on the Use of Restorative Justice Programmes in Criminal Matters. United Nations, New York.

United Nations, 2005. United Nations A/Conf.203/1: Eleventh United Nations Congress on Crime Prevention and Criminal Justice, Bangkok 18–25 April. United Nations, New York.

United Nations Committee on the Rights of the Child, 2007. General Comment No. 10: Children's Rights in Juvenile Justice, Forty-fourth Session, 15 January–2 February. Office of the High Commissioner for Human Rights, Geneva.

United Nations General Assembly, 1985. United Nations Standard Minimum Rules for the Administration of Juvenile Justice. United Nations, New York.

United Nations General Assembly, 1989. United Nations Convention on the Rights of the Child. United Nations, New York.

United Nations General Assembly, 1990a. United Nations Guidelines for the Prevention of Juvenile Delinquency. United Nations, New York.

United Nations General Assembly, 1990b. United Nations Rules for the Protection of Juveniles Deprived of Their Liberty. United Nations, New York.

United Nations General Assembly, 2012. Report of the Committee on the Rights of the Child. United Nations, New York.

Wacquant, L., 2009. Punishing the Poor: The Neo-Liberal Government of Social Insecurity. Duke University Press, Durham, NC.

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 13

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser—Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 13

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 13 *xliii*

VOLUME 13

Kant, Immanuel (1724–1804) 1
Volker Gerhardt

Kantian Ethics and Politics 5
Matthias Kaufmann

Key, Valdimer Orlando (1908–63) 11
Walter D Burnham

Keynes, John Maynard (1883–1946) 14
Robert Skidelsky

Keynesianism and the Question of State Interventionism 20
Robert Salais

Kimura, Motoo (1924–94) 24
James F Crow

Kin Selection 26
Andy Gardner

Kindling 32
Gordon C Teskey and Michael E Corcoran

Kinship in Anthropology 36
Joanna Overing, Paolo Fortis, and Margherita Margiotti

Kinship, Evolution of 44
Drew Gerkey

Kinship, Formal Models of 53

Dwight Read

Kinship Terminology 61
Dwight Read

Klein, Melanie (1882–1960) 67
Nancy J Chodorow

Knowledge, Anthropology of 70
Dale F Eickelman

Knowledge (Explicit, Implicit and Tacit): Philosophical Aspects 74
Martin Davies

Knowledge, Geography of 91
Peter Meusburger

Knowledge Representation 98
Ernest Davis

Knowledge Society, History of 105
Nico Stehr

Knowledge, Sociology of 111
Volker Meja

Kohlberg, Lawrence (1927–87) 119
Gertrud Nunner-Winkler

Köhler, Wolfgang (1887–1967) 123
Viktor Sarris

Kraepelin, Emil (1856–1926) 127
Paul Hoff

Kuhn, Thomas S. (1922–96) 133
Paul Hoyningen-Huene and Simon Lohse

Kula Ring, Anthropology of 139
Frederick H Damon

Kuznets Curves 144
Tommaso Luzzati

Labeling Theory, History of 150
Bernard B Berk

Labor, Division of 156
Wolfgang Littek

Labor, Geography of 163
Hege M Knutsen, Sylvi B Endresen, Ann C Bergene, and David Jordhus-Lier

Labor History 169
Marcel van der Linden

Labor Market Institutions and Labor Market Outcomes in Transition Economies 174
Hartmut Lehmann and Alexander Muravyev

Labor Market Policy 183
Daniel Clegg

Labor Supply 190
Bradley T Heim

Labor Unions 199
Wolfgang Streeck

Laboratory Studies: Historical Perspectives 205
Park Doing

Lacan, Jacques (1901–81) 211
August Ruhs and Jeanne Wolff Bernstein

Laicization, History of 216
Jean Baubérot

Land Change: The Merger of Land Cover and Land use Dynamics 220
William J McConnell

Landscape 224
Kenneth R Olwig

Land Tenure 231
Parker Shipton and Daivi Rodima-Taylor

Land-Use Indicators 238
Karl-Heinz Erb

Language Acquisition 245
Brian MacWhinney

Language and Ethnicity 251
Günther Schlee

Language and Gender 256
Sarah Gormley

Language and Labor Market Success 260
Ingo E Isphording

Language and Literature 266
Nigel Fabb

Language and Poetic Structure 272
Kristin Hanson

Language and Social Class 281
Suzanne Romaine

Language and Society 288
Cliff Goddard and Anna Wierzbicka

Language and Thought: The Neo-Whorfian Hypothesis 293
Antoni Gomila

Language Contact 300
Brian D Joseph

Language Development, Theories of 307
David W Carroll and Eleni Pinnow

Language Development: Emergentist Theories 313
William O'Grady

Language Development: Evolutionary Theories 319
Mark Atkinson and Kenny Smith

Language Development: Influence of Socio-Economic Status 324
Erika Hoff and Krystal M Ribot

Language Disorders, Treatment and Remediation of 329
Dimitrios S Kasselimis and Constantin Potagas

Language Families, Archaeology and History of 337
Peter Bellwood

Language Learning Impairment 344
Paula Tallal and Sabine Heim

Language Policy 353
Ofelia García

Language Processes in Delirium and Dementia 360
Joël Macoir, Yves Turgeon, and Robert Laforce, Jr

Language Processing, Functional Magnetic Resonance Imaging of 368
Jie Yang and Steven L Small

Laplace, Pierre Simon, Marquis de (1749–1827) 381
Bernard Bru

Large-Scale Social Surveys, Analysis of 385
Elaine Zanutto and Andrew Gelman

Lashley, Karl Spencer (1890–1958) 391
Richard F Thompson

Latent Structure and Causal Variables 394
Peter Spirtes

Lateralization of Language as Demonstrated by Brain Imaging Procedures 398
Brigitte Stemmer

Latin-American Legal Systems 410
Javier Couso

Latour, Bruno (1947–) 414
Lisa Garforth

Law and Aging 420
Israel (Issi) Doron

Law and Courts in Authoritarian States 427
Peter H Solomon, Jr

Law and Democracy 435
Grażyna Skąpska

Law and Development 443
David M Trubek

Law and Economics 448
Steven Shavell

Law and Economics: Empirical Dimensions 454
Brian J Broughman and Kenneth G Dau-Schmidt

Law and Emotion 461
Susan A Bandes

Law and Everyday Life 468
Patricia Ewick

Law and Geography 474
Nicholas Blomley and Joshua Labove

Law and Morality: A Continental European Perspective 479
Robert Alexy

Law and Morality: An Analytical Perspective 483
Tom Campbell

Law and Neuroscience 489
Owen D Jones and Matthew Ginther

Law and People with Disabilities 497
Janet E Lord, Deepti S Raja, and Peter Blanck

Law and Risk 504
Jonathan R Nash

Law and Society 509
Lora M Levett and Adina M Thompson

Law and Society: Development of the Field 515
Bryant G Garth

Law and Technology 520
Ruth L Okediji

Law and Terrorism in the US and Beyond 528
William C Banks

Law: Anthropological Aspects 535
Sally E Merry and Matthew C Canfield

Law as an Instrument of Social Change 542
Catherine R Albiston and Gwendolyn M Leachman

Law as Constitutive 550
Roger Cotterrell

Law, Autopoiesis in 554
Ralf Rogowski

Law, Bioethics, and Biotechnology 557
Michele Goodwin and Allison M Whelan

Law: Change and Evolution 565
Lawrence M Friedman

Law: Economics of Its Public Enforcement 570
Alexander Volokh

Law Firms 578
David B Wilkins

Law: History of Its Relation to the Social Sciences 585
W Tim Murphy

Law, Mobilization of 590
Calvin Morrill, Mayra Feddersen, and Stephen Rushin

Law, Sociology of 598
Carol A Heimer and Jaimie Morse

Law: The Socio-Legal Perspective 604
Lauren Edelman and Marc Galanter

Laws, Conflict of 614
Symeon C Symeonides

Lawyers, Regulation of 619
Laurel S Terry

Lawyers: Social Organization of the Profession 628
Fiona M Kay

Lazarsfeld, Paul Felix (1901–76) 635
Christian Fleck

Leader–Member Exchange Theory 641
Berrin Erdogan and Talya N Bauer

Leadership 648
S Alexander Haslam, Stephen D Reicher, and Michael J Platow

Leadership Education: Theory and Practice 655
Robert Emmerling, Basak Canboy, Ricard Serlavos, and Joan MB Foguet

Leadership in Organizations, Sociology of 664
Manfred Kets de Vries and Alicia Cheak-Baillargeon

Learning and Instruction: Social-Cognitive Perspectives 670
Felice Carugati and Anne-Nelly Perret-Clermont

Learning and Memory, Neural Basis of 677
Hans J Markowitsch and Angelica Staniloiu

Learning Disability and Social Work 684
Robyn Munford and Garth Bennie

Learning: Mathematical Learning Theory, History of 690
Sandy Lovie and Pat Lovie

Learning: Organizational 695
Silvia Gherardi

Learning Theories and Educational Paradigms 699
Robert D Tennyson and Ashley Volk

Learning to Learn 712
Claire E Weinstein, Jaimie M Krause, Nancy Stano, Taylor W Acee, M Krause Jaimie, and Nancy Stano

Legal Bases of Social Work 720
Michael Preston-Shoot

Legal Culture and Legal Consciousness 726
Susan S Silbey

Legal Education 734
Kim Economides

Legal Formalism 740
Daniel A Farber

Legal Issues: Public Opinion 747
Robert J MacCoun and Karin D Martin

Legal Personality 753
Gregory A Mark

Legal Pluralism 757
John Griffiths

Legal Positivism 762
Brian Z Tamanaha

Legal Process and Social Science: United States 767
John Monahan

Legal Realism 772
John Henry Schlegel

Legal Reasoning and Argumentation 776
Geoffrey Samuel

Legal Reasoning: Computational Models 784
Henry Prakken

Legal Scholarship 792
Edward Rubin

Legal Systems, Classification of 799
David S Clark

Legal Systems: Private 807
Stuart Henry

Legal Traditions of Southeast Asia 812
Andrew J Harding

Legalization: The Expansion of Law's Domain 818
Jethro K Lieberman

Legitimacy, Sociology of 823
Cathryn Johnson and Lesley Watson

Leisure Activities Choices among Adolescents 829
Maria K Pavlova and Rainer K Silbereisen

Leisure and Cultural Consumption: The European Perspective 837
Philippe Coulangeon

Leisure and Cultural Consumption: US Perspective 845
John P Robinson and Timothy Triplett

Leisure, Sociology of 852
Ken Roberts

Lesbian and Gay Activism in Latin America 858
Jordi Díez

Lesbian, Gay, Bisexual and Transgender Aging 864
Sue Westwood

Lesbian, Gay, Bisexual and Trans-sexual Butch-Femme Subcultures 869
Julia A Puckett and Heidi M Levitt

Lesbian, Gay, Bisexual and Trans-sexual Drag Culture 875
Shae D Miller and Verta Taylor

Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview 879
Don Barrett

Lesbian, Gay, Bisexual and Trans-sexual: Commercial Sex 882
Jason H Prior and Penny Crofts

Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures 887
Francisco I Surace and Heidi M Levitt

Lesbians: Social and Economic Situation 893
MV Lee Badgett, Laura Durso, and Alyssa Schneebaum

Levels of Analysis in Social Psychology 899
Willem Doise and Joaquim Pires Valentim

Levi-Strauss, Claude (1908–2009) 904
Dominique Bouchet

Lewin, Kurt (1890–1947) 911
Carl-Friedrich Graumann

Lex Mercatoria 915
Christine Windbichler

Lexical Consolidation 919
Jakke Tamminen

Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects 925
Matthew A Lambon Ralph

Lexical Semantics 930
Dirk Geeraerts

Lexicology and Lexicography 937
Wolfgang Klein

Lexicon 942
James Pustejovsky

CONTRIBUTORS TO VOLUME 13

Taylor W. Acee
Texas State University — San Marcos, San Marcos, TX, USA

Catherine R. Albiston
University of California, Berkeley School of Law, Berkeley, CA, USA

Robert Alexy
Christian-Albrechts-Universität, Kiel, Germany

Mark Atkinson
Language Evolution and Computation Research Unit, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, UK

M.V. Lee Badgett
Center for Public Policy & Administration, University of Massachusetts Amherst, Amherst, MA, USA

Susan A. Bandes
DePaul University College of Law, Chicago, IL, USA

William C. Banks
Syracuse University, Syracuse, NY, USA

Don Barrett
California State University San Marcos, San Marcos, CA, USA

Jean Baubérot
École pratique des hautes études, Sorbonne, Paris, France

Talya N. Bauer
Portland State University, Portland, OR, USA

Peter Bellwood
Australian National University, Canberra, ACT, Australia

Garth Bennie
Ministry of Education, Palmerston North, New Zealand

Ann C. Bergene
Work Research Institute, Oslo, Norway

Bernard B. Berk
California State University, Los Angeles, CA, USA

Jeanne Wolff Bernstein
Psychoanalytic Institute of Northern California, San Francisco, CA, USA; Postdoctoral Psychoanalytic and Psychotherapy Program, New York University, New York, NY, USA; and Wien, Austria

Peter Blanck
Syracuse University, Syracuse, NY, USA

Nicholas Blomley
Simon Fraser University, Burnaby, BC, Canada

Dominique Bouchet
University of Southern Denmark, Denmark; and School of Creative Arts, James Cook University, Townsville, QLD, Australia

Brian J. Broughman
Maurer School of Law, Indiana University, Bloomington, IN, USA

Bernard Bru
Université René Descartes Paris V, France

Walter D. Burnham
The University of Texas, Austin, TX, USA

Tom Campbell
Charles Sturt University, Kingston, ACT, Australia

Basak Canboy
Ramon Llull University — ESADE Business School, Leadership Development Research Centre, Sant Cugat (Barcelona), Spain

Matthew C. Canfield
New York University, New York, NY, USA

David W. Carroll
University of Wisconsin-Superior, Superior, WI, USA

Felice Carugati
Alma Mater Studiorum Università di Bologna, Bologna, Italy

Alicia Cheak-Baillargeon
INSEAD Global Leadership Centre, Fontainebleau Cedex, France

Nancy J. Chodorow
University of California, Berkeley, CA, USA

David S. Clark
Willamette University College of Law, Salem, OR, USA

Daniel Clegg
School of Social and Political Science, Edinburgh, UK

Michael E. Corcoran
University of Saskatchewan, Saskatoon, SK, Canada

Roger Cotterrell
University of London, London, UK

Philippe Coulangeon
Observatoire Sociologique du changement (OSC), SciencesPo/CNRS, Paris, France

Javier Couso
Facultad de Derecho, Universidad Diego Portales, Santiago, Chile

Penny Crofts
University of Technology Sydney, Broadway, NSW, Australia

James F. Crow
University of Wisconsin, Madison, WI, USA

Frederick H. Damon
University of Virginia, Charlottesville, VA, USA

Kenneth G. Dau-Schmidt
Maurer School of Law, Indiana University, Bloomington, IN, USA

Martin Davies
University of Oxford, Oxford, UK

Ernest Davis
New York University, New York, NY, USA

Jordi Díez
University of Guelph, Guelph, ON, Canada

Park Doing
Cornell University, Ithaca, NY, USA

Willem Doise
Faculté de Psychologie et des Sciences de l'Éducation, Université de Genève, Genève, Switzerland

Israel (Issi) Doron
University of Haifa, Haifa, Israel

Laura Durso
Center for American Progress, Washington, DC, USA

Kim Economides
Flinders Law School, Flinders University, Adelaide, SA, Australia

Lauren Edelman
University of California - Berkeley, Berkeley, CA, USA

Dale F. Eickelman
Dartmouth College, Hanover, NH, USA

Robert Emmerling
Ramon Llull University — ESADE Business School, Leadership Development Research Centre, Sant Cugat (Barcelona), Spain

Sylvi B. Endresen
University of Oslo, Oslo, Norway

Karl-Heinz Erb
Institute of Social Ecology Vienna, Alpen-Adria Universität Klagenfurt, Vienna, Graz, Austria

Berrin Erdogan
Portland State University, Portland, OR, USA

Patricia Ewick
Clark University, Worcester, MA, USA

Nigel Fabb
School of Humanities, University of Strathclyde, Glasgow, UK

Daniel A. Farber
School of Law, University of California, Berkeley, CA, USA

Mayra Feddersen
University of California, Berkeley, CA, USA

Christian Fleck
University of Graz, Austria

Joan M.B. Foguet
Ramon Llull University — ESADE Business School, Leadership Development Research Centre, Sant Cugat (Barcelona), Spain

Paolo Fortis
Centre for Amerindian, University of St Andrews, St Andrews, UK

Lawrence M. Friedman
Stanford University, Stanford, CA, USA

Marc Galanter
University of Wisconsin - Madison, Madison, WI, USA

Ofelia García
Graduate Center, The City University of New York, New York, NY, USA

Andy Gardner
University of St Andrews, St Andrews, UK

Lisa Garforth
School of Geography, Politics and Sociology, Newcastle upon Tyne, UK

Bryant G. Garth
University of California, Irvine, CA, USA

Dirk Geeraerts
Faculty of Arts, KU Leuven, Leuven, Belgium

Andrew Gelman
Columbia University, New York, NY, USA

Volker Gerhardt
Humboldt University, Berlin, Germany

Drew Gerkey
University of Maryland, Annapolis, MD, USA

Silvia Gherardi
University of Trento, Via Verdi, Trento, Italy

Matthew Ginther
Candidate in Law and Neuroscience, Vanderbilt University, Nashville, TN, USA

Cliff Goddard
University of New England, Armidale, NSW, Australia

Antoni Gomila
Campus Carretera Valldemossa, Palma de Mallorca, Spain

Michele Goodwin
University of California — Irvine, Irvine, CA, USA

Sarah Gormley
University of Manchester, Manchester, UK; and University of Birmingham, Birmingham, UK

Carl-Friedrich Graumann
University of Heidelberg, Heidelberg, Germany

John Griffiths
University of Groningen, Groningen, The Netherlands

Kristin Hanson
University of California, Berkeley, CA, USA

Andrew J. Harding
National University of Singapore, Singapore

S. Alexander Haslam
School of Psychology, University of Queensland, QLD, Australia

Bradley T. Heim
Indiana University, Bloomington, IN, USA

Sabine Heim
Center for Molecular and Behavioral Neuroscience, Rutgers University-Newark, Newark, NJ, USA

Carol A. Heimer
Northwestern University, Evanston, IL, USA; and American Bar Foundation, Chicago, IL, USA

Stuart Henry
San Diego State University, San Diego, CA, USA

Erika Hoff
Florida Atlantic University, Davie, FL, USA

Paul Hoff
Psychiatric University Hospital Zurich, Zurich, Switzerland

Paul Hoyningen-Huene
Center for Philosophy and Ethics of Science, Institute of Philosophy, Leibniz University of Hannover, Hannover, Germany

Ingo E. Isphording
IZA — Institute for the Study of Labor, Bonn, Germany

M. Krause Jaimie
The University of Texas at Austin, Austin, TX, USA

Cathryn Johnson
Emory University, Atlanta, GA, USA

Owen D. Jones
New York Alumni Chancellor's Professor of Law & Professor of Biological Sciences, Vanderbilt University, Nashville, TN, USA; and Director, MacArthur Foundation Research Network on Law and Neuroscience

David Jordhus-Lier
University of Oslo, Oslo, Norway

Brian D. Joseph
The Ohio State University, Columbus, OH, USA

Dimitrios S. Kasselimis
University of Crete, Crete, Greece; and Eginition Hospital, National and Kapodistrian University of Athens, Athens, Greece

Matthias Kaufmann
Martin-Luther-Universität Halle-Wittenberg, Halle (Saale), Germany

Fiona M. Kay
Queen's University, Kingston, ON, Canada

Manfred Kets de Vries
INSEAD Global Leadership Centre, Fontainebleau Cedex, France

Wolfgang Klein
Max Planck Institut für Psycholinguistik, Nijmegen, The Netherlands

Hege M. Knutsen
University of Oslo, Oslo, Norway

Jaimie M. Krause
The University of Texas at Austin, Austin, TX, USA

Joshua Labove
Simon Fraser University, Burnaby, BC, Canada

Robert Laforce
Clinique Interdisciplinaire de Mémoire, Québec, QC, Canada

Matthew A. Lambon Ralph
School of Psychological Sciences, University of Manchester, Manchester, UK

Gwendolyn M. Leachman
Williams Institute, UCLA School of Law, Los Angeles, CA, USA

Hartmut Lehmann
University of Bologna, Bologna, Italy; and IZA, Bonn, Germany

Lora M. Levett
University of Florida, Gainesville, FL, USA

Heidi M. Levitt
The University of Massachusetts Boston, Boston, MA, USA

Jethro K. Lieberman
New York Law School, New York, NY, USA

Wolfgang Littek
University of Bremen, Bremen, Germany

Simon Lohse
Center for Philosophy and Ethics of Science, Institute of Philosophy, Leibniz University of Hannover, Hannover, Germany

Janet E. Lord
Syracuse University, Baltimore, MD, USA

Pat Lovie
Keele University, Keele, Staffordshire, UK

Sandy Lovie
University of Liverpool, Liverpool, UK

Tommaso Luzzati
Università di Pisa, Pisa, Italy

Robert J. MacCoun
University of California, Berkeley, CA, USA

Joël Macoir
Faculté de Médecine, Université Laval, Québec, QC, Canada

Brian MacWhinney
Carnegie Mellon University, Pittsburgh, PA, USA

Margherita Margiotti
University of Bristol, Bristol, UK

Gregory A. Mark
Rutgers-Newark School of Law, NJ, USA

Hans J. Markowitsch
University of Bielefeld, Bielefeld, Germany

Karin D. Martin
John Jay College of Criminal Justice, CUNY, New York, NY, USA

William J. McConnell
Michigan State University, East Lansing, MI, USA

Volker Meja
Memorial University of Newfoundland, St. John's, NL, Canada

Sally E. Merry
New York University, New York, NY, USA

Peter Meusburger
Heidelberg University, Heidelberg, Germany

Shae D. Miller
University of California, Santa Barbara, CA, USA

John Monahan
University of Virginia School of Law, Charlottesville, VA, USA

Calvin Morrill
University of California, Berkeley, CA, USA

Jaimie Morse
Northwestern University, Chicago, IL, USA

Robyn Munford
School of Health and Social Services, Massey University, Palmerston North, New Zealand

Alexander Muravyev
Higher School of Economics, St. Petersburg Campus, St. Petersburg, Russia; and IZA, Bonn, Germany

Nancy Stano
The University of Texas at Austin, Austin, TX, USA

Jonathan R. Nash
Emory University Law School, Atlanta, GA, USA

Gertrud Nunner-Winkler
Max Planck Institute for Psychological Research, Munich, Germany

William O'Grady
University of Hawai'i, Honolulu, HI, USA

Ruth L. Okediji
William L. Prosser Professor of Law, University of Minnesota Law School, Minneapolis, MN, USA

Kenneth R. Olwig
Swedish University of Agricultural Sciences, Alnarp, Sweden

Joanna Overing
University of St Andrews, St Andrews, UK

Maria K. Pavlova
Center for Applied Developmental Science (CADS), Friedrich Schiller University Jena, Jena, Germany

Anne-Nelly Perret-Clermont
Université de Neuchâtel, Neuchâtel, Switzerland

Eleni Pinnow
University of Wisconsin-Superior, Superior, WI, USA

Michael J. Platow
Research School of Psychology, Australian National University, Canberra, ACT, Australia

Constantin Potagas
Eginition Hospital, National and Kapodistrian University of Athens, Athens, Greece

Henry Prakken
Utrecht University, Utrecht, The Netherlands; and University of Groningen, Groningen, The Netherlands

Michael Preston-Shoot
University of Bedfordshire, England

Jason H. Prior
University of Technology Sydney, Broadway, NSW, Australia

Julia A. Puckett
The University of Massachusetts, Boston, MA, USA

James Pustejovsky
Brandes University, Waltham, MA, USA

Deepti S. Raja
Syracuse University, Washington, DC, USA

Dwight Read
University of California, Los Angeles (UCLA), Los Angeles, CA, USA

Stephen D. Reicher
School of Psychology, University of St. Andrews, Scotland, UK

Krystal M. Ribot
Florida Atlantic University, Davie, FL, USA

Ken Roberts
University of Liverpool, Liverpool, England, UK

John P. Robinson
University of Maryland, College Park, MD, USA

Daivi Rodima-Taylor
Boston University, Boston, MA, USA

Ralf Rogowski
University of Warwick, Coventry, UK

Suzanne Romaine
Merton College, Oxford, UK

Edward Rubin
Vanderbilt University, Nashville, TN, USA

August Ruhs
Medical University of Vienna, Wien, Austria

Stephen Rushin
University of Illinois, Champaign, IL, USA

Robert Salais
IDHE, Cachan, France

Geoffrey Samuel
Kent Law School, UK; and Law School, Sciences-Po, Paris

Viktor Sarris
Johann-Wolfgang-Goethe University, Frankfurt am Main, Germany

Günther Schlee
Max Planck Institute for Social Anthropology, Halle (Saale), Germany

John Henry Schlegel
State University of New York, Buffalo, NY, USA

Alyssa Schneebaum
Vienna University of Economics and Business, Vienna, Austria

Ricard Serlavos
Ramon Llull University — ESADE Business School, Leadership Development Research Centre, Sant Cugat (Barcelona), Spain

Steven Shavell
Harvard University, Cambridge, MA, USA

Parker Shipton
Boston University, Boston, MA, USA

Rainer K. Silbereisen
Center for Applied Developmental Science (CADS), Friedrich Schiller University Jena, Jena, Germany

Susan S. Silbey
School of Humanities, Arts, and Social Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Grażyna Skapska
Kraków, Poland

Robert Skidelsky
Warwick University, Coventry, UK

Steven L. Small
University of California, Irvine, CA, USA

Kenny Smith
Language Evolution and Computation Research Unit, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, UK

Peter H. Solomon
University of Toronto, Toronto, ON, Canada

Peter Spirtes
Carnegie Mellon University, Pittsburgh, PA, USA

Angelica Staniloiu
University of Bielefeld, Bielefeld, Germany

Nancy Stano
The University of Texas at Austin, Austin, TX, USA

Nico Stehr
Zeppelin University, Friedrichshafen, Germany

Brigitte Stemmer
Université de Montréal, Montreal, QC, Canada; and Brock University, St. Catherines, ON, Canada

Wolfgang Streeck
Max Planck Institute for the Study of Societies, Cologne, Germany

Francisco I. Surace
The University of Massachusetts Boston, Boston, MA, USA

Symeon C. Symeonides
Willamette University College of Law, Salem, OR, USA

Paula Tallal
Salk Institute for Biological Studies, La Jolla, CA, USA; and Center for Human Development, University of California San Diego, La Jolla, CA, USA

Brian Z. Tamanaha
Washington University in St. Louis, St. Louis, MO, USA

Jakke Tamminen
Royal Holloway, University of London, Egham, Surrey, UK

Verta Taylor
University of California, Santa Barbara, CA, USA

Robert D. Tennyson
College of Education and Human Development, University of Minnesota, Minneapolis, MN, USA

Laurel S. Terry
Penn State Dickinson School of Law, Carlisle, PA, USA

Gordon C. Teskey
Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada

Adina M. Thompson
University of Florida, Gainesville, FL, USA

Richard F. Thompson
University of Southern California, Los Angeles, CA, USA

W. Tim Murphy
London School of Economics, London, UK

Timothy Triplett
Urban Institute, Washington, DC, USA

David M. Trubek
University of Wisconsin—Madison, Brooklyn, NY, USA

Yves Turgeon
Clinique Médicale Carleton-sur-Mer, Carleton-sur-Mer, QC, Canada

Joaquim Pires Valentim
Faculdade de Psicologia e de Ciências da Educação, Universidade de Coimbra, Coimbra, Portugal

Marcel van der Linden
International Institute of Social History, Amsterdam, The Netherlands

Ashley Volk
College of Education and Human Development, University of Minnesota, Minneapolis, MN, USA

Alexander Volokh
Emory University School of Law, Atlanta, GA, USA

Lesley Watson
Emory University, Atlanta, GA, USA

Claire E. Weinstein
The University of Texas at Austin, Austin, TX, USA

Sue Westwood
University of Keele, Keele, Staffordshire, UK

Allison M. Whelan
University of Minnesota, Minneapolis, MN, USA

Anna Wierzbicka
Australian National University, Canberra, ACT, Australia

David B. Wilkins
Harvard Law School, Cambridge, MA, USA

Christine Windbichler
Humboldt Universität zu Berlin, Berlin, Germany

Jie Yang
University of California, Irvine, CA, USA

Elaine Zanutto
University of Pennsylvania, Philadelphia, PA, USA

This page intentionally left blank

K

Kant, Immanuel (1724–1804)

Volker Gerhardt, Humboldt University, Berlin, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 8071–8075, © 2001, Elsevier Ltd.

Abstract

Kant, Immanuel (1724–1804) German philosopher; academic career at the University of Königsberg. Early interest in natural science led to publications in natural history. Treatise on *First Principles of Metaphysical Cognition* (1755) prefigures the object of Kant's critical philosophy: the conditions of the possibility of human knowledge. 'Metaphysical' is that knowledge which exceeds empirical experience. Kant's transcendental philosophy is professedly a 'metaphysics of metaphysics' in which critical standards for metaphysics are established. Primary condition of human knowledge is the impossibility of knowing 'things-in-themselves'; these are concepts with no access to absolute truth about the world. Our knowledge depends on such concepts, however, and is hence relative, since determined by our cognitive capacities. The *Critique of Pure Reason* (1781) develops these theses systematically. The *Critique of Practical Reason* (1788) draws practical consequences of the first Critique: neither worldly good nor evil, but rather the human will is the foundation of morality. The *Critique of Judgment* (1790) contains theories of aesthetic judgment and of living nature. Kant's political philosophy includes a conception of law, morality, and politics in which human rights guarantee freedom and equality in nation-states bound by international treaty. Kant's powers failed before the completion of his philosophy of nature.

'The character of a true philosopher is such that the philosopher does nothing other than to exercise his natural powers and capacities – this namely through the searching investigation of critique.' This claim cannot be put more humbly – it is the very claim which guided Kant on his way to the *Critique of Pure Reason* and effected a 'revolution in style of thought' which continues until this day. According to Kant, this exercise and development of the best human powers deserves the title of philosophy only then, however, if such activity emanates from independent thinking: 'my philosophy must be grounded in my own self, and not in the understanding of others.' Philosophy cannot merely serve the satisfaction of individual curiosity, but rather should be related to the 'interests of humanity.' At the outset of his critical career, Kant wrote that he would deem himself 'more useless than a common laborer' if he couldn't believe that philosophy would contribute to 'the establishment of the rights of humanity.'

Kant was born on 22 April 1724 in Königsberg into a family of manual laborers of very little means. His schooling took place in the strict spirit of Pietism. At the age of 16, Kant began his studies at the university of his native city, financing himself in part by giving private lessons and playing at the billiard table. The breadth of his subjects of study – mathematics, natural science, theology, philosophy, and Latin – was later to be reflected in the wide spectrum of Kant's fields of inquiry. Initially, his interest in modern physics was particularly strong; he viewed Newton's *Philosophiae naturalis Principia mathematica* as the paradigm of rigorous science.

Kant's first academic publication, *Thoughts on the True Estimation of Living Forces* (1747), owes its importance above all to its central claim, which is very much indicative of Kant's later philosophy. Here the 24-year-old attempts to settle a dispute between Cartesians and Leibnizians by proposing a mediating position relative to the point of view of each party. Kant exclaimed: 'I have already drawn up the path which I would like to follow. I shall set upon it, and ... nothing should prevent me from pursuing it further.' Having lost the basis of his financial support through the death of his father, Kant left the university in 1746 in order to earn a living as private tutor in various families. During this time, Kant had occasion to pursue several extensive studies in natural science, the results of which he published anonymously in 1755 under the title *Universal Natural History and Theory of the Heavens*. Here, both the birth of the cosmos and solar systems, as well as the development of the earth up until the appearance of life and the rational animal, man, is deduced solely from physical laws. In positing the creation of heavenly bodies out of rotating nebula, Kant anticipated Laplace's theory; and the insight that spiral nebula are actually very distant galaxies secured Kant a sure place in the history of astronomy. In the year 1755, Kant returned to the university. Having attained a doctoral degree with a meditation 'On Fire' (*De igne*), he earned his *venia legendi* with a postdoctoral treatise on 'First Principles of Metaphysical Cognition.' In this work Kant marked out that field of problems with which he was to occupy himself for his entire life. Even the so-called 'transcendental philosophy' of the critical period remains connected to the 'first principles,' in that it poses the question after

the 'conditions of the possibility' of our knowledge. Although these 'conditions of possibility' are no longer located in a divine understanding – later, Kant came to understand them as the functional conditions of human understanding and in this respect as *a priori* – they nevertheless continue to be the object of metaphysics.

For Kant, all scientific knowledge is metaphysical insofar as it requires us to go, by means of concepts, beyond purely empirical experience. This is always the case in those passages where Kant speaks of knowledge in general, of 'world' and 'reality,' of morality, beauty, or history. Understood in this sense, transcendental philosophy is a 'metaphysics of metaphysics.' Moses Mendelssohn's famous statement that Kant was a 'Crush-All' of metaphysics should therefore not be so construed, as if all metaphysics had been discredited by critique. Actually, Kant disabled only the claims of a theorational and empirist metaphysics, while setting 'critical' standards which all scientific metaphysics must be able to satisfy.

In the over 25-year period between Kant's postdoctoral thesis of 1755 and the publication of *Critique of Pure Reason* in 1781, Kant wrote several works concerning the foundation of theoretical and practical metaphysics. He also experimented with empirically designed procedures, and began to show a certain affinity for skeptical methods – an affinity which was to prove lasting. Two works from this period give us the impression that Kant might have considered abandoning metaphysics altogether: the elegantly parleying *Observations on the Feeling of the Beautiful and the Sublime* (1764) and the satirical *Dreams of a Spirit-seer* (1766). Kant's appearance of having an antimetaphysical stance must have been strengthened for his contemporaries when, having completed yet another obligatory dissertation 'On the Form and Principles of the Sensible and Intelligible World' (*De mundi sensibilis atque intelligibilis forma et principiis*) in 1770, Kant published nothing worthy of mention for 11 years. In retrospect, we can recognize the approach of critical thought in this second dissertation: space and time are introduced as human forms of intuition which do not belong to things as understanding apprehends them. At this point, however, Kant still thought human understanding capable of a kind of cognition which could apprehend the nature of things as a divine understanding might.

The continuing 'critical' insights of the following years forced Kant to renounce the claim that we are able to know the nature of things. From a human perspective, nothing can be said about how 'things-in-themselves' are constituted. At this point in Kant's philosophical development, such 'things-in-themselves' as a divine understanding might be able to know can no longer serve us as the idea of a substance which could provide a basis for the things we can know. It does not even make sense to pose the question concerning the reality of 'things-in-themselves,' for they are mere concepts – we do not even know if anything corresponds to them. As such, the 'thing-in-itself' is simply an expression of our own limitation. The 'thing-in-itself' may serve as a limiting concept, however, in order to prevent us from deeming human knowledge absolutely certain. Even the pure concepts of understanding and reason do not have access to absolute truth.

Kant by no means contests the possibility of knowledge; he is not an agnostic on this count. Rather, he emphasizes the relativity of human knowledge: every insight, even in the objective natural sciences, is functionally relative to human cognitive capacities. We perceive reality in the manner in which it 'appears' to us. Every objective perception is dependent on the sensible intuition in space and time, and remains in this respect limited to experience. But renouncing experience leads either to analytical judgments, which only express what is already contained in concepts, or it misleads us to speculation, in which all certainty is lost. Yet the speculative thinking of pure reason is not sterile. Under conditions of disciplined self-critique, pure reason may contribute to the guidance and ordering of empirical knowledge (its 'regulative function'), it can ward off dogmatic claims (its 'critical function'), and it can set practical goals in those areas where certain knowledge is not attainable anyway. In this last case, pure reason becomes 'practical reason.'

The *Critique of Pure Reason* (1781), in which Kant systematically developed these insights, initially received little attention. The second edition (1787), elucidated by the Prolegomena (1783) and expanded by the *Grounding for the Metaphysics of Morals* and the *Metaphysical Foundations of Natural Science*, gave rise to a sea change, the likes of which philosophy had seldom seen. Kant's experimental-philosophical proposal for the foundation of the results of human knowledge proceeded no longer from the character of the object, but rather from the conditions of the subject – an innovation which came to be understood as a revolution in science. The 'cardinal propositions' of human knowledge cannot be extracted from 'things-in-themselves,' but rather have their fundaments in our very acts of cognition, and only become the fundament of nature *qua* 'appearance' through human activity. This went very well with the heightened self-confidence of the Enlightenment. The critique of reason declares man originator of his world – without contradicting the results of natural science or questioning the traditional claims of the belief in divine origins. At the same time, the delimitation of knowledge by means of critique serves to assign belief a genuine place. God and world, freedom and the person come to be interpreted as concepts of pure reason; they do not denote any object, but are imperative for the conceptual self-determination of the individual.

This attempt to ground human procedural capacities *a priori* had consequences for practical action; Kant showed these in the *Critique of Practical Reason* (1788). In ethics, too, the cardinal propositions are not grounded in the world, but are the product of our own capacities. It is not possible to speak of any preexisting 'good' or 'evil' in the world: one must seek the origin of these terms solely in the human will. Insofar as this will conceive itself as a rational one, it determines by itself what is to count as morally good. The principle of the will is therefore that of self-determination (autonomy) according to the principle of practical reason in the will itself. So long as we acknowledge a commitment to this principle – and we have already done so when we ask ourselves what we 'should' do – we understand ourselves as free and autonomous. In this way, we have the possibility to orient the principles of our action (maxims) on rational insight and submit them to the general self-jurisdiction of reason, which consists in the categorical

imperative. Like the categorical elements of the perception of nature, the categorical imperative does not emanate from external authority but from man's own conception of himself as a rational being.

No new virtues are made compulsory in Kant's critical refoundation of ethics. Instead, principles from Platonic, Aristotelian, and Stoicist ethics are made subordinate to specific criteria of reason. In the last analysis, the emancipated individual alone must determine these criteria. Justification of behavior through one's own insight (and therefore through one's own reason), which first came into philosophical consciousness through Socrates, ascends to its full conceptual conclusion here. The autonomy of ethical self-determination aspired to since antiquity, while being defined over and against empirical explanation, thereby gains a pointedly individual expression. The 'categorical imperative' has meaning only in relation to the 'maxims' or 'subjective principles' of the individual. It is to be noted that justification always refers only to the 'form' or conceptual part of a moral judgment. The concept alone is meaningless without content or 'material,' however, and this is also the case in ethics.

There are therefore always certain 'preferences' involved in that 'material' which must be adduced to every moral judgment and hence to every moral act. Feelings and passions belong to the moral world. Thus, there must be some feeling which gives impetus to moral action – Kant calls it the 'feeling of reverence.' He is careful not to conflate the necessary motivation of acting with its justification; as in Plato, justification may only proceed from rational insight. Solely by means of this kind of insight can man experience himself as free.

Kant's third critique, the *Critique of Judgment* (1790), contains in Part 1 a grounding of aesthetics, once again departing from human standards alone. Aesthetic judgment is the expression of a harmony of cognitive powers which stimulate the individual pleasurably: whenever sensibility and reason come together and mutually intensify one another, an aesthetic experience is possible. Aesthetic experience also expresses itself in a 'judgment' in which the individual, existing in unrestricted subjectivity, communicates himself to others. Kant frees aesthetic judgment from all claims of scientific truth, moral valuation, and purely sensible satisfaction. In aesthetic experience, everything depends on one's own self-autonomous ('heautonomous') judgment. This faculty of judgment binds all sensible and mental powers together in itself and is capable of producing its own pleasure. What is more, it demands expression among those who are like-minded. Hence the faculty of judgment does not at all isolate the individual, but is the very thing which brings him into a conscious relation with those who are congenial to him.

In Part 2 of the *Critique of Judgment* Kant draws up a theory of living nature. The individual constantly judges nature in analogy to his own experience as an active being. An organism is perceived in accordance with its functional purpose, 'as if' a rational and end-oriented will were operating within it – a will which the individual is admittedly only acquainted with through himself. Kant's theory of the organism attempts to grasp the dynamic complex unity of living beings. In so doing, he draws upon the principle of causation whenever this is possible. In this vein, Kant establishes the definition that the parts of the organism 'join one another in the unity of the Whole by virtue of being mutually the cause and effect of one another's form.' Indeed, a living product of nature stands in itself 'reciprocally as cause and effect.' Kant there by made the way for a theory of 'self-organization' which found much attention in the twentieth century, in biology as well as in physics and sociology. The third critique contains the most important impulses for the following generation of German philosophers, particularly for Schelling and Hegel. These philosophers sought to overcome the subjective limitation of aesthetic and teleological judgments emphasized by Kant within the framework of an extensive objectivity of nature or of mind.

Once he had completed his critical works, the elderly Kant went about the business of drawing systematic consequences for particular areas of research. He only managed this, however, in the domain of practical action. Against the resistance of the Prussian Censor, Kant showed these consequences for religious belief (*Religion Within the Limits of Reason Alone*, 1793) and published in 1796/77 the *Metaphysics of Morals*, containing not only a 'doctrine of virtue' for self-responsible action but also doctrines of law and state. Together with the brief essay *Perpetual Peace*, this work exercised direct influence on the political opinions of Kant's contemporaries.

Kant put the productivity of his principle of critical justification to the test in numerous smaller writings on political philosophy and the philosophy of history, also concerning the concept of mankind or the relation among the sciences. Though he added no new model to the pre-existing political constructions of his time, he did develop a new principle of legitimization for political rule under the auspices of a conception of law and the idea of international treaty. He showed how, under general conditions of law, morality, and politics are connected through the medium of the principle of publicity without being derived from one another. He probably was the first to see that the action of states is not only to be justified with regard to the obligations to its own citizens, but is rather to be measured by 'cosmopolitan' standards which require international protection. Where once was the primarily religiously motivated idea of a Christian and European peace order, Kant placed the concept of an international peace order founded upon legal institutions – a concept upon which was drawn in the foundation of the League of Nations in the twentieth century.

Kant's theory of world peace was an innovation in political theory. It was based upon a historically and economically well-founded diagnosis of the state of mankind on the verge of the complete colonialization of the earth. There is no longer a single political entity on the globe which is independent from the political events in any other given part of the world. Because the surface of the earth offers only limited space, people can no longer make way for one another. War, the historical importance of which Kant recognized, can therefore no longer be a means for political ends. The French Revolution had shown that a people's claim to freedom is capable by itself of causing overdue changes in law. Hence Kant arrived at a comprehensive conception of the right of morality and politics on the basis of 'human rights.' These rights, which guarantee freedom and equality for all, are to be made binding within states by means of a republican and representative constitution, and are to be secured among different states through a system of treaties.

A stop is put to colonial politics by 'world citizens' rights'; at the same time, each individual citizen has the basic right to humane treatment, even in states other than his own. The individual states remain intact as indispensable units of social self-determination, but they are compelled by the dynamics of self-binding law to make internal reforms and to rely upon a federal alliance in their external relations. In his brief treatise on *Perpetual Peace*, Kant sketched a theory of politics appropriate to modern conditions of action through the combination of juridical, moral, rhetorical, and power-political elements.

Kant was prevented from completing his system of natural philosophy by numerous obligations which resulted from the broad public debate over his works. Inspired by the French Revolution, Germany experienced a euphoric decade of critique before the turn of the eighteenth century. Kant was thus exposed not only to a growing following, but also to a rapidly formed and primarily religiously motivated polemic. He gave proof of his literary ability in his equally polemical responses to this polemic. It is an ability which, in Kant's main works of critique, only reveals itself to those who have penetrated the difficult coherence of his thought.

In 1796 Kant ceased to teach and in 1801, at the age of 77, retired. His physical strength, already weak, deteriorated rapidly thereafter. Kant could hardly influence the edition of his lectures on anthropology, pedagogy, logic, and geography which were prepared by his students. Nevertheless, he continued working on the yet incomplete philosophy of nature. In doing so, Kant returned once again to the premises of his epistemological critique and conceived a unity of man and nature which bears a remarkable resemblance to the systems of Schelling and Hegel. While writing, Kant's strength suddenly failed him. The speculative notes break off with increasing regularity, and comments on everyday incidents are interspersed among Kant's notes. These writings, which today are collected and published under the title *Opus postumum*, document the way in which Kant's body gradually failed his mind.

Kant's final years were marked by a deterioration of physical and mental powers which was nevertheless born with dignity. On 12 February 1804, Kant died of old age in his city of birth. Despite numerous offers from other universities, Kant never left Königsberg. Yet he always took an active interest in world events, and kept company with businessmen, administration officials, and officers. He knew the travel literature of his time so well that one might think he had been to Java, America, or London himself. He was an enthusiastic supporter of the ideas of the French Revolution, both in the daily society gathered at his table and in public. In Kant's main œ*uvre* of critique as well as in his wide-ranging biological, social, and political interests, one driving incentive of Kant's philosophy is manifest: the exploration of the possibilities and boundaries of man.

In his lectures on logic Kant often summarized the 'cosmopolitan importance' of philosophy in four central questions: (a) 'What can I know?' (b) 'What should I do?' (c) 'What may I hope for?' (d) 'What is man?' To the first question, Kant added by way of explanation, 'answer *metaphysics*; to the second, *morality* answers; to the third, *religion;* and to the fourth, *anthropology*.' The true meaning of this catalogue of questions is only recognized then when Kant adds that the last question encompasses all the others. In conclusion Kant writes that one could basically count metaphysics, morality, and religion to anthropology, 'since the first three questions are related to the last.' Thus Critical Philosophy appears to be, above all, a comprehensive inventory of the best human faculties – and these are the faculties of reason.

See also: Cognitive Psychology: History.

Bibliography

Allison, H.E., 1990. Kant's Theory of Freedom. Cambridge University Press, Cambridge UK.
Allison, H.E., 1983. Kant's Transcendental Idealism. An Interpretation and Defense. Yale University Press, New Haven, CT.
Beck, L.W., 1974. Kants 'Kritik der praktischen Vernunft'. Ein Kommentar. Fink Verlag, Munich, Germany.
Cassirer, E., 1918. Kants Leben und Lehre. B. Cassirer, Berlin.
Friedmann, M., 1992. Kant and the Exact Sciences. Harvard University Press, Cambridge, MA.
Gerhardt, V., 1995. Immanuel Kants Entwurf 'Zum ewigen Frieden' Eine Theorie der Politik. Wissenschaffliche Buchgesellschaft, Darmstadt, Germany.
Guyer, P., 1987. Kant and the Claims of Knowledge. Cambridge University Press, Cambridge, UK.
Guyer, P., 1989. The unity of reason. Pure reason as practical reason in Kant's early concept of the transcendental dialectic. The Monist 72, 139–167.
Guyer, P., 1992. The Cambridge Companion to Kant. Cambridge University Press, Cambridge, UK.
Höffe, O., 1983. Immanuel Kant. C H Beck. Munich, Germany.
Kaulbach, F., 1969. Immanuel Kant. De Gruyter, Berlin.
Kersting, W., 1984. Wohlgeordnete Freiheit. Immanuel Kants Rechts-und Staatsphilosophie. De Gruyter, Berlin/New York.
Langer, C., 1986. Reform nach Prinzipien. Untersuchungen zur politischen Theorie Immanuel Kants. Klett-Cotte, Stuttgart, Germany.
O'Neill, O., 1989. The Constructions of Reason: An Exploration of Kant's Practical Philosophy. Cambridge University Press, Cambridge.
Prauss, G., 1973. Kant. Zur Deutung seiner Theorie von Erkennen und Handeln. Kiepenneus und Witsch, Cologne, Germany.
Shell, S.M., 1996. The Embodiment of Reason: Kant on Spirit, Generation, and Community. University of Chicago Press, Chicago.
Sommer, M., 1988. Identität im Übergang: Kant. Suhrkamp, Frankfurt am Main, Germany.
Wood, A.W., 1978. Kant's Rational Theology. Cornell University Press, Ithaca, NY.

Kantian Ethics and Politics

Matthias Kaufmann, Martin-Luther-Universität Halle-Wittenberg, Halle (Saale), Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Kantian ethics judges an action as morally good if it is done autonomously out of duty and its maxim could be the foundation of a general law. Neither the purpose nor the outcome of an action have any impact on its moral value. This is why Kantian ethics normally is seen as contrary to any type of consequentialist ethics. But Kant tries to reconcile the human wish for happiness with his concept of morality in his theory of the 'highest good,' and in the Doctrine of Virtues he explains in which sense the outcome of an action does matter. Politics, according to Kant, always has to be strictly under the rule of law as it works to achieve the happiness of the people as an end in itself. Implementing the rule of law also in international politics, Kant hopes that international relations will move slowly toward eternal peace.

In his ethics, Kant pointed out that our ability to follow the moral law as rational beings, that is, our capacity to act freely without being determined by natural inclinations, puts us under the obligation to act out of respect for the moral law according to which we should act so that the maxim of the will could be a universal law. Our own happiness, as a natural end, can never be a moral motive; we may hope only to be rewarded for our efforts in eternal life. Nevertheless, we ought to take care of our own well-being because this makes it easier for us to act morally, and in his treatment of different virtues and vices, Kant put a great deal of emphasis on self-respect. Politics, according to Kant, always has to be under the rule of law. Although Kant admited that the aim of politics is the happiness of the people, he insisted that this must never be contrary to the principles of law, which he defined as the compatibility of one person's outer freedom with the freedom of all others. In international politics, Kant postulated a federation of free states that – he hoped – would move slowly toward eternal peace.

Kantian ethics judges an action as morally good if it is done autonomously out of duty and its maxim could be the foundation of a general law. Neither the purpose nor the outcome of an action have any impact on its moral value. This is why Kantian ethics normally is seen as contrary to any type of consequentialist ethics. Politics, according to Kant, always has to be strictly under the rule of law as it works to achieve the happiness of the people as an end in itself. Implementing the rule of law in international politics as well, Kant hoped that international relations would move slowly toward eternal peace. (Because it is not in all cases clear whether Kant meant 'right' or 'law,' and because he sometimes seemed to make use of the double meaning of the German word *Recht*, it is more appropriate to use the word 'right' instead of 'law,' as Mary Gregor did in her translation; Kant, 1996.)

Kant drew a clear distinction between the fields of morality and legality. Hence, it is possible to discuss these topics separately, without denying the important connections that are due to the common origin of their key principles in (practical) reason while the nature of these connections is still under debate. (cf, e.g., Kersting, 1993: pp. 175ff.; Flikschuh, 2010: pp. 62ff.; Guyer, 2002; Willaschek, 2002)

Ethics of Duty – Ethics of Autonomy

Although, as with any other label, it is an oversimplification to reduce Kantian ethics either to an ethics of autonomy or to a mere ethics of duty, it might help to find an initial approach to say that somehow it is both (Baum, 2005). The connection between these two central concepts will be clarified later. They both signify, however, that Kant did not develop a eudaimonian (happiness-based) or any type of consequentialist ethics. Kant did not think that happiness is unimportant for human life; to the contrary, to pursue one's happiness is a moral duty (Himmelmann, 2003; Allison, 2011). But he deliberately rejected ancient eudaimonian ethics as a possible foundation for morality, and centered his postcritical works on moral philosophy around the concepts of duty, autonomy, and moral law (Wood, 1999; Timmermann, 2009; Forschner, 1993; Wike, 1994). Simplifying again, it seems helpful to characterize Kantian ethics within three steps, more or less identifying each with one of the central Kantian works on moral philosophy.

Extracting and Vindicating the Categorical Imperative: The Groundwork of the Metaphysics of Morals

In his first larger postcritical work on ethics, starting from the statement that nothing can be called good without restrictions except a good will, Kant showed that a will is good if and only if it acts autonomously out of duty and according to the categorical imperative (CI). Everything other than a good will may be used with bad intentions. Even some of the virtues can be held by immoral people as the "coolness of a scoundrel" proves to us (Kant, 1996: p. 50). By saying that lack of success does not diminish the moral value of good will "as something that has its full worth in itself" *as long as good will acts "with its greatest efforts"* (Kant, 1996: p. 50), Kant immediately answered some of the most tenacious objections to his ethics.

The main topic of the groundwork of the metaphysics of morals (GMM) is an attempt to clarify the relations between duty, autonomy, and the law of morals. First, Kant showed that the concept of duty is the best way to explain what is meant by a good will; particularly, how the moral worth of an action does not lie in its expected effect but in the principle of volition. An action is morally good if it is not only done in *conformity*

with duty but *from duty* (Kant, 1996: p. 53). Duty is defined as "the necessity of an action from respect for law" (Kant, 1996: p. 55) (according to the principle that I ought to act in such a way that my maxim could become a universal law) and respect is defined as a feeling that is "not received by means of influence" but "self-wrought by means of a rational concept and therefore specifically different from all feelings … that can be reduced to inclination or fear" (Kant, 1996: p. 56). In Kant's view, the question of whether our maxims could become a universal law provides us, "inexperienced in the course of the world" as we are, with an easy test of whether the maxim of our volition is morally good. This is what "common human reason" has "always before its eyes and uses as the norm for its appraisals," although it "admittedly does not think so abstractly in a universal form" (Kant, 1996: p. 58). It is important to notice Kant's regard for common human reason and the morality of the common people, which is, according to his own testimony, due to the influence of Rousseau (Schneewind, 1998: pp. 487ff.; Kant, 1902: ff. XX, 44). Kant did not see the CI "act only in accordance with that maxim through which you can at the same time will that it become a universal law" (Kant, 1996: p. 73), as an invention of philosophers, but rather he saw it as a precise and short formula of how ordinary people think about morality. What he claimed, however, was that – until now – no philosopher had found a clear and precise description of what common reason thinks about ethics.

Nevertheless, common human reason might indeed find some difficulties in the abstractness of this account, just as philosophers have until now. This can be seen from the comment of Friedrich Schiller who felt immoral because he loved his friends from inclination and not from duty. Hence, it should be mentioned that actions in conformity with duty that are motivated by inclination are not immoral. We just have no means to judge whether they are moral. Kant, however, distrusted inclination, which he characterized in the *Critique of Practical Reason* (*CPrR*) as "blind and servile, whether it is kindly or not" (Kant, 1996: p. 235).

On the other hand, as Kant explained from the outset of the second part of his *Groundwork*, we never really know whether we, or anybody else, acts from duty. If we look closely enough at all empirical actions, even at those "from love of humankind" that are generally in conformity with duty, we "come upon the dear self, which is always turning up" (Kant, 1996: p. 62). Therefore, we require an "isolated metaphysics of morals" that is made for all rational beings. Thus, there are the two results: (1) that morality is something for rational beings and (2) that a will is a good will not from love of mankind or any other inclination but from its accordance with the moral law. With good will reason itself becomes practical.

An ongoing subject that is discussed several times in this and in the later works is the relation of happiness to morality (Hills, 2009). Kant constantly argued that morality cannot be based on any principle about happiness. He had different reasons for saying this. First, because instinct is a much better guide to happiness than reason, it would have been an unnecessary and useless effort of nature to provide us with reason if it were for the sake of happiness. Although this argument has been criticized for its obscure 'naturalism,' one may in fact agree with the view that happiness is a natural aim for rational beings, so it is useless to set the pursuit of happiness as a norm. According to another argument "it is a misfortune that the concept of fortune is such an indeterminate concept that, although every human being wishes to attain this, he can still never say determinately and consistently with himself what he really wishes and wills" (Kant, 1996: p. 70). Furthermore, because it is difficult to predict the result of an action, it would be difficult to decide whether it is right or wrong, whereas, in Kant's view, it is very easy to know how the categorical imperative handles this. However, it is important to notice that morality and happiness are by no means necessarily in conflict. They are just logically independent, insofar as happiness is the most general term for an object of the will, that is, for the matter of the will, while the will's moral goodness is due to its form (*CPrR*, Kant, 1996: p. 160). It is even a moral duty to preserve one's own happiness just to avoid the temptation to act immorally out of poverty or envy.

The rejection of happiness as the basic moral end in itself leads Kant to the characterization of the CI as the one and only unconditionally valid imperative. In contrast to the imperatives of skill and prudence (prudence teaches us how to achieve happiness), there is no external end that justifies the categorical imperative. Another result is the special status of *rational* beings as ends in themselves. Kant's hypothetical construction of a 'kingdom of ends' (*Reich der Zwecke*), which is somehow included in the concept of every rational being, may be the best tool to show how closely duty, autonomy, and the moral law are interrelated in his moral philosophy. In the kingdom of ends, every rational being, including God, that is, every person, is seen as a lawgiver of the moral law. This is the reason why they are autonomous. Every finite rational being has the duty to follow this law. Every rational being is an end in itself and thus benefits from the moral law. Because of its deeply egalitarian character, obviously influenced by the Rousseauian view of the *volonté générale*, it might be better to follow Mary Gregor's suggestion for an alternative translation of *Reich der Zwecke* by "commonwealth of ends" (Kant, 1996: p. 83, on the "moral force of reciprocity," cf Korsgaard, 1996: pp. 194f.). In the discussion over the past few decades on animal and medical ethics, the question has been raised whether the circle of those who benefit from the moral law should be expanded to include beings that are not rational or only potentially rational (Nussbaum, 2006; Potter, 2005).

Because of the inordinate reliance on the *Groundwork* in the attempt to clarify Kant's moral philosophy, the charges of formality and emptiness have been laid against Kant. An integral reading of his works shows, however, that this charge is misplaced. It even becomes clear that it would be misleading to label Kant a 'rule ethicist.' True, maxims can be tested as to their universality, but on the ground of maxims alone it is not always possible to decide what to do in a particular situation. Kant was perfectly aware of this, as his stressing of the importance of the faculty of practical judgment and his casuistic questions in the *Doctrine of Virtue (DV) of the Metaphysics of Morals* make clear. Before we deal with these issues, we must look at why the CI is valid or why we should be moral, in Kant's words, how the categorical imperative is possible (cf Guyer, 1998). Although

Kant's main thrust on this issue is found in the *CPrR*, it is worth seeing how he dealt with the question in the *Groundwork*.

The Moral Law, Free Will, and the Highest Good: The Critique of Practical Reason

In the third part of the *Groundwork* that points to a critique of pure practical reason, Kant posed the following question: "But why, then, ought I to subject myself to this principle and do so simply as a rational being, thus also subjecting to it all other beings endowed with reason?" (Kant, 1996: p. 96). Here he answered this question by removing a "kind of circle" (Kant, 1996: p. 97, Schönecker, 1998). It seems that, on the one hand, we must be free from causal determination to be subject to the moral law, and on the other hand, we must declare that we are subject to the moral law because we are free. Freedom (i.e., autonomy), however, consists of the self-constraint of the will through the moral law.

Kant found his solution by making the distinction between two 'standpoints.' In one, we are in the world of senses, having only cognition of appearances, never cognition of things in themselves. In the other, we are, especially in a practical context, members of the world of understanding (Kant, 1996: p. 98). However, "as a rational being, and thus as a being belonging to the intelligible world, the human being can never think of the causality of his own will otherwise than under the idea of freedom" (Kant, 1996: p. 99), and this means submitting to the moral law. This being said, Kant admitted that we can never explain *how* pure reason becomes practical (Kant, 1996: p. 104). This seems to be the problem of his moral philosophy in general: The 'good will' through which pure reason becomes practical is in the intelligible world of understanding. Then it is difficult to understand how it can be *my* will. Or it is *my* will, and then it is difficult to see how it can be separated completely from all my emotional attitudes. In its essence, this problem remains more or less the same from the GMM until the *Metaphysics of Morals* (Timmermann, 2009; Engstrom, 2010). Even if we say that in accordance with the principle of charity we should read these passages in the sense that both parts are the result of abstraction from the 'disturbing' elements of the other side, it is Kant himself who has driven these abstractions to their extreme. To talk of two standpoints does not help much while there can be no explanation of their interrelationship.

The connection between the freedom of will and the rule of the moral law – the first is what makes the law hold, the second is what makes freedom known to us (*CPrR*, Kant, 1996: p. 140) – is one of the main topics of the *CPrR* in which Kant maintained several times that, with moral law, we have in practical contexts as a reality what could in theoretical philosophy only be proved to be not impossible in the solution of his 'third antinomy' in the *Critique of Pure Reason* (Kant, 1902: ff. III, 366ff.; cf Wood, 1999: pp. 172ff.; Ertl, 1998).

In a comment in section 7 of the chapter on the principles of pure practical reason, Kant explained the fundamental law of pure practical reason, that is, the CI, as a "fact of pure reason" (Kant, 1996: pp. 164f.) that is not empirically found. It is not quite clear what is meant by this cryptic formula, which is repeated several times within the *CPrR* and other Kantian works, sometimes with slight variations and has been criticized as well as defended (Kleingeld, 2010). But it appears not to mean that something is revealed to pure reason or to us by pure reason but rather that the use of pure reason in practical contexts implies that we are under obligation by moral law.

Happiness is the natural object of desire of human beings, that is, of all rational and finite beings, whereas the moral law rules the formal structure of the will, happiness, like any other material content, cannot be the incentive for moral law. Kant explicitly rejected the moral relevance of any distinction between higher and lower pleasures. Nevertheless, Kant tried to reconcile the human wish for happiness with his concept of morality in his theory of the 'highest good,' where he found in the doctrine of Christianity – in contrast to the teachings of the Epicureans and the Stoics (Kant, 1996: pp. 241ff.) – a chance to give to those who are worthy of happiness the "hope of some day participating in happiness to the degree that we have been intent upon not being unworthy of it" (Kant, 1996: p. 244). To support this hope, he proposed in his doctrine of postulates not only the freedom of will but also the immortality of our soul that continues its way toward perfection even after our death and the existence of an almighty god who can and will give deserved rewards to those who lived in conformity with moral law. Perhaps one cannot be sure that a *postulated* god can give *real* hope to everybody and an incentive to follow the moral law. Perhaps, instead, some may find this incentive in the hope for glory or in the hope to contribute to the moral improvement of the world or in something similar.

Living the Categorical Imperative: *The Doctrine of Virtues*

In the *DV* of his late masterpiece the *Metaphysics of Morals*, Kant dealt with the different types of duties that cannot be ruled by outer laws – duties of man concerning himself and his perfection and duties concerning the happiness of others. In the *DV*, it becomes clear that for Kant there are obligatory ends for human beings: This indicates that his moral philosophy, which surely is not a variation of consequentialism, is not a pure *Gesinnungsethik* either, that is, the outcome of an action does matter, although it is not always in our hands what this outcome might be (Herman, 2011). To be sure, that does not mean that the moral quality of an action is determined by anything else other than the quality of the motivation and the universality of the maxim in question. This includes the idea, however, that the agent must do everything he or she can to make morality (in the form of moral actions) a reality in the world. Because we have a duty to follow the CI, we have a duty to pursue our own perfection and the happiness of others. We also have duties toward ourselves. These are, again, due to our 'double existence' in the world of appearances and the intelligible world. We have duties toward ourselves as rational beings, that is, toward the humanity in us (section 3, Kant, 1996: p. 544).

A second important achievement of the *DV* compared with earlier works is the casuistic element that demonstrates how Kant's ethics might be applied, at least in some cases. There are also considerations on the use of judgment (*Urteilskraft*) in the *CPrR* (Kant, 1996: pp. 194ff.). But here, "the rule of judgment under laws or pure practical reason is this: ask yourself whether, if the action you propose were to take part, you could indeed regard it as possible through your will" (*CPrR*,

Kant, 1996: p. 196, on moral judgment see Herman, 1993). In his later work, however, Kant used the humanity formula, that is, our duty to respect the moral law and our duty to respect the humanity in ourselves as well as every man as an end in itself to deal with particular subjects. Among these questions and subjects are suicide – that is, to be seen as "debasing humanity in one's person" and therefore morally forbidden (section 6, Kant, 1996: p. 546), as well as "defiling oneself by lust" (Kant, 1996: pp. 548ff.), avarice (Kant, 1996: pp. 555ff.), and servility – "self-esteem is a duty of the human being to himself" (section 11, Kant, 1996: p. 558) – as well as also duties like beneficence and gratitude (Kant, 1996: pp. 571ff.) and vices like envy and malice (Kant, 1996: pp. 576ff.). Kant also dealt with friendship (Kant, 1996: pp. 584ff.). Particularly in the 'casuistic questions' that are added in different places, one can find that Kant's views on these subjects were somehow strict. They are more diversified than might be expected after reading the *Groundwork* and even the *CPrR*. Nevertheless, a remaining problem appears to be that Kant supplied no guidelines on how to interpret the moral law in 'hard cases,' although it might be said that this cannot be the task of virtue ethics in Kant's understanding (Esser, 2004: p. 267). This partly is due to his view that a law court, which he often used as a metaphor for the internal judgment of the conscience (*DV* section 13, Kant, 1996: p. 560), must *apply* the law via imputation, not *interpret* it – that is, the court has to decide whether this individual case falls under the law, not to explain what the law means. This seems to have been a common view of eighteenth-century continental legal theory. On the other hand, Kant may be the first philosopher who recognized that to apply a rule we need a faculty of its own, which he called the faculty of judgment (Esser, 2004: pp. 258ff.), a discovery that is sometimes attributed to Wittgenstein's reflections in rule-following.

Right and Politics

According to Kant, any ethically acceptable politics has to take place under the rule of right. It is the duty of 'moral politicians' who are opposed to hypocritical 'political moralists' to do their best to bring constitutions of states and political relations between states "into conformity with natural right, which stands before us as a model in the idea of reason" (Toward Perpetual Peace, TPP, Kant, 1996: p. 340). Nevertheless, there is an aim of politics on its own, that is, the happiness of the citizen (TPP, Kant, 1996: p. 351) – something that might be surprising to some of Kant's critics. Therefore, pivotal to Kant's legal and political philosophy is, first, his treatment of right and how to determine its realm and, second, his hope on historical development toward peace and republican government. The key engine for bringing about legal progress is the public use of reason (Toward Perpetual Peace, TPP, Kant, 1996: p. 351).

Right as Compatibility of Freedom

Defining right in the *Introduction to the Doctrine of Right* (*DR*) of his *Metaphysics of Morals* as the condition for the compatibility of one person's outer freedom with that of the other (Kant, 1996: p. 387), Kant gave an account of every human being's innate right of freedom as "independence from being constrained by any other's choice" (Kant, 1996: p. 393). In fact, according to Kant, "right is connected with an authorization to use coercion" (Kant, 1996: p. 388) as long as this coercion is opposed to a wrongful "hindrance to freedom in accordance with universal laws" and so is understood as "a hindering of a hindrance to freedom" (Kant, 1996: p. 388). 'Strict right' that is "not mingled with anything ethical" can even "be represented as the possibility of a fully reciprocal use of coercion that is consistent with everyone's freedom in accordance with universal laws" (Kant, 1996: pp. 388f.). The strict reciprocity of rightful coercion shows how crucial the egalitarian element is in Kant's legal philosophy, together with the emphasis on freedom from any other's arbitrary coercion.

Kant's *Doctrine of Right* begins with private right "concerning what is externally mine or yours in general." According to Kant's view, individual possession already exists in the state of nature (section 9, Kant, 1996: pp. 409f.) because the earth, which is originally in common possession, has to be divided according to the "originally united will," which is in accordance with the postulate of pure practical reason that there can be no res nullius (Byrd and Hruschka, 2010: pp. 132ff.). There have been vehement discussions about whether this view together with the general fact that private right is treated before public right justify looking at Kant as a possessive individualist (Saage, 1994; Zotta, 2000) and whether Kant's statement that this possession is only provisional and 'conclusive possession' is possible "only on a law of a common will" (section 9, Kant, 1996: p. 410) means that the common will just confirms possessions won in the state of nature or whether it rearranges and redistributes possession according to principles of justice, such that Rousseauian and therefore egalitarian parts would gain more weight. Kant rejected Locke's founding of property on work, because work is something 'accidental' to the land that cannot decide whose property it is (sections 15 and 17) and because it bases provisional possession on first occupation and conclusive possession on the common will (cf Williams, 1983: p. 94; cf Kersting, 1993: pp. 337ff.). Perhaps one might say that although Kant's primary concern is with individual freedom, which implies appropriation in the state of nature – according to his definition of right – he nevertheless has a sense of public welfare and distributive justice because freedom is not a privilege, but rather it is a universal right (Rosen, 1993; Westphal, 2002; Kaufmann, 2005; Ripstein, 2009: pp. 267ff.).

Because in the state of nature, property and any right are insecure "however well disposed … human beings may be" (section 44, Kant, 1996: p. 456), we have the moral and legal duty to enter a state of public right (section 42). What Kant described in his *DR* is 'the state in idea' that "serves as a norm for every actual union into a commonwealth" (section 45, Kant, 1996: p. 457). Kant clearly preferred a republic with division of powers that is the only alternative to despotism and "any true republic is and can only be a system representing the people" (DR section 52, Kant, 1996: p. 481), which is obviously incompatible with the absolutist and feudalist systems of his times. Therefore, there is another discussion on whether he really rejects every resistance against unjust regimes so bluntly as it looks like in the published

works (DR section 49 A, Kant, 1996: pp. 463f.; OCS, Kant, 1996: p. 303) or whether there might be some more or less hidden hints that he sees it in a more nuanced manner (Beck, 1993; Westphal, 1992). Anyway, it seems to be clear that Kant prefers reform of the constitution (Goyard-Fabre, 1996: pp. 217f.) and for this he claims again and again the transcendental principle of publicity and of freedom of the pen (OCS, Kant, 1996: p. 302).

According to an interesting and attractive interpretation, Kant accepted the political systems of his time – and perhaps he would do the same with those of our time – mainly because they provide us with the possibility of a change toward more just societies that are closer to the state in the idea (Brandt, 1982; Kersting, 1993). But this change has to be done by the sovereign via reform and not by the people via revolution (DR section 49 A, Kant, 1996: p. 465). As we have to accept the state of affairs in the distribution of property at least provisionally for being able to enter a civil state with distributive justice (DR section 41), we have to accept the actually existing legality for being able to transform it by rightful means until it comes close to a true republic. On an international level, we have to accept large parts of the actual distribution of power for being able to approach eternal peace via a commonwealth of free republics. The common tool for this treatment of nonperfect social situations is the permissive law that allows us to live in circumstances that are contrary to the principles of right as long as a quick change by means of right is impossible and as long as there is a serious effort to improve the state of affairs. Since Kant's use of the term 'permissive law' in different places is rather complicated (Hruschka, 2004; Byrd and Hruschka, 2010: pp. 94ff.) and at the same time not that independent from the natural law tradition as sometimes suggested, there have been intense debates about this interpretation, which nevertheless has not been discarded on the whole (Kaufmann, 2005).

Toward Perpetual Peace

One of the most influential and effective elements in Kant's political philosophy might be his proposal of an international legal order that could be able to guarantee eternal peace that is 'a suspension of hostilities' (TPP, Kant, 1996: p. 317). There had been considerations how to realize a peaceful world through a council of the sovereigns like the *Projet pour rendre la paix perpétuelle en Europe* of the Abbé St Pierre. What Kant did – again under the influence of Rousseau – is to connect the international order with the inner structure of the state. His first 'definitive article for perpetual peace' therefore says that "the civil constitution of every state shall be republican" (TPP, Kant, 1996: p. 322) and the second that "the right of nations shall be based on a federalism of free states" (Kant, 1996: p. 325). This 'surrogate' for a world republic – which seems to be unachievable under current conditions – provides a way to maintain the sovereignty of the participating states with the rule of right, the only alternative to the unacceptable rule of war, "right cannot be decided by war and its favorable outcome" (Kant, 1996: pp. 327f.).

On the other hand, Kant is not a categorical pacifist. He accepts that wars had played an important role in the development of mankind (*Idee zu einer allgemeinen Geschichte in weltbürgerlicher Absicht*, Kant, 1902: ff., VIII, 24ff.), and he would prefer the warlike state of nature to a repressive world monarchy (TPP, Kant, 1996: p. 336). He is not an unrealistic dreamer who expects world peace within a short time. Anyway, as Doyle (Doyle, 1986) and others have shown, it was not an unrealistic assumption that democratically governed states tend to demonstrate much less bellicose behavior toward each other than other forms of government. Kant rather thought, or hoped, that with the help of the "great artist nature" (*die große Künstlerin Natur*, TPP, Kant, 1996: p. 331), human societies would move slowly but steadily toward the federation of free republics guaranteeing eternal peace. Kant's idea seems to be that even immoral actions ultimately lead to an improvement of the legal order without discharging the culprit of his responsibility. Somehow behind the back of the agents who are looking for their personal advantage, the artist nature works in a way like Smith's 'invisible hand' toward this federalism of republics that might even turn into a world republic. Anyway, this historical development prepares conditions under which persons may act morally much more easily than it is for them now in a situation that is closer to the state of nature.

Kant did not consider this "mechanism of nature" (TPP, Kant, 1996: p. 331.) as a natural law. Rather it is a projection of practical reason, a hope we must have because we have a duty to move toward eternal peace. Elsewhere, Kant said, "if we can't prove that a development will take place but can't prove the contrary either, a need of reason (*Vernunftbedürfnis*) gives us the right to expect what seems more rational" (Kant, 1902: ff., VIII, Kleingeld, 1995: p. 94).

See also: Justice: Philosophical Aspects; Kant, Immanuel (1724–1804); Law and Morality: A Continental European Perspective; Law and Morality: An Analytical Perspective; Moral Reasoning in Psychology; Political Theory, History of; Research Ethics: Research; Social Work Ethics; Virtue Ethics.

Bibliography

Allison, H., 2011. Kant's Groundwork for the Metaphysics of Morals. a Commentary. Oxford University Press, Oxford.

Baum, M., 2005. Freiheit und Verbindlichkeit in Kants Moralphilosophie. In: Philosophia Practica Universalis. Jahrbuch für Recht und Ethik 13, 31–43.

Beck, L.W., 1993. Kant's two conceptions of the will in their political context. In: Beiner, R., Booth, W.J. (Eds.), Kant & Political Philosophy. The Contemporary Legacy. Yale University Press, New Haven, CT.

Brandt, R., 1982. Das Erlaubnisgesetz, oder: Vernunft und Geschichte in Kants Rechtslehre. In: Brandt, R. (Ed.), Rechtsphilosophie der Aufklärung. de Gruyter, Berlin.

Byrd, B.S., Hruschka, J., 2010. Kant's Doctrine of Rights. A Commentary. Cambridge University Press, Cambridge, UK.

Doyle, M.W., 1986. Liberalism and world politics. American Political Science Review 80, 1151–1169.

Engstrom, St, 2010. Reason, desire, and the will. In: Denis, L. (Ed.), Kant's Metaphysics of Morals. A Critical Guide. Cambridge University Press, Cambridge, UK, pp. 28–50.

Ertl, W., 1998. Kants Auflösung der 'dritten Antinomie. Alber, Freiburg.

Esser, A., 2004. Eine Ethik für Endliche. Kants Tugendlehre in der Gegenwart. Fromman-Holzboog, Stuttgart-Bad Cannstatt.

Flikschuh, K., 2010. Justice without virtue. In: Denis, L. (Ed.), Kant's Metaphysics of Morals. A Critical Guide. Cambridge University Press, Cambridge, UK, pp. 51–70.
Forschner, M., 1993. Moralität und Glückseligkeit in Kants Reflexionen. In: Forschner, M. (Ed.), Über das Glück des Menschen. Wissenschaftliche Buchgesellschaft, Darmstadt, pp. 107–150.
Goyard-Fabre, S., 1996. La philosophie du droit de Kant. Librairie philosophique J. Vrin, Paris.
Guyer, P., 1998. The possibility of the categorical imperative. In: Guyer, P. (Ed.), Kant's Groundwork of the Metaphysiscs of Morals. Rowman & Littlefield, Lanham, MD.
Guyer, P., 2002. Kant's deduction of the principles of right. In: Timmons, M. (Ed.), Kant's Metaphysics of Morals. Interpretative Essays. Oxford University Press, Oxford, pp. 23–64.
Herman, B., 1993. The Practice of Moral Judgment. Harvard University Press, Cambridge, MA.
Herman, B., 2011. The difference that ends make. In: Jost, L., Wuerth, J. (Eds.), Perfecting Virtue. New Essays on Kantian Ethics and Virtue Ethics. Cambridge University Press, Cambridge, UK, pp. 92–115.
Hills, A., 2009. Happiness in the Groundwork. In: Timmermann, J. (Ed.), Kant's Groundwork of the Metaphysics of Morals. A Critical Guide. Cambridge University Press, Cambridge, UK, pp. 29–44.
Himmelmann, B., 2003. Kants Begriff des Glücks. De Gruyter, Berlin.
Höffe, O. (Ed.), 2000. Grundlegung zur Metaphysik der Sitten. Ein kooperativer Kommentar. Vittorio Klostermann, Frankfurt/Main, Germany.
Hruschka, J., 2004. The permissive law of practical reason in Kant's metaphysics of morals. Law and Philosophy 23, 45–72.
Kant, I., 1902. Kants gesammelte Schriften. Königlich Preußischen Akademie der Wissenschaften ed. Walter de Gruyter, Berlin.
Kant, I., 1996. Practical Philosophy (M. Gregor, Trans.). Cambridge University Press, Cambridge, UK.
Kaufmann, M., 2005. Was erlaubt das Erlaubnisgesetz – und wozu braucht es Kant? In: Philosophia Practica Universalis. Jahrbuch für Recht und Ethik 13, 195–219.
Kersting, W., 1993. Wohlgeordnete Freiheit. Suhrkamp, Frankfurt/Main, Germany.
Kleingeld, P., 1995. Fortschritt und Vernunft: Zur Geschichtsphilosophie Kants. Königshausen & Neumann, Würzburg, Germany.
Kleingeld, P., 2010. Moral consciousness and the 'fact of reason'. In: Reath, A., Timmermann, J. (Eds.), Kant's Critique of Practical Reason. A Critical Guide. Cambridge University Press, Cambridge, UK, pp. 55–72.
Korsgaard, Ch. M., 1996. Creating the Kingdom of Ends. Cambridge University Press, Cambridge, UK.
Nussbaum, M., 2006. Frontiers of Justice. Belknap Press, Cambridge, Mass./London, England.
Potter, N., 2005. Kant on duties to animals. In: Philosophia Practica Universalis. Jahrbuch für Recht und Ethik 13, 299–311.
Ripstein, A., 2009. Force and Freedom. Kant's Legal and Political Philosophy. Harvard University Press, Cambridge, MA/London, England.
Rosen, A.D., 1993. Kant's Theory of Justice. Cornell University Press, Ithaca, NY.
Saage, R., 1994. Eigentum, Staat und Gesellschaft bei Immanuel Kant. Nomos Verlagsgesellschaft, Baden-Baden, Germany.
Schneewind, J.B., 1998. The Invention of Autonomy. A History of Modern Moral Philosophy. Cambridge University Press, Cambridge, UK.
Schönecker, D., 1998. Kant: Grundlegung III. Die Deduktion des kategorischen Imperativs. Alber, Freiburg.
Timmermann, J., 2009. Acting from duty: inclination, reason and moral worth. In: Timmermann, J. (Ed.), Kant's Groundwork of the Metaphysics of Morals. A Critical Guide. Cambridge University Press, Cambridge, UK, pp. 45–62.
Westphal, K., 1992. Kant on the state, law, and obedience in the alleged 'Antirevolutionary writings'. Journal of Philosophical Research XVII, 383–427.
Westphal, K., 2002. A Kantian justification of possession. In: Timmons, M. (Ed.), Kant's Metaphysics of Morals. Interpretative Essays. Oxford University Press, Oxford, pp. 89–109.
Wike, V.S., 1994. Kant on Happiness in Ethics. State University of New York Press, Albany, NY.
Willaschek, M., 2002. Which imperatives for right? On the non-prescriptive character of juridical laws in Kant's metaphysics of morals. In: Timmons, M. (Ed.), Kant's Metaphysics of Morals. Interpretative Essays. Oxford University Press, Oxford, pp. 65–87.
Williams, H., 1983. Kant's Political Philosophy. St. Martin's Press, New York.
Wood, A.W., 1999. Kant's Ethical Thought. Cambridge University Press, Cambridge, UK.
Zotta, F., 2000. Immanuel Kant. Legitimität und Recht. Alber, Freiburg.

Key, Valdimer Orlando (1908–63)

Walter D Burnham, The University of Texas, Austin, TX, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 8080–8082, © 2001, Elsevier Ltd.

Abstract

Valdimir Orlando Key, Jr. was a major figure in the development of modern empirical analysis of American political parties and electoral politics. Among his noteworthy contributions were (1) a comprehensive, magisterial analysis of Southern politics in the USA during its apartheid era shortly before that was brought to an end. The methodology employed involved a combination of 'thick description' based on interviews and other sources in each of the eleven states under study, as well as imaginative graphic use of aggregate electoral data. Key's analysis of the region's influence on national policymaking was a significant contribution, and even more so was his analysis of the nature of one-party regimes and their consequences for democratic politics. Similar constraints on effective partisan representation and integration were explored for the non-Southern states as well some years later. (2) A systematic engagement with the often apolitical, ahistorical (or acontextual) models developed by early national election studies (survey research), and a challenge to their dominant conclusions that voters were largely incapable of fulfilling the roles specified by democratic theory. (3) The first systematic study of critical elections/realignments as major vectors of transformation in the electoral and political system.

Biographical Summary

Valdimir Orlando Key, Jr. was born in Austin, Texas on 13 March 1908. He spent his childhood in the very small West Texas town of Lamesa, later obtaining a BA from the University of Texas at Austin in 1929, and an MA degree at the same institution in 1930. V. O. Key, Jr. then moved to the University of Chicago – then perhaps the premier institution for studying American politics in the United States – and received his doctorate there in 1934. His early professional years were devoted to teaching and, later, to work in Washington for the Social Science Research Council and, subsequently, as part of the technical staff of the National Resources Planning Board during the heyday of the New Deal. Further Washington experience came his way during World War II with a three-year stint at the US Bureau of the Budget (now the Office of Management and Budget). His academic career took him first to Johns Hopkins University and thence to Yale and Harvard Universities, the latter from 1951 until his untimely death in Cambridge, Massachusetts, on 4 October 1963. The long-standing illness to which he eventually succumbed did not prevent him from moving to the very front rank of American political scientists of his generation. The reasons for this fame – a reputational survey of 1965 by A. Somit and J. Tanenhaus placed him then in a clear first position – will be concisely discussed below.

Scientific Contribution

As a Southerner, and at a time when the South was far more a region apart from the rest of the American polity than it is now, Key proved the ideal scholar to provide a comprehensive description and analysis of the South's *apartheid* political order, and its massive influence on national politics, very shortly before its demise. This study, *Southern Politics in State and Nation* (1949), provided a detailed portrait of politics in each on the eleven former Confederate States at the time of writing, as well as analyses of the region's influence on congressional politics; and – an important contribution in its own right – an extended summary statement on the nature and consequences of one-party systems for the representation of public opinion through elections. The craftsmanship is meticulous and innovative, with a mixture of 'thick description' of persons and forces at work based on extensive interviewing, along with the skilful deployment of charts and maps based upon aggregate returns from referenda and primary and general-election returns. More than half a century later, *Southern Politics* still repays the reader's attention.

Quite a lot of Key's basic approaches to the study of American politics, and his personal value commitments, can be found in these pages. First, and obviously, lies the exceptional historic and contemporary importance of 'sectionalism' (major regional differences based on economy, political subculture and experience) in the study of American politics. This point is worth stressing, since with the advent of national scientific surveys of American voters and their behavior in and after the 1960s, this dimension strikingly faded from view in Americanist political science for a protracted period. Secondly, *Southern Politics* underscored the importance of history and historical experience in shaping the confines of today's electoral politics. It is thus perfectly appropriate to view Tennessee politics in the 1940s (and later) in light of the polar oppositions that were displayed in a June 1861 referendum on the state's secession from the Union. This concern to cite present reality in its developmental context was also, however, to run counter to mainstream preoccupations within political science, as the quest for 'scientific regularities' variously led to embracing concepts and models derived from social psychology or economics coming to the fore.

Finally, Key's democratic value commitments, which informed his entire published output across his professional lifetime, are presented in clear view in this magisterial work. Several points about how these views were expressed deserve mention here. Key made no pretension to 'grand theory,' and imported no elaborate conceptual schemes from great Europeans such as Marx or Weber (see Marx, Karl (1818–83);

Weber, Max (1864–1920)). Like many in his and a somewhat earlier generation of American political scientists, his intellectual focus was American to the core, and concentrated on the gap between the democratic promise embedded in the national political culture on one hand, and real-world antidemocratic practices and realities on the other. These latter obviously abounded in the Southern politics he studied, and this gap represented the chief analytic problematic of his lifetime. As he demonstrated in great detail, in the kind of disorganized and exclusionist politics of his South the have-mores profited and the have-lesses lost out as a matter of empirical fact.

At the same time, Key's was a mind which was strikingly subtle and nuanced. Not for him were the sledgehammer blows of Beard (1913), or other notable writers in the American Progressive tradition. Rather, he very carefully spelled out the structure of a given state of political affairs, equally carefully pointed out the vastness of the gap between ideals and empirical reality, and invited his readers to reflect on this gap and how it might be overcome. He was an optimist about the extent this could be accomplished in time – as scholars shaped by the political crises of the 1960s and 1970s often were not. Thus, his hopes for his native region rested to a significant degree on the longer-term effects of industrialization and urbanization on its society and politics. In retrospect, these hopes were to be realized, though only in the context – he was no prophet, after all – of a massive federal intervention in the 1960s, the 'Second Reconstruction,' fuelled by the civil-rights revolution; and, not least, by the USA's foreign-policy requirements in a postcolonial world.

His later major studies displayed his awareness that there were other impediments to the realization of his ideal, the expansion of the electorate's effective role in politics through the instrument of party competition and political choice. In *American State Politics* (1956), Key devoted his attention to several of these impediments: the impact of direct primaries and other regulations of party activities on party organization, including the multiplication of uncontested seats in legislative bodies; what he calls the 'lottery of the long ballot,' i.e., the separate election of a wide range of state office-holders; and the shifting of major state elections away from coincidence with presidential elections and 'national tides' to off years, where, incidentally, voter turnout is nearly always substantially smaller than in presidential years. Key was clearly concerned about the implications of such developments for the integrity of parties as vehicles for individual and collective choice at elections. As **Table 1** below shows on just one of these dimensions, his forebodings have been more than amply realized in the long generation since his death.

Table 1 Insulation effects: states having at least some gubernatorial elections coinciding with presidential elections, 1936–2000

Year	*No. of states*	*Total no. of US reps. in these states*	*Total US reps. in these states (%)*
1936	35	313	72.0
1956*	29	225	51.7
1984	13	58	13.3
2000	11	52	12.0

*Year of publication of *American State Politics*

In his last years, V. O. Key increasingly devoted his attention to the challenges he saw in the development of survey research, both academic and applied. This was to issue in two works, *Public Opinion and American Democracy* (1961) and – posthumously – *The Responsible Electorate* (1966), and a number of articles where criticism was perhaps more pointed. He was vigorously opposed to political-science approaches that appeared to eliminate politics from their analytic schemes. He was alarmed by the potential for manipulating the public that was already (if by today's standards dimly) visible in applied 'market research' by candidates and their media and polling staffs. For, in his view, public opinion ultimately was an 'echo chamber'; and the quality of its output was heavily dependent upon the input provided by political elites and their consultants. Finally, he was convinced that the survey-research portrait of voters as in the main know-nothings (or know-very-littles) drastically underestimated voters' practical political intelligence and capacity for purposive, rational pursuit of political choice in accordance with their interests. On one major point raised in *The Responsible Electorate* – the distinction between interelection 'switchers' and self-Identified partisan independents – his view was later to receive substantial confirmation in a 1981 study of panel surveys by Morris Fiorina (1981). In any event, Key's last work made its own real contribution to the subsequent explosion of 'issue-voting' studies and, eventually, to a much more sophisticated and politically-centered survey-research enterprise.

Assessment of Key's Professional Impact and Current Importance

As the consummate professional that he was, V. O. Key devoted himself to the well-being and improvement of the political-science profession in countless ways. One example, at a time when statistical knowledge within the profession was still pretty much at a primitive level, was his little 1954 volume, *A Primer of Statistics for Political Scientists*. Another was his textbook, *Politics, Parties and Pressure Groups* (Crowell, NY, 1942–1964), which was far the best in its field and extended his teaching to generations of students. He sought to expand the frontiers of knowledge, not only through these and other works, but through myriad contacts with his peers and service to such organizations as the Social Science Research Council and the American Political Science Association.

It may be said that, for two decades or so after his death, Key remained more revered than followed by the political-science mainstream. His was always a sterling example of outstanding craftsmanship in specifying and executing important research projects that provided new insights into the American electoral process; of an enormous mastery of the data of American politics past and present; and a most exceptional capacity to break new ground by using it in novel ways. At the same time, regionalism and history largely fell out of mainstream concerns; his older South rapidly became, or seemed to become, a museum piece; and for a long time, analysis of

aggregate election data seemed to be under W. S. Robinson's 1950 ban on the uses of ecological correlation. Since then, as problematics and professional interests have shifted somewhat, regionalism and historical depth in the explanation of current states of affairs have seen a notable comeback, and a whole field, American Political Development, has emerged into view. At the same time, ecological regression has emerged as a valuable estimation technique for analyzing voter transitions, e.g., during the critical realignment of the 1850s.

In a real sense, Key's work is more significant today than it seemed to be for some years after his death. Nowhere is this more true than in one major area of research by political scientists and historians alike: the study of critical realignment as a major change process across the span of American electoral, and more broadly political, history. As of 1991 (the last formal count), more than 500 books and articles had been devoted to this subject. While some historians had recognized qualitatively the exceptional turning-point significance of such elections as 1800, 1828, 1860, 1896, and 1932. Key was the first to provide an examination of this as an empirically specifiable problem in his truly seminal 1955 article, 'A theory of critical elections' (1955). It is noteworthy that Key never saw fit to explore the subject further, and one can doubt that he would find the elaboration of models in this field always to his taste. But his was the pathbreaking innovation that truly founded a vast research enterprise that is still underway, and has also attracted the attention of non-American researchers as well (Martin, 2000).

Bibliography

Bensel, R.F., 1980. Sectionalism and American Political Development. Madison, WI.

Cummings, M.C. (Ed.), 1988. V. O. Key, Jr. and the Study of American Politics. Washington.

Fiorina, M.P., 1981. Retrospective Voting in American National Elections. New Haven, CT.

Gienapp, W.E., 1987. The Origins of the Republican Party, 1852–1856. New York.

Key, V.O., 1949. Southern Politics in State and Nation. Knopf, New York.

Key, V.O., 1955. A theory of critical elections. Journal of Politics. 17, 3–18.

Key, V.O., 1956. American State Politics. New York.

Key, V.O., 1959. Secular realignment and the party system. Journal of Politics. 21, 198–210.

Key, V.O., Munger, F., 1959. Social determinism and electoral decision: The case of Indiana. In: Burdick, E., Brodbeck, A. (Eds.), American Voting Behavior. Free Press, Glencoe, IL, pp. 281–299.

Key, V.O., 1961. Public Opinion and American Democracy. New York.

Key, V.O., 1961. Public opinion and the decay of democracy. Virginia Quarterly 37, 481–494. Rev.

Key, V.O., 1966. In: Cummings, M.C. (Ed.), The Responsible Electorate. American Political Science Association, Cambridge, MA.

Kleppner, P., 1987. In: Appendix: Realignment theory after Key. Change and Continuity in Electoral Politics, 1893–1928. Westport, CT, pp. 239–250.

Martin, P., 2000. Comprendre les évolutions électorales: La théorie des realignements revisitée. Paris.

Natchez, P., 1985. Images of Voting/Visions of Democracy. New York.

Keynes, John Maynard (1883–1946)

Robert Skidelsky, Warwick University, Coventry, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 8082–8089, © 2001, Elsevier Ltd.

Abstract

John Maynard Keynes (1883–1946), British economist and civil servant. The most influential economist of the twentieth century, he made important contributions to philosophy, political thought, and arts policy. He invented, almost single-handedly, a new branch of economics, macroeconomics, or the theory of what determines output as a whole, as contrasted with microeconomics which studies the forces determining the composition of a given output. He did so in order to destroy the implicit assumption of 'classical' economics that a competitive market economy would always ensure full employment and thus to justify government intervention to maintain full employment. The task seemed the more urgent in the wake of the Great Depression of 1929–1933. The result was The General Theory of Employment, Interest and Money, whose publication in 1936 is conventionally taken to be the start of the Keynesian Revolution. Its main policy fruit was the commitment, till the mid-1970s, by most Western governments, to maintain high and stable levels of employment in their countries. The lack of secure microfoundations for Keynesian macrotheory, as well as the somewhat contentious implications of that theory for the role of governments, has continued to make Keynesian economics controversial; while economic historians have disputed the contribution of Keynesian policy to the 'golden age' of global prosperity running from the 1950s to the early 1970s.

John Maynard Keynes was the most influential economist of the twentieth century. He invented, almost single-handedly, a new branch of economics, macroeconomics, or the theory of what determines output as a whole, as contrasted with microeconomics which studies the forces determining the composition of a *given* output. He did so in order to destroy the implicit assumption of 'classical' economics that a competitive market economy would always ensure full employment; and thus to justify government intervention to maintain full employment. The task seemed the more urgent in the wake of the Great Depression of 1929–33. The result was *The General Theory of Employment, Interest and Money* (hereafter GT), whose publication in 1936 is conventionally taken to be the start of the Keynesian Revolution. Its main policy fruit was the commitment, till the mid-1970s, by most Western governments, to maintain high and stable levels of employment in their countries. The lack of secure microfoundations for Keynesian macrotheory, as well as the somewhat contentious implications of that theory for the role of governments, has continued to make Keynesian economics controversial; while economic historians have disputed the contribution of Keynesian policy to the 'golden age' of global prosperity running from the 1950s to the early 1970s.

His Life

John Maynard Keynes was born at 6 Harvey Road, Cambridge, England on 5 June 1883 into an academic family. He was the eldest son of John Neville Keynes, a logician and economist, and Florence Ada Brown, the daughter of a Congregational divine. He had an outstandingly successful school career at Eton College which was followed by an equally glittering undergraduate one at King's College, Cambridge, where he gained a first-class honours degree in Mathematics, wrote papers on the medieval theologian Peter Abelard and the Conservative political philosopher Edmund Burke, and became President of the Cambridge Union and the University Liberal Club. Of crucial importance to his intellectual and moral formation was his election, in 1902, to the Cambridge Apostles, an exclusive 'conversation' society, where he fell under the influence of the philosopher G. E. Moore, and which brought him the friendship of Lytton Strachey. Moore's *Principia Ethica* remained his 'religion under the surface' for the rest of his life. It taught that the highest forms of civilized life were friendship, aesthetic enjoyments, and the pursuit of truth.

In 1906 Keynes was placed second in the Civil Service Examination, his worst marks being in economics, which he had studied briefly under Alfred Marshall (see Marshall, Alfred (1842–1924)). After 2 years in the India Office, in which he wrote a thesis on probability in his spare time, he started lecturing on monetary economics at Cambridge; in 1909, his thesis won him a fellowship at King's College, which remained his academic home for the rest of his life. His membership of the Bloomsbury Group, a commune of Cambridge-connected writers and painters who lived in the Bloomsbury district of London, dates from the start of his friendship with the painter Duncan Grant, Lyttton Strachey's cousin, in 1908. In 1913 he published his first book, *Indian Currency and Finance,* and served on the Royal Commission on Indian Finance and Currency.

Keynes helped to avert the collapse of the gold standard in the banking crisis of August 1914 which accompanied the outbreak of the first world war. From January 1915 to June 1919 he was a temporary civil servant at the Treasury, showing a notable ability to apply economic theory to the practical problems of war finance. He was against military conscription, and would have been a conscientious objector had his Treasury work not exempted him from military service. When Lloyd George succeeded Asquith as Prime Minister, Keynes (in January 1917) became head of the Treasury's new 'A' Division, set up to manage external finance. He helped to build up the system of Allied purchases in external markets,

while chafing at Britain's growing dependence on American loans and the failure to arrange a compromise peace. Keynes was chief Treasury representative at the Paris Peace Conference in 1919, where he tried unavailingly to limit Germany's bill for reparations, and to promote an American loan for the reconstruction of Europe. His resignation from the Treasury on 5 June 1919 was followed by the publication, in December, of *The Economic Consequences of the Peace*, the book which first brought him international fame. A bitter polemic, informed by both moral passion and economic argument, against the Allied policy of trying to extort from Germany an indemnity which it could not pay, it reflected his revulsion against Lloyd George's leadership in both war and peace, and his fears for the future of European civilization. Unless the Versailles Treaty were drastically revised, 'vengeance, I dare predict, will not limp' (Keynes, 1919, p. 170)

Between the wars, Keynes's life was divided between Cambridge, London, and East Sussex. He was a spectacularly successful investment Bursar of King's College, and despite some major reverses, made about $400 000 for himself over his lifetime – worth $12m. today – out of which he financed a fine collection of pictures and rare books and the building of the Cambridge Arts Theatre in 1935. In London, where he rented a house at 46 Gordon Square, he was, at various times, on the Boards of five investment and insurance companies, the chief one being the National Mutual Life Assurance Company, which he chaired from 1921 to 1937. Between 1923 and 1931 he was chief proprietor and chairman of the Board of the weekly journal, the *Nation and Athenaeum*, contributing regular articles on financial and economic topics. (He remained chairman of the Board of the *New Statesman and Athenaeum* when the two journals merged in 1931.) Between 1911 and 1945, he edited the *Economic Journal*. In the 1920s his ideas on economic policy permeated Whitehall through monthly meetings of the Tuesday Club, a dining club started by his friend and stockbroker, Oswald Falk. In the 1930s, he sought to influence policy through his membership of the Prime Minister's Economic Advisory Council. In 1925, he took the lease of Tilton, a farmhouse in East Sussex, next door to Charleston, where Duncan Grant lived with the painter Vanessa Bell. This move coincided with his marriage to the Russian ballerina, Lydia Lopokova, who gave his life the emotional stability it had previously lacked, and which provided the necessary background for sustained intellectual effort.

In the 1920s, the postwar European inflations, succeeded in Britain by heavy unemployment, formed the background to his two theoretical books *A Tract on Monetary Reform* (1923) and *A Treatise on Money* (1930) which dealt with the causes and consequences of monetary instability, and their remedies. These theoretical exercises were punctuated by two notable polemical pamphlets *'The Economic Consequences of Mr. Churchill'* (1925) and *'Can Lloyd George Do It?'* (1929), written with Hubert Henderson. The first attacked Churchill's decision, as Chancellor of the Exchequer, to put the pound back on the gold standard at an overvalued exchange rate against the dollar; the second was a plea for a large public-works program. Reconciled to Lloyd George in 1926, Keynes attempted to provide the Liberal Party with a social philosophy of the 'middle way' between individualism and state socialism, suitable for an inflexible industrial structure. Regulation of demand, he would later write, was the only way to maintain capitalism in conditions of freedom (Keynes, 1936, p. 381, 1940, p. 123).

The Great Depression of 1929–32, together with technical flaws in *A Treatise on Money*, took Keynes back to the theoretical drawing board. What now seemed to be needed was not an explanation of Britain's 'special problem' of persisting unemployment, but an explanation of how aggregate output could collapse, labor remain unused, and global depression persist in a world in which resources remained scarce. From the autumn of 1931 to the summer of 1935, Keynes worked on a new book of theory, initially entitled *'The Monetary Theory of Production*.' He was helped not just by older economists like Ralph Hawtrey and Dennis Robertson, but by Roy Harrod and a 'Cambridge Circus' of young disciples led by Richard Kahn. There was just one major pamphlet in these years, *'The Means to Prosperity*,' written in June 1933. On a trip to the United States in 1934 to study the New Deal first-hand, he wrote: 'Here, not in Moscow, is the economic laboratory of the world.'

The GT, published in February 1936, tried to demonstrate both that 'underemployment equilibrium' was logically possible, and how monetary manipulation by the central bank combined with an extensive 'socialization' of investment could maintain full employment. The first proposition, in particular, divided the economics profession, since it rejected the 'classical' thesis of a self-equilibrating economy. The publication of the GT confirmed the breach between Keynes and Dennis Robertson and Hubert Henderson, his chief collaborators in the 1920s; at the same time it marked the birth of a 'Keynesian school' of economics led by Richard Kahn and Joan Robinson (*see* Robinson, Joan (1903–83)) at Cambridge, Roy Harrod and James Meade at Oxford, and Nicholas Kaldor at the London School of Economics. In the United States, the GT supplied the younger generation of (mainly Harvard-trained) economists with a theoretical rationale for the New Deal. Keynes himself joined in the fierce controversies which his book generated, even though he was severely incapacitated, from May 1937 to March 1939, with heart disease. As another European war, and with it, the return to deficit-financed full employment became increasingly likely, Keynes sought to win acceptance for his Revolution by showing how the managment of aggregate demand to avert depression could just as easily be used to contain inflation in a war economy.

The upshot was his pamphlet *'How to Pay for the War*' (1940), which won the approval of his arch critic Friedrich Hayek (*see* von Hayek, Friedrich A. (1899–1992)), and whose logic influenced Kingsley Wood's war budget of 1941. Restored to a semblance of health by his doctor, Janos Plesch, Keynes himself returned to the Treasury in August 1940 as unpaid adviser to the Chancellor of the Exchequer, and remained its dominating force for the rest of his life. Elevated to the House of Lords as Baron Keynes of Tilton, his influence was felt in the Beveridge Report on Social Security in 1942, and in the Employment White Paper of 1944 which pledged the government to maintain a 'high and stable level of employment' after the war. In 1942, Keynes became chairman of the Council for the Encouragement of Music and the Arts (CEMA), a wartime innovation which, transformed into the Arts Council of Great Britain shortly before he died, inaugurated permanent state patronage of the arts.

The American demand that, in return for Lend-Lease, Britain scrap its imperial preference system after the war, inspired Keynes to his last great constructive effort, his plan for an International Clearing Union (1942). This was designed to shift balance-of-payments adjustment from debtor to creditor countries, so as to avoid the externally generated deflationary shocks which had spread depression under the gold standard. The Bretton Woods Agreement of 1944, which set up a system of fixed, but adjustable exchange-rates and two new institutions, the International Monetary Fund and the International Bank for Reconstruction and Development, fell short of Keynes's hopes. The abrupt US cancellation of Lend–Lease in August 1945 led him to undertake the fifth of his six Treasury missions to the United States to secure an American grant or interest-free loan of $5bn. Forced to accept a semi-commercial loan for an amount far less than he had requested, Keynes gave a brilliant defence of his policy in his last speech to the House of Lords on 7 December 1945.

On 21 April 1946, worn out by his labors, he suffered a fatal heart attack at Tilton, a little short of his 63rd birthday. In an imposing memorial service at Westminster Abbey, Clement Attlee, the prime minster, headed a list of mourners drawn from all walks of life.

Major Contributions

Keynes's main contributions to economics can be grouped under three heads: the theory of deep economic fluctuations; the theory of effective demand; and the technique of stabilization policy based on the second.

The Theory of Economic Fluctuations

This cannot be wholly understood apart from Keynes's distinctive theory of probability. In his *A Treatise on Probability* (1921), his main contention was that probability is not statistical, but is a logical relationship between the premise and conclusion of an argument which fell short of demonstrative proof. Like Frank Knight (who however used a statistical theory of probability), Keynes distinguished between risk, or situations when agents can form probabilities and thus insure against loss, and uncertainty, when 'there is no scientific basis on which to form any calculable probability.' He wrote in 1937: "Thus the fact that our knowledge of the future is fluctuating, vague and uncertain, renders wealth a peculiarly unsuitable subject for the methods of the classical economic theory' (Keynes, 1937, p. 113). This may be contrasted with the modern mainstream view that agents maximize with known probability distributions, so that market-clearing models are the most appropriate for analyzing economic behavior. Keynes himself identified as a tacit axiom of the classical theory of the self-regulating economy that 'at any given time facts and expectations were ... given in a definite and calculable form' (Keynes, 1937, p. 112).

Keynes's theoretical work in the 1920s is devoted to unpackaging the Quantity Theory of Money (QTM) which ruled out, by assumption, the possibility that monetary disturbances can have 'real' effects on the economy. This might be true in the long-run, Keynes remarked acidly, 'but in the long-run we are all dead' (Keynes, 1923, p. 65). Keynes used the quantity of money *equation* to show that changes in the quantity of money can unsettle business expectations and the distribution of income, thereby causing short-run fluctuations in the level of business activity. In doing so, he emphasized the use of money as a 'store of value' or hedge against uncertainty, and the distinction between fixed and flexible prices. Thus in the *Tract on Monetary Reform* he wrote: 'The *fact* of falling prices injures entrepreneurs; consequently the *fear* of falling prices causes them to protect themselves by curtailing their operations' (i.e., adding to their cash balances) (Keynes, 1923, p. 34). In the *Treatise on Money*, he argued that the disturbing effects of money can arise not just from inflationary or deflationary policies, but from *autonomous* changes in business sentiment. Much of the *Treatise* is devoted to analyzing the consequences of short-run changes in the 'propensity to hoard' or 'velocity of circulation,' which Keynes identifies, by means of his 'Fundamental Equations,' with the emergence of windfall profits and losses, and a disequilibrium between saving and investment. Keynes shows how disequilibrium prices produce an oscillation of boom and slump round a (notional) full-employment equilibrium. But such equilibrium is achieved accidentally, in the absence of appropriate monetary policy.

The chief theoretical point which emerges from these two works is the instability in what Keynes would later call the liquidity-preference schedule. His claim was that, in a variety of situations readily encountered in economic life, hoarding money or savings can yield superior utility to investing them, and that this, combined with 'sticky' wage and price contracts, can explain quite prolonged lapses from full employment. His contribution up to 1930 was largely a summation of the monetary theories of Irving Fisher (see *Fisher, Irving* (*1867–1947*)) and Knut Wicksell, and is properly regarded as a contribution to the theory of economic fluctuations, with increased emphasis on volatile expectations. Like Fisher, Keynes advocated a monetary policy of keeping domestic prices stable; this might be incompatible with adherence to the international gold standard, an institution which Keynes dismissed as a 'barbarous relic' (Keynes, 1923, p. 138).

The Theory of Effective Demand

Keynes's revolutionary contribution to economics lies not in his theory of business fluctuations but in his theory of how aggregate output is determined. This is the subject matter of his GT. The QTM was a theory of the price level, when the levels of output and employment were explained in real, nonmonetary terms. Following a suggestion from Ralph Hawtrey, Keynes agreed in 1930 that "it will probably be difficult in the future to prevent monetary theory and the theory of short-period supply from running together" (Keynes, 1930, p. 146). This was the start of macroeconomics.

The core of the GT is the theory of effective demand. It tells how, when something happens to reduce aggregate demand, aggregate supply and aggregate demand are reconciled, not through falls in the price-level of output, but by a reduction in its quantity. It is a short-period theory: prices adjust only after, and as a result of, the fall in output and employment, and even then not completely. In the short-run, the volume of output in

an economy is determined by the money demand for that output. The GT is the theory of what determines the level of money demand for output.

In Keynes's notation $Y = C + I$. This tautology simply means that National Income (Y?) is equal to the sum of the expenditure on National Output, divided into consumption spending (C) and investment spending (I). Consumption expenditure is governed by Keynes's 'consumption function,' based on the 'psychological law' that consumption (and therefore saving, which is $Y - C$) is a stable fraction of total income, and that the marginal propensity to consume is less than unity. In a progressive economy the gap beween consumpion and production must be filled by investment if full employment is to be maintained. The amount of investment is determined by the relationship between what the investment is expected to yield over its lifetime ('the marginal efficiency of capital') and the cost of borrowing the money needed to finance it (the rate of interest). Both are subject to uncertain expectations. Consumption is the stable element in demand; investment the volatile element. Thus, what the community wishes to save may at a particular moment exceed what businessmen want to invest. The revolutionary claim of the GT is that when this happens, the divergent plans are brought into equality by a reduction in the level of income (output). Moreover, if the consumption function is known, it is possible to show, by a mathematical formula (Kahn's multiplier), by how much the community's income or alternatively employment must change to reconcile divergent saving and investment plans. The novelty of Keynes's treatment, as compared not just with the older theory, but his own *Treatise on Money*, was to demonstrate that an economy might be in equilibrium, with saving equal to investment, and aggregate demand equal to aggregate supply, at less than full employment.

Critics fastened on the most contentious feature of the new doctrine, the theory of 'under-employment equilibrium.' They denied that this was a genuine equilibrium; it was a disequilibrium state, frozen by the assumption that certain prices were fixed. Flexible interest rates and wage rates would always ensure the desired level of employment, whatever was happening to aggregate demand.

Keynes could have defended his position by saying that his fixed-price assumptions were realistic but he set out to expose the logical flaws underlying the implicit classical price-adjustment story. He denied, first of all, that the interest rate was, or could be, determined in the market for saving and investment, since the quantity of saving depends on the level of current income. This was the basis of his view that it is changes in income that equilibrate saving and investment plans. In place of the classical theory of interest, he offered his liquidity-preference theory of interest, which makes interest the price for giving up cash. The greater the liquidity-preference of wealth-holders, the higher the yield they will demand for switching from cash into bonds or other securities. Insofar as the rate of interest *does* move in response to a fall in investment demand, it will be in the wrong direction, since the "same circumstances which lead to pessmistic views about future yields (of capital assets) are apt to increase the propensity to hoard" (Keynes, 1937, p. 118).

As far as wage flexibility is concerned, Keynes agreed with the orthodox view that the *real* wage is inversely related to the quantity of employment, but denied that real wages are determined in the labor market by money-wage bargains, since an all-round reduction in money-wages might, by reducing the general price level, leave the real wage unaffected. The effect of a 'flexible wages policy' thus depended entirely on its impact on the components of aggregate demand-consumption and investment – which might well be unfavorable.

Keynes conceded that, in time, the decline in employment (or money-wage rates) would, by decreasing the amount of money required to satisfy the business demand, lower the rate of interest required to satisfy the "propensity to hoard" (Keynes, 1937, p. 118); he also thought that, after a slump, a shortage of capital goods would develop which would revive profit expectations" (Keynes, 1936, pp. 317–318). In short, a slump might eventually bring about the price adjustments which according to the classical economists were supposed to prevent the slump from occurring. He doubted, though, whether even longer-run forces would ensure a gravitational pull toward full employment. 'We oscillate,' he wrote, "round an intermediate position, appreciably below full employment and appreciably above the minimum employment a decline below which would endanger life" (Keynes, 1936, p. 254).

Despite nonacceptance by the leading economists of key elements in Keynes's macroeconomic model (Dennis Robertson, in particular, made a damaging attack on his theory of interest), critics were driven to argue for their positions within the analytical framework Keynes had set up. Macroeconomic analysis became part of the economists' tool kit, even if Keynes's own conclusions were rejected.

Stabilization Technique

Keynes's three main theoretical books were all aimed at providing justifications for conscious attempts to stabilize output and employment at a high level. The quest for output stabilization led directly to the development of National Income and Expenditure Accounts, to enable governments to estimate the size of the 'output' gap – the gap between what the economy was producing and what it could produce at full employment (or, in later refinement, what it could produce when it was growing to trend) – which needed to be plugged by extra spending. Keynes also insisted that the aggregate demand and supply framework of the GT could be used to work out how much spending needed to be withdrawn from an economy to prevent inflation if aggregate demand exceeded aggregate supply.

In principle, stabilization policy can be either monetary or fiscal, or some combination of both. In practice, Keynes rejected the monetary route to managing demand. He did not deny that investment was interest-elastic. But he thought that interest-rate changes, produced by varying the quantity of bank credit, were too slow-moving in their effects on investment; and he was worried that if interest rates were raised to check a boom it would be difficult to bring them down to check a slump, since the expectation of falling rates would increase the 'propensity to hoard.' So he advocated a policy of permanently cheap money, to be buttressed by capital controls and 'under-funding' of the National Debt. This left fiscal policy as the main instrument for preventing both slumps

and inflationary booms. Keynes's fiscal philosophy must be distinguished from Keynesian fiscal policy as understood (and sometimes practiced) after his death. It was based on a sharp distinction between the government's 'ordinary' or current-account budget and what he called the 'capital' budget, made up not just of government spending on capital projects but the economy's total investment. The government's ordinary budget should be balanced at all times, with a surplus earmarked for debt repayment in 'normal' times. It was the task of the capital budget to balance the 'national accounts.' Keynes believed that a great deal of investment was already 'socialized,' that is, subject to public policy even if it was technically private. As he put it in 1943: "If two-thirds or three quarters of total investment is carried out or can be influenced by public or semi-public bodies, a long-term programme of a stable character should be capable of reducing the range of fluctuation to much narrower limits than formerly ... If this is successful it should not be too difficult to offset small fluctuations by expediting or retarding some items in this long-term programme" (Keynes, 1943, p. 322).

Finally, Keynes attached only minor importance to exchange-rate adjustments. His Clearing Union Plan (1942) provided for fixed, but adjustable rates, and this was a feature of the Bretton Woods Agreement (1944). Keynes's attitude – like that of most economists of his day – was governed by price-elasticity pessimism. He preferred countries to use, if necessary, capital controls to protect their balance of payments in the framework of a monetary regime of fixed exchange rates, low tariffs, and automatic creditor lending through international institutions.

Keynes was not an ivory-tower theorist. His theorizing was controlled by real world events. His own investment experience enabled him to identify the 'speculative' motive for holding money; his sense of political fragility led him to concentrate on the economics of stabilization; his civil service experience enabled him to turn theories into workable plans. His economic theorizing was less directly, but still importantly, controlled by his philosophical beliefs, particularly by his conception of the good life and of the conditions of just exchange.

The Legacy

Economics was transformed by its encounter with Keynes. Macroeconomics became for many years a dominant part of the subject; macroeconomic forecasting the main tool of government policy. However, Keynes's doctrines never won universal acceptance, and key aspects of his theoretical and policy legacy have been challenged. The debated issues can be grouped under four heads.

Acceptance of Keynes's Theory

What Keynes bequeathed was not the same as what was accepted. The first theoretical breach came with demonstrations by Pigou (1942) and by Modigliani (1944) that the Keynesian slogan 'quantities adjust, not prices' was true only if money wages were rigid. This became the basis of the 'neoclassical synthesis,' which grafted Keynesian macroeconomics on to classical theory, but left the rigidities unexplained. Milton Friedman's counter-attack was both methodological and theoretical. On the one hand, he argued that it was not acceptable to posit ad hoc supply and demand functions. Second, his own application of neoclassical standards of method to Keynes's aggregate equations – seen in his permanent income hypothesis (1957), the stable demand for money function (1956, 1963), and the theory of the 'natural rate' of unemployment (1968) – undermined the case for Keynesian stabilization policy. Economies were more cylically stable than Keynes had supposed; multipliers were small, or nonexistent; government manipulation of aggregate demand had no permanent 'real' effects, but only raised the inflation rate. These 'policy ineffectiveness' propositions were to be hardened still further by the 'rational expectations' school of Robert Lucas and Thomas Sargent. The tendency of Friedman's critique (popularly called 'monetarism') was to reinsert an updated version of the Quantity Theory of Money into the heart of macroeconomics. It revived the pre-Keynesian notion (adumbrated by Keynes himself in the *Tract on Monetary Reform*) that the most important macroeconomic function of governments was to keep stable the purchasing power of money.

Acceptance of Keynesian Policy

Contrary to widespread mythology, this was patchy. Both the British government and the US Administration committed themselves to targetting 'high' levels of employment, but it is often asserted that US policy only became Keynesian in the early, and German policy in the late, 1960s, and that both Keynesian episodes were fairly brief. Much of this discussion begs the question of what one means by 'Keynesian' policy. Running a budget deficit is no more a sign of Keynesian virtue than is running a budget surplus in boom conditions anti-Keynesian. An alternative argument is that Keynes's influence was exerted not so much through national policies as through the US willingness to provide the rest of the world with reserves and liquidity. However, this likewise begs the question of how 'Keynesian' this willingness was.

The Impact of Policy on Events

For a long time, the canonical view was that Keynesian demand-management policies and Keynesian-inspired institutions (the Bretton Woods system) were mainly responsible for the uniquely successful employment and growth performance of most countries in the 1950s and 1960s. Today most of the credit for the 'golden age' is given to opportunities for 'catching up' with American technology, recession-proof military spending by the US, and high levels of 'social' spending.

In the 1950s and 1960s it was common to argue that Keynesian policy helped to save capitalism by removing the scourge of mass unemployment. (In Marxist terms, it legitimized the capitalist order.) By the 1970s it was being argued that it endangered the long-run survival of capitalism by producing rising inflation, an expanding public sector, and increasingly draconian wage and price controls.

The Current Debate Over Keynesian Economics

According to the monetarist-cum-rational expectations schools, Keynesian economics failed the predictive test: it led to

inflation, still worse 'stagflation.' In the 1970s, Keynesian policies were attacked for ignoring the existence of a 'natural' rate of unemployment, and (by the Virginia or Public Choice school) for assuming that politicians wanted to maximize the collective social welfare, rather than their own individual utilities. Taken together, these two attacks offered a forceful argument against the use of discretionary fiscal and monetary policy to balance economies.

The use of models of economies with nominal rigidities is still general, but whether these rigidities are to be taken as given is much more questioned than in 1950s and 1960s. The dominant 'supply side revolution' of the 1980s chiefly was concerned to dissolve rigidities seen as institutional by deregulating labor markets. Against this, the 'new Keynesians' explained how sticky prices are rational because of transactions and information costs, and how shocks to demand can destroy both physical and human capital. These explanations seemed both to strengthen and weaken the case for Keynesian macroeconomic policy. On the one hand, they gave renewed intellectual respectability to stabilization policy. On the other hand, by explicitly introducing inflation into their analyses, they conceded the existence of a rate of unemployment (theso-called NAIRU or Non-Inflation Accelerating Rate of Unemployment) below which unemployment could not be pushed by manipulating demand. Finally, against the mainstream profession's use of Bayesian statistics and decision theory to model agents' behavior, a minority school of 'Post-Keynesians' continues to assert the fundamental nature of Keynes's attack on the rationality axiom.

An interim judgment on the Keynesian Revolution would be that the main body of classical economics was too well-entrenched to be overthrown by the frontal assault he mounted. The notion that uncertainty was at the heart, rather than at the financial margins, of economic processes, proved too subversive of the science economics claims to be, to be acceptable. Most economists would say today of Keynes what Marshall said of Jevons: "His success was aided even by his faults ... he led many to think he was correcting great errors; whereas he was really only adding important explanations." Whether this will be the final verdict is still questionable.

Bibliography

Davidson, P., 1992. Would Keynes be a new Keynesian? Eastern Economic Journal 18 (4), 449–463.

Harrod, R.F., 1951. The Life of John Maynard Keynes. Macmillan, London.

Hicks, J.R., 1974. The Crisis in Keynesian Economics. Blackwell, Oxford, UK.

Keynes, J.M., 1919. The Economic Consequences of the Peace, vol. 2. Collected Writings of John Maynard Keynes. Macmillan and Cambridge University Press, London.

Keynes, J.M., 1923. A Tract on Monetary Reform, vol. 4. Collected Writings of John Maynard Keynes. Macmillan and Cambridge University Press, London.

Keynes, J.M., 1930. The General Theory and After: Part I Preparation, vol. 8. Collected Writings of John Maynard Keynes. Macmillan and Cambridge University Press, London.

Keynes, J.M., 1936. The General Theory of Employment, Interest and Money, vol. 7. Collected Writings of John Maynard Keynes. Macmillan and Cambridge University Press, London.

Keynes, J.M., 1937. The General Theory and After: Part II Defence and Development, vol. 14. Collected Writings of John Maynard Keynes. Macmillan and Cambridge University Press, London.

Keynes, J.M., 1940. Activities 1939–1945: Internal War Finance, vol. 22. Collected Writings of John Maynard Keynes. Macmillan and Cambridge University Press, London.

Keynes, J.M., 1943. Activities 1940–1946: Shaping the Post-War World, vol. 27. Collected Writings of John Maynard Keynes. Macmillan and Cambridge University Press, London.

Leijonhufvud, A., 1969. Keynes and the Classics: Two Lectures on Keynes' Contribution. IEA, London.

Moggridge, D.E., 1992. Maynard Keynes: An Economist's Biography. Routledge, London.

O'Donnell, R.M. (Ed.), 1991. Keynes as Philosopher-Economist. Macmillan, London.

Patinkin, D., 1976. Keynes's Monetary Thought: A Study of its Development. Duke University Press, Durham, NC.

Schumpeter, J.A., 1952. Ten Great Economists: from Marx to Keynes. Allen and Unwin, London.

Skidelsky, R., 1983. John Maynard Keynes: Hopes Betrayed 1883–1920. Macmillan, London.

Skidelsky, R., 1992. John Maynard Keynes: The Economist as Saviour 1920–1937. Macmillan, London.

Skidelsky, R., 1995. Keynes and the quantity theory of money. In: The Quantity Theory of Money from Locke to Keynes. Edward Elgar, Cheltenham, UK.

Skidelsky, R., 2000. John Maynard Keynes: Fighting for Britain 1937–1946. Macmillan, London.

Walters, A., 1987. Friedman, Milton. The New Palgrave Dictionary of Economics. Macmillan, London.

Keynesianism and the Question of State Interventionism

Robert Salais, IDHE, Cachan, France.

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 8089–8092, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Keynesianism assigns to the State the responsibility to regularly intervene in the operation of the economy so as to maintain its performance at a high level. It stresses the fine-tuning of short-term fluctuations by discretionary policies. It was not born with Keynes (who probably would be hostile to it), but with macroeconomics, national accounting and statistics in the post-war period. Stagflation of the 1970s was its downfall. Today paradoxically the attacks from the 'rational anticipations' school converge with new readings of Keynes. A government of the economy, limited in scope and in ambition, is permissible via a rule-based macroeconomics.

This reprint from the 1st ed offers the view on Keynes and the so-called Keynesianism as it was held before the recession of 2008 and the following years. Therefore the chapter is a historical document and at the same time proof that the knowledge production in economics and other social sciences is not always in accord with developments just around the corner. Even former opponents of Keynes started to think differently about his role and the appropriateness of an economic policy which follows some of his advices.

Keynesianism, whatever its variants, assigns to the State the responsibility to regularly intervene in the operation of the economy so as to maintain its performance at a high level of employment and growth (Hall, 1989). Keynesianism believes in the effectiveness of this intervention. From Keynes' analysis of the preference for liquidity and the economic crisis of the 1930s it has retained a definite skepticism concerning the effectiveness of monetary policy. Investment and consumption are held to be little affected by interest rates; it therefore suffices to have a monetary policy that maintains interest rates at a sufficiently low level. On the other hand, Keynesianism stresses counter-cyclical management of overall demand by means of fiscal and budget policies: financing expenditures by public borrowing in case of recession; seeking a budget surplus to hold down inflation in the rising phase of the cycle. Fiscal constraint can be relaxed to finance welfare programs. This is a decisive change with respect to classical economics that deemed a balanced budget to be the highest priority. The Keynesian stimulus works via a sort of self-fulfilling prophecy concerning investment and demand. Prewar policy held to the contrary that the modest means available to the government was a guarantee for democracy, allowing it to stand up to populist excesses and to unlimited demands from special-interest groups.

Keynes would have been hostile to Keynesianism. In all likelihood he would have espoused, although in different terms, the criticism expressed by Friedman (1968) and Lucas (1976) towards the Keynesians' conception of State intervention in the economy. But this criticism, albeit radical (and essentially well taken), is aimed less at the principle than at the modalities of an effective State intervention. Keynesianism, for all its '*naïveté*,' has managed to make this intervention accepted. This is the paradox that ensures a long life for intellectual and political debate about Keynes' writings. This contribution is divided into three sections: a history of the reception encountered by Keynesian ideas: criticism of these ideas, and the relevance of Keynes today.

The Reception of Keynesian Ideas

Keynesianism was not born with Keynes. Hicks (1937) notably formalized the IS-LM curves that jointly determine the global income and the rate of interest, via the relations between investment, savings, demand and supply of money. Samuelson (1948) provided the analytical frameworks of macroeconomic policy. Put briefly, the synthesis thus established, neoclassical in spirit, resituates Keynes' work in the conventional framework of market equilibrium. It eliminates the role of agents' expectations, which was decisive for Keynes (as well as eliminating the role of the 'state of confidence'), and replaces it with a mechanistic vision of the economy that is reduced to a few global parameters. Phillips (1958) then added the curve linking wage increases, rising prices and the unemployment rate. The development of national accounting and statistical apparatus provided the basis for econometric modeling. Little by little economists, following in this respect the theory of economic policy elaborated by Tinbergen (and drawing politicians and political discourse in their wake), came to espouse the use of macroeconomic models for the quantitative evaluation of State policy. Keynesian policy is fundamentally a policy designed to fine-tune short-term fluctuations in the economy. If short-term forecasts given by a model are accurate, then the long-term evaluation of the effects of a policy is presumed to be reliable. The long-term outcome is engendered by a carefully constructed chain of short-term periods. Keynes thought, on the contrary, that the agents' view of the long term had a direct effect on their current behavior.

From the 1930s to the 1950s the reception of Keynesian ideas spread through three channels: the influence of professional economists' expertise on political decision making (seeking new explanations for the economic crisis); new demands formulated by political parties (themselves conditioned by previous experience); and the capacity of Keynesian ideas to spawn broad political coalitions. The configuration of history fostered the convergence of these three channels: the heritage of the war economy (when the 'young' Keynesians had

demonstrated the effectiveness of State organization and intervention); the needs of reconstruction (which called for a high level of public investment), and the expansion of social policies aimed at achieving full employment. In the light of the prewar risk of confrontation between the labor agenda (leading to nationalization of the means of production) and that of capitalism (*laissez-faire* liberalism), Keynesianism offered a compromise. Full employment and sustained growth could be achieved in a capitalist economy through the judicious use of macroeconomic management (without interfering with the management prerogatives of private capital). Labor could see full employment guaranteed, and capital could continue to control investment.

But in order for Keynesian policy to be effective two structural conditions were necessary: a system of appropriate institutions; and a common language and beliefs shared among the actors in the economy. These conditions were fulfilled to various degrees depending on the country – more highly developed in Great Britain and France though differently, than in the United States or Germany. The State must be linked to the actors by a set of institutions that structure the flow of information and resources. These institutions ensure a common assessment of the economic situation and provide security for anticipations (notably via property rights and contract law). A relationship of confidence between the central bank, the State, and financial markets is crucial for the financing of public debt and establishing regulation of the banking system. At the same time, shared views of the expected and legitimate role of the State must spread throughout the economy and society. This 'conventional' foundation creates the state of confidence enabling each actor to constitute favorable anticipations (of demand and investment). The combination of these anticipations is self-fulfilling of the forecast projected by the macroeconomic model; it allows State intervention to attain its goals. This double system of institutions and shared conventions paradoxically ensures the effectiveness of Keynesian policy. Without this system, Keynesian policy is nothing. This is what the Keynesians forgot. They consequently come to believe in the State as a *deus ex machina*. In neglecting the question of expectations they left the way open for the school of rational anticipations. The return in force of market economics and neo-classical rationality was the outcome.

Criticism of Keynesianism

A series of attacks, some old (Hayek, Buchanan) and others recent or renewed (Friedman, Lucas, Sargent), converged in the late 1960s and early 1970s, leading to a collapse of Keynesianism.

Hayek and Buchanan had paved the way, one by rejecting any sort of validity for macroeconomics in itself, the other by demolishing the myth of the State as an enlightened despot. Keynesianism opens the floodgates of inflation by facilitating uncontrolled creation of credit. Then come wage and price control policies that eliminate inflation at the cost of much higher unemployment than that which would have resulted under a stable currency regime (Hayek). Economics can be seen as a science that gives people greater control over their environment, but it has no command over politics. Governments are more interested in collecting the maximum number of votes than in pursuing the public good (Buchanan).

The Phillips relationship had given Keynesian policy its missing link: that of setting wages. Fine tuning efforts could focus on the inflation/unemployment dilemma: how much unemployment is needed to ensure price stability? The stagflation of the 1970s (high unemployment and high inflation occurring simultaneously) was its downfall. Friedman reintroduced anticipations. He showed that the trade-off between inflation and unemployment is only temporary: it stems from inflation that is not anticipated by agents. Eventually unemployment returns to its natural rate, which is determined by structural socioeconomic characteristics, but at the price of higher inflation. Discretionary monetary policy cannot durably stabilize either interest rates or unemployment at low levels. Keynesians situate themselves in a world where everyone anticipates that prices will remain stable, and where this anticipation persists no matter what happens to prices and wages. This world has vanished. Friedman updated his 1948 program: replacing a discretionary monetary policy with public adoption by the monetary authorities of a set rule. In the present instance this means: setting a growth target for the money supply (using a specified monetary aggregate); avoiding sudden changes, not overreacting to events. In 1948 this 'monetarist' framework had received no attention. Economists had adopted the highly flattering position of 'advisor' to the managers of the economy (and thereby justifying the development of their profession). The contrast between the massive unemployment of the Great Depression and the success of the post-war economy had persuaded the economy's managers of the efficacy of systematic State intervention. 'Friedmanian' rules are now widely followed by monetary authorities (like the Federal Reserve Bank or the European Central Bank).

Lucas's criticism was aimed at the use of macroeconometric models. The theory of economic policy (Tinbergen, 1952) underlying this use is invalidated by the fact that actors adapt their rules of decision to modifications in economic policy. The variables controlling economic policy are thus endogenous to the system; the structure of the models varies with economic policy. No reliable projection of future behavior and trends can be obtained by extrapolation of series and correlation with the past. In order for a change in the rules to have a chance of being effective (i.e., leading to foreseeable trends) it must be publicly debated and understood by the actors in the economy. Once again, despite a few pockets of resistance, this criticism ultimately dealt a fatal blow to substantive macroeconomics 'defined by itself to be the body of expertise … devoted to the development and refinement of forecasting and policy evaluation methods' (Lucas, 1980, p. 250). This discipline failed to foresee the return of economic growth at the dawn of the twenty-first century. But this has not prevented the teaching of macroeconomics to continue in academia.

Now that the din of the dogmatic battles has ceased, a more balanced assessment is possible. The structuring role of the anticipations of agents is now recognized. They introduce an irreducible element of unpredictability. What counts in the dynamics of the economy is action and coordination.

Economic policy must abandon the aim of influencing the course of events by substantial and discretionary means. Its field of action relates to the rules that serve as the framework of reference for private forecasts and actions. At bottom, for these rules to be effective they must be debated and made public. They leave room for interpretation in action and for innovation. They provide a framework that can be publicly revised and adjusted in keeping with the results obtained. This does not imply, however, an absence of concern for joblessness and full employment, but rather a clearer consciousness of the properties of various policy instruments and how they complement each other. The unemployment rate can no longer be a direct target of economic policy. Measuring an unemployment rate under full employment is meaningless. Firms and employers have many options other than unemployment at their disposal for adjusting to fluctuation in the demand for products and to the outlook for innovation. Satisfactory allocation of social risks is the task of welfare economics. Providing a framework of stability and security that safeguards the anticipations and terms of agreements concluded – this should be the essence of economic policy (Lucas, 1976). But can this modest objective satisfy the body of professional economists, who are much more numerous than in the past and permanently accustomed to the comforts and remuneration of policy expertise?.

The Relevance of Keynes Today

Keynesianism is dead, but what of the writings of Keynes himself? In his ferocious criticism of Tinbergen in 1939 Keynes already expressed the arguments that were later expanded by Lucas (who did so without making any reference to the writings of Keynes). This criticism originates in Keynes's *Treatise on Probability* (1921). As early as this work he understood that the problem is a double one, that of measurement and that of 'passing from statistical description to inductive generalisation' (Keynes, 1939, p. 566). If all causal factors must be measurable for economic policy to be subjected to statistical tests and forecasting, 'it withdraws from the operation of the method (of Tinbergen) all those economic problems where political, social and psychological factors, including such things as government policy, the progress of invention and the state of expectation, may be significant' (1939, p. 561). Keynes was likewise prudent concerning the arithmetic of the multiplier. Lastly, for Keynes the prime target of monetary policy was the value that actors accord to the conventions that they follow on financial markets. It is the state of confidence that is important, and not directly interest rates and unemployment.

'Neo-Keynesians' have counter-attacked. They integrated into the Keynesian scheme rational anticipations and the possibility of exogenous shocks and of constraints on individual supplies and demands. But in so doing they remain open to these criticisms of Keynes. And by insisting on the rigidity of costs, their recommendation of wage flexibility paradoxically converges with that of liberals attached to the idea of reducing the natural unemployment rate. The 'post-Keynesians' (Davidson, 1991) insist on the radical and non-probabilistic uncertainty that surrounds the decision-making of economic agents. They thus move closer to the institutional (and not solely political) content of Keynes' message. They attempt to express his relevance to the current context. This uncertainty stems from an instability in the overall trajectory of the economy. Institutions must be created to construct a 'corridor' that restricts the range of directions that actions can take. Following Keynes' intuition on inference and probability, it should probably be added that institutions allow people to make judgement and inference on situations of action which are well-founded and secure.

It is time to leave behind the paralysis of opposing schools and dogmatic assumptions (in particular regarding States vs. the market, or the nature of the market, or rationality). Too often a worthwhile idea is rejected, on the pretext that it comes from the opposing school. Little by little a macroeconomics of rules is expanding in which government of the economy by the State is permissible (a posthumous victory for Keynesianism). But, as thought Keynes, this government must be limited in scope and in ambition (Skidelsky, 1999). The dynamics of the economy are above all the result of the actions, anticipation and coordination of the actors. These actors are diverse in terms of their identities, their means and ends (and in the relationship they establish between means and ends). The knowledge they have acquired of their terrain renders their action irreplaceable in a context of uncertainty. A framework of democratically debated, stable and foreseeable rules seems necessary for the implementation of collectively effective actions. But the debate is shifting to the status of these rules (are they prescriptive rules, or guidelines open to interpretation in action?) and to the status of macroeconomic variables (objectives, constraints, or simple by-products of market activity?).

See also: Culture and Economy; Keynes, John Maynard (1883–1946); Monetary Policy Since the 2007–2009 Financial Crisis.

Bibliography

Davidson, P., 1991. Is probability theory relevant for uncertainty? A post-Keynesian perspective. Journal of Economic Perspectives 5, 129–143.

Friedman, M., 1968. The role of monetary policy. American Economic Review 58, 1–17.

Hall, P.A. (Ed.), 1989. The Political Power of Economic Ideas. Keynesianism Across Nations. Princeton University Press, Princeton, NJ.

Hicks, J.R., April 1937. Mr Keynes and the 'Classics': a suggested interpretation. Econometrica 147–168.

Keynes, J.M., 1921. A Treatise on Probability. Macmillan, London (1973 reprinted in The Collected Writings of J. M. Keynes, Vol. VIII. Macmillan, London).

Keynes, J.M., 1936. The General Theory of Employment, Interest and Money. Macmillan, London (1973 reprinted in The Collected Writings of J. M. Keynes, Vol. VII. Macmillan, London).

Keynes, J.M., 1939. A review of A Method and its Application to Investment Activity by J. Tinbergen (1939 Statistical Testing of Business-cycle Theories, tome I, League of Nations, Geneva). The Economic Journal 49, 558–568 (1973 reprinted in Collected Writings. Macmillan, London, Vol. XIV).

Leijonhufvud, A., 1968. On Keynesian Economics and the Economics of Keynes – A Study on Monetary Theory. Oxford University Press, Oxford, UK.

Lucas, R.E., 1976. Econometric policy evaluation: a critique. In: Lucas, R.E. (Ed.), 1981 Studies in Business-Cycle Theory. MIT, Cambridge, MA.

Lucas, R.E., 1978. Unemployment policy. In: Lucas, R.E. (Ed.), 1981 Studies in Business-Cycle Theory. MIT, Cambridge, MA.

Lucas, R.E., 1980. Rules, discretion and the role of the economic advisor. In: Lucas, R.E. (Ed.), 1981 Studies in Business-Cycle Theory. MIT, Cambridge, MA.

Phillips, A.W., 1958. The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. Economica 25, 283–299.
Posner, Richard A., 2009. 'How I Became a Keynesian': Second Thoughts in the Middle of a Crisis. In: The New Republic September 23, 2009.
Posner, Richard A., 2009. A Failure of Capitalism: The Crisis of '08 and the Descent Into depression. Harvard University Press, Cambridge, Mass.
Samuelson, P.A., 1948. Economics. MIT, Cambridge, MA.
Skidelsky, R., 1999. The conditions for the reinstatement of Keynesian policy. In: Pasinetti, L.L., Schefold, B. (Eds.), The Impact of Keynes on Economics in the 20th Century. Edward Elgar, Cheltenham, UK.
Skidelsky, Robert, 2009. Keynes: The Return of the Master. Allen Lane, London.
Tinbergen, J., 1952. On the Theory of Economic Policy. North-Holland. Amsterdam.

Kimura, Motoo (1924–94)

James F Crow, University of Wisconsin, Madison, WI, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 8092–8094, © 2001, Elsevier Ltd.

Abstract

Motoo Kimura (1924 –94) was a leading Japanese geneticist and the successor to Fisher, Haldane, and Wright in developing the mathematical theory of evolution. His scientific life can be divided into two parts. In the first part he used his mathematical skills to extend the existing theory, especially where stochastic processes are involved. In the second part, he developed the neutral theory – the idea that the great majority of changes at the nucleotide level are driven by mutation and random drift rather than natural selection. The theory, at first controversial, is now largely accepted for molecular changes and has led, among other things, to the molecular clock.

Motoo Kimura was Japan's foremost population and evolutionary geneticist. He is best known for his neutral theory of molecular evolution, which asserts that the great majority of evolutionary changes at the DNA level are driven by mutation and random survival of mutant genes, rather than by natural selection. He was also a major contributor to the mathematical theory of evolution and population genetics.

Life

Motoo Kimura was born on 13 November 1924, in Okazaki, Japan. Early in his life he was fascinated by plants and had a deep interest in their structure. He also displayed an early talent for mathematics. He was admitted to the National High School in Nagoya in 1942. Here, thanks to a knowledgeable and sympathetic teacher, he studied the chromosomes of lilies. This institution also provided his first contact with biometry, and he realized for the first time that he could use his mathematical skills to study biological problems. In 1944 he was admitted to the Kyoto Imperial University.

Normally Kimura would have enrolled in the Faculty of Agriculture to study with the internationally respected geneticist, Hitoshi Kihara. But agriculture students were being called into military service, while botany students were not. So, on Kihara's advice, Kimura enrolled in the Faculty of Science. By the time of his graduation the war was over, and he joined Kihara who encouraged his study of mathematical problems. During this time, Kimura had no specific duties and took the opportunity on his own to read systematically in mathematical genetics. His mathematics, formidable as it sometimes seems, was mainly self-taught. Life in wartime Japan was miserable, and remained so for some time afterward. Kimura was able to supplement a sparse physical and intellectual diet by visiting his cousin, a quantum physicist who lived in the suburbs and could supply both food and scientific talk. Kimura hoped for a theory of population genetics comparable to that of physics.

In 1949, he joined the staff of the National Institute of Genetics in Mishima, and retained this position for the rest of his life. Thanks to help from some American geneticists studying the effects of the atomic bomb in Hiroshima, he was able to obtain support for graduate study in the USA. After a year at Iowa State College Kimura transferred to the University of Wisconsin in 1954 as the author's student. This also provided an opportunity for daily association with America's most eminent population geneticist, Sewall Wright, who had recently moved to the University of Wisconsin.

Except for his graduate years and several short periods abroad, Kimura spent the remainder of his life in the National Institute of Genetics in Mishima, Japan. Late in his life, he developed amyotrophic lateral sclerosis and died on his 70th birthday, 13 November 1994.

Major Contributions

Kimura's scientific life can be divided into two periods. The first began while he was still a student and continued until 1968. During this time he made major contributions to the mathematical theory of population genetics. In 1968, he introduced his neutral theory of molecular evolution. Although he never stopped making theoretical contributions, his major efforts from this time on were devoted to further developments and defense of the neutral theory.

The mathematical theory of evolution had its heyday in the period roughly from 1920 to 1950. It was almost completely dominated by three men, R.A. Fisher and J.B.S. Haldane in England, UK and Sewall Wright in the US. Kimura became the logical successor. Much of Kimura's work utilized the stochastic equations of the Russian mathematician, A. Kolmogrov. Kimura showed a remarkable inventiveness in solving these difficult equations and applying them creatively to significant evolutionary problems.

Among the problems that Kimura solved are the probability that a new mutant gene will ultimately spread through the population; the number of generations required for this process; if the mutation is lost, the number of generations it persists before loss; the number of individuals that carry a mutation during the time until it is fixed or lost; and the average age of a mutant gene segregating in a population.

At the same time he contributed to many other areas of population genetics theory. These included a comparison of inbreeding systems, showing that systems minimizing the immediate increase in homozygosity are not the best in the long run; a further development of Fisher's 'fundamental theorem of natural selection' with more explicit treatment of gene interactions; a new and influential model of population structure, the 'stepping stone' model, in which migrants are

restricted to neighboring colonies; the efficiency of rank-order selection in changing gene frequencies and removing deleterious mutations from the population.

By 1968 the subject of molecular evolution was in a period of rapid development, which still continues. This was the year of Kimura's bombshell. He argued that the great majority of amino acid and nucleic acid changes are selectively neutral. Instead of natural selection as the main directive force, these changes occur by mutation and whether they persist or are lost is a matter of chance. Remarkably, Kimura's early work on stochastic processes turned out to be preadapted for the study of molecular evolution. He had a set of ready-made tools available, and he proceeded to exploit them with great vigor.

The neutral theory made a number of qualitative and quantitative predictions. Among these was that molecular changes occur at a rate independent of structural and physiological changes; Kimura noted that sharks, whose external form has hardly changed in an enormous time period, have as much molecular variation as rapidly evolving animal species, such as mammals. There were a number of similar predictions. One of the simplest and most remarkable is that the rate of neutral gene substitution in the population is equal to the rate of mutation of this gene per individual. This has permitted predictions of evolutionary rates; departures from these predictions provide evidence for natural selection. Rates that are slower than the neutral expectation can be attributed to 'selective constraint'; that is, the amino acids in this region are so specifically fitted to their function that any replacement doesn't work as well. Thus, essentially all mutations are harmful and are removed by natural selection (sometimes called purifying selection). Several proteins, for example some histones, are in this category. Any change in the molecular structure is harmful. In contrast, regions that evolve faster than the neutral rate are attributed to positive selection. An appreciable number of the mutations are beneficial. One of the best understood examples is the DNA region governing tissue transplants.

The neutral theory has provided a rational basis for the idea of a 'molecular clock' guiding the rate of evolution. The rough constancy of the mutation rates for comparable genes in different organisms – in particular that they are much more stable than the forces of selection – has led to the clock assumption becoming the basis for construction of phylogenetic trees, a field of great activity as molecular data are increasingly used.

At first, Kimura's theory was rejected out of hand by most evolutionists. But gradually, over the years, it has come to be widely accepted. It is now realized that a great deal of the DNA of higher organisms has no known function and that the actual genes constitute a small part of the DNA. At present it is accepted that most nucleic acid substitutions follow the neutral paradigm. As for amino acid changes, the jury is still out. Clearly some changes follow neutral kinetics, others clearly are selected, and the proportions are yet to be sorted out.

An Assessment

Let it be said at the outset that Kimura's work has had its greatest impact on evolution theory rather than on the social and behavioral sciences. His work is very general and is applicable to the whole living world, of which that of greatest interest to social scientists, human evolution, and human behavioral evolution, in particular, is only a part. Molecular evolution has yet to make the impact on human behavioral evolution that it is certain to make one day.

Kimura's greatest contribution has been to impose a stochastic viewpoint on the study of evolution. Beneath surface regularities (and irregularities) of form and function are a multitude of molecular changes, governed by the laws of chance. At present, research in behavioral genetics and evolution is based mainly on deterministic theory. Problems such as separating genetic from environmental effects are usually studied by such means as observing adopted children or identical twins reared apart, and the techniques are usually those derived from analysis of variance. But as evolutionary and population genetics theory becomes further developed and richer, and as the molecular basis of behavioral traits (including those in humans) are better understood, stochastic aspects will play an increasingly important role.

The study of natural selection and adaptation has long had difficulty in formulating clearly testable and falsifiable hypotheses. This has been particularly evident in the abundance of 'just so' stories; because we can think of a plausible reason for a structure or function doesn't necessarily mean that it arose for this reason (although it often has). One attribute of the neutral theory is that it can provide a null hypothesis for testing causes of evolutionary changes. For example, as mentioned above, the neutral theory makes a simple evolutionary rate prediction, against which departures can be measured. Thus far, this kind of reasoning has been used mainly for genes with physiological functions, but as the molecular basis of behavioral changes becomes better understood this kind of analysis will be increasingly applicable.

During his lifetime, Motoo Kimura received virtually every honor for which an evolutionary geneticist is eligible. In addition to numerous honors both in Japan and elsewhere, he was a foreign member of the United States National Academy of Sciences, the Royal Society of London, and the Chevalier de L'Ordre National du Mérite in France. He received honorary doctorates from the Universities of Chicago and Wisconsin.

Kimura's lasting impact is twofold. The first, best appreciated by population geneticists, is his enrichment of population genetic and evolutionary theory, especially as these are influenced by stochastic processes. The second, appreciated by a much larger audience, is his discovery of the enormous role that random processes play in molecular evolution. Kimura never denied the importance of Darwinian natural selection as the main evolutionary force shaping form and function, but beneath this, he emphasized, is an enormous play of random changes at the molecular level. In addition to Darwin's 'survival of the fittest,' he added the slogan, 'survival of the luckiest.'

See also: Adaptation, Fitness, and Evolution; Darwin, Charles Robert (1809–82); Evolution, History of; Fisher, Ronald A (1890–1962); Galton, Sir Francis (1822–1911); Human Evolutionary Genetics; Lorenz, Konrad (1903–89).

Bibliography

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK.

Kimura, M., 1994. Population Genetics, Molecular Evolution, and the Neutral Theory. University of Chicago Press, Chicago.

Kin Selection

Andy Gardner, University of St Andrews, St Andrews, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The theory of kin selection is the foundation for the study of social evolution. Here I discuss the historical origins of the theory, provide a mathematical overview, review the various mechanisms that give rise to kin selection, and explain its relevance to the understanding of eusociality.

The Origins of Kin Selection

The basic idea of kin selection was formulated by Charles Darwin, in *The Origin of Species* (Darwin, 1859; *see* Darwinism). It arose as the solution to a puzzle posed by sterile workers in social insect colonies; specifically, how to explain the striking adaptations of these individuals, given that they cannot reproduce and hence appear to be beyond the reach of natural selection. Darwin realized that this puzzle is similar to a problem faced by farmers who rear cattle for their meat. After slaughtering an animal, preparing the meat, and finally discovering that it is particularly delicious, the farmer is unable to breed from the slaughtered individual to ensure that subsequent generations of the herd will inherit the superior meat. However, he/she can breed from the slaughtered individual's close relatives who, on account of their co-ancestry, are likely to share the heritable tendency for delicious meat. Darwin suggested that something similar happens in the social insects. If a variant worker behavior or morphology arises that provides a reproductive benefit to the individual's fertile family members then, because those family members have a strong likelihood of sharing the same heritable tendency, it may be preserved for posterity by the action of natural selection.

Later, R.A. Fisher provided a kin selection explanation for the evolution of distastefulness in gregarious insect larvae in *The Genetical Theory of Natural Selection* (Fisher, 1930). He suggested that, although distastefulness provides no benefit to a predated larva, it may dissuade the predator from eating the larva's siblings. Fisher remarked that the ensuing selective benefit is only half as great as it would be had the individual herself avoided predation, and this appears to be the first use of a coefficient of relatedness to translate fitness effects for relatives into a selective benefit for a focal actor. However, Fisher did not elaborate the principle for more distant degrees of relationship. And he explicitly neglected such 'indirect effects' in the derivation of his fundamental theorem of natural selection, which he used as justification for the view that natural selection leads organisms to appear designed to maximize their personal reproductive success. In 1955, in a popular science article, J.B.S. Haldane gave a verbal account of how selection would operate on a rare gene that made its carrier jump into a river to save a drowning relative, showing that the gene could only be selectively favored if the degree of relationship were sufficiently close (Haldane, 1955). However, he seems not to have appreciated that this principle could be used to understand social evolution more generally.

Thus, it fell to W.D. Hamilton to develop and usefully apply a proper mathematical theory of kin selection, more than a century after Darwin had first formulated the idea. As an undergraduate at the University of Cambridge in the 1950s, Hamilton had been dismayed by the rampant 'for the good of the species' thinking that had come to dominate evolutionary biology in general and the study of social behavior in particular. So he took up the mantle of kin selection where Fisher had left it, focusing his doctoral studies upon the genetical evolution of altruistic behavior. In a series of articles published over 10 years during the 1960s and 1970s (Hamilton, 1963, 1964, 1967, 1970, 1972), Hamilton laid the theoretical foundations of kin selection, devised a classification of social behaviors, derived a general statement of how direct and indirect fitness effects jointly determine the response to natural selection, showed that this leads organisms to appear designed to maximize their inclusive fitness rather than their personal fitness, and applied this understanding to make seminal contributions to the topics of altruism, sex allocation, and eusociality. Spurred by Hamilton's work, John Maynard Smith coined the term 'kin selection' (Maynard Smith, 1964), and Eric Charnov immortalized Hamilton's contribution by naming the core mathematical result of kin selection theory 'Hamilton's rule' (Charnov, 1977). The theory of kin selection now provides the foundation for the study of social evolution (Frank, 1998).

The Genetical Theory of Kin Selection

The canonical elucidation of kin selection theory takes Price's (1970) equation of natural selection as its starting point (e.g., Hamilton, 1970). The following exposition owes heavily to Queller (1992) and is more fully developed by Gardner et al. (2011). Price's equation encapsulates the action of natural selection as the part of genetic change that is ascribable to the differential fitness of individuals:

$$\Delta_{\mathrm{NS}}E(g) = \mathrm{Cov}(v, g). \qquad [1]$$

This states that the change in the population average (or expectation, E) of any genetical character that owes to the action of natural selection is equal to the covariance (Cov) of relative fitness (v) and genetical character (g) across all the individuals in the population. Here, the genetical character

may represent the frequency of a particular gene in the individual's genome (in which case $E(g)$ is the population frequency of that gene), the heritable component of any phenotypic trait of interest (i.e., a genetic 'breeding' value), or any other weighted sum of genic values. That is, if individuals with a greater genetical character value tend to have higher fitness, then the covariance is positive ($\mathrm{Cov}(v, g) > 0$) and natural selection acts to favor an increase in average character value ($\Delta_{\mathrm{NS}}E(g) > 0$). Whereas, if individuals with a greater genetical character value tend to have lower fitness, then the covariance is negative ($\mathrm{Cov}(v, g) < 0$) and natural selection acts to favor a decrease in the average character value ($\Delta_{\mathrm{NS}}E(g) < 0$). If there is no tendency for individuals with a greater genetical character value to have higher or lower fitness, then the covariance is zero ($\mathrm{Cov}(v, g) = 0$), and natural selection favors neither an increase nor a decrease in the average character value ($\Delta_{\mathrm{NS}}E(g) = 0$).

Assuming that there is some heritable variance in the population (i.e., $\mathrm{Var}(g) > 0$) then, without loss of generality, one can write $\mathrm{Cov}(v, g) = \beta(v, g)\mathrm{Var}(g)$, where $\beta(v, g) \equiv \mathrm{Cov}(v, g)/\mathrm{Var}(g)$. This coefficient $\beta(v, g)$ turns out to have a useful interpretation: if one were to make a plot of relative fitness (v) against genetic character (g) for all the individuals in the population, then $\beta(v, g)$ is equal to the slope of the straight line $E(v|g) = E(v) + \beta(v, g)\,(g - E(g))$ fitted through these (v, g) data by the method of least squares (Figure 1(a)). Since variances are non-negative, the condition for natural selection to favor an increase in average character value ($\mathrm{Cov}(v, g) > 0$) may be written as:

$$\beta(v, g) > 0 \qquad [2]$$

Note that I have not assumed that there is a straight-line relationship between genes and fitness in the population. Indeed, there may be an arbitrarily complex relationship between genes and fitness, and the least-squares fit may fail to capture lots of this complexity (Figure 1(a)). However, it is only the least-squares fit that natural selection cares about. That is, natural selection favors an increase in the average character value if and only if the slope of the least-squares linear regression is positive, and this translates into an actual response to selection in proportion to the heritable variance that is present in the population.

This association between genetic value and fitness is statistical and need not reflect a direct causal relationship. For example, if we consider that the genetic character leads the individual to enact altruism toward her neighbor, then the direct impact of the character on her own fitness will be negative, but she may enjoy an indirect benefit owing to her neighbor also carrying the gene and directing altruistic behavior back toward her. The overall slope of relative fitness against genetic character can be partitioned into these separate direct and indirect effects using multiregression methods. That is, instead of fitting a straight line through two dimensional data, as was described earlier, the idea is to fit a plane $E(v|g, g') = E(v) + \beta(v, g|g')\,(g - E(g)) + \beta(v, g'|g)\,(g' - E(g'))$ through the three-dimensional data set (v, g, g'), where g' denotes the genetic character of the focal individual's neighbor (Figure 1(b)). This leads to the multiregression partition:

$$\beta(v, g) = \beta(v, g|g') + \beta(v, g'|g)\,\beta(g', g). \qquad [3]$$

Considering an altruistic trait, the impact of the individual's character value (g) upon her own fitness (v), holding fixed the character value of her neighbor (g'), defines the cost of altruism for the actor, i.e., $\beta(v, g|g') \equiv -c$. The impact of the neighbor's character value (g') upon the focal individual's own fitness (v), holding fixed her own character value (g), defines the benefit of altruism for the recipient, i.e., $\beta(v, g'|g) \equiv b$. Finally, the association between the genetic character values of the focal individual (g) and her neighbor (g') defines the kin selection coefficient of relatedness, i.e., $\beta(g', g) \equiv r$. Hence, from eqn [3], the condition for natural selection to favor an increase in altruism – or, indeed, any genetic character of interest – is:

$$-c + br > 0, \qquad [4]$$

which is Hamilton's rule of kin selection. This formally separates the direct fitness effect of a character ($-c$) from the indirect

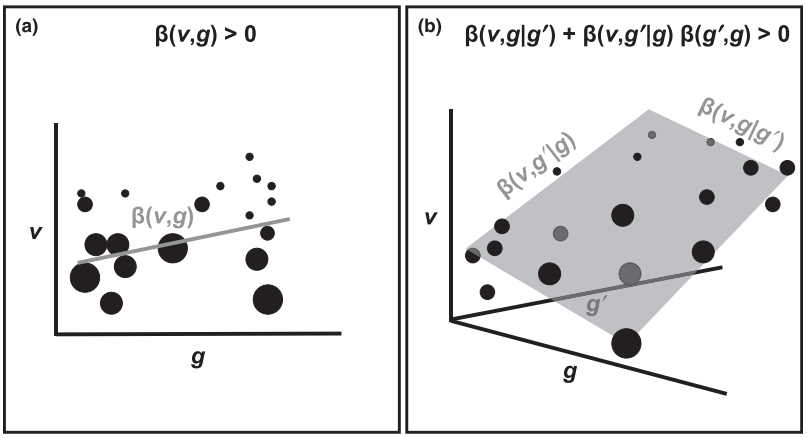

Figure 1 The genetical theory of kin selection. (a) Natural selection favors any heritable trait that is positively associated with fitness ($\beta(v, g) > 0$). (b) This association can be broken down into the direct impact of the trait on its bearer's fitness ($\beta(v, g|g')$) and the indirect impact of the trait on its bearer's fitness ($\beta(v, g'|g) \times \beta(g', g)$).

fitness effect of that character (rb). The part of natural selection corresponding to direct fitness effects may be termed 'direct selection,' and the part that corresponds to indirect fitness effects may be termed 'indirect selection' or 'kin selection.' That is, kin selection is a part of natural selection, driven by differences in individual fitness, rather than an additional evolutionary process.

In the above derivation, I have assumed that the focal individual has only a single social partner. However, the same methods are readily extended to consider multiple social partners, acting in a variety of roles. This yields a Hamilton's rule of the form $-c + \Sigma_{i \in I}\, b_i\, r_i$, where b_i and r_i are the benefit and relatedness, respectively, corresponding to the individual's ith social partner (Grafen, 2006). I have also assumed that the focal individual and her social partner are interchangeable. That is, they are exactly equivalent in all ways except for their genetic character values. More generally, social partners may belong to different classes, such as male and female or young and old or queen and worker. However, kin selection methods are also readily applied to such class structured populations (Taylor, 1990). Note that I have made no assumption that genes for altruism are additive in their action. There may be any degree of dominance within loci, epistasis between loci, synergy between individuals and frequency dependence at the population level and Hamilton's rule still holds. Recurring claims that Hamilton's rule breaks down in these scenarios are based upon a confusion between the additivity of a linear regression model versus the additivity of an evolutionary genetics model (reviewed by Gardner et al., 2011).

Here, I have focused on altruism, but Hamilton's rule applies when any heritable trait is favored by natural selection. Social behaviors are classified according to their impact on the fitness of the actor and the recipient (Table 1): altruistic behaviors decrease the fitness of the actor ($c > 0$) and increase the fitness of the recipient ($b > 0$); selfish behaviors increase the fitness of the actor ($c < 0$) and decrease the fitness of the recipient ($b < 0$); mutually beneficial behaviors increase the fitness of both the actor ($c < 0$) and the recipient ($b > 0$); and spiteful behaviors decrease the fitness of both the actor ($c > 0$) and the recipient ($b < 0$). In each case, the condition for natural selection to favor the social behavior is $-c + br > 0$. Note that spiteful behavior ($b < 0$ and $c > 0$) is only favored if there is negative relatedness between actor and recipient ($r < 0$, such that $-c + br > 0$). Negative relatedness sounds unusual, but simply means that the actor and recipient are less related than average (Hamilton, 1970). Nonsocial behaviors are those that do not have fitness consequences for individuals other than the actor ($b = 0$), and hence are favored if they improve the fitness of the actor ($c < 0$) and are disfavored if they reduce the fitness of the actor ($c > 0$).

The above derivation of Hamilton's rule has taken the 'personal fitness' or 'neighbor-modulated fitness' approach to kin selection (Figure 2). This considers the impact of an individual's own genetic character (g) and that of her social partner (g') upon her own fitness (v). An alternative – but exactly equivalent – approach is the 'inclusive fitness' approach, which considers the impact of an individual's own genetic character (g) upon her own fitness (v) and the fitness of her social partner (v'; Figure 2). The only difference between these two approaches is that the benefit term, conceptualized as $b = \beta(v, g'|g)$ under the personal fitness view, is reconceptualized as $b = \beta(v', g|g')$ under the inclusive fitness view. Thus, the response to selection may be written as $\beta(v, g|g') + \beta(v', g|g')r = \beta(v + v'r, g|g')$. This describes the partial effect of the individual's own genetic character (g) upon the relatedness-weighted sum of the individual's relatives' fitnesses ($v + v'r$), i.e., the impact that the individual has upon her own inclusive fitness.

Table 1 Classification of social behaviors

		Impact upon recipient	
Impact upon actor		+	–
	+	Mutual benefit	Selfishness
	–	Altruism	Spite

Behaviors that increase the fitness of both actor and recipient (+/+) are termed *mutually beneficial*; behaviors that increase the fitness of the actor and decrease the fitness of the recipient (+/–) are termed *selfish*; behaviors that decrease the fitness of the actor and increase the fitness of the recipient (–/+) are termed *altruistic*; and behaviors that decrease the fitness of both actor and recipient (–/–) are termed *spiteful*. See West et al. (2007) for more discussion.

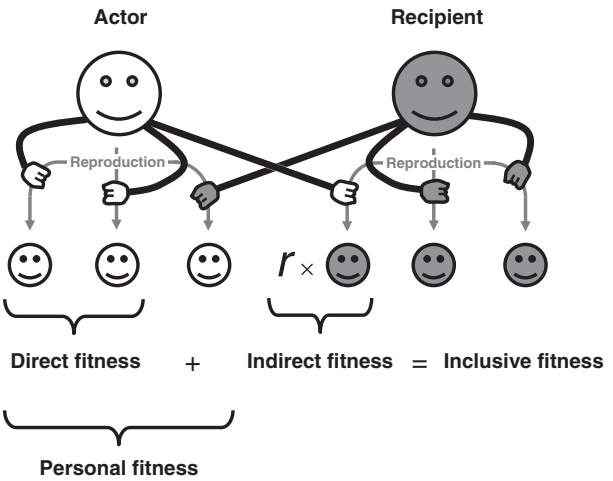

Figure 2 Personal and inclusive fitness. Personal (or neighbor-modulated) fitness describes a focal individual's lifetime reproductive success (unshaded offspring), some of which may be attributed to the action of the individual's own genes (unshaded hands), and some of which may be attributed to the action of the genes carried by the individual's social partners (shaded hands). A necessary and sufficient condition for genes to be favored by natural selection is that they are positively associated with personal fitness. However, individuals do not behave as if they are trying to maximize their personal fitness, because they do not have complete control over this quantity. Inclusive fitness describes the impact that a focal individual's genes (unshaded hands) have on her own fitness (unshaded offspring) and the fitness of her genetic relatives (shaded offspring), each increment or decrement being weighted by the coefficient of relatedness between the individual and her recipient. A necessary and sufficient condition for genes to be favored by natural selection is that they are positively associated with inclusive fitness. Moreover, individuals do behave as if they are trying to maximize their inclusive fitness, because they have complete control over this quantity.

Inclusive Fitness

The dynamics of natural selection – operating via direct and indirect (kin selected) effects – can be understood either in terms of personal fitness or in terms of inclusive fitness. These two approaches are exactly equivalent and yield the same predictions. However, the adaptive design that emerges from the action of natural selection can only be understood in terms of inclusive fitness. The organism is not generally adapted to maximize her personal fitness, because – in the context of social interactions – she does not have full control over this quantity. However, she is adapted to maximize her inclusive fitness, which is under her sole control, the effects of her social partners having been explicitly stripped away. Consequently, inclusive fitness is a crucial concept in the study of Darwinian adaptation (*see* Adaptation, Fitness, and Evolution). The formal connection between the dynamics of natural selection and the optimization of inclusive fitness has been shown by Grafen (2006).

This point is best explained with an illustration. Consider a gene that makes an insect worker lay fewer eggs and work harder to improve the reproductive success of her queen. From a personal fitness approach we would say that sometimes the gene will be present in a worker, in which case it will be associated with reduced fitness, because it makes her lay fewer eggs; but sometimes the gene will be present in a queen, in which case it will be associated with increased fitness, because the workers in the queen's colony will also tend to carry the gene, which makes them work harder. Consequently, the gene for worker altruism will be favored if the former direct effect is outweighed by the latter indirect effect. However, we cannot regard the insect worker as striving to maximize her personal fitness, as this would never allow for worker altruism to evolve. Instead, we can take an inclusive fitness approach and say that altruism enacted by the worker reduces her direct fitness, by making her lay fewer eggs; but it also improves her indirect fitness as a consequence of increasing the reproductive success of the queen; and, consequently, worker altruism is favored if the former direct effect is outweighed by the latter indirect effect. Crucially, the worker can be regarded as striving to maximize her inclusive fitness, as it is under her complete control.

Often, the personal fitness approach is the easier option when the aim is simply to capture the dynamics of natural selection in a mathematical model. However, by emphasizing an actor and the causal impact that her social behavior has upon the reproductive success of her social partners, the inclusive fitness approach is more useful for empirical research in behavioral ecology. Moreover, because humans are naturally better at reasoning about agents and strategies than about the dynamics of gene frequencies, the inclusive fitness approach – which yields a simple maximizing agent interpretation – provides a powerful way of thinking about social evolution. The notion of an agent with an agenda is implicit within the intentional language employed by evolutionary biologists: selfishness, altruism, conflicts of interest, etc. For these reasons, the inclusive fitness approach to kin selection has dominated the empirical study of social evolution.

Mechanisms of Kin Selection

Kin selection occurs whenever there is a genetic association between the actor and the recipients of her social behavior. Hamilton (1964) suggested three basic mechanisms for ensuring such a genetic association: population viscosity; kin recognition; and the greenbeard effect (Figure 3).

First, population viscosity occurs when individuals do not disperse far from their place of origin, so that neighboring individuals tend to be closely related, and hence even indiscriminate social behavior can lead to kin selection (Figure 3(a)). Because this requires no special cognition or recognition mechanism, population viscosity has been suggested to provide a very general explanation for cooperative behavior. However, a problem with this idea is that viscosity may also be associated with intense local competition for resources, which can inhibit the evolution of cooperative behavior. In fact, in the simplest models of viscous populations the relatedness-enhancing and competition-enhancing effects of reduced dispersal exactly cancel so that there is no overall impact of the rate of dispersal on the evolution of cooperation (Taylor, 1992). This surprising result has generated a great deal of theoretical interest in exploring factors that act to decouple the relatedness and competition consequences of dispersal, so that cooperation can evolve in viscous populations.

One such factor is budding dispersal. If individuals disperse in groups, alongside social partners, then they may maintain high relatedness with their neighbors while also enjoying the competition-relaxation benefits of dispersal. This idea has been tested empirically, using experimental evolution of

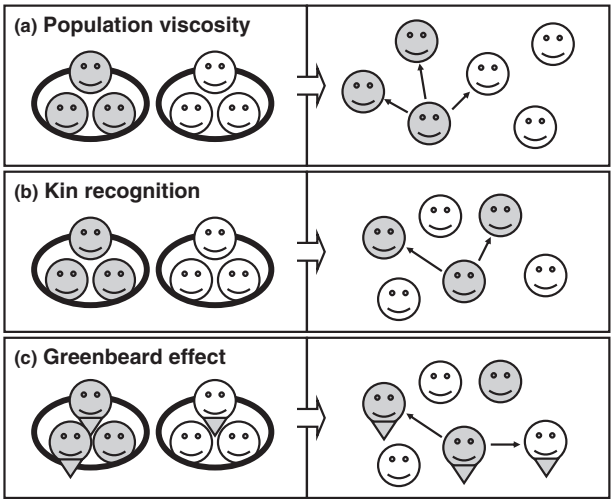

Figure 3 Mechanisms of kin selection. (a) Population viscosity: if individuals do not move far during their lifetime, then they will tend to be surrounded by kin (shaded), and hence even indiscriminate social behavior will give rise to kin selection. (b) Kin recognition: if individuals use cues that carry information about kinship to mediate their social behavior, such as remembering with whom they shared a nest (shaded) and acting altruistically toward those individuals, then this will give rise to kin selection. (c) Greenbeard effect: if the gene controlling social behavior is associated with a phenotypic marker, such as a beard, then carriers of the gene can identify which of their social partners also carry copies, independently of whether or not these social partners are kin (shaded).

Pseudomonas aeruginosa bacteria. In iron-limited medium, bacterial cells may release iron-scavenging molecules called siderophores that bind iron and facilitate its uptake by cells. However, the cell that takes up the bound iron need not be the cell that manufactured and released the siderophore, so siderophore production is a form of public-goods cooperation and cheating cells that do not produce siderophores may free-ride on the cooperation of the other cells in the medium. Kummerli et al. (2009) varied the pattern and rate of dispersal of cells between flasks of medium and showed that, while cooperation was disfavored under both high and low individual dispersal, it was favored under budding dispersal.

Second, kin selection may also occur when individuals use environmental or genetic cues to recognize which of their social partners are their genealogical kin and adjust their social behavior according to this information (Figure 3(b)). Environmental kin recognition occurs when an individual makes use of an environmental cue that is associated with her having greater kinship to her social partners. For example, the rule 'behave altruistically toward individuals who are in my nest, and behave aggressively toward individuals who are not in my nest' uses the position of the recipient relative to the individual's own nest as a proxy for degree of kinship. In contrast, genetic kin recognition occurs when an individual makes use of a genetic cue. For example, if blue eyes were relatively rare in the population, then it is likely that two neighbors who shared blue eyes would do so because they were genealogically close kin, rather than simply by chance. Consequently, individuals may be favored to behave altruistically toward those neighbors whom they most closely resemble.

However, genetic kin recognition appears to be inherently unstable. This is because, while only relatively rare genetic markers provide a good indicator of genealogical closeness, common genetic markers may enjoy a selective advantage owing to their carriers being more often identified as kin and receiving benefits of altruism. Hence, the polymorphism at marker loci that is required for genetic kin recognition to work may be eroded by natural selection, until every individual carries the same alleles, which no longer give any information about genealogical kinship (Crozier, 1986). This may explain why genetic kin recognition is relatively rare in the natural world. One possibility for maintaining genetic kin recognition is if the genetic marker loci mediate host–parasite interaction, as this is a potent source of balancing selection that can help to maintain rare alleles. This may explain why genetic kin recognition, when it does occur, appears to be based on immune-system genes such as the multihistocompatibility complex in humans (Wedekind et al., 1995) and extraembryonic-membrane genes in parasitoid wasps (Giron and Strand, 2004).

Third, the greenbeard effect is similar to genetic kin recognition, but occurs when the same gene (or two closely linked genes) encodes both the marker and the social behavior (Figure 3(c)). The idea here is that two individuals who share a genetic marker in common may also share genes for social behavior that are linked to that marker, irrespective of whether they are genealogically close kin. The greenbeard mechanism was devised by Hamilton (1964) as a thought experiment to illustrate that genetic similarity, and not genealogical kinship per se, is the key to kin selection. The name was coined by Dawkins (1976), who gave a hypothetical example of a gene that causes its carrier to grow a green beard and to act altruistically toward other individuals with green beards. However, empirical examples have been found in the natural world. For example, bacteriocins are toxins produced by bacteria, often at great cost to the producer cell, that poison neighboring cells. The bacteriocin gene is tightly linked to a gene for immunity to that bacteriocin (they are often carried together on a plasmid), and so other carriers of the bacteriocin gene survive the killing and benefit from reduced competition for resources (Gardner and West, 2010). Greenbeards are susceptible to invasion by 'falsebeards,' that enjoy the benefits of bearing a beard but without paying the costs of the social behavior (Gardner and West, 2010). For example, a bacterium that enjoys immunity to bacteriocins without committing itself to the costs of bacteriocin production. Such falsebeards are readily created by sex and recombination between beard and behavior loci, and this may explain why most empirical examples of greenbeards are found in the asexual world of microbes.

Kin Selection and Eusociality

Obligate eusociality is a permanent division of reproductive labor, whereby both reproductive and helper castes are required for totipotency. This system of biological organization is known only in the social hymenoptera (ants, wasps, and bees) and the social cockroaches (termites). The problem of eusociality is to explain why individuals of the helper caste have evolved to give up any chance of future reproductive success in order to serve the needs of their reproductive colony mates. This is the most striking example of altruistic behavior in the natural world and has consequently received much attention from kin selection researchers.

Historically, most attention has been given to the 'haplodiploidy hypothesis' for eusociality. Most of the origins of obligate eusociality have occurred in the social Hymenoptera, a group that is noted for its haplodiploid mode of inheritance. These organisms produce diploid daughters sexually, with one genome being contributed from the mother and one from the father. But they produce haploid sons asexually, from unfertilized eggs. Hamilton (1964, 1972) famously suggested that the apparent predisposition for haplodiploid organisms to eusociality owes to the inflated relatedness between full sisters that emerges under this mode of inheritance. Under both diploidy and haplodiploidy, the relatedness of a female to her daughter is one half ($r = 1/2$), but whereas the relatedness between sisters under diploidy is also one half ($r = 1/2$) the relatedness between sisters under haplodiploidy is three quarters ($r = 3/4$). Hence, while females in diploid populations are ambivalent over whether to rear a sister or a daughter, females in haplodiploid populations prefer to rear sisters over daughters, and this potentially paves the way for the evolution of eusociality.

However, this argument does not really work. Hamilton (1964, 1972) was aware that the relatedness between sister and brother is correspondingly lower under haplodiploidy, and Trivers and Hare (1976) pointed out that this actually cancels any benefit of rearing siblings (sisters and brothers)

rather than own offspring (daughters and sons), assuming an even sex ratio. Craig (1979) subsequently showed that even a female-biased sex ratio does not resolve this problem, as the increased relatedness to siblings (mostly sisters) is exactly countered by the increased reproductive value of males (which is higher, owing to the rarer-sex effect, in a population with a female-biased sex ratio). Haplodiploidy can be made to promote altruistic sib-rearing if there are 'split sex ratios,' whereby some colonies are relatively female biased and others are relatively male biased (Grafen, 1986), but it can also be made to inhibit altruistic sib-rearing, and in most cases its impact is meager (Gardner et al., 2012).

Instead, the key to eusociality appears to be strict lifetime monogamy, which ensures that maternal siblings are full siblings, and hence – on average – as related to the actor as are her own offspring (Boomsma, 2009). That is, it is not necessary for relatedness of siblings to exceed that of offspring, but only that relatedness of siblings match that of offspring, such that any slight efficiency benefit of sib-rearing is sufficient to provide a selective advantage for eusociality. Although not all eusocial organisms are strictly monogamous – the highly polyandrous honeybee providing a striking counterexample – phylogenetic analysis reveals that strict lifetime monogamy is the ancestral state for all eusocial hymenoptera and was the mating system in which eusociality always initially arose (Hughes et al., 2008). The predisposition for the hymenoptera to evolve eusociality may simply be due to their ancestral sperm-storage ability, which enables a singly mated female to produce multiple batches of offspring without needing to remate. Moreover, the monogamy hypothesis also explains why termites have evolved eusociality (their societies are founded by a monogamous mating pair), despite their diploid mode of inheritance.

Summary

Darwin formulated the concept of kin selection in *The Origin of Species*. More than a century later, Hamilton developed a mathematical theory of kin selection, and showed that natural selection leads organisms to appear designed to maximize their inclusive fitness. Hamilton's rule of kin selection emerges as a simple partition of natural selection into its direct and indirect components and is as general as the genetical theory of kin selection itself. There are three major mechanisms by which kin selection operates: population viscosity, kin recognition, and the greenbeard effect. Kin selection is the key to understanding altruistic behavior and the evolution of eusociality.

See also: Adaptation, Fitness, and Evolution; Human Cooperation, Evolution of; Kinship, Evolution of; Sociobiology: Overview.

Bibliography

Boomsma, J.J., 2009. Philosophical Transactions of the Royal Society B 364, 3191–3208.
Charnov, E.L., 1977. Journal of Theoretical Biology 66, 541–550.
Craig, R., 1979. Evolution 33, 319–334.
Crozier, R.H., 1986. Evolution 40, 1100–1101.
Darwin, C.R., 1859. The Origin of Species. John Murray, London.
Dawkins, R., 1976. The Selfish Gene. Oxford University Press, Oxford.
Fisher, R.A., 1930. The Genetical Theory of Natural Selection. Clarendon Press, Oxford.
Frank, S.A., 1998. Foundations of Social Evolution. Princeton University Press, Princeton.
Gardner, A., West, S.A., 2010. Evolution 64, 25–38.
Gardner, A., West, S.A., Wild, G., 2011. Journal of Evolutionary Biology 24, 1020–1043.
Gardner, A., Alpedrinha, J., West, S.A., 2012. American Naturalist 179, 240–256.
Giron, D., Strand, M.R., 2004. Proceedings of the Royal Society, London B Biology Letters 271, S395–S398. (Suppl.).
Grafen, A., 1986. Journal of Theoretical Biology 122, 95–121.
Grafen, A., 2006. Journal of Theoretical Biology 238, 541–563.
Haldane, J.B.S., 1955. New Biology 18, 34–51.
Hamilton, W.D., 1963. American Naturalist 97, 354–356.
Hamilton, W.D., 1964. Journal of Theoretical Biology 7, 1–52.
Hamilton, W.D., 1967. Science 156, 477–488.
Hamilton, W.D., 1970. Nature 228, 1218–1220.
Hamilton, W.D., 1972. Annual Review of Ecology, Evolution, and Systematics 3, 193–232.
Hughes, W.O.H., Oldroyd, B.P., Beekman, M., Ratnieks, F.L.W., 2008. Science 320, 1213–1216.
Kummerli, R., Gardner, A., West, S.A., Griffin, A.S., 2009. Evolution 63, 939–949.
Maynard Smith, J., 1964. Nature 201, 1145–1147.
Price, G.R., 1970. Nature 227, 520–521.
Queller, D.C., 1992. Evolution 46, 376–380.
Taylor, P.D., 1990. American Naturalist 135, 95–106.
Taylor, P.D., 1992. Evolutionary Ecology 6, 352–356.
Trivers, R.L., Hare, H., 1976. Science 191, 249–263.
Wedekind, C., Seebeck, T., Bettens, F., Paepke, A.J., 1995. Proceedings of the Royal Society, London B 260, 245–249.
West, S.A., Griffin, A.S., Gardner, A., 2007. Journal of Evolutionary Biology 20, 415–432.

Kindling

Gordon C Teskey, Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada
Michael E Corcoran, University of Saskatchewan, Saskatoon, SK, Canada

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by S. Schreiber, volume 12, pp. 8094–8098, © 2001, Elsevier Ltd.

Abstract

Intermittent application of brief trains of electrical stimulation to focal areas of the brain triggers epileptiform discharge (afterdischarges) and the gradual development of behavioral seizures, a phenomenon known as kindling. Because the intensity of the electrical stimulation is not altered, kindling must be due to a change in the responsiveness of the nervous system itself. Thus kindling results in an extreme form of sensitization, to the extent that the kindled brain is inordinately responsive to a variety of potential convulsants. Another important characteristic of kindling is that the increase in the nervous system's susceptibility to seizures is extraordinarily long-lasting and may be considered permanent. Moreover, kindling allows a high degree of experimental control, as the researcher can target any nuclei in the brain and apply stimulation at will. Kindling has proven to be one of the most robust, reliable, and dramatic forms of neural plasticity yet described, and over the past half century, it has attracted an enormous amount of experimental attention.

While Graham Goddard's seminal paper in 1967 is rightly credited as the first systematic description of the kindling phenomenon there were several earlier observations of kindling-like phenomenon. Descriptions by Watanabe in 1936, by Whieldon and Van Harreveld in 1950, from Jose Delgado's laboratory at Yale in the late 1950s and early 1960s all preceded Goddard. But it was Graham Goddard who first fully appreciated, named, and then with the help, in the early days, of several others, including Ron Racine, Dan McIntyre, Juhn Wada, John Pinel, Mac Burnham, Michael Corcoran, Peter Cain, and Bob Adamec, effectively promoted and explored kindling.

The Discovery of Kindling

There are several noteworthy discoveries in the history of neuroscience that arose from serendipitous observations and the discovery of kindling was one such observation that began from research on an entirely different problem. Graham Goddard's doctoral research in the early 1960s was concerned with the effects of electrical stimulation of the amygdala on conditioning in rats. Due to an equipment malfunction, Goddard accidentally delivered repeated trains of fairly high-intensity stimulation to the amygdala of some rats, one of which developed seizures after a few repetitions. The rat continued to display seizures even after the equipment had been repaired and low-intensity stimulation was delivered. Goddard realized that this might be significant when he found that the low-intensity stimulation given to experimentally naive rats did not cause seizures. In other words, it was the *repeated application* of stimulation that led to the elicitation of seizures, not the stimulation itself. In an independent and elegant series of experiments, Ronald Racine (1972a,b) carefully monitored electroencephalogram activity during kindling and convincingly demonstrated that stimulation has to elicit an afterdischarge (AD) in order to kindle seizures. Goddard further surmised that the brain itself might have been changing in response to the repeated application of stimulation, and that this phenomenon might in turn have something to do with the mechanisms of learning. It is worth noting the Goddard's view of the emerging phenomenon was that it could give insight into the normal functioning of the brain. This is important because the initial questions that he formulated where predicated on the notion of discovering a relation between kindling and normal learning and memory, rather than kindling as a model of seizures and epilepsy.

The Properties of Kindling

In order for the phenomenon to potentially be related to learning and memory, it should show the following properties: Different areas should have different susceptibilities; the effect should be persistent; it should be found in all mammals; and intermittent stimulation should be more effective than massed stimulation. Furthermore, the phenomenon should not depend on the electrical stimulation itself, and it should not be due to damage cause by the electrode or the passing of current.

Differential Susceptibility of Brain Areas

Kindling rate is operationally defined as the number of stimulations delivered to brain that results in some clearly defined end point such as the first fully generalized (stage 5) seizure. Within rat brain there is a relationship between the structure being kindled and its kindling rate. The following forebrain areas are ordered from slowest kindling rate to fastest kindling rate: dorsal hippocampus, ventral hippocampus, amygdala, piriform cortex, and perirhinal cortex. Many more brain structures than those listed here have been kindled and all have shown the phenomenon except one; the cerebellum. Although many of us have tried, it seems that the cerebellum is incapable of generating an epileptiform discharge. Note that kindling rate, as defined above, is based on a behavioral seizure and that means the elicited AD must spread to motor circuits and

ultimately to the spinal cord and alpha motor neurons. Thus, kindling rate is a function of the propagation of ADs from the site of stimulation to other structures. Kindling rate is a different phenomenon from seizure susceptibility or AD threshold, which is defined as the minimal amount of current that can elicit an AD. The hippocampus appears to have the lowest AD threshold but takes a large number of stimulations to kindle. Thus different brain areas have both different kindling rates and seizure thresholds that appear to be independent from each other.

Persistence

Several publications have documented that kindling, like learning, is highly persistent (Wada et al., 1974). Rats that have been kindled in the hippocampus or amygdala rested for long periods of time (e.g., 24 weeks) and then rekindled showed remarkable savings. Rats typically displayed much shorter AD durations and less-severe behavioral seizures on the first rekindling stimulation compared to the response to the last stimulation before the rest period. However, they rapidly rekindle and only a few stimulations are required before they displayed the behavioral end point they achieved before the rest interval (Dennison et al., 1995). The effects of kindling are highly persistent compared to the life span of the animal.

Generalizability

Kindling has been demonstrated in all species studied to date. Seizures can be kindled in lizards, frogs, mice, gerbils, rats, rabbits, guinea pigs, cats, dogs, squirrel monkeys, rhesus monkeys, and several species of baboon. Thus kindling is a very robust and seemingly universal phenomenon in all vertebrates tested. But has kindling occurred in humans?

There are two published accounts of kindling in people. Sramka et al. (1977) described the case of a 47-year-old woman who carried depth electrodes located in three structures of the left thalamus and was stimulated daily to relieve pain. During the third week of stimulation, she developed spontaneous mild movements in the right half of her face and parts of her right hand. Although the electrodes were removed, she continued to have seizures and paroxysmal brain activity, and the seizures became secondarily generalized. Monroe in 1982 also clearly described kindling in a person. On the face of it, these data suggest that the human brain is indeed susceptible to kindling. So why are there not more reports of kindling in people? It is, of course, possible that physicians have observed kindling after electrical stimulation of the human brain but have refrained from reporting it, for fear of prosecution. In confirmation of this possibility, Dan McIntyre has described private conversations with neurologists who stated that they have unintentionally kindled seizures with stimulation via depth electrodes in some patients.

What might occur more obviously in some populations of people with epilepsy is an intensification of ictal phenomena that could be due to kindling-like events or to the neurobiological mechanisms underlying kindling. Neurologists have argued that the progressive increases in ictal phenomena observed in patients with tumor-related seizures can be attributed to a kindling-like effect. Similarly, withdrawal seizures can intensify with repeated episodes of detoxification from ethanol in humans, an observation interpreted as a kindling effect. In balance, we think it is a certainty that a kindling-like process occurs in the human brain.

Stimulation Interval

Goddard et al. (1969) were also concerned with identifying the optimal stimulation schedule for kindling seizures, and they found that kindling was unlikely with continuous stimulation or stimulation presented at interstimulation intervals below 20 min, whereas the high probability of kindling did not vary at interstimulation intervals of 8 h to 7 days. That is, continuous or too frequent stimulation in adult rats either fails to kindle seizures or results in a long-lasting decrease in seizure susceptibility that interferes with subsequent kindling. The cardinal consideration is that stimulation not be applied at too short an interval; otherwise kindling will fail (Racine et al., 1973).

Brain Damage

Normal learning and memory obviously does not result from the irritative effects of deposited metals from electrodes or outright tissue damage, so it was important for the early kindlers to show that electrode deposition and/or damage did not result from repeated stimulations. There was prior work demonstrating that seizures could be triggered by electrolysis of metallic ions into brain tissue, and it is likely that some deposition of ions occurs even with the balanced waveforms used for kindling. Might not kindling then be due simply to local irritative effects produced by passage of current through metal electrodes? Perhaps the most convincing evidence against a role for irritative effects is found in the observation that kindling of one brain site results in *transfer* of seizure susceptibility to a second brain site. This is measured by a significant facilitation in the rate of secondary-site kindling as compared to naive controls. If kindling is due solely to induction of an irritative focus restricted to the primary site of stimulation, it is unclear how this could drive epileptiform events outside the focus and result in the development of behavioral seizures, which presumably reflect activity produced in part in spinal circuits. Thus, transfer clearly indicates that kindling involves development of multiple areas of enhanced epileptogenicity, and this development must depend, in part at least, on synaptic transmission.

Unfortunately trying *not* to find brain damage is akin to testing the null hypothesis and as all students of research methodology know an inadequate sample size and/or poor measurement technique will ensure that no difference between comparison groups is found. With respect to damage, it has been shown that kindled seizures do result in neuronal loss (Cavazos et al., 1994), astrocyte hypertrophy (Khurgel and Ivy, 1996), and other indies that the brain is compromised. Since epilepsy, but not learning is associated with these sorts of changes this leads us to consider kindling as a model of seizures and epilepsy.

Kindling as a Model for Seizures and Epileptogenesis

The generality of kindling raises two important questions: Does kindling capture relevant features of the clinical epilepsies – is kindling a valid 'animal model' of epilepsy?

Kindling provides strong homology – an animal model – for many aspects of seizures and epilepsy – but just which aspects? This is a complex question, because kindling may be used in a multiple ways. Because kindling is both a phenomenon and a technique, we summarize here the phenomenological properties as well as the technical aspects with respect to the clinical condition. We suggest that kindling can be used to model: (1) the *creation* of a seizure focus, (2) the *occurrence* of focal and secondarily generalized seizures (ictal behaviors), (3) either the *progression* in severity of seizures (epileptogenesis) or the *stability* of seizures over time, (4) focal seizures of *unknown cause*, (5) *multiple* seizure generators, (6) pharmacoresistance and conditioned tolerance, and (7) the consequences of multiple seizures, including cognitive dysfunction, personality disturbances, and psychoses (interictal behaviors).

Creation of a Seizure Focus

In the case of posttraumatic epilepsy, tissue that was presumably once normal becomes disrupted in a manner that eventually gives rise to spontaneous seizures. It is evidently not the case that the acute reactive ictal events that occur during the immediate posttraumatic period necessarily cause the appearance of a chronic epileptic condition. Furthermore, the risk factors for reactive seizures are different from those for chronic epilepsy. Thus, some other mechanism, working over months or years, and causing some structural reorganization of the injured tissue eventually results in spontaneous seizures. In the experimentally naive animal, direct application of electrical stimulation that does not initially give rise to an AD can, if it is repeated over several sessions, eventually come to give rise to an AD (Flynn and Teskey, 2007). In this way repeated subthreshold stimulation could model the 'irritation' and lower seizure threshold until seizures emerge.

The Occurrence of Focal and Secondarily Generalized Seizures

Kindling is perhaps most obviously a model for seizures, because seizures can be evoked following brief trains of suprathreshold electrical stimulation. In patients with epilepsy, where activity is generated from a particular focus, the seizures or convulsions are remarkably similar to those shown by the kindled animal stimulated in the same site. Manifestations of both tonic and clonic behaviors are readily induced in the kindled animal. Hence, the initial seizures kindled from amygdaloid or cortical loci resemble either simple partial or complex partial attacks (Racine's stages 1–3), and later seizures that secondarily generalize resemble tonic–clonic attacks (Racine's stages 4 and 5) (Racine, 1972b).

The Progression in Severity of Seizures (Epileptogenesis)

The progression of seizure severity in epileptic populations is somewhat controversial because the widespread use of anticonvulsant drugs prevents an uncontaminated examination of the phenomenon. However, before the introduction of anticonvulsant drugs, it was observed that as seizures recurred there was an increasing tendency for minor attacks to lead to major ones and for the interval between attacks to decrease. The famous quote "seizures beget more seizures" captures this general tendency. It is probably the case that certain subpopulations of people with epilepsy tend to show seizure progression, whereas other subpopulations tend to show stability in seizure severity. The aspects of kindling that model seizure progression are as follows: (1) the unambiguous progression of electrographic and seizure severity with repeated stimulations; (2) the occurrence of focal seizures in one hemisphere leading to the development of independent discharge at sites in the contralateral hemisphere; (3) increased transsynaptic excitability; and (4) the increase in neuronal burst-firing.

Note that kindled seizures are typically elicited, whereas epileptic seizures in patients occur spontaneously. However, some species of animal (e.g., rats, gerbils, cats, dogs, baboons) can be kindled to the point where they exhibit spontaneous seizures, and this contradicts the argument that kindling lacks relevance because the seizures are induced and not spontaneous. Furthermore, the spontaneous seizures that patients express, although not induced by trains of electrical stimulation, are induced by *some* stimulus, probably internal but possibly external (although not obvious). In this way, kindling can be thought to model focal seizures of unknown cause.

Multiple Seizure Generators

The technique of concurrent multiple-site kindling leads to a more rapid progression and greater seizure severity, and this is probably related to the higher incidence of seizures, more severe seizures, and more frequent seizures in individuals with multiple independent spike foci (Teskey et al., 1999), as often seen in patients with several brain tumors.

Pharmacoresistance and Conditioned Tolerance

Focal epilepsy is often refractory to treatment with anticonvulsant drugs. Although the pathophysiologic mechanisms underlying this resistance are unknown, Potschka and Löscher (2005) have identified a proportion of amygdaloid kindled Wistar rats as not responding to anticonvulsant drugs. These rats did not respond to phenytoin with an increase in ADT and were also shown to be resistant to many other major anticonvulsant drugs. A retrospective analysis of nonresponders revealed that plasma concentrations of phenytoin, kindling parameters, electrode location, and placement of electrodes are not responsible for the nonresponsiveness; instead it is related to genetically determined differences in the expression of the multidrug transporter P-glycoprotein. Nonresponding animals may thus prove to be useful in determining the acquired pharmacoresistance that comes with repeated seizures.

In terms of anticonvulsant tolerance, Post et al. (2005) have observed that although carbamazepine, diazepam, and valproate are initially highly effective in suppressing seizures kindled with suprathreshold stimulation, over a brief period of time the drugs lose effectiveness. Tolerance is not observed in animals that receive equivalent doses of the anticonvulsants after each seizure. The loss of anticonvulsant efficacy is thus an associative

phenomenon, termed contingent tolerance, and differs from conventional pharmacokinetic tolerance. After establishment of contingent tolerance to the anticonvulsant effects of drug, anticonvulsant effects can be restored by triggering a number of seizures without the drug or even triggering seizures with the drug given postictally. However, introducing a seizure-free or drug-free period does not reverse contingent tolerance. Thus kindling can again serve as a useful tool for investigating the acquisition of anticonvulsant tolerance in the clinical setting.

The Consequences of Multiple Seizures – Interictal Behavioral Dysfunction

It has been said the behavioral comorbidities with epilepsy are the rule rather than the exception. Cognitive and behavioral abnormalities occur with alarming frequencies in people with epilepsy, including deficits in verbal and visual memory that may not be directly linked to the occurrence of seizures. Furthermore, changes in fear and anxiety are some of the most common emotional disturbances associated with epilepsy of temporal lobe origin. It is often difficult to determine whether behavioral impairments are a result of seizures or other factors, such as medication. Furthermore, given the difficulties with the stigma associated with epilepsy, it is important to ensure that any causal relation between seizures and persistent aberrant behavior is real and unimagined. The fact that certain epileptic-related interictal behaviors can disappear after successful surgical treatment implies that there is a relation between the two. We and others have used kindling to study the relation between epileptogenesis and cognitive change and emotionality (Teskey and Corcoran, 2009).

Conclusion

Kindling involves a profound and permanent increase in the brain's susceptibility to seizures. It is associated with long-lasting changes in sensory, cognition, emotional, and motor behavior and, in some species, with the development of spontaneous seizures. We conclude that kindling is a valid model of clinical epilepsies.

See also: Electroencephalography: Basic Principles and Applications; Epilepsy; Neural Plasticity.

Bibliography

Cavazos, J.E., Das, L., Sutula, T.P., 1994. Neuronal los induced in limbic pathways by kindling: evidence for induction of hippocampal sclerosis by repeated brief seizures. Journal of Neuroscience 14, 3106–3112.

Dennison, Z., Teskey, G.C., Cain, D.P., 1995. Persistence of kindling: effect of partial kindling, retention interval, kindling site and stimulation parameters. Epilepsy Research 21, 171–182.

Flynn, C., Teskey, G.C., 2007. Reduction of seizure thresholds following electrical stimulation of sensorimotor cortex is dependent on stimulation intensity and is not related to synaptic potentiation. Neuroscience 149, 263–272.

Goddard, G.V., McIntyre, D.C., Leech, C.K., 1969. A permanent change in brain function resulting from daily electrical stimulation. Experimental Neurology 25, 295–330.

Goddard, G.V., 1967. Development of epileptic seizures through brain stimulation at low intensity. Nature 214, 1020–1021.

Khurgel, M., Ivy, G.O., 1996. Astrocytes in kindling: relevance to epileptogenesis. Epilepsy Research 26, 163–175.

Monroe, R.R., 1982. Limbic ictus and atypical psychoses. Journal of Nervous and Mental Disease 170, 711–716.

Post, R.M., Zhang, Z.-J., Weiss, S.R.B., Xing, G., Obeng, K., 2005. Contingent tolerance and cross tolerance to anticonvulsant effects of amygdala-kindled seizures. In: Corcoran, M.E., Moshe, S.L. (Eds.), Kindling 6. Plenum Press, New York, pp. 305–314.

Potschka, H., Löscher, W., 2005. Mechanisms of pharmacoresistance in the phenytoin-resistant kindled Wistar rat. In: Corcoran, M.E., Moshe, S.L. (Eds.), Kindling 6. Plenum Press, New York, pp. 315–323.

Racine, R.J., 1972a. Modification of seizure activity by electrical stimulation 1. After-discharge threshold. Electroencephalography and Clinical Neurophysiology 32, 269–279.

Racine, R.J., 1972b. Modification of seizure activity by electrical stimulation 2. Motor Seizure. Electroencephalography and Clinical Neurophysiology 32, 281–294.

Racine, R.J., McIntryre, W., Gartner, J.G., Levitan, D., 1973. Rates of motor seizure development in rats subjected to electrical brain and interstimulation interval effects. Electroencephalography and Clinical Neurophysiology 35, 553–556.

Sramka, M., Deslak, P., Nadvornik, P., 1977. In: Sweet, W.H. (Ed.), Neurosurgical Treatment in Psychiatry, Pain and Epilepsy. University Park Press, Baltimore, p. 651.

Teskey, G.C., Corcoran, M.E., 2009. Interictal behavioural comorbidities in a model of epilepsy. In: Schwartzkroin, P. (Ed.), Encyclopedia of Basic Epilepsy Research, vol. 3. Elsevier, Oxford, UK, pp. 1254–1260.

Teskey, G.C., Thiessen, L.J., Gilbert, T.H., 1999. Alternate-site kindling in the guinea-pig results in accelerated seizure progression and generalization. Epilepsy Research 34, 151–159.

Wada, J.A., Sato, M., Corcoran, M.E., 1974. Persistent seizure susceptibility and recurrent spontaneous seizures in kindled cats. Epilepsia 15, 465–478.

Watanabe, T., Nomura, Y., 1936. Effects of the thyroid gland upon the behaviour and learning ability of rats. Japanese Journal of Psychology 11, 329–335.

Whieldon, J.A., Van Harreveld, A., 1950. Cumulative effects of minimal cortical stimulations. Electroencephalography and Clinical Neurophysiology 2, 49–57.

Kinship in Anthropology

Joanna Overing, University of St Andrews, St Andrews, UK
Paolo Fortis, Centre for Amerindian, University of St Andrews, St Andrews, UK
Margherita Margiotti, University of Bristol, Bristol, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J. Overing, volume 12, pp. 8098–8105, © 2001, Elsevier Ltd.

Abstract

This article outlines the development of kinship studies in anthropology from their beginning to our days. It reviews classic debates on descent and marriage, the role of gender studies in rethinking kinship categories and the more recent contributions approaching the subject from the perspective of the body, aesthetics, and new reproductive technologies. In doing so, this article provides a critical approach to assess the fundamental and ongoing contribution of the kinship studies to the discipline of anthropology.

Studies in kinship, and the related institutions of marriage and the family, have until recently been central to anthropological investigation and debate. The problem of handling the topic of kinship cross-culturally is the key to understanding the historical development of a large majority of anthropology's central analytic concepts, theories, and methods. Deliberations over the puzzles of kinship and marriage gave rise to the discipline's most sophisticated technical and theoretical elaborations, and its most virulent, ever-present controversies. It was also assumed to be the area of technical competence, the most demanding of rigor in thought, through which anthropology could best defend its scientific respectability. As a result, kinship was that aspect of social life that became the linchpin for the unfolding of all the grand paradigms of thought within anthropology, whether it be Morgan's narrative of evolutionism (1871), Malinowski's functionalism (1930), Radcliffe-Brown's structural-functionalism (1952), or Lévi-Strauss' structuralism (1969). As Fox (1967) could still comment, "kinship is to anthropology what logic is to philosophy or the nude is to art; it is the basic discipline of the subject." The situation has, however, changed.

Nowadays at least two major questions can be asked. First, why did anthropologists predominantly prioritize kinship over politics, economics, and religion? Second, why lately have they shown a decided *lack of interest* in kinship, or rather in 'kinship' as it was conceived traditionally within anthropology? The answers will take us to problems in the major presuppositions underlying anthropology's highly valued analytic constructs of kinship and society. Although it has been the case that kinship from the start has always been the object of strident debate, it is only since the 1960s that the legitimacy of the topic itself began to be queried. In more recent years the process of doubting has accelerated due to very basic epistemic shifts within anthropology in the wake of feminism and other modes of disciplinary self-inspection about its claims to knowledge. One prevalent conclusion forthcoming from such transformations is that much of received kinship theory is no longer seen as justifiable.

The Narrative of Kinship

Law and Order by Another Name

The primary puzzle for the anthropologist throughout the first half of the twentieth century was how to explain the maintenance of order within the 'simple societies' of far-flung regions where anthropologists conducted their research. Such societies lacked the basic law and order organizing institutions of Western society. They had no government to speak of, no marketplace, as we know it, and no law courts, police, or armies. It was clear that they did not compartmentalize their social life into the distinct and separate institutions that we recognize as kinship, economics, politics, and religion. Anthropologists found instead that these peoples used the *idiom of kinship* to frame most of their activities, including those with political, economic, and religious intent. Analytically, the step from this insight was to view kinship to be the major institution of 'tribal' societies, and the kinship tie to be the one that compelled all others in social relations. Kinship, as the strongest of social bonds, became seen as the basis through which 'primitive' societies maintained order; it was through kinship ties that people created relations of social solidarity. Thus, 'social structure,' that is, those rules regulating the kinship, marriage, and residential institutions of a people that endow social role and identity, and which, therefore, perpetuate societal relationships, became anthropology's proper object of inquiry (e.g., see Radcliffe-Brown, 1965). Everything else, a society's morals, law, etiquette, religion, politics, and education, was to be studied as but an aspect of social structure (Radcliffe-Brown, 1965), or in other words its kinship system.

The emphasis anthropologists placed upon the problem of 'societal order' cannot be stressed too much. As Firth comments, the perception of order was fundamental to their investigation (Firth, 1961). An underlying concern was over what could replace the authority of government in 'simple' societies, and the answer was to view kinship as having this coercive power. In other words, it was the kinship system that came to be understood as playing the same official function among 'primitive' peoples as government does in western societies. The reasoning leading to such a conclusion is circular,

for kinship as a system became defined as the primary source for the rules and regulations that provided for order and continuity within 'native' society. It was presumed that the status, role, rights, duties, and obligations of a person in a 'simple society' were forthcoming from and ascribed by the person's place within the kinship system. The primary societal organization of these societies was then understood to be 'kinship based,' and it was the 'kinship polity' (see Fortes, 1969) forthcoming from either patrilineality or matrilineality that was the key concern. The slippage was simple: we have government, while they have the politics of kinship.

The Plot Thickens: The Distinction between the Domestic and the Jural

The 'law and order' thrust of traditional kinship studies, which equated the kinship system (of 'primitives') to 'society' itself, also came to include the critical distinction between the 'domestic' and 'jural' domains of kinship. The realm of the domestic was composed of relations of filiation between parents and children; it was the domain of the hearth, the family, the husband, his wife, and his children. In Fortes' terms (1969), it was the domain where the 'axiom of amity' reigned. In contrast, within the 'jural' domain the everyday relations of amity and filiation were for the most part irrelevant, for the principles of descent and lineage ruled its membership and provided the backbone for its jural structures of dominance and subordination. It was the jural domain that comprised the polity and, thus, provided society with its order and continuity. The prescriptions and regulations of the kinship polity (composed of its 'corporate descent groups') were what ruled and constrained 'primitive' people. A man's jural status, rights, and obligations within society were in essence provided by his place within the lineage of his birth. Genealogy determined one's political status, and one's rights to land and other entitlements.

Given the above narrative of the place of kinship in 'native' societies, we can understand that the received wisdom of kinship theory until as late as the 1970s was that unilineal descent systems were necessary as a sticking plaster of 'primitive' societal order – despite increasing evidence to the contrary. As Radcliffe-Brown (1965) asserts, "unilineal institutions in some form, are almost, if not entirely, a necessity in any ordered social system." Even Lévi-Strauss, in his major critique of descent theory in kinship studies, finds the existence of unilineal descent essential to the logic of his model of elementary structures of marriage exchange. He says that this is because the social cohesion of elementary systems of kinship that are premised on the notion of groups of men exchanging wives requires a rule of descent, for the groups themselves must be defined by such a 'stable' rule of descent (Lévi-Strauss, 1969). One should be aware, however, that Lévi-Strauss began his formidable attack upon the prevalent view of 'descent as societal order' in 1949, and that his stress upon alliance over descent as the salient ordering principle of 'primitive society' did not become part of mainstream debate until the 1960s (e.g., Leach, 1961; Needham, 1964), but also see the earlier debate between Radcliffe-Brown (1953) and Dumont (1953). The concern of Lévi-Strauss, it is to be noted, was with creating a minimal model of society by showing the ways in which kinship groups were integrated through rules of marriage exchange, that is, he wished to demonstrate how classifications of kinship and marriage logically provided a broader level of societal integration than that achieved through rules of descent alone.

The rather chauvinistic reductionism of the structural functionalist and structuralist grand paradigms of society and societal ordering, where society itself was to be equated with male structures of order and control, took another couple of generations to unveil and unravel.

The Question of Definitional Rigor

Kinship, Descent, and Marriage

In itself, a fuller ethnographic record began steadily to undermine many of the major analytic constructs of kinship and societal order, particularly the idea that 'primitive society' was universally based upon exogamous, corporate, land-holding, unilineal descent group structures. By the 1950s and 1960s reports of field research especially from the Pacific on kinship systems that were not premised on a unitary rule of unilineal descent but upon cognatic, bilineal, ambilineal, or double descent principles became legion (cf Bohannan and Middleton, 1968). The debate over the unitary view of unilineal descent basically was closed by the influential article by Scheffler (1966), who was able to demonstrate through the ethnography by then at hand that notions of descent were used among different peoples, and by the same people, toward highly varied ends, and not necessarily toward that of corporate group structure.

By the 1970s, for instance, through the ethnography from Amazonia that began to enter mainstream debates on kinship and marriage, it became clear that the notion of descent itself could hardly be declared a universal principle of 'primitive' social ordering, for there were peoples who did not recognize a principle of descent as relevant for any social or political purpose (e.g., see Rivière, 1969; Overing, 1975). Even kinship and marriage as analytical constructs per se came also under attack, especially the notion of achieving any type of unitary view of either (but also, see Leach's (1961) earlier hacheting of these sacred constructs of societal ordering). We find Needham (1971) announcing flamboyantly that "there is no such thing as kinship, and it follows that there can be no such thing as kinship theory!" He was referring to 'minimal' definitions of kinship when framed in the context of genealogically reckoned jural rights, such as their allocation and transmission from one generation to the next. Cross-culturally, the ethnographic evidence could not uphold a totalizing view that assumed a predictable relationship between the cultural constructs of kinship and its classification, social roles, rights, and obligations, and the allocation of individuals to particular types of social groups. It was this unitary package of kinship as part and parcel of particular politico-jural orders that earlier anthropology had indeed upheld.

Even more courageous for the times are Rivière's (1971) queries into the analytic concept of marriage. He argues against any jural definition of marriage, and suggests instead that the institution of marriage first be viewed structurally, as one of many relationships conceived possible between men

and women. It was his reading of the ethnographic literature that anthropologists had been defining institutions cross-culturally as 'marriage' when the said institutions in fact had "no feature in common other than that they are concerned with the conceptual roles of male and female" (Rivière, 1971). In other words, to understand what marriage is for any given people the question of the cultural construction of gender relationships must be understood, rather than the jural relations between groups of men that entail their exchange of women.

It is significant that Rivière's fieldwork experience had been with indigenous peoples of Amazonia, for whom anything approaching a 'jural' relationship usually would be stretching the point, as too would lineages and descent group ordering as normally discussed in the literature. It was as difficult to find corporate land-holding groups among Amazonian people, as the elders who might rule them. There were no *groups* of men forming ties of alliance through the exchange of their women. Instead, ties of marriage, which moreover were highly salient to Amazonian constructs of sociality, were linked to a principle of cognation and not descent. As a consequence the contributions of Amazonian specialists, in line with Dumont's (1957) reading of marriage alliance in India, played a major part in the later reinterpretation and unraveling of both alliance theory as first formulated by Lévi-Strauss and descent theory as formulated by Fortes and Radcliffe-Brown.

The Structural Analysis of Kinship Terminologies: The Power of the Paradigm

It is by now obvious that we cannot achieve an analytic definition of the construct of kinship and related institutions that would be both universally adequate and at the same time respectful of indigenous understandings and knowledge. In short, anthropology cannot, even if it wished, arrive at a universal definition of kinship. Part of the dilemma is of course linguistic, insofar as most of the important analytic terms of anthropology have not only been highly abstract but also continued to carry the complex historical baggage of western thought and practice. Terms, such as society, community, family, kinship, descent, lineage, or structure, function, and system, are to be used at the peril of totally eluding another people's understanding of what they are doing socially – which, one would think, is the very *raison d'etre* of the anthropological task. As anthropology slowly came to realize, any 'definitional rigor' to be achieved through the use of any of these terms is well-nigh impossible. Such constructs tend to sit within particular and forceful paradigms of social order, and therefore carry all the litter of such grand narratives. Structural-functionalism was followed by structuralism, and the structural analysis of kinship terminologies was a particularly obvious case of a paradigm so powerful that its highly reductive results killed for the time being the possibility of interesting further advances being made in kinship theory, the very area where anthropology was once so creative and rich in debate.

The overwhelming allure of structural analyses of kinship can be ascribed to the power of their methods, which wed anthropology to advances in modern linguistics, a field considered to have become the most scientific of the human sciences. The methods and models of formal analysis were so enlightening of previously misunderstood structures that they gave promise to a mathematical rigor and definitional clarity that would transform anthropology into a 'true science.' Their initial success was so stupendous (e.g., see Lounsbury, 1968) that they seemingly made child's play of previous attempts to provide order to the complicated structures of many 'native' kin terminologies. Anthropologists were taught to be more rigorous in discerning the logical differences between systems. Such sophistication in method was greatly needed in anthropology, and for its example we can be only grateful.

However, as Overing (1987) has noted, *the method* became confused with worldview. The logic of the method through sleight of hand became equated to the logic of terminological use, and thus also with indigenous thought itself. As a result, we arrived once more at the 'universal' to what Schneider so aptly derides as the anthropological 'Doctrine of the Genealogical Unity of Mankind!' (Schneider, 1984) or the genealogical meaning of kinship terms. Kinship is, everywhere, first and foremost about genealogical relatedness (see especially, Scheffler, 1978), a resoundingly uninteresting conclusion to come in the wake of such dazzling structuralist performances. It was also a judgment that was highly suspect. It appears that the method itself allowed for no other interpretation because the meanings of terms were made to fit not only the domain of logic, but also western commonsense notions about what kinship is. The demand itself of formal analysis for logical rigor reduces the number of its elements that account for meaning to very few: affinity as well as consanguinity can be allowed, but not the further complications of what such notions might possibly mean from the indigenous point of view. It was the niggling doubt over this issue (where are all the people?) that drove many anthropologists away from the technical chore of analyzing kinship logics. As Campbell (1985) remarks, the very abstract level at which structural analysis operates is about the tenth remove from anything going on in daily practice and thought.

In somewhat more recent formal analyses of kinship terminologies in the hands of scholars, such as Héritier (1981) we moved in the more promising direction of not assuming indigenous meaning. The aim instead is to specify the structural possibilities of a given classificatory system that in turn frame or limit the choices made in the world of reality. In short, the exercise is no longer directed toward the task of uncovering hidden layers of meaning, but to explore the interplay between structure and the semantics of actual use. This, indeed, was a road worth traveling.

Is There Hope for Kinship?

Schneider (1984) advises anthropologists not only to stop looking for 'kinship,' but also all the other meta-cultural categories of economics, politics, and religion which, together with kinship, form a quartet of societal domains. This was an idea embedded in European culture that has been willy-nilly incorporated into the European social scientist's analytic schemes. While we can heartily agree with Schneider's full blown rampage against the anthropological treatment of the topic of kinship, such concordance does not, however, entail total dismissal of the study of those social relationships, and

their classifications, that were once, more or less, subsumed under the label of 'kinship.' People do bear children, and there *is* a social framework through which they do so, and through which these children are raised to become adult members of human social groups. The members of these social groups follow particular practices in the course of which relationships that are highly significant for them are developed, as too are very interesting ways of thinking about them. With all this we can agree. The overriding question still remains – how do we understand and translate such practices, relationships, and ways of thinking?

Happily, since the 1980s, anthropologists have developed a myriad of ways to approach subjects that formerly would have been classified under the general rubric of 'kinship.' While nowadays the topic of something called 'kinship' does not loom large in the literature, such issues as 'self,' 'agency,' 'gender,' 'body,' 'intimacy,' and the life of values and affect, do. The topics of personhood, emotions, and aesthetics are much more likely to take their place in the titles of doctoral theses than those of kinship, affinity, and jural rules. Thus, we see that the 'technical' language of anthropology has been transformed in the wake of the shifts of attention away from something we once called the 'jural-political domain,' with its contrast to something we labeled 'the domestic group.' Attention is removed from the notions of social structure and prescriptive behavior, and those of rights and obligations. In their place, idioms of equality and inequality are now being explored, and such values as nurture, sharing, generosity – and those of peace and violence. The stress is upon ambiguity, process, the everyday, and a multiplicity of voices, rather than upon grand structures, rules and regulations. The emphasis tends to be upon context and the performative, and not hidden rules of practice and thought. Such shifts in direction have often been undertaken in a spirit of rebellion, by feminists, but also many others, against the ethnocentrisms underpinning the narratives of anthropology. The gain has been that anthropologists are presenting very different pictures, certainly in their richness, of the ways other people view, act, and experience the world. In short, it is rich ethnography, not grand theory, that drives our students and most of our colleagues today.

The Impact of Gender Studies

It has also made all the difference to the ethnographic endeavor to include the subject of women, and especially women's voices. In the process we have time and again discovered a rich symbolic world that demands an understanding of the interplay of men and women, and their respective knowledges, that makes travesty of the simplistic 'male only' models of yesteryear. It is only by understanding both male and female (often complementary or conflicting) perspectives that gender relations among another people can begin to be comprehended. As gender studies progressed, it became absolutely clear that the authoritative voice of the ethnographer was insufficient, as too that of the male 'informant.' Nor can we speak from the perspective of a generic female, for there is no such thing. There is ever a plurality of voices, a seemingly helter-skelter chiming that provides, significantly, fertile layers of evaluative contextualization, the recognition of which has in the end changed our anthropological visions of culture and society.

All those kinship structures through which men established important relationships with each other through the exchange of their muted women, which became the model of 'society'; all those 'political' roles and statuses through which men who controlled the political domain came to be the knowledge holders of their culture – these were the topics that once were recognized as primary anthropological concerns. From the 1970s onward, once women were inserted firmly within the context of 'establishment' discourse in anthropology, the consequence was a true 'Kuhnian' shift of paradigm.

One major contribution of the early writings on gender was to make anthropology clearly aware of its exclusion of women as a topic for study, and how such oversight was, moreover, linked to the male bias of dominant theoretical and methodological assumptions of the discipline, and beyond that to major strands of western thought in which anthropology played its part. In *Women, Culture & Society*, its authors show how women, like men, transculturally "are social actors whose goals and strategies are intrinsic to the process of social life" (Rosaldo and Lamphere, 1974), and that "even in situations of overt sex role asymmetry women have a good deal more power than conventional theorists have assumed" (Rosaldo and Lamphere, 1974). It became quickly obvious that major presuppositions and procedures of the discipline needed to be reconsidered radically in order to include the agency of women. In other words, studies of social life and its organization needed to include women as actors and speakers and not as mere passive, muted objects of men. It slowly came to be seen that women as well as men played a major role in the creation of social ties of community life.

There is by now a large amount of ethnography telling of indigenous peoples among whom the women, both ideologically and in practice, control their own labor, and the products of it, as shown in the recent literature on Amazonia. The ethnography of Lowland Central and South America has shown that activities of both men and women – different but complementary – reproduce sociality (Overing, 1986; McCallum, 2001; Belaunde, 2001). Gender is seen as a matter to be created, and formed, and not as a fixed attribute of bodies. The topic of corporeal substances is often found to be the key to understanding the gendering of bodies. In Amazonia, it is particularly blood and menstruation that play a crucial role, as Belaunde has demonstrated:

> Blood among indigenous Amazonian people brings together the biological, the intellectual and the spiritual, because the management of the flow of blood during the life cycle is a constitutive process, not only of fertility, but also of health, labor, creativity, well being, spirituality, personal identity and interethnic relations. To understand the difference between men and women from an Amazonian perspective necessarily entails a comprehension of the scope of corporeality, created through dietary and seclusion practices, and of predation and communication with cosmic beings, linked to the flow of blood.
>
> Belaunde, 2005: 18–19

Likewise, other substances, such as breast milk, contribute to the formation of gendered bodies. For example, among the Kuna, baby boys and girls are differently treated in relation to such maternal substance, which is not only conceived as

nurture, but also provides their bodies with gender features (Margiotti, 2010).

Values – and the structures of equality and inequality linked to them – may not be so straightforward, but ambiguous, as among the Hagen of New Guinea where there is constant play between egalitarian and hierarchical principles (e.g., Strathern, 1988). There are also now many ethnographic examples showing that the quality of the relations between the genders may well be subject to perspective, where females have a very different view from males of the strength of their respective roles and participation within the social life of the community (e.g., Keesing, 1987). This stress in studies of gender upon the critical importance of presenting perspective has led to energetic debate over the epistemological foundations of anthropology, which in turn has transformed the question, the topics, and the methodology of the discipline – and in the end its own self image as having the right of authorial privilege.

The Emergence of the Body and Aesthetics

The creation, gendering, and more generally, the treatment of bodies in different phases of the life cycle, has become an increasingly important topic in the study of indigenous socialities in the last 30 years and has provided crucial new insights into their kinship systems. By looking at how bodies are fabricated through daily processes of nurturing, conviviality, the observance of taboos, ritual seclusion and body decorations, Americanist anthropologists have been able to gain a better understanding of how Amerindians conceive themselves as living in groups of people related by substances. In brief, to recreate and maintain society it is essential to make human bodies that share the same qualities and affects. Often this process of creating kinship is understood to be in contraposition to opposing predatory forces, such as those of animals, that seek to incorporate human beings into their society by way of attaching their own, animal, affects onto the bodies of newborn humans (e.g., Londono Sulkin, 2000, 2012). Hence, understanding indigenous ideas of the social entails understanding their ideas and practices relating to bodies (Seeger et al., 1979). It is therefore of paramount importance to understand how, in each specific case, bodies become the focus of ritual practices and processes of feeding. In short, we are realizing the importance of understanding the specific ideas of the body that each society holds. The practices of bodily fabrication are usually aimed at imparting young people with the specific knowledge that teaches them to create 'bodies that know' (McCallum, 1996), that is, endowed with those moral human capacities that distinguish humans from other beings and that form the basis for the constitution of kinship. The process of fabricating the bodies of kinspeople entails the skillful management of multiple relationships with alterity. This is a topic that has interested a number of scholars in recent years, and which has been dealt with through an alternating focus on how relations are managed both inside the group of living human beings and outside in the wider space of neighboring groups, enemies, allies, and nonhuman beings (Overing, 1985b; Londoño Sulkin, 2000, 2012; Viveiros de Castro, 2001; Vilaça, 2002).

A further elaboration, in recent years, has been that of unraveling how decorations and ornamentations are key practices for the fabrication of bodies, which indeed endows them with those qualities that render them human and thus capable of regenerating kinship relations. This is a growing field of study that has the merit of showing how aesthetics is an extremely productive field to refine our understanding of indigenous forms sociality, and ideas of personhood and agency. Bodily decorations, such as body painting, tattoos, featherwork, beadwork, and clothing, are often ways of rendering visible the inner qualities of persons (Strathern, 1979), their capacity of fostering creative forms of sociality (Overing, 1989), and women's capacities to bear children (Gow, 1999). Indigenous notions of aesthetics are indissolubly linked to the creation of bodies, and therefore their study sheds light on how persons and social praxes are constituted (Fortis, 2010). Body painting and other forms of decoration are not only means of creating human bodies, they also make bodies resistant to predatory attacks from malevolent beings (Lagrou, 2007), or endow them with the capacities to act in social contexts inhabited by foreign others (Ewart, 2007). It has therefore become increasingly evident how ideas and practices originally considered, from a western perspective, as pertaining exclusively to the field of aesthetics, are in fact part of the whole fabric of everyday life, which includes, but is not limited to, those categories that were traditionally the object of kinship studies.

Reproductive Medicine

As Schneider (1984) argued, an axiom critical to kinship theory has been that the social and cultural attributes of kinship are derivative of the biological relations of reproduction. A reaction to Schneider's work has been a bourgeoning interest in the meanings of kinship in Euro-American contexts. Reproductive medicine and biotechnological advances have provided a fascinating context for the analysis of kinship meanings and practices aimed at forging new families (Strathern, 1992, 2005; Edwards, 2000; Carsten, 2000; Edwards and Salazar, 2009). Thus, one new 'anthropology of kinship' analyses Euro-American understandings of relatedness by putting on the map topics, such as adoption and foster care, surrogate motherhood, and gamete donation in national and international contexts (Ragoné, 1994, 2000; Howell, 2006). This further illuminates the link between kinship and new forms of commodification of bodies and reproductive material, as well as issues of race and inequality (Birenbaum-Carmeli and Inhorn, 2009). By the same token, these new kinship studies have signed a definitive rupture with old paradigms of nature and culture divide, rethinking ethnographically ideas of the natural and the artificial (Strathern, 1992). On a theoretical level, some of these studies confirm the importance of procreation in western ideas of kinship. Procreation is viewed as a transaction of genetic material. This material is 'passed down' across generations to form new relatives (Strathern, 1992, 2005). Idioms of genetics are therefore employed to refer to notions of origin and belonging, which are key to western ideas about kinship (Edwards and Strathern, 2000). Likewise, technological novelties in the field of reproduction and the new forms of families therefore emerging are interpreted through kinship ideas concerning the importance of

localities, naming, nurturance, and affects. Nurturing and bringing up children are equally constitutive to kinship, and remain key topics to understand changing ideas and practices in contemporary societies.

Kinship by Another Name? Networks of Relationships and Personal Kind Terms

We find that Schneider and Needham, as pioneers in the deconstructing of key concepts of kinship theory, were merely tapping the surface of a modernist creation for which the very notion of kinship was but one aspect of a multifractaled edifice filled with assumptions about society and the social order, that in turn were tied to networks of ideas about the relation of the family to other societal institutions, and the relationship between the sexes, the private and the public, the dominant and the subordinate, all of which were implicated further by presuppositions about the nature of human existence, and its progress, which in addition were premised upon ontological notions of the priority of reason over the emotions, and so it continued through an enormous number of other dualisms and bundles of relations pertinent to the western imagery of society and the world, and the elements of which they were comprised. Our notion of kinship carried with it the interarticulations of this entire structure.

The interesting lesson that has been learned more recently through changing the types of questions anthropologists' ask is that other people's views of the social relationships of everyday life are as enmeshed as our own within wider networks of meaning. They also include ontological presuppositions about the nature of human existence and capabilities for sociality that are linked to bundles of relationships and arrays of ideas about the world that are as complex as our own. How do we then understand these other sets of linkages, so unlike our own? That has become a primary question. The quest for understanding their interconnections has in part required the unpeeling of our presuppositions about reproductory and biological processes, parenting, the nature of the material world, and the interconnections of all of these things to what we call Society (cf Strathern, 1992).

Overing (1985a) has suggested that to understand better the complexity of indigenous social thought, we should change the label of what we have been calling 'kinship terms' to 'personal kind terms.' This involves a radical switch in perspective that concomitantly raises the conceptual status of these terms to one more closely aligned to the indigenous view and practice. As with many of our own constructs of 'natural kinds,' 'personal kind terms' are also highly abstract, philosophically important concepts that defy unitary definition (e.g., see Overing, 2006). They share the open-endedness and elusiveness that is typical of all abstract terms that comprise complex relational properties. The difference is that personal kind terms do not refer to the world of nature, which is the western domain of competence, but to qualities of personal relationships, the area that indigenous people have opted for theoretical elaboration (cf Horton, 1979). We have reduced their personal kind terms to our own very weak language of kinship, one that speaks of 'consanguinity,' 'affinity,' 'social category,' and 'amity,' which is often a bad mistake. These terms, as used in everyday life, can have metaphysical weight that goes well beyond our western notions of a 'biological' relationship or one through marriage. Moreover, the quality of the relationship they demonstrate, as, for instance, one of nurturing, teaching, treachery, competition, or predation, may override a more 'physical' kind of relating. Will this man or woman work tranquilly with me, or have predatory designs that will make me ill?

Certainly in Amazonian ethnography, the emphasis today in investigations of peoples' use of relationship terminologies is often upon the metaphysical and/or moral loading of the classifications (e.g., Teixeira-Pinto, 1997; Viveiros de Castro, 1992; Overing, 1985a). For instance, Kidd writes about the Enxet of Paraguay that:

> Their understanding of why they act as they do centres on their concept of the waxok, an aspect of the self that is both intensely private and inherently social. They insist that their social behaviour – both appropriate and inappropriate – can be explained by the physical – or metaphysical – state of the waxok. Furthermore, because the waxok is also the centre of cognition, people can also consciously transform it so as to enable themselves to act in either a self-centred or other-regarding manner. It is an explanation that, I believe, we should take seriously if we want to understand indigenous social life and it is one that finds its root in the practice of child-raising, in the creation of 'good/beautiful' people who have been taught not only how to think but how to feel. It is this waxok-centred combination of thinking and feeling that enables the Enxet to act appropriately and which, ultimately, guides them as they strive to generate sociality and engender tranquillity.
>
> Kidd, 1999

The peoples of Amazonia have been difficult for anthropologists to believe, and to an interesting extent still are (cf Overing, 1999). For many of these people the personal ties of parenting, nurturing, sharing, and pooling are not premised primarily on a notion of 'biological linkage' or a 'linkage through blood,' or membership within a jural group, but which instead are generated over time through consistent and processual action. As such, the personal quality of a relationship as made manifest through everyday practice is paramount to its classification. In other words, the ongoing quality of a relationship (as in the 'growing' of a child by a parent and child, the mutual care of brothers, or sisters, or of any other personal relation) may have generative value in a material sense that goes far beyond the minimal possibilities that are endowed through the act of sex (e.g., Gow, 1991). For sure, the personal kind terms that are here being applied to their nearest and dearest, and to those further a field, are about reproductory possibilities, but it is not 'reproduction' in the sense that the anthropologist imagined it in traditional kinship theory.

Kinship Is Alive and Well!

In fact, we can say that kinship studies are as alive and well as always in anthropology, in that the personal relationships and activities of parenting, nurturing, and the processes of generating and gendering of bodies into social adulthood hold center attention. Such studies are, however, unrecognizable as pertaining to the kinship theory of yesteryear. A similar

reorientation of concentration is as clear in Melanesian studies as in those of Amazonia (e.g., Strathern, 1988) where personal relationships are discussed through categories very different from the former ones of prescriptive rule, roles and statuses, and social structure, and where the talk instead, as among Amazonianists, is upon the indigenous understandings of such matters as gender distinctions, the contents of the self and the mastery of them, and the construction of human bodies; it is about indigenous ambiguities over the nature of personhood and the various possibilities of agency in this world and others, and the elaborate relation of these issues to indigenous practice, the life of affect, and their metaphysics. In other words, a dialogue is being created between us, the anthropologists, and them, the peoples of New Guinea or Amazonia, over what it means to be human in this world, or female humans, male humans, or betwixt and between? What does it mean to be social beings in this world? How do we go about attaining this state? If for other people the western grand distinctions between society and nature do not hold, what other possibilities of an interesting kind are there? These are very different questions from those asked in mainstream anthropology in the 1970s and 1980s.

See also: Ancestors, Anthropology of; Clan; Family: The Anthropology of the Concept and Its History; Genealogy in Anthropology; Marriage.

Bibliography

Belaunde, L.E., 2001. Viviendo bien: género y fertilidad entre los Airo-Pai de la Amazonía peruana. Centro Amazónico de Antropología y Aplicación Práctica, CAAAP, Lima.

Belaunde, L.E., 2005. El recuerdo de Luna: Género, sangre y memoria entre los pueblos amazónicos. Fondo Editorial de la Facultad de Ciencias Sociales, Lima.

Birenbaum-Carmeli, D., Inhorn, M.C., 2009. Assisting Reproduction, Testing Genes. Global Encounters with the New Biotechnologies. Berghahn Books, Oxford, UK.

Bohannan, P., Middleton, J. (Eds.), 1968. Kinship and Social Organization. Natural History Press, Garden City, NY.

Carsten, J., 2000. Cultures of Relatedness: New Approaches to the Study of Kinship. Cambridge University Press, Cambridge, UK.

Dumont, L., 1953. The Dravidian kinship terminology as an expression of marriage. Man 53 (224), 143.

Dumont, L., 1957. Hierarchy and marriage alliance in South Indian kinship. Occasional Papers of the Royal Anthropological Institute no.12.

Edwards, J., 2000. Born and Bred. Idioms of Kinship and New Reproductive Technologies in England. Oxford University Press, Oxford, UK.

Edwards, J., Salazar, C., 2009. European Kinship in the Age of Biotechnology. Berghahn Books, Oxford, UK.

Edwards, J., Strathern, M., 2000. Including our own. In: Carsten J (Ed.), Cultures of Relatedness: New Approaches to the Study of Kinship. Cambridge University Press, Cambridge, UK, pp. 149–166.

Ewart, E., 2007. Black paint, red paint and wristwatch: the aesthetics of modernity among the Panará in Central Brazil. In: Ewart, E., O'Hanlon, M. (Eds.), Body Arts and Modernità. Sean Kingston Publishing, Wantage, pp. 36–52.

Firth, R., 1961. Elements of Social Organization. Beacon, Boston [1951. Watts & Co., London].

Fortes, M., 1969. Kinship and the Social Order: The Legacy of Lewis Henry Morgan. Routledge & Kegan Paul, London.

Fortis, P., 2010. The birth of design: a Kuna theory of body and personhood. Journal of the Royal Anthropological Institute 16 (3), 480–495.

Fox, R., 1967. Kinship and Marriage: An Anthropological Perspective. Penguin, Harmondsworth, UK.

Gow, P., 1991. Of Mixed Blood. Clarendon Press, Oxford, UK.

Gow, P., 1999. Piro designs: painting as meaningful action in an Amazonian lived world. Journal of the Royal Anthropological Institute 5, 229–246.

Héritier, F., 1981. L'Exercice de la Parenté. Gallimard: Seuil, Paris.

Holy, L., 1996. Anthropological Perspectives on Kinship. Pluto Press, London.

Horton, R., 1979. Material-object language and theoretical language: towards a Strawsonian sociology of thought. In: Brown, S.C. (Ed.), Philosophical Disputes in the Social Sciences. Harvester Press, Brighton, UK.

Howell, S., 2006. The Kinning of Foreigners: Transnational Adoption in a Global Perspective. Berghahn Books, Oxford, UK.

Keesing, R., 1987. Ta'ageni: women's perspective on Kwaio society. In: Strathern, M. (Ed.), Dealing with Inequality: Analysing Gender Relations in Melanesia and beyond. Cambridge University Press, Cambridge, UK.

Kidd, S.,1999. Love and hate among the people without things: the social and economic relations of the Enxet people of Paraguay (Ph.D. thesis). University of St Andrews, UK.

Lagrou, E., 2007. A fluidez da forma: arte, alteridade e agência em uma sociedade Amazônica. Topbooks, Rio de Janeiro.

Leach, E.R., 1961. Rethinking Anthropology. Athlone Press, London.

Lévi-Strauss, C., 1969. The Elementary Structures of Kinship. Beacon, Boston [1949. Les Structures Élémentaires de la Parenté Presse. Universitaires de France, Paris].

Londoño-Sulkin, C., 2000. Though it comes as evil, I embrace it as good: social sensibilities and the transformation of malignant agency among the Muinane. In: Overing, J., Passes, A. (Eds.), The Anthropology of Love and Anger: The Aesthetics of Conviviality in Native Amazonia. Routledge, London, pp. 170–186.

Londoño-Sulkin, C., 2012. People of Substance. An Ethnography of Morality in the Colombian Amazon. Anthropological Horizons. University of Toronto.

Lounsbury, F., 1968. The structural analysis of kinship semantics. In: Bohannan, P., Middleton, J. (Eds.), Kinship and Social Organization, first ed. Natural History Press, Garden City, NY, pp. 125–148.

McCallum, C., 1996. The body that knows: from Cashinahua epistemology to a medical anthropology of Lowland South America. Medical Anthropology Quarterly 10 (3), 347–372.

McCallum, C., 2001. Gender and Sociality in Amazonia: How Real People Are Made. Berg, Oxford.

Malinowski, B., 1930. Kinship. Man 30, 9–29.

Margiotti, M., 2010. Kinship and the saturation of life among the Kuna of Panamá (Ph.D. thesis). University of St Andrews, UK.

Morgan, L.H., 1871. Systems of Consanguinity and Affinity of the Human Family. Smithsonian Institution, Washington, DC.

Needham, R., 1964. Descent, category, and alliance in Siriono society. Southwestern Journal of Anthropology 20, 229–240.

Needham, R., 1971. Remarks on the analysis of kinship and marriage. In: Needham, R. (Ed.), Rethinking Kinship and Marriage. Tavistock Publications, London, pp. 1–34.

Overing, J., 1985a. Today I shall call him 'Mummy': multiple worlds and classificatory confusion. In: Overing, J. (Ed.), Reason and Morality. Tavistock Publications, London, pp. 152–179.

Overing, J., 1985b. There is no end of evil: the guilty innocents and their fallible God. In: Parkin, D. (Ed.), The Anthropology of Evil. Basil Blackwell, Oxford, pp. 244–278.

Overing, J., 1986. Men control women? the 'Catch 22' in gender analysis. International Journal of Moral and Social Studies 1 (2), 135–156.

Overing, J., 1987. Translation as a creative process: the power of the name. In: Holy, L. (Ed.), Comparative Anthropology. Blackwell, Oxford, UK, pp. 70–87.

Overing, J., 1989. The aesthetics of production: the sense of community among the Cubeo and Piaroa. Dialectical Anthropology 14, 159–175.

Overing, J., 1999. Elogio do cotidiano: a confiança e a arte da vida social em uma comunidade amazónica (In praise of the everyday: trust and the art of sociality in an Amazonian community). Mana 5 (1), 81–108.

Overing, J., 2006. The stench of death and the aromas of life. Tipiti, Journal of the Society for the Anthropology of Lowland South America 4 (1 & 2).

Overing, K.J., 1975. The Piaroa, a People of the Orinoco Basin. Clarendon Press, Oxford, UK.

Ragoné, H., 1994. Surrogate Motherhood: Conception in the Hearth. Westview Press, Boulder.

Ragoné, H. (Ed.), 2000. Ideologies and Technologies of Motherhood: Race, Class, Sexuality, Nationalism. Routledge, New York.

Radcliffe-Brown, A.R., 1953. Dravidian kinship terminology. Man 53 (169), 112.

Radcliffe-Brown, A.R., 1965. Structure and Function in Primitive Society. Free Press, New York [1952. Cohen & West, London].

Rivière, P., 1969. Marriage Among the Trio: A Principle of Social Organisation. Clarendon Press, Oxford, UK.

Rivière, P.G., 1971. Marriage: a reassessment. In: Needham, R. (Ed.), Rethinking Kinship and Marriage. Tavistock Publications, London, pp. 57–74.

Rosaldo, M.Z., Lamphere, L., 1974. Introduction. In: Rosaldo, M.Z., Lamphere, L. (Eds.), Woman, Culture and Society. Stanford University, Stanford, CA, pp. 17–42.
Scheffler, H.W., 1966. Ancestor worship in anthropology or, observations on descent and descent groups. Current Anthropology 7, 541–551.
Scheffler, H.W., 1978. Australian Kin Classification. Cambridge University Press, Cambridge, UK.
Schneider, D.M., 1984. A Critique of the Study of Kinship. University of Michigan Press, Ann Arbor, MI.
Seeger, A., Da Matta, R., Viveiros de Castro, E., 1979. A construção da pessoa nas sociedades indígenas Brasileiras. Boletim Do Museu Nacional 32, 2–19.
Strathern, M., 1979. The self in self-decoration. Oceania 4, 241–257.
Strathern, M., 1988. The Gender of the Gift: Problems with Women and Problems with Society in Melanesia. University of California Press, Berkeley, CA.
Strathern, M., 1992. After Nature: English Kinship in the Late Twentieth Century. Cambridge University Press, Cambridge, UK.
Strathern, M., 2005. Kinship, Law and the Unexpected: Relatives Are Always a Surprise. Cambridge University Press, Cambridge, UK.
Teixeira-Pinto, M., 1997. Ieipari. ANPOCS/HUCITEC, São Paulo, Brazil.
Vilaça, A., 2002. Making kin out of others in Amazonia. Journal of the Royal Anthropological Institute 8, 347–365.
Viveiros de Castro, E., 1992. From the Enemy's Point of View: Humanity and Divinity in an Amazonian Society. University of Chicago, Chicago.
Viveiros de Castro, E., 2001. GUT feelings about Amazonia: potential affinity and the construction of sociality. In: Rival, L., Whitehead, N. (Eds.), Beyond the Visible and the Material: The Amerindianization of Society in the Work of Peter Rivière. Oxford University Press, Oxford, pp. 19–43.

Kinship, Evolution of

Drew Gerkey, University of Maryland, Annapolis, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article introduces core concepts and theories in the evolution of kinship and reviews empirical research that applies these perspectives to understand individual physiology, behavior, and social structure. After presenting theoretical definitions and methods used to measure genetic relatedness, the article reviews the mechanisms individuals use to recognize one another as kin. Research on inclusive fitness and kin selection demonstrates how relatedness between individuals affects individual actions and social interactions. This work establishes connections between the evolution of kinship and research on multilevel selection and major transitions in the evolution of complexity. A brief review suggests how the unique features of human kinship can illustrate the interactions between evolution, culture, and multiple pathways of inheritance.

Introduction

Kinship is a fundamental component of evolutionary theory. In the natural sciences, the term is used to define relationships between individuals who are genetically related (*see* Kin Selection). Two individuals who are kin possess similar genetic material due to a shared common ancestor. Therefore, kinship emerges from one of the evolution's fundamental conditions: the process of transmitting and inheriting genetic material from one generation to the next. Because modes of reproduction can vary from one species to the next, precise measures of genetic relatedness between individuals must derive from the particular processes of genetic replication and transmission for each species. Yet, because all species reproduce, species-specific measures of genetic relatedness, their effects on the physiology and behavior of individuals, and the social structures arising from interactions among kin can be compared across species to investigate general evolutionary patterns and principles. In the broadest sense, all living organisms are kin because all share a common ancestor, although the precise connections among contemporary and extinct species are often far in the past and only partially known. Thus, kinship can be understood as a continuum, extending from investigations of an organism's physiology and behavior in the present and near past to the deepest evolutionary histories of all life forms. The fact that core principles of kinship can be extended across the entire range of this continuum illustrates the importance of kinship in evolutionary theory.

This article introduces core concepts and theories in the evolution of kinship and briefly reviews the diverse empirical research that applies these perspectives to understand individual physiology, behavior, and social structure. The first section focuses on the essence of kinship relatedness, explaining how this term is calculated and the variations that arise from different modes of reproduction across species. The second section examines how relatedness between individuals affects social interactions by reviewing research on inclusive fitness and kin selection. The third section describes mechanisms of kin recognition, which are fundamental to linking calculations of relatedness to individual actions and social interactions. The fourth section presents research on common social structures that emerge from kin selection, including reproduction, parental investment, cooperative breeding, and group formation. The fifth section shifts toward the broader end of the kinship continuum to establish connections between the evolution of kinship and deeper evolutionary histories, including research on multilevel selection, genomic imprinting, and major transitions in the evolution of complexity. Finally, the article concludes with human kinship, suggesting how the various ways our species defines relatedness can illustrate the interactions between evolution, culture, and multiple pathways of inheritance.

Relatedness

Relatedness between two individuals is the essence of kinship. Evolutionary scientists define relatedness in genetic currencies, because the process of genetic replication and transmission across generations is a fundamental component of evolution. Although Darwin understood that the act of reproduction transmitted traits from one generation to the next, he and his contemporaries famously lacked accurate, detailed knowledge of these processes (*see* Evolution, History of). Gregor Mendel's research on genetic inheritance and hybridization was published several years after Darwin's *On the Origin of Species*, but its importance was not understood until research on genetics began to coalesce in the early years of the twentieth century. Even then, it was not until several decades later – through the research of Ronald A. Fisher, Sewall Wright, J.B.S Haldane, and others – that the theory of evolution by natural selection was integrated with population genetics to establish the 'modern synthesis' between heredity and evolution. A key development leading up to the modern synthesis was Wright's coefficient of relationship, which defined the relatedness between two individuals as the sum of correlation coefficients calculated for every genealogical path connecting them to a common ancestor. In most cases, this correlation coefficient equals the probability that two individuals share a gene via descent from a common ancestor, above the random expectation that individuals share a gene due to membership in a population or species. This probability is now most often known as the coefficient of relatedness, or simply r.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.81041-7

The coefficient of relatedness can range in value from one – between identical twins, clones, or one's relatedness to self – to values negligibly greater than zero – between individuals whose closest common ancestor lies many generations in the past. How one calculates r depends on the species' mode of reproduction. For example, in a species that reproduces asexually by budding offspring from a parent, the coefficient of relatedness between parents and offspring – and among siblings – will equal one. In this case, because all of the offspring's genes are received from the parent, there is 100% probability that a gene found in one will appear in the other (setting aside variation introduced by mutation). However, in a diploid species that reproduces sexually – like our own – each offspring inherits half of his or her genes from each parent. In this case, a simple way to calculate the coefficient of relatedness is to (1) define the genealogical pathways that connect two individuals, (2) value each link in the path according to the probability of inheriting a gene, (3) multiply the probabilities for each link on a path connecting two individuals, and (4) sum the product of probabilities for each path. For example, two full siblings are connected by a genealogical path via the father ($0.5 \times 0.5 = 0.25$) plus a path via the mother ($0.5 \times 0.5 = 0.25$). When we add the product of probabilities along each path ($0.25 + 0.25$), we obtain a coefficient of relatedness equal to 0.5. This method for calculating coefficients of relatedness works for all sexually reproducing diploid species, but must be altered to produce correct results for species with different modes of reproduction.

In practice, there are multiple ways one can measure coefficients of relatedness. First, one can collect genealogical information for all individuals in a population, and then use this information to apply the simple method described above, known as the 'pedigree method' (Queller, 2010). This method is accurate as long as one obtains genealogical information for each individual with sufficient generational depth. Unfortunately, in many cases it is difficult to obtain genealogies that go beyond a few generations, limiting the practical usefulness of the pedigree method in some contexts. Fortunately, advances in genetic sequencing have enabled additional techniques based on variable neutral genetic markers (Weir et al., 2006). For example, short tandem repeats of DNA (microsatellites) or single-nucleotide polymorphisms (SNPs) can be compared between two individuals to ascertain their coefficient of relatedness with high accuracy. Although this method can be useful when genealogical information is absent or difficult to obtain, specialized equipment and skills for genetic sequencing are necessary, posing difficulties for researchers who lack the ability to implement these techniques.

Kin Selection and Inclusive Fitness

Measuring genetic relatedness between individuals is important because this information can be used to understand the evolution of social interactions. An individual's actions not only affect his or her own reproductive success, they may also affect the reproductive success of others who interact with that individual. William D. Hamilton developed the concept of inclusive fitness to capture the significance of genetic relatedness for the evolution of these social interactions (Hamilton, 1964) (*see* Kin Selection; Sociobiology: Overview). Whereas an individual's direct fitness is determined by his or her success in producing offspring, an individual's inclusive fitness combines an individual's reproductive success (direct fitness) with the effects that individual's actions have on the reproductive success of related individuals (indirect fitness). The core insight of inclusive fitness is that a trait that appears to decrease an individual's direct fitness may actually increase that individual's inclusive fitness, resolving apparent contradictions between the evolutionary process of natural selection and the persistence of physical traits, behaviors, and social structures that impose net costs on an individual's direct fitness.

Hamilton applied the concept of inclusive fitness to predict the evolution of such costly characteristics with a simple inequality: $c < b \times r$. This inequality states that a physical trait, behavior, or social structure that imposes net costs on an individual's direct fitness (c) can evolve via natural selection if those costs are less than the benefits of that characteristic for other individuals (b), multiplied by the coefficient of relatedness between them (r). Because the coefficient of relatedness varies between 0 and 1, it effectively devalues the indirect benefits of the costly characteristic as genetic relatedness declines between two individuals. For example, a costly characteristic that reduces an individual's direct fitness by 1 unit but increases the reproductive success of a full sibling ($r = 0.5$) by 3 units can evolve via natural selection because the indirect benefit for the individual ($3 \times 0.5 = 1.5$) exceeds the direct cost of the physical trait, behavior, or social structure. However, that same costly characteristic would be unlikely to evolve as a result of interactions between half siblings ($r = 0.25$) because the indirect benefits ($3 \times 0.25 = 0.75$) do not exceed the direct costs ($c = 1$) for the individual. Although a complete explanation for the evolution of a specific costly trait in a particular species at a given moment in time would involve more complex models from population genetics and evolutionary game theory, the simple inequality above – often known as Hamilton's rule – provides a useful starting point for investigating the relationship between evolution and kinship that can be applied across a wide range of species and contexts.

Understandably, researchers applying Hamilton's rule tend to focus on variation in coefficients of relatedness. However, it is important not to lose sight of the fact that other elements of the inequality – costs for direct fitness and benefits for indirect fitness – can play an equally important role. Although coefficients of relatedness are often equal for a variety of social relationships, factors that determine direct and indirect fitness can vary significantly across these relationships, particularly in species with multistage life histories. For example, in a diploid, sexually reproducing species like ours, coefficients of relatedness between parents and offspring ($r = 0.5$) equal coefficients of relatedness between full siblings ($r = 0.5$). Yet full siblings are almost always closer in age and the periods each depends on his or her parents for support often overlap, creating the potential for conflicts of interest that are more intense than those between parents and offspring. In this example, the greater age differences between parents and offspring can effectively reduce the direct costs of a particular trait, behavior, or social structure for the parent and increase the indirect benefit via aid to the offspring. Direct costs may be lower for parents because older

individuals often possess greater resources, benefit only marginally from small increases in resources relative to offspring, and hold fewer remaining opportunities to reproduce. In contrast, offspring often possess limited resources, benefit greatly from small increases in resources due to key life history trade-offs at early stages of development, and hold many opportunities to reproduce once reaching adulthood, all of which can significantly increase the indirect benefits of a costly trait held by parents. These asymmetries in age, resources, sensitivity to trade-offs, and reproductive potential are less pronounced among full siblings, meaning that a similar costly trait held by one sibling may be less likely to evolve due to the indirect benefits of providing aid to another sibling who may be competing for parental resources. Variations in coefficients of relatedness will always be important for investigating kinship and the evolution of social interactions, but the example above illustrates why it is important for researchers to also explore factors that affect the other terms of Hamilton's inequality – the costs and benefits of direct and indirect fitness for physical traits, behaviors, and social structures.

Kin Recognition

With sufficient genealogical or genetic data, it is not too difficult for researchers to calculate coefficients of relatedness and investigate how kinship affects social interactions. For the kin who are the focus of this research, however, the information necessary for calculating coefficients of relatedness is often quite limited, raising an important question: How do individuals recognize one another as kin? Because coefficients of relatedness are unobservable, species have developed various kin recognition mechanisms that exploit contextual cues, physical traits, social structures, and other circumstances that correlate reliably with genetic relatedness. On a general level, a kin recognition mechanism includes four essential components (Penn and Frommen, 2010). First, a kin recognition mechanism requires a phenotypic or contextual cue that is sufficiently variable among individuals, is correlated with genetic differences, and either remains stable across an individual's life span or is at least present during a critical phase when the recognition mechanism operates. Second, a mechanism requires a process of recognition, where an individual perceives a cue and compares it to some template that defines how that cue corresponds to a degree of relatedness. Third, the results of the recognition process are used to apply decision rules that specify how an individual will act according to relatedness. Fourth, these decision rules are assessed and used as the basis of actions that are biased in some way by the relatedness between the two individuals. These four components can take many different forms, but kin recognition mechanisms can be classified based on the kinds of cues they utilize: (1) contextual cues based on spatial proximity and familiarity from past interactions and (2) phenotypic matching via visual, auditory, and olfactory cues.

Contextual Mechanisms

Although coefficients of relatedness are not perceptually salient, individuals can often exploit consistent social structures and environmental patterns to recognize kin. These contextual mechanisms of kin recognition can include spatial proximity due to nests or other persistent residential groups as well as familiarity developed through repeated interactions. A mother who gives birth to offspring can be confident she is related to those offspring, and to the extent those offspring are raised together and interact repeatedly, they can assume they are related to one another. Fathers, however, are often in a more precarious position to make such assumptions, depending on the mating strategies prevalent in the social group (Widdig, 2007). Even in species with a high prevalence of monogamy, extra-pair mating can occur, reducing the effectiveness of spatial proximity as a mechanism of kin recognition for fathers. This variation in the effectiveness of contextual mechanisms opens possibilities for diverging strategies among individuals within a species. One parent may benefit from circumstances that cause another to overestimate relatedness, while the other would benefit from a more accurate assessment, creating dynamic interactions, as in the case of cuckoldry and other forms of sexual conflict involving paternity certainty and kin recognition (Arnqvist and Rowe, 2005).

Interestingly, contextual mechanisms of kin recognition often provide opportunities for conflict and exploitation between members of different species, as in the case of brood parasitism. Cuckoo species (family Cuculidae) provide an iconic example, laying their own eggs in the nests of other birds and leaving them to be raised by the unsuspecting parents of the other species. Brood parasitism is often successful due to the victim's reliance on contextual mechanisms of kin recognition such as the presence of eggs in one's own nest. Many species have evolved counter-strategies to discriminate their offspring from parasite offspring, in turn resulting in more elaborate attempts at deception by the species practicing brood parasitism (Feeney et al., 2012). Research on incest avoidance and familiarity provides another iconic example of contextual mechanisms of kin recognition, this time with a confluence rather than a conflict of interests among individuals. Building on the ideas of Edvard Westermarck, scholars have examined contexts where individuals raised together display reduced levels of attraction and mating success compared to those raised apart, even when individuals raised together are genetically unrelated. Coresidence is hypothesized as a mechanism allowing individuals to recognize closely related kin (e.g., siblings) and avoid the deleterious effects of mating with them.

Phenotypic Matching

Just as members of one species resemble each other more than members of another species, it is possible that more closely related individuals will resemble one another to a greater extent than distantly related individuals. If differences in genetic relatedness are correlated with salient phenotypic characteristics, individuals may be able to recognize kin by assessing phenotypes of other individuals. Penn and Frommen distinguish between three kinds of phenotypic matching: familiarity dependent, familiarity independent, and self-referential (2010: pp. 12–15). Familiarity dependent matching occurs when an individual assesses the phenotype of another individual at one time, and then compares this assessment with perceptions of

that same individual's phenotype at a later time. Put simply, individuals learn what phenotypic traits are associated with kin, and then use these traits to identify those kin as they encounter them again across the life span. Thus, familiarity-dependent mechanisms of kin recognition rely on contextual mechanisms discussed earlier – such as spatial proximity – to identify phenotypic traits of kin. These phenotypic traits are then used to identify kin in absence of the original contextual mechanisms, as when siblings disperse from a natal nest.

Phenotypic matching can also be extended to unfamiliar individuals. By comparing the phenotypic traits of unfamiliar individuals with phenotypic traits of familiar kin, an individual can extend contextual mechanisms beyond their spatial and temporal restrictions. For example, a contextual mechanism for recognizing siblings that relies on the spatial proximity of a nest and temporal proximity of a litter could lead siblings to become familiar with one another, and then recognize one another by sight, sound, or smell after dispersing from the nest (familiarity-dependent matching). If these visual, auditory, or olfactory cues are somewhat consistent across litters reared by the same parents, then these cues could also be used by individuals to recognize older or younger siblings from separate litters. Because these litters did not share the same nest, neither contextual nor familiarity-dependent mechanisms would allow them to recognize one another as siblings. Failure to recognize siblings from different litters could have significant negative consequences for inclusive fitness, as in the case of inbreeding depression. Individuals who are able to develop familiarity-dependent phenotypic cues of relatedness and apply these cues to unfamiliar individuals (familiarity-independent matching) can avoid mating with close relatives and escape the deleterious effects of inbreeding.

While potentially effective in a variety of circumstances, these two kinds of phenotypic matching depend on contextual mechanisms to develop templates that associate particular traits with kin. A third kind of phenotypic matching does not depend on any contextual or familiarity mechanisms; it simply depends on an individual's ability to compare its own phenotype to the phenotypes of unfamiliar individuals. Such self-referential phenotypic matching mechanisms are limited to those visual, auditory, and olfactory traits that an individual can perceive in itself and identify in others. Self-referential mechanisms may be particularly effective when context and familiarity do not allow individuals to accurately assess relatedness, such as when individuals are reared alone, in litters where genetic relatedness varies due to paternity uncertainty, or in litters that include unrelated individuals from the same or another species (Hauber and Sherman, 2001).

Although these recognition mechanisms may not always capture precise coefficients of relatedness, it is clear that individuals in many species are successful at using kin recognition mechanisms to guide social interactions and act in ways that increase their inclusive fitness. Evidence of effective kin recognition can be found in a variety of taxa, including humans (Kaminski et al., 2009; Park et al., 2008), nonhuman primates (Silk, 2002; Rendall, 2004), mammals (Griffin and West, 2003; Mateo, 2003), birds (Nakagawa and Waas, 2004; Zelano and Edwards, 2002), fish (Ward and Hart, 2003; Griffiths, 2003), invertebrates (Gherardi et al., 2012), and microbes (Strassmann et al., 2011). In some species, multiple mechanisms may be used in tandem. In other species, mechanisms may operate at specific development stages instead of across the life span. Understanding these processes of kin recognition and investigating their evolution is an important line of inquiry for further research.

Kinship and Social Structure

To the extent individuals are able to recognize kin, natural selection should favor physiological traits, behavioral strategies, and social structures that utilize this information to maximize inclusive fitness within the particular constraints of the individual's environment. Hamilton's research is considered a fundamental breakthrough in our understanding of the evolution of sociality because it provides a clear concept (inclusive fitness) and a simple rule ($c < b \times r$) that allows us to investigate adaptations that emerge from interactions among individuals with varying degrees of genetic relatedness (Hamilton, 1997). A comprehensive review of the literature building on Hamilton's work is beyond the scope of this article (Gardner et al., 2011; Queller, 2011). Instead, it will suffice to highlight some broad domains of social structure that are shaped by the evolution of kinship, starting with the foundations of mating and parental care and extending to more elaborate social structures involving cooperation, group defense, and fission–fusion dynamics.

Considering the fundamental role of the transmission and inheritance of genetic information in evolution, it should not be surprising that kinship is important for our understanding of reproduction. Genetic relatedness between individuals shapes reproductive strategies at all stages, from choosing potential mates to producing and caring for offspring (Pizzari and Gardner, 2012). In sexually reproducing species, mating with close kin (incest) can have significant, negative effects on offspring survival and reproductive success (inbreeding depression), limiting the inclusive fitness of close kin who mate with one another. There is evidence that mechanisms of kin recognition help related individuals avoid incest in a variety of species (Pusey and Wolf, 1996), including birds (Nakagawa and Waas, 2004), humans (DeBruine, 2005; Rantala and Marcinkowska, 2011), mammals (Mateo, 2002), and insects (Tabadkani et al., 2012). One of the most extensively investigated examples concerns the role of major histocompatibility complex (MHC) genes in regulating mating preferences (Bernatchez and Landry, 2003; Penn and Potts, 1999; Tregenza and Wedell, 2000). MHC molecules vary among individual members of a species, and this diversity plays an important role in immune function, helping to distinguish an individual's own cells from pathogens and other foreign cells. Genes that code for MHC molecules are transmitted from parents to offspring, and the molecules can be perceived by smell, providing an opportunity for kin recognition via phenotypic matching mechanisms. In addition to the value of avoiding incest, the benefits of MHC diversity for immune function can lead individuals to seek mating partners with dissimilar MHC molecules. There is evidence that MHC genes play a role in partner choice in humans (Havlicek and Roberts, 2009; Wedekind, 2007), primates (Knapp et al., 2006; Setchell and

Huchard, 2010), birds (Zelano and Edwards, 2002), fish (Eizaguirre and Lenz, 2010; Reusch et al., 2001), and reptiles (Miller et al., 2009; Olsson et al., 2005).

Parents of all species invest energy, time, and material resources to produce offspring, and the patterns of parental investment that emerge can be explained by measuring coefficients of relatedness, investigating mechanisms of kin recognition, and documenting physiological and behavioral traits that are biased by these mechanisms. Genetic relatedness plays a key role in evolutionary theories of parental investment, helping researchers navigate the potential conflicts and confluences of interest among individuals with varying degrees of kinship. The interactions of strategies among kin in the context of reproduction and parental care are complex even when focusing on parents and offspring. These complexities grow significantly when more distantly related kin become involved. Research on cooperative breeding and alloparental care illustrates the importance of kinship for understanding the evolution of social structures involved in reproduction and parental investment. Cooperative breeding occurs when individuals sacrifice some of their own reproductive effort in order to contribute to the reproductive effort of other individuals as alloparents (Komdeur et al., 2008). Alloparents are often – but not always – genetically related to the parents they assist, and evidence suggests that alloparents can recover the direct fitness costs of these strategies through a combination of indirect fitness benefits and enhanced mating opportunities, diminished predation risk, and other benefits of group membership (Bergmüller et al., 2007; Clutton-Brock, 2002; Lukas and Clutton-Brock, 2012).

While the indirect fitness benefits of cooperative breeding strategies are an intriguing example of the influence of kinship on social structure, other benefits of group membership found in a wider range of species are also shaped by genetic relatedness. Some of the earliest investigations of kinship and social behavior focused on the production of alarm calls in social birds and rodents. Individuals are vulnerable to predation while foraging outside their burrows or nests, and alarm calls are used to alert group members to the presence of a predator. Because the alarm also attracts the attention of the predator, an individual who sounds the alarm may be exposed to a higher risk of predation than an individual who spots a predator and silently seeks safety. Evidence from a variety of studies suggests that associations between genetic relatedness and alarm calls – mediated by factors such as age, sex, dominance rank, and migration rates – underlie the evolution of this antipredation strategy (Blumstein and Armitage, 1997; Hollén and Radford, 2009), although other factors that affect risks to alarm callers can also be important (Shelley and Blumstein, 2004).

Evidence suggests that kinship can also drive the formation and dissolution of social groups. Fission–fusion dynamics (i.e., the dynamic process of merging and splitting between and among subgroups within a population) reflect demographic growth and the availability of resources, and kinship often – but not always – plays an important role in identifying which neighboring groups will combine or which subgroups will cleave apart (Aureli et al., 2008). African savannah elephants live in social groups that vary frequently in spatial location and composition, and genetic relatedness among older females predicts patterns of fission and fusion (Archie et al., 2006). Similar results arise from research on fission–fusion dynamics in giraffes (Carter et al., 2013), bats (Popa-Lisseanu et al., 2008), rhesus macaques (Widdig et al., 2006), and dolphins (Frère et al., 2010). Even when the boundaries and composition of groups remain more or less stable over time, individuals often disperse from one group to the next according to factors such as age, sex, and dominance rank (Ronce, 2007). These patterns of dispersal have consequences for the relationship between kinship and social structure, particularly the balance between cooperation and competition among kin and non-kin (Lawson Handley and Perrin, 2007; Hatchwell, 2010; Lehmann and Rousset, 2010).

Although kinship shapes processes and patterns of reproduction, parental investment, cooperative breeding, defense, and group formation in many species, there are some species that integrate these domains of social life to a much greater extent, embedding individual actions within elaborate social structures shaped by the evolution of kinship. Eusocial species – defined as those that exhibit a reproductive division of labor, cooperative breeding, and overlapping generations within the social group (Wilson, 1971) – are perhaps the most striking examples of kinship's influence on the evolution of physiology, behavior, and social structure. Many eusocial species combine sterile worker castes with reproductive 'queens' that monopolize mating opportunities and benefit from extensive alloparental care, cooperative resource production, and territorial defense provided by other members of the social group. In these cases, some individuals sacrifice their own opportunities to reproduce and devote their energy to increasing the reproduction of a single related female, deriving inclusive fitness benefits entirely from indirect reproduction. Indeed, Hamilton's original formulation of inclusive fitness and kin selection theory was intended in part to explain the evolution of eusociality in a number of insect species (ants, bees, wasps), which Darwin singled out as a troublesome, potential contradiction to his theory of evolution by natural selection. Since Hamilton, scholars have continued to refine definitions of eusociality (Crespi, 2005; Lacey and Sherman, 2005), nominate additions to the list of eusocial species – including mole rats (Burda et al., 2000), beetles (Biedermann and Taborsky, 2011), shrimp (Duffy and Macdonald, 2010), and humans (Wilson, 2012) – and debate the role of kin selection, group selection, and other evolutionary mechanisms underlying the evolution of eusociality (Gardner and Ross, 2013; Hughes et al., 2008; Wilson and Hölldobler, 2005).

Kinship, Multilevel Selection, and Major Transitions

Hamilton's research on inclusive fitness and kin selection was among the first to demonstrate the advantages of exploring conflicts and confluences of interest across varying levels of biological and social hierarchy. Whereas evolutionary theories of cooperation and conflict had primarily focused on calculating costs and benefits at the level of individual organisms and groups of organisms, Hamilton's approach utilized what has come to be called 'the gene's-eye' perspective (Dawkins, 1976). This approach entails calculating costs and benefits of a trait from the perspective of a gene that codes for the

trait rather than the interests of the individual who carries the gene, opening up the potential for cooperation and conflict to occur at multiple levels simultaneously, including within individuals. This approach revolutionized the study of social interactions because physiology, behaviors, or social structures that entail an apparent cost for an individual can nonetheless evolve due to benefits accruing to the putative gene underlying the trait. Prior to the work of Hamilton, George C. Williams, Robert Trivers, and others, it was common to explain traits that were costly for an individual in terms of their benefits to the individual's social group or species (Williams, 1966; Trivers, 2002). Although the surge in research adopting the gene's-eye perspective tipped scholarly consensus on the evolution of cooperation and competition toward explanations that emphasized the interests of genes and individuals at the expense of groups, these debates have been rekindled by a number of researchers who emphasize the importance of natural selection operating at the group level (Sober and Wilson, 1998; Nowak et al., 2010), provoking in turn a strong response from proponents of Hamilton's approach (Abbot et al., 2011; Bourke, 2011b).

One important, positive outcome of this debate has been a renewed interest in multilevel selection (Okasha, 2006) and theories that articulate how evolutionary dynamics operate simultaneously across multiple levels of biological and social hierarchy. Although scholars continue to disagree about the relative importance of natural selection operating at the levels of genes, individuals, and groups in empirical contexts, there is widespread agreement that evolution can, in principle, operate at multiple levels. This consensus emerges in large part from the work of John Maynard Smith and Eörs Szathmáry on 'major transitions' in evolution (Maynard Smith and Szathmáry, 1997). A major transition occurs when previously independent units (genes, cells, individuals) combine to form groups, lose the ability to reproduce independently of those groups, and become interdependent units whose evolutionary fate depends in large part on their ability to function cohesively. Although such transitions have clearly occurred in the past – including transitions from prokaryotes to eukaryotes, protists to multicellular organisms, asexual reproduction to sexual reproduction, and solitary to eusocial species – evolutionary scholars still struggle to understand the specific conditions, dynamics, and thresholds that bring about major transitions (Calcott and Sterelny, 2011). The conceptual and methodological tools developed to understand the evolution of kinship – including coefficients of relatedness, kin recognition mechanisms, and inclusive fitness – have a key role to play in this effort because they help scholars link conflicts and confluences of interest at different levels of hierarchy to modes of reproduction and forces of selection operating across these levels, from genes to individuals to groups (Birch, 2012; Bourke, 2011a).

Whereas research on multilevel selection and major transitions extends inquiry outward, beyond the tangible lines dividing individuals to investigate less perceptible boundaries of social groups, others continue to pursue the implications of the gene's eye perspective for understanding equally imperceptible dynamics within individuals. Research on genomic imprinting and kinship is a particularly intriguing example. Genomic imprinting refers to instances when a gene is expressed (or not) depending on whether it is inherited from a mother or a father (Pomiankowski, 2002). David Haig proposes an explanation for patterns in genomic imprinting that derives inspiration from Hamilton's work, suggesting potential conflicts of interest between maternal and paternal genes can be used to predict the presence of imprinted genes and their effects on physiology and behavior (Haig, 2004). For example, genes regulating parental investment might be expected to demand greater resources from the mother when inherited from the father than when inherited from the mother, due to the difference in costs and benefits of maternal investment from the perspective of each parent. The best documented example concerns a gene encoding for insulin-like growth factor 2 (IGF2) in mammals, which is only expressed when inherited from the father and compels the offspring to demand greater resources from the mother in-utero (Reik et al., 2003). IGF2 and other examples of genomic imprinting illustrate ongoing cooperation and conflict at the genetic level, alluding to deeper evolutionary histories of conflicting interests that were mitigated, but not entirely resolved, by major transitions (Burt and Trivers, 2008).

Kinship and Culture

This review of research on the evolution of kinship has focused on describing key concepts, methods, and theories at a general level in order to illustrate the potential for complementary research across species. However, in order to support consilience across the divide between the natural and social sciences (Wilson, 1998), it is important to consider insights from the long tradition of research on human kinship, which is often unfamiliar to evolutionary scholars in the natural sciences. Although biological dimensions of human kinship have long been acknowledged and investigated by social scientists, cross-cultural variations in the way people define and recognize kin – as well as implications for social structures – have been the primary focus of research (Parkin and Stone, 2004). A complete review of the cultural and biological dimensions of human kinship is beyond the scope of this article (see Cronk and Gerkey, 2007; Kurland and Gaulin, 2005). Here it will suffice to briefly reframe research on human kinship in relation to the conceptual structure for the evolution of kinship offered above, moving from kinship terminologies to some common social structures that emerge from the interaction of biological and cultural definitions of relatedness. In this task, brevity is accomplished by focusing on insights from the growing body of research by social scientists (particularly anthropologists) who integrate evolutionary and cultural perspectives (Shenk and Mattison, 2011).

While evolutionary scholars define relatedness in genetic currencies, social scientists define relatedness by following the cultural categories utilized by the people with whom they work. Cultural definitions of relatedness vary dramatically from one place to the next, evident in different kin terminologies that lump and split social relationships into categories that often depart from those defined by biological coefficients of relatedness. Although this variation is impressive, it is important to acknowledge that the observed variation in kin

terms does not exhaust the full range of logical possibilities for lumping and splitting kin (Kemp and Regier, 2012). In other words, cross-cultural variations in kin terminologies exhibit patterns, and these patterns may provide clues to understanding why people's cultural definitions of relatedness differ from relatedness defined in genetic currencies. Kin terminologies often split or lump kin according to a limited number of factors, including gender (male or female), laterality (paternal or maternal), generation (older or younger), and genealogical distance (close or distant) (Jones, 2010; Jordan, 2011).

Although kin terminologies establish cultural calculations of relatedness that certainly differ from biological calculations of relatedness, this disconnect is often quickly – and perhaps mistakenly – assumed to mean that concepts and hypotheses derived from evolutionary theory are irrelevant for understanding human kinship. However, it is interesting to note that the same factors people use to split or lump categories of kin often direct our attention to consistent variations among individuals and social structures that may have consequences for inclusive fitness. In other words, kin terminologies may convey information about the fitness consequences of interacting with kin – culturally defined – by combining coefficients of genetic relatedness with factors crucial to calculating the costs and benefits of interactions with different individuals. Recall that evolutionary scientists studying inclusive fitness and kin selection in other species have warned against focusing exclusively on variation in coefficients of relatedness and ignoring factors that affect costs to direct fitness and benefits to inclusive fitness, which are both important parts of Hamilton's inequality. Similarly, it may be unwise for scholars studying human kinship to use disconnects between cultural and biological definitions of relatedness as a reason to ignore evolutionary theory's relevance. Instead, the wide variety and consistent patterns in human kin terminologies may provide a guide for developing and testing hypotheses about the relationship between cultural and biological definitions of relatedness and exploring the importance of inclusive fitness in human social structures.

Evolutionary anthropologists have begun to investigate the intersections of cultural and biological definitions of relatedness in a variety of social domains where human kinship is important. Some have used inclusive fitness and kin selection to investigate the evolution of marriage practices, including monogamy (Fortunato and Archetti, 2010), polygyny (Borgerhoff Mulder et al., 2006), and polyandry (Starkweather and Hames, 2012) (*see* Marriage Systems, Evolution of). Others have synthesized research on nonhuman primate social dynamics with anthropological research on kinship to understand the role of exogamy – marrying outside one's kin group – in extending cooperation to and restricting competition among wider social circles (Chapais, 2008). Similar approaches have been used to understand human mating, marriage, and child care as cooperative breeding strategies that have parallels in nonhuman primates and other mammals (Hrdy, 2011).

Unilineal descent systems – where membership in a group is traced along paternal or maternal lineages – provide another common pattern in human kinship that can be understood by applying concepts and tools from evolutionary theory. Researchers have found that higher levels of extra-pair mating and paternity uncertainty are associated with matrilineal descent systems, perhaps reflecting the fact that calculations of genetic relatedness are more accurate across generations when traced via maternal lines due to diminished paternity certainty among men (Hartung, 1985). Similarly, others have found that the emergence of matrilineal or patrilineal descent in a particular place may be tied to the differential influence of inherited property – whether land, livestock, or status – on reproductive success of daughters and sons. When inheritance favors the reproductive success of sons, patrilineal descent is common, and when inheritance favors daughters, matrilineal descent is common (Fortunato, 2012; Mattison, 2011). A shift in the kind of property inherited can even lead to a shift in descent, as when African Bantu-speaking matrilineal horticulturalists became patrilineal pastoralists (Holden and Mace, 2003). Such instances of change in social structures tied to kinship often reveal underlying dynamics that can be understood using evolutionary theory (Leonetti and Nath, 2009).

While pursuing evolutionary research on human kinship, it is important to distinguish culture from behavior. Whereas behaviors are actions that can be directly observed and documented, culture is best defined as socially learned and transmitted information and ideas (Durham, 1992). For example, kin terminologies are socially learned and transmitted ideas about how to recognize kin that are distinct from behavioral interactions among kin, which may of course be informed by kin terminologies but may also be influenced by evolved predispositions. This distinction between culture and behavior is not intended to hold culture and evolution as opposed analytically. Rather, it is intended to define behavior as the domain for explanation, and culture and evolution as two potential pathways through which the information that informs behavior can be transmitted from one generation to the next (Cronk, 1999). For example, one of the primary functions of kin terminologies throughout the world is to define who is and is not allowed to marry and produce children. These prohibitions against incest may operate in tandem with evolved kin recognition mechanisms that help individuals avoid the biological costs of inbreeding. However, these prohibitions often depart from those defined by biological costs, applying asymmetrically – to kin on either the paternal or maternal side of the family, when both are equally related genetically – or extending to distantly related kin who pose little danger for costs of inbreeding. Rather than functioning simply to avoid inbreeding, kin terminologies often structure marital relationships that maintain or transform the boundaries between social groups. The practices of marrying within one's kin group – endogamy – or outside one's kin group – exogamy – often function to negotiate social structures, ascribe rights and obligations within groups, and form alliances between groups. Indeed, cultural marriage practices shaped by these aspects of social structure may even come into conflict with evolved kin recognition mechanisms aimed at avoiding inbreeding. A classic example concerns a Taiwanese 'sim-pua' marriage practice that betrothed children at an early age and transferred future brides to reside in the households of their future husbands. Despite the cultural prescriptions of these practices, researchers have found that partners in these marriages experienced less marital

harmony and produced fewer children than partners who were raised in separate households (Lieberman, 2009; Kushnick and Fessler, 2011; Wolf and Durham, 2004). In this case, evolved kin recognition mechanisms constrain the variation of cultural kinship practices, evident when we document associations between particular behaviors (incidences of marital strife, offspring produced), cultural practices (sim-pua marriages), and conditions that trigger evolved propensities (coresidence during early childhood).

This approach to studying human kinship can be used to explore interactions between culture and evolution as part of the growing research on gene-culture coevolution (Richerson and Boyd, 2004) (*see* Human Cooperation, Evolution of). Although many aspects of human kinship are unlikely to have close parallels in other species – such as kin terminology and unilineal descent – these exceptions can be valuable for understanding fundamental social dynamics that apply more broadly. For example, eusociality and cooperative breeding are found in relatively few species, but their unique features allow us to test predictions about inclusive fitness and multilevel selection that should, in principle, explain the conditions that lead to the presence or absence of both novel and common social structures. Although many important questions along these lines remain unanswered, this review of the evolution of kinship has highlighted some key concepts, theories, and empirical studies that provide a solid foundation for future work.

Acknowledgment

DG was supported by the University of Maryland and NSF Award #DBI-1052875 to the National Socio-Environmental Synthesis Center.

See also: Human Cooperation, Evolution of; Incest Prohibition, Origin and Evolution of; Kin Selection; Marriage Systems, Evolution of; Sociobiology: Overview.

Bibliography

Abbot, P., et al., 2011. Inclusive fitness theory and eusociality. Nature 471 (7339), E1–E4 author reply E9–E10.
Archie, E.A., Moss, C.J., Alberts, S.C., 2006. The ties that bind: genetic relatedness predicts the fission and fusion of social groups in wild African elephants. Proceedings of the Royal Society B: Biological Sciences 273 (1586), 513–522.
Arnqvist, G., Rowe, L., 2005. Sexual Conflict. Princeton University Press, Princeton.
Aureli, F., et al., 2008. Fission? Fusion dynamics: new research frameworks. Current Anthropology 49 (4), 627–654.
Bergmüller, R., et al., 2007. Integrating cooperative breeding into theoretical concepts of cooperation. Behavioural Processes 76 (2), 61–72.
Bernatchez, L., Landry, C., 2003. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? Journal of Evolutionary Biology 16 (3), 363–377.
Biedermann, P.H.W., Taborsky, M., 2011. Larval helpers and age polyethism in ambrosia beetles. Proceedings of the National Academy of Sciences of the United States of America 108 (41), 17064–17069.
Birch, J., 2012. Collective action in the fraternal transitions. Biology & Philosophy 27 (3), 363–380.
Blumstein, D.T., Armitage, K.B., 1997. Does sociality drive the evolution of communicative complexity? A comparative test with ground-dwelling sciurid alarm calls. The American Naturalist 150 (2), 179–200.
Borgerhoff Mulder, M., Nunn, C.L., Towner, M.C., 2006. Cultural macroevolution and the transmission of traits. Evolutionary Anthropology: Issues, News, and Reviews 15 (2), 52–64.
Bourke, A.F.G., 2011a. Principles of Social Evolution. Oxford University Press, Oxford.
Bourke, A.F.G., 2011b. The validity and value of inclusive fitness theory. Proceedings of the Royal Society B: Biological Sciences 278 (1723), 3313–3320.
Burda, H., et al., 2000. Are naked and common mole-rats eusocial and if so, why? Behavioral Ecology and Sociobiology 47 (5), 293–303.
Burt, A., Trivers, R.L., 2008. Genes in Conflict: the Biology of Selfish Genetic Elements. Harvard University Press, Cambridge.
Calcott, B., Sterelny, K. (Eds.), 2011. The Major Transitions in Evolution Revisited. MIT Press, Cambridge.
Carter, K.D., et al., 2013. Fission–fusion dynamics in wild giraffes may be driven by kinship, spatial overlap and individual social preferences. Animal Behaviour 85 (2), 385–394.
Chapais, B., 2008. Primeval Kinship: How Pair-Bonding Gave Birth to Human Society. Harvard University Press, Cambridge.
Clutton-Brock, T., 2002. Breeding together: kin selection and mutualism in cooperative vertebrates. Science (New York) 296 (5565), 69–72.
Crespi, B.J., 2005. Social sophistry: logos and mythos in the forms of cooperation. Annales Zoologici Fennici 42 (6), 569–571.
Cronk, L., 1999. That Complex Whole: Culture and the Evolution of Human Behavior. Westview Press, Boulder.
Cronk, L., Gerkey, D., 2007. Kinship and descent. In: Dunbar, R.I.M., Barrett, L. (Eds.), Oxford Handbook of Evolutionary Psychology. Oxford University Press, Oxford, pp. 463–478.
Dawkins, R., 1976. The Selfish Gene. Oxford University Press, Oxford.
DeBruine, L.M., 2005. Trustworthy but not lust-worthy: context-specific effects of facial resemblance. Proceedings of the Royal Society B: Biological Sciences 272 (1566), 919–922.
Duffy, J.E., Macdonald, K.S., 2010. Kin structure, ecology and the evolution of social organization in shrimp: a comparative analysis. Proceedings of the Royal Society B: Biological Sciences 277 (1681), 575–584.
Durham, W.H., 1992. Co-evolution: Genes, Culture, and Human Diversity. Stanford University Press, Stanford.
Eizaguirre, C., Lenz, T.L., 2010. Major histocompatibility complex polymorphism: dynamics and consequences of parasite-mediated local adaptation in fishes. Journal of Fish Biology 77 (9), 2023–2047.
Feeney, W.E., Welbergen, J.A., Langmore, N.E., 2012. The frontline of avian brood parasite-host coevolution. Animal Behaviour 84 (1), 3–12.
Fortunato, Laura, 2012. The evolution of matrilineal kinship organization. Proceedings of the Royal Society B: Biological Sciences 279 (1749), 4939–4945.
Fortunato, L., Archetti, M., 2010. Evolution of monogamous marriage by maximization of inclusive fitness. Journal of Evolutionary Biology 23 (1), 149–156.
Frère, C.H., et al., 2010. Home range overlap, matrilineal and biparental kinship drive female associations in bottlenose dolphins. Animal Behaviour 80 (3), 481–486.
Gardner, A., West, S.A., Wild, G., 2011. The genetical theory of kin selection. Journal of Evolutionary Biology 24 (5), 1020–1043.
Gardner, Andy, Ross, L., 2013. Haplodiploidy, Sex-Ratio Adjustment, and Eusociality. The American Naturalist, pp. E60–E67.
Gherardi, F., Aquiloni, L., Tricarico, E., 2012. Revisiting social recognition systems in invertebrates. Animal Cognition 15 (5), 745–762.
Griffin, A.S., West, Stuart A., 2003. Kin discrimination and the benefit of helping in cooperatively breeding vertebrates. Science (New York) 302 (5645), 634–636.
Griffiths, S.W., 2003. Learned recognition of conspecifics by fishes. Fish and Fisheries 4 (3), 256–268.
Haig, D., 2004. Genomic imprinting and kinship: how good is the evidence? Annual Review of Genetics 38, 553–585.
Hamilton, W.D., 1964. The genetical evolution of social behaviour. I. Journal of Theoretical Biology 7 (1), 1–16.
Hamilton, W.D., 1997. Narrow Roads of Gene Land: the Collected Papers of W. D. Hamilton. In: Evolution of Social Behavior, vol. 1. Oxford University Press, Oxford.
Hartung, J., 1985. Matrilineal inheritance – new theory and analysis. Behavioral and Brain Sciences 8 (5), 661–670.
Hatchwell, B.J., 2010. Cryptic kin selection: kin structure in vertebrate populations and opportunities for kin-directed cooperation. Ethology 116 (3), 203–216.
Hauber, M.E., Sherman, P.W., 2001. Self-referent phenotype matching: theoretical considerations and empirical evidence. Trends in Neurosciences 24 (10), 609–616.
Havlicek, J., Roberts, S.C., 2009. MHC-correlated mate choice in humans: a review. Psychoneuroendocrinology 34 (4), 497–512.
Holden, C.J., Mace, R., 2003. Spread of cattle led to the loss of matrilineal descent in Africa: a coevolutionary analysis. Proceedings of the Royal Society B: Biological Sciences 270 (1532), 2425–2433.

Hollén, L.I., Radford, A.N., 2009. The development of alarm call behaviour in mammals and birds. Animal Behaviour 78 (4), 791–800.

Hrdy, S.B., 2011. Mothers and Others: the Evolutionary Origins of Mutual Understanding. Harvard University Press, Cambridge.

Hughes, W.O.H., et al., 2008. Ancestral monogamy shows kin selection is key to the evolution of eusociality. Science (New York) 320 (5880), 1213–1216.

Jones, D., 2010. Human kinship, from conceptual structure to grammar. Behavioral and Brain Sciences 33 (05), 367–381.

Jordan, F.M., 2011. A phylogenetic analysis of the evolution of Austronesian sibling terminologies. Human Biology 83 (2), 297–321.

Kaminski, G., et al., 2009. Human ability to detect kinship in strangers' faces: effects of the degree of relatedness. Proceedings of the Royal Society B: Biological Sciences 276 (1670), 3193–3200.

Kemp, C., Regier, T., 2012. Kinship categories across languages reflect general communicative principles. Science (New York) 336 (6084), 1049–1054.

Knapp, L.A., Robson, J., Waterhouse, J.S., 2006. Olfactory signals and the MHC: a review and a case study in *Lemur catta*. American Journal of Primatology 68 (6), 568–584.

Komdeur, J., et al., 2008. The evolution and ecology of cooperative breeding in vertebrates. In: Encyclopedia of Life Sciences. John Wiley & Sons, Ltd, Chichester, pp. 1–8.

Kurland, J.A., Gaulin, S.J.C., 2005. Cooperation and conflict among kin. In: The Handbook of Evolutionary Psychology. John Wiley & Sons, Hoboken, pp. 447–482.

Kushnick, G., Fessler, D.M.T., 2011. Karo Batak cousin marriage, Cosocialization, and the Westermarck hypothesis. Current Anthropology 52 (3), 443–448.

Lacey, E.A., Sherman, P.W., 2005. Redefining eusociality: concepts, goals and levels of analysis. Annales Zoologici Fennici 42 (6), 573–577.

Lawson Handley, L.J., Perrin, N., 2007. Advances in our understanding of mammalian sex-biased dispersal. Molecular Ecology 16 (8), 1559–1578.

Lehmann, L., Rousset, F., 2010. How life history and demography promote or inhibit the evolution of helping behaviours. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 365 (1553), 2599–2617.

Leonetti, D.L., Nath, D.C., 2009. Age at first reproduction and economic change in the context of differing kinship ecologies. American Journal of Human Biology: The Official Journal of the Human Biology Council 21 (4), 438–447.

Lieberman, D., 2009. Rethinking the Taiwanese minor marriage data: evidence the mind uses multiple kinship cues to regulate inbreeding avoidance. Evolution and Human Behavior 30 (3), 153–160.

Lukas, D., Clutton-Brock, T., 2012. Life histories and the evolution of cooperative breeding in mammals. Proceedings of the Royal Society B: Biological Sciences 279 (1744), 4065–4070.

Mateo, J.M., 2002. Kin-recognition abilities and nepotism as a function of sociality. Proceedings of the Royal Society B: Biological Sciences 269 (1492), 721–727.

Mateo, J.M., 2003. Kin recognition in ground squirrels and other rodents. Journal of Mammalogy 84 (4), 1163–1181.

Mattison, S.M., 2011. Evolutionary contributions to solving the "matrilineal puzzle": a test of Holden, Sear, and Mace's model. Human Nature (Hawthorne, N.Y.) 22 (1–2), 64–88.

Maynard Smith, J., Szathmáry, E., 1997. The Major Transitions in Evolution. Oxford University Press, Oxford.

Miller, H.C., et al., 2009. Influence of major histocompatibility complex genotype on mating success in a free-ranging reptile population. Proceedings of the Royal Society B: Biological Sciences 276 (1662), 1695–1704.

Nakagawa, S., Waas, J.R., 2004. O sibling, where art thou? A review of avian sibling recognition with respect to the mammalian literature. Biological Reviews 79 (1), 101–119.

Nowak, M.A., Tarnita, C.E., Wilson, E.O., 2010. The evolution of eusociality. Nature 466 (7310), 1057–1062.

Okasha, S., 2006. Evolution and the Levels of Selection. Oxford University Press, Oxford.

Olsson, M., et al., 2005. MHC, health, color, and reproductive success in sand lizards. Behavioral Ecology and Sociobiology 58 (3), 289–294.

Park, J.H., Schaller, M., Van Vugt, M., 2008. Psychology of human kin recognition: heuristic cues, erroneous inferences, and their implications. Review of General Psychology 12 (3), 215–235.

Parkin, D., Stone, L., 2004. Kinship and Family: an Anthropological Reader. Wiley-Blackwell, London.

Penn, D.J., Frommen, J.G., 2010. Kin recognition: an overview of conceptual issues, mechanisms and evolutionary theory. In: Kappeler, P.M. (Ed.), Animal Behavior: Evolution and Mechanisms. Springer, Heidelberg, pp. 55–85.

Penn, D.J., Potts, W.K., 1999. The evolution of mating preferences and major histocompatibility complex genes. The American Naturalist 153 (2), 145–164.

Pizzari, T., Gardner, Andy, 2012. The sociobiology of sex: inclusive fitness consequences of inter-sexual interactions. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 367 (1600), 2314–2323.

Pomiankowski, A., 2002. Genomic imprinting. In: The Oxford Encyclopedia of Evolution. Oxford University Press, Oxford, pp. 423–428.

Popa-Lisseanu, A.G., et al., 2008. Highly structured fission–fusion societies in an aerial-hawking, carnivorous bat. Animal Behaviour 75 (2), 471–482.

Pusey, A., Wolf, M., 1996. Inbreeding avoidance in animals. Trends in Ecology & Evolution 11 (5), 201–206.

Queller, D.C., 2010. Kin selection and relatedness. In: Breed, M., Moore, J. (Eds.), Encyclopedia of Animal Behavior. Elsevier, pp. 247–252.

Queller, D.C., 2011. Expanded social fitness and Hamilton's rule for kin, kith, and kind. Proceedings of the National Academy of Sciences of the United States of America 108 (Suppl.), 10792–10799.

Rantala, M.J., Marcinkowska, U.M., 2011. The role of sexual imprinting and the Westermarck effect in mate choice in humans. Behavioral Ecology and Sociobiology 65 (5), 859–873.

Reik, W., et al., 2003. Regulation of supply and demand for maternal nutrients in mammals by imprinted genes. The Journal of Physiology 547 (Pt 1), 35–44.

Rendall, D., 2004. Kin bias: proximate and functional processes. In: Chapais, B., Berman, C.M. (Eds.), Kinship and Behaviour in Primates. Oxford University Press, Oxford, pp. 295–316.

Reusch, T.B., et al., 2001. Female sticklebacks count alleles in a strategy of sexual selection explaining MHC polymorphism. Nature 414 (6861), 300–302.

Richerson, P.J., Boyd, R., 2004. Not by Genes Alone: How Culture Transformed Human Evolution. University of Chicago Press, Chicago.

Ronce, O., 2007. How does it feel to be like a rolling stone? Ten questions about dispersal evolution. Annual Review of Ecology, Evolution, and Systematics 38 (1), 231–253.

Setchell, J.M., Huchard, E., 2010. The hidden benefits of sex: evidence for MHC-associated mate choice in primate societies. BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology 32 (11), 940–948.

Shelley, E.L., Blumstein, D.T., 2004. The evolution of vocal alarm communication in rodents. Behavioral Ecology 16 (1), 169–177.

Shenk, M.K., Mattison, S.M., 2011. The rebirth of kinship. Human Nature 22 (1–2), 1–15.

Silk, J., 2002. Kin selection in primate groups. International Journal of Primatology 23 (4), 849–875.

Sober, E., Wilson, D.S., 1998. Unto Others: the Evolution and Psychology of Unselfish Behavior. Harvard University Press, Cambridge.

Starkweather, K.E., Hames, R., 2012. A survey of non-classical polyandry. Human Nature (Hawthorne, N.Y.) 23 (2), 149–172.

Strassmann, J.E., Gilbert, O.M., Queller, D.C., 2011. Kin discrimination and cooperation in microbes. Annual Review of Microbiology 65, 349–367.

Tabadkani, S.M., Nozari, J., Lihoreau, M., 2012. Inbreeding and the evolution of sociality in arthropods. Die Naturwissenschaften 99 (10), 779–788.

Tregenza, T., Wedell, N., 2000. Genetic compatibility, mate choice and patterns of parentage: invited review. Molecular Ecology 9 (8), 1013–1027.

Trivers, R.L., 2002. Natural Selection and Social Theory: Selected Papers of Robert Trivers. Oxford University Press, Oxford.

Ward, A.J.W., Hart, P.J.B., 2003. The effects of kin and familiarity on interactions between fish. Fish and Fisheries 4 (4), 348–358.

Wedekind, C., 2007. Body odours and body odour preferences in humans. In: Dunbar, R.I.M., Barrett, L. (Eds.), Oxford Handbook of Evolutionary Psychology. Oxford University Press, Oxford, pp. 315–320.

Weir, B.S., Anderson, A.D., Hepler, A.B., 2006. Genetic relatedness analysis: modern data and new challenges. Nature Reviews. Genetics 7 (10), 771–780.

Widdig, A., et al., 2006. Consequences of group fission for the patterns of relatedness among rhesus macaques. Molecular Ecology 15 (12), 3825–3832.

Widdig, Anja, 2007. Paternal kin discrimination: the evidence and likely mechanisms. Biological Reviews of the Cambridge Philosophical Society 82 (2), 319–334.

Williams, G.C., 1966. Adaptation and Natural Selection. Princeton University Press, Princeton.

Wilson, E.O., 1998. Consilience: the Unity of Knowledge. Alfred A. Knopf, New York.

Wilson, E.O., 1971. The Insect Societies. Harvard University Press, Cambridge.

Wilson, E.O., 2012. The Social Conquest of Earth. W. W. Norton, New York.

Wilson, E.O., Hölldobler, B., 2005. Eusociality: origin and consequences. Proceedings of the National Academy of Sciences of the United States of America 102 (38), 13367–13371.

Wolf, A.P., Durham, W.H. (Eds.), 2004. Inbreeding, Incest, and the Incest Taboo. Stanford University Press, Stanford.

Zelano, B., Edwards, S.V., 2002. An MHC component to kin recognition and mate choice in birds: predictions, progress, and prospects. The American Naturalist 160 (Suppl.), S225–S237.

Kinship, Formal Models of

Dwight Read, University of California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Kinship involves social interactions and forms of social organization based on systems of culturally constructed social relations expressed linguistically through the kin terms constituting a kinship terminology. The organization and structure for these social relations expressed in a kinship terminology is formally modeled here as being part of an axiomatic theory in which a few, primary kinship concepts are the equivalent of axioms in a mathematical theory. The cultural knowledge embedded in a kinship terminology therefore is part of a cultural theory – analogous to a mathematical theory – that expresses the culturally salient kinship properties derived from the primary kinship concepts.

Introduction: The Domain of Kinship

Kinship involves social interactions and forms of social organization based on systems of culturally constructed social relations expressed linguistically through the kin terms constituting a kinship terminology. Formal modeling of structural patterning in the social relations expressed through kin terms began with Lewis Henry Morgan's systematic study of kinship terminologies in the mid-1800s. Since then, the analysis of kinship systems has been central to anthropological theorizing, though some aspects such as what constitutes the cultural criteria by which one person is the kin of another person are still topics of considerable disagreement. Regardless, all societies have a corpus of terms – a kinship terminology – that expresses for culture bearers their cultural concepts regarding kinship. A kin term referentially designates, in a culturally meaningful and mutually understood manner, the kin relation one person may have to another person, thereby identifying those who are kin to speaker and to each other.

The British anthropologist Rodney Needham usefully divided the domain of kinship into three aspects: (1) behavior – what people actually do, (2) rules – what people say are the rules or jural restrictions on behavior, and (3) categories and relations among categories. These divisions have content varying from less to more abstract and the third is the focus here for the formal modeling of kinship systems.

Modeling of categories and relations among categories has generally considered one of two related analytical frameworks: (1) the structure and organization of a kinship terminology or (2) the implication of prescriptive marriage rules for the structural integration of social units such as families, lineages, clans, or even whole societies. The first framework will be the focus of this article. For this framework, it has generally been assumed – though incorrectly – that kin terms are primarily linguistic labels for categories of genealogical relations. While kin terms have this function, it is secondary to the categories being derived from the implications of the primary kinship concepts embedded in a kinship terminology.

The organization and structure of the kinship relations expressed through a kinship terminology will be formally modeled by considering a few, primary kinship concepts to be the equivalent of axioms in a mathematical theory. The cultural knowledge embedded in a kinship system therefore incorporates a cultural theory – analogous to a mathematical theory – that expresses the culturally salient kinship properties derived from these primary kinship concepts. These concepts have roots both in (without being determined solely by or limited to) the biological facts of reproduction and in culturally defined marriage systems that define conditions for offspring to be fully recognized as social members of a society or a social stratum. Marriage rules – the cultural stipulation of the social facts sufficient for assigning a complete social identity to the offspring of a woman – structure the interconnections among social groups either positively through prescriptive marriage rules or negatively through incest taboos.

Primary Kinship Concepts

We begin by identifying culturally salient concepts of relatedness that are the axioms for cultural theories of kinship relations. These concepts derive from the universally recognized statuses of *motherhood*, *fatherhood*, and *spousehood*, each having different, culturally specific criteria for its implementation. In its default mode, motherhood involves 'mothering' – the positive caring, nurturing, and feeding behaviors that a female directs toward those she recognizes as her offspring – but 'mothering,' hence motherhood, is not limited to behaviors directed in this manner and can arise in conjunction with other practices such as adoption, suckling, and/or coresidence, among others.

Similarly, what constitutes fatherhood is culture-specific and the status can, but need not always, be initiated through cultural recognition of, and the meaning assigned to, a male's biological role in a woman becoming pregnant. In some societies, the biological role of semen in pregnancy is not given importance in local theories of reproduction and pregnancy is culturally attributed to other actions a man engages in that relate to establishing the social identity of a newborn.

Motherhood (fatherhood) leads to the concept of a *mother relation* (*father relation*) through associating mothering behavior with the dyad of two interacting individuals that is the locus for the behavior associated with motherhood. Cognizing a mother relation in this sense is not just a human capacity but occurs among the macaques (Dasser, 1988). The father relation in human societies, however, does not have a counterpart among

nonhuman primates since the status of fatherhood does not exist for most nonhuman primate species.

We can define the idea of a mother relation (father relation) through the ensemble of behaviors, B_{Mo}, associated with motherhood (B_{Fa}, associated with fatherhood). We define a *mother relation*, M, as follows.

Definition of Mother Relation M: For female person *s* and person *t*, let *s*M*t* (read '*s* is the mother of *t*') be true when *t* is recognized as the target for *s* when engaging in the mothering behaviors in B_{Mo}; that is, *s* takes on the motherhood status vis-a-vis *t* (see left side of **Figure 1** in gray) by virtue of engaging in the behaviors making up B_{Mo} with *t*, the recipient of those behaviors.

A *father relation*, F, may be defined in a similar manner using the behaviors in B_{Fa}.

There are several boundary conditions included in the cultural instantiation of these two relations: (1) if *s*M*t* or *s*F*t* is true, then $s \neq t$, (2) for no persons *s*, *t* is it the case that *s*M*t* and *s*F*t* are simultaneously true, and (3) for each person, *u* there is presumed to be a person *s* and a person *t* such that *s*M*u* and *t*F*u*. Another condition, corresponding to the assertion that a single female has the primary responsibility for the care and well-being of a child (Schneider, 1961), is that for persons *s*, *t*, and *u*, *s*M*t* and *u*M*t* are simultaneously true only when $u = s$.

Since both of these statuses require that there be a person who is the recipient of the behaviors, it follows that there is a reciprocal *daughter relation* and *son relation* for the mother and father relations definable from the perspective of the recipient of the B_{Mo} and B_{Fa} behaviors. The daughter relation can be defined as follows.

Definition of Daughter Relation D: For persons *s* and *t*, let *s*D*t* (read '*s* is the daughter of *t*') be true when *s* is female and either *t*M*s* or *t*F*s* is true.

A *son relation* may be defined in a similar manner.

The M and F relations are conceptually independent since *s* can be recognized as the mother of *t* without a person being recognized as father of *t*, or vice versa. However, cultural kinship systems are bilateral, meaning that the mother relation and the father relation are both axiomatic for generating kinship concepts. By itself, this would lead to conceptually disconnected sets of kinship relations. Matrilateral relations (those who are related to speaker through a sequence of one or more relations beginning with the mother relation and then followed by a parental or sibling relation when there is more than one relation in the sequence) and patrilateral relations (defined similarly except beginning with the father relation) would lack a conceptual connection.

The cultural institution of marriage conceptually forms a connection between the mother relation and the father relation, thereby removing disconnectedness. Through marriage, a spousehood status (see top part of **Figure 1**) is defined using culture-specific behavioral criteria that determine a symmetric *spouse relation* in a manner analogous to how the mother relation is determined from the motherhood status. Associating the social identity of offspring with

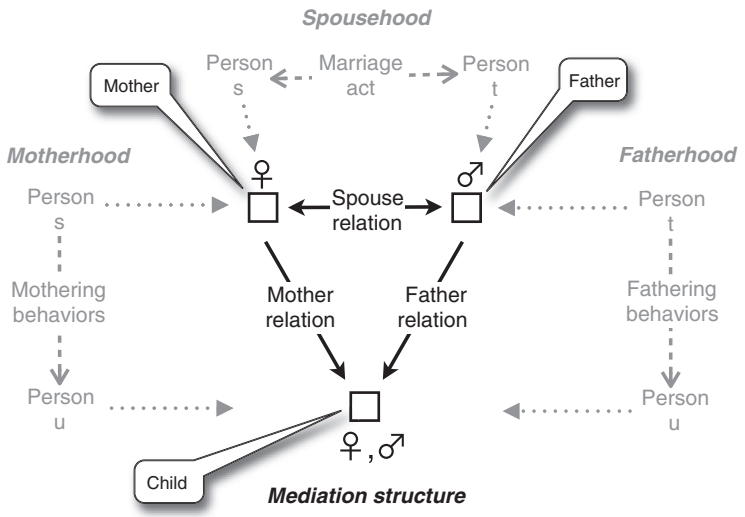

Figure 1 Motherhood and fatherhood are statuses determined by behaviors directed to another person (dashed arrows). Spousehood is a status determined by a marriage act (dashed arrow). Each is the basis for identifying when a relation holds between two persons. The three relations form a mediation structure (El Guindi, Fadwa, Read, Dwight, 1979. Mathematics in structural theory. Current Anthropology 20, 761–790; Read, Dwight, 2010a. Mathematical representation of cultural constructs. In: Kronenfeld, D., Bernardo, G., De Munck, V.C., Fischer, M.D. (Eds.), A Companion to Cognitive Anthropology. Wiley-Blackwell Publishing, Malden, MD) for the three categories (shown by boxes) labeled *mother*, *father*, and *child* with the indicated gender attributes. The structural opposition between the *mother* category and the *father* category is expressed through the *spouse* relation and each of the categories in opposition is linked to a mediating category (bottom box) through the *mother* and *father* relations, respectively. Persons *s*, *t*, and *u* may be categorized as indicated by the dotted arrows, thereby conceptually making *s mother of u* and *t father of u* and, reciprocally, *u* is *child of* s and *u* is *child of t*. The *child* category (lower box) has both male and female attributes as required for a mediating category. The spouse categorization conceptually makes *s spouse of t* and *t spouse of s*. Generally the biological sex of the person being categorized agrees with the gender of the category, though in some societies such as the *LoDagaba* of Ghana, Africa, a male person – a mother's brother – is categorized as a kind of mother and correspondingly is referred to by the term *madeb* whose literal meaning is 'male mother' (Goody, Jack, 1959. The mother's brother and the sister's son in West Africa. The Journal of the Royal Anthropological Institute of Great Britain and Ireland 89, 61–86). In several African societies such as the *Nandi* of Kenya, female fatherhood arises through female husbands (Oboler, Regine S., 1980. Is the female husband a man? Woman/woman marriage among the Nandi of Kenya. Ethnology 19, 69–88). Note: Terms, symbols, and arrows in gray refer to phenomena in the material domain and those in black refer to concepts in the cultural domain.

marriage ensures that there is at least a presumed, if not actual, spouse relation between a person who culturally instantiates the mother relation, if any, and a person who culturally instantiates the father relation, if any, for an offspring to have a complete social identity.

The mother, father, and child categories and the spouse, mother, and father relations are conceptually integrated through a *mediation structure* (El Guindi and Read, 1979; Read, 2010a). A mediation structure consists of an open, mediating category having a pair of attributes with opposite values that is linked to a pair of closed categories, each of which has associated with it one of the pair of opposite-valued attributes associated with the mediating category (see central part of **Figure 1** in black). The child category has the pair of opposite attributes, *male*, *female*, and is open since it categorizes the outcome of a biological birth. The mother category has the attribute *female*, the father category has the attribute *male*, and each is linked to the child category through the mother relation and the father relation, respectively. They are closed as there are culture-specific criteria for being culturally recognized as a mother or a father. The mother and father categories are, then, categories in opposition and conceptually one goes from the mother category to the child category using the mother relation and then from the child category to the father category using the inverse of the father relation, and vice versa. This link is culturally enabled through the institution of marriage and the associated spouse relation defined by *mother of child of father* is *spouse* is *father of child of mother*.

More succinctly, the spouse relation expresses the opposition between the mother and father categories by the computations, *spouse of mother is father* and *spouse of father is mother*. These computations map the female gender of mother to the male gender of father. The derived kinship system based on mother, father, and spouse will be bilateral. Also, it follows that a person can be categorized as father without reference to the biological facts of reproduction, as occurs in groups such as the Tiwi of Australia where *father* is the man currently married to *mother* regardless of his biological relationship to her child. Yet another implication is that the axiomatic foundation of kinship can neither be reduced to biological nor to cultural properties alone, as it necessarily incorporates both.

Family Structure

Kinship systems universally recognize a *sibling relation* derived from the biological fact that multiple offspring can be produced through reproduction. While each offspring may be categorized as *child* in the mediating structure shown in **Figure 1**, *siblinghood status* arises through behaviors between offspring framed by having at least one person in common categorized as mother or father. A sibling relation may be defined as follows.

Definition of Sibling Relation SB: For persons s and t, let sSBt (read 's is the sibling of t') be true when there is a u and a v such that both of uMs and uMt and both of vFs and vFt are true. The sibling relation expands the mediating structure (**Figure 2(a)**) into what we will call a *family structure* (see **Figure 2(b)**). It should be noted that this structure neither defines nor determines the empirical form of actual families. Rather, it illustrates the structural relationships among the axiomatic kinship relations that are integral to the concept of family.

Next we begin the construction of a *kinship domain* by first constructing a Family Space derived from the family structure. Then we add a Genealogical Space (GS) and a Kin Term Space and finally we indicate the structural relationships among these spaces.

Family Space

Kin terms identify a kinship relation between speaker and target person, whereas the primary kinship relations introduced above are from the perspective of an observer noting the behavior between the two members of a dyad. Shifting from an observer to an actor or agent perspective corresponds to introducing the universal concept of self into the family structure and changing, for example, the relation, 's is mother of t,' into, 's is my mother,' with t categorized as self. Introducing *self* transforms the Family Structure of relations into a Family Space of positions and their connections to *self*.

Two Family Spaces are possible, depending on whether the sibling relation is derived from the parent and child relations (see **Figure 3(a)**), or is primary (see **Figure 3(b)**). Both concepts regarding sibling relations are empirically found in

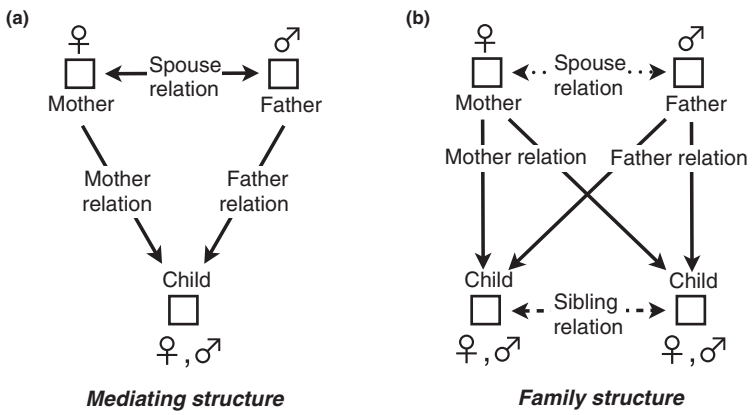

Figure 2 (a) The mediating structure from **Figure 1**. (b) Modified mediating structure with a sibling relation. The structure incorporates all of the elements for the concept of a family. The structure is that of a nuclear family, but actual families have empirical form depending on the cultural instantiation of the categories in the family structure.

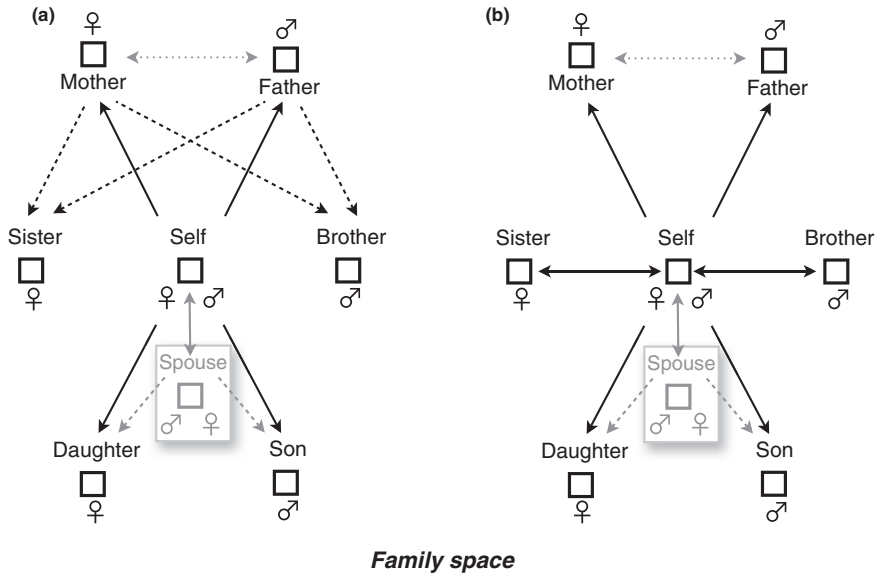

Figure 3 (a) and (b) Solid arrows (black or gray) are the primary relations. Spouse is in a third dimension and connected to *self* by the spouse relation (double headed arrow). *Daughter* of *spouse* is *daughter* and *son* of *spouse* is *son* (dashed arrows). The mother and father positions are connected with the spouse relation (dotted, double-headed arrow). (a) Family Space based on constructing the sibling relations through parent: *sister* is *child* of *parent* and *brother* is *son* of *parent* (dashed arrows). (b) Family Space when the sibling relation is a primary relation (double-headed arrows).

human societies (see Marshall, 1983 for ethnographic examples of axiomatic sibling relations).

Genealogical Space

The GS consists of *genealogical paths* constructed from the relations in the Family Space through (1) each relation in the Family Space is a genealogical path and (2) if P and Q are genealogical paths, then the product PQ defined through concatenation is a genealogical path. GS consists of all possible products of the form $P = P_1, P_2, \ldots P_n$, $1 \leq n < \infty$, where each P_i, $1 \leq i \leq n$, is a relation in the Family Space.

For persons s and t, a *genealogical claim* connecting s to t consists of a list of persons, beginning with s and ending with t, such that all adjoining persons in the list satisfy one of the relations in the Family Space. A genealogical path, P, may (potentially) be instantiated recursively over a group G of persons as a genealogical claim from person s to some person t in G as follows. Instantiate *self* with s and find s_1 in G (if any) such that $s\, P_1\, s_1$ is true. If $n = 1$, set $t = s_1$ and the list, $[s, t]$, is the desired genealogical claim. Otherwise, recursively instantiate *self* with s_1 and find s_2 in G (if any) such that $s_1\, P_2\, s_2$ is true. If $n = 2$, set $t = s_2$ and $[s, s_1, t]$ is the desired genealogical claim. Otherwise, continue recursively in this manner. If there is a person s_i in G with $s_{i-1}\, P_i\, s_i$ true for all i, $1 \leq i \leq n$, then the genealogical path P determines the genealogical claim $[s, s_1, s_2, \ldots, s_{n-1}, t]$ connecting s and $t = s_n$. Conversely, each genealogical claim corresponds to a genealogical path.

Kin Term Space and Kin Term Maps

The size of the GS scales with k^n, where k is the number of relations used to form genealogical paths and n is the length of a genealogical path, hence GS is cognitively complex even for short genealogical paths. In all societies, this complexity is reduced to categories of genealogical paths corresponding to the 15–30 kin terms in a kinship terminology. Unlike folk taxonomies, though, kin relations can be computed directly by culture bearers from other kin terms using the knowledge they have of their kinship terminology as a system of interconnected terms. As reported in numerous ethnographic accounts, the calculations have the form of a *kin term product* (see Figure 4) defined as follows.

Definition of a Kin Term Product: If person s (properly) refers to person t using the kin term L and person t (properly) refers to person u using the kin term K, then the kin term M, if any, that

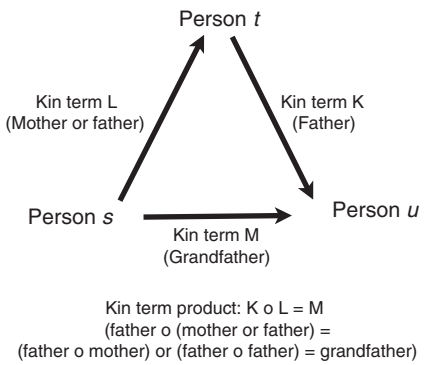

Figure 4 Definition of the kin term product, K o L = M, for the kin term L used by person s to refer to person t and the kin term K used by person t to refer to person u. The product is illustrated by the kin terms in parentheses. The English kin term *grandfather* is determined by the product of the kin term *father* with the kin term *parent* = (*mother*, *father*).

person *s* uses to (properly) refer to person *u* is the product of the kin terms K and L: M = K o L.

The kin term product makes it possible to generate the kin terms in a kinship terminology starting with the kinship relations in the Family Space. The lexeme K is a (*primary*) *kin term* corresponding to the Family Space relation R when there is cultural agreement that person *s* may (properly) refer to person *t* by K when *t*R*s* is true. For example, English speakers agree that it is proper for *s* to refer to *t* as 'my parent' when either '*t* is my mother' (*t*M*s*) or '*t* is my father' (*t*F*s*) is true for *s*, so *parent* becomes a primary kin term for English speakers. A new kin term may be constructed from a primary kin term K and an already identified kin term L by letting the kin term product K o L determine the relation person *s* has to person *u* when *s* refers to person *t* by the term L and *t* refers to *u* by the term K (see **Figure 4**). Naming this kin term product incorporates it as a kin term in the kinship terminology. The same name may be used for more than one product as part of the kinship concepts embedded in the kinship terminology. For English speakers, for example, *grandfather* is the name for the kin term product *father* o *mother* and it is also the name for the kin term product *father* o *father* (see **Figure 4**).

We visually display the structure of a kinship terminology with a *kin term map* (see **Figure 5**) made by letting each kin term be a node and drawing an arrow corresponding to a primary kin term (with a different arrow form for each primary kin term) from a kin term K to the kin term L when L is the product of the primary term with the kin term K. The kin term maps for the American/English kinship terminology and the terminology of the Shipibo Indians, a horticultural group in the eastern part of Peru, are shown in **Figure 5**. Structural differences in their respective kinship terminologies are immediately apparent. The Shipibo terminology has symmetry between matrilateral and patrilateral kin reflected in the symmetry of the women's designs used to decorate pottery, clothing, faces, and other parts of the body that relate to therapeutic aspects of curing and bewitching shamanistic practices (Roe, 2004). The American terminology has a contrast between lineal and collateral kin reflected in the practice of a woman and her children taking on the name of her husband even though lineal descent groups are not culturally recognized.

Generating a Kinship Terminology

The procedure for generating new kin terms through the kin term product suggests that kinship terminologies may share, despite their structural differences (see **Figure 5**), a common scheme for the generation of a kinship terminology as an abstract algebra consisting of the ordered triplet, <T, o, S>, where T is the set of kin terms including self, o is the kin term product interpreted as a binary operation over T with self an identity element (0 is included in T to ensure that o is closed over T), and S is a set of ethnographically validated structural equations giving the algebra its particular structure as displayed in the kin term map. Analysis of a wide variety of kinship terminologies leads to the following construction schema expressed as a sequence of steps (see Leaf and Read (2012) for a more detailed presentation; examples of generating a terminology can be found in Read (1984) (American terminology), (2010b) (Dravidian terminology), (2012) (Kung San terminology), Read and Behrens (1990) (English, Shipibo, and Trobriand terminologies), Bennardo and Read (2007) (Tongan terminology); and Leaf and Read (2012) (Kariera and Punjabi terminologies)).

Step 1: Generate an ascending structure of kin terms based on either the generating set (1) A = {self, P}, where P is an ascending primary term or (2) A = {self, P, G}, where G is a primary sibling term, along with the structural equations G o G = G ('sibling of sibling is sibling') and P o G = P ('parent of sibling is parent') included in S. Inclusion of a primary sibling term accounts for the defining structural properties of the terminologies that Morgan distinguished

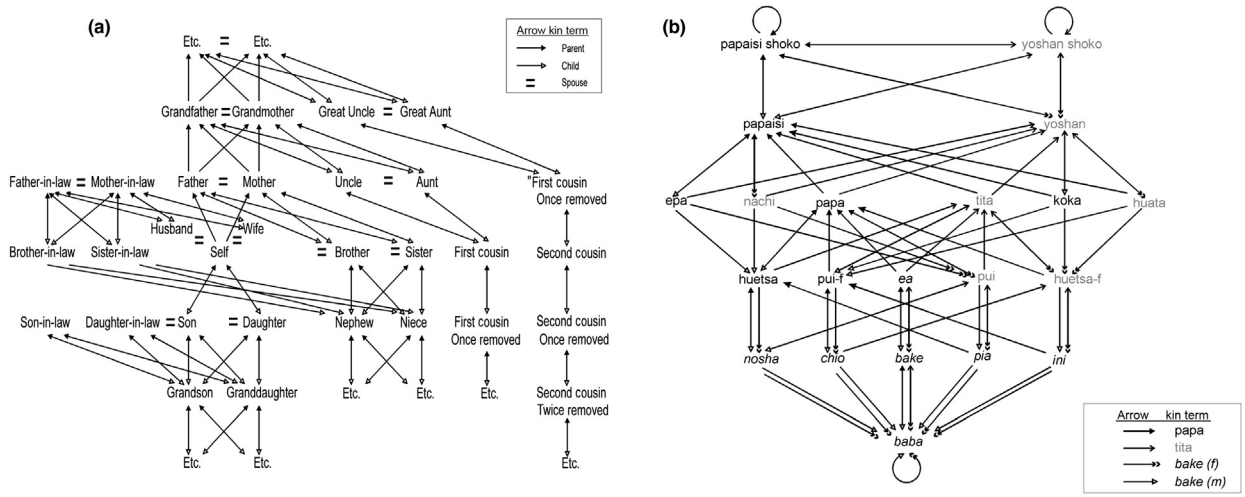

Figure 5 (a) Kin term map of the American/English kinship terminology based on *parent*, *child*, and *spouse* as the primary terms. (b) Kin term map of the Shipibo Indians of eastern Peru based in *ea* ('self'), *papa* ('father'), *tita* ('mother'), and *bake* ('child') as the primary kin terms. Affinal kin terms have not been included.

as *classificatory terminologies*. For some terminologies, either the structural equation $A^n = 0$ or $A^n = A^{n-1}$, $n > 2$, that limits producing new kin terms in an ascending direction is included in S, where $A^n = A \circ A \circ \ldots \circ A$ (n times).

Step 2: Generate an isomorphic, descending structure of kin terms based on either the generating set (1) D = {self, C}, where C is a descending primary term or (2) D = {self, C, G*}, where G* is a primary sibling term distinct from G. (The sibling terms G and G* have interpretation as *ascending sibling* and *descending sibling*, respectively, though they are usually referred to as older sibling and younger sibling.) Include in S, the isomorphic version of any structural equation included in Step 1.

Step 3: Form a single structure with generating set G = A ∪ D. Include in S, the structural equation P o C = self ('parent of child is self' for consanguineal relations) that defines P and C to be reciprocal kin terms. If G, G* are in G, then include the structural equations G o G* = self = G* o G ('ascending sibling of descending sibling is self is descending sibling of ascending sibling') that define G and G* to be reciprocal kin terms.

Step 4: Introduce 'sex marking' of kin terms either by (1) adding sex marking elements m and f to S that are right-identities in S or (2) making an isomorphic copy of the structure generated in Step 3, but with a new term, self*, replacing self in the isomorphic copy. The structure and its isomorphic copy are interpreted as male-marked and female-marked terms, respectively. These two structures are joined to form a single structure of both male-marked and female-marked kin terms. Different ways the structures may be joined correspond to major structural differences among the classificatory terminologies.

Step 5: Introduce affinal terms by either (1) adding a spouse primary term, Sp, to S, along with appropriate structural equation such as Sp o P = P, Sp o Sp = Sp, and so on or (2) defining a kin term in S to also be an affinal term through a marriage rule, such as the prescriptive marriage rule in many Australian societies specifying that a man marries one of the women he refers to by such-and-such a kin term (e.g., the kin term for a cross-cousin).

Step 6: Introduce terminology specific kin term rules that locally modify the structure determined from Steps 1–5.

Step 7: Introduce any relevant culture-specific kin term distinctions that arise from usage of the terminology, such as the term with translation, 'younger brother of mother,' in the Tongan terminology due to inheritance rules.

The terminologies generated in this manner using the primary concepts of a kinship terminology have been found to have structure isomorphic to the structure shown in a kin term map for that terminology, hence the schema presented here can be considered to determine a theory explanatory of the properties of kinship terminologies.

Relations among the Family Space, the GS, and the Kin Term Space

The relations among the Family Space, the GS, and the Kin Term Space are shown in Figure 6 and derived as follows. Both the GS and the Kin Term Space are constructed from the relations in the Family Space, hence there is a natural mapping, h, of the GS into the Kin Term Space defined as follows. Map each path of length 1 to its corresponding primary kin term; for example, for the English terminology, h: M → *mother*. For all other paths, map the relations in the genealogical path to the kin term product of the corresponding primary kin terms and reduce this product to a kin term in the Kin Term Space. Thus for the genealogical path [M, B, D], h:[M, B, D] → *daughter* o *brother* o *mother* = *daughter* o *uncle* = *cousin*. (Note: products of kin terms are written from right to left to allow a product such as *mother* o *father* to be read as the kin term product 'mother of father,' whereas a genealogical path is written from left to right to allow a list such as [F, M] to be read as the claim, 'ego's father's mother.')

The mapping is a *homomorphism* since for genealogical paths $P = [P_1, \ldots, P_n]$ and $Q = [Q_1, \ldots, Q_m]$, $h{:}PQ = h{:}[P_1, \ldots, P_n, Q_1, \ldots, Q_m] = h{:}Q_m \circ \ldots \circ h{:}Q_1 \circ h{:}P_n \circ \ldots \circ h{:}P_1 = h{:}[Q_1, \ldots, Q_m] \circ h{:}[P_1, \ldots, P_n] = h{:}Q \circ h{:}P$. It is a *surjective* (onto) mapping since any kin term, K, in the Kin Term Space may be written as a product of primary kin terms, hence the genealogical path determined by a sequence of relations constructed with each relation in the sequence mapped by h to the corresponding primary term in the product definition of K will be mapped to K by h.

Predicted genealogical definitions of kin terms are given by the inverse mapping, h^{-1}, and the predictions match empirical observations. The mapping, h, it should be noted, ensures that computations with kin terms will be consistent with genealogical computations.

Formal Representation of Computations over the GS

Most symbolic systems devised for representing genealogical paths are essentially notation systems. One of the more elaborated symbolic representations of this kind is the Grafik symbol system (Atkins, 1974), but it does not represent a computational system for doing computations with genealogical relations. The latter has been developed formally by distinguishing between genealogical claims that conceptually link one person to another through a chain of persons connected by parent and/or child relations and an abstract positional schema for expressing the genealogical path embedded in a genealogical claim (Lehman and Witz, 1974). Culture bearers do not just make genealogical claims; they also do genealogical computations with the genealogical relations that provide the connections among the adjacent persons in a genealogical claim. Lehman and Witz define an abstract *Primary Genealogical Space* (PGS) as the multiplicative semigroup generated from four filiation symbols P^m, P^f, ${}^mQ$, ${}^fQ$ that represent the relations 'has male parent,' 'has female parent,' 'is child of male person,' and 'is child of female person,' respectively, and are derived from the upward lineation symbol P and its inverse downward lineation symbol Q, all subject to the structural equations $(P^x)({}^yQ) = 0 = ({}^xQ)(P^x)$, where x, y = m (male) or f (female) and $x \neq y$ (1974: p. 119). (Their PGS is essentially what is being called GS here.) The first equation says that a male (female) parent cannot also be a female (male) person having a child and the second eliminates genealogical

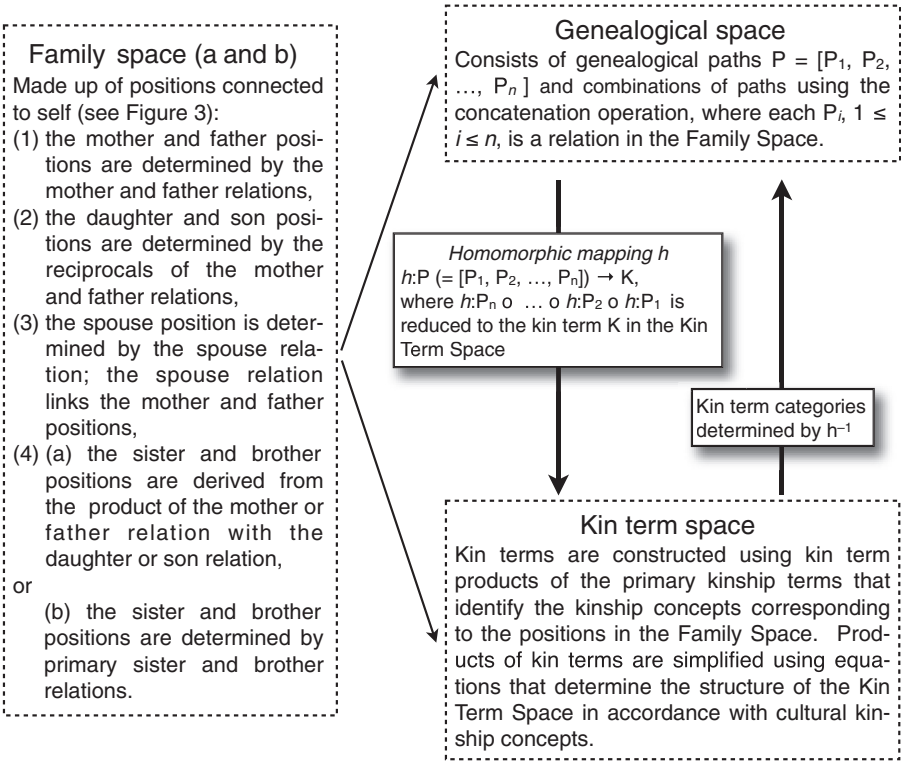

Figure 6 Relationships among the Family Space, the Genealogical Space (GS), and Kin Term Spaces. The GS may be mapped homomorphically onto a Kin Term Space, implying that the structure of the GS is carried over to a Kin Term Space. Genealogical definitions are determined by the inverse of the homomorphism mapping from GS onto Kin Term Space.

claim loops that go from a person to a child and back to the starting person. They show that from PGS a positional schema can be defined that formally represents the process of genealogical reckoning.

Summary

The kinship domain has been formally represented here as being composed of a Family Space of positions defined with respect to self, a GS of genealogical paths defined as sequences of relations from the Family Space, and a Kin Term Space defined through kin term products of the primary kin terms determined from the positions and relations in the Family Space and subject to equations that structurally express kinship concepts structuring a particular kinship terminology. The formal representation establishes that the GS can be mapped homomorphically onto a Kin Term Space, meaning that calculations made with genealogical paths will be consistent with computations in a Kin Term Space.

The formal representation relates to empirical observations through the systematic elicitation of kin terms from culture bearers (Leaf, 2006) conducted by asking about kin term products of the primary kin terms with the primary kin terms and then products with any elicited kin terms. The elicitation establishes that a terminology is conceptually bounded since any sequence of kin term products with a primary kin term eventually terminates, either because the next product does not yield a new kin term or it yields a previously elicited term (or possibly a variant on a previously elicited term such as great … great grandparent in English).

The homomorphism from the GS onto a Kin Term Space and the algebraic structure of the Kin Term Space together imply that the terminology provides a symbolic way, through the kin term product, to compute kinship relations in a manner consistent with genealogical reckoning. Thus the relationship between calculating with genealogical paths and the symbolic kin term computations with the kin term product is analogous to the relationship between calculating with counting numbers (where a counting number represents the cardinality of a set of objects) and symbolic number computations using the addition operation with the natural numbers. In both situations, a formal representation makes evident how new concepts are generated from a few concepts taken as self-evident, namely the counting number 1 and the idea of a successor number in the case of arithmetic and the positions and relations of the Family Space and the idea of the kin term product as a way to construct a new kinship relation in the case of kinship systems. The formal representation lays the foundation for exploring the interplay of kinship with other domains whose cultural properties are expressed using kinship concepts.

Bibliography

Atkins, John R., 1974. Grafik: a multipurpose kinship metalanguage. In: Ballonoff, P. (Ed.), Genealogical Mathematics. Mouton, Paris.

Bennardo, Giovanni, Read, Dwight, 2007. Cognition, algebra, and culture in the Tongan kinship terminology. Journal of Cognition 7, 49–88.

Dasser, Verena, 1988. Mapping social concepts in monkeys. In: Byrne, R.W., Whiten, A. (Eds.), Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans. Oxford University Press, New York.

El Guindi, Fadwa, Read, Dwight, 1979. Mathematics in structural theory. Current Anthropology 20, 761–790.

Goody, Jack, 1959. The mother's brother and the sister's son in West Africa. The Journal of the Royal Anthropological Institute of Great Britain and Ireland 89, 61–86.

Leaf, Murray, 2006. Experimental analysis of kinship. Ethnology 45, 305–330.

Leaf, Murray, Read, Dwight, 2012. The Conceptual Foundation of Human Society and Thought: Anthropology on a New Plane. Lexington Books, Lanham, MD.

Lehman, Frederick K., Witz, Klaus, 1974. Prolegomena to a formal theory of kinship. In: Ballonoff, P. (Ed.), Genealogical Mathematics. Mouton, Paris.

Marshall, Mac (Ed.), 1983. Siblingship in Oceania. University Press of America, Lanham, MD.

Oboler, Regine S., 1980. Is the female husband a man? Woman/woman marriage among the Nandi of Kenya. Ethnology 19, 69–88.

Read, Dwight, 1984. An algebraic account of the American kinship terminology. Current Anthropology 25, 417–449.

Read, Dwight, 2010a. Mathematical representation of cultural constructs. In: Kronenfeld, D., Bernardo, G., De Munck, V.C., Fischer, M.D. (Eds.), A Companion to Cognitive Anthropology. Wiley-Blackwell Publishing, Malden, MD.

Read, Dwight, 2010b. The generative logic of Dravidian language terminologies. Mathematical Anthropology and Cultural Theory 3 (7).

Read, Dwight, 2012. How Culture Makes Us Human: Primate Evolution and the Formation of Human Societies. Left Coast Press, Walnut Creek, CA.

Read, Dwight, Behrens, Clifford, 1990. KAES: an expert system for the algebraic analysis of kinship terminologies. Journal of Quantitative Anthropology 2, 353–393.

Roe, Peter, 2004. At play in the fields of symmetry: design structure and shamanic therapy in the Upper Amazon. In: Washburn, D.K., Crowe, D.W. (Eds.), Symmetry Comes of Age: The Role of Pattern in Culture. University of Washington Press, Seattle, WA.

Schneider, David M., 1961. Introduction: the distinctive features of matrilineal descent groups. In: Schneider, D.M., Gough, K. (Eds.), Matrilineal Kinship. University of California Press, Berkeley, CA.

Relevant Websites

http://www.ausanthrop.net/research/kinship/ – research, resources and documentation on the study of kinship.

http://kaes.anthrosciences.net/ – University of Kent, Centre for Social Anthropology and Computing.

http://sourceforge.net/projects/kam/ – Kinship Algebra Modeller.

http://www.umanitoba.ca/faculties/arts/anthropology/kintitle.html – Kinship and Social Organization: An Interactive Tutorial.

Kinship Terminology

Dwight Read, University of California, Los Angeles (UCLA), Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by N. Evans, volume 12, pp. 8105–8111, © 2001, Elsevier Ltd.

Abstract

Kinship terminologies consist of the terms used to reference culturally recognized kinship relations between persons. These terms have been assumed to identify categories of genealogical relations (despite ethnographic evidence to the contrary), and kinship terminologies are classified using differences in genealogical referents of kin terms. Recent analysis, however, by building on ethnographically validated procedures for computing kin relations from kin terms without reference to genealogy makes evident the underlying generative logic for the structure of kinship terminologies. Making the generative logic of terminologies explicit provides a more rigorous comparative basis for the study of kinship terminology systems.

Introduction

All societies have terms that are used to address and to refer to one's kin (*see* Kinship in Anthropology). Address terms are the expressions used to address a person during social discourse that reflect the kinship relation between speaker and listener. Multiple terms can be used within a society for the same kinship relation, with the word form varying according to the aspect of the relationship between speaker and listener being identified by the address term. Using an address term presumes the parties involved already know the kinship relation between them. English speakers, for example, might address the woman who has the kinship relation of *mother* to speaker by mom, mommy, mother, mama, ma, mum, mummy, and so on, depending on the age of speaker and the emotional relationship between speaker and listener being foregrounded in their conversation.

Terms of reference, in contrast, overtly express the kinship relation understood to apply between the individuals of concern. Collectively, the terms of reference used by societal members form what anthropologists refer to as a kinship terminology. By a *kinship terminology* is meant, then, the terms used to refer to, or express, the corpus of culturally recognized kinship relations one individual can have to another individual in a particular society. Kinship terminologies differ from one society to another not only because of language differences, but, what is more important, due to cultural differences regarding the meaning, content, structure, and organization of the kinship relations expressed linguistically through the kin terms comprising a kinship terminology.

A kinship terminology is not just a nomenclature system for the kinship relations recognized in a particular society, as it expresses both the organization of, and conceptual interrelations among, the kinship relations making up a kinship terminology (Leaf and Read, 2012). The terminology linguistically expresses a society's indigenous theory of kin relations through the semantic content of kin terms and through the culturally understood conceptual relationships among the kinship terms. The terminology can thus be viewed as an idea system composed of the conceptual relations structurally linking the terms in a kinship terminology into a conceptually bounded system of kin relations. For this reason, formal studies of kinship aimed at elucidating the kinship ideas and concepts that are part of the cultural milieu of a society focus on the terms of reference making up a kinship terminology rather than the terms of address.

Consanguineal and Affinal Kinship Terms

Within a kinship terminology, there is generally a single term corresponding to each culturally recognized kinship relation, though two individuals may have more than a single kinship relation between them, as occurs, for example, when a male and a female who are already related as kin marry and thereby take on a new kinship relation as well. The corpus of kinship terms making up a kinship terminology is typically divided analytically into two parts: *consanguineal* kin terms based on the assumption that the primary kinship relations are determined through biological procreation (hence kin relations are said to be blood relations) and *affinal* kin terms that designate kinship relations determined through marriage. The division into consanguineal and affinal terms with procreation and marriage the basis for the contrast has, however, led to extensive controversy regarding what is meant culturally by kinship and kinship relations, and is not a satisfactory division. In many societies, kin relations can be established by criteria other than procreation and marriage, such as *sponsorship* (which includes adoption, a godparent relationship, and kinship through suckling (El Guindi, 2011)), a name giving-name receiving relationship, food sharing, or co-residence, among other means (Sahlins, 2013).

For English speakers, the kinship terminology includes the consanguineal terms father, mother, son, daughter, brother, sister, uncle, aunt, nephew, niece, and cousin, and the affinal terms include father-in-law, mother-in-law, son-in-law, daughter-in-law, brother-in-law, sister-in-law – the terms marked with the '-in-law' suffix. Some English kin terms, such as aunt and uncle, are both consanguineal ('my parent's siblings') and affinal ('my parent's sibling's spouse') and therefore said to be defined disjunctively, while the English kin term cousin does not have an affinal form and the English kin term spouse is only an affinal term.

Systematic Study of Kinship Terminologies

Kinship terminologies are universal: there is no society that does not have a set of reference terms used to express the kinship relations recognized in that society. Kinship terminologies vary in the number of terms from about 10 or so to around 30 terms and differ not only in the number of kinship relations that are recognized linguistically, but in the interrelationships among these kin terms. The systematic study of kinship terminologies traces back to the seminal work of Lewis Henry Morgan (1871), in the mid-nineteenth century, who carried out a worldwide study of structural differences among kinship terminologies by asking respondents for the kin terms that a terminology user would employ for each of a series of genealogical relations. The relations were aimed at determining the way the kin terms in a terminology refer to lineal and collateral genealogical relations defined with respect to speaker. Morgan assumed that kinship relations are determined through procreation and marriage, hence a kinship terminology represented a particular group's organization of the domain of genealogically defined kinship relations. Subsequently, anthropologists have generally assumed that kin terms are linguistic labels for categories of genealogical relations determined through procreation as it is culturally understood.

Morgan's systematic study of kinship terminologies began with his interest in the terminology of the Seneca (Iroquois) Indians, as their terminology made kinship distinctions lacking an obvious genealogical basis. For example, rather than having a single term referring just to genealogical mother (as happens with the English term *mother*), the Seneca term *no-yeh'* refers not only to genealogical mother but also to mother's mother's daughter, mother's mother's mother's daughter's daughter, and so on. Thus *no-yeh'* refers both to a lineal relation (genealogical mother) and to collateral relations (mother's mother's daughter, mother's mother's mother's daughter's daughter, and so on), contrary to the English and many other terminologies. (A sexually symmetric pattern occurs with the Seneca term *hä-nih'* used to refer to genealogical father, genealogical father's father's son, genealogical father's father's father's son's son, and so on.) In addition, the terms, *ah'je* and *ka-gä* used to refer to genealogical older sister and younger sister, respectively, were also used to refer to the genealogical daughter of any person referred to as *no-yeh'* or *hä-nih'*; that is, the terminology used the logic among the kin terms that a *ka-yä-wan-da* ('daughter') of a *no-yeh'* ('mother') or *hä-nih'* ('father') is an *ah'je* ('older sister') or *ka-gä* ('younger sister'), thereby expressing, without reference to genealogy, a consistently expressed conceptual relationship among the kin terms *ka-yä-wan-da*, *no-yeh'* or *hä-nih'* and *ah'je*/*ka-gä*. (A sexually symmetric pattern occurs with the kin terms for genealogical son and genealogical older and younger brother.) From the perspective of the kin terms, the category of genealogical relations associated with these kin terms derives from this logic expressing the relationship among the kin terms and not the reverse.

Classification of Kinship Terminologies

Most kinship theorists consider that genealogy is the foremost driver of kinship relations and so the terminology linguistically expresses their conceptual organization of genealogical relations, and so the terms are linguistic labels for a priori categories of genealogical relations. The Seneca pattern of including both lineal and collateral genealogical relations under the same kin term – a pattern that occurred in a number of other terminologies elicited by Morgan through his schedules – led him to refer to terminologies that systematically included collateral with lineal genealogical relations as *classificatory terminologies* and all other terminologies that distinguished lineal from collateral genealogical relations, such as is the case with the English terminology, as *descriptive terminologies*.

Early kinship theorists expanded on Morgan's two part typology, culminating in George Peter Murdock's (1949) extensive, comparative work that led to a six part typology still widely used today, based on shared genealogical features in ego's generation. Each of his six types is named for a society whose terminology is an exemplar of that type. Murdock divided the descriptive terminologies into Sudanese (also known as *bifurcate collateral* due to paternal and maternal genealogical cousins being terminologically distinguished) and Eskimo (also known as *lineal* due to all genealogical cousins having the same term and distinguished from genealogical siblings). The classificatory terminologies were divided into Hawaiian (also known as *generational* due to genealogical siblings and cousins not being distinguished), Iroquois (also known as *bifurcate merging* due to genealogical parallel cousins (see below) being distinguished from genealogical cross cousins (see below) and grouped with genealogical siblings), Crow (a bifurcate merging terminology in which the same term is used for genealogical father's sister and father's sister's daughter), and Omaha (a bifurcate merging terminology in which the same term is used for genealogical mother's brother and mother's brother's son). Subsequently, Iroquois has been divided into Iroquoian versus Dravidian terminologies (the latter used by Dravidian speakers in southern India) based on different ways in which genealogical parallel and cross relations are recognized in the terminology.

Parallel and *cross relations* refer to two patterns for the sex of sequential, genealogical positions in a genealogical pathway connecting ego and alter, with kinship relations having the same pattern being given similar sociological treatment by the users of the terminology. For example, many terminologies distinguish between a genealogical *parallel cousin*, who is the genealogical offspring of anyone referred to as a 'brother' of a 'father' or a 'sister' of a 'mother' (sometimes referred to, but incompletely, as 'the offspring of the same sex sibling of a parent') and a genealogical *cross cousin*, who is the genealogical offspring of anyone referred to as a 'brother' of a 'mother' or a 'sister' of a 'father' (sometimes referred to, but incompletely, as 'the offspring of the opposite sex sibling of a parent'). When the parallel/cross distinction is recognized in the terminology in the parental generation, parallel cousins are usually included terminologically with siblings, hence, from a sociological viewpoint, they are not marriageable under an incest taboo forbidding marriage with siblings, whereas cross cousins may be prescribed as the ideal choice for marriage partners.

Other terminology types have been distinguished when other patterns for differences among kinship terminologies have been recognized. While Murdock's typology has helped organize the variety of kinship terminologies, it is also widely

recognized as being inadequate as it assumes similarity in the categorization of genealogical relations in ego's generation implies similarity in the categorization of all other genealogical relations. Using Murdock's criteria, for example, the terminology of the !Kung San, a hunter – gatherer group in northwest Botswana, is classified as an Eskimo type of terminology along with the English terminology, yet except for having a single term for all genealogical cousins, the !Kung San terminology has little resemblance to the English kinship terminology.

Relationship between Terminologies and form of Social Organization

Part of the motivation for organizing the worldwide corpus of kinship terminologies was the assumption that kinship terminology variation would relate to variation in forms of social organization, especially differences in marriage systems. Evolutionary schemes for past systems of social organization were hypothesized, assuming that current kinship terminology distinctions reflected past forms of social organization and associated marriage rules (*see* Kinship, Evolution of). Superficially, bifurcate merging terminologies appeared to correlate with unilineal descent systems, at least when a society had a moiety form of social organization in which the society is divided into two exogamous descent groups. (A descent group consists of all persons who trace through genealogical parents of the same sex to a recognized ancestor when tracing is through genealogical fathers or to an ancestress when tracing is through genealogical mothers.) With descent groups, a man and his terminologically identified brothers are in the same descent group and similarly for a woman and her terminologically identified sisters. With a moiety system, a person's parallel cousins are in the same descent group as that person, hence are not marriageable, whereas that person's cross cousins are in the opposing moiety, hence marriageable. Thus there are parallels between a bifurcate merging (classificatory) terminology with a cousin marriage rule and a moiety form of social organization, but neither bifurcate merging terminologies nor cousin marriage rules are restricted to societies with a moiety form of social organization and each may occur in societies without organization as descent groups. As Rodney Needham concluded, the "assumption that societies with similar terminologies … are thereby sociologically similar" (1971, p. 22) does not hold up to ethnographic evidence.

Absent a connection between terminology and social system through genealogy, there is no need to assume genealogy is the basis of kinship relations. Numerous ethnographers have commented, for the group they are studying, that kinship and kinship relations are far broader and richer than what is implied under the assumption of kinship relations being created through procreation – even when taking into account the cultural construal of procreation – and then expressed through genealogy.

Genealogical Hypothesis and Kinship Terminologies

Early challenges to this genealogical assumption considered kin terms to identify social categories determined through commonality in social relations. The social category argument logically collapsed under Floyd Lounsbury's (1965) formal demonstration, using rewrite rules, of the genealogical logic underlying the Trobriand Islander's kinship terms, contrary to Edmund Leach's (1958) claim that their terminology made no sense from a genealogical perspective and could only be understood as identifying social categories that were part of their social system. (Rewrite rules (Lounsbury, 1964) specify, for the genealogical definition of a kin term, how the primary genealogical relation in that definition may be rewritten to obtain the other genealogical relations in the kin term definition.) Lounsbury's formalism, however, had a critical flaw; he had to assume that each of the kin terms in classificatory terminologies had a primary genealogical meaning that was then extended to other genealogical relations. This *extensionist hypothesis*, though, leads to circular arguments since the hypothesized extensions lack cultural grounding. The hypothesized extensions are neither recognized as such by the users of the terminology, nor are criteria given establishing when one society will make use of the hypothesized extensions in its kinship terminology and another will not. Why English speakers, for example, do not make use of the hypothesized extensions but the Trobriand Islanders do so, is not accounted for.

The genealogical assumption for kinship terminology systems was subsequently criticized extensively by David Schneider (1984) because, he argued, it reflected a western, folk science based on genitor viewed as physical father and genetrix as physical mother and not universal properties of kinship systems. Schneider's critique led to the extreme position of denying that kinship existed as a culturally salient analytical category, despite the fact that all societies make a fundamental social and conceptual distinction between kin and non-kin. What Schneider identified, however, was not the nonexistence of kinship, but of kinship universally derived from relations determined through procreation and expressed using biologically grounded genealogical criteria. Schneider recognized that kinship should be seen in a manner other than through the genealogical lens that had been used since the time of Morgan, but he neither identified what that different lens might be nor recognized that an alternative to assuming the primacy of genealogy for understanding kinship relations was being developed at the same time he was critiquing the assumption of a genealogical basis for the kinship relations expressed through the kin terms making up a kinship terminology.

Structural Logic of Kinship Terminologies

Missing from the genealogical assumption of kinship terminologies has been a way to account for the structural logic whose existence was made evident through the formalism of the descriptive system of rewrite rules discussed by Harold Scheffler and Floyd Lounsbury (1971). Their work made it evident that kin terms are logically interrelated, but failed to identify its underlying structural basis. In a paper published in the 1970s, Murray Leaf (1971) showed that it is possible to elicit a kinship terminology without making use of the genealogical method of elicitation developed by W.H.R. Rivers (1910[1968]) and used extensively in ethnographic research on kinship terminologies. Leaf showed how he was able to

elicit the kin terms of the Punjabi terminology through the kind of kin term calculations that have been widely reported in the ethnographic literature as the means by which users of kinship terminology systems compute kinship relations without reference to genealogy. Two persons can compute the kinship relation they each have to the other simply by each person knowing his or her kinship relation to a third person, even absent knowledge of the genealogical connections among these three persons and without needing to know the genealogical definitions of the kin terms.

The computation is essentially like the computation discussed earlier for the Seneca kin terms. More precisely, and illustrated here with the English kinship terminology, if person A does not know his kin relation to person C, but person A (properly) refers to person B as, say, *uncle* and person B (properly) refers to person C as, say, *daughter*, then from those two kin relations person A, as a user of the English kinship terminology, knows to (properly) refer to person C as *cousin*, without needing to know the genealogical relations involved, if any, among persons A, B, and C (C might be the adopted daughter of B), and without needing to know the genealogical definition of the English kin terms *uncle*, *daughter*, and *cousin*. We may refer to an indigenous calculation of this sort as a *kin term product* (Read, 1984) and write the equation: *daughter* of *uncle* is *cousin* (or more formally, *daughter* o *uncle* = *cousin*, where 'o' stands for a binary product operation). Beginning with the primary kin terms that relate to family relations, the kinship terminology may be elicited by first using kin term products of the primary terms to elicit nonprimary kin terms and then by using products of the primary terms with any elicited terms until no more terms are elicited. The result of the elicitation may be shown as a graph, referred to as a *kinship map* when the graph shows the outcome of the process of eliciting kinship relations, or as a *kin term map* when the graph shows the formal relations among the kin terms expressed through the kin term product equations obtained through the elicitation process (see Figure 1). The kinship map and the kin term map are isomorphic data models showing the structure of the kinship terminology that results from the structural logic embedded within the terminology. The rewrite rules developed by Lounsbury are a consequence of this structure, but the structure is not evident from the rewrite rules.

Generative Basis of Kinship Terminologies

By making explicit the structural form of a kinship terminology, we can then analytically determine its generative basis, if any, by determining, if possible, the generating kin terms for the structure and the structural equations underlying its form, with the requirement that the generating terms and structural equations have cultural saliency, as the goal is not a formal

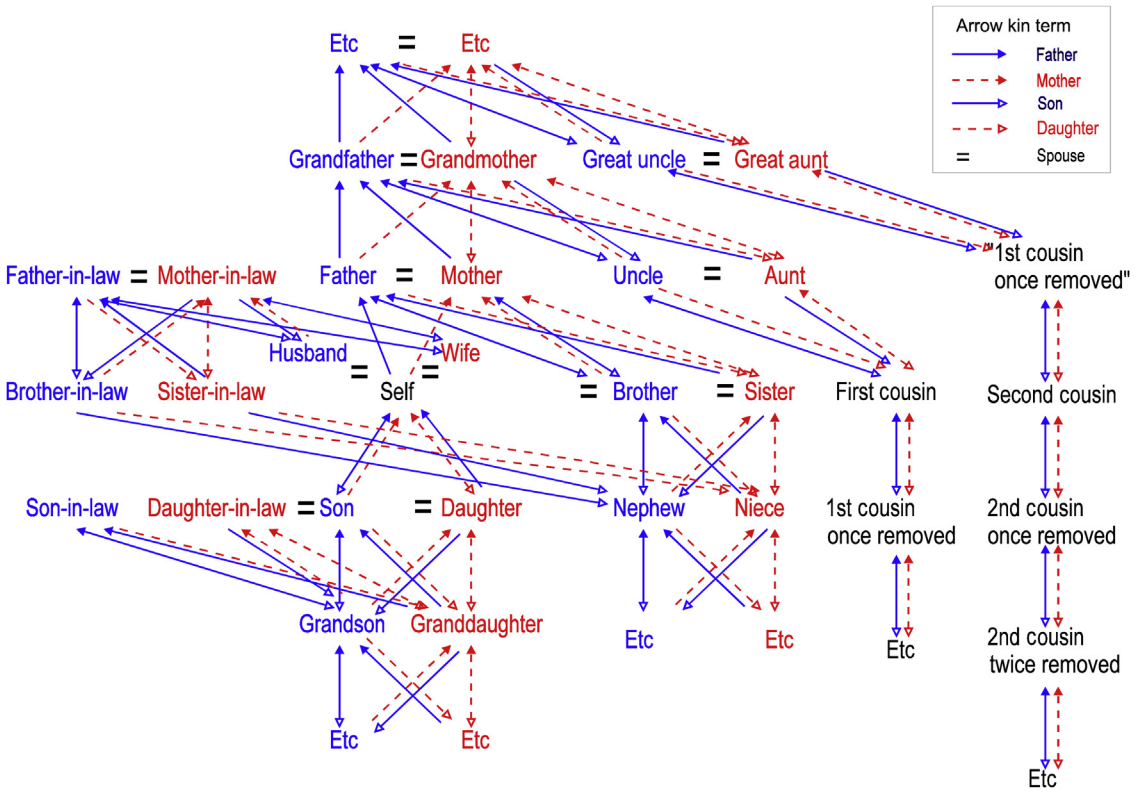

Figure 1 Kin term map for the American Kinship Terminology, based on the primary kin terms father, mother, son, daughter, and spouse. Male terms are blue, female terms are red, and neutral terms are black. Etc indicates that the kin term map continues in the same way without any structural changes.

representation, per se, but a formal account of the cultural kinship knowledge embedded within a kinship terminology (Read, 2007; *see* Formal Models of Kinship). Thus this analytical method is based on a falsifiable theory about a cultural domain, namely that kinship terminologies have a logically consistent and culturally salient structural form (Leaf and Read, 2012). Failure to find a culturally salient set of generating kin terms and structural equations for the structure shown in a kin term map would falsify the universality of the theory. To date, no counter examples have been found.

Structural Properties of Kinship Terminologies

The structural analysis of kinship terminologies has consistently found structural equations that are culturally meaningful and ethnographically justified. The analysis also makes evident the extent to which properties of a kinship terminology are determined by its internal logic versus considerations external to the terminology. For example, the distinction between descriptive and classificatory terminologies is structurally the consequence of which one of two ways that a sibling relation may be conceptualized is culturally salient: (1) a sibling is the *other child* of one's parents or (2) siblings have the *same parents*. The first concept implies that the kin term referring to the sibling relation is a derived kinship concept: *child* o *parent* = *sibling*, where the instantiation of a kin term product is understood to exclude the speaker. The second implies that the kin term referring to the sibling relation is a primary kin term and corresponds to societies in which sibling relations are central in the domain of kinship relations. The first concept also implies that sibling is not a generating term and a parent kin term (sex-marked in most terminologies, but not in the English terminology) is the primary (consanguineal) generating term. This leads to the descriptive terminologies. The second concept implies that sibling is also a generating term, which then leads to the classificatory terminologies. The second definition for a sibling relation also provides the logical basis for A.R. Radcliffe-Brown's (1950) *equivalence of siblings*, proposed by him as a way to account for the pattern of kinship relations for genealogical mother and genealogical father found in the Seneca and other classificatory terminologies. Radcliffe-Brown, however, did not account for when the equivalence of siblings was culturally salient; we now know that it is culturally salient when their concept of siblings corresponds to the second concept given earlier.

By working out the structural basis of kinship terminologies, we can distinguish the features that are the consequence of its generative logic, such as having kin terms that distinguish between older and younger sibling, or are due to distinctions embedded in the kinship terminology for social and cultural reasons, such as a distinction between older and younger mother's brother in the Tongan terminology that relates to patterns of inheritance (Bennardo and Read, 2007). In addition, rather than assuming features of terminologies are simply added to or removed from a terminology in an evolutionary account, change in features must be related to changes in the structural logic of the terminology (*see* Kinship, Evolution of). For example, terminologies in Australia distinguish between older and younger sibling regardless of the sex of speaker, whereas in the Polynesian terminologies the older/younger distinction generally depends on whether the sibling is a parallel (same sex) or a cross (opposite sex) sibling with regard to speaker. This difference is a consequence of differences in the structural logic of the terminologies from these two regions and so an evolutionary account linking these terminologies must be formulated at the level of changes in the structural logic of terminologies. Using a structural logic approach also makes it possible to form a typology of kinship terminologies based on differences in their respective sets of generating terms and structural equations (Read, 2013a), thereby clarifying what are structurally feasible evolutionary changes in kinship terminologies, as has been shown for the Polynesian terminologies (Read, 2013b). The evolutionary pathway of the Polynesian terminologies determined through their structural analysis is consistent with the work by historical linguists on changes in the word form of the kin terms, but requires different root terminologies for the Polynesian and Oceanic terminologies than the one posited in historical linguistic accounts.

Conclusions

Historically, the study of kinship terminologies began with the assumption that they are mainly names for genealogical categories whose origin is external to the terminology (*see* Family and Kinship, History of). Kinship relations were assumed to be determined through procreation and expressed genealogically. This approach left unspecified how the categories were determined and has been contradicted by ethnographic research that repeatedly finds that kinship ideas and kinship systems are not based in an obvious way on genealogy (which is not to say that kin terms do not relate to genealogy as it is always possible to provide genealogical definitions of kin terms in a post hoc manner). Reaction against the genealogical hypothesis led to the untenable assumption that kinship does not exist, rather than viewing kinship terminologies as cultural objects in their own right. The latter makes it possible to identify the structure embedded in a kinship terminology by identifying the way the kin terms are interrelated through the culturally salient kin term product. This, in turn, has led to analytical determination of the generative basis of the structure of a kinship terminology, thereby bringing the formal analysis of kinship terminologies back to the ethnographic study of kinship systems and the meaning of kinship relations. The analytical approach that has thereby developed makes possible a vibrant and deep-seated understanding not only of kinship terminologies as the expression of the kinship ideas of a particular society, but more broadly of what is meant by culture as a constructed reality, for the kinship terminologies of human societies are cultural constructs through which a kinship domain is formulated that is not the codification of already existing patterns of behavior. The analytical study of kinship terminologies now has the potential of making evident the interrelationship between kinship and social systems in a sound and rigorous manner.

See also: Family and Kinship, History of; Kinship in Anthropology; Kinship, Evolution of; Kinship, Formal Models of.

Bibliography

Bennardo, G., Read, D., 2007. Cognition, algebra, and culture in the Tongan kinship terminology. Journal of Cognition 7, 49–88.

El Guindi, F., 2011. Milk and blood: kinship among Muslim Arabs in Qatar. Anthropos 107, 545–555.

Leach, E., 1958. Concerning Trobriand clans and the kinship category "tabu". In: Goody, J. (Ed.), The Developmental Cycle in Domestic Groups. Cambridge University Press, Cambridge.

Leaf, M., 1971. The Punjabi kinship terminology as a semantic system. American Anthropologist 73, 545–554.

Leaf, M., Read, D., 2012. The Conceptual Foundation of Human Society and Thought: Anthropology on a New Plane. Lexington Books, Lanham.

Lounsbury, F., 1964. A formal account of Crow- and Omaha-type kinship terminologies. In: Goodenough, W. (Ed.), Explorations in Cultural Anthropology. McGraw Hill, New York.

Lounsbury, F., 1965. Another view of the Trobriand kinship categories. American Anthropologist 67, 142–185.

Morgan, L.H., 1871. Systems of Consanguinity and Affinity in the Human Family. The Smithsonian Institute, Washington, DC.

Murdock, G.P., 1949. Social Structure. Macmillan, New York.

Needham, R., 1971. Remarks on the analysis of kinship and marriage. In: Needham, R. (Ed.), Rethinking Kinship and Marriage. Tavistock, London.

Radcliffe-Brown, A.R., 1950. Introduction. In: Radcliffe-Brown, A.R., Forde, D. (Eds.), African Systems of Kinship and Marriage. Oxford University Press, Oxford.

Read, D., 1984. An algebraic account of the American kinship terminology. Current Anthropology 25, 417–449.

Read, D., 2007. Kinship theory: a paradigm shift. Ethnology 46, 329–364.

Read, D., 2013a. A new approach to forming a typology of kinship terminology systems: from Morgan and Murdock to the present. Structure and Dynamics 6. http://www.escholarship.org/uc/item/0ss6j8sh.

Read, D., 2013b. Reconstructing the proto-Polynesian terminology: kinship terminologies as evolving logical structures. In: Mcconvell, P., Keen, I., Hendery, R. (Eds.), Kinship Systems: Change and Reconstruction. University of Utah Press, Salt Lake City.

Rivers, W.H.R., 1910(1968). The genealogical method of anthropological enquiry. In: Rivers, W.H.R. (Ed.), Kinship and Social Organization. The Athlone Press, New York.

Sahlins, M., 2013. What Kinship Is … and Is Not. University of Chicago Press, Chicago.

Scheffler, H.W., Lounsbury, F., 1971. A Study in Structural Semantics: The Siriono Kinship System. Prentice-Hall, Englewood Cliffs.

Schneider, D., 1984. A Critique of the Study of Kinship. University of Michigan Press, Ann Arbor.

Klein, Melanie (1882–1960)

Nancy J Chodorow, University of California, Berkeley, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 8111–8113, © 2001, Elsevier Ltd.

Abstract

This article provides a short biography of the psychoanalyst who has had, after Freud, more impact than any other. It traces her moves from Vienna to Budapest to Berlin and, finally, to London, where she spent most of her professional life. Klein was a founder of one branch of child analysis. She took seriously Freud's hypothesis of the death drives. She also developed an original developmental theory, in which the child's major psychic task is the management of anxiety arising from destructive fantasies and wishes and from virulent envy. In the first, paranoid-schizoid, position, persecutory anxiety is prominent and splitting the major defense. Later, depressive anxiety and the fear of harming the loved object accompany the depressive position. Klein theorized the internal world, and the processes of splitting, projection, and introjection that create that world. Throughout life these are the challenges facing any human being as well as shaping and undermining social formations and processes. The central datum of Kleinian theory is unconscious fantasy.

Within psychoanalysis worldwide, Melanie Klein has had more impact than any other psychoanalyst besides Freud. Her wide-ranging contributions not only reshaped psychoanalytic understandings of mental life; they also have fundamental import for social scientific understandings of human action and the creation of meaning. That Klein is indisputably the most important psychoanalyst after Freud also points to the unusual prominence and participation of women in psychoanalysis from almost the beginning (see Chodorow, 2000).

Born in Vienna on 30 March 1882, Melanie Reizes Klein moved to Budapest in 1910, where she began to read Freud. She entered analysis and analytic training under Freud's colleague Sandor Ferenczi (on Klein's biography, see Grosskurth, 1986; Segal, 1979). She moved to Berlin in 1921, joining the Berlin Psychoanalytic Society and entering analysis with Karl Abraham. Her first formal work as a child analyst began in Berlin. At the invitation of Ernest Jones, Klein visited London in 1925 to lecture on child analysis, and she subsequently moved to England in 1926. Until her death in London on 22 September 1960, she was an active clinician, a leading and innovative child analyst, and a prolific writer who had many followers and colleagues.

Understanding Melanie Klein's life requires an understanding of the history of psychoanalysis. Klein had no direct contact with Freud, but was trained and supported by Ferenczi, Abraham, and Jones, the leaders of the main psychoanalytic centers outside of Vienna and Freud's most important colleagues. These several centers allowed more possibility for independence of thought and action than Vienna itself. While Klein was pioneering the practice and theory of child analysis in Berlin and London in the 1920s, Anna Freud was engaged in a parallel project, based on somewhat different understandings and methods, in Vienna. After Freud developed cancer in 1923, he came to depend almost entirely on his daughter as both professional heiress and personal companion and caretaker. Klein's theory, technique, and stature as a child analyst directly challenged Anna Freud's primacy in this same arena, and Freud was drawn to support his daughter. Thus, even as Klein considered herself the true theoretical heir to Freud, her work from the outset was strongly contested by Anna Freud and her father.

As a result of the Anschluss of March 1938, in which Hitler, with the approval of many Austrians, seized and incorporated Austria into Nazi Germany, the Freuds moved to London, thereby setting the stage for a confrontation and theoretical and organizational sharpening of differences. Freud died in 1939, but the followers of Anna Freud and those of Melanie Klein in London spent the latter war years in 'Controversial Discussions' (King and Steiner, 1991), with an 'Independent' Group holding a middle ground and developing its own theories. Debate posed Klein's espousal of Freud's late instinct theory, with its focus on aggression, her theories of early infant development, her advocacy of the analysis of very young children, and her insistence that the analysis of children (and adults) should begin immediately from interpretation of unconscious fantasy, against Anna Freud's focus on ego defenses and ego functions, her following of the traditional structural and libidinal phase theories, and her advocacy of analyzing from the most surface layer of defense rather than from the deepest layer of unconscious fantasy. The result of the 'Controversial Discussions' was a settling in of a Kleinian Group, led by Klein until her death and still an active force in British psychoanalysis (Spillius, 1988 includes writings on theory and technique by Klein's students and later colleagues).

Since the 1940s, no psychoanalytic center has developed without grappling with Klein. Kleinians continue to be central to British psychoanalysis, and Kleinian thinking has shaped psychoanalysis throughout Latin America. Kleinian ideas were initially anathema in the USA, where emigre analysts were closely allied with Anna Freud, but since the 1980s, Kleinian thinking has become widely influential. In Europe too, psychoanalytic theory grapples with, accepts, or challenges Klein.

Kleinian theory is a complex, intertwined whole resting on several foundations (see Klein, 1975). This account focuses on Kleinian theory rather than on Klein's founding contributions to child analysis and psychoanalytic clinical technique. As a psychoanalyst, Klein like Freud begins from the discovery and investigation of the dynamic unconscious, the postulate that conscious thinking only expresses a small fraction of what goes on in the human mind. The beginning datum of Kleinian theory is the psychic reality, or internal reality, of what Klein

called 'unconscious fantasy.' The dynamic unconscious consists of repressed or split off fantasies that have a powerful effect in shaping conscious life, feeling, and behavior. Unconscious fantasy operates from the very beginning of life and consists of affect-laden stories or proto-stories about the self, the body, and one's relation to others. These stories are not linguistic or narrative, but are forms of implicit knowing and being that portray in direct, unmediated, affectively charged ways, a mental picture of an inner world of self and others. Klein described how this internal reality is created and its constituent effects on all bodily experience, interpersonal relations, and social and cultural experience (for an extended reading of Klein along these lines, see Chodorow, 1999b). It is an 'object-relations theory,' picturing the self, or ego, in relation to others, or objects. Klein's method of play analysis and work with very young children led her to a 'developmental theory' based on changes in prevalent unconscious fantasies. The management of anxiety about 'aggression' in self and others is a central goal of most unconscious fantasy.

Although Freud discussed the ego's relations to its objects and the effects of the object upon the ego – for instance, in his claim that the ego contains a history of object choices – Klein is considered the original object-relations theorist. For Klein, the ego is object-related from birth, all drives are directed toward objects rather than toward gratification and release, and all psychological life revolves around internal pictures of the ego in relation to its objects. Objects are both external – actual people like mother, father, or analyst with whom the self interacts – and internal, parts of the outside world that are unconsciously experienced as within the self. Objects in this 'internal world' may be 'part-objects' – an aspect of another, a body part of another, like the mother's breast or a fantasy of her inside organs – or 'whole objects' – others in all their different aspects. Psychic life especially consists in the building of an internal world through what the ego does psychologically with objects, using such mechanisms as 'splitting,' 'projection,' and 'introjection.' Through these processes, internal and external realities are continuously recreated in relation to one another.

Whereas for Freud, repression of unacceptable thoughts and feelings is the key defense, for Klein, splitting, introjection, and projection are the central defenses. Klein's developmental account revolves around the uses of these defensive processes. In the 'paranoid-schizoid position,' the self and the earliest internal objects – initially the mother's breast, and throughout life, objects that represent the breast – are psychically experienced as either all good or all bad. The ego's goals are to keep bad and good parts of the self and object separate. The ego attempts to preserve its own goodness by keeping good, or loved, objects within, and ejecting and rejecting bad, or hated, objects. 'Projective identification' then leads the ego to feel that the object, now containing bad parts of the self, is persecutory. To control the now persecutory objects, they may be reintrojected, in turn bringing hatred back into the self, which projects badness back into the object. And so on. Split-off good feelings and parts of the self may also be defensively projected into the object, so that the self can feel protected and loved. 'Persecutory anxiety' – the fear of being attacked by bad objects or of annihilation of the self – is the dominant anxiety of the paranoid-schizoid position and leads to the spiral of splitting, projection, and introjection that characterize it. The paranoid-schizoid position is paranoid because of the fear of attack it contains and schizoid because it is based on separating different aspects of the self and the other.

When the child reaches the 'depressive position,' he or she recognizes people as whole objects and painfully realizes that an attack on the bad object is also an attack on the good one, since they are one and the same. Guilt arises, along with the desire to make 'reparation.' Reparation is one root of creativity throughout life as well as of the desire to parent and take care of others. 'Depressive anxiety' – the fear that one will damage the good object – is the characteristic anxiety of the depressive position. Unlike Freud's libidinal stage theory, Klein's theory of positions is not a stage theory in which the achievement of the depressive position involves the giving up of the paranoid-schizoid. All people who have reached the depressive position continue throughout life to move between the two positions.

A complex inner world of self and objects constitutes unconscious fantasy and is the basis of 'transference.' For Klein, transference is not just the putting of early feelings about a parent onto the analyst but a total way of being in the world. Although it is mainly investigated in analysis, transference gives meaning to and animates all experience. It functions as an internal prism through which the world is perceived, becomes subjectively meaningful, and is an object of action. Klein, and especially later Kleinians, also show how forms of thinking themselves transferentially express psychic structure and psychic fantasy, for example, whether a patient can make connections among thoughts, can keep anxiety-laden ideas in mind or must split them off and evacuate them, can keep ideas whole or must destroy or fragment them, can take in ideas from the analyst, can symbolize generatively, and other similar processes.

For Klein, the most anxiety-laden and intolerable unconscious fantasies concern aggression. Klein holds to the later Freudian dual-drive theory, rejected by many analysts, of the life and death instincts. The death instinct constitutes a threat of annihilation or fragmentation from within, creating basic existential anxiety from the beginning of life. The basic wish of the psyche is to get rid of this threat, leading to the projection that creates the first bad object. Persecutory anxiety thus arises originally as a direct response to the internal threat of destruction from within and thereafter also in response to the aggression that has been through projective identification put into objects. The paranoid-schizoid and depressive positions are both organized largely in response to anxieties generated by these fears of annihilation, aggression, and hatred that emerge from self and object.

A major manifestation of the death instinct is found in envy, which Klein describes as the most toxic emotion. Klein discovered the importance of envy in patients' attempts to destroy and dismantle the help offered by the analyst. Envy originates in the perception that the other possesses and enjoys something good and desirable – originally imaged as the creativity, fullness, and bounty of the maternal breast. The envious impulse is to spoil or destroy that goodness. Envy is felt especially virulently against those who appear not to be envious, those who can enjoy creativity and happiness in others. The envious cannot feel gratitude or serenity, and in fact

envy is perhaps the least tolerable of emotions, because its goal is to destroy or spoil not that which is bad, but that which is good. Envy is manifested in attempts to destroy the offerings of parents, of teachers, of analysts, of any whom seem contented and happy. In literature, Iago represents envy in its purest form.

Although she did not explore these implications herself, Klein's work has profound implications for social science (see e.g., Chodorow, 1999b). Klein's account of unconscious fantasy and transference challenges the sociological assumption that the material or structural reality of politics, society, and culture come first and create or determine human life. By documenting how people create the subjective meaning of the world from within, through the dynamics of unconscious fantasy, and how perception, action, and thought result from a transferential shaping that has real effects, Kleinian theory demonstrates that action in the external world is not only taken through the conscious, rational assessment of options, nor is it determined through structural or material constraints. Action is as much a result of an internal, psychologically created picture as of society or culture, and meanings are as much created psychodynamically through affectively charged fantasy as they are culturally presented.

Splitting, projection, and introjection are psychological mechanisms that characterize many social processes. Many of the major political events, crises, and controversies of our time can only be fully understood if the operation of rage, splitting, projection, and introjection is taken into account (see Chodorow, 1998). For example, political and cultural demonization of the enemy is psychologically based on projection and introjection, in which all good resides in the nation, the ethnos, the group, or the political party, and all bad resides in those who are not part of this group. Splitting bad from good, and containing all good within the self while putting all bad into the other, underpins all virulent racism, nationalism, ethnic conflict, and attempts at ethnic cleansing or genocide. Splitting may also be used to ward off elements in the self that are threatening because they are too attractive. Thus, misogyny and homophobia express splitting, projection, and denial, in this case of misogynist and homophobic men's feminine identifications or desires for other men – attractions and desires that are too anxiety-laden to be contained within the self. Finally, Klein describes the great envy of women's creative insides and the psychic threat of images of the mother who contains everything good and powerful – babies, milk, the father's penis, and the ability to give and take life. Such images of maternal power and the mother's body help to explain the psychological intensity of abortion politics and perhaps the extreme violence of a few anti-abortion activists (for a brief overview of Kleinian theory in relation to gender, see Chodorow, 1999a).

Kleinian theory is therefore a basis both of a theory of psychoanalytic treatment and a description of how the human mind develops and operates. It describes how people create themselves and give the world subjective meaning. It explains motivations for action and perception. It is thus an essential component of all investigations of society, politics, and culture.

Bibliography

Chodorow, N.J., 1998. The enemy outside: thoughts on the psychodynamics of extreme violence with special attention to men and masculinity. Journal for the Psychoanalysis of Culture and Society 3 (1), 25–38.

Chodorow, N.J., 1999a. From subjectivity in general to subjective gender in particular. introductory essay to Melanie Klein's mourning and its relation to manic-depressive states. In: Bassin, D. (Ed.), Female Sexuality: Contemporary Engagements. Jason Aronson, Northvale, NJ, 241–250.

Chodorow, N.J., 1999b. The Power of Feelings. Yale University Press, New Haven, CT.

Chodorow, N.J., 2000. Psychoanalysis and women psychoanalysts. In: de Mijolla-Mellor, S. (Ed.), Women in the History of Psychoanalysis. Karnac Books, London.

Grosskurth, P., 1986. Melanie Klein: Her World and Her Work. Knopf, New York.

King, P., Steiner, R. (Eds.), 1991. The Freud–Klein Controversies, 1941–45. Routledge, London and New York.

Klein, M., 1975. The Writings of Melanie Klein, vols. 1–4. The Hogarth Press, London.

Segal, H., 1979. Klein. William Collins Sons and Co, Glasgow, UK.

Spillius, E.B. (Ed.), 1988. Melanie Klein Today: Developments in Theory and Practice, vols. 1 and 2. Routledge, London and New York.

Knowledge, Anthropology of

Dale F Eickelman, Dartmouth College, Hanover, NH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The anthropology of knowledge concerns the categories and organizing concepts through which people make sense of, and act on, their social and natural environments. It is the study of the connections between social organization and content in cultural traditions – a key problem in social thought since the emergence of the social sciences in the nineteenth century. Since the late twentieth century, the complex links to cognition and brain science have further enriched explanations of how knowledge is perceived and communicated.

The anthropology of knowledge concerns the categories and organizing concepts through which people make sense of, and act on, their social and natural environments. It is the study of the connections between social organization and content in cultural traditions – a key problem in social thought since the emergence of the social sciences in the nineteenth century. Since the late twentieth century, the complex links to cognition and brain science have further enriched explanations of how knowledge is perceived and communicated.

With the overall framework of the anthropology of knowledge, some scholars have been concerned primarily with the ethnographic analysis of particular social and cultural groups, the relationship between their beliefs and social structure, and how some bodies are made to appear more central and legitimate than others – as in colonial and authoritarian settings.

Beginning in the 1970s, the anthropology of knowledge increasingly included a concern with how culturally valued cognitive styles – sets of basic, implicit master patterns of language and thought on the basis of which other patterns of knowledge are subsequently acquired – were defined and transmitted. It also focused on how these patterns of knowledge shape, and in turn how they are legitimated by, configurations of power and economic relations among groups and classes in different societies.

Since the late twentieth century, emphasis has shifted from the anthropological study of classification and categorization within cultural traditions – and debates over what is common to humankind and other species – to the relationship between the means of communication and the content of cultural traditions and to how socially valued knowledge is produced and made to seem legitimate and natural to some but is imposed on others – as in colonial settings. This 'thinking about knowing,' as the anthropology of knowledge has recently been called, necessarily engages anthropologists to collaborate closely with brain scientists, biologists, and neuroscientists in addition to historians and philosophers (Marchand, 2010). Memory and forgetting does not work the same way for groups as it does for individuals; groups make memories through food, body sculpture, landscapes, rituals, and monuments (Assmann, 2011).

First Approaches: The Late Nineteenth and Early Twentieth Centuries

Most nineteenth century formulations of the problem of knowledge focused on what entire social groups shared, both those depicted as intensely local and embodied and those shared beyond the limits of particular societies. Thus, representations of aboriginal Australian knowledge depend on seeing landscapes firsthand and on adults who have reached a certain status in their life cycle in which they can impart to others how features of the landscape were created during the Dreaming by the "transformational acts of creative beings" (Tonkinson, 2011: p. 331). Other forms of knowledge, such as geometry and mathematics, were seen as more abstract, not tied to specific peoples and places and more advanced. Earlier formulations of the anthropology of knowledge also implicitly or explicitly assumed an evolutionary schema of social development from elementary to more complex societies, a notion as common to Charles Darwin as it was to Karl Marx. Thus William Robertson Smith (1972), a Biblical scholar whose work profoundly influenced later sociologists and anthropologists, including Émile Durkheim and E.E. Evans-Pritchard, saw a close relationship between the stage of development of a social group and the nature of its intellectual, religious, and moral life. For example, Smith argued that to understand the message of the prophet Ezekiel, the social organization of his time had to be delineated, together with the cosmology and modes of thought contemporary with it. If Ezekiel's prophetic message was found to be different from that of Samuel, Smith argued, it is not because of any contradiction in the Bible but because Samuel addressed himself to groups and situations at different stages of human development.

This dual sense of evolution in history and in the complexity of social forms was basic to Durkheim's *Elementary Forms of the Religious Life* (1915) and his *Primitive Classification* (1963), coauthored with his nephew, Marcel Mauss. Durkheim's goal was to study the most elementary societal forms in order to discern the basic social elements out of which ideas of time, space, causality, category, and person are constructed. Using the example of Australian aborigines, which he considered the best-documented elementary society, Durkheim argued that religious thought, like all forms of thought, derive from collective social action.

Durkheim sought to synthesize the two major prior approaches to categories of thought, the apriorist and the empiricist. The apriorist approach, exemplified by Immanuel Kant, viewed categories of thought as immanent in the human mind. The empiricist approach, exemplified by John Locke, saw these categories of thought as derived from experience, with individuals independently constructing similar perspectives. Mediating between these two approaches, Durkheim argued that categories of understanding were neither innate nor derived from

individual experiences. Instead, they were the product of a collective consciousness created and sustained through shared ritual action. This social basis of knowledge and categorization applied as much to the classification of objects and the natural world as it did to religious thought, at least in elementary societies. This is why the embodied knowledge, as represented by Australian aborigines – connecting landscapes to the Dreaming, to rituals, to perceptions of time and place, and to the relationships among individuals and groups – merited special attention.

For Durkheim, the notion of the collective consciousness (*conscience collective*) joined assumptions of the fundamental faculties of the human mind with the creation of collective representations. Durkheim's argument was elegant, but in synthesizing the main currents of European social philosophy, he was unable to demonstrate the primacy of the social in causing forms of classification. At the least, his argument suggested an analogy between the social order and forms of knowledge, and his psychological assumptions concerning collective representations were quickly superseded. Nonetheless, his work profoundly influenced successive generations of anthropologists interested in the ethnographic representation of cosmologies and religious beliefs and their relation to the social order.

The Ethnography of Knowledge

A classic example of how the work of Robertson Smith and Durkheim influenced ethnographic writing in the first half of the twentieth century is E.E. Evans-Pritchard's study of the Nuer of the southern Sudan (1940). Evans-Pritchard cites neither of these predecessors in his text, although their influence is evident in the structure of his argument. He depicted the Nuer as a cattle-herding people with a simple material life and minimal social and political differentiation, highly dependent on their local environment and seasonal migrations. Evans-Pritchard described these in detail, indicating how ecological conditions shaped Nuer social life and cosmology, or ideas of religion, causality, and the order of things. He showed how the structure of Nuer lineages, age-grades, and political units corresponded to their ideas of time, space, and political relationships.

Like Durkheim, Evans-Pritchard allowed readers to infer that Nuer categories and classifications derived from the structure of their social and political life. As with Durkheim, however, all Evans-Pritchard demonstrated was common schemas or homologies of form among a traditional social structure, ideas of time and space, and cosmology. The anthropological tradition represented by Evans-Pritchard remained dominant through the mid-twentieth century, inspiring other work on cosmology and social structure in other African societies, including John Middleton on the Lugbara of Uganda and Godfrey Lienhardt on the Dinka of the southern Sudan.

The roots of Durkheim's influence continued well beyond mid-century and could be seen in the work of anthropologists such as Edmund Leach and Mary Douglas (1966). Such studies suggest the extent to which people go to maintain or distinguish conceptual kinds and boundaries among animals, plants, genders, foods, peoples, and types of activity. Claude Lévi-Strauss (1966) also focused on the forms of classification implicit in both the kinship structures and in the myths of elementary societies, exploring similarities across cultural traditions that suggested common structures in myths, for example, akin to those studied by structural linguists. Lévi-Strauss's work encouraged a focus on how analogies are made between different cultural domains.

Some anthropologists such as Maurice Godelier (1986) sought to join Durkheimian concerns with categories and classification to a Marxist-inspired linking of ideas of knowledge to modes of production. As with Durkheim's work, however, the emphasis in these studies remained primarily in the classificatory schema implicit in systems of knowledge and in analogies between symbolic representations and material/economic life rather than in how people created, imposed, and altered ways of knowing.

Psychology, Brain Science, and Anthropology

Until the 1920s, the lines between psychology and anthropology significantly overlapped. It was often psychologists more than anthropologists who engaged in comparative work across societies and explored the extent to which humans thought alike. Some anthropologists argued that human intellectual and cognitive processes were essentially similar and varied only with the level of evolution of their culture. Others, such as the French philosopher Lucien Lévy-Bruhl (1857–1939), argued that there was a great divide between the prerational and primitive forms of thought and the rational thought of modern societies. Lévy-Bruhl's (1922) argument had a strong resonance for the French colonial lobby because, with minor adaptations, it even offered to explain the presumed inability of the Muslim societies of North Africa to abandon social forms and beliefs supposedly nonadaptive to the colonial and modern world rather than a form of resistance irrelevant to the presumed great divide between traditional and modern.

After the 1920s, however, there was an increasing disciplinary separation between psychological and anthropological approaches to cognition and knowledge. Psychologists, more than anthropologists, explored issues of cognition and the capacities and processes by which individuals learn, classify, and make associations and analogies. Psychologists thus focused on how the brain is hard-wired for certain thought processes, including the ability to learn languages. Anthropologists dealt more with culturally learned forms of knowledge and classification. Only a few anthropologists such as Gregory Bateson (1972) continued to envision a common ground between the two disciplines.

As Dan Sperber (1985) noted, anthropological work always had two dimensions. Some anthropologists have been more interested in the interpretation and analysis of particular cultural and social groups. Others have been more concerned with the cultural abilities and dispositions of humankind in general. Studies in cognitive anthropology, as the latter focus became known, underscored this distinction. These were the anthropologists most inclined to explore the link with psychology. The 1960s again brought scholars from the two disciplines into close collaboration. In one key study, for example, color perceptions, a domain that anthropologists previously assumed to be culturally transmitted, became recognized as having a more limited range of variation than was initially supposed, suggesting underlying cognitive principles (Berlin and Kay, 1969).

The focus of the Durkheimian tradition of the anthropology of knowledge was always in the outer frame of knowledge – how classificatory schemes for understanding society and the world related to the social order. The work of the cognitive anthropologists studied the order internal to particular schemes of knowledge – such as botany, color, disease types, firewood, and kinship and relationship systems. Michael Cole (1996) called this renewed collaboration between anthropologists and psychologists the second psychology, in which the two disciplines explored the frontier between culture and mind in understanding cognitive development; the acquisition of literacy and numeracy (Wagner, 1993); and language acquisition. In this vein, Dorothy Holland and Naomi Quinn (1987), for example, use the term *schema* to refer to standard learned patterns of activity such as going to the post office by which members of a society learn cultural patterns and habits to accomplish routine tasks, freeing individuals to focus instead on less predictable situations.

Intellectuals

Another component of the anthropology of knowledge is the role of intellectuals in society and the social production of knowledge. Until recent decades, anthropologists were concerned primarily with the bodies of cultural knowledge shared by all members of a society. Anthropologists such as Paul Radin (1957) in American anthropology and Marcel Griaule (1965) in the French tradition portrayed intellectuals as possessing an especially intense awareness of the sacred center of their society and an ability to reflect and explain its categories of knowledge. A classic essay by Victor Turner (1967) about the knowledge and character of one of his principal informants on Ndembu symbolism, comparing him to an Oxford don, also broke from an earlier tradition in which individual producers of knowledge were subsumed in larger collectivities.

An influential approach to defining intellectuals in relation to theories of modernization and modernity was Edward Shils (1972). For him, intellectuals are those imputed with the innate need to understand and articulate ideas and concepts that are more general and profound than those faced in practical, everyday concerns. For Shils, all intellectual work originally arose from religious occupations, but religious orientations in modern times have attracted a diminishing share of the creative capacities of an elite oriented toward more modernconcerns. Shils's approach assumed that not all knowledge was equally shared, and it also encouraged anthropologists and other scholars to think of various types of intellectuals and their relationships with the different political and social contexts in which knowledge is produced and valued. Whether those people were called intellectuals or entrepreneurs, anthropologists shifted their focus away from discerning classificatory categories and cosmologies and toward the economic and political contexts in which knowledge is produced and shaped.

In this shift to understanding knowledge and its social distribution as process and as historically rooted, the strategies used for comprehending elementary societies – to invoke Durkheim's term – were adapted to depict ideas of knowledge in advanced, complex societies. Thus Pierre Bourdieu (1967) compared the cognitive style implicitly learned at the Sorbonne to that transmitted by Bororo elders in the Brazilian Amazon. For Bourdieu, the verbal maneuvers learned by students preparing for the *leçon* at the Sorbonne furnished a model for the 'right' mode of intellectual activity in the French context as much as the dualistic spatial layout of Bororo villages embodied Bororo cosmology as interpreted by dominant elders. The dualistic method of the *leçon*, which establishes a tradition of two views on any subject, is subsequently applied to discussion of a wide range of intellectual problems, to the exclusion of alternative, less culturally valued approaches. But Bourdieu's analogy between the spatial plan of Bororo villages in the Brazilian Amazon region and the cognitive styles implicit in French and German higher learning also contributed to erasing the former assumptions concerning how people think in elementary and advanced societies, with all the assumptions implicit in such terms. Since the 1990s there has been much more emphasis on the shared factors of grounded cognition – an awareness that "the bodily, cognitive, neural, and social mechanisms that permit and constrain knowledge transmission are conjointly operative and mutually contingent" (Cohen, 2010: p. S194). Academic disciplines differ in how much emphasis they place on these different factors, but increasingly they recognize the complex tasks involved in learning to listen, see, weave, perform, and communicate. As Yehuda Elkhana (1981: 27–28) noted more than three decades ago, the notion of a great divide between Western and non-Western, traditional and modern, prescientific and scientific, and literate and nonliterate depended on a distinction between the content and processes of thinking. This distinction is less clear, even in terms of thinking about the European scientific revolution of the seventeenth century. As Elkana notes, one mode of thought was not totally obliterated by the other, and modes of thought have many constituent elements that "shift over time and vary among different cultures."

Knowledge and Power

Bourdieu's work on education, social class, and knowledge in the 1960s linked together several long-standing themes in the anthropology of knowledge but also indicated a shift from earlier concerns with categorization to the practice of creating and distributing knowledge. Bourdieu (1977), like Maurice Bloch (1985), Richard Fardon (1985), Maurice Godelier (1986), and others related knowledge and social classifications to the distribution and control of resources. In the specific case of education in France, Bourdieu depicted how the style and forms of knowledge of French higher education converted social classifications into school classifications, relating mastery of the symbolic classifications and habits of thought of formal schooling to both class and social capital.

The theme of the historical problem of change in dominant ideologies of higher education and how these ideologies were related to the conflicting interests of class and social groups in complex societies was raised at the beginning of the century. Durkheim's (1977) neglected lectures on the evolution of higher education in France are a case in point. It was only by the 1970s, however, that such ideas rose to the forefront of anthropological thought. Studies such as Bourdieu's on higher education in France and others on themes such as the internal logic and changing social context of Islamic ideas of knowledge (Eickelman, 1985) and the process by which

dominant ideas of religion and tradition shifted in Indonesia (Hefner, 1985) joined the long-standing anthropological concern with tacit knowledge in religious experience with how social actors derive meaning from public symbolism and continually shape and renew bodies of knowledge in differing social and historical circumstances.

Fredrik Barth's (1987, 2002) accounts of local variations in the cultural traditions of the Ok of highland New Guinea, like recent multidisciplinary studies of the relationships among minds, bodies, and knowledge, erode the earlier distinction between elementary and complex societies and between implicit and embodied knowledge and explicit knowledge. The Ok use ritual occasions to create and alter systems of knowledge as much as seminars and innovative lectures accomplish the same objective in contemporary Western societies. Departing from earlier accounts such as those of Evans-Pritchard about the Nuer, Barth emphasized the processes of thought and innovation at work within a system of knowledge, producing multiple variations and developments.

Approaches such as Barth's and those of Tonkinson (2011) concerning Australian aborigines also have profound implications for the idea of culture and knowledge. Rather than viewing cultures and bodies of knowledge as implicitly and internally homogeneous, shared, and sharply bounded, they are instead conceived as existing in a perpetually interactive social organization of communication in which individuals reproduce and alter representations and symbols.

Current trends in the anthropology of knowledge thus encompass two related concerns. One approach builds on new collaborations among anthropology, psychology, geography, and other disciplines, exploring new ways of relating public and institutionalized cultural expression to cognition and mental representations (Shore, 1996). The other approach focuses on the social production of knowledge and its relation to power, value, communications, and control. The earlier concern with the values, cosmology, and ways of knowing that are common to all people in a given society has given way to how ideas and innovation are communicated in more bounded and explicit fields of knowledge. Such studies include the diffusion of knowledge and ideas in Silicon Valley, in the genome project, and in biogenetic engineering, and the studies also consider the implications of the new communications media on ideas of religious and political authority. The processes by which change is embodied in implicitly shared ideas of gender and sexuality and of authority and responsibility also form a component of current concerns in the anthropology of knowledge.

See also: Anthropology and History; Anthropology: Overview; Collective Beliefs, Sociological Explanation of; Collective Memory, Anthropology of; Cultural Critique, Anthropological; Cultural Relativism, Anthropology of; Culture: Contemporary Views; Ethnography; Ethnology; Historical Thought and Historiography: Current Trends; Interpretation in Anthropology; Knowledge (Explicit, Implicit and Tacit): Philosophical Aspects; Knowledge, Sociology of; Linguistic Anthropology; Modernity: Anthropological Aspects; Phenomenology: Philosophical Aspects; Piaget's Theory of Human Development and Education; Psychological Anthropology; Tradition, Anthropology of.

Bibliography

Assmann, J., 2011. Communicative and cultural memory. In: Meusberger, P., Heffernan, M., Wunder, E. (Eds.), Cultural Memories: The Geographical Point of View. Springer, Heidelberg, pp. 15–27.

Barth, F., 1987. Cosmologies in the Making: A Generative Approach to Cultural Variation in Inner New Guinea. Cambridge University Press, Cambridge, UK.

Barth, F., 2002. An anthropology of knowledge. Current Anthropology 43, 1–18.

Bateson, G., 1972. Steps to an Ecology of Mind. Ballantine Books, New York.

Berlin, B., Kay, P., 1969. Basic Color Terms: Their Universality and Evolution. University of California Press, Berkeley, CA.

Bloch, M., 1985. From cognition to ideology. In: Fardon, R. (Ed.), Power and Knowledge: Anthropological and Sociological Approaches. Scottish Academic Press, Edinburgh, UK.

Bourdieu, P., 1967. Systems of education and systems of thought. International Social Science Journal 19, 338–358.

Bourdieu, P., 1977. Outline of a Theory of Practice. Cambridge University Press, New York.

Cohen, E., 2010. Anthropology of knowledge. Journal of the Royal Anthropological Institute (N.S.), S193–S202.

Cole, M., 1996. Cultural Psychology: A Once and Future Discipline. Harvard University Press, Cambridge, UK.

Douglas, M., 1966. Purity and Danger: An Analysis of Concepts of Pollution and Taboo. Frederick A. Praeger, New York.

Durkheim, É., 1915. The Elementary Forms of the Religious Life. George Allen and Unwin Ltd, London.

Durkheim, É., 1977. The Evolution of Educational Thought. Routledge and Kegan Paul, London.

Durkheim, É., Mauss, M., 1963. Primitive Classification. University of Chicago Press, Chicago. (Original work published 1903.)

Eickelman, D.F., 1985. Knowledge and Power in Morocco: The Education of a Twentieth Century Notable. Princeton University Press, Princeton, NJ.

Elkhana, Y., 1981. A programmatic attempt at an anthropology of knowledge. In: Mendelsohn, E., Elkana, Y. (Eds.), Sciences and Culture: Anthropological and Historical Studies of the Sciences. Kluwer, Boston, pp. 1–76.

Evans-Pritchard, E.E., 1940. The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People. Clarendon Press, Oxford, UK.

Fardon, R. (Ed.), 1985. Power and Knowledge: Anthropological and Sociological Approaches. Scottish Academic Press, Edinburgh, UK.

Godelier, M., 1986. The Mental and the Material. Verso, London.

Griaule, M., 1965. Conversations with Ogotêmmeli. Oxford University Press for the International African Institute, London. (Original work published 1948.)

Hefner, R., 1985. Hindu Javanese: Tengger Tradition and Islam. Princeton University Press, Princeton, NJ.

Holland, D., Quinn, N., 1987. Cultural Models in Language and Thought. Cambridge University Press, Cambridge, UK.

Lévi-Strauss, C., 1966. The Savage Mind. University of Chicago Press, Chicago.

Lévy-Bruhl, L., 1922. La mentalité primitive. Librairie Félix Alcan, Paris.

Marchand, T., 2010. Making knowledge: explorations of the indissoluble relation between minds, bodies, and environment. Journal of the Royal Anthropological Institute (N.S.), S1–S21.

Radin, P., 1957. Primitive Man as a Philosopher. Dover Books, New York. (Original work published 1927.)

Robertson Smith, W., 1972. Lectures on the Religion of the Semites. Schocken Books, New York. (Original work published 1889.)

Shils, E., 1972. The Intellectuals and the Powers. University of Chicago Press, Chicago.

Shore, B., 1996. Culture in Mind: Cognition, Culture, and the Problem of Meaning. Oxford University Press, New York.

Sperber, D., 1985. On Anthropological Knowledge. Cambridge University Press, Cambridge, UK.

Tonkinson, R., 2011. Landscape, transformations, and immutability in an aboriginal Australian culture. In: Meusberger, P., Heffernan, M., Wunder, E. (Eds.), Cultural Memories: The Geographical Point of View. Springer, Heidelberg, pp. 329–345.

Turner, V., 1967. Muchona the hornet, interpreter of religion. In: Turner, V. (Ed.), The Forest of Symbols: Aspects of Ndembu Ritual. Cornell University Press, Ithaca, NY. (Original work published 1957.)

Wagner, D., 1993. Literacy, Culture, and Development: Becoming Literate in Morocco. Cambridge University Press, New York.

Knowledge (Explicit, Implicit and Tacit): Philosophical Aspects

Martin Davies, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Explicit knowledge is knowledge that the knower can make explicit by means of a verbal statement; implicit knowledge is knowledge that is not explicit. Chomskyan tacit knowledge of rules can be understood as requiring a particular structure in the content-involving explanations of transitions between internal representations. States of tacit knowledge are different from beliefs and states of explicit knowledge in that their content is at least partly nonconceptual. Several applications of the notion of tacit knowledge in philosophical theory are reviewed. Chomsky's notion of tacit knowledge of the rules of a generative grammar provided the theoretical background for research on artificial grammar learning (AGL). Results and disputes in AGL research are reviewed, with particular reference to requirements on testing for conscious versus unconscious knowledge. No test is guaranteed to meet the requirements but tests can be validated if they are used in the context of psychological theories about qualitative differences between conscious and unconscious processes.

Explicit knowledge is knowledge that the knower can make explicit by means of a verbal statement: "Someone has explicit knowledge of something if a statement of it can be elicited from him by suitable enquiry or prompting" (Dummett, 1991). Implicit knowledge can then be defined simply as knowledge that is not explicit. On this construal, implicit knowledge corresponds roughly to what Polanyi (1967) calls 'tacit knowing': "we can know more than we can tell."

Tacit knowing in Polanyi's sense encompasses a variety of phenomena, such as the ability to recognize something (e.g., a person's face) even though one cannot describe in context-independent terms (e.g., without saying, "I know that Bob looks like *this*") how one recognizes it. Polanyi's tacit knowing also includes the holistic diagnostic skills that an experienced and humane psychologist or psychiatrist brings to a clinical encounter. Diagnostic judgment depends on 'context-dependent practical knowledge' (Thornton, 2013).

While explicit knowledge can be expressed and communicated linguistically, Polanyi's tacit knowing cannot be captured in language without reliance on context-dependent or demonstrative elements (e.g., I know that to tie my shoelaces I move my fingers *thus* (in this way)). But tacit knowing can still be communicated and shared. The mechanism of transmission is not telling and understanding, but rather performance and imitation (learning by example). The notion of socially and culturally shared knowledge in the service of solving theoretical and practical problems (perhaps in the form of shared knowledge about what will happen, and about what to do, in various situations) is of evident importance for the social sciences. Specifically, shared theoretical and practical knowledge can be seen to play a foundational role in the establishment and development of institutions (Mantzavinos, 2001; Mantzavinos et al., 2004).

A different notion of tacit knowledge, introduced by Chomsky (1965; see also Fodor, 1968), has been of central importance for philosophy and cognitive science over the last half century, and this will be the focus of the first four sections of this article. An account of Chomskyan tacit knowledge will be given and the relationship between tacit knowledge and other psychological states will be discussed (Section An Account of Tacit Knowledge). Two distinctions that are drawn using the terminology of 'explicit' versus 'implicit' will be explained (Section Representation and Rules: Explicit and Implicit): information that is explicitly stored versus information that is implicitly represented (implied by explicitly stored information) and rules that are explicitly stored versus rules that are implicit in a system ('wired in'). Four examples of the use of the notion of tacit knowledge in philosophical theory will be reviewed (Section Tacit Knowledge in Philosophy).

Chomsky's notion of tacit knowledge provided the theoretical background for empirical research using the artificial grammar learning (AGL) task (Section Tacit Knowledge in Psychology: The AGL Task) and this research program will be described (Section Implicit Learning and Tacit Knowledge of Rules). The primary interest of AGL research for this article is that it makes use of a distinction between conscious and unconscious knowledge – a contemporary rendition of the explicit versus implicit knowledge distinction. Thus, empirical research using the AGL task can be seen as a chapter in the scientific study of consciousness. Theoretical proposals about testing for conscious knowledge will be discussed (Section Conscious and Unconscious Knowledge) and recent work on conscious (explicit) and unconscious (implicit) knowledge using AGL and related experimental paradigms will be reviewed (Section Conscious Knowledge in the Context of Psychological Theory).

Chomsky on Tacit Knowledge of Rules

In a famous passage, Chomsky (1965) said, "Obviously, every speaker of a language has mastered and internalized a generative grammar that expresses his knowledge of his language. This is not to say that he is aware of the rules of the grammar or even that he can become aware of them." This internalization of the rules or principles of a generative grammar is usually described as tacit knowledge and the claim that each language user tacitly knows a generative grammar has been a core element of Chomskyan linguistics and cognitive science (e.g., Chomsky, 1986: "the central concern becomes knowledge of language"). Tacit knowledge in this Chomskyan sense counts

as implicit according to our initial definition, because a speaker who has tacit knowledge of the rules or principles of a grammar is usually unable to provide a verbal statement of those rules or principles. (A student of linguistics may be able to state the rules that he tacitly knows. But, even in this case, the tacit knowledge is not simply transformed from implicit to explicit. The linguistics student's state of explicit knowledge will probably be based in part on reading books and listening to lectures and will be somewhat independent from – not constituted or underpinned by, and not a manifestation of – his or her state of tacit knowledge.)

Chomsky on the Faculty of Language

We have introduced the notion of tacit knowledge as being Chomskyan. But, before proceeding, it is appropriate to acknowledge that, in more recent presentations (e.g., Chomsky, 2000, 2005; Hauser et al., 2002), Chomsky describes his project in rather different terms, and from a 'biolinguistic perspective.'

According to these more recent presentations, the object of study is an individual's internal language (I-language), the attained state of the individual's faculty of language. The faculty of language itself is conceived as a component or organ of the mind/brain, with its initial state (studied under the heading 'universal grammar') fixed by genetic endowment. At the heart of the faculty of language are a collection of primitive items (the lexicon or vocabulary) and a procedure that generates complex structures (expressions) from the lexical items. The faculty of language interfaces with sensorimotor (perceptual-articulatory) systems and with conceptual-intentional systems and each expression, E, is a pair <PHON(E), SEM(E)>, in line with the conception of language as a system of sound–meaning pairings.

It seems clear that, on this account, the attained state of the faculty of language is not a state of explicit or tacit knowledge about the I-language; it is the I-language itself. This attained state of the faculty of language is not a state of knowledge about a set of lexical items and about complex expressions built from those items; rather, it is constituted by the lexicon and the generative procedure themselves. And the procedure generates, not representations of complex expressions, but the expressions themselves.

Someone might suggest that the primary function of language is interpersonal communication and that the internally generated expressions of the I-language should be regarded as representations of external tokens of sentences in communicative use. Consequently, the state of the faculty of language should still be regarded as a body of knowledge – knowledge about external language. But Chomsky would utterly reject that suggestion. Following on from some discussion of remarks by Jacob and Luria, Chomsky (2005) describes the internally generated expressions as "the infinite variety of internal structures that enter into thought, interpretation, planning, and other human mental acts, and that are sometimes put to use in action, including the externalization that is a secondary process if the speculations just reviewed turn out to be correct."

There is some debate as to whether there has really been a radical change in Chomsky's substantive position (Collins, 2004; Matthews, 2006). But, however that may be, many questions arise. For example, even if the attained state of the faculty of language is the I-language, rather than being a body of tacit knowledge about the I-language, subjects still have explicit knowledge about their own languages, and some of this knowledge seems to have a first-person character. If Chomsky is correct, then this first-personal linguistic knowledge cannot be explained in terms of the subject's drawing on states of tacit knowledge embodied in the faculty of language. So one continuing project for foundational research on knowledge of language is to provide and evaluate accounts of first-personal linguistic knowledge that do not appeal to tacit knowledge (Smith, 2008).

With this acknowledgment of the nature of Chomsky's recent descriptions of his project, we return to 'Chomskyan' tacit knowledge of rules.

Quine's Challenge

Quine challenged Chomsky's introduction of the notion of tacit knowledge by making use of the distinction between behavior that is guided by a rule and behavior that merely conforms to a rule. As Quine (1970) used the notion, guidance by a rule requires explicit knowledge of that rule: "the behavior is not *guided* by the rule unless the behaver knows the rule and can state it." In contrast, as the notion of conformity is used here, conformity to a rule does not require knowledge of the rule. Behavior conforms to a rule if it fits the pattern that the rule requires. Thus, a subject can behave in a way that conforms to a rule although his behavior is not guided by the rule: "Fitting is a matter of true description; guiding is matter of cause and effect."

Chomsky's tacit knowledge of a rule is supposed to be less demanding than the kind of guidance that requires explicit knowledge of the rule; but it cannot be equated with mere conformity to the rule. In fact, conformity to a rule is neither necessary nor sufficient for tacit knowledge of the rule. Conformity is not necessary for tacit knowledge, because the psychological state of tacit knowledge of a rule does not guarantee effective use of that knowledge to generate behavior that conforms to the rule. In Chomsky's terms, competence – which is a state of knowledge – must be distinguished from performance. (This distinction between having knowledge and making use of that knowledge holds good for explicit, as well as for tacit, knowledge. Sometimes – all too often – we possess explicit knowledge that is relevant to the task at hand and yet the knowledge is not put to use in our behavior.)

More importantly for present purposes, conformity to a rule is not sufficient for tacit knowledge, because a tacit knowledge claim is not offered as a summary description of a pattern in behavior but, rather, as a putative explanation of behavior. There will always be alternative sets of rules that require just the same behavior for conformity, and thus provide equally adequate summary descriptions of a battery of conforming behavior. But it is part of the idea of tacit knowledge that a speaker's actual behavior might be correctly explained in terms of tacit knowledge of one set of rules, rather than in terms of tacit knowledge of an alternative set of rules to which the behavior also conforms.

Quine (1970) expressed skepticism about Chomsky's conception of tacit knowledge as intermediate between explicit guidance and mere conformity: "Implicit guidance is a moot enough idea to demand some explicit methodology." It is at this point that Quine posed his challenge. He insisted that, if an attribution of tacit knowledge is an empirical claim that goes beyond a summary of conforming behavior, then it should be possible to indicate what kinds of evidence would count in favor of, or against, that attribution. He also insisted that this evidence should involve the subject's own behavior.

Quine's demand is challenging because there is no unique set of rules to which a subject's behavior conforms. But there are two points to be made in response. First, there can be no a priori limit on the kinds of evidence that might be relevant to an empirical claim. So it is not legitimate to restrict evidence that could support an attribution of tacit knowledge to the behavior of the very subject to whom the attribution is being made. Second, the more general point about evidence is a fair one. There should be something to be said about what kinds of evidence would support an attribution of tacit knowledge. But more fundamental than the question about evidence is the question what the correctness of such an attribution would consist in – what a state of tacit knowledge is.

An Account of Tacit Knowledge

We can sketch one kind of answer to this fundamental question about tacit knowledge (Davies, 1987; Evans, 1981; Peacocke, 1989) by using an example that involves, not rules of syntax, but rules of pronunciation for simple three-letter strings. Suppose that there are five initial consonants, 'B,' 'D,' 'F,' 'H,' 'K'; five central vowels, 'A,' 'E,' 'I,' 'O,' 'U'; and five final consonants, 'M,' 'N,' 'P,' 'T,' 'Z,' Thus, there are 125 three-letter strings. Suppose that each letter is assigned a single phoneme by a letter–sound rule, such as the 'B' → /b/ rule: if a string begins with the letter 'B,' then its pronunciation begins with the phoneme /b/. The 15 letter–sound rules logically determine pronunciations for all the 125 strings.

Now consider a person reading these letter strings aloud. Behaviorally, the person produces a spoken pronunciation when a letter string is visually presented; for example, the person produces the pronunciation /bæm/ (rhymes with 'ham') when the letter string 'BAM' is presented. Suppose that this person's behavior – vocally producing pronunciations in response to the visual presentation of letter strings – conforms to the 15 letter–sound rules. We assume that there are psychological processes underpinning this behavior, and that these processes involve transitions from internal representations of presented letter strings to internal representations of to-be-produced pronunciations. These transitions can also be described as conforming to the letter–sound rules: When an input represents a three-letter string, the output represents the pronunciation that is determined by the rules. In this setting, our fundamental question is: What would it be for this person (or, indeed, for a mechanical reading-aloud system) to have (or embody) tacit knowledge of the 15 letter–sound rules?

The answer to this question, in outline, is that tacit knowledge of the rules requires a particular structure in the causal explanations of the transitions from internal representations of letter strings (orthographic representations) to internal representations of pronunciations (phonological representations). Consider first the abstract structure of the logical determination of pronunciations by the letter–sound rules. Each rule (e.g., the 'B' → /b/ rule) makes a common contribution to the logical determination of the pronunciations of 25 three-letter strings (e.g., those beginning with 'B'); and the pronunciation of each letter string is logically determined by three letter–sound rules. Tacit knowledge of the 15 letter–sound rules then requires a causal-explanatory structure that mirrors this structure of logical determination. A state of tacit knowledge of a letter–sound rule makes a common contribution to the causal explanations of 25 transitions from orthographic to phonological representations; and the causal explanation of each transition involves three states of tacit knowledge of letter–sound rules.

Clearly, conformity to the rules is not sufficient for tacit knowledge of the rules according to this outlined account. For example, mere conformity to the rules does not guarantee that there is any common factor in the causal explanations of the 25 transitions from orthographic representations of letter strings beginning with 'B' to phonological representations of their pronunciations (beginning with /b/). Conformity to the rules might be achieved although, for each of the 25 strings beginning with 'B,' the transition from an orthographic to a phonological representation depended on causal processes that were specific to that string.

The Structure Required for Tacit Knowledge

The kind of structure that tacit knowledge requires, in the causal explanations of transitions between psychological states, is not mysterious. Stepping back from internal representations and tacit knowledge for a moment, consider transitions between ordinary conscious beliefs (or states of explicit knowledge) about presented letter strings and conscious beliefs (or, again, states of explicit knowledge) about to-be-produced pronunciations. In particular, consider these transitions in two subjects, Syst and List. Suppose that Syst has explicit knowledge of the 15 letter–sound rules and is able to draw on this knowledge to work out – that is, to infer – the pronunciations of all 125 three-letter strings. For each letter–sound rule, explicit knowledge of it figures as a common factor in Syst's inferences about the pronunciations of 25 three-letter strings and, for each of the 125 three-letter strings, Syst's inference draws on explicit knowledge of three letter–sound rules. Suppose that List, in contrast, lacks explicit knowledge of the letter–sound rules but has 125 separate pieces of explicit knowledge, one for each three-letter string – from explicit knowledge that string #1, 'BAM,' is pronounced /bæm/ through to explicit knowledge that string #125, 'KUZ,' is pronounced /kʌz/ (rhymes with 'fuzz'). We can say that List has explicit knowledge of 125 pronunciation rules, but each of these rules applies to only one three-letter string.

Syst and List assign the same pronunciations to the 125 letter strings, but there are substantial differences in the structure of the causal explanations of their transitions from beliefs about letter strings to beliefs about pronunciations. For example, there is a single psychological state of Syst (namely, explicit knowledge of the 'B' → /b/ letter–sound rule) that

makes a common contribution to the causal explanation of each of Syst's beliefs about the pronunciations of letter strings beginning with the letter 'B.' But there is no such common factor in the causal explanations of List's beliefs. Similarly, the psychological states that figure in the explanations of Syst's beliefs about the pronunciations of some letter strings (e.g., that 'BAM' is pronounced /bæm/ and that 'DEN' is pronounced /den/) are together sufficient to explain Syst's beliefs about the pronunciations of other letter strings ('BAN,' 'BEM,' and four others); but nothing similar is true of List.

These differences in causal-explanatory structure between Syst and List may be manifested in many ways. For example, List may lose (e.g., by forgetting) his explicit knowledge about the pronunciation of just one-letter string (e.g., 'BAM'), while retaining his knowledge about the pronunciations of the other 124 strings. But if Syst loses some of his relevant explicit knowledge and no longer believes that the letter string 'BAM' is pronounced /bæm/, then he will probably also lose his beliefs about several other letter strings (all strings with 'B' as initial consonant, or all strings with 'A' as central vowel, or all strings with 'M' as final consonant). Similarly, List may revise his beliefs about the pronunciations of individual letter strings in a piecemeal way; but Syst is likely to revise his beliefs about the pronunciations of letter strings in a systematic way.

We can use these descriptions of the role of explicit knowledge of rules in Syst's and List's transitions between conscious beliefs to clarify the kind of causal-explanatory structure that tacit knowledge of rules requires. A state of explicit knowledge (e.g., of a letter–sound rule, in the case of Syst) is a state with a particular content and it makes a common contribution to the content-involving causal explanations of many inferential transitions between beliefs; the transitions are explained as transitions between beliefs with their particular contents. In a similar way, a state of tacit knowledge is a state with a particular content and it makes a common contribution to the content-involving causal explanations of many inference-like transitions between representations; the transitions are explained as transitions between states with particular representational properties (Peacocke, 2008). These representations might not be beliefs but, rather, representations or states of information of the kind that figure in information-processing psychology. Indeed, states of tacit knowledge are often states of information-processing systems within a subject – systems whose internal operations are not normally accessible to the subject's consciousness.

In general, a state of tacit knowledge is a state that figures as a common factor in content-involving causal explanations of transitions between representations or states of information. Where different sets of rules require the same transitions for conformity (the case that concerned Quine), the correct attribution to a subject of tacit knowledge of one set of rules rather than another requires a particular structure in the causal explanations of transitions between representations within the subject. For example, correct attribution of tacit knowledge of the 15 letter–sound rules rather than the 125 one-case pronunciation rules requires the Syst-like causal-explanatory structure rather than the List-like structure (Davies, 1987; Evans, 1981; Peacocke, 1989).

Once an account of tacit knowledge in terms of causal-explanatory structure is in place, it is a relatively straightforward matter to give examples of empirical evidence that would count in favor of – and other evidence that would count against – the attribution to a subject of tacit knowledge of a particular set of rules, such as rules of a language. Indeed, some of this evidence meets Quine's additional requirement of concerning the behavior of the subject to whom the attribution is being made. Relevant behavioral evidence could come from experimental studies of language acquisition, language perception, and language impairment following brain damage, while further evidence would be available from neural imaging.

Tacit Knowledge, Beliefs, and Nonconceptual Content

While it is fairly natural to describe tacit knowledge of rules (such as syntactic rules, in Chomsky's example) as a case of 'knowing more than we can tell,' it is also possible to query the appropriateness of the use of the term 'knowledge' here. Knowledge is usually taken to require belief and to be connected with reliability of the method by which the belief is produced and with notions of justification, warrant, or entitlement to believe. Information-processing mechanisms whose operation is not accessible to consciousness may well meet requirements of reliability, and subjects may have an epistemic entitlement to rely on such mechanisms in forming beliefs (Burge, 2003). Thus, a belief in whose production states of tacit knowledge are implicated may be justified or warranted. But states of tacit knowledge themselves fall outside the scope of epistemic justification or warrant because they are not beliefs.

Several considerations may be offered in support of the claim that states of tacit knowledge are different from beliefs. One is that a subject is usually unaware of his or her own states of tacit knowledge and, indeed, unable to become aware of them. But this, by itself, is not clearly decisive because it seems possible to make sense of the notion of an unconscious belief – a belief of which one is not aware, and perhaps even a belief of which one cannot become aware. Second, states of tacit knowledge are limited in their causal role and influence. It is characteristic of beliefs that a belief may figure in many different ways in a subject's thinking, reasoning, and planning, depending on the subject's other beliefs and desires. A belief is "at the service of many distinct projects" (Evans, 1981). States of tacit knowledge, in contrast, are not inferentially integrated with the subject's beliefs (Stich, 1978); they are inferentially isolated, often with causal roles that are limited to an encapsulated and domain-specific information-processing system (a module; Fodor, 1983). However, this again may not be quite decisive, because we can make sense of the ideas of fragmented rather than integrated minds, and of beliefs that are compartmentalized, guiding behavior in only some – perhaps very few – contexts (Egan, 2008). Some monothematic and circumscribed delusions may be beliefs of this kind (Davies et al., 2001).

A third consideration supporting a distinction between states of tacit knowledge and beliefs is that believing requires the possession of concepts. No one can believe that emus do not fly without having some concept of emus and of flying (and

negation). But, in general, the subject who has tacit knowledge of a rule does not need to grasp or possess the concepts that a theorist would use to specify the rule that is tacitly known. This is particularly clear in the case of tacit knowledge of rules of language (e.g., rules of syntax). Every normal speaker has tacit knowledge of such rules but only people who have studied linguistics grasp all of the concepts that are needed to specify the rules. Thus, states of tacit knowledge are psychological states with nonconceptual content (or, at least, with content that is partly nonconceptual). The subject of the state does not need to possess (still less, to deploy) all the concepts that would be required to specify the content of the state (Davies, 1989). Beliefs, in contrast, are the paradigm example of psychological states with conceptual content.

Critiques of Tacit Knowledge

Quine is certainly not the only philosopher to have adopted a critical view of the notion of tacit knowledge as it figures in linguistics, psychology, and cognitive science.

Some philosophers have argued that Chomsky's claims about tacit knowledge involve a conceptual confusion, because the notion of a rule of language belongs with the idea of a normative practice in which people advert to rules to justify, criticize, or excuse their actions (Baker and Hacker, 1984). But the notion of a tacitly known rule should be distinguished from the notion of a rule that figures in a normative practice. Once the distinction is made, there remains an important question about the relationship between the rules that are tacitly known by individual speakers and the normative practices in which those speakers participate. But that question does not indicate any conceptual incoherence in the very idea of tacit knowledge.

More recently, Bennett and Hacker (2003) have argued that the information-processing paradigm that is shared by cognitive neuroscience and cognitive science rests on a philosophical mistake. They say that psychological powers (e.g., perceiving, thinking, feeling) that are properly attributed to whole human beings should not be attributed to parts of human beings. According to this critique, neither the brain (in cognitive neuroscience) nor an information-processing module (in cognitive science) is an appropriate subject of psychological predicates. Here, we should acknowledge that, in the sciences of the mind, systems smaller than a whole human being are credited with psychological states and processes such as inference-like transitions between representations, mediated by states of tacit knowledge. But it is not clear that this is a philosophical mistake, for these psychological attributions can surely be distinguished from attributions of perceptual experiences, thoughts, and feelings. (We have just seen that states of tacit knowledge are importantly different from beliefs.)

It has also been argued that the ideas of tacit knowledge of rules, and of unconscious information processing in general, are problematic because of a deep connection between intentionality (the representational nature of thoughts and beliefs) and consciousness. This critique culminates in the claim that "there are brute, blind, neurophysiological processes and there is consciousness; but there is nothing else" (Searle, 1990). Specifically, according to this critique, there is no unconscious information processing and there are no states of tacit knowledge that are inaccessible to consciousness. But, once again, a response to the critique can begin from a distinction. Even if we accept that there is an important connection between intentionality and consciousness at the level of the whole human being, states of tacit knowledge and representations within cognitive modules are not beliefs, and a philosophical account of their representational nature may be quite different from an account of the intentionality of beliefs. (For a defense of the appeal to tacit knowledge in psychological explanation against an earlier critique (Ryle, 1949), see Fodor, 1968.)

There remain questions about the relationship between the psychological properties of whole conscious human beings – which belong at the folk-psychological or personal level of description – and the information-processing properties of component systems of the mind and brain – which belong at the subpersonal level of cognitive and neural mechanisms (Dennett, 1969). But there is no reason to assume that answering these questions in a way that highlights what is distinctive about the mental lives of conscious persons will require a radically negative assessment of cognitive science (Davies, 2000).

Representation and Rules: Explicit and Implicit

Our starting point was that explicit knowledge is knowledge that the knower can make explicit by means of a verbal statement. Thus, explicitness is a matter of the subject being able to present information in linguistic form; it is not a matter of how the information is stored in the subject's mind or brain when it is not being called upon.

Suppose that someone knows, in the everyday sense of that term, the axioms of some theory. Provided that this knowledge can be verbally articulated, it counts as explicit knowledge. Now consider some theorem that is derivable from those axioms. The person who knows the axioms may well, with some suitable inquiry and prompting, be able to see that the theorem follows from the axioms and to state it verbally. For example, someone who explicitly knows some elementary arithmetical facts may be able to work out, and to state, that $68 + 57 = 125$. On Dummett's (1991) account, this latter piece of knowledge, even though the sum is computed when needed rather than being stored in memory, counts as no less explicit than the stored elementary facts from which it is derived. Indeed, it counts as explicit knowledge even before the knower works it out; for the personal-level notion of explicit knowledge is defined in terms of the possibility of eliciting a verbal statement by inquiry or prompting. Thus, once some propositions are classified as explicitly known, the category of explicit knowledge also includes at least some of the as-yet-undrawn consequences of those propositions.

However, the terms 'explicit' and 'implicit' are also used to mark a number of distinctions that are specifically related to the subpersonal-level storage and processing of information. Two such distinctions are explained in this section.

Inference and Implicit Representation

Dennett (1983) introduced a quite different distinction from Dummett's when he said, "Let us have it that for information to

be represented *implicitly*, we shall mean that it is *implied* logically by something that is stored explicitly." On this usage, the as-yet-undrawn consequences of propositions that count as explicitly stored would be classified as implicitly represented. In fact, relative to any given notion of explicit storage of information, it is possible to define a whole family of notions of implicit representation, differing over the inferential resources that can be used in drawing out consequences from the information that is explicitly stored.

There is no uniquely correct account of explicit storage but one natural idea is that explicit storage involves the use of a language-like code (a subpersonal-level echo of the requirement for explicit knowledge). A second idea that is fairly natural in the context of theories of information storage and processing is that explicitly stored information has to be accessed before it can be used. A problem in accessing stored information may result in a failure to use information that is relevant to the task at hand. Putting the two ideas together, we have a picture in which explicit storage of information is a matter of syntactically structured inscriptions on pages that are not always easy to locate. Drawing out the consequences of explicitly stored information could then be conceived as manipulating inscriptions in ways that are analogous to the formal procedures involved in the derivation of theorems from axioms in a logical system (the computational model of the mind; Fodor, 1987).

Beginning from this notion of explicit storage of information, we could consider information that can be arrived at by deductive inference from information that is explicitly stored. If elementary arithmetical facts were explicitly stored in a person's memory, then the information that $68 + 57 = 125$ would count as being implicitly represented. Thus, the subpersonal-level contrast between explicit storage and implicit representation could distinguish between two cases that would both count as explicit knowledge at the personal level. This deductive notion of implicit representation could be refined by distinguishing, for example, between information that can be arrived at from the explicitly stored information by unrestricted deductive inference, by deductive inference subject to some set of human limitations, or by deductive inference that is within the compass of some specified individual. Alternatively, we could go beyond deductive inference and allow methods of induction (or rule extraction) and abduction (inference to the best explanation). Then, for example, if only facts about particulars were explicitly stored, generalizations or rules that subsumed those facts would count as implicitly represented.

Rule-Explicit and Rule-Implicit Systems

The notions of explicit storage and implicit representation belong with mechanistic theories about the storage and processing of information. These notions could be applied to any information-processing system and they have no special connection with human subjects. Even in the case of human subjects, we have already seen that subpersonal-level explicit storage is not necessary for personal-level explicit knowledge. Explicitly known information might be either explicitly stored or implicitly represented. It is also the case that explicit storage – even explicit storage within a human subject – is not sufficient for explicit knowledge. Information might be explicitly stored within a human subject yet not available to the subject for verbal report, not accessible to the subject's consciousness, and not conceptualized by the subject.

Tacit knowledge of a rule may involve explicit storage of this kind. An internal representation encodes the rule in a syntactically structured format and a process of searching for and accessing that representation is required before tacit knowledge of the rule can contribute causally to any transitions between input and output representations. For example, letter–sound rules could be explicitly stored within an information-processing system for reading aloud words and other pronounceable letter strings (see the implemented computational model of reading aloud described by Coltheart et al., 2001). In such a case, the system is said to be rule-explicit, and there is a clear gap between having tacit knowledge of the rules (that is, having the rules represented in the system) and using that tacit knowledge.

While the account of tacit knowledge in terms of causal-explanatory structure allows for tacit knowledge that is a matter of explicit storage, it certainly does not require that tacitly known rules should always be explicitly stored. A state of tacit knowledge is a state that figures as a common factor in causal explanations of transitions between representations. The use of an explicitly stored rule can figure as such a causal common factor, but so can a component processor (a 'transition mediator'; Peacocke, 1989) within an information-processing system. The transition corresponding to the rule can be 'wired in' to the system, which is then described as rule-implicit (Fodor and Pylyshyn, 1988).

The distinction between rule-explicit and rule-implicit systems is quite different from the distinction (in the previous subsection) between explicit storage and implicit representation of information. Rule explicitness is a matter of explicit storage but rule implicitness (that is, a rule that is 'wired in' to the system) is not a matter of implicitly represented information (that is, information that is implied by what is explicitly stored). Partial analogies for both contrasts can be found by considering a formal theory in which 15 proper (that is, not purely logical) axioms state the letter–sound rules, and theorems specifying the pronunciations of the 125 three-letter strings can be derived. If the axioms are analogous to what is explicitly stored in an information-processing system, then the theorems that are implied by the axioms are analogous to what is implicitly represented.

Each proper axiom makes a common contribution to the proofs of 25 theorems specifying the pronunciations of letter strings and, in the proof of each such theorem, three proper axioms figure. This structure of derivational common factors would be preserved if the theory were formulated, not with 15 proper axioms, but instead with 15 proper rules of inference (e.g., From: String X begins with letter 'B,' infer: The pronunciation of X begins with the phoneme /b/). Derivationally, a rule of inference is a transition mediator and the distinction between proper axioms and proper rules of inference in a formal theory provides a partial analogy for the distinction between explicitly stored rules and rules that are implicit in ('wired in' to) an information-processing system.

Tacit Knowledge in Philosophy

The starting point for this article was Dummett's (1991) definition of explicit knowledge as knowledge that the knower can make explicit by means of a verbal statement. Explicit knowledge was contrasted with implicit knowledge, conceived as Polanyi's tacit knowing, and with Chomsky's tacit knowledge. We have given an account of tacit knowledge as having a characteristic causal-explanatory role: a state of tacit knowledge makes a common contribution to content-involving causal explanations of transitions between representations. States of tacit knowledge are different from beliefs and states of explicit knowledge in that the subject of a state of tacit knowledge does not need to possess all the concepts that would be required to specify the content of the state. We have also explained the subpersonal-level distinctions between information that is explicitly stored and information that is implicitly represented and between a tacitly known rule that is explicitly stored and a tacitly known rule that is embodied in a component processor (transition mediator).

In this section, we review four cases where the notion of tacit knowledge figures in philosophical theory.

Tacit Knowledge of Compositional Semantic Theories

Philosophers of language have an interest in compositional semantic theories for natural languages; that is, theories that reveal how the meanings of complete sentences depend on the meanings of their constituent words and the ways in which those words are put together. In such theories, proper axioms specify the meanings of words and basic modes of combination of words, and theorems that are derivable from those axioms specify the meanings of complete sentences. While it is plausible that ordinary speakers of a natural language have, or can come to have, explicit knowledge of some of these theorems, it is also plausible that ordinary speakers do not need to have, and usually do not have, explicit knowledge of all the axioms of a semantic theory for their language. What, then, is the relationship between an ordinary language user and the axioms of a semantic theory for his or her language?

One possible answer to this question is that the epistemological relationship between the language user and the theory extends no further than the theorems that specify meanings for complete sentences, and that the role of the axioms is merely descriptive. If a semantic theory has just a finite number of proper axioms (one for each word in the language and one for each basic mode of combination), from which meaning specifications for complete sentences can be derived, then those axioms provide a statable summary description of the language user's capacity to understand complete sentences.

Another possible answer makes use of the notion of explicit knowledge of the axioms, but in an indirect way. Davidson (1984) says that a condition of adequacy for a semantic theory is that what it states should be "something knowledge of which would suffice for interpreting utterances of speakers of the language to which it applies." Clearly, if explicit knowledge of the theorems that specify meanings for complete sentences would suffice for interpreting utterances of speakers, then so would explicit knowledge of the axioms from which those theorems can be derived.

Because the explicit knowledge that would suffice for interpretation or understanding is not knowledge that ordinary speakers have, this second answer shares an important property with the first answer. Neither answer proposes a relationship between ordinary speakers and the axioms of a compositional semantic theory that would provide an explanation of the fact that ordinary users of a language are able to understand sentences that they have never encountered before. An explanation of this fact is available, however, if language users have tacit knowledge of the axioms of a compositional semantic theory for their language. If a hitherto unencountered sentence is built out of familiar words in familiar ways, then understanding of the new sentence can draw on the language user's tacit knowledge of meaning-specifying axioms for those words and modes of combination (Evans, 1981; Davies, 1987).

The appeal to tacit knowledge here has an additional theoretical advantage of assigning a determinate meaning even to sentences that will never be used. An alternative solution to the 'meaning without use' problem appeals, not to tacit knowledge, but simply to extrapolation from the used sentences to unused sentences: "Use determines some meanings, those meanings determine the rules, and the rules determine the rest of the meanings" (Lewis, 1992). However, this alternative solution to the problem has an unattractive consequence.

Suppose that a speaker PB uses (or equally, suppose that a population of speakers use) a finite set S of sentences from a language (for which there is a compositional semantic theory), but that PB has learned the meanings of these sentences by dipping into a phrase book. Let us stipulate that PB treats each sentence in S as an unstructured string of letters or sounds, the meaning of which he needs to memorize. Consequently, PB has a separate piece of tacit or explicit knowledge about the meaning of each sentence in the set. We can say that PB has tacit or explicit knowledge of a noncompositional semantic theory with an axiom for each sentence in S. But PB's tacit or explicit knowledge about meanings is not adequate for understanding any sentence beyond the set S.

A semantic theorist could fit a compositional semantic theory to the sentences in S and extrapolate to sentences outside that set. But it is plausible that sentences outside S do not belong to the phrase-book 'language' that PB has learned, and that sentences outside S have no determinate meaning for PB – no determinate meaning in PB's language. Consequently, the appeal to tacit knowledge is to be preferred over the appeal to extrapolation (Schiffer, 1993, 2006). As Loar (1981) says, "the Chomskyan idea of the internalization of the generative procedures of a grammar has got to be invoked to…make sense of literal meaning."

Near the beginning of this article (Section Chomsky on the Faculty of Language), we noted that the notion of tacit knowledge does not figure in Chomsky's recent descriptions of his project (Chomsky, 2000, 2005; Hauser et al., 2002). Chomsky's account of the attained state of the faculty of language does not absolutely exclude the possibility that the subject has tacit knowledge of linguistic rules or principles – perhaps, tacit knowledge of a compositional semantic theory for her own language – embodied in some component of the mind/brain other than the faculty of language. But it is a significant question for future research whether appeal to a generative procedure can do the same work in philosophy

of language as the appeal to tacit knowledge of a compositional semantic theory.

A more general question in the same area is whether Chomsky's distinction between competence and performance can be retained if the attained state of the faculty of language is not to be conceived as a body of knowledge. Collins (2007) argues that the distinction does not depend on an epistemological – or even on a representational – understanding of competence. Linguistic cognition – and thence, linguistic behavior – depends on many component systems of the mind/brain. As Hauser et al. (2002) say, "a neuroscientist might ask: What components of the human nervous system are recruited in the use of language in its broadest sense? Because any aspect of cognition appears to be, at least in principle, accessible to language, the broadest answer to this question is, probably, 'most of it'…this conception is too broad to be of much use." Thus, reflecting normal scientific practice, the study of the faculty of language focuses on one component of the mind/brain "under idealization of its integration with other systems." The notion of competence "is intended to abstract one system [the faculty of language] out of an ensemble of systems whose integration underlies performance" (Collins, 2007).

Tacit Knowledge and Concept Possession

The notion of tacit knowledge also figures in philosophical theories about possession of concepts. One possible proposal in this area would be that possession of some concepts – particularly, concepts of logical operations such as conjunction and disjunction, but not only those – consists in tacit knowledge of logical or proper rules of inference that codify the inferential potential that is lodged in those concepts. Another possible proposal would be that possession of some concepts consists in tacit knowledge of definitions of those concepts.

Peacocke (2008) calls a state of tacit knowledge that is required for possession of a concept an 'implicit conception' and he proposes that, at least in some cases, an implicit conception is or includes a state of tacit knowledge of the contribution to truth conditions that is made by the concept. For example, in the case of the logical concept of (inclusive) disjunction, expressed by 'or,' this would not be tacit knowledge of rules of inference, but tacit knowledge that any proposition (thought content) of the form '*A or B*' is true if and only if either *A* is true or *B* is true. Such a state of tacit knowledge makes a common contribution to the explanations of the subject's evaluation for truth of many specific thought contents involving the concept (in the case of the example, a common contribution to the subject's truth-evaluation of many specific contents of the '*A or B*' form). The state of tacit knowledge of truth conditions can also contribute to the explanation of the subject's assessment of a principle or rule of inference as being valid or invalid. This proposal helps to explain the fact that even a primitive (nonderived) principle or rule of inference for a concept is susceptible of rational justification, a kind of justification that draws on the subject's grasp of the concept in question.

Here it is important to distinguish between two conditions that a principle might meet. First, it might be that personal-level acceptance of the principle can be justified by drawing on one's grasp of a particular concept. Second, it might be that personal-level acceptance of the principle is necessary for possession of the concept. Many principles that meet the first condition do not meet the second condition. In fact, some philosophers doubt that there are any principles at all that meet the second condition (Williamson, 2006). Suppose, though, that we were to allow (as conceptual-role theories of concept possession maintain) that, for a given concept, there are principles or rules of inference that do meet the second condition. That is, personal-level acceptance of, or commitment to, those principles or rules is necessary for possession of the concept. Peacocke (2008) argues that this could not provide a complete account of concept possession.

His argument appeals to "the phenomenon of new principles." There can be a principle involving a particular concept, which has two properties. First, the principle meets the first condition in the previous paragraph; that is, a subject can rationally justify acceptance of the principle by drawing on his grasp of the concept. Second, the principle is new, in the sense that it does not follow from the principles or rules already mentioned in the putative conceptual-role account of possession of the concept. These properties are together problematic for the conceptual role account because, according to that account, possession of the concept consists in personal-level acceptance of or commitment to principles or rules from which the new principle does not follow.

Rational justification of new principles has a ready explanation, however, if concept possession consists (in whole or in part) in tacit knowledge of the concept's contribution to truth conditions. Peacocke (2008) draws a parallel between explaining rational justification of new principles and explaining understanding of new sentences: "I am inclined to think that the Phenomenon of New Principles is as decisive an argument against personal-level conceptual role theories as the phenomenon of understanding sentences one has never encountered before is decisive against theories of meaning which do not proceed compositionally."

Implicit Rules in Connectionist Networks

A third point at which the notion of tacit knowledge figures in philosophical theory concerns the information processing that takes place in connectionist networks. A network consists of units, organized in layers, and weighted connections between units in successive layers. Input and output representations are patterns of activation across units in the input and output layers, respectively. Processing in a network is a matter of activation being passed from the input units to the output units, usually via a layer of hidden (i.e., neither input nor output) units. The activation of units in each successive layer is determined by the activation of units in the previous layer and the weights on the connections between the two layers. Thus, the precise weights on the connections are crucial for the network's performance of its task (e.g., generating an output representation of the correct pronunciation, given an input representation of a letter string). In this sense, a network's 'knowledge' about its task domain is embodied in the weights on its connections.

In a pioneering contribution to the development of connectionist models of cognitive processes, McClelland et al. (1986) highlighted a difference between connectionist models and earlier models by suggesting that connectionist networks "may provide a mechanism sufficient to capture lawful behavior, without requiring the postulation of explicit but inaccessible rules." In another influential paper, Rumelhart and McClelland (1986) said, "Using knowledge in processing is no longer a matter of finding the relevant information in memory and bringing it to bear; it is part and parcel of the processing itself." These (and many similar) passages draw attention to significant features of connectionist networks. But it is important to see that they do not, by themselves, reveal a difference between processing in networks and processing that is guided by tacitly known rules.

Earlier (Section Rule-Explicit and Rule-Implicit Systems), we distinguished between rule-explicit and rule-implicit systems. Tacit knowledge of a rule may involve explicit storage: a representation that encodes the rule must be searched for and accessed before the tacit knowledge can be used. We also noted earlier that explicit storage is not sufficient for explicit knowledge. Tacit knowledge of rules may involve explicit storage of information that is not available to the subject for verbal report and not accessible to the subject's consciousness. This kind of tacit knowledge ('finding the relevant information in memory and bringing it to bear'; 'explicit but inaccessible rules') can, indeed, be contrasted with the way in which knowledge about a task domain is embodied in a connectionist network.

But a rule may be both tacitly known and 'wired in' to a system, because a component processor, no less than an explicitly stored representation, can make a common contribution to the content-involving explanations of many transitions between representations. Peacocke (2008) makes the same point about implicit conceptions. At the subpersonal level, the content of an implicit conception might be 'explicitly formulated in a language of thought' or 'grounded in the operation of a processor.' When a tacitly known rule is implicit or 'wired in,' it is well described as 'part and parcel of the processing.' Thus, the connectionist architecture of units and activations, connections and weights, does not exclude the possibility that tacit knowledge of rules may be embodied in networks.

If we consider again the very simple example of transitions from representations of three-letter strings to representations of their pronunciations, it is certainly possible to describe a connectionist network in which tacit knowledge of each of the 15 letter–sound rules is embodied in a particular set of weighted connections, so that the causal-explanatory structure is like that in Syst. And it is possible to describe a different network in which tacit knowledge of each of the 125 one-case pronunciation rules is embodied in a particular set of weighted connections, so that the causal-explanatory structure is like that in List (Davies, 1995). With more realistic tasks (e.g., when a proportion of the letter strings have irregular or exception pronunciations), and with networks that are more typical of contemporary connectionist research (e.g., Plaut et al., 1996), it is usually not possible to identify component processors (sets of weighted connections) that embody letter–sound rules. Nevertheless, it may be possible to find theoretically significant commonalities in the processing of similar items, and the pattern of commonalities may be preserved across different networks (e.g., with different numbers of units and connections) that perform the same task (Shea, 2007).

Tacit Theory and Simulation in Everyday Psychological Understanding

A fourth example is provided by the debate between the theory theory and mental simulation approaches to everyday psychological understanding (Davies and Stone, 1995). On both sides of the debate, it is agreed that we have a capacity to predict, interpret, and explain the mental lives of other people. What is the basis of this capacity? According to the theory theory, we draw on a rich body of knowledge about psychological matters – a theory, in a suitably inclusive use of that term. According to the simulation theory, in apparent contrast, we identify with the other person in imagination and replicate or reenact aspects of the other person's mental life. If the other person is thinking about interest rates and the stock market, then we think about those same financial matters (we 'co-cognize' with the other person) rather than drawing on a rich body of psychological principles concerned with how people think about financial matters (Heal, 1986, 1998).

The theory theory faces objections if its claim is that ordinary people with a normal capacity for everyday psychological understanding have a rich body of explicit knowledge about psychological matters. Ordinary people are generally not able to articulate many principles of psychological theory: "[W]hy, one wonders, should it be so difficult to articulate laws if we appeal to them all the time in our interpretive practice?" (Goldman, 1989). The obvious response on behalf of the theory theory is to say that the rich body of knowledge is tacit knowledge rather than explicit knowledge. Thus, Stich and Nichols (1992) draw attention to the 'dominant explanatory strategy' in cognitive science, which is to appeal to "an internally represented 'knowledge structure'…which serves to guide the execution of the capacity to be explained." On their account, the theory theory says that everyday psychological understanding is explained by the presence of a 'largely tacit psychological theory.'

A concern about the shape of the debate may then arise. The theory theory and the simulation theory were supposed to be competing accounts of the basis of our capacity for everyday psychological understanding. But, depending on what it takes for a theory to be tacitly known, might not the simulation theory turn out to be consistent with – even a version of – the tacit theory theory? Blackburn (1992) gives expression to this concern that the tacit theory theory may be so undemanding as to subsume the supposed simulation alternative: "If we are good at something…then we can be thought of as making tacit (very tacit) use of some set of principles that could, in principle, provide a description of a device…that is also good at it."

The concern would be justified if tacit knowledge were to be conceived in the way that Blackburn envisages. The debate between the tacit theory theory and the mental simulation theory would collapse if the attribution of tacit knowledge of principles to a person with a particular capacity required no

more than that explicit knowledge of the principles should suffice for having the capacity, or no more than that behavioral manifestations of the capacity should conform to the principles, or no more than that it should be as if the person knew the principles.

The account of tacit knowledge given in this article (Section An Account of Tacit Knowledge) is certainly more demanding than those trivializing accounts. But Heal (1994) has argued that, even on that account of tacit knowledge, the mental simulation theory turns out to be a version of the tacit theory theory. In fact, she argues more generally that, if a mechanism is used to simulate the operation of other mechanisms of the same kind so as to permit predictions about them, then the mechanism used for simulation embodies tacit knowledge of theoretical principles about how mechanisms of that kind operate.

Suppose (to take an extremely simple example) that a subject, S, begins with the information about another person, O, that O believes the proposition that interest rates are falling (*p*). Suppose, too, that S wonders whether, if the question of the truth of the disjunctive proposition that interest rates are falling or the stock market is rising (*p or q*) arises, O will believe that it is true. In order to generate a prediction about O, S co-cognizes with O and, in suppositional or imaginative mode, thinks that *p* and considers whether the proposition *p or q* is true. Drawing on her own tacit knowledge of the inference rule of '*or*'-introduction or her own tacit knowledge of the contribution made to truth conditions by the concept of (inclusive) disjunction, S concludes that *p or q* is true. Then, on the basis of this simulation exercise, S predicts that O will believe *p or q*.

Clearly, S could proceed in the same way in a multitude of actual or potential cases. Beginning from the information that a subject X believes a proposition *A*, S would arrive at the prediction that X will believe a proposition of the '*A or B*' form if the question of that proposition's truth arises. Furthermore, S's transition would have the same causal explanation in each case. Thus, Heal (1994) argues, S has tacit knowledge of the general psychological principle that if X believes *A* then X will also believe *A or B* – just as the tacit theory theory would say.

Davies and Stone (2001) respond to Heal's argument with two points. First, the attribution to S of tacit knowledge of the psychological principle that if X believes *A* then X will also believe *A or B* is unmotivated. Simulation theorists, including Heal, agree that the use of mental simulation to arrive at psychological predictions about other people depends on tacit or explicit knowledge or acceptance of some principle linking the simulating subject's own thinking with other people's thinking. One example of a principle of this kind is (Sim–Other): If, within the scope of a simulation exercise that begins from the thought *C*, I arrive at the thought *D* then, if O believes *C* then *ceteris paribus* O will also believe *D*. The transition in S's simulation exercise from the thought that interest rates are falling (*p*) to the thought that interest rates are falling or the stock market is rising (*p or q*), together with S's acceptance of the principle (Sim–Other), already provides an explanation of S's proceeding from the information that O believes *p* to the prediction that O will believe *p or q*. No additional knowledge, tacit or explicit, about psychological matters is required.

The second point that Davies and Stone (2001) make is that the attribution to S of tacit knowledge of the psychological principle that if X believes *A* then X will also believe *A or B* is not licensed by the account of tacit knowledge. Heal allows that some component of S's subpersonal-level reasoning system embodies tacit knowledge about the inferential or truth-conditional properties of disjunction and she proposes that this same information-processing component also embodies tacit knowledge of the principle that if X believes *A*, then X will also believe *A or B*. But states of tacit knowledge figure in content-involving explanations and the component in question does not play the right causal-explanatory role to embody tacit knowledge of that principle. A state of tacit knowledge of the psychological principle would explain transitions between representations under their content-involving descriptions. These would have to be representations with contents of the forms 'X believes *A*' and 'X believes *A or B*.' But the information-processing component in question mediates transitions between states whose contents do not concern believing at all but instead (in our example) concern financial matters. The transition between representations with contents 'interest rates are falling' and 'interest rates are falling or the stock market is rising' is, of course, entirely appropriate to the causal-explanatory role of a state of tacit knowledge about the inferential or truth-conditional properties of disjunction. But it is not appropriate to the causal-explanatory role of a state of tacit knowledge of a psychological principle.

Tacit Knowledge in Psychology: The AGL Task

Half a century ago, Chomsky's notion of tacit knowledge of the rules of a generative grammar provided the theoretical background for empirical research on artificial grammar learning (AGL). Chomsky's early work on formal language theory also provided the main tool that has been used in this research.

In a typical AGL experiment, the grammar in question is a regular or finite state grammar. Chomsky (1956) described finite state grammars, phrase structure grammars, and transformational grammars, and showed that finite state grammars cannot capture the syntax of English. Chomsky and Miller (1958) provided a more extensive formal account of finite state grammars and the languages that they generate. Although finite state grammars do not include a grammar for English, they illustrate well enough the core notions that Chomsky associates with the faculty of language: a finite vocabulary and a procedure that generates infinitely many finite complex structures (expressions) from the vocabulary items.

Finite State Grammars

A finite state grammar can be represented graphically by a state diagram (a diagram for a finite state automaton); an example is provided in **Figure 1**. There is a finite set of states – in the example, four states (S_1, S_2, S_3, S_4) apart from the start and end states (S_0 and $S_{0'}$); and arrows indicate possible transitions between states. The vocabulary is a finite set of items – in the example, the set of letters {P, S, T, V, X}. Beginning in the start state, S_0, each successive transition adds

a letter to the string. The state that the system is in fully determines which letters can be added to the string (e.g., in state S_2, S or X can be added). No memory for earlier transitions, or for letters already in the string, is required.

A string of letters is an expression of the language generated by the grammar just in case it corresponds to a path through the state diagram, from the start state to the end state. Thus, the strings TTS and VVS are expressions of the language generated by the grammar corresponding to the state diagram in Figure 1, as are TPTS, TPPTS, TPPPTS, and so on, and TTXVPS, TTXVPXVPS, TTXVPXVPXVPS, and so on.

A grammar is canonically specified by a set of rules and the rules of a finite state grammar have the form: A → *a* B, where A denotes a state other than the end state, *a* denotes a vocabulary item, and B denotes a state. The rules of a finite state grammar correspond to the possible transitions in the state diagram that represents the grammar graphically. Each rule specifies the state in which the transition applies, the vocabulary item that is added to the string, and the state following the transition. Thus, the rules of the grammar represented by the state diagram in Figure 1 include: $S_0 \rightarrow T\,S_1$, $S_0 \rightarrow V\,S_3$, $S_1 \rightarrow P\,S_1$, $S_1 \rightarrow T\,S_2$, $S_2 \rightarrow X\,S_3$, $S_2 \rightarrow S\,S_{0'}$ and so on. The rules can be read as saying that the symbol on the left of the arrow can be replaced by the symbols on the right (they are 'rewrite' rules). A string, *w*, is said to be generated by the grammar if and only if it is possible to begin with the symbol for the start state and reach *w* (followed by the symbol for the end state) by a finite sequence of rule applications. Thus, for example, a derivation of the string TTS would proceed from S_0 to $T\,S_1$ (by the first rule above), then to $TT\,S_2$ (by the fourth rule), and finally to $TTS\,S_{0'}$ (by the sixth rule). For the purpose of displaying the derivations of strings, a more perspicuous notation would use the null symbol to denote the end state. Rules would have the form A → *a* B (as before) or A → *a*, and the sixth rule would be simply $S_2 \rightarrow S$.

Learning and Classification

In a typical AGL experiment, subjects learn to classify letter strings as being grammatical or ungrammatical. In general, there are many ways in which subjects may learn to classify items as belonging or not belonging to some category. In one kind of case, subjects may be provided with a principle or rule of classification. That is, a teacher may make a rule explicit and the subjects may acquire explicit knowledge of the rule and use that explicit knowledge in making their classifications.

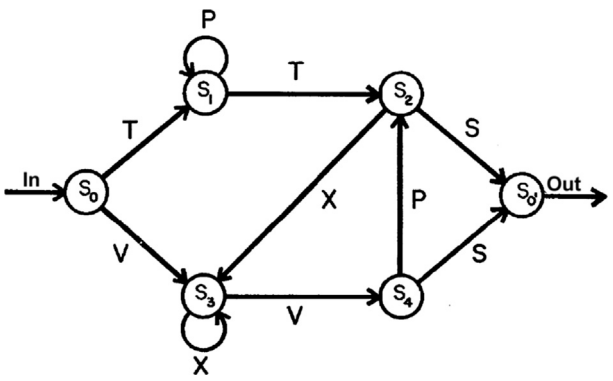

Figure 1 Diagram of a finite state grammar.

In other cases, subjects may simply be provided with a set of exemplars of the category to study or memorize. After a period of exposure to the exemplars, subjects may achieve performance on the classification task that is significantly better than would be predicted if they were guessing (or tossing a coin). In principle, their classifications might even be reliably correct. A subject who has attained this level of performance may also be able to state a rule to which his or her classification performance reliably conforms. That is, even without being explicitly told any rule of classification, the subject may acquire explicit knowledge of a rule. In such a case, it may be (though it need not be) that it is because the subject has acquired explicit knowledge of the rule that his or her classifications are reliably correct.

It may, however, happen that, following the period of exposure to exemplars, the subject achieves classification performance that substantially, or even reliably, conforms to a rule of classification, although the subject is quite unable to state the rule. Indeed, the subject may be unable to provide any illuminating account of how he or she is making the classification. Such a case fits Polanyi's (1967) conception of tacit knowing – knowing more than one can tell. The subject is able to – has learned to – make the classification, but cannot tell – cannot make explicit – how to make the classification.

Cases of this kind are sometimes described as examples of implicit learning: the subject "learns about the structure of a fairly complex stimulus environment, without necessarily intending to do so, and in such a way that the resulting knowledge is difficult to express" (Dienes and Berry, 1997). Implicit learning is an unconscious and unintentional process (Shanks, 2005) – in contrast to learning by testing or evaluating explicitly formulated hypotheses, for example. Because the description of implicit learning seems to fit learning (or, more generally, acquisition) of our first language, it is natural to ask whether implicit learning results in Chomskyan tacit knowledge of a rule (or a set of rules) of classification.

Implicit Learning and Tacit Knowledge of Rules

At the beginning of a seminal paper in which the term 'implicit learning' was introduced, Reber (1967) wrote: "In recent years, the model of the verbal organism as an imitative and generalizing mechanism has been largely replaced by a model that characterizes him as a 'sentence generating machine' who has learned a 'generative grammar' in some implicit fashion." In the learning phase of his first AGL study, Reber (1967, Experiment 2) presented subjects with 20 strings generated by the finite state grammar in Figure 1. In this phase, there was no mention of grammars or rules and the subjects were required simply to memorize the presented strings. At the beginning of the test phase of the study, subjects were told that the 20 strings had been generated by a set of grammatical rules, but they were provided with no further information about the rules. Subjects were then asked to classify new strings as either grammatical or ungrammatical. Each new string was presented twice and the striking finding was that 79% of the subjects' classification responses were correct (with 69% of the new strings classified correctly on both presentations).

A separate experiment (1967, Experiment 1) showed that, following an extensive learning phase (in which, again, there

was no mention of grammars or rules), subjects were unable to answer any questions about what the rules governing the strings might be. Unless they were given considerable prompting, subjects were unable even to say which letters the strings could begin with (T or V) or end with (only S). Reber concluded that the learning phase of the experiment provided subjects with abstract information about the rule-governed structure of the presented strings; and that this information was used to classify new strings in the test phase, although no verbal account of the rules could be elicited from subjects. In short, Reber's view was that, through a process of implicit learning, tacit knowledge of rules was acquired.

Skepticism about Implicit Learning

It has been argued, however, that it would be possible for subjects to achieve the better-than-chance classification performance that is typically demonstrated in the test phase of an AGL experiment without acquiring knowledge of the rules of the grammar used to generate the strings presented in the learning phase. One proposal is that information about pairs of adjacent letters (bigrams) that occur in the presented strings (perhaps along with information about allowable first and last letters) would be sufficient to account for subjects' classification performance. Furthermore, the results of a recognition memory test support the hypothesis that, in the learning phase, subjects acquire explicit knowledge about which bigrams occurred in the presented strings. Thus, an alternative to Reber's view is that what is acquired during the learning phase and used during the test phase is not tacit knowledge of abstract rules but explicit knowledge of specific fragments (Perruchet and Pacteau, 1990). In a review of research on implicit learning, Shanks (2005) supports this alternative proposal: "participants merely focus on and learn about small letter chunks. In the test stage they endorse (i.e., call grammatical) any item composed mainly of familiar letter pairs."

Here it is relevant that there are findings of transfer of classification performance to strings that are generated by the same finite state grammar, but from a new set of letters (Reber, 1969), or even to strings of stimuli presented in a different modality from those in the learning phase (e.g., letters vs tones; Altmann et al., 1995). Reber explains transfer in terms of abstract knowledge (even if not knowledge of the basic rules of the grammar) acquired during the learning phase. An alternative proposal is that transfer can be explained in terms of abstraction or analogical reasoning in the test phase from explicit knowledge, about specific fragments (e.g., bigrams or trigrams) or about specific complete exemplars, acquired in the learning phase. Shanks (2005) argues that the balance of evidence favors the alternative proposal and concludes: "In a nutshell, performance in AGL experiments appears to be fully explained without the need to impute unconscious rule-learning."

AGL in Patients with Amnesia

In comparing Reber's view with alternatives, we have contrasted implicit abstract knowledge with explicit specific knowledge. But the implicit versus explicit distinction and the abstract versus specific distinction are not to be conflated.

Patients suffering from amnesia cannot (by definition) retain new explicit knowledge but, in AGL experiments, they show very similar classification performance to that of age-matched controls (Knowlton and Squire, 1994, 1996). This finding is enough to rebut any suggestion that acquisition of explicit knowledge in the learning phase is essential for better-than-chance classification performance in the test phase of AGL experiments. But the fact that patients with amnesia cannot rely on explicit knowledge acquired in the learning phase to classify new strings presented in the test phase does not settle the question whether the patients rely on abstract or specific information or both.

Knowlton and Squire showed that, in both patients and controls, classification performance was influenced by the occurrence in test strings of bigrams and trigrams (chunks) that occurred frequently in strings presented in the learning phase, despite the fact that the patients with amnesia scored only marginally above chance on a recognition memory test for chunks that had appeared in the learning phase. Classification performance was also influenced by whether the test strings conformed to the rules of the grammar and there was some transfer of performance to test strings composed from a new set of letters (though at a lower level than with test strings composed from the original set of letters). At least in the case of patients with amnesia, it is not plausible that transfer of classification performance depends on abstraction or analogy in the test phase, based on explicit knowledge about chunks or complete strings retained from the learning phase. Rather, it seems that patients acquire specific information about chunks and complete strings presented in the learning phase, and also more abstract information that supports transfer.

Knowlton and Squire's (1994, 1996) findings do not exclude the possibility that subjects with unimpaired memory – and particularly, perhaps, healthy younger subjects – may acquire in the learning phase, and use in the test phase, information that is available for verbal report. But the findings, and the similar performance of patients and age-matched controls, certainly suggest that, even if in normal healthy subjects some of the relevant information is available for verbal report, not all of it is. As Reber (1989) puts the point: "The implicitly acquired epistemic contents of mind are always richer and more sophisticated than what can be explicated."

Conscious and Unconscious Knowledge

It will be clear from the previous section that the dialectical context of AGL experiments is different from that of typical appeals to tacit knowledge in philosophy. Philosophical appeals to tacit knowledge typically occur in contexts where it is not in dispute that relevant explicit knowledge is lacking. In AGL research, in contrast, the claim that what has been acquired in the learning phase is implicit or tacit knowledge, rather than explicit knowledge, is highly contested. It is also a feature of AGL research that the question of what has been acquired in the learning phase is often formulated in terms of conscious knowledge versus unconscious knowledge – a contemporary version of the explicit knowledge versus implicit knowledge contrast. Thus, considerable importance

attaches to the empirical assessment of conscious versus unconscious knowledge.

Two Criteria for a Test of Conscious Knowledge

Shanks and St John (1994) argue that, in studies that aim to demonstrate that unconscious knowledge (e.g., Chomskyan tacit knowledge of rules) has been acquired through implicit learning, two criteria must be met in order to exclude the possibility that what subjects have acquired is really conscious knowledge. First, if the subjects' awareness of a particular kind of information is tested, then this must be the same kind of information that is drawn on in the subjects' performance: the information criterion. For example, in Reber's (1967) study, subjects' conscious knowledge of the rules governing the presented strings was tested. So it is important that it should be information about those rules that was drawn on in subjects' classification of new strings as grammatical or ungrammatical. Clearly, no tacit knowledge or implicit learning is demonstrated if one kind of information (e.g., about bigrams that occurred in strings presented in the learning phase) is drawn on in subjects' performance but subjects are found to be unaware of some other kind of information (e.g., about the rules of the finite state grammar).

The second criterion is that tests of subjects' conscious knowledge should be sufficiently sensitive to reveal all of the relevant conscious knowledge: the sensitivity criterion. Here it is relevant that, in research on memory, both recall and recognition have been regarded as explicit tests of memory (Schacter, 1989), with recognition being generally more sensitive than recall. Thus, if subjects' performance on a classification test were to draw on information about grammatical rules, then simply asking subjects what the rules are (a test of free recall) would meet the information criterion, but might be relatively insensitive to subjects' actual conscious knowledge of rules. For example, subjects might not volunteer a statement of conscious knowledge about which they were not confident. If a test of conscious knowledge used in an AGL experiment does not meet both the information criterion and the sensitivity criterion then task-relevant conscious knowledge acquired during the learning phase may be underestimated – and implicit learning may be correspondingly overestimated.

Exhaustiveness and Exclusiveness

Lessons for research on implicit learning have also been drawn from Reingold and Merikle's (1990) discussion of research on unconscious perception. They note that one requirement for an adequate measure of conscious awareness is that it should be exhaustive (a requirement very similar to the sensitivity criterion). Given a finding that performance apparently outruns conscious perception (in Reingold and Merikle's case) or conscious knowledge (in the AGL case), it cannot be concluded that there is perception or learning without awareness if the test of conscious awareness is not exhaustive.

A second requirement for a test of conscious awareness is that it should be exclusive; that is, results on the test should be influenced only by conscious, and not also by unconscious, perception or knowledge. Given a finding that performance apparently does not outrun conscious perception or knowledge, it cannot be concluded that there is no perception or learning without awareness if the test of conscious awareness is not exclusive (Reingold and Merikle, 1990). In the AGL case, specifically, results on a test of conscious awareness of what has been learned should be influenced only by conscious knowledge, and not contaminated by unconscious knowledge.

As Shanks and St John (1994) in effect pointed out, if a test of conscious knowledge is not exhaustive then it may underestimate conscious knowledge and overestimate implicit learning. Now we should add that if a test is not exclusive, then it may overestimate conscious knowledge and underestimate implicit learning. Different theorists put more stress on the exhaustiveness (sensitivity) requirement or on the exclusiveness requirement and this may reflect, in part, their view as to whether conscious or unconscious processing enjoys theoretical primacy. One influential view is that, in human beings, conscious learning is the default case and 'unconscious learning has not yet been satisfactorily established' (Shanks and St John, 1994; see also Shanks, 2005). Reber, on the other hand, defends the primacy of the implicit (Reber, 1990).

Validating a Test of Conscious Knowledge

Unless a test of conscious knowledge is, and is known to be, both exhaustive and exclusive, it cannot, by itself, be used to adjudicate evenhandedly a putative case of implicit learning. We have already seen that a test of conscious knowledge based on free recall is unlikely to be exhaustive. Tests based on recognition memory (rather than recall), which we have mentioned as being more sensitive to conscious knowledge, may not be exclusive because responses may be contaminated by unconscious processes. For example, recognition judgments are influenced by perceptual fluency. Subjects are more likely to 'recognize' a test word falsely as having occurred on a previously studied list if their perception of the test word is facilitated by very brief prior presentation of the same word under conditions (masking) that preclude conscious perception (Jacoby and Whitehouse, 1989).

The fact that a recognition memory test of conscious knowledge is not exclusive illustrates the more general point that distinctions between psychological tasks (e.g., between the so-called explicit and implicit tests of memory) do not coincide with distinctions between psychological processes (Jacoby, 1991). Recognition has been regarded as an explicit test of memory, but the results of a recognition test may be influenced by unconscious, as well as by conscious, processes. More generally, we should not assume that an experimental task (particularly, a task used as a test of conscious knowledge) is 'process pure,' that is, that it involves only conscious processes or else only unconscious processes. There is no a priori guarantee that any particular test of conscious knowledge will be exhaustive and exclusive.

Reingold and Merikle (1990) argue for an approach on which a putative test (or 'operational definition') of conscious awareness is validated by converging empirical evidence. In their view, the theoretical importance of the distinction between conscious and unconscious processes lies in the existence of qualitative differences between the two. Sometimes, indeed, conscious and unconscious processes are not just qualitatively different, but opposed to each other. For example, the

recognition memory experiment briefly mentioned two paragraphs back involved two conditions. In one (unaware condition, described above), presentation of a test word was preceded by a very brief masked presentation of the same word. In the aware condition, in contrast, presentation of a test word was preceded by a presentation of the same word that was long enough to allow conscious perception. As described above, subjects in the unaware condition were more likely to 'recognize' the test word falsely as having occurred on a previously studied list (by comparison with an appropriate baseline condition in which there was no prior presentation of the test word). But subjects in the aware condition were less likely (than in the baseline condition) to 'recognize' the test word falsely as having occurred on the previously studied list.

Thus, there is the prospect that a putative test of conscious knowledge can be validated in the context of psychological theory to the extent that the test correctly predicts processing of qualitatively different – perhaps opposed – kinds. In the next section, we describe some AGL research that adopts the approach recommended by Reingold and Merikle.

Conscious Knowledge in the Context of Psychological Theory

In the learning phase of an AGL experiment, subjects may acquire knowledge about the structure of the presented strings. This may be knowledge of rules or knowledge about individual strings. Possible examples of knowledge of rules include not only knowledge of the rules of the finite state grammar that was used to generate the strings, but also knowledge of simpler rules such as that strings can begin with T or V, that all strings end with S, and that P can be repeated. Possible examples of knowledge about individual strings presented in the learning phase include knowledge that the string VXXVS was presented, that a string including the bigram TS was presented, and that a string including the trigram VPX was presented. (The distinction between the two kinds of knowledge is not sharp. If a string beginning with VX was presented in the learning phase, then strings can begin with VX.) The knowledge acquired in the learning phase may be drawn on in the test phase as subjects make judgments about test strings – judgments that the strings are grammatical (that is, conform to the rules that were used to generate the strings presented in the learning phase) or ungrammatical.

Conscious and Unconscious Judgment Knowledge

Dienes and Scott (2005) distinguish between structural knowledge acquired during the learning phase and what they call 'judgment knowledge' arrived at and deployed in the test phase – for example, the knowledge that the string VVPXXVS, presented in the test phase, is grammatical. The aim of their study was to assess knowledge as conscious or unconscious, in the context of psychological theories about conscious and unconscious processes.

Dienes and Scott commend two confidence criteria for unconscious knowledge, the guessing criterion and the zero-correlation criterion (see also Dienes, 2008, 2012). A subject who is asked to judge whether strings of letters are grammatical or not may, on some trials, believe him- or herself to have no basis for the judgment and, consequently, he or she may have no confidence in his or her judgment; he or she may describe him- or herself as guessing. (Strictly speaking, he or she may have confidence of 50% in his or her judgment – he or she might as well be tossing a coin.) If the subject's guessing responses are correct at a rate that is better than chance, then this is evidence (that is, prima facie but defeasible evidence) of unconscious judgment knowledge (according to the guessing criterion).

On other trials, the subject may believe him- or herself to be knowledgeable to some extent and may have some degree of confidence in his or her judgment. If the subject's guessing responses and these more confident responses together reveal a correlation between accuracy and confidence, then this is evidence of some conscious judgment knowledge. If there is no correlation between accuracy and confidence but the subject's overall performance is better than chance then this is, again, evidence of unconscious judgment knowledge (according to the zero-correlation criterion).

Clearly, the use of the guessing criterion and the zero-correlation criterion may provide evidence of both unconscious and conscious judgment knowledge. For example, a subject's guessing performance may be better than chance and his or her more confident judgments may be more likely to be correct than his or her guesses. Dienes and Scott (2005) investigated judgment knowledge by asking subjects in the test phase of an AGL experiment to classify each string as grammatical or ungrammatical and then to indicate their degree of confidence in the classification judgment (from 50 to 100% confident). Overall, subjects' confidence ratings provided evidence of both unconscious and conscious judgment knowledge.

Conscious and Unconscious Structural Knowledge

In principle, it would seem that conscious or unconscious judgment knowledge and structural knowledge could be combined in three ways. In the case of guessing, there is no conscious judgment knowledge or conscious structural knowledge. But conscious judgment knowledge can be combined with either unconscious or conscious structural knowledge. When a subject's judgment knowledge is conscious but his or her structural knowledge is unconscious, the phenomenology may be that of intuition.

Dienes and Scott (2005) investigated structural knowledge by asking subjects to indicate the source of their judgment knowledge, choosing between four options: guess (it seemed to the subject that the judgment had no basis whatsoever); intuition (the subject had some confidence in the judgment but had no idea why it was right); rules (the subject felt that he or she based the judgment on some rule or rules acquired in the learning phase – rules that he or she could state if asked); and memory (the subject felt that he or she based the judgment on memory for particular items or parts of items from the learning phase). The 'guess' and 'intuition' responses provided evidence that the subject's structural knowledge was unconscious, while the 'rules' and 'memory' responses provided evidence that it was at least partly conscious. Some more recent studies include an additional potential source of

subjects' judgment knowledge: familiarity (the subject felt that he or she based the judgment on a feeling of familiarity but had no idea what the familiarity itself was based on). The 'familiarity' response, like the 'intuition' response, provides evidence of unconscious structural knowledge (Scott and Dienes, 2008; see Dienes et al., 2011, for a review).

It is independently plausible on theoretical grounds that the acquisition and use of conscious structural knowledge would make demands on working memory and executive function. Thus, having subjects perform an executive working memory load task (e.g., generating random numbers) during the learning phase of an AGL experiment should disrupt the acquisition of conscious structural knowledge by comparison with unconscious structural knowledge. If 'rules'/'memory' responses indicate conscious structural knowledge, then one can predict that subjects who performed a load task during the learning phase should be less likely (than subjects who did not perform the load task) to attribute the source of their judgments in the test phase to rules or memory. This prediction was confirmed in Dienes and Scott's (2005) study and, to that extent, the use of the 'rules'/'memory' versus 'guess'/'intuition' contrast to distinguish between conscious and unconscious structural knowledge was validated.

The study by Dienes and Scott (2005) investigated the distinction between conscious and unconscious structural knowledge in the context of psychological theory about working memory and executive function. But the load task manipulation did nothing to validate the distinction between conscious and unconscious judgment knowledge.

The Serial Reaction Time Task

Converging evidence to support the distinction between conscious and unconscious judgment knowledge has come from a different implicit learning paradigm, the serial reaction time (SRT) task (Destrebecqz and Cleeremans, 2001). In the learning phase, subjects see a sequence of stimuli on a computer screen. Each stimulus appears at one of several (e.g., four) possible locations on the screen and the subject's task is simply to press a key corresponding to the stimulus location as quickly as possible. In fact, the sequence of stimuli follows a repeating pattern, in which the locations are determined by second-order conditional rules. That is, each position is determined by the previous two positions (e.g., positions 4 and then 2 are always followed by position 3). Learning is revealed in the pattern of reaction times, which decrease as the learning phase progresses, increase sharply when a new sequence is introduced, and then drop back to the lower level when the original sequence returns.

In the test phase, there are two conditions, motivated by Jacoby's (1991) 'process dissociation framework.' The leading idea is that the use of conscious knowledge can be controlled, whereas the use of unconscious knowledge is automatic. In one condition of the test phase (inclusion), subjects are asked to generate a sequence that resembles the training sequence as much as possible. In this condition, automatic use of unconscious knowledge is facilitatory. Subjects can perform the task by drawing on either conscious or unconscious knowledge about which position would be a legal continuation of the sequence (e.g., that 3 would be a legal continuation of …42). Consequently, legal continuations in the inclusion condition do not provide evidence that distinguishes between conscious and unconscious knowledge acquired during the learning phase.

In a second – theoretically crucial – condition (exclusion), subjects are asked to generate a sequence that avoids the regularities of the training sequence. In this condition, controlled use of conscious knowledge and automatic use of unconscious knowledge are set against each other. To perform the task, subjects need to make controlled use of their conscious knowledge about which position would be a legal continuation. Specifically, they need to continue in a way that avoids that position. In this condition, automatic use of unconscious knowledge interferes with, rather than facilitates, performance of the exclusion task. Thus, legal continuations in the exclusion condition provide evidence of unconscious knowledge acquired during the learning phase. (Relevant evidence would include the proportion of legal continuations being above a suitable baseline and there being no fewer legal continuations in the exclusion than in the inclusion condition.)

The results of experiments using the SRT task do not all point in the same direction. But, given appropriate conditions in the learning phase, evidence from the exclusion condition in the test phase does support the hypothesis that unconscious knowledge about legal continuations is acquired (Destrebecqz and Cleeremans, 2001; Fu et al., 2008). Furthermore, it can be argued that the relevant distinction for the application of Jacoby's (1991) process dissociation framework to sequence generation in the test phase of the SRT task is not between conscious and unconscious structural knowledge, but between conscious and unconscious judgment knowledge. Specifically, the relevant distinction is between intuition and guessing, because conscious judgment knowledge that is based on unconscious structural knowledge can still be controlled (Fu et al., 2010).

Summary and Continuing Research

When the distinction between explicit and implicit knowledge is interpreted as that between conscious and unconscious knowledge, empirical research using the AGL task can be seen as a chapter in the scientific study of consciousness. In this research, attributions of tacit knowledge of rules, and of unconscious knowledge more generally, have been contested and various requirements for a test of conscious knowledge have been proposed.

There is no test of conscious knowledge for which we have a guarantee that it will meet the demanding requirements of exhaustiveness and exclusiveness – revealing all the relevant conscious knowledge while remaining uncontaminated by unconscious knowledge. But putative tests may be validated by converging evidence in the context of psychological theory. Dienes and colleagues have proposed tests for conscious judgment knowledge (based on subjects' confidence ratings) and conscious structural knowledge (based on subjects' source attributions for their judgments). As briefly described in the preceding section, they have taken steps to validate the tests by appeal to converging evidence from (respectively) the exclusion condition of an experiment based on Jacoby's

(1991) process dissociation framework and an executive working memory load task (see Dienes, 2012, for a review).

Recent and continuing research on implicit learning, using the AGL task, is investigating new measures of conscious awareness such as wagering (Persaud et al., 2007; Dienes and Seth, 2010), and comparing different measures (Wierzchoń et al., 2012). Research also continues on feelings of familiarity and their relationship to perceptual fluency (Scott and Dienes, 2008, 2010; Dienes et al., 2011) and the AGL paradigm has been adapted to investigate cross-cultural differences in unconscious processes (Kiyokawa et al., 2012). In other work, researchers have developed implemented computational models of implicit learning tasks, including the AGL and SRT tasks (see Cleeremans and Dienes, 2008, for a review).

Finally, the methods of AGL are being used, in conjunction with formal language theory, to investigate learning – particularly, implicit learning – of grammars with greater expressive power than regular (that is, finite state) grammars (Rohrmeier et al., 2012). (Recall that Chomsky (1956) showed that finite state grammars cannot capture the syntax of English.) Some of this research extends to AGL in preverbal infants and nonhuman animals and addresses the questions whether the human capacity to acquire supraregular grammars is biologically distinctive and what its neural substrate might be (Fitch et al., 2012; Fitch and Friederici, 2012).

Acknowledgment

I am grateful to Zoltan Dienes, David Lobina, and Crysostomos Mantzavinos for their comments and advice.

See also: Meaning and Rule Following; Reference and Representation: Philosophical Aspects.

Bibliography

Altmann, G.T.M., Dienes, Z., Goode, A., 1995. Modality independence of implicitly learned grammatical knowledge. Journal of Experimental Psychology: Learning, Memory, and Cognition 21, 899–912.

Baker, G.P., Hacker, P.M.S., 1984. Language, Sense and Nonsense: A Critical Investigation into Modern Theories of Language. Blackwell, Oxford.

Bennett, M.R., Hacker, P.M.S., 2003. Philosophical Foundations of Neuroscience. Blackwell, Oxford.

Blackburn, S., 1992. Theory, observation and drama. Mind & Language 7, 187–203 (Reprinted in: Davies and Stone (Eds.), 1995).

Burge, T., 2003. Perceptual entitlement. Philosophy and Phenomenological Research 67, 503–548.

Chomsky, N., 1956. Three models for the description of language. IRE Transactions on Information Theory 2 (3), 113–124.

Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.

Chomsky, N., 1986. Knowledge of Language: Its Nature, Origin, and Use. Praeger, New York.

Chomsky, N., 2000. New Horizons in the Study of Language and Mind. Cambridge University Press, Cambridge.

Chomsky, N., 2005. Three factors in language design. Linguistic Inquiry 36, 1–22.

Chomsky, N., Miller, G.A., 1958. Finite state languages. Information and Control 1, 91–112.

Cleeremans, A., Dienes, Z., 2008. Computational models of implicit learning. In: Sun, R. (Ed.), The Cambridge Handbook of Computational Psychology. Cambridge University Press, Cambridge.

Collins, J., 2004. Faculty disputes. Mind & Language 17, 503–533.

Collins, J., 2007. Linguistic competence without knowledge of language. Philosophy Compass 2, 880–895.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., Ziegler, J., 2001. DRC: a dual route cascaded model of visual word recognition and reading aloud. Psychological Review 108, 204–256.

Davidson, D., 1984. Inquiries into Truth and Interpretation. Oxford University Press, Oxford.

Davies, M., 1987. Tacit knowledge and semantic theory: can a five per cent difference matter? Mind 96, 441–462.

Davies, M., 1989. Tacit knowledge and subdoxastic states. In: George, A. (Ed.), Reflections on Chomsky. Blackwell, Oxford.

Davies, M., 1995. Two notions of implicit rules. Philosophical Perspectives. AI, Connectionism, and Philosophical Psychology, vol. 9, pp. 153–183.

Davies, M., 2000. Interaction without reduction: the relationship between personal and sub-personal levels of description. Mind and Society 1, 87–105.

Davies, M., Coltheart, M., Langdon, R., Breen, N., 2001. Monothematic delusions: towards a two-factor account. Philosophy, Psychiatry, & Psychology 8, 133–158.

Davies, M., Stone, T., 2001. Mental simulation, tacit theory, and the threat of collapse. Philosophical Topics 29, 127–173.

Davies, M., Stone, T. (Eds.), 1995. Folk Psychology: The Theory of Mind Debate. Blackwell, Oxford.

Dennett, D.C., 1969. Content and Consciousness. Routledge and Kegan Paul, London.

Dennett, D.C., 1983. Styles of mental representation. Behavioral and Brain Sciences 83, 213–226 (Reprinted in: The Intentional Stance. MIT Press, Cambridge, MA, 1987).

Destrebecqz, A., Cleeremans, A., 2001. Can sequence learning be implicit? New evidence with the process dissociation procedure. Psychonomic Bulletin & Review 8, 343–350.

Dienes, Z., 2008. Subjective measures of unconscious knowledge. Progress in Brain Research 168, 49–64.

Dienes, Z., 2012. Conscious versus unconscious learning of structure. In: Rebuschat, P., Williams, J.N. (Eds.), Statistical Learning and Language Acquisition. de Gruyter Mouton, Berlin.

Dienes, Z., Berry, D., 1997. Implicit learning: below the subjective threshold. Psychonomic Bulletin & Review 4, 3–23.

Dienes, Z., Scott, R., 2005. Measuring unconscious knowledge: distinguishing between structural knowledge and judgment knowledge. Psychological Research 69, 338–351.

Dienes, Z., Scott, R.B., Wan, L., 2011. The role of familiarity in implicit learning. In: Higham, P.A., Leboe, J.P. (Eds.), Constructions of Remembering and Metacognition: Essays in Honour of Bruce Whittlesea. Palgrave Macmillan, Basingstoke.

Dienes, Z., Seth, A., 2010. Gambling on the unconscious: a comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. Consciousness and Cognition 19, 674–681.

Dummett, M., 1991. The Logical Basis of Metaphysics. Duckworth, London.

Egan, A., 2008. Seeing and believing: perception, belief formation and the divided mind. Philosophical Studies 140, 47–63.

Evans, G., 1981. Semantic theory and tacit knowledge. In: Holtzmann, S.H., Leich, C.M. (Eds.), Wittgenstein: To Follow a Rule. Routledge and Kegan Paul, London (Reprinted in: Collected Papers. Oxford University Press, Oxford, 1985).

Fitch, W.T., Friederici, A.D., 2012. Artificial grammar learning meets formal language theory: an overview. Philosophical Transactions of the Royal Society B: Biological Sciences 367, 1933–1955.

Fitch, W.T., Friederici, A.D., Hagoort, P., 2012. Pattern perception and computational complexity: introduction to the special issue. Philosophical Transactions of the Royal Society B: Biological Sciences 367, 1925–1932.

Fodor, J.A., 1968. The appeal to tacit knowledge in psychological explanation. Journal of Philosophy 65, 627–640.

Fodor, J.A., 1983. The Modularity of Mind. MIT Press, Cambridge, MA.

Fodor, J.A., 1987. Psychosemantics: The Problem of Meaning in the Philosophy of Mind. Appendix, 'Why there still has to be a language of thought'. MIT Press, Cambridge, MA.

Fodor, J.A., Pylyshyn, Z.W., 1988. Connectionism and cognitive architecture: a critical analysis. Cognition 28, 3–71.

Fu, Q., Dienes, Z., Fu, X., 2010. Can unconscious knowledge allow control in sequence learning? Consciousness and Cognition 19, 462–474.

Fu, Q., Fu, X., Dienes, Z., 2008. Implicit sequence learning and conscious awareness. Consciousness and Cognition 17, 185–202.

Goldman, A.I., 1989. Interpretation psychologized. Mind & Language 4, 161–185 (Reprinted in: Davies and Stone (Eds.), 1995).

Hauser, M.D., Chomsky, N., Fitch, W.T., 2002. The faculty of language: what is it, who has it, and how did it evolve? Science 298, 1569–1579.

Heal, J., 1986. Replication and functionalism. In: Butterfield, J. (Ed.), Language, Mind and Logic. Cambridge University Press, Cambridge (Reprinted in: Davies and Stone (Eds.), 1995).

Heal, J., 1994. Simulation vs. theory theory: what is at issue? In: Peacocke, C. (Ed.), Objectivity, Simulation and the Unity of Consciousness: Current Issues in the Philosophy of Mind, Proceedings of the British Academy, vol. 83. Oxford University Press, Oxford.

Heal, J., 1998. Co-cognition and off-line simulation: two ways of understanding the simulation approach. Mind & Language 14, 477–498.

Jacoby, L.L., 1991. A process dissociation framework: separating automatic from intentional uses of memory. Journal of Memory and Language 30, 513–541.

Jacoby, L.L., Whitehouse, K., 1989. An illusion of memory: false recognition influenced by unconscious perception. Journal of Experimental Psychology: General 118, 126–135.

Kiyokawa, S., Dienes, Z., Tanaka, D., Yamada, A., Crowe, L., 2012. Cross cultural differences in unconscious knowledge. Cognition 124, 16–24.

Knowlton, B.J., Squire, L.R., 1994. The information acquired during artificial grammar learning. Journal of Experimental Psychology: Learning, Memory, and Cognition 20, 79–91.

Knowlton, B.J., Squire, L.R., 1996. Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. Journal of Experimental Psychology: Learning, Memory, and Cognition 22, 169–181.

Lewis, D., 1992. Meaning without use: reply to Hawthorne. Australasian Journal of Philosophy 70, 106–110.

Loar, B., 1981. Mind and Meaning. Cambridge University Press, Cambridge.

Mantzavinos, C., 2001. Individuals, Institutions, and Markets. Cambridge University Press, Cambridge.

Mantzavinos, C., North, D.C., Shariq, S., 2004. Learning, institutions, and economic performance. Perspectives on Politics 2, 75–84.

Matthews, R.J., 2006. Knowledge of language and linguistic competence. Philosophical Issues Philosophy of Language 16, 200–220.

McClelland, J.L., Rumelhart, D.E., Hinton, G.E., 1986. The appeal of parallel distributed processing. In: Rumelhart, D.E., McClelland, J.L., the PDP Research Group (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1. Foundations. MIT Press, Cambridge, MA.

Peacocke, C., 1989. When is a grammar psychologically real? In: George, A. (Ed.), Reflections on Chomsky. Blackwell, Oxford.

Peacocke, C., 2008. Truly Understood. Chapter 4, 'Implicit conceptions'. Oxford University Press, Oxford.

Perruchet, P., Pacteau, C., 1990. Synthetic grammar learning: implicit rule abstraction or explicit fragmentary knowledge? Journal of Experimental Psychology: General 119, 264–275.

Persaud, N., McLeod, P., Cowey, A., 2007. Post-decision wagering objectively measures awareness. Nature Neuroscience 10, 257–261.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S., Patterson, K., 1996. Understanding normal and impaired word reading: computational principles in quasi-regular domains. Psychological Review 103, 56–115.

Polanyi, M., 1967. The Tacit Dimension. Routledge and Kegan Paul, London.

Quine, W.V.O., 1970. Methodological reflections on current linguistic theory. Synthese 21, 386–398.

Reber, A.S., 1967. Implicit learning of artificial grammars. Journal of Verbal Learning and Verbal Behavior 6, 855–863.

Reber, A.S., 1969. Transfer of syntactic structure in synthetic languages. Journal of Experimental Psychology 81, 115–119.

Reber, A.S., 1989. Implicit learning and tacit knowledge. Journal of Experimental Psychology: General 118, 219–235.

Reber, A.S., 1990. On the primacy of the implicit: comment on Perruchet and Pacteau. Journal of Experimental Psychology: General 119, 340–342.

Reingold, E.M., Merikle, P.M., 1990. On the inter-relatedness of theory and measurement in the study of unconscious processes. Mind & Language 5, 9–28.

Rohrmeier, M., Fu, Q., Dienes, Z., 2012. Implicit learning of recursive context-free grammars. PLoS One 7 (10), e45885.

Rumelhart, D.E., McClelland, J.L., 1986. On learning the past tenses of English verbs. In: McClelland, J.L., Rumelhart, D.E., the PDP Research Group (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 2. Psychological and Biological Models. MIT Press, Cambridge, MA.

Ryle, G., 1949. The Concept of Mind. Hutchinson, London.

Schacter, D.L., 1989. Memory. In: Posner, M.I. (Ed.), Foundations of Cognitive Science. MIT Press, Cambridge, MA.

Schiffer, S., 1993. Actual-language relations. Philosophical Perspectives Language and Logic 7, 231–258.

Schiffer, S., 2006. Two perspectives on knowledge of language. Philosophical Issues Philosophy of Language 16, 275–287.

Scott, R.B., Dienes, Z., 2008. The conscious, the unconscious, and familiarity. Journal of Experimental Psychology: Learning, Memory, and Cognition 34, 1264–1288.

Scott, R.B., Dienes, Z., 2010. Fluency does not express implicit knowledge of artificial grammars. Cognition 114, 372–388.

Searle, J., 1990. Consciousness, explanatory inversion, and cognitive science. Behavioral and Brain Sciences 13, 585–595.

Shanks, D.R., 2005. Implicit learning. In: Lamberts, K., Goldstone, R.L. (Eds.), Handbook of Cognition. SAGE Publications Ltd, London.

Shanks, D.R., St John, M.F., 1994. Characteristics of dissociable human learning systems. Behavioral and Brain Sciences 17, 367–395.

Shea, N., 2007. Content and its vehicles in connectionist systems. Mind & Language 22, 246–269.

Smith, B.C., 2008. What remains of our knowledge of language? Reply to Collins. Croatian Journal of Philosophy 8, 57–76.

Stich, S., 1978. Beliefs and subdoxastic states. Philosophy of Science 45, 499–518.

Stich, S., Nichols, S., 1992. Folk psychology: simulation or tacit theory? Mind & Language 7, 35–71 (Reprinted in: Davies and Stone (Eds.), 1995).

Thornton, T., 2013. Clinical judgment, tacit knowledge, and recognition in psychiatric diagnosis. In: Fulford, K.W.M., Davies, M., Gipps, R.G.T., Graham, G., Sadler, J.Z., Stanghellini, G. (Eds.), The Oxford Handbook of Philosophy and Psychiatry. Oxford University Press, Oxford.

Wierzchoń, M., Asanowicz, D., Paulewicz, B., Cleeremans, A., 2012. Subjective measures of consciousness in artificial grammar learning task. Consciousness and Cognition 21, 1141–1153.

Williamson, T., 2006. Conceptual truth. Proceedings of the Aristotelian Society Supplementary Volume 80, 1–41.

Knowledge, Geography of

Peter Meusburger, Heidelberg University, Heidelberg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The geography of knowledge studies spatial variations in the generation, dissemination, and application of various categories of knowledge. Scholars in this area are interested in the places and spaces of knowledge generation and the impact of local milieus (environments) and spatial relations on the generation and diffusion of knowledge. They analyze the reasons for spatial disparities of knowledge and their consequences for society and economy.

Introduction

The geography of knowledge examines spatial variations in the generation, dissemination, and application of various categories of knowledge. Interest focuses on the spatiality of learning and problem solving and on the impact that places, spatial relations, local milieus, and political and socioeconomic structures have on the generation and dissemination of knowledge. Scholars analyze the reasons for spatial disparities of knowledge and their consequences for society and economy. Research focuses on relationships between knowledge and power, knowledge and competitiveness, knowledge and migration, the ethnic and cultural dimension of knowledge, and other issues related to spatial variations of knowledge. This article first discusses the spatiality of learning processes and some of the main reasons why spatial differences of knowledge are so persistent over time. It then turns to methodological issues of localizing knowledge, and introduces various categories of knowledge. In its main part it focuses on the process of communicating different categories of knowledge from A to B; coalitions between knowledge and power; the attractiveness of power centers for highly skilled professionals; and the distribution of knowledge-intensive jobs in the hierarchy of the urban system.

Spatiality of Knowledge

Spatiality of Learning Processes

Learning processes and problem solving do not take place in a social, political, or economic vacuum. They are influenced by a multitude of factors whose local interaction results in a spatial context, milieu, or environment. A knowledge environment, as a context of action, consists above all of knowledgeable actors capable of interacting with one another at a location. It is shaped also by infrastructure, resources, organizational structures, and rules, stimuli, informal expectations, and pressures influencing the activities and value systems of people. What counts is the 'local buzz' (Asheim et al., 2007; Bathelt et al., 2004; Storper and Venables, 2004), the intensity and quality of interaction among these elements at a given location and time period. In each domain (e.g., scientific disciplines, production of jewelry, financial services, software development, arts), talented individuals need specific knowledge milieus within which they can work on their creative ideas or follow their professional ambitions. Within each domain, knowledge environments at various locations differ from each other and build a ranking of attractiveness. Outstanding knowledge milieus are able to keep their high reputation only if they can attract the best available agents in their domain from elsewhere. The spatial mobility of scholars, highly educated professionals, creative artists, brilliant students, and distinguished top managers is therefore one of the core research questions in the geography of knowledge.

Knowledge cultures, learning and research conditions, spatial arrangements of infrastructure, and innovative milieus should not be conceived of as a variable determining the activities of people in a location and having predictable effects on all actors involved, but rather as a potential or local offer of opportunities that enable interested, skilled, and motivated people to achieve certain results. Not all individuals will perceive the opportunities or benefit from them. The relation between a knowledge milieu and the actions of individuals is full of serendipities. The discourses and social interactions in a certain location and the effects of a knowledge environment cannot be forecast, but only analyzed after interactions and problem solving have taken place or after certain stages of a research process or a career have ended.

Spatial Differences of Knowledge

New knowledge always starts as local knowledge. It is created in particular places and contexts. Local knowledge becomes regional or global knowledge only if it is understood and accepted elsewhere. The speed at which new knowledge diffuses over space depends on the category of knowledge, the interest that the producer (inventor) has in sharing his or her knowledge, the reputation of the institution within which the new knowledge is produced, the platforms on which the new knowledge is introduced, the codes (e.g., languages and mathematical formulas) and channels by which the information is presented, the prior knowledge that the receivers of the new knowledge need in order to understand the message, the availability of the technology necessary for applying or reproducing certain types of knowledge, and the varying inclination of people to accept a specific kind of knowledge (for details see Meusburger, 2009). Prior knowledge is not something people possess. It is constantly developed in a way similar to the knowledge spiral described by Nonaka and Takeuchi (1995: p. 71). Many spatial

disparities of knowledge evolved from keeping knowledge secret as long as possible. Keeping knowledge secret and exploiting information gaps are well-established ways to be more successful than competitors or gain large profits.

Spatial disparities in knowledge, competence, professional skills, and technological standards have existed since early human history. Key factors in the competitiveness and survival of social systems include skilled, experienced, highly educated, and well-informed elites; informational advantage of merchants and political institutions; inventiveness and creativity; superior technologies giving military supremacy to a territory or state and increasing the efficiency, productivity, and transport capacity of industries; seminal scientific achievements; and a highly skilled workforce. Superior knowledge makes it possible to evaluate situations more realistically than less-informed competitors are able to do, to foresee upcoming problems and unintended consequences of actions earlier than others, to increase productivity, and to replace existing technologies with newer, better performing ones. However, power and knowledge are employed and exercised mainly within organizations and institutions 'through a netlike organization' (Foucault, 1980: p. 98; see also Blackler, 1995; Glückler, 2013); without the help of institutions, organizations, and networks an individual agent is rarely able to transform knowledge into power.

Geographers argue that many spatial disparities of knowledge are not ephemeral, transitional phenomena but remarkably persistent over time. Centers of knowledge (and power) have shifted over the centuries, but spatial differences of knowledge have never disappeared and are a fundamental structural attribute of any society with an elaborate division of labor. Time lags in the circulation and adoption of new knowledge and spatial mobility of highly and lowly educated people have repeatedly created new spatial disparities of knowledge. Traditional neoclassical economists used to assert that codified knowledge is a public and tradable commodity and that new communication devices and transport technologies would eliminate or diminish spatial disparities of knowledge. These misjudgments have resulted from an inadequate concept and definition of knowledge, a neglect of the role of the spatial context, an excessively naive model of communication, and failure to recognize the importance of the time dimension. In many competitive situations, it is not knowledge per se that counts but rather the point in time when new information is acquired or when new knowledge and skills are generated or utilized. The economic value of knowledge and information varies with time. On the stock market, receiving information just minutes or even seconds in advance is enough to net huge profits. In academia, days or weeks can decide which author goes down in history as the one credited with a given scientific discovery.

Methodological Problems of Localizing Knowledge

The first crucial question for empirical work on the geography of knowledge is how intangibles such as knowledge can be localized. The answer depends on the research question, the scale of inquiry, and the category of knowledge investigated. Knowledge is not floating in the air; it is bound to agents or bearers. With human bearers of knowledge, the places of birth, study, work, and residence, as well as the spaces of interaction are usually documented in biographies. This background reveals the environments in which a person was socialized and trained and the places where she or he pursued creative ideas or had professional breakthroughs. It gives insight into a person's networks, career mobility, and other place-specific information. Material bearers of knowledge (e.g., intelligent robots) and immaterial ones (e.g., theoretical concepts, scientific methods, patents, ideologies) can be related to the places where they were invented and later diffused. It is possible to localize the places where exceptional research results have been achieved, new philosophical schools founded, new scientific questions asked, new scientific instruments used for the first time, or seminal scientific discussions conducted (Meusburger and Schuch, 2012).

The second question is: which categories of knowledge are or are not suitable for localization in space and on various scales of investigation. Unlike the meaning of the terms *educational attainment* and *educational achievement*, whose spatial variations can be studied with relatively well-defined indicators for small spatial entities, the meaning of the word *knowledge* is broad and elusive. It is quite clear that not all categories of knowledge can be localized. Scholars of social sciences and humanities focusing on the actions (behavior) of individual human beings regard knowledge primarily as capability or capacity for social action (Stehr, 2001). From that point of view, knowledge is rooted in individual persons and becomes manifest in their cognitive and intellectual capacities, professional competence and achievements, and performance in problem solving and decision making. At this micro level of analysis, it is possible to study the spatial dimension of many categories of knowledge. But at higher levels of spatial aggregation, knowledge can be grasped only with the help of indicators representing *input* in knowledge generation (e.g., research resources) and *outcomes* of learning, research (e.g., scientific publications, patents, high-tech products), problem solving and experience. The challenge is to find indicators representing such outcomes of learning and achievements. Such indicators may include professional achievements and skills, scientific awards, research results, inventions, patents, high-tech products, scientific reputation of a university or research institute, new organizational structures, specific institutional rules, cultural value systems, and other outcomes of learning and not learning.

Categorizations and Definitions of Knowledge

Knowledge can be theorized, categorized, defined, and operationalized in many different ways, depending on the discipline, research questions, theoretical approach, and scale of inquiry. Philosophical questions such as 'what is (objective) knowledge,' 'what is truth,' 'what is reality,' and 'what are the differences between opinion, belief, and knowledge' have been discussed in thousands of publications (for details see Abel, 2008). The discussion here is confined to some of those categories and definitions that are important for the explanation of spatial variations of knowledge.

Knowledge Is Not Equivalent to Information

As a large part of knowledge can be codified and because codified knowledge can be transformed into information, some authors, especially mainstream neoclassical economists, have argued for decades that knowledge and information are more or less the same (for an overview and critique of this approach see Ancori et al., 2000: p. 256; Cowan et al., 2000: p. 221; Meusburger, 2009, 2013: pp. 16–17). Granted from the viewpoint of the person who generates new knowledge or tries to communicate his or her knowledge to other people the distinction between information and knowledge becomes blurred. From the standpoint of a message's recipient, however, the difference between knowledge and information is rather distinct.

The fact that codified knowledge is made public, shared with others, and made available for free does not mean that it is understood, accepted, or used by all who have access to it. Ready access to information is only one side of the coin. The other side is about the cognitive abilities, interests, training, educational achievement, and prejudice of recipients and the milieu they are embedded in.

As soon as the communication process between a generator and receiver of new knowledge, the spatial dissemination of information, and spatial disparities of knowledge become topics of research, a distinction between knowledge and information becomes indispensable. The discrepancy between knowledge and information is also manifest in the contemporary simultaneity of the superabundance of information and scarcity of knowledge.

Categorizations of Knowledge

According to Abel (2008), it is important to distinguish first between a narrow and a broad notion of knowledge. The narrow concept of knowledge "refers to knowledge obtained by a methodically well-regulated procedure bound to justification, truth, and verification. It is essential with such knowledge that one be able to talk about it and that it be communicable, transferable, intersubjectively verifiable, and interchangeable" (p. 12). Scientific knowledge falls into this category. "The broad notion of knowing and knowledge refers to the ability to adequately grasp what something is about (e.g., what a sentence or a picture is about) on the one hand and the domain of human capacities, competencies, skills, practices, and proficiencies on the other" (p. 12).

A second categorization (Abel, 2008: p. 13) distinguishes among the following four types:

1. *Everyday knowledge* (e.g., knowing where the bus station is, knowing the price of a good, knowing the results of yesterday's soccer matches).
2. *Theoretical knowledge* (e.g., knowing that the sum of the angles in a triangle equals 180°).
3. *Action knowledge* (e.g., knowing how to drive a car, play violin, or perform a heart surgery).
4. *Moral or orientational knowledge* (e.g., knowing what is right and wrong). This kind of knowledge is generated and cultivated by religions, ideologies, and collective memories. Its truth claims are not open to scientific reasoning (for details see Abel, 2008: p. 12). For pragmatic reasons, categories 1, 2, and 3 are sometimes subsumed in the category called 'factual knowledge.' The categories factual knowledge and orientational knowledge are frequently used when the mutual relations between power and knowledge are an issue.

A very popular distinction is between explicit (codified) and implicit (tacit) knowledge. Tacit knowledge means those aspects of knowing that are implicit in situations of perceiving, speaking, thinking, and acting but not made explicit, i.e., not disclosed on the surface (for details see Collins, 2001). Some authors use the concept of tacit knowledge as a kind of umbrella term for nonverbal and nonpropositional knowledge. Unfortunately, "the terminology and meaning of 'tacitness' in the economics literature [have] drifted far from its original epistemological and psychological moorings [and have] become unproductively amorphous" (Cowan et al., 2000: p. 213).

Verbal knowledge (knowledge that can be articulated by using linguistic expressions) can be distinguished from nonverbal knowledge (e.g., visual, pictorial, or musical knowledge). One can also make a distinction between propositional knowledge (knowledge that can be articulated in a linguistic proposition such as 'Picasso was a painter' or 'Berlin is the capital of Germany') and nonpropositional knowledge that cannot really be grasped by words (such as knowing how to understand a bodily movement but not being able to put it into words). "Forms of nonpropositional knowledge also become manifest in a person's practical skills, pictures, shapes, sounds, gestures, or mental images" (Abel, 2008: p. 25). Very important is the differentiation between knowledge relating to matters of fact and knowledge based on skills and abilities.

When spatial dissemination of knowledge becomes an issue, researchers also need a vertical differentiation of codified knowledge (Meusburger, 2013). After first considering the time and costs of acquiring the prior knowledge necessary to understand particular information, one should distinguish between lower and higher levels of codified knowledge. The gap between knowledge and information grows with each level. The lowest category, everyday knowledge, is easily understood by almost everyone. Whoever has access to this kind of knowledge is able to use it, provided he or she understands the language. This type of knowledge corresponds more or less to the traditional neoclassical economist's standard view that codified knowledge is equivalent to information, that codified knowledge is a public, tradable, and spatially mobile commodity, and that spatial disparities of knowledge are only short lived.

All other, higher levels of codified knowledge require the information recipient to possess a certain body of prior knowledge (sometimes years of study, professional experience, and research) that will enable this person to understand the available information, evaluate its relevance and usefulness for achieving goals, and integrate it into his or her own knowledge base. Having access to higher levels of codified knowledge is not equivalent to gaining knowledge.

Only everyday knowledge, which is the least important type for economic competitiveness and innovation, is ubiquitously available (at least theoretically) if it is assumed that all places have the necessary technical equipment and that all people

have an equivalent inclination to receive the message. Workplaces of actors commanding levels of codified knowledge higher than everyday knowledge tend to be spatially concentrated. It goes without saying that other categories of knowledge not discussed in this article are also frequently used in geography of knowledge.

The Process of Communicating Different Categories of Knowledge from A to B

The spatial diffusion of new knowledge and the migration of scholars and highly skilled professionals are key issues in geography of knowledge. However, the spatial dissemination of knowledge is much more complex and more place-dependent than the simple sender–receiver model of information theory suggests.

The most efficient or reliable way to transfer specialized high-level knowledge or competencies from place A to place B is through migration of those people who generated or currently command that knowledge. No other attempts at transferring high-level knowledge (in the form of texts, figures, or instruments) guarantee that the knowledge will be fully understood, accepted, or applied by the receivers as the sender intended. When studying the spatial circulation of different categories of knowledge, one has to take into account the cognitive processes of sender and receiver and the intervening variables and learning loops within the communication process. Many authors overemphasize the knowledge producer's ability to codify knowledge into information. They also underestimate or neglect the cognitive processes of the receiver and overlook the role of prior knowledge, a receiver's ability to read, understand, and accept certain kinds of information and to integrate it into his or her knowledge base. The producers and transmitters of knowledge have limited influence on the extent to which their message is understood, accepted, and processed or the way it is translated and interpreted elsewhere (Meusburger, 2009, 2013).

Higher levels of knowledge can be transferred from one person to another only if the receiver's prior knowledge, training, skills, experience, or capabilities are broadly equivalent to those of the sender. A certain type of knowledge may be perfectly codified in text and equations, published in international journals, and offered for free on the Internet, but it may be understood by only 100–200 theoretical physicists worldwide. The rest of the world's population may not have acquired the prior knowledge and experience necessary to read and comprehend the text, formulas, and equations and to integrate this new information into their own knowledge base.

Persons who have not completed years of study and research in molecular biology or theoretical physics have little or no use for the available scientific publications in these fields. Some types of scientific knowledge cannot be simply transferred from A to B. To be assimilated, they must be replicated in B with expensive experiments in sophisticated laboratories. Owing to these preconditions, higher levels of knowledge often circulate only among the upper echelons of the urban hierarchy, among university towns, or among centers of an economic domain, and they leapfrog large areas. The generation of new high-level knowledge generally precipitates new knowledge divides (e.g., digital divide).

Each step of the communication process illustrated in Meusburger (2009: p. 35) has a high degree of actor-, community-, and place-dependent contingency letting some information pass in its original meaning but withholding, transforming, or distorting other information. In this context, the term *filter* is used as a metaphor for various cognitive processes, mind-sets, and power relations that influence the selectivity of perception; the processing, evaluation, and interpretation of incoming information; and the enactment of knowledge as practice.

Relations between Knowledge and Power

In the study of the persistence and continuous reproduction of knowledge disparities between centers and peripheries, two issues have become crucial: coalitions between knowledge and power, and the attractiveness of power centers for highly skilled professionals. Why do agents representing the highest echelons of power and knowledge so frequently build coalitions? And why do they seek (relational) proximity to each other? On the one hand, it is in the interest of power to bind knowledge into networks of assent. Both instrumental power (i.e., the capacity to make others do one's will and the ability to control, regulate, and dominate) and associational power (i.e., the power to do things by acting in concert or by using institutional mediation) depend on experts, scientists, research institutions, intelligence services, and think tanks to achieve and maintain influential status for themselves. Their knowledge, expertise, and competence are needed in order to examine a situation as precisely as possible, to perceive upcoming problems earlier than others, to offer solutions to problems that have to be solved and to retain political, economic, military, or cultural supremacy.

Upper echelons of political power also need the support of the representatives of orientational knowledge (priests, intellectuals, artists, propagandists, and ideologues) who share their norms, values, and collective memories; vindicate their actions; and legitimate their power. Experts of orientational knowledge are trained and experienced in the art of influencing, convincing, and manipulating people. Their task is not to analyze a 'real situation' or to seek 'objective facts' but rather to make moral judgments; interpret events; create beliefs, mind-sets, and collective memories; sustain the internal cohesion and motivation of their social system; mobilize loyalty; and excuse actions criticized by opponents.

Advisers, counselors, scientists, experts, priests, and intellectuals derive part of their power from their claim to know better than the majority of people (Stehr, 1992), to foresee the consequences of an intended action, and to be able to forecast future events and developments. When having to make important decisions, those in power in earlier centuries engaged dream interpreters, oracles, and priests, all of whom claimed they could communicate with the gods (or ancestors) or represent God's will on Earth. Today, power brokers and decision makers rely on think tanks, scientists, and other experts.

Just as power depends on knowledge, many bearers of factual knowledge and orientational knowledge need the

support of power and seek relational proximity to power. Scientists and experts in various domains often have to prove their usefulness to power in order to secure the privileges and resources they need for their work. Often the state, a political party, center of economic power, or prestigious journal lends authority to experts. Power often ignores those experts who do not serve its interests. Therefore, experts authorized by state agencies may be opposed by experts supporting the interests of local people. Most experts and intellectuals want to be concerned with the essential in existence; they strive for influence, recognition, and reputation. Network centrality, or relational proximity to power, demonstrates authority, importance, and prestige. It determines access to valued resources and information unavailable to people at the margin. Proximity to and trustful relations with those in power increase the opportunities of experts and scholars to influence important political decisions. Even dissidents and other intellectuals opposing those in power have to operate in centers of power in order to elicit response in the media and create a viable alternative.

Those in power have always attempted to exercise control over institutions of knowledge generation and information distribution, fortunately with limited success. Over the centuries, knowledge was frequently used as a means to discipline, dominate, exclude, and manipulate people (Foucault, 1969). Production and distribution of knowledge and information are often associated with censorship, forgery, secrecy, and propaganda. Many national memories and histories are socially constructed according to the interests of dominant groups and contested by other groups. The representations of the culture, history, and geography of colonies by imperial powers (Said, 1978; Gregory, 1998) are perfect examples of how culture (knowledge) underwrites power and how power elaborates culture (knowledge).

Distribution of Knowledge-Intensive Jobs in the Hierarchy of the Urban System

The vertical division of labor (bifurcation of skills), the differentiation of complex large systems, the professionalization and bureaucratization of labor relations, and various other trends induced a spatial concentration of power and knowledge in upper levels of the urban hierarchy. This uneven distribution of jobs by skill level also influences the selective migration of employees with higher or lower levels of educational achievement, as well as the uneven distribution of creative milieus. Important decision makers, scientists, experts, artists, and intellectuals striving for prestige, influence, and success sooner or later find their place of work in one of the centers of their domain or will create a new center in their chosen field. Knowledge centers attract a large share of their highly educated workforce from other regions, and rural or peripheral areas lose most of their talented youths to the centers. Both processes significantly transform the places concerned and contribute to the long persistence of center-periphery disparities.

Why do workplaces of highly educated employees in various economic sectors tend toward such high spatial concentration in a few centers, whereas low-skilled routine activities in production and administration tend toward decentralization as shown in Tables 1 and 2?

Table 1 Spatial concentrations of jobs for Hungarian university graduates in Budapest and the capitals of the 19 Hungarian counties (1980)

Economic sectors or professions offering jobs for university graduates	*Percentage of jobs located in Budapest and the capitals of the 19 Hungarian counties*[a]
Foreign trade	98.7
Personal and business services	96.1
Organization and development of housing	96.0
Science and research	88.7
Cultural services	88.6
Banking and finance	87.8
Transport and traffic	86.6
Telecommunications	86.5
Public administration	80.3
Machinery	79.9
Light industry	74.9

Note: Special analysis of the 1980 Hungarian census.
[a]Budapest and the capitals of the 19 counties accounted for 35.4% of the national population in 1980.
Source: Meusburger, P., 1997. Spatial and social inequality in communist countries and in the first period of the transformation process to a market economy: the example of Hungary. Geographical Review of Japan 70 (Ser. B), 132. Reprint with permission.

Table 2 The concentration of high- and low-qualified jobs in Hungarian towns and villages up to 5000 inhabitants (1980)

Professions and occupations of persons with a university degree	*Percentage of all Hungarian jobs located in cities and villages smaller than 5000 inhabitants*[a]
Journalists	0.3
Leading organs of state power and central administration	0.4
Foreign trade	0.4
Community services	0.6
Banking and finance	0.7
Telecommunications	0.8
Leaders of party and mass organizations	1.0
Law and public order	1.4
Jobs for university graduates in total	7.4

Note: Special analysis of the 1980 Hungarian census.
[a]Villages up to 5000 inhabitants accounted for 32% of the national population in 1980.
Source: Meusburger, P., 1997. Spatial and social inequality in communist countries and in the first period of the transformation process to a market economy: the example of Hungary. Geographical Review of Japan 70 (Ser. B), 133. Reprint with permission.

The first explanation draws on organization theory and the meritocratic principle. No social system can completely avoid faulty analyses, misjudgments, or wrong decisions. But the fewer mistakes it makes in important situations of problem solving and the more adaptive it is, the more likely its long-term competitive survivability will be. In social systems, knowledge, skills, and experience have a function similar to that of redundancy in technical systems. They reduce uncertainty and enhance the chances of survival. A social

system, institution, or state acts in its own interest when it fills its vital positions of information processing, decision making, planning, coordination, and control with well-educated, experienced persons. In particular, people and subsystems whose positions are fraught with uncertainty and whose decisions have abiding consequences for the entire system require advanced skills, superior knowledge, and vast experience.

In a dynamic and competitive society, the skills and knowledge needed for the chief functions of a social system striving for success will always be scarce. Therefore, the first crucial question is where to locate scarce knowledge and skills within a social system's structure of decision making. In bureaucratic systems and in systems confronted with low degrees of uncertainty, decision making, problem solving, research, and planning shift to the upper levels of the system's hierarchy, so the lower levels are characterized by routine activities and workplaces for people with lower skills. In systems dealing with a dynamic, complex environment and dealing with constantly changing, unpredictable, one-time transactions, lower ranking and decentralized positions, too, need people having a high level of competence and skills (for details see Mintzberg, 1979).

The second question is where to locate valuable and scarce knowledge and skills spatially. Routine activities for which there are plans, rules, and common procedures call for fewer learning processes, lower skills, and less face-to-face contacts with other organizations than nonroutine activities do. Routine activities are coordinated predominantly by indirect contacts (phone, emails, and letters). Therefore, they can be situated in a great number of relatively smaller cities, even rural areas, as long as the traditional location factors such as transportation and wage costs make it feasible. The harder it is for decisions to be governed by guidelines, plans, regulations, or so-called hard information, the more dynamic and competitive the social environment, and the greater the uncertainty about the future, the more necessary it is to have frequent and spontaneous face-to-face contacts with knowledgeable and well-informed decision makers and highly skilled specialists of other organizations and domains. At the highest levels of management the need for spontaneous face-to-face contacts and valuable networks can be met only in a few large centers offering a large diversity and density of knowledgeable specialists and prominent decision makers of various domains. Trustful relations and frequent face-to-face contacts with leading decision makers in government, research, finance, consulting, media, and big business offer access to early knowledge about important tendencies and innovations and provides crucial information to an organization's ranking members facilitating their quick response and adaptation to new situations and developments.

The second explanation for the spatial concentration of jobs for important decision makers and highly specialized professionals is that places and milieus, especially centers and peripheries, have symbolic meanings. The center of a social system or of a domain is defined as the place where its highest authority is located. By definition, the place of maximum access to the system as a whole, a center is associated with power, authority, dominance, prestige, control, and influence. It claims preeminence with regard to expertise and competence. The term periphery symbolizes dependency, marginality, backwardness, and lower prestige.

Further Perspectives

As knowledge concerns a broad range of activities, a large number of emerging fields and further perspectives could be mentioned. From a geographer's point of view, at least three issues should receive more attention. The first comprises the questions: What is the essence of knowledge milieus and creative environments? How can the impact of environments on knowledge generation be theoretically conceptualized and empirically scrutinized? What kinds of processes and policies help to create knowledge-based regional advantages? Are creative milieus bound to certain types of cities? If they emerge in rural peripheries, too, then under which external conditions and in which domains will they form?

Most of the research in the geography of knowledge has dealt with positively valued aspects of knowledge generation, such as research, innovations, or cultural and intellectual exchange. But social relations are not based only on selfless knowledge sharing, transparency, and open innovation. A second group of new perspectives could highlight the darker sides of knowledge generation, e.g., manipulation and censorship of information, cultural imperialism, spying, using asymmetries of knowledge for fraud, and so on. Many hegemonic practices are based on asymmetries of knowledge and the exploitation of knowledge gaps.

A third group of future perspectives can be seen in the field of spatial disparities of orientational knowledge, e.g., in spatial disparities of the predominance of ideologies, mentalities, collective memories, and prejudice, and in clashes of different bodies of orientational knowledge in ethnically diverse areas.

See also: Authority, Social Theories of; Center–Periphery Relationships; Education, Geography of; Elites, Anthropology of; Epistemic Communities and Collaborative Research; Expertise, Anthropological Approaches To; Hegemony: Anthropological Aspects; Inequality: Comparative Aspects; Information and Knowledge, Organizational; Intelligence, Prior Knowledge, and Learning; Knowledge (Explicit, Implicit and Tacit): Philosophical Aspects; Knowledge Society, History of; Knowledge, Anthropology of; Knowledge, Sociology of; Learning: Organizational; Organizations: Authority and Power; Power in Society; Power; Research and Development in Organizations; Science, Geography of; Social Scientists as Experts and Public Intellectuals; Space and Social Theory in Geography; Tacit Knowledge, Psychology of.

Bibliography

Abel, G., 2008. Forms of knowledge: problems, projects, perspectives. In: Meusburger, P., Welker, M., Wunder, E. (Eds.), Clashes of Knowledge: Orthodoxies and Heterodoxies in Science and Religion, Knowledge and Space, vol. 1. Springer, Dordrecht, The Netherlands, pp. 11–33.

Ancori, B., Bureth, A., Cohendet, P., 2000. The economics of knowledge: the debate about codification and tacit knowledge. Industrial and Corporate Change 9, 255–287.

Asheim, B., Coenen, L., Vang, J., 2007. Face-to-face, buzz, and knowledge bases: sociospatial implications for learning, innovation, and innovation policy. Environment and Planning C: Government and Policy 25, 655–670.

Bathelt, H., Malmberg, A., Maskell, P., 2004. Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. Progress in Human Geography 28, 31–56. http://dx.doi.org/10.1191/0309132504ph469oa.

Blackler, F., 1995. Knowledge, knowledge work and organizations: an overview and interpretation. Organization Studies 16, 1021–1046.

Collins, H.M., 2001. What is tacit knowledge? In: Schatzi, T., Knorr Cetina, K., von Savigny, E. (Eds.), The Practice Turn in Contemporary Theory. Routledge, London, pp. 107–119.

Cowan, R., David, P.A., Foray, D., 2000. The explicit economics of knowledge: codification and tacitness. Industrial and Corporate Change 9, 211–253.

Foucault, M., 1969. The Archaeology of Knowledge. Pantheon Books, London.

Foucault, M., 1980. In: Gordon., C. (Ed.), Power/Knowledge. Selected Interviews & Other Writings 1972–1977. Pantheon, New York.

Glückler, J., 2013. The problem of mobilizing expertise at a distance. In: Meusburger, P., Glückler, J., El Meskioui, M. (Eds.), Knowledge and the Economy, Knowledge and Space, vol. 5. Springer, Dordrecht, pp. 95–112.

Gregory, D., 1998. Power, Knowledge and Geography: An Introduction to Geographic Thought and Practice. Blackwell, Oxford.

Maskell, P., Malmberg, A., 1999. Localised learning and industrial competitiveness. Cambridge Journal of Economics 23, 167–185.

Meusburger, P., 2009. Spatial mobility of knowledge: a proposal for a more realistic communication model. disP – The Planning Review 177 (2), 29–39.

Meusburger, P., 2013. Relations between knowledge and economic development: some methodological considerations. In: Meusburger, P., Glückler, J., El Meskioui, M. (Eds.), Knowledge and the Economy, Knowledge and Space, vol. 5. Springer, Dordrecht, pp. 15–42.

Meusburger, P., Schuch, T. (Eds.), 2012. Wissenschaftsatlas of Heidelberg University. Spatio-temporal Relations of Academic Knowledge Production. Bibliotheca Palatina, Knittlingen.

Mintzberg, H., 1979. The Structuring of Organizations. A Synthesis of the Research. Prentice Hall, Englewood Cliffs, NJ.

Nonaka, I., Takeuchi, H., 1995. The Knowledge-Creating Company. How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, New York, Oxford.

Said, E.W., 1978. Orientalism. Routledge and Kegan Paul, London.

Stehr, N., 1992. Experts, counsellors and advisors. In: Stehr, N., Ericson, R.V. (Eds.), The Culture and Power of Knowledge. Inquiries into Contemporary Societies. Walter de Gruyter, Berlin, New York, pp. 107–155.

Stehr, N., 2001. Wissen und Wirtschaften. Die gesellschaftlichen Grundlagen moderner Ökonomie (Knowledge and Economic Activities: The Societal Foundations of Modern Economies). Suhrkamp, Frankfurt am Main.

Storper, M., Venables, A.J., 2004. Buzz: face-to-face contact and the urban economy. Journal of Economic Geography 4, 351–370.

Relevant Website

http://link.springer.com/search?facet-series=%227568%22&facet-content-type=%22Book%22 – For more titles published in the series "Knowledge and Space".

Knowledge Representation

Ernest Davis, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In artificial intelligence, knowledge representation is the study of how the beliefs, intentions, and value judgments of an intelligent agent can be expressed in a transparent, symbolic notation suitable for automated reasoning. From a purely computational point of view, the major objectives to be achieved are breadth of scope, expressivity, precision, support of efficient inference, learnability, robustness, and ease of construction. Knowledge-based techniques have been applied successfully for many computational tasks including text interpretation and cognitive robotics. Many different general architectures have been used for knowledge representation, including first-order logic, other formal logics, semantic networks, and frame-based systems. The representation of temporal knowledge is both a problem of central importance in knowledge representation and an archetype of the kinds of issues that arise in developing representations for various domains. The use of machine learning techniques for the automatic construction of knowledge bases and knowledge representations is difficult, but has achieved some degree of success.

Knowledge Representation: Ideal and Reality

The knowledge-based approach to artificial intelligence (AI) conjectures that a large component of any AI system with broad intelligence will be a knowledge base: a transparent, symbolic representation of the agent's beliefs, intentions, and value judgments, together with reasoning processes that modify the knowledge base in rational ways. Perceptual and motor processes connect the knowledge base to the outside world. The field of knowledge representation (KR) is the study of how such a knowledge base can be constructed (van Harmelen et al., 2008).

The representation of knowledge in a knowledge base should be symbolic, transparent, and modular. That is, the atomic elements of a knowledge base are symbols, each of which has an associated meaning. It is possible to identify small substructures of the knowledge base that correspond to specific beliefs, judgments, and intentions. For example, the belief that Holbein painted a portrait of Thomas More in oil might be represented in a predicate calculus formula like

$$\exists_x \text{ Oilpainting}(x) \ \wedge \ \text{Painter}(\text{Holbein}, x) \wedge \text{Depicts}(x, \text{ ThomasMore})$$

or in a fragment of a semantic net (Figure 1).

Here the symbols are '∃' (meaning 'there exists'), 'x' (a variable), '∧' (meaning 'and'), 'OilPainting', 'Painter', 'Depicts', 'Holbein', and 'ThomasMore'. The parentheses and the comma indicate the syntactic structure.

Not every belief of the agent can be explicitly expressed in the knowledge base. For example, someone who knows that Holbein painted a portrait of Thomas More will also know that Holbein knew how to paint, that Thomas More was painted, that Thomas More was born before Holbein's death, and so on. These ancillary beliefs are not each separately represented in the knowledge base. Rather, they are derived as needed by combining the particular fact with general knowledge about painting.

The construction of a broadly intelligent artificial agent is the ultimate objective for AI, but realistic short-term projects focus on specific, narrowly defined tasks. For such a limited project, the problem of KR is one of finding data structures that express the knowledge needed for the particular application.

Thus broadly construed, the problem of KR would appear to be equivalent to computer programming generally. Any program, after all, uses some data structures and incorporates some kind of knowledge. The field of KR is distinguished in its focus on a broad range of fundamental knowledge, particularly commonsensical world knowledge; on representations that are highly expressive, particularly of various natural forms of partial knowledge; and on forms of reasoning that arise in intelligent tasks. As AI is the study of tasks that are easy for people to carry out but difficult to automate, so is KR the design of data structures that express knowledge that people deal with easily, but are hard to encode in a data structure. The design of a KR can be very roughly divided into five parts.

1. Architecture: The design of general-purpose schemas for KR and procedures for reasoning with these schemas.
2. Content: The characterization of the knowledge that is to be represented, particularly for fundamental categories such as time, space, and communication.

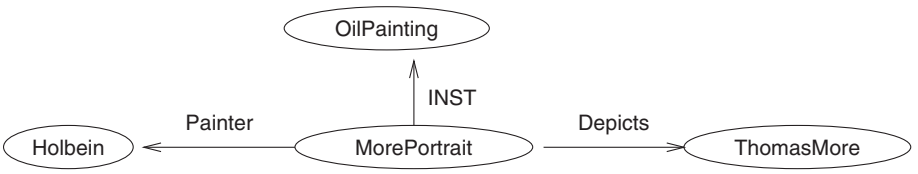

Figure 1 Fragment of a semantic net.

3. Implementation: The specifics of instantiating the representation as a data structure and the reasoning process as a procedure.
4. Interface: The connection of the knowledge base to an external application.
5. Learning: Constructing the knowledge base, or parts of it, automatically by the analysis of a data corpus.

Desiderata

The design of a KR will generally aim toward the following objectives:

Scope: A KR for a limited application should cover all relevant aspects of the relevant domains. For a broadly intelligent aspect, an adequate KR theory must, first, describe the specifics of the representation of fundamental and universal human thought, such as time, space, simple physical categories, the human mind, and human relations (Davis, 1990; McCarthy, 1968), and, second, describe a general representational architecture that is applicable to any comprehensible object of thought.

Expressivity: The KR should be able to express the forms of partial and incomplete knowledge that naturally arise for the desired application.

For example, a standard form of representation is a collection of tables of values, known in computer science as a 'relational database'. Table 1 displays a table of paintings and a table of artists.

However, even within its narrow topical scope, such a representation may not achieve reasonable expressivity, because there is no way to express plausible partial states of knowledge such as

Delacroix was a generation older than Monet.
Rembrandt painted many oil paintings.
The Sistine Chapel and the sculpture of David were created by the same artist.

If such partial states of knowledge are important in the application, then a more expressive language must be found.

Here is another example: it is tempting to represent spatial information using a representation that corresponds to a single picture of the scene, sometimes called a diagrammatic representation (Glasgow et al., 1995). Such representations are easily understood and support efficient inference. However, they cannot be used to represent partial information. For instance, an agent might know that A is a mile from B and that A is a mile from C without knowing the distance from A to C. However, it is not possible for a single picture to display the distances from A to B and from A to C without committing itself to some particular distance between B and C. Therefore, this and other similar states of knowledge cannot be represented in a diagrammatic representation.

Table 1 Relational database of paintings and artists

Work	*Medium*	*Genre*	*Artist*
Last Supper	Fresco	Religious	da Vinci
Melancholia	Drawing	Allegory	Durer
Thomas More	Oil	Portrait	Holbein
Mona Lisa	Oil	Portrait	da Vinci

Artist	*Country*	*Birth*	*Death*
Durer	Germany	1471	1528
Holbein	Germany	1497	1543
da Vinci	Italy	1452	1519
O'Keefe	USA	1887	1986

Precision: The meaning of a KR should be unambiguous, and, as far as the subject matter allows, well defined.

Support for reasoning: It must be possible to model the processes of reasoning involved in cognition in terms of well-defined operations over the KR. For example, as mentioned above, an agent who knows that Holbein painted a portrait of More should be able to conclude that Holbein knew how to paint, that More was born before Holbein died, and so on.

Efficiency of reasoning: The implementation of reasoning operations over the KR must also be efficient. Depending on the application, it may be necessary to draw conclusions from a large knowledge base in a fraction of a second.

Learnability: The KR should support modules that can learn from experience and can generalize individual cases to general rules.

Support for uncertainty: To be useful, a knowledge base cannot be limited to information known with certainty. Rather, the KR must be able to express relative degrees of certainty, and reasoning modules must use these (Pearl, 1988; Halpern, 2003).

Robustness: The information in a knowledge base will sometimes be wrong or corrupted. The effects of an error in the knowledge base should be limited in scope.

Ease of construction: Ideally, the information in a large knowledge base can be collected automatically or semi-automatically from existing data sets. If the knowledge base must be constructed manually, then a notation that can be used by users with little technical expertize is obviously preferable to one that requires a KR expert. The less the human labor, especially expert labor, required to build a knowledge base, the better.

Interface: A KR should support a suitable interface to the user or the application.

Expressivity and precision are features both of the architecture and of the content of a knowledge base. Scope is chiefly a feature of content, although expressivity and therefore architecture do bear on scope; some domains inherently require a very expressive language. Robustness, ease of construction, and support for reasoning, learning, and uncertainty are primarily features of the architecture. Efficiency of reasoning depends on the architecture, procedures, and class of problems being solved.

Applications

KR and automated reasoning have been applied, with greater and lesser degrees of success, to a very wide range of intelligent tasks. Here we will discuss briefly the representational issues involved in two particular important categories of applications: the interpretation of natural language text and action planning for robotic systems.

Text Interpretation and Information Retrieval

Many useful computational tasks involving text, such as document retrieval, can be effectively carried out based purely on superficial features of the text, such as word counts, with no representation of the meaning of the text or knowledge of the domain it refers to. However, tasks such as answering questions necessarily engage with the meaning of the text; therefore, performing these well requires a representation of the text and of the underlying domain. Other tasks such as machine translation, summarization, and information extraction can be carried out without using KR to an accuracy that may be useful but cannot be carried out with very high quality. Basic components of text interpretation, especially disambiguation, often require the application of world knowledge. For example, in interpreting the sentence "Jane knocked on Susan's door, but she did not answer", disambiguating 'she' to mean 'Susan' requires understanding the conventions of answering knocks. (By contrast, if the sentence ended, 'but she did not get an answer', then 'she' would refer to Jane.)

Attempts to use knowledge-based techniques in text understanding vary widely in terms of the depth of the knowledge represented, the breadth of the subject matter, and the degree to which the knowledge base is created manually as opposed to assembled automatically. At one end of the spectrum are systems that form a complete logical representation of the explicit and implicit content of the text, and that draw on a detailed knowledge base in carrying out inferences from the text (Balduccini et al., 2008); these necessarily are highly limited in scope, and require massive human labor. At the other end are systems whose knowledge base is derived purely from statistical analysis of a corpus of texts; the knowledge base is essentially a table of correlations between words (Dumais, 2005).

As of the time of writing, perhaps the greatest successes have been achieved by information extraction systems that lie somewhere between these two extremes. Information extraction systems are designed to extract from text information of a restricted, often predetermined form, such as the employee, company, and position in a business page story about hiring a new executive. The systems use a combination of general linguistic analysis, specific linguistic patterns, pre-existing online language tools such as WordNet, and basic background knowledge of the domain such as a hierarchy of semantic categories and semantic constraints (e.g., an employee must be a person). Some parts of these are constructed manually for the particular task; some are drawn from existing manually created tools; some are developed automatically, using machine learning techniques, both supervised and unsupervised.

Robotics and Planning

The application of knowledge-based techniques to robot control, sometimes called 'cognitive robotics,' has focused primarily on high-level planning, although researchers have also looked at integrating automated reasoning with lower level manipulation (Reiter, 2001; Ghallab et al., 2004; LaValle, 2006). Central issues in cognitive robotics include the following:

- *Causal modeling*: Developing a theory that characterizes how the world changes, both in response to the robot's actions, and as a result of processes independent of the robot.
- *Modeling perception*: A high-level model of what the robot can expect to perceive while executing a plan, and how it can use the perceptions to guide its actions.
- *Developing a high-level language of plans*: In the simplest case, extensively studied, a plan is simply a sequence of actions to be executed, but in richer domains, a plan may involve conditionals, loops, interruptions, concurrency, and so on.
- *Projection*: Predicting what will happen if the robot attempts to execute a specified plan.
- *Planning*: Constructing a plan that achieves a specified goal.
- *Multiagent planning*: Constructing a plan for an environment with multiple agents. Agents may be cooperative, independent, or antagonistic; if cooperative, there may or may not exist a centralized control (the boss).

A number of autonomous robotic systems, including the Mars rover, have made some use of knowledge-based reasoning.

Architectures

A number of KR architectures are briefly surveyed below. The particular architectures included here have been included, in part because of their importance in KR research and applications, and in part to illustrate the range of possible approaches to KR.

Logic-Based Systems

A logic-based system represents knowledge as a set of sentences in a logical language. Reasoning is implemented as inference procedures that carry out deductive proofs authorized within the logical theory (Geneserth and Nilsson, 1987).

A logical theory characteristically consists of three parts:

- A syntax, which specifies what kinds of symbols are used, and how they are combined to form sentences.
- A semantics, which specifies how the meaning of a sentence relates to the meanings of the symbols it contains.
- A proof theory, which specifies what combinations of sentences constitute deductive proofs.

For example, first-order logic (also known as the predicate calculus) is the most widely used logic in KR. In first-order logic, there are seven kinds of symbols:

- Constant symbols represent individual entities. For example, the symbol Ad1517 may represent the year 1517.
- Predicate symbols represent properties or relations. For example, Painter may represent the two-place relation of a person painting a painting. Precedes may represent the relation of one event preceding another.
- Function symbols represent mappings. For example, BirthOf may represent the one-place mapping of a person to the event of his/her birth. CreationOf may represent the mapping of an artifact to the event of its creation.
- The Boolean operators are '$\wedge$' meaning 'and'; '$\vee$' meaning 'or'; '$\sim$' meaning 'not'; and '$\Rightarrow$' meaning 'implies'.
- Variables such as x, y, z represent indeterminate entities.
- The quantifiers are '$\forall$' meaning 'for all' and '$\exists$' meaning 'there exists'.

- Finally, the comma and the open and close parentheses indicate the structure of the sentence.

Constant, predicate, and function symbols are known as nonlogical symbols; they may be chosen, and their meanings assigned, by the designer of the knowledge base. Boolean operators, quantifiers, variables, and grouping symbols are known as logical symbols; their meanings are fixed by the logic. The distinction is roughly analogous to the distinction between open and closed categories in linguistics, or to the distinction between identifiers and reserved words in programming languages.

The syntactic rules for first-order logic establish that these symbols can be put together to form sentences such as the following:

$$\forall x,y\ \text{Painter}(x,y) \Rightarrow \text{Precedes}(\text{BirthOf}(x),\text{CreationOf}(y)) \quad (1)$$

The semantic rules establish that, given the above assignments of meanings to the symbols Painter, Precedes, BirthOf, and CreationOf, the sentence is true just if the birth of every painter precedes the creation of all his/her works. The proof theory establishes that, given sentence (1) above and the sentence Painter(Holbein,MorePortrait), it is possible to infer the following sentence:

Precedes(BirthOf(Holbein),CreationOf(MorePortrait))

This inference is valid regardless of the meaning of the nonlogical symbols (Mates, 1972).

Many other logics have also been used for KR. Some of these extend first-order logic to achieve greater expressivity. Modal logics define a variety of operators on sentences. For instance, given a base sentence, such as 'Bill Clinton is President,' sentences such as 'Bill Clinton was President,' 'Al Gore knows that Bill Clinton is President,' or 'Bill Clinton ought to be President' can be generated by applying temporal, epistemic, or deontic operators (Turner, 1984). Fuzzy logic allows the use of vague terms that hold to greater and lesser degrees; for example, 'George Washington was tall' is reasonably true, while 'Wilt Chamberlain was a tall person' is absolutely true (Zadeh, 1987). Probabilistic logics allow sentences to be tagged with a measure of certainty (Halpern, 2003).

Other KRs use logics less expressive than the predicate calculus in order to carry out inference more efficiently. For example, in the propositional calculus, the only logical symbols are Boolean operators, and a separate symbol must be used for each atomic proposition. Thus there is one symbol for the proposition, 'Leonardo painted the Last Supper,' another for the (false) proposition, 'Rodin painted the Last Supper,' and so on. General rules like formula (1) above can then only be stated by stating every instance for every pair of a painter and a painting. Despite this proliferation of symbols and formulas, however, as long as the underlying domain is not too large, it is often more effective to use the propositional calculus than the predicate calculus, because the best methods of inference known are much faster for the propositional calculus than the predicate calculus (Gomes et al., 2008).

The chief virtues of logic-based representation are precision and modularity. The meaning of a formula in a logic-based representation is completely and exactly fixed by the logic and the meaning of the symbols included. It is independent of any other formula in the knowledge base, and any other aspect of the knowledge base. Moreover, in a monotonic logic, the validity of any proof depends only on the formulas used in the proof and not on any other aspect of the knowledge base. Another important virtue is that procedures for deductive inference over logic-based knowledge bases are known. These procedures characteristically generate only valid inferences and all valid inferences (in the technical jargon, they are sound and complete), and, are in practice, they can often be implemented so as to find useful answers quickly.

Their chief drawback is difficulty of use. Building a knowledge base in a logic-based representation inevitably requires a large investment of expert labor. Moreover, even experts in KR find the encoding of complex domains in logical languages surprisingly difficult and error-prone. Logic-based languages also tend to be fragile. In principle, any inconsistency in any part of a logical theory invalidates the entire theory, and in practice small and subtle inconsistencies in foundational parts of a logic-based knowledge base can lead to bizarre and unpredictable results.

Semantic Networks

A semantic network consists of labeled nodes connected by labeled arcs (Figure 2). A node may represent an individual

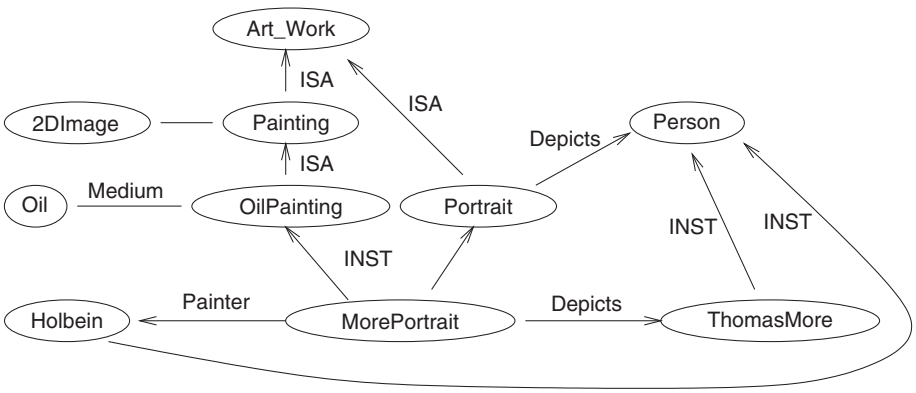

Figure 2 Semantic network consisting of labeled nodes connected by labeled arcs.

such as Holbein or a category such as oil paintings. An arc represents a two-place relation between nodes, such as the relation 'painter' between a work of art and a person.

The most important arcs in any semantic network are those labeled 'INST' and 'ISA.' An INST link goes from an individual *I* to a category *C* and asserts that *I* is an instance of *C*. An ISA link goes from category *C* to category *D* and asserts that *C* is a subset of *D*. Implicitly, therefore, if there is a chain of ISA links from *C* to *D*, then it can be inferred that *C* is a subset of *D*. If there is an INST link from *I* to *C* and then a chain of ISA links from *C* to *D*, then it can be inferred that *I* is an instance of *D*. For instance, in Figure 2, it can be inferred that oil paintings are works of art, and that the portrait of Thomas More is a work of art.

The central reasoning mechanism in a semantic network is inheritance: a category and its instances inherit the properties of the categories that contain them. For instance, in Figure 2, the node MorePortrait inherits the Medium arc to Oil from OilPainting, the Form arc to 2D-Image from Painting, and the Depicts arc to Person from Portrait. That is, the system can infer that the portrait of More is a two-dimensional image from the facts that it is a painting, and that all paintings are two-dimensional images.

The advantages of the semantic network architecture are that it is easy and natural to use, its meaning can be defined precisely, and the inheritance mechanism is easy to implement and efficient. The chief drawback is that it is not very expressive. In addition, the apparent simplicity of the representation can end up working against precision; an inexpert user can easily construct a semantic network that looks plausible, but that uses the same notation with two different meanings (Woods, 1975).

The minimal architecture described above is often supplemented with additional features, such as 'meta-arcs' that connect two arcs and describe the logical relations between them (Sowa, 2000). However, the more such features are added, the less simple and natural the representation becomes.

Frame-Based Systems

A frame represents either an individual or a category, and associates with its subject a collection of pairs of a slot with a corresponding filler. A slot is the name of a feature of an individual, and a filler for a slot of an individual is the value of that feature; for example, the Color slot for StatueOf Liberty has filler Green. The filler of a slot for a category contains information about the possible fillers of that slot for individual instances of the category. This can take many different forms; it can be a particular value, a range of values, a constraint on values, or a function that computes the value.

For instance, in frame OilPainting in Figure 3, the value of the Date slot is a constraint asserting that the date of a painting must take place within the painter's lifetime. (The specific syntax used here is invented, but the features are typical of frame-based systems.) As in semantic nets, frames inherit fillers through INST links, which connect an individual to a category, and through ISA links, which connect a subcategory to a supercategory. Other forms of inference in frame-based systems include computing or verifying the value of a slot using inherited procedures or constraints and categorizing a new

```
Frame MorePortrait
  Inst:OilPainting;
  Genre:Portrait;
  Depicts:ThomasMore;
  Shape:Rectangle (29.5 inch, 23.7 inch);
  Surface:Wood;
  Date:1527;
  Painter:Holbein;
  Owner:Frick;
  Image:more.gif {file with displayable image};
EndFrame

Frame OilPainting
  Isa:Painting;
  Medium:Oil;
  Surface: <OR> [Canvas, Wood];
  Date: <CONSTRAINT> [@.Painter.Birth <@.Date ≤ @.Painter.Death];
  Painter: <OR> [ <INSTANCE> Person, <INSTANCE> Workshop];
  Owner: <OR> [Lost, Destroyed, <INSTANCE> Person, <INSTANCE> Institution];
EndFrame
```

Figure 3 Frame-based systems.

individual as an instance of a category based on its features (Minsky, 1975) (Borgida et al., 1989).

Architectures for Large-Scale Systems

Several representational architectures are available for use in large-scale KR. The best known large-scale representational project is CYC, which has been under development since 1984 by a team led by Douglas Lenat (1995). The representation language is called CYCL. CYC is a commercial product, and the system as a whole is proprietary. Although substantial parts, including the structure of CYCL, the upper ontology, and a fraction of the content are openly available, very little has been published about CYC, and scientists outside the project have consistently found it difficult to determine how it can be effectively used, what content it contains, and how it has been evaluated.

Other representational architectures for large-scale systems include KIF (Knowledge Interchange Format) developed at Stanford and KQML (Knowledge Query and Manipulation Language) developed at the University of Maryland. All these contain a central logical component with many extensions of various kinds.

Fundamental Domains

The most difficult problems facing the designer of a knowledge base are those of content: what knowledge is needed for a given application, what concepts are needed, and how can the knowledge be effectively encoded in a given architecture?

KRs for specialized applications generally involve a protracted collaboration between the designers of the knowledge base and the domain experts. The interaction brings to light central concepts, distinctions, and forms of inference in the applications that are not evident to either group at the start (Friedland et al., 2004).

In encoding nonspecialized, commonsense knowledge, a popular methodology is to begin by defining a microworld, a small, coherent domain of study. Aspects of the real world that lie outside the microworld are ignored or dealt with ad hoc. A collection of commonsensically obvious inferences in the microworld is assembled. The researcher constructs a formal model that suffices to justify these inferences, defines symbols to express the concepts used in the inferences, and expresses the knowledge involved in the inferences in terms of these symbols in the selected representational format (Davis, 1990, 1998).

Time

The issues involved in content may be illustrated in a discussion of representations of time, certainly the most ubiquitous domain in automated reasoning.

Ontologically, time may be taken to be discrete or continuous, to be linear or branching, and to consist of points or of intervals. Which choice is made depends on the applications being considered. For instance, in a chess-playing program, it is natural to take time to be discrete, as nothing significant happens between successive states of the board, and to be branching, to express the options available to each player at each stage. In reasoning about physical processes, it is usually desirable to take time to be continuous, in order to describe continuous change, and linear, since no choices are being made. (Uncertainty about the future, as opposed to choices about the future, is easily accommodated in a linear theory of time.)

Many different representations have been devised for time. Some of the most popular include the following:

- *Time stamps*: If a linear model of time is used, and clock times are known precisely and are important, then each event can be marked by date and clock time.
- *Numerical constraints on time*: If a linear model of time is used, and clock times are important but are not fully known, then a reasonable representation can use a collection of symbolic constraints between the times of events, such as 'The date of the painting of More's portrait comes after Holbein's birth and before his death.'
- *The situation calculus*: This theory uses a branching, discrete model of time. There are three primary types of entities: situations, which are essentially instants of time; fluents, which are propositions whose value changes over time, such as 'Obama is President'; and actions, which cause one situation to change to a resultant situation. The two principal symbols are Result(s,a), which maps a situation s and an action a to the situation that results if a is performed in s, and Holds(s,f), the assertion that fluent f holds in situation s (McCarthy and Hayes, 1969).
- *The event/interval calculus*: Another representation for time popular in KR research is based around intervals and events. There are 13 possible temporal relations between intervals; interval i may entirely precede interval j or the end of i may coincide with the beginning of j and so on. The interval language contains primitives to describe these 13 relations, plus the relation Occurs(i,e), meaning event e occurs in interval i (Allen, 1983).
- *Tense logic*: This was developed in the philosophical literature to formalize the tenses of natural language. It is a modal logic, containing operators like 'φ will always be true,' 'φ has been true at some time in the past,' and so on (van Benthem, 1983).
- *Dynamic logic*: This representation is popular in formal theories of programming languages. It uses a discrete, branching model of time. Statements have the form 'after p has been executed, φ will be true', where p is an action or a computer program.

Other Domains

Other commonsense domains that have been extensively studied in the KR literature include the following:

- Spatial reasoning, in particular the development of reasoning systems that can effectively use qualitative information about shape and other spatial relations, rather than requiring exact geometric information.
- Spatiotemporal reasoning, in particular reasoning based on the fact that spatial relations change continuously over time.
- *Physical reasoning*: Here the challenge is to characterize the physical world in the terms in which the human agent encounters it, sometimes called the mesoscopic scale.
- *Mental states of agent*: Knowing, believing, hoping, fearing, and so on. The chief challenge here is to develop a theory

that is precise enough to support useful reasoning but not so idealized as to be entirely unrealistic.
- Communication between agents.

Learning and Knowledge Representation

In view of the immensity of the task of hand encoding a knowledge base of all commonsense knowledge, or even of developing a representation language adequate for such a knowledge base, and in view of the great success that machine learning has had in the past 20 years in many different tasks, the proposal has naturally been made that machine learning techniques be employed to develop both the representation language and the content for a knowledge (Etzioni et al., 2008).

Almost all such recent projects work from texts gathered from the World Wide Web. Projects differ widely in terms of the kinds of information that is gathered:

- The program may have a fixed set of categories and relations and may gather specific instances of these. For instance, the program may know of the category Person and the relation Father, and learn from text that Odysseus was a person and that Odysseus was the father of Telemachus.
- The program may learn both new relations and new instances; for instance, it may learn that Athlete and Team are categories, that Babe Ruth is an athlete, and that PlaysFor is a relation.
- The program may additionally learn semantic relations between categories and relations. For instance, it may learn that Athlete is a subcategory of Person and that PlaysFor is a relation between Athlete and Team.
- The program may learn more complex logical rules or probabilistic correlations, such as, 'If food F is made from ingredient G and G contains chemical C, then F contains chemical C.'

In general, such projects work by building up, side-by-side with the conceptual knowledge base, a collection of linguistic patterns that indicate the target relation and that can be matched against the text. For instance, if a phrase in the text matches the pattern '<*Class*> such as <*Instances*>,' it suggests that the enumerated instances in the text may be examples of the specified category. The accuracy and the coverage of any individual pattern may be low, but if many such patterns are used and are applied over a large corpus of text, the combined evidence may be substantial.

See also: Axiomatic Theories; Concept Learning and Representation: Models; Deductive Reasoning Systems; Logics for Knowledge Representation; Nonstandard Reasoning.

Bibliography

Allen, J., 1983. Maintaining knowledge about temporal intervals. Communications of the ACM 26 (11), 832–843.
Balduccini, M., Baral, C., Lierler, Y., 2008. Knowledge representation and question answering. In: van Harmelen, F., Lifschitz, V., Porter, B. (Eds.), Handbook of Knowledge Representation, Elsevier, Amsterdam, pp. 779–820.
Borgida, A., Brachman, R., McGuiness, D., Resnick, L., 1989. CLASSIC: a structural data model for objects. In: Proc. SIGMOD '89. Oregon, Portland, pp. 59–67.
Davis, E., 1990. Representations of Commonsense Knowledge. Morgan Kaufmann, San Mateo, California.
Davis, E., 1998. The naive physics perplex. Artificial Intelligence Magazine 19, 51–79.
Dumais, S., 2005. Latent semantic analysis. Annual Review of Information Science and Technology 38, 188–230.
Etzioni, O., Banko, M., Soderland, S., Weld, D., 2008. Open information extraction from the web. Communications of the ACM 51, 68–74.
Friedland, N., et al., 2004. Project Halo: toward a digital Aristotle. Artificial Intelligence Magazine 25, 29–47.
Genesereth, M., Nilsson, N., 1987. Logical Foundations of Artificial Intelligence. Morgan Kaufmann, San Mateo, California.
Ghallab, M., Nau, D., Traverso, P., 2004. Automated Planning: Theory and Practice. Morgan Kaufmann, San Mateo California.
Glasgow, J., Narayanan, N., Chandrasekaran, B., 1995. Diagrammatic Reasoning. MIT Press, Cambridge, Massachusetts.
Gomes, C., Kautz, H., Sabharawal, A., Selman, B., 2008. Satisfiability solvers. In: van Harmelen, F., Lifschitz, V., Porter, B. (Eds.), Handbook of Knowledge Representation, Elsevier, Amsterdam, pp. 89–134.
Halpern, J., 2003. Reasoning About Uncertainty. MIT Press, Cambridge.
LaValle, S., 2006. Planning Algorithms. Cambridge University Press, Cambridge.
Lenat, D., 1995. CYC: a large-scale investment in knowledge infrastructure. Communications of the ACM 38 (11), 33–38.
Mates, B., 1972. Elementary Logic. Oxford University Press, Oxford.
McCarthy, J., 1968. Programs with common sense. In: Minsky, M. (Ed.), Semantic Information Processing. MIT Press, Cambridge, Massachusetts, pp. 403–418.
McCarthy, J., Hayes, P., 1969. Some philosophical problems from the standpoint of artificial intelligence. In: Metzler, B., Michie, D. (Eds.), Machine Learning 4. Edinburgh U. Press, Edinburgh.
Minsky, M., 1975. A framework for representing knowledge. In: Winston, P. (Ed.), The Psychology of Computer Vision. McGraw-Hill, New York, pp. 211–280.
Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo, California.
Reiter, R., 2001. Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems. MIT Press, Cambridge, MA.
Sowa, J., 2000. Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks/Cole, Pacific Grove, California.
Turner, R., 1984. Logics for Artificial Intelligence. Wiley, New York.
van Benthem, J., 1983. The Logic of Time. Reidel Pubs, Dordrecht.
van Harmelen, F., Lifschitz, V., Porter, B., 2008. Handbook of Knowledge Representation. Elsevier, Amsterdam.
Woods, W., 1975. What's in a link: foundations for semantic networks. In: Bobrow, D., Collins, A. (Eds.), Representation and Understanding: Studies in Cognitive Science. Academic Press, New York, pp. 35–82.
Zadeh, L., 1987. Commonsense and fuzzy logic. In: Cercone, N., McCalla, G. (Eds.), The Knowledge Frontier: Essays in the Representation of Knowledge. Springer-Verlag, New York, pp. 103–136.

Knowledge Society, History of

Nico Stehr, Zeppelin University, Friedrichshafen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Past theories of society, choose to designate, quite properly, those attributes of social relations that are constitutive of the specific nature of that society as identifying labels. Thus, such names as 'capitalist' society or 'industrial' society were created. For the same reasons, the now-emerging form of society represents a 'knowledge' society because the constitutive mechanism or the identity of modern society increasingly is driven by 'knowledge'. Knowledge has become the central economic resource and source of additional economic growth. The article on knowledge society explores the emergence, development, and current state of theorizing about modern societies as knowledge societies.

Past theories of society, choose to designate, quite properly, those attributes of social relations that are constitutive of the specific nature of that society as identifying labels. Thus, such names as 'capitalist' society or 'industrial' society were created. For the same reasons, the now-emerging form of society represents a 'knowledge' society because the constitutive mechanism or the identity of modern society increasingly is driven by 'knowledge'. Knowledge has become the central economic resource and source of additional economic growth (Drucker, [1968]1972: p. 40; Böhme and Stehr, 1986; Stehr, 2002). (I utilize observations about the history of the idea of modern society as a knowledge society first summarized in Stehr (1984). I have added considerable material to the earlier reflections.)

The term knowledge society first appears, as far as I can see, in the name of 'benevolent' societies in England in the nineteenth century. The purpose of one of these societies, the *Provident Knowledge Society* under the patronage of Lord Derby, Lord Shaftesbury, and other distinguished men "is to make regular weekly saving a national habit, and so to increase the facilities for saving that it shall be as easy for a man to put by a small sum as it is now for him to spend that sum in beer or spirits. The Society has published a series of tracts upon *Penny Banks, Pensions, Insuring One's Life*, and other subjects, which are intended to promote thrift and forethought among the working classes; and of these tracts many thousands have been circulated" (The British Medical Journal, 1875: p. 283; also The Scottish Historical Review, 1920). Inasmuch as the benevolent knowledge societies in England saw their mission to enlighten and educate the uninformed public, especially members of the working class, there is at least some resemblance to features of the modern day knowledge societies that also rest on the broad dissemination of knowledge, information and knowledgeability throughout the society.

Early Usages of the Term Knowledge Society

In its contemporary, social scientific version, the term knowledge society and the idea that modern societies are knowledge societies are much more recent. The usage and development of the idea of that *modern society* is a knowledge society date to the early 1970s and more prominently to the 1980s and later.

One of the first social scientists to employ the term 'knowledgeable society' is Robert E. Lane (1966: p. 650). He justifies the use of this concept by pointing to the growing societal relevance of *scientific knowledge* and defines a knowledgeable society, in a 'first approximation', as one in which its members

> (a) inquire into the basis of their beliefs about man, nature, and society; (b) are guided (perhaps unconsciously) by objective standards of veridical truth, and, at the upper levels of education, follow scientific rules of evidence and inference in inquiry; (c) devote considerable resources to this inquiry and thus have a large store of knowledge; (d) collect, organize, and interpret their knowledge in a constant effort to extract further meaning from it for the purposes at hand; (e) employ this knowledge to illuminate (and perhaps modify) their values and goals as well as to advance them. Just as the 'democratic' society has a foundation in governmental and interpersonal relations, and the 'affluent society' a foundation in economics, so the knowledgeable society has its roots in epistemology and the logic of inquiry.

Lane's conception of a knowledgeable society is coupled rather closely to the promise of a particular theory of science and reflects, also, the great optimism or fear, as the case may be, in the early 1960s which suggested that science would somehow allow for the possibility of a society in which common sense would be replaced by scientific reasoning. That is, as Lane stresses in his definition, members of the knowledgeable society are guided in their conduct, if only subconsciously, by the standards of 'veridical truth'.

The very promise Lane attached to scientific knowledge as the motor that would transform modern society into a knowledgeable society is the reason for other social scientists to reflect about the emergence of a 'technical state'. One prominent social scientist who warned about the possibility of a technical state is Helmut Schelsky. Schelsky was one of the most eminent social scientists of the postwar era in Germany. His influence reached well beyond the narrow boundaries of academic social science. His influence was based on the distinctive understanding of the evolution of modern society and the terms he coined to capture core cultural and socio-structural developments. The theories he offered were distinctly middle range and not of a grand design. Lane and Schelsky share in the confidence that scientific knowledge will assume dominance

throughout society and in the case of the political system reduce decision to technical matters. A brief quote from Schelsky ([1961]1965: p. 459) transports the essence of the meaning of the notion of the technical state: "The 'technical state or government' deprives … democracy of its substance. Technical and scientific decisions cannot be subject to democratic decision making, for in this manner they only become ineffective." Helmut Schelsky adds what is perhaps obvious, namely once political decisions are executed based on scientific knowledge alone, the basic premise of democratic participation is eliminated: that is, the assumption that reason and reasoned judgments are equitably distributed throughout society.

Peter Drucker and Daniel Bell's Conceptions of the Knowledge Society

In *The Age of Discontinuity*, Peter Drucker (1968) explicitly and consistently employs the term *knowledge society*. The general thesis places knowledge "as central to our society and as the foundation of economy and social action" (Drucker, [1968] 1992: p. 349). In this most general sense, Drucker's concept of a knowledge society resonates closely with the idea of modern society as a knowledge society, as explicated in subsequent decades. Although Drucker's emphasis of knowledge is, in many ways, pioneering, it is not evident, however, whether he attributes to the knowledge principle, in the late 1960s at least, the same centrality for society as does, for example, at about the same time, Daniel Bell. Nor is there any direct evidence of cross fertilization between Drucker and Bell. Nonetheless, both theorists contributed with their reflections about the rise in the significance of knowledge as an economic resource displacing leading observations about the existence of a *managerial capitalism* immediately after World War II (cf Chandler, 1977; Burnham, 1941).

Drucker goes to considerable and informative lengths to describe some of the novel features and functions of contemporary knowledge. His definition of what counts as knowledge in a knowledge society is not restricted, as is the case for Lane's knowledgeable society, to scientific knowledge. Drucker's account of the emergence of knowledge societies and the knowledge-based economy must be counted among Drucker's original contributions to the theory of the knowledge society. The popular response to the questions about the origins of the knowledge society and the knowledge-based economy typically refers (see Drucker, [1968]1972: p. 279) to the more technically demanding and more complex tasks of jobs. Drucker rejects the 'demand' induced transformation of industrial society and instead points out that "the direct cause of the upgrading of the jobs is … the upgrading of the educational level of the entrant into the labor force." Moreover, Drucker suggests that the stimulus for the increasing demand for knowledge-based work has to do less with changing technological regimes, the growing complexity and specialization of the economy, or enhanced functional steering and coordination needs. The growth of knowledge workers has more to do with the substantial extension in the working life span of individuals and the enhanced knowledge with which individuals come to the labor market in the first place. If we follow Drucker, it is not so much the *demand* for labor and particular skills as the result of more complex and exacting jobs, but a societal transformation that results in the *supply* of highly skilled labor that in turn underlies the transformation of the world of work.

Among the surprising even startling properties of the transformation of the labor market is that the American economy at least was able to "satisfy the expectations of all these people with long years of schooling…. As a result of the change in supply, we now have to create genuine knowledge jobs, whether the work itself demands it or not. For a true knowledge job is the only way to make highly schooled people productive…That the knowledge worker [Drucker first proposed the term knowledge worker in 1959 in his *Landmarks of Tomorrow*] came first and knowledge work second – that indeed knowledge work is still largely to come – is a historical accident. From now on, we can expect increasing emphasis on work based on knowledge, and especially skills based on knowledge" (Drucker, [1968]1972: pp. 285–286; also Drucker, 1992: p. 91).

For a contemporary contrast and comparison between Peter Drucker's notion of the knowledge society and self-organizing knowledge labor as *cognitive capitalism* (cf Boutang, [2007] 2011) and the more recent Italian Marxist inspired analysis of capitalism's cognitive phase (Hardt and Negri, 2000), see Peters and Reveley (2012).

Daniel Bell (1968: p. 198) also employs the term 'knowledge society' in the context of his discussion of the emergence of *postindustrial society*, a designation he ultimately prefers. Bell at times uses the concept knowledge society interchangeably with the notion of 'postindustrial society'. For example, he exclaims at one point in his discussion that "the postindustrial society, it is clear, is a knowledge society" (Bell, 1973: p. 212). The basic justification for such an equivalence is, of course, that 'knowledge is a fundamental resource' of postindustrial society. As a matter of fact, Bell (1973: p. 37) indicates that he could have substituted 'knowledge society' for 'postindustrial society' because either term, and others, for example, 'intellectual society' (Bell, 1964: p. 49), might be just as apt in describing at least some salient aspects of the emerging structure of society that he proposes to examine in his study of the structure and culture of modern society.

Radovan Richta's (1969) theory of the scientific-technical revolution constitutes the socialist counterpart to Bell's theory of society. Bell (1973: p. 212) argues that postindustrial society is a knowledge society for two major reasons: (1) "the sources of innovation are increasingly derivative from research and development (and more directly, there is a new relation between science and technology because of the centrality of theoretical knowledge)", and (2) "the weight of the society – measured by a larger proportion of gross national product and a larger share of employment – is increasingly in the knowledge field." The pace and scale of the translation of knowledge into technology provides the basis for the possibility of modernity. Thus, if there is a "radical gap between the present and the past, it lies in the nature of technology and the ways it has transformed social relations and our ways of looking at the world" (Bell, 1968: p. 174).

Daniel Bell's theory of postindustrial society, Radovan Richta's theory of scientific technological revolution, Helmut Schelsky's ([1961]1965) theory of scientific civilization, or John K. Galbraith's (1967) theory of the new industrial state,

adopt quite readily those principles of theory construction characteristic of classical (sociological) theories of society. Among such principles is not merely a considerable aura of optimism about the future, and the conviction that there is a radical, even revolutionary, break in social relations but often also rather deterministic, law like assumptions about the tempo, the course and the direction of social transformations.

What is evident in both Peter Drucker's and Daniel Bell's designation of modern society as knowledge societies is their common assumption that the significant transformations that bring about new constitutive principles of social relations and hence a new societal configuration are changes within the economic system that then radiate throughout society and affirm the dominance of the economy. As a result, the emergence of knowledge or postindustrial societies out of industrial societies has not undermined in the eyes of Drucker and Bell, the preeminence of the economy. In the course of further reflections about knowledge societies, hypotheses about a reversal of the equation of sub- (i.e., the material foundations) to superstructure (the cultural or cognitive foundations) become a primary research agenda.

From the late 1970s onward, the analysis of the modern society as knowledge societies became more sophisticated looking both backward to the existence of past knowledge societies as well as forward in time toward major social transformations of contemporary society (Böhme and Stehr, 1986). The theory of the knowledge society was developed alongside and competed with theories of modern society such as the *information society*, the *risk society* or the *network society*.

What justifies the designation of the emerging society as a knowledge society rather than, as is at times the case, a science society (e.g., Kreibich, 1986), an information society (e.g., Nora and Minc, 1980), a network society (Castells, 1996; Schiller, 1999), a postindustrial society (Bell, 1964), or as technological civilization (Richta et al., 1969)? A number of reasons are relevant for the choice of the term. Much of the discussion about the information society is animated by a concern with the "production, processing, and transmission of a very large amount of data about all sorts of matter – individual and national, social and commercial, economic and military" (Schiller, 1981: p. 25). But every society has to transmit information. Less, or very little, is said about the genesis of the substance of information, the media of communication, especially human ones, or the reasons for the demand for, and changes brought about by, the contents of the information, which is communicated. Nor are discussions about the information society usually concerned with questions of solidarity or domination in modern society and whether any economic effects of this spread of information and communication technologies cannot just as well be accommodated within more conventional economic discourse, namely "as phenomena best understandable in terms of long-established and familiar market-based criteria" (Schiller, 1981: p. xii). The category 'postindustrial society' is misleading because 'industry' or manufacturing does not vanish in modern societies. True, industry is transformed, it is host to many fewer jobs but it is misleading to think that it will disappear. The contribution of the manufacturing sector in terms of the value generated remains unchanged.

Mature Theories of the Knowledge Society and Its Antecedents

The historical emergence of 'knowledge societies' does not occur suddenly; it represents not a revolutionary development, but rather a gradual process during which the defining characteristic of society changes and a new one emerges. Even today, the demise of societies often is as slow as is their beginning. Social transformations rarely occur in spectacular leaps. Nonetheless, proximity to significant social, economic, and cultural changes seems to assure that what comes into view appears to be particularly significant and exceptional. The disruption of routines displaces orientations but it remains "difficult to pinpoint the actual crystallization of a new state of affairs so as to identify, both clearly and unequivocally, the emergence of a new society and its new modes of behavior" (Narr, [1979]1985: p. 32).

Knowledge societies do not come about as the result of a simple unimodal unfolding and in an unambiguous fashion. Knowledge societies do not turn into some kind of one-dimensional social figurations. Knowledge societies become similar by remaining or even becoming dissimilar. New technological modes of communication and transportation break down the distance between groups and individuals but the isolation between regions, cities, and villages remains. The world opens up and creeds, styles and commodities mingle, yet the walls between convictions of what is sacred do not come tumbling down. The meaning of time and place erodes while boundaries are celebrated.

Modern society was, until recently, conceived primarily in terms of property and labor. Labor and property (capital) have had a long association in social, economic, and political theory. Work is seen as property and as a source of emerging property. In the Marxist tradition, capital is objectified, encapsulated labor. On the basis of these attributes, individuals and groups were able or constrained to define their membership in society. In the wake of their declining importance in the productive process, especially in the sense of their conventional economic attributes and manifestations, for example as 'corporeal' property such as land and manual work, the social constructs of labor and property themselves are changing. While these features of labor and property certainly have not disappeared entirely, a new principle, 'knowledge', has been added which, to an extent, challenges as well as transforms property and labor as the constitutive mechanisms of society.

The foundations of economic growth or the "wealth-creating activities will be neither the allocation of capital to productive uses nor 'labor' – the two poles of nineteenth- and twentieth century economic theories … The representative social groups will be neither the 'capitalist' nor the 'worker' … The ruling group will be the knowledge workers, knowledge executives, knowledge professionals, and knowledge entrepreneurs" (Drucker, 1992: p. 94; also Stehr, 1992, 2002; Grundmann and Stehr, 2012).

Dating Modern Knowledge Societies

Radovan Richta et al. (1969: p. 276) date the beginning of the profound transformation of modern society, through the

scientific and technological revolution, from the 1950s. Daniel Bell (1973: p. 346) argues, even though he points out that it really is foolhardy to give precise dates to social processes, that the 'symbolic' onset of the postindustrial society can be traced to the period since the end of World War II, for it was during this era that a new consciousness about time and social change began to emerge. In contrast, Block and Hirschhorn (1979: p. 368), who ask about the beginning of the emergence of the new productive forces typical for postindustrial society, namely, knowledge, science, and technology, and in particular the period when they began to make a decisive difference in production, reach the conclusion that the 1920s already saw such a qualitative shift occurring. In the 1920s, at least in the United States, the input of labor, time, and capital were constant or had begun to decrease while output had begun to rise. In economic terms therefore, knowledge has become the crucial source of (added) value.

Of course knowledge has always had a function in social life; as a matter of fact, one could speak justifiably of an anthropological constant: human action is knowledge based. Social *groups* and social roles of all types depend on, and are mediated by, knowledge. Relations among *individuals* are based on knowledge of each other. Similarly, power has frequently been based on advantages in knowledge, not only on physical strength. And, last but not least, societal reproduction is not merely physical reproduction but, in the case of humans, always cultural, i.e., reproduction of knowledge. In short, knowledge is not confined to a particular social system (cf Luhmann, 1990: p. 147). In retrospect, one is able to describe a variety of ancient societies as knowledge societies, for example, ancient Israel, which was a society structured by its religious law like Tora-knowledge. Ancient Egypt was a society in which religious, astronomical, and agrarian knowledge served as the organizing principle and the basis of authority.

Contemporary society may be described as a knowledge society based on the penetration of all its spheres of life by knowledge. As a matter of fact, knowledge is not coupled to a particular social system. Marxist theories of society have always assigned decisive importance to the forces or means of production for societal development since "man's understanding of nature and his mastery over it by virtue of his presence as a social body … appears as the great foundation-stone (*Grundpfeiler*) of production and of wealth", so that general knowledge becomes a direct force of production (Marx, [1939–1941]1973: p. 705). Contemporary Marxist theories, especially through the notion of the Scientific-Technological Revolution developed by Radovan Richta and others, have analyzed scientific and technical knowledge as the principal motor of change. Max Weber's seminal inquiry into the unique features of Western civilization stresses the pervasive use of reason to secure the methodical efficiency of social action. The source of rational action and, therefore, of rationalization is found in particular intellectual devices. The theory of industrial society, as developed by Raymond Aron (1962), which encompasses both socialist and capitalist forms of economic organization, stresses first and foremost the extent to which science and technology shape the social organization of productive activities and, therefore, other forms of life in society. More recent theories of postindustrial society and similar efforts at forecasting the course of social evolution of industrial society, in particular those by Daniel Bell, have elevated, as we have seen, theoretical knowledge as the axial principle of society.

The Restrictive Concept of Knowledge

Knowledge represents a capacity to act (see Stehr, 1984). The knowledge referred to in many of the early theories of the knowledge society, and the groups of individuals that acquire influence and control with it, are conceptualized narrowly. Paradoxically perhaps, there is a tendency to overestimate the efficacy of 'objective' technical-scientific or formal knowledge and the immediate 'performative' capacity of knowledge. The initial theories of modern society that evolve into the theory of the knowledge society lacked sufficient detail and scope in their conceptualization of the 'knowledge' supplied, the reasons for the demand of more and more knowledge, the ways in which knowledge travels, the rapidly expanding groups of individuals in society who, in one of many ways, live off knowledge, the many forms of knowledge that are considered as pragmatically useful, failed to distinguish between information and knowledge, the application of knowledge to knowledge, the commodification of knowledge and the various effects which knowledge may have on social relations. Peter Drucker's (1993: pp. 52–53) idea that there has been "a radical change in the meaning of knowledge" is illuminating: "In both West and Asia knowledge had always been seen as applying to *being*. Almost overnight it came to be applied to *doing* (my emphasis)."

Contemporary theories of knowledge societies advocate a conception of scientific knowledge that is distant to any notion associated with technological or scientific determinism, especially the one-sided or unidirectional forms of determinism. The constraining features of science and technology are, however, by no means underestimated. But in contradistinction to most arguments in favor of technological and scientific determinism, and the theories of society associated with such views, the crucial point about knowledge societies is that science and technology have simultaneously strong features that allow for resistance to homogeneous transformation (Stehr, 2013). That is, science and technology have important enabling features that increase the number of available strategies, heighten flexibility or affect the ability of the powerful to exercise control and constraining forces that limit choices, reduce options and enhance the fragility of major social institutions, and limit the 'power' of knowledge and impose penalties and risks (e.g., Stehr, 2001). It is therefore by no means contradictory to maintain that knowledge societies can simultaneously become more standardized and more fragile. Generally, it is important to avoid overstating the extent to which science and technology are forces that merely are means of control and regulation and therefore constrain human agency and limit social action. They do, but that is only part of their consequences. The other part is represented by the 'opposite' because science and technology enter relational fields of social action and can assume quite different values or outcomes especially for opposing social forces and purposes.

The Economic System of the Knowledge Society

The emergence of knowledge societies signals first and foremost a radical transformation in the *structure of the economy*. Productive processes in *industrial society* are governed by a number of factors, which appear to be on the decline in their significance as conditions for the possibility of a changing, particularly growing economy. The common denominator of the changes in the structure of the economy seems to be a shift from an economy driven and governed, in large measure, by 'material' inputs into the productive process and its organization to an economy in which transformations in productive and distributive processes are determined much more by 'symbolic' or knowledge-based inputs.

The economy of industrial society is initially and primarily a *material economy* and then changes gradually to a monetary economy; for example, Keynes' economic theory, particularly his *General Theory* (1936), reflects this transformation of the economy of industrial society into an economy affected to a considerable extent by monetary matters. But as more recent evidence indicates, the economy Keynes described now becomes a (nonmonetary) *symbolic economy* (cf Drucker, 1980: p. 8). The changes in the structure of the economy and its dynamics are increasingly a reflection of the fact that *knowledge* becomes the leading dimension in the productive process, the primary condition for its expansion and for a change in the limits to economic growth in the developed world. In the knowledge society, most of the wealth of a company is increasingly embodied in its creativity and knowledge. With the exception of the most standardized commodities and services, factors other than "the amount of labor time or the amount of physical capital become increasingly central" (Block, 1985: p. 95) to the economy of advanced societies (see especially Drucker, 1986; Lipsey, 1992).

One of the few extensive and early empirical contributions to the analysis of the nature and the emergence of a knowledge society and the knowledge-based economy (knowledge industry) in particular more than five decades ago may be found in the work of the economist Fritz Machlup (1962, 1981; 1984). His analysis is constituted by an elaborate apparatus of empirical, often census type information about observable shifts and trends, such as in the occupational structure, intended to show that a knowledge society is indeed emerging and can therefore be documented in its own self-exemplifying terms, that is, in a quantitative (rational) manner. The most recent data generated as part of the Machlup research programme to quantify the overall expenditures for knowledge production in the United States may be found in Rubin and Huber's (1986) study. Their attempt to measure knowledge incomes and expenditures connected with commodities and services constitutes the explicit effort to extend Fritz Machlup's (1962) investigation into the proportion of the Domestic Economic Product that goes to knowledge production.

The Nature of Knowledge Societies

Advanced theories of the knowledge society do not argue that knowledge societies become uniform social and intellectual entities. They allow, for example, for the coexistence, even interdependence, of historically distinct forms of social organization and thought. Knowledge societies do not spell the end of ideology or irrationality. Nor is scientific knowledge, as a cultural ensemble, merely a way of deciphering the world; it is also a model for the world (capacity to act).

However, in order to demonstrate and appreciate fully the significance of knowledge for societies and social action generally, and for advanced societies in particular, one first has to formulate a sociological concept of knowledge. One must be able to differentiate between what is known, the content of knowledge, and knowing itself. What is it, then, that we know? Some examples taken from the *Oxford Dictionary of Current English* indicate the following instances: "Every child knows that two and two make four. He knows a lot of English. Do you know how to play chess? I don't know whether he is here or not." These examples show that knowing is a relation to things and facts, but also to laws and rules. In any case, knowing is some sort of participation: knowing things, facts, and rules is 'appropriating' them in some manner, including them into our field of orientation and competence. A very important point, however, is that knowledge can be objectified, that is, the intellectual appropriation of things, facts, and rules can be established symbolically, so that in the future in order to know, it is no longer necessary to get into contact with the things themselves but only with their symbolic representations. This is the social significance of language, writing, printing, and data storage. Modern societies have made dramatic advances in the intellectual appropriation of nature and society. There is an immense stock of objectified knowledge that mediates our relation to nature and to ourselves. In a general sense, this advancement has been called, in other contexts, modernization or rationalization. This secondary nature is overgrowing the primary nature of humans. The real and the fictional merge and become indistinguishable; theories become facts and not vice versa, that is, facts do not police theories.

It is only after one acquires a sense of the societal significance of such opposites and oppositions that the full sociological significance of knowledge begins to emerge. Such a perspective assures that one realizes the extent to which knowledge can form the basis for authority, that access to knowledge becomes a major societal resource and the occasion for political and social struggles.

Applied to contemporary society, the question becomes whether knowledge can provide the robust principle for social hierarchies and stratification, for the formation of inequality structures, for the distribution of chances of social and political influence, and for the nature of personal life and, finally, whether knowledge may also prove to be a normative principle of social cohesion and integration even though the variations and alterations in the reproduction of knowledge appear to be enormous. Paradoxically, efforts to entrench necessity in history or eliminate chance from history have produced, at least at the collective level, it seems, its opposite. The role of chance at the collective level continues to be part of the way society comes to be organized. Knowledge societies are (to adopt a phrase coined by Adam Ferguson) the result of human action but not necessarily of deliberate human design. Knowledge societies emerge as adaptations to persistent but evolving needs and changing circumstances of human conduct.

Among the most significant transformations in circumstances that face human conduct is the continuous 'enlargement' of human action.

See also: Culture and Economy; Information Society, Geography of; Information Society; Knowledge (Explicit, Implicit and Tacit): Philosophical Aspects; Knowledge, Anthropology of; Knowledge, Geography of; Knowledge, Sociology of; Scientific Knowledge, Sociology of.

Bibliography

Aron, Raymond, 1962. Eighteen Lectures on Industrial Society. Free Press, New York.

Bell, Daniel, 1964. The post-industrial society. In: Ginzberg, Eli (Ed.), Technology and Social Change. Columbia University Press, New York, pp. 44–59.

Bell, Daniel, 1968. The measurement of knowledge and technology. In: Sheldon, Eleanor B., Moore, Wilbert E. (Eds.), Indicators of Social Change. Concepts and Measurements. Russell Sage Foundation, Hartford, CT, pp. 145–246.

Bell, Daniel, 1973. The Coming of Post-Industrial Society. A Venture in Social Forecasting. Basic Books, New York.

Block, Fred, 1985. Postindustrial development and the obsolescence of economic categories. Politics and Society 14, 416–441.

Block, Fred, Hirschhorn, Larry, 1979. New productive forces and the contradictions of contemporary capitalism. Theory and Society 17, 363–395.

Böhme, Gernot, Stehr, Nico (Eds.), 1986. Knowledge Society. D. Reidel Publishing, Dordrecht.

Boutang, Moulier Y., 2011[2007]. Cognitive Capitalism. Polity, Cambridge.

Burnham, James, 1941. The Managerial Revolution. What Is Happening in the World. John Day, New York.

Castells, Manuel, 1996. The Information Age: Economy, Society and Culture. Volume I: The Rise of the Network Society. Blackwell, Oxford.

Chandler, Alfred D., 1977. The Visible Hand. The Managerial Revolution in American Business. Harvard University Press, Cambridge, MA.

Drucker, Peter F., 1972[1968]. The Age of Discontinuity. Guidelines to Our Changing Society. Harper & Row, New York.

Drucker, Peter F., 1980. Toward the next economics. The Public Interest, (Spring), 4–18.

Drucker, Peter F., 1986. The changed world economy. Foreign Affairs 64, 768–791.

Drucker, Peter F., 1992. The post-capitalist world. The Public Interest 109, 89–100.

Drucker, Peter F., 1993. The rise of the knowledge society. Wilson Quarterly 17, 52–71.

Galbraith, John K., 1967. The New Industrial State. Houghton Mifflin, New York.

Grundmann, Reiner, Stehr, Nico, 2012. Experts: The Knowledge and Power of Expertise. Routledge, London.

Hardt, Michael, Negri, Antonio, 2000. Empire. Harvard University Press, Cambridge, MA.

Keynes, John M., 1936. The General Theory of Employment, Interest and Money. Macmillan, London.

Kreibich, Rolf, 1986. Die Wissenschaftsgesellschaft. Von Galilei zur High-Tech Revolution. Suhrkamp, Frankfurt am Main.

Lane, Robert E., 1966. The decline of politics and ideology in a knowledgeable society. American Sociological Review 31, 649–662.

Lipsey, Richard G., 1992. Global change and economic policy. In: Stehr, Nico, Ericson, Richard V. (Eds.), The Culture and Power of Knowledge: Inquiries into Contemporary Societies. de Gruyter, Berlin and New York, pp. 279–299.

Luhmann, Niklas, 1990. Die Wissenschaft der Gesellschaft. Suhrkamp, Frankfurt am Main.

Machlup, Fritz, 1962. The Production and Distribution of Knowledge in the United States. Princeton University Press, Princeton, NJ.

Machlup, Fritz, 1981. Knowledge and Knowledge Production. Princeton University Press, Princeton, NJ.

Machlup, Fritz, 1984. The Economics of Information and Human Capital. Princeton University Press, Princeton, NJ.

Marx, Karl, 1939–1941. Grundrisse der Kritik der politischen Ökonomie. Europäische Verlagsanstalt, Frankfurt am Main.

Narr, Wolf-Dieter, 1985[1979]. Toward a society of conditioned reflexes. In: Habermas, Jürgen (Ed.), Observations on the Spiritual Situation of the Age. MIT Press, Cambridge, MA, pp. 31–66.

Nora, Simon, Minc, Alain, 1980. The Computerisation of Society. MIT Press, Cambridge, MA.

Peters, Michael A., Reveley, James, 2012. Retrofitting Drucker: knowledge work under cognitive capitalism. Culture and Organization. http://dx.doi.org/10.1080/14759551.2012.692591.

Provident Knowledge Society, 1875. The British Medical Journal 1, 283.

Richta, Radovan, et al., 1969. Civilization at the Crossroads: Social and Human Implications of the Scientific and Technological Revolution. International Arts and Sciences Press, White Plains, NY.

Rubin, Michael R., Huber, Mary T., 1986. The Knowledge Industry in the United States, 1960–1980. Princeton University Press, Princeton, NJ.

Schelsky, Helmut, 1965[1961]. Der Mensch in der wissenschaftlichen Zivilisation. In: Schelsky, Helmut (Ed.), Auf der Suche nach der Wirklichkeit. Gesammelte Aufsätze. Diederichs, Düsseldorf, pp. 439–480.

Schiller, Herbert I., 1981. Who Knows: Information in the Age of the Fortune 500. Ablex, Norwood, NJ.

Schiller, Dan, 1999. Digital Capitalism. Networking the Global Market System. MIT Press, Cambridge, MA.

Stehr, Nico, 1984. The Knowledge Society. Sage, London.

Stehr, Nico, 1992. Experts, counselors and advisors. In: Stehr, Nico, Ericson, Richard V. (Eds.), The Culture and Power of Knowledge. Inquiries into Contemporary Societies. de Gruyter, Berlin, pp. 107–155.

Stehr, Nico, 2001. The Fragility of Modern Societies: Knowledge and Risk in the Information Age. Sage, London.

Stehr, Nico, 2002. Knowledge and Economic Conduct: The Social Foundations of the Modern Economy. University of Toronto Press, Toronto.

Stehr, Nico, 2013. Die Freiheit ist eine Tochter des Wissens. Velbrück Wissenschaft, Weilerswist.

The Fenwick Improvement of Knowledge, 1920. The Scottish Historical Review 17, 219–224.

Knowledge, Sociology of

Volker Meja, Memorial University of Newfoundland, St. John's, NL, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C. Camic, volume 12, pp. 8143–8148, © 2001, Elsevier Ltd.

Abstract

This article examines the history and the development of the sociology of knowledge, the subfield of sociology focused on the relationship between the processes and products of human cognition (including ideas, ideologies, scientific theories, religious, philosophical and political doctrines, moral beliefs, mental categories, cultural and organizational discourses, and the forms and practices of everyday knowing) and their sociocultural origins and consequences. Sociological examination of this relationship has involved linking intellectual products and processes to factors such as sociohistorical contexts, material and cultural conditions, modes of production, power relationships, institutional arrangements, internal organizational processes, and various social positions and interests.

Most inquiries comprehended under the classification of sociology of knowledge as an academic field were familiar long before the term was devised. They have to do with the circumstances that foster – or hamper – the acquisition and reproduction of knowledge and with the circumstances that account for intellectual designs taken or offered as true, notwithstanding the absence of philosophically adequate warrant for such claims. The issues arise in various forms throughout the philosophical canon, hedging about the borders of knowledge by excluding categories of beliefs whose status is seen as a function of factors unrelated to epistemological grounds.

This article examines the history and the development of the sociology of knowledge, the subfield of sociology focused on the relationship between the processes and products of human cognition (including ideas, ideologies, scientific theories, religious, philosophical and political doctrines, moral beliefs, mental categories, cultural and organizational discourses, and the forms and practices of everyday knowing) and their sociocultural origins and consequences. Sociological examination of this relationship has involved linking intellectual products and processes to factors such as sociohistorical contexts, material and cultural conditions, modes of production, power relationships, institutional arrangements, internal organizational processes, and various social positions and interests.

The stages identified in the development of the sociology of knowledge are, first, the long period from antiquity to the early twentieth century, during which thinkers moved from frequent general claims about the social bases and effects of various types of ideas to more specific causal statements, seen in Marx's analysis of the material foundations of ideology and in Durkheim's studies of the social origins of the categories of human thought; second, the period from the early 1920s to the early 1960s, when the sociology of knowledge emerged as a distinct theory-and-research subfield, initially in the writings of Karl Mannheim and Max Scheler, which accented its epistemological and political implications, and then in the more research-oriented work of R.K. Merton and others; third, the period to the present, during which the sociology of knowledge, after an initial decline, has undergone a revival and has been pulled in contrary directions by various agendas.

The sociology of knowledge is a summative term for an assortment of theoretical statements and empirical studies that deal with the relationship between different processes and products of human cognition and other sociocultural factors. On the organizational map of the discipline of sociology in Europe and in North America, the sociology of knowledge has generally occupied a marginal position, with its warrant and scope so contested that it may be argued "that the history of the subdiscipline … has been the history of its various definitions" (Berger and Luckmann, 1966: p. 4). Focusing on manifestations of knowledge that include ideas, ideologies, scientific theories, religious and political doctrines, moral, ethical and philosophical beliefs, world-views, mental categories, cultural and organizational discourses, languages, social norms, and the forms and practices of everyday knowing, sociologists of knowledge (whether self-identified or retrospectively claimed for the area by later contributors) have concerned themselves with the sociocultural consequences of this range of intellectual products and processes and, particularly, with their origins and transformations – developments that have been linked to sociohistorical contexts, material and cultural conditions, modes of production, power relationships in various forms, institutional arrangements, internal organizational processes, and a vast array of different social positions and interests. Sociologically relevant reflection on this set of issues may be periodized in three broad stages.

Antecedents of the Sociology of Knowledge

While observations about the social foundations and consequences of various cognitive products and processes have a complex history that extends back to the Greeks and forward through such Renaissance figures as Machiavelli and Francis Bacon, the modern roots of the sociology of knowledge are generally dated from the French and Scottish Enlightenment of the eighteenth century, when thinkers like Helvetius, Holbach, and Voltaire turned attention to the social bases and effects of religious – as contrasted with scientific – conceptions, while Montesquieu, Rousseau, and Condorcet adumbrated the broader thesis that human consciousness is socially

determined. Working in this tradition, Destutt de Tracy advocated a 'science of ideas' that would trace human thoughts to their sensory and material conditions, ultimately making possible the construction of a new morality and a perfected political order. For this science, he coined the term 'ideology' (in 1796) – a neologism that his critics swiftly converted into a pejorative word for doctrinaire dreams and illusions (Eagleton, 1991; Kennedy, 1978; Larrain, 1979).

Francis Bacon was stringently ascetic in his determination of what counts as scientific knowledge, but he also put his faith in the civil history that was supposed to guide statesmen; the eighteenth-century Scottish moral philosopher, Adam Ferguson, developed his distinction between the 'history of the individual' that grounded scientific knowledge and the 'history of the species' that featured the stages of history and other synthetic conceptions that made him interesting to later commentators. In John Stuart Mill's *Logic of the Moral Sciences* (1843) there is a succession of chapters on the 'physical' and on the 'historical' sciences. Notwithstanding Mill's ingenious designation of the latter as representing the 'inverse deductive' method, there is little doubt that he did not consider the latter kinds of theory – exemplified by Saint-Simon and Comte – to be 'scientific' in the strict sense that governs his entire treatise, although he clearly finds such theory relevant to social understanding.

In the early nineteenth century, Auguste Comte declared the evolution of modes of human thought the main topic of the new discipline of sociology, erecting the field on the Law of Three Stages, viz., the assertion "that each of our leading conceptions – each branch of our knowledge – passes successively through three different theoretical conditions: the theological or fictitious; the metaphysical, or abstract; and the scientific, or positive" (1975[1830]: p. 71). Substituting French positivism for German idealism, G.W.F. Hegel (1967[1807]) likewise presented the history of society as a sequence of intellectual stages – as successively developing forms of human consciousness, dialectically progressing to a condition of absolute (genuinely scientific) knowledge.

French and Scottish Enlightenment philosophy recognized that all social differences have social causes and are thus potentially subject to human control. Social, economic, and political factors determine the genesis, structure, and content of human consciousness. Hence the philosophers of the Enlightenment anticipated what became a major thrust of the sociology of knowledge. In the main, however, philosophers have tried to show that such a sociology is neither possible nor desirable. Immanuel Kant allowed that perception always involves categories of understanding and concepts, but he viewed the understanding itself as a priori. Empiricists of the most widely varied persuasions have argued that (scientific) knowledge is warranted by direct experience not dependent on social conditions. At best, nontheoretical factors might influence the genesis of ideas, but not the structure and content of thought itself.

From the mid-nineteenth to the early twentieth century, early modern anticipations of the sociology of the knowledge were elaborated in many directions, particularly in connection with the development of 'classical' sociological theory. Rejecting Hegelian idealism for historical materialism, Karl Marx conceptualized the connections between human thought and social conditions in several ways. In his most encompassing formulation, he asserted that dialectically changing "modes of production [formed] the basis of all history" and that the "origins and growth [of] all the different theoretical products and forms of consciousness, religion, philosophy, ethics, etc., etc." – the idealistic superstructure – were explainable 'from that basis' (Marx and Engels, [1846]1970: pp. 57–58). Taking the social relations of production as 'the realm of real history,' Marx's early work cast this superstructure as the derivative 'realm of ideology,' using the latter term in the pejorative sense of 'illusions' and distorted 'inversions' of reality that 'veil the class struggle' and thus promote existing patterns of class domination (Marx and Engels, [1846]1970: pp. 47, 119; Marx, [1852]1963: pp. 46–47). Marx also linked particular ideas to specific 'material interests' and held that "the class which has the means of material production at its disposal has control … over the means of mental production," so that "the ideas of the ruling class are in every epoch the ruling ideas," as these are developed by 'conceptive ideologists' perfecting the illusions that benefit their class (Marx and Engels, [1846]1970: pp. 64, 65, 99). Marx's later work, in contrast, presented ideology less as a superstructural inversion of reality than as a reflection of inversions already embedded in particular modes of production. Characterizing capitalism itself, for example, as an 'enchanted and perverted world,' he held that, whatever their class interests, conceptualizers under capitalism generally 'remain more or less in the grip of the world of illusion,' a world that 'corresponds' with "the forms of illusion in which they move about and find their daily occupations" ([1894]1967: pp. 827–831; for discussion, see Eagleton, 1991; Mepham, 1979). Subsequently, many of Marx's followers overlaid this negative view of the social functions of ideology with a broader "sense of ideology as the mental forms within which men and women fight out their social conflicts" (Eagleton, 1991: p. 90). This is a tendency evident in the work of Bernstein, Lenin, and Lukács, and in Hoare and Smith's conception of ideologies as forces that, under sociohistorical conditions that can "organize human masses, and create the terrain on which men move, acquire consciousness of their positions, struggle, etc" ([1948–51] 1971: p. 377).

Emile Durkheim is also a forerunner of the sociology of knowledge, although he did not fully succeed in offering a comprehensive model of the process of classification he sought to present. Coming from the French tradition, Durkheim approached knowledge from a different angle. Extending his concern with the social bases (and functions) of such 'collective representations' as religion and morals to the 'extralogical origin' of 'logical thought,' Durkheim argued that the 'mental operations' by which members of preliterate societies "classify … things, events, and facts about the world into kinds and species [are] modeled on [their] form of social organization," i.e., that "the first logical categories were social categories; the first classes of things were classes of men, into which these things were integrated" (Durkheim and Mauss, [1903]1963: pp. 3, 4, 8, 82). Generalizing the point, Durkheim subsequently proposed that the form and the content of all the 'categories of human thought' – of the most basic 'conception[s] of time, space, cause, number, and so on,' even in their modern scientific incarnations – "depend on the way in

which the collectivity is organized, upon its morphology, its religious, moral, and economic institutions, and so on" ([1912]1995: pp. 14–16). Controversial to this day, the claim proved widely influential, inspiring work by Halbwachs (1925), Granet (1934), and others. Durkheim and Mauss, by researching the classification processes of preliterate societies, reached the conviction that all important categories of human thought are socially conditioned. Their thesis that the classification of things reproduces the classification of people (e.g., into clans) is often still the starting point for sociological research.

The Sociology of Knowledge as a Specialty Area

In the 1920s the decisive impulses for the emergence of the sociology of knowledge as a theory-and-research area with its own name came from Max Scheler (1874–1928) and Karl Mannheim (1893–1947). Their separate efforts led to the creation of *Wissenssoziologie* as a symptomatic product of a deep-rooted crisis in the social and historical consciousness of Weimar Germany, where historicism, relativism, and skepticism were vigorously debated (Meja and Stehr, 1982, 2014). What was often a naïve belief in progress came under attack by distrust and skepticism. The sociology of knowledge evolved in this situation as the doctrine of the rise and fall of social processes, relations, and depictions proper to the cultural sphere of knowledge (Scheler), and especially the doctrine of the link between knowledge (or thought) and the social context within which it occurs (Mannheim). In contrast to the Marxist critique of ideology, the sociology of knowledge lays less emphasis on the concealing intention of assertions than on the varied structures of consciousness possessed by the different subject types in history and society.

Even though Karl Mannheim is widely seen as the main founder of the sociology of knowledge, it was Max Scheler who coined the term 'Wissenssoziologie,' as an extension of his interests in philosophy. Scheler's efforts generated a research program in the sociology of culture – and especially the sociology of knowledge – that was by no means constrained within the limits of the philosophical context he postulated. Seeking to recognize "the social–historical development of [even] the highest types of human knowledge" but to "escape [from the] philosophical relativism" that had led contemporaries to deny the truth-claims of ideas "conditioned by a historical and sociological standpoint," he challenged Marx's thesis that one "can 'explain' th[e] ideal world from the world of real history" by asserting that "the history of real factors" – understood as a sequence in which kinship, political, and finally economic factors successively hold centrality – "does not at all determine the positive meaning-content of works of mind, though it does hinder, release, delay, or hasten [their] development," opening and closing "the sluices of the spiritual stream" ([1924]1980: pp. 41–42, 53–54; see also Frisby, 1983).

Mannheim's ambitious program was to lay the foundation for a sociological analysis of knowledge (Meja and Stehr, 1999). Like Scheler, he took a broader view of the Marxist concept of superstructure and argued that psychological, biological, intellectual and religious factors, as well as the purely economic, lay behind it. Investigating the social conditions of the various forms of knowledge, he excluded mathematics and portions of the 'exact sciences' from his analysis, seeking to problematize the 'social conditioning' of the forms and contents of 'every product of thought' – 'perceptions and interpretations,' moral concepts of 'duty, transgression, and sin,' political 'opinions, statements, [and] propositions,' and the total 'modes of thought' of different epochs, cultures, and social groups (see Pels, 1996). According to Mannheim, such thought-products stand in 'multiple interconnections' with 'the structure of historical reality,' with 'the social position of given groups' – a phrase primarily denoting not only class position, but also including "generations, status groups, sects, occupational groups, schools, etc." – and with "the historical–social situation of the intellectually active and responsible members of the group" ([1929]1936a: pp. 56–57, 81–82, 86; [1929]1936c: pp. 264–265, 276). In putting forth this view, Mannheim hoped to lay the foundation for a wide-ranging academic field that would subject "the relationship between thought and existence" to systematic 'empirical research' ([1929]1936a: p. 69; [1929]1936c: p. 264; see also 1936b), and he engaged in such research of himself, most notably in his study on the problem of generations ([1928]1952a), on economic ambition ([1930]1952a), competition ([1928]a), structures of thinking ([1922–24]1982a) and similar topics in the sociology of culture, as well as his later work in political sociology, notably on the theme of planning ([1935]1940). Above all, Mannheim is known for his methodological and substantive work, beginning with his seminal essay on German conservative thought ([1925]1986; see Kettler et al., 1984). Mannheim's most famous book is *Ideology and Utopia* ([1929]1936a; Kettler et al., 1984), in which he elevated the concept of 'ideology,' despite its reductionist and polemical associations, to a central place in the interpretation of all modes of knowledge. Mannheim set for the sociology of knowledge a large agenda, borne from his belief that examination of "the relationship between the intellectual positions held and the social position occupied" would necessarily reveal that 'every point of view' – Marxism included – 'is particular to a social situation.' From this it followed, for Mannheim, that "the thought of all parties in all epochs is of an ideological character," where he took 'ideological' to mean 'narrow' and 'one-sided' – properties with important epistemological and political implications. By exposing the one-sidedness of all perspectives, the sociology of knowledge would make possible 'a new type of objectivity,' a dynamic synthesis of different viewpoints that would bring political rivals into dialogue, checking their pernicious tendency to unmask each another's ideas as mere illusions (1936a: pp. 77–85; [1929]1936b: p. 5; see also Elias, 1971; Kettler and Meja, 1995; Kettler et al., 2008).

Ideology and Utopia opens with a discussion of the crisis in politics caused by all parties now making use of the device once reserved to the Marxists. Each unmasks the positions advanced by the others as nothing but interest-bound ideology, so that none can retain any confidence in the objectivity of even their own perceptions and claims, This is the impasse that the sociology of knowledge will help overcome, in part by reassuring each and all parties about

the perspectival character of all political knowledge. The intervention is meant to foster exchanges between the parties that lead to mediation, although these exchanges can only approximate the 'free exchange of ideas' projected by liberalism. In his discussion and critique of the socialist–communist conception of the theory–practice relationship, for example, Mannheim characterizes a process wherein actors representing 'real' social factors mutually take one another's measure and cumulatively create a sequence of 'real' situations that sequentially provide the starting point for the next phase of development. Quite apart from Marxism, Mannheim contends, we know at least since Carl Schmitt that this is a more adequate account of parliamentary political institutions than the liberal one. Schmitt clearly exposed the liberal illusion of parliaments as societies for the discovery of truth through discussion. The Marxist version of this 'real' contestation, he points out, envisions a moment of revolutionary overturn, when all contradictions are dialectically overcome. As his analysis proceeds, however, leaving the exposition of the different ideological standpoints behind, it becomes clear that his culminating conception of a politics that has been brought beyond critical impasse through the intervention of the sociology of knowledge amounts to a regularization of that 'real' process, without the dramatic disruptions and presumed transcendent reconciliations achieved, according to Marxists, by the anticipated revolutionary overturn.

The sociology of knowledge dispute, even while lasting only a few years, until Hitler's rise to power, focused upon the most important methodological and epistemological confronting German sociology in the Weimar period. While the more-than-30 major papers published in response to Mannheim's book (Meja and Stehr, 1982, [1990]2014) cover a considerable range of theoretical positions, the majority of responses and critiques come from Marxist or socialist scholars of various orientations, among them the orthodox Marxists Karl Wittfogel, Otto Neurath, and Albert Fogarasi; the anti-positivist Marxists Max Horkheimer, Herbert Marcuse, and Ernst Lewalter; the religious socialist Paul Tillich; and the Social Democrats Hannah Arendt, Guenther Stern (later Guenther Anders), and Hans Speier. The interest in the relationship between Marxism and the sociology of knowledge constitutes the major point of unity in the response to Mannheim's book. The positions range from seeing in the sociology of knowledge a sophisticated reformulation of the materialist conception of history to attributing its popularity to a neutering and betrayal of Marxism.

Mannheim thought that the sociology of knowledge could play a key role in intellectual and political life by showing that social conditions help to give rise to political philosophies, ideologies, cultural achievements, and competing systems of thought, thereby introducing reason into politics. His agenda drew broad attention to the sociology of knowledge in Europe and in America and set the new field on an unsteady course. In Europe, his epistemological and political views became the focus of extensive critical discussion and made the field, for decades, largely a site of *meta*-theoretical controversy (Frisby, 1983; Wolff, 1993; Meja and Stehr, 1988). In America, where engagement with sociology-of-knowledge type issues had long been underway (Fuhrman, 1980), Mannheim's call for empirical studies had greater resonance and fueled efforts to constitute the sociology of knowledge as a specialized subdiscipline dedicated to 'fruitful empirical research,' rather than to debate over 'philosophical' questions of relativism, objectivity, and the like (quoting Wirth, 1936: pp. xxi, xxvii; see also Kettler and Meja, 1995: pp. 193–246; see also Kettler et al., 2008).

Particularly significant here are Robert K. Merton's writings in the early 1940s, which sought to move the sociology of knowledge from 'speculative insight' to 'rigorous test,' by purging Mannheim's ideas of "their epistemological impedimenta" and associating the field with a defined "list of problems [open to] empirical investigation" ([1945]1968a: p. 562; [1941]1968b: p. 542). To this end, Merton elaborated a "paradigm for the sociology of knowledge," in which he replaced Mannheim's "amorphous category of knowledge" with a systematic inventory of different types and aspects of 'mental productions' and then differentiated the many possible sociocultural or 'existential bases' for human thought-products ([1941]1968a: p. 552; 1968b: pp. 514–515). Most distinctively, Merton gave prominence to the question (which he faulted his predecessors for neglecting) "*how* are mental production related to the existential basis?" – distinguishing, in reply, a wide range of possible modes of relationship, analysis of which "promise [d] to take research in the sociology of knowledge from the plane of general imputation to that of testable empirical inquiry" ([1945]1968b: pp. 515, 537).

This promise proved difficult to realize, however. In the course of the next 20 years, students of Merton occasionally took up the challenge – Gouldner (1965) with research on Plato's social theory as a response to the social tensions present in ancient Greece, Coser (1965) with work on the general societal and institutional conditions that favored the development of Western intellectuals. These studies joined writings by Mills ([1943]1964) on the social roots of American pragmatism (see also Mills, 1963), by Znaniecki (1940) and Shils ([1955–68] 1972) on intellectuals, by members of the Frankfurt school on the ideology of capitalist societies (see especially Marcuse, 1964), and by a scattering of other scholars on a miscellany of topics (see the collection by Curtis and Petras, 1970) to steer the sociology of knowledge in more empirical directions. But these remained sporadic efforts, intermixed with works offering programmatic definitions of the sociology of knowledge and further discussion of Mannheim's more philosophical concerns (see, e.g., DeGré, [1939–43]1985; Stark, 1958.

Kurt H. Wolff (Mannheim's student in Frankfurt, his translator and editor in America, and – though not a disciple – loyal to Mannheim and the sociology of knowledge project even in exile) is an interesting and central figure here (Wolff, 1959, 1968, 1974, 1976, 1983, 1993; see also Backhaus and Psathas, 2007). Mannheim's belief was that the crisis of universal distrust can only be countered by radical doubt, and that this doubt will paradoxically eventuate in a new ontology and in a new foundational science of being. From this point of view, sociology of knowledge is not a positive discipline with consequences for methodology: it is philosophy carried on by more appropriate means. This is the motif to which Kurt H. Wolff's mediation has sought to give expression, life, and relevance. Hardly accessible to

'insiders' of 'American sociology,' this motif may nevertheless be audible in a sociological profession that has been adopted as home by many students deeply troubled by a new sense of social crisis and by a corresponding disbelief in the methodological asceticism of empirical inquiry. Mannheim's questions are seen by Wolff to revolve around Nietzsche's ghastly discoveries, now confirmed by the manifest discrediting of all claims to speak with the authority of reason in the spheres of utmost importance to human action. The exposure of such claims as 'rationalization' and 'ideology' itself proceeds by rational means, of course, at first embodied in shrewd political assaults and eventually in the two genealogical inquiries of psychoanalysis and sociology of knowledge.

For Wolff Mannheim's sociology of knowledge epitomizes the courage to reshape and to pursue this destructive course without abandoning the goal of finding a new affirmation. To avoid misunderstanding, it should be emphasized that Wolff was not content to spin out metaphysical suppositions to resolve these gravest of all doubts. He was strict in his understanding of the work that sociologists of knowledge must do, and generous in his appreciation of quite detailed examples of the empirical craft. The point is not to encourage speculation, but to foster reflexive attention to the meanings of the work that sociologists of knowledge professionally do, locating it in the context of purposes that need to be subserved by knowledge and of problems that must be handled wisely. If knowledge is what our sociological analyses show it to be, what is the point of questions that Wolff regarded as central to Mannheim's sociological project and that Wolff asks us to receive as integral to the tradition of inquiry initiated by Mannheim?

Such work as Wolff's distanced empirically oriented sociologists from the field, even as various sociology-of-knowledge type questions resurfaced in the sociology of religion, the sociology of education, the sociology of occupations and professions, the sociology of generations, the sociology of public opinion and mass communication, and the sociology of science (Coser, 1968: p. 432), as well as in research by historians of science and historians of ideas (see especially Ringer, 1969).

The era that began with the writings of Scheler and Mannheim and extended to the mid-1960s remains, nonetheless, the most active and perhaps fruitful period in the sociology of knowledge as a specialty area, and it stamped the field with a set of basic assumptions that were common to most work in the field at the time. These included, first, the assumption of a division between sociocultural factors, on the one side, and the substantive contents of thought-products and processes, on the other, where the former factors were 'external' to the realm of knowledge, while the latter were 'internal'; second, the supposition that the 'internal' factors were easily interpretable by the sociological investigator, their meanings more-or-less transparent in the statements where they were contained; and, third, the belief that sociological explanations of knowledge chiefly involved appeals to 'external' factors at the macrolevel, i.e., to societal-wide economic, political, and cultural developments, or combinations thereof (Camic and Gross, 2000; see also Kuklick, 1983).

Recent Developments

For a variety of reasons, largely exogenous to the area itself, interest in the sociology of knowledge diminished for two decades beginning in the early 1960s. This was due partly to a continuation of the tendency, already evident earlier, for partial incorporation of sociology of knowledge topics into other sociological subfields. Expanding interest in the sociology of science proved especially consequential here, for in the process of focusing on scientists' career patterns and on the organizational and normative conditions that sustain scientific institutions, sociologists of science effectively denied – despite contrary suggestions in the influential work of Kuhn (1962) – that there was even much possibility for a "sociology of the conceptual and theoretical contents of science" and related domains of knowledge (Ben-David, 1971: p. 14). In the face of this challenge, Berger and Luckmann attempted to redirect the sociology of knowledge to the study of "what is taken for granted as 'knowledge' [by] the man in the street" (1966: p. 3), an emphasis neglected by the classical sociology of knowledge. With reference to the phenomenology of everyday life of Alfred Schutz and the philosophical anthropology of Arnold Gehlen, Berger and Luckmann deliberately set aside the classical approaches of Marx, Mannheim, and to some extent also Scheler, with their epistemological and methodological concerns and their focus on absolutely valid knowledge, instead taking as their primary theme everything that counts as knowledge in society, including ideology, 'false consciousness,' propaganda, art and science (Gross, 2010: esp. 216–217; see also 2013). Their work may be regarded as an application of Durkheim's dictum that social facts must be considered as 'things,' that everything in society is a social product, a social construction. Social Facts are *sui generis*, the effect or the creation of human activities, but they are most often the unanticipated consequences of human agency or behavior.

Compounding these developments within specialty areas, including ethnomethodology, another subfield then on the rise (Maynard and Clayman, 1991), were broader trends in the social sciences during this period. Most far-reaching among these was a growing commitment to quantitative research, based increasingly on types of survey data that seemed remote from the traditional concerns of sociologists of knowledge. This trend, to be sure, coexisted with a heightening of interest in theory and in historical scholarship, but the Marxist renaissance that was then underway generally pulled this interest toward questions about the dynamics of modes of production and state apparatuses and toward debates about the role and concept of social class (and, subsequently, about gender and race). While new theories of ideology did emerge (see, in particular Althusser, 1969, 1971; Therborn, 1980), they did so, almost without exception, as appendages to work on class and state structures; as such, their effect was more to divert attention from, than to turn it to, examination of the sources and consequences of cognitive products and processes in their own right. In this context, little new work in the sociology of knowledge came forth, aside from reinterpretations of significant figures from earlier periods, especially Mannheim (see Simonds, 1978; Loader, 1985) and various editions of their original writings.

Just as exogenous developments during the two preceding decades threatened to imperil the sociology of knowledge, intellectual trends since the 1980s have resuscitated the field. Most generally these have included: burgeoning interest in the interdisciplinary field of cultural studies, also encouraged by the posthumous publication of Mannheim's early writings (1982a [1922–24]) in the sociology of culture (see Barboza, 2005, 2009; Corsten, 2010) and stimulated by a declining focus on Marxist structuralism and the growing influence of works on discourse and linguistic theory, poststructuralism, and postmodernism (for discussion, see McCarthy, 1996); the effort within Marxism itself to transform general claims about ideology into empirical studies of discursive formations and practices (Laclau and Mouffe, 1985; Thompson, 1990); the move within feminist scholarship to problematize knowledge and the standpoint and social position of knowledge producers (Harding, 1991; Smith, 1987; see also Kettler and Meja, 1993); the rediscovery of Mannheim's methodological reflections on the interpretation of word-views, which Mannheim called the 'documentary' method of interpretation (1952b[1921–22]: pp. 51–63; 1982[1924]: pp. 141–288), as one of the essential methods for interpretive, or reconstructive, qualitative sociological research (Bohnsack, 2007, 2008; Weller, 2005); the increased production of sociologically informed studies in the history of science and intellectual history (see Shapin, 1982; Toews, 1987); and the redirection of the sociology of science toward the examination of the contents of scientific knowledge and the social processes shaping them (see Collins, 1983). Reinforcing these broad trends has been an expansion of interest in European social theorists whose works give considerable attention to knowledge-production processes, particularly Michel Foucault and his archeological and genealogical studies of the reciprocal constitution of various knowledge systems and webs of power relations (see Prado, 1995), Pierre Bourdieu and his writings on the dynamics of intellectual and scientific fields (see Ringer, 1990; Camic, 2013), as well as Norbert Elias' figurational sociology and Niklas Luhmann's systems theory, both influenced by Mannheim's sociology of knowledge (Elias, 1971, 1981, 2009; Luhmann, 1980, 2008).

From this confluence of exogenous developments has come a renewed concern with the sociology of knowledge, a concern that is currently pulling the field in the direction of several different approaches. The first of these is the broad-constructionist approach, exemplified in programmatic statements by McCarthy (1996), Meja and Stehr (1988), Stehr and Meja (2005), and Swidler and Arditi (1994), and in studies by Griswold (1986), Lamont (1987), and Wuthnow (1989). Pointing the sociology of knowledge outward toward the contemporary sociology of culture rather than toward the study of knowledge specialists, this perspective treats "political and religious ideologies as well as science and everyday life, cultural and organizational discourses along with formal and informal types of knowledge" (Swidler and Arditi, 1994: p. 306); and while amenable to a middle-level or institutional analysis of the social sources and consequences of these various forms of knowledge, it often retains the underlying suppositions of the 1920–65 period, viz. assumptions about the division between external (sociocultural) and internal (thought-content) factors, about the transparency of internal meanings to the sociological investigator, and about the explanatory importance of macrolevel external factors. Moreover, while strongly committed to empirical research, this approach has also favored reopening some of the *meta*-theoretical issues that occupied earlier generations (see the contributions in Stehr and Meja, 2005).

In these ways, the broad-constructionist approach diverges from the recent narrow-constructionist alternative. Found in the work of Camic (1983, 1992; Camic and Xie, 1994), Collins (1998), Kusch (1995), and Shapin (1994), this second approach focuses the sociology of knowledge mainly on men and women who specialize in the production of ideas and on the particular social processes by which their ideas emerge and develop – a move that, in effect, transforms the field into a sociology of ideas. This perspective has tended to reject the core assumptions of the older sociology of knowledge, building instead on scholarship that argues that sociocultural processes are as much internal to the content of ideas as they are external (Bloor, 1976; Shapin, 1992), that the meanings of ideas are only understandable to an investigator after careful contextual reconstruction (Skinner, 1969), and that local, micro-level settings are often the main sites for the development of ideas (Geertz, 1983; Whitley, 1984). Like the broad-constructionist approach, this narrow-constructionist perspective presently provides a foundation for several lines of empirical research (see Camic and Gross, 2000).

Inspired by developments in the history of knowledge, sociologists of knowledge have also engaged in empirical analyses of the 'manufacture' of natural-scientific knowledge. The most important research tools have been ethnographies of laboratory life (Latour and Woolgar, 1979; Knorr-Cetina, 1981, 1999). These studies have led to a reassessment of the limitations of various types of knowledge, of the conditions of their origin, and of their structural qualities. Traditional theoretical assumptions about the specific rationality of scientific knowledge have been reassessed (Woolgar, 1988; Serres and Latour, 1995; Latour, 1999).

Lewis Coser observed that the ambitious project of the founders of the sociology of knowledge remains a continuing challenge, not an obsolete burden. In his survey of the history and prospects of the sociology of knowledge he writes:

> The sociology of knowledge was marked in its early history by a tendency to set up grandiose hypothetical schemes. These contributed a number of extremely suggestive leads. Recently its practitioners have tended to withdraw from such ambitious undertakings and to restrict themselves to somewhat more manageable investigations. Although this tendency has been an antidote to earlier types of premature generalizations, it also carries with it the risk of trivializations. Perhaps the sociology of knowledge of the future will return to the more daring concerns of its founders, thus building upon the accumulation of careful and detailed investigations by preceding generations of researchers.
>
> Coser, 1968: p. 432

Such a return of sociology knowledge practitioners to the more daring concerns of the founders cannot not exclusively result from intellectual and scientific efforts. Merton, for example, therefore emphasizes that interest in sociology of knowledge issues is itself socially conditioned. One important factor, as Merton argued, is that

> With increasing social conflict, differences in the values, attitudes, and modes of thought of groups develop to the point where the orientation which these groups previously had in common is overshadowed by incompatible differences. Not only do there develop distinct universes of discourse, but the existence of any one universe challenges the validity and legitimacy of the others.
>
> Merton, 1968b: p. 511

In this sense the reconstructions and interpretation of the sociology of knowledge by the generation of sociologists succeeding the generation of the founders, as well as the period of temporary waning of interest in sociology of knowledge issues discussed above, coincided with particular economic, political and social conditions as well as academic and disciplinary circumstances that, in their relative tranquility, stood in market contrast to the crisis conditions of the 1920s. The present renewal of interest in some of the major issues first articulated by the founders of the sociology of knowledge, in contrast, 'reflects' (just as in the earlier period in which the classical sociology of knowledge arose) the experience of crisis in present-day society (Beck, [1986]1992; Stehr, 2001).

Knowledge has always played a significant role in human life. Human action has to a greater or lesser extent always been steered by knowledge. Power, for example, has rarely been based exclusively on brute physical force, but almost always also involves a knowledge advantage. At present, however, knowledge is assuming greater significance than ever before. In spite of a simultaneous rise in interest in not-knowing and in knowledge gaps (Beck and Wehling, 2012; Gross, 2010), advanced societies may be regarded as 'knowledge societies' (Stehr, 1994). A thoroughgoing scientization of all spheres of human life and action, the transformation both of traditional structures of domination and of the economy, as well as the growing impact and influence of experts are all indications of the rapidly increasing role of knowledge in the organization of contemporary modern societies.

See also: Constructivism/Constructionism: Methodology; Comte, Auguste (1798–1857); Culture, Sociology of; Durkheim, Émile (1858–1917); Foucault, Michel (1926–84); History of Science; History of Science: Constructivist Perspectives; Ideology (Sociology of); Intellectuals: History of the Concept; Intellectuals, Sociology of; Knowledge Representation; Mannheim, Karl (1893–1947); Marx, Karl (1818–83); Positivism: Sociological; Science and Religion; Science, Sociology of; Sociological Theory; Sociology: Overview.

Bibliography

Althusser, L., 1969. For Marx. Penguin Press, London.
Althusser, L., 1971. Lenin and Philosophy and Other Essays. New Left Books, London.
Backhaus, G., Psathas, G., 2007. The Sociology of Radical Commitment: Kurt H. Wolff's Existential Turn. Lexington Books, Lanham, MD.
Barboza, A., 2005. Kunst und Wissen. Die Stilanalyse in der Soziologie Karl Mannheims.
Barboza, A., 2009. Karl Mannheim. Klassiker der Wissenssoziologie. UVK Verlagsgesellschaft, Konstanz.
Beck, U., 1992[1986]. Risk Society. Sage, London.
Beck, U., Wehling, P., 2012. The problem of not-knowing: an emerging area of social and political conflict in reflexive modernity. In: Rubio, F.D., Baert, P. (Eds.), The Politics of Knowledge. Routledge, New York.
Ben-David, J., 1971. The Scientist's Role in Society: A Comparative Study. Prentice-Hall, Englewood Cliffs, NJ.
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, Garden City, NY.
Bloor, D., 1976. Knowledge and Social Imagery. Routledge & Kegan Paul, London.
Bohnsack, R., 2007. Rekonstruktive Sozialforschung: Einführung in Qualitative Methoden, seventh ed. Budrich, Opladen.
Bohnsack, R., 2008. The interpretation of pictures and the documentary method. Forum: Qualitative Social Research 9 (3).
Camic, C., 1983. Experience and Enlightenment: Socialization for Cultural Change in Eighteenth Century-England. University of Chicago Press, Chicago.
Camic, C., 1992. Reputation and predecessor selection. American Sociological Review 57, 421–445.
Camic, C., 2013. Bourdieu's two sociologies of knowledge. In: Gorski, P. (Ed.), Bourdieu and Historical Analysis. Duke University Press, Durham, NC.
Camic, C., Gross, N., 2000. The new sociology of ideas. In: Blau, J. (Ed.), The Blackwell Companion to Sociology. Blackwell, Oxford, UK.
Camic, C., Xie, Y., 1994. The statistical turn in American social science – Columbia University, 1890 to 1915. American Sociological Review 59, 773–805.
Collins, H.M., 1983. The sociology of scientific knowledge – studies of contemporary science. Annual Review of Sociology 9, 265–285.
Collins, R., 1998. The Sociology of Philosophies: A Global Theory of Intellectual Change. Belknap Press of Harvard University Press, Cambridge, MA.
Comte, A., 1975[1830]. Auguste Comte and Positivism: The Essential Writings. Harper & Row, New York.
Corsten, M., 2010. Karl Mannheims Kultursoziologie. Campus, Frankfurt and New York.
Coser, L.A., 1965. Men of Ideas: A Sociologist's View. Free Press, New York.
Coser, L.A., 1968. Sociology of Knowledge. In: International Encyclopedia of the Social Sciences, vol. 8. Macmillan, New York, 428–435.
Curtis, J.E., Petras, J.W. (Eds.), 1970. The Sociology of Knowledge: A Reader. Duckworth, London.
DeGre, G., 1985[1939–43]. The Social Compulsions of Ideas: Toward a Sociological Analysis of Knowledge. Transaction Books, New Brunswick, NJ.
Durkheim, E., 1995[1912]. The Elementary Forms of Religious Life. Free Press, New York.
Durkheim, E., Mauss, M., 1963[1903]. Primitive Classification. University of Chicago Press, Chicago.
Eagleton, T., 1991. Ideology: An Introduction. Verso, London.
Elias, N., 1971. Sociology of knowledge: new perspectives. Sociology 5, 149–168, 226–252.
Elias, N., 2005[1981]. Knowledge and power. Interview with Peter Ludes. In: Stehr, N., Meja, V. (Eds.), Society and Knowledge, second rev. ed. Transaction Publishers, New Brunswick, NJ.
Elias, N., 2009. Essays on the Sociology of Knowledge and the Sciences. UCD Press, Dublin (Works, vol. 14).
Frisby, D., 1983. The Alienated Mind: The Sociology of Knowledge in Germany, 1918–1933. Heinemann Educational Books, London.
Fuhrman, E.R., 1980. The Sociology of Knowledge in America, 1883–1915. University of Virginia Press, Charlottesville, VA.
Geertz, C., 1983. Local Knowledge: Further Essays in Interpretive Anthropology. Basic Books, New York.
Gouldner, A.W., 1965. Enter Plato. Basic Books, New York.
Gouldner, A.W., 1970. The Coming Crisis of Western Sociology. Basic Books, New York.
Granet, M., 1934. La Pensee Chinoise. La Renaissance du Livre, Paris.
Griswold, W., 1986. Renaissance Revivals City, Comedy and Revenge Tragedy in the London Theatre, 1576–1980. University of Chicago Press, Chicago.
Gross, M., 2010. Ignorance and Surprise: Science, Society, and Ecological Design. MIT Press, Cambridge, MA.
Gross, N., 2010. Pragmatism, phenomenology, and twentieth century sociology. In: Calhoun, C. (Ed.), Sociology in America. University of Chicago Press, Chicago and London.
Halbwachs, M., 1925. Les Cadres Sociaux de la Mémorie. F. Alcan, Paris.
Harding, S.G., 1991. Whose Science? Whose Knowledge? Thinking from Women's Lives. Cornell University Press, Ithaca, NY.
Hegel, G.W.F., 1967[1807]. The Phenomenology of Mind. Harper & Row, New York.
Hoare, Q., Smith, G.N. (Eds.), 1971. Selections from the Prison Notebooks of Antonio Gramsci. Lawrence & Wishart, London.
Kennedy, E., 1978. A Philosopher in the Age of Revolution: Destutt de Tracy and the Origins of Ideology. University of Pennsylvania Press, Philadelphia.
Kettler, D., Meja, V., 1993. Their 'own peculiar way': Karl Mannheim and the rise of women. International Sociology 8 (1).

Kettler, D., Meja, V., 1994. 'That typically German kind of sociology which verges on philosophy: the dispute about 'ideology and utopia' in the United States. Sociological Theory 23 (3).
Kettler, D., Meja, V., 1995. Karl Mannheim and the Crisis of Liberalism: The Secret of These New Times. Transaction Publishers, New Brunswick, NJ.
Kettler, D., Meja, V., Stehr, N., 1984. Karl Mannheim and conservatism: the ancestry of historical thinking. American Sociological Review 49 (1).
Kettler, D., Meja, V., Stehr, N., 1990. Rationalizing the irrational: Karl Mannheim and the besetting sin of German intellectuals. American Journal of Sociology 95 (6).
Kettler, D., Loader, C., Meja, V., 2008. Karl Mannheim and the Legacy of Max Weber. Retrieving a Research Programme. Ashgate, Aldershot, England.
Knorr-Cetina, K., 1981. The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science. Pergamon Press, New York.
Knorr-Cetina, K., 1999. Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press, Cambridge, MA.
Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Kuklick, H., 1983. The sociology of knowledge: retrospect and prospect. Annual Review of Sociology 9, 287–310.
Kusch, M., 1995. Psychologism: A Case Study in the Sociology of Philosophical Knowledge. Routledge, London.
Laclau, E., Mouffe, C., 1985. Hegemony and Socialist Strategy: Towards a Radical Democratic Politics. Verso, London.
Lamont, M., 1987. How to become a dominant French philosopher – the case of Jacques Derrida. American Journal of Sociology 93, 584–622.
Larrain, J., 1979. The Concept of Ideology. University of Georgia Press, Athens, GA.
Latour, B., 1999. Pandora's Hope: Essays on the Reality of Science Studies. Harvard University Press, Cambridge, MA.
Latour, B., Woolgar, S., 1979. Laboratory Life: The Social Construction of Scientific Facts.
Loader, C., 1985. The Intellectual Development of Karl Mannheim: Culture, Politics, and Planning. Cambridge University Press, Cambridge.
Luhmann, N., 1980. Gesellschaftsstruktur und Semantik: Studien zur Wissenssoziologie der modernen Gesellschaft. Suhrkamp, Frankfurt.
Luhmann, N., 2008. Ideenevolution: Beiträge zur Wissenssoziologie. Suhrkamp, Frankfurt.
Mannheim, K., 1936a[1929]. Ideology and utopia. In: Mannheim, K. (Ed.), Ideology and Utopia: An Introduction to the Sociology of Knowledge. Routledge & Kegan Paul, London.
Mannheim, K., 1936b. Preliminary approach to the problem. In: Mannheim, K. (Ed.), Ideology and Utopia: An Introduction to the Sociology of Knowledge. Routledge & Kegan Paul, London.
Mannheim, K., 1936c[1929]. The sociology of knowledge. In: Mannheim, K. (Ed.), Ideology and Utopia: An Introduction to the Sociology of Knowledge. Routledge & Kegan Paul, London.
Mannheim, K., 1940[1935]. Man and Society in an Age of Reconstruction. Routledge & Kegan Paul, London.
Mannheim, K., 1952a[1921–30]. Essays on the Sociology of Knowledge. Routledge & Kegan Paul, London.
Mannheim, K., 1952b[1921–22]. On the interpretation of 'Weltanschauung'. In: Mannheim, K. (Ed.), Essays on the Sociology of Knowledge. Routledge & Kegan Paul, London.
Mannheim, K., 1956. Essays on the Sociology of Culture. Routledge & Kegan Paul, London (early 1930s).
Mannheim, K., 1982a[1922–24]. Structures of Thinking. Routledge & Kegan Paul, London.
Mannheim, K., 1982b[1924]. A sociological theory of culture and its knowability (conjuctive and communicative thinking). In: Mannheim, K. (Ed.), Structures of Thinking. Routledge & Kegan Paul, London.
Mannheim, K., 1986[1925]. Conservatism: A Contribution to the Sociology of Knowledge. Routledge & Kegan Paul, London.
Marcuse, H., 1964. One-dimensional Man: Studies in the Ideology of Advanced Industrial Society. Beacon, Boston.
Marx, K., 1963[1852]. The 18th Brumaire of Louis Bonaparte. International Publishers, New York.
Marx, K., 1967[1894]. Capital, vol. 3. International Publishers, New York.
Marx, K., Engels, F., 1970[1846]. The German Ideology. International Publishers, New York.
Maynard, D.W., Clayman, S.E., 1991. The diversity of ethnomethodology. Annual Review of Sociology 17, 385–418.
McCarthy, E.D., 1996. Knowledge as Culture: The New Sociology of Knowledge. Routledge, London.
Meja, V., Stehr, N. (Eds.), 1982. Der Streit um die Wissenssoziologie, 2 vols. Suhrkamp, Frankfurt.
Meja, V., Stehr, N., 1988. Social science, epistemology, and the problem of relativism. Social Epstemology 2 (3), 263–271.
Meja, V., Stehr, N. (Eds.), 1999. The Sociology of Knowledge, 2 vols. Cheltenham, UK and Northampton, MA (An Elgar Reference Collection).
Knowledge and politics. In: Meja, V., Stehr, N. (Eds.), The Sociology of Knowledge Dispute. Routledge, London.
Mepham, J., 1979. The theory of ideology in capital. In: Mepham, J., Ruben, D.-H. (Eds.), Issues in Marxist Philosophy, Epistemology, Science, Ideology, vol. 3. Harvester, London.
Merton, R.K., 1968a[1941]. Karl Mannheim and the sociology of knowledge. In: Merton, R.K. (Ed.), Social Theory and Social Structure. Free Press, New York.
Merton, R.K., 1968b[1945]. The sociology of knowledge. In: Merton, R.K. (Ed.), Social Theory and Social Structure. Free Press, New York.
Mills, C.W., 1963[1943]. Power, Politics, and People: The Collected Essays of C. Wright Mills. Oxford University Press, New York.
Mills, C.W., 1964. Sociology and Pragmatism. Oxford University Press, New York.
Pels, D., 1996. Karl Mannheim and the sociology of scientific knowledge: toward a new agenda. Sociological Theory 14 (1), 30–48.
Pfadenhauer, M., 2013. The New Sociology of Knowledge. The Life and Work of Peter L. Berger.
Prado, C.G., 1995. Starting with Foucault: An Introduction to Genealogy. Westview Press, Boulder, CO.
Ringer, F.K., 1969. The Decline of the German Mandarins: The German Academic Community, 1890–1933. Harvard University Press, Cambridge, MA.
Ringer, F.K., 1990. The intellectual field, intellectual history, and the sociology of knowledge. Theory and Society 19, 269–294.
Scheler, M., 1980[1924]. Problems of a Sociology of Knowledge. Routledge & Kegan Paul, London.
Serres, M., Latour, B., 1995. Conversations on Science, Culture, and Time. University of Michigan Press, Ann Arbor.
Shapin, S., 1982. History of science and its sociological reconstructions. History of Science 20, 157–211.
Shapin, S., 1992. Discipline and bounding: the history and sociology of science as seen through the externalism–internalism debate. History of Science 30, 333–369.
Shapin, S., 1994. A Social History of Truth: Civility and Science in Seventeenth-Century England. University of Chicago Press, Chicago.
Shils, E., 1972[1955–68]. The Intellectuals and the Powers and Other Essays. University of Chicago Press, Chicago.
Sica, A., 2010. In: Calhoun, C., Merton, Robert K. (Eds.), Merton, Mannheim and the Sociology of Knowledge. Columbia University Press, New York.
Simonds, A.P., 1978. Karl Mannheim's Sociology of Knowledge. Clarendon, Oxford.
Skinner, Q., 1969. Meaning and understanding in the history of ideas. History and Theory 8, 3–53.
Smith, D., 1987. The Everyday World as Problematic: A Feminist Sociology. University of Toronto Press, Toronto.
Stark, W., 1958. The Sociology of Knowledge: An Essay in Aid of a Deeper Understanding of the History of Ideas. Free Press, Glencoe, IL.
Stehr, N., 1994. Knowledge Societies. Sage, London.
Stehr, N., 2001. The Fragility of Modern Societies: Knowledge and Risk in the Information Age.
Stehr, N., Meja, V. (Eds.), 2005. Society and Knowledge: Contemporary Perspectives in the Sociology of Knowledge & Science, second rev. ed. Transaction Publishers, New Brunswick, NJ.
Swidler, A., Arditi, J., 1994. The new sociology of knowledge. Annual Review of Sociology 20, 305–329.
Therborn, G., 1980. The Ideology of Power and the Power of Ideology. NLB, London.
Thompson, J.B., 1990. Ideology and Modern Culture: Critical Social Theory in the Era of Mass Communication. Polity Press, Cambridge, UK.
Toews, J.E., 1987. Intellectual history after the linguistic turn. American Historical Review 92, 879–907.
Weller, V., 2005. Karl Mannheim und die dokumentarische Methode. In: Zeitschrift für qualitative Bildungs-Beratungs-und Sozialforschung, 6, p. 2.
Whitley, R., 1984. The Intellectual and Social Organization of the Sciences. Oxford University Press, Oxford, UK.
Wirth, L., 1936. Preface. In: Mannheim, K., Ideology, Utopia (Eds.), An Introduction to the Sociology of Knowledge. Harcourt Brace Jovanovich, New York.
Wolff, K.H., 1959. The sociology of knowledge and sociological theory. In: Gross, L. (Ed.), Symposium on Sociological Theory. Row, Peterson, Evanston, IL, pp. 587–590.
Wolff, K.H., 1968. Versuch zu einer Wissenssoziologie. Luchterhand, Berlin.
Wolff, K.H., 1974. Trying Sociology. John Wiley & Sons, New York.
Wolff, K.H., 1976. Surrender and Catch: Inquiry and Experience Today. Reidel, Dordrecht.
Wolff, K.H., 1983[1942–65]. Beyond the Sociology of Knowledge. University Press of America, Lanham, NY.
Wolff, K.H., 1993. From Karl Mannheim, second expanded ed. Transaction Publishers, New Brunswick, NJ.
Woolgar, S. (Ed.), 1988. Knowledge and Reflexivity: New Frontiers in the Sociology of Knowledge. Sage, London.
Wuthnow, R., 1989. Communities of Discourse. Harvard University Press, Cambridge, MA.
Znaniecki, F., 1940. The Social Role of the Man of Knowledge. Columbia University Press, New York.

Kohlberg, Lawrence (1927–87)

Gertrud Nunner-Winkler, Max Planck Institute for Psychological Research, Munich, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 8151–8155, © 2001, Elsevier Ltd.

Abstract

First, a brief overview of Kohlberg's biography and some reminiscences by friends and colleagues highlighting some core characteristics of his personality are presented. His basic assumptions about the nature of human beings, of morality, and of psychology as a social science are briefly touched upon. Then, central features of this theory of moral development are explicated: the concept of a developmental logic, his description of the moral stages, the methods used, the mechanisms assumed to advance development, and implications for pedagogical intervention. Empirical support is discussed concerning the assumptions of the sequentiality, hierarchy and universality of moral development. Finally, some critical issues are taken up in the light of recent empirical findings: the judgment action gap and Kohlberg's descriptions of children's moral understanding. A new reading of moral development is suggested in which cognitive and motivational aspects are disentangled.

Kohlberg was born on 25 October 1927, in Yonkers, New York, the son of a Protestant mother – a well-known chemist – and a Jewish father – a successful self-made businessman with strong civic commitment and of conservative leaning. Kohlberg grew up in a liberal Jewish community. In 1945 he joined the US Merchant Marines; he then helped to smuggle East European Jewish refugees to Palestine, was seized by the British army, and was interned in Cyprus. He fled to Palestine and spent some time in a kibbutz – pondering issues of justice and violence. In 1947 he went to the University of Chicago, reading moral philosophy from Plato to Dewey, studying Locke, Mill, and Jefferson in search of a basic moral principle that might bridge disagreements – and discovered Kant's imperative: 'treat every human being as an end in himself.' He did graduate work in psychology, studying with Bettelheim, Rogers, Gewirtz, Morris, Newgarten, Havighurst, and Strauss – 'the first American scientist who attempted to integrate Piaget's structural theory with Mead's symbolic interactionism' (Kohlberg 1984, p.vii).

After two years of clinical work he started research for his doctoral dissertation on the development of moral judgment that he finished in 1958. From 1959 to 1961 he was assistant professor at Yale University; 1961–62 fellow at the Institute for Advanced Study in the Behavioral Sciences in Palo Alto; 1962–68 professor at the University of Chicago; 1968–87 professor at Harvard University. Having an incurable disease, on 17 January 1987 Kohlberg ended his life.

Kohlberg changed US psychology. His influence is due to intellectual brilliance, but also to an inspiring personality attracting scientists from all over the world. Kuhmerker et al. (1991) elicited personal letters from friends and colleagues describing what Kohlberg had meant for their lives. Some core characteristics emerge: Kohlberg was 'loveable' – and had 'a remarkable ability to insure this characteristic in those who worked with him ... all of whom had not only a peculiarly fanatical interest in the study of moral development but also a deep personal affection for Lawrence Kohlberg, the man.' (Grimley) Kohlberg was persistent; what struck Kegan was 'that an idea can be compelling enough and an individual devoted enough for the two to spend 25 years together in a relationship as fresh and combustible today as when it began.' Kohlberg was stern – demanding extraordinary efforts of himself and of others; he 'would call another scoring workshop whenever he felt his assistants were working less than 80 hours a week.' His commitment was contagious: 'At the Center of Moral Education there was a sense of being at the right place at the right time – that terribly new and important breakthroughs were just around the corner ... that questions of social justice were on everyone's mind and we thought we had the answers – or nearly had the answers.' He was 'a visionary' who kept 'the major questions in focus.' Had it not been for him 'many of us would not have become interested in psychology at all' (Rest, 1986).

For Kohlberg the unity of moral judgment and action was not just a theoretical stipulation – it was his very personality. A colleague recalls an occasion when he had initiated a resolution against the invasion to Cambodia. When the question arose who would accompany him in handing it over in Washington:

> ... the meeting became silent. Motionless. Eerily quiet. These giants of education who daily proclaimed the virtues of equality, of human rights, of democratic citizenship and had themselves just voted overwhelmingly in favor of the resolution now sat in perfect pose, looking at their shoes. Numbed and dazzled by this educational experience. I finally noticed one lone figure somewhat unkempt in attire standing next to me. Although I had met him a number of prior occasions, that was the time I really met Larry. (Sprinthall)

Kohlberg addressed important issues authentically and in an innovative way. This carried others along. Thus, Kohlberg's theory springs also from collective endeavors – initiated, inspired, and directed by Kohlberg, but advanced by his students, friends, and colleagues as well.

Kohlberg's Approach

Kohlberg's approach ran counter to the basic assumptions about the nature of human beings, of norms, and of science dominant at the time. Humans were seen as basically passive (shaped by reinforcement schedules in behaviorism, by parental authorities in psychoanalytic theories, or by the gradual unfolding of genetic endowments in maturationism) and primarily motivated by self-serving concerns (avoiding external or superego sanctions). In contrast, Kohlberg – following Piaget – viewed

people as active interpreters of their world and also motivated by intrinsic concerns (for mastery, truth, or other values).

Norms were seen as externally given contingent facts. Kohlberg, instead, viewed norms as propositions raising the validity claim of rightness, that is, in need of and capable of being justified. Psychology followed a natural science epistemology: copying or explaining reality is aimed for by standardizing measurement procedures or experimentally isolating and manipulating variables. In contrast, Piagetian analyses seek to trace universal developmental changes in subjects' interactive constructions of reality; hermeneutic interviews aim at objectivity by guaranteeing not the physical identity of the stimulus, but the identity of its meaning. Kohlberg was one of the pioneers in bringing this Piagetian cognitive frame to bear on academic psychology in the USA.

Also, his interest in moral reasoning was new. Up until then, only normconforming behavior (not moral action) or moral emotions (not moral judgment) had been studied – both from an observer's point of view. Kohlberg, instead, looked at morality from a firstperson perspective. And he did so with a specific – not uncontested – understanding of morality by focusing on judgment he excluded questions of moral character, of habitualized moral dispositions, of virtues. By defining morality on deontic terms he excluded questions of the good life that he left to ethics and/or to the personal domain. By identifying morality with a justice procedure he largely neglected content. By describing this procedure as 'moral musical chairs' he sketched a hypothetical role-taking procedure that remains entirely monological.

Moral Development

The Stage Model

Piaget had described young children's moral development as progress from a heteronomous stage (characterized by egocentrism; unilateral respect; a focus on consequences rather than on intention; an understanding of rules as objective, absolute and unalterable; a belief in immanent justice) to an autonomous morality produced by peer interaction and cognitive decentering. From his study of adolescent development Kohlberg derived a refined description of the stage model and of developmental mechanisms.

Development consists in changes in the reasoning structure justifying judgments of 'right' or 'wrong' and in the concerns motivating norm conformity. Its structural core is a widening of the sociomoral perspective. Universally, there are three levels with two 'hard' stages each. The model is based on the following assumptions:

- each stage is a structured whole that is qualitatively different from the others;
- the sequence of stages is invariant, irreversibly progressive without stage skipping;
- the stages form a hierarchy: higher stages are better differentiated and integrated structures that allow morally better solutions to dilemmas.

The stages are described as follows:

At the preconventional level (typical for most children under nine) situations are judged from a concrete individual perspective. Rules are seen to be external to the self. 'Right' is defined as obeying rules and authority, avoiding punishment, not doing physical harm (Stage 1) or as meeting one's own interests and letting others do the same (Stage 2); the reasons for doing right are to avoid punishment and to bow to the superior power of authorities (Stage 1) or to serve one's own needs or interests (Stage 2).

At the conventional level (typical of most adults) morality is judged from a 'member of society' perspective and is identified with socially shared internalized norms. 'Right' means living up to the expectations of one's group (Stage 3) or doing one's duty in society (Stage 4); the reasons for doing right are the need to be 'good' in one's own or others' eyes and caring for others (Stage 3) or the need to maintain a good conscience by meeting defined obligations (Stage 4).

At the postconventional level (reached by few adults only) moral values are defined from a 'prior to society' perspective in terms of self-chosen universal principles. 'Right' is upholding legal contracts (Stage 5) or is defined in terms of universal ethical principles such as equality and respect for the dignity of human beings as individuals (Stage 6). The reasons for doing right are obligations to the social contract (Stage 5) or a commitment to principles that have been rationally understood and accepted as valid (Stage 6).

Tests of Moral Development

In the Moral Judgment Interview (MJI, Colby and Kohlberg 1987) hypothetical moral dilemmas are given which allow of no 'right' answer (e.g., Should a man steal a drug to save his dying wife?). Subjects are first to decide what the protagonist ought to do (uphold the law or preserve life?) and then to justify and defend this decision against moral counterarguments advanced by the interviewer who aims at eliciting the subject's highest competence level. The responses are matched to structurally similar criterion judgments listed for each stage in the scoring manual.

In the Social Reflection Measure (SRM) Gibbs et al. (1992) developed a brief interview and coding manual that can be used for younger children. Objective tests have also been developed in which subjects rank their preference for a set of standardized responses that express stage-specific considerations (Rest's Defining Issue Test, DIT; Lind's Moral Judgement Test, MUT). Preference measures produce scores that are consistently (about one stage) higher than spontaneously produced answers, yet do correlate in medium range with scores in MJI.

Mechanisms of Development

In contrast to social learning and psychoanalytic theories which assume identification or modeling as the primary developmental mechanisms, cognitive theories see development as a result of internal restructuring in response to cognitive conflict. There is experimental evidence in support of this mechanism (cf Walker, 1986). The need to reconcile conflicting views may be the reason why taking leadership roles advances moral development.

This understanding of developmental mechanisms has inspired pedagogic interventions. Classroom discussions of moral dilemma tend to produce (albeit slight) moral growth.

The institutionalization of just communities, that is, structures allowing for democratic self-government, have been shown to improve the moral atmosphere of schools or prisons considerably and generate feelings of reciprocal responsibility (Higgins et al. in Kurtines and Gewirtz 1984).

Empirical Support and Criticisms of the Stage Model

Sequence

Age and stage have been shown to correlate in cross-sectional studies. Regression or stage skipping have rarely been found in longitudinal studies and not been induced in experimental studies designed to produce moral development (Snarey, 1985; Walker, 1986). Thus, the claim of a developmental sequence and of an invariant order stands.

Hierarchy

There is conceptual and empirical support for the claim of hierarchy. Conceptually, the stage sequence comprises an increasing universalization of perspectives (isolated individual; dyad; group; a given society; a democratically organized society; humanity at large) and an expansion of relevant concerns (negative and positive consequences to the actor; intentions; role obligations; societal institutions and legality; universal rights). Postconventional thinking makes it possible to integrate and balance these different concerns in view of universal principles of equality and respect for the individual. Empirically, subjects prefer higher and reject lower stage reasoning.

Universality

Kohlberg has been accused of ethnocentrism and gender bias in claiming universal validity for a stage model that was initially developed from interviews with males in the USA only. Two issues are at stake: first, is there a universal core to morality, and second, is Kohlberg's stage sequence universal? The first issue is dealt with theoretically in moral philosophy and practically in endeavors to win worldwide acceptance for human rights. The second issue sparked extensive empirical research (*see* Gender Differences in Personality and Social Behavior; Ethnocentrism).

The universal presence of Stages 1–4 is reported in a review of 45 studies from 27 widely diverse cultures involving over 5000 subjects. Postconventional thinking was found in urban samples only, not in traditional tribal or village societies either in Western or non-Western cultures (Snarey, 1985). Principled thinking thus seems a 'metalevel' feature necessary for adjudicating conflicts between subculture-specific norms and not required in normatively integrated and isolated cultures.

Gilligan (1982) has claimed the existence of two moralities: a rigid justice orientation more typical for males (corresponding to Stage 4) and a flexible morality of care and responsibility more typical for females (corresponding to Stage 3). Empirically, reviews of research involving 19 000 subjects have shown that either there are no stage differences between the sexes (Lind et al., 1987) or else they tend not to disadvantage women, or to disappear when education and employment are controlled for (Walker, 1984). Conceptually, Gilligan starts from a more encompassing understanding of morality than Kohlberg by including questions about the good life. Also, there are some theoretical confusions: 'care' often is experienced as a 'duty' by women, and flexibility – if not a mere reflection of powerlessness – seems a correlate of a modern secularized moral understanding ('ethics of responsibility,' Max Weber) rather than of sex membership.

Another View of Moral Development?

One core issue concerns the relationship between judgment and action. This problem has been dealt with in several ways. First, the claim was raised and empirically supported that consistency will increase with moral development (Kohlberg and Candee, 1984; Blasi, 1980). This makes sense given that subjects referring to self-serving concerns will be assigned to lower stages. Later, responsibility judgments (expressing the personal relevance of moral judgments, Higgins et al., 1984; Blasi, 1984), and the distinction between types A (considering pragmatic concerns) and B (exclusively focusing on moral values) have been introduced (Kohlberg, 1984; Colby and Kohlberg, 1987). These additions are remedies to deal with the fact that individuals differ in intensity of moral commitment – as is especially clearly evidenced by Colby's and Damon's (1994) 'moral exemplars.'

A more far reaching theoretical revision, however, may be called for in view of recent studies on children's moral understanding. Turiel (1983) and Nucci and Turiel (1993) have shown that young children distinguish between moral rules (enjoying universal, unalterable, intrinsic validity) and conventional or religious rules (that depend on authorities or God and are valid only for in-group members). Nevertheless, most young children expect a wrongdoer to feel good and only gradually begin to ascribe negative emotions (Nunner-Winkler and Sodian, 1988). Until about 6 or 7 years (i.e., before self-reflexive role-taking is developed) these predict moral behavior. By about 10 years almost all children (even those low in moral motivation) say they expect a wrongdoer to feel bad. In justifying these negative emotions most children tend to refer neither to impending sanctions nor to empathic identification with the victim but to the fact that a wrong had been committed. These departures from Kohlberg's description of preconventional thinking result from differences in problems addressed and measurement procedures used: Turiel and Nucci and Lee questioned children's understanding of norm validity. Nunner-Winkler focused on the intensity and type of moral motivation. Kohlberg asked for action recommendations in moral dilemmas. Until they begin to care about morality children will guide these along prudential rather than moral considerations.

A new reading of moral development suggests itself. Universally, children acquire an early understanding of the categorical ought: they know that moral norms are intrinsically valid and motives should not be self-serving. Thus, from early on, all have a good cognitive grasp of the constitutive feature of the moral domain – its deontic character. Moral competence, however, requires more. Sociocognitive development (which constitutes the structural core of Kohlberg's stage descriptions) and the acquisition of knowledge are needed for passing judgments in more complex situations. Moral motivation is needed to translate knowledge into action. These requisites are acquired in later and differential learning processes.

Kohlberg's Legacy

In discussing Kohlberg's contribution, Rest (in Kuhmerker et al., 1991) notes that the stage concept may dwindle away and additional variables (e.g., content, motivation, or character) may prove of import. What will remain is Kohlberg's part in the paradigm shift from behaviorism to cognitive psychology. More specifically, what will remain is his emphasis on the cognitive side of morality; his striving for reconstructing it from a participant perspective and linking it with philosophical quests for justifiability; his insistence that intrinsic motivation exists (still an urgent claim given rational choice reductionism); his emphasis that there is more to morality than 'being friendly with your friends' (Rest, 1986). In theory and practice Kohlberg was committed to an understanding of morality that demands building just institutions and treating all human beings with equal respect.

See also: Dewey, John (1859–1952); Erikson, Erik Homburger (1902–94); Gender Differences in Personality and Social Behavior; Hall, Granville Stanley (1844–1924); James, William (1842–1910); Kant, Immanuel (1724–1804); Moral Development and Education; Moral Development, Cultural Differences In; Moral Reasoning in Psychology; Morality 'East' and 'West': Cultural Concerns; Piaget's Theory of Human Development and Education; Piaget, Jean (1896–1980).

Bibliography

Clark Power, F., Higgins, A., Kohlberg, L., 1989. Lawrence Kohlberg's Approach to Moral Education. Columbia University Press, New York.

Colby, A., Kohlberg, L., 1987. The Measurement of Moral Judgment. vol. 1: Theoretical Foundations and Research Validation. vol. 2: Standard Issue Scoring Manual. Cambridge University Press, New York.

Gibbs, J.C., Basinger, K.S., Fuller, D., 1992. Moral Maturity. Measuring The Development of Sociomoral Reflection. Erlbaum, Hillsdale, NJ.

Kohlberg, L., 1969. Stage and sequence. The cognitive-developmental approach to socialization. In: Goslin, D.A. (Ed.), Handbook of Socialization Theory and Research. Rand McNally, Chicago, pp. 347–480.

Kohlberg, L., 1981. Essays on Moral Development: vol. 1. The Philosophy of Moral Development. Moral Stages and The Idea Of Justice. Harper & Row, San Francisco.

Kohlberg, L., 1984. Essays on Moral Development: vol. II: The Psychology of Moral Development. The Nature and Validity of Moral Stages. Harper & Row, San Francisco.

Kohlberg, L., 1987. Child Psychology and Childhood Education. Longman, New York.

Kohlberg, L., Levine, C., Hewer, A., 1983. Moral Stages: A Current Formulation and a Response to Critics. Basel.

Kuhmerker, L., Gielen, U., Hayes, R.L. (Eds.), 1991. The Kohlberg Legacy for the Helping Professions. REP Books, Birmingham, AL.

Kurtines, W.M., Gewirtz, J.L., 1984. Morality, Moral Behavior, and Moral Development. Basic Issues in Theory and Research. Wiley, New York.

Kurtines, W.M., Gewirtz, J.L. (Eds.), 1991. Handbook of Moral Behavior and Development. Erlbaum, Hillsdale, NJ. vol. 1: Theory. vol. 2: Research, vol. 3: Application.

Modgil, S., Modgil, C. (Eds.), 1986. Lawrence Kohlberg. Consensus and Controversy. Falmer Press, Philadelphia.

New Directions for Child Development, 1990. The legacy of Lawrence Kohlberg 47, 1–106.

Nunner-Winkler, G., 1998. Development of moral understanding and moral motivation. In: Weinert, F.E., Schneider, W. (Eds.), Individual development from 3 to 12. Findings from the Munich Longitudinal Study. Cambridge University Press, New York, pp. 253–290.

Rest, J.R., 1986. Moral development. Advances in Research and Theory. Praeger, New York.

Köhler, Wolfgang (1887–1967)

Viktor Sarris, Johann-Wolfgang-Goethe University, Frankfurt am Main, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Wolfgang Köhler, a German–US psychologist, was one of the founders of the Gestalt school of psychology in the beginning of the twentieth century. The scope of his scientific contributions was extremely broad, ranging from sensory psychophysics to theoretical physics, and the logic of discovery, and from perceptual psychology to animal cognition, and human values. Köhler's experiments with chimpanzees made him internationally famous; together with Max Wertheimer (1880–1943) and Kurt Koffka (1886–1941) he became an influential scholar first in Germany and then, after his emigration (1935), also in the USA. Early in his career, Köhler demonstrated the ape's so-called 'transposition' of perceived stimulus relationships; however, his speculative neurophysiology of perception was refuted on both experimental and theoretical grounds. Today his work in general is seen as a major breakthrough in animal problem solving as well as in human perception and cognition. After World War II, Köhler received numerous honors; and he was president of the American Psychological Association for the 1958–59 term.

Wolfgang Felix U. Köhler, a leading German–US psychologist in the twentieth century, was born to Franz E. Köhler, a German gymnasium (high-school) teacher and his wife Wilhelmine on 21 January 1887 in Reval (today, Tallinn), Estonia, at that time a part of the Russian Empire. Köhler's experiments with apes made him internationally renowned early during his long scientific career. As a cofounder of a school called Gestalt psychology, together with Max Wertheimer and Kurt Koffka, he became an influential scholar many years before he left Europe to join the faculty of Swarthmore College, Pennsylvania in the USA (see Table 1).

Major Events in Wolfgang Köhler's Life and Work

After his graduation from high school in 1905, Köhler undertook university studies in philosophy, history, and natural science in Tübingen, Bonn, and then Berlin, where he concentrated on psychology under Carl Stumpf's supervision. Besides his studies in experimental psychology, especially in psychoacoustics with Stumpf, he continued to pursue his interests in natural science and physics under Max Planck and Walther Nernst. In his doctoral dissertation, which he finished in 1909, he combined physical acoustics with psychoacoustics. This research, published in the same year in the *Zeitschrift für Psychologie*, demonstrated his great talent as a scientist. His findings forced his teacher Carl Stumpf to revise and do further research on his own psychoacoustical theories.

As an assistant at Frankfurt's institute of psychology (1909–13) under its director Friedrich Schumann, Köhler worked with Max Wertheimer and Kurt Koffka, who also had been trained in natural science and experimental psychology. This cooperation led to a life-long exchange among these three men, each of them involved in the advancement of psychological theory and research. Köhler worked on the then widely accepted '*constancy hypothesis*,' continued his psychoacoustical investigations in Frankfurt, and in his paper on '*unnoticed sensations and errors of judgment*' (1913) merged his individual line of research with that of Max Wertheimer's pioneering Gestalt work (King and Mich. Wertheimer, 2005).

From 1913 to 1920, Köhler conducted research on the mentality of apes, chickens, and children at the anthropoid

Table 1 Chronology: major events and dates in Wolfgang Köhler's life and work

1887	Born on 21 January 1887 in Reval (Tallinn), Estonia
1905–09	University studies in Tübingen, Bonn, and Berlin
1909	Doctoral dissertation thesis with Carl Stumpf at Berlin's Institute of Psychology
1909–13	Assistant at Frankfurt's Institute of Psychology; collaboration with Max Wertheimer (1880–1943) and Kurt Koffka (1886–1941)
1911	*Privatdozent* (lecturer) at Frankfurt
1913–20	Director of the Anthropoid Station on Tenerife; first epoch-making publications on 'transposition' and 'insight' in the chimpanzee
1920	Book: Treatise on Gestalt phenomena in natural sciences (*Die physischen Gestalten in Ruhe und im stationären Zustand: Eine naturphilosophische Untersuchung*)
1921	Professor at Göttingen. Founding of the journal *Psychologische Forschung* (*Psychological Research*), together with Max Wertheimer, Kurt Koffka, and others; major journal of Gestalt psychology (until 1938)
1922–35	Full professor of psychology and director of Berlin's Institute of Psychology
1925	Book: *The Mentality of Apes* (trans. from the 1921 German book)
1929	Book: *Gestalt Psychology* (later ed. 1947)
1935	Final clash with the Nazi authorities in Berlin and emigration to the USA; psychology professor at Swarthmore College until 1958
1938	Book: *The Place of Value in a World of Facts*
1944	Influential research publication on 'figural after-effects' (together with Hans Wallach)
1954–56	Perceptual research at the *Institute for the Advanced Studies* in Princeton, NJ
1958–59	President of the *American Psychological Association* (USA)
1959–67	Perceptual and psychophysiological research at Dartmouth College, New Hampshire (USA)
1967	Deceased on 11 June 1967 in Enfield, New Hampshire (USA)

research station of the Prussian Academy of Sciences on Tenerife, Canary Islands. He began his investigations with perceptual and problem-solving tasks, demonstrating the ape's and the chicken's size constancy and their transposition of perceived stimulus relationships. He also demonstrated the chimpanzee's fabrication and use of simple 'tools.' Most famous became the success of one of his apes named Sultan in fitting together two sticks to make them long enough to reach the goal object, a banana. Köhler (1917, 1921) also did extensive work on the chimpanzees' ability to solve a problem by stacking boxes on top of each other, thus demonstrating, once again, 'insightful' animal behavior. Köhler (1920; see **Table 1**) expanded the Gestalt-theoretical perspective when he wrote his treatise on '*Die physischen Gestalten in Ruhe und im stationären Zustand*' ('The physical configurations at rest and in stationary states'), a book in which he aimed to integrate basic principles in psychology and physics in a new psychoneural solution to the mind-body problem (see Section Köhler's Accomplishments and His Impact on Psychology). He became a professor at Göttingen's psychology institute for a short time (1921) before he followed Carl Stumpf in the chair at Berlin.

In 1922 Köhler became full professor of philosophy and director of the Psychological Institute at the University of Berlin, where he worked together with other eminent scientists like Max Wertheimer and Kurt Lewin. Here the international fame of the Berlin School of Gestalt psychology grew until its peak in the early 1930s. It was during this time that the Gestalt principles of the organization of perception were published and extended to new research issues (Wertheimer, 1923) like memory, cognition, motivation, and aesthetics.

During these years Köhler also lectured abroad several times, in the USA at Clark University (1925–26) and at Harvard University (1934–35), in South America in Uruguay (1930), as well as in Brazil and Argentina (1932). When Köhler encountered the harassment of university scholars by the Nazi government in 1933, he fought racial discrimination and struggled against the political authorities of those days; he may have been the only psychology professor in Germany who protested in public against the totalitarian system (Henle, 1978). When in 1935 his assistants were fired, he left Germany and followed his colleagues into exile in the USA.

In 1935 Köhler became a psychology professor at a small but prestigious undergraduate college, Swarthmore, near Philadelphia. There he continued his research, but on a smaller scale as his efforts were hampered by the absence of advanced facilities and the lack of a doctoral program at this institution (King and Mich. Wertheimer, 2005; cf also Ash, 1995; Sarris, 1995).

Nevertheless, Köhler, together with his American colleagues and students, published important new findings on figural after-effects and on related perceptual phenomena; he also studied the cortical activity associated with pattern vision (see Section Köhler's Accomplishments and His Impact on Psychology below), and wrote three further books on Gestalt psychology, all in English (Köhler, 1958).

After World War II, Köhler received worldwide recognition for his life work, including several honorary doctorates from American and European universities. He was a recipient of the APA's (American Psychological Association) Distinguished Scientific Contribution Award and was elected a Member of the National Academy of Sciences. He was also named an 'honorary member' (*Ehrenbürger*) of the Free University of Berlin, where he lectured several times, first as a visiting professor. He was elected president of the APA for the 1958–59 term, an unusual honor for an emigré scientist in the USA (Sarris, 2002).

Köhler's Accomplishments and His Impact on Psychology

Gestalt theory had begun after the turn of the twentieth century as a scientific revolt, first against Wilhelm Wundt's ruling elementaristic ('atomistic') approach, and later opposed to John Broadus Watson's – then dominant – behaviorism. It took several decades before Gestalt theory became a fully acknowledged perspective in psychology.

Köhler's ideas about thinking were opposed strongly to those of the other schools of psychology. Some of his research was specifically directed against E.L. Thorndike's influential behavioristic interpretation of learning and problem solving in animals. According to the latter view, a correct behavioral response occurs in the course of blind trial-and-error activity; on repeated trials the correct response is reinforced and thus strengthened simply by association in an 'automatic' way (Thorndike's 'law of effect'). Köhler, in sharp contrast to that approach, argued that Thorndike's puzzle-box task did not allow any perceptual-cognitive insight into the whole procedural arrangement and thus represents no possibility of real problem solving; 'insight' requires a situation which is completely 'natural' and 'visible' to the animal (Köhler, 1921).

According to Köhler (1951), problem solving is based on the organization of meaningful relations. It begins with a situation and a goal that cannot be directly reached (the so-called 'detour' problem). The process of problem solving requires a reexamination of the situationally given materials which must be 'reorganized,' quite analogously to the reversal of perceptual patterns. Köhler was criticized by his opponents for not considering the influence of past experience on problem solving. He did not deny the role of past experience in principle, but considered it of minor significance in insightful learning. This issue has still not been fully resolved today.

Basic to Köhler's perceptual research was the study of stimulus relationships, starting with his comparative work on transposition in apes, chickens, and children (Köhler, 1917); for example, a chicken is presented with two gray samples, a darker one, *A*, and a lighter one, *B* (the positive stimulus): after successful discrimination training the chicken is tested with a new pair of samples, *B* and *C* (*C* being lighter than *B*). The majority of Köhler's hens chose *C* over *B*, thus showing relative choice behavior ('transposition'). Köhler obtained similar experimental results with his chimpanzees as subjects, using size as well as brightness as stimulus dimensions, thereby demonstrating the priority of 'structural' properties in perception and learning. From the perspective of Gestalt theory, the effective stimulus for a given perception or

perceptual choice is always relational in character; what is perceived and reacted to is fundamentally a given 'pattern,' which is conceived of as the interaction (relation) between a particular stimulus and its surround. This idea, although challenged by the behaviorist school, has been proved to be accurate, by and large; it fits with the Gestaltists' emphases on the importance of the *frame of reference* in perception (cf Sarris, 1994, 2006).

Köhler's interest in the physical processes in the brain, which dates back to his years on Tenerife (or even before), generated an influential book on a special psychoneural theory of perception. This monograph (1920) dealt with the neurophysiological processing during a given experience, in an attempt to relate the laws of physics to the brain's physiochemical characteristics. It was to this end that he later directed his research on figural after-effects and so-called satiation (Köhler and Wallach, 1944). His explanation for these much – investigated phenomena was based on the notion that direct electric currents are generated in the cortex, resulting in different active 'fields' for various perceptual phenomena (Köhler and Held, 1949). However, this theory was not accepted widely by the scientific community; on the contrary, it was later refuted on both experimental and theoretical grounds in criticisms of Köhler's speculative psychophysical isomorphism.

Köhler expanded the Gestalt principle of perceptual and cognitive organization to the province of memory and learning by assuming that recognition and recall are functionally another case of grouping. When a form or pattern is perceived it tends to be subsequently recognized and recalled in a well-structured, organized way; memory traces in the brain are conceived of as neurophysiological correlates of recognition and recall.

Köhler (1947) developed several criticisms of classic associationist psychology. For instance, he attacked the 'law' of association by contiguity as being merely a mechanistic principle:

> Two processes A and B happen to occur together and, whatever the nature of A and B may be, a bond is formed between them! I do not know a single law in physics or chemistry which could in this respect be compared with the law of contiguity.
>
> Köhler (1947)

Köhler argued that meaningful pairs of words are always associated more readily than are pairs of nonsense elements as the former can more easily ('naturally') be organized into a unified pattern ('Gestalt'). According to this account, recognition and recall depend on an interaction between the assumed memory traces that are relationally determined (Köhler, 1951; see the reference details in Sarris, 2006).

Köhler's and his followers' work gradually influenced the direction and content of modern psychology. Gestalt psychology turned out to offer an important alternative to behaviorism and its related perspectives, especially in the fields of perception and cognition, since the 1960s in the twentieth century.

Köhler was a research-oriented scientist aiming at an integrative and multidisciplinary development of psychology as a science. His work is now seen as a theoretical and experimental breakthrough in animal problem solving and in human perception and cognition. Nowadays modern psychophysical and neurobiological theories are being used to reexamine the classic Gestalt issues.

Some of the central tasks of the brain – such as figure-ground coding, perceptual organization, object identification, and the formation of object memory – concern the integration of separately coded stimulus features into coherent and meaningful object representations. These fundamental problems of explaining the basis of perception and cognition, namely the understanding of the computational or informational significance of neurobiological activity, have not been solved so far. Nowadays these issues are approached by the concerted efforts of different disciplines (notably, psychophysics, neurobiology, and neurocomputing; e.g., Spillmann, 2012).

Köhler also envisioned the science of psychology as an expanding area to be developed according to an ethical understanding (Köhler, 1969). His analysis of values and human ethics led Köhler to reject the commonly held assumption that values are merely subjective; on the contrary, his position offers an alternative to the relativism that seems to dominate discussions of values in science and life.

Final Notes

Köhler died on 11 June 1967, at the age of 80 at his home in Enfield, NH, USA. There has been a Wolfgang Köhler Society on Tenerife, Spain, since 1992; and the Wolfgang Köhler Memorial Lectures are organized biannually by the *Deutsche Gesellschaft für Psychologie*, Today there is also a "Wolfgang Köhler Primate Research Center" at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany. Germany's national psychology association. The Wolfgang Köhler Archives are housed in the Archives of the Philosophical Society of America in Philadelphia. The complete bibliography of Köhler is published in Henle (1971).

See also: Cognitive Psychology: History; Heider, Fritz (1896–1988); Lewin, Kurt (1890–1947); Primate Behavior: Significance for Understanding Humans; Psychology, History of (Twentieth Century); Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Ash, M.G., 1995. Gestalt Psychology in German Culture 1890–1967. Cambridge University Press, Cambridge, UK.

Henle, M. (Ed.), 1971. The Selected Papers of Wolfgang Köhler. Liveright, New York.

Henle, M., 1978. One man against the Nazis – Wolfgang Köhler. American Psychologist 33, 939–944.

King, D.B., Wertheimer, Mich, 2005. Max Wertheimer & Gestalt Theory. Transaction, New Brunswick.

Köhler, W., 1920. Die physischen Gestalten in Ruhe und im stationären Zustand. Vieweg, Braunschweig.

Köhler, W., 1917. Die Farbe der Sehdinge beim Schimpansen und beim Haushuhn (The color of visual objects in the chimpanzee and in the chicken). Zeitschrift für Psychologie 77, 248–255.

Köhler, W., 1921. Intelligenzprüfungen an Menschenaffen (Intelligence Testing of Apes). Springer, Berlin (Trans. into English, 1925 The Mentality of Apes. Harcourt, Brace, New York).

Köhler, W., 1947. Gestalt Psychology. Liveright, New York.
Köhler, W., 1951. Relational determination in perception. In: Jeffress, L.A. (Ed.), Cerebral Mechanisms in Behavior. Wiley, New York.
Köhler, W., 1958. Perceptual organization and learning. American Journal of Psychology 71, 311–315.
Köhler, W., 1969. The Place of Value in a World of Facts. Liveright, New York.
Köhler, W., Held, R., 1949. The cortical correlate of pattern vision. Science 110, 414–419.
Köhler, W., Wallach, H., 1944. Figural after-effects: an investigation of visual processes. Proceedings of the American Philosophical Society 88, 269–357.
Sarris, V., 1994. Contextual effects in animal psychophysics: comparative perception. Behavioral and Brain Sciences 17, 763–764.
Sarris, V., 1995. Max Wertheimer in Frankfurt: Beginn und Aufbaukrise der Gestaltpsychologie (Max Wertheimer in Frankfurt: Beginnings and Foundational Dilemmas of Gestalt Psychology). Pabst, Lengerich.
Sarris, V., 2002. Wolfgang Köhler (1887-1967) - ein großer Psychologe des 20. Jahrhunderts. In: Böhme, G. (Ed.), Die Frankfurter Gelehrtenrepublik. Schulz-Kirchner, Idstein, pp. 183–198.
Sarris, V., 2006. Relational Psychophysics in Humans and Animals: A Comparative-Developmental Approach. Psychology Press, London.
Spillmann, L. (Ed.), 2012. Max Wertheimer on Perceived Motion and Figural Organization. MIT Press, Cambridge, MA.
Wertheimer, M., 1923. Untersuchungen zur Lehre von der Gestalt (Investigations concerning the theory of Gestalt' patterns). Psychologische Forschung 4, 301–350. (Full transl into English in Spillmann, 2012).

Kraepelin, Emil (1856–1926)

Paul Hoff, Psychiatric University Hospital Zurich, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article – after giving an overview on Emil Kraepelin's life – discusses Kraepelin's concept of mental illness in general and psychiatric research in particular. Although he was hardly interested in philosophical presuppositions of psychiatric work, his publications prove the influence of major theoretical concepts like realism, naturalism and parallelism. The basic idea of 'natural disease entities' in psychiatry and the development of Kraepelin's clinical nosology over a period of nearly 50 years is then described before focusing on the significant influence Kraepelin still has on present-day psychiatry, especially with regard to psychiatric diagnoses and research. However, in recent years the 'neo-Kraepelinian' perspective has also been criticized substantially especially with regard to the nosological dichotomy of schizophrenic and affective psychoses. A thorough knowledge and balanced interpretation of Kraepelin's work is indispensable for a clear understanding of this debate and for its further development.

Biographical Note

Emil Kraepelin was born on 15 February 1856 in Neustrelitz (Mecklenburg, West Pomerania, Germany). He studied medicine in Leipzig and Würzburg from 1874 until 1878. During his medical studies in Würzburg, Kraepelin worked as a guest student at the psychiatric hospital under the directorship of Franz Rinecker. He began his professional life in 1878 working with Bernhard von Gudden at the District Mental Hospital in Munich, where he stayed until 1882. Kraepelin then moved to Leipzig, where he worked with Paul Flechsig and Wilhelm Erb. He was promoted to university lecturer there in 1883.

His lifelong personal and scientific relationship with Wilhelm Wundt began in Leipzig. Encouraged by Wundt, Kraepelin, at the age of 27, wrote his *Compendium of Psychiatry* in 1883, the precursor of his influential textbook *Psychiatry*, which was published in nine editions between 1883 and 1927. Kraepelin stayed in contact with Wundt by correspondence and paid him several visits before Wundt's death in 1920; in many of his publications he emphasized and acknowledged the importance of this relationship.

In 1884 Kraepelin married Ina Schwabe. After a short period of employment in Leubus (in Silesia) and Dresden, Kraepelin was appointed professor of psychiatry at the University of Dorpat in 1886. In 1891 he took over the chair of psychiatry at the University of Heidelberg. From 1903 until 1922 Kraepelin was ordinary professor of psychiatry in Munich where, in 1904, he opened the new building of the psychiatric hospital of the Ludwig Maximilian University, the main part of which is still in use. Despite the difficult conditions caused by World War I, Kraepelin founded the *Deutsche Forschungsanstalt für Psychiatrie* (the German Research Institute for Psychiatry) in Munich in 1917 to encourage and improve psychiatric research (Weber, 1991). During his time in Munich, Kraepelin's colleagues at the university hospital and the *Forschungsanstalt* included Alois Alzheimer, Franz Nissl, Korbinian Brodmann, Walter Spielmeyer, August Paul von Wassermann, and Felix Plaut – to mention but a few. In 1924 the *Deutsche Forschungsanstalt* was integrated into the Kaiser Wilhelm Society and became in 1945 the Max Planck Institute for Psychiatry as part of the Max Planck Society. Emil Kraepelin died in Munich on 7 October 1926. Kraepelin's memoirs were published in German (1983) and in English (1987).

Kraepelin's Concept of Psychiatry as a Science

Three important authors who strongly influenced Kraepelin's understanding of psychiatry have to be mentioned: Griesinger, Kahlbaum, and Wundt.

Wilhelm Griesinger had marked the turning point in nineteenth-century psychiatry by calling for thorough clinical and pathophysiological research based on the premise that 'mental illness is an illness of the brain.' But Griesinger's theory was by no means as simple as this one statement, which is so often quoted. He held differentiated views on the problem of somato- and psychogenesis, although favoring the first in the case of what was later to be called 'endogenous psychoses.'

Karl Ludwig Kahlbaum continued the traditions of French psychopathology as represented by Jean-Pierre Falret and Antoine Laurent Jessé Bayle. He developed a clinically orientated research method in the second half of the nineteenth century in Germany, strongly focusing on the course of illness. This, like Griesinger's theory, was believed to be a victory over the speculative concepts of the so-called 'romantic medicine' and even more so an argument for a critical distance to the position of the harsh and unreflected 'somaticists,' who were no less dogmatic than some of the 'psychicists.' Kahlbaum had clearly recognized the conceptual and methodological differences between neuroanatomical and psychopathological approaches. With 'progressive paralysis of the insane' as an example he explained the route from the 'syndrome-course unit' (*Syndrom-Verlaufs-Einheit*) to the – postulated – etiologically based 'disease unit' (*Krankheitseinheit*).

The influence of Wilhelm Wundt, one of the founders of experimental psychology, on Kraepelin can hardly be overestimated. Wundt's aim, on the one hand, was to establish psychology as a kind of natural science using an experimental approach to collect data. In this line of thought, he criticized the highly speculative approach of 'philosophy of nature' in the

sense of Schelling and Schleiermacher. On the other hand, he did not agree with materialism or association psychology in the sense of Herbart. Especially in his earlier writings, Wundt clearly favored a parallelistic point of view in the mind–body problem. The young Kraepelin was highly impressed by these Wundtian ideas since they allowed experimental research to be successfully applied in psychology without ignoring the epistemological differences between the mental and the physical. Along the years, Kraepelin modified Wundt's concepts by extracting what he regarded useful for empirical research in clinical psychiatry. That is why Wundt's psychology, viewed through the 'filter' of Kraepelin's texts, seems much more unified and straightforward than it really was. In brief: Kraepelin simplified and, so to speak, 'smoothed out' Wundt's concept, but he did not falsify it.

Kraepelin was not profoundly interested in the philosophical basis and implications of psychiatry. His view of what (natural) science was, and what impact it had or should have on social and political developments, was, as Engstrom (1991, 1995) has shown convincingly, typical for the way natural scientists saw themselves at the turn of the last century. However, despite his skeptical attitude toward continuous philosophical reflections in psychiatry, the following major theoretical frameworks can be identified as underlying – seldom explicitly, most often very implicitly – Kraepelin's concept: realism, parallelism, experimental approach, naturalism.

Realism

Kraepelin, like most of his contemporaries in academic psychiatry, believed in an independently existing 'real world' that includes other people and their healthy or disturbed mental processes. Kraepelin repeatedly pointed out that the psychiatric researcher has to describe objectively what 'really' exists and what 'nature presents to him' – the formulations differ, but the essence is always a strictly realistic philosophy. The consequences for psychiatric nosology are evident: such realism will lead to the concept of 'natural disease entities' which exist independently of the researcher. The scientist describes what he finds, or – in stronger terms – describes 'given things.' His own activity in constructing scientific hypotheses or diagnostic entities is underestimated or may even be overlooked.

Parallelism

Kraepelin advocated psychophysical parallelism. Like Wilhelm Griesinger, whom he admired for his critical attitude toward speculative psychiatric theories, he disapproved of reductionistic materialism which *identifies* mental events with neurobiological processes. Kraepelin spoke about two kinds of phenomena, somatic and psychological, which are decidedly different, but closely connected. He defended the existence of mental phenomena against all kinds of what he, like Karl Jaspers, called 'brain mythologies.' Contrary to Wundt, however, Kraepelin, although calling himself a parallelist, did not enter the widely ramified philosophical debate on this issue. In particular, he did not critically differentiate between parallelism and interactionism, nor did he realize that any strictly defined parallelism makes it more than doubtful if mental phenomena still may be regarded as an independent sphere and not just as having a one-to-one relationship with the somatic level. The latter option, of course, comes close to (causal) determinism.

As a consequence of his ambivalent, if not blurry position in the mind–body debate, there indeed is an implicit tendency toward monism in Kraepelin's writings, particularly when it comes to his ideas about psychology as a natural science. But it should be emphasized that this 'monistic tendency' was not a metaphysical one, but a weak version of what may be called 'methodological monism,' insofar as Kraepelin decidedly favored the quantitative methods brought forward by the natural sciences (Verwey, 1985).

Experimental Approach

Kraepelin argued that the psychological and psychophysiological experiment should become a major scientific tool to reach a more profound understanding of disturbed, but also of healthy mental processes. Both Wundt and Kraepelin realized the difference between a physical and a psychological experiment, but in their view the experimental design as such did not differ significantly in both areas. Kraepelin seems to have considered the experimental approach a kind of guarantee for the scientific status of psychiatric research. From this it may be inferred that he rated the experimental approach higher than the mere description of clinical phenomena, although, of course, he accepted and even stressed the latter method as indispensable tool, especially if combined with follow-up examinations. In conclusion, Kraepelin developed and maintained a skeptical attitude toward subjective – especially biographically determined – aspects of mental disorders, which could not be studied experimentally.

Naturalism

In his early writings – mainly in those on forensic topics – Kraepelin clearly expressed his opinion that a priori ideas (in Kant's sense), freedom of the will, and unquestionable moral values do not exist. For him, man is nothing but a part of nature, and, consequently, anything man can do is a product of this natural existence. This position brought Kraepelin very close to what is nowadays called 'evolutionary naturalism' (Hoff, 1998). Later in his life, he became somewhat more cautious concerning these matters, but there is no reason to believe that he substantially changed his mind. This naturalistic, 'antimetaphysical' point of view, of course, made Kraepelin feel sympathetic with Darwinist and biologistic theories, although he always rejected oversimplifications such as those in the monistic theories of Ernst Haeckel, Jakob Moleschott, and Ludwig Büchner.

Kraepelin and Degeneration Theory

There are strong links between the basically naturalistic attitude of most academic psychiatrists in the end of the nineteenth century and the increasing influence of degeneration theory, although both aspects should, of course, not be mixed up. This

issue is discussed here in some detail, because there has been substantial criticism toward Kraepelin's broad, albeit not fully uncritical acceptance of degeneration theory. His position was even said to have carried "overtones of protofascism" (Shepherd, 1995a,b).

Degeneration theory is far from being only a psychiatric or even medical issue. It had gained wide influence not only in scientific, but also in philosophical and political circles in the last decades of the nineteenth century. As for psychiatry, it had important roots in French psychopathology, especially in the writings of Bénédict Augustin Morel and Valentin Magnan. The central idea of this concept was that in 'degenerative' illness there is a steady decline in mental functioning and social adaptation from one generation to the other. For example, there might be an intergenerational development from nervous characters to major depressive disorder, then to overt psychotic illness and, finally, to severe and chronic cognitive impairment, i.e., dementia. It should be noted, however, that this theory has always been a vague and speculative concept, which was brought forward decades before the rediscovery of Mendelian genetics and their application to medicine in general and to psychiatry in particular (Chamberlin and Gilman, 1985; Dowbiggin, 1985; Hermle, 1986; Liegeois, 1991; Pick, 1989).

Kraepelin and with him most of the contemporary authors of psychiatric textbooks used arguments derived from degeneration theory broadly. He made a special reference to them with regard to manic-depressive illness, paranoia and personality disorders. However, like in his Swiss colleague Eugen Bleuler in Zurich, Kraepelin's attitude toward degeneration theory was not straightforward positive, but included critical arguments. For example, although commenting approvingly on the basic ideas of Cesare Lombroso's 'criminal anthropology,' Kraepelin did not accept the idea of overt 'stigmata degenerationis,' by which individual persons could be identified as being 'degenerated' simply by their physical appearance (Zubin et al., 1985).

To sum up this aspect, there is no doubt that Kraepelin, on the one hand, in many respects accepted degeneration theory and implemented it into the debate on etiology and pathogenesis of mental disorders. On the other hand, it is not appropriate to draw a *simple and direct* line from earlier versions of degeneration theory to the horrible crimes against the mentally ill that were committed during the period of National Socialism. A differentiated view is needed here, which will only be reached by thorough and unbiased research (Hoff, 2008).

Kraepelin's Psychiatric Nosology

Basic Aspects

On the clinical level, Kraepelin changed the details of his diagnostic system over and over again. On the basic level, however, his nosology showed remarkable stability over time: between the second and the ninth editions of his textbook (i.e., from 1887 to 1927) Kraepelin did *not* change the central postulate: He stated that the essential features of all psychotic disorders will eventually be classified in a 'natural' system, no matter which scientific method is applied, pathological anatomy, etiology or clinical symptomatology including long-term course of illness. All these approaches will necessarily converge in the same 'natural disease entities,' simply because they are 'natural kinds' that are detected, not constructed by research (see **Figure 1**).

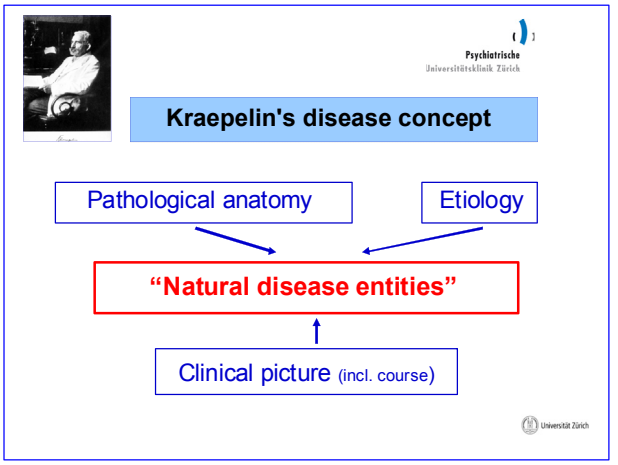

Figure 1 Kraepelin's disease concept.

Kraepelin limited this strong hypothesis to a certain extent in three theoretical papers, written between 1918 and 1920: 'Ends and means of psychiatric research' ('*Ziele und Wege der psychiatrischen Forschung*'; 1918), 'Research in the manifestations of mental illness' ('*Die Erforschung psychischer Krankheitsforme*'; 1919) and 'Clinical manifestations of mental illness' ('*Die Erscheinungsformen des Irreseins*'; 1920). In these papers Kraepelin took into account contemporary arguments such as Karl Birnbaum's differentiation between pathogenetic and pathoplastic factors in mental illness or Robert Gaupp's hypothesis of the possibility of psychogenic delusion; Kraepelin now acknowledged the value of defining certain syndromes as a middle course between nosologically unspecific symptoms and specific diseases. But – and this is essential – at no time did he abandon his postulate of underlying distinct and natural disease entities (Hoff, 1994, 1995).

Clinical Aspects

Kraepelin's clinical nosology is best separated into three periods. The early period, 1880–91, is characterized by the search for a reliable and valid psychiatric system between clear-cut naturalistic beliefs and the methodological framework of experimental psychology in the sense of Wundt. As for nosology, Kraepelin slowly moved away from earlier nineteenth-century concepts which he criticized as unreliable and ill-defined from a clinical and, especially, prognostic point of view. In these years the term '*dementia praecox*' was not yet used. A group of clinically heterogeneous paranoid and hallucinatory psychoses tending to chronicity was labeled '*Wahnsinn*' (insanity). It probably resembled the cases now known as schizophrenic psychoses developing residual states. In addition, Kraepelin introduced '*Verrücktheit*' (madness) as a chronic psychosis with a better prognosis, that explicitly did not lead to residual states. The affective

psychoses were split into three groups: melancholia, mania, and periodical or 'circular' psychosis.

In the middle period, 1891–1915, Kraepelin's thinking reached its most systematic and influential level regarding its clinical as well as scientific implications: Kraepelin significantly broadened his clinical experience and self-consciously created a complete nosological system. He finalized his concept of 'natural disease entities' as discussed above. The main clinical result of this period – first clearly proposed in the sixth edition of 1899 – was the well-known dichotomy of endogenous psychoses: that is, the separation of *dementia praecox* with – as he saw it – a poor prognosis, from manic-depressive illness (bipolar disorder) with a good, or at least better, prognosis. With respect to *dementia praecox*, he supposed an organic defect as the basis of the illness, a kind of 'autointoxication,' leading to the destruction of cortical neurons. The patient's personality may promote the development of the psychotic illness, but it is not a central pathogenetic factor; contrary to most other nosological areas, 'degeneration' was believed to be of much less importance in *dementia praecox*. 'Paraphrenia' was conceptualized as a psychosis with acute and heterogeneous clinical symptomatology, including the development of lasting deficits. Its separation from typical cases of *dementia praecox* was justified by the postulated absence of massive disturbances of volition and by a much lesser degree of affective flattening.

In manic-depressive illness the etiology was said to be much less clear than that of *dementia praecox*. Kraepelin proposed a genetically determined irritability of 'normal' affects, so that the psychosis itself emerged from certain predisposing 'basic states' (*Grundzustände*). The concept of degeneration was an important part of this hypothesis. In this period, Kraepelin integrated different types of circular or recurrent affective illness into the overarching concept of manic-depressive insanity (*Manisch-depressives Irresein*) (6th edition 1899).

Kraepelin's concept of paranoia was modified several times in this period. After the very broad concept of *Verrücktheit* of the early editions of *Psychiatry*, which was of restricted clinical use, he significantly narrowed it, especially in the 5th edition of 1896. Here, paranoia was defined as a severe and chronic delusional illness without constant alteration of personality and volition; the existence of abortive or benign cases was denied up to the 7th edition (1903/04). In the 8th edition (1915) this very rigid concept was broadened again, but not to such a degree as in the early versions. Kraepelin now accepted cases with low severity and a comparably good prognosis, but he maintained the strict separation of *dementia praecox* and paranoia.

In Kraepelin's view – typical of his way of thinking within the theoretical framework of degeneration theory – disorders of personality resulted from a circumscript retardation of psychological development. He argued that, since some patients with personality disorders reach a 'normal' or mature level of affective and cognitive functioning and some don't, it was not justified in this field to postulate clear-cut disease processes as, for example, in *dementia praecox*.

In his later period, 1916–26, Kraepelin reacted to several commentaries and criticisms of his nosology, for example, Alfred Erich Hoche's syndromatic theory and Ernst Kretschmer's suggestion to supplement the Kraepelinian system with a multidimensional approach. He moved toward an internal broadening of his system by reformulating his disease concept as discussed above; he combined a more differentiated view of pathogenesis and the role of individual psychological factors with a strong and unchanged position on the existence of 'natural disease entities.' Compared to this differentiation of the theoretical basis during these years, changes in clinical nosology were small.

Kraepelin and Twenty-First Century Psychiatry

Kraepelin's psychiatry became so influential during his lifetime, because it offered a pragmatical, clinical and prognosis-oriented nosology, developed by a self-confident author who focused on straightforward quantitative and naturalistic research methods. He claimed to abandon speculative aspects of psychiatry as far as was possible, although he himself did – unintentionally – 'import' a number of implicit theoretical and, in part, highly speculative aspects into psychiatry.

In the years after World War II, Kraepelinian ideas – and 'biological' concepts in general – lost much of their influence and were discussed mainly from the historical point of view. However, Kraepelin's concepts were 'rediscovered' in the 1960s and 1970s when 'biological psychiatry' – precursor of the neuroscience approach of our days – became the most influential (and most self-confident) field of psychiatric research. In particular, researchers and clinicians from the English-speaking countries viewed themselves as 'Neo-Kraepelinians' (Blashfield, 1984). But 'Neo-Kraepelinianism' was (and is) not a scientific theory in the strict sense. It has always been a heterogeneous set of concepts, trying to strengthen the basis for neurobiological research in clinical psychiatry. Central to it is the intention to identify the biological basis – the 'natural' basis, in Kraepelin's words – of mental disorders.

But these findings may well lead to a nosological system quite different from Kraepelin's, a topic which has attracted much interest in recent years. Many authors have postulated a 'denosologization' of psychiatric research, especially with regard to neurobiological research (Van Praag et al., 1987) or, more recently, a 'deconstruction' of diagnostic categories in psychiatry (Gaebel and Zielasek, 2008; Read et al., 2004; Van Os and Tamminga, 2007). The main idea here is to prevent nosological prejudices from hampering research by restricting the interpretation of data to the limits of 'classical' concepts, the most prominent of which being the Kraepelinian dichotomy of endogenous psychoses. Taking the recent literature into account, these authors have doubtlessly been right to predict that neurobiological findings will often fail to prove the traditional clinical boundaries. However, such an approach, although continuously modifying Kraepelinian nosology, does not stand in real contrast to his basic ideas. Kraepelin himself often stressed that all diagnostic criteria and categories were due to change according to the actual state of the art in psychiatric research – leaving, which is essential, his central postulate of the existence of 'natural disease entities' untouched.

'Neo-Kraepelinian' authors are – as was Kraepelin – at risk of overestimating the explanatory power of neurobiological findings, for example, by generally rendering biological data and criteria more reliable and valid than psychopathological or social ones. Nosological prejudices do, of course, exist in psychopathological and social sciences-approaches as well as in neurobiological ones. The essential point is that they always obstruct research, no matter what their theoretical underpinnings may be.

In the context of 'Neo-Kraepelinianism,' the operationalized diagnostic systems *ICD-10* and *DSM-IV* should be considered (APA, 2000; WHO, 1991). Their main impetus is to improve the reliability of psychiatric diagnoses by establishing clear diagnostic criteria and algorithms. They strongly favor descriptive methods by focusing on the behavioral level. Consequently they are skeptical against heuristic methods and try to avoid the interpretation of symptoms wherever possible. In addition, contrary to Kraepelin's nosological system which included a number of etiological hypotheses and, albeit mostly implicitly, epistemological presuppositions, operationalized diagnostic systems avoid etiological considerations, though not fully consequently for example, the ICD-10 groups F0 and F1 carry the main etiological factor in their very name (F0: 'Organic, including symptomatic, mental disorders;' F1: 'Mental and behavioral disorders due to psychoactive substance use'). It is a common misnomer to call this approach 'atheoretical' since no scientific concept can be 'atheoretical,' not in psychiatry and not in any other field. 'Etiologically neutral, where possible' would be much more appropriate. In spite of their theoretical differences, both Kraepelinian nosology and modern diagnostic systems run the risk of uncritically 'reifying' their constructs called 'diagnostic entities' ('reification,' derived from the Latin word res for thing or object ('Sache'), is an epistemological term meaning the identification of a concept with a 'real' entity which exists fully independently from the research and his conceptual framework, e.g., when the concept of schizophrenia is believed to be identical with or, at least, to prove the existence of a objectively (neurobiologically) detectable disease entity). This is why the authors of DSM-IV, for example, warn their users not to regard diagnostic categories as definite or 'natural' entities, not as 'natural kinds,' but as conventions which need further verification – or falsification. Consequently, operationalized diagnostic manuals are continuously modified according to new empirical evidence or conceptual developments (Möller, 2009).

In summary, Kraepelin's psychiatry was not only highly influential in his lifetime, but gained wide acceptance toward the end of the twentieth century, as the focus of psychiatric research again shifted to neurobiological topics. It is of crucial importance that present-day psychiatry – be it 'neo-Kraepelinian' or not – adopts a comprehensive view of Kraepelin's scientific work, embedded in a thorough knowledge of its complex conceptual history (Bennett and Hacker, 2003; Berrios and Hauser, 1988; Honnefelder and Schmidt, 2007; Hoff, 2010, 2012; Noll, 2011; Palm and Möller, 2011; Zivanovic and Nedic, 2012). Otherwise, as well as benefiting from Kraepelin's ideas, psychiatry will also be restricted by their conceptual limitations.

See also: Bipolar (Manic-Depressive) Disorder; Bleuler, Eugen (1857–1939); Dementia: Psychiatric Aspects; Depression; Diagnostic Processes in Clinical Psychology and Psychotherapy; Freud, Sigmund (1856–1939); Janet, Pierre (1859–1947); Normality and Mental Health; Nosology in Psychiatry; Philosophy, Psychiatry, and Psychology; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Schizophrenia; Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

APA, 2000. Diagnostic and Statistical Manual of Mental Disorders, fourth ed., text rev. American Psychiatric Association (APA), Washington, DC.

Bennett, M.R., Hacker, P.M.S., 2003. Philosophical Foundations of Neuroscience. Blackwell Publishing, Malden.

Berrios, G.E., Hauser, R., 1988. The early development of Kraepelin's ideas on classification: a conceptual history. Psychological Medicine 18, 813–821.

Blashfield, R.K., 1984. The Classification of Psychopathology – Neo-Kraepelinian and Quantitative Approaches. Plenum Press, New York.

Chamberlin, E., Gilman, S. (Eds.), 1985. Degeneration. The Dark Side of Progress. Columbia University Press, New York.

Dowbiggin, I., 1985. Degeneration and hereditarianism in French mental medicine 1840–1890 – psychiatric theory as ideological adaptation. In: Bynum, W.F., Porter, R., Shepherd, M. (Eds.), Anatomy of Madness, People and Ideas, vol. I. Tavistock, London, pp. 188–232.

Engstrom, E.J., 1991. Emil Kraepelin: psychiatry and public affairs in Wilhelmine Germany. History of Psychiatry 2, 111–132.

Engstrom, E.J., 1995. Kraepelin – Social Section. In: Berrios, G.E., Porter, R. (Eds.), A History of Clinical Psychiatry. The Origin and History of Psychiatric Disorders. Athlone, London, pp. 292–301.

Gaebel, W., Zielasek, J., 2008. The DSM-V initiative "deconstructing psychosis" in the context of Kraepelin's concept on nosology. European Archives of Psychiatry and Clinical Neuroscience 258 (Suppl. 2), 41–47.

Hermle, L., 1986. Die Degenerationslehre in der Psychiatrie. Fortschritte der Neurologie und Psychiatrie 54, 69–79.

Hoff, P., 1994. Emil Kraepelin und die Psychiatrie als klinische Wissenschaft. Ein Beitrag zum Selbstverständnis psychiatrischer Forschung. Springer, Berlin.

Hoff, P., 1995. Kraepelin – Clinical Section. In: Berrios, G.E., Porter, R. (Eds.), A History of Clinical Psychiatry. The Origin and History of Psychiatric Disorders. Athlone, London, pp. 261–279.

Hoff, P., 1998. Emil Kraepelin and forensic psychiatry. International Journal of Law and Psychiatry 21, 343–353.

Hoff, P., 2008. Kraepelin and degeneration theory. European Archives of Psychiatry and Clinical Neuroscience 258 (Suppl. 2), 12–17.

Hoff, P., 2010. Sicht der Klassiker: der Affekt in den Psychosekonzepten von Emil Kraepelin und Eugen Bleuler. Psychiatrie & Psychotherapie 6, 98–103.

Hoff, P., 2012. Psychopathologie und die Identität des Faches Psychiatrie. In: Böker, H., Seifritz, E. (Eds.), Psychotherapie und Neurowissenschaften. Integration, Kritik, Zukunftsaussichten. Huber, Bern, pp. 567–580.

Honnefelder, L., Schmidt, M.C., (Hrsg.) 2007. Naturalismus als Paradigma. Wie weit reicht die naturwissenschaftliche Erklärung des Menschen? University Press, Berlin.

Kraepelin, E., 1883–1927. Psychiatrie. Nine editions: 1883, Leipzig: Abel; 1887, Leipzig: 1887, 1889 Leipzig: Abel; 1893, Leipzig: Abel; 1896, Leipzig: Barth; 1899 (2 vols.), Leipzig: Barth; 1903/04 (2 vols.), Leipzig: Barth; 1909–15 (4 vols.), Leipzig: Barth; 1927 (2 vols.). Barth, Leipzig.

Kraepelin, E., 1918. Ziele und Wege der psychiatrischen Forschung. Zeitschrift für die gesamte Neurologie und Psychiatrie 42, 169–205.

Kraepelin, E., 1919. Die erforschung psychischer krankheitsformen. Zeitschrift für die gesamte Neurologie und Psychiatrie 51, 224–246.

Kraepelin, E., 1920. Die Erscheinungsformen des Irreseins. Zeitschrift für die gesamte Neurologie und Psychiatrie 62, 1–29.

Kraepelin, E., 1983. In: Hippius, H., Peters, G., Ploog, D. (Eds.), Lebenserinnerungen. Springer with the collaboration of P. Hoff and A. Kreuter, Heidelberg, Berlin, New York. English translation (1987): Memoirs. Springer, Berlin.

Liegeois, A., 1991. Hidden philosophy and theology in Morel's theory of degeneration and nosology. History of Psychiatry 2, 419–427.

Möller, H.J., 2009. Development of DSM-V and ICD-11: tendencies and potential of new classifications in psychiatry at the current state of knowledge. Psychiatry and Clinical Neurosciences 63, 595–612.

Noll, R., 2011. American Madness: The Rise and Fall of Dementia Praecox. Harvard University Press, Cambridge.

Palm, U., Möller, H.J., 2011. Reception of Kraepelin's ideas 1900–1960. Psychiatry and Clinical Neurosciences 65, 318–325.

Pick, D., 1989. Faces of Degeneration: A European Disorder 1848–1918. Cambridge University Press, Cambridge.

Read, J., Mosher, L.R., Bentall, R.P. (Eds.), 2004. Models of Madness. Psychological, Social and Biological Approaches to Schizophrenia. Brunner-Routledge, Hove.

Shepherd, M., 1995a. Two faces of Emil Kraepelin. British Journal of Psychiatry 167, 174–183.

Shepherd, M., 1995b. Kraepelin and modern psychiatry. European Archives of Psychiatry and Clinical Neuroscience 245, 289–295.

Van Os, J., Tamminga, C., 2007. Deconstructing psychosis. Schizophrenia Bulletin 33, 861–862.

Van Praag, H.M., Kahn, R.S., Asnis, G.M., Wetzler, S., Brown, S.L., Bleich, A., Korn, M.L., 1987. Denosologization of biological psychiatry or the specificity of 5-HT disturbances in psychiatric disorders. Journal of Affective Disorders 4, 173–193.

Verwey, G., 1985. Psychiatry in an Anthropological and Biomedical Context – Philosophical Presuppositions and Implications of German Psychiatry 1820–1870. Reidel, Dordrecht.

Weber, M.M., 1991. Ein Forschungsinstitut für Psychiatrie. Die Entwicklung der Deutschen Forschungsanstalt für Psychiatrie München 1917–1945. Sudhoffs Archiv 75, 74–89.

World Health Organization, 1991. Tenth Revision of the International Classification of Diseases, Chapter V (F): Mental and Behavioural Disorders (Including Disorders of Psychological Development). Clinical Descriptions and Diagnostic Guidelines. WHO, Geneva.

Zivanovic, O., Nedic, A., 2012. Kraepelin's concept of manic-depressive insanity: one hundred years later. Journal of Affective Disorders 137, 15–24.

Zubin, J., Oppenheimer, G., Neugebauer, G., 1985. Degeneration theory and the stigma of schizophrenia (Editorial). Biological Psychiatry 20, 1145–1148.

Kuhn, Thomas S. (1922–96)

Paul Hoyningen-Huene and Simon Lohse, Center for Philosophy and Ethics of Science, Institute of Philosophy, Leibniz University of Hannover, Hannover, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P. Hoyningen-Huene, volume 12, pp. 8171–8176, © 2001, Elsevier Ltd.

Abstract

The first section describes Kuhn's life and works. The second discusses Kuhn's theory of scientific development. This theory consists of a developmental scheme for the basic natural sciences that comprises the consecutive phases of prenormal, normal, and extraordinary science. The major differences between Kuhn's theory and the antecedent tradition are discussed, most notably Kuhn's insistence that scientific communities are the principal actors of scientific development. In the last section, Kuhn's influence on the social and behavioral sciences is discussed and assessed with special attention to his deep influence on the sociology of scientific knowledge.

Life and Works

Thomas S. Kuhn was born 18 July 1922 in Cincinnati, Ohio. He grew up in the State of New York and enrolled at Harvard University in 1940, where he majored in physics. Additionally, he attended philosophy courses and was particularly impressed by the theoretical philosophy of Immanuel Kant. In 1943, Kuhn earned a Bachelor's degree in physics from Harvard. During his graduate work in theoretical solid-state physics with John H. van Vleck (who later won a Nobel prize), James Conant, then president of Harvard, introduced him to the history of science. In 1947, while preparing a course on the development of mechanics prior to Galileo and puzzling over passages of Aristotle's physics, he discovered something that set the agenda of his further work. To understand the development of mechanics, Kuhn tried to find out which elements of modern mechanics were traceable back to Aristotle and which elements were first introduced in the sixteenth and seventeenth centuries. However, when Kuhn studied Aristotle with contemporary physical concepts in mind, much of Aristotle's physics, his conception of motion (kinēsis) in particular, appeared to be either false or even outright meaningless. Apparently, Galileo practically needed to start from scratch. However, this picture of Aristotelian physics, though prevalent in the first half of the twentieth century, seemed somewhat peculiar: Why would one of the greatest logicians of all time, an exceptional philosopher and otherwise perceptive scientist, who is widely regarded as one of the founders of biology, have been almost completely wrong in respect to physics? And why was Aristotelian physics, nevertheless, canonical for almost two millennia? These inconsistencies did not resolve until Kuhn managed to develop a different reading of Aristotle. He avoided the framework of modern physical concepts and tried to understand Aristotelian physics from an internal perspective. In other words, Kuhn developed a hermeneutical approach to the original texts which avoided semantic distortions through the import of modern scientific concepts as far as possible. By changing the ascribed meaning of some of the Aristotelian key concepts (in some cases only slightly and in others more intensively), a whole new world of physics became accessible, namely that of Aristotle and his contemporaries. In retrospect Kuhn described this episode as formative for his entire career in history and philosophy of science (Kuhn, 2000: pt. 3).

After completing his PhD in theoretical physics in 1949, Kuhn turned entirely to the history of science (motivated by a primary interest in philosophical questions though) where he began applying the insights he gained through his experience in Aristotelian physics. From 1948 to 1951, he was a Junior Fellow of the Harvard Society of Fellows where he met Willard Van Orman Quine, a then senior fellow in the society, and Karl R. Popper, who gave the James Lectures at Harvard in 1950. In this period Kuhn read gestalt psychology and Jean Piaget's work as well as Benjamin Lee Whorfs linguistic ideas, and he became acquainted with Ludwig Fleck's *Entstehung und Entwicklung einer wissenschaftlichen Tatsache* (*Genesis and Development of a Scientific Fact*), all of which promoted the development of his own ideas (Kuhn, 1962: preface; Kuhn, 2000: pt. 3). From 1951 to 1956 Kuhn was Assistant Professor for General Education and History of Science at Harvard. From 1956 to 1964, he was a member of the faculty of the University of California at Berkeley, mainly in the History of Science Department. During his years at Berkeley, academic discussion with Wittgenstein scholar Stanley Cavell, whom he already had met at Harvard, had a crucial impact on the development of Kuhn's nascent main work *The Structure of Scientific Revolutions* (*SSR*). In 1957, Kuhn published *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought*, which became a very successful textbook. Kuhn spent the year 1958–59 as a Fellow at the Center for Advanced Study of the Behavioral Sciences in Stanford where he became aware of the different patterns typical of the interactions among natural scientists on the one hand, and among social scientists on the other. The difference typically concerns a broad consensus of natural scientists with respect to the foundations of their discipline, in contrast to social scientists who typically dispute them. This observation became very important when Kuhn wrote *SSR*. In 1960/1961, Kuhn finished the first draft of *SSR* and received important feedback from philosophers such as Paul Feyerabend (cf Hoyningen-Huene, 1995, 2006) and Ernest Nagel as well as physicist Pierre Noyes and historian John Heilbron, who was Kuhn's graduate student at that time. The completed book was published in 1962. *SSR* is the most widely read book ever

published in the history and philosophy of science, and it also had a tremendous impact in the social and behavioral sciences, most of all in the sociology of scientific knowledge. By now, more than 1 million copies of the book have been sold in English, and it has been translated into at least 26 languages. Its key term, paradigm, is now a household word. Some of its other central concepts like normal science, paradigm shift, or incommensurability are now commonly used in a large number of diverse disciplines. However, many of the interpretations of the book (or its parts) were vigorously rejected by Kuhn as deep misunderstandings of what he intended to say. In 1964, Kuhn joined Princeton University as a Professor of History of Science; he was also a member of the Institute of Advanced Study from 1972 to 1979. He met Carl Gustav Hempel, who was more open minded with respect to Kuhn's ideas than most contemporary philosophers, and became an important discussion partner for Kuhn during his time at Princeton.

Kuhn's historical work after *SSR* mainly concerned the genesis of quantum mechanics. *Sources for the History of Quantum Physics* (Kuhn et al., 1967) contains an inventory of relevant material including interviews with many of the contributors to quantum mechanics. *Black Body Theory and the Quantum Discontinuity, 1894–1912* (Kuhn, 1978) is a controversial narrative of the introduction of the quantum hypothesis. Some of Kuhn's papers both on the history and the philosophy of science are contained in his *The Essential Tension*, published in 1977. In 1979, he became Professor for Philosophy and History of Science at the Massachusetts Institute of Technology until he retired in 1991. From the 1980s on, Kuhn worked on a book whose final working title was *The Plurality of Worlds: An Evolutionary Theory of Scientific Development*. In this book, he tried to elucidate the foundations of scientific change through a new theory of meaning of empirical concepts, especially of taxonomic kind terms. Furthermore, he elaborated on the evolutionary metaphor for scientific development with which *SSR*'s last Chapter ended. Kuhn could not finish the book due to his death on 17 June 1996 in Cambridge, MA.

In 2000, *The Road since Structure* was published, which contains some of Kuhn's substantial philosophical papers between 1970 and 1993 and an extensive autobiographical interview from 1995. In summary, it can be stated that Kuhn is a distinguished historian of science and one of the most influential philosophers of science in the twentieth century. Even in 2007, 45 years after the publication of *SSR*, Kuhn belongs to the 40 most cited authors of books in the humanities and the social sciences (THE, 2009).

Kuhn's Theory of Scientific Development

Kuhn's main subject in the philosophy of science was a developmental scheme, or a schematic description, of scientific development that primarily concerned the basic (i.e., curiosity-driven) natural sciences. The intended description is schematic in the sense that it presents scientific development as a succession of different phases whose main features are independent of the specific subject matter of the respective science. For each one of these phases a specific mode of scientific practice is characteristic. This theory was first presented in *SSR* but was later refined in various respects.

The developmental scheme that Kuhn suggests should hold true with only minor variations for all basic scientific disciplines. It consists of the following: Before reaching maturity, nascent scientific fields are typically characterized by controversies between competing schools. Consensus lacks among the practitioners of the emerging field. Each one of these schools has a particular view of the respective research domain, and typically these views derive from extrascientific sources. This phase of scientific development can be referred to as 'prenormal.'

The competition among the schools could come to an end when one group produces an exemplary solution to a preeminent research problem with two characteristics: it is sufficiently unprecedented to attract the members of the other schools, and it is sufficiently open ended to leave enough interesting problems for further scientific work. These model solutions, for example, Lavoisier's oxygen theory or Maxwell's electrodynamics, are referred to as paradigms. They serve to guide research implicitly in the succeeding period called 'normal science.' When a particular science has reached its first normal science phase it is said to have matured.

Normal science is characterized by a broad consensus of the practitioners of the field on fundamental questions, and consequently, by a particular mode of research. This mode of research can be described by a five-dimensional analogy to puzzle solving where exemplars of puzzles include chess problems and crossword puzzles. The analogy concerns

1. The existence of regulations constraining acceptable approaches to and solutions of problems. In the case of science, these regulations are given only implicitly, namely as part and parcel of the paradigmatic solutions.
2. The expectation of the solubility of appropriately chosen problems. In the case of science, this expectation is fed by the conviction that the appropriateness of the paradigmatic solutions extends beyond the particular problems they are solutions of.
3. Missing intentions of fundamental innovation of the guiding regulations. In the case of science, normal science intends to exploit the cognitive potential intrinsic to the paradigmatic solutions, and does not intend to overturn them.
4. Absence of test or confirmation of the guiding regulations. In the case of science, this is due to the fact that in normal science, failure to solve a research problem is seen as a failure of the researcher and not of the regulations that guide research. Either the research problem was not chosen well or the researcher was unable to solve it despite its solubility.
5. Individual motivations: to prove oneself an expert puzzle solver. In the case of science, this is evidenced by the extremely elaborate system by which scientific achievements are rewarded (e.g., scientific prizes including the Nobel Prize, honorary degrees, invited lectures, nominations).

Normal science is always confronted with anomalies, that is, with phenomena or problems that behave contrary to the expectations supplied by the paradigm. Anomalies do not usually call the validity of the guiding regulations of normal

research into question. But under special circumstances they may, and then they become 'significant anomalies.' In such cases, the practice of science changes into 'extraordinary science' or 'science in crisis.' It aims at amending or even overthrowing the still binding regulations. Its research focuses on the significant anomalies and their context. Extraordinary science resembles prenormal science in that it tends to develop competing schools. However, it is more focused than prenormal science as all these different schools focus on the same set of significant anomalies and they all thrive to retain as much as possible from the earlier period of normal science. There are four possible ways in which extraordinary science can come to an end. (1) Eventually, it is possible to resolve the crisis triggering anomalies within the existing regulations. (2) The anomalies in question are disregarded if possible, that is, if the continuation of normal science is possible without interference. (3) Similarly to a speciation process in biology, a new scientific subfield is created in which the formerly troublesome cases are treated (Wray, 2011). (4) Research leads to a new paradigm that is accepted by the scientific community because it can lead to a new phase of normal science: the occurrence of a scientific revolution. Scientific revolutions in Kuhn's sense are thus "the tradition-shattering complements to the tradition–bound activity of normal science" (Kuhn, 1962: 6). Kuhn describes the features of scientific revolutions by characterizing the relation between successive traditions of normal science as 'incommensurable.'

In *SSR*, this incommensurability has three different aspects: First, through a scientific revolution, a change in the field of scientific problems that must be addressed by any theory of that domain occurs as well as a change in the field of problems that are legitimately addressed. Problems whose solutions are vitally important to the older tradition may be abandoned as obsolete or even unscientific; problems that did not exist, or whose solution was considered trivial, may gain extraordinary significance in the new tradition. Along with the problems, quite often the standards imposed upon scientifically admissible solutions change. For instance, in classical physics fundamental theories are required to be deterministic; this requirement vanishes with the quantum-mechanical revolution. A second aspect of Kuhn's (1962) incommensurability concept concerns scientific methods and concepts. After a revolution, many of the older concepts and methods are still used, but in slightly modified ways. The change of concepts discussed in *SSR* is of prime importance. It has an extensional and an intensional aspect. The extensional aspect consists of the movement of objects belonging to the extension of one concept into the extension of another concept, the two concepts being mutually exclusive. The striking example that Kuhn uses in *SSR* is the change of the concept of a planet in the Copernican revolution. After the revolution, the earth, for instance, is a planet, whereas the sun and the moon are not planets anymore. The intensional aspect of the change of concepts consists of a change of meaning of the respective concepts. This is the case since the properties of the objects that are subsumed under these concepts change. As the third and most fundamental aspect of incommensurability, Kuhn claims in *SSR* that scientists belonging to different paradigms conduct their research in different worlds. But what does it mean to say that the world changes with a revolution? This question is not answered in *SSR*, although Kuhn has clearly seen its urgency there.

To understand this aspect of Kuhn's theory, it is important to distinguish between two meanings of the term 'world' in Kuhn's philosophy, namely the phenomenal world and the world in-itself. It is the phenomenological world to which we actually have access in everyday life or in science. We can perceive and describe such a world, and in such a world there are octopuses, lecture halls, and electrons, for example. Such a world has a certain conceptual structure, for instance the aforementioned categories. Kuhn supposes that these concepts are of human origin, that is, we impose a structure on the world by means of these concepts, instead of just reading off these concepts from the world itself, as a more familiar (realist) story wants us to believe. Although it is not possible to impose any and every structure on the world, clearly more than just one is possible, which implies a certain degree of historical contingency of these structures. This is illustrated by the historical change of these conceptual structures that are embedded in paradigms. As a consequence, paradigms are constitutive of a perceptually and conceptually subdivided world. In other words, the subjects of knowledge contribute to the constitution of the objects of knowledge (by means of paradigms) insofar as they structure the world of these objects.

The second sense of the term 'world' in *SSR* is obtained by asking what is left if one subtracts all these human (subject-sided) contributions, all this perceptual and conceptual structuring from the world in the first sense. Then, one is left with a world that is completely independent from our perceptions and conceptions, a world – as one might say – that is purely object-sided. But we have, according to Kuhn, no access whatsoever to this purely object-sided world. According to Kuhn, what we, or a particular scientific community, describes as reality is a – but not the only possible – phenomenal world. Due to the historical change of these worlds, Kuhn's theory has sometimes been described as 'Kant on wheels.' Note, however, that numerous philosophers of science are extremely skeptical about these apparently idealist elements in Kuhn's philosophy and reject the cogency of this line of argument (e.g., Godfrey-Smith, 2003; Preston, 2008).

Due to incommensurability, we must, according to Kuhn, rethink the concept of scientific progress in the natural sciences. First, scientific progress is not cumulative, due to conceptual changes during revolutions. Cumulativity implies that something that is a part of science at some point in time will, neglecting small corrections, remain a part of science forever. However, the conceptual changes that occur during revolutions are much more than slight corrections of the existing body of knowledge. They amount to a sometimes far-reaching conceptual reorganization, and typically to a changed ontological perspective. This holds true in spite of the fact that very often much of the older knowledge is retained, especially for practical applications, and that, in the case of quantitative theories, limiting relations exist between the older and the newer theories. Furthermore, Kuhn denies that scientific progress is an approach to truth. Instead of conceiving scientific progress as a teleological process, that is, one that is goal directed, we should think of scientific

progress in an analogous way as Darwinian evolutionary theory conceives of evolution (Reydon and Hoyningen-Huene, 2010). Darwinian evolutionary theory states that there is no goal of evolution toward which it is directed. In a similar way, in scientific development there is no 'set goal' which would be 'a permanent fixed scientific truth' that science approaches (Kuhn, 1962: 173). However, Kuhn stresses that this does not imply at all that there is no progress in the sciences. There is progress in the sciences, though not in the form of an approximation to truth. Rather, there is progress in the form of 'an increase in articulation and specialization' of scientific knowledge (Kuhn, 1962: 172). Therefore, the widespread characterization of Kuhn's theory as relativist is simply false.

There are several breaks between Kuhn's theory and the previous tradition in the philosophy of science, some of which explain the extraordinary resonance his theory found in the social and behavioral sciences. Seven of them will be discussed briefly. The first is Kuhn's altered view of scientific progress that has just been given. Second, this view implies the untenability of forms of realism which assert that science at least approximately describes what is really 'out there,' independently of any observer. Instead, theories describe the world in terms of concepts that are historically contingent (but not arbitrary!) and which may change in the future. Third, because of this change of basic scientific concepts through revolutions, the classical conception of reductionism is also hardly tenable. In this classical conception, the redefinition of the concepts of the reduced theory by means of the concepts of the reducing theory played a key role. But if incommensurability prevails, some of the required redefinitions are impossible because of meaning shifts. Fourth, many of *SSR*'s assertions stood in marked opposition to Popper's critical rationalism. For Popper, good science depends on the constant questioning of all issues in a scientific field. Kuhn's normal science, as seen from Popper's perspective, is therefore bad science since it is not continuously directed toward critically testing the guiding regulations. Instead, some anomalies are not even pursued, but are ignored or even pushed aside. Alternatives to the paradigmatic framework are also typically ignored. Yet in Kuhn's view, critical evaluation of fundamental theories is restricted to the periods of extraordinary science, and even then, scientific practice is not simply an attempt to falsify theories by confronting them with observational statements about nature. Rather, theory evaluation is a comparative procedure in which at least two theories are assessed with respect to their scientific potential. Thus, according to Kuhn, theory falsification as described by Popper is a stereotype that does not occur in the actual history of science. A fifth break results from Kuhn's concern for the actual history of science and its assigned role in philosophy of science. Kuhn argues that the abovementioned historical–hermeneutical approach to the history of science can provide us with important data that a reasonable philosophy of science should incorporate. This is frequently referred to as the 'historical turn in philosophy of science.' Pre-Kuhnian philosophy of science, in contrast, unknowingly depends on a stereotypic picture of scientific development, which has its origin in a presentist interpretation of past scientific theories via modern scientific concepts. A sixth consequence of Kuhn's theory is the abolishment of the idea that science is determined by logic and observational data alone, or by the scientific method, as construed as a set of rigorous rules. This idea has dominated the understanding of modern science from its very beginning. According to Kuhn, however, exemplary problem solutions guide scientific research in its normal phase. Their cognitive potential for research is not exploited by explicit (or fully explicable) rules, but rather by implicit analogies. New problems are identified in the light of solved ones, and new solutions are judged as legitimate in a similar manner.

Finally, and most importantly in our context, the principal agents of science according to Kuhn are communities, not individuals. In the philosophy of science tradition before Kuhn, no one questioned the identity of the principle agent of science. Though armed with the distinction between individuals and communities, it is obvious that in the older tradition, the individual scientist was seen as the principal actor. The possibility of two scientists rationally disagreeing was not permitted, nor were there discussions about the gradual formation of scientific consensus out of disagreement, or of the development of disagreement from a previously established consensus. In other words, scientific communities were barely present in this discussion. Science was primarily seen as a one-person game, but for Kuhn it is, at its heart, a social enterprise. It is communities who are the ultimate evaluators of knowledge claims. It is communities who ultimately decide between competing theories and their associated modes of research, and it is communities who are seen by themselves and by others as responsible for some domain of scientific knowledge. Kuhn, therefore, repeatedly refers to the 'sociological base of (his) position' (for references, see Hoyningen-Huene, 1993), but at the same time, he insists that this fact does not invalidate the essential epistemological component of his theory, that is, its normative element. How is this to be understood?

Kuhn characterizes scientific communities by the values they hold. Typically, these values comprise accuracy, scope, consistency, fruitfulness, and explanatory power, among others. These values are instrumental in the evaluation of the application of theories and, partly derivatively, in the evaluation of the theories themselves. However, these communal values are not so strictly defined that each member of the community reaches exactly the same evaluation in a concrete case of application. Individuals may differ in the relative weight they attribute to these values as well as in the specific articulation of any one of them. Yet in normal science, these differences typically do not become apparent; they remain latent. However, with the advent of significant anomalies, these latent differences between scientists belonging to the same community may become manifest due to different evaluations of the anomalies, resulting in dissent. The dissent may continue with respect to the different candidates proposed to resolve the crisis. A new consensus about some theory and its ability to solve outstanding problems can only emerge if, and when, (almost) all members of the community prefer this theory over its competitors, although their individual differences with respect to the values have not fully disappeared. Thus, during crisis resolution, the individual differences of the members of the community become insignificant: it is the communal core of the values that prompts the choice of the new paradigmatic theory. Thus, the choice of the new theory is essentially an event that must be described in sociological terms,

as opposed to psychological ones. However, the normative dimension of theory choice has not disappeared. The reasons for the choice may still be normatively evaluated: scientific values are an essential part of the reasons for choosing some theory. They are good reasons if they promote the aims of science, that is, the comprehensive and systematic description and explanation of nature. This does apply to the values mentioned above (i.e., accuracy, fruitfulness, consistency, and the like) (Hoyningen-Huene, 1992).

Kuhn's Impact on the Behavioral and Social Sciences

Kuhn's work is extremely widely received in the social and behavioral sciences. Before Kuhn, mainstream social science adopted a picture of the natural sciences that was based mainly on logical empiricism, and many social scientists attempted to assimilate their disciplines to that picture. However, success was limited. Neither the predictive nor the explanatory power of a natural science like physics could be achieved. In addition, methodological controversy did not cease to exist even if in some field a particular school became dominant for some period. In this situation, the alternative view of the natural sciences that Kuhn offered also seemed to suggest, as a corollary, a different view of the social sciences. More specifically, as Kuhn incorporated elements in his description of the natural sciences which had been barred by the earlier tradition in the name of the objectivity of science, such as group processes and the like, the difference between the natural and the social sciences seemed to have decreased. In this way, the already comparatively high interest that social scientists have in the philosophy of science, given the epistemologically insecure status of their disciplines, became even more intense in the case of Kuhn's philosophy. At least to some authors, Kuhn seemed to suggest a new understanding of the social sciences.

Three general facts about the reception of Kuhn's work in the social and behavioral sciences are striking. First, the reception was by and large uncritical, and in many cases very superficial. For example, there are literally hundreds of articles in which reference to Kuhn appears to be a purely rhetorical device, making no substantial point. Second, Kuhn's developmental model was intended to apply only to the basic disciplines of the natural sciences. Kuhn says very little about the social sciences, basically only that they are typically in the preparadigmatic state (Kuhn, 1962, 1977). This did not prevent many social and behavioral scientists from applying parts of Kuhn's theory to different aspects of their disciplines. Third, Kuhn's ideas have been used for the most diverse purposes. On the basis of his work, even outright contradictory assertions have been made. This is partly due to the fact that Kuhn's writings, especially *SSR*, have resulted in several very different interpretations, which derives from the inner tensions, ambiguities, and metaphors that his works contains. In particular, Kuhn's evaluation of the rationality of science appeared as highly ambivalent, and his use of the paradigm concept seemed at times to cover just about everything that scientists deal with. Thus, Kuhn's texts appear to have a high degree of plasticity, a fact that has been subsequently deplored by its author. Only a very close and patient reading of the texts can reduce this plasticity. On top of the plasticity of Kuhn's texts, social scientists have been seduced to read many things into Kuhn's work due to their own agendas. Thus, many of the claims in the literature about what Kuhn does or does not support are not actually contained in Kuhn's texts.

The predominant use of Kuhn's work in the social and behavioral sciences has concerned the current status of some discipline or its history. Is the discipline still in its prenormal state or does it already have a paradigm? In other words, is it a 'mature' science that has reached the state of normal science? Has it had paradigms in the past? Are the changes that the discipline experienced in the past Kuhnian revolutions? Could competing theories, say in sociology or political science, be described as incommensurable? Or are the social sciences fundamentally different from the natural sciences in that they are 'multiple paradigm sciences?' As mentioned above, these questions received the most diverse answers by different authors. Elements from Kuhn's theory were used by some to defend mainstream social science, and by others to develop a critical attitude toward it, or to defend or attack alternatives. Some authors tried to show that on Kuhnian terms, a given social science closely resembles a natural science, whereas others tried to demonstrate the opposite, and so on. No consensus has been reached in any of the social or behavioral sciences about these and other related questions.

Regarding the history of science, there is, perhaps surprisingly, not much of a reception of Kuhn's *meta*-historical work. This may be partly due to the same reasons that Kuhn did not apply his characteristic concepts (like paradigm, normal science) to his own historical work. Historians often try to avoid heavily theory-laden concepts when doing historical work due to the imminent distortions of what can be obtained from the sources themselves. In addition, historians in the second half of the twentieth century have typically been quite critical about any generalizations on the course of history (such as a developmental scheme).

In the sociology of scientific knowledge, however, Kuhn's impact has been tremendous. Pre-Kuhnian sociology of science either discussed the social structure of science or the impact of science upon society. The scientific content seemed to lie entirely outside the scope of sociology because of the purported objectivity of science, which precluded any direct social contingency. The most relevant sociological question asked with respect to scientific knowledge was 'What are the social conditions of a scientific community that are conducive to the production of such knowledge?' Answers to this question typically consisted of the description of a purportedly universal value system which holds true in scientific communities. This value system was seen as virtually guaranteeing the desired outcome of the activities of the community's members. However, with the advent of Kuhn's work, this view was challenged in various respects. The production of scientific knowledge now seemed not at all determined by time-independent, universal methodological rules. Instead of binding rules, research is guided by paradigms that include values that only influence, but do not determine scientific behavior. In addition, individual members of the community differ in their specific assimilation of the communal values and their relative weights. In other words, scientific knowledge now seemed socially contingent which licenses and even appears to demand sociological analysis. Various schools developed

analytical tools for this purpose, and a great many case studies of contemporary and past science were performed in the emerging field of Science and Technology Studies (STS). Typically, the content of science was now seen as the result of negotiations among scientists in which human interests of a personal or political nature, or social positions and power relations within scientific communities, played an important, and sometimes even decisive, role. Along these social constructivist lines, scientific knowledge can no longer be seen as objective in the traditional sense of the term. Due to an intense debate within sociology and interventions mainly from philosophers, these positions showed rapid development that continues until today. It should be noted, however, that despite all of the changes in these positions, Kuhn himself remained extremely skeptical of them. In his view, the epistemological dimension of scientific knowledge, that is, the possibility of its normative evaluation from an internal perspective, must be an essential part of an integrated image of science. It is only in this manner that the object-sided aspects of experience, which constrain the possibility space for scientific theories, can be properly taken into consideration. This aspect of Kuhn's theory had been entirely dismissed by many sociologists of scientific knowledge (Kuhn, 1991, 1992).

See also: Critical Rationalism; Logical Positivism and Logical Empiricism; Objectivity: Philosophical Aspects; Relativism: Philosophical Aspects; Science, History of; Social Constructivism.

Bibliography

Godfrey-Smith, P., 2003. Theory and Reality: An Introduction to the Philosophy of Science. University of Chicago Press, Chicago.

Hoyningen-Huene, P., 1992. The interrelations between the philosophy, history and sociology of science in Thomas Kuhn's theory of scientific development. British Journal for the Philosophy of Science 43, 487–501.

Hoyningen-Huene, P., 1993. Reconstructing Scientific Revolutions. Thomas S. Kuhn's Philosophy of Science (A.T. Levine, Trans.). University of Chicago Press, Chicago.

Hoyningen-Huene, P., 1995. Two letters by Paul Feyerabend to Thomas S. Kuhn on a draft of the structure of scientific revolutions. Studies in History and Philosophy of Science 26, 353–387.

Hoyningen-Huene, P., 2006. More letters by Paul Feyerabend to Thomas S. Kuhn on proto-structure. Studies in History and Philosophy of Science 37, 610–632.

Kuhn, T.S., 1957. The Copernican Revolution: Planetary Astronomy in the Development of Western Thought. Harvard University Press, Cambridge, MA.

Kuhn, T.S., 1962[1970]. The Structure of Scientific Revolutions, second ed. University of Chicago Press, Chicago.

Kuhn, T.S., 1977. The Essential Tension: Selected Studies in Scientific Tradition and Change. University of Chicago Press, Chicago.

Kuhn, T.S., 1978. Black Body Theory and the Quantum Discontinuity, 1894–1912. Clarendon Press, Oxford, UK.

Kuhn, T.S., 1991. The road since structure. In: Fine, A., Forbes, M., Wessels, L. (Eds.), Proceedings of the 1990 Biennial Meeting of the Philosophy of Science Association. PSA 1990, vol. 2. Philosophy of Science Association, East Lansing, MI, pp. 3–13.

Kuhn, T.S., 1992. The Trouble with the Historical Philosophy of Science. Robert and Maurine Rothschild Distinguished Lecture, November 19, 1991, An Occasional Publication of the Department of the History of Science. Harvard University Press, Cambridge, MA.

Kuhn, T.S., 2000. The Road since Structure. Philosophical Essays, 1970–1993, with an Autobiographical Interview, (Conant, J., Hangeland, L., Ed.). University of Chicago Press, Chicago.

Kuhn, T.S., Heilbron, J.L., Forman, P., Allen, L., 1967. Sources for the History of Quantum Physics. An Inventory and Report. American Philosophical Society, Philadelphia, PA.

Lakatos, I., Musgrave, A. (Eds.), 1970. Criticism and the Growth of Knowledge. Cambridge University Press, London.

Preston, J., 2008. Kuhn's "The Structure of Scientific Revolutions": A Reader's Guide. Continuum, London and New York.

Reydon, T.A.C., Hoyningen-Huene, P., 2010. Discussion: Kuhn's evolutionary analogy in *the Structure of scientific revolutions* and "the road since structure". Philosophy of Science 77, 468–476.

Wray, K.B., 2011. Kuhn's Evolutionary Social Epistemology. Cambridge University Press, Cambridge.

Relevant Website

http://www.timeshighereducation.co.uk/ (28.02.12.) – THE (Times Higher Education), 2009. Most Cited Authors of Books in the Humanities, 2007.

Kula Ring, Anthropology of

Frederick H Damon, University of Virginia, Charlottesville, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by N.D. Munn, volume 12, pp. 8176–8179, © 2001, Elsevier Ltd.

Abstract

Approaching the centennial of Malinowski's initial fieldwork and writing about the kula ring in southeastern Papua New Guinea, this article reviews the history of research on the institution, describes its fundamental forms and outlines some of the questions that have been and will be addressed to the institution as a new generation of scholars begin to redefine this classic region in the history of anthropology.

First set out in detail in Malinowski's classic *Argonauts of the Western Pacific* (1922), the 'kula ring' – a translocal, long distance exchange network connecting islands off the southeast coast of Papua New Guinea (**Figure 1**) – remains a classic ethnographic reference point for the study of non-Western exchange practices and the development of exchange theory. Malinowski's work was followed by more limited contributions by Fortune (1932), Roheim (1950), and Belshaw (1955), but it was not until the late 1960s and 1970s when new field research yielding significant publications on kula was carried out (see Leach and Leach, 1983). Some of those first returnees continue to publish on the institution (e.g., Damon, 2002), and a new set of scholars are examining the region with fresh eyes (beginning with Kuehling, 2005). Archaeological research provisionally locates the region in the larger context of the human occupation of the Pacific in general and the Austronesian expansion in particular (Bickler, 2006; Ambrose et al., 2012). Today the region is once again being subjected to Western-inspired resource extraction and incursion analogous to that which occurred during the transition from the nineteenth to the twentieth century, yet local people now occupy most modern state offices, diesel boats and

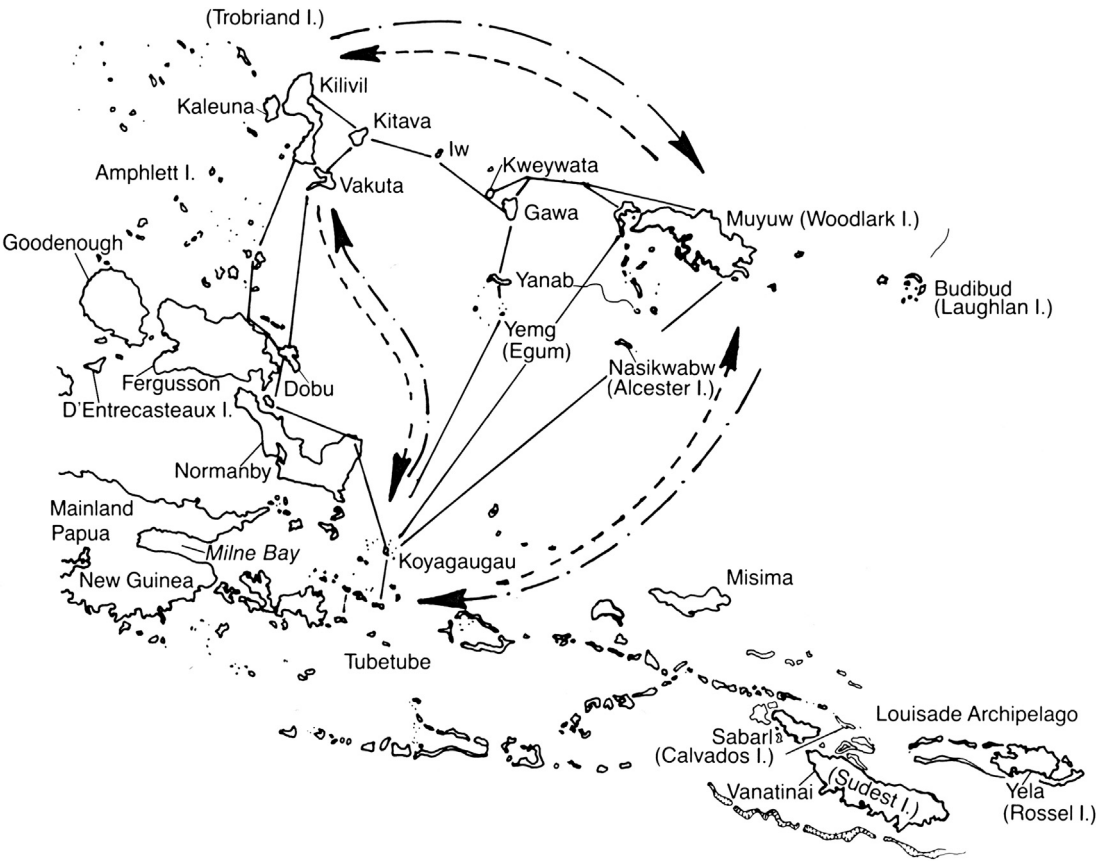

Figure 1 The kula ring. Standard English names are given in parentheses.

fiberglass dinghies motored with outboards facilitate much of the interisland communication, and most islands have vastly expanded populations. However, the kula institution remains robust, absorbing as much of the new as it is being absorbed by it. This article explains the key features of kula, updating Malinowski's account where necessary on the basis of the more recent work and modifying some common oversimplifications; in conclusion, it comments briefly on anthropological approaches to kula.

Some Defining Features of Present-Day Kula

The exchange of white (*Conus*) armshells and red (*Spondylus* or *Chama*)-shell necklaces in opposite directions around the islands centers the modern kula exchange (**Figure 1**). The valuables are gendered, all are qualitatively graded but only middle and upper ranked shells have personal names. They circulate in formal relationships denoted by a word which may be translated as 'path,' 'way,' or even 'meaning,' ked/keda/eda and cognates. There are other kinds of kula media (see Section Adjustment of Some Common Oversimplifications of Kula), one of which, pig tusks, no longer circulates but for some reason became collected by the hundreds on Goodenough Island. However, only the two shell types currently circle the region establishing its basic sociogeography. Kula transactors, however, should exchange directly in a limited sector with partners or potential partners in participant communities on either side of them (although illicit bypassing occurs). Only occasionally and usually just with the very largest valuables do people know exactly how specific transects close. The problem the institution addresses is the rank of its transactors. Their primary aim is to increase their ranks by means of the ranked shell exchanges. Rank is usually stated as the height or visibility of a person's name. Transactors obtain shells of one category on visits to partners in one direction and later give them to their own visitors from the opposite kula direction. Typically, these practices engage long-term debt involving a year or much more between a gift and its equivalent repayment. A major transaction with multiple, additional shell exchanges pinned to its passage may go on for years, and big men work toward this result, often eschewing exact returns in order to keep debts, relationships, active. Nevertheless, less prestigious exchanges may occasionally be completed with a visitor who brings the appropriate type of shell for immediate exchange, thus avoiding the risks of default inherent in time and distance. Defaults, however, are common, and many a high name has tumbled into obscurity.

Although kula participants are primarily men, recent research points to the contemporary importance of women all across Muyuw (Woodlark Island), in the Bwanabwana region (Macintyre, 1983a), and Dobu (J. Leach in Leach and Leach, 1983, p. 17). Nevertheless, women participants rarely sail on kula voyages and it is very rare for a woman to become as significant in the institution as some men regularly do.

Research from the 1970s has also revealed a crucial type of shell ownership not noted by Malinowski (but probably operative in his time) called 'kitoum/kitomu' (or cognates). It is fundamental to the *modus operandi* of kula. A shell's first owner is its maker. Owners may do as they wish with their kitoum shells, keeping them for strategic use in the main kula competition; turning them into significant quantities of food, pigs, magic, or cash; using them to pay off another person's kula debt, and so acquire the relationship that debt made; or, on some islands, using them to replace people who may have been killed by witchcraft or other actions. On some islands, men formally have to replace their deceased wives with one or more kitoum, if the woman has moved from her village to his. Among the more intense actors within particular locales, these transfers become purposely enmeshed in marriage and kinship, and so intergenerational relationships become part of the category's significance. A shell transferred as a kitoum becomes another person's kitoum, thus overriding a valuable's understanding as an armshell or necklace, and in isolated cases, a valuable may even reverse customary shell directionality (counterclockwise for armshells, clockwise for necklaces). If used in normal kula exchange, when one gives away one's kitoum its equivalent return becomes, on receipt, the replacement for the original kitoum, the prior shell becoming the kitoum of the equivalent's donor. Thus particular kitoum represents powers or capacities that may change form from one thing to another, from an armshell to a necklace, armshell/kitoum to a pig, boat, or one clan or subclan's possession to another's (Damon in Leach and Leach, 1983; Munn, 1977, 1986; Weiner, 1988, 1992 cf also Godelier, 1996, pp. 90–91). Kitoum shells, empty of prior obligations, may be said to provide an element of freedom in transactions and so variously aid owners in the building of their positions. They anchor a kind of capital circuit.

Adjustment of Some Common Oversimplifications of Kula

Following Malinowski's emphasis, a common assumption is that modern kula entails a ritual exchange of just the two types of shell set apart as a system from trade with other nonritual goods (see Section Anthropological Approaches to Kula). But a variety of evidence indicates that such a sharp line cannot be drawn between different items produced or acquired for exchange. For instance, in Bwanabwana and Muyuw (**Figure 1**), pigs are intrinsic to kula (there called 'kune') although they obviously do not circulate around the whole ring (e.g., Macintyre, 1983a, pp. 105–106). Overseas directional transactions of Gawan-built canoes can also be described as 'kula-ing' one's canoe. Armshells from southern canoe recipients become the canoe producers' kitoum; with the exchange finished, canoes are transferred from their makers by means of shell valuables (Munn, 1986). Moreover, the institution is the ultimate expression of regional interdependency which, if not always visible, becomes clear in crisis contexts. The 1997–98 extreme El Nino Southern Oscillation first created a desperate situation on the small islands between Muyuw and the Trobriands, then eventuated a massive movement of yams, as food and seed, from Muyuw to Gawa, Kweywata, and Iwa. That movement of resources swamped the usual patterns of reciprocity in the region (Iwa, for example, is packed with various nut trees which it regularly gives as gifts to other islanders). Participants explained this situation by ignoring

the lower-tiered practices. They couched their actions in the requirements of the kula exchange: without that movement of food and seed stock Iwa would cease playing its role in the chained relational network that is the first condition of the region (see Damon, 2008).

If kula media are not as exclusive as sometimes supposed, so also the circulation of the shells themselves is not exclusive to Massim islands participating in the ring: other islands make use of these shells, leaking them into and draining them out of the kula region, thus affecting kula circulation. Indeed, as Malinowski reported, a source of the necklaces (Rossel Island) is itself far south of the ring. Until the 1960s, Maylu people from along the southeastern coast of Papua New Guinea used to sail deep into the northern reaches of the ring to dive for conus shells for their own different use of the valuables. People from Muyuw, who directed them to their reefs, remember this as a time and context for an entirely different exchange of regionally defined products compared to those that regularly circulated within the region as defined by the reciprocal exchange of armshells for necklaces. Kula is in this sense an open not a closed system, and thus very much like Lévi-Strauss's model of generalized exchange (Levi-Strauss, 1969, see also Weiner, 1976).

Neither is kula static. In its present form, which can be tentatively dated to *c.*1400, it is a modern development undoubtedly affected by British colonization. One key change has been the widening of participation in kula. Formerly confined to men of influence, it may be argued that the kula effloresced, becoming increasingly 'democratized' as endemic warfare gave way to British 'pacification' (Macintyre, 1983b; Young, 1983). Of course change is ongoing; for instance, in the 1990s the increased use of Western boats by Bwanabwanans was reducing the southward extension of canoe transactions, thus negatively affecting Gawan acquisitions of kitoum from canoes. As of 2012, the last two-way voyage of an eastern kula ring large kula canoe (anageg) was in 2002. The canoe builders on Gawa, Kweywata, and Yanaba are trying to adapt a new canoe structure to their old role in the network of islands. Interestingly, this new form comes out of the very southeastern appendage of the region replicating the myth if not the reality of the 'original' anageg form, which, while largely built in Gawa, Kweywata, and to a lesser extent Yanaba, most people consider it to stem from the Sudest (Vanatinai) Island area.

At another level, change is inherent in the way individual kula paths and routings are constituted. A path of partners is initiated and maintained only through the passage of shells: without shells a path disappears. *Pace* Malinowski (1922, p. 83), partnerships are not necessarily 'permanent.' In a politics that entails the constant testing of trust and building of productive support systems, shifting personal relations alters paths across the ring. Nor are routes between communities fixed. Over time kula's competitive practices can yield not just individual path alterations, but also alterations in routes. Thus by the 1970s Duau (north-east Normanby) villages no longer did much kula directly with Bwanabwana: new southern coastal villages of people from the interior inserted themselves into the route (Thune in Leach and Leach, 1983, p. 352). In the 1970s the hierarchy of Trobriand chiefs played miniscule roles in the kula; that is not true now since, for example, virtually every significant kula actor in Muyuw, in the northeast sector of the ring, has or desires to have a relationship with the Trobriand paramount chief, Pulayasi.

This dynamic view of kula also illuminates noncircular aspects of its sociogeography. A path may never attain circularity because closure can be produced only by making reliable connections via the passage of shells. Although participants invariably describe the total system by means of the clockwise and counterclockwise movement of the two valuables, only the very highest ranked valuables occasionally travel in closing circles. Two that were said to be doing that in the 1970s and 1980s, the armshell Klibulouboul and necklace Senubet, were not so traveling over the last 15 years. Most valuables remain organized by unified lines of intentionality for only a few moves before being defined by different motives. Understanding, influencing, and reacting to diverse intentions are in fact in the center of the contest that is the kula.

Just as kula participation and paths are subject to change, so also is the larger organization marked by variability throughout the ring. Without question kula shells act as a kind of currency and so play a profound mediating role. But whereas many currency systems operate by means of an absolute equivalency, a single hierarchizing value, and although one may view the kitoum category in this guise, the fundamental structure of the institution forms patterns of discontinuity. The complementary differences between the two shell types, armshell and necklace, in many parts of the region phrased as male and female, respectively, symbolize if not generate an elaboration of differences. The system runs on niches, minor and large productive differences between islands and parts of islands. Additionally, there are different articulations with marital and other domestic exchanges. In Bwanabwana, kula is not sharply separable from marriage exchanges, but in the Trobriands the two are kept apart (Macintyre, 1983a; Weiner in Leach and Leach, 1983). Elsewhere, articulations vary. Gawan men take shells from their kula paths to give to affines for the latter's paths; shell returns ideally go back to donors and the original paths (Munn, 1986). But on Vakuta, shells taken off paths for affinal recipients in return for gifts of pigs become these affines' kitoum, even though they may not be the donors' kitoum (Campbell in Leach and Leach, 1983, pp. 206–207); presumably donors must find path replacements elsewhere. These sorts of local differences, among many others, have ripple effects on translocal shell movements (cf also Wagner, 1989, p. 270; Persson, 1999). These facts are not just the result of analytical distance. Since socially significant numbers of people travel from one island to another consciousness of cultural variation is intrinsic to the social order, and in the experience of some of the recent ethnographers at least, it is evident that communities regularly select particularly smart youths for significant interisland experience. In the 1970s, some of the elders running the system in its northeast corner worried that the relatively recent expansion of the United Methodist Church, due to a particularly active SIL (Summer Institute of Linguistics) missionary Bible translator, was pulling away so many of the smart younger men that in the near future there would not be enough shrewd people to keep the system intact. In 2012, one of those people who had been pulled

away, but who had now retired from the Church, agreed that this was a problem.

Anthropological Approaches to Kula

Serious focus upon kula first emerged within two formative studies of anthropological thinking on exchange: Malinowski's (1922) first major work developing his use-oriented functionalism and Mauss' (1950) classic essay on the gift. Basic problems posed by typical, pre-1970s kula commentaries drew on Malinowski's questionable dichotomy between kula as a ceremonial practice involving the transaction of 'useless' things and the economic trade of 'useful' things (for a brief critique see Young, 1983, p. 8). The distinction between ceremonial (kula) and nonceremonial trade was also the prototype for the later development of more complex models of economies with multiple separate 'exchange spheres' (cf Firth, 1957, p. 218).

The same Western economic dualism framed the search for some single explanatory function, or 'purpose' of kula such as 'prestige competition' (J.W. Leach in Leach and Leach, 1983, pp. 5–6). In particular, it fueled the concern with finding an economic function for kula by putting it in the service of the 'useful' commerce – for instance, as establishing 'the peace of the market' enabling utilities to circulate (Singh Uberoi, 1962, p. 148; see Macintyre's 1983b critique of the association of kula with peacemaking). The period's most serious reexamination combined some of these ideas in a more broadly based structural–functional study (Singh Uberoi, 1962) which viewed Trobriand kula as a sociopolitical practice periodically releasing men from internal descent group relations to establish interisland bonds.

It is, however, Mauss' concerns that pervade contemporary kula analyses. Mauss situated kula in a theory of moral–juridical contract empowered by bonds between persons and things. His attention to the cultural symbolism and ontology of 'person–thing' relations in gift exchange (rather than simply to exchange relations between things, as in the 'exchange spheres' model) has been influential in recent exchange and kula studies. Godelier (1996 [1999]) synthesizes much of the 1970–90s word derived from Mauss. Mauss' contrast between 'archaic' gift exchange and the modern Western market-based system is also one root of current arguments involving another dualism, that of 'gift' and 'commodity' exchange. Kula figures prominently in these arguments – for instance, in Gregory's (1982) widely cited, if problematic, model of gift exchange as the transaction of 'inalienable objects' in contrast to the 'alienable objects' of commodity exchange. Some would say that the dynamics of kitoum may be expressed in terms of alienated powers, and if so this would suggest the difficulties of applying this dichotomous framework to kula. This, however, may be misreading what is intended by the discussion of alienation. That idea is not so much about people being separated from objects as it is about being separated from their capacities, their powers. In the kula, people are separated from their kitoum, their boats, their pigs, their brothers, sisters, and children. But when these disjunctions find their replacements in kitoum, the absolute pivots of kula exchange, their powers are not so much lost as transformed into other modalities. It is no wonder then that kula actors marvel about where 'profit' comes from in the Western system that now encompasses them. Many of these are people who have the cultural memory of at least 100 years of Western intrusion and who can recount this time period's fantastic transformation in wealth, that is capital, the difference between the picks and shovels of 1910 and Caterpillar Wheel Loaders of today, and the tankers and other massive steel ships that traverse this region to and fro between Indonesian oil and Australian twenty-first-century mechanization. While the institution condenses many features of Indo-Pacific societies scholars will contemplate well into the future, it also continues to hold up a mirror for Western practices long ago anticipated in Malinowski's first masterpiece.

No single orientation dominates contemporary kula studies. Rather, kula has been subject to varied forms of rethinking: for example, in terms of a Marxian concern with grounding circulation in production and labor value (Damon, 1980), a symbolic formulation of action, spacetime, and value (Munn in Leach and Leach, 1983; Munn, 1986), or gendered models of power and social reproduction (Weiner, 1992). Persson's (1999), following to some extent Damon (1990), detailed a comparison of selected Massim societies, and thus approached kula through the study of regional variation (see also Damon, 2005). Among the analyses that the future foretells are a richer understanding of the anthropogenic nature of the region's ecology and an understanding that will embed its cultural forms in the social variation evident from at least Bali to West and the sacrificial ranking systems that created what we know of Polynesian sociality to the east. Since its introduction into anthropology at the beginning of the twentieth century, kula has been a sounding board for anthropological issues and continues to be so as a new set of scholars take its bearing amid the classics in a new century.

See also: Exchange in Anthropology; Malinowski, Bronislaw (1884–1942); Mauss, Marcel (1872–1950); Money: Anthropological Aspects; Oceania, Sociocultural Overviews: Melanesia; Southeast Asia: Sociocultural Aspects.

Bibliography

Ambrose, W., Petchey, F., Swadling, P., Beran, H., Bonshek, E., Szabo, K., Bickler, S., Summerhayes, G., 2012. Engraved prehistoric Conus shell valuables from southeastern Papua New Guinea: their antiquity, motifs and distribution. Archaeology of Oceania 47, 113–132.

Belshaw, C., 1955. In search of wealth: a study of the emergence of commercial operations in the Melanesian society of southeast — Papua. American Anthropological Association Memoir 80, 57 (1). pt. 2.

Bickler, S.H., 2006. Prehistoric stone monuments in the northern region of the kula ring. Antiquity 80, 38–51.

Damon, F.H., 1980. The Kula and generalised exchange: considering some unconsidered aspects of the elementary structures of kinship. Man 15, 267–292.

Damon, F.H., 1990. From Muyuw to the Trobriands: Transformations along the Northern Side of the Kula Ring. University of Arizona Press, Tuscon, AZ.

Damon, F.H., April–June 2002. Kula valuables, the problem of value and the production of names. L'Homme 162, 107–136.

Damon, F.H., 2005. 'Pity' and 'ecstasy:' the problem of order and differentiated difference across kula societies. In: Mosko, Mark, Damon, Fred (Eds.), On the Order of 'chaos' Social Anthropology & the Science of 'Chaos'. Berghahn Books, pp. 79–107 (Chapter 4).

Damon, F.H., 2008. On the ideas of a boat. From forest patches to cybernetic structures in the outrigger sailing Craft of the eastern kula ring, Papua New Guinea. In: Sather, Clifford, Kaartinen, Timo (Eds.), Beyond the Horizon. Essays on Myth, History, Travel and Society. Studia Fennica Anthropologica 2. Finnish Literature Society, Helsinki, pp. 123–144.

Firth, R., 1957. The place of Malinowski in the history of economic anthropology. In: Firth, R. (Ed.), Man and Culture. Routledge and Paul, London.

Fortune, R.K., 1932. Sorcerers of Dobu. Routledge and Sons, London.

Godelier, M.F., 1996. L'enigme du don. Librairie Artheme Fayard (1999. The Enigma of the Gift). University of Chicago, Chicago.

Gregory, C.H., 1982. Gifts and Commodities. Academic Press, London.

Kuehling, S., 2005. Dobu: Ethics of Exchange on a Massim Island. University of Hawaii Press, Papua New Guinea.

Leach, E., Leach, J.W. (Eds.), 1983. The Kula: New Perspectives on Massim Exchange. Cambridge University Press, Cambridge, UK.

Lévi-Strauss, Claude, 1969 (Translated from the French). In: Bell, James Harle, von Sturmer, John Richard, Needham, Rodney (Eds.), The Elementary Structures of Kinship. Beacon Press, Boston.

Macintyre, M., 1983a. Changing Paths: An Historical Ethnography of the Traders of Tubetube. Ph.D. thesis (unpubl.), Australian National University, Canberra, Australia.

Macintyre, M., 1983b. Warfare and the changing context of *kune* on Tubetube and Papua–New Guinea. Journal of Pacific History 18, 11–34.

Malinowski, B., 1922. Argonauts of the Western Pacific. Routledge and Sons, London.

Mauss, M., 1950. Essai sur le don. Formes et raison de l'echange dans les sociétés archaiques. In: Sociologie et anthropologie, first ed. Presses Universitaires de France (1990. The Gift: The Form and Reason for Exchange in Archaic Societies). Routledge and Kegan Paul, London.

Munn, N., 1977. The spatiotemporal transformation of Gawa canoes. Journal de la société des Océanistes 33, 39–53.

Munn, N.D., 1986. The Fame of Gawa: A Symbolic Study of Value Transformation in a Massim (Papua New Guinea) Society. Cambridge University Press, Cambridge, UK.

Persson, J., 1999. Sagali and the Kula: A Regional Systems Analysis of the Massim. Department of Sociology, Lund University, Lund, Sweden.

Roheim, G., 1950. Psychoanalysis of a Matrilineal Society. Psychoanalysis and Anthropology. International Universities Press, New York.

Singh Uberoi, J.P., 1962. Politics of the Kula Ring: An Analysis of the Findings of Bronislaw Malinowski. Manchester University Press, Manchester, UK.

Wagner, R., 1989. The exchange context of the kula. In: Damon, F.H., Wagner, R. (Eds.), Death Rituals and Life in the Societies of the Kula Ring. Northern Illinois University Press, De Kalbe, IL.

Weiner, A., 1992. Inalienable Possessions: The Paradox of Keeping While Giving. University of California Press, Berkeley, CA.

Weiner, A.B., 1976. Women of Value, Men of Renown. University of Texas, Austin, TX.

Weiner, A.B., 1988. The Trobrianders of Papua New Guinea. Holt Rhinehart and Winston, New York.

Young, M., 1983. The Massim: an introduction. The Journal of Pacific History 18, 4–10.

Kuznets Curves

Tommaso Luzzati, Università di Pisa, Pisa, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explains what the Environmental Kuznets Curve is, how was it born, what are its theoretical underpinnings, and the findings of the empirical literature. It discusses the main weaknesses of the debate, also highlighting an inconsistency that has been ignored so far. This article is not a survey of the literature. Rather, it attempts to provide a systematic and original presentation of (hopefully) all the ingredients of a fairy tale, telling us that economic growth will eventually result in environmental improvement.

What Is the EKC?

"Does inequality in the distribution of income increase or decrease in the course of a country's economic growth?" is one of the main questions of the presidential address given in 1954 by Simon Kuznets at the 67th annual meeting of the American Economic Review (Kuznets, 1955, p. 1). His answer was that inequality might increase with per capita income at an early stage of economic development and eventually decrease as income continues to grow. In graphical terms, the relationship between inequality and real per capita income might have a 'bell-shaped' or 'inverted-U' curve. When reading the article, one realizes how cautious Kuznets was in suggesting this relationship. The scarcity of data led him to define the trends he found as "perilously close to pure guesswork," and to affirm that his articulated "attempt to explain these dimly discernible trends may surely seem foolhardy" (Kuznets, 1955, p. 5), and to describe his paper as "perhaps 5 percent of empirical information and 95 percent speculation" (Kuznets, 1955, p. 26). Actually, his main goal was to stimulate research in a topic that he viewed as "central to much of economic analysis and thinking" (Kuznets, 1955, p. 26) and calling "for a shift from market economics to political and social economy" (Kuznets, 1955, p. 28).

The prudence used by Kuznets did not characterize much of the literature that, beginning in the early 1990s, looked for an inverted-U relationship between various forms of environmental pressures/states and income per capita; that is, a relationship predicting the economic growth harms the environment in the early stage of economic development and improves it after a given level in per capita income is reached (see **Figure 1**). This relationship was named after Kuznets as the Environmental Kuznets Curve. Dasgupta S. et al. (2002, p. 147) attributed the coining of the term to Grossman and Krueger (1993). Actually, Grossman and Krueger did not use the term EKC in the 1993 paper nor in its preliminary version, the NBER working paper appeared in 1991. Panayotou (1993) was the first to use this term in a published paper. The EKC hypothesis was rapidly taken by many economists as a stylized fact—For instance, at the very beginning of the debate, Panayotou (1993: Preface) affirmed that "an environmental Kuznets curve is an empirical reality" and Selden and Song (1994, p. 147) wrote that "there is an emerging consensus that at least some forms of pollution exhibit inverted-U, or 'Kuznets' relationships with economic development." The EKC was used to maintain that (market driven) economic growth is needed for sustainability. Beckerman (1992, p. 491), in an often quoted sentence, wrote that "in the end the best – and probably the only – way to attain a decent environment in most countries is to become rich." Clearly, the EKC hypothesis can be traced back to the controversy on the environmental limits to economic growth started by the famous Club of Rome report (Meadows et al., 1972) that showed how unregulated growth could become a major threat to human welfare. The policy orientation of the EKC debate is expressed well by the quote of a paper by Grossman and Krueger (1995, p. 353) who commenced by asking "Will continued economic growth bring ever greater harm to the earth's environment? Or do increases in income and wealth sow the seeds for the amelioration of ecological problems?"

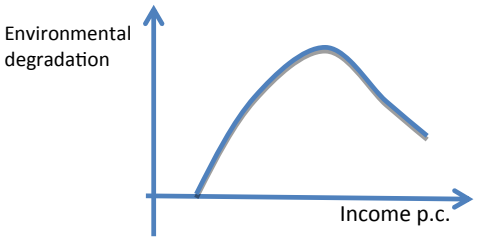

Figure 1 Environmental Kuznets curve.

Grossman and Krueger (1991) are actually those who started the EKC literature with a paper about the environmental impacts of the North American Free Trade Agreement (NAFTA) agreement, closely followed by other seminal contributions, particularly Shafik and Bandyopadhyay (1992), Hettige et al. (1992), Holtz-Eakin and Selden (1992), Panayotou (1993), Selden and Song (1994), and Grossman and Krueger (1995). The topic gained immediate popularity and, since then, a huge number of articles have been published. One of the reasons for this success was that, since the 1990s, large datasets with environmental indicators became eventually available, despite the urgent need for them was already advocated in the 1950s by the pioneering contributions on enviromental degradation of the German institutional economist K.W.Kapp. (see Luzzati, 2009).

Theoretical Underpinnings for a Bell-Shaped Relationship between Per Capita Income and Environmental Degradation

(Environmental degradation and quality are thought by economists as latent constructs. Empirical analyses attempted to capture them by using indicators of environmental pressure and, less often, state. It has to be emphasized that also income 'inequality' in the original Kuznets curve can be captured by different indicators. Its meaning, however, is much more precise, than that of the vague idea of environmental quality.) Following Grossman and Krueger (1991), economic growth is usually thought to affect environmental quality through the following effects: 'scale,' 'composition,' and 'technique.' 'Scale' refers to the *size* of an economy and has an adverse effect on environmental quality since increasing output requires more input and involves more emissions and waste. 'Composition' refers to the structure of the economy, that is, to the *output mix* that is produced, which varies according to the specialization of a country. Economic growth will be environmentally beneficial if it causes the economy to move toward less polluting outputs. 'Technique' refers to *how* a product is obtained, which includes both the input mix and the production process. Technological development associated with economic growth might help in substituting 'dirty' inputs and processes with cleaner ones. In terms of the neoclassical production model, a change in the input mix corresponds to a move along the isoquant while a more efficient process corresponds to a downward shift of the isoquant.

One can think of composition and technique as *channels* through which economic growth can eventually become beneficial to the environment. Which are then the *driving forces* that have been thought to operate through these three channels? To answer to this question it is interesting to read Panayotou (1993, p. 1) who affirms that "At higher levels of development, structural change towards information-intensive industries and services, coupled with increased environmental awareness, enforcement of environmental regulations, better technology and higher environmental expenditures, result in levelling off and gradual decline of environmental degradation."

This quotation summarizes the major ideas thought to predict an EKC curve, that is

1. economic growth brings along structural change ('composition' effect) and
2. technological development, which makes available a wider range of 'solutions' and lowers intervention costs. In terms of consumer theory this can be thought of as a decreasing price for environmental services.
3. Economic growth increasingly harms the capacity of the natural environment as resource provider and absorption sink, involving an increasing awareness of the environmental problems which is also fostered by
4. the advancement in knowledge, which improves our understanding of human impacts on the environment and their feedbacks on human welfare.
5. Finally it is often assumed that the environmental quality is a luxury good; that is, its income elasticity is greater than one and needs not be demanded at low levels of income.

Points (2) to (5) refers to the technique effect which is supposed to operate because economic growth brings both more potential for lowering the human impact on the environment (2) and more willingness to pay for environmental services (3), (4), and (5). Technique effect is then thought of as rooted in consumer theory predicting a more than proportional relationship between income and demand for environmental services. We can easily observe this with the help of Figure 2 that depicts a standard consumer choice problem. Suppose that preferences can be represented by a Cobb–Douglas utility function $U = CQ$, where C is a bundle of consumption products and Q is environmental quality. Q depends on what is freely provided by Nature (and affected by human activities), α, and by resources spent to improve environmental quality, E. For the sake of simplicity, suppose $Q = \alpha + E$. Figure 2 shows the problem of the consumer who has to allocate her/his income between E and C, having the utility function $U = C(\alpha + E)$ and facing a price for C equal to p_C and a price for E equal to p_E. The broken line containing points 'a,' 'b,' 'c,' 'd,' is the income–consumption curve that shows the optimal solutions when income changes. For low levels of income, the consumer is forced into corner solutions, e.g., point 'a,' where the demand of E is constrained to zero, while the optimal one would be negative, in 'a_0.' From 'b' onwards income becomes sufficient for the consumer to spend in E.

The effect of environmental degradation (3 above) is captured by decreases in α, which forces the consumer to spend more on E: the indifference-curve map changes so that the income-consumption line moves left (not shown), involving, e.g., point 'd_1' as an optimal choice. A similar change in the indifference-curve map can be thought of as caused by a better knowledge and awareness (4 above). Note that, due to the decrease of α, the utility in d_1 might be lower than the utility in 'c.'

The effect of technological improvement (2 above) can be thought of as lowering the price for the intervention on the environment, p_E, involving a steeper slope of the budget constraint, which in turn moves the income consumption line to the left (not shown) and makes it steeper (not

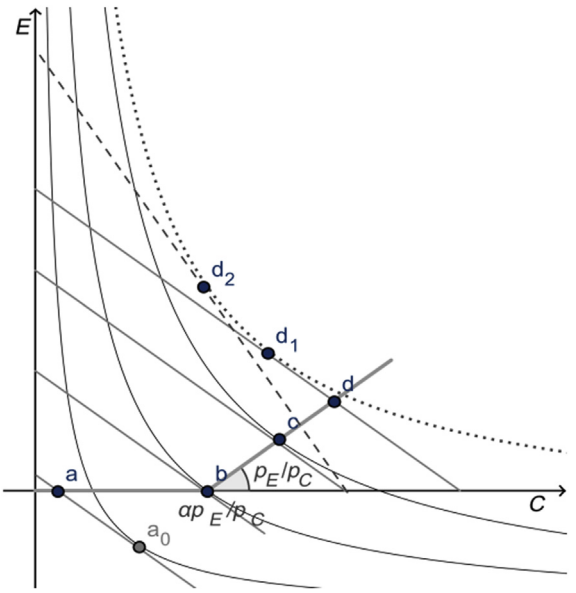

Figure 2 The increase in the demand for environmental quality.

shown). As a consequence, the demand for E becomes higher. Given the same level of income as in 'c,' the optimal choice with technological progress and environmental deterioration would be 'd_2.'

Following these lines of thought, the bell-shaped curve between environmental quality and per capita income is easily obtained. In the first stage of development, the willingness to give up consumption for getting better environmental quality is very low or absent. As long as income grows, the environment becomes increasingly deteriorated and willingness to pay increases. We become more aware of environmental problems while technology advancement makes environmental interventions less. Obviously, none of these conditions is necessary. What is needed is that willingness to pay for environmental services increases strongly enough to stimulate mechanisms able to offset the damages produced by the income growth via the 'scale' effect. Among those mechanisms one can include tighter regulation and incentives to 'green' investments, emergence of less environmentally damaging consumption and defensive expenditures. It is worth noting that other factors have been put forward in favor of the EKC hypothesis, such as economies of scale in the pollution control techniques (e.g., Andreoni and Levinson, 2001) or institutional arrangements (e.g., Panayotou, 1997). These factors allow more effective law enforcement and regulations, lower levels in subsidies hiding resource scarcity, higher probability of accepting environmental concerns, and better regimes of property rights. Concerning the latter point, some authors argue that a larger extent of private property in rich countries allows for more efficient allocation involving lower negative externalities and environmental degradation (e.g., Cropper and Griffiths, 1994).

Using a different combination of the mentioned ideas, a large number of theoretical models were used, including growth models attempting to accommodate the dynamic interactions between the environment and the economy. These theoretical models show how an inverted-U relationship between environmental degradation and income can emerge. It has, however, to be admitted that the question is an empirical one. Building a theoretical model, although exciting from an intellectual point of view, can eventually prove to be rather useless if its predictions depend on parameters that are inconsistent with empirical evidence. Since data on the parameters used in the models are often not available, most of the debate on the EKC has focused on an empirical estimate of the relation between per capita income and various environmental indicators, as will be discussed in the next section.

Empirical Analysis

The EKC was born as an empirical question and remained as such. Most of the studies are regression analyses of the empirical relationship between per capita income and various environmental indicators. Regression analysis is about looking for models that are able to predict a dependent variable with the behavior of other variables, which are exogenous and independent from each other. In this case, this would imply 'explaining' (predicting) environmental degradation with models that take into account all potentially relevant variables and relationships, including the possibility of having more than one dependent variable.

This, however, clashes with the very narrative of the EKC, since the debate started with the explicit purpose of challenging the view according to which economic growth brings along environmental deterioration. In other words, the purpose was to understand whether the net income effect becomes eventually beneficial to the environment. This is done by univariate regressions, with per capita income as the sole, independent variable, that is, by estimating a reduced-form model (Grossman and Krueger, 1995, p. 359) of a complicated web of determinants and relationships.

Such a research strategy has advantages and disadvantages. On the one hand it bypasses the issue of data availability when including many variables but on the other it introduces many potential problems concerning the econometrics, in particular due to omitting relevant variables and simultaneity problems, due to the feedback from environmental quality to income and more generally, to sustainability (Arrow et al., 1995). In any case, a more standard approach, e.g., including other explanatory variables, is not consistent with the question the debate aims to answer, net income effect on the environment.

The econometric studies of the EKC vary according to different aspects, namely,

1. indicators (and databases) used to proxy income and environment, particularly if they refer to different geographical levels (city, regions, country, world)
2. units included in the sample
3. shape of the relation tested (quadratic, cubic, nonparametric)
4. the use of variables other than income as regressors
5. econometric bad practices (or even 'mistakes').

The choice of income indicator is relevant when the units of analysis belong to countries with different currencies. Some studies have used GDP at market exchange rates, which, however, does not necessarily reflect what individuals can buy domestically with their income, especially in economies not particularly open. For this reason the most used indicator is income expressed in purchasing power parities.

Several distinctions can be made to classify environmental indicators. A major one, which is not always acknowledged in the EKC studies, is between 'pressure' and 'state' indicators, referring, respectively, to pressures exerted by society on the environment, e.g., emissions or resource use, and the quality of the environment, e.g., concentration of pollutants or resource stocks. This distinction is central to the indicator framework adopted by many agencies, the Pressure-State-Response framework and its evolutions. It was developed by OECD at the beginning of the 1990s. (For details see http://www.fao.org/ag/againfo/programmes/en/lead/toolbox/Refer/EnvIndi.htm or http://ia2dec.ew.eea.europa.eu/knowledge_base/Frameworks/doc101182/). One also has to consider the geographical, urban, country, or worldwide level. Especially in the early studies, evidence of an EKC emerges for local pollutants (water and air) and not for more global ones. This is in accordance with the progressive move out of the urban areas of pollutant productive activities (due to rent increases). It is also consistent with the idea that local impacts are more likely to get tackled, particularly if their abatement costs are low.

A very relevant source of differences is the sample type. Also due to lack of long time series, data used are usually panels of time series observations for different units (Cities, States, Countries). Including or excluding units (or changing the time span) greatly affects the results, concerning both the existence and estimated per capita income of the turning point (that is the p.c. income level after which environmental degradation decreases). Obviously it is very important to exclude outliers, that is, units with peculiarities such that the others cannot imitate their pattern.

Since the purpose was to look for an inverted-U relationship, some of the earliest studies tested for a quadratic relationship between environmental indicators and income. The cubic relationship became soon more popular since it allows for more possible shapes, depending on the estimated values of the parameters. To further avoid the problem of forcing data into a specific (time) frame, since the 2000s (Millimet et al., 2003) some EKC analyses used semi-parametric methods. With these methods the dependent variable is estimated as a smooth and continuous (possibly nonlinear) function of the regressor, allowing for freedom in the shape of the relationship.

Despite having introduced the EKC as a reduced form story, with income as the unique regressor, in most studies other regressors have been added basically to improve the fit of the estimates. Such a strategy (a kind of compromise between a structural form model and an univariate regression) gives results that are difficult to interpret. Once the significance and effect of some variables are isolated, what interpretation has to be given to the estimated coefficients of per capita income? For instance, a time trend is often found statistically significant and interpreted as a proxy for technological progress. Is it not growth in income that drives technological advancement and the other factors what ultimately make economic growth environmentally beneficial? Moreover, several explanatory variables are used without specifying either a simultaneous equation system from which the reduced form could be obtained or providing solid justifications involving the risk of ad-hocery.

Finally, some methodological and econometric bad practices have been pointed out. For instance, Luzzati and Orsini (2009) have emphasized that there are no reasons for using environmental pressure indicators in per capita terms, as is customary in the EKC literature. This is wholly inconsistent not only with using concentration indicators in absolute terms, but also with common sense. Actually it is not good news to observe pressure reductions in per capita terms if accompanied by a population growth that more than offsets pressure reductions. Nature does not care about per capita emissions. Total impacts are relevant. Theoretical models have acknowledged this (see, e.g., de Bruyn and Opschoor, 1997, p. 256; Aghion and Howitt, 1998, pp. 151–171; and the review by Brock and Taylor, 2005). The same is acknowledged by the EKC hypothesis, which is told as a story between total human pressure and per capita income.

Several authors raised econometric criticisms, particularly Stern (e.g., 2004). Early studies disregarded issues like heteroskedasticity, simultaneity, omitted variable bias, and time series properties. The latter is particularly relevant since nonstationarity variables easily give rise to spurious correlation. The unit roots test and the cointegration analysis, as commonly used in more recent works on EKC, prerequisite any serious EKC study.

Given the large differences in premises and methods, the expected outcome is a large variability of results of the empirical EKC studies. Actually a large spectrum of different patterns emerged, monotonic increasing and decreasing, inverted U-shaped, and N-shaped (Panayotou, 2003, pp. 62–67 contains a large summary table). Differences emerged also concerning the estimate of the turning point for the same environmental indicators, if there is one. Sometimes, actually, the turning point is only theoretical, since it lies outside the range of actual observations that is of an income level not experienced.

Hence the skepticism expressed by Copeland and Taylor (2004, p. 8) "about the existence of a simple and predictable relationship between pollution and per-capita income" is still valid.

An Assessment of the EKC Literature

The debate on the EKC has been highly controversial, acknowledging many caveats that have also brought forth favorable evidence. The paper by Arrow et al. (1995) in *Science* was perhaps the first critique to the emerging EKC literature. Since then, several scholars raised important issues, most of which are briefly reported above, that make overall the EKC hypothesis rather weak. Part of this weakness depends also on the EKC narrative being somehow naïve. There are several reasons for such a claim. For instance, the narrative contains the idea that poor societies do not harm the environment, an idea that is falsified by many examples of which the Easter Island is just one. Furthermore, treating the natural environment as a 'luxury' could be correct from the viewpoint of persons living in relatively rich societies. This might not be the case for people whose subsistence is secured outside the market in the natural environment (see Martinez Alier, 2002). Moreover, it is not obvious that relative increases in the service sector carry reductions in the material and energy throughput and in environmental impacts. For instance, air transportation is within the service sector. Consider also that weightless products such as software require both matter and energy for using them.

Even more importantly, it is somehow naïve to expect that the vague and complex concept of 'environmental degradation' can adequately be captured by a bunch of available indicators. Data availability is a major constraint in applied research. When data are available, it is because society is already conscious of those problems and could already have attempted to formulate some kind of policy response. Problems emerge and possible explanations are put forward collecting statistics to validate hypothetical explanations. For this reason, one should not be surprised to see that some available environmental indicators in some part of the world have improved over the course of time (and development). The question, however, is whether the adopted remedies have actually diminished the environmental impact or simply transformed into other (perhaps unknown) impacts. In other words, progresses on some environmental indicators could have been obtained at the expense of other environmental problems that have not yet attracted our attention. Hence, substitution/shifting of impacts could be more relevant than income growth for explaining progress in some environmental indicators.

Shifting is the basis of the 'new toxics' scenario which claim that as pollutants are cleaned up new ones emerge. Dasgupta et al. (2002, p. 148) attribute this hypothesis to the "pessimistic school" rather than considering it just as a hypothesis. I by the way do not find it fruitful to mix scientific reasoning and emotional concepts such as 'optimism' or 'pessimism.' Shifting is also relevant in explaining evidence of environmental improvement in many urban areas. Activities with high environmental impacts tend to move out of the densely populated areas. The same can be applied to international trade. As mentioned, the whole debate was born to assess the impact of trade liberalization between Mexico and the US. Many papers attempted to assess to which extent improvements in environmental indicators of rich countries can be attributed to migration of intensive pollution toward poorer countries, the so-called 'Displacement' or 'Pollution Haven' hypothesis. Since structural change in rich countries is not accompanied by equivalent change in the consumption structure, it is not surprising that some evidence in favor of the displacement idea has emerged (see e.g., Dinda, 2004). Part of structural change can be generically attributed to the low labor costs in poor countries associated with globalization. Migration of highly polluting industries can be attributed to weaker environmental regulations in poor countries, which in turn might put pressures on rich countries to mitigate the strength of their environmental standards ('race to the bottom hypothesis').

A few studies attempted to remedy the substitution issue by using indicators of overall impact, mainly material use (e.g., de Bruyn and Opschoor 1997; Eurostat 2002; Canas et al., 2003) and energy (e.g., Agras and Chapman, 1999; Suri and Chapman, 1998; Cole et al., 1997; Luzzati and Orsini, 2009). Energy is one factor "whose direct and indirect environmental significance is indisputable" (Jänicke et al., 1989, p. 25), the obvious reason being that our interferences with nature also arise from the power that energy gives us to move and extract matter. (For instance, by simply pushing a button, we are able to turn on engines that pump out water at unprecedented rates – rates which put water tables at serious risk of irreversible damage, no matter how clean the source of energy that we use.) Clearly, no 'best' indicator of complex phenomena such as environmental anthropogenic pressure can exist. Energy, however, is a rather good one due to its generic nature. When using energy in absolute terms, the EKC estimates do not support the EKC hypothesis at the world level, for panel data level (unless one is willing to include outliers), and for most countries (Luzzati and Orsini 2009, 2014).

A further step in this discussion is about impacts. Suppose that a genuine EKC emerges for environmental pressures, where genuine is supposed to mean total pressures and as depending on 'shifting,' that is on substituting one kind of pressure with another ones or moving the pressures to other regions. What can we predict about the state of the environment? An issue that has been acknowledged since the early stage of the debate is that of irreversibility. Even if pressures decrease, they might have already surpassed thresholds and have provoked irreversible changes in the environment with high costs to human society. 'Tunneling' the EKC was often called for, that is, implementing policies in order to shift down the EKC and avoid reaching too high levels of pressure.

The issue, however, becomes now to understand what 'too high' means. This would require not only to understand the complex relationship between pressures and the state of the environment, but also between the latter and the feedback on the economy and its impact on human well-being and sustainability. A precautionary approach might be a wise route to follow, particularly, if accompanied by the help of life sciences. If we did this we would discover that already in the 1930s a relationship was found between efficiency and size of animals, the so-called mouse-to-elephant or Kleiber's curve (e.g., Miller, 1986). This curve predicts that the higher the size, the greater the efficiency of animals. (An echo of this can be found in the economic debate over the Jevons' paradox (known also as rebound effect).) While waiting to see whether this also applies or not to efficiency induced by technological improvements, a precautionary approach would be for rich countries starting a diet to reduce their material obesity.

Acknowledgment

I am particularly grateful to Alison Testroete who helped considerably to improve the clarity of this article. Errors remain my own.

See also: Carrying Capacity of the Environment; Ecological Economics; Energy and Society; Environment and Health.

Bibliography

Aghion, P., Howitt, P., 1998. Endogenous Growth Theory. MIT Press, Cambridge.

Andreoni, J., Levinson, A., 2001. The simple analytics of the environmental Kuznets curve. Journal of Public Economics 80 (2), 269–286.

Arrow, K., Bolin, B., Costanza, R., Dasgupta, P., Folke, C., Holling, C.S., Jansson, B.O., Levin, S., Mäler, K.G., Perrings, C., Pimentel, D., 28 April 1995. Economic growth, carrying capacity, and the environment. Science 268, 520–521.

Agras, J., Chapman, D., 1999. A dynamic approach to the environmental Kuznets curve hypothesis. Ecological Economics 28 (2), 267–277.

Beckerman, W., 1992. Economic growth and the environment: whose growth? Whose environment? World Development 20, 481–496.

Brock, W., Taylor, S.M., 2005. Economic growth and the environment: a review of theory and empirics. In: Aghion, P., Durlauf, S. (Eds.), Handbook of Economic Growth. Elsevier, Amsterdam.

de Bruyn, S.M., Opschoor, J.B., 1997. Developments in the throughput-income relationship: theoretical and empirical observations. Ecological Economics 20 (3), 255–268.

Canas, A., Ferrao, P., Conceicao, P., 2003. A new environmental Kuznets curve? Relationship between direct material input and income per capita: evidence from industrialised countries. Ecological Economics 46 (2), 217–229.

Cole, M.A., Rayner, A.J., Bates, J.M., 1997. The environmental Kuznets curve: an empirical analysis. Environment and Development Economics 2, 401–416.

Cropper, M., Griffiths, C., 1994. The interaction of populations, growth and environmental quality. American Economic Review, Papers and Proceedings 84, 250–254.

Dasgupta, S., Laplante, B., Wang, H., Wheeler, D., 2002. Confronting the environmental Kuznets curve. Journal of Economic Perspectives 16, 147–168.

Dinda, S., 2004. Environmental Kuznets curve hypothesis: a survey. Ecological Economics 49 (4), 431–455.

Eurostat, 2002. Material Use in the European Union 1980–2000. Indicators and Analysis. Working Papers and Studies. Prepared by Weisz, H., Fischer-Kowalski, M., Amann, C., Eisenmenger, N., Hubacek, K., and Krausmann, F. Eurostat, Office for Official Publications of the European Communities, Luxembourg.

Grossman, G.M., Krueger, A.B., 1991. Environmental Impacts of a North American Free Trade Agreement. NBER Working Paper, No. 3914. NBER, Cambridge, MA.

Grossman, G.M., Krueger, A., 1993. Environmental impacts of the North American free trade agreement. In: Garber, P. (Ed.), The U.S.-Mexico Free Trade Agreement. MIT Press, Cambridge, pp. 13–56.

Grossman, G.M., Krueger, A.B., 1995. Economic growth and the environment. Quarterly Journal of Economics 110, 353–377.

Hettige, H., Lucas, R., Wheeler, D., 1992. The toxic intensity of industrial pollution: global patterns, trends and trade policy. American Economic Review 82 (2), 478–481.

Holtz-Eakin, D., Selden, T.M., 1992. Stoking the Fires? CO_2 Emissions and Economic Growth. NBER Working Paper No. 4248, Cambridge, MA.

Jänicke, M., Mönch, H., Ranneberg, T., Simonis, U.E., 1989. Structural change and environmental impact. Intereconomics 24, 24–35.

Luzzati, T., 2009. Human needs, sustainable development, and public policy: learning from K.W. Kapp (1910–1976). In: Salvadori, N., Opocher, A. (Eds.), Long-Run Growth, Social Institutions and Living Standards. Edward Elgar, pp. 306–322.

Kuznets, S., 1955. Economic growth and income inequality. The American Economic Review 45 (1), 1–28.

Luzzati, T., Orsini, M., 2009. Investigating the energy-environmental Kuznets curve. Energy 34 (3), 291–300.

Luzzati, T., Orsini, M., 2014. Reassessing the Energy- and CO2- EKC curve, manuscript presented at the biannual conference of the International Society for Ecological Economics. Iceland.

Martinez Alier, J., 2002. The Environmentalism of the Poor: A Study of Ecological Conflicts and Valuation. Edward Elgar, Cheltenham.

Meadows, Donella H., et al., 1972. The Limits to Growth: A Report to the Club of Rome. Universe Books, New York.

Millimet, D.L., List, J.A., Stengos, T., 2003. The environmental Kuznets curve: real progress or misspecified models? Review of Economics and Statistics 85, 1038–1047.

Miller, A.T., 1986. Energy Metabolism. F.A. Davis Company, Philadelphia, PA.

Panayotou, T., 1993. Empirical Tests and Policy Analysis of Environmental Degradation at Different Stages of Economic Development. Working Paper WP238, Technology and Environment Programme. International Labour Office, Geneva.

Panayotou, T., 1997. Demystifying the environmental Kuznets curve: turning a black box into a policy tool. Environment and Development Economics 2, 465–484.

Panayotou, T., 2003. Economic Growth and the Environment. Chapter 2 of Economic Survey of Europe, 2003. No. 2. UNECE. http://www.unece.org/fileadmin/DAM/ead/pub/032/032_p2.pdf.

Selden, T.M., Song, D., 1994. Environmental quality and development: is there a Kuznets curve for air pollution? Journal of Environmental Economics and Environmental Management 27 (2), 147–162.

Shafik, N., Bandyopadhyay, S., 1992. Economic Growth and Environmental Quality: Time Series and Cross-Country Evidence. World Bank Policy Research Working Paper WPS 904 (Background Paper for World Development Report 1992. World Bank, Washington, DC.

Stern, D.I., 2004. The rise and fall of the environmental Kuznets curve. World Development 32 (8), 1419–1439.

Suri, V., Chapman, D., 1998. Economic growth, trade and the environment: implications for the environmental Kuznets curve. Ecological Economics 25, 195–208.

L

Labeling Theory, History of

Bernard B Berk, California State University, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Labeling theory emerged as the dominant perspective in the study of deviance in the 1960s, though its origins can be traced to Durkheim. Labeling theory, influenced by symbolic interactionism, dramatically transformed the field by redefining what constituted deviance and what was significant to understand about deviance. It dominated the study of deviance for a quarter of a century. While its zenith has passed, it remains, as an aspect of societal reaction theory, an enduring perspective that continues to generate research refining the perspective and extending it to new areas under related concepts such as stigma.

The study of deviance has been shaped by a variety of streams of influence over the years each pulling the field in different directions. The earliest roots of the study of deviance can be traced to Emile Durkheim's (1897) pioneering investigation of suicide, over a century ago, in which he established a relationship between the organization of society and rates of suicide. From the onset, a dominant stream of influence has been this 'positivistic' perspective where (1) 'deviance' referred to norm-violating behavior, such as crime, delinquency, suicide, alcoholism, etc., and (2) research was oriented toward identifying the social causes or etiology of the norm-violating behavior as well as its consequences upon the larger society.

In the 1960s, however, the traditional positivistic perspective was eclipsed by what has been characterized as the labeling, societal reaction, or the interactionist perspective. The roots of this perspective can be traced to a second line of investigation of Durkheim (1895) in *The Normal and the Pathological*, which shifted attention from how society generated nonconforming behavior to how and why society created criminals through societal reaction. Durkheim's primary focus in this work was on macroanalysis, establishing the functional paradigm, and applying the paradigm to the study of the (positive) contributions of crime/punishment in establishing and maintaining social order and cohesion in society.

However the societal reaction perspective did not gain traction until over half a century later in the 1960s, where it was redirected away from functionalism and macroanalysis toward microanalysis with a focus on social process rather than social structure. This shift in focus was largely the result of the very strong influence that both symbolic interactionism and the Chicago School had on the field of deviance and redirected the societal reaction perspective toward microanalysis and a focus on social interaction. The works of early pioneers such as Tannenbaum (1938), Lemert (1951, 1967), Becker (1963), Goffman (1961, 1963), and Scheff (1966) brought labeling theory front and center to a position of preeminence in the field. Labeling theory generated much interest, research, and excitement in the field and continued to dominate the study of deviance for almost the next quarter of a century. It gave rise to a variety of theories and research studies which redirected the traditional positivistic study of deviance toward new issues and concerns. Though its popularity began to wane by the 1980s, its legacy remains a vital perspective today in the study of deviance. Since the late 1990s there has been a resurgence of articles and books focusing on evaluating labeling theory, reinterpreting and refining the theory, and applying it to new areas of study. Labeling theory is thoroughly embedded in the fabric of deviance theory today. Several outstanding recent histories of labeling theory have documented the resurgence of deviance theory (Plummer, 2011; Matsueda, 2014).

There also is no all-encompassing labeling perspective. Because it developed in a piecemeal fashion, in a variety of areas, focusing on different issues, its relationship to the larger body of sociological theory, such as societal reaction theory or symbolic interactionism, has not been clearly explicated. In addition to its macro and micro focuses, it also assumed quite different forms as it was applied to different types of deviance and to different facets of the processes of deviance. It did not evolve into a coherent body of theory, incorporating both macro and micro concerns, integrating the disparate and seemingly unrelated studies.

One of the first attempts at integrating the disparate strands of many diverse studies which sought to bring them together in a more unified way was an iconic text by Earl Rubington and Martin Weinberg (1968) which described this emerging body of theory and research as the 'interactionist' perspective. It undertook efforts to incorporate both the more narrow micro concerns of labeling theory, focusing on the effects of labeling on the individual, while at the same time including research studies that were broader in scope than what might have been typically regarded as labeling theory. Studies of institutions of social control such as the court, police, prisons, and social work agencies as well as studies dealing with the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03161-5

emergence of subcultures and social movements began to dot the landscape. These studies, however, required a broader framework than what the term 'interactionist' perspective might convey. Therefore, a more encompassing conceptual framework is needed to contain the range of concerns in this array of theories and studies. However, the effect of symbolic interactionism on societal reaction theory was to keep the focus on the micro and social psychological aspects of the labeling process. Whereas Durkheim focused on the macro aspects of societal reaction in maintaining the social system as did later scholars Erikson (1966) and Moral Panic theorists (Nachman and Goode, 1994), who focused on the larger implications of societal definition and reaction. The more generic 'societal reaction' perspective, which subsumes both the narrower micro concerns of labeling theory as well as broader macro perspectives, will be employed as the primary framework in this analysis.

The rise or fall in popularity of various paradigms in the study of deviance often reflects their popularity in sociology generally. In both mainstream sociology as well as deviance theory, the popularity of various paradigms is influenced by the tenor of the times. Times of social unrest usually bring conflict theory into prominence, as it did in the late 1960s and 1970s, when cities were aflame and blood was running in the streets. Functional models of society often fade from favor as they do not reflect the circumstances of those times. Whereas during periods of relative calm and prosperity or increasing cultural diversity such as the late 1950s and 1960s, symbolic interactionism and images of negotiated social orders rise to prominence, as they mirror the social conditions of the times. Therefore, changing times often lead to changing paradigms within a discipline. These short-lived cycles, in deviance theory have been noted as 'generational effects' (Plummer, 2011) or paradigm shifts (Kuhn, 1970).

Symbolic interactionism left its substantial mark on labeling theory. In addition, the orientation of the labeling theory got continually refocused as it was applied to different forms of deviance, such as crime, delinquency, mental illness, drug use, gangs, sexual orientation, or disability. This fragmentation of labeling theory became so pronounced that some critics did not view it as a theory but only a loosely tied together body of propositions which acted more like a sensitizing perspective than a theory.

Nonetheless labeling theory dramatically shifted attention from the traditional positivistic concerns of viewing deviance as behavior which violated norms and seeking the social causes of rule-violating behavior. While employing the same term, *deviance*, they radically redefined both the term and the field of study. They asserted that what deviance fundamentally refers to is *actions* or *people* that elicit *social disapproval, condemnation, rejection, or exclusion*. Furthermore, they asserted sociologists can never deduce in advance what will be sanctioned by idealized versions of the norm. Norms were reconceptualized from 'objectively determinable' to 'subjectively problematic' (Rubington and Weinberg, 1968). Moral norms must be studied in action as they are created, invoked, or applied in everyday interaction, and that you can never tell what the norm is until people are actually negatively reacted to by others and this is highly variable (Becker, 1963). It is other's reaction to them that makes them deviant, not the act itself. Furthermore, people can elicit social condemnation without violating social norms, and such individuals are still to be regarded as deviant from this perspective. This is why *positive deviance or hidden deviance is* an oxymoron. It is not the act that the person engages in that makes them deviant. It is only the reactions of others that confer their deviant status upon them, no matter their actual behavior. Hence the meaning of the term deviant shifts from *behavior* in which the individual engages, in the positivist perspective, to their *status* in the group and how they are defined by others in the labeling perspective. That is determined by the reactions of other members in the group. Thus the concerns and research efforts of labeling theorists were then directed to *the social definitions and causes of labeling as well as the consequences of labeling especially on the individuals' subsequent behavior, self, identity, and possible career of deviance*. Though the variable nature of applying rules and labels to individuals was commonly investigated, the variable responses to the labeling by prospective deviants, was less well examined, with few exceptions (Goffman, 1963).

Labeling theory also shifted attention from the harmful effects of deviance on society, frequently presumed by positivists, to the harmful effects and consequences of official control and labeling on the lives of individuals who are designated deviants by society. In addition, the emergence of labeling theory served to raise questions about value positions which are implicit in particular ways of conceptualizing and studying deviance (Mills, 1943) by incorporating the perspectives of deviants, as Becker's (1964) title, *The Other Side*, aptly conveys. It gave a type of legitimacy or validity to the perspective of deviants. Including the perspectives of deviants also bears the mark of the Chicago School's emphasis on obtaining rich ethnographies on the lives of deviants giving them a seat at the table in the dialogue. There are even current pleas (Fitch, 2010) urging reconnecting labeling theory to its early roots to Dewey by reemphasizing that isolating, stigmatizing, and condemning individuals creates an underclass contrary to democratic ideals. The morality of labeling cuts both ways. Other aspects of labeling theory also reflect the strong influence of symbolic interactionism, particularly theorists such as Mead (1934), as they shaped a view of social life as fluid, reality as socially constructed, meaning and order as negotiated, and took a less deterministic nature of social interaction with a focus on social process rather than on social structure in place of the stricter determinism characteristic of functional/positivistic approaches.

Societal Reaction Perspective and Labeling Theory

Central to sociological analysis is 'part to whole analysis' and in this vein it would be helpful to place labeling theory in a larger contextual perspective. It is asserted that labeling theory cannot be fully appreciated apart from its connection to societal reaction theory and that it can be viewed as a subbranch of societal reaction theory, which includes micro, meso, and macro levels of analysis. Furthermore, the various levels of the perspective are intimately intertwined. Societal reaction theory can be viewed as comprising three streams of interrelated studies and theories: (1) microanalysis encompassing studies dealing with the traditional concerns of labeling theory, (2) mesoanalysis exemplified by studies dealing with larger focuses and

structures such as agencies of social control, professions, social movements, etc., and (3) macroanalysis dealing with the interconnections of labeling processes to the larger social structure and social system, perhaps even extending to the world order and global processes related to patterns of social control.

1. Labeling theory would fall into a more narrow focus on the micro aspects of societal reaction theory such as consequences of labeling on the individual, their self-concept, identity, life chances, and subsequent deviant actions and entrenchment into deviant careers.

Tannenbaum (1938) was among the first to redefine the concept of deviant and redirect study to the effects of labeling and how it creates what Lemert (1967) calls 'secondary' deviance; deviance that is a product of the labeling process itself. Whereas positivists had focused their attention on the initial rule-breaking act or primary deviance. The labeling apparatus could be viewed as a system of amplification of deviance, which paralleled concerns of mainstream sociologists with the process of socialization, more aptly described as resocialization. Becker's (1963, 1964) focus was more on the creation and application of the rules by moral entrepreneurs and enforcers, the various contingencies involved in labeling, and the importance of social definitions. The labeling approach led to a reconceptualization of norms as 'objectively determinable' to 'subjectively problematic' (Rubington and Weinberg's, 1968). Becker (1963), in his study of marijuana use and Scheff (1966, 1984,1999) in his analysis of mental illness, both highlighted the importance of learning the deviant role, which further targeted the amplification process of elaborating and stabilizing deviance. Reflecting similar concerns, Goffman in *Asylums* (1961) elucidated the systematic efforts of institutions of social control to transform individuals' identities into deviant ones, in order to more effectively manage them in a bureaucratic system. This work addressed social control institutions and more encompassing institutions he described as 'total' institutions. His *Stigma* (1963) called attention to the central place of stigma and the critical role that social acceptance plays in the study of deviance, not only in the lives of deviants but in the underlying threat of embarrassment and loss of face that everyone faces at every moment in everyday interaction. All individuals share with deviants, though in a lesser degree, the experience of managing stigma. When social rejection becomes more extreme and spills over into other roles and situations, it reflects the terrain of deviance. Normals can usually shed the devaluation at the termination of the interaction sequence as it does not usually follow them into new situations as it may among deviants. Thus the 'meaning' as a deviant the person comes to have for others is constructed through social interaction by the treatment the individual receives, which is critical to creating deviants. The meaning is revealed in the process of social interaction where the individual is constructed out of the actions and reactions of others.

This tradition was often characteristic of subsequent research in criminology where a plethora of studies emerged with their focus on the consequences of formal systems of social control on recidivism. Here the findings show somewhat consistent effects of labeling and the formal processing of criminals on subsequent law-breaking behavior and reincarceration. Some studies, however, showed more inconsistent effects and suggested reformulation of the theory by specifying the conditions under which labeling is likely to be effective, in order to more adequately test labeling theory (Paternoster and Iovanni, 1989).

More recent studies extend the theory to new areas such as gambling, victim perceptions, and drug use. In addition, they explore differences in formal and informal labeling, the varied conditions under which labeling takes place, the type of sanctions imposed, age at which labeling occurs, the duration of the effects of labeling, areas impacted by labeling, as well as differences in the effectiveness of labeling on various groups differing in gender, age, ethnicity, etc. The studies showed significant differences and point to the necessity of further research characterized by more refined formulations more adequate controls over these variables. Efforts were undertaken to identify how the effects of labeling are mediated, whether through (1) identity change or (2) cumulative structural disadvantage through the exclusionary process of labeling. Both contribute to explaining the variance. More specification is needed of the types of deviance, sanctions, and conditions under which labeling takes place, so a more precise understanding of how the process of labeling influences outcomes. The introduction of more sophisticated techniques such as modeling theory has increased the explanatory power of labeling theory in delinquency (Hayes, 1997). The research and the concerns raised by labeling theory, despite the lack of highly controlled studies, has led to an increased use of alternatives to traditional methods of social control in society such as drug courts, home monitoring, alternatives to incarceration, and techniques of radical nonintervention (Schur, 1973).

This pattern of research was also characteristic of studies in the area of mental illness where research explored the consequences of labeling on symptom remission, rehospitalization, and the stigma the individual experienced, though more inconsistent effects were found with respect to symptom remission and rehospitalization (Scheff, 1967, 1999; Gove, 1980; Link et al., 2001). Alternatives to hospitalization such as deinstitutionalization were undertaken on a mass scale in states like California, though these programs failed to provide help for the individuals in the community which ultimately led to an increase of the mentally ill among the homeless and now prisons and jails have become the largest mental hospitals in the nation. Self-help and twelve-step programs have arisen around an enormous range of 'addictions' reflecting some deprofessionalization of the management of deviance countering somewhat the trends toward the medicalization of deviance. Studies in the area of mental illness also urge a more thoughtful reformulation of the theory in order that more precise predictions can be made.

The most recent direction that labeling theory and microanalysis has taken, spearheaded by the work of Link et al. (1989, 2001), has been under the terms of stigma and stigmatization. Stigma studies tend to show consistent negative effects of labeling and have been linked to self-esteem, earning power, ties in the community, psychological symptoms, and satisfaction with life (Markowitz, 1998).

Plummer (2011) noted that while many current researchers do not explicitly identify with the labeling perspective or locate the relevance of their findings to the tradition, their findings fall

squarely in labeling theory paradigm. Ordinary sorrow is transformed into depressive disorder that illustrates medicalization of deviance as life history criminology, which relates subsequent criminality to involvement in the criminal justice system is old wine in new bottles. Also one cannot ignore the impact of labeling theory on other disciplines such as psychology, psychiatry, social work, special education, etc. (Corrigan, 2004). Numerous studies raise serious questions about the nosology and types of mental illness of the field of psychiatry, based on the Diagnostic Statistical Manual, calling into question not only its reliability but also its validity. The goodness of fit between these fundamental diagnostic categories and the symptoms manifested by individuals is very poor, lending credence to Scheff's claims and affirming social constructionist's interpretations of the term 'mental illness.'

All the above studies and efforts fall into this more narrow focus on societal reaction traditionally associated with labeling theory.

2. Mesoanalysis attempts to link labeling perspectives with larger structures in society such as agencies of social control, social movements, and the emergence of subcultures or professions.

More encompassing studies on labeling processes and societal reaction take into account the creation of social types, the consequences of large formal agencies of social control (Goffman, 1961) such as courts, police, prisons and apply these perspectives to other bureaucratic systems such as school systems and the tracking of students or social work agencies. These studies investigate social processes and systems with a much broader scope than just focusing on the effects of labeling. Studies in cultural criminology, public reactions to crime, and the role of media in shaping public responses to crime also reflect mesolevel concerns of societal reaction theory. Developing an understanding of how these institutions operate as social systems, such as the "Theory of Office" (Rubington and Weinberg, 1968), as well as their role in the larger society in managing deviance also demonstrates linkages to larger structures in society sometimes edging the labeling perspective toward macroanalysis.

In mesoanalysis, this social 'meaning' key to the term deviant is made more structural by introducing the concept of *status* as defining the deviant, and along with the concept of 'master status,' deviant status is placed in a larger structural context of analysis or in Goffman's terms a larger frame. Taking it further, deviance can be seen as a form of social stratification based on 'status groups' in Weber's terms referring to the prestige, respect, or dignity of the person so as to make social acceptance possible through status claims. Celebrities would form the highest tier, ordinary people an intermediate tier, and deviants fall at the bottom of the hierarchical system. Therefore deviance can be studied within a structure of stratification and inequality of prestige or social acceptance in the same way other systems of inequality are examined in mainstream sociology.

Mesolevel research also examines the effects of labeling systems on both the deviants and on other institutions in society. This tradition of social constructionism has embodied the study of whole professions such as psychiatry, medicine, and the pharmaceutical industry and their roles in medicalizing deviance. Some of the richest studies on professions have emerged from studies in the field of deviance. These studies integrate 'structural' aspects of large-scale organizations and institutions with micro level interpersonal and individual effects of managing deviance.

3. Macro concerns in societal reaction theory had its roots in Durkheim (1895), and reemerged in the 1960s with the resurgence of labeling theory. Aspects of macroanalysis can be found in Erikson (1966), Scheff (1966, 1984, 1999) where the larger functions of labeling deviance or mental illness were examined, and in Moral Panic analyses (Nachman and Goode, 1994) which focused on larger societal reactions. Studies that examined aspects of deviance as precursors to social movements such as gay pride or the discrimination associated with mental illness, like the emergence of these movements were efforts to cope with deviant status in society. These studies and theories embodied macro perspectives.

Labeling concerns are also linked to macroanalysis by including the society's system of social types in their analyses. The boxes or categories within which individuals are placed are larger societal constructions. Furthermore, Scheff's (1966, 1999) analysis of mental illness as a 'residual' category (placing all forms of deviance that do not fit other established types or categories), permits the larger structural system's integrity by coping with the range of unacceptable behavior that does not neatly fit into the conventional system of social types, and ties his analysis to macro concerns of system maintenance. Furthermore, Scheff's (1966, 1999) assertion that the function of labeling is to (1) protect the status quo and (2) preserve the underlying social reality, also speaks to the issues raised by Durkheim in his macroanalysis of deviance. Less well explored, but equally critical to a larger macroanalysis, are Scheff's and Goffman's notion of psychiatric symptoms reflecting the *taken-for-granted residual rules of society*. They are a reflective prism on the assumptive world of that social system. Thus psychiatry becomes a way of policing a certain part of the moral terrain. This extends Durkheim's macroanalysis of crime to mental illness. For Durkheim creating criminals is a way to define moral boundaries and similarly for Scheff, labeling people as mentally ill is a way also to define moral boundaries of the taken-for-granted world.

Another stream of influence flowing from societal reaction perspectives arose as the similarity between the plight of deviants and other oppressed or stigmatized groups came to be recognized in social movement theory. The connecting of gay rights, the disability community, prisoner rights movements, etc. were seen as instances of civil rights struggles and the larger issues of equality and civil rights came to the forefront forging an alliance between larger concerns of social acceptance and the power of the state. Increasing convergences between issues studied by deviant theorists and those in mainstream sociology can be observed. The plight of deviants and threats to their civil rights arose along with other stigmatized and oppressed minorities became an increasingly important issue in recent times. This was foreshadowed by Durkheim's analysis of the growing importance of the individual and individual rights in modern society as the sphere of the collective morality diminished

with industrialization. Parsons pushed the analysis of modernity further and viewed increasing democratization as a consequence of modern society. Thus these movements may reflect a larger transformation to attain dignity and rights for all in society even among the lowest of the low, social outcasts.

Within this macro tradition are also studies demonstrating relationships between labeling systems and macro structures. For example, there has been research linking increasing rates of incarceration to the creation of prisons as the new Jim Crow system. Some view increased incarceration as a result of the later stages of declining empires linking labeling processes to the larger world order. Thus labeling may be driven by much larger macro concerns than deviance theorists have yet explored. Conflict/radical theorists have sought to link increasing incarceration to the late stages of capitalism as a way of coping with the bourgeoning unemployment. Studies investigating the current decriminalizing of marijuana bear directly upon Durkheim's hypothesis that societies seek to maintain 'optimum' amounts of crime through efforts at loosening boundaries when there are excessive numbers of deviants, as was the case before the repeal of prohibition in the U.S. Current mass releases of inmates from prison also reflect society's efforts to balance the creation of criminals with social resources. In addition, investigations of efforts by society to create new categories of deviants such as terrorists and extending that label to eco-terrorism or animal rights terrorism links labeling processes to efforts of state or corporate control to employ labeling as a means of extending their power and control in society. Thus issues of societal reaction fit neatly into aspects of conflict/radical perspectives of deviance as well and merge easily into macroanalysis.

Falling squarely in the macro tradition were investigations and studies which explored variations in labeling processes by type of society (Pfohl, 1981; Raybeck, 1988) which focused specifically on macro–micro relationships between narrower concerns of labeling and social control at the micro level and broader issues of social organization societal reaction, with an emphasis on macro–micro integration.

These trends in research and theory illustrate the various strands of labeling and societal reaction theory and the range of societal reaction theory. Overall the trends in research over the last half century has been moving toward expanding the range of societal reaction theory, identifying convergences in different areas of deviance, and proposing refinements in the theory by specifying the relevant variables and types of controls necessary to isolate the effects of various systems of labeling. There are calls for efforts to integrate the various strands of societal reaction theory into a more encompassing and coherent perspective and designed research to examine the overall system of social control.

Not only may it be impossible to separate aspects of labeling theory as issues seem to be strongly interrelated. It would be artificial to separate them as independent lines of inquiry without seeing their interconnection. It may be true that separating deviance theory from general sociological theory may prove increasingly unfruitful since the problem of deviance (disorder) is the opposite side of the coin of the 'problem of order' which has been a foundational concern since the inception of the discipline. A coherent sociology will attempt to unify instead of fragment the field into subdisciplines. Separating rather than integrating disciplines may result in fracturing our grasp of the world. Each science is a strategy of abstracting from a complex world. Otherwise everything would have to be said at once to capture the immense complexity of the world. On the other hand, by separating disciplines, we lose the interconnectedness of nature.

If labeling theory is to advance today, it must consolidate a variety of disparate concerns into a more coherent overall theory. Currently it sensitizes researchers to specific issues in the study of deviance without integrating its various facets showing the relationships among the diverse concerns on both a macro and micro level that have been embraced by labeling theorists in order to move from a perspective to a theory.

Deviance theory has not developed as a cumulative endeavor where each research study builds upon previous studies and theories are modified on the basis of empirical findings, so there is a close correspondence between the state of research and theory. If the field is to advance significantly the various strands of labeling theory, positivism, and conflict theory have to be pulled together within the various subareas, and then in turn has to be integrated into mainstream sociological theory.

Acknowledgment

The author acknowledges helpful comments made by Thomas J. Scheff.

See also: Applied Criminology and Criminal Justice; Constructivism and Constructionism: Methodology; Individuals in Society: History of the Concept; Social Constructivism; Social Control and Deviance, History of.

Bibliography

Becker, Howard S., 1963. Outsiders: Studies in the Sociology of Deviance. The Free Press, New York.

Becker, Howard S., 1964. The Other Side: Perspectives on Deviance. Free Press, New York.

Corrigan, Patrick, 2004. How stigma interferes with mental health care. American Psychologist 59 (7), 614–625.

Durkheim, Emile, 1895. Rules of the Sociological Method. Felix Alcan, Paris, France.

Durkheim, Emile, 1897. Suicide: A Study in Sociology. Felix Alcan, Paris, France.

Erikson, Kai T., 1966. Wayward Puritans: A Study in the Sociology of Deviance. Wiley and Sons, New York.

Fitch, Frank, 2010. Laggards, labeling and limitations: re-connecting labeling deviance theory with Deweyan pragmatism. Philosophical Studies in Education 41, 17–28.

Goffman, Erving, 1961. Asylums: Essays on the Social Situation of Mental Patients and Other Inmates. Doubleday, New York.

Goffman, Erving, 1963. Stigma: Notes on the Management of Spoiled Identity. Prentice Hall, Englewood Cliffs, NJ.

Gove, Walter R., 1980. The Labeling of Deviance: Evaluating a Perspective, second ed. Sage, Beverly Hills, CA.

Hayes, Hennessey D., 1997. Using integrated theory to explain the movement into juvenile delinquency. Deviant Behavior 18 (2), 614–625.

Kuhn, Thomas, 1970. The Structure of Scientific Revolutions, second ed. University of Chicago Press, Chicago.

Lemert, Edwin M., 1951. Social Pathology: A Systematic Approach to the Study of Sociopathic Behavior. McGraw Hill, New York.

Lemert, Edwin M., 1967. Human Deviance, Social Problems and Social Control. Prentice Hall, Englewood Cliffs, NJ.

Link, Bruce G., Cullen, Francis, Struening, Elmer, Shrout, Patrick, Donrenwend, Bruce, 1989. A modified labeling theory approach to mental disorders: an empirical assessment. American Sociological Review 54 (3), 400–423.

Link, Bruce G., Struening, Elmer L., Neese-Todd, Sheree, Asmussen, Sara, Phelan, Joe C., 2001. Stigma as a barrier to recovery: the consequences of stigma for the self-esteem of people with mental illnesses. Psychiatric Services 52 (12), 1621–1626.

Markowitz, Fred E., 1998. The effects of stigma on the psychological well-being and life satisfaction of persons with mental illness. Journal of Health and Social Behaviors 39, 335–347.

Matsueda, Ross L., 2014. The natural history of labeling theory in Farrington. In: David, P., Murray, Joseph (Eds.), Labeling Theory: Empirical Tests. Transaction, NJ.

Mead, George H., 1934. Mind, Self and Society from the Standpoint of a Social Behaviorist. University of Chicago Press, Chicago.

Mills, C. Wright, 1943. The professional ideology of social pathologists. American Journal of Sociology 49 (2), 165–180.

Nachman, Ben-Yehuda, Goode, Eric, 1994. Moral Panics: The Social Construction of Deviance. Blackwell, Cambridge, MA.

Paternoster, Raymond, Iovanni, Leeann, 1989. The labeling perspective and delinquency: an elaboration of the theory and an assessment of the evidence. Justice Quarterly 6 (3), 359–394.

Pfohl, Stephen, 1981. Labeling criminals. In: Ross, Laurence H. (Ed.), Law and Deviance. Sage, Beverly Hills, CA, pp. 65–97.

Plummer, Ken, 2011. The labelling perspective forty years on. In: Peters, Helge, Dellwing, Michael (Eds.), Langweiliges Verbrechen (Boring Crimes). VS Verlag, Webaden.

Raybeck, Douglas, 1988. Anthropology and labeling theory: a constructive critique. Ethos 16 (4), 371–397.

Rubington, Earl, Weinberg, Martin S., 1968. Deviance: The Interactionist Perspective. Macmillan, New York.

Scheff, Thomas J., 1966, 1984 second ed., 1999 third ed. Being Mentally Ill: A Sociological Theory. Aldine, Chicago.

Scheff, ThomasJ., 1967. Mental Illness and Social Process. Harper and Row, New York.

Schur, Edwin, 1973. Radical Non-Intervention: Rethinking the Delinquency Problem. Prentice Hall, Englewood Cliffs, NJ.

Tannenbaum, Frank, 1938. Crime and the Community. Columbia University Press, New York.

Labor, Division of

Wolfgang Littek, University of Bremen, Bremen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Labor (or work) denotes an activity necessary to secure human existence. Division of labor refers to the separation of tasks and the specialized allocation to different workers. It is an active force in social structuring, economic development, and the everyday experience of individuals. Whereas changes within premodern societies have been slight, the rise of capitalist industrialization has caused a tremendous expansion of the detailed division of labor in large-scale manufacturing accompanied by a complex social division of labor in society. Various forms as well as the dynamics of structural change of the division of labor are discussed. An overview of main theories concludes the article.

Division of labor, which refers to the system of differentiation and specialization of work tasks, is a feature of social structuring found in all human societies. In its forms it varies widely and changes over time. The division of labor creates interdependencies between the single specialized actors. Families, enterprises, and organizations, as well as national economies and international trade relations, are seen as structures of division of labor. Whereas simple societies show very little differentiation into specialized tasks or occupations, some highly developed societies of today discern among more than 40 000 different occupations, with additional specializations on top. It is helpful to distinguish between a narrow, technical-economic understanding and a broader, social understanding of 'division of labor.'

The article will briefly consider the complexity of the concept by looking at different levels and forms of the division of labor and at its role in the development of human history. The main attention will be given to the division of labor in modern industrial-capitalist societies. It will become clear that the division of labor underlies a permanently ongoing structural change, which implies that by no means may it be considered a static concept.

Labor

Labor or work refers to the execution of tasks that serve a useful purpose both to those performing them and to society at large to secure their existence and preserve the environment. Only humans perform work in the sense of using creativity, conceptual, and analytical thinking together with manual abilities to transform nature or to be of use to society. A distinction generally is made between work and labor. Work is connected more with positive aspects. In conjunction with creativity and interaction, it is seen as being a human need and as forming a basis for self-respect. The word 'labor' is derived from the Latin (laboris) and implies toil and distress. Hence, labor alludes to arduous work, done under some duress, and often under control of others (Standing, 1999; *see* Work, Sociology of). But the differential use of these terms is in fact not that strict. With regard to the specialization of tasks, however, it is common to speak of 'division of labor,' and not of 'division of work.'

Divisions of Labor: Concepts, Levels, and Forms

The fact that no society exists without at least some division of labor does not imply that this is caused by natural differences (such as biological differences between women and men, by race or age, or by differing personal aptitudes, for example). Any existent division of labor is always human-made and its forms are socially shaped. Throughout human history the division of labor has grown increasingly diversified, but only rather recently has it reached the complexity we are familiar with due to the growth of production and trade, the industrialization process, and the development of modern societies with all their specific institutions.

Concepts

It is necessary to recognize that for 'division of labor,' two different concepts are in use:

1. In a narrow sense, the concept is used in an economic or technological context and describes the breakdown of a complex productive task into a number of detailed and simpler tasks. This is known also as 'detailed' or 'technical division of labor.' The most renowned example in social sciences is that of Adam Smith (1776) for pin needle production. By splitting up the traditional holistic pin production by one craftsman into particular subtasks for different workers, such as cutting the wire, making the head, the body, packaging, etc., a sharp increase in output was reached. This sharp increase in productivity was the ultimate reason for the separation and specialization of tasks. It marks the introduction of modern manufacturing and led to the development of large industries.
2. Much broader is its use in a social context meaning because the labor performed by a person reflects his or her social standing in a group or in society. Recognized here is the connection of the division of labor with social differentiation. On the societal level, everyone is affected by the great differences in social esteem, economic power, and the distribution of wealth due to his or her standing in the existing division of labor. Any achieved social division of labor represents a result of the 'struggle for survival' in which individuals are unevenly successful (*see* Inequality, Social; Differentiation: Social). This social concept includes the forces of social cohesion or disintegration in connection with

the division of labor, or the importance played by solidarity and morale in a society (this aspect was first, and powerfully, developed by Emile Durkheim (1893/English 1984).

In a schematic form, the two concepts may be summarized thus:

Detailed division of labor	*Social division of labor*
Division in connection with *economic or technological reasons* (Based on society-specific institutions for work; resulting in separated subtasks, dependency, necessity of coordination)	*Division existing along social criteria* (Based on gender, power, skills, cultural laws, etc.; resulting in social differentiation, social inequality, necessity of social integration)

Levels

Clearly, 'division of labor' is a complex term and also may refer to different levels of human activity. The division can extend from the household or family work on the microlevel, through work organizations on an intermediate (meso-)level, to divisions in society at large on the macrolevel, and finally to the entire world on the global level. Examples of divisions of labor on the various levels are the domestic division of labor; the organizational division of labor in and between workshops, enterprises, or administrations; the occupational or regional division of labor in a country; the division of labor in production networks stretching nowadays often across national boundaries; and finally, the international division of labor, where economic globalization is affecting workers in the developing as well as the developed world.

In schematic form, the levels may be differentiated as in this overview:

Levels of the division of labor			
Microlevel	*Mesolevel*	*Macrolevel*	*Globallevel*
For example, domestic or familial division of labor	For example, division of labor in organizations, enterprises, administrations (or between them)	For example, occupational, regional, product-related, or national division of labor	For example, international division of labor, international trade

Forms

Furthermore, different forms, or types, of division of labor are to be recognized. Certainly, to the oldest forms belongs the sexual or gendered division of labor, that is, the division of tasks between women and men. This reaches far beyond the household and the family. It is clearly of importance in the area of work but is recognizable as well in social, political, cultural, and religious functions. Even in most modern societies, more or less discrimination exists by which women do not have equal access to functions in these fields (*see* Gender, Economics of).

Other forms of division may be based on age, ethnicity, heritage, or skills. Or they may be due to the size of workshops, to regional expansion of markets, to national characteristics, or to rules in international trade and diversification. The importance of specific forms has changed very much in the history of human societies, and many forms have newly developed.

Here is a schematic overview of the (more important) forms of the division of labor:

	Forms of the division of labor
Division of labor by	Sex (or gender) Age Ethnicity Occupation Skill Hierarchy Heredity Specialization and organization in the workshop or enterprise Space or region Vertical or horizontal specialization within a society International trade

The Need for Coordination and Cooperation

Specializations and subdivisions in work tasks, however, create the necessity to 'unify' the fragmented and specialized subtasks again in order to achieve an integral result. Any fragmented subtask alone in a division of labor makes no sense for the larger social context. The more the tasks have been fragmented, the more they become dependent on coordination and cooperation. This 'unification' is reached differently at the different levels or at the various forms of the division of labor.

In the domestic division of labor, this 'unification' is reached by reciprocity within the family. With the upcoming of a number of specialized occupations in earlier societies, coordination was achieved by exchange of the various products. As a medium for this exchange, markets developed. They again gained enormously in importance with the invention of money as a generalized exchange medium.

Differently achieved is the coordination of divided subtasks in large-scale manufacturing, namely by entrepreneurial planning, which is based on domination and hierarchy. Between the various enterprises (and most other actors) in an economy, coordination is achieved again by free markets. (Only in the case of former socialist countries, the coordination was done by national planning.) In the international division of labor, the unifying process is achieved by foreign trade (which itself may be based either on free exchange or on coercion, as in colonialism).

In sum, the examples show that, basically, there exist two principles to gain a 'unification' for the subdivisions into partial work tasks: these are either the 'market' (with free choice) or 'planning' (based on domination or coercion).

Division of Labor and Socioeconomic Development

Viewing the long sweep of human history, the contrast between very simple and vastly extended complex social structures springs to mind. The extent of social differentiation is closely connected to the extent of the division of labor. Both accompany the shifts toward more complex societies.

Very little change occurred during the longest period in the history of humanity, which began almost 250 000 years ago with foraging as the means to secure survival. Most likely only some division of labor by sex and age existed. The agricultural revolution (between 8000 and 5000 BC) brought about a first important change. Humans transitioned from a nomadic to a settled lifestyle as farmers. A division of labor was prompted to store and provide food between growing seasons. This led to the rise of a leisured upper class and the creation of the first cities as places for exchange. The production of a surplus beyond self-provisioning then gave room to the emergence of specialized, artistic occupations (*see* Societies, Types of). Based on specialization and the concentration of power and artisanal aptitudes, the first 'high cultures' with some extension in the division of labor developed in some parts of the world, e.g., Mesopotamia, Egypt, later China, Middle and South America.

In addition, for a long time a large part of the world comprises societies characterized by rather little division of labor, as the overwhelming portion of productive tasks are performed as self-provisioning by farming. The family in these societies is also the productive unit for survival. With some development in agrarian cities, only a small part of the population is engaged in specialized occupations such as crafts, trading, religion, or warrior functions. Tradition is of eminent importance, and rarely are there any changes in the overall division of labor over long periods of time.

However, no linear progression occurred in human history toward ever greater complexity, no continuous evolution from simple to modern societies. Not infrequently, the division of labor was based on coercion and power. Unfree labor, as in slavery, was widespread in order to pursue labor tasks. Slavery, a system under which people are treated as property to be bought and sold and are forced to work, is still in existence in some parts of the world and was eradicated only in the late nineteenth century in several countries, including the United States (*see* Slavery as Social Institution). Feudalism in Europe and Japan during the Middle Ages was another system with division of labor based on coercion and power. Such societies were characterized by a generally accepted cultural legitimation for existent inequalities. The distribution of functions in the labor process in these societies is based on 'ascription' (being born into one's role and status), not on personal 'achievement.'

In sum, relatively static forces were dominant in the division of labor for thousands of years in human history. This came to an abrupt end when capitalism and industrialization in Europe began to transform the world (beginning in England in the early eighteenth century) (*see* Capitalism: Global; Industrialization). The last 300 years alone have seen by far more changes in the division of labor than all previous ages. Compared with the development up to that time, industrial-capitalist transformation marks a quantum leap for the division of labor. It rightly was labeled '*The Great Transformation*' in a famous book by Karl Polanyi (1944[2001]).

The Division of Labor in Modern Societies

The Institutional Frame

The capitalist-industrial mode of production was established on the basis of a number of institutions that previously did not exist. Along with the Enlightenment, a civil society was forming, empowered not least by the American and the French revolutions. The idea of a free and responsible individual spread. As a result, the notion that one person could be the owner of the labor power of another person (as in slavery) lost its legitimacy. Labor became now executed as free labor, at the disposition of the individual himself/herself. Labor now turned into a commodity: labor power is traded on labor markets, and the resulting division of labor is based on free will. This is so common to us nowadays in societies functioning under capitalist principles that it might seem to be natural to many people. But it is only a very recent invention of mankind.

Of course, not all work, not all division of labor, underlies capitalist principles even in today's modern societies. For the working person, the social relations and the social orientations in which the tasks are embedded are important. Aside from wage labor in the private sector of the economy, domestic labor and communal labor play a significant role as well. And, people in fact do work for a whole range of rewards and incentives, not only measured by money (Pahl, 1988).

However, it were capitalist principles and state regulation that caused the significant change from self-provisioning or artisanal labor to modern industrial and postindustrial working conditions. The major specific conditions relevant for the division of labor in capitalism and modern civil society, therefore, will be outlined below in short to help understand their uniqueness. (Socioeconomic literature here is abundant. Most is based, more or less, on the first grand analysts Marx (1867/English 1976) or Weber (1922/English 1968); *see* Capitalism: Global).

Separation of Household and Workplace

With the introduction of manufacturing in special premises, the domestic labor and gainful employment become separated. A previously unknown division is introduced between the public sphere of work and the private sphere of home. The function of the family (or the household) changes dramatically, as it is no longer a combined unit of production and reproduction. The division implies the separation between unpaid labor in the household and paid labor in the public workplace. This is responsible for a thorough change in the sexual division of labor, as it provides a new basis for power relations in the family and for the gender typing in employment outside. From the beginning, women are either excluded from paid work or only obtain access to less paid, less demanding jobs, but rarely to leading functions.

Concentration of Labor in Manufacture and the Factory

The assembling of great numbers of workers in one place, under single supervision, is the prerequisite to splitting up complex craft work, executed by one expert, into a multitude of detailed tasks performed by various persons. With this detailed division of labor, enormous gains in productivity are possible. This usage of labor carries a number of (economic) advantages: unskilled workers, who receive relatively low wages and are easily replaceable, can be engaged; the specific skill developed by a person in the specific subtask tends to get higher than that of a person performing the complete series of tasks; and finally, control over speed and output is easy. Fragmentation into simple repetitive functions is also the basis for the subsequent mechanization and automation, which again drives productivity.

With the introduction of capitalist manufacture, the 'separation of conception from execution' found its crystallization. However, some part of the productivity gains always are taken up by the efforts necessary for coordination, organization, supervising, and control of the partial tasks (i.e. management). Plus, a serious price has to be paid for rising productivity in this type of division of labor – namely, alienation and loss of motivation on the side of the workers in such repetitive jobs with little space for proudness and self-esteem as a 'producer' (*see* Alienation, Sociology of).

Private Property

In the process of industrialization, property acquires a new social function and is important for the division of labor in two respects. First, it provides for the owners and for those who are engaged in planning and controlling the legitimacy to enforce the detailed division of labor. Second, it provides the basis for the accumulation of wealth. Profit becomes the main motive for organizing a labor process, which was not necessarily the case in traditional societies. Capitalists ultimately only have the 'exchange value' in mind, whereas provision of a product or a service useful for the consumer ('use value,' in the terminology of Karl Marx) no longer is the main impetus for any production process. This implies that the social function and the importance of money change completely: from a generalized exchange medium on local markets to an abstract carrier of value and power that makes the world go round (*see* Money, Sociology of).

Free Labor

A labor market, where labor can be sold and bought as a commodity, was not common in human history. Quite frequently, societies in earlier times depended on the use of unfree labor (in slavery, serfdom, subjugation to a ruler). Those are the typical characteristics of a 'forced division of labor' into which persons often also got by birth. Modern development relies on formally free labor, which is free for contracting with an employer. The workers sell their labor power. This implies that social relations of a double character become relevant for wage labor: on the labor market, the relation is formally one between equals, whereas inside the workshop, the relation is one of subjugation under domination and power (Rueschemeyer, 1986; *see* Labor Supply).

Liberalization of Markets and Competition

The importance of liberalism for greatly increasing the wealth-creating capacity of a nation was strongly emphasized by classical 'political economics' during the Enlightenment of the eighteenth century. Unrestrained by administrative rules of government, a 'free market' would encourage producers to specialize in activities where they have a relative or absolute productive advantage. This implies pleading for political activity that aims at a specific design of national and international division of labor. Similarly, 'free trade' is seen to be to the benefit of all countries participating because of the advantages of comparative costs (main authors: David Ricardo, Adam Smith). Through specialization, producers benefit from greater dexterity, more efficient use of materials, time, expertise, and technology. At the same time, the invisible hand of competition penalizes ineffective producers by removing them from the market. This thinking is still powerful nowadays with the dominance of neoclassical economics almost worldwide, even though critics have widely pointed at the unrealistic assumptions inherent in the picture of the 'free' market.

Application of Science, Technological Change, and Automation

In the pursuit of profit and under the constant pressure of competition, science is systematically and permanently applied for the creation of ever more sophisticated machines and procedures to render production processes more efficient or develop new products. In contrast to traditional craft work, technological change now becomes a permanent feature of working life. Human labor that remains in the work process is forced to fill the gaps left by machinery, new technologies, and organization, and to adapt constantly to changing requirements.

Division of Labor and the Dynamics of Change

Change has remained a fundamental feature of the division of labor ever since industrialization altered occupational structures and skill requirements in an unprecedented manner. Today, ongoing changes in the division of labor on all levels, and in all dimensions, are mainly due to three megatrends (Graßl, 2000), which shortly will be dealt with: (1) the change from industrial society to service and information society, (2) changes in the gendered division of labor, and (3) globalization and international division of labor.

Change to Postindustrial Information Societies

In the last decades, highly developed societies have been leaving behind industrial production as the major field for employment. The 'service society' and the 'informational economy' have taken the ground, described, e.g., as '*The Rise of the Network Society*' (Castells, 2000). With information technologies, the occupational, regional, and international division of labor has expanded into new fields. Provision of knowledge-driven services, evaluations, information, guidance, analysis, finance, even caretaking, etc., now outnumbers industrial production work by far.

Increases in the professional level of many jobs and the professions as such gain in importance in highly developed countries. Lowly skilled, tightly prescribed work decreases in importance, and simple labor is often either relocated to less developed countries or eliminated altogether through automation.

Changes in Work Organization

Assembling people in large numbers and assigning them detailed tasks requires organizational planning to reach the common goal. As a result, organizations have become an indispensable feature of modern life in many areas, and 'organization theory' and 'management' have developed into indispensable fields of their own over the last century (*see* Organizations, Sociology of).

At the outset of industrialization, early enterprises were organized very much by patriarchic ruling principles or by the example set by the military. With the rise of capitalism, rationality and market principles grew vastly in importance. Aside from analyzing the emergence of legal rationality with the modern state, Max Weber in his '*Economy and Society*' (1922/

English 1968) worked out, more generally, the importance of rational behavior as a basis for modern administration and organizations and for modern work life in general.

Regarding production work, '*The Principles of Scientific Management*' by Frederick Taylor (first 1911, numerous reprints) then became highly influential for the improvement of industrial efficiency. At the core of Taylor's principles is the strict separation of conception (or planning work) from execution. Taylor advocated a division of labor that relies on tight control and hierarchy. All decision making about the 'one best way' of how to perform a task rests solely with management, thus destroying skills and use of their discretion on the side of the workers. Taylor's whole system based the division of labor on a low level of trust.

The industrialist Henry Ford added the disciplining power of the assembly line to the work process. Furthermore, he combined mass production with the cultivation of mass markets. Since then, bureaucratic organization, Taylorism, and Fordism have been so successful that they shaped the form of the division of labor and working conditions in industries all over the world, even in the (former) communist or state-socialist countries. (The latter, however, did not just seize the unique opportunity to realize a form of division of labor based on different, social principles, e.g., Sayer and Walker, 1992.) Many social scientists agree that the widespread application of Taylorist and Fordist principles of work reorganization led to '*The Degradation of Work in the Twentieth Century*' (as analyzed in a seminal study by Harry Braverman, 1974). The ongoing impact on working conditions in various countries is being discussed more fully, e.g., in the ensuing 'labor process debate.'

The grounds for a '*New Division of Labor*' (Littek and Charles, 1995/reprint 2011) arose from a number of far-reaching changes in the markets, a new quality of competition, and growing pressure for flexibility and for customer services. Toward the end of the twentieth century, this was labeled by some scientists as the 'end of mass production.' Cost-reduction, higher efficiency, and quality-gains were sought in 'lean production.' Also, it was found out more and more that hierarchical command structures tend to be not utterly effective in a volatile economic environment. The more it becomes necessary to trigger creativity and to improvise in the workplace, the more democratic, as a rule, cooperation becomes. The thourough digitaliszation of work life in the last decades has enormously facilitated and accelerated such processes.

Elements that have gained in importance in leading business enterprises for doing one's job are decentralization, networking, quality circles, upgraded skills, motivation, self-organization, outsourcing, and autonomy of subunits of an enterprise. Even internally, competition and market principles instead of rigid prescriptions from the hierarchy get employed to coordinate the single work tasks in the division of labor. All of these bring responsibility back to the work floor. These are conditions based much more on trust as a means of coordination of partial work tasks. Of course, and this has to be underlined, all these new forms of organizing the labor process are ultimately steered by the search for excellence of performance under the overriding goal of optimal profit-making. That creates a built-in pressure not only in the highly competitive national and international environment but often just as well within the single enterprise and its subunits.

Globalization and International Division of Labor

Today, division of labor in practically every country is also greatly affected by the globalization of economic activities. The international division of labor has changed tremendously. A 'world economy' in which capital accumulation proceeded through the world has existed in the West at least since the sixteenth century. But, historically new, since the late twentieth century, is the emergence of a 'global economy.' This 'global economy' has the capacity to function as a unit in real time on a planetary scale. It is based on the new infrastructure provided by information and communication technologies.

A number of essential elements are present. First is that capital is now managed around the clock in globally integrated financial markets working in real time, which gives it a 'hypermobility.' Second, labor has become, in several respects, a global resource: firms may choose to locate in places worldwide to find the labor supply they need (be it in terms of skills, costs, or social control); firms almost everywhere may also solicit needed skilled labor from everywhere; and labor will enter any market on its own initiative, from wherever it originates. Third, markets for goods and services become increasingly globalized. Production processes incorporate components that are produced (or designed, or sold) in many different geographic locations by different firms (e.g., Castells, 2000; Vallas, 2012).

As a result, working conditions and the division of labor have changed massively. The opening up of markets has made it possible for large corporations to relocate substantial parts of their business operations overseas, with drastic reduction of work opportunities in the home countries. Even more pronounced have been the effects on working conditions in the developing world, where female and male workers seldom receive any labor protection at all (Vallas, 2012; Mies, 1999). In sum, the spatial division of labor quasi has exploded within a few decades.

Division of Labor, Social Differentiation, and Cohesion

Social differentiation is so closely connected with the division of labor that both could almost be treated as synonymous. The growing complexity of the division of labor under modern industrialism and postindustrialism caused a similarly wide expansion of social differentiation in society. Along with this expansion, social and economic interdependence has increased enormously: Each individual in society is dependent on so many others for his or her existence (*see* Differentiation: Social).

Social differentiation itself embraces various aspects of which the more important ones shall be examined:

1. As a major element in differentiation, gender remains a cause for inequality. Even though patriarchy has lost ground and equal opportunity rights are enforced in most modern societies, the effects of a gendered division of labor are still to be found in practically all societies. These often include lower income, confinement to less demanding jobs, reduced access to leading functions, broader employment in precarious jobs, part-time work, and so on.

On the basis of social policy programs, the chances for equal employment in many countries have improved in recent

decades. But, studies have found that a multitude of factors so far prevent the arrival at definite equal opportunity (Lundberg, 2010; Vallas, 2012 – for USA). Also, most societies differ significantly in their specific 'gender culture' (Pfau-Effinger, 2004 – for Europe; *see* Gender, Economics of). This continuing actual discrimination is explained by the fact that the household continues to be the basic social unit around which people conduct their lives, and an unequal distribution of tasks in household and public employment between women and men continues to be practiced, more or less, throughout the world (e.g., Pahl, 1988; Mies, 1999).

2. Class and social stratification are likewise closely connected to the division of labor. They are relevant as a precondition, as well as a result, of the division of labor. With regard to class, most social scientists would agree that the number of classes needed to describe the social structure in modern complex societies has to be extended beyond the two basic ones (bourgeoisie and wage labor) (*see* Class: Social). Marx had in mind. Often the concept is replaced altogether by social stratification. Nevertheless, ownership of productive property is important as a precondition for the right to organize labor, or otherwise being organized. Between the higher and the lower strata, there exists a strong fragmentation from an elitist core to outright precarious jobs or even unemployment. The gains from ones economic activity, however, shape the wealth distribution. As a result of the differentiated division of labor, the inequality of the wealth distribution is increasing in most countries over time (e.g., Standing, 1999; *see* Inequality, Social; Social Stratification).
3. Finally, social integration and cohesion become much more important with the growing diversion and inequality that follow the extension of the division of labor (*see* Integration, Social). Cooperation and coordination are necessary to counterbalance the diverging forces and to secure at least a minimum of social integration (*see* Cooperation: Sociological Aspects). Social integration on the societal level is based, above all, on two institutions: trust, solidarity, motivation, and shared values on the one hand, and power on the other. (Power is based on legitimacy in modern states, not merely on coercion.) Due to the vast increase in the division of labor with the rise of modern industrial societies, a change from 'mechanical solidarity' to 'organic solidarity' as a basis for social integration has been prominently outlined by Emile Durkheim (1893/English 1984; *see* Solidarity, Sociology of). In most highly developed countries the strategy nowadays is to secure solidarity and shared values along with legitimate power: First, a more or less powerful social policy aims at providing a minimum of social security, and second, some form of social regulation is applied against too harsh effects of the market economy. Thus, a socially responsible division of labor (as an important aspect of sustainability) may be seen as the guiding vision for modern countries.

Division of Labor in Social Thought

Because the division of labor is a fundamental structuring principle of societies and has so tremendously changed especially in the last 200–300 years, it has played an important role in theories about society and social change. Interest in division of labor was particularly strong in the eighteenth and nineteenth centuries, when the unprecedented change from traditional agrarian to industrial-capitalist societies occurred. That was also the time of the formation of political economy and sociology. Later, the division of labor as a category in social theory was somewhat neglected. But, in the last few decades, interest has risen again, mainly due to

- a broad expansion of studies into women and work, the sexual division of labor, paid and unpaid employment, and the role of women in society (*see* Gender, Economics of);
- studies on the globalization of work and its new division because the world economy in general has grown increasingly global, production networks now stretching across national boundaries and continents, with effects on working conditions in literally every country; and
- studies on the increased pressure for labor flexibility and precarious jobs in the new economy, which indicate a loss of dependable employment conditions and of elements that formerly structured work like defined time, space, acquired skills or occupation, or tenure (*see* Work and Industry, Sociology of; Work, Sociology of).

In the history of theory on the division of labor, contributions from a number of key authors ought to be mentioned (*see* Division of Labor: History of the Concept):

1. Ancient Greek philosophy, and particularly Plato and Aristotle, recognized the necessity of a functional division of labor, reasoning that individuals have different natural capacities, but together are able of producing a harmonious whole. Economically as well as politically, this is of advantage for the social welfare of the state. The socializing effect similarly stood in the foreground for Aristotle when he noted that no society made up of only two physicians could emerge, but a society containing a physician, a farmer, and a few other occupations would be able to exist.
2. At the beginning of modern times, Adam Smith (1723–90) is of the utmost importance for the development of the discipline that became known as political economy. No one has ever placed more emphasis on the division of labor. But, whereas for Plato the level of specialization was externally determined (by the natural inequality of human abilities), division of labor for Smith is the most effective factor for driving economic progress. In his book '*An Inquiry into the Nature and Causes of the Wealth of Nations*' (1776), Smith explored the possibilities for raising productivity along two avenues: the detailed division of labor within the manufacturing process and the liberalization of markets in the economy as a whole. He recognized the egoistic pursuit of individual interests as the driving force of economic activity, which is coordinated by the 'invisible hand of the market.' But Smith also recognized that the breakdown of a production into specialized subtasks carried the danger of 'mental mutilation' of the workers being confined to a single repetitive task.
3. For his contemporary Adam Ferguson (1723–1816), under the impression of the upcoming new commercial civilization, social differentiation through the division of labor and

occupational specialization stood in the foreground. He considered the division of labor as a specific structuring activity for human solidarity, but also the concept of alienation is prefigured in his work. Thus, his views became influential for Marx and Durkheim later on.

4. Both the economic and the social aspects of the division of labor were analyzed by Karl Marx (1818–83). He developed much of his work as '*A Critique of Political Economy*' (this is the subtitle of his seminal publication '*Capital*,' Vol.1, 1867). In a way similar to the classical theorists, Marx elaborated the enormous potential of division of labor for productivity in manufacture and in big industry. But, although drawing on very similar insights, he came to different conclusions. He emphasized that there is a fundamental difference between division of labor among autonomous parties ('division of labor in society') and division of labor taking place under the authority and in the interest of one dominant party ('division of labor in manufacture'). The contrast between 'anarchy' of the market and 'despotism' in the workshop was highlighted by Marx as an essential feature of the 'capitalist mode of production.' He was also strongly concerned with de-skilling and alienation of workers as a result of the specific fragmentation and domination of work in manufacture.
5. Emile Durkheim (1858–1917) directed attention mainly to the sociological aspects in his '*The Division of Labor in Society*' (1893) for which above all he gained reputation. In contrast to more pessimistic views, he recognized that the fragmentation and specialization inherent in the division of labor does not have to create social disintegration. Durkheim's argument was that in small-scale societies, with only a limited division of labor, people were bound together by 'mechanical solidarity.' More complex modern societies, on the other hand, are bound together by 'organic solidarity,' as their advanced division of labor acts as the basis for much interdependency and bears a greater potential for social integration based on more individual rights. As a threat inherent in the modern development, he observed the danger of possible 'anomie' and of a 'forced division of labor' in a society.
6. In his analysis of the modern (Western) world, Max Weber (1922) (similar to Marx) worked out the specifics of the capitalist division of labor. He also emphasized the role of private property, profit as the motive for production, and the application of rational organizing principles. However, he emphasized different aspects. Weber was interested mainly in fundamental prerequisites of 'modernity' among which legitimate power and 'rationalization' of all spheres of social life (including dependable rational administration in the form of 'bureaucracy') are of importance, as contrasted to incontrollable and personalized execution of power in premodern societies.

See also: Alienation, Sociology of; Capitalism: Global; Cooperation: Sociological Aspects; Differentiation: Social; Division of Labor: History of the Concept; Gender, Economics of; Industrialization; Inequality, Social; Labor Supply; Modernity; Modernization, Sociological Theories of; Organizations, Sociology of; Slavery as Social Institution; Social Stratification; Societies, Types of; Solidarity, Sociology of; Work and Industry, Sociology of; Work, Sociology of.

Bibliography

Braverman, H., 1974. Labor and Monopoly Capital. The Degradation of Work in the 20th Century. Monthly Review Press, New York.
Castells, M., 1996/second ed. 2000. The Rise of the Network Society. Blackwell, Malden, MA.
Durkheim, E., 1893. De la division du travail social. Alcan, Paris (English: 1984. The Division of Labour in Society. Macmillan, London).
Graßl, H., 2000. Strukturwandel der Arbeitsteilung. Universitätsverlag, Konstanz (Structural Change of the Division of Labor).
Karl, P., 1944. The great transformation: the political and economic origins of our time. Boston, 1957, 2001.
Littek, W., Charles, T. (Eds.), 1995/reprint 2011. The New Division of Labour. Emerging Forms of Work Organization in International Perspective. de Gruyter, Berlin.
Lundberg, S., 2010. The changing sexual division of labour. In: Solow, R.M., Touffut, J.-P. (Eds.), The Shape of the Division of Labour: Nations, Industries and Households. Edward Elgar, Cheltenham, UK, pp. 122–148.
Marx, K., 1867. Das Kapital I (English: 1976. Capital, Vol. 1. Penguin, Harmondsworth, UK).
Mies, Maria, 1999. Patriarchy and Accumulation on a World Scale. Women in the International Division of Labor, new seventh ed. Zed Books (Macmillan), London & New York.
Pahl, R.E. (Ed.), 1988. On Work. Historical, Comparative and Theoretical Approaches. Blackwell, Oxford, UK.
Pfau-Effinger, B., 2004. Development of Culture, Welfare States and Women's Employment in Europe. Ashgate, Aldershot.
Rueschemeyer, D., 1986. Power and the Division of Labour. Blackwell, Oxford, UK.
Sayer, A., Walker, R., 1992. The New Social Economy: Reworking the Division of Labor. Blackwell, Cambridge, MA.
Smith, A., 1776/reprint 1939. An Inquiry into the Nature and Causes of the Wealth of Nations. Random House, New York.
Standing, G., 1999. Global Labour Flexibility. Seeking Distributive Justice. Macmillan, London.
Vallas, S.P., 2012. Work. A Critique. Polity Press, Cambridge, UK/Malden, USA.
Weber, M., 1922. Wirtschaft und Gesellschaft (English: 1968. Economy and Society. Bedminster Press, New York).

Labor, Geography of

Hege M Knutsen and Sylvi B Endresen, University of Oslo, Oslo, Norway
Ann C Bergene, Work Research Institute, Oslo, Norway
David Jordhus-Lier, University of Oslo, Oslo, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Labor geography has been shaped in a 'dialogue' between social theory and empirical studies that encompass the real-world experiences of contemporary labor geographers. Three themes recur in debates on this subfield: the agency of labor, constraining and enabling structures, and spatiality. In reviewing these themes, the authors present studies undertaken by labor geographers.

Introduction

As conceptualized by Andrew Herod in the mid-1990s, there are two distinctively different ways of studying labor in human geography. One takes a passive approach to the agency of labor, as exemplified in both neoclassical economic geography in the 1960s and 1970s and Marxist economic geography in the 1980s. The neoclassical approach treated labor merely as an input factor, while the Marxist approach saw workers as reactive agents or a pseudo-commodity, subject to the dynamics of capital accumulation (Herod, 1997). To emphasize their passive views of labor, Herod refers to these approaches as geography of labor. Labor geography, on the other hand, brings labor to the forefront and marks a shift in focus from understanding and representing the interests of capital to understanding and representing the interests of labor. Labor geography examines labor as an agent that acts intentionally and proactively for its own benefit and is capable of shaping the economic geography of capitalism based on its self-reproduction on a daily and generational basis. Moreover, based on the notions that labor is not only local, that capital is not only global, and that space can be a source of power, labor geography provides an approach to place, space, and scale that can explain the plight of workers and opens the way for analyses of contestation and resistance.

Labor geography is a normative discipline that has engaged a number of activist scholars. Over the years, it has widened its thematic approach to cover most significant debates in human geography (Wills, 2009).

Jonas (2009) distinguishes between two main strands of labor geography. One strand is primarily concerned with 'labor as an economic and political agent for itself,' while the other is concerned with the structures and constraints under which labor exists 'despite itself' (Jonas, 2009: p. 62), addressing how workers experience the labor market and are affected by, and respond to, economic restructuring and labor control regimes. When these strands are taken together, the main topics in labor geography are the mobilization of labor, the labor market, and the labor process. Emerging interests in embodied work and emotional labor form part of the latter. While labor mobilization clearly attracts attention to agency, the topics of labor market and labor process point to structures and constraints. However, as labor geography has developed, these themes have become integrated.

Here, the authors review themes and concepts that have characterized the development of labor geography. They first explain the agency of labor in the making of economic landscapes. Second, they review work on the constraining and enabling structures that condition the agency of labor, focusing on processes of flexibility and intensification of work; gender, class, and ethnicity dynamics; labor regimes; and global value chains and production networks. In doing so, the authors draw on geographical work preceding labor geography proper, as well as more recent efforts to view workers' agency as embedded and variegated. Finally, they highlight how geographers have used spatial concepts to develop sophisticated arguments about labor and its prospects in a globalizing economy.

Agency of Labor to the Forefront

Herod proposed that workers are in fact active agents in the making of economic landscapes (Herod, 2001a), and he has explored this proposition empirically through several studies. For example, he has shown that workers and their organizations have been pivotal in determining where economic activities such as offloading containers were to be done (Herod, 1998). In other studies, he found that local worker mobilization in a single plant could have global effects on entire production networks (Herod, 2001b). Herod and other geographers have documented that labor is an active agent in the political economy.

Most empirical studies of labor geography are case studies of particular trade union struggles. A noteworthy example is the attempt of Battersea and Wandsworth Trade Union Council's Organising Centre, a multiunion initiative in Southwest London, to move away from traditional industry-specific organizing toward a territorially defined community activism in the early 2000s (Wills and Simms, 2004). Another is Tufts' (2007) study of hotel unionism in Toronto, where Local 75 of UNITE-HERE combined local and international tactics to enforce pattern bargaining in a metropolitan hotel market. Unqualified success stories are few; nevertheless, the studies convey how a defined group of organized labor (e.g., in a local

campaign) manages to achieve certain gains through mobilizing in innovative ways. This focus may be explained by the fact that many labor geographers are also dedicated activists who nurture long-standing engagements with unionists and take part in political debates about the need for political reform in the union movement. By providing examples that function as inspiration for other workers and activists, they run the risk of depicting labor as a 'freer agent' than is the case for most workers of the world.

While labor geography has come some way in exploring the agency of labor in its expression through trade unionism, labor geographers have only recently started exploring other, less organized and more individualized forms of agency. Among the exceptions is Rogaly (2009). He uses the experiences of Indian short-term migrants to explore individual worker strategies to negotiate the terms of work and employment. Still, nonunionized workers, workers in the informal sector, and labor migrants have yet to attract the level of rigorous analysis that organized labor has received.

While workers may have agency to make small, yet important, improvements in their everyday working lives, Mitchell (2011) argues that labor geographers have overemphasized the power of workers to effect structural changes of the kind that affect the economic landscape. Hence, studies of the structures under which workers live and work are important to uncover how forces within capital and the state may be served by workers' powerlessness.

Structural Constraints and Enabling Structures

Labor geographers now increasingly emphasize that the agency of labor is subject to structural constraints that vary according to context. Doreen Massey's work in the 1980s can be considered a forerunner in this respect (Peck, 2013). Her approach to labor, capital, and the state in explaining economic restructuring and job loss is relational and based on a critical reading of political economy: the geography of industry is an object of struggle, the world is not simply a product of the requirements of capital, and labor is both a locational factor based on its price and an active agent shaping investment decisions (Massey, 1984). This way, labor shapes the economic landscape.

Also, applying a political-economy approach, Jamie Peck inspires a new generation of geographers. His critique of mainstream labor market (segmentation) theory and his emphasis on contextual factors in explaining working conditions and unemployment make 'Work-Place: The social regulation of labor markets' (1996) a watershed. When neoliberals seek to construct labor markets as commodity markets, their powerful discourse is designed to ease deregulation and flexibility. Building on Marx and Polanyi, Peck stresses the social nature of human labor and productive activity and emphasizes the placeness of work. Structural constraints are prominent in Peck's work. However, stressing that labor markets are situated in different regulative contexts should inspire work on enabling structures that have been underemphasized in labor geography so far.

The last three decades have witnessed intensification of market rule, keener competition, and thereby enhanced commodification of labor. Labor geographers delve into how intensification of work and greater demands on flexibility including increased subcontracting and offshoring of work affect labor in terms of access to work, changes in the work process, working conditions, employment conditions, and how agency is put to work. They are also concerned with opportunities and efforts taken to counter or ease negative effects of increasing competition on access to work, wages, and working conditions (Coe et al., 2009).

Labor geographers have undertaken extensive research on labor market intermediaries, such as temporary staffing agencies. A prominent example is the work of The Geographies of Temporary Staffing Unit at the University of Manchester. This research group is analyzing and mapping the undertakings of this industry, including its worldwide expansion, the way in which it seeks to shape labor markets, its impact at various geographical scales, as well as various attempts at regulating its activities (Coe et al., 2009). Acknowledging that traditional trade unionism is ineffective in parts of the 'new economy,' Brenner and Dean's (2000) study of Silicon Valley investigates new forms of labor organization and employee representation in a context where nonstandard employment has become the norm. In a study of postapartheid Namibia, Endresen (2010) demonstrates how employers, by recruiting temporary workers through labor market intermediaries, may distance themselves from responsibility for their workers and spur worker alienation.

Gender, Class, Ethnicity

A body of the 'structural-constraints strand' in labor geography delves into gender, race, and ethnicity. While gender relations were integral to industrial relocation in Massey's 'Spatial Divisions of Labour' (1984), Hanson and Pratt (1995) develop this further by problematizing gender-based occupational segregation in a study of workplace–domestic sphere relations in Worcester, Massachusetts. They examine how structures of employment opportunity are constructed in place. Local norms and discourses on gender identities such as familial responsibilities and women's capabilities to work, as well as racial differences, are central to this. While women have agency to move both in terms of identity and space, analyses of containment and mobility show how difficult this is for women in their study area. In another study exploring what poststructuralist theories and discourse analysis can bring to theories of labor market segmentation, Pratt (1999) shows how discourses on Filipina workers in Vancouver reinforce their marginalization in the labor market. Hanson and Pratt (1995) argue that the experience of being a woman depends on positionality, in terms of, for instance, class, race, and religion. This notion is further developed in Linda McDowell's work on the intersectionality of gender, class, and ethnicity and in her studies of embodied forms of work in the service economy. Applying an 'intersectionality approach' in studying migrants, McDowell et al. (2007) explain hyperexploitation and labor segregation at a hotel workplace in London and how employee identity becomes subsumed within the workplace, effectively limiting opportunity for resistance.

Labor Regimes

Labor regimes are understood as a complex of institutions, rules, and practices through which relations between labor and capital are regulated, both at the workplace and in society at large. Hence, labor geographers' analyses emphasize place, space, and scale and are concerned with how conditions outside the workplace influence how labor responds to poor working conditions (Jonas, 1996, 2009; Kelly, 2002; Knutsen and Hansson, 2010).

As defined by Jonas (1996, 2009), the concept of local labor control regime represents a shift in explanatory focus away from the politics of *work*places to that of work*places*. Jonas's analyses of the sphere of reproduction stretch to the politics of the place of residence and the politics of consumption. Formed by local conditions as influenced by extralocal institutions and regulations, local regimes reflect practices constructed in places. Underlining these arguments, Kelly (2002) found, in a study in the Philippines, that local regulatory regimes differ from national ones as they are mediated through a range of institutions and that the relations that constitute local labor control regimes are informal and fluid.

Labor in Chains and Networks

Recent theorizing aims to reembed labor as an active agent in the concrete landscapes and structures circumscribing their role as change makers (Cumbers et al., 2008; Coe and Jordhus-Lier, 2010). This has included studies on labor in global value chains and global production networks where the ability of workers and groups of workers to engage in decision making around their work and employment have often been seen as limited (Riisgaard and Hammer, 2011). Within the global production networks framework, Cumbers et al. (2008) demonstrated a way to unpack the agency of labor and distinguish between the collective and the individual agency of labor.

In the global value chains approach, Hale and Wills (2005) were among the first to treat labor as active agents embedded within structural constraints and enabling structures. They demonstrate how labor is affected by organizations of production in global supply chains more generally, regarding industrial upgrading as well as labor standards. In order to address working conditions, they consider it vital to understand industry structure, management of supply chains, and political institutions. For labor to meet the challenges of globalization, they propose new labor internationalism and new forms of workers' organizations that link workers and trade unions in producing locations to local consumers and activists, as well as consumers and activists in important markets of the final products and in the home countries of the lead firms.

The agency of labor is embedded in the structures of the state and capital, but to develop this further requires that the agency of the state and the agency of capital are also analyzed. Reembedding exercises have particular implications for the classical research subject in labor geography: trade unions. There has thus been an increasing interest in labor unions' links to other structures. Rutherford (2013) focuses on the relations between labor and the state. Wills and Simms (2004) investigate the links between trade unions and other organizations in civil society, whereas Brenner and Dean (2000) and Coe et al. (2009) discuss the significance of the growing importance of labor market intermediaries for organized labor. Cumbers et al. (2008) explain how global union federations become spaces of class struggle that can be co-opted into global production strategies, thereby reinforcing exploitation.

Other studies on labor's agency in chains and networks focus on workers' sources of structural power and associational power and how structural power can be mobilized by associational power (Riisgaard and Hammer, 2011). The structural power of labor arises from labor's location in chains and networks while associational power results from formation of collective organizations. These analyses of power are highly contextualized in addressing the nature of the labor process, capital–labor relations, government policy of labor, and development of collective organization, always with an eye to the international and transnational links of capital and collective organizations. These studies provide concrete examples of how the agency of labor can be put to power in chains and networks and illustrate the cumulative character of labor geography.

Spatiality (of) Struggles

Labor geography has developed in an era defined by globalization and neoliberal hegemony. Within this larger context, labor geographers have contributed to general debates in human geography on the nature of space and place and the concept of geographical scale; these debates emerged from the basic insight that space is socially constructed.

Depictions of globalization often rest on an assumption that global space is reserved for capital while workers live and act locally in place (Castree, 2000). However, global restructuring is characterized by profound reconfigurations of the scalar organization of society. Scaling and rescaling are political processes, a view shared by Swyngedouw (1997) and Kelly (2002), whose particular concern is that geographical scales are constructed strategically, in political struggles for power and control. Thereby, scales are continuously shaped and reconfigured, and the construction of scale becomes a vital part of class struggle. Herod (2001b) was among the first to suggest that the concept of scale was useful as a tool to investigate how trade unions have been able to institutionalize lasting victories across space in ways that counter the geographic unevenness of capitalism. Since then, labor geographers have arguably been among the most influential in developing the concept of scale within human geography. The scalar strategies of labor are not confined to efforts of 'scaling up.' Rather, case studies show how an interdependent set of agendas meet in the more successful cases, where local mobilization, national institutions, and global campaigns can reinforce each other.

Labor geographers thus investigate the spatial dimension of the capital–labor relation; a common sentiment is that labor resistance to neoliberal hegemony cannot be confined to the local scale. Accordingly, labor needs to pay attention to how geography has a constitutive role in class struggle.

No geographical scale is 'the natural' scale at which the capital–labor relation should be constituted (Swyngedouw, 1997; Kelly, 2002), although scale is often formalized, or fixed, such as in the area of trade unionism. Agreements establish at which scales bargaining should take place: the workplace, at local, national, or international levels. Scalar struggles may involve efforts to confine counterparts to manageable scales in order to control and subordinate, but scalar strategies may also be liberating. For instance, workers may 'jump scales' (Jones, 1998), i.e., avoid constraints by engaging at other scales.

Some examples may clarify these arguments. Castree (2000) discusses the construction of geographical scales during the Liverpool dock dispute. Union procedures got in the way of striking dockworkers; their answer was to instigate and coordinate the dispute themselves, outside the formal unions. The dockers thereby transformed a local issue into a global one. The dock dispute is an example of grassroots' internationalism outside formal unions, at the national and global scales.

The many ways in which capital may construct spaces and thereby undermine the collective strength of labor are demonstrated by Kelly (2002) in his study from the Philippines. Alongside the regular spatial labor control strategy of (re) locating production to places with no traditions of unionization, Kelly identifies multiple other facets of spatial control of labor: individualizing negotiations; constructing the workplace as the container for dispute resolution; establishing the industrial estate as a separate, closed space; constructing imagined national/ethnic communities; and distancing workplace from homeplace by using migrant workers – in other words, by confining labor to manageable scales.

Scales may be invoked in scalar narratives as agents provide accounts of dramatic events (Swyngedouw, 1997). Defining the scale at which regrettable events, such as plant closures, are triggered and by whom, is an example of framing. In Bergene's (2007) case, the closure of a Malaysian textile factory in Namibia caused not only a contestation around who caused the closure, but also at which scale it was triggered. In this process, both factory owners and the Namibian government tried to contain the Namibian working class and the unions to the national scale by stressing that upscaling the dispute would jeopardize investments and even retard national development.

Similarly, according to Swyngedouw (1997), different agents may invoke different scales strategically in their attempt to situate and explain social events. Jumping scales thus also entail reshaping the discourses within which struggles are constituted, and the politics of scale is rendered a representational strategy (Jones, 1998).

Cox's (1998) twin concepts of spaces of dependence and spaces of engagement are of seminal importance and of particular significance to the politics of scale. Both labor and capital strive to secure their conditions for survival in particular places, i.e., spaces of dependence. These spaces are constantly jeopardized by ongoing social struggles, impelling agents to construct spaces of engagement, which often entail organizing across space.

The view that labor, particularly unionized labor, needs to widen its spaces of engagement given the global character of capitalism is widely accepted. However, as Castree (2000) notes, the global scale cannot a priori be given priority. In his view, the significance of any given scale is not the result of its 'size,' but rather of the processes operating within and across several scales. Hence, the strategies of labor ought to be multiscalar. The global scale is increasingly important, but this fact does not render local and national scales irrelevant. Researchers should not seek to identify which scale was more important in deciding the outcome of a dispute but rather analyze the ways in which actions at various scales complemented, or in some cases contradicted, each other. Despite these qualifications, the tendencies toward global capitalism have led researchers to call for an upscaling of trade union strategies to the global scale. Widening spaces of engagement through coordinated action by establishing extensive alliances requires building solidarity. When studying acts of solidarity, Johns' (1998) distinction between accommodationist and transformatory solidarity is useful. Accomodationist solidarity accommodates workers to capitalism, whereas transformative solidarity involves questioning capitalist social relations. As such, transformatory solidarity entails a commitment to universal class interests over and above the spatially derived interests that accommodationist solidarity seeks to safeguard.

Conclusion

Put simply, what labor geographers have in common is a desire to enhance knowledge that is relevant to advancing the interests of labor. Research is done for the sake of labor, whether labor's agency is directly or indirectly brought to the forefront of the analyses. As such, labor geography is a political endeavor. In early work, it is easy to discern a distinction between 'labor geography proper,' accentuating the active agency of labor but largely shunning issues of structural constraints, and 'labor geography political-economy style' studying how labor's agency is constrained by capital–labor relations, regulations, norms, and values as a prerequisite to find ways to put agency to work.

In both strands, studies are empirically oriented, seeking understanding through case studies. This orientation has its advantages in the thoroughness with which contexts can be analyzed. However, 'single' case studies make theorizing difficult as one cannot distinguish necessary from contingent or contextual conditions in the same manner as in comparative case studies. Hence, in order to develop and refine theory, labor geographers should increasingly work comparatively as this approach broadens the base from which to infer. Attempts at theorizing work and workers' conditions should be even bolder.

The relational approach of agency as embedded in structures of the state and capital strengthens the case for enhanced emphasis on the structures and constraints that labor faces. Moreover, current approaches to labor in value chains and production networks, as well as current studies on labor in the context of keener competition and intensification of work, suggest that the two strands might blend such that the binary between structure and agency dissolves.

The concept of structures encompasses concrete rules, roles, and social institutions as well as mechanisms and processes underpinning social life. Structures may not only constrain, but also enable. The very structure that enables one actor may constrain another. Understanding the empirical manifestations of this duality demands a continuous deconstruction of the concept of structure, building on social theory that would enrich the field. Moreover, a better grasp on norms and values representing structural constraints and enabling structures calls for more use of discursive approaches in labor geography. As labor geography proper has been preoccupied with the active agency of labor, and the political-economy strand has primarily dealt with structural constraints, merging the two might yield a much-needed improved understanding of enabling structures.

Labor geographers have started to unpack the agency of labor and the state, but more needs to be done on how labor as individual agents relate to constraining and enabling structures. The agency of labor is predominantly seen as a positive force in labor geography. With increasing labor mobility, negative aspects such as racism in the workplace need more attention.

Still little has been done to unpack the agency of capital; for example, it is time to dig into power relations in the formation and implementation of formal regulations at all levels, from the workplace to the international level. Labor geography would benefit from more studies of how workers can influence the state and attain better control of the regulations that affect their work life. More attention to working conditions and agency during crises, as well as to regulatory regimes and the agency of those who are pushed into unemployment, is also called for.

In sum, labor geography is a subdiscipline on the way to maturity, where the defining trait during the formative years, the focus on agency, is still strongly present, but where structural constraints have increasingly gained attention the last decade. This leads to a more intense dialogue with social theory in explaining such constraints, a dialogue that is most explicit when spatiality is on the table. Here is where the major contribution of human geographers to the understanding of work and workers' life worlds is most visible.

See also: Agency Theory; Commodity Chains; Critical Realism in Geography; Economic Geography; Gender and Feminist Studies in Geography; Industrial Geography; Labor Market Policy; Labor Unions; Marxist Geography; Network Analysis; Political Geography; Scale in Geography; Social Movements, Geography of; Space and Social Theory in Geography; Spatial Thinking in the Social Sciences.

Bibliography

Bergene, A.C., 2007. Trade unions walking the tightrope in defending workers' interests: wielding a weapon too strong? Labor Studies Journal 32, 142–166.

Brenner, C., Dean, A., 2000. Labour in the new economy: lessons from the labour organization in silicon valley. In: Carré, F., Ferber, M., Herzenberg, S., Golden, L. (Eds.), Nonstandard Work: The Nature and Challenges of Changing Employment Arrangements. IRRA Series, Illinois, pp. 361–376.

Castree, N., 2000. Geographical scale and grassroots internationalism: the liverpool dock dispute, 1995–1998. Economic Geography 76, 272–292.

Coe, N.M., Jordhus-Lier, D.C., 2010. Constrained agency? Re-evaluating the geographies of labour. Progress in Human Geography 35, 211–233.

Coe, N.M., Johns, J., Ward, K., 2009. Agents of casualization? The temporary staffing industry and labour market restructuring in Australia. Journal of Economic Geography 9, 55–84.

Cox, K.R., 1998. Spaces of dependence, spaces of engagement and the politics of scale, or: looking for local politics. Political Geography 17, 1–23.

Cumbers, A., Nativel, C., Routledge, P., 2008. Labour agency and union positionalities in global production networks. Journal of Economic Geography 8, 369–387.

Endresen, S.B., 2010. We order 20 bodies. Labour hire and alienation. In: Bergene, A.C., Endresen, S.B., Knutsen, H.M. (Eds.), Missing Links in Labour Geography. Ashgate, Farnham, pp. 211–224.

Hale, A., Wills, J., 2005. Threads of Labour. Blackwell, Oxford.

Hanson, S., Pratt, G., 1995. Gender, Work, and Space. Routledge, London.

Herod, A., 1997. From a geography of labor to a labor geography: labor's spatial fix and the geography of capitalism. Antipode 29, 1–31.

Herod, A., 1998. Discourse on the docks: containerization and inter-union disputes in US ports, 1955–85. Transactions of the Institute of British Geographers, New Series 23, 177–191.

Herod, A., 2001a. Labor Geographies: Workers and the Landscapes of Capitalism. Guilford Press, New York.

Herod, A., 2001b. Labor internationalism and the contradictions of globalisation: or, why the local is sometimes still important in a global economy. Antipode 33, 407–426.

Johns, R.A., 1998. Bridging the gap between class and space: U.S. workers solidarity with Guatemala. Economic Geography 74, 252–271.

Jonas, A.E.G., 1996. Local labour control regimes: uneven development and the social regulation of production. Regional Studies 30, 323–338.

Jonas, A.E.G., 2009. Labour control regime. In: Kitchin, R., Thrift, N. (Eds.), International Encyclopedia of Human Geography. Elsevier, Amsterdam, pp. 59–65.

Jones, K.T., 1998. Scale as epistemology. Political Geography 17, 25–28.

Kelly, P.F., 2002. Spaces of labour control: comparative perspectives from Southeast Asia. Transactions of the Institute of British Geographers 27, 395–411.

Knutsen, H.M., Hansson, E., 2010. Theoretical approaches to changing labour regimes in transition economies. In: Bergene, A.C., Endresen, S.B., Knutsen, H.M. (Eds.), Missing Links in Labour Geography. Ashgate, Farnham, pp. 155–169.

Massey, D., 1984. Spatial Divisions of Labour: Social Structures and the Geography of Production. Macmillan, Basingstoke.

McDowell, L., Batnitzky, A., Dyer, S., 2007. Division, segmentation, and interpellations: the embodied labors of migrant workers in a greater London hotel. Economic Geography 83, 1–25.

Mitchell, D., 2011. Labor's geography: capital, violence, guest workers and the post-World War II landscape. Antipode 43, 563–595.

Peck, J.A., 1996. Work-Place. The Social Regulation of Labor Markets. Guilford Press, New York.

Peck, J.A., 2013. Making space for labour. In: Featherstone, D., Painter, J. (Eds.), Spatial Politics. Essays for Doreen Massey. RGS-IBG Book Series. Wiley Blackwell, Chichester, pp. 99–114.

Pratt, G., 1999. From registered nurse to registered nanny: discursive geographies of Filipina domestic workers in Vancouver, B.C. Economic Geography 75, 215–236.

Riisgaard, L., Hammer, N., 2011. Prospects for labour in global value chains: labour standards in the cut flower and banana industries. British Journal of Industrial Relations 49, 168–190.

Rogaly, B., 2009. Spaces of work and everyday life: labour geographies and the agency of unorganised temporary migrant workers. Geography Compass 3, 1975–1987.

Rutherford, T.D., 2013. Scaling up by law? Canadian labour law, the nation-state and the case of the British Columbia Health Employees Union. Transactions of the Institute of British Geographers 38, 25–35.

Swyngedouw, E., 1997. Neither global nor local. In: Cox, K.R. (Ed.), "Globalization" and the Politics of Scale. Spaces of Globalization. Reasserting the Power of the Local. Guildford Press, London, pp. 1–23.

Tufts, S., 2007. Emerging labour strategies in Toronto's hotel sector: toward a spatial circuit of union renewal. Environment & Planning A 39, 2383–2404.

Wills, J., 2009. Labour geography. In: Gregory, D., Johnston, R., Pratt, G., Watts, M.J., Whatmore, S. (Eds.), The Dictionary of Human Geography. Wiley Blackwell, Chichester, p. 404.

Wills, J., Simms, M., 2004. Building reciprocal community unionism in the UK. Capital and Class 28, 59–84.

Relevant Websites

http://www.clb.org.hk/en/ – China Labour Bulletin seeks to defend and promote workers' rights in China. Their website presents research reports and news on labor issues in China.

http://www.corpwatch.org – Corpwatch monitors the activities of multinational companies.

http://www.sed.manchester.ac.uk/geography/research/gotsu – The Geographies of Temporary Staffing Unit (GOTSU) at the University of Manchester presents their research on the temporary staffing industry as well as links to relevant companies and institutions and a reference database.

http://www.ilo.org – The International Labor Organization (ILO) provides information on international labor standards as well as access to country profiles, statistics, and databases.

http://www.labourstart.org – LabourStart is a website for unionism and updates on political developments in the world of work.

Labor History

Marcel van der Linden, International Institute of Social History, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 8171–8175, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

'Labor history' has two meanings: it denotes the history of labor movements and of the working classes at large. This article discusses the rise of labor movement history in the late nineteenth century, the growth of working-class historiography during the twentieth century, the origins of the current paradigmatic crisis in labor history, the discipline's core controversies, and its current 'globalization.'

Labor history has two descriptions. In a narrow sense it denotes the history of social movements of wage earners, including their leaders, their organizations (e.g., mutual aid societies, trade unions, consumer and producer cooperatives, and working-class parties), and their collective actions (petitions, strikes, boycotts, electoral campaigns, etc.). In a broad sense, labor history comprises the history of the working classes at large, including the history of family life, demography, everyday culture, leisure activities, housing, religion, migration, and so on (*see* Working Classes, History of). In the early phases of its development as an academic discipline since the 1870s, practitioners of labor history predominantly used the narrow concept.

Beginnings

Labor history as an academic practice came into being after the classical Wage Fund theory had become discredited. According to this theory (of which many variations have existed) there is at any point in time a fixed amount of money out of which wages can be paid; it was therefore "plainly impossible, in any given state of capital and population, to bring about any genuine and permanent rise of wages, otherwise than in the slow course of generations" (Webb and Webb, 1897: II, p. 603). Trade union action could thus have two results. Either it would lead to a redistribution of wages from one segment of the working class to another without the total available amount increasing: "the increased wages secured for merely absorb a larger share of the capital, leaving less to be distributed among others." Or, trade union action was more or less useless: "If wages are below equilibrium, a union can raise them, thus doing what ought to be done and in time would be done by the market itself" (McCulloch, 1851: pp. 50–52). In the 1870s the decline of the Wage Fund theory started with criticisms by John Stuart Mill and others.

With the decline of the Wage Fund theory trade unions and strikes came to be seen in a different light. While most scholars previously had considered trade union activities as false in theory and pernicious or futile in practice, it now became a much more serious topic deserving careful analysis, both contemporary and historical. Pioneer efforts were, among others, Lujo Brentano's *On the History and Development of Gilds, and the Origins of Tradeunions* (London 1870), Richard T. Ely's *The Labor Movement in America* (New York 1886), and the *History of Trade Unionism* (London 1894) by Sidney Webb and Beatrice Webb (born Potter).

For a long time two patterns of interpretations were dominant. On the one hand there was the liberal tradition, represented by, for instance, John R. Commons and Selig Perlman in the United States, the Hammonds in Britain, or scholars like Lujo Brentano, Gerhard Schulze-Gävernitz, and others in Germany. They reconstructed the development of labor movements as the history of the civil emancipation – and consequently the integration – of the working class within capitalism.

The other 'classical' tradition was, of course, the socialist approach in its many forms, varying from moderates to radicals, and including people like the Webbs in Britain, Edouard Dolléans in France, Eduard Bernstein and Carl Grünberg in Germany, or Edvard Bull in Norway. What these socialists had in common was their interpretation of labor history as a history of attempts to transcend capitalism – in whatever way.

Both 'classical' approaches emphasized the labor movements' institutional aspects. Classical labor history consisted of the application of the approaches, methods, formats, and styles of the traditional historiography of ideas and politics to the field of labor history. Insofar as it had addressed 'social issues,' it had been an offshoot of economics and economic history; and insofar as it had addressed the movement itself and its organizations, leaders, and ideas, it had been an offshoot of political or intellectual history. Both versions of the classical approach "tended to produce both a model and an accepted version of history, both national and international, which ranged from an informal but not very flexible to a formal and highly inflexible orthodoxy" (Hobsbawm, 1984: p. 4). I should add, however, that some authors were more or less dissident. There have always been historians who paid attention to atypical aspects, for example, Werner Sombart in his essays on the Italian proletariat (1895); John Le Breton Hammond and his wife Barbara Bradby in their trilogy *The Village Labourer, 1760–1832* (1911), *The Town Labourer, 1760–1832* (1917), and *The Skilled Labourer, 1760–1832* (1919), or Ivy Pinchbeck in her book on *Women Workers and the Industrial Revolution, 1750–1850* (1930).

Broad Interpretation

The classical approaches, whatever their precise substance, served as syntheses for large groups of labor historians. These

old syntheses began to be undermined in the 1950s and the 1960s – a process that continued with even greater force in the following decades. In Britain, Asa Briggs, Eric Hobsbawm, Edward Thompson, and others tried to contextualize workers' struggles. As Hobsbawm wrote in 1964, these historians accentuated "the working classes as such (… and) the economic and technical conditions that allowed labor movements to be effective, or which prevented them from being effective" (Hobsbawm, 1964: p. vii). Somewhat later, similar trends became apparent in other countries. In the United States, Ira Berlin, David Brody, Herbert Gutmann, and David Montgomery, among others, were pioneers.

In France, where the writers of 'evenemential' (*événementiel*) labor movement history had long been opposed by the followers of the *Annales*, whose structural and serial historiography had largely 'left the people out' (to borrow Raphael Samuel's expression), some rapprochement was, especially after 1968, apparent; that was, however, such a slow process that Perrot felt the need to remark in 1979: "(In French historiography) the study of the workers movement has polarized historians for a long time and eclipsed other problems, such as the development of the working class and its culture. However, this is changing rapidly" (Perrot, 1979: p. 150). In Germany, the *Strukturgeschichte*, which was founded in the 1950s by Werner Conze and others (and which incorporated some elements from the National Socialist *Volkssoziologie*, including its focus on the 'structured totality' of the social 'order'), was modified and transformed into a modern approach combining Weberian and Marxian elements by Hans-Ulrich Wehler, Jürgen Kocka, and others since the late 1960s.

Gradually, and especially since the 1960s, a broader approach to labor history has thus become widely accepted. The discipline started to incorporate studies of, for example, social mobility, migration, demography, labor processes, schooling, households, urbanization, social insurance, or sexuality. Race, ethnicity, and gender became generally accepted as essential elements of any analysis, but a full integration into mainstream historiography has not yet been accomplished.

Parallel to labor history's thematic expansion, its geographical scope has widened. Labor history originated in Western Europe and North America, although, early on, there have been important contributions from other parts of the world. From the 1920s it became established in the Soviet Union and later in other Communist countries, and gradually struck roots in Latin America (especially in Argentina, Brazil, and Mexico), Africa (especially South Africa), the Middle East, India, and Australasia. Labor history's present situation is uneven. On the one hand it has become an acknowledged and integral part of historiography in general and of social history in particular. However, on the other hand, its academic conjuncture is strongly linked to political changes, like the collapse of 'socialism' in the Soviet Union and Eastern Europe, and the rise of new workers' movements in Asia.

Both thematic and geographical expansions have made labor history an extremely rich and diversified field. At present, labor history overlaps into women's history, urban history, agrarian history, cultural anthropology, folklore, social economics, history of technology, historical government theories, industrial relations, history of law, and the like. While this disciplinary expansion clearly embodies a wealth of intellectual opportunities, it also leads to a lack of orientation. Already in 1973, Thompson observed that "The central concern of history, as a relevant humane study – to generalize and integrate and to attain a comprehension of the full social and cultural process – becomes lost" (Thompson, 1973).

Conceptual Plurality

Labor history in the broad sense reached its peak during the 1970s when former student militants began to write their theses and books. The productivity of this generation was immense and impressive. Perhaps there would never again be written so much about labor history in so few years as was written at that time. Numerous substantive controversies enlivened the discipline: How should one conceptualize social classes? Are Karl Marx and Max Weber compatible? Is 'labor aristocracy' a useful concept? Have 'organized' labor movements suppressed the interests of unskilled workers? How can one integrate women's history and labor history? What is the societal, cultural, and political importance of the working classes? A wide range of issues were hotly debated, both in professional journals and in the general media.

This peak period did not last long. Most importantly, the organizing categories of the 1960s and 1970s were increasingly called into doubt, partly because they contained continuities with the founding period. Frequently the teleology of the first phase had simply been reserved. While the 'classic' labor historians had been inspired by an optimistic perspective, celebrating the emergence of mass organization and stressing the dynamics of unity and organization, the generation of the 1960s and 1970s tended to ask what had gone wrong with the movement. From a methodological point of view an 'epistemology of absence' became predominant: "Rather than seeking to explain the presence of radically varying dispositions and practices, [labor historians] have concentrated disproportionately on explaining the absence of an expected outcome, namely the emergence of a revolutionary class consciousness among the Western working class" (Somers, 1989: p. 325).

Ever since, many labor historians regard the state of their discipline as a protracted crisis. First, the emerging paradigms of women's and ethnic history showed that there had been giant blank spots on labor history's map and that the filling in of these spots made a complete rewriting of the old narratives unavoidable. And second, the unilinear conception of class consciousness that had been dominant for a long time was questioned. "Once it was established that class was a construction and not a predetermined consequence of structural forces, then the notion that it could be seen as the product of an autonomous culture reacting to certain political and economic circumstances ultimately had to be rethought. Class and class action, it became apparent, could work in many different and contradictory ways" (Price, 1991: p. 253).

As a result of the growing uncertainty about its organizing categories, labor history is beginning to lose its character as a 'discipline.' On the one hand, the distinction with contiguous disciplines, such as women's studies, ethnic studies, anthropology, institutional economics, sociology, or social psychology, is beginning to dissolve, and on the other hand,

conceptual difficulties and political disappointments both have stimulated postmodern approaches, especially in the United States.

In North America and Western Europe, including Britain, the debate between labor historians is therefore dominated by a paradigmatic crisis – a crisis that, as a matter of fact, has apparently not always resulted in a strongly reduced output of the 'older generation' of labor historians, but has often caused a diminished interest among students. In Eastern Europe and the former Soviet Union, the problems are even more serious. The collapse of the so-called 'socialism' has led to the almost complete demise of classical labor historiography, since the latter is with justification perceived as a constituent of the old dictatorship; generally Communist regimes used a distorted 'official' history of the national workers' movement as part of their ideology of legitimation.

At the same time, exciting new developments are taking place. In Latin America labor history is booming (French, 2000), and also in Asia it is rapidly advancing. After a group of Indian intellectuals had already published a series of very innovative collections of essays on the social history of the 'subaltern' classes, which focused especially on the peasants, now a generation of sophisticated labor historians has emerged in that country. Interest in labor history is increasing in several other Asian countries as well, parallel to the emergence of large groups of new wage laborers. Partly, these new labor historians grapple with the same conceptual questions as their colleagues in the rich countries, but that does not prevent their intense research activity.

Core Controversies

There seem to be four interconnected central issues dominating debates within the discipline.

Central Categories

The geographical and thematical broadening of labor history's scope, plus a wider time frame (which now starts to include earlier periods, roughly from the thirteenth century on) makes it necessary to rethink elementary categories, including the discipline's most central notion: labor. In the Anglo-Saxon world, this term has two meanings. It refers simultaneously to work and to workers' movements.

Labor historians have traditionally conceived 'work' as wage labor, performed by legally-free individuals in factories, mines, ports, etc., thus reflecting occupational and gendered views of classical labor movements. Implicitly not included were (1) other forms of free waged labor like domestic service, office work, or voluntary military and police service; (2) unfree waged labor of indentured or otherwise bonded workers; (3) hidden wage labor, for example, through self-employment; and (4) nonwaged labor for subsistance or market purposes. Labor history's broadened perspective suggests a reconceptualization of 'work' as any human effort adding use value to goods and services; use value is understood here as "a good or service that could sustain any activity carried on by a person other than its producer – whether or not we approve of the activity" (Tilly and Tilly, 1998: p. 22–23). This excludes 'purely destructive, expressive, or consumptive acts,' which might be considered as "antiwork" (Tilly and Tilly, 1998: p. 23). Such a Catholic approach makes it possible to overcome old fixations of an institutional or Eurocentric nature.

Traditional labor history has defined a workers' movement as "all of the organized activity of wage earners to better their own conditions either immediately or in the more or less distant future. In all countries it has run along three lines – political, economic, and cooperative" (Commons, 1932: p. 682). A broadening of the concept of 'work' and 'worker' would imply that the notion of a workers' movement has to be broadened as well, and that it should include collective survival activities of all previously excluded categories of workers, whether these are covered by political parties, trade unions, and cooperatives, or not.

Class/Gender/Ethnicity

The relationship between class, gender, and ethnicity has been discussed for several decades. While most labor historians originally considered 'class' as central, and 'gender' and 'ethnicity' and 'race' as categories of a secondary nature, gradually a new consensus has grown which stresses that gender identity, race identity, and class identity have "shaped each other" (Roediger, 1993: p. 131). But despite this new understanding, the *simultaneous* historical analysis of the three identities has proved to be extremely difficult. Following Curthoys one is tempted to see a fundamental indeterminacy problem here: "Trying to keep just two of these concepts in play has proved extremely difficult. […] But if keeping two such concepts in play is hard enough, look what happens when the third concept, be it ethnicity or class or sex, is brought seriously into play. The system, the analysis, becomes too complex to handle" (Curthoys, 1991: p. 15).

Structure versus Actor

A second fundamental indeterminacy problem has a strong influence on labor history debates as well: the relationship between the actor-oriented and the structure-oriented perspective. Classical labor history focused primarily on the agents of social changes: workers, their organizations, and leaders. More recent approaches have paid much attention to broader developments, like, for example, demographic shifts, socioeconomic structures, or governmental bureaucracies. Although few would contest that both modes of interpretation are essential for labor history, their integration is highly problematic.

Even authors who try to combine the two levels (e.g., by paying equal attention to the *longue durée* and short-term events), have to accept that they are in fact telling two separate but combined stories instead of constructing one integrated narrative. The more a scholar focuses on one aspect, the more he or she tends to lose sight of the other aspect. Elton once observed: "Despite attempts to deny this, it [quantitative history] can effectively operate only by suppressing the individual by reducing its subject matter to a collectivity of human data in which the facts of humanity have real difficulty in surviving" (Elton, 1983: pp. 118–119). The reverse could be said as well.

Great versus Small Stories

Classical labor history interpreted the history of working classes and labor movements against the background of major long-term material and cultural changes: for example, the general process of modernization or the rise of capitalism. Since the 1970s this perspective has been challenged from at least two sides.

First, there has grown a strong tendency to criticize such broad interpretative frameworks as 'centrist.' It has been argued that such a way of seeing "positions historical phenomena either to the center or the margins of the historical process, according to their supposed role in the 'great transformation.' " Against this unilinear approach, the critics juxtaposed the emphasis on experiences and processes that do not fit in general schemes. Such a 'de-centered' research strategy is expected to lead to "a greater awareness of the uniqueness, difference, and otherness of historical phenomena, something which is rather obscured by the application of universalizing, passepartout categories (such as role, economic growth, bureaucracy)" (Medick, 1987: p. 82).

Second, classical labor history has been disputed by the so-called 'linguistic turn.' This point of view claims that the backgrounds and interests of the members of a workers' movement cannot explain its development. Instead the language used by the movement's members should be considered as the driving force, as "a complex rhetoric binding together, in a systematic way, shared premisses, analytical routines, strategic options, and programmatic demands" (Jones, 1983: p. 107, cf Scott, 1988: pp. 53–67). Discourse analysis therefore replaces the study of larger material and cultural developments.

Both challenges have directed scholarly attention to previously neglected aspects of labor history. But they have also contributed to the discipline's fragmentation. If discourse is disconnected from its context and considered as 'non-referential' (Jones, 1983: p. 22), or if the concept of historical master processes is contested, then labor history becomes the study of disjointed processes and texts, confirming the lack of a 'central concern' deplored by Thompson (1973). An alternative way of developing labor history further might be through maintaining the idea of master processes and of the contextual situatedness of discursive activities, but by simultaneously taking 'non-central' and linguistic aspects seriously (Kocka, 1997).

Globalization

While the development of labor history is so problematic and uneven, in addition and in connection to this, another trend is apparent: the strong increase in the need to go beyond the national boundaries. The more we know about the history of the working classes in an increasing number of countries, the more tempting it is to place the various national developments in a broader context. Two basic reasons for this can be put forward.

In the first place, we can only discover what is specific and what is generic in our own history by looking beyond national borders. A second important reason for the crossing of national borders is that the object of the investigation is hardly concerned with these borders. Working-class formation is not neatly restricted to certain national borders, but is a process in which voluntary and forced immigration and emigration have a great influence. Dramatic developments in one country may cause turbulences in other countries; strike waves often have a transnational character; new forms of campaigning are imitated elsewhere; national labor movements communicate with each other, learn from each other, and create international organizations.

The growing need for studies that look beyond national borders has resulted in a tremendous growth of the number of contributions to transnational labor history, which started in the 1970s. Yet the work of most labor historians is still characterized by a distinct 'methodological nationalism' that prestructures their writings, with an often rather strict separation of studies of the United States and of other countries.

See also: Gender History; Labor Unions; Social History; Social Inequality in History (Stratification and Classes); Social Movements, History of: General; Socialism: Historical Aspects; Working Classes, History of.

Bibliography

Alexander, R.J., Parker, E.M., 2005a. A History of Organized Labor in Uruguay and Paraguay. Praeger, Westport, CT.
Alexander, R.J., Parker, E.M., 2003. A History of Organized Labor in Brazil. Praeger, Westport, CT.
Alexander, R.J., 2002. A History of Organized Labor in Cuba. Praeger, Westport, CT.
Alexander, R.J., Parker, E.M., 2005b. A History of Organized Labor in Bolivia. Praeger, Westport, CT.
Alexander, R.J., 2003. A History of Organized Labor in Argentina. Praeger, Westport, CT.
Arnesen, E., 2004. The Human Tradition in American Labor History. SR Books, Wilmington, DE.
August, A., 2007. The British Working Class, 1832–1940. Pearson Longman, Harlow, England.
Brody, D., 2005. Labor Embattled: History, Power, Rights. University of Illinois Press, Urbana.
Behan, T., 1997. The Long Awaited Moment: The Working Class and the Italian Communist Party in Milan, 1943–1948. P. Lang, New York.
Berger, S., 2000. Social Democracy and the Working Class in the Nineteenth and Twentieth Century Germany. Longman, Harlow, England.
Belshaw, J.D., 2002. Colonization and Community: The Vancouver Island Coalfield and the Making of the British Columbian Working Class. McGill-Queen's University Press, Montréal.
Commons, J.R., 1932. Labor movement. In: Encyclopaedia of the Social Sciences, vol. 8. Macmillan, New York, pp. 682–696.
Curthoys, A., 1991. The three body problem: feminism and chaos theory. Hecate 17 (1), 14–21.
Clark, A., 1995. The Struggle for the Breeches: Gender and the Making of the British Working Class. University of California Press, Berkeley.
Dubofsky, M., 2000. Hard Work: The Making of Labor History. University of Illinois Press, Urbana.
Donkin, R., 2010. The History of Work. Palgrave Macmillan, Houndmills, Basingstoke, Hampshire.
Elton, G.R., 1983. Two kinds of history. In: Fogel, R.W., Elton, G.R. (Eds.), Which Road to the Past? Two Views of History. Yale University Press, New Haven, CT, pp. 71–121.
Ellis, P.B., 1996. A History of the Irish Working Class. Pluto Press, London.
French, J.D., 2000. The Latin American labor studies boom. International Review of Social History 45, 279–308.
Fink, L., 2011. Workers across the Americas: The Transnational Turn in Labor History. Oxford University Press, New York.
Frader, L.L., Rose, S.O., 1996. Gender and Class in Modern Europe. Cornell University Press, Ithaca, NY.

Freund, B., 1995. Insiders and Outsiders: The Indian Working Class of Durban, 1910–1990. Heinemann, Portsmouth, NH.
Guharāẏa, S., 2007. Calcutta tramwaymen: A Study of Working Class History. Progressive Publishers, Kolkata.
Hobsbawm, E.J., 1964. Labouring Men: Studies in the History of Labour. Weidenfeld and Nicholson, London.
Hobsbawm, E.J., 1984. Worlds of Labour: Further Studies in the History of Labour. Weidenfeld and Nicholson, London.
Heerma, V.L., Linden, M., 2002. Class and Other Identities: Gender, Religion and Ethnicity in the Writing of European Labor History. Berghahn Books, New York.
Haverty-Stacke, D.T., Walkowitz, D.J., 2010. Rethinking U.S. Labor History: Essays on the Working-class Experience, 1756–2009. Continuum, New York.
Hindman, H.D., 2002. Child Labor: An American History. M.E. Sharpe, Armonk, NY.
Jones, G.S., 1983. Languages of Class: Studies in English Working Class History, 1832–1982. Cambridge University Press, Cambridge, UK.
Kocka, J., 1997. New trends in labour movement historiography: a German perspective. International Review of Social History 42, 67–78.
Kessler-Harris, A., 2007. Gendering Labor History. University of Illinois Press, Urbana.
Kessler-Harris, A., 2003. Out to Work: A History of Wage-earning Women in the United States. Oxford University Press, Oxford, England.
Kirk, N., 1998. Change, Continuity and Class: Labour in British Society, 1850–1920. Manchester University Press, Manchester, UK.
Le, B.P., 1999. A Short History of the U.S. Working Class: From Colonial Times to the Twenty-first Century. Humanity Books, Amherst, NY.
Linden, M., 2008. Workers of the World: Essays toward a Global Labor History. Brill, Leiden.
Lockman, Z., 1994. Workers and Working Classes in the Middle East: Struggles, Histories, Historiographies. State University of New York Press, Albany.
McCulloch, J.R., 1851. A Treatise on the Circumstances Which Determine the Rate of Wages and the Condition of the Labouring Classes. London.
Medick, H., 1987. Missionaries in the row boat'? Ethnological ways of knowing as a challenge to social history. Comparative Studies in Society and History 29, 76–98.
Melancon, M.S., Pate, A.K., 2002. New Labor History: Worker Identity and Experience in Russia, 1840–1918. Slavica Publishers, Bloomington, IN.
McIvor, A., 2001. A History of Work in Britain, 1880–1950. Palgrave, Houndmills, Basingstoke, Hampshire.
McLeod, H., 1993. Religion and Irreligion in Victorian England: How Secular Was the Working Class? Headstart History, Bangor.
MacDowell, L.S., Radforth, I.W., 2000. Canadian working Class History. Canadian Scholars' Press, Toronto.
Palmer, B.D., 1990. Descent into Discourse: The Reification of Language and the Writing of Social History. Temple University Press, Philadelphia, PA.
Perrot, M., 1979. The three ages of industrial discipline in nineteenth-century France. In: Merriman, J.M. (Ed.), Consciousness and Class Experience in Nineteenth-Century Europe. Holmes & Meier, New York and London, pp. 149–168.
Price, R., 1991. The future of British labour history. International Review of Social History 36, 249–260.
Roediger, D., 1993. Race and the working-class past in the United States: multiple identities and the future of labor history. In: Van Der Linden, M. (Ed.), The End of Labour History? Cambridge University Press, Cambridge, UK, pp. 127–143.
Scott, J.W., 1988. Gender and the Politics of History. Columbia University Press, New York.
Somers, M.R., 1989. Working class formation—19th century patterns in Western Europe and the United States. Contemporary Sociology 18, 325–329.
Subaltem Studies, 9 vols, 1982–96. Oxford University Press, New Delhi.
Schneider, D., 1994. Trade Unions and Community: The German Working Class in New York City, 1870–1900. University of Illinois Press, Urbana.
Silbey, D., 2005. The British Working Class and Enthusiasm for War, 1914–1916. Frank Cass, New York.
Thompson, E.P., 4 October 1973. Responses to Reality. New Society.
Tilly, C., Tilly, C., 1998. Work under Capitalism. Westview Press, Boulder, CO, and Oxford, UK.
Usner, D.H., 2009. Indian Work: Language and Livelihood in Native American History. Harvard University Press, Cambridge, MA.
Van, H.C.E., Schaffner, H.A., 2003. Work in America: An Encyclopedia of History, Policy, and Society. ABC-CLIO, Santa Barbara, CA.
Webb, S., Webb, B., 1897. Industrial Democracy, 2. vols. Longmans, Green, New York and Bombay, India.

Labor Market Institutions and Labor Market Outcomes in Transition Economies

Hartmut Lehmann, University of Bologna, Bologna, Italy; and IZA, Bonn, Germany
Alexander Muravyev, Higher School of Economics, St. Petersburg Campus, St. Petersburg, Russia; and IZA, Bonn, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article we study whether labor market institutions matter for labor market performance using a new and unique dataset that covers the countries of Eastern Europe and Central Asia, which in the last two decades experienced radical economic and institutional changes. We show that institutions indeed matter and that deregulation of labor markets improves their performance. We also find several significant interactions between different institutions, which are in line with the idea of beneficial effects of reform complementarity and broad reform packages.

Introduction

This article is on the evolution of labor markets and labor market institutions and policies in the countries of Central and Eastern Europe as well as of Central Asia over the last two decades. As the dynamics of labor market outcomes in these transition economies (TEs) has drawn considerable attention already in the early years of transition and resulted in a voluminous literature easily available elsewhere, this article only briefly touches on this topic. Instead, we focus on the much less explored issue of the evolution of labor market institutions, which are among candidate explanations for the very diverse trajectories of labor markets in the region. We consider recent contributions that attempt to assess the effect of labor market institutions on labor market performance of TEs using unique aggregate country-level data, which we have hand collected.

Labor market institutions have taken center stage in the literature on unemployment ever since the seminal contributions of Nickell (1997) and Layard and Nickell (1999). Employment protection, the unemployment benefit system, and taxes on labor are institutions that profoundly affect wage and price setting in the labor market and thus determine the unemployment rate in the short and long run. The proper design of labor market institutions is thus crucial when pursuing macroeconomic policies that guarantee sustained growth and the highest level of employment possible. Layard and Nickell have taught us that a good design does not necessarily mean that all institutions are as 'flexible' as possible. Instead, the right interplay of institutions, often reinforced with an aggressive stance on active labor market policies, is far more important to guarantee low unemployment. Our work on the role of labor market institutions in TEs contributes to the empirical literature on the crucial link of these institutions and labor market performance.

We look at five labor market outcomes that are of particular interest in the literature when assessing labor market performance: the employment-to-population ratio, the unemployment rate, the long-term unemployment rate, the youth unemployment rate, and the size of the informal sector in the economy. (An early seminal contribution to this literature covering OECD countries is by Nickell (1997) who looks at the first four outcome variables only. Unfortunately, there are no reliable data on informal employment for all the transition economies and we take the size of the informal sector as a proxy for informal employment. (Gimpelson and Zudina (2011) for Russia, Lehmann and Pignatti (2007) for Ukraine show that the two variables are highly correlated over time.) These five outcomes will constitute the dependent variables whose determinants we will analyze in our empirical analysis. The determinants of labor market performance are many; in this article we restrict ourselves to labor market institutions and labor market policies, in particular employment protection legislation, the tax wedge on labor, union density, the level of unemployment benefits, and their duration as well as expenditures on active labor market policies. (For a recent survey of the link between labor market institutions and policies and labor market performance in OECD countries, see Bassanini and Duval (2009).) For reasons discussed below, neither the minimum wage nor wage bargaining is included in the list of determinants of labor market performance in TEs.

Traditional measures of labor market performance are the first four mentioned outcomes, while informality has only recently been linked to labor market institutions and policies in the literature on labor markets in less developed economies (Lehmann and Muravyev, forthcoming). Informality and informal employment pose a major challenge to policy makers in the transition countries of Central and Eastern Europe and the Commonwealth of Independent States. While it is difficult to precisely estimate the size of these phenomena, there can be no doubt that in these areas of the world a large part of economic activity is not registered or only partially registered and that many workers enter employment relationships that provide only partial or no protection against unemployment, illness, and old age (see, e.g., Slonimczyk, 2012, 2013; Lehmann and Pignatti, 2007). Whether labor market institutions and policies have an impact on the size of the informal sector and thus on informal employment is an important but hardly researched topic in the literature. Our work on this topic, summarized in this article, thus provides a unique contribution when trying to answer this question for TEs.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.94032-7

The bulk of the available evidence concerning the impact of institutions and policies on the first four measures of labor market performance is based on data from two dozen OECD countries. Only in recent years have some scholars started to look at the role of labor market institutions and policies in less developed economies (Botero et al., 2004; Feldmann, 2008). To a considerable extent, such interest stems from a much larger variation in institutions and labor market policies as well as in labor market outcomes across such an extended list of countries, both in the cross-section and time dimension. (For example, Djankov and Ramalho (2009, p. 11) state: "Developing countries present an exciting venue for studying the impact of regulatory reforms, including of labor reforms. A number of countries, especially in Eastern Europe, have recently undergone significant reforms to make labor regulation more flexible.") In addition, data from less developed countries, in principle, can help reveal whether the previously obtained conclusions for the OECD economies can be generalized to other regions of the world. However, the potential of non-OECD countries to contribute to the economics literature in general has not yet been fully realized because of only very limited availability of data. In the subset of non-OECD countries that interests us here, the transition countries, changes in labor market outcomes as well as changes in institutions and policies are more marked over time than they are in mature OECD countries (Boeri and Lehmann, 1999), providing thus a natural testing ground of the theoretical considerations that link labor market institutions and labor market outcomes.

The literature on OECD countries highlights the large heterogeneity in labor market outcomes even in the presence of very similar labor market institutions. The explanation for this large heterogeneity is often tied to the complementary nature of the interaction of labor market institutions when institutions are simultaneously liberalized (see, e.g., Coe and Snower, 1997). By analyzing interactions of different institutions and policies we provide some additional support for the complementary nature of simultaneous liberalization of various labor market institutions in the rather different economic context of transition countries. This analysis is, however, only undertaken for the first four labor market outcomes but not for the size of the informal sector.

Macroeconometric analysis of labor market developments using country-wide data certainly has its critics since it has difficulties in pinning down causal effects. Nevertheless, we consider this analysis worthwhile for at least two reasons. First, establishing links between some labor market institutions or policies and measures of labor market performance that are robust across specifications provides some solid evidence regarding the beneficial effects of flexible labor market institutions in the medium to long run. Second, in non-OECD countries the impact of labor market institutions on labor market outcomes, and in particular on informality, has been analyzed nearly exclusively from a labor supply perspective (see, for example, the chapters in Frölich et al., forthcoming). In contrast, our macroeconometric analysis of the determination of the five labor market outcomes also encompasses the labor demand side, at least partially.

The Evolution of Labor Market Outcomes in Transition Economies

In the literature studying the region, it is common to classify countries into three major groups: Central Eastern Europe (CEE, embracing the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Slovak Republic, Slovenia), South Eastern Europe (SEE, which includes Albania, Bosnia and Herzegovina, Bulgaria, Croatia, Macedonia, Montenegro, Romania, and Serbia), and the Commonwealth of Independent States (CIS, which until recently included 12 out of 15 constituent republics of the former USSR, namely Armenia, Azerbaijan, Belarus, Georgia, Kazakhstan, Kyrgyzstan, Moldova, Russian Federation, Ukraine Tajikistan, Turkmenistan, and Uzbekistan with Georgia officially leaving the organization in August 2009). We will for the most part stick to this classification, commenting on smaller subgroups of countries or single economies when necessary and/or important. (In particular, we will split the CIS into European CIS (Belarus, Moldova, Russia, Ukraine), Caucasus (Armenia, Azerbaijan, and Georgia), and Central Asia, separate the Baltic states from the rest of the CEE group, and single out Bulgaria and Romania in the SEE group.)

The data suggest some striking patterns of labor market outcomes in the transition countries during the last decades. The key factor behind this dynamics was the drop in output that these countries experienced at the outset of transition. **Figure 1** presents data on GDP per capita across the subgroups of countries. Although the decline in output was significant everywhere, it was particularly dramatic in the CIS and Baltic states. The group of European CIS also experienced the longest recession. The decline in output could not occur without important consequences for the labor market. As the adjustment mechanisms have been thoroughly studied elsewhere (Svejnar, 1999; Boeri and Terrell, 2002; Earle and Sabirianova, 2002; Lehmann et al., 1999; Lehmann and Wadsworth, 2000), we will only illustrate key stylized facts. The employment-to-population ratio is substantially smaller in SEE than in the other transition regions (see **Figure 2**). It is u-shaped for CEE in the 1990s, indicating an upturn in labor demand in the later part of transition, while in SEE it shows a strong downward trend until 2004. The unemployment rates also exhibit interesting patterns even if we average the rates within regions (**Figure 3**). It is noteworthy that unemployment rates have been higher in CEE than in the CIS for the most part of the period considered even though the employment-to-population ratios hardly differ. The other important feature that should be mentioned is the large drop in the unemployment rate between 2004 and 2008 in CEE and SEE, whereas the unemployment rates drop gently in the CIS. Long-term and youth unemployment rates are far higher in SEE than in the other two transition regions. The largest drop in both rates between 2004 and 2008 can be observed in CEE and SEE.

As far as the size of the informal sector is concerned, it is very high in the CIS, high in SEE, and moderate in CEE. This pattern stays the same over the time horizon that Schneider et al. (2010) consider, namely from 1999 to 2007. In **Table 1** we report this informality statistic from their paper for each country for the years that we use in our empirical analysis. Inspection of the table shows that in 1999 Georgia and

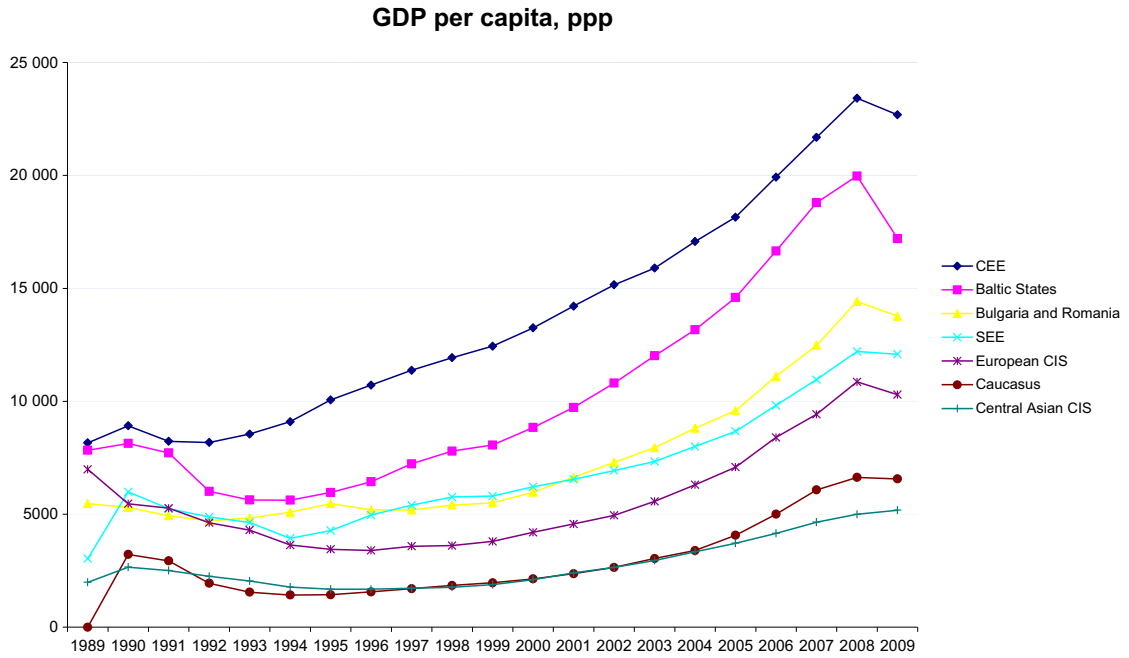

Figure 1 GDP per capita across the subregions of Central Eastern Europe and Central Asia.

Azerbaijan had the highest share of informality with 68% and 61% of GDP, respectively, followed by Ukraine, Belarus, and the Russian Federation. The Central Asian Republics have also high shares of informality. These numbers clearly point to the importance of activities in the grey economy in the CIS. With the exception of the Baltic States the numbers for CEE are relatively modest, in particular for the Czech and Slovak Republics, while the countries in SEE take a middle position. The numbers indicate at any rate that relative to advanced OECD countries the informal sector is of greater importance in TEs. Table 1 also demonstrates that between 1999, a year of recession, and 2007, when we have an economic boom, the size of the informal sector is declining in all transition countries, pointing to an anticyclical pattern of informality in the region. We can thus moot that workers take up informal activities more intensely when the economy is in recession and move more into formal employment in boom times.

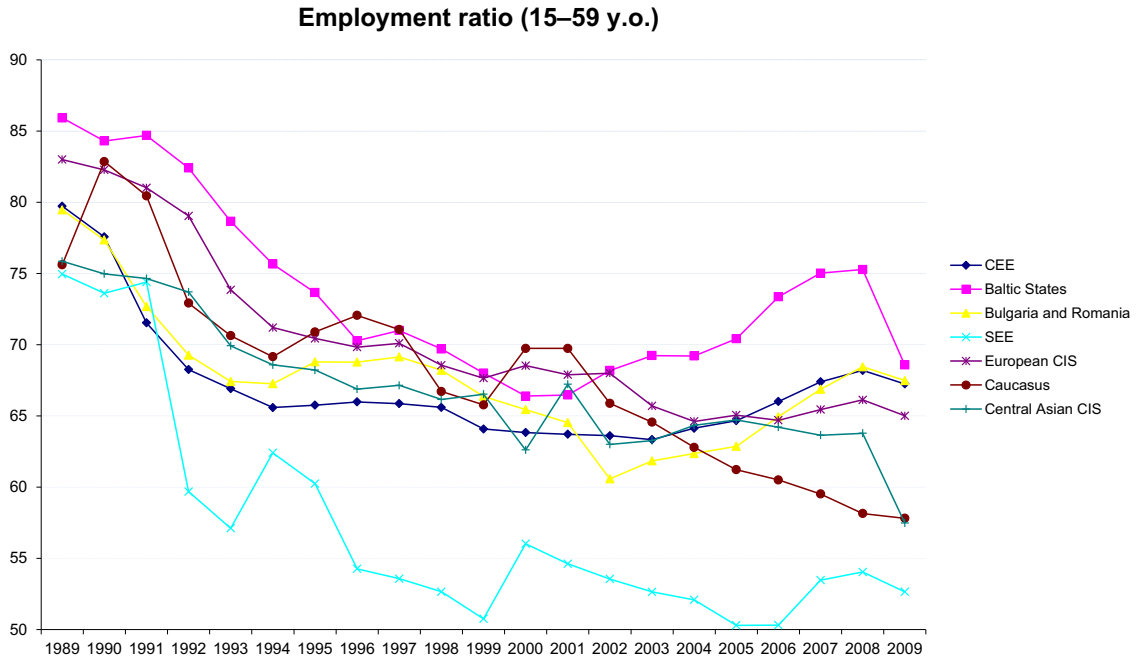

Figure 2 The employment-to-population ratio across the subregions of Central Eastern Europe and Central Asia.

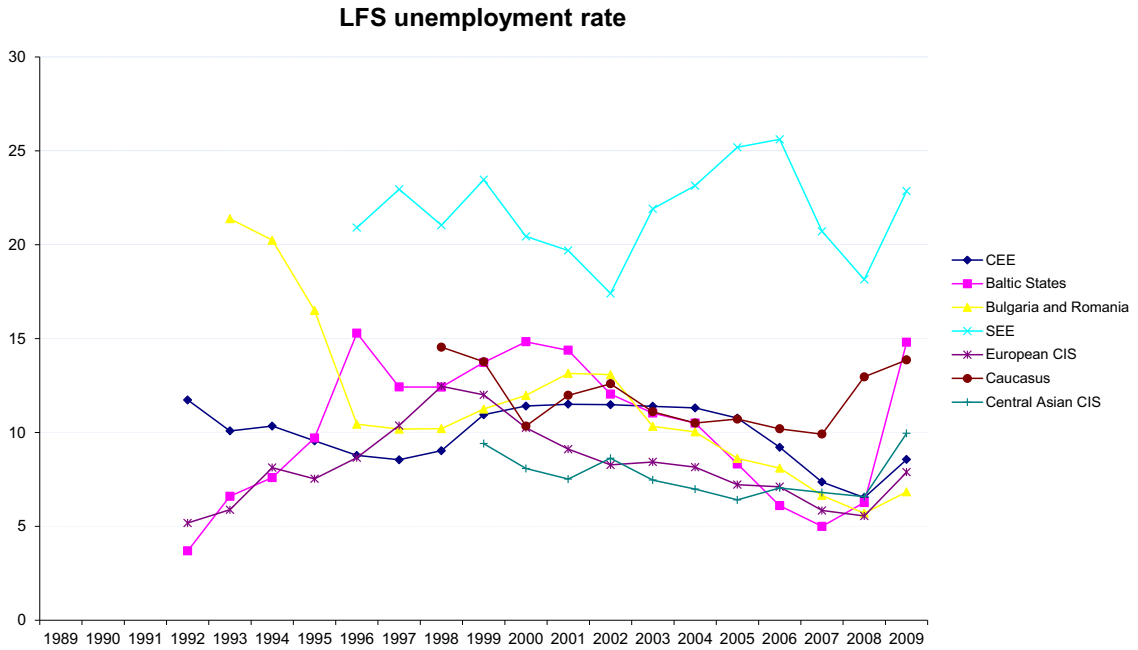

Figure 3 The unemployment rate across the subregions of Central Eastern Europe and Central Asia. Source: TransMonEE database, UNICEF.

Table 1 The size of informal economy (in % of official GDP) in transition countries

Country	*1999*	*2003*	*2007*
CEE			
Czech Republic	19.3	18.7	17
Hungary	25.4	24.4	23.7
Poland	27.7	27.5	26
Slovak Republic	18.9	18.3	16.8
Slovenia	27.3	26.4	24.7
Baltic States			
Estonia		31.4	29.5
Latvia	30.8	29.4	27.2
Lithuania	33.8	32	29.7
Bulgaria and Romania			
Bulgaria	37.3	35.6	32.7
Romania	34.3	32.8	30.2
SEE			
Albania	35.7	34.4	32.9
Bosnia and Herzegovina	34.3	33.5	32.8
Croatia	33.8	32.1	30.4
Macedonia	39	38.4	34.9
European CIS			
Belarus	48.3	47	43.3
Moldova	45.6	44.6	
Russian Federation	47	43.6	40.6
Ukraine	52.7	49.7	46.8
Caucasus			
Armenia	46.6	43.9	41.1
Azerbaijan	61	59.1	52
Georgia	68.3	65.9	62.1
Central Asia			
Kazakhstan	43.8	41.1	38.4
Kyrgyz Republic	41.4	40.5	38.8
Tajikistan	43.5	42.1	41

Source: Schneider, F. Buehn, A., Montenegro, C.E., 2010. New estimates for the shadow economies all over the world. International Economic Journal 24 (4), 443–461.

Data

Before giving a descriptive analysis of the labor market institutions and policies we briefly describe our data sources. Our empirical analysis is based on data from several principal sources. The first source is the database of the IZA Program Area 'Labor markets in emerging and transition economies,' which is a new hand-collected dataset that provides essential information about the evolution of labor markets in the countries of Central Europe and Central Asia. It includes 27 countries of the region and spans 4 years, 1995, 1999, 2003, and 2007. The database contains four key variables characterizing labor market outcomes and six key variables describing labor market institutions. The latter include employment protection legislation (OECD version II), expenditures on active labor marker programs, tax wedge, unemployment benefit size (measured as average benefit to average wage), unemployment benefit duration, and union density. The details about the variables included in the database can be found in Lehmann and Muravyev (2012).

The second source is data on the size of informal economy taken from the paper by Schneider et al. (2010). This source provides estimates for 162 countries, including Eastern European and Central Asian countries over 1999 to 2007. This is a unique dataset providing comparable estimates for most countries of the world based on the MIMIC estimation method. From this source, we take information on the size of informal economy in transition countries in 1999, 2003, and 2007.

Finally, we have added some key macroeconomic variables from the World Bank database (http://data.worldbank.org/), such as GDP growth rate and inflation. These variables are commonly used in macroeconometric labor regressions for various robustness checks.

General Descriptive Statistics and the Evolution of Labor Market Institutions

Descriptive statistics for the key variables characterizing labor market outcomes as well as labor market institutions is shown in Table 2. The size of the informal economy is large in the region (above 36%), but varies considerably across the countries. The other labor market outcomes are somewhat weaker in the sampled countries as compared with the developed economies (see also Lehmann and Muravyev, 2012). However, there is also considerable variation across the countries and over time. The variables measuring labor market institutions and policies are, in general, at lower levels than in mature market economies, especially of Western Europe. The EPL appears to be relatively rigid, at the level of 2.41. The tax wedge is non-negligible, although less than in most mature market economies. Unemployment benefit is rather small, 25% of the average wage, and much less than in developed countries. The duration of unemployment benefit is 11 months, on average. Again, this is less than in most high-income countries, especially of Western Europe. Union density is at the level of 43%, which is considerable.

Figure 4 shows the evolution of the six variables that proxy labor market institutions and policies for the three subsets of transition countries and, as benchmarks, for the EU15 and the U.S. The index for employment protection legislation is similar among the three regions of the transition area and the EU15 whilst the index is much lower for the U.S., pointing to a difference in employment protection that is of an order of magnitude between the latter country and the other regions analyzed. It is also striking that between 1999 and 2007 the indices hardly change for CEE, the EU15, and the U.S.; in contrast we see a substantial fall in the CIS region. The upper right panel demonstrates a much lower union density for the U.S. and the near constancy of union density in the EU15 and the U.S. while in transition countries we see a large fall in CEE and the CIS. Nevertheless, at the end of the period SEE and the CIS still have a substantially higher percentage of workers with union membership than does the EU15. Tax wedges on labor are high and remain high for CEE, SEE, and the EU15 throughout the period, while in CIS countries it shows a precipitous fall from more than 40% to close to 30%. The introduction of low flat income tax rates in many CIS countries in the first years of the new century seems to have driven the lowering of the tax wedge in this region, approaching the constant low tax wedge that we see in the U.S.

The last three panels show measures of active and passive labor market policies. Only in the EU15 are expenditures on active labor market policies substantial reaching to nearly 1% of GDP. These expenditures are much lower in all transition regions, in particular in SEE and the CIS. There is some slight decrease over time in all regions apart from SEE, which is related to an expanding economy in the latter half of the period. It is also noteworthy that the U.S. spends consistently less than CEE. The last two panels deal with the generosity of the unemployment benefit system. The EU15 has by far the most generous system since a long duration of benefits is combined with a relatively high replacement rate. Among the transition regions SEE is the most generous, the CIS is the least generous since the replacement rates are close to negligible over the whole period and the duration of benefits falls dramatically in the latter half of the period.

Figure 4 demonstrates why we can speak of the U.S. labor market as the archetypical flexible labor market since the measures for EPL, union density, and the tax wedge on labor are all very low. Many researchers link the superior GDP growth of the U.S. economy between 1993 and 2007 to the flexibility of the U.S. labor market. At the same time Figure 4 shows a clear tendency toward a more flexible labor market in the transition region, and in particular in the CIS. We, therefore, ask whether this increase in flexibility has brought with it a better labor market performance, i.e., more employment, less unemployment, and less informality. We now turn to the methodology employed to answer this question.

Table 2 Descriptive statistics

Variable	*Mean*	*SD*	*Min*	*Max*	*N*
Employment ratio (ER)	64.23	8.69	31.20	79.90	97
Unemployment rate (UR)	11.34	6.30	4.00	36.70	79
Long-term unemployment rate (LTUR)	52.66	17.53	14.60	92.60	70
Youth unemployment rate (YUR)	24.84	14.18	7.80	65.70	72
Size of the informal sector (INFORM)	36.64	11.24	16.80	68.30	73
EPL	2.41	0.59	0.38	4.10	85
ALMP	0.20	0.18	0.01	0.83	79
TAX	38.97	5.53	23.00	53.20	90
DENSITY	43.46	21.20	13.18	100.00	85
BEN_DURATION	11.25	5.54	0.00	24.00	89
BENEFIT_SIZE	25.20	11.39	0.00	60.00	89
INFLATION	−112.14	553.66	−5096.71	262.12	107
GDP_GROWTH	1.11	0.23	0.48	2.00	112

Source: Database of the IZA Program Area 'Labor markets in emerging and transition economies.' See Lehmann, H., Muravyev, A., 2012. Labour market institutions and labour market performance: what can we learn from transition countries? Economics of Transition 20 (2), 235–269.

Methodology

Our analysis of the links between labor market outcomes, including the size of informal economy on the one hand and labor market institutions and policies on the other hand, draws heavily on the standard macro-regressions proposed in the seminal study by Nickell (1997). In that study, labor market outcome variables are related, in a panel regression framework, to a set of variables measuring institutions and policies, as well as to the change in inflation. We proceed in an essentially similar fashion by considering, in the baseline specification, six variables characterizing institutions and policies.

We note that our results do not necessarily have a causal interpretation as both institutions and policies may be shaped by labor market outcomes, for example, via the mechanism of elections (Blanchard, 2006). Nevertheless, we try to address endogeneity (at least some of its sources) by controlling for

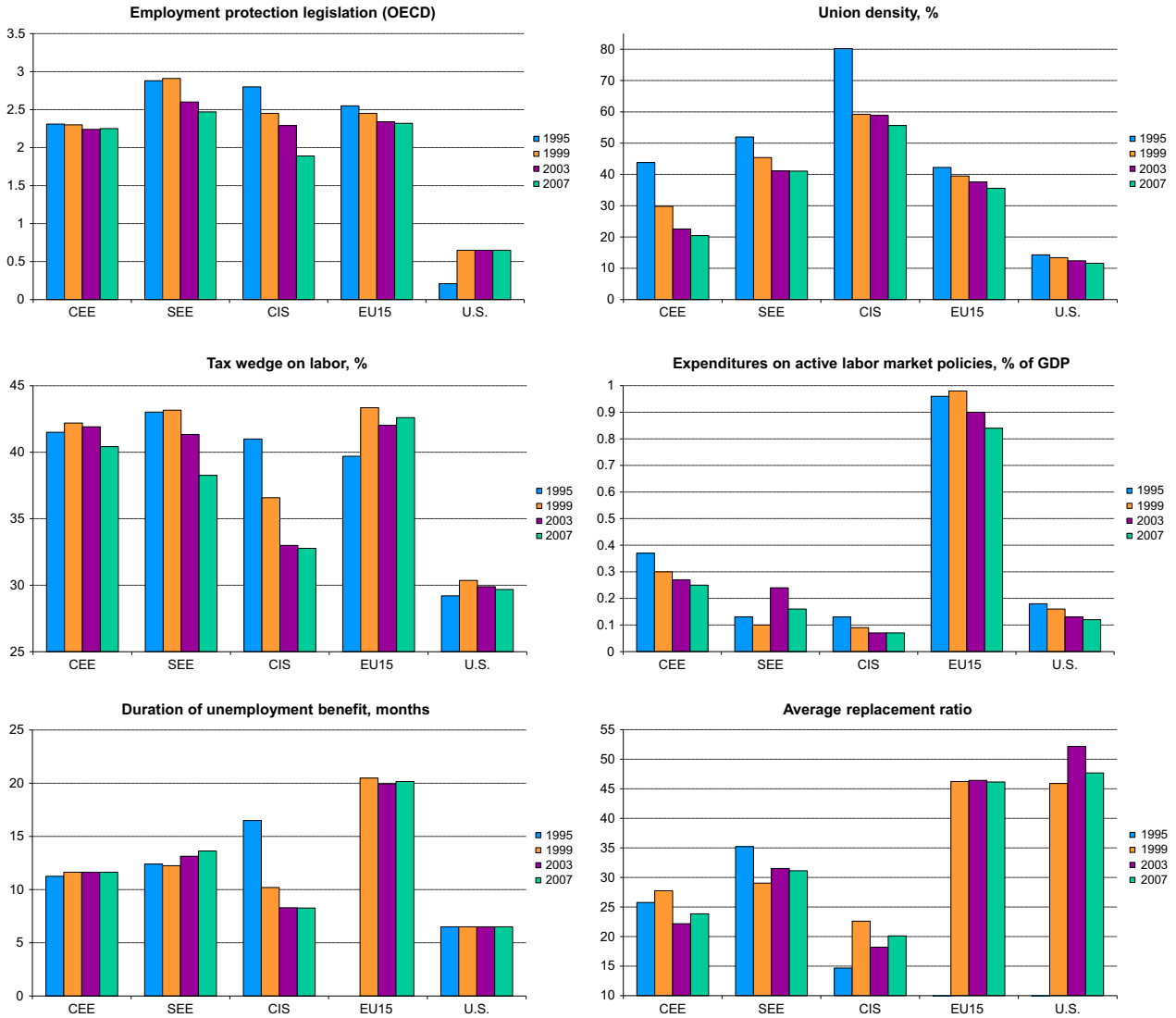

Figure 4 Labor market institutions in Central Eastern Europe and Central Asia. Source: Database of the IZA Program Area 'Labor markets in emerging and transition economies.'

omitted factors (including unobserved characteristics of countries) by using random- or fixed-effects specifications of our regression model. These are necessary as the paucity of the degrees of freedom does not allow inclusion of many potentially relevant explanatory variables. The baseline regression eqn [1] can then be written in the following way:

$$LMO_{it} = \alpha_i + \beta_1 EPL_{it} + \beta_2 ALMP_{it} + \beta_3 TAX_{it} + \beta_4 DENSITY_{it} + \beta_5 BEN_DURATION_{it} + \beta_6 BENEFIT_SIZE_{it} + \beta_7 INFLATION_{it} + \beta_8 GDP_GROWTH_{it} + \varepsilon_{it} \quad [1]$$

where index i denotes countries and index t denotes time, $t \in \{1999, 2003, 2007\}$, *LMO* stands for labor market outcomes (As much of the previous studies, we do not apply logarithmic transformation to the dependent variables in the model.), *EPL* measures the strictness of employment protection legislation, *ALMP* is expenditures on active labor market programs, in % of GDP, *TAX* is the tax wedge on labor, *DENS* measures union density, *BENEFIT_SIZE* stands for the average unemployment benefit replacement rate, *BEN_DURATION* stands for the maximum duration of unemployment benefits, *INFLATION* is the change in inflation over the previous year, *GDP_GROWTH* is the cumulative growth of GDP in the 3 years before labor market outcomes are measured, and ε is a white noise disturbance. In addition, we consider a specification which includes time effects that capture common macro-trends in the countries sampled.

As can be seen from the specification of eqn [1], one substantial difference from the study by Nickell (1997) is that we do not include variables measuring union coverage rates and bargaining type – again, mostly for data reasons, but also due to the difficulties in interpreting these variables in less-developed countries. (For example, how would one interpret data on bargaining in a country where trade unions with high membership rates are effectively controlled by the

government? It is therefore no surprise that the World Bank did not provide statistics on the coverage rates and bargaining type in the CIS countries (World Bank, 2005).) We, however, believe that we capture the essential aspects of wage setting with our union density variable since it is regarded as the most important of the related factors (Eichhorst et al., 2008). We also do not include the minimum wage. While its effect on employment is controversial in the literature (see, for example, Card and Krueger, 1994; Neumark and Wachter, 1992, 2008), it certainly affects labor supply and demand in countries where it is binding and enforced. In many transition countries of the CIS the real minimum wage is, however, so low throughout the period that it is far below the subsistence minimum and thus does not influence workers' behavior. In addition, where the minimum wage is not negligible it is often not enforced due to weak legal instruments.

In previous work (Lehmann and Muravyev, 2012), from which we report here, we also looked at the question whether reform complementarities improved labor market performance; that is, whether reforming more than one labor market institution brought larger benefits than confining the reform effort to one institution or policy. There is convincing evidence in OECD countries, at length debated in the literature, which shows that countries with similar single institutions might experience very different labor market outcomes because these single institutions are interacted with other single institutions, which in turn differ from country to country. A classic example of the importance of institutional complementarities is shown in Nickell (1997) when discussing the level of unemployment benefits. Ceteris paribus, a high level of unemployment benefits should imply a higher unemployment rate, but when interacted with a short duration of these benefits and large expenditures on ALMP, the institutional mix applied in Sweden, generous unemployment benefits might result in better labor market performance.

In Lehmann and Muravyev (2012), we add pairwise interactions of labor market institutions and policies to the baseline specification for the full sample of transition countries. Given the low number of degrees of freedom, only one such interaction is inserted at a time. Following Nickell et al. (2005), the interaction terms are defined in the form of products of deviations of the institutional variables from their sample means. This analysis of reform complementarity is done for the first four labor market outcomes, that is, Lehmann and Muravyev (2012) exclude the size of the informal sector from the analysis.

Empirical Results

We present the results of estimating the baseline specification for all five labor market outcomes, employment-to-population ratio, unemployment rate, long-term unemployment rate, youth unemployment rate, and informality in **Table 3**. These results should be interpreted as the medium-to-long-term effects of changes in labor market institutions and policies. The regression in column 1 shows the effect of labor market institutions on the employment-to-population ratio. The results suggest a negative and statistically significant effect of EPL as well as of the tax wedge on the employment-to-population ratio. Column 2 shows the results for the unemployment rate as the dependent variable. The only statistically significant coefficient here is on benefit duration. It has a negative sign, which most economists would consider counterintuitive: increasing the duration of unemployment benefits leads to a reduction in the unemployment rate. (Longer benefit duration makes the search less costly. Longer search in turn might result in productive job matches that last longer than in the absence of longer search. Hence a longer benefit duration might depress the stock of unemployment via smaller inflows into unemployment.) Note, however, that the coefficient is only significant at the 10% level. Column 3 shows the results for the long-term unemployment rate. The results are similar to those in column 2. Among labor market institutions, the only statistically significant effect is on benefit size, which has a counterintuitive negative sign. The results for youth unemployment, reported in column 4, are more in accord

Table 3 Baseline estimation results

	ER	*UR*	*LTUR*	*YUR*	*INFORMAL*
EPL	−4.127[b] (1.487)	0.636 (0.652)	1.723 (3.296)	4.704[c] (2.326)	1.848[b] (0.680)
ALMP	0.845 (2.737)	−1.431 (2.962)	−4.425 (6.502)	−5.736 (6.329)	0.007 (1.892)
TAX	−0.434[a] (0.126)	0.126 (0.149)	−0.276 (0.413)	1.049[a] (0.286)	0.152 (0.089)
DENSITY	−0.023 (0.058)	0.015 (0.025)	0.042 (0.193)	−0.119 (0.133)	0.038 (0.036)
BEN_DURATION	−0.203 (0.210)	−0.235[c] (0.116)	0.022 (0.636)	−0.301 (0.378)	−0.153[c] (0.077)
BENEFIT_SIZE	0.200[c] (0.097)	−0.077 (0.085)	−0.616[b] (0.249)	−0.227 (0.242)	−0.017 (0.057)
INFLATION	−0.039 (0.030)	0.008 (0.017)	0.093[c] (0.047)	0.014 (0.045)	−0.025 (0.020)
GDP_GROWTH	−12.060[b] (4.992)	−9.303[b] (3.306)	11.933 (17.329)	0.845 (11.568)	−6.247[c] (3.119)
INTERCEPT	101.716[a] (10.205)	20.243[b] (8.986)	62.597[c] (36.124)	−13.328 (20.321)	32.037[a] (6.725)
R^2	0.39	0.41	0.19	0.21	0.66
N	70	67	62	63	57

Notes: Dependent variables: Employment rate (ER), unemployment rate (UR), long-term unemployment rate (LTUR), youth unemployment rate (YUR), and the size of the informal sector (INFORMAL). The results are obtained using fixed-effects estimators with cluster-robust standard errors (clustering by country). Standard errors are reported in parentheses. R^2 refers to the within-variation in the data.
[a]Significance level at 1%.
[b]Significance level at 5%.
[c]Significance level at 10%.

Table 4 Interactions between institutions and their complementarities – four measures of labor market performance

	Employment rate	*Unemployment rate*	*Long-term unemployment rate*	*Youth unemployment rate*
ALMP_TAX	0	Comp[b]	Comp[b]	0
ALMP_DENS	0	Comp[b]	Comp[c]	Comp[b]
ALMP_BENF	0	0	0	0
ALMP_BEND	Comp[b]	0	0	0
ALMP_EPL	0	Comp[c]	Comp[c]	Comp[b]
EPL_TAX	0	0	0	0
EPL_DENS	0	0	0	0
EPL_BENF	0	0	0	0
EPL_BEND	0	0	0	0
TAX_DENS	0	0	0	0
TAX_BENF	0	0	0	0
TAX_BEND	0	0	Comp[b]	Comp[a]
DENS_BENF	0	0	0	0
DENS_BEND	0	0	0	0
BENF_BEND	0	0	0	0

[a]Complementary at the 1% significance level.
[b]Complementary at the 5% significance level.
[c]Complementary at the 10% significance level.
Source: Based on Lehmann, H., Muravyev, A., 2012. Labour market institutions and labour market performance: what can we learn from transition countries? Economics of Transition 20 (2), 235–269.

with economic theory. Here, there are two statistically significant coefficients – on EPL and tax wedge. Both are positive, implying that higher taxes on labor as well as stricter EPL are associated with higher youth unemployment rate. The results for informality, reported in column 5, suggest that informal economy expands with higher EPL and shrinks with longer benefit duration. These last results are plausible since stricter employment protection will depress formal employment and thus boost informality. The conventional interpretation of the impact of longer benefit duration, i.e., keeping workers longer in unemployment, gives us smaller flows into informal employment and thus a lower size of the informal sector.

When we use the alternative specification that includes time dummies capturing trends common to all countries of the region, most of the results reported in Table 3 stay intact. Some results are different, however. In particular, we observe a negative and statistically significant coefficient on ALMP in the regression with youth unemployment. This suggests that spending on active labor market policies is effective in tackling the problem of unemployment among the young workers in the countries under study. Importantly, we still observe detrimental effects of a higher tax wedge and a stricter EPL on labor market outcomes in theTEs.

We finally give a qualitative assessment of the issue of reform complementarity in Table 4, which summarizes the results of Lehmann and Muravyev (2012). Raising expenditures on ALMP will have a stronger impact on unemployment and long-term unemployment if the labor market is characterized by a relatively low tax wedge (row 1 of Table 4). This result is intuitively quite plausible. If, for example, an ALMP measure targeted at the long-term unemployed is meant to rebuild their human capital, such a measure will result in larger outflows into employment if labor is taxed less. A similar argument can be made with the interaction of ALMP and EPL. ALMP measures are meant to make the unemployed fit for unsubsidized regular employment. If EPL is relatively unregulated firms will be more willing to hire from the unemployment pool, a result that is shown in row 5 of Table 4. Very convincing strikes us also the complementary effect of lowering the tax wedge and shortening benefit duration also strikes us as very intuitive. According to row 12 of Table 4, this complementarity will boost outflows from long-term and youth unemployment. It is these results that are particularly interesting from a policy point of view, as they imply that broad reform packages in transition countries that tackle more than one institution or policy will result in better labor market performance than reforming only one institution at a time.

Conclusion

There has been considerable heterogeneity in the evolution of labor markets in TEs over the last two decades, especially at the level of separate countries. The countries of the CIS appear by far the most diverse in terms of economic development and labor market dynamics, with increasing divergence in the most recent years.

The set of labor market institutions that TEs had at the beginning of transition was rather atypical of mature market economies, but has changed considerably over the last two decades. There has been considerable liberalization of labor regulations in TEs, more pronounced than in the old EU member states. By now, most TEs have established sets of labor market institutions and policies, which are similar to those existing in mature market economies.

A growing literature suggests the importance of labor market institutions in the determination of labor market outcomes in TEs. This is particularly true of EPL, the tax wedge, and ALMP. Overall, data from TEs suggest that deregulation of labor markets improves their performance. There is also some

evidence (see Lehmann and Muravyev, 2012) from TEs that reforming two institutions jointly or applying broad reform packages will generate larger benefits than a partial reform focusing on a single labor market institution.

This article also provides first evidence that labor market institutions may matter for informality in the transition countries. This is particularly true of employment protection legislation and unemployment benefit schemes. These effects need to be taken into careful consideration when formulating government labor market policies bent on increased formalization of the economy.

Our empirical analysis, which covers the precrisis years, establishes the robust result that more flexible labor market institutions translate into better labor market performance in the medium to long run. The liberalizing agenda pursued by the Bretton Woods institutions, therefore, seems to have some validity for TEs, at least in 'normal' times.

See also: Employment and Labor, Regulation of; Unemployment: Structural; Youth Joblessness.

Bibliography

Bassanini, A., Duval, R., 2009. Unemployment, institutions, and reform complementarities: re-assessing the aggregate evidence for OECD countries. Oxford Review of Economic Policy 25 (1), 40–59.

Blanchard, O., 2006. European unemployment: the evolution of facts and ideas. Economic Policy 21 (45), 5–59.

Boeri, T., Lehmann, H., 1999. Unemployment and labor market policies in transition countries. Journal of Comparative Economics 27 (1), 1–3.

Boeri, T., Terrell, K., 2002. Institutional determinants of labor reallocation in transition. Journal of Economic Perspectives 16 (1), 51–76.

Botero, J., Djankov, S., La Porta, R., Lopez-de-Silanes, F., Shleifer, A., 2004. The regulation of labor. Quarterly Journal of Economics 119 (3), 1339–1382.

Card, D., Krueger, A.B., 1994. Minimum wages and employment: a case study of the fast food industry in New Jersey and Pennsylvania. American Economic Review 84 (5), 772–793.

Coe, D.T., Snower, D.J., 1997. Policy complementarities: the case for fundamental labor market reform. IMF Staff Papers 44 (1), 1–35.

Djankov, S., Ramalho, R., 2009. Employment laws in developing countries. Journal of Comparative Economics 37 (1), 3–13.

Earle, J., Sabirianova, K., 2002. How late to pay? Understanding wage arrears in Russia. Journal of Labor Economics 20 (3), 661–707.

Eichhorst, W., Feil, M., Braun, C., 2008. What Have We Learned? Assessing Labor Market Institutions and Indicators. IZA Discussion Papers 3470. Institute for the Study of Labor (IZA).

Feldmann, H., 2008. Business regulation and labor market performance around the world. Journal of Regulatory Economics 33 (2), 201–235.

Frölich, M., Kaplan, D., Pages, C., Rigolini, J., D. Robalino (Eds.). Social Insurance and Labor Markets: How to Protect Workers while Creating Good Jobs. Oxford University Press, forthcoming.

Gimpelson, V., Zudina, A., 2011. Informal Workers in the Russian Economy: Who Are They and How many? (In Russian: Neformaly v Rossijskoj Ekonomike: Skolko Ih i Kto Oni?) Higher School of Economics Working Paper WP3/2011/06. Higher School of Economics, Moscow.

Layard, R., Nickell, S., 1999. Labor market institutions and economic performance (chapter 46). In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, first ed., vol. 3. Elsevier, Amsterdam, pp. 3029–3084.

Lehmann, H., Muravyev, A. Labor market institutions and informality in transition and Latin American countries. In: Frölich, M., et al. (Eds.), Social Insurance and Labor Markets: How to Protect Workers while Creating Good Jobs. Oxford University Press, forthcoming.

Lehmann, H., Wadsworth, J., Acquisti, A., 1999. Grime and punishment: job insecurity and wage arrears in the Russian Federation. Journal of Comparative Economics 27 (4), 595–617.

Lehmann, H., Wadsworth, J., 2000. Tenures that shook the world: worker turnover in Russia, Poland and Britain. Journal of Comparative Economics 28 (4), 639–664.

Lehmann, H., Muravyev, A., 2012. Labour market institutions and labour market performance: what can we learn from transition countries? Economics of Transition 20 (2), 235–269.

Lehmann, H., Pignatti, N., 2007. Informal Employment and Labor Market Segmentation in Transition Economies: Evidence from Ukraine. IZA Discussion Papers 3269. Institute for the Study of Labor (IZA).

Neumark, D., Wascher, W.L., 1992. Employment effects of minimum wages and subminimum wages: panel data on state minimum wage laws. Industrial and Labor Relations Review 46 (1), 55–81.

Neumark, D., Washer, W.L., 2008. Minimum Wages. The MIT Press, Cambridge, Massachusetts.

Nickell, S., 1997. Unemployment and labor market rigidities: Europe versus North America. Journal of Economic Perspectives 11 (3), 55–74.

Nickell, S., Nunziata, L., Ochel, W., 2005. Unemployment in the OECD since the 1960s. What do we know? Economic Journal 115 (500), 1–27.

Schneider, F., Buehn, A., Montenegro, C.E., 2010. New estimates for the shadow economies all over the world. International Economic Journal 24 (4), 443–461.

Slonimczyk, F., 2013. Informality as a Stepping Stone. Higher School of Economics Working Paper WP3/2013/05. Higher School of Economics, Moscow.

Slonimczyk, F., 2012. The effect of taxation on informal employment: evidence form the Russian flat tax reform. Research in Labor Economics 34, 55–100.

Svejnar, J., 1999. Labor markets in the transitional Central and Eastern European economies. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics. North-Holland, Amsterdam.

World Bank, 2005. Enhancing Job Opportunities: Eastern Europe and the Former Soviet Union. The World Bank, Washington DC.

Labor Market Policy

Daniel Clegg, School of Social and Political Science, Edinburgh, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Labor market policy (LMP) refers to a set of instruments through which governments assist individuals facing difficulties in the labor market, especially but not exclusively unemployed people. The scope and nature of government activities in this area varies considerably between developed countries, and has evolved over time. Alongside an extensive literature that seeks to evaluate the effects of LMP interventions, there has been a growing interest in recent years in the determinants of this important component of modern states actions to promote the welfare of their citizens.

Introduction

In modern societies, the majority of the working age population relies on dependent employment for access to regular income, and – arguably – for a sense of identity and full social integration. The ability to find and keep a job in a competitive labor market is thus a crucial determinant of individual welfare, while in the absence of public support all but the most transitory periods of unemployment can quickly lead to poverty as well as a range of other personal problems. A government purporting to promote the life chances of citizens cannot therefore ignore the socioeconomic risks individuals face due to the operation modern labor markets.

In the first decades after World War II, high rates of uninterrupted economic growth and a political commitment to full employment ensured that most workers in developed economies had limited exposure to the risk of prolonged unemployment. Since the early 1980s, however, lower growth and the reassertion of low inflation as the principal aim of macroeconomic policy has resulted in unemployment rising in many rich countries, reaching levels last seen in the 1930s even in periods of cyclical expansion. In parallel, rapid deindustrialization has placed those with low or obsolete skills at heightened competitive disadvantage in the labor market. It is as governments have sought to support these victims of structural economic change, and craft instruments for managing labor market risks at the individual level, that labor market policy (LMP) has become an increasingly significant field of public action.

This article introduces the field of LMP and surveys some of the most important themes in its academic study. It first defines the boundaries of LMP and describes the variety of interventions conventionally subsumed by the concept. It next discusses the evolution of LMP in developed countries, emphasizing the substantial cross-national variation in the scope of government activities in this area as well as some common trends in their development over time. A final section reviews the two major scientific approaches to LMP, namely analyses that seek to evaluate the impact of LMP and those that seek to explain its evolution.

Defining LMP

The variety of public policies that impact on the demand or supply side of the labor market, and thus indirectly on the extent and distribution of labor market risks, is very large. Public policies in areas as different as trade, regional development, taxation, working-time regulation, education, or child care, to name but these, might be considered labor market interventions. By convention, however, the concept of LMP has a narrower scope, limited to policies that incur direct governmental expenditure and are targeted directly on groups facing particular difficulties in the labor market.

While the targeting criterion may be useful in delimiting the field of LMP conceptually, it contradicts actual practice in most developed countries, and leads to certain anomalies in the statistical recording of LMP. For example, whenever youth unemployment has risen sharply, governments have tended to react by introducing large youth employment programs, often involving temporary subsidies for private-sector employers when hiring any person under the age of 25 years. Though such measures have a target group far broader than other LMPs, they are recorded in international data on LMP expenditure and participation on the justification that the entire demographic group can be considered as disadvantaged in the labor market. By contrast, measures that subsidize private-sector employers hiring any person below a certain wage level are excluded from these statistics on the grounds of being insufficiently targeted. This is despite the fact that the beneficiaries of the two types of measure largely overlap, each is presented in public discourse as a measure to combat youth unemployment, and they are in all probability in competition for a share of the same public budget allocation. As governments have in recent years turned increasingly to more general measures of labor cost or wage subsidization (see below), expenditure data based on conventional definitions of LMP may thus give a somewhat misleading impression of declining public support for vulnerable groups in the labor market.

Limiting the definition of LMP to interventions incurring public expenditure also carries implications. In addressing youth unemployment, for example, governments in many countries have in recent years placed as much emphasis on regulatory reforms as expenditure measures. New types of employment contract with extended probationary periods are an instance of public policies very directly addressed to the problems facing groups disadvantaged in the labor market but using other policy levers than monetary investments and incentives. Indeed, the issue of employment protection legislation (EPL), and its interaction with expenditure-based LMP, has been central to recent debates around the management of labor market risk, particularly in Europe (see below). While

there may be good intellectual grounds for excluding EPL, or minimum wage legislation, from the formal definition of LMP, it is important to acknowledge that these regulations are equally crucial institutional determinants of labor market performance.

Despite being in some senses overly restrictive, the conventional definition of LMP still encompasses a range of very different public policy instruments. EUROSTAT (2006) identifies three broad types of action, which it calls LMP 'services,' 'measures,' and 'supports.' LMP services refer essentially to the intermediation work of the public employment service (PES), including publication of vacancies and referral of jobseekers as well as more intensive case-management work with unemployed people. Also focused on facilitating transitions from one labor market status to another, LMP measures designate a series of more interventionist policy instruments, including programs involving the direct creation of additional jobs for the unemployed in the public and parapublic sectors and subsidies to private employers that seek to incentivize recruitment of disadvantaged groups. Training programs that aim to enhance the employability of disadvantaged groups through instruction in formal educational contexts and/or in workplaces are a third important type of labor market measure, a category that also includes job sharing schemes, support for unemployed individuals seeking to start their own business and specific measures of occupational rehabilitation for those with reduced work capacity due to illness or disability.

Rather differently, LMP supports provide individuals with cash transfers to compensate them financially for economic disadvantage resulting from their labor market situation. The most important type of support is provided through unemployment compensation, including both unemployment insurance (UI) benefits that are paid to eligible unemployed individuals up to a maximum duration of unemployment, and unemployment assistance (UA) benefits that are paid to those who remain unemployed thereafter or who are not eligible for UI in the first place, for example due to an inadequate work or contribution history. The UI benefits to which an individual is entitled is often related to their previous earnings in employment, while eligibility for UA benefits depends on a means test, meaning the income and assets of an individual, or their household, must be below a specified level to allow them to receive support.

Three additional types of financial compensation for disadvantaged workers are also generally considered to belong to the category of LMP supports. The first are one-off lump-sum payments that individuals may receive from public authorities (rather than from their employers) in the event of their being made redundant due to adverse economic conditions. The second are early retirement benefits, which compensate older workers assumed to have little chance of finding replacement employment for their total or partial retirement from the labor market, up to the age of their entitlement to a regular old-age pension. In some cases the payment of these benefits is made conditional on the retiree's employer hiring a replacement employee from an LMP target group such as young jobseekers. Finally, LMP supports also include measures that compensate short-time working arrangements or 'partial unemployment' linked to a temporary downturn in a firm's economic activities as a result of cyclical or seasonal fluctuations. These policies, which are seen as a way of preventing redundancies by facilitating working-time adjustments, have been mobilized extensively in many countries in the current global economic crisis.

The categorization of LMP into services, measures, and supports offers an alternative to an older approach to classifying these policies, still favored by the Organisation for Economic Cooperation and Development (OECD), the distinction between 'active' and 'passive' LMP. The notion of active LMP maps onto the EUROSTAT categories of services and measures, and has spawned an acronym – ALMP – that is widely used in the academic and policy literature. Passive LMP, in this approach, refers to instruments that compensate individuals financially for the loss or absence of employment, in other words what EUROSTAT considers LMP supports. Though widely used, the active–passive distinction is problematic in a number of ways. Most obviously, the terms are rather loaded; from the respective synonyms of active (energetic, lively, vigorous…) and passive (flaccid, inert, submissive…), it is clear they carry very different evaluative connotations concerning the desirability of these different types of instrument, all of which are arguably important parts of a balanced LMP portfolio. If the distinction is understood to differentiate between policies that facilitate positive transitions and those that compensate negative situations, then it neglects the fact that payment of unemployment benefits itself helps to assist labor market transitions by giving jobseekers the breathing space necessary to find jobs well-matched to their skills and professional aspirations, with potentially beneficial effects for productivity levels in the economy as a whole (Clasen, 1999). Inversely, many supposedly active policies may result in labor market transitions that are far from positive, placing individuals in dead-end schemes that may be stigmatizing for the individual beneficiary and do nothing to improve their future employment prospects.

It remains the case, however, that the field of LMP covers policy instruments of a very different character. The distinction between 'entitlement' and 'discretionary' policies that is commonly used in the American budgetary process is of relevance here. Any individual meeting the existing eligibility requirements has a legal entitlement or right – enforceable in court if necessary – to UI or UA. By contrast, receiving LMP services, participating in measures or benefiting from other supports, such as early retirement measures, is not usually an individual entitlement. These other types of LMPs are thus generally discretionary measures, where public expenditure determines the scope of individual participation and not vice versa. As such, governments are far more likely to make frequent adjustments to discretionary LMP than they do to unemployment benefits. The peculiar nature of unemployment benefits as entitlement programs also explains why they have more often than other LMPs been considered part of social policy, something reflected in some countries in them being separately administered from LMP measures and services. But understandings of the respective functions of, and appropriate interactions between, different LMPs have also evolved over time in developed countries, with changing socioeconomic conditions and political priorities.

LMP Development in Time and Space

The first LMPs to become established components of government action across developed countries were unemployment benefits. Modern unemployment compensation has its origins in mutual insurance arrangements established in the second half of the nineteenth century, usually by trade unions in particular localities or economic sectors. In the early twentieth century, as part of the gradual construction of welfare states, governments in most developed countries introduced compulsory UI schemes at the national level (Alber, 1981). An exception was the Nordic countries, where UI remains voluntary to this day, provided through funds managed by trade unions but heavily subsidized by the state. This has provided a strong incentive for workers in Nordic countries to join trade unions, and is one of the factors behind the high rates of union density still found in these countries.

Antecedents of the modern PES, a nationally coordinated system of public labor exchanges generally developed alongside compulsory UI, performing a range of functions – registering unemployment, monitoring availability for work, and publicizing vacancies – crucial for the integrity of a public unemployment benefit system. With the exception of sporadic experimentations with targeted public works programs at times of high unemployment – and especially during the Great Depression of the 1930s – up to the mid-1970s the scope of LMP interventions was largely limited in the majority of developed countries to unemployment benefits and labor market intermediation services provided through the PES. A notable exception was Sweden, where an ambitious policy of professional training and support for the functional and geographical mobility of workers was from the early 1950s a cornerstone of the famed Rehn–Meidner model, which sought to balance full employment, price stability, growth, and equity. Partly in response to the success of this model, and partly out of a perceived need to facilitate economic modernization, labor market measures emphasizing professional training and requalification were also initiated on a rather smaller scale in countries such as France, Germany, and Italy during the 1960s (Bonoli, 2010).

As unemployment rose across the developed world from the mid-1970s onward, LMP development entered a distinctive new phase. Experimentation with a panoply of LMP measures began in most countries, often with the intention of limiting the fallout from the abandonment of a commitment to full employment. Many measures involved the direct placement of young or long-term unemployed people in socially useful community employment or in large training programs, both of which had the advantage of enabling governments to make quick reductions in the politically sensitive unemployment rate. At the other end of the age distribution, the 1980s also saw a steep increase in expenditure on early retirement, as governments sought to contain unemployment through reductions in the labor supply of older workers. This so-called 'labor shedding' strategy, which was particularly popular in Continental European countries, enjoyed the support not only of trade unions but also of many employers, for whom it offered the possibility of externalizing the cost of profoundly restructuring their workforces (Ebbinghaus, 2006). Allied to escalating expenditure on unemployment benefits, this expansionism saw LMP come to represent an increasingly significant part of total government expenditure. In Denmark, most spectacularly, LMP expenditure peaked in 1993 at more than 7% of GDP.

The OECD 'Jobs Study' (OECD, 1994), published in the wake of the recession of the early 1990s, called for a new direction to be taken in LMP development, and has been an important influence on policy debates ever since. Focused on the problem of structural unemployment, the OECD started from the premise that "policies and systems have made economies rigid, and stalled the ability and even willingness to adapt" (OECD, 1994: 1). More explicitly than in the past, it established an opposition between the logic of LMP measures and services, on the one hand, and supports, on the other, arguing for the need to shift the focus of LMP from the latter to the former. With respect to unemployment benefits specifically, it also argued that benefit entitlements should be reduced and tighter behavioral controls imposed upon recipients to reduce disincentives to work. The concept of labor market 'activation,' which is variously interpreted as increasing the share of expenditure on 'active' instruments in total LMP disbursements or enhancing the work-related conditionality of unemployment benefits, has its intellectual roots in these propositions. Rhetoric aside, however, the Jobs Study was actually also rather critical of many 'active' LMPs, advocating less emphasis on public sector employment and labor market training relative to job-search assistance and well-targeted subsidies to private employers, as well as a closer administrative articulation of LMP measures with benefit provision. The introduction of competition in LMP delivery, through the development of 'quasimarket' mechanisms in the PES and increased contracting-out to private employment service providers, was also promoted.

The Jobs Study's recommendations have been embraced to differing extents in LMP development over the last decade. There has indeed been a marked shift in the composition of expenditure on LMP measures toward what Bonoli (2010) calls 'employment assistance' and 'incentive-reinforcement' approaches, and up to the onset of the 2008 economic crisis expenditure on LMP measures has also tended to decline, driven on the one hand by unemployment falling back from its peak following the recession of the early 1990s, and on the other by governments turning to more general policies to 'make work pay,' such as the subsidization of nonwage labor costs in all lower wage employment to stimulate demand at the bottom end of the labor market and/or the provision of wage supplements to low-paid individuals through the tax or benefit system. 'Active' and 'passive' LMP have also become much more closely integrated at the level of administration and delivery in most countries (Clasen and Clegg, 2011), while 'contractualism' has gained ground as a principle governing relationships between both LMP beneficiaries and the PES, on the one hand, and the financers and providers of LMP measures and services, on the other (Sol and Westerveld, 2005).

By contrast there has been less-aggressive retrenchment of unemployment benefits than the OECD advocated. Across the older members of the European Union, where the most generous unemployment benefit systems could be found, the net replacement rate (NRR) of unemployment benefits remained broadly stable between the middle of the 1990s and the end of the 2000s (Van Vliet and Caminada, 2012). Though

significant declines in replacement rates were seen in some countries – such as the Nordic states and the Netherlands – elsewhere there were substantial improvements in unemployment benefit generosity, as countries like Italy enhanced their traditionally very rudimentary system of LMP supports. The protectiveness of unemployment compensation arrangements has, however, been challenged by more implicit forms of disentitlement, particularly as the deregulation of fixed-term employment contracts – also advocated in the Jobs Study – has resulted in a growing share of people entering unemployment after only brief periods in work, and therefore not having made sufficient social contributions to be entitled to much, or any, support from UI. Early retirement benefits have also been scaled back substantially since the mid-1990s as concern has grown over the sustainability of unfavorable dependency ratios in a context of rapid aging across the developed world.

If the absence of substantial direct retrenchment of generous unemployment benefit systems in Europe can be partly explained by the political risks for governments venturing down this path, it also reflects a more positive appreciation of their role in contemporary LMP than that articulated by the Jobs Study. This is reflected in the rise of the concept of 'flexicurity,' which has become the European Commission's guiding framework for LMP development in the 2000s (European Commission, 2007). Inspired by the Danish policy system, flexicurity is premised on the potential for combining limited regulation of employment contracts, generous and encompassing unemployment benefits and an extensive portfolio of LMP measures to simultaneously offer flexibility for firms and security for workers, while also avoiding the labor market segmentation that results from restrictive EPL. With the rise of the concept of flexicurity, the combined income and employability security provided by LMP has been celebrated over the job security provided by labor market regulation. At least in principle, flexicurity thus seems to move beyond sharp distinctions between active and passive policies that characterized LMP debates for much of the 1980s and 1990s.

While the preceding discussion has traced the broad trends in LMP development, which are largely similar across developed countries, there are also substantial variations in the intensiveness and profile of government actions in this field. Comparative cross-national data on LMP expenditure, and on distinct institutional indicators such as unemployment benefit replacement rates, can help to illustrate these, although they do have their limitations. Expenditure data must be used with caution, not the least given the classificatory anomalies discussed in Section Defining LMP. Meanwhile, as illustrated in **Figure 1**, NRR indicators that take into account the tax treatment of benefits and additional entitlements that recipients may claim in parallel give a rather different picture of unemployment benefit generosity from gross replacement rate (GRR) indicators, which do not. Whether the reference wage used for the calculation of replacement rates is that of an average worker (AW), as in **Figure 1**, or the more traditionally used average production worker (APW) wage can also influence comparisons, as the distance between the AW and APW wage also varies cross-nationally. Most importantly, though, replacement rates are only one of a range of possible parameters on which the design of unemployment benefit systems varies, and can be assessed.

Qualifications notwithstanding, **Figure 1** demonstrates the considerable cross-national variation in unemployment benefit generosity. In the developed countries shown, NRRs range from three-quarters of average wages in Denmark to merely a quarter in Italy. With regard to GRRs, we see the familiar clustering of the English-speaking countries among a group of less-generous systems, though this is attenuated when taxes and additional benefits are included in the calculation. On the NRR indicator, the two Nordic countries are revealed to be the most generous, in keeping with their highly developed welfare states.

Figure 2 illustrates the diversity in the size and composition of overall LMP expenditure across the same countries. As in **Figure 1** we see a clustering of the English-speaking countries, in a group of particularly low overall spenders on LMP. Denmark stands out as an exceptionally high spender on LMP, something that is furthermore not driven by high unemployment; at 3.8%, the 2007 unemployment rate in Denmark was below that of any of the three English-speaking countries. LMP supports are the biggest category of expenditure in all countries except Sweden and the UK, who are otherwise extremely different in their overall commitment to LMP. In Denmark too,

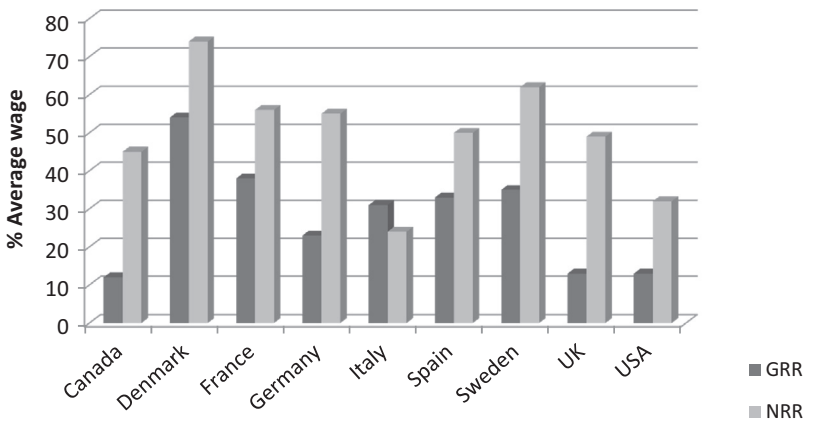

Figure 1 Gross and net unemployment benefit replacement rates, 2007. Source: OECD benefits and wages.

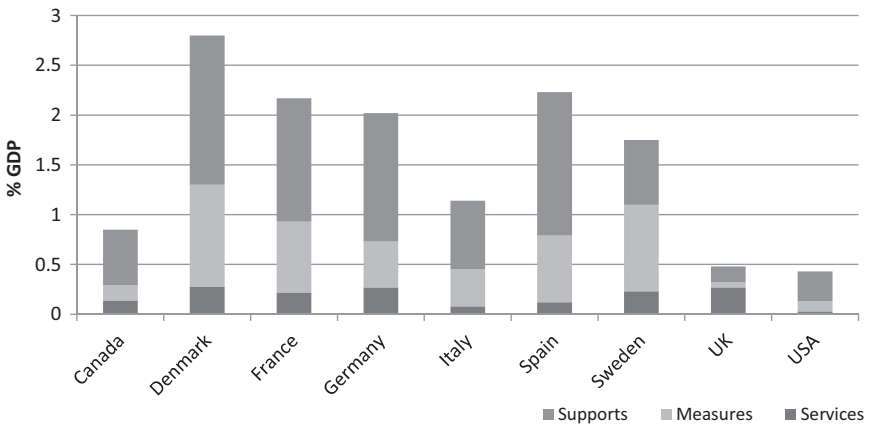

Figure 2 Labor market policy expenditure by category, 2007. Source: own calculations from OECD labor market policy database.

though, combined expenditure on services and measures is close to expenditure on supports. Denmark and Sweden thus stand out as active welfare states, which nonetheless retain comparatively generous unemployment benefit systems.

Evaluative and Explanatory Perspectives on LMP

A large scientific literature has developed alongside the emergence of ALMP as an increasingly crucial component of government action. Next to descriptive studies that enhance our knowledge of LMP across the developed world, and increasingly in transition economies too (Auer et al., 2008), two discrete bodies of work can be distinguished. An evaluative literature, heavily dominated by labor market economists, has sought to assess the impact of different programs and policy instruments, both at the individual and the aggregate level. An explanatory literature, more grounded in political science and political sociology, focuses on the drivers of LMP development, often seeking to account for the very variable levels of commitment to these policies across developed countries.

The literature on LMP evaluation draws on a diverse and increasingly sophisticated range of evaluation approaches and methodologies. A basic distinction is between micro- and macroevaluations, the former interested in the impact of policies on individual outcomes such as employment and earnings, the latter on aggregate outcomes such as employment and unemployment rates. Individual-level evaluations of LMP were long rendered difficult by the lack of adequate microdata, itself a reflection of an absence of a developed 'evaluation culture' in government, particularly in European countries. This problem is abating with time, however, and a recent meta-analysis of evaluations of LMP measures (Card et al., 2010) was able to draw on 97 studies conducted between 1996 and 2007, many providing both short- and medium-term estimates of the impacts of measures. In line with many previous findings, more positive impacts were reported for job-search assistance measures and private-sector employment subsidies than for public sector job creation. Contrary to the findings reported in the Jobs Study, however, positive effects were also found for some training programs, notably when a longer time horizon was adopted.

An even larger number of microevaluations have been published on LMP supports, and especially unemployment compensation provisions. On the whole these depart from the basic premise of neoclassical job-search theory that unconditional provision of unemployment benefits will discourage or lengthen an individual's job search, prolonging unemployment duration and negatively impacting on employment and earnings. Empirical studies have been able to exploit major changes in the unemployment benefit systems to explore these assumptions. A review of evidence on unemployment duration suggests that although the predicted effects are found in many studies, they are typically modest in size (Howell and Rehm, 2009). Furthermore, it is not certain that unemployment duration is the only relevant evaluation parameter. A systematic review of evidence on the effects of unemployment benefit sanctions showed that while there were often discernible effects on exits from unemployment, impacts on individual earnings and employment were far less positive (Griggs and Evans, 2010). Macroevaluations have also explored the relationship between unemployment benefit provisions and unemployment, mostly using quantitative data on relevant policy parameters and outcomes to compare countries with different benefit arrangements. Given the limited number of countries for which reliable data of these kind exist these studies face a 'small N' problem, making it difficult to isolate the causal impact of unemployment benefits on observed variations in unemployment rates, or clarify in which direction any such causality runs. These studies also typically use only one or a few indicators of benefit entitlement, the choice of which can strongly determine results (Howell and Rehm, 2009).

Macroevaluations of LMP measures typically have a more national focus, seeking to identify the employment effects of interventions net of 'displacement' effects, where a beneficiary returns to work only at the expense of an employed person, and 'deadweight' effects, where a beneficiary achieves a positive outcome that they would have achieved even in the absence of the intervention. This literature takes a generally dimmer view of the effectiveness of LMP measures than work based on microevidence (Calmfors, 1994). Such findings do not necessarily disprove the value of such measures, however, as their intended impacts may be largely distributional. More broadly, if evaluative studies – whatever their design or policy

focus – depart from a narrow emphasis on labor market outcomes to investigate the impact of LMP on social outcomes such as poverty and social inclusion, health and even crime, then a rather richer view of their benefits may emerge.

Alongside the impact of changes in socioeconomic environments, the diffusion of evaluative evidence on 'what works' by international agencies may play some role in shaping LMP development in different national contexts. This is particularly the case in the European Union, where since the late 1990s the European Employment Strategy has been developed as an instrument for the explicit supranational coordination of domestic LMP (Goetschy, 1999). However, as LMP impacts upon the wage labor nexus that is central to social relations in capitalist economies, it is likely that the most important determinants of the scope and nature of LMP cross-nationally are political. Extensive LMP 'decommodifies' workers, making them less dependent on selling their labor power in the market, and thereby enhancing their relative bargaining power. In the broader welfare state literature, the significance for the development of generous unemployment compensation of the 'power resources' of the sociopolitical representatives of workers – social democratic parties and trade unions – has long been stressed (Esping-Andersen, 1990). This argument has recently been extended to LMP more broadly (Swank, 2011), though without any systematic testing of the political determinants of different types of LMP measure, which could a priori be expected to appeal to different constituencies in different ways.

The left power approach is not uncontested in the literature on the determinants of LMP, however. It has been argued that social democratic governments in fact do little to positively influence commitments to LMP when employment relationships are also relatively protected, and a divergence in interests opens up between their core electorate of workers ('insiders') and the unemployed and those more precariously attached to the labor market ('outsiders') (Rueda, 2007). With limited risk of becoming unemployed and existing bargaining power, insiders have little to gain from increased expenditure on LMP supports or measures. Indeed, they arguably have something to lose from such policies, in the form of higher taxes and – in the case of LMP measures, at least – intensified competition from outsiders. While the empirical basis for this argument remains controversial, the insider–outsider framework taps into the often ambivalent attitude of the employed population to the unemployed that has without doubt characterized the politics of LMP throughout modern history.

A different challenge to the left power perspective has come from literature on employers' interests in LMP. It has been argued that in protecting workers from socioeconomic risks, LMPs like unemployment compensation simultaneously benefit employers who rely on risk-taking by workers, and particularly their investment in the highly specific skills required in much manufacturing employment (Estevez-Abe et al., 2001). Employers' in manufacturing – long dominant in employer representation in many countries – could on these grounds be expected to have strong preferences for protective UI. While this claim should perhaps not be overdrawn, more nuanced accounts of employers' positions on LMP certainly add something to our understanding of the drivers of policy development. Faced with the prospect of state intervention in this area employers have worked to shape policy development in their preferred direction, in many countries gaining representation alongside trade unions in parapublic institutions through which the policy sector is governed (Mares, 2003). Where such policy institutions are established, furthermore, employers are likely to be socialized into attitudes that acknowledge the benefits as well as the costs of these policies for firms.

This last point illustrates how the development of LMP is crucially shaped by already existing institutional arrangements. Accounts of such effects – too varied and numerous to summarize adequately here – have generally been provided by detailed studies of policy development in one or a few countries. With reference to the UK and Germany, Clasen (1994), for example, showed how the initial financing and governance structure of unemployment benefits shaped policy reactions to mounting unemployment in the 1980s. Schmid et al. (1992) similarly showed the crucial importance of the financing system for LMP supports and measures, which varies considerably across developed countries, on their deployment across the economic cycle. These studies and others like them help explain why governments facing similar labor market problems, armed with identical evaluation evidence and subject to equivalent political pressures and demands still often make substantially different choices in the field of LMP.

Conclusion

The field of LMP covers a very wide array of instruments through which governments intervene on behalf of citizens encountering difficulties in the labor market. Long restricted largely to provision of unemployment compensation and placement services, government interventions in this area have become increasingly diverse since the early 1980s. While commonalities can be seen in the long-run development of LMP in developed countries, striking cross-national differences in the scope and nature of government action remain. As governments in transition and developing countries build up activities in this field, they can draw on a large evaluation literature that has sought to identify the effects of different instruments. This expansion of LMP in turn provides new empirical cases for researchers seeking to better understand the political and institutional dynamics of policy development in this area.

See also: Labor Supply; Poverty Policy; Skills Training: Public Policy Aspects; Unemployment: Structural; Welfare Policy Comparisons; Welfare Retrenchment.

Bibliography

Alber, J., 1981. Government responses to the challenge of unemployment: the development of unemployment insurance in Western Europe. In: Flora, P., Heidenheimer, A. (Eds.), The Development of Welfare States in Europe and America. Transaction Publishers, London, pp. 151–183.

Auer, P., Efendioğlu, Ü., Leschke, J., 2008. Active Labour Market Policies around the World: Coping with the Consequences of Globalization, second ed. ILO, Geneva.

Bonoli, G., 2010. The political economy of active labour market policy. Politics & Society 38 (4), 435–457.
Calmfors, L., 1994. Active labour market policy and unemployment – a framework for the analysis of crucial design features. OECD Economic Studies 22, 7–47.
Card, D., Kluve, J., Weber, A., 2010. Active labour market policy evaluations: a meta-analysis. The Economic Journal 120, F452–F477.
Clasen, J., 1994. Paying the Jobless: A Comparison of Unemployment Benefit Systems in Great Britain and Germany. Ashgate, Aldershot.
Clasen, J., 1999. Beyond social security: the economic value of giving money to unemployed people. European Journal of Social Security 1 (2), 151–180.
Clasen, J., Clegg, D., 2011. Regulating the Risk of Unemployment: National Adaptations to Post-Industrial Labour Markets in Europe. Oxford University Press, Oxford.
Ebbinghaus, B., 2006. Reforming Early Retirement in Europe, Japan and the USA. Oxford University Press, Oxford.
Esping-Andersen, G., 1990. The Three Worlds of Welfare Capitalism. Polity, Cambridge.
Estevez-Abé, M., Iversen, T., Soskice, D., 2001. Social protection and the formation of skills: a reinterpretation of the welfare state. In: Hall, P., Soskice, D. (Eds.), Varieties of Capitalism: The Institutional Foundations of Comparative Advantage. Oxford University Press, Oxford, pp. 145–183.
EUROSTAT, 2006. Labour Market Policy Database Methodology. Office for the Official Publications of the European Communities, Luxembourg revision of June 2006.
European Commission, 2007. Towards Common Principles of Flexicurity: More and Better Jobs through Flexibility and Security (COM(2007) 359 Final). Commission of the European Communities, Brussels.
Goetschy, J., 1999. The European employment strategy: genesis and development. European Journal of Industrial Relations 5 (2), 117–137.
Griggs, J., Evans, M., 2010. Sanctions within Conditional Benefit Systems: A Review of Evidence. Joseph Rowntree Foundation, York.
Howell, D., Rehm, M., 2009. Unemployment compensation and high European unemployment: a reassessment with new benefit indicators. Oxford Review of Economic Policy 25 (1), 60–93.
Mares, I., 2003. The Politics of Social Risk: Business and Welfare State Development. Cambridge University Press, Cambridge.
OECD, 1994. The OECD Jobs Study. OECD, Paris.
Rueda, D., 2007. Social Democracy Inside Out: Partisanship and Labor Market Policy in Advanced Industrialized Democracies. Oxford University Press, Oxford.
Schmid, G., Reissert, B., Bruche, G., 1992. Unemployment Insurance and Active Labor Market Policy: An International Comparison of Financing Systems. Wayne State University Press, Detroit.
Sol, E., Westerveld, M., 2005. Contractualism in Employment Services: A New Form of Welfare State Governance. Kluwer Law, The Hague.
Swank, D., 2011. Activating workers? The political economy of active social policy in postindustrial democracies. Research in the Sociology of Work 22, 9–51.
Van Vliet, O., Caminada, K., 2012. Unemployment Replacement Rates Dataset Among 34 Welfare States, 1971–2009. Leiden University, Leiden. Neujobs Special Report, 2.

Relevant Websites

http://ec.europa.eu/eurostat – EUROSTAT.
http://ec.europa.eu/social – European Commission Employment, Social Affairs and Inclusion.
http://www.eu-employment-obsevratory.net – European Employment Observatory.
http://www.ilo.org – International Labour Organisation.
http://www.oecd.org – OECD.
http://stats.oecd.org – OECD Statistics Portal.

Labor Supply

Bradley T Heim, Indiana University, Bloomington, IN, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J. Pencavel, volume 12, pp. 8195–8204, © 2001, Elsevier Ltd.

Abstract

Studies of labor supply analyze how individuals' decisions of whether to work for pay, and how much to work, are affected by wages, incomes, and other variables. After presenting some empirical regularities, theoretical models are discussed, including static and life cycle models, individual and household models, and models that incorporate income taxation. Finally, the large body of empirical research is reviewed. Though in some areas (such as the size of the intertemporal substitution elasticity and the impact of the Earned Income Tax Credit) a consensus has emerged, there remains considerable disagreement over the magnitudes of many key labor supply parameters.

In economics, the study of 'labor supply' has come to mean the analysis of work behavior where the questions addressed include: How many hours do people work for pay and what factors affect work hours? What proportion of the population works for pay and what factors affect this proportion? The first question, about the number of hours or weeks that individuals work in the labor market, concerns the intensive margin of work. The second question, whether people work or not, concerns the extensive margin of work. The extensive margin may be measured by the employment–population ratio or by the labor force participation rate.

The questions above stress work 'for pay.' This excludes voluntary work and unpaid household work. In fact, data are available on both voluntary work (see Freeman, 1997) and household work (see Juster and Stafford, 1991), and the very same conceptual framework summarized below that is used to study paid work is also employed to understand these other types of work. However, this article concentrates on work for pay and, subsequently, whenever the word 'work' is used, it should be understood that 'work for pay' is meant. Most of the research by economists during recent decades has addressed the supply of labor to the labor market as a whole, not the supply of labor to particular occupations or to particular firms or industries. Research on labor supply to particular occupations or to particular firms uses models similar to the one sketched below to understand the supply of labor to the entire labor market, although there will necessarily be some differences in the analysis. For a review of some research on occupational labor supply, see Freeman (1986). For a study of the labor supply to a single firm, see Nelson (1973). In this article, the research reviewed concerns the supply of labor to the labor market as a whole.

Key Empirical Regularities

Any satisfactory theory of work behavior should explain the following empirical regularities:

1. Men work more than women.
2. In North America and Western Europe, white men work more than black men.
3. In recent years and in relatively rich economies, those who have been at school longer work more.
4. Unmarried men work less than married men while, until recently at least, married women work less than unmarried women.
5. Work hours and employment–population ratios fall in recessions and rise in booms though this procyclical movement is stronger for low-skill than high-skill workers.
6. Over the past century, in North America and Western Europe, for most types of workers, work hours have fallen and vacations have lengthened. A particularly instructive illustration of this for the USA is provided in **Figure 1**, which graphs the usual length of the workday by wage decile. The unbroken line in this graph draws upon information from various surveys in the 1890s, while the broken line provides information from a special Current Population Survey in 2004. The length of the workday has fallen for all workers except those at the highest wage decile. Indeed, whereas low-wage workers once worked the longest hours, high-wage workers now tend to do so. **Figure 1** describes US women, but the data for US men are similar (see Costa, unpublished papera).
7. Among men, work rises with age until middle age and then falls.
8. Although comparisons across countries are frustrated by differences in data collection methods and definitions, it appears that, in the 2000s, average hours worked in the USA exceeded those in most European countries.
9. Average hours worked by self-employed people tend to exceed those by employees (see Aronson, 1991).

No doubt, instances will be found where these patterns are contradicted, but these empirical findings constitute the norm. A rich literature exists in economics on these issues.

One empirical regularity has received much less attention, the fact that individuals synchronize their work schedules: slightly more than 80% of full-time workers in the USA start work between 6 and 9 a.m. while slightly less than 80% finish between 3 and 6 p.m. and these fractions have remained stable since the 1980s (Mellor, 1986); about 78% of all those employed do not work weekends, a number that has increased slightly since the 1980s (Smith, 1986). Presumably there are

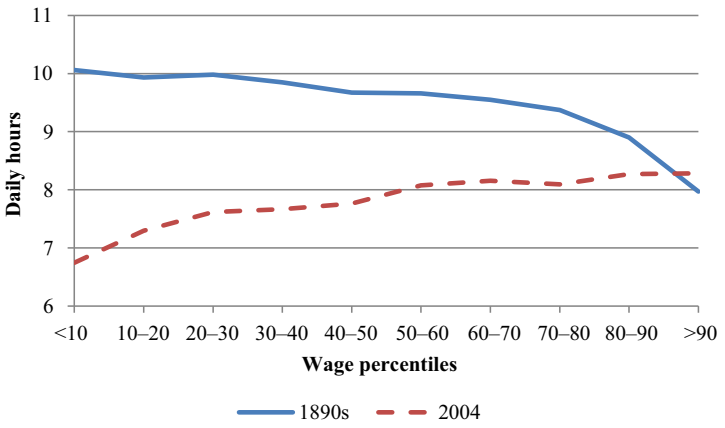

Figure 1 Usual length workday by hourly wage percentile, women aged 18 to 64, 1890s and 2004.

strong complementarities in work and leisure activities (i.e., the benefits to one individual from engaging in a given activity are greater if other individuals also engage in that activity at the same time) although little research effort has been directed toward identifying the key complementarities. Weiss (1996) has been one of the few to grapple with this issue.

An Informal Statement

Classifying the Variables

In essence, economists think in terms of an individual's time spent at work as depending upon four classes of variables: (1) pay per unit time at work; (2) other income; (3) certain demographic characteristics that are observed by researchers; and (4) characteristics of the individual that are unobserved by researchers.

Consider these variables in turn. First, the dependent variable: time spent at work. There are many different dimensions of work time, including annual hours of work, annual weeks worked, weekly hours, daily hours, and whether to work or not.

Now the independent variables. 'Pay per unit time at work' is designed to measure the incentive to work or, equivalently, the price of not working. It is usually measured by an individual's average hourly earnings. Economists usually presume that posttax earnings matter, not pretax earnings. Where average hourly earnings do not vary with time at work, then the use of hourly earnings is appropriate to characterize the incentive to work for pay. However, there are a number of instances where the wage per hour is not fixed but varies (often increases) with work hours, so average hourly earnings may provide only a local indicator of work incentives. For instance, in many countries, by statute or by union-negotiated collective bargaining agreements, or by the operation of market forces, wage premiums are paid for long hours worked per week. Or job offers come in tied wage-hours packages: one employer may offer a job specifying a wage of US$10 per hour for 40 h per week while another may specify a wage of US$12 per hour for 42 h. In these instances, the incentive to work is not provided by a single wage rate that describes all work opportunities.

The variable 'other income' is sometimes represented by the individual's rent, dividends, and interest, so-called nonlabor income. This type of income for an individual is often the consequence of labor income earned at an earlier age and, therefore, will tend to be higher for those who habitually work long hours. In this way, high nonlabor income will tend to be a result of, rather than a cause of, long work hours. Hence, some researchers prefer to use some measure of wealth at a young age (if the data are available to measure it). Or 'other income' is sometimes represented by the labor income of other family members.

It is routine to examine the relationship between time at work and wages holding constant certain demographic characteristics, such as gender and race, characteristics that are typically included in surveys and that are available to researchers. Even allowing for differences in wages and other income, work patterns vary by race and gender, suggesting either systematic differences in 'tastes for work' or different work and nonwork opportunities for these groups.

Finally, it is recognized that much of the variation in work behavior across individuals remains unaccounted for after holding constant all the variables that are usually provided in large data sets. Expressed differently, variables unobserved by researchers 'explain' most of the variation in work across individuals. These unobservables are sometimes called 'tastes for work,' but it should be understood that this is simply giving a name to our ignorance.

The Nature of the Available Data

It is important in understanding any empirical labor supply research to recognize the essential source of variation in the data that is being exploited to measure the effects of various variables on work. One source of variation is across people: the correlation between work and other variables (especially wages and other income) is computed across people observed at approximately the same time. This correlation is often calculated separately for men and women and for different racial groups. A different source of variation is over time: here, in essence, the average values of each individual's work and other variables are first computed, then variations over time around each individual's average in work and other variables are calculated to form the relevant correlations.

Although it is straightforward to compute these correlations, they are sometimes difficult to interpret. Consider the effect of wages on work behavior, a subject that has been the focus of most of economists' research on labor supply. An ideal experiment from which to infer the effect of wages on work would be to select people in the population at random, assign wages to them prior to their work choices, and then observe whether and how much they work. In this experiment, the parts played by the random selection of the population and the prior assignment of wages to people are very important. Unfortunately, nature does not provide us with these ideal experimental conditions.

First, the wages that people face at any moment are determined in part by individuals' past decisions about how long to stay at school and what skills to acquire. These past schooling and training decisions may have been made with the expectation of capitalizing on them subsequently by working longer hours in the labor market. People who are less averse to market work are more likely to stay at school longer and to acquire the skills that result in relatively high wages later in life. In this event, the correlation between work hours and wage rates incorporates a causation that runs from preferences about work hours to wages rather than the other way around. Wages are not assigned to people in the manner described by our ideal experiment; in part, wage differences across people reflect people's prior decisions and their past and current preferences.

Second, when computing the correlation between work hours and wage rates, the wage rates facing people who choose not to work for pay are usually unobserved. Expressed differently, information on the wage offers that nonworkers are rejecting is not available to researchers. This problem is typically addressed by computing the hours–wage correlation among those working positive hours only. However, this generates the problem that the sample of people who choose to work does not represent a random sample of the population but a highly nonrandom and self-selected sample. The requirement of our ideal experiment that the correlation between work and wages be computed for people who constitute a random sample of the population is not met. This is one manifestation of what has become known as the problem of 'sample selection' and it has been emphasized in Heckman's influential research (e.g., Heckman, 1979). The sample selection problem arises whenever factors unobserved by the researcher affect both (1) the selection of the sample over which the hours–wage correlations are to be computed and (2) the hours of work decisions of the individuals. In this case, the ideal experimental condition – the random selection of the sample – is not being satisfied, and the meaning of the hours–wage correlation is subject to more than one interpretation (as shown succinctly by Heckman, 1993).

Therefore, the data available to researchers to compute the effect of wages and other variables on work behavior probably do not satisfy certain ideal experimental conditions, and this frustrates measuring the behavioral parameters governing the impact of these variables on work choices. However, the closer the variation across individuals or across time is to this ideal, the more reliable the resulting estimates are likely to be. For example, estimates from a study that utilizes variation both across individuals and across time, that resulted from policy changes that impacted some areas but not others, will be more reliable than a study that utilizes only cross-sectional variation in gross wages that resulted from individual schooling and occupation choices.

Conceptual Framework

Static Models of Labor Supply

To model an individual's work decision, let his or her preferences for work hours (h) and consumption (c) be embodied in a utility function $U = U(c, h; X, e)$, where X denotes fixed characteristics of the individual such as gender and race that are observed by researchers. Characteristics affecting preferences that are not observed by researchers are given by 'e'. The individual's income–expenditure constraint is expressed as $p \cdot c = w \cdot h + y$, where p is the price of consumer goods, y is nonlabor income, and w is average hourly earnings.

In the simplest and most tractable of cases, p, w, and y are fixed to the individual and independent of choices of c and h. In this instance, provided U has some convenient technical properties, selecting c and h to maximize U subject to the income–expenditure constraint yields an expression that relates hours worked, h, to the budget constraint variables that are given to the individual, namely, p, w, and y, and to the fixed characteristics conditioning his or her preferences, X and e: $h = f(p, w, y; X, e)$. This relationship is the labor supply function and it provides a formal justification for the characterization given in Section Conceptual Framework, whereby time spent at work is related to four classes of variables: w, y, X, and e. (Because the budget constraint may be written $c = (w/p) \cdot h + (y/p)$, it is evident that there are not three independent price and income variables, namely, p, w, and y, but only two such variables, namely, w/p and y/p.) As usual in economics, for the key empirical propositions, it is not essential that any individual behaves in this easily caricatured, maximizing manner, but that it facilitates the derivation of relationships and provides a convenient benchmark if we assume that the individual behaves as if he were deliberately optimizing. Less terse explanations of this model can be found in many good textbooks of Ehrenberg and Smith (2011) or Borjas (2012).

This framework applies to the individual who chooses to work a positive number of hours ($h > 0$) and to the individual who chooses not to work at all ($h = 0$). In the case of the decision of whether to work, the individual who is currently not working compares the wage per hour he would receive if he worked, the 'offered market wage,' with his reservation wage that is defined as the implicit value of his nonwork time. When his reservation wage exceeds his offered wage, he does not work for pay ($h = 0$); when his reservation wage falls short of his offered wage, he works some positive hours in the market ($h > 0$). Therefore, the higher is the offered wage, the greater the likelihood of an individual working for pay.

The previous statement concerns an individual's work decisions. Many people live in households where joint work decisions have to be made. In these cases, the standard treatment postulates a household utility function that depends upon the household's consumption and the work time of each household member. If incomes and expenditures are pooled, then each household member's work behavior will depend not

only on his wage rate but also on the wage rates of other household members.

This household formulation has been the subject of criticism for failing to specify how the household utility function is derived and for being silent on how resources are allocated within the household. In addition, many of the implications of this formulation tend to be rejected empirically (see Chiappori and Donni, 2009). Some recent models, then, start with the utility functions of individual household members each of which factor into the behavior of the household.

One version of these models characterizes household decision making as a bargaining process. Each individual has a level of welfare that can be enjoyed in the absence of cooperation with other members of the household. This is the individual's 'threat point' and is sometimes equated with the welfare the individual would enjoy if he or she left the household. Provided the household remains intact, anything that increases an individual's threat point will increase the individual's share of the benefits from remaining in the household. For instance, if the market wages paid to unmarried women rise, this increases the wife's threat point in bargaining with her husband and, provided the wife remains married, she should enjoy a greater share of resources within the household. This model has been used not only to understand husband–wife decision making and the growing incidence of divorce, but also intergenerational behavior in developing countries when adult children live with their parents. A formal treatment of these issues is found in McElroy (1990) and a more accessible discussion is in Lundberg and Pollak (1996).

Another version of these models does not take a stand on the form of the decision making process within the household, but postulates that the household reaches efficient decisions with respect to its consumption and work behavior. 'Efficiency' here means that, given prices and wages, the work and consumption decisions of the household members have exploited all opportunities to enhance their welfare without cost so that one person's welfare can be improved only by lowering another member's welfare (Chiappori, 1988). Recently, such models have been extended to account for labor force nonparticipation, nonlinear budget constraints, and discrete labor supply choices (Donni, 2003; Blundell et al., 1997).

Life Cycle Models of Labor Supply

When understanding work behavior over time, it becomes important to consider life cycle or dynamic models of labor supply.

In such models, there are two sources of variation in wages to which individuals may adjust their labor supply. The first is the life cycle variation in wages for a given individual; that is, the evolutionary changes in wages as a person ages. Typically, as an individual ages, their wages rise, sometimes at a decreasing rate, until later in life when wages begin to fall. The second source of variation in wages is the difference in wages across people at the same stage in their life cycle. This second source of variation in wages is called the parametric difference in wages across individuals. At any moment, differences in work behavior among people will be a mixture of responses of labor supply to, first, the path of wages as a given individual ages and, second, differences across individuals in wages observed at the same age.

The first of these responses may be observed if the work behavior and wages of a cohort of people born at about the same time are followed as they age. The proportional change in hours worked that follows a proportional change in wages as these individuals age is called the intertemporal substitution elasticity. It is the life cycle effect of wages on work.

The second of these responses may be observed by sorting individuals by age group and computing the correlation between work and wages across people within each age group. These people are about the same age, but they are born in different years (as they belong to different cohorts) so that the observations on their work and wages must be drawn from different calendar years. This correlation incorporates the response of labor supply to parametric differences in wages across people.

Labor Supply and Taxes

Because changes in income and payroll taxes, as well as changes in welfare benefits, are tantamount to changes in w (and sometimes in y, too), the impact of these tax and welfare policies depend on whether, and by, how much hours worked and employment–population ratios increase or decrease as w is altered. As a result, features of tax and transfer policies are often incorporated into labor supply models.

If t is the tax rate, w represents gross (that is, pretax) hourly earnings, and y is gross (pretax) nonlabor income, then net or posttax wages are defined as $w^n = (1-t) \cdot w$ and net or posttax nonlabor income is $y^n = (1-t) \cdot y$. If each individual responds to net wage rates and net nonlabor income, then the relevant budget constraint involves the posttax variables: $p \cdot c = w^n \cdot h + y^n$.

The problem posed for researchers measuring the effect of income taxes on labor supply is that, in most income tax systems, tax rates change with income (which, in turn, changes with work hours) so that the net wage rate varies as an individual alters his work hours. This means that the incentive to work is not represented by a value for hourly earnings that is fixed to the individual and independent of their work choice. Instead, there is no single value that represents the incentive to work because the net hourly wage at, say, 20 h of work per week will diverge from the net hourly wage at 40 h per week. In this instance, the budget constraint, $p \cdot c = w^n \cdot h + y^n$, is no longer linear.

It is not only the explicit income tax system that produces nonlinearities in the budget constraint. Most income-maintenance programs designed to provide support for low-income people also result in nonlinear budget constraints for individuals eligible for such programs. This is because most income-maintenance programs do not transfer a fixed sum that is independent of the recipient's income. On the contrary, most transfer programs are income conditioned: less money is transferred to the recipient as the recipient's income rises, until a point is reached, the so-called break-even point, where transfers are reduced to zero and the individual moves off the welfare program and is faced with just the (positive) income tax system. The rate by which transfer income is reduced as the welfare recipient earns one more dollar is the welfare system's

implicit tax rate. Virtually all implicit tax rates are very high, often as high as 100% when transfer income is reduced by one dollar for every dollar that the individual earns, so that budget constraints of low-income people are highly nonlinear.

Empirical Research

A large amount of empirical research has been directed toward the relationship between hours and wages.

In labor economics research in the 1970s and 1980s, the most common source of information to examine the relationship between hours worked and average hourly earnings consisted of large cross-sectional data sets; that is, data on individual people observed at approximately the same calendar time. In essence, after controlling for differences among people in many characteristics, the residual correlation was computed between time spent at work and average hourly earnings. In other words, the impact of wages on work was inferred from differences among people in their wages and work after controlling for other variables.

The most common finding in this research was that, among adult men, hours worked and wage rates were mildly negatively correlated; whereas, among adult women, hours worked and wage rates were positively correlated. A summary and evaluation of these studies appear in Killingsworth (1983), Pencavel (1986), and Killingsworth and Heckman (1986). However, the particular results were disturbingly fickle: relatively small changes in the composition of the sample or in estimating methods sometimes had large effects on the magnitude of the work–wage correlations. Mroz (1987) provides an excellent demonstration of this. Nevertheless, some recent papers that have employed similar methodologies (Blau and Kahn, 2007; Heim, 2007) have found that the correlation between hours and wages among married women has dropped substantially in the past two decades, suggesting that wages have less of an impact on hours among this group now than was the case in the 1970s and 1980s.

The first study to use data that followed people over time and estimate the two different types of labor supply responses in a life cyle model was MaCurdy (1981). His estimates of the intertemporal substitution elasticity for adult US men were in the order of 0.23, meaning that a 10% growth in wages as an individual ages would induce a 2.3% increase in annual hours worked. His estimates were imprecise, but more recent estimates for both men and women derive similar values for the intertemporal substitution elasticity (see Pencavel, 1998, unpublished paper).

When individuals are sorted by age and the correlation between work and wages computed across people holding age approximately constant, the subsequent estimate of the impact of wages on work is often (though not invariably) positive: for both US men and women, a representative estimated elasticity of annual hours with respect to wages is of the order of 0.10. (The elasticity of hours with respect to wages is defined as the proportional change in hours induced by a proportional increase in wages.)

Whereas in the research in the 1970s and 1980s it was common to compute negative estimates of labor supply elasticities, in the 1990s the estimates were more frequently positive and appeared less sensitive to modifications in specification and methods. Why should this be? First, there was a better understanding about the need to distinguish conceptually among different possible wage effects on work behavior. In particular, researchers differentiate the incentives for an individual to work at ages when his wages are relatively high (this is the life cycle effect) from the incentives to work that come from wage differences arising for reasons that are independent of age (this is the effect of parametric differences in wages). Second, the 1990s research had an important time series dimension, which may be a critical feature given that, over the past 20 years or so, the structure of wages has changed considerably. This generates a variation in wages that was denied to researchers in the 1970s and 1980s who relied exclusively on wage differences among people at a given moment.

Though there remains some disagreement among economists regarding the value of the elasticity of hours worked with respect to wages, it is undeniable that positive estimates of the effect of wages on work would help to explain many features of changes in work behavior over the past 25 years in the USA. Work levels have increased more for women than men, they have increased more for well-educated people than poorly educated people, and they have increased more for married people than for single people. In each of these three comparisons, the wage structure has changed favoring those who work more: the wages of women have risen relative to those of men, the wages of well-educated workers have risen relative to those with less education, and the wages of married people have risen relative to the wages of unmarried people. One simple story compatible with these facts is that changes in technology and other factors have caused changes in the wage structure and – if work–wage relationships are positive – these wage changes have induced relatively more work from those whose wages have risen relative to those whose wages have fallen.

Positive work–wage estimates may also explain some of the empirical regularities listed in Section An Informal Statement. Men work more than women and, other things equal, men are paid more per hour than women. White men work more than black men and, other things equal, white men are paid more than black men. Those with more schooling work more and have higher average hourly earnings. Unmarried men work less than married men, and unmarried men tend to have lower wages than married men. Unmarried women work more than married women, and unmarried women tend to have higher hourly pay than married women. The work–age patterns for men broadly correspond to their wage–age patterns. No doubt, there are other competing explanations for these regularities, but the naive positive association between work and wages across these groups is striking.

Other Empirical Findings

Labor Supply and Taxes

One aspect of empirical labor supply research that has occupied a large amount of attention is the impact of income taxation and transfer programs on work behavior.

Studies of this type done in the 1980s and 1990s often attempted to reproduce the essential features of individuals'

posttax and posttransfer budget constraints and explicitly estimate the parameters of the economic model. When the tax system can be characterized as generating a posttax budget constraint for an individual in which posttax wage rates are continually falling as more hours are worked (i.e., a budget constraint that is strictly concave from below as implied by a smoothly progressive tax system), the statistical procedures for estimating the effect of wages and taxes on work behavior are relatively straightforward and tractable (Hausman, 1981). When, however, the posttax budget constraint has kinks and net wages may rise over some regions as hours increase, the statistical methods become more complex and the results more sensitive to alternative estimation methods. Blundell and MaCurdy (1999) survey results from these studies. Though the wage effects estimated for married women are typically larger than those found for men, and though most studies find negative income effects, the variation in the estimates is striking. The essential source of the difficulty for researchers is that most data sets provide little information about the income tax and transfer programs that households face. Consequently, with the information available on gross wages and income, in effect, researchers try to complete the income tax and transfer algorithms for each individual, but the inaccuracies in this process and the errors introduced must be extensive. In addition, some of these methods implicitly constrain the estimates to imply positive wage effects and negative income effects (MaCurdy et al. (1990)).

A similar problem exists in trying to determine the impact of welfare or income maintenance programs on labor supply. A popular hypothesis is that some people reduce their work levels and, therefore, their market earnings to become eligible for welfare. This is revealed, it is argued, by the fact that welfare participation (the fraction of households receiving welfare payments) is higher in those states in the USA where the welfare program parameters imply relatively generous benefits. The problem with this argument is that, even in the absence of any work incentive effects of the welfare system, an association between welfare participation and the parameters of the welfare system are to be expected. Where welfare program parameters imply relatively generous welfare benefits, households will accept opportunities to raise their disposable income and participate in the welfare program, even if they do not change their work behavior. The association between welfare participation and welfare program parameters (especially the welfare system's implicit tax rate) has to be specified carefully to identify work incentive effects (see Ashenfelter, 1983). Unfortunately, because very few differences across states exist in the statutory implicit tax rate on welfare income, little persuasive evidence on the work disincentive effects can be gleaned from state to state variation in policies.

Much of the recent research on labor supply has been devoted to examining the impact of changes in particular program parameters on labor supply and labor force participation. A popular focus of study has been the Earned Income Tax Credit (EITC) (or a related credit in the UK, the Working Families Tax Credit) which provides a sizable tax credit for low-income working individuals. These studies typically use estimation strategies that compare a group impacted by the policy change (often single mothers) to a comparison group similarly situated but not affected by the change. A recent survey by Eissa and Hoynes (2006) notes that studies on the EITC generally find that labor force participation rates increase in response to an expansion of the credit, but that hours of work are largely unaffected.

Some researchers have moved away from examining the impact of posttax wages on hours of work, and instead attempt to measure the impact on tax rates on earned incomes. The rationale for this switch is that hours of work may be a poor measure of work effort for some individuals (such as executives and those working on commission) who have a large amount of discretion over their work activities. These studies (surveyed in Saez et al., 2012) also often compare individuals who were and were not impacted by a change in tax rates to identify the impact of the change, and typically find that earned incomes increase in response to a decrease in tax rates.

Family Labor Supply Decisions

The employment probabilities of husbands and wives do not appear to be independent of one another. To show this, consider households in the USA where both the husband and the wife are aged between 25 and 64 years. (The data are drawn from the 2012 March Current Population Survey, which asks questions about labor force behavior in 2011.) **Table 1** presents the joint and unconditional probabilities of market work of these husbands and wives. The probability that both the wife and the husband work is 0.633 and the probability that the wife works but the husband does not is 0.071. Hence, the sum of these two joint probabilities is the unconditional probability of the wife working which is 0.704 (given in the bottom line of **Table 1**). Similarly the probability that the wife does not work but the husband does is 0.242 and (as already noted) the probability that both the wife and the husband work is 0.633. Hence, the sum of these two joint probabilities yields the unconditional probability of the husband working which is 0.875 (given in the last column of **Table 1**).

Table 1 Joint and unconditional probabilities of employment of husbands and wives in the USA, 2011

		Probability of wife		
		Not working	*Working*	*Unconditional probability*
Probability of husband	Not working	0.054	0.071	0.125 (Probability of husband not working)
	Working	0.242	0.633	0.875 (Probability of husband working)
Unconditional probability		0.296 (Probability of wife not working)	0.704 (Probability of wife working)	

From these joint and unconditional probabilities can be derived the conditional probabilities and these are listed in Table 2. For instance, the probability that the wife works conditional on the husband working is 0.723, given in the first line of Table 2. Or, the probability that the husband works conditional on the wife also working is 0.899, given in the third line of Table 2. If both the husband and wife's employment probabilities were independent, then the conditional employment probabilities in Table 2 would equal the unconditional probabilities in the final column and the final line of Table 1. In fact, the conditional probabilities of the wife working are 0.723 and 0.568 (given in the first two lines of Table 2), whereas the unconditional probability of the wife working is 0.704 (from Table 1). The conditional probabilities of the husband working are 0.899 and 0.818 (given in the last two lines of Table 2) while the unconditional probability of the husband working is 0.875 (from Table 1). These data suggest that employment independence is closer to being satisfied for husbands than it is for wives. However, for both husbands and wives, the probability that one is employed is greater if his or her spouse is also employed. Consistent with this, working husbands and wives tend to retire at, or close to, the same time (see Hurd, 1990; Costa, unpublished paperb).

There are a number of reasons why employment probabilities of husbands and wives are positively correlated. First, this may reflect something about marital choices: women less averse to market work tend to marry men with similar work preferences. Second, the correlation may be that each party's responses to their shared economic environment. For instance, if both husbands and wives work more in households possessing little nonlabor income, then the positive correlation in work behavior may reflect the common influence of this source of wealth. Third, the correlation may be the outcome of both marriage patterns and economic environments. For example, it has already been reported that those men facing higher wages are more likely to work in the market than lower wage men and those women facing higher wages are more likely to work in the market than lower wage women (1). Furthermore, high-wage men tend to marry high-wage women (2). Together (1) and (2) will generate a positive correlation between the employment probabilities of husbands and wives.

The positive correlation in the market work of husbands and wives may also reflect cross-wage effects. That is, hold the wages of the husband constant: then, do those husbands married to relatively high-wage wives work more than those husbands married to relatively low-wage wives? Conversely, hold the wages of the wife constant: do those wives married to relatively high-wage husbands work more than those wives married to relatively low-wage husbands? This issue has arisen in the last few decades as a possible reason for the growth in the employment–population ratio of married women. A popular explanation has focused on the slow growth in the wages of married men since the early 1970s: are married women working more because the wages of their husbands have increased so little or, in some instances, have actually decreased? The broad trends in the key variables support such an interpretation because there has been a steady increase in the market work of married women while, for the past 25 years or so, median real wages of married men have increased hardly at all.

Indeed, the work decisions of husbands and wives do seem to be related to the wages of their spouses. In particular, when one spouse's real pay falls, the other spouse tends to work more (see Juhn and Murphy, 1997; Pencavel, 1998, unpublished paper; Heim, 2009). However, these cross-wage effects – the relation between one individual's work behavior and his or her spouse's real earnings – are much less powerful than the own-wage effects – the relation between each individual's work behavior and his or her own real wages. According to these findings, the employment of married women has increased principally because their real offered wages have been increasing and the movements of their husband's earnings have been a factor, but a factor of less importance in accounting for the movements in their work levels.

If women were working more to augment the stagnant pay of their husbands, then the increases in the work of married women should be greatest for those men whose wages have fallen, not for those men whose wages have risen. In fact, the largest increases in work among women have occurred for those women married to high-wage men and the real earnings of high-wage men have tended to increase in real terms since the 1970s. By contrast, the market work of wives of low-wage men has increased much less and yet the real wages of low-wage men have tended to fall. These facts are reported in Juhn and Murphy, (1997).

Hence this evidence suggests that husbands and wives tend to substitute their work time for one another: when the wage rate of one party increases (decreases), that person works more (less) and his wife or her husband works less (more). The positive correlation in the employment probabilities of husbands and wives is driven more by marriage patterns and by own-wage effects than by cross-wage effects.

Table 2 Conditional probabilities of employment of husbands and wives in the USA, 2011

Probability of wife working conditional on husband working	=Probability of both husband and wife working divided by the unconditional probability of husband working	=0.633/0.875	=0.723
Probability of wife working conditional on husband not working	=Probability of wife working and husband not working divided by the unconditional probability of husband not working	=0.071/0.125	=0.568
Probability of husband working conditional on wife working	=Probability of both husband and wife working divided by the unconditional probability of wife working	=0.633/0.704	=0.899
Probability of husband working conditional on wife not working	=Probability of husband working and wife not working divided by the unconditional probability of wife not working	=0.242/0.296	=0.818

Work Behavior of Older People

One of the most striking of long-term developments in work behavior is the decline in employment of older men. Of men 65 years and older, 68% were in the labor force in 1890, 42% in 1940, 31% in 1960, and fewer than 20% in the mid-1980s. Interestingly, from the mid-1980s to 2010, this decline appears to have flattened out or reversed slightly (Kromer and Howard, 2013). Similar declines in hours of work among those men employed are also observed. By contrast, the labor force participation rate of women aged 65 years and over has changed relatively little: it was 8% in 1890, 8–10% in the 1980s and 1990s, and 12.5% in 2010.

The drop in the amount of work allocated by older men to the market is one definition of retirement. Another definition, often accompanying the first, turns on the receipt of pension or social security income. It is tempting to convert this association into a causal relationship: retirement in the sense of the reduction in labor supply among older workers is induced by the opportunity to draw on annuities from accumulated pension funds.

Because the dominant retirement income for the median elderly household is social security income, a much researched issue has been the impact of social security on retirement. For most of the past 50 years or so, social security wealth has been increasing over time while male labor force participation rates have been falling so the aggregate time series movements in these variables are compatible with a causal role for public pensions. However, monthly income from social security was first paid in 1940 to a relatively small fraction of the elderly population at which time labor force participation rates of men of 65 years and over had already fallen by more than 20% since the beginning of the century. As Costa has noted (1998: pp. 20–21), in the USA and other countries (Britain, Germany, and France), retirement trends do not display a discontinuity around years when, as a consequence of new legislation, public pension benefits start being paid.

However, a role for social security affecting retirement behavior is suggested from an examination of the retirement patterns of the members of a given cohort as they age. In particular, since the mid-1960s when early social security benefits were first introduced, there has been a spike in male labor force participation rates at age 62 years. It is difficult to understand why there should be such a spike at 62 years other than the discontinuous retirement incentives embodied in the Social Security benefit rules.

In addition, certain private pension programs create incentives to retire before the age of 65 years. Defined-benefit pension plans specify the pension annuity as a fixed proportion of a worker's highest annual earnings and often the actuarial value of such defined-benefit plans is higher for people younger than 65. In contrast, defined-contribution plans embody no incentives for early retirement, since pension wealth grows until the funds are withdrawn. Because the fraction of workers covered by defined-benefit pension programs has been falling and that of defined-contribution pension plans increasing, the prevalence of early retirement incentives in private pension programs has been on the decline.

Nevertheless, the growth in social security and in private pensions represent specific manifestations of a more general phenomenon; namely, the increase in wealth that many cohorts have enjoyed this century. This increase in wealth discourages market work at some time during an individual's life and it may be efficient for the typical individual to take this increased leisure not in the form of fewer work hours each year, but in the form of fewer years at work and of earlier retirement.

Differences in Work–Wage Elasticities

In undergraduate textbooks, for many years it has been common to draw graphs of the relationship between hours worked and wages that imply a positive association at relatively low wages and a negative association at relatively high wages. This nonlinear relationship is called a 'backward bending labor supply curve.' Although this curve is a staple of every economics student's education, the empirical support for this relationship is meager. In some instances, researchers have specified the relationships in such a way that a nonlinear and nonmonotonic relationship between wages and work hours was not investigated. In other instances, however, researchers have explored the possibility of this backward-bending curve and found little evidence for it.

With the richer and more recent data, the issue has come up for re-examination. It surfaced because the impact of wages on work appeared to be larger for men with relatively low levels of schooling than for well-educated men, and it was a short step to move from this observation to a specification in which the impact of wages on labor supply fell as wages rose. Indeed, according to one set of estimates, the elasticity of annual hours worked with respect to wages is calculated to be of the order of 0.33 at the average wage of men with less than a high school education, and it falls to 0.14 at the average wage of men with more than 16 years of schooling (see Pencavel, unpublished paper). Note, however, that although the slope of the hours–wage relationship becomes less positive at higher wages, it never becomes negative, except for a very few people with extremely high wages. A similar result is obtained by Juhn and Murphy (1997), though Kapteyn et al. (1990) and Bloemen (2009) find evidence of a backward bending labor supply function for larger fractions of married men.

Conclusion

Over the past 50 years or so, there has been a very large volume of research on labor supply by economists. This is because the work decisions of individuals represent one of the essential components of labor market analysis (in the classical metaphor, these labor supply decisions represent 'one blade of the scissors,' the other blade being the hiring and employment decisions of firms) and because questions involving labor supply are closely tied to important issues of the public policy (such as the effect of income taxes on how much individuals work). This research has been increasingly sophisticated in economic and statistical modeling. In addition, much more is known today about work patterns than was known in 1960. Some components of the economist's standard model of labor supply (such as the intertemporal substitution elasticity), and the labor force participation impacts of some policies (such as the EITC) have been estimated with confidence. However, there remains some disagreement over the values of other key

parameters such as the effect across different individuals of increases in wages on hours worked. Most economists probably believe that, for the typical worker, increases in wages generally induce relatively small increases in work hours although researchers would refrain from confident statements about precise magnitudes.

See also: Economic Geography; Employment and Labor, Regulation of; Labor, Division of; Unemployment: Structural; Work, Anthropology of; Work, Sociology of.

Bibliography

Aronson, R.L., 1991. Self-employment: A Labor Market Perspective. ILR Press, Ithaca, NY.

Ashenfelter, O., 1983. Determining participation in income-tested social programs. Journal of the American Statistical Association 78, 517–525.

Blau, F.D., Kahn, L.M., 2007. Changes in the labor supply behavior of married women: 1980–2000. Journal of Labor Economics 25, 393–438.

Bloemen, H.G., 2009. An empirical model of collective household labour supply with non-participation. The Economic Journal 120, 183–214.

Blundell, R., Chiappori, P.-A., Magnac, T., Meghir, C., 1997. Collective labor supply: heterogeneity and non-participation. Review of Economic Studies 74, 417–445.

Blundell, R., MaCurdy, T., 1999. Labor supply: a review of alternative approaches. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3a. North-Holland, Amsterdam, pp. 1559–1696.

Borjas, G., 2012. Labor Economics, sixth ed. McGraw-Hill/Irwin, Boston, MA.

Chiappori, P.-A., 1988. Rational household labor supply. Econometrica 56, 63–90.

Chiappori, P.-A., Donni, O., 2009. Non-unitary Models of Household Behavior: A Survey of the Literature. IZA Discussion Paper No. 4603.

Costa, D., 1998. The Evolution of Retirement: An American Economic History, 1880–1990. University of Chicago Press, Chicago, IL.

Costa, D. The Wage and the Length of the Work Day: From the 1890s to 1991. Department of Economics M.I.T. and National Bureau of Economic Research, unpublished paper a.

Costa, D. The Rise of Joint Retirement, 1940–1998, unpublished paper b.

Donni, O., 2003. Collective household labor supply: non-participation and income taxation. Journal of Public Economics 87, 1179–1198.

Eissa, N., Hoynes, H.W., 2006. Behavioral responses to taxes: lessons from the EITC and labor supply. In: Poterba, J. (Ed.), Tax Policy and the Economy, vol. 20. MIT Press, Cambridge, MA, pp. 73–110.

Ehrenberg, R.G., Smith, R.S., 2011. Modern Labor Economics: Theory and Public Policy, eleventh ed. Prentice Hall, Upper Saddle River, NJ.

Filer, R.K., Hamermesh, D.S., Rees, A.E., 1996. The Economics of Work and Pay, sixth ed. HarperCollins, New York.

Freeman, R.B., 1986. Demand for education. In: Ashenfelter, O., Layard, R. (Eds.), Handbook of Labor Economics. North-Holland, Amsterdam, pp. 357–386.

Freeman, R.B., 1997. Working for nothing: the supply of volunteer labor. Journal of Labor Economics 15, S140–S166.

Hausman, J.A., 1981. Labor Supply. In: Aaron, H.J., Pechman, J. (Eds.), How Taxes Affect Economic Behavior. Brookings Institution, Washington, DC.

Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–162.

Heckman, J.J., 1993. What has been learned about labor supply in the past twenty years? American Economic Review, Papers and Proceedings 83, 116–121.

Heim, B.T., 2007. The incredible shrinking elasticities: married female labor supply, 1987–2002. Journal of Human Resources 42, 881–918.

Heim, B.T., 2009. Structural estimation of family labor supply with taxes: estimating a continuous hours model using a direct utility specification. Journal of Human Resources 44, 350–385.

Hurd, M., 1990. The joint retirement decision of husbands and wives. In: Wise, D.A. (Ed.), Issues in the Economics of Aging. University of Chicago Press, Chicago, IL, pp. 231–254.

Juhn, C., Murphy, K.M., 1997. Wage inequality and family labor supply. Journal of Labor Economics 15, 72–97.

Juster, F.T., Stafford, F., 1991. The allocation of time: empirical findings, behavioral models, and problems of measurement. Journal of Economic Literature 39, 471–522.

Kapteyn, A., Kooreman, P., van Soest, A., 1990. Quantity rationing and concavity in a flexible household labor supply model. The Review of Economics and Statistics 72, 55–62.

Killingsworth, M.R., 1983. Labor Supply. Cambridge University Press, Cambridge, UK.

Killingsworth, M.R., Heckman, J.J., 1986. Female labor supply: a survey. In: Ashenfelter, O., Layard, R. (Eds.), Handbook of Labor Economics. North-Holland, Amsterdam, pp. 103–204.

Kromer, B., Howard, D., 2013. Labor Force Participation and Work Status of People 65 Years and Older. American Community Survey Briefs 11-09.

Lundberg, S., Pollak, R.A., 1996. Bargaining and distribution in marriage. Journal of Economic Perspectives 10, 139–158.

MaCurdy, T.E., 1981. An empirical model of labor supply in a life-cycle setting. Journal of Political Economy 89, 1059–1085.

MaCurdy, T., Green, D., Paarsch, H., 1990. Assessing empirical approaches for analyzing taxes and labor supply. Journal of Human Resources 25, 415–490.

McElroy, M.B., 1990. The empirical content of Nash-bargained household behavior. Journal of Human Resources 25, 559–583.

Mellor, E.F., 1986. Shift work and flextime: how prevalent are they? Monthly Labor Review 109, 14–21.

Mroz, T.A., 1987. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. Econometrica 55, 765–799.

Nelson, P., 1973. The elasticity of labor supply to the individual firm. Econometrica 41, 853–866.

Pencavel, J., 1986. Labor supply of men: a review. In: Ashenfelter, O., Layard, R. (Eds.), Handbook of Labor Economics. North-Holland, Amsterdam, pp. 3–102.

Pencavel, J., 1998. The market work behavior and wages of women. Journal of Human Resources 33, 771–804.

Pencavel, J. A Cohort Analysis of the Association between Work and Wages among Men. Department of Economics Stanford University, Stanford, CA, unpublished paper.

Saez, E., Slemrod, J., Giertz, S., 2012. The elasticity of taxable income with respect to marginal tax rates: a critical review. Journal of Economic Literature 50, 3–50.

Smith, S.J., 1986. The growing diversity of work schedules. Monthly Labor Review 109, 7–13.

Weiss, Y., 1996. Synchronization of work schedules. International Economic Review 37, 157–179.

Labor Unions

Wolfgang Streeck, Max Planck Institute for the Study of Societies, Cologne, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 8204–8210, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editors.

Abstract

Labor unions are interest associations of workers in waged employment. The article summarizes their treatment in the social sciences and describes their rise in the industrial and political history of Western societies. It outlines the main dimensions of union organization and union growth and summarizes the literature on modern unions as collective actors in their three principal arenas: the labor market, the workplace, and the polity. The article concludes with a discussion of the prospects of unionism today.

Labor unions are interest associations of workers in waged employment. They are formed to improve the market situation and the life chances of their members, by representing them in the labor market, at the workplace, and in the polity, and in particular by collectively regulating their members' terms of employment. Unions emerged in the transition to industrial society in the nineteenth century, together with the defeudalization of work, the rise of free labor markets, and the commodification of labor. While employing modern means of formal organization, they represent an element of traditional collectivism in a market economy and society. Unions have taken a wide variety of forms and adopted different strategies in different historical periods, countries, and sectors. They are therefore favorite subjects of comparative social science.

Unions in the Social Sciences

The literature on unions is vast and extends into several disciplines. In the Marxist and socialist tradition, authors in the late nineteenth and early twentieth centuries debated whether and how unionism might contribute to the overthrow of capitalism or, to the contrary, its stabilization. Democratic reformers like Beatrice and Sidney Webb explored the potential contribution of unionism to the governance of a democratic industrial society. Institutional economists, such as Commons and Perlman in the United States or Brentano in Germany, tried to understand the economic causes and consequences of unionization. Neoclassical economists like Böhm-Bawerk explored the capacity of unions to override market prices and achieve a lasting redistribution of resources, while their successors in the 1960s applied theories of monopolistic competition to analyze the impact of collective bargaining on wages (Chamberlain). Early econometric analyses studied the impact of economic fluctuations on union organization and union activities, including strikes (Ashenfelter). The discipline of industrial relations, which grew after World War II in the US (Dunlop, 1993) and Britain (Flanders, Clegg), focused on unions as participants in collective bargaining under the 'New Deal' and in the postwar 'mixed economy.' Sociologists took an early interest in the transformation of unions into centralized and bureaucratic mass organizations with a staff of full-time officials (Cassau). Postwar American research on union democracy was in particular interested in the sources of communist influence in unions, while macro-sociological studies have investigated the origins of unionism in the context of social protest movements and their integration in the modern nation-state (Tilly). Political sociology and political science display a lasting interest in the role of unions in the polities of industrialized democracies. Economists study the economic effects of unions, in particular with respect to relative factor prices, inflation, productivity, and employment (Freeman and Medoff, Calmfors and Driffill, Soskice). Similar concerns figured prominently in historical institutionalist research and theory on neocorporatism in the political economy of advanced European welfare states (Schmitter, Crouch). Labor law has long been concerned with the rights of unions vis-à-vis employers, and with reconciling collective bargaining and collective industrial agreements with civil law and the freedom of trade and contract. Constitutional lawyers, in particular on the European continent, have tried to clarify the status of unions and collective bargaining in relation to the state and public legislation.

The Rise of Unionism in Industrial Countries

Unions emerged in conflict with both economic liberalism and political authoritarianism, striving simultaneously for economic regulation and political freedom. As a result, they were originally treated as conspiracies against free trade, the state, or both. But as unions established themselves as effective labor market cartels, they also became suppliers of labor to those employers willing to deal with them, and in this capacity slowly turned into 'managers of industrial discontent' (Flanders). Similarly, while unions represented the interests of a class opposed to the advance of capitalism, their toleration in the course of democratization contributed to the transformation of liberal into organized capitalism, and was central to the institutionalization of compromise between capital and labor. The integration of unionism and working class collectivism in the institutional framework of modern industrial societies turned unions into legitimate intermediary organizations between their members on the one hand and employers and the state on the other, and transformed class conflict into joint regulation of work and employment. It represents the foremost example of the stabilization of the capitalist economy through democratic politics and institution building.

Early unions saw themselves as democratic organizations of self-help and self-government of workers independent of the feudal or bourgeois predemocratic state. Often they belonged to broader labor movements that included political parties, consumer cooperatives, mutual assistance funds, educational associations, sports clubs, etc. While unions generally resented interference of the state and the law in their internal organization and activities, they differed widely in structure and ideology. Thus, syndicalist and anarcho-syndicalist unions, which in a number of countries remained significant well into the twentieth century, regarded themselves as constituent units of a direct democracy of producers set to replace both capitalist employers and the bureaucratic apparatus of the modern state. These traditions, which equally opposed capitalism and parliamentary democracy and favored direct action over both collective bargaining and political-electoral lobbying, culminated in the militant council movements of World War I and the revolutions following it, especially in the Soviet Union.

Integration of unions in democratic capitalism, and union recognition by governments and employers, was greatly advanced by the two World Wars. Economic mobilization and the governance of the war economy required the collaboration of union leaders, who in many countries came to be co-opted into positions of quasi-public authority. Also, enlisted soldiers had to be promised a better life in a fairer society upon their return from the battlefields, and in defeated countries traditional elites were replaced in the aftermath of war by liberal or socialist governments. Generally, the end of World War I brought political democratization and, precipitated by the threat of socialist revolution, widespread acceptance of collective bargaining. But the first postwar settlement proved fragile in countries like Germany, Japan, Italy, and Spain, where unionism was soon suppressed by authoritarian regimes and replaced with state-controlled mass organizations of workers that were put in charge of administering state social policy. Similarly, in the Soviet Union, workers councils were incorporated in a repressive state machinery and unions turned into 'transmission belts' from the state to the working class. In the United States, by comparison, the New Deal extended union organizing rights while the Swedish Social Democratic government of the 1930s and the British war cabinet of the 1940s began to develop the contours of the labor-inclusive Keynesian welfare state of the second postwar settlement after 1945.

The 'golden age' of capitalism in the second half of the twentieth century saw the worldwide ascendance of a 'mature' type of unionism, centralized at the level of the national state and pursuing economic and social policy goals through collective bargaining and political-electoral lobbying within the confines of capitalism and parliamentary democracy. This development was part of the consolidation of both democratic capitalism and the nation-state in the countries under American hegemony, where legal recognition of unions and free collective bargaining, extensive social welfare provision, a sizable public sector, and politically guaranteed full employment made possible the coexistence of liberal democracy and the market economy. The normalization of unionism under postwar 'embedded liberalism' (Ruggie) coincided with national regulation and standardization of the employment relationship and of the status of wage earners as distinguished from employers and the self-employed, which in turn corresponded to the advance of industrial mass production. Far-reaching legal and political regulation of the labor market, introduced to insulate employment and employment conditions as much as possible from economic fluctuations, stabilized union power. The worldwide outburst of labor militancy in the late 1960s, which took national union leaders by surprise, further added to union power, as governments, still feeling politically constrained to provide for full employment, more than ever depended on union cooperation for restoring economic and political stability.

Economic, political, and social change in the 1980s and 1990s caused a roll-back of unionism in most countries. In the 1980s, the British government proved that elections could be won even with high levels of unemployment. Under the leadership of the post-New Deal United States, governments also found less politically and economically costly ways to combat inflation compared to 'social contracts' with union leaders who had often turned out unable to deliver on their promises of wage moderation. Low inflation and high unemployment coincided with deep industrial restructuring, in response to economic internationalization and a new wave of technological change. Gradually this undermined the Fordist organization of work and the normalized employment relationship that unions had helped shape and with which they had learned to live comfortably. Also, the breakdown of Communism in the late 1980s eliminated the systemic opposition to capitalism that had made Western governments willing to make concessions to unions after the two World Wars. While private sector unionism has almost disappeared in the world's leading economy, the United States, it is, however, still firmly entrenched in many European countries, due to strong institutional supports left over from the social compact of the postwar era (Baccaro and Howell, 2011).

Union Organization and Union Growth

The first unions were local protest movements of the 'working poor.' In the mid-nineteenth century, they were superseded by formally organized craft unions of skilled workers which thrived particularly in the liberal political environment of early industrializing Britain and America. Commanding considerable market power, craft unions represented their members mainly through economic action. In addition to maintaining sizable mutual assistance funds, many of them were cartels that unilaterally fixed the price of their members' services. From the state, craft unions did not expect much apart from freedom of organization. While politically liberal, they were socially exclusive. Not only was membership in them reserved to the skilled. To control the labor supply, draft unions also limited apprenticeship training, prevented employers from hiring nonunion members for craft jobs ('pre-entry closed shop' for specific 'job territories') and defended an organization of work that ensured demand for their members' skills (job control unionism).

Where craft unions became firmly established, they survived beyond the second wave of unionization at the end of the century, when the growing masses of unskilled factory workers began to get organized. Given the weak market position of their

members, the 'general' unions that undertook to organize the large factories of the beginning age of mass production had to place their hope on the strength of large numbers. They therefore tried to recruit as many members as possible regardless of skill, although where craft unionism was strong, they mostly remained unions of the unskilled. While craft unions preferred a 'voluntarist' mode of operation keeping the state at arm's length, the new general unions emphasized political action and pressed for social legislation and state intervention in the economy.

In Anglo-America, where industrialization and democracy had come early, craft and general unions coexisted well into the twentieth century, with the former deeply influencing the practices and ideologies of the latter. By comparison, in late industrializing countries with authoritarian governments and narrowly limited political and economic opportunities for a market pursuit of collective interests, craft unions were mostly absorbed in inclusive 'industrial unions' organized by economic sectors. As these had originated in often-bitter political struggles especially for universal suffrage, their unionism was embedded in universalistic concepts of social and political reform that differed profoundly from craft particularism or American 'business unionism.' Frequently the new unions were allied to political parties, especially socialist and Catholic, and while craft unions had been fragmented by skill, industrial unionism was often fragmented by political orientation. Indeed as authoritarian governments had initially denied them the opportunity to represent their members economically, some of the new unions remained for long subordinate to political parties.

Craft unions, while narrowly based and normally small, were highly centralized where their members faced national labor markets. With the rise of the nation-state industrial unions, representing their members through political action for organizational rights, favorable labor laws, and social welfare legislation, also developed nationally centralized structures. In principle, these were usable for centralized collective bargaining as well, in solidaristic egalitarian pursuit of nationally standardized wages and employment conditions. However, getting members to support common demands in spite of often widely different situations at their workplaces required strong organizations. While political ideology helped, the capacity of national unions to aggregate the interests of a diverse membership depended in large part on whether employers were supportive of negotiation at national level. In countries with an established craft legacys, bargaining remained mostly decentralized and fragmented as industrial unions had to allow their members independent pursuit of sectional interests, on the model of and in competition with skilled unions. Sometimes, especially in the United States, unskilled unions managed to establish rights to 'job property' for their members as well and enforced on employers their own version of a closed shop, under which new hires had to join the union within a short period or lose their job (post-entry closed shop).

Union organizational development in the nineteenth and twentieth centuries included the rise of national federations of local and sectoral unions, which occurred in close correspondence with the formation of the modern nation-state and the establishment of liberal democracy. It also involved the growing independence of unions, not just from governments but also from friendly political parties, enabling even the industrial unions of the European continent to deal with democratically elected governments of conservative political complexion. Related to this was the increasing reliance of unions on collective bargaining with employers as their main mode of operation, with either individual firms or employers, associations. This presupposed the elimination of 'yellow' company unions controlled by employers as well as recognition of collective bargaining in labor law, for both of which government support was crucial. Collective bargaining, in turn, required that unions established organizational authority over strikes – the collective withdrawal of the labor power of workers from employers – just as political recognition of unions demanded that strikes were called only for economic purposes, abandoning the syndicalist tradition of the 'general' political strike. Union organizational development further involved integrating the growing number of white-collar workers, either into separate unions or inside encompassing industrial unions, which followed different trajectories in different countries and generally remained much less successful than the organization of blue-collar industrial workforces.

The literature offers a variety of explanations for the growth and decline of union membership over time and the differences in union density rates between sectors and countries. Unions tend to be strong in the core sectors of export-oriented manufacturing, like metal engineering, in large establishments, and in the public sector, while they are weak in private services and in small firms. But cross-country differences are considerable, and indeed on hardly any parameter of social structure do developed industrial societies differ as much. While in some Scandinavian countries around 80% of wage earners are union members, in France and the US density had fallen to about 10% by the end of the twentieth century. And while union membership relative to the workforce declined during the 1980s in the United Kingdom and the Netherlands, it increased in Belgium and Finland. Fluctuations in membership over time can also be enormous, as they were in Britain where in the 3 years between 1920 and 1923, union density declined from 45% to 30%.

Econometric studies have found that in the United States and Britain in particular, union density tended to increase with nominal wages and inflation, while it was negatively affected by unemployment and an already high level of unionization (Bain and Elsheikh). However, studies of this sort cannot explain the variation in unionization between countries, or ruptures in historical continuity. Institutionalist approaches take into account that union growth includes the acquisition by unions of recognized status, based on agreements with employers or the state helping unions recruit members ('union security'). Often such status is gained in exchange for political and economic moderation. Differences in union density then reflect different opportunities for unions to organize and keep members ('workers do not organize unions; unions organize workers'). In addition to the closed shop in its many official and unofficial versions – which makes membership compulsory by making access to employment dependent on it – employers may assist unions in collecting their membership subscriptions. Similarly, favorable legislation on union recognition or workplace representation through works councils may improve unions' ability to recruit. In some countries, like

Switzerland, nonmembers are legally obliged to pay union dues, and in Belgium and Denmark unions administer parts of the public social insurance system, enabling them to discriminate against nonmembers and thereby create additional incentives to join.

More generally, the organizing problems of unions may be described in terms of generic problems of collective action in the rational pursuit of group interests. Union achievements, like higher wages or favorable social policy legislation, are often collective goods in that they cannot be withheld from workers who have not contributed to their production. Apart from sentiments of fraternal solidarity, there is then little incentive for workers to join a union, even if they find the outcomes of union activities desirable. Indeed, given the risk that others will 'free-ride,' members may contribute in vain, as the union may not find enough support to be successful. Like other collective organizations, unions therefore typically offer potential members 'outside inducements' (Olson), i.e., benefits other than the collective goods for which they were originally founded, which unlike these can be limited to members (for example, health care benefits). They also strive to make membership less voluntary and obligatory, often with the help of the state or the employers.

Unions as Collective Actors

As intermediary organizations, modern unions represent their members in the labor market, the workplace, and the polity. Depending on their membership, their organizational structures and traditions, and the constraints and opportunities offered by employers and the state, different unions attach different significances to the three arenas and coordinate their activities in them differently.

Unions in the Labor Market

Unionism, and indeed much of modern labor law, assumes that individual workers are at a disadvantage vis-à-vis employers, and that redress of this imbalance requires that workers combine to face their employer as a collectivity. By regularizing wages and conditions of employment, sometimes also through state legislation, unions try to take wages and conditions out of competition, to make the market more predictable and protect workers' living conditions from its fluctuations. In this sense, Polanyi regards unionism as an element of the social countermovement against the 'satanic mill' of the 'self-regulating market' of early liberalism.

Economists have long debated whether unions can raise wages above the market level. Clearly, they can get nominal wage increases that exceed the increase in productivity, which depending on the government's monetary policy may result in either inflation or unemployment. With respect to the wage structure, in fragmented union systems skilled unions may use their independence to defend wage differentials, causing a wide wage spread between skill groups and industries. Fragmented ('competitive') bargaining may also enable well-organized groups of strategically placed unskilled workers to make high relative wage gains, which in turn may result in a 'leapfrogging' pattern of general, and potentially inflationary, wage increases. Wage differentials are also high in economies in which unions have little or no influence.

By comparison, industrial or general unions typically entertain an egalitarian ideology and strive to compress wage differentials, raising the wages of their less well-paid members more than those of the others. The result tends to be increased capital intensity and unemployment at the lower end of the labor market. Industrial unions therefore favor extensive training and retraining, funded by employers or the state, to increase the productivity of the less skilled to a level where they can earn the high wages negotiated for them. In favorable circumstances, this may lead to a general upgrading of a country or sector's production system, in the direction of more skill-intensive and less price-competitive production.

Since the 1970s, there has been a debate on the relationship between centralization of collective bargaining and various dimensions of macro-economic performance, such as inflation, growth, and employment. In the 1970s, centralized bargaining by national or industrial unions was believed by many to be more economically beneficial than fragmented bargaining, as it made it more difficult for bargainers to disregard the costs of redistributive bargaining or productivity-restraining work rules. Later it was suggested that bargaining agents in a decentralized wage setting system were more likely to be disciplined by market pressures, the problem being intermediate levels of centralization and unions, both large enough to override the market and small enough to be able to externalize the costs of their behavior to society. Statistical analyses have remained inconclusive and suggest that the economic effects of bargaining structures are highly conditional on economic and institutional circumstances.

Unions in the Workplace

Comparative research suggests that employers are willing to agree to centralized bargaining if this helps them keep unions out of the workplace. Workplace unionism, especially in a craft tradition, detracts from managerial prerogative and may interfere with management attempts to increase productivity. On the other hand, even industrial unions need to be present in the workplace, if only to recruit members, monitor the implementation of collective agreements, and generally insert themselves in the governance of the employment relationship and the 'wage–effort bargain' inherent in the labor process. Balancing central and workplace representation, or external and internal union organization, was and is one of the main problems of unionism and industrial relations.

Least of all, employers seem to like multiunionism in the workplace. This has often led to agreement with industrial unions on unified arrangements for plant-level workforce representation, like in Italy. In some countries, like the US, single-union representation is established by a legal procedure. In others, workplace unionism is merged into institutions of legally based participation, giving works councils elected by the workforce as a whole rights to information, consultation, and codecision-making while ensuring that management has a single counterpart with which to negotiate. While works councils in most countries cannot call strikes, they tend to have close relations with internal and external unions, not the least since most of their members are usually also union members.

In a number of Asian countries, unions have evolved into 'enterprise unions' that are only loosely federated above the enterprise level. This corresponds to an environment of closed internal labor markets and lifetime employment of core workforces. Enterprise unions tend to be highly cooperative with managerial efforts to increase productivity and competitiveness as they represent the close community of fate between their members and the enterprise that employs them. Often their officials are managers whose career is advanced by serving for a time in a union position. Works councils in European countries also tend to cooperate with management. Unlike enterprise unions, however, they exist next to an external union that can call strikes without regard to the competitive position of individual workplaces.

The economic effects of different forms of workplace unionism are still being debated. For some time, decentralized unionism and workplace bargaining were associated with high strike rates whereas centralized and politically influential unions, like those in Scandinavia, were found to be highly effective without much recourse to strikes. Especially where distributive wage bargaining is moved to the national or the industrial level, or where distributive conflict is suspended by jointly faced competitive market pressures, workplace representation was shown to enhance efficiency, among other things by enabling workers to voice complaints as an alternative to quitting. The same effect has been attributed to works councils. By comparison, in countries with adversarial industrial relation traditions, external labor markets, decentralized wage bargaining, freedom of workplace unions to call strikes, and a history of craft unionism and restrictive practices, workplace unionism came to be regarded in the 1960s and 1970s as a source of inefficiency. Here employers began to devise their own structures of workplace cooperation and 'worker involvement,' in the context of the developing practice of 'human resource management,' which often but not necessarily presupposes or aims at a 'union-free environment.'

Unions in the Polity

In most European countries, the rise of unionism was closely intertwined with nation-building and state formation. T.H. Marshall regards unionism as a step in the evolution from civil, to political, to social rights of citizenship, with unions using political rights to collective organization in the pursuit of social rights, albeit in the civil sphere of the marketplace and through freely negotiated – collective – contracts ('industrial citizenship'). As early repression of unions gave way to inclusion in shared public spaces (Crouch), unions became firmly integrated in the social and political order of the Western European nation-state, especially under the second postwar settlement after 1945.

The high point of union inclusion in advanced industrial countries was the 'neocorporatist' period of the 1970s and early 1980s. Especially in countries with a tradition of political unionism and centralized collective bargaining, unions were able to get major concessions from governments and employers in 'political exchange' for wage restraint, which in turn was to enable governments to protect full employment with Keynesian means without having to accept excessive rates of inflation ('incomes policies'). Among the concessions gained by unions in this period were legislation on union recognition and access to the workplace; various other forms of 'union security'; legal extension of collective agreements to all firms in an industry; higher social security and public pension benefits; and industrial democracy legislation. Neocorporatism was more likely to work where governments were of a social democratic complexion; but most conservative governments were also eager not to antagonize the unions. The more politically powerful unions became under corporatist arrangements, the lower as a rule was the strike rate as strong unions were able to get in the political arena what others had to struggle for, with uncertain event, in the industrial arena.

The historically close involvement of union movements with their respective nation-states explains to a large extent why unions and collective bargaining have so far resisted economic pressures for institutional convergence. In fact, the structure of unionism and its status inside its institutional context are an important dimension distinguishing between the different 'models of capitalism' that have attracted attention in the 1990s. The lasting symbiosis between union movements and national states also accounts in large parts for the weakness of international unionism. Such weakness is found even in the European Union, where in spite of deep economic integration national unions have remained highly distinctive in their organizational structures and *modus operandi*, and are jealously guarding their independence.

The Prospects of Unionism Today

Much of the literature on unions in the early 1960s observed and predicted a decline of unionism in leading industrial societies. In most Western countries, with the notable exception of the United States where union decline continued unabated, this was contradicted by the 'hot autumns' of 1968 and 1969, which were followed by a sharp reversal of membership losses and by unprecedented influence of unions in politics. In the 1990s, however, union membership and power were again declining worldwide. With the American private economy on the way to effective deunionization, the British government in the 1980s attacked unions' legal and institutional supports, with considerable success (Davis, 2009). Outside the Anglo-American world, union influence was weakened by the end of politically guaranteed full employment combined with intensified international competition, while the growing importance of the service sector and the 'knowledge economy' and the rise of a more diverse and better educated workforce made it more difficult for unions to recruit new members. Even where institutional conditions remained unchanged, unionism tended to be increasingly confined to the traditional sectors of the economy, which have long been shrinking in size and importance (Streeck, 2005).

In their history unions have become deeply involved in the functional and social integration of the modern nation-state: in the national standardization of employment practices, the organization of redistributive solidarity at national level, and the political governance of the national economy. The decay

of the second postwar settlement coincides with strong market pressures for more flexible modes of governance, less political interference, and internationalization of economic transactions. It is not clear whether and how unions, and indeed the nation-state itself, will find a response to the denationalization of the global economy. With their membership aging, unions may become defenders of an industrial employment and social security regime that no longer fits the needs of the rising postindustrial sector of the workforce. Being unattractive to the winners of structural change in the highly skilled and female segments of the labor force, unions also seem less than successful in organizing the new underclass of low-paid service workers, many of them immigrants, that has been growing for some time in many industrial societies.

See also: Class Consciousness; Class: Social; Conflict: Organizational; Division of Labor: History of the Concept; Industrial Relations, History of; Labor History; Labor Market Policy; Labor Supply; Labor, Division of; Work, History of; Work, Sociology of.

Bibliography

Bain, G., Price, R., 1980. Profiles of Union Growth: A Comparative Statistical Portrait of Eight Countries. Blackwell, Oxford, UK.

Baccaro, L., Howell, C., 2011. A common neoliberal trajectory: the transformation of industrial relations in advanced capitalism. Politics & Society 39 (4), 521–563.

Crouch, C., 1993. Industrial Relations and European State Traditions. Clarendon Press, Oxford, UK.

Crouch, C., Pizzorno, A., 1978. The Resurgence of Class Conflict in Europe. Macmillan, London.

Davis, G.F., 2009. Managed by the Markets: How Finance Re-Shaped America. Oxford University Press, Oxford.

Dunlop, J.T., 1993. Industrial Relations Systems, rev. ed. Harvard Business School Press, Boston, MA.

Ebbinghaus, B., Visser, J., 2000. Trade Unions in Western Europe since 1945. Macmillan, London.

Lipset, S.M. (Ed.), 1986. Unions in Transition: Entering the Second Century. Institute for Contemporary Studies, San Francisco.

Marks, G., 1989. Unions in Politics: Britain, Germany and the United States in the Nineteenth and Early Twentieth Centuries. Princeton University Press, Princeton, NJ.

Rogers, J., Streeck, W. (Eds.), 1995. Works Councils: Consultation, Representation, and Cooperation in Industrial Relations. The University of Chicago Press, Chicago.

Streeck, W., 2005. The sociology of labor markets and trade unions. In: Smelser, N.J., Swedberg, R. (Eds.), The Handbook of Economic Sociology, second ed. Princeton University Press, Princeton.

Laboratory Studies: Historical Perspectives

Park Doing, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by K.K. Cetina, volume 12, pp. 8222–8228, © 2001, Elsevier Ltd.

Abstract

The main purpose of laboratory studies has long been taken as epistemological. Seen as the *coup-de-grace* for constructivist philosophy of science, early practitioners of laboratory studies were seen to have disproved demarcationist theories of scientific fact production. They were supposed to have done so by revealing the messy, intertwined, socio-technical world inside the laboratory itself. Because laboratory studies were seen in this way, they initially garnered sensational attention, but later were seen to be less important since the epistemological project was considered either to have been completed or to be a methodological non sequitur. This view of laboratory studies has obscured their actual importance and their continuing contributions to studies of science and technology. If laboratory studies are seen not as epistemological projects, but rather as investigations into technoscientific agency, their real and sustained importance to science and technology studies (STS) emerges. In this view, it can be seen that early laboratory studies set the stage for, and subsequent studies fleshed out and further challenged, debates regarding material agency, distributed organizational agency, and cultural agency in technoscience; debates that animate the field of STS to this day.

The idea of exploring scientific laboratories in real time as a means of understanding the nature of science began in the late 1970s with several ethnographies of laboratories in the USA. Since then, laboratory studies have played a founding role in science and technology studies (STS) and influenced a spectrum of established fields in the humanities from history to literary criticism. First, laboratory studies were a driving force in the rise of the Sociology of Scientific Knowledge (SSK), a movement that asserted on philosophical grounds that scientific knowledge was sociological. Laboratory studies were enrolled in, and sometimes enthusiastically enrolled themselves in, proving the manufactured nature of scientific fact-making at the 'actual site of scientific action' – the laboratory floor (Knorr Cetina, 1983: 117). This movement garnered a lot of attention, disrupting relationships between sociology, anthropology, history, and philosophy of science, instigating new approaches in the humanities, and challenging the relationship between the sociology of science and science itself – culminating in the 'science wars' of the 1990s. Since that time, however, after all the sensational sturm und drang, the epistemological project in laboratory studies has quietly died down, and, depending on the author, the goal of proving that scientific knowledge is sociological is seen as either complete, failed, or a methodological non sequitur. As a result, the kind of programmatic attention afforded to the initial laboratory studies has not been as prominent in subsequent studies of laboratory science. Regardless of this lack of wider attention, however, laboratory studies continue to make up a rich field of inquiry that has built in important ways upon the instigations effected by the initial studies. In order to see this importance, however, one must not look at laboratory studies as epistemological projects, as they were initially and sensationally held to be. Rather, they should be viewed as investigations into technoscientific agency. Indeed, if one casts such an eye back over the history of laboratory studies, it can be seen that their real importance was missed the first time around, and their actual contributions have run beneath the radar ever since. If laboratory studies are looked at through a lens of agency, rather than considered for their self-proclaimed or attributed contributions to epistemology, they can be seen as important precursors and developers of influential concepts regarding human and technological actions, which have been taken up in the broader field of STS. Far from being micro-sociological studies, the early laboratory studies developed sophisticated conceptions of agency that spurred conceptual growth in studies of science and technology. Laboratory studies that followed the initial studies built upon these conceptions in ways that continue to challenge science studies to this day. If one can get over programmatic preoccupations with epistemology, one can see that agency is where the action is in laboratory studies – and where it always has been.

Machines, Conventions, Culture: Agency in the 'First Author' Laboratory Studies

The epistemological project in laboratory studies can be traced to two initial projects, Bruno Latour and Steve Woolgar's *Laboratory Life: The Social Construction of Scientific Facts* and Karin Knorr Cetina's *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science* (Latour and Woolgar, 1979; Knorr Cetina, 1981). The authors of these works enthusiastically enrolled themselves in the project of reworking understandings of the nature of science. In both of them, the stated goal of the ethnographic incursions behind the walls of laboratory science was to problematize demarcationist theories of science (efforts to draw a clear, logical separation between objective scientific knowledge and nonscientific beliefs) by unveiling the sociality at the heart of knowledge production. This project fit nicely with conceptualizations put forth by SSK scholars who had engaged with Wittgenstein's philosophical provocations and certain interpretations of Kuhn (Bloor, 1975; Pinch, 1982). Indeed, these first ethnographies of laboratory practice were considered the *coup-de-grace* for the

constructivist triumph, displaying what the philosopher-sociologists had been saying all along – that science was social to its core.

For their influential *Laboratory Life*, Latour and Woolgar entered The Salk Institute in San Diego on a mission. They held that their self-imposed naiveté about the actual nature of science would hold them in good methodological stead as they observed how facts were made by not biasing them toward the scientists' own accounts of their work. Their ethnographic conclusions, undermining the scientists' view that they were observing nature in the laboratory, were taken up readily, although often contentiously, in fields ranging from sociology and anthropology, to history, literature, and critical theory. Interestingly, while the terminus of the deconstructionist project was held up as the valorous result of *Lab Life* at the time, Latour and Woolgar's approach to agency can be seen to have set in motion a more important, longer lasting, and better sustained research arc in the STS field as a whole. While pursuing their stated interest in showing how facts are constructed, and while specifically putting aside some possible modes of agency, including the "administrative organization of laboratory work," the "influence of such organization on creativity," the "influence of laboratory organization on scientists' careers," and the "nature of communication and pattern of information flow" (Latour and Woolgar, 1979: 40), Latour and Woolgar focused in their early study on a mode of agency that would form the basis of the later formulation of the actor-network theory of Callon and Latour that would become a driving force in studies of science and technology – material agency. This agency is expressed through their notion of the 'inscription devices,' which Latour and Woolgar see at work throughout the laboratory. An inscription device is "any item of apparatus or particular configuration of such items which can transform a material substance into a figure or diagram which is directly usable" by scientists (Latour and Woolgar, 1979: 52). The crux of their conception of material agency is that:

> An important consequence of this notion of inscription device is that inscriptions are regarded (by scientists) as having a direct relationship to "the original substance." The final diagram or curve thus provides the focus of discussion about properties of the substance. The intervening *material activity* and all aspects of what is often a prolonged and costly process are bracketed off in discussions about what the figure means.
>
> Latour and Woolgar, 1979: 52, emphasis added

The most sustainably influential part of *Laboratory Life* was not its epistemological assertions about the construction of facts, but rather the conception that, while the scientists and technicians were busy writing down notes and figures in reference to other published and nonpublished notes and figures, the machines were busy writing too. Even if one does not take this 'material activity' as impugning the status of scientific facts (as many do not – including Latour himself later in his career), this germination of the idea of material agency, that technology itself has a hand in the shaping of human affairs and understandings of the nature of machines, would go on to enliven laboratory studies and studies of science and technology for decades to come, and also would lead to the removal of the word 'social' in the second edition of *Laboratory Life* (Latour and Woolgar, 1986).

In *The Manufacture of Knowledge*, Karin Knorr-Cetina also asserts explicitly that her project upended the demarcationist conception of scientific knowledge. Her view, as laid out in her programmatic essay 'Ethnographic Study of Scientific Work,' is that:

> The problem-solving model suggests that science is a successful tool for adequately coping with the world, even though we may not want to think of it as correctly representing the world. The vision behind this model is that of a decreasing stock of problems constituted by what we do not yet know about how to handle the world-out-there. The vision behind the constructivist programme as I conceive of it is that of a potentially increasing stock of problems created by science in the process of secreting an unending stream of entities and relations that make up 'the world.' The 'unknown world' as an intentional object of science is itself a function of a constantly changing scientific practice, of what at every moment of scientific work emerges as the known world. But this known world is a cultural object, a world identified and embodied in our language and our practices.
>
> Knorr Cetina, 1983: 135

For her part, Knorr Cetina points to two modes of agency that provide for the construction of knowledge. The first is similar to the material agency to which Latour and Woolgar point. For Knorr Cetina, technology, in the form of instrumentation that embodies and asserts previously made decisions, produces artifacts in the laboratory, not direct observations of nature. As with Latour and Woolgar, "instrumentally accomplished observations intercept natural courses of events" (Knorr Cetina, 1983: 119). This 'interception' is the manifestation of the material agency of the instrument(s). The scientific laboratory, then, "consists of materializations of earlier scientific selections; this accounts for its artifactual character" (Knorr Cetina, 1983: 121). For Knorr Cetina, however, this put-in-place technology is not as autonomous as it was for Latour and Woolgar. The instrumentation is itself dependent on previous human decisions. Elucidating this mode of agency was one of Knorr Cetina's important contributions to laboratory studies, and as with Latour, Knorr Cetina re-explored and elaborated on this core conception in future work (revising it more toward the autonomous agent that Latour and Woolgar proposed), which became very influential in studies of science and technology at large. This early divide between the laboratory studies of Latour and Woolgar and Knorr Cetina can be seen to presage debates between actor-network theorists, social constructivists of technology (SCOT), and non-SCOT for decades to come in STS.

In *Manufacture of Knowledge*, Knorr Cetina (1981) also pursues a second mode of agency specifically left unexplored by Latour and Woolgar, which she refers to as the 'trans-epistemic field.' Not satisfied with the loose and ubiquitous concept of scientific community, Knorr Cetina asserts that the larger and wider, supposedly external, organizational ties that, for example, have to do with funding and the administration of publications are not neatly cordoned off from the internal intellectual ties that bind a community of scientific

practitioners. In an early review of laboratory studies, Knorr Cetina notes that:

> Scientists, even specialty colleagues, may confront each other in their everyday exchanges in 'non-scientific' roles in which they administer money or decide on professional careers. Similarly, a government official or sales agent may negotiate with the scientific specialist the methods to be used in a research project, or the proper interpretations of a measurement result.
>
> Knorr Cetina, 1983: 132

For Knorr Cetina, the crucial question is in what sense these 'trans-epistemic connections' are relevant to 'the process of inquiry' in science. Her answer is that such connectivity is the second agent of fact construction. "These relationships emerge as the locus in which the decision criteria corresponding to laboratory selections … are defined and negotiated" (Knorr Cetina, 1983: 132). As with material agency, future laboratory studies would utilize and build upon Knorr Cetina's initial exploration of such distributed organizational agency, as she herself also would in her own studies. Regardless of how her assertions about the epistemic implications of her observations were pushed by her and taken up by others, this exploration into material agency and distributed organizational agency proved to be of longer-lasting influence in studies of science and technology, as both actor-network theorists and social constructivists of technology (now SCOTS) wrestled with incorporating systems of organization into their respective frameworks.

While Latour, Woolgar, and Knorr Cetina actively engaged themselves in the epistemological project of laboratory studies, Michael Lynch and Sharon Traweek made no such epistemological proclamations, although they were variously enrolled in the project by others (Lynch, 1985; Traweek, 1988). Using an ethnomethodological approach, which itself does not hold that scientific knowledge is social in the same sense as Latour and Woolgar, and Knorr Cetina proclaimed, Lynch focused on the contingent decision making of scientists on the 'shop floor' of the laboratory. This decision making was done by individuals in interaction with other individuals and investigative materials, and each successfully negotiated decision became the basis for further decision making, thus locking in the conversationally negotiated contingency. These conversational negotiations were not random, however, and here Lynch brings in a form of agency different from that of Latour, Woolgar, and Knorr Cetina. Yes, individual scientists are in the foreground of Lynch's analysis, but conversation is used by individuals to assert their positions, or will, on the shop floor of science, and in order to do so certain conventions of language and conversation must be adhered to. Here, the structure of the conversational conventions themselves, however loose, becomes an agent in the decision-making chain of laboratory practice. Although such conversational negotiations could lead to a wide variety of possible outcomes, owing to the momentary and contingent interactions, such interactions could not simply be random, or infinitely malleable. Rules of language were governing forces in interactions that built up to scientific claims. In this work, widely considered the epitome of micro-sociological studies, structural agency was in fact present, and would be taken up and pursued by later laboratory studies researchers.

Sharon Traweek states at the outset of her book *Beamtimes and Lifetimes* that exploring epistemology is not the purpose of her study, noting that her book is about how physicists "have forged a research community for themselves, how they turn novices into physicists, and how their community works to produce knowledge" (Traweek, 1988: 1). In pursuit of her project, Traweek brings out several different kinds of agency that are different from those featured in the other early laboratory studies, in that they are rooted in broader conceptions of culture. Traweek does see a role for a kind of material agency at the laboratory, albeit a different one from Latour, Woolgar, and Knorr-Cetina. For Traweek, technologies in the laboratory transmit not their own essence or logic, or reified versions of previous human decision making, instead they are cultural forces. The cultural agencies that she describes throughout her book can be, she notes, are asserted through the scientific instrumentation of particle physics detectors. For Traweek.

> (d)etectors express diverse styles of research, different strategies for success. Above all, they embody rules for separating noise, which arises from imperfections in the experimental process, from data, which carry authentic messages from nature; nevertheless, detectors ideally become invisible, transparent scientific instruments for reading news from nature. I see detectors as the material embodiment of the high energy physics culture.
>
> Traweek, 1988: 158

For Traweek, the cultural aspects of agency include gender, national culture, the culture of high-energy physics professionalization, and cultural conceptions of time itself. Her ethnographic approach to gender, working from the inside of the scientific laboratory outward and including her cross-national comparison of US and Japanese laboratories, was an innovation that spurred much important work in STS. Her inclusion of organizational practices, both national and professional, as an active part of conceptions of action in the laboratory among scientific practitioners is in line with Knorr Cetina's push to consider trans-epistemic fields, and was an important impetus for such considerations in subsequent laboratory studies. Her invocation of time as a conceptual driving force for scientific action was unique and creative. Overall, Traweek brought into play a cultural conception of agency, where individual practitioners negotiated cultural forces; forces that could also be transferred through the detectors themselves. These larger forces could be manipulated to a certain extent, as when certain stereotypes were broken successfully in the laboratory. But more so than the other early authors of laboratory studies, Traweek pointed to a prime-moving, nonmaterial agency at the laboratory that came from an inaccessible, and to a large extent immutable, cultural realm. As with the initial incursions into material agency and distributed organizational action, the role of cultural forces would also occupy actor-network theorists and SCOT(S) and non-SCOT(S) adherents as they endeavored to account for such obviously powerful, yet seemingly ineffable, driving forces.

Developing Agency: The Continuing Build-Up of Laboratory Studies

Following on the initial laboratory studies, subsequent researchers have built up, further refined, and further challenged conceptions of technoscientific agency. Karin Knorr Cetina went on to expand her conception of material agency, including both Latour and Woolgar's conception of independent material action and also Traweek's sensibility with regard to culture(s). Knorr Cetina codified a concept of 'epistemic cultures,' whereby different disciplines in science, by virtue of their differing instrumental configurations, produce distinctive "ontolog[ies] of objects and machines," as well as larger "'epistementalities' that prevail in the larger society – beliefs and attitudes about the correct distribution of knowledge, the rights of access to it, the requirement and meaning of transparency, etc." (Knorr Cetina, 2001: 8235). Thus molecular biology with its dispersed bench work practices and wet laboratories is culturally, epistemologically, and ontologically different from particle physics, with its large, centralized accelerators and detectors and the unified actions of hordes of scientists. Here, the machines themselves and cultural forces combine to distinguish the laboratory work of different scientific fields. Wrestling with Knorr Cetina's concept of trans-epistemic fields, but focusing on dynamics of change, Park Doing has proposed the concept of 'epistemic politics' to connect laboratory members' disciplinary negotiations (negotiations over laboratory resources, instrumentation, operations, and agreements on the proper nature experimentation) to modes of organizational reporting and strategies for sustained laboratory support, including funding, academic posts, public awareness and support, economic leverage, and control of peer review in scientific publication (Doing, 2009).

Considering trans-epistemic fields in a different way, Martina Merz has proposed the notion of 'technological platform' – like the concept of 'nanotechnology' itself – as a dispersed imperative that laboratories and organizations draw upon to bring together funding sources and intellectual resources (Merz, 2010). Continuing to consider organization in laboratory practice, Vincent Simoulin (2007) has also proposed redefining the notion of scientific community based on his study of interactions of 'research collectives' revolving around a synchrotron, and Hackett et al. (2004) have referred to the 'research assemblages' interacting at a Tokomak fusion laboratory. Morgan Jouvenet (2007) has been concerned with both distributed organization and professional culture in nanotechnology laboratories. Benjamin Sims and Christopher Henke have connected the larger politics of nuclear weapons stockpiling to the operation of, and assessments of the validity of, nuclear fusion laboratories (Sims and Henke, 2012). Sims also has brought in the driving force of 'safety' as a cultural imperative that determines conceptions of laboratory instrumentation and operations (Sims, 1999). Pushing into further explorations of larger cultural concepts, Traweek has expanded upon her previous notions of agency, building on her initial study as well as others by elaborating on the relationships between instrumentation, organization, culture, and individual self-conceptions that drive action in scientific laboratories. Traweek has opened her notions of agency to include conceptions of colonial culture, thus relating 'big science' laboratories to colonial centers of power analyzable with the tools of postcolonial cultural theory and initiating similar movements in studies of science and technology in general (Traweek, 1996, 2004). Most recently, following the impetus from feminist studies of science and critical theory, a movement is beginning to emerge that calls for an approach to studying laboratories that is tuned into and driven by the normative notion of political and cultural 'matters of care' (Puig de la Bellacasa, 2011).

Further wrestling with material agency, Matthew Hubert (2007) has explored relationships between identity and instrumentation in the hybridization of nanotechnology laboratories. Dominique Vinck and Gloria Zarama (2007) have wrestled with similar considerations during the merging of nanotechnology and microtechnology laboratories. Knorr Cetina and Merz have posed a particular challenge to considerations of material agency by following the work of theoretical physicists (Knorr Cetina and Merz, 1997; Merz and Knorr Cetina, 1997). What kind of agency is at play when no material at all is involved in the scientific work? Expanding conceptions of material agency from the instrumental, Thomas Gieryn has called for exploring the role of space in laboratories (Gieryn, 2000). How does the material agency not just of instrumentation and technological infrastructure, but also of the architecture of laboratories determine actions and conceptions of work inside of them? Thinking the other way, studies have also begun to wrestle with the role of the virtual laboratory in a world of increasingly linked and real-time communications (Sims et al., 2013). What is the implication for considering the laboratory as a place of no 'space' at all, where "the electronic space becomes a *workbench* for participants?" (Knorr Cetina, 2001: 8237). And, picking up on the early strand of conversational agency, researchers have explored the role of skepticism in real time laboratory practice and also pushed again on conversational modes of convergence in laboratory decision making (Owen-Smith, 2001; Sormani, 2014).

Carrying forth Traweek's initial impetus to consider the different 'kinds' of physicists in the course of a career, laboratory studies researchers have also sought to bring the ubiquitous presence of technicians in laboratories into considerations of technoscientific agency (Barley and Bechky, 1994; Barley, 1996; Lynch and Jordan, 1995; Orr and Barley, 1997; Doing, 2004; Bechky, 2006; Wylie, 2009). A driving question for these studies is whether technicians are fundamentally different agents than scientists themselves. Considering this question has challenged some long-held notions of both skill and expertise in STS as well as considerations of material agency. For SSK writers, tacit skill is the cornerstone of theories of the 'enculturation' of scientific knowledge (Collins, 1985). Uncodefiable tacit ability is the means by which the social is 'injected' into technical knowledge production by scientists, defining what it means to be a technical expert. Recent laboratory studies, however, have challenged that notion by noting how 'skill' is used as a political category variously asserted and performed by scientists and technicians as a means of cordoning off workplace authority and control (Doing, 2004) and managing disparate groups of laboratory interests (Wylie, 2013), or as an accounting device to rationalize practical action (Lynch and Jordan, 1995). If

tacit knowledge is not the epistemological cornerstone of enculturation, then on what basis is technical expertise sociological? A further challenge stemming from considering technicians in laboratories involves exploring material agency. What does it mean for material agency if scientists, and sometimes technicians themselves, see technicians as inscription devices? If a human is carrying out the material inscription, does that mean that a material agent need not be material after all, that some other, more specified, quality of action is what determines the 'materiality' of technological force?

In all of these investigations, agency is never a one-way street, or even a two-way street. Rather, these empirical investigations into technoscientific action in the laboratory are more in line with dialectical considerations of agency from social theory, such as Anthony Giddens' 'structuration' and Pierre Bourdieu's conception of the relationship between 'habitus' and 'field.' Actions are taken by motivated agents who variously conceive of themselves as able to act or as constrained by various forces 'large' and 'small,' 'human' and 'technological.' Any account of what actually drives the action has to wrestle with this kind of recursion. Since laboratory studies practitioners are so closely confronted with their subjects' accounts of themselves, this kind of recursion is necessarily at the fore of accounts of agency emerging from this kind of research. Of course, there is no 'grand unified theory' of agency – as continuing theoretical debates in social theory and STS make evident. But just how individual action, material agency, distributed organizational agency, and cultural forces manifest in recursive dialectic relation in the course of technoscientific practice at the laboratory has been the core project of laboratory studies from the beginning, even if was missed at the time, with many and new results arising from continuing investigations. As Vinck summarizes the supple kind of agency that drives the field:

> A laboratory is … not just an organisation that coordinates individuals or groups with recognised skills working on an identified socio-technical production, but a system in which objects, instruments and skills are re-opened, a system that is assessed through its ability to further exploration and extend questioning. It (is) especially interesting to examine the way these laboratories reconfigure the 'resources' they implement (money, groups of researchers, instruments, and partner relations in particular) in order to integrate these in new registers, and knowledge projects. These laboratories merge and redirect pre-existing socio-technical entities; they disconnect and transform these entities, settling them finally in a new universe of phenomena and relations.
>
> Vinck, 2007: 163

With such explorations into the relations between individual action, material agency, distributed organizational agency, and cultural 'forces,' laboratory studies continue to challenge theoretical debates between actor-network theorists, SCOT(S) adherents, non-SCOT(S) practitioners, cultural studies of science and technology, and newer proponents of the 'mangle' of practice (Pickering, 1995), all of which make up grist for the mill of STS research.

Conclusion

Key questions regarding technoscientific agency have been driving forces for, and still reverberate in, STS today. What is the nature and role of material agency? How might distributed organizations or systems constrain or affect technoscientific action? What role is there for larger cultural 'forces'? How might all of these agential impetuses conflict, relate, or recursively interact and intercede with one another? From the beginning, laboratory studies have put forth and wrestled with these kinds of questions. Since their main contribution was seen to have been with regard to the epistemology of scientific facts, rather than making sense of the steering of technoscientific realms, the importance of laboratory studies investigations to these kinds of questions was missed at the outset, and then largely ignored. Even so, laboratory studies practitioners first set the stage for theoretical debates about agency in STS, and then quietly and empirically dug into sophisticated conceptions of action that continue to be at the forefront of theoretical currents in the field to this day. The laboratory is, and has always been, where the technoscientific action is. Consequently, laboratory studies are, and continue to be, valuable resources for STS practitioners.

See also: Cultural Studies of Science; Ethnography in Applied Social Research; Knowledge, Sociology of; Nanotechnology: Social and Cultural Aspects; Science and Technology Studies, Ethnomethodology of; Science and Technology, Anthropology of; Science, Sociology of; Scientific Knowledge, Sociology of; Technology, Social Construction of; Visualization in Science and Technology.

Bibliography

Barley, S.R., 1996. Technicians in the workplace: ethnographic evidence for bringing work into organizational studies. Administrative Science Quarterly 41 (3), 404–441.

Barley, S.R., Bechky, B., 1994. In the backrooms of science: the work of technicians in science labs. Work and Occupations 21 (1), 85–126.

Bechky, B., 2006. Talking about machines, thick description, and knowledge work. Organization Studies 27, 1757–1768.

Bloor, D., 1975. The structure of scientific inference. Studies in History and Philosophy of Science, Part A 5 (4), 382–395.

Collins, H.M., 1985. Changing Order: Replication and Induction in Scientific Practice. Sage, Beverley Hills and London.

Doing, P., 2004. 'Lab hands' and the 'Scarlet O': epistemic politics and (scientific) labor. Social Studies of Science 34 (3), 299–323.

Doing, P., 2009. Velvet Revolution at the Synchrotron. MIT Press, Cambridge, MA.

Gieryn, T.F., 2000. A space for place in sociology. Annual Review of Sociology 26, 463–496.

Hackett, E.J., Conz, D., Parker, J., Bashford, J., DeLay, S., 2004. Tokamaks and turbulence: research ensembles, policy and technoscientific work. Research Policy 33 (5), 747–767.

Hubert, M., 2007. Hybridations instrumentales et identitaires dans la recherche sur les nanotechnologies. Revue d'Anthropolie des Connaissances 1 (2), 243–266.

Jouvenet, M., 2007. La culture du "bricolage" instrumental et l'organisation du travail scientifique enquête dans un centre de recherche en nanosciences. Revue d'Anthropologie des Connaissances 1 (2), 189–219.

Knorr Cetina, K., 1981. The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science. Pergamon Press, Oxford.

Knorr Cetina, K., 1983. The ethnographic study of scientific work: towards a constructivist interpretation of science. In: Knorr Cetina, K., Mulkay, M. (Eds.), Science Observed: Perspectives on the Social Study of Science. Sage, London, pp. 115–140.

Knorr Cetina, K., 1999. Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press, Cambridge, MA.

Knorr Cetina, K., 2001. Laboratory Studies. Historical Perspectives. International Encyclopedia of the Social & Behavioral Sciences. Elsevier, Amsterdam.

Knorr Cetina, K., Merz, M., 1997. Floundering or frolicking – how does ethnography fare in theoretical physics? (And what sort of ethnography?). Social Studies of Science 27 (1), 123–131.

Latour, B., Woolgar, S., 1979. Laboratory Life: The Social Construction of Scientific Facts. Sage, London.

Latour, B., Woolgar, S., 1986. Laboratory Life: The Construction of Scientific Facts, second ed. Princeton University Press, Princeton.

Lynch, M., 1985. Art and Artifact in Laboratory Science: A Study of Shop Work and Shop Talk in a Research Laboratory. Routledge and Kegan Paul, London.

Lynch, M., Jordan, K., 1995. Instructed actions in, of and as molecular biology. Human Studies 18 (2–3), 227–244.

Merz, M., 2010. Reinventing a laboratory: nanotechnology as a resource for organizational change. In: Kaiser, M., Kurath, M., Massen, S., Rehmann-Sutter, C. (Eds.), Governing Future Technologies: Nanotechnology and the Rise of an Assessment Regime, Sociology of the Sciences Yearbook, vol. 27. Springer, Dordrecht, pp. 3–19.

Merz, M., Knorr Cetina, K., 1997. Deconstruction in a "thinking" science: theoretical physicists at work. Social Studies of Science 27 (1), 73–111.

Orr, J.E., Barley, S.R. (Eds.), 1997. Between Craft and Science: Technical Work in U.S. Settings. IRL Press, Ithaca, NY.

Owen-Smith, J., 2001. Managing laboratory work through skepticism: processes of evaluation and control. American Sociological Review 66 (3), 427–452.

Pickering, A., 1995. The mangle of practice: time, agency, and science. University of Chicago Press, Chicago.

Pinch, T., 1982. Kuhn – the conservative and radical interpretations: are some Mertonians 'Kuhnians' and some Kuhnians 'Mertonians'? Social Studies of Science 27 (3), 465–482.

Puig de la Bellacasa, M., 2011. Matters of care in technoscience: assembling neglected things. Social Studies of Science 41 (1), 85–106.

Simoulin, V., 2007. Une communauté instrumentale divisée… et réunie par son instrument. Revue d'Anthropologie des Connaissances 1 (2), 221–241.

Sims, B., 1999. Concrete practices: testing in an earthquake-engineering laboratory. Social Studies of Science 29 (4), 483–518.

Sims, B., Henke, C.R., 2012. Repairing credibility: repositioning nuclear weapons knowledge after the cold war. Social Studies of Science 42 (3), 324–347.

Sims, B., Sinitsyn, N., Eidenbenz, S., 2013. Visualization and modeling of structural features of a large organizational email network. In: Proceedings of the 2013 IEEE/ASM International Conference on Advances in Social Networks Analysis and Mining, pp. 787–791.

Sormani, P., 2014. Respecifying Lab Ethnography: An Ethnomethodological Study of Experimental Physics. Ashgate, Surrey, UK.

Traweek, S., 1988. Beamtimes and Lifetimes: The World of High Energy Particle Physics. Harvard University Press, Cambridge, MA.

Traweek, S., 1996. Kokusaika, Gaiatsu, and Bachigai: Japanese physicists' strategies for moving into the international political economy of science. In: Nader, L. (Ed.), Naked Science: Anthropological Inquiry into Boundaries, Power, and Knowledge. Routledge, Boston, pp. 174–197.

Traweek, S., 2004. Keizu to Nendaiki: making history in Tsukuba science city. In: Matas, J.A. (Ed.), Sociology of Science in the New Millennium. Harvard University Press, Cambridge, MA.

Vinck, D., 2007. Back to the laboratory as a knowledge production space. Revue d'Anthropologie des Connaissances 1 (2), 160–166.

Vinck, D., Zarama, G., 2007. La fusion de laboratoires. Revue d'Anthropologie des Connaissances 1 (2), 267–296.

Wylie, C.D., 2009. Preparation in action: paleontological skill and the role of the fossil preparatory. In: Brown, M., Kane, J., Parker, W. (Eds.), Methods in Fossil Preparation: Proceedings of the First Annual Fossil Preparation and Collections Symposium, pp. 3–12.

Wylie, C.D., 2013. Invisible Technicians: A Sociology of Scientific Work, Workers, and Specimens in Paleontology Laboratories. (Ph.D. thesis). Clare Hall, Department of History and Philosophy of Science, University of Cambridge.

Lacan, Jacques (1901–81)

August Ruhs, Medical University of Vienna, Wien, Austria
Jeanne Wolff Bernstein, Psychoanalytic Institute of Northern California, San Francisco, CA, USA; Postdoctoral Psychoanalytic and Psychotherapy Program, New York University, New York, NY, USA; and Wien, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Jacques Lacan was one of the most influential psychoanalysts in France. Lacan reinterprets Freud's early texts under a linguistic lens, concluding that the unconscious is structured like a language and that unconscious processes like condensation and displacement are equivalent to linguistic processes of metaphor and displacement. Lacan emphasized that the human being is a divided subject, split between conscious and unconscious processes, between language and speech, and constituted in and through language. Lacan redefines human existence as living in between the realms of the imaginary, the symbolic, and the real, covering all aspects of life that are visually and symbolically represented and capturing those that lie beyond grasp through the register of the real. Lacan's psychoanalytic theory is a theory of lack in which the human subject is held in his longings between a complex interplay of need/demand and desire. Lucan deduced that because every human subject is born dependent upon the Other, his subjectivity is shaped by the desire and the law of the Other since it is for and through the Other that human desire is founded. The father as the symbolic representative of law and culture and responsible for the separation of the child from his incestuous desire for and by the mother, is crucial in Lacan's reconceptualization of Freud's oedipal complex. For Lacan, the oedipal complex is not a developmental stage but a permanent structure through which the human being negotiates his desire within or outside the law.

Jacques Lacan, born on 13 April 1901 (Paris), was the youngest child of a family of Orleans-based vinegar merchants and was raised by a strictly Catholic mother and a father who was perceived to be rather weak. He was the eldest of four children, with one sister and two younger brothers, of whom one died, and the other joined a Benedict monastery where he became a priest in 1929. Lacan pursued his early studies at the prestigious Jesuit *College Stanislav* and began his medical studies in 1926, specializing in psychiatry at the *Sainte Anne Hospital*, Paris. During his medical studies Lacan broke with Catholicism and developed a growing interest for philosophy, especially the works of Karl Jaspers, Martin Heidegger, and Spinoza, and he attended seminars about Hegel given by Alexandre Kojeve. He also professed a strong interest in literary studies and art, in particular, in the literary works of James Joyce and the surrealist movement. He was close friends with Andre Breton and Salvador Dali and published some of his earliest writings in the surrealist magazine *Minotaure*. Had it not been for his particular French character, one who was interested in philosophy, history, and language, Lacan might have not been as successful as he was in introducing psychoanalysis to France, a country that had been highly suspicious of the foreign roots of psychoanalysis.

In 1931, Lacan received his license to become a forensic psychiatrist and in 1932, he started his training analysis with Rudolph Loewenstein. Within the same year he finished his dissertation on "The Paranoid Psychosis and Its Relation to the Personality" in which he analyses the case of a woman, 'Aimée,' who later turned out to be the mother of his initial companion and later rival Didier Anzieu, who became a prominent psychoanalyst in his own right. Aimée (the pseudonym for Anzieu's mother) had attempted to kill a famous actress whom she had imagined had spread slander about her, and Lacan had interpreted this attempted murder as Aimée's unconscious wish to merge her own image with that of an idealized other one, whom Aimée had thought to destroy in a wish for self-punishment. Aimée's case burgeoned Lacan's long-standing interest in paranoiac themes, such as jealousy, persecution, grandiosity and the essential importance of the image and its influence on the identity formation of the subject. In writing about this case that he had seen at the *St. Anne Hospital*, Paris, Lacan was greatly influenced by Freud's (1922) paper *Some Neurotic Mechanisms in Jealousy, Paranoia and Homosexuality*, which he also translated from German to French, a practice, incidentally, he maintained throughout his life, of translating papers from German or English to French, which were of importance for his psychoanalytic studies and publications. From Freud's text and the case of Aimée, Lacan took the idea that a paranoiac knowledge existed at the core of each subject and that the confusing image of the Other strongly shaped, as he would later formulate in his concept of *The Mirror Stage*, the formation of the ego (the 'moi imaginaire'). A painter, like Salvador Dali, strongly agreed with Lacan's basic tenet that a paranoiac mechanism exerted a strong force in the construction of each subject.

In addition to his collaborations with the surrealist writers and painters of his epoch, Lacan was also fascinated during his medical studies by his teacher, Gaetan Gatian de Clearambault, who had been a well-known forensic psychiatrist and an avid collector of photographs of North African female garments. However, he had become even more prominent through his 'spectacular' suicide since he shot himself in the head, sitting in front of a mirror, after he had learned that he was destined to lose his eyesight due to a retina infection. The tragic irony of Clearambault's specular suicide is not lost on most readers of Lacan's work since it powerfully illustrates the aggression underlying the narcissistic ego formation, which Lacan discusses in his concept of the mirror stage.

After 6 years of analysis, Lacan left his analysis with Rudolf Loewenstein in disagreement. He nonetheless joined

the Parisian Psychoanalytic Society (SPP) in 1938 and became a member of the SPP. At the 14th Congress of the International Psychoanalytic Association (IPA) in Marienbad in 1936, he presents his first analytic paper on "The Mirror Stage as Formative of the Function of the Ego." His presentation is disrupted by Ernest Jones, the president of the IPA, after 10 min, because he was unwilling to extend Lacan's slated presentation time. Lacan leaves the congress in protest and presents the paper 13 years later in its full and extended version at the Zurich IPA-Congress in Switzerland. Jones' autocratic behavior at this highly politicized congress set yet another signal to Lacan of the increasing influence the ego psychologists would hold over the developing psychoanalytic discourse. His own former training analyst, Rudolf Loewenstein in conjunction with Heinz Hartmann, David Rapaport, and Ernst Kris were pursuing a psychoanalytic path that was already highly suspect to Freud himself, and which focused on the powerful character of the ego and the therapeutic benefits that would be gained with the strengthening of the ego. In line with Freud's (1937) own late misgivings about the seeming normality of the ego, Lacan unmasks the ego as an illusionary and deceiving structure that alienates the subject from his or her unconscious desire. Instead of strengthening the ego, Lacan advocates the dismantling of the ego as a means of setting free the subject's unconscious desire, allowing the subject to speak its own truth. In 1949, Lacan writes:

> The mirror stage is a drama whose internal thrust is precipitated from insufficiency to anticipation- and which manufactures for the subject, caught up in the lure of spatial identification, the succession of phantasies that extends from a fragmented body-image to a form of its totality that I shall call orthopaedic-and, lastly, to the assumption of an armour of an alienating identity, which will mark with its rigid structure the subject's entire mental development. (p. 4)

Lacan defines 'The Mirror Stage' both as a developmental process and a permanent, 'rigid' structure through which the young 'infans' (that means without language) forms an image of his body through the *Gestalt* in the mirror which promises a more unified and complete body image than the child's actual motoric capacity. The young infant typically greets this more unified image with jubilation (*jouissance*) because the child recognizes in the mirror a body that holds out the illusion of a unity and mastery that he cannot yet physically experience but which sets him up on an anticipatory level for a desire to reach for a more complete specular version of himself than he is capable of. The tension between the perceived image and the body creates a phantasma of a body felt to be in pieces (*fantasme du corps morcelé*) which appears in psychotic regressions or also at moments in the psychoanalytic treatment when certain levels of awareness are reached. The unified image the child prematurely obtains is always at risk to fall apart and threatens to haunt the subject in his own subjectivity. While the mirror stage was held to be initially a primarily specular process, where the infant identifies with the image projected ahead of him and mistakes that image to be himself, Lacan later also always emphasizes that this imaginary process unfolds in the presence of an Other, a third who speaks to the child and identifies him as the child in the mirror. "Look who is there, there is little Peter." Through the language of the Other, typically the mother or father who holds the infant in her/his arms, a symbolic identity is introduced intertwined with the imaginary one. In the moment of the parent saying or expressing to the child something like "This is you!," a triangulation takes shape in terms of the imaginary, symbolic, and real registers, the image, the word, and the referent. With the mirror stage, Lacan not only expands Freud's concept of the ideal ego and ego ideal, but he also provides a radically new idea about the process of identification as a specular, desiring, and alienating process where the 'I is the Other.' By unmasking the ego as an agent of deception, Lacan also pays tribute to Freud's late realization that the ego is not normal but shares many characteristics with the psychotic, as Freud describes in his 1937 text, *Analysis Terminable and Interminable*.

During World War II, and while The Société Psychanalytique de Paris was closed by the Nazis, Lacan served in the French army at the Val-de-Grace military hospital in Paris. By the beginning of the war, Lacan had been married for 6 years to Marie-Louise Blondin and had fathered two children, his daughter Caroline, born in 1934 followed by his son Thibault in 1939. In 1940, a third child, Sibylle, was born. However, a year later, Lacan fathers another girl, Judith, with another woman, Sylvie Bataille, the estranged wife of George Bataille, whom he marries in 1953 after having divorced Blondin beforehand. Judith, his youngest daughter, will become later very important for the transmission of her father's work together with her husband Jacques-Alain Miller. He will be in charge of Lacan's publications and distributions of seminars after Lacan's death in 1981.

Due to his appeal for a return to Freud and his earlier writings, Lacan comes to the forefront of the international psychoanalytic scene in the 1950s. This appeal is linked to an intensive focus on the philosophical thoughts of Heidegger, the linguistics of Ferdinand de Saussure, and the structuralistic concepts of Lévi-Strauss. Lacan's postwar writings are increasingly focused on the theory of the symbolic register, language, speech and the fundamental concept of the Other as a representative of the law, culture, and the storehouse of signifiers. Rereading Freud's early texts, like *The Interpretation of Dreams* (1900), *The Psychopathology of Everyday Life* (1901), and *Jokes and Their Relation to the Unconscious* (1905), through a lens of structuralism and linguistics, brings into sharper focus the fact that the human subject is primarily constituted in and through language and that this language which precedes the subject, speaks the subject as well.

In the above mentioned essays, Freud not only discovers the laws and rules of the *primary process* as the structure of the language-like psychoanalytic unconscious, but he also proves himself to be a forerunner of modern linguistics by representing conceptions that relate closely to the later school of semiotics, in particular to the work of Saussure and Roman Jakobson.

With the discovery of the mechanisms of condensation and displacement which are so essential for the work of dreams, jokes, and parapraxes and which will later be conceptualized as metaphor and metonymy in the field of linguistics, Freud provided not only the basics of a rhetoric which illustrated the poetic aspects of the unconscious but also the principles of signification in any language.

Identifying the unconscious as a linguistic mechanism led to the understanding that thinking and speaking were not hierarchized activities, but processes which were mutually constituting themselves. The human language does not consist of deep structures that shape thinking or the oral expression of words, but it is instead constituted by what Saussure calls 'signs,' which are elements that combine sounds with concepts. The concept is called the signified and the sound of the word is called signifier. The sign is not a primary name that identifies an external phenomenon, but it consists instead of a combination of a signifier and signified, which are related to one another in an arbitrary way. Language is thus no longer thought to be representational, but primarily conceived in terms of articulation, in terms of sounds relating to one another in an arbitrary fashion. Language, conceptualized in terms of articulatory principles, is governed by a constitutive law of difference, represented by the symbolic order, and held in contrast by the constitutive law of similarity, which is instituted by the register of the imaginary. The relationship between signifier and signified is arbitrary and codetermined by social conventions/understanding/agreement. For instance, signifiers and signifieds alike exist because they are something, which others are not. Sounds, in other words, single letters can only function as signifiers as long as they differentiate themselves from others. A thought or concept can only become a concept/thought as long as it is different and separate from another. Binary opposition and differentiation mark the structure of an articulated speech.

Unlike Saussure, who privileged the signified over the signifier, Lacan reverses this connection, and sets the signifier over the signified, underscoring thereby the importance of sound over an agreed-upon social meaning. Lacan also wanted to place particular importance upon the signifier's ability to shift between meanings because without this ability, a joke or a dream cannot properly function. We hear, for instance, the word 'wait' and can hear the word 'weight' at the same time, and it is this sliding between sounds in each language that accounts for the multiple layers of meanings between words embedded in each language. In addition, Lacan wanted to emphasize that speech comes before thinking in the development of the human being, and that the human subject is already spoken and phantasized about long before his actual birth. He is impregnated by the parents' conscious and unconscious wishes, desires, and expectations and is thus intrinsically captured in his being through the language of the Other. The human subject is not dominated by language, but he is constituted in and through language.

Lacan places at the forefront the double birth of the human being – of the 'I' (Ego/moi) through the specular identificatory process during the mirror stage, and of the subject (je) through the determination of the signifier in the acquisition of language. The human subject is however furthermore alienated through his premature birth and his subsequent fundamental helplessness which Lacan connects with Heidegger's notion of 'Seinsverfehlung' and translates as 'manque-a-etre,' which is akin to the English neologism 'Want-to-be.' Lacan's psychoanalytic theory can be conceptualized as a theory of lack for which culture is destined to complement the deficits of the human drives and his essential drivenness.

With his theory of the double birth of the human being as an imaginary ego and a symbolic subject (corresponding to the myths of Narcissus and Oedipus), Lacan introduces an object relations theory with medial dimensions running parallel to Freud's drive theory. As categories of representation the imaginary and symbolic realms capture the internal and external world of the human being in permanent modifications, while the real as a third dimension covers all those aspects of the human world that cannot be symbolized, are not mediated and are beyond human grasp. Lacan will devote the last part of his theoretical work to the workings of this register of the real, and devote his clinical focus upon the various aspects of the human's use of *jouissance* as a kind of excess pleasure going beyond images and language, not undergoing symbolic castration, which implies the acceptance of an inevitable lack of being.

In contrast to Freud, but in agreement with Melanie Klein, Lacan situates the oedipal complex in the early development of the child. During the third phase of the mirror stage, around the 14th month of his life, the baby enters into the part of the early stage of the oedipal complex as he is gradually identifying his bodily 'moi' with the image projected of himself by the outside world. Since the young infant is born prematurely and is therefore totally dependent upon the help of the Other, he is dependent upon the desire of the mother. In experiencing this desire the child strives to be the object of the desire of the Other, thus identifying with the fundamental maternal lack and unconsciously with the desire to become the mother's phallus. Eventually, the father intervenes into this imaginary triangle as one who spoils this game of illusionary omnipotence between mother and child. He intervenes as the representative of the law and culture and forbids the child his phallic identification and thereby also prohibits the mother to use her child as her phallic substitute. Lacan identifies the appearance of the father as the representative of the law as the second step in the oedipal complex.

During the third phase of the oedipal complex, the child is supposed to identify with the father and to submit himself to the symbolic order, the order of language. The father is identified as the carrier of the word and the representative of the law, but he has to be credible in this symbolic function in order to be accepted as the 'nom-du-père.' 'Nom' and 'Non' – the two signifiers sound alike and combine the function of the father in one word as the one who has to act in his own name when he says 'no' to the mother/child phallic dyad. The mother has to accept the father as a representative of the law so that the child can accept the name of the father as well. If there is no maternal recognition of this symbolic father and if the child refuses the law, the imaginary predominates and the child remains submitted to the mother (neurosis or perversion). In the worst case, which is the case of psychosis, not only is the process of *separation* lacking, but also the process of what Lacan defines as *alienation:* the child functions as a real part of the dyadic and symbiotic mother–child relationship and therefore will not really arrive at a clear differentiation between the self and the Other.

If, however, both mother and child accept the paternal law, which does not have to be represented by a real father figure, but can be assumed by another figure or instance filling out this symbolic position, the child identifies with this third element supposed to carry the phallus as something the mother is

longing for outside of the child himself. The symbolic father is the one to place the phallus in its rightful position. He marks the phallus both as a desire of the mother and as a signifier of a principal lack. Nobody really possesses the phallus, as the identification of the phallus with the masculine penis proves to be the result of an illusionary misjudgment. The child is no longer the one the mother lacks, it is other and instead of being a desired object it becomes a desiring subject. Through an identification with this Other, the child enters into the category of having and leaves behind the primacy of being.

The movement from being to having is a symbolic castration: The father castrates the child by differentiating him from the phallus and by separating him from the mother. With this shift the child becomes a subject who differentiates himself from the other two subjects. As the child moves from the imaginary to the symbolic register, his counterpart also changes from being a double to becoming a speaking Other. However, even the symbolic castration does not abolish the internal subjective division because the subject remains still alienated after his specular identification, since he can neither find a 'Seinsgrundlage' (ontological fixation) in speech. He is also only represented by the signifier referring to another signifier, or, as Lacan writes, that "… a signifier is that which represents a subject for another signifier" (1978: p. 207).

When the subject says 'I,' his division becomes obvious, he is at the same time the subject of his enunciation and subject of his expression. Lacan identifies the mature subject as a permanently divided subject and symbolizes that division through a bar of the subject: $\$$.

As the incestuous desire for the mother is being repressed through the installment of the name of the father, the phallus as a sign of lack is also repressed. The repression of the phallus is equivalent to primary repression, which constitutes the unconscious and initiates secondary repressive mechanisms. In this constitutive process, the signifier of lack, of a border, and of a prohibition is anchored in the unconscious. The infantile wish to be joined with the primary Other is forever repressed, which does not mean that the human subject does not continue to keep looking for it and to find only substitutes for it in its place. In this way, the wish turns into the unsatisfiable desire which so characterizes Freud's notion of a wish and which is part and parcel of Freud's definition of the unconscious. Since desire is embedded in this open system of language – where a signifier always points toward another signifier, which means that meaning can never be limited, Lacan comes to the conclusion that the unconscious is not just made of desires (wishes), but that it is also constructed like a language. The unconscious can be seen as a site of repressed and displaced signifiers, which began to be formed through the repression of the fundamental signifier of the phallus.

The signifiers belong to an order of letters – sound letters – and are therefore less bound to a system of meanings or sense. Lacan references Freud's idea that the dream uses signifiers through condensation and displacement to shift and distort meanings to evade the censor, and this work of distortion is less guided by a system of contents but rather by a system of formal structures. The unconscious is a system of writings as Freud already described it in Letter 52 to Wilhelm Fliess in 1896 when he identified the unconscious as an apparatus of 'Niederschriften' (written inscriptions). Language and writing come from the outside, come from the Other who speaks. For this reason, Lacan also defined the unconscious as The Other who speaks (and writes) in us and by whom we are also spoken (and written).

As a consequence of these considerations, Lacan differentiates between *need*, *demand*, and *desire* which constitutes his detailed elaboration and reworking of Freud's drive model. To begin with, one has to remember that each drive is a partial drive, which is bound to a particular erotogenic zone. Even though the genital drive is the one to privilege sexual strivings, other drives are also sexually bound. Each drive is also an amalgam of *eros* and *thanatos* because the moment the drive is satisfied, it works toward its own self-destruction. As Freud already described in *Beyond the Pleasure Principle*, the human being is also mortal through his sexual procreation. In order to repress the destructive part of the drive, the drive has to be blocked, tamed, and pushed back. In Lacan's writings, the drive is primarily bound to the lack (*manque*) to the exigencies of life (*Not des Lebens*) as Freud formulated it. The drive expresses itself initially as a need which forms a closer link to the psychosomatic than to the psychic life because it does not necessarily require a subjective experience to be registered intrapsychically, as is the case in an insulin coma for instance, where the body's need for glucose bypasses any intrapsychic awareness.

Pure need aims at an object and finds satisfaction in it. The infant's first needs can only be satisfied through another person, primarily hunger which is satisfied through nursing and here the infant encounters for the first time a forceful power over which the mother wields because she can give and withdraw her breast as she sees fit. Lacan focuses on this power which installs itself prior to the satisfaction of needs and transcends the satisfaction or deprival of needs, which in turn points to the dynamics of what Lacan calls 'the demand.' Demand is not really concentrated upon real objects but is essentially focused upon a demand for love and exerts itself through presence and absence, through giving and withdrawing the longed-for love. The early helplessness of the young infant is the source of this demand which Freud had described as follows:

> The biological factor is the long period of time during which the young of the human species is in a condition of helplessness and dependence. Its intra-uterine existence seems to be short in comparison with that of most animals, and it is sent into the world in a less finished state…The biological factor, then, establishes the earliest situations of danger and creates the need to be loved which will accompany the child through the rest of its life.
>
> 1926: p. 155

The power, which distorts the access to the object, has a tendency to debase and minimize the need for satisfaction and to destroy the demand for love within the same act. A consistent misrecognition for a demand of love can have disastrous consequences as one can see in cases of anorexia nervosa. Lacan asserts that in anorexia nervosa, the child is always given what the Other one has (nutrition) but is not given what the Other one does not have, which is love. As a result, the child refuses any nourishment and sets into play a complex dynamic of refusal and desire.

Desire is the third term Lacan enlists in his triadic structure of need, demand, and desire and it is most akin to Freud's concept of a wish (Wunsch) that aims in the unconscious for something that cannot be satisfied. Lacan's assertion that "the desire is the desire of the Other," which means that we receive the desire through the words of the Other and that what we desire is always other, is always different.

While working on these different psychoanalytic concepts in the 1950s and developing a new body of psychoanalytic thinking through a simultaneous return to Freud's early texts, Lacan encounters during these productive years several conflicts with other psychoanalytic institutions. In the early 1950s a crisis erupts in the 'Société Psychanalytique de Paris,' which eventually leads to a division that is primarily based upon the question of lay analysis. Due to his code of analytic practice, in particular, the variable duration of analytic sessions, Lacan joins the liberal group around Daniel Lagache – much to the discontent of his own orthodox institute. This group develops into the 'Sociéte Francaise de Psychoanalyse' and is established as the second Parisian Association in 1953. In Lacan's opposition to American ego psychology he is strongly supported by the child psychoanalyst Francoise Dolto. However, as a result of its controversial stance, the group loses its membership in the IPA without either wanting or noticing it.

After long and difficult years of negotiations for reentry, the IPA demands Lacan's and Francoise Dolto's expulsion as trainees. In 1964, Lacan finally founds his own school called 'Ecole Freudienne de Paris.' Thirty-four articles by Lacan are published as 'Ecrits' in 1966. During 1966–76, Lacan frequently travels throughout the United States for the purpose of lecturing. Due to Lacan's teachings, the University of Paris VIII founds a department for Psychoanalysis in 1969, offering a psychoanalytic doctoral program. In some way influenced by the students' revolts of 1968, Lacan develops his theory of discourse as an attempt to insert his concept of the unconscious into social relationships. In 1969, Lacan also introduces the so-called 'passe' for the training of psychoanalysts. This special type of accreditation, however, is not recognized by everyone and, as a result, some of the members leave the school in order to establish their own psychoanalytic society ('Quatrième groupe').

During the last years of his life, Lacan is preoccupied with topology and mathematics in order to grasp the dimension of the real with its elements of the 'thing' (la chose, near to Freud's 'Ding') and the object 'a,' attempting also to elaborate the very connections between the categories of the real, the imaginary, and the symbolic. At the same time his 'Ecole Freudienne de Paris' is increasingly weakened as an institution due to the assumption that Lacan is no longer capable of acting in his own name, because of his son-in-law's, Jacques-Alain Miller's, supposed takeover of Lacan's school. In 1980, the school is finally closed as an act for a new beginning. Shortly before Lacan dies due to a malignant intestinal tumor on 9 September 1981, the school is refounded as 'Ecole de la Cause Freudienne.' (See for further biographical discussion: Elisabeth Roudinesco, 1999.)

See also: Eating Disorders; Freud, Sigmund (1856–1939); Hegel, Georg Wilhelm Friedrich (1770–1831); Infancy and Childhood: Emotional Development; Klein, Melanie (1882–1960); Language Development, Theories of; Levi-Strauss, Claude (1908–2009); Linguistics: Overview; Psychoanalysis, History of; Psycholinguistics: Overview; Saussure, Ferdinand de (1857–1913); Self: Philosophical Aspects; Stress in Childhood: Effects on Development.

Bibliography

Freud, S., 1892–1899. Extracts from the Fliess papers. In: Strachey, J. (Ed. and Trans.). The Edition of the Complete Psychological Works of Sigmund Freud, vol. 1. Hogarth Press, London.

Freud, S., 1900. The interpretation of dreams. In: Strachey, J. (Ed. and Trans.). The Edition of the Complete Psychological Works of Sigmund Freud, vol. 4. Hogarth Press, London.

Freud, S., 1922. Some neurotic mechanisms in jealousy, paranoia and homosexuality. In: Strachey, J. (Ed. and Trans.). The Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 18. Hogarth Press, London, pp. 221–234.

Freud, S., 1926. Inhibitions, symptoms and anxiety. In: Strachey, J. (Ed. and Trans.). The Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 20. Hogarth Press, London, pp. 77–178.

Freud, S., 1937. Analysis terminable, interminable. In: Strachey, J. (Ed. and Trans.). The Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 23. Hogarth Press, London.

Lacan, J., 1977. Ecrits (A. Sheridan). W.W. Norton & Company, New York.

Lacan, J., 1932/1975. De la psychose paranoiaque dans ses rapports avec la personalite suivi d' Ecrits sur la paranoia. Editions du Seuil, Paris.

Lacan, J., 1978. In: Jacques-Alain Miller (Ed.), The Four Fundamental Concepts of Psycho-analysis (Alan Sheridan, Trans.). W.W. Norton & Company, New York.

Roudinesco, E., 1999. Jacques Lacan. Columbia University Press, New York.

Laicization, History of

Jean Baubérot, École pratique des hautes études, Sorbonne, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Secularization is nowadays the preferred term to describe the movement of societies away from the cultural or political dominion of a particular religion or of a sometimes competing set of religious beliefs. This article argues that a distinguish is usefully made between secularization and laicization, the latter being the largely political process that accompanied the growth of toleration and religious liberty in Western Europe after the end of the wars that accompanied the Reformation. The concept and its history are explored in this article.

During the second half of the twentieth century, sociologists often used the notion of *secularization*. This notion was basically seen as the decrease of religion's influence on modern society. Nowadays, not only can one notice that the decrease is not complete but one can also criticize the notion of *secularization* for being too broad. Often, this notion includes both a cultural and a political process, causing some confusion: for Gill (2008) in several cases, the origins of 'religious liberty' are essentially based on 'the political interest' of rulers. For a better understanding, one might want to distinguish two processes, a dominant cultural and social one (secularization) and a prevailing political one (laicization), without ignoring the interactions between the two.

When the process of secularization is predominant, the relative decline of religion's influence takes place in parallel with other evolutions, with no major tensions between religious and political, or other social forces. Certainly, religious changes, as well as economic, political, and other changes, may produce internal tensions. But, when all is said and done, the triumphant forces participate in the same social dynamic. Therefore, there is no important clash between the changes within the religious sphere and other social changes. For that matter, Scandinavian countries are a typical example of a dominant process of secularization. The switch to Lutheran Protestantism, particularly linked to the Bible's translation, favored the development of a national culture. Theological arguments ('theology of the two reigns,' absence of essential differences between clerics and laymen, rejection of papacy, etc.) have allowed the autonomy of the Nation-State toward religion not to provoke important conflicts. Moreover, a joint and progressive democratization of State and Church has taken place. A national church has emerged – this continues to be the symbol of the identity of nations that have been largely secularized.

The prevalence of process of laicization differs: the tensions between various social forces take on the aspect of an open-conflict where the religious becomes a political stake. The symbolization of the national culture, among others, is politically controversial. The State either imposes religion to society (*clericalism*) or it prevents religion from being the symbol of national identity (*laicization*). The French case is the main example of such a process.

Revolution in France and After: The First Threshold of Laicization

In France, the roots of this conflict are the results of the religious wars (sixteenth century) as well as the persecution of Protestants and dissenting Catholics (Jansenists) (end of the seventeenth century and eighteenth century). As a reaction to this, a certain antireligious dimension appeared in the French philosophy of Enlightenment. This spirit differed from the English Enlightenment (or the German *Aufkärung*). The latter appeared as a confrontation within a diversified Protestantism, wanting an 'enlightened religion;' the former directly attacked the Roman Catholic Church, '*l'infâme*' as Voltaire wrote, who was a deist.

In 1789, religion pervaded French society. The monolith Roman Catholic faith was made compulsory for the French people. The Catholic Church was an inclusive institution with power over other institutions, which often only existed in embryonic states. The monarchy had a religious justification with the theory of the divine right of kings. The king possessed a politico-religious power and the defense of Roman Catholicism was one of the monarchy's basic duties. But gradually monarchy also established a large autonomous political and administrative structure. A Gallican movement developed the idea of a French Catholic Church, which was autonomous from the Pope, protected and controlled by the State.

The first collapse came with the *Declaration of Human Rights*. It proclaimed the principle of religious freedom (1789), which was progressively applied to minorities. The Constitution divested the monarchy from its religious character. The monarch had a contract with the nation. The fact that the King did not respect this contract led to the conflictuous establishment of the Republic (1792). The state no longer had religious foundations. At that time, a breach between the Catholic Church and the Revolution appeared. There were three reasons for this: the proclamation of religious freedom, the bill of the Civil Constitution of the Clergy (1790), which fostered Gallicanism, and the establishment of civil registers for births, marriages, and deaths. The 'religion of the Fathers' could not embody the new society any longer, and the antirevolutionary propaganda became associated with Roman Catholicism. By the year II (autumn 1793), the institution of revolutionary cults went with

a violent struggle (period of the Terror) directed against revealed religions.

Napoleon Bonaparte inherited a chaotic situation. He thought it would be possible to restore the Roman Catholic Church and maintain some of the Revolution's great achievements. A Concordat was signed with the Pope (1801). The Roman Catholic Church was proclaimed 'religion of the great majority of the French' but did not have the status of a state religion; the 'Organic Articles' recognized two Protestant Churches (Lutheran and Reformed) (1802) and the Jewish religion (1807). In 1804, the Civil Code had no reference about religion: civil laws were generally secular (*laïque*) even so France was relatively unsecularized.

It is possible to assess this situation, using the abstract notion of the first threshold of laicization (constructed on the basis of Weber's ideal type), which was the permanent result of revolutionary upsets and is marked by three characteristics:

1. *Institutional fragmentation*: Roman Catholicism was no longer an inclusive institution. Religious activities were more and more clearly distinguished from 'profane' activities. Education and health gradually became autonomous sectors, with specific institutions that were going to develop progressively.
2. *Recognition of legitimacy*: The Revolution did not destroy religious needs that continued to exist objectively. The service provided by religion was a public service and the state paid ministers of recognized religions. Religions were socially recognized as a foundation of social morality.
3. *Religious pluralism*: The state recognized four religions; it protected and controlled them for they could satisfy the religious needs of their followers and develop moral values. Other religions were more or less tolerated. Relative freedom was even given to those who decided to do without 'the help of religion.'

The French Revolution did not succeed in establishing a durable system of separation of Church and State whereas the American Revolution did. In the United States, the Declaration of Independence stipulates that all human beings *are created equal, that they are endowed by their Creator with certain unalienable Rights*, which is not the case in the French Declaration of Rights. Why? Because in the United States, religious pluralism allows the dissociation of God from every church.

In France, religious monopoly would let the Roman Catholic Church become the owner of human rights, being the mediator between God and the Humans. This creates a structural difference in the long run. The American situation may have specific characteristics linked to the major function of religion: a social space for cultural pluralism. Stephen Warner (1993) indicates the main differences:

1. Religion is constitutive for some American subcultures.
2. Religion in America has historically promoted the formation of associations among mobile people.
3. Religion in America serves as a refuge of free association and autonomous identity, a 'free social space.'
4. The second generation of immigrants often transmutes ethnicity into religion because it allows assimilating and conserving their identity.

From the First Threshold to the Second Threshold of Laicization

In France, the compromise of the first threshold of laicization delineates a logic that dominated for almost a century. But the situation was not static. At first, this was due to the fact that at the beginning of the nineteenth century, a number of aspects of the first threshold were not entirely concretized: other institutions were not sufficiently developed to become totally autonomous. Then, especially during the second half of the nineteenth century, there was a growing political–religious conflict between the proponents of revolutionary values and those who considered France mainly as a Roman Catholic nation.

The first conception won. During the 1880s, the Republicans introduced a religiously neutral state educational system, which included a *morale laïque* (1882). At the end of the century, a new crisis arose from the Dreyfus case. The impression prevailed that the Republic was threatened by the *clerical peril*. The important law of 1901 on voluntary societies excluded religious congregations from the right to associate freely. The separation of Church and State appeared to be inevitable and took place in a climate of violent confrontation. But two models of separation were possible. One model was antireligious and combative. However, it was refused, especially by Aristide Briand and Jean Jaures, socialist leaders, who hoped to pass a law that would bring peace and eventually make it possible to fight social problems. A liberal model prevailed, ensured the freedom of conscience, guaranteed the free exercise of religion, and respected the self-organization of each church (art. 1 and 4) even though it neither recognized nor paid subsidies to any religion (art. 2). This marks the end of the system of *recognized cults*.

Even in liberal dominance, this model of separation, after the religiously neutral educational state system, found a new logic. This can be seen as the second threshold of laicization. As for the first threshold, there are three characteristics:

1. *Institutional dissociation*: Religion was entirely optional because it was no longer considered as the institutional structure of society but rather as a mode of free association;
2. *Social neutrality concerning religious legitimacy*: Religious needs became a private matter and public morality officially *laïque*. The question as to usefulness of religion for society was no longer publicly relevant. (In logic of the conflict, the opponents to Roman Catholicism asserted that religion and religious morality were dangerous for the Republic and its ideals. Neutrality was a compromise.)
3. *Freedom of religion*: Various religious societies belonged to the sphere of public liberties. The state guaranteed each citizen freedom of conscience and allowed each one to meet with religious societies or associations.

The separation of 1905 was marked by a twofold aspect: the victory of the *laïque* camp in the hard and long political–religious conflict of the *two Frances* and an implicit covenant with opponents. The conflict ceased progressively because religion was not a political problem any longer. The *Résistance* united unbelievers and believers in a common struggle. By 1945, the bishops held that there could be a positive meaning to the term *laicity* and, the following year, *laicity* was

officially proclaimed by the Constitution. It was a tripartite government – communist, socialist, and MRP (a political party that is close to Christian Democracy) – that included *laicity* in the Constitution.

Laicization in Various Countries

Other countries also show cases of processes of laicization with analogies with the French case as well as specificities. Here are three examples. At first, Belgium became an independent state in 1831, in a 'unionist' atmosphere, which gathered Catholics and Liberals, rather than anticlerics. The outcome was a regime of separation-recognition, accompanied by 'recognized cults' (but no Concordat) which tended to resemble the first threshold of laicization in France – this was possible because the situation was more democratic than that under Napoleon Bonaparte. In 1879, the Liberals attempted to create a *Laïque* school but the Catholic Party, victorious at the 1884 elections, progressively abolished these measures. Therefore, Belgium was never situated in a similar logic to the French second threshold but slowly integrated *laicity* and *laïque* morals in its system of 'recognized cults': 'nonconfessional' morals classes, *laïque* advisers in hospitals or prisons, *laïque* youth fetes that were similar to the first communion, etc.

For a very long time, Spain lived through a refusal of laicization of any sort. The 1912 Constitution declared that "the Nation's religion is (…) the Apostolic and Roman Catholic religion, the only real one." But in the 1931 Constitution, the Spanish Republic imposed both elements of the first French threshold (abolition of the Catholic religious monopoly, institution of the civil marriage) and of the second French threshold (separation, *laïque* schools, suppression of the religious symbols on public buildings) and even a few measures that go beyond the second threshold in France (for example, chaplaincies were suppressed). During the Civil War, this laicization resembled more the situation of the French Revolution than that of the Third Republic. The issue lied in Franco's national Catholicism, which reaffirmed Roman Catholicism as 'the only real' religion. Nevertheless, since 1978, there has been no State religion but 'relations based on cooperation with the Catholic Church and other confessions.'

The Italian situation of the mid-twentieth century was more ambivalent than the Spanish one but was not *laïque* either. The 1948 national reconciliation respected the agreements of Latran (1929), which had been a compromise between the Fascist regime and the Catholic Church. On the other hand, at the same time, it asserted the mutual independence of the State and Catholicism (State religion). This lasted until 1971, when the constitutional court asserted the priorities of constitutional norms on the agreements with the Catholic Church. After that, Catholicism ceased to be the State religion, divorce was authorized, and in 1984 the Concordat was renegotiated and an official pluralism began to exist.

Today's Spanish and Italian systems tend to resemble a laicity that stands between the first and the second French thresholds of laicization. On the other hand, England possesses an established Church that coexists with a maintained Catholicism as well as nonconformist Churches. This constitutes an original variation of the process of secularization. As early as the end of the seventeenth century, politico-religious conflicts no longer took place. In Germany, after the Thirty Years' War, the main process was that of secularization but it also integrated elements of laicization (*Kulturkampf*, in particular).

We then could think that the French situation is the most laïque but this is not the case. The Mexican process of laicization was built in direct opposition to the Catholic Church (Blancarte, 2008). Established since 1859, the separation between State and Church in Mexico became even more drastic and controversial toward religion in 1917. The Constitution showed, for example, that State education 'should remain absolutely away from any religious doctrine' and should fight 'fanaticism and prejudices.' It banned monastic orders, made sure that religious ceremonies only took place in Churches, which 'will always be under the authority's supervision,' and it refused the ministers of cult the right to vote, etc. Some measures were progressively eased, but the Mexican laicity remained stricter than the French one until 1992. Perhaps one could consider it as an anticlerical system rather than an example of laicity.

The Turkish process of laicization is also a stronger one (Kuru, 2009). In 1924, Mustapha Kemal Ataturk removed the caliphate and closed many religious orders. *Laicity* was officially stated in the Constitution in 1937. At that time, religious signs were restricted in public space. In Turkey, even if a moderate Islamic party governs now (2012), the state is separated from religion. But Islam, the largely prevailing religion, is supervised by the State Department of Religious Affairs (*Dianet*) and is not really separated from the state.

New Problems and the European Construction

Nowadays, whatever the degree reached in laicization, various societies must face new problems. For example, a dynamic element in the laicization process encouraged the development of laws about morals, independently from religious standards (like divorce in Catholic countries). Country rulers had believed in the progress of science and its capacity to maintain the social link. But, presently, above all, science is understood as a *techno-science*, the functional efficiency of which is undeniable but which, if it helps to resolve moral questions, also created new ones, difficult to resolve because of their intrinsic power. The development of genetics is significant, and in bioethical committees, some religious representatives are asked to give their point of view. In today's society, there are various debates on morals: questions concerning abortion, gay marriage, euthanasia, etc., are brought up.

In several secular states, similar to the United States and South American countries, the question is: Is religion a legitimate basis of morals lawmaking in a democracy? Many people contest the ban of life's choice or some pressures which are exercised to force them to accept religious ethical standards as 'natural' ones for all.

But for political current affairs, the prevailing debates often concern directly laicity and Islam: for some immigrants' children, this opposes the secularized culture that is taught at school and an orthodox Islam, transmitted by preachers, which seems to them a possibility to discover their roots through

original tradition. In France, but also in other countries, the *burqa* and *headscarves affair* are a typical example. Those supporting a ban argue that Islamic clothes are a symbol of women's repression, although the women themselves protest that it is 'their right and their religion.' But these affairs find significance, also by showing how difficult it can be for the immigrant populations of Muslim faith or origin – who often settled in the outskirts of a city, areas that are prone to break up and directly subject to increase employment and the crisis of the Welfare State – to integrate.

Moreover, the relations between the State and religion, in most of the countries of the European Union, tend to become a specific case of a larger European model. Italian Professor Silvio Ferrari defines it with the following coordinates (Ferrari, 1998):

1. The State is impartial toward the various individual religious citizens.
2. A religious subsector is singled out within the public sector. This may be understood as a playing or protected area field. This is the area in which lay most of the differences between France and the other European countries, because in France this area is not institutional.
3. The State has the right to intervene in this area only to see that the players respect the rules of the game and the boundaries of the playing field.

This model is deeply rooted now in the political and legal culture of contemporary Western Europe. It is ingrained to such an extent, that it has become the basis for the very formulation of the laws in force concerning religious liberties and church–state relationships. This may be seen both in internal conventions and in the constitutional laws of the main European nations. These conventions and laws are formulated to ensure the impartiality of the public powers, which are legally bound to respect the freedom to profess any religious faith (besides art. 9 of the European Convention of Human Rights) and to avoid any discrimination based on religion (art. 14 ECHR). With such a supranational legal situation, secularization can be put side by side with elements of laicity by diminishing the differences linked to the existence of an official religion.

To conclude, the ideal type of laicity can first be characterized as a twofold refusal: the refusal of an official religion, the standards and prescriptions of which would be imposed on all by means of civil law with the state's affirmation of religious neutrality, and the refusal of an official atheism which would limit religious freedom. More positively, laicity can be seen as a close relation between the freedoms of conscience and the freedom of thought. The freedom of conscience implies that every individual has the right to believe or not to believe, to retain or to change his beliefs. Therefore, the freedom of conscience leads to the pluralism of beliefs and philosophies within society. The freedom of thought implies that none of these beliefs or philosophies are privileged or imposed by the State, to the detriment of other convictions. The State is to respect them all and is to make sure that they can all be expressed freely.

See also: Secularization: History of the Concept; Secularization.

Bibliography

Baubérot, J., 2009. Les laïcités dans le monde, third rev. ed. PUF, Paris.
Baubérot, J., 2010. Histoire de la Laïcité en France, fifth rev. ed. PUF, Paris.
Baubérot, J., Milot, M., 2011. Laïcités sans frontières. Le Seuil, Paris.
Blancarte, R., 2008. Laicidad y laicismo en America latina. Estudios Sociologicos XXVI (76), 139–164.
Bowen, J., 2007. Why the French Don't like Headscarves? Islam, the State, and Public Space. Princeton University Press, Princeton.
Cady, L.E., Shakman Hurd, E. (Eds.), 2010. Comparative Secularisms in a Global Age. Palgrave Macmillan, New York.
Ferrari, S., 1998. Tolerance, religion and the law in contemporary Europe. In: Ferrari, V., Helier, Th, de Tullio, E. (Eds.), Citizenship and Immigration, Giuffre editor, pp. 159–177.
Gill, A., 2008. The Political Origins of Religious Liberty. Cambridge University Press, Cambridge.
Kuru, A.T., 2009. Secularism and State Policies toward Religion. The United States, France, and Turkey. Cambridge University Press, Cambridge.
Les laïcités dans les Amériques. In: Milot, M. (Ed.), avril–juin 2009. Archives de Sciences Sociales des Religions, vol. 146.
Warner, R.S., March 1993. Work in progress toward a new paradigm for the sociological study of religion in the United States. American Journal of Sociology 98 (5), 1044–1093.

Land Change: The Merger of Land Cover and Land use Dynamics

William J McConnell, Michigan State University, East Lansing, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article considers the merging of scientific perspectives directed at understanding land-use practices and the social and natural environments that shape, and are shaped by, them. While land cover refers to the biophysical attributes of the earth's surface and immediate subsurface, land use refers to the human imprint at a given location, both in terms of the modification of the land cover that would otherwise exist, and of the driving forces for such modification. Land change models aim to explain past land change, often with a view to projecting possible future dynamics.

The Integrative Study of Land Change

Land change science is concerned with land-use dynamics and with the social and natural environments that shape – and are shaped by – them (Turner et al., 2007). Previously known as land-use and land-cover change, conveying the conjoint nature of this branch of human–environment research, the current appellation reflects the coalescence of an integrated field of study aimed at estimating past human impacts on the earth, often with a view to projecting possible futures.

While closely related, the terms land use and land cover refer to quite different dimensions of the earth's surface that have traditionally been investigated in distinct scientific traditions. Land cover refers to the biophysical attributes of the earth's surface and immediate subsurface. This includes vegetative layers (both natural and artificial), exposed soil or rock, water and ice, as well as artificial structures, such as buildings and pavement. While some land covers are considered pristine or at least minimally touched by human use (e.g., remote tropical rain forests), others are recognized as transformed or radically altered (e.g., urban areas). Land use, meanwhile, refers to the human intent to which the land surface is put, and the actions taken to achieve that intent. While use registers as an imprint on a given location in terms of the change in land cover that would otherwise exist, it also includes the motivations, constraints, and enabling conditions that give rise to the use. Land cover and land use each imply both states and processes, and have multiple and overlapping properties.

Because of the compound nature of the questions relating land-use practices to land-cover states and changes, addressing them involves an amalgam of analytical perspectives spanning, for example, the social and physical sciences, quantitative and qualitative approaches, and academic and policy concerns. Each of the research traditions involved brings its own theoretical perspectives and modes of inquiry, and their coupling is not always easy. In particular, the need to merge locally specific explanations of how and why people use land in certain ways with broader patterns observed at regional or global scales brings into sharp relief a long-standing conundrum of scientific inquiry – scale dependence: the dynamics of land change vary depending on the spatial and temporal scale at which they are observed (Easterling and Kok (2002); *see* Scale in Geography). Nevertheless, the enterprise is bound together by the desire to understand how biophysical conditions, particularly 'initial' land cover, shape human activities (opportunities and constraints), and how these activities, or land uses, alter those preexisting conditions, thereby shaping future social and ecological possibilities, as well as the feedbacks, thresholds, and other attributes of such complex systems.

Origins and Motivations of Land Change Science

The relationship between humans and the natural world appears to be a universal concern of Western thought, occupying a central place in Greco-Roman, Arab, and European mythology and philosophy. Modern attempts to characterize the impact of human activities on the earth include George Perkins Marsh's seminal work *Man and Nature* (1864), and multidisciplinary efforts such as *Man's Role in Changing the Face of the Earth* (Thomas, 1956) and *The Earth as Transformed by Human Action* (Turner et al., 1990). More recently, the Millennium Ecosystem Assessment (UNEP, n.d.) sought, inter alia, to quantify the impact of land use on the delivery of ecosystems goods and services. The coalescence of the field of land change science owes much to the development of earth system science over the two past decades, reflecting a growing appreciation for the key role of land change in biogeochemical cycles. At the same time, attention has increasingly focused on the consequences of global environmental change in terms of the sustainability of livelihood systems, of the vulnerability of particular groups of people and places, and of the societal perceptions of and responses to change (Kasperson et al., 1995). These concerns require understanding the human systems – economic, political, cultural, and sociotechnical – associated with human activities at particular places and times (Moran, 2000). Other important questions driving land change research concern the loss of biological diversity through the destruction of habitat and the issue of land rights in the global competition for increasingly scarce land-based resources required to feed rapidly growing demand for food, fuel, and industrial inputs (Foley et al., 2011) .

These concerns are of more than academic interest; policy is actively being crafted to limit human impacts on the environment at both regional and global scales. Many of these policies focus squarely on land use, such as the reducing emissions from deforestation and forest degradation (REDD+) mechanism, which is intended to limit carbon emissions that occur

when forest is converted to agricultural or other uses. The effectiveness of such international environmental agreements depends on sound scientific understanding of the impact of land use, in order to target critical areas and to monitor compliance. Accordingly, research programs have been developed at national and international levels, including the Global Land Project (globallandproject.org), under the aegis of the International Geosphere-Biosphere Programme and the International Human Dimensions Programme (GLP, 2005). Land change research constitutes a bridge between the understanding of the biophysical aspects of global change and the human dimensions of sustainability (*see* Sustainability and Sustainability Science).

Two fundamental, linked, sets of questions underlie land change research: the first concerns the nature, location, and rates of anthropogenic modification of land cover; while the second seeks to understand the reasons behind land-use practices leading to such modifications. The first set of questions is generally examined at relatively broad scales, often using satellite imagery to determine spatial and temporal patterns of change, while the explanation of land-use practices generally requires local scale research, and often involves interviews with land managers and direct observation of land management practices. As answers to these two sets of questions take shape, it becomes possible to model the trajectories of change that have been observed, and to project future land cover outcomes. The following sections treat these issues in more detail.

Land-Cover Dynamics

Our understanding of land-cover dynamics has advanced rapidly over the past two decades, through technological and methodological advances in our ability to observe the current state of the earth's surface from space, and to reconstruct past land-cover patterns with paleological techniques.

Remote sensing of the earth refers to the use of instruments carried on aircraft and earth-orbiting satellites that detect energy reflected or emitted from the earth's surface in different portions of the electromagnetic spectrum. The land cover of most inhabited portions of the earth has been periodically recorded through aerial photography since the 1950s, facilitating the production of land-cover maps. Since the 1970s, satellite remote sensing data have provided more frequent views of large areas (*see* Remote Sensing; Geographic Information Systems; Geographic Information Systems and Remote Sensing).

Land-cover research has utilized three sets of satellite imagery, though the distinction between them is disappearing. One set of satellite instruments (e.g., those carried on the Landsat, SPOT, and ERS platforms) provides relatively high spatial resolution, sampling the earth's surface on the order of tens of meters, along swathes of 100–200 km. While such imagery is capable of distinguishing several types of land cover at scales closely corresponding to the case studies needed to understand land use, these satellites take weeks to observe the entire surface of the earth, and views of land-cover are often inhibited by cloud cover or other atmospheric conditions, especially in the tropics.

Other satellite systems, originally intended for the observation of weather, acquire daily views of the earth. Short series of these daily views may be combined to create cloud-free composite images for weekly or monthly intervals. These data sources previously suffered from very limited spectral and spatial resolution, but still yielded valuable information on seasonal and longer-term climatic patterns, such as the El Niño Southern Oscillation. More recently, daily imagery is being acquired by sensors with improved spatial and spectral resolution (e.g., NASA's MODIS instrument), further enhancing the ability to detect temporal patterns. These data have helped to support the rapidly growing field of phenological analysis, separating the background of natural cycles and rhythms to more reliably detect anthropogenic land-cover change.

Over the past decade, a third set of imagery has become widely available, mainly through private sector satellite systems, offering very fine-grained views of the earth's surface once available only to the defense/military intelligence community. These commercial systems deliver imagery with pixels corresponding to less than 10 m, and even less than 1 m, on the ground. Concurrent advances in computing power and storage capacity have made the analysis of such imagery for small areas feasible for even moderately equipped laboratories or individuals.

As the data from the various satellite systems are better integrated, the gap will close between the global coverage and local case studies using aerial photography and high-resolution imagery. Attempts at such integration have highlighted the disparities between analytical schemes for characterizing land cover, as well as the difficulties of merging information about biophysical and social dimensions of land change (Rindfuss et al., 2004).

In addition to the very detailed record of land-use and related land-cover impacts provided by remote sensing over the second half of the twentieth century, our knowledge of prior land-use impacts is also growing through the analysis of pollen in sediment cores and other paleoecological and archaeological techniques. Such techniques have enabled the estimation of the earth's land cover 6000 years ago (GAIM, 2000), and detailed reconstructions have been made for the past three centuries by projecting current land cover back in time using historical agricultural and demographic information (Klein-Goldeweik, 2001; Ramankutty and Foley, 1999). The convergence of remotely sensed and other historical sources with prehistoric reconstructions will provide a much fuller picture of the earth's land cover throughout the anthropocene (Crutzen and Stoermer, 2000), yielding valuable insights on long-term trajectories of change, and of possible path dependency in the evolution of regional land-cover patterns (*see* Anthropocene Epoch).

Land-Use Dynamics

Elements of land-use practices have been a staple topic of investigation for numerous disciplines in the social sciences, including anthropology, archaeology, demography, economics, geography, history, political science, and sociology. Through different lenses, scholars in these fields have sought to explain how and why people in different parts of the world have

developed technology and social systems in response to local environmental conditions, how these human actions have, in turn, shaped the local environment, and how people subsequently respond to these changing conditions (Redman, 1999). Thousands of case studies have been undertaken, usually concerning communities or small regions, and efforts are increasingly being made to synthesize these cases in order to extract general principles and to develop a unified theory. While the history of land use in Europe is well documented, land-cover transformation has been perceived to be most dramatic in the tropics in recent years, and a great deal of attention has been accorded these regions, with particular attention to agricultural 'land grabs' (Lambin and Meyfroidt, 2011).

A broad range of factors has been proffered to explain local land-use practices, with perhaps the greatest effort spent examining agricultural change and the role of population growth (*see* Agroecology and Agricultural Change). The now-famous work of British economist Thomas Malthus, who predicted famine as an inevitable result of population growth in Europe, has fostered huge debate within the social sciences. Another major landmark was the work of Ester Boserup 1965[1993], who turned the Malthusian thesis on its head, attributing agricultural development in Asia to land pressure driven by population growth. Subsequent refinements have described a range of land-use change trajectories that may emerge from land pressure and suggested how the driving forces of change are mediated by social and environmental factors at many levels (Roy Chowdhury and Turner, 2006).

In addressing the relationship between population growth and land degradation, political economic analyses hold external forces responsible when land managers engage in environmentally destructive land-use practices. Likewise, cases have been articulated of population growth leading to the rehabilitation of natural resources, including afforestation. A great deal of work has been done to examine the role of demographic factors in land use change, with particular emphasis on government-subsidized settlement programs in Latin America, and rural–urban migration in Africa and Asia. Results suggest complex relationships between land-cover dynamics and the provision of infrastructure (especially roads), family development cycles, urban wage employment, and other factors that vary strongly by region (e.g., Lambin and Geist (2006); *see* Property Rights).

n the case of forest cover, global scale empirical evidence reveals a complex relationship between forest cover and population density, suggesting that as population density rises in newly settled regions, forest cover has often been removed for the purposes of agricultural expansion. Another period ensues during which wood fuel is replaced by other sources (e.g., fossil fuels) and agricultural intensification allows for the production of sufficient food on less land. Eventually affluence rises to the point where attention to environmental rehabilitation becomes feasible and afforestation becomes the dominant trend, in what has become known as the forest transition (Rudel et al., 2005).

In order to synthesize the findings of the hundreds of cases, studies in the peer-reviewed literature, a number of *meta*-studies have been conducted on the processes of deforestation (Geist and Lambin, 2002), desertification (Geist and Lambin, 2005), and agricultural intensification (Keys and McConnell, 2005).

Land Change Models

Perhaps the main challenge of land change science is to forge robust linkages between the broad views of land-cover change provided by remote sensing and other techniques, and the fine-grained understanding of land-use dynamics. In weaving these together, land change modeling examines the observed paths, or trajectories, and often projects them into the future. While this may not always lead to prediction, it can suggest what might happen should present trends continue and, perhaps more importantly, suggest other possible futures. There are significant obstacles to achieving this integration of land-use and land-cover perspectives, but advances in geographic information technologies and new analytical techniques are enabling significant advances (see Spatial Analysis in Geography).

Most early models of land-cover change were either empirical (mimicking the observed relationships between variables, such as forest cover and population density), or mechanistic (describing a limited range of landscape processes according to simple equations) (Kaimowitz and Angelsen, 1998). While such models have the advantage of requiring minimal empirical evidence (e.g., maps of forest cover and population density), their utility is limited by the assumptions that such relationships will persist, and that they are the main determinants of land change (Lambin, 1994). However, it is increasingly recognized that these conditions (particularly in the socioeconomic realm) are mutable, requiring consideration of alternate scenarios outside the empirical record (Aspinall, 2004). In order to incorporate the information gleaned through research on land-use dynamics more realistically, a new generation of dynamic simulation models has undergone rapid development over the past decade that model the behavior – and even the learning processes – of individual actors, or agents, in the landscape (Parker et al., 2003; *see* Agent-Based Modeling and Associated Statistical Aspects).

Summary and Outlook

Land change science is an inherently interdisciplinary field of inquiry, bringing to bear a broad range of perspectives to explicate and model the highly complex interplay of human and natural forces shaping the transformation of the earth's surface. The maturation of the field over the last decade has been marked by its burgeoning breadth of experience and deepening responsibility to the challenge of sustainability. The increasing proliferation of case studies will continue to engender more sophisticated and innovative synthesis efforts, building on developments in *meta*-study techniques across the disciplines (Magliocca et al., 2014; *see* Sustainability and Sustainability Science). These syntheses will benefit from the trend toward the use of continuous variables in case studies, describing both the land change outcomes and the factors involved in creating them in finer detail. Developments in the field will be driven by societal needs to manage the flow of ecosystem goods and services, with accounting for carbon stocks, for example, now crucial to the enforcement of international environmental agreements (e.g., REDD+) and to financial instruments (e.g., carbon credits) intended to stabilize

the earth's atmosphere. Technological advancement will continue to occur both in the availability of historic (i.e., declassified) and contemporary (especially private sector) remotely sensed imagery, as well as through more powerful computing hardware and software, with broad-ranging high-speed Internet access giving virtually anyone the tools needed to detect, describe, and explain changes in the landscape. Together, these advances will contribute to the development of increasingly robust theoretical understanding of this crucial aspect of human–nature interaction.

See also: Agricultural Sciences and Technology; Land-Use Indicators; Planning Issues and Sustainable Development; Sustainable Agriculture.

Bibliography

Aspinall, R., 2004. Modelling land use change with generalized linear models – a multi-model analysis of change between 1860 and 2000 in Gallatin Valley, Montana. Journal of Environmental Management 72 (1–2), 91–103.

Boserup, E., 1965[1993]. The Conditions of Agricultural Growth: The Economics of Agrarian Change under Population Pressure. Earthscan Publications, London.

Crutzen, P.J., Stoermer, E.F., 2000. The anthropocene. Global Change Newsletter 41, 17–18.

Easterling, W., Kok, K., 2002. Emergent properties of scale in global environmental modeling – are there any? Integrated Assessment 3 (2–3), 233–246.

Foley, J.A., Ramankutty, N., Brauman, K.A., Cassidy, E.S., Gerber, J.S., Johnston, M., Mueller, N.D., O'Connell, C., Ray, D.K., West, P.C., Balzer, C., Bennett, E.M., Carpenter, S.R., Hill, J., Monfreda, C., Polasky, S., Rockstrom, J., Sheehan, J., Siebert, S., Tilman, D., Zaks, D.P.M., 2011. Solutions for a cultivated planet. Nature 478, 337–342. http://dx.doi.org/10.1038/nature10452.

GAIM, 2000. Summary Report for the GAIM 6000 yr BP Experiment and BIOME 6000. Global Analysis, Integration and Modeling. http://gaim.unh.edu/Structure/Key-Issues/Biome6000/index.html.

Geist, H.J., Lambin, E.F., 2002. Proximate causes and underlying driving forces of tropical deforestation. Bioscience 52 (2), 143–150.

Geist, H.J., Lambin, E.F., 2005. Dynamic causal patterns of desertification. Bioscience 54 (9), 817–829.

Global Land Project, 2005. Science Plan and Implementation Strategy. IGBP Report No. 53; IHDP Report No. 19. IGBP Secretariat, Stockholm, Sweden.

Kaimowitz, D., Angelsen, A., 1998. Economic Models of Tropical Deforestation: A Review. Center for International Agroforestry Research, Bogor, Indonesia.

Kasperson, J.X., Kasperson, R., Turner II, B.L. (Eds.), 1995. Regions at Risk: Comparisons of Threatened Environments. United Nations University Press, Tokyo.

Keys, E., McConnell, W.J., 2005. Global change and the intensification of agriculture. Global Environmental Change 15 (4), 320–337.

Klein-Goldeweik, K., 2001. Estimating global land cover change over the past 300 years: the HYDE database. Global Biogeochemical Cycles 15 (2), 417–433.

Lambin, E.F., 1994. Modelling Deforestation Processes: A Review. The European Commission, Luxembourg. TREES Series B: Research Report No. 1.

Lambin, E.F., Geist, H.J., 2006. Land-Use and Land-Cover Change. Local Processes and Global Impacts. Global Change – the IGBP Series. Springer-Verlag, Berlin-Heidelberg.

Lambin, E.F., Meyfroidt, P., 2011. Global land use change, economic globalization, and the looming land scarcity. Proceedings of the National Academy of Sciences 108 (9), 3465–3472.

Marsh, G.P., 1864[1965]. Man and Nature: or, Physical Geography as Modified by Human Action. Belknap Press, Cambridge, MA.

Moran, E., 2000. Human Adaptability: An Introduction to Ecological Anthropology. Westview Press, Boulder, CO.

Magliocca, Nicholas R., Rudel, Thomas K., Verburg, Peter H., McConnell, William J., Mertz, Ole, Gerstner, Katharina, Heinimann, Andreas, Ellis, Erle C., 2014. Synthesis in land change science: methodological patterns, challenges, and guidelines. Regional Environmental Change. http://dx.doi.org/10.1007/s10113-014-0626-8.

Parker, D., Manson, S.M., Janssen, M., Hoffman, M.J., Deadman, P., 2003. Annals of the Association of American Geographers 94, 313–337.

Ramankutty, N., Foley, J.A., 1999. Estimating historical changes in global land cover: croplands from 1700 to 1992. Global Biogeochemical Cycles 13 (4), 997–1027.

Redman, C.L., 1999. Human Impact on Ancient Environments. University of Arizona Press, Tucson, AZ.

Rindfuss, R.R., Walsh, S.J., Turner II, B.L., Fox, J.F., Mishra, V., 2004. Developing a science of land change: challenges and methodological issues. Proceedings of the National Academy of Sciences 101 (39), 13976–13981.

Roy Chowdhury, R., Turner II, B.L., 2006. Reconciling Agency and structure in empirical analysis: smallholder land use in Yucatán, Mexico. Annals of the Association of American Geographers 96 (2), 302–322.

Rudel, T.K., Coomes, O.T., Moran, E.F., Achard, A., Anglesen, A., Xu, J., Lambin, E.F., 2005. Forest transitions: toward a global understanding of land use change. Global Environmental Change 15, 23–31.

Thomas Jr., W.L. (Ed.), 1956. Man's Role in Changing the Face of the Earth. University of Chicago Press, Chicago.

Turner II, B.L., Clark, W.C., Kates, R.W., Richards, J.F., Matthews, J.T., Meyer, W.B. (Eds.), 1990. The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere over the Past 300 Years. Cambridge University Press, Cambridge, UK.

Turner II, B.L., Ali, A.M.S., 1996. Induced intensification: agricultural change in Bangladesh with implications for Malthus and Boserup. Proceedings of the National Academy of Sciences USA 93, 14984–14991.

Turner II, B.L., Lambin, E.F., Reenberg, A., 2007. The emergence of land change science for global environmental change and sustainability. Proceedings of the National Academy of Sciences 104 (52), 20666–20671.

UNEP, Millennium Ecosystem Assessment. United Nations Environment Program http://www.unep.org/maweb/en/index.aspx (accessed 25.06.14.), n.d.

Landscape

Kenneth R Olwig, Swedish University of Agricultural Sciences, Alnarp, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Geography, as a modern scholarly discipline, has its roots in the European Renaissance and Enlightenment rediscovery, and rethinking, of the ancient Greco-Roman classics. Though the word landscape is of Germanic origin, its use in geography must be understood in this cosmopolitan disciplinary context, and notably in relation to the key ancient Greek geographical concept of *choros*, meaning 'land,' as the 'land' in landscape. There are two classical paradigms for landscape study: one exemplified by the narrative, humanities-oriented approaches of Herodotus and Strabo, and the other deriving from the mathematical and scientific approach of the geometer, cartographer, geographer, and astrologer Claudius Ptolemy. This article will trace the entangled history of the two as they manifested themselves in diverging ideas of landscape as a meta-disciplinary concept.

Introduction

The Greek Herodotus (*c*.484–425 BC), who helped found history, ethnography, and geography as interlinked studies, and for the latter Greco-Roman geographer Strabo (64/63 BC–*c*.AD 24), *choros* (*also spelled chora and khora*) referred to a particular place, country, or 'land,' with its language and customary laws – e.g., a Greek city-state. *Chorography* meant the largely narrative written description (*-graphy* meaning writing) of a historically constituted land/country, or *choros*. Cartographic representation, by contrast, was foundational to the Alexandrian Greco-Roman geometer, geographer, and astrologer Claudius Ptolemy's (*c*.AD 100–178) conception of geography. Rediscovered in the Renaissance, his geography was promoted by Enlightenment scientists such as Britain's Isaac Newton (1643–1727) and Germany's Bernhardus Varenius (1622–50), as a template upon which modern scientific geography subsequently was significantly built (Warntz, 1989; Martin and James, 1993).

For Ptolemy, 'geography' literally was the graphic science of locating earthly positions on the surface of a globe according to celestial coordinates, whereas 'chorography' involved the demarcation of areas on the globe. For him, as a cartographer, '-graphy' meant imitation/representation through a graphic form, which in the Renaissance was interpreted to mean cosmographic 'pictorial form' (Olwig, 2008). Ptolemy wrote

> Geography looks at the position rather than the quality, noting the relation of distances everywhere. [.] Chorography needs an artist, and no one presents it rightly unless he is an artist. […] In Geography one must contemplate the extent of the entire earth, as well as its shape, and its position under the heavens. […I]n short all things having regard to our earthly habitation.
>
> Ptolemy, 1991, p. 26

Ptolemy's approach reconfigured *choros* in terms of a mathematical, scientific (and astrologic) logic suitable to Roman imperial governance (van Paassen, 1957). A Ptolemaic globe, the prototype for modern cartography, is made by creating a geometric framework of lines of latitude and longitude and then situating locations within it using celestial astronomical observations. Cartographic maps are made by projecting this global space onto the two-dimensional, gridded, and scaled space of a 'graticule' drawn upon a flat and usually square surface. Because, according to Euclidean geometry, lines and points in space are ideal adjuncts to thought found nowhere on terrestrial earth (and are hence utopian), a line has no width and a point no circumference. There is thus no theoretical limit to the points of location that can be plotted on the lines crossing the map's gridded graticule, and there will therefore always be space between them. To demarcate an area on a globe or map one must thus, for both practical and theoretical reasons, connect the dots of the plotted positions, filling in the space between them using the graphic skill, or 'art' of an artist in the archaic sense of an artisan, to create an outline of an area. This activity involves a process of skillful translation from the infinite geometrical integers of a gridded, ideal, Euclidean space, to a finite shape that is analog with a continuous physical terrestrial phenomenon, such as a shoreline. The same skill was required for topographic mapping (**Figure 1**). Today's standard dictionaries continue to define chorography in terms of the art of cartographically defining regions and places, i.e., as: "the art of describing or mapping a region or district," and topography as "the art or practice of graphic delineation in detail usually on maps or charts of natural and man-made features of a place or region especially in a way to show their relative positions and elevations" (Merriam-Webster, 1996: chorography, topography). Through such mapping processes the meaning of *choros* and *topos* underwent a critical change from being a historically and socially constituted land and place, to a technique of representation within a fixed and unchanging ideal geometric space. This technique was useful not only for astrology, but also for commercial and military purposes, as when navigating across an open sea, and for predicting the general location of climate belts, the supposed racial characteristics of peoples that are determined by living in such environs, and other natural phenomena. The technique was therefore vital to the dawning age of world exploration and it was duly embraced by the inheritors of Enlightenment science (Livingstone, 1992). It also made possible the contemporary spatial enclosure of places, so that common lands belonging to political communities with use rights to the material properties of their varied places (*topos*) could be transformed into private properties enclosed as land surface within the Euclidean,

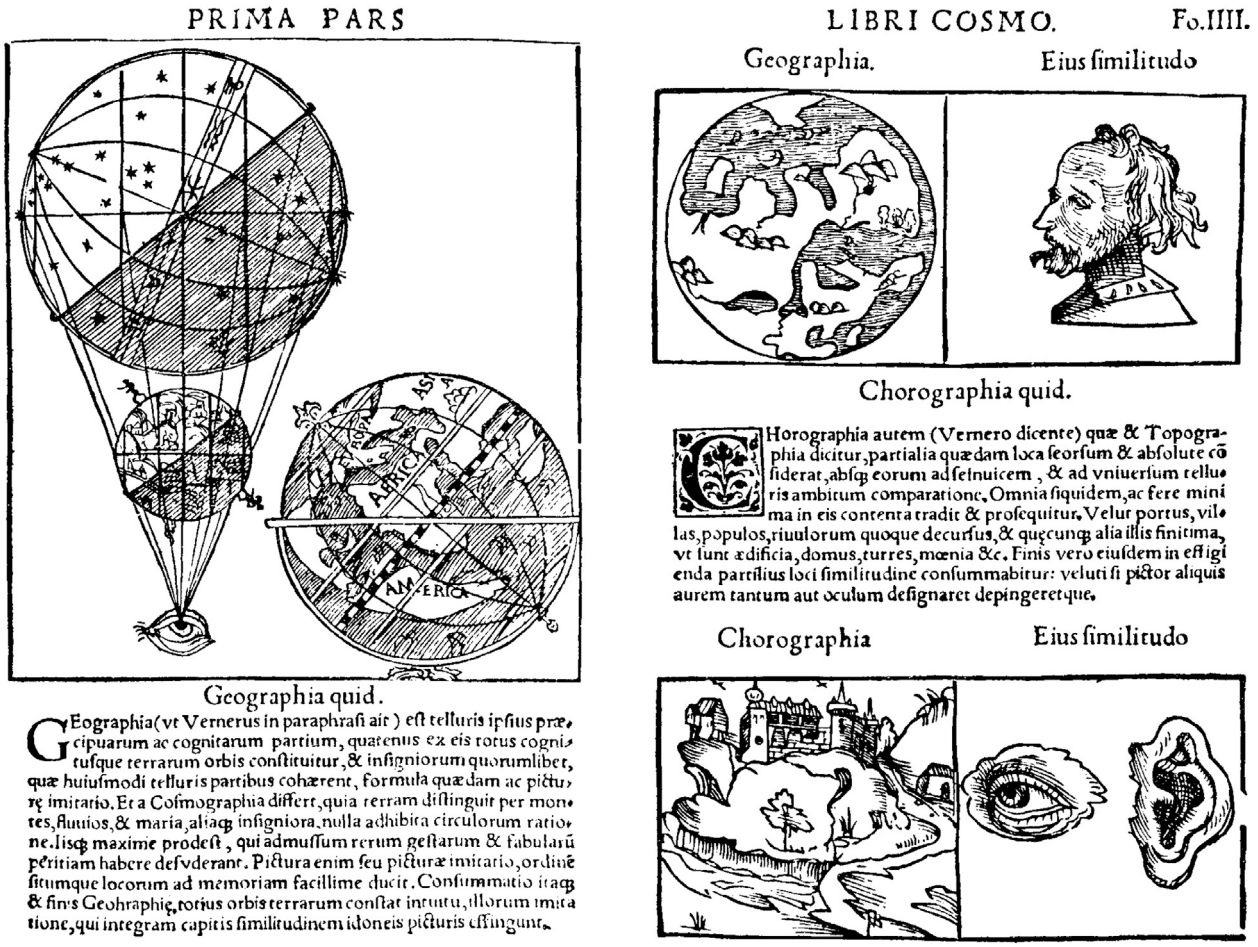

Figure 1 Ptolemy's division of geography into geography proper, chorography, and topography, from Peter Apianus, *Cosmographia* (Antwerp, 1540). The globe of the earth represents geography proper, the chorographic regions of the earth, top right, are analogous to the face.

chorographic space of a cadastral map. This, in turn, facilitated the buying, selling, and agricultural specialization of land as a physical area of property, and thus the transition from a feudal to a modern exchange money economy (Cosgrove, 1984).

The Meanings of Landscape

A landscape was originally, in the Germanic languages, including English, a place (in the original sense of *choros*) having, like a Greek city-state, the character of a country or land, but usually subordinate to a greater authority, as in the Netherlands' federation (Olwig, 2002). The word *pays*, as in *paysage*, likewise meant land as a country, but unlike in the Germanic languages, it could not be confused with land surface, soil, or terrain. The suffix '-scape' in landscape is equivalent to '-ship,' meaning, among other things, 'character,' 'shape,' or 'constitution' as in *friendship* (the characteristic quality shared by friends constituting friendship) (Merriam-Webster, 1996: landscape, -ship). The suffix '*-age*' in the French *paysage* in the Romance languages has the same meaning. A sixteenth-century painting of a landscape polity making up the Dutch federation would seek to capture its '-ship,' character, shape, or constitution as a land – something of political importance due to the imperial, centralizing, threat of Spain. Landscape painting, in origin, thus involved portraiture of the people, customs and history of a land comparable to written depictions of places or lands in contemporary chorography. The pan-European popularity of Netherlands' landscape paintings led to landscape's establishment as an artistic genre alongside portraiture and history painting. But just as Ptolemaic cartography transformed the meaning of chorography, turning it into a form of cartographic representation, the techniques of Ptolemaic cartography changed the meaning of landscape when they were applied to the visual portraiture of landscape.

The Renaissance rediscovery of Ptolemaic cartography provided a basis for the development of pictorial perspective within a spatial framework inspired, at least in part, by Ptolemaic chorography and topography (Cosgrove, 1984). As noted, Ptolemy's notion of the '-graphy' in chorography was interpreted in the Renaissance as referring to a 'pictorial' form of representation involving art, and this idea helped spur the use of spatial perspectival techniques derived from Ptolemaic mapping to create artistic representations in,

for example, landscape painting and theater scenography (Olwig, 2002, 2008). This was done, simply put, by changing the angle of projection from top-down to more horizontal. This scenic representation required a reconfiguration of the grid used to plot cartographic locations, so that its lines no longer form rectangular shapes, but rather appear to taper inward toward an infinite vanishing point. Art is then used to plot and delineate a perspectival scene by connecting points within this ideal spatial framework to form recognizable shapes. As the eye is drawn inward to the point of infinity an illusion of gradually changing scalar depth is created on the flat surface (**Figure 2**).

Just as there is a discontinuity between locational points plotted upon a graticule of a chorographic map, there is also a discontinuous break between chorographic areas plotted upon maps of different scales. Scale derives from *scala*, the Latin for ladder, and this discontinuity is comparable to the break between the rungs of a ladder. The same problem of discontinuity also arises when demarcating topographical layers on a topographical map, so that mountains, for example, are not represented with smooth sides, but appear more like a wedding cake seen from above. It nevertheless proved possible to create an illusion of gradually changing spatial depth through the use of spatial perspective, thereby smoothing the discontinuities of scale. The same problem of discontinuity applies to temporal scale, so that no matter how many historical maps one might make of differing periods in an area's history, there will always be a temporal gap between them, meaning that one must surmise, using the art of the historian, what has happened between each mapped layer of history (Darby, 1962). The use of spatial perspective in theater scenery, however, made it possible to create a realistic illusion of temporal change occurring within the unified time and space of the theater (Olwig, 2002).

The application of central point perspectival techniques to depictions of landscapes had the same effect as the application of Ptolemaic geography to chorography. Landscape thus underwent a similar critical change from being an historically and socially constituted land and place, to being first a form of representation, and then the world understood as scenery, as perceived in such a representation (Cosgrove, 1984). Especially in Italy, these techniques were often used to depict ideal landscape scenes inspired by Biblical and mythical Greco-Roman stories, rather than actual mundane places, as in the Netherlands. The term landscape could thus now be used to refer not to an existing place, but to a form of spatial representation, as well as subject matter perceived in such representations, so that landscape became synonymous with scenery or terrain per se. It thereby became a universalized form of representation applicable to any phenomena, including scientific models of natural processes. However, just as a change in the projection of a map creates significantly different images of the world, so too do changes in the projection of a perspectival representation. The scientific objectivity of a map (and its scale) and perspectival depiction are thus somewhat illusory – a matter of 'point of view.'

Like the Ptolemaic map, the perspectival representation of *choros* as landscape scenery had implications for science. Already in the sixteenth century it had become common to name Ptolemaic Atlases 'theaters,' as with Abraham Ortelius' *Theatrum Orbis Terrarum* (*Theater of the Lands of the World*) (Ortelius, 1595 (orig.) 1570), and they typically included both top-down and more horizontal views, resembling the world as seen as the scenery on the then newly invented perspectival theater stage (**Figure 3**). Prophets of modern science, such as Bernard le Bovier de Fontenelle (1657–1757) and Antoine-Nicolas De Condorcet (1743–94), envisioned science in terms of a theater stage, upon which temporal scientific and social progress occurs in changing historical stages, as in a theatrical drama (Olwig, 2002). The perspectival scenery on the stage floor thus represented the land surface, soil, and geological/topographical shape of a land (still described as 'scenery' in some modern geology texts), and the stage scenery upon it represented the flora, fauna, and cultural features of the land, whereas climate was located in the area above stage. This form of representation allows for a shift in the meaning of land from the place or *choros* of a polity (e.g., Scotland) to simply soil (e.g., wetland). Landscape could thereby be stripped of its link to social phenomena, and become a 'physical' or natural phenomenon, to be studied by natural scientists, or it could represent a synthesis of physical and cultural landscapes, illustrating the interaction between society, flora and fauna, geology, and climate (**Figure 4**).

Alexander von Humboldt (1769–1859), one of the most famous and widely read scientists of his time, is often regarded as a founder of modern geography (Martin and James, 1993). His

Figure 2 Illustration from Albrecht Dürer, *Underweyssung der Messung* (Nuremberg, 1538). The male draughtsman is using the techniques of cartographic surveying to create an illusion of three dimensional, bodily space, as represented by a woman, but also by the landscape scene seen through the window, and, in a circular way, the entire drawing, which itself is done in perspective.

Figure 3 This plate from Ortelius' late sixteenth-century Atlas, *Theater of the Lands of the World*, illustrates how a more perspectival and horizontal view of the world merged with a more top-down cartographic representation of a region.

success owed in part to his understanding of the value of landscape, together with cartography, chorography, and topography, as a means of studying, explicating, and popularizing natural history. Traveling the globe Humboldt was able to map and correlate physical and cultural geographical phenomena according to the chorographic and the scenic structure of landscape. He thereby became a founding figure in fields such as climatology, oceanology, and plant geography, while using landscape artists to capture and communicate the visual character of places. Humboldt, furthermore, was not blind to the importance of the more literary tradition of Herodotus and Strabo, writing:

> In classical antiquity the earliest historians made little attempt to separate the description of lands from the narration of events the scene of which was in the areas described. For a long time physical geography and history appear attractively intermingled.
>
> von Humboldt quoted by Sauer, 1969, orig. 1925

Humboldt's writings, in turn, inspired the young Charles Darwin to undertake his voyage with the Beagle (Livingstone, 1992). Darwin's resultant theory of evolution then charted the developmental stages of life itself in terms of an interaction between species and their landscape environment. Landscape was thus not so much a subject of particular study, as a metastructure, embodying chorography, topography, cartography, and cosmography, which was used to organize and transmit knowledge, embodying an entire cosmology as outlined in Humboldt's seminal *Cosmos: A Sketch of a Physical Description of the Universe* (Humboldt, 1849–58).

By the beginning of the nineteenth century most European states had developed a centralized government within relatively clear chorographically mapped boundaries, encompassing a relatively unified people speaking a national language (Tang, 2008). This was not the case for Germany, which was divided into a myriad of polities ranging from large kingdoms, such as Prussia, to small landscape polities (*Landschaft*), wherein mutually unintelligible German dialects often were spoken. Groups of German-speaking people were also dispersed across especially Eastern Europe, often in isolated enclaves. The nineteenth century was a time of European nation-state building, and it was especially in Germany that a geographical discipline was needed, especially in the schools, to help make sense of Germany as a nation in the world. Landscape, with its ability to unite chorography, topography, and cartographic spatial relations, played a particularly important role in this, at a time when German landscape art also flowered. This might explain why, during the nineteenth century, German geography dominated the discipline internationally, starting with Humboldt, and why landscape played such an important role in that geography. Toward the turn of the twentieth century, when geography became formalized as a university discipline, the subject was therefore layered into differing areas of specialized study within the physical and/or human sciences which, when linked together, could be seen to form an overarching, metatheoretical, physical, and cultural landscape totality. Thus, in 1925 when Carl Sauer was founding the new geography department at the University of California, Berkeley, he laid the department's groundwork with a magisterial survey of Western geography, dominated by German geographers, entitled *The Morphology of Landscape* (Sauer, 1969).

Figure 4 Thomas Hobbes' pioneering 1651 work on the constitution of the state in terms of its matter, form and power used a perspectival image of a landscape as commanded by the vision of the head of state.

The Collapse of Landscape's Disciplinary Hegemony

Geography is still divided into physical and human sections, reflecting the division between the physical and the cultural landscape within an overarching landscape totality, but this structure has become increasingly strained since World War II. Particularly Germany's ideological excesses led many non-German geographers to question the German foundations of their field.

The idea that the space inhabited by a nation should be demarcated within the chorographic space of map was, as noted, foundational to the nation-state. The post-Darwinian landscape paradigm suggested an interaction between this national space's physical landscape and its cultural landscape, changing in stages through time in a process of evolution according to the geographical doctrine of 'environmental determinism.' This interaction was foundational to nationalist *Blut und Boden* (German for 'blood and soil') ideology, which contended that the character of the human races developing within the national territory was determined by their historical interaction with the national natural environment, and which argued for the maintenance of national racial purity through the exclusion of non-German nationality, especially those deemed to be racially inferior. This same ideology suggested that superior national races would 'naturally,' according to evolutionary doctrine, grow and expand, and therefore need *Lebensraum* or 'living space' – a relational space with disastrous consequences for the peoples whose lands the Nazis sought to colonize. These ideologies, however, were by no means exclusive to the Germans; they were also used to justify imperial expansion by other nations. As an ideal model the interaction of environmental factors within the chorographic space of landscape was heuristically useful for the development of evolutionary theory, especially as an island in the Galapagos presented a neatly bounded chorographic space. But the German state could not be comprehended as a large-scale chorographic island. Landscape art created a perspectival illusion of a unified national space, and chorography provided a means of mapping that space. This space and its boundaries, however, were illusory and utopian, and could not be applied, without becoming dystopian, to an area made up of a large number of heterogeneous territories with varying histories, populations, languages, and religions, and its borderland areas of mixed nationalities.

Landscape in Post-World War II Geography

Nationalist ideological excesses led to a questioning of the discipline's established paradigm. A leading figure in the post-war revisioning of geography was Richard Hartshorne, an American who, among other things, sought to cleanse geography of landscape. Hartshorne had studied in Germany and his approach to geography was grounded in German ideas of science, or *Wissenshaft*, and geographical theory, particularly the 'chorology' of Alfred Hettner. *Wissenshaft* was a broad notion of science in which the rigor of the natural sciences was also sought brought to bear on social phenomena, thereby ostensibly creating objective knowledge. From this perspective the nationalistic ideological excesses owed to a misuse of objective science for subjective ideological purposes, and it would thus be possible to avoid ideological misuse through rigorous *Wissenshaft*. Hartshorne's solution, following this logic, was to retain the *chorographic* approach, but transform it from a descriptive graphic study to an empirical 'spatial' science concerned with regional study. Hartshorne's chorology would be devoted to the correlation of a diversity of data within areas demarcated by the geographer according to analytical, rather than historical, criteria. This was to be an 'idiographic' approach concerned with specific facts or processes occurring in unique places, as distinct from generalizable (nomothetic) laws (Hartshorne, 1939), thus preventing the formulation of causal laws, applicable anywhere, such as those of environmental determinism, *Blut und Boden*, or *Lebensraum*. Hartshorne's chorology excluded the study of landscape with its concern for the historical interaction of society and its environment in a given historically constituted area. In English, he further argued, landscape meant a scenic phenomenon, perceived as extending into infinity, not a bounded area, as in the German language and in German geography. Hartshorne's redefinition of geography suited his own personal interest in economic geography, and hence geography as basically a social science, rather than

an environmental science. Hartshorne became the leading theoretician of mainstream geography in the post-war English-speaking world (Hartshorne, 1959), but his focus on chorology was eventually eclipsed. This happened largely in the 1960s through a shift to the nomothetic study of spatial location in the tradition of Ptolemy's idea of geography as a locational science. By this time spatial science's own Nazi past had been forgotten (Rössler, 1989), and it was largely believed, furthermore, that 'objective' positivistic science, like *Wissenschaft*, would be a guarantee against ideological misuse. The rise of geography as spatial science thereby marginalized landscape as a paradigmatic framework for mainstream English language geography. This was not the case, however, with geography in Germany. In many continental European countries within the sphere of influence of traditional German geography, where rigorous *Wissenschaft* was believed to mitigate ideological misuse, the traditional German approach to landscape, combining the study of the physical and cultural landscape through time, remained important and contributed to the subsequent development of landscape ecology. The notion of landscape, however, took a very different turn with regard to the geography emanating from Berkeley under the stewardship of Sauer. Here, landscape developed along a different path from either geography as the science of space, or traditional German landscape geography.

The Berkeley School Landscape Studies

The critique of landscape mounted by Hartshorne did not fit the approach to landscape taken in practice by Sauer and his colleagues and students as they developed it over time (Sauer, 1969; Denevan and Mathewson, 2009). The reason is that Berkeley geography followed an interdisciplinary humanities-oriented approach more in the vein of *choros* (which he spelled *chore*) as understood by Herodotus and Strabo, than Ptolemy's chorography, which they criticized. Sauer's landscape geography was thus wedded to the study of history, anthropology, archaeology, and place, even if it included a concern with natural science and cartography. Sauer severely criticized environmental determinism and focused upon human agency in shaping differing landscapes, viewing landscape as a mental, social, and physical 'phenomenon.' The conference and book, *Man's Role in Changing the Face of the Earth*, gives a good impression of the breadth, depth, and impact of the 'Berkeley school' and its British counterpart in the school of H.C. Darby's historical geography (Thomas, 1956). Landscape geography, nevertheless, led a subaltern existence within a discipline increasingly dominated by various permutations of spatial social science. A new approach to landscape geography, however, did develop in the interstices between Berkeley geography and spatial science.

Landscape Perception

During the 1960s David Lowenthal, a geographer and historian who had studied both at Berkeley and with Hartshorne at the University of Wisconsin, effectively turned Hartshorne's critique of landscape on end. Sauer had defined the study of landscape in 'phenomenological' terms, meaning that landscape was not constituted analytically through science, but was rather, following Sauer, a 'phenomenon' that was perceived as a basic epistemological category by laymen. Lowenthal took this idea and applied it to Hartshorne's understanding of landscape as a scenic perception of the environment. Lowenthal argued that the study of ordinary people's environmental perceptions was necessary if geographers wished to understand human attitudes and behavior (Lowenthal, 1961). Similar approaches were taken at this time by humanities-oriented American geographers with an affiliation to Berkeley, such as Yi-Fu Tuan and J.B. Jackson, thus helping to generate a subfield concerned with the perception of landscape and the environment. At this time landscape, conceptualized largely as scenery, became an object of public management in nature conservation, cultural heritage, agricultural policy, and physical planning. The British geographer Jay Appleton, working in this context, combined the landscape perception approach with evolutionary biology, arguing that the frequent use of a dark foreground in landscape painting reflected 'man's' prehistory as hunters seeking 'refuge' in the shadows where they looked out over a landscape 'prospect' while stalking game (Appleton, 1975).

Appleton's return to a variant of Darwinian environmentalism sparked a counterreaction among some British geographers, notably Denis Cosgrove, with a more humanities-based and social constructionist orientation. Cosgrove argued that the perception of landscape in scenic terms as prospect was not an outcome of evolutionary biology, but rather of the Renaissance development of the techniques of perspectival representation. Focusing on Venice, he saw a vital connection between the rediscovery of Ptolemaic cartography, the surveying practices of land drainage and enclosure, the development of landscape architecture, and the ability to understand, appreciate, and transform the world as perspectival spatial scenery – an approach to landscape that jibed well with geography conceptualized as spatial science (Cosgrove, 1984). Working with Stephen Daniels, who took an art and gardening history perspective, and the Canadian-British geographer James Duncan, who applied a semiotic perspective, Cosgrove pioneered an approach, influential well beyond the bounds of geography, defining landscape principally as a form of visual representation (Cosgrove and Daniels, 1988).

For Cosgrove, the concept of landscape originated with perspectival representation through which people learned not only to perceive the world as landscape scenery, but also to shape it as landscape scenery, as in landscape architecture. This analysis, however, did not take into account that landscape existed as place and polity (*choros*), prior to its representation in perspectival space, and that it still was regarded this way in many areas of Europe, particularly in the Nordic region (Olwig, 2002; Jones and Olwig, 2008). This shortcoming does not mean that the scenic, perspectival, reconceptualization of landscape as space has no importance, but it makes it necessary to understand the substantive relationship between the way places conceptualized in terms of the older sense of landscape are socially and legally constituted, and governed, in terms of, for example, customary use rights to commons rather than property rights to enclosed land, or the governance of federated polities rather than that of a centralized state (Olwig, 2002; Cosgrove, 2004). The substantive approach to landscape has helped ameliorate French philosophy-inspired criticism of the visual bias in landscape studies (Jay, 1994), as well as critique

by those who see landscape as the embodiment of practice and labor (Ingold, 2000; Mitchell, 1996), or who are inspired by nonrepresentational geographical theory (Wylie, 2007).

Conclusion

Landscape has played a prominent role in geography as an overarching metaconcept through which the character, shape, and constitution of a land, understood as *choros*, has been understood. In one sense it has been perceived, in the tradition of Herodotus and Strabo, as the interlinked object of historical, ethnographical, and geographical narrative closely linked to the humanities. In another sense, it has been defined as a mode of scientific spatial representation encompassing cartographic chorography, topography, cosmography, and geography as a locational spatial science, and the scenic phenomena thus represented. The two senses, overlapping through time, have not just provided a framework for geographical study, but also a medium through which people in society have comprehended, for example, their place as commoners in a landscape polity, or as private property owners and citizens of a nation-state. This is evidenced today by the European Landscape Convention, as promulgated by the Council of Europe and ratified by most European states, which defines landscape as "an area, as perceived by people, whose character is the result of the action and interaction of natural and/or human factors." It calls for states "to recognise landscapes in law as an essential component of people's surroundings, an expression of the diversity of their shared cultural and natural heritage, and a foundation of their identity" (Europe, 2000). The Convention, in turn, has stimulated recent scholarly interest in the link between landscape, human rights, and democracy (Egoz et al., 2011; Jones and Stenseke, 2011). Landscape thus continues to be of relevance to geographers, and to society at large, even if its meaning has changed through time, resulting in an entangled complex concept that is as good to think 'with,' as 'about,' because the transformations of its meaning over time and space reflect the historical transformations of society and its geography.

See also: Cultural Geography; Environmental Determinism; Geographic Thought, History of; Geography: Overview; Historical Geography; Place in Geography; Regional Geography.

Bibliography

Appleton, J., 1975. The Experience of Landscape. Wiley, London.
Cosgrove, D., 1984. Social Formation and Symbolic Landscape. Croom Helm, London.
Cosgrove, D., February 19, 2004. Landscape and Landschaft lecture delivered at the "Spatial Turn in History" symposium German historical institute. GHI Bulletin 35, 57–71.
Cosgrove, D., Daniels, S. (Eds.), 1988. The Iconography of Landscape. Cambridge University Press, Cambridge.
Darby, H.C., 1962. The problem of geographical description. Transactions and Papers, publ. 30, The Institute of British Geographers 30.
Denevan, W.M., Mathewson, K. (Eds.), 2009. Carl Sauer on Culture and Landscape: Readings and Commentaries. Louisiana State University Press, Baton Rouge.
Egoz, S., Pungetti, G., Makhzoumi, J. (Eds.), 2011. The Right to Landscape: Contesting Landscape and Human Rights. Ashgate, Aldershot.
Europe, Council of, 2000. European Landscape Convention, Florence. CETS No. 176. Council of Europe, Strasbourg.
Hartshorne, R., 1939. The Nature of Geography. Association of American Geographers, Lancaster, PA.
Hartshorne, R., 1959. Perspective on the Nature of Geography. Rand McNally, Chicago.
Humboldt, A. von, 1849–58. Cosmos: A Sketch of a Physical Description of the Universe. Henry G. Bohn, London.
Ingold, T., 2000. The temporality of landscape. In: Ingold, T. (Ed.), The Perception of the Environment: Essays in Livelihood Dwelling and Skill. Routledge, London, pp. 189–218.
Jay, M., 1994. Downcast Eyes: The Denigration of Vision in Twentieth-Century French Thought. University of California Press, Berkeley.
Jones, M., Olwig, K.R. (Eds.), 2008. Nordic Landscapes: Region and Belonging on the Northern Edge of Europe. University of Minnesota Press, Minneapolis.
Jones, M., Stenseke, M. (Eds.), 2011. The European Landscape Convention: Challenges of Participation. Landscape Series 13. Springer, Dordrecht.
Livingstone, D.N., 1992. The Geographical Tradition: Episodes in the History of a Contested Enterprise. Blackwell, Oxford.
Lowenthal, D., 1961. Geography, experience and imagination: towards a geographical epistemology. Annals, Association of American Geographers 51 (3), 241–260.
Martin, G.J., James, P.E., 1993. All Possible Worlds: A History of Geographical Ideas. John Wiley, New York.
Merriam-Webster, 1996. Collegiate Dictionary. Merriam-Webster, Springfield, MA.
Mitchell, D., 1996. The Lie of the Land: Migrant Workers in the Californian Landscape. University of Minnesota Press, Minneapolis.
Olwig, K.R., 2002a. Landscape, Nature and the Body Politic. University of Wisconsin Press, Madison.
Olwig, K.R., 2002b. Landscape, place and the state of progress. In: Sack, R.D. (Ed.), Progress: Geographical Essays. Johns Hopkins University Press, Baltimore, pp. 22–60.
Olwig, K.R., 2008. Has "Geography" always been modern?: *choros*, (non)representation, performance, and the landscape. Environment and Planning A 40, 1843–1861.
Ortelius, A., 1595 (orig.) 1570. Theatrum Orbis Terrarum. Antwerp.
van Paassen, C., 1957. The Classical Tradition of Geography. J.B. Wolters, Groningen.
Ptolemy, C., 1991. The Geography. Dover, New York.
Rössler, M., 1989. Applied geography and area research in Nazi society: central place theory and planning, 1933–1945. Environment and Planning D: Society and Space 7, 419–431.
Sauer, C., 1969. Land and Life: A Selection from the Writings of Carl Ortwin Sauer. University of California Press, Berkely.
Tang, C., 2008. The Geographic Imagination of Modernity: Geography, Literature, and Philosophy in German Romanticism. Stanford University Press, Standford.
Thomas, W.L., 1956. Man's Role in Changing the Face of the Earth. University of Chicago Press, Chicago.
Warntz, W., 1989. Newton, the Newtonians, and the Geographia Generalis Varenii. Annals of the Association of American Geographers 79, 165–191.
Wylie, J., 2007. Landscape. Routledge, London.

Land Tenure

Parker Shipton and Daivi Rodima-Taylor, Boston University, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Land tenure, connecting to much-debated issues of human territoriality and sovereignty, can be approached at any level of scale from domestic to cosmic. As presented in this article, land tenure moves beyond the confines of conventional English legal terms like 'rights and duties' and political–economic ones like public or private 'property,' or collective or individual 'ownership.' A wider array of concepts, taking real access, attachment, and mobility into account, is needed to capture the new practices and institutions for recording land-use patterns that arise locally or are produced by diverse directed tenure reforms. Particularly in postauthoritarian settings, land property claims are becoming central in continuing struggles over belonging and identity. New forms of decollectivization as well as intensifying land grabbing for agri industries produce novel patterns of exclusion and debates over territory and jurisdiction. Where migration is concerned, and where persons absent as well as present can make land claims, scholars and practitioners alike encounter yet unstudied forms of tenure. Terms and translations – for people, things, and links between – are critical throughout.

Introduction: Struggles Over Land and Meaning

Land can refer to space or place, often denoting material substance deemed immovable, and sometimes connoting belonging. In some tongues, it takes several distinct words to translate, and corresponding terms may or may not include water, plants, animals, people, and built or dugout structures. 'Tenure' (from L. *tenere*, to hold) or landholding need not refer just to a particular form of legal, political, or economic regime (such as freehold, fiefhold, or collective village ownership) but can indicate human–land attachment in a broader sense, taking account of unofficial, illegal, or occasional use, and of persons living, dead, or unborn.

"Nearly all social struggles in Antiquity," claimed Max Weber, "were essentially for the ownership and use of land" (1988, p. 343). He might say the same for recent times. Particular regions have hallmark forms of land tension and strife, simplified and sharpened in mass media if not also in scholarly representation. Such are ethnoreligious struggles over autochthony in the Levant and Northern Ireland; forced transmigration in Indonesia; latifundio, minifundio, squatting, and private militias ('goon' squads) in Hispanic America; small-tribe survival struggles in Amazonia; landlord–tenant exploitation in South Asia; plantation oppression in the Caribbean and island southeast Asia; collectivization and decollectivization in Russia and China and states in their hegemonies; farmer–herder conflicts in the American west or the Sahel; and individual titling that threatens women, children, and the rural poor in the African tropics. But none of these regions monopolizes the tenure issue for which it is famed. Tasks for social scientists and humanists include finding particularities and exceptions, bringing to light indigenous practical solutions, and assessing the performance of authorities that claim jurisdiction or sovereignty.

Historical Background

Records of land claims and transfers have been dug up among some of the earliest known writings in Mesopotamia from 4000 to 5000 years ago, along with royal edicts canceling subjects' debts and reversing land forfeitures to creditors. Plato's and his student Aristotle's debate about the merits of public versus private landholding in the Aegean *polis* echoes two millennia later in European and American traditions. John Locke, Adam Smith, and Herbert Spencer are still widely cited by private property's upholders, while the young Jean-Jacques Rousseau (esp. 1965), Karl Marx, Pierre Proudhon, and (in America) Henry George are champions of its critics.

Evolutionist Political Philosophy

The spread of European imperial power and the westward expansion of Euro-American settlement in the late nineteenth and early twentieth centuries, accompanied as they were by wholesale expropriation or disenfranchisement of foragers, mobile herders, and swidden (slash-and-burn, long fallow) farmers, led sociologists and anthropologists into the study of land tenure as their disciplines coalesced. Jurist Henry Maine's vision of Roman property law's and human history's evolution from ascribed 'status' to achieved 'contract' was qualified by lawyer-anthropologist Lewis Henry Morgan's conclusion in *Ancient Society* (1877) that individual property was an inadequate basis for civilization's further progress. Societies moved, wrote Morgan, from *societas*, i.e., ties based on kinship and other particular ties, to *civitas*, ties based on territorially defined polity, but this did not always mean upward progress either. Karl Marx and Friedrich Engels, inspired by Rousseau and Morgan among others, envisaged history as a one-way flow through stages of prefeudal tribalism, feudalism, capitalism, socialism, and finally a future communism with abolition of private property and state government.

Evolutionist theories based on a largely imaginary prehistory and assumptions of unilinear change continue to resound in legal, political, and economic circles. But twentieth-century anthropology made it more apparent how tenure change can zigzag and backtrack, and how many different forms of tenure can be 'modern' at once, sometimes in close coexistence.

Persons, Things, and Systems

Early ethnography in Australia and Native America, and Bronislaw Malinowski's (1935) work on Trobriand Island gardening in Melanesia in the 1910s, revealed indigenous concepts in which human groups and spaces (however defined)

were deemed aspects of each other, often inextricable from spiritual beings or forces. Research in colonial Africa in the mid-twentieth century depicted how land tenure tied tightly into cognition, language, and translation; kinship and friendship; political control and hegemony; religion and ritual; and esthetics and myth (Meek, 1949; Biebuyck, 1963). Monarchs' and chiefs' control over land (as in Buganda or Ashantiland) appeared to be part and parcel with control over people, but their rights excluded alienation by sale and other rights often identified with ownership. Landholding was justified by clearing, residing on, or working on the land; by conquest; by membership in a kin or local group; by chiefly assignment and tribute; or a mix of these. Multiple rights of different groups to use land for different purposes, for example farming, grazing, and thatch, firewood, or clay collection, overlapped in complex ways not captured by terms like 'property' or 'ownership.' In 'acephalous' (politically uncentralized) societies like those of Tiv in Nigeria or Chagga in Tanzania, land territories were often found closely linked to genealogy of clans or lineages, shifting and stretching as these moved, grew, or shrank. Here and elsewhere in Africa, land was spoken of as an active interest of not only the living, but also the dead and unborn. To Audrey Richards, Lucy Mair, and other British-trained social scientists, land tenure appeared as part of complex, functioning cultural systems. Max Gluckman's southern Africa-based work on 'hierarchies of estates' over land, for instance of use, transfer, and administration, influenced a generation of tenure studies. Gluckman (1965) observed that property is not just about rights of people over things, but also about duties between people with respect to things.

Structure and Process of Inequity: Critical Rethinkings

In the 1960s and early 1970s, Latin American-inspired scholarship on dependency and 'modes of production,' in a Marxist vein, criticized the exploitation of rural and tropical peripheries by metropolitan capitals by agribusiness or plantation owners and through the co-optation of comprador (buyer) elites. South America, the continent with the longest history of European colonialism, had, some observed, the most uneven distribution of landholdings (Orlove and Custred, 1980). Parisian Marxist structuralists in the late 1970s revised the view that capitalism eroded all other forms of tenure and exchange (quasifeudalism, family farm holding, etc.), contending instead that it encapsulated and preserved these for its own profit.

In the 1980s and 1990s, social scientists with political–economic interest turned their attention to attempts by national regimes to monopolize power over land relations, particularly in Africa. State attempts to nationalize, privatize, or socialize landholding now looked like raw or subtle power grabs (Downs and Reyna, 1988; LeBris et al., 1991). Large-scale resettlement efforts, as in Ethiopia or Indonesia, seemed to some analysts like attempts to remove or disorient potential dissidents. Attempts to 'villagize' scattered farmers or sedentarize mobile herders appeared as attempts more easily to tax, conscript, and regulate them (Scott, 1998). Economic anthropologists devoting new attention to farmer perspectives about contingency and safety found that consolidating scattered holdings into larger blocks, in the name of efficiency or mechanized monocropping, could stymie prudent efforts of family farmers to maintain parcels on varied lands to reduce their exposure to risks like plagues, floods, or droughts. 'Family and community dynamics and their agro-economic rationales came under closer scrutiny (McCay and Acheson, 1987; Netting, 1993). Links between material and symbolic concerns became more apparent (Goheen and Shipton, 1992; Shipton, 1994).' Postmodernist and poststructuralist scholars, meanwhile, eschewed the notion of timeless tenure 'systems,' preferring to see complexity, indeterminacy, negotiability, contestability, and historical contingency, in human interaction, and to draw attention to questions of differential power (Hann, 1998; von Benda-Beckmann et al., 2006). In town and country, they and others criticized tenure regimes based on race, class, and gender inequities (Lawrence and Low, 1990; LeBris et al., 1991).

Toward the turn of the millennium, a plurality of newly empowered indigenous voices in southern countries helped force scholars, administrators, and jurists concerned with land tenure everywhere to reassess their most basic terminology and philosophical premises (Agarwal, 1994). Assumptions of human dominion over land, of individualistic rights and duties, and of the sanctity of arbitrarily drawn borders and boundaries that blocked human movement all came under criticism.

Metaphor and Rhetoric

Human communication about land draws heavily on metaphoric imagery from the human body (headwaters, heartlands, fertility, etc.), kinship (motherland, fatherland), domicile (gateway to Asia), and other human-centered sources. Such metaphorizing imbues poetry and politics, while imagery of 'grounding' or 'terra firma' imbues legal and scientific discourse. English terms like 'property,' 'rights,' or 'ownership' sometimes carry culture-bound overtones of civilization, justice, or morality. To speak of land as a commodity denies psychological attachments to space and place. To speak of land as a 'resource' implies it exists to serve human use. Between people who assume that humans can own land and master it, and those who assume on the contrary that humans belong to the land and must respect it and conform to its requirements, communication easily goes awry.

People who derive their livelihood from land in different ways perceive and communicate about it in different geometric and geological idioms. Whether in the Andes, the Sahel, or Lapland, transhumant herders who must focus on linear access routes tend to clash with farmers who focus on fields along the way, when animals eat or trample crops. Seasonal and tidal wetlands and mudflats (as in the Niger delta, in the South Sudan's Sudd or along the Senegal River) are frequently contested by farmer-fishers, ecologists, and developers using different terminology.

Land and Tenure Reform

While land reform, in English, conventionally refers to redistribution of holdings, land *tenure* reform refers to changes in the more basic rules and understandings by which land is occupied, used, or transferred. Among many kinds of tenure reform are proclamation of sovereignty or eminent domain, implementation of titling (i.e., registration), and change in rules about transfer (loan, rental, sale, inheritance, etc.). Land reforms have a long history (Tuma, 1965). Dozens of

twentieth-century national tenure reforms sought to individualize or collectivize landholding.

Their histories are full of ironies and surprises. Frequently, reforms aimed at improving distribution of holdings have ended up worsening it, whether because of unequal information, implementer bias, deceit, or reneged deals. Moreover, reforms aimed at boosting production have often reduced it instead, for instance because new owners or occupants lacked farming experience or acquired the land only for speculation. Land tenure is so tightly bound up with kinship, religion, and other aspects of culture that a newly imposed tenure regime seldom expunges what was there before (Shipton, 1994). More often, it just crowds new procedures together with old, allowing the wealthy, powerful, astute, or well connected to gain at others' expense by manipulating confused or conflicting principles.

There are exceptional cases of successful land and tenure reform. One celebrated example is Taiwan's land redistribution in 1953, swiftly and efficiently implemented after a period of Japanese occupation had weakened local resistance. Attempts to emulate it elsewhere have seldom succeeded.

Bureaucratic Simplification

Bureaucratic administration, tending to push toward abstraction and universalism in its publicly professed goals, simplifies complex realities of human culture and ecology, sometimes beyond recognition, to make people or spaces susceptible to manipulation for taxation, military conscription, etc. Empty spaces on maps, or designations like 'vacant' or 'underproductive,' help justify land grabs, disfavoring swidden farmers, transhumant herders, foragers, and other occasional users. Titles issued to individuals ignore claims of dependents. But subjected people may thwart imposed tenure regimes, for instance by continuing to transfer fields without informing state authorities, as happened in Kenyan and other African cases of titling reforms over the second half of the twentieth century. Real and paper tenure diverge further.

Nowhere do real people live their lives consistently by capitalist or socialist principles, whatever they may think they do. In agropastoral Africa, for instance, land farmed by families or individuals may be expected to be opened for grazing by the animals of some broader public after harvest. Property shifts between public and private or remains subject to multiple tenure 'systems' simultaneously.

Collectivization and Villagization

Collective ownership and production have been part of utopian visions since Plato's *Republic* or before, and many different attempts have been made to bring these into reality in monasteries or other single- or mixed-sex religious or secular communes, sometimes modeled upon families. Mexico's *ejidos* after the 1910 revolution, and Israel's kibbutzim more recently, have been closely watched by admirers and critics alike. Marxist-inspired attempts to collectivize landholding under state aegis in the Soviet Union between 1917 and 1931, and in China after 1949, have been emulated in various parts of the world, notably eastern Europe, Vietnam and Cambodia (Kampuchea), Cuba, Tanzania, Ethiopia, and Lusophone African countries including Angola and Mozambique.

While late twentieth-century Chinese history has shown that some forms of state-organized collective farming can raise productivity, improve distribution of food and other vital resources, or both, if starting from a feudal or quasifeudal baseline, people in few parts of the world have been able to emulate socialist tenure reforms with lasting success. Commonly, work-incentive problems get in the way, and local populations chafe under centralized bureaucratic control of their lives.

Notable failures have been attempts to concentrate rural people into villages where the land could not sustain them, as in Tanzania's compulsory villagization from 1973 to 1976 (an experiment duplicated in Ethiopia a few years later). Similarly, savannah borehole well projects, for instance in Botswana or across the Sahel, have led to overconcentration of animals and overgrazing. Two policy conclusions on herders and farmers are clear enough: (1) people who live spread out, or mobile, usually have good reasons for wanting to stay that way; and (2) good intentions, in land tenure reforms, are about as likely to produce harmful as helpful results.

Titling and Individuation

Another kind of tenure reform is the titling or registration of individual claims for the first time. 'Private' property is often used to mean individual property, but the two need not be the same. State and local governments and landholding companies have several ways of registering land titles and transactions. Among the most commonly applied in former British colonies is the Torrens system, devised in Australia and New Zealand in the 1860s, where it allowed grid allotments for new settlers.

Where people already live, titling land in the names of individuals, under state guarantee, does not necessarily increase security of tenure as often assumed. It may decrease it. Instructive was the General Allotment Act (or Dawes Act) of 1887, by which the US government allowed for the awarding of individual land titles to Native Americans (or First Peoples). Between 1887 and 1932, about two-thirds of Native American lands were lost to immigrant settlers and speculators, who were not slow to exclude its former inhabitants and users.

British, French, and Belgian colonial administrations appear to have paid this story little heed in launching land titling programs in their colonies. Early twentieth-century titling initiatives were expanded to nationwide scale in the Kenyan model established under the Swynnerton Plan in 1954 and passed into law in 1957. That model, which still continues, was being emulated elsewhere in Africa at the turn of the millennium.

At least as often as not, land titling in new states and perhaps old seems to speed the concentration of holdings into fewer hands. Individuals newly freed of social constraints on selling, mortgaging, or pledging erstwhile family or clan land do not always make wise decisions. Speculators and accumulators use urban capital and rural off-farm incomes to buy out disadvantaged rural people in periods of drought, plague, or unfavorable prices.

The mortgage, a deadlined pledge linking land or other property to credit, plays a key part in this process (Shipton, 2009). Borrowers who fail to repay by a specified date forfeit their land to the lender or lender's agent. The mortgage appears regularly in histories of rural dispossession, not least in England, where law about it has changed many times, and in North America. While seemingly vital to urban and industrial economic life, in rural

areas mortgage loans are more treacherous to farming people, particularly where rainfall, infrastructure, and crop markets are unreliable. Although rich like poor may lose lands, mortgage foreclosure contributes heavily to class stratification. Inland tropical settings often lack sufficient industrial or commercial opportunity to absorb newly landless or land-poor people. Even where land titling is accepted, mortgaging is morally debated in many cultural and legal traditions. *Shari'a*, the sacred law of Islam, forbids mortgaging as sin.

When, and whom, is titling likely to help? Threatened indigenous minorities frequently seek titles to defend their home territories – and by extension, their livelihood and way of life – from more numerous and powerful outsiders who threaten to dispossess or displace them. These may include fortune seekers like ranchers and rubber tappers, as in Amazonia or Malaysia; homesteaders and land speculators, as in the American west; planters, as in the Caribbean; or politicians and their favorites, as in equatorial Africa.

Land titling is never a politically neutral process of recording information. Usually someone gains and someone loses. Gainers are often richer, better-connected, and better-schooled people; erstwhile land borrowers or land clients, who may become owners; and people who emigrate or seek to speculate in land without having to occupy or farm it. Losers include ancillary rightholders and dependents, and members of fast-reproducing families. Just as rural people have circumvented or resisted state directives to villagize or collectivize their landholding under socialist policies, as in Tanzania or Ethiopia, they have reasserted the rights of collectivities – clans, lineages, age sets, or voluntary associations – under individualistic capitalist ones, as in Kenya or Côte d'Ivoire in the late twentieth and early twenty-first centuries.

Land Rights and Tenure Security in Postsocialist Settings

Recent postsocialist decollectivization and privatization reforms have profoundly affected landholding patterns and rural institutions. Emerging from the need to dismantle the collective institutions imposed upon rural production in the socialist era, mass scale decollectivization efforts have had uneven and ambiguous results. While in some cases encouraging entrepreneurial initiative and small-scale farming, they have also led to legal and administrative ambiguities and a proliferation of alternative institutional arrangements with their own specific histories. The effects of the privatization reforms on the rise of viable smallholder production are particularly unclear and contextually diverse. In Russia, for example, privately held family farms have mostly failed to take off and former collective farms have continued operating as privatized joint-stock operations, characterized by complex and overlapping rental contracts and managerial reliance on the former elites (Miller and Heady, 2003; Lerman et al., 2004).

Experience from other former Soviet republics indicates similar continuities that build on pre-socialist traditions and cultural norms. In Estonia, popular aversion toward large-scale landholding had developed both in response to the earlier feudal estates of foreign overlords as well as more recent soviet collectivization reforms of the 1940s (Abrahams and Kahk, 1994). As all three Baltic republics – Estonia, Latvia, and Lithuania – had undergone forcible nationalization of agricultural lands during the socialist period, postsocialist mass restitution of land restored the family farms. Many of these still face severe structural constraints including unreliable markets, limited government subsidies, and poor access to agricultural credit (the last often a mixed blessing). In postsocialist contexts, land markets are often functioning poorly because of bureaucratic barriers and delays in legislative changes. In some cases, however, family farms have been relatively successful in surviving the massive tenure reforms. Radical socialist collectivization efforts in China like Mao's expansive 'people's communes' failed to eliminate households as farming units, and after Deng Xiaoping's decollectivization reforms of the 1970s household-based landholding made its return through lease contracts. Novel market opportunities in postsocialist China have helped raise smallholder production and local entrepreneurial initiative (Zhang and Donaldson, 2010) and fostered the somewhat paradoxical coexistence of kin-based moral economy and sharing with commercialized family farms (Brandstädter, 2003).

Both socialist and postsocialist land tenure reforms have significantly affected ecological systems and biodiversity. Socialist collectivization has brought along forcible sedentarization of pastoralists in many areas of Siberia, Mongolia, and elsewhere. The migration and mobility patterns of herder communities of the Soviet northeast have been profoundly altered by enclosures of pastures and establishment of stationary farming complexes, often turning the pastoralist subsistence economy into attempts at large-scale ranching for reindeer and other animals. The postsocialist decollectivization reforms have imposed their own organizational patterns for managing privatized herds, paradoxically reinforcing informal arrangements and alternative institutions for resource management (see Stammler and Ventsel, 2003 for a discussion of privatization reforms among Nentsy and Dolgan reindeer herders; Upton, 2012 for an analysis of endogenous and donor-initiated institutions of pastoralist management of Mongolia's Gobi region). The institutional changes brought along by the decollectivization reforms have also been associated with increasing abandonment of agricultural land. A comparative study of land abandonment in the first decade of postsocialist transition in several East European countries showed that the abandonment rates were lowest in the areas with least change in land use institutions from the socialist era, whereas countries with many new institutional regulations and delays in harmonizing legislative frameworks suffered from the highest rates of abandonment – for example, in Latvia, some 42% of all agricultural land in use in 1990 had been abandoned by 2000 (Prischepov et al., 2012: p. 1). The loss of crop and forest cover frequently affects first more marginal communities and ecosystems and can be associated with decreasing incomes from cultivation and a growing importance of nonfarm activities, as evidenced in Albania (Müller and Munroe, 2008).

In many postsocialist areas, the imposition of Western property models like private titling and land registration has failed to promote security of tenure or facilitate agricultural development. The rights-based models that seek land restitution to the former owners have failed to consider the multiple overarching entitlements and nodes of social control that characterize the lived experience of socialist and postsocialist agrarian communities (Humphrey, 1983). In the centrally planned economy, land property was governed by

administrative institutions and decrees rather than juridical ones, and it comprised an intricate mix of individual and collective obligations and entitlements that were not directly handled through the legal system (Hann, 2003; Verdery, 2003). Promoting exclusive land rights, postsocialist tenure reforms have not accounted for the intermediate layers of entitlement and social control between the landholder and the state that were present in socialist property systems, thereby overlooking actual property claims and mechanisms of their legitimation. Basic land use distinctions, like one between wet rice fields and upland rain-fed land in Vietnam, can involve different norms and modes of allocation, but such distinctions do not always survive tenure reforms (Sikor, 2006). Postsocialist landholding, mixing older and newer kinds of entitlements and collectivities, has been vulnerable to easy manipulation by different actors involved, complicating land restitution processes. Land property can thus not only become an important part of forging new social identities and political alliances, but at the same time inspire new kinds of competition and conflict.

The Contested Formalization of Land Claims

Particularly in postauthoritarian settings, land property claims are becoming central in continuing struggles over belonging, inclusion, and identity. Liberalization and decentralization reforms have brought new attention to issues of autochthony and tradition, highlighting the role of local forms of organization and authority in contestations over belonging and resource access (Geschiere, 2009). Competing public authorities strive to gain legitimacy by familiar idioms and vocabularies from local traditions and national bureaus, diversifying the rights and duties at different levels and scales as they variously attempt to formalize and regulate land matters locally and regionally (Lund, 2008). Hence, in Africa, south of the Sahara, a trend of 'informal formalization' of land rights, where new practices, institutions, and techniques for recording land rights and transactions arise locally, in parallel to the official ones of the state (Benjaminsen and Lund, 2003). These can involve novel-written contracts and testimonies and public oral witnessing for property claims well recognized yet outside the formal legal system, as shown by Lavigne Delville (2002) for francophone rural Africa. While creative, these processes have resulted in some fragmented land regimes, sometimes with seeming internal contradictions. Arguably, though, authority and property rights bolster each other in circular fashion: institutions or regimes that confer property rights sometimes seem, just by doing so, to gain an aura of legitimacy in the public eye (Sikor and Lund, 2009).

Displacement, Agrarian Encroachments, and Land Grabbing

Large numbers of people sometimes need to resettle as a result of development programs like irrigation schemes and hydroelectric dams, of military travel and war, and of environmental hardships like drought, or like chemical spills or radioactive contamination. Late twentieth-century anthropological studies of resettlement, notably by Elizabeth Colson, Thayer Scudder, and collaborators among Gwembe Tonga in Zambia (formerly Northern Rhodesia), showed how resettled people adjust to extreme stress and trauma in patterned and predictable phases. Among their problems is shaken confidence in whatever leaders are seen to have allowed the misfortune to happen. Helping or allowing communities or neighborhoods to move intact, insofar as their members wish, seems to mitigate the ill effects by keeping at least their social 'landscape' familiar. But no kind of human 'development' intervention is harder to orchestrate, or more prone to produce catastrophe, than forcing resettlement.

Many contemporary 'foraging' peoples reside in difficult environments not of their choosing, for instance in southern and central Africa and northern Canada, where 'race' discrimination has been a factor. Foraging, swidden farming, and mobile herding peoples worldwide suffered a varyingly steady diminution of their lands, and sometimes severe dislocation, from before the eighteenth until the late twentieth centuries, often in episodes where racism played a part (Ingold et al., 1988). The history of Native American foraging and farming peoples after Columbus – a story of military conquest, deceit, misrepresentation of consent, and treaty reneging by Euro-Americans – hit one of its low points in President Andrew Jackson's Indian Removal Act of 1830, which forced Cherokee, Seminole, and other settled agricultural Native Americans from their southeastern homes to Oklahoma, at the loss of thousands of lives. In another kind of extreme case, Australian law did not recognize the rights of indigenous (or Aboriginal) peoples to land until 1976. In the twentieth century, swidden farmers in tropical and subtropical zones worldwide faced serious threats to their livelihoods under encroachment from sedentary farmers, agribusinesses and other concerns, and under their own population growth. Loss of land for these peoples, as for foragers and pastoralists, has often meant disruption of culture and society in myriad, sometimes indirect ways.

Land grabbing for large-scale biofuel and export food cultivation has dramatically escalated in many parts of the world in recent years. Large monoculture plantations funded by foreign investment are threatening smallholder food security particularly in Africa, a continent often erroneously seen as a primary reservoir of underutilized land. The recent expansion in the production of biofuels, related to globally rising food and energy costs and growing demand for alternative energy sources, is adversely affecting African ecosystems and biodiversity, and weakening the livelihoods of poor and marginal groups (Matondi, Havnevik, and Beyene, 2011). Because of land tenure insecurity and complex and overlapping land rights, local communities are especially vulnerable to the lack of transparency in large-scale land acquisition contracts that often marginalize local populations. Local-level research is needed to evaluate the political, economic, and environmental impacts of the rapidly escalating land grabbing on poor smallholder communities worldwide.

Special Issues in Irrigated and Flood-Prone Lands

Irrigated lands sustain some of the world's highest rural population densities. But nowhere is control over land more hotly political than in irrigated areas, where central authorities commonly seek to dictate settlement, land and water allocation, inheritance, etc. Where flood control is a main concern, as

in much of irrigated China and Bangladesh, bureaucratic polities have tended to be highly centralized and authoritarian. Indeed, need for coordinating large numbers of people for flood control along rivers, and not just for irrigation, was among the main reasons why many such centralized, authoritarian state regimes arose as early as they did in Mesopotamia, Egypt, China, the Indus Valley, and elsewhere. Where river flooding poses less threat, as over much of agrarian Africa, regimes attempting to emulate highly centralized, authoritarian models have tended to enjoy less popular support, often proving fragile and short-lived – particularly where ethnic and linguistic mixing compounds the adjustment problems of living in close, compulsory synchrony.

Irrigation schemes frequently involve inequalities in access to water. The privilege of upstream location in the trench network, implying superior control over water, often coincides, as cause or effect, with personal or familial wealth and power of other kinds. The centripetal draw of high farm yields tempting farmers into irrigation schemes is sometimes counteracted by the centrifugal push of struggles over control over community and family life (as for instance in the Jahally Pachar rice scheme in The Gambia). Other problems common in irrigated areas include canal silting, and sluice gate aging and failure, difficulties of drainage and, where drainage fails, soil salinization – any of which can force resettlement.

Advocacy and Practice

The activist international defense of land rights of threatened indigenous minorities, particularly in sparsely settled regions pressed for new resource exploitation by outsiders (Amazonia, Alaska, central Africa, northern Australia, Irian Jaya, and elsewhere), crystallized in the Copenhagen-based International Work Group for Indigenous Affairs in 1968, London-based Survival International, in 1969, and Cambridge, Massachusetts-based Cultural Survival, Inc., in 1972, as well as other human rights defense groups. A worldwide movement of indigenous human rights groups, mutually emboldening, has raised public consciousness and achieved changes in national legislation like the Alaska Native Claims Settlement Act in 1971, Australia's Aboriginal Land Rights (Northern Territories) Act in 1976, and from Canada, the demarcation of Nunavut at the end of the century as a largely autonomous territory for First Peoples. Still, the International Labor Office's Convention 107, claiming that 'the right of ownership, collective or individual, of the members of the populations concerned over the lands which these populations traditionally occupy shall be recognized,' and the World Bank's tribal peoples policy, issued in 1982, have mainly been honored in the breach (Bodley, 1999). Participating in the active design of constructive tenure processes along lines more faithful to indigenous systems and aspirations, and showing respect to indigenous terminology, remains a challenge. Still, documents like the United Nations Declaration of the Rights of Indigenous People, passed in 2007, indicate that more principles of fairness, of respect, and (in this case) of resistance to colonization and dispossession have been making it into official policy at high levels and wide scales (United Nations, 2008).

New Frontiers

As social scientists and humanists look beyond simplistic dichotomies like public and private, collective and individual, or rights and duties, they perceive many yet unstudied forms of tenure. These may offer important solutions to work-incentive problems typical of socialism, on the one hand, and maldistribution problems typical of capitalism on the other. Other terms, concepts, and policy frameworks are needed that respect access, stewardship, and sentiments of attachment, that allow for possibilities of local autonomy and that permit human population flows between territories.

Shifting attention between micro- and macroscale, geographers, archaeologists, and others have sought to connect their work on the ground with findings newly accessible from 'remote sensing,' of landforms and land occupation and use, from air and space. Rural land tenure remains better studied than urban and periurban, and ground tenure better than water tenure. Border and boundary issues are land tenure issues on a large scale. Problems of land quality and distribution can be addressed not only by altering tenure regimes within nations, but also by allowing people to move between nations or continents more freely. Land tenure is not just about getting in or staying, but also about getting out.

Human territoriality and accommodation, if much studied, remain open for speculation on causes (Malmberg, 1980). A challenge for social and biological scientists alike is to connect the understandings about territoriality observable from human action on the land, with understandings becoming obtainable through ethology and the electrochemical study of the human brain. Human biology has yet produced few if any conclusive findings about territorial instincts in humans, but enough species exhibit recognizable social and cultural patterning in their territorial relations to make it seem unlikely that human spatial perceptions and attachments result from uniquely human social or cultural processes alone.

Meanwhile, as more languages have become involved in the academic study of land and tenure, the reexamination of the most familiar, but deeply cultural and potentially exotic, concepts and categories – like 'land,' 'property,' and 'rights' – has continued.

See also: Agriculture, Economics of; Irrigation Societies; Property Rights, History of; Property: Legal Aspects; Rights: Legal Aspects; Rural Sociology.

Bibliography

Abrahams, R., Kahk, J., 1994. Barons and Farmers: Continuity and Transformation in Rural Estonia, 1816–1994. Faculty of Arts, University of Gothenburg, Sweden.

Agarwal, B., 1994. A Field of One's Own: Gender and Land Rights in South Asia. Cambridge University Press, Cambridge, UK.

Benjaminsen, T., Lund, C. (Eds.), 2003. Securing Land Rights in Africa. Frank Cass, London.

Biebuyck, D. (Ed.), 1963. African Agrarian Systems: Studies Presented and Discussed. Oxford University Press, Oxford for the International African Institute, London.

Bodley, J.H., 1999. Victims of Progress, fourth ed. Mayfield Publishing, Mountain View, CA.

Brandstädter, S., 2003. The moral economy of kinship and property in southern China. In: Hann, C., et al. (Eds.), The Postsocialist Agrarian Question. LIT Verlag, Münster, pp. 419–440.

von Benda-Beckmann, F.K., von Benda-Beckmann, K., Wiber, M.G. (Eds.), 2006. Changing Properties of Property. Berghahn Books, Oxford and New York.
Downs, R.E., Reyna, S.P. (Eds.), 1988. Land and Society in Contemporary Africa. University Press of New England, Hanover, NH.
Geschiere, P., 2009. The Perils of Belonging: Autochthony, Citizenship, and Exclusion in Africa and Europe. University of Chicago Press, Chicago.
Gluckman, M., 1965. The Ideas in Barotse Jurisprudence. Yale University Press, New Haven, CT.
Goheen, M., Shipton, P. (Eds.), 1992. Rights over Land: Categories and Controversies. Africa 62 (3), special issue.
Hann, C., et al. (Eds.), 2003. The Postsocialist Agrarian Question. LIT Verlag, Munster.
Hann, C.M. (Ed.), 1998. Property Relations: Renewing the Anthropological Tradition. Cambridge University Press, Cambridge, UK.
Humphrey, C., 1983. Karl Marx Collective: Economy, Society and Religion in a Siberian Collective Farm. Cambridge University Press, Cambridge, UK.
Ingold, T., Riches, D., Woodburn, J. (Eds.), 1988. Hunters and Gatherers: Property, Power and Ideology. Berg, Oxford, UK.
Lavigne Delville, P., 2002. When farmers use pieces of paper to record their land transactions in francophone rural Africa: insights into the dynamics of institutional innovation. The European Journal of Development Research 14 (2), 89–108.
Lawrence, D.L., Low, S.M., 1990. The built environment and spatial form. Annual Review of Anthropology 19, 453–505.
LeBris, E., Le Roy, E., Mathieu, P. (Eds.), 1991. L'Appropriation de la Terre en Afrique Noire: Manual d'Analyse de Décision et de Question Foncières. Editions Karthala, Paris.
Lerman, Z., Csaki, C., Feder, G., 2004. Agriculture in Transition: Land Policies and Evolving Farm Structures in Post-Soviet Countries. Lexington Books, Lanham, MD.
Lund, C., 2008. Local Politics and the Dynamics of Property in Africa. Cambridge University Press, New York.
Malinowski, B., 1935. Coral Gardens and Their Magic: A Study of the Methods of Tilling the Soil and of Agricultural Sites in the Trobriand Islands. George Allen and Unwin, London.
Malmberg, T., 1980. Human Territoriality: Survey of Behavioural Territories in Man with Preliminary Analysis and Discussion of Meaning. Mouton, The Hague, The Netherlands.
Matondi, P., Havnevik, K., Beyene, A., 2011. Biofuels, Land Grabbing and Food Security in Africa. Zed Books, London, New York.
McCay, B.M., Acheson, J.M. (Eds.), 1987. The Question of the Commons: The Culture and Ecology of Communal Resources. University of Arizona Press, Tucson, AZ.
Meek, C.K., 1949. Land Law and Custom in the Colonies, second ed. Oxford University Press, London.
Miller, L., Heady, P., 2003. Cooperation, power and community: economy and ideology in the Russian countryside. In: Hann, C., et al. (Eds.), The Postsocialist Agrarian Question. LIT Verlag, Münster, pp. 257–292.
Morgan, L.H., 1877. Ancient Society. Henry Holt, New York.
Müller, D., Munroe, D., 2008. Changing rural landscapes in Albania: cropland abandonment and forest clearing in the postsocialist transition. Annals of the Association of American Geographers 98 (4), 855–876.
Netting, R., 1993. Smallholders, Householders: Farm Families and the Ecology of Intensive, Sustainable Agriculture. Stanford University Press, Stanford, CA.
Orlove, B.S., Custred, G. (Eds.), 1980. Land and Power in Latin America: Agrarian Economies and Social Processes in the Andes. Holmes and Meier, New York.
Prischepov, A., et al., 2012. Effects of institutional changes on land use. Environmental Research Letters 7.
Rousseau, J.-J., 1965. Discours sur l'Origine et les Fondements de l'Inégalité parmi les Hommes (Essay on the Origin and Foundations of Human Inequality). Gallimard, Paris.
Scott, J.C., 1998. Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed. Yale University Press, New Haven, CT.
Shipton, P., 1994. Land and culture in tropical Africa: soils, symbols, and the metaphysics of the mundane. Annual Review of Anthropology 23, 347–377.
Shipton, P., 2009. Mortgaging the Ancestors: Ideologies of Attachment in Africa. Yale University Press, New Haven, CT.
Sikor, T., 2006. Politics of rural land registration in post-socialist societies: contested titling in villages of Northwest Vietnam. Land Use Policy 23, 617–628.
Sikor, T., Lund, C. (Eds.), 2009. The Politics of Possession: Property, Authority, and Access to Natural Resources. Wiley-Blackwell, Oxford, UK.
Stammler, F., Ventsel, A., 2003. Between neo-liberalism and dirigisme: approaches to reindeer herding in Yamal and Sakha. In: Hann, C., et al. (Eds.), The Postsocialist Agrarian Question. LIT Verlag, Münster, pp. 321–362.
Tuma, E.H., 1965. Twenty-six Centuries of Agrarian Reform: A Comparative Analysis. University of California Press, Berkeley, CA.
United Nations, 2008. United Nations Declaration on the Rights of Indigenous Peoples.
Upton, C., 2012. Adaptive capacity and institutional evolution in contemporary pastoral societies. Applied Geography 33, 135–141.
Verdery, K., 2003. The Vanishing Hectare: Property and Value in Postsocialist Transylvania. Cornell University Press, Ithaca.
Weber, M., 1988. The Agrarian Sociology of Ancient Civilizations (R.I. Frank, Trans.). Verso, London.
Zhang, Q., Donaldson, J., 2010. From peasants to farmers: peasant differentiation, labor regimes, and land-rights institutions in China's agrarian transition. Politics & Society 38 (4), 458–489.

Land-Use Indicators

Karl-Heinz Erb, Institute of Social Ecology Vienna, Alpen-Adria Universität Klagenfurt, Vienna, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Land use denotes the sum of human activities and arrangements aimed at harnessing terrestrial ecosystem services. Land use is a central element of society–nature interactions and a pervasive driver of global environmental change. Metrics for land use refer to two aspects: area extent of certain land-use types and the intensity of land use within a certain land-use type. Land-use intensity encompasses metrics for inputs to the land-based production system, for output for the production systems, and system-level changes induced by land use that impact upon ecological and social systems (e.g., impacts on ecosystem parameters). This article presents different prominent biophysical indicators for land use, and discusses their reference systems (production vs consumption, spatially explicit vs administrative units) as well as their position in the drivers-pressures-state-impact-response (DPSIR) indicator system framework.

Land use denotes the sum of human activities and arrangements aimed at harnessing services provided by terrestrial ecosystems. Land use is a central element of society–nature interactions and a pervasive driver of global environmental change (Turner et al., 2007). One of the central functions of land use is the provision of resources to society, an essential and not substitutable resource for many purposes. This includes, in particular, the provision of food, feed, fuel, fibers, and also clean water and air (Millennium Ecosystem Assessment, 2005). Besides these provisioning functions, land provides many other services to society: it hosts infrastructure, absorbs wastes and emissions, and represents a source of recreational, esthetical, spiritual, and cultural values (Millennium Ecosystem Assessment, 2005). But land not only provides services to human society, it also represents the basis of terrestrial biodiversity (e.g., by providing habitats and gene reserves for plants, animals, and micro-organisms, above and below ground). Furthermore, terrestrial ecosystems are major components of the global climate systems and strongly connect to biogeochemical cycles of, e.g., carbon, nitrogen, phosphorous, and water. By using the land, human societies alter ecological pattern and processes in order to maximize the output of certain services, often at the expense of others (Foley et al., 2005). Land use directly and indirectly relates to many of the grand sustainability challenges (Reid et al., 2010) and is at the center of the three 'Rio conventions' by the United Nations (the United Nations Convention on Biological Diversity, CBD, the United Nations Convention to Combat Desertification, UNCCD, and the United Nations Framework Convention on Climate Change, UNFCCC). Land use is also a central research theme of Social Ecology (*see* Social Ecology).

Land is a finite resource (the terrestrial ice-free surface amounts to approximately 130 Mio km^2), and more than three quarters of it are currently used by humans (Sanderson et al., 2002), with more or less strong effects on ecosystem properties, depending on the type and intensity of use. One of the most ostensible effects of land use are land-cover changes. Land cover denotes the observed biological and physical attributes of Earth's terrestrial surface, such as vegetation or human-made artifacts (*see* Land Change: The Merger of Land Cover and Land use Dynamics).

Metrics of Land Use

Indicators are symbolic representations designed to communicate a property or trend in a system that is inaccessible to direct description due to its internal complexity (Hák et al., 2007). Indicators provide information on the state of a phenomenon of interest and fulfill, in general, three purposes: (1) simplification, (2) quantification, and (3) communication. Indicators have to fulfill the criteria of soundness, relevance, and practicability (e.g., data availability). This article discusses biophysical sustainability indicators for land use (*see* Sustainability and Sustainability Science).

In general biophysical land-use indicators refer to two aspects of land use. (1) The extent of land use, usually measured in m^2, ha, or km^2, and its change in time such as the expansion or contraction of a certain land use at the expense of another. (2) The intensity of land use. Land-use intensity encompasses the levels of socioeconomic inputs (e.g., labor, resources, water, energy, or capital) and/or altered output (value or quantity) per unit area (Erb et al., 2013; Kuemmerle et al., 2013). There is no unequivocal unit for measuring land-use intensity, but it strongly depends on the perspective and system of reference (see below). It is important to note that changes in intensity need not necessarily result in changes in land cover but occur within the same land-cover type.

Area-Related Indicators

The most intuitive metrics for land use refer to the extent of a certain land use. A prominent example is the extent of cropland and its expansion into pristine forests, a currently still widespread trend in particular in the tropics; it gained attention due to its far-reaching effects on global biodiversity and perturbations of the global carbon cycle. Other examples are the extent of forest, of organically farmed cropland, of areas with natural protection status, of fragmented areas, or of areas subject to or restored from degradation, within a given territory, and the respective changes over time. These metrics allow to draw conclusions on state and trends of the environment.

Methodologically, the development of area-related land-use indicators relies on the use of nominal scales and the unambiguous assignment of a discrete unit (class) to land-use types. The use of nominal information is powerful as it allows to quantified, map and traced land use through space and time. This, however, is complicated by the fact that land-use activities usually consist of different sub-activities (e.g., cropland agriculture consists of sowing, plowing, harrowing, weeding) which are often cultivar and purpose specific and thus performed with different intensity. Different land-use activities are often highly integrated, encompassing different types of land use, such as with agroforestry. Unambiguous attribution of land-use types is thus often intricate. Moreover, nominal scales are limited in analyzing continuous, functional processes, such as quality aspects of land use or land-use intensity (Erb, 2012).

Area-related land-use indicators are closely linked to the concept of land cover (*see* Land Change: The Merger of Land Cover and Land use Dynamics) and, in general, discern major land-use types such as cropland (arable land and permanent crops), infrastructure areas (urban and transport), forests used for forestry, grasslands, areas devoid of land use (wilderness), and protected areas (e.g., natural parks, nature reserves). As large areas cannot be unambiguously attributed to all of these classes, a certain fraction of land area is often labeled 'other land' without further specification. Land-use classifications are especially powerful in cases where land use closely relates to a specific with land cover, such as is the case with infrastructure use or cropping. In such cases, one can capitalize on the remarkable progress made by remote sensing techniques, which allow generating cheap, consistent, spatially explicit land-cover datasets with wall-to-wall representation, usually compiled in Geographic Information System (*see* Geographic Information Systems and Remote Sensing).

In cases where land use is not bound tightly to a certain land-cover type or ecosystem, the development and application of area-related indicators for land use is much less straightforward and prone to misconceptions. Prominent examples are the definitions of forests (land cover) and forestry (land use). First, definitions of forests differ significantly among different data sources, referring to different combinations of characteristics such as fraction of crown closure or tree height. This results in difficulties to track the extent of forests (Grainger, 2008). Second, forests can be used for other purposes than timber harvest, e.g., grazing, and individual plots of used forests might be tree-less, as is, e.g., the case with clear-cut areas in dense forests which are allowed to regrow. Such examples that lead to complex problems of attribution are not the exception, but the rule, and apply to almost all broad types of land use: Fallow areas as intrinsic features of less-extensive arable land farming or livestock that grazes in many ecosystems, including forests, semi-deserts, natural and artificial grasslands are but examples to illustrate this (Erb, 2012).

Structural Landscape Elements

In landscape ecology, indicators for landscape structures, i.e., the form of structural elements, have been developed and successfully applied to socioecological analyses (Forman and Godron, 1986). Several indicators for landscape heterogeneity have been developed in this discipline. Heterogeneity relates to differences and diversity within a landscape and discerns elements such as patches, matrix, and corridors. Patches are identified as the spatial units of a landscape, which are embedded in the matrix and may be connected by corridors. From the spatiotemporal characteristics of patches and corridors, indicators for, e.g., the connectivity of the landscape are developed: A high spatial connectivity, usually beneficial for species richness in an area, is given either when the patches are sufficiently close that movement can occur among them, or that there is some corridor along which the organisms can move. In general, human influence is found to result in a simplification and geometrization of landscape pattern, with detrimental effects on the provision of ecosystem services and biodiversity (Millennium Ecosystem Assessment, 2005). In consequence, metrics of the size, shape, number, and configuration of patches and corridors are used as surrogate indicators for biodiversity endangerment and/or land-use intensity.

Land-Use Intensity Indicators

A focus on the extent of land use neglects important aspects that relate to many, benign as well as detrimental, consequences of land use. The quantification of the intensity of land use deserves particular attention with this regard, as it is recognized as one of the most pervasive factors of global environmental change (Foley et al., 2005; Ellis et al., 2013). For instance, empirical analyses revealed that effects of land-use intensity (changes) on many Earth system components, such as energy flows or micro climate, can be of similar magnitude as land-cover changes (Luyssaert et al., 2014).

In recent decades, due to the widespread availability of spatially explicit land-cover datasets, the major focus of research on land use lies with land cover (changes), while other land-use changes (sometimes denoted 'subtle,' i.e., occurring within land-cover types) gained less attention (Erb et al., 2013; Kuemmerle et al., 2013). This is surprising, because studies on the intensity of land use reach back to the founders of economics such as Malthus, Riccardo and Von Thünen and re-gained impetus in the 1960s and 1970s with the seminal paper by E. Boserup on the *Conditions of Agricultural Growth* (Boserup, 1965). Studies of land cover are usually dominated by natural science approaches, while social sciences typically focus on land use, integrating these two perspectives is high on the agenda of land system science (GLP, 2005).

Land-use intensity is characterized by three core elements of measure (Erb et al., 2013; Kuemmerle et al., 2013): (1) inputs to the production system, (2) outputs from the production system, and (3) changes in ecosystem properties. All these elements are differently pronounced in different disciplines. While (1) and (2) directly relate to concepts such as the socioeconomic metabolism (Erb, 2012) and are established research themes in land-related research, e.g., not only in economics in the form of the production functions (in monetary units), but also in geography or anthropology, and (3) refers to (changes) in system states of ecosystems, a research strand that gained impetus in particular in natural-science-oriented research strands (Erb et al., 2013).

Output Intensity

Output indicators encompass the outputs of land to society, most prominently the amount of produce per time period and

unit area (e.g., tons of cereals harvested, milk produced, or timber removed from forests per hectare and year). While measuring output appears straightforward, the development of robust and comparable indicators is intricate. These intricacies related to (1) the unit of measurement and (2) the methodology used to measure output consistently (Erb et al., 2013). Results of output intensity measurements can vary considerably when different units are used, e.g., mass fresh weight, mass dry matter, volume, energy, calorific value, or monetary value per area and year, and with regard to the reference system: agricultural areas or industrial sectors. Indicators for meat production, e.g., can be expressed as live animals per unit area (stocking density), meat output per fodder input (feed conversion ratio), or the carcass weight of animals, all of which yield different insights in the intensity of the land-based production process. Moreover, comparability of results is also hampered due to the fact that the yields in agriculture and forestry can vary widely due to differences in climate, soil conditions, or management, and are specific for individual cultivars or cultivar groups. This renders the simple use of yield data as surrogates for output intensity problematic.

Methods have been developed that take these differences in production conditions into account. Among those, the 'yield gap' method has gained attention. 'Yield gaps' denote the difference between agricultural yields at the plot, farm, or regional level to a reference yield that is attained under similar conditions of production, i.e., in the same region, but with optimal management (e.g., Lobell et al., 2009). Furthermore, on cropland the measurement of output appears to be intuitive and simple, but it is often complex: Because in most production systems a considerable fraction of the total cropland is periodically left fallow to maintain the soil fertility, it is important to relate the outputs to the full production cycle to generate comparable values. Such data, however, are usually not readily available or are highly uncertain (Kuemmerle et al., 2013).

Input Intensity

Metrics of input intensity quantify inputs to the production system. Here, the substitution of capital, labor, and skills against land is at the center of attention, which stimulated the development of land-use intensity indices based on (combinations of) input data, such as fertilizer application or labor input per unit cropland and year, mechanization, etc. One pivotal input indicator for input intensity is cropping intensity, i.e., the land actually under production (counted each time it is harvested in multicropping) divided by the total cropland extent (Boserup, 1965). For cropland with fallows cycles, this indicator is between 0 and 1, for multicropping, >1. The focus on this indicator is based on the observation that, under preindustrial conditions, the annual crop yield of a parcel of land can scarcely be enhanced, and only an increased cropping frequency can increase production. Similarly, rotation length is an important intensity indicator in forestry. The cropping intensity index gained particular attention in anthropology and was expanded by, e.g., indicators that consider the different lengths in fallow and cropping period or indices for used technologies (Shriar, 2000).

Intensity Indicators at the Ecosystem Level

The systematic observation of the effects of land-use intensity on the ecological patterns and processes, such as nutrient cycling, biodiversity, net primary production (NPP; *see* Energy Flow Analysis), carbon storage, water quality, and water retention capacity, is an important component for a comprehensive analytical framework for land-use research (GLP, 2005). Metrics that relate to alterations of system properties can also be used to develop indicators for land-use intensity.

Quantifying the effects of changes in the ecosystem parameters and consequential alterations of the provision of ecosystem services, thus, appears central for developing indicators of system-level outcomes. While the measurements of effects of such alterations on the provision of ecosystem services have proven to be difficult (Haines-Young et al., 2012), indicator systems that aim at quantifying these changes at the system level have been suggested. These approaches assess the differences between the actual and potential states of the ecosystem, and use this 'distance' as a surrogate indicator for land-use intensity (**Figure 1**). This notion is closely related to the ecological definition of disturbance (Pickett and White, 1986). Examples for such approaches include the spatial inventories of the density of human artifacts on the Earth's surface (Sanderson et al., 2002), or the land use and disturbance intensity index proposed for forestry, as based on the ecological law of self-thinning (Luyssaert et al., 2011).

In line with this notion, the concept of 'hemeroby' describes gradients of human influence on landscape and flora and is

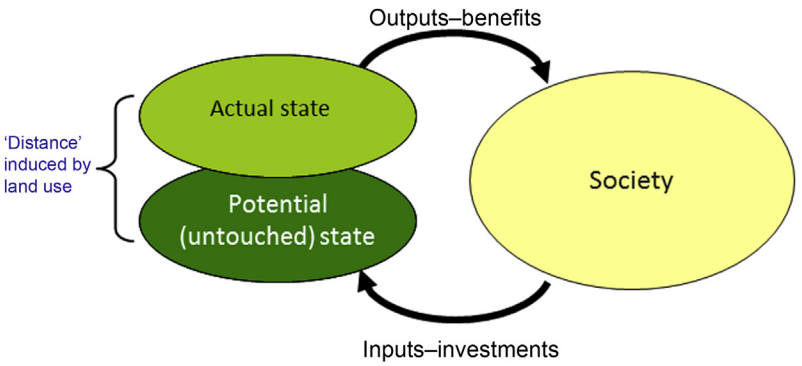

Figure 1 Indicators for land-use intensity integrate socioeconomic investments and inputs to land, outputs, and benefits gained from land and the system-level changes induced by management, which can be quantified as 'distance' between potential and actual state.

also conceptualized along the lines of disturbance and gained attention in land-use planning (Naveh and Lieberman, 1990) and the integration of land-use-related aspect such as biodiversity loss in Life Cycle Analyses (LCA). The hemeroby concept closely relates to the concept of naturalness and classifies landscape elements on an ordinal scale from levels of no human influence ('ahmemerob') to purely artificial and the related ecosystems and vegetation ('metahemerob') in varying graduations. Input data include the share of life-forms (e.g., trees, shrubs, grasses) and introduced species of vascular plants in the prevailing plant association, as well as by parameters that describe human impacts on ecosystems such as land-use types, plant communities, and soils.

Another surrogate indicator is the indicator 'ecological integrity' (Pimentel et al., 2000). Ecological integrity refers to the stability of biotic communities and relates to the idea of naturalness (see above) and 'ecosystem health,' in analogy to human health. Land use is conceptualized as reducing the degrees of freedom for natural processes to self-organize, i.e., their integrity; thus, major emphasis is put on the supporting ecosystem function such as NPP, soil formation, or nutrient cycling.

One prominent indicator that relies on a comparison of potential and actual state of a fundamental ecosystem process, i.e., NPP, is the land-use intensity indicator 'human appropriation of net primary production' (HANPP). This indicator, proposed in the late 1980s by Vitousek et al. (1986) and was further developed since (Haberl et al., 2007a; Krausmann et al., in press) and closely relates to and expands the concept social metabolism. HANPP assesses changes in the availability of NPP (*see* Energy Flow Analysis) for ecological processes induced by (1) alterations of the productivity of vegetation that result from land use and (2) extraction of NPP from ecosystems through biomass harvest. One example is forest clearing for agricultural purposes: the replacement of forests with arable land considerably affects (and predominantly decreases) NPP. Second, harvest removes NPP from the ecosystem in the form of biomass, thereby reducing the amount of available biomass remaining in the ecosystem for all heterotrophic food chains or for biomass accumulation. These two processes can either be expressed in the unit of biomass, carbon, or energy flow per year and area assessed by calculating the difference between the NPP that would prevail (hypothetically) in the absence of land use and the NPP that actually remains in ecosystems, i.e., after harvest. Because NPP is a central resource for society as well as for natural processes, HANPP is a valuable socio-ecological concept that systematically integrates socioeconomic (biomass harvest) and ecological information (NPP) and thus is accessible for socioeconomic as well as ecological research (Erb et al., 2009). A particular important aspect of HANPP is the close interrelation between energy flows in ecosystems and biodiversity, asserted in the so-called 'species-energy hypothesis: HANPP, i.e., a metric of the alteration of energy flows in ecosystems, has been found to be a robust land-use indicator for biodiversity endangerment (Haberl et al., 2007b).

Many more system-level indicators have been developed and could be used as system-level indicators for land-use intensity indicators, e.g., the effect of land use on terrestrial carbon stocks, on nitrogen, or water cycles. Datasets, however, that would allow for straightforward assessments are still rare, particularly at the supranational level (Kuemmerle et al., 2013).

Integrated Indicators

A few land-use indicators have been proposed that integrate two or more dimensions of land use. Agricultural yield, for instance, is already a combination of output and input metric: produce per land area subject to production. The indicator 'energy return on investment' (EROI; *see* Energy Flow Analysis), important in system-level and product-level approaches alike, balances energetic inputs and outputs to and from land and can illustrate the net-effects of intensification strategies and allows for insights in the interrelation of land-use intensity and efficiency. Moving beyond analyses of input–output relations, one can go for analyzing the interrelation between intended and unintended effects of land use. However, only a few examples exist: HANPP, e.g., allows to systematically linking land-use outputs (harvest) with the energetic change caused in the ecosystem, i.e., to calculate the HANPP-efficiency of land use. Similar developments are required as they would provide the basis for systematic analysis of the trade-offs and synergies related to land use such as land competition (Smith et al., 2010) and effects on climate change mitigation (*see* Land Change: The Merger of Land Cover and Land use Dynamics), or food security.

Spatial Reference of Land Use Information

Both, social and natural processes are highly diverse in space, which renders the generation of land-use indicators intricate. Most datasets related to natural processes, and thus also land cover, are available in spatial grids. Besides thematic problems of homogenous, comparable legends, maps also differ with regard to geographic protection and spatial resolution, and need to be reconciled in GIS environments (*see* Geographic Information Systems and Remote Sensing). However, pixels, or grid cells, are not system entities of socioeconomic processes. Land-use datasets, in contrast, are usually available for administrative units, which are the decisive spatial units for the organization and management of natural resource use. The development of spatial data which bridge the different spatial scales of ecological, social, and socioeconomic processes is required for the development of land-use indicator systems (Liverman et al., 1998).

Production and Consumption Indicators

Many land-use indicators refer to the ecosystem or the place of land use, and thus follow a production perspective. Alternative indicators that follow a consumer's perspective have been proposed that systematically link production and consumption. The principle of these accounting schemes, related to the concept of LCA, is displayed in **Figure 2**. These frameworks allow to study drivers of land use and land-cover change (demand for commodities is among the most important land-use drivers) as well as phenomena such as displacement of land

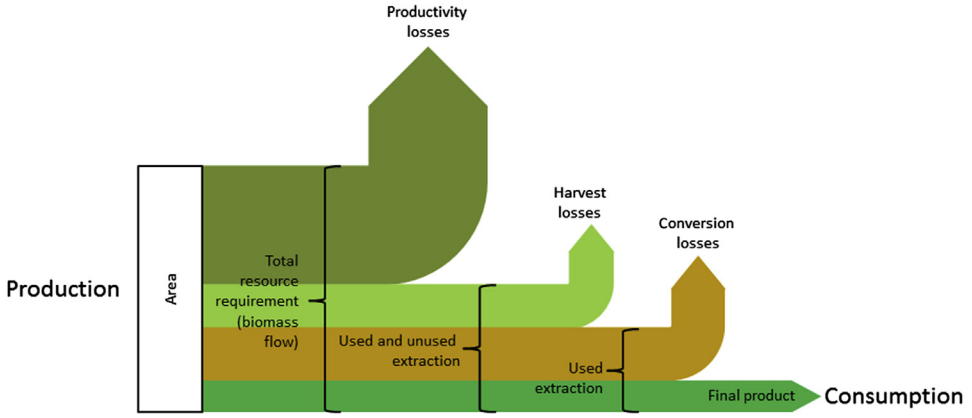

Figure 2 Production–consumption interlinkages. Arrow sizes are proportional to global flows of biomass from Haberl, H., Erb, K.H., Krausmann, F., Gaube, V., Bondeau, A., Plutzar, C., Gingrich, S., Lucht, W., Fischer-Kowalski, M., 2007a. Quantifying and mapping the human appropriation of net primary production in earth's terrestrial ecosystems. Proceedings of the National Academy of Sciences USA 104, 12942–12947.

use or leakage (Meyfroidt et al., 2013; *see* Land Change) induced by the international trade of commodities.

A simple consumer-based land-use indicator is the Ecological Footprint. Instead of calculating the extent of land use within a territory, it assesses the area demand required by society to sustain its apparent consumption of resources and goods, i.e., domestic production plus imports minus exports (*see* Material Flow Analysis; Energy Flow Analysis). Ecological footprint studies evaluate how much land is needed to produce the biomass consumed, to host infrastructure, and to absorb wastes, above all, CO_2 generated by a human population, and compare this footprint to the global or national average per-capita area availability of land. In its original version, an Ecological Footprint analysis uses the unit 'global hectares' by calculating the area that would be necessary to maintain these flows on the global average, i.e., using global average yields per cultivars in any given year and weighing different land-use types according to their contribution of global biomass flows.

Whereas on a global level the interpretation of a footprint calculation is straightforward, as neither trade nor differences in agricultural yields distort the picture, the interpretation of ecological footprint results is intricate at sub-global levels. Furthermore, ecological footprint assessments do not provide information on the actual size of land under use, but rather on the share of global biological production a society requires (expressed in the unit 'global hectares'). In consequence, the footprint approach has been modified in order to be able to monitor land requirements whilst avoiding this central shortcoming of ecological footprint analyses. Such approaches assess the extent of land use domestically and abroad in a manner compatible with other forms of monitoring land use, e.g., remote sensing and resource use, such as MFA and EFA (Erb, 2004; *see* Material Flow Analysis; Energy Flow Analysis).

These intricacies as well as the problem of multiple land use led to the development of the embodied HANPP (eHANPP) concept (Haberl et al., 2012). The eHANPP concept systematically links the amount of HANPP caused by land use to the consumption of final land-based products. By focusing on NPP, this concept allows to overcome one of the central weaknesses of land demand studies: the extent of land use is usually not a good indicator for the associated ecological impact. For instance, areas can be used intensively or extensively, and environmental damage is strongly associated with the intensity of certain land-use practices. Furthermore, land can be used several times a year (e.g., intercropping, agroforestry), while annual NPP can only be used in once if no stocks are to be depleted. In contrast to resource use indicators, such as those derived with MFA techniques (*see* Material Flow Analysis), fossil–energy related CO_2 emissions or the ecological footprint, eHANPP is only loosely correlated to GDP, which can be explained by improvements in agricultural efficiency associated with GDP growth (Haberl et al., 2012). eHANPP studies thus allow to cover land use aspects that are not grasped by other resource-use indicator systems.

Indicator Systems

A widely applied Indicators classification system for environmental and sustainability concerns is the pressure-state-response (PSR) scheme put forward by the OECD, or its extended version, the DPSIR scheme (driving forces-pressures-states-impacts-response) framework by the European Environmental Agency. This framework allows to categorize indicators by distinguishing different levels or reference systems of indicators: (1) driving forces and (2) pressures on the environment (in land use studies these two are often denoted as ultimate and proximate causes), (3) state indicators (including system-level indicators for natural processes or variables describing the natural environment), (4) impacts (effects on ecological or social systems, mostly unintended), and (5) responses (activities following undesired impacts that can be targeted at any part of the chain between driving forces and impacts). A sound indicator system consists of the balanced set of these five types of indicators, as none indicator can grasp the multitude of dimensions related to sustainability (Hák et al., 2007).

The DPSIR scheme gained particular attention in land-use planning and decision making as it well addresses the policy life cycle (Hák et al., 2007), schematically illustrated in **Figure 3**.

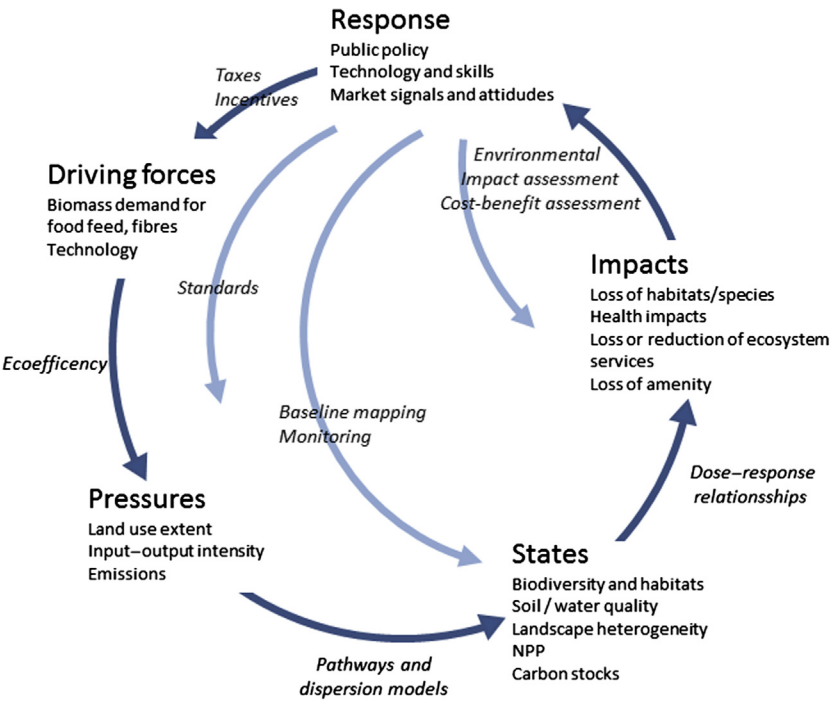

Figure 3 The DPSIR scheme, associate information flows and policy actions. Source: Compiled from Hák, T., Moldan, B., Dahl, L., 2007. Sustainability Indicators: A Scientific Assessment, Scope. Island Press.

The DPSIR scheme allows identifying a broad spectrum of indicators. The above-discussed biophysical indicators can be embedded in the DSPIR scheme and complemented by economic and social indicators. Biophysical indicators that depict the production–consumption disparity operationalize driving force indicators, next to indicators for socioeconomic development such as GDP, profit contribution of farms, or income distribution. Indicators that refer to land-use extent and input–output intensity represent pressure indicators. System-level intensity as well as spatial heterogeneity are state indicators that closely link to impacts such as biodiversity loss, degradation, or the loss of other ecosystem services. Here also social indicators such as health impacts and their social distribution. Responses include activities focusing on impacts (e.g., restoration programs), states (conservation programs), pressures (standards for agricultural inputs, harvest allowances), and driving forces (e.g., technological development programs).

Conclusion and Outlook

Land use provides the biophysical basis of human societies and is at the same time a pervasive factor of global environmental change. Land represents a confined resource, of which almost three quarters are currently used for infrastructure, cropping, grazing, or forestry, and at varying degrees of intensity. The demand for land is anticipated to increase, however, due to a growing and increasingly affluent world population that aims at satisfying consumption needs. Improving our still limited understanding of land use (change), its driving forces and constraints as well as its impacts on ecological as well as on social systems, is a central mandate of sustainability science.

Land-use-indicator systems that allow to measure and monitor the many aspects of society–nature interlinkages related to terrestrial ecosystems play a central role in this context. The development of such systems can build on a rich tradition and approaches established in disciplines such as ecology, geography, anthropology, and economics. Furthering these approaches by bridging disciplinary boundaries and making use of the plurality of scientific perspectives is particularly fruitful, but also required to close knowledge gaps, reduce uncertainty, and enhance our forecasting and monitoring capabilities. The nature of the present sustainability challenges, such as climate change, biodiversity conservation or food security, requires interdisciplinary, integrated approaches in order to produce effective, useful information for policymakers and practitioners, as well as for stakeholders.

Acknowledgment

I gratefully acknowledge funding from ERC-Stg-2010 263522 LUISE.

See also: Energy Flow Analysis; Food Security and 'Green Revolution'; Geographic Information Systems and Remote Sensing; Land Change: The Merger of Land Cover and Land use Dynamics; Material Flow Analysis; Social Ecology; Sustainability and Sustainability Science.

Bibliography

Boserup, E., 1965. The Conditions of Agricultural Growth: The Economics of Agrarian Change under Population Pressure. Earthscan, London.

Ellis, E.C., Kaplan, J.O., Fuller, D.Q., Vavrus, S., Goldewijk, K.K., Verburg, P.H., 2013. Used planet: a global history. Proceedings of the National Academy of Sciences USA 110, 7978–7985. http://dx.doi.org/10.1073/pnas.1217241110.

Erb, K.-H., 2004. Actual land demand of Austria 1926–2000: a variation on ecological footprint assessments. Land Use Policy 21, 247–259.

Erb, K.-H., 2012. How a socioecological metabolism approach can help to advance our understanding of changes in land-use intensity. Ecological Economics 76, 8–14.

Erb, K.-H., Haberl, H., Jepsen, M.R., Kuemmerle, T., Lindner, M., Müller, D., Verburg, P.H., Reenberg, A., 2013. A conceptual framework for analysing and measuring land-use intensity. Current Opinion in Environmental Sustainability 5, 464–470.

Erb, K.-H., Krausmann, F., Gaube, V., Gingrich, S., Bondeau, A., Fischer-Kowalski, M., Haberl, H., 2009. Analyzing the global human appropriation of net primary production – processes, trajectories, implications. An introduction. Ecological Economics 69, 250–259.

Foley, J.A., DeFries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., Chapin, F.S., Coe, M.T., Daily, G.C., Gibbs, H.K., others, 2005. Global consequences of land use. Science 309, 570.

Forman, R.T.T., Godron, M., 1986. Landscape Ecology. Wiley.

GLP, 2005. Global Land Project: Science Plan and Implementation Strategy. IGBP Secretariat.

Grainger, A., 2008. Difficulties in tracking the long-term global trend in tropical forest area. Proceedings of the National Academy of Sciences USA 105, 818–823.

Haberl, H., Erb, K.H., Krausmann, F., Gaube, V., Bondeau, A., Plutzar, C., Gingrich, S., Lucht, W., Fischer-Kowalski, M., 2007a. Quantifying and mapping the human appropriation of net primary production in earth's terrestrial ecosystems. Proceedings of the National Academy of Sciences USA 104, 12942–12947.

Haberl, H., Erb, K.-H., Plutzar, C., Fischer-Kowalski, M., Krausmann, F., 2007b. Human appropriation of net primary production (HANPP) as indicator for pressures on biodiversity. In: Sustainability Indicators. A Scientific Assessment. Island Press, Washington, DC, pp. 271–288.

Haberl, H., Steinberger, J.K., Plutzar, C., Erb, K.-H., Gaube, V., Gingrich, S., Krausmann, F., 2012. Natural and socioeconomic determinants of the embodied human appropriation of net primary production and its relation to other resource use indicators. Ecological Indicators 23, 222–231.

Haines-Young, R., Potschin, M., Kienast, F., 2012. Indicators of ecosystem service potential at European scales: mapping marginal changes and trade-offs. Ecological Indicators 21, 39–53.

Hák, T., Moldan, B., Dahl, L., 2007. Sustainability Indicators: A Scientific Assessment, Scope. Island Press.

Krausmann, F., Erb, K.-H., Gingrich, S., Haberl, H., Bondeau, A., Gaube, V., Lauk, C., Plutzar, C., Searchinger, T.D., 2013. Global human appropriation of net primary production doubled in the 20th century. Proceedings of the National Academy of Sciences USA 110, 10324–10329. http://dx.doi.org/10.1073/pnas.1211349110.

Kuemmerle, T., Erb, K., Meyfroidt, P., Müller, D., Verburg, P.H., Estel, S., Haberl, H., Hostert, P., Jepsen, M.R., Kastner, T., Levers, C., Lindner, M., Plutzar, C., Verkerk, P.J., van der Zanden, E.H., Reenberg, A., 2013. Challenges and opportunities in mapping land use intensity globally. Current Opinion in Environmental Sustainability 5, 484–493.

Liverman, D., Moran, E.F., Rindfuss, R.R., Stern, P., 1998. People and Pixels: Linking Remote Sensing and Social Science. National Academy Press, Washington, DC, 256p.

Lobell, D.B., Cassman, K.G., Field, C.B., 2009. Crop yield gaps: their importance, magnitudes, and causes. Annual Review of Environment and Resources 34, 179.

Luyssaert, S., Hessenmöller, D., von Lüpke, N., Kaiser, S., Schulze, E.D., 2011. Quantifying land use and disturbance intensity in forestry, based on the self-thinning relationship. Ecological Applications 8, 3272–3284.

Luyssaert, S., Jammet, M., Stoy, P.C., Estel, S., Pongratz, J., Ceschia, E., Churkina, G., Don, A., Erb, K., Ferlicoq, M., Gielen, B., Grünwald, T., Houghton, R.A., Klumpp, K., Knohl, A., Kolb, T., Kuemmerle, T., Laurila, T., Lohila, A., Loustau, D., McGrath, M.J., Meyfroidt, P., Moors, E.J., Naudts, K., Novick, K., Otto, J., Pilegaard, K., Pio, C.A., Rambal, S., Rebmann, C., Ryder, J., Suyker, A.E., Varlagin, A., Wattenbach, M., Dolman, A.J., 2014. Land management and land-cover change have impacts of similar magnitude on surface temperature. Nature Climate Change 4, 389–393.

Meyfroidt, P., Lambin, E.F., Erb, K.-H., Hertel, T.W., 2013. Globalization of land use: distant drivers of land change and geographic displacement of land use. Current Opinion in Environmental Sustainability, Human Settlements and Industrial Systems 5, 438–444.

Millennium Ecosystem Assessment, 2005. Ecosystems and Human Well-being: Current State and Trends, vol. 1. Island Press, Washington, DC.

Naveh, Z., Lieberman, A.S., 1990. Landscape ecology: theory and application. Landscape Ecology: Theory and Applications.

Pickett, S.T.A., White, T.A., 1986. The Ecology of Natural Disturbance and Patch Dynamics. Academic Press.

Pimentel, D., Westra, L., Noss, R.F., 2000. Ecological Integrity: Integrating Environment, Conservation, and Health. Island Press.

Reid, W.V., Chen, D., Goldfarb, L., Hackmann, H., Lee, Y.T., Mokhele, K., Ostrom, E., Raivio, K., Rockström, J., Schellnhuber, H.J., others, 2010. Earth system science for global sustainability: grand challenges. Science 330, 916–917.

Sanderson, E.W., Jaiteh, M., Levy, M.A., Redford, K.H., Wannebo, A.V., Woolmer, G., 2002. The human footprint and the last of the wild. BioScience 52, 891–904.

Shriar, A., 2000. Agricultural intensity and its measurement in frontier regions. Agroforestry Systems 49, 301–318.

Smith, P., Gregory, P.J., van Vuuren, D., Obersteiner, M., Havlík, P., Rounsevell, M., Woods, J., Stehfest, E., Bellarby, J., 2010. Competition for land. Philosophical Transactions of the Royal Society B: Biological Sciences 365, 2941–2957.

Turner, B.L., Lambin, E.F., Reenberg, A., 2007. The emergence of land change science for global environmental change and sustainability. Proceedings of the National Academy of Sciences USA 104, 20666–20671.

Vitousek, P.M., Ehrlich, P.R., Ehrlich, A.H., Matson, P.A., 1986. Human appropriation of the products of photosynthesis. BioScience 36, 368–373.

Relevant Websites

http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Agri-environmental_indicator_-_land_use_change – Agroecological Land-Use Indicators by Eurostat.

http://ia2dec.ew.eea.europa.eu/knowledge_base/Frameworks/doc101182/ – The DPSIR Framework by the EEA.

http://www.fao.org/docrep/w4745e/w4745e00.htm#Contents – FAO: Land Quality Indicators and Their Use in Sustainable Agriculture and Rural Development.

http://faostat.fao.org/site/377/default.aspx#ancor – Land Use Database by the FAO.

http://cfpub.epa.gov/eroe/index.cfm?fuseaction=detail.viewInd&lv=list.listbyalpha&r=219665&subtop=225 – United States Environmental Protection Agency – Report on the Environment, Land Use.

Language Acquisition

Brian MacWhinney, Carnegie Mellon University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Adults often underestimate the size and scope of the social, cognitive, and motoric challenges that children face when learning language. In fact, language learning is the most complex task that any of us will master, and the fact that we all succeed in learning our native language reflects the extent to which language is shaped to conform to the abilities we all possess.

The ability to learn language is present in almost every human child. Children who are born blind have no problem learning to speak, although they may occasionally be confused about words for colors or locations. Children who are born deaf readily acquire a rich system of signs, as long as they are exposed to native sign language speakers. Even a child, like Helen Keller, who has lost both hearing and sight, can still acquire language through symbols expressed in touch and motion. Children with neurological disorders, such as brain lesions or hydrocephalus, often acquire complete control over spoken language, despite some early delay (MacWhinney et al., 2000). Children with the most extreme forms of mental retardation are still able to acquire the basic units of human communication. Given this pervasiveness and inevitability of first-language acquisition, we often tend to take the process of language learning for granted. But language is the most complex skill that a human being can master. The fact that nearly all of us succeed in this task indicates how remarkably well the structure of language adapts to our underlying abilities. Language is immensely complex, but its complexity is accessible to all of us because it is uniquely crafted to fit our human nature (Christiansen and Chater, 2008).

We can look at language acquisition from different perspectives. Of these, the three most fully developed are those of the linguist, the psychologist, and the educator. Linguists tend to think of language as having a universal core from which individual languages select out a particular configuration of features, parameters, and settings (Chomsky, 1982). From this perspective, child language is an interesting slice of the universal pie. The shape of this slice is presumably limited by both formal universal constraints and the child's mental abilities or developmental status.

Psychologists (MacWhinney, 1998) look at language learning from a very different perspective. To the psychologist, language acquisition is a window on the operation of the human mind. This window allows us to view the structure and functioning of neural circuits in the brain. It also allows us to understand how these circuits support processes such as reinforcement, generalization, imagination, and thinking. To understand these processes better, psychologists conduct controlled experiments in which children learn new words, sounds, and rules. They may measure these processes using neural imaging techniques, or they may simply study the changes in the language of the child across time.

Educators and parents are typically interested in understanding how to use language learning to promote children's overall social, emotional, and cognitive development. Working within the tradition of sociocultural theory, as articulated by Vygotsky (1934), these researchers examine the ways in which children use words in specific games and interactions with their parents as a way of acquiring the patterns of their culture (Nelson, 1998).

Important and engaging though these adult perspectives may be, the best way to appreciate the dynamics of language development is to assume the perspective of the child. In particular, we need to understand the specific challenges that language learning presents to the child and the ways in which each is overcome. For detailed data on actual interactions between children and their parents, readers are encouraged to browse the many recordings and transcripts to be found at http://childes.talkbank.org.

Early Audition

William James (1890) described the world of the newborn as a "blooming, buzzing confusion." However, on the auditory level, the newborn's world is remarkably well structured. The cochlea and auditory nerve provide extensive preprocessing of signals for pitch and intensity. In the 1970s, researchers discovered that human infants were specifically adapted at birth to perceive contrasts such as that between 'p' and 'b,' as in *pit* and *bit*. However, subsequent research showed that even chinchillas are capable of making this distinction (Werker, 1995). Thus, it appears that much of the basic structure of the auditory world can be attributed to fundamental processes in the mammalian ear.

Beyond this basic level of auditory processing, it appears that infants have a remarkable capacity to record and store sequences of auditory events. For example, if a 6-month-old hears a string of nonsense syllables such as *badigudibagadigudigagidu* repeated many times, the parts that are repeated will stand out and affect later listening. In this example, the repeated string is *digudi*. If infants are trained on such strings, the repeated segment will first pop out, and then they will grow tired of this sound and will come to prefer to listen to new sound strings rather than those that have the boring old *digudi* string (Thiessen, 2009). These habituation effects are strongest for stressed syllables and syllables immediately following stressed syllables.

Recent studies of these effects in auditory memory suggest that we are born with an ability to store and recall the sounds of human language. It is as if the infant has a tape recorder in the auditory cortex that records input sounds, replays them, and accustoms the ear to their patterns.

Babies also demonstrate preferences for the language that resembles the speech of their mothers. Thus, a French infant will prefer to listen to French, whereas a Polish infant will prefer to listen to Polish (Jusczyk, 1997). In addition, babies demonstrate a preference for their own mother's voice, as opposed to that of other women. Together, these abilities and preferences suggest that during the first 8 months, the child is remarkably attentive to language. Although children are not yet learning words, they are acquiring the basic auditory and intonational patterns of their native language. As they sharpen their ability to hear the contrasts of their native language, they begin to lose the ability to hear contrasts not represented in the native language. If the child is growing up in a bilingual world, full perceptual flexibility is maintained. However, if the child is growing up in a monolingual world, flexibility in processing is gradually traded off for quickness and automaticity.

Early Articulation

During the first 3 months, a baby's vocalizations are nothing more than cries and vegetative adaptations. However, around 3 months, at the time of the first social smile, babies begin to make the delightful little sounds that we call 'cooing.' These sounds have no particular linguistic structure, but their well-integrated intonation makes them sure parent pleasers. By 6 months, the baby is producing structured vocalizations, including a larger diversity of nasals, vowel types, and syllables with the canonical consonant–vowel (CV) structure. The basic framework of early babbling seems to be constructed on top of patterns of noisy lip smacking that are present in many primates (MacNeilage, 1998). These vocal gestures include some form of vocal closure followed by a release with vocalic resonance. Essentially, this is the CV syllable in which a consonant is followed by a vowel.

Until the sixth month, deaf infants continue to babble normally. However, by the age of 9 months, deaf infants lose their interest in babbling. This suggests that earlier babbling is sustained largely through proprioceptive and somaesthetic feedback, as babies explore the various ways in which they can play with their mouth. After 6 months, babbling relies increasingly on auditory feedback. During this period, the infant tries to produce specific sounds to match up with specific auditory impressions. It is at this point that deaf children no longer find babbling entertaining, since they cannot obtain auditory feedback. These facts suggest that, from the infant's point of view, babbling is essentially a process of self-entertainment.

In the heyday of behaviorism, researchers analyzed the development of babbling in terms of reinforcement theory (Mowrer, 1960). They thought that process of producing reinforcing sounds like the sounds made by the child's mother would lead a Chinese baby to babble the sounds of Chinese, whereas a Quechua baby would babble the sounds of Quechua. This was the theory of 'babbling drift.' However, closer observation of the babbling of 8-month-olds indicates that virtually no such drift occurs. By 12 months, there is some slight drift in the direction of the native language, as the infant begins to acquire the first words. Proponents of universal phonology have sometimes suggested that all children engage in babbling all the sounds of all the world's languages. Here, again, the claim seems to be overstated. Although it is certainly true that some English-learning infants will produce Bantu clicks and Quechua implosives, not all children produce all of these sounds.

The First Words

The child's ability to produce the first word is based on three earlier developments. The first is the infant's growing ability to record the sounds of words. The second is the development of an ability to control vocal productions that occurs in the late stages of babbling. The third is the general growth of the symbolic function, as represented in play, imitation, and object manipulation. Piaget (1954) characterized the infant's cognitive development in terms of the growth of representation or the 'object concept.' In the first 6 months of life, the child is unable to think about objects that are not physically present. However, as infants learn more about objects, they become able to associate their properties with their own actions and other features of the context. In this way, subtle cues can be used to dredge up fuller representations from memory. For example, a child may see a dog's tail sticking out from behind a chair and realize that the rest of the dog is hiding behind the chair. This understanding of how parts relate to wholes supports the child's first major use of the symbolic function. When playing with toys, the 12-month-old will begin to produce sounds such as room or bambam that represent properties of these toys and actions. Often these phonologically consistent forms (Peters, 1983) appear before the first real words. Because they have no clear conventional status, parents may tend to ignore these first symbolic attempts as nothing more than spurious productions or babbling.

Even before children have produced their first conventional word, they have already acquired an ability to comprehend as many as 10 conventional forms. They learn these forms through frequent associations between actions, objects, and words. Parents often realize that the prelinguistic infant is beginning to understand what they say. However, they are hard pressed to demonstrate this ability convincingly. Researchers deal with this problem by bringing infants into the laboratory, placing them in comfortable high chairs, and asking them to look at pictures, using the technique of visually reinforced preferential looking (Schafer and Plunkett, 1998). A name such as dog is produced across loudspeakers. Pictures of two objects are then displayed. In this case, a dog may be on the screen to the right of the baby and a car may be on the screen to the left. If the child looks at the picture that matches the word, a toy bunny pops up and does an amusing drum roll. This convinces babies that they have chosen correctly and they then do the best they can to look at the correct picture on each trial. Some children get fussy after only a few trials, but others last for 10 trials or more at one sitting and provide reliable evidence that

they have begun to understand a few basic words. Many children show this level of understanding by the tenth month – often 2–3 months before the child has produced a recognizable first word.

Producing the first word is a bit like stepping out on stage. When babbling for their own entertainment, the only constraints that infants face are those arising from their own playfulness and interest. If they want to play around with a particular articulatory pattern or sound, there are no problematic consequences. However, when faced with the task of producing word forms, the child's articulation has to be extremely accurate and within conventional limits. Because the motoric aspects of speech production are so demanding, the requirement to produce conventional forms represents an enormous barrier to the child's learning of language. As a result, the forms of early words often deviate radically from the adult standard. The many simplifications that the 1-year-old introduces to adult phonology are well known to students of phonological development (Vihman and Croft, 2007). Children tend to drop unstressed syllables, producing *hippopotamus* as *poma*. They repeat consonants, producing *water* as *wawa*. And they simplify and reduce consonant clusters, producing tree as pee. These phonological processes echo similar processes found in the historical development and dialectal variation of adult language. What is different in child language is the fact that so many simplifications occur at once, making so many words difficult to recognize.

Early Semantics

Since Plato, scholars have considered the task of figuring out word meaning to be a core intellectual challenge. Quine (1960) illustrated the problem by imagining a scenario in which a hunter is out on safari with a native guide. Suddenly, the guide shouts 'Gavagai!' and the hunter, who does not know the native language, has to quickly infer the meaning of the word. Does it mean 'Shoot now' or 'There's a rhino' or perhaps even 'It got away'? Without some additional cues regarding the likely meaning of the word, how can the hunter figure this out?

The problem facing the toddler is similar to that facing the hunter. Fortunately, the toddler has some good cues to rely on. Foremost among these cues is the parent's use of joint attention and shared eye gaze to establish common reference for objects and actions. If the father says *hippo* while holding a hippopotamus in his hand, the child can use the manual, visual, verbal, and proxemic cues to infer that the word *hippo* refers to the hippopotamus. A similar strategy works for the learning of the names of easily produced actions such as falling, running, or eating. It also works for social activities such as *bath* or *bye-bye*. The normal child probably understands the role of shared eye gaze even before learning the first words. At 3 months, children maintain constant shared eye gaze with their parents. In normal children, this contact maintains and deepens over time. For autistic children, contact is less stable and automatic. As a result, autistic children may be delayed in word learning and the general growth of communication.

Shared attention reference is not the only cue toddlers use to delineate and pick out the reference of words. They also use the form of utterances to derive the meanings of new words. For example, if toddlers hear 'Here is a zav,' they know that *zav* is a common noun. However, if they hear 'Here is Zav,' then they know that *Zav* is either a proper noun or perhaps the name of a quantity. If they hear 'I want some zav,' they know that *zav* is a quantity and not a proper or common noun. Cues of this type can give a child a rough idea of the meaning of a new word (Tomasello, 2003). Other sentential frames can give an even more precise meaning. If the child hears, 'This is not green, it is chartreuse,' then it is clear that *chartreuse* is a color. If children hear, 'Please don't cover it, just sprinkle it lightly,' then they know that *sprinkle* is a verb of the same general class of cover. The use of cues of this type leads to a fast, but shallow, mapping of new words to new meanings.

As the child's stock of words grows, it becomes harder to keep words apart from each other. To solve this problem, children have to choose between two opposing strategies: overgeneralization and undergeneralization. It is extremely easy to detect overgeneralizations. If the child calls a tiger by the name *kitty*, there is clear evidence for overgeneralization. Undergeneralization is much more difficult to detect. For example, a child may use the word *dog* to refer only to the family dog and not to any other dog. Or a child may use the word *car* to refer only to cars parked outside a certain balcony in the house and not cars in any other context. Such undergeneralizations can only be detected if one takes careful note of the contexts in which a child avoids using a word.

Undergeneralizations are soon corrected, as they hear the relevant word used in a wider variety of contexts. Overgeneralizations can be corrected by supplying the name for the unknown object, such as the word *tiger* in the case of this example. In most cases, children are happy to learn these new names so that they can properly disambiguate things that are obviously so different as cats and tigers (Merriman and Bowman, 1989).

Word Combinations

Throughout the second year, the child struggles with perfecting the sounds and meanings of the first words. For several months, children produce isolated single words. However, eventually they need to go beyond his/her one-word level. They need to associate predicates such as want, more, or go with arguments such as cookie or mommy. The association of predicates to arguments is the first step in syntactic development. As in the other areas of language development, these first steps are taken in a very gradual fashion. Before producing a smooth combination of two words such as *my horsie*, children will often string together a series of single-word utterances that appear to be searching out some syntactic form. For example, a child might say *my*, *that*, *that*, *horsie* with pauses between each word. Later, the pauses will be gone and the child will say *that horsie*, *my horsie*. This tentative combination of words involves groping on both intonational and semantic levels. On the one hand, children must figure out how to join words together smoothly in production. On the other hand, they also must figure out which words can meaningfully be combined with which others.

As was the case in the learning of single words, this learning is guided by earlier developments in comprehension. We

have to assess children's early syntactic comprehension by controlled experiments in the laboratory. Here, as in the study of word learning, researchers have used the preferential looking paradigm. To the right of the child, there is a TV monitor with a movie of Big Bird tickling Cookie Monster. To the child's left, there is a TV monitor with a movie of Cookie Monster tickling Big Bird. The experimenter produces the sentence 'Big Bird is tickling Cookie Monster.' If children look at the matching TV monitor, they are reinforced and a correct look is scored. Using this technique, researchers have found that 17-month-olds already have a good idea about the correct word order for English sentences. This is about 5–6 months before they begin to use word order systematically in production.

The grammar of the first word combinations is extremely basic (MacWhinney, in press). The child learns that each predicate should appear in a constant position vis-à-vis the arguments it requires. For example, in English, the word more appears before the noun it modifies and the verb run appears after the subject with which it combines. Slot-filler relations can control this basic type of grammatical combination. Each predicate specifies a slot for the argument. For example, *more* has a slot for a following noun. When a noun, such as *milk*, is selected to appear with more, that noun fills the slot opened up by the word *more*. The result is the combination *more milk*. Later, the child can treat this whole unit as an argument to the verb *want* and the result is *want more milk*. Finally, the child can express the second argument of the verb *want* and the result is *I want more milk*. Thus, children build up longer sentences bit by bit, creating a more complex grammar. This type of word-based learning is present even in adults. In languages with strong morphological marking systems, word-based patterns specify the attachment of affixes, rather than just the linear position of words. In fact, most languages of the world make far more use of morphological marking than does English. In this regard, English is a rather exotic language.

Missing Glue

The child's first sentences are almost all incomplete and ungrammatical. Instead of saying, 'This is Mommy's chair,' the child produces only 'Mommy chair' with the possessive suffix, the demonstrative, and the copula verb all deleted. Just as the first words are full of phonological deletions and simplifications, the first sentences include only the most important words, without any of the glue. In some cases, children simply have not yet learned the missing words and devices. In other cases, they may know the 'glue words' but find it difficult to coordinate the production of so many words in the correct order.

These early omissions provide evidence for two major processes in language development. First, children try to make sure that the most important and substantive parts of the communication are not omitted. Unfortunately, children make this evaluation from their own, egocentric perspective. In an utterance like *Mommy chair*, it is not clear whether the child means 'This is Mommy's chair' or 'Mommy is sitting in the chair,' although the choice between these interpretations may be clear in context (Bloom, 1973). The second factor that shapes early omissions is phrasal frequency. Children tend to preserve frequent word combinations such as *like it* or *want some*. These combinations are often treated as units, producing errors such as 'I like it the ball' or 'I want some a banana.'

Productivity

Productivity can be demonstrated in the laboratory by teaching children names for new objects. For example, we can show a child a picture of a funny-looking creature and call it a *wug*. As we noted before, the positioning of the word *wug* after the article 'a' induces the child to treat the word as a common noun. Children can then move from this fact to infer that the noun *wug* can pluralize as *wugs*, even if they have never heard the word *wugs*.

Three-year-olds also demonstrate some limited productive use of syntactic patterns for new verbs. However, children tend to be conservative and unsure about how to use verbs productively until about age of 5 years. From the child's perspective, these laboratory experiments with strange new toys and new words tend to encourage a conservative approach. As they get older and braver, children start to show productive use of constructions such as the double object, the passive, or the causative (Brooks et al., 1999). For example, an experimenter can introduce a new verb like *griff* in the frame 'Tim griffed the ball to Frank' and the 5-year-old will productively generalize to 'Tim griffed Frank the ball.'

The control of productivity is based on two complementary sets of cues: semantics and co-occurrence. When children hear *a wag*, they correctly infer that *wag* is a count noun. In fact, because they also see a picture of a cute little animal, they infer that *wag* is a common, count, name for an animate creature. These semantic features allow them to generalize their knowledge by producing the form *wugs*. However, we could also view this extension as based on co-occurrence learning. The child learns that words that take the indefinite article also form plurals. On the other hand, words that take the quantifier some do not form plurals. In this way, the child can use both semantic and co-occurrence information to build up knowledge about the parts of speech. This knowledge can then be fed into existing syntactic generalizations to produce new combinations and new forms of newly learned words. The bulk of grammatical acquisition relies on this process.

The Logical Problem of Language Acquisition

The problem with productivity is that it produces overgeneralization. For example, an English-speaking child will soon learn to form the past tense of a new verb by adding one of the variant forms of *-ed*. This knowledge helps the child produce forms such as *jumped* or *wanted*. Unfortunately, it may also lead the child to produce an error such as *goed*. When this occurs, we can say that the child has formulated an overly general grammar. One way of convincing the child to reject the overly general grammar in which goed occurs is to provide the child with negative feedback. This requires the parent to tell the child, 'No, you can't say goed.' The problem here is that children seem to ignore parental feedback regarding the form of language. If the child calls a hamburger a hot dog, the parent can say, 'No, it is a hamburger.' The child will accept this type of semantic correction. But children are notoriously resistant to being corrected for formal grammatical features.

The fact that children tend to ignore formal correction has important consequences for language acquisition theory. In the 1960s, work in formal analysis convinced some linguists that the task of learning the grammar of a language was impossible unless negative feedback was provided. Since negative feedback appeared to be unavailable or unused, this meant that language could not be learned without some additional innate constraints. This argument has led to many hundreds of research articles exploring the ways in which children's learning places constraints on the form of grammar. Referring back to Plato's *Republic* and *the Allegory of the Cave*, Chomsky and others have characterized the task of language learning as a logical problem, much as Quine characterized the word learning Gavagai problem.

In fact, children have more resources available to them than Chomsky seems to suggest. Using these resources, the child can recover from overgeneralization without negative feedback. In the case of goed, everyone agrees that recovery is easy. All the child has to do is to realize that there is only one way of producing the past tense of go and that is went. In other words, the irregular form went comes to block production of the overregularized form goed. Here, recovery from overgeneralization is based on the competition between the regular pattern and the irregular form.

Consider another example. Suppose that a child decides that the verb *recommend* patterns like the verb *give*. After all, both verbs involve a beneficiary and an object being transferred. However, only *give* allows a double object construction, as in 'John gave the library the book.' Most people find 'John recommended the library the book' ungrammatical. One solution to this error is to avoid making it in the first place. However, a more general solution is to record the strength of the competing syntactic patterns. The correct way of saying 'John recommended the library the book' is to say 'John recommended the book to the library.' This correct formulation should be strengthened whenever it is heard. As the strength of the frame for the verb recommend grows in comparison to the ungrammatical frame, the use of the competing frame is blocked. This solution assumes that the child realizes that the two frames are in competition. It may be that reaching this realization requires some attention to syntactic form. However, this solution does not require children to pay attention to corrective feedback. Instead, they only need to attend to correct sentences and to make sure that they understand that these are competing ways of saying roughly the same thing.

Emergentist Accounts

In the 1980s, a sharp division existed between accounts of language acquisition that emphasized learning and those that emphasized innate knowledge of universal grammar. At the start of the twenty-first century, there is no longer a clear separation between alternative approaches. Instead, the challenge facing researchers is to formulate a set of mechanisms that is rich enough to account for the details of these emergent processes and that is grounded on the facts of human neurophysiology. To achieve this, many psychologists rely on the formalisms of neural network theory, which is also known as connectionism (Rumelhart and McClelland, 1986). This framework uses large numbers of units and the connections between these units to capture the patterns of language. This weblike architecture of nodes and connections is intended explicitly to resemble the structure of the human brain with neurons, synapses, and weights on synaptic connections. Neural network models have provided good accounts for the acquisition of morphology, phonology, and syntax (MacWhinney, 1999).

A major limitation of many current models is their failure to pay sufficient attention to the child's perspective on learning. In the late 1990s, researchers began to correct this problem. Models that rely on concepts derived from dynamic systems theory (Smith, 1999) and the theory of attention (Merriman, 1999) have proven useful in accounting for word learning by focusing on the ways in which concrete situations provide learners with cues. Newer formal models also relate word learning to children's physical control of their own body and their own perspective as a causal agent (MacWhinney, 2008). These new developments may also allow us to understand how language works to embed children into the fabric of social life (Nelson, 1998).

Conclusion

Language is a unique marker of humanity. It distinguishes the human species from the rest of the creation, and it allows us to share our thoughts and feelings. Language is the most complex skill that any of us will ever master. Despite this complexity, nearly every human child succeeds in learning language. This suggests that language is optimally shaped to mesh with the abilities of the human mind and body. On the one hand, the universals of human language match up with our neurological, cognitive, and physical abilities. At the same time, parents provide rich contextual and emotional support to guide children through the process of language discovery. By studying language learning, we learn more about universals of human language, the shape of social interaction, and the structure of the human mind.

See also: First Language Acquisition, Linguistic Theory of; Language Learning Impairment.

Bibliography

Bloom, L., 1973. One Word at a Time: The Use of Single Word Utterances. Mouton, The Hague.

Brooks, P.J., Tomasello, M., Dodson, K., Lewis, L.B., 1999. Young children's overgeneralizations with fixed transitivity verbs. Child Development 70, 1325–1337.

Chomsky, N., 1982. Some Concepts and Consequences of the Theory of Government and Binding. MIT Press, Cambridge, MA.

Christiansen, M., Chater, N., 2008. Language as shaped by the brain. Behavioral and Brain Sciences 31, 489–558.

James, W., 1890. The Principles of Psychology. Holt, Rinehart, and Winston, New York.

Jusczyk, P., 1997. The Discovery of Spoken Language. MIT Press, Cambridge, MA.

MacNeilage, P., 1998. The frame/content theory of evolution of speech production. Behavioral and Brain Sciences 21, 499–546.

MacWhinney, B., 1998. Models of the emergence of language. Annual Review of Psychology 49, 199–227.

MacWhinney, B., 2008. How mental models encode embodied linguistic perspectives. In: Klatzky, R., MacWhinney, B., Behrmann, M. (Eds.), Embodiment, Ego-Space, and Action. Lawrence Erlbaum, Mahwah, pp. 369–410.

MacWhinney, B. Item-based patterns in early syntactic development. In: Herbst, T. (Ed.), Valency Relations. Springer, Berlin, in press.
MacWhinney, B. (Ed.), 1999. The Emergence of Language. Lawrence Erlbaum Associates, Mahwah, NJ.
MacWhinney, B., Feldman, H.M., Sacco, K., Valdes-Perez, R., 2000. Online measures of basic language skills in children with early focal brain lesions. Brain and Language 71, 400–431.
Merriman, W., 1999. Competition, attention, and young children's lexical processing. In: MacWhinney, B. (Ed.), The Emergence of Language. Lawrence Erlbaum, Mahwah, NJ, pp. 331–358.
Merriman, W., Bowman, L., 1989. The mutual exclusivity bias in children's word learning. Monographs of the Society for Research in Child Development 54 (3–4), 1–129.
Mowrer, O., 1960. Learning Theory and the Symbolic Processes. Wiley, New York.
Nelson, K., 1998. Language in Cognitive Development: The Emergence of the Mediated Mind. Cambridge University Press, New York.
Peters, A., 1983. The Units of Language Acquisition. Cambridge University Press, New York.
Piaget, J., 1954. The Construction of Reality in the Child. Basic Books, New York.
Quine, W., 1960. Word and Object. MIT Press, Cambridge, MA.
Rumelhart, D.E., McClelland, J.L., 1986. Parallel Distributed Processing. MIT Press, Cambridge, MA.
Schafer, G., Plunkett, K., 1998. Rapid word learning by 15-month-olds under tightly controlled conditions. Child Development 69, 309–320.
Smith, L., 1999. Children's noun learning: how general processes make specialized learning mechanisms. In: MacWhinney, B. (Ed.), The Emergence of Language. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 277–304.
Thiessen, E.D., 2009. Statistical learning. In: Bavin, E. (Ed.), Cambridge Handbook of Child Language. Cambridge University Press, Cambridge, pp. 35–50.
Tomasello, M., 2003. Constructing a First Language: A Usage-Based Theory of Language Acquisition. Harvard University Press, Cambridge.
Vihman, M., Croft, W., 2007. Phonological development: toward a "radical" templatic phonology. Linguistics 45, 683–725.
Vygotsky, L., 1934. Thought and Language. MIT Press, Cambridge, MA.
Werker, J., 1995. Exploring developmental changes in cross-language speech perception. In: Gleitman, L., Liberman, M. (Eds.), An Invitation to Cognitive Science. Language, vol. 1. MIT Press, Cambridge, MA, pp. 87–106.

Language and Ethnicity

Günther Schlee, Max Planck Institute for Social Anthropology, Halle (Saale), Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

After defining ethnicity and the ethnic group, this article states that language is usually high up on the list of features most often used as ethnic markers. When it is used as the sole or defining criterion of ethnic identity, the question of the relationship between language and ethnicity becomes tautological. Against this premature identification of linguistic group and ethnic group, cases of ethnic groups are cited which define themselves as such and are perceived as such by their neighbors and do not share one language. Other examples illustrate that there is no clear correlation between feelings of otherness/hostility and level of linguistic difference. The role of language and beliefs about language are discussed in the context of the evolution of the nation-state. Macro-categories like language families are also found to have a variety of roles in identity discourses. Some originally linguistic classifications are used with genocidal intention, although they lack a linguistic or historical foundation by modern academic standards. Written and spoken language can also have different meanings for ethnic identification. After acquiring a second language, communities may continue to use the old alphabet, also for the newly acquired language, or they may use the newly acquired writing system even for their original language.

As 'language' is defined elsewhere in this volume, only 'ethnicity' needs to be defined here before the relationship between the two is discussed.

Ethnicity is a form of collective identity, and thus belongs to the same class of phenomena as religious affiliation, lineage, clan, or class membership. It is the awareness of belonging to an ethnic group, and the belief that others belong to other groups of this kind. Ethnic groups differ from other groups, like voluntary associations or age-sets, by the following criteria: they comprise people of both genders and all ages, or, as Elwert (1989) puts it, entire families. Even where exogamy extends far beyond the nuclear family, ethnic groups comprise enough such exogamous units to guarantee self-sufficiency in biological reproduction: marriage partners can be provided within the ethnic unit. This means that ethnic groups have the potential to recruit members by birth, and, indeed, that is how they recruit most of their members. It does not mean that this is the only form of recruitment and that there is no ethnic conversion.

The criterion of self-sufficiency in biological reproduction sets the lower limit of the size of an ethnic group. The upper limit is set by the criterion of alterity. As ethnic groups define themselves and are defined in contradistinction to other such groups, an ethnic group can never comprise all of humankind. Ethnic groups tend to be medium- to large-scale human aggregates, typically ranging from thousands to millions. Many ethnic groups, including those which do not have a 'nation-state' to their name, exceed many independent states in the size of their populations. 'Ethnic group' and 'nation' form a pair of concepts, the first stressing cultural aspects, and the second political ones. 'Nation' includes the additional element of statehood in one or the other form. One can say that an ethnic group plus a state in which it predominates or a claim to such separate statehood is a nation. The fact that political rights like self-determination are attributed to nations in the political and juridical discourse does not help to define which ethnic group qualifies as a nation and which one does not. Any ethnic group can claim to be a nation and become one through such a claim. The only difference between a mere ethnic group and a nation without state is that the latter claims political sovereignty and the former does not necessarily do so. This distinction may not be very helpful in solving practical political issues, but social science does not provide a more helpful one.

Ethnicity is a cultural construct and its construction materials are cultural as well. There are other cultural constructs which are partly based on biological givens, like gender on biological sex, but in all cases in which ethnicity claims to be based on biological facts like discontinuities in the distributions of genes (race), such claims can be shown to be without a biological basis. Genes may cluster in different ways, but they never mark ethnic boundaries. Ethnic groups are never based on race, but very frequently on beliefs about race. Another element ethnic groups tend to attribute to themselves, often counterfactually, is a venerable age, and therefore ethnicity can in some ways be said to be a pseudo-biological category: ethnic groups view themselves as natural and stable over time like biological species. Other parts of culture, in addition to beliefs about race, which play a part in the construction of ethnic groups, include the reference to real or imagined historical events, peculiarities of customs, and the like. No two ethnic groups have a set of criteria which is alike, but typically ethnic discourses make reference to a plurality of such criteria. If only one such criterion is stressed, like religious affiliation, one would not speak of an ethnic group, but, rather, of a religious one.

For our present topic it is of special importance that language tends to be (but by no means always is) high up in the catalog of cultural criteria used to mark the boundaries between ethnic groups. This makes the question of the relationship between language and ethnicity appears tautological in some cases. Where language is used as the primary marker of an ethnic boundary (Barth, 1969), ethnic and linguistic units tend to be congruent. In post-World War I Europe, the language census has in numerous cases simply replaced the declaration of political will of local communities in effecting boundary corrections between nation-states (Dench, 1986). On the other hand, through the school system and the exclusive use of the majority language for administration, national languages or languages which enjoy a special status in certain areas have

been imposed on linguistic minorities (Gellner, 1981). Congruence between linguistic communities and ethnic units/nations has thus been promoted on the interstate level by adjusting political boundaries to language distribution areas and on the intrastate level by extending a language policy to the boundaries (e.g., Alsace-Lorraine, Silesia). In spite of the fact that these processes have been at work for some time, still only a small fraction of the world's nations is linguistically homogeneous (Ra'anan, 1989).

While in some cases ethnic and linguistic units are coterminous, at the other extreme we find ethnic units, well defined by other criteria, which have nothing to do with language. The Garre of northeastern Africa conceive of themselves as one ethnic group because of their common genealogy and clan organization, their shared pastoral nomadic culture, and their desire, not always fulfilled, to keep pace with each other. The Garre are divided into four linguistic clusters which crosscut other criteria of differentiation like clanship. Some of them speak an Oromo dialect close to the one of the Boran, while some speak Af Rahanweyn, yet others Af Garreh Kofar. The latter two are closely related Somali-like languages but are kept clearly apart by their speakers. There are also Garre who speak Somali proper. Oromo is a different language well beyond comprehension. It belongs to the same lowland branch of the East Cushitic languages as the Somali-type languages, but internal differentiation within this branch is high. The fact that the Garre are also divided between three nation-states (Kenya, Ethiopia, and Somalia) has nothing to do with this linguistic differentiation, since speakers of all four languages are found among the Garre of all three states. The only language which is spoken exclusively by Garre appears to be Af Garreh Kofar, but to the outside observer it is difficult to distinguish that language from Af Rahanweyn which is spoken by hundreds of thousands of non-Garre, namely Somali of the Rahanweyn clan-cluster. It does happen that Garre who do not share one of these Cushitic languages are obliged to converse with each other in languages from totally different language families, like Swahili (Bantu) or English (Germanic) which they have acquired at school, an institution frequented by only a minority of them for mostly short periods.

If there are enough other elements (political, economic, and cultural) on which a feeling of ethnic belonging and commonality is based, such an ethnic unit can tolerate a high degree of linguistic diversity. If these other foundations are absent, no amount of linguistic homogeneity can prevent disruption. Even within a language formerly perceived as one, now three different languages can be distinguished if the political will so dictates. This is exemplified by the split of the Serbo-Croatian language into Serbian, Croatian, and Bosnian.

The close association between a nation and its national language is a fairly recent development of political ideology even in Europe, and more so in other parts of the world to which this idea has been exported. Originally the term mother tongue (*Muttersprache*) referred to the language learned from the mother or wet nurse as distinct from the language of the male sphere, of officialdom and scholarship, i.e., Latin (Ahlzweig, 1994). Through romanticism and the evolving linguistic nationalism this term later acquired political overtones. The modern national states of Europe have, in their formative period, taken little account of the map of language distribution. Personalities claimed in retrospect as forerunners of German, Russian, or other nationalisms, in their real lives often preferred to speak French (e.g., Frederick II of Prussia, Tsar Peter the Great). The ethnicization of the state, its culturalization, and its evolving association with a particular language was a gradual process in the development of modern Europe. Linguistic obstacles were sometimes artificially raised between nations. As their national standard, the Norwegians have chosen from among their many dialects the one most different from Danish.

Not only languages but also language families have been politicized by identifying them with groups of people who were often ascribed racial characteristics, as a result of which they were either accorded privileges, or denied such privileges and even such basic rights as freedom and the right to live. Nowadays nobody outside the lunatic fringes would use the term Semite, to the extent that it is still used at all, in any other sense than speaker of a Semitic language. But there are other such terms of Biblical derivation like Cushites and Hamites which still continue in political careers which were unforeseen by the late eighteenth-, nineteenth-, and twentieth-century European historical linguists who applied them to clusters of languages they found to derive from shared ancestral forms. One can hear Oromo nationalists in Ethiopia arguing for the inclusion of minority groups and their territories into Oromia on the grounds that, although they are not Oromo, they are also Cushites. Ideologies about the repulsion of Hamitic intruders are at the root of the genocide against the Tutsi in Rwanda, although the Tutsi, in fact, speak the same Bantu language as their Hutu murderers, and historical theories that they ever spoke a different language are questionable to say the least. Ironically, too, the term Hamitic is no longer used by linguists. This may serve as a recent example that linguistic classifications, once they cross the boundaries of the academic sphere, can develop a dynamic of their own and that such a dynamic can be quite a murderous one.

Linguistic nationalism often appeals to romantic notions about the relationship between a particular language and other aspects of the culture of its speakers. The language we speak is believed to influence our thinking in a subtle way, giving language a much deeper cultural implication than that of a mere boundary marker between different cultural/ethnic units. Such ideas, which go back to Johann Gottfried Herder (1744–1803), cannot be explained away by pointing to the ways in which they are instrumentalized by nationalists. In Herder's view, a language is the 'collective treasure' of a nation and, since each language shapes the perspective on the world in a different way, the intellectual wealth of mankind consists in the diversity of its languages, i.e., in the number of such treasures (Schlesinger, 1991: p. 12f). Or, in the words of a modern Herderian: "It is ethnic and linguistic diversity that makes life worth living. It is creativity and beauty based upon ethnic and linguistic diversity which makes man human" (Fishman, 1989: p. 15). Very differentiated assumptions about the interrelationship between language and thought have also been put forward by Humboldt (1988), while a more simplistic view, basically a linguistic determinism, was proposed much later by Whorf (1962). Whorfians have also claimed the earlier Sapir as one of their persuasion and speak of a Sapir–Whorf hypothesis (i.e., one's thinking is both enabled and limited by the

structure of the languages we speak). A fuller analysis of the interrelationship of language and thought or language and cognition would go beyond the scope of an article on language and ethnicity (for a summary see Schlesinger, 1991).

The above-mentioned possibilities, that an ethnic group is completely coterminous and identified with the community of speakers of a certain language, and the opposite case, of a total lack of such a congruence and the absence of any association between an ethnic group and the language(s) spoken by it, are two extremes of a range of variation. Between these extremes we find language and ethnicity interacting in a great number of ways. To examine how language relates to ethnicity, and which languages/ethnic identifications persist in which functions in a plurilingual and culturally heterogeneous setting, i.e., to study the ecology of languages, the following variables need to be taken into account:

1. Demographic variables, like the sizes of language communities and their growth rates, and their compact or dispersed form of settlement including the forms of mixture with speakers of other languages;
2. Political and juridical variables, like the status of a language and the institutions in which it is used, and the presence or absence of ethnic stratification of the labor market with which language status often correlates;
3. Ethno-cultural variables, like the association of a language with a group's history or religion, and the potential of a language in terms of range of vocabulary and tradition of writing;
4. Interactional variables like communicational mobility and degree of multilingualism of speakers, specialization of the functions of speech varieties in diglossic and polyglossic settings, and routines of interactions between ethnic groups/speech communities;
5. Variations on topic, to explore the influence of different topics on the choice of a language;
6. Predominant/exclusive use of a language in the public/private sphere, and the linguistic distance between the languages in contact (Haarmann, 1986: p. 11ff).

This list is not meant to be exhaustive.

Some of these variables are illustrated. In contact situations, languages may undergo functional differentiation. The original language in bilingual communities may become restricted to sacral functions. Ethnicity and language choice should not be studied in isolation from other dimensions of social identity like gender and age. Minority people may revert in their old age to the language of their childhood, while their professional lives have been dominated by languages of wider currency (Dow, 1991). The past interaction between languages, their differences in potential and prestige in different semantic domains, and the relationships between the speakers of those languages and the cultural values attached to their ethnic identities are reflected by loanwords. Contrary to the widespread assumption that loanwords are typical for specialized concepts, these often extend far into the basic vocabulary and may designate parts of the body or kinship terms (Haarmann, 1986: p. 186ff; Schlee, 1994). In intensive language contact situations, borrowing is not restricted to the lexicon and there may be light or heavy structural borrowing (Thomason and Kaufman, 1988: p. 74f).

Written language and writing systems show a stronger tendency to be maintained as sacred, and/or symbols of ethnic identity, than does spoken language. Aramaic-speaking Christians became speakers of Arabic, which they also used for writing, and in which they produced a considerable literature, but for a long time they continued to use their earlier Syriac alphabet because they identified the Arabic alphabet with the Qur'an and Islam. The mirror case is provided by Muslims in the post-*Reconquista* Spain, who wrote Spanish in Arabic characters. Jews throughout Europe for centuries wrote the respective majority languages of the countries in which they lived, with a sprinkling of Hebrew words, in Hebrew characters. Where the new alphabet, like the Arabic alphabet in non-Arabic countries converted to Islam, is associated with holy texts, the opposite may occur. Many different languages have been written in the Arabic script, and other writing systems, some of them ancient, have fallen into disuse (Lewis, 1998).

Also linguistic variation within the performance of one speaker in the same speech act can highlight aspects of language and ethnicity. A narrator may switch in a story back and forth between languages and/or more or less standardized/creolized varieties of these languages to depict characters to whom different ethnic and other social identities are attributed. Also, in presenting themselves, polyglot people, or people who master more than one register of a language, can play with options of identification rather than being tied to one form of speech and the ethnicity, or more generally the identity, it connotes (LePage and Tabouret-Keller, 1985).

Switching and the need to choose a language are phenomena which occur only in polyglot settings. Moving from what happens within an individual to the global perspective, we notice that monolingual and polyglot settings are very unevenly distributed, or, put in a more gradualist fashion, that the number of languages per surface area varies a great deal if we move from one region to another. The number of languages is roughly correlated with the level of difference between them. Where we find many languages in the same area, often these represent different language families, or, when they are of the same family, they belong to quite different branches. Combining the number of languages with the levels of difference between them, we can say that zones adjacent to each other can differ a great deal in their linguistic diversity. Often linguistic fragmentation is found in rough territory, typically in mountain ranges. The many and often only quite distantly related Native American languages, which used to be spoken in the west of the United States would fit this pattern. But geography does not explain all of the variation. Mountains and other obstacles to communication seem to favor heterogeneity, and both the Alps and the Caucasus have a high level of diversity. But in the Alps we have just three language families – Romance, Germanic, and Slavic – which are all related to each other on the next higher level as Indo-European languages; and this is nothing in comparison to the levels of diversity we find in the Caucasus. And there are other high-diversity zones like the Lower Omo area of southwestern Ethiopia, which are flat or hilly at the most, and nevertheless have a smaller 'particle size' in everything from languages to clans and ethnic groups. There we find many Nilotic, Omotic, and Cushitic languages numbering only a few thousand speakers each.

In some cases this linguistic diversity is reflected by the political map, whereas in other cases it is not. Where linguistic and political units have a higher level of congruence, the causality behind this phenomenon may have worked in one direction or the other. Linguistic proto-nationalism may cause linguistically homogenous areas to unite politically, or the state may compose one language on its territory (e.g., Modern Greek 'Katharevousa') or develops incentive structures which cause one language to win. Applicants for the civil service in Kazakhstan, for example, have to be fluent in Kazakh, although in practice Russians still predominate in day-to-day administrative business. But this may change with the increasing number of Kazakh speakers in the administration.

But why is there a fragmentation belt with hundreds of languages belonging to three different macro-families across Africa roughly along the Sahel zone and the savannah belt south of it, while across north Africa even before the spread of Arabic only one language family (Afro-Asiatic) was clearly dominant and south of it the comparatively recent spread of Bantu languages covers most of the continent? The conditions which lead to the rise of diversity and its preservation are far from having been comprehensively explored.

Conclusions

Apart from all these riddles about higher and lower degrees of diversity, we also find some universals. One of these universal features of human beings is the capacity to deal with diversity. Since the disappearance of other human species such as the Neanderthal, we have been left alone. There are no other hominids apart from Homo 'sapiens sapiens.' Still, we seem to be biologically equipped to deal with a high degree of internal diversity. (Communication with other species, such as canines and ungulates, in a symbiotic system (pastoralism), would be a different topic, though not a less interesting one.) Stephen Levinson (2006) has proclaimed the existence of a human interaction engine. Starting from observations about 'first contacts' between people who have no knowledge about each other and do not share a language, and from his experience with deaf-mute people, he has come to the conclusion that human beings start to create sign systems instantly and spontaneously. How has this engine evolved? And why do we have the capacity to learn 10 languages or more, and why are we able to translate and interpret? Why do we possess the basic skills for doing anthropology? Nature does not pay for luxuries. These features must have offered an evolutionary advantage. (This human interaction engine must be something on the same level of generality as the 'ability to read and share intentions,' stressed by Tomasello et al. (2005: p. 690). It is one of the prerequisites for the development of languages and language-based complex collaboration. We cannot derive it from language, because it must have been there before.) The game we are studying, the game of identification and reidentification, of learning the features which define friend and foe, of manipulating these delineations, of asserting or hiding who we are, of expressing ourselves in one or the other code, thereby including or excluding people from our audience or defining those to whom we appeal for solidarity and those against whom we make these appeals, making group differences narrower or wider, might be a very old game. Playing this game well often was a matter of survival, and many who did not play well are no longer with us. We have evolved with this game and in response to it. Cultural and linguistic diversity is natural. It has existed for a long time, and we are biologically adapted to it.

See also: Ethnic Identity, Psychology of; Ethnicity and Ethnic Groups: Historical Aspects; Ethnicity, Sociology of; Ethnicity: Anthropological Aspects; Ethnology; Language and Social Class; Linguistic Typology; Writing Systems; Wundt, Wilhelm Maximilian (1832–1920); von Humboldt, Wilhelm (1767–1835).

Bibliography

Alber, Erdmute, 2008. Töchter als Pflegetöchter. Elternschaftskonzepte in Nordbenin, Westafrika. In: Bormann, Lukas, Kügler, Joachim (Eds.), Töchter (Gottes), Studien zum Verhältnis von Kultur, Religion und Geschlecht. [Bayreuther Forum Transit 8]. Lit-Verlag, Berlin, pp. 144–156.

Ahlzweig, G., 1994. Muttersprache–Vaterland: die deutsche Nation und ihre Sprache. Westdeutscher Verlag, Opladen, Germany.

Barth, Fredrik, 1959. Political Leadership among the Swat Pathans. Athlone Press, London.

Barth, Fredrik (Ed.), 1969. Ethnic Groups and Boundaries. Allen and Unwin, London.

Comaroff, Jean, Comaroff, John L., 2009. Reflections on the anthropology of law, governance and sovereignity. In: von Benda-Beckmann, Franz, von Benda-Beckmann, Keebet, Eckert, Julia (Eds.), Rules of the Law and Laws of Ruling. Ashgate, Farnham, pp. 31–59.

Coope, Robert L., Spolsky, Bernard (Eds.), 1991. The Influence of Language on Culture and Thought. Mouton de Gruyter, Berlin.

Dench, Geoff, 1986. Minorities in the Open Society: Prisoners of Ambivalence. Routledge and Kegan Paul, London.

Donahoe, Brian, Eidson, John, Feyissa, Dereje, Fuest, Veronika, Hoehne, Markus, Nieswand, Boris, Schlee, Günther, Zenker, Olaf, 2009. The Formation and Mobilization of Collective Identities in Situations of Conflict and Integration. Max Planck Institute for Social Anthropology Working Paper No. 116. Max Planck Institute for Social Anthropology, Halle/Saale.

Dow, James R. (Ed.), 1991. Language and Ethnicity. Benjamins, Amsterdam.

Elwert, Georg, 1989. Nationalismus und Ethnizität. Über die Bildung von Wir-Gruppen. Kölner Zeitschrift für Soziologie und Sozialpsychologie 3, 440–464.

Enfield, N.J., Levinson, C., Stephen, 2006. Introduction: human sociality as a new interdisciplinary field. In: Enfield, N.J., Levinson, Stephen, C. (Eds.), Roots of Human Sociality: Culture, Cognition and Interaction. Oxford, Berghahn Books, New York, Oxford, pp. 1–35.

Esser, Hartmut, 11 February 1999. Inklusion und Exklusion. Lecture at Bielefeld University.

Fardon, Richard, Furniss, Graham (Eds.), 1994. African Languages, Development, and the State. Routledge, London.

Feyissa, Dereje, 2011. Playing Different Games: The Paradox of Anywaa and Nuer Identification Strategies in the Gambella Region, Ethiopia. Berghahn Books, New York, Oxford.

Fishman, Joshua A., 1989. Language and Ethnicity in Minority Sociolinguistic Perspective. Multilingual, Clevedon, PA.

Gellner, Ernest, 1981. Muslim Society. Cambridge University Press, Cambridge, UK.

Giddens, Anthony, 1979. New Rules of Sociological Method: A Positive Critique of Interpretative Sociologies. Hutchinson, London.

Gigerenzer, Gerd, 2007. Gut Feelings: The Intelligence of the Unconscious. Viking, New York.

Girke, Felix, 2009. The Ädamo of the Kara: Rhetoric in Social Relations on the Lower Omo. Doctoral thesis. Martin-Luther-Universität Halle-Wittenberg, Halle.

Goody, Jack, 1976. Production and Reproduction: A Comparative Study of the Domestic Domain. Cambridge University Press, Cambridge.

Haarmann, Harald, 1986. Language in Ethnicity. A View of Basic Ecological Relations. Mouton de Gruyter, Berlin.

Hann, Chris (Ed.), 1994. When History Accelerates: Essays on Rapid Social Change, Complexity and Creativity. Athlone, London.

Hechter, Michael, 1987. Principles of Group Solidarity. University of California Press, Berkeley.

Humboldt, Wilhelm Von, 1988. On Language: The Diversity of Human Language Structure and its Influence on the Mental Development of Mankind. Cambridge University Press, Cambridge, UK.

LePage, R.B., Tabouret-Keller, Andree, 1985. Acts of Identity: Creole Based Approaches to Language and Ethnicity. Cambridge University Press, Cambridge, UK.

Levinson, Stephen C., 2006. On the human "interaction engine". In: Enfield, N.J., Levinson, Stephen C. (Eds.), Roots of Human Sociality: Culture, Cognition and Interaction. Berghahn Books, New York, Oxford, pp. 39–69.

Lewis, Bernard, 1998. The Multiple Identities of the Middle East. Weidenfeld and Nicholson, London.

Montville, Joseph V. (Ed.), 1989. Conflict and Peacemaking in Multiethnic Societies. Lexington Books, Lexington, MA.

Nieswand, Boris, 2007. Ghanaian Migrants in Germany and the Status Paradox of Migration: A Multi-sited Ethnography of Transnational Pathways of Migrant Inclusion. Doctoral thesis. Martin-Luther-Universität Halle-Wittenberg, Halle.

Ra'anan, Uri, 1989. The nation-state fallacy. In: Montville, Joseph V. (Ed.), Conflict and Peacemaking in Multiethnic Societies. Lexington Books, Lexington, MA, pp. 5–20.

Riker, William H., 1962. The Theory of Political Coalitions. Yale University Press, New Haven.

Sahlins, Marshal, 1961. The segmentary lineage: an organization of predatory expansions. American Anthropologist 63, 322–345.

Schlee, Günther, 1989. Identities on the Move: Clanship and Pastoralism in Northern Kenya. Manchester University Press, Manchester.

Schlee, Günther, 1994. Loanwords in Oromo and Rendille as a mirror of past interethnic relations. In: Fardon, Richard, Furniss, Graham (Eds.), African Languages, Development, and the State. Routledge, London, pp. 191–212.

Schlee, Günther, 2003. Identification in violent settings and situations of rapid change. Africa 73 (3), 333–342.

Schlee, Günther, 2008. How Enemies Are Made: Towards a Theory of Ethnic and Religious Conflicts. Berghahn Books, New York, Oxford.

Schlee, Günther, Heady, Patrick, 2010. Terminology and practice: European kinship in a world-wide perspective. In: Heady, Patrick, Kohli, Martin (Eds.), Kinship and Social Security in Contemporary Europe, vol. 3, pp. 347–374.

Schlesinger, I.M., 1991. The wax and wane of Whorfian views. In: Cooper, Robert L., Spolsky, Bernard (Eds.), The Influence of Language on Culture and Thought. Mouton de Gruyter, Berlin.

Stiglitz, Joseph E., 2002. Information and the change in the paradigm in economics. The American Economic Review 92 (3), 446–501.

Tomasello, Michael, Carpenter, Malinda, Call, Josep, Gehne, Tanya, Moll, Henrike, 2005. Understanding and sharing intentions: the origins of cultural cognition. Behavioral and Brain Sciences 28, 675–735.

Thomason, Sarah G., Kaufman, Terrence, 1988. Language Contact, Creolization, and Genetic Linguistics. University of California Press, Berkeley, CA.

Whorf, Benjamin L., 1962. Language, Thought, and Reality: Selected Writings. Wiley, New York.

Language and Gender

Sarah Gormley, University of Manchester, Manchester, UK; and University of Birmingham, Birmingham, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Coates, volume 12, pp. 8288–8292, © 2001, Elsevier Ltd.

Abstract

Gender is an important construct in all societies, involving the binary distinction male–female, and divisions made on the basis of what it is to be a woman or a man, and what is taken to be masculine or feminine. Language plays a key role in both constructing and sustaining these divisions. The study of language and gender has produced a body of diverse explanations for gender differentiation in language use. While early work on language and gender employed quantitative methods and categorized speakers according to biological sex, within recent research qualitative, ethnographic approaches have predominated as the field has witnessed a shift from viewing gender as a fixed social category to theorizing gender as a dynamic, performative social construct.

Gender is an important construct in all societies, involving the binary distinction male–female, and divisions made on the basis of what it is to be a woman or a man, and what is taken to be masculine or feminine. Language plays a key role in both constructing and sustaining these divisions. The study of language and gender has produced a body of diverse explanations for gender differentiation in language use. These differing, even conflicting, explanations reflect the theoretical developments in the discipline, as a shift has taken place from viewing gender as a fixed social category to theorizing gender as a dynamic, performative social construct. It is through tracing these developments and their corresponding explanations for how women and men speak that this article addresses the interplay between language and gender.

Sociolinguistic Variationist Studies

Early work on language and gender emerged from the sociolinguistic investigations of language variation in the 1960s and 1970s. Labov's (1966, 1972) social stratification studies examined particular linguistic variables in accordance with the social class and particular geographical location of their informants, assessing which speakers used the most prestigious, 'standard,' forms and which speakers used 'nonstandard' vernacular forms. What came to light from these studies was that as well as class, gender was affecting phonological and grammatical patterning. Labov found that in each social class, women used prestige forms more than men, and that this was particularly telling in the lower middle-class grouping. Identifying what he terms hypercorrection, or the overuse of prestige forms by the female speakers in that group, Labov argued that the lower middle-class females were insecure about their status and therefore hypercorrect in order to gain prestige. For the male speakers, Labov argued that there appeared to be connotations of masculinity associated with nonstandard forms in the male working-class group, suggesting that covert prestige was associated with the use of vernacular forms by the working-class male speakers. This notion of overt and covert prestige was investigated by Trudgill (1972) who reported that, like the female speakers of Labov's New York studies, the female speakers in his study in Norwich in the UK tended to use nonstandard, vernacular forms less frequently than the male speakers. Trudgill argued that these differences in language result from status-consciousness, with women being more status-conscious than men. According to Trudgill, women are more status-conscious because their social position is less secure than men, and therefore signaling one's social status linguistically may be more important for women.

Alternative explanations for gender differences in language use emerged from social network theory (Milroy and Milroy, 1978; Milroy, 1980). From this approach, the differences between speakers who use more prestige forms and those who use more vernacular forms result from group membership. Social network theory suggests that rather than social status, it is group pressures from within the communities in which individuals live and work that have the power to enforce linguistic norms, but that these vary among communities since individual speakers have differing amounts of social ties within a particular group; for example, in some communities it may be that men have a wide range of social contacts whereas women have fewer social contacts, while in other communities men may have more restricted social contacts than women. A further explanation for gendered differentiation in language use derives from what has been termed a market forces view (Talbot, 2010). From this perspective, economic factors are seen as the crucial determining factor which affects use of the prestigious or vernacular variants. Nichols (1998), for example, demonstrates that gender differences in employment require some speakers, but not others, to use more standard and less vernacular variants. The text often regarded as the foundation for the discipline of language and gender studies, however, is Lakoff's (1975) *Language and Woman's Place*. Lakoff's work heralds the first key development in the discipline: the approach categorized by the term the deficit approach (Talbot, 2010).

Deficit and Dominance

Lakoff (1975) claims that there is a distinct 'women's language.' According to Lakoff, as a result of sex-role socialization, where girls and boys learn to talk in ways appropriate to their gender,

females are socialized into using inferior language characterized by weakness, uncertainty, and excessive politeness. The use of such deficient language, Lakoff argues, disadvantages women and subjects them to linguistic discrimination. The features Lakoff identifies as characterizing women's language include lexical items, discourse particles, and intonation patterns. Lexical items comprise precise color terms such as teal green, affective adjectives such as adorable, and superpolite forms such as the avoidance of expletives and the use of euphemism. Discourse particles and intonation patterns include hedges and tag questions which mitigate the force of an utterance, rising intonation and emphatic stress. While influential in drawing attention to gender and language use, Lakoff's work has been robustly criticized. Methodologically, for example, Lakoff's study is flawed since her observations are based on anecdote, taken from her own observations with no recorded empirical evidence to support her assertions. Furthermore, a simple one-to-one correlation between form and function is problematic and easily challenged; Holmes (1984), for example, demonstrates that tag questions have different functions, including the expression of solidarity. However, the most fundamental problem with Lakoff's approach is that in characterizing women's speech as deficient, men's language is positioned as the norm. Despite its flaws, however, *Language and Women's Place* set the agenda for language and gender studies, and it is in response to this work that the second key development in the discipline emerged: the dominance approach.

The dominance approach to language and gender is typified by the work of Zimmerman and West (1975) and West (1984, 1998). Unlike the deficit approach, the dominance approach does not take the view that women's speech is deficient; instead women are viewed as oppressed by patriarchy. From a dominance approach, it is the significant social, economic, and political power that men have which results in women being dominated. Zimmerman and West (1975) recorded single-sex and mixed-sex interactions, examining elements of the sequential organization of conversation including turn-taking, interruptions, overlaps, and silences. They found that in single-sex interactions interruptions were uncommon, however in male–female interactions, interruptions were frequent, with men responsible for the overwhelming majority of interruptions. Zimmerman and West concluded that men dominate in mixed-sex interaction by breaching women's right to the conversational floor. In a later study, West (1984, 1998) continued the focus upon interruptions, but the empirical data in this study were taken from dyadic doctor–patient interactions. West's findings were that in the encounters between male physicians and their patients, the majority of the interruptions were physician initiated, however, in the encounters between female physicians and their patients, the majority of the interruptions were patient initiated. West concluded that the conventional asymmetry between doctor and patient interaction is reversed when female physicians are involved, suggesting that gender can take primacy over status. West and Zimmerman and West thus see interruptions as a way of enacting power, viewing asymmetrical patterns in women's and men's language use as enactments of male privilege. There have, however, been several criticisms leveled at the dominance approach. Researchers came to view the dominance model as becoming a deficit model, portraying men's speech as the norm and women's speech as deviant. Furthermore, the model of patriarchal power came to be seen as problematic, implying that all men oppress all women. This notion of patriarchal power effectively categorizes women and men into two homogenous groups one of which is powerful and the other powerless, and therefore fails to take into account variation in groups of men and groups of women. The most central reaction against the dominance approach, however, arose as researchers became increasingly unhappy with the negative portrayal of women perpetuated by the dominance approach. The third key development in language and gender studies, the difference approach, emerged from this dissatisfaction.

The Difference Approach

The difference approach to language and gender became popular in the 1980s and 1990s, and is considered to have originated with the work of Maltz and Borker ([1982]1998). The focus of Maltz and Borker's work is on the problems of miscommunication that they suggest existing research shows when women and men interact. Their approach to the issue of miscommunication was heavily influenced by Gumperz's (1982a,b) work on interethnic communication. Gumperz argued that if cultural backgrounds differ, then miscommunication will occur because individuals will be drawing upon divergent uses and interpretations of linguistic strategies. Maltz and Borker applied this conceptualization of miscommunication to mixed-sex interaction. According to Maltz and Borker, girls and boys are socialized into two different subcultures, and it is within these respective subcultures that girls and boys learn to become competent speakers. Since girls and boys will thus develop different communicative norms, mixed-sex interaction should be seen as cross-cultural communication and miscommunication will therefore frequently occur. According to this view, women and men characteristically draw on different linguistic strategies: while women orient themselves to the person they are talking to such as through the use of personal and inclusive pronouns and actively seeking and giving back-channel support, men's speech is characterized by verbal posturing, capturing an audience through storytelling and arguing in order to assert dominance.

Researchers working on single-sex interaction have found the difference model particularly useful. Coates' (1988, 1996) work on all-female interaction provides a prime example of the difference approach. Coates argues that one of the key distinctions between male and female communicative styles is that men's conversation style is based on competitiveness whereas women's conversational style is based on cooperativeness, and her work on all-female interaction identifies numerous examples of this cooperativeness. According to Coates, the difference approach is necessary when analyzing single-sex interaction, since examining the speech of groups of all female speakers as simply the talk of an oppressed group misses the complexities and the strengths of women's talk, and makes it particularly difficult to positively reevaluate female speech styles.

Tannen's (1990) work on the difference between female and male linguistic performance is based on Maltz and Borker's two cultures theory. According to Tannen, in mixed-sex interaction individuals interpret the linguistic performance of others according to their own subcultural norms. As men and women thus evaluate utterances in disparate ways, miscommunication occurs. Tannen puts forward a number of dichotomies in her explanation of these communicative difficulties. Such dichotomies include status versus connection, whereby through engaging in conversation men expect to establish status in a hierarchy whereas women seek to establish connection and intimacy, and report talk versus rapport talk, where men are concerned with the exchange of information during talk, while women are concerned with building rapport. In order to understand each other better and to create more harmonious relationships, Tannen argues that women and men simply need to learn each other's subcultural linguistic norms. Tannen's book has, however, been widely criticized for its lack of attention to power differences. Cameron (1998), for example, points out that Tannen's approach is based on the notion that change will simply occur if men and women learn to understand the differences in each other's linguistic behavior, but, she argues, this fails to acknowledge that power roles are constitutive of gender differentiation. Although the difference approach has been valuable for language and gender research as it allows for a positive reevaluation of the speech styles that have been associated with women, in recent years the way in which gender has been conceptualized within the field has involved moving away from the binary opposition of male and female and seeing difference as the major goal of language and gender research.

The Dynamic Approach

The most recent approach to language and gender views gender as a dynamic, active concept. Rather than perceiving individuals as being of a particular gender, individuals are perceived of as doing gender. This shift from seeing gender as a verb rather than a noun moves analysis on from seeing gender as something that is simply learned during childhood socialization to a perspective from which gender is a constant process of performance.

Performativity

Butler's (1990, 1993) work on gender and performativity has been particularly influential in recent language and gender studies. According to Butler, gender is a repeated performance: "[t]he materiality of sex is constructed through a ritualised repetition of norms" (1993, p. x) However, this performance does not mean that one can perform anything one wishes, since, as Butler argues, "a wilful and instrumental subject, one who decides on its gender is clearly not its gender from the start and fails to realise that its existence is already decided *by* gender" (1993, p. x). Gender as a performative social construct, or the repeated performance of a range of behaviors associated with a particular sex, is therefore, for Butler, a constant process, as it is with preexisting gender norms that individuals negotiate. This move from binary oppositions to a more sophisticated conceptualization of gender has led to more nuanced concerns with the ways that individuals reaffirm, negotiate with, and contest what they hypothesize as appropriate behavior in particular contexts. The notion of a community of practice (henceforth CofP) has been particularly important for language and gender researchers in thinking through the concept of individuals negotiating, confirming, and attesting practices and interpretations in local contexts, developing a sense of their own gendered identity.

Communities of Practice

The CofP approach was first introduced into language and gender studies by Eckert and McConnell-Ginet (1992a,b). Eckert and McConnell-Ginet define a CofP as a group of people who are brought together through a joint engagement on a task and who, in so doing, jointly construct a range of speech styles, ways of interacting, and shared meanings. From a CofP perspective, individuals belong to multiple and wide-ranging CofPs, for example, at home, at work, in one's hobbies and interests, and therefore an individual's gender identity is fluid and variable, produced and reproduced during social interaction within these different CofPs. Indeed, in much work which takes the CofP as its methodological approach, the construction of identity as plural and variable has been a central analytical focus (for example, Bucholtz, 1999; Eckert, 2000; Mendoza-Denton, 2008). The CofP model allows for variations in women's and men's speech rather than making global, homogenizing statements, as there has been an observable move from large-scale quantitative studies such as those conducted by Labov to qualitative, ethnographic studies.

Indirect Indexicality

Many studies of language and gender which focus upon interaction have found the notion of indirect indexicality (Ochs, 1992) useful for overcoming the problematic correlation of linguistic form, function, and the sex of the speaker evidenced in earlier approaches (e.g., Kiesling, 2003; Mullany, 2007). In her work on indexicality, Ochs (1992) argues that the relation between language and gender is mediated through a web of social meanings. Direct indexicality is defined as language use where gender is explicitly encoded by linguistic items such as *Mr*, *Miss*, *boy*, *girl*, *man*, *woman*. Indirect indexicality is defined as language use where particular linguistic styles are encoded with stereotypically gendered meanings. From this perspective, the association of women with tag questions, for example, or the expectation that women are more linguistically polite than men operates only as *ideological* expectations.

Discourses of Gender/Gendered Discourses

Poststructuralist theory has been heavily influential in language and gender scholarship, most notably the work of Michel Foucault (1972, 1979, 1981) and his conceptualization of discourse as social practice. According to Foucault, discourses are "practices that systematically form the objects of which they speak" (1972, p. 49). On this view, Cameron and Kulick (2003, p. 16) argue, discourses are to some extent

"language-dependent practices of definition, classification, explanation, and justification"; in other words, socially constructed bodies of knowledge. As Mills (2004) points out, a discursive structure can be identified by the stability and regularity of the ideas, opinions, concepts, ways of thinking and behaving which are formed within a particular context and that act as a form of social control. According to Foucault, then, discourses structure the individual's sense of reality, of the world, of oneself, and of others, and therefore the ordering human beings impose on the world is through the categorization and interpretation of objects and experiences according to the discursive structures available to them. In terms of gender, Mills (2004) points out that discourses of femininity and masculinity demarcate the boundaries within which women and men perform their gender identities. Influenced by Foucault's conceptualization of discourse, Coates (1997, p. 311), for example, argues that there are competing discourses of femininity in circulation, she argues "there is no single unified way of doing femininity, of being a woman …different discourses give us access to different femininities…we are unwittingly involved in the ceaseless struggle to define gender." Taking the view that there is a potential multiplicity of femininities and masculinities that individuals take up, reject, or negotiate with has allowed researchers to produce more sophisticated analyses of how gender is constructed in both spoken and written data.

New Directions

The CofP model continues to be a popular approach to the study of language and gender in the twenty-first century. However, Mills and Mullany (2011) have pointed out that at present few studies relate the norms established within the particular CofP under study to wider social norms, therefore future research is needed which attempts to relate small-scale, local studies to their wider social context. An even wider, international perspective is also being developed within the field. Not only is research widening to non-English-speaking communities, but also in her work on language, gender, and globalization, McElhinny (2007) highlights the necessity for language and gender research to attend more thoroughly to the study of the articulation of language, gender, and various aspects of global economy. As the construction and negotiation of gender and the central role language plays in this construction become increasingly more complex and multifaceted, language and gender research continues to develop more nuanced approaches to the study of the interplay between language and gender.

See also: Discourse and Indentity; Language and Ethnicity; Language and Social Class; Linguistic Anthropology; Linguistic Fieldwork; Politeness and Language; Quantitative Linguistics; Sociolinguistics.

Bibliography

Bucholtz, M., 1999. Why be normal? Language and identity practices in a community of nerd girls. Language and Society 28 (2), 203–225.

Butler, J., 1993. Bodies that Matter: On the Discursive Limits of "Sex". Routledge, London.

Butler, J., 1990. Gender Trouble: Feminism and the Subversion of Identity. Routledge, New York.

Cameron, D., 1998. "Is there any ketchup, Vera?": Gender, power and pragmatics. Discourse and Society 9 (4), 435–455.

Cameron, D., Kulick, D., 2003. Language and Sexuality. Cambridge University Press, Cambridge.

Coates, J., 1988. Gossip revisited. In: Coates, J., Cameron, D. (Eds.), Women in Their Speech Communities. Longman, London, pp. 94–122.

Coates, J., 1997. Competing discourses of femininity. In: Kotthoff, H., Wodak, R. (Eds.), Communicating Gender in Context. Benjamins, Amsterdam, pp. 285–313.

Coates, J., 1996. Women Talk. Oxford, Blackwell.

Eckert, P., 2000. Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High. Oxford, Blackwell.

Eckert, P., McConnell-Ginet, S., 1992a. Communities of practice: where language, gender and power all live. In: Hall, K., Bucholtz, M., Moonwomon, B. (Eds.), Locating Power: Proceedings of the Second Berkeley Women and Language Conference. Berkeley Women and Language Group, Berkeley, CA, pp. 89–99.

Eckert, P., McConnell-Ginet, S., 1992b. Think practically and look locally: language and gender as community-based practice. Annual Review of Anthropology 21, 461–490.

Foucault, M., 1981. The order of discourse. In: Young, R. (Ed.), Untying the Text: A Post-Structural Anthology. Routledge & Kegan Paul, Boston.

Foucault, M., 1979. Discipline and Punish: The Birth of the Prison. Penguin, Harmondsworth.

Foucault, M., 1972. The Archaeology of Knowledge. Tavistock, London.

Gumperz, J., 1982a. Discourse Strategies. Cambridge University Press, Cambridge.

Gumperz, J., 1982b. Language and Social Identity. Cambridge University Press, Cambridge.

Holmes, J., 1984. Women's language: a functional approach. General Linguistics 24 (3), 149–178.

Kiesling, S., 2003. Prestige, cultural models, and other ways of thinking about underlying norms and gender. In: Holmes, J., Meyerhoff, M. (Eds.), The Handbook of Language and Gender. Oxford, Blackwell, pp. 509–527.

Labov, W., 1966. The Social Stratification of English in New York City. CAL, Washington.

Labov, W., 1972. Sociolinguistic Patterns. University of Pennsylvania Press, Philadelphia, PA.

Lakoff, R., 1975. Language and Woman's Place. Harper and Row, New York.

Maltz, D.N., Borker, R.A., [1982]1998. A cultural approach to male-female miscommunication. In: Coates, J., (Ed.), Language and Gender: A Reader. Blackwell, London, 417–434.

McElhinny, B., 2007. Words, Worlds and Material Girls: Language, Gender, Globalization. Mouton De Gruyter, Berlin, New York.

Mendoza-Denton, N., 2008. Homegirls: Language and Cultural Practice Among Latina GirlGangs. Blackwell, Malden.

Mills, S., 2004. Discourse, second ed. Routledge, London.

Mills, S., Mullany, L., 2011. Language, Gender and Feminism. Routledge, London.

Milroy, J., Milroy, L., 1978. Belfast: change and variation in an urban vernacular. In: Trudgill, P. (Ed.), Sociolinguistic Patterns in British English. Arnold, London, pp. 19–36.

Milroy, L., 1980. Language and Social Networks. Blackwell, Oxford.

Mullany, L., 2007. Gendered Discourse in the Professional Workplace. Palgrave Macmillan, Basingstoke.

Nichols, P., 1998. Black women in the rural south: conservative and innovative. In: Coates, J. (Ed.), Language and Gender: A Reader. Blackwell, Oxford, pp. 55–63.

Ochs, E., 1992. Indexing gender. In: Duranti, A., Goodwin, C. (Eds.), Rethinking Context: Language as an Interactive Phenomenon. Cambridge University Press, Cambridge, pp. 335–358.

Talbot, M., 2010. Language and Gender, second ed. Polity, Cambridge.

Tannen, D., 1990. You Just Don't Understand: Women and Men in Conversation. William Morrow, New York.

Trudgill, P., 1972. Sex, covert prestige and linguistic change in urban British English of Norwich. Language in Society 1, 179–195.

West, C., 1998. When the doctor is a 'Lady': power, status and gender in physician-patient encounters. In: Coates, J. (Ed.), Language and Gender: A Reader. Blackwell, Oxford, pp. 396–412.

Zimmerman, D., West, C., 1975. Sex-roles, interruptions and silences in conversation. In: Thorne, B., Kramarae, C., Henley, N. (Eds.), Language, Gender and Society. Newbury House, Rowley, pp. 89–101.

Language and Labor Market Success

Ingo E Isphording, IZA – Institute for the Study of Labor, Bonn, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article summarizes three different strands of the literature that address the labor market effects of language-related human capital: (1) A general importance is demonstrated in the empirical evidence on earnings and employment effects of literacy as the ability to productively use written information. Significant effects are found for developed and developing countries, leading on an aggregated macrolevel to a positive relationship between literacy and economic development; (2) The ongoing globalization leads to an increased demand for foreign-language proficiency to reduce search and information costs and overcome cultural barriers in the trade of services and goods and tourism. Against the background of scarce skill supply, employers are willing to pay significant wage premia, especially for global and local lingua franca; (3) For international migrants, destination language skills display both a prerequisite for and outcome of successful integration. Investments into destination language skills are highly rewarded by wage returns and higher employment probabilities and act as the medium of translation to apply premigration human capital in the destination country labor market.

Introduction

Proficiency in a language is one of the most basic parts of an individual's human capital portfolio. The productive functions of language are manifold, in both the labor market and everyday life. Language is used to express one's emotional status, providing the counterpart with necessary information on the otherwise unobservable inner state of mind, as well as expressing one's own identity and demands. Language is used as the main medium of social interaction. While this interaction does not necessarily contain any information conveyed, it provides the basic conditions for a productive environment. Language is the medium used for the record-keeping, organization, and processing of information and facts. Finally, the concept of linguistic determinism in linguistics states that language determines the way in which individuals are able to think and structure their own thoughts (Crystal, 2010).

The importance of language skills in the labor market has been growing during recent decades, in a process that is likely to continue for some time. The reduced demand for manual labor and the ongoing transformation of the labor market toward increasingly information-based production have drastically increased the need for language and literacy skills. Especially with the emergence and diffusion of information and communications technology usage into any niche of the labor market, there are hardly any occupations do not require a minimum of speaking, writing, and reading skills.

Possessing a sufficient level of literacy and language proficiency is far from being self-evident. Literacy rates across countries differ substantially, with an even greater heterogeneity within countries. Literacy and language proficiency are crucially related to socioeconomic status, education, health, and nativity. This interrelationship between an individual's socioeconomic environment and the language-related human capital introduces a large heterogeneity in language skills. Given the productive value of language skills, this heterogeneity is likely to be translated into labor market disadvantages for individuals possessing insufficient levels of literacy and language proficiency.

This article aims to provide an overview of the empirical evidence on how literacy and language skills depend on socioeconomic characteristics and affect individual labor market outcomes. The empirical evidence is summarized via three different angles that have gained attention in the literature: Section Literacy and Productivity deals the importance of language in general, as the main medium of communication, recording and storing of information, and the expression of demands; Section Foreign Languages in the Labor Market summarizes the literature on the importance of foreign languages in modern interdependent economies; and Section Destination Language Skills of Migrants provides an overview on the literature on the importance of the destination language acquisition of immigrants for the economic and societal integration. Finally, Section Conclusion concludes.

Literacy and Productivity

As the main medium of everyday communication, language plays a naturally important role in shaping the economic success of an individual, although the analysis of determinants and labor market advantages of skills in one's own mother tongue lacks a natural theoretical framework and is ultimately data-driven. Language skills in one's mother tongue affect the productivity in the work place in the form of literacy. UNESCO defines literacy as:

> the ability to identify, understand, interpret, create, communicate and compute, using printed and written materials associated with varying contexts. Literacy involves a continuum of learning in enabling individuals to achieve his or her goals, develop his or her knowledge and potential and participate fully in community and wider society.
>
> UNESCO, 2004

Literacy therefore represents the productive functions of one's language ability. In recent years, the data availability on the relationship between adult literacy and labor market

outcomes has improved substantially. In particular, the OECD has pushed forward data collection on literacy skills, with three subsequent international surveys now collecting literacy scores based on test booklets, as well as comprehensive information on socioeconomic living conditions: the International Adult Literacy Survey (IALS) conducted in 1994–98, the Adult Literacy and Lifeskills (ALL) Survey conducted in 2003–08, and the Program for the International Assessment of Adult Competencies (PIAAC) conducted in 2012. These data sets provide the basis for a range of studies analyzing the determinants and labor market effects of adult literacy.

The evidence suggests that basic literacy skills are associated with distinctive wage premia. McIntosh and Vignoles (2001) provide estimates for positive wage and employment effects for UK, noting that around 20% of the British population had severe literacy difficulties in 1999. These literacy difficulties are associated with a sizable wage penalty of about 15%. Using Canadian data, Charette and Meng (1998) show that literacy has a significant positive effect on wages, with their findings indicating that literacy is closely associated to formal education. Nonetheless, the wage effect goes beyond the returns of formal education, with literacy skills providing additional information in explaining wage differentials.

The relationship between literacy and economic well-being holds specific importance for developing newly industrialized countries, given that literacy is crucially dependent on a well-developed and functioning education system. Therefore, developing countries might display significantly lower-literacy rates than developed countries, although appropriate data are scarce in these countries. de Baldini Rocha et al. (2011) analyze the effect of literacy on earnings in Brazil, finding both significant employment and earnings effects of becoming literate. This evidence on the microlevel is complementary to macro-evidence on the relationship between literacy rates and the economic development of regions and countries. Aggregated literacy rates provide insights into the human capital stock of a country. Coulombe and Tremblay (2006) use data from the IALS study to explain the growth of 14 OECD countries by the literacy of labor market entrants. They show that literacy has a significant positive effect on growth indicators and conclude that raising the average level of literacy (e.g., by investing into women's education) is more important than focusing on high levels for restricted groups. Hanushek and Woessmann (2008) provide an overview of the relationship between cognitive skills (including literacy), individual labor market outcomes and economic growth, highlighting the importance of educational policy measures for countries' development. Following this importance for the economic development of countries, literacy has entered the UN Human Development Index as a main dimension of measurement.

Foreign Languages in the Labor Market

In an increasingly interconnected globalized world, international communication plays a growing and crucial role as a prerequisite for international factor movements, international trade, foreign direct investments, financial flows, and human migration. This leads to a worldwide growing demand for foreign language proficiency that is not always met by skill supply.

The acquisition of reading and speaking skills in a foreign language can be modeled as the economic decision weighing off the expected benefits and costs of speaking a new language. Formalized models are provided by Selten and Pool (1991) and Church and King (1993). The expected benefits of speaking a new language combine utility gains through being able to communicate in a certain language. The benefits and incentives of learning a new language can be divided into internal and external motivations. Internal motivations mainly refer to the nonmarket value of languages. Speaking a new language offers potential utility gains through easier access to new information and an increased population of potential communication partners. Furthermore, it also offers increased reputation, prestige, and recognition among peers. For some, the learning process itself might lead to utility gains through being a joyful process.

From an economic perspective, this nonmarket value of languages is accompanied by a market value related to the relative importance of a language to a country. This relative value is dependent on a range of factors: it is shaped by the applicability of a language in direct communication with foreign customers in tourism and services. Its value is related to its potential in international trade to reduce search, translating information costs when dealing with foreign information. In particular, the latter role of languages in shaping international trade flows has gained significant attention in the economic literature. Although English acts as lingua franca in modern international trade, knowledge in the trade partner's original language remains a crucial variable, especially in the development of long-term business partnerships and overcoming further cultural hurdles. Further, lower-linguistic barriers ease the promotion of trade and commerce by immigrants through easier access to the destination country, as discussed later in Destination Language Skills of Migrants. A range of studies proxies these linguistic barriers by summary statistics of linguistic distance between two trade partners, using grammatical information (Lohmann, 2011) or information on pronunciation of word lists (Isphording and Otten, 2013). Common findings of these studies are significantly reduced trade flows by higher-linguistic hurdles.

On the microlevel, language barriers are translated into expected benefits of learning a new language by the willingness of employers to pay for foreign language proficiency. Individuals have to weigh these potential utility gains from learning a new language against the costs associated to the language acquisition. These costs come in terms of direct monetary costs for language classes and indirectly by spent effort and foregone wages. The size of the costs is related to a range of correlates. Apart from a fundamental and unobservable heterogeneity in the ability to learn languages, the difficulty of learning a new language is influenced by the dissimilarity between the first language and the acquired language. Dissimilarities between languages raise large barriers in the acquisition process. While it might be comparably easy to learn a language from one's own language family (e.g., German, English, and Swedish from the Germanic language family, or Romanian, Spanish, and Italian sharing the Romance language family), the acquisition becomes much more difficult for linguistically distant languages, e.g., for a European attempting to learn Mandarin or Japanese. Isphording and Otten (2013) provide an extensive

analysis of the role of linguistic borders in language acquisition. Another crucial determinant of the costs of language acquisition is the age at which the language acquisition takes place. Linguistic and neurobiological research has identified a crucial 'critical period' in early adolescence until when language acquisition takes place almost effortlessly. The ability to acquire new languages subsequently decreases, in a process that continues into adulthood. As second-language education is typically associated with secondary schooling, most of the foreign-language acquisition takes place after this critical period. Acquiring a new language in adulthood faces an even more reduced ability to learn.

A rational and utility-maximizing individual will weigh these costs of language acquisition against expected benefits of speaking the new language. Ginsburgh and Prieto-Rodriguez (2011) provide an overview on returns to foreign languages in the EU. Their results indicate distinctive wage premia paid for foreign languages, with a strong heterogeneity across different countries and the wage distribution. Knowledge of foreign languages mainly pays off in high-skilled occupations. While English as the main lingua franca is rewarded in any country, in Southern European countries it competes in importance with Romance languages, which seem to play the role of local lingua franca, such as French in Spain. This importance is backed up by Isphording (2013), who estimates large returns to foreign-language skills within a sample of immigrants in Spain. Again, the highest returns are paid for English skills, closely followed by the returns to French skills.

Furthermore, Williams (2011) finds significant positive returns to foreign-language skills in Western European countries, which he relates to the importance of incoming tourism from specific language groups. Here, a similar logic as in the case of international trade applies, whereby tourism flows increase the applicability of language skills in the service sector, as well as employers' willingness to pay for proficiency in a specific language.

Destination Language Skills of Migrants

The best-researched field regarding the labor market effects of languages is the importance of destination language skills of international migrants. Language skills are the most important and visible part of the human capital an immigrant acquires during his or her migration history. The consideration to invest into proficiency in the destination country language heavily influences the economic success in the host country. However, even prior to the migration decision, the expected costs of later language acquisition, e.g., in the form of perceived linguistic borders, shape the initial decision to migrate itself, thereby influencing aggregated migration flows along linguistic borders (Adsera and Pytlikova, 2012).

As the decision to invest in the destination language is a conscious one (contrary to the acquisition of a first language, but similar to the acquisition of foreign languages discussed in the former section), it is appropriately analyzed in a human capital framework, weighing off expected benefits and costs of the language acquisition. Chiswick and Miller (1995) distinguish the determinants of language fluency of immigrants into three categories related to the exposure to the destination language, the efficiency in learning a new language, and the economic incentives derived from being able to communicate in the destination language. Exposure and learning efficiency, dependent on individual, institutional, and situational characteristics, shape the costs in terms of effort or monetary costs for language classes, while economic incentives display the expected benefits formed by increased opportunities for economic integration and social participation.

The exposure to the destination language can be divided into prior to and postmigration exposure. Prior to migration, emigrants might learn about the destination language by being educated in secondary school and reading nontranslated literature and watching dubbed movies. Postmigration, exposure to the destination language will mainly be a nondirected process dependent on the time of residence in the country. The exposure is also dependent on an individual's socioeconomic situation, with the labor market status, working full-time, being unemployed or out of the labor force, influencing the degree of exposure. The decrease in exposure when individuals are not or unemployed is especially pronounced when living in a socioeconomic enclave. Such a linguistic enclave further reduces incentives to acquire proficiency in the destination country by restricting the range of opportunities where this language can be applied.

Immigrants differ in their ability or efficiency in acquiring the destination language. This ability is shaped by cognitive ability, education that might be influenced by selection processes in migration. A major factor influencing the ability to learn new languages is the age at arrival, given that this determines the ability to acquire the new language according to the linguistic-based critical period hypothesis discussed earlier. Immigrants arriving in young ages before this critical period in early adolescence pick up the destination language almost effortlessly and reach a near-native level of proficiency. For immigrants arriving after the critical period, the ability to learn new languages is distinctively decreasing.

A second linguistically based factor influencing the ability to acquire proficiency in the destination language is the linguistic distance between the mother tongue and the language to acquire. The linguistic origin turns proves a major determinant of destination language proficiency, outweighing further socioeconomic correlates and determinants. The effect of linguistic distance interacts with the effect of age at arrival: while young immigrants are unlikely to be affected by their linguistic origin, the additional costs in the language acquisition associated to a distant linguistic origin are substantial for later arrivers who start acquiring the new language after the critical period (Isphording and Otten, 2013).

The third group of influencing factors (besides exposure and efficiency) constitutes the economic incentives of language proficiency in the destination language. The willingness to invest into the language proficiency mirrors the expected benefits. While parts of these benefits and utility gains will be due to an increased potential in social and cultural participation, economists tend to focus especially on labor market advantages.

The estimation of labor market effects of language skills is confounded by a number of empirical pitfalls, measurement error, unobserved variable bias, and simultaneity. First, with objective assessments of language skills being a rare exception,

studies addressing the destination language proficiency of migrants are typically restricted to self-reported information on language skills. This subjective information is likely to be prone to a severe measurement and misclassification error. Second, analogously to the estimation on returns to education, estimates for effects of language proficiency are likely to be biased by omitting unobservable ability from the regressions. Both wages and language skills are potentially positively affected by an unobservable (to the researcher) ability.

Recent studies have attempted to address these empirical issues by instrumental variable regressions. This technique relies on the assumption that there exists an observable variable that affects the language acquisition but is unrelated to the sources of the empirical pitfalls, namely the measurement error, the unobserved ability, and the labor market outcomes itself. With this instrument variable, exogenous levels of language skills can be predicted, which are then used to estimate the actual effect of language proficiency.

Due to the omnipresence of language in everyday life, such an instrument variable is hard to find. Initially, Chiswick and Miller (1995) proposed using parental education, family composition, and family ties abroad as valid instruments. Their estimates for the US, Canada, and Israel indicate severely downward biased ordinary least squares (OLS) estimates, but distinctive wage premia estimated by IV between 30 and 50% for speaking English 'Well' or 'Very Well.' Dustmann and Van Soest (2002) attempt to address the measurement error by using time-adjusted averages and lags and leads of self-reported measures of language proficiency from panel data of the German Socio-Economic Panel. Their sample is restricted to German 'guest workers,' immigrant blue-collar workers from Mediterranean countries, enlisted to fulfill the excess labor demand after World War II. They conclude that estimates from uncorrected OLS regressions are severely downward biased. Their results indicate a wage premium for speaking German 'Well' or 'Very Well' by about 11–14%, which exceeds the OLS estimates by a factor of 2–3.

Using data from the 1990 US Census, Bleakley and Chin (2004) focus on childhood immigrants in the United States who arrived at an age younger than 17. Their identification relies on the so-called critical period of language acquisition, based on the empirical evidence of a decreasing ability to learn languages after reaching adolescence. Linguists and neurobiologists state that children are able to acquire a language effortlessly up to a native level until this critical period, while this ability decreases distinctively thereafter. Bleakley and Chin use this structural break to compare young and old arriving immigrants, while additionally controlling for whether the immigrants originates from an English- or non-English-speaking country. By using this double comparison groups, they again state that uncorrected OLS estimates are severely downward biased. In their estimates, an increase in language ability on a 0–3 scale (0 = not at all, 3 = Very well) is associated with a wage premium of around 30%. Using more recent data from the American Community Survey 2010, Isphording and Sinning (2012) extend their analysis to the overall population of immigrants in the United States, substituting the age at arrival with the exposure to English by time of residence. Their results confirm the initial results for child migrants and for both child and adult migrants.

While these causal IV estimations provide a robust estimate of the importance of language skills for the economic integration of immigrants into a host country, they provide little insight into the channels through which the language skills affect labor market outcomes. Indeed, only a few studies have gone into detail in this regard. Bleakley and Chin (2004) identify schooling as a major channel through which language skills affect the earnings of child migrants. Child migrants from non-English-speaking countries arriving in early years possess on average 2 years of additional schooling compared to their late-arriving counterparts. The additional schooling is able to explain almost the complete causal effect of language on wages. Isphording and Sinning (2012) confirm the importance of education as a mediating channel for child migrants, although their results neglect this importance for immigrants arriving in adulthood (who acquired their education in the sending country).

Chiswick and Miller (2010) address the occupational choice of immigrants regarding the language requirements of occupations as a mediating channel of language skills on wages. They combine data from the 2000 US Census with scores for English requirements measures taken from the O*NET database. Their results indicate a strong effect of language probability on the job choice, whereby good job matches in terms of language skills and requirements are rewarded by wage premia. Berman et al. (2003) use Israeli survey data to assess a potential complementarity between immigrant language proficiency and further human capital. Their results indicate a strong heterogeneity in the effect of language proficiency on wage growth across occupations. As the language acquisition is only relevant for high-skilled professions, the authors interpret this as evidence for the complementarity between language proficiency and further human capital. Chiswick and Miller (2003) use 1991 Canadian Census data to show a similar complementarity between premigration schooling and experience and the destination language skills. Language skills act as the translating medium to apply premigration acquired human capital into the host country labor market.

Aside from these studies focusing on the earnings effects of wages, further studies have shown additional effects of language proficiency on labor-market related outcomes. Aldashev et al. (2009) document the impact of German language use on labor supply decisions and unemployment probability using German household data from the Socio-Economic Panel. They conclude that language proficiency significantly affects the probability of labor market participation and subsequent employment. Dustmann and Fabbri (2003) analyze the language skills of ethnic minorities using UK survey data. Besides significant earnings effects, they also find a higher employment probability of around 20% for individuals who are proficient in English.

Finally, besides heterogeneity in the level of proficiency, a second mechanism can arise through which language differences can affect labor market outcomes. Differences in language do not necessarily mean vertical differences in proficiency, but can also appear in the form of a horizontal differentiation through dialects or accents, speech patterns, or choice of words. These differences can act as a signal providing information on social status and ethnicity and allow for taste-based or statistical

discrimination. Grogger (2011) provides evidence of an influence of racial speech patterns in the United States. Using unique data from the 1997 National Longitudinal Survey of Youth (NLSY) augmented with anonymously rated speech patterns, he identifies a significant wage penalty of around 12% for speaking African-American English instead of standard American English.

Conclusion

This article aimed at providing an overview of the economic literature addressing the relationship between language and productivity. Three different strands of literature have been summarized: (1) studies addressing the literacy of the native population and its effect on individual labor market outcomes and aggregated growth and development; (2) the labor market effects of foreign language proficiency against the background of a growing demand by globalization; and (3) the importance of destination language proficiency for the social and economic integration of immigrants.

All three strands of the literature provide comprehensive evidence for the crucial role played by language-related human capital in economic processes. As the key medium of communication, language crucially shapes economic outcomes on both the individual and aggregated level.

The general importance of language-related human capital becomes apparent with empirical evidence on the labor market effects of adult literacy. For both developed and developing countries, studies demonstrate a positive effect of literacy on wages and employment probabilities that go beyond the effect of associated formal education. This microevidence is accompanied by macroevidence linking the economic growth to literacy rates as a crucial driver of economic development.

The specific importance of language in the labor market has been fueled by ongoing globalization. International trade in goods and services, tourism, and migration flows increase the importance of foreign-language skills. Foreign-language skills are crucial in reducing search and information costs, as well as establishing long-run business relations. This is especially true for English as the commonly accepted lingua franca in international trade. Proficiency in foreign languages is rewarded by employers though significant wage premia, especially in occupations related to services, trade, and tourism.

Finally, the best-researched field against the background of language skills and the labor market is the importance of destination language skills for international migrants. The proficiency in the destination country is both a prerequisite for and an outcome of a successful integration into the host country. Immigrants weigh off costs and expected benefits to decide on their investment in terms of the effort and opportunity costs of language acquisition. Determinants of language skills, such as linguistic origin, age at migration, individual education and ability, and enclave density, shape a heterogeneous cost structure, while expected benefits are given by labor market advantages, in terms of higher wages and employment probabilities. Wage effects of language skills are heterogeneous across the occupational distribution, with requirements for language skills strongly differing between blue- and white-collar occupations. Language skills are shown to be complementary to further forms of premigration acquired human capital, acting as the medium of translation to apply home-country knowledge in the host-county labor market.

See also: Foreign Language Teaching and Learning; Language Development, Theories of; Language Development: Influence of Socio-Economic Status; Language and Thought: The Neo-Whorfian Hypothesis; Migration, Economics of.

Bibliography

Adsera, Alicia, Pytlikova, Mariola, 2012. The Role of Language in Shaping International Migration. CReAM Discussion Paper Series 1206. Centre for Research and Analysis of Migration (CReAM), Department of Economics, University College London.

Aldashev, Alisher, Gernandt, Johannes, Thomsen, Stephan L., June 2009. Language usage, participation, employment and earnings: evidence for foreigners in West Germany with multiple sources of selection. Labour Economics, Elsevier 16 (3), 330–341.

de Baldini Rocha, Maúna Soares, Ponczek, Vladimir, August 2011. The effects of adult literacy on earnings and employment. Economics of Education Review, Elsevier 30 (4), 755–764.

Berman, Eli, Lang, Kevin, Siniver, Erez, June 2003. Language-skill complementarity: returns to immigrant language acquisition. Labour Economics, Elsevier 10 (3), 265–290.

Bleakley, Hoyt, Chin, Aimee, May 2004. Language skills and earnings: evidence from childhood immigrants. The Review of Economics and Statistics, MIT Press 86 (2), 481–496.

Charette, Michael F., Meng, Ronald, August 1998. Literacy and numeracy. Canadian Journal of Economics, Canadian Economics Association 31 (3), 495–517.

Chiswick, Barry R., Miller, Paul W., April 1995. The endogeneity between language and earnings: international analyses. Journal of Labor Economics, University of Chicago Press 13 (2), 246–288.

Chiswick, Barry R., Miller, Paul W., October 2003. The complementarity of language and other human capital: immigrant earnings in Canada. Economics of Education Review, Elsevier 22 (5), 469–480.

Chiswick, Barry, Miller, Paul, January 2010. Occupational language requirements and the value of English in the US labor market. Journal of Population Economics, Springer 23 (1), 353–372.

Church, Jeffrey, King, Ian, May 1993. Bilingualism and network externalities. Canadian Journal of Economics, Canadian Economics Association 26 (2), 337–345.

Coulombe, Serge, Tremblay, Jean-François, August 2006. Literacy and growth. The B.E. Journal of Macroeconomics, De Gruyter 6 (2), 1–34.

Crystal, D., 2010. The Cambridge Encyclopedia of Language, third ed. Cambridge University Press.

Dustmann, Christian, Fabbri, Francesca, July 2003. Language proficiency and labour market performance of immigrants in the UK. Economic Journal, Royal Economic Society 113 (489), 695–717.

Dustmann, Christian, Soest, Arthur Van, April 2002. Language and the earnings of immigrants. Industrial and Labor Relations Review, ILR Review, Cornell University, ILR School 55 (3), 473–492.

Ginsburgh, Victor A., Prieto-Rodriguez, Juan, April 2011. Returns to foreign languages of native workers in the EU. Industrial and Labor Relations Review, ILR Review, Cornell University, ILR School 64 (3), 599–618.

Grogger, Jeffrey, 2011. Speech patterns and racial wage inequality. Journal of Human Resources, University of Wisconsin Press 46 (1), 1–25.

Hanushek, Eric A., Woessmann, Ludger, September 2008. The role of cognitive skills in economic development. Journal of Economic Literature, American Economic Association 46 (3), 607–668.

Isphording, Ingo E., December 2013. Returns to foreign language skills of immigrants in Spain. LABOUR, Centre for Economics and International Studies 27 (4), 443–461.

Isphording, Ingo E., Otten, Sebastian, May 2013. The costs of Babylon – linguistic distance in applied economics. Review of International Economics, Wiley Blackwell 21 (2), 354–369.

Isphording, Ingo E., Sinning, Mathias, 2012. The Returns to Language Skills in the US Labor Market. IZA Discussion Papers 7080. Institute for the Study of Labor (IZA).

Lohmann, Johannes, February 2011. Do language barriers affect trade? Economics Letters, Elsevier 110 (2), 159–162.

McIntosh, Steven, Vignoles, Anna, July 2001. Measuring and assessing the impact of basic skills on labour market outcomes. Oxford Economic Papers, Oxford University Press 53 (3), 453–481.

Selten, Reinhard, Pool, Jonathan, 1991. The distribution of foreign language skills as a game equilibrium game equilibrium models. In: Social and Political Interaction, vol. 4. Springer.

UNESCO, 2004. The plurality of literacy and its implications for policies and programs. UNESCO Education Sector Position Paper 13.

Williams, Donald R., July 2011. Multiple language usage and earnings in Western Europe. International Journal of Manpower, Emerald Group Publishing 32 (4), 372–393.

Language and Literature

Nigel Fabb, School of Humanities, University of Strathclyde, Glasgow, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The linguistic processing of poetry involves focus on the verbatim form of the line as a whole unit in working memory, with epistemic and affective consequences. The form of text can also be part of its meaning. Language is used in narrative to structure the represented world and engage us in specific ways with its fictional people. Linguistic form in literature can be different in interesting ways from ordinary language: attenuated, deviant, and with added regularities.

The Significance of Verbatim Language in Literature

When a person processes language, sounds, words, and syntactic structures are identified; we might distinguish between the verbatim form of language (the actual sounds and words, in the order in which they are pronounced), the underlying form (the complex linguistic representation which underlies the verbatim form and gives it its linguistic structure), and the gist – the meaning which can be extracted from the verbatim/underlying form. Much of the processing of the sounds or letters which form the linguistic input is devoted to extracting gist; that is, the verbatim (and underlying) form serves essentially to get to the stage of extracting a coded meaning. However, in the processing of a literary text, the verbatim form of the text takes on a particular importance, needing to be attended to and remembered. That is, the actual sounds, words, and surface textual organization are (in principle) important in the processing of the language of literature. Roman Jakobson made this point in a seminal article on literary linguistics (Jakobson, 1960); he identified various functions of language, including the poetic function (which is the dominant function in literature), and suggested that the poetic function was best performed by texts which manifested a 'focus on the message,' by which he meant a focus on the verbatim linguistic form of the text.

The study of the language of literature is thus dominated by the study of the specific verbatim form of a text, rather than by the meanings which are coded in it: the reader or hearer's attention to the verbatim form is often seen as providing different kinds of meaning or other effects than are produced in ordinary language processing. However, other aspects of linguistic form, which are not verbatim, are also potentially relevant: in some traditions, aspects of underlying (rather than surface verbatim) phonological form appear to be relevant to how meter, rhyme, and alliteration are processed; *see chapter by Hanson for further discussion*.

Linguistic Form in Literature

In some kinds of literature, particularly in verse, patterns appear in the language of the text. These patterns can be regular and predictable, in which case we say that they are imposed by rule, or they can emerge less predictably in the text. Consider, for example, the type of pattern called 'alliteration,' in which nearby words begin with the same consonant. In Old English poetry, the line is divided into two halves, with a rule that at least one word in the first half must alliterate with at most one word in the second half:

Oft Scyld Scēfing sceathena thrēatum
Monegum mægthum meodo-setla oftēah

> often Scyld the son of Scef, the enemy troops / of many nations the mead-benches took away.
>
> *Beowulf* (lines 4–5).

This is a regular pattern, a form which is imposed on the text by rule. Alliteration also appears in modern English verse, but never predictably: a line may have several alliterating words in it but there is no requirement for alliteration nor is there any regulation of alliteration of the kind we have just seen in Old English. Consider, for example, these lines by the seventeenth-century poet John Dryden in which the second line alliterates internally on 'd' but the surrounding lines do not alliterate at all.

> And therefore 'tis no wonder if we see
> They dot on dowdies and deformity;
> Ev'n what they cannot praise they will not blame
> But veil with some extenuating name:
>
> John Dryden *Concerning the Nature of Love* 1685 (lines 141–144).

Dryden's text demonstrates alliteration as an irregular pattern, which emerges in the text rather than being imposed by rule on the text.

Both of the quoted texts demonstrate another kind of form, meter, which is differently regulated in each case; meter governs the overall length of the line, and controls its rhythm. In Beowulf, each line is divided into halves, with a normative pattern where each half line has two stressed syllables, and the two stressed syllables in the first half alliterate and with the first stressed syllable in the second half; there are many variations on this basic pattern, but every line shows some variation on it. This meter is 'strong stress meter.' In Dryden's text every line has 10 syllables and even-numbered syllables tend to be stressed: this meter is 'iambic pentameter.' The Dryden text further demonstrates (regulated) rhyme: in every pair of lines, the final words rhyme. The Dryden text also

 http://dx.doi.org/10.1016/B978-0-08-097086-8.53056-6

demonstrates emergent (and irregular) parallelism, where adjacent parts of text have the same syntactic structure involving related but different words: 'they cannot praise – they will not blame.' In some literature, parallelism is strictly regulated; in English it is not. (For more detail on linguistic accounts of the regulation of linguistic form in verse *see chapter by Hanson*.)

Why do literary texts have added form? Jakobson (1960) argued that these added forms – both regular and irregular – function to draw attention to the verbatim language of the text, and hence manifest the poetic function. Richard Bauman (1975) suggests that the production of verbal art should be seen as the performance of texts which can be evaluated, and that the added forms offer criteria for evaluation; for example, the author is evaluated on how he or she meets the formal requirements for a text of a particular type.

A third possibility relates to the stylistic notion of 'markedness,' which depends on a norm which is transgressed, with consequences for interpretation. Where regular forms are stipulated, markedness can be produced by deviations from strict regularity: in many traditions, the regular forms set a norm but allow variations. Thus English iambic pentameter normatively requires 10 syllables to the line and all and only even-numbered syllables are stressed; many lines in this meter legitimately deviate somewhat from this, thus allowing markedness to emerge. (For example, in Dryden's verse, there are often fewer than five stresses in the line, even though there are always five even-numbered syllables; the stress patterns change from line to line even though the lines are always in the same meter.) Markedness can thus be produced by variation against a regular formal norm, but it can also be produced by a form emerging against a norm of its being absent; thus the alliterating pattern in Dryden's text is marked in a text which otherwise does not manifest alliteration in any regular way. The forms of verse thus allow markedness to emerge.

Markedness is a standard notion in formalist literary criticism, and takes on a specific interest when we consider it in the context of relevance theory, as a theory of communication (Sperber and Wilson, 1986). Communication is governed by the principle of relevance, which guarantees that if greater processing effort is demanded (e.g., by a text) then there will be greater cognitive effects. The key cognitive effect is that an implicature is produced. When a meaning is implied by a text it is an implication. A subset of implications are implicatures: an implicature is a communicated implication, and the derivation of implicatures by an audience is a derivation of implications which they take the author to be communicating. Thus the generation of a wider range of implicatures increases the range of what the audience can assume is being communicated by the text. Markedness produces implicatures. The unpredictability of markedness characteristically demands greater processing effort, and thus should always demand of the audience a greater degree of inferencing leading to a wider range of implicatures than would be demanded by an equivalent unmarked text. Because literary texts have added form, and because added form always permits the emergence of markedness, literary texts thus encourage their audiences to draw a wider range of implicatures.

Attenuated Linguistic Form in Literature

Literary texts can thus have 'more form' than ordinary language (lineation, meter, rhyme, etc.) but on the other hand, literary texts can also have 'less form' than ordinary language, generally by virtue of the literary text having an attenuated syntax. Thus, for example, Lorimer (1954: 548) describes (Persian) Bakhtiari verse, saying: "the language is frequently reduced to a sort of telegraphese, in which words or phrases are frequently juxtaposed without any nexus, or often any predicate." Something similar can be seen in this stanza by Blake:

> And what shoulder, & what art,
> Could twist the sinews of thy heart?
> And when thy heart began to beat,
> What dread hand? & what dread feet?
> William Blake 'The Tyger' (Blake's punctuation).

Attenuation of syntax can emerge in literary texts through concatenation, listing, the omission of conjunctions, a dispreference for complex sentence structures (e.g., fewer embedded clauses), omission of other syntactic units ('ellipsis'), and a greater freedom of reordering of elements within the clause.

Where syntax is attenuated, the text as code is attenuated, and so the text codes less meaning: that is, logical forms or propositions are less easily decoded from the syntactic organization of the words. Bloch (1974) proposed that attenuated syntax was a characteristic of ritual language (which is often one of the subtypes of verbal art), and he suggested that attenuated form in ritual language functions as a means of control by authority: "ritual is a kind of tunnel into which one plunges, and where, since there is no possibility of turning either to right or left, the only thing to do is to follow." However, this overemphasizes the degree to which the syntax of a text determines what it communicates: as we know from pragmatic theory, what the text codes is only partial evidence for communicated meaning, but meanings can be inferred from other kinds of evidence associated with the text in a cultural context. Attenuated form can liberate rather than limit interpretation, by demanding greater inferential activity from the audience (for a general discussion of theories of ritual speech, see Keane, 1997).

The syntax of verse can be attenuated relative to ordinary syntax, but it can also violate principles of ordinary syntax, either by its extreme attenuation, or by reorderings of words or phrases in ways which are not normally permitted. Examinations of these radical reorderings include Austin (1984) on the syntax of English Romantic poetry, or O'Neil (2001) on the very radical reorderings found in medieval Icelandic poetry, and these accounts usually conclude that where syntactic rules or conditions are broken, they are not broken randomly, but are broken in ways which specifically refer to the underlying linguistic forms of the language (for more discussion *see Hanson chapter*). More generally, there is a view that where the syntactic (or phonological) form of literature varies from ordinary language, it varies in a way which is dependent not only on general linguistic principles, but on the specific principles of that language. (For critical discussion of the possibility that literary language is a development of ordinary language, see Fabb, 2010.)

The Processing of Literary Linguistic Form

There are two broadly distinct processing tasks in producing or receiving language, which we might distinguish as coding and inferencing. As far as the hearer or reader is concerned, the coding tasks involve the identification of the sound-structure of the utterance, identification of words, of syntactic relations between words, and thereby producing a sentence structure which codes a 'logical form,' from which propositions can be derived. The inferencing tasks take the propositions coded by the sentence as partial evidence of the communicator's intention, add other propositions drawn from memory, and develop implicatures (implications which are intended by the communicator) from this. In the coding stage of processing, the verbatim text must be identified (i.e., the actual sounds, words, order), but only in order to determine the logical form, or gist, of the linguistic input; the gist may be retained in memory but there is little reason to retain a memory of the verbatim form, or to focus on the verbatim form as such. However, (following Jakobson) we can say that in literary texts, we are expected to focus on the verbatim form, for reasons discussed earlier.

It is likely that in addition to extracting the coded meanings from a literary text, some cognitive process assigns specific literary forms to the verbatim text, whether regular or irregular. For example, when reading or listening to Dryden's text, the alliteration in the line may be specifically identified by the processor as a pattern, as are the meter and rhyme pattern which hold throughout. The literary formal patterns which the processor may identify are based on the verbatim text. This raises interesting questions for working memory, which I now consider, using the model of Baddeley and Hitch (Baddeley, 2012). In this model, the seen or heard linguistic input is integrated with knowledge of language (phonology, syntax, etc.) into structured verbatim linguistic material, in working memory (in the Baddeley–Hitch model, this is done in a subpart of working memory called the episodic buffer, which integrates material). Only a small amount of material can be held – about 15 syntactically connected words; this can be understood by hypothesizing that working memory can hold about four chunks or episodes: when holding language, those chunks might be individual unconnected words, but if the words are connected then several words can be combined into a single chunk. When we listen to language, this limited amount of material is constantly replenished as new material comes in and old material is lost. In verse, many of the added forms (particularly the regular added forms) hold relative to the line or in some cases a pair of lines: meter regulates the whole line, rhymes tend to be placed in specific positions relative to the line, alliteration stretches across a line, parallelism holds between half-lines or between whole lines, and so on. This suggests that if these added forms are to be processed, the line must be treated as a whole unit, verbatim, in working memory. We might begin with the position that words are fed in a continuous stream into and out of working memory capacity in ordinary language. In contrast, the hypothesis is that verse is processed quantally, line by line, with the first word in a line held in working memory until the whole line is processed. Once processed, the line as a whole can be dispensed with (perhaps retaining the final word in memory as a rhyme). The view that the line is treated as a whole unit in working memory is usually based on the fact that the line is sufficiently short that a whole line could be contained verbatim within working memory capacity (Hogan, 1997; Fabb, 2014). In fact, where lines are short, pairs of lines might be treated as whole units, and this would make sense in those traditions where parallelism between lines is strictly regulated and must be processed as such.

The evidence that the formal patterns in the verbatim text must be specifically processed alongside the decoding for gist comes from experimental work on processing ease and memory. We know that regular forms make the processing of the language easier (Reber et al., 2004), at least in part because the forms restrict the selection space for words. In processing language, the hearer must discover what word is being spoken, which involves matching the sound input to the known word. The retrieval of the known word is made easier when the candidates are more restricted: that is, when the selection space is restricted. For example, in Blake's text the hearer knows that the fourth line must end in a word rhyming with 'beat.' This means that the rhyme means that the word 'feet' can more easily be drawn from the lexicon to match the sounds which are heard (and thus to recognize that those sounds code that word). Another kind of evidence that formal patterns must be specifically processed comes from the fact that when an alliterating pattern is repeated later in the text, words adjacent to the earlier pattern prime the selection of words adjacent to the later pattern (Lea et al., 2008). These experiments show that verbatim form must be specifically processed, alongside processing for gist, because it has an influence on the decoding process.

Why should it be interesting that lines are treated as a whole verbatim unit in working memory? This is a claim which has been made before, in various guises, and usually it is a claim which suggests that conscious attention is focused on the line, that the contents of the line, separated off from the text, are given a privileged status, presumably with affective and epistemic consequences. These claims remain untested, but fit well with traditional views of literature, thus matching newer cognitive and linguistic approaches with more traditional conceptions of poetry (see, for example, Boyd, 2012: 41).

Issues of form and working memory usually arise for verse, but it is worth asking what interest they have for literary prose. We know that literary prose can also have added form, including, for example, alliteration, rhyme, parallelism, and rhythmicity, though these forms are usually irregular (in part because regular form depends on lineation to provide frames against which the regularities can be formulated). It is likely that literary prose also requires attention to the verbatim language of the text, but in the absence of lineation, this raises unanswered questions about how prose is handled by working memory capacity.

The Assignment of Form by Inferencing

There is good reason to think that forms are assigned to the linguistic material being processed during the early stages of processing (alongside the decoding process), presumably by technical means similar to the ways in which linguistic form is assigned to the material. Indeed, the possibility *(Kristin*

Hanson chapter, Language and Poetic Structure) that some of this processing of form depends on underlying linguistic form fits with the possibility that the processing of linguistic form is closely connected with general phonological, lexical, and syntactic processing. However, these stages of processing are not characteristically open to introspection, and so we must also consider the fact that literary form – unlike linguistic form generally – is often quite salient, being explicitly recognized as such by the audience. This is best handled by suggesting that form can be assigned in a second way to the text, by inferencing (Fabb, 2002). For example, in Dryden's text cited above, the identification of alliteration is a specific part of the covert processing of the text (irrespective of whether it is explicitly noticed), and at the same time may be explicitly recognized as alliteration by the literarily competent audience. That is, the hearer or reader may draw the implicature that 'this line alliterates,' based, as with all implicatures, on the evidence provided by the text in a context. Similarly, the audience can produce explicit thoughts about the text that 'it rhymes,' 'it is in iambic pentameter,' 'it is in heroic couplets,' and so on. Form thus holds of the text as the content of an implicature (about the text) for which the text provides evidence. Two interesting consequences follow from focusing on form as implicated. First, it means that like the content of any implicature, form can be vague, ambiguous, or contradictory; in principle, form can even be 'metaphorical' – a form may be implied for a text which is difficult to literally assign to it. Second, it means that by treating form as a content, we can better understand the close interaction of form and content in a literary text, how forms can be symbolic, and so on.

Narrative and Language: Representation and Engagement

In this section, I comment on three aspects of narrative where there are issues of specific linguistic interest.

As Labov and Waletzky (1967) noted, narrative is defined by sequence, where verbal sequence mimics event sequence: the most minimal narrative represents an event leading to another event, and in verbal narratives the clause representing the first event precedes the clause representing the second event. An important way in which language functions in narrative relates to the narrative as a sequence, and how the sequence is divided into subsequences. The subsequencing operates at different levels. At the highest level, a narrative characteristically represents a complication (e.g., something goes wrong) which is resolved (e.g., it is put right), and the moment of shift between these often involves linguistic markedness, for example, a repetition, or some unexpected form or pattern (this is the moment of 'evaluation' of the narrative for Labov and Waletzky, 1967). This markedness appears to function to draw the audience's attention to this most important point in the story. At an intermediate level, longer narratives tend to be composed of distinct episodes (each usually with its own demarcated time, place, and participants), and episode boundaries are characteristically linguistically marked, for example, with a distinct connective such as 'so' or 'then,' by a shift of tense, or some other linguistic change whose semantic content may be neutralized or de-emphasized in favor of the structural function of marking a boundary (Hymes, 1981). At the sentence level, narratives make a distinction between clauses which express the sequence of events and clauses which indicate context or circumstance for events; the former are called 'storyline clauses' and are characteristically linguistically distinctive, for example, having a higher level of transitivity (in an extended sense of transitivity based around an actor making a change to another actor or thing, the basis of many narratives; Hopper and Thompson, 1980).

Another aspect in which language functions interestingly in narrative is in how the language enables the reader to keep track of which represented people are involved in events, and the sequencing and location of those events. This involves linguistic analysis of the use of pronouns, tense, and time markers, and other deictic items, to investigate how readers construct a mental representation of a fictional world (Emmott, 1997). As with all aspects of literary language, literary narrative can break or bend the expected rules, to produce particular effects. Thus, consider the way in which Charlotte Brontë begins *Jane Eyre*, by referring to location, time, and participants in a way which suggests (impossibly) that they are already part of the reader's mental representation of the narrated world.

> There was no possibility of taking a walk that day. We had been wandering, indeed, in the leafless shrubbery an hour in the morning; but since dinner (Mrs. Reed, when there was no company, dined early) the cold winter wind had brought with it clouds so sombre, and a rain so penetrating, that further out-door exercise was now out of the question.
>
> I was glad of it: I never liked long walks, especially on chilly afternoons: dreadful to me was the coming home in the raw twilight, with nipped fingers and toes, and a heart saddened by the chidings of Bessie, the nurse, and humbled by the consciousness of my physical inferiority to Eliza, John, and Georgiana Reed.
>
> Charlotte Brontë *Jane Eyre*, chapter 1 (first two paragraphs).

The same event can be represented by different types of syntactic construction: by an active or passive sentence, by a noun phrase, and so on. The choice between these can either emphasize or de-emphasize the role of actors in an event. In the second paragraph of the text above note the use of 'the coming home' (not: I came home) and 'a heart saddened' (not: my heart saddened). In both cases, the narrator, Jane, goes unmentioned as an agent or a possessor/experiencer, in a verbal echo of her more general removal and marginalization in the represented world of the novel. The analysis of how events are represented by different 'choices' within the linguistic system – both choices at a syntactic level and more generally in ways of speaking or stylistic choice – often forms part of a critical theoretical account of society, power, and inequality, often in media language rather than strictly literary language, and in this case is part of critical discourse analysis (Fairclough, 2010).

A third aspect of language which is of particular interest in narrative relates to the factuality or reliability of the represented events. In many types of narrative, events are represented by mediating them through a fictional character, who may or may not be explicitly represented as the narrator. Narrators may use the language of the text to indicate that events are

being mediated through an observer. Thus in the following text, the events are partly narrated by an (presumably reliable) external narrator, and partly narrated from the point of view of (the less reliable) Elizabeth herself.

> And with these words he hastily left the room, and Elizabeth heard him the next moment open the front door and quit the house.
> The tumult of her mind, was now painfully great. She knew not how to support herself, and from actual weakness sat down and cried for half-an-hour. Her astonishment, as she reflected on what had passed, was increased by every review of it. That she should receive an offer of marriage from Mr. Darcy!
>
> Jane Austen Pride and Prejudice ch.34.

The use of the combination 'was now' is particularly interesting (and common): it represents the events both as in the past (was) from the external narrator's point of view, and in the present (now) from the character's point of view. The reader is perhaps caught in both these perspectives at the same time. Linguistic devices of this kind for mediating the representation of events, and putting the reader or listener into a complex conspiracy with the characters, are common in literature (Banfield, 1982).

Effects of Language in Literature

From the perspective of literary studies, the primary interest in the language of literature is in the effects that specific kinds of language choices have on the reader or listener. In particular, can we explain why literature is characteristically valued as a way of using language? One possibility has already been explored, which is that language is used to produce markedness: a norm is set and then violated or varied from, thus requiring particular processing effort in the audience, which in turn delivers a wider range of interpretive effects. Thus the forms of language produce a wider range of meanings.

A sequence of psychological experiments have shown that various manipulations of language which are characteristic of literary texts can produce effects of truth or familiarity or pleasure. Thus McGlone and Tofighbakhsh (2000) show that aphorisms which rhyme are considered to be truer than aphorisms which do not. Reber et al. (2004) show that all kinds of regular poetic forms make language processing easier, and that the easing of processing can produce as side effects the sense that what is coded by the text is more true, is familiar (even if it is in fact novel), and is liked more. Thus the regular forms of language in literature appear to have an inherent capability to produce these epistemic and aesthetic effects.

Related questions may arise about other aspects of language in literature. For example, Rhodes and Donaldson (2008) show that syntactic contexts which result in the pairing of two words can produce the effect of 'unitizing' those words, as though they are a connected pair (even when they are not strictly related in meaning). This in turn can produce the effect of familiarity, as though this pair is already familiar (even when it is not). Though this is another quite low-level epistemic effect, it may nevertheless constitute a characteristic effect of the frequent use of parallelism in literature. Thus in Dryden's text quoted earlier, a parallel structure creates a pairing of 'praise' and 'blame' (already quite familiar in its own right) but also creates a pairing of 'dowdies' and 'deformity' which in itself is quite novel. Nevertheless, psychological work on unitization suggests that this pairing may produce a small 'epistemic jag,' a small effect of familiarity in the reader, which perhaps contributes to the overall aesthetic effect of the text (the term adapted from 'arousal jag,' Berlyne, 1960).

Finally, it is worth asking whether aspects of the language of literature encourage or enable creativity in authors. The processing of added literary form, in both verse and prose, appears to operate alongside ordinary language processing. Syntactic structure matches neither the line nor the limits of working memory (syntactic processing depends more indirectly on working memory, not requiring, for example, each sentence to be processed as a whole). The imposition of other types of form on the language of literature raises questions about the composition of literary texts. Fabb (2009) suggests that literary texts (and particularly verse) may in some cases be composed by concatenating verbal material rather than generating it through the normal syntactic processes. More generally, it is worth asking whether literary composition interferes with ordinary language production in ways which give rise to kinds of inventiveness, formal surprise, and creativity characteristic of literary composition.

See also: Discourse and Indentity; Language and Poetic Structure; Lexicon; Semantics.

Bibliography

Austin, Timothy R., 1984. Language Crafted. A Linguistic Theory of Poetic Syntax. Indiana University Press, Bloomington.
Banfield, Ann, 1982. Unspeakable Sentences. Narration and Representation in the Language of Fiction. Routledge & Kegan Paul, Boston.
Bauman, Richard, 1975. Verbal art as performance. American Anthropologist 77, 290–311.
Baddeley, Alan, Wilson, Barbara A., 2002. Prose recall and amnesia: implications for the structure of working memory. Neuropsychologia 40, 1737–1743.
Baddeley, Alan, 2012. Working memory: theories, models, and controversies. Annual Review of Psychology 63, 1–29.
Berlyne, Dan E., 1960. Conflict, Arousal, and Curiosity. McGraw-Hill Education, New York.
Bloch, Maurice, 1974. Symbols, song, dance and features of articulation. Is religion an extreme form of traditional authority? European Journal of Sociology 15, 55–81.
Boyd, Brian, 2012. Why Lyrics Last. Evolution, Cognition and Shakespeare's Sonnets. Harvard UP, Cambridge, MA.
Emmott, Catherine, 1997. Narrative Comprehension. A Discourse Perspective. Oxford University Press, Oxford.
Fabb, Nigel, 2002. Language and Literary Structure: The Linguistic Analysis of Form in verse and Narrative. Cambridge University Press, Cambridge.
Fabb, Nigel, 2009. Why is verse poetry? PN Review 36, 52–57.
Fabb, Nigel, 2010. Is literary language a development of ordinary language? Lingua 120, 1219–1232.
Fabb, Nigel, 2014. The verse-line as a whole unit in working memory, ease of processing, and the aesthetic effects of form. Royal Institute of Philosophy Supplement 75, 1–22.
Fairclough, Norman, 2010. Critical Discourse Analysis: The Critical Study of Language, second ed. Longman, London.
Hogan, Patrick C., 1997. Literary universals. Poetics Today 18, 223–249.
Hopper, Paul. J., Thompson, Sandra A., 1980. Transitivity in grammar and discourse. Language 56, 251–299.
Hymes, Dell, 1981. 'In Vain I Tried to Tell you.' Essays in Native American Ethnopoetics. University of Pennsylvania Press, Philadelphia, PA.

Jakobson, Roman, 1960. Closing statement: linguistics and poetics. In: Sebeok, T. (Ed.), Style in Language. MIT Press, Cambridge, MA, pp. 350–377.

Keane, Webb, 1997. Religious language. Annual Review of Anthropology 26 (1997), 47–71.

Labov, William, Waletzky, Joshua, 1967. Narrative analysis: oral versions of personal experience. In: Helm, June (Ed.), Essays on the Verbal and Visual Arts. Proceedings of the 1966 Annual Spring Meeting of the American Ethnological Society. University of Washington Press, Seattle, pp. 12–44.

Lea, R., Rapp, D., Elfenbein, A., Mitchel, A., Romine, R., 2008. Sweet silent thought: alliteration and resonance in poetry comprehension. Psychological Science 19, 709–716.

Lorimer, D.L.R., 1954. The popular verse of the Bakhtiāri of S.W. Persia -I. Bulletin of the School of Oriental and African Studies 16, 542–555.

McGlone, Matthew S., Tofighbakhsh, Jessica, 2000. Birds of a feather flock conjointly (?): Rhyme as reason in aphorisms. Psychological Science 11 (5), 424–428.

O'Neil, Wayne, 2001. Grammar games: unscrambling Skaldic syntax. In: Kenstowicz, Michael (Ed.), Ken Hale. A Life in Language. MIT Press, Cambridge, pp. 339–354.

Reber, Rolf, Schwarz, Norbert, Winkielman, Piotr, 2004. Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? Personality and Social Psychology Review 8, 364–382.

Rhodes, Sinéad M., Donaldson, David I., 2008. Electrophysiological evidence for the effect of interactive imagery on episodic memory: encouraging familiarity for non-unitized stimuli during associative recognition. NeuroImage 39, 873–884.

Sperber, Dan, Wilson, Deirdre, 1986. Relevance: Communication and Cognition. Blackwell, Oxford.

Language and Poetic Structure

Kristin Hanson, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A fundamental relationship between language and poetic structure is that between grammar and verse forms such as meter, rhyme and alliteration, and syntactic parallelism. Like a speaker of a language, a poet composing in such a verse form has intuitive knowledge of a finite system of constraints which defines the form, within which infinitely many various and new instances of the form can be created. Research making such knowledge explicit has revealed just how directly such forms draw on grammatical structure, inspiring and supporting an explicit cross-linguistic theory of verse forms relating them to universal grammar.

Language is the material out of which poetry is made. It is a complex phenomenon, having a characteristic structure involving a lexicon and a grammar with both phonology and syntax, as well as characteristic uses such as communication. All of these aspects of language play interesting roles in poetic structure. This article will focus mainly on grammar, and its relationship to characteristic aspects of poetic structure such as meter, rhyme and alliteration, and syntactic parallelism. These define core forms of versification found across the world's languages (Jakobson, 1960; Kiparsky, 1973, 1987; Fabb, 1997) and create poetic effects even when they are not used with sufficient regularity to constitute versification. In recent decades they have been the subject of striking new directions of research as a result of new directions of research in grammar.

Grammar and Verse

Versification by definition imposes formal constraints on language beyond those language conforms to intrinsically. To describe those constraints for any given verse form is a surprisingly challenging problem, because individual lines within a form are all different from one another, often artfully so for specific effects. It is thus a problem, as Jakobson (1960) influentially put it, of abstracting the invariant 'verse design' from the variation embodied in all the 'verse instances.' Even for those verse forms, especially in written traditions, for which some constraints have been articulated, the problem remains a challenging one, in that poets' actual practice may not match the articulated constraints: lines which violate them may be included in compositions in the form, at the same time that lines which conform to them may never be used. Such mismatches suggest that articulated constraints may not constitute a complete and true description of what a poet instinctively knows about a form.

Construing the problem of describing a verse form as one of making explicit unconscious as well as conscious knowledge a poet has of what defines the form makes it, as Halle and Keyser (1972) point out, analogous to the problem of describing grammar itself. In Chomsky's (1965, 1972) influential formulation, grammar is knowledge a fluent speaker of a language has, partly consciously and partly unconsciously, of what would or would not be well-formed instances of the language. In phonology, for example, the subsystem of grammar dealing with sound structure (*see* Phonology), an English speaker knows not only that the combination of sounds in *brick* is an actual word and that in *blick* is not, but also that *blick* is a possible word whereas **bnick* is not; the '*' indicates the latter's ill-formedness. Similarly in syntax, an English speaker knows that a sentence like *Colorless green ideas sleep furiously* is a possible sentence of English while **Furiously ideas green colorless sleep the* is not. To describe the grammar of a language, Chomsky argues, is to make explicit the constraints which constitute such knowledge. The research program within linguistics which adopts that goal is 'generative linguistics' (*see* Generative Grammar), and its analog for verse is 'generative poetics.'

Describing a verse form involves grammar in substantive aspects as well. Although meter, rhyme and alliteration, and even syntactic parallelism are often described metaphorically as contributing 'music' to poetry, the entities which define them – particular configurations of vowels and consonants, syllables and stress, noun phrases and verb phrases – are specific to grammar. This means that as research into grammar has sought to discover the fundamental entities and organizational principles of grammars, new formal structures have become available to descriptions of verse forms, sometimes leading to striking new insights into familiar forms. By the same token, research into verse forms has suggested and supported hypothesized entities and principles of grammar.

Two ideas which have shaped grammatical research have been particularly significant for research into versification. One is the assumption, articulated by de Saussure (2006) and continued in generative grammar, that grammar is at some level purely formal: phonology is not reducible to sound, and syntax is not reducible to meaning. The entities of grammar defining verse forms may thus sometimes prove to be surprisingly abstract.

The other is the conceptualization of the grammar of any particular language as one instance of a common phenomenon of human language. Although this idea also has a long history, including notably for poetics the interest in it taken by Jakobson and his circle (Toman, 1995), it was given a particularly galvanizing (though not uncontroversial) formulation in Chomsky's (1965, 1972) suggestion that the most basic formal properties of grammars might be essentially similar across all languages because they are innately determined (see also Pinker, 1994). This hypothesis of 'universal grammar' motivates the relevance of wide-ranging cross-linguistic phenomena to the formalization of any single language's grammar. For

versification, this means that evidence for particular ways of analyzing the structures figuring in a verse form may come from languages unrelated to that of the verse in question, and it means that perception of some aspects of verse may be as much instinctive as learned (Kiparsky, 1987).

The observation that the goals and vocabulary of the description of verse are similar to those of the description of grammar invites the deeper question of whether there can be a theory of versification, as there are theories of grammar, and if so, what specific relationship of versification to grammar such a theory might posit. After all, there are differences as well as similarities: we can usually tell in an instant whether language is verse or not. Again it was Jakobson (1960) who influentially suggested that it is specifically the sequencing of entities which are in some way equivalent in language which constitutes verse. Interpreting this in light of the formal and universalist understanding of grammar characteristic of generative theories, Kiparsky (1973, 1987) ventured one of the first testable theories of versification, hypothesizing that verse forms are constituted by repetition of those phonological and syntactic structures which figure in universal grammar. What these structures are and how they are organized may be different in important ways from what Jakobson assumed, however; for example, the linear sequencing he emphasized is crucially understood within generative theory to interact with hierarchical structure.

One of the simplest observations in support of this hypothesis about the relationship of verse to grammar is that each of the major forms of versification corresponds to a major subsystem of grammar. Let us turn, then, to some specific examples of descriptive and theoretical claims following this line of inquiry, and then briefly consider some of its consequences and limitations.

Meter

The phonology of every language involves two distinct though interrelated subsystems, one dealing with rhythm, such as the quantitative grouping of sounds into stressed and unstressed syllables, and the other dealing with what is sometimes termed 'melody,' referring not only to tones but also to the qualitative features of segments which make, for example, a [p] different from a [b] or an [i] different from an [e]. Verse forms which impose constraints on rhythmic structure beyond those given by the phonology are meters.

The dominant meter of the modern English poetic tradition, the iambic pentameter (from Greek ἰαμβικός πεντάμετρος, a line of five iambs, metrical feet of two syllables, one 'short' then one 'long'), affords a well-developed example of how changes in the goals and substance of grammatical description have wrought changes in the conceptualization of a familiar verse form. The English iambic pentameter has a long tradition of being described as a line having a basic form of five iambic feet (in (a) separated by '|'), each consisting of an unstressed syllable ('x') followed by a stressed one ('/'):

a. x / | x / | x / | x / | x /

What 'unstressed' and 'stressed' mean in this context are matters of debate, but a line like (b i) below plausibly has a rhythmic pattern similar to (a) on most interpretations. Other lines, however, like (b ii) cited in Kiparsky (1975), vary from the pattern in (a) and are felt to be more complex. Variations as in (b ii) are, in this tradition, described as involving 'substitution' of different kinds of feet with different patterns of syllables and stress for the basic iambs of (a), such as trochees ($^{/x}$), spondees ($^{//}$), phyrrics (^{xx}), or anapests ($^{x\,x\,/}$), as shown:

b. x / |x / | x / | x / | x /

i. So long as men can breathe and eyes can see
(Shakespeare, Sonnets 18, 13)

/ x| / / | x x| / /|x /

ii. Pluck the keen teeth from the fierce tiger's jaws,
(Shakespeare, Sonnets 19, 3)

x /| x x /|x x |x /| x x /

iii. How many a holy and obsequious tear
(Shakespeare, Sonnets 31, 5)

What Halle and Keyser, building on Jespersen (1933), point out is that unless such a description makes explicit not only what substitutions are found, but also what ones are not found, it fails to capture what the poet knows that leads to use of some variations but avoidance of other imaginable ones, by hypothesis on grounds of their metricality and unmetricality, respectively. The construct in (c), for example, sounds like prose, not verse, and in spite of its 10 syllables has a rhythmic structure not found in any lines in Shakespeare's sonnets. Yet as shown, it can be analyzed into basic and substituted feet just as the lines in (b) can; nothing in the traditional description of iambic pentameter captures (c)'s fundamental difference of unmetricality:

c. x / |x x|x / | x x|/ x

*Imaginable lines are avoided (construct).

An adequate description of a meter, Halle and Keyser suggest, should model a poet's intuitions regarding metricality as well as complexity.

Much research has gone into the question of how to realize this goal. Most fundamentally, Halle and Keyser (1972) suggest that a meter must be formally distinguished from phonology. As language, a line of verse has a rhythmic structure defined by phonology (Chomsky and Halle, 1968), regardless of any meter. For the English iambic pentameter, they suggest, the meter itself can be formalized in two components: an abstract template composed of weak ('w') and strong ('s') metrical positions; and correspondence rules defining allowable mappings of the phonological structure of the line into the template. A scansion is, then, not a representation of a stress pattern, but a mapping of a stress pattern – or more precisely, of a phonological form which is realized as one – into the template in accordance with the rules. And the challenge for a description of a meter is to specify the right template and correspondence rules.

For example, Kiparsky (1977), building on insights of Liberman and Prince (1977) that linguistic rhythm itself involves w and s positions which form constituents ('⋀'),

proposes that at its lowest level the English iambic pentameter has a template of five feet as in (d):

d.

One correspondence rule maps one syllable into each position, as stated in (f i) below. As for stress, Kiparsky (1975, 1977) argues that Shakespeare avoids lines like the imaginable (e) which is rhythmically similar to (b ii) but has polysyllabic words where (b ii) has monosyllables:

e. *Pluck imménse teeth from enráged tigers' jaws (construct)

This contrast suggests that it is specifically stress relations within words that the meter regulates, as expressed informally in the correspondence rule in (f ii):

f. i. Each metrical position corresponds to a syllable.
ii. No weak metrical position corresponds to a syllable which is strong relative to another syllable in the same word.

Such a model of the iambic pentameter describes the lines in (b) as metrical. In (g) the stress regulated by (f ii) is marked, and as (g i, ii) show, the lines in (b) can be mapped into the template in (d) in accordance with (f). The model describes the lines in (c) and (e), in contrast, as unmetrical; as (g iii, iv) show, they cannot be mapped into (d):

g. i. So long as men can breathe and eyes can see
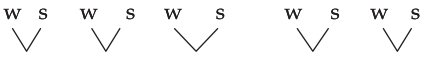

ii. Pluck the keen teeth from the fierce tíger's jaws

iii. *Pluck imménse teeth from enráged tígers' jaws
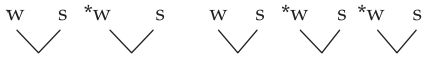

iv. *Imáginable lines are avóided

The model thus posits a verse design that allows a varied range of rhythmic structures in its verse instances, but not an unlimited one. Moreover, it does so in a way that establishes for each line an explicit relationship between its phonologically defined rhythmic structure and a metrical template, the relationship from which metrical effects arise.

The template in (d) and the constraints in (f) are only illustrative of basic elements of a description of the English iambic pentameter. There are systematic exceptions to (f) as well as additional strictures; there are phenomena of elision (reduction of two syllables to one used in (b iii)), extrametricality (extra syllables at edges), catalexis (unfilled metrical positions), resolution (multiple syllables in single metrical positions), caesurae (syntactic breaks within lines), and enjambment (absence of expected syntactic breaks at ends of lines) to be accounted for, not to mention refinements pertaining to stress itself. The details are different for every poet and genre (Kiparsky, 1975, 1977; Tarlinskaja, 1976); as Kiparsky (1977) observes, describing an individual poet's practice within a metrical tradition is analogous to describing dialects within a language. Shakespeare's claimed avoidance of lines like (e), for example, is categorical in his sonnets but not in his plays (Youmans, 1983; Golston, 1998; Hanson, 2006). And of course, English has meters other than the iambic pentameter, and languages across the world have other meters still.

Crucially, however, the goals, methods, and substance of generative grammar which shape the model in (d) and (f) have invited this metrical variety to be explored in a unified and grounded way. As noted, Kiparsky's (1977) formalization of (d) and (f ii) built on Liberman and Prince's (1977) theory of linguistic rhythm itself as composed of constituents pairing strong and weak elements, and further posited that those constituents are organized hierarchically to form a complex periodic structure of which stress patterns are one manifestation. That theory has inspired cross-linguistic studies of stress and revealed a rich range of structures and principles which figure in the grammar of rhythm (Hayes, 1995; Kager, 1999). Often inspired by and as well as inspiring metrics, this research has shaped further questions and answers arising from models like (d) and (f).

What, for example, are possible templates cross-linguistically? Hanson and Kiparsky (1996) suggest that templates stylize the rhythmic structure of language, and are therefore likewise composed of binary, headed constituents, though hierarchical organization of those constituents can produce surprising complexity; the iambic pentameter, for example, arguably has an inherently ambiguous template more like (h) than simply (d) (Kiparsky, 1977; Piera, 1981; Hanson, 1996, 2006 and others), with cola (groups of feet) and lines (groups of cola) themselves being constituents with a weak–strong form:

h. 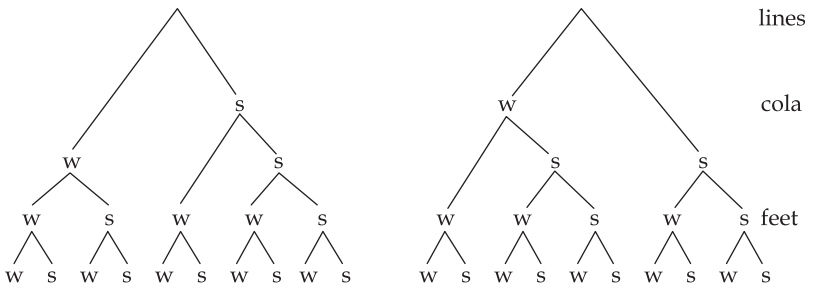

Fabb and Halle (2008) make similar claims about the hierarchical structure of the English iambic pentameter, though drawing on different phonological assumptions, using a different notation (the bracketed grids of Halle and Idsardi, 1995), and arguing for a different ontological status for the structure, with what functions like a template in their theory being generated by parsing of individual lines. Golston and Riad (2000) argue that templates are unnecessary, but only because, they argue, meters can be described directly through phonology.

What are possible correspondence rules? The familiar possibility of 'feminine endings,' the inclusion of an extra unstressed syllable at the edge of a line as in (i i), affords an early example of a metrical phenomenon inspiring a phonological analysis, specifically Hayes' (1981) now well-established hypothesis that in many languages a final constituent may be 'extrametrical' ('<>'), and left outside the basic groupings which define stress patterns. This allows reconceptualization of feminine endings as in (i i) as special correspondence rules grounded in phonology (Kiparsky, 1977); and a parallel reconceptualization of catalexis ('ø') as in (i ii) has followed suit (Kiparsky, 1991):

i. i. Make haste, the bridegroom he is come already

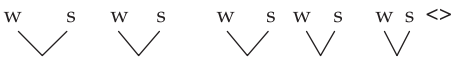

(*Romeo and Juliet* 4.4.27)

ii. ø Dead art thou! Alack, my child is dead,

(*Romeo and Juliet* 4.5.63)

More generally, the universal grammar of rhythm posits more entities than just syllables, including moras and phonological feet; more contrasts in prominence than just stress; and more ways that constituency matters to rhythm than just defining such contrasts (Hayes, 1995; Kager, 1999; McCarthy and Prince, 1993). Different ways have been suggested as to how all these combine in meters, sometimes with surprising results. For example, some forms of the modern English iambic pentameter permit 'resolution,' mapping of a pair of syllables into a single metrical position when stress and quantity are configured in particular ways, as in (j) (Kiparsky, 1977):

j. Where all the kindred of the *Capu*lets lie

(*Romeo and Juliet* 4.1.112)

Hanson and Kiparsky (1996) argue that an iambic pentameter with such resolution does not have correspondence rule (f i) mapping a syllable into each metrical position but rather a rule mapping at most a phonological foot into each, thereby relating it to other English meters permitting variation in syllable count, including Hopkins' Sprung Rhythm (Kiparsky, 1989), as well as to meters of Old English, Latin, Greek, Finnish, and other languages. All of these have different rules for stress placement than (f i), raising issues of formalization of that rule as well. From a different perspective, Fabb and Halle (2008) offer an account of resolution in Greek meters through rules that count moras. The development of an explicit theory of how the universal grammar of rhythm seeks to align edges of constituents (McCarthy and Prince, 1993) has similarly allowed traditional constraints on line endings and on placement of caesuras in meters of Greek, Serbo-Croatian, or French, for example, to be formalized as correspondence rules of a kind a grammar might posit (Hayes, 1988; Hanson, 2009; Hayes et al., 2012), and helped to explain why even in the absence of such rules in the English iambic pentameter, for example, different caesura placements have very different effects.

In sum, the universal theory of grammatical rhythm is so well-developed at this point (which is not to say agreed upon) that it has been possible to propose cross-linguistic theories of meter based on it. Hanson and Kiparsky (1996) derive unrelated metrical systems in English and Finnish from common principles; Deo and Kiparsky (2011) show how such principles shape metrical borrowing and change in Arabic, Persian, and Urdu; and Duffell (2008) and Ecke (2009) point to where they have shaped the history of the English metrical tradition. Fabb and Halle (2008) stands as the first book-length study to attempt to formalize meters across a range of languages including English, French, Spanish and Italian, Greek, Arabic, Sanskrit, Latvian, and Hebrew from a consistent theoretical perspective derived from generative phonological theory.

At the same time, while meter has become better understood through better understanding of grammatical structure, many open questions remain. For example, knowledge of a meter clearly includes recognition not only of the limits of well-formedness, but also of specific effects of metrical complexity that can be created within those limits (Halle and Keyser, 1972). How is that knowledge expressed in models like that in (d) and (f) (Youmans, 1983; Freidberg, 2002)? What role does the frequency of occurrence of particular structures play in the conceptualization of a meter as a whole (Gasparov, 1987; Golston, 1998; Hayes et al., 2012)? Interestingly, while such questions pose challenges for metrical description, they do not necessarily challenge the theoretical assumption of the close relationship of meter to grammar; rather, parallel questions have been posed about modeling grammar itself.

Finally, explicit metrical description and theory have brought renewed interest to the metaphorical characterization of meter as 'music.' The abstract rhythmic structures claimed for language are in important ways formally like those of music (Lerdahl and Jackendoff, 1983), and the problem of aligning language to a metrical template is in important ways like that of aligning it to a tune, or 'textsetting' (Liberman, 1975). Research into textsetting has helped explain metrical preferences, extend generative models of English meters from lines to stanzas (Hayes and MacEachern, 1998), challenge arguments for aperiodic meters (Deo, 2007), and reveal similarities and differences in meter and textsetting in languages from English, French, and Spanish (Dell and Halle, 2009; Rodríguez-Vázquez, 2010) to Tashlhiyt Berber (Dell and Elmedlaoui, 2008).

Alliteration and Rhyme

Meters interact with, but are formally distinct from, verse forms which impose constraints on melodic structure such as rhyme and alliteration (Hayes, 1988; Hanson, 2003; Ecke, 2009). While the latter are sometimes collectively referred to as 'sound patterning' (Fabb, 1997), the term is somewhat unsatisfactory since it is not clear why these forms have any better claim to be described as 'sound' than meter does, at the same time that sound is not really what any such forms involve, so much as phonological structure, as poetry in sign languages emphasizes. A more appropriate collective term for them, in light of the specific phonological structures they engage with, might therefore be '(poetic) melody,' parallel to '(poetic) meter.'

Again it was Jakobson (1963) who, in an influential analysis of Germanic alliteration, showed how explicit characterization of such forms can lead to discoveries about their relationships to grammar. In Old English verse, for example, syllables in certain positions must alliterate, beginning either with the same consonants, as in (k i), or with potentially different vowels, as in (k ii):

k. i. þa *m*iddangeard *m*oncynnes Weard ('Cædmon's
'then middle earth mankind's Guardian' Hymn')
ii. *e*ce Drihten *æ*fter teode ('Cædmon's
'eternal Lord afterwards made' Hymn')

Jakobson observes that no characterization of alliteration in terms of repeating sounds can generalize across this differential treatment of consonants and vowels, and Kiparsky (1968, 1973) characterizes it in terms of syllable structure instead (*see* Suprasegmentals). The most sonorous element in a syllable is its 'nucleus,' in English typically, though not always, the vowel; segments which precede the nucleus form the syllable's 'onset' while those which follow the nucleus form its 'coda.' The onset in 'trait,' for example, is [tr]; in 'rate' it is [r], and in 'ate' it is empty. Hence, the alliteration in (k) can be characterized as requiring alliterating syllables to have the same initial onset segment, a requirement vacuously satisfied in vowel-initial syllables which have no such segment.

While Germanic alliteration involves further complexities when variation in it across the tradition is taken into account (Minkova, 2003), Kiparsky's specification of its locus in terms of syllable structure stands as a model for parallel exploration of where repetition is required in other melodic forms. Assonance as in (l i), for example, would seem to require repetition in the syllable nucleus, and slant rhyme as in (l ii) to require repetition of the final segment in the syllable coda for some poets (Kiparsky, 1973) and in the syllable as a whole for others (Hanson, 2003):

l. i. Time let me hail and *climb*
Golden in the heyday of his *eyes*
(Dylan Thomas, 'Fern Hill')
ii. The lips of time leech to the fountain *head*
Love drips and gathers, but the fallen *blood*
('The Force that through the Green Fuse Drives Flower')

Full rhyme in its dominant form in English, though not in all forms, involves repetition in a domain which extends beyond the syllable to include segments in following unstressed syllables (Rickert, 1978):

m. i. Do not go gentle into that good *night*
Rage, rage against the dying of the *light*
('Do Not Go Gentle into that Good Night')
ii. Round the grief of the *ages*
Who pay no praise or *wages*
('In My Craft or Sullen Art')

That domain's best definition – possibly in terms of phonological feet, possibly in terms of phonological words – stands as an open question (Steriade, 2011), but one which has received recent attention especially in light of poets' exuberant engagement with it in hip hop and rap, not only in English but also in Japanese and other languages (Katz, 2014; Kawahara, 2007). Finally, even when syllable structure does not seem to be involved, as in the elaborate Welsh *cynghanedd*, a verse form in which consonant sequences are repeated in a specific order, other grammatical principles still seem to be involved, with Hammond (2012), for example, relating the form as illustrated in (n) to principles governing root-and-pattern morphology in Semitic languages, where a root like 'k-t-b' might occur with one set of vowels in the form *kataba* 'he wrote' and another in *kitaabun* 'book':

n. A*gwrdd drwst* | a *gerdda drá*w ('Y Gwynt')
'mighty commotion passing yonder'

Across the variety and complexity in which melodic repetitions defining verse forms are found in poetry, then, while much remains to be explored, what does not seem to be found is, as Kiparsky (1973, 1987) suggests, imaginable locations for them which would have no grammatical definition at all.

Poetic melody seems also to involve grammatical structure with regard to variation in what counts as repetition. In some forms, similarity rather than complete identity seems to suffice:

o. Deep with the first dead lies London's daughter,
Robed in the long fr*iends*,
The grains beyond age, the dark veins of her mother.
Secret by the unmourning water
Of the riding Th*ames*,
After the first death, there is no other.
('A Refusal to Mourn the Death, by Fire, of a Child in London')

Explicit description and theorization of such variation, too, has progressed in conjunction with key discoveries about melody in language. First, segments are not monolithic, but composed of distinctive features defining classes of segments (Trubetzkoy, 1969; Chomsky and Halle, 1968) (*see* Phonology). Second, those features' distribution and realization may be affected by their phonological context, so that individual lexical items' underlying phonological representations may not be their surface forms (Chomsky and Halle, 1968). These concepts have been invoked to explain melodic variation in a variety of ways, both in traditions where it is conventionally recognized and in ones where it is only more recently acknowledged. For example, in Irish, consonants are traditionally specified as belonging to

certain rhyme classes, and segments in corresponding positions in rhymes must belong to the same class; one such class is the voiced stops [b,d,g], so that *éc* 'death' (in which the 'c' spells [g]), for example, can rhyme with *bét* 'misdeed' (in which the 't' spells [d]) (Murphy, 1961). Malone (1987) shows that these same voiced stops are nasalized in certain contexts, making the rhyme class a clear instance of an equivalence in verse being one which figures in grammar. Moreover, these voiced stops can alliterate with nasalized stops in the same place, so that *bo* 'cow,' for example, may alliterate with *mban* 'white (gen.pl.).' Malone (1987) suggests that here, it is because the initial sound of the latter is underlyingly [b], even though the nasalization makes it different on the surface, that the two can alliterate, in an instance of a different way in which Kiparsky (1968, 1973) had suggested that poets' practice might reflect equivalences defined by grammar, namely, by being based on underlying phonological representations rather than on surface forms.

In English, variation in melodic forms has traditionally been criticized rather than codified; but in one of the earliest investigations from a generative perspective, Zwicky (1976) suggests that in rhymes of English rock lyrics, corresponding segments may differ by one feature, as [n] and [m] in Dylan Thomas' rhyme in (o) of *friends/Thames* differ in their place of articulation; he also notes that some segments may have no correspondents, as [d] does not in that same rhyme. Here too, focusing on individual poets' practices in the way studies of meter have done has revealed relationships with grammatical structure. Golston (1997), for example, relates the possibility of segments having no correspondents as used regularly in Bob Dylan's rhymes to generalizations about English syllable structure; and Hanson (2003) relates the possibility of featural differences like that between [m] and [n] as used regularly in Robert Pinsky's slant rhymes to parallel neutralizations in English grammar, where [n] and [m] are not distinguished before oral stops, for example, since nasals must be articulated in the same place as a following stop, as in *ramp, rant, rank* [ŋk]. Steriade (2003, 2009) argues that similar practices are found among Romanian poets even when no comparable neutralizations exist in the language, and suggests that, more generally, featural differences in poetic melody reflect the perceptual similarity of segments in certain positions, the same similarity which shapes both phonological alternations and historical alterations cross-linguistically.

Collectively, the foregoing ideas and the studies which support them suggest that melodic forms show a good deal of variety across verse instances in the same way that metrical forms do. Fully formalizing their verse designs, however, has generally received less attention, with fewer studies working out explicit models of possible complexities and their limits in the practice of individual poets, let alone characterizing them as variations on an overall tradition. A particularly interesting theoretical impetus to such studies has come, however, from cross-linguistic studies of reduplication. These systematic processes in which all or part of the beginnings or ends of words are copied to derive new words or forms of words were precisely what Kiparsky (1973) pointed to in support of his suggestion that the equivalences that define verse forms are those that figure in grammar, with the reduplicative Gothic past tense in (p i), for example, where 'ai' spells [ɛ], showing exactly the same pattern as the Germanic alliteration in (k ii), and rhyme involving similar parallels as in (p ii) (Yip, 1999):

p.	i.	Gothic past tense:	haitan 'call' slepan 'sleep' aukan 'increase'	hai-hait 'called' sai-slep 'slept' ai-auk 'increased'
	ii.	Chamorro intensives:	dankolo 'big' bunita 'pretty'	dankolo-lo 'really big' bunita-ta 'very pretty'

As the intricate variety of reduplication systems has begun to be studied and theorized (McCarthy and Prince, 1994; Inkelas and Zoll, 2005), cross-linguistic theories of poetic melody analogous to those advanced for meter have begun to be explored (Holtman, 1994), and poetic melody has begun to be recognized as a source of insight into reduplication as poetic meter has been for metrical phonology.

Syntactic Parallelism

Finally, not only phonology but also morphology and syntax define verse forms, through requirements of parallel structure from line to line (Jakobson, 1966). Especially influential for English has been the syntactic parallelism used in ancient Hebrew poetry (Alter, 2011), illustrated by the passage in (q) translated by Jakobson (1966):

q.	ʔitti with me	milləbanon from Lebanon	kallah bride
	ʔitti with me	milləbanon from Lebanon	tabo ʔi come!
	tasuri depart	mero ʔ from the peak of	amanah Amanah,
	mero ʔs from the peak of	sənir Senir	wəhermon and Hermon,
	mimmə¿onot from the lairs of		ʔarayot lions
	mehaŕ rey from the mountains of ('Song of Solomon' 4:8)		ń merim leopards!

Here the last two lines transparently repeat a common morphological and syntactic structure: the individual words differ, but the grammatical classes of their component parts and the wholes these form are identical: prepositions, plural nouns, etc.

Again, however, making the verse design explicit by specifying what, exactly, repeats in general raises interesting grammatical questions noted by Jakobson (1966) and reframed from an early generative perspective by Kiparsky (1973). The first two lines, for example, repeat individual words up until their last ones, which then differ, being the vocative *kallah* 'bride' and the imperative *tabo ʔi* 'come'; nonetheless,

these words share a function of direct address which entails morphological similarities, both being bare forms. The third and fourth lines also differ, in that there is a verb, *tasuri* 'depart' in the third which is absent from the fourth, and a conjoined noun in the fourth, *wəhermon* 'and Hermon,' which is absent from the third. From the point of view of syntax, however, conjunction (Conj) produces an instance of the same syntactic category of each conjunct on its own, conjunction itself being a form of parallelism; and in any case a preposition (P) which requires a noun phrase (NP) object will be indifferent to the internal structure of the noun phrase. Hence, in English, for example, 'from Mary' and 'from Mary and John' are identical at one level of structure as the rudimentary sketches in (r) show:

r.

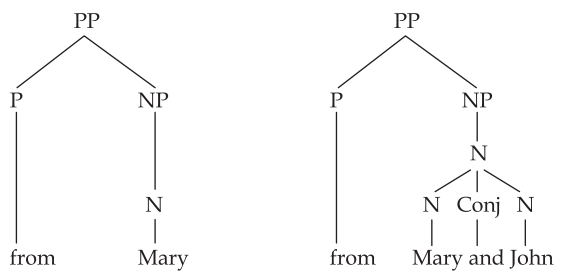

Moreover, when clauses (S) are conjoined, if a verb (V) is absent from the surface structure of the second clause, the verb of the first will nonetheless be understood to be present in it at some level of analysis: in 'Mary ate fish and John meat,' for example, 'ate' will be understood as the verb of the second clause, suggesting that both clauses maintain a consistent syntactic structure roughly as in (s), where 'ø' represents the absent verb:

s.

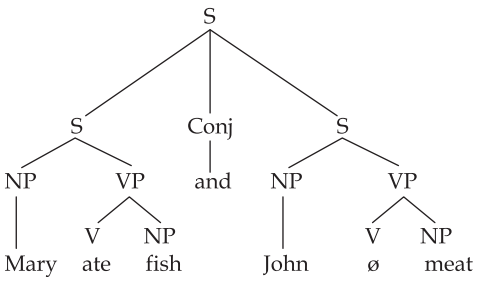

Thus, syntactic parallelism like that between lines three and four seems to involve morphological or syntactic identity at a somewhat abstract level, one more complex than any linear repetition of words or phrases or their types.

On the hypothesis of Kiparsky (1987), that level should be one that morphological and syntactic theories must posit to account for grammatical knowledge itself, independent of any poetic practice. Perhaps because English poetry only rarely uses syntactic parallelism in a sustained way, this hypothesis has not been so fully explored in the scholarly literature in English as its counterpart has been for other verse forms. Syntactic parallelism is, however, used as a verse form in diverse ways in many diverse languages – alongside the Hebrew, Russian, and Chinese discussed by Jakobson (1966), for example, it is used in Finnish, Galician, and many Native American languages (Bright, 1984; Hymes, 1998). Coupled with more recent theories of syntax, therefore, it affords rich material for exploration.

Conclusion

Versification represents only one of many ways in which language contributes to poetic structure. Not just the grammar but also the lexicon may be the source of significant patterns of repetition, engaging complex structures in storage of items (Tsur, 1992; Pinker, 1994). For every subsystem of grammar which defines a verse form by constraining some element of that system, other elements of that subsystem, as well as elements of other subsystems, remain free to create poetic effects of other kinds: phrasal rhythms (Cureton, 1992), syntactic suspensions (Austin, 1984), and more. And even the most traditional elements of verse are not necessarily used pervasively as verse forms, but sometimes only fragmentarily, or in traditional terminology 'ornamentally' rather than 'structurally.' In Shakespeare's sonnets, for example, alliteration and syntactic parallelism come and go within a form versified by meter and rhyme, in a manner which is not fundamentally different from the way meter and rhyme come and go in poems of Eliot which are not versified at all (Hanson and Kiparsky, 1997). Even as only a limiting case, however, discoveries of the relation of versification to grammar provide remarkable evidence of how forms which seem the height of artifice may in fact be rooted in ways of perceiving which, if the hypothesis of universal grammar is correct, operate as naturally as vision or hearing (Fodor, 1983). Such discoveries have consequences for poetic effects of all kinds, whether traditional or innovative, pervasive or fragmentary, natural or genuinely artificial.

Perhaps more strikingly, the focus on versification here has not touched on such fundamental functions of language as communication (*see* Communicative Competence: Linguistic Aspects), which are particularly complex and interesting in the case of poetry (Fabb, 2002). To some extent, this reflects the observation that communication is not unique to language: it has been argued by Sperber and Wilson (1986), for example, that in ordinary language, the phonological and syntactic parsing of a sentence makes only a minimal contribution to what it actually communicates, which in fact depends on an open-ended, context-dependent process of general reasoning which, unlike the structural analysis of grammar, is common to how information of all kinds is interpreted. At the same time, however, there is at least one way in which versification invites its own kind of reasoning. As Wimsatt (1954) notes, the repetitions which constitute versification consistently invite the question of how whatever is designated by similar forms might themselves be alike. In this way, as Jakobson (1956) suggests, there is a metaphorical way of thinking inherent in poetic structure, which is not only a matter of its meaning, but also of its form.

See also: Communicative Competence: Linguistic Aspects; Generative Grammar; Language and Literature; Lexicon; Markedness in Linguistics; Parts of Speech; Phonology; Pragmatics, Linguistic; Suprasegmentals.

Bibliography

Alter, R., 2011. The Art of Biblical Poetry. Basic Books.
Austin, T., 1984. Language Crafted: A Linguistic Theory of Poetic Syntax. Indiana University Press, Bloomington.
Bright, W., 1984. American Indian Linguistics and Literature. Mouton, Berlin.
Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.
Chomsky, N., 1972. Language and Mind, enl. ed. Harcourt, Brace, Javonovich, New York.
Chomsky, N., Halle, M., 1968. The Sound Pattern of English. Harper and Row, New York.
Cureton, R., 1992. Rhythmic Phrasing in English Verse. Longman, London.
Dell, F., Elmedlaoui, M., 2008. Poetic Meter and Musical Form in Tashlhiyt Berber Songs. Rüdiger Köppe Verlag, Cologne.
Dell, F., Halle, J., 2009. Comparing musical textsetting in French and in English songs. In: Aroui, J.-L., Arleo, A. (Eds.), Towards a Typology of Poetic Forms. John Benjamins, Amsterdam.
Deo, A., 2007. The metrical organization of Classical Sanskrit verse. Journal of Linguistics 43, 63–114.
Deo, A., Kiparsky, P., 2011. Poetries in contact: Arabic, Persian and Urdu. In: Lotman, M. (Ed.), Frontiers of Comparative Metrics. Peter Lang, Bern.
Dresher, E., Friedberg, N. (Eds.), 2006. Formal Approaches to Poetry: Recent Developments in Metrics. Mouton de Gruyter, Berlin.
Duffell, M., 2008. A New History of English Metre. Legenda, Oxford.
Ecke, J., 2009. The Metrical Imagination: Middle English Metrics and Alliterative Standardization (Ph.D. dissertation). English, UC Berkeley.
Fabb, N., 1997. Linguistics and Literature. Blackwell, Oxford.
Fabb, N., 2002. Language and Literary Structure. Cambridge University Press, Cambridge.
Fabb, N., Halle, M., 2008. Meter in Poetry: A New Theory. Cambridge University Press, Cambridge.
Fodor, J., 1983. The Modularity of Mind. MIT Press, MA, Cambridge.
Frawley, W., 1984. A note on the phonology of slant rhyme. Language and Style 17 (1), 44–47.
Freeman, D.C. (Ed.), 1970. Linguistics and Literary Style. Holt, Rinehart and Winston, New York.
Freeman, D.C. (Ed.), 1981. Essays in Modern Stylistics. Methuen, London.
Freidberg, N., 2002. Metrical Complexity in Russian Verse (Ph.D. dissertation). University of Toronto.
Gasparov, M.L., 1987. A probability model of verse. Style 21, 322–358.
Golston, C., 1997. The Geometry of Rhyme. Ms., California State University, Fresno.
Golston, C., 1998. Constraint-based metrics. Natural Language and Linguistic Theory 16, 719–779.
Golston, C., Riad, T., 2000. The phonology of Classical Greek meter. Linguistics 38, 1–69.
Halle, M., Idsardi, W., 1995. General properties of stress and metrical structure. In: Goldsmith, J.A. (Ed.), The Handbook of Phonological Theory. Blackwell, Oxford.
Halle, M., Keyser, S.J., 1972. The iambic pentameter. In: Wimsatt, W.K. (Ed.), Versification: Major Language Types. New York University Press, New York. Reprinted in Freeman, 1981.
Hammond, M., 2012. The phonology of Welsh cynghanedd. Lingua 122, 386–408.
Hanson, K., 2009. Metrical alignment. In: Aroui, J.-L., Arleo, A. (Eds.), Towards a Typology of Poetic Forms. John Benjamins, Amsterdam.
Hanson, K., 2006. Shakespeare's lyric and dramatic metrical styles. In: Dresher, B.E., Friedberg, N. (Eds.), Formal Approaches to Poetry: Recent Developments in Metrics. Mouton de Gruyter, Berlin.
Hanson, K., 2003. Formal variation in the rhymes of Robert Pinsky's The Inferno of Dante. Language and Literature 12 (4), 309–337.
Hanson, K., 1996. From Dante to Pinsky: a theoretical perspective on the history of the modern English iambic pentameter. Rivista di Linguistica 9 (1), 53–97, 1997.
Hanson, K., Kiparsky, P., 1996. A parametric theory of poetic meter. Language 72, 287–335.
Hanson, K., Kiparsky, P., 1997. The nature of verse and its consequences for the mixed form. In: Reichl, K., Harris, J. (Eds.), Prosimetrum: Cross-Cultural Perspectives on Narrative in Verse and Prose. D.S. Brewer, Cambridge.
Hayes, B., 1981. A Metrical Theory of Stress Rules (Ph.D. dissertation). MIT.
Hayes, B., 1988. Metrics and phonological theory. In: Newmeyer, F. (Ed.), Linguistics: The Cambridge Survey. Cambridge University Press.
Hayes, B., 1995. Metrical Stress Theory: Principles and Case Studies. University of Chicago Press.
Hayes, B., MacEachern, M., 1998. Quatrain form in English folk verse. Language 74, 473–507.
Hayes, B., Wilson, C., Shisko, A., 2012. Maxent grammars for the metrics of Shakespeare and Milton. Language 88, 691–731.
Holtman, A., 1994. A constraint-based approach to rhyme. In: Bok-Bennema, R., Cremers, C. (Eds.), Linguistics in the Netherlands, 1994. John Benjamins, Amsterdam, 49–60.
Hymes, D., 1998. When is oral narrative poetry? Pragmatics 8 (4), 475–500.
Inkelas, S., Zoll, C., 2005. Reduplication: Doubling in Morphology. Cambridge University Press.
Jakobson, R., 1960. Closing Statement: Linguistics and Poetics. Reprinted in Rudy, 1962–1981.
Jakobson, R., 1963. On the So-Called Vowel Alliteration in Germanic Verse. Reprinted in Rudy, 1962–1981.
Jakobson, R., 1966. Grammatical Parallelism and Its Russian Facet. Reprinted in Rudy, 1962–1981.
Jakobson, R., 1956. Two Aspects of Language and Two Types of Aphasic Disturbances. Reprinted in Rudy, 1962–1981.
Jespersen, O., 1933. Notes on metre. In: Linguistica. Levin and Munksgaard, Copenhagen.
Kager, R., 1999. Optimality Theory. Cambridge University Press.
Katz, J., 2014. Hip Hop Rhymes Mirror Phonological Typology. Ms., University of California, Berkeley.
Kawahara, S., 2007. Half-rhymes in Japanese rap lyrics and knowledge of similarity. Journal of East Asian Linguistics 16, 113–144.
Kiparsky, P., 1968. Metrics and Morphophonemics in the Kalevala. Reprinted in Freeman, 1970.
Kiparsky, P., 1973. The Role of Linguistics in a Theory of Poetry. Reprinted in Freeman, 1981.
Kiparsky, P., 1975. Stress, syntax and meter. Language 51, 576–616.
Kiparsky, P., 1977. The rhythmic structure of English verse. Linguistic Inquiry 8, 189–247.
Kiparsky, P., 1987. On theory and interpretation. In: Fabb, N., et al. (Eds.), The Linguistics of Writing. Methuen, New York.
Kiparsky, P., 1989. In: Kiparsky, P., Youmans, G. (Eds.), Sprung Rhythm.
Kiparsky, P., 1991. Catalexis. Ms. Stanford University.
Kiparsky, P., Youmans, G. (Eds.), 1989. Phonetics and Phonology. Rhythm and Meter, vol. 1. Academic Press, San Diego.
Lerdahl, F., Jackendoff, R., 1983. A Generative Theory of Tonal Music. MIT Press, Cambridge, MA.
Liberman, M., 1975. The Intonational System of English. Garland, New York.
Liberman, M., Prince, A., 1977. On stress and linguistic rhythm. Linguistic Inquiry 8, 249–336.
Malone, J., 1987. Muted euphony and consonant matching in Irish verse. General Linguisticsl 27, 133–143.
McCarthy, J., Prince, A., 1994. A generalized theory of alignment. Yearbook of Morphology 1993. Kluwer, Dordrecht, pp. 79–153.
Minkova, D., 2003. Alliteration and Sound Change in Early English. Cambridge University Press, Cambridge.
Murphy, G., 1961. Early Irish Metrics. Royal Irish Academy, Dublin.
Piera, Carlos, 1981. Spanish Verse and the Theory of Meter (Ph.D. dissertation). UCLA.
Pinker, S., 1994. The Language Instinct. William Morrow and Co., New York.
Rickert, W.E., 1978. Rhyme terms. Style 12, 35–46.
Rodríguez-Vázquez, R., 2010. The Rhythm of Speech, Verse and Vocal Music: A New Theory. Peter Lang.
Rudy, S. (Ed.), 1962–81. Selected Writings of Roman Jakobson, 8 vols. Mouton, The Hague.
Saussure, F.de, 2006. Writings in General Linguistics (C. Sanders and M. Pires, Trans.). Oxford Univeristy Press, Oxford.
Sperber, D., Wilson, D., 1986. Relevance: Communication and Cognition. Harvard University Press, Cambridge, MA.
Steriade, D., 2003. Knowledge of similarity and narrow lexical override. BLS 29, 583–598.

Steriade, D., 2009. The P-map and perceptibility effects. In: Hanson, K., Inkelas, S. (Eds.), The Nature of the Word: Studies in Honor of Paul Kiparsky. MIT Press, Cambridge.
Steriade, D., 2011. Rhyming Evidence for Intervals.
Tarlinskaja, M., 1976. English Verse: Theory and History. Mouton.
Toman, J., 1995. The Magic of a Common Language. MIT Press, Cambridge, MA.
Trubetzkoy, N., 1969. Principles of Phonology. University of California Press, Berkeley (C. Baltaxe, Trans.).
Tsur, R., 1992. Toward a Theory of Cognitive Poetics. North Holland, Amsterdam.
Wimsatt, W.K., 1954. One relation of rhyme to reason. In: Wimsatt, W.K. (Ed.), The Verbal Icon. Kentucky University Press, Lexington.
Yip, M., 1999. Reduplication as alliteration and rhyme. GLOT International 4 (8), 1–7.
Youmans, G., 1983. Generative tests for generative meter. Language 59, 67–72.
Youmans, G., 1989. Milton's meter. In: Kiparsky, P., Youmans, G. (Eds.), Phonetics and Phonology: Vol. 1: Rhythm and Meter. Academic Press, San Diego, pp. 341–380.
Zwicky, A., 1976. Well, this rock and roll has got to stop. Junior's head is hard as a rock. CLS 12, 676–697.

Language and Social Class

Suzanne Romaine, Merton College, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social class has occupied a central place in the study of language variation since the rise of the discipline of sociolinguistics during the latter half of the twentieth century. Sociolinguists have correlated sociolinguistic variables such as postvocalic /r/ (e.g., car) with social factors such as class, sex, style, age, etc. A major finding of urban sociolinguistic work is that differences among social class dialects are quantitative and not qualitative. Generally speaking, the use of nonstandard forms increases the less formal the style and the lower one's social status, with men's scores higher than women's.

Early Studies of Language and Social Class

Since the rise of the discipline of sociolinguistics during the latter half of the twentieth century, social class has occupied a central place in the study of language variation as a key factor offering important insights into social structure. Until the 1960s, most studies of variability were concerned primarily with regional variation or dialectology, following a tradition established in the nineteenth century. These studies concentrated their efforts on documenting the rural dialects which were believed would soon disappear. Only during the latter half of the twentieth century would the concern for status-based differences in language become a primary rather than a secondary focus, when sociolinguists turned their attention to the language of cities, where an increasing proportion of the world's population lives in modern times. The rise of urbanization is connected with an increase in social stratification reflected in linguistic variation.

Research focusing on social dialects is sometimes referred to as social dialectology and occupies a central place in quantitative sociolinguistic research on urban speech varieties, beginning with William Labov's (1966) work in New York City. Although linguists were aware before of differences in language use tied to social class, Labov was the first to introduce a systematic methodology for investigating social dialects and produced the first large-scale sociolinguistic survey of an urban community. Earlier, for instance, Alan Ross (1954) suggested that certain vocabulary differences in English could be classified as U (upper class) or non-U (lower class), e.g., *serviette* (non-U) versus *table-napkin* (U), one of the best known of all linguistic class-indicators of England at the time. Labov, however, concentrating primarily on pronunciation differences, called phonological variables, showed how idiolects (or the speech of individuals) considered in isolation might seem random, but the speech community as a whole behaved regularly. Previous investigations had concluded that the speech of New Yorkers appeared to vary in a random and unpredictable manner. Sometimes they pronounced the names *Ian* and *Ann* alike and sometimes they pronounced postvocalic /r/ (i.e., *r* following a vowel) in words such as car, while at other times they did not. This fluctuation was termed 'free variation' because there did not seem to be any explanation for it. Labov's study and subsequent ones modeled after it, however, showed that when such free variation in the speech of and between individuals was viewed against the background of the community as a whole, it was not free, but rather conditioned by social factors such as social class, age, sex, and style in predictable ways.

Unlike previous dialectological studies, which generally relied on one person (usually an older male) as representative of a particular area, most sociolinguists choose a sample of persons with different social characteristics. Labov's survey was based on tape-recorded interviews with 103 people who had been chosen by random sample as being representative of the various social classes, ages, ethnic groups, etc. to be found in New York City. This approach solved the problem of how any one person's speech could be thought of as representing a large urban area. Thus, while idiolects (or the speech of individuals) considered in isolation might seem random, the speech community as a whole behaved regularly. Using these methods, one could predict that a person of a particular social class, age, sex, etc. would pronounce postvocalic /r/ a certain percent of the time in certain situations. Through the introduction of these new methods for investigating social dialects by correlating sociolinguistic variables with social factors, sociolinguists have been able to build up a comprehensive picture of social dialect differentiation in the United States and Britain in particular, and other places, where these studies have since been replicated.

Methods

In order to demonstrate a regular relationship between social and linguistic factors, we have to be able to measure them in a reliable way. Of the principal social dimensions sociolinguists have been concerned with (e.g., social class, age, sex, style, and network), social class has probably been the most researched. Patterns of social class differentiation are often assumed to be fundamental and other so-called sociolinguistic patterns of variation, e.g., stylistic and gender variation, are regarded as derivative of them. Many sociolinguistic studies have followed the generally accepted sociological practice of relying on indicators of social status such as education, occupation, housing, income, etc. to group individuals into social classes, and then looked to see how certain linguistic features were used by each group. The method used in New York City to study the linguistic features was to select easily quantifiable

items, in particular, phonological variables such as postvocalic /r/, which was either present or absent. This was one of the first features to be studied in detail by sociolinguists.

Results

Phonological Variables

Varieties of English can be divided into two major groups with respect to their treatment of postvocalic /r/: those that are r-pronouncing (rhotic) and those that are not r-pronouncing (nonrhotic). Today in Britain accents that have lost postvocalic /r/ as a result of linguistic change generally have more prestige than those, like Scottish English, that preserve it. In many parts of the United States the reverse is true, although this has not always been the case. Table 1 compares the pronunciation of postvocalic /r/ in New York City with that of Reading, England. The results show that in New York City the lower one's social status, as measured in terms of factors such as occupation, education, and income, the fewer postvocalic /r/'s one uses, while in Reading the reverse is true. Like many features investigated by sociolinguists, the pronunciation of postvocalic /r/ shows a geographically as well as socially significant distribution. There are also internal linguistic constraints affecting the realization of postvocalic /r/. New York City speakers pronounced /r/'s in word-final position (e.g., *floor*) more frequently than they did in word-internal position (e.g., *fourth*).

Just as the diffusion of linguistic features may be halted by natural geographical barriers, it may also be impeded by social class stratification. However, the boundaries between social dialects tend for the most part not to be absolute. The pattern of variation for postvocalic /r/ shows 'fine stratification' or continuous variation along a linguistic dimension (in this case a phonetic one) as well as an extralinguistic one (in this case social class). The indices go up or down in relation to social class, and there are no sharp breaks between groups. A major finding of urban sociolinguistic work is that differences among social dialects are quantitative and not qualitative.

Sociolinguists have studied many other variables in English and other languages that show similar sociolinguistically significant distributions. Peter Trudgill (1974a) found similar differences in Norwich in an urban dialect study modeled after the New York City research. Table 2 shows the results for three of his phonological variables: (ing), (t), and (h). The numbers show the percentage of nonstandard forms used by different class groups. The variable (ing) refers to alternation between alveolar /n/ and a velar nasal /ŋ/ in words with *-ing* endings such as *reading*, *singing*, etc. The lower a person's social status, the more likely he or she is to use a higher percentage of alveolar rather than velar nasal endings. This is often referred to popularly as 'dropping one's *g*'s', and is a well-known marker of social status over most of the English-speaking world, including New York City, where Labov also studied it.

Table 1 Percent of postvocalic /r/s pronounced

New York City	*Reading*	*Social class*
32	0	Upper-middle class
20	28	Lower-middle class
12	44	Upper-working class
0	49	Lower-working class

Adapted from Romaine, S., 2000. Language in Society. An Introduction to Sociolinguistics. Oxford University Press, Oxford, p. 68.

Table 2 Percent of nonstandard forms in Norwich

(*ing*)	(*h*)	(*t*)	*Social class*
31	6	41	Middle-middle class
42	14	62	Lower-middle class
87	40	89	Upper-working class
95	59	92	Middle-working class
100	61	94	Lower-working class

Adapted from Trudgill, P., 1974b. Sociolinguistics. An Introduction. Penguin, Harmondsworth, p. 48.

The variable (h) refers to alternation between /h/ and lack of it in words beginning with <h> such as *heart*, *hand*, etc. Most urban accents in England do not have initial /h/ or are variable in their usage of it. For these speakers who 'drop their *h*'s', *art* and *heart* are pronounced the same. Again, the lower a person's social status, the more likely he or she is to drop *h*'s. Speakers in the north of England, Scotland, and Ireland, however, retain /h/ , as do speakers of American English. The variable (t) refers to the use of glottal stops instead of /t/, as in words such as *bottle*, which is sometimes stereotypically spelled as *bot'le* to represent the glottalized pronunciation of the medial /t/. Most speakers of English glottalize final /t/ in words such as *pat*, and no social significance is attached to it. In many urban dialects of British English, however, glottal stops are more widely used, e.g., in London and Glasgow.

By comparing the results for the use of glottal stops in Norwich with those for (ing) and (h), some interesting conclusions can be drawn about the way language and social class are related in this English city. Looking first at frequency, even the middle class in Norwich uses glottal stops very frequently, i.e., almost 50% of the time, but this isn't true of (h). There is much more social stigma attached to *h*-dropping. Phonetician Alexander Ellis (1869) regarded *h*-dropping as tantamount to 'social suicide,' and more than a hundred years later J.C. Wells (1982: p. 254) described it as the 'single most powerful pronunciation shibboleth in England.' There is of course no reason to assume that every instance of variation in language will correlate with social structure in the same way or to the same extent. Most sociolinguistic variables have a complicated history, and the social significance of linguistic features may vary over time. Some variables will serve to stratify the population more finely than others, and some cases of variation do not seem to correlate with any external variables, e.g., the variation between /i/ and /ɛ/ in the first vowel of *economic* is probably one such instance. Phonological variables tend to show fine stratification and there is more socially significant variation in the pronunciation of English vowels than in consonants.

As in the case of postvocalic /r/, what is socially significant is how frequently a person uses glottal stops in particular linguistic and social contexts. The use of glottal stops is particularly socially stigmatized in medial position, e.g., *bottle*, *butter*. Glottal stops are more likely to occur according to the hierarchy of linguistic environments in Figure 1.

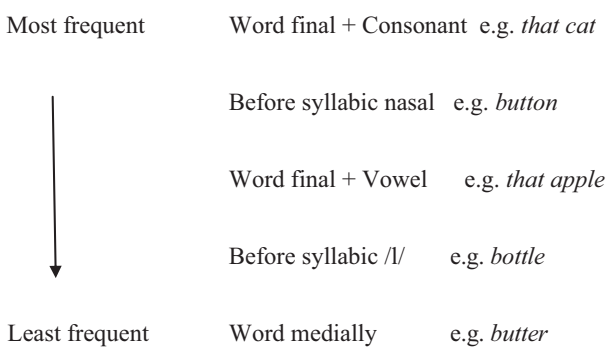

Figure 1 Linguistic environments for glottal stops.

Although all speakers are affected by the same internal constraints in the same way, they apply at different frequency levels, depending on social class membership and other external factors. Table 3 shows the incidence of glottal stops in relation to social class in Glasgow for all environments compared with that occurring only in medial position. Class I is the highest and contains professional people, while Class III is the lowest and contains unskilled workers. The results show that glottal stops are the norm for this community (74.3%) if we look at all the environments. Even the highest social class uses glottal stops nearly half the time and the lowest class, almost all the time. However, if we look at medial position, the highest social class uses no glottal stops in this environment, while the lowest class uses 68.8%.

The use of glottal stops has been noted as a feature of Glasgow speech since at least the late nineteenth century, but more recent sociolinguistic research shows that this feature is spreading from working-class urban speech, particularly that of London, and making rapid inroads into the speech of middle-class speakers in many other areas of southeastern England and Wales, and even as far as New Zealand, where it had not been heard before (Docherty et al., 2006). Hence, a feature once regarded as a vulgarism associated with the working class is becoming respectable (Wells, 1994: p. 201). Given its frequency in many areas of Britain across the social class spectrum, it would no longer be correct, strictly speaking, to classify this feature as nonstandard.

Grammatical Variables

Although grammatical variables have been less frequently studied than phonological ones, they have tended to show sharp stratification; that is, they often display a large social barrier between the middle class and the working class, as observed in the case of *h*-dropping. Table 4 compares the results of a study of a grammatical variable in Detroit and Norwich. The variable concerns the use of nonstandard third-person singular present tense verb forms without *-s*, e.g., *he go*. Only working-class speakers use these forms with any great frequency, and this is more often the case in Norwich than in Detroit. The gap between the middle- and working-class norms is also greater in Norwich than in Detroit, reflecting perhaps greater social mobility in the American social system. There are also other varieties of British English, e.g., in parts of the north, southwest, and south Wales, where the present tense paradigm is regularized in the opposite direction and all persons of the verb take *-s*, i.e., *I goes, you goes, he goes*, etc.

Another common grammatical feature of nonstandard English is so-called double or multiple negation, which involves the use of two or more negative elements where standard English requires only one. Compare nonstandard *he ain't got no money* with standard English *he hasn't got any money* or *he has no money*. This variable also has a long history going back centuries, but during the eighteenth century grammarians condemned the construction as nonstandard. The use of *ain't* has a similar history. This resulted in a decline in use of both *ain't* and multiple negation among the educated. In spite of correction by generations of teachers, both features have persisted in many nonstandard varieties. Table 5 shows the percent of multiple negative forms among African-American and white speakers in Detroit. There is sharp class stratification, with middle-class speakers regardless of race rarely using this feature. Overall, however, speakers of African-American English use multiple negation more frequently than whites.

Lexical and Discourse Variables

Sociolinguists have also extended the search for correlations between language and social class to encompass lexical and

Table 3 Percent of glottal stops in Glasgow in all environments compared to word medial position

Class	*I*	*IIa*	*IIb*	*III*	*Total*
All environments	48.4	72.9	84.3	91.7	74.3
Medial position	0	7.2	42.5	68.8	29.6

Adapted from Macaulay, R.K.S., 1977. Language, Social Class, and Education. A Glasgow Study. University of Edinburgh Press, Edinburgh, p. 45, 48.

Table 4 Percent of verbs without *-s* in Norwich and Detroit

Social class	*Detroit*	*Norwich*
Upper-middle class	1	0
Lower-middle class	10	2
Upper-working class	57	70
Middle-working class	–	87
Lower-working class	71	97

Adapted from Trudgill, P., 1974b. Sociolinguistics. An Introduction. Penguin, Harmondsworth, p. 44.

Table 5 Percent of multiple negation in Detroit

	Black	*White*
Upper-middle class	0	6.3
Lower-middle class	25.0	6.3
Upper-working class	41.1	31.8
Lower-working class	73.6	56.3

Adapted from Shuy, R., Wolfram W.A., Riley, W.K., 1969. Field Techniques in an Urban Language Study. Center for Applied Linguistics, Washington, DC.

discourse variables. Macaulay (2002), for instance, found that middle-class Scottish speakers in Ayr and Glasgow used adverbs ending in *-ly* (e.g., *slowly*) more than twice as frequently as working-class speakers. Other distinctions between the middle and working classes involved the use of *very*, which was almost exclusively a middle-class word, and *quite*, also used more predominantly by the middle class. There were also differences in the kinds of evaluative adjectives, with half of the adjectives used by the working class consisting of simple words of approval or disapproval such as *good*, *bad*, *nice*, etc. By contrast, the middle-class speakers used a wider variety of adjectives that were never used by working-class speakers, such as *horrible*, *messy*, *impressive*, *interesting*, *unattractive*, *fantastic*, *tremendous*, *impeccable*, and others. These differences are of a somewhat different nature than the ones previously discussed, where social class correlated closely with the boundary between standard and nonstandard usage. The higher incidence of adjectives and adverbs used by middle-class speakers served to emphasize the speaker's stance or personal viewpoint, while the working-class speakers on the other hand tended to leave their hearers to infer their point of view.

Studies such as these are also different in kind from Ross's early study of social class differences in vocabulary, where the variables investigated consisted of different words used to refer to the same thing, e.g., *lavatory* versus *toilet*. While some of the differences between so-called U (upper class) and non-U seem outdated and old-fashioned (e.g., U *looking glass* vs non-U *mirror* or U *wireless* vs non-U *radio*), and it has become increasingly common to think that access to higher education has increased social mobility and brought about a classless society, there is still considerable anxiety about social class and language in Britain. The issue of U and non-U language erupted in 2007 when Prince William broke up with then girlfriend Kate Middleton. The press speculated that perhaps she and her family, especially Kate's mother Carole, were just too middle class for the Royal family. Among Carole Middleton's linguistic faux pas the press commented on were her use of non-U *toilet* and *pardon* instead of *lavatory* (or the abbreviation *loo*) or *what*? Journalists speculated that perhaps a toilet had come between William and Kate, and headlines referred to a *toiletgate* scandal. Another linguistic misstep Carole Middleton made was to say 'Pleased to meet you' when she met William's grandmother, the Queen, instead of U 'How do you do?'

This incident indicates the enduring nature of linguistic insecurity and the divide between the upper and middle classes, despite the fact that some of the forms marking the boundary have changed. Today's non-U word or phrase may have been yesterday's U. *Toilet*, for instance, was a very smart word for the Edwardians, because it came from French, which was regarded as a prestige language. It went out of fashion, however, when their servants adopted it (compare also *serviette*, another term of French origin). Electronic corpora of modern English containing information about social class such as the British National Corpus can provide a handy way of checking the persistence of some of the social markers Ross mentioned (Romaine, 2008). The corpus contains one hundred million words of spoken and written British English from the 1990s. The spoken texts, comprising ten million words, include informal, unscripted conversation by speakers of different ages, regions, and social class. The British National Corpus categorized the speakers into four groups based on occupation: AB (top or middle management, administrative or professional), C1 (junior management, supervisory or clerical), C2 (skilled manual workers), and DE (semiskilled and unskilled manual workers). Although only 20% of the material in the spoken component of the corpus has been coded for speakers' social class, the corpus can still provide some insight into modern usage.

Table 6 shows the social distribution for *settee/couch/sofa* and *lounge/living room/sitting room*. The number of instances per million words is given for each term; boldface indicates the social group showing the highest usage for a particular term. The results show that social class differentiation in modern Britain is more nuanced than suggested by the simple division between U and non-U. While some words may distinguish the upper class from the rest, others may separate the working class from the lower-middle, or the middle-middle from the upper-middle classes. Ross, for instance, regarded *sofa* and *sitting room* as U and *settee* or *couch* and *lounge* as non-U. Looking first at variation in terms for the item of furniture, all four social groups use both *settee* and *sofa*; interestingly, the term *couch* does not occur for the highest and lowest social group. The lowest social group strongly favors the non-U term *settee*; the highest social group uses that term least. The term *sofa* occurs most frequently among class C2 followed by DE, but is less often used by two highest classes AB and C1. As for the room where this item of furniture is found, all social groups use all three terms. The middle and lower-middle classes (C1 and C2), however, are the greatest users of the supposedly non-U term *lounge*. The highest group leads in the use of the term *sitting room*, and the lowest in the use of the term *living room*. Thus, the upper class displays a tendency to sit on a settee in the sitting room, while the working class is more likely to sit on a settee in the living room, and the middle class to sit either on a sofa or couch in the lounge. These results display a rather mixed picture. While they appear to offer some support for the continuing U status of *sitting room* versus *lounge* and the non-U status of *settee* and *couch*, *sofa* does not appear to be an indicator of U status.

Often collocations, i.e., co-occurrences of words in context, are even more powerful indicators of social class than words in isolation. In 2012 the phrase *kitchen supper* triggered a national debate about class when Prime Minister David Cameron's Cabinet Office Minister remarked that the meals Cameron offered at his Downing Street flat to would-be Conservative

Table 6 Social distribution of *settee/sofa/couch* and *lounge/living room/sitting room* in the British National Corpus: hits/million words

Social class	*Settee*	*Sofa*	*Couch*	*Lounge*	*Living room*	*Sitting room*
AB	12.32	2.7	0	11.09	12.32	**13.55**
C1	18.02	2.57	5.5	32.18	9.01	9.01
C2	13.98	**8.39**	**8.39**	**48.93**	13.98	5.59
DE	**31.21**	4.46	0	8.92	**22.9**	8.92

p is less than or equal to .01; distribution is significant for *settee/sofa/couch*.
p is less than or equal to .001; distribution is significant for *lounge/living room/sitting room*.
Adapted from Romaine, S., 2008. Corpus linguistics and sociolinguistics. In: Lüdeling, A., Kytö, M. (Eds.), Corpus Linguistics. An International Handbook. Mouton de Gruyter, Berlin, p. 103.

donors were just 'kitchen suppers'. It meant relaxed, friendly meals in Mr Cameron's home rather than official dinners in a formal setting such as a Downing Street dining room. Terms for meals display social differentiation, with the evening meal typically referred to as *dinner* or *supper* by the upper class, but as *tea* by the working class, who use the term *dinner* for a midday meal. For the upper class *tea* is a late afternoon meal with sandwiches and baked goods like scones, and the midday meal is *lunch*.

Social Class Differentiation in Relation to Other Sociolinguistic Patterns

The intersection of social class and stylistic continua is one of the most important findings of quantitative sociolinguistics: namely, if a feature occurs more frequently in working-class speech, then it will occur more frequently in the informal speech of all speakers. There are also strong correlations between patterns of social stratification and gender, with a number of now classic findings emerging repeatedly. One of these sociolinguistic patterns is that women, regardless of other social characteristics such as class, age, etc., tend to use more standard forms than men.

Table 7 shows the results for the variable (ing) in Norwich in relation to social class, style, and sex. The scores represent the percent of nonstandard forms used by men and women in each social group in four contextual styles, i.e., when reading a word list, reading a short text, formal speech, and casual speech. Generally speaking, the use of nonstandard forms increases the less formal the style and the lower one's social status, with men's scores higher than women's. Although each class has different average scores in each style, all groups style-shift in the same direction in their formal speech style; that is, in the direction of the standard language. This similar behavior can be taken as an indication of membership in a speech community sharing norms for social evaluation of the relative prestige of variables. All groups recognize the overt greater prestige of standard speech and shift toward it in more formal styles.

Summing up these sociolinguistic patterns involving social class, gender, and style, sociolinguists would reply to the question of who is likely to speak most nonstandardly in a community: working-class men speaking in casual conversation. Conversely, middle-class women speaking in more formal conversation are closest to the standard. In Table 7, for instance, we can see that middle-middle-class women in word-list style never use the nonstandard form, while lower-working-class men use it all of the time. Note, however, that the differences between men and women are not equal throughout the social hierarchy. For this variable they are greatest in the lower-middle and upper-working class. Such patterns reveal basic linguistic fault lines in a community, and are indicative of the uneven spread of the standard and its associated prescriptive ideology in a speech community. Similar results have been found in other places, such as Sweden and the Netherlands (Nordberg, 2005).

Table 7 Social class, style, and sex differentiation in (ing) in Norwich

Class		*Style word list*	*Reading*	*Formal speech*	*Casual speech*
Middle-middle	M	0	0	4	31
	F	0	0	0	0
Lower-middle	M	0	20	27	17
	F	0	0	3	67
Upper-working	M	0	18	81	95
	F	11	13	68	77
Middle-working	M	24	43	91	97
	F	20	46	81	88
Lower-working	M	66	100	100	100
	F	17	54	97	100

Adapted from Trudgill, P., 1974b. Sociolinguistics. An Introduction. Penguin, Harmondsworth, p. 112.

Social Class Differentiation and Language Change

Because variability is a prerequisite for change, synchronic variation may represent a stage in long-term change. By examining the way in which variation is embedded into the social structure of a community, we can chart the spread of innovations just as dialect geographers mapped variation and change through geographical space. Sociolinguists have distinguished between 'change from above' and 'change from below' to refer to the differing points of departure for the diffusion of linguistic innovations through the social hierarchy. Change from above is conscious change originating in more formal styles and in the upper end of the social hierarchy, and change from below is below the level of conscious awareness originating in the lower end of the social hierarchy. Gender is a critical variable too. Women, particularly in the lower-middle class, often lead in the introduction of new standard forms of many of the phonological variables studied in the United States, the UK, and other industrialized societies such as Sweden, while men tend to lead in instances of change from below.

Because language, like fashion, is always changing, keeping up linguistically or sartorially with the proverbial Joneses is a never-ending task. Indeed, the Joneses may themselves decide to adopt once highly stigmatized features like the glottal stop, a behavior that has been referred to as 'linguistic slumming.' By contrast, in their effort to be linguistically correct, the middle and lower classes may hypercorrect and overshoot the norm of the group whose language they are aiming at, or use a false analogy resulting in an incorrect form. In the case of postvocalic /r/ in New York City, the lower-middle class shows the most radical style shifting, exceeding even the highest status group in their use of the standard forms in the most formal style. Another type of hypercorrection actually results in the production of incorrect forms. In New York City highly stigmatized pronunciations such as 'toity toid street' (for 33rd Street) are popularly associated with the so-called Brooklyn accent (although in fact are more widespread throughout working class speech of the city). Speakers who wish to dissociate themselves from this way of speaking often 'correct' forms containing the sound <oi>, so that *toilet* becomes *terlet* and *boil* becomes *berl*. Speakers alter these sounds because they are aware of the 'toity toid' stereotype, which suggests to them that <oi> is an incorrect pronunciation of <er>, so they correct all

cases where <oi> occurs to <er>. In cases like *terlet*, this produces a form that is not only hypercorrect, but also in fact nonexistent in the more prestigious form of speech that these speakers are trying to emulate. These hypercorrections often become so common that they too have now become stereotypes. People who drop *h*'s sometimes insert them in before words beginning with a vowel where they do not occur in the standard, resulting in hypercorrect forms such as *hextra* (*extra*), *hever* (*ever*), etc.

Hypercorrection happens because the behavior of the lower-middle class is governed by their recognition of an exterior standard of correctness and their insecurity about their own speech. They see the use of postvocalic /r/ as a prestige marker of the highest social group. In their attempt to adopt the norm of this group, they manifest their aspirations of upward social mobility, but they overshoot the mark. The clearest cases of hypercorrection occur when a feature is undergoing change in response to social pressure from above, i.e., a prestige norm used by the upper class. In New York City the new /r/ pronouncing norm is being imported into previously nonrhotic areas of the eastern United States. Hypercorrection by the lower-middle class accelerates the introduction of this new norm. The variable (ing), on the other hand, has been a stable marker of social and stylistic variation for a very long time and does not appear to be involved in change and hence does not display hypercorrection.

Alternatives to Social Class

Despite the successes of these investigations many sociolinguists have increasingly questioned the use of socioeconomic indices to determine speakers' class backgrounds and have explored alternative approaches to grouping people into discrete social classes based on indicators such as income, occupation, education, etc. This method assumes a consensus view of social class membership and that those in the same group will behave similarly. Some have argued, however, that social class differentiation is grounded in conflict rather than consensus. The concept of social class also applies more readily to western industrialized societies than to traditional rural village-based societies. Starting with the individual rather than the group allows us to overcome some of these limitations and to see patterns of variation and linguistic choices as a way of actively constructing social meanings and identities rather than merely reflecting membership in predetermined social categories.

Social Networks

One alternative to a class-based approach emphasizes the nature of contacts and networks in a society. Networks may cut across social class boundaries and reveal differences within social classes. Lesley Milroy's (1980) work in Belfast revealed important linguistic differences among working-class networks. Networks in which individuals interact locally within a well-defined territory and whose members are linked to each other in several capacities, e.g., as kin, neighbor, workmate, etc., act as a powerful influence on the maintenance of local norms. Two women, Hannah and Paula, who live in the same type of housing in the same area of Belfast and have similar employment, nevertheless, behaved quite differently from one another linguistically. Hannah is much more standard in her speech than Paula. The differences reflect their socialization patterns. Paula, whose speech is more nonstandard, is a member of a local bingo-playing group and has extensive kin ties in the area. Hannah has no kin in the area and does not associate with local people. In general, those with high network scores indicating the strength of association with the local community used more local, nonstandard forms of speech. Those whose networks were more open and less locally constrained used more standard speech.

Although there is a tendency for middle-class speakers to have looser networks than the working class, dense networks may also be found at the upper levels of society. The British upper class, for example, has been educated at elite schools and at Oxford and Cambridge, which has given rise to an equally distinctive speech variety that does not indicate their regional origins. More men than women had dense networks in Belfast, which suggests an explanation for some of the patterns of sex differentiation described earlier. The network approach has also been applied in non-Western settings such as Africa and Brazil. In Brazil, for example, it was used to study the extent to which rural migrants to urban areas assimilated to urban standard speech norms. Change has been slower for migrant women, who have fewer social contacts than men. The notion of network is potentially more useful than social class because it applies equally well to multilingual and monolingual settings.

Communities of Practice, Habitus, and the Linguistic Marketplace

Penelope Eckert's (2000) study of Jocks and Burnouts, two self-defined groups of high school students in Detroit, revealed that differing orientations to the school culture were correlated with opposing linguistic norms and styles of clothing. The Jocks aligned themselves with the middle-class values of the school and saw themselves as headed for college. The Burnouts oriented themselves to the local working-class world outside school, tended not to participate in school activities, and were not planning to attend college. Eckert situated her research within a theoretical framework drawing on notions such as community of practice and habitus. The concept of community of practice is closely related to that of social network. A community of practice shares ways of doing things, ways of talking, beliefs, and values. A habitus is a set of dispositions developed over the course of life through being embedded in a particular social position that influences patterns of behavior, affect, and thought. The notion of habitus is associated with the related concept of linguistic marketplace, both of which were developed by French sociologist, Pierre Bourdieu (1991). Linguistic practices function as commodities that are assigned value in a linguistic marketplace. Highly valued ways of speaking can be converted into actual economic capital by providing access to desirable career opportunities and employment. Conversely, nonstandard usage is an index of distance from or resistance to mainstream social and linguistic norms. The Burnouts, for instance, made significantly greater use of multiple negation than the Jocks, but a subgroup of

Burnouts (referred to as 'burned-out burnout girls') led the entire school, including burnout boys, in their use of this nonstandard feature.

These studies illustrate the importance of adopting a more dynamic view of language and of understanding how speakers use variable linguistic resources to construct identities in line with class-based language ideologies, i.e., beliefs and attitudes about language.

See also: Corpus Linguistics; Dialectology; Discourse and Indentity; Language and Society.

Bibliography

Bourdieu, P., 1991. Language and Symbolic Power. Polity Press, Cambridge.

Docherty, G., Hay, J., Walker, A., 2006. Sociophonetic patterning of phrase-final /t/ in New Zealand English. In: Warren, P., Watson, C.I. (Eds.), Proceedings of the 11th Australian International Conference on Speech Science & Technology. Australian Speech Science & Technology Association Inc, pp. 378–383.

Ellis, A.J., 1869. On Early English Pronunciation, with Especial Reference to Shakespeare and Chaucer. N. Trübner & Co., London.

Eckert, P., 2000. Linguistic Variation as Social Practice. Blackwell, Oxford.

Labov, W., 1966. The Social Stratification of English in New York City. Center for Applied Linguistics, Washington, DC.

Labov, W., 1972. Sociolinguistic Patterns. University of Pennsylvania Press, Philadelphia.

Macaulay, R.K.S., 1977. Language, Social Class, and Education. A Glasgow Study. University of Edinburgh Press, Edinburgh.

Macaulay, R.K.S., 2002. Extremely interesting, very interesting, or only quite interesting? Adverbs and social class. Journal of Sociolinguistics 6 (3), 398–417.

Milroy, L., 1980. Language and Social Networks. Blackwell, Oxford.

Milroy, L., Gordon, M., 2003. Sociolinguistics. Method and Interpretation. Blackwell, Oxford.

Nordberg, B., 2005. Social stratification in the present-day Nordic languages I: Swedish. In: Bandle, O., Braunmuller, K., Jahr, E.-H., Karker, A., Naumann, H.-P. (Eds.), The Nordic Languages 2: An International Handbook of the History of the North Germanic Languages. Mouton de Gruyter, Berlin, pp. 1759–1772.

Romaine, S., 2000. Language in Society. An Introduction to Sociolinguistics. Oxford University Press, Oxford.

Romaine, S., 2008. Corpus linguistics and sociolinguistics. In: Lüdeling, A., Kytö, M. (Eds.), Corpus Linguistics. An International Handbook. Mouton de Gruyter, Berlin, pp. 96–111.

Ross, A.S.C., 1954. Linguistic class indicators in present-day English. Neuphilologische Mitteilungen 16, 171–185.

Shuy, R., Wolfram, W.A., Riley, W.K., 1969. Field Techniques in an Urban Language Study. Center for Applied Linguistics, Washington, DC.

Trudgill, P., 1974a. The Social Differentiation of English in Norwich. Cambridge University Press, Cambridge.

Trudgill, P., 1974b. Sociolinguistics. An Introduction. Penguin, Harmondsworth.

Wells, J.C., 1982. Accents of English, 3 vols. Cambridge University Press, Cambridge.

Wells, J.C., 1994. The cockneyfication of RP? In: Melchers, G., Johannesson, N.-L. (Eds.), Nonstandard Varieties of Language. Almqvist and Wiksell International, Stockholm, pp. 198–205.

Wolfram, W.A., 1969. A Sociolinguistic Description of Detroit Negro Speech. Center for Applied Linguistics, Washington, DC.

Language and Society

Cliff Goddard, University of New England, Armidale, NSW, Australia
Anna Wierzbicka, Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 8305–8310, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

The vocabulary, grammatical structure, and usage conventions of any language are linked in innumerable ways with the social, cultural, and historical experience of its speakers. Drawing on examples from many languages, this article demonstrates the nature and range of these links. It highlights the culture-specific nature of many words, including terms for social categories, emotions, and value concepts. It explains the notion of cultural key words and the significance of lexical elaboration. In the domain of grammar, the article shows how grammatical constructions and marking may express and reflect culture-related meanings. In the realm of language in use, topics considered include discourse particles and interjections, linguistic routines, speech genres and speech styles, and broader discourse styles, demonstrating diverse ways in which language use can be related to differing cultural norms, values, and attitudes.

Introduction

Every language provides its speakers with a unique, intricate, and constantly changing medium of expression. The structure of any language embodies a myriad of prepackaged meanings, a large proportion of which are language- and culture-specific, in two respects: first, in not having exact counterparts in other languages of the world and second, in reflecting, embodying, and helping to perpetuate a particular social, cultural, and historical experience. The purpose of this article is to demonstrate this in relation to three linguistic domains – words and phraseology, grammar, and language in use – and to delineate some of the conceptual and methodological issues involved. A somewhat similar theme is pursued under the aegis of the Sapir–Whorf hypothesis (*see* Sapir–Whorf Hypothesis), which concerns the influence of language on worldview. This article, however, is concerned with more subtle and fine-grained effects.

These issues have been of pressing concern to language study from at least the time of Johann Herder and Wilhelm von Humboldt. They were carried into American linguistics and anthropology in the early twentieth century by Franz Boas and then Edward Sapir. Similar concerns continued to occupy anthropology, especially cognitive and symbolic anthropology, in the second half of the twentieth century; but in linguistics, these concerns became marginalized with the ascendency of Chomskyan generative grammar. Though, to some extent, the flame was kept alight by the ethnographic wing of sociolinguistics in the work of Dell Hymes and associates (cf Bauman and Sherzer, 1974; Gumperz and Hymes, 1986). In the 1980s and 1990s, however, the dominance of generative grammar began to wane, leading to what one commentator dubbed 'the greening of linguistics,' with renewed interest in pragmatics, functionalism, anthropological linguistics, and cross-cultural semantics (cf Foley, 1997; Goddard, 1998). In a parallel development, mainstream psychology became more interested in ethnopsychology and in the new field of cultural psychology (Shweder, 1990; Shweder and Sullivan, 1993). These developments set the stage for a revitalization of studies into the interconnections between language, culture, and society.

A complex of methodological problems connected with cross-cultural semantic analysis confronts this emergent interdisciplinary field. How can one ensure that any set of semantic descriptors is not biased by the language from which it is drawn? How to achieve the maximum resolution of complex meanings? How can the process of semantic analysis be freed from subjectivity and vagueness? According to the most well-developed model of cross-cultural semantic analysis, the 'natural semantic metalanguage' (NSM) approach, the most promising solution to these problems depends on identifying a viable set of universal semantic primes (or primitives), i.e., simple concepts that have linguistic exponents (as words or wordlike elements) in all languages (cf Wierzbicka, 1996; Goddard and Wierzbicka, 1994, 2002). Cross-linguistic research indicates the existence of at least 60 semantic primes, including (using English exponents) I, YOU, SOMEONE, SOMETHING, PEOPLE, DO, HAPPEN, SAY, THINK, KNOW, WANT, GOOD, BAD, BIG, SMALL, THIS, ONE, TWO, OTHER, THE SAME, WHEN/TIME, WHERE/PLACE, BECAUSE, IF, CAN, NOT, LIKE. This restricted and universal vocabulary of simple meanings can be used as a kind of 'semantic bridge' in terms of which to articulate complex language- and culture-specific meanings in maximum detail. Because it is grounded in simple words of ordinary language, this approach makes hypotheses about subtleties of meaning accessible to the intuitions and judgments of native speakers, thus improving the verifiability of the analyses. The NSM approach is not universally accepted, however, and much valuable work in cross-cultural semantics is still conducted in essentially informal terms.

Words and Phraseology

Even between genetically related languages spoken by the people of relatively similar cultures (such as the languages of Europe), there are sizable semantic differences in all aspects of the lexicon. Sapir (1949) was thus right when he said,

"vocabulary is a very sensitive index of the culture of a people." Most people have no difficulty in accepting the culture-specific nature of words for items of material culture and for social institutions, but many find it hard to accept the extent of cross-linguistic differences in terms for value concepts, emotions, speech acts, and social categories, to mention only a few examples.

Often it is possible to nominate certain highly salient and deeply culture-laden words in a language as 'key words' of that culture, i.e., words whose meanings can provide a key to understanding whole domains of cultural attitudes, values, and behaviors (Williams, 1976; Wierzbicka, 1997). For example, it can be argued that among terms for value concepts, English *freedom*, Japanese *wa* 'harmony,' and Malay *kesetiaan* 'loyalty,' 'faithfulness' have key word status. Among ethnopsychological concepts, the same can be argued for English *mind*, Malay *hati* 'heart,' and Russian *duša* 'soul.' Single-word glosses are never adequate, however, and a reasonable understanding of the key words of any culture requires dedicated and disciplined semantic investigation. Cultures vary across time, as well as across space; so, a close analysis of meaning shifts can be revealing of sociohistorical change. For example, it can be shown that the English word *friend* has weakened considerably in meaning since Shakespeare's time, when it designated an exceptional, intimate, and life-long bond of trust, affection, and mutual assistance. This shift can be seen as linked with a shift toward more flexible and broadly based (if 'shallower') mode of social relationships.

Languages differ not only in their inventories of culture-specific words, but also in the number of words they have for speaking about a particular domain of meaning. A relatively high number of words in a single domain (lexical elaboration) are often reflective of sociocultural facts. For example, compared with most indigenous cultures, European languages have a very large stock of expressions to do with measuring and reckoning time (words such as *clock*, *calendar*, *date*, *second*, *minute*, *hour*, *week*, *Monday*, *Tuesday*, etc., *January*, *February*, etc.). At the same time, there are linguistic universals of time, such as WHEN, AFTER, and BEFORE (cf Keesing, 1994).

Two further ways in which vocabulary and phraseology can reflect sociocultural facts concern the innovation of new words and the borrowing of words from other languages. In general, new words and phrases are indexes of social change; for example, the expressions *housemate* and *online* reflect social and technological developments of mainstream English-speaking culture of the late twentieth century. Linguistic borrowings index the extent and mode of cultural influence, as witness, for example, the massive borrowing of Chinese words into Japanese, of Latin and Greek words into English, or Sanskrit and Arabic words into Malay. Some scholars argue that permeability to such a borrowing is itself an index of the social openness of a community.

Grammar

Proponents of linguistic relativity argue that in as much as speaking a language obliges individuals to continually attend to particular semantic and conceptual distinctions, this must to some extent 'shape' the habitual thinking of its speakers. Approaching the issue this way means focusing on obligatory grammatical rules and processes. On the other hand, it can be argued that optional choices in grammar, i.e., aspects over which speakers may exercise some choice and control, are more likely to reflect subtle sociocultural meanings. A familiar example is the contrast between the so-called 'intimate' and 'formal' second-person pronouns in European languages, such as French *tu* versus *vous*. Much more extensive is the honorific system of Japanese, whereby verbs carry suffixes which indicate degrees of respect felt by the speaker in relation to the individuals being spoken about. In Slavic languages, systems of expressive derivation reflect the central place of warmth and affection in Slavic cultures; in Polish, for example, a personal name like *Maria* can have as many as 10 different derivatives (*Marysia*, *Marysieńka*, *Maryśka*, *Marysiuchna*, and so on), each reflecting a slightly different emotional attitude and mood. It will be useful to look in some detail at two more examples, from Russian and English.

'Fatalism' in Russian Grammar

The Russian language has a large family of impersonal dative-infinitive constructions, which refer to inexplicable things that happen to people against their will or irrespective of their will (Wierzbicka, 1992). For example, the sentence *Ne budet tebe pasporta* meaning 'There will be no passport for you' involves a dative pronoun, a negated existential verb, and a noun in the genitive case. The sentence conveys a subtle message, not fully evident in the English translation. Essentially, a construction of this kind rules out the possibility of someone being able to obtain something beneficial and desired, implying at the same time that this is due to the fact that someone in a superior position does not want it to happen.

Another construction combines negation, an infinitive verb, and a human noun in the dative case, for example, *Ne vidat' tebe ètix podarkov* meaning 'You will never see those presents.' This construction is not so common in everyday speech, but is highly characteristic of Russian folk literature and folk speech: *byliny* (folk epic), *plač* (lamentation), in folk song, proverbs, and sayings. It conveys the impression that a desired outcome can never eventuate due to some elusive and mysterious force or power that is, so to speak, 'above' humanity.

A third construction combines a human noun in dative case with a mental verb in the 3sg reflexive form, for example, *Segodnja mne vspomnilas' Praga-sady* meaning 'Today I was reminded of Prague, of its gardens.' This depicts a mental event simply happening inside a person inexplicably, and, in a sense, irresistibly. The most important Russian expression of this sort is the ubiquitous *xočetsja* lit. 'It wants itself to me,' which suggests a spontaneous and involuntary desire. In some ways, this parallels English constructions like *It occurred to me*… or *It seemed to me*…, but whereas the pattern has a limited scope in English, in Russian it is both fully productive and extremely common.

In these and other constructions, Russian furnishes an excellent example of the grammatical realization of a semantic theme; in this case, the theme of *sud'ba* ('fate'), of not being in

control, of living in a world which is unknowable and which cannot be rationally controlled.

Personal Autonomy and the English Wh-Imperative

Compared with many languages and cultures, English makes only scant use of its imperative construction, e.g., a simple *Do this*!, whose core meaning is 'I want you to do this.' Much more frequent is the use of one of a brace of more elaborate locutions (the so-called wh-imperatives), such as *Will you do this? Would you do this? Could you do this? Would you mind doing this?* and so on. These constructions can be seen as hybrids: in form they are questions, but since they convey the speaker's wish that the addressee do something, they have a similar function to imperatives.

It can be argued that the English wh-imperatives represent a linguistic reflection of the importance, in Anglo culture, of a cultural principle that can be labeled 'personal autonomy.' Roughly speaking, this calls upon people to respect the right of other people to make their own decisions about what they will and will not do. To use a bare on-the-record imperative could be seen to infringe this principle, hence the potentially confronting message that 'I want you to do something' is embedded into a more complex configuration, which acknowledges the addressees' autonomy by inviting them to say whether or not they will comply. In form, at least the addressee is presented with the option of refusing.

Similar examples of grammatical elaboration being driven or motivated by culture-specific pragmatics or cultural values could be given from many languages and cultures (cf Enfield, 2004). Often such phenomena arise from the routinization of particular culture-specific patterns of usage.

Language in Use

Discourse Particles and Interjections

These elements exist in the border zone between grammar and discourse. While they are usually not obligatory from a strictly grammatical point of view, they are often an essential part of native-like fluency. Interjections stand as independent utterances, whereas discourse particles are integrated into the syntax of a larger utterance; but both express a speaker's here-and-now feelings, thoughts, wants, and reactions to what is happening or to what is being said. When their semantics are examined in close detail, both types of element tend to be highly indicative of sociocultural priorities, values, and attitudes.

The very common Japanese particle *-ne*, for example, conveys the message 'I think you would say the same,' consistent with the Japanese cultural value placed on harmony and consensus-seeking. English interjections and interjectional phrases such as *well* and *on the other hand* reflect the Anglo emphasis on balance and tolerance of opposing views. A multiplicity of Polish discourse particles (such as *ależ*, *skądże*, and *przecież*) and related interjectional phrases express disagreement, exasperation, and impatience, consistent with the Polish preference for spontaneous expression of emotion and lack of any inhibition on expressions of conflict or disagreement.

Linguistic Routines

These are fixed, formulaic utterances or sequences of utterances used in standardized communicative situations, for example, greetings, partings, thanks, excuses, condolences, compliments, jokes, curses, small talk, and so on. They can range in size from a single word to lengthy interchanges, and usually occur very frequently. The overall meaning of a routine cannot be 'read off' from the literal meaning of the individual words involved, e.g., *How do you do?* is not a question about health.

In general, routines are highly culture-specific both in form and in the way they relate to the sociocultural context. To illustrate, one can compare some fixed expressions in English and Ewe (Ghana and Togo, West Africa). In many societies, when one realizes that something good has happened to another person, it is usual to say something expressing one's own good feelings at the news. For example, in English one might say *Congratulations!* or *Well done!*; both expressions imply that the addressee is responsible to some extent for the happy event. In Ewe, on the other hand, appropriate things to say refer to *Máwú* 'God,' *tạgbéwó* 'ancestors,' or *ŋú-wò núwó* 'beings,' expressing messages such as 'God is strong,' 'Ancestors are strong,' or 'Beings around you have worked.' Although one function is to register and display one's happiness at the good event, the Ewe formulae are not focused on the individuals concerned. This reflects the religious belief system, which holds that everything that happens is ultimately the work of *Máwú* who may work in diverse ways through the ancestors or other spirits and divinities.

Speech Genres

Bakhtin (1986) defined speech genres as 'relatively stable and normative forms of the utterance' into which we 'pour' our everyday speech. The concept is a very broad one, subsuming examples that differ in size and complexity as widely as (to illustrate from English) *joke*, *toast*, *speech*, *letter*, *eulogy*, *submission*, and *lecture*. Bakhtin stressed that the repertoire of genres available to a speech community changes according to historical and sociocultural conditions.

This conveniently can be illustrated with the (now passé) Polish 'microgenre' of the *kawał* (plural *kawały*) which, roughly speaking, was a kind of conspiratorial joke. Most *kawały* were political, expressing national solidarity vis-à-vis foreign powers, which have been so much a part of the Polish historical experience: the foreign partitioning powers in the nineteenth century, the Nazi occupation during World War II, the Soviet-imposed communist regime in postwar Poland. *Kawały* circulated widely, the anonymous creations of an oral culture. Like English *jokes*, they were intended to promote pleasant togetherness, i.e., to make the speaker and the addressee feel good together, but one did not value a *kawał* for its ingenuity, novelty, or sophistication (as one did, and does, *dowcipy* 'witty jokes'). More important was the feeling it gave of belonging to an in-group, the implication being 'I can say this to you, but there are some people I could not say this to.' For example, in 1981 after martial law had been imposed in an effort to suppress the Solidarity movement, every new demonstration, strike, or protest was ascribed to 'Solidarity extremists.' A *kawał* from that time gave examples from a 'TV

Dictionary' as follows: 'two Poles – an illegal gathering; three Poles – an illegal demonstration; 10 million Poles – a handful of extremists.' Clearly, though the Polish genre *kawał* may overlap in some respects with the English genre of a *joke*, it is not identical with its nearest English counterpart. Equally clearly, the precise nature of the Polish genre was a response to (and an index of) social conditions of Poland.

Speech Styles

Speech styles represent specialized 'ways of speaking,' that are recognized (and, usually, named) by the speech community as appropriate for use with certain categories of interlocutors or in certain situations. When examined in detail, both the stylistic conventions and social parameters of the speech styles can show much about sociocultural conditions. One common pattern can be illustrated from Yankunytjatjara (Central Australia), which has a special indirect speech style *tjalpawangkanyi* (*wangkanyi* 'talk') for use with people whose kin relationships call for constrained behavior, and a boisterous joking or teasing style (*inkatjingani*) for use with people whose kin relationships imply mutual acceptance and a lack of any power relationships (Goddard and Wierzbicka, 1997). *Tjalpawangkanyi* has a distinctive vocal delivery (softer, slower, higher pitch) and an exaggerated rising intonation, as if to give the impression that the speaker is merely musing aloud. Direct references to the addressee, including imperatives and vocatives, are carefully avoided. Overt expressions of denial, refusal, or disagreement are also scrupulously avoided. There is liberal use of discourse particles and interjections that express uncertainty, hesitation, and minimization. Another striking feature, reported also of respectful speech styles in many other places, is generality of reference. Speakers avoid using any expression which unambiguously indicates a person, place, or thing, preferring vague locutions like 'that one (person)' and 'around here.'

The social meaning of *tjalpawangkanyi* refers to the Yankunytjatjara social emotion *kunta*. Usually glossed as 'shame,' 'embarrassment,' or 'respect,' this involves a sense of social difference and discomfort in the other person's presence, and the desire to avoid doing anything which might cause the other person to think anything bad about one. The strongest *kunta* is evoked by the respectful *umari* ('avoidance') relationship between a man and his parents-in-law, which absolutely prohibits personal contact of any kind. Less-severe *kunta* is felt in the presence of the siblings or cousins of *umari*, and in other relationships where propriety is important, the very relationships for which *tjalpawangkanyi* is appropriate. In this culture, then, the indirect speech style can be seen both as a kind of partial avoidance and a way of giving voice to *kunta*. In other societies, indirect speech styles can serve quite different purposes.

Yankunytjatjara joking styles are at the other end of the spectrum to *tjalpawangkanyi*. They bend, flaunt, or even parody the normal conventions of interaction. All the linguistic forms excluded from *tjalpawangkanyi*, including imperatives, vocatives, contradiction, exclamations, and sensitive vocabulary items, flourish in this style. Participants delight in banter, mock insult, and make the most of any chance to defy, challenge, or demean each other. Sometimes there will be sexual innuendo and risqué comments that consciously play on kin-role expectations. In extreme versions, such banter becomes blatantly obscene. Joking, uninhibited speech styles are reported from many societies around the world. Although at a general level one can say that he or she expresses 'solidarity,' 'intimacy,' or the like, the rich social meanings involved cannot really be summed up in a few words, and detailed study reveals subtle but important differences.

Discourse Styles

At an even broader level, entire speech communities can be characterized as having certain tacitly shared understandings about how it is appropriate to speak in particular, culturally construed situations. In some societies it is normal for conversations to be loud, animated, and bristling with disagreement, while in others people prefer to avoid contention, to speak in well-considered phrases, and guard against exposure of their inner selves. In some societies, silence is felt to be awkward, while in others it is welcomed. In some societies, it is considered very bad to speak when another person is talking, while in others this is an expected part of a co-conversationalist's work (Tannen, 1981). Researchers in the ethnography of communication and in cross-cultural pragmatics are concerned to spell out the differing norms, values, and attitudes (sometimes called cultural scripts), which underlie and explain these behavioral differences (Saville-Troike, 1989).

Cultural norms about expressing emotions (what one feels) vary widely. For example, in mainstream Anglo societies, the expression of negative feelings is relatively constrained (cf 'cheerful' English routines such as *How are you? I am fine*). In Russian and Polish societies, by contrast, a greater value is placed on the immediate and spontaneous expression of emotions, 'bad feelings' as well as good ones. In yet other societies (for example, in Malay and Japanese societies), the cultural ideal is to maintain a calm equanimity of expression, moderating the expression of good feelings as well as bad ones. These differences are reflected also in facial expressions and other aspects of nonverbal communication. Needless to say, these characterizations are only approximate. Many modulations and variations are found across cultures.

Even within a single language group, there can be significant differences. For example, it can be argued that contemporary Anglo-American culture has developed some distinctive routines and expressions for expressing enthusiasm, optimism, and a 'positive attitude': *Wow! Great! How nice! That's fantastic! I had a terrific time! It was wonderful! Have a nice day!* Sociological studies of cultural attitudes toward emotions have also shown that Americans place an exceptional emphasis on enthusiasm.

Along with differences in emotional expression, other broad parameters of social variation in discourse style concern the expression of conflict or opposition between interlocutors' views (what one thinks) or wishes (what one wants). Japanese society, for example, is a society in which overt expression of differences and disagreements is discouraged, on account of the prevailing cultural value of *wa* (roughly, 'harmony'). For example, a well-known article by a Japanese author (Ueda, 1974) is titled *Sixteen Ways to Avoid Saying 'No' in Japanese*. In Jewish culture (Schiffrin, 1984; Blum-Kulka, 1982), on the

other hand, an overt and intense clash of views and evaluations is not only considered nonthreatening, but is often actively welcomed.

Differences like these not only contribute powerfully to the texture of everyday social interaction and to the character of interpersonal relationships, but also, arguably, exert an influence, through socialization, on the formation of personality types.

Editor's Note

A widely-cited and current text in Language and Society is Martin Montgomery (2013). Goddard and Wierzbicka's entry assures us that any language is inseparable from "the social, cultural and historical experience of its speakers." The language of conflict in the Middle East is a current case in point and an explicit theme in the work of the Palestinian linguist Yasir Suleiman. See his A War of Words: Language and Conflict in the Middle East (Cambridge University Press, 2004) or his 2013 book, Arabic in the Fray: Language Ideology and Cultural Politic (Edinburgh University Press). The leading journal in this field is Language in Society, published by Cambridge University Press and currently in its 43rd volume; all of the most frequently cited papers in this area are published there. Recent work in the area has focused on language and globalization; multilingualism; and language, social policy, and ideology. Examples of current work in these areas of research would include: Nikolas Coupland (2011), Ingrid Piller and Jinhyun Cho (2013), and Michal Krzyzanowski and Ruth Wodak (2011).

See also: Cognitive Anthropology; Communicative Competence: Linguistic Aspects; Cultural Evolution: Overview; Language and Social Class; Lexical Semantics; Pidgin and Creole Languages; Politeness and Language; Semantics; Sociolinguistics.

Bibliography

Bakhtin, M., 1986. In: Emerson, C., Holquist, M. (Eds.), Speech Genres and Other Late Essays (V.W. McGee, Trans.). University of Texas Press, Austin, TX.

Bauman, R., Sherzer, J. (Eds.), 1974. Explorations in the Ethnography of Speaking. Cambridge University Press, London.

Blum-Kulka, S., 1982. Learning to say what you mean in a second language: a study of the speech act performance of learners of Hebrew as a second language. Applied Linguistics 3, 29–58.

Coupland, Nikolas (Ed.), 2011. The Handbook of Language and Globalization. Blackwell, Oxford.

Enfield, N.J. (Ed.), 2004. Ethnosyntax. Oxford University Press, Oxford, UK.

Foley, W.A., 1997. Anthropological Linguistics. An Introduction. Blackwell, Oxford, UK.

Goddard, C., 1998. Semantic Analysis: A Practical Introduction. Oxford University Press, Oxford, UK.

Goddard, C., Wierzbicka, A. (Eds.), 1994. Semantic and Lexical Universals: Theory and Empirical Findings. Benjamins, Amsterdam.

Goddard, C., Wierzbicka, A., 1997. Discourse and culture. In: van Dijk, T.A. (Ed.), Discourse as Social Interaction, Discourse Studies: A Multidisciplinary Introduction, vol. 2. Sage, London.

Goddard, C., Wierzbicka, A. (Eds.), 2002. Meaning and Universal Grammar. Benjamins, Amsterdam.

Gumperz, J.J., Hymes, D. (Eds.), 1986/1972. Directions in Sociolinguistics. The Ethnography of Communication. Blackwell, Oxford, UK.

Keesing, R.M., 1994. Radical cultural difference: anthropology's myth? In: Pütz, M. (Ed.), Language Contact and Language Conflict. Benjamins, Amsterdam.

Krzyżanowski, M., Wodak, R., 2011. Political strategies and language policies: the European Union Lisbon strategy and its implications for the EU's language and multilingualism policy. Language Policy 10 (2), 115–136.

Montgomery, M., 2013. An Introduction to Language and Society, third ed. Routledge, Oxford.

Piller, I., Cho, J., 2013. Neoliberalism as language policy. Language and Society 41 (1), 23–44.

Sapir, E., 1949. In: Mandelbaum, D. (Ed.), Selected Writings of Edward Sapir in Language, Culture and Personality. University of California Press, Berkeley, CA.

Saville-Troike, M., 1989. The Ethnography of Communication, second ed. Blackwell, Oxford, UK.

Schiffrin, D., 1984. Jewish argument as sociability. Language in Society 13, 311–335.

Shweder, R.A., 1990. Cultural psychology: what is it? In: Stigler, J.W., Shweder, R.A., Herdt, G. (Eds.), Cultural Psychology: Essays on Comparative Human Development. Cambridge University Press, Cambridge, UK, pp. 1–43.

Shweder, R.A., Sullivan, M.A., 1993. Cultural psychology: who needs it? Annual Review of Psychology 4, 487–523.

Suleiman, Y., 2004. A War of Words: Language and Conflict in the Middle East. Cambridge University Press.

Suleiman, Y., 2013. Arabic in the Fray: Language Ideology and Cultural Politic. Edinburgh University Press.

Tannen, D., 1981. New York Jewish conversational style. International Journal of the Sociology of Language 30, 133–149.

Ueda, K., 1974. Sixteen ways to avoid saying 'no' in Japanese. In: Condon, J.C., Saito, M. (Eds.), Intercultural Encounters with Japan: Communication, Contact and Conflict, first ed. Simul Press, Tokyo, pp. 185–192.

Wierzbicka, A., 1992. Semantics, Culture and Cognition. Oxford University Press, Oxford, UK.

Wierzbicka, A., 1996. Semantics: Primes and Universals. Oxford University Press, Oxford, UK.

Wierzbicka, A., 1997. Understanding Cultures through Their Key Words: English, Russian, Polish, German, Japanese. Oxford University Press, New York.

Williams, R., 1976. Keywords. Oxford University Press, New York.

Language and Thought: The Neo-Whorfian Hypothesis

Antoni Gomila, Campus Carretera Valldemossa, Palma de Mallorca, Spain

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E. Hunt, volume 12, pp. 8310–8315, © 2001, Elsevier Ltd.

Abstract

During the last decade, a renewed interest in linguistic relativism has given rise to an exponentially growing body of literature concerning the influence of language on cognition. The current concern is not whether language influences cognition, but rather how much, and to what extent. After briefly characterizing Whorf's linguistic relativism, the current neo-Whorfian hypothesis is introduced, and a selection of the most relevant and influential studies is reviewed. The conclusion contends that the influence of language on cognition is not homogeneous, but depends on how much recoding is brought about during linguistic development.

Linguistic relativism finds its roots in Romanticism: in the reaction to the supremacist attitudes of the Enlightment thinkers, who were in the business of establishing hierarchies of languages in order to find the 'perfect' one (generally, French). Romantic thinkers, such as Herder, viewed peoples ('Volks') as incommensurable historical entities, whose worldviews are somehow condensed in their respective languages. In other words, language was viewed as the accumulated wisdom and repository of experience of a particular People.

This Romantic attitude found its proper place in the development of Anthropology as a science in the late nineteenth century. A privileged area of research in this regard was the documentation of the Amerindian languages as their speakers were being exterminated. This work naturally gave rise to the linguistic relativism hypothesis: what one can think is constrained or molded by the language one speaks. Whorf's originality (1956), in this regard, lies in his particular way of arguing for this view and in his effort to provide evidence for it, instead of taking it as an obvious postulate.

Whorf's reasoning sides with the functionalist American tradition. Following William James' views, Whorf describes the infant development as the process of making sense of the 'booming, buzzing' confusion of early sensory experience. Language provides a set of 'ready-made' categories. By learning a language, then, we acquire a categorical system that allows us to make sense of our experience. If different languages 'carve nature at different joints,' speakers of different languages will come to experience the world differently. To put the point in antirealist terms, as sometimes Whorf did, they will come to experience different worlds. Language, from this point of view, is not just a communicative tool but also a representational one.

This simplistic argument is not coherent, though: it does not apply to language the general point about sensory confusion. Language is thought to provide a way to structure one's experience, as if it were a salient part of one's experience in the first place, as if the sounds of language were not part of that initial confusion. In addition, there could be categories common to all languages. In other words, Whorf's relativistic hypothesis is grounded in taken-for-granted assumptions.

It is also true that Whorf suggests in some passages that thought is carried out in the particular language one speaks: "Thought takes place in a language, be it English, Sanskrit or Chinese" (Whorf, 1956: p. 283). But this is not a core aspect of linguistic relativism as such. The latter just requires that what can be thought is configured by the categories of one's language, that one's global conceptualization of experience derives exclusively from one's linguistic experience: the doctrine of linguistic determinism.

The contemporary neo-Whorfian versions of linguistic relativism (Lucy, 1992; Gumpertz and Levinson, 1996; Levinson, 2003; Casasanto and Boroditsky, 2008) avoid the epistemological and metaphysical assumptions of Whorf, and are not committed to the contention that thinking takes place in a natural language. The focus is on the central argument for linguistic relativism, which can be summarized this way:

Premise 1: Linguistic diversity: languages differ in their lexical and morphosyntactic rules and categories.

Premise 2: Linguistic determinism: the linguistic ways of categorizing human experience determine the cognitive ways of categorizing it.

Conclusion: The categorical structure of thought varies according to the language of the thinker.

Notice that this is more a deductive schema than a well-specified hypothesis. All sentences in this reasoning admit of different degrees of modal strength, while for any of them some sort of linguistic relativism follows. Thus, linguistic diversity may be underlined (Levinson, 2003) or downplayed, for instance, as parametric variation of linguistic universals (Chomsky, 2007). In the same vein, the second premise affords a strong reading (all cognitive categories are linguistic) or a weak one (at least some cognitive categories depend upon the language acquired) (Kay and Kempton, 1984). Understood this way, Whorf provides a concrete program of empirical research, to find out whether, and to what extent, it is true that linguistic structure influences cognition: select a way in which languages differ, and see whether speakers of those languages differ in some nonspecifically linguistic cognitive task (perception, memory); then, pay attention to how such cognitive abilities appear in ontogeny, to confirm whether differences appear only due to language mastering. Such a program has given rise in the last decade to a lively neo-Whorfian program in contemporary cognitive science, whose main contributions are novel paradigms, rigorous experimental control, and piecemeal examination.

In what follows, I will select the most outstanding recent contributions in this area. Then, I will discuss which conclusion is best supported by the evidence – the influence of language on cognition not being into question, but rather how much influence it has, in which areas, and by which mechanism. The conclusion will turn out to be that relativist influences do not happen across the whole cognitive system, but just in late-developing, more abstract, cognitive domains. They might also show up, though, in early-learned perceptual categories, as long as they involve recoding of sensory experience. There is no need for radical constructivist assumptions ('no perception without language') to accept Whorfian effects: they may just involve overcoming initial implicit preferences, which are not linguistically coded, and making some perceptual aspects more salient because of the way language drives attention to them, as cognitive development benefits from the abstract recoding fostered by lexical development (Gomila, 2012).

Color Terms Influence Color Perception and Memory

The Basic Color Terms of Berlin and Kay (1969) made a long-lasting impact by challenging the relativistic hypothesis. They claimed that color perception is independent of the color terms of one's language, and also, they downplayed linguistic diversity, claiming that languages differ in their color vocabularies in a systematic way. The outcome of the debate that followed was not a rejection of the notion of basic color terms, but rather a heightened awareness of the need to be more careful in experimental research in two directions: in introducing greater experimental control when looking for nonverbal cognitive effects of such different vocabularies; and in carrying out an ambitious world survey of color vocabularies, in order to assess the claim of a universal pattern of increasing differentiation of colors, the World Color Survey (Kay and Regier, 2003), which is made available through the online World Atlas of Language Structures (Kay and Maffi, 2011). It was realized, though, that linguistic diversity is an independent issue: linguistic differences may have a cognitive effect, even if these differences are not arbitrary or purely conventional.

A great deal of effort focused on color discrimination tasks to assess whether the categorical effect – the fact that discrimination is easier between category than within category – in color perception is sensitive to linguistic differences in establishing categorical borders. Thus, Kay and Kempton (1984) supported the role of linguistic categorization in perception through different categorical perception effects across languages. In Tarahumara, a Mexican indigenous language, there is just one term for the part of the spectrum that we divide between blue and green. Kay and Kempton reasoned that if language influences cognition, speakers of English would distinguish more strongly between color samples that they call 'green' and 'blue' than between color samples that fall into the same color category. They posited that this would be true, even if the differences in physical terms were the same in all instances, while for speakers of Tarahumara that would not be the case. The experimental paradigm consisted in three color samples, and the participants had to decide which one was the most different. As hypothesized, the Tarahumara speakers – lacking a verbal distinction – discriminated among the examples in relation to their physical differences, while the English speakers saw the similarities and differences along their lexical categories. Kay and Kempton reasoned that this could be due to the fact that the strategy of the participants could be that of labeling the samples in the first place, calling each sample 'blue' or 'green' – a strategy not available to the Tarahumara speakers. The sample that was categorized differently from the other two would be the one chosen as most different, even if the physical differences among the different samples were the same.

To test this explanation, they devised a second experiment. To block that strategy, this time, the three color instances were not presented simultaneously, but through a moving window that allowed the participants to see two of the three samples at the same time: the one on the left and the one in the center, or the one in the center and the one on the right. In each trial, the participant had to determine whether the left example was 'greener' than the one in the middle and whether the right example was more 'bluish' than the one in the middle. In order to answer which difference was the greatest: the one between the two 'greens' or the one between the two 'blues.' In so doing, the center instance was labeled both 'green' and 'blue' in the same trial, so that the discrimination could not be due to its categorization as one of the two. In this setting, the effect found in the first experiment disappeared, and speakers of English performed as the speakers of Tarahumara. Language, then, can be said to be influential in how we perceive similarities and differences.

This comparison also makes it clear that we do not perceive colors only when we have a term to denote them; it is not that without language we would not have color experiences. It is rather that language plays a role in structuring our perceptual space, particularly in memory tasks. Thus, Berinmo speakers from Papua New Guinea exhibit enhanced color discrimination from memory across Berinmo category boundaries, but not across English boundaries, while English speakers show the reverse pattern (Robertson et al., 2000). In general, if two colors are the extension of the same term in a particular language, speakers of that language will judge the two colors to be more similar and will be more likely to confuse them in memory than speakers of a language with a term for each color. These differences develop early in infants and coincide with the acquisition of color terms (Roberson et al., 2004).

Further studies also confirmed the Kay and Kempton's verbal interference effect: crosslinguistic differences in similarity judgments and recognition memory can be affected by direct verbal interference. This suggests that it is language, itself, that is involved 'online' during these tasks. Verbal interference has also been shown to affect speeded color discrimination and visual search tasks across the English blue/green boundary (Drivonikou et al., 2007; Gilbert et al., 2006).

But such linguistic effects have been found not just in nonverbal memory tasks, but also in discriminative tasks. Winnaver and his colleagues have recently compared Russian and English speakers on a simple similarity judgment task, focused on the blue part of the spectrum. It turns out that in Russian two 'basic' terms (according to the Berlin and Kay definition) correspond to the English 'blue': 'golubóy' and 'siniy.' The former is typically applied to the sky, the latter to the sea. Winaver et al. (2007) showed their participants three chips

of the blue spectrum: one being the sample, and two options from which to choose the one that's more similar to the sample. They reasoned that if linguistic representations are used to deal with ambiguous color samples – as claimed by critics of relativistic effects – then no language effects should appear in this task, given that it is an objective one that participants can solve with high accuracy. Reaction times were measured as an indication of the relative easiness of the task. Finally, the task had to be carried out under two conditions: with and without verbal interference. The main hypothesis concerns the effect of the categorical boundary on Russian speakers. If color vocabulary has cognitive effects, performance of Russian speakers should differ from English speakers with regard to perceptual discrimination performance across the boundary. Additionally, verbal interference should affect only Russian speakers. Therefore, it is the cross-category trials – when the participants had to choose between a 'golubóy' item and a 'siniy' item – in which critical differences were to be expected. Russian speakers should be faster at cross-category trials than within-category trials, whereas no difference is expected by English speakers. The results nicely confirmed the predictions: Russian speakers were faster at judging similarity cross-categorially than within-categorially, and verbal interference affected more Russian cross-category than within-category judgments.

Spatial Terms Influence Nonverbal Spatial Tasks

Languages offer different kinds of terms to talk about space. Levinson (2003) has distinguished between three main areas of spatial cognition which can also be found in language: deixis, topology, and frames of reference. Deixis concerns distance from ego ('this' versus 'that') and direction ('coming' versus 'going'), and involves radial rather than vector characterizations (it is possible to 'come here' from any direction). Topological distinctions concern relations of contact, containment, or propinquity between a 'figure' object and a 'ground' landmark. Finally, the frame of reference provides a way to specify directions – taking ego as polar coordinates – and it is thus most relevant for orientation and navigation.

Bowerman studied some crosslinguistic differences in topological terms (Bowerman and Choi, 2003). Thus, for instance, Choi and Bowerman compared how English and Korean speakers differ in which spatial relationships they pay attention to. In English, the prepositions 'on,' 'in,' and 'under' specify the possible relations between an object and a container. Korean, though, lacks a similar specification, but it marks morphologically whether an object is tightly or loosely limited by another (for instance, a finger by a ring versus a picture and the wall), an aspect of the situation which English does not specify. Their comparative studies revealed that English and Korean pay different attention to these spatial relations, concluding that acquisition of spatial semantics of the terms of one's language influences children's categorization of spatial relations. This does not mean that babies learning English are blind to the tight fit versus loose fit containment relations: all babies are sensitive to this difference by 9 months of age (McDonough et al., 2003). It is rather that – given that the language they learn does not pay attention to such differences – those differences lose cognitive saliency later on. A study with deaf children, though, suggests that without linguistic input, some spatial relationships take much longer to be understood (Gentner et al., 2013).

As regards frames of reference, three different lexical systems have been distinguished (Levinson, 2003):

1. Object-centric or intrinsic framework terms are terms that specify intrinsic positions of landmark objects to extract a direction (often by metaphorical projection of body parts). In English, such terms as 'heads' and 'tails' (mostly for coins, but extensible to other objects), 'front' and 'back,' or 'top' and 'bottom' specify a particular part of a landmark that can be used for orientation.
2. Egocentric or relative framework terms are indexical terms whose meanings require attention be paid to the position of the speaker. In English, some of the previous terms can be so used, if they specify a part of the speaker (whose corresponding direction, though, moves with the moving subject). Such terms as 'in front of,' 'to the right of,' or 'behind' are of this kind (as in 'it's in front of you,' 'the seat to my right,' or 'put it behind the table').
3. Allocentric or absolute framework terms are terms that specify absolute positions in space. In English, terms such as north or south are of this kind (as in 'travel North' or 'it's more to the South').

Most languages combine two or more of these systems. While in English allocentric spatial terms seem to be exceptional, they seem to be the rule in languages such as Tzeltal (a Mayan language); some Australian aboriginal languages, such as Guugu Yimithirr and Kuuk Thaayorre; and some Asian languages, such as rural Tamil, Longgu, and Arrernte (Levinson and Wilkins, 2006, which is a survey of linguistic diversity in space and movement description). Notice that such absolute directions are more arbitrary and abstract than intrinsic or relative terms, since they cannot be directly perceived; they do not depend upon endpoints or milestones, and orientation requires constant tracking of one's position in a fixed spatial map.

Brown and Levinson (2009) have studied the Tzeltal linguistic community of Tenejapa. The Tzeltal spatial vocabulary lacks terms corresponding to 'left' and 'right,' but has words corresponding to absolute positions, formed by abstracting from the slope of their land. In this case, absolute position terms correspond to 'uphill' (which is South) and 'downhill' (which is North), so that they say things like: 'Give me the uphill glass,' or 'the rope is downhillwards the bottle,' regardless of current position of the speaker. For orientations corresponding to East and West, the term is the same, corresponding to 'across.' Tenejapan people are oriented by reference to this framework at all times, while they can also use an intrinsic system of reference, with terms that refer to body parts. The question is whether this preference for an independent frame of reference plays a cognitive influence beyond language.

According to Levinson and Brown, the answer is yes: speakers of 'allocentric' languages differ from speakers of 'egocentric' languages in nonverbal tasks of spatial reasoning, of visual memory, and of gesture. The differences occur in forms related to their different lexical systems, thus providing

evidence for a role of language in cognition. Levinson and Brown set several different nonverbal spatial tasks for speakers of these languages, and compared their performance to that of speakers of 'intrinsic' and 'egocentric' languages (such as Dutch). Tasks required memory of spatial configuration, motion, and path direction. Thus, in a spatial reversal task, Tzeltal speakers saw the spatial ordering of four items on a table, for instance, and were asked to remember the spatial ordering of three of them. They had to reconstruct that 'same' ordering on another table, which they approached from the opposite side (a 180° turn). The Mayan put the objects in the same allocentric ordering as the original set (what was to the left was again to the left, despite the subject reversal), while Dutch control participants would go for the same egocentric ordering (what was to the left was now to the right). Similarly, in a spatial-reasoning task, participants were observed to find out whether they preferred a transitive ordering of objects in terms of a left to right (egocentric) or a north to south (allocentric) frame of reference. Again, the Dutch used the egocentric and the Tenejapans used the allocentric framework.

Notice that the strategy consists in presenting an ambiguous task, which is why it was described as the 'same,' as opposed to the same. The instructions could be understood both ways. Therefore, the evidence does not show that any of the participants could not solve the task the other way. In fact, further research showed that both groups could solve the task both ways (Li et al., 2005). The results, though, do show a different preferential understanding of the task, related to the dominant spatial vocabulary.

Numerals Are Required for Counting

Numbers, or rather, numerals, have been another outstanding area of research. Numbers are abstract entities, which cannot be grasped directly, but through a symbolic means of representation. Numerals are the terms language offers to denote numbers; for arithmetics, other, more specific symbols may be required (functions, operations), in addition to a graphical numerical representation. Languages also differ in their numeral systems, so the question arises: to what extent do these differences matter for arithmetic competence?

There is a broad consensus that some numerical concepts are independent of language. There is evidence that we share some basic notions with animals, such as 'more or less' and 'numerosity,' given that prelinguistic infants and some nonhuman animals can discriminate the numerosity of small groups of objects and can recognize that one is larger or smaller than another (Starkey et al., 1990).

However, when it comes to exact numerical concepts larger than four, it is clear that they are acquired through language. It is a prominent example of the concepts that depend on language, at least at two levels: developmentally, in that the term comes first and then the child has to somehow come to understand its meaning; and processually, given the evidence that their linguistic representation is involved in calculations and arithmetic tasks (Spelke and Tsivkin, 2001). The authors of this study found that bilingual subjects trained on new number facts in one language, recalled those facts faster and more reliably when tested in that language, than when tested in their other language.

As with color, though, languages may vary a lot in their numbering systems. Some languages may have a very short numeral repertoire, as in 'one-two-many,' while others avail themselves a generative system that guarantees a numeral for each possible number, such as that found in Western languages. Peter Gordon has focused on one of the 'one-two-many' languages: the Pirahä, an Amazonian language spoken by a tribe of hunter-gatherers (Gordon, 2004). The question is whether Pirahä speakers – having terms just for one, two, and many – are somehow impaired in arithmetic functions, such as counting, addition, or subtraction. According to Gordon, they are unable reliably to tell the difference between four and five objects placed in a row. In other words, Pirahä speakers perform like nonverbal infants and nonenculturated primates, lacking a symbolic means for counting. The procedure Gordon used was a sort of imitation game: he laid out a random number of familiar objects, like sticks and nuts, in a row, and the participant (there were seven of them), was supposed to do the same. For one, two, and three objects, the Pirahäs' row matched Gordon's sample. But when the number of objects was from 4 to 10, Pirahäs just approximated the number, with increasing deviation as the row grew longer. In another task, participants were shown several boxes, with different quantities of fishes depicted on top. Seconds later, they failed to take this difference into account in order to remember which box kept a hidden object. Similarly, for rhythmic tapping they could imitate tapping on the floor of up to three taps, they failed to mimic strings of four or five taps. Gordon observes that the Pirahä do not need to count in daily life, and concludes that this evidence demonstrates that numerical concepts beyond three are acquired through the acquisition of the corresponding numerals.

Similar results were obtained by Pierre Pica with speakers of Mundurukú (Pica et al., 2004), another Amazonian language which lacks numerals beyond 5, even though Pica and collaborators avoid a relativistic conclusion. While speakers of Mundurukú are able to compare and add large approximate numbers beyond 5, they fail exact arithmetic with numbers over 4 or 5. Again, this evidence is interpreted in terms of a dual system of numerical competence: a basic nonverbal system of number approximation, and a language-based counting system of exact number and arithmetic. However, it again remains to be established what the precise role language plays in the understanding of numbers. This must be determined in order to know if the language-based system is not a real improvement on numerical cognition (as Pica thinks), or if abstract understanding of concepts as language-dependent is a relativist effect (as Gordon affirms). The question is as follows: which relationship holds between the two systems, once the recursive number system made possible by numerals is in place? Is it a complementary one? Or is it rather a transformative one, given that exact number understanding opens up a whole new world of cognitive possibilities? At least part of the answer to this latter question will depend on how number understanding is developed. In other words, what does it take to grasp a recursive, potentially infinite system, like the number one?

A way to address this question can be found in Pica's observation that numeral expressions in Mundurukú are long,

often having as many syllables as their corresponding quantity. The words for three and four are polymorphemic: 'ebapug' means 'your two arms plus one,' and 'eba-dipdip' means 'your two arms plus one plus one.' This suggests the rudiment of a way to name quantities in a systematic way, by establishing a correspondence between quantity and syllabic structure. Whereas it is not a feasible procedure in general, it makes clear that counting is the required ability to master. Pica suggests that other languages of the Mundurukú family use a base-2 system of numerals – in which the corresponding terms of 'eba' get repeated – but even in that case, the most revelatory aspect is that it suggests a basic form of counting that is related to body parts. Similarly, the expression for 5 corresponds to 'a hand' or a 'handful.'

This required emphasis on counting as the key ability to go from numerosity and approximate quantity to exact number prompts us to refer in this context to Saxe's (1982) study of Oksapmin, as a sort of middle, transitional case. Oksapmin is a Papua New Guinea linguistic community, which at the time of study was undergoing a cultural change related to the introduction of money. This made counting practices more relevant, and required an amplification of the numeric system to deal with the new arithmetic operations involved in the use of money. According to Saxe, the traditional Oksapmin number system for counting is also finite and based on the body, but this time 27 consecutive numerals are available. As a matter of fact, the terms constitute an order sequence of body parts, starting with the thumb on one hand, and proceeding through the upper periphery of the body and then down to the other side. No other symbolic representation is available: number 14 corresponds to the nose, and the forearm is 7. This numeric system was used for counting and measuring. However, it was not put to traditional arithmetic use, and no term exists for mathematical concepts of division, or for rational and irrational numbers.

During the 1970s, money was introduced to the Oksapmin community, through paid work and commerce. This created a denomination problem, given the shortness of the numerical system. The adaptation that took place to deal with the new social need for referring to currency was a base-20 system, that recursively uses the basic terms in successive cycles, as required. Such a change went hand in hand with a reorganization of arithmetic thought: beginning with the new coin counting practices in commerce, for instance, arithmetic functions – such as addition and subtraction – later emerged. The culmination was the emergence of mental calculus – the ability to count without physical support (fingers in our case; upper body parts in the Oksapmin's case). Saxe (1982) found evidence of this cognitive transformation in the making, by comparing four groups of subjects with respect to addition and subtraction problems. The groups included men who had owned a trade store for at least 2 years; men who had returned from a period of work at a plantation but did not own a trade store; young adults who had not been to a plantation, and, therefore, had no money experience; and older adults who had only a cursory level of experience with the money economy. He could witness the process by which practice with money pushes for a recursive open-ended numerical system that is increasingly used for mental calculation. Addition problems were solved in that way by the most experienced subjects. A procedure of starting with the first figure and then following the bodily series for the other figure was used by less experienced subjects. This followed more simple strategies – like the double enumeration one – in which each figure, which corresponds to a numeral-body part, is mapped onto the corresponding body part to begin with, and then a procedure for establishing correspondence is tried. In general, while all groups could do addition with coins, subtraction with coins was difficult for the groups not habituated to money. Without coins, performance of all groups differed in proficiency, in a rank that went from nonceiling competence for experienced traders, to null for nonexperienced older adults. Nonexperienced adults, though, performed at a higher than chance level, an indication of a greater facility for learning.

In summary, Oksapmin exemplifies the central idea of the 'conceptual metaphor' theory: that understanding of abstract concepts is achieved in terms of more concrete ones. Most importantly, these concrete ones are 'image schemas' and spatial egocentric concepts based on the body (Núñez, 2008). While these basic schemas are thought to be primitive, there are many of them. Languages freely choose from this pool of conceptual resources in a rather arbitrary manner, which, however, has strong cognitive effects. In other words, numerals do not create number understanding out of nothing. When numerals are encountered by children in their process of language acquisition, they make sense of them in terms of one of the basic body schemas available that the group is already using to make sense of numbers. In this way, such schemas scaffold the individual cognitive development. Numerical problems can only be dealt with to a certain extent in such a system. Purely mental processes result from internalization of such practices, which also limits the scope of tractable arithmetic problems. Change in these systems is also possible by standard processes of cultural innovation, but what is kept constant is the bodily grounding of conceptual understanding.

Linguistic Differences in Temporal Concepts

As observed by Whorf, languages differ in terms of the time units they distinguish. As important as that is, languages also differ in the different aspects of time they require a speaker to morphologically mark (e.g., tense, aspect, and mode), and in terms of the means for temporal deixis they make available (e.g., now, tomorrow, and later). Moreover, most temporal concepts – just like abstract mathematical notions – are grounded in spatial metaphors (which entail different frames of reference), and languages also differ in which conceptual metaphors they use (Casasanto and Boroditsky, 2008).

For example, Boroditsky studied whether the preference for absolute spatial orientation also influenced these other cognitive areas (Boroditsky and Gaby, 2010). She gave participants a set of pictures showing some kind of temporal process (like an aging man or a growing crocodile) and asked them to arrange the pictures according to temporal order. They tested the participants in two separate places, differently positioned according to cardinal orientation. Previous studies had shown that time direction in English and Hebrew speakers is

conceived in terms of writing direction: left to right for English, right to left for Hebrews (Fuhrman and Boroditsky, 2010). Given that speakers of Kuuk-Thaayorre lack terms for left and right, they did not arrange cards from left to right, more often than from right to left. The pattern of their responses was rather east to west orientation (metaphorically grounded in the daily process of the sun): if they were placed facing South, they placed the cards left to right, but if they were placed facing North, they placed the cards right to left. Thus, speakers of Kuuk Thaayorre exhibit a greater navigational ability and spatial knowledge than speakers of languages with relative frames of reference. They keep track of where they are, even in unfamiliar landscapes or inside unfamiliar buildings, because otherwise they could not talk about space.

Many other metaphors are used in different languages: speakers of English tend to use a horizontal spatial metaphor (past is behind, future is ahead), whereas speakers of Chinese use a vertical metaphor (future is down, past is up) (Boroditsky, 2001). The demonstration that this is not just a *façon de parler*, but that it reveals how events in time are conceived is found in the example that, given a spot in space directly in front of the subject to stand for today, when asked to place a spot for yesterday and tomorrow, English speakers use a horizontal axis, while Chinese speakers use a vertical one (Boroditsky, 2007). A similar metaphorical understanding was found by Whorf in the Hopi language, which revealed a circular understanding of time.

Núñez's work on the Ayamaras' conceptual metaphors for time also deserves mention (Núñez and Sweetser, 2006). The Ayamara linguistic community is over 2 million people; they live in the Andes highlands of Bolivia, Perú, and northern Chile. They reverse the central Western spatial metaphor time: instead of the projection 'front of ego: future/back of ego: past,' they use exactly the converse one. The basic word for front, 'nayra,' is also a basic expression meaning past, whereas 'qhipa,' which is the basic word for back, also means future. Thus, for example, 'nayra mara,' which translates literally as 'front year' refers to the previous year. Again, this is not a purely linguistic difference (a sort of frozen metaphor), but it affects how Aymaras understand temporal relations. In this case, Núñez found the evidence in the deictic gestures that accompany Aymaras talk of past and future: deictic temporal reference involved pointing to one's back to refer to future events and to one's front to refer to past ones, with distance from one's self-indicating temporal distance.

Conclusion

We have just sampled the recent neo-Whorfian literature. But it is enough to find a common pattern in the evidence: language has a deep impact on cognitive development. It is not that all cognitive development is driven by language – that's too strong a position; but when linguistic coding plays a role on categorization, linguistic differences give rise to cognitive difference. These are not absolute differences, but rather force fields, that make some differences more salient and relevant.

Language exerts its influence on cognition through linguistic development. It changes the initial, sensorimotor, and imagistic medium of the mental representation of a restricted implicit system into a more powerful explicit system. Or it facilitates attention to dimensions of experience for which no initial preference is set. Lexical labeling is not just a matter of tagging a previously and independently constituted symbolic concept, but the way symbolic concepts are grasped (Lupyan, 2008). It has to be kept in mind, though, that the greater impact of language on cognition is made evident, not in comparing speakers of two languages, but in comparing verbal minds to nonverbal ones (Gomila, 2012).

See also: Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Logic and Linguistics.

Bibliography

Berlin, B., Kay, P., 1969. Basic Color Terms: Their Universality and Evolution. University of California Press, Berkeley, CA, and Los Angeles, CA.

Boroditsky, L., 2001. Does language shape thought? English and Mandarin speakers' conceptions of time. Cognitive Psychology 43, 1–22.

Boroditsky, L., 2007. Do English and Mandarin speakers think differently about time? In: Love, B.C., McRae, K., Sloutsky, V.M. (Eds.), Proceedings of the 30th Annual Conference of the Cognitive Science Society. Cognitive Science Society, Austin, TX, pp. 64–70.

Boroditsky, L., Gaby, A., 2010. Remembrances of time east. Absolute spatial representation of time in an Australian aboriginal community. Psychological Science 21, 1635–1639.

Bowerman, M., Choi, S., 2003. Space under construction: language-specific spatial categorization in first language acquisition. In: Gentner, D., Goldin-Meadow, S. (Eds.), Language in Mind. MIT Press, pp. 387–427.

Brown, P., Levinson, S., 2009. Language as mind tools: learning how to think through speaking. In: Guo, J., Lieven, E.V., Budwig, N., Ervin-Tripp, S., Nakamura, K., Ozcaliskan, S. (Eds.), Crosslinguistic Approaches to the Psychology of Language: Research in the Traditions of Dan Slobin. Psychology Press, New York, pp. 451–464.

Casasanto, D., Boroditsky, L., 2008. Time in the mind: using space to think about time. Cognition 106, 579–593.

Chomsky, N., 2007. Biolinguistic explorations: design, development, evolution. International Journal of Philosophical Studies 15, 1–21.

Drivonikou, G.V., Kay, P., Regier, T., Ivry, R.B., Gilbert, A.L., Franklin, A., Davies, I.R.L., 2007. Further evidence that Whorfian effects are stronger in the right visual field than the left. Proceedings of the National Academy of Sciences 104, 1097–1102.

Dryer, M.S., Haspelmath, M., 2011. The World Atlas of Language Structures Online, Chapter 132–135. Max Planck Digital Library, Munich. http://wals.info.

Fuhrman, O., Boroditsky, L., 2010. Cross-cultural differences in mental representations of time: evidence from an implicit nonlinguistic task. Cognitive Science 34, 1430–1451.

Gentner, D., Özyürek, A., Gürcanli, Ö., Goldin-Meadow, S., 2013. Spatial language facilitates spatial cognition: evidence from children who lack language input. Cognition 127, 318–330.

Gilbert, A.L., Regier, T., Kay, P., Ivry, R.B., 2006. Whorf hypothesis is supported in the right visual field but not the left. Proceedings of the National Academy of Sciences 103, 489–494.

Gomila, A., 2012. Verbal Minds: Language and the Architecture of Cognition. Elsevier, London.

Gordon, P., 2004. Numerical cognition without words: evidence from Amazonia. Science 306, 496–499.

Gumpertz, J.J., Levinson, S.C. (Eds.), 1996. Rethinking Linguistic Relativity. Cambridge University Press, Cambridge, UK.

Kay, P., Kempton, W., 1984. What is the Sapir-Whorf hypothesis? American Anthropologist 86, 65–79.

Kay, P., Maffi, L., 2011. Color terms. In: Dryer, M.S., Haspelmath, M. (Eds.), The World Atlas of Language Structures Online, Chapter 132–135. Max Planck Digital Library, Munich. http://wals.info/chapter/132 (accessed 17.08.2011.).

Kay, P., Regier, T., 2003. Resolving the question of color naming universals. Proceedings of the National Academy of Sciences 100, 9085–9089.

Levinson, S.C., 2003. Space in Language and Cognition: Explorations in Cognitive Diversity. Cambridge University Press, Cambridge.

Levinson, S.C., Wilkins, D.P. (Eds.), 2006. Grammars of Space: Explorations in Cognitive Diversity. Cambridge University Press, Cambridge.

Li, P., Abarbanell, L., Papafragou, A., 2005. Spatial reasoning skills in Tenejapan Mayans. In: Proceedings from the 27th Annual Meeting of the Cognitive Science Society. Lawrence Erlbaum, Hillsdale, NJ.

Lucy, J., 1992. Language Diversity and Thought: A Reformulation of the Linguistic Relativity Hypothesis. Cambridge University Press, Cambridge.

Lupyan, G., 2008. The conceptual grouping effect: categories matter (and named categories matter more). Cognition 108, 566–577.

McDonough, L., et al., 2003. Understanding spatial relations: flexible infants, lexical adults. Cognitive Psychology 46, 229–259.

Núñez, R., 2008. Mathematics: the ultimate challenge to embodiment. Truth and the Grounding of Axiomatic Systems. In: Calvo, P., Gomila, T. (Eds.), Handbook of Cognitive Science: An Embodied Approach. Elsevier, 2012.

Núñez, R., Sweetser, E., 2006. With the future behind them: convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construals of time. Cognitive Science 30, 133–146.

Pica, P., Lener, C., Izard, V., Dehaene, S., 2004. Exact and approximate arithmetic in an Amazonian indigenous group. Science 306, 499–503.

Roberson, D., Davies, I., Davidoff, J., 2000. Color categories are not universal: replications and new evidence from a stone-age culture. Journal of Experimental Psychology: General 128, 369–398.

Roberson, D., Davidoff, J., Davies, I.R., Shapiro, L.R., 2004. Colour categories and category acquisition in Himba and English. Journal of Experimental Psychology: General 133, 554–571.

Saxe, G., 1982. Developing forms of arithmetical thought among the Oksapmin of Papua New Guinea. Developmental Psychology 18, 583–594.

Spelke, E., Tsivkin, S., 2001. Language and number: a bilingual training study. Cognition 78, 45–88.

Starkey, P., Spelke, E.S., Gelman, R., 1990. Numerical abstraction by human infants. Cognition 36, 97–127.

Whorf, B.L., 1956. Language, Thought, and Reality. MIT Press, Cambridge, MA.

Winaver, J., Witthoff, N., Frank, M.C., Wu, L., Wade, A.R., Boroditsky, L., 2007. Russian blues reveal effects of language on color discrimination. Proceedings of the National Academy of Sciences 104, 7780–7785.

Language Contact

Brian D Joseph, The Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S.G. Thomason, volume 12, pp. 8315–8319, © 2001, Elsevier Ltd.

Abstract

With examples drawn from a variety of languages but especially those found in the Balkans, this article discusses language contact from a variety of perspectives. The author first places contact between speakers of different languages in its sociolinguistic context, then briefly traces the history of the study of language contact, followed by a discussion of the mechanisms of language contact, its manifestations, and its effects.

Introduction

Although language exists in the minds of individual speakers, it is also very much a social phenomenon, involving more than one speaker engaged with another in acts of communication. Thus, speakers of a language, any language, (except, of course, those with a small number of speakers who are not in a position to talk with one another) are always interacting with other speakers. Generally, the interaction takes place between speakers of the same language, but what does 'same language' mean here? It is known that no two speakers of what is ostensibly the same language speak exactly identically, and that, moreover, the line between different languages and different dialects of the same language is often very hard to draw and is generally open both to interpretation and to external manipulation, for example, for political or nationalistic purposes.

Thus in the normal course of language use, contact between speakers of nonidentical language systems takes place again and again. Contact between speakers of different dialects is at one end of a continuum of degree of identity in such situations; at the other end is contact between speakers of different languages. Moreover, even pairs of different languages may show different degrees of similarity, when one is set in opposition to another, so that Russian and Ukrainian are in many ways more similar to each other than either is to Hindi or to Thai. Linguists usually refers to such contact situations as 'language contact' or 'dialect contact,' as the case may be, even though it must be emphasized that the contact is between speakers in concrete communicative situations and not between languages or dialects as systems in the abstract. Moreover, even where one might recognize a construct such as a 'macrolanguage' to subsume all varieties of a given speech form or in cases where there is a standard language that is the default that one thinks of when referring to 'language X,' the speaker-to-speaker contact that is generally characterized as 'language contact' actually involves a dialect, that is, the particular variety of the language, the specific dialect, that each speaker is using in that communicative encounter.

The study of language contact has a long history, inasmuch as a mixing in of words and even other linguistic elements was recognized by ancient observers; the Romans, for instance, were aware of the presence of Greek words in their Latin, and certain spelling conventions, such as the use of the letter *y*, were used at one time by Romans writing Latin to mark words of Greek origin. And, as a true science of language emerged in the West in the nineteenth century, through the determination of the history of various classical languages of Europe, especially Latin and Greek, and their genealogical – a term often used now instead of the more traditional 'genetic,' because of the unwanted biological implications of the traditional term – relationship to other languages in Europe, such as Gothic and Irish, and in Asia, such as Old Persian and Sanskrit, the distinction between 'inherited' forms and structures, i.e., those passed down through regular language transmission, and 'borrowed' forms and structures, i.e., those that diffused into a language from an external linguistic system, took on a special significance. This distinction was a key part of the reasoning behind the establishment of regular sound correspondences between two or more languages (and the recognition of regular sound changes that follows from the identification of such correspondences) and of the demonstration of language relationships. To elaborate briefly, the point is that once one is able to isolate across two or more languages words and/or morphemes that are cognate, i.e., that have sprung from the same source and are inherited into each of the languages in question, and that by virtue of their cognacy show systematic correspondences of sound between them, it becomes possible to argue that forms that do not fit the patterns of correspondence may be later entries from some language into one or more of the languages being compared. For instance, a comparison of Latin *dent-* and English *tooth* reveals two consonant correspondences, $d \sim t$ and $t \sim \theta$, that recur in many other words (e.g., ***d**ecem* ~ ***t**en*, ***d**omus* 'house' ~ ***t**imber* ('building' in old English), ***t**res* ~ ***th**ree*, ***t**erra* '(dry) land' ~ ***th**irst*, inter alia) and that therefore provide a basis for determining that matches like ***d**ent-* ~ ***d**ental* or ***t**riplus* ~ ***t**riple* are due to contact-based influence, despite the identity of the relevant elements; indeed, some might say 'because of' rather than 'despite,' on the assumption that with matchings that are identical in one set and nonidentical in another set, the nonidentical ones are more likely to reflect true cognacy and the identical ones more likely to reflect a contact-related development.

This distinction was in part reflected in the controversy in the nineteenth century between the family tree modeling of language relationships, as espoused by August Schleicher ('*Stammbaumtheorie*,' cf. Schleicher, 1853), and the wave theory modeling ('*Wellentheorie*') of Johannes Schmidt (cf Schmidt, 1872). Schleicher's model reflected system-internal innovations that caused breaks in speech communities and ultimately

resulted in splits into multiple dialects and eventually multiple languages, while Schmidt's model reflected innovations that diffused across the (relatively clean) lines of descent of the Stammbaum model, thus marking changes that involved contact between speakers of differentiated speech communities, i.e., that spread through dialect or language contact. Although Schmidt's contact-based model seems to be at odds with Scheicher's line-of-descent model, they are really modeling different kinds of linguistic innovations, essentially changes induced by system-internal factors (Schleicher) and those induced by system-external factors (Schmidt). The importance of recognizing system-external factors is thus set against what is known about language-internal paths to differentiation; as the example of *tooth/dental/dent-* above shows, one can only understand what is due to contact by having a clear picture of what is inherited (and vice versa, one might well say, so that what is inherited becomes clear(er) once what is due to external influence is identified).

Nonetheless, despite the early importance that the recognition of contact between languages has played in the study of language in general but especially the study of language change, it is fair to say that in recent years, the study of language contact has come into its own, aided by the emergence of the study of pidgin and creole languages as a distinct subarea within linguistics on the one hand, and the recognition of the importance of bi- and multilingualism on a psychological and social level on the other. There is now a separate subdiscipline of 'contact linguistics,' complete with textbooks (e.g., Thomason, 2001; Winford, 2003; Matras, 2009), handbooks (e.g., Goebl et al., 1997/2008, Hickey, 2010), specialized studies (e.g., Thomason and Kaufman, 1988; Myers-Scotton, 2002), focused collections (e.g., Thomason, 1997; Siemund and Kintana, 2008), and the like. In what follows, the key concepts and constructs associated with the study of language contact are presented, by way of illustrating and explicating how vital the recognition of language contact is to the understanding of language in general, and how pervasive the effects of language contact can be.

Mechanisms of Language Contact

Language contact must begin with different language systems, and thus with some degree of bilingualism (using this as a cover term for any number of languages involved) either on an individual basis, where one speaker has abilities in two (or more) different languages, or on a societal basis, where groups of speakers of different languages live coterritorially, interacting with one another in the community at large and existing side by side (to be understood figuratively and broadly, so as to cover as well cases in which different language groups are essentially segregated in different parts of a city or region). In cases of individual bilingualism, especially when the speaker in question has a nativelike command of both languages, the language contact can be said to take place in the mind of the speaker. In societal bilingualism, however, the contact takes place between different speakers so that there truly is a social dimension to the contact. But in such a situation, in order for speakers of different languages to interact linguistically, at least one must have to some degree learned the language of the other, and in that phrase 'to some degree,' the seeds of language change through language contact are to be found. It is important to emphasize 'linguistic' interaction, because made-up and highly iconic gestures can be a basis for some communication, much as picture writing can be in the absence of a shared writing system; such gestures would foster communication between groups but without affecting their languages.

That is, in most instances of contact between speakers of different languages, one of the speakers involved is less than fully proficient in the language being used, typically having learned the language as a second language, an add-on to already well-entrenched language abilities. This second-language learning will often have come in adulthood.

At this point, an aside, but an important one, is in order. It is a common observation that younger children and perhaps even teens seem to be able to gain a native or nativelike command of a second language in ways and to a degree that older adults cannot; rather, adults seem generally to struggle with gaining such abilities. And, it is common, though somewhat controversial, to explain this observation in terms of a 'critical period' developmentally for language acquisition that allows younger speakers to gain a native command of a given language; such a period is usually thought to last into one's early teens or thereabouts, so that one could distinguish between the presumably perfect acquisition of a native language and the less-than-perfect command that typically comes with language learning later in life. While this hypothesis is doctrine among some linguists, there are adults who gain such a fine command of a second language that they can pass as native speakers, suggesting that other factors may be involved in second-language acquisition and that the occurrence of differential outcomes for child and adult language learners is not a purely biological developmental matter. That is, differential learning outcomes may not be an effect specific to the learning of language but could be due to differential motivations for learning on the part of the different groups, e.g., just for communication or for social indexing as well, or differential general learning abilities and strategies employed by the different individuals in the different groups.

This said, second-language learning for adults and even for many bilingual children will necessarily involve the second language being learned through the filter of an already existing language system, the learner's native language. This means that a speaker's native language can have an effect on a second language that one comes to speak later in life, after the native language is solidly in place in one's mind. These effects are often referred to as 'substratum effects' (the native language being a substratum that underlies production of a secondarily acquired language) or as 'interference' or 'transfer' effects (the native language patterns interfering with or transferring onto the production of the secondarily acquired language). Such substratum/interference/transfer effects can be observed in the difficulties second-language speakers can have learning to use elements with no direct analog in their native language. At the level of sound, for instance, one can note the persistence of native pronunciation patterns into a second language, i.e., speaking with an 'accent,' as when English speakers have trouble with the voiced velar fricative ([γ]) of modern Greek, and pronounce a name like [andiγóni] ('Antigone') as [andigóni], or Greek speakers have trouble with the voiceless palatal

fricative ([ʃ]) of English, and pronounce a name like [ʃejn] ('Shane') as [sejn], since these sounds are not part of the speaker's native language phonological inventory. Similarly, Russian speakers have trouble learning to indicate definiteness and indefiniteness in English through the use of the articles *the* and *a(n)*, since Russian has no such morphemes, relying instead on word order to express such discourse-sensitive notions.

Such difficulties in second-language learning are ubiquitous and almost inevitable, given that there is always a first language in place when an older speaker learns a second language. Moreover, these effects are important to note, as they give an indication of how a language in the broad sense could be affected by contact. The operative phrase here is 'in the broad sense,' since such substratum interference generally has no effect on the grammars of *native* speakers. Rather, the target language is changed only in the mouths (and minds) of those acquiring and using it secondarily; the French spoken with an American English accent by a university student studying French as a foreign language does not affect in any way the French spoken by a Parisian or by a Quebecois(e). Nonetheless, such interference might be relevant for contact-induced change in a couple of ways.

First, if the numbers of second-language users and their demographic concentration are sufficient, interference on an individual basis, repeated in speaker after speaker coming at the learning of a second language, perhaps out of necessity due to conquest or trade encounters or forced – or even voluntary – relocation, with the same first language background, can lead to the emergence of a new dialect of the target language, and thus to change in the overall makeup of that language. This has been the case with the emergence of Indian English, for instance, and, for that matter, with most so-called 'New Englishes' (see Schneider, 2003).

Second, a powerful force in language contact is accommodation (see, e.g., Giles, 1973; Giles and Johnson, 1987; Sachdev and Giles, 2004, among many other works, for discussion of accommodation), the process by which speakers alter their behavior, in this case their linguistic behavior, in the direction of that of their interlocutor(s), partly to enhance communication and partly to avoid uncomfortable encounters and to promote the saving of face for all participants. Thus a native speaker might well adjust his or her pronunciation of particular words in the direction used erroneously by a nonnative second-language speaker so as not to embarrass the nonnative speaker or draw attention to his or her error. Or, the native speaker might avoid complex syntactic constructions on the assumption that they might be too difficult for the second-language speaker to understand or respond to. In this way, the usage of the native speaker and the second-language speaker might converge in certain respects on the second-language speaker's usage, moving in a direction away from the native norm, or might reach some sort of 'compromise' variety between the two language forms. If the interaction is just a one-off event, there is really no effect on the target (native) language, but if such encounters are commonplace, and again if the numbers of second-language speakers are high enough, there could well be an overall effect on the usage of the native language speakers, especially in mixed company involving both native speakers and second-language speakers.

Another type of interference effect arises in cases in which a speaker is obliged in the society at large to use a language other than his or her native language on a regular basis, to the extent where the second-language patterns affect the native language patterns for that individual. Such effects can be called 'reverse interference' (see Joseph, 2009; Friedman and Joseph, 2015: Chapter 3), since it is the reverse of the more usual substratum situation described above. Cook (2003: 1) calls it 'backward transfer,' also 'reverse transfer', but 'reverse interference' is preferable because it does not always involve the transfer of features per se, but can involve just some influence (as with the Portuguese-on-English example given below).

Although not as well studied as substratum interference, reverse interference is no less real. A well-known case has to do with voice onset time (VOT), as described by Major (1992), who discovered that native speakers of English living in Brazil, i.e., in a Portuguese-dominant environment, showed VOT in their English stops that approximated but did not necessarily match exactly the Portuguese VOT, thus a sort of compromise between English and Portuguese VOT. Similarly, Hussein (1994) demonstrated that native Arabic speakers immersed for many years in an English-dominant environment had English-like vowel lengthening before voiced stops in their Arabic. Such 'reverse' effects also show up in speech perception; Caramazza et al. (1973) found that French–English bilinguals perceived stops in both languages in terms of the VOT of their second language, English, even though they did not show reverse production influence from English into their French, suggesting that this overall phenomenon is perhaps more complex than might first be supposed. A final example shows that reverse interference effects can be found in syntax as well as in phonology, for Young People's Dyirbal (YD), as described by Schmidt (1985: 230), shows evidence of English word order in the productions of YD speakers, saying "the exceptionally-free T[raditional]D[yirbal] word order is rigidified in YD as an A[gent]-V[erb]-O[bject] pattern as in English."

As mechanisms of change through language contact, interference (substratum/transfer) and reverse interference are not completely distinct. They have in common, of course, the fact of one language affecting another, but it may be a matter in each of the dominant language affecting the lesser one; 'dominant' here does not mean socially dominant, but rather cognitively or psychologically dominant in the perception of the individual speakers. Thus, what might make it possible for a speaker's native language to be affected by a secondarily learned language is that the speaker uses the second language in a wide range of contexts, almost to the exclusion of the native language. That is, native language attrition may be involved with the second language becoming the speaker's default language. In this way, the social dimension of bilingualism and the individual dimension converge, make possible change within the individual, and therefore change within the social group as well.

Manifestations of Language Contact

As the discussion in the previous section indicates, the main manifestation of language contact is the influence of one language system over another. The most straightforward effect is addition to the lexicon, what is traditionally called 'borrowing,' producing 'borrowings' or 'loanwords.' Terminology referring to 'borrowing' or 'loans' is widely used, but

is infelicitous for various reasons. For one thing, the donor language does not 'lose' anything and is not diminished in any way and 'borrowed/lent' items/features are not returned. One might make an exception to this latter claim in the case of reborrowings, where a word passes from one language into another and then, at a later stage back into the original donor language, as with English *tennis*, originally from French *tenez*! 'hold! receive!,' but now the source of present-day French *tennis*; however, even in such a case, there is really only a historical oddity to marvel at, not an overt decision by speakers to give back and thus repatriate some word or other. Furthermore, as traditionally used, 'borrowing' was (almost) exclusively lexical in nature (as suggested also by the other term, 'loanword'); linguists have come to realize, though, that more can be passed ('borrowed/lent') between languages than just lexical items, as is discussed more fully below. Despite the shortcomings of the traditional terminology, its use is commonplace, and not likely to change any time soon.

Loanwords have been classified in various ways. One classification schema focuses primarily on the form of the loan. Haugen (1950), for instance, drew a difference between 'importation' and 'substitution,' where the key distinction is whether markers of foreignness are present (in 'importations') or not (in 'substitutions'). English borrowings from Norman French with initial *v*-, like *veal*, would be importations, since the French sound was maintained more or less as is, even though at the time of borrowing, [v] was not found word-initially in English. A recent borrowing into English from (Mexican) Spanish such as *taco*, with aspirated [t^h] in the English in place of a Spanish unaspirated [t], which does not occur in initial position in English, would be an example of a substitution.

There can also be classification based primarily on the content of the loan, as in the schema of Bloomfield (1933). Bloomfield's distinction was between, on the one hand 'cultural borrowings,' those arising via the exchange, often mutual, of terminology for culturally based objects between speakers of different languages representing different cultures, as with many English words from American Indian languages, such as *moccasin*, from an Algonquian language designating at first just the soft leather shoe worn by the Indians, and on the other hand 'intimate borrowings,' those not obviously linked to cultural objects, such as *veal* noted above. Additionally, there can be classification focusing primarily on the motivation for the loan, as in the schema of Hockett (1958), who distinguished between 'need borrowings,' echoing Bloomfield's cultural type, but with the motivation of 'needing' a word for a (new) cultural item at issue, and 'prestige borrowings,' where the motivation is the 'prestige' that the borrowing language speakers accord to material from the donor language.

These typologies have been quite influential, but are not without some problems. For one thing, by focusing on form, Haugen's classification does not build in the social context for the loans, even though there is almost always the social dimension to borrowing of contact between speakers. An exception to this latter statement must be made for so-called 'learnèd borrowings' from an earlier, often (perceived as) stylistically elevated, stage of a language, as seen, for instance, in Latin borrowings into Romance languages, such as Spanish *digito* 'digit, number' from Latin *digitum* (here Spanish also has the form that developed by regular sound changes from *digitum*, namely, *dedo* 'finger'), and also in the Bulgarian and Russian use of Church Slavonic words, the adoption into Modern Greek *dimotiki* (low-style, colloquial language) of *katharevousa* (high-style, ancient-Greek-based) words, or Sanskritisms in modern Indo-Aryan languages (so-called 'tatsama' words). Moreover, the types listed above are not necessarily discrete; a cultural/need loan might be undertaken for reasons of (Hockettian) prestige or be associated with (Bloomfieldian) intimate contact. Also, noncultural/nonneed loans are not always a matter of prestige, at least not obviously so. For instance, the Albanian dialect of Greece known as Arvanitika, spoken in Greece for some 600 years, is the source of the diminutive suffix *–zə* that occurred in the early twentieth century at least in the Greek town of Megara, in Attica, as in *liγaza* 'a little,' formed from Greek *liγa* 'a little' with the Arvanitika diminutive added on; this occurred even though Albanian, and Albanians, have never been accorded much overt prestige by Greek speakers.

One may do well, then, to recognize a type of loanword that goes beyond these classificatory boundaries and focuses on the milieu in which the borrowing takes place. Friedman and Joseph (2014, 2015), for instance, have suggested that the medium of conversation must be recognized as crucial in borrowings and that in face-to-face interactions of an ongoing and sustained sort, various types of conversationally rooted borrowings transcending need or prestige and emerging just from repeated conversational encounters will occur. Such loans can be referred to as 'ERIC' loans, an acronym for 'essentially rooted in conversation' (that also honors Eric P. Hamp, historical linguist and contact linguist (and Balkanist) par excellence). Discourse-based items such as confirmatory words, hedges, attitudinal markers, and the like can be borrowed in such circumstances. In the Balkans, for instance, one finds 'yes' passing from Slavic (*da*) into Romanian and from Greek (*malista*) into Aromanian as well as negatives being borrowed, as with the Greek prohibitive negator *mi* 'don't…!' and the negative word *oxi* 'no' entering Southern Aromanian (as *mi* and *ohi*).

Further, there are some forms that are inextricably tied to conversation. For instance, a vocative element signaling unceremonious address, rather like 'hey you' in English, has spread all over the Balkans from a Greek starting point (with most forms deriving ultimately from the ancient Greek *mōré*, vocative of 'foolish'), as discussed in Joseph (1997); a sampling of the relevant forms is as follows:

Turkish	*bre*, *bire*, *be*
Albanian	*ore*, *or*, *mor*, *more*, *moj*, *ori*, *mori*, *moré*, *mre*, *voré*, *bre*, *be*
Aromanian	*More*, *mori*, *bre*, *be*, *or* …
Romanian	*bre*, *mă*, *măi*
Bulgarian/ Macedonian	*more*, *mori*, *bre*, *be*, *or* …
Romani	*more*, *mori*, *bre*, *be*, *or* …
Greek	*moré*, *bre*, *vre*, *re*, *aré*, *maré*, *marí*, *oré*, *voré*, *or*, *mor*, *mo*, etc.

Also, there are exhortatives, rather like English 'c'mon,' that have spread widely, such as Albanian, Bulgarian, and Macedonian *ela* 'c'mon' from Greek *ela* 'c'mon,' or the widespread *(h)a(γ)de*, from Turkish *(h)aγ de*.

Another classificatory schema for loanwords has to do with what is referred to as 'borrowability,' and it is suggested that there are some words that are less likely to be borrowed. At least as far back as Swadesh (1950) and his list of basic vocabulary items, linguists have worked with the assumption that a core of basic lexemes would be more resistant to borrowing than nonbasic items; this core varies from linguist to linguist, but includes items such as primary body parts ('head,' 'foot,' 'mouth,' etc.), omnipresent aspects of the natural world ('sun,' 'water,' 'cloud,' etc.), basic verbal actions ('eat,' 'say,' 'sit,' etc.), immediate kinship terms ('father,' 'mother,' 'brother,' etc.), and various closed-class or grammatical forms such as numerals, pronouns ('I,' 'this,' etc.), adpositions, complementizers, connectives, and the like. Still, it is possible to find examples of the borrowing of such items: kin-term loanwords occur among languages of the Balkans, e.g., Turkish *baba* 'father' ⇒ Greek *babás*, Albanian *baba*; Greek *patéra* ⇒ Aromanian *patera*; Turkish *dayı* 'maternal uncle' ⇒ Albanian *dajë*, Macedonian *daja*, as do pronominal loans, e.g., Greek *m(u)* 'my' ⇒ Aromanian *–m;* Turkish *bu* 'this' ⇒ Ottoman-era Edirne Greek *bu* 'this'; postpositions, e.g., Turkish *karşı* (*karşi* in Balkan Turkish) 'opposite' ⇒ Albanian, Aromanian, Bulgarian, Macedonian, and Ottoman-era Edirne Greek *karshi* (spelled differently in the various languages so that 'karshi' here is a cover form for various similar-sounding realizations of the Turkish word; it is noteworthy that the Aromanian and Greek forms were used as postpositions, while in the other languages, the form was adapted to the native pattern of prepositional use); numerals, e.g., Greek *efta, oxto, enja* '7, 8, 9' ⇒ Romani *efta, ohto, enna/enja*; complementizers, e.g., Turkish (<Persian) *zira* (std), *zere* (dial) 'because' ⇒ Bulgarian and Macedonian *zer* (with variants *zerem, zare, zerja, zeri*); as well as connectives, e.g., Macedonian *i* 'and' ⇒ Aromanian and Romani *i*.

Some linguists have seen constraints on borrowing as tied to the grammatical nature of the element, with bound morphology, both derivational and inflectional, being hard to borrow. Yet once again, counterexamples can be found. The Turkish agentive/occupational suffix *–çI/-cI* (the symbol *–I* here represents a high vowel that harmonizes for frontness and rounding with the final vowel of the stem that the suffix attaches to) has spread all over the Balkans (cf Greek *taksi-dzis* 'taxi driver,' Macedonian *lov-džija* 'hunter'), and the Turkish plural suffix *–lar/-ler* still occurs in present-day Albanian (e.g., *baballar* 'fathers,' *efendiler* 'sirs') and could be found in wider use as well in nineteenth-century Bulgarian.

One element of the debate over constraining borrowing has focused on whether only concrete material can be borrowed or whether structural aspects of a language can also be affected by contact with another language. Clear examples of structural patterns passing from one language to another, sometimes running counter to trends in place in the borrowing language, however, can be found. Turkish borrowed finite complementation with the complementizer *ki* from Persian, in contradistinction to the otherwise nonfinite-only complementation patterns in the language, and postpositional borrowing into prepositional languages is noted above, with Turkish *karşı* and in Ottoman-era Edirne Greek with Turkish *gibi* 'like' also. And, going beyond just the borrowing of the syntactic requirements of a lexical item along with the item itself, it has been argued that the wholesale replacement of nonfinite complementation by finite complementation in the Balkans, a characteristic found especially strongly in Albanian, Aromanian, Bulgarian, Greek, Macedonian, and Romani, and also in western Rumelian dialects of Turkish in contact with Greek and Bulgarian, is due to intensive contact among speakers of these languages, with modeling and copying of finite structure, even down to details as to the kinds of complementation (indicative as well as subjunctive), and accommodation to nonnative varieties both playing a role in the transfer of innovative patterns from one language to another (see Joseph, 1983/2009 for discussion).

Seeking constraints of a purely linguistic nature on borrowing and on transfer across languages more generally, therefore, may be an exercise in futility. Rather, it may well be, as Thomason and Kaufman (1988) claim, that under the right social conditions (e.g., having to do with intensity of contact), anything can be borrowed. The prevalence and apparent ease of lexical borrowings, therefore, as well as the seeming greater difficulty in borrowing bound morphology and structural patterns, may simply be a function of the fact that the social conditions leading to the intense contact needed for borrowing of morphology and syntax arise less often than the casual contact needed for lexical borrowing.

As a final point regarding lexical borrowing, it should be noted that besides loanwords per se, there are other kinds of lexical effects arising from language contact. In particular, contact can lead to loan translations, also known as 'calques,' in which the structure of a word in the source language is copied in the borrowing language but realized with native language material. For instance, Latin *compassio* 'compassion' was the model for German *Mitleid*, based on an equating of *com-* with *mit-* and *passio* with *leid-*. Even phrases can be calqued; English *it goes without saying* is a calque on the French *il va sans dire*, where the odd diathesis of the English (in particular, passivelike meaning but not passive form, i.e., not '… without being said'), can be attributed to the French model (where such diathesis is normal). It perhaps goes without saying here that calquing requires a certain degree of bilingualism on the part of the recipient language speakers, for otherwise the morpheme-by-morpheme or word-by-word identifying and translating would not be possible. A more subtle kind of lexical effect is seen in 'loan blends,' or 'hybrid' forms, where a slight or partial effect can be detected; for instance, in the Greek of southern Albania, the adjective *ekonomikós* 'economic,' with initial *e-* instead of the expected *i-* (as in standard Modern Greek, *ikonomikós*), seems to be a result of influence from Albanian, where the corresponding adjective is *ekonomík* – the placement of the stress in Greek reveals that this is not just a wholesale borrowing from Albanian but rather shows Albanian initial vocalism grafted onto the Greek word. Similarly subtle, but no less real, is contact-induced 'iso-semy,' where meanings shift in one language under the influence of the semantics of parallel forms; for instance, Greek *anixto-*, literally meaning 'open(ed),' is used with color terms to mean 'light-' (e.g., *anixto-kokino-* 'light red'), apparently due to the influence of the parallel use of Turkish *açık* 'open,' and also 'light-' with colors.

Effects of Contact

One common outcome of contact is simplification, and particularly striking in this regard is the creation in certain contact situations of pidgin languages - somewhat streamlined and highly simplified versions of linguistic systems put to use for rudimentary but essential communication, for instance, when commerce is involved, between groups of speakers having no common language between them. But contact can also lead to complexification, not just when pidgins are expanded functionally to other situations and become elaborated grammatically as creoles, accommodating to new functions, but also when borrowing brings in new structural possibilities alongside existing structures and adds expressive possibilities to a language. Thus it is probably safer to give the most typical outcome of contact-induced change as convergence between the languages involved. The term 'convergence' is used here to refer to an outcome of similarity between two languages, but it can also, sometimes somewhat confusingly and thus unfortunately, be used for the process by which such similarity arises. Convergence means that via borrowing, interference, transfer, calquing, and blending, languages in contact come to be more like one another along particular parameters, whether it is a matter of shared sounds or sound patterns, shared structures, shared words, shared phraseology, or whatever. The convergence is not necessarily unidirectional, of course, as the discussion above indicates, but convergence it is.

In some instances, the convergence is so great and to such a pervasive degree that the languages affected by contact seem to share a significant proportion of their structure. In such cases, especially where there is multilateral and multidirectional multilingualism and more than just casual contact, but rather sustained contact on an ongoing intimate, conversational basis, one can find a whole group of languages that come to share structure as well as lexical items. Such widespread areal convergence is often referred to as a 'sprachbund' (a borrowing from German) or 'linguistic area,' and while several sprachbunds (some linguists borrow the German plural here, *Sprachbünde*) have been identified around the world, e.g., in South Asia, in the Pacific Northwest of North America, in the Amazon, and elsewhere, the best known and most studied such convergence zone is the Balkans, where languages alluded to repeatedly herein – Albanian, Aromanian, Bulgarian, Greek, Macedonian, Romani, and (Western Rumelian dialects of) Turkish – converge not only on numerous lexical items of a noncultural sort but also on numerous points of structure including:

1. a reduction in the nominal case system, especially a falling together of genitive and dative cases;
2. the formation of a future tense based on a reduced, often invariant, form of the verb 'want';
3. the use of an enclitic (postposed) definite article, typically occurring after the first word in the noun phrase;
4. analytic adjectival comparative adjective formations;
5. marking of personal direct objects with a preposition;
6. double determination in deixis, that is a demonstrative adjective co-occurring with a definite article and a noun (thus, 'this-the-man');
7. possessive use of dative enclitic pronouns;
8. the use of verbal forms to distinguish actions on the basis of real or presumed information source, commonly referred to as marking a witnessed/reported distinction but also including nuances of surprise (admirative) and doubt (dubitative);
9. the reduction in use of a nonfinite verbal complement ('infinitive') and its replacement by fully finite complement clauses;
10. the pleonastic use of weak object pronominal forms together with full noun phrase direct or indirect objects ('object doubling'); and
11. the formation of the 'teen' numerals as DIGIT-'on'-TEN.

And there are far more convergent features when one looks just at the Balkans in very localized pairs or triples of languages rather than aiming for a broader coverage. This has led some researchers, most notably Friedman and Joseph (2015), to argue that even two languages can form a sprachbund, if enough convergence is observed, and that the difference between intense contact involving two languages and the sprachbund sort of convergence effect is a matter of difference of degree not difference of kind. Again, the social conditions on the ground are the key to the outcome, not some predetermined or arbitrary limit or constraint that a linguist might want to impose.

For all the fact that contact so often leads to convergence, there are contact situations in which the outcome is divergence between the languages involved, or at least, a failure to converge. In the Balkans, even with all the convergent forces at work, and even with considerable shared structure between Romani and other Balkan languages, Romani phonology has maintained its distance from the phonology of other coterritorial languages. In particular, Romani still has its voiceless aspirated stops, and thus diverges from the other languages in the region as being the only language in the Balkans with distinctively aspirated stops. It is not that phonology is immune to contact effects: the Aromanian spoken in Greece, for instance, has adopted Greek fricatives in loanwords without nativizing (substituting for) them and the fricatives have spread into some native Aromanian words even; the same can be said for some of the Macedonian dialects in Greece and Albania). Rather, it is the relative social isolation of Romani speakers and the one-way bilingualism that characterizes interactions the Roma have with speakers of other languages may have helped to promote the maintenance of a distinctive Romani character through its phonology, with the social factors therefore contributing to divergence as an outcome of contact. Once again, the social dimension cannot be ignored in a full consideration of what goes on in a language contact situation, and may be more of a determinant than anything purely linguistic in nature.

Conclusion

To sum up, it must be admitted that contact between speakers is a powerful force in shaping languages and thus language contact is an important synchronic fact about speakers' lives as well as an important force in language change, leading to what is referred to as 'externally driven' or 'externally motivated'

change. But in a sense it must be seen as intrinsically connected to internally driven change, what Andersen (1973) refers to as 'evolutive change,' not wholly divorced from it (Andersen (1973) refers to contact-induced change, or at least one manifestation of it, namely, dialect contact, as 'adaptive' change; Labov (2007) uses the rather apt terms 'transmission,' for internally driven change, and 'diffusion,' for externally motivated change). That is, establishing a genealogical relationship between languages is only possible if the correspondences observed between languages that are due to contact (borrowing) can be sifted out so that the inherited correspondences can be detected. Similarly, the contact-induced correspondences stand out only once a reasonably clear picture is developed of what the inherited material in a given language is. Thus the two types of historical connections between languages, a genetic/genealogical relationship and a contact relationship, are complementary, not competing, ways of modeling language history, and can be seen as two sides of the same coin, what Hamp (1977) has called the 'twin faces of diachronic linguistics,' as both are needed in the elucidation of the sources of similarities and differences between languages observable today. Even with the emphasis here on the Balkans (due in large part to this being the region where the author's expertise lies, though as a hotbed of language contact, the Balkans represent one of the best exemplars from which to draw general lessons about contact between speakers of different languages), it should be clear that the recognition of language contact enriches not only the historical understanding of languages but also the synchronic understanding.

See also: Areal Linguistics; Biological Preconditions for Language Development; Bloomfield, Leonard (1887–1949); Dialectology; Foreign Language Teaching and Learning; Language Acquisition; Language Development, Theories of; Multilingual Language Development; Pidgin and Creole Languages; Second Language Acquisition.

Bibliography

Andersen, Henning, 1973. Abductive and deductive change. Language 49, 765–793.
Bloomfield, Leonard, 1933. Language. Holt, Reinhart and Winston, New York.
Caramazza, Alfonso, Yeni-Komshian, Grace H., Zurif, E.B., Carbone, E., 1973. The acquisition of a new phonological contrast: the case of stop consonants in French-English bilinguals. Journal of the Acoustical Society of America 54, 421–428.
Cook, Vivien, 2003. Introduction: the changing L1 in the L2 user's mind. In: Cook, Vivien (Ed.), Effects of the Second Language on the First. Multilingual Matters, Clevedon, pp. 1–18.
Friedman, Victor A., Joseph, Brian D., 2014. Lessons from Judezmo about the Balkan Sprachbund and contact linguistics. International Journal of the Sociology of Language 226, 3–23.
Friedman, Victor A., Joseph, Brian D., 2015. The Balkan Languages. Cambridge University Press, Cambridge.
Giles, H., 1973. Accent mobility: a model and some data. Anthropological Linguistics 15, 87–105.
Giles, H., Johnson, P., 1987. Ethnolinguistic identity theory: a social psychological approach to language maintenance. International Journal of the Sociology of Language 68, 69–99.
Goebl, Hans, Nelde, Peter H., Stary, Zdenek, Wölck, Wolfgang (Eds.), 1997/2008. Kontaktlinguistik/Contact Linguistics/Linguistique de Contact. Ein internationales Handbuch zeitgenössischer Forschung/An International Handbook of Contemporary Research/Manuel international des recherches contemporaines. de Gruyter, Berlin, 1997: 1. Halbband (vol. 12.1)/2008: 2. Halbband (vol. 12.2).
Hamp, Eric P., 1977. On some questions of areal linguistics. In: Whistler, K., et al. (Eds.), Proceedings of the 3rd Annual Meeting of the Berkeley Linguistics Society. Berkeley Linguistics Society, Berkeley, pp. 279–282.
Haugen, Einar, 1950. The analysis of linguistic borrowing. Language 26, 210–231.
Hickey, Raymond (Ed.), 2010. The Handbook of Language Contact. Wiley-Blackwell, Hoboken, NJ.
Hockett, Charles F., 1958. A Course in Modern Linguistics. The Macmillan Company, New York.
Hussein, Lutfi, 1994. Voicing Dependent Vowel Duration in Standard Arabic and its Acquisition by American Students of Arabic (Ph.D. dissertation). Ohio State University.
Joseph, Brian D., 1983/2009. The Synchrony and Diachrony of the Balkan Infinitive: A Study in Areal, General, and Historical Linguistics. Cambridge University Press, Cambridge (reissued in paperback, 2009).
Joseph, Brian D., 1997. Methodological issues in the history of the Balkan Lexicon: the case of Greek *vré/ré* and its Relatives. In: Friedman, Victor, Belyavski-Frank, Masha, Pisaro, Marc, Testen, David (Eds.), *Balkanistica* 10 (Studies Dedicated to the Memory of Zbigniew Gołąb, March 19, 1923–March 24, 1994), pp. 255–277.
Joseph, Brian D., 2009. Broad vs. Localistic dialectology, standard vs. Dialect: the case of the Balkans and the drawing of linguistic boundaries. In: Tsiplakou, Stavroula, Karyolemou, Marianna, Pavlou, Pavlos (Eds.), Language Variation – European Perspectives II: Selected Papers from the 4th International Conference on Language Variation in Europe (ICLaVE 4), Nicosia, June 2007. John Benjamins, Amsterdam, pp. 119–134.
Labov, William, 2007. Transmission and diffusion. Language 83, 344–387.
Major, Roy C., 1992. Losing English as a first language. Modern Language Journal 76, 190–208.
Matras, Yaron, 2009. Language Contact. Cambridge University Press, Cambridge.
Myers-Scotton, Carol, 2002. Contact Linguistics: Bilingual Encounters and Grammatical Outcomes. Oxford University Press, Oxford.
Sachdev, Itesh, Giles, Howard, 2004. Bilingual accommodation. In: Bhatia, Tej K., Ritchie, William C. (Eds.), Handbook of Bilingualism. Blackwell, Oxford, pp. 353–378.
Schleicher, August, August 1853. Die ersten Spaltungen des indogermanischen Urvolkes. Allgemeine Zeitung für Wissenschaft und Literatur, 786–787.
Schmidt, Annette, 1985. Young People's Dyirbal: An Example of Language Death from Australia. Cambridge University Press, Cambridge.
Schmidt, Johannes, 1872. Die Verwandtschaftsverhältnisse der indogermanischen Sprachen. Weimar, H. Böhlau.
Schneider, Edgar, 2003. The dynamics of new Englishes: from identity construction to dialect birth. Language 79, 233–281.
Siemund, Peter, Kintana, Noemi (Eds.), 2008. Language Contact and Contact Languages. John Benjamins, Amsterdam.
Swadesh, Morris, 1950. Salish internal relationships. International Journal of American Linguistics 16, 157–167.
Thomason, Sarah G. (Ed.), 1997. Contact Languages: A Wider Perspective. John Benjamins, Amsterdam.
Thomason, Sarah G., 2001. Language Contact. Edinburgh University Press, Edinburgh.
Thomason, Sarah G., Kaufman, Terrence, 1988. Language Contact, Creolization, and Genetic Linguistics. University of California Press, Berkeley.
Winford, Donald, 2003. An Introduction to Contact Linguistics. Wiley-Blackwell, Chichester.

Language Development, Theories of

David W Carroll and Eleni Pinnow, University of Wisconsin-Superior, Superior, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Language acquisition by children is robust. Even children with relatively limited exposure to adult speech acquire the major properties of their native language early in childhood. Theoretical approaches to language development have emphasized our biological preparedness for language, the speech presented to children, children's cognitive readiness, and their ability to identify regularities in the speech signal. Contemporary research suggests that several interacting factors are likely to be responsible for children's language acquisition. At the same time, there has been a renewed interest in the biological foundations of language, albeit with an emphasis on the social dimensions of language.

A variety of theoretical approaches to language acquisition have emerged in recent decades. Although the field was initially the province of nativist approaches that claimed an innate basis for language acquisition, in recent years, a diverse array of views have been discussed and debated. Despite the absence of a clear consensus on theoretical matters at the present time, some of the elements of a comprehensive model have come into view. In this article, we consider five theoretical approaches to language acquisition: nativism, child-directed speech, statistical learning, lexical stress, and cognitive development.

Nativist Approaches

The nativist approach to child language emphasizes the role of innate factors in language acquisition (Lenneberg, 1967). The approach most often is associated with the linguist Noam Chomsky, who has asserted that child language acquisition cannot be explained entirely by reference to children's language experience. Chomsky's primary argument for the nativist position has been *the poverty of stimulus argument* (Chomsky, 1980). This argument claims that there is not enough information in the language that is presented to children to enable them to infer the underlying properties of their native language. As a consequence, Chomsky concluded that children needed to have some innate knowledge that, together with the data acquired from the environment, enables the child to acquire the native language.

Several different sources of data have been used to support the nativist explanation. Bickerton (1984) has examined the language of agricultural workers brought into Hawaii to work on sugar plantations. Because the workers were not fluent in the native language, they needed to acquire as much of the language as they could on the fly. What resulted was a *pidgin*, a rudimentary form of language that exists when speakers of mutually unintelligible languages are in close contact with one another. Pidgins are a linguistic free-for-all, with word order often based on the speakers' native language. Interestingly, children of these immigrants developed *creoles*, which were grammatically more sophisticated. The observation that children of immigrants, receiving rudimentary language exposure, were able to produce a more complex version of language than their parents suggested to Bickerton the presence of an innate grammatical system that guides language acquisition.

Similar conclusions emerged from the work of Goldin-Meadow (1982) with congenitally deaf children between 13 months and 4 years of age. Although the children were not exposed to any conventional sign language, they developed their own form of sign language, which was called *homesign*, which shared grammatical properties with signed and spoken languages. Moreover, the milestones that children achieved in homesign (e.g., one-sign utterances about 18 months, followed by two- and three-sign utterances) were similar to those achieved by children exposed to sign. Goldin-Meadow concluded that when language experience was minimal, deaf children might create a gestural system that shares many of the properties with sign languages.

Additional evidence for the nativist view came from studies of sign language acquisition in Nicaragua. The studies took advantage of some political changes in the country in the late 1970s, when the Sandanistas took over control of the government but provided no formal education for deaf children. As with the homesign children, Nicaraguan children began to develop a rudimentary sign system that resembled a pidgin language. Children exposed to the pidgin sign language, however, appeared to acquire a more full-fledged sign language similar to the creoles studied by Bickerton. Whereas the pidgin sign varied considerably from user to user, the creole sign was more consistent. These observations, coupled with those of Bickerton and Goldin-Meadow, suggest that, in situations in which language exposure is limited, individuals are capable of developing or creating a form of communication approaching a full-fledged language.

Although these studies are an important source of data, they pertain to fairly extreme degrees of environmental language neglect. Thus, they leave open the question of to what extent more mundane language experience may shape language acquisition in children. That is, can we identify any relationships between the type of language that parents typically provide to their children and the child's subsequent acquisition of language? The poverty of stimulus argument represents a challenge for more environmentally oriented theorists to demonstrate that the language ordinarily presented to children guides the child's language progress. One line of approach is represented in the next section, when we consider the nature and impact of child-directed speech.

Role of Child-Directed Speech

Parents speak differently to children acquiring language than they do to older children or adults. There are modifications in adult speech to children at every linguistic level. Adults use exaggerated intonation, clearer articulation, shorter sentences, and a greater number of diminutives (e.g., *doggie*) when speaking to children. They also show a greater preponderance of directives and questions than is common in adult speech. Several of the aspects of child-directed speech (sometimes called *motherese*) appear to have the effect of attracting and maintaining the attention of young children. Others may serve to direct or control the behavior of children.

A more challenging question is to assess the effect of child-directed speech on language acquisition. Gathercole and Hoff (2007) identified three central questions pertaining to the impact of language input on language acquisition. First, what information about grammar can the child extract from child-directed speech? At present, there is little evidence that language input promotes language acquisition. For example, although there is some evidence that shorter sentences may promote acquisition, several studies instead suggest that presenting children with longer utterances advances their syntactic development.

Child-directed speech also provides an opportunity for children to receive corrective feedback on their utterances. For example, when a child uses an ungrammatical form, parents may provide feedback in the form of more grammatically correct language. The data on this issue, however, are not clear. Some studies suggest that parents are more likely to correct their children's utterances on semantic rather than syntactic grounds, thus providing minimal support for grammatical development. Other studies found that adults sometimes provide subtle feedback that provides an opportunity for children to recast their utterances in grammatical form. Even here, though, it is unclear whether the feedback is provided often enough to facilitate grammatical acquisition.

Studies that suggest that the input children receive may serve as data for distributional learning may be more promising. For example, children acquiring English will be exposed to numerous utterances that conform to the grammatical requirements of the language. Children gradually may induce these grammatical properties over time. Thus, English-speaking children often will hear utterances in which the word *a* is followed by an object (e.g., *a bird*), whereas the word *some* is followed by a substance (e.g., *some playdough*). Such regularities permit the child to infer some of the basic grammatical properties of the language from the input they receive. How children might do so is discussed more fully in the section dealing with statistical learning.

A second question is whether child-directed speech controls either the sequence or the speed with which children construct the grammar. Regarding sequence, the question is whether the order in which children acquire different aspects of language can be predicted from their frequency of use. For example, we know that English-speaking children acquire the active voice prior to the passive voice and that the active voice is much more common in adult speech. Children learning languages in which the passive voice is more common, however, appear to acquire the construction more rapidly, suggesting that input may influence the developmental sequence. On the other hand, studies of the child's acquisition of grammatical morphemes, however, do not support the assertion that frequency of input is the primary factor in the timing of acquisition. English-speaking children typically acquire these morphemes based on their linguistic complexity more than their frequency of presentation to children. For example, the plural morpheme (*balls*), which encodes the notion of number, is typically acquired earlier than the third-person regular tense (*he works*), which encodes both number and time.

There is more support for the notion that input influences the speed of acquisition. Here studies of bilingual children are most revealing. English-Spanish bilingual children show greater facility of language with English when they had the greatest amount of input in English, and the reverse is true for those who more commonly heard Spanish. These results converge with those just discussed that suggest that frequency of usage – rather than more active processes, such as corrective feedback – may provide important clues for the child's construction of grammar.

Third, is the speech presented to children sufficient to explain the child's acquisition of grammar? Few scholars beyond diehard behaviorists believe that input is sufficient to explain the child's language acquisition. Gathercole and Hoff identified two important factors – cognitive understanding and linguistic complexity – that also play a role. Grammatical features that encode complex concepts typically are acquired later, regardless of input. Moreover, how concepts are encoded in the language matters: When a concept has a more complex expression in a given language, acquisition is slower. For example, a *wh*-question in English such as *Why won't Harold call me back?* is syntactically complex, requiring speakers to begin with a *wh*-word (*who, what, where*, etc.), invert the noun phrase and auxiliary verb, and negate the auxiliary verb. Other languages may encode such questions through the simpler process of elevated intonation at the end of the sentence. All things equal, when the expression is more complex, acquisition is slower.

On balance, recent research suggests an increasing acceptance of the role of input and the declining support for the nativist view. This development, however, is tempered with the recognition that input is not sufficient. We also need to understand how child-directed speech interacts with the linguistic complexity of the forms and the cognitive preparedness of the child. Nonetheless, there is increasing recognition that children are able to acquire key aspects of language from the data they routinely experience.

Statistical Learning

Statistical learning mostly has been used to investigate whether statistical regularities within a language can explain some aspects of language acquisition with experientially developed expectancies. Simple frequency of an item is not enough for statistical learning to occur. Instead, a transitional probability must be used (transitional probability equals frequency of cooccurrence of two stimuli divided by the total frequency of one of the stimuli) to examine whether statistical patterns could serve as a cue to word boundaries. Infants can

successfully exploit statistical cues such as transitional probabilities to determine word boundaries. When first described in the mid-1990s, statistical learning offered a unique and important contribution highlighting the crucial role that experience can play in language acquisition – this includes the novel suggestion that language acquisition may not be purely innate. In contrast to Universal Grammar, proponents of statistical learning argue that learning is central to language acquisition and that learning mechanisms based on statistical regularities explain similarities across languages. Also suggested is that language learning is constrained: Listeners are attuned to certain types of probabilities and dependencies but not others. This portion of the theory examines how listeners are able to find reliable statistical patterns that will aid them in statistical learning and ignore spurious correlations. Although statistical learning seems to be effective for all ages, attentional resources do seem to be crucial for statistical learning.

Statistical learning uses innate learning mechanisms to find statistical regularities in a stimulus, exploit these regularities and process the signal using the regularities. Although other types of stimulus-dependent learning (e.g., perceptual learning) is fairly clear that the underlying changing representations are phonemic (or featural) in nature, statistical learning does not make a precise prediction about the nature of representations. Some theorists have argued that the statistical regularities extracted from the signal are abstracted. These theories, however, claim that it is the statistical regularities that are abstracted and not specifics about the language or grammar that is being learned.

Not surprisingly, research on statistical learning also has examined the degree to which learning can be generalized. This research generally has focused on what is called algebraic learning: learning nonadjacent predictive dependencies. Seven-month-old infants could learn simple 'sentences' (with a pattern of ABA or ABB) made of nonsense syllables. Infants could discriminate between grammatically consistent items and ungrammatical items at test – even when they were tested on novel stimuli. This suggests that the learning of predictive dependencies (statistical or algebraic) can be generalized to novel stimuli.

Statistical learning tends to be flexible in the degree of consistency needed for learning to occur. Twelve-month-old infants presented with an artificial grammar still showed learning effects when only 83% of the stimuli was of the same pattern (17% of the stimuli were taken from a different artificial grammar). The tolerance to conflicting cues, however, does have a limit and learning was impaired when infants heard an artificial grammar that was consistent only 67% of the time (with 33% different artificial grammar). Similarly, Saffran et al. (1996) found that transitional probabilities do not need to be perfectly predictive for statistical learning to occur. In a forced choice test between words (tp = 0.31–1.0), part words (tp = 0.1–0.2), and nonwords (tp = 0.0) listeners showed sensitivity to transitional probability. Participants could better distinguish words from nonwords than words from part words. The higher the transitional probability for the words, the better participants were able to discriminate words from nonwords and part words – however, learning of the words was demonstrated even at lower transitional probabilities.

Statistical regularities in language beyond transitional probabilities and algebraic rules facilitate language acquisition. For example, probabilistic phonotactics is the occurrence or cooccurrence of a phoneme(s) within a word in a given position. Phonotactics is specific to a language and phonotactic probability can range from zero (for illegal phoneme combinations) to much higher (for frequently cooccurring phonemes). Knowledge of phonotactics is available very early in language development. Nine-month old infants listen longer to lists of nonwords that are consistent with the phonotactic rules of their native language (e.g., English-learning infants would listen longer to 'spoll' than to 'mboll' because 'mb' is an illegal word onset in English). Even before speech production, infants are sensitive to the phonotactic rules and combinations that govern their language. Infants, however, are not just sensitive to legal and illegal sound sequences, they also are sensitive to phonotactic probabilities. For example, 9-month-old infants prefer to listen to monosyllables that occur frequently in their native language (have a high phonotactic probability) than monosyllables that occur infrequently (have a low phonotactic probability). The low phonotactic probability monosyllables *did* meet the legality constraints of the infants' native language – this indicates that infants are sensitive to more fine-tuned phonotactic detail than simple gross measures of legality. In general, children under the age of six are able to learn high phonotactic probability nonwords faster than low phonotactic probability nonwords. One potential reason for better learning of the high-phonotactic-probability nonwords is that they sounded 'word-like' due to their high probability, and this facilitated word learning in the children. The low-phonotactic-probability nonwords, however, did not sound like existing words and were ignored as unlikely to be relevant linguistic stimuli. This suggests that phonotactics might be used as a metric in word learning and lexical decision tasks.

Statistical learning is not without its critics though. Johnson and Jusczyk (2001) have argued that although statistical learning may have some predictive power, there is more to speech segmentation than statistics. In the same paradigm as the previously mentioned statistical learning studies, when given a choice, infants relied more heavily on stress cues and coarticulation cues than statistics. These results seem to indicate that listeners use all cues available to them. This finding is not really condemning statistical learning, but rather it points out how robust language learning is and how language processing relies on more than one cue.

Lexical Stress

Stress is a potentially rich source of linguistic information about speech that may be fruitful for language acquisition. In a general linguistic sense, stress refers to the accentuation of a unit within a larger context (e.g., syllables within a word or words within a sentence). The focus of the current section is lexical stress or the syllabic stress patterns within a word. The stressed syllable within a word varies from other syllables on some objective measures: pitch (fundamental frequency), loudness (amplitude of the speech signal), and utterance timing (durational patterning). There are a variety of ways that

lexical stress may influence speech perception. Stress can help a listener correctly discern the meaning of a word. For example, if we consider 'contract' to be a noun, the stress is on the first syllable (*CONtract*); if, however, we use the verb form, the stress falls on the second syllable (*conTRACT*). Overall, these quantitative differences make syllable stress a salient cue within the speech signal. The salience has led to the hypothesis that stress plays a crucial role in the segmentation of words within the speech signal.

Although the segmentation of words within a larger linguistic context seems like an easy task, it is not. Speech's actual acoustic signal contains no silences or inherent cues to word boundaries. This lack of invariance in speech has been a cornerstone of the speech perception literature since its inception. Similarly, the invariance problem poses a significant challenge to models of language acquisition. Given the lack of consistency in the speech signal, how are infants able to acquire the language? One early solution to this problem was the idea that lexical stress may serve as a cue to word boundaries. In English, more than half of polysyllabic words have an initially stressed syllable. (This proportion is even higher in other Germanic languages, notably Dutch, Icelandic, Czech, and German. Uralic languages such as Finnish and Hungarian also have high proportions of primary syllabic stress.) Other languages, such as Polish, typically contain a stressed penultimate syllable. Regardless of language, lexical stress often can provide valuable information into the location of word boundaries.

Lexical stress may be an important cue for language acquisition, but what information are infants extracting from their native language? One specific theory that addresses this issue is the native-language acquisition hypothesis. This hypothesis posits that the initial sensitivity that infants show to rhythmic class forms the basis for learning about the larger rhythmic structure of the language. This rudimentary knowledge about the language structure could allow for infants to use rhythmic cues to segment speech.

The stress pattern of stimuli also may influence the perception of individual phonemes. For example, when presented with a stimulus with initial syllabic stress, native English speakers were able to identify a target phoneme more quickly than if the initial syllable was unstressed. The decreased reaction time occurred regardless of whether the phoneme occurred in a stressed or an unstressed syllable. Other evidence, however, suggests that individual phonemes are identified more easily if they occur within a stressed syllable (perhaps due to the acoustic dimensions of stress discussed earlier). Overall, however, stressed syllables are judged as clearer and more intelligible than nonstressed syllables when excised from their surrounding linguistic context. Taken together, the enhanced perception of stressed syllables offer insight into why stress is a reliable cue for determining word boundaries.

Together with other cues within the speech signal, stress may aid in the acquisition of language. Infants show a remarkable sensitivity to stress patterns within words very early in life. In the first few days of life, infants can discriminate between stressed and unstressed syllables. One important study examined infants' preference for words with a stress pattern that mirrored their native language. Jusczyk et al. (1994) found that English-learning 9-month-old nonspeaking infants listened longer to words that had a strong initial syllable (this pattern, however, was not found in 6-month-old infants). Together, these findings suggest that infants' preference for their native language's stress pattern is determined by their experience with the language (and not an innate preference). This listening pattern and preference may be rudimentary evidence for how infants initially learn to segment words from the larger speech signal.

As infants' develop their linguistic skills, however, lexical stress plays a decreasing role in segmentation and recognition. Later work by Jusczyk et al. suggests that by 11 months of age, infants are no longer relying exclusively on lexical stress to determine word boundaries. As the infants' understanding of the language increases, they begin to rely on multiple cues to determine word boundaries. The salient nature of lexical stress may be an important initial cue to word learning that serves as a bootstrap to learning other cues that may aid in speech perception.

For example, the statistical regularities within words and this regularity may be even more crucial when combined with lexical stress. Using the same paradigm as Saffran's statistical learning, when given a choice, infants relied more heavily on stress cues and coarticulation cues than statistics. Other work demonstrates that when linguistic cues are inconsistent (phonotactic and statistical information conflicts with stress pattern), infants will use stress cues to determine word boundaries and will ignore the conflicting cues. Similarly, the acquisition of an artificial lexicon was influenced by the lexical stress patterns of the stimuli. Stimuli with stress patterns that are consistent with the language learner's native language were more easily learned. Stress pattern might be a crucial cue for word segmentation and may trigger some form of word learning mechanisms. Similar data with adult listeners is consistent with the hypothesis that listeners rely on stress patterns. When listeners are presented with nonsense consonant strings and are asked to rate their 'goodness' (based on their first language), listeners use both statistical cues and stress cues to make their decisions. Thus, these cues seem to have an additive effect in the recognition and processing of spoken language. Interestingly, these cues seem to be relevant to listeners throughout their linguistic development. Additionally, this research suggests that words with primary stress are accessed more rapidly than unstressed syllables. These results seem to indicate that listeners use all cues available to them (statistical probabilities, phonotactic constraints, and stress patterns). Language is robust and, accordingly, language acquisition capitalizes on the multitude of cues.

Cognitive Approaches

Cognitive approaches to language acquisition deemphasize innate, domain-specific linguistic factors in favor of broader cognitive processes. The general assumption underlying these studies is that cognitive development serves as a prerequisite for language development. That is, children need to be able to understand a given concept before they are able to master its linguistic expression. Exploration of the relationships between cognitive and language development have occurred on several fronts.

One set of studies has examined language acquisition in children with cognitive disabilities. For example, research indicates that the language development of children with Down syndrome tends to be proportional to their cognitive deficits (Abbeduto et al., 2007): Children with more severe cognitive deficits typically experience more severe linguistic deficits. In contrast, some researchers have suggested that children with Williams syndrome, a disorder caused by the deletion of genes on chromosome 7, may have language abilities that are advanced relative to their cognitive development. Early work fit this pattern: children with Williams syndrome, although often experiencing significant cognitive delays, appeared to retain a higher level of linguistic skills, particularly syntactic skills. As a consequence, some scholars posited that these children possessed an innate language-specific domain.

More recent studies, however, suggest more proportionate linguistic deficits for these children (Mervis and Becerra, 2007). Current evidence is consistent with the notion that although children with Down and Williams syndromes have somewhat-different linguistic profiles – the former are weaker in syntax but stronger in vocabulary, whereas the latter are stronger in syntactic skills but weaker in visual-spatial abilities – both groups appear to experience delays in language development that are proportionate to their cognitive development.

Another line of inquiry has explored relationships between measures of cognitive development and various linguistic achievements. For example, some studies examined whether the achievement of object permanence, the awareness that objects continue to exist even if no longer present, precedes the acquisition of words such as *all gone* and *more* that may be used to encode that understanding. Research generally has found correlations between cognitive and language development. The evidence that cognitive achievements are necessary prerequisites for linguistic advances, however, is relatively weak: In some studies, cognitive milestones tend to cooccur with, rather than precede, linguistic achievements.

A third and perhaps more promising approach is represented by cross-linguistic studies. Johnston and Slobin (1979) found that the order in which children acquired locative expressions, such as *in*, *on*, *under*, *beside*, and *back*, was similar for children acquiring Italian, Serbo-Croatian, and Turkish, suggesting that the order of acquisition reflected the conceptual complexity of different spatial concepts. In contrast, other studies found that children acquiring Hebrew learn to form questions considerably earlier than children who are acquiring English. Presumably this disparity is related to the structural complexity of questions in English, which as already noted requires a number of linguistic operations. In contrast, in Hebrew, yes–no questions merely require speakers to use a rising intonation with declarative sentences.

Research studies suggest that cognitive development may serve as a general pacesetter for language development. It may be that cognitive achievements serve as a prerequisite for corresponding language achievements. Even so, language acquisition depends not only on cognitive readiness but also on the complexity of the forms that different languages use to encode particular concepts. It also may depend, as we have seen previously, on the frequency of exposure to particular linguistic items. Johnston and Slobin (1979) presented the metaphor of a waiting room as a way of expressing these relationships. Consider a room with an entry door and an exit door. The former represents the child's acquisition of certain cognitive prerequisites; this puts the child in the waiting room. Opening the exit door depends on other factors, such as the frequency of exposure to and the linguistic complexity of the item.

Conclusion

Although the modern study of language development began with nativist approaches, the field has moved away from strictly biological approaches in subsequent decades. Contemporary researchers, although acknowledging a biological component to language, have investigated how children might exploit the data typically presented to them in the course of language acquisition. Coupled with attention to children's cognitive readiness and the linguistic complexity of various aspects of language, studies of statistical learning may provide an important part of a more comprehensive account of language acquisition than is presently available.

At the same time, there has been a resurgence of interest in biological approaches. Two active areas of research pertain to mirror neurons and the FOXP2 gene. FOXP2 (Forkhead box protein P2) is the first gene to be identified as directly relevant to speech and language. Patients with an inherited genetic mutation of the FOXP2 gene show developmental verbal dyspraxia (difficulty producing the highly coordinated motor movements required by speech productions) and limited problems with language comprehension. In songbirds, a mutation of the FOXP2 gene results in reduced vocal learning. FOXP2 demonstrates that the ability to comprehend and produce language may be rooted in our evolutionary history.

Although FOXP2 likely is crucial to understanding the evolutionary development of language, mirror neurons focus are more immediately influential on individual language development. Mirror neurons are active when a person performs a behavior and when that behavior is observed in another individual. For example, mirror neurons in the motor cortex are active both when someone writes her name, as well as when she watches other people write their names. Currently, mirror neurons are known to influence imitation, action understanding, theory of mind, learning, and, of course, language. Although research on mirror neurons and linguistic processing is in its infancy, current hypotheses suggest that mirror neurons first facilitated communicative development in prelinguistic gestural communication. As humans' ability to communicate became more sophisticated, mirror neurons are thought to help grow vocabularies. For example, children's vocabulary increases with their ability to vocally mirror new words – a skill that some researchers link directly to mirror neuron activation. Mirror neurons may be an important beginning step in understanding the neural underpinnings of language development while also reinforcing the importance of social factors of language.

In many ways, biological theories of language acquisition are reminiscent of the early innate theories of linguistic development. Current biological theories, however, also stress the

importance of the linguistic environment. Thus, it appears that the development of language is not nature or nurture, but rather is a complex interaction of all the factors discussed in this article. Further advances in the theoretical explanation of language development continue apace. We can expect fuller examination of the interplay of cognitive, social, and learning abilities along with additional scrutiny of what newer biological approaches may convey.

Bibliography

Abbeduto, L., Warren, S.F., Connors, F.A., 2007. Language development in Down syndrome: from the prelinguistic period to the acquisition of literacy. Mental Retardation and Developmental Disabilities Research Reviews 13, 247–261.

Bickerton, D., 1984. The language bioprogram hypothesis. Behavioral and Brain Sciences 7, 173–221.

Chomsky, N., 1980. Rules and Representations. Columbia University Press, New York.

Gathercole, V.C.M., Hoff, E., 2007. Input and the acquisition of language: three questions. In: Hoff, E., Shatz, M. (Eds.), Blackwell Handbook of Language Development. Blackwell, Malden, MA, pp. 107–127.

Goldin-Meadow, S., 1982. The resilience of recursion: a study of a communication system developed without a conventional language model. In: Wanner, E., Gleitman, L.R. (Eds.), Language Acquisition: The State of the Art. Cambridge University Press, Cambridge, UK, pp. 51–77.

Johnson, E.K., Jusczyk, P.W., 2001. Word segmentation by 8-month-olds: when speech cues count for more than statistics. Journal of Memory and Language 44, 548–567.

Johnston, J.R., Slobin, D.I., 1979. The development of locative expressions in English, Italian, Serbo-Croatian, and Turkish. Journal of Child Language 6, 529–545.

Jusczyk, P.W., Cutler, A., Redanz, N., 1994. Preference for the predominant stress patterns of English words. Child Development 64, 675–687.

Lenneberg, E.H., 1967. Biological Foundations of Language. Wiley, New York.

Mervis, C.B., Becerra, A.M., 2007. Language and communication development in Williams syndrome. Mental Retardation and Developmental Disabilities Research Reviews 13, 3–15.

Saffran, J.R., Newport, E.L., Aslin, R.N., 1996. Word segmentation: the role of distributional cues. Journal of Memory and Language 35, 606–621.

Language Development: Emergentist Theories

William O'Grady, University of Hawai'i, Honolulu, HI, USA

© 2015 Elsevier Ltd. All rights reserved.

Special symbols

¬ Negation.
∀ Universal quantification.

Abstract

The course of linguistic development is shaped by processing pressures that reflect the interaction of internal factors such as burden on working memory with external factors relating to experience. Two case studies help reveal the central role of internal pressures in determining both the properties of language and the manner in which those properties emerge over time, opening the door to a new perspective on development.

Introduction

Much of the research program of contemporary linguistics focuses on two fundamental questions.

- Why do languages have the particular properties that they do?
- How do those properties emerge in the course of language acquisition?

Emergentism holds that the answer to both questions lies in the interaction of forces that exist independently of language. One such force, on which I will concentrate here, is processing cost.

The central thesis of this article is that processing pressures are largely responsible for the course of development – the series of changes that bring a learner's language in alignment with the language of his or her community. This idea can be stated informally as follows.

1. Processing determinism
   Processing pressures shape the course of development.

This thesis rules out the possibility that development is the product of maturational factors and/or environmental pressures unrelated to processing, but is neutral with respect to the nature of the mechanisms that emerge in the course of language acquisition. Although I take those mechanisms to be processing routines rather than conventional grammatical rules (e.g., O'Grady, 2005), nothing turns on that point here, and I will explore the consequences of processing determinism in theory-neutral terms to the extent possible.

Processing-related pressures come from two very different sources. On the one hand, there are external factors that arise from the environment, including the composition and quantity of the input to which learners are exposed. On the other hand, there are internal pressures that stem from the burden that particular linguistic phenomena place on working memory. I will consider each type of pressure in turn, before examining the question of whether development follows a uniform course despite apparent differences among individual learners.

External Pressures: The Role of Input

Most emergentist work on development maintains a strong input orientation. An early and very influential example of this comes from the classic Competition Model. Explicitly presented as a processing-based approach to development (MacWhinney, 1987: 301), the Competition Model focuses on the relative weighting of the input cues that help identify sound-meaning mappings. It makes a simple but daring prediction: by the virtue of its facilitatory effect on processing, cue validity should determine the order in which particular aspects of language are acquired (Bates and MacWhinney, 1987: 171).

In the classic model, cue validity reflects the interaction of two factors – availability (how often a cue is present in the input when a particular mapping is being computed) and reliability (how often the cue leads to the correct result). In English, for instance, the cue of noun–verb order, exemplified in (2), scores high in validity when it comes to identifying a sentence's subject.

2. Rob likes Lucy

Not only is noun–verb order almost always available in English (the vast majority of sentences have a noun to the left of the verb), it is also highly reliable (a preverbal noun almost always corresponds to the subject). In contrast, cues such as case (*he* vs *him*, *she* vs *her*) or verbal agreement are less frequently available, and are hence less valid. These facts are reflected in development in the predicted way: children learning English initially rely primarily on word order to distinguish subjects from direct objects. In other languages, of course, other cues have a higher validity (e.g., animacy in Italian, and case in Turkish), shaping development in those languages accordingly (Slobin and Bever, 1982; Bates et al., 1984).

The Competition Model is just one of several streams of emergentist thought that focus on the role of input in shaping development. Usage-based and constructivist approaches to language acquisition (e.g., Tomasello, 2003; Goldberg, 2006;

MacWhinney, 2010; Ambridge and Lieven, in press) place special emphasis on children's pattern-finding skills. Research on statistical learning (Saffran et al., 1996, 2008; Thompson and Newport, 2007) focuses on learners' ability to track transitional probabilities (the likelihood that the occurrence of X will result in the appearance of Y), a phenomenon that is also the object of connectionist modeling (Elman, 2002; Chater and Manning, 2006; Chang et al., 2006). Dynamic Systems Theory (van Geert and Verspoor, in press) investigates linguistic development by considering how proficiency changes over time in response to the interaction of experience, maturation, and previous learning.

It makes sense to expect input to have a major role in shaping development. All other things being equal, frequent sound-meaning mappings should emerge and become entrenched before their less frequent counterparts. And this is in fact more or less what happens (Menn, 2000; Jurafsky, 2003; Ferreira, 2003); for a general review, see Jaeger and Tily (2011).

However, processing pressures that come from the input tell only part of the story. There is also a network of internal efficiency-related processing considerations that help to shape development in significant ways, alone and in conjunction with input. Unfortunately, the effect of these factors is often difficult to discern because of the presence of parallel external pressures. The development of relative clauses is a case in point.

A frequent finding in the acquisition literature is that children learning English find subject relative clauses easier to produce and understand than direct object relative clauses (Diessel and Tomasello, 2005; Friedmann et al., 2009; O'Grady, 2011).

3a. Subject relative clause:
the boy [that _ hugged the girl]

3b. Direct object relative clause:
the boy [that the girl hugged _]

There are a number of promising processing accounts for this asymmetry. For example, Grodner and Gibson (2005) propose that processing cost is determined, at least in part, by the amount and type of material intervening between the gap inside the relative clause and the associated 'filler' (*the boy* in our examples). The sole intervening element in the subject relative clause in (3a) is the complementizer *that*. In the direct object relative clause in (3b), in contrast, two major constituents (the NP *the girl* and the verb *hug*) intervene between the filler and the gap – increasing the burden on working memory and contributing to later mastery.

Unfortunately, however, the force of this explanation is somewhat compromised by a possible frequency-related input effect. At least some corpus research reports that subject relative clauses are in general more frequent than direct object relative clauses, and the asymmetry is overwhelming when one takes into account animacy: relative clauses that modify an animate noun, such as those in (3), are far more likely to be of the subject type than the object type (e.g., Kidd et al., 2007). Thus input factors too could be responsible for the early mastery of subject relative clauses.

Confounds between internal processing cost and frequency of occurrence in the input are to be expected: all other things being equal, language use should favor patterns that incur less processing cost. How then can we ensure that processing determinism in development is not simply input determinism?

The Case for Internal Factors

The key to establishing an independent role for internal processing factors lies in showing they have effects distinct from those predicted by external factors. Various phenomena offer the opportunity to pursue this line of inquiry, and I will consider two such cases here, both involving comprehension. In this regard, I depart from the practice, common in much of the literature on development, of focusing just on production. This departure is justified for two reasons.

First, advances in comprehension drive language acquisition. Without the input that comes from understanding the linguistic patterns that are exemplified in the speech of others, there would be no development. Moreover, development is possible in the absence of production, as the investigation of mutism has documented (Lenneberg, 1962; Hickok, 2009).

Second, as I will try to demonstrate in what follows, key insights into the nature of development come from the study of patterns for which there are few opportunities for use, especially in the speech of children. An exclusive focus on production in such cases would exclude them from consideration, an obviously undesirable result.

Scope Preferences in First Language Learners of English and Korean

A signature feature of natural language is its capacity for the expression of negation and quantification. The sentence in (4) offers a simple example.

4. Kim didn't buy all the pencils.

The dominant interpretation of this sentence is that Kim bought only some of the pencils (the so-called '*not* > *all* reading'), not that no pencils were purchased by Kim (the '*all* > *not* reading'). Interestingly, the reverse is true in Korean (Subj = subject marker; DO = direct object marker; Pst = past tense; Decl = declarative).

5. Kim-i motun yenphil-ul an sa-ss-ta.
Kim-Subj all pencil-DO not buy-Pst-Decl
'Kim bought none of the pencils.' (All the pencils went unpurchased.)

Very few sentences containing a negated verb and a universally quantified direct object occur in parental speech in either Korean or English (O'Grady, 2013). Yet scopal interpretations appear to develop in a systematic way, following a particular (but different) course in each language. Whereas Korean-speaking children manifest a strong preference for the *all* > *not* interpretation from the outset (Han et al., 2007), English-speaking children initially permit both the *all* > *not* and the *not* > *all* interpretations (Musolino et al., 2000; Musolino and Lidz, 2006).

As I have argued elsewhere (e.g., O'Grady, 2013), processing pressures offer an explanation for both why Korean and English differ in their scopal preferences and why the emergence of those preferences follows a different path in each language. Let us first consider the case of Korean.

The two key steps in processing the Korean scopal pattern can be schematically depicted as follows. (The symbol ∀ stands for universal quantification, and the symbol ¬ for negation.)

6a. Step 1: The processor encounters the quantified NP.
Kim all the pencils ….
∀x (pencils)

6b. Step 2: The processor encounters the negative.
Kim all the pencils not buy
∀x (pencils) ¬ → *all* > *not*

As illustrated here, the quantified NP is encountered and interpreted early in the sentence (step 1). This leaves just one option at the point at which the processor comes upon the negative: the *all* > *not* interpretation, as shown in (6b). Derivation of the alternative *not* > *all* option can be achieved only through revision of that initial interpretation, as depicted below.

7. Kim all the pencils not buy
∀x (pencil) ¬→ *all* > *not* (first interpr.) → *not* > *all* (revised interpr.)

It seems reasonable to suppose that revisions to a previously derived interpretation are costly (O'Grady 2013 and the references cited there) – hence children's avoidance of this option in Korean.

Matters are very different in English. Because the negative precedes the quantifier, the processor comes upon it first and registers its presence in the sentence.

8. Step 1: The processor encounters the negative.
Kim didn't buy
¬

Upon subsequently encountering the quantified NP, the processor has two options: allow the negative to scope over the quantifier (the *not* > *all* interpretation) or avoid any such interaction (the *all* > *not* interpretation).

9. Step 2: The processor encounters the quantified NP and computes scope.
Kim didn't buy all the pencils
¬ ∀x (pencils) → *not* > *all*
→ *all* > *not*

Because there are no grounds for believing that either option is inherently more costly, neither should be favored in the absence of relevant input. As summarized in **Table 1**, this aligns exactly with the initial state of development for English, which is characterized by acceptance of both interpretations.

In sum, processing pressures offer a straightforward explanation both for the difference in scopal preferences exhibited by Korean and English and for the developmental profile seen in each language. The reason that Korean strongly favors the *all* > *not* interpretation and the reason that children initially seize on that option is that it has a lower processing cost than its *not* > *all* counterpart. And the reason that English is not constrained in this way and that children initially allow both the *not* > *all* and the *all* > *not* interpretations is that the two options are comparable in processing cost.

Table 1 The developmental sequences for scope in English and Korean

English	*Korean*
1. Both *not* > *all* & *all* > *not* accepted 2. Strong preference for *not* > *all*	1. Strong preference for *all* > *not* 2. No change

Table 2 Developmental sequences for reflexive and plain pronouns

Reflexive pronouns	*Plain pronouns*
1. Correct interpretation	1. Possible misinterpretation as a reflexive 2. Correct interpretation

Of course, this does not mean that there is no role for input. Somehow, English speakers later come to favor the *not* > *all* interpretation, presumably on the basis of experience, however sparse, that is accumulated over a period of many years. (Recall that the preference is not manifested in the judgments of school-age children.) This fits well with the idea that the course of development is shaped by a combination of internal and external processing factors.

Reflexive Pronouns

Children learning English manifest a curious asymmetry in their interpretation of pronouns, with higher early success on reflexive pronouns (*himself*, *herself*, etc.) than on their plain counterparts (*him*, *her*, etc.). Of particular interest is the type of error most commonly made: plain pronouns are misinterpreted as reflexive pronouns, but not vice versa (Conroy et al., 2009; O'Grady, in press; **Table 2**).

Frequency of occurrence offers no explanation for this asymmetry, as plain pronouns occur far more commonly than their reflexive counterparts in caregiver input. In a search of the Child Language Data Exchange System (CHILDES) database, I uncovered 1836 instances of *him*, *her*, and *them* in maternal speech to Adam, Eve, and Sarah, compared with just 17 instances of *himself*, *herself*, and *themselves*.

In contrast, processing pressures offer a compelling explanation for the developmental facts, as the two types of pronoun differ in terms of the resources needed to locate the appropriate antecedent. As is standardly observed, reflexive pronouns typically require a 'local' antecedent – almost always an argument in the same clause, as illustrated in (10).

10. [**Mickey** tickled **himself**].
↑_____|

In contrast, the search for an antecedent in the case of plain pronouns focuses on referents that are prominent in the discourse, without regard for locality (Song and Fisher, 2007; Foraker and McElree, 2007; among many others).

11. Mickey tickled **him**
↖_____|

As frequently noted (e.g., Reinhart, 2006: 181ff; Reuland, 2011: 127), a search through discourse often involves greater processing cost than is required for the interpretation of a reflexive pronoun, whose antecedent is predictably located in the same clause.

Internal processing pressures thus provide a straightforward explanation for children's difficulty with plain pronouns, even after exposure to ample input. Of course, this does not mean that there is no role for experience – after all, the right forms have to be learned (*himself* rather *hisself*, *myself* rather than *meself*). In the important matter of interpretation, however, internal processing pressures do the heavy lifting.

Summary

The two case studies that we have been considering are instructive for two reasons. First, from the point of view of typology, they illustrate how processing pressure can shed light on why languages have the particular properties that they do. Korean favors the *all* > *not* interpretation of scopal patterns because that reading is easier to process than the alternative. English, like countless other languages, requires local antecedents for reflexive pronouns because that interpretation minimizes the processing cost associated with resolving referential dependencies and so on.

Second, the same processing forces that favor patterns of a particular type shape the emergence of language in children, establishing the course of early development. In some cases (perhaps relative clauses), internal processing pressures act in conjunction with the input, but there are also telling cases in which they act on their own, neutralizing the potential influence of input (as in the case of pronoun interpretation) and even making up for the absence of relevant experience (as in the case of scopal patterns).

How Uniform Is Development?

Although it is sometimes suggested that development is highly uniform (e.g., Brown, 1973; Pienemann, 2005), much recent work points in the opposite direction. Dynamic Systems Theory, in particular, has emphasized both intralearner variability and interlearner variation, raising the possibility that no two learners proceed at the same rate, make the same errors, or attain proficiency in the same manner (de Bot et al., 2007; van Dijk et al., 2011; Verspoor and van Dijk, 2011). What is the reality?

There is some reason to think that developmental diversity, however widespread, is circumscribed by the influence of processing pressures.

12. The Uniformity Thesis
    Development is uniform, where processing cost is relevant.

Three points call for immediate clarification.

First, the Uniformity Thesis has nothing to say about differences in *rate* of development – a well established and well documented parameter of ontogenetic variation. The cost of maintaining and accessing a vocabulary of, say, 150 words is the same, regardless of whether those words are acquired by age 18 months or by age 24 months. The cost of computing the agreement relationship between a verb and subject (*A boy is/*are outside*) is the same when acquired at 30 months as it is when learned at 36 months. These and other differences unrelated to processing cost fall outside the scope of processing determinism (and, by extension, the Uniformity Thesis).

Second, a large portion of the variation across learners that is reported in the literature arguably reflects differences in language use that are present independent of development. This is especially so in the case of spontaneous speech – a very common source of data in studies of developmental order (Brown, 1973; Pienemann, 2005; Bassano and van Geert, 2007). Even mature monolingual speakers of a language differ in their ability to express themselves in extemporaneous speech, and there can be fluctuations in performance literally from moment to moment. Variation and instability are no doubt magnified in learners, whose processing routines are less entrenched than those of mature speakers and hence even more susceptible to the factors that make speech difficult in the first place. Conclusions based on production data therefore need to be interpreted cautiously, and tempered by the insights afforded by controlled experiments, including those involving comprehension.

Third, the Uniformity Thesis does not predict that learners will follow identical paths of development regardless of input; rather, it predicts uniformity in the effect of input. An obvious case in point involves lexical development, whose precise course varies considerably from child to child, but nonetheless appears to reflect experience quite faithfully (Huttenlocher et al., 1991; Hoff and Naigles, 2002). That is why, to take a trivial example, children growing up in America learn the words *truck*, *elevator*, *cookie*, and *apartment* before *lorry*, *lift*, *biscuit*, and *flat*, whereas the reverse is true for children growing up in the United Kingdom.

The study of morphosyntax offers comparable examples. It would not be surprising to find earlier emergence of *shall* in the United Kingdom than in the United States, or a preference for *I don't have any milk* over *I haven't any milk* in American children but not in their British counterparts. In these and other cases, development is presumably as uniform as the input that drives it.

If all of this is right, then the case for Uniformity depends crucially on the effect of internal processing factors, especially those that operate without the confounding influence of input. Do such factors shape development in a uniform way across children and even across languages? If the phenomena considered in this article are typical, the answer may well be yes.

- Children learning Korean uniformly show an early preference for *all* > *not* scope. In contrast, English-speaking children initially permit either interpretation.
- English-speaking children sometimes interpret plain pronouns as if they were reflexives (e.g., Conroy et al., 2009), but errors in the reverse direction are uniformly avoided.

Findings of this sort point toward a fundamental similarity in those aspects of development that are shaped by internal processing pressures, consistent with the Uniformity Thesis.

Conclusion

Development is a signature phenomenon of language, and its study offers the promise of insights into the mechanisms that make language possible in the first place. The emergentist approach to this challenge takes as its starting point the idea that processing cost shapes both the properties of language and the manner in which those properties are acquired. A key component of the particular proposal that I have put forward is a distinction between two quite different determinants of processing cost, one rooted in internal pressures relating to burden on working memory and the other reflecting the composition and quantity of learners' experience. The two case studies considered here point to the crucial role of internal factors in shaping the emergence of language, at times in the absence of relevant input. Moreover, these studies raise the possibility that development may unfold in a more uniform manner than is usually assumed. For now, though, these conclusions remain tentative, awaiting the further work that is needed to fully assess their viability.

See also: Anaphora; Child-Directed Speech: Influence on Language Development; Emergent Properties; First Language Acquisition, Linguistic Theory of; Language Acquisition; Language Development, Theories of; Negation, Linguistics of; Vocabulary Acquisition.

Bibliography

Ambridge, B., Lieven, L. A (neo-)constructivist account of child language acquisition. In: MacWhinney, B., O'Grady, W. (Eds.), Handbook of Language Emergence. Wiley, Boston, in press.

Bassano, D., van Geert, P., 2007. Modeling continuity and discontinuity in utterance length. Developmental Science 10, 588–612.

Bates, E., MacWhinney, B., Caselli, C., Devescovi, A., Natale, F., Venza, V., 1984. A crosslinguistic study of the development of sentence interpretation strategies. Child Development 55, 341–354.

Bates, E., MacWhinney, B., 1987. Competition, variation, and language learning. In: MacWhinney, B. (Ed.), Mechanisms of Language Acquisition. Erlbaum, Mahwah, NJ, pp. 157–193.

Brown, R., 1973. A First Language: The Early Stages. Harvard University Press, Cambridge, MA.

de Bot, K., Lowie, W., Verspoor, M., 2007. A dynamic systems theory approach to second language acquisition. Bilingualism. Language and Cognition 10, 7–21.

Chang, F., Dell, G., Bock, K., 2006. Becoming syntactic. Psychological Review 113, 234–272.

Chater, N., Manning, C., 2006. Probabilistic models of language processing and acquisition. Trends in Cognitive Science 10, 335–344.

Conroy, A., Takahashi, E., Lidz, J., Phillips, C., 2009. Equal treatment for all antecedents: how children succeed with principle B. Linguistic Inquiry 40, 446–486.

Diessel, H., Tomasello, M., 2005. A new look at the acquisition of relative clauses. Language 81, 882–906.

van Dijk, M., Verspoor, M., Lowie, W., 2011. Variability and DST. In: Verspoor, M., deBot, K., Lowie, W. (Eds.), A Dynamic Approach to Second Language Development: Methods and Techniques. John Benjamins, Amsterdam, pp. 55–84.

Elman, J., 2002. Generalization from sparse input. In: Andronis, M., Debenport, E., Pycha, E., Yoshimura, K. (Eds.), Proceedings of the 38th Regional Meeting of the Chicago Linguistic Society: The Panels. Chicago Linguistic Society, Chicago, pp. 175–200.

Ferreira, F., 2003. The misinterpretation of noncanonical sentences. Cognitive Psychology 47, 164–203.

Foraker, S., McElree, B., 2007. The role of prominence in pronoun resolution: active versus passive representations. Journal of Memory and Language 56, 357–383.

Friedmann, N., Belletti, A., Rizzi, L., 2009. Relativized relatives: types of intervention in the acquisition of A-bar dependencies. Lingua 119, 67–88.

van Geert, P., Verspoor, M. Dynamic systems and language development. In: MacWhinney, B., O'Grady, W. (Eds.), Handbook of Language Emergence. Wiley, Boston, in press.

Goldberg, A., 2006. Constructions at Work. Oxford University Press, Oxford, UK.

Grodner, D., Gibson, E., 2005. Consequences of the serial nature of linguistic input for sentential complexity. Cognitive Science 29, 261–290.

Han, C., Lidz, J., Musolino, J., 2007. V-raising and grammar competition in Korean: evidence from negation and quantifier scope. Linguistic Inquiry 38, 1–48.

Hawkins, J., 2004. Efficiency and Complexity in Grammars. Oxford University Press, Oxford, UK.

Hickok, G., 2009. Understanding language without ability to speak. Talking Brains: News and Views on the Neural Organization of Language. http://www.talkingbrains.org/2009/03/understanding-language-without-ability.html (accessed 10.05.14.).

Hoff, E., Naigles, L., 2002. Children use input to acquire a lexicon. Child Development 73, 418–433.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., Lyons, T., 1991. Vocabulary growth: relation to language input and gender. Developmental Psychology 27, 236–248.

Jaeger, T.F., Tily, J., 2011. On language 'utility': processing complexity and communicative efficiency. Wiley Interdisciplinary Reviews: Cognitive Science 2, 323–335.

Jurafsky, D., 2003. Probabilistic modeling in psycholinguistics: linguistic comprehension and production. In: Bod, R., Hay, J., Jannedy, S. (Eds.), Probabilistic Linguistics. MIT Press, Cambridge, MA, pp. 39–95.

Kidd, E., Brandt, S., Lieven, E., Tomasello, M., 2007. Object relatives made easy: a cross-linguistic comparison of the constraints influencing young children's processing of relative clauses. Language and Cognitive Processes 22, 860–897.

Lenneberg, E., 1962. Understanding language without ability to speak: a case report. Journal of Abnormal and Clinical Psychology 65, 419–425.

Lewis, R., Vasishth, S., Van Dyke, J., 2006. Computational principles of working memory. Trends in Cognitive Sciences 10, 447–454.

MacWhinney, B., 1987. The competition model. In: MacWhinney, B. (Ed.), Mechanisms of Language Acquisition. Erlbaum, Mahwah, NJ, pp. 249–308.

MacWhinney, B., 2010. Language development. In: Bornstein, M., Lamb, M. (Eds.), Developmental Science: An Advanced Textbook, sixth ed. Psychology Press, New York, pp. 389–423.

Menn, L., 2000. Let's face a simple question: why is canonical form simple? Brain and Language 71, 157–159.

Musolino, J., Lidz, J., 2006. Why children aren't universally successful with quantification. Linguistics 44, 817–852.

Musolino, J., Crain, S., Thornton, R., 2000. Navigating negative quantificational space. Linguistics 38, 1–32.

O'Grady, W., 2005. Syntactic Carpentry: An Emergentist Approach to Syntax. Erlbaum, Mahwah, NJ.

O'Grady, W., 2011. Relative clauses: processing and acquisition. In: Kidd, E. (Ed.), The Acquisition of Relative Clauses: Processing, Typology and Function. John Benjamins, Amsterdam, pp. 13–38.

O'Grady, W., 2013. The illusion of language acquisition. Linguistic Approaches to Bilingualism 3, 253–285.

O'Grady, W. Anaphora and the case for emergentism. In: MacWhinney, B., O'Grady, W. (Eds.), Handbook of Language Emergence. Wiley, Boston, in press.

O'Grady, W. Processing Determinism. Language Learning, to appear.

Pienemann, M., 2005. An introduction to processability theory. In: Pienemann, M. (Ed.), Cross-linguistic Aspects of Processability Theory. John Benjamins, Amsterdam, pp. 1–60.

Reinhart, T., 2006. Interface strategies: reference-set computation. MIT Press, Cambridge, MA.

Reuland, E., 2011. Anaphora and Language Design. MIT Press, Cambridge, MA.

Saffran, J., Aslin, R., Newport, E., 1996. Statistical learning by 8-month old infants. Science 274, 1926–1928.

Saffran, J., Hauser, M., Seibel, R., Kapfhamer, J., Tsao, F., Cushman, F., 2008. Grammatical pattern learning by human infants and cotton-top tamarin monkeys. Cognition 107, 479–500.

Slobin, D., Bever, T., 1982. Children use canonical sentence schemas: a crosslinguistic study of word order and inflections. Cognition 12, 229–265.

Soderstrom, M., Conwell, E., Feldman, N., Morgan, J., 2009. The learners as statistician: three principles of computational success in language acquisition. Developmental Science 12, 409–411.
Song, H., Fisher, C., 2007. Discourse prominence effects on 2.5-year-old children's interpretation of pronouns. Lingua 117, 1959–1987.
Thompson, S., Newport, E., 2007. Statistical learning of syntax: the role of transitional probability. Language Learning and Development 3, 1–42.
Tomasello, M., 2003. Constructing a Language: A Usage-Based Theory of Language. Harvard University Press, Cambridge, MA.
Verspoor, M., van Dijk, M., 2011. Visualizing interactions between variables. In: Verspoor, M., de Bot, K., Lowie, W. (Eds.), A Dynamic Approach to Second Language Development: Methods and Techniques. John Benjamins, Amsterdam, pp. 85–98.

Relevant Websites

psyling.psy.cmu.edu/ – Brian MacWhinney.
http://www.rug.nl/staff/c.l.j.de.bot/ – Kees de Bot.
http://www.rug.nl/staff/m.h.verspoor/ – Marjolijn Verspoor.
http://www-personal.umich.edu/~ncellis/NickEllis/Home.html – Nick Ellis.
http://www.paulvangeert.nl/ – Paul van Geert.
http://www.angl.hu-berlin.de/confslecs/AILA-ReN-Langscape-Symposium_2013/call-for-papers – Sociocultural Theory and Emergentism.
http://www.rug.nl/staff/w.m.lowie/ – Wander Lowie.
http://www.ling.hawaii.edu/william-ogrady – William O'Grady.

Language Development: Evolutionary Theories

Mark Atkinson and Kenny Smith, Language Evolution and Computation Research Unit, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Accounts of language development can be broadly divided into two opposing camps: nativist theories and statistical learning accounts. We review the claims made by each about the role of the socially acquired linguistic data in language acquisition and consider the implications of each account on the likely evolutionary history of the language-learning capacity.

Our species is characterized by a uniquely powerful communication system: anything we can think, we can convey to others using language. No other species has a communication system that comes close to offering this level of expressive power; instead every other system is limited to the communication of a small number of possible messages on a tightly circumscribed set of topics (e.g., sources of dangers, sources of food, or availability as a mate). Human language is also relatively rare in that it is learned (see Fitch, 2000, for a discussion of the scarcity of learned communication in the natural world): we acquire the language of our speech community based on exposure to the linguistic behavior of that community. Explaining the evolution of language therefore amounts to explaining the evolution of the uniquely human capacity to acquire a language from linguistic data. So how did the capacity for this unique communication system evolve in our species and why is it not more widespread in the natural world?

Accounts of language development can be broadly divided into two opposing camps: nativist theories, emphasizing the role of innate, domain-specific knowledge relating only to language, and statistical learning accounts, emphasizing the richness of linguistic data and the role of domain-general learning procedures. Both views of language development accept that socially provided language input is necessary, but the function of this input in each account is very different: in nativist accounts its role is merely to activate the linguistic knowledge already present within an individual learner; in statistical learning accounts it is the sole source of linguistic knowledge. These two views of language learning also make very different claims about the likely evolutionary history of the language-learning capacity, which we review here.

Nativism and the Evolution of Language Learning

Nativist accounts of language acquisition date back to the late 1950s and are mostly associated with the theories of Noam Chomsky (e.g., Chomsky, 1965). Nativist theories seek to explain how children acquire the formal rules describing a language's grammar from linguistic input which is, they argue, deficient in a number of crucial ways (*see* Language Acquisition). In particular, the data provided to children as input forms but a small sample of the infinitely many possible sentences which could be produced by a language that is capable of being infinitely expressive. Furthermore, certain grammatical structures are argued to be unlearnable without 'negative evidence,' explicit, impermissible counterexamples to grammatically well-formed sentences. The solution to this problem of acquiring a language from degenerate data is the proposed existence of 'Universal Grammar,' UG. UG is an innate, genetically encoded module that provides the complex, abstract and formal knowledge necessary to acquire the grammatical system of any language. Acquiring a language then requires very little learning, if indeed 'learning' is thought to happen at all (Piattelli-Palmarini, 1989). In addition to providing elements such as a language's lexicon (considered a relatively superficial aspect of the language-learning problem), the linguistic input simply serves to guide the child to a particular grammar from within the set of possibilities defined by UG (*see* Language Development, Theories of). Under one influential account (Chomsky, 1981), this is achieved through the setting of a small number of 'parameters,' each of which specify one or more structural properties of a language and can be set using relatively simplistic linguistic data, the trigger for setting that parameter. This process has been likened to the setting of a number of 'switches' within the 'switch box' provided by UG. For instance, one parameter might determine whether a language has a fixed or free word order, with its setting dependent on whether the input received is from a language with more rigid ordering of words in its sentences, such as English, or a syntactically more flexible language, such as Hungarian. UG therefore defines a space of possible human language types, one for each possible combination of parameters: in principle, it might be possible to determine the parameters of UG by inspecting the range of languages we see in the world and how they vary (Baker, 2001), since this variation should by definition be constrained by UG. 'Linguistic universals,' ubiquitous features thought to exist in all natural languages, are therefore a result of genetically encoded constraints on possible languages within the individual.

Where did UG, this rich, complex, species-wide biological capacity, come from? The definitive argument for UG as a product of evolution by natural selection is provided by Pinker and Bloom (1990), who argue that such a trait can only be explained as the result of an adaptionist process, as:

> The only successful account of the origin of [any] complex biological structure is the theory of natural selection.
>
> Pinker and Bloom, 1990: p. 708

In their theory, the uniquely social ecology of humans puts a premium on the ability to communicate knowledge and internal states to others. Language provides the apparatus to do this (*see* Biological Preconditions for Language Development), and individuals whose genes equip them with the ability to learn a complex system for expressing such knowledge using language, and understanding others when they do so, will be at a selective advantage: thanks to the many benefits of sharing hard-won knowledge with friends and family, such individuals will be more likely to survive and reproduce, leading to the gradual evolution of the modern language faculty. Pinker and Bloom identify a number of features of language which, they argue, exhibit the appearance of design for this function. For example, they argue that major lexical categories (such as nouns and verbs) are used to distinguish basic ontological categories, phrase structure and overt case marking provide surface cues to the structure of the underlying propositions which language is used to convey (i.e., who did what to who), and verb affixes signal temporal aspects of events (whether it happened in the past, will happen in the future, or is happening now). Pinker and Bloom go on to identify scenarios where language exhibiting these design features might offer a selective advantage. They propose that the hunter-gatherer lifestyle hypothesized to characterize our recent evolutionary predecessors would favor an ability to communicate information about toolmaking, ecology, and the behavior of plants and animals:

> Devices for communicating precise information about time, space, predicate–argument relations, restrictive modification, and modality are not wasted in such efforts.
>
> Pinker and Bloom, 1990: p. 724

Secondly, they emphasize the value of language in supporting social interactions, which put

> a premium on the ability to convey such socially relevant abstract information as time, possession, beliefs, desires, tendencies, obligations, truth, probability, hypotheticals, and counterfactuals… Furthermore, in a group of communicators competing for attention there is a premium on the ability to engage, interest and persuade listeners. This in turn encourages the development of discourse and rhetorical skills and the pragmatically relevant grammatical devices that support them.
>
> Pinker and Bloom, 1990: p. 725

Therefore, according to Pinker and Bloom, the human capacity to acquire complex language evolved as a direct consequence of the payoff it provided in relation to these communicative functions. As clarified in Pinker (2010), this adaptive account of the evolution of the human capacity to learn language fits within the broader explanatory framework of evolutionary psychology (and indeed has been hailed as the crowning success of that school of thought by Tooby et al., 2005): the capacity for language is only one of a suite of cognitive adaptations arising from the human cognitive niche, characterized by our social interdependence and reliance on knowledge acquired from others (*see* Human Language, Evolution of).

Why, under this account, does UG also constrain the kinds of languages that can be learned? According to Pinker and Bloom, the evolutionary drive for expressive power was tempered by the requirement for speaker–hearer parity. Any arbitrary communication system would have been advantageous, but it would have been necessary for it to be conventionalized; that is, shared among the members of a group. Individual elements of such a system would originally have to be learned, but once conventionalized, might gradually become innate (i.e., hardwired into individuals) due to the 'Baldwin effect': the increased genetic coding of advantageous acquired traits over a large number of generations (Baldwin, 1896). Nativization of conventional aspects of language structure removes the risk of failing to learn and paves the way for innate constraints on language learning as found in UG.

This adaptationist account of the evolution of UG has received support from models of the evolution of communication systems, most notably from Martin Nowak and colleagues (Nowak et al., 1999, 2000, 2001), who present a series of models exploring the conditions under which grammatical language and UG might evolve under natural selection. Nowak et al. (2000) compare the fitnesses of individuals pursuing two possible learning strategies: holistic learners, who learn unitary agrammatical utterances to describe complex events, and compositional learners, who learn a system of labels for the component parts of complex events (e.g., individual words for objects and actions) and describe complex events using combinations of those simpler labels. They then calculate the number of labels that can be sustained in a population of such learners (which is dependent on the accuracy of learning) and the probability of successful communication between two individuals drawn from such a population (i.e., the probability that, for a given event, both individuals will know the word or words required to describe that event). Using this approach, they are able to show that the compositional learning strategy will be selected for (i.e., will lead to more successful communication) whenever the total number of complex events to be communicated about exceeds a certain threshold size: when there is a small number of events to be labeled (and/or words in a compositional system are harder to learn than unitary utterances), it pays to have a nongrammatical, holistic system of communication. On the other hand, if the number of events to be communicated about is large, then holistic systems become unwieldy (you are unlikely to have encountered the labels for every event) and the possibility of generalization afforded by the compositional strategy pays off: while it is unlikely that you and your communicative partner have heard precisely this complex event described before, you probably know the words for describing the component parts and can therefore successfully communicate anyway. While this model is extremely simple, it confirms the central intuition behind Pinker and Bloom's paper, namely that the large set of things that humans need to talk about will lead to evolutionary pressure in favor of learners who can acquire a rule-based system for expressing those concepts, that is, a compositional language.

Nowak et al. (2001) then consider the evolution of a restrictive UG, of the sort discussed above, which constrains learners to considering a relatively small set of possible languages (perhaps characterized by a small set of parameters).

They provide two general results. Firstly, there is selection in favor of a language faculty which reduces the size of the search space to levels that allow a population to converge on a shared grammar, as predicted by Pinker and Bloom (1990): if the space of possible languages considered by learners is very large, members of a population may end up arriving at different grammars on the basis of the same linguistic data and therefore fail to communicate; this can be avoided by reducing the size of the space of possible languages to below the 'coherence threshold,' which allows a population to reliably converge on a shared language. However, this selection in favor of a constrained UG has its limits, and will not lead to the most constrained possible learner: once the space of possible grammars has been reduced sufficiently to meet the coherence threshold, there is selection against being overly constrained. Essentially, there is an advantage to remaining flexible: an overly constrained learner risks being born into a population of less-constrained learners who use a language which the more constrained learner cannot acquire. Nowak et al. (2001) conclude that this balance of selective pressures will lead to learners whose UG is constrained enough to permit convergence within a population, but which is nevertheless permissive, allowing them to learn the largest possible set of languages within the limits of convergence. This result matches the prediction provided by Pinker and Bloom: constraints on language learning facilitate communication, at least up to a point.

Criticisms of the Adaptationist Account

Perhaps surprisingly, the adaptationist account of the evolution of UG has been criticized from within the nativist camp: most notably, Noam Chomsky, the founder of the theory of UG, has been notoriously reluctant to offer an account of the origins of this innate and language-specific organ of the human mind, preferring instead to offload the explanatory burden on the notion of inherent constraints on the nature of computation:

> Language is clearly a computational system. Therefore, it makes sense to seek the role of general principles of computational efficiency that apply far more broadly and may well be rooted in organism-independent natural law. This is particularly natural given the very little we know about the evolution of language. When a generative process (GP) emerged in some individual, perhaps by some slight rewiring of the brain, it should have been as simple as possible, the minimal change from the earlier state, presumably by mutation. Furthermore, there were no selectional pressures at that initial stage, so we should therefore expect the emergent GP to be determined solely by natural law.
>
> Chomsky, 2011: p. 271

Other work that takes evolutionary theory more seriously has identified potential problems with the evolutionary account put forward by Pinker and Bloom (1990). In a series of articles (e.g., Christiansen and Chater, 2008; Chater et al., 2009), Christiansen and Chater outline what they see as two key problems with the adaptationist account. Firstly, nativist accounts assume that UG is shared across our species, as the 'Universal' in 'Universal Grammar' would suggest, and as the ability of any child to acquire any human language seems to testify. Given that our species began dispersing across the planet 100 000 years ago, with concomitant diversification of the languages of those populations, they question why the same evolutionary mechanisms which lead to the development of UG would not continue to operate postdispersion, fine-tuning the 'U'G of each population to specialize in the acquisition of the languages of that region. While some accounts do suggest that there might indeed be subtle variations in the ability to acquire linguistic features (Dediu and Ladd, 2007), there is as yet no evidence for substantial geographic differences in the ability to acquire languages that might be predicted under this kind of theory. Secondly, Chater et al. (2009) use a computational model to demonstrate that biological evolution cannot encode innate constraints for linguistic features that change rapidly as a result of external factors (e.g., language contact): the Baldwin effect requires a reasonably stable target of assimilation; however, languages change fast, too fast to allow the Baldwin effect to operate.

However, much of the opposition to adaptationist accounts arises not from the internal logic of their account, but the rejection of their premise that language is fundamentally a biological capacity, that language learning is necessarily under strong innate constraints, and that the data a learner receives is deficient. It is to these alternative accounts that we now turn.

Statistical Learning and the Evolution of Learnability

Alternatives to the nativist accounts of language development focus on the human capacity for statistical learning, and make very different claims about the evolution of the human capacity to learn language. Statistical learning refers to the process of acquiring information about the environment – in this case, the linguistic environment – by identifying rules and statistical regularities in a rich sample of data provided by that environment. For language, this involves identifying features such as phonemes, syllables, word meanings, lexical categories, and syntax from the kinds of rich, raw linguistic data that children are exposed to (Romberg and Saffran, 2010; Rebuschat and Williams, 2012). Importantly, language learning is assumed to be possible without abstract innate knowledge of the learning task: rather, learning involves the application of domain-general learning procedures to multiple domains (see Saffran and Thiessen, 2007, for a review).

If language acquisition is a result of the application of domain-general learning processes that apply to learning about many aspects of the world (e.g., the social and physical environment) as well as language, then what can be said about the evolution of the human capacity to learn language? Under statistical learning accounts, language learning is not a separate capacity in need of a unique explanation: humans are unusual in the extent and ubiquity of their reliance on learned behaviors in many domains, and in particular on socially learned behaviors (behaviors that are learned from others). The challenge then becomes to explain why humans rely so much on (social) learning in general, including the social learning of their communication system. When framed

like this, the evolutionary account implicit in statistical learning seems quite close to the notion of a uniquely human cognitive niche proposed by Pinker (2010), but assumes that the human adaptation to this cognitive niche is a unique propensity for sophisticated social learning, rather than a suite of independent, domain-specific modular adaptations. Indeed, we may be able to gain an insight into the ancient roots of the human capacity for statistical learning by adapting techniques from comparative psychology, studying homologous statistical learning capabilities in other species, including other primates (Fitch and Hauser, 2004), rodents (Toro and Trobalon, 2005), and birds (Stobbe et al., 2012) (*see* Human Language: Evolutionary Precursors). Such an approach, in determining which elements of our capacity for language we share with other species and which are uniquely human, may also go some way to reconciling statistical learning accounts of the evolution of language development with nativist ones (Hauser et al., 2002).

Statistical learning accounts therefore assume that one of the key motivations for nativist accounts of language learning – the notion that language acquisition is easy, but would be impossible without innate knowledge of how languages work – is based on a failure to appreciate the power of domain-general statistical learning techniques. A recent strand of work also explores how the socially learned nature of language offers another potential solution to the 'mystery of fit' between language and language learners (Christiansen and Chater, 2008): rather than language learners adapting themselves to the challenge of language acquisition, languages adapt to suit their learners, minimizing the challenge of language acquisition. Languages persist by a process of cultural transmission (*see* First Language Acquisition, Developmental Psychology of), with the linguistic output of one generation forming the input for social learning in the next (*see* Evolution and Language: Cultural Transmission). Because language is transmitted in this way, language learners can have a direct influence on the structure of the language that subsequent generations of learners must learn, because that language passes through them:

> The structure of a language is under intense selection because in its reproduction from generation to generation, it must pass through a narrow bottleneck: children's minds… Language operations that can be learned quickly and easily by children will tend to get passed on to the next generation more effectively and intact than those that are difficult to learn. So, languages should change through history in ways that tend to conform to children's expectations.
>
> Deacon, 1997: p. 110

As a result of this process of filtering through the minds of language learners, languages will gradually evolve to reflect the biases inherent in language learning. Consequently, individual language learners are only faced with acquiring languages which have already been shaped for learning by previous generations:

> Human children appear preadapted to guess the rules of syntax correctly, precisely because languages evolve so as to embody in their syntax the most frequently guessed patterns.
>
> Deacon, 1997: p. 122

This pressure arising from repeated learning can also explain the presence of grammatical rules of the sort that Nowak et al. (2000) sought to explain by appealing to natural selection: the central observation by Nowak et al. that compositional linguistic rules facilitate generalization also means such rules are favored during language transmission, as they allow the learner to recover the language of the previous generation from limited data (Kirby, 2001; Zuidema, 2003; Brighton et al., 2005).

In addition to these pressures arising from learnability, language use will also leave its imprint on language design (Christiansen and Chater, 2008), as languages will also be subject to pressures resulting from both the language users' need for expressivity and preference for minimal processing costs (Kirby, 1999). Languages are so well designed for communication not because of the biological evolution of the language faculty under natural selection, but because of the accumulation of modifications made by people using language to communicate, which then show up in the linguistic data subsequent learners acquire their language from. Again, this is closely related to Pinker and Bloom's idea that language is adapted for communication: the only difference is the mechanism driving adaptation is the cultural evolution of languages themselves, rather than the adaptation of the language faculty on a biological timescale.

This account of language development, and its explanation of how languages adapt to their learners and users, also makes very different claims about the languages we would expect to observe around the world and any commonalities they share. Nativist accounts describe a comparatively transparent relationship between the parameter settings in individuals and the structural features of the languages they speak, and between the constraints of UG and linguistic universals. Statistical learning accounts, however, predict a more complex link between learner and language. As weak biases of learners can ultimately have strong effects on a language as it is repeatedly transmitted from generation to generation (see Smith, 2011, for review), a language universal may reflect an amplification of these weak, underlying biases rather than any absolute constraint on a language learner. There should therefore be very many more learnable, usable languages than those we can observe in the world now, and so the current typological distribution of world languages may only be a small subset of all plausible languages.

The relationship between the biases of individual language learners and the features of languages is also thought to be further obscured by mediating sociocultural factors, such as the number of users of a language, the complexity of a speech community's social structure, and the proportion of users who speak it as a second language, all thought to influence its linguistic characteristics. For example, a larger number of speakers of a language is proposed to result in the selection of more 'simple' grammatical features, such as more regular tense marking or lexical, rather than morphological, case marking, as the language is transmitted across generations (e.g., Dale and Lupyan, 2012; Nettle, 2012).

Conclusion

Humans have a unique capacity to acquire complex, expressive, grammatical language. This capacity must have some basis in

uniquely human biology, but precisely what this biological contribution is remains open for debate. We have reviewed two possibilities here. The capacity to acquire language may be an adaptation, a domain-specific component of the human mind evolved under natural selection due to the communicative payoff associated with the ability to learn language. Alternatively, languages themselves might evolve as they are passed from person to person, with language acquisition being the result of the application of less specialized, domain-general learning procedures. The resolution of this debate will depend in part on further painstaking exploration of the processes of language development in individuals, the evolutionary dynamics arising from the complex interplay between the biological and cultural processes that shape language, and the comparative psychology of statistical learning. In both cases, the evolution of the human capacity to acquire language is likely an evolutionary consequence of the uniquely social human ecology, characterized by the exchange of knowledge between individuals via language and social learning.

See also: Biological Preconditions for Language Development; First Language Acquisition, Developmental Psychology of; Human Language, Evolution of; Language Development, Theories of.

Bibliography

Baker, M., 2001. The Atoms of Language: The Mind's Hidden Rules of Grammar. Basic Books, New York.

Baldwin, J.M., 1896. A new factor in evolution. American Naturalist 30, 441–451.

Brighton, H., Smith, K., Kirby, S., 2005. Language as an evolutionary system. Physics of Life Reviews 2 (3), 177–226.

Chater, N., Reali, F., Christiansen, M.H., 2009. Restrictions on biological adaptation in language evolution. Proceedings of the National Academy of Sciences, USA 27, 1015–1020.

Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.

Chomsky, N., 1981. Lectures on Government and Binding: The Pisa Lectures. Foris, Dordrecht.

Chomsky, N., 2011. Language and other cognitive systems. what is special about language? Language Learning and Development 7, 263–278.

Christiansen, M.H., Chater, N., 2008. Language as shaped by the brain. Behavioral and Brain Sciences 31, 489–509.

Dale, R., Lupyan, G., 2012. Understanding the origins of morphological complexity: the linguistic niche hypothesis. Advances in Complex Systems 15 (3 & 4), 1150017.

Deacon, T., 1997. The Symbolic Species. Penguin, London.

Dediu, D., Ladd, D.R., 2007. Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, *aspm* and *microcephalin*. Proceedings of the National Academy of Sciences, USA 104, 10944–10949.

Fitch, W.T., 2000. The evolution of speech: a comparative review. Trends in Cognitive Sciences 4 (3), 258–267.

Fitch, W.T., Hauser, M.D., 2004. Computational constraints on syntactic processing in a nonhuman primate. Science 303, 377–380.

Hauser, M.D., Chomsky, N., Fitch, W.T., 2002. The faculty of language: what is it, who has it, and how did it evolve. Science 298, 1569–1579.

Kirby, S., 1999. Function, Selection and Innateness: The Emergence of Language Universals. Oxford University Press, Oxford.

Kirby, S., 2001. Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. IEEE Transactions on Evolutionary Computation 5 (2), 102–110.

Nettle, D., 2012. Social scale and structural complexity in human languages. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 367 (1597), 1829–1836.

Nowak, M.A., Krakauer, D.C., Dress, A., 1999. An error limit for the evolution of language. Proceedings of The Royal Society of London. Series B, Biological Sciences 266 (1433), 2131–2136.

Nowak, M.A., Plotkin, J.B., Jansen, V.A.A., 2000. The evolution of syntactic communication. Nature 404, 495–498.

Nowak, M.A., Komarova, N.L., Niyogi, P., 2001. Evolution of universal grammar. Science 291, 114–117.

Piattelli-Palmarini, M., 1989. Evolution, selection and cognition: from "learning" to parameter setting in biology and in the study of language. Cognition 31, 1–44.

Pinker, S., 2010. The cognitive niche: coevolution of intelligence, sociality, and language. Proceedings of the National Academy of Sciences, USA 107, 8993–8999.

Pinker, S., Bloom, P., 1990. Natural language and natural selection. Behavioral and Brain Sciences 13 (4), 707–784.

Rebuschat, P., Williams, J., 2012. Introduction: statistical learning and language acquisition. In: Rebuschat, P., Williams, J. (Eds.), Statistical Learning and Language Acquisition. Mouton de Gruyter, Boston, pp. 1–12.

Romberg, A.R., Saffran, J.R., 2010. Statistical learning and language acquisition. Wiley Interdisciplinary Reviews: Cognitive Science 1 (6), 906–914.

Saffran, J.R., Thiessen, E.D., 2007. Domain-general learning capacities. In: Hoff, E., Shatz, M. (Eds.), Handbook of Language Development. Blackwell, Cambridge, pp. 68–86.

Smith, K., 2011. Learning bias, cultural evolution of language and the biological evolution of the language faculty. Human Biology 83, 261–278.

Stobbe, N., Westphal-Fitch, Gesche, Aust, Ulrike, Tecumseh Fitch, W., 2012. Visual artificial grammar learning: comparative research on humans, kea (*Nestor notabilis*) and pigeons (*Columba livia*). Philosophical Transactions of the Royal Society B: Biological Sciences 367 (1598), 1995–2006.

Tooby, J., Cosmides, L., Barrett, H.C., 2005. Resolving the debate on innate ideas: learnability constraints and the evolved interpenetration of motivational and conceptual functions. In: Carruthers, P., Laurence, S., Stich, S. (Eds.), The Innate Mind: Structure and Content. Oxford University Press, pp. 305–337.

Toro, J.M., Trobalon, J.B., 2005. Statistical computations over a speech stream in a rodent. Perception and Psychophysics 67 (5), 867–875.

Zuidema, W.H., 2003. How the poverty of the stimulus solves the poverty of the stimulus. In: Becker, S., Thrun, S., Obermayer, K. (Eds.), Advances in Neural Information Processing Systems, vol. 15. MIT Press, Cambridge, MA, pp. 51–58.

Language Development: Influence of Socio-Economic Status

Erika Hoff and Krystal M Ribot, Florida Atlantic University, Davie, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Family socioeconomic status (SES) is related to language development in multiple domains, throughout childhood, and in adulthood as well. Children from lower SES homes show lower levels of language and communicative skill than children from higher SES homes beginning in infancy. The gap between lower and higher SES children persists and sometimes widens with age. Substantial evidence locates the source of these differences in children's language experience. Children from lower SES homes are talked to less, and the speech they hear is less supportive of language development. The evidence that SES-related differences in parents' speech cause SES-related differences in children's language development is consistent with theoretical approaches that give experience an important role in the language acquisition process.

The goal of this article is to summarize the evidence that socioeconomic status (SES) is related to language development and to review the literature that explains why SES has this effect. Family SES is a powerful predictor of many aspects of child development and, relatedly, of many aspects of parenting behavior (Hoff et al., 2002; Keating and Hertzman, 1999; Linver et al., 2002; National Research Council and Institute of Medicine, 2000). This article focuses on the relation of SES to children's language development and to the antecedent SES-related differences in how parents use language in interaction with their children. The first section of this article provides a definition of SES. The second and third sections review the evidence that SES influences children's language acquisition and does so via the influence of SES-related differences in the speech adults address to children. A final section considers the theoretical implications of the observed relations among SES, language experience, and language development.

Definitions of Socioeconomic Status

SES refers to an individual's or household's relative position in a social hierarchy. Traditionally, this hierarchy has been defined in terms of access to, or control over, wealth, prestige, and power (Mueller and Parcel, 1981). More recently, SES has been conceptualized in terms of three sorts of capital or resources that families have access to: financial capital (material resources), human capital (nonmaterial resources, such as knowledge and skills), and social capital (connections to a larger social group) (Coleman, 1988). Researchers have measured SES using indicators of parents' level of education, household income, and prestige of occupation, or using a composite index based on multiple indicators (Hoff et al., 2012). Each component of SES may be associated with some level of risk for poor developmental outcomes because both US and international data suggest that greater disparities and/or greater predictive power are associated with the use of multiple indicators (Fernald et al., 2011; Halle et al., 2009). Often, however, research has used maternal education as the sole index of SES on the logic that education is highly correlated with other measures, is stable, and appears to be the best single predictor of child outcomes (Hoff et al., 2012). Furthermore, some research indicates that outside of extreme poverty, maternal education is the component of SES most strongly related to parenting measures (Bornstein et al., 2003; Hoff et al., 2002).

Relations between SES and Language Development

Throughout childhood, higher SES children have more advanced language skills than lower SES children of the same age. SES-related differences in children's communicative behavior have been found as early as infancy, are consistently found in the preschool and school-age years, and appear in adulthood as well. For instance, Halle et al.'s (2009) study of a nationally representative sample of approximately 11 000 infants born in the United States in 2001 found SES disparities in language skills (e.g., expressive vocalization and naming objects) in infants as early as 9 months of age, and the SES-related disparities widen by 24 months of age (e.g., in receptive and expressive vocabulary and listening/comprehension). In a public school sample, Oller and Eilers (2002) found SES-related differences in kindergarten and 5th grade children's scores on standardized tests of oral vocabulary equal to almost one standard deviation. Among adults, Hoff-Ginsberg (1991) found that education level was related to the amount of speech, the lexical richness, and the syntactic complexity of utterances that mothers addressed to an interviewer, and Van den Broeck (1977) found education level was related to the syntactic complexity of the speech used by adults in conversation.

SES-related differences in children's language skill have been observed across multiple domains of language including vocabulary, grammar, communication, narrative skill, phonological awareness, and processing speed.

SES Differences in Vocabulary

A difference in vocabulary size is the most reliably observed SES-related difference in language skill. A vocabulary gap has been found as early as the toddler years (Halle et al., 2009; Hart and Risley, 1995). The gap further widens by age 4 and then

remains fairly stable through the school years (Farkas and Beron, 2004). The magnitude of differences related to SES can be large. Hart and Risley (1995) found that by 3 years of age, the mean cumulative recorded vocabulary for the higher SES children in that sample was over 1000 words while for the lower SES children it was close to 500; SES accounted for 36% of variance in vocabulary in this sample. Arriaga et al. (1998) found that 80% of a sample of low-SES children between 18 and 30 months scored below the 50th percentile in productive vocabulary, using a test normed on a mid- to high-SES reference group. Comparing children of college-educated mothers to children of high school-educated mothers, Hoff-Ginsberg (1998) found SES-related differences in the size of the vocabularies they used in spontaneous speech. SES accounted for 5% of the variance in children's vocabulary in this sample.

Other studies using spontaneous speech, maternal report measures, and standardized tests show similar findings, with the magnitude of the SES-related differences depending on the range of SES in the sample studied (Dollaghan et al., 1999; Pan et al., 2005; Rescorla, 1989). For example, SES accounted for less than 1% of the variance in vocabulary assessed via maternal report using the MacArthur Communicative Development Inventory (MCDI) in a fairly homogeneous, middle-class sample (Fenson et al., 1994). Snow (1999) reported that SES differences are greater for measures of productive vocabulary than comprehension vocabulary, but Wells (1986) reported that the clearest SES-related difference is in oral comprehension. These vocabulary differences have long-term consequences. Vocabulary is a fundamental predictor of school success, and vocabulary deficits are one of the primary reasons that lower SES children enter school at greater risk for failure than their high-SES peers (Snow et al., 1998).

SES Differences in Grammar

SES-related differences in grammatical development have also been found. Children from higher social strata have been found to produce longer responses to adult speech (McCarthy, 1930), score higher on standardized tests that include measures of grammatical development (Dollaghan et al., 1999), produce more complex utterances in spontaneous speech as toddlers (Arriaga et al., 1998) and at age 5 years (Snow, 1999), and perform significantly better on measures of productive and receptive syntax at age 6 years (Huttenlocher et al., 2002). As an indicator of the magnitude of these effects on grammatical development, the low-income sample studied by Snow (1999) had an average mean utterance length (MLU) at age 3;9 that would be typical of children more than a year younger according to norms based on a middle-class sample. At age 5;6, they had an average MLU typical of middle-class children aged 3;1. Arriaga et al. (1998) found that 70% of low-income children were below the 50th percentile on the MCDI measure of sentence complexity. The SES-related differences are not in whether or not children can use complex structures in their speech, but in the frequency with which they do so (Huttenlocher et al., 2002; Tough, 1982).

Effects of SES on grammatical development are not as large or reliable as effects on vocabulary; some studies have found no relation between SES and measures of children's syntactic development (Hoff-Ginsberg, 1998; Hoff and Tian, 2005; Jackson and Roberts, 2001; Noble et al., 2005). A reason for these inconsistent findings is suggested by the finding in Vasilyeva et al. (2008) that SES was not related to measures of early mastery of basic syntactic rules but was related – in the same sample of children – to later differences in the use of more complex structures.

SES Differences in Communication

Studies of school-aged children find SES-related differences in the communicative purposes to which language is put, such that children with less educated parents less frequently use language to analyze and reflect, to reason and justify, or to predict and consider alternative possibilities than children with more educated parents. It has been suggested that the structural differences in children's language associated with SES may be a by-product of these functional differences (Tough, 1982). In an early sociological analysis of class and language, Bernstein (1970) similarly argued that higher and lower SES children do not differ in language knowledge but the differences in the communicative burden carried by language give rise to differences in the amount, the structure, and the lexical repertoire of the speech produced.

SES-related differences in school-aged children also appear in the ability to communicate meaning through language and to draw meaning from language – sometimes referred to as speaker and listener skills (Lloyd et al., 1998). In the referential communication task, which requires children to describe one item in an array of objects so that a visually separated listener with the same array can identify that item, lower SES children are less able than higher SES children to produce sufficiently informative messages and to use information in messages addressed to them to make correct choices (Lloyd et al., 1998). Children from lower socioeconomic strata also perform less well than higher SES children in solving mathematics word problems. That some of this difference is attributable to language ability, not mathematical ability, is suggested by evidence from one sample of lower SES children who showed poorer performance than higher SES children in word problems and in tests of verbal skills but who did not differ in their performance in math calculations (Jordan et al., 1992).

In sum, the findings are clear that children from different socioeconomic strata display different levels of language skill across multiple domains of language. It is also true that the specifics of SES-language relations and the size of the effects depend on the particular language outcome and the range of SES under consideration.

Other SES-Related Differences

In addition to these SES-related differences in vocabulary, grammar, and communication, there are SES-related differences in children's narrative skills, in their phonological awareness, and in their speed of language processing. The narratives produced by lower SES children are less sophisticated than the narratives produced by middle-class children of the same age, when assessed in terms of topic coherence and

independence from the nonlinguistic context (Heath, 1983; Vernon-Feagans et al., 2001). Lower SES children show lower levels of phonological awareness than middle-class children, with the size of the SES-related difference increasing from the age of 2 to 5 years (Bowey, 1995; Lonigan et al., 1998; McDowell et al., 2007). Lower SES children are also slower in accessing the words they know (Fernald et al., 2012).

In sum, a substantial body of evidence argues that the effect of SES on children's early language skills is pervasive and robust. The size of the effect differs depending on the range of SES being considered, the age of the children, and the aspect of language under consideration, but differences as large as 1 standard deviation are frequently observed and larger differences are also possible. Using age as a measure of the size of the SES-related difference, there are findings that children from low-income backgrounds are a year or more behind their middle-class counterparts (e.g., Snow, 1999).

It is important to note that neither low- nor high-SES children are monolithic in their language skill, that substantial individual differences exist within both populations, and that the distributions of skills among lower and higher SES children overlap (Hoff, 2003a,b; Pan et al., 2005). It is also important to note that low-SES children have unique linguistic strengths that are not captured by the procedures and tests most frequently used in research. Arguments for the unique skills of lower SES children tend to be in the form of ethnographic studies of specific populations and sometimes of particular uses of language that are unique to those populations. For example, Heath (1983) described the narrative skills displayed by very young boys in a rural, lower SES African-American community in the southeastern United States. Boys under the age of 4 years were able to hold the floor and engage the attention of adults with the stories they told, using a variety of poetic devices, sound effects, and accompanying movement. The narrative cohesion that is part of standard assessments of children's narrative skill is not a valued property of stories in this community. Similar skills in low-SES African-American boys have been described by Vernon-Feagans et al. (2001), who also described the children's narratives as being jointly constructed with other, older children – in contrast to the narrative performances of mainstream children, which are often monologues. Research with both European-American and African-American children has described other complex skill sets that lower SES children must master including those used in teasing interchanges (Miller, 1986), in ritualized insults (Abrahams, 1962), and in the combination of improvisational rhymes, steps, and hand claps that form the 'steps' performances of preadolescent African-American girls (Gilmore, 1986).

In addition to their description of the particular skills of particular groups, these studies make the broader point that there is substantial sociocultural variation in norms for language use that can make children from nonmainstream backgrounds look deficient when viewed from the perspective of mainstream expectations, while they are not at all deficient according to the norms for their own group. A similar argument applies to the interpretation of standardized test scores, which are frequently developed and normed using middle-class reference groups and may not tap the skills of children from other backgrounds. In the extreme, this line of reasoning argues that any comparison of language development across socio-economic strata is conceptually ill-founded because lower SES children are learning a different style of language use than higher SES children (see Hoff, 2013). The argument is that lower SES children have different language skills than higher SES children, but they do not have deficient language skills. It is possible, however, for there to be both stylistic differences and differences in language knowledge and skill, and the weight of the evidence argues that differences in knowledge and skill exist. Only two studies have reported SES-related differences in children's language in the direction of lower SES children being more advanced on measures of language knowledge. Both used the MCDI, and both attributed their findings to lower SES mothers' tendencies to overestimate their children's abilities (Feldman et al., 2000; Fenson et al., 1994).

Sources of SES-Related Differences in Language Development

The observed relations between SES and child language could be the result of several factors including (1) biologically based differences in children's abilities, caused by genes or health; (2) global effects of differences in family functioning and home environments (Linver et al., 2002); and/or (3) specific effects of differences in language-learning experiences (Hoff and Naigles, 2002; Hoff-Ginsberg, 1998). A substantial body of evidence argues for the third hypothesis that SES-related differences in children's language skills are the result of SES-related differences in their language experience.

Evidence of SES-related differences in mothers' child-directed speech abounds. Compared to mothers with more education, mothers with less education talk less to their children, and the nature of the speech they address to children is less supportive of language development than is the speech of more educated mothers. Lower SES mothers address speech to their children more frequently for the purpose of directing their children's behavior and less frequently for the purpose of eliciting and maintaining conversation (Hart and Risley, 1995; Hoff, 2006). In talking to their children, lower SES mothers make use of a smaller vocabulary, less complex syntactic structures, and less variety among the syntactic structures used, compared to higher SES mothers (Hoff, 2003a,b; Huttenlocher et al., 2007). Lower SES mothers have been found to spend less time in mutual activities with their children than do middle-SES mothers, and their conversations do not necessarily rely on their children's speech (e.g., Bee et al., 1969; Farran and Haskins, 1980).

Multiple studies have found that the properties characteristic of higher SES mothers are positive predictors of children's language development – even within SES (Hoff, 2006; Huttenlocher et al., 2010). Children who experience more responsive verbal environments, who hear more speech, who hear a richer vocabulary, and who hear a greater variety of syntactic structures acquire language more rapidly than children who hear less speech and who hear speech with lower levels of these supportive features (Hoff, 2006). One study found that properties of maternal speech fully mediated an SES-related difference in 2-year-olds' vocabulary (Hoff,

2003a,b). Evidence that the relation between input and language acquisition is causal, and not just a reflection of genetically based similarity in the verbal skills of mothers and children, comes from studies of teacher input effects. Children whose teachers provide more language-advancing input progress more in their language over the course of the school year than children with teachers whose language use is less supportive (Dickinson and Porche, 2011; Huttenlocher et al., 2002).

The sources of SES-related differences in language skills among older children are less directly studied. It is clear, however, that school does not eliminate the early-appearing differences. Rather, language skills at school entry predict later language and literacy skills (Dickinson et al., 2003; Muter et al., 2004; NICHD Early Child Care Research Network, 2005). SES-related differences in children's language skills at school entry are thought to play a large role in explaining the persistent SES-related achievement gap (Hoff, 2013).

Theoretical Implications of Relations Between SES and Language Development

Each of the candidate explanations of the SES–child language relation derives from a broader scope theory of how and to what extent the environment influences development in general, and language acquisition in particular. One view is that development, in particular language development, unfolds following a genetic blueprint (e.g., Pinker, 2002). The alternative view, of course, is that the environment plays a substantial role, and the evidence that SES-related differences in children's language development have their source in differences in their experience supports the latter view. Pointing to the environment, however, does not explain how the environment exerts its influence. It could be that the effect of the environment is global. That is, supportive environments benefit all aspects of development and unsupportive environments impede them. The fact that the effects of SES are pervasive and cross developmental domains is consistent with this global effect model.

A variety of evidence argues for the principle of environmental specificity, however. According to this principle, different aspects of the environment influence different aspects of development (Wachs, 1991). For example, the aspects of maternal behavior that predict language development are different from the aspects of maternal behavior that predict play development (Lyytinen et al., 2003; Tamis-LeMonda and Bornstein, 1994). Within language, the aspects of experience that influence the development of communicative skill are different from the aspects of experience that influence the development of language knowledge (Bernicot and Roux, 1998; Hoff-Ginsberg, 1998; Wachs and Chan, 1986). It is also likely that the input requirement for word learning and the input requirement for learning grammar differ, and that different types of words and different aspects of grammar further make their own requirements of experience. One reason that SES effects are more reliably observed in vocabulary than grammar and more for later than earlier grammatical acquisitions may have to do with the different requirements these acquisitions make and the relation of SES of each of those properties of input. For example, Huttenlocher et al. (2010) have argued that SES differences appear for later and not earlier grammatical development because early acquisitions depend on features of input that are always available while later acquisitions depend on features of input that vary as a function of education.

Conclusion

To sum, there is ample evidence in the literature that both the environments in which children acquire language and the rate of their language development vary as a function of family SES. It has been documented extensively that mothers from different socioeconomic strata interact with and talk to their children differently, and that these factors influence the rates at which children acquire language. The weight of the evidence suggests that SES affects children's opportunities for communicative interaction and the availability of language input with the consequence that, even after effects of language style are taken into account, the rate of children's language development differs as a function of SES. The evidence that SES-related differences in parents' speech cause SES-related differences in children's language development is consistent with theoretical approaches that give experience an important role in the language acquisition process.

Bibliography

Abrahams, R., 1962. Playing the dozens. The Journal of American Folklore 75, 209–220. http://dx.doi.org/10.2307/537723.

Arriaga, R.K., Fenson, L., Cronan, T., Pethick, S.J., 1998. Scores on the MacArthur communicative Development Inventory of children from low- and middle-income families. Applied Psycholinguistics 19, 209–225.

Bee, H.L., Van Egeren, L.F., Streissguth, A.P., Nyman, B.A., Leckie, M.A., 1969. Social class differences in maternal teaching strategies and speech patterns. Developmental Psychology 1, 726–734.

Bernicot, J., Roux, M., 1998. La structure et l'usage des énoncés: comparison d'enfants uniques et d'enfants second nés. In: Bernicot, J., Marcos, H., Day, C., Guidetti, M., Rabain-Jamin, J., Laval, V., Babelot, G. (Eds.), De l'usage des gestes et des mots chez les enfants. Armand Colin, Paris, pp. 157–178.

Bernstein, B.B., 1970. A socio-linguistic approach to social learning. In: Williams, F. (Ed.), Language and Poverty. Markham, Chicago, pp. 25–61.

Bornstein, M.H., Hahn, C.-S., Suwalsky, J.T.D., Haynes, O.M., 2003. Socioeconomic status, parenting, and child development: the Hollingshead four-factor index of social status and the socioeconomic index of occupations. In: Bornstein, M.H., Bradley, R.H. (Eds.), Socioeconomic Status, Parenting, and Child Development. Erlbaum, Mahwah, NJ, pp. 29–82.

Bowey, J.A., 1995. Socioeconomic status differences in preschool phonological sensitivity and first-grade reading achievement. Journal of Educational Psychology 87, 476–487. http://dx.doi.org/10.1037/0022-0663.87.3.476.

Coleman, J.S., 1988. Social capital in the creation of human capital. The American Journal of Sociology 94, S95–S120.

Dickinson, D.K., McCabe, A., Anastasopoulos, L., Peisner-Feinberg, E.S., Poe, M.D., 2003. The comprehensive language approach to early literacy: the interrelationships among vocabulary, phonological sensitivity, and print knowledge among preschool-aged children. Journal of Educational Psychology 95, 465–481. http://dx.doi.org/10.1037/0022-0663.95.3.465.

Dickinson, D.K., Porche, M.V., 2011. Relation between language experiences in preschool classrooms and children's kindergarten and fourth-grade language and reading abilities. Child Development 82, 870–886. http://dx.doi.org/10.1111/j.1467-8624.2011.01576.x. PMid:21413936.

Dollaghan, C.A., Campbell, T.F., Paradise, J.L., Feldman, H.M., Janosky, J.E., Pitcairn, D.N., et al., 1999. Maternal education and measures of early speech and language. Journal of Speech, Language and Hearing Research 42, 1432–1443.

Farkas, G., Beron, K., 2004. The detailed age trajectory of oral vocabulary knowledge: differences by class and race. Social Science Research 33, 464–497.

Farran, D.C., Haskins, R., 1980. Reciprocal influence in the social interaction of mothers and three-year-old children from different socioeconomic backgrounds. Child Development 51, 780–791.

Feldman, H.M., Dollaghan, C.A., Campbell, T.F., Kurs-Lasky, M., Janosky, J.E., Paradise, J.L., 2000. Measurement properties of the MacArthur communicative development inventory at ages one and two years. Child Development 71, 310–322.

Fenson, L., Dale, P.S., Reznick, J.S., Bates, E., Thal, D.J., Pethick, S.J., 1994. Variability in Early Communicative Development. Monographs of the Society for Research in Child Development 59 Serial No. 242.

Fernald, A., Weisleder, A., Marchman, V., 2012. The Poverty of the Stimulus Has Developmental Consequences. Paper presented at the Boston University Conference on Language Development. Boston, Massachusetts, November 2–4.

Fernald, L.C.H., Weber, A., Galasso, E., Ratsifandrihamanana, R., 2011. Socioeconomic gradients and child development in a very low income population: evidence from Madagascar. Developmental Science 14, 832–847.

Gilmore, P., 1986. Sub-rosa literacy: peers, play and ownership in literacy acquisition. In: Schieffelin, B.B., Gilmore, P. (Eds.), The Acquisition of Literacy: Ethnographic Perspectives. Ablex, Norwood, NJ, pp. 155–168.

Halle, T., Forry, N., Hair, E., Pepper, K., Wandner, L., Wessel, J., Vick, J., 2009. Disparities in Early Language and Development: Lessons from the Early Childhood Longitudinal Study – Birth Cohort (ECLS-b). Child Trends, Washington, DC.

Hart, B., Risley, T.R., 1995. Meaningful Differences in the Everyday Experiences of Young American Children. Paul H. Brookes Publishing, Baltimore, MD.

Heath, S.B., 1983. Ways with Words: Language, Life, and Work in Communities and Classrooms. Cambridge University Press, London.

Hoff, E., 2003a. The specificity of environmental influence: socioeconomic status affects early vocabulary development via maternal speech. Child Development 74, 1368–1378.

Hoff, E., 2003b. Causes and consequences of SES-related differences in parent-to-child speech. In: Bornstein, M.H., Bradley, R.H. (Eds.), Socioeconomic Status, Parenting, and Child Development. Lawrence Erlbaum Assoc., Mahwah, NJ, pp. 147–160.

Hoff, E., 2006. How social contexts support and shape language development. Developmental Review 26, 55–88.

Hoff, E., 2013. Interpreting the early language trajectories of children from low-SES and language minority homes: Implications for closing achievement gaps. Developmental Psychology 49, 4–14. http://dx.doi.org/10.1037/a0027238.

Hoff, E., Laursen, B., Bridges, K., 2012. Measurement and model building in studying the influence of socioeconomic status on child development. In: Lewis, M., Mayes, L. (Eds.), A Developmental Environmental Measurement Handbook. Cambridge University Press, Cambridge, England.

Hoff, E., Laursen, B., Tardif, T., 2002. Socioeconomic status and parenting. In: Bornstein, M.H. (Ed.), Handbook of Parenting: Ecology and Biology of Parenting, vol. II. Lawrence Erlbaum, Mahwah, New Jersey, pp. 161–188.

Hoff, E., Naigles, L., 2002. How children use input to acquire a lexicon. Child Development 73, 418–433.

Hoff, E., Tian, C., 2005. Socioeconomic status and cultural influences on language. Journal of Communication Disorders 38, 271–278.

Hoff-Ginsberg, E., 1991. Mother–child conversation in different social classes and communicative settings. Child Development 62, 782–796.

Hoff-Ginsberg, E., 1998. The relation of birth order and socioeconomic status to children's language experience and language development. Applied Psycholinguistics 19, 603–629.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., Levine, S., 2002. Language input at home and at school: relation to child syntax. Cognitive Psychology 45, 337–374.

Huttenlocher, J., Vasilyeva, M., Waterfall, H.R., Vevea, J.L., Hedges, L.V., 2007. The varieties of speech to young children. Developmental Psychology 43, 1062–1083.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., Hedges, L.V., 2010. Sources of variability in children's language growth. Cognitive Psychology 61, 343–365. http://dx.doi.org/10.1016/j.cogpsych.2010.08.002. PMid:20832781.

Jackson, S.C., Roberts, J.E., 2001. Complex syntax production of African American preschoolers. Journal of Speech, Language & Hearing Research 44, 1083–1096.

Jordan, N., Huttenlocher, J., Levine, S., 1992. Differential calculation abilities in young children from middle and low-income families. Developmental Psychology 28, 644–653.

Keating, D.P., Hertzman, C., 1999. Developmental Health and the Wealth of Nations. Guildford, New York.

Linver, M.R., Brooks-Gunn, J., Kohen, D.E., 2002. Family processes and pathways from income to young children's development. Developmental Psychology 38, 719–734.

Lloyd, P., Mann, S., Peers, I., 1998. The growth of speaker and listener skills from five to eleven years. First Language 18, 81–104.

Lonigan, C.J., Burgess, S.R., Anthony, J.L., Barker, T.A., 1998. Development of phonological sensitivity in 2- to 5-year-old children. Journal of Educational Psychology 90, 294–311.

Lyytinen, P., Eklund, K., Lyytinen, H., 2003. The plan and language behavior of mothers with and without dyslexia and its association to their toddlers' language development. Journal of Learning Disabilities 36, 74–86.

McCarthy, D., 1930. The Language Development of the Preschool Child. The University of Minnesota Press, Minneapolis.

McDowell, K.D., Lonigan, D.J., Goldstein, H., 2007. Relations among socioeconomic status, age, and predictors of phonological awareness. Journal of Speech, Language, and Hearing Research 50, 1079–1092.

Miller, P., 1986. Teasing as language socialization and verbal play in a white working-class community. In: Schieffelin, B.B., Ochs, E. (Eds.), Language Socialization across Cultures. Cambridge University Press, Cambridge, England, pp. 199–212.

Mueller, C.W., Parcel, T.L., 1981. Measures of socioeconomic status: alternatives and recommendations. Child Development 52, 13–30.

Muter, V., Hulme, C., Snowling, M.J., Stevenson, J., 2004. Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: evidence from a longitudinal study. Developmental Psychology 40, 665–681. http://dx.doi.org/10.1037/0012-1649.40.5.665. PMid:15355157.

National Institute of Child Health and Human Development Early Child Care Research Network, 2005. Pathways to reading; the role of oral language in the transition to reading. Developmental Psychology 41, 428–442. http://dx.doi.org/10.1037/0012-1649.41.2.428. PMid:15769197.

National Research Council and Institute of Medicine, 2000. From Neurons to Neighborhoods: The Science of Early Childhood Development. National Academy Press, Washington, DC.

Noble, K.G., Norman, M.F., Farah, M.J., 2005. Neurocognitive correlates of socioeconomic status in kindergarten children. Developmental Science 8, 74–87.

Oller, D.K., Eilers, R.E. (Eds.), 2002. Language and Literacy in Bilingual Children. Multilingual Matters, Clevedon.

Pan, B.A., Rowe, L.M., Singer, J.D., Snow, C.E., 2005. Maternal correlates of growth in toddler vocabulary production in low-income families. Child Development 76, 763–782.

Pinker, S., 2002. The Blank Slate. Viking Press, New York.

Rescorla, L., 1989. The language development survey: a screening tool for delayed language in toddlers. Journal of Speech and Hearing Disorders 54, 587–599.

Snow, C.E., 1999. Social perspectives on the emergence of language. In: MacWhinney, B. (Ed.), The Emergence of Language. Erlbaum, Mahway, NJ, pp. 257–276.

Snow, C.E., Burns, M.S., Griffin, P., 1998. Preventing Reading Difficulties in Young Children. National Academy Press, Washington, DC.

Tamis-LeMonda, C.S., Bornstein, M.H., 1994. Specificity in mother-toddler language-play relations across the second year. Developmental Psychology 30, 283–292.

Tough, J., 1982. Language, poverty, and disadvantage in school. In: Feagans, L., Farran, D.C. (Eds.), The Language of Children Reared in Poverty: Implications for Evaluation and Intervention. Academic Press, New York, pp. 3–18.

Van den Broeck, J., 1977. Class differences in syntactic complexity in the Flemish town of Maaseik. Language in Society 6, 149–181.

Vasilyeva, M., Waterfall, H., Huttenlocher, J., 2008. Emergence of syntax: commonalities and differences across children. Developmental Science 11, 84–97.

Vernon-Feagans, L., Hammer, C.S., Miccio, A., Manlove, E., 2001. Early language and literacy skills in low-income African American and Hispanic children. In: Neuman, S.B., Dickinson, D.K. (Eds.), Handbook of Early Literacy Research. Guilford, New York, pp. 192–210.

Wachs, T.D., 1991. Environmental considerations in studies with nonextreme groups. In: Wachs, T.D., Plomin, R. (Eds.), Conceptualization and Measurement of Organism-environment Interaction. American Psychological Association, Washington, DC, pp. 44–67.

Wachs, T.D., Chan, A., 1986. Specificity of environmental action, as seen in environmental correlates of infants' communication performance. Child Development 57, 1464–1474.

Wells, G., 1986. The Meaning Makers: Children Learning Language and Using Language to Learn. Heinemann, Portsmouth, NH.

Language Disorders, Treatment and Remediation of

Dimitrios S Kasselimis, University of Crete, Crete, Greece; and Eginition Hospital, National and Kapodistrian University of Athens, Athens, Greece
Constantin Potagas, Eginition Hospital, National and Kapodistrian University of Athens, Athens, Greece

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Language disturbances often occur after lesion – usually stroke – in the left hemisphere. In this article we aim to review methods of treatment and remediation of aphasic deficits. We first discuss the underlying neural mechanisms of recovery after stroke. We then focus on the several traditional approaches to aphasia rehabilitation and further refer to contemporary noninvasive, such as constrained-induced therapy, and invasive, such as pharmacological augmentation, treatments. Issues on efficacy of aphasia therapy are also discussed.

Introduction

Pierre-Marie claimed that there is a single form of language impairment, in the sense that only Wernicke's aphasia manifests symptoms that reveal a real cognitive deficit, and subsequently a breakdown of the neural language network. Patients with *aphemie*, described earlier by Broca, were considered to be individuals who, while suffering from anarthria, retained their core language-processing abilities. Global aphasia was the result of a lesion affecting both posterior temporal cortices, thus producing aphasia, and 'Marie's Quadrilateral,' a region including the insula, the underlying white matter, and subcortical nuclei, thus producing anarthria (Lecours et al., 1992; Tesak and Code, 2008). Marie's view seems nowadays to be oversimplified, since accumulating data from neuroimaging and lesion studies reveal a rather complex image with regard to language function and linguistic impairments. Nevertheless, his argument still stands: a linguistic disorder must have a cognitive component in order to be defined as such. Communications problems may arise from a variety of clinical conditions, such as dysarthria, stuttering, dysprosodia, etc. However, in the present article we will focus on acquired language disturbances due to underlying cognitive impairments resulting from focal or diffuse brain lesions. The generic term 'aphasia' is often used to describe such disturbances, which are usually due to a cerebrovascular accident (CVA). In the following sections we will discuss two major issues: recovery from stroke, and the variety of approaches and methods with regard to aphasia therapy.

Recovery after Stroke

Brain's capacity for reorganization after injury demonstrated in several studies (see, for example, the remarkable case of Alex, published by Vargha-Khadem in 1997, a 9-year-old boy who regained linguistic abilities after left hemispherectomy) clearly indicates that plasticity is the essence of recovery. On the basis of the relevant literature, it can be asserted that post-stroke recovery of cognitive functions is a gradual process that mainly takes place during a time-window of 1 year post-onset. The majority of stroke patients will show signs of significant recovery during the first 6 months (see, for example, Nicholas et al., 1993). The first 3 months after the CVA are considered to be the subacute post-stroke period (e.g., Gottesman and Hillis, 2010). Studies on cognitive impairment resulting from CVA suggest that the greater bulk of recovery can be observed within this particular timeframe (Demeurisse et al., 1980; Engelter et al., 2004; Hachioui et al., 2013; Lazar et al., 2010; Laska et al., 2001; Pedersen et al., 1995; Pickersgill and Lincoln, 1983; Sarno and Levita, 1971; Stone et al., 1992). According to general consensus, aphasic patients can be characterized as chronic after the 6-month post-onset period. However, there are findings reported in the literature that contradict the notion that recovery is restricted in the 3-month post-stroke period suggested above (e.g., Smania et al., 2010). Furthermore, it must be clarified that even if spontaneous recovery is limited in the acute and post-acute phase, improvement of linguistic abilities can be achieved also in chronic aphasia, provided that an appropriate treatment is implemented (Basso, 2003). It should be further noted that the rate and time spectrum of recovery is not uniform for all patients and heavily depends on individual differences. Several factors, such as lesion type, volume and locus, age, and initial severity, may affect the extent and rate of recovery. In general, advanced age, increased initial severity, cerebrovascular etiology, and extended lesions are related to poor prognosis (Kertesz and McCabe, 1977; Lazar et al., 2008; Maas et al., 2012; Pedersen et al., 1995, 2004; Reineck et al., 2005).

Observed recovery depends also on the specific function or aphasic syndrome studied. For example, in the case of patients with nonfluent aphasia due to left frontal lesions, clinical practice indicates that recovery of motor abilities usually precedes that of speech output. But even when making comparisons between more similar aspects of brain function, such as two components of verbal communication, one can find variability. One study recruited a mixed sample of 74 fluent and nonfluent aphasic patients, in order to investigate recovery of expressive and receptive abilities. During a period of 1 year, the patients were tested three times on several measures of spontaneous speech and sentence comprehension. There was no overall clinical improvement with regard to spontaneous speech. On the other hand, there was found considerable, significant improvement in sentence comprehension tests (Prins et al., 1978). It has been also demonstrated that different aspects of language,

such as phonology, semantics, and syntax components can be differentially affected (Hachioui et al., 2012), and follow distinct recovery patterns, with verbal communication being the latest to recover, during the first 6 months post -stroke (Hachioui et al., 2011, 2013; Smania et al., 2010). Another study by Laska et al. (2001) assessed the evolution aphasia subtypes from onset till 18 months post-stroke, as well as the recovery rate for each aphasic syndrome, by examining a group of aphasic patients in the acute stage, 3, 6, and 18 months after the occurrence of CVA. Recovery was observed for a great proportion of the whole aphasic sample (with one-fourth of the patients having completely recovered). Concerning evolution patterns for the different aphasia subtypes within the first 18 months post-stroke, most global aphasias evolved to mixed nonfluent, most Wernicke's aphasia to conduction, mixed, fluent, or transcortical sensory, and most patients initially diagnosed with conduction aphasia recovered completely. It should be, however, noted that the authors do not provide any information on whether patients received any kind of therapy or not, thus there is still a possibility of existing confounding factors with regard to outcome in the particular study. In other words, improvement cannot be certainly attributed to spontaneous or treatment-induced recovery, or even to a combination of both. Earlier research on the natural course of aphasia has also shown a gradual evolution reflected in taxonomic changes. Different patterns and rates of recovery among syndromes have also been observed (Sarno and Levita, 1979). Kertesz and McCabe (1977) studied 93 aphasic patients during the first year of recovery. While there were different patterns of recovery, i.e., patients with global aphasia showing in general rather poor prognosis in contrast with conduction, Broca's, and Wernicke's aphasics, the authors observed a general trend toward anomic aphasia, which was a common end-stage of evolution. Moreover, in accordance with Laska et al. (2001), their findings suggested that complete recovery occurred rather frequently among conduction aphasics, as well as patients with anomic and transcortical aphasia. A somewhat contradictive conclusion is drawn from the study of Nicholas et al. (1993), who performed consecutive aphasia assessments of 21 patients with 6-month intervals during a period of 2 years post-onset. Even though their findings indicated a significant improvement in communicative functions at least for up to 18 months post-stroke, the authors failed to find any changes with regard to aphasia classification in the aforementioned 2-year period. Further evidence suggesting stability of aphasia syndromes in the course of time is provided by Basso et al. (1982).

From a neurological perspective, a brain lesion causing various cognitive deficits should not be viewed as a stable and permanent condition, but as a snapshot of the ongoing recovery course. Longitudinal functional imaging studies show that brain reorganization is a highly dynamic process – similar to learning processes in healthy individuals – based on the emergence of new connections between the remaining modules of the shattered system (Rijntjes, 2006). In the early stages of the acute phase, there are two phenomena that play an important role in recovery: penumbra and diaschisis. Penumbra refers to underirrigation of a perilesional region, the penumbra zone, where neural cells are still partially functioning. It tends to diminish within the hyperacute phase, either spontaneously or aided by pharmaceutical treatment (Potagas et al., 2013; Ginsberg, 2003). Diaschisis is a notion originally introduced by von Monakow (1914), and refers to the supposed loss or decrease of function in the early period in structurally unaffected brain regions connected to the damaged area, in ipsilateral and contralateral areas (see also Feeney and Baron, 1986; Finger et al., 2004). Regression of these two phenomena is related to clinical recovery in the early period after stroke. A recent study by Saur et al. (2006) showed that language recovery in aphasia occurs in three stages: (1) reduced activation of the remaining left perisylvian regions in the acute phase, (2) increased activation of homologue right hemisphere regions, and (3) normalization of activation. However, it is suggested that reorganization processes are differentiated on the basis of lesion localization (Rijntjes, 2006) and type of impairment, with speech recovery depending on gradual left hemisphere activation changes and auditory comprehension recovery depending on both left and right temporal lobe activation (Price and Crinion, 2005). Few studies thus far have combined functional Magnetic Resonance Imaging (fMRI) and Diffusion Tensor Imaging (DTI) data to investigate the reorganization of brain mechanisms supporting language functions in the context of aphasia recovery. While such studies are limited to case reports (e.g., Heller et al., 2005), the necessity of longitudinal group studies is stressed by many authors. Crinion and Leff (2007) argue that such studies with larger patient samples and more detailed behavioral and lesion characterization are essential for the clinical diagnosis and treatment of aphasia, in the sense that they could elucidate the relationship between neuronal reorganization and recovery in chronic aphasic individuals.

Approaches to Aphasia Therapy

The interest on aphasia therapy grew rather slowly and reached a climax several decades after the seminal publications by Broca (1861, 1865), Wernicke (1874), and Lichtheim (1885), in which language disturbances were originally described in detail and defined as aphasic syndromes. Little attention was paid on rehabilitation during the first reign of the associationist theory, in the second half of the nineteenth century. However, after World War I several clinicians started working on possible treatment methodologies, which were mainly based on the idea that patients with aphasia had to relearn language abilities that have been lost, usually through repetition and drilling tasks. Aphasia rehabilitation started receiving more attention after World War II, due to the great numbers of rather young soldiers who suffered language disturbances after gunshot wounds. During this period the need for therapy with regard to older stroke patients with aphasia became more evident. The issue of aphasia was gaining popularity; the media started providing the public with relevant information, the request for rehabilitation increased, and several intervention techniques were devised (for a historical overview on the matter, see Basso, 2003). Up until now, several approaches have emerged, which will be briefly described below. The different approaches developed in the twentieth century obviously reflect distinct perspectives on aphasia. When a theory for the biological substrate of language and the underlying pathophysiological

mechanisms of aphasia is already formulated, it is inevitable for the methods of assessment and treatment to be defined on the basis of the theoretical hypotheses.

It has been suggested that the foundations of the so-called *classic* approach (otherwise called 'stimulation' or 'stimulation-facilitation' approach) have been established by the ideas of the holistic school, according to which language is a complex indivisible cognitive function that cannot be divided in terms of cortical representation to discrete subserving brain modules (Basso, 2003). The main representative of the particular view is Hildred Schuell. According to her view, aphasia may be seen as a clinical condition determined by two factors. The first one is the severity of the aphasic disorder and the second is the accompanying deficits in other cognitive domains, such as perceptual and sensorimotor. It is thus the degree of aphasia severity that differs among patients, and the distinct syndromes that may be observed can be considered as an artifact due to the aforementioned comorbid deficits. In sum, there are not truly different aphasia subtypes, but different degrees of severity and various types of co-existing impairments (Schuell et al., 1962). The second assertion of the stimulation approach is that it's not the linguistic function per se that is lost in aphasia, rather than access to it. This notion is based on the voluntary/automatic dichotomy, which in turn is supported by the common observation that linguistic behavior of aphasic patients is far from constant, in the sense that even severely impaired patients can produce speech under certain situations. There are many anecdotal reports of severely nonfluent patients producing utterances in a rather effortless way, when in any kind of stress. The patient reported by Jackson (1878/1958), suddenly shouting 'fire!' is such a case. Other examples are patients swearing when irritated. The famous Monsieur Tan is said to fluently utter the curse '*Sacré nom de Dieu*.' There was also a patient in our clinic with a severe nonfluent aphasia due to which he could only communicate through single words. He had right hemiplegia, therefore writing was quite difficult for him (note he was right-handed). When asked to copy a sentence with his left hand, he eloquently replied cursing at one of the authors (DK), using a four-word utterance. In the context of stimulation therapy, since aphasia cannot be further divided in distinct syndromes, therapy could be unitary for all patients, aiming to enhance access to language components. Given that aural comprehension is a key feature of language processing, rehabilitation is facilitated via auditory stimulation (Duffy and Coelho, 2001). Repetition of the stimulus is essential, while the complexity and length of the verbal material used in intervention sessions must be adjusted to the patient's abilities (Schuell et al., 1964).

The *behavioral* approach is based on operant conditioning, a notion introduced by the pioneer behaviorist B.F. Skinner. According to the Behaviorism's view only external stimuli and observed behavior can be studied directly, while brain function cannot, thus being beyond the scope of psychological research. Skinner (1957) claimed that verbal behavior is not different from any other aspect of human behavior; therefore it is subject to the same rules and can be modified in a similar way. By applying the principles of operant conditioning in aphasia therapy, the therapist can formulate a so-called *programmed instruction*, i.e., a carefully designed intervention program in which both the therapist's and the patient's behaviors are predetermined, in order for the latter's verbal behavior to be essentially modified, via manipulation of the external stimuli. In order to achieve this modification, the therapist makes use of two techniques: fading and shaping. Fading refers to facilitation of a response that is presumed to exist in the patient's repertoire, via facilitation based on selected antecedent stimuli. When the patient becomes able to provide the particular response without any help, the facilitation is gradually faded away. Shaping refers to facilitation of an unavailable response by manipulation of similar existing responses (for more details, see Basso, 2003).

The *functional reorganization* approach was introduced by Luria in the mid-twentieth century (Luria, 1963, 1970; Luria et al., 1969). Luria believed that linguistic functions are supported by a neural network consisting of several brain regions, each of them related to different components of verbal processing. Consequently, the type of aphasia resulting from brain damage would be dependent on the lesion locus. In terms of rehabilitation, the specifics of the impaired linguistic component must be firstly identified. Then, an intervention program focusing at the lost ability is designed, where aid is provided by the therapist in order for the patient to achieve maximum performance on a related language task. The program consists of specific tests which must be strictly followed. The final goal is functional reorganization; other, previously non-language related neural networks to be involved in linguistic processing, usually through a shift from a lower/automatic to a higher/voluntary level. Similar theoretical underpinnings are related to the *processing* approach. This view is based on cognitive models created to explain normal linguistic functions. Such models argue that each language component reflects a complex cognitive function supported by a network of distinct modules, not necessarily linked to an exact site in the brain, that process different kinds of verbal information (Hillis, 2001). The first step for rehabilitation is to identify the deficit and the underlying impaired module. Intervention then progresses by focusing on remediation of the affected cognitive processes and/or compensation through the spared modules (Hillis, 2001, 2002).

The *pragmatic* approach (also called 'functional communication' approach) is concerned with the patient's general communicative ability. Instead of assessing the accuracy of verbal output, therapists that adopt this view are rather interested in the efficacy of the message per se. Language is defined as a means of verbal communication, always examined within the particular real-life context. Therefore, aphasia is a communication disorder, rather than just a language breakdown. On the basis of this alternative perspective, the final goal of therapy is to enhance the ability of the patient to convey meaning to others, by any means available. Several studies have pointed out that aphasic patients' general communication capacity often seems to be better preserved compared to specific linguistic functions (Foldi et al., 1983; Kadzielawa et al., 1981; Wilcox et al., 1978). This is also evident in everyday clinical practice. Based on our experience, when not formally tested, aphasic patients can communicate more effectively, expressing feelings and ideas by using verbal along with nonverbal means, such as gestures. A remarkable

communicative capacity can be seen, for example, during self-help group sessions, where severely nonfluent patients engage in conversations, demonstrating a contrasting image with regard to their assessment-based clinical profile. It should be, however, noted that impairment-focused therapy can improve communication capacity in real-life circumstances (Carragher et al., 2012). The best-known and most widely used pragmatic approach is PACE (Promoting Aphasic's Communicative Effectiveness), developed by Davis and Wilcox (1985). In PACE-based intervention programs, the therapist aims to enhance the patient's communicative strategies, verbal or not. Every session must include conveyance of new information between the patient and the therapist, both of whom participate equally in this exchange, by sending and receiving messages. Patients are encouraged to use any means necessary, including words, drawing, writing, and gestures, in order to communicate. A step further within the context of the pragmatic approach would be to extend treatment, by including real-life communication in natural environment and involving partners in therapy (e.g., Kagan, 1998; Lubinski, 2001; Nykänen et al., 2013; Wilkinson et al., 2010). It should be, however, noted that the particular approach, although interesting, has been criticized on grounds of poor ecological value, scarce evidence in the literature, and limited clinical utility for the whole spectrum of aphasic disorders (Basso, 2003).

According to the *neurolinguistic* approach, language is governed by an internal system of organization rules. Post-stroke aphasic deficits are thus analyzed in the framework of a linguistic theory. The first comprehensive linguistic interpretation of aphasia was introduced by Jakobson (1964), who suggested two integral components of language processing: selection, which refers to the operation of selecting the appropriate word among alternatives on a similarity axis, and combination, which refers to the operation of combining the selected words along the contiguity axis. The model further enriched by other dichotomies and deficits in different axons correspond to specific aphasic phenomena (such as paraphasias or agrammatism) or even taxonomic categories. Several authors consider the linguistic theory to be essential for aphasia therapy (e.g., Hatfield, 1972; MacMahon, 1972). Given that a particular linguistic impairment is at the core of the treatment design, it has been claimed that a detailed analysis of distinct features of the aphasic syndrome allows the formulation of a highly specialized intervention program, strictly targeted at the specific deficit (Weniger et al., 1980). Therapists in favor of this approach often build their intervention strategies on the basis of linguistic principles, such as Chomsky's (2002) ideas of competence/performance dichotomy (e.g., Weigl, 1961; Weigl and Bierwisch, 1970) or transformational grammar (e.g., Naeser, 1975). Overall, the neurolinguistic approach has been extensively implemented through intervention strategies aiming at rehabilitating various syntactic (Helm-Estabrooks and Nicholas, 2000; Loverso et al., 1988; Naeser, 1975; Thompson, 2001; Weniger et al., 1980; Wilk and Paradis, 1993), phonemic (Naeser et al., 1986; Weniger et al., 1980; Wilk and Paradis, 1993), and lexical (Weniger et al., 1980; Wilk and Paradis, 1993) deficits. Nevertheless, the contribution of neurolinguistics in aphasia remediation remains under question, since the particular approach suffers many limitations (e.g., rather vague suggestions for founding therapy on linguistic principles; see also Basso, 2003).

There are several heterogeneous therapies that were developed in Boston, and were based on the neoassociationist theory. Basso (2003) groups them under the generic label *neoassociationist* or *neoclassic* approach, while acknowledging this heterogeneity among them. The theoretical background of these therapies is that language is subserved by a complex neural net that is composed of cortical faculties, each specialized for a particular linguistic function, interconnected by association with white matter fibers. Different types of aphasia are due to lesions affecting different parts of the perisylvian language network. Therefore, remediation techniques vary depending on the syndrome, and there are many therapy programs specially designed for specific taxonomic categories or discrete aphasic symptoms (Helm-Estabrooks and Albert, 1991). In general, the therapist focuses on the level of the observed disorder (i.e., whether it is phonemic, lexical, syntactic, semantic, etc.), and intervention is heavily based on stimulation. Among the various types of treatment within the spectrum of neoclassic theory, one can find several different views on language function and aphasia rehabilitiation. For example, according to the Language-Oriented Treatment (Bandur and Shewan, 2001; Shewan and Bandur, 1986), language can be separated in distinct modalities, which in turn are further divided in mutually exclusive areas. The areas are set in order of increasing difficulty and then the therapist attempts to remediate a specific modality, by using cues hierarchically organized in terms of effectiveness (see, for example, Best et al., 2013). Melodic Intonation Therapy (Albert et al., 1973; Norton et al., 2009; Sparks, 2001; Sparks et al., 1974; Sparks and Holland, 1976) utilizes the right hemisphere's music processing abilities, by using intonation and rhythm in the context of a hierarchically structured program, in order to facilitate sentence production. The neoassociationist umbrella also includes many other treatment methodologies, designed specifically to rehabilitate particular symptoms or aspects of a defined syndrome, such as Treatment for Wernicke's Aphasia (Helm-Estabrooks and Albert, 1991), Treatment of Aphasic Perseveration (Helm-Estabrooks et al., 1987), Helm Elicited Language Program for Syntax Stimulation (Helm-Estabrooks, 1981; Helm-Estabrooks et al., 1981), and Voluntary Control of Involuntary Utterances (Helm-Estabrooks and Albert, 2004; Helm and Barresi, 1980).

There are always new methods suggested for aphasia rehabilitation, more or less invasive. A rather new one is Constrained-Induced Therapy, originally developed by Taub (Taub, 1980; Taub et al., 1998, 1993, 1999) for remediating motor deficits, and used in aphasia rehabilitation for about a decade (Cherney et al., 2008; Meinzer et al., 2012). This view stresses the importance of forcing the patient to use his/her impaired ability to convey a message, excluding other routes of communication (see, for example, studies by Kirmess and Maher, 2010; Pulvermüller et al., 2001; Szaflarski et al., 2008). More invasive therapies have also been implemented in aphasia. Pharmacotherapy, for example, has provided rather promising (even though preliminary) evidence that, when combined with noninvasive

remediation methods, can be beneficial with regard to post-stroke aphasia's outcome (Berthier et al., 2011; Cherney, 2010). Several dopaminergic, cholinergic, noradrenergic, and glutamatergic agents have been used in this field (Berthier et al., 2009, 2011; Gill and Leff, 2014; Moscowitch et al., 1991; Seniów et al., 2009; Small and Llano, 2009; Tanaka et al., 1997; Walker-Batson et al., 2001; Whiting et al., 2007). Overall, research on pharmaceutical treatment of aphasia has provided useful findings indicating improvement in a wide range of linguistic abilities, such as naming, articulation, comprehension, and phonological skills (Berthier et al., 2011). Another biological approach to language remediation is related to the use of Transcranial Magnetic Stimulation (TMS) in aphasia. Based on the presumption that post-stroke overactivity of the right hemisphere homologues may be maladaptive for recovery, TMS has been applied to right cortical areas, in order to suppress their excitability (Martin et al., 2004; Hamilton et al., 2011). Recent studies' preliminary results suggest that there may be a possible beneficial role of TMS application combined with traditional language therapy with regard to aphasia outcome (Barwood and Murdoch, 2013; Naeser et al., 2005a,b; Szaflarski et al., 2011). Alternative methods of cortical stimulation have also been used in this context (e.g., Cherney et al., 2012).

Efficacy of Aphasia Therapy

There are many remediation techniques implemented in aphasia therapy, that may refer to the basic approaches described above, or extend beyond them, including group therapy (Basso, 2003), psychosocial-driven intervention strategies (Code et al., 2003), or even therapy programs that utilize computers (Palmer et al., 2012, 2013), humor (Simmons-Mackie and Schultz, 2003), or dogs (LaFrance et al., 2007). However, the crucial question is whether aphasia therapy is beneficial or not. There is no clear-cut answer in the relevant literature. There are controversial views that argue in favor either of high efficacy or limited (or even naught) effectiveness of aphasia therapy (e.g., Kertesz and McCabe, 1977; but see also Basso, 2003 for a discussion). In any case, there is robust evidence that different kinds of therapy are beneficial at least for some patients. As mentioned above, new intervention methodologies, such as TMS and pharmacological augmentation have provided promising results, by showing post-therapy improvement of several linguistic skills, such as naming and comprehension (e.g., Barwood and Murdoch, 2013; Berthier et al., 2011; Gill and Leff, 2014; Martin et al., 2004; Naeser et al., 2005a,b; Walker-Batson et al., 2001; Whiting et al., 2007). But older, more conventional therapies have been proven to be rather efficient (Aftonomos et al., 1999; Basso et al., 2011; Nickels, 2002; Shewan and Kertesz, 1984; Szaflarski et al., 2008), even for chronic (Basso and Caporali, 2004; Basso and Macis, 2011; Marangolo et al., 2013; Moss and Nicholas, 2006; Pulvermüller and Berthier, 2008) and severe aphasia (Basso, 2010). Furthermore, there is evidence that therapy may induce neuroplasticity in acute and also in chronic stages (Marcotte et al., 2012, 2013; Meinzer et al., 2004; Schlaug et al., 2009). An overall view on aphasia therapy efficacy has been recently offered by Basso (2003). The author performs a thorough review of the available literature, including meta-analyses, group, and single-case studies and concludes that remediation of language disorders appear to be generally efficient. A final remark should be made on the time frame of an intervention program in relation to effectiveness and final outcome. With regard to the starting point of therapy, it has been suggested that remediation strategies should be implemented at the acute phase (Godecke et al., 2012; Worall et al., 2013). Furthermore, several authors stress the importance of increased intensity, which is considered to be extremely beneficial for rehabilitation (Basso, 2005; Basso and Caporali, 2001; Bhogal et al., 2003; Code and Petheram, 2011; Godecke et al., 2012; Meinzer et al., 2004; Pulvermüller and Berthier, 2008; Sage et al., 2011; but see also Cherney et al., 2011; Marshall, 2008).

See also: Aphasia; Brain Asymmetry; Brain Damage: Neuropsychological Rehabilitation; Caregiver Burden; Cerebral Cortex; Classical Tests for Speech and Language Disorders; Disability: Social and Psychological Aspects; Functional Brain Imaging of Language processes; Jackson, John Hughlings (1835–1911); Language Processing, Functional Magnetic Resonance Imaging of; Lateralization of Language as Demonstrated by Brain Imaging Procedures; Neural Plasticity; Proper and Common Names: Impairments of Anomia; Speech Processing, The Cortical Organization of; Speech Production, Neural Basis of.

Bibliography

Aftonomos, L.B., Appelbaum, J.S., Steele, R.D., 1999. Improving outcomes for persons with aphasia in advanced community-based treatment programs. Stroke 30 (7), 1370–1379.

Albert, M.L., Sparks, R.W., Helm, N.A., 1973. Melodic intonation therapy for aphasia. Archives of Neurology 29 (2), 130.

Basso, A., 2003. Aphasia and Its Therapy. Oxford University Press, New York.

Basso, A., 2005. How intensive/prolonged should an intensive/prolonged treatment be? Aphasiology 19 (10–11), 975–984.

Basso, A., 2010. "Natural" conversation: a treatment for severe aphasia. Aphasiology 24 (4), 466–479.

Basso, A., Capitani, E., Zanobio, M.E., 1982. Pattern of recovery of oral and written expression and comprehension in aphasic patients. Behavioural Brain Research 6 (2), 115–128.

Basso, A., Caporali, A., 2001. Aphasia therapy or the importance of being earnest. Aphasiology 15 (4), 307–332.

Basso, A., Caporali, A., 2004. Targeted intervention for multiple language disorders: a case study. Journal of Neurolinguistics 17 (5), 383–401.

Basso, A., Cattaneo, S., Girelli, L., Luzzatti, C., Miozzo, A., Modena, L., Monti, A., 2011. Treatment efficacy of language and calculation disorders and speech apraxia: a review of the literature. European Journal of Physical and Rehabilitation Medicine 47 (1), 101–121.

Basso, A., Macis, M., 2011. Therapy efficacy in chronic aphasia. Behavioural Neurology 24 (4), 317–325.

Bandur, D.L., Shewan, C.M., 2001. Language oriented treatment: a psycholinguistic approach to aphasia. In: Chapey, R. (Ed.), Language Intervention Strategies in Aphasia and Related Neurogenic Communication Disorders. Lippincott Williams & Wilkins, New York, pp. 629–662.

Barwood, C.H.S., Murdoch, B.E., 2013. rTMS as a treatment for neurogenic communication and swallowing disorders. Acta Neurologica Scandinavica 127 (2), 77–91.

Berthier, M.L., Green, C., Lara, J.P., Higueras, C., Barbancho, M.A., Dávila, G., Pulvermüller, F., 2009. Memantine and constraint-induced aphasia therapy in chronic poststroke aphasia. Annals of neurology 65 (5), 577–585.

Berthier, M.L., Pulvermüller, F., Dávila, G., Casares, N.G., Gutiérrez, A., 2011. Drug therapy of post-stroke aphasia: a review of current evidence. Neuropsychology Review 21 (3), 302–317.

Best, W., Greenwood, A., Grassly, J., Herbert, R., Hickin, J., Howard, D., 2013. Aphasia rehabilitation: does generalisation from anomia therapy occur and is it predictable? A case series study. Cortex 49 (9), 2345–2357.

Bhogal, S.K., Teasell, R., Speechley, M., 2003. Intensity of aphasia therapy, impact on recovery. Stroke 34 (4), 987–993.

Breier, J.I., Juranek, J., Maher, L.M., Schmadeke, S., Men, D., Papanicolaou, A.C., 2009. Behavioral and neurophysiologic response to therapy for chronic aphasia. Archives of Physical Medicine and Rehabilitation 90 (12), 2026–2033.

Broca, P., 1861. Perte de la parole, ramollissement chronique de destruction partielle du lobe antırieur gauche du cerveau. Bulletins de la Societe d'Anthropologie de Paris 1861, 235–238.

Broca, P., 1865. Sur le siege de la faculte du language articule. Bulletins de la Societe Anthropologique de Paris 1865, 377–393.

Carragher, M., Conroy, P., Sage, K., Wilkinson, R., 2012. Can impairment-focused therapy change the everyday conversations of people with aphasia? A review of the literature and future directions. Aphasiology 26 (7), 895–916.

Cherney, L.R., 2010. Remediation of language disorders in adults. In: Whitaker, H. (Ed.), Concise Encyclopedia of Brain and Language. Elsevier, Oxford, pp. 451–454.

Cherney, L.R., Harvey, R.L., Babbitt, E.M., Hurwitz, R., Kaye, R.C., Lee, J.B., Small, S.L., 2012. Epidural cortical stimulation and aphasia therapy. Aphasiology 26 (9), 1192–1217.

Cherney, L.R., Patterson, J.P., Raymer, A.M., 2011. Intensity of aphasia therapy: evidence and efficacy. Current Neurology and Neuroscience Reports 11 (6), 560–569.

Cherney, L.R., Patterson, J.P., Raymer, A., Frymark, T., Schooling, T., 2008. Evidence-based systematic review: effects of intensity of treatment and constraint-induced language therapy for individuals with stroke-induced aphasia. Journal of Speech, Language, and Hearing Research 51, 1282–1299.

Chomsky, N., 2002. Syntactic Structures. Mouton, The Hague.

Code, C., Eales, C., Pearl, G., Conan, M., Cowin, K., Hickin, J., De Bleser, R., 2003. Supported self-help groups for aphasic people: development and research. In: Papathanasiou, I., De Bleser, R. (Eds.), The Sciences of Aphasia: From Therapy to Theory. Pergamon, Amsterdam, pp. 188–200.

Code, C., Petheram, B., 2011. Delivering for aphasia. International Journal of Speech-Language Pathology 13 (1), 3–10.

Crinion, J.T., Leff, A.P., 2007. Recovery and treatment of aphasia after stroke: functional imaging studies. Current Opinion in Neurology 20 (6), 667–673.

Davis, G., Wilcox, M., 1985. Adult Aphasia Rehabilitation: Applied Pragmatics. NFER-Nelson, Windsor.

Demeurisse, G., Demol, O., Derouck, M., De Beuckelaer, R., Coekaerts, M.J., Capon, A., 1980. Quantitative study of the rate of recovery from aphasia due to ischemic stroke. Stroke 11 (5), 455–458.

Duffy, J.R., Coelho, C.A., 2001. Schuell's stimulation approach to rehabilitation. In: Chapey, R. (Ed.), Language Intervention Strategies in Aphasia and Related Neurogenic Communication Disorders. Lippincott Williams & Wilkins, New York, pp. 341–382.

Engelter, S.T., Gostynski, M., Papa, S., Frei, M., Born, C.C., Ajdacic-Gross, V., Lyrer, P.A., 2004. Prevalence and severity of aphasia due to first ischemic stroke: a prospective population based study. Stroke 35 (6), E195.

Feeney, D.M., Baron, J.C., 1986. Diaschisis. Stroke 17 (5), 817–830.

Finger, S., Koehler, P.J., Jagella, C., 2004. The Monakow concept of diaschisis: origins and perspectives. Archives of Neurology 61 (2), 283–288.

Foldi, N.S., Cicone, M., Gardner, H., 1983. Pragmatic aspects of communication in brain-damaged patients. In: Segalowitz, S. (Ed.), Language Functions and Brain Organization. Academic Press, New York, pp. 51–86.

Gill, S.K., Leff, A.P., 2014. Dopaminergic therapy in aphasia. Aphasiology 28 (2), 155–170.

Ginsberg, M.D., 2003. Adventures in the pathophysiology of brain ischemia: penumbra, gene expression, neuroprotection. Stroke 34 (1), 214–223.

Godecke, E., Hird, K., Lalor, E.E., Rai, T., Phillips, M.R., 2012. Very early poststroke aphasia therapy: a pilot randomized controlled efficacy trial. International Journal of Stroke 7 (8), 635–644.

Gottesman, R.F., Hillis, A.E., 2010. Predictors and assessment of cognitive dysfunction resulting from ischaemic stroke. The Lancet Neurology 9 (9), 895–905.

Hachioui, H.E., Lingsma, H.F., van de Sandt-Koenderman, M.E., Dippel, D.W., Koudstaal, P.J., Visch-Brink, E.G., 2013. Recovery of aphasia after stroke: a 1-year follow-up study. Journal of Neurology 260 (1), 166–171.

Hachioui, E.H., van de Sandt-Koenderman, M.W., Dippel, D.W., Koudstaal, P.J., Visch-Brink, E.G., 2011. A 3-year evolution of linguistic disorders in aphasia after stroke. International Journal of Rehabilitation Research 34 (3), 215–221.

Hachioui, H.E., van de Sandt-Koenderman, M.W., Dippel, D.W., Koudstaal, P.J., Visch-Brink, E.G., 2012. The ScreeLing: occurrence of linguistic deficits in acute aphasia post-stroke. Journal of Rehabilitation Medicine 44 (5), 429–435.

Hamilton, R.H., Chrysikou, E.G., Coslett, B., 2011. Mechanisms of aphasia recovery after stroke and the role of noninvasive brain stimulation. Brain and Language 118 (1), 40–50.

Hatfield, F.M., 1972. Looking for help from linguistics. International Journal of Language & Communication Disorders 7 (1), 64–81.

Heller, S.L., Heier, L.A., Watts, R., Schwartz, T.H., Zelenko, N., Doyle, W., Devinsky, O., 2005. Evidence of cerebral reorganization following perinatal stroke demonstrated with fMRI and DTI tractography. Clinical Imaging 29 (4), 283–287.

Helm, N.A., Barresi, B., 1980. Voluntary control of involuntary utterances: a treatment approach for severe aphasia. Clinical Aphasiology 10, 308–315.

Helm-Estabrooks, N., 1981. Helm Elicited Language Program for Syntax Stimulation. PRo-ed, Austin, Texas.

Helm-Estabrooks, N., Albert, M.L., 1991. Manual of Aphasia Therapy. Pro-ed, Austin, TX.

Helm-Estabrooks, N., Albert, M.L., 2004. Voluntary control of involuntary utterances. In: Helm-Estabrooks, N., Albert, M.L. (Eds.), Manual of Aphasia and Aphasia Therapy. Pro-ed, Austin, TX, pp. 193–200.

Helm-Estabrooks, N., Emery, P., Albert, M.L., 1987. Treatment of aphasic perseveration (TAP) program: a new approach to aphasia therapy. Archives of Neurology 44 (12), 1253.

Helm-Estabrooks, N., Fitzpatrick, P.M., Barresi, B., 1981. Response of an agrammatic patient to a syntax stimulation program for aphasia. Journal of Speech and Hearing disorders 46 (4), 422–427.

Helm-Estabrooks, N., Nicholas, M., 2000. Sentence Production Program for Aphasia. Pro-Ed, Austin, TX.

Hillis, A.E., 2001. Cognitive neuropsychological approaches to rehabilitation of language disorders. In: Chapey, R. (Ed.), Language Intervention Strategies in Aphasia and Related Neurogenic Communication Disorders. Lippincott Williams & Wilkins, New York, pp. 513–523.

Hillis, A.E., 2002. The Handbook of Adult Language Disorders: Integrating Cognitive Neuropsychology, Neurology, and Rehabilitation. Psychology Press, New York.

Jakobson, R., 1964. Towards a linguistic typology of aphasic impairments. In: De Reuck, A.V., O'Connor, M. (Eds.), Disorders of Language. Churchill Livingstone, London, pp. 21–42.

Kadzielawa, D., Dabrowska, A., Nowakowska, M.T., Seniow, J., 1981. Literal and conveyed meaning as interpreted by aphasics and non-aphasics. Polish Psychological Bulletin 12 (1), 57–62.

Kagan, A., 1998. Supported conversation for adults with aphasia: methods and resources for training conversation partners. Aphasiology 12, 816–830.

Kertesz, A., McCabe, P., 1977. Recovery patterns and prognosis in aphasia. Brain: A Journal of Neurology 100, 1–18.

Kirmess, M., Maher, L.M., 2010. Constraint induced language therapy in early aphasia rehabilitation. Aphasiology 24 (6–8), 725–736.

LaFrance, C., Garcia, L.J., Labreche, J., 2007. The effect of a therapy dog on the communication skills of an adult with aphasia. Journal of Communication Disorders 40 (3), 215–224.

Laska, A.C., Hellblom, A., Murray, V., Kahan, T., Von Arbin, M., 2001. Aphasia in acute stroke and relation to outcome. Journal of Internal Medicine 249 (5), 413–422.

Lazar, R.M., Minzer, B., Antoniello, D., Festa, J.R., Krakauer, J.W., Marshall, R.S., 2010. Improvement in aphasia scores after stroke is well predicted by initial severity. Stroke 41 (7), 1485–1488.

Lazar, R.M., Speizer, A.E., Festa, J.R., Krakauer, J.W., Marshall, R.S., 2008. Variability in language recovery after first-time stroke. Journal of Neurology, Neurosurgery & Psychiatry 79 (5), 530–534.

Lecours, A.R., Chain, F., Poncet, M., Nespoulous, J.-L., Joanette, Y., 1992. Paris 1908: the hot summer of aphasiology or a season in the life of a chair. Brain and Language 42, 105–152.

Lee, J.B., Cherney, L.R., 2008. The changing "face" of aphasia therapy. SIG 2 Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders 18 (1), 15–23.

Lichtheim, L., 1885. On aphasia. Brain 7, 433–485.

Loverso, F.L., Prescott, T.E., Selinger, M., 1988. Cueing verbs: a treatment strategy for aphasic adults (CVT). Journal of Rehabilitation Research and Development 25 (2), 47–60.

Lubinski, R., 2001. Environmental systems approach in adult aphasia. In: Chapey, R. (Ed.), Language Intervention Strategies in Aphasia and Related Neurogenic Communication Disorders. Lippincott Williams & Wilkins, New York, pp. 269–296.

Luria, A.R., 1963. Restoration of Function after Brain Injury. Pergamon Press, Oxford.

Luria, A.R., 1970. Traumatic Aphasia. Mouton, The Hague.
Luria, A.R., Naydin, V.L., Tsvetkova, L.S., Vinarskaya, E.N., 1969. Restoration of higher cortical function following local brain damage. In: Vinken, P.J., Bruyn, G.W. (Eds.), Handbook of Clinical Neurology. North-Holland Publishing Company, Amsterdam, pp. 368–433.
Maas, M.B., Lev, M.H., Ay, H., Singhal, A.B., Greer, D.M., Smith, W.S., Harris, G.J., Halpern, E.F., Koroshetz, W.J., Furie, K.L., 2012. The prognosis for aphasia in stroke. Journal of Stroke and Cerebrovascular Diseases 21 (5), 350–357.
MacMahon, M.K.C., 1972. Modern linguistics and aphasia. International Journal of Language & Communication Disorders 7 (1), 54–63.
Marangolo, P., Fiori, V., Caltagirone, C., Marini, A., 2013. How conversational therapy influences language recovery in chronic non-fluent aphasia. Neuropsychological Rehabilitation 23 (5), 715–731.
Marcotte, K., Adrover-Roig, D., Damien, B., de Préaumont, M., Genereux, S., Hubert, M., Ansaldo, A.I., 2012. Therapy-induced neuroplasticity in chronic aphasia. Neuropsychologia 50 (8), 1776–1786.
Marcotte, K., Perlbarg, V., Marrelec, G., Benali, H., Ansaldo, A.I., 2013. Default-mode network functional connectivity in aphasia: therapy-induced neuroplasticity. Brain and Language 124 (1), 45–55.
Marshall, R.C., 2008. The impact of intensity of aphasia therapy on recovery. Stroke 39 (2), e48.
Martin, P.I., Naeser, A., MariaTormos, J., Nicholas, M., Kurland, J., Fregni, F., Seekins, H., Doron, K., Pascual-Leone, A., 2004. Transcranial magnetic stimulation as a complementary treatment for aphasia. Clinical Neurophysiology 15, 16.
Meinzer, M., Elbert, T., Wienbruch, C., Djundja, D., Barthel, G., Rockstroh, B., 2004. Intensive language training enhances brain plasticity in chronic aphasia. BioMed Central Biology 2 (1), 20.
Meinzer, M., Rodriguez, A.D., Rothi, L.J.G., 2012. The first decade of research on constrained-induced treatment approaches for aphasia rehabilitation. Archives of Physical Medicine and Rehabilitation 93 (1), S35–S45.
Moscowitch, L., McNamara, P., Albert, M.L., 1991. Neurochemical correlates of aphasia. Neurology 41 (1), 410.
Moss, A., Nicholas, M., 2006. Language rehabilitation in chronic aphasia and time postonset a review of single-subject data. Stroke 37 (12), 3043–3051.
von Monakow, C., 1914. Die Lokalisation im Grosshirn und der Abbau der Funktion durch kortikale Herde. J.F. Bergmann, Wiesbaden, Germany.
Naeser, M.A., 1975. A structured approach teaching aphasics basic sentence types. International Journal of Language & Communication Disorders 10 (1), 70–76.
Naeser, M.A., Haas, G., Mazurski, P., Laughlin, S., 1986. Sentence level auditory comprehension treatment program for aphasic adults. Archives of Physical Medicine and Rehabilitation 67 (6), 393–399.
Naeser, M.A., Martin, P.I., Nicholas, M., Baker, E.H., Seekins, H., Kobayashi, M., Theoret, H., Fregni, F., Maria-Tormos, J., Kurland, J., Doron, K.W., Pascual-Leone, A., 2005a. Improved picture naming in chronic aphasia after TMS to part of right Broca's area: an open-protocol study. Brain and Language 93 (1), 95–105.
Naeser, M.A., Martin, P.I., Nicholas, M., Baker, E.H., Seekins, H., Helm-Estabrooks, N., Cayer-Meade, C., Kobayashi, M., Theoret, H., Fregni, F., Maria-Tormos, J., Kurland, J., Doron, K.W., Pascual-Leone, A., 2005b. Improved naming after TMS treatments in a chronic, global aphasia patient–case report. Neurocase 11 (3), 182–193.
Nicholas, M.L., Helm-Estabrooks, N., Ward-Lonergan, J., Morgan, A.R., 1993. Evolution of severe aphasia in the first two years post onset. Archives of physical medicine and rehabilitation 74, 830–836.
Nickels, L., 2002. Therapy for naming disorders: revisiting, revising, and reviewing. Aphasiology 16 (10–11), 935–979.
Norton, A., Zipse, L., Marchina, S., Schlaug, G., 2009. Melodic intonation therapy: shared insights on how it is done and why it might help. Annals of the New York Academy of Sciences 1169 (1), 431–436.
Nykänen, A., Nyrkkö, H., Nykänen, M., Brunou, R., Rautakoski, P., 2013. Communication therapy for people with aphasia and their partners (APPUTE). Aphasiology 27 (10), 1159–1179.
Palmer, R., Enderby, P., Cooper, C., Latimer, N., Julious, S., Paterson, G., Dimairo, M., Dixon, S., Mortley, J., Hilton, R., Delaney, A., Hughes, H., 2012. Computer therapy compared with usual care for people with long-standing aphasia poststroke a pilot randomized controlled trial. Stroke 43 (7), 1904–1911.
Palmer, R., Enderby, P., Paterson, G., 2013. Using computers to enable self-management of aphasia therapy exercises for word finding: the patient and carer perspective. International Journal of Language & Communication Disorders 48 (5), 508–521.
Pedersen, P.M., Jørgensen, H.S., Nakayama, H., Raaschou, H.O., Olsen, T.S., 1995. Aphasia in acute stroke: incidence, determinants, and recovery. Annals of Neurology 38 (4), 659–666.
Pedersen, P., Vinter, K., Olsen, T.S.O.J., 2004. Aphasia after stroke: type, severity and prognosis. Cerebrovascular Diseases 17 (1), 35–43.
Pickersgill, M.J., Lincoln, N.B., 1983. Prognostic indicators and the pattern of recovery of communication in aphasic stroke patients. Journal of Neurology, Neurosurgery & Psychiatry 46 (2), 130–139.
Potagas, C., Kasselimis, D.S., Evdokimidis, I., 2013. Elements of neurology essential for understanding the aphasias. In: Papathanasiou, I., Coppens, P., Potagas, C. (Eds.), Aphasia and Related Communication Disorders. Jones & Bartlett Learning, Burlington, MA, pp. 23–46.
Price, C.J., Crinion, J., 2005. The latest on functional imaging studies of aphasic stroke. Current Opinion in Neurology 18 (4), 429–434.
Prins, R.S., Snow, C.E., Wagenaar, E., 1978. Recovery from aphasia: spontaneous speech versus language comprehension. Brain and Language 6 (2), 192–211.
Pulvermüller, F., Berthier, M.L., 2008. Aphasia therapy on a neuroscience basis. Aphasiology 22 (6), 563–599.
Pulvermüller, F., Neininger, B., Elbert, T., Mohr, B., Rockstroh, B., Koebbel, P., Taub, E., 2001. Constraint-induced therapy for chronic aphasia after stroke. Stroke 32, 1621–1626.
Reineck, L.A., Agarwal, S., Hillis, A.E., 2005. "Diffusion-clinical mismatch" is associated with potential for early recovery of aphasia. Neurology 64 (5), 828–833.
Richter, M., Miltner, W.H., Straube, T., 2008. Association between therapy outcome and right-hemispheric activation in chronic aphasia. Brain 131 (5), 1391–1401.
Rijntjes, M., 2006. Mechanisms of recovery in stroke patients with hemiparesis or aphasia: new insights, old questions and the meaning of therapies. Current Opinion in Neurology 19 (1), 76–83.
Sage, K., Snell, C., Lambon Ralph, M.A., 2011. How intensive does anomia therapy for people with aphasia need to be? Neuropsychological Rehabilitation 21 (1), 26–41.
Sarno, M.T., Levita, E., 1971. Natural course of recovery in severe aphasia. Archives of physical medicine and rehabilitation 52 (4), 175–178.
Sarno, M.T., Levita, E., 1979. Recovery in treated aphasia in the first year post-stroke. Stroke 10 (6), 663–670.
Saur, D., Lange, R., Baumgaertner, A., Schraknepper, V., Willmes, K., Rijntjes, M., Weiller, C., 2006. Dynamics of language reorganization after stroke. Brain 129 (6), 1371–1384.
Schlaug, G., Marchina, S., Norton, A., 2009. Evidence for plasticity in white-matter tracts of patients with chronic Broca's aphasia undergoing intense intonation-based speech therapy. Annals of the New York Academy of Sciences 1169 (1), 385–394.
Schuell, H., Jenkins, J.J., Carroll, J.B., 1962. A factor analysis of the Minnesota test for differential diagnosis of aphasia. Journal of Speech and Hearing Research 5 (4), 349.
Schuell, H., Jenkins, J.J., Jimenez-Pabon, E., 1964. Aphasia in Adults. Harper & Row, New York.
Seniów, J., Litwin, M., Litwin, T., Leśniak, M., Członkowska, A., 2009. New approach to the rehabilitation of post-stroke focal cognitive syndrome: effect of levodopa combined with speech and language therapy on functional recovery from aphasia. Journal of the Neurological Sciences 283 (1), 214–218.
Shewan, C.M., Bandur, D.L., 1986. Treatment of Aphasia: A Language-Oriented Approach. Taylor & Francis, London.
Shewan, C.M., Kertesz, A., 1984. Effects of speech and language treatment on recovery from aphasia. Brain and language 23 (2), 272–299.
Simmons-Mackie, N., Schultz, M., 2003. The role of humour in therapy for aphasia. Aphasiology 17 (8), 751–766.
Skinner, B.F., 1957. Verbal Behavior. Appleton-Century-Crofts, New York.
Small, S.L., Llano, D.A., 2009. Biological approaches to aphasia treatment. Current Neurology and Neuroscience Reports 9 (6), 443–450.
Smania, N., Gandolfi, M., Aglioti, S.M., Girardi, P., Fiaschi, A., Girardi, F., 2010. How long is the recovery of global aphasia? Twenty-five years of follow-up in a patient with left hemisphere stroke. Neurorehabilitation and Neural Repair 24 (9), 871–875.
Sparks, R.W., 2001. Melodic intonation therapy. In: Chapey, R. (Ed.), Language Intervention Strategies in Aphasia and Related Neurogenic Communication Disorders. Lippincott Williams & Wilkins, New York, pp. 703–717.
Sparks, R., Helm, N., Albert, M., 1974. Aphasia rehabilitation resulting from melodic intonation therapy. Cortex 10 (4), 303–316.
Sparks, R.W., Holland, A.L., 1976. Method: melodic intonation therapy for aphasia. Journal of Speech and Hearing Disorders 41 (3), 287.
Stone, S.P., Patel, P., Greenwood, R.J., Halligan, P.W., 1992. Measuring visual neglect in acute stroke and predicting its recovery: the visual neglect recovery index. Journal of Neurology, Neurosurgery & Psychiatry 55 (6), 431–436.
Szaflarski, J.P., Ball, A.L., Grether, S., Al-fwaress, F., Griffith, N.M., Neils-Strunjas, J., Newmeyer, A., Reichhardt, R., 2008. Constraint-induced aphasia therapy stimulates language recovery in patients with chronic aphasia after ischemic stroke. Medical Science Monitor: International Medical Journal of Experimental and Clinical Research 14 (5), CR243–CR250.

Szaflarski, J.P., Vannest, J., Wu, S.W., DiFrancesco, M.W., Banks, C., Gilbert, D.L., 2011. Excitatory repetitive transcranial magnetic stimulation induces improvements in chronic post-stroke aphasia. Medical Science Monitor: International Medical Journal of Experimental and Clinical Research 17 (3), CR132–CR139.

Tanaka, Y., Miyazaki, M., Albert, M.L., 1997. Effects of increased cholinergic activity on naming in aphasia. The Lancet 350 (9071), 116–117.

Taub, E., 1980. Somatosensory deafferentation research with monkeys: implications for rehabilitation medicine. In: Ince, L.P. (Ed.), Behavioral Psychology in Rehabilitation Medicine: Clinical Applications. William & Wilkins, New York, pp. 371–401.

Taub, E., Crago, J.E., Uswatte, G., 1998. Constraint-induced movement therapy: a new approach to treatment in physical rehabilitation. Rehabilitation Psychology 43 (2), 152–170.

Taub, E., Miller, N.E., Novack, T.A., Cook 3rd, E.W., Fleming, W.C., Nepomuceno, C.S., Connell, J.S., Crago, J.E., 1993. Technique to improve chronic motor deficit after stroke. Archives of Physical Medicine and Rehabilitation 74 (4), 347–354.

Taub, E., Uswatte, G., Pidikiti, R., 1999. Constraint-induced movement therapy: a new family of techniques with broad application to physical rehabilitation: a clinical review. Journal of Rehabilitation Research & Development 36 (3), 237–251.

Tesak, J., Code, C., 2008. Milestones in the History of Aphasia: Theories and Protagonists. Psychology Press, New York.

Thompson, C.K., 2001. Treatment of underlying forms: a linguistic specific approach for sentence production deficits in agrammatic aphasia. In: Chapey, R. (Ed.), Language Intervention Strategies in Aphasia and Related Neurogenic Communication Disorders. Lippincott Williams & Wilkins, New York, pp. 605–628.

Vargha-Khadem, F., Carr, L.J., Isaacs, E., Brett, E., Adams, C., Mishkin, M., 1997. Onset of speech after left hemispherectomy in a nine-year-old boy. Brain 120 (1), 159–182.

Walker-Batson, D., Curtis, S., Natarajan, R., Ford, J., Dronkers, N., Salmeron, E., Lai, J., Unwin, D.H., 2001. A double-blind, placebo-controlled study of the use of amphetamine in the treatment of aphasia. Stroke 32 (9), 2093–2098.

Weigl, E., 1961. The phenomenon of temporary deblocking in aphasia. Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 14, 337–363.

Weigl, E., Bierwisch, M., 1970. Neuropsychology and linguistics: topics of common research. Foundations of Language 1–18.

Weiller, C., Isensee, C., Rijntjes, M., Huber, W., Müller, S., Bier, D., Dutschka, K., Woods, R.P., Noth, J., Diener, H.C., 1995. Recovery from Wernicke's aphasia: a positron emission tomographic study. Annals of Neurology 37 (6), 723–732.

Weniger, D., Huber, W., Stachowiak, F.J., Poeck, K., 1980. Treatment of aphasia on a linguistic basis. In: Sarno, M.T., Hook, O. (Eds.), Aphasia: Assessment and Treatment. Almquist and Wiksell, Stockholm, Sweden, pp. 149–157.

Wernicke, C., 1874. Der Aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis. Cohn & Weigert, Breslau.

Whiting, E., Chenery, H.J., Chalk, J., Copland, D.A., 2007. Dexamphetamine boosts naming treatment effects in chronic aphasia. Journal of the International Neuropsychological Society 13 (06), 972–979.

Wilcox, M.J., Davis, G.A., Leonard, L.B., 1978. Aphasics' comprehension of contextually conveyed meaning. Brain and Language 6 (3), 362–377.

Wilk, J.J., Paradis, M., 1993. Linguistic foundations of rehabilitation methods. In: Paradis, M. (Ed.), Foundations of Aphasia Rehabilitation. Pergamon Press, Oxford, pp. 101–193.

Wilkinson, R., Bryan, K., Lock, S., Sage, K., 2010. Implementing and evaluating aphasia therapy targeted at couples' conversations: a single case study. Aphasiology 24 (6–8), 869–886.

Worall, L., Papathanasiou, I., Sherratt, S., 2013. Therapy approaches to aphasia. In: Papathanasiou, I., Coppens, P., Potagas, C. (Eds.), Aphasia and Related Communication Disorders. Jones & Bartlett Learning, Burlington, MA, pp. 93–111.

Language Families, Archaeology and History of

Peter Bellwood, Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The relationship between the archeological and linguistic records of human activity is very broad, extending from modern human language origins amongst our Paleolithic forebears onward into the usage of 'texts' (both written and oral) to assist interpretation of the archeological record of recent human history. This article, however, will be more specific, and is concerned essentially with late prehistory, prior to the development of written records, and with the question of how the world's major language families have originated and spread. It is also concerned with how linguistic information on this topic can be integrated with the archeological record of human expansion and migration during the past 10 000 years. This time span is chosen because extant language families shed little light on human affairs during the Pleistocene (2.6 million to 12 000 years ago), given that languages change constantly in ways that have slowly but inevitably erased traces of once-existing Pleistocene linguistic landscapes.

This article is concerned with the warm and relatively wet post-glacial period of time known as the Holocene (10 000 BC to now), the period within which agriculture and unstoppable growth in the size of the human population have underpinned the world that we inhabit today, replete with its different languages, cultures, and biological populations. All extant language families, of hunters and farmers alike, trace their dispersal histories back to some point within this time span, to specific episodes of human migration and dispersal. Human societies and migrations, of course, existed and occurred during the Pleistocene, as did languages remotely ancestral to all existing language families. The Pleistocene can be investigated archeologically, but we cannot trace its languages in any direct way, except through comparative reconstructions of macrofamilies (e.g., Nostratic, Amerind, Dene-Caucasian – see Gell-Mann and Ruhlen, 2011). The results from using a macrofamily approach do not intersect easily with the archeological record, and this topic is not considered further here.

Language families of wide distribution across the world are shown in **Figures 1** and **2**. All existed, including Indo-European, to virtually their pre-AD 1500 limits before the recording of any actual history, whether it be from the Rigveda, Herodotus, Julius Caesar, or the British Empire. It is not possible to link the pre-AD 1500 distribution of any major language family to any episode of state formation in human history. All were created by the small-scale societies of prehistory, amongst societies of stone-using and sometimes metal-using farmers, pastoralists, and hunter-gatherers. Major language spreads in recent colonial history – English, Spanish, Russian, Chinese, and so forth – should not be confused with the origins of whole language families. Like these recent spreads, the major element behind the spread of whole language families in prehistory was always a movement of actual speakers of the ancestral languages, not purely language shift through secondary learning across linguistically unrelated communities. However, such migrant native speakers always mixed genetically with indigenous populations, when they existed. Some, such as the early Austronesians of the Pacific, were able to enter empty territories and thus colonize pristine landscapes.

Why do language families exist? Why are English and Bengali quite closely related, at least to a linguist, even if such relationship today would be totally puzzling to monolingual native speakers of those two languages? Johann Reinholt Forster and William Jones both knew the answer in the late eighteenth century (the former in 1774 while on Cook's second Pacific voyage for Austronesian, the latter in 1786 for Indo-European). As Jones put it so famously: "… no philologer could examine them all three [Sanskrit, Greek and Latin], without believing them to have sprung from some common source, which, perhaps, no longer exists" (Jones, 1993: 175). Reconstructing the common source is therefore essential if we are to broach the "why do language families exist" question.

Today, language history and the archeological record can be studied in combination to recover history at two major (but clearly overlapping) levels:

1. at the level of the individual language, ethnolinguistic group, or historical community; and,
2. at the level of the language family, or major internal subgroup within a family.

With regard to (1), it is very difficult to trace the identity into deep levels of prehistory of a specific ethnolinguistic or historical population (e.g., Celts, Greeks, Etruscans, Kurds, Jews), unless one is dealing with a very isolated region or an island where one can assume there has been no substantial population replacement or mixing during the period in question. Good examples of the latter would be certain Pacific islands, for example Easter Island or New Zealand, both fairly isolated since their first human settlements less than 1000 years ago by Polynesians (Kirch and Green, 2001). However, this article is not primarily concerned with such society- or culture-specific correlations.

Instead, it focuses on the study of languages as members of genetically constituted and evolving families. The member languages within a language family have inherited a common set of retentions from a so-called 'protolanguage' that can be reconstructed in part through the methodology of comparative linguistics. Such protolanguages, and their reconstructed sets of

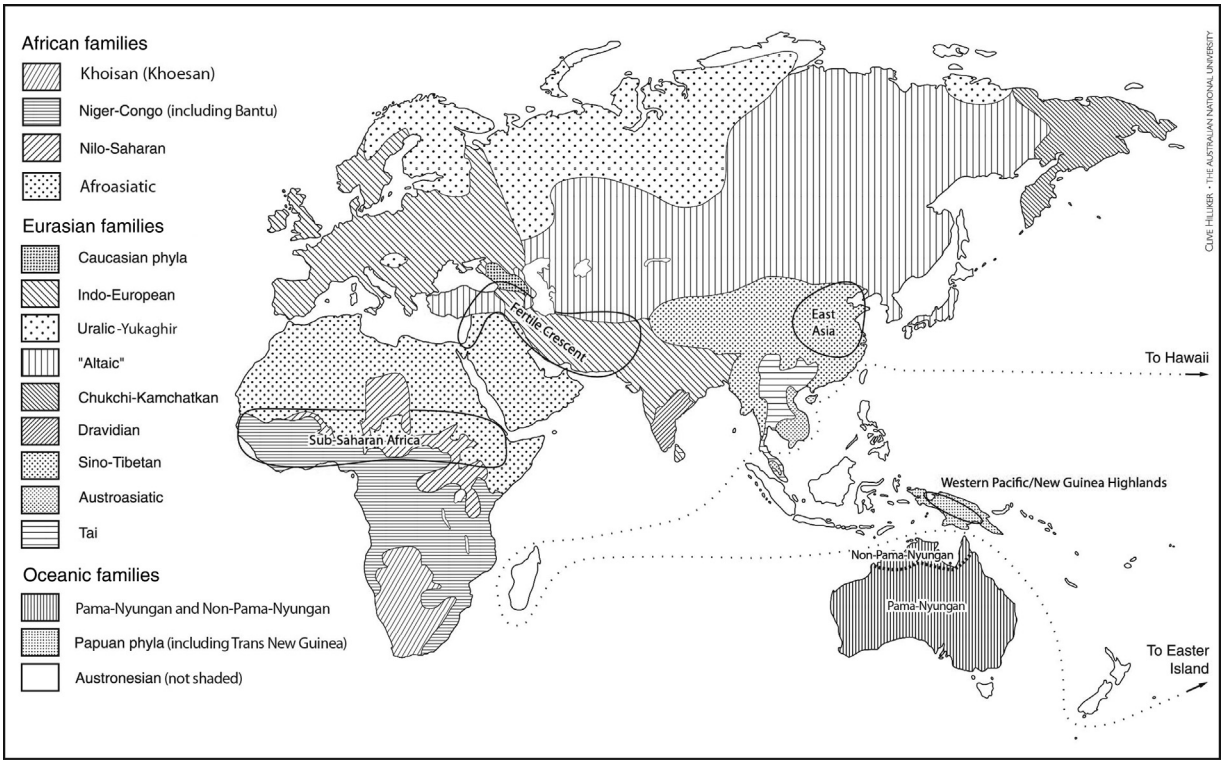

Figure 1 The major language families of the Old World, overlaying the most significant regions of early agricultural development. The Fertile Crescent is suggested as the homeland region for Indo-European and Afroasiatic; Sub-Saharan Africa for Nilo-Saharan pastoralism and Bantu; East Asia for Sino-Tibetan, "Altaic" (including Turkic), Sino-Tibetan, Tai, Austroasiatic, Hmong-Mien and Austronesian; New Guinea Highlands for Trans New Guinea. The Pama-Nyungan languages of Australia are also shown. Original map by Clive Hilliker, ANU.

lexical meanings for items of material culture, can sometimes be associated with well-defined and well-dated archeological traditions as they spread, evolved and interacted through time and space. It is necessary to avoid circular reasoning here, but because the archeological and linguistic records of prehistoric human affairs are generated from completely independent data sets (languages on the one hand, the archeological record on the other), it is possible to compare reconstructions from both disciplines carefully in order to generate testable and convincing 'best explanations' (Fogelin, 2007).

Historical reconstruction at this level tends to be organized such that language families (e.g., Indo-European, Austronesian) are foregrounded as the major foci of enquiry, rather than archeological complexes. This is because language families are usually more sharply defined and reveal clearer patterns of phylogenetic inheritance than do archeological complexes (*see* Cultural Evolution: Phylogeny versus Reticulation). In such situations, archeology tends to be used to support or refute historical linguistic questions. For instance, where was the Indo-European homeland located, what lifestyle did its inhabitants enjoy, and when? We examine this question in more detail later. In addition, however, some archeological complexes of particularly wide distribution, internal homogeneity, and short time span, such as the Linearbandkeramik (LBK) early Neolithic of Central Europe (*c*.5400–5000 BC) and the Lapita cultural complex of the western Pacific (*c*.1300–800 BC), are also sometimes foregrounded as requiring a paleolinguistic identity. For instance, did the LBK correlate with early Indo-European linguistic dispersal through central Europe? Did Lapita correlate with Austronesian (Oceanic subgroup) linguistic dispersal through Melanesia into Polynesia?

To understand how the data of historical linguistics and archeology might be compared against each other to improve understanding of the human past, it is first necessary to state clearly the abilities and limitations of the two disciplines.

Language as a Source of Information on Human Prehistory

The branch of linguistics which is of most interest to prehistorians is that known as comparative historical linguistics, in which the structures and vocabularies of present-day or historically recorded languages are compared to identify families, and subgroups within these families. The methodology of comparative linguistic reconstruction is precise. Like the methodology of cladistics, as applied in biology, its main goal at the within-family level is to identify shared innovations which can identify language subgroups. Such subgroups comprise languages which have shared a common ancestry, separate from other languages with which they are more distantly related. Languages which comprise a subgroup share descent from a common 'protolanguage,' this having been in many cases a chain of related dialects. The protolanguages of subgroups within a family can sometimes be organized into

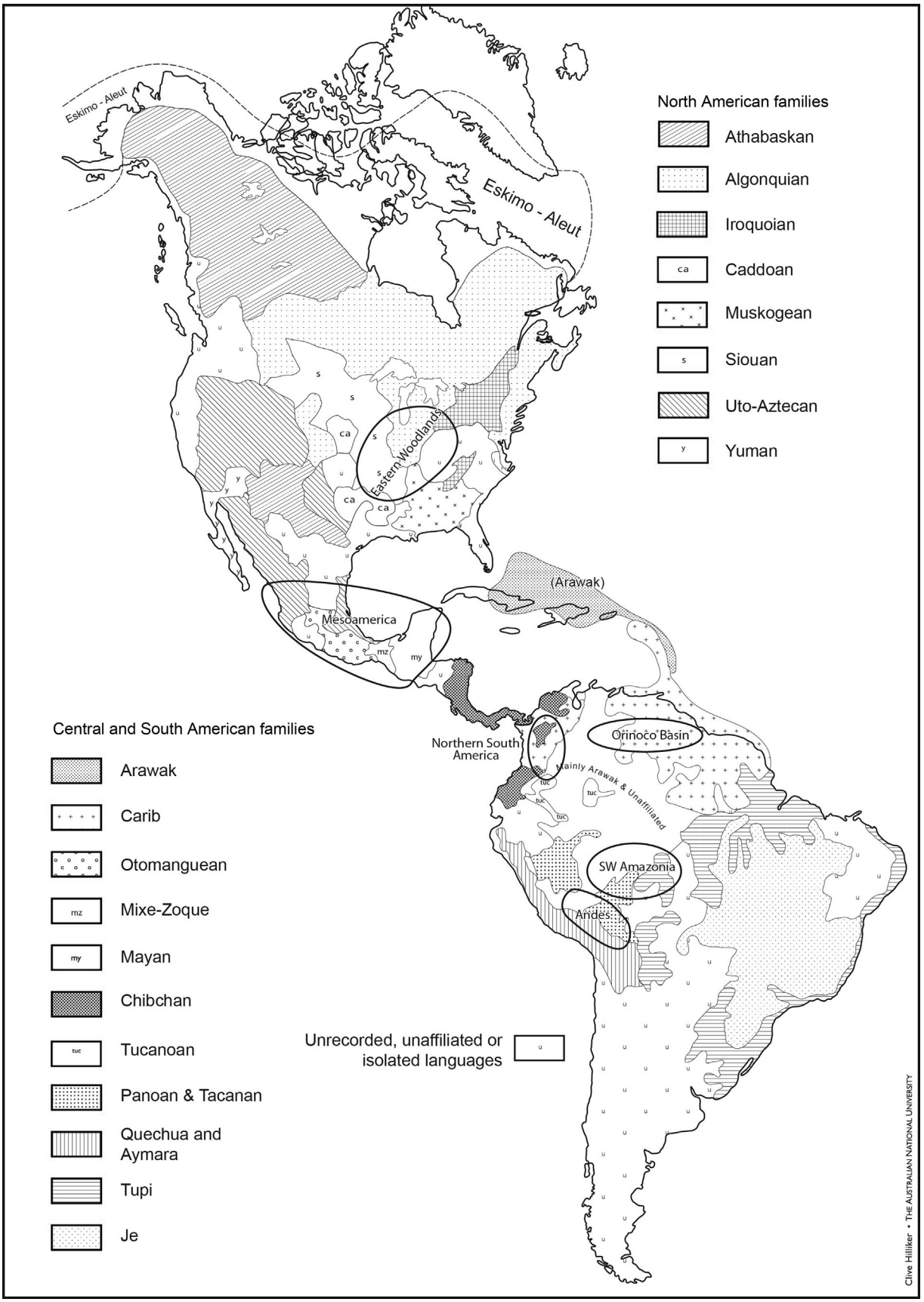

Figure 2 The major language families of the New World, overlaying the most significant regions of early agricultural development (Piperno, D., 2011. Plant cultivation and domestication in the New World tropics. Current Anthropology 52 (Suppl. 4), S453–S470). The Eastern Woodlands are suggested as the homeland for the Siouan and Iroquoian language families; Mesoamerica for Uto-Aztecan, Otomanguean, Mixe-Zoque, and Mayan; the Andes for Quechua and Aymara; southwest Amazonia for Tupi and Panoan; the Orinoco Basin for Carib; and northern South America possibly for Chibchan (and Arawak?). Original map by Clive Hilliker, ANU.

a family tree of successive linguistic differentiations (not always sharp splits, unlike real tree branches), and for some families it is possible to postulate a relative chronological order of subgroup formation. For instance, many linguists believe that the separation between the Anatolian languages (including Hittite) and the rest of Indo-European represents the first identifiable differentiation in the history of that family. Likewise the separation between the Formosan (Taiwan) languages and the rest of Austronesian (Malayo-Polynesian) represents the first identifiable differentiation within Austronesian.

The vocabularies of reconstructed protolanguages (e.g., Proto-Indo-European, Proto-Austronesian) can sometimes provide remarkable details on the locations and lifestyles of ancient ancestral communities, with many hundreds of ancestral terms and their associated meanings reconstructable in some instances. There have also been many attempts to date language subgroup history by a variety of methods based on lexical comparison, including glottochronology, Bayesian phylolinguistics, and the analysis of Levenshtein lexical distances (see discussions in Renfrew et al., 2000; Greenhill and Gray, 2012; Holman et al., 2011). Such methods always require calibration against real time using historical records, and can be affected by sociolinguistic situation and the intensity of ancient borrowing. Debates continue about their validity (Donohue et al., 2012), but there can be no doubt that the chronological answers they give often accord well with archeological dating, particularly for language families whose roots are suspected to correlate with the regional beginnings of agriculture (Bellwood, 2005).

The other major source of linguistic variation, apart from modification through descent, is that termed by linguists 'borrowing' or 'contact-induced change.' This operates between different languages, and often between languages in completely unrelated families. Borrowing, if identified at the protolanguage level, can be as much an indicator of the homeland of a language family or subgroup as can genetic structure. It can also reflect important contact events in language history.

Archeology as a Source of Information on Human Prehistory

Archeology is concerned mainly with the recovery and interpretation of the material remains of the human past, and the environmental contexts in which those remains were originally deposited. Such remains may be dated, and grouped into regional complexes of related components. Such complexes can then be compared with other complexes, and the boundaries between such complexes can be studied carefully. Some are sharply bounded, hence possible candidates for correlation with an ethnolinguistic group, while others are simply nodes of relative homogeneity in a kaleidoscope of ever-shifting patterning. Archeology alone cannot pinpoint ethnicity, unless of course it operates in an environment associated with literacy and the availability of written records. Even then, ambiguity can plague interpretation, as in the modern debate in British archeological literature about the definition and archeological history of the Celts (Collis, 2003). Any correlations between the archeological and linguistic records will always require care.

How Can Language Family History and Archeological Prehistory Be Correlated?

Correlation of the archeological and linguistic records is not always a simple matter. However, correlations can be made when language family distributions correspond with the distributions of delineated archeological complexes, particularly when the material culture and environmental vocabulary reconstructed for a given family at the protolanguage level correspond with material culture and its environmental correlates as derived from the archeological record. Many reconstructed protolanguages, for instance, have vocabularies which imply the existence of crucial categories within human economic and technological behavior, such as agriculture, domestic animals, pottery, and metallurgy, these all being identifiable in the archeological record.

The concept of the language family is sturdier than that of the archeological culture. This is important, because linguists have come to a remarkable level of agreement on the classification of the world's language families. Apart from a small number of creoles, mostly a result of recent historical European colonization and population translocation, the vast majority of the world's language families are clearly bounded in a classificatory sense and not beset with huge numbers of 'mixed' languages. As an example, the Indian subcontinent has been a region of interaction between the speakers of Indo-European and Dravidian languages for at least five millennia. Languages within these two families have borrowed extensively from each other, to the extent that the subcontinent is often referred to by linguists as a 'linguistic area' (a zone of widespread areal diffusion). Nevertheless, the subcontinent is not covered by languages which are half Indo-European, half Dravidian, chaotically mixed. This means that language families are coherent entities, capable of maintaining coherence and independence through long periods of time. As such, they are believed to carry traceable records of human history and to be associated, in their origins, with homeland regions and processes of population dispersal. Although some populations have changed their languages in the past, it is unlikely that language shift, as opposed to an actual dispersal of ancestral speakers of a protolanguage, could have been the main mode of dispersal of any major language family (Ostler, 2005). All of this means that the origins and dispersals of the protolanguages from which language families are created should correlate with major population movements, frequently on a scale which should be visible in the archeological record (Bellwood, 2008, 2009, 2010, 2013).

Some Examples of Language Family Origin and Dispersal Histories with Claimed Archeological Correlations

In the remainder of this article we examine some of the major language families of the world, and discuss recent literature pertaining to their origin locations and histories of dispersal. This literature is of course vast, and readers will find that much is reviewed in *The Encyclopedia of Global Human Migration* (Ness and Bellwood, 2013), and in Bellwood (2013). Many archeologists and linguists today recognize that some very

large language families could owe their initial creations to population dispersal as a result of population growth following on from the development of agriculture. If this is so, then the homelands of these particular families, to be discussed individually below, can be expected to overlap with the most important regions of early agriculture, as indeed seems to be the case for the Middle East, East Asia, and Mesoamerica (**Figure 1**). However, it is important to remember that many language families are associated totally with hunting and gathering populations, and presumably always have been, so their histories obviously will not involve this factor. Such hunter-gatherer families with recent histories of expansion include the Pama-Nyungan languages of Australia, Athabaskan in Canada and the US Southwest, and Eskimo-Aleut. In these cases, explanations for their expansion must also be based on factors of demographic growth and technological advantage, albeit not involving food production. Other families, such as Uralic, Algonquian, and Uto-Aztecan, have both agricultural and hunter-gatherer populations. In some of these cases it is possible that former agricultural peoples actually became hunters and gatherers in agronomically difficult environments. Examples include the southern Maoris of New Zealand and the Great Basin Uto-Aztecans (Hill, 2002; Bellwood, 2005).

As noted above, it is apparent that some languages and subgroups (but not whole language families) have been recorded as spreading over large distances in historical times, under conditions of statehood, religious evangelism, and colonialism. Thai in Thailand, the Chinese languages in southern China, Arabic in North Africa and the Middle East, and of course English and Spanish in their recent overseas colonies all come to mind here. In the twenty-first century, it is also apparent that lingua francas and national languages can spread rapidly as a result of educational policy, literacy, mass media, and sociolinguistic status. But it is more difficult to imagine such processes of language adoption as being of great significance amongst the small-scale societies of preurban prehistory. Nevertheless, many prehistorians have suggested that language replacement processes of this type, whereby people adopt a language deemed to be of high social status and abandon their original language, have been instrumental in the spread of some families. One such family is Indo-European, for which some linguists and archeologists have long agreed on a homeland in the steppes north of the Black Sea, followed by a spread into Europe by Late Neolithic and Bronze Age pastoral peoples with domesticated horses and wheeled transport (Anthony, 2007). According to this view, these people undertook their migrations into Europe between 4500 and 2500 BC, dominating and absorbing the older Neolithic societies in the process.

This view of Bronze Age conquest and language replacement for Indo-European dispersal has been challenged by the archeologist Colin Renfrew (1987, 1996), who opts instead for an association of early Indo-European languages with early Neolithic farming dispersal into Europe from Anatolia (Turkey). This view has been strengthened in recent years by new evidence for some quite dramatic population dispersal, associated with farming, commencing around 6500 BC from Anatolia into northern Greece and the Balkans (Özdoğan, 2013; and *see* Phylogeny and Reticulation in Prehistory for further discussion of relevant archeology and genetics). Farmers typically have larger populations than hunter-gatherers, and if farmers are not enclosed by other farming populations, or if they live in a region surrounded essentially by low-density hunter-gatherers, then expansion is a likely outcome, exactly as in the recent European colonial frontiers in Australia and western North America. So, while the agricultural dispersal hypothesis is no more 'provable' than many other hypotheses to explain language family origins, it does at least have the strong supporting factor of a historically demonstrable mechanism which can allow and encourage population expansion to occur. Such expansion need not mean extinction of all hunter-gatherers. In many ethnographic situations, hunter-gatherers have survived in the interstices of agricultural or pastoralist landscapes, perhaps for millennia.

Put simply, the farming dispersal hypothesis, as expressed in more detail elsewhere (Bellwood and Renfrew, 2002; Diamond and Bellwood, 2003; Bellwood, 2005, 2013), would see the protolanguages for Indo-European and Afroasiatic located in the wheat, barley, cattle, sheep, and goat early farming zone in the Middle East, with dispersals occurring, far to the east and west, mainly between 6500 and 3000 BC. Cyprus was actually colonized by boat from the Asian mainland as early as 9000 BC. By 4000 BC, Indo-European speakers had reached the British Isles. They penetrated much earlier toward Pakistan and Turkmenistan, commencing around 7000 BC, but only entered the upper Ganges basin around 3000 BC. Other ancient and extinct languages such as Sumerian and Elamite must also have existed, perhaps quite widely, in the Middle East well before 3000 BC, but these later became replaced by members of the more successful language families, especially the Semitic (Afroasiatic) and Iranian (Indo-European) languages. Some linguists have also suggested that Elamite was related to the existing Dravidian family of languages in southern India and Sri Lanka, which perhaps spread from northwestern regions of the subcontinent after the appearance there, around 3000 BC, of millet agriculture and cattle pastoralism.

Early Afroasiatic speaking populations carried agriculture, sheep, and goats from the Levant into northwestern Africa about 6500 BC, where they met existing populations of Nilo-Saharan speakers who had spread as hunter-gatherers through what is now the southern and eastern Sahara Desert during a wet climatic phase starting *c.* 9000 BC. These Nilo-Saharan speakers in turn adopted pastoralism, and some later moved southward into Kenya and East Africa. Further west, Niger-Congo expansion began also during the early Holocene wet phase in the southwestern Sahara, but Bantu subgroup speakers spread most dramatically over much of central and southern Africa after 1000 BC, in association with millet agriculture, cattle pastoralism, and the use of iron. This expansion led to considerable retraction of the original distribution of the large and much diversified group of languages termed Khoisan, such that they only survived amongst hunters and pastoralists in the southern parts of the continent when Europeans arrived. Bantu languages today dominate almost all of Sub-Saharan Africa, except for the far south.

Likewise, Afroasiatic and Indo-European languages dominate most regions of northern Africa and western Eurasia, except for those regions of northern Eurasia settled during the Holocene by speakers of Uralic languages, originally hunter-gatherers. Many of these expansions, of course, took place in historical circumstances and are very well recorded. But the basic fact remains that the overall pre-AD 1500 distributions of both Afroasiatic and Indo-European, and probably Uralic too (except for Finnish and Hungarian), were laid down long before the days of Cheops, Sargon of Agade, Alexander the Great, Jesus Christ, or Mohammed. None of these people, or the religions associated with some of them, can be highlighted as prime movers.

Moving across to the opposite side of Eurasia and following the same logic, the Sino-Tibetan, Austroasiatic, Austronesian, Tai, and Hmong-Mien families would all have begun their dispersal from the region of rice and millet cultivation in central and southern China, focused in the middle and lower Yellow and Yangzi valleys, between 5000 and 2000 BC. Austronesians eventually colonized the greater part of the Pacific after 1500 BC, after migrating through Island Southeast Asia via Taiwan and island-hopping around much of western Melanesia, particularly New Guinea, where numerous Papuan language families have maintained dominance to the present as a result of indigenous developments in food production. The languages of northern tropical Australia probably reflect a deep Pleistocene ancestry, but the Pama-Nyungan language family that covers most of the continent beyond the tropics had a much more recent origin, and perhaps reflected the impact of new population arrivals from India or Indonesia around 4000 years ago, bringing dogs (dingoes) and a heavy emphasis on a backed and microlithic stone tool technology.

Across the Pacific, in the Americas, the Mayan, Otomanguean, Mixe-Zoque, Uto-Aztecan, and Chibchan language families probably spread as a result of agricultural developments in Greater Mesoamerica after 3500 BC. In South America the picture is a little more diffuse, but some of the major Andean and Amazonian families might have spread as a result of the widespread establishment of maize and manioc agriculture after about 2000 BC. Examples here include Quechua and Aymara in the Andes, and lowland Amazonian families such as Arawak, Carib, and Tupi.

Archeologically, these suggested language radiations associated with early agricultural societies should be reflected in the distributions of some very widespread archeological complexes. In particular, it has been noted in many regions that the archeological complexes of early agricultural phases are much more widespread and homogeneous in content than the highly regionalized complexes of later periods. This appears to be the case in some parts of early Neolithic Europe, East Asia, and the Pacific, and amongst the early formative cultures of the Americas. There is simply not space in this article to refer to all of this archeological evidence.

Indeed, the author freely admits that the above reconstructions are not universally accepted, as discussed in more detail elsewhere (see Bibliography). It should also be noted that disagreements are not just between disciplines, but also occur within them. For example, not all linguists or archeologists believe that Austronesian languages spread by migration from southern China through Taiwan (e.g., Donohue and Denham, 2010; see reply in Bellwood, 2011). There is also another powerful source of information on human prehistory, one not really acknowledged in this article, and this is human genetics and skeletal morphometrics. As I sit in front of my computer writing this article, yet another genetics article crosses my desk, by Irina Pugach et al., (2013), suggesting that gene flow occurred between northern Australia and India around 4000 years ago. Needless to say, the authors draw the same archeological conclusion as do I, that new populations arrived in northern Australia with their genes, their dogs, and a different (but not entirely unique) stone technology (they do not consider linguistics). This movement, if verifiable, had nothing whatsoever to do with agriculture. But, like that of the Athabaskan and Eskimo-Aleut hunter-gatherers of the North American middle and late Holocene, it had everything to do with economy and technology, and perhaps also a degree of geographical retreat by the earlier inhabitants of the territories that these groups entered. In the US Southwest, for instance, Apachean (Athabaskan) speakers entered former Pueblo territories after AD 1300 because the former maize farming inhabitants retracted their distributions in the face of drought, warfare, overpopulation, or all in combination. The Inuit Eskimo were able to colonize coastal regions of Arctic, Canada, and Greenland within the past millennium because their Paleoeskimo predecessors had virtually disappeared from these territories.

Language Contact and Cultural Contact

Finally, language and archeology correlations can be sought not only for the origins and dispersal histories of language families, but they can also reflect the contacts which took place from time to time between languages in different families and subgroups. For instance, the Austronesian speakers in pockets of coastal New Guinea have been in intense contact for upward of 2500 years with the speakers of Papuan languages in several unrelated families. This has led to a great deal of contact-induced change and language shift, and it is therefore not surprising to discover that distributions of material culture often cross-cut language boundaries. The archeological record, however, suggests that quite strong differences in material culture would have distinguished Papuan and Austronesian societies 3000 years ago, at the time of the Lapita archeological spread through much of the western Pacific. The Lapita spread, with its distinctive stamped pottery, lithic and shell artifacts, was probably associated with the initial Austronesian colonization of many of the western Pacific Islands. But it is significant that it appears initially to have avoided the island of New Guinea itself, where Austronesian speakers even today are found only in a few pockets of coastal distribution (Kirch and Green, 2001; note that late Lapita sites do exist in southeastern New Guinea).

Linguist Robert Dixon (1997) once suggested that the overall history of the major language families of the world

can be separated into short periods of widespread dispersal, when the families were actually founded through migration, interspersed with long periods like that described for New Guinea, when populations of quite different linguistic and cultural origin interacted over the long term, whether peacefully or belligerently. This hypothesis resembles the theory of punctuated equilibrium as applied to the biological evolution of species.

In this regard, not all of the speakers of some of the major language families are of tightly defined and geographically restricted genetic origin. For instance, the speakers of Austronesian languages range from Southeast Asians to Melanesians and Polynesians. The speakers of Indo-European languages range from northern Europeans to northern Indians. The speakers of Afroasiatic languages range from Ethiopians to Palestinians and Israelis. It is possible, but rather unlikely, that these differences represent no more than natural selection operating since the initial population dispersal which founded the language family in question. It is far more likely that they reflect population mixing, not always paralleled by an equivalent amount of language mixing. In other words, language families can have a life of their own, as can nodes of biological variation. Many population dispersals must have incorporated large numbers of the existing inhabitants of the newly settled regions, with consequent genetic effects stamped on later generations. This does not mean that there can be no correlations between variations in language and biology within our single and unified human species, but we must be aware that any correlations will not always be clear-cut and obvious. They must be teased apart with care.

See also: Cultural Evolution: Phylogeny versus Reticulation; Evolution and Language: Cultural Transmission; Evolution and Language: Overview; Evolution and Language: Phylogenetic Analyses; First Language Acquisition, Developmental Psychology of; Global Migration; Human Language, Evolution of; Indo-European Languages; Language Contact; Linguistic Anthropology; Migrations, Colonizations, and Diasporas in Archaeology.

Bibliography

Anthony, D., 2007. The Horse, the Wheel, and Language. Princeton University Press.

Bellwood, P., 2005. First Farmers. Blackwell, Oxford and Malden, MA.

Bellwood, P., 2008. Archaeology and the origins of language families. In: Bentley, A., Maschner, H., Chippindale, C. (Eds.), Handbook of Archaeological Theories. Altamira, Lanham, pp. 225–243.

Bellwood, P., 2009. The dispersals of established food-producing populations. Current Anthropology 50, 621–626.

Bellwood, P., 2010. Language families and the history of human migration. In: Bowden, John, Himmelmann, Nikolaus, Ross, Malcolm (Eds.), A Journey through Austronesian and Papuan Linguistic and Cultural Space: Papers in Honour of Andrew K. Pawley. Pacific Linguistics, Canberra, pp. 79–93.

Bellwood, P., 2011. Holocene population history in the Pacific region as a model for world-wide food producer dispersals. Current Anthropology 52 (S4), 363–378.

Bellwood, P., 2013. First Migrants: Ancient Migration in Global Perspective. Wiley-Blackwell, Boston and Oxford.

Bellwood, P., Renfrew, C. (Eds.), 2002. Examining the Farming/Language Dispersal Hypothesis. McDonald Institute, Cambridge.

Collis, J., 2003. The Celts: Origins, Myths, Inventions. Tempus, Stroud.

Diamond, J., Bellwood, P., 2003. Farmers and their languages: the first expansions. Science 300, 597–603.

Dixon, R., 1997. The Rise and Fall of Language. Cambridge University Press, Cambridge, UK.

Donohue, M., Denham, T., 2010. Farming and language in Island Southeast Asia; reframing Austronesian history. Current Anthropology 51, 223–256.

Donohue, M., Denham, T., Oppenheimer, S., 2012. New methodologies for historical linguistics? Diachronica 29, 505–522.

Fogelin, L., 2007. Inference to the best explanation. American Antiquity 72, 603–625.

Gell-Mann, M., Ruhlen, M., 2011. The origin and evolution or word order. Proceedings of the National Academy of Sciences of the United States of America 108, 17290–17295.

Greenhill, S., Gray, R., 2012. Basic vocabulary and Bayesian phylolinguistics. Diachronica 29, 523–537.

Hill, J., 2002. Proto-uto-Aztecan Cultivation and the Northern Devolution. In: Bellwood, P., Renfrew, C. (Eds.), Examining the Farming/Language Dispersal Hypothesis. McDonald Institute, Cambridge, pp. 331–340.

Holman, E., Brown, C., et al., 2011. Automated dating of the world's language families based on lexical similarity. Current Anthropology 52, 841–875.

Jones, W., 1993. The third anniversary discourse. In: Pachori, S.S. (Ed.), Sir William Jones: a Reader. Oxford University Press, Delhi, pp. 172–178.

Kirch, P., Green, R., 2001. Hawaiki, Ancestral Polynesia. Cambridge University Press, Cambridge, UK.

Ness, I., Bellwood, P. (Eds.), 2013. Encyclopedia of Global Human Migration, 5 vols (vol. 1, Prehistory). Wiley-Blackwell, Malden (Massachusetts, USA) and Oxford.

Ostler, N., 2005. Empires of the Word. Harper Perennial, London.

Özdoğan, M., 2013. In: Ness, Bellwood (Eds.), Anatolia and the Balkans: Archaeology, vol. 1, Prehistory, pp. 139–145.

Piperno, D., 2011. Plant cultivation and domestication in the New World tropics. Current Anthropology 52 (Suppl. 4), S453–S470.

Pugach, I., Delfin, F., et al., January 2013. Genome-wide data substantiate Holocene gene flow from India to Australia. Proceedings of the National Academy of Sciences of the United States of America (Early ed.).

Renfrew, C., 1987. Archaeology and Language. Jonathan Cape, London.

Renfrew, C., 1996. Language families and the spread of farming. In: Harris, D. (Ed.), The Origins and Spread of Agriculture and Pastoralism in Eurasia. UCL Press, London, pp. 70–92.

Renfrew, C., McMahon, A., Trask, L. (Eds.), 2000. Time Depth in Historical Linguistics, 2 vols. McDonald Institute, Cambridge.

Language Learning Impairment

Paula Tallal, Salk Institute for Biological Studies, La Jolla, CA, USA; and Center for Human Development, University of California San Diego, La Jolla, CA, USA
Sabine Heim, Center for Molecular and Behavioral Neuroscience, Rutgers University-Newark, Newark, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Approximately 7.5% of kindergarten children exhibit significant delays in language learning of unknown origin – a condition often referred to as language learning impairment (LLI). Children with LLI are susceptible for reading failure and other academic achievement difficulties, school dropout, as well as social and emotional problems. Here we provide a characterization of LLI on multiple levels, namely behavioral assessment, etiological models drawing on current psychological and biological domains, as well as remedial intervention. The article concludes with a brief synopsis of future directions regarding basic and applied research in LLI.

Early reports of children who failed to learn language at or near the expected age, but with no history of hearing loss, intellectual disability, acquired brain lesions, or developmental anomalies of the speech organs, date back to the late nineteenth century. English and French authors usually published their brief descriptions under the umbrella of 'congenital aphasia' or 'aphasie congénitale,' a phrase coined by deaf instructor Léon Vaïsse in 1866. The German literature used the designation 'Hörstummheit' (hearing mutism), introduced in 1886 by Raphael Coën, then a 'Spracharzt' (literally, 'speech doctor'). Based on extensive clinical observations, the speech doctors proposed etiological models and designed therapy programs for children with severe limitations in language output, laying the ground for further refinement and conceptual progress of the condition. In the 1950s, efforts in differentiating deficits in language production from those involving both comprehension and production resulted in the terms 'expressive developmental aphasia' and 'receptive–expressive developmental aphasia.' Partially motivated by the fact that the prefix 'a' classically denotes absence, the term 'dysphasia' (indicating difficulties in language) appeared in the 1960s, and 'developmental dysphasia' became the most frequent classification for language impairment of unknown origin about two decades later (Benton, 1978; Tallal and Piercy, 1978). However, none of these classifications succeeded as appropriately descriptive terminology. Over the years, several other phrases were created to describe the syndrome, such as 'deviant language,' 'delayed language,' 'developmental language disorder,' and 'specific language impairment,' all consistently emphasizing the developmental rather than acquired nature of the disorder as well as the specificity of the deficit to the linguistic domain. Problems in language acquisition secondary to other developmental disabilities, such as hearing loss or deafness, mental retardation, autism, evidence of neurological illness (such as seizure disorder or hemiplegia), and structural or motor deficits affecting the oral musculature (such as cleft palate or cerebral palsy), were traditionally excluded from the classification under consideration.

Although 'specific language impairment' has become a popular notion in the contemporary Anglo-American literature, several researchers prefer the most recent classification 'language learning impairment' (LLI), to acknowledge the results from longitudinal studies that demonstrate a continuum between early spoken and subsequent written language learning difficulties in children (Rissman et al., 1990; Snowling et al., 2000; Tallal et al., 1997). Empirical findings that point to neurobehavioral similarities between language impairment and reading disability (or 'developmental dyslexia') have further strengthened the broader LLI concept (cf Bishop and Snowling, 2004; Heim and Benasich, 2006). However, it is important to emphasize that while children with developmental language disorders often continue on to have literacy problems, this is not the case for all children. Furthermore, there are many children with dyslexia who do not appear to have had an early history of or concurrent oral language impairment. Thus, the classification of LLI (which we will use in this article) should not be taken to be synonymous with developmental dyslexia.

Disorders of language can occur for manifold reasons, and rank among the most prevalent of all developmental disabilities. It has been estimated that approximately 20% of kindergarten and school-age children suffer from deficits in oral and/or written language (Beitchman et al., 1986b; S.E. Shaywitz and Shaywitz, 2001). The most common known causes of language disturbances are hearing loss (including intermittent hearing loss resulting from chronic otitis media), intellectual disability, and motor or structural defects affecting the oral musculature. However, even when language disorders secondary to these other conditions are excluded, approximately 7.5% of kindergarten children exhibit significant delays in language learning of unknown origin (Tomblin et al., 1997).

Diagnosis

To date, there is no definitive neurological, genetic, or other biological marker of LLI that can be applied at the individual level to diagnose the condition. As such the differential diagnosis of LLI is based on behavioral evaluations that include audiological, neurological, psychological, educational, speech, and language, and in older children, literacy assessments. The *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition, Text Revision (DSM-IV-TR; American Psychiatric Association, 2000) distinguishes between two basic categories of developmental language disorder in the context of otherwise

typical development: 'expressive language disorder,' principally affecting language production, and 'mixed receptive–expressive language disorder,' encompassing both language comprehension and production deficits. Comprehension and/or production problems may manifest within one or more of the components of language, including phonology (organization of speech sounds), morphology (inflectional and lexical word formation), semantics (meaning of language forms), syntax (phrase and sentence formation), and pragmatics (interactive conversational skills). Because psychometric measures of expressive and receptive language have not shown sufficient validity for making the distinction between these two forms of language impairment, and because individual children may change in the extent to which they show expressive and receptive problems across development, the forthcoming fifth edition of the DSM (DSM-5; expected release in May 2013) resigned the expressive versus mixed-language disorder category. Instead the designation 'specific language impairment' (SLI) along with the more general classification 'language impairment' (LI) has been proposed. The latter may be seen as synonymous with the term LLI, indicating that language learning disturbances, across development, generally affect both spoken and written language skills, may occur as a primary condition, or may coexist with other developmental disorders (cf Heim and Benasich, 2006). Accordingly, while many young children with LLI concurrently exhibit difficulties in the ability to clearly and correctly produce one or more of the sounds in speech, such articulation difficulties often occur in the absence of receptive or expressive language problems. Furthermore, many developmental cognitive disorders, such as intellectual impairment, autistic spectrum disorder, learning disability, central auditory processing disorder, as well as genetic syndromes, such as Down syndrome, Fragile X syndrome, Klinefelter syndrome, and Williams syndrome generally include delay in language development.

Children with LLI also have been found to be at risk for behavioral and emotional disorders, such as attention deficit hyperactivity disorder, anxiety, or depression (Beitchman et al., 1986a; Benasich et al., 1993; Tallal et al., 1989a). Moreover, LLI children are more likely to play alone and have lesser and poorer quality peer relationships. A recent longitudinal study suggests that behavioral difficulties (conduct issues and hyperactivity) in language-impaired individuals decrease to typical population levels from childhood to adolescence, whereas emotional problems are still present in adolescents with a history of LLI. In contrast, social (peer) difficulties become more prominent as these children grow older (St. Clair et al., 2011). Conversely, elevated occurrence rates of impaired language development have been observed in children receiving services for emotional and behavioral disorders (Benner, 2005; Benner et al., 2009).

Longitudinal research following children with LLI from the preschool through elementary school years has demonstrated a striking link between early oral language disturbances and subsequent academic achievement problems, especially dyslexia, which is characterized by reading, writing, and spelling deficits (Rissman et al., 1990; Snowling et al., 2000; Tallal et al., 1997). In line with this, family risk studies in dyslexia suggest that well-developed language skills help children to compensate for reading and spelling problems (Snowling et al., 2003, 2007). When comparing children classified as 'language impaired' with those diagnosed as 'dyslexic,' it becomes obvious that both groups are characterized by a variety of oral language deficits, specifically phonological processing and language comprehension difficulties (cf Bishop, 1997; Stark and Tallal, 1988). Impairments in these parameters of language do not always match, and different models exist to explain the overlap and distinction between the two disorders (cf Bishop and Snowling, 2004; Messaoud-Galusi and Marshall, 2010). It is still a matter of considerable debate, however, whether the relationship between oral language and literacy impairments derives from speech-specific mechanisms or from more basic neural-processing deficits (cf Heim and Benasich, 2006; Tallal and Gaab, 2006).

Etiology

Considerable attention has been focused on trying to understand the causes of LLI. Given the lack of homogeneity among children showing delayed language development, and the diversity of co-occurring symptomatologies, it is unlikely that a single etiology is responsible. Etiological research has been concerned with linguistic, perceptual/cognitive, genetic, and neurobiological mechanisms.

Linguistic Studies

There have been considerable advances made in understanding the specific linguistic deficits of children with language impairments. Particular patterns of difficulty have been observed within the linguistic components of phonology, morphology, and syntax. For instance, English-speaking preschool children with LLI omit tense-related morphemes (e.g., third person singular -s, past-tense -ed) more often than younger, typically developing children with a comparable mean length of utterance. Although the growing LLI child may demonstrate an increased use of these morphemes, some of these children still experience problems upon entering elementary education programs (cf Rice, 2004). Another area of difficulty and prolonged period of mastery is the correct use of English articles. Polite et al. (2011) found that 5-year-olds with LLI make less use of definite articles in particular, relative to younger, typically developing individuals. Furthermore, children with LLI tend to utter sentences of simpler syntactic structure and exhibit a weakness in comprehending complex sentences (cf L.B. Leonard, 1998). Pragmatic or social communication skills in LLI are usually of milder degree and less affected than the formal aspects of language (Tomblin et al., 2004). This is in contrast to children with autism spectrum disorder who generally exhibit more severe pragmatic and social communication deficits (both verbal and nonverbal) than do children with LLI.

Overall, children with LLI appear to develop the structures of language along a similar linguistic trajectory to that seen in younger children with typical language development, albeit considerably more slowly (Curtiss et al., 1992). To date, there is little evidence that the affected children produce linguistic errors that are deviant or substantially different from younger individuals matched for language status. Rather, LLI children

take much longer to progress through the stages of typical language development. Not all linguistic aspects in LLI, however, may show onset delays and protracted growth periods on the same timescale (cf Rice, 2004) and variations may also occur as a function of the severity of the disorder (Law et al., 2008).

Crosslinguistic research of childhood language disorders has failed to find common (universal) linguistic profiles across individuals struggling to learn different languages. Rather, the common feature of LLI crosslinguistically appears to be that whichever linguistic structures are acoustically brief, unstressed, or of weak phonetic substance, and occur less frequently in a particular language, are the most difficult for typical children to learn, and also the most delayed in individuals with LLI (cf L.B. Leonard, 1998). The apparent similarity in the order of linguistic development across populations and language environments gives strong support for a potent metric of sheer difficulty (whether representational or perceptual) that is imposed on the learning of linguistic structures and contents, and most vulnerable to language deficits.

Perceptual/Cognitive Studies

The most extensive etiological studies to date have utilized behavioral methodologies and have resulted in two classes of theoretical accounts: the 'linguistic specific' versus the 'domain general' account. The 'domain general' approach posits that general mechanisms (attention, sensory processing, sequencing, memory, executive functions) important for learning are at the core of the language learning deficits seen in children with LLI. A prominent exponent of the latter class is the 'rate processing constraint' hypothesis, which posits a direct impact of processing speed on learning in general. Because speech perception and comprehension require exquisite temporal resolution at high spectral accuracy over extended time periods (cf Hickok and Poeppel, 2007; Shannon, 2005), rate-processing constraints would be expected to particularly impact linguistic processing. In this view, central auditory processing constraints, specifically in the areas of auditory attention, sequencing, and serial memory, may result from a more basic impairment in the rate of neural information processing (cf Heim and Benasich, 2006; Tallal, 2004; Tallal and Gaab, 2006; Tallal et al., 1993). Tallal and colleagues have shown that many children with phonologically based speech, language, and reading disorders need significantly more neural processing time (hundreds of milliseconds instead of tens of milliseconds) to process rapidly changing sensory stimuli. They have hypothesized that this dynamic sensory constraint has a particularly detrimental effect on phonological processing because many acoustic cues occur in parallel and/or in rapid succession within syllables and words, which typically spans time windows in the tens of milliseconds range. It is important to note that if a child is poor at resolving the acoustic differences between phonemes, he/she likely will have great difficulty becoming aware that words can be broken into smaller units (speech sounds) and ultimately map onto letters and letter strings. This so-called 'phonological awareness' deficit is considered to be at the heart of reading impairments, including dyslexia (cf Melby-Lervåg et al., 2012). In addition to phonological processing and awareness, a lower level acoustic rate constraint has been shown to impact speech segmentation (McNealy et al., 2010) as well as the perception of short and unstressed grammatical forms, specifically grammatical morphology (Fellbaum et al., 1995).

In school-age children and adults with language-based learning impairments, enhanced dynamic auditory/visual sensitivity has been found to be linked to superior reading skills (Conlon et al., 2004; Helenius et al., 1999; cf Stein, 2001). Importantly, significant positive correlations between sensory processing rate and language/literacy skills have also been reported in unselected individuals from childhood to adulthood (Talcott et al., 2000; Walker et al., 2006). These findings provide strong evidence that individual differences in dynamic sensitivity across sensory modalities play an important role in the development of both oral and written language across the entire range of development. As such, children with language disturbances of unknown origin may fall at the disadvantageous end of a normal distribution, rather than representing a distinct impairment.

In addition to research being concerned with fine-grained acoustic analysis at the intrasyllabic level (in the tens of milliseconds range), other authors have focused on low-frequency temporal acoustic cues (on the order of hundreds of milliseconds). These acoustic cues, called 'time–energy envelopes,' are carried along with the phonological structure of syllables, and play a crucial role in speech rhythm identification, and thus in typical speech perception are important for processing prosodic cues in language. Several studies have posited deficits in acoustic processing of the time–energy envelope as a candidate mechanism underlying developmental language and reading disorders (Corriveau et al., 2007; Goswami et al., 2002; Richardson et al., 2004).

Constraints in the rate of processing have also been examined on a cognitive level, beyond perception. For instance, Miller et al. (2001) showed that school-age children with LLI exhibited about 14% slower response times than their typically developing peers on a wide array of tasks, some of which did not involve language, such as simple motor reaction, visual search, and mental rotation demands.

Processing limitations in working memory and executive function have also been postulated to account for the language deficits observed in LLI (cf Montgomery et al., 2010). In their seminal work, replicated by many other researchers, Gathercole and Baddeley (1990, 1993) observed that many children with LLI have great difficulty in repeating multisyllabic pseudowords. This has been attributed to impairment in phonological storage, affecting acquisition of vocabulary and grammar, as well as literacy achievement later in life. In a more recent study, L.B. Leonard et al. (2007) found that verbal working memory made a larger contribution to language performance than general processing speed.

The proponents of the 'linguistic-specific' hypothesis strongly dispute these data. They exclude any role for nonlinguistic processing constraints as a possible etiology for LLI. Alternatively, they support a modular view of language processing and propose that developmental language disorders are caused by failure of innate, 'language specific,' modules to function normally (cf L.B. Leonard, 1998; Pennington and Bishop, 2009).

Genetic Studies

LLI often aggregates in families (Flax et al., 2003; Tallal et al., 1989b, 2001; Tomblin, 1989), occurs more often in monozygotic than in dizygotic twins (Bishop et al., 1995; Tomblin and Buckwalter, 1998), and may be genetically transmitted (Bartlett et al., 2002, 2004; Monaco, 2007; SLI Consortium, 2002, 2004; Vernes et al., 2008). Infants born into families with a history of LLI are, thus, at an elevated risk for developing language/literacy impairments themselves. With the goal of identifying the developmental precursors to LLI, Benasich and colleagues (cf Benasich and Choudhury, 2012; Heim and Benasich, 2006) followed two groups of infants prospectively through the early school years: infants born into families with a positive history of LLI (FH+) and age-mates without such a family history (FH−). Using convergent methodologies at the behavioral and brain-electrical level (see Section Neurobiological Studies) the specific aim was to determine whether individual differences in perceptual/cognitive processing abilities, established in infancy, predicted subsequent language development and disorders. At age 7.5 months, FH+ infants tended to need longer processing times for nonverbal acoustic events than FH− controls. Follow-up measures indicated that two variables – nonverbal acoustic processing rate established at this early age and gender – correctly predicted over 90% of those children who at age 3 years scored within the 'impaired' range on the verbal (but not performance) scale of the Stanford Binet test of intelligence (Thorndike et al., 1986). In other words, those infants (particularly males) who exhibited the slowest auditory processing rates were subsequently most likely to be delayed in language development (Benasich and Tallal, 1996). Perhaps the most striking finding of this study, however, was that these data were independent of family history and thus clearly demonstrate the crucial aspect of sensory timing and gender (being male) for early language development. Moreover, when these children were followed into their early elementary school years, a toddler's rate of receptive (but not expressive) language development was found to be a significant predictor of subsequent reading skills (Flax et al., 2009). This is consistent with other studies demonstrating a link between early spoken language impairments and literacy delay (Catts et al., 2008; Snowling et al., 2000; Whitehouse et al., 2009).

The finding that LLI appears to run in families has also led to a number of molecular-genetic studies focusing on the search for specific genes that may be linked to the condition. Bartlett et al. (2002, 2004) collected extensive behavioral and genetic data on five multigenerational pedigrees of Celtic ancestry that segregate language impairment. A genome-wide categorical linkage analysis was conducted under both dominant and recessive models by use of three phenotypic classifications: clinical diagnosis of language impairment, spoken language test scores below the normal range, and substantial discrepancy between reading and nonverbal cognitive ability. Despite these pedigrees being selected based on spoken language impairment, significant linkages were found in a region of the long arm (q) of chromosome 13 (13q21), and only with the recessive reading discrepancy model. This result was confirmed in a North American sample selected according to the same phenotype criteria.

As mentioned earlier in this article, language delay is a core disability in autistic spectrum disorder, and those children with autism who do learn to talk often show language profiles similar to that of LLI children (Kjelgaard and Tager-Flusberg, 2001). Family studies have found that even in siblings and twins of children with autism who are not autistic, a disproportionate number experience language difficulties reminiscent of the phenotype of language impairment of unknown origin (Fombonne et al., 1997; Freeman et al., 1989; Szatmari et al., 1993). It is of considerable interest, therefore, that a significant linkage to chromosome 13q21 also has been found in a number of genetic studies of autism (Collaborative Linkage Study of Autism, 2001a,b). Furthermore, when families were subdivided by whether the autistic child and his/her sibling were concordant for presence or absence of phrase speech delay (defined as onset of phrase speech later than 36 months), the linkage signal on 13q21 was only increased in the speech-delayed group (Collaborative Linkage Study of Autism, 2001b). Intriguingly, the same genetic marker in both the LLI and autism studies yielded the highest linkage scores on this particular chromosomal region (Bartlett et al., 2004). These genetic studies provide strong evidence not only for a continuum between developmental language and reading disorders (Bartlett et al., 2002), but also a potential overlap with autism (Bartlett et al., 2004).

Neurobiological Studies

Over the past three decades, there has been a spurt in literature applying several electrophysiological and neuroimaging methodologies to the study of the neuronal correlates of LLI. Because a dominant role of the left hemisphere has been noted for both acoustic and linguistic functions that rely on high-temporal precision information processing (Schwartz and Tallal, 1980; Zatorre and Belin, 2001), neurobiological studies have specifically focused on hemispheric asymmetry in LLI. Results from magnetic resonance imaging (MRI) have indicated volume reduction in left-hemisphere language structures (such as the *planum temporale*) and/or absence of the typical leftward asymmetry in developmental disorders of oral and written language (Gauger et al., 1997; Herbert et al., 2005; Hynd et al., 1990; Plante et al., 1991). The issue of these digressions is not conclusive as there have been other contradictory observations (Preis et al., 1998; Rumsey et al., 1997). A promising approach is provided by a more recent MRI study, suggesting that LLI children with relatively smaller, symmetrical brain structures may show deficits across a wide range of language and reading functions, whereas those children with larger, more asymmetrical brain structures tend to exhibit less difficulty, with relatively preserved oral and written language comprehension (Leonard et al., 2006).

Functional MRI (fMRI) studies have been conducted mainly in school-age children and adults with a diagnosis of dyslexia rather than SLI. These have revealed a hypoactivation of left-hemisphere posterior regions during reading processes (Hoeft

et al., 2007; Shaywitz et al., 2002; Shaywitz et al., 1998; Temple et al., 2001). Affected individuals have also been found to show hyperactivation in anterior regions, such as the inferior frontal gyrus of the left and right hemisphere, which has been suggested to reflect compensatory mechanisms to overcome the disruption of posterior cortical networks (Georgiewa et al., 2002; Hoeft et al., 2007; Shaywitz et al., 2002; Shaywitz et al., 1998). In line with this, the strength of right prefrontal activity during a reading task taxing phonological awareness positively predicted dyslexic children's reading progress 2.5 years later (Hoeft et al., 2011). In a recent study, Raschle et al. (2012) reported that the posterior hypoactivation is already a characteristic in the prereading child at risk for developing language/literacy problems by virtue of family history. Furthermore, reduced posterior activity was accompanied by lower performance in an auditory phonological memory test, regardless of familial risk. This important result indicates that rather than being a consequence of lack of reading experience or brain reorganization due to reading difficulty, this distinctive pattern of hypoactivation of left-hemisphere posterior regions may be a neurobiological constraint on subsequent literacy development.

Electrophysiological techniques such as electroencephalography (EEG) and magnetoencephalography (MEG) have the advantage of providing temporal resolution most compatible with probing the real-time neural processing involved in language. In EEG work exploring the neuronal correlates of processing 'slow' temporal cues (hundreds of milliseconds) important for speech prosody, typically reading children had earlier and larger right-hemisphere responses, whereas poor readers exhibited dominant left-hemisphere activity. Electrocortical asymmetry contributed significantly to the variability in phonological awareness (Abrams et al., 2009). A similar finding has been obtained in adults with dyslexia, who showed impaired performance in phonological domains and decreased temporal stability of low-frequency magnetocortical responses in the right hemisphere evoked by slow modulations in nonspeech stimuli (Hämäläinen et al., 2012).

Electrophysiological studies probing the neural correlates of processing 'fast' temporal cues (tens of milliseconds) essential for tracking formant transitions in speech often show deviances in the left hemisphere of language/literacy-impaired individuals. Using MEG, the left-hemispheric sources underlying the magnetic wave around 200 ms after onset of a tone or syllable were found to be located more anterior in dyslexia children than in typical peers (Heim et al., 2000). This has been taken to indicate that the source positions might index activity within the *planum temporale* in the unimpaired children, whereas in the dyslexia children sources might be tied to sites anterior to the planum. When listening to brief and unstressed grammatical forms embedded in various language contexts, LLI children showed attenuated electrocortical responses from 200 to 400 ms after stimulus onset over left temporal sites, and enhanced responses over the right hemisphere. The processing pattern tended to be reversed in children with typical language development (Shafer et al., 2001). Such studies are suggestive of a reduced dynamic sensitivity in the posterior perisylvian region of the left hemisphere. As a consequence, less efficient cortical areas or sites might be recruited for auditory reception, contributing to the rate-processing differences observed in individuals with a history of language-based learning disability. More direct support for the 'rate processing constraint' hypothesis is provided by a recent EEG study evaluating high-frequency auditory oscillations evoked by fast-paced tone doublets (Heim et al., 2011). Oscillatory rhythms in school-age children with LLI did not differ from those in typically developing peers for the first tone, but negatively digressed in magnitude and temporal stability for the second tone of the doublet. Explanations for this pattern may include the fact that neural masses cannot be engaged in a well-synchronized and orderly fashion in LLI when rapid processing is needed. Alternatively, the recovery time of the cell assemblies firing to the first stimulus may be more variable in the LLI group, leading to greater latency jitter of the second response.

Divergences in brain-electrical processing patterns to rapid tone sequences have already been observed between 6-month-old infants with a family history of LLI and their age-mates from control families (Benasich et al., 2006; Choudhury and Benasich, 2011). The at-risk group exhibited smaller neuronal discrimination responses than the control group in the left hemisphere. Moreover, electrocortical parameters in the latency range around 250 ms after stimulus onset have been found to be predictive of language abilities at 16, 24, 36, and 48 months of age above and beyond family risk status. It may well be the case, however, that encoding deficits for fine-grained acoustic structures are already present on a subcortical level. By recording brainstem potentials to fast tonal sweeps, Basu et al. (2010) found the very early neural activity to be degraded in the majority of children comprising the LLI group. Similar evidence emerged from studies involving poor readers, showing weaker neuronal encoding of syllable fine structure than typically reading children (Banai et al., 2009; Hornickel et al., 2012). It remains to be determined whether and how such mechanisms ultimately contribute to other linguistic deficits associated with LLI.

Remedial Intervention

Remediation of LLI is generally provided by speech language pathologists, psychologists, or special education teachers. Traditionally, speech and language therapy has been guided by a detailed linguistic assessment of the pattern of receptive as well as expressive errors made by an individual child using both standardized testing instruments and language sampling methodologies. Remediation has focused on helping the child to use age-appropriate linguistic structures in a variety of contexts. In the literacy domain, intervention protocols have been concerned with explicitly training children to recognize and manipulate the speech sounds within words (called phonological awareness, see Section Perceptual/Cognitive Studies) in an attempt to improve reading decoding skills, as well as methods aimed at improving fluency and reading comprehension. Longitudinal research has shown that children with LLI generally require

approximately 2 years of traditional intervention to achieve the equivalent of 1 year of gain in oral and written language development (cf L.B. Leonard, 1998; Lyon and Moats, 1997). In a study monitoring the acquisition of grammatical morphemes by preschool children with LLI during intervention, L.B. Leonard et al. (2006, 2008) reported only modest gains, even after 96 sessions of treatment. Because the performance increases were steady and maintained 1 month after completion, the authors concluded that maturation may have had an important influence, suggesting an arduous and protracted acquisition process of grammatical morphemes. Other factors such as nonlinguistic skills may also have mediated the outcome. Drawing on research showing significant contributions of working memory and executive function to language performance, L.B. Leonard et al. (2007) recommended that intervention in LLI should be broad-based, concurrently training both linguistic forms and general information processing. Greater efficiency in information processing may help the affected child to better learn the targeted linguistic structures.

Electrophysiological studies of sensory cortices in animals have supported the notion that functional brain circuits are shaped by experience and can be altered through specific, temporally cohesive training regimens (cf Buonomano and Merzenich, 1998; de Villers-Sidani and Merzenich, 2011). This approach has led to the design of neuroplasticity-based remedies, which assume that the temporal precision of neural coding can be enhanced by intense training in an optimal learning environment. Merzenich and Tallal (Merzenich et al., 1996; Tallal et al., 1996) reported the results from the first series of controlled studies designed to test this hypothesis. Children with LLI worked on a set of computerized audiovisual training exercises, in which rapid transitions in speech and language were either presented with natural speech (control group) or speech that had been prolonged in time and differentially amplified (experimental group). As children in the experimental group progressed through the exercises and performance improved, the modified acoustic stimuli were presented at rates and amplitude levels closer and closer to those that occur in natural speech. In addition to the acoustically modified speech tasks, the experimental group engaged in an exercise designed to individually adapt to speed up their rate of nonlinguistic acoustic processing, while the control group performed another computer exercise that did not target processing speed. Both groups of children benefited significantly from intervention in speech discrimination and language comprehension. However, participants in the experimental group, who completed training with a focus on enhancing fine-grained acoustic analysis, exhibited not only an acceleration in auditory rate processing, but also significantly greater gains than the control group on measures of speech and language. Subsequent studies that implemented this form of neuroplasticity-based remedy in LLI revealed collateral improvements in the literacy domain, as well as changes in metabolic brain activity (Gaab et al., 2007; Temple et al., 2003; cf Tallal, 2004). Moreover, LLI individuals and typically developing children benefited from this training on measures of brain-electrical activity underlying selective auditory attention (Stevens et al., 2008). Using audiovisual training without linguistic material, Kujala et al. (2001) reported faster reaction times and enhanced electrocortical responses to sound changes in reading-impaired children. Importantly, the nonlinguistic training generalized to the level of reading decoding performance.

Overall, such remediation studies, together with those focusing on phonological awareness training (cf Snowling and Hulme, 2011), have stimulated intervention research that aims to develop more scientifically validated approaches for the treatment of disturbances in oral and written language learning. Several training programs (presented as computer games) have been designed and made commercially available to enhance perceptual, cognitive, language, and literacy skills in various age groups (Fast ForWord-Language and Literacy Series at http://www.scilearn.com/, Cogmed Working Memory Training at http://www.cogmed.com/, LearningRx at http://www.learningrx.com/). Crucially, for children without access to resources for intense, individualized speech therapy, computerized training programs have proven to be equally as effective as one-on-one speech therapy or academic enrichment provided for a comparable amount of time (Gillam et al., 2008).

Goals for Future Research

The brief overview presented here underscores the complexity of LLI, which involves multiple levels of analysis, ranging from the molecular level to overt behavior. It seems likely that the precursors of LLI can be observed very early in life, and also that language deficits persist, usually in subtle form, into later childhood, adolescence, and conceivably into adulthood. Children with LLI are susceptible for reading failure and other academic achievement difficulties, school dropout, as well as social and emotional problems. Therefore, the impact of LLI on society is substantive (cf Heim and Benasich, 2006). Future research should continue to focus on etiology, specifically on the neurophysiological, genetic, and environmental influences that contribute mutually to the condition. In this vein, research is underway aiming to characterize how nonlinguistic neural processes (attention, dynamic processing, memory, and other executive functions) interact developmentally with language systems in the brain – during typical and atypical language development. Prospective studies following infants at risk for LLI and control peers longitudinally into adolescence have shown particular promise to address these questions. Such studies are still few in number, however. Moreover, the integration of different levels of analysis, including cognitive methodologies, sophisticated neuroimaging and electrophysiological procedures, as well as molecular-genetic techniques will help to narrow down the candidate causal and contributory mechanisms for LLI. Determining whether LLI represents a distinct disability or rather lies at the disadvantageous end of a continuum of interindividual differences will also remain an important focus of research. The ultimate goal will be to achieve a theoretical understanding of the multiple factors affecting the language acquisition process. This may in turn lead to

the development of more effective training protocols for accelerating oral as well as written language learning in all individuals, specifically those with LLI.

See also: Language Acquisition; Multilingual Language Development; Neural Basis of Language Development; Neurophysiological Markers of Language Development; Speech Perception.

Bibliography

Abrams, D.A., Nicol, T., Zecker, S., Kraus, N., 2009. Abnormal cortical processing of the syllable rate of speech in poor readers. Journal of Neuroscience 29, 7686–7693.

American Psychiatric Association (APA), 2000. Diagnostic and Statistical Manual of Mental Disorders: Fourth Edition, Text Revision, DSM-IV-TR. Author, Washington, DC.

Banai, K., Hornickel, J., Skoe, E., Nicol, T., Zecker, S., Kraus, N., 2009. Reading and subcortical auditory function. Cerebral Cortex 19, 2699–2707.

Bartlett, C.W., Flax, J.F., Logue, M.W., Smith, B.J., Vieland, V.J., Tallal, P., Brzustowicz, L.M., 2004. Examination of potential overlap in autism and language loci on chromosomes 2, 7, and 13 in two independent samples ascertained for specific language impairment. Human Heredity 57, 10–20.

Bartlett, C.W., Flax, J.F., Logue, M.W., Vieland, V.J., Bassett, A.S., Tallal, P., Brzustowicz, L.M., 2002. A major susceptibility locus for specific language impairment is located on 13q21. American Journal of Human Genetics 71, 45–55.

Basu, M., Krishnan, A., Weber-Fox, C., 2010. Brainstem correlates of temporal auditory processing in children with specific language impairment. Developmental Science 13, 77–91.

Beitchman, J.H., Nair, R., Clegg, M., Ferguson, B., Patel, P.G., 1986a. Prevalence of psychiatric disorders in children with speech and language disorders. Journal of the American Academy of Child Psychiatry 25, 528–535.

Beitchman, J.H., Nair, R., Clegg, M., Patel, P.G., Ferguson, B., Pressman, E., Smith, A., 1986b. Prevalence of speech and language disorders in 5-year-old kindergarten children in the Ottawa-Carleton region. Journal of Speech and Hearing Disorders 51, 98–110 (Erratum published in 1987, Journal of Speech and Hearing Disorders 52, 94).

Benasich, A.A., Choudhury, N., 2012. Timing, information processing, and efficacy: early factors that impact childhood language trajectories. In: Benasich, A.A., Fitch, R.H. (Eds.), Developmental Dyslexia: Early Precursors, Neurobehavioral Markers, and Biological Substrates. Brookes Publishing, Baltimore, MD, pp. 99–118.

Benasich, A.A., Choudhury, N., Friedman, J.T., Realpe-Bonilla, T., Chojnowska, C., Gou, Z., 2006. The infant as a prelinguistic model for language learning impairments: predicting from event-related potentials to behavior. Neuropsychologia 44, 396–411.

Benasich, A.A., Curtiss, S., Tallal, P., 1993. Language, learning, and behavioral disturbances in childhood: a longitudinal perspective. Journal of the American Academy of Child and Adolescent Psychiatry 32, 585–594.

Benasich, A.A., Tallal, P., 1996. Auditory temporal processing thresholds, habituation, and recognition memory over the 1st year. Infant Behavior and Development 19, 339–357.

Benner, G., 2005. Language skills of elementary-aged children with emotional and behavioral disorders. Great Plains Research: A Journal of Natural and Social Sciences 785. Available from: http://digitalcommons.unl.edu/greatplainsresearch/785.

Benner, G.J., Mattison, R.E., Nelson, J.R., Ralston, N.C., 2009. Types of language disorders in students classified as ED: prevalence and association with learning disabilities and psychopathology. Special Education and Communication Disorders Faculty Publications 27. Available from: http://digitalcommons.unl.edu/specedfacpub/27.

Benton, A.L., 1978. The cognitive functioning of children with developmental dysphasia. In: Wyke, M.A. (Ed.), Developmental Dysphasia. Academic Press, New York, NY, pp. 43–62.

Bishop, D.V., 1997. Uncommon Understanding: Development and Disorders of Language Comprehension in Children. Psychology Press, Hove, UK.

Bishop, D.V., North, T., Donlan, C., 1995. Genetic basis of specific language impairment: evidence from a twin study. Developmental Medicine and Child Neurology 37, 56–71.

Bishop, D.V., Snowling, M.J., 2004. Developmental dyslexia and specific language impairment: same or different? Psychological Bulletin 130, 858–886.

Buonomano, D.V., Merzenich, M.M., 1998. Cortical plasticity: from synapses to maps. Annual Review of Neuroscience 21, 149–186.

Catts, H.W., Bridges, M.S., Little, T.D., Tomblin, J.B., 2008. Reading achievement growth in children with language impairments. Journal of Speech, Language, and Hearing Research 51, 1569–1579.

Choudhury, N., Benasich, A.A., 2011. Maturation of auditory evoked potentials from 6 to 48 months: prediction to 3 and 4 year language and cognitive abilities. Clinical Neurophysiology 122, 320–338.

Collaborative Linkage Study of Autism, 2001a. An autosomal genomic screen for autism. American Journal of Medical Genetics 105, 609–615.

Collaborative Linkage Study of Autism, 2001b. Incorporating language phenotypes strengthens evidence of linkage to autism. American Journal of Medical Genetics 105, 539–547.

Conlon, E., Sanders, M., Zapart, S., 2004. Temporal processing in poor adult readers. Neuropsychologia 42, 142–157.

Corriveau, K., Pasquini, E., Goswami, U., 2007. Basic auditory processing skills and specific language impairment: a new look at an old hypothesis. Journal of Speech, Language, and Hearing Research 50, 647–666.

Curtiss, S., Katz, W., Tallal, P., 1992. Delay versus deviance in the language acquisition of language-impaired children. Journal of Speech and Hearing Research 35, 373–383.

Fellbaum, C., Miller, S., Curtiss, S., Tallal, P., 1995. An auditory processing deficit as a possible source of SLI. In: McLaughlin, D., McEwen, S. (Eds.), Proceedings of the 20th Annual Boston University Conference on Language Development. Cascadilla Press, Somerville, MA, pp. 204–215.

Flax, J.F., Realpe-Bonilla, T., Hirsch, L.S., Brzustowicz, L.M., Bartlett, C.W., Tallal, P., 2003. Specific language impairment in families: evidence for co-occurrence with reading impairments. Journal of Speech, Language, and Hearing Research 46, 530–543.

Flax, J.F., Realpe-Bonilla, T., Roesler, C., Choudhury, N., Benasich, A., 2009. Using early standardized language measures to predict later language and early reading outcomes in children at high risk for language-learning impairments. Journal of Learning Disabilities 42, 61–75.

Fombonne, E., Bolton, P., Prior, J., Jordan, H., Rutter, M., 1997. A family study of autism: cognitive patterns and levels in parents and siblings. Journal of Child Psychology and Psychiatry 38, 667–683.

Freeman, B.J., Ritvo, E.R., Mason-Brothers, A., Pingree, C., Yokota, A., Jenson, W.R., McMahon, W.M., Petersen, P.B., Mo, A., Schroth, P., 1989. Psychometric assessment of first-degree relatives of 62 autistic probands in Utah. American Journal of Psychiatry 146, 361–364.

Gaab, N., Gabrieli, J.D., Deutsch, G.K., Tallal, P., Temple, E., 2007. Neural correlates of rapid auditory processing are disrupted in children with developmental dyslexia and ameliorated with training: an fMRI study. Restorative Neurology and Neuroscience 25, 295–310.

Gathercole, S.E., Baddeley, A.D., 1990. Phonological memory deficits in language disordered children: is there a causal connection? Journal of Memory and Language 29, 336–360.

Gathercole, S.E., Baddeley, A.D., 1993. Phonological working memory: a critical building block for reading development and vocabulary acquisition? European Journal of Psychology of Education 8, 259–272.

Gauger, L.M., Lombardino, L.J., Leonard, C.M., 1997. Brain morphology in children with specific language impairment. Journal of Speech, Language, and Hearing Research 40, 1272–1284.

Georgiewa, P., Rzanny, R., Gaser, C., Gerhard, U.J., Vieweg, U., Freesmeyer, D., Mentzel, H.J., Kaiser, W.A., Blanz, B., 2002. Phonological processing in dyslexic children: a study combining functional imaging and event related potentials. Neuroscience Letters 318, 5–8.

Gillam, R.B., Loeb, D.F., Hoffman, L.M., Bohman, T., Champlin, C.A., Thibodeau, L., Widen, J., Brandel, J., Friel-Patti, S., 2008. The efficacy of Fast ForWord Language intervention in school-age children with language impairment: a randomized controlled trial. Journal of Speech, Language, and Hearing Research 51, 97–119.

Goswami, U., Thomson, J., Richardson, U., Stainthorp, R., Hughes, D., Rosen, S., Scott, S.K., 2002. Amplitude envelope onsets and developmental dyslexia: a new hypothesis. Proceedings of the National Academy of Sciences of the United States of America 99, 10911–10916.

Hämäläinen, J.A., Rupp, A., Soltész, F., Szücs, D., Goswami, U., 2012. Reduced phase locking to slow amplitude modulation in adults with dyslexia: an MEG study. Neuroimage 59, 2952–2961.

Heim, S., Benasich, A.A., 2006. Developmental disorders of language. In: Cicchetti, D., Cohen, D.J. (Eds.), Developmental Psychopathology: Risk, Disorder, and Adaptation, second ed., vol. 3. Wiley, Hoboken, NJ, pp. 268–316.

Heim, S., Eulitz, C., Kaufmann, J., Füchter, I., Pantev, C., Lamprecht-Dinnesen, A., Matulat, P., Scheer, P., Borstel, M., Elbert, T., 2000. Atypical organisation of the auditory cortex in dyslexia as revealed by MEG. Neuropsychologia 38, 1749–1759.

Heim, S., Friedman, J.T., Keil, A., Benasich, A.A., 2011. Reduced sensory oscillatory activity during rapid auditory processing as a correlate of language-learning impairment. Journal of Neurolinguistics 24, 539–555.

Helenius, P., Uutela, K., Hari, R., 1999. Auditory stream segregation in dyslexic adults. Brain 122, 907–913.

Herbert, M.R., Ziegler, D.A., Deutsch, C.K., O'Brien, L.M., Kennedy, D.N., Filipek, P.A., Bakardjiev, A.I., Hodgson, J., Takeoka, M., Makris, N., Caviness Jr., V.S., 2005. Brain asymmetries in autism and developmental language disorder: a nested whole-brain analysis. Brain 128, 213–226.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nature Reviews Neuroscience 8, 393–402.

Hoeft, F., McCandliss, B.D., Black, J.M., Gantman, A., Zakerani, N., Hulme, C., Lyytinen, H., Whitfield-Gabrieli, S., Glover, G.H., Reiss, A.L., Gabrieli, J.D., 2011. Neural systems predicting long-term outcome in dyslexia. Proceedings of the National Academy of Sciences of the United States of America 108, 361–366.

Hoeft, F., Meyler, A., Hernandez, A., Juel, C., Taylor-Hill, H., Martindale, J.L., McMillon, G., Kolchugina, G., Black, J.M., Faizi, A., Deutsch, G.K., Siok, W.T., Reiss, A.L., Whitfield-Gabrieli, S., Gabrieli, J.D., 2007. Functional and morphometric brain dissociation between dyslexia and reading ability. Proceedings of the National Academy of Sciences of the United States of America 104, 4234–4239.

Hornickel, J., Anderson, S., Skoe, E., Yi, H.G., Kraus, N., 2012. Subcortical representation of speech fine structure relates to reading ability. Neuroreport 23, 6–9.

Hynd, G.W., Semrud-Clikeman, M., Lorys, A.R., Novey, E.S., Eliopulos, D., 1990. Brain morphology in developmental dyslexia and attention deficit disorder/hyperactivity. Archives of Neurology 47, 919–926.

Kjelgaard, M.M., Tager-Flusberg, H., 2001. An investigation of language impairment in autism: implications for genetic subgroups. Language and Cognitive Processes 16, 287–308.

Kujala, T., Karma, K., Ceponiene, R., Belitz, S., Turkkila, P., Tervaniemi, M., Näätänen, R., 2001. Plastic neural changes and reading improvement caused by audiovisual training in reading-impaired children. Proceedings of the National Academy of Sciences of the United States of America 98, 10509–10514.

Law, J., Tomblin, J.B., Zhang, X., 2008. Characterizing the growth trajectories of language-impaired children between 7 and 11 years of age. Journal of Speech, Language, and Hearing Research 51, 739–749.

Leonard, C., Eckert, M., Given, B., Virginia, B., Eden, G., 2006. Individual differences in anatomy predict reading and oral language impairments in children. Brain 129, 3329–3342.

Leonard, L.B., 1998. Children with Specific Language Impairment. MIT Press, Cambridge, MA.

Leonard, L.B., Camarata, S.M., Pawłowska, M., Brown, B., Camarata, M.N., 2006. Tense and agreement morphemes in the speech of children with specific language impairment during intervention: phase 2. Journal of Speech, Language, and Hearing Research 49, 749–770.

Leonard, L.B., Camarata, S.M., Pawlowska, M., Brown, B., Camarata, M.N., 2008. The acquisition of tense and agreement morphemes by children with specific language impairment during intervention: phase 3. Journal of Speech, Language, and Hearing Research 51, 120–125.

Leonard, L.B., Ellis Weismer, S., Miller, C.A., Francis, D.J., Tomblin, J.B., Kail, R.V., 2007. Speed of processing, working memory, and language impairment in children. Journal of Speech, Language, and Hearing Research 50, 408–428.

Lyon, G.R., Moats, L.C., 1997. Critical conceptual and methodological considerations in reading intervention research. Journal of Learning Disabilities 30, 578–588.

McNealy, K., Mazziotta, J.C., Dapretto, M., 2010. The neural basis of speech parsing in children and adults. Developmental Science 13, 385–406.

Melby-Lervåg, M., Lyster, S.A., Hulme, C., 2012. Phonological skills and their role in learning to read: a meta-analytic review. Psychological Bulletin 138, 322–352.

Merzenich, M.M., Jenkins, W.M., Johnston, P., Schreiner, C., Miller, S.L., Tallal, P., 1996. Temporal processing deficits of language-learning impaired children ameliorated by training. Science 271, 77–81.

Messaoud-Galusi, S., Marshall, C.R., 2010. Introduction to this special issue exploring the overlap between dyslexia and SLI: the role of phonology. Scientific Studies of Reading 14, 1–7.

Miller, C.A., Kail, R., Leonard, L.B., Tomblin, J.B., 2001. Speed of processing in children with specific language impairment. Journal of Speech, Language, and Hear Research 44, 416–433.

Monaco, A.P., 2007. Multivariate linkage analysis of specific language impairment (SLI). Annals of Human Genetics 71, 660–673.

Montgomery, J.W., Magimairaj, B.M., Finney, M.C., 2010. Working memory and specific language impairment: an update on the relation and perspectives on assessment and treatment. American Journal of Speech-Language Pathology 19, 78–94.

Pennington, B.F., Bishop, D.V., 2009. Relations among speech, language, and reading disorders. Annual Review of Psychology 60, 283–306.

Plante, E., Swisher, L., Vance, R., Rapcsak, S., 1991. MRI findings in boys with specific language impairment. Brain and Language 41, 52–66.

Polite, E.J., Leonard, L.B., Roberts, F.D., 2011. The use of definite and indefinite articles by children with specific language impairment. International Journal of Speech-Language Pathology 13, 291–300.

Preis, S., Jäncke, L., Schittler, P., Huang, Y., Steinmetz, H., 1998. Normal intrasylvian anatomical asymmetry in children with developmental language disorder. Neuropsychologia 36, 849–855.

Raschle, N.M., Zuk, J., Gaab, N., 2012. Functional characteristics of developmental dyslexia in left-hemispheric posterior brain regions predate reading onset. Proceedings of the National Academy of Sciences of the United States of America 109, 2156–2161.

Rice, M.L., 2004. Growth models of developmental language disorders. In: Rice, M.L., Warren, S.F. (Eds.), Developmental Language Disorders: From Phenotypes to Etiologies. Erlbaum, Mahwah, NJ, pp. 207–240.

Richardson, U., Thomson, J.M., Scott, S.K., Goswami, U., 2004. Auditory processing skills and phonological representation in dyslexic children. Dyslexia 10, 215–233.

Rissman, M., Curtiss, S., Tallal, P., 1990. School placement outcomes of young language impaired children. Journal of Speech-Language Pathology and Audiology 14, 49–58.

Rumsey, J.M., Donohue, B.C., Brady, D.R., Nace, K., Giedd, J.N., Andreason, P., 1997. A magnetic resonance imaging study of planum temporale asymmetry in men with developmental dyslexia. Archives of Neurology 54, 1481–1489.

Schwartz, J., Tallal, P., 1980. Rate of acoustic change may underlie hemispheric specialization for speech perception. Science 207, 1380–1381.

Shafer, V.L., Schwartz, R.G., Mor, M.L., Kessler, K.L., Kurtzberg, D., Ruben, R.J., 2001. Neurophysiological indices of language impairment in children. Acta Otolaryngolica 121, 297–300.

Shannon, R.V., 2005. Speech and music have different requirements for spectral resolution. International Review of Neurobiology 70, 121–134.

Shaywitz, B.A., Shaywitz, S.E., Pugh, K.R., Mencl, W.E., Fulbright, R.K., Skudlarski, P., Constable, R.T., Marchione, K.E., Fletcher, J.M., Lyon, G.R., Gore, J.C., 2002. Disruption of posterior brain systems for reading in children with developmental dyslexia. Biological Psychiatry 52, 101–110.

Shaywitz, S.E., Shaywitz, B.A., 2001. The neurobiology of reading and dyslexia. Focus on Basics: Connecting Research and Practice 5. Available from: http://www.ncsall.net/?id=278.

Shaywitz, S.E., Shaywitz, B.A., Pugh, K.R., Fulbright, R.K., Constable, R.T., Mencl, W.E., Shankweiler, D.P., Liberman, A.M., Skudlarski, P., Fletcher, J.M., Katz, L., Marchione, K.E., Lacadie, C., Gatenby, C., Gore, J.C., 1998. Functional disruption in the organization of the brain for reading in dyslexia. Proceedings of the National Academy of Sciences of the United States of America 95, 2636–2641.

SLI Consortium, 2002. A genomewide scan identifies two novel loci involved in specific language impairment. American Journal of Human Genetics 70, 384–398.

SLI Consortium (SLIC), 2004. Highly significant linkage to the SLI1 locus in an expanded sample of individuals affected by specific language impairment. American Journal of Human Genetics 74, 1225–1238.

Snowling, M., Bishop, D.V., Stothard, S.E., 2000. Is preschool language impairment a risk factor for dyslexia in adolescence? Journal of Child Psychology and Psychiatry 41, 587–600.

Snowling, M., Gallagher, A., Frith, U., 2003. Family risk of dyslexia is continuous: individual differences in the precursors of reading skill. Child Development 74, 358–373.

Snowling, M.J., Hulme, C., 2011. Evidence-based interventions for reading and language difficulties: creating a virtuous circle. British Journal of Educational Psychology 81, 1–23.

Snowling, M.J., Muter, V., Carroll, J., 2007. Children at family risk of dyslexia: a follow-up in early adolescence. Journal of Child Psychology and Psychiatry 48, 609–618.

Stark, R.E., Tallal, P., 1988. Language, Speech, and Reading Disorders in Children: Neuropsychological Studies. Little Brown, Boston, MA.

St. Clair, M.C., Pickles, A., Durkin, K., Conti-Ramsden, G., 2011. A longitudinal study of behavioral, emotional and social difficulties in individuals with a history of specific language impairment (SLI). Journal of Communication Disorders 44, 186–199.

Stein, J., 2001. The magnocellular theory of developmental dyslexia. Dyslexia 7, 12–36.

Stevens, C., Fanning, J., Coch, D., Sanders, L., Neville, H., 2008. Neural mechanisms of selective auditory attention are enhanced by computerized training: electrophysiological evidence from language-impaired and typically developing children. Brain Research 1205, 55–69.

Szatmari, P., Jones, M.B., Tuff, L., Bartolucci, G., Fisman, S., Mahoney, W., 1993. Lack of cognitive impairments in first-degree relatives of children with pervasive developmental disorders. Journal of the American Academy of Child and Adolescent Psychiatry 32, 1264–1273.

Talcott, J.B., Witton, C., McClean, M.F., Hansen, P.C., Rees, A., Green, G.G., Stein, J.F., 2000. Dynamic sensory sensitivity and children's word decoding skills. Proceedings of the National Academy of Sciences of the United States of America 97, 2952–2957.

Tallal, P., 2004. Improving language and literacy is a matter of time. Nature Reviews Neuroscience 5, 721–728.

Tallal, P., Allard, L., Miller, S., Curtiss, S., 1997. Academic outcomes of language impaired children. In: Hulme, C., Snowling, M. (Eds.), Dyslexia: Biology, Cognition and Intervention. Whurr, London, UK, pp. 167–181.

Tallal, P., Benasich, A.A., 2002. Developmental language learning impairments. Development and Psychopathology 14, 559–579.

Tallal, P., Dukette, D., Curtiss, S., 1989a. Behavioral/emotional profiles of preschool language-impaired children. Development and Psychopathology 1, 51–67.

Tallal, P., Gaab, N., 2006. Dynamic auditory processing, musical experience and language development. Trends in Neurosciences 29, 382–390.

Tallal, P., Galaburda, A.M., Llinás, R.R., von Euler, C. (Eds.), 1993. Temporal Information Processing in the Nervous System: Special Reference to Dyslexia and Dysphasia, vol. 682. New York Academy of Sciences Press, New York, NY.

Tallal, P., Hirsch, L.S., Realpe-Bonilla, T., Miller, S., Brzustowicz, L.M., Bartlett, C., Flax, J.F., 2001. Familial aggregation in specific language impairment. Journal of Speech, Language, and Hearing Research 44, 1172–1182.

Tallal, P., Miller, S.L., Bedi, G., Byma, G., Wang, X., Nagarajan, S.S., Schreiner, C., Jenkins, W.M., Merzenich, M.M., 1996. Language comprehension in language-learning impaired children improved with acoustically modified speech. Science 271, 81–84.

Tallal, P., Piercy, M., 1978. Defects of auditory perception in children with developmental dysphasia. In: Wyke, M.A. (Ed.), Developmental Dysphasia. Academic Press, New York, NY, pp. 63–84.

Tallal, P., Ross, R., Curtiss, S., 1989b. Familial aggregation in specific language impairment. Journal of Speech and Hearing Disorders 54, 167–173.

Temple, E., Deutsch, G.K., Poldrack, R.A., Miller, S.L., Tallal, P., Merzenich, M.M., Gabrieli, J.D., 2003. Neural deficits in children with dyslexia ameliorated by behavioral remediation: evidence from functional MRI. Proceedings of the National Academy of Sciences of the United States of America 100, 2860–2865.

Temple, E., Poldrack, R.A., Salidis, J., Deutsch, G.K., Tallal, P., Merzenich, M.M., Gabrieli, J.D., 2001. Disrupted neural responses to phonological and orthographic processing in dyslexic children: an fMRI study. Neuroreport 12, 299–307.

Thorndike, R.L., Hagen, E.P., Sattler, J.M., 1986. The Stanford-Binet Intelligence Scale, fourth ed. Riverside Publishing, Chicago, IL.

Tomblin, J.B., 1989. Familial concentration of developmental language impairment. Journal of Speech and Hearing Disorders 54, 287–295.

Tomblin, J.B., Buckwalter, P.R., 1998. Heritability of poor language achievement among twins. Journal of Speech, Language, and Hearing Research 41, 188–199.

Tomblin, J.B., Records, N.L., Buckwalter, P., Zhang, X., Smith, E., O'Brien, M., 1997. Prevalence of specific language impairment in kindergarten children. Journal of Speech, Language, and Hearing Research 40, 1245–1260.

Tomblin, J.B., Zhang, X., Weiss, A., Catts, H., Weismer, S.E., 2004. Dimensions of individual differences in communication skills among primary grade children. In: Rice, M.L., Warren, S.F. (Eds.), Developmental Language Disorders: From Phenotypes to Etiologies. Erlbaum, Mahwah, NJ, pp. 53–76.

Vernes, S.C., Newbury, D.F., Abrahams, B.S., Winchester, L., Nicod, J., Groszer, M., Alarcón, M., Oliver, P.L., Davies, K.E., Geschwind, D.H., Monaco, A.P., Fisher, S.E., 2008. A functional genetic link between distinct developmental language disorders. New England Journal of Medicine 359, 2337–2345.

de Villers-Sidani, E., Merzenich, M.M., 2011. Lifelong plasticity in the rat auditory cortex: basic mechanisms and role of sensory experience. Progress in Brain Research 191, 119–131.

Walker, K.M., Hall, S.E., Klein, R.M., Phillips, D.P., 2006. Development of perceptual correlates of reading performance. Brain Research 1124, 126–141.

Whitehouse, A.J., Line, E.A., Watt, H.J., Bishop, D.V., 2009. Qualitative aspects of developmental language impairment relate to language and literacy outcome in adulthood. International Journal of Language and Communication Disorders 44, 489–510.

Zatorre, R.J., Belin, P., 2001. Spectral and temporal processing in human auditory cortex. Cerebral Cortex 11, 946–953.

Language Policy

Ofelia García, Graduate Center, The City University of New York, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.A. Fishman, volume 12, pp. 8351–8355, © 2001, Elsevier Ltd.

Abstract

This article reviews the field of language policy, and presents a historical overview of the theoretical perspectives. It reviews the activities of classical language planning from the mid-twentieth century to the more critical and post-structural perspective of the field today. By presenting a framework of motivations for language policy, the article also presents different cases of types of language policy.

The field of language policy can be considered a branch of macro-sociolinguistics, also known as sociology of language. Not only has the field of language policy evolved in relation to changes in the social sciences, including sociolinguistics, and our epistemologies in the twenty-first century, but also the naming of the field itself has shifted. Originally scholars referred to the field as *language planning*, focusing on what could be done by the state to systematically promote linguistic change (Cooper, 1989; Fishman et al., 1968; Fishman, 1971; Haugen, 1959, 1966; Kaplan and Baldauf, 1997). Cooper (1989) defines language planning as "*deliberate* efforts to influence the behaviour of others with respect to the acquisition, structure, or functional allocations of their language codes" (p. 45, my italics). But post-structuralist and critical scholarship has challenged the idea that a community's way of speaking could be planned and changed by authoritative agencies. There are beliefs, practices, and even regulations that impact the ways in which people use language. Thus, the naming of the field itself has shifted to *language policy*, as it recognizes the multiple forces that influence behavior toward language (*LP*) (Ricento, 2006; Shohamy, 2006; Spolsky, 2004, 2012). Some scholars, however, have preferred to speak about *language policy and planning* (*LPP*) (Hornberger and Ricento, 1996; Hornberger, 2006) or *language policy and language planning* (LPLP) (Wright, 2004), suggesting that both terms are needed in order to capture their distinctive roles – one referring to language changes by the state or authoritative organizations, the other communicating behaviors and beliefs or attitudes toward language that shape the way that language is managed and used in society.

The Beginnings: Classic Language Planning

Language planning became an important field of study during the 1960s and 1970s, in what Ricento (2000: 206) has called *classic language planning*. Einar Haugen, a Scandinavian language specialist is credited with having used the term *language planning* for the first time in 1959 to refer to "the activity of preparing a normative orthography, grammar, and dictionary for the guidance of writers and speakers in a non-homogeneous speech community" (p. 8). The term itself was first used by Uriel Weinreich in a 1957 seminar at Columbia University (Cooper, 1989: 29).

In 1965 Joshua Fishman published his now classic article entitled "Who speaks what language to whom and when," the basis for the development of macro-sociolinguistics or sociology of language, which focuses on the social organization of language behavior, "including not only language usage per se, but also language attitudes and overt behaviors toward language and toward language users" (Fishman, 1972: 1). Fishman argued, and has continued to do so, that social action was needed on behalf of languages and its speakers, and that language could be planned just in the same way that there is economic planning.

Around the same time, in a study of Modern Norwegian, Einar Haugen defined language planning as consisting of four prongs: (1) norm selection, (2) codification, (3) implementation of language functions by spread, and (4) elaboration of functions to meet language needs (1966). Norm selection refers to the formal role that language was to have in society; for example, as an official or national language. Codification includes the choice of linguistic forms, including standardization and graphization. Implementation has to do with whether the policy would be accepted and followed. Finally, elaboration refers to continued linguistic reforms that have societal aims such as modernization, purification, and other stylistic reforms.

In 1968, Fishman, Ferguson, and Das Gupta published *Language Problems of Developing Nations*, perhaps what could be considered the first formal text in the field of language policy. It was the 'problems' of developing nations, and specifically the fact that national identity and national language did not coincide with either ethnic identity nor language use, that in many ways fueled the development of the field (Jernudd and Das Gupta, 1971). Language planning was put in the service of 'solving' the language 'problems' of the newly independent African and Asian states, by assessing the sociolinguistic situation and prescribing certain changes in how language was to be used (Das Gupta and Ferguson, 1977). Jernudd and Das Gupta (1971: 211) summarize the charge of language planning by saying that it is "a political and administrative activity for solving language problems in society." It is clear that these early language planners strongly believed in the social intervention of the state on behalf of languages and their users.

Language planning was originally studied as consisting of two types:

1. *Corpus planning* refers to changes in the linguistic form of the language itself through standardization (standardizing language forms), graphization (developing a writing system), and modernization (coining new words and terms).

2. *Status planning* refers to changes in the functions of language in order to elevate its prestige and increase the uses of a language.

These terms are attributed to the German sociolinguist, Heinz Kloss (1969).

Although described as two processes, Joshua A. Fishman insists that this distinction is clearer in theory than in practice, and that it is only possible to engage both types jointly (Fishman, 2006).

From the very beginning of language planning activities, there was questioning of whether language could indeed be planned, as reflected in the title of Rubin and Jernudd's influential early book, *Can Language be Planned?* (1971). The questioning of whether it was possible to definitely plan the form that a language takes, as well as its uses, emerged from the field itself. Einar Haugen had referred to the *ecology of language*; that is, to the "interactions between any given language and its environment" (1972: 57). Haugen defines that environment as being social and natural, in part psychological, and in part sociological. And he states: "The ecology of language is determined by the people who learn it, use it, and transmit it to others" (p. 57). By putting language in the hands of people and in the interactions between individual choices and socio-psychological possibilities and restrictions, deliberate language planning by governments and authoritative agencies takes a backward step, as the more flexible concept of language policy began to overtake the field.

From Language Planning to Language Policy

In 1989, Robert Cooper published his influential book, *Language Planning and Social Change*. With the now classic question, "What actors attempt to influence what behaviors of which people for what ends under what conditions by what means through what decision-making process with what effect?" (p. 98), Cooper changed the lens of language planning from one of mere description to one engaged in theorizing social action. Language planning, said Cooper, was not only in the hands of governments and authoritative bodies as the Académie Française, but also in the hands of individuals (like Ben Yehuda in the case of revitalization of Hebrew in Israel) and in the hands of communities and groups (such as women opposing sexist language). Cooper also introduced a third type of language planning, beyond corpus and status planning, that he termed *acquisition planning*. This type of language planning focuses on increasing the number of users of the language through two means: (1) the teaching of foreign/second languages, (2) the use of the home language as medium of instruction so as to spur language maintenance or to reacquire the language. Because of the importance of education in acquisition planning, Kaplan and Baldauf (1997) have called it language-in-education planning (LiEP), and many scholars refer to it as language education planning (Shohamy, 2006).

What is important about Cooper's contribution is that it expands the classical language planning activity. Beyond states and authoritative agencies, other agents are included as language planners, and the activity is not restricted to a single thing. As we cited in the introduction of this article, Cooper defined language planning as: "deliberate efforts to influence the behaviour of others with respect to the acquisition, structure, or functional allocations of their language codes" (p. 45). Cooper's definition talks about "influencing" linguistic 'behavior,' rather than changing it (since the activity may just modify or maintain behaviors). The emphasis in Cooper's work is on influencing behaviors and not on solving problems.

At around the same time James Tollefson (1991) published *Planning Language, Planning Inequality*. Interestingly enough, the subtitle of the book was *Language Policy in the Community*. In this important book, Tollefson advances a central idea in the shift from classical language planning to what we know today as language policy. He says:

> In this book language planning-policy means the institutionalization of language as a basis for distinctions among social groups (classes). That is, language policy is one mechanism for locating language within social structure so that language determines who has access to political power and economic resources. Language policy is one mechanism by which dominant groups establish hegemony in language use.
>
> p. 16

Tollefson points to the fact that language actions and choices are always constrained by ideological or structural (class) means, having to do with power, hegemony, and dominance. Structural constraints can block understanding of inequality, leading then to ideology. And he signals that language is an arena for struggle, since social groups exercise power through it. Tollefson adds: "Language policy is a form of disciplinary power. Its success depends in part upon the ability of the state to structure into the institutions of society the differentiation of individuals into 'insider' and 'outsiders'" (Tollefson, 1991: 207). Thus, Tollefson posits an alternative conception of language policy referring to both governmental and nongovernmental activity and focusing on the dynamic relationship between social structural relations and language practices that could result in structural equality. In keeping with Foucault's concept of governmentality, Tollefson views the operation of power as not limited to the state, but also to diverse practices.

Building on the critical work of Tollefson, and influenced by the work of Pierre Bourdieu, other scholars have maintained that all language behaviors and attitudes are always ideological and enmeshed in social systems of domination and subordination of groups, having to do not only with ethnicity, but also with class, gender, and race (Irvine and Gal, 2000). Bourdieu (1991) refers to habitus as a "system of dispositions common to all products of the same conditionings" (p. 59) by which the material form of life is "embodied and turned into second nature" (p. 63). This explains why speakers who do not control language practices that are considered 'standard' consider these to be more credible or legitimate than those they do speak. Language policy thus signals ideological positions and boundary markers of group identity for inclusion and exclusion (Kroskrity, 2000), and for constructing and effacing an other (Suleiman, 2004).

Power is said to operate in practices and ideologies and beliefs, and not only in laws and regulations. It is precisely these new understandings that have expanded the field of language policy from its original conceptualization as simply 'planning' at the state level and carried out by official language 'managers,' to one that today incorporates families,

faith communities, employers and educators (Spolsky, 2009). Spolsky points out (2009: 8):

> Many countries and institutions and social groups do not have formal or written language policies, so that the nature of their language policy must be derived from a study of their language practice or beliefs. Even where there is a formal, written language policy, its effect on language practices, is neither guaranteed nor consistent.

Thus, Spolsky (2004) expands our understandings of language policy to encompass three components:

1. *Language management* – also known as language planning, language intervention, language engineering, or language treatment, and referring to direct efforts to modify language practice beyond just government.
2. *Language practices* – the habitual pattern of selecting among varieties that make up the linguistic repertoire; that is, what people actually do.
3. *Language beliefs* or ideology – the values assigned to language and language use; that is, what people think should be done.

The first component, language management, was suggested by Jernudd (1993) to expand language planning from its emphasis on state as decision maker, to include more bottom-up and discourse-based planning. Jernudd (1997: 136) comments that language management includes:

> a broad range of different sociolinguistic situations at different levels of enlargement (from nation to firm), of a broad range of different interests and populations (from women to refugees), under widely different communicative circumstances (of media, channels, information processing), and foremost, of the different ideological and real, global and local sociopolitical conditions.

The variability of features inherent in language, the different ways in which language can be used at different times and with various interlocutors, as well as the various attitudes and beliefs held about language features and practices, is what makes language policy possible. Because choices must be made about all of this, either by individuals and groups themselves, or by outsiders, language policy has to be understood as a most important component of sociolinguistics.

Language choice and language beliefs are not neutral. As Pavlenko and Blackledge (2004: 1, 2) propose:

> [L]anguage choice and attitudes are inseparable from political arrangements, relations of power, language ideologies, and interlocutors' views of their own and others' identities. Ongoing social, economic, and political changes affect these constellations, modifying identity options offered to individuals at a given moment in history and ideologies that legitimize and value particular identities more than others.

Bringing the classical language planning strand together with the more flexible language policy strand, Shohamy (2006) points to the important distinction between 'overt policy' – having to do with the use of language to influence sociolinguistic use and norms, and thus social life – and 'covert or hidden policy' – not explicitly addressing language itself but relying on the discursive power of language to have the same social effects. Kaplan and Baldauf (1997) had already noticed that there has always been unplanned language policy and planning which goes unnoticed and unrecorded (p. 299). This has been the focus of Hornberger's work and especially of her important contribution on this topic – *Indigenous Literacies in the Americas: Language Planning from the Bottom Up* (1996). Hornberger's text and her use of the phrase 'bottom-up,' remind us that language policy does not always flow from the top-down. As Sutton and Levinson (2001: p. 3) have said, individuals and groups "engage in situated behaviors that are both constrained and enabled by existing structures, but which allow the person to exercise agency in the emerging situation." Thus, both the structure and the activity are parts of a complex process of dynamic construction in multiple directions and with multiple stakeholders. The line of power is not always linear from top to bottom since speakers have agency and language policy interacts with ideologies, global, national, and local forces. This more dynamic view of language policy, has spurred the recent focus on ethnographic studies of language policy (Hornberger and Johnson, 2007, Johnson, 2009; McCarty, 2011). These more grounded studies of language policy have taken a lens of language not as a system of structures, but of language as embodied practice, as interactions and practices, as languaging (Becker, 1995). Thus, language policy has taken up the attitudes, uses and management of languaging, and its pluralization, translanguaging (García, 2009). Translanguaging refers to the entire linguistic repertoire of bilingual people as they adjust their fluid discourse to different social, cultural, and political contexts, thus giving voice to new social realities.

Spolsky (2004) makes clear that if language policy takes on an ecological lens, two things follow:

1. Language policy must consider not just those language varieties that are recognized, but the entire sociolinguistic repertoire of the community.
2. Language policy cannot be considered except in relationship to the "social, political, economic, cultural, religious and ideological context that makes up human society" (p. 218).

But it is important to consider not only what language planning is and how it evolved, but also what have been the goals of the enterprise and why it is that this field has emerged and that these efforts have happened. We turn now to the goals and motivations for language planning, providing the context for some of the language policy efforts that we have been describing.

Goals and Motivations for Language Policy

Goals and motivations are different aspects of the same thing. Goals refer to what the action of language planning directs us toward. Motivation, however, is why we're impelled toward that goal. While classical language planning was more directed toward specific goals, the development of the field has emphasized how motivations are ideological and political. I first review the goals from a classical language planning perspective and then turn to how psychological and linguistic

motivations for language planning interact with socio-political and economic ones.

Goals: Linguistic and Social Goals

Nahir (1984) identified 11 goals for language planning:

1. *Language purification*, that is, maintaining linguistic consistency and standards of a language, usually through the development of prescriptive grammars and dictionaries.
2. *Language revival*, language revitalization and language reversal, encompassing efforts at restoring the language.
3. *Language reform*, that is, changing the orthography, spelling, lexicon, or grammar of a language in order to facilitate language use.
4. *Language standardization* for effective communication, accomplished usually through pedagogical grammars and dictionaries.
5. *Language spread*, an attempt to increase the number of speakers of a language, usually by having speakers shift to another language.
6. *Lexical modernization*, that is, expanding the capacity of a language to deal with new concepts and technology;
7. *Terminological unification*, also known as term planning, and having to do with development of equivalent terminology across geographic areas, especially terms having to do with medicine, science, industries, aviation and maritime navigation, and technology.
8. *Stylistic simplification*, attempts to make text more readable and less complex in lexicon and syntax.
9. *Interlingual communication* to facilitate communication between members of different speech communities.
10. *Language maintenance*, having to do with the preservation of a language.
11. *Auxiliary code standardization*. Here Nahir (1984) refers to "signs for the deaf, place names, and rules of transliteration and transcription" (p. 318).

All of these are important goals of the language planning enterprise and may be seen this way today. But what are the motivations for language planning? Why is it that these goals are pursued? What are the wishes and desires that motivate government, agencies, families, special groups, and individuals to pursue, adopt, or negotiate language policies? These questions are better answered by the scholarship of Joshua A. Fishman.

Motivations: Independence versus Interdependence

Fishman (2000, 2006) has pointed out that all efforts of language planning boil down to a super-factor of *independence versus interdependence*. By independence Fishman means whether the social group wants to be considered self-sufficient and autonomous of another. Interdependence refers to wanting to be affiliated with another social group. He develops a motivational theory for language policy that is divided into these two super-factors.

Language Policy for Independence

According to Fishman, *language policy for independence* is carried out through four processes: (1) ausbau, (2) uniqueness, (3) purification, and (4) classicization. Some of the cases that follow are from Fishman (2006), whereas others are from other sources, and especially from García and Fishman (2012).

1. *Ausbau*. This process builds on the work of Heinz Kloss (1967) who spoke about *abstandsprache* (language by virtue of linguistic distance) and *ausbausprache* (language by virtue of development). For Fishman the process of ausbau refers to language policy where the purpose is to build a language away from another; that is, to distance it from another. To contextualize this process, I provide the reader with some examples.
   a. Noah Webster developed the American English dictionary in 1783 as a way of distancing American English from British English.
   b. Urdu and Hindi adopted different scripts to differentiate one from the other. Hindi uses the Devanagari script, while Urdu uses a modified version of the Perso-Arabic script, with Nasta'liq the preferred calligraphic style. Although Hindi borrows strongly from Sanskrit, Urdu draws on Persian and Arabic.
   c. Norway has spurred several reforms in order to differentiate Norwegian from Danish. The result has been two written standards – Bokmål (book language) and Nynorsk (new Norwegian).
   d. In Central Asia the case of Azerbaijani/Turkish and Tajik/Farsi are important cases of *ausbau* by which languages that had been previously considered to be one language are now claimed to be autonomous from others, as they are linked to specific ethnic identities, thus claiming to be separate nations needing a state. For example, shortly after independence in Azerbaijan, the state language was first identified, in 1992, as 'Turkish' (türk dili). But there were also competing names proposed, 'Azerbaijani Turkic' and 'Azerbaijani.' In the 1995 constitution, the name 'Azerbaijani' was chosen. Thus, renaming the language emphasized the link between language and a separate Azerbaijani nationhood. Similarly, in Tajikistan, the 1989 language law referred to the language as 'Tajik (Farsi).' But in 1999 the word 'Farsi' was removed from the state language law, emphasizing separateness and autonomy as a nation of Tajikistan.
   e. This same process of ausbau has operated in the Slavic World (Hroch, 2010) by naming Serbian, Croatian, Bosnian, and Montenegrin, for example, when they were all previously considered Serbo-Croatian.
2. *Uniqueness*. Some language policy efforts aim to make a language more heavily one's own.
   a. Johannes Aavik (1880–1973) selected 'artificial' syllables for Estonian, ensuring that it would not be mistaken for Finnish or swallowed up by Russian.
3. *Purification*. This type of language policy is driven by fear of the language being contaminated. The first two examples have to do with languages, but the next two have to do with languaging; that is, the action of using language.
   a. The Loi Toubon in France that was passed in 1994 was meant to protect French from English. The law states: "By virtue of the Constitution, French is the language of the Republic, and the French language is a fundamental element of the character and heritage of France." As such, French is mandated in all official government

publications, advertisements, workplace, commercial contracts, and government-financed schools.

b. The promotion of Hebrew as a result of the revitalization that accompanied the formation of the state of Israel was accompanied by a ban on Yiddish, which was the language of most Ashkenazi Jews in the state. The Pure Language Society was set up in 1889 to stamp out Yiddish because of fear that it would contaminate the Hebrew of Israel.

c. The efforts of language academies such as the Académie Française and the Real Academia Española in the seventeenth and eighteenth centuries. This last one had a motto that said to "limpia, fija y da esplendor" (cleanses, stabilizes, and gives splendor).

d.The management of bad language, as with swearing or crude language, obscenities, profane language, blasphemy, racist, and sexist language.

4. *Classicization* refers to the policies that attempt to keep the language very close to its classical source. Examples follow:
   a. As explained in point 1, Hindi draws on Sanskrit, whereas Urdu draws on Arabic and Persian. Both languages attempt in this way to get closer to their classical source.
   b. Modern Tamil has attempted to shed Sanskrit loanwords in order to show continuity with its classical past as Old Tamil. This was the platform of the Pure Tamil movement of the early twentieth century.

Language Policy for Interdependence

In contrast to the four processes identified above, *language policy for interdependence* follows four different courses of action: (1) einbau, (2) internationalization, (3) regionalization, and (4) vernacularization. Again, examples are from Fishman (2006) and from García and Fishman (2012).

1. *Einbau*. This process is the opposite of ausbau. The language development is not meant to distance the languages from each other; instead, it brings them closer. There are many examples of this process:
   a. The Republic of Moldova declared itself independent in 1991 after the dissolution of the Soviet Union. The Language Law speaks of a Moldovan-Romanian linguistic identity, and little by little the few distinctions that once existed between Romanian and Moldovan are disappearing. Linguists and politicians have been acknowledging that the two languages are identical, although many continue to call their language Moldovan.
   b. Chinese authorities insist that the many languages spoken by Han Chinese are all 'dialects' of Chinese, despite the fact that they are all mutually unintelligible and could be considered different languages.
2. *Internationalization*. Some language policies are meant to bring language into compliance with the international community or as a focus of interest in the international community.
   a. Under the rule of Mustafa Kemal Atatürk, Turkish was westernized by adopting the Latin script, and adopting many French influences. In 1928, the Arabic-based Ottoman Turkish alphabet was replaced by the Latin alphabet. Afterward, a Turkish Language Association was established to replace loanwords of Arabic and Persian origin with Turkish equivalents.
   b. Although Kazakh, Kyrgyz, and Tajik are still written in Cyrillic, Azerbaijani, Turkmen, and Uzbek are written in the Latin alphabet, emphasizing their growing internationalization.
   c. The choice of the Latin script for the Oromo language in Ethiopia, as opposed to the Ge'ez used for Amharic and Tigrinya, has put Oromo in an internationalization path.
   d. The British Council and the US government have supported the greater use of English in international communication (Phillipson, 1992). The Goethe Institute, the Japanese Foundation, and the Instituto Cervantes do the same for German, Japanese, and Spanish, respectively.
3. *Regionalization*. Some language policies are meant to bring people from similar regions together.
   a. Bahasa Malaysia and Bahasa Indonesia derive from Bahasa Melayu (Malay) and despite their differences, borrow broadly from each other.
4. *Vernacularization*. Language policies that favor popular usage have the capacity to bring languages closer together.

But not all language planning has to do with choosing between independence and interdependence. Beyond the model proposed by Fishman (2006), we can identify motivations as stemming from three other supra-factors – exclusion, inclusion and cultural autonomy. I turn now to the ideological aspects of inclusion and exclusion that motivates language policy.

Motivations: Exclusion, Inclusion, and Cultural Autonomy

Seen from an ideological angle, language policies respond to three axes that have more to do with the speakers of a language than with language itself – exclusion, inclusion, and cultural autonomy.

Language Policy for Exclusion

When the motivation for language policy is exclusion, it is carried out through three processes:

1. *Shaming*. These language policies, often the result of language-in-education policies, result in stigmatization and exclusion of minority groups. They were prevalent in colonial contexts, and continue to be exercised in many states today.
   a. In Ghana, students who spoke African languages during colonial times were forced to wear a sign on their chest that said: "I'm stupid, I spoke a vernacular on the school's premises today."
   b. Indigenous children throughout the world are subjected to monolingual education policies with a purpose of shaming students who speak these languages. This is also the case of autochthonous minorities throughout Europe, of immigrant children throughout the world, and of other minorities. This shaming also includes those who speak varieties of language that are stigmatized, for example, African Americans in the USA.
   c. Until recently, Deaf children were educated through oralism which insisted on speech production and forbade them to use sign language.

   d. Latino students in the USA are often made to think that they speak a corrupted 'Spanglish,' when the translanguaging they perform is the discursive norm of all bilingual communities.
2. *Othering*. Language minorities today are not openly excluded from society and education with 'sticks.' Instead, other systems exclude those who do not conform linguistically.
   a. Students in the UK and the USA are excluded from educational opportunities when they do not do well in standardized examinations that test the use of standard English.
   b. Immigrants who do not speak the language of the state are excluded from citizenship opportunities by language tests. For example, the Dutch government requires that immigrant applicants take a Dutch language proficiency exam.
   c. In the UK and the USA, the use of standard English in education and assessment responds to a political ideology of monolingualism. In the UK, the Swann Report (1985) established policy that there should not be any 'separate provision' for language minority children and that bilingual education should not be supported. In the USA the No Child Left Behind Act requires states to assess student development of standard English.
3. *Silencing*. Language policy often attempts to silence other voices, erasing the power of people.
   a. Basing its dominance on its shape as a perfect hexagon, France has maintained that only French is tied to French identity. As a result, the other languages of France – Basque, Breton, Catalan, Corsican, German, Occitan, and Flemish – have been silenced until recently. The presence of regional languages in the regions where they are spoken was recognized by the French constitution (Article 75-1, 23 July 2008). Nevertheless, France's hesitation to ratify the European Charter for Regional or Minority Languages has to do with its fear to extend the expression of French identity to languages other than French.
   b. The banning of signing at schools for the deaf in the 1880s was motivated by an ideology that oral speech was the only way to pray to God, and thus, for deaf children to be saved.
   c. When Algeria obtained independence from France in 1962, it declared Classical Arabic the sole official and national language, rejecting Algerian Arabic and the Berber languages (also known as *Tamazight*).
   d. The Education Act of 1870 in Wales prohibited Welsh children from speaking Welsh at school, and required them to learn English.
   e. In the USA, California, Arizona, and Massachusetts have passed legislation banning bilingual education and the use of languages other than English as a means of instruction. And the word 'bilingual' has been silenced in all public discourse.

Language Policy for Inclusion

Even though most language education policy efforts have focused on exclusion of groups and peoples, we can also identify policies that do the opposite; that is, that include. *Language policy for inclusion* is carried out through three different means:

1. Bilingualism/multilingualism
   a. In Canada, the declaration of the Official Languages Act (1969), made English and French languages of Canadian identity.
   b. Singapore has adopted four official languages – Mandarin, Malay, Tamil, and English. The first three were chosen for identity purposes of the three major ethno-linguistic groups. And yet, these languages are mother tongues of only 17% of the Malay, Indian, and Chinese population. Malay is the mother tongue of 12% of Malays, Tamil of 5% of Indians, and Mandarin of .1% of the Chinese population (Rubdy, 2005: 57). English, however, was selected as a neutral 'other language' to serve as the language of interethnic communication.
2. Language teaching
   a. Some languages are taught to give advantage to people. This is the case especially of English throughout the world. In Africa, it is the case of Swahili in Eastern Africa, Wolof in Senegal, Akan in Ghana, Hausa in Northern Nigeria, and Lingala in Zaire.
   b. In some bilingual education programs, translanguaging is accepted as the norm of language practices in bilingual communities, and used to teach.
3. Diversity and pluralism
   a. In the 1970s and 1980s, many advocated for the use of nonsexist language, a use that is the norm today.
   b. In *Lau versus Nichols* (1974), the US Supreme Court ruled that educating English Language Learners in a language they did not understand was not equal educational opportunity and that something had to be done about the language of instruction.

Language Policy for Cultural Autonomy

Cultural autonomy refers to choices people make to be free "from domination by outsiders, not freedom to be independent" (García and Fishman, 2012: 143).

1. Since the Statute of Autonomy that was passed in 1979 after Francisco Franco's death, Catalan has been the only language used in the government of the *Generalitat de Catalunya* (the government), and must be used as medium of instruction, since it is the 'lengua propia' (own language) of the people. Since 1983, Catalonia has followed a policy of 'linguistic normalization' intended to protect and extend the use of Catalan.
2. In 1974 Québec declared French the sole official language of Québec, the language of Québecois identity, and Bill 101, the Charter of the French Language in 1977, made French the language of work, business, and education in Québec.

Conclusion

Depending on the ecological relationship between the social, historical, and political context and language, language policies act differently. The global world of today, with its interconnectedness made possible by technology produces different kinds of

language policies. Often the language policies are in conflict with each other, responding to the different wishes and motivations of sectors in society. The field has evolved from one where descriptions abounded and the state was in charge to one that is developing a theory of social action motivated by the struggle created by power and inequality. With a critical ethnographic lens, language policy is now equipped to teach us much about the ideological, socio-structural, and historical bases of language use, language attitudes, and language management.

See also: Areal Linguistics; Code Switching: Linguistic; Corpus Linguistics; Dialectology; Indo-European Languages; Language Contact; Language and Social Class; Linguistic Typology; Pidgin and Creole Languages; Sign Language: Psychological and Neural Aspects.

Bibliography

Becker, A.L., 1995. Beyond Translation: Essays toward a Modern Philosophy. University of Michigan Press, Ann Arbor.

Bourdieu, P., 1991. Language and Symbolic Power. Harvard University Press, Cambridge, MA.

Cooper, R.L., 1989. Language Planning and Social Change. Cambridge University Press, Cambridge.

Das Gupta, J., Ferguson, C., 1977. Problems of language planning. In: Rubin, J., Jernudd, B., Das Gupta, J., Fishman, J.A., Ferguson, C. (Eds.), Language Planning Processes. Mouton de Gruyter, Berlin, pp. 3–8.

Fishman, J.A., 1965. Who speaks what language to whom and when? La Linguistique 2, 67–88.

Fishman, J.A., 1971. In: Rubin, J., Jernudd, B.H. (Eds.), The Impact of Nationalism on Language Planning: Some Comparisons Between early Twentieth-Century Europe and More Recent Years in South and Southeast Asia, pp. 3–20.

Fishman, J.A., 1972. The Sociology of Language: An Interdisciplinary Social Science Approach to Language in Society. Newbury House Publishers, Rowley, MA.

Fishman, J.A., 2000. The status agenda in corpus planning. In: Lambert, R.D., Shohamy, E. (Eds.), Language Policy and Pedagogy: Essays in Honor of A. Ronald Walton. John Benjamins, Amsterdam, pp. 43–51.

Fishman, J.A., 2006. Do Not Leave Your Language Alone: The Hidden Status Agendas within Corpus Planning in Language Policy. Erlbaum, Mahwah, NJ.

Fishman, J.A., Ferguson, C., Das Gupta, J. (Eds.), 1968. Language Problems of Developing Nations. John Wiley and Sons, New York.

Fishman, J.A., García, O. (Eds.), 2010. Handbook of Language and Ethnic Identity Disciplinary and Regional Perspectives, second ed., vol. 1. Oxford University Press, Oxford.

Fishman, J.A., García, O. (Eds.), 2011. Handbook of language and ethnic identity. The Success and Failure Continuum, vol. 2. Oxford University Press, Oxford.

García, O., 2009. Bilingual Education in the 21st Century: A Global Perspective. Blackwell/Wiley, Malden, MA and Oxford.

García, O., 2010. In: Fishman, J.A., García, O. (Eds.), Languaging and Ethnifying, pp. 519–534.

García, O., Fishman, J.A., 2012. Power-sharing and cultural autonomy: some sociolinguistic principles. International Journal of the Sociology of Language 213, 143–148.

Haugen, E., 1959. Planning for a standard language in Norway. Anthropological Linguistics 1 (3), 8–21.

Haugen, E., 1966. Linguistics and language planning. In: Bright, W. (Ed.), Sociolinguistics. Mouton, The Hague.

Haugen, E., 1972. The Ecology of Language. Stanford University press, Stanford.

Hornberger, N., 1996. Indigenous Literacies in the Americas: Language Planning from the Bottom Up. Mouton de Gruyter, Berlin.

Hornberger, N., 2006. Frameworks and models in language policy and planning. In: Ricenton, T. (Ed.), Language Policy: Theory and Method. Blackwell, Malden, MA, pp. 24–41.

Hornberger, N., Johnson, D.C., 2007. Slicing the onion ethnographically: layers and spaces in multilingual language education policy and practice. TESOL Quarterly 41, 509–532.

Hornberger, N., Ricento, T.K. (Eds.), 1996. Language Planning and Policy. TESOL Quarterly 30 (3) (special issue).

Hroch, M., 2010. The Slavic World. In: Fishman, J.A., García, O. (Eds.), Handbook of Language and Ethnic Identity, vol. 1. Oxford University Press, Oxford and New York, pp. 269–284.

Irvine, J., Gal, S., 2000. Language ideology and linguistic differentiation. In: Kroskrity, P. (Ed.), Regimes of Language: Ideologies, Polities and Identities. School of American Research Press, and Oxford: James Currey, Santa Fe, NM, pp. 34–84.

Jernudd, B.H., 1993. Language planning form a management perspective: an interpretation of findings. In: Jahr, E.H. (Ed.), Language Conflict and Language Planning. Mouton de Gruyter, Berlin, pp. 133–142.

Jernudd, B.H., 1997. The revolution of sociolinguistics. In: Paulston, C.B., Tucker, R.G. (Eds.), The Early Days of Sociolinguistics. Summer Institute of Linguistics, Dallas, pp. 131–138.

Jernudd, B.H., Das Gupta, J., 1971. Towards a theory of language planning. In: Rubin, J., Jernudd, B. (Eds.), Can Language Be Planned? University Press of Hawaii, Honolulu, pp. 195–216.

Johnson, D.C., 2009. Ethnography of language policy. Language Policy 8, 139–159.

Kaplan, R.B., Baldauf, R.B., 1997. Language Planning: From Practice to Theory. Multilingual Matters, Clevedon, UK.

Kloss, H., 1967. Types of multilingual communities: a discussion of ten variables. In: Lieberson, S. (Ed.), Explorations in Sociolinguistics. Indiana U., Bloomington, pp. 7–17. Originally published in The Sociolinguistic Inquiry, 36(2), Spring (1966).

Kloss, H., 1969. Research Possibilities on Group Bilingualism: A Report. International Center for Research on Bilingualism, Quebec.

Kroskrity, P. (Ed.), 2000. Regimes of Language: Ideologies, Polities and Identities. School of American Research Press, & Oxford: James Currey, Santa Fe, NM.

McCarty, T.L. (Ed.), 2011. Ethnography and Language Policy. Routledge, New York and London.

Nahir, M., 1984. Language planning goals: a classification. Language Problems and Language Planning 8, 294–327.

Pavlenko, A., Blackledge, A., 2004. Negotiation of Identities in Multilingual Contexts. Multilingual Matters, Clevedon, UK.

Phillipson, R., 1992. Linguistic Imperialism. Oxford University Press, Oxford.

Ricento, T., 2000. Historical and theoretical perspectives in language policy and planning. Journal of Sociolinguistics 4, 196–213.

Ricento, T. (Ed.), 2006. An Introduction to Language Policy: Theory and Method. Blackwell, Malden, MA.

Rubin, J., Jernudd, B. (Eds.), 1971. Can Language Be Planned? Sociolinguistic Theory and Practice for Developing Nations. East-West Center and University of Hawaii Press, Honolulu.

Rubdy, R., 2005. Remaking Singapore for the New Age: official ideology and the realities of practice in language-in-education. In: Lin, A., Martin, P.W. (Eds.), Decolonisation, Globalisation. Language-in-Education Policy and Practice. Multilingual Matters, Clevedon, UK, pp. 55–73.

Shohamy, E., 2006. Language Policy: Hidden Agendas and New Approaches. Routledge, London and New York.

Spolsky, B., 2004. Language Policy. Cambridge University press, Cambridge.

Spolsky, B., 2009. Language Management. Cambridge University press, Cambridge.

Spolsky, B. (Ed.), 2012. The Cambridge Handbook of Language Policy. Cambridge University Press, Cambridge.

Suleiman, Y., 2004. A War of Words: Language and Conflict in the Middle East. Cambridge University Press, Cambridge.

Sutton, M., Levinson, B.A.U. (Eds.), 2001. Policy as Practice: Toward a Comparative Sociocultural Analysis of Educational Policy. Ablex, Westport, CT.

Tollefson, J.W., 1991. Planning Language, Planning Inequality: Language Policy in the Community. Longman, London.

Woolard, K., Schieffelin, B., 1994. Language ideology. Annual Review of Anthropology 23, 55–82.

Wright, S., 2004. Language Policy and Language Planning: From Nationalism to Globalisation. Palgrave, New York.

Language Processes in Delirium and Dementia

Joël Macoir, Faculté de Médecine, Université Laval, Québec, QC, Canada
Yves Turgeon, Clinique Médicale Carleton-sur-Mer, Carleton-sur-Mer, QC, Canada
Robert Laforce, Jr., Clinique Interdisciplinaire de Mémoire, Québec, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, we report current data on language and communication disorders in the most frequent primary progressive syndromes of dementia, namely Alzheimer's disease, vascular cognitive impairment, and dementia with Lewy bodies. Language deficits are also the core features of three variants of primary progressive aphasia, namely the nonfluent/agrammatic, semantic, and logopenic variants. This survey reveals that, like other cognitive functions, language is highly vulnerable to delirium and neurodegenerative diseases. In some of them, language deficits result from impairment to linguistic processes per se. In others, they are the direct consequence of impairments affecting working memory and executive functions.

Introduction

Delirium is an acute state of confusion characterized by disturbances in consciousness and cognitive functioning. Although delirium is known to affect mainly attentional and perceptual abilities, language disturbances have also been reported. In clinical settings, delirium must be differentiated from dementia, a syndrome characterized by multiple deficits in cognition, which is the direct consequence of physiological changes. Dementia affects attention, perception, praxis, and higher mental functions such as episodic memory and executive functions. The core feature of dementia is impairment of memory, but specific language deficits as well as language difficulties caused by impairments of attention and/or executive functions are also found in the major forms of dementia. In this article, we briefly present the general characteristics of delirium and major forms of dementia. In each section, we also discuss disorders of language associated with each condition.

Language Processes in Delirium

Acute confusional state is used interchangeably with delirium and both conditions have been referred to as transient cognitive disorders. Delirium typically has a sudden onset, sharp rise, fluctuating course, and brief duration. Symptoms may arise abruptly (e.g., following a stroke or head injury) or can be gradual (e.g., in older adults receiving multiple medications). Reports on the onset and course of delirium are numerous and some have maintained that delirium is worse at night, a phenomenon known as sundowning (Evans, 1987). In general, delirium develops over periods of time that vary from a few hours to a few days. If left untreated, symptoms may last several days or weeks, even months, especially in individuals with coexisting cognitive disorders due to a persistent or evolving medical condition.

A number of conditions can cause delirium. The most common etiologies for delirium include chemical intoxication (e.g., alcohol intoxication and withdrawal), metabolic disorders (e.g., hypoglycemia), structural brain diseases (e.g., vascular infarction), infections (e.g., anticholinergic), and environmental factors (e.g., prior recent acute-care hospital admission).

Current diagnostic criteria for delirium are more commonly based on the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision* (DSM-IV-TR) (American Psychiatric Association, 2000). The DSM-IV-TR criteria focus on three essential features of delirium: disturbances of consciousness (i.e., reduced awareness of the environment) with reduced attention (Criterion A); a change in cognition or the development of a perceptual disturbance that is not accounted for by a preexisting, established, or evolving dementia (Criterion B); and a disturbance that develops over a short period of time and fluctuates over the course of the day (Criterion C). Criterion D refers to etiological features of delirium.

Delirium is ordinarily associated with changes in sleep–wake patterns, disorientation to time and space, disordered or incoherent speech, visual hallucinations, and marked alterations in psychomotor activity. Individuals suffering from delirium commonly display neuropsychological deficits in tasks involving attention and concentration, memory, language, and spatial processes. Evidence suggests that spatial and executive impairments are key features of delirium (Inouye et al., 2006).

Chédru and Geschwind (1972) provided one of the first descriptive studies of cognitive disorders in delirium including speech and language. These authors reported that impaired reading, writing, and verbal fluency were the most common linguistic abnormalities in delirium. They also determined that individuals suffering from delirium exhibited naming and repetition abnormalities but that spontaneous speech was relatively preserved. In 1994, Wallesch and Hundsalz compared a group of individuals suffering from delirium with a group diagnosed with probable Alzheimer's disease (AD) on naming and language comprehension tasks. They found a comparable number of naming errors in both groups but unrelated naming errors were significantly more frequent in the delirium group. These authors also showed that some of the overall unrelated naming errors in participants presenting with delirium were caused by visual misperceptions, thus supporting nonlinguistic influences on language.

A number of studies have found writing deficits in participants presenting with delirium. Macleod and Whitehead (1997) have shown that writing is the most impaired of the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54031-8

language modalities in delirium. In a 6-month study, they compared the performance of a small group of individuals diagnosed with acute delirium with that of a control group. Their results provided further support for the predominance of writing impairments in delirium, which are mainly characterized by motor agraphia and the reduplication of letters. They also showed that reading was the least impaired of linguistic functions. Aakerlund and Rosenberg (1994) also found that all individuals who developed postoperative delirium had severely impaired writing, with features such as reluctance to write, motor disability, and spatial disturbances. Their results suggest that assessment of writing ability might be useful in the diagnosis of delirium.

Language Processes in Dementia

Diagnostic classifications of dementia syndromes identify various subtypes based on characteristics such as typical symptoms at presentation, progression and course of the disease, psychiatric and behavioral features as well as presumed causes (e.g., general medical condition, persistent effects of a substance, multiple etiologies). Because of the absence of biological markers or simple diagnostic methods, early detection of dementia relies on various assessments performed to rule out other possible causes and identify specific aspects of a disease (i.e., differential diagnosis). Neuropsychological testing plays an important role in the assessment of individuals with cognitive impairment. A large number of studies have sought to identify neuropsychological features that distinguish the different forms of dementia. What emerges from this work are descriptions of cognitive functioning, which contain some distinctions that are useful for the differential diagnosis of dementia. For example, there are remarkable differences between patients with AD and vascular cognitive impairment (VCI) with respect to long-term verbal memory and executive functions. However, these patients may present with similar language, constructional ability, attention, and memory deficits. Moreover, overlap in cognitive deficits is often seen between syndromes, possibly due to the heterogeneity of neuropsychological manifestations in the early stages of the disease (e.g., Benedet et al., 2012), as well as methodological problems, including the application of diagnostic criteria, selection of neuropsychological tests (aimed at large cognitive domains instead of more precise components of cognition), and lack of adequate matching of patients on age and education. In addition, the definition of dementia continues to evolve and recent criteria have made it possible to distinguish between different presentations of primary progressive aphasia (PPA).

With respect to language impairments, neuroepidemiological and neuropsychological studies propose clinical linguistic profiles usually associated with common forms of dementias. Recent neuropsycholinguistic studies have also made a major contribution to better characterization of language deficits in dementia by specifically identifying functional localization of impaired and preserved processing components and subcomponents of the linguistic processing system.

In the following sections we will review current data on language and communication disorders in the most frequent primary progressive syndromes of dementia, namely AD, VCI, dementia with Lewy bodies (DLB), as well as three forms of PPA, namely the nonfluent/agrammatic variant (nfvPPA), semantic variant (svPPA), and logopenic variant (lvPPA). The more salient characteristics of language disorders in these diseases are presented in Table 1.

Table 1 Salient characteristics of language disorders in the major forms of dementia

Language production	*Language comprehension*	*Reading*	*Spelling*	*Semantics*
1. Alzheimer's disease (early and moderate stages of the disease)				
Anomia	Deficit for words Possible deficit for sentences	Surface dyslexia	Surface dysgraphia	Access deficit and/or degraded semantic representations
2. Vascular cognitive impairment[a]				
Anomia	Deficit for sentences	Acquired dyslexia not specified	Acquired central dysgraphia Possible peripheral dysgraphia	Access deficit and/or degraded semantic representations
3. Dementia with Lewy bodies				
Anomia	Deficit for sentences	Unknown	Unknown	Access deficit and/or degraded semantic representations
4. Nonfluent/agrammatic variant of primary progressive aphasia				
Anomia Apraxia of speech Agrammatism	Deficit for syntactically complex sentences	Preserved in the early stages	Deep dysgraphia Agrammatism	Unimpaired
5. Semantic variant of primary progressive aphasia				
Anomia	Deficit for words	Surface dyslexia	Surface dysgraphia	Degraded semantic representations
6. Logopenic variant of primary progressive aphasia				
Anomia Impairment in repetition of long and complex sentences Phonological errors	Deficit for sentences	Phonological dyslexia	Phonological dysgraphia	Unimpaired

[a]Language deficits in vascular cognitive impairment generally depend on the location of prior strokes or white matter lesions.

Language Processes in AD

McKhann et al. (2011) recently revised the classification criteria for dementia due to AD. The diagnosis of probable AD requires that an individual present with a dementia syndrome established through a combination of history-taking from the patient and a knowledgeable informant as well as an objective cognitive assessment performed with a screening test or a neuropsychological assessment battery. The authors also differentiate between two types of AD presentation, according to the most prominent cognitive deficits. In the amnestic form, the most common presentation of AD, the deficits mainly include impairment in learning and recall of recently learned information. In the nonamnestic form, patients may present with prominent visuospatial (visual agnosia, prosopagnosia, simultagnosia, and alexia), executive function (impaired reasoning, judgment, and problem solving), or language (word-finding problems) deficits. This represents a major evolution from previous criteria where only the amnestic subtype of AD was considered.

Similar to other dementing syndromes, AD does not begin with global impairment in cognitive functions but usually progresses through different stages. The most pervasive characteristic of AD is undoubtedly progressive episodic memory loss associated with a decline in other cognitive areas. With respect to language, the typical course of AD is characterized mainly by: (1) word-finding problems in the early stages of the disease, (2) severe reduction, stereotyped utterances, important anomia, agraphia, and comprehension difficulties in moderate stages of the disease, and (3) impairment of all linguistic abilities in the severe and final stages. By using cognitive models of language processing, neuropsycholinguistic studies have significantly contributed to the identification of deficits affecting specific linguistic abilities in AD.

According to these studies, the main impact of the disease is on semantic memory, a component of explicit long-term memory that contains permanent information related to objects, concepts, words, and their meanings. Because of its central role in the processing of language, the deficit affecting semantic memory leads to substantial difficulties in word comprehension and production. Therefore, when tested with comprehensive batteries of language tests, individuals with AD usually show a poor performance in various tasks such as picture naming, word–picture matching, and semantic questionnaires, even in the early stages of the disease. Moreover, the consistency for individual items across tests (Hodges and Patterson, 1995; i.e., the same items are preserved or impaired across tests) and the consistency of impairment (Hodges et al., 1992b; i.e., the same deficit is observed in production and comprehension using visually and verbal-based materials) are suggestive of a deterioration in semantic representations instead of impaired access to preserved semantic representations (Rogers et al., 2006). This question is controversial, however, and other researchers have reported experimental data that tend to support a normal organization of semantic memory and an impaired ability to consciously retrieve items from the semantic network, at least in early stage AD (e.g., Rogers and Friedman, 2008). The relative preservation of superordinate knowledge as compared to specific and distinctive knowledge (e.g., knowing that a zebra is an animal but not that it has stripes or lives in Africa) is also suggestive of a partially degraded semantic network, combined with a deficit of explicit retrieval (Duarte et al., 2009). Furthermore, with respect to semantic memory, studies have also investigated whether the semantic loss in AD affects some categories more than others. Most of these studies conclude that individuals with AD perform worse on biological concepts (animals, fruits, vegetables, etc.) than on artifacts (vehicles, furniture, tools, etc.) (e.g., Garrard et al., 2001).

Articulatory, phonological, and syntactic abilities are usually considered unscathed until the final stage of AD. Some studies, however, showed that disease progression might impact these abilities. For example, Croot and her colleagues (Croot et al., 2000) have shown that in less typical cases of AD, phonological and articulatory abilities may be impaired, sometimes as a selective deficit. Syntactic processes required for sentence production (e.g., Bates et al., 1995) and sentence comprehension (e.g., Waters and Caplan, 1997) may also be compromised in some cases of AD, even in the early stages of the disease. Whether this deficit is really syntactic in nature or is more explicable in terms of a working memory disorder or in terms of semantic interpretative processing remains unresolved.

Studies on reading and writing impairments in AD also reflect a relative lack of consistency in findings. According to the standard clinical portrait of AD, reading and spelling abilities are considered largely preserved in patients, even at rather advanced stages of the disease. In contrast to this classical conception, several single-case and group studies using cognitive neuropsychological or parallel-distributed-processing models have shown that the ability to read aloud and to spell words is frequently affected in AD. These patients often demonstrate characteristics of surface dyslexia and surface dysgraphia: they are better at reading/spelling orthographically regular than irregular words (i.e., words with exceptional or unpredictable sound-to-spelling and spelling-to-sound correspondences), they show a preserved ability to read/write nonwords, and they tend to produce regularization errors in reading (e.g., bread → brid) and phonologically plausible errors in spelling (e.g., crane → crain). Group studies, however, have shown that different patterns of written language impairment may be observed in AD, especially according to the progression of the disease (Forbes et al., 2004; for a review on agraphia, see Neils-Strunjas et al., 2006; for a review on alexia, see Noble et al., 2000).

Language Processes in VCI

Recently, Gorelick et al. (2011) provided a new statement about vascular contributions to cognitive impairment and dementia, according to which VCI "represents the spectrum of cognitive impairment associated with frank stroke, vascular brain injury, or subclinical disease ranging from the least severe (i.e. vascular mild cognitive impairment) to the most severe (i.e. vascular dementia; VaD) clinical manifestations" (p. 2676). VCI may result from an array of causes. It includes all dementias following ischemic or hemorrhagic cerebrovascular accidents, i.e., single vascular insult (most typically to a critical area of the brain), repeated cerebral infarction (also referred to as multi-infarct dementia), and chronic ischemia without discrete infarction. According to Gorelick et al. (2011),

a diagnosis of probable VaD requires: (1) the demonstration of a loss of cognitive ability, associated with imaging evidence of cerebrovascular disease; and (2) the absence of a history of gradually progressive cognitive deficits before or after the stroke that suggests the presence of a nonvascular neurodegenerative disorder.

In contrast with the extensive literature on language deficits in AD, less is known about the patterns of cognitive deterioration in other forms of dementia. Consequently, the identification of clear neuropsychological differences between the major progressive dementias remains difficult to establish. In VCI, for example, most of the reported descriptions related to cognitive deficits concern behavioral manifestations and large cognitive domains. The most commonly reported cognitive impairments in VCI are bradyphrenia (slowed cognition), dysexecutive symptoms (e.g., difficulty in problem solving and planning), and decreased initiation and spontaneity. There are very few studies that have systematically described language disorders in VCI. Most of them were conducted in an attempt to differentiate VCI from AD. Many of these studies suggested that there was no significant difference in language function between the two syndromes (e.g., Fahlander et al., 2002). However, contradictory results were reported by other authors. For example, some reported that patients with VCI performed better than individuals with AD in picture naming tasks (Barr et al., 1992), while others found no differences between groups (Graham et al., 2004b) or the opposite pattern (Lukatela et al., 1998). With respect to semantic processing, individuals with VCI appeared to be more affected than those with AD (Graham et al., 2004a). Controversial results were also reported with respect to syntactic processing in VCI. According to most studies, the comprehension of complex sentences is affected in people with dementia because of impaired working memory and executive functioning. In some of these studies (e.g., Vuorinen et al., 2000), there was little or no difference in syntactic comprehension between individuals with AD and VCI, while in others (Grossman et al., 1996), individuals with VCI performed better than those with AD.

To date, very little is known about the nature and prevalence of reading and writing deficits in VCI. A few recent studies that included reading and writing tasks in the neuropsychological assessment found that both abilities appeared to be impaired in VCI. For example, using the Clinical Dementia Rating scale in newly diagnosed patients, Chan et al. (2008) found more reading difficulty in VCI than in AD. Other studies, however, found no differences between the two groups in reading tasks (Vuorinen et al., 2000). For written spelling, a few investigations (Forbes and Venneri, 2003) are suggestive of peripheral impairments (e.g., difficulty in writing letters and copying sentences). Others found that, as compared to individuals with AD, patients with VCI produced more spelling errors and grammatically less complex sentences, therefore suggesting a central origin to the deficit (Graham, 2000).

Finally, as compared to AD, the spontaneous speech of individuals with VCI is less empty and conveys more information but they tend to produce shorter and less grammatically complex phrases (Powell et al., 1988). The same authors have also shown that patients with VCI often present with abnormal prosody and articulation.

Language Processes in DLB

DLB is a frequent form of dementia, commonly regarded as a Parkinson-plus syndrome that has the clinical and pathological features of both AD and Parkinson's disease. In 2005, a DLB Consortium revised the criteria for the clinical and pathologic diagnosis of DLB (McKeith et al., 2005). According to these criteria, the central clinical diagnostic feature of DLB is the presence of dementia, as well as deficits on tests of attention, executive function, and visuospatial ability whereas memory could remain unimpaired, at least in the early stages of the disease. Two of the following three core features are sufficient for a diagnosis of probable DLB: (1) fluctuating cognition with pronounced variations in attention and alertness, (2) recurrent visual hallucinations that are typically well formed and detailed, and (3) spontaneous features of Parkinsonism. Moreover, the presence of one or more suggestive features (e.g., severe neuroleptic sensitivity) can reinforce the diagnosis of DLB.

There are few papers on the cognitive symptoms of DLB (e.g., Tröster, 2008). Most of them conclude that these symptoms are not specific to the disease. DLB is associated with impairment in all areas of cognition, with visuoperceptual/spatial deficits predominating. Almost no study has specifically addressed language deterioration in DLB. Language is usually considered unimpaired in the early stages of the disease but impairments in confrontation naming have been reported with progression. This impairment is considered similar to what has been observed in patients with AD (e.g., Lambon Ralph et al., 2001), while some authors (e.g., Ferman et al., 2006) have reported that picture naming is better preserved in DLB than AD and could therefore be used as a differential diagnosis feature. Lambon Ralph et al. (2001) have shown that semantic impairment is affected in AD and DLB. However, while patients with AD showed equivalent deficits in every modality, patients with DLB demonstrated more severe deficits when they were presented with pictures than with words, a performance interpreted by the authors as the result of a combination of semantic and visuoperceptual impairments. Individuals with DLB also present syntactic impairment in sentence comprehension, a deficit attributed (Gross et al., 2012) to working memory and executive dysfunctions. A similar executive origin was proposed to explain their difficulty in organizing narrative speech in both production (Ash et al., 2011) and comprehension (Ash et al., 2012) modalities. Finally, very little is known as yet about reading and spelling abilities in DLB.

Language Processes in PPA

PPA is a progressive disorder of language associated with atrophy of the frontal and temporal regions of the left hemisphere, typically resulting from a neurodegenerative disease. PPA is a heterogeneous condition, in which the most prominent clinical feature is a difficulty with language (progressive impairment of language production, object naming, syntax, or word comprehension), while other cognitive domains are not affected at onset and in the early stages of the disease (Mesulam et al., 2012). A broad ranging International Consensus Group recently published recommendations for

the diagnosis and classification of PPA (Gorno-Tempini et al., 2011). These recommendations provide a classification of PPA and its three main variants, namely the nfvPPA, svPPA, and lvPPA. In the following subsections, we will briefly review current data on language and communication disorders in these three variants of PPA.

Language Processes in nfvPPA

According to the classification proposed by Gorno-Tempini et al. (2011), at least one of the following two core features must be present to detect nfvPPA: (1) effortful, halting speech with inconsistent speech sound errors and distortions (apraxia of speech), and (2) agrammatism in language production. Moreover, at least two of the following three additional features must be present: (1) impaired comprehension of syntactically complex sentences, (2) spared single-word comprehension, and (3) spared object knowledge. Imaging abnormalities in the left posterior frontoinsular region support the diagnosis.

Patients with nfvPPA complain of difficulty expressing themselves and finding words (Ash et al., 2010). Their speech fluency remains relatively preserved in the early stage of the disease, although the spontaneous speech of some patients is markedly slow, hesitant, and effortful (an example of spontaneous speech is presented in **Figure 1**). As the disease progresses, speech fluency decreases and articulation and prosody become more affected. Thus individuals with nfvPPA may present apraxia of speech, sometimes as initial symptoms of the disease, characterized by deficits in motor planning and sequencing of speech movements. These deficits are responsible for the production of phonological errors as well as the slow rate and abnormal pauses in connected speech (Wilson et al., 2010). The second core feature for the diagnosis of nfvPPA is the presence of agrammatism in production, characterized by difficulty in inflecting verbs (Thompson et al., 2012), by the omission or substitution of closed-class words (Ash et al., 2010), and by difficulty in sentence construction involving embedded clauses (Deleon et al., 2012). Patients with nfvPPA may also present with deficits in comprehension of syntactically complex sentences (Peelle et al., 2008), while single-word comprehension is usually unimpaired. Finally, less is known about written language in nfvPPA. Reading abilities usually remain good, although in some patients the disease progression is accompanied by the production of phonological errors (Rohrer et al., 2010). Written production appears to be more impaired than reading, affecting the majority of patients. Frequently reported manifestations of dysgraphia involve written agrammatism and the production of non-phonologically plausible paragraphias (Sepelyak et al., 2011).

Language Processes in svPPA

The svPPA is a clinical syndrome that results from a degenerative disease of the temporal lobes, leading to the selective impairment of semantic memory. According to the classification proposed by Gorno-Tempini et al. (2011), the following two core features must be present to establish a diagnosis of svPPA: (1) impaired confrontation naming, and (2) impaired single-word comprehension. Moreover, at least three of the following additional features must be present: (1) impaired object knowledge, particularly for low-frequency or low-familiarity items, (2) surface dyslexia or dysgraphia, (3) spared repetition, and (4) spared speech production (grammar and motor speech). With respect to neuroimaging, svPPA is supported by the presence of abnormalities in the ventral and lateral portions of the anterior temporal lobes bilaterally, although damage is usually greater on the left.

Spontaneous speech in patients with svPPA is fluent, well articulated and grammatically correct but presents many signs of word-finding difficulties (see **Figure 1**): miscarried sentences, latencies, circumlocutions, and occasional semantic paraphasias (Hodges et al., 1992a). Repetition is typically well preserved for words, nonwords, and sentences (Hodges and Patterson, 2007). Patients usually complain of a loss of memory for words. This deficit is obvious in tests of naming to definition and confrontation naming in which patients mainly produce semantic errors consisting of the production of a superordinate (lion → animal) or a category coordinate (lion → tiger) concept. This pattern of errors reflects a loss of attribute knowledge along with preservation of general superordinate information (Hodges and Patterson, 1995). Patients are also impaired on tests of verbal and nonverbal semantic memory. Spoken and written single-word comprehension is affected in tasks of word–picture matching, synonymy judging, word sorting, etc. The centrality of the semantic deficit is also confirmed by tests exploring nonverbal semantic knowledge such as tests of semantic relatedness judgment on pictures, picture sorting, etc. Deficit in picture naming and single-word comprehension is influenced by the familiarity of the items in the early stage of svPPA, and extends to more typical concepts with the progression of the disease (Hodges and Patterson, 2007). Whether the semantic deficit in svPPA has a greater effect on some category of concepts or some type of words is also discussed in the literature. For example, category-specific semantic impairment is rarely observed in svPPA but individuals have been described with greater impairment for living than for nonliving things (Barbarotto et al., 1995). In some patients, performance was better for abstract than for concrete concepts (Macoir, 2009) but the reverse dissociation was also shown (Hoffman and Lambon Ralph, 2011). Syntax processing is considered preserved in svPPA but impairment in verb inflection has been reported in a few studies (Cortese et al., 2006), suggesting a semantic contribution to morphological processing. Reading and spelling impairments are other clinical features of svPPA. Patients demonstrate surface dyslexia and surface dysgraphia characteristics (Brambati et al., 2009). A strong association between these impairments and the extent of the semantic loss was demonstrated in word reading (e.g., Funnell, 1996) and word spelling (e.g., Macoir and Bernier, 2002).

Language Processes in lvPPA

The lvPPA was recently identified as a distinct form of PPA. The two core features essential to the diagnosis of lvPPA are the presence of anomia in spontaneous speech and confrontation naming, and impaired repetition of sentences and phrases (Gorno-Tempini et al., 2011). At least three of the following additional features must be present: (1) production of phonological errors, (2) preservation of semantic memory, (3) preservation of articulation and prosody, and (4) absence of

1.a Narrative discourse of a patient with nfvPPA

When asked to describe the "Cookie Theft picture" from the BDAE (Goodglass and Kaplan, 1983) (a boy is falling from a stool while trying to catch cookies. His mother wipes the dishes while the sink is flooding), GC a 63 year-old woman with nfvPPA answered: '*There's…there's a b-b-boy…there's waber…the the floor'*. This is all she produced with over 120 seconds. Sentences were short. Language was non-fluent, effortful and marked by phonological paraphasias.

1.b Spontaneous written description of the problems encountered in everyday life by SC, a patient with svPPA (Bier and Macoir, 2010)

I can prepare and eat all my meals very well . . . However, I've lost many words related to things that I eat. For example, when I want to eat spaghetti, I know how to prepare it: to make the sauce and the whole meal. Visually, I know what I have to put into it to make it. However, I don't remember all of the words that correspond to what I need . . . The only words I remember since I made the recipe 4 days ago are: two cans of tomatoes, one can of tomato juice, rice, beef, pepper, and a dozen other small things I can see in my head. I don't remember their names but I know where they are in my kitchen . . . It's when I go to the restaurant that my symptoms are obvious. When I read the menu, I don't understand anymore. It is only what I eat regularly in two restaurants, that I can recognize instantly when I read [the menu]. But if I go back to a restaurant after many months, I don't remember the name of what I ate, even when I read the menu.

1.c Sentence repetition in a patient with lvPPA

DG, a 74 year-old women with lvPPA performed flawlessly in a single-word (e.g., grocery; pelican; chocolate) and short sentence (the plane was gone; the sun shines in the sky) repetition task. However, her performance was substantially affected (incomplete sentences, omission or substitution of words, no response) when she was asked to repeat long and syntactically complex sentences.
Examples:

Stimulus: We will give her as soon as she will claim it
Response: We, we, … if she wants it

Stimulus: The chicken was eaten by the big black dog
Response: the dog and the chicken … I don't know

Bier, N., Macoir, J., 2010. How to make a spaghetti sauce with a dozen of small things that I can not name: a review of the impact of semantic memory deficits on everyday action. Journal of Clinical and Experimental Neuropsychology 32 (2), 201–211.

Goodglass, H., Kaplan, E., 1983. The Assessment of Aphasia and Related Disorders. Lea & Febiger, Philadelphia.

Figure 1 Examples of language deficits in the three variants of primary progressive aphasia.

agrammatism. With respect to neuroimaging, measures of cortical thinning have revealed a predominant left posterior perisylvian or parietal atrophy in individuals with lvPPA. This variant is highly reminiscent of posterior forms of conduction aphasias and has been shown to harbor underlying AD pathology.

Spontaneous speech of patients with lvPPA is slowed, with frequent word-finding pauses and production of phonological errors. Compared to nfvPPA, utterances are simple but grammatically correct, and motor control of speech is not affected. Confrontation naming is less impaired than in the semantic variant, and semantic processing is well preserved. Although repetition is spared for short single words, patients encounter substantial difficulty repeating long and complex sentences (see **Figure 1**), particularly low-probability ones for which word substitutions are noted, suggesting that patients use a semantic rather than a phonological route to repeat (Gorno-Tempini et al., 2008). This impairment in repetition is not linguistic in nature but rather attributed to a reduction in verbal short-term memory resources. A similar functional origin was proposed to account for the impairment in sentence comprehension also found in lvPPA patients (Amici et al., 2007), whose performance was influenced by length but not by syntactic complexity. With respect to syntactic and morpho-syntactic processing per se, individuals with lvPPA show neither a deficit in production of verb tense nor problems in the production of grammatical sentences in connected speech (Thompson et al., 2013). To our knowledge, only Brambati

et al. (2009) specifically explored reading abilities in lvPPA. Results from this study show a pattern of phonological dyslexia, characterized by severe impairment in reading nonwords, accompanied by mild impairment for regular and irregular words, suggesting a deficit mainly localized in the grapheme-to-phoneme conversion route of reading. According to Faria et al. (2013), patients with lvPPA usually present with phonological dysgraphia, characterized by an impaired phonology-to-orthography conversion route and partially impaired access to orthographic word forms. The same spelling profile was also reported in a majority of cases by Sepelyak and her colleagues (Sepelyak et al., 2011), although the authors noted that other forms of dysgraphia are also possible.

Conclusions

In this article, we briefly reviewed impairments of spoken and written language processes accompanying the major forms of primary progressive dementia as well as VCI. These deficits are often prominent symptoms of the disease and may occur early in the process. We have shown that substantial information is available for spoken language disturbances in AD as well as in nfvPPA and svPPA. In contrast with the extensive literature on these dementing illnesses, relatively little is known about the patterns of spoken language deterioration in VCI and DLB. For example, the few available descriptions of cognitive deficits in these diseases mainly concern behavioral manifestations and large cognitive domains. More specific abilities are almost completely ignored and it is still difficult to relate functional deficits to language components. The concept of lvPPA is also quite recent and further research is needed to complete the linguistic description of the syndrome.

Knowledge about written language deficits in the major forms of dementia and PPA, and their contribution to the overall language syndrome are much less established in the literature. For example, very little is known about written language in nfvPPA and lvPPA and, to the best of our knowledge, no authors have addressed reading and spelling abilities in DLB. Further research should address this linguistic domain in order to complete the clinical profile of the different syndromes and thus provide additional features for the differential diagnosis.

The concept of dementia continues to evolve and this was recently reflected in 2011 by an avalanche of new criteria for various syndromes. Current studies have refined core features and made it possible to distinguish between different forms of PPA, for example. An improvement in clinical diagnosis should result from the study of more specific subcomponents of the different cognitive domains. In this respect, cognitive neuropsychology is certainly useful in that it allows the fractionation of many cognitive domains into specialized subcomponents that can be selectively impaired in a neurological affection. Such an approach has led to extensive descriptions of language deficits in AD and svPPA. Further research is needed to better characterize the effect of the progression of the disease on various language abilities and components. Overall, better characterization of language deficits in dementia and PPA would also be very useful for behavioral treatment approaches, as well as for counseling and educating patients, families, and caregivers.

See also: Aging and Memory; Agnosia; Agraphia (Acquired Dysgraphia); Alzheimer's Disease, Neural Basis of; Alzheimer's Disease: Behavioral and Social Aspects; Aphasia; Apraxia; Brain Damage: Neuropsychological Rehabilitation; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Cognitive Neuropsychology, Methods of; Cognitive Neuroscience; Dysexecutive Syndromes; Dyslexias, Acquired and Agraphia; Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects; Neuropsychological Testing; Psycholinguistics: Overview; Semantic Memory; Sentence Comprehension, Psychology of.

Bibliography

Aakerlund, L.P., Rosenberg, J., 1994. Writing disturbances: an indicator for post-operative delirium. International Journal of Psychiatry in Medicine 24, 245–257.

American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders: Text Revision, fourth ed. American Psychiatric Association, Washington.

Amici, S., Brambati, S.M., Wilkins, D.P., Ogar, J., Dronkers, N.L., Miller, B.L., Gorno-Tempini, M.L., 2007. Anatomical correlates of sentence comprehension and verbal working memory in neurodegenerative disease. Journal of Neuroscience 27, 6282–6290.

Ash, S., McMillan, C., Gross, R.G., Cook, P., Morgan, B., Boller, A., Dreyfuss, M., Siderowf, A., Grossman, M., 2011. The organization of narrative discourse in Lewy body spectrum disorder. Brain and Language 119, 30–41.

Ash, S., McMillan, C., Gunawardena, D., Avants, B., Morgan, B., Khan, A., Moore, P., Gee, J., Grossman, M., 2010. Speech errors in progressive non-fluent aphasia. Brain and Language 113, 13–20.

Ash, S., Xie, S.X., Gross, R.G., Dreyfuss, M., Boller, A., Camp, E., Morgan, B., O'Shea, J., Grossman, M., 2012. The organization and anatomy of narrative comprehension and expression in Lewy body spectrum disorders. Neuropsychology 26, 368–384.

Barbarotto, R., Capitani, E., Spinnler, H., Trivelli, C., 1995. Slowly progressive semantic impairment with category specificity. Neurocase 1, 107–119.

Barr, A., Benedict, R., Tune, L., Brandt, J., 1992. Neuropsychological differentiation of Alzheimer's disease from vascular dementia. International Journal of Geriatric Psychiatry 7, 621–627.

Bates, E., Harris, C., Marchman, V., Wulfeck, B., Kritchevsky, M., 1995. Production of complex syntax in normal ageing and Alzheimer's disease. Language and Cognitive Processes 10, 487–539.

Benedet, M.J., Montz, R., Delgado-Bolton, R.C., 2012. Alzheimer's disease and vascular dementia: neuropsychological differentiation in clinical practice. Clinical Gerontologist 35, 88–104.

Brambati, S.M., Ogar, J., Neuhaus, J., Miller, B.L., Gorno-Tempini, M.L., 2009. Reading disorders in primary progressive aphasia: a behavioral and neuroimaging study. Neuropsychologia 47, 1893–1900.

Chan, M., Lim, W.S., Sahadevan, S., 2008. Stage-independent and stage-specific phenotypic differences between vascular dementia and Alzheimer's disease. Dementia and Geriatric Cognitive Disorders 26, 513–521.

Chédru, F., Geschwind, N., 1972. Disorders of higher cortical functions in acute confusional states. Cortex 8, 395–411.

Cortese, M.J., Balota, D.A., Sergent-Marshall, S.D., Buckner, R.L., Gold, B.T., 2006. Consistency and regularity in past-tense verb generation in healthy ageing, Alzheimer's disease, and semantic dementia. Cognitive Neuropsychology 23, 856–876.

Croot, K., Hodges, J.R., Xuereb, J., Patterson, K., 2000. Phonological and articulatory impairment in Alzheimer's disease: a case series. Brain and Language 75, 277–309.

Deleon, J., Gesierich, B., Besbris, M., Ogar, J., Henry, M.L., Miller, B.L., Gorno-Tempini, M.L., Wilson, S.M., 2012. Elicitation of specific syntactic structures in primary progressive aphasia. Brain and Language 123, 183–190.

Duarte, L.R., Marquié, L., Marquié, J.-C., Terrier, P., Ousset, P.-J., 2009. Analyzing feature distinctiveness in the processing of living and non-living concepts in Alzheimer's disease. Brain and Cognition 71, 108–117.

Evans, L.K., 1987. Sundown syndrome in institutionalized elderly. Journal of the American Geriatrics Society 35, 101–108.

Fahlander, K., Wahlin, A., Almkvist, O., Bäckman, L., 2002. Cognitive functioning in Alzheimer's disease and vascular dementia: further evidence for similar patterns of deficits. Journal of Clinical and Experimental Neuropsychology 24, 734–744.

Faria, A.V., Crinion, J., Tsapkini, K., Newhart, M., Davis, C., Cooley, S., Mori, S., Hillis, A.E., 2013. Patterns of dysgraphia in primary progressive aphasia compared to post-stroke aphasia. Behavioural Neurology 26, 21–34.

Ferman, T.J., Smith, G.E., Boeve, B.F., Graff-Radford, N.R., Lucas, J.A., Knopman, D.S., Petersen, R.C., Ivnik, R.J., Wszolek, Z., Uitti, R., Dickson, D.W., 2006. Neuropsychological differentiation of dementia with Lewy bodies from normal aging and Alzheimer's disease. The Clinical Neuropsychologist 20, 623–636.

Forbes, K.E., Shanks, M.F., Venneri, A., 2004. The evolution of dysgraphia in Alzheimer's disease. Brain Research Bulletin 63, 19–24.

Forbes, K.E., Venneri, A., 2003. A case for case: handling letter case selection in written spelling. Neuropsychologia 41, 16–24.

Funnell, E., 1996. Response biases in oral reading: an account of the co-occurrence of surface dyslexia and semantic dementia. The Quarterly Journal of Experimental Psychology 49, 417–446.

Garrard, P., Lambon Ralph, M.A., Watson, P.C., Powis, J., Patterson, K., Hodges, J.R., 2001. Longitudinal profiles of semantic impairment for living and nonliving concepts in dementia of Alzheimer's type. Journal of Cognitive Neuroscience 13, 892–909.

Gorelick, P.B., Scuteri, A., Black, S.E., Decarli, C., Greenberg, S.M., Iadecola, C., Launer, L.J., Laurent, S., Lopez, O.L., Nyenhuis, D., Petersen, R.C., Schneider, J.A., Tzourio, C., Arnett, D.K., Bennett, D.A., Chui, H.C., Higashida, R.T., Lindquist, R., Nilsson, P.M., Roman, G.C., Sellke, F.W., Seshadri, S., 2011. Vascular contributions to cognitive impairment and dementia: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 42, 2672–2713.

Gorno-Tempini, M.L., Brambati, S.M., Ginex, V., Ogar, J., Dronkers, N.F., Marcone, A., Perani, D., Garibotto, V., Cappa, S.F., Miller, B.L., 2008. The logopenic/phonological variant of primary progressive aphasia. Neurology 71, 1227–1234.

Gorno-Tempini, M.L., Dronkers, N.F., Rankin, K.P., Ogar, J.M., Phengrasamy, L., Rosen, H.J., Johnson, J.K., Weiner, M.W., Miller, B.L., 2004. Cognition and anatomy in three variants of primary progressive aphasia. Annals of Neurology 55, 335–346.

Gorno-Tempini, M.L., Hillis, A.E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S.F., Ogar, J.M., Rohrer, J.D., Black, S., Boeve, B.F., Manes, F., Dronkers, N.F., Vandenberghe, R., Rascovsky, K., Patterson, K., Miller, B.L., Knopman, D.S., Hodges, J.R., Mesulam, M.M., Grossman, M., 2011. Classification of primary progressive aphasia and its variants. Neurology 76, 1006–1014.

Graham, N.L., 2000. Dysgraphia in dementia. Neurocase 6, 365–376.

Graham, N.L., Emery, T., Hodges, J.R., 2004a. Distinctive cognitive profiles in Alzheimer's disease and subcortical vascular dementia. Journal of Neurology, Neurosurgery, and Psychiatry 75, 61–71.

Graham, N.L., Patterson, K., Hodges, J.R., 2004b. When more yields less: speaking and writing deficits in nonfluent progressive aphasia. Neurocase 10, 141–155.

Gross, R.G., McMillan, C.T., Chandrasekaran, K., Dreyfuss, M., Ash, S., Avants, B., Cook, P., Moore, P., Libon, D.J., Siderowf, A., Grossman, M., 2012. Sentence processing in Lewy body spectrum disorder: the role of working memory. Brain and Cognition 78, 85–93.

Grossman, M., D'Esposito, M., Hughes, E., Onishi, K., Biassou, N., White-Devine, T., Robinson, K.M., 1996. Language comprehension profiles in Alzheimer's disease, multi-infarct dementia, and frontotemporal degeneration. Neurology 47, 183–189.

Hodges, J.R., Patterson, K., 1995. Is semantic memory consistently impaired early in the course of Alzheimer's disease? Neuroanatomical and diagnostic implications. Neuropsychologia 33, 441–459.

Hodges, J.R., Patterson, K., 2007. Semantic dementia: a unique clinicopathological syndrome. Lancet Neurology 6, 1004–1014.

Hodges, J.R., Patterson, K., Oxbury, S., Funnell, E., 1992a. Semantic dementia. Progressive fluent aphasia with temporal lobe atrophy. Brain 115, 1783–1806.

Hodges, J.R., Salmon, D.P., Butters, N., 1992b. Semantic memory impairment in Alzheimer's disease: failure of access or degraded knowledge? Neuropsychologia 30, 301–314.

Hoffman, P., Lambon Ralph, M.A., 2011. Reverse concreteness effects are not a typical feature of semantic dementia: evidence for the hub-and-spoke model of conceptual representation. Cerebral Cortex 21, 2103–2112.

Inouye, S.K., Zhang, Y., Han, L., Leo-Summers, L., Jones, R., Marcantonio, E., 2006. Recoverable cognitive dysfunction at hospital admission in older persons during acute illness. Journal of General Internal Medicine 21, 1276–1281.

Lambon Ralph, M.A., Powell, J., Howard, D., Whitworth, A.B., Garrard, P., Hodges, J.R., 2001. Semantic memory is impaired in both dementia with Lewy bodies (DLB) and dementia of Alzheimer's type (DAT): a comparative neuropsychological study and literature review. Journal of Neurology, Neurosurgery & Psychiatry 70, 149–156.

Lukatela, K., Malloy, P., Jenkins, M., Cohen, R., 1998. The naming deficit in early Alzheimer's and vascular dementia. Neuropsychology 12, 565–572.

Macleod, A.D., Whitehead, L.E., 1997. Dysgraphia and terminal delirium. Palliative Medicine 11, 127–132.

Macoir, J., 2009. Is a plum a memory problem? Longitudinal study of the reversal of concreteness effect in a patient with semantic dementia. Neuropsychologia 47, 518–535.

Macoir, J., Bernier, J., 2002. Is surface dysgraphia tied to semantic impairment? Evidence from a case of semantic dementia. Brain and Cognition 48, 236–244.

McKeith, I.G., Dickson, D.W., Lowe, J., Emre, M., O'Brien, J.T., Feldman, H., Cummings, J., Duda, J.E., Lippa, C., Perry, E.K., Aarsland, D., Arai, H., Ballard, C.G., Boeve, B., Burn, D.J., Costa, D., Del Ser, T., Dubois, B., Galasko, D., Gauthier, S., Goetz, C.G., Gomez-Tortosa, E., Halliday, G., Hansen, L.A., Hardy, J., Iwatsubo, T., Kalaria, R.N., Kaufer, D., Kenny, R.A., Korczyn, A., Kosaka, K., Lee, V.M.Y., Lees, A., Litvan, I., Londos, E., Lopez, O.L., Minoshima, S., Mizuno, Y., Molina, J.A., Mukaetova-Ladinska, E.B., Pasquier, F., Perry, R.H., Schulz, J.B., Trojanowski, J.Q., Yamada, M., 2005. Diagnosis and management of dementia with Lewy bodies: third report of the DLB Consortium. Neurology 65, 1863–1872.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack, C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., Mohs, R.C., Morris, J.C., Rossor, M.N., Scheltens, P., Carrillo, M.C., Thies, B., Weintraub, S., Phelps, C.H., 2011. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia 7, 263–269.

Mesulam, M.-M., Wieneke, C., Thompson, C., Rogalski, E., Weintraub, S., 2012. Quantitative classification of primary progressive aphasia at early and mild impairment stages. Brain 135, 1537–1553.

Neils-Strunjas, J., Groves-Wright, K., Mashima, P., Harnish, S., 2006. Dysgraphia in Alzheimer's disease: a review for clinical and research purposes. Journal of Speech, Language, and Hearing Research 49, 1313–1330.

Noble, K., Glosser, G., Grossman, M., 2000. Oral reading in dementia. Brain and Language 74, 48–69.

Peelle, J.E., Troiani, V., Gee, J., Moore, P., McMillan, C., Vesely, L., Grossman, M., 2008. Sentence comprehension and voxel-based morphometry in progressive nonfluent aphasia, semantic dementia, and nonaphasic frontotemporal dementia. Journal of Neurolinguistics 21, 418–432.

Powell, A.L., Cummings, J.L., Hill, M.A., Benson, D.F., 1988. Speech and language alterations in multi-infarct dementia. Neurology 38, 717–719.

Rogers, S.L., Friedman, R.B., 2008. The underlying mechanisms of semantic memory loss in Alzheimer's disease and semantic dementia. Neuropsychologia 46, 12–21.

Rogers, T.T., Ivanoiu, A., Patterson, K., Hodges, J.R., 2006. Semantic memory in Alzheimer's disease and the frontotemporal dementias: a longitudinal study of 236 patients. Neuropsychology 20, 319–335.

Rohrer, J., Rossor, M., Warren, J., 2010. Syndromes of nonfluent primary progressive aphasia A clinical and neurolinguistic analysis. Neurology 75, 603–610.

Sepelyak, K., Crinion, J., Molitoris, J., Epstein-Peterson, Z., Bann, M., Davis, C., Newhart, M., Heidler-Gary, J., Tsapkini, K., Hillis, A.E., 2011. Patterns of breakdown in spelling in primary progressive aphasia. Cortex 47, 342–352.

Thompson, C.K., Cho, S., Hsu, C.-J., Wieneke, C., Rademaker, A., Weitner, B.B., Mesulam, M.-M., Weintraub, S., 2012. Dissociations between fluency and agrammatism in primary progressive aphasia. Aphasiology 26, 20–43.

Thompson, C.K., Meltzer-Asscher, A., Cho, S., Lee, J., Wieneke, C., Weintraub, S., Mesulam, M.-M., 2013. Syntactic and morphosyntactic processing in stroke-induced and primary progressive aphasia. Behavioural Neurology 26, 35–54.

Tröster, A.I., 2008. Neuropsychological characteristics of dementia with Lewy bodies and Parkinson's disease with dementia: differentiation, early detection, and implications for "mild cognitive impairment" and biomarkers. Neuropsychology Review 18, 103–119.

Vuorinen, E., Laine, M., Rinne, J., 2000. Common pattern of language impairment in vascular dementia and in Alzheimer disease. Alzheimer Disease & Associated Disorders 14, 81–86.

Wallesch, C.W., Hundsalz, A., 1994. Language function in delirium: a comparison of single word processing in acute confusional states and probable Alzheimer's disease. Brain and Language 46, 592–606.

Waters, G.S., Caplan, 1997. Working memory and on-line sentence comprehension in patients with Alzheimer's disease. Journal of Psycholinguistic Research 26, 377–400.

Wilson, S.M., Henry, M.L., Besbris, M., Ogar, J.M., Dronkers, N.F., Jarrold, W., Miller, B.L., Gorno-Tempini, M.L., 2010. Connected speech production in three variants of primary progressive aphasia. Brain 133, 2069–2088.

Language Processing, Functional Magnetic Resonance Imaging of

Jie Yang and Steven L Small, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Functional imaging has now been used to study the neurobiology of language for more than 20 years. In this article, we review the current literature from functional magnetic resonance imaging in the areas of phonological processing, word comprehension, sentence processing, and discourse comprehension. While this literature confirms a major role of the classical perisylvian regions of the left hemisphere in language processing, it has become increasingly apparent that there are important contributions to language processing from a wide network of brain areas, including many regions in the right hemisphere, sensory–motor regions, and other areas – particularly in the nonperisylvian regions of the temporal and parietal lobes. Of particular recent interest has been the role of sensory and motor cortices in semantic representation and the integration between verbal and nonverbal information during communication.

Introduction

Functional magnetic resonance imaging (fMRI) has become a widely used method to investigate the neurobiology of language. The major advantage of fMRI, particularly with respect to earlier methods such as positron emission tomography and single-photon emission computed tomography, is that it is completely noninvasive, requires no radiation exposure, and if performed correctly, is without serious risks. Furthermore, the images it produces have a very high spatial resolution (as detailed as 1 mm). Besides the confined space of an MRI scanner, which makes it difficult to perform certain types of tasks in MRI experiments, functional MRI studies require that the person being scanned remain relatively motionless, rendering certain research paradigms more difficult to use in fMRI studies. For example, overt language generation tasks are infrequently used in such studies, since they can engender significant head movement. When fMRI studies involve covert language production or passive reading or listening, it is not possible to measure the behavior of participants directly. Instead, researchers often perform separate investigations outside of the scanner to collect behavioral data that can be related to the imaging data. This solution, while introducing confounds of its own, is often preferable to using procedures that introduce head motion, require metalinguistic processing, or add extraneous motor or sensory tasks to the primary task of interest (Small and Nusbaum, 2004).

Stimulus paradigms commonly used to elicit behavior in fMRI experiments fall into three general categories: block designs, event-related (ER) designs, and more recently continuous narrative designs. In a block design, two or more conditions are alternated in blocks, with each block containing many stimuli (and possibly responses) of a consistent type. The duration of each block is typically less than 1 min and depends on the time it takes for the MRI scanner to acquire a single volume of data (usually the same as the repetition time, or TR, with a duration of 1.5–3 s). For example, if a block lasts for 20 s and the TR = 2, then there will be 10 functional brain volumes collected per block. Stimuli in a block are presented sequentially and consist of a single experimental condition repeated continuously for the duration of the block. By alternating blocks containing different conditions, the fMRI signal can be compared across conditions, yielding information about the functional anatomy of the behavior tested. In some cases, one condition includes only a visual fixation point or nothing at all, and the other condition(s) can be compared to this 'rest,' yielding a widespread pattern of brain activity in the 'active' condition when compared to the 'rest' condition. Another type of design includes two 'active' conditions that differ in minor ways, such that the 'subtraction' (statistical contrast) is thought to represent a 'cognitive element' of interest. This latter approach has come under some scrutiny in recent years, due to a number of probably fallacious assumptions of this 'subtraction' method (Amaro and Barker, 2006; Logothetis, 2008; Small and Nusbaum, 2004). The main advantage of the block design is the very high signal-to-noise ratio obtained by repeating the same stimuli over and over (Buxton et al., 1998; Friston et al., 1999). This approach has some important drawbacks, however, such as the sensitivity of the signal to drift (e.g., head motion) and, perhaps most importantly, the consequences for behavioral aspects of the study, including the inability to characterize the processing of a single stimulus at a time, to randomize different types of stimuli, and to avoid strong expectations (e.g., list monitoring effects).

In contrast to block designs, ER designs involve presentation and measurement of individual stimuli. In such experiments, stimuli of different types are typically presented in a pseudorandom order. Two different types of ER experimental designs are commonly used, slow and rapid. In slow ER experiments, stimulus presentation is followed by a sufficiently long interstimulus interval (ISI) to allow the brain's hemodynamic response (the basis of the blood oxygen level-dependent signal that underlies the fMRI signal) to rise and fall completely in response to the stimulus. It typically takes about 4–8 s for the hemodynamic response to reach its maximum (depending on responsive brain areas and stimulus intensity) and another 5–10 s for it to return to baseline (Friston et al., 1999; Katwal et al., 2012; Thompson and Siegle, 2009). The advantage of this slow ER design is that each individual stimulus response is clearly distinguishable and subject to individual analysis. The disadvantage of the slow ER design is that there are very

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54033-1

few stimuli that can be presented, given the excessive time devoted to each one. On the other hand, combinations of individual events can be easily compared to rest or to other combinations of events in a variety of ways, taking advantage of individual features of the events.

To increase the numbers of stimuli presented in any given study, a number of researchers developed the 'rapid' ER design (Buckner et al., 1998; Burock et al., 1998; Maccotta et al., 2001; Wager et al., 2005), in which the individual stimuli are presented in relatively rapid succession, but separated by a variable ISI (or 'jitter') that permits mathematical separation of the overlapping hemodynamic responses. Both types of ER designs make it possible to analyze the brain effects related to individual trials, and when subjects are asked to make explicit responses to each stimulus, the trials can be evaluated with respect to different response characteristics (e.g., reaction times, errors). Other advantages include the possibility of randomizing the order of stimulus conditions (Rosen et al., 1998) and decreased sensitivity to head motion artifacts (Huang et al., 2002). The main drawback of ER designs is that the statistical power is lower than in block designs due to the fact that the signal change following a signal stimulus presentation is small.

Finally, some investigators have been experimenting with continuous narrative designs, in which naturalistic stimuli, such as movies or audiovisual stories, are presented for a relatively long duration, without any requirement by the subject other than attention to the stimulus. Each stimulus is coded beforehand for stimulus events of interest, e.g., hand movements (Hasson et al., 2004; Skipper et al., 2009) or semantic categories (Huth et al., 2012; Cukur et al., 2013), and then various methods are used to extract ER information from the raw times series. Although quite novel, this approach to stimulus presentation allows for the investigation of language in a real-world context without metacognitive influences.

Main Findings from fMRI Studies of Language Processing

Phonological Processing

Phonological processing involves detecting and discriminating differences in speech sounds for the purpose of speech understanding, and in producing the appropriate sounds in speech production. Various tasks have been used to investigate phonological processing, including repetition of words and syllables (Calvert et al., 1999; Riecker et al., 2000; Wildgruber et al., 2001), discriminating syllables or letters (Belin and Zatorre, 2003; Hugdahl et al., 2003; Jancke et al., 2002; Jancke and Shah, 2002; Jessen et al., 1999; Joanisse and Gati, 2003; Sekiyama et al., 2003), reading pseudowords that are pronounceable but have no meaning (Kotz et al., 2002; Mechelli et al., 2000; Meyer et al., 2002; Poldrack et al., 1999), discriminating pseudowords from real words (Binder et al., 2000; Cappa et al., 1998; Herbster et al., 1997; Hickok et al., 2003; Hugdahl et al., 2003; Meyer et al., 2002), judging whether words rhymes with each other (Booth et al., 2002; Buchanan et al., 2000; Heim and Friederici, 2003; Scott et al., ;2003), and recalling letters (Bunge et al., 2001; Burton et al., 2000, 2001, 2005; Burton and Small, 2006; Cohen et al., 1997; Hautzel et al., 2002; Rypma et al., 1999).

By way of introduction, it is important to note that the literature on 'phonological processing' contains a bewildering array of different phonological tasks, and that brain activation in the service of such 'phonological processing' depends critically on the task (Poeppel, 1996). In this section, we survey the frontal, parietal, and temporal regions involved in a variety of task-dependent phonological processes.

Frontal and Parietal Regions

A number of cortical regions in the frontal and parietal lobes appear to be involved in aspects of phonological processing. The frontal regions include the Rolandic operculum, the lower precentral gyrus, the gyral area directly ventral to the termination of the central sulcus (the 'guenon,' Bouchard et al., 2013), the dorsal part of the pars triangularis of the inferior frontal gyrus (IFG), and the junction between the orbital part of IFG and the middle frontal gyrus (MFG).

Among these regions, the guenon, ventral precentral gyrus, and Rolandic operculum appear to be involved in sensory–motor control during phonological processing (Vigneau et al., 2006). The upper guenon is in proximity to the primary motor area for the mouth and might contribute to low-level aspects of speech motor control. Several studies have found that this area is also activated during auditory discrimination of syllables (e.g., Hickok and Poeppel, 2004), listening to syllables (Wilson et al., 2004), and syllable production (Wildgruber et al., 2001). These findings suggest that simulation of motor activity might be necessary for perception of syllables (but see Tremblay and Small, 2011). The ventral part of the precentral gyrus is recruited by tongue and laryngeal movements (Riecher et al., 2000) and may be dedicated to coordination of fine movements of the pharynx and tongue (Poldrack et al., 1999) as well as working memory for phonological information (e.g., Bunge et al., 2001; Hautzel et al., 2002). The Rolandic operculum is activated in syllable/pseudoword articulation (Heim et al., 2002) and may reflect sensory–motor integration.

Previous studies have found that the pars opercularis is involved in phonological processing (Jessen et al., 1999; Mechelli et al., 2000; Sekiyama et al., 2003). This area appears to be important for working memory and/or cognitive control (Novick et al., 2005; Snyder et al., 2007) and is active on such tasks as recall lists of letters or syllables after a short delay (Bunge et al., 2001; Cohen et al., 1997; Hautzel et al., 2002; Rypma et al., 1999), and it might play an important role in executive control (Fuster, 1998). For example, in Snyder et al. (2007), participants performed a semantic and a phonological judgment. In each task, they were asked to select which of two targets best matched a probe. Both tasks included low-conflict and high-conflict trials. The results showed that the posterior left IFG showed increased activation with high versus low inhibition demand, regardless of whether the task was semantic or phonological. The anterior part of MFG and the pars triangularis of IFG are activated in syllable counting (Poldrack et al., 1999) and syllable identification (Sekiyama et al., 2003). The pars triangularis of the IFG and SMG are thought to be connected by the arcuate fasciculus (Catani et al., 2005) as well as by short connections (Duffau et al., 2003). Co-activations have been found in both pars triangularis and SMG during working-memory tasks (Cohen et al., 1997; Hautzel et al.,

2002; Rypma et al., 1999) and SMG might be related to phonological storage. The connection between the pars opercularis and the SMG suggests a frontal–temporal loop for phonological information retrieval and manipulation based on task demands. Finally, the anterior insula is recruited by dichotic presentation of syllables (Jäncke and Shah, 2002) and rhyme detection for written words (Booth et al., 2002).

Taken together, these findings suggest that the frontal and parietal regions are involved in the sensory–motor integration, retrieval, and manipulation of phonological information.

Temporal Regions

The temporal regions involved in phonological processing include primary auditory regions (within the lateral fissure and comprising parts of Heschl's gyrus and the superior temporal gyrus (STG), including planum polare and planum temporale (PT)), the anterior and middle parts of the STG, the middle part of the middle temporal gyrus (MTG), and the posterior part of the inferior temporal gyrus (ITG). These temporal regions are segregated along the superior temporal sulcus (STS).

Findings from nonhuman primate studies indicate that primary auditory regions consist of several parts (Hackett et al., 2001). A core field is comprised of auditory area I, rostral area, and rostrotemporal area with independent parallel inputs from the ventral division of the medial geniculate complex. The core field is surrounded by a belt field that consists of several areas, such as the caudolateral auditory belt, the caudomedial auditory belt, the anterior lateral auditory belt, and the rostrotemporolateral auditory belt. Anatomical tracing studies in monkeys found connections projecting from the anterior belt to the ventrolateral prefrontal cortex and from the caudal (posterior) belt to the dorsolateral prefrontal cortex. This provides evidence for ventral and dorsal processing streams within the auditory cortex (Kaas and Hackett, 2000). The dual processing streams in the auditory cortex are also found in the human brain. Anterior regions of the superior temporal cortex respond to native speech sounds and intelligible speech hierarchically, indicating an antero-lateral gradient along which the complexity of stimuli increases, from tones to words and sentences, while the posterior regions of the superior temporal cortex (e.g., PT) are involved in the processing of spectro-temporally complex sounds (for a review, see Rauschecker and Scott, 2009).

Recent fMRI studies have found that the anterior STG and Heschl's gyrus are involved in selective attention to sounds (Jäncke et al., 2001) and are activated in rhyming tasks (Fernandez et al., 2004). Regions in the left mid-posterior STS are found to perform phonetic and phonological analysis during speech perception (Turkeltaub and Coslett, 2010). The posterior STG is involved in phonological processing, independent of semantic processing. For instance, Graves et al. (2008) used an auditory pseudoword repetition task and found significantly reduced neural activity in the left posterior STG. This suggests that the left posterior STG participates specifically in lexical phonology. PT and posterior STG are connected with the precentral gyrus and pars opercularis of the IFG by the arcuate fasciculus (e.g., Parker et al., 2005). PT is involved in syllable perception (Jäncke et al., 2002; Joanisse and Gati, 2003; Sekiyama et al., 2003), auditory selective attention (Hugdahl et al., 2003), and covert speech production (Buchsbaum et al., 2001; Hickok et al., 2003; Okada et al., 2003). In speech perception and production, the PT and posterior STG in the temporal lobe, and the precentral gyrus and frontal operculum in the frontal lobe, are activated together. These findings are consistent with the (revised) motor theory of speech perception (Liberman and Whalen, 2000) and suggest a perception–action cycle permitting motor-sound-based phoneme representation (Rauschecker and Tian, 2000).

Semantic Processing of Words

Semantic processing of words is typically investigated by single word reading or listening (Binder et al., 1996, 2000; Cohen et al., 2002; Fiez et al., 1999; Moore and Price, 1999; Small et al., 1996; Specht and Reul, 2003), semantic categorization judgment (i.e., making a decision whether a presented stimulus is a member of a specific semantic category, such as living versus nonliving, or natural versus artificial (Binder et al., 1996, 2003; Braver and Bongiolatti, 2002; Buchanan et al., 2000; Heim et al., 2002; Noesselt et al., 2003; Poldrack et al., 1999; Scott et al., 2003), semantic association judgments involving words (Adams and Janata, 2002; Booth et al., 2002; Chee et al., 2000; Davis et al., 2004; McDermott et al., 2003; Noppeney and Price, 2004; Vandenberghe et al., 1996) or pictures (Adams and Janata, 2002; Chee et al., 2000; Vandenberghe et al., 1996), word generation (Buckner et al., 2000; Crosson et al., 1999), semantic retrieval (i.e., producing an object property) (Heun et al., 2000; James and Gauthier, 2004; Kelley et al., 2002; Thompson-Schill et al., 1999), and semantic priming (i.e., examining how the exposure to a stimulus influences the semantic processing of a later stimulus) (Kotz et al., 2002; Wagner et al., 2000; Wager and Smith, 2003).

Frontal Regions

In the frontal lobe, the pars orbitalis of the IFG appears to be important for online retrieval of semantic information (Demb et al., 1995). This region is involved in semantic categorization (Adams and Janata, 2002; Binder et al., 2003; Braver and Bongiolatti, 2002; Buchanan et al., 2000; Noesselt et al., 2003; Noppeney and Price, 2004; Poldrack et al., 1999), semantic retrieval (Kelley et al., 2002; James and Gauthier, 2004; Jobard et al., 2011), word generation (Adams and Janata, 2002; Gurd et al., 2002; Kelley et al., 2002; Wagner et al., 2001), and semantic association (Adams and Janata, 2002; Booth et al., 2002; Gurd et al., 2002; McDermott et al., 2003). The roles of the pars triangularis and pars opercularis of the IFG in semantic processing are controversial. Poldrack et al. (1999) found that performance of a semantic decision task preferentially elicited the ventral aspect of the left IFG (pars orbitalis/pars triangularis), and performance of a phonological task activated the dorsal part of the IFG (pars triangularis/pars opercularis). The authors claimed that the pars opercularis was relatively selective for phonological processing, whereas the pars triangularis was more relevant for semantic processing. However, recent findings and meta-analysis results have indicated that the pars opercularis appears to play a role in semantic functions,

whereas different portions of the pars triangularis may play various roles, with the dorsal part more involved in phonological processing and the ventral part more involved in semantic processing (Vigneau et al., 2006).

Parietal and Temporal Regions

In the parietal and temporal lobes, several brain regions are found to be involved in lexical–semantic processing, including the angular gyrus (AG), MTG, and ITG. All these areas are commonly activated during comprehension of spoken and written language (Price, 2012; Spitsyna et al., 2006).

Angular Gyrus

AG has been shown to be involved in the semantic processing of written words (Booth et al., 2002), auditory words (Ahmad et al., 2003; Chee et al., 1999), and pictures (Seghier et al., 2010). The anatomical location of AG adjoins visual, spatial, auditory, and somatosensory association areas, and this region plays a role in complex information integration and knowledge retrieval. Some researchers claim that AG might be the best candidate for a high-level, supermodel integration area in the human brain (Geschwind, 1965). It is possible that AG is at the top of a processing hierarchy for semantic retrieval and integration (Binder et al., 2009). A recent fMRI study investigated the functional subdivisions of the left AG for semantic processing and for default mode processing (Seghier et al., 2010). Their findings showed the middle AG to be involved in semantic association regardless of stimuli, and that this region overlaps with parts of the 'default network' (for an introduction to the default network, see Anticevic et al., 2012). Dorsal AG was primarily involved in meaning access of visual stimuli, and ventral AG was involved when making semantic judgments on visual stimuli.

Middle and Inferior Temporal Gyri

Activation in MTG and ITG occurs during semantic judgments (e.g., whether a target word is semantically related to a probe word) on auditory words, written words, and pictures. The anterior parts of MTG and ITG appear to be involved in retrieval and integration of semantic information independent of input modality (e.g., Awad et al., 2007; Jobard et al., 2007). The posterior MTG is sensitive to increased working-memory load, and processing word meaning from auditory and visual modalities simultaneously (Dick et al., 2009; Holle et al., 2010; Kircher et al., 2009; Robins et al., 2009).

Fusiform Gyrus

The fusiform gyrus (FG) is involved in word reading and word listening, when studied in contrast with pseudoword reading and nonword listening (Fiebach et al., 2002; Moore and Price, 1999, but see McCandiss et al., 2003). FG plays an important role in semantic processing of both pictures and words. Previous studies have found that anterior FG responds specifically to word meaning and the semantic context in which the words are presented, while posterior FG is activated during recognition of object pictures (e.g., Ishai et al., 1999; Okada et al., 2000; Haxby et al., 2001). FG also becomes active during reading in different modalities, such as tactile and visual in blind and sighted subjects (Büchel et al., 1998). Nobre et al. (1994) claimed that conceptual knowledge of words might be accessed directly in the inferior temporal lobe. FG is thought to correspond to the 'basal temporal language area' described by Lüders et al. (1991) based on neurophysiological recordings in epilepsy patients. The FG and parahippocampal gyrus are often activated together during semantic processing, and this suggests an interface between semantic memory (e.g., memory about conceptual knowledge) and episodic memory processing (Levy et al., 2004).

Connectivity

A classic model of frontal–temporal connections is the Broca–Wernicke–Geschwind model (Geschwind, 1970). According to the model, the arcuate fasciculus connects Wernicke's area in the left posterior STG with Broca's area in the left IFG. Recent studies have suggested a more extensive set of connections between language temporal and frontal areas. One recent view is that temporal–parietal areas incorporating AG, FG, and the temporal pole (TP) work with the pars orbitalis of the IFG to constitute a fronto-parieto-temporal network that is involved in semantic processing (Vigneau et al., 2006). The anatomical basis of this network includes the inferior longitudinal fasciculus that links posterior STS/angular sulcus (i.e., the portion of the STS that extends into the AG) to the TP and basal temporal language areas (Catani et al., 2002) and uncinate fasciculus connecting the TP and orbital frontal cortex (Catani et al., 2002; Horwitz et al., 1998; Nieuwenduys et al., 1988). This network is thought to play a role in semantics by constructing word meaning based on integrated knowledge from externally generated (e.g., vision and audition) and internally generated (e.g., emotion and long-term memory) information. A second network connecting the pars opercularis and the AG via the arcuate fasciculus is thought to be involved in working memory about word meaning. In this network, the AG may play a role in establishing semantic associations, whereas the pars opercularis might be more involved in semantic selection driven by task demands (Wagner et al., 2001).

Sentence Processing

Sentence comprehension requires integrating single word meanings into sentence-level meanings. Many linguists believe that this requires constructing sentence structures as an intermediary (Goldberg, 2006; Jackendoff, 2002; Pinker, 1999), although psycholinguistic data have provided substantial counterevidence against this modular perspective (Farmer et al., 2006; for a review, see Tanenhaus and Hare, 2007). Various tasks are used in fMRI studies of sentence understanding, such as passive listening or reading sentences (Crinion et al., 2003; Goel et al., 2000; Kansaku et al., 2000; Meyer et al., 2004), plausibility judgments of sentence meaning (Baumgaertner et al., 2002; Kuperberg et al., 2000; Zysset et al., 2002), emotional judgments (Kotz et al., 2003), mental imagery (Just et al., 2004), and generating sentences' final words (Kircher et al., 2001). Tasks probing syntactic processing include processing sentences with complex syntactic structures and detecting grammatical errors (Ben Shachar et al., 2004; Caplan, 2001; Constable et al., 2004; Cooke et al., 2002; Dapretto and Bookheimer, 1999; Embick et al., 2000; Luke et al., 2002). In this section, we will review

the brain regions involved in cognitive control and selection, processing structure and meaning, prosody and figurative language comprehension.

Cognitive Control and Selection during Sentence Comprehension

In the frontal lobe, the pars opercularis and pars triangularis of the IFG both seem to be involved in sentence comprehension. An area in the pars opercularis, located more caudally, is said to be involved in syntactic processing, such as comprehending complex sentences containing syntactic movement (Ben Shachar et al., 2004) and detecting grammatical errors (Embick et al., 2000). The pars triangularis is recruited during semantic processing related to grammatical processing, such as lexical categorization (Adams and Janata, 2002; Buchanan et al., 2000; Heim et al., 2002).

Although the frontal lobe is undoubtedly involved in sentence processing at some level, this remains somewhat in dispute. It is suggested that the functional organization of the frontal lobe might be determined by process type instead of stimulus type (e.g., verbal or nonverbal information) (e.g., Bornkessel-Schlesewsky and Schlesewsky, 2013; Wager and Smith, 2003). Numerous scholars have argued that Broca's region is involved in general processes of cognitive control rather than in linguistic computation per se (e.g., Stowe et al., 1998, 2005; Thompson-Schill et al., 1997, 2005). For instance, Novick et al. (2005) found an increase in left IFG activity for processing ambiguous sentences compared to unambiguous sentences. Their explanation was that sentences with ambiguity require selection among competing alternatives and that IFG helps coordinate this cognitive mechanism.

A meta-analysis of working-memory studies (Owen et al., 2005) indicated that the left IFG is recruited during selection, comparison, and judgment of stimuli. The authors suggest that the left ventrolateral IFG responds in a modality-independent manner to a variety of explicit task demands (e.g., explicit retrieval).

Processing of Structure and Meaning

In the temporal lobe, the middle MTG, posterior MTG, inferior STS, and posterior STG are activated during sentence processing.

The middle MTG is involved in comprehending sentences with simple or complex structures (Ben Shachar et al., 2004; Kansaku et al., 2000) and sentences that are syntactically or semantically incoherent (Luke et al., 2002). The posterior MTG appears to play a role in processing complex sentences (e.g., sentences that contain relative clauses) (Constable et al., 2004; Cooke et al., 2002). The middle MTG is also important in some semantic tasks (e.g., semantic association judgment) based on auditory and visual word presentation (Binder et al., 2000; Davis et al., 2004; Specht and Reul, 2003). The posterior MTG appears to be involved mainly in sentence *reading*, but it can also be activated by auditory presentation of sentences when subjects are asked to generate a mental image of the sentence content (Just et al., 2004). Current evidence suggests that MTG is an area that is of critical importance for using word meaning to understand sentences (Cooke et al., 2002; Dronkers et al., 2004; Vigneau et al., 2006), and it might also be important for cooperation between visual and auditory language processing networks during sentence processing.

The posterior STS is also sensitive to syntactic complexity (Constable et al., 2004), grammatical errors (Embick et al., 2000), syntactic generation (Kircher et al., 2001), and coherence in sentences (Kuperberg et al., 2000; Luke et al., 2002). One role of this area probably relates to the linkage of linguistic structure to meaning (Homae et al., 2002) and to the integration of linguistic material at the sentence level. The posterior STS also appears to play a role in the processing of theory of mind (TOM) (Gallagher et al., 2000) and inference development during narrative comprehension (Saxe and Kanwisher, 2003).

The inferior bank of the left STS is involved in sentence listening within meaningful discourse but not during unintelligible speech (Crinion et al., 2003; Kansaku et al., 2000; Meyer et al., 2004). It is also activated during syntactic plausibility judgment in sentence reading (Luke et al., 2002) and might be important for processing grammatical prosody (von Kriegstein et al., 2003).

The anterior temporal lobes are also active during sentence comprehension (Just et al., 2004; Kircher et al., 2001; Kuperberg et al., 2000; Luke et al., 2002; Stowe et al., 2005). Some researchers have claimed that the anterior temporal lobes play a critical role in the process of combining words into semantic chunks (Kintsch, 1998).

Figurative Language Comprehension

Figurative language refers to words and groups of words that exaggerate or alter the usual meanings of the component words. Figurative language may involve exaggerations and analogy to similar concepts or other contexts. The main fMRI studies on figurative language focus on metaphor, idiom, and irony comprehension.

Metaphor uses conceptual knowledge from a concrete semantic domain (e.g., a tangible thing or an image) to represent conceptual knowledge in an abstract semantic domain (Lakoff and Johnson, 1980). The figurative meaning of a metaphor can diverge from its literal meaning such that it is not obviously related to its original meaning. One important historical question, suggested by investigations of individuals with brain injuries (e.g., Monetta et al., 2006; Tompkins, 1990), is whether the right hemisphere plays a special role in comprehending metaphor. fMRI studies of metaphor processing give a mixed answer to this question, suggesting that right hemisphere involvement in metaphor comprehension can be influenced by several factors. One important factor relates to conventionality and novelty. In conventional metaphors, the figurative meaning is more salient and easy to access, whereas in novel metaphors, the literal meaning is more salient and the access of figurative meaning requires inference. According to the Graded Salience Hypothesis (Giora and Fein, 1999), the left hemisphere is more involved in processing conventional metaphors and the right hemisphere is more involved in the comprehension of novel metaphors. This view is supported by several fMRI studies. Lee and Dapretto (2006) showed that processing of conventional metaphors elicited stronger effects only in the left IFG, MTG, and IPL compared with literal comprehension, while Diaz et al. (2011) showed that comprehending novel metaphors evoked strong bilateral

effects, including activation in the right posterior MTG, TP, and IFG. Mashal et al. (2005) found that the right posterior MTG was involved in novel metaphor comprehension. Direct comparison showed that novel metaphors elicited stronger bilateral activations that included effects in the right frontal regions (Ahrens et al., 2007; Mashal et al., 2007) and the temporal regions (Mashal et al., 2007). Other differences have been found between novel and conventional metaphor comprehension. For example, Eviatar and Just (2006) revealed that conventional metaphors showed stronger activation than literal utterances not only in the left IFG and left ITG, but also in the right ITG. Several studies indicate that processing novel metaphors shows stronger effects only in the left brain regions compared with literal processing (e.g., Mashal et al., 2009; Rapp et al., 2004; Schibata et al., 2007). A recent meta-analysis using activation likelihood estimation found that the significant effects in the right hemisphere during metaphor processing occur when the metaphorical meaning is novel (Yang, 2012). This result supports the right hemisphere hypothesis on processing broad semantic fields and integrating distant concepts (Beeman, 2005).

Idioms are 'dead' metaphors (highly conventionalized metaphors), in which the figurative meaning is comprehended with regard to a common use of the expression. In idiom, the figurative meaning is directly (and holistically) retrieved from memory. fMRI studies of idiom comprehension have confirmed a bilateral brain network for processing idiomatic meaning. Zempleni et al. (2007) found processing of idiomatic sentences to involve IFG bilaterally and the left MTG. In Lauro et al. (2008), comprehension of idiomatic sentences elicited greater bilateral activations than comprehension of literal sentences. Effective connectivity analysis showed that the prefrontal cortex on both sides plays a crucial role in the selection between alternative sentence meanings. It is possible that the right hemisphere is involved in maintaining simultaneously various interpretations (Kacinik and Chiarello, 2007). In line with this, Mashal et al. (2008) found that compared to the idiomatic interpretation of idiomatic sentences, literal interpretation of idiomatic sentences elicited stronger activations in the right MTG, STG, and bilateral anterior insula lobe. The authors claimed that the right hemisphere is responsible for maintaining the nonsalient meaning (literal meaning), independent of figurativeness.

Irony is the use of words to express implied meaning other than and especially opposite to the literal meaning. Irony is widely used in daily conversation. Several fMRI studies have investigated the neural substrates of irony comprehension. Eviatar and Just (2006) found that listening to ironic statements results in stronger activation than to literal statements in the right STG and MTG. Wakusawa et al. (2007) investigated irony processing in real social situations. Participants were shown pictures depicting daily communicative situations and asked to judge situational appropriateness and literal correctness. The results showed the right TP to be involved in irony-specific processing, independent of task demand, while the medial orbitofrontal cortex plays a bigger role during irony comprehension only in the situational judgment task. The authors claimed that the right TP is involved in automatically assessing situational context during irony comprehension, whereas the medial orbitofrontal cortex is involved in judging the situational context. In Shibata et al. (2010), participants read a short scenario that consisted of five sentences. The first four sentences explained the situation of the protagonists. The fifth connoted an ironic, literal, or unconnected meaning. Then participants judged whether or not the final sentence expressed irony. Results indicated that both ironic and literal situations led to widespread bilateral brain activation, including the SFG, MFG, IFG, medial prefrontal cortex, STG, IPL, caudate, and thalamus, as well as the left insula and the left amygdala. Direct comparison between the ironic and the literal conditions showed higher activation for the ironic sentence in the right medial prefrontal cortex, the right precentral gyrus, and the left STG. Spotorno et al. (2013) found that verbal irony comprehension activated brain regions involved in the processing of TOM, including the bilateral temporal–parietal junction, the medial prefrontal cortex, and the precuneus. These results suggest that right fronto-temporal regions play important roles in irony comprehension, and that brain regions involved in emotion processing and processing of TOM contribute to understanding of irony.

In summary, sentence comprehension involves a brain network including the bilateral frontal and temporal regions for sentence processing. The temporal regions are important for comprehending sentences with complex meaning and syntactic structures, while the frontal regions are important for comprehension tasks that require cognitive control.

Discourse Processing

Discourse comprehension requires not only understanding of words and sentences, but also retrieval of world knowledge, semantic integration of current information with prior context, and coherence checking. An extended bilateral network is engaged during discourse comprehension as individuals make inferences and develop interpretations.

The tasks employed in the investigation of the neurobiology of discourse comprehension include reasoning (Caplan and Dapretto, 2001), topic judgment (Caplan and Dapretto, 2001), evaluation of connotation (Rapp et al., 2004), coherence judgment (Ferstl and von Cramon, 2001, 2002; Ferstl et al., 2005), attention to voice identity (Crinion et al., 2003), and moral judgment (Nichelli et al., 1995).

Previous studies have shown that auditory discourse comprehension activates a widespread bilateral fronto-temporal network, including the bilateral anterior temporal lobe, STG/STS, precentral gyrus, and pars triangularis and opercularis of IFG, MFG, and pre-SMA. Whether in a discourse context or not, speech perception per se is far more left lateralized, activating left MTG, posterior STS, left ITG/FG, pars triangularis of the left IFG, and the anterior temporal lobe bilaterally. Coherence checking during discourse comprehension activates the anterior temporal lobes bilaterally, the posterior STS bilaterally, as well as the middle MTG, pars opercularis, dorsal medial prefrontal cortex, and precentral gyrus, all on the left.

The bilateral middle and posterior MTG, STG, and STS are involved in semantic processing during discourse comprehension. The posterior portion of MTG might support a process providing initial access to meaning representations, such as

activating features and first-order associations of input words (Bookheimer, 2002; Demonet et al., 2005; Xu et al., 2005). The middle STG is activated during listening to permuted narratives compared with reversed narratives (Kansaku et al., 2000), making plausibility judgment of coherent sentences (Kuperberg et al., 2000), and comprehension of stories compared with listening to sentences of pseudowords (Ferstl and von Cramon, 2001). The middle STS plays a role in text comprehension (Kansaku et al., 2000), dialogue processing (Homae et al., 2003), and understanding coherent sentences (Kuperberg et al., 2000).

Compared to the middle and posterior temporal lobes, the anterior STG, STS, MTG, and TP (bilaterally) are multimodal association areas and appear to be responsible for integrating semantic, syntactic, and episodic information into a meaning representation of discourse content.

The pars opercularis and triangularis of the left IFG are involved in semantic selection that inhibits competing concepts while selecting one concept for action or for consciousness. The regions are co-activated with bilateral STG during word and sentence comprehension. In discourse comprehension, frontal areas appear to help process sentence-level context (Baumgärtner et al., 2002). The right pars opercularis and the anterior part of the right insula are involved in manipulation of information in working memory and decision-making. The involvement of these inferior frontal areas does not appear to be specific for one particular language component, but rather to be involved in cognitive control functions (Novick et al., 2005; Snyder et al., 2007). The right MFG is involved during auditory selective attention, independent of language stimuli (Jäncke and Shah, 2002; Poldrack et al., 1999; Sekiyama et al., 2003).

Finally, medial areas such as the orbital gyrus are involved during coherence checking. The orbital gyrus is activated in emotion processing in the context of verbal humor (Ferstl et al., 2005; Goel and Dolan, 2001, 2004; Siebörger et al., 2007) and in the integration of a current sentence with background knowledge in comprehension of funny cartoons (Maguire et al., 1999).

Additional Aspects of Language Processing in the Brain

Embodied Language Comprehension

The degree to which sensory–motor information is necessary for understanding the meanings of words and sentences has been the source of tremendous theoretical and experimental interest over the past decade.

Relevant theories can be placed on a continuum from not-at-all necessary ('unembodied') to crucial (strongly 'embodied'). Unembodied theories claim that sensory–motor information plays no causal role in semantic interpretation and there is an arbitrary relationship between the reference and the format of semantic information (Collins and Loftus, 1975; Landauer and Dumais, 1997; Levelt, 1989).

Some theories posit that although the format of semantic representation is amodal, it does not mean that there is a hard boundary between semantic information and sensory–motor information. These theories can be regarded as 'secondary embodiment,' which suggests that semantic representation is associated with sensory–motor systems but that sensory–motor representations are not causally involved in comprehension (Mahon and Caramazza, 2008; Patterson et al., 2007). According to these more integrated theories, activity in sensory and motor areas during semantic processing is secondarily affected by the connections mediating among semantic and sensory and motor information. Some adherents of this view propose the existence of a semantic 'hub' integrating information from sensory–motor locations (Rogers et al., 2004; Patterson et al., 2007).

A fourth view, of 'weak embodiment,' proposes that sensory–motor information partly contributes to semantic representation and is not just accessed secondary to abstract semantic information (Simmons and Barsalou, 2003; Vigliocco et al., 2004). In this view, semantic information relies on association areas that integrate modal information, but this information is distinct from raw experience processed by primary sensory–motor cortices.

Views that embrace strong embodiment claim that low-level sensory and motor information is activated in primary cortical areas and this activation is in service of semantic processing. They have generally shown the former but not established the latter. Sensory–motor systems are used to represent semantic information and can be modulated by semantic processing (Glenberg and Robertson, 2000; Zwaan, 2004). For example, Pulvermüller (1999) proposed that brain regions involved in processing of word-form and semantic related sensory–motor information 'wire together' during word meaning acquisition, and thus these areas 'fire together' during word comprehension. Gallese and Lakoff (2005) proposed that the same neural substrates are used for perceiving/doing, imaging, and language comprehension.

Thus far the neuroimaging evidence supports a degree of embodiment. Several findings indicate that semantic processing activates brain areas anterior to the primary sensory–motor areas involved in direct experience (Chatterjee, 2008, 2010; Martin et al., 2000; Thompson-Schill, 2003; Wallentin et al., 2005; Willems et al., 2010a,b). For example, Bar et al. (2001) showed that when participants viewed pictures about objects, areas in the ventral stream become active. The areas are anterior to those involved in visual shape perception. Wallentin et al. (2005) found that sentences describing actual and fictive motions activate brain areas anterior to MT/V5. Willems et al. (2010a,b) showed that comprehending action language elicits activity in the premotor cortex instead of the primary motor cortex. Tremblay and Small (2011b) found that different parts of the ventral premotor cortex were involved during action observation and action sentence comprehension. In another study, Tremblay and Small (2011a) found that the left ventral premotor cortex was significantly active during speech perception and production, but that activation was scaled to articulatory complexity only during speech production. This suggests an incompletely specified efferent motor signal during speech perception. Taken together, these findings indicate that sensory–motor information may be abstracted from direct experience and question the necessity of simulation during semantic processing. Complex conjunctions across modalities can represent more schematic and abstract information (Chatterjee, 2010).

Some recent studies have suggested that individual differences can shape the sensory–motor activations during language comprehension. Beilock et al. (2008) and Lyons et al. (2009) found that experienced ice-hockey players have stronger activations than nonhockey players in the left premotor cortex when they process sentences about hockey actions, but not when they process sentences about daily actions. Willems et al. (2010) indicated that handedness can influence motor activity during comprehension of hand verbs. Right-handed participants showed activity in the left premotor cortex during action verb comprehension, while left-handed participants indicated activity in the right premotor cortex. These results suggest that language comprehension is an embodied process and there is variability between individuals.

Nonverbal Information in Language Comprehension

During language comprehension, nonverbal information, such as facial expressions, gestures, and body movements, plays an important role. Recent fMRI studies provide evidence for the nonverbal information processing.

One important type of nonverbal information is gestures (for reviews, see Andric and Small, 2012; Goldin-Meadow and Alibali, in press). Studies about co-speech gestures (Dick et al., 2009; Green et al., 2009; Holle et al., 2008, 2010; Hubbard et al., 2009; Kircher et al., 2009; Skipper et al., 2009) and emblems (e.g., Lotze et al., 2006; Montgomery et al., 2007; Nakamura et al., 2004) have shown that IPL, SMG, and intraparietal sulcus (IPS) are strongly involved, though the laterality of SMG and IPS are not consistent across studies (Green et al., 2009; Lotze et al., 2006; Willems et al., 2007). In the frontal lobe, the bilateral premotor cortex is activated when viewing emblems (Nakamura et al., 2004; Montgomery et al., 2007; Villarreal et al., 2008) and co-speech gestures (Kircher et al., 2009). Distinct roles appear to exist for the ventral and dorsal premotor cortex. For example, Kircher et al. (2009) indicated that the ventral premotor cortex becomes active when people view metaphoric gestures, while Hubbard et al. (2009) found that the bilateral dorsal premotor cortex is activated when viewing beat gestures. The ventral premotor cortex might be sensitive to the semantic contribution of gesture (Skipper et al., 2007; Willems et al., 2007). Taken together, the premotor and parietal areas are involved in general gesture processing.

The IFG and lateral temporal cortex are involved in processing emblems that communicate meanings independent of speech (Goldin-Meadow, 1999; Gunter and Bach, 2004). The pars triangularis of IFG is thought to be important for interpreting gesture meaning as it does in interpreting language meaning (Kircher et al., 2009; Willems et al., 2007). The pars opercularis is also activated during gesture processing, particularly during co-speech gesture (Dick et al., 2009; Green et al., 2009; Kircher et al., 2009) and mouth and hand actions (Binkofski and Buccino, 2004; Rizzolatti and Craighero, 2004). In the temporal lobe, lateral temporal areas also respond to speech accompanied by gestures. The posterior STS responds to co-speech gestures and might play a role in integrating iconic gestures and speech (Holle et al., 2010). The bilateral posterior MTG is activated when people see metaphoric or iconic gestures (Green et al., 2009; Kircher et al., 2009; Willems et al., 2009). The anterior STG is activated in co-speech gesture processing (Skipper et al., 2007, 2009; Green et al., 2009; Straube et al., 2011) and emblem processing (Lotze et al., 2006) and might be involved in context processing. In a word, the IFG and lateral temporal cortices might respond to meanings independent of modality (Andric et al., in press). This suggests that interpretations of speech and gesture share a common neural substrate.

In addition to the investigation gesture processing during speech comprehension, researchers have also found that visual processing of mouth movements profoundly influences human speech perception. For example, watching a speaker's mouth movements improves speech comprehension, especially for normal listeners in noisy environments and for the hearing impaired (Grant et al., 1998; Sekiyama et al., 2003; Sumby and Pollack, 1954). Several brain regions have been found to be involved in cross-modal integration during speech perception, including STS, IPS, IFG, insular lobe, and Heschl's gyrus (Calvert et al., 1999, 2000, 2001; Callan et al., 2003, 2004; Olson et al., 2002; Beauchamp et al., 2004; Mottonen et al., 2004; Pekkola et al., 2005). These findings indicate that merging information from different modalities enables faster and more accurate discrimination than unimodal stimuli.

Conclusion and Further Research Directions

From the above review, two general conclusions can be seen in the current fMRI studies about language comprehension.

The first is that single language components (e.g., phonological processing) cannot be isolated functionally or anatomically. No function is localized in one particular brain region, but instead, each component is supported by a network that contains several regions and dynamic interactions among the regions. Connectivity analysis methods, such as functional connectivity and effective connectivity analysis, are used to examine the correlations or influences between regions. In addition, methods such as diffusion tensor imaging are used to detect structure connectivity. Further work using techniques such as quantitative meta-analysis should help elaborate connectivity patterns across brain regions (Cauda et al., 2011; Robinson et al., 2009).

The second is that understanding words and sentences, whether literal or figurative, requires representing and/or making use of semantic information, and how this information is encoded in the human brain remains a critical area for research. The notion that sensory–motor information plays a role in these processes is not much in dispute, but the nature and degree of this role is not known. On the one hand, sensory–motor information related to semantic content of the language stimuli is involved in language comprehension. The involvement of sensory–motor regions has been found to be flexible and context dependent. Yet the exact contribution of the sensory–motor system in comprehension, particularly related to its causal role in comprehension, still remains unclear. Furthermore, nonverbal information plays an important role. For example, visual perception of actions such as mouth movements and gestures can influence language

comprehension, and brain networks involved in understanding these actions are not distinct from those involved in understanding words and sentences. Future research will focus on the contextual influence on the integration of language and gesture processing.

See also: Embodiment Theory; Functional Brain Imaging of Language processes; Functional Magnetic Resonance Imagery, Analysis of; Gesture in Linguistics; Lateralization of Language as Demonstrated by Brain Imaging Procedures.

Bibliography

Adams, R.B., Janata, P., 2002. A comparison of neural circuits underlying auditory and visual object categorization. NeuroImage 16, 361–377.

Ahmad, Z., Balsamo, L.M., Sachs, B.C., Xu, B., Gaillard, W.D., 2003. Auditory comprehension of language in young children: neural networks identified with fMRI. Neurology 60, 1598–1605.

Ahrens, K., Liu, H.L., Lee, C.Y., Gong, S.P., Fang, S.Y., Hsu, Y.Y., 2007. Functional MRI of conventional and anomalous metaphors in Mandarin Chinese. Brain and Language 100, 163–171.

Amaro Jr., E., Barker, G.J., 2006. Study design in fMRI: basic principles. Brain and Cognition 60, 220–232.

Andric, M., Solodkin, A., Buccino, G., Goldin-Meadow, S., Rizzolatti, G., Small, S., Brain function overlaps when people observe emblems, speech, and grasping. Neuropsychologia 51, 1619–1629.

Andric, M., Small, S.L., 2012. Gesture's neural language. Frontiers and Psychology 3 (99).

Anticevic, A., Cole, M.W., Murray, J.D., Corlett, P.R., Wang, X.J., Krystal, J.H., 2012. The role of default network deactivation in cognition and disease. Trends in Cognitive Sciences 16, 584–592.

Awad, M., Warren, J.E., Scott, S.K., Turkheimer, F.E., Wise, R.J., 2007. A common system for the comprehension and production of narrative speech. Journal of Neuroscience 27, 11455–11464.

Bar, M., Tootell, R.B.H., Schacter, D.L., Greve, D.N., Fischl, B., Mendola, J.D., Rosen, B.R., Dale, A.M., 2001. Cortical mechanisms specific to explicit visual object recognition. Neuron 29, 529–535.

Baumgaertner, A., Weiller, C., Buchel, C., 2002. Event-related fMRI reveals cortical sites involved in contextual sentence integration. NeuroImage 16, 736–745.

Beeman, M., 2005. Bilateral brain processes for comprehending natural language. Trends in Cognitive Sciences 9, 512–518.

Beilock, S.L., Lyons, I.M., Mattarella-Micke, A., Nusbaum, H.C., Small, S.L., 2008. Sports experience changes the neural processing of action language. Proceedings of the National Academy of Sciences of the United States of America 105, 13269–13273.

Beauchamp, M.S., Argall, B.D., Bodurka, J., Duyn, J.H., Martin, A., 2004. Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nature Neuroscience 7, 1190–1192.

Belin, P., Zatorre, R.J., 2003. Adaptation to speaker's voice in right anterior temporal lobe. NeuroReport 14, 2105–2109.

Ben Shachar, M., Palti, D., Grodzinsky, Y., 2004. Neural correlates of syntactic movement: converging evidence from two fMRI experiments. NeuroImage 21, 1320–1336.

Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cerebral Cortex 19, 2767–2796.

Binder, J.R., Frost, J.A., Hammeke, T.A., Rao, S.M., Cox, R.W., 1996. Function of the left planum temporale in auditory and linguistic processing. Brain 119, 1239–1247.

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S.F., Springer, J.A., Kaufman, J.N., Possing, E.T., 2000. Human temporal lobe activation by speech and nonspeech sounds. Cerebral Cortex 10, 512–528.

Binkofski, F., Buccino, G., 2004. Motor functions of the Broca's region. Brain and Language 89, 362–369.

Bookheimer, S., 2002. Functional MRI of language: new approaches to understanding the cortical organization of semantic processing. Annual Review of Neuroscience 25, 151–188.

Bookheimer, S.Y., Zeffiro, T.A., Blaxton, T.A., Gaillard, W., Theodore, W.H., 2000. Activation of language cortex with automatic speech tasks. Neurology 55, 1151–1157.

Booth, J.R., Burman, D.D., Meyer, J.R., Gitelman, D.R., Parrish, T.B., Mesulam, M.M., 2002. Modality independence of word comprehension. Human Brain Mapping 16, 251–261.

Bornkessel-Schlesewsky, I., Schlesewsky, M., 2013. Reconciling time, space and function: a new dorsal-ventral stream model of sentence comprehension. Brain and Language 125, 60–76.

Bouchard, K.E., Mesgarani, N., Johnson, K., Chang, E.F., 2013. Functional organization of human sensorimotor cortex for speech articulation. Nature 495, 327–332.

Braver, T.S., Bongiolatti, S.R., 2002. The role of frontopolar cortex in subgoal processing during working memory. NeuroImage 15, 523–536.

Brockway, J.P., 2000. Two functional magnetic resonance imaging fMRI tasks that may replace the gold standard, Wada testing, for language lateralization while giving additional localization information. Brain and Cognition 43, 57–59.

Brown, S., Laird, A.R., Pfordresher, P.Q., Thelen, S.M., Turkeltaub, P., Liotti, M., 2009. The somatotopy of speech: phonation and articulation in the human motor cortex. Brain and Cognition 70, 31–41.

Buchanan, T.W., Lutz, K., Mirzazade, S., Specht, K., Shah, N.J., Zilles, K., Jäncke, L., 2000. Recognition of emotional prosody and verbal components of spoken language: an fMRI study. Brain Research. Cognitive Brain Research 9, 227–238.

Büchel, C., Price, C., Friston, K., 1998. A multimodal language region in the ventral visual pathway. Nature 394, 274–277.

Buchsbaum, B.R., Hickok, G., Humphries, C., 2001. Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. Cognitive Science 25, 663–678.

Buckner, R.L., Goodman, J., Burock, M., Rotte, M., Koutstaal, W., Schacter, D., Rosen, B., Dale, A.M., 1998. Functional-anatomic correlates of object priming in humans revealed by rapid presentation event-related fMRI. Neuron 20, 285–296.

Buckner, R.L., Koutstaal, W., Schacter, D.L., Rosen, B.R., 2000. Functional MRI evidence for a role of frontal and inferior temporal cortex in amodal components of priming. Brain 123, 620–640.

Bunge, S.A., Ochsner, K.N., Desmond, J.E., Glover, G.H., Gabrieli, J.D.E., 2001. Prefrontal regions involved in keeping information in and out of mind. Brain 124, 2074–2086.

Burock, M.A., Buckner, R.L., Woldorff, M.G., Rosen, B.R., Dale, A.M., 1998. Randomized event related experimental designs allow for extremely rapid presentation rates using functional MRI. Neuroreport 9, 3735–3739.

Burton, M.W., Locasto, P.C., Krebs-Noble, D., Gullapalli, R.P., 2005. A systematic investigation of the functional neuroanatomy of auditory and visual phonological processing. Neuroimage 26, 647–661.

Burton, M.W., Noll, D.C., Small, S.L., 2001. The anatomy of auditory word processing: individual variability. Brain and Language 77, 119–131.

Burton, M.W., Small, S.L., 2006. Functional neuroanatomy of segmenting speech and nonspeech. Cortex 42, 644–651.

Burton, M.W., Small, S.L., Blumstein, S.E., 2000. The role of segmentation in phonological processing: an fMRI investigation. Journal of Cognitive Neuroscience 12, 679–690.

Buxton, R.B., Wong, E.C., Frank, L.R., 1998. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. Magnetic Resonance in Medicine 39, 855–864.

Calvert, G.A., Brammer, M.J., Bullmore, E.T., Campbell, R., Iversen, S.D., David, A.S., 1999. Response amplification in sensory-specific cortices during crossmodal binding. NeuroReport 10, 2619–2623.

Calvert, G.A., Campbell, R., Brammer, M.J., 2000. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Current Biology 10, 649–657.

Calvert, G.A., Hansen, P.C., Iversen, S.D., Brammer, M.J., 2001. Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. NeuroImage 14, 427–438.

Caplan, D., 2001. Functional neuroimaging studies of syntactic processing. Journal of Psycholinguistic Research 30, 297–320.

Caplan, R., Dapretto, M., 2001. Making sense during conversation: an fMRI study. NeuroReport 12, 3625–3632.

Callan, D.E., Jones, J.A., Munhall, K., Callan, A.M., Kroos, C., Vatikiotis-Bateson, E., 2003. Neural processes underlying perceptual enhancement by visual speech gestures. NeuroReport 14, 2213–2218.

Callan, D.E., Jones, J.A., Munhall, K., Kroos, C., Callan, A.M., Vatikiotis-Bateson, E., 2004. Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. Journal Cognitive Neuroscience 16, 805–816.

Cappa, S.F., Perani, D., Schnur, T., Tettamanti, M., Fazio, F., 1998. The effect of semantic category and knowledge type on lexical – semantic access: a PET study. NeuroImage 8, 350–359.

Catani, M., Howard, R.J., Pajevic, S., Jones, D.K., 2002. Virtual *in vivo* interactive dissection of white matter fasciculi in the human brain. NeuroImage 17, 77–94.

Catani, M., Jones, D.K., ffytche, D.H., 2005. Perisylvian language networks of the human brain. Annals of Neurology 57, 8–16.

Cauda, F., Cavanna, A.E., D'agata, F., Sacco, K., Duca, S., Geminiani, G.C., 2011. Functional connectivity and coactivation of the nucleus accumbens: a combined functional connectivity and structure-based meta-analysis. Journal Cognitive Neuroscience 23, 2864–2877.
Chatterjee, A., 2008. The neural organization of spatial thought and language. Seminars in Speech and Language 29, 226–238.
Chatterjee, A., 2010. Disembodying cognition. Language and Cognition 2, 79–116.
Chee, M., Caplan, D., Soon, C.S., Sriram, N., Tan, E., Thiel, T., Weekes, B., 1999. Processing of visually presented sentences in Mandarin and English studied with fMRI. Neuron 23, 127–137.
Chee, M.W.L., Weekes, B., Lee, K.M., Soon, C.S., Schreiber, A., Hoon, J.J., Chee, M., 2000. Overlap and dissociation of semantic processing of Chinese characters, English words, and pictures: evidence from fMRI. NeuroImage 12, 392–403.
Cohen, L., Lehericy, S., Chochon, F., Lemer, C., Rivaud, S., Dehaene, S., 2002. Language-specific tuning of visual cortex? Functional properties of the visual word form area. Brain 125, 1054–1069.
Cohen, J.D., Peristein, W.M., Braver, T.S., Nystrom, L.E., Noll, D.C., Jonides, J., Smith, E.E., 1997. Temporal dynamics of brain activation during a working memory task. Nature 386, 604–607.
Collins, A.M., Loftus, E.F., 1975. A spreading-activation theory of semantic processing. Psychological Review 82, 407–428.
Cooke, A., Zurif, E.B., DeVita, C., Alsop, D., Koenig, P., Detre, J., Gee, J., Pinango, M., Balogh, J., Grossman, M., 2002. Neural basis for sentence comprehension: grammatical and short-term memory components. Human Brain Mapping 15, 80–94.
Constable, R.T., Pugh, K.R., Berroya, E., Mencl, W.E., Westerveld, M., Ni, W., Shankweiler, D., 2004. Sentence complexity and input modality effects in sentence comprehension: an fMRI study. NeuroImage 22, 11–21.
Crinion, J.T., Lambon-Ralph, M.A., Warburton, E.A., Howard, D., Wise, R.J., 2003. Temporal lobe regions engaged during normal speech comprehension. Brain 126, 1193–1201.
Crosson, B., Radonovich, K., Sadek, J.R., Gökçay, D., Bauer, R.M., Fischler, I.S., Cato, M.A., Maron, L., Auerbach, E.J., Browd, S.R., Briggs, R.W., 1999. Left-hemisphere processing of emotional connotation during word generation. NeuroReport 10, 2449–2455.
Cukur, T., Nishimoto, S., Huth, A.G., Gallant, J.L., 2013. Attention during natural vision warps semantic representation across the human brain. Nature Neuroscience 16, 763–770.
Dapretto, M., Bookheimer, S.Y., 1999. Form and content: dissociating syntax and semantics in sentence comprehension. Neuron 24, 427–432.
Davis, M.H., Meunier, F., Marslen-Wilson, W.D., 2004. Neural responses to morphological, syntactic, and semantic properties of single words: an fMRI study. Brain and Language 89, 439–449.
Demb, J.B., Desmond, J.E., Wagner, A.D., Vaidya, C.J., Glover, G.H., Gabrieli, J.D.E., 1995. Semantic encoding and retrieval in the left inferior prefrontal cortex: a functional MRI study of task difficulty and process specificity. Journal of Neuroscience 15, 5870–5878.
Démonet, J.-F., Thierry, G., Cardebat, D., 2005. Renewal of the neurophysiology of language: functional neuroimaging. Physiological Reviews 85, 49–95.
Diaz, M.T., Barrett, K.T., Hogstrom, L.J., 2011. The influence of sentence novelty and figurativeness on brain activity. Neuropsychologia 49, 320–330.
Dick, A.S., Goldin-Meadow, S., Hasson, U., Skipper, J.I., Small, S.L., 2009. Co speech gestures influence neural activity in brain regions associated with processing semantic information. Human Brain Mapping 30, 3509–3526.
Dronkers, N.F., Wilkins, D.P., Van Jr., V.R., Redfern, B.B., Jaeger, J.J., 2004. Lesion analysis of the brain areas involved in language comprehension. Cognition 92, 145–177.
Duffau, H., Gatignol, P., Denvil, D., Lopes, M., Capelle, L., 2003. The articulatory loop: study of the subcortical connectivity by electrostimulation. NeuroReport 14, 2005–2008.
Embick, D., Marantz, A., Miyashita, Y., O'Neil, W., Sakai, K.L., 2000. A syntactic specialization for Broca's area. Proceedings of the National Academy of Sciences of the United States of America 97, 6150–6154.
Eviatar, Z., Just, M.A., 2006. Brain correlates of discourse processing: an fMRI investigation of irony and conventional metaphor comprehension. Neuropsychologia 44, 2348–2359.
Farmer, T.A., Christiansen, M.H., Monaghan, P., 2006. Phonological typicality influences on-line sentence comprehension. Proceedings of the National Academy Science USA 103, 12203–12208.
Ferstl, E.C., von Cramon, D.Y., 2001. The role of coherence and cohesion in text comprehension: an event-related fMRI study. Brain Research. Cognitive Brain Research 11, 325–340.
Ferstl, E.C., von Cramon, D.Y., 2002. What does the fronto-medial cortex contribute to language processing: coherence or theory of mind? NeuroImage 17, 1599–1612.
Ferstl, E.C., Rinck, M., von Cramon, D.Y., 2005. Emotional and temporal aspects of situation model processing during text comprehension: an event-related fMRI study. Journal of Cognitive Neuroscience 17, 724–739.
Fernandez, B., Cardebat, D., Demonet, J.F., Joseph, P.A., Mazaux, J.M., Barat, M., Allard, M., 2004. Functional MRI follow-up study of language processes in healthy subjects and during recovery in a case of aphasia. Stroke 35, 2171–2176.
Fiebach, C.J., Friederici, A.D., Müller, K., von Cramon, D.Y., 2002. fMRI evidence for dual routes to the mental lexicon in visual word recognition. Journal of Cognitive Neuroscience 14, 11–23.
Fiez, J.A., Balota, D.A., Raichle, M.E., Petersen, S.E., 1999. Effects of lexicality, frequency, and spelling-to-sound consistency on the functional anatomy of reading. Neuron 24, 205–218.
Friston, K.J., Holmes, A.P., Price, C.J., Buchel, C., Worsley, K.J., 1999. Multisubject fMRI studies and conjunction analyses. NeuroImage 10, 385–396.
Friston, K.J., Zarahn, E., Josephs, O., Henson, R.N., Dale, A.M., 1999b. Stochastic designs in event-related fMRI. Neuroimage 10, 607–619.
Fuster, J.M., 1998. Linkage at the top. Neuron 21, 1223–1224.
Gallagher, H.L., Happé, F., Brunswick, N., Fletcher, P.C., Frith, U., Frith, C.D., 2000. Reading the mind in cartoons and stories: an fMRI study of theory of mind in verbal and nonverbal tasks. Neuropsychologia 38, 11–21.
Gallese, V., Lakoff, G., 2005. The brain's concepts: the role of the sensory-motor system in conceptual knowledge. Cognitive Neuropsychology 22, 455–479.
Geschwind, N., 1965. Disconnection syndromes in animals and man. Brain 88, 237–294, 585–644.
Geschwind, N., 1970. The organization of language and the brain. Science 170, 940–944.
Giora, R., Fein, O., 1999. On understanding familiar and less familiar figurative language. Journal of Pragmatics 31, 1601–1618.
Glenberg, A.M., Kaschak, M.P., 2003. The body's contribution to language. In: Ross, B.H. (Ed.), The Psychology of Learning and Motivation. Academic Press, San Diego, CA, pp. 93–126.
Glenberg, A.M., Robertson, D.A., 2000. Symbol grounding and meaning: a comparison of high-dimensional and embodied theories of meaning. Journal Memory and Language 43, 379–401.
Glover, G.H., 1999. Deconvolution of impulse response in event-related BOLD fMRI. NeuroImage 9, 416–429.
Goel, V., Dolan, R.J., 2001. The functional anatomy of humor: segregating cognitive and affective components. Nature Neuroscience 4, 237–261.
Goel, V., Dolan, R.J., 2004. Reciprocal neural response within lateral and ventral medial prefrontal cortex during hot and cold reasoning. NeuroImage 20, 2314–2321.
Goel, V., Buchel, C., Frith, C., Dolan, R.J., 2000. Dissociation of mechanisms underlying syllogistic reasoning. NeuroImage 12, 504–514.
Goldberg, A., 2006. Constructions at Work. Oxford Univ. Press, New York.
Goldin-Meadow, S., 1999. The role of gesture in communication and thinking. Trends in Cognitive Sciences 3, 419–429.
Goldin-Meadow, S., Alibali, M.W., Gesture's role in speaking, learning, and creating language. Annual Review of Psychology 64, 257–283.
Grant, K.W., Walden, B.E., Seitz, P.F., 1998. Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. Journal of the Acoustical Society of America 103, 2677–2690.
Graves, W.W., Grabowski, T.J., Mehta, S., Gupta, P., 2008. The left posterior superior temporal gyrus participates specifically in accessing lexical phonology. Journal of Cognitive Neuroscience 20, 1698–1710.
Green, A., Straube, B., Weis, S., Jansen, A., Willmes, K., Konrad, K., Kircher, T., 2009. Neural integration of iconic and unrelated coverbal gestures: a functional MRI study. Human Brain Mapping 30, 3309–3324.
Gurd, J.M., Amunts, K., Weiss, P.H., Zafiris, O., Zilles, K., Marshall, J.C., Fink, G.R., 2002. Posterior parietal cortex is implicated in continuous switching between verbal fluency tasks: an fMRI study with clinical implications. Brain 125, 1024–1038.
Gunter, T.C., Bach, P., 2004. Communicating hands: ERPs elicited by meaningful symbolic hand postures. Neuroscience Letters 372, 52–56.
Hackett, T.A., Preuss, T.M., Kaas, J.H., 2001. Architectonic identification of the core region in auditory cortex of macaques, chimpanzees, and humans. Journal of Comparative Neurology 441, 197–222.
Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R., 2004. Intersubject synchronization of cortical activity during natural vision. Science 303, 1634–1640.
Hautzel, H., Mottaghy, F.M., Schmidt, D., Zemb, M., Shah, N.J., Muller-Gartner, H.W., Krause, B.J., 2002. Topographic segregation and convergence of verbal, object, shape and spatial working memory in humans. Neuroscience Letters 323, 156–160.
Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.
Heim, S., Friederici, A.D., 2003. Phonological processing in language production: time course of brain activity. NeuroReport 14, 2031–2033.

Heim, S., Opitz, B., Friederici, A.D., 2002. Broca's area in the human brain is involved in the selection of grammatical gender for language production: evidence from event-related functional magnetic resonance imaging. Neuroscience Letters 328, 101–104.
Herbster, A.N., Mintun, M.A., Nebes, R.D., Becker, J.T., 1997. Regional cerebral blood flow during word and nonword reading. Human Brain Mapping 5, 84–92.
Heun, R., Jessen, F., Klose, U., Erb, M., Granath, D.O., Freymann, N., Grodd, W., 2000. Interindividual variation of cerebral activation during encoding and retrieval of words. European Psychiatry 15, 470–479.
Hickok, G., Buchsbaum, B., Humphries, C., Muftuler, T., 2003. Auditory-motor interaction revealed by fMRI: speech, music, and working memory in area Spt. Journal of Cognitive Neuroscience 15, 673–682.
Hickok, G., Poeppel, D., 2000. Towards a functional neuroanatomy of speech perception. Trends in Cognitive Sciences 4, 131–138.
Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92, 67–99.
Holle, H., Gunter, T.C., Rüschemeyer, S.A., Hennenlotter, A., Iacoboni, M., 2008. Neural correlates of the processing of co-speech gestures. NeuroImage 39, 2010–2024.
Holle, H., Obleser, J., Rueschemeyer, S.-A., Gunter, T.C., 2010. Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. NeuroImage 49, 875–884.
Homae, F., Hashimoto, R., Nakajima, K., Miyashita, Y., Sakai, K.L., 2002. From perception to sentence comprehension: the convergence of auditory and visual information of language in the left inferior frontal cortex. NeuroImage 16, 883–900.
Homae, F., Yahata, N., Sakai, K.L., 2003. Selective enhancement of functional connectivity in the left prefrontal cortex during sentence processing. NeuroImage 20, 578–586.
Horwitz, B., Rumsey, J.M., Donohue, B.C., 1998. Functional connectivity of the angular gyrus in normal reading and dyslexia. Proceedings of the National Academy of Sciences of the United States of America 95, 8939–8944.
Huang, J., Carr, T.H., Cao, Y., 2002. Comparing cortical activations for silent and overt speech using event-related fMRI. Human Brain Mapping 15, 39–53.
Hubbard, A.L., Wilson, S.M., Callan, D.E., Dapretto, M., 2009. Giving speech a hand: gesture modulates activity in auditory cortex during speech perception. Human Brain Mapping 30, 1028–1037.
Hugdahl, K., Thomsen, T., Ersland, L., Morten, R.L., Niemi, J., 2003. The effects of attention on speech perception: an fMRI study. Brain and Language 85, 37–48.
Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76, 1210–1224.
Ishai, A., Ungerleider, L.G., Martin, A., Schouten, J.L., Haxby, J.V., 1999. Distributed representation of objects in the human ventral visual pathway. Proceedings of the National Academy of Sciences of the United States of America 96, 9379–9384.
Jackendoff, R., 2002. Foundations of Language. Oxford Univ. Press, New York.
James, T.W., Gauthier, I., 2004. Brain areas engaged during visual judgments by involuntary access to novel semantic information. Vision Research 44, 429–439.
Jäncke, L., Buchanan, T.W., Lutz, K., Shah, N.J., 2001. Focused and nonfocused attention in verbal and emotional dichotic listening: an FMRI study. Brain and Language 78, 349–363.
Jäncke, L., Shah, N.J., 2002. Does dichotic listening probe temporal lobe functions? Neurology 58, 736–743.
Jäncke, L., Wustenberg, T., Scheich, H., Heinze, H.J., 2002. Phonetic perception and the temporal cortex. NeuroImage 15, 733–746.
Jessen, F., Erb, M., Klose, U., Lotze, M., Grodd, W., Heun, R., 1999. Activation of human language processing brain regions after the presentation of random letter strings demonstrated with event-related functional magnetic resonance imaging. Neuroscience Letters 270, 13–16.
Joanisse, M.F., Gati, J.S., 2003. Overlapping neural regions for processing rapid temporal cues in speech and nonspeech signals. NeuroImage 19, 64–79.
Jobard, G., Vigneau, M., Mazoyer, B., Tzourio-Mazoyer, N., 2007. Impact of modality and linguistic complexity during reading and listening tasks. NeuroImage 34, 784–800.
Jobard, G., Vigneau, M., Simon, G., Tzourio-Mazoyer, N., 2011. The weight of skill: Interindividual variability of reading related brain activation patterns in fluent readers. Journal of Neurolinguistics 24, 113–132.
Just, M.A., Newman, S.D., Keller, T.A., McEleney, A., Carpenter, P.A., 2004. Imagery in sentence comprehension: an fMRI study. NeuroImage 21, 112–124.
Kaas, J.H., Hackett, T.A., 2000. Subdivisions of auditory cortex and processing streams in primates. Proceedings of National Academy of Sciences USA 97, 11793–11799.
Kacinik, N.A., Chiarello, C., 2007. Understanding metaphors: is the right hemisphere uniquely involved? Brain and Language 100, 188–207.
Kansaku, K., Yamaura, A., Kitazawa, S., 2000. Sex differences in lateralization revealed in the posterior language areas. Cerebral Cortex 10, 866–872.
Katwal, S.B., Gore, J.C., Gatenby, J.C., Rogers, B.P., 2012. Measuring relative timings of brain activities using fMRI. Neuroimage 66C, 436–448.
Kelley, W.M., Macrae, C.N., Wyland, C.L., Caglar, S., Inati, S., Heatherton, T.F., 2002. Finding the self ? An event-related fMRI study. Journal of Cognitive Neuroscience 14, 785–794.
Kircher, T.T., Brammer, M., Tous, A.N., Williams, S.C., McGuire, P.K., 2001. Engagement of right temporal cortex during processing of linguistic context. Neuropsychologia 39, 798–809.
Kircher, T., Straube, B., Leube, D., Weis, S., Sachs, O., Willmes, K., Konrad, K., Green, A., 2009. Neural interaction of speech and gesture: differential activations of metaphoric co-verbal gestures. Neuropsychologia 47, 169–179.
Kintsch, W., 1998. Comprehension: A Paradigm for Cognition. Cambridge University Press, Cambridge.
Kotz, S.A., Cappa, S.F., von Cramon, D.Y., Friederici, A.D., 2002. Modulation of the lexical-semantic network by auditory semantic priming: an event-related functional MRI study. NeuroImage 17, 1761–1772.
Kotz, S.A., Meyer, M., Alter, K., Besson, M., von Cramon, D.Y., Friederici, A.D., 2003. On the lateralization of emotional prosody: an event-related functional MR investigation. Brain and Language 86, 366–376.
Kuperberg, G.R., McGuire, P.K., Bullmore, E.T., Brammer, M.J., Rabe-Hesketh, S., Wright, I.C., Lythgoe, D.J., Williams, S.C.R., David, A.S., 2000. Common and distinct neural substrates for pragmatic, semantic, and syntactic processing of spoken sentences: an fMRI study. Journal of Cognitive Neuroscience 12, 321–341.
Lakoff, G., Johnson, M., 1980. Metaphors We Live by. The University of Chicago Press, Chicago.
Landauer, T.K., Dumais, S.T., 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review 104, 211–240.
Lauro, L.J., Tettamanti, M., Cappa, S.F., Papagno, C., 2008. Idiom comprehension: a prefrontal task? Cerebral Cortex 18, 162–170.
Lee, S.S., Dapretto, M., 2006. Metaphorical vs. literal word meanings: fMRI evidence against a selective role of the right hemisphere. Neuroimage 292, 536–544.
Levelt, W.J.M., 1989. Speaking: From Intention to Articulation. MIT Press, London, England.
Levy, D.A., Bayley, P.J., Squire, L.R., 2004. The anatomy of semantic knowledge: medial vs lateral temporal lobe. Proceedings of the National Academy of Sciences of the United States of America 101, 6710–6715.
Liberman, A.M., Whalen, D.H., 2000. On the relation of speech to language. Trends in Cognitive Science 4, 187–196.
Logothetis, N.K., 2008. What we can do and what we cannot do with fMRI. Nature 453, 869–878.
Loubinoux, I., Carel, C., Alary, F., Boulanouar, K., Viallard, G., Manelfe, C., Rascol, O., Celsis, P., Chollet, F., 2001. Within-session and between-session reproducibility of cerebral sensorimotor activation: a test-retest effect evidenced with functional magnetic resonance imaging. Journal of Cerebral Blood Flow and Metabolism 21, 592–607.
Lotze, M., Heymans, U., Birbaumer, N., Veit, R., Erb, M., Flor, H., Hals- band, U., 2006. Differential cerebral activation during observation of expressive gestures and motor acts. Neuropsychologia 44, 1787–1795.
Luke, K.K., Liu, H.L., Wai, Y.Y., Wan, Y.L., Tan, L.H., 2002. Functional anatomy of syntactic and semantic processing in language comprehension. Human Brain Mapping 16, 133–145.
Lüders, H., Lesser, R.P., Hahn, J., Dinner, D.S., Morris, H.H., Wyllie, E., Godoy, J., 1991. Basal temporal language area. Brain 114, 743–754.
Lyons, I.M., Mattarella-Micke, A., Cieslak, M., Nusbaum, H.C., Small, S.L., Beilock, S.L., 2009. The role of personal experience in the neural processing of action-related language. Brain and Language 1123, 214–222.
Maccotta, L., Zacks, J.M., Buckner, R.L., 2001. Rapid self-paced event-related functional MRI: feasibility and implications of stimulus- versus response-locked timing. Neuroimage 14, 1105–1121.
Maguire, E.A., Frith, C.D., Morris, R.G.M., 1999. The functional neuroanatomy of comprehension and memory: the importance of prior knowledge. Brain 122, 1839–1850.
Mahon, B.Z., Caramazza, A., 2008. A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. Journal of Physiology Paris 102, 59–70.
Mashal, N., Faust, M., Hendler, T., 2005. The role of the right hemisphere in processing nonsalient metaphorical meanings: application of principal components analysis to fMRI data. Neuropsychologia 43, 2084–2100.
Mashal, N., Faust, M., Hendler, T., Jung-Beeman, M., 2007. An fMRI investigation of the neural correlates underlying the processing of novel metaphoric expressions. Brain and Language 100, 115–126.
Mashal, N., Faust, M., Hendler, T., Jung-Beeman, M., 2008. Hemispheric differences in processing the literal interpretation of idioms: converging evidence from behavioral and fMRI studies. Cortex 44, 848–860.

Mashal, N., Faust, M., Hendler, T., Jung-Beeman, M., 2009. An fMRI study of processing novel metaphoric sentences. Laterality 14, 30–54.
Martin, A., Wiggs, C.L., Ungerleider, L.G., Haxby, J.V., 2000. Neural correlates of category specific knowledge. Nature 379, 649–652.
McCandliss, B.D., Cohen, L., Dehaene, S., 2003. The visual word form area: expertise for reading in the fusiform gyrus. Trends in Cognitive Sciences 7, 293–299.
McDermott, K.B., Petersen, S.E., Watson, J.M., Ojemann, J.G., 2003. A procedure for identifying regions preferentially activated by attention to semantic and phonological relations using functional magnetic resonance imaging. Neuropsychologia 41, 293–303.
Mechelli, A., Friston, K.J., Price, C.J., 2000. The effects of presentation rate during word and pseudoword reading: a comparison of PET and fMRI. Journal of Cognitive Neuroscience 12, 145–156.
Meyer, M., Alter, K., Friederici, A.D., Lohmann, G., von Cramon, D.Y., 2002. FMRI reveals brain regions mediating slow prosodic modulations in spoken sentences. Human Brain Mapping 17, 73–88.
Meyer, M., Steinhauer, K., Alter, K., Friederici, A.D., von Cramon, D.Y., 2004. Brain activity varies with modulation of dynamic pitch variance in sentence melody. Brain and Language 89, 277–289.
Monetta, L., Ouellet-Plamondon, C., Joanette, Y., 2006. Simulating the pattern of right-hemisphere-damaged patients for the processing of the alternative metaphorical meanings of words: evidence in favor of a cognitive resources hypothesis. Brain and Language 96, 171–177.
Montgomery, K.J., Isenberg, N., Haxby, J.V., 2007. Communicative hand gestures and object-directed hand movements activated the mirror neuron system. Social Cognitive and Affective Neuroscience 2, 114–122.
Moore, C.J., Price, C.J., 1999. Three distinct ventral occipitotemporal regions for reading and object naming. NeuroImage 10, 181–192.
Mottonen, R., Schurmann, M., Sams, M., 2004. Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study. Neuroscience Letters 363, 112–115.
Nakamura, A., Maess, B., Knosche, T.R., Gunter, T.C., Bach, P., Friederici, A.D., 2004. Cooperation of different neuronal systems during hand sign recognition. NeuroImage 23, 25–34.
Nichelli, P., Grafman, J., Pietrini, P., Clark, K., Lee, K.Y., Miletich, R., 1995. Where the brain appreciates the moral of a story. NeuroReport 6, 2309–2313.
Nieuwenhuys, R., Voogd, J., van Huijzen, C., 1988. The Human Central Nervous System. A Synopsis and Atlas. Springer-Verlag, Berlin.
Nobre, A.C., Allison, T., McCarthy, G., 1994. Word recognition in the human inferior temporal lobe. Nature 372, 260–263.
Noesselt, T., Shah, N.J., Jancke, L., 2003. Top-down and bottom-up modulation of language related areas—An fMRI study. BioMed Central Neuroscience 4, 13.
Noppeney, U., Price, C.J., 2004. Retrieval of abstract semantics. NeuroImage 22, 164–170.
Novick, J.M., Trueswell, J.C., Thompson-Schill, S.L., 2005. Cognitive control and parsing: reexamining the role of Broca's area in sentence comprehension. Cognitive, Affect, and Behavioral Neuroscience 5, 263–281.
Okada, T., Tanaka, S., Nakai, T., Nishizawa, S., Inui, T., et al., 2000. Naming of animals and tools: a functional magnetic resonance imaging study of categorical differences in the human brain areas commonly used for naming visually presented objects. Neuroscience Letters 15, 33–36.
Okada, K., Smith, K.R., Humphries, C., Hickok, G., 2003. Word length modulates neural activity in auditory cortex during covert object naming. NeuroReport 14, 2323–2326.
Olson, I.R., Gatenby, J.C., Gore, J.C., 2002. A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. Brain Research. Cognitive Brain Research 14, 129–138.
Owen, A.M., McMillan, K.M., Laird, A.R., Bullmore, E., 2005. N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. Human Brain Mapping 25, 46–59.
Parker, G.J., Luzzi, S., Alexander, D.C., Wheeler-Kingshott, C.A., Ciccarelli, O., Lambon Ralph, M.A., 2005. Lateralization of ventral and dorsal auditory-language pathways in the human brain. NeuroImage 24, 656–666.
Patterson, K., Nestor, P.J., Rogers, T.T., 2007. Where do you know what you know? The representation of semantic knowledge in the human brain. Nature Reviews Neuroscience 8, 976–987.
Pekkola, J., Ojanen, V., Autti, T., Jaaskelainen, I.P., Mottonen, R., Tarkiainen, A., Sams, M., 2005. Primary auditory cortex activation by visual speech: an fMRI study at 3 T. NeuroReport 16, 125–128.
Pinker, S., 1999. Words and Rules. Harper, New York.
Poeppel, D., 1996. A critical review of PET studies of phonological processing. Brain and Language 55, 317–351.
Poldrack, R.A., Wagner, A.D., Prull, M.W., Desmond, J.E., Glover, G.H., Gabrieli, J.D., 1999. Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. NeuroImage 10, 15–35.
Price, C.J., 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage 62, 816–847.
Pulvermüller, F., 1999. Words in the brain's language. Behavioral and Brain Sciences 22, 253–336.
Rapp, A.M., Leube, D.T., Erb, M., Grodd, W., Kircher, T.T., 2004. Neural correlates of metaphor processing. Brain Research. Cognitive Brain Research 203, 395–402.
Rauschecker, J.P., Scott, S.K., 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature Reviews Neuroscience 12, 718–724.
Rauschecker, J.P., Tian, B., 2000. Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proceedings of the National Academy of Sciences of the United States of America 97, 11800–11806.
Riecker, A., Ackermann, H., Wildgruber, D., Meyer, J., Dogil, G., Haider, H., Grodd, W., 2000. Articulatory/phonetic sequencing at the level of the anterior perisylvian cortex: a functional magnetic resonance imaging fMRI study. Brain and Language 75, 259–276.
Rizzolatti, G., Craighero, L., 2004. The mirror-neuronsystem. Annual Review of Neuroscience 27, 169–192.
Robins, D.L., Hunyadi, E., Schultz, R.T., 2009. Superior temporal activation in response to dynamic audio-visual emotional cues. Brain and Cognition 69, 269–278.
Robinson, J.L., Laird, A.R., Glahn, D.C., Lovallo, W.R., Fox, P.T., 2009. Meta-analytic connectivity modeling: delineating the functional connectivity of the human amygdala. Human Brain Mapping 31, 173–184.
Rogers, T.T., Lambon Ralph, M.A., Garrard, P., Bozeat, S., McClelland, J.L., Hodges, J.R., Patterson, K., 2004. Structure and deterioration of semantic memory: a neuropsychological and computational investigation. Psychological Review 111, 205–235.
Rosen, B.R., Buckner, R.L., Dale, A.M., 1998. Event-related functional MRI: past, present, and future. Proceedings of the National Academy of Sciences of the United States of America 95, 773–780.
Rypma, B., Prabhakaran, V., Desmond, J.E., Glover, G.H., Gabrieli, J.D., 1999. Load-dependent roles of frontal brain regions in the maintenance of working memory. NeuroImage 9, 216–226.
Saxe, R., Kanwisher, N., 2003. People thinking about thinking people. The role of the temporo-parietal junction in ''theory of mind''. NeuroImage 19, 1835–1842.
Scott, S.K., Leff, A.P., Wise, R.J., 2003. Going beyond the information given: a neural system supporting semantic interpretation. NeuroImage 19, 870–876.
Seghier, M.L., Fagan, E., Price, C.J., 2010. Functional subdivisions in the left angular gyrus where the semantic system meets and diverges from the default network. Journal of Neuroscience 30, 16809–16817.
Sekiyama, K., Kanno, I., Miura, S., Sugita, Y., 2003. Auditory-visual speech perception examined by fMRI and PET. Neuroscience Research 47, 277–287.
Shibata, M., Abe, J., Terao, A., Miyamoto, T., 2007. Neural mechanisms involved in the comprehension of metaphoric and literal sentences: an fMRI study. Brain Research 1166, 92–102.
Shibata, M., Toyomura, A., Itoh, H., Abe, J., 2010. Neural substrates of irony comprehension: a functional MRI study. Brain Research 1308, 114–123.
Siebörger, F., Ferstl, E., von Cramon, Y., 2007. Making sense of nonsense: An fMRI study of task induced inference processes during discourse comprehension. Brain Research 1166, 77–91.
Simmons, K.W., Barsalou, L.W., 2003. The similarity-in-topography principle: reconciling theories of conceptual deficits. Cognitive Neuropsychology 20 (3/4/5/6), 451–486.
Skipper, J.I., Goldin-Meadow, S., Nusbaum, H.C., Small, S.L., 2007. Speech-associated gestures, Broca's area, and the human mirror system. Brain and Language 101, 260–277.
Skipper, J.I., Goldin-Meadow, S., Nusbaum, H.C., Small, S.L., 2009. Gestures orchestrate brain networks for language understanding. Current Biology 19, 1–7.
Small, S.L., Nusbaum, H.C., 2004. On the neurobiological investigation of language understanding in context. Brain and Language 89, 300–311.
Small, S.L., Noll, D.C., Perfetti, C.A., Hlustik, P., Wellington, R., Schneider, W., 1996. Localizing the lexicon for reading aloud: replication of a PET study using fMRI. NeuroReport 7, 961–965.
Snyder, H.R., Feigenson, K., Thompson-Schill, S.L., 2007. Prefrontal cortical response to conflict during semantic and phonological tasks. Journal of Cognitive Neuroscience 19, 761–775.
Sotillo, M., Carreti, L., Hinojosa, J.A., Tapia, M., Mercado, F., Lopez-Martin, S., Albert, J., 2005. Neural activity associated with metaphor comprehension: spatial analysis. Neuroscience Letters 373, 5–9.

Specht, K., Reul, J., 2003. Functional segregation of the temporal lobes into highly differentiated subsystems for auditory perception: an auditory rapid event-related fMRI-task. NeuroImage 20, 1944–1954.

Spitsyna, G., Warren, J.E., Scott, S.K., Turkheimer, F.E., Wise, R.J.S., 2006. Converging language streams in the human temporal lobe. Journal of Neuroscience 26, 7328–7336.

Spotorno, N., Cheylus, A., Van Der Henst, J.-B., Noveck, I.A., 2013. What's behind a p600? integration operations during irony processing. PLoS ONE 8, 66839.

Stowe, L.A., Broere, C., Paans, A., Wijers, A., Mulder, G., Vaalburg, W., Zwarts, F., 1998. Localising components of a complex task: sentence processing and working memory. Neuroreport 9, 2995–2999.

Stowe, L.A., Haverkort, M., Zwarts, F., 2005. Rethinking the neurological basis of language. Lingua 115, 997–1042.

Straube, B., Green, A., Bromberger, B., Kircher, T., 2011. The differentiation of iconic and metaphoric gestures: common and unique integration processes. Human Brain Mapping 32, 520–533.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America 26, 212–215.

Tanenhaus, M.K., Hare, M., 2007. Phonological typicality and sentence processing. Trends Cognitive Sciences 11, 93–95.

Thompson, W.K., Siegle, G., 2009. A stimulus-locked vector autoregressive model for slow event-related fMRI designs. Neuroimage 46, 739–748.

Thompson-Schill, S.L., 2003. Neuroimaging studies of semantic memory: inferring "how" from "where". Neuropsychologia 41, 280–292.

Thompson-Schill, S.L., Aguirre, G., D'Esposito, M., Farah, M.J., 1999. A neural basis for category and modality specific of semantic knowledge. Neuropsychologia 37, 671–676.

Thompson-Schill, S.L., Bedny, M., Goldberg, R.F., 2005. The frontal lobes and the regulation of mental activity. Current Opinion in Neurobiology 15, 219–224.

Thompson-Schill, S.L., D'Esposito, M., Aguirre, G.K., Farah, M.J., 1997. Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. Proceedings of the National Academy of Sciences of the United States of America 94, 14792–14797.

Tompkins, C.A., 1990. Knowledge and strategies for processing lexical metaphor after right or left hemisphere brain damage. Journal of Speech and Hearing Research 33, 307–316.

Tremblay, P., Small, S.L., 2011a. On the context-dependent nature of the contribution of the ventral premotor cortex to speech perception. Neuroimage 57, 1561–1571.

Tremblay, P., Small, S.L., 2011b. From language comprehension to action understanding and back again. Cerebral Cortex 21, 1166–1177.

Turkeltaub, P.E., Coslett, H.B., 2010. Localization of sublexical speech perception components. Brain and Language 114, 1–15.

Vandenberghe, R., Price, C., Wise, R., Josephs, O., Frackowiak, R.S.J., 1996. Functional anatomy of a common semantic system for words and pictures. Nature 383, 254–256.

Vigliocco, G., Vinson, D.P., Lewis, W., Garrett, M.F., 2004. Representing the meaning of object and action words: the featural and unitary semantic space hypothesis. Cognitive Psychology 48, 422–488.

Vigneau, M., Beaucousin, V., Hervé, P.Y., Duffau, H., Crivello, F., Houdé, O., Mazoyer, B., Tzourio-Mazoyer, N., 2006. Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. Neuroimage 304, 1414–14132.

Vigneau, M., Beaucousin, V., Hervé, P.Y., Jobard, G., Petit, L., Crivello, F., Mellet, E., Zago, L., Mazoyer, B., Tzourio-Mazoyer, N., 2011. What is right hemisphere contribution to phonological, lexico-semantic, and sentence processing? Insights from a meta-analysis. NeuroImage 541, 577–593.

Villarreal, M., Fridman, E.A., Amengual, A., Falasco, G., Gerscovich, E.R., Ulloa, E.R., Leiguarda, R.C., 2008. The neural substrate of gesture recognition. Neuropsychologia 46, 2371–2382.

von Kriegstein, K., Eger, E., Kleinschmidt, A., Giraud, A.L., 2003. Modulation of neural responses to speech by directing attention to voices or verbal content. Brain Research. Cognitive Brain Research 17, 48–55.

Wager, T.D., Smith, E.E., 2003. Neuroimaging studies of working memory: a meta-analysis. Cognitive, Affective, and Behavioral Neuroscience 3, 255–274.

Wager, T.D., Vazquez, A., Hernandez, L., Noll, D.C., 2005. Accounting for nonlinear BOLD effects in fMRI: parameter estimates and a model for prediction in rapid event-related studies. Neuroimage 25, 206–218.

Wagner, A.D., Koutstaal, W., Maril, A., Schacter, D.L., Buckner, R.L., 2000. Task-specific repetition priming in left inferior prefrontal cortex. Cerebral Cortex 10, 1176–1184.

Wagner, A.D., Pare-Blagoev, E.J., Clark, J., Poldrack, R.A., 2001. Recovering meaning: left prefrontal cortex guides controlled semantic retrieval. Neuron 31, 329–338.

Wakusawa, K., Sugiura, M., Sassa, Y., Jeong, H., Horie, K., Sato, S., Yokoyama, H., Tsuchiya, S., Inuma, K., Kawashima, R., 2007. Comprehension of implicit meanings in social situations involving irony: a functional MRI study. NeuroImage 37, 1417–1426.

Wallentin, M., Ellegaard, T., Ostergaard, S., Ostergaard, L., Roepstorff, A., 2005. Motion verb sentences activate left posterior middle temporal cortex despite static context. NeuroReport 16, 649–652.

Wildgruber, D., Ackermann, H., Grodd, W., 2001. Differential contributions of motor cortex, basal ganglia, and cerebellum to speech motor control: effects of syllable repetition rate evaluated by fMRI. NeuroImage 13, 101–109.

Willems, R.M., Hagoort, P., Casasanto, D., 2010a. Body-specific representations of action verbs: neural evidence from right- and left-handers. Psychological Sciences 21, 67–74.

Willems, R.M., Toni, I., Hagoort, P., Casasanto, D., 2010b. Neural dissociations between action verb understanding and motor imagery. Journal of Cognitive Neuroscience 22, 2387–2400.

Willems, R.M., Ozyurek, A., Hagoort, P., 2007. When language meets action: the neural integration of gesture and speech. Cerebral Cortex 17, 2322–2333.

Willems, R.M., Ozyurek, A., Hagoort, P., 2009. Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. Neuroimage 47, 1992–2004.

Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nature Neuroscience 7, 701–702.

Xu, J., Kemeny, S., Park, G., Frattali, C., Braun, A., 2005. Language in context: emergent features of word, sentence, and narrative comprehension. NeuroImage 25, 1002–1015.

Yang, J., 2012. The role of the right hemisphere in metaphor comprehension: a meta-analysis of functional magnetic resonance imaging studies. Human Brain Mapping. http://dx.doi.org/10.1002/hbm.22160.

Zatorre, R.J., 2001. Neural specializations for tonal processing. Annals of the New York Academy of Scienes 930, 193–210.

Zempleni, M.Z., Haverkort, M., Renken, R.A., Stowe, L., 2007. Evidence for bilateral involvement in idiom comprehension: an fMRI study. NeuroImage 343, 1280–1291.

Zwaan, R., 2004. The immersed experiencer: toward an embodied theory of language comprehension. Psychology of Learning and Motivation 44, 35–62.

Zysset, S., Huber, O., Ferstl, E., von Cramon, D.Y., 2002. The anterior frontomedian cortex and evaluative judgment: an fMRI study. NeuroImage 15, 983–991.

Laplace, Pierre Simon, Marquis de (1749–1827)

Bernard Bru, Université René Descartes Paris V, France

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous editions, volume 12, pp. 8368–8372, © 2001, Elsevier Ltd.

Abstract

Trained in the grand Analysis of Euler, d'Alembert, and Lagrange, Laplace found himself at a scientific crossroads which was unique in the modern history of science. In fact, in the last quarter of the eighteenth century, Laplace was to solve in succession a large number of the problems left unresolved by Newtonian theory. Celestial mechanics, which had seemed limited to the description of general phenomena, was able to explain the system of the world down to its finest details and with unlimited precision; scientific determinism was thus possible and Laplace's *Celestial Mechanics* was a decisive proof of this. This constantly enhanced precision could reach to a fine degree the then inaccessible areas of the moral sciences. The second Laplacian message, the *Analytical Theory of Probability*, allowed the scholar to assess data from the whole of nature, and in the long term reveal the 'constant causes' which affect and determine human society. It could also, in the absence of a large number of observations, guide man in his choices by eliminating arbitrariness and error as far as possible. Laplace's considerable scientific work, spanning the years from 1770 to 1827, was to consist in constantly asserting that mathematical analysis could allow Mankind to become master of its universe with an unparalleled precision and efficacy.

Pierre-Simon de Laplace was born on 23 March 1749 in Beaumont-en-Auge into a well-to-do family of the Norman landed gentry which occasionally married into the nobility. Destined for the priesthood, he studied first at the College de Beaumont under the direction of the Benedictines of Saint-Maur. Endowed with an excellent memory and a keen intellect, Laplace seems to have excelled at the College, where he received the classical education of the time – the great Latin and French authors, in particular Virgil, Cicero, and Racine, whom he liked to cite and who are perhaps the origin of his obvious liking for well-turned phrases and his passion for physical astronomy. At the age of 16, he entered the College of Arts at the University of Caen for the 2 years of philosophy prior to the Degree in Theology. Here he discovered mathematical analysis in his dealings with natural philosophy, Euler, Clairaut, d' Alembert, and Daniel Bernoulli, and at the same time discovered his astonishing aptitude for handling formulae. On the advice of his tutors, he abandoned theology and moved to Paris in 1769, where he was very quickly noticed by d' Alembert. From then on, he was never to leave either Paris or analysis.

Programme of Studies

As Professor of Mathematics at the Ecole Royale Militaire, Laplace was literally to inundate the Académie des Sciences with mathematical papers, some 15 in 4 years, certain ones of which ran to well over one 100 pages in quarto. On 24 April 1773, he was elected *adjoint mécanicien* at the Académie des Sciences in Paris. The subjects which he tackled were those which he would continue to explore for the whole of his scientific life, a life which was to prove exceptionally long, especially as he would continue publishing original and fundamental papers right up to his death in Paris on 5 March 1827.

Becoming very quickly aware of his intellectual superiority, Laplace set about tackling all the problems of his day in the three principal academies where exact sciences were fostered – Paris, Berlin, and St Petersburg. Indeed, one need only read the titles of his first papers to get a good idea of the state of mathematical sciences in the last quarter of the eighteenth century. First and foremost came differential and integral calculus, and in particular the study of differential equations and partial differential equations which governed the mechanics of solids as well as fluids and the stars. In a paper presented in 1773 (Laplace, 1912, vol. 9, pp. 5–68), it was Laplace who was the first to demonstrate that in very general cases one could determine the whole integral of linear partial differential equations, and that in exceptional cases it was still possible to obtain from them particular solutions in the form of *defined integrals* that is, *integral transforms* which fitted with the initial data. This method would allow him, for example, in 1809, to solve the equation of heat in more general cases than those envisaged by Fourier.

The second subject tackled by Laplace equally concerns the solutions of differential systems, but in those cases where whole integrability is beyond reach, such as in the theory of the Moon or the large planets. These theories had not only eluded Newton, who had not used the infinitesimal algorithm in its full extension, but had also eluded Euler, Clairaut, d' Alembert, and Lagrange who were ingenious rivals in their use of it. Laplace, right from the outset, was to demonstrate his extraordinary faculty – often spoken of in terms of intuition – for reducing by approximation the irreducible difficulties present in equations dealing with the problems of multiple bodies. In March 1773, Laplace presented a paper (Laplace, 1912, vol. 8, pp. 201–275) on the 'principle of universal gravitation' in which he showed that Euler and Lagrange had been wrong in their planetary theory and that he, Laplace, at the age of twenty-four, through Newton's single principle that 'the planets gravitate against each other,' could show the absence of secular inequalities in planetary mean motions, so that the planets could not accelerate their motion around the sun irreversibly by themselves. Furthermore, they only deviated from their perfectly elliptical trajectories through periodic and temporary inequalities, thus assuring a stability in the 'system of the world' which Laplace would encounter again in all areas of his work. The mechanics of the planets is quasi-stable, and

the God of Newton played no useful role in maintaining them within rational limits. The foundations for Laplace's astronomical work were established, a work which he never once doubted and which he would pursue from then on, right up to his death.

The third subject, the 'theory of chance,' is more surprising coming from a protégé of d'Alembert, but it was of great topical interest in 1770. It was in fact in this year that the great academies suddenly resumed this subject, which had to some extent been abandoned for 50 years, and which was now brought back into fashion by works in political arithmetic – notably those of Süssmilch – and at the same time by the requirements of astronomy. Lambert, Euler, Daniel Bernoulli, Lagrange, and Condorcet simultaneously published important papers on *probability calculus*. It was therefore inevitable that the young Laplace would turn his efforts to it, and his first contributions were magnificent, notably the great paper 'on the probability of causes' published in 1774 (Laplace, 1912, vol. 8 pp. 27–65), which marked the beginning of Laplacian mathematical statistics. In this paper, one of the classic works of the theory, Laplace developed the analytical results of the proportionality principle from the probability of causes to that of effects, and, for the first time, proposed, as a criterion for estimation, a quantity known only from observations subject to unavoidable errors, this criterion serving to minimize the sum of potential errors through their respective probability. This method, which Laplace called the *method of situation* and which appeared to him to be the result of the nature of things, should, he believed, be applied to research in astronomical fields. In fact, it led to formidable mathematical difficulties which Laplace would, in the next 50 years, work on breaking down into a series of fundamental works which would culminate in 1812 in the publication of the *Analytical Theory of Probability* (Laplace, 1820).

Analysis, *celestial mechanics*, and the *theory of probability* were three overlapping subjects which Laplace did not invent and which he already found in the academic papers of his time, but which he grappled with and never abandoned, to such an extent that he was to change the view scholars had formed of the world which surrounded them, a physical world, of course, but also a moral world from the moment when it could be converted into numbers.

Celestial Mechanics

The years which were to follow were a long succession of scientific triumphs. One after another, Laplace solved the most difficult problems in celestial mechanics posed since Newton, problems which had foiled the attempts of the most eminent scholars of the eighteenth century to such an extent that it was sometimes doubted whether the single principle of attraction could really account for the inequalities of motion – such as the secular inequality of the Moon, the inequalities of Jupiter and Saturn, the theory of the satellites of Jupiter, and so on – which were being brought to light by the increasingly precise observations of modern astronomy. Laplace's works as a whole – which were spread over 20 years – were brought together in a vast treatise, *Celestial Mechanics* (Laplace, 1825), the publication of which began in 1799, and which Laplace would continue to expand and amend up to the end of his life. Concurrently, in 1796 he published *Exposition of the System of the World* (Laplace, 1984), which was aimed at a wider readership and presented, in layman's terms, the principal results which demonstrated with the utmost clarity that Newton's law of universal gravity 'represents all celestial phenomena, right down to their smallest details' and that 'through this, empiricism has been entirely removed from Astronomy, which is now a major problem of mechanics.'

However, Laplace was well aware of the fact that whole areas of the system of the world existed where empiricism seemed difficult to exclude entirely, such as the theory of oceanic or atmospheric tides; in other words, the areas where 'irregular causes' occurred in an anarchic fashion which could, if one did not bear them in mind, conceal the true causes of phenomena and their fixed order. Laplace encountered this problem from the beginning of the 1770s to the moment when the theory of errors began to take shape. Observations of all parts of nature, even the most precise, were affected by 'unavoidable errors' which could prove themselves predominant if left to accumulate by chance, and remove all scientific value from the apparent objectivity. According to Laplace, it was therefore necessary to multiply the observations, adjust the value of the uncertain elements progressively, and reveal gradually the phenomena which were 'until this point shrouded in the errors of observations.' This progressive revealing could not be done without a method, and this is precisely the purpose of Laplace's probability calculus.

The Analytical Theory of Probability

For Laplace, probability theory was first conceived as a means of critically analysing data from observation, a critical analysis which gradually became more refined and revealed the true system of the world. Two periods in Laplace's development are usually distinguished. The first, which runs from 1774 to 1785, saw the development of Laplace's first method. In the spirit of Bayes, but essentially asymptotic in that it was independent of the special nature of *a priori* laws, it did not allow Laplace to deal with the major problem of the theory of errors, that is, to determine in the best possible way the elements known from a large number of indirect observations. It did, however, allow him to solve the population theory problems posed by Daniel Bernoulli and Condorcet by giving political arithmetic the scientific rigor it lacked, left open as it was to the most trivial empirical digressions; for example, in deciding whether in truth more boys were born in London than in Paris for the same number of births, or even whether the population of France was increasing or decreasing.

The second Laplacian asymptotic theory, this time non-Bayesian, began in 1810 after its author had solved one of the major problems of probability theory, the *central limit problem* as it is called today, where under very general conditions the total sum of errors is asymptotically normal, which allows the calculation of probability even if the particular law of errors is unknown. This remarkable theorem and its applications in the *method of least squares*, published the following year, makes up Chapter IV of Book II of the *Analytical Theory of Probability*, published by Laplace in 1812 (Laplace, 1820), and

bringing together his work on probability as a whole. This second great treatise from Laplace would be expanded with various different chapters until 1825. From 1816, Laplace gradually extended the applications of probability calculus into increasingly wider fields, anticipating a large number of the applications which the twentieth century would develop. The theory of probability, therefore, became for Laplace 'the most felicitous addition to the ignorance and weakness of the human spirit' as he asserts in conclusion to his *Philosophical Essay on Probability* (Laplace, 1986), the first edition of which dates back to 1814 and the last to 1825.

Laplace did not limit the possibilities for the application of probability calculus to traditional fields such as games, the theory of errors, and annuities. From the moment when a shred of calculation seemed to him possible, whatever the field and however tenuous it may be, he got to grips with it and generally exceeded previous results by a great deal. For example, inspired by the works of Daniel Bernoulli, Laplace moved from population theory to a general reflection on the chance series of draws from a ballot box which, at first, only display disorder and chaos and gradually adjust and give way to the simplest and most admirable order, the irregular causes cancelling each other out to display the eternal action of constant causes. These reflections on order in chaos anticipate the work on statistical physics at the beginning of the twentieth century. What is more, Laplace, inspired by Condorcet, proposed a theory of decisions in majority voting where the basic concepts of modern test theory can be seen. Further still, inspired once again by Daniel Bernoulli, Laplace established a general result for the concave utility functions, and applied it to questions of maritime insurance and so on. One can also see many more things in the *Analytical Theory* (Laplace, 1820), which make this work a monument of science for all time. (See Bibliography for complete studies on this subject.)

Career

After the flamboyant beginnings we have seen, Laplace was very quickly recognized by his peers, and if his arrogance had caused him some trouble at the beginning, he had an exceptional career. Elected in 1783 to the position of Associate Academician, he became a member of the Académie des Sciences in Paris in 1785. During the French Revolution, Laplace retired to a property near Paris to draft out *Celestial Mechanics* and kept himself at a distance from events. But with the fall of Robespierre, he returned to Paris and took an active part in the creation of three great scholarly institutions emerging out of the Revolution – the Institut de France, the Bureau de Longitudes, and the Ecole Polytechnique. Bonaparte, who was concerned about his image and the scientific glory of France, was to make Laplace, albeit short-lived, his First Minister of the Interior and soon one of the highest dignitaries of the Empire, the Chancellor of the Senate, and Count of the Empire.

In 1814 Laplace voted with the majority of the Senate for the forfeiture of the Emperor. He was immediately made a peer of France, then marquis, by King Louis XVIII. In the numerous positions he held, Laplace made himself the herald and the defender of exact sciences. He involved himself little in politics, and when he did, only with questions he could reduce to numbers.

In his lifetime, Laplace enjoyed an immense scientific prestige. As founder of the Société d' Arcueil in 1807, he addressed all the scientific debates of his time, and created a brilliant school which would make his mark on part of his century, up to the point at which Laplacian physics of corpuscular fluids was superseded by positive mathematical physics and experimental physics (on this subject, see Gillispie, 1997).

In 1788, Laplace married Charlotte de Courty (1769–1862) with whom he had two children, Charles (1789–1874), general and peer of France who died without descendants, and Sophie (1792–1813) who died in childbirth. Angelique (1813–89), the only daughter of Sophie, was brought up by her grandparents and married the marquis de Colbert Chabanais and had three children, through whom the family line was established.

Influence

Laplace was one of the most influential men of science of all time, as much through his scientific work as through his two expository works which met with considerable success. There was no public or private library which did not possess them, and Laplace remains among the most cited scholars in present-day literature. His contribution to the formal assertion of mechanical determinism is sometimes summarized thus: 'All events, even those which by their insignificance do not seem to be due to the great laws of nature, are a consequence of them which is as necessary as the revolutions of the sun.' This introductory sentence from the *Philosophical Essay* (Laplace, 1986) of 1814 which can already be found in one of Laplace's first philosophical texts written in 1773 is without doubt the most frequently occurring scholarly quotations in the whole of world literature. In fact, as is well known, the assertion of determinism is a recurrent theme in the history of human thought, and countless authors in the eighteenth century in particular could be credited with it. Laplace's originality lies in producing a brilliant proof; that the tiniest disturbances in the system of the world – for example, the slow acceleration of the Moon – are a necessary consequence of the single best established law of nature, that of universal gravity. It was because such a deduction was possible, as Laplace demonstrated, that determinism could be asserted and, even more, that it was then a question of one of the highest scholarly ambitions of humanity. Science in the nineteenth century became a race for precision made possible by the local determinism of phenomena in the whole of nature.

However, Laplace's influence extends well beyond that particularly in the social and moral sciences of the nineteenth century, most notably in the work of Quetelet. Indeed, 'Laplacian determinism' was not limited to determinist phenomena. In the system of the world, there are numerous phenomena which are subject to irregular accidental causes which constantly alter the course of these phenomena. Necessity, order, and determinism are nevertheless not absent from them. For example – and it is an example dear to Pastor Süssmilch which Daniel Bernoulli took up again – the sex at

birth does not seem to be subject to any determinism, yet one observes that always or almost always slightly more boys are born than girls as soon as one considers quite a large number of births. Laplace concludes that the invariability involved here is 'a result of regular causes which override the anomalies due to the chance' of irregular causes (Laplace, 1912, vol. 7, p. xlviii). Chance is thus furnished with a special meaning; it is the combination of irregular causes which take effect in all directions, without favoring any one in particular. Regular causes, on the other hand, compel nature along enforced paths which become apparent in the long term. This is Laplace's *theory of constant causes* which one can clearly see in the works of other authors, for example Buffon, Moivre, and the Bernoullis, but also in the works of the ancient Stoics. It is Laplace, however, who, just as in the case of determinism in minor astronomical perturbations, provides this theory with a mathematical provision which justifies its implementation. This is the full import of his theorem of 1810; the variations due to *irregular causes* are enclosed in narrow limits determined by the Laplace formula, which gradually move closer to genuine elements in such a way that the constant causes determine them. The analysis of causes which Adolphe Quetelet undertook of the 'phenomena which organised beings present us with' is no different, except in that it integrates into the Laplacian constant causes–accidental causes schema a third type of causes, *variable causes* which, for example, produce seasonal effects, causes which Poisson, Cournot, and Bienaymé had envisaged in order to account for irregularly periodic statistical data. Quetelet's 'average man,' and the 'tendencies to crime,' can only be understood in the perspective of the constant causes in Laplace's *Essay*. In a still more obvious way, Poisson's 'law of large numbers' is a triumphant metaphor for Laplacian theory, or at least for certain aspects of it – showing as it does the order of constant causes within the apparent chaos of accidental causes.

The decision-making aspect of Laplace's work – the theory of probability seen as 'common sense reduced to calculation' – experienced a less glorious fate. After the Laplace school – Poisson, Cournot, and Bienayme – had attempted to give it a more acceptable form in the first half of the nineteenth century, it was for a long time rejected, and with this rejection went an increasingly significant part of Laplace's work on probability. Causes were multiple and, moreover, 'variable' since this aspect a century later was reinstated at the highest level in the way that it is best known today – the theory of games, the theory of decisions, the theory of tests, and judgments based on samples. For Laplace, chance was also the name for our ignorance, an ignorance which was often incurable owing to those multiple accidental causes which distort the natural order of things and of our reason, so that the eminent role of the theory of chance, of probability calculus, is to 'criticize' the statistical data from all parts of nature; and even in the absence of data, by analogy, image, model, and calculation when possible, the theory still allowed one to find one's way more confidently through the maze of social and economic life. In France one would have to wait for Borel and his theory of games before one considered afresh along with Laplace that the theory of probability 'leaves nothing arbitrary in the choice of opinions and in the choice of decisions to make, every time one can determine, in one's way, the most advantageous choice.' Laplace's legacy – to perfect everything through the powerful means that mathematical analysis places at the scholar's disposal – is in no danger of being forgotten.

See also: Bayesian Statistics; Bayesian Theory, History, Applications, and Contemporary Directions; Elicitation of Probabilities and Probability Distributions; Errors in Statistical Assumptions, Effects of; Probability Density Estimation; Probability Theory: Formal; Probability Theory: Interpretations; Probability and Chance: Philosophical Aspects; Socialism: Historical Aspects; Statistics, History of; Test Theory: Applied Probabilistic Measurement Structures; Unemployment: Structural.

Bibliography

Gillispie, C.C., 1997. Pierre-Simon de Laplace 1749–1827. A Life in Exact Science. Princeton University Press, Princeton, NJ.

Haid, A., 1998. A History of Mathematical Statistics from 1750 to 1930. Wiley, New York.

Laplace, P-S de, 1820. Theorie Analytique des Prob (Analytical Theory of Probability). Ve Courcier, Paris.

Laplace, P-S de, 1825. Traité de Mécanique Céleste (Celestial Mechanics). Gauthier-Villars, Paris.

Laplace, P-S de, 1912. Oeuvres Compèltes de Laplace (Complete Works), vol. 14. (1878–1912). Ve Courcier, Paris.

Laplace, P-S de, 1984. Exposition du Systéme du monde (The System of the World). Artheme Fayard, Paris.

Laplace, P-S de, 1986. Essai philosophique sur les probabilités (Philosophical Essay on Probability). Bourgois, Paris.

Stigler, S.M., 1986. The History of Statistics: The Measurement of Uncertainty Before 1900. Harvard University Press, Cambridge, MA.

Large-Scale Social Surveys, Analysis of

Elaine Zanutto, University of Pennsylvania, Philadelphia, PA, USA
Andrew Gelman, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E. Zanutto, A. Gelman, volume 12, pp. 8386–8392, © 2001, Elsevier Ltd.

Abstract

Large-scale social surveys are an important source of information for a wide range of topics. Analysis of such surveys can be challenging. Given the difficulties of constructing a sampling frame, implementing probability sampling, and obtaining a high response rate, it is no surprise that analysts must often make major efforts to adjust for differences between sample and population, using tools including weighting, imputation, regression modeling, and poststratification.

Quick Recommendations

Large-scale social surveys are an important source of information for a wide range of topics. In analyzing such surveys, it is important to be aware of the complexity of the sampling design and the data adjustments that are used by survey organizations, including weighting to adjust for differences between sample and population and imputation to fill in missing responses. For estimating population means and proportions, the analyst should calculate weighted averages, using survey weights. In regression analysis, it is best to include, among the predictors, the variables that are used in the design and the weighting. Hierarchical models can be used to obtain inferences for small subpopulations, and poststratification can be used to aggregate cell-level estimates to obtain inferences for larger segments of the population. In general, analyzing survey data as if they had been obtained through simple random sampling can result in biased point estimates and underestimated standard errors. These calculations can often be adjusted reasonably well by using design effects reported in the survey documentation or in previous analyses. For more involved analysis, it is advisable to use statistical packages that are specifically designed to correctly handle complex survey data, and to explicitly adjust for variables with mismatch between sample and population rather than simply relying on unit-level weights.

Background

Large-scale social surveys provide an inexpensive and easily accessible source of information on a wide range of topics including education, health, economics, demography, politics, and criminal justice. These surveys may serve as a primary data source, a preliminary data source to be analyzed prior to conducting a more tailored survey, or in combination with smaller, more specific data sources. After summarizing the advantages of using large-scale social surveys, the authors discuss how to confront the challenges they present.

There are many advantages to analyzing data from large-scale social surveys. The data are usually extensive in content and coverage of the target population. Sample sizes are often very large and the data are often easily obtained from the sponsoring agency on CD-ROM or from the Internet, in many cases at no cost to the researcher. Moreover, these surveys are typically conducted by large government agencies or private survey organizations that have experienced interviewers, well-designed and tested survey instruments, and the expertise to design efficient sampling plans. As a result, the response rates are typically high (e.g., 75–90%) and sophisticated nonresponse adjustments are often incorporated into the data to try to reduce nonresponse bias that results from systematic differences between respondents and nonrespondents. The sample data are also often reweighted, using poststratification, to agree with known population distributions of characteristics such as age, sex, and race. Finally, many surveys contain panel (longitudinal) components in which the same sample units are interviewed repeatedly over time, allowing researchers to follow individual changes over time.

The analysis of large-scale social survey data also presents several difficulties. One of the greatest technical challenges arises because sampling in these surveys is usually carried out according to complex multistage designs that combine stratification, cluster sampling, and unequal probability sampling. Many surveys that are repeated on a regular basis also rotate respondents in and out of the sample. Often, point estimates of simple estimands such as population means and proportions can be calculated using a minimal amount of information about the sampling design, but standard error calculations must incorporate the key sampling design elements explicitly. Survey weights, provided along with the survey data, contain important sampling design information. Correct interpretation and use of these weights is key to the analysis of large-scale social survey data, and particularly key to standard error estimation. Analysts must also be aware of the impact of nonresponse adjustments – such as weighting class adjustments for unit nonresponse and imputation for item nonresponse – on standard error estimates. Complex multistage sampling must also be taken into consideration when calculating confidence intervals for population parameters. In particular, the degrees of freedom used when calculating confidence intervals using a t-distribution depend on the sampling design, not solely on the sample size. Methods of incorporating survey weights into estimates of simple population parameters such as means, totals, and proportions are well established; however, the appropriate approach is less obvious when fitting more complex statistical models such as linear regression models and loglinear models. The authors address each of these issues in the following sections.

Complex Sample Designs

Although simple random samples are the easiest to design and analyze, large-scale social surveys often take advantage of complex multistage sampling designs to increase the precision of the resulting estimates. These designs typically involve multiple stages of sampling. For example, the first stage of sampling may select a sample of counties; the second stage may select a sample of subregions within each sampled county; the third stage may select a certain number of sample units (e.g., individuals, households, or hospitals) within each sampled subregion. Units selected at the first stage of sampling are called primary sampling units (PSUs). Units selected at the final stage are called ultimate sampling units. One reason for multistage sampling is to minimize the travel costs of interviewers in face-to-face surveys. Another reason is that a sampling frame (a list of all the ultimate sampling units in the population) may not exist for the target population, but may be constructed sequentially as needed. For example, a list of all school-age children in a state may not be readily available, but a list of all school districts in a state may be available; then for a sample of school districts, a list of all the schools in those districts may be obtained; then for a sample of schools, a list of children may be obtained.

Multistage designs often involve cluster sampling in which groups of sampling units are selected, rather than individual units. Examples of cluster sampling include selecting several students in a class, several households in a city block, or several patients in a hospital. Cluster sampling tends to increase the variability of survey estimates compared with a simple random sample of the same size. This occurs because units in the same cluster tend to be correlated and, as a result, each additional unit sampled within the same cluster provides less additional information about the overall population parameters than an element selected at random from the population. However, due to travel costs, frame construction costs, and other overhead costs associated with each sampled cluster, a simple random sample may be much more expensive than a cluster sample of the same size. Therefore, given a fixed budget, a cluster sample may provide the most efficient estimates.

Another common design element used in large-scale multistage surveys is stratification, which involves dividing the population into nonoverlapping segments, called strata, and sampling independently within each stratum. Typical stratifying factors include region of the country and urbanicity (urban, suburban, and rural). If the variable of interest tends to take on different values, on average, in each stratum, then stratified sampling can lead to more efficient estimates. In particular, the sample can be optimally allocated so that a larger sample is obtained in strata where there is larger variability in the variable of interest. Another reason for stratification is to ensure adequate representation of all segments of the population, avoiding the possibility of a simple random sample that, by chance, omits part of the population. Often certain strata, such as those containing a large proportion of racial minorities or elderly people, are sampled at a disproportionately greater rate (oversampled) in order to ensure an adequate sample size for research hypotheses that involve these groups. For example, the British Crime Survey oversamples inner city areas by a factor of two (Mirrlees-Black et al., 1998). Sometimes a census is conducted in some strata, called self-representing or take-all strata. This type of stratification is particularly useful for very skewed populations that contain a few large units and many small units (e.g., Hidiroglou, 1986). Stratification can also be combined with cluster sampling. For example, neighborhoods may be stratified by geographic region and then a sample of households within neighborhoods may be selected; or schools may be stratified by size and then a sample of classes may be selected from each school.

Unequal probability sampling is often used when sampling clusters. In particular, clusters are usually selected with probabilities proportional to their size (PPS) sampling, with larger clusters having greater probabilities of selection. The size measure used will depend on the survey, but, as examples, a school can be measured in terms of the number of students enrolled, a company can be measured in terms of the numbers of employees, a farm can be measured in terms of acreage, and a city block can be measured in terms of the number of households. PPS sampling will lead to more efficient estimates whenever cluster totals are proportional to the size of the cluster, which is often the case. For example, larger schools tend to have more students with learning disabilities, larger companies tend to have larger revenue, larger farms tend to have a larger crop yield, larger city blocks tend to have more single parents. Unequal probability sampling is also sometimes used for convenience. For example, random-digit-dial telephone surveys actually select households with probabilities proportional to the number of voice telephone lines in the household. Selecting a simple random sample of households would require a list of the individual households, rather than a list of telephone numbers, and this may be much more difficult to obtain or construct than a list of telephone numbers.

Finally, many large surveys that are repeated on an ongoing basis (e.g., monthly) use rotation groups. Rather than select new nonoverlapping samples each month, rotation groups remain in the sample for several months in a row (Fienberg and Tanur, 1983; Iachan and Jones, 1987). Every month, a new rotation group enters the survey and an old one drops out so that each month only a fraction of the sample units are new. This reduces costs by not having to draw a new sample each month (there are administrative costs with drawing a new sample, locating the new sample members, explaining the survey to them, etc.) and also provides some longitudinal information. Rotation groups present no complication to cross-sectional analysis of a single wave of data but can complicate analyses that combine several months of data since the same sample units may appear multiple times in these combined datasets. Analysis should then account for correlations between repeated measurements on some sample units. In contrast, some surveys, called rolling surveys (Kish, 1998), are specifically designed with nonoverlapping samples each month so that the resulting data can be analyzed for a single month or combined over several months to form larger datasets.

Survey Weights

The simplest analysis of survey data assumes that all elements in the population are equally likely to be included in the sample. Unfortunately, this is not generally true for two

reasons. First, complex sampling designs often lead to unequal selection probabilities, as discussed in the previous section. Second, even if units are equally likely to be selected, they generally have different probabilities of response; for example, in household surveys, women are usually easier to reach and more likely to cooperate than men.

Survey weights are used for correcting for known or expected differences between sample and population; each unit i is given a weight, w_i, and the population mean of any variable y is estimated by the weighted sample average, $Sum_i\ w_i y_i / Sum_i w_i$.

As the formula indicates, only the relative values of the weights are relevant. The weights are usually calculated by the survey producers and provided along with the survey data. An informal interpretation of these weights is that w_i is proportional to the number of units in the population represented by respondent i. Estimates of population proportions are calculated using the same formula, where y_i is a variable taking on values of 0 and 1.

Survey weights in large-scale surveys are often constructed so that sample estimates agree with current census estimates of key variables such as age, sex, and race, through a process called poststratification and raking (Oh and Scheuren, 1983; Gelman and Carlin, 2001). Voss et al. (1995) detail the various design and weighting strategies used by various national polling organizations for preelection surveys. Survey weights are also often adjusted to account for nonresponse, through a process called weighting class adjustment (Oh and Scheuren, 1983; Kalton and Kasprzyk, 1986; Bethlehem, 2001).

In most complex multistage samples, sampled units have different survey weights. However, it is possible to design these samples so that all sampled units have the same weight. These designs are called self-weighting (Cochran, 1977: p. 91) because the usual unweighted estimators, the same ones used for simple random samples, will produce unbiased estimates of population parameters such as means and proportions. However, the usual unweighted standard error estimation formulas are not necessarily appropriate in these designs. Therefore, although self-weighting designs may simplify point estimation calculations, they do not simplify standard error estimation calculations.

Finally, many surveys produce separate sets of weights for different types of analyses. For example, the US General Social Survey (GSS) (Davis and Smith, 1992) provides separate weights for household-level estimates and for individual-level estimates. The GSS is designed to give each household an equal probability of inclusion in the sample so that the sample is self-weighting for household-level variables. However, since only one individual in each household is interviewed, people living in large households are less likely to be interviewed than people in smaller households. Therefore, even though each household has an equal chance of being selected, individuals have different chances of being selected depending on the size of their households, with these inclusion probabilities varying by the type of survey (Gelman and Little, 1998). Longitudinal studies may also have separate cross-sectional and longitudinal weights. For example, the 1993 US Panel Study on Income Dynamics (PSID) provides both 'cross-sectional analysis weights' and 'longitudinal analysis weights' (Hill, 1992). Cross-sectional analysis weights can be used to obtain unbiased estimates of population statistics for a given year. Longitudinal analysis weights, on the other hand, can be used to estimate models of changes over time. In particular, the 1993 PSID longitudinal analysis weights assign nonzero weights only to individuals who were respondents during each of the preceding 5 years. The Canadian Survey of Income and Labour Dynamics also provides both longitudinal and cross-sectional weights (Statistics Canada, 1997). When using surveys with multiple sets of weights, consult survey documentation for advice on appropriate choice of survey weight.

Standard Errors

Although estimates of simple population parameters such as means and proportions can easily be constructed as weighted averages using survey weights, the corresponding standard error estimation procedures are not included in most standard statistical packages. When the data have been collected from a survey with a complex sample design, standard error estimates of survey statistics derived under simple random sampling assumptions generally underestimate the true standard error, resulting in confidence intervals that are too narrow and hypothesis tests that are too likely to reject the null hypothesis. In particular, most standard statistical packages can handle 'analytic' weights that are used in regression estimation when the data values themselves each have different (known) variances, and 'frequency' weights that are used to store data concisely by recording the number of times an observation is duplicated in the dataset rather than recording each observation individually, but these weights should not be confused with survey weights. Incorrectly treating survey weights as analytic weights or frequency weights will generally result in unbiased point estimates but underestimated standard errors. Moreover, ignoring the weights altogether and using a standard statistical package to perform an unweighted analysis of complex survey data will generally yield biased estimates of population parameters and underestimated standard errors. Furthermore, standard statistical packages do not account for stratification, cluster sampling, or adjustments such as post-stratification and raking. Brogan (1998) and Wang et al. (1997) illustrate some of the potential consequences of incorrectly analyzing survey data using standard statistical packages.

As a measure of the effect of the sampling design on the variance of survey estimates, some surveys report design effect estimates. The design effect (deff) (Kish, 1965: p. 162; Lohr, 1999: p. 239) measures the increase (or decrease) in variance obtained using the given sample design relative to the variance that would have been obtained if a simple random sample of the same size had been used. For example, Davis and Smith (1992: p. 41) report that the average design effect for 60 variables in the GSS is 1.76, with slightly larger design effects on average for demographic variables, which tend to have more clustering and smaller design effects, on average, for attitude variables. Similarly, Hill (1992: p. 67) reports design effects for estimates from the PSID ranging from 0.95 to 2.53, with most in the range 1.1–2.2. A design effect of 1.76 indicates that the variance of an estimate from the current design is 1.76 times greater than the variance that would have been obtained using a simple random sample of the same size. The design effect can be used to determine the 'effective sample

size' by dividing the actual sample size by the design effect. The effective sample size is the number of observations that would yield the same level of precision from a simple random sample. For example, a design effect of 1.5 implies that the precision obtained by sampling 1500 units under the current complex sampling design could be obtained by sampling only 1000 in a simple random sample design. The design effect does not, however, account for differential costs of the two sampling designs.

Several software packages are available to analyze data from complex multistage surveys. Summaries and reviews of these packages are provided by Lepkowski and Bowles (1996), Cohen (1997), Carlson (1998), and Zaslavsky (2000). Stata (StataCorp, 1999) and the latest release of SAS (An and Watts, 1998) are multipurpose statistical packages that include procedures for analyzing complex survey data. Because variance estimation calculations for sampling designs that use without-replacement unequal probability sampling are complex, most survey software packages approximate these variances using formulae appropriate for unequal with-replacement sampling. This approximation is usually acceptable since, for these survey designs, the resulting variance estimates will slightly overestimate the true variance, resulting in slightly conservative confidence intervals and hypothesis tests. Finally, survey software packages generally use variance formulae that are technically only appropriate when the survey weights represent only the sampling design and do not incorporate nonresponse or poststratification adjustments. A few packages (WesVarPC and VPLX, see Lepkowski and Bowles, 1996) use a variance estimation method call replication or resampling (Wolter, 1985: Chapter 2; Korn and Graubard, 1999: p. 29) that can incorporate the impact of nonresponse weighting adjustments and poststratification adjustments. To take advantage of these capabilities, data users must be provided with 'replicate weights' (e.g., Korn and Graubard, 1999: p. 34) by the data producers. For example, the US Census Bureau releases replicate weights for the Survey of Income and Program Participation.

Finally, many large-scale surveys provide generalized variance functions (Korn and Graubard, 1999: p. 218; Lohr, 1999: pp. 308–310) in summary reports of the survey data. To reduce publication costs and calculation effort, these surveys do not publish standard errors for every estimate in the summary report. Instead, generalized variance functions are provided to enable data users to calculate standard errors for their estimates of interest. These functions estimate the standard error of an estimate as a function of the estimate. These standard error estimates are very approximate and their results should be interpreted with caution. Direct calculation of standard error estimates using the raw data, survey weights, and other sampling design information is always preferable to using generalized variance functions. However, generalized variance functions permit standard error estimation when either the raw data or the sampling design information cannot be released for confidentiality reasons.

Confidence Intervals

When constructing confidence intervals for population parameters using survey data, it is usually assumed that survey estimates of means and proportions are approximately normally distributed, so that the usual normal confidence intervals or, for small sample sizes n, t-distribution intervals with $n - 1$ degrees of freedom are appropriate (Cochran, 1977: p. 27). For complex multistage sampling, the degrees of freedom for t-intervals is typically equal to the number of sampled PSUs minus a correction, c, where $c = 1$ for unstratified cluster sampling and c equals the number of strata for stratified cluster sampling (Korn and Graubard, 1999: p. 62). The normality approximation may break down for estimates of proportions that are very close to 0 or 1. In these situations, Korn and Graubard (1999: p. 65) provide approximations to exact binomial confidence intervals that can be used with data from complex multistage sampling schemes.

Statistical Modeling with Survey Data

Much has been written on the use of the sample design and weights when fitting statistical models such as linear regression, logistic regression, loglinear models, and survival models to survey data (Fienberg, 1980; Groves, 1989: Section 6.7; Little, 1991; Pfeffermann, 1993; Lohr and Liu, 1994; Winship and Radbill, 1994; Korn and Graubard, 1995; Lemeshow et al., 1998). In general, it is recommended to include, as explanatory variables in any regression-type analysis, all variables that are used in the design or the weighting of the survey. These variables include stratum indicators, cluster indicators, and demographics (such as age, sex, ethnicity, and education) that are used to construct the weights. It is thus important in such an analysis to know how the weights were constructed, not just their numerical values. If all these variables are included, one can fit the model without survey weights (see DuMouchel and Duncan, 1983; Korn and Graubard, 1999: p. 177). It may then be necessary to put the weighting information back into the model – after having estimated the parameters – to estimate population quantities of interest. For example, Gelman and Little (1997) present an example in which estimates are obtained for each of the states in the United States, summing over demographic categories, which requires the use of census demographic breakdowns for each state. Difficulties can arise when the number of clusters or strata is large relative to the sample size or when stratum or cluster indicators are not released in public-use data for confidentiality reasons. It can also make sense to perform both weighted and unweighted analyses (DuMouchel and Duncan, 1983; Groves, 1989: p. 290; Lohr, 1999: p. 365; *see also* Sample Surveys: Model-based Approaches). Large differences between the two analyses may indicate model misspecification.

Modeling is also used for 'small-area estimation' in which inferences are made about subsets of the population (e.g., small geographic regions or demographic segments) whose sample sizes are so small that simple weighted mean estimates are too variable to be useful. Hierarchical Bayes and related methods have been useful here (Fay and Herriot, 1979; Dempster and Raghunathan, 1987), and one area of ongoing research is combining these methods with classical weighting adjustments (Gelman, 2007). The method of multilevel regression and poststratification (Lax and Phillips, 2009) combines small-area estimation (using hierarchical modeling

to get estimates for small subgroups of the population via partial pooling) and population-level information to sum up to larger groups, for example, to estimate state-level opinions from national surveys.

Missing Data

In a sense, all sampling is about missing data, since the goal is to learn about the entire population from a subset. Usually, however, missing data in sample surveys refers to unit nonresponse, which occurs when some sampled units do not participate at all in the survey, either because they refuse or cannot be reached for interview, and item nonresponse, which occurs when a sampled unit responds to some but not all questions in the survey. Survey organizations usually adjust for unit nonresponse by weighting, as discussed earlier in this article.

Item nonresponse is typically handled in one of three ways (Little and Rubin, 1987). In 'complete-case' analysis, one works only with the data from respondents who provided responses to all questions of interest. 'Available-case' analysis is the other extreme, where, for each question, one uses all the responses in the data. Both these approaches have the serious problem that item nonresponses do not generally occur completely at random, so it can be misleading to simply ignore them (Rubin, 1976). A more flexible approach to item nonresponse is 'imputation,' filling in values for missing responses.

Imputation can be done in a number of ways, and for large-scale surveys, it is often performed by the survey organization itself. Missing values may be imputed using single or multiple imputations. In single imputation (Kalton and Kasprzyk, 1986; Little and Rubin, 1987: Section 4.5), a single value is imputed for each missing value. In multiple imputations (Rubin, 1987), m plausible values are imputed for each missing value to reflect the uncertainty about the missing data (usually m is between 5 and 10). These plausible values create m completed datasets that can each be analyzed separately as if they had complete response. Simple combining rules (Rubin, 1987: p. 76) can be used to combine results from each of these m separate analyses into one overall estimate along with its standard error. If data have been imputed using single imputation, special variance formulae (e.g., Rao, 1996) are needed to account for the extra variability introduced into survey estimates due to the uncertainty in the missing values. Most large-scale surveys are currently singly imputed, but multiple imputations is becoming more popular because of the relative ease of analyzing multiply imputed datasets compared to singly imputed data. For example, the US National Health and Nutritional Examination Survey plans to release a multiply imputed dataset (Schafer et al., 1993, 1996).

Related to imputation are various data-adjustment and correction procedures. For example, survey questions such as income that have occasional very high, unreliable responses can be 'topcoded' – that is, truncated at threshold values. Difficulties with analyzing topcoded data can sometimes be avoided by working with medians or quantiles (e.g., upper fifth, second fifth, etc.), as these are less sensitive than averages to extreme values.

Summary

Large-scale social surveys are an important source of information for a wide range of topics. In analyzing such surveys, it is important to be aware of the complexity of the sampling design and the data adjustments that are used by survey organizations, including weighting to adjust for differences between sample and population and imputation to fill in missing responses. For estimating population means and proportions, the analyst should calculate weighted averages, using survey weights. In regression analysis, it is best to include, among the predictors, the variables that are used in the design and the weighting. Hierarchical models can be used to obtain inferences for small subpopulations.

In general, simple standard error estimates, which assume simple random sampling, are too small because of stratification, clustering, and weighting. These calculations can often be adjusted reasonably well by using design effects reported in the survey documentation or in previous analyses. For more involved analysis, it is advisable to use statistical packages that are specifically designed to correctly handle complex survey data.

See also: Longitudinal Data: Event–History Analysis in Discrete Time.

Bibliography

An, A.B., Watts, D.L., 1998. New SAS procedures for analysis of sample survey data. In: (SUGI) SAS User Group International Proceedings. SAS Institute, Cary, NC.

Bethlehem, J.G., 2001. Weighting adjustments for ignorable nonresponse. In: Groves, R., Eltinge, J., Little, R.J.A. (Eds.), Survey Nonresponse. Wiley, New York.

Brogan, D.J., 1998. Pitfalls of using standard statistical software packages for sample survey data. In: Armitage, P., Cotton, T. (Eds.), Encyclopedia of Biostatistics, vol. 5. Wiley, New York, pp. 4167–4174.

Carlson, B.L., 1998. Software for statistical analysis of sample survey data. In: Armitage, P., Colton, T. (Eds.), Encyclopedia of Biostatistics. Wiley, New York.

Cochran, W.G., 1977. Sampling Techniques, third ed. Wiley, New York.

Cohen, S.B., 1997. An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. The American Statistican 51, 285–292.

Davis, J.A., Smith, T.W., 1992. The NORC General Social Survey: A User's Guide. Sage, Newbury Park, CA.

Dempster, A.P., Raghunathan, T.E., 1987. Using a covariate for small area estimation: a common sense Bayesian approach. In: Platek, R., Rao, J.N.K., Sarndal, C.E., Singh, M.P. (Eds.), Small Area Statistics: An International Symposium. Wiley, New York, pp. 77–90.

DuMouchel, W.H., Duncan, G.J., 1983. Using sample survey weights in multiple regression analyses of stratified samples. Journal of the American Statistical Association 78, 535–543.

Fay, R.E., Herriot, R.A., 1979. Estimates of income for small places: an application of James–Stein procedures to census data. Journal of the American Statistical Association 74, 269–277.

Fienberg, S.E., 1980. The measurement of crime victimization: prospects for panel analysis of a panel survey. The Statistician 29, 313–350.

Fienberg, S.E., Tanur, J.M., 1983. Large-scale social surveys: perspectives, problems, and prospects. Behavioral Science 28, 135–153.

Gelman, A., 2007. Struggles with survey weighting and regression modeling (with discussion). Statistical Science 22, 153–188.

Gelman, A., Carlin, J.B., 2002. Poststratification and weighting adjustments. In: Groves, R.M., Dillman, D.A., Eltinge, J.L. (Eds.), Survey Nonresponse. Wiley, New York, pp. 289–302.

Gelman, A., Little, T.C., 1997. Poststratification into many categories using hierarchical logistic regression. Survey Methodology 23, 127–135.

Gelman, A., Little, T.C., 1998. Improving on probability weighting for household size. Public Opinion Quarterly 62, 398–404.
Groves, R.M., 1989. Survey Errors and Survey Costs. Wiley, New York.
Hidiroglou, M.A., 1986. The construction of a self-representing stratum of large units in a survey design. The American Statistician 4, 27–31.
Hill, M.S., 1992. The Panel Study of Income Dynamics: A User's Guide. Sage, Newbury Park, CA.
Iachan, R., Jones, B.I., 1987. Classification and properties of rotation sampling designs. Journal of Official Statistics 3, 169–181.
Kalton, G., Kasprzyk, D., 1986. The treatment of missing survey data. Survey Methodology 12, 1–16.
Kish, L., 1965. Survey Sampling. Wiley, New York.
Kish, L., 1998. Space/time variations and rolling samples. Journal of Official Statistics 14, 31–46.
Korn, E.L., Graubard, B.I., 1995. Examples of differing weighted and unweighted estimates from a sample survey. The American Statistician 49, 291–295.
Korn, E.L., Graubard, B.I., 1999. Analysis of Health Surveys. Wiley, New York.
Lax, J.R., Phillips, J.H., 2009. How should we estimate public opinion in the states? American Journal of Political Science 53, 107–121.
Lemeshow, S., Letenneur, L., Dartigues, J.-F., Lafont, S., Orgogozo, J.-M., Commenges, D., 1998. Illustration of analysis taking into account complex survey considerations: the association between wine consumption and dementia in the PAQUID Study. American Journal of Epidemiology 148, 298–306.
Lepkowski, J., Bowles, J., 1996. Sampling error software for personal computers. The Survey Statistician 35, 10–17.
Little, R.J.A., 1991. Inference with survey weights. Journal of Official Statistics, 405–424.
Little, R.J.A., Rubin, D.B., 1987. Statistical Analysis with Missing Data. Wiley, New York.
Lohr, S.L., 1999. Sampling: Design and Analysis. Duxbury Press, Pacific Grove, CA.
Lohr, S.L., Liu, J., 1994. A comparison of weighted and unweighted analyses in the National Crime Victimization Survey. Journal of Quantitative Criminology 10, 343–360.
Mirrlees-Black, C., Budd, T., Patridge, S., Mayhew, P., 1998. The 1998 British Crime Survey, England and Wales. Home Office, London. Home Office Statistical Bulletin 21/98.
Oh, H.L., Scheuren, F., 1983. Weighting adjustment for unit nonresponse. In: Madow, W.G., Olkin, I., Rubin, D.B. (Eds.), Incomplete Data in Sample Surveys, Volume 2, Theory and Bibliographies. Academic Press, New York.
Pfeffermann, D., 1993. The role of sampling weights when modeling survey data. International Statistical Review 61, 317–337.
Rao, J.N.K., 1996. On variance estimation with imputed survey data. Journal of the American Statistical Association 91, 499–506.
Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592.
Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.
Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A., Rubin, D.B., 1996. The NHANES III multiple imputation project. In: ASA Proceedings of the Section on Survey Research Methods. American Statistical Association, Alexandria, VA, pp. 28–37.
Schafer, J.L., Khare, M., Ezzati-Rice, T.M., 1993. Multiple imputation of missing data in NHANES III. In: Proceedings of the Bureau of the Census Annual Research Conference. US Bureau of the Census, Suitland, MD, pp. 459–510.
StataCorp, 1999. Stata Statistical Software: Release 6.0. Stata Corporation, College Station, TX.
Statistics Canada, 1997. Survey of Labour and Income Dynamics Microdata User's Guide. Catalog No. 75M0001GPE. Statistics Canada, Ottawa, ON.
Voss, D.S., Gelman, A., King, G., 1995. The polls—a review—preelection survey methodology: details from nine polling organizations, 1988 and 1992. Public Opinion Quarterly 59, 98–132.
Wang, S.T., Yo, M.L., Lin Li, Y., 1997. Consequences of analyzing complex survey data using inappropriate analysis and software computing packages. Public Health 111, 259–262.
Winship, C., Radbill, L., 1994. Sampling weights and regression analysis. Sociological Methods and Research 23, 230–257.
Wolter, K.M., 1985. Introduction to Variance Estimation. Springer-Verlag, New York.
Zaslavsky, A., 2000. Summary of Survey Analysis Software. American Statistical Association Survey Research Methods Section Webpage. http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html.

Lashley, Karl Spencer (1890–1958)

Richard F Thompson, University of Southern California, Los Angeles, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 8392–8394, © 2001, Elsevier Ltd.

Abstract

Karl Lashley was the most important and influential American Scientist in the first half of the twentieth century in the development of modern biological psychology. He was born in Davis, West Virginia on 7 June 1890 and died in Poitiers, France on 7 August 1958. He obtained his PhD at Johns Hopkins with H.S. Jennings and also worked with Adolf Meyer and John B. Watson. Lashley developed the lesion-behavioral methodology to analyze brain substrates of learning and memory, and revolutionized the field. His analysis of the problem of serial order in behavior was a precursor of modern cognitive neuroscience. His critical analysis of all extant theories of brain, learning, and behavior, including his own, was so compelling that none of these theories survived. On the other hand, he contributed an extremely important empirical literature on the role of the visual cortex in visual discriminations, on recovery of function following brain injury, on the organization of thalamocortical projection systems, and on cortical cytoarchitectonics. 'Eminent psychologist with no earned degree in psychology, famous theorist who specialized in disproving theories, including his own, inspiring teacher who described all teaching as useless'.

It is probably fair to say that Karl Lashley was the most important and influential American scientist in the first half of the twentieth century in the development of modern biological (and even experimental) psychology. His work and writings had a profound influence on such fields as brain substrates of learning and memory, the study of serial behaviors, complex cognitive processes, and our understanding of basic brain organization, particularly of the cerebral cortex.

Karl Lashley was born in Davis, West Virginia on 7 June 1890. His parents were themselves successful – his father was variously a storekeeper, postmaster, banker, and mayor of Davis and his mother, energetic and intellectual with literary and artistic interests. Lashley grew up in Davis, except for a rather extraordinary trek; the family took to the gold rush in the Klondike in 1898–99.

Lashley was an only child and acknowledged a great debt to his mother for nurturing his intellectual interests. He had learned to read well by the age of four. His mother taught in a country school until Karl was born but maintained a center for adult education in her home where women (and Karl) came for instruction in such subjects as civics, Greek poetry, and drama.

As a boy Karl read widely from the extensive family library but also displayed great interest in plant and animal life, wandering through the local fields and woods, collecting butterflies, snakes, frogs and snails, and keeping pet mice and raccoons.

Karl entered a private school at the age of four and later attended public schools in Davis, where he graduated from high school at the age of 14. He entered the University of West Virginia but did not qualify as a freshman because his high school was not accredited. He spent a preparatory year which he felt was wasted. He initially planned to study Latin but was instead assigned English as a major. To fill a vacancy in his schedule he enrolled in a course in Zoology taught by John Black Johnston, a neurologist and the only teacher of zoology at the university. This course and Johnston had a profound influence on Karl. He later wrote that 'within a few weeks in his class I knew that I had found my life's work.'

Lashley received his BA at West Virginia in 1910. He then entered the master's degree program at the University of Pittsburgh as a teaching fellow. During his year there in biology, he attended a course in experimental psychology by Karl Dallenbach, a teaching fellow in psychology. To quote Dallenbach: 'Lashley's laboratory was on the floor above mine. Though he had never taken a course in psychology, he was permitted, because he was a fellow, to elect my laboratory course in experimental psychology. The class was small and we worked intimately together upon the various experiments. Lashley was intensely interested and was the outstanding student in the class—as one might expect being a graduate student in an undergraduate class, but he was more than a run-of-the-mine graduate student. He showed in that course the promise that he later fulfilled' (Beach, 1961, p. 170).

The following summer Lashley worked at the Carnegie Laboratories at Cold Spring Harbor, where he conducted basic studies on a ciliate. A very distinguished zoologist at Johns Hopkins, H.S. Jennings, was so impressed with this work that he offered Lashley a fellowship. At Hopkins, Lashley worked with Jennings on paramecia and completed a PhD thesis on inheritance in the asexual reproduction of Hydra (1914). Lashley took a minor in psychiatry with Adolf Meyer, a pioneering figure in biological psychiatry and another minor (in psychology) with John B. Watson. The influence of Watson on Lashley was profound.

In a letter written much later to Ernest Hilgard at Stanford, Lashley described taking a seminar with Watson in 1914. Watson called their attention in the seminar to the writings of Bechterev on conditioned reflexes, which they translated. 'In the spring I served as an unpaid assistant and we constructed apparatus and did experiments, repeating a number of their experiments. Our whole program was then disrupted by the move to the lab in Meyer's Clinic. There were no adequate animal quarters there. Watson started work with infants as the next best material available. I tagged along for awhile but disliked the babies and found me a rat lab in another building. We accumulated a considerable amount of experimental material

on the conditioned reflex that was never published. Watson sought the basis of a systemic psychology and was not greatly concerned with the reaction itself' (Hilgard, personal communication, 1984).

The conditioned reflex thus came to form the basis of Watson's behaviorism. Lashley, on the other hand, had become interested in the physiology of the reaction and the attempt to trace conditioned reflex paths through the central nervous system. At the end of his career, Lashley was reputed to have said that anyone who knows American psychology today (*c*. 1958) knows that its value derives from biology and from Watson.

During the period at Hopkins, Lashley worked with Shepherd Ivory Franz at St. Elizabeth's Hospital in Washington. Together, they developed a new approach to the study of brain mechanisms of learning and memory and published landmark papers on the effect of cortical lesions on learning in rats. From this time (1916) until 1929 Lashley systematically used the lesion method in the attempt to localize memory traces in the brain. Following Watson (and Pavlov), Lashley conceived of the brain as a massive reflex switchboard, with sequential chaining of input–output circuitries via the cerebral cortex as the basis of memory. This work culminated in his classic monograph *Brain Mechanisms of Intelligence* (1929). He was also President of the American Psychological Association that year. His presidential address, and his 1929 monograph, destroyed the switchboard reflexology theory of the brain function and learning as it existed at that time.

In complex mazes, rats were impaired in learning and memory in proportion to the degree of cerebral cortex destroyed, independent of locus. Lashley employed the terms 'mass action' and 'equipotentiality' more to describe his results than as a major theory. He did not deny localization of function in the neocortex; indeed, his work on visual discriminations strongly supported localization (see Lashley, 1942). Using anatomical and physiological methods, he was perhaps the first to demonstrate the point-to-point projection from the lateral geniculate nucleus to the visual cortex. Rather, he argued that the neural substrate for higher order memorial functions, he often used the term 'intelligence,' as in complex maze learning in rats, was widely distributed in the cerebral cortex. He did attempt to develop a theory of cortical field-like interference patterns as a mechanism to account for equipotentiality in perception and learning, but he found it wanting. 'He remarked that he had destroyed all theories of behavior, including his own' (Hebb, 1959, p. 149). Lashley was perhaps the most formidable critical thinker of his time, successfully demolishing all major theories of brain-behavior extant at the time, from Pavlov to the Gestalt psychologists, to his own views.

Lashley's influence was such that the field concerned with brain substrates of learning and memory went into decline for a period of years. The studies by Lashley had critics. Walter Hunter, for example, noted that the more cortex destroyed, the more sensory areas are destroyed and the fewer sensory cues are available to the animal. Part of the problem is that maze learning is a very complex task; the notion that a simple localized engram or memory trace could serve to code the memory for a maze, seems unrealistic today. In fact, simple memories do appear to have engrams that are quite localized in the brain: conditioning of discrete responses like eyeblink in the cerebellum and conditioned fear in the amygdala. The brain structure most involved in spatial learning and memory is not the neocortex but rather the hippocampus, at least in rats. However, Lashley's view is still held today in the sense that some cognitive scientists argue that complex memories are widely distributed in the brain. The jury is still out on this matter, no distributed memories having been found.

Near the end of his career, looking over his lifetime of research on memory, Lashley (1950) concluded that: 'This series of experiments yielded a good bit of information about what and where the memory trace is not. It has discovered nothing directly of the real nature of the memory trace. I sometimes feel, in reviewing the evidence of the localization of the memory trace, that the necessary conclusion is that learning is just not possible. It is difficult to conceive of a mechanism that can satisfy the conditions set for it. Nevertheless, in spite of such evidence against it, learning sometimes does occur' (pp. 477–478).

Lashley's positive contributions were extraordinary, including an insightful analysis of 'instinctive' behaviors, analysis of sexual behaviors, characterization of the patterns of thalamocortical projections, a rethinking of cortical cytoarchitectonics, etc. Lashley's first professional position was as instructor of psychology at the University of Minnesota (1917). Robert Yerkes was to become the chair of the psychology department there and arranged Lashley's hire. However, Yerkes became much involved in the organization of the National Research Council in Washington and did not move to Minnesota. Lashley left after 1 year, returning to Baltimore where he resumed his work with John Watson. Lashley returned to Minnesota in 1920, where he was promoted from instructor to full professor in 3 years. Lashley remained at Minnesota until 1926, when he moved to the University of Chicago.

In 1936 James B. Conant, then the new president of Harvard, appointed an ad hoc committee to find the best psychologist in the world. Lashley, then at the University of Chicago, was chosen and hired in 1937. He became Director of the Yerkes Laboratory of Primate Biology in Florida but maintained his Chair at Harvard, traveling to Cambridge once each year to give his 2-week graduate seminar. The roster of eminent psychologists and neuroscientists who worked with Lashley is without parallel.

Lashley was an extraordinary teacher, albeit most of his teaching was individual. He was of the view that those who need to be taught can't learn and those who can learn don't need to be taught. Following is a description of his teaching at Minnesota by a former student: 'I saw Karl for the first time when I was taking a General Psychology course. An assembly lecture was given to all psychology classes once every two weeks, with different professors giving one lecture to about 200 students. Karl, tall, lean, sardonic held up a frog, made the muscles jump, grinned at us and talked informally in his rather precise voice. We learned more in 40 minutes from him than we had from the other eight professors and we all filed out of the hall wanting to know when we could take a course from this thin, splendid lecturer, with the pince-nez on a black string. We quickly learned that we couldn't. He hated teaching formally and was available for only a few graduate students' (Beach, 1961, p. 182).

Although Lashley is not widely recognized as one of the sources for the development of the current field of cognitive

neuroscience, his extraordinary paper on the problem of serial order in behavior clearly anticipated this field (see Bruce, 1994).

'This is the essential problem of serial order: the existence of generalized schemata of action which determine the sequence of specific acts, acts which in themselves or in their associations have no temporal valence' (Lashley, 1951, p. 122).

Lashley never wrote directly about the nature of consciousness. According to his student Roger Sperry (personal communication) he was interested in the problem of consciousness but refused to write about it, considering it to be perhaps something of an epi- phenomenon. He did, however, speculate about the nature of mind: 'Mind is a complex organization, held together by interaction processes and by time scales of memory, centered about the body image. It has no distinguishing features other than its organization ... there is no logical or empirical reason for denying the possibility that the correlation (between mental and neural activities) may eventually show a complete identity of the two organizations' (Lashley, 1958, p. 542).

Lashley received essentially all the honors available to a psychologist: President of the American Psychological Association; election to membership in the Society of Experimental Psychologists; British Psychological Association; Harvey Society; US National Academy of Sciences; Royal Society, London (Foreign Member); American Philosophical Society; American Academy of Arts and Sciences; granted the Howard Crosby Warren Medal by the Society of Experimental Psychologists for outstanding scientific contributions to psychology; the Daniel Girard Elliot Medal in Zoology by the National Academy of Sciences and the William Baly Medal in Physiology from the Royal College of Physicians. About the only honor he did not receive is the Nobel Prize, and many felt he should have done so. Sir Francis Walshe, a most distinguished neurologist, gave the following tribute: '... one of the most brilliant exponents of those problems that include psychology and physiology of the brain. He moved in this difficult territory with an experimental flair, an intellectual poise, and a competence he shared with few. His contributions to science were of a major order, and lesser men than he have been Nobel Prize men' (Beach 1961, p. 191).

Frank Beach (1961) offered a favored tribute: 'Eminent psychologist with no earned degree in psychology. Famous theorist who specialized in disproving theories, including his own. Inspiring teacher who described all teaching as useless' (Beach, 1961, p. 163).

Karl Lashley died in Poitiers, France, on 7 August 1958.

See also: Experimentation in Psychology, History of; Learning and Memory, Neural Basis of; Occipital Lobe (Visual Cortex): Functional Aspects; Psychologism, History of; Serial Order, Representation of; Topographic Maps; Watson, John Broadus (1878–1958).

Bibliography

Beach, F.A., 1961. A biographical memoir [of Karl Lashley]. Biographical Memoirs of the National Academy of Sciences 35, 162–204.

Bruce, D., 1994. Lashley and the problem of serial order. American Psychologist 49, 93–103.

Hebb, D.D., 1959. Karl Spencer Lashley, 1890–1958. The American Journal of Psychology 72, 142–150.

Lashley, K.S., 1929. Brain Mechanisms and Intelligence. University of Chicago Press, Chicago.

Lashley, K.S., 1942. The problem of cerebral organization in vision. Biographical Symposia 7, 301–322.

Lashley, K.S., 1950. In search of the engram. Society of Experimental Biology, Symposium 4, 454–482.

Lashley, K.S., 1951. The problem of serial order in behavior. Cerebral Mechanisms in Behavior. John Wiley & Sons, New York, pp. 112–136.

Lashley, K.S., 1958. Cerebral organization and behavior. Proceedings of the Association for Research in Nervous and Mental Diseases 36, 1–18.

Latent Structure and Causal Variables

Peter Spirtes, Carnegie Mellon University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 12, pp. 8395–8400, © 2001, Elsevier Ltd. with an updated Bibliography section supplied by the Editor.

Abstract

A 'latent variable' in a statistical model is a random variable that is unmeasured (although not necessarily unmeasurable). There are three main reasons for introducing latent variables into a statistical model. One reason is to include in the model features of interest that are not directly measurable, or were not measured. A second reason is that, in some circumstances, latent variables models can be used to construct estimators that are more efficient than those constructed from nonlatent variable models. A third advantage is that, in some circumstances, latent variables models can be used to construct estimators of manipulation statistics that are unbiased, unlike those constructed from nonlatent variable models. (A manipulation statistic is a statistic in a population in which some of the random variables have values externally forced upon them, e.g., as in a randomized experiment.) However, introducing latent variables into a model also introduces a host of modeling difficulties, including possibly making the model underidentified, and greatly complicating the search for the correct causal model (which in many cases is needed to construct a consistent estimator of manipulation statistics).

A 'latent variable' in a statistical model is unmeasured, although not necessarily unmeasurable. Many types of statistical models contain latent variables, including factor analytic models, item response models, some structural equation models, Rasch models, and finite mixture models. The section 'Linear Structural Equation Models' describes latent structural equation models, which are representative examples of latent variable models. (Latent structural equation models include factor analytic models as a special case. *See* Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis and Factor Analysis and Latent Variable Models in Personality Psychology.) The section 'Introducing Latent Variables' describes the purpose of introducing latent variables into a model. The section 'Construction of Latent Variable Models' describes methods of constructing latent variable models, and problems that arise with the introduction of latent variables. The section 'Conclusion' is a brief summary, and The section 'Pointers to Literature' gives pointers to the literature.

Linear Structural Equation Models

There are several different methods of parameterizing structural equation models. The one described here is essentially the same as the one used in Bentler (1985). The variables in a linear structural equation model (SEM) can be divided into two sets, the 'error variables' or 'error terms,' and the substantive variables. The error terms are latent, and some of the substantive variables may be latent as well. A structural equation model consists of a set of structural equations, one for each substantive variable, and the distributions of the error terms; together these determine the joint distribution of the substantative variables. The structural equation for a substantive variable Xi is a linear equation with Xi on the left-hand side of the equation, and the direct causes of Xi plus an error term ε_i on the right-hand side of the equation. (In some representations, some substantive variables occur only on the right-hand side of structural equations and do not have a corresponding error variable.) A SEM contains information about both the joint probability distribution over the substantive variables, and about the causal relations between the substantive variables.

Figure 1 contains an example of a latent variable SEM that is a model of mathematical test scores discussed in Spirtes et al. (2001). The original data set came from Mardia et al. (1979). The test scores for 88 students in five subjects (Mechanics, Vector Algebra, Algebra, Analysis, and Statistics) are the measured variables. The latent substantive variables are Algebra Skill, Vector Algebra Skill, and Real Analysis Skill. The distribution of the test scores is approximately normal. In Model M of Figure 1, the free parameters are the linear coefficient a, b, c, and d, and the variances and means of the error terms ε_M, ε_V, ε_{Al}, ε_{An}, ε_S, δAl, δAn, and δV. (The coefficients in the structural equations for Mechanics, Algebra, and Statistics have been fixed at 1 to ensure identifiability, which is explained below.)

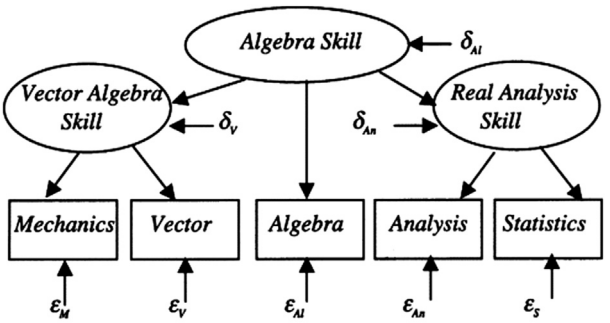

$$\textit{Mechanics} = 1 \times \textit{Vector Algebra Skill} + \varepsilon_M$$
$$\textit{Vector} = a \times \textit{Vector Algebra Skill} + \varepsilon_V$$
$$\textit{Algebra} = 1 \times \textit{Algebra Skill} + \varepsilon_{Al}$$
$$\textit{Analysis} = b \times \textit{Real Analysis Skill} + \varepsilon_{An}$$
$$\textit{Statistics} = 1 \times \textit{Real Analysis Skill} + \varepsilon_S$$
$$\textit{Vector Algebra Skill} = c \times \textit{Algebra Skill} + \delta_V$$
$$\textit{Real Analysis Skill} = d \times \textit{Algebra Skill} + \delta_{An}$$
$$\textit{Algebra Skill} = \delta_{Al}$$

Figure 1 Model M.

Structural equation models are a subset of graphical models (*see* Bayesian Graphical Models and Networks). Each SEM is associated with a graph (known as a path diagram) that represents the causal structure of the model and the form of the linear equations, where there is a directed edge from X to Y ($X \rightarrow Y$) if the coefficient of X in the structural equation for Y is nonzero (i.e., X is a direct cause of Y), and there is a bi-directed edge between the error terms ε_X and ε_Y if and only if the covariance between the error terms is nonzero. In path diagrams, latent substantive variables are enclosed in ovals, and measured variables are enclosed in rectangles.

SEMs and Probability Distributions

Let θ be a set of parameter values for SEM N that contains just the parameter values for the linear coefficients that have not been fixed, and the covariances of the error terms (i.e., the means of the error terms are not included in θ). In a SEM N, θ determines a covariance matrix over the substantive variables which will be denoted by $\sum(\theta)$ (Bollen, 1989). A covariance matrix $\sum$ is represented by a SEM N if there is a set of parameter values θ for the parameters in N such that $\sum(\theta) = \sum$. From the joint distribution over the variables, it is also possible to calculate conditional distributions, e.g., the conditional covariance matrix is just the partial covariance matrix if the error terms are normally distributed.

If there are no two distinct sets of parameter values θ and θ' for a model N such that $\sum(\theta) = \sum(\theta')$ then the model is *identified*. If a model is identified then there are a variety of methods for estimating the parameters, and a number of programs that perform the estimation (Bentler, 1985; Jöreskog and Sörbom, 1998). Also, if a model is identified, it is possible to evaluate it by testing it against a nested model using a χ^2 test with the appropriate degrees of freedom, or by assigning one of a variety of kinds of scores to the model (Bentler, 1985; Jöreskog and Sörbom, 1998).

In addition to implying a distribution over the substantive variables in the model, a SEM implies the total effect of each variable on each other variable (Bollen, 1989). A SEM is *acyclic* when its path diagram contains no directed path from a variable to itself. For acyclic SEMs, there is a simple rule for calculating the total effect of A on B. Suppose that the edge $X \rightarrow Y$ in the path diagram is labeled by the linear coefficient of Y in the structural equation for X (e.g., the edge from *Real Analysis Skill* to *Analysis* is labeled by b). Then the total effect of A on B is the sum over all directed paths from A to B, of the product of the edge labels along each path (e.g., the total effect of *Algebra Skill* on *Real Analysis* is $b \times d$).

The meaning of the total effect of A on B can be described as follows. In contrast to conditioning, a manipulated probability distribution is not a distribution in a subpopulation of an existing population, but is a distribution in a (possibly hypothetical) population formed by externally forcing a value upon a variable in the system. For example, among the subgroup of students who scored above 65 on the *Analysis* test (which was quite a high score), the mean *Statistics* score was above the mean for the whole class. On the other hand it is possible to manipulate the scores of all of the students by teaching them the answers to the questions on the *Analysis* test before they take it. In that case the mean *Statistics* score would remain the same for the whole class. So the mean *Statistics* score conditional on the *Analysis* score being above 65 would not equal the mean *Statistics* score, given that the score of the *Analysis* test was manipulated to be above 65. The total effect of A on B is just the change in B resulting from a manipulation that forces a unit change in A.

Of course, how an actual manipulation would affect other variables would depend upon how the system is manipulated. If the *Analysis* test score were changed by practice and studying to improve *Real Analysis Skill*, then one might expect that the *Statistics* test score might change as well. In that case, the manipulation has changed *Statistics*, but not through the mechanism of changing the *Analysis* score. The total effect of *Analysis* on *Statistics* is the resulting change in *Statistics* due to a kind of ideal manipulation of *Analysis*, in which the only variable in the system directly affected by the manipulation of *Analysis* is *Analysis* itself. Strotz and Wold (1960) suggested that an ideal manipulation of a variable X to a fixed value x could be represented by replacing the structural equation for X with a new equation $X = x$. The manipulated distribution is then just the distribution for the new structural equation model. Figure 2 shows the model that results from the manipulation of *Analysis*. The path diagram of the manipulated population is formed by breaking all of the edges into the manipulated variable.

Unlike a conditional distribution which is completely determined by the population joint distribution, a manipulated distribution depends upon both the population joint distribution and the true path diagram (or structural equations). In particular, two models with different path diagrams can represent the same set of joint probability distributions, but differ in their predictions of the total effect of manipulating a variable. Figure 3 shows an alternative to Model M of Figure 1, where the edge between *Real Analysis Skill* and *Algebra Skill* has been reversed. In Model M, the total effect of *Algebra Skill* on *Real Analysis* is nonzero, while in Model M′, the total effect of *Algebra Skill* on *Real Analysis* is zero. However, both M

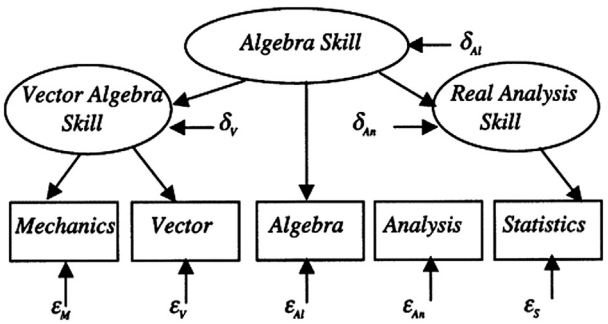

$$\textit{Mechanics} = 1 \times \textit{Vector Algebra Skill} + \varepsilon_M$$
$$\textit{Vector} = a \times \textit{Vector Algebra Skill} + \varepsilon_V$$
$$\textit{Algebra} = 1 \times \textit{Algebra Skill} + \varepsilon_{Al}$$
$$\textit{Analysis} = 100$$
$$\textit{Statistics} = 1 \times \textit{Real Analysis Skill} + \varepsilon_S$$
$$\textit{Vector Algebra Skill} = c \times \textit{Algebra Skill} + \delta_V$$
$$\textit{Real Analysis Skill} = d \times \textit{Algebra Skill} + \delta_{An}$$
$$\textit{Algebra Skill} = \delta_{Al}$$

Figure 2 Model M after *Analysis* is manipulated to 100.

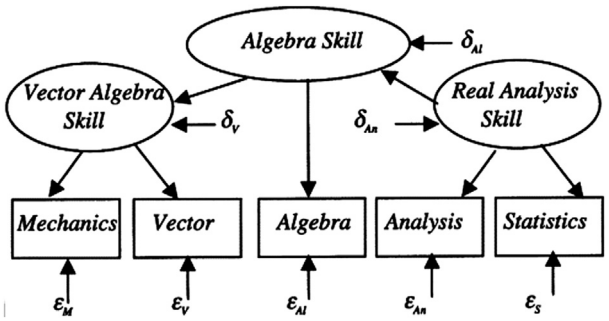

Figure 3 Model M.

and M′ represent the same set of distributions over the substantive variables. If Model M′ were considered to be compatible with background knowledge, the fact that M′ passes a statistical test or is assigned a high score is not sufficient evidence that M′ can be used for calculating total effects; it is also important to know whether there are any other plausible models compatible with background knowledge (such as M) that would also pass a statistical test or receive a high score, and predict a different total effect.

Introducing Latent Variables

There are three main reasons for introducing latent variables into a model. The first is that there may be variables of interest that either were not measured, or cannot be measured without error. The second is that a latent variable model may be necessary in order to represent correctly both the probability distribution over the measured variables, and the total effects that the measured variables have on each other. The third is that a latent variable model may be a more parsimonious representation of the distribution over the observed variables than any model without latent variables.

Measurement Error

Whenever there are measurement errors, the actual measurements are observed, but the quantity of interest that is being measured is a latent variable. For example, in Figure 1, *Algebra Skill*, *Real Analysis Skill*, and *Vector Algebra Skill* are measured indirectly by test scores. The test scores are not perfect measures of the corresponding latent quantities, but are also dependent on the error terms, i.e., there is measurement error in this model. Hence *Algebra Skill*, *Real Analysis Skill*, and *Vector Algebra Skill* are latent variables. If the goal is to estimate the total effect of *Algebra Skill* on *Statistics*, and the variance of ε_{Al} is not zero, then regressing *Statistics* on the measured variable *Algebra*, instead of on the latent variable *Algebra Skill*, gives a biased estimate of the total effect. However, under the assumption that the latent variable Model M is correct, estimations of the parameters of M can be used to calculate an unbiased estimate of the effect of *Algebra Skill* on *Statistics*.

Representing Causal Relationships Between Variables

The second main reason for introducing latent variables into a model is to represent correctly the total effects of the variables

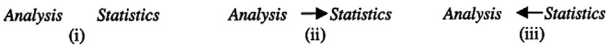

Figure 4 SEMs without latent variables.

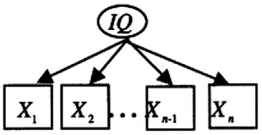

Figure 5 A parsimonious latent variable model.

upon each other. Suppose that Model M is the correct representation of the causal relations between the variables in the path diagram. In that case, the total effect of *Analysis* on *Statistics* is zero, the total effect of *Statistics* on *Analysis* is zero, but *Analysis* and *Statistics* are correlated. The possible SEMs (with no correlated errors) when there are no latent variable and only two measured variables *Analysis* and *Statistics*, are shown in Figure 4. None of these nonlatent variable models correctly represents both the correlation between *Analysis* and *Statistics*, and the zero total effects of each variable on the other. So in this case, a latent variable model is needed for the correct representation.

A bi-directed edge between the ε_{Al} and ε_S also correctly models the probability distribution and the total effects and does not explicitly introduce a latent variable that is a direct cause of both *Analysis* and *Statistics*; however, it does introduce such a latent variable implicitly. Assuming that a correlation between ε_{Al} and ε_S is produced either by ε_{Al} causing ε_S, ε_S causing ε_{Al}, or some third latent variable causing both of them, a bi-directed edge between ε_{Al} and ε_S also implies that there is some latent variable (possibly ε_{Al} or ε_S) that is a direct cause of both *Analysis* and *Statistics*.

Parsimonious Models

In some cases a latent variable model provides a more parsimonious (i.e., fewer parameters and higher degrees of freedom) representation of a distribution than any nonlatent variable model that represents the same distribution. Consider the model in Figure 5, where one latent variable (e.g., *IQ*) causes many measured indicators (e.g., various intellectual tasks.) It follows from results in Spirtes et al. (2001) that the latent variable model is more parsimonious than any other SEM nonlatent variable model that represents the same distribution (assuming all of the parameters of the nonlatent variable model are not constrained). Other things being equal, from a small sample size, the more parsimonious latent variable model could be used to provide a prediction of xn (e.g., grades in first year in college) from $x1, \ldots, xn - 1$ (e.g., scores on many different kinds of tests) that has a smaller mean squared error than any nonlatent variable model.

Construction of Latent Variable Models

Construction of SEMs often proceeds through a combination of the application of background knowledge with either

nonautomated or automated searches through the space of alternatives. The searches look for a model or a set of models with high posterior probabilities (*see* Bayesian Statistics), or high scores, or that pass a statistical test. Even without latent variables, the problem is a difficult one because the space of models to be searched grows exponentially with the number of variables. The search space may be constrained by background knowledge, but it is still typically enormously large.

Allowing latent variables into models makes the problem of constructing good models even more difficult. This is because estimating, testing, scoring, and searching for latent variable SEMs is fundamentally more difficult than the corresponding operations for acyclic directed graph SEMs (i.e., SEMs without latent variables, cycles, or correlated errors). The space of possible acyclic directed graph SEMs is finite, but the space of latent variable models is infinite. An acyclic directed graph SEM is always identified. A SEM with latent variables may not be identified, and the only known algorithm for determining if it is identified has a running time that grows exponentially with the number of variables. It is always easy to calculate the number of degrees of freedom for an acyclic directed graph SEM, but it is often difficult to calculate the number of degrees of freedom for a SEM with latent variables (Geiger and Meek, 1999).

While the estimation, testing, and scoring problems with latent variable models have been solved for special cases, and there are heuristic rules for other cases (Bollen, 1989), practical general solutions are not available.

Conclusion

There are three main reasons for introducing latent variables into models: using latent variables can lead to a more parsimonious model that improves prediction; when there is measurement error the natural representation of the quantities of interest is as latent variables; in some cases the only correct representation of the joint distribution of the observed variables and of the total effects of variables on each other is with models that contain latent variables (or bi-directed edges which imply the existence of latent variables).

While latent variable models have some advantages over nonlatent variable models, construction of a latent variable model is often more difficult than construction of nonlatent variable models. It is generally more difficult to estimate, test, score, and search for latent variable models than for nonlatent variable models.

Pointers to Literature

Sewell Wright (1921) introduced a form of structural equation model called path models. Bollen (1989) provides an overview and a brief history of linear structural equation models with latent variables, and Bartholomew and Knott (1999) describe a wide variety of latent variable models. Rubin (1977) discusses manipulations, which are also discussed in Spirtes et al. (2001), Pearl (1995), and Pearl (2000). Heckerman et al. (1999) and Spirtes et al. (2001) discuss the causal interpretation of latent variable models and describe a variety of automated search techniques. Acyclic directed graph structural equation models are special cases of Bayesian networks and graphical models.

Bibliography

Bartholmew, D., Knott, M., 1999. Latent Variable Models and Factor Analysis. Charles Griffin, London, UK.

Bentler, P., 1985. Theory and Implementation of EQS: A Structural Equations Program. BMDP Statistical Software, Los Angeles, CA.

Bollen, K.A., 1989. Structural Equations with Latent Variables. John Wiley, New York.

Geiger, D., Meek, C., 1999. Quantifier elimination for statistical problems. In: Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers, San Francisco, CA, pp. 226–235.

Gelman, Andrew, Carlin, John B., Stern, Hal S., Dunson, David B., Vehtari, Aki, Rubin, Donald B., 2014. Bayesian Data Analysis. CRC press, Boca Raton, FL.

Hagenaars, Jacques A., McCutcheon, Allan L., 2002. Applied Latent Class Analysis. Cambridge University Press, Cambridge, UK.

Heckerman, D., Meek, C., Cooper, G., 1999. A Bayesian approach to causal discovery. In: Glymour, C., Cooper, G. (Eds.), Computation, Causation and Discovery. MIT Press, Cambridge, MA.

Jöreskog, K., Sörbom, D., 1998. LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language. Scientific Software, Mooresville, IN.

Loehlin, John C., 2004. Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis. L. Erlbaum Associates, Mahwah, NJ.

Mardia, K., Kent, J., Bibby, J., 1979. Multivariate Analysis. Academic Press, New York.

Muthén, Bengt O., 2002. Beyond sem: general latent variable modeling. Behaviormetrika 29 (1), 81–118.

Pearl, J., 1995. Causal diagrams for empirical research. Biometrika 82, 669–709.

Pearl, J., 2000. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, UK.

Rubin, D., 1977. Assignment to treatment group on the basis of a covariate. Journal of Educational Statistics 2, 1–26.

Skrondal, Anders, Rabe-Hesketh, Sophia, 2004. Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. CRC Press, Boca Raton, FL.

Spirtes, P., Glymour, C., Scheines, R., 2001. Causation, Prediction, and Search, second ed. MIT Press, Cambridge, MA.

Strotz, R., Wold, H., 1960. Recursive versus nonrecursive systems: an attempt at synthesis. Econometrica 28, 417–427.

Whittaker, J., 1990. Graphical Models in Applied Multivariate Statistics. Wiley, New York.

Wright, S., 1921. Correlation and causation. Journal of Agricultural Research 20, 557–585.

Lateralization of Language as Demonstrated by Brain Imaging Procedures

Brigitte Stemmer, Université de Montréal, Montreal, QC, Canada; and Brock University, St. Catherines, ON, Canada

Published by Elsevier Ltd.

Abstract

Are statements such as 'Language is in the left hemisphere,' 'The first language is in the left and the second in the right hemisphere,' 'The brain of children recovers language perfectly after brain damage,' 'The adult brain cannot recover language functions and chronic aphasia does not improve over time,' 'Language functions are more lateralized to the left in men and more bilaterally in women' truths or myths? This article will discuss these and other issues related to the contribution the left and right hemispheres make to certain aspects of language. As will be shown, there is not always a straightforward answer to seemingly simple questions.

Introduction

Hemispheric asymmetry or laterality can be found across processing domains and contemporary species. Anatomical asymmetry refers to asymmetries concerning the brain structure and functional asymmetry (often referred to as lateralization) relates to domains such as perception, attention, motor control, memory, emotion, and language. It must be understood that hemispheric asymmetry is not an all or nothing phenomenon but rather operates on a continuum. Even if one hemisphere is dominant in performing a task or processing information, the other hemisphere usually has at least some competence in these domains (Hellige, 2008).

The notion that the left cerebral hemisphere deals with language is usually associated with the nineteenth-century neurologists Paul Broca and Carl Wernicke. Based on the observation that some patients with lesions on the left side of their brain showed language problems, speech production was associated with the *left* frontal region of the brain (left posterior inferior frontal gyrus, Broca's region) and speech perception with the *left* temporal region of the brain (left posterior superior temporal gyrus, Wernicke's region) (see Figure 1). The consequences of these landmark postulates are still present today and are reflected, for example, in the well-known language impairments called Broca's aphasia or Wernicke's aphasia.

Since the early lesion studies with brain damaged patients, the advent of functional neuroimaging techniques (*see* Functional Brain Imaging of Language processes) made it possible to identify the hemisphere and the regions therein that are active while a person is performing a task. As will be seen below, not all neuroimaging studies support the findings of lesion or behavioral studies. One reason is that while some of these neuroimaging techniques can provide relatively accurate information on where in the brain something happens, they do not tell us what the function or role of the activated hemisphere or brain region is, or how or whether these regions interact with other brain regions. Assigning functions or roles to an activated brain region is ultimately an interpretation by the researcher usually based on clever experimental designs and integration with findings from previous research. As pointed out by Petersen and Fiez (Petersen and Fiez, 1993: 513), it should not be forgotten that "A functional area of the brain is not a task area; there is no 'tennis forehand area' to be discovered. Likewise, no area of the brain is devoted to a very complex function; 'attention' or 'language' is not localized in a particular Brodmann area or lobe. Any task or 'function' utilizes a complex and distributed set of brain areas." Another issue the reader should be aware of is that the majority of neuroimaging studies has investigated language at the word level and, in comparison, there are relatively fewer studies investigating the sentence level and even fewer studies at the text or discourse level. In addition, although the way language is represented in the brain and the way it is processed are two different issues, this distinction is not always reflected in neuroimaging studies.

Considering these limitation, instead of talking about the lateralization of language in general, it would thus be more accurate to talk about the contribution the hemispheres make to *specific aspects* of language representation and processing. (For an overview of the language network in the left hemisphere and the anatomical regions referred to in the text see Figure 1.)

Lateralization at the Word Level

Imagine that you read a string of symbols such as *ptakopgx* (a non word), *siglowark* (a pseudoword), or *house* (a word). With astonishing speed, about 200–250 ms after seeing these symbols or hearing the corresponding sounds, your brain has determined whether this is a nonword, a word that does not make sense such as a pseudoword or a known word. Approximately 400 ms after reading or hearing a word, your brain has assigned some meaning to it. Which regions in the brain are implicated in these fast processes and is one hemisphere more implicated than the other? The classical language model advanced by Broca and Wernicke assigned language processing (words, sentences) to the left hemisphere and was based on the observation of a few patients and the study of their postmortem brains. More than 100 years after Broca's and Wernicke's famous insights, the first language study was published using positron emission tomography (PET) that allowed 'watching' which brain areas were active during auditory and visual word processing in healthy subjects (Petersen et al., 1988; Petersen and Fiez, 1993). These early and also later neuroimaging studies only partially confirmed the classical model. Summarizing the results of the numerous

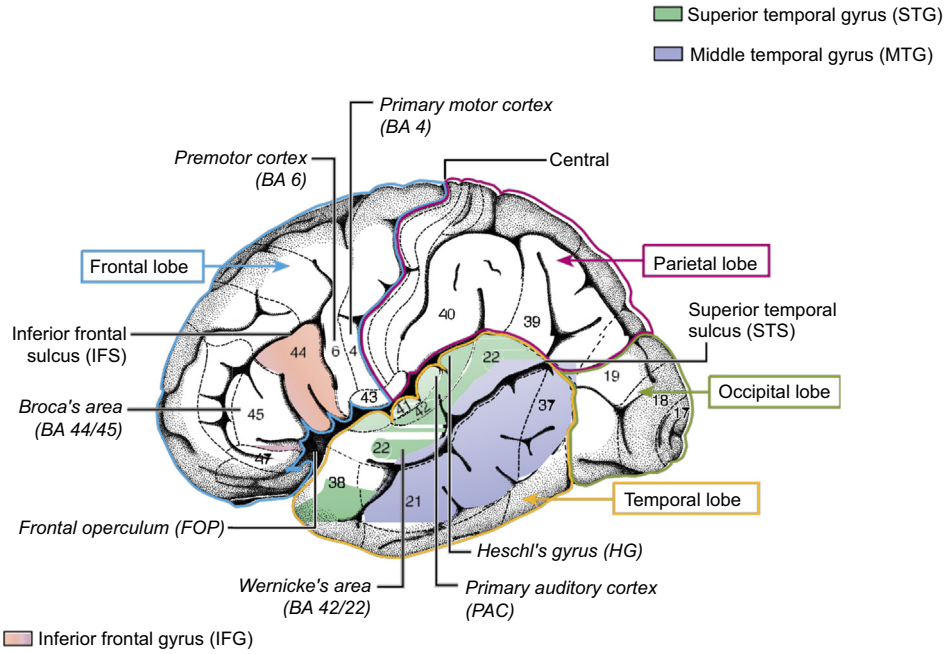

Figure 1 Anatomical and cytoarchitectonic details of the left hemisphere. The different lobes (frontal, temporal, parietal, and occipital) are marked by colored borders. Major language-relevant gyri (IFG, STG, MTG) are color coded. Numbers indicate language-relevant Brodmann Areas (BA), which Brodmann (1909) defined on the basis of cytoarchitectonic characteristics. The coordinate labels superior/inferior indicate the position of the gyrus within a lobe (e.g., superior temporal gyrus) or within a BA (e.g., superior BA 44; the superior/inferior dimension is also labeled dorsal/ventral). The coordinate labels anterior/posterior indicate the position within a gyrus (e.g., anterior superior temporal gyrus; the anterior/posterior dimension is also labeled rostral/caudal). Broca's area consists of the pars opercularis (BA 44) and the pars triangularis (BA 45). Located anterior to Broca's area is the pars orbitalis (BA 47). The frontal operculum (FOP) is located ventrally and more medially to BA 44, BA 45. The premotor cortex is located in BA 6. Wernicke's area is defined as BA 42 and BA 22. The primary auditory cortex (PAC) and Heschl's gyrus (HG) are located in a lateral to medial orientation. Friederici, A.D., 2011. The Brain Basis of Language Processing: From Structure to Function. Physiological Reviews 91, 1357–1392.

neuroimaging studies that have since then been published is difficult due to different experimental paradigms and stimuli. Vigneau and colleagues (Vigneau et al., 2006, 2011) took on this challenging task and statistically contrasted and combined results from neuroimaging studies investigating mostly single word processing but also some sentence processing in healthy subjects. These meta-analyses showed that not only the left hemisphere but also the right hemisphere activated in single word processing albeit to a much lesser extent. Right-hemisphere activation was found in about one-third of the studies and this activation was most commonly accompanied by left-hemisphere activation. This has been interpreted as supporting the view that the right hemisphere involvement is dominated by interactions between the two hemispheres. In more than two-thirds of the studies, left-hemisphere activation was reported mostly without concurrent activation of the right hemisphere thus suggesting that left hemisphere involvement occurs mainly within the left hemisphere without much interaction with the right hemisphere (Vigneau et al., 2011). It was further noted that right-hemisphere participation in frontal brain regions was not always specific for a given language component suggesting that other cognitive processes such as working memory were also recruited by the verbal tasks.

After this global overview, we will now look in more detail at the contribution of the left and/or right hemisphere in response to specific processing tasks or demands.

Perceptual, Phonological, and Motor processing

When we perceptually process visual material (e.g., passively reading words), auditory material (e.g., listening to words), or when we acoustically analyze incoming speech sounds, regions in both hemispheres are recruited (specifically extrastriate occipital cortices, primary auditory cortices, and temporal brain regions including and surrounding the primary auditory cortex, respectively) (Petersen and Fiez, 1993; Price, 2012). Both hemispheres (especially motor regions and other regions surrounding Broca's area) are also activated when the task involves motor output of language and articulatory coding such as repeating words aloud. In short, processing related to word perception, acoustic analysis of speech sounds and articulation recruits both hemispheres. This, however, does not mean that the regions activated in both hemispheres are specific to speech; it has been shown that they also activate in response to other nonspeech material.

Regions of the left hemisphere alone (temporoparietal and cingulate regions) seem to be implicated in phonological coding processes (Petersen and Fiez, 1993). Other functions that have been associated with left lateralization only are acoustic complexity of speech (anterior part of superior temporal cortex), disambiguation of speech sounds (frontal and premotor areas), and the functional integration of acoustic processing (Price, 2012).

Lexicosemantic Processing

There is more to understanding or producing language than perceptual or phonological processes. We must also assign meaning (semantics) to words (lexis) and the question thus is what the contribution of each hemisphere is to lexicosemantic processing. Generally, most imaging studies support the view that lexical semantic knowledge is widely distributed in the brain with the left hemisphere being more involved than the right hemisphere (Binder et al., 2009; Vigneau et al., 2006, 2011; **Figure 2**). Activation of the left hemisphere alone was reported in most tasks that implicated reading words or listening to words for semantic content. Other tasks targeting semantic word processing such as word retrieval, word generation (*Name everything you can buy in a supermarket*), or word categorization also activated the left hemisphere, and especially left frontal brain regions (pars opercularis and orbitalis; dorsomedial prefrontal cortex) (Binder et al., 2009; Vigneau et al., 2006). Comparing words with different semantic content (e.g., reading words versus pseudowords; categorization of written words) showed activations in left temporal brain regions (posterior temporal gyrus). It was suggested that these activations are related to letter and grapheme processing in conjunction with the phonological network operational during reading (Vigneau et al., 2006). When associating meaning with words, one may also ask whether the *type* of meaning encoded in the words influences hemisphere activation. Many (but not all) non-imaging studies have suggested that abstract words (such as *creativity*) are processed in the right hemisphere and concrete words (such as *bottle*) in the left hemisphere. Although this has been supported by studies using the transcranial magnetic stimulation technique (TMS) (Hartwigsen and Siebner, 2011), the issue surrounding abstract and concrete concepts may also entail perception, action, and emotional systems which, in turn, would influence hemisphere activation (for a discussion see Kiefer and Pulvermüller, 2012). As it stands now, the issue needs further clarification.

In addition to frontal and anterior temporal regions, posterior temporal, and parietal regions have also been associated with semantic processing. The left fusiform gyrus, a region in the left temporooccipital brain area and best known for its implication in face recognition, has been related to the semantic

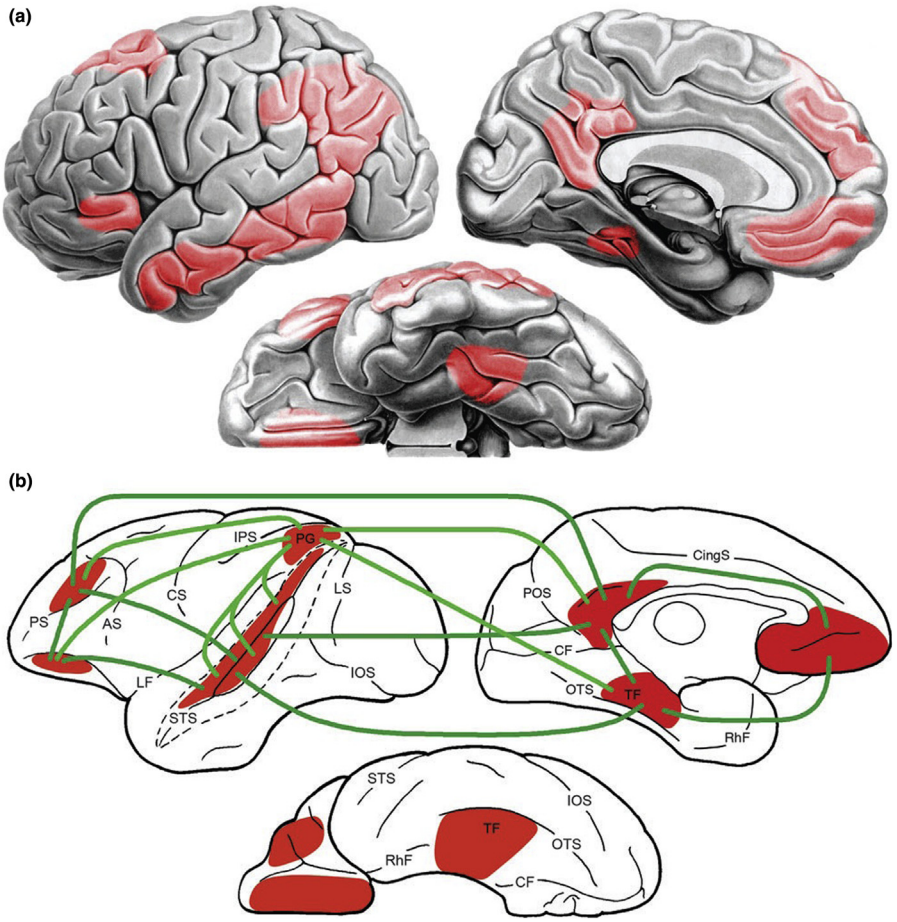

Figure 2 The diagrams show the large-scale semantic network of the human brain (a) and the probable homologous network in the macaque monkey brain (b). The green lines show the connections between different brain regions in the macaque monkey brain. Compared to the human semantic system, the cortical networks in the macaque monkey brain are strikingly similar in location and function (for details see Binder et al., 2009). Figure taken from Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where is the semantic system? A critical review andmeta-analysis of 120 functional neuroimaging studies. Cerebral Cortex 19, 2767–2796.

processing of words and pictures (Vigneau et al., 2006). The parietal cortex of both hemispheres, and specifically the angular gyrus as one of the most consistently activated regions in the left hemisphere, has been viewed as a high-level integration area that occupies a top position in the processing hierarchy underlying concept retrieval and conceptual integration in both hemispheres (Geschwind, 1965; Binder et al., 2009; Vigneau et al., 2006). Some structures in both hemispheres, and especially the posterior cingulate gyrus and the hippocampal system, seem to act as an interface between semantic retrieval systems and brain systems that encode episodic information (Binder et al., 2009). Although activation of both hemispheres has been observed, left hemisphere activation is nevertheless frequently dominant.

Taken together, processing related to word perception, acoustic analysis of speech sounds and articulation relies on both hemispheres. The processing of phonological codes, acoustically complex speech and its functional integration, and the disambiguation of speech sounds seem to rely solely on the left hemisphere. While right-hemisphere activation can be observed with lexicosemantic processing, it is nevertheless the left hemisphere that seems dominant. The degree to which the two hemispheres are involved depends on the type of processing, the material that is being processed and the implication of other cognitive systems (such as attention and memory). It should be mentioned that although most studies related brain activation to semantic word processing, it has nevertheless been difficult to tease apart what is purely semantic processing and what is due to increased task demands and thus increased involvement of attention and memory networks, or phonological processing demands (Binder et al., 2009).

Lateralization at the Sentence Level

Language usually occurs as a stream of speech forming utterances which, in turn, are made up of words that may string together to form sentences. What is the contribution of each hemisphere in sentence or syntactic processing? This is not an easy question to answer. For one thing, most studies use sentences that compare meaningful complex sentences (e.g., *Sophia, who had a broken leg, was given a lift home by Peter*) with syntactically simpler sentences (*He gave her a lift home*) or with nonmeaningful sentences (*Mo soodled lorn en saweek*). This makes the distinction between syntactic and semantic processing components difficult (Cappa, 2012; Vigneau et al., 2006). In addition, effort and processing load may also play a role and involve other systems such as attention or memory. Other issues concern methodological decisions (e.g., the resolution power of the scans; the type of analysis; the focus on the processing or the linguistic material involved). Considering these limitations, it is not surprising that meta-analyses report activation of largely overlapping brain regions during syntactic and semantic processing (Vigneau et al., 2006, 2011). About half of the studies included in a meta-analysis investigating sentence and text processing tasks reported bilateral activation and the other half left-hemisphere activation only (Vigneau et al., 2011). Bilateral activation was seen in response to mental imagery with auditory sentences, plausibility judgment on the semantic content of a sentence, and comparison of complex to simple sentences (Vigneau et al., 2011). When the left hemisphere alone was activated, two regions in the frontal lobe seemed dedicated to sentence processing. These regions were identified in the middle frontal region (left posterior middle frontal gyrus) and in Broca's area (inferior frontal region, specifically the dorsal part of the pars opercularis). As the middle frontal gyrus is also activated during working memory processes, it was suggested that its recruitment in language tasks was related to working memory implicated in the comprehension of complex linguistic material rather than solely to sentence processing (Vigneau et al., 2006). The left inferior frontal gyrus, which houses Broca's area, is the other region that seemed dedicated to syntactic processing. Aside from the frontal lobe, the temporal lobes of the left hemisphere also activate during sentence processing. In fact, Vigneau et al. (2006) found that the left temporal cortex was more frequently activated in sentence processing (and overlapping with semantic regions) than the left frontal cortex.

Processing spoken sentences involves not only processing syntactic and semantic information but also acoustic features that are used to express linguistic prosody such as intensity, vocal pitch, rhythm, or rate of utterance. Linguistic prosody is mainly expressed by intonation, that is rising and falling variations in pitch sequences within speech, and is crucial for the interpretation of sentences (for example, distinguishing questions from nonquestions in English or extracting lexical meaning in tonal languages like Chinese). Most lesion studies have assigned a special role to the right hemisphere for processing suprasegmental prosody. Neuroimaging studies support this view and specify that the right hemisphere is involved in the processing of pitch information during sentence processing (Gandour et al., 2000, 2004; Plante et al., 2002; Minagawa-Kawai et al., 2011). If, however, pitch is used to distinguish lexical meaning (as in tonal languages), then a left-hemisphere processing bias emerges (Minagawa-Kawai et al., 2011).

To sum up, there is a large overlap of brain regions that activate during syntactic as well as semantic processing. This is not surprising considering the difficulties to clearly distinguish syntactic and semantic processing components. Bilateral activation is frequently observed in response to tasks that involve metacognitive judgments on sentence characteristics or complexity issues such as comparing complex and simple sentences. Some brain regions in the frontal lobe (including Broca's area) seem dedicated to syntactic processing although it seems that temporal regions in the left hemisphere are more frequently activated than frontal regions. The right hemisphere is dominant in processing linguistic prosody in nontonal languages.

Lateralization at the Pragmatic Level

Until now we have discussed lateralization of language at the word and sentence level without paying much attention to the context in which language is uttered. When we communicate in real life, however, we use language as action performed with words. The meaning of words (and larger linguistic units) is not viewed as an isolated entity but as an entity that gets

meaning by interpreting it in the context in which it occurs. An utterance such as *Same procedure as last year* may be humorously or ironically uttered to refer to the repetitive gathering for New Year's dinner or refer to the annual changing of my winter tires. What the speaker conveys with this utterance depends on, for example, the context, the situational circumstances and the listener's interpretation. The way we use these linguistic (as well as nonlinguistic) signs in interaction for the purpose of communication has been referred to as pragmatics, and the underlying brain correlates and processes that bring about pragmatic behavior have been coined neuropragmatics. Based on findings from patients whose right hemisphere is damaged and who subsequently show problems with different aspects of pragmatics, it has traditionally been assumed that pragmatics is the realm of the right hemisphere. With the advent of brain imaging techniques, this assumption has been called into question. Two aspects of pragmatics have received particular attention, that is the comprehension of text and discourse and of figurative language.

Text and Discourse Comprehension

One aspect of text comprehension that has been investigated with brain imaging are the neural substrates underlying coherence building, that is the processes that link the elements of a text in a meaningful way. For example, the following two utterances form a coherent text: (1) *The court convicted him to lifelong imprisonment;* (2) *The victim was brutally tortured and killed*. We link these utterances in a meaningful way by making inferences from our knowledge of the world (murder is a crime that is punished). Brain imaging studies investigating various types of coherence building report different findings. Some describe right hemisphere, some left hemisphere, and some bilateral activation. These discrepancies are possibly due to the different types of coherence and inferences investigated and differences in experimental paradigms (Ferstl, 2010; Stemmer, 2013).

Figurative Language

An area of research that has been the focus of numerous brain imaging studies is the way we communicate the 'unsaid' as reflected in metaphors or other figurative language (irony, sarcasm, idioms, proverbs). Traditionally, it was assumed that the processing mechanisms underlying figurative (*He lives on the moon*) and nonfigurative language (*He lives in Petersburg*) are different and that this difference is reflected in different hemisphere activation: nonfigurative language is processed by the left hemisphere and figurative language by the right hemisphere. While numerous lesion studies supported these hypotheses, no clear-cut picture has emerged from neuroimaging studies. Some studies report left hemisphere activation only, while others report either right or bilateral activation (for summaries see Bohrn et al., 2012; Schmidt and Seger, 2009; Ferstl, 2010). Heterogeneity of the experimental paradigms, individual variability and analysis methods are likely reasons for these discrepant findings. Despite these heterogeneities, researchers have attempted to find convergence. In a qualitative review, it was suggested that for metaphor processing the left hemisphere is activated when frozen, conventional or familiar metaphors are used and the right hemisphere activates when novel metaphors are used possibly due to nonfamiliarity and processing difficulty (Schmidt and Seger, 2009). In a quantitative review, bilateral activation with a bias to the left hemisphere was reported for figurative language when compared to literal language (Bohrn et al., 2012). This is surprising as numerous individual studies do report right hemisphere activation. It is possible that the meta-analysis did not produce a statistically significant activation cluster in the right hemisphere because many different regions activated in the right hemisphere. It was further reported that novel metaphors activated a bilateral network and this was interpreted as indicating processes of set-shifting and hypothesis generation. Familiar idioms or proverbs recruited the right hemisphere while the processing of irony and sarcasm activated prefrontal regions in both hemispheres and frontotemporal regions in the right hemisphere only (Bohrn et al., 2012). As other studies have shown that these regions are frequently activated in theory of mind and mentalizing tasks (that is recognizing the intentions and desires of others), it was suggested that irony and sarcasm rely more on theory of mind and mentalizing functions compared to the other types of figurative language.

Generally, it is difficult to reconcile the numerous brain imaging studies. What seems to emerge is that figurative and nonfigurative language share similar language networks in the left hemisphere – possibly including the default language network that is active whenever linguistic material is processed. In addition, depending on the type of stimuli, the task demands and characteristics of the listener, figurative language will recruit networks in the right hemisphere possibly reflecting other nonlinguistic processing mechanisms (such as attention, memory, organization and integration of information, and so forth).

Brain Connectivity within the Left Hemisphere

As detailed in the previous sections, the left hemisphere has a dominant role in the functional neural networks for processing different aspects of language, especially at the word and sentence level. Within the left hemisphere, the language network includes frontal regions (Broca's area in the inferior frontal gyrus) and temporal regions (Wernicke's area in the superior temporal gyrus) (see **Figure 3**). For these regions to be functional, they need to correspond with other brain regions via so-called fiber tracts or pathways (white matter). There is general consensus that major pathways connecting important language areas run above (dorsally) or below (ventrally) the Sylvian fissure, which separates the frontal from the temporal lobe (Catani et al., 2005; Catani and Mesulam, 2008; Friederici, 2009, 2012; Saur et al., 2008, 2010; Weiller et al., 2009). Some of these pathways show hemispheric asymmetries already in early infancy (Geng et al., 2012). For example, it has been shown that the arcuate fasciculus connecting temporal with frontal brain regions is already asymmetric in the first 2 years of life. It is still a matter of debate where exactly these fiber tracts start and end and how they branch out. While most researchers agree that there is a ventral stream for sensory percept and a dorsal stream for meaning, the details of their functional significance is still

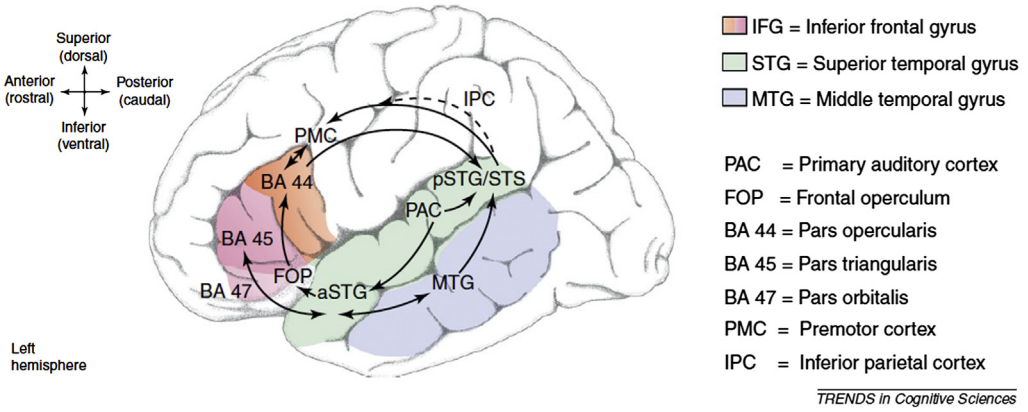

Figure 3 The cortical language circuit (schematic view of the left hemisphere). The major gyri involved in language processing are color coded. In the frontal cortex, four language-related regions are labeled: three cytoarchitectonically defined Brodmann [39] areas (BA 47, 45, 44), the premotor cortex (PMC) and the ventrally located frontal operculum (FOP). In the temporal and parietal cortex the following regions are labeled: the primary auditory cortex (PAC), the anterior (a) and posterior (p) portions of the superior temporal gyrus (STG) and sulcus (STS), the middle temporal gyrus (MTG) and the inferior parietal cortex (IPC). The solid black lines schematically indicate the direct pathways between these regions. The broken black line indicates an indirect connection between the pSTG/STS and the PMC mediated by the IPC. The arrows indicate the assumed major direction of the information flow between these regions. During auditory sentence comprehension, information flow starts from PAC and proceeds from there to the anterior STG and via ventral connections to the frontal cortex. Back-projections from BA 45 to anterior STG and MTG via ventral connections are assumed to support top-down processes in the semantic domain, and the dorsal back-projection from BA 44 to posterior STG/STS to subserve top-down processes relevant for the assignment of grammatical relations. The dorsal pathway from PAC via pSTG/STS to the PMC is assumed to support auditory-to-motor mapping. Furthermore, within the temporal cortex, anterior and posterior regions are connected via the inferior and middle longitudinal fasciculi, branches of which may allow information flow from and to the mid-MTG. Friederici, A.D., 2012. The cortical language circuit: from auditory perception to sentence comprehension. Trends in Cognitive Sciences 16, 262–268.

a matter of debate. Ventral pathways connect the left temporal cortex with the left prefrontal cortex via the extreme fiber capsule system and the uncinate fasciculus and have been associated with mapping sound to meaning, the acquisition of natural grammar, and syntactic processing. Dorsal pathways connect the left posterior temporal lobe with the left posterior frontal lobe via the arcuate fasciculus and the longitudinal fasciculus and have been associated with mapping sound onto articulatory based representations, processing complex syntactic operations possible due to an increase in syntactic working memory load, and may also serve higher level semantic and syntactic language functions (Friederici, 2011; Hickok and Poeppel, 2004, 2007; Saur et al., 2008; Weiller et al., 2009, 2011).

Although the left dorsal and ventral streams have mostly been discussed as separate systems, it has been suggested that the language-relevant brain areas can exchange information bidirectionally and in parallel via dorsal and ventral pathway systems and form a kind of feedback loop that needs to be passed at least once during language processing (Weiller et al., 2011).

Anatomical Asymmetries and Language Functions

To the naked eye, the left and right halves of the brain look very similar. Although the two hemispheres have a predominantly equivalent structure, anatomical (structural) asymmetries have been reported in various regions of the brain, and this not only in the adult but also in the immature human as well as in the ape brain (Hiscock and Kinsbourne, 2008). It should be noted that brain asymmetry is not necessarily a static feature and may vary in the general population. Empirical data suggest that experiential and environmental factors as well as behavioral and gender-related factors can modulate brain asymmetry (Renteria, 2012).

The important question, however, what these anatomical asymmetries mean at a functional level is not always clear. (Note that functional asymmetry is often referred to as lateralization).

The Planum Temporale

In 1968, Geschwind and Levitsky (1968) reported a leftward asymmetry of the planum temporale surface area (a region located in the superior temporal gyrus behind Heschl's gyrus). This and subsequent confirmation of these findings was interpreted by some researchers as neuroanatomical evidence for the left hemisphere's dominance for speech. Forty-five years later, the issue is still controversial. Opposing this view are observations that the incidence of left-hemisphere planum temporale asymmetry is around 60–80% while the incidence of left-hemisphere language lateralization has been estimated to be around 96% or higher (Dorsaint-Pierre et al., 2006). In addition, the measurement and boundaries to adjacent regions in individual studies differ and not all studies have reported asymmetries (Dorsaint-Pierre et al., 2006; Mazoyer and Tzourio-Mazoyer, 2004; Steinmetz, 1996). Considering the currently unclear situation, studies that link language dominance with anatomical differences in the planum temporale should thus be interpreted with caution until further evidence emerges.

The Auditory Cortex and Heschl's Gyrus

Another asymmetry in the auditory cortex concerns Heschl's gyrus whose volume is larger in the left than the right hemisphere. Although many individuals show this asymmetry, not all individuals do. This has been interpreted as suggesting that the relationship between structural asymmetry and language function in Heschl's gyrus is not always mandatory (Dorsaint-Pierre et al., 2006). Some research indicates that the anatomy and asymmetry of Heschl's gyrus in the auditory cortex may be a contributing predictor for learning foreign speech sounds. This is based on the observation that fast phonetic learners displayed a higher density of fiber tracts (white matter) in Heschl's gyrus than slow phonetic learners. At the same time, fast learners exhibited a leftward asymmetry in parietal lobe volumes while slow learners showed a more superior location of Heschl's gyrus (Golestani and Zatorre, 2004).

The Inferior Frontal Gyrus and Broca's Region

Anatomical differences between the two hemispheres have also been described in the inferior frontal gyrus, a brain region that in the left hemisphere entails Broca's area. These differences have been found at the cellular level, at the receptor level (important for signal transduction), and with fiber tracts (connecting brain regions). In the left hemisphere, these anatomical asymmetries have generally been interpreted as correlates for the lateralization of brain functions including language (Amunts, 2008; Amunts et al., 2010; Renteria, 2012). Already in 1980, Galaburda hypothesized (without brain imaging techniques) that the anatomical asymmetry of the pars opercularis, a region that is part of Broca's area, is a correlate of the functional lateralization of speech production (Galaburda, 1980 cited in Amunts et al., 2012). Others have speculated that the pars opercularis is use-dependent such that in the expression of speech, patients with left-hemisphere speech may employ more the left motor-related regions and patients with right-hemisphere speech the right motor-related regions (Dorsaint-Pierre et al., 2006). It has further been observed that the asymmetry in Broca's region (which is larger in the left hemisphere compared to its right-sided homologue) seems to increase with age with the parcs opercularis reaching adult levels around 11 years of age and the pars triangularis around 5 years of age. The observation that syntactic processes activate the pars opercularis and semantic processes activate the pars triangularis, and both processes develop at a different age in children has given rise to the *speculation* that the anatomically asymmetric development has a functional correlate (Amunts, 2008; Amunts et al., 2012).

Asymmetries in Pathological Populations

Abnormalities in asymmetries (including in the planum temporale) have also been found in certain pathologies such as developmental dyslexia, specific language impairment (SLI) and schizophrenia. Schizophrenia patients show reduced asymmetry in the volume of the left planum temporale, left Heschel's gyrus, and the right anterior cingulate gyrus (Crow, 2008; Honea et al., 2005; Kawasaki et al., 2008; Renteria, 2012). Auditory hallucinations in schizophrenia have been associated with increased blood flow in areas specialized for language. In SLI, abnormal asymmetry in Broca's area, atypical rightward asymmetry in anterior, and posterior peri-Sylvian areas (including the planum temporale) have been described and hypothesized to give rise to weak language skills (Badcock et al., 2012). Most neuroimaging studies have focused on relating patterns of activation in specific brain regions in response to specific tasks with regional anatomical deviances. Only recently have connectivity patterns between brain regions and the neurochemistry in these regions been investigated. It has, for example, been found that patients with autistic spectrum disorder – characterized by social, language, and other cognitive disorders – show an atypically connected brain as well as neurochemical abnormalities (Anagnostou and Taylor, 2011). Generally, however, the exact mechanism by which abnormal structural asymmetries in pathological populations may contribute to deviant language or communicative problems is still unknown and a matter of debate.

To sum up, several anatomical (structural) asymmetries of brain regions that belong to the language network have been described such as the planum temporale, Heschl's gyrus, and the inferior frontal gyrus including Broca's region. Abnormal asymmetries have also been described in certain pathologies such as developmental language disorders, schizophrenia, or autistic spectrum disorder. Although attempts have been made to relate such asymmetries to normal or abnormal language functions, it is currently not clear what exactly this relationship is.

Lateralization in Special Populations

The question whether special language skills implicate the two hemispheres in different ways has been a matter of wide interest, for example, in deaf people who use sign language or in people speaking two or more languages.

Sign-Language Users

The question what or how the two hemispheres contribute to language processing has most frequently been investigated with regard to spoken language with the vocal tract as the primary articulator. What, however, happens if the articulator consists of hands and nonmanual components (e.g., mouth) such as used by sign-language users? The first question that needs to be answered is whether the brains of deaf people look different from those who can hear. This question is currently underexplored but from the few studies that investigated this issue, it seems that landmark regions, such as the auditory cortex, seem to be similar in volume although there may be subtle differences in white matter connections (Campbell et al., 2008). At a functional level, the language regions in the left hemisphere seem just as important in deaf and hearing native signers as in spoken language users. This also pertains to impaired language processing: damage to the left hemisphere can produce aphasia in signers just as it does in spoken language users, whereas right hemisphere damage usually does not.

While most researchers agree that sign production is strongly left-lateralized, the contribution of the two hemispheres to sign perception is still controversial as findings do not always converge. For example, while some researchers report left-hemisphere dominance for signed language perception, others

have argued that the right hemisphere seems to be more recruited than the left hemisphere by native American sign-language users compared to written language processing of hearing nonsigners (Neville et al., 1998; for a summary see Campbell et al., 2008; MacSweeney et al., 2008). Explanations for these differences might be discrepancies in the experimental paradigms or differences between languages of the study participants.

Hemisphere differences have also been observed when information is conveyed by facial expression. For example, when British sign-language users use a headshake (or the manual equivalent) to signal negation, only the right hemisphere seems involved while emotional facial expressions can be processed bilaterally or predominantly in the left hemisphere. In hearing the emotional facial expressions of nonsigners are processed predominantly in the right hemisphere. This difference between sign language and spoken language users probably reflects that the face serves a wide range of functions in signers and, as a consequence, has resulted in cerebral reorganization (MacSweeney et al., 2008).

Multilingual Populations

The question where and how different languages are represented and processed in the brain of multilingual speakers has received considerable attention. Early studies investigated this question by electrically stimulating the exposed surface of the brain while multilingual patients underwent epilepsy surgery. The left hemisphere was involved in the representation of different languages when the brain was electrically stimulated during naming and reading tasks. The sites identified were widely distributed and overlapped only partially; there were sites at which one language alone was represented or sites at which both languages overlapped to different degrees (Ojemann and Whitaker, 1978; Rapport et al., 1983). Only two patients were reported that showed right-hemisphere involvement; one who generally showed right-hemisphere dominance for language and the other whose bilateral representation was probably due to a congenital lesion in the left-Sylvian fissure.

Subsequent neuroimaging studies using the fMRI or PET technique provided inconsistent results. Some early studies reported no interhemispheric difference for different languages (Klein et al., 1999), others reported representation of second languages in the left hemisphere alone (Kim et al., 1997) or in both hemispheres (Dehaene et al., 1997), or different findings for different participants (Chee et al., 2001). These inconsistencies are probably due to methodological issues such as different scanning techniques, different stimuli, tasks, task demands, and experimental procedures. Other reasons relate to different subject characteristics (age of acquisition, proficiency level, length of exposure to the second language).

Despite the inhomogeneity of studies, reviewing numerous neuroimaging studies targeting the neural bases for first and second language processing, Perani and Abutalebi (2005) concluded that a second language seems to be acquired through the same neural devices as a first language. Furthermore, tasks that engage specific aspects of linguistic processing activate similar brain patterns during first and second language processing. The authors further concluded that the most important variables that influence and modify the elicited brain patterns and the way a language is represented are language proficiency, age of acquisition, and the amount of exposure.

Differences in the linguistic structure of two languages have also been shown to influence language representation and processing in early and highly proficient bilinguals. Investigating syntactic processing in English and Spanish bilinguals, Kovelman et al. (2008) reported differentiated neural patterns of activation of the two languages and these differences were predictable from differences in the linguistic structures typical of the two languages.

Most neuroimaging studies report predominantly left hemisphere activation of both the first and the second language in the language-specific regions. These findings converge with behavioral studies that have shown similar left hemisphere dominant lateralization for the two languages in proficient bilinguals and stronger lateralization for the first than the second language in nonproficient (late) bilinguals (Hull and Vaid, 2007). Neuroimaging studies have not investigated language lateralization in young bilingual infants. Behavioral studies suggest that within bilinguals the two languages show functionally similar patterns in the sense that if one language in infant bilinguals is bilaterally organized so is the other, and when the first language of late bilinguals is left hemisphere dominant, so is the other (Hull and Vaid, 2007).

Lateralization in Pathological Populations

Structural asymmetries in pathological populations have already been discussed in Section Asymmetries in Pathological Populations. In the following, we will focus on the contribution of the hemispheres during the recovery process and how to determine language lateralization before surgery.

Hemispherectomy in Childhood

There are only few cases of children described in the literature whose entire hemisphere has been surgically removed (hemispherectomy) and the language tests in these children vary widely and are often limited to the word or sentence level. It is thus problematic to draw generalization from these individual cases. Nevertheless, the reported cases are witness to the extraordinary ability of the brain to recover and reorganize. Children born with left hemisphere pathology and subsequent left hemispherectomy may develop near-normal speech and communication abilities. The situation seems different for children who develop pathology after 5 years of age. In these children, after left hemispherectomy, receptive language functions were often spared while productive language functions recovered slowly and incomplete (summary in Stemmer, 2006). In the right hemisphere, a variable pattern of language networks activated in the inferior frontal regions when words had to be generated, while a more consistent pattern was observed in temporal and precentral regions (Liégeois et al., 2008).

For a long time, it was believed that this recovery potential is limited to children and does not apply to adults. Undoubtedly, recovery and reorganization in children can be very dramatic but it also occurs in adulthood, even until late in life, albeit to a lesser degree.

Aphasia

Aphasia is a loss or impairment of the ability to appropriately produce and/or comprehend language, usually due to left-sided brain damage such as a stroke. Clinical recovery from aphasic syndromes in adults can be observed over days, weeks, months, or years after the incident.

Acute Phase after Brain Damage

Neuroimaging studies have tried to elucidate the neural mechanisms underlying early recovery. In the early acute phase (up to about 6 months) after a stroke, recovery of language function depends on how well blood supply returns to the tissue that has (at least partially) been preserved. This is called reperfusion. It further depends on how well edema (abnormal accumulation of fluid) at the lesion site are resolved and how well the metabolism at implicated but distant sites of the infarct is restored (Hillis et al., 2008; Lee and Hillis, 2008; for a summary see Berthier et al., 2011). Empirical data suggest that in acute aphasia after a stroke the main mechanism underlying recovery is the restoration of brain networks around the damaged site in the left hemisphere. Although contribution of the right hemisphere does occur, it is frequently minor unless the left hemisphere language networks are severely damaged (Heiss and Thiel, 2006). The more efficient the left hemisphere networks, the less the contribution of the right hemisphere (Winhuisen et al., 2005, 2007).

Chronic Phase after Brain Damage

Contrary to early beliefs that there is no language recovery the first year after the incident, we now know that significant improvement of language functions can still occur for many years later. Unlike the acute phase, the main mechanisms that drive recovery in the chronic phase seem to be the recruitment of additional brain areas (compensation) and displacement of a primary language area (reorganization) (for a summary see (Berthier et al., 2011; Meinzer et al., 2008, 2011)). The reorganization of neural networks can be triggered spontaneously or through intensive training, pharmacological treatments, or electromagnetic stimulation (repetitive TMS; transcranial direct current stimulation). Both hemispheres seem to be involved in the reorganization of brain networks in chronic poststroke aphasia but due to individual variability and the heterogeneity of the patient population and lesion characteristics, it has been difficult to specify the role that each hemisphere occupies in this process (for details see Berthier et al., 2011).

Defining Hemispheric Lateralization in Presurgical Patients

When patients need to undergo neurosurgical intervention, for example, to resect epilepsy foci or a tumor, one tries to preserve neural substrate that is essential for important functions such as aspects of language. For a long time, the gold standard to determine in which hemisphere language is lateralized in these patients was the presurgical Wada procedure. During this procedure, one hemisphere is anesthetized while the patient has to perform simple language tasks such as naming objects or following simple commands. If the left hemisphere is anesthetized and the patient is unable to perform the language task, this suggests left-sided dominance for the task. Most neurosurgical patients tested with the Wada procedure have been shown to be left hemisphere dominant. Right-hemisphere dominance for language in neurosurgical patients has been estimated to range from 2 to 10% (for a summary see Drane et al., 2012).

Comparisons between the rather invasive and thus limited Wada test and the noninvasive fMRI procedure has shown that there is a good agreement between the two procedures. The fMRI procedure may even outperform the Wada procedures due to its less limiting time constraints and more flexible application (Binder, 2011; Dym et al., 2011). In addition, rather than a 'pass or fail' judgment, fMRI may improve prediction accuracy (such as the risk for postoperative naming deficits) when combined with other noninvasive measures and more realistically predict the actual outcomes (Binder, 2011).

Lateralization in Males and Females

A popular belief is that language functions are more lateralized to the left side of the brain in men and represented more bilaterally in women. Unfortunately, scientific support for this statement is ambiguous. Although a number of brain imaging studies have reported more bilateral activity during language processing in women compared to men, meta-analyses of these studies did not find a significant difference of language lateralization in men or women, regardless of whether they were adults or children (for a summary see Wallentin, 2009). Aside from methodological differences and high-individual variation, many studies that report differences relied on a relatively small sample size while studies that did not find differences were based on larger samples. Studies investigating low-level processing (verb generation, word completion, or semantic word judgment tasks) reported no lateralization differences in men or women while others using rhyming judgments or passive listening to spoken narratives found more bilateral activation in women compared to men (for a summary see Stemmer, 2006).

Gender differences at the structural level have also been documented. In males, the left planum temporale is larger than its right homologue, whereas women do not show this difference. Whether the corpus callosum (CCA), a large fiber bundle that connects the two hemispheres, is larger in females than in males has been controversial. While some researchers conclude that the sex-related CCA difference is a myth (e.g., Wallentin, 2009), Ardekani et al. (2012) measured the cross-sectional area of the CCA and reported a larger CCA in females compared to males. Regardless of the controversies surrounding structural brain asymmetries in males and females, an important point is that it is currently not clear what such differences mean in terms of their functional significance such as language lateralization.

Genetics and Lateralization

Some developmental language disorders such as dyslexia or SLI have been associated with genetic abnormalities (Ramus and

Fisher, 2009; Newbury et al., 2010). It is further assumed that genes are involved in determining cerebral asymmetry (Bishop, 2013). The question thus asked is whether there is a link between the genetic underpinnings of developmental language disorders and cerebral asymmetry.

Neuroimaging studies have shown anatomical (structural) asymmetries (see Section Asymmetries in Pathological Populations) as well as a poorly functionally lateralized brain such as reduced left hemisphere activity during word generation tasks in developmental language disorders. Several scenarios have been suggested that may link genetic influence, asymmetry, and language problems. One scenario is that genes influence language disorders by affecting cerebral lateralization. Another scenario suggests that the same genes that affect language disorders also affect lateralization. Finally, genes influence language disorders and the disorder results in weak laterality. (See **Figure 4**; for details see Bishop, 2013.) Research on linking the genetics of language disorders with the genetics of cerebral lateralization is still in its infancy and it is currently not possible to draw well-founded conclusions. Although genetic abnormalities have been found in some developmental language disorders and genes are most probably involved in determining cerebral asymmetry, the role of these genes, their interaction with, and the influence of nongenetic factors on cerebral asymmetry is currently open.

Lateralization across a Life Span

An old question that has attracted much research is whether there are any age-related changes in language lateralization. Despite the large corpus of data accumulated, the evidence is often difficult to interpret and ambiguous. The main models that have been advanced in relation to whether language lateralization changes across the life span are the developmental invariance model, the progressive lateralization models, and the progressive degeneration model (for a summary see Hiscock and Kinsbourne, 2008). The developmental invariance model posits that the left hemisphere is specialized for language at the time of or even prior to birth and no changes will occur during further development. The progressive lateralization models postulate that the bilateral (functional) symmetry of the infant's brain gradually lateralizes to the left hemisphere at some stage

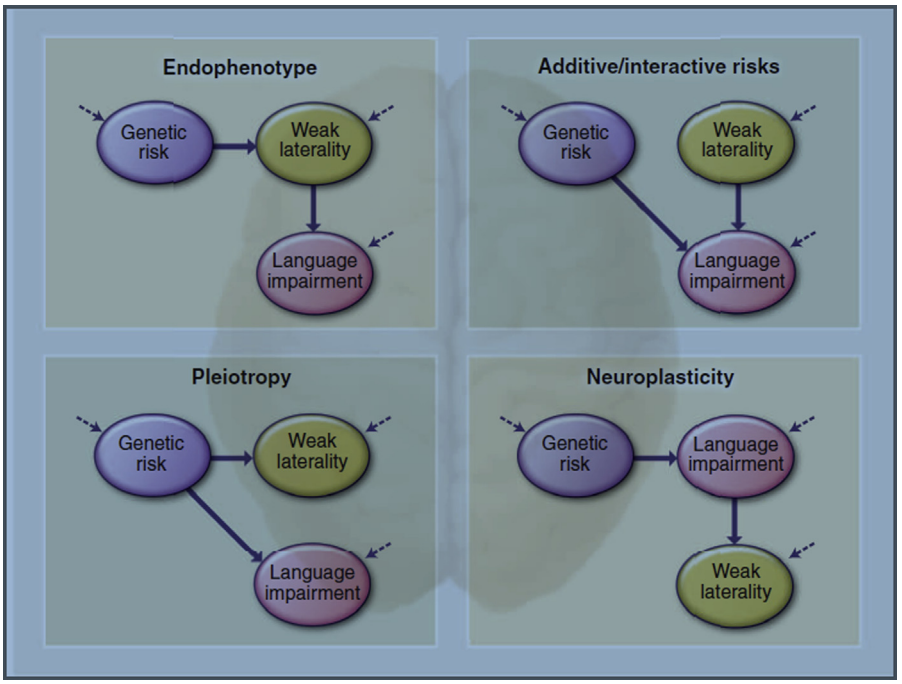

Figure 4 Four path diagrams of the association between weak cerebral lateralization for language and language and literacy problems. These causal models are simplified abstractions and not mutually exclusive, but they illustrate the differing predictions about the pattern of associations that might be found between genotypes, language and literacy impairment, and cerebral lateralization. The endophenotype model depicts the situation in which genes that influence risk for language impairment do so by affecting cerebral lateralization. This model predicts that cerebral lateralization should be at least as heritable as language impairment, with the same genes affecting both traits. The pleiotropy model similarly assumes that the same genes that lead to risk for language impairment also affect cerebral lateralization, but there is no direct causal link: Weak laterality and language impairment co-occur because they have common origins, not because one causes the other. In the additive/interactive risks model, the genetic risk factors for language impairment do not affect lateralization. However, weak laterality, which could have genetic and/or nongenetic origins, exerts an independent causal influence on language impairment, which may add or interact with other genetic risk factors. In the neuroplasticity model, cerebral lateralization has no causal effect on language; rather, language impairment influences how the brain develops and is associated with weaker cerebral lateralization. The dotted arrows indicate that, for each construct, there will be sources of variation in addition to those depicted in the model. Bishop, D.V., 2013. Cerebral asymmetry and language development: cause, correlate, or consequence? Science 340, 1230531.

during development (e.g., around puberty or by 5 years of age), or during maturity, or even into senescence. The progressive degeneration model suggests that the function of one hemisphere (usually the right) degenerates more rapidly than the function of the other thus leading to an asymmetry. While most evidence seems to speak against the developmental invariance model, general problems, such as insufficient data on functional maturation in relation to asymmetry or methodological limitations with very young or very old populations, make it difficult to obtain a clear picture.

See also: Asymmetry of Body and Brain; Functional Brain Imaging of Language processes.

Bibliography

Amunts, K., 2008. Architectonic language research. In: Stemmer, B., Whitaker, H.A. (Eds.), Handbook of the Neuroscience of Language. Academic Press/Elsevier, Amsterdam, San Diego.

Amunts, K., Lenzen, M., Friederici, A.D., Schleicher, A., Morosan, P., Palomero-GALLAGHER, N., Zilles, K., 2010. Broca's region: novel organizational principles and multiple receptor mapping. PLoS Biology 8, e1000489.

Amunts, K., Morosan, P., Hilbig, H., Zilles, K., 2012. Chapter 36-Auditory system. In: Juergen, K.M., George, P. (Eds.), The Human Nervous System, third ed. Academic Press, San Diego.

Anagnostou, E., Taylor, M., 2011. Review of neuroimaging in autism spectrum disorders: what have we learned and where we go from here. Molecular Autism 2, 4.

Ardekani, B.A., Figarsky, K., Sidtis, J.J., 2012. Sexual Dimorphism in the Human Corpus Callosum: An MRI Study Using the OASIS Brain Database. Cerebral Cortex.

Badcock, N.A., Bishop, D.V.M., Hardiman, M.J., Barry, J.G., Watkins, K.E., 2012. Co-localisation of abnormal brain structure and function in specific language impairment. Brain and Language 120, 310–320.

Berthier, M.L., Pulvermüller, F., Davila, G., Casares, N.G., Gutierrez, A., 2011. Drug therapy of post-stroke aphasia: a review of current evidence. Neuropsychology Review 21, 302–317.

Binder, J.R., 2011. Functional MRI is a valid noninvasive alternative to Wada testing. Epilepsy and Behavior 20, 214–222.

Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cerebral Cortex 19, 2767–2796.

Bishop, D.V., 2013. Cerebral asymmetry and language development: cause, correlate, or consequence? Science 340, 1230531.

Bohrn, I.C., Altmann, U., Jacobs, A.M., 2012. Looking at the brains behind figurative language—A quantitative meta-analysis of neuroimaging studies on metaphor, idiom, and irony processing. Neuropsychologia 50, 2669–2683.

Campbell, R., Macsweeney, M., Waters, D., 2008. Sign language and the brain: a review. Journal of Deaf Studies and Deaf Education 13, 3–20.

Cappa, S.F., 2012. Imaging semantics and syntax. Neuroimage 61, 427–431.

Catani, M., Jones, D.K., Ffytche, D.H., 2005. Perisylvian language networks of the human brain. Annals of Neurology 57, 8–16.

Catani, M., Mesulam, M., 2008. The arcuate fasciculus and the disconnection theme in language and aphasia: history and current state. Cortex 44, 953–961.

Chee, M.W., Hon, N., Lee, H.I., Soon, C.S., 2001. Relative language proficiency modulates BOLD signal change when bilinguals perform semantic judgments. Neuroimage 13, 155–163.

Crow, T.J., 2008. Chapter 29-The Torque Defines the Four Quadrants of the Human Language Circuit and the Nuclear Symptoms of Schizophrenia Identify Their Component Functions. Handbook of the Neuroscience of Language. Elsevier, San Diego.

Dehaene, S., Dupoux, E., Mehler, J., Cohen, L., Paulesu, E., Perani, D., Van De Moortele, P.F., Lehericy, S., Le Bihan, D., 1997. Anatomical variability in the cortical representation of first and second language. Neuroreport 8, 3809–3815.

Dorsaint-Pierre, R., Penhune, V.B., Watkins, K.E., Neelin, P., Lerch, J.P., Bouffard, M., Zatorre, R.J., 2006. Asymmetries of the planum temporale and Heschl's gyrus: relationship to language lateralization. Brain 129, 1164–1176.

Drane, D.L., Roraback-Carson, J., Hebb, A.O., Hersonskey, T., Lucas, T., Ojemann, G.A., Lettich, E., Silbergeld, D.L., Miller, J.W., Ojemann, J.G., 2012. Cortical stimulation mapping and Wada results demonstrate a normal variant of right hemisphere language organization. Epilepsia 53, 1790–1798.

Dym, R.J., Burns, J., Freeman, K., Lipton, M.L., 2011. Is functional MR imaging assessment of hemispheric language dominance as good as the Wada test? A meta-analysis. Radiology 261, 446–455.

Ferstl, E.C., 2010. Neuroimaging of text comprehension: where are we now? Italian Journal of Linguistics 22, 61–88.

Friederici, A.D., 2009. Pathways to language: fiber tracts in the human brain. Trends in Cognitive Sciences 13, 175–181.

Friederici, A.D., 2011. The brain basis of language processing: from structure to function. Physiological Reviews 91, 1357–1392.

Friederici, A.D., 2012. The cortical language circuit: from auditory perception to sentence comprehension. Trends in Cognitive Sciences 16, 262–268.

Galaburda, A.M., 1980. La region de Broca: observations anatomiques faites un siecle apres la mort de son decoveur. Revue Neurologique 136, 609–616.

Gandour, J., Tong, Y., Wong, D., Talavage, T., Dzemidzic, M., Xu, Y., Li, X., Lowe, M., 2004. Hemispheric roles in the perception of speech prosody. Neuroimage 23, 344–357.

Gandour, J., Wong, D., Hsieh, L., Weinzapfel, B., Van Lancker, D., Hutchins, G.D., 2000. A crosslinguistic PET study of tone perception. Journal Cognitive Neuroscience 12, 207–222.

Geng, X., Gouttard, S., Sharma, A., Gu, H., Styner, M., Lin, W., Gerig, G., Gilmore, J.H., 2012. Quantitative tract-based white matter development from birth to age 2 years. NeuroImage 61, 542–557.

Geschwind, N., 1965. Disconnexion syndromes in animals and man. I. Brain 88, 237–294.

Geschwind, N., Levitsky, W., 1968. Human brain: left-right asymmetries in temporal speech region. Science 161, 186–187.

Golestani, N., Zatorre, R.J., 2004. Learning new sounds of speech: reallocation of neural substrates. NeuroImage 21, 494–506.

Hartwigsen, G., Siebner, H.R., 2011. Probing the involvement of the right hemisphere in language processing with online transcranial magnetic stimulation in healthy volunteers. Aphasiology 26, 1131–1152.

Heiss, W.D., Thiel, A., 2006. A proposed regional hierarchy in recovery of post-stroke aphasia. Brain and Language 98, 118–123.

Hellige, J.B., 2008. Chapter 25-Interhemispheric Interaction in the Lateralized Brain. Handbook of the Neuroscience of Language. Elsevier, San Diego.

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92, 67–99.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nature Reviews Neuroscience 8, 393–402.

Hillis, A.E., Gold, L., Kannan, V., Cloutman, L., Kleinman, J.T., Newhart, M., Heidler-Gary, J., Davis, C., Aldrich, E., Llinas, R., Gottesman, R.F., 2008. Site of the ischemic penumbra as a predictor of potential for recovery of functions. Neurology 71, 184–189.

Hiscock, M., Kinsbourne, M., 2008. Chapter 24-Lateralization of Language across the Life Span. Handbook of the Neuroscience of Language. Elsevier, San Diego.

Honea, R., Crow, T.J., Passingham, D., Mackay, C.E., 2005. Regional deficits in brain volume in schizophrenia: a meta-analysis of voxel-based morphometry studies. American Journal of Psychiatry 162, 2233–2245.

Hull, R., Vaid, J., 2007. Bilingual language lateralization: a meta-analytic tale of two hemispheres. Neuropsychologia 45, 1987–2008.

Kawasaki, Y., Suzuki, M., Takahashi, T., Nohara, S., Mcguire, P.K., Seto, H., Kurachi, M., 2008. Anomalous cerebral asymmetry in patients with schizophrenia demonstrated by voxel-based morphometry. Biologial Psychiatry 63, 793–800.

Kiefer, M., Pulvermüller, F., 2012. Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. Cortex 48, 805–825.

Kim, K.H.S., Relkin, N.R., Kyoung-Min, L., Hirsch, J., 1997. Distinct cortical areas associated with native and second languages. Nature 388, 171–174.

Klein, D., Milner, B., Zatorre, R.J., Zhao, V., Nikelski, J., 1999. Cerebral organization in bilinguals: a PET study of Chinsese-English verb generation. NeuroReport 10, 2841–2846.

Kovelman, I., Baker, S.A., Petitto, L.A., 2008. Bilingual and monolingual brains compared: a functional magnetic resonance imaging investigation of syntactic processing and a possible "neural signature" of bilingualism. Journal of Cognitive Neuroscience 20, 153–169.

Liégeois, F., Connelly, A., Baldeweg, T., Vargha-Khadem, F., 2008. Speaking with a single cerebral hemisphere: fMRI language organization after hemispherectomy in childhood. Brain and Language 106, 195–203.

Lee, A.W., Hillis, A.E., 2008. Chapter 40-The Pharmacological Treatment of Aphasia. Handbook of the Neuroscience of Language. Elsevier, San Diego.

MacSweeney, M., Capek, C.M., Campbell, R., Woll, B., 2008. The signing brain: the neurobiology of sign language. Trends in Cognitive Sciences 12, 432–440.

Mazoyer, B.M., Tzourio-Mazoyer, N.G., 2004. Planum temporale asymmetry and models of dominance for language: a reappraisal. Neuroreport 15, 1057–1059.

Meinzer, M., Flaisch, T., Breitenstein, C., Wienbruch, C., Elbert, T., Rockstroh, B., 2008. Functional re-recruitment of dysfunctional brain areas predicts language recovery in chronic aphasia. Neuroimage 39, 2038–2046.

Meinzer, M., Harnish, S., Conway, T., Crosson, B., 2011. Recent developments in functional and structural imaging of aphasia recovery after stroke. Aphasiology 25, 271–290.

Minagawa-Kawai, Y., Cristià, A., Dupoux, E., 2011. Cerebral lateralization and early speech acquisition: a developmental scenario. Developmental Cognitive Neuroscience 1, 217–232.

Neville, H.J., Bavelier, D., Corina, D., Rauschecker, J., Karni, A., Lalwani, A., Braun, A., Clark, V., Jezzard, P., Turner, R., 1998. Cerebral organization for language in deaf and hearing subjects: biological constraints and effects of experience. Proceedings of the National Academy of Sciences of the United States of America 95, 922–929.

Newbury, D., Fisher, S., Monaco, A., 2010. Recent advances in the genetics of language impairment. Genome Medicine 2, 6.

Ojemann, G.A., Whitaker, H.A., 1978. The bilingual brain. Archives of Neurology 35, 409–412.

Perani, D., Abutalebi, J., 2005. The neural basis of first and second language processing. Current Opinion in Neurobiology 15, 202–206.

Petersen, S.E., Fiez, J.A., 1993. The processing of single words studied with positron emission tomography. Annual Review of Neuroscience 16, 509–530.

Petersen, S.E., Fox, P.T., Posner, M.I., Mintun, M., Raichle, M.E., 1988. Positron emission tomographic studies of the cortical anatomy of single-word processing. Nature 331, 585–589.

Plante, E., Creusere, M., Sabin, C., 2002. Dissociating sentential prosody from sentence processing: activation interacts with task demands. Neuroimage 17, 401–410.

Price, C.J., 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage 62, 816–847.

Ramus, F., Fisher, S.E., 2009. Genetics of language. In: Gazzaniga, M.S. (Ed.), The Cognitive Neurosciences, fourth ed. MIT Press, Cambridge, MA.

Rapport, R.L., Tan, C.T., Whitaker, H.A., 1983. Language function and dysfunction among Chinese- and English-speaking polyglots: cortical stimulation, Wada Testing, and clinical studies. Brain and Language 18, 342–366.

Renteria, M.E., 2012. Cerebral asymmetry: a quantitative, multifactorial, and plastic brain phenotype. Twin Research and Human Genetics 15, 401–413.

Saur, D., Kreher, B.W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M.-S., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, W., Rijntjes, M., Hennig, J., Weiller, C., 2008. Ventral and dorsal pathways for language. Proceedings of the National Academy of Sciences 105, 18035–18040.

Saur, D., Schelter, B., Schnell, S., Kratochvil, D., Küpper, H., Kellmeyer, P., Kümmerer, D., Klöppel, S., Glauche, V., Lange, R., Mader, W., Feess, D., Timmer, J., Weiller, C., 2010. Combining functional and anatomical connectivity reveals brain networks for auditory language comprehension. Neuroimage 49, 3187–3197.

Schmidt, G.L., Seger, C.A., 2009. Neural correlates of metaphor processing: the roles of figurativeness, familiarity and difficulty. Brain and Cognition 71, 375–386.

Steinmetz, H., 1996. Structure, function and cerebral asymmetry: *in vivo* morphometry of the planum temporale. Neuroscience & Biobehavioral Reviews 20, 587–591.

Stemmer, B., 2006. Imaging brain lateralization: words, sentences, and influencing factors in healthy, pathological, and special populations. In: Keith, B. (Ed.), Encyclopedia of Language & Linguistics, second ed. Elsevier, Oxford.

Stemmer, B., 2013. Neuropragmatics. In: Huang, Y. (Ed.), Handbook of Pragmatics. Oxford University Press, Oxford. Retrieved from Oxford Handbooks Online, http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199697960.001.001/oxfordhb-9780199697960-e-003.

Vigneau, M., Beaucousin, V., Herve, P.-Y., Jobard, G., Petit, L., Crivello, F., Mellet, E., Zago, L., Mazoyer, B., Tzourio-Mazoyer, N., 2011. What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing? Insights from a meta-analysis. Neuroimage 54, 577–593.

Vigneau, M., Beaucousin, V., Herve, P.Y., Duffau, H., Crivello, F., Houde, O., Mazoyer, B., Tzourio-Mazoyer, N., 2006. Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. Neuroimage 30, 1414–1432.

Wallentin, M., 2009. Putative sex differences in verbal abilities and language cortex: a critical review. Brain and Language 108, 175–183.

Weiller, C., Bormann, T., Saur, D., Musso, M., Rijntjes, M., 2011. How the ventral pathway got lost – and what its recovery might mean. Brain and Language 118, 29–39.

Weiller, C., Musso, M., Rijntjes, M., Saur, D., 2009. Please don't underestimate the ventral pathway in language. Trends in Cognitive Sciences 13, 369–370.

Winhuisen, L., Thiel, A., Schumacher, B., Kessler, J., Rudolf, J., Haupt, W.F., Heiss, W.D., 2005. Role of the contralateral inferior frontal gyrus in recovery of language function in poststroke aphasia: a combined repetitive transcranial magnetic stimulation and positron emission tomography study. Stroke 36, 1759–1763.

Winhuisen, L., Thiel, A., Schumacher, B., Kessler, J., Rudolf, J., Haupt, W.F., Heiss, W.D., 2007. The right inferior frontal gyrus and poststroke aphasia: a follow-up investigation. Stroke 38, 1286–1292.

Latin-American Legal Systems

Javier Couso, Facultad de Derecho, Universidad Diego Portales, Santiago, Chile

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Latin-American legal systems' is a label that groups together a diverse set of 20 countries that share a common historical background, geographical proximity, similar cultural traits, and, often, an important degree of economic interaction. As a consequence of these features, they generally exhibit similar patterns in their law. They also differ in important ways, however, such as regarding the degree to which formal state law penetrates social, economic, and political life; the different roles played by constitutions and judicial systems; and the extent to which they exhibit legal pluralism. After accounting for similarities and differences among the Latin-American legal systems, the article describes the role that International Human Rights Law is playing in the legal life of the region, as well as the convergence with Anglo-American legal tradition, in both Public and Private Law.

I

The expression 'Latin-American legal systems' is an intellectual construct designed to group together a diverse set of countries that share a common historical background, geographical proximity, similar cultural traits, and – often – an important degree of economic interaction. As a consequence of these features, they generally exhibit similar patterns in their law.

Historically, the 20 Latin-American states (Argentina, Brazil, Bolivia, Chile, Colombia, Costa Rica, Cuba, Ecuador, El Salvador, Guatemala, Haiti, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Dominican Republic, Uruguay, and Venezuela) were classified as belonging to the 'Civil Law Tradition' (Merryman, 1969), given the influence that Continental European Law had in the region since the colonial era and throughout the formative years as independent nations – in the first half of the nineteenth century – particularly in the domains of Civil and Procedural Law. As members of this legal tradition, most of the Latin-American countries exhibited a strong attachment to codified law and a corresponding opposition to judge-made law and to the doctrine of 'stare decisis.' Also in line with the Continental European tradition, the systems of this region gave a prominent role to legal academics in the development of law, through the elaboration of doctrine. Further characteristics of the Latin-American legal systems include the sway that legal formalism had over judges in their adjudicative role, and the hierarchical structure of the judiciary. Finally, many observers emphasize the chronic inability of most of the formal legal systems of the region to actually govern the social, political, and economic relations of their countries (García Villegas, 2009; Nino, 1992). Indeed, ever since the colonial practice of formally recognizing the legitimacy of royal decrees but then openly disobeying them ('*La ley se acata, pero no se cumple*'), the incapacity of formal state law to penetrate society has been a constant concern in many Latin-American countries.

Although the assessment that the Latin-American legal systems are the childbirth of Continental European Law is largely right when it comes to Private Law, this is not the case with regard to Public Law, where the United States' constitutional model had an enormous impact (Gargarella, 2013). This influence explains why all states of the region have presidentialist regimes and why several of them have a federalist system (Brazil, Mexico, and Argentina). Thus, the notion that Latin-American legal systems are simply a subset of the 'Civil Law Tradition' has to be qualified, and attributed to the fact that, until recently, legal analysts of the region focused mostly on Private Law.

II

In spite of the common traits exhibited by the legal systems of Latin America, it has to be noted that there are significant differences in terms of the relevance of formal state law in each of the countries of the region (Sieder, 2002). This is intimately related to the degree of state formation exhibited by the Latin-American nations (López-Alves, 2000). Thus, while some countries have managed to consolidate constitutional regimes where state law reaches most of the national territory (Argentina, Chile, Costa Rica, and Uruguay), others exhibit weak states, where the formal legal system is unable to reach vast segments of society. These places are instead ruled by informal practices (Haiti and Paraguay) or, in some cases, by indigenous law (Bolivia, Guatemala, and, to a lesser extent, Ecuador and Peru). The rest of the states of the region fall in between the poles of the prevalence of state law or the coexistence of the latter with indigenous and other types of informal law.

An additional difference that can be observed among Latin-American states is the degree of consolidation of the rule of law – and, more specifically, judicial independence – exhibited by each of the countries of the region. Thus, for example, a report by the 'World Economic Forum' ranks Chile, Costa Rica, and Uruguay within the top 30 states in the world in terms of judicial independence, while it found Paraguay and Venezuela to be the countries with the least independent judiciaries of the 139 nations surveyed (*Global Competitiveness Report 2010–2011*). These rankings suggest a disparate reality within the region when it comes to judicial autonomy from the government; in some countries the judiciary is able to do its work with a reasonable degree of autonomy from the executive branch, but in others the judiciary is completely dominated by the executive.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86136-X

Something similar happens with the degree of corruption exhibited by the different countries of the region – a proxy for adherence to legality by political authorities – as reported by the organization 'Transparency International.' According to this organization's *Annual Report 2012*, Latin-American states are distributed within a wide range in terms of perceived corruption, with Uruguay and Chile ranking among the 20 most transparent states while Venezuela and Haiti rank among the 10 most corrupt.

The indicators just mentioned suggest that the gap between 'law in the books' and 'law in action' also varies widely in Latin America. The same goes for the prevalence of legal pluralism. Concerning the latter, and given the fact that there is quite some variation in the percentage of indigenous peoples living in the different Latin-American states (from over 50% of the population in Bolivia, to less than 1% of the population in Uruguay), it is not surprising that the degree to which indigenous law coexists with formal state law also varies immensely. In Bolivia, Guatemala, and Ecuador, for example, the role of indigenous law is extremely important, while in Uruguay, Costa Rica, and Chile it is more limited.

III

In the previous sections we have seen that – in their formative years – most Latin-American countries embraced codification as a key feature of their state-building efforts. The goal was to affirm national identity, as well as to introduce clarity and coherence to Private Law (Mirow, 2004). This was seen as a necessary step to enhance commercial transactions and to better regulate family relations, given the chaotic mass of colonial law, local regulations, and informal practices prevalent at the time of independence from imperial rule. The codification of Civil Law – which was done in most countries of the region during the last decades of the nineteenth century – was followed by the introduction of codes of Civil Procedure, codes organizing the judicial branch, and codes of Criminal Law and Criminal Procedure.

In the case of Public Law, the adoption of written constitutions was done much earlier, that is to say, soon after independence from imperial rule. As stated above, the model for most of the region's Constitutional Law was the United States, given that it was the only nonmonarchical country in existence at the time. Having said this, the chronic political instability experienced by most Latin-American states at the time meant that existing constitutions were constantly replaced by new ones – or were profoundly amended – as a result of revolutions, civil wars and other types of generalized violence. This explains the astonishing number of constitutions that Latin America has had over the last two centuries (Elkins et al., 2009). The high number of constitutions has not, however, always translated into limited government, due to the heavy use of 'states of emergence' throughout the region's history (Loveman, 1993). Even though the chronic reality of a proliferation of formal constitutional charters and a lack of substantial constitutional rule of law just noted was the norm during the nineteenth century, a few countries eventually managed to establish relatively stable and meaningful constitutional regimes (such as Chile and Argentina, at the turn of the century).

Observers of Latin-American Public Law attribute many nations' difficulties in consolidating stable and effective constitutional regimes to a number of factors, such as geographic and ethnic fragmentation, a profoundly unequal social structure and the lack of a 'culture of rights' (Rosenn, 1990; Schor, 2006). Others argue that the root cause of the lack of a strong constitutional culture in the region is the persistence of a long colonial legacy characterized by a centralized, authoritarian, and hierarchical exercise of political authority (Adelman, 1999).

In spite of the problems just noted, toward the middle of the twentieth century a sizable number of states in the region managed to consolidate constitutional regimes that were gradually introducing a democratic rule (this was the case in Argentina, Brazil, Chile, Colombia, Peru, Venezuela, and Uruguay). Parallel to this process it should be noted that, starting with the Mexican Constitution of 1917, many countries of the region adopted what is known as 'social constitutionalism,' that is to say, they introduced into their constitutional charters various social rights, such as the right to work, education, health care, housing, and so on.

In the decades after World War II, the trend toward the inclusion of new rights continued, as many Latin-American states ratified a large set of both universal and regional human rights treaties; these would later exert an important influence in domestic constitutional law (Franklin, 2008). The introduction (in the 1970s) of a regional system for the protection of fundamental rights, the Inter-American Human Rights Court and Commission, represented a crucial development in the Latin-American legal systems, since it provided a regional arena where individuals can seek judicial remedies when their own states have violated their human rights and the domestic courts have failed to protect them (Shaver, 2010).

Even a brief account of the Latin-American legal systems has to provide an account of the emergence – during the first decade of the current century – of what has been labeled 'radical constitutionalism.' This emerged in the so-called 'Bolivarian republics' of Venezuela, Bolivia, and Ecuador, as a result of popular frustration with the inability of the previous constitutional regimes to make decisive progress toward reducing poverty and inequality, as well as their negligence in controlling corruption (Couso, 2014). The lack of legitimacy of liberal constitutionalism in the above-mentioned countries was seized upon by a group of highly charismatic leaders (Hugo Chavez, in Venezuela; Evo Morales, in Bolivia; and Rafael Correa, in Ecuador) who, after gaining control of the executive and legislative branches, initiated an aggressive program of constitutional reform explicitly aimed at the political and social 'rebirth' of their respective nations. Given that the constitutional charters in place at the time did not contemplate referendums to ask the people about the need to adopt a new constitutional order, they used executive decrees calling for these referendums to call for constitutional assemblies that eventually passed new fundamental charters. These radical constitutional experiences purposefully weakened the principle of separation of powers (especially to the detriment of the autonomy of the judiciary); expanded the power of the executive branch; and – in the case of Venezuela and Ecuador – substituted bicameralism for unicameralism in the legislative branch, all transformations based on the notion that a 'unity' of

powers was needed to deliver radical social change (Viciano and Martínez Dalmau, 2010). The actual impact of this type of radical constitutionalism is yet to be seen, but in the view of some observers it has already done much to weaken the rule of law in the 'Bolivarian republics.'

IV

The adoption by most countries of the region of (an ever expanding) International Human Rights Law, as well as the Inter-American Human Rights system, have been important sources of transformation of the very conception of law in the region, leading to the weakening of the legal formalism characteristic of Latin-American legal systems and the development of a more 'justice-oriented' conception of law. Indeed, as López Medina (2004) has suggested, the processes noted above, coupled with the influence of a human rights–based understanding of law typical of postwar Continental European legal doctrine (Shapiro and Stone Sweet, 2002) have led to the increasing abandonment by Latin-American legal actors of the legal positivism advocated by Hans Kelsen that was so dominant in this region for most of the twentieth century. This change in the conception of law and legality has also provoked a change in the understanding of the role of the courts, from one in which they were largely deferent to legislative sovereignty, to one which openly embraces the judicial review of the constitutionality of legislation, a doctrine which has in turn led to growing degrees of judicialization of politics (Couso et al., 2010; Sieder et al., 2005).

The developments mentioned above have profoundly transformed the role played by the courts in many of the legal systems of Latin America (particularly in Colombia, Costa Rica, Argentina, and, to a lesser extent, in Brazil, Mexico, Chile, and Peru), a process which in turn has changed the expectations of legal academics, lawyers, and regular citizens about the role that law and courts should play in the social, political, and economic life of their nations. Another consequence of the increasing power of Latin-American courts and, in particular, of the incipient development of judge-made law has been the introduction in some countries of mild forms of *stare decisis* (the doctrine of precedent), something that suggests a degree of convergence with the Common Law tradition.

The convergence just noted has also been facilitated by the large number of legal academics and practicing lawyers who – starting in the 1970s – have received graduate legal education in the United States, the United Kingdom, and other former British colonies. Thus, after a century in which Latin-American legal scholars and practitioners almost exclusively attended Continental European Law schools for postgraduate legal education (French and Italian ones, for legal training in Private Law; German ones, for legal training in Criminal Law), in the last decades of the twentieth century growing numbers of them started to attend LL.M.s and S.J.D.s in Anglo-American institutions (Dezalay and Garth, 2002).

V

The legal actors of the Latin-American legal systems have changed significantly over the last decades, due in part to the developments described in the preceding section. The most relevant transformation in this domain is the growing centrality of judges vis-à-vis legal academics, a process that would have been unthinkable earlier in Latin America's legal history, where the latter dominated the articulation of a legal discourse then invoked by practitioners and judges. Of course, given the significant variations already noted among the different countries of the region, the degree to which judges have become key actors in the development of law and legal doctrine through their decisions varies widely. Thus, for example, in Argentina and Colombia this trend is most noticeable, while in Uruguay and Peru legal academics continue to overshadow judges.

The growing relevance of judges in Latin America is also the result of large programs of judicial reform backed by the World Bank and other international organizations, within the context of transitions to democracy and economic globalization. These programs, which introduced in many Latin-American countries 'councils of the magistracy,' have not, however, always been successful (Prillaman, 2000).

Another area of the Latin-American legal field where one can observe significant change is in legal education, which has become more diversified, professionalized, and globalized in many states of the region (Pérez-Perdomo and Gómez, 2008). Indeed, as opposed to the classical legal education performed by successful legal practitioners who would teach for a few hours a week using the traditional model of formal lectures based on a close exegesis of codified law and a few treatises, a number of Latin-American countries have managed to introduce a truly professional legal education, with specialized law professors fully dedicated to teaching and researching (this is particularly the case in Colombia, Chile, Brazil, Mexico, and Argentina). Although it is still too soon to assess the impact that this dramatic change in legal education will have in the said countries, there is already evidence that the trend throughout the region is toward this professionalized model.

Related to changes in legal education is the introduction in recent decades of programs for the training of judges, both as a requisite for admission to judgeship and for the judges' continuing education. One important example at hand is Chile's Judicial Academy, introduced in the 1990s (Skaar, 2010), and based on the French model.

VI

As this article shows, Latin-American legal systems exhibit some common features but also have important differences. The similarities are the result of three centuries of common imperial history that left an important legacy in the social, political, and legal structures of the countries of the region, as well as in their cultures. The differences are the result of the distinct political, economic, and constitutional paths followed by each of the 20 Latin-American states over the last two centuries.

Furthermore, it is important to highlight the convergence with Anglo-American law that in recent decades can be detected in almost all Latin-American legal systems, due to both economic and legal globalization (Friedman and Pérez-Perdomo, 2003). This convergence is now apparent in both Public and Private Law. In the case of the former, it is expressed

in the fact that many Latin-American constitutional and High Courts often cite decisions by the United States' Supreme Court, while in the case of the latter, it is expressed by the adoption of contract, tort, and family law doctrines and judicial decisions developed in the Anglo-American legal world. Finally, convergence has been heavily driven by the growing importance of International Human Rights Law in most Latin-American legal systems.

See also: Civil Law; Constitutional Courts; Indigenous Rights; International Law and Treaties; Judges; Legal Formalism; Legal Positivism; Postcolonial Law.

Bibliography

Adelman, Jeremy (Ed.), 1999. Colonial Legacies: The Problem of Persistence in Latin. American History. Routledge.

Couso, Javier, Hone's, Alexandra, Sieder, Rachel (Eds.), 2010. Cultures of Legality: Judicialization and Political Activism in Latin America. Cambridge University Press.

Couso, Javier, 2014. Radical Democracy and the "New Latin American Constitutionalism". SELA, Yale Law School.

Dezalay, Yves, Garth, Bryant G., 2002. The Internationalization of Palace Wars. Lawyers, Economists, and the Contest to Transform Latin American States. University of Chicago Press.

Elkins, Zachary, Ginsburg, Tom, Melton, James, 2009. The Endurance of National Constitutions. Cambridge University Press.

Franklin, James, 2008. Shame on you: the impact of human rights criticism on political repression in Latin America. International Studies Quarterly 52 (1), 187–211.

Friedman, Lawrence, Pérez-Perdomo, Rogelio (Eds.), 2003. Legal Culture in the Age of Globalization: Latin America and Latin Europe. Stanford University Press.

García Villegas, Mauricio, 2009. Normas de papel. Cultura del incumplimiento de reglas. Siglo del Hombre, Dejusticia.

Gargarella, Robert, 2013. Latin American Constitutionalism, 1810–2010. The Engine Room of the Constitution. Oxford University Press.

López-Alves, Fernando, 2000. State Formation and Democracy in Latin America, 1810–1900. Duke University Press.

López Medina, Diego, 2004. Teoría Impura del Derecho. La transformación de la cultura jurídica latinoamericana. Legis-Universidad de Los Andes-Universidad Nacional de Colombia.

Loveman, Brian, 1993. The Constitution of Tyranny: Regimes of Exception in Spanish America. University of Pittsburgh Press.

Merryman, John H., 1969. The Civil Law Tradition: An Introduction to the Legal Systems of Europe and Latin America. Stanford University Press.

Mirow, Matthew, 2004. Latin American Law: A History of Private Law and Institutions in Spanish America. University of Texas Press.

Nino, Carlos, 1992. Un país al margen de la ley: estudio de la anomia como componente del subdesarrollo argentino. Emece.

Pérez-Perdomo, Rogelio, Gómez, Manuel, 2008. Innovaciones en la educación jurídica latinoamericana. Derecho y Democracia: Cuadernos Unimetanos 15, 4, 5.

Prillaman, William, 2000. The Judiciary and Democratic Decay in Latin America. Declining Confidence in the Rule of Law. Praeger.

Rosenn, Keith, 1990. The success of constitutionalism in the United States and its failure in Latin America: an explanation. University of Miami Inter-American Law Review 22, 1–39.

Schor, Miguel, 2006. Constitutionalism through the looking glass of Latin America. Texas International Law Journal 41, 1.

Shapiro, Martin, Stone Sweet, Alec, 2002. On Law, Politics, and Judicialization. Oxford University Press.

Shaver, Lea, 2010. The Inter-American human rights system: an effective institution for regional rights protection. Washington University Global Studies Law Review 9, 639.

Sieder, Rachel, Schjolden, Line, Angell, Alan (Eds.), 2005. The Judicialization of Politics in Latin America. Palgrave MacMillan.

Sieder, Rachel (Ed.), 2002. Multiculturalism in Latin America: Indigenous Rights, Diversity and Democracy. Palgrave Press.

Skaar, Elin, October 2010. Un análisis de las reformas judiciales de Argentina, Chile y Uruguay. América Latina Hoy 34, 147–186.

Transparency International, Annual Report 2012.

Viciano, Roberto, Martínez Dalmau, Rubén, 2010. Aspectos Generales del Nuevo Constitucionalismo Latinoamericano. In: El nuevo constitucionalismo en América Latina. Corte Constitucional del Ecuador.

World Economic Forum, Global Competitiveness Report 2010–2011.

Latour, Bruno (1947–)

Lisa Garforth, School of Geography, Politics and Sociology, Newcastle upon Tyne, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Beginning by presenting his 2005 book *Reassembling the Social* as a key text for social scientists, this article introduces Latour's social material philosophy and exploration of a reality consisting of actants, networks, and lively relationships. It puts the metaphysical Latour side-by-side with the one interested in questions of environment, politics, and practical methods. It aims to pull diverse elements of Latour's writings and associations together into a useful summary, while remaining faithful to his theoretical commitments to heterogeneity, translation, mutation, and hybridity.

Introduction: Bruno Latour Reassembles the Social

There is something slightly uncomfortable about summarizing the work of Bruno Latour for an encyclopedia of the social sciences. Latour suggests in his book *Reassembling the Social* (2005) that there are only 'two drawbacks' to be found in the idea of a 'science of the social': 'the word 'social,' and the word 'science' (2005: 2). This combative relationship with the disciplines closest to home for a supposed sociologist and anthropologist is an integral part of Latour's project. He seeks a new route out of the contradictions and epistemological impasses of modern thought, especially its desire to separate out purified entities instead of engaging directly with hybridity and connectedness. One part of this project has involved taking apart the idea of a homogeneous and authoritative science with special access to the real world. He shows instead how sciences in the plural are grounded, situated practices that build up their claims out of multiple relationships with different kinds of materials and methods. Another part of his project, more unsettling for social scientists, challenges the notion that society is a distinctive entity, separate from nature and from technological artifacts. Instead, Latour leaves us with a motley collective of humans and nonhumans, of social, technological, and natural issues all mixed up together, and he offers us the instruction that we must learn anew, every time, the specific and contingent ways in which entities associate together.

For some in the human and social sciences, especially in the multidisciplinary field of science studies, Latour's provocative approach has been invigorating and transformative. His work has also been taken up by philosophers trying to move beyond some of the dead ends of postmodernism/post-structuralism and toward frameworks foregrounding lively materialities (Bennett, 2010; Harman, 2009). Within the mainstream of the social sciences, however, Latour's work has often been met with indifference and sometimes hostility. This is a loss, I think, especially to social theory. His work goes to the very heart of what we mean when we talk about society, and it stimulates us to think in new ways about the most fundamental logics and dynamics of social action. He gives us a rigorous alternative to the structure–agency dyad and asks us to question how much we cling, often despite ourselves, to some version of a society *sui generis* to ground social inquiries. He gives us ways of exploring social stability and social change in terms of contingent, constructed networks, and ways of rethinking social action to include lively nonhuman actors. As well as trying faithfully to describe the hybrid worlds we inhabit, he has developed hopeful and robustly realistic proposals about how we can find better ways of living in and with them. Latour has a lot to offer social science, even if he does not like the label.

I begin here with *Reassembling The Social* because it speaks most directly of all Latour's writing to and about social scientists. It is not always a comfortable conversation, but it is an important one. It asks us to reimagine the stuff of our subject matter in terms of diverse and open-ended associations between human, technological, and natural entities, across which the lively capacity for action and the objective capacity to endure are distributed in surprising ways. *Reassembling* offers social scientists a way in to Latour's wider project, which as we will see is simultaneously ontological, methodological, and political. Across his extensive writings he weaves and reweaves some core propositions. There is the refusal of binary categories (nature/society, objective/subjective, facts/values) in favor of a rich picture of hybrids, networks, and relations. There is the insistence that nothing can be reduced to anything else. There is a rejection of the idea that some important things (identities, actions, ideas) are 'socially constructed' while other important things (objects, physical forces, nature) are external and separate from us. *Reassembling* very clearly links these ontological themes back to questions of methods and how we find out about the modes of association that make up the messy worlds we inhabit.

Who Is Latour?

Two recent full-length studies take different approaches to Latour's plural intellectual stances and unusual academic biography (Blok and Jensen, 2011; Harman, 2009). Neither fully resolves the question of his disciplinary identity and I do not propose to do so here. Instead, I start with an overview of his current professional affiliation and projects and then work backward to indicate some of the contexts, alliances, collaborations, and institutions through which he has forged his ideas.

Latour is currently Professor and director of the Theory of Actor-network and Research in Digital Environments (TARDE) program at Sciences Po (Institut d'études Politiques de Paris) in Paris, where he founded the médialab research center for digital

social sciences and is director of the innovative interdisciplinary Experimentation in Arts and Politics (SPEAP) program. Both ventures speak to his growing interests in the intersections between disciplines and the connections between different forms of inquiry. In 2010, he was awarded a European Research Council Grant to pursue new work on an 'anthropology of the moderns'. This has recently appeared in the form of a book, *An Inquiry into Modes of Existence* (2013), and a collaborative Web site (AIME). They explore (and invite collaborators to explore) the plurality of worlds or realities that in different ways organize how we live, what we know, and who we are. In law, religion, science and aesthetics we can discover different sets of conditions for deciding on what is true and false, good and bad, what matters and what is unimportant. These are conditions that can only be seen in all their strangeness if we accept that we were never really modern, that is, that there was never a stable objective world out there and that human agency alone does not make history. The book develops concerns that were present in some of Latour's earliest writing (*Irreductions*, available in English in print as an appendix to Latour (1988); see Tresch (2013) and Harman (2009)), and revitalizes the philosophical and theological studies he pursued as an undergraduate at the University of Bourgogne in Dijon and later at Tours, where he received his doctorate in 1975.

In 2013, Latour won the Holbert Prize for outstanding scholarly work in the humanities and social sciences. Here he was identified as a sociologist and anthropologist, linking him back to the disciplinary affiliations of the earlier parts of his career. Between 1982 and 2006, Latour was based at the Centre de Sociologie de l'Innovation at the Ecole Nationale Supérieure des Mines, Paris. During this period, he published some of his best known books, including his intervention in the philosophy of modernity, *We Have Never Been Modern* (1993); his reflections on science studies, the 'science wars' and the limits of social constructionism, *Pandora's Hope* (1999); his argument for a sociology of associations, *Reassembling the Social* (2005); and his proposal for new political institutions fit to deal with nature and environmental crisis, *Politics of Nature* (2004). With Callon, Law, and others he developed under the eventual rubric of actor-network-theory (ANT) some new approaches to explaining how scientific and technical things are made and circulated. Against the grain of dominant models in the philosophy of science, Latour argued that scientific claims, like technical objects, are not discovered but rather practically and painstakingly constructed. Once a fact or a technological object starts to hold together, its path to legitimacy and acceptance is not about people recognizing its necessary or self-evident truth. Rather it is about making the fact or object more widely relevant or useful by extending the contexts in which it is relevant or the network within which it makes sense.

These arguments emerged from Latour's work in the field of sociology of science, starting with his first published book, *Laboratory Life* (Latour and Woolgar, 1986). Its findings were rooted in an ethnographic study of Roger Guillemin's neuroendocrinology laboratory group at the Salk Institute for the Biological Sciences in La Jolla, California, where Latour spent two postdoctoral years (1975–77). *Laboratory Life* sends back reports from the everyday life of scientists. It explains the enormous effort and the great diversity of actors (texts, machines, and biological things, as well as scientists) that are needed to translate material substances into written traces, and thereby to stake a scientific claim to have discovered the truth about a bit of the physical world. In his later work, *Science in Action* (1987), Latour clarifies his agonistic model of the construction of scientific facts, that is, the combative and contesting nature of scientific practice. He also shows how expensive it is, in terms of both money and effort, to contest and challenge a scientific claim once it has traveled widely, been stripped of its 'modalities' or contingencies, and been reduced by usage simply to its inputs and outputs.

Latour often claims to be working on a 'practical metaphysics' (2005: 50). His varied academic biography shows how this claim is rooted in a career made up of different disciplinary orientations, multiple empirical encounters, and diverse methods of inquiry. He has been a laboratory ethnographer, peering over the shoulders of endocrinologists, and a philosopher of knowledge, making general claims about the nonexistence of external reality. He has been the historian of a modernity he claims did not happen, and an analyst of the future institutions that might help to usher in a new politics of nature. He has studied Biblical exegesis and worked in international development in Cote d'Ivoire (during civilian national service in the 1970s; see Tresch, 2013; Harman, 2009). Across all these engagements, he now says, he "never had any other interest than the exploration of the many ways there are to find the truth of a situation" (Holberg, 2013).

Metaphysics and Methods: from Culture and Nature to ANT and Ants

The multiplicity of truth. As the quote above might suggest, the Latour currently taking center stage is primarily located in relation to metaphysical issues, that is, questions of "the basic structure of the world" (Latour, 2005: 51). In an article prefacing his most recent book, Latour narrated his "coming out" as a philosopher (2010). In 2009, Harman's *Prince of Networks* (2009) read Latour as a neglected metaphysicist. Latour wears the mantle well. His work pokes from different directions at how humans and their technologies manage to produce workable truths about the fundamental nature of existence. Metaphysics has something to do with what seems to exceed, transcend, or precede the physical, and hence becomes deeply troublesome and deeply interesting after the emergence of modern science, issues that Latour explored since *We Have Never Been Modern* (1993). And if metaphysics now sometimes seems to be a leftover category for aspects of philosophy that do not fit neatly into the boxes of ethics, epistemology, aesthetics or logic (Van Inwagen, 2013: pt 4), then perhaps it is a good home for Latour's commitment to thinking with and about hybrids, quasi-objects, collectives, and heterogeneous cosmologies.

To start with the metaphysical Latour is to start with his arguments about what the world is like. It is to encounter a vivid ontology in which all things are thoroughly lively and relational. Latour's ontology is best apprehended as a kind of monism (Bennett, 2010) which attempts to understand reality in terms of a single flat plane rather than starting with distinctions between types of things. It is especially hostile to binaries and dualisms. In Latour's cosmology there is no nature

separate from culture. Society is not an object *sui generis*. The capacity for autonomous agency is not restricted to humans or compared to objective things and forces. Instead we find a plurality of concrete, irreducible entities, or 'actants.' Actants are constituted through their connections. Individual people *are* networks, made up of material substances and physical processes; of thoughts and self-awareness; of the things we take into ourselves and the arrangements and others outside us who make us what we are. People *make* alliances too, by persuading others, both human and nonhuman, to lend their interests to building and stabilizing a set of connections that can achieve something new or make something happen: in the language of actor-networks, through acts of enrollment or translation.

All entities, human and nonhuman, are lively and agential if they have the capacity to affect and effect other things. Events and identities are not reducible to single causes or determinants, but are contingent outcomes of an assembly of forces and connections, more or less temporary congealings or stabilizations of diverse and dynamic forces. If things seem to run on familiar tracks, if objects, institutions, and actors appear durable and stable, then the networks are bedded down and largely invisible. This does not mean, however, that the things we treat as real and reliable in our world do not depend on ongoing work to maintain them. If they are to matter, to remain stable and effective and meaningful, all networks involve work – that is, active doing, commitment, activity, and effort. Latour then proposes a kind of relational realism in which connection is the substance of the world. It is also a dynamic kind of realism. In order to be and to matter, things have to be practiced or performed. For Latour, "an actant is always an event" (Harman, 2009: 17). The more connected a thing is and the more work that goes into making it up, the more constructed it is. And the more connected and well-constructed a thing is, the more real and robust it is (Latour, 2005). Here we find some of the most interesting and also the most contentious elements of Latour's thinking. Reality is not the opposite but rather the outcome of construction and relatedness. This approach challenges the assumptions of (scientific) realism as well as rejecting the implications of (social) constructionism. As a result it has been the object of ferocious critique from both sides (Latour, 1999).

Both approaches are wrong, Latour argues, because of a fundamental historical misstep – the emergence of the self-conscious moderns and their powerful but deluded model for understanding the world. In *We Have Never Been Modern* (1993), Latour explains that the 'modern Constitution' involves two opposing dynamics. A rigorous conceptual process of purification divides the world into separate realms of nature and culture, divides thought into contrasting epistemologies relating to either objective facts or subjective meanings, divides things and materiality from beings and agency. But the very impetus of modernity is to mix things up. Processes of hybridization and the making of imbroglios are the practical lifeblood of the moderns. From the backroom experiments of the earliest natural philosophers to the most elaborate industrial enterprises, from the discovery of oxygen to the manufacture of automobiles, our social life constantly articulates human interests, technological apparatus, natural resources, and practical know-how. Purification, according to Latour, is a smoke screen. By refusing to recognize its own monsters and mixtures, modernity in fact allows them to be produced. But if modernity rested on the apparent human capacity to stand apart from and objectively know nature, then it never really was. We have never been modern.

Here we see very clearly the metaphysical Latour and can begin to understand why this work has been important and innovative. But this metaphysical Latour, however interesting, perhaps threatens to overshadow equally compelling aspects of his work. Firstly, there is the ecological Latour who has engaged since *Pandora's Hope* in 1999 with how we nonmoderns can more honestly describe the hybrid techno-collectives we have made, revise our methods for living with environmental imbroglios, and devise a 'parliament of things' for managing the natural-social collective (Latour, 1999; see Chapter 6). *Politics of Nature* (2004) takes up these questions with gusto, developing a new focus on ecological philosophy and politics that continues to drive much of Latour's work (2013). Here he argues that environmentalism took a wrong turn when it tried to save nature or bring it into politics. Both moves simply reiterate a division between human values and natural objects that does not exist. Environmental crisis, Latour tells us, is a way of noting that many lively hybrid objects are insisting on recognition, and we cannot answer to them from within a modern constitution that keeps facts and values separate. The book goes on to describe the logics of a new constitution and new skills that could help us to better recognize and manage all the elements of our cosmos. Here Latour outlines a cosmopolitics (the term comes from the philosopher Isabelle Stengers; see Stengers 2010) that is unabashed in its desire to find better ways of living with the worlds we create.

Secondly, the metaphysical Latour threatens to overshadow the Latour of methods. This Latour emerges in the work on science studies with the lab ethnographer watching bioscientists at work in *Laboratory Life*. This methodological focus is developed in *Science in Action*, which formalizes 'rules of method' (1987: 258–259) to help interested analysts learn to 'follow scientists and engineers around,' watch 'science in the making,' and observe how the disorderly mixtures of the world are turned into orderly patterns and artifacts (1987: 15). Latour's capacity to work fluidly across the empirical and the conceptual, to link metaphysics and methods, is arguably the most distinctive aspect of his project. Working sometimes from concrete empirical observations in the midst of the action, sometimes reconstructing case studies from the history of science, the emphasis is on accurately describing how people get on with the task at hand in particular settings. Latour's laboratory anthropologies made a dazzling methodological contribution to science and technology studies, helping to push the field towards the observation of practice and what scientists do, and away from the abstract certainties of the philosophy of science and the retrospective constructions of what scientists say.

Later, the Latour of *Reassembling* sought to set out new methods for a 'sociology of association' and for a descriptive social science more generally. Although its subtitle is 'An Introduction to Actor-Network-Theory,' the book explicitly asks to be read as a 'travel guide' or a 'how to book' (2006: 17). If society does not exist as a discrete entity that preexists social things (action, identity, processes, change), then it cannot account of itself for anything. For Latour there are 'no powerful

agents hidden in the dark' that can explain things (2004: 228). We will have to identify new tools and tactics to find out how the social gets put together. Latour argues that this means describing all the elements and tracing all the connections in a network or assemblage, starting with the actants involved in it. This means new methods and even a new job description for social scientists. We should not see ourselves as combative critics who reveal underlying structures and forces, or as clever interpreters of peoples' meanings. Turning the lofty ontological claims associated with ANT into something more modest and humble, Latour instead invites social scientists to see themselves as ants: "blind, myopic, workaholic, trail-sniffing, and collective traveler[s]" (2005: 9). As ants we would travel very slowly across and through the terrain of a network, noting all the connections, taking no shortcuts. Real knowledge rests on faithful attention to what is there and what all the actors (human and nonhuman) are doing. Whatever we think of Latour's philosophical and ontological pronouncements, his methodological prescriptions are always rigorous and often surprising. And he reminds us that the currency of our own methods, like those in any actor-network, is effort and work, not unique epistemological insights or special cognitive properties.

Following Latour Around: Alliances, Impacts and Further Travels

Latour's influences combine the philosophical with the methodological. He has drawn productively on Whitehead's pragmatic philosophy and insistence on the primacy of process and event, as well as on Deleuze's invocation of a flat world composed of multiple assemblages. Latour's conversations with the philosopher Michel Serres (Latour and Serres, 1995) draw out a shared concern with how translations are made between different kinds of knowledge and actions, and a particular interest in 'quasi-objects' and other in-between entities. His work draws on ethnomethodological sociology to develop a focus on practical action and lay reasoning (especially the work of Harold Garfinkel), and his laboratory studies drew him into an engagement with ethnographic and participant observation methods in anthropology. Beyond this, however, Latour is mainly critical of mainstream sociology. He is dismissive of the French positivist tradition founded by Emile Durkheim, and frequently suggests that sociology would have done better to adopt Durkheim's contemporary Gabriel Tarde as an alternative founding father. He is also opposed to critical theory, Marxist sociology, and some strands of postmodernism, although there are parallels with post-structuralism and especially Foucault throughout his writing (for an interesting account of this see Kendall and Wickham, 1999). He has a complex relationship with Pierre Bourdieu, whose early emphasis on exploring a theory of practice and ethnographic approach shares much with Latour's project, but whose later structuralist and critical turn is relentlessly criticized (2005).

As we have seen, Latour is closely associated with ANT or actor-network-theory. This approach emerged in science studies in the 1980s to account for things – entities, relationships, institutions – in terms of their connections. John Law (1999: 3) has said that ANT can be understood as the "ruthless" application of the relational and antiessentialist logics of semiotics beyond language to the material sphere (Law, 1993: 3). ANT highlights the performative nature of reality, showing how what appears durable, patterned, and stable is rather the contingent outcome of constant acts of making and remaking. ANT is probably best understood not as fixed theoretical position but as a loose assemblage of related concerns and assertions. It has indeed reassembled itself repeatedly, not least through critical reflection on its own precepts (Latour, 1999; Law, 1999). Latour himself has been very active in ensuring that ANT does not ossify into a framework of abstract concepts. He has half jokingly 'recalled' the label (Latour, 1999), both taking it back and suggesting the alternative name actant rhizome ontology. However, as we have seen, Latour then reclaimed the name and the method, at the same time puncturing some of its more grandiose and obscure connotations. Thus ANT continues to trouble our ontological assumptions at the same time as Latour asserts that it is nothing more than adopting the humble and modest methods of ants.

Latour has recently said that he was overwhelmed by the 'after-sale service' required in the light of the success and notoriety of ANT (Tresch, 2013: 303). This involved addressing critiques from both scientific realists and social constructionists in relation to his relational realist approach. The former group, fighting what have become known as the 'science wars,' allied Latour with deconstructionism and antifoundationalism, and accused him of an absurd rejection of the real (Sokal and Bricmont, 1998). At the same time, Latour's insistence that objectivity concerns precisely how things object, and his willingness to treat nonhuman actants symmetrically with human ones, seemed to many sociologists of scientific knowledge a step too far toward granting agency to unreflexive, unintending entities (Collins and Yearley, 1992). But all the while Latour and ANT were being avidly taken up by researchers within science and technology studies and beyond. In STS (science and technology studies) Latour and ANT reshaped the field, offering new ways of thinking about knowledge as real *because* it is constructed and new ways of studying practical action in specific sites. ANT has been attractive to geographers interested in rethinking the physical/social divide and in the spatial implications of network approaches. ANT is also increasingly present in organization and management studies, where its focus on the active work of ordering and organizing multiple actors in complex chains and networks offers an alternative to approaches based on structures and levels.

Conclusion

Latour's project moves between ontological questioning, empirical analysis, situated methods, and political speculation. He gives us radically new ways of thinking about a mixed-up world that are becoming more insistently relevant as Western societies begin the process of facing climate change. He shows how we 'nonmoderns' have inherited an amazing capacity to generate novelty and hybrids and an equally startling failure to recognise and take responsibility for them (2004; 2013). In the

face of a new reckoning with our transformed 'natural' environments, Latour is right to argue that this denial cannot go on. Philosophically, he gives us a real alternative to the circular nature of postmodern debates about epistemology and signification and a chance to rethink life and materiality. Methodologically, he pushes us as social scientists to be more faithful to what we can observe in the world, but to always be on the alert for the "slight surprise of action" and the need to recognise what is genuinely new in our world. (1999: 266). It is not quite clear that the social sciences have recognized it yet, but Latour's intervention in our worlds and our ways of thinking about them is already making transformative associations across our fields.

For those interested in following Latour's ideas back to the texts they came from, I offer here a highly selective reading guide. One place to begin is with the important trilogy of books on scientific practice, science studies and knowledge politics that span roughly the first twenty years of Latour's career. They are: *Laboratory Life* (Latour and Woolgar 1986, first published 1979); *Science in Action* (1987), and *Pandora's Hope* (1999).

Then there are three important philosophical and political texts produced at roughly 10-year intervals investigating the wrong turn taken by the modern constitution and its implications for social science and environmentalism. They are: *We Have Never Been Modern* (1993), *Politics of Nature* (2004), and *An Inquiry into Modes of Existence* (2013).

Looking only at the major books tends to focus on big theoretical claims at the expense of a sense of how often they emerge out of intense and generative engagements with very particular objects, people, and case studies. Here the reader might be interested in Latour's compelling exploration of Pasteur and his microbes in *The Pasteurization of France* (1988), or his examination of a guided transport system that never got built in Paris in *Aramis, or, the Love of Technology* (1996). It also misses Latour's collaborations with others through cowriting, conversation and contributions to edited collections. With Peter Weibel, Latour edited a wide-ranging and multifaceted collection of essays on how things become objects of political and civic concern, *Making Things Public: Atmospheres of Democracy* (2008); and we have already mentioned his conversations with Serres in their 1994 collection.

And it misses Latour's collaborations with others through cowriting, conversation and contributions to edited collections:

Focusing too narrowly on Latour's books can also obscure the fact that for many readers, especially in STS, Latour is best known as an essayist. His shorter pieces serve as excellent introductions. Latour's style and his interest in concrete objects and issues suit the form exceptionally well. He is adept at taking an idea and a reader for a rhetorical walk, provoking and entertaining along the way, and returning us to where we thought we had started only to find that we are somewhere else entirely. Here I want to mention three examples, spanning some twenty years.

The essay 'Give Me a Laboratory and I Will Move the World' (1983) takes apart the idea that scientific knowledge is either the outcome of neutral and pure methods or a product of social or political interests. Instead, it suggests, we might try seeing sociology in the very content and doing of science, and recognize microobjects in the lab as brand new sources of large-scale power.

In 'Where are the Missing Masses?' (1992), Latour dwells on mundane objects, including a seat belt and a 'Berliner key.' We delegate to them not only practical tasks but also morality – these objects keep us safe and maintain social borders. Latour argues that humans are neither wholly in control of our technologies nor dominated by the object(ive) worlds we create. Rather, we are cocreators and coinhabitants of a human/nonhuman collective. Ought not our social theory treat the missing masses more like we treat people?

In 'Why Critique Has Run Out of Steam,' Latour asks the social sciences to leave behind the deconstructive and debunking epistemology that has dominated from critical Marxism to postmodernism – without adopting its apparent opposite, a naïve realism concerned with external matters of fact. Instead we should deal with "matters of concern" and cultivate a "stubbornly realist attitude" (2004: 232) that is at the same time capable of dealing creatively with a constructed world.

See also: Environmental Movements; Foucault, Michel (1926–84); Garfinkel, Harold (1917–2011); Modernization and Modernity in History; Objectivity: Philosophical Aspects; Postmodernism: Philosophical Aspects; Practical Reasoning: Philosophical Aspects; Relativism: Philosophical Aspects; Science, History of; Social Constructivism.

Bibliography

Bennett, J., 2010. Vibrant Matter: A Political Ecology of Things. Duke University Press, Durham, NC.

Blok, A., Jensen, T.E., 2011. Bruno Latour: Hybrid Thoughts in a Hybrid World. Routledge, London.

Collins, H.M., Yearley, S., 1992. Epistemological Chicken. In: Pickering, A. (Ed.), Science as Practice and Culture. Chicago University Press, Chicago, pp. 301–326.

Harman, G., 2009. Prince of Networks: Bruno Latour and Metaphysics. Re-press, Melbourne.

Holberg Prize Website, 2013. To Find the Truth of a Situation: An Interview with Bruno Latour. http://holbergprisen.no/en/bruno-latour/interview.html. (accessed 10.11.13.).

Kendall, G., Wickham, G., 1999. Using Foucault's Methods. Sage, London.

Latour, B., 1983. Give me a laboratory and I will move the world. In: Knorr, K., Mulkay, M. (Eds.), Science Observed. Sage, London, pp. 141–170.

Latour, B., 1987. Science in Action: How to Follow Scientists and Engineers through Society. Harvard University Press, Cambridge, MA.

Latour, B., 1988. The Pasteurization of France. Harvard University Press, Cambridge, MA (Originally published in French 1984).

Latour, B., 1992. Where are the missing masses: the sociology of a Few Mundane Artifacts. In: Bijker, W.E., Law, J. (Eds.), Shaping Technology/Building Society: Studies in Sociotechnical Change. MIT Press, pp. 225–258.

Latour, B., 1993. We Have Never Been Modern (C. Porter, Trans.). Harvard University Press, Cambridge, MA.. First published 1991.

Latour, B., 1996. Aramis, or the Love of Technology. Harvard University Press, Cambridge, Mass (Originally published in French 1992).

Latour, B., 1999a. On recalling ANT. In: Law, J., Hassard, J. (Eds.), Actor Network Theory and after, Blackwell and the Sociological Review. Oxford and Keele, pp. 15–25.

Latour, B., 1999b. Pandora's Hope: Essays on the Reality of Science Studies. Harvard University Press, Cambridge, MA.

Latour, B., 2004a. Politics of Nature: How to Bring the Sciences into Democracy (C Porter, Trans.). Harvard University Press, Cambridge, MA.

Latour, B., 2004b. Why critique has run out of steam: from matters of fact to matters of concern. Critical Inquiry 30, 225–248.

Latour, B., 2005. Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, Oxford.
Latour, B., 2010. Coming out as a philosopher. Social Studies of Science 40 (4), 599–608.
Latour, B., 2013. An Inquiry into Modes of Existence: An Anthropology of the Moderns. Harvard University Press, Cambridge, MA.
Latour, B., Serres, M., 1995. Conversations on Science, Culture, and Time (R. Lapidus, Trans.). University of Michigan Press, Ann Arbor (Originally published in French as Elaircissements 1990).
Latour, B., Weibel, P. (Eds.), 2008. Making Things Public: Atmospheres of Democracy. MIT Press, Cambridge, MA.
Latour, B., Woolgar, S., 1986. Laboratory Life: The Construction of Scientific Facts (first ed. published as Laboratory Life: The Social Construction of Scientific Facts in 1979), second ed. Princeton University, Press, Princeton, NJ.
Law, J., 1999. After ANT: complexity, naming and topology. In: Law, J., Hassard, J. (Eds.), Actor Network Theory and after, Blackwell and the Sociological Review. Oxford and Keele, pp. 1–14.
Sokal, A., Bricmont, J., 1998. Intellectual Impostures. Profile, London. Stengers, I. 2010. Cosmopolitics I. Minneapolis, University of Minnesota Press.
Tresch, J., 2013. Another turn after ANT: an interview with Bruno Latour. Social Studies of Science 43 (2), 302–313.
van Inwagen, P., Winter 2013. Metaphysics. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy, Edition. http://plato.stanford.edu/archives/win2013/entries/metaphysics/.

Relevant Websites

http://anthem-group.net/category/bruno-latour/ – The ANTHEM Group blog initially focused on putting Latour's work into conversation with Heidegger but now offers lively and wide-ranging discussions more broadly around philosophy, technology and ANT.
http://modesofexistence.org – Collaborative website for *An Inquiry into Modes of Existence*.
http://www.holbergprisen.no/en/bruno-latour – Interview and information about Latour from the Holberg Prize website.
http://www.bruno-latour.fr/ – Latour's Website.

Law and Aging

Israel (Issi) Doron, University of Haifa, Haifa, Israel

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by L.A. Frolik, volume 12, pp. 8426–8429, © 2001, Elsevier Ltd.

Abstract

Law and aging is a new and developing field of knowledge. It is both multi- and interdisciplinary by nature. It combines theoretical and practical knowledge from the fields of law, gerontology, and geriatrics. It focuses not only on legal interests of older persons and their family members, but also on substantive legal issues, which are of importance in old age (e.g., guardianship, social security, or health law). As described in this article, law and aging is a rich and diverse field, which has novel elements in both international law as well as municipal and local law.

What Is 'Law and Aging'?

The phrase 'law and aging' can actually carry various meanings (Morgan, 2010). It can mean 'elder law,' which is a specific field within the legal discipline that focuses on the legal issues relevant to the older population, their families, and caregivers. This could be called a 'population-based' or 'client-focused' definition: it focuses on the target population and clientele, and defines its field accordingly. It assumes that 'older persons' can be identified as a unique social group, and that this social group has unique legal interests that can be defined and categorized as elder law. According to this approach, the clients are the one who define the field of practice (Kohn and Spurgeon, 2010).

Unlike the population/client-centered approach, a different conceptual approach to elder law can be found in a 'value/topic-based' definition: it focuses on core values (e.g., autonomy, dignity, or quality of life), and organizes a 'list' of substantive legal topics that ensure these values in old age (NAELA, 2010). Historically, fields such as poverty law, social security law, housing law, or health law were all considered a part of 'elder law.' In the United States, for example, elder law was developed to some extent through 'topic-specific' practitioners (e.g., in the field of social security or health law), who transformed and broadened their interests beyond their initial expertise in order to promote the human rights (HR) and dignity of older persons.

A different conceptual approach focuses less on the legal interests of older persons or the substantive fields of law that relate to them. A more sociological approach to the field of law and aging (also known as 'jurisprudential gerontology' or 'geriatric jurisprudence') focuses on the social interaction between aging processes, the social construction of old age, and legal institutions. Its origins are in the sociology of law, and it is interested in power relationships, social phenomena such as ageism, and how law and ageism interact in real life (Eglit, 2004).

For the purposes of this project, we adopt a broad and inclusive definition. Law and aging in this article is meant to be a broad, 'umbrella' term, referring to a multi- and interdisciplinary field of study, which includes the different and diverse empirical and conceptual interactions between legal institutions, norms, and theories, with the social phenomena of aging and its accompanying multidisciplinary science of gerontology and geriatrics. In this sense, a guardianship order over an older person handed by local courts; the psychological and economic implications of mandatory retirement laws; policy research about the rates of involvement of older drivers in road accidents with reference to the legal regulation of drivers licenses; or the study of the medical dilemmas around physician assisted suicide – are all well within the field of law and aging.

It is worth noting that there are critics who question the justification of the existence of 'elder law' or 'law and aging' as a valid field of legal knowledge. The most basic argument in this context is that law should not be different for a person over 80 years of age as opposed to a person of 45 years of age. Moreover, to the extent that law is indeed different, it only reflects – and reinforces – ageism, and hence should be avoided as a proper policy. Moreover, 'older persons,' as a group can only be characterized in their diversity and incoherence, therefore, there cannot be a rational justification for this field of knowledge. These criticisms have merit. However, in the real world, old age remains a strong socio-legal-policy construct (like gender and race). Ignoring this reality or turning away from the role that law plays in constructing old age, does not serve older persons or our understating nor can it allow us to change or legally combat ageism (Herring, 2009).

A Short History of Law and Aging

The field of 'law and aging' (or 'elder law') is relatively new (Bogutz, 2009; Frolik, 1993). It was 'recognized' as a distinct field within the discipline of law, only 40 years ago. Its historical roots were based on various legal and social interests. On the legal side, one base was lawyers specializing in wills and estate planning, which was broadened into the complicated Medicaid planning. In light of the dramatic change of demographic aging, these legal practitioners sought professional recognition of their specialty, as well as tools for professional development. A different interest base was also found in lawyers dealing with poverty law, and rights of the poor and elderly, in fields such as social security, housing, and health law. Lawyers and legal scholars began to recognize a need for unique and special attention for the legal needs and interests of older persons and their families, especially as the baby-boom generation started to face old age and legal issues around aging.

But not only practitioners were interested in this new field, academics, legal scholars, and gerontologists were also starting to look into the interactions of law and aging from both theoretical and empirical dimensions, and from an interest in developing novel fields of law. These efforts and recognitions culminated in the establishment of various professional organizations and nongovernmental organizations (NGOs). For example, the American National Senior Citizens Law Center was founded in 1972; the American Bar Association's Commission on Law and Aging was established in 1978; and the American National Academy of Elder Law Attorneys was formed in 1987 (Morgan, 2010). These associations were soon accompanied by the development of academic courses and law clinics and eventually by law journals in the field (e.g., *The Elder Law Journal* (first issue published 1993); or the *Marquette Elder's Advisor Law Review* (1999)).

This American legal development was slowly acknowledged and adopted in other common-law and English-speaking countries as well. In Canada, for example, consciousness of the existence of this field began to crystallize in the 1990s (Beaulieu and Spencer, 2000; Doron, 2003a; Soden, 2005). This has culminated in the establishment of the Canadian Centre of Elder Law Studies in the early 2000s, which has initiated academic conferences, research studies, and scholarship in the field. Finally, in Australia, a center for elder law was established at the University of Western Sydney at the beginning of the new millennium, and alongside it, a new *Elder Law Review* journal was founded (first issue published 2002) (Field, 2005).

European jurisdictions (in contrast with North America) have only very recently 'joined' this trend (Evrard and Lacour, 2012). Books or articles referring directly to the field of 'elder law' or 'law and aging' – as such – have only very recently begun to be published. Interest in the field was starting to emerge in European and other countries, including the United Kingdom (Herring, 2009; McDonald and Taylor, 2006; Meenan and Broadbent, 2007), Israel (Doron, 2007), and more recently in France, Belgium (Evrard and Lacour, 2012), and Germany.

Mapping the Field of Law and Aging: A Multidimensional Model

Over the years various theoretical perspectives have been proposed to provide a broad conceptual understanding of the field. Such perspectives have included: a feminist approach (Dayton, 2008); a therapeutic approach (Kapp, 2008); a law and economics approach (Kaplan, 2008); a disability studies approach (Surtees, 2008); a later life planning approach (Frolik, 2008); an equity approach (Hall, 2008); or a mental health theory approach (Schmidt, 2008).

One of the models that provide a broad and inclusive perspective on the field is known as the 'multidimensional model.' This model is an ongoing project, which continues to be developed and studied (Doron, 2003a, 2008a; Doron and Kohn, 2011). This model is presented in **Figure 1** below.

As shown in **Figure 1**, and as described by Doron and Kohn (2011), the model consists of five different elements, as follows: The 'Legal Principles Dimension' recognizes that rights of older persons can be viewed through the lens of general and nonage specific principles and doctrines. For example, values and equality and antidiscrimination can be applied in order to promote the rights of older persons without having to 'create' novel or unique rules. One of the more well-known examples within this dimension is the struggle against age discrimination in old age. Hence, promoting rights of older persons through antiage discrimination legislation is a clear implementation of 'core' legal values such as equality or dignity. These values do not attempt to 'treat' older persons 'differently' but rather argue that they should be treated like any other member of the society and not discriminated against only due to chronological age.

The 'Protective Dimension' reflects the understanding that sometimes the universal dimension is not enough, and some unique and specific rules are needed to protect older persons. Such legal measures are especially relevant to combat or address complex social issues such as elder abuse, self-neglect, or poverty in old age. Well-known examples for this dimension can be found in adult guardianship legal instruments, adult

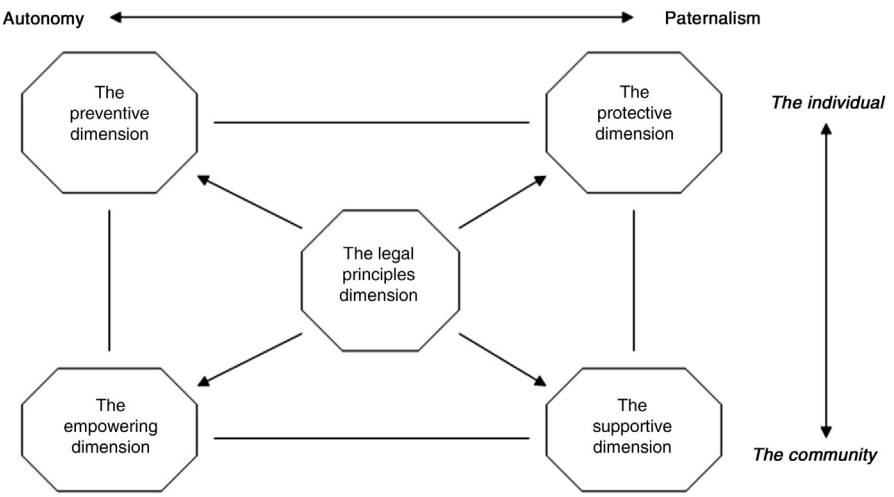

Figure 1 The multidimensional model. Source: Doron, I., 2003b. A multi-dimensional model of Elder Law: an Israeli example. Ageing International 28 (3), 242–259.

protective services mechanisms, mandatory reporting laws, or quality assurance regulations of nursing homes. Such laws authorize socio-legal interventions that can prevent and discontinue abuse or neglect of older persons, while providing solutions that enable providence of care.

The 'Supportive Dimension' mirrors social gerontological knowledge regarding the importance of informal care and social networks to the quality of life and health of people in old age. This knowledge presents a challenge to legislatures and policy makers to both mandate and encourage family members to care for their older parents and family members. Such legal policies may allow family members to take paid leave in order to care for the older parents; they may provide tax incentives or discounts for informal cares that provide care for dependent older adults; or, they can provide paid respite and relief for the caregivers themselves. By providing the tools for promoting and supporting social support systems, systems that the gerontological literature indicates can play an important role in promoting older persons' well-being and elder self-determination. These third dimension laws have the potential to improve significantly the lives of older adults.

The 'Preventive Dimension' stresses the important role law can serve as a preventive social tool in old age. In many countries, law now provides ample tools for various realities of old age, such as planning for the loss of mental capacity, or planning for the need for long-term care. Some of the well-known instruments of this dimension include legal tools such as 'durable powers of attorney' (also known as 'enduring/continuing powers of attorney') – which allow individuals to plan for future financial decision making. Others – such as health care proxies or living wills – allow individuals to plan for health care decision making. Specifically, a health care proxy (also known as a power of attorney for health care) allows individuals to appoint someone else to make decisions for him/her in the event he/she is unable to do so for himself/herself. All these legal tools can play a crucially important role in enabling older individuals to make autonomous, individualistic choices about their own lives. These personal planning tools are flexible in that they can be easily adjusted to correspond to the particular needs and desires of the individual. They also have the potential to be sensitive to ethnic, cultural, and moral differences in that they do not prescribe a particular course of action, but rather permit decisions to be made based on individual preferences and values.

Finally, the 'Empowering Dimension' stresses the known fact that merely adding rights to the law on the books is insufficient to change reality. One needs measures such as advocacy, legal representation, knowledge creation, and distribution, in order to actually make a difference in the lives of older persons. An example of an empowering legal policy is mandatory representation of older persons in guardianship proceedings. In light of the far-reaching legal consequences of imposing a guardianship over an individual, most US states require the allegedly incapacitated person be represented by either legal counsel or a guardian-ad-litem at the hearing (American Bar Association Commission on Law and Aging, 2009). Such counsel is important both for remedying potential power imbalances, and for ensuring that those subjected to guardianship proceeding are afforded due process of law.

The multidimensional model emphasizes the diversity and variability of legal systems' responses to aging and the concerns of older adults. The model suggests that the field of law and aging is composed of many distinct types of laws designed to meet the needs of older adults. The model also suggests that elder law reflects a range of political and philosophic approaches to meeting the needs of older persons. The field is by no means directed by a single viewpoint. Accordingly, it should not be considered either 'individualistic' or 'paternalistic,' nor should it be portrayed as promoting 'negative' as opposed to 'positive' rights, or as favoring the individual over the family, or as privileging the private over the public. Rather, elder law encompasses a broad range of possible approaches and perspectives, and a given legal system's approach to 'elder law' can be fully appreciated only by understanding how that system weighs the different components – or dimensions – of the field of law and aging.

International Elder Law

In recent years the field of law and aging has undergone what can be described as the 'internationalization' of the field. This has been the outcome of a call made by various scholars (e.g., Doron, 2006; Doron and Apter, 2010; Doron and Spanier, 2012; Rodriguez-Pinzon and Martin, 2002; Tang and Lee, 2006).

Comparative Elder Law

Various legal systems around the world approach the range of law and aging issues in very different ways. In recent years there has been a growing interest in studying and comparing these different legal approaches by using comparative law methodologies. Comparative law can be a powerful tool to both better understand one's own legal system and values while realizing that other legal options do exist.

There are many recent examples of comparative elder law research. For example, a comparative legal study in the field of elder guardianship revealed totally different conceptual approaches to this legal institution. On the one hand, one can find a 'North American' approach to adult guardianship, which is focused on 'substitute decision-making' models; while on the other hand, a more European/Continental approach to adult guardianship is focused on assisted/shared decision-making models (Doron, 2002).

Global and Regional HR of Older Persons

Since World War II there has been a proliferation of HR instruments. Interestingly, older persons are one of the few populations that still do not have an HR convention of their own. An analysis and review of the international rights of older persons published in 2002 by Rodriguez-Pinzon and Martin (2002) summarized the legal field in this area as follows:

> A strategy to have a comprehensive legal instrument on elderly rights is missing at the international level in both universal and regional systems. There are very few provisions in international law that

> directly address elderly rights. There are isolated efforts by certain international bodies to systematically refer to the rights of the elderly when interpreting their corresponding conventions …
>
> However, there is no specific international body with the mandate to focus on the rights of the elderly. Nor is there an elderly rights convention in place. It is in fact the only vulnerable population that does not have a comprehensive and/or binding international instrument addressing their rights specifically. (p. 1008)

However, recent years have been characterized by a movement to change this reality both on the global/UN level, as well on the regional levels (e.g., Organization of American States (OAS), African Union (AU), and Europe). At the UN level, a symbolic and important point of reference was made at the UN Expert Meeting Group that was held in Bonn, Germany (April 2009). The international group of experts in the field of elder law held in-depth debates that resulted in various recommendations with regard to the advancement of the rights of older person, including a specific reference to the need for a new convention:

> A convention on the rights of older persons would add additional weight in furthering, deepening and more precisely defining the rights of older persons. A convention would create obligatory and binding international law. Similar to the adoption of various other human rights instruments, member states would undertake a threefold commitment when adopting such a convention: to respect, to protect and to fulfill the rights enshrined in the relevant text.

This recommendation was followed by a series of UN resolutions (UN GA R. 65/182) culminating in the establishment of the Open Ended Working Group (OEWG) on strengthening the protection of the HR of older persons. The OEWG held its first meeting in New York in April 2010, and held its fourth meeting in August 2013. Only time will tell to what extent this working group will be successful in moving forward the future International Convention for the Rights of Older Persons (OEWG, 2013).

While at the UN level the OEWG is still struggling with the question as to whether there is a need or a justification for a new convention, some regional organizations and agencies have move forward much quicker. The most startling and promising advancement was made in this area at the OAS. The General Assembly, through its resolution AG/RES. 2654 (XLI-O/11), "Protecting the Human Rights of Older Persons," adopted at the fourth plenary session on 7 June 2011, resolved, in operative paragraph 1, the following:

> To request the Permanent Council to establish a working group composed of national representatives and experts from the academic sector and civil society, as well as from international organizations and specialized agencies, to prepare a report by the end of 2011, analyzing the situation of older persons in the Hemisphere and the effectiveness of binding universal and regional human rights instruments with regard to protection of the rights of older persons.

Meanwhile, the AU has also moved forward very quickly on this front. Starting in 2007, the African Commission on Human and Peoples' Rights, decided to establish a Focal Point on the Rights of Elderly Persons in Africa (which was later transformed into a Working Group on the Rights of Older Persons and Persons with Disabilities in 2009 (*ACHPR/Res.143 (XXXXV) 09*)). Among others, a primary role of this working group was "Spearheading the process of drafting a Protocol on the Rights of Older Persons for submission to the AU Policy Organs for consideration and adoption as soon as possible." A preliminary Draft Protocol was presented to the African Commission during the 48th session held in 2010 in Banjul, Gambia.

As of this time, none of the above international or regional initiatives have transformed into final or binding legal instruments. However, there is no doubt that it is only a matter of time until these initiatives will eventually crystallize into an official and binding HR instrument for older persons, a development that will change the international landscape of law and aging.

Private International Law

Globalization and the movement of older persons around the world for various reasons (old age migration; sun-belt migration; elder-care migration, and more) have given rise to the need for private international law solutions for new and complex legal issues concerning older persons. For example: which law will decide the validity of a will made by an American widower, who moved to live in Sweden in old age with his newly met partner, but has property in Canada? Or should the Japanese courts honor advanced health care directives made by an English citizen under English law, but who suffered a stroke during a visit to his son in Japan? And how should the tax and social security authorities in Germany deal with an Argentinean citizen, who worked in Greece for many years, but decided to retire in Germany, and receives a pension from Argentina? These questions and more can and should be answered by the substantive law in the field of private international elder law.

An example in this field can be found in the Convention on the International Protection of Adults. This international private-law convention "applies to the protection in international situations of adults who, by reason of impairment or insufficiency of their personal facilities, are not in a position to protect their interests." The convention "aims to shield the vulnerable members of society by determining which state – that of their citizenship or of their current residence – may assert jurisdiction over them" (Fagan, 2002: p. 331). While not explicitly stated, it is clear that this convention is of significant, if not primary, relevance to older persons.

Municipal Elder Law

A novel development in the field of law and aging has begun to look into the legal potential existing in the powers of municipalities, cities, and villages to regulate and establish rights for their older residents. In an era of globalization, it is no simple matter to claim that decentralization and the broadening of the jurisdiction of local government are appropriate methods of promoting the rights of the older population. Nevertheless,

recent studies suggest that the power of local elder law is not incidental or residual. Local elder law can potentially be used as a strategic legal tool to change the position of older people in their communities. Local governments in various places around the world can play an important role as both leader and champion for community-based legislative initiatives to promote the rights of their older residents.

In a series of studies conducted in Israel, Canada (Doron, 2009), and the US (Dayton and Doron, 2012; Doron and Dayton, 2012), it was found that despite the potential legal powers that exist in the hands of local authorities, very little attention is given to the rights of older persons in this field. For example, in a national US study, the authors concluded that: "Unfortunately, our findings in this study suggest that local law is not currently fulfilling its potential as a means to expand the rights of older citizens. Few cities across the country appear to have taken more than minor steps in the direction of developing clear, ordinance-based policies that will truly foster age-friendly cities and committed public agendas for achieving the objective of 'aging in place' on a large scale" (Dayton and Doron, 2012: p. 116). Moreover, as described by the researchers: "The ordinances we found, that do explicitly acknowledge the needs and rights of seniors, tend almost entirely to fall within the most traditional spheres of local regulatory authority, such as land use laws and taxation. Many ordinances focus on age as a negative attribute; some arguably work to isolate and stigmatize age by relegating congregate housing options for compromised seniors to commercial districts or other nonresidential areas" (p. 116).

Jurisprudential Gerontology/Geriatric Jurisprudence

The article began by presenting the various definitional approaches to law and aging. One of the key differences between the 'elder law' as opposed to the 'law and aging' approaches is their relationship to gerontology and to other nonlegal disciplines. Gerontology (i.e., the science of aging) has been characterized by being inter- and multidisciplinary. It includes psychological, sociological, anthropological, and many other disciplinary perspectives on aging. Hence, the scientific and disciplinary cooperation between gerontology theories and empirical findings with legal knowledge has developed into what is known as 'jurisprudential gerontology' or 'geriatric jurisprudence.' This multi- and interdisciplinary approach has been gaining force in the current development of the field of law and aging (Doron, 2006; Kapp, 2003).

In a recent article, Doron and Meenan (2012) argued that geriatric jurisprudence is essential for the further development of the field of law and aging. Firstly, they describe what geriatric jurisprudence is: "It is a new interdisciplinary research approach that attempts to combine jurisprudence with geriatrics. It is not only the outcome of the meeting of lawyers, geriatricians, and gerontologists, but is the attempt to break down the disciplinary borders: an integration of the fields of knowledge based in geriatrics, gerontology, and law. The different scientific approaches enrich each other and enable one to better conceptualize and theorize complex socio-medical realities concerning older people" (Doron and Meenan, 2012).

Secondly, they provide the justification for the need for such an approach to the field of law and aging: "[G]erontology and geriatrics are, by nature, not only multi- but also interdisciplinary fields. One cannot, for example, address issues such as age discrimination in health care or advanced health care directives without taking into account not only the medical, psychological, or social conditions of older persons, but also the relevant laws and legal precedents that exist in these fields. Moreover, law is an important tool for social policy and social change. Law is an important ingredient of the social and cultural fabric of the society. One cannot fully understand the social and medical dimensions that are relevant to the lives of older persons without taking into account the legal implications whether speaking of empowerment, preemption, protection, or redress" (Doron and Meenan, 2012: p. 3).

Conclusion: The Future of Law and Aging

Law and aging is still an emerging field of law and society scholarship, struggling to be recognized as a legitimate unique field of knowledge. In this section we ask: what can we foresee in the coming years? In what directions will elder law develop? What are the new and promising directions that attorneys and advocates can promise to older persons? Some general trends can be identified and recommendations made as follows.

The Growing Interest in and Development of the Field

In many parts of the world 'law and aging' is still an unknown territory. It is not taught in law schools, nor in gerontological or geriatric studies. Even in countries where elder law and law and aging are known as distinct fields of law, they are still under-researched, and there is much room to further develop the empirical data and theoretical foundations of the field. From this perspective, it is clear that much academic and practical work is still needed in this field in order to establish its theoretical foundations.

More specifically, there is place for empirical studies that provide the necessary data to better understand the ways in which older persons experience law in daily life, and how judges and lawmakers construct old age in their legislative work and their case rulings. Greater cooperation is needed between scholars and practitioners to establish intellectual cross-fertilization of knowledge and practice. Finally, a stronger and richer body of 'third sector'/NGOs that specialize in this field and serve as advocates and activists of rights of older persons should be further developed.

International HR

It is quite clear that the coming years will bring great changes to the field of law and aging in the international arena, both globally and regionally. Both in the Americas and in Africa, the regional international bodies are deeply engaged in the establishment of new and novel HR instruments for rights of older persons. On the UN level, the opposing forces are still strong, yet there is a steady movement forward, with a push from NGOs around the world to continue and work for the establishment of a new binding HR convention for older persons.

These new developments on the international arena will directly influence the future of law and aging in many societies. Preliminary findings from supernational and international legal tribunals (e.g., The European Court of Justice and the European Court of Human Rights) reveal that older persons are using these legal forums and that they have the potential to serve as instruments for social change and the promotion of HR of older persons (Doron, 2013).

Legal Politics

The field of law and aging is also going to be strongly affected by the politics of aging (Hudson, 2010; Binstock, 2007; Vincent, 1999). This can be viewed from different angles. For example, one angle is the growing trend of national and local political parties of pensioners and senior citizens as well as elder-rights social movements (Kohn, 2010, 2012). These new political parties and social movements have begun to gain power and achieve political success, which can be translated into legislation, legal reform, and increased interest in the field. A different angle can be the socioeconomic construction of 'generational warfare,' especially around the funding of pensions, and wealth distribution between the younger and older generations.

The politics of aging can be reflected in legislative initiatives regarding rights of older persons. One such example can be found in Israeli politics and the legislative developments around Israel's Senior Citizens' Rights Act. A study conducted on this specific legislation (Doron, 2008b) concluded that: "The analysis of Israel's *Senior Citizens Act* affords no reason for political-gerontology optimism. The interest-driven reality of politics is reflected in the ways in which the law is 'brought to life' shortly before general elections, only to be allowed to 'die' again soon after. Prime Ministers and Ministers of Labor and Social Affairs show no interest in senior citizens, but use the law to create a caring image that lacks foundation. MKs (Parliament Members) are happy to laud Israel's older population. However, they are not willing to engage in a real struggle in the face of opposition from the Prime Minister or the Ministers of Finance or Social Affairs. Lastly, the legislature seems not interested in the actual implementation of the law, since image is far more important than the lives of old people" (Doron, 2008b: pp. 196–197).

As the interest in the 'gray force' of politics continues to grow, it will be important to continue to study how these political forces 'translate' into legislation, regulation, and legal rights of older persons.

Ageism and Law as a Tool for Social Change

Finally, an important role of the 'law and aging' field can be to provide tools for advocates of social change in combating ageism. Awareness of the social construction of old age and its manifestations in law – both negative and positive – is relatively recent (e.g., Eglit, 2004; Whitton, 1997). Yet, since the introduction of the term 'ageism' by Professor Robert Butler in the late 1960s, there has been a growing body of literature and research indicating the scope of ageism on all socio levels. Hence, similar to the role of legal scholars and scholarship in other movements for social change, for example, the feminist or the civil rights movements, the field of law and aging can provide an important foundation for the next civil rights movement, that of the older population. However, in order to serve that role, empirical and theoretical foundations need to be further established and strengthened – yet another challenge for the field of law and aging looking forward.

See also: Age Policy; Age, Sociology of; Age: Anthropological Aspects; Aging and Older People, Geography of; Geriatric Psychiatry; Human Rights and Social Work; Old Age Work Participation; Pension Policy; Population Aging: Economic and Social Consequences; Retirement and Encore Adulthood: The New Later Life Course.

Bibliography

Beaulieu, M.B., Spencer, C., 2000. Older Adult's Personal Relationships and the Law in Canada. Law Commission of Canada, Ottawa.

Binstock, R.H., 2007. Older people and political engagement: from avid voters to "cooled out marks." Generation 30 (4), 24–31.

Bogutz, A.D., 2009. Elder law: A personal perspective. In: Doron, I. (Ed.), Theories on Law and Ageing: The Jurisprudence of Elder Law. Springer, Berlin, pp. 1–10.

Dayton, K., 2008. A feminist approach to elder law. In: Doron, I. (Ed.), Theories on Law and Ageing. Springer, Berlin, pp. 45–58.

Dayton, K., Doron, I., 2012. Municipal elder law: a Minnesota perspective. The Elder Law Journal 20 (1), 33–70.

Doron, I., 2002. Elder guardianship kaleidoscope: a comparative legal perspective. International Journal of Law, Policy and the Family 16 (3), 368–398.

Doron, I., 2003a. Mental incapacity, guardianship, and the elderly: an exploratory study of Ontario's Consent and Capacity Board. Canadian Journal of Law and Society 18 (1), 131–148.

Doron, I., 2003b. A multi-dimensional model of elder law: an Israeli example. Ageing International 28 (3), 242–259.

Doron, I., 2006. Elder law: current issues and future frontiers. European Journal on Ageing 3 (1), 60–66.

Doron, I., 2007. Law and ageing in Israel: the development of a new field of law. Journal of International Ageing, Law and Policy 2, 33–66.

Doron, I., 2008a. A multi-dimensional model of elder law. In: Doron, I. (Ed.), Theories on Law and Ageing. Springer, Berlin, pp. 59–74.

Doron, I., 2008b. Law and older people - the rise and fall of Israel's senior citizens' act. Journal of Aging and Social Policy 20 (3), 353–375.

Doron, I., 2009. Time for municipal elder rights law: an Anglo-Canadian perspective. Canadian Journal on Ageing 27 (3), 301–313.

Doron, I., 2013. Older Europeans and the European Court of Justice. Age & Ageing 42, 604–608.

Doron, I., Apter, I., 2010. The debate around the need for an international convention on the rights of older persons. The Gerontologist 50 (5), 586–593.

Doron, I., Dayton, K., 2012. "Thinking locally" law, aging and municipal government findings from a national survey. Temple Political and Civil Rights Law Review 21 (2), 365–386.

Doron, I., Kohn, N.A., 2011. Aging and law: using a multi-dimensional model to understand the legal response to aging. In: Chen, S., Powell, J.L. (Eds.), Aging in Perspective and the Case of China: Issues and Approaches. Nova, New York, pp. 71–86.

Doron, I., Meenan, H., 2012. Time for geriatric jurisprudence. Gerontology 58 (3), 193–196.

Doron, I., Spanier, B., 2012. International elder law: the future of elder law. In: Doron, I., Soden, A. (Eds.), Beyond Elder Law: New Directions in Law and Aging. Springer Publications, Berlin, pp. 125–148.

Eglit, H., 2004. Elders on Trial: Age and Ageism in the American Legal System. University Press of Florida, Gainsville, FL.

Evrard, A., Lacour, C., 2012. A European approach to developing the field of law and ageing. In: Doron, I. (Ed.), Theories on Law and Ageing. Springer, Berlin, pp. 149–174.

Fagan, A.R., 2002. An analysis of the convention on the international Protection of adults. The Elder Law Journal 10, 329–337.

Field, S., 2005. Issues facing older Australians: legal, financial and societal. Journal of International Aging, Law and Policy 1, 95–115.
Frolik, L., 1993. The developing field of elder law: a historical perspective. The Elder Law Journal 1 (1), 1–20.
Frolik, L., 2008. Later life legal planning. In: Doron, I. (Ed.), Theories on Law and Ageing. Springer, Berlin, pp. 11–31.
Hall, M.I., 2008. Equity theory: responding to the material exploitation of the vulnerable but capable. In: Doron, I. (Ed.), Theories on Law and Ageing. Springer, Berlin, pp. 107–120.
Herring, J., 2009. Older People in Law and Society. Oxford University Press, Oxford.
Hudson, R.B. (Ed.), 2010. The New Politics of Old Age Policy, second ed. Johns Hopkins University Press, Baltimore, MD.
Kaplan, R.L., 2008. A law and economics approach. In: Doron, I. (Ed.), Theories on Law and Ageing. Springer, Berlin, pp. 75–92.
Kapp, M., 2003. The Law and Older Persons: Is Geriatric Jurisprudence Therapeutic? Carolina Academic Press, Durham, NC.
Kapp, M., 2008. A therapeutic approach. In: Doron, I. (Ed.), Theories on Law and Ageing. Springer, Berlin, pp. 31–44.
Kohn, N., 2010. The lawyer's role in fostering an elder rights movement. William Mitchell Law Review 37 (1), 49–79.
Kohn, N., 2012. A civil rights approach to elder law. In: Doron, I. (Ed.), Theories on Law and Ageing. Springer, Berlin, pp. 19–33.
Kohn, N.A., Spurgeon, E.D., 2010. Elder law teaching and scholarship: an empirical analysis of an evolving field. Journal of Legal Education 59 (3), 414–431.
McDonald, A., Taylor, M., 2006. Older People and the Law. Policy Press, Bristol, UK.
Meenan, H., Broadbent, G., 2007. Law, ageing and policy in the United Kingdom. Journal of International Aging, Law and Policy 2, 67–102.
Morgan, R.C., 2010. The future of elder law practice. William Mitchell Law Review 37 (1), 1–48.
National Academy of Elder Law Attorneys, 2010. What Is Elder Law? http://www.naela.org/Public_WhatIsElderLaw.aspx?Internal=true (accessed 15.02.2010).
OEWG, 2013. UN Open Ended Working Group on Ageing for the Purpose of Strengthening the Protection of the Human Rights of Older Persons. http://social.un.org/ageing-working-group/.
Rodriguez-Pinzon, D., Martin, C., 2002. International human rights status of elderly persons. American University International Law Review 18, 915.
Schmidt, W.C., 2008. Law and aging: mental health theory approach. In: Doron, I. (Ed.), Theories on Law and Ageing. Springer, Berlin, pp. 121–144.
Soden, A., 2005. Advising the Older Client. LexisNexis, Markham, ON.
Surtees, D., 2008. What can elder law learn from disability law? In: Doron, I. (Ed.), Theories on Law and Ageing. Springer, Berlin, pp. 107–120.
Tang, K.L., Lee, J.J., 2006. Global social justice for older people: the case for an international convention on the rights of older people. British Journal of Social Work 36 (7), 1135.
Vincent, J.A., 1999. Politics, Power and Old Age. Open University Press, Buckingham, UK.
Whitton, L.S., 1997. Ageism: paternalism and prejudice. DePaul Law Review 46, 453.

Relevant Websites

http://www.americanbar.org/groups/law_aging.html – ABA (American Bar Association) Commission on Law and Aging.
http://www.bcli.org/ccel – Canadian Centre for Elder law Studies.
http://www.globalelderlaw.blogspot.com – Global Elder Law Blog.
http://www.naela.org/ – National Academy of Elder Law Attorneys.
http://www.nsclc.org – National Senior Citizens Law Center.

Law and Courts in Authoritarian States

Peter H Solomon, Jr., University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Authoritarian rulers need law and courts as instruments of rule, to perform such functions as social control, legitimation, administrative accountability, and encouragement of investment and trade, but more than their democratic colleagues they find it difficult to sacrifice political power. Consequently, they fashion weak and dependent courts, opt for fragmented jurisdiction (with special tribunals for matters of regime interest), readily readjust the jurisdiction of courts and judicial discretion when they are provoked, or create formally independent courts that are undermined by informal practices. Authoritarian rulers prefer narrow positivist understandings of law to avoid constraints but may face pressures to meet international standards for the protection of business or human rights. Rule by law rather than any version of rule of law characterizes the role of law in authoritarian states.

Law and courts play major roles in modern authoritarian states, as they do in democratic ones. At a minimum they perform the same basic functions of social regulation and the resolution of disputes, and they do so in a reasonable fashion in many instances. At the same time, though, law and its enforcement through courts often connect to the distribution and exercise of power, and it is through this nexus that significant differences across regime types arise.

As a rule, authoritarian states lack the features that define democracy, such as free and fair elections, observance of civil liberties, universal suffrage, and effective control of the state by elected officials. But the essence of authoritarianism in its various forms (including competitive authoritarianism and so-called hybrid regimes) lies in the concentration of political power in the hands of an autocrat or an oligarch. Whereas in Western democracies the law may serve as a constraint on the power of rulers as well as an instrument of rule, in authoritarian regimes the instrumental role tends to dominate, in some cases (and ideally) to the exclusion of a limiting role. Thus, while theorists may speak of 'rule of law (ROL)' as a feature of democracy (meaning not only procedural guarantees for the citizen but also constraints on leaders), the most convenient shorthand for authoritarian regimes is 'rule by law', a phrase that emphasizes law's role as an instrument of rule (Ginsburg and Moustafa, 2008).

One should stress that the availability of law as an instrument of rule for authoritarian rulers does not mean that they will act lawfully all or most of the time. They and their states may still "predate on their subjects, [or] violate property rights, and [they] are unbound by the strictures of law" (Darnden, 2008: p. 38). The same rulers may also promulgate laws that are not designed to be obeyed (Darnden, 2008: p. 38).

But the reality is that as much as authoritarian leaders may wish to maximize their power and avoid law's potential limits, in practice they end up making compromises. This is unavoidable if law and courts are to perform some of the functions that the leaders value, including legitimation of their rule. Moreover, leaders must maintain elite cohesion and popular consent, including the consent of liberalizing constituencies engaging in social mobilization (Svolik, 2012). Finally, in the late twentieth and early twenty-first centuries, leaders are facing global pressures to meet standards of legal protection, at least for business activities if not also for human rights.

Not surprisingly, then, in some authoritarian states the role and functions of law (even its meaning) are contested and involve tensions, if not also conflict. Such tensions are especially prominent in regimes where leaders claim legitimacy through the use of quasi-democratic institutions or in regimes that some might call hybrid or competitive authoritarian (Levitsky and Way, 2010). Such tensions also characterized authoritarian states in nineteenth-century Europe (e.g., Germany and Russia), where liberal political forces promoted judicial power to balance that of the executive.

This article starts with an examination of the role of law and the functions it may play in authoritarian contexts. It goes on to consider the alternative approaches that rulers take to the power and independence of courts, so as to maximize outcomes favorable to the regime in cases that matter to it. Then, the article illustrates ways that the pressures of the global economy and polity influence choices made by authoritarian rulers. Finally, it addresses briefly two unresolved issues – the utility of stretching the concept of ROL to fit some authoritarian states; and how to explain variation in the role of law and courts in authoritarianism.

The Law

For the authoritarian ruler, the most convenient understanding of law is limited to his own legal enactments, whatever they are called. The law consists of the laws of the state, especially the edicts or decrees of the ruler himself. However, few autocrats have had the luxury of gaining a monopoly on law in their time. Most feudal and early modern rulers in Europe faced competition from the realm of the sacred, or the domain of canon law, where they had little or no authority. (Norms of customary law were usually built into the laws of the state). Moreover, the idea of a source of secular law beyond the state and ruler – 'natural law' – had been in circulation at least since its eloquent articulation by the Roman orator Cicero. To be sure, the later Roman code of the Emperor Justinian placed the ruler above the law, satisfying the authoritarian urge. But even

the monarchs of pre-Enlightenment Europe confronted forms of legal pluralism and the idea of law beyond the state (Berman, 2003; Tamanaha, 2004).

From the Enlightenment, natural law theory became a justification and even a tool for limiting the power of monarchs and moving toward constitutional monarchy and democracy. While gaining popularity in France and America, the theory faced strong competition in Germany and Central Europe, where what would become known as legal positivism gained sway and became entrenched in the lasting conception of rechtsstaat. In Wilhelmian Germany, a liberal autocracy, the law consisted of the laws and not of preexisting rights or values, as some theorists of ROL averred (Ledford, 1993). In the twentieth century, most authoritarian regimes followed a positivist view of law as comprising merely of the laws, but this view was soon challenged by the new international law of human rights, itself an emanation of natural law theory. Once again, as we shall see, many authoritarian rulers had to yield part of their cherished monopoly of law, though in particular instances they resisted.

It is not uncommon for authoritarian regimes to have legislatures that, while controlled by the ruler(s) (or ruling party), have responsibility for adopting laws, thereby giving them a kind of functional legitimacy even though bureaucratic politics may account for the content of laws more often than legislative activity. But there remains the problem of the role of pronouncements that come directly from the leader(s), which may in practice take precedence over laws regardless of their legal status. Moreover, in some states (e.g., USSR/Russia), regulations issued by bureaucratic agencies often carry more weight than the laws that they are supposed to implement. The weakness (or absence) of legal hierarchy was a particular problem in former communist countries, and it has continued in many of their successors (Solomon, 2008). Finally, there may be occasions (even frequent ones) where informal practices or norms trump the formalities of law, legal procedures, and institutions. Conventional wisdom says that this happens more often in authoritarian than democratic states, if only because authoritarian states tend to be weaker.

Functions of Law in Authoritarian States

In authoritarian states, the primary role of law is that of an instrument of rule. It is, of course, one of many instruments, along with coercion or repression applied directly by police or the military. Of course, law can also serve as an instrument in the hands of individuals and social groups seeking to change policies or institutions, as part of the process known as legal mobilization. But in authoritarian states the balance among these functions favors the state, usually strongly. There remains room for different uses of the law by rulers. In some states the law delineates the privileges of particular groups or castes, thereby defining the social structure in a discriminatory way (e.g., Tsarist Russia before 1861 and Germany before 1877). In contrast, law can serve as an instrument of social change, even revolutionary change, as exemplified in the changes in property relations associated in Stalin's collectivization of agriculture and seizure of land. After the resulting upheavals and flight of peasants to the cities, Stalin turned to law again to stabilize the situation, inter alia through restrictions on movement of population (Solomon, 1996).

In their pioneering volume of essays *Rule by Law: The Politics of Courts in Authoritarian Regimes*, Tom Ginsburg and Tamir Moustafa identified four functions that law and courts often play in authoritarian states. Leaders use courts to: "(1) establish social control and sideline political opponents; (2) bolster a regime's claim to 'legal' legitimacy; (3) strengthen administrative compliance within the state's own bureaucracy machinery…; [and] (4) facilitate trade and investment…" (Ginsburg and Moustafa, 2008: Introduction). While these functions of law are commonplace and unremarkable in democratic states, they have special meaning in authoritarian ones and are not always easy to achieve, notwithstanding their increasing importance in some of them.

Managing the boundaries of politics and addressing challenges to them, both substantive and procedural, is a core task faced by rulers in all kinds of regimes, but the function has traditionally involved more consistent and intense activities by police and courts in authoritarian contexts than in democratic ones. Typically, authoritarian leaders restrict both the content of political discourse and the means of expressing views. They are also less likely to compromise with challengers by either accepting their demands or incorporating them into the political system. Just as in democratic states, so in the authoritarian ones the dividing lines between acceptable and unacceptable challenges and challengers are fluid, and the boundaries of politics change over time, depending upon the vulnerability felt by regime leaders. Thus, in the post-Soviet world peaceful demonstrations for causes that seemed acceptable before the Orange Revolution of 2004 in Ukraine appeared more threatening in the years that followed that event. After experiencing opposition protests on the streets across the country following the flawed parliamentary elections of 2011, the Russian government sought to stymie civil society initiatives through laws forcing nongovernmental organizations (NGOs) to publicize their reliance on foreign funding. The regime also engineered the prosecution and conviction of a leading opposition figure on dubious, if not trumped up, charges (Solomon, 2013).

The mechanisms used by authoritarian regimes for reacting to challengers they define as political deviants resemble those used in democratic regimes, but their content may differ. The political trial, whether held in regular courts or special tribunals, is more likely to reflect a strong bias in favor of conviction, and sometimes involves specific instructions. In the case of the famous Stalinist show trials in 1937 and 1938, where the accusations were contrived, the accused were pressured through torture into performing parts set out in elaborate theatrical scripts (Solomon, 1996). All of the accused were convicted and most were executed, and a similar fate awaited half of those prosecuted for state crimes in the special tribunals of the political police. In contrast, after 1964, Brazil's military dictatorship prosecuted a large number of its political opponents in military courts with the participation of civilian judges; over a 15-year period more than half of the accused were acquitted, and nearly two-thirds of those convicted received sentences of under 4 years imprisonment (Pereira, 2005). These examples point to the variety of ways that law and courts may be used in political policing and in the degree of judicial autonomy.

A second function of law, and reason to empower courts and even to make them appear autonomous, is the cultivation

of legitimacy for the regime. Authoritarian rulers cannot long rely on force alone; they need consent and support from the population, not to mention support from elites and comrades at arms. As a rule, stable and long lasting authoritarian regimes develop legal forms and institutions (including parliaments and courts, even elections of a sort) that curtail the image of arbitrariness in the system of rule. One of these is constitutions, and another more recent invention is constitutional courts.

Many, if not most, authoritarian regimes in the twentieth century chose to adopt constitutions, if only to set down guidelines for public administration and to give their rule normative expression. While constitutions are judicially enforceable in only a minority of authoritarian states (and even there within limits), the constitutions still matter for the rulers' relationships with the public and elites alike. A recent review of authoritarian constitutions identified four broad roles that they could fulfill. They can act as operating manuals, laying out the distribution of functions and responsibilities within the public realm. They can act as billboards, advertising the regime's purported strengths, values, and goals. They can serve as blueprints, outlining aspirations for a better future. And/or they can act as window dressing, describing institutions that lack power and rights that are nonenforceable. The constitutions in most countries have performed a mixture or blend of these functions, and the balance might change over time; in almost all cases, adoption of these constitutions led to some legitimation for the regime (Ginsburg and Simpser, unpublished book manuscript). Even the (in)famous Stalin Constitution of 1936, known for its articulation of rights that had no bearing on reality, performed the other functions in a limited way. It gave the Soviet public a sense that the period of disruption was over and promised stable and better times to come. For foreign observers attracted by the impressive list of rights, albeit to be observed only 'in the interests of the working class', the window dressing function dominated (Solomon, 1996).

To the extent that constitutions define the structure of power and locate the ruler on the institutional map, they also influence the way that informal politics is structured and determine patterns of patronage. Thus, constitutions that center power in a presidency give that institution larger-than-life importance in the world of informal institutions, often of crucial import in authoritarian contexts.

To cultivate legitimacy and to appear to have normal, even democratic regimes, authoritarian rulers not only issue constitutions but also create constitutional courts. This is especially likely when regimes enter a liberalizing phase or are trying to present their regimes as democratizing. This was the case in Egypt (in 1979); in Indonesia (in 2001); in Russia (1991), in Ukraine (1996); in Kazakhstan (1991); and in Belarus (1994). Even if judges on such courts act cautiously, taking initiative only in approved realms and within the parameters set by the regime, they may still prove too much of challenge for authoritarian rulers. Thus, it was no surprise when Kazakhstan in 1995 eliminated its constitutional court only a few years after its establishment. In the case of Egypt, the leader Mubarek was content as long as the court facilitated his economic program; when it began enforcing rights he quickly appointed a new Head of the court and expanded its membership to make the court compliant with his policies (Moustafa, 2007). For its part the Constitutional Court of the Russian Federation, after getting too involved in politics in 1993 and suffering temporary suspension, steered a careful course, avoiding decisions that contradicted the interests of the Russian President. The Court was able to carve out significant areas of action (e.g., defining the federal system and economic rights) and to develop a meaningful jurisprudence. Its biggest problem lay in securing implementation of its decisions by other governments, ministries, and courts alike (Trochev, 2008).

A third function of law and courts in authoritarian states is the promotion of administrative accountability. Whether they use regular courts, administrative courts, or tribunals within the executive branch, leaders derive two-edged benefits from administrative justice. On the one hand, they gain the support of members of the public who can successfully challenge decisions of officials that seem illegal; on the other hand, they are able to detect counterproductive, not to mention illegal, conduct on the part of officials and thereby improve the performance of government.

Authoritarian leaders, past and present, have come to value administrative justice and have been prepared to empower judges in this realm. Early examples include the Supreme Administrative Law Court of Prussia, which from 1876 provided legal controls over the actions of officials in an authoritarian state and helped bring their performance closer to the standards of the *rechtsstaat*, this despite a fixation on procedure. Within Tsarist Russia the Judicial Reform of 1864 produced, inter alia, a version of administrative justice in which citizens could bring suits against officials to commissions consisting of higher officials, although they had to collect damages directly from the officials themselves. Despite following the French model of separating administrative justice from the courts, this system worked reasonably well (Ledford, 2004; Solomon, 2008).

In Mexico under the rule of the Partido Revolucionario Institucional (PRI), or Institutional Revolutionary Party in English, citizens could complain to the courts about violations of their rights by officials through the mechanism of an amparo suit. The result was often satisfaction in individual cases, but rarely did the decisions have a broader impact because they were binding only between the parties (Magaloni, 2008).

For legal reformers in the USSR in the era of perestroika, strengthening the embryonic system of administrative justice was a priority (the USSR had relied on complaints to an administrative body – the Procuracy – rather than the courts). Soon after the country's disintegration Russia ended up with a capacious system of administrative justice, in which citizens could challenge in court the legality of almost any decision taken by an official, as well as the bulk of bureaucratic regulations. This area of the administration of justice worked particularly well in practice, and the bulk of petitioners won their cases in court and gained satisfaction in real life (Solomon, 2004).

The fourth function of law and courts for authoritarian rulers lies in the promotion of trade and foreign investment, by making investors confident that their property and interests will be protected. The rulers' desire to protect property rights and to create investor confidence lies at the root of the adoption of legal reforms in Asian countries like Singapore and China (Rajah, 2012). In contrast, the failure of law and courts to protect property rights in post-Communist countries like Russia and Ukraine make it clear that the economic interests of

state officials, reflected in state ownership of strategic economic firms and in a network of corrupt relationships, can trump concerns with creating adequate protection of property rights. At the same time, the creation of reliable forums for the resolution of business disputes and fair material and procedural law for them to apply does not mean that courts across the board will be autonomous or that other rights of citizens will be protected. As we shall see, concern with the needs of investors can lead to a bifurcation within the legal system, where parts operate close to international standards and other parts fail to meet minimal requirements of fairness (Fuller, 1969).

Courts

Authoritarian rulers need and use law to govern. As a result, they often end up empowering courts and judges by giving them some jurisdiction that matters (for example, political crime, administrative justice, regulation of business, and even constitutional questions). But authoritarian rulers do not easily yield power, and they may feel the need to ensure that judges deliver the 'correct' decisions in cases that matter to them. At the same time, to secure legitimacy, rulers need courts that have (or appear to have) a modicum of autonomy, if not some actual independence – that is, freedom from external influences. In short, authoritarian rulers experience tension between the need for some judicial autonomy and their desire to control the outcomes of cases.

Models of Judicial Power and Independence in Authoritarian States

In authoritarian states, past and present, one finds four common patterns or models of judicial power and independence, and sometimes combinations of them (Solomon, 2007). The first pattern, encountered in the USSR throughout most of its history, features both courts that were politically marginal (that is, given jurisdiction over a minimum of politically sensitive questions) and judges who were not independent (they faced multiple lines of dependency). Courts dealt with disputes among private persons and ordinary criminal matters, but not constitutional matters; they heard only minimal administrative cases, adjudicated hardly any business disputes (which were instead placed in the hands of executive tribunals), and until 1960 handled relatively few cases of political crimes, most of which were handled either by special bodies or military tribunals. Judges were dependent upon both political authorities at the same level of government and superiors within the courts. They faced regular evaluation and continuous need for reappointment (Solomon, 1996; Solomon and Foglesong, 2000).

The second model is what I call the Spanish solution, or a 'fragmented judicial system'. In Spain in the late Franco years the judiciary as a whole was independent; it had the normal set of institutional protections but its members lacked power or jurisdiction on all matters of interest to the government. These matters were placed in the care of a separate set of tribunals, whose adjudicators lacked the protections supplied to judges (such as life appointment and firing only for real causes). Judges on courts were part of a well-organized judicial bureaucracy where regular evaluations encouraged conformist behavior, but they did not suffer from outside interventions. This model may well embrace Salazar's Portugal and Mussolini's Italy, and it corresponds to a degree to Nazi Germany before World War II. In fact, it was in describing Germany that Ernst Frankel developed the concept of 'dual state' whereby some spheres of public life remained subject to traditional legal norms and others were politicized. The fragmented judicial system is especially common in states that have had a strong judicial bureaucracy and/or legal tradition (Toharia, 1975; Fraenkel, 1941).

Third come courts that are relatively independent and have politically meaningful jurisdiction created by authoritarian rulers. This model tends to generate tension or conflict, for when judges fail to support regime interests, they may face consequences. Wilhelmian Germany after the reforms of 1878 is one example (Ledford, 1993). Another is Tsarist Russia after the Judicial Reform of 1864, a place where judges (sometimes with the help of juries) made decisions that were intolerable to the Tsar and his circle, including the famous acquittal of a young female terrorist who shot and wounded the Governor of St. Petersburg. Not surprisingly, adjustments were made, including the removal of politically sensitive crimes to military courts, the declaration of regional states of emergency, and the cultivation of cooperative judges through increasing management of their careers (Solomon, 2008a).

Lastly, there are countries where courts are formally independent and empowered, but informal practices ensure that judges do not rule against regime interests. This category seems to include Singapore, which has convinced international business that it has a sound legal system, even though judges pay heed to regime interests, especially when political opposition is involved (Silverstein, 2008; Rajah, 2012). To a considerable degree, the laws (for example of libel) are written to facilitate this image. Another country that fits this model is post-Soviet Russia under Vladimir Putin. In this particular authoritarian state, courts have considerable autonomy and power, but informal practices ensure that Russian judges normally do the bidding of powerful persons in cases that matter to them. They do this sometimes for money, more often to keep the heads of their courts content with their behavior and to avoid any attempts to have them dismissed (Solomon, 2008a).

In Russia and many other post-Soviet states, the awarding of acquittals to the accused in criminal cases is discouraged by special reviews of the judges who give them and their frequent reversals on appeal, a negative in the evaluation of judicial performance and sometimes grounds for dismissal. In civil cases involving business, one side often has more political clout than the other, and the trial judge must pay heed to this reality to avoid conflict with the chief judge. Formally no judge may be dismissed except for cause (ethical breach, truly weak performance) and after review by his peers. In reality, the chairs of courts have sufficient leverage over the commissions charged with the discipline of judges that they can engineer the dismissal of any judge on a pretext (Solomon, 2010; Trochev, unpublished paper).

The first and second models (weak/dependent courts; fragmented judicial system) are both stable and straightforward;

the third (supposedly independent and empowered courts in authoritarian context) is unstable, featuring a constrained judiciary that faces threats to its independence and power. The fourth (courts that are independent formally but not in practice) involves a gap between formal institutions and reality, which may engender public cynicism and mistrust of the courts. It is also possible to envisage situations where there might be combinations of these models.

International Influence: Leverage at Work

Complex configurations of the relationship between law and courts and political power are becoming more common in the wake of international pressures that impinge on authoritarian leaders. In the late twentieth and the twenty-first centuries, few authoritarian states can survive without close ties to other countries, at least in their own regions if not further afield. And with those ties (for example, in the realm of business) go expectations about the nature of laws and the protection of the property of investors, if not also about human rights and the fairness of adjudication more generally. The closer the business ties with other countries or international organizations, the more leverage the latter can exert on authoritarian leaders so that they feel compelled to provide at least the appearance of laws and legal institutions that meet world standards.

The example of Singapore is well known, a country that has gone out of its way to court international business with promises of legal protections and a discourse about law that has earned praise for the country's commitment to law. At the same time, the administration of criminal justice, especially vis-a-vis rivals for power in various spheres, be they religious, journalistic, or groups of lawyers, is harsh and arbitrary. Through the criminal justice system Singapore's leaders have created an illiberal and repressive political order. To some observers, this contrast between the role of law in business and in criminal justice resembles the model of fragmented legal system, or a version of a dual state, where legality infuses some, but not all, areas of public life, and might be classified as 'authoritarian ROL' (Rajah, 2012). Another authoritarian state that seems to be experiencing a bifurcation in the state of its law and courts is China (mainland or communist China). Like Singapore, China has gone out of its way to transform its law relating to international business, trade, and investment. Many of these changes were either preparations for or consequences of China's accession to the World Trade Organization (WTO) in December 2001. Joining the WTO imposed obligations upon China to produce a liberal trading environment while protecting the rights of others (for example, in intellectual property rights). This entailed massive rewriting of domestic legislation on many subjects. At the same time, implementation of some new laws, especially regarding intellectual property, has proven slow and difficult. Intellectual property (IP) violations remain rampant; even the US government is critical of the mechanisms used by China to police property rights. At the same time, foreign businesspersons in China can face criminal prosecutions for breaches of legal norms that in most other countries would lead only to civil suits (Ya Qin, 2007; Bath, 2011).

While making concessions vis-a-vis the law of trade, customs, and intellectual property, China has maintained its authoritarian approach to law as an instrument of social control. China combines elements of a dynamic capitalist economy with an authoritarian and unreformed partocratic state. While its courts and judges remain even in formal terms far from independent and fully 'embedded' in the political order, under outside pressure China has engaged in elements of legal and judicial reform, aimed in the words of one observer at making courts "display a rational and professional outlook in the Western sense" (Liu, 2006). Yet the persistence and further flowering of informal practices ensures that judges pay heed to the interests of political authorities and conform to the norms of their peers. The situation of defense counsel is no better, as they are expected in practice to give precedence to cooperation with law enforcement over conducting a meaningful defense (Solomon, 2010).

At least three informal practices ensure that state interests predominate within the administration of criminal justice in China. First, individual judges (whatever protections they may have) rarely render decisions, especially in harder cases, which are left formally to three judge panels, whose members then informally consult with the head and senior judges of the court and, if necessary, judges on a higher court. Secondly, both the conviction and sentencing of offenders are influenced by periodic campaigns against particular kinds of crime. Thirdly, defense counsel who insisted on defending clients conscientiously and going beyond their conventional supporting role face the possibility of persecution by police, including beatings, illegal detentions, and even the opening of criminal cases against them (He, 2012; Solomon, 2010).

When handling civil disputes involving officials, Chinese judges readily change their policies to suit political needs. Legal reforms in the 1990s encouraged citizens to challenge the legality of official actions and thereby make judges key decision makers. But when too many petitions reached higher courts, authorities tried to prevent any threat to social stability by pressing trial court judges to achieve mediated outcomes that could not be appealed (Minzer, 2011).

In short, in China outside pressures to provide an adequate legal environment for business pushed authoritarian rulers to adopt a blend of a fragmented legal system and one governed by informal practices.

In contrast to China, the main outside influence on the administration of justice in Russia derived not from international business norms but from the global promotion of human rights. Because of its geographical location and its first leaders' desire to demonstrate democratic credentials, in 1996 the Russian Federation joined the Council of Europe, an institution devoted to making its members observe the European Convention on Human Rights. It is also equipped with a court to enforce its provisions: the European Court of Human Rights receives and rules on complaints coming from citizens in its member states and makes decisions that call for state authorities to both pay compensation and address the underlying problems that produce rights violations.

Over the years residents of the Russian Federation have brought thousands of cases to the European Court of Human Rights (ECHR), including complaints about violations associated with criminal procedure. Hundreds have been heard by the court and for the most part, rulings have been against the government of Russia. The government has usually paid the

fines to the petitioners, but it has not fulfilled the second part of its obligations, failing to take effective measures to reduce future violations. Russia continues to have the largest absolute number of complaints registered at the Court, although other countries (e.g., Ukraine and Turkey) may have the similar rates per capita population (Trochev, 2009).

This continuing saga has frustrated Russian leaders, at a minimum embarrassing them, at worst angering them. Why should a sovereign state representing a former great power subject itself to assessment by representatives of the international community? By Putin's second term in office (2005) Russia's leaders began having the decisions of the ECHR translated into Russian and circulated to most judges, in the hope that this would reduce violations and complaints. This did not happen; the number of complaints kept growing, in part because of the presence of NGOs that assisted Russians in bringing claims to the ECHR. For that reason, Russian leaders created an alternative domestic procedure for certain kinds of complaints: in April 2010, Russian courts were given the power to hear claims for compensation for violation of a party's right to a trial, and to provide funds for compensation. (A substantial share of the complaints to the ECHR coming from the Russian Federation had involved violations of this right.)

While the new procedure may have reduced the flow of cases to the ECHR by as much as 25%, it did not address the embarrassment felt by some Russian jurists and politicians of having their country's practices subject to criticism by a foreign body. The festering resentment boiled over during 2011. After the European Court of Human Rights dared to criticize not just the practice of an ordinary Russian agency or court but a decision of the Constitutional Court itself, its chair Valery Zorkin made the daring proposal that Russia and its Constitutional Court should be guided by its own understanding of the requirements of the Russian Constitution, rather than by the ECHR. Zorkin cited a decision from the German Constitutional Court stating that the lawmaker must have the right to avoid international legal obligations that would violate its own constitution (and therefore sovereignty). Zorkin's words aroused stern reactions from other Russian jurists, who understood that this approach violated Russia's treaty obligations. Not long after, the acting chairman of the Federation Council 'Senator' Torshin proposed legislation that would eliminate the priority Russian law gives to the ECHR, notwithstanding the fact that this would violate the Russian Constitution and expose Russia to possible expulsion from the Council of Europe. Torshin's draft was not adopted (Solomon, 2013).

The story of Russia's experience with the European Court of Human Rights constitutes an unusual instance of authoritarian rulers' confronting the exercise of power by a court outside their realm and resenting its exercise by that court. This ironic situation represents another example of how the development of global standards of justice along with institutions to enforce them could place burdens on and limit the power of authoritarian as well as democratic rulers.

Authoritarian 'ROL'?

One issue that has arisen in the study of law and courts in authoritarian states is whether one can or should speak of ROL in authoritarian states, at least with adjectives attached. In serious scholarship there is a distinction between a minimal, thin, or formal variety of ROL and the deep, thick, or substantive one. The thin or formal versions call for basic, but far from inevitable, features including: open and legitimate procedures for law making; consistent legal hierarchy (regulations subordinate to and implementing laws); transparency of laws and regulations; laws that are prospective, clear, consistent, and stable (if not also fair); and enforcement of the laws that is consistently just and impartial. The thick or substantive version of the ROL adds an important element to the minimal version: the law must also have a moral basis and be based on a normative foundation beyond the state and its laws, such as religion, natural law, or the international law of human rights. In the thick version of ROL, pravo (jus) consists of more than the laws (leges). It features legal principles and rights that the leaders of the state may not violate or contradict and that take precedence over their legal enactments. This implies that some Court or courts have the responsibility (and power) to decide when laws are not lawful (Tamanaha, 2004; Kleinfeld, 2006; Fuller, 1969).

As we have seen, authoritarian rulers are typically reluctant to yield major power to courts, and to my knowledge no authoritarian regime to date has met the standards for thick or substantive ROL. (None of the four models of courts and political power implies this). Nor have most of them approached meeting all the demands of thin or formal ROL, and it is not surprising that the editors of a pioneering book on courts under authoritarianism chose the title "Rule by Law". However, it may be theoretically possible for an authoritarian regime to approximate the requirements of thin ROL. Arguably, law and courts in Wilhemian Germany on the eve of World War I came close. The prevailing theory of *rechtsstaat* was consistent with thin ROL, and the forces of liberalism in Germany, including the legal profession, did what they could to promote it. Whether even the thin version of ROL could last for a long time in an authoritarian regime is far from clear, for there would be inevitable conflicts that could lead to either movement away from authoritarianism or from law.

For some scholars who study non-Western societies with authoritarian regimes, the misfit between ROL and normal politics seems to be based, at least in part, on cultural biases. Thus, the thick or substantive version of ROL, which prescribes limits on the power of rulers to violate individual rights, may well reflect values that are central in Western civilization but less important in other cultures. If one accepts this premise, then it may follow that one should not treat ROL as defined by Western scholars, in either version, as the only acceptable standard for non-Western countries, especially those with authoritarian regimes. For example, what should one call a state like Singapore that seems to meet the standards of thin ROL in the business sphere but not in matters of social control? To Jothie Rajah, one needs the designation 'authoritarian ROL' (Rajah, 2012). In an effort to identify models that have empirical purchase, Randall Peerenboom pictures three lesser versions of ROL in between full or liberal ROL and rule by law. He calls them communitarian ROL; neo-authoritarian ROL; and statist socialist ROL. Two of his qualified models of ROL include some limits of the power of the rulers (central to the spirit of ROL), while one does not. Two of them place collective

rights ahead of individual ones; and one gives more emphasis to strengthening the state. Peerenboom spells out each of the five models in considerable detail, so that other scholars can judge their utility (Peerenboom, 2002).

Variations among Authoritarian States: Competition and Informal Practices

Another issue in the law–authoritarianism relationship is explaining variations in the role of law and courts in authoritarian states. A number of factors might correlate with particular models of courts in authoritarian states, including: the age of a regime; its stability; the identity of the ruling group (individual, military oligarchy, a political party), the amount of competition it tolerates, and the strength of the state (or its formal institutions).

The factor that has received the most attention from analysts thus far is competition, especially the emergence within authoritarian states of new elite groups or more intense competition among elites. As I have detailed elsewhere, a series of studies has correlated judicial empowerment with political competition, in countries within Latin America and Asia, as well as within Europe centuries ago, although the relationship is not universal (Solomon, 2007; Trochev, 2010). Moreover, when the power of the ruler or ruling group is less certain, judges are more likely to resist fulfilling their demands, whether through the pursuit of a narrow positivist reasoning, strategic defection, or off bench political activity (Osiel, 1995; Helmke, 2004; Hilbink, 2007; Trochev and Ellett, unpublished paper).

Without doing a rigorous comparative study, one can still construct hypotheses about the other factors. Rule by a political party and/or with a discrete ideology may well connect to the first model of courts, where rulers keep full control over courts and do not empower them. Personal dictatorships, especially in countries that already had an autonomous judiciary, seem more likely to adopt the Spanish solution of a fragmented judicial system. This was the situation in Portugal under Salazar and Italy under Mussolini. The third model, in which autonomous legal institutions coexist with authoritarian rule and generate tension and conflict, is characteristic of older, stable regimes where there is some (liberalizing) group pushing for reform (Halliday and Karpik, 1997).

But it is the fourth model, where rulers create formal institutions that promise judicial autonomy but allow informal practices to undermine it, that may prove most in accord with the realities of the twenty-first century and most likely to gain prominence in the future. This approach is consistent with weak formal institutions, a characteristic of many authoritarian states in the developing world and post-Soviet space. It also fits with high levels of clientelism and corruption, another common (though far from universal) characteristic of authoritarian states (Zaloznaya, unpublished paper). At the same time, this model has the advantage of allowing authoritarian rulers to cultivate good reputations and public relations while retaining control over the administration of justice when needed. Remaining to be answered are such challenging questions as when and why many authoritarian states are weak, why they feature large doses of clientelism, and whether these conditions have a consistent impact on law and courts.

Conclusion

This analysis of the role of law and courts in authoritarian states and regimes underscores both their utility to authoritarian rulers as an instrument of rule (ruling by law) and the likelihood that the actual use of law and courts will make them yield some political power. In addition to needing law as a source of legitimation at home, authoritarian leaders often faced international pressures, if only to encourage trade and investment. It is difficult for authoritarian leaders to find a comfortable balance between their aspirations for control and the need to allow courts to rule against their interests, and most of the solutions to this dilemma involve contradictions and conflict. As a result, the administration of justice tends to be less stable and predictable in authoritarian states than in democratic ones. Rarely does it meet the standards of ROL even in its least demanding form. Rule by law remains the best description of its role.

Acknowledgment

I am pleased to thank Tom Ginsburg, Rosann Greenspan, Lucan Way, and Jurgen Zarusky for helpful comments on an earlier draft, and Milosz Zak for research assistance.

See also: Administrative Law: The US and Beyond; Attitudes Political: Authoritarianism and Toleranced; Constitutional Courts; Constitutions; Courts and Adjudication; Human Rights: Legal and Sociolegal Perspectives; Judicial Independence; Law as an Instrument of Social Change; Legal Positivism; Natural Law; Nazi Law; Political Trials; Rule of Law (and Rechtsstaat); Socialist Law.

Bibliography

Bath, V., 2011. China, international business, and the criminal law. Asian-Pacific Law and Policy Journal 131 (1), 1–35.

Berman, H.J., 2003. Law and Revolution II: The Impact of the Protestant Reformations on the Western Legal Tradition. Harvard University Press, Cambridge.

Darnden, K., 2008. The integrity of corrupt states: graft as an informal state institution. Politics & Society 36 (1), 35–60.

Fraenkel, E., 1941. The Dual State: A Contribution to the Theory of Dictatorship (E. Shils, Trans.). Octagon Books, New York.

Fuller, L.L., 1969. The Morality of Law, revised ed. Yale University Press, New Haven.

Ginsburg, T., Moustafa, T. (Eds.), 2008. Rule by Law. Cambridge University Press, New York.

Ginsburg, T., Simpser, A. Constitutions in Authoritarian Regimes, unpublished book manuscript.

Halliday, T.C., Karpik, C., 1997. Lawyers and the Rise of Western Political Liberalism. Europe and North America from the Eighteenth to Twentieth Centuries. Clarendon Press, Oxford.

He, X., 2012. Black hole of responsibility: the adjudication committee's role in a Chinese court. Law & Society Review 46 (4), 681–712.

Helmke, G., 2004. Courts under Constraints: Judges, Generals and Presidents in Argentina. Cambridge University Press, New York.

Hilbink, L., 2007. Judges beyond Politics in Democracy and Dictatorship: Lessons from Chile. Cambridge University Press, New York.

Kleinfeld, R., 2006. Competing definitions of the rule of law. In: Carothers, T. (Ed.), Promoting the Rule of Law Abroad: In Search of Knowledge. Carnegie Endowment for International Peace, Washington, pp. 31–74.

Ledford, K.F., 1993. Lawyers, liberalism, and procedure: the German imperial justice laws of 1877–79. Central European History 26 (2), 165–193.

Ledford, K.F., 2004. Formalizing the rule of law in Prussia: the supreme administrative law court, 1876–1914. Central European History 37 (2), 203–2224.

Levitsky, S., Way, L., 2010. Competitive Authoritarianism: Hybrid Regimes After the Cold War. Cambridge University Press, Cambridge and New York.

Liu, S., 2006. Beyond global convergence: conflicts of legitimacy in a Chinese lower court. Law and Social Inquiry 31 (1), 75–106.

Magaloni, B., 2008. Enforcing the autocratic political order and the role of courts. In: Ginsburg, Moustafa (Eds.), Rule by Law: The Politics of Courts in Authoritarian Regimes. Cambridge University Press, New York, pp. 180–206.

Minzer, C.F., 2011. China's turn against the law. American Journal of Comparative Law 59, 935–984.

Moustafa, T., 2007. The Struggle for Constitutional Power: Law, Politics and Economic Development in Egypt. Cambridge University Press, Cambridge.

Osiel, M.J., 1995. Dialogue with dictators: judicial resistance in Argentina and Brazil. Law and Social Inquiry 20 (2, Spring), 481–560.

Peerenboom, R., 2002. China's Long March toward Rule of Law. Cambridge University Press, Cambridge.

Pereira, A.W., 2005. Political (In)justice: Authoritarianism and the Rule of Law in Brazil, Chile and Argentina. University of Pittsburgh Press, Pittsburgh, PA.

Rajah, J., 2012. Authoritarian Rule of Law: Legislation, Discourse, and Legitimacy in Singapore. Cambridge University Press, New York.

Silverstein, G., 2008. Singapore: the exception that proves rules matters. In: Ginsburg, T., Moustafa, T. (Eds.), Rule by Law: The Politics of Courts in Authoritarian Regimes. Cambridge University Press, New York, pp. 73–101.

Solomon Jr., P.H., 1996. Soviet Criminal Justice under Stalin. Cambridge University Press, New York.

Solomon, Jr., P.H., 2004. Judicial power in Russia: through the prism of administrative justice. Law and Society Review, 38 (3), 549–580.

Solomon Jr., P.H., October 2007. Courts and judges in authoritarian regimes. World Politics 60 (1), 122–145.

Solomon Jr., P.H., March 2008. Law in public administration: how Russia differs. The Journal of Communist Studies and Transition Politics 24 (1), 15–35.

Solomon Jr., P.H., 2008a. Judicial power in authoritarian states: the Russian experience. In: Ginsburg, Moustafa (Eds.), Rule by Law: The Politics of Courts in Authoritarian Regimes. Cambridge University Press, New York, pp. 261–283.

Solomon Jr., P.H., 2010. Authoritarian legality and informal practices: judges, lawyers and the state in Russia and China. Communist and Post-Communist Studies 43 (4), 351–362.

Solomon Jr., P.H., 2013. Courts, law and policing under Medvedev: many reforms, modest change, new voices. In: Black, J.L., Johns, M. (Eds.), Russia after 2012: From Putin to Medvedev to Putin—continuity, Change, or Revolution?, pp. 19–41.

Solomon, Jr., P.H. Putin's New Presidency and the Criminal Law, unpublished.

Solomon Jr., P.H., Foglesong, T., 2000. Courts and Transition in Russia: The Challenge of Judicial Reform. Westview, Boulder, CO.

Svolik, M.W., 2012. The Politics of Authoritarian Rule. Cambridge University Press, New York.

Tamanaha, B.Z., 2004. On the Rule of Law: History, Politics, Theory. Cambridge University Press, Cambridge.

Toharia, J.J., 1975. Judicial independence in an authoritarian regime: the case of contemporary Spain. Law and Society Review 9 (3, Spring), 475–496.

Trochev, A., 2008. Judging Russia: The Role of the Constitutional Court in Russian Politics, 1990–2006. Cambridge University Press, New York.

Trochev, A., 2009. All appeals lead to strasbourg? unpacking the impact of the European court of human rights in Russia. Demokratizatsiia 17 (2, Spring), 145–178.

Trochev, A., 2010. Meddling with justice: competitive politics, impunity, and distrusted courts in post-Orange Ukraine. Demokratizatsiia 18 (2), 18–34.

Trochev, A. How Judges Arrest and Acquit: Soviet Legacies in Post-Communist Criminal Justice, unpublished paper.

Trochev, A., Ellett, R. Rethinking Judicial Autonomy: Off-Bench Resistance against Blatant Interference, unpublished paper.

Ya Qin, J., 2007. The impact of WTO accession on China's legal system: trade, investment, and beyond. Wayne State University Law School Legal Studies Research Paper Series 07–15.

Zaloznaya, M. Does Authoritarianism Breed Corruption? Reconsidering the Relationship between Authoritarian Governance and Extra-Legal Behavior in Bureaucracies, unpublished paper.

Law and Democracy

Grażyna Skąpska, Kraków, Poland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The complex and dynamic relationship between law and democracy has been a subject of interest since ancient times. In the contemporary world, the principle of democratic rule of law and the concept of law's legitimacy express the reciprocal and circular links between law and democracy. However, this relationship is not symmetrical. The fundamental task of law consists in controlling and even 'taming' democratic government against the possible misuse of power by the political majority. The relationship between law and democracy is increasingly problematic in all relevant dimensions: democratic participation in legislation, social and political subordination to law, and legal control of democratic government in order to protect rights and liberties of citizens. Contemporary challenges have been answered by new developments characterizing law: the growth in the power of the judiciary and the empowerment of citizens. The latter takes the form of the broad interpretation of the nondiscrimination principle and the protection of individual, and more recently also collective, rights. Therefore one can speak about the broadening and deepening of democracy. These new developments have been enhanced by the collapse of authoritarian and totalitarian regimes in Europe and elsewhere. However, one observes new phenomena consisting in the growing importance of subnational and international processes and actors. They present a new challenge to both democracy and law, and result in the fragmentation of relations between them.

The Relationship between Law and Democracy

The relationship between law and democracy is complex and dynamic. In general terms, democracy can be defined as rule by the *demos* ('the people,' i.e., by equal human beings who resolve their conflicts on the basis of procedures and laws agreed upon by themselves), and the law as more or less systematically ordered, formalized, and rational rules of conduct backed by the threat of force by those in political power when these rules are violated. What is decisive in the deep links between democracy and the law is the fact that the *demo*s – 'the people' – is the political source of the law to which it, in turn, subordinates itself. This takes the succinct form of constitutional principle of the democratic rule of law, present in a direct or indirect form in every liberal democratic constitution. In the contemporary world, the principle of democratic rule of law on the one hand, and the concept of law's legitimacy in democratic, liberal societies on the other, express the fundamental and reciprocal links between democracy and law. Thus, the interconnections between democracy and law are circular and twofold: the law legitimates the political power of democratic government, and democratic government presents the source of law, and makes it legitimate.

In Niklas Luhmann's autopoietic systems theory, these connections are reflected in the democratic liberal constitution as the highest binding law whose source is the democratic will of the people. The constitution establishes a 'structural coupling' between two differentiated subsystems (or partial systems) of society: the legal and political (Luhmann, 2004).

However, the democratic rule of law principle that expresses the relationship between law and democracy in contemporary societies implies that the position of law is higher and stronger. It is a task of law to protect not only its own goals, but also to protect democracy against its own 'vices,' against the untamed struggle of political interests, ambitions, wars of political factions. Therefore in the modern world courts of final instance increasingly decide not only legal, but also politically sensitive, matters.

These issues are becoming increasingly complex, and the concept of the democratic rule of law, as well as the idea of the democratic legitimacy of law, are undergoing changes inspired by the struggles for the protection of rights by individual and collective actors ever more conscious of their subjectivity. Thus, in modern liberal societies it is important to devise systems of autonomous and general rules to enable the protection of individuals and social groups struggling for their rights and the fulfillment of their interests against the misuse of power by democratic majorities, and to enable them to participate in the political decision-making process. In turn, the mutual links between democracy and law modified by the protection of rights have fundamentally important consequences for the legitimacy of law.

The relationship between democracy and law is further challenged by recent political, economic, and social phenomena: the emergence of supranational political organizations, like the European Union, the rapid growth of internationally active transnational corporations and nonstate organizations, and the growing importance of subnational, regional, and even local governments, associations, and organizations of local interests. To this, one should add the growing cultural pluralism of the modern world, which implies differentiated conceptualizations and interpretations of basic rights. Furthermore, the democratic rule of law debate is reinvigorated by the collapse of authoritarian and totalitarian regimes. These developments are inspiring new conceptualizations of government and law. They present a challenge to the classical democratic rule of law concept, and call into question the concepts of the political legitimacy of law and the legal legitimacy of government. The concept of *governance* rather than government has emerged on the agenda of political scientists, referring to the phenomenon of multilevel and complex systems of decision making, with participation by nonstate actors. New types of laws are debated by legal scholars, politicians, and lawyers. One speaks about the emergence of a third, and even a fourth legal order: a legal order of supranational associations of states, and a legal order that emerges in

form of agreements or regulations outside the state framework, affecting activities of supra-, inter-, and substate public, half-public, and private actors, across state boundaries and jurisdictions. Cultural pluralism and the differentiation of interpretations of the protection of human rights contributes to the debate on the democratic rule of law as well.

Important questions are emerging regarding the possible democratic deficit of these complex systems of governance on the regional (as in the case of the European Union) or even global level, and consequently, the possible deficit of the democratic legitimacy of nonstate law. Therefore, although the relationship between democracy and law has been debated in the West since ancient times, worldwide contemporary processes of globalization, the growing recognition of cultural pluralism, and the increasing structural complexity of the modern world greatly contribute to this debate, introducing new topics and new perspectives. Of particular importance as already mentioned, was the collapse of totalitarian and authoritarian regimes in the twentieth century in Europe and elsewhere. This article tries to identify the areas of common interest and disagreement in the debates on democracy and law, to summarize classical arguments on the reciprocal relationships between them, to shed some light on new incentives for the democracy and law debate, which result from the collapse of totalitarian an authoritarian regimes, and to reflect on the contemporary issues and challenges to classical concepts and institutional arrangements.

Areas of Common Interest and Disagreement

Many scholars and politicians, ancient, modern, and contemporary, have maintained that democracy, being a political source of law, cannot exist without law, since it would otherwise turn into anarchy. It is also argued that democracy represents a fundamental factor in the development of the law, since it has the unique potential to reflect and react to changing political and social reality. Democracy also contributes to the validity and legitimacy of the law: the law holds society together through the power of enforcement, but the legitimacy of political coercion derives from democratic legislative procedures. It must be stressed, however, that questions have been raised since ancient times about the possible misuse of law by a democratic power, in the form of a threat posed to individual sovereignty and creativity by the majority represented by the government. Presently, the relationship between majoritarian democracy as a source of law is critically reviewed from the perspective not only of individual, inviolable rights, but also of collective rights related to gender, and ethnic, racial, or religious minorities. This theoretical debate is translated into debate on the specific legal arrangements 'taming' democracy, on quotas or thresholds, i.e., on the visibility and 'voice' of minorities in democratic election and the lawmaking process, on the sensitivity and reflexivity of democratic legislation of the life experiences and lifeworlds of those, whose 'voice' could be silenced by the majority. New interpretations of such principles as nondiscrimination and equal participation and representation, equal treatment, or equal protection of rights appear which try to encompass different lifestyles, lifeworlds, and life experiences on the one hand, yet are rooted in 'local knowledge' – socially experienced interpretations of equal participation or exclusion – on the other.

In contrast to the flexibility and reflexivity of democracy as a political system, a certain 'gravity' is one of the law's characteristic features. Therefore, from the standpoint of the law, its relationship to democracy was never perceived as unquestionable or nonproblematic. The potential sources of disagreement were identified by David Held (Held, 1987: 2). From the point of view of political science, the primary disagreement lies in the definition of 'the people.' From the perspective of the law, this disagreement is summarized in a question: "Does the 'rule by' entail the obligation to obey?" (Held, 1987: 2). Answer to this question govern the principle of a democratic rule of law, or of a democratic state founded on the rule of law (democratic *Rechtsstaat*).

Positions taken in debates on these problems depend upon the model or variant of democracy, and on the type of society – whether it is a relatively simple, premodern society of the ancient city-state type, a functionally differentiated modern society, or a complex and pluralistic, culturally differentiated society characteristic of late- or postmodernity. There has been a process of broadening and deepening of the democracy concept – i.e., the broadening concept of rule by 'the people' that includes an ever-greater part of the society in the form of a popular franchise, and the deepening of a concept of the democratic state founded on the rule of law. Another related, controversial issues regards direct democracy in the form of referendum as a possible source of law.

At the beginning of the twenty-first century further new issues broaden the areas of interest and disagreement. The burning question emerges concerning *demos*, who are to be defined as 'the people' in the complex modern world characterized by overlapping organizational networks and competing loyalties that cross state boundaries. There appear also questions of the legitimacy of nonstate law, and the obligation to obey differentiated and complex laws issued on many levels of social organization and agreed upon by many procedures, not necessarily democratic in the classical sense.

Democracy and Law in Classic Debates: From the Ancient *Polis* to Liberal Democracy

Thinking about democracy and law in the West has been shaped by the issues and questions raised by ancient philosophers and politicians with reference to the aims, ideals, and the constitution of the *polis*, the ancient Greek city-state. These issues and questions do not just present a source of modern thinking about the relationship between law and democracy. They were also reflected in institutional arrangements, characteristic of liberal democratic regimes.

Three such issues are most relevant for deliberations on democracy and law: the civic and political culture (civic virtues); the rule of law, due process, and constitutional government; and the idea of superior argument in lawmaking and its enforcement. The latter idea demonstrates a foundation of the most influential contemporary theory on law and democracy, i.e., the theory of procedural democracy.

Among the ancient thinkers and politicians, one should mention Pericles, Aristotle, and Plato. Pericles emphasized that

respect for the law results from the lawmaking process by people who simultaneously subject themselves to an obligation to obey the law. The citizen's law, and respect for such law, lies in contrast to the arbitrary will of tyranny and subjugation to it through the use of brute force.

In his *Politics*, Aristotle (384–322 BC) observed that, if liberty is the first defining principle in the model of the democratic city-state, it had to be closely related to an equality of opportunity to hold office: only equal participation in legislation protects the freedoms and liberties of citizens. Therefore, liberal democracy relies on the possibility that every citizen has an equal opportunity to rule, which also justifies obedience to laws made by a democratic majority, even if free citizens want to live 'as they like' (Aristotle, 1981: 195–198).

On the other hand, the most important criticism of democratic rule was raised by Plato. In one of the famous dialogues – cited by sociologist of law Vilhelm Aubert in his book *In Search of Law* – Plato (427–347 BC) challenges democratic rule by raising the question: Why is the tyranny of the majority better than that of an oligarchy or of an individual? This question returns in every debate on the issue of majority democratic institutions protecting individuals and minorities from a potential majority abuse of power (Aubert, 1983: 6, 7).

Despite the intellectual potential and contemporary relevance of arguments raised by ancient thinkers, the institutional forms and ideas linked with the ancient *polis* have their obvious limitations. First and foremost, equality and freedom were restricted to those inhabitants of the city-state who were regarded as citizens, i.e., adult, male, and Athenian born. Women, immigrants and, above all, slaves, were excluded from this category. These ideas lacked the modern liberal fundamental conviction that all human beings are 'individuals,' and bearers of inalienable rights. Ultimately, their empirical terms of reference were based on a relatively small and structurally simple, homogeneous society whose economy was predominantly based on slave labor, and whose members could directly participate in legislative and juridical proceedings because of the small numbers involved and their affinity with one another. Nevertheless, the issues raised by ancient thinkers and politicians have set a framework for debate, as well as an ideal for institutional arrangements in modern times. In the contemporary debates references to these ancient issues continue to be made with regard to less authoritarian, more horizontal, and communicative models of political decision- and lawmaking procedures.

Questions and arguments concerning the relationship between democracy and the law have been especially important under conditions of rapidly growing social differentiation, of changes in social integration accompanied by diffusion of ownership, of a developing free-market economy involving the formation of a relatively broad class of producers and merchants on the one hand, and the growing complexity and centralization of an impersonal, bureaucratic state apparatus on the other. These conditions characterize the processes of social modernization and the formation of liberal democracy as known in the West. These questions and arguments are reflected in basic institutional arrangements, predominantly in the constitutions of contemporary liberal democracies.

There are two important issues connecting law and democracy within the set of institutional arrangements defined as liberal democracy: conflicting interests and the protection of individual rights.

In a rapidly changing, diffused society characterized by conflicting interests, relations between citizens and the state are reshaping the concept of the social contract. They require specific forms of law – through which compromises among conflicting interests can be struck. Moreover, they require a law with attributes of generality and autonomy in order to reflect changing social values and power relations. In the words of Roberto Mangabeira Unger, in liberal societies the law somehow accommodates antagonistic interests, and its procedures are such that anyone might find it in his or her own interests to subscribe to them regardless of the ends one happens to be seeking. Therefore, "The liberal type of social organization generates, and is reinforced by, a style of consciousness whose substance is the image of society as an arena of conflicting subjective interests" (Unger, 1976: 68–69).

Ideas concerning the notion of social contract as a foundation for political institutional arrangements, which accommodate antagonistic interests, and the delegation of legal sovereignty crucial for the validity and legitimacy of law and its observance are to be found in the writings of Thomas Hobbes (1588–1679). These ideas were developed and supplemented by John Locke (1632–1704), Charles Louis de Secondat, Baron de Montesquieu (1689–1755), Jean Jacques Rousseau (1712–1778), James Madison (1751–1836), Jeremy Bentham (1748–1832), James Mill (1773–1836), and John Stuart Mill (1806–1873).

Generally, according to proponents of the classic model of liberal democracy, the activities of a democratic government are to be limited and controlled, since the sole aim of the state is to preserve the 'life, liberty, and estate' of its citizens (Locke, *Two Treatises on Government*), or, as emphasized by John Stuart Mill, to protect the free development of the individuality of every citizen. This is possible in the form of a limited, constitutional government, subordinate to the law, whose activities are also subject to the mutual checks and balances of autonomous branches of government. In the highly developed Doctrine of Montesquieu, these are the checks and balances of the legislative, executive, and judiciary powers (Montesquieu, *The Spirit of Laws*).

In the public sphere of relations between citizens and the state, the idea of a contract concluded by free and formally equal citizens, together with the extended (initially to citizens understood as male property owners) franchise, legitimizes the political and legal sovereignty of the government. These ideas are closely linked with the notion of liberty, i.e., freedom within legal limits, and with the notions of natural, universal, and inalienable public rights of citizens whose activities are governed by reason, and who conclude a political contract and participate, directly or through their representatives, in the legislative process.

Classical thinkers located the source and foundation of individual rights and liberties in the idea of natural law as 'higher' than man-made, positive law. Thus they developed a concept of law, which limits political authority, even democratic political authority, with respect to human rights. However, the notion of natural law and its binding impact on positive law executed by a representative government was interpreted in various ways. Natural laws were defined as those discovered by

reason, i.e., by rational citizens acting under conditions of freedom, according to German philosopher Immanuel Kant. They were also understood to be of divine origin (Locke, 1960: 311). For Montesquieu, natural laws discovered by reason were to be interpreted in the context of local customs and traditions (Montesquieu, 1952: 32). The utilitarians Jeremy Bentham and James Mill declared theories of natural laws as the maximization of individual satisfaction whose source was the natural tendency of human beings to pursue individual happiness and interests and to avoid suffering (Bentham, *Fragment on Government*, James Mill, *An Essay on Government*).

One may conclude that with respect to protection of individual rights in the private and public domain, in the classic model of liberal democracy based on the primary importance of protecting individual property rights and the utilitarian maximization of individual profits, the 'haves' are protected first and foremost, and individual goals and interests dominate the constitutional, democratic state. The concept of equality proposed by the classic thinkers, and written into the first liberal constitutions, was therefore initially still very restrictive. Only persons who were property owners and male had a right to vote or to hold public office. Only John Stuart Mill wrote of a right to the development of individuality that made reference to women and minorities, and was connected with a postulated broad franchise (John Stuart Mill, *On Liberty*).

Different ideas about contract, the legitimacy of the law, civic virtues, civic commitment, and public government were proposed by the first representatives of the republican version of democracy. First of all, the notion of contract was here understood not as a legal fiction, but as a real contract closed by the people in the form of public debate and referendum on the most important public issues. The issue of civic virtue and of civic duties, and the notion of a *citoyen* as an unselfish, committed participant in public affairs was emphasized by Jean Jacques Rousseau. Accompanying this was the idea of a permanent social contract and direct participation in its formation as the foundations of social and political integration, the legitimacy of the law, and its observance. Here the claim to the legitimacy of a legal order could only be gained through the integrative force of the united will of all free and equal citizens. Advocating direct participation in legislation, and defining the equality of citizens as economic independence and material security, Rousseau opted for the model of a discursive democracy and referendum as its institutional form (Rousseau, 1973).

Liberal Democracy and Law: Contemporary Issues

"Liberal democracy was the first great secular religion in the Western history" writes Harold Berman (Berman, 1983: 32). However, large-scale, complex, densely populated, culturally pluralistic, and differentiated modern societies have developed new interest in and discussion about the relationship between democracy and the law. Four issues are salient in the current debates on law and liberal democratic institutional arrangements. The first is posed by concerns with the principle of equality, both formal and actual, which results in changes in the idea of justice, legal reasoning, forms of law and the concept of rights. The second issue is the growing differentiation of law because of its adaptation to the complex and pluralistic, modern democratic society and efforts to include local experiences of collective actors, defined as minorities, in interpreting universal legal standards and principles. Third is the issue of the deeper justification of law, and of the interpretation of rights that could be binding for democratic legislation and law application, similar to how the doctrine of natural law justified the law in the early modern times. Finally, there appear concerns about the 'bottom up' and 'from above' processes of lawmaking, on the possible 'democratic deficit' in the supranational or a local lawmaking, and the overlapping legal systems in the contemporary world. Such concerns are awakened by the new modes of participation in legislation in form of deliberations between many governmental and nongovernmental bodies, and the state- and nonstate actors within the complex system of governance.

One of the main questions raised in modern debates on liberal democracy concerns the concepts of equality and justice – no longer defined in formal, but in substantive terms such as protection of equal opportunity and positive discrimination based on the idea of equal participation guaranteed every citizen by the democratic state. Initially debates on social justice and equal participation in public life, and the role of the state in their protection, pointed to the growing scope and range of state intervention and its consequences for law, on the one hand, and the civic engagement of the democratic electorate on the other. These are especially vivid with regard to the welfare state and concern policy-oriented legal reasoning based on an emphasis upon broad standards of fairness and the social responsibility of the state.

It is argued that the political organization of the democratic welfare state challenges the universality and autonomy of the law, as well as its democratic legitimacy. The law becomes an instrument of state intervention, characterized by a growing number of particular entitlements to definite substantive goals. Authors, such as Niklas Luhmann in Germany, have pointed to the specific consequences for democracy and law as a result of state interventionism. According to Luhmann, political inclusion – i.e., the subordination of an ever-growing range of social relations to the intervention of the political democratic system of the welfare state – imposes new public tasks on the state, which is considered responsible for the well-being and quality of life of its citizens. This results in rising social demands for new standards and new regulations of ever broader and more particular issues, from minimum wage, to the functioning of kindergartens, to the regulation of public places not in the form of citizens' rights, but citizens' claims on the state (Luhmann, 1981: 38–40).

This phenomenon is described as the 'socialization of the law,' on the one hand, and 'juridification' or 'legalization' of social relations on the other, that result from the claims of the democratic electorate directed on state law. These new developments lead to changes in the form of legal regulation and legal intervention, and of relations among the legislative, judicial, and executive branches. The integrity and autonomy of the legal system is challenged by increasingly specific demands for interventionist legal regulation and the necessity of ad hoc reactions to changing social needs. The 'flood' of changing legal regulations comes to be characteristic of interventionist state activity that transform social needs into claims on the state. Moreover, instrumental legislation is produced to an ever-increasing extent

by governmental executive agencies – and not by the democratic legislature – in the form of administrative prescriptions. The reasons for such instrumental regulations are given as concerns for technical efficiency and economic effectivity, in the fulfilling of state tasks: thus, the technical standards on which the regulations are justified strongly limits the public debate on the contents of the law, possibly resulting in the growing passivity of a democratic electorate. Therefore, the legitimacy of the modern, instrumental law depends in an ever-greater degree on the efficiency of the interventionist state. Therefore, not only growing demands for social justice, transformed into ever more particular claims on the state but also the limited public debate present serious challenges to the relationship between law and democracy.

The demands for social justice, characteristic of modern democracy, are presently not only taking the institutional form of interventionist legislation. They are to an ever more visible degree characterized by the stress on the nondiscrimination, and the equal participation principle. Classic liberal approaches to the principles of equality and equal treatment are challenged by citizens for whom formal equal treatment is not acceptable, because of the specific and differentiated social experiences, specific lifeworlds that require particular attention from law. These issues were initially raised by feminist jurisprudence, and also by critical race scholars and critical legal scholars. Equality cannot be reduced to sameness, writes a representative of the feminist critique of law, arguing for accommodation of women experiences into law and policy (Fletcher, 2002: 149).

Thus, according to the feminist critique of liberal law, western liberal legal systems "… reflect the characteristics of their predominantly male architects: the legislators and judges" (Barnett, 1998: 9). In effect, in a liberal democratic society "… the opinion of the majority is precisely what underpins the sexism and inequality which women have traditionally suffered and continue to experience, despite the many formal legal reforms" (Barnett, 1998: 106). More precisely, the feminist critique of liberal law is directed against its 'objectivism' and formalism, its lack of sensitivity to the specific lifeworlds and experiences of women, and also against the dominant, i.e., hegemonic legal discourse, characterized by the conceptualization and interpretation of rights imposed by the male-dominated legislatures and judiciary.

The struggle to incorporate into democratic legislation greater equality, differentiated experiences, sensitivities, and abilities, as well as different worldviews and discourses, has already resulted in the introduction of new institutional arrangements: new standards and principles in form of the quotas regulating the representation of women in the lawmaking bodies, above all in the parliaments, in the redefinition of crimes and procedural rules, in the introduction of new standards and reasons for regulations on pornography, prostitution, or sexual harassment. The reasons on which legislation and law interpretation is based shall express women's experiences, their possible suffering, in order to change the regulations and law application in a democratic, but nevertheless male dominated, society. The new regulations on participation in democratic legislation are based on the concept of empowerment of those who were actually, and not necessarily formally, excluded from the public sphere, like women, racial and ethnic, or religious minorities. These new legal standards and regulations indicate a process of the deepening and broadening of liberal democracy, the reconceptualization of the democratic rule of law principle, and rethinking of the legitimization issue that result from the growing sense of subjectivity by those, whose voice was difficult to hear in the public debate. This, in turn, further leads to the differentiation of legal standards and regulations, and the growing complexity of law in contemporary liberal democracies.

The increasingly more technical legal reasoning, as well as the growing differentiation and complexity of legal standards, and the increasing importance of the executive branch of government are subjects of deep concern with regard to the democratic legitimacy of law and its binding force. There appears a question about the deeper justification of law and the interpretation of legal rule that would be able to encompass different lifeworlds, experiences, lifestyles. Thus, in late-modern democratic society, there are demands for the interpretation of law to be more reflexive, and context sensitive, in order to create a nonfictive 'living law' responding not only to rapid social change, but to social differentiation and the growing public awareness of rights. It is noted that such concepts and principles as the rule of law, human dignity protection, the division of powers, human rights, and definitions of citizenship are far from clear cut, and are not even clearly defined by the law itself. On the other hand, however, they have great symbolic value and, as such, they respond to gaps or inadequacies in the law.

The demands for 'reflexive law' in democratic society are strongly supported by social philosophers who advocate a procedural and communicative rationality-based approach to law, especially constitutional law (Habermas, 1993), and by lawyers who argue that new forms of law are characteristic of the contemporary world. According to these authors, in the globalizing world the constitution of the global social systems relates the law to the particular systems (e.g., economic, scientific, or religious) at issue. Therefore, in line with some arguments raised by Niklas Luhmann, here the constitution is examined in terms of structural coupling between law and another social subsystem as the constitutional object (or the social field the law regulates) (Teubner, 2004a,b, 2012). Such propositions are supported by political scientists, who argue that democracy itself is changing in the contemporary world. Instead of classical arrangements of liberal and majoritarian democracy, new forms of government emerge that include public participation in form of deliberations and engagement of nongovernmental organizations (NGOs), half-private, half-public actors, associations of citizens, employers, and employees in the lawmaking procedures. Another important feature of late modernity is the impact on lawmaking of lobbying groups, NGOs, or spontaneous social movements representing civil society. These new developments seem to be both inevitable and benign, but the issue arises as to how a social contract is possible if one considers the pluralism of values and lifestyles and the increasing differentiation of collective interests characteristic of late modernity. One may plausibly argue that these phenomena, and the struggle of collective interests (e.g., in the form of lobbying for particular legislation) could have a deteriorating effect on the democratic legislature and would put in question the legitimacy of law.

In an attempt to answer the question of legitimacy Jürgen Habermas postulates the discursive and procedural legitimacy of legal and political institutions, which are to be justified not in common values, but in procedures regulating free and unconstrained public debate on values and norms, open to every citizen at any time (Habermas, 1986, 1993). According to this view, only those norms can be considered justified that would stand up to unlimited and unrestricted debate. The stress is placed here on engagement in the public debate of civil society – a necessary link between the private and public sphere – that answers contemporary demands for interpretation of law close to local experiences, and the equal application of universal legal standards. Only then the law as well as democratic institutional arrangements have a chance to recover their institutional background. According to such arguments, civil society and public debate create a strong connection between the democratic political system and the system of law. Public debate imports local meanings and values, and particular experiences into interpretations of general and universal legal principles. However, it leaves open the issue of when compromise is impossible. A prerequisite for the functioning of such a model of procedural and discursive democracy is the revival of virtues, of the republican spirit, or habits of the heart characteristic of civil society, which make citizens participate and present their arguments, but also tolerate arguments presented by others. Within this model, proceduralization also means a growing stress on the role of the constitutional court regarding its relation to the political legislative process, and its activity as key in the formation of democratic will. Hence, the fundamental principles of democratic liberal constitution and constitutional control of political process are the important institutional arrangements that are able to control and tame the conflicting interests when consensus is not possible. Hence, one speaks about the constitutionalization of the political process in contemporary liberal democracies.

On the other hand, however, the new forms of government, known as *governance*, obviously have a great importance for current conceptualizations of law and its legitimacy. Legitimacy is given as the result of bi- and multilateral deliberations of various stakeholders, and reflects a consensus or a compromise among the various interests at stake. These new developments question not only the democratic legitimacy of law in the classical sense, but also the capacity of law to control and tame the struggle of the interests of actors engaged in the deliberations, and to protect the public good as well as citizens' rights. These new developments further contribute to the growing importance of judicial review – above all, of constitutional review – as not only controlling the application of law in individual cases, but also as controlling the conformity of governmental decisions and parliamentary legislation with the constitution. Instead of a clear separation of powers, a functional interdependence of the legislative, judicial, and executive branches of government characterizes modern democracies.

The model of procedural democracy and the proceduralization of law leaves some questions open. In light of this model's propositions, the legal and political systems of modern democratic societies are not only flexible but also fragile, their institutional arrangements and interpretations of norms depend on the results of debates and deliberations, or are left unresolved because common understanding or compromise cannot be reached. Here two questions emerge: (1) about stable and reliable foundations for the adjudication and legitimation of the democratic state ruled by law, and (2) about constraints on political and judicial deliberations.

In contemporary democracies, stable foundations for the legal as well as political system are sought in the concept of rights. This increasingly important approach to law – called 'new personalism' (Friedman, 1990: 44–46) – refers to inalienable human rights as a higher standard for evaluation of democratic and legal system functioning. This would serve as a final test according to which the quality of democratic and legal order would be measured. Therefore, the relationship between democracy and law in modern, complex, and pluralistic societies must be analyzed and evaluated with reference to universal standards – encompassed by human rights – interpreted and applied in the changing social and differentiated cultural conditions of the modern world. Protection of human rights presents also the most important constraints to political decision making and law application.

The Contribution of the Experience of Authoritarian and Totalitarian Regimes to the Democracy and Law Debate

The recent peaceful revolutions in Central and Eastern Europe were also called 'lawful' revolutions, since the observance of law, and the establishment of the rule of law was on the agenda of democratic dissent in that part of the world. However, the democratization process following the collapse of authoritarian and totalitarian regimes not only in Europe but also in Latin America, and most recently in Egypt or Tunisia, raised some fundamentally important legal issues, or reawoke some debates highly relevant for the analysis of the relationship between democracy and law.

Especially in Europe and Latin America, it was the important issue of inalienable human rights, and the idea of laws that are higher then man-made, positive law and therefore cannot be changed by regular, democratic procedures. Second, was an issue of control of the democratic legislature, i.e., the place and role of judicial review, constitutional review, and the review of supranational courts and tribunals (of which the European Tribunal of Human Rights presents a model example) of the democratic lawmaking process. Third, the issue was raised of 'transitional justice' i.e., of settling accounts with past human rights violations, to deal with the complicated legal matters and to invent procedures and institutions that would help to restore the sense of law and justice in these societies.

In Europe, the abuse of legality by Nazi, fascist, and communist regimes necessitated a new, rights-centered concept of legality and the rule of law. The issue of fundamental human rights as a higher legal order to which the positive law must subordinate itself was famously raised by the necessity of new constitutions in societies on the path to democratic reconstruction. The example was set by the German Basic Law of 1949 which – in its opening Article One – refers to the principle of inalienable human dignity and human rights, which are directly binding, and which are fundamental for every human community, for peace and justice in the world. The next 20 articles of that Basic Law that declare human rights and

liberties protection are pronounced inalienable. The issue of human rights was on the agenda of all dissident movements in the East Central Europe, and elsewhere. One speaks about the human rights revolution on a global scale, which had its roots in the 1948 Universal Declaration of Human Rights, the 1950 European Convention of Human Rights, and the international conventions of the 1960s – the International Covenant on Civil and Political Rights, and the International Convention on Economic, Social, and Cultural Rights. These conventions within a framework of the Helsinki peace process proved to be of salient importance for democratic dissent. As the source of human rights, and of human dignity, the classical concept of natural law has been evoked, again first in the German Federal Republic after World War II. Later this concept was referred to in East Central Europe. However, the revival of natural law as the source of human dignity and human rights opened a new set of questions and challenging issues in the debate on law and democracy. Among them the most salient follow who has the legitimacy of natural law principles interpretation, the interpretation of the human dignity principle and of human rights content in the modern, complex, pluralistic, and democratic society. This question concerns the concept of justiciable human rights substantively, and the specific mechanisms and institutional devices for safeguarding them, in the system of democratic government.

This put on the agenda again the issue of the judicial control of democratic lawmaking, and the place of the judiciary within the democratic political system. One speaks about the 'judicialization of politics,' or 'global expansion of judicial power' (Tate and Vallinder, 1995). According to Said Arjomand, one important consequence of political reconstruction presents the judicial activism (Arjomand, 2003: 8). This phenomenon is characteristic of the so-called era of 'new constitutionalism' since 1989 that is "marked by a mixture of increasingly judicialized legislation by parliaments and administrative organs and legislative jurisprudence by constitutional courts and supranational judiciary organs" (Arjomand, 2003: 9). As had been stressed with regard to post-Nazi Germany and postcommunist Eastern Europe, as well as in postauthoritarian Southern Europe and elsewhere, constitutional courts and tribunals with rights jurisdiction built into them, in some cases also the institution of an ombudsperson play a key role in the transition to democracy and protection of the rule of law centered on human rights protection, as guardians of the liberal democratic constitution (Bryde, 1999: 28–29). This puts in question the classical separation of powers in liberal democratic regimes (Stone Sweet, 2000: 139–152).

Finally, one has to reckon with past legacies and expose the truth about the former regime. At stake is the principle of the law-governed state, interpreted under conditions of mass-scale human right violations committed by the former regime. Thus, the term transitional justice was invoked and interpreted as a period of time when the former perpetrators of atrocities and their accomplices are put on trial (Teitel, 2000).

There are important contributions to the debate on democracy and law, as well as to the very concept of the rule of law resulting from posttotalitarian and postauthoritarian experiences with 'transitional justice.' These take form of new rights among which the 'right to truth,' to compensation of losses and restitution of rights presents their crucial, institutional crystallizations.

A prominent lawyer, the Executive Director of the Inter-American Institute on Human Rights in Latin America, argues that facing and reckoning past human rights abuse constitutes an obligation for new governments. They need to investigate, prosecute, and punish; they also need to disclose to the victims and society all that can be reliably discovered about the circumstances of such crimes, including the identity of the perpetrators and instigators. Such an obligation is translated into a society's right to truth (Mendez, 2009: 39). This right emerged in an age of gross violations of human rights, giving rise to high numbers of perpetrators, instigators and collaborators, and victims, and was formulated to deal with an unprecedented technical capacity of hiding the truth. Thus, the right to truth comprises a civic right of a new type, a response by a civic society to the increasing possibilities of concealing atrocities, destroying evidence, and silencing people – behavior typical of modern dictatorships and totalitarian regimes. As such, it is a most important contribution to interpretation of the law-governed state principle after the collapse of a totalitarian or dictatorial regime.

The right to truth is based on a belief that the truth about the past, if disclosed and disseminated, brings both restitution of dignity to victims and retribution for the past atrocities. In East Central Europe, such an obligation stems from the conviction that 'living within the truth' is a crucial part of the moral capital with which these societies initiated their transformation toward democracy. If taken seriously, the right to truth serves as crucial guidance in the process of differentiation of the new political system from the previous one. Therefore, the truth about the past is constitutive for new democracies. It contributes to the restoration of victims' dignity and the protection of their rights. It also comprises an initial step in the process of public enlightenment, of public self-reflection, as well as differentiation from the past.

Moreover, the right to truth represents a specific form of the broader civic right to information enshrined in every democratic liberal constitution. Further, the right to truth is directly and imminently connected with another one: the right to remembrance and to present one's own experiences, contributing to the collective memory in a democratic society. The right to remembrance places democratic governments under an obligation to document the particular experiences of individuals and broaden the official version of national history by local experiences. Wrongs must be documented, evidence collected, and the 'white spots' of a nation's history filled in order to render the collective memory more democratic and, above all, include the memories of victims. In order to meet this obligation and protect the right to remembrance, new 'institutes of national remembrance,' or similar institutions, have been established. If protected, these rights to truth and to remembrance contribute to the process, which could be labeled 'public enlightenment.'

Interestingly, it is the civil society, whose struggles against policies of past concealment contribute to this process of public enlightenment and large-scale public learning, thus guaranteeing protection of the right to truth and to remembrance. As already witnessed during global processes of reckoning with the past, the quantum of victimization has seldom been exposed by governmental or international bodies (Bassiouni, 1996: 11). Likewise in Eastern Europe that task has primarily been undertaken, with all its understandable limitations, by NGOs

such as Charter 77 in Czechoslovakia, or the Committee for the Defense of Workers of the late 1970s and early 1980s in Poland, or the Memorial organization in Russia thanks to which we learn about atrocities committed by the USSR. There are also organizations of victims of the tyrannical regimes, of slaves condemned to work in Siberian or other labor camps, of owners deprived of their property, of mothers, of dedicated journalists, and of committed researchers struggling to expose the truth and not forget it.

However, there are great dilemmas and important ambiguities to reckon too. One must be aware that democracy is open and vulnerable: open to any organized interests, and therefore vulnerable because these could be antidemocratic and/or even represent those who played a crucial role in the functioning of the previous system. One has also to acknowledge the fact that past reckoning brings much temptation to the new, democratic governments. It is a temptation to get rid of political opponents, if only one can find any evidence, however feeble, that they collaborated in the functioning of the former regime. Therefore, it is important that the past-reckoning activities be subordinated to transparent procedures. In this respect, the role of courts, especially of constitutional courts, has been stressed (Solyom, 2003: 133–161).

Conclusions

In the contemporary world, the relationship between law and democracy is increasingly problematic in all relevant dimensions: democratic participation in legislation, social and political subordination to law, and legal control of democratic majority, here called the taming of democracy by law, in order to protect rights and liberties of citizens. The contemporary challenges have been answered by new developments characteristic of law, that, in turn, result not only in the growing controlling power of the judiciary but also in the empowerment of citizens. The latter takes the form of the broad interpretation of the nondiscrimination principle and the protection of individual, and recently also collective, rights. Therefore one can speak about broadening and deepening democracy, accompanied by the parallel growing importance of rights. These new developments have been enhanced by the collapse of authoritarian and totalitarian regimes in Europe and elsewhere. However, one observes new phenomena consisting in the growing importance of sub- and international processes, presenting new challenges to both democracy and law and resulting in the fragmentation of relations between them. Therefore, the relationship between law and democracy in the contemporary world needs further deep attention and analysis.

See also: Affirmative Action: Comparative Policies and Controversies; Civil Liberties and Human Rights; Civil Rights; Constitutional Courts; Critical Race Theory; Deliberation and Democracy; Democratic Transitions; Equality and Inequality: Legal Aspects; Feminist Legal Theory; Race and the Law; Social Justice: Historical and Theoretical Considerations.

Bibliography

Aristotle, 1981. Politics. Penguin, Harmondsworth.
Arjomand, Said Amir, March 2003. Law, political reconstruction and constitutional politics. International Sociology 18 (1), 7–32.
Aubert, Vilhelm, 1983. In Search of Law. Sociological Approaches to Law. Martin Robertson, Oxford.
Barnett, Hilaire, 1998. Introduction to Feminist Jurisprudence. Cavendish Publishing Limited, London-Sydney.
Bassiouni, M.C., 1996. Searching for peace and achieving justice: the need for accountability. Law and Contemporary Problems LIX (4).
Berman, Harold J., 1983. Law and Revolution. The Formation of the Western Legal Tradition. Harvard University Press, Cambridge, Massachusetts.
Bryde, Brun-Otto, 1999. The role of constitutional jurisprudence in German democracy. In: Weick, G. (Ed.), Competition or Convergence: The Future of European Legal Culture. Peter Lang, Frankfurt, pp. 23–29.
Fletcher, Ruth, 2002. Feminist legal theory. In: Banakar, Reza, Travers, Max (Eds.), An Introduction to Law and Social Theory. Hart Publishing, Oxford- Portland, Oregon, pp. 135–154.
Friedman, Lawrence B., 1990. The Republic of Choice: Law, Authority and Culture. Harvard University Press, Cambridge, Massachusetts.
Habermas, Jürgen, 1986. Law as medium and law as institution. In: Teubner, G. (Ed.), Dilemmas of Law in the Welfare State. de Gruyter, Berlin.
Habermas, Jürgen, 1993. Faktizität und Geltung. Suhrkamp, Frankfurt am Main.
Held, David, 1987. Models of Democracy. Polity Press, Basil Blackwell, Cambridge.
Held, Davi, 1995. Democracy and the Global Order. Cambridge Polity Press.
Locke, John, 1960. Two Treaties on Government. org. ed. 1690. Cambridge University Press, Cambridge.
Luhmann, Niklas, 1981. Politische Theorie im Wohlfahrtstaat. Guenther Olzog Verlag, Muenchen-Vien.
Luhmann, Niklas, 2004. Law as a Social System. Oxford University Press.
Mendez, Juan, 2009. "Right to truth" Latin-American contribution. In: Karstedt, Susanne (Ed.), Legal Institutions and Collective Memories. Hart Publishing, Oxford and Portland, Oregon, pp. 39–62.
Mendez, Juan E., 2009. An emerging right to truth. In: Karstedt, Susane (Ed.), Legal Institutions and Collective Memories. Hart Publishing, Oxford and Portland, Oregon, pp. 39–62.
Montesquieu, 1952. The Spirit of Laws. William Benton, Chicago.
Rousseau, Jean-Jacob, 1973. The social contract. org. ed. 1762. In: The Social Contract and Discourses. (G.D.H.Cole, Trans.). Everyman, London.
Slaughter, Anne-Marie, 2004. Sovereignty and power in a networked world order. Stanford Journal of International Law 40, 283–327.
Solyom, Laszlo, March 2003. The role of constitutional courts in the transition to democracy, with special reference to Hungary. International Sociology 18 (1), 133–161.
Sweet, Alec Stone, 2000. Governing with Judges. Constitutional Politics in Europe. Oxford University Press, Oxford.
Tate, C. Neal, Vallinder, Torbjoern (Eds.), 1995. The Global Expansion of Judicial Power. New York University Press.
Teitel, Ruti G., 2000. Transitional Justice. Oxford University Press, Oxford.
Teubner, Gunther, 2004a. Societal constitutionalism: alternatives to state-centered constitutional theory. In: Joerges, C., Sand, I.-J., Teubner, G. (Eds.). Hart Publishing, pp. 3–28.
Teubner, Gunther, 2004b. Global private regimes: neo-spontaneous law and dual constitutions in autonomous sectors. In: Ladeur, K.-H. (Ed.), Public Governance in the Age of Globalization. Ashgate, pp. 71–87.
Teubner, Gunther, 2012. Constitutional Fragments: Societal Constitution and Globalization. Oxford University Press.
Teubner, Gunther, 2013. The law before its law. German Law Journal 14 (2).
Unger, Roberto Mangabeira, 1976. Law in Modern Society. Toward a Criticism of Social Theory. Free Press, New York.

Law and Development

David M Trubek, University of Wisconsin–Madison, Brooklyn, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Law and development' refers both to organized efforts to transform legal systems in developing countries to foster economic, political, and social development and to the academic projects stimulated by these efforts. Begun in the mid-twentieth century, law and development's dominant ideas and projects have changed over time as theories of development and agency priorities have changed. Law and development efforts accelerated in the 1990s as international financial institutions began to emphasize the rule of law. As the twenty-first century dawned, ideas and projects of the twentieth century were assessed and critiqued, new themes emerged, the field fragmented, and consensus disappeared.

Law and development refers to organized efforts to transform legal systems in developing countries to foster economic, political, and social development. The term also refers to the academic projects associated with these efforts. Begun in the mid-twentieth century, law and development's dominant ideas and projects have changed over time as theories of development and development agency priorities have changed. Twice in the twentieth century, a consensus formed and guided practice. As the twenty-first century dawned, ideas and projects of the twentieth century were assessed and critiqued, new themes emerged, the field fragmented, and consensus disappeared.

In the twentieth century, law and development focused on the use of domestic law as a tool to facilitate economic growth, and priority was given to law as an instrument to empower developmental states. Toward the end of that century, this approach gave way to a focus on law as the underpinning of markets and a restraint on the state (Trubek and Santos, 2006).

In the twenty-first century, new ideas and forces are once again transforming the field. The increasing impact of global forces has drawn more attention to the role of global institutions and actors, new ideas about how development works have led to calls for more use of law as a framework for public–private coordination and experimentation, and interest in law as an element of development itself has led to stress on guarantees of basic rights. Finally, recognition that much twentieth-century law and development effort was based on weak or faulty knowledge (about both the role of law and the possibilities for reform) led to calls for more research, an evidence-based approach to planned change, and horizontal learning among developing nations.

Law and Development in the Twentieth Century: Domestic Law and Economic Growth

The idea that a nation's legal system affects its economic and social prospects can be traced back as far as the eighteenth century. In the nineteenth century, German sociologist Max Weber argued that European legal systems were an important factor in the development of capitalism (Trubek, 1972). But it was only in the twentieth century that governments and international institutions concerned with development began to organize systematic legal reform projects. The 1960s saw the start of support for legal reform efforts by international development agencies and the beginnings of academic study of law and development. By the late 1970s, agency interest had waned and the academic enterprise stagnated. However, in the 1980s, new ideas about law's role in development began to emerge. This spurred increasing interest by development agencies; the pace of support for systemic reform picked up, and academic study revived (Davis and Trebilcock, 2008). International development agencies spent billions on legal reform, and scholars in several disciplines other than law began to study the role of law in development.

The second half of the twentieth century saw the emergence of various theories and approaches. The focus was on domestic legal systems, and the goal was increasing economic growth. Three major themes can be identified, and they are discussed in this section.

Law Can Be an Instrument to Be Used by Developmental States to Foster Change

The first law and development movement stressed the need for change in economic and social relations and expressed hope that new laws could induce such change. This body of thought drew on modernization theory and the view – widely held by economists around midcentury – that a strong state was needed to manage the economy and transform society (Thomas, 2006). Law, in this view, could serve as a positive instrument of change. Law could increase state capacity. It could offer incentives for people and institutions that are 'modern' and would promote growth. It could create disincentives for those who resist change and cling to traditional values (Davis and Trebilcock, 2008). In what Trubek and Santos (2006) call the 'first moment' of law and development, there was a consensus on the importance of the state and the need for modernization, and this consensus helped structure a development policy that focused on the use of law to strengthen the state and transform society.

Law May Be a Barrier to Economic Development

From the beginning, even those who had a positive view of law understood that the wrong kind of legal rules and practices could reduce incentives for investment and increase the cost of innovation. Concern about the negative impact of law grew in

importance toward the end of the century as development agencies lost faith in state intervention and began to put more stress on the role of markets and on the need for deregulation (Kennedy, 2006).

Law Should Be a Framework to Facilitate Private Decision Making

As development economics turned away from a belief in state-led initiatives, more and more emphasis was placed on the role of law as a framework within which private actors would make economic decisions. In the second 'moment' of law and development, a consensus formed that priority should be placed on the role of law as the underpinning of markets and a means of restraining excessive state intervention. This new consensus guided reform efforts in the 1990s and beyond. Scholars stressed that, to function properly, markets require a complex infrastructure of institutions and rules, including legal rules such as the laws of contract and property (North, 1991; Dam, 2006; Trebilcock and Daniels, 2008). They also recognized the importance of qualified legal professionals and judges in ensuring that the laws are effective (Trebilcock and Daniels, 2008). Markets may also require regulations, like antitrust and securities law (Dam, 2006; Trebilcock and Daniels, 2008). Much of the interest in legal transplants and in judicial reform for development, which made up the bulk of international support for law reform in the late twentieth century, was based on the importance of creating such a 'neutral' framework.

The Mixed Legacy of the Twentieth Century

The legacy of the twentieth century is a complex one. First, the hoped-for academic field of law and development never materialized. In 1974, Trubek and Galanter (1974) criticized scholars' excessive closeness to development agencies and called for greater academic detachment. This article, intended by its authors to strengthen the field, ironically helped bring about its demise by questioning core assumptions and weakening the link between funding agencies and the academic world. For a decade or more, scholarly work on law and development declined. But, by the late 1980s, academic interest increased and several other disciplines, including economics, took up the topic. However, nothing like a 'field' took shape. While the number of studies and publications grew exponentially, there still is no association that brings law and development scholars and practitioners together, no agreement on canonical texts, and no publication that is widely read and commands broad respect. Brian Tamahana (2011) argues that such a field is neither possible nor desirable, while others hope for greater coordination and integration of knowledge in the future.

Second, the experience of reform was mixed as best: some projects failed, and some transplanted laws did not 'take.' Some decided the whole enterprise was fundamentally flawed and that organized efforts by external actors to promote legal reform were doomed to fail. They argued that it was impossible for external actors to understand how developing countries' legal systems work and/or to craft workable reform strategies (Tamahana, 2011). Others were more sanguine: they believed that law could matter and that properly planned reforms could have an effect. But they admitted that the empirical base for reform was too thin and called for more research to identify what works and what does not (Davis and Trebilcock, 2008).

Third, conflicts arose among the several ideas that dominated the period. There are tensions between the idea of the use of law by a strong developmental state, on the one hand, and a deregulatory push, on the other; between law as an instrument to change behavior and law as a neutral framework; and between deregulation and the need to create new institutions (Kennedy, 2006). Law and development scholars had trouble incorporating the lessons of the East Asian miracle: while countries like Taiwan and Korea experienced rapid and inclusive growth, their legal systems did not conform to the precepts of the law and development orthodoxy of the 1990s (Ohnesorge, 2006; Jayasuriya, 1999).

Finally, lawyers and economists diverged in their views about methodology, their interpretation of data, and the policies they derived from the data. One of the most important changes in law and development in the twentieth century was the discovery of the field by economists. The economics profession paid scant attention to law in the 1960s and 1970s, but starting in the 1980s development economists began to take law seriously. This led to a series of studies, including several about the importance of law in the creation of financial markets. One series widely discussed were the studies purporting to show that legal systems whose origins were in the common law were superior for financial development than those based on the civil law (La Porta et al., 1998, 2008). Lawyers found these studies less than persuasive, pointing out numerous flaws in the reasoning employed and suggesting the need for alternative approaches. In a very influential article, legal scholar Katherina Pistor showed that some of the studies used simple models that presume that lightly regulated markets are optimal, fail to capture the complexity of real legal systems and the normative choices they encode, misunderstand the relationship between the history of a legal system and its ability to incorporate specific rules, and rely on questionable assumptions about causal relations between law and economic outcomes. Rejecting the methods used in such studies, she prefers an approach partially inspired by comparative law and sociolegal studies (Pistor, 2009). All these issues and many others continue to be debated in the academy and in practice (Davis and Trebilcock, 2008).

Law and Development in the Twenty-First Century: Continuity and Change

Thus, as the twenty-first century dawned, law and development was faced with unresolved issues and unfinished tasks. Leading scholars looked for lessons to be derived from the expansion of reform efforts in the 1990s and early 2000s, tried to reconcile tensions among the different ideas about law and about development in the twentieth century, started to deal with new trends and phenomena, and called for the creation of a systematic body of knowledge that could serve as a reliable guide to reform efforts (Davis and Trebilcock, 2008; Trubek and Santos, 2006). Others suggested it would be best to give the whole thing up as a bad job (Tamahana, 2011). For those who felt that law could

make a difference in development, and who believed that externally supported reform might still be desirable, there was a lot of experience to digest as well as new ideas to incorporate and new challenges to face.

Evaluating the Experience of the Twentieth Century

Much of the literature in the early twenty-first century focused on the experiences of the prior decades, codifying what was known, pointing to gaps in knowledge, and critiquing assumptions behind theories and policies. Between 2002 and 2011, at least seven important books on law and development in general, along with numerous specialized studies, were published in the United States alone. Some of the general studies looked at past reform efforts (Jansen and Heller, 2003; Carothers, 2003) and underscored the limitations of knowledge about law and development and the obstacles to effective reform, including the difficulties of transplanting legal institutions from one country to another. Others offered an overview of the legal reforms needed for capitalist growth and summarized the state of knowledge accumulated in the twentieth century about specific areas of law and key legal institutions (Dam, 2006; Trebilcock and Daniels, 2008; Trebilcock and Prado, 2011). These studies looked at the role that could be played by reforms in private law, regulation, criminal justice, the role of the judiciary, legal education, and the legal profession and summarized what was known about reforms that had been carried out in these areas. The literature included a critical strand: Dezalay and Garth (2002) placed law and development within the context of global forces and the hegemony of neoliberal development economics, while authors in the volume edited by Trubek and Santos (2006) outlined several critiques of mainstream twentieth-century law and development policies and practices. They noted that the supposedly neutral framework of private law could favor certain groups over others and affect the distribution of income. They also contended that the reform effort of the 1990s had favored economic goals over social protection and inclusion.

New Ideas, New Forces, New Institutions

The new century also saw the emergence of new ideas in development economics, new forms of state activism, and new issues for the field of law and development. New approaches to development economics and the emergence of new kinds of developmental states required new ideas about the role of law. Trubek et al. (2013) produced a study of the revival of industrial policy and other new forms of state activism in Brazil and suggested that these new strategies called for yet another reassessment of law's role in development. Global forces, including international economic law, began to play a greater role in the shaping of domestic legal systems. And there were calls to go beyond the instrumental view of law as a tool for development and treat a functioning legal system with basic guarantees as an element of what should properly be called 'development.' Law and development scholars recognized the limits of their knowledge and called for more empirical research, while the capacity for such work by scholars in the Global South began to expand.

Law Should Facilitate Experimentation and Innovation

By and large, twentieth-century law and development experts assumed that the path to development was known and the challenge was to create instruments and institutions that would help move nations down that path. For those who saw law as an instrument of state control, knowledge about the path forward would be provided by state planners, and the law would implement the plan. Those who rejected planning thought that unrestricted markets would make the optimal economic choices, so the right way to foster development was to create markets and leave them alone (Trubek et al., 2013).

However, in the twenty-first century, many are coming to believe that neither planners nor markets working alone can find the optimal path. Rather, strategies must evolve and investment choices must be made through public–private partnerships and processes of iterative experimentation (Rodrik, 2004; Sabel, 2007; Sabel and Reddy, 2003; Hausmann et al., 2007). These processes require new forms of governance and law. In such an 'experimentalist' developmental state, law can be neither a simple tool for direct state invention nor merely a neutral framework for private decisions (Pires, 2008). Rather, the law should seek to establish partnerships between public and private sectors and institutionalize a process of mutual search for innovative solutions and optimal developmental paths (Trubek et al., 2012).

Law Is Increasingly Affected by Global Forces

Twenty-first century law and development must deal with the growing impact of global forces on the law (Dezalay and Garth, 2002; Trubek et al., 1994). There are three major forces at work. The first is the availability of global models such as the formulae for law and development promoted by the World Bank (Davis, 2004; Davis and Kruse, 2007; Santos, 2009). Whether or not a country receives funding from the Bank, the models it promotes have an influence on national thinking. Second, national lawmakers must take account of the role that law plays in determining national competitiveness. The more a country's development strategy depends on foreign investment, the more its laws will be subject to scrutiny by foreign investors, who will compare the legal environment for development in various nations before deciding where to invest. Today, investors have many options and law becomes an important factor in the effort to attract capital. The third global force is the growth of transnational law. Increasingly, a nation's legal order is affected by norms originating outside of its borders. Whether they are norms of regional bodies (like the North American Free Trade Agreement or Mercosur) or of global institutions like the World Trade Organization, national legal orders are subject to constraints from higher levels of governance (Trachtman and Thomas, 2009; Shaffer et al., 2008; Santos, 2012).

Law Itself Is Part of Development

All the major ideas about law and development in the twentieth century saw law as a *means* to some other goal, whether economic growth or social protection. But recently, scholars have argued that the existence of the 'rule of law' is a goal in itself, a necessary part of the process of empowerment and capability enhancement that constitutes 'development' (Sen, 1999). This

means that legal protection for constitutional values and human rights, including economic and social rights, must form part of the law and development agenda along with economic law and judicial reform (Sen, 1999; Rittich, 2004).

Recognition of the Need for Evidence-Based Policies

As more evidence about the failures of twentieth-century reform efforts accumulates, there are increasing calls to get beyond abstract debates and develop empirical evidence concerning what works and what doesn't work (Carothers, 2003; Jansen and Heller, 2003; Davis and Trebilcock, 2008). There is very little empirical work of any kind on the role of law in developing countries, yet the whole law and development enterprise requires such knowledge. That will include developing tools to diagnose problems and measure the results of reforms. The creation of cross-country legal indicators by agencies like the World Bank is a reflection of the search for such tools. But this process is still in its infancy, and there are questions about some of the indices being used and the policies derived from them (Davis, 2004; Davis and Kruse, 2007). Indicators can be misleading if they rely on only the formal written law, not the law as it is actually applied, or are drawn from user surveys that are not truly representative (Santos, 2009). Moreover, even when the data reflected in the indicator are accurate, sometimes policy makers make questionable leaps from the data by proposing reforms that are either inappropriate or beyond the capacity of the government to carry out.

Increasing Law and Development Capacity in the Global South

Developing countries need to make a quantum leap in their capacity for sociolegal research. One of the most important developments in recent years is the growth of law and development capacity in the Global South. While scholars in developing countries have always been aware of the role that law might play in development, their capacity to study and analyze these issues has been limited. Law and development calls for an interdisciplinary approach and demands empirical research. The legal academies in much of the Global South were ill equipped for such work. They did little systematic research and what was done was largely doctrinal. Toward the end of the twentieth century, this began to change: legal scholars in some countries acquired interdisciplinary knowledge and empirical skills, law schools supported research and teaching on development issues, and some funds for empirical research became available. Legal scholars from some developing countries began to exchange ideas and share experiences: for example, in 2010, scholars from Brazil, Russia, India, and China met in São Paulo to discuss selected law and development topics (Schapiro and Trubek, 2012). As capacity grows, and as more cooperation between the legal scholars of developing countries occurs, 'law and development' may be institutionalized in the Global South and the kind of 'horizontal learning' among developing nations illustrated by the São Paulo meeting may increase.

Conclusion: Fragmentation and an Uncertain Future

While reform continues and academic work on the role of law in development is expanding in the Global North and the Global South, there is no consensus on the nature of the challenges or the appropriate policies to follow. Twice in the history of law and development, a dominant law and development paradigm emerged, which served as a guide to policy makers (Trubek and Santos, 2006). But today the 'field' is fragmented in many different ways. While there is more academic work than ever, no approach is dominant and no policy consensus can be found. A review of the current debates suggests fragmentation along three axes. The first is the basic approach that scholars take to the law and development project itself. The second is disciplinary, with economists and lawyers often looking at similar issues but using different methods and reaching differing conclusions. The third is substantive: while work on law and development once brought together all relevant topic areas, today these issues are often taken up in separate academic silos with little communication among them.

Attitudes toward the Project: Skeptics, Critics, and Optimists

In a masterful review of the recent academic literature, Kevin Davis and Michael Trebilcock (2008) divided the field into skeptics who have serious doubts about the project and optimists who recognize problems but remain hopeful about the enterprise. In fact, the field seems to be divided among *three* groups that take very different approaches to the whole idea of externally supported legal reform as a tool for development. The first are the skeptics who believe that either law does not matter for development or, if it does, the external actors are incapable of understanding what is needed and will always get it wrong (Tamahana, 2011). The second are the critics: they know that law matters. They think that external actors know what they are doing and the reforms they promote serve identifiable aims. The problem, say the critics, is not with the effectiveness of the norms but with who loses and who benefits from them. For example, some see external reforms working, but in a way that may weaken social protections and benefit global capital (Rittich, 2004). Finally, the optimists believe that law matters, that legal reforms can be crafted that will benefit all in developing societies, and that external agencies, properly guided by empirically informed research, can make a difference (Davis and Trebilcock, 2008).

Disciplinary Divide

The division between lawyers and economists that started in the twentieth century has continued. Thus, when leading proponents of the 'legal origins literature' restated their position concerning the superiority of countries that adopted the common law (La Porta et al., 2008), legal scholars, including Katerina Pistor and John Ohnesorge, expanded the critique of the methods used and conclusions reached by the economists (Pistor, 2009; Ohnesorge, 2009).

Topical Fragmentation

A final source of fragmentation can best be described as topical. As interest in development grew, various legal disciplines began to explore the relevance of their topic for development broadly defined. Thus, we saw the growth of interest in trade and

development, finance and development, human rights and development, women's rights and development, and so on. While these topical specializations enriched our knowledge of the prospects and pitfalls of law in development, they often became independent specializations in themselves, thus contributing to further fragmentation of the field.

An Uncertain Future

Academic interest in law and development in the Global North has never been stronger, and capacity in the Global South has grown. These developments promise to provide insights that will influence reform. Yet the proliferation of studies and the fragmentation of the field along multiple axes suggest that we may never see the emergence of another dominant paradigm like those 'moments' that existed in the past.

See also: Globalization: Legal Aspects; Law and Society: Development of the Field; Law: Change and Evolution; Postcolonial Law; Rule of Law (and Rechtsstaat).

Bibliography

Carothers, T., 2003. Promoting the Rule of Law Abroad. Carnegie Endowment for International Peace, Washington, DC.

Dam, K., 2006. The Law-Growth Nexus: The Rule of Law and Economic Development. Brookings, Washington, DC.

Davis, K., 2004. What can the rule of law variable tell us about the rule of law reform? Michigan Journal of International Law 26 (1), 141–162.

Davis, K., Kruse, M.B., 2007. Taking the measure of law: the case of doing business project. Law & Social Inquiry 32 (4), 1095–1119.

Davis, K., Trebilcock, M., 2008. The relationship between law and development: optimists versus skeptics. American Journal of Comparative Law 56 (4), 895–946.

Dezalay, Y., Garth, B., 2002. Global Prescriptions: The Production, Exportation, and Importation of a New Legal Orthodoxy. University of Michigan Press.

Hausmann, R., Rodrik, D., Sabel, C.F., 2007. Reconfiguring industrial policy: a framework with application to South Africa. Unpublished Manuscript. Available from: http://www.hks.harvard.edu/fs/drodrik/publications.html.

Jansen, E.G., Heller, T.C., 2003. Beyond Common Knowledge: Empirical Approaches to the Rule of Law. Stanford University Press, Stanford.

Jayasuriya, K. (Ed.), 1999. Law, Capitalism and Power in Asia: The Rule of Law and Legal Institutions. Routledge, New York.

Kennedy, D., 2006. The rule of law, political choices, and development common sense. In: Trubek, D.M., Santos, A. (Eds.), The New Law and Economic Development: A Critical Appraisal. Cambridge University Press, New York.

La Porta, R., Lopez-de-Silalnes, F., Shleifer, A., Vishny, R.W., 1998. Law and finance. Journal of Political Economy 106 (6), 1113–1155.

La Porta, R., Lopez-de-Silianes, F., Shliegfer, A., 2008. The economic consequences of legal origins. Journal of Economic Literature 46 (2), 285–322.

North, D., 1991. Institutions. Journal of Economic Perspectives 5 (1), 97–112.

Ohnesorge, J., 2006. Developing development theories: law and development orthodoxies and the northeast Asian experience. University of Pennsylvania Journal of International Economic Law 28 (2), 219–308.

Ohnesorge, J., 2009. Legal origins and the tasks of corporate law in economic development: a preliminary exploration. Brigham Young University Law Review (6), 1619–1634.

Pires, R., 2008. Promoting sustainable compliance: styles of labour inspection and compliance outcomes in Brazil. International Labor Review 147 (2–3), 199–229.

Pistor, K., 2009. Rethinking the "law and finance" paradigm. Brigham Young University Law Review 2009 (6), 1647–1670.

Rittich, K., 2004. The future of law and development: second generation reforms and the incorporation of the social. Michigan Journal of International Law 26 (1), 199–244.

Rodrik, D., 2004. Industrial policy for the twenty-first century. In: Rodrik, D. (Ed.), One Economics, Many Recipes: Globalization, Institutions of Economic Growth. Princeton University Press, Princeton, pp. 99–152.

Sabel, C.F., 2007. Bootstrapping development: rethinking the role of public intervention in promoting growth. In: Lee, Victor, Swedberg, Richard (Eds.), On Capitalism. Stanford University Press, Stanford, pp. 305–341.

Sabel, C.F., Reddy, S. 2003. Learning to Learn: Undoing the Gordian Knot of Development Today. Columbia Law and Economics Working Paper No. 308. Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=944619.

Santos, A., 2009. Labor flexibility, legal reform and economic development. Virginia Journal of International Law 50 (1), 43–106.

Santos, A., 2012. Carving out policy autonomy for developing countries in the world trade organization: the experience of Brazil and Mexico. Virginia Journal of International Law 52 (3), 551–632.

Schapiro, M., Trubek, D.M., 2012. Direito e desinvolvimento: um diálogo entre os BRICs. Saraiva, Sao Paulo.

Sen, A., 1999. Development as Freedom. Knopf, New York.

Shaffer, G., Sanchez, M.R., Rosenberg, B., 2008. The trials of winning at the WTO: what lies behind Brazil's success. Cornell International Law Journal 41 (2), 383–502.

Tamahana, B., 2011. The primacy of society and the failure of law and development. Cornell International Law Journal 44 (2), 209–248.

Thomas, C., 2006. Max Weber, Talcott Parsons and the sociology of legal reform: a reassessment with implications for law and development. Minnesota Journal International Law 15 (2), 383–425.

Thomas, C., 2011. Law and neoclassical economic development: toward an institutionalist critique of institutionalism. Cornell Law Review 101 (4), 967–1024.

Trachtman, J., Thomas, C. (Eds.), 2009. Developing Countries in the WTO Legal System. Oxford University Press, Oxford.

Trebilcock, M.J., Daniels, R.J., 2008. Rule of Law Reform and Development. Edward Elgar Publishing.

Trebilcock, M.J., Prado, M.M., 2011. What Makes Poor Countries Poor: Institutional Determinants of Development. Edward Elgar Publishing.

Trubek, D.M., 1972. Max Weber on law and the rise of capitalism. Wisconsin Law Review 3, 720–754.

Trubek, D.M., Galanter, M., 1974. Scholars in self estrangement: reflections on the crisis in law and development studies. Wisconsin Law Review 4, 1062–1104.

Trubek, D.M., et al., 1994. Global restructuring and the law: studies of the internationalization of legal fields and the creation of transnational arenas. Case Western Reserve Law Review 44 (2), 407–498.

Trubek, D.M., Santos, A., 2006. The New Law and Economic Development: A Critical Appraisal. Cambridge University Press, New York.

Trubek, D.M., Countinho, D., Shapiro, M., 2012. Toward a New Law and Development: New State Activism in Brazil and the Challenge for Legal Institutions. World Bank Legal Review Paper No. 1207. Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2144939.

Trubek, D.M., et al., 2013. Law, state and the new developmentalism: an introduction. In: Trubek, D.M., et al. (Eds.), Law and the Developmental State: The Brazilian Experience in Global Context. Cambridge University Press, Cambridge.

Law and Economics

Steven Shavell, Harvard University, Cambridge, MA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 12, pp. 8436–8442, © 2001, Elsevier Ltd.

Abstract

Economic analysis of law is concerned with (a) determination of the effects of legal rules and (b) evaluation of the desirability of the effects of legal rules, with respect to well-specified definitions of social welfare. This article surveys the approach as it applies to basic areas of law – accident, property, contract, and criminal law – as well as to the litigation process. The economic approach is also contrasted with traditional analysis of law, under which the effects of legal rules are not usually systematically assessed.

Under the economic approach to the analysis of law, two basic questions about legal rules are addressed: descriptive questions, concerning the effects of legal rules on behavior and outcomes; and evaluative questions, concerning the social desirability of the effects of legal rules. In answering these questions, the method employed is that used in economic analysis generally. Namely, individuals and firms are ordinarily presumed to be forward-looking and rational, and the framework of welfare economics is adopted to assess the social desirability of outcomes. The field of economic analysis of law may be traced significantly to Bentham (1789), but lay essentially dormant until the contributions of Coase (1960), Becker (1968), Calabresi (1970), and Posner (1972). The field is now rapidly growing, although it is far from mature (one indication being lack of empirical work). To illustrate the approach, this article first focuses on accident law, briefly considers other areas of law, and concludes with a section on basic foundations and criticisms of the economic approach.

Economic Analysis of Law Illustrated: Accident Law

By accident law is meant the law governing liability for accidents, that is, the rules determining when a party who causes an accident must pay for the harm done. Economic analysis of this branch of law centers on three issues: incentives toward safety; insurance and compensation of accident victims; and litigation-related costs.

Incentives Toward Safety

A major effect of the liability system is that it fosters the taking of precautions against accidents. Suppose that a precaution will lower accident-caused harm for which a party would be liable from $3000 to $2000, that is, by $1000, but costs less than this amount, say $500. The party would then be likely to take the $500 precaution, as it would reduce the party's liability expense by $1000. Such logic underlies the conclusion that, under many forms of liability, parties will be led to take socially desirable precautions.

An important qualification about the general point that liability creates incentives toward safety applies to accidents caused by firms' products. In particular, a firm's interest in its reputation may lead it to take proper precautions even in the absence of liability. If a firm's water heaters tend to fail and consumers know about this, they will not pay as much for the heaters, and thus the firm will have a motive to reduce the risk of heater failure in order to avoid having to accept a lower price. Note, however, that this argument depends on consumers' obtaining information about product risk.

Insurance and Compensation

The role of insurance in the context of accidents and liability is of substantial importance due to the widespread ownership of insurance. A major form of insurance is liability insurance, which provides coverage against legal liability. Because liability insurers pay for much of the losses for which injurers are found liable, it might initially be thought that liability insurance largely negates the incentives toward safety inherent in liability. However, some such incentives are preserved under liability insurance because insurers often link premiums or conditions of coverage to the adequacy of precautions, raise premiums on the basis of accident history, and offer only partial coverage against liability.

In addition to liability insurance, standard insurance for victims, that is, their (private or public) medical, life, disability, and property insurance, is of significance. The prevalence of victims' insurance limits the need for the liability system as a means of assuring victims compensation for accidents.

Litigation-related Costs

The litigation-related costs of the liability system are the legal fees and associated expenses (including litigants' time and effort) borne by parties in resolving disputes that arise when harm occurs. Litigation-related costs are high; for every dollar received by a victim, it appears that over a dollar is spent delivering the dollar to the victim.

Evaluation Under the Economic Approach

Under the economic approach, the liability system is considered to be socially worthwhile where its social benefits exceed its litigation-related costs. The social benefits of the liability system do not lie significantly in compensation of victims, because standard forms of insurance for victims are a cheaper means of compensation than the liability system. Rather, the

social benefits of the liability system reside largely in its influence on accident rates. If this accident reduction effect is sufficient to outweigh litigation-related costs, the liability system is socially worthwhile. For example, the liability system might be socially worthwhile in the area of industrial pollution – perhaps the desire to avoid liability induces firms to reduce substantially polluting activity. However, the liability system might not be worthwhile in the area of automobile accidents – perhaps the prospect of liability does not affect driving behavior, which is mainly influenced by fear of personal injury in accidents.

Traditional Analysis Contrasted with Economic Analysis

The traditional view of the liability system is that its primary effect, and its major warrant for existence, is the compensation of victims of harm. This view is at odds with the economic view. On one hand, as just emphasized, compensation is to a great extent accomplished by standard forms of insurance; and, because insurance is a less expensive means of compensation than the liability system, it would be economically unwise to employ the liability system for the purpose of achieving compensation. On the other hand, the inducement of safer behavior is usually not paid serious attention under the traditional view of the liability system. Hence, the prescriptions for use of liability under the two views may conflict. Under the economic view, but ordinarily not under the traditional view, the recommendation is that the scope of liability be reduced where liability has little influence on accident frequency, and that the scope of liability be increased where liability would substantially reduce accident frequency.

A second, related aspect of the traditional view is that liability is intended to ensure the attainment of corrective justice in the classic sense that a wrongdoer should be punished by being made to pay the victim for harm done. From the economic perspective, this conception of the purpose of the liability system is problematic even in a descriptive sense. As stressed above, a party who is found liable usually does not pay a judgment himself but has his liability insurer pay; if there is punishment, it might be that the party is not fully covered or that his liability insurance premiums may rise. Hence, the degree to which a wrongdoer is punished owing to liability is not direct; it is attenuated and translated in character by liability insurance. Additionally, the victim often does not receive the payment made by the liability insurer; rather it is the victim's insurer that frequently obtains the payment (as reimbursement for the payment that the insurer made to the victim earlier). In all, then, the liability system does not achieve corrective justice in the way contemplated by traditional analysts, because they overlook the effects of liability and of victims' insurance.

Economic Analysis of Other Areas of Law

In this section, the contours of economic analysis of other major areas of law will be indicated.

Property Law

A fundamental topic in economic analysis of property law is the justification for the very existence of property rights. From the economic point of view, these rights are said to exist because they promote incentives to work, to maintain and improve things, and to trade; and, as well, because the rights reduce problems of wasteful and destructive efforts to take things and to prevent takings. Some economic literature traces historical instances of the emergence of property rights (especially in land) and certain recent property rights developments (for example, in the broadcast spectrum) to these social advantages. Property rights are seen as beneficial due to their salutary effects, not to any intrinsic belief that a person should own the fruits of his labor.

Given the general basis for the existence of property rights, many issues pertaining to property law have been addressed from the economic perspective. One concerns acquisition of property rights in things not yet owned, such as fish in the sea and undiscovered mineral deposits. Here, a theme is that the 'finders-keepers' rule has the advantage that it creates incentives to find things. Yet these incentives may be socially excessive: individuals may engage in duplicative activities to find things (a large number of fishermen may compete for fish that a smaller number could easily catch; many parties may drill oil wells, even though a small number of wells would be sufficient to extract the oil from a reservoir). When this is so, it is suggested that regulation (limiting the catch of fish, 'unitizing' ownership of an oil reservoir) may be desirable to curb the problem of excessive effort.

Another issue of interest concerns external effects associated with the use of property. These 'externalities' may be detrimental, such as pollution or noise creation, or beneficial, such as beautification of land or spraying to kill mosquitoes. Bargaining among affected parties may sometimes resolve externality problems. Suppose that a factory causes annual pollution harm of \$10 000 to its neighbor but can prevent the pollution by installing scrubbers at an annual cost of \$1000. In the absence of legal rules preventing pollution or imposing liability for it, the victim of the pollution might be expected to pay the factory to obtain scrubbers in order to avoid suffering harm. The victim should be willing to pay more than \$1000 (perhaps \$2000) for the factory to obtain scrubbers. Hence, a legal rule requiring the factory to install scrubbers, or a rule imposing liability for harm, may not be needed to achieve the result that the factory obtain scrubbers. The general possibility that externality problems may be avoided through bargaining, and that legal rules may not be necessary for their resolution, is known as the Coase Theorem, and was advanced in Coase (1960). However, much economic writing discusses reasons why bargaining may not cure externality problems: costs of bargaining (especially significant when the number of affected parties is large), breakdowns in negotiation (often arising when one side misgauges the other's situation), victims' lack of knowledge of danger (suppose the pollution is not apparent). Because bargaining will often fail to solve externality problems, attention has been paid to the use of legal rules (particularly liability rules, regulatory requirements, property rights) to accomplish their amelioration.

A very different issue in economic analysis of property law involves public property, which is to say, property like roads and libraries that the government needs to supply because the private sector will not provide them, or not in appropriate quantity. In this connection, the legal system sometimes allows the government to take land through its powers of eminent

domain, but requires it to pay compensation for its takings. According to economic analysis, the power of eminent domain may be socially desirable where government would have undue difficulty in making purchases. A standard example is that, in trying to purchase land for a road, government would be likely to be stymied by even a few landowners on the road's planned path (landowners might be tempted to hold out for strategic reasons). The requirement that government compensate for takings cannot be justified as it sometimes is, as an implicit form of insurance against takings; for, in the absence of a compensation requirement, individuals doubtlessly would purchase insurance against takings. A possible advantage of the requirement of compensation for takings is that it may discourage government from ill-advised takings, though that argument has been criticized.

Another area of property law concerns intellectual property rights: the law of patents, copyright, and trade secrets. The main theme of economic analysis of this body of law is that intellectual property rights have the beneficial effect of spurring innovations, but the detrimental effect associated with high prices and lower than socially desirable sales of goods incorporating the innovations. For example, economic thinking suggests that the copyright doctrine of 'fair use,' permitting for instance limited excerpting from a book for inclusion in teaching materials, might be socially desirable. This excerpting probably does not reduce book sales or the financial motive to author books (indeed, limited excerpting might increase sales of a book because it serves as a form of advertising), whereas the excerpting benefits teaching, because it means that teaching materials can be rapidly assembled without added cost. Most rules of intellectual property law are viewed against the background of their influence on incentives to innovate and on the sale and dissemination of goods embodying innovations.

Contract Law

A primary issue addressed in economic analysis of contract law is that of contract formation. The basic rule of contract formation is that a contract is legally recognized when and only when both sides have given explicit assent, such as by signing a document. This rule is said to be desirable for two reasons. First, it obviously enables parties to make a contract. Second, it protects parties from becoming bound against their wishes due to their having engaged in negotiations; this protection against unwanted obligations is beneficial because, without it, the negotiations that lead to the making of contracts would be inhibited.

Another issue surrounding contract formation concerns legal duties to disclose information. Economic analysis emphasizes that the social desirability of disclosure depends on the situation. For example, disclosure of a material defect (such as a leaky basement) in the home that the owner seeks to sell tends to be beneficial; for if the buyer knows about the defect, he can take steps to avert harm (avoid storing valuables in the basement or repair it). But should an oil company that learns, through costly investigative effort, that oil lies under a parcel of land, be required to disclose this information to the seller of the parcel? Perhaps not: the effect of such a disclosure obligation would be to make oil companies pay substantially more for parcels of land that they learn are romising, and thus to discourage expensive investigation of the location of oil deposits. Note from the foregoing that the economic analysis of legal rules about disclosure obligations concerns their effects on outcomes and does not derive from a possible moral call to tell the truth.

The most developed aspect of economic analysis of contract law deals with enforcement of contractual agreements. Enforcement is accomplished mainly by requiring parties who commit breach to pay the victims of breach for harm, to pay them 'damages.' One effect of the requirement to pay damages is that it induces contractual performance, which tends to raise the value of contracts to the parties and to society. A less obvious advantage of damage payments is that they constitute an escape hatch that parties can use when contractual performance becomes difficult, for they can breach and pay damages rather than bear very high costs to perform. This escape hatch element of damage payments also raises the value of contracts, as it makes parties more willing to assume contractual obligations. The escape hatch feature of damages for breach is also socially advantageous – it is not socially desirable for parties to perform when the cost of so doing outstrips the benefit to the recipient of performance. These points and others (notably, concerning risk allocation, and incentives to invest) regarding the virtues of payment of damages for breach have been analyzed intensively in the economic literature on contracts.

The orientation of economic analysis of contractual enforcement, through damage payments for breach, is very different from that of traditional legal analysis. Under the latter, damage payments for breach tend not to be regarded as incentives toward performance or as implicit escape hatches. Damage payments are seen primarily as compensation for harm or as proper desert for the wrong of breaking a promise. It should be added that, under the economic view, breach of a contract should not necessarily be identified with breaking a promise. The contracts that are written are not interpreted as detailed promises that parties truly want to be kept, but rather as incomplete promises that are only rough guides for behavior, and that the parties do not want to govern when performance would be very difficult.

Economic analysis of contracts has also been concerned with specific classes of contracts, including principal and agent contracts, insurance contracts, financial contracts, and donative contracts; the literature on some of these contractual contexts is now highly refined.

Litigation

One aspect of the economic analysis of litigation describes the motive to bring suits in terms of the potential plaintiff's costs of suit, the likelihood of success at trial, and the amount that would be obtained in the event of success.

Another element of the analysis of suit concerns the issue of whether the number of suits is socially excessive (is there a litigation explosion?) or perhaps socially inadequate. In this regard, it is observed that when a person considers suit, he does not factor in, as a cost to himself, the defendant's legal costs or the state's costs. This indicates that plaintiffs' incentives to bring suit may be socially excessive and thus that suit should be curtailed or barred in some domains. Yet a person considering suit will not usually take into account the deterrent value of the

suit – the message that suit will send, which will affect the behavior of others in the future – as well as any additional, wider social benefits. This divergence between the private and the social benefits from suit suggests that in some contexts the number of suits brought might be inadequate, and that public promotion of suit might be desirable.

Given that suit has been brought, the question arises whether the parties will settle their dispute or proceed to trial. The thrust of economic analysis of this question is that settlement is likely when the beliefs of the two sides about the trial outcome are similar, but that trial is likely when the plaintiff is much more optimistic than the defendant. For example, suppose that the beliefs of both sides are identical – each thinks the plaintiff would definitely obtain $100 000 at trial – and that the trial expenses of each would be $10 000. Then the plaintiff should be willing to accept as little as $90 000 (that is, $100 000 minus his litigation costs) in a settlement, and the defendant should be willing to pay as much as $110 000 ($100 000 plus his litigation costs). Hence, there should be room for settlement (any amount in between $90 000 and $110 000). If, however, the plaintiff's estimate of his winnings is much higher, say $200 000, than the $100 000 the defendant expects to pay at trial, the plaintiff would demand at least $190 000, which exceeds the $110 000 the defendant would be willing to pay. Accordingly, trial would be likely.

Economic analysis of litigation has also begun to address topics beyond the general ones of the bringing of suits and of settlement decisions. Among the topics considered are disclosure of information before trial, appeal of trial outcomes, class actions, and alternative dispute resolution.

Law Enforcement and Criminal Law

An additional area of economic analysis concerns public enforcement of law: the use of enforcement agents (such as police, safety inspectors, auditors) to detect violations of law; and the imposition of penalties for violations.

A theme of the literature on law enforcement is that the magnitude of penalties should be inflated from the level that would be appropriate were detection of violations certain. Inflation of penalties is needed to maintain deterrence, in effect to compensate for the possibility that a violator will not be detected. For example, if a polluting firm that causes harm of $10 000 is detected only a third of the time, then the fine that is imposed when the firm is detected should be not $10 000, but this amount multiplied by three, or $30 000. For, if the fine when the firm is detected is $30 000, the firm's probability-discounted, or average, fine is one-third of $30 000, or $10 000, providing it with incentives not to pollute similar to those that would exist if it paid a certain fine equal to the $10 000 harm.

Another point of emphasis in enforcement literature is that, because law enforcement is expensive, it will often be desirable for society not to spend so much on enforcement as to detect violations with high probability: it may be best, all things considered, to conserve on enforcement resources even though this means that many violators will escape detection. To combat the problem of inadequate deterrence that might accompany a low level of enforcement, multiplied penalties can be applied, as just discussed. Thus, there is appeal in employing an enforcement strategy that involves significant chances of escaping punishment combined with high levels of penalty. However, a problem with this low probability-high penalty strategy is that high penalties may be infeasible. Monetary penalties cannot exceed the assets of violators, which may be quite modest. Also, jail sentences can only be so long, and sanctions that conflict with retributive justice – sanctions that are out of proportion to the gravity of a bad act – might be resisted by the public.

A further issue examined in the economic literature on enforcement is the socially desirable choice between fines and imprisonment as forms of penalty. Here, it is generally said that fines are preferable to prison sentences, because fines do not themselves deplete social resources (but rather transfer command over resources from violators to the state). Whereas, imprisonment does diminish social resources, for prisons are expensive to operate, deprive individuals of their liberty, and prevent individuals from participating in the labor force. Therefore, the economic prescription is that fines be employed as the form of penalty when possible. But when fines cannot be used to deter, because the appropriate fines exceed the assets of violators (the typical robber could not be deterred by the threat of fines, given his level of assets), imprisonment should frequently be employed as the form of penalty. A closely connected point is that when fines cannot be used to deter, imprisonment may be useful as a penalty not only because it may deter, but also because it will incapacitate, that is, prevent individuals from doing further harm while in prison.

The foregoing conclusions and associated ones have been applied to criminal law. It has been suggested that many of the undesirable acts that are punished under criminal law (robbery, murder, rape) have the feature that civil suit and fines would not suffice to achieve adequate prevention of the acts. Therefore, imprisonment is often necessary to deter and to incapacitate those who commit the acts. Further, the magnitude of sanctions under criminal law has been related to the likelihood of detection of the acts, among other elements. Additionally, various doctrines of criminal law have been interpreted as desirable from the economic perspective. For example, that the sanction for murder committed in the heat of passion is less than that for premeditated murder is said to be rational: the ability to deter acts carried out in the heat of passion is relatively low, implying that it would be a mistake for society to incur the costs of the higher level of sanctions that are imposed for planned acts for which deterrence is more effective. Economic analysis of criminal law is concerned generally with the efficacy and the social costs of enforcement and the imposition of sanctions, and does not usually view punishment as a means of achieving retributive justice or other ideas of desert.

Additional Areas of Research

Economic analysis has been brought to bear on a host of other areas of law. These include many business-related areas of law, such as corporate law, tax law, antitrust law, and bankruptcy law, as well as, increasingly, diverse other areas, such as family law, anti-discrimination law, and constitutional law. Moreover, economic analysis has addressed questions surrounding the role of legislative bodies in formulating and enacting legal rules

and the role of courts and regulatory agencies in applying legal rules.

Foundations of, and Criticism about, Economic Analysis of Law

Economic analysis of law is premised on the general assumptions of the discipline of economics. These assumptions and their relationship to economic analysis of law in particular are sketched here. Certain commonly encountered criticisms of economic analysis of law are also mentioned.

Basic Assumptions

With regard to prediction of behavior, the usual assumption made in economics is, as noted at the outset, that parties are forward-looking and rational. This assumption is sometimes criticized as unrealistic. However, the assumption is usually made with the understanding that, although it is best for predicting central tendencies in behavior, various psychological and cognitive biases also influence behavior, and that these sometimes should be taken into explicit account. For example, the propensity to underestimate certain classes of risk, and thus for individuals not to take proper precautions or to insure adequately against them, has been recognized in economic analysis of law.

With regard to the evaluation of outcomes under the economic framework, the well-being or 'utility' of a person is basic. Economists' conception of utility is completely general and reflects not only the material pleasures of life to a person, but also, for example, the influence on his/her happiness of the treatment of others. From the utilities of individuals, a measure of social welfare is constructed, but there is no single, objective measure of social welfare that analysts study. Thus, utilitarianism is just one among a continuum of measures of social welfare that could be examined. The only significant presumption that is ordinarily made is that the measure of social welfare depends exclusively on the utilities of individuals. This assumption is consistent with concerns for equity in the distribution of utility and wealth, and economists have studied the implications of such concerns in depth.

Notions of Fairness and the Law

As has been seen in Sects. 1 and 2, classic notions of fairness, such as corrective and retributive justice, typically are omitted from the evaluation of legal rules under the economic framework, whereas these notions are traditionally viewed as of great significance to the assessment of the law. The essential reason that, under welfare economics, the notions of fairness are not accorded intrinsic importance is the assumption that they do not directly enter into individual's wellbeing. For instance, whether punishment is in proportion to the seriousness of a crime is ordinarily assumed not to affect individuals' utilities *per se;* rather, punishment may affect individuals' wellbeing through its deterrent or incapacitative effects. Because satisfaction of notions of fairness is presumed not to raise individuals' wellbeing in a direct manner, granting these notions independent weight in the evaluation of outcomes would tend to alter social decisions in ways that lower individuals' wellbeing.

Nevertheless, several qualifications to the last paragraph should be made. First, the assumption that notions of fairness do not matter to individuals may not always be apposite: individuals may have tastes for adherence to notions of fairness (individuals might feel happier if punishments fit crimes). To the extent that that is so, satisfaction of a notion of fairness properly enters into individual wellbeing and thus into social welfare, just as satisfaction of a taste for a material good does. (Notice that the importance of notions of fairness as personal tastes, being contingent on what the tastes of individuals happen to be, is different from the intrinsic importance accorded to conceptions of what is fair and right under deontological philosophical views.) Second, notions of fairness tend to have a desirable functional role (punishment only in proportion to the gravity of bad acts tends to discourage bad acts at relatively low social cost). Thus, advancing notions of fairness and inculcating them in the population (perhaps partly through adoption of legal rules that embody them) may serve to promote social welfare.

Income Distributional Equity and the Law

It may have been noted in Sects. 1 and 2 that the income distributional effects of legal rules were not mentioned as relevant to their evaluation under welfare economics, even though, as noted in Sect. 3.1, distributional equity does enter into the assessment of social welfare. The reason that the income distributional effects of legal rules are usually not considered in their evaluation is that economic analysis suggests that the income tax system (combined with income transfer programs) is a better means of achieving distributional objectives than the legal system. The income tax system is overtly redistributive, reaches all individuals, and is relatively cheap to administer. The legal system is not well designed to redistribute: it directly affects only those individuals who are involved in litigation; and even among them, the legal system is difficult to employ to effect redistribution, for a given class of litigants is often comprised of individuals with widely varying incomes (consider the class of victims of automobile accidents). Furthermore, the legal system is a very expensive device for transferring income. Hence, according to economic analysis, the legal system should not be used as a tool to achieve distributional goals, and if legal rules turn out to have undesirable distributional effects, these can be remedied through adjustment of the income tax system.

Economic Explanation of the Law

Finally, a strand of economic analysis should be mentioned claiming that the legal rules that are observed can be explained as those which best advance social welfare. This hypothesis seems attractive at a very gross level of description (for instance, that liability is imposed for causing harm, rather than for doing good, is explainable in the sense that such liability discourages harmful acts); and sometimes the hypothesis is appealing at a fairly detailed level. However, many, if not most, economic analysts hold a nuanced view of the economic rationality of the law, for many legal rules do not have obvious economic rationales, and a number undoubtedly reduce social wellbeing.

See also: Age Policy; Age, Sociology of; Aging and Health in Old Age; Differential Aging; Discrimination and the Law; Population Aging: Economic and Social Consequences; Psychiatry: Informed Consent and Care; Retirement and Health; Retirement, Economics of.

Bibliography

Becker, G.S., 1968. Crime and punishment: An economic approach. Journal of Political Economics 76, 169–217.

Bentham, J., 1789. An introduction to the principles of morals and legislation. In: Bentham, J. (Ed.), 1973. The Utilitarians. Anchor, Garden City, New York.

Bouckaert, B., De Geest, G. (Eds.), 2000. Encyclopedia of Law and Economics. Elgar, Cheltenham, UK.

Calabresi, G., 1970. The Costs of Accidents. Yale University Press, New Haven, CT.

Coase, R., 1960. The problem of social cost. Journal of Law and Economics 3, 1–44.

Cooter, R., Ulen, T., 1997. Law and Economics, second ed. Addison-Wesley, Reading, MA.

Kaplow, L., Shavell, S., 2001a. Economic analysis of law. In: Auerbach, A., Feldstein, M. (Eds.), Handbook of Public Economics. Elsevier, Oxford, UK.

Kaplow, L., Shavell, S., 2001b. Fairness versus welfare. Harvard Law Review 114, 961–1388.

Miceli, T., 1997. Economics of the Law. Oxford University Press, New York.

Newman, P. (Ed.), 1998. New Palgrave Dictionary of Economics and the Law. Stockton, New York.

Polinsky, A.M., 1989. An Introduction to Law and Economics, second ed. Little Brown, Boston.

Posner, R.A., 1972. Economic Analysis of Law, first ed. Little Brown, Boston.

Posner, R.A., 1998. Economic Analysis of Law, fifth ed. Aspen, New York.

Shavell, S., 1987. Economic Analysis of Accident Law. Harvard University Press, Cambridge, MA.

Law and Economics: Empirical Dimensions

Brian J Broughman and Kenneth G Dau-Schmidt, Maurer School of Law, Indiana University, Bloomington, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article surveys the importance of empirical work in the economic analysis of law. It describes the history of economics as a positive science and then explains how the positivist approach has in recent years led to an increased emphasis on empirical scholarship within the field of law and economics.

Introduction

Since the beginning of the twentieth century, economists have endeavored to conduct economics as a 'positive science.' Under this project, assumptions about reality are combined with the logic of optimization theory to deduce positive predictions about how people will respond to changes in the relevant economic environment. Empirical work testing either the realism of the assumptions or the veracity of the predictions is essential to this undertaking, to ensure that the economic model is a useful abstraction of the examined problem.

As in the consideration of other problems, empirical work is essential to the economic analysis of law. Indeed, because many legal topics are outside the traditional subject matter studied by economists, the traditional assumptions of the neoclassical economic model do not always fit well in the analysis of these problems, and empirical work is especially important to make sure that the logical conclusion of the economic model is relevant to the examined problem. Also, owing to the overlap of the examined subject matter with that studied by other disciplines, the need to examine basic assumptions and the shared statistical methods, empirical work in law and economics can serve as a touchstone for multidisciplinary discourse on law.

Economic Analysis

Economics first developed as a discipline in the seventeenth century, in close association with political science and sociology. During this early period of development, the social sciences enjoyed considerable overlap in examined problems and methodology. Although some works from this time can be identified as being distinctly 'economic,' early social scientists such as Adam Smith (1723–90), Karl Marx (1818–83), and John Stuart Mill (1806–73) hypothesized about economic, social, and political institutions with equal skill and aplomb. This close association between the social sciences continued into the beginning of the twentieth century with the work of institutional economists such as John R. Commons, and economically oriented sociologists such as Emile Durkheim and Max Weber. Indeed, Weber, one of the most distinguished sociologists of his time, held several academic positions as an economist.

However, early in the twentieth century, economics began to separate from the other social sciences in two important ways. First, economics began to diverge in terms of its examined subject matter. Economists began to focus more on studying the machinations of the market, leaving subjects such as social relationships and political economy to the other disciplines. Since the work of Gary Becker (1968, 1976), however, this trend has been reversed, so that now economists study a wide variety of social and political phenomena, including the law. Second, economics began to diverge in terms of its adopted methodology. Although all of the social sciences adopted modern statistical methods for analyzing quantitative data, economists also undertook to develop a 'positive science' of economics under the neoclassical model.

Positive Economics

Since early in the twentieth century, most economists have endeavored to develop economics as a positive science analogous to the physical sciences. The ultimate goal of a 'positive science' is to produce theories that describe observed phenomena accurately, rather than offer a normative evaluation of the desirability of the phenomena. Such theories are devised by making assumptions about the world that abstract the essential features of the examined problem, and then using logic to deduce predictions about the as yet unobserved events. The validity of the theory is then confirmed through empirical work, which examines the predicted events and tests the statistical significance of results consistent with the theoretical model. This deductive process, working from theory to data, can be distinguished from the inductive process used by many other social scientists who tend to analyze the data before inductively deriving their theories. Accordingly, for the majority of the twentieth century, most economists aspired to conduct economics as a positive science that deduces empirical predictions from theory and then tests those predictions against data.

The Neoclassical Model

In the development of economics as a positive science, the dominant paradigm has been the 'neoclassical model.' The neoclassical model consists of a set of assumptions that economists make in analyzing problems. This set of assumptions achieved paradigmatic dominance in economics because they proved useful in analyzing market behavior, and because they were mathematically tractable. Under the neoclassical model, economists generally assume that the appropriate unit for

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86049-3

analysis is the individual consumer or firm, individuals and firms rationally choose among available alternatives to select the option that maximizes utility or profits, people's preferences are exogenous to the model, and transaction and information costs are zero. The economists' notion of 'rationality' assumes that people have preferences with respect to all possible states of the world that these preferences are reflexive and transitive, and that the people will select the option they most prefer. The assumption of free information is usually referred to as the assumption of 'perfect information.' It has deceptively broad-ranging implications in that it implies that each person knows the quality of each good and the price of all alternatives, and indeed has the knowledge necessary to enter the market and produce the good if the market for the good enjoys supercompetitive profits. The assumption of zero transaction costs also has potentially far-reaching implications since, as Ronald Coase (1960) has pointed out, in a world of zero transaction costs, everything happens at once. To be sure, the above characterization of the neoclassical model is a crude abstraction from reality. Many of the primary advances in the field of economics over the past century respond to such concerns by refining the neoclassical model to account for monopoly power, asymmetric information, nonrational behavior, principle agent conflicts, and transaction costs. These refinements allow a researcher to model an economic problem with more realism. Nonetheless, the neoclassical model remains a basic starting point for economic theory.

The conduct of economics as a positive science using the neoclassical model tends to separate the work of economists from that of other disciplines. The adoption of the perspective of the individual as the unit of analysis, and the assumption that preferences are exogenous, preclude consideration of the effect of social norms and the process of socialization that are so important to many sociological analyses. The construction of the concept of individual utility based merely on an ordinal preference ranking was undertaken with the purpose of avoiding some of the stickier problems in psychology, thus making the concept 'psychology-free.' Although the assumption of 'rational choice' has become popular in explaining results or building models in many disciplines, including sociology and political science, the economists' well-defined concept of rationality is generally narrower than that used in other disciplines. Moreover, since the other social sciences rely on inductive logic, working from data to theory, they generally make fewer assumptions about what constitutes rational choice. Accordingly, some sociologists have argued that they study rather than assume human rationality.

The Importance of Empirical Work

The assumptions of the neoclassical model are clearly unrealistic, and this fact raises the importance of empirical work in economics. Every analysis of a problem is necessarily an abstraction from reality, ignoring some of the complexities of the real word to focus on others. The art of modeling is to know which simplifying assumptions one can make and still capture the essence of the problem. Of course, an important test of how well a theory captures the essence of a problem is how well it fits the available data. Milton Friedman (1953) has argued that the realism of a model's underlying assumptions is of little importance as long as the model yields good predictive results. Indeed, Friedman has gone so far as to argue that, paradoxically, the most important hypotheses have the most unrealistic assumptions, because a hypothesis is important if it abstracts the crucial elements of a phenomenon in such a way as to explain much of the phenomenon by assuming little. Other economists have argued that there are purposes to modeling beyond mere prediction, such as explaining the phenomenon or providing a catalog of knowledge for future work. As a result, they have argued that there are reasons for considering the realism of economic assumptions that go beyond the simple predictive value of the model.

The Economic Analysis of Law

Beginning with the pioneering work of Gary Becker (1968, 1976), economists have sought to broaden the scope of their inquiry to all aspects of society. This movement has brought economics into a variety of subject matters studied by other disciplines, including the family, education, discrimination, politics, and – of particular relevance to this article – law.

Interest in applying economic analysis to law, however, has not been limited to economists. Indeed traditional legal scholars, and even judges, have shown interest in using economics to help guide legal analysis and law reform efforts. Economic models can help untangle the incentives and predict the efficiency consequences associated with different legal rules. Viewed in this way, economic analysis provides a normative criterion – efficiency – for legal scholars and judges. Despite the fact that law and economic theory has, since the 1960s, had a large impact on legal scholarship, empirical testing of such theory had been relatively limited until recently. In the past decade, however, there has been a large increase in empirical scholarship within law and economics.

Foundational Research in Law and Economics

The law and economics movement had its origin in the work of Gary Becker, Ronald Coase, and Guido Calabresi. Becker (1968) used simple principles of labor economics to analyze the criminal's decision of whether to engage in crime, and derived implications for optimal penalties and optimal enforcement of the law. Coase's (1960) work established the proposition known as the 'Coase Theorem,' showing that, in a world of zero transaction costs, the distribution of resources is invariant to the rule of law. However, Coase later made clear that he considers his work concerning how to minimize the effects of transaction costs in assigning the rule of law under the more realistic assumption of positive transaction costs to be more important. The 'Coase Theorem with positive transaction costs' posits that the efficient distribution of resources will not necessarily occur under any legal rule and that society should minimize the impact of transaction costs by assigning entitlements away from the party that can most efficiently remedy any externality. Calabresi (1970) wrote a classical study analyzing the costs of accidents and suggesting ways in which tort law might be formulated to minimize those costs.

After these initial contributions, an extensive body of work, primarily by legal scholars, has applied the principle of wealth maximization to various legal problems, explaining either how the law can be understood as promoting wealth maximization (Posner, 1998; Rubin, 1977), or how it might be amended to better achieve that objective. The former approach – finding the underlying efficiency within the common law – is sometimes referred to as the positive or Chicago school of law and economics; while the later approach – highlighting market failures and suggesting legal reform to correct such failures – is sometimes referred to as the normative or Yale school of law and economics (Parisi, 2004). These labels, however, can be confusing since some arguments within the so-called positive school are effectively normative justification of the status quo, sometimes critiqued as 'just-so stories' (Schroeder, 2000; Hovenkamp, 1990). In this article, we use the label 'positive' as it is typically used by economists.

This body of research by legal scholars has primarily focused on the common law. Notable examples highlight the tradeoff between negligence and strict liability in tort law (Shavell, 1980); provide a theory for efficient breach of contract (Birmingham, 1969); and analyze the distinction between property and liability rules (Calabresi and Melamed, 1972). Cooter and Ulen (2012) provide a textbook overview of foundational research in law and economics, and several additional contributions are summarized and excerpted by Katz (1998).

Empirical Work in Law and Economics

Until recently, empirical work in the economic analysis of law has been rather slow to follow these theoretical advances. Nonetheless, some notable early studies have been done directly on issues important to the economic analysis of law. One of the earliest significant empirical works in law and economics was Isaac Ehrlich's (1975) work showing a deterrent effect from capital punishment. Ehrlich's work prompted enormous academic and public debate on the subject, and his results were later disputed by other scholars. A number of scholars have done empirical work in an attempt to verify or disprove the fundamental implication of Coase's work that, when transaction costs are negligible, the distribution of resources is invariant with respect to the rule of law. Among these works, perhaps the most notable for the purposes of this article is Robert Ellickson's (1986) on the distribution of cattle in Shasta County, California, under alternative rules of free range and nonfree range. Ellickson found that Coase was correct that changes in the rule of law did not affect the distribution of cattle; however, Ellickson attributed this finding more to the sociological reason that neighbors had informal means of dealing with animal trespasses that were not influenced by changes in the law, than Coase's reasoning that neighbors would just bargain around inefficient assignments of entitlements. Kip Viscusi (1983, 1992) has also made important empirical contributions in the economic analysis of legal issues through his analysis of the valuation of life and limb and the relative magnitude of various risks. Viscusi's work is important in evaluating the efficiency of tort law, workers' compensation laws, and government regulation of various workplace and environmental risks. Finally, John Donohue (1998) has done important empirical work documenting employment discrimination and the positive impact of law in ameliorating this problem.

The above were certainly not the only notable empirical studies during this time. Nonetheless, the relative paucity of empirical work during this period is worth noting. Writing in 1989, Posner (1989: 58–59) laments that "it continues to be the case that a disproportionate share of economic studies of law lack a substantial empirical component." This problem continued a decade later. Using data from 1972 to 2002, Landes (2003) finds that the frequency of empirical work in peer-reviewed law and economics journals is significantly lower than in economics journals generally. By overlooking the vast body of law and economics research published in law reviews, which are significantly less likely to publish empirical work, Landes (2003) likely understates the gap in empirical research. At least during the twentieth century, law and economics research was subject to less empirical scrutiny than most work in the broader discipline of economics.

There may be several overlapping reasons for the relatively low level of empirical output during the above time period, which possibly include a lack of training in empirical methods among legal academics, a lack of resources dedicated to such work, a lack of preexisting databases during this period for empirical analysis in law, a lack of prestige and incentives attached to such work (Landes, 2003), and perhaps even some fear of exposure to falsification through replication on the part of other scholars. But whatever the exact reason the situation has changed markedly in the 10 years since Landes' (2003) study.

In the past decade, law and economics scholars have produced empirical work at a much higher rate than previously. This is part of a broader move toward empirical scholarship within legal academia generally (George, 2006); but the effect is particularly pronounced within the law and economics community. Following Landes' methodology, we calculate the percentage of papers published from 2011 through the first issue of 2013 in *The Journal of Legal Studies* (JLS) – a leading law and economics publication – which contain empirical content. Out of a total of 37 JLS articles during this recent period, we found that 24 (or 65%) contained a substantial empirical component. By contrast from 1972 to 2002, Landes (2003) only found that 38% of JLS articles contained empirical content. This difference in the rate of empirical scholarship – 38% vs 65% – between the two periods is statistically significant at the 1% level (difference of means t-test), and suggests that there has been a meaningful change in the type of scholarly output produced by the law and economics community. We do not believe this result is limited to JLS, but rather reflects a broader trend in the law and economics community.

Recent empirical work in law and economics takes a number of different approaches. Some articles use data to test predictions from theory, while other projects – such as the following example investigating the cost of medical malpractice – examine policy implications stemming from law reform. Using a comprehensive database of claim outcomes from medical malpractice suits brought in Texas from 1988 to the present, Black et al. (2005) find no evidence of a malpractice insurance crisis; damage awards did not increase from the late 1980s to the present. Furthermore, damage caps – introduced

in Texas in 2003 – appear to have little impact on the overall cost of health care (Paik et al., 2012). These studies have important policy implications for tort reform.

Other scholars investigate whether patent protection stimulates innovation and restricts output as predicted by economic theory. For example, Moser (2012) uses a large data set of British and American inventions presented at world's fairs between 1851 and 1915, and finds the majority of such inventions were not protected by patent. In a separate study, Lakdawalla and Philipson (2012) present data from the pharmaceutical industry showing that output often fails to rise after patent expiration, and suggesting that this result is due to reduced incentives for marketing nonpatented drugs. Both Moser (2012) and Lakdawalla and Philipson (2012) find results contrary to basic predictions from economic theory, and highlight a need for a more refined economic theory of IP protection.

Most empirical work in law and economics is quantitative in nature, and typically involves large-n databases. While working with such data has a number of benefits related to replication and hypothesis testing, it may cause law and economics scholars to overlook topics where quantitative data are unavailable. In contrast to this trend, some empirical work in law and economics does involve qualitative research. Baker and Griffith (2006, 2007), for example, investigate the market for directors' and officers' (D&O) liability insurance policies by conducting semistructured interviews with professionals in the D&O insurance industry. The interviews provide an excellent insight – impossible to glean from quantitative data – into how insurance underwrites price and monitor the risk of corporate governance failure (Baker and Griffith, 2006, 2007). We believe law and economics empirical work would benefit if more projects took advantage of qualitative evidence, possibly alongside quantitative data, to help ensure that the theory is grounded in reality. This quibble aside, the recent increase in empirical research in law and economics is noteworthy.

We consider a number of developments that can help explain why empirical work has become more prevalent within the law and economics community over the past 10 years. First, over the past decade there has been a large increase in joint degree law professors who in addition to a JD have a Ph.D in economics or a related field (Hersch and Viscusi, 2012). Scholars with an economics Ph.D are trained in modern statistical methods and bring with them professional norms of the economics community, including a general acceptance of a positivist approach to research. Furthermore, the increasing presence of Ph.D scholars in law schools has brought greater acceptance within the legal academy of publishing in disciplinary and interdisciplinary journals as opposed to law reviews, raising the reward and lowering the risk of doing empirical work.

Second, under the deductive method favored by positive economics one first needs theory before empirical testing is even possible. Once there are multiple competing models analyzing a given subject, however, there arises an increased need for empirical testing to determine which model more accurately predicts behavior. For example, consider Waldfogel (1998), who uses data on the selection of cases for trial to test two prominent and competing hypotheses – Priest and Klein (1984) vs Bebchuk (1984) – for the choice between settlement and trial. Waldfogel's empirical research was only possible after the underlying theoretical work had been developed. Thus, even within the positivist framework it makes sense that early scholarship in a new field, such as law and economics, would be primarily theoretical, and would then be followed by more empirical testing as the field matures, and as novel theoretical contributions are harder to make. The relative need for theoretical as opposed to empirical research has changed as law and economics has matured.

Third, changes or discontinuities in the law can be particularly useful for purposes of statistical identification. Identification is the strategy by which an empirical study isolates causation as opposed to endogenously driven correlation. In practice, valid identification is a difficult hurdle to clear; however, as statistical methods have substantially improved in recent years the economics community increasingly requires valid identification for publication in top journals. Interestingly, changes in the law or discontinuities in legal treatment are some of the most widely used identification strategies by economic researchers. The use of law for identification purposes has brought several econometricians to the law and economics field. One notable example of such work is Greenstone et al. (2006), which uses a 1964 change in US securities regulation to isolate the effect of mandatory disclosure. The new law extended mandatory Securities Exchange Commission (SEC) disclosure to some, but not all, over-the-counter (OTC) companies. This creates a natural experiment since there is both a control group (the OTC companies unaffected by the new law), and a treatment group (the OTC companies required to comply with SEC disclosure starting in 1964). The study found that treatment group companies experienced positive abnormal returns around the passage of the new law, suggesting that investors value mandatory disclosure (Greenstone et al., 2006). Studies of this sort – which use a change or discontinuity in the law for purposes of statistical identification – are part of a growing trend, increasing both the quantity and quality of empirical research in law and economics.

A final consideration leading to more empirical work in law and economics is the emergence of specialized journals and numerous conferences devoted to empirical legal studies (ELS). The *Journal of Empirical Legal Studies* was first published in 2004 and the Conference on Empirical Legal Studies was first held in 2006. Both the conference and the journal have already had a large impact on legal scholarship, and have been followed by a number of law review symposiums devoted to ELS (Eisenberg, 2011). To be sure, the ELS movement represents a broad trend in legal academia and is not exclusively linked to law and economics. Nonetheless, there is a large overlap between ELS and the law and economics movement, creating additional publication and presentation outlets for empirical research in law and economics.

Empirical Work in Law and Economics as a Touchstone for a Multidisciplinary Discourse on Law

It has become increasingly apparent that empirical work in law and economics can serve as a touchstone for interdisciplinary discourse on law. This is true for several reasons. First,

economics shares the same accepted statistical methods for analyzing data with other social science disciplines, facilitating dialogue between empirical researchers studying the law regardless of the researchers' background. Second, grounding in empirical research encourages law and economics scholars to relax some of the assumptions of the neoclassical model. The traditional economic model of individual rational choice does not always capture the essence of legal problems in the way that it did in the study of atomistic markets.

By relaxing some of the assumptions of the neoclassical model, economists have found that their work becomes less distinct from that of theoretical work in other disciplines. This is true not only for the reason that such economic work is inductive rather than deductive in nature, working from the data to theory, but also because relaxation of some of the assumptions of the neoclassical model allows for the consideration of work from other disciplines. In particular, some law and economics scholars have found connections between their work and the work of sociologists who study law, while other law and economics scholars have made use of empirical work out of the behavioral sciences.

Sociological Influences in the Economic Analysis of Law

Some law and economics scholars have drawn on the work of sociologists. Interest among law and economics scholars in the influence of social norms and the role of law in the socialization process is one example of this trend. In order to examine these phenomena, the economist usually has to relax one or more of the traditional assumptions of the neoclassical model. For example, to examine the influence of social norms on people's behavior, the economist might adopt a perspective of analysis that takes into account group or community interests, rather than just individual interests. Similarly, in examining the role of law in the socialization process, some economists have found it useful to relax the assumption that people's preferences are exogenously determined. With the relaxation of these assumptions, it is also necessary generally to adopt an objective function other than individual utility, wealth, or profit for maximization in the economic model – for example, net social benefits or social welfare.

Although some very interesting work in the study of norms from the economic perspective has been done by McAdams (1997), Cooter (1995), and Posner (1996), perhaps the seminal work on the subject was the work previously alluded to by Robert Ellickson (1986). To test the Coase theorem, Ellickson surveyed land owners in Shasta County, California, as to the number of animal trespass cases they had experienced under different legal regimes, and how those incidents were resolved. Ellickson found that the allocation of resources was indeed invariant to the legal rule, as predicted by Coase. However, Ellickson (1986) found that this result, obtained not because transaction costs were low and the parties had bargained around the law to reach an efficient allocation of resources, but instead because resort to the law was costly, and as an alternative the affected parties were committed to a norm of cooperation with their neighbors, regardless of the rule of law. As a result, Ellickson concluded that law and economics would benefit by considering previous work on such questions in sociology, and the law and society movement. Taking his own suggestion to heart, Ellickson (1998) has used empirical work from other disciplines on norms of 'fairness' and 'civic responsibility' to analyze a variety of legal problems including rent control, the doctrine of unconscionability, the implied covenant of good faith and fair dealing, the duty to rescue, political and judicial processes, and the First Amendment.

Another area in which the economic analysis of law has shown connections to work in sociology is the analysis of the role of law in the process of socialization. In this area, one of the present authors has written an economic analysis of the criminal law as a preference-shaping policy (Dau-Schmidt, 1990). If one relaxes the neoclassical assumption that preferences are exogenous to the model, and instead assumes that criminal law is intended to discourage preferences for criminal behavior, as well as to raise the costs of engaging in such behavior, then one can explain many of the fundamental doctrines of criminal law. Further work on how the law shapes preferences or helps establish new norms – sometime referred to as the 'expressive function of law' – has been done by Cooter (1998) and Sunstein (1996); and empirical testing of the expressive function of law is found in Feldman (2009) (suggesting based on field experiments that the expressive impact of trade secret laws operates primarily through their impact on morality), Funk (2007) (laws removing the obligation to vote led to significantly lower levels of voter turnout in Switzerland, even though the former voting requirement was only symbolic and never enforced by fine), and Wittlin (2011) (finding that seat belt laws increase seat belt usage independent of enforcement). This work lends itself naturally to integration with work from sociology, because it raises questions concerning how law impacts socialization and how law shapes both norms and individual preferences.

Behavioral Law and Economics

Other economists have drawn on the empirical work of behavioral psychologists in their analysis of law. In 'laboratory' experiments, behavioral psychologists such as Amos Tversky and Daniel Kahneman (Tversky and Kahneman, 1982) have established that the human decision-making process typically varies in predictable ways from the simple model of rational choice assumed in the neoclassical economic model. In particular, they have established that people commonly exhibit 'bounded rationality' in that they can only process a finite amount of information and thus must rely on rules of thumb to help them make decisions, 'bounded willpower' in that they sometimes do things that are clearly not in their long-term interests, and 'bounded self-interest' in that they care about other people and whether the treatment in their relationship is reciprocally 'fair.' Tversky and Kahneman have also established that people evaluate outcomes based on the change they represent from an initial 'reference point,' weighting potential losses from the reference point more heavily than potential gains. This finding is referred to as 'prospect theory' in the literature, and it can be exploited to manipulate people's decision-making process by suggesting the initial reference point to 'frame' possible choices as gains or losses.

As with the influence of sociology on the economic analysis of the law, there has been a boom in the number of scholars who seek to use 'behavioral economics' in their analyses of the

law. Early proponents of such work include Thomas Ulen (1989) and Robert Ellickson (1989), but they have since been joined by others including Russell Korobkin (see Korobkin and Ulen, 2000), and Christine Jolls et al. (1998). These scholars argue that amendment of the assumptions of the neoclassical model concerning individual rational choice according to the empirical findings of behavioral economics is necessary to improve the predictive power of the economic model. Their critique is fundamental, in that it goes back to basic assumptions about the purposes of law, and how people will react to and enforce the law. For example, Jolls et al. (1998) have argued that the Coase theorem is often empirically false, because empirical work in behavioral economics suggests that people sometimes develop attachments to objects they possess, and this 'endowment effect' may mean that people must sometimes be paid more to part with an entitlement than they would pay to possess the entitlement. Accordingly, people will not bargain around the assignment of entitlements in the way described by Coase. Similarly, Korobkin has argued that information heuristics will affect the efficacy of tort damages in providing efficient incentives to take care. Because work in behavioral economics shows that people are often overly optimistic in that they undervalue the chances of something bad happening, and overvalue the chances of something good happening, this suggests that, even if tort damages are set at efficient levels, people will not account adequately for the chance of accident and so will not take efficient care.

An interesting behavioral update to Becker's (1968) classic study of deterrence comes from a recent empirical study by Lee and McCrary (2005). Becker (1968) assumes that the same level of deterrence can be achieved either by increasing the likelihood of punishment or by increasing the length of sentence. Lee and McCrary (2005) take advantage of the fact that young offenders are treated as adults – and are thus subject to substantially longer periods of incarceration – after their eighteenth birthday. Thus, Becker's theory would predict a large deterrent effect in comparing 18-year-olds to 17-year-olds. Instead, Lee and McCrary find that the severity of punishment has very little impact on offending rates, whereas increasing the likelihood and immediacy of detection can have a large deterrent effect. Potential offenders appear to be extremely myopic or impatient, and do not behave consistently with standard rational actor assumptions. From a positivist perspective, this suggests that Becker's theory of deterrence, while interesting, is not a predictively useful abstraction of reality. Future theoretical work is needed to better account for the behavioral pattern found in Lee and McCrary's (2005) work.

Conclusion

This article surveys empirical research in the economic analysis of law. We describe the history of economics as a positive science and then explain how the positivist approach has in recent years led to an increased emphasis on empirical law and economics scholarship. We discuss trends regarding empirical methodologies used in law and economics, and also explain how empirical work in law and economics can serve as a touchstone for a multidisciplinary discourse on law.

See also: Behavioral Economics; Law and Economics; Law and Society; Positivism, Sociological.

Bibliography

Baker, Tom, Griffith, Sean, 2006. The missing monitor in corporate governance: the directors' & officers' liability insurer. Georgetown Law Journal 95, 1795–1842.

Baker, Tom, Griffith, Sean, 2007. Predicting corporate governance risk: evidence from the directors' and officers' liability insurance market. University of Chicago Law Review 74, 487–544.

Bebchuk, Lucian, 1984. Litigation and settlement under imperfect information. RAND Journal of Economics 15, 404–415.

Becker, Gary, 1957. The Economics of Discrimination. University of Chicago Press, Chicago.

Becker, Gary, 1968. Crime and punishment: an economic approach. Journal of Political Economy 76, 169–217.

Becker, Gary, 1976. The Economic Approach to Human Behavior. University of Chicago Press, Chicago.

Birmingham, Robert, 1969. Breach of contract, damage measures, and economic efficiency. Rutgers Law Review 24, 273.

Black, Bernard, Silver, Charles, Hyman, David, Sage, William, 2005. Stability, not crises: medical malpractice claim outcomes in Texas, 1988–2002. Journal of Empirical Legal Studies 2, 207–259.

Calabresi, Guido, 1970. The Costs of Accidents: A Legal and Economic Analysis. Yale University Press, New Haven, CT.

Calabresi, Guido, Melamed, Douglas, 1972. Property rules, liability rules, and inalienability: one view of the cathedral. Harvard Law Review 85, 1089–1128.

Coase, Ronald, 1960. The problem of social cost. The Journal of Law and Economics 3, 1–44.

Cooter, Robert, 1995. Decentralized law for a complex economy: the structural approach to adjudicating the new law merchant. University of Pennsylvania Law Review 144, 1643.

Cooter, Robert, 1998. Expressive law and economics. Journal of Legal Studies 27, 585.

Cooter, Robert, Ulen, Thomas, 2012. Law & Economics, sixth ed. Addison-Wesley.

Cotterell, Roger, 1995. Sociological interpretations of legal development. European Journal of Law and Economics 2, 347–359.

Dau-Schmidt, Kenneth G., 1990. An economic analysis of the criminal law as a preference shaping policy. Duke Law Journal 1990, 1–38.

Dau-Schmidt, Kenneth G., 1991. Relaxing traditional assumptions and values: toward a new multidisciplinary discourse on law. Syracuse Law Review 42, 181–195.

Dau-Schmidt, Kenneth G., 1997. Economics and sociology: the prospects for an interdisciplinary discourse on law. Wisconsin Law Review 1997, 389–419.

De Geest, Gerrit, 1995. Toward an integration of economic and sociological approaches. European Journal of Law and Economics 2, 301–308.

Donohue, John, 1998. The legal response to discrimination: does law matter? In: Garth, Sarat (Eds.), How Does Law Matter? Northwestern University Press, Chicago, pp. 45–75.

Ehrlich, Isaac, 1975. The deterrent effect of capital punishment: a question of life and death. American Economic Review 81, 397–417.

Eisenberg, Theodore, 2011. The origins, nature, and promise of empirical legal studies and a response to concerns. University of Illinois Law Review 2011, 1713–1738.

Ellickson, Robert C., 1986. Of Coase and cattle: dispute resolution among neighbors in Shasta County. Stanford Law Review 38, 623–687.

Ellickson, Robert C., 1989. Bringing culture and human frailty to rational actors: a critique of classical law and economics. Chicago-Kent Law Review 65, 23–55.

Ellickson, Robert C., 1998. Law and economics discovers social norms. The Journal of Legal Studies 27, 537–552.

Feldman, Yuval, 2009. The expressive function of trade secret law: legality, cost, intrinsic motivation, and consensus. Journal of Empirical Legal Studies 6, 177–212.

Friedman, Milton, 1953. The methodology of positive economics. In: Friedman (Ed.), Essays in Positive Economics. University of Chicago Press, Chicago, pp. 1–38.

Funk, Patricia, 2007. Is there an expressive function of law? An empirical analysis of voting laws with symbolic fines. American Law and Economics Review 9 (1), 135–159.

George, Tracey, 2006. An empirical study of empirical legal scholarship: the top schools. Indiana Law Journal 81, 141–161.

Greenstone, Michael, Oyer, Paul, Vissing Jorgensen, Annette, 2006. Mandated disclosure, stock returns, and the 1964 securities act amendments. The Quarterly Journal of Economics 121, 399–460.

Heise, Michael, 1999. The importance of being empirical. Pepperdine Law Review 26, 807–834.
Hersch, Joni, Viscusi, W. Kip, 2012. Law and economics as a pillar of legal education. Review of Law & Economics 8 (2), 487–510.
Hovenkamp, Herbert, 1990. Positivism in law & economics. California Law Review 78, 815–852.
Jolls, Christine, Sunstein, Cass R., Thaler, Richard, 1998. A behavioral approach to law and economics. Stanford Law Review 50, 1471–1550.
Kahneman, Daniel, Tversky, Amos, 1979. Prospect theory: an analysis of decision under risk. Econometrica 47, 263–292.
Katz, Avery, 1998. Foundations of the Economic Approach to Law. Foundation Press, New York.
Korobkin, Robert, Ulen, Thomas, 2000. Law and behavioral science: removing the rationality assumption from law and economics. California Law Review 88, 1051.
Lakdawalla, Darius, Philipson, Tomas, 2012. Does intellectual property restrict output? An analysis of pharmaceutical markets. The Journal of Law and Economics 55, 151–187.
Landes, William, 2003. The empirical side of law and economics. University of Chicago Law Review 70, 167–180.
Lee, David, McCrary, Justin, 2005. Crime Punishment, and Myopia. NBER Working Paper No. 11491.
Levitt, Steven, Dubner, Stephen, 2006. Freakonomics: A Rogue Economist Explores the Hidden Side of Everything. HarperCollins, New York.
Macneil, Ian R., 2000. Other sociological approaches. In: Bouckaert, Geest, De (Eds.), Encyclopedia of Law and Economics. Edward Elgar, Cheltenham, UK.
McAdams, Richard, 1995. Cooperation and conflict: the economics of group status production and race discrimination. Harvard Law Review 108, 1005–1089.
McAdams, Richard, 1997. The origins, development and regulation of norms. Michigan Law Review 96, 338–433.
Moser, Petra, 2012. Innovation without patents: evidence from world's fairs. The Journal of Law and Economics 55, 43–74.
Paik, Myungho, Black, Bernard, Hyman, David, Silver, Charles, 2012. Will tort reform bend the cost curve? Evidence from Texas. Journal of Empirical Legal Studies 9, 173–216.
Parisi, Francesco, 2004. Positive, normative and functional schools in law and economics. European Journal of Law and Economics 18, 259–272.
Posner, Eric, 1996. The regulation of groups: the influence of legal and nonlegal sanctions on collective action. University of Chicago Law Review 63, 133–197.
Posner, Richard, 1989. The future of law and economics: a comment on Ellickson. Chicago-Kent Law Review 65, 57.
Posner, Richard, 1998. Economic Analysis of Law, fifth ed. Aspen Law and Business, New York.
Priest, George, Klein, Benjamin, 1984. The selection of disputes for litigation. Journal of Legal Studies 13, 1–55.
Rubin, Paul, 1977. Why is the common law efficient? The Journal of Legal Studies 6, 51–63.
Schroeder, Jeanne, 2000. Just so stories: Posnerian methodology. Cardozo Law Review 22, 351.
Shavell, Steven, 1980. Strict liability versus negligence. Journal of Legal Studies 9, 1–25.
Sunstein, Cass, 1996. On the expressive function of law. University of Pennsylvania Law Review 144, 2021–2053.
Tversky, Amos, Kahneman, Daniel, 1982. Judgment under uncertainty: heuristics and biases. In: Tversky, Slovic, Kahneman (Eds.), Judgment under Uncertainty. Cambridge University Press, Cambridge, UK.
Ulen, Thomas, 1989. Cognitive imperfections and the economic analysis of law. Hamline Law Review 12, 385–410.
Viscusi, W. Kip, 1983. Risk by Choice: Regulating Health and Safety in the Workplace. Harvard University Press, Cambridge, MA.
Viscusi, W. Kip, 1992. Fatal Tradeoffs: Public and Private Responsibilities for Risk. Oxford University Press, New York.
Waldfogel, Joel, 1998. Reconciling asymmetric information and divergent expectations theories of litigation. Journal of Law and Economics 41, 451–476.
Wittlin, Maggie, 2011. Buckling under pressure: an empirical test of the expressive effects of law. Yale Journal on Regulation 28, 419.

Law and Emotion

Susan A Bandes, DePaul University College of Law, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The field of law and emotion draws from a range of disciplines in the sciences, social sciences, and humanities to shed light on the emotions that pervade the legal system. It utilizes insights from these disciplines to identify and assess the implicit and explicit assumptions about emotion that animate legal reasoning, legal norms, legal doctrine, the behavior of legal actors, and the structure of legal institutions.

Introduction

The field of law and emotion draws from a range of disciplines in the sciences, social sciences, and humanities to shed light on the emotions that pervade the legal system. It utilizes insights from these disciplines to identify and assess the implicit and explicit assumptions about emotion that animate legal reasoning, legal doctrine, and the behavior of legal actors. In light of law's focus on influencing social norms and on structuring effective and just institutions, one development that holds enormous promise is the growing interdisciplinary interest in collective decision-making and in the emotional dynamics of groups. Work in the affective sciences (an emerging interdisciplinary field composed of psychologists, neuroscientists, anthropologists, and other researchers concerned with the role of emotion in human behavior and decision-making) on how emotion and cognition interact is another rich vein for legal scholars interested in the assessment of responsibility and blame, the role of morality in law, and a host of other areas. By reevaluating legal doctrine and policy in light of relevant knowledge from psychology, neuroscience, sociology and other fields, law and emotion scholarship aims to contribute to a more informed and effective framework for refining legal doctrine and reforming legal institutions.

Most fundamentally, law and emotion scholarship challenges standard legal accounts of rational deliberation and legal reasoning. As a descriptive matter, it challenges the deeply ingrained belief that emotion is a hindrance to – or even the antithesis of – rationality. It emphasizes that this belief is increasingly untenable in light of the emerging consensus in the sciences and social sciences that emotion is an inextricable part of the cognitive process. On the normative level, it explores the appropriate role of emotion in the identification and implementation of legal norms and in the deliberative process for juries, judges, and other legal actors.

Second, at the level of specific doctrine, law rests on myriad assumptions, both explicit and implicit, about how emotion influences behavior and about how to channel emotion to influence decision-making in desirable ways. Law and emotion scholarship identifies and scrutinizes these assumptions in light of evolving knowledge about the role of emotion in decision-making across a range of disciplines.

Finally, the field increasingly focuses on emotion not only as an internal and individual phenomenon but also as an essential component of social and institutional dynamics. It thus sheds light on collective decision-making, on the conditions that lead to legal change, and on the dynamics of institutional behavior and institutional reform.

Emergence of the Field

The emergence of modern law and emotions scholarship was fueled by several interrelated trends. Beginning in the 1980s and early 1990s, the topic of emotion began attracting renewed scholarly attention in fields like philosophy (Nussbaum, 1990), psychology (Clore et al., 1994), and sociology (Thoits, 1989). At the same time, cognitive scientists progressed in understanding emotion's role in judgment by studying patients with brain abnormalities or injuries that impair emotional functions (e.g., Damasio, 1994, 1999). More recently, the field of cognitive neuroscience began to employ powerful neuroimaging technology such as functional magnetic resonance imaging and positron emission tomography scans to investigate these dynamics (see, generally, Sinnott-Armstrong, 2008; Jones et al., 2009). The study of the role of emotion in cognitive processes has in recent years become a focal point for a number of disciplines, generating a complex, provisional and rapidly evolving body of knowledge. Although many disciplines hold pieces to the puzzle of emotion's role, there is 'no umbrella discipline to bridge the findings' of these disparate fields (Panksepp, 1998: p. 5).

Concurrent with the explosion of interest in emotion in the sciences and social sciences, legal scholarship grew more open to relevant insights from other disciplines. In addition, epistemological challenges to notions of judicial objectivity first raised by the legal realists were renewed and extended by critical race theorists and other scholars (Abrams and Keren, 2010: pp. 2005–2006). Feminist jurisprudence, in particular, was an important force behind emergent law and emotions scholarship, arguing that law tended to privilege a version of reason that exiled important qualities like empathy and compassion; qualities that were regarded as soft and feminine and therefore lacking in rigor (Minow and Spelman, 1988). The necessary next steps, as Bandes (1996, 2000) argued, were to move beyond the rehabilitation of certain emotions, clarify that no particular emotional stance is likely to be uniformly desirable or undesirable, and turn to more specific consideration of the value of emotions, whether positive or negative, in defined

contexts. Recently the scope of law and emotions scholarship has expanded considerably. It now covers the range of positive and negative emotions, incorporates findings from a wide range of disciplines, and explores multiple doctrinal areas as disparate as securities regulation, risk regulation, personal injury law (Madeira, 2006), foreclosure law (White, 2010), family law, contract law, and trademark law, to name a few.

Definitional Issues: What Is Emotion? What Are Emotions?

One continual challenge of this emerging field is defining its central term: emotion. Terms for particular emotions, such as remorse, disgust, or fear likewise elude fixed definitions, both across and within disciplines. As Jerome Kagan points out, for example, a neuroscientist seeking to measure fear responses in lab rats, an anthropologist studying fear responses to the approach of strangers, and a child psychologist studying separation anxiety all use the term 'fear,' but describe different agents, targets, contexts, methodologies, and research goals, and these need to be specified to facilitate interdisciplinary interchange (Kagan, 1998: pp. 14–38). A common error for jurists and legal scholars is to deploy such terms as if they have stable meanings not only within individual disciplines but when transported to the legal realm (Bandes, 2009b: pp. 8–13). Ideally, those who deploy the category 'emotion' or terms describing emotions in the legal realm will clarify their working definitions, the context in which the terms are being used (Kagan, 2007: p. 41), and what legal consequences flow from the use of the terms.

In the conventional legal view, emotions are unpredictable, untamable bursts of feeling that are too ephemeral to study and too instinctual to channel or educate. However, current research across several disciplines paints a very different picture. Although any definition must be provisional, emotions can be better described as a set of evaluative and motivational processes, distributed throughout the brain, that assist us in appraising and reacting to stimuli, and that are formed, interpreted, and communicated in social and cultural context. They influence the way we screen, categorize, and interpret information; influence our evaluations of the intentions or credibility of others; and help us decide what is important or valuable. They drive us to care about the outcome of our decision-making and motivate us to take action, or refrain from taking action, on the situations we evaluate. In short, they are pervasive influences on human behavior, and therefore essential for a full understanding of legal behavior and institutions.

Identifying and Investigating Law's Affective Dimensions

Scholars have investigated the role of emotion in a wide range of legal contexts. Maroney's (2006) law and emotion taxonomy provides a valuable organizing framework for scholars across the range of disciplines. Recent comprehensive overviews of the field by Abrams and Keren (2010) and Bandes and Blumenthal (2012) convey the range of the scholarship. Three main approaches to the study of law and emotion are discussed below.

The Role of Emotion in Doctrine

Legal doctrine sometimes rests on explicit assumptions about emotion. For example, in the criminal law, the practice of imposing shaming punishments is premised on the assumption that such punishments will encourage offenders to change their ways. In the law of evidence, the excited utterance exception to the hearsay requirement is premised on the assumption that a statement delivered in an excited state brought on by a startling event is more likely to be reliable. More often, law's assumptions about emotion are implicit. This makes them more difficult to identify, but no less likely to lead to the imposition of legal consequences. For example, Calhoun (2000) argues that objections to same-sex marriage rest upon implicit assumptions about who is capable of romantic love. Lynch and Haney (2011) have found that in capital sentencing proceedings, the jury's perception of whether the defendant displayed appropriate remorse is one of the main factors – and perhaps the most important factor – in determining whether the death penalty is imposed. More subtly, assumptions about what counts as emotion may influence legal doctrine in ways that go undetected. For example, appeals for mercy may be coded as emotional and off-limits, whereas appeals for vengeance or retribution may be coded as garden-variety legal arguments (Bandes, 2009a: pp. 497–498).

Much of the importance of the scholarship on law and emotion lies in exploring the ways in which the categories of reason and emotion are deployed and in "challenging the notion of a neutral, emotionless baseline" in legal theory and practice (Bandes, 1996: p. 370). Once the assumptions about emotional dynamics that pervade the law are identified, their accuracy can be evaluated in light of scientific and social-scientific knowledge. Whether a gap between law and science ought to lead to revising doctrine is a legal question, but one that ought to be made in light of the relevant available information about how emotions work.

The Role of Emotion for Various Legal Actors

There has long been available a robust empirical literature on jury decision-making (see e.g., Diamond et al., 2010), which has recently begun looking specifically at juror emotion (see e.g., Sundby, 2003) and at how the collective dynamics of the jury room differ from the emotions of individual jurors (see Lynch and Haney, 2011). It has more recently been joined by a literature on the role of emotion in negotiation (e.g., Ryan, 2005). Nevertheless, the question of what role emotion plays or ought to play in the reasoning process of particular legal actors or legal entities – for example jurors as a collective entity, legislators (see Sanger, 2012), or judges – is ripe for investigation. The emotions of judges have received scant attention until recently, when the emotional dynamics of judicial decision-making attracted substantial interest, prompted in part by controversy over the role of empathy in judging (West, 2012). Scholarship has focused both on how judges deliberate (see, e.g., Posner, 2008) and on how the emotional dynamics of judicial deliberation affect judges (see, e.g., Maroney, 2011b; discussing how judges do and should regulate their emotions).

The Role of Emotion in the Reasoning Process

Although it is widely agreed that emotion plays a role in the cognitive process, the question of precisely what that role entails is the subject of lively debate and substantial research across disciplines. It is clear that emotion influences not only 'the *content* of cognition' but also '*how* people think' (Forgas, 2001: p. xiv). They affect what stimuli draw attention, what meaning is attached to observations, and what details remain in memory (Phelps, 2012: p. 19). Feigenson identifies several paths through which emotion may influence judgment, including: (1) by affecting people's strategies for processing information; (2) by biasing the perception, recall or evaluation of facts in a particular direction; (3) as a source of informational cues to proper attribution of blame; and (4) through anticipation of future emotions that might follow from a judgment (Feigenson, 2010: pp. 46–47).

Emotion's Informational Role

Emotions may affect judgment directly, by providing informational cues about the judgment target. This path to emotion's influence on judgment is described in the literature as 'affect-as-information' (Clore et al., 1994). (The terms 'affect' and 'emotion' are often used interchangeably.) For example, a witness's testimony may elicit unease or mistrust; a crime may elicit anger toward the defendant and sympathy toward the victim; a judicial decision may elicit outrage; a victim's testimony may elicit sympathy, a public event may evoke fear or sympathy that triggers a push for legislation. At bottom, whether the information provided by these emotional reactions is relevant or useful is a legal question, dependent in part on the legal context and the goals of the proceeding (for example anger toward a defendant might be relevant to sentencing but not to guilt). More broadly, it is a jurisprudential question about what factors and values ought to influence legal judgment. Studies lend increasing support to the conclusion that emotion is integrally involved in moral judgment. The question of whether 'moral' emotions and capacities, such as moral indignation, outrage, disgust, compassion, and empathy, ought to influence legal norms and judgments (Kahan, 2000; Nussbaum, 2000) is intertwined with longstanding debates about the role of moral reasoning in legal judgment.

Emotion's effects on information processing

Studies have shown in several contexts that heightened emotion increases the likelihood of a criminal conviction, at least in mock juries. There is evidence that 'gruesome' photographic evidence increases mock jurors' negative emotional state, and usually leads to increased conviction rates. Mock jurors have been found to be unable to disregard emotionally charged evidence when instructed to, and the emotional content of such evidence has been tied to increased guilt and sentence ratings. Victim impact statements have been found to elicit anger at the defendant, and that anger has been linked to increased sentence severity (Nadler and Rose, 2003: pp. 24–26; Salerno and Bottoms, 2009: pp. 284–285). Feigenson and Park (2006) and Maroney (2006) summarize many other studies of emotion's impact on legal judgments. As Feigenson suggests, emotion may influence judgment in these contexts through several paths. For example, the emotions affecting the decision-maker may influence strategies for processing information. Sadness has been linked with careful information processing. Anger, on the other hand, has been found to exert a negative effect on judgment – leading to shallower information processing, greater reliance on heuristics and biases, and greater confidence in the accuracy of one's judgment. Alternatively, a decision-maker's emotions may bias the perception, recall, or evaluation of facts in a particular direction. They may lead the decision-maker to seek out emotion-congruent information, and to ignore or downplay incongruent information. For example, anger at a criminal defendant may lead jurors to focus on evidence that validates their negative reactions and to discount mitigating evidence. Much work remains to be done identifying the specific emotions at play in various contexts and their mediating effects on judgment.

Identifying and Interpreting Emotion

One of the central roles of legal decision-makers is to interpret and evaluate the intentions, beliefs, and motivations of others. The affective sciences and social sciences offer several promising approaches to studying the dynamics of communicating and interpreting emotion.

Nonverbal Expression of Emotion

The study of the communication of emotion through facial expression and body language is a burgeoning field. As noted above, significant legal consequences may turn on how facial expressions or body language are interpreted – for example on whether a defendant looks remorseful (Weisman, 2009). Researchers in this field focus on issues highly germane to law, including whether facial expression of emotion is culture-specific, and whether facial display conveys information about the expresser's social motives and intentions (Niedenthal et al., 2006: p. 116).

Cultural Expectations for the Feeling and Display of Emotion

A second approach to understanding the dynamics of communicating and interpreting emotion is the study of schemas and cultural expectations about how emotion ought to be expressed in various contexts (see, e.g., Hochschild (1983) on feeling rules and display rules). The credibility of a range of legal actors may be influenced by whether they display the emotions expected from those in their role (e.g., whether a rape complainant expressed the emotions the decision-maker expects from a rape victim (Schuster and Propen, 2011)).

Empathy

A third research area focusing on the communication and interpretation of emotion is the social neuroscience of empathy, which sheds light on how individuals understand the minds, desires, and motivations of others (for an excellent compendium, see Decety and Ickes, 2011). This area also has important implications for law.

The role of empathy in law has been much discussed recently, specifically in the judicial context, as mentioned above. The debate, in law as in other disciplines, has been bedeviled by

definitional ambiguities. Neuroscientific studies of empathy are helping to disambiguate the neural underpinnings of various types of empathy, some of which also go by such names as sympathy and compassion (Batson, 2009). These definitional ambiguities are relevant to legal and political debates over the role of empathy in judging. For example, some concerns about judicial use of empathy are based on the notion that empathy entails a sentimental attachment to – or a desire to offer assistance to – one of the parties to a controversy and is thus inconsistent with impartiality. This concern may be well-founded if empathy means an 'impulse to respond with sensitivity and care to the suffering of another' (Batson, 2009: p. 9). However, if empathy consists of understanding the thoughts and feelings of another (Batson, 2009: p. 4), then it is, arguably, an essential capacity for judges (Posner, 2008: p. 117; West, 2012).

Moreover, the dynamics of empathy are relevant to law in much broader contexts. If empathy is the capacity to understand the thoughts and feelings of another, it becomes clear that it is not an occasionally employed sentiment or a sentiment directed solely at those less fortunate, but a tool that is in constant (albeit not always accurate) use in legal decision-making. Thus it is important to investigate how empathy is employed in a variety of legal contexts. For example, studies suggest not only that juries have difficulty empathizing with defendants from other racial and demographic backgrounds (Lynch and Haney, 2011), but also that juries in homicide cases tend to feel more empathetic toward some homicide victims than others and that this selective empathy influences their verdicts (Sundby, 2003). The science and social science of empathy also point the way toward the possibility of addressing empathy deficits or selective empathy; empathy appears to be a capacity that is eminently trainable (Feshbach and Feshbach, 2011).

Implications for Implementation and Reform

As the foregoing discussion illustrates, the relationship between law and emotion is multifaceted. Therefore, the reforms to which insights from law and emotion scholarship point may take many forms. One approach to reform is revision of legal doctrine that rests on erroneous assumptions about emotion. An even broader challenge is to channel and regulate the emotions of legal actors, and to apply the insights of emotion theory to the structure and function of legal institutions. The following section briefly discusses several possible approaches to legal reform. It briefly discusses the revision of doctrine and the management of the emotions of legal actors. It then turns to the question of how legal institutions might be structured in light of insights about the role of emotion in legal and moral decision-making.

Doctrinal Revision

Where legal doctrine relies on erroneous assumptions about emotions or emotional dynamics, is ought to be reconsidered. To take a straightforward example, shaming sanctions are meant to deter and rehabilitate, and if it can be established that they have the opposite effect – that instead they stigmatize and marginalize offenders in a way that interferes with both these goals – then shaming sanctions serve no legitimate penal purpose and should no longer be imposed (Massaro, 2000). Whether a disconnect between legal standards and scientific or social-scientific knowledge requires a change in legal doctrine is ultimately a legal question, but one that should be informed by relevant scientific and social-scientific knowledge.

Managing Emotions

On the individual level, numerous studies show that people are able to regulate or manage the emotions they experience, whether by reconstructing aspects of a situation, shifting attention, reappraising reactions to a situation, or avoiding situations that might elicit a particular emotion (see Maroney, 2011a; reviewing the literature and applying it to the ability of judges to manage their emotions). Much of the primary emotion regulation and management research thus far has focused on the coping strategies and psychological health of individuals. Thus, an important avenue for law and emotion research will be to move this line of research into decision-making generally and legal decision-making in particular.

Educating and Channeling Emotions through Structural Reform

A fertile area for research is the question of whether, and under what circumstances, emotions can be educated. Both unconscious and conscious emotions influence legal decision-making, and either type of emotional influence may be desirable or deleterious, depending on the context. Indeed it increasingly appears that all decision-making has both unconscious, intuitive components and slower, more conscious components. Emotion-based capacities such as empathy, as discussed above, appear to be highly educable. There is some evidence that this educative capacity exists not only for emotions in conscious awareness, but even for intuitive, unconscious emotions, implicit biases, and stereotypes (Blumenthal, 2010: pp. 196–198). To construct legal institutions capable of channeling and educating both conscious and unconscious emotions, it is necessary to understand the complex interplay between the two. This requires examination of dual-process theories generally and in the context of moral decision-making.

Dual-Process Theories

Most current theories of cognition adopt a dual-process approach, involving some combination of quick, intuitive judgments and slower, more deliberative judgments (Kahneman, 2011). Much of the debate centers on how the two processes interact. The least nuanced dual-process models treat the two systems as quite separate, and treat the quick, intuitive process as inferior – a path to be taken only when there is insufficient time to deliberate properly. One pitfall to watch for in the dual-process model literature is slippage between the concepts of intuition and emotion. Scholarship in law, political science, and other disciplines, to some extent abetted by the neuroscientific literature itself, too often equates 'intuitive' with 'emotional' and 'deliberative' with 'reasonable,' replicating the problematic dichotomy between emotion and reason. Not all quick, intuitive judgments are emotional in nature, and emotion influences slow, deliberative judgments in addition to quick intuitive ones. Fortunately, dual-process

models are increasingly sophisticated, recognizing that the processes are highly interconnected and interdependent. Two of the most prominent dual-process theories are discussed below in the context of moral decision-making.

Moral Decision-Making

Greene and colleagues (e.g., Greene et al., 2001) suggest that some moral decisions "engage emotional processing to a greater extent than other(s), and these differences in emotional engagement affect people's judgments" (p. 2106). They argue that emotion automatically and unconsciously influences personal but not impersonal moral decisions. This theory has interesting implications for jury decision-making in general, for more specific issues like the use of videotaped testimony, and for the validity of mock jury studies, among other issues.

In Haidt's (2001) social-intuitionist theory, moral decision-making is intuitive, fast, unconscious, and automatic. According to this approach, stimuli lead to instant, affect laden feelings of moral approval or disapproval and these feelings cause moral judgments (p. 814). Haidt argues that moral intuitions are difficult to override, and that conscious reasoning is often merely a post-hoc way to explain our initial moral intuitions. Nevertheless he identifies several means of overriding or at least revising these strong intuitive responses. These include using conscious verbal reasoning and reframing a situation in a way that triggers a competing intuition. If Haidt is correct that moral judgment comes first, followed by a reasoning process that is designed to explain rather than revise the initial intuition, his model has, at minimum, important descriptive implications for our understanding of how judging and other deliberative processes occur.

Dual-process dynamics can help shed light on how intuitive emotions might be channeled or tempered by the legal system where appropriate. Along these lines, Sunstein (2009) has explored the emotion of moral indignation as a species of intuitive processing that provides valuable information to the legal system but that often needs to be tempered by more reflective reasoning, and has considered how various legal institutions and processes (including homicide trials, administrative and legislative risk regulation, and punitive damage assessments) ought to be structured to ensure that both intuition and deliberation play their appropriate roles. Moreover, Haidt (2001) and others argue that moral judgment flourishes not in isolation but in situations of social interchange, particularly with those who hold different perspectives. The implications of dual-process models for the structure of deliberative institutions like the jury and the judicial system are well worth exploring.

Emotion in Group and Institutional Settings

The conventional focus across disciplines on emotions as private and internal feelings has been enormously limiting, both in general and in the legal context. Just as emotions are shaped by social and institutional context, institutions are constructed in light of assumptions about emotional dynamics. In the legal realm, as elsewhere, these assumptions are ripe for scrutiny in light of a growing body of interdisciplinary knowledge about emotion in collective and institutional settings (Bandes, 2009b: p. 394).

Recently, a number of disciplines have begun to focus on emotion in social context. The sociology of emotion is now a burgeoning field (Turner and Stets, 2005). Psychologists are increasingly interested in emotion and social cognition, in emotion's role in group dynamics (Parkinson et al., 2005), and in the interaction between emotion and culture (Niedenthal et al., 2006), an interest they share with anthropologists (Lutz and White, 1986). Political scientists are exploring the role of emotion in democratic deliberation (Westen, 2007; Krause, 2008). The emerging field of affective neuroscience is examining the neural dynamics of emotional interchange (Panksepp, 1998). Criminologists are looking at the role of emotion in a variety of contexts, including the formation of punitive attitudes, the dynamics of violence, and the structure of criminal justice institutions such as police departments and prisons (Karstedt et al., 2011). Viewing emotions in their broader social and cultural context permits a fuller understanding of their role in the law and points the way toward a rich set of questions for future scholars in the field of law and emotion, several of which are discussed below.

Emotion and Social Norms

Emotions such as shame, fear, trust, and the desire for approval are intimately involved in the development, communication, and enforcement of the norms animating law. Legal scholars have engaged with the emotional aspects of norm-creation in a few discrete areas, notably on the topic of why and how we punish (see, e.g., Murphy, 2003; Dubber, 2006). But emotions are implicated in a wide array of basic social norm questions, such as the creation of familial norms, norms of respect for intellectual property, norms about the moral force of mortgage obligations (White, 2010), and questions of why citizens obey the law or feel themselves part of a polity (Deigh, 2000; Dubber, 2006; Gewirtzman, 2009).

Emotion Cultures

Law is permeated with implicit social and cultural expectations about the feeling and display of emotion (for discussion generally of emotion cultures, feeling norms, and display norms, see Hochschild, 1983; Thoits, 1989). These include, for example, expectations about who is capable of romantic love and therefore marriage (Calhoun, 2000) or about what rape victims or domestic violence victims ought to feel and how they ought to express their feelings (Schuster and Propen, 2011) or about what pregnant women ought to feel toward their unborn babies (Guthrie, 2008). The study of emotion cultures also encompasses the culture of particular legal institutions, such as the bench (Maroney, 2011a), the courtroom (Lynch and Haney, 2011) the law school classroom (Harris and Shultz, 1993), and the law firm (Levit and Linder, 2010).

Effects of Group Dynamics on the Emotions of Individuals

What effect does group participation have on the emotions of individual group members? For instance, does jury deliberation increase or decrease the emotional effect of various types of evidence on individual jurors? These questions are ripe for jury researchers.

The Unique Dynamics of Group-Level Emotion

Increasing empirical research is underway on group-level emotion, intergroup emotion, and emotional contagion (Parkinson et al., 2005: pp. 87–143). This research ought to be of enormous interest to jury researchers, as well as to those studying group decision-making in legal institutions more generally, and to those studying the effects of public opinion on legislation and other legal initiatives.

Emotion and Institutional Structure

Institutions are, inevitably, constructed in light of assumptions about emotional dynamics, and these assumptions should be illuminated and evaluated. Knowledge of emotional dynamics can be utilized to restructure incentive systems and construct more effective legal institutions. For example, Minow (2000) has examined the respective roles of divorce courts, mediation and arbitration tribunals, and truth and reconciliation commissions in channeling, educating, exacerbating, and responding to emotions. Ahmed and Braithwaite (2011) have investigated how institutional management of shame and pride affects bullying in the workplace.

Emotion and the Mobilization of Rights

Emotion theory can shed important light on how rights are conceived, how perceptions of fairness and unfairness arise, how political persuasion works (Westen, 2007; Krause, 2008), and what drives and sustains legal change and reform. Emotion can bind commitments to social and legal structures or provoke or sustain a challenge to those structures (Goodwin et al., 2001; Turner and Stets, 2005). Conversely, legal developments can create emotional momentum that leads to social change (see, e.g., Gould 2001, discussing the mobilizing effect of the US Supreme Court's decision in *Bowers v. Hardwick* (1986)).

Some Areas for Future Study

Underexplored Doctrines

These include First Amendment issues such as the regulation of offensive speech; the emotional dynamics animating the separation of powers doctrine; the role of emotional attachment to property in the measure of just compensation or in adverse possession doctrine; the role of trust in contract law, and assumptions about affective bonds undergirding the law of wills and trusts.

Underexplored Legal Actors

Studies of juries should move beyond the focus on individual jurors and take increasing account of the dynamics of the jury as a collective entity. The salutary trend toward studying judicial emotion should be expanded to include studies of how emotion affects decision-making by legislators, regulators, attorneys, and other legal actors.

Emotion and Persuasion

Emotion's role in rhetoric and persuasion is often dismissed as a form of manipulation or pandering, yet emotion plays an essential role in effective communication and persuasion. What role does, and should, emotion play in the persuasiveness of various forms of legal argument, including judicial opinions and dissents, the arguments of advocates, and jury deliberation?

Emotion and Legislation

In addition to studying legislators, scholars might focus on the relationship between emotion and legislation. For example, what is the appropriate role of emotion in providing an impetus for legislation? What is the role of legislation in acknowledging or giving voice to constituent emotion? (See, for example, Sanger's (2012) discussion of legislation requiring the provision of stillborn birth certificates.)

Strategies for Educating Emotion

In light of increasing evidence that institutions can be structured to promote a range of values and emotional attributes, debate about the structure and function of legal institutions should encompass the question of which values and attributes are worth promoting in particular contexts and how to do so.

Conclusion

The law offers an unparalleled opportunity for emotion researchers: a rich occasion to explore emotion as it operates in a complex set of institutions designed to reflect, channel, and educate human behavior. Legal theory and practice stand to benefit tremendously from this body of knowledge, and indeed, the legal system proceeds without this essential insight into human behavior at its peril, and to the detriment of those it serves. Legal doctrines and legal institutions too frequently rely on unexamined or demonstrably faulty assumptions about emotion and its effects on human behavior. Law and emotion scholarship has often focused on illuminating the role of emotion in law, which is an essential aspect of the project, but one that must be increasingly accompanied by interdisciplinary investigation, by debate about whether the disconnect between assumptions and evidence interferes with the administration of justice, and by concrete proposals for implementation and reform.

Acknowledgments

This article is adapted from Susan A. Bandes and Jeremy A. Blumenthal, 2012. Emotion and the law. *Annual Review of Law and Social Science* 8, 161–181. http://dx.doi.org/10.1146/annurev-lawsocsci-102811-173825.

See also: Critical Race Theory; Empirical Legal Studies; Judges; Juries; Law and Everyday Life; Law as an Instrument of Social Change; Law, Mobilization of; Law: History of Its Relation to the Social Sciences; Legal Culture and Legal Consciousness; Legal Personality; Punishment: Social and Legal Aspects; Science and Law.

Bibliography

Abrams, K.R., Keren, H., 2010. Who's afraid of law and the emotions? Minnesota Law Review 94, 1997–2074.

Ahmed, E., Braithwaite, J., 2011. Shame, pride and workplace bullying. In: Karstedt, S., Loader, I., Strang, H. (Eds.), Emotions, Crime and Justice. Oxford University Press, Oxford, pp. 55–79.

Bandes, S.A., 1996. Empathy, narrative, and victim impact statements. The University of Chicago Law Review 63, 361–412.

Bandes, S.A. (Ed.), 2000. The Passions of Law. NYU Press, New York.

Bandes, S.A., 2009a. Repellent crimes and rational deliberation: emotion and the death penalty. Vermont Law Review 33, 489–518.

Bandes, S.A., 2009b. Victims, "closure", and the sociology of emotion. Law and Contemporary Problems 72 (2), 1–26.

Bandes, S.A., Blumenthal, J.A., 2012. Emotion and the law. Annual Review of Law and Social Science 8, 161–181.

Batson, C.D., 2009. These things called empathy: eight related but distinct phenomena. In: Decety, J., Ickes, W. (Eds.), The Social Neuroscience of Empathy. MIT Press, Cambridge, pp. 3–15.

Blumenthal, J.A., 2010. A moody view of the law: looking back and looking ahead at law and the emotions. In: Bornstein, B., Wiener, R. (Eds.), Emotion and the Law: Psychological Perspectives. Springer, New York, pp. 185–210.

Bowers v. Hardwick, 478 U.S. 186 (1986).

Calhoun, C., 2000. Making up emotional people. In: Bandes, S.A. (Ed.), The Passions of Law. NYU Press, New York, pp. 217–240.

Clore, G.L., Schwarz, N., Conway, M., 1994. Cognitive causes and consequences of emotion. In: Wyer, R.S., Srull, T.K. (Eds.), Handbook of Social Cognition, second ed. Erlbaum, Hillsdale, NJ, pp. 323–417.

Damasio, A., 1994. Descartes' Error: Emotion, Reason, and the Human Brain. Penguin Putnam, New York.

Damasio, A., 1999. The Feeling of what Happens: Body and Emotion in the Making of Consciousness. Harcourt, Orlando.

Decety, J., Ickes, W. (Eds.), 2011. The Social Neuroscience of Empathy. MIT Press, Cambridge.

Deigh, J., 2000. Emotion and the authority of law: variation on themes in Bentham and Austin. In: Bandes, S.A. (Ed.), The Passions of Law. NYU Press, New York, pp. 285–308.

Diamond, S.S., Rose, M.R., Butler, K.M., 2010. Goffman on the jury: real jurors' attention to the "Offstage" of trials. Law and Human Behavior 34, 310–323.

Dubber, M.D., 2006. The Sense of Justice: Empathy in Law and Punishment. NYU Press, New York.

Feigenson, N., 2010. Emotional influences on judgments of legal blame: how they happen, whether they should, and what to do about it. In: Bornstein, B., Wiener, R. (Eds.), Emotion and the Law: Psychological Perspectives. Springer, New York, pp. 45–96.

Feigenson, N., Park, J., 2006. Emotions and attributions of legal responsibility and blame: a research review. Law and Human Behavior 30 (2), 143–161.

Feshbach, N.D., Feshbach, S., 2011. Empathy and education. In: Decety, J., Ickes, W. (Eds.), The Social Neuroscience of Empathy. MIT Press, , Cambridge, pp. 85–97.

Forgas, J.P. (Ed.), 2001. Feeling and Thinking: The Role of Affect in Social Cognition. Cambridge University Press, Cambridge.

Gewirtzman, D.N., 2009. Our founding feelings: emotion, commitment, and imagination in constitutional culture. University of Richmond Law Review 43, 623–683.

Goodwin, J., Jasper, J., Polletta, F. (Eds.), 2001. Passionate Politics: Emotions and Social Movements. University of Chicago Press, Chicago.

Gould, D., 2001. Rock the boat, don't rock the boat, baby: ambivalence and the emergence of militant AIDS activism. In: Goodwin, J., Jasper, J., Polletta, F. (Eds.), Passionate Politics: Emotions and Social Movements. University of Chicago Press, Chicago, pp. 135–157.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D., 2001. An fMRI investigation of emotional engagement in moral judgment. Science 293, 2105–2108.

Guthrie, C., 2008. Constitutional rights and the psychology of regret. Southern California Law Review 81, 877–903.

Haidt, J., 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological Review 108 (4), 814–834.

Harris, A., Shultz, M., 1993. A(nother) critique of pure reason: toward civic virtue in legal education. Stanford Law Review 45 (6), 1773–1806.

Hochschild, A.R., 1983. The Managed Heart: Commercialization of Human Feeling. University of California Press, Berkeley and Los Angeles.

Jones, O., Buckholtz, J., Schall, J., Marois, R., 2009. Brain imaging for legal thinkers: a guide for the perplexed. Stanford Technology Law Review 5.

Kagan, J., 1998. Three Seductive Ideas. Harvard University Press, Cambridge.

Kagan, J., 2007. What Is Emotion? History, Measures, and Meanings. Yale University Press, New Haven.

Kahan, D.M., 2000. The progressive appropriation of disgust. In: Bandes, S.A. (Ed.), The Passions of Law. NYU Press, New York, pp. 63–79.

Kahneman, D., 2011. Thinking Fast and Slow. Farrar, Straus and Giroux, New York.

Karstedt, S., Loader, I., Strang, H., 2011. Emotions, Crime and Justice. Oxford University Press, Oxford.

Krause, S., 2008. Civil Passions: Moral Sentiment and Democratic Deliberation. Princeton University Press, Princeton.

Levit, N., Linder, D.O., 2010. The Happy Lawyer: Making a Good Life in the Law. Oxford University Press, New York.

Lutz, C., White, G.M., 1986. The anthropology of emotions. Annual Review of Anthropology 15, 405–436.

Lynch, M., Haney, C., 2011. Mapping the racial bias of the white male capital juror: jury composition and the "empathic divide". Law & Society Review 45 (1), 69–102.

Madeira, J.L., 2006. Lashing reason to the mast: understanding judicial constraints on emotion in personal injury litigation. UC Davis Law Review 40, 137–197.

Maroney, T.A., 2006. Law and emotion: a proposed taxonomy of an emerging field. Law and Human Behavior 30 (2), 119–142.

Maroney, T.A., 2011a. Emotional regulation and judicial behavior. California Law Review 99, 1485–1555.

Maroney, T.A., 2011b. The persistent cultural script of judicial dispassion. California Law Review 99, 629–681.

Massaro, T., 2000. Show (some) emotions. In: Bandes, S.A. (Ed.), The Passions of Law. NYU Press, New York, pp. 80–120.

Minow, M.L., 2000. Institutions and emotions: redressing mass violence. In: Bandes, S.A. (Ed.), The Passions of Law. NYU Press, New York, pp. 265–284.

Minow, M.L., Spelman, E.V., 1988. Passion for justice. Cardozo Law Review 10, 37–76.

Murphy, J.G., 2003. Getting Even: Forgiveness and Its Limits. Oxford University Press, New York.

Nadler, J., Rose, M.R., 2003. Victim impact testimony and the psychology of punishment. Cornell Law Review 88, 419–456.

Niedenthal, P.M., Krauth-Gruber, S., Ric, F., 2006. Psychology of Emotion: Interpersonal, Experiential and Cognitive Approaches. Psychology Press, New York.

Nussbaum, M.C., 1990. Love's Knowledge: Essays on Philosophy and Literature. Oxford University Press, New York.

Nussbaum, M.C., 2000. "Secret sewers of vice": disgust, bodies and the law. In: Bandes, S.A. (Ed.), The Passions of Law. NYU Press, New York, pp. 17–62.

Panksepp, J., 1998. Affective Neuroscience: The Foundations of Human and Animal Emotions. Oxford University Press, New York.

Parkinson, B., Fischer, A.H., Manstead, A.S.R., 2005. Emotions in Social Relations: Cultural, Group, and Interpersonal Processes. Psychology Press, New York.

Phelps, E.A., 2012. Emotion's impact on memory. In: Nadel, L., Sinnott-Armstrong, W. (Eds.), Memory and Law. Oxford University Press, New York, pp. 8–26.

Posner, R.A., 2008. How Judges Think. Harvard University Press, Cambridge.

Ryan, E., 2005. The discourse beneath: emotional epistemology in legal deliberation and negotiation. Harvard Negotiation Law Review 10, 231–285.

Salerno, J.M., Bottoms, B.L., 2009. Emotional evidence and jurors' judgments: the promise of neuroscience for informing psychology and law. Behavioral Sciences & the Law 27 (2), 273–296.

Sanger, C., 2012. 'The birth of death': stillborn birth certificates and the problem for law. California Law Review 100, 269–312.

Schuster, M.L., Propen, A.D., 2011. Victim Advocacy in the Courtroom: Persuasive Practices in Domestic Violence and Child Protection Cases. Northeastern University Press, Boston.

Sinnott-Armstrong, W., 2008. Moral Psychology Volume 2: The Cognitive Science of Morality: Intuition and Diversity. MIT Press, Cambridge.

Sundby, S.E., 2003. The capital jury and empathy: the problem of worthy and unworthy victims. Cornell Law Review 88, 343–381.

Sunstein, C.R., 2009. Some effects of moral indignation on law. Vermont Law Review 33, 405–433.

Thoits, P., 1989. The sociology of emotions. Annual Review of Sociology 15, 317–342.

Turner, J.H., Stets, J.E., 2005. The Sociology of Emotions. Cambridge University Press, New York.

Weisman, R., 2009. Being and doing: the judicial use of remorse to construct character and community. Social & Legal Studies 18 (1), 47–69.

West, R., 2012. The empathetic turn. In: NOMOS LIII: Passions and Emotions, pp. 243–288.

Westen, D., 2007. The Political Brain: The Role of Emotion in Deciding the Fate of the Nation. Public Affairs, New York.

White, B.T., 2010. Underwater and not walking away: shame, fear, and the social management of the housing crisis. Wake Forest Law Review 45, 971–1023.

Law and Everyday Life

Patricia Ewick, Clark University, Worcester, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Historically, law and everyday life have been defined in opposition to one another. Over the past century, the boundary separating law from everyday life has become blurred. Increasingly, socio-legal scholars understand law and everyday life to be mutually constituted in a dynamic and emergent process wherein law is produced and applied in the everyday domain of social action. At the same time, the practices, relationships, and meanings that make up the everyday are themselves, in part, constructed by law or legality.

Law and Everyday Life

Law and everyday life have been defined historically in opposition to one another. Everyday life has typically denoted that sphere of social action characterized by being familiar, routine, pragmatic, and particular: the world of work, play, family, and leisure. By contrast, law has been conceptualized as formal, timeless, general, and transcendent. In the context of this opposition, law was understood to radiate downward toward everyday life to resolve conflicts and impart order. Everyday relationships, interactions, and practices were thought to constitute the objects upon which law operated with varying degrees of success. Alexis De Tocqueville articulated such a view of law and the everyday life over 200 years ago:

> The language of the law thus becomes, in some measure, a vulgar tongue; the spirit of the law, which is produced in the schools and courts of justice, gradually penetrates their walls into the bosom of society, where it descends to the lowest classes, so that at last the whole people contract the habit and tastes of the judicial magistrate. (Tocqueville, 1959: pp. 2–20)

Over the past century, the boundary separating law from everyday life has become increasingly blurred. With the blurring of this boundary, the presumed direction of effect from law ('the schools and courts of justice') downward to 'society' has likewise been challenged. Increasingly, socio-legal scholars understand law and everyday life to be mutually constituted in a dynamic and emergent process wherein law is produced and applied in the everyday domain of social action. At the same time, the practices, relationships, and meanings that make up the everyday life are themselves, at least in part, constructed by law or legality.

Legal Realism: The Everyday Life of the Law

In the early twentieth century, a perspective called legal realism rejected the prevailing formalist perspective on law (*see* Legal Realism; Legal Formalism). Whereas legal formalism saw the law as having an internal logic that led to determinant outcomes, legal realism posited that law is more than the rules and pronouncements contained in judicial decisions or statutes. Legal realists understood that legal rules were not simply and mechanically applied by judges, lawyers, police, and regulators. They proposed to account for the gap between 'law on the books' and 'law in action,' that is, between the law as envisioned by its makers and law as it is actually enforced on a daily basis.

Oliver Wendell Holmes (1897), a distinguished jurist and philosopher of law, postulated that the life of the law is experience, not logic. The realist view of law as action and experience assumes that law has a life outside of legal texts and that the contours of law's life are shaped by the interests, and limited by the vision, of its agents. By studying the impediments to full enforcement of law, legal realists implicitly rejected de Tocqueville's assertion that the 'habits and tastes' of the magistrate inevitably penetrate into the bosom of society. In radiating outward, the law encounters myriad obstacles to its full and intended enforcement. The obstacles might include the bounded rationality, political, or economic interests, or even the biographical idiosyncrasies of legal agents. In short, in entering the realm of lived experience, legal rules are articulated within particular social and historical contexts that shape law's effects. This proposition authorized an examination of those contexts and constraints, thus creating an opening to the everyday in the study of law.

Law First Perspective

Legal realism – even while it acknowledged the life, and not merely the spirit of the law – assiduously maintained the boundary between the law and everyday life, retaining what has been called a 'law first' orientation (Sarat and Kearns, 1993). Even Holmes – after declaring the logic of law to be a myth – closes his essay 'The Path of the Law' (1897) with an invocation of law's universality and distance from everyday life: "The remoter and more general aspects of the law are those which give it universal interest. It is through them that you…connect your subject with the universe, and catch an echo of the whole, a glimpse of the unfathomable process, a hint of the universal law" (pp. 405, 406). The law may make its way in the world, but it still inhabits the lofty reaches of the universe. Moreover, the view that the law is distinct from the everyday underwrote the legal realists' conviction that the law could be used instrumentally to bring about progressive social reform.

Karl Llewellyn and E. Adamson Hoebel in their classic book *The Cheyenne Way* (1941) proposed that law performed a dual function in ordering everyday life. Frequently, the law's prescriptions operated beneath the level of notice or consciousness "to make [people] go round in more or less clear ways." In this regulative function, the law created routines and norms that structured people's practices and expectations. At times, however, the so-called drift of everyday life produced divergent interests, expectations, and norms that upset the orderliness of social life. Llewellyn and Hoebel referred to these instances as 'trouble cases' and proposed that a central function of the law is to process these cases to restore social order. 'Law,' they wrote, "exists for the event of breach of law and has a major portion of its essence in the doing of something about such a breach" (Llewellyn and Hoebel 1941: p. 20). How law responds to such breaches – how it selects cases of trouble, in what way it redresses these situations and with what consequences – ultimately comes to comprise the law. Accordingly, Llewellyn and Hoebel recommended empirical examination of the selection, processing, and resolution of 'trouble cases' as a way of elucidating and understanding the law. Research in this tradition focused on the everyday work of legal professionals, examining such processes as judicial decision making, prosecutorial discretion, or police arrest practice within specific organizational and political contexts.

When the everyday world is examined within this tradition, it is only at the boundary that separates law from society, in those settings where citizens brought their disputes, claims, and grievances to the attention of legal professionals. Social scientific research examined, for instance, the sorts of claims and awards made in small claims courts (Yngvesson and Hennessey, 1975), the advice given in lawyer's offices (Sarat and Felstiner, 1995), or police arrest discretion (Black, 1971). In short, the everyday was included in these analyses of law, only in those instances when its familiar routines broke down to the point where the law was mobilized to restore order, redress an injury, or protect a right.

In *The Common Place of Law* (1998), Ewick and Silbey point out the problem with the 'law first' approach.

> The most serious limitation of this institutionally centered law-first perspective is that, by reckoning the boundary of law to correspond neatly to its formal institutional boundary, we drastically narrow our vision. By focusing our scholarly gaze on exchanges across that boundary, we fail to ask how people go to that hypothesized boundary. We exclude from observation that which needs yet to be explained: how, where, and with what effect law is produced in and through common place social interactions within neighborhoods, workplaces, families, schools, community organizations and the like. (1998: p. 20)

Everyday Life versus the Law

This 'law first' perspective shifted somewhat as scholars began to consider normative systems that function as an alternative or precursor to law – such as custom or religion. Whereas legal realists depicted the law as intervening in social life to create and restore order, Paul Bohannon (1973) presented law as emerging out of customary practices. He claimed that law resulted from a double institutionalization of such practices. This process of formalization granted to custom a force and legitimacy that eventually led to law having a life and dynamic of its own. Hence, according to Bohannon, while law attains an existence apart from the everyday, its source lies within the social. Its relative independence did more than simply reflect custom; as a result of the restatement of the customary in the process of formalization, the law is 'always out of phase with society' (Bohannon 1973: p. 311). Bohannon did not see the gap between the legal and the social as a problem, but rather as a fertile source of change and growth: "It is the fertile dilemma of law that it must always be out of step with society; but that people always… attempt to reduce the lack of phase. Custom must either grow to fit the law or it must actively reject it; law must grow to fit the custom, or it must suppress it" (1973: p. 311). Thus law and everyday are located in different 'realms,' yet they are inextricably linked to one another.

The 'law first' perspective was also displaced in research that examined citizens' decisions to turn to the law. Studying the daily enforcement practices of legal agents revealed the extent to which most western legal systems are reactive, that is, rely on citizens to mobilize the law. Obviously, this is true in the case of civil law where citizens seek legal remedies for private disputes. It is also true in regard to criminal law, however. In a classic study of police arrest decisions, for instance, Donald Black (1971) found that approximately 87% of all police–citizen encounters were initiated by citizens either in the role of victims or witnesses. In mobilizing both civil and criminal law, citizens are thus responsible for shaping the number and type of cases that the law processes. Recognition of their crucial role in determining legal caseloads, shifted research toward the question of how citizens interpret situations and make the decision to turn to law.

In this tradition, Felstiner et al. (1980–1981) developed a model of the interpretive process that leads up to a citizen's decision to seek legal remedies. To begin with, problems and situations (potential 'trouble cases') are not self-evidently injurious or legal. Therefore, Felstiner et al. refer to the first stage in this process as 'naming,' or defining a situation or event as injurious. 'Blaming' refers to the identification of a party who is responsible for the problem. Finally, 'claiming' entails bringing the grievance or complaint to a third party (which may or may not be a legal body) to obtain remedy. Although the model postulates an individual decision maker, it stimulated interest in the cultural meanings and values that might compete with interpretations favoring legal action at any point in the process. It also suggested that the relationship between law and everyday life was uneven and variable. Certain kinds of issues – for instance, environmental harm – are resistant to being transformed into legal disputes; while others – automobile accidents – yield a much greater number of legal claims. In considering these matters, socio-legal research extended its analyses to the communities and cultures that exist beyond the legal arena and led to a greater appreciation of how everyday life becomes illegible or uninterpretable within the grid of law and legal categories.

Researchers in the law versus everyday life tradition were also particularly interested in those instances where the law interacted with alternative normative systems. Legal anthropologist Sally Falk Moore (1972) examined the relationship between law and custom in the context of delimited structured everyday interactions, or what she referred to as social fields. She defined social fields as patterns of relationship and interaction that are ongoing and which generate rules that coerce or induce compliance. While such fields are defined by their capacity to generate norms and customs, their boundaries are porous and thus they are susceptible to external norms, including the law. Because of this openness, Moore characterized social fields as 'semiautonomous.' Actors within these fields are constrained by legal norms but not determined by them. Like Bohannon, Moore further blurred the boundary between the two spheres by highlighting how they might mutually reinforce or undermine one another. Drawing from an analysis of the garment industry, Moore noted that union bosses routinely refrained from enforcing the law when the workload was heavy and overtime was needed. In exchange, the garment contractors would give gifts such as liquor and dresses. Something that looked like friendships emerged over time and became essential for the smooth functioning of the industry. Moore points out, however, that these informal exchange norms did not supersede the law. Even in relaxing labor laws that stipulated hours of work, the law was operating, although not as its framers intended. Even if the labor boss did not enforce the law, everyone understood that he could. The law empowered him; and the possibility of his exercising legal rules explained the pattern of gifts and fictive friendships. Moore thus summed up the ongoing interpenetration of law and custom: "A court or legislature can make custom law. A semi-autonomous field can make law its custom" (Moore, 1972: p. 79).

A similar relationship was described by Stewart Macaulay in a classic article entitled 'Non-contractual Relations in Business: A Preliminary Study' (1963). Macaulay found that law suits for breach of contract are relatively rare among businessmen. The paramount reason for this included the availability of effective nonlegal sanctions that operated within the business community, such as gossip, giving or withholding of advice, or expertise. But more importantly, the businessmen's reluctance to use the law in lieu of these alternatives was due to the mutual desirability of maintaining ongoing relationships and the sense that invoking formal law would interfere with that goal. Robert Ellickson found that neighboring cattle ranchers in California rarely invoked the law in disputes involving trespass. Like Macaulay's businessmen and Moore's garment contractors, the ranchers, according to Ellickson, do not want, or necessarily need, the law to achieve order.

Others examined the cultural and symbolic implications of resorting to law in the context of alternative normative realms. In her ethnography of a small Baptist community in Georgia, Carol Greenhouse found that the religious commitments of the members of this community tended to suppress disputes and steer members away from secular legal processes in the event that disputes did occur (1986). Their outright rejection of law, she wrote, "…is rooted in the logic of sacredness that obviates not only applications of human authority, but human authority itself" (p. 206). Robert Kidder (1990) studied other 'law-avoiding' societies – including the Amish and the Japanese – and similarly discovered, in each case, an incompatibility between the community's ideological or spiritual commitments and law.

The research on law avoidance thus served to correct the sense that law was paramount in ordering relationships in modern societies. By proposing that law is only one of variously competing normative orders available to citizens for achieving orderly relationships, this research further underscored the need for a serious consideration of everyday lives, relationships, and beliefs of citizens.

Although this research abandoned an exclusive or primary focus on formal law – considering informal community norms and practices as well – it continued to conceive of law and everyday life as distinct and relatively autonomous from one another. Indeed, by focusing on law avoidance, the relationship between the formal law and the informal norms of everyday life was depicted as not merely distinct but oppositional. Subsequent research and theory has challenged the idea that law and everyday life are autonomous or oppositional, suggesting instead that the two exist in a dynamic and mutually constitutive relationship.

Law in Everyday Life; Everyday Life in Law

In the last two decades of the twentieth century, socio-legal scholars began to examine everyday life not simply as an object of legal action or as a reservoir of potential legal cases but as a domain in which law circulates, is appropriated, invoked, resisted, and produced. This view of the law rejects the notion of law as a mere instrument to be used or avoided by citizens for achieving social or private ends. Law and the everyday are seen as mutually constitutive, that is, literally 'making up' each other in an ongoing dynamic process.

Barbara Yngvesson (1989) has argued that law is invented by citizens in local struggles over the meaning of fundamental cultural symbols such as 'community,' 'neighborhood,' or 'rights.' As citizens act on their understanding, law "emerges in the exchanges and the struggles of everyday people, in neighborhood fights and lovers' quarrels, in the framing of claims and complaints to the police, the court, neighborhood justice, and other agencies, and in the response to these agencies to such claims." (p. 1709)

Citizens do not, in other words, simply mobilize the police or go to court with their preconstituted legal cases. Before the formal legal apparatus has been invoked, the cases have been fashioned out of the popular legal culture circulating in everyday life. The results of that fashioning come to constitute what cases are processed, decided, and contested in formal legal settings. Rather than imagining law as existing apart from social relations, or conceiving of it as produced solely by groups of powerful lawmakers (such as legislatures), in this constitutive view, law is understood as emerging from the 'bottom up' as a continuing product of practical reason and action.

Of course, the shape and content of everyday life is fundamentally dependent upon law as well. Even when we act in

situations that seem devoid of legal meaning or involve a self-conscious avoidance of law, we do not escape the constitutive power of law. The law creates roles, relationships, and obligations. It 'names the world.' To routinely enact the roles of wife, doctor, or consumer, or any social role, tacitly implicates the law. In fact, the appearance of the routine, the familiar, and the quotidian is in no small measure a consequence of those unspoken legal meanings.

This view of law and everyday life would take issue with the conclusions of research, such as Macaulay's (1963) and Ellickson's (1991). In demonstrating the absence of law in the ongoing relationship between businessmen or cattle ranchers, this research fails to note that the very roles and relationships in which their subjects act are legal creations. Moreover, the concepts of private property or profit, for instance, draw upon legally produced and sanctioned meanings. As Sarat and Kearns (1993) have written, "one is apt to catch more than a fleeting glimpse of law's presence in the putatively alegal responses" (p. 47).

Recognizing the pervasive presence of law outside of formal institutional settings, Ewick and Silbey (1998) used the concept of legality to augment the concept of formal law. By legality, they refer to the meanings, sources of authority, and cultural practices that are commonly recognized as legal, regardless of who employs them or for what ends. In this rendering, people may invoke and enact legality in ways neither approved nor acknowledged by the law. Legality operates as both an interpretative framework and a set of resources with which and through which the social world (including that part known as the law) is constituted. In sum, they conceive legality as an emergent structure of social life that manifests itself in diverse places, including both formal institutional settings and the everyday lives of groups and communities.

Studying legality wherever it manifests itself has led to the broadening of subjects and sites of research. One example of research that has systematically explored the presence of law in everyday life is Hendrik Hartog's analysis (Hartog, 1993) of the eighteenth century diary of Abigail Bailey. Hartog's reading of Abigail Bailey's diary reveals how this woman, over the course of many years, struggled to make sense of her marriage, her husband's sexual abuse of their daughter, and their separation and eventual divorce, as well as her own religious beliefs regarding her duties as a wife and mother. Hartog demonstrates that although law is conspicuously absent from the diary, this narrative of personal tragedy and change is incomprehensible without reference to legal categories such as the prevailing law of coverture (a woman's loss of legal rights or personality upon marriage). Abigail Bailey's perception and assessment of her situation and her daughter's experiences, were conditioned upon her understanding of the legitimacy of a husband's desires and the priority of his rights. Because the law established a husband as a virtual sovereign within his family, it was difficult to question or oppose openly her husband's actions as inappropriate.

The recognition that law and everyday life are mutually constitutive has led to an interest in legal consciousness, ideology, and resistance (*see* Legal Culture and Legal Consciousness). In fact, some have argued that the fact that our glimpse of law in everyday life is 'fleeting' gets to the heart of law's power and to the power that law confers. John Brigham observes that "Law is hard to see unless it wants to be seen" (2009: p. 382). The reasons for this are twofold: First, professional legal actors benefit from a vision of law as esoteric and remote from daily life. More importantly, when law constitutes the contours of life without being implicated in doing so, it naturalizes the power and privilege it produces. According to Antonio Gramsci (1971) when law 'disappears' itself from daily life, its power becomes hegemonic. The key to the hegemonic power of law is the maintenance of what Gramsci called 'contradictory consciousness,' an experience of the world in which an unbridgeable chasm exists between the experience of everyday life in which people encounter the conditions of work and lived social relations and the official accounts of how things work. Hegemonic power operates not by top–down social control or the imposition of a single vision of the world, but by disqualifying the significance of everyday experience through reification, wherein historically specific processes and behaviors are 'misrecognized' as having a temporal and spatial transcendence.

The study of the hegemonic power of law has led to an interest in how the law is embedded in citizens' consciousness, how they imagine and 'enact' law. This would include not only how and when citizens mobilize formal law, but also how they may invoke, interpret, avoid, and deploy law and legal concepts in daily interactions even when no formal or official actors or rules are involved. Socio-legal scholars who have studied the legal consciousness of ordinary citizens have found evidence of such contradictory consciousness in which the internal connection between legality and particular experiences is severed in ways that insulate legal institutions from critique. The reified version of the world becomes the standard of objective reality against which the subjective, the concrete, and the immediate (the 'here and now-ness' of everyday life) are assessed and interpreted. Ewick and Silbey (1998), for instance, found that American citizens often locate legality in the legal rules and regulations that seem to produce effects independent of human action. One man they interviewed, explained why he failed to lodge a complaint of discrimination after having been paid less than others. He explained that his employer used 'a grading system' to determine salaries and there was nothing he could do. His injustice claim was thus thwarted by an impersonal metric that seemingly could not be challenged.

Yet, hegemony is never total. It is, rather, a "process of continuous creation which, given its massive scale, is bound to be uneven in the degree of legitimacy it commands and to leave some room for antagonistic cultural expressions to develop" (Adamson, 1983: p. 174). Indeed, within the contradictions of consciousness lie the possibilities for challenge and resistance.

For example, David Engel's study of the Education for All Handicapped Children Act of 1975 (Engel, 1993) rejects a traditional analysis that would have traced the ways in which this legislation unilaterally constrained or defined the educational experiences of children, the meaning of their physical conditions, or the interactions that occurred between parents and educators. Instead, he examines how this legislation and everyday life within this community are

interdependent and mutually shaped this social domain. At times, Engel notes, the law reinscribed 'commonsense' images of disability and difference, constituting the meaning and content of those social categories as they operated within the community. Even though the intended effect of the legislation was to challenge the community's perception of these students as different, – by mandating educational inclusion – the law's challenge to common prejudices was only partial. Engel observes, for instance, that in order to receive the benefits of educational inclusion guaranteed under this legislation, the law required an initial determination by educational experts that a child was different, in other words disabled or abnormal, and thus subject to the law's protection. He also shows that the operation of the law in marking differences did not itself go unchallenged. Parents resisted this requirement, with varying degrees of success, by rejecting the holistic term 'disabled' and insisting on more specific and limited descriptions of each child's condition (such as 'speech impairment' or 'cerebral palsy'). This sort of resistance was possible because the legislation did not simply reinstate existing authority relations. The statute, by prescribing parental involvement, made such challenges possible. The legal rights of handicapped children were not simply bestowed from above or asserted from below, but emerged as a product of a complex array of social relations – part legislation, part community conceptions of difference, part professional authority, and part parental desires for their children's inclusion.

Research such as Engel's illustrates that law's presence in everyday life, while a source of its hegemonic power, also exposes law to forms of resistance, subversion, and alteration. While much of this resistance is tactical and opportunistic, remaining as fleeting and ephemeral as the power it opposes, it may also take the form of an organized challenge. In their analysis of the development of the alternative birth movement, Beckett and Hoffman (2005) argue that the movement was paradoxically stimulated by the efforts of organized medicine to mobilize law to prevent the licensure of midwives. Having entered the discursive field, law – in particular that law that 'privileged' organized medicine – became available as a resource for challengers. "The fact that many legislators explained their decision to vote in favor of midwifery licensure in terms of the very categories that are thought to privilege medicine indicates that hegemony is indeed contestable and transformable. Cultural meanings, images, and associations are not fixed, but (somewhat) fluid; this fluidity is an important resource for movement activists resisting sociolegal hegemony" (p. 164).

Recognizing the fallibility and vulnerability of law that circulates in everyday life, Kitty Calavita (2001) claims that law can have deconstitutive effects, that is, it can undermine its own ideological categories. Analyzing the public rejection and ridicule of an Italian Supreme Court decision that overturned a rape conviction on the grounds that the alleged victim was wearing blue jeans (which the court assumed could not be taken off without cooperation by their wearer), she argues that law exposed its own 'ideological blunders' and the outmoded moral vision it endorses. She concludes that such deconstitutive events are more likely to occur during 'unsettled cultural periods' during which there are explicit challenges to entrenched ways of thinking and doing (Swidler, 1986: p. 273).

The fate of law and everyday life in such unsettled periods is the question examined by David Engel and Jaruwan S. Engel in *Tort, Custom and Karma: Globalization and Legal Consciousness in Thailand* (2010). They begin by asking "What happens to law in the lives of ordinary people when a society undergoes rapid change, economic development, and integration into global markets?"(p. 3). Based on in-depth interviews with 35 people from Chiangmai who had been injured as a result of the conduct of another person, the authors trace the transformation of legal culture and consciousness in contemporary Thailand as a result of globalization. They depict a situation in which both state and customary law are increasingly irrelevant and, in the case of state law, an illegitimate way of ordering social relations.

By the 1930s, Thailand had adopted a legal system that conformed to the model of a liberal legal state. Recognizing individual subjects, valuing neutrality in judging, and adopting written laws and opinions, this state law operated evenly across delimited spatial and temporal boundaries. Thailand also had a customary legality which was, by contrast, defined around the idea of sacred centers – located in the heart of family, village, and community – whose authority diminished as it radiated outward. This customary legality was animated by ideas of karma, spirits, fate, and collective harm, and it was based on these concepts that fault and remedy were defined. Despite their differences, state and customary legality were mutually reinforcing. Operating as a court of last resort to enforce decisions made at the local level, state law institutionalized, and thus reinforced, customary remedies. At the same time, state law derived its own legitimacy from its connections to customary legal forms.

This interlegality thus sustained the multiple legal orders. At the same time, as Engel and Engel's research demonstrates, such interlegality also exposed both legal orderings to a shared vulnerability. The rapid social changes in the late twentieth century meant that people work and travel far from their villages and thus from sacred centers and customary authority. Injuries are no longer interpreted within the discourse of malevolent spirits posing risks of collective harm and thus warranting collectively enforced remedies. Instead, injuries are considered karmic in origin and thus the responsibility of the injured person. Moreover, within this cosmology, to pursue legal compensation through state law is thought to be a delusion and a form of attachment that will only lead to further suffering. Thus the displacements of globalization have led to the individuation of injuries and a withering of both state and customary law, and with them the possibility of obtaining justice through legal means.

Conclusion

Socio-legal research on law and everyday life began by conceiving of the two spheres as relatively autonomous. Research in this tradition asserted the primacy of law in generating norms and stressed its regulative function in processing 'trouble cases' and restoring order to the everyday world. By the middle of the twentieth century, however, scholars began to question the distinctiveness of law and the everyday. Law not only performs a restorative function when

norms are violated, its categories and discourses constituted the order that is occasionally disrupted by trouble. People's identities and relationships are crafted out of legal ideas; the material world – including bodies – is a product of law. Thus, law was seen operating in the routine and the familiar; the personal and the intimate.

Eventually, scholars came to recognize that this constitutive relationship was mutual: as law circulated throughout social life, it too, was shaped by these encounters 'on the ground.' Persons and groups could reject, avoid, appropriate, translate, or simply ignore the law. In doing so, the content, form, and reach of law itself was produced. By exposing law's presence in everyday interactions, these insights generated considerable interest in the ways that collective actors could transform law (McCann, 1994) and how even relatively powerless persons could resist the force of law through their everyday acts of resistance (Sarat, 1988; Ewick and Silbey, 1998).

Most recently, scholars have noted the effects of globalization and neo-liberal governance on both the meaning of 'everyday life' and of 'law.' Globalization connotes a particular configuration of nation-state and capital such that the source and reach of power are extended to overcome both time and space. This means, among other things, that power can be exercised remotely and in such a way as to exceed the contours of ordinary life. Similarly, globalization challenges the idea of sovereignty and with it traditional notions of law. In many ways, globalization has thus disrupted not only the relationship between law and the everyday, it has destabilized the categories themselves. These developments suggest that it is not sufficient to simply state that there is a constitutive relationship between law and everyday life, without specifying the nature of the relationship or acknowledging that it can wither or be curtailed. As Engel and Engel observe, in an age of neo-liberal governance and globalization, "[T]he inactivity of law in the lives and consciousness of ordinary people may be its most important feature…" (2010: p. 161). What had been definitively fixed (initially, that law and the everyday are separate; subsequently, that they are mutually constitutive) is now seen to be empirically contingent. Although the lives of ordinary citizens may be suffused with law and legal accountability, other actors – such as transnational corporations – can move across and beyond legal jurisdictions, effectively floating free of both law and the everyday world on the ground. Thus, while the law may be 'all over' for some actors (Sarat, 1988) it is not evenly so, nor is it uniform across time and space. New research in the relationship between law and everyday life must, then, account for the absence as well as the presence of law in a globalized world.

See also: Conflict and Conflict Resolution, Social Psychology of; Dispute Resolutions in Economics; Disputes, Social Construction and Transformation of; Empirical Legal Studies; Everyday Life, Anthropology of; Law and Society: Development of the Field; Legal Culture and Legal Consciousness; Legal Formalism; Legal Realism; Legal Systems: Private.

Bibliography

Adamson, W., 1983. Hegemony and Revolution: A Study of Antonio Gramsci's Political and Cultural Theory. University of California Press, Berkeley, CA.

Beckett, K., Hoffman, B., 2005. Challenging medicine: law, resistance and the cultural politics of childbirth. Law & Society Review 39 (1), 125–168.

Black, D., 1971. The social organization of arrest. Stanford Law Review 23, 1087–1111.

Bohannon, P., 1973. The differing realms of the law. In: Black, D., Mileski, M. (Eds.), The Social Organization of Law. Seminar Press, pp. 306–317.

Brigham, J., 2009. Seeing jurisdiction: some jurisprudential issues arising from law being "…all over". Law and Policy 31, 381–403.

Calavita, K., 2001. Blue Jeans, Rape and the 'De-Constitutive' Power of Law. Law & Society Review 35 (1), 89–116.

Ellickson, R., 1991. Order without Law: How Neighbors Settle Disputes. Harvard University Press, Cambridge, MA.

Engel, D., 1993. Origin myths: narratives of authority, resistance, disability and law. Law & Society Review 27 (4), 785–826.

Engel, D., Engel, J., 2010. Tort, Custom and Karma: Globalization and Legal Consciousness in Thailand. Stanford University Press, Stanford, CA.

Ewick, P., Silbey, S., 1998. The Common Place of Law. University of Chicago Press, Chicago, IL.

Felstiner, W.L.F., Abel, R., Sarat, A., 1980–1981. The emergence and transformation of disputes: naming, blaming and claiming. Law & Society Review 15 (3–4), 631–654.

Gramsci, A., 1971. Selections from the Prison Notebooks. Lawrence and Wishart, London.

Greenhouse, C., 1986. Praying for Justice: Faith, Order, and Community in an American Town Cornell. University Press, Ithaca, NY.

Hartog, D., 1993. Abigail Bailey's coverture: law in a married woman's consciousness. In: Sarat, K. (Eds.), Law and Everyday Life. University of Michigan Press, Ann Arbor, MI.

Holmes, O.W., 1897. The path of law. Harvard Law Review 10, 457–478.

Kidder, R., 1990. Managing ideologies: harmony as ideology in Amish and Japanese societies. Law & Society Review 24 (4), 895–922.

Llewellyn, K., Hoebel, E.A., 1941. The Cheyenne Way: Conflict in Case Law in Primitive Jurisprudence. University of Oklahoma Press, Norman, OK.

Macaulay, S., 1963. Non-contractual relations in business: a preliminary study. American Sociological Review 28, 55.

McCann, M.W., 1994. Rights at Work: Pay Equity Reform and the Politics of Legal Mobilization. University of Chicago Press, Chicago, IL.

Moore, S.M., 1972. Law and social change: the semi-autonomous field as an appropriate subject of study. Law and Society Review 7, 719–746.

Sarat, A., 1988. 'The law is all over': power, resistance, and the legal consciuosness of the welfare poor. Yale Journal of Law and Humanities 2, 343–380.

Sarat, A., Felstiner, W.L.F., 1995. Divorce Layers and Their Clients: Power and Meaning in the Legal Process. Oxford University Press, New York.

Sarat, A., Kearns, T.R. (Eds.), 1993. Law in Everyday Life. University of Michigan Press, Ann Arbor, MI.

Swidler, A., 1986. Culture in action: symbols and strategies. American Sociological Review 51, 273–286.

Tocqueville, A., 1959. Democracy in America. Vintage, New York.

Yngvesson, B., 1989. Inventing law in local settings: rethinking popular legal culture. Yale Law Journal 98, 1691–1709.

Yngvesson, B., Hennessey, P., 1975. Small claims, complex disputes: a review of the small claims literature. Law & Society Review 9 (2), 219–274.

Law and Geography

Nicholas Blomley and Joshua Labove, Simon Fraser University, Burnaby, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article replaces the previous edition article by N. Blomley, volume 12, pp. 8481–8485, © 2001, Elsevier Ltd.

Abstract

The article will overview changing theorizations of the relationships between law and geography, beginning with the 'regional' accounts of the early years of the twentieth century, where laws were seen as structured by geographic and environmental conditions. These are compared with the 'impact analyses' of mid-century, where law was seen as structuring space, and with a contemporary legal geography, where the relationship between space and law is seen as mutually constitutive, political, and socially consequential. Substantive empirical foci will include the regulation of public space and international immigration and border enforcement. It is suggested that this scholarship offers useful insights into a broader array of questions relating to law, social relations, power, and social space.

'Law and geography' should be defined as the analysis of the manner in which law, as a set of processes, texts, and practices, is shaped by the geographic dimensions of social and political life, and the ways in which the geography of social life is in turn structured by law. Defined thus, the systematic study of the relationship between law and geography has a long, although uneven, pedigree. Over the last century, scholars have defined both 'law' and 'space,' and the relationship between the two, in opposed ways, sequentially foregrounded space and then law. The first decade of the twenty-first century, however, has seen considerable interest in transcending the binary divide between law and space.

The article will overview changing theorizations of the relationships between law and geography, beginning with the 'regional' accounts of the early years of the twentieth century, where laws were seen as structured by geographic and environmental conditions. These are compared with the 'impact analyses' of the mid-twentieth century, where law was seen as structuring space, to contemporary debates in a self-defined legal geography, where the relationship between space and law is seen as mutually constitutive, political, and socially consequential. This scholarship offers useful insights into a broader array of questions relating to law, social relations, power, and social space.

Regional Studies of Law

The earliest 'legal geographies' were concerned with describing and explaining the regional diversity of legal systems. Eminent eighteenth-century scholars such as Montesquieu (1748/1949) could thus be classified as early 'legal geographers.' Regional studies of law, however, were to come into their own in the early years of the twentieth century, with the development of comparative law and the sociology of law. The American comparative legal scholar John Wigmore, in particular, undertook an extensive three-volume survey entitled *A Panorama of the World's Legal Systems* (1928). With Kocourek in 1918, he coedited a collection on legal evolution that included papers written by lawyers, sociologists, and geographers, which perhaps is the definitive summary of the regional field (Kocourek and Wigmore, 1918).

The central claim among these authors appears to be that spatial variations in law are a function of local environmental factors. For some, the latter is defined as the physical environment. Picard (1918), for example, points to variations in climate, soil, forest cover, and distance from the sea as important in explaining legal diversity: "All these and other variations of land and atmosphere show their traces in the jural system … Brazil and Russia, because they are different geographic regions, will therefore have many differences in law" (p. 170). In keeping with the assumptions of the day, other scholars point to 'race' as an explanatory factor in legal variation. Similar regional schools also seem to have developed in interwar France and Germany.

Impact Analysis

If regional studies read from geography to law, an influential paper by American geographer Whittlesey (1935) inverted this relationship in his call for the exploration, as the title put it, of the 'impress of effective central authority on the landscape.' Trade tariffs or embargoes, land tenure systems, and the regulation of resource exploitation were all identified as laws that result in 'landscape modification.' Thus, in understanding a particular region, geographers were urged to consider those 'phenomena engendered by political forces' (p. 97). The geographer's task, it seems, was now to focus on law's spatial 'impacts,' while taking law itself as an external 'given.'

Such a definition of the relation between law and geography (which Clark, 1989, terms 'impact analysis') proved influential until the late 1980s. The tendency was to treat space as a surface upon which laws instrumentally impact. For Wije (1990, pp. 26, 30), for example, "Law is a means of achieving social control by regulating behavior or by altering behavioral incentives while geography can study the spatial ramifications of this behavior. … Law can be considered as a geographic… variable."

Both traditions offer specific insights. A regional analysis is useful to the extent that it reminds us of the ways in which law, despite appearances of homogeneity, can in fact be different and diverse. In reminding nonlawyers of the undeniable importance of laws in social life and spatial forms and patterns,

impact analyses are also useful. We only need to reflect on the direct and indirect ways in which, for example, the urban landscape is a product of countless legal decisions, rulings, and mediations relating to real property, administrative law, and so on. Yet some commentators found these analyses partial and limited in terms of the ways in which they theorized law, space, and their interrelation (cf. Clark, 1989; Blomley, 1994).

Legal Geography

The 1990s, consequently, saw the emergence of a third school of legal-geographic inquiry that, in building upon but departing from the ideas discussed here, attempted to redefine both law and space; it also moved beyond the dualistic thinking that relates one to the other. The work of Gordon Clark, especially his *Judges and the Cities* (1985), and Nicholas Blomley, in his *Law, Space and the Geographies of Power* (1994), was important here. Identifying itself as 'legal geography,' this scholarship can be traced within geography, law, sociolegal studies, anthropology, and elsewhere. While far from unified, much of it drew from a set of emergent claims concerning law, space, and their mutual relationship. Essentially, legal geographers productively brought together parallel retheorizations of law and space.

First, in keeping with much recent scholarship, law itself was defined as more than a set of positive, external commands; it was to be understood as implicated in an array of social relationships. For example, law can produce many social and political effects, as it is able to define relationships, confer status, and selectively empower. This is important, given the reach of law in our lives, defined either formally (in the sense of statutes, judicial interpretation, etc.) or informally (the more messy, but no less important, legal sensibilities that we carry in our heads and express in our daily practices). As Twining (2000, p. 1) notes, law is omnipresent:

> One way of introducing students to the study of law is to ask them to read every word in a newspaper and to mark all the passages that they think deal with law or are "law-related" … Even those who adopt a narrow conception of law find it on every page … On the arts page, pornography, copyright, defamation, and other issues relating to freedom of expression arise along with licensing, charities law, taxation, and the ubiquitous contract. The advertisements are permeated with legal words and phrases. And so on through the paper.

Law can usefully be understood not only as a set of operative controls, but also as a repertoire of cultural and political meanings through which citizens can negotiate and interact with one another. Legal discourses offer a 'relatively plastic medium for refiguring the terms of past settlements over legitimate expectations and for expressing aspirations for new forms of entitlement' (McCann, 1994). These discourses were seen as important, shaping the ways in which people conceive of the political realm and their place within it. For some critical scholars, law was seen not as socially benign or neutral, but as confining and oppressive, naturalizing hierarchy and limiting progressive alternatives (Peller, 1985).

What, then, of space? In tandem with legal scholarship, the 'spatial turn' in social theory saw Marxist, feminist, and poststructuralist theory arguing that space was to be understood not as an objective, a priori phenomenon, but as both socially produced and socially constitutive (cf. Lefebvre, 1979). This, again, was significant, given the claim for the wide-ranging importance of space to social life, relations, and identities. An everyday activity, such as a visit to a shopping mall, includes spatial practices (such as the act of shopping), spatial representation (the advertisements and designs within the mall, often designed to conjure up images of other exotic places), place (or perhaps the absence thereof, given the ways in which corporate domination and mall design conjoin to produce 'placeless' spaces), and scale (marketing and distribution networks through which 'global' products are channeled to our local stores).

Such spaces are again not inert, but rather socially and politically consequential. Examples of the social and political dimensions of space can be traced at various levels of inquiry. For example, simple reflection reveals the ways in which interpersonal spatial categories such as 'near' and 'far' are not abstract concepts but socially contingent, as the concept of 'personal space' reveals, as well as being culturally diverse. However, the categories are also more complicated. The workings of racism, for example, depend in part upon spatial arrangements that place racialized groups identified as unlike 'us' at a spatial remove from those who we hold 'near and dear.'

Viewed thus, the retheorizations of both space and law would seem to have opened up considerable common ground. Both law and space are produced through social and political action. In other words, they are not transcendental, 'natural' categories. Law and space can be experienced at many different levels. Both are deeply representational, entailing stories, discourses, and images; in this way, they provide an important vocabulary for representing and organizing the world. Consider that which is evoked by concepts such as the 'Third World,' for example, or 'blind justice.' Both are also corporeal. Law is deeply concerned with the body – for example, the way that bodies are displayed (pornography, obscenity, etc.); what can be done to bodies (murder, rape, abortion, tattoos, etc.); and the movement, confinement, and control of bodies (e.g., through trespass, imprisonment, or capital punishment). The body is also spatially constituted and conceived – think of the different expectations concerning the comportment of men and women in public space, for example. Law and space are also material, beyond the body; that is, we experience both not only through representations but also in our encounters with material objects. Law is not only texts and words and images. It is also bullets, fences, prisons, 'For Sale' signs, doors, and windows. Similarly, space is made manifest and significant through, for example, the arrangement of chairs in a room, or the socially differentiated ways in which people set a table.

Separately, then, a considerable body of scholarship had argued for the respective significance of space and law along these lines. Yet, surprisingly, until the 1990s, there had been no systematic attempt to connect the two. This was the task that legal geography set itself, with the insistence that in doing so, we learned something new about law's work in the world. Social and political life, legal geographers argue, can be usefully thought of as simultaneously spatial and legal: "'law' and

'geography' do not name discrete factors that shape some third prelegal, aspatial entity called society. Rather the legal and the spatial are, in significant ways, aspects of each other …" (Delaney et al., 2001, p. xviii). If law and space, separately, are important to social life, then their combined effects are worth looking at. A shopping mall, for example, is simultaneously a space and a bundle of legal practices and discourses. So, for example, the mall's significance to constitutional rights of expression is not an abstract one: it is the fact that people enter and use a physical space that gives it practical relevance. But the space of the mall itself is legally constituted, reliant upon the workings of property, contract, and labor law, as well as zoning and other factors.

These insights, and their ramifications, have fed a growing body of literature. Special issues of journals include *Urban Geography* (11, 5, and 6, 1990), *Stanford Law Review* (48, 1996), *Journal of Historical Geography* (2000, 28), *Law/text/culture* (2005, 9), and *International Journal of Law in Context* (6, 3, 2010). Book-length overviews include Holder and Harrison (2003), Blomley et al. (2002), Sarat et al. (2003), Taylor (2006), Benda-Beckmann et al. (2009), and Braverman et al. (forthcoming). Useful overviews can be found in Blomley (2003, 2008), Manderson (2005), and Blank and Rosen-Zvi (2010).

Legal-geographic research has been conducted on a number of phenomena, including racism (Delaney, 1998), nationalism (Darian-Smith, 1995), property (Blomley, 1998, 2005; Graham, 2010), public space (Mitchell, 1996), climate change (Benson, 2010; Osofsky, 2009), globalization (Mahmud, 2007; Berman, 2002; Robertson, 2011), Israel and Palestine (Kedar, 2001; Braverman, 2009), and colonialism (Harris, 2008; Mawani, 2003). Generalizing about this diverse literature is not easy. While much is avowedly critical in orientation, some is more applied. Much of the work has been focused on particular case studies, making generalization more difficult. One possible distinction (Delaney, 2003) is that between 'Law-in-space' and 'Space-in-law.' The former concerns the ways in which law – statutes, judicial interpretations, the practice of legal officials, and so on – helps produce space. So legal actors, such as judges, can usefully be thought of as geographers, to the extent that their interpretations and edicts make space. Property is an obvious example: the spatial architectures of land ownership, both public and private, in which we all live and negotiate an everyday basis, have been made through generations of legal practice on the part of judges, surveyors, lawyers, police officers, municipal officials, and, of course, owners and nonowners.

Law in this sense does not 'impact' upon space but rather helps constitute it, to give it meaning and significance. This is not straightforward; it often entails complicated forms of interpretation and disagreement. Erbsen (2011) describes the multiple spaces that are produced by constitutional practice in the United States, arguing that many discrete problems on which scholars have focused – such as the rights of US military detainees abroad, the role of federal law on Indian reservations, and the extraterritorial reach of state law – reflect a deeper spatio-legal ambiguity in how the US Constitution allocates power over different kinds of spaces.

One very important way in which law produces space is through its designation of boundaries, marking distinctions between spaces (e.g., the public and private world) (Blandy and Sibley, 2010). This can be highly localized (e.g., property boundaries) or globalized (e.g., international law, state boundaries, etc.). These boundaries are said to be consequential:

> Boundaries *mean*. They signify, they differentiate, they unify the insides of the spaces that they mark. *What* they mean refers to constellations of social relational power. And the form that this meaning often takes – the meaning that social actors confer on lines and space – is *legal* meaning. How they mean is through the authoritative inscription of legal categories, or the projection of legal images and stories on to the material world of things (Delaney et al., 2001, p. xviii).

However, others have questioned the manner in which boundaries generate legal meaning, arguing for more nuanced and ambiguous reception processes (e.g., Blomley, 2005).

'Space-in-law,' conversely, attempts to come at the geographies of law from the other direction, tracing the ways in which legal practice and discourse are shot through with spatial tropes and metaphors. Legal actors frequently engage in spatial representation, being concerned with the delineation of zones, boundaries, and domains. When a judge designates a shopping center as private, from this perspective, she draws upon (and helps reproduce) a particularly consequential 'map' of the social world. When property rights are talked about in terms of autonomy, boundaries, and separation, we also rely upon a geographic set of metaphors.

These legal representations touch all aspects of legal life, including property, constitutional law, contract, crime, and intergovernmental law. Don Mitchell (1996), for example, outlines the ways in which judges and other legal actors help define the boundaries between public and private space in US cities, and regulate the activities that are deemed permissible in such spaces. This, he notes, has had particularly oppressive consequences for those people who effectively lack private space – the homeless. Legal action and interpretation in this sense can actively create certain spatial arrangements, practices, or representations. So, for example, David Delaney (1998) analyzes legal decisions concerning the legality of restrictive covenants in the United States, designed to exclude black residents from white neighborhoods. Necessarily, this entailed various construals of urban space by the judiciary. For Delaney, the effect of such apparently objective mappings was to position the judge as an active participant in the construction of ethnic spatial segregation. Housing covenants and similar approaches to restrict property ownership have been outlawed in the United States, through Shelley v. Kraemer and McGhee v. Sipes. The Supreme Court held that the creation and maintenance of private racial covenants resulted in a state action that violated federal law, in this case the Fourteenth Amendment of the US Constitution.

However, some see the analytical divide between law-in-space and space-in-law as increasingly confining, to the extent that it invites a restrictive view of the relationship between space and law. Rather than treating law and space as autonomous categories and then tracing the effect of one upon the other, both should be treated as recursively interrelated, it is argued. A 'legal' concept such as 'jurisdiction,' for example, is also

irreducibly spatial. The prison and the shopping mall are both 'spatial' and 'legal,' such that it is impossible to separate the one from the other. Some legal geographers argue that we need a new language that takes us beyond such dichotomies, allowing us to discern the reciprocal materializations and significations of law, space, society, and power. Delaney (2010) envisions the two coexisting as the 'nomosphere,' where law operates as a tool for 'world-making.' For Delaney, the nomosphere does not start from the spatial and move into the legal, but treats each as already there, intertwined in powerful and engrained ways. Such a view refuses a view of space as simply material, and of law as exclusively representational. The neologism usefully extends the metaphor of geographing to identify law as instantiating space. If law is instrumental in making space, space too provides an apparatus for law. Similarly, Blomley (2003) has argued for an attention to the 'splice,' a composite that is both space and law. We can see an attention to the recursivity of the spatio-legal in relation to scholarship that rethinks state borders and public space.

In a decade marked by proclamations to defend territory and sovereignty, border sites have become testing grounds for new legal instruments in security, exclusion, and the War on Terror at the opening of the twenty-first century. As a result, identifying borders as a theoretical site of inquiry takes legal geography seriously and problematizes 'the line' as embedded in a larger series of practices and policies. Rather than imagining borders as static edges, new scholarship is bridging law and geography to think about these spaces as relational sites where 'power may be extended through,' bringing the border to a whole set of new locations far from the edges of the nation-state (Allen, 2011). In doing so, these scholars extend the border beyond a place of transit and examine the multiplicity of laws and jurisdictions that coalesce at borders, on immigrant bodies, and far off the shorelines of the nation-state. The border, consequently, is not an inert, fixed line, but a fluid and moving zone, productive of, and constituted by, legal practices.

Scholars such as Mat Coleman couple immigration law with ethnography to consider the way that US immigration law enforcement influences individual mobility and what Stuart Elden has called "the spatial extent of sovereignty" (Elden, 2009). In Coleman's work on post-9/11 immigration enforcement in the US South, border enforcement has been relocated away from the edges of the nation-state, as the US Department of Homeland Security has deputized local law enforcement to serve as federal immigration agents (Coleman, 2005, 2009). Coleman documents border enforcement practices moving beyond the border, with a complicated set of laws that carry the border to the migrant's body. In this way, post-9/11 immigration law carves out a new space, a spontaneous border, through passport checks on buses and biometric data collection in small communities far from the edges of the nation-state. Dubbed the 'thousand little borders' approach to immigration control by legal scholar Huyen Pham (2004), the displacement of border-work creates borders wherever migrants and the state cross paths. Coleman suggests that both borders and law are in motion, and his work on two particular laws (287(g) referenda and the Secure Communities program) applies issues of law and jurisdiction to space and territory. In so doing, he importantly shows the way each is produced through the other – that immigration law and border security help, as Joe Painter puts it, to 'effectuate the territory' and in turn produce complex spaces of interlegality (Painter, 2010).

But how is this mobile border itself experienced? Alison Mountz explores moments of legal and spatial liminality as asylum seekers leave one state yet do not fully arrive in the territory of another. Refugees are consigned to an interstitial spatial-legal moment – a period marked by "waiting, limbo, disruption of life before and after and legal and jurisdictional ambiguity that inhibits access to rights and protections encoded in domestic and international law" (Mountz, 2011, p. 381). Her 2010 book, *Seeking Asylum: Human Smuggling and Bureaucracy at the Border*, looks at the challenges that arise when migrants leave home and wait in detention, secondary screenings, and administrative holding centers. Through case studies of the United States, Canada, and Australia, Mountz shows that the border is not simply moving, as in Coleman's United States, but also telescoping. She likens the experience to a 'long tunnel,' whereby asylum applicants are kept in an intermediary space between home and legal personhood (Mountz, 2010). Much like the long jetway greeting passengers at the end of a flight, the long tunnel assumes a new space between 'here' and 'there' – an exceptional space somehow outside the territory for the purposes of asylum.

Another example of the recursive nature of law and space can be seen in scholarship on public space (Layard, 2010). The sidewalk and the plaza are equally 'spatial' (entailing borders, representations of space, spatial practices, material objects, and bodies in motion) and 'legal' (implicating concepts and practices relating to citizenship, police powers, rights, privacy, etc.). Scholarship that is fully attentive to the complicated imbrication of the two offers particularly powerful insights. Take, for example, the work of Beckett and Herbert (2010) on the regulation of public space in Seattle. Eschewing generalized claims for regulation, they note the use of very particular spatial tactics of control and exclusion, tracing their development, reach, and effects. To control space, it seems, is to organize and materialize law. Another example comes from Blomley's (2011) work. He avoids discussion of 'public space' as an abstraction, focusing on the sidewalk as a particular site. Avoiding generalized claims concerning 'law' in the abstract, he focuses on those actors (judges, city engineers, etc.) who literally produce the sidewalk and govern and interpret its uses and functions. In so doing, he reveals a very particular (and overlooked) legal logic. Yet this is not simply a 'legal' account. He also traces the ways in which the space of the sidewalk, and its constituent parts and users, is visualized, ordered, and produced, in a constant tension between fixity and placement, on the one hand, and flow and circulation, on the other.

Legal geography offers important and valuable insights. Thinking of the legal (such as it relates to property or crime) in terms of space, or the spatial (such as a boundary or place) in terms of law, promises new and enriched insight. Understandings of both space and law are changed, and new questions posed. This is important in an analytical sense: if we are interested in how spatial arrangements are put together or sustained over time, recognizing the role of law is important, and it has an ethical and social justice dimension that scholars need to be attuned to. For better or worse, legal geographies have political

and moral effects. Designating the mall as 'private property' can close down certain possibilities. Yet workers, seeking to picket their employers within the mall, may also use a legal language of rights, freedom, and citizenship to challenge such geographies.

See also: Critical Legal Studies; Globalization: Legal Aspects; International Law and Treaties; Law and Democracy; Law and Development; Law and Economics: Empirical Dimensions; Law and Everyday Life; Law and Society: Development of the Field; Law as an Instrument of Social Change; Legal Pluralism; Legal Scholarship.

Bibliography

Allen, John, 2011. Topological twists: power's shifting geographies. Dialogues in Human Geography 1 (3), 283–298.

Beckett, Katherine, Herbert, Steve, 2010. Banished: The New Social Control in Urban America. Oxford University Press, Oxford.

Benda-Beckmann, Franz von, Keebet von Benda-Beckmann, Anne Griffiths (Eds.), 2009. Spatializing Law: An Anthropological Geography of Law in Society. Ashgate, Farnham, England.

Benson, Melinda Harm, 2010. Regional initiatives: scaling the climate response and responding to conceptions of scale. Annals of the Association of American Geographers 100, 1025–1035.

Berman, Paul Schiff, 2002. The globalization of jurisdiction. University of Pennsylvania Law Review 151, 311–529.

Blandy, Sarah, Sibley, David, 2010. Law, boundaries, and the production of space. Social and Legal Studies 19 (3), 275–284.

Blank, Yishai, Rosen-Zvi, Issi, 2010. The spatial turn in legal theory. HAGAR: Studies in Culture, Polity and Identities 10 (1), 39–63.

Blomley, Nicholas, 1994. Law, Space and the Geographies of Power. Guilford, New York.

Blomley, Nicholas, 1998. Landscapes of property. Law and Society Review 32 (3), 567–612.

Blomley, N., Delaney, D., Ford, R.T. (Eds.), 2002. The Legal Geographies Reader: Law, Power, and Space. Blackwell, Oxford.

Blomley, Nicholas, 2003. From "what?" to "so what?" Law and geography in retrospect. In: Holder, J., Harrison, C. (Eds.), Law and Geography. Current Legal Issues, vol. 5. Oxford University Press, pp. 17–33.

Blomley, Nicholas, 2005. Flowers in the bathtub: boundary crossings at the public–private divide. Geoforum 36, 281–296.

Blomley, Nicholas, 2008. How law matters to political geography. In: Cox, K., Low, M., Robinson, J. (Eds.), The SAGE Handbook of Political Geography. SAGE Publishers, New York.

Blomley, Nicholas, 2011. Rights of Passage: Sidewalks and the Regulation of Public Flow. Routledge, New York.

Braverman, Irus, 2009. Planted Flags: Trees, Land and Law in Israel/Palestine. Cambridge University Press, Cambridge.

Braverman, Irus, Blomley, Nicholas, Delaney, David, Kedar, Alexander (Eds.) forthcoming. The Expanding Spaces of a Law: A Timely Legal Geography. Stanford University Press.

Clark, Gordon L., 1985. Judges and the Cities: Interpreting Local Autonomy. University of Chicago Press, Chicago, IL.

Clark, Gordon L., 1989. The geography of law. In: Peet, R., Thrift, N. (Eds.), The New Models in Human Geography. Unwin Hyman, London, UK, pp. 310–337.

Coleman, Mathew, 2005. U.S. statecraft and the U.S.–Mexico border as security/economy nexus. Political Geography 24 (2), 185–209.

Coleman, Mathew, 2009. What counts as the politics and practice of security, and where? Devolution and immigrant insecurity after 9/11. Annals of the Association of American Geographers 99 (5), 904–913.

Darian-Smith, Eve, 1995. Law in place: legal mediations of national identity and state territory in Europe. In: Fitzpatrick, P. (Ed.), Nationalism, Racism and the Rule of Law. Dartmouth, Aldershot, UK, pp. 27–44.

Delaney, David, 1998. Race, Place and the Law – 1836–1948. University of Texas Press, Austin, TX.

Delaney, D., 2003. Beyond the word: law as a thing of this world. In: Holder, J., Harrison, C. (Eds.), Law and Geography. Oxford University Press, Oxford.

Delaney, David, 2010. The Spatial, the Legal and the Pragmatics of World-Making: Nomospheric Investigations. Routledge, New York.

Delaney, D., Blomley, N., Ford, R.T., 2001. Where is law? In: Blomley, N., Delaney, D., Ford, R.T. (Eds.), The Legal Geographies Reader: Law, Power, and Space Blackwell. Oxford, pp. xiii–xxii.

Elden, S., 2009. Terror and Territory. University of Minnesota Press, Minneapolis, MN.

Erbsen, Allen, 2011. Constitutional spaces. Minnesota Law Review, 1168–1267.

Graham, Nicole, 2010. Lawscape: Property, Environment and the Law. Routledge, New York.

Harris, Doug, 2008. Landing Native Fisheries: Indian Reserves and Fishing Rights in British Columbia, 1849–1925. UBC Press, Vancouver.

Holder, J., Harrison, C. (Eds.), 2003. Law and Geography. Oxford University Press, Oxford.

Kedar, Alexander, 2001. The legal transformation of ethnic geography: Israeli law and the Palestinian landholder, 1948–68. New York University Journal of International Law and Policy 33, 923–1000.

Kocourek, Albert, Wigmore, John Henry (Eds.), 1918. Formative Influences of Legal Development. Little Brown, Boston, MA.

Layard, Antonia, 2010. Shopping in the public realm: a law of place. Journal of Law and Society 37 (3), 412–441.

Lefebvre, Henri, 1979. Reflections on the politics of space. Antipode 8 (2), 30–37.

Mahmud, Tayyab, 2007. Geography and international law: a postcolonial mapping. Santa Clara Journal of International Law 5, 525.

Manderson, Desmond, 2005. Interstices: new works on legal spaces. Law/text/culture 9, 1–10.

Mawani, Renisa, 2003. 'Law, space and the making of Stanley Park, 1859–2001. Law/Text/Culture 7, 98.

McCann, Michael W., 1994. Rights at Work: Pay Equity Reform and the Politics of Legal Mobilization. University of Chicago Press, Chicago, IL.

Mitchell, Don, 1996. Political violence, order and the legal construction of public space: power and the public forum doctrine. Urban Geography 17 (2), 152–178.

Montesquieu, B.de., 1748/1949. The Spirit of the Laws. Hafner Press, New York.

Mountz, Alison, 2010. Seeking Asylum: Human Smuggling and Bureaucracy at the Border. University of Minnesota Press, Minneapolis, MN.

Mountz, Alison, 2011. Where asylum-seekers wait: feminist counter-topographies of sites between states. Gender, Place, Culture 18 (3), 381–399.

Osofsky, Hari M., 2009. Scaling "local": the implications of greenhouse gas regulation in San Bernardino County. Michigan Journal of International Law 30, 689–705.

Painter, Joe, 2010. Rethinking territory. Antipode 45 (2), 1090–1118.

Peller, Gary, 1985. The metaphysics of American law. California Law Review 73, 1151–1290.

Pham, Huyen, 2004. The inherent flaws in the inherent authority position. Florida State University Law Review 31 (4), 965–1003.

Picard, E., 1918. Factors of legal evolution. In: Kocourek, A., Wigmore, J.H. (Eds.), Formative Influences of Legal Development. Little Brown, Boston, MA, pp. 163–181.

Robertson, Sean, 2011. Postdevelopmental Properties in the Age of the Exception: The Political and Affective Lives of the Traditional Environmental Knowledge of Plateau Peoples in British Columbia. Unpublished Doctoral Dissertation, Department of Geography, Simon Fraser University, Burnaby, BC, Canada.

Sarat, A., Douglas, L., Umphrey, M.M. (Eds.), 2003. The Place of Law. University of Michigan Press, The Amherst Series in Law, Jurisprudence and Social Theory.

Taylor, William, 2006. The Geography of Law: Landscape, Identity and Regulation. Hart Publishing, Oxford.

Twining, W., 2000. Globalization and Legal Theory. Butterworths, London.

Whittlesey, D., 1935. The impress of effective central authority upon the landscape. Annals of the Association of American Geographers 25, 85–97.

Wigmore, John H., 1928. A Panorama of the World's Legal Systems. West Publishing Company, St Paul, MN.

Wije, C., 1990. Applied law and applied geography. Operational Geographer 8 (1), 26–31.

Law and Morality: A Continental European Perspective

Robert Alexy, Christian-Albrechts-Universität, Kiel, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The relation between law and morality is one of the main problems of legal philosophy. In a discussion that reaches back more than 2000 years, two positions have prevailed: positivism and nonpositivism. The main argument in favor of nonpositivism and against positivism is the claim of law to correctness. This claim implies that law has a dual nature, which comprises a real or authoritative dimension as well as an ideal or critical dimension. The dual nature of law is the basis for the Radbruch formula, which says, in its shortest form, that extreme injustice is not law.

Under the general rubric of the concept and nature of law, one of the most fundamental problems is the relation between law and morality. Two positions have competed for attention for over two millennia: positivism and nonpositivism.

Separation Thesis and Connection Thesis

All positivists defend the separation thesis. In its most general form, it says that there is no necessary connection between the law as it is and the law as it ought to be. Or, in a more precise formulation, it states that there is no necessary connection between legal validity or legal correctness, on the one hand, and moral merits or moral correctness, on the other. By contrast, all nonpositivists defend the connection thesis, which says that there is a necessary connection between legal validity or legal correctness, on the one hand, and moral merits or moral correctness, on the other (Alexy, 2008b, 284–5). This implies that in order to determine the concept and nature of law, all positivistic theories are confined to two elements, namely, authoritative issuance and social efficacy. The characterization of nonpositivistic theories includes a third element as well, correctness of content (Alexy, 2002a, 3–4).

Forms of Positivism and Nonpositivism

Which thesis, the separation thesis or the connection thesis, is more defensible? Both lend themselves to a variety of interpretations. The answer to our main question turns on these various interpretations.

Within positivism, the distinction between exclusive and inclusive legal positivism is the most important difference where the relation between law and morality is concerned. Exclusive positivism, as defended most prominently by Joseph Raz, maintains that morality is necessarily excluded from the concept of law (Raz, 2009, 47). Inclusive positivism, as defended, for instance, by Jules Coleman, says that morality is neither necessarily excluded nor necessarily included. The inclusion is declared to be a contingent or conventional matter turning on what the positive law in fact says (Coleman, 1996, 316). This implies that the relation between law and morality in both cases, that of inclusive positivism as well as that of exclusive positivism, is determined solely by what is authoritatively issued and socially efficacious, that is, by social facts. Inclusive positivism is a form of positivism because it claims that the initial decision in a particular legal system to include morality in the law is contingent or conventional (Alexy, 2012, 4). Nonpositivism argues not only–against exclusive positivism–that morality is not necessarily excluded, but also–against inclusive positivism–that it is necessarily included; nonpositivism is therefore contrary to both forms of positivism.

The differences within nonpositivism are no less important than the differences within positivism. Of special importance for the debate over the concept and nature of law are the differences that stem from different effects on legal validity that are attributable to moral defects. Nonpositivism can determine the effect on legal validity that stems from moral defects or demerits in three different ways. It might be the case that legal validity is lost in all cases, that legal validity is lost in some cases but not in others, or, finally, that legal validity is affected in no way at all (Alexy, 2008b, 287).

The first position, according to which every moral defect yields legal invalidity, is the most radical version of nonpositivism. This position might be characterized as 'exclusive nonpositivism' in order to express the idea that each and every moral defect precludes legal validity. With this, in cases of moral defects, it excludes social facts from the sources of law. Augustine gives us a classical statement of this view when he says that "a law that was not just would not seem to me to be a law" (Augustinus, 2006, 86). A recent example is Deryck Beyleveld and Roger Brownsword's thesis that "immoral rules are not legally valid" (Beyleveld and Brownsword, 2001, 76).

The radical counterpart of exclusive nonpositivism is superinclusive nonpositivism. Superinclusive nonpositivism goes to the other extreme. It maintains that legal validity is affected in no way whatever by moral defects. At first glance, this seems to be a version of positivism, not nonpositivism (Waldron, 1996, 1566). This first impression, however, is recognized as misleading as soon as one sees that there exist two different sorts of connection between law and morality: a classifying and a qualifying connection (Alexy, 2002a, 26). These two sorts of connection are distinguished by the effects of moral defects. The effect of a classifying connection is the loss of legal validity. By contrast, the effect of a qualifying connection is legal defectiveness, which does not, however, reach so far as to undermine legal validity. It does, however, create a legal obligation or at least an empowerment on the part of appellate courts to quash

unjust judgments of lower courts. Immanuel Kant's combination of the postulate of '[u]nconditional submission' (Kant, 1996, 506) to the positive law with the idea of a necessary subordination of the positive law to nonpositive law can be read as a version of superinclusive nonpositivism (Alexy, 2008b, 289; 2010, 176). The same is true of Aquinas' thesis that a tyrannical law is law but 'not law simpliciter' (Aquinas, 1962, 947), or, as John Finnis puts it, "not law in the focal sense of the term 'law'" (Finnis, 2011, 364).

The third version of nonpositivism, inclusive nonpositivism, is found between the extremes of exclusive nonpositivism and superinclusive nonpositivism. Inclusive nonpositivism claims neither that moral defects always undermine legal validity, as represented by exclusive nonpositivism, nor that they never do, as represented by superinclusive nonpositivism. It claims that moral defects undermine legal validity under some conditions but not under others.

Inclusive nonpositivism is given its most prominent expression in the Radbruch formula, which, in its most compressed form, runs as follows: Extreme injustice is not law (Radbruch, 2006, 7; Alexy, 2008a, 427–8). According to this formula, moral defects undermine legal validity if and only if the threshold of extreme injustice is transgressed. Injustice below this threshold is included in the concept of law as defective but valid law. This means that inclusive nonpositivism includes a considerable degree of positivity, that is to say, a commitment to what is authoritatively issued and socially efficacious. The nonpositivism of inclusive nonpositivism consists, first, in establishing an outermost border of law and, second, in qualifying immoral or unjust law as not only morally but also legally defective. The practical consequences of establishing an outermost border become evident when one looks at the application of the Radbruch formula by German courts after the defeat of National Socialism in 1945 and after the collapse of the German Democratic Republic in 1989 (Alexy, 2008a, 428–32). A practical consequence of qualifying immoral or unjust law as not only morally but also legally defective is that appellate courts acquire the possibility of quashing unjust judgments of lower courts owing to their legal defectiveness.

The Argument from Correctness

The existence of two forms of positivism and three forms of nonpositivism shows that the debate between positivism and nonpositivism concerns far more than a mere contest between two monolithic positions, often presented as the opposition of 'legal positivism' and 'natural law.' And this is not all. Things become even more complicated as soon as one takes account of the fact that not only positivism as well as nonpositivism are in themselves complex. Complexity is also manifest in the structure of arguments for and against the different forms of positivism and nonpositivism. The Archimedean point of this structure is the argument from correctness. All other arguments revolve around this center.

The argument from correctness says that individual legal norms and individual legal decisions as well as legal systems as a whole necessarily lay claim to correctness. Ronald Dworkin has objected that the question of whether the representatives of law raise any claims is a question of fact, and not a question of necessity (Dworkin, 2006, 200). This objection can be rejected if it is possible to show that the claim to correctness is necessarily implicit in law, quite apart from the intentions of its representatives. Here the idea is to show that the explicit negation of the claim of correctness leads to a contradiction (Alexy, 2002a, 35–9). An example is the fictitious first article of a constitution that reads: X is a sovereign, federal, and unjust republic. This article is somehow absurd. The absurdity stems from a contradiction between what is implicitly claimed in framing a constitution, namely, that it be just, and what is explicitly declared, namely, that it is unjust. Now justice counts as a special case of correctness, for justice is nothing other than correct distribution and compensation. Therefore, the contradiction in our example is not only a contradiction with respect to the dichotomy of just and unjust but also a contradiction with respect to the dichotomy of correct and incorrect.

What is more in the aforementioned example of the fictitious first article of a constitution, the contradiction there between what is explicit and what is implicit is necessary. It could be avoided only if one were to abandon the implicit claim. But to do this would represent the transition from a legal system to naked power relations (Alexy, 1998, 213–4). Thus, our example shows that law and the claim to correctness are connected not only by contingent, prudential reasons, but also – and this goes a good bit further – by reasons necessary in nature. This connection is by no means confined to such fundamental acts as framing a constitution. It is present everywhere in the legal system, and may be illustrated by the absurdity found in decisions such as the following: The accused is sentenced to life imprisonment, which is an incorrect interpretation of prevailing law.

In order to establish a necessary connection between law and morality, it is not enough that the claim to correctness be necessarily raised by law. Over and above this, it is necessary that its content necessarily refer to morality.

The claim of law to correctness would not necessarily refer to morality if it were possible that it refer exclusively to social facts, that is, to what has been authoritatively issued and is socially efficacious. A claim with this content would count as a purely positivistic claim to correctness. Hard cases, however, illustrate that the positivistic interpretation of the correctness argument gives rise to serious problems. Hard cases occur when the positive law – that is, the authoritative or source-based reasons – allows for more than one decision. The decision to be made in such an 'open' sphere is a decision on a normative matter that cannot be based on standards of positive law, for if it could be based on such standards, it would not be a decision in an 'open' sphere. If it is to be based on any standards at all – that is, if it is not to be a merely arbitrary decision, a decision that would contradict the claim to correctness – it must be based on other normative standards. Legal decisions regularly concern questions of distribution and compensation. Questions of correct distribution and compensation are questions of justice. Questions of justice, however, are moral questions. In this way, the open texture of law, taken together with the nature of legal questions, implies that the claim to correctness raised in legal decision-making necessarily refers not only to authoritative or source-based reasons but also to moral reasons. This implies that the claim to correctness necessarily raised in law leads to a necessary inclusion of morality in law (Alexy, 2007, 49–50).

The Dual Nature of Law

Hans Kelsen has argued against nonpositivism that it presupposes "an absolute morality, that is to say, a morality that is valid everywhere and at all times" (Kelsen, 1967, 68 trans. alt), and, Kelsen goes on, no such absolute morality exists. One might call this the 'argument from relativism' (Alexy, 2002a, 53–5). The question of whether Kelsen's objection based on relativism is cogent is a question of interpretation. If one understands the objection as saying that there is no instance in which only one correct moral answer can be given, then the objection fails. There are cases of severe interference with human rights in which only one moral answer is correct or true, namely, that the interference violates human rights and is, for that reason, morally wrong (Alexy, 2012, 8–13). The existence of such cases suffices as an epistemological or meta-ethical basis of nonpositivism. If, however, one interprets Kelsen's objection as saying that there exist a number of cases, even a considerable number, in which 'reasonable disagreement' (Rawls, 1993, 55) about what is morally right or wrong is possible, then Kelsen's argument is based on a correct epistemological or meta-ethical thesis, but, and this is the point, it would then no longer suffice as an argument against nonpositivism. Nonpositivism is compatible with reasonable disagreement if it is possible to arrive at an approximation to truth or correctness in discourse and if there are at least some cases in which only one moral answer is right, that is, in which reasonable disagreement is not possible. Approximation to truth or correctness is possible because rational practical discourse is possible, and there are cases in which only one moral answer is possible for the reason that other results in these cases are discursively impossible – as, for instance, the legal status of a slave or the abolition of religious freedom (Alexy, 1989, 187–208).

The existence of reasonable disagreement means that there are a considerable number of social problems that cannot be resolved by moral argument alone. One might call this the 'problem of practical knowledge'. The problem of practical knowledge can be resolved only by means of legally regulated procedures that guarantee a decision. This is the step from morality to positive law, as described, for instance, by Kant (Kant, 1996, 456). What is more, the problem of practical knowledge is not the only problem that can be resolved only by means of positive law. A second problem is the problem of enforcement. If it is possible to violate the law without running any risk, and if some take advantage of this possibility, then compliance with the regulation is no longer assured. In short, procedures that provide for the enforcement of law are necessary. In addition, there is the problem of organization. A modern society can be effectively organized only by means of positive law.

The need to solve these three problems and, therefore, the necessity of positivity, that is, of authoritative issuance and social efficacy, stems from the moral requirements of avoiding the costs of anarchy and civil war and assuring the advantages stemming from social coordination and cooperation. As moral reasons, these reasons are elements of the content of the claim of law to correctness. This implies that the claim of law to correctness qua claim to moral correctness necessarily comprises elements of positivity. This is not to say, however, that the claim to correctness comprises solely elements of positivity. This is the mistake of superinclusive nonpositivism. The claim to substantial correctness – that is, first and foremost, the claim to justice – does not disappear once law is institutionalized. It lives on in the law. For this reason, one has to distinguish two stages or levels of correctness: first-order correctness and second-order correctness.

First-order correctness concerns justice as such. Second-order correctness is more comprehensive. It refers both to justice and to positivity. Justice represents the ideal or critical dimension of law, positivity its real, factual, or institutional side. The claim of law to correctness qua second-order claim unites the real and ideal dimensions of law. It is an expression of the dual nature of law (Alexy, 2010, 173–4).

The dual nature of law implies that law necessarily comprises two principles: the principle of justice and the principle of legal certainty. The principle of legal certainty is a formal principle. It requires a commitment to what is authoritatively issued and socially efficacious. The principle of justice is a material or substantive principle. It requires that the decision be morally correct. These two principles, as is the case with principles generally, may collide, and they often do. Neither can ever supplant the other completely, that is to say, in all cases. On the contrary, the dual nature of law demands that they be seen in correct proportion to each other. This, in turn, can only be achieved by balancing. The idea of an outermost border of law is a result of such balancing, that is, of balancing the principles of legal certainty and justice.

Outermost Border

The Radbruch formula that, in its shortest form, says that extreme injustice is not law is the classical expression of the idea of an outermost border of law (Radbruch, 2006, 7; Alexy, 2008a, 428). This formula represents, first and foremost, a rejection of the positivistic thesis that 'any kind of content might be law' (Kelsen, 1967, 198). Kelsen illustrates this thesis with the following remark: "According to the law of totalitarian states, the government is empowered to confine in concentration camps persons with rejected convictions, religion, or race and to force them to do any sort of work whatever, even to kill them" (Kelsen, 1967, 40, trans. altered). Killing persons in concentration camps for reasons of rejected convictions, religion, or race is a clear case of extreme injustice. Therefore, according to the Radbruch formula, norms that empower legal officials to do such things cannot be valid law. From the positivistic point of view, the situation is different. If these norms have been authoritatively issued and are socially efficacious, they are valid law. This is also true of inclusive positivism in instances where the positive law does not actually refer, in a socially efficacious way, to moral principles that exclude the killings. From the nonpositivistic point of view, everything depends on balancing the principles of legal certainty and justice. The principle of legal certainty stands for the view that the norm in Kelsen's example is legally valid, the principle of justice stands for the opposite result. The determination arrived at by balancing essentially depends on the intensity of interference with each of the colliding principles (Alexy, 2002b, 102). Not to conceive of the norm in question as invalid would count as an extremely intensive

interference with the principle of justice, for justice comprises human rights, and the interference with human rights in Kelsen's example is extreme. On the other hand, the loss in legal certainty if one were to conceive the norm as invalid would be quite limited. Even in totalitarian states, a great number of norms do not exceed the threshold of extreme injustice. Thus, the result of balancing is determined, and this result reflects precisely the Radbruch formula, that is to say, inclusive nonpositivism as defined by the formula. By contrast, superinclusive nonpositivism as well as exclusive nonpositivism must fail. Superinclusive nonpositivism fails, for it gives too little weight to justice, and exclusive nonpositivism fails, for it gives too little weight to legal certainty.

Participant and Observer

This justification of the Radbruch formula, it might be objected, is flawed on the ground that it is based on normative arguments, namely, the principles of justice and legal certainty. The question of the nature of law, the objection continues, concerns what law is, and what a thing is cannot be established by means of normative arguments (Raz, 1996, 7; Marmor, 2005, 778).

To reply to this objection, the distinction between the observer's perspective and the participant's perspective is fundamental (Alexy, 2002a, 25). An observer poses questions and adduces arguments on behalf of a position that reflects the way in which legal questions are actually decided in a legal system, whereas the participant poses questions and adduces arguments with an eye to what he deems to be the correct answer to the legal question at hand. The observer's perspective is defined by the question 'How are legal decisions actually made?' the participant's by the question 'What is the correct legal answer?'

Positivism reflects quite well the observer's perspective (Alexy, 2002a, 27–35). From this perspective, what the law is depends exclusively on what has actually been issued and is socially efficacious. By contrast, the participant's perspective presupposes nonpositivism. For a participant, the law is what it is correctly taken to be. What it is correctly taken to be depends, however, not only on social facts but also on moral reasons. In this way, normative arguments make their way into what the law is as soon as the participant's perspective is taken up. What is more, this perspective must be taken up – for law is possible without observers but not without participants.

See also: Ethics and Values; Justice: Philosophical Aspects; Kantian Ethics and Politics; Philosophy of Law.

Bibliography

Alexy, R., 1989. A Theory of Legal Argumentation (R. Adler, N. MacCormick, Trans.). Clarendon Press, Oxford.
Alexy, R., 1998. Law and correctness. Current Legal Problems 51, 205–221.
Alexy, R., 2002a. The Argument from Injustice. A Reply to Legal Positivism (B. Litschewski Paulson, S.L. Paulson, Trans.). Clarendon Press, Oxford.
Alexy, R., 2002b. A Theory of Constitutional Rights. Trans Rivers, J. Oxford University Press, Oxford.
Alexy, R., 2007. An answer to Joseph Raz. In: Pavlakos, G. (Ed.), Law, Rights and Discourse. The Legal Philosophy of Robert Alexy. Hart Publishing, Oxford, pp. 37–65.
Alexy, R., 2008a. A defence of Radbruch's formula. In: Freeman, M.D.A. (Ed.), Lloyd's Introduction to Jurisprudence. Sweet & Maxwell and Thomson Reuters, London, pp. 426–443.
Alexy, R., 2008b. On the concept and the nature of law. Ratio Juris 21, 281–299.
Alexy, R., 2010. The dual nature of law. Ratio Juris 23, 167–182.
Alexy, R., 2012. Law, morality, and the existence of human rights. Ratio Juris 25, 2–14.
Augustinus, 2006. In: Brachtendorf, J., Decroll, V.H. (Eds.), De libero arbitrio. Schöningh, Paderborn.
Aquinas, T., 1962. Summa theologiae. Paoline, Turin.
Beyleveld, D., Brownsword, R., 2001. Human Dignity in Bioethics and Biolaw. Oxford University Press, Oxford.
Coleman, J., 1996. Authority and reason. In: George, R.P. (Ed.), The Autonomy of Law. Essays on Legal Positivism. Clarendon Press, Oxford, pp. 287–319.
Dworkin, R., 2006. Justice in Robes. Harvard University Press, Cambridge, MA.
Finnis, J., 2011. Natural Law and Natural Rights, second ed. Oxford University Press, Oxford.
Kant, I., 1996. The metaphysics of morals (M.J. Gregor, Trans. and ed.) pp. 353–603. In: Kant, I.,. Practical Philosophy. Cambridge University Press, Cambridge.
Kelsen, H., 1967. Pure Theory of Law (M. Knight, Trans.). University of California Press, Berkeley.
Marmor, A., 2005. Debate. Anales de la cátedra Francisco Suárez 39, 769–793.
Radbruch, G., 2006. Statutory lawlessness and supra-statutory law. Trans. Litschewski Paulson, B. and Paulson, S. L. Oxford Journal of Legal Studies 26, 1–11.
Rawls, J., 1993. Political Liberalism. Columbia University Press, New York.
Raz, J., 1996. On the nature of law. Archives for Philosophy of Law and Social Philosophy 82, 1–25.
Raz, J., 2009. The Authority of Law, second ed. Oxford University Press, Oxford.
Waldron, J., 1996. Kant's legal positivism. Harvard Law Review 109, 1535–1566.

Law and Morality: An Analytical Perspective

Tom Campbell, Charles Sturt University, Kingston, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Noting the similarities in content, form, and function between law and morality, the article examines significant differences between these two systems of normative social order. Having distinguished positive and critical morality, the article discusses whether (1) law ought to incorporate the positive morality of the citizens of that jurisdiction, (2) the critical moral principles that we adopt to criticize positive morality ought to be the same as those we use to criticize law, and (3) the role of human rights ought to keep law within morally acceptable bounds. This involves discussing the theoretical disagreements between legal positivism and natural law theory in terms of the 'separability thesis,' and touches on the controversial role of human rights as an essential moral minimum for valid law.

As complementary and yet partly competing systems of normative social order, law and morality have much in common and much to distinguish them. In content, form, and social function, law and morality exhibit similar features that can not only be mutually supportive but also generate social and political conflict, particularly in large states characterized by moral pluralism and constant social change.

There is substantial overlap in the content of the moral and legal norms of all societies, and law may readily be seen as both an expression of, and an influence on, a society's morality. Judges are normally expected to have regard to moral considerations when interpreting law, and, in many democratic countries, judges have power to override legislation that conflicts with those basic moral rights, such as freedom of expression, which are enshrined in the constitution of their polity. Further, legislatures are expected to make law in the light of the moral beliefs of the population. Conversely, the established law of a country has some effect on what its inhabitants regard as morally right and wrong by reinforcing and altering behavioral expectations.

Moreover, legal systems depend, in varying degrees, on securing the conformity of citizens to their requirements, which in turn depends, in part, on satisfying their moral expectations as well as their self-interest. In turn, the moral constraints that citizens impose on each other and on themselves through the informal processes of social approval and disapproval are crucially reinforced by the sanctions of criminal law and the remedies for such matters as breach of contract and negligently causing harm to others.

On account of their overlapping social functions, legal and moral discourse shares a whole range of basic concepts. Justice, fairness, rights, utility, equality, and freedom are the common vocabulary of both normative systems, which suggests that moral and legal obligations cannot always be clearly distinguished, and explains why in less developed societies there is little or no institutional divide in their normative structure between law and morality.

These commonalities and interactions of law and morality must be kept in mind when examining attempts to distinguish law and morals. They suggest that law and morality are parts of a wider system of normative order arising out of the elements of conflict and coordination that characterize human society. That said, there are important differences between law and morality that call for analysis, explanation, and evaluation. In particular, there is a recurrent and ongoing tension between the law of authoritarian states and the morality of oppressed majorities and between the law of democratic states and the morality of politically excluded minorities. Conceptualizing and explaining these conflicts and examining how they ought to be resolved are important themes in the social and behavioral sciences.

Critical and Positive Morality

Focusing on evident differences between law and morality, several key variables emerge: (1) law is *mandatory*, usually coercive, whereas morality is voluntary and often unenforceable; (2) law is an *institutionalized* system of norms, with specialized officials and procedures, whereas morality is individualistic, informal, and diffuse in its operation (MacCormick, 2007); (3) law is *public* and associated with the state, whereas morality is *private* and a matter for the individual to determine for herself; (4) law gives a significance to *precedents* and *sources* more than is the case with morality; and (5) laws may be *unjust*, but there is something incongruous in the notion of an unjust moral principle.

Some of these variables can be analyzed by distinguishing between two senses of 'morality.' The first sense refers to those norms of conduct that are generally observed and routinely used to criticize conduct in a given social group. This is sometimes called 'customary' or 'positive' morality (Austin, 1954: p. 122).

The second sense of 'morality' relates to the activity of criticizing and evaluating, from an external point of view, the norms and practices that make up both positive morality and the law of a society. It is this 'critical' morality (Hart, 1961: pp. 176–180), in which the content and form of positive moralities and actual legal systems are assessed morally, that gives rise to some of the perceived differences between law and morality. This happens where individuals and groups adopt a critical normative approach to the form and content of laws and legal systems. One of these differences is that critical morality is a matter of personal moral opinion (and cannot therefore be

coerced or forced on others), while the form and content of law (and indeed of positive morality) are matters of empirical fact. Critical morality may be contrasted with both legal and positive moral systems of norms, with law and positive morality sharing the characteristics that typify ongoing cultural structures and critical morality standing outside and above these phenomena. One contentious issue that arises here is whether, from the moral point of view, a person's critical morality takes precedence over their assumed moral obligation to obey the law. Precisely what that moral point of view may be is a controversial issue within moral philosophy, but we may take it to be the point of view that seeks to determine the ultimate criteria for assessing the worth of human conduct from an impartial or universal perspective (Baier, 1958).

Having distinguished positive and critical morality, we can frame the questions that arise in relation to law and morality in a more sophisticated way. Thus we may raise, as a matter of critical morality, (1) whether law ought to incorporate the positive morality of the citizens of that jurisdiction, (2) whether the critical moral principles that we adopt to criticize positive morality ought to be the same as those we use to criticize law, and (3) whether there is such a thing as 'critical law,' in contrast to 'positive law,' similar to the distinction between critical and positive morality, as most natural law theorists contend.

Legal Positivism and Natural Law

The relationships between law and morality have long been prime sites for the articulation of competing legal philosophies, particularly classical natural law and legal positivist approaches to the understanding and evaluation of law and legal systems. However, as we have seen, the relationships involved are varied in type and substance, and it is by no means clear that the traditional packages of theses associated with the classical theories still hang together. Indeed, there is a real prospect for conceptual realignments that render some of the historical debates on law and morality outmoded.

Legal positivism, which received its classical formulation in the work of Jeremy Bentham (1748–1832) and John Austin (1790–1859), emphasizes the contrasts between law and morality, and the political subordination of morals to the will of the sovereign (Austin, 1954). Following on from Hobbes' analysis of political authority, which he based on the requirements of mutual survival in a predatory social world, the state is seen as imposing a morally arbitrary rule of commands so as to secure peace between warring factions, religions, and moralities. This characteristically modern theory of the state replaced the medieval European synthesis of government, religion, and morality that received its classical formulation in the *Summa Theologica* of Thomas Aquinas, whose definition of law as "a rational ordering of things which concern the common good; promulgated by whoever is charge with care of the community" (Aquinas, 1966: Art. 4) sums up the fusion of law and morality in the natural law tradition.

Austin, through whose work legal positivism became the reigning orthodoxy in nineteenth century Anglo-American legal theory, presents his position in a series of definitions designed to identify actual or 'positive' (in the sense of observable) law in empirical terms, such as commands, sanctions, and habitual obedience. Law is defined as the general commands of a sovereign, that is, the person or group that is habitually obeyed in a certain territory and does not himself habitually obey anyone else. A command is the expression of a wish relating to the conduct of another and it is backed by threats of force, with the legal obligation being the likelihood that failure to comply with a command will be followed by the imposition of sanctions. Each state has one such sovereign (which may comprise an individual or a collectivity) who is at the giving (but not the receiving) end of such commands. The existence of this unitary sovereign is the core of the classical positivist analysis of the modern sovereign state.

It is a defining feature of this version of legal positivism that there is no necessary connection between law and morality (Hart, 1961: p. 181). The nature of the necessity embedded in the classical positivist separation of law and morality is based intellectually on Hume's distinction between 'is' (statements about what is and is not the case) and 'ought' (statements about what ought or ought not to be the case). In Austin's ringing words, "the existence of law is one thing, its goodness or badness another" (Austin, 1954: p. 184). Legal positivists insist that it is possible to separate out the description of law from our moral evaluation and criticism of that law: legal argument is not moral argument, legal authority is not moral authority, and a law is a law whether it is just or unjust, or good or evil. All this applies only to what is called 'positive law,' that is, the law as it is actually applied in human societies; but legal positivists contend that, strictly speaking, this is the only genuine law that exists.

Such a dichotomy depends on our capacity to identify law in an empirical manner, and this Austin does with his analysis of law as the command of a political sovereign. In the terminology of a more recent and equally famous legal positivist, H.L.A. Hart, the law of a jurisdiction can be identified by means of its 'rule of recognition' (Hart, 1961: pp. 92–93), which lays down the criteria used by legal officials to determine legal 'validity,' (i.e., what is and is not part of the law of that particular jurisdiction). In Austin's case, that rule identifies the commands of the political sovereign, the person or body that is habitually obeyed in a particular location, as the criterion that is necessary and sufficient for the determination of legal validity.

A similar approach is taken by the classical legal positivists to positive morality. The rules of conduct, which are widely followed in a given society and are backed by the sanctions of public approval and disapproval, are seen as a distinct system of identifiable moral norms analogous to positive law, although always overlapping with the positive laws of the same society in terms of form and content. Both of these positive systems are distinguished from 'natural law,' a term that Austin applies to what is actually morally right and wrong in human conduct, which he himself, following Bentham, takes to be the utilitarian standard of acting or refraining from acting so as to maximize human pleasure and minimize human pain. While such natural laws can be regarded, in positivist terms, as the commands of the ultimate sovereign – namely, God – Austin, and legal positivists generally, draw on utility as the basis for their critique of positive law and not as a test of legal validity. The classical natural law position, in contrast, is that knowledge of right and wrong as manifest in God's law, or discovered

through human reason or conscience, plays an essential part in the identification and understanding of all the positive laws that generate moral obligations.

For some natural lawyers, this means that there can be no such thing as an 'unjust law,' so that if a positive law is morally unjust, then it cannot be a genuine law. However, John Finnis, currently the most influential natural law theorist in the classical tradition, rejects this analysis and sees natural law theory as presenting the role of positive law as a social institution that is necessary to give concrete application to the general moral principles that derive from the idea of human flourishing (Finnis, 1980/2011). Other contemporary natural law theorists make the stronger contention that, while individual laws may or may not be just, the idea of law, as distinct from the content of particular laws, is necessarily connected to justice, for law as an institution must not only claim to be just but also seek to be so, otherwise it could not be a legal system (Beyleveld and Brownsword, 1989; Raz, 2009: p. 4).

A crucial matter of interpretation at this point is whether the role of morality in both classical approaches is an exemplar of critical or of positive morality. In the case of legal positivism, there seems no doubt that the utilitarian morality used to assess legal systems is a critical morality. Indeed, Bentham calls this activity 'censorial jurisprudence.' In the case of natural law theory, the situation is less clear because the moral methodology of natural law theory concerns 'nature' this includes human behavior, particularly normative behavior, which is conduct that involves the control of natural appetites in a way that makes social life and individual spiritual development possible. In effect, this means that universal positive morality has a key evidential role in determining what is in accordance with right reason. However, classical natural law theory often presents as a critical morality, and modern natural law theories, which are standardly uncoupled from theological backgrounds or moral deference to what is in accordance with nature, hold that it is a critical morality, albeit one that is central to the very idea of law (Simmonds, 2007).

Thus, some current understandings of both the natural law and legal positivist traditions appear to share the view that their role is to provide a moral critique of positive law, the remaining differences between them being about what sort of critical morality should feature in that critique. Legal positivists tend to adopt a utilitarian morality, and natural lawyers hold to a more communitarian morality focusing on human nature and the common good. A recent example of this latter position is the program for revitalization put forward by Robin West, who, drawing on the work of John Finnis and Martha Nussbaum (2001), calls for the adoption of a liberal and progressive natural law theory that recognizes that "Our happiness, our well-being, our capabilities, our flourishing, our functioning, and even our much maligned pleasure, cannot be reduced to the sum of the utilities we seek" (West, 2011: p. 44). Another question remains, however: how is it that the preferred critical morality does or should find its way into the content and practice of positive law.

All this does not mean, however, that modern versions of the classical theories are simply presenting competing schemes of critical morality with the utilitarianism of legal positivism pitted against secular versions of naturalism. Indeed, both approaches have outgrown their original moral affiliations. Exactly what is now at stake between the two approaches is best explained by reference to more recent controversies over what is called 'the separability thesis.'

Separation and Separability

With the introduction of Hart's rule of recognition, the classical debate about law and morality has come to be formulated in terms of what is known as the 'separability thesis.' This goes beyond the positivist idea that real law can always be traced to social sources, (e.g., commands, enactments, or court decisions) and takes us to the idea of a 'rule of recognition' that gives the official criteria for identifying those social sources by which norms and principles are 'recognized' as legal norms or principles.

The separability thesis, as formulated by Jules Coleman (1989), is the view that a rule of recognition need not include moral criteria, such as justice or utility, but can consist entirely of empirical identifiables, such as human commands, court decisions, or legislative enactments. Because this thesis possibility is compatible with all actual rules of recognition, including moral criteria as a matter of empirical fact, Coleman's minimalist form of legal positivism is compatible with what is called 'soft' (Hart, 1994, Postscript) or sometimes 'inclusive' legal positivism (Waluchow, 1994). This contrasts with the 'hard' or 'exclusive' variety of legal positivism, which admits only nonmoral criteria into anything that counts as criteria of validity in a rule of recognition. According to soft or inclusive legal positivism, it is a contingent matter how far a legal system officially incorporates moral considerations into its criteria of legal validity through the social practices that constitute lawmaking. This view is still regarded as positivist because it requires that morality enters into legal reasoning through an empirically identifiable rule of recognition that licenses this practice. Hard positivism also accepts that moral criteria are, and indeed ought to be, utilized in lawmaking through legislation, but insists that what makes that morality or anything else into law is its adoption by lawmakers authorized by a rule of recognition. The difference between soft and hard legal positivism thus comes down to whether or not legal officials, such as judges, are, in their jurisdiction, authorized to make law through the direct application of their own moral views. Important empirical issues then arise for social science as to whose morality finds its way into the law, and how far the morality within the law stems from legislative rather than judicial activity.

Viewed as a descriptive thesis, inclusive legal positivism is clearly superior as all legal systems enable some role for moral judgments in the adjudication that determines legal validity. Common-law systems permit judicial development of law in the light of broad moral principles, albeit subject to the theoretical right of legislators to amend the common law. Many legal systems incorporate a bill of rights with general moral principles that courts are obliged to interpret, something that includes some reference to the moral preferences of those who give concrete application to such vague moral standards as equality and freedom. This seems to accord more with a natural law approach, since it involves using what are perceived to be fundamental moral rights at a point where law and morality become inseparable (Dworkin, 1996).

The moral criteria in actual rules of recognition often serve to incorporate within law the customary or positive morality, or some identified aspect of it, rather than the personal or critical moral judgments of judges and other officials. This may be regarded as upholding what are regarded as reasonable expectations. However, since in modern complex societies there is seldom any indisputable agreed content to positive morality on almost any topic, in effect inclusive legal positivism legitimates the use of critical moral judgment in the process of adjudication, which in turn facilitates the impact of the political and self-interested views of judges. For the positivist to claim that there is a conceivable society in which this is the case is a relatively minor conceptual point to make about law and morality.

The retreat into soft, or inclusive, positivism is therefore explicable, especially as legal positivists have historically placed great weight on providing a descriptive theory of all developed legal systems. However, legal positivism can be understood in a more normative light as an approach that is seeking to establish a conception of a 'good' system of law, thus turning the separability thesis into a prescriptive claim that rules of recognition should not include moral terms that cannot be put into practice requiring moral judgments about what they require by way of legally required conduct. Hard, or exclusive, positivism may be viewed as a more or less disguised prescription that is recommending that a rule of recognition ought not to include moral criteria and that all positive law should be framed in such a way that it can be understood and applied without recourse to the moral opinions of citizens or judges (MacCormick, 1989; Campbell, 1996: p. 72; Waldron, 2001).

This morally normative approach to the separation of law and morality is associated with a complex political philosophy that points to the function of authoritative rules in human society (Raz, 1979; Shapiro, 2011). Social cooperation and coordination, dispute resolution, planning, and maintenance of public order, it is alleged, require rules and processes that are followed and enforced according to an authorized formulation that leaves as little room as possible for doubt and controversy as to their content and practical implications. These analyses often relate to the utility of rules as social devices that give effect to conventions that make it possible for individuals and groups to anticipate each other's conduct with considerable reliance in complex social circumstances. The utility of these rules and associated norms is relatively independent of their precise content since what matters is that there is some generally recognized and readily understandable convention, even if its clear content is morally controversial (Postema, 1982).

While these considerations explain why it may be socially important for some laws to be clear and applicable without controversy as to their meaning, the moral usefulness of other laws, such as those forbidding clearly harmful conduct, is more dependent on a common understanding and its moral acceptability on the part of those to whom it is addressed. Even here, however, there is a role for official determination as to what conduct is punishable and what is merely open to social criticism. This is because there can be controversy as to the precise nature of the harm in question. The need for definitive rulings derives from the existence of moral disagreement, for in fact there is rarely any one coherent system of positive morality, particularly where a culture encourages critical approaches and individuals have some autonomy in making up their own minds as to what they believe ought to be punished or recompensed. The political need to reach a working agreement on matters with the potential to generate conflict and disorder gives a morally significant role for the positivist idea of law as a system of mandatory rules whose authority depends on their adoption as law and not on their specific content.

Other moral arguments in favor of such a formalist conception of law are presented by Lon Fuller (Fuller, 1964), who demonstrates the connection between what he calls the governance of rules – rules that are clear, promulgated, prospective, practicable, and stable – and a society in which there is respect for autonomous moral agents who can reasonably be held responsible for their actions and accept that their obligations are part of a mutually beneficial political system. Fuller calls his theory a procedural form of natural law. Its formal requirements of 'good' law are part of a morality of mutual respect. He goes further than this and argues that a government that abides by the requirements of formally good procedural law will thereby tend to produce good substantive law. In a classic exchange with Fuller, H.L.A. Hart pointed out that Fuller's procedural natural law is 'compatible with great iniquity' (Hart, 1957). However, the general thesis that there are moral reasons for government in accordance with rules has clear affinities with prescriptive legal positivism, and it can be argued that Hart himself includes elements of this prescriptive approach (in a largely descriptive analysis) when he describes the benefits that arise for social order when the informal 'law' of simple societies is supplemented by 'secondary rules' that enable official determinations about what the primary rules of a society are and who has the right to interpret and change them (Hart, 1961: pp. 77–96).

Enforcing Morality

Analyses of the relationship between law and morality are often framed in terms of whether or not law should enforce morality, usually the dominant positive morality of a society. It is generally accepted that there is a sense in which morality cannot be enforced because truly moral conduct involves the willing acceptance of moral obligations. Thus, Immanuel Kant defined moral goodness as doing what is morally right because it is morally right and not, for instance, because the law requires it of us.

Further, it is rarely argued, as a matter of critical morality, that the law should require conformity to all and only conduct that is morally correct. Natural law theory tends to identify justice and the common good as that part of morality that is binding on human beings in a way that makes it appropriately enforced through positive law. The theory also requires governments to fill out positive law with lots of morally neutral ways that are required to give specific application to other natural law values. Classical legal positivists apply the general principle of utility to all moral questions, but point out that choosing a law is not making the same moral choice as choosing how to behave as an individual. Law involves sanctions and restricts liberty of action, which human beings enjoy and use to achieve their own happiness. Therefore, there are always moral reasons against enacting a law. This approach was developed by John Stuart Mill (1910) into the famous harm

principle according to which only harm-producing conduct may be prohibited, and then only when the negative effects of enforcing conduct are outweighed by the positive effects in reducing harm. He confined this harm to that which is caused to others, an antipaternalist positions that is a dominant strand in all liberal political philosophies.

The highly individualistic approach encouraged by the utilitarian moral theory is countered by various forms of communitarian analyses of society, including natural law theory. This conflict is well illustrated in a classic debate between Lord Devlin (1965) and H.L.A. Hart (1963) on the matter of decriminalizing homosexual conduct between consenting adults. Devlin argues that the basic institutions of a society, such as marriage, have to be protected in the interests of that society as a whole; therefore, it is treasonable to allow forms of conduct that undermine such foundational social institutions, and it follows that society has a right to enforce particular types of sexual organization. This was a response to Hart, who takes the view that sexual conduct of this sort is a purely private matter so that any disgust that the majority might feel toward such conduct has no bearing on the moral question of whether or not homosexual conduct should be a criminal offense.

Debates of this sort may be approached in a number of ways. Sociologically, Devlin's communitarian point can be investigated to see whether it is the case that the preservation of monogamous heterosexual marriage, as the dominant social institution with respect to sex and reproduction, does require prohibition of incompatible forms of sexual conduct. Other arguments center on the concept of toleration, the idea that it is morally desirable to allow people to follow out their chosen way of life even when you think that they are wrong to behave in this way. The arguments for toleration are complex (Mendus, 1988). For some people, toleration is an acknowledgment of the significance of moral goodness. For others, it is simply morally right not to interfere with the conduct of others when such conduct does not harm our interests. For others, it is a response to moral disagreement and the difficulty of reaching an objective decision about which competing form of life is morally superior. A more extreme form of this position is that choices between competing ways of life cannot themselves be moral choices, since all moral choices have to assume values that arise from a particular set of cultural circumstances. All these approaches either permit or encourage a pluralism of positive moralities that make it impossible for the law to enforce morality without a choice being made as to whose morality or which morality is to be enforced.

Responses to the problem of choosing which morality to enforce can be broken down into three categories:

1. Those that aspire to a morally neutral principle as a guide to which part of morality to translate into law. Mill's principle that only acts that cause harm to others may be criminalized has been presented in this light, but it is clearly part of the utilitarian position that causing pain to another is an immoral act, which is not a morally neutral position, especially when it excludes other possible grounds for moral disapproval.
2. Those that distinguish between private and public spheres, arguing that the morality that pertains to law is public morality, that is, morality about political and state relationships, while private has to do with the individual and her freely chosen associates. This shifts the argument to what is public and what is private, a principal battlefield being the sphere of family relationships. Feminists have powerfully pointed out that to classify the family as private is to leave this male-dominant sphere unregulated. Ultimately, 'private' comes to mean simply those matters that the law ought not to deal with, making it uselessly circular as a principle for determining the proper sphere of the law.
3. Those that seek to identify a certain number of basic rights that, while not morally neutral, allow for an extensive range of different forms of life to flourish as each individual and group pursues its own conception of what is good. This is the position developed by John Rawls in *A Theory of Justice* (1971) and enunciated by Ronald Dworkin in a series of works that represent the most sustained recent attempts to identify a morality that the law ought to enforce (1977, 1985). Dworkin's position is that judges should, within the confines of established legal decisions, exercise their own moral judgment in determining a point of law; this is a qualified form of natural law that requires judges to invalidate positive laws that are incompatible with natural law in the form of fundamental rights, which makes it a necessary condition of legality that a mandatory rule does not violate a human right, but still leaving considerable scope for laws whose content is determined solely by elected legislatures, often on the grounds that the right to self-determination is itself a human right (Waldron, 2006).

This third position arises from the emergence of the idea that there are human rights, a form of morally important universal rights, which establish legal limits to the powers of governments on the basis of certain fundamental moral requirements. Accepting that there are such moral limits and the desirability of enforcing them legally does not, however, enable us to achieve actual agreement as to what these rights require in practice. There is therefore a moral question as to who determines the specific content and the relative priority of the rights that are to be legally enforced. This takes us to constitutional questions that are crucial to contemporary arguments about law and morality.

If we believe that there is a moral method that enables us to identify human rights in an objective manner, then it is reasonable to enact by plebiscite or legislative process a general bill of rights that is then filled out and applied by courts in their determination as to whether the law produced by elected assemblies does or does not violate these rights. This may be regarded as a qualified natural law position since it assumes an identifiable lawlike morality particularly related to fundamental justice. It therefore makes sense to have rules of recognition that list these rights and leave it to the critical morality of the judiciaries to give them concrete application.

If we are skeptical about our capacity to solve disagreements about fundamental rights in this way and give high priority to the right of self-determination, then we can draw on a prescriptive form of hard legal positivism and suggest that it is for elected assemblies and popular plebiscites to determine the content of all law in light of the critical moral judgments of the people as a whole, and the duty of the courts to apply such laws

without reference to their own moral views, in which case the morality that is in the law is determined independently of and externally to the process of law application.

This might require legislatures to turn to the values that they believe give the best moral interpretation to the human rights identified for implementation through formally good positive rules that can be applied without further moral adjudication to concrete social circumstances. A natural law alternative posits that it is humanly possible to reach morally correct answers to such disputed matters and that judiciaries are in the best position to discover the correct moral strength and scope to give to the human rights whose interpretation is entrusted to them. Legal positivists regard this as a confusion of morality and law, but this need not be the case if the aim is the division of legislative power between legislatures and courts rather than a strict institutional separation of legislative and adjudicative powers. The choice between these two perspectives is one of the most important contemporary moral choices regarding the best way of conceptualizing and institutionalizing distinctions between morality and law.

See also: Law and Morality: A Continental European Perspective.

Bibliography

Aquinas, T., 1966. Summa Theologica (1269). Blackfriars, London.
Austin, J., 1954. In: Hart, H.L.A. (Ed.), The Province of Jurisprudence Determined (1832). Weidenfeld & Nicolson, London.
Baier, K., 1958. The Moral Point of View. Cornell University Press, Ithaca, New York.
Bentham, J., 1970. In: Burns, J.H., Hart, H.L.A. (Eds.), An Introduction to the Principles of Morals and Legislation. The Athlone Press, London.
Beyleveld, D., Brownsword, R., 1989. Law as Moral Judgment. Sweet & Maxwell, London.
Campbell, T., 1996. The Legal Theory of Ethical Positivism. Dartmouth, Aldershot.
Coleman, J., 1989. Negative and positive positivism. Journal of Legal Studies 11, 139–164.
Devlin, P., 1965. The Enforcement of Morals. Oxford University Press, Oxford.
Dworkin, R., 1977. Taking Rights Seriously. Duckworth, London.
Dworkin, R., 1985. A Matter of Principle. Harvard University Press, Cambridge.
Dworkin, R., 1996. Freedom's Law: The Moral Reading of the American Constitution. Harvard University Press, Cambridge, MA.
Finnis, J.M., 1980. Natural Law and Natural Rights. Clarendon Press, Oxford.
Finnis, J.M., 2011. Natural Law and Natural Rights, second ed. Clarendon Press, Oxford.
Fuller, L.L., 1964. The Morality of Law. Yale University Press, New Haven.
Fuller, L.L., 1958. Positivism and the separation of law and morals: a reply to professor Hart. Harvard Law Review 71, 597–639.
Hart, H.L.A., 1957. Positivism and the separation of law and morals. Harvard Law Review 71, 593–629.
Hart, H.L.A., 1961. The Concept of Law. Clarendon Press, Oxford (second ed. 1994).
Hart, H.L.A., 1963. Law, Liberty and Morality. Oxford University Press, Oxford.
Hart, H.L.A., 1994. The Concept of Law, second ed. Clarendon Press, Oxford.
MacCormick, N., 1989. The ethics of legalism. Ratio Juris 2, 184–193.
MacCormick, N., 2007. Institutions of Law: An Essay in Legal Theory. Oxford University Press, Oxford.
Mendus, S., 1988. Justifying Toleration: Conceptual and Historical Perspectives. Cambridge University Press, Cambridge.
Mill, J.S., 1910. Utilitarianism, Liberty and Representative Government. Dent, London.
Nussbaum, M.C., 2001. Women and Human Development. Cambridge University Press, Cambridge.
Postema, G.J., 1982. Coordination and convention at the foundations of law. Journal of Legal Studies XI, 165–203.
Rawls, J., 1974. A Theory of Justice. Clarendon Press, Oxford.
Raz, J., 1979. The Authority of Law. Clarendon Press, Oxford.
Raz, J., 2009. Between Authority and Interpretation. Oxford University Press, Oxford.
Shapiro, S., 2011. Legality. Harvard University Press, Cambridge/London.
Simmonds, N., 2007. Law as a Moral Idea. Oxford University Press, Oxford.
Waldron, 2001. Normative (or ethical) positivism. In: Coleman, J. (Ed.), Hart's Postscript. Oxford University Press, Oxford.
Waldron, J., 2006. The core case against judicial review. Yale Law Journal 115, 1346.
Waluchow, W.L., 1994. Inclusive Legal Positivism. Clarendon Press, Oxford.
West, R., 2011. Normative Jurisprudence. Cambridge University Press, Cambridge.

Law and Neuroscience

Owen D Jones, New York Alumni Chancellor's Professor of Law & Professor of Biological Sciences, Vanderbilt University, Nashville, TN, USA; and Director, MacArthur Foundation Research Network on Law and Neuroscience
Matthew Ginther, Candidate in Law and Neuroscience, Vanderbilt University, Nashville, TN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses how the intersection of perennial legal questions and new neuroscientific advances has fueled the emergence of a new field: law and neuroscience. It provides an overview of issues, discussing both the promise and the limitations.

Introduction

In barely a decade, a new field of law and neuroscience (sometimes called 'neurolaw') has emerged. Its quick growth is evident not only from the increasing frequency with which attorneys offer to introduce neuroscientific evidence in court, but also from the sharp rise in conceptual and empirical scholarship, the proliferation of neurolaw conferences, the founding of multiple international neurolaw societies, the increase in neurolaw course offerings (in law schools and beyond), the creation of joint degree programs, and the investments by public and private entities that have fueled interdisciplinary collaborations and research (Jones et al., 2013a,b).

As with any new field, there is wheat and chaff; a considerable proportion of the early efforts has involved developing ways to identify, evaluate, and separate the two. Given the technical nature of neuroscientific capabilities, a legal system needs guidance on how to chart a course that maximizes the promise of neuroscientific insights and minimizes the potential perils of incorporating them. Some of the best work in law and neuroscience to date (e.g., Morse and Roskies, 2013) partners experts from both fields to move forward in a cautious but deliberate – instead of purely reactive – way (Jones, 2013a).

Three cases, from three different countries, illustrate some of the many contexts in which neuroscientific evidence has – for better or for worse – influenced legal proceedings. In 2008, prosecutors in India secured the conviction of a woman for the murder of her fiancé (*Maharashtra v. Sharma*, 2008). The prosecution presented key evidence – data collected from the defendant's brain – that supposedly corroborated her suspected involvement in the murder. In 2009, an Italian court reduced the sentence of a woman convicted of murdering her sister (and attempting to murder her parents) from life in prison to 20 years in prison. The court cited, as relevant to this reduction, structural brain abnormalities as well as low levels of an enzyme responsible for maintaining proper levels of various neurotransmitters in the brain, a combination that is thought to increase violent tendencies (Feresin, 2011). In 2012, the United States Supreme Court concluded that it was unconstitutional for juvenile offenders to be sentenced to mandatory life in prison without the possibility of parole (*Miller v. Alabama*, 2012). In explaining its decision, the Court cited (alongside other lines of evidence) neuroscientific findings regarding the differences between adolescent and adult brains.

Cases, such as these, and the growth of scholarship and public interest that surround neurolaw issues generally, suggest that neuroscience will (again for better or for worse) ultimately play an increasing role in legal proceedings. What accounts for this? It likely results from the combination of two distinct factors that have rapidly converged over the past decade or so.

First, the legal system has long struggled to answer important questions – questions about what goes on in people's heads, and why – that society demands be answered. For example: Was the defendant responsible? What did the defendant intend? How competent is the individual to stand trial? Is this witness lying? What capacity did the defendant have to act differently? And will the person be a repeat offender? Correctly answering each of these questions is essential to the just enforcement of the law. Yet to various degrees answering them requires legal decision makers to act as amateur mind readers – of questionable accuracy.

Second, robust advancements in cognitive neuroscience, driven in large part by technological developments related to brain imaging, appear to offer some promise in helping to answer these perennial questions (Posner and Raichle, 1994). Perhaps, the thinking goes, structural or functional brain measures can aid judgments about whether someone is lying, responsible, and likely to recidivate. Perhaps, some hope, technology can increase the accuracy of our conclusions by providing objective, clear, and quantifiable data.

What neuroscience can and cannot provide to the legal system is, of course, a matter of debate across multiple fora (Wagner, 2010; Weisberg et al., 2008; Morse, 2011; Greene and Cohen, 2004; Morse and Roskies, 2013, Freeman and Goodenough, 2009; Jones et al., 2014; Greely and Illes, 2007). Which is as it should be – because the potential for misunderstanding or misuse is (as with any technological or scientific advancements) concerning and real. The purpose of this article is therefore to introduce some of the potential promise of neurolaw, as well as to raise some cautions and limitations, lest hopes outpace capabilities.

Technological Capacities

Although it is by no means the only neuroscientific technique of importance, functional magnetic resonance imaging (fMRI) – which provides unparalleled ability to link brain function with brain structure – is attracting increasing attention in legal arenas. Introduced in the early 1990s, fMRI takes advantage of

physical properties of oxygenated and deoxygenated blood to measure changes in blood flow to brain regions. This change in blood flow is linked to changes in neural activity. By analyzing the change in blood flow across the brain and across the experiment, neuroscientists can create compelling evidence linking perception and behavior to neural activity in specific regions of the brain (Friston et al., 1998). This has allowed investigators to create cognitive models that are of interest to the law, such as the brain systems that support social decision making, deception, and cognitive capacity (Rilling and Sanfey, 2011; Kozel et al., 2005; Bunge et al., 2002; MacDonald, 2000).

In addition to linking neural activity in a brain region to cognitive and perceptual processes, advancements in fMRI have provided investigators with the ability to study the functional connectivity between brain regions and to decode patterns in neural activity that can detect, describe, and predict perception and cognition in human subjects. For example, brain connectivity studies have been used to reveal the limited influence of top-down cognitive control systems in adolescents (Luna et al., 2010) and pattern analysis studies have revealed patterns of brain activity that can reliably indicate that subjects recollect stimuli presented to them (Johnson et al., 2009).

fMRI is not the only technology used for legally relevant neuroscience research. Positron emission tomography (PET), diffusion tensor imaging (DTI), electroencephalography (EEG), and repetitive transcranial magnetic stimulation (rTMS) have all been used to inform legal decisions or questions in a variety of litigation or research contexts. Each of these technologies provides the investigator with unique sets of tools (the details of which are beyond the scope of this article) that can observe or manipulate the human brain.

Caveats and Cautions of Brain Imaging

However remarkable, neuroscientific technologies and methods – like all techniques and methods – have their limitations. And, given the stakes involved, it is particularly important that legal systems have informed and clear-eyed understandings of those limitations. Here we divide the major concerns regarding functional imaging evidence into two categories: those involving mistaken facts and those involving mistaken inferences.

In the realm of mistaken facts, it is easy to misunderstand what brain images actually represent (Jones et al., 2009). For example, fMRI brain images, which appear to show brain regions as 'lit up,' are not in fact the brain scanner's output of where neural activity is located. Indeed, fMRI provides no direct measure of neural activity at all. Instead, the different colors on an image indicate different results of statistical analyses performed on minute changes in blood flow in a set of brain regions, across many trials of an experiment (and, usually, across many subjects). On one hand, there are many reasons to believe that increased blood flow to neurons is strongly associated with increased neural activity. Changes within brain tissue in the demand for blood are thought to be a reliable proxy for changes in the activity of neurons within the brain. On the other hand, the exact relationship of blood flow and neural activity is poorly understood.

The most prevalent of mistaken inferences about fMRI reflects the assumption that if a region shows activation with a particular task or behavior then that region is therefore causing that particular task or behavior. Yet, as in other contexts, correlation is not causation. In fact, fMRI is mainly a technique of discovering correlations, which must still be further investigated before claims of causality can be made legitimately (Logothetis, 2008).

Another important example of mistaken inference follows from overlooking the crucial distinction between group and individual data. Most brain research is concerned not with what is true about any given individual, but with what is true about the average individual within the studied group. Group averaged data do not necessarily provide an accurate description of what is occurring in each individual. To the extent that law is often engaged in adjudicating the futures of individuals before the court, one must exercise great care and caution when trying to reason one's way to an individual's brain from a general finding grounded in the brain data of groups (Faigman et al., 2014). For example, if a group-level study demonstrated that those telling lies have increased activity in a particular brain region, we could not say with confidence that any given individual will also demonstrate increased activity in that same region when telling a lie.

The combination of mistakes of fact and mistakes of inference yields a quite reasonable concern that jurors and others in the legal system may afford more weight to neuroscientific evidence – particularly vivid, colorized images – than the underlying science warrants. Some early studies seemed to indicate that jurors were likely to be prejudiced by such evidence (Gurley and Marcus, 2008; Weisberg et al., 2008). Later studies have painted a more complicated picture, in which jurors are generally more discerning than originally thought (Saks et al., 2014), though they are still likely, in some contexts, to give undue weight to brain images (Saks et al., 2014; Schweitzer et al., 2011).

The main takeaway is that in neuroscientific evidence the law encounters a highly technical domain, the chief products of which – however vivid – must first be contextualized within a series of interrelated caveats and cautions, and then be evaluated – on a case-by-case basis – within the specific legal context in which the science may, or may not, directly aid the decisions at hand.

Ways in Which Neuroscience May Aid Law

The 'law,' of course, is not one thing; it is rather a vast web of institutions and rules. The bulk of these are specific to defined jurisdictions in which they operate. In the United States, for example, most criminal law is a matter of state rather than federal jurisdiction. And the state laws can vary (and often do vary) as much as do the states themselves.

Because of this, it is difficult to speak generally about the role and influence of neuroscience in the law. Nonetheless, one of us has proposed that there are seven main (and sometimes overlapping) ways that neuroscience can aid law (Jones, 2013b):

1. *Buttressing* – by increasing juror confidence in a conclusion to which other, nonneuroscientific evidence already

independently points (such as in the context of 'diminished capacity' determinations);
2. *Challenging* – by calling into question or contradicting either other evidence in a case or a relevant legal assumption (such as those reflected in certain evidentiary rules);
3. *Detecting* – by identifying the existence of legally relevant facts (such as injuries, lies, or pain);
4. *Sorting* – by separating people into useful categories (such as those most likely to respond to drug rehabilitation);
5. *Intervening* – by providing new methods to achieve legal goals (such as through pharmacological interventions that would help to reduce recidivism);
6. *Explaining* – by illuminating decision pathways with information that may lead to more informed and less biased decisions (such as in the context of third-party punishment decisions); and
7. *Predicting* – by improving law's ability to estimate probabilities of future behavior (such as future violence).

For purposes here, however, a rough and different cut into two main functions of neuroscience in legal arenas will serve. Consequently, the Section Neurolaw and Individuals considers ways in which neuroscience can help to discover facts that are legally relevant (such as: How injured is this person's brain?). The Section Neurolaw and Groups considers ways in which neuroscience can inform broader policy making (such as: To what extent are adolescents similar to or different from adults?).

Neurolaw and Individuals

Legal systems – in their courts, for instance – are often engaged in making decisions about the rights, transgressions, and lives of individuals. Often the content and nature of those decisions depend intimately on which version of hotly contested facts is the more accurate one.

Multiple sources of evidence may bear on the process of deciding what – at least for the purposes of law – those facts really are. For example, when a key legal issue is whether a defendant is legally insane, a court may hear and weigh conflicting evidence from clinical psychologists, psychiatrists, employers, neighbors, and the like.

Although the behavioral evidence is of course crucial in such contexts, neuroscientific evidence (of considerable brain tissue degradation, for example) could help the legal system choose which version of contested facts is most likely. This section illustrates a number of contexts in which the main value to law of neuroscientific evidence is in helping it to determine the facts, frequently in the context of litigation.

Brain Injury and Pain

The intersection of neuroscience and law arguably originated with the use of neuroscience to better define and detect brain injury. This occurred primarily in the realm of personal injury (or 'tort') law, which is a type of civil action. A successful personal injury claim must typically demonstrate: (1) actual injury; and (2) a causal link between the defendant's actions and the claimant's injuries.

Due to the dearth of knowledge about brain function, proving a causal link between the defendant and the injury, or even demonstrating an actual injury, was impossible for many actual and potential claimants. As an example, imagine two different car accidents caused by equally negligent actors. One accident results in broken legs, but no trauma to the head, while the other results in massive trauma to the brain, without breaking any bones. Despite the likelihood that the latter will be lifelong and life changing, it is significantly more difficult to establish the injury, link it to the act, and obtain a judgment that is commensurate with the injury (Zitter, 1997; Luce, 2012).

Because neuroscience provides the tools and techniques to detect and classify brain injury and estimate a prognosis, it has been increasingly called upon to inform legal claims involving brain injuries. Advances in neuroscience currently appear to provide better specificity and sensitivity when diagnosing and assessing these injuries (Lewine et al., 2007).

This is important because traumatic brain injuries (TBI) are quite common. The Centers for Disease Control estimates that TBI contributes to one third of all injury-related deaths in the United States and that, in that country alone, 1.7 million people sustain a TBI each year, 52 000 of whom die shortly thereafter. And lawsuits involving TBI have become increasingly prominent. For example, many former players of American-style football have sued over brain injuries allegedly sustained across their careers. Corroborating evidence for the causes of the cognitive and affective impairments they allege, which is hard to demonstrate while still alive, has come from investigation of the brains of deceased players, which often exhibit chronic traumatic encephalopathy (CTE) (McKee et al., 2009).

In addition to providing for the improved detection and classification of brain injury, neuroscience also provides evidence that could potentially alter how we define brain injury in the first place. In large part, the law recognizes pain as an injury. However, the subjective nature of pain has influenced how the law treats pain. New neuroscientific evidence may come to challenge this in at least two ways.

In the first instance, legal systems have struggled with the fact that pain is a subjective experience, difficult to prove, and possible to fake (Pryor, 1991). A common way to circumvent this problem is to require indicia – such as trauma or disease – beyond the subjective reporting of pain. This approach is only satisfactory, however, when such objective indicia are available and consistent with the magnitude of subjective pain. In cases of chronic pain, such as fibromyalgia, it is very difficult for those suffering to establish that they have the subjective experience of chronic pain. And, even though 10% of the population suffers from different forms of chronic pain, those who bring claims related to their chronic pain are frequently thought to be malingerers or outright fakers. Brain imagers are attempting to develop methods to more reliably observe the subjective experience of pain (Rainville et al., 1997). As these methods improve, they may help to identify pain when it is present and to raise doubts when it is not.

In the second instance, brain imaging data are undermining the law's long-standing divide between physical and psychological pain. Jurisdictions vary, but it is generally the case that remedies for psychological pain are not available without demonstrating a physical harm. This rule likely arose to serve

an evidentiary purpose. But neuroscience is calling into question one of the basic premises behind the rule. Specifically, recent neuroimaging results reveal that psychological pain engages the same brain regions as physical pain (Eisenberger, 2012). In addition, to the extent that certain brain disorders (such as posttraumatic stress disorder (PTSD) and other anxiety disorders) are classified as psychological, brain imaging research may demonstrate the physical basis for the disorders (Shen, 2013). This can frame a discussion wherein psychological and physical traumas are potentially seen as one and the same.

Cognitive Capacity and Intent

In most legal systems, criminal responsibility requires both a bad act and a culpable state of mind. Neither alone is sufficient. For example, a person judged to be insane is not ordinarily treated the same as a noninsane person who committed the identical act. The substance, procedures, and burdens of using insanity to escape criminal culpability vary widely across jurisdictions (Simon and Ahn-Redding, 2008). Despite the variability, there are three primary legal mechanisms by which neuroscience can influence questions regarding the culpability of a defendant who has committed a proscribed act.

One mechanism is for the defendant to claim that his condition satisfies his jurisdiction's insanity defense. Insanity defenses are created by courts or by legislatures and allow for defendants to argue that they are excused from full criminal responsibility on the basis of mental disease or defect. The type or extent of the disease or defect required for an insanity defense varies. A majority of US jurisdictions use the 'M'Naghten rule.' The rule, which originated in a nineteenth-century English case (and is named after the defendant in that case), requires a defendant to establish that he lacked the ability to understand the nature and quality of his actions. Because the conscious experience of understanding the appropriateness of one's own actions is a complex and poorly understood concept, neuroscience has been of limited use in operationalizing such a rule. However, as neuroscience begins to uncover the mechanisms supporting human consciousness, its contribution toward these deliberations may increase. A sizable minority of US jurisdictions and a number of other countries use a test focused on the question of the defendant's volitional control as opposed to his cognitive capacity (as is the case with the M'Naghten rule). The neurological disorder of frontal lobe dysfunction (FLD) is a useful example of the difference between these two rules.

Individuals with FLD suffer from severe deficiencies in impulse control and therefore cannot conform their behavior to the expectations of the law and society as easily as do non-FLD individuals (Brower and Price, 2001). Nonetheless, these individuals maintain the ability to understand the criminal nature of their actions because the area of cortex responsible for this cognitive capacity is seemingly preserved in cases of FLD. As a result, individuals with FLD do not ordinarily meet the requirements of legal insanity under the M'Naghten rule but – because their ability to conform their behavior is deficient – they may be successful in making a claim of insanity in those jurisdictions that use volitional tests. Even then, however, neuroscience may ultimately be of limited use, because it has yet to provide a neurobiological model of volition that is consistent with its legal understanding. In other words, volition as the law understands and defines it does not map comfortably onto an identical concept in neuroscience. This explains both the criticism surrounding the use of neuroscience to establish legal insanity and the limited impact of neuroscience on the success of such claims.

Despite the limited influence of neuroscience in assisting claims of legal insanity, neuroscience continues to be used to challenge criminal responsibility by challenging the prosecution's prima facie case. Generally, to establish its prima facie case, the prosecution must establish a level of culpable intent accompanying the acts, often referred to as *mens rea*. While the gradations of *mens rea* vary across jurisdictions, the most severe criminal statutes generally demand that the defendant acted with either general or specific intent with regards to the resulting harm. With increasing frequency defendants will argue that brain abnormalities, visualized through brain imaging, provide evidence that the defendant was incapable of formulating the required intent and therefore should be found not guilty of the charges (Edersheim et al., 2012).

Finally, as of now defense attorneys have found the use of neuroscience to be most useful at the sentencing phase of a trial. In many jurisdictions and for many crimes, trials will take place in two parts. In the first part fact finders evaluate whether the defendant is guilty of the charges brought against him. In the second part, if the defendant is found guilty, a judge or a jury will determine the appropriate sentence. For a variety of evidentiary reasons, the bar for admissibility of evidence is set much lower for this latter stage. Taking advantage of this, some attorneys have started to introduce neuroscientific evidence in an attempt to argue that the defendant should be given a more lenient sentence. While empirical studies have yet to demonstrate a significant effect of the use of neuroscience at this stage of the trial, anecdotal evidence indicates that it has already influenced sentencing in some cases (Jones and Shen, 2012).

Lie Detection and Recognition

Prospects for brain-based lie detection have received great attention – both because the idea of reliably detecting lies is inherently alluring and because criminal defense lawyers have, on at least several (so far unsuccessful) occasions, tried to introduce fMRI-based lie detection evidence into trials. Prior to the use of brain scans to uncover deception, the most recent efforts to introduce physiological measures as evidence of truth or deception have involved the so-called 'polygraph' – which measures peripheral indicators of cognitive states such as blood pressure, heart rate, and skin conductivity. The reliability of these measures as indicators of lies is a source of much debate. Although the polygraph is still used widely in interviews and interrogations, US courts routinely exclude polygraph evidence (in light of, among other things, a 2003 report of the United States National Academy of Sciences, which labeled it unreliable and biased).

fMRI techniques for identifying lies seek to bypass indirect physiological indicators, which can be skewed by countermeasures or nonindicative anxiety, and find direct neural signatures of the cognitive processes involved in deception. Proponents argue that this would reduce the risk of both false

positives and false negatives. Although some data suggest that these assessments can have accuracy rates approximating 90%, other data suggest a far lower accuracy, approaching or no better than chance. Beyond this, several robust critiques note problems with ecological validity, the 'real-worldness' of the studies (e.g., Farah et al., 2014). For example, almost all studies to date involve individuals from limited demographics practicing relatively low-stakes deception without engaging countermeasures. Moreover, they are almost exclusively limited to 'instructed lie' experimental designs, in which subjects are explicitly told to try to deceive the researchers.

The extent to which instructed lies are meaningfully like 'real' lies, or like real lies under high stakes (such as the risk of imprisonment, for instance) is questionable. One respected neuroscientist, Anthony Wagner, has studied all published studies of fMRI-based lie detection and has concluded that, at the present time, the literature does not provide an accurate indicator of the sensitivity or specificity of fMRI-based lie detection (Wagner, 2010). The few courts to consider the matter have agreed with this assessment, and have ruled such evidence inadmissible, though it may become admissible in the future if methodological short-comings can be overcome (*U.S. v. Semrau*, 2010; *Wilson v. Corestaff Services*, 2010).

An area of active research closely related to deception concerns use of fMRI (or EEG) to detect actual autobiographical recognition of various stimuli, such as pictures of scenes or people, names, and the like. Early work in the field appears remarkably promising (Rissman et al., 2010). And although these techniques are also subject to various important limitations (such as vulnerability to countermeasures) the techniques have potential application in a number of legal contexts, should the limitations be overcome effectively.

Future Behavior: Assessing Dangerousness

Many jurisdictions across the globe depend on assessments of dangerousness when deciding sentence or considering the release of imprisoned criminals. Historically, these assessments have come in the form of expert reports (from clinical psychiatrists, for example). Actuarial assessments, however, are increasingly prevalent alternatives, widely thought to be more accurate (Grove et al., 2000).

Because risk of future dangerousness is, essentially, a question of the proclivities of the brain, many hope that neuroscience can and will improve the accuracy of these assessments. In this vein, one recent study compared 4-year recidivism rates of former prisoners who had been brain-scanned while in prison. It found that one neural assessment using fMRI provided a robust level of predictive power, even when controlling for other risk factors and psychological measures (Aharoni et al., 2013). Whether these assessments are viable in practical use, and the extent to which they enable individualized assessments, as distinct from group-averaged ones, remains to be seen.

Neurolaw and Groups

The prior section illustrated contexts in which neuroscience may provide information about individuals (such as a defendant) that can aid legal decision makers (such as courts) in making decisions relevant to those individuals alone. But there are also times, in contrast, when neuroscience may provide information about groups (such as adolescents, addicts, or eyewitnesses) that aid legal decision makers (such as legislators or courts) in making or refining broad policies. This section provides several examples.

Adolescent Justice

Children are everywhere recognized as not fully responsible for their behaviors. Adults are presumptively responsible. In between lie adolescents, for whom determining the appropriate level of responsibility has long vexed legal systems, as they balance a society's desire for individualized justice with both the difficulty of reliably assessing the relative maturity of individuals and the virtues of establishing easily administrable bright-line rules.

For nearly all the twentieth century, most western jurisdictions created specialized frameworks to process adolescent offenders. The primary characteristics of these regimes were an abundance of decision-maker discretion, a focus on rehabilitation and intervention, and very limited use of retributive punishment. The 1980s and 1990s, however, witnessed a global shift toward a less paternalistic and increasingly punitive justice system for adolescents (Junger-Tas and Decker, 2006). This shift is largely attributable to increasing crime rates, increasing fear within the population, and increasingly widespread beliefs that adolescents were morally responsible for their actions.

More recently, the retributive approach to juvenile justice has been losing favor with policy makers. That shift coincides with (and may or may not be partially attributable to) marked advances in understanding the adolescent brain. For example, a large body of neuroscientific evidence has clarified the ways in which the adolescent brain is categorically different from a fully developed brain. Structural analysis and fiber tract analysis have demonstrated that the human brain continues to develop and mature until around age 25. fMRI studies have revealed strong functional differences – between adolescent and adult brain systems – in delaying reward, impulse control, and emotional processing. Combining these findings with the behavioral findings, many neuroscientists conclude that adolescents do not have the full capacity for cognitive control over their behaviors and are not nearly as able as adults to modulate their socially and emotionally aroused inclinations with deliberative processes (Steinberg, 2008).

Of course, the legal implications of such evidence are a source of much debate. Some critics argue that the neuroscience is of little use, if not completely irrelevant, because it adds nothing to what we already knew from the social science research – not to mention lay observation of adolescents. We already knew, they note, that adolescents are irrational risk takers, susceptible to peer influence. Other critics question the utility of the information in deciding what to do with a specific adolescent – who might be much more or much less immature than average (Maroney, 2010; *Roper v. Simmons*, 2005 Scalia dissenting). In contrast, proponents of informing legal policy making with neuroscientific evidence

argue that it both deepens and makes more concrete and vivid our appreciation of precisely how different adolescents are, and why (Steinberg, 2012).

Despite the uncertainty surrounding the legal weight of such evidence, neuroscience has begun to be introduced as evidence in the youth-related policy decisions made by courts and legislatures. Of course, the causal role of neuroscience in these decisions is disputed.

For example, between 2005 and 2012 the United States Supreme Court issued three decisions detailing the protections that the Constitution affords adolescent offenders. In each case the Court was provided with group-based neuroscientific evidence (among much other evidence). And in each case the Court clarified or expanded the group-wide reach of those protections. The first opinion did not mention neuroscience (*Roper v. Simmons*, 2005), the second mentioned it in a footnote (*Graham v. Florida*, 2010), and the third referred to 'cutting-edge neuroscience' in the text (*Miller v. Alabama*, 2012). It is unknown whether the reference had an actual effect on the outcome, or instead was a citation of convenience.

Separately, a State Senator from Washington reports that neuroscientific evidence about the development of the human brain is what finally convinced her colleagues to pass bills related to childhood education, after 7 years of considering and rejecting similar bills unaccompanied by neuroscientific evidence. While we cannot truly know the impact of neuroscience on any given legislator, the State Senator reported that those who had been most ardently opposed to spending money on childhood education bills came up to her and said that the neuroscience helped them 'get it' (Gehrman, 2007).

Whether in fact the neuroscience had a causal role, these two examples illustrate ways in which group-level neuroscientific evidence appears to be affecting some discussions at the broad policy-making level.

The Limits of Human Judgment and Memory

Neuroscience can also be relevant, at the broad group level, when helping us understand – so as better to develop policies to combat – various frailties that can impede the fair and efficient provision of justice. Two examples suffice to make the point.

First, how might we better understand and combat bias in legal decision makers? The initial step is to better understand the mechanisms by which key legal decisions – such as those by jurors determining guilt, and by judges determining a sentence – are made. Neuroscience has already begun discovering the brain bases of these processes (Buckholtz et al., 2008; Buckholtz and Marois, 2012). A next step is to deploy neuroscientific techniques to deepen our ability to understand and perhaps to combat the influences on and origins of prejudice and bias. Neuroscience has already moved in this direction, helping to uncover what is happening, neurally, during behaviors that reflect unconscious bias (De Martino, 2006; Stanley et al., 2008). Indeed some recent studies have found that brain imaging can be superior to psychological measures at uncovering implicit biases (Korn et al., 2012). Of course, scanning the brains of jurors and judges prior to trials will probably remain an unworkable (or undesirable) proposition, but by revealing and understanding the neural nature of bias, and our susceptibility to it, neuroscience may begin to address what has long been seen as an intractable issue in our criminal justice system.

Second, consider the limitations of eyewitness testimony. Eyewitness testimony is both gravely important and frequently incorrect. For example, of the first 250 prisoners cleared off death row by DNA evidence in the United States, eyewitness testimony was a factor in more than 75% of the convictions. This presents a unique peril to justice because, in the realm of eyewitness testimony, the true threat may come not from those who purposefully undermine the truth, but instead from those who firmly believe that their false memories are true.

Neuroscience has helped to demonstrate that – contrary to assumption – humans are not neurologically capable of maintaining memories of events in a manner similar to a tape recorder that can be played back for the benefit of the court (Schacter and Loftus, 2013). Neuroscience studies have also demonstrated, promisingly, that it is in some contexts possible to differentiate between true and false memories, using brain data, even when the subject does not know the difference (Slotnick and Schacter, 2004). Moreover, neuroscience has helped to reveal that the act of retrieving a memory (as in the course of an investigation, for example) creates a susceptible version of the memory that can be distorted (Nader and Einarsson, 2010). Learning more about how memories can be stored, retrieved, and modified has the potential to affect not only the legal system's policies about how memories (and the jury instructions regarding them) are handled in the courtroom; it may also aid efforts to constructively reform the process of collecting and documenting eyewitness testimony in the course of an investigation.

Conclusion: The Future of Law and Neuroscience

There is no denying that recent and rapid advances in cognitive neuroscience are fueling extraordinary discoveries about the operating machinery of human experience, perception, decision making, and behavior. Brain imaging technology in particular is enabling previously unimaginable capabilities: to noninvasively observe the structure and function of a healthy human brain, both at rest and on task.

What these new capabilities will ultimately reveal about our brains, and about us, remains unclear. Equally unclear, but crucially important, is how legal systems across the world will engage these new capabilities in a manner that improves the system's ability to make high-stakes decisions, about real people, under conditions of marked uncertainty.

Fortunately, there are many efforts underway – individual and collective, public and private, national and international – to help and to guide this engagement. In the end, legal systems must evaluate neuroscientific evidence in an informed, robust, and skeptical but open-minded way – in order to navigate successfully between both the promise and the peril of incorporating neuroscientific insights into legally relevant decision making.

See also: Criminal Liability and Criminal Defenses; Eye Movements in Reading; Functional Brain Imaging of Language Processes; Juvenile Justice: International Law and Children's Rights; Law: History of Its Relation to the Social Sciences; Magnetic Resonance and Molecular Imaging in Psychiatry; Moral Reasoning in Psychology; Prefrontal Cortex Development and Development of Cognitive Function; Science and Law; Scientific Evidence: Legal Aspects; Torts in Law; Trials.

Bibliography

Aharoni, E., Vincent, G.M., Harenski, C.L., Calhoun, V.D., Sinnott-Armstrong, W., Gazzaniga, M.S., Kiehl, K.A., 2013. Neuroprediction of future rearrest. Proceedings of the National Academy of Sciences 110, 6223–6228.

Brower, M.C., Price, B.H., 2001. Neuropsychiatry of frontal lobe dysfunction in violent and criminal behaviour: a critical review. Journal of Neurology, Neurosurgery & Psychiatry 71, 720–726.

Buckholtz, J.W., Marois, R., 2012. The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. Nature Neuroscience 15, 655–661.

Buckholtz, J.W., Asplund, C.L., Dux, P.E., Zald, D.H., Gore, J.C., Jones, O.D., Marois, R., 2008. The neural correlates of third-party punishment. Neuron 60, 930–940.

Bunge, S.A., Dudukovic, N.M., Thomason, M.E., Vaidya, C.J., Gabrieli, J.D., 2002. Immature frontal lobe contributions to cognitive control in children: evidence from fMRI. Neuron 33, 301–311.

De Martino, B., 2006. Frames, biases, and rational decision-making in the human brain. Science 313, 684–687.

Edersheim, J.G., Brendel, R.W., Price, B.H., 2012. Neuroimaging, diminished capacity and mitigation. In: Simpson, J.R. (Ed.), Neuroimaging in Forensic Psychiatry: From the Clinic to the Courtroom. John Wiley & Sons, Chichester, pp. 163–194.

Eisenberger, N.I., 2012. The neural bases of social pain: evidence for shared representations with physical pain. The University of Chicago Law Review 74, 126–135.

Faigman, D.L., Monahan, J., Slobogin, C., 2014. Group to individual (G2i) inference in scientific expert testimony. The University of Chicago Law Review 81.

Farah, M.J., Hutchinson, J.B., Phelps, E.A., Wagner, A.D., 2014. Functional MRI-based lie detection: scientific and societal challenges. Nature Reviews Neuroscience 15 (2), 123–131.

Feresin, E., 2011. Italian court reduces murder sentence based on neuroimaging data, 01 September. Nature News Blog viewed 20 September 2013. http://blogs.nature.com/news/2011/09/italian_court_reduces_murder_s.html.

Friston, K.J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M.D., Turner, R., 1998. Event-related fMRI: characterizing differential responses. NeuroImage 7, 30–40.

Freeman, M.D., Goodenough, O.R., 2009. Law, Mind and Brain. Ashgate Publishing, Surrey.

Gehrman, E., 2007. From neuroscience to childhood policy. Harvard Gazette viewed 5 September, 2013. http://news.harvard.edu/gazette/story/2007/12/from-neuroscience-to-childhood-policy/.

Graham v. Florida, 560 U.S. 48 (2010).

Greely, H.T., Illes, J., 2007. Neuroscience-based lie detection: the urgent need for regulation. American Journal of Law & Medicine 33, 377–432.

Greene, J., Cohen, J., 2004. For the law, neuroscience changes nothing and everything. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 359, 1775–1785.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., Nelson, C., 2000. Clinical versus mechanical prediction: a meta-analysis. Psychological Assessment 12, 19.

Gurley, J.R., Marcus, D.K., 2008. The effects of neuroimaging and brain injury on insanity defenses. Behavioral Sciences & the Law 26, 85–97.

Johnson, J.D., McDuff, S.G.R., Rugg, M.D., Norman, K.A., 2009. Recollection, familiarity, and cortical reinstatement: a multivoxel pattern analysis. Neuron 63, 697–708.

Jones, O.D., 2013a. Foreword. In: Morse, S.J., Roskies, A.L. (Eds.), A Primer on Criminal Law and Neuroscience. Oxford University Press, New York.

Jones, O.D., 2013b. Seven ways neuroscience aids law. In: Battro, A., Dehaene, S., Singer, W. (Eds.), Neurosciences and the Human Person: New Perspectives on Human Activities. Pontifical Academy of Sciences, Vatican City.

Jones, O.D., Shen, F.X., 2012. Law and neuroscience in the United States. In: Spranger, T.M. (Ed.), International Neurolaw: A Comparative Analysis. Springer-Verlag, Berlin, Germany.

Jones, O.D., Buckholtz, J.W., Schall, J.D., Marois, R., 2009. Brain imaging for legal thinkers: a guide for the perplexed. Stanford Technology Law Review 5, 10–09.

Jones, O.D., Marois, R., Farah, M.J., Greely, H.T., 2013a. Law and neuroscience. Journal of Neuroscience 33 (45), 17624–17630.

Jones, O.D., Wagner, A.D., Faigman, D.L., Raichle, M.E., 2013b. Neuroscientists in court. Nature Reviews Neuroscience 14 (10), 730–736.

Jones, O.D., Schall, J.D., Shen, F.X., 2014. Law and Neuroscience. Aspen Publishers.

Junger-Tas, J., Decker, S.H. (Eds.), 2006. International Handbook of Juvenile Justice. Springer, Berlin.

Korn, H.A., Johnson, M.A., Chun, M.M., 2012. Neurolaw: differential brain activity for black and white faces predicts damage awards in hypothetical employment discrimination cases. Social Neuroscience 7, 398–409.

Kozel, F.A., Johnson, K.A., Mu, Q., Grenesko, E.L., Laken, S.J., George, M.S., 2005. Detecting deception using functional magnetic resonance imaging. Biological Psychiatry 58, 605–613.

Lewine, J.D., Davis, J.T., Bigler, E.D., Thoma, R., Hill, D., Funke, M., Sloan, J.H., Hall, S., Orrison, W.W., 2007. Objective documentation of traumatic brain injury subsequent to mild head trauma: multimodal brain imaging with MEG, SPECT, and MRI. The Journal of Head Trauma Rehabilitation 22, 141–155.

Logothetis, N.K., 2008. What we can do and what we cannot do with fMRI. Nature 453, 869–878.

Luce, R., 2012. Proving a 'mild' traumatic brain injury: a complex but no longer impossible task. Vermont Bar Journal 38, 1–8.

Luna, B., Padmanabhan, A., O'Hearn, K., 2010. Brain and cognition. Brain and Cognition 72, 101–113.

MacDonald, A.W., 2000. Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. Science 288, 1835–1838.

Maharashtra v. Sharma, No. 508/07 (Ct. Sess. Pune, India 12 June 2008) http://court.mah.nic.in/courtweb/orders/pundcis/orders/201501005082007_1.pdf.

Maroney, T., 2010. The false promise of adolescent brain science in juvenile justice. Notre Dame Law Review 85, 89.

McKee, A.C., Cantu, R.C., Nowinski, C.J., Hedley-Whyte, E.T., Gavett, B.E., Budson, A.E., Santini, V.E., Lee, H.-S., Kubilus, C.A., Stern, R.A., 2009. Chronic traumatic encephalopathy in athletes: progressive tauopathy following repetitive head injury. Journal of Neuropathology and Experimental Neurology 68, 709.

Miller v. Alabama, 132 S.Ct. 2455 (2012).

Morse, S., 2011. Avoiding irrational neurolaw exuberance: a plea for neuromodesty. Mercer Law Review 62, 837–859.

Morse, S.J., Roskies, A.L. (Eds.), 2013. A Primer on Criminal Law and Neuroscience. Oxford University Press, New York.

Nader, K., Einarsson, E.Ö., 2010. Memory reconsolidation: an update. Annals of the New York Academy of Sciences 1191, 27–41.

Posner, M.I., Raichle, M.E., 1994. Images of Mind. Scientific American Library/Scientific American Books, New York.

Pryor, E.S., 1991. Compensation and the ineradicable problems of pain. Workers' Compensation Law Review 14, 307.

Rainville, P., Duncan, G.H., Price, D.D., Carrier, B., Bushnell, M.C., 1997. Pain affect encoded in human anterior cingulate but not somatosensory cortex. Science 277, 968–971.

Rilling, J.K., Sanfey, A.G., 2011. The neuroscience of social decision-making. Annual Review of Psychology 62, 23–48.

Rissman, J., Greely, H.T., Wagner, A.D., 2010. Detecting individual memories through the neural decoding of memory states and past experience. Proceedings of the National Academy of Sciences of the United States of America 107, 9849–9854.

Roper v. Simmons, 543 U.S. 551 (2005).

Saks, M.J., Schweitzer, N.J., Aharoni, E., 2014. The impact of neuroimages in the sentencing phase of capital trials. Journal of Empirical Legal Studies 11, 105–131.

Schacter, D.L., Loftus, E.F., 2013. Memory and law: what can cognitive neuroscience contribute? Nature Publishing Group 16, 119–123.

Schweitzer, N.J., Saks, M.J., Murphy, E.R., Roskies, A.L., Sinnott-Armstrong, W., Gaudet, L.M., 2011. Neuroimages as evidence in a mens rea defense: no impact. Psychology, Public Policy, and Law 17, 357–393.

Shen, F.X., 2013. Neuroscience, mental privacy, and the law. Harvard Journal of Law and Public Policy 36, 653–713.

Simon, R.J., Ahn-Redding, H., 2008. The Insanity Defense the World Over. Lexington Books, Lanham, MD.

Slotnick, S.D., Schacter, D.L., 2004. A sensory signature that distinguishes true from false memories. Nature Neuroscience 7, 664–672.
Stanley, D., Phelps, E., Banaji, M., 2008. The neural basis of implicit attitudes. Current Directions in Psychological Science 17, 164–170.
Steinberg, L., 2008. A social neuroscience perspective on adolescent risk-taking. Developmental Review 28, 78–106.
Steinberg, L., 2012. Should the science of adolescent brain development inform public policy? Issues Sci. Technol 28, 67–78.
U.S. v. Semrau, WL 6845092 (2010).
Wagner, A., 2010. Can neuroscience identify lies?. In: A Judge's Guide to Neuroscience: A Concise Introduction. University of California, Santa Barbara, pp. 13–25.
Weisberg, D.S., Keil, F.C., Goodstein, J., Rawson, E., Gray, J.R., 2008. The seductive allure of neuroscience explanations. Journal of Cognitive Neuroscience 20, 470–477.
Wilson v. Corestaff Services, 900 N.Y.S.2d 639 (2010).
Zitter, J.M., 1997. Excessiveness or adequacy of damages awarded for injuries to head or brain. American Law Reports 50, 1–216.

Relevant Websites

http://www.lawneuro.org/bibliography.php – Law and Neuroscience Bibliography.
http://www.lawneuro.org/ – MacArthur Foundation Research Network on Law & Neuroscience.
http://www.lawneuro.org/listserv.php – Neurolaw News.

Law and People with Disabilities

Janet E Lord, Syracuse University, Baltimore, MD, USA
Deepti S Raja, Syracuse University, Washington, DC, USA
Peter Blanck, Syracuse University, Syracuse, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M. Jones, volume 12, pp. 8465–8470, © 2001, Elsevier Ltd.

Abstract

This article describes the progressive development of disability law, particularly in light of international developments that are transforming domestic disability law frameworks. It provides an overview of the conceptualization of disability in law, charting the shift from a medical model to a social model of disability and more recently to a human rights model. The abstract covers the implications of the Convention on the Rights of Persons with Disabilities (CRPD), and the processes and approaches for translating international disability law standards into national level law. It sheds light on the contemporary disability rights landscape in the light of dramatic legislative action promoted by the adoption of the CRPD.

Introduction

People with disabilities represent 15% of the world's population, many of whom experience discrimination and exclusion and live in poverty, with restricted access to basic goods and services (World Health Organization and World Bank, 2011). The United Nations adopted the Convention on the Rights of Persons with Disabilities (the Convention or CRPD) and its Optional Protocol in 2006, with the aim of establishing a comprehensive framework to promote and protect the rights of persons with disabilities (CRPD, 2006; Optional Protocol, 2006). The Convention elaborates minimum legal standards to address disability discrimination and to promote equality for persons with disabilities in all spheres of life – political, social, economic, cultural, and civil (Schur et al., 2013). The Optional Protocol allows individuals and groups to seek redress at the international level when these standards are not met, and provides for an independent committee of experts to undertake investigations into serious breaches of disability rights as set forth in the Convention.

At the time the Convention was proposed, in 2001, very few countries around the world had any disability law or policy framework. Its drafters recognized that, throughout the world, national disability law and policy remained underdeveloped or, worse, contained explicitly discriminatory provisions that strip persons with disabilities of their rights (Stein and Lord, 2012). In many respects, law was a tool for exclusion and oppression of persons with disabilities. Fewer than 50 countries had disability discrimination laws, while many countries limited the rights of persons with disabilities through laws and policies, that for instance, excluded persons with disabilities from exercising their right to vote, placed arbitrary restrictions on their right to certain types of employment or limited their ability to take part in the judicial process as witnesses, jurors, or even claimants. In other examples, disability was addressed narrowly, as a social welfare matter, typically coordinated at the government level by ministries of health or social welfare (Degener and Quinn, 2002).

Since 2006, many countries have adopted the standards in the Convention through the process of ratification (UN Enable). As of 26 August 2013, 133 countries have ratified and 156 have signed the Convention (UN Enable). Ratification is an important step in recognizing disability rights at the national level. Countries around the world are undertaking comprehensive and cross-cutting reviews to assess the compatibility of their laws to the Convention and, where needed, to reform and develop laws to bring the national legal system into alignment with the Convention's standards. Law reform is one step in ensuring full respect for the enjoyment of the rights of persons with disabilities; yet other measures are needed in order to effect real change. Budgets must be allocated to ensure adequate implementation of laws; the public must be made aware of the rights of persons with disabilities to ensure that their actions do not obstruct persons with disabilities from enjoying their rights; and monitoring mechanisms must be put into place to protect disability rights (Lord and Stein, 2008). In addition, persons with disabilities and their representative advocacy organizations must have the capacity to claim their rights and to participate in law, policy, and programming (Lord et al., 2012).

This article considers the progressive development of disability law, particularly in the light of developments at the international level that have promoted the transformation of domestic disability law frameworks. It first provides an overview of the conceptualization of disability in law, focusing on transitional perspectives and examining current trends. It charts the shift from a medical model of disability to a rights-oriented social model of disability in law and policy. Thereafter, the article considers the CRPD framework and its implications for law reform. Then, it addresses the processes and approaches for translating international disability law standards into national level law. The article concludes with observations on the contemporary disability rights landscape in the light of dramatic legislative action promoted by the adoption of the CRPD.

Conceptualizing Disability in Law

The conceptualization of disability in law and policy has progressed, in the post–World War II era, from a limited, medicalized, and welfare-oriented understanding to a rights-based, sociocontextual perspective. The implications of these perspectives are significant and help to explain the gradual

enlargement of the disability law landscape. The various models of disability are discussed in turn.

Models of Disability and Their Influence on Disability Law

As stated by Kimberlé Williams Crenshaw, "law … embodies and reinforces ideological assumptions about human relations that people accept as natural or even immutable" (cited in Berg, 1999: p. 4). The conceptualization of and cultural discourse on disability in society has evolved overtime, and these changes are reflected in the nature and scope of laws that impact persons with disabilities.

The early understanding of disability centered on its perceived biological origins, wherein impairment resulted from an injury, disease, or other medical condition (Logue and Blanck, 2010). This 'medical model' of disability focused on the impairment as a deviation from or disfigurement of the 'normal' human body, and conceived of a person to be less capable than his or her peers without disabilities. The medicalization of disability is derived from assumptions and expectations of persons without disabilities about those experiencing life with a disability. This resulted in the notions of the disabled being weaker, less able to undertake and effectively accomplish activities such as going to school and getting a job, and more dependent on others for their subsistence (Berg, 1999; Blanck, 2004; Ingram, 2006; Muller, 2011; Stein, 2007).

With the cultural discourse focused on the presumed deficiencies of persons with disabilities and their resulting economic needs, the medical model shaped disability welfare laws around the world. Social welfare laws for persons with disabilities often equate disability with the inability to work and earn an income, and create government benefits that support people economically as well as provide them access to necessary medical, rehabilitation, and support services, for example, through medical insurance, cash benefits, and personal assistance services (Bagenstos, 2004; Blanck et al., 2013; Jones and Marks, 2001). While social welfare laws for persons with disabilities were established early on in the twentieth century, they continue to provide government benefits and assistance in most countries despite the newer rights-based legislations.

Beginning in the 1970s, disability advocates began to question and dismantle the dominant medical view of disability, and raised a different reason for the exclusion of the disabled in society. They countered that disability was a social, rather than a medical, condition, caused by the challenges in interacting with environments and systems, which are not designed to accommodate human differences and are thus inaccessible to persons with disabilities (Gray et al., 2003). The 'social model' of disability challenged the presumption that benefits and welfare were the legal solutions to the exclusion of people with disabilities, and instead advocated for legislation that would remove barriers to their full social and economic participation. The social model of disability is seen as a 'minority model' of disability by some legal scholars, in that it defines persons with disabilities as a group, like other minorities, that has been willfully excluded from mainstream society due to stigma, negative attitudes, and stereotypes about their abilities and capacities (Bickenbach, 1999). The social model of disability shaped the next iteration of disability-related legislation with a focus on antidiscrimination and equal access (Blanck, 2014).

Antidiscrimination law aims to provide protections and safeguards to minority groups and promote equality of opportunity, along with resources to seek compensatory damages (Jones and Marks, 2001). As the social model of disability became the dominant view of disability, countries adopted antidiscrimination into their disability legislation either as the foundational premise or within sections targeted at employment and access to public and private goods and services. The Americans with Disabilities Act passed in 1990 has been widely viewed as the standard bearer of disability civil rights and antidiscrimination laws that were enacted in many countries in the 1990s (Blanck, 2006, 2008; Stein, 2007). In many countries, antidiscrimination law is coupled with legislation establishing quotas for persons with disabilities in education, employment, political representation, housing, and other domains (Byrnes, 2011; Degener, 2005; Kim, 2011) as a corrective measure to level the playing field. The social model of disability is related to the independent living movement in western societies where persons with disabilities fought for their right to live within inclusive communities independently, rather than be segregated in institutions or other exclusionary arrangements (Blanck, 2006).

The social model of disability focuses on the corrective measures required in a society to break down environmental and attitudinal barriers to participation, and has conceptually organized the fight for civil rights for persons with disabilities. However, with time scholars and advocates have proposed a more comprehensive approach that looks beyond civil rights and liberties and frames disability as a human rights imperative. Stein and Stein (2007) describe this approach as a way to go beyond seeking only first generation civil and political rights through antidiscrimination legislation, to seeking first and second generation rights that include equality of opportunity on the "full spectrum of social, cultural, and economic measures" (p. 1205). A comprehensive 'human rights model' of disability will advance laws that aim to recognize the human rights of persons with disabilities; that is, "the fundamental, universal and indivisible principles by which every single human being can gain justice and equality" (Albert and Hurst, 2006: p. 2). The human rights model of disability has been the founding element behind the CRPD, the first international treaty on disability. The next section discusses the application of human rights standards to disability law.

As described above, different conceptualizations and cultural realities often lead to different legal responses to disability. Despite their differences, these legal responses frequently coexist within legal systems; countries use a combination of welfare, antidiscrimination, affirmative action, and human rights–based legal approaches for persons with disabilities. Most countries use a variety of legal mechanisms such as constitutional law, civil rights law, and criminal law to employ and implement these differing legal approaches (Degener, 2005).

Increasingly, countries are turning to constitutional law to incorporate the human rights model of disability and articulate the recognition of the human rights of persons with disabilities in the highest law of the land. In some countries, notably South Africa, Uganda, Zambia, and Thailand (United Nations, 2011), disabled peoples organizations (DPOs) have

used constitution-building processes to leverage disability rights. Ecuador introduced a chapter on disability within its 2008 Constitution while also issuing a Presidential Executive Decree in 2007 to promote the development of disability programs across all sectors (United Nations, 2011). Kenya provides another example of constitutional change to advance disability rights and the CRPD through Article 54 in its 2010 Constitution which "seeks to minimize barriers to equalization of opportunities in all aspects of social-cultural, economic and political life" for persons with disabilities (United Nations, 2011).

Criminal laws that focus on persons with disabilities demonstrate the contradictory, complex, and varying contextualizations of disability prevalent in society. Many such laws concern persons with disabilities as victims of crime, stemming from the increasing recognition of hate crime, violence, and abuses against persons with disabilities (International Network of Women with Disabilities, 2011; Hughes et al., 2012; Piggott, 2011; Shakespeare, 2012). On the other hand, some types of disabilities, especially psychosocial disabilities, have been criminalized and penalized through laws that mandate involuntary confinement and forced institutionalizations (Perlin and Szeli, 2012; Szmukler, 2010).

Emergence of International Human Rights Law Standards on the Rights of Persons with Disabilities

The rationale for the development of a disability-specific human rights treaty was the virtual invisibility of persons with disabilities from the human rights system and the reality that an estimated one billion persons with disabilities around the world are excluded from society and are routinely denied access to education, employment, health care services, and basic needs (World Health Organization and World Bank, 2011: p. 29). Moreover, persons with disabilities are far more likely than others to live in poverty.

International human rights law did little to address the human rights of persons with disabilities prior to the adoption of the CRPD. No international human rights treaty comprehensively addressed the multitude of barriers experienced by persons with disabilities. While few states had developed disability rights law and policy frameworks, where they did have disability-related legislation, often they explicitly sanctioned disability discrimination; for example, denying persons with disabilities the right to vote, obtain an education, serve as a juror, open a bank account, own property, or work in certain sectors (Quinn, 2009). Welfare-oriented approaches to disability law and policy served to construct the development of holistic, rights-oriented legal protections (Degener and Quinn, 2002).

The UN General Assembly adopted the Convention on the Rights of Persons with Disabilities by general consensus in 2006, opening for signature by states parties on 30 March 2007. It attained the requisite 20 ratifications in short order, triggering its entry into force on 3 May 2008.

The CRPD is the first legally binding international human rights treaty to address the rights and fundamental freedoms of one billion persons and to set forth a framework for comprehensive domestic legislation to address disability rights (CRPD, 2006). The terms of the CRPD require the reform of legal systems, justice sectors, and institutions in meeting its obligations. Significantly, it goes beyond this traditional focus and brings within its ambit nonformal and decentralized justice and administrative systems and processes and recognizes the role that local communities play in rights protection and the duties of private as well as public actors in fostering inclusion (Lord et al., 2012). Its potential for shaping national disability law and policies and transforming institutions is far reaching.

The CRPD is responsive to the legal void that existed at the international level in disability law. The CRPD charts a progressive and comprehensive course for the development and reform of domestic disability law and policy in alignment with international human rights principles. It includes monitoring mechanisms intended to build the capacity of national governments, national human rights institutions, and DPOs in advancing disability rights.

UN Disability Convention Structure and Content

The CRPD comprises 25 preambular paragraphs and 50 articles. Its preamble lays out the historical progression of international standards on disability and highlights issues of importance, including the understanding of disability as a socially constructed phenomenon. The 50 operative articles in the Convention provide a comprehensive legal framework for the consideration of civil, political, economic, social, and cultural rights within the specific context of disability. It is the most detailed framing of the rights of persons with disabilities found in any treaty and serves to amplify the protections to which persons with disabilities are entitled in other, non-disability-specific human rights instruments (Kayess and French, 2008).

The CRPD has an introductory set of provisions that outline its purpose (Article 1), set out key definitions (Article 2), and articulate general articles that apply and must be interpreted and applied across provisions of the treaty (Articles 3–9). One of these general provisions is Article 5, Equality and Nondiscrimination, which prohibits discrimination on the basis of disability and requires the provision of reasonable accommodation. Article 9, Accessibility, creates obligations for accessibility in a variety of contexts. Articles 10–30 enumerate specific substantive rights of the Convention, including civil, political, economic, social, and cultural rights. The CRPD establishes a system of national and international level monitoring and implementation (Articles 31–40), and, like other treaties, contains provisions that govern its operation (Articles 41–50).

The general requirements set forth in Article 4 make clear the need to ground CRPD obligations in national law, policy, and programming, in consultation with persons with disabilities. It requires states parties to consult with and involve persons with disabilities in developing and implementing legislation and policies and in decision-making processes (CRPD, 2006: art. 4). Accordingly, the CRPD affirms the rights of stakeholders and their representative organizations to be heard, and signals recognition that participation and inclusion in decision-making is a precondition of legal empowerment.

The CRPD establishes national level monitoring requirements with three dimensions. It calls for cross-governmental coordination in recognition that disability is a cross-cutting issue, requires independent monitoring to be performed by a national human rights or disability rights institution, and calls for stakeholder participation in domestic monitoring (CRPD, 2006: art. 33). A committee on the Rights of Persons with Disabilities – the CRPDs treaty monitoring body – is tasked with monitoring implementation by states parties through its oversight of the mandatory reporting requirement and the issuance of general recommendations for the state party concerned. The Optional Protocol to the CRPD, consisting of 18 articles, gives the Committee competence to examine individual complaints with regard to alleged CRPD violations by parties to the protocol (Optional Protocol, 2006). It allows states parties to opt into participation in individual and group communications procedures and establishes an inquiry procedure, both of which are overseen by the Committee. These mechanisms are vehicles for empowering disability advocates and their representative organizations to bring complaints for individual violations as well as systemic abuses to the attention of the CRPD for review. Finally, the CRPD establishes a periodic meeting of a Conference of States Parties to consider issues relating to implementation (CRPD, 2006: art. 40). This mechanism allows state parties and disability advocates to engage in dialogue regarding implementation and best practices in law and policy reform consistent with the CRPD.

In sum, the CRPD sets forth general obligations familiar to human rights treaties – requiring national law reform and domestic incorporation of its provisions. It also provides a blueprint for national level disability rights advocacy and action. Significantly, the Convention contains a diverse set of obligations that, when surveyed, constitute a map of advocacy and empowerment opportunities at the local, national, and international levels. In so doing, the CRPD plots a course for a human rights practice that goes beyond traditional, top-down law reform-oriented approaches to embrace legal empowerment and transformation that is locally driven, community focused, and civil society oriented (Lord et al., 2012).

Translating International Standards into National Legislation

Although more than 130 countries have enthusiastically supported, ratified, and signed the CRPD, many face challenges in implementing the Convention due to disempowering social contexts for persons with disabilities and underdeveloped legal systems. This reality raises a variety of issues, among them the varied perspectives on the nature and definition of disability, and the differentiation between developed and developing countries in capacities for implementation and distinctions in approach to domestic legislative reform to bring about human rights implementation. These challenges are considered in turn.

Defining Disability in Law

Defining disability in law and policy is complex and no universal definition exists. Even within a single legal system, a multitude of definitions may be used to define disability in discrete contexts. Defining disability is highly relevant for identifying the class of individuals for whom disability rights protection is accorded. The CRPD does not provide a concrete definition of disability, but states that "persons with disabilities include those who have long-term physical, mental, intellectual or sensory impairments which in interaction with various barriers may hinder their full and effective participation in society on an equal basis with others" (CRPD, 2006: art. 1). As such, the understanding of disability in the CRPD sets a baseline, from which states may expand disability rights protection to cover, for instance, disability discrimination in the case of an individual who is temporarily disabled.

The CRPD makes clear that the term 'person with disabilities' in the Convention 'includes' individuals with long-term impairments and thus 'persons with disabilities' may indeed encompass a broader category of persons at the national level. Further, impairments are referred to in the CRPD as 'physical, mental, intellectual, or sensory' in recognition that disability is diverse with the implication that legislation must provide protection to disabled persons whatever their impairment be. Legislation excluding a given category of individuals with disability from protection, for instance, voting exclusions based on mental disability, is prohibited (CRPD, 2006: arts. 1, 12, 29).

This conceptualization demands a shift in how policy makers have reflected on disability. For example, disability laws in some countries define disability with a finite list of observable impairments, ignoring hidden disabilities that arise from chronic or mental illness, learning and cognitive difficulty, and traumas due to accidents, and disregarding environmental and interactional issues completely (Byrnes, 2011).

Disability Inclusion in Legal Reform Assistance

The provisions and motivations under Article 32, which covers international cooperation, promote the potential for state parties to assist one another with challenges in crafting required innovations and facilitating access to resources as an aid in implementation (Guersey et al., 2007). Promoting the international exchange of technical knowledge and best practices, while providing resources for capacity building, is intended to aid government workers and development practitioners in their efforts to ushering in improvements to and reforms of policies, programs, and practices. Addressing disability in international aid and development funding and support will assist low- and middle-income states parties to provide far-reaching and inclusive programs that meet their goals and obligations under the CRPD.

The barriers faced by persons with disabilities in developing countries restrict their access to information and awareness of their rights, a situation also faced by persons living in poverty and other disadvantaged groups in these countries. As emphasized in the *World Report on Disability*, persons with disabilities are likely to live in poverty and are often restricted in their access to education, employment, transportation, and health care; and often live in isolation from the wider community (World Health Organization and World Bank, 2011: p. 39). The social determinants of legal empowerment, therefore, are severely restricted for persons with disabilities. The need for awareness-building initiatives, including legal literacy programs,

participatory human rights education tied to action at the local level, legal aid bureaus, and inclusive development programming aiming to integrate persons with disabilities into mainstream programming (such as health, economic development, democracy and governance, and education) are important and implicit in the CRPD framework.

Preconditions for Law Reform

Disability awareness-raising is recognized as a core precondition of legislating for disability rights (CRPD, 2006: art. 8). As with other marginalized groups, persons with disabilities were generally excluded from participating in early disability law and policy efforts. As a result, numerous disabled people's organizations are now focusing their attention on building an understanding of legal rights and, more specifically, how to claim these rights through legal action (Golub and McQuay, 2001).

The CRPD recognizes in Article 8 that combating stigma and stereotyping rooted in disability discrimination is essential to laying a foundation for effective law and policy change (CRPD, 2006: art. 8). The focus is on disability awareness to create a culture of legal equality combined with explicit legal mandates to require public consultation as well as accessibility measures in the context of law and policy making.

Legislating Disability Rights

Many of the concepts set forth in the CRPD require states parties to bring about law and policy reforms (or the development of disability law and policy) to align with the shift to the human rights model required by the CRPD. This shift in framework presents challenges whilst creating opportunities for progressive change and legal innovation (Kanter, 2007). The requirements for legal reform to align with CRPD obligations hinge on the application of treaties with the domestic legal system of a country. It also depends upon the general state of the domestic legal framework in relation to its recognition of disability rights.

Some legal systems mandate legislative enactment to incorporate an international treaty into the domestic legal order (e.g., United Kingdom, New Zealand). In some countries, however, treaty obligations assume legislative supremacy and are automatically part of the national legal order on ratification (e.g., Argentina, the Netherlands). In other countries, elements of both approaches are evident in that some standards are self-executing and do not require additional domestic action to be enforceable, while other standards require legislative action or need clarification to be domestically enforceable.

Law review and reform play a significant role in all these systems. The CRPD requires states to undertake the following actions in the context of law reform: (1) states parties must adopt all appropriate legislative, administrative, and other measures for the implementation of the CRPD; and (2) states parties must adopt all appropriate measures, including legislation, to modify or abolish existing laws, regulations, customs, and practices that constitute discrimination against persons with disabilities (CRPD, 2006: art. 4).

The general obligation in the CRPD to give effect to its provisions in the domestic legal order implies the undertaking of comprehensive legal review in order to assess alignment with the treaty. Where gaps or inconsistencies appear, such review exercises identify areas requiring corrective measures in the form of legal development or amendment (Lord and Stein, 2008). That is to say, treaty implementation requires amendment of laws that are inconsistent as well as measures to put into place additional laws to advance implementation. The starting point for such review generally is an assessment of (1) the Constitution; (2) existing antidiscrimination legislation, whether disability-specific or general in scope; (3) existing comprehensive disability rights legislation, such as a national disability law; and (4) human rights acts.

The CRPD does not specify the precise format of implementation and thus allows an amount of flexibility, given the variation in legal systems and legislative frameworks among state parties. There are options that states may consider; for example, comprehensive nondiscrimination legislation that includes protection from discrimination on the basis of disability in one or many spheres, or a comprehensive disability rights law that addresses protection from disability discrimination as well as other types of measures. Equally, a state might introduce reforms and additional measures in discrete areas; for instance, reforming electoral codes and employment law together with new legislation to address gaps in education, transportation, or other areas to bring about wholesale change. In most instances, a combined approach will be necessary to produce full compliance and alignment with the obligations of the CRPD.

In view of the fact that the CRPD is a far reaching instrument and given that disability is a cross-cutting issue with relevance across a legal framework, a legal review will entail an assessment of a range of legislation, beyond disability-specific legislation or human rights legislation (Lord and Stein, 2008). Legislation may be usefully grouped into three categories of particular relevance for disability rights assessment: (1) Legislation of a disability-specific nature, such as a comprehensive national law on disability or an inclusive education statute or a mental health law; (2) legislation that relates to all persons or general groups of persons, but which may make particular reference to or provision for disability, such as antidiscrimination law, social security law, guardianship laws, education laws, criminal codes, or electoral codes; and (3) legislation that does not address persons with disabilities, but which is relevant to persons with disabilities, such as tax laws, construction laws, family laws, intellectual property laws, and banking and contract laws.

Finally, legal assessment for the purpose of disability law reform requires the participation of persons with disabilities and their representative organizations, as recognized in the CRPD (CRPD, 2006: art. 4(3)). In this sense, the process of law reform is fundamental to its outcome. The lived experience of persons with disabilities provides a sound basis for law and policy reform that is responsive to the specific needs of individuals and offers the best chance for human rights implementation (Blanck, 2014).

Conclusion

The first decade of the twenty-first century marked a transformative period in the progressive development of disability law, both at the international level as a consequence

of the drafting and adoption of the CRPD, and at the national level, as the result of significant reforms in domestic disability law. This period has witnessed a number of notable changes at the domestic level, where disability is emerging as a salient human rights issue and the rights of persons with disabilities are increasingly understood as worthy of attention and protection. Notable are constitutional law developments in countries where disability is a prohibited ground of discrimination and efforts across the world to adopt comprehensive disability antidiscrimination laws. The drafters of the CRPD recognized that, throughout the world, national disability law and policy frameworks were underdeveloped or, worse, contained explicitly discriminatory provisions that strip persons with disabilities of their rights. Several countries have focused their attention on adopting sign language as a national language in legislation. Others are working to modify guardianship laws to comply with the supported decision-making framework and legal capacity provisions of the CRPD. Still other countries are reviewing electoral codes and Internet usage, implementing regulations to enhance their accessibility for persons with disabilities. Advocates are likewise challenging outmoded national laws before domestic and regional courts, making claims about discriminatory educational systems that segregate students with disabilities and provide them with inferior education or challenging policies of institutionalization that inhibit the social inclusion of persons with disabilities. In the light of these recent progressive developments and growing adoption of the CRPD, the dynamic pace of disability law reform and legal development shows little sign of slowing.

Acknowledgments

Support for this project in part is from grants to BBI (see http://bbi.syr.edu), such as: The National Institute on Disability and Rehabilitation Research (NIDRR), U.S. Department of Education, for (i) "Southeast ADA Center," Grant No. H133A110021, and (ii) "Center on Effective Delivery of Rehabilitation Technology by Vocational Rehabilitation Agencies," Grant No. H133A090004. This article is adapted from presentations made at the "Persons with Disabilities–Innovation and Empowerment through the Convention on the Rights of Persons with Disabilities" session, World Bank Law, Justice and Development Week, Washington, DC, 2011. The authors wish to thank participants in the panel session for their inputs. The authors wish to thank Robert J. Borrelle for his assistance in preparing this paper for publication.

See also: Afrocentric Approaches to Social Work; Disability, Geography of; Disability: Social and Psychological Aspects; Disability: Sociological Aspects; Empirical Legal Studies; International Law and Treaties.

Bibliography

Albert, B., Hurst, R., 2006. Disability and a human rights approach to development. Briefing Paper, Disability Knowledge and Research Publications. http://www.daa.org.uk/uploads/pdf/Disability,%20Development%20&%20HR.pdf (accessed 12.04.13.).

Bagenstos, S.R., 2004. The future of disability law. Yale Law Journal 114, 1–84. http://www.yalelawjournal.org/pdf/114-1/Bagenstos_FINAL.pdf (accessed 12.04.13.).

Berg, P.E., 1999. Ill/legal: interrogating the meaning and function of the category of disability in antidiscrimination law. Yale Law and Policy Review 18 (1), 1–51.

Bickenbach, J., 1999. Minority rights or universal participation: the politics of disablement. In: Jones, M., Basser Marks, L.A. (Eds.), Disability, Divers-Ability and Legal Change. Martinus Nijhoff, The Hague, pp. 101–115.

Blanck, P., 2004. Justice for all? Stories about Americans with disabilities and their civil rights. Journal of Gender, Race & Justice 8, 1–30.

Blanck, P., 2006. Americans with disabilities and their civil rights: past, present, and future. University of Pittsburgh Law Review 66, 687–719.

Blanck, P., 2008. The right to live in the world: disability yesterday, today, and tomorrow. Texas Journal on Civil Liberties and Civil Rights 13 (2), 369–403.

Blanck, P., 2014. eQuality: Web Equality for People with Cognitive Disabilities. Cambridge University Press, New York, NY.

Blanck, P., Hill, E., Siegal, C., Waterstone, M., 2013. Disability Civil Rights Law and Policy, Cases and Materials. Thomson/West Publishers, St. Paul, MN.

Byrnes, A., 2011. Disability discrimination law and the Asian and Pacific Region. In: ESCAP, Expert Group Meeting on Harmonization of National Legislation with the Convention on the Rights of Persons with Disabilities in Asia and the Pacific. Bangkok, June 8-10, 2009. http://e.unescap.org/sdd/issues/disability/crpd/files/Paper-III-Andrew-20110121.pdf (accessed 11.04.13.).

Convention on the Rights of Persons with Disabilities, 2006. G.A. Res. 61/106, 46 I.L.M. 433.

Degener, T., 2005. Disability discrimination law: a global comparative approach. In: Lawson, A., Gooding, C. (Eds.), Disability Rights in Europe: From Theory to Practice. Hart, London, pp. 87–106.

Degener, T., Quinn, G., 2002. A survey of international, comparative and regional disability law reform. In: Breslin, M.L., Yee, S. (Eds.), Disability Rights Law and Policy: International and National Perspectives. Transnational, New York, pp. 25–45. http://www.dredf.org/international/degener_quinn.html (accessed 24.07.12.).

Golub, S., McQuay, K., 2001. Legal empowerment: good governance and poverty reduction. In: Office of the General Counsel (Ed.), Law and Policy Development at the Asian Development Bank. Asian Development Bank, Manila, pp. 7–164.

Gray, D.B., Gould, M., Bickenbach, J.E., 2003. Environmental barriers and disability. Journal of Architectural and Planning Research 20 (1), 29–37.

Guersey, K., Nicoli, M., Ninio, A., 2007. Convention on the Rights of Persons with Disabilities: Its Implementation and Relevance for the World Bank. Social Protection Discussion Paper No. 0712. The World Bank, Washington, DC, pp. 45–46. http://siteresources.worldbank.org/SOCIALPROTECTION/Resources/SP-Discussion-papers/Disability-DP/0712.pdf (accessed 14.04.13.).

Hughes, K., Bellis, M.A., Jones, L., Wood, S., Bates, G., Eckley, L., et al., 2012. Prevalence and risk of violence against adults with disabilities: a systematic review and meta-analysis of observational studies. The Lancet 379 (9826), 1621–1629.

Ingram, D., 2006. Antidiscrimination, welfare, and democracy: towards a discourse-ethical understanding of disability law. Social Theory and Practice 32 (2), 213–248.

Jones, M., Marks, L.A.B., 2001. Law and people with disabilities. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Elsevier Science Ltd, Philadelphia, PA, pp. 8475–8480.

Kanter, A., 2007. The promise and challenge of the United Nations Convention on the rights of persons with disabilities. Syracuse Journal of International Law & Commerce 34 (287), 291.

Kayess, R., French, P., 2008. Out of the darkness into the light? Introducing the convention on the rights of persons with disabilities. Human Rights Law Review 8 (1), 1–27.

Kim, C.K., 2011. Comparative perspectives on disability employment policy. International Review of Public Administration 15 (3), 27–35.

Logue, L., Blanck, P., 2010. Race, Ethnicity, and Disability: Veterans and Benefits in Post-Civil War America. Cambridge University Press, New York, NY.

Lord, J.E., Stein, M.A., 2008. The domestic incorporation of human rights law and the United Nations convention on the rights of persons with disabilities. University of Washington Law Review 83, 449.

Lord, J.E., Raja, Samant D., Blanck, P., 2012. Beyond the orthodoxy of rule of law and justice sector reform: a framework for legal empowerment and innovation through the Convention on the rights of persons with disabilities. In: Cisse, H., Muller, S., Thomas, C., Wang, C. (Eds.), The World Bank Legal Review, Legal Innovation and Empowerment for Development, vol. 4. The World Bank, Washington, DC, pp. 45–66.

Muller, J.F., 2011. Disability, ambivalence, and the law. American Journal of Law and Medicine 37 (4), 469–521.

Optional Protocol to the Convention on the Rights of Persons with Disabilities, opened for signature March 30, 2007, 46 I.L.M. 433.

Perlin, M.L., Szeli, E., 2012. Mental health law and human rights: evolution and contemporary challenges. In: Dudley, M., Siolve, D., Gale, F. (Eds.), Mental Health and Human Rights: Vision, Praxis, and Courage. Oxford University Press, Oxford, pp. 80–94.

Piggott, L., 2011. Prosecuting disability hate crime: a disabling solution? People, Place & Policy Online 5 (1), 25–34.

Quinn, G., 2009. Resisting the 'temptation of elegance': can the convention on the rights of persons with disabilities socialize states in rights behavior? In: Quinn, G., Arnardóttir, O.M. (Eds.), The UN Convention on the Rights of Persons with Disabilities: European and Scandinavian Perspectives. Martinus Neijhoff, Leiden, pp. 215–256.

Schur, L., Kruse, D., Blanck, P., 2013. Are People with Disabilities Sidelined or Mainstreamed? Economic, Political, and Social Inclusion. Cambridge University Press, New York.

Shakespeare, T., 2012. Disability, hate crime and violence. In: Roulstone, A., Mason-Bish, H. (Eds.), The Lancet, vol. 380 (9845), p. 878.

Stein, M.A., 2007. Disability human rights. California Law Review 95, 75–121. http://scholarship.law.wm.edu/cgi/viewcontent.cgi?article=1285&context=facpubs (accessed 9.04.13.).

Stein, M.A., Stein, P.J.S., 2007. Beyond disability civil rights. Hastings Law Journal 58, 1203–1240. http://digitalcommons.ilr.cornell.edu (accessed 12.04.13.).

Stein, M.A., Lord, J.E., February 27, 2012. Forging effective international agreements: lessons from the UN Convention on the Rights of Persons with Disabilities. In: Heymann, J., Cassola, A. (Eds.), Making Equal Rights Real: Taking Effective Action to Overcome Global Challenges. Cambridge University Press, Cambridge.

Szmukler, G., 2010. How Mental Health Law Discriminates Unfairly Against People with Mental Illness, Online Video, Gresham College. http://www.gresham.ac.uk/lectures-and-events/how-mental-health-law-discriminates-unfairly-against-people-with-mental-illness (accessed 11.04.13.).

The International Network of Women with Disabilities, 2011. Violence against women with disabilities. In: Barbara Faye Waxman Fiduccia Papers on Women and Girls with Disabilities. Center For Women Policy Studies, Washington, DC. http://www.centerwomenpolicy.org/programs/waxmanfiduccia/2011OnlineSeriesBarbaraWaxmanFiduccia.asp (accessed 14.05.13.).

United Nations, 2011. Compilation of Legislative Measures Undertaken in the Implementation of the Convention on the Rights of Persons with Disabilities: 2011 Update (CRPD/CSP/2011/CRP.5). Available at: http://www.un.org/disabilities/documents/COP/crpd_csp_2011_crp.5.doc (accessed 24.05.12.).

World Health Organization and World Bank, 2011. World Report on Disability. http://www.who.int/disabilities/world_report/2011/report/en/ (accessed 25.07.12.).

Relevant Websites

http://globalforumljd.org – Bringing Together Distinguished Legal Scholars, Practitioners, and Advocates on Current Issues in Disability Law and Policy, Global Forum on Law, Justice and Development, Community of Practice on Access to Justice for Persons with Disabilities.

http://www.ohchr.org/en/hrbodies/crpd/pages/crpdindex.aspx – Committee on the Rights of Persons with Disabilities.

http://https//nfb.org/law-symposium – Official UN Website on Persons with Disabilities. Jacobus tenBroek Disability Law Symposium.

http://www.un.org/disabilities – Official Website for the Conference of States Parties. United Nations Enable.

http://www.un.org/disabilities/default.asp?id=1433 – Official Website for the CRPD Committee. Conference of States Parties to the CRPD.

Law and Risk

Jonathan R Nash, Emory University Law School, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Bora, volume 12, pp. 8470–8474, © 2001, Elsevier Ltd.

Abstract

The issue of risk regulation arises in several areas of law. Risk regulation figures prominently in the area of environmental law, which deals with risks created by discharges of pollution. Although it is often not acknowledged, risk also has a role in criminal law, including (as discussed in greater detail below) how to balance the risk of convicting an innocent person against the risk of releasing a guilty one. This article will first consider risk in contradistinction to uncertainty, and will canvass various types of risks that societal actors face. Second, it will discuss private and public responses to risk. Third, it will examine types of public regulation enacted in response to risk. The fourth part will discuss the role various branches of government play with respect to risk regulation. Fifth, the article will consider paradigmatic risk analysis and risk management, the role of experts, and the role of public opinion in setting a risk regulatory agenda. Finally, it will discuss the precautionary principle, a normative principle grounded in concerns over risk that has been ascendant especially in the international legal arena.

The issue of risk regulation arises in several areas of law. Risk regulation figures prominently in the area of environmental law, which deals with risks created by discharges of pollution. Although it is often not acknowledged, risk also has a role in criminal law, including (as discussed in greater detail below) how to balance the risk of convicting an innocent person against the risk of releasing a guilty one.

Our discussion will first consider risk in contradistinction to uncertainty, and will canvass various types of risks that societal actors face. Second, it will discuss private and public responses to risk.

Our discussion will then focus on public regulation in response to risk. The fourth part will discuss the role the three branches of government play with respect to risk regulation. Fifth, our discussion will consider paradigmatic risk analysis and risk management, the role of experts, and the role of public opinion, in setting a risk regulatory agenda. Finally, it will discuss the precautionary principle, a normative principle grounded in concerns over risk that has been ascendant especially in the international legal arena.

What Is Risk?

Risk arises when there are multiple outcomes to which one might be subject. No single outcome will occur with certainty. Rather, each outcome – and therefore the payout associated with it – will arise with a certain probability.

Economists contrast 'true risk' with uncertainty. When we face risk, we do not know with certainty the outcome, but we do know the probability distribution of possible outcomes. In contrast, uncertainty arises when we do not even know the probability distribution of the possible outcomes (Knight, 1921: pp. 19–20). Because of limitations on our knowledge and understanding, global warming is often said to present great uncertainty, rather than simply risk. In practice, it may sometimes be difficult to know whether one really faces risk or uncertainty; many legal discussions conflate the two concepts.

The fact that, in a situation of risk, we know the probability distribution of possible outcomes means that we can calculate the 'expected value' of the risky scenario. The expected value equals the sum, for all possible outcomes, of the product of (1) the payout (whether positive or negative) from that outcome and (2) the probability that that outcome will come to pass (Nash, 2013: p. 1285). This cannot be done in the face of uncertainty.

Risk may arise in many different contexts. Engaging in certain activities may give rise to risks: For example, driving subjects drivers, passengers, and pedestrians to the risk of bodily harm, while investing in stocks and bonds may subject investors to economic risk. Economic cycles may subject businesses to risks. Contact with certain toxic substances may subject people to health risk. Nature in its extremes (e.g., hurricanes, droughts, or floods) may also put certain people at risk, if they live in the affected geographic area.

Note that different societies may assess the extent and importance of the same risk very differently. For example, the EU generally sees genetically modified organisms to pose a greater risk to human health than does the US. On the other hand, the US takes a more cautionary approach with respect to the threat of terrorism than does the European Union (EU) (Wiener, 2003: pp. 225–230). Also, in the US but less so the EU, the risk of allowing a guilty person to go free is considered more acceptable than is the risk of allowing an innocent person to be imprisoned or put to death (Nash, 2012: p. 199). Indeed, discussions and analyses of risk often raise complicated questions of morality (Ericson and Doyle, 2003: pp. 1–9).

To be sure, these differences in approaches to different risks appear in formal law. However, it is very likely that they indicate the existence of, and indeed emanate from, different cultural conceptions of what is risk and how best to respond to it. Indeed, recent studies confirm that, even within a single country, an individual's cultural values inform that individual's perceptions of risk (see, e.g., Kahan and Braman, 2006).

How Can the Law Deal with Risk?

The law takes different approaches to the various kinds of risk that societal actors face. The first option is to do nothing at all. This leaves the risk in the hands of those who bear it. In turn, those who bear risk face a number of alternatives. The first is for the societal actor to accept some combination of (1) reducing his or her (or its) exposure to the risk in question, and (2) bearing the downside of the risk if that comes to pass. Another alternative is for the societal actor to obtain insurance from a private provider. An insurer agrees to accept an insured's risk in return for payment based on the expected value of the risk and some premium. An insurer is able to take on risks from numerous insureds provided that the risks are sufficiently diversified; risks that are diversified are highly unlikely to produce extreme negative outcomes at one time (Masur and Nash, 2010: pp. 417–418). (Note that private insurers are also subject to government regulation.)

A second government option is to require societal actors who face a certain kind of risk to show the financial ability to cover a risk should a bad outcome in fact occur. For example, the US federal Resource Conservation and Recovery Act and the regulations thereunder require owners and operators of facilities that treat, store, or dispose of hazardous wastes to demonstrate 'financial assurance' for – that is, that there will be sufficient funds to cover – closure and postclosure care of their facilities, the clean up of any accidental releases of hazardous materials, and compensation to third parties for any resulting injury or property damage. Financial assurance may come in different forms, including bonds and private insurance (see *Code of Federal Regulations*, tit. 40, pt. H). In other cases, the law specifically calls for private actors to obtain private insurance. This has generally been the case for US automobile drivers and owners of nuclear plants, and now in the US is the case for health insurance under the Affordable Health Care Act.

A third government option is to provide government-backed insurance. This may come to pass where private insurers are loath to insure the risk in question, perhaps because the risk is nondiversified. For example, as a prerequisite to obtaining a mortgage backed by a federal agency on property located in a flood zone, one has to obtain government-issued flood insurance.

A fourth government option is to mandate disclosure of information pertinent to the risk – perhaps both as to the likely outcomes and the scope of possible downside scenarios. This empowers societal actors to make more knowledgeable decisions about whether to expose themselves to risks, and about those risks to which they ought to accept, and decline, exposure. The US federal government has taken this approach with respect to investment risk under the federal securities laws (Nash, 2010: p. 1030).

A fifth government approach is to use ex post liability regimes to create incentives to reduce risks. For example, tort law often imposes absolute liability on property owners who use their property in inherently risky ways.

A sixth approach is government regulation that tries to limit risk ex ante. These regulations are often promulgated by government agencies. Examples include environmental, health, and workplace safety regulations. The ex ante regulatory approach to risk has grown substantially over the past few decades (Section Paradigmatic Regulatory Risk Analysis and Management below discusses how regulators deal with risk.)

A final option is to have the courts assess a risk and then enter a ruling that deals with the assessed risk according to a governing legal rule. As one example, criminal courts in common law countries generally hold bail hearings and decide whether to grant bail – and, if so, at what amount to set bail – based upon the defendant's risk of flight or danger to the community. For another example, in deciding whether to grant a plaintiff's request for a preliminary injunction, a court will consider (among other things) the risk of irreparable harm to the plaintiff if the injunction is not granted (Nash, 2013: p. 1317).

Branches of Government and Risk Regulation

The US system typically divides government into legislative, executive, and judicial branches. What role might we expect each branch to play in risk regulation?

The legislative branch conceivably can put in place ex ante procedures, or ex post liability rules that create ex ante incentives, regarding risk. In practice, however, the legislature may be reluctant to enter the risk arena directly – perhaps especially with respect to ex ante procedures – fearing that a lack of expertise may generate ineffective and/or undesirable laws. Instead, the legislature may see fit to delegate power to an agency (either within the executive branch or independent of the political branches) to regulate as it sees fit (within the confines of legislation).

In turn, agencies that receive regulatory power from the legislature then undertake risk regulation. The discussion in Section Paradigmatic Regulatory Risk Analysis and Management below details how agency regulators typically proceed with risk regulation.

At the federal level, regulatory action by executive branch agencies is also subject to the strictures of Executive Orders issued by the President (see, e.g., Executive Order 12866, Regulatory Review, 58 Fed. Reg. 51735 (30 September 1993) (Clinton era Executive Order establishing principles for, among other things, risk assessment)). Compliance with Executive Orders is overseen by a unified executive branch regulatory office within the White House – the Office of Management and Budget's Office of Information and Regulatory Affairs (see, e.g., Percival, 1991: pp. 182–184).

At times, decisions that may have impacts on risk must be made quickly by agencies charged with enforcement. For example, police may have to decide whether to storm a location with the goal of apprehending a hostage-taker and free hostages, but at the same time putting at risk police and the hostages.

The judicial branch may use the common law process to put in place rules that implicate risk. More commonly today, the legislature may empower the courts to hear challenges to risk regulations issued by agencies.

Should Government Respond to Public Perceptions, or Expert Analyses, of Risk?

As the discussion in Section Paradigmatic Regulatory Risk Analysis and Management below elucidates, the approach to risk regulation that currently dominates in the US calls for

identification and measurement of the relevant risks should be undertaken by experts in the field. Another possibility, however, is for the government to respond not to expert assessments of risk, but rather to public calls for risk regulation.

Proponents of risk regulation based on expert risk assessment highlight how lay people tend to wildly overestimate some risks and to wildly underestimate other risks. For example, many believe that the public (or many members of the public) tend to overestimate the risks associated with vaccinations, but to underestimate the risks posed by radon gas in basements. Thus, proponents of expert-based risk regulation conclude, lay-based risk regulation will be based not on actual risks, but rather on *public perceptions* of risk. Along similar lines, they highlight that, insofar as they are not grounded in reason and evidence, people's risk perceptions are likely to be highly malleable, and thus subject to special interests groups' efforts to mold them as they see fit.

In response, proponents of a role for the public in risk regulation point to the importance of the democratic process in governance. For one thing, government ought to be responsive to people's needs. For another, government risk regulation that takes into account public opinion will have greater legitimacy.

One intermediate version of risk regulation that shows up in some US regulatory regimes is to allow the government agency to propose regulations as it sees fit, presumably based upon expert risk assessment. Then, however, the legislature empowers private societal actors to file suit in court to compel the regulator to add to its regulatory agenda (Biber, 2008: pp. 466–467). Note that such a regime may screen out situations where the societal actor lacks evidence of the existence of a real risk; the court may require such evidence before awarding the petitioning private actor the relief it seeks.

The EU also employs risk assessment methodology. Although EU risk assessments have generally tended to be more qualitative than quantitative (as compared to the US approach), there seems to be growing convergence between the two (see Wiener, 2004: pp. 82–83).

Paradigmatic Regulatory Risk Analysis and Management

This section provides a somewhat stylized summary of how government regulators respond to risk and uncertainty. (This discussion draws on the discussion in Nash (2008): pp. 498–504 and Nash (2012): pp. 192–197.) The paradigmatic regulatory approach to risk divides risk analysis into two components: risk assessment and risk management. Risk assessment goes to the question of the extent to which there is a risk. Risk management, in contrast, asks how best to respond to that risk. The justification behind this distinction is a perceived dichotomy between the types of questions that risk assessment is designed to address and those that risk management is designed to address. The question of the extent to which an activity or the use of a substance poses a risk is viewed as a matter that is appropriately resolved using the best available science. In contrast, the question of how best to respond to the existence of some risk is viewed as a policy question. Risk assessment thus seeks to isolate scientific inquiry from a balancing of policy concerns. According to theory, only after risk assessment reveals a risk should risk managers be called upon to decide what, if anything, to do about that risk.

The recommended dichotomy between risk assessment and risk management means that different actors should resolve these questions, even if both sets of actors are experts in their respective areas. Experts – with different areas of expertise – are called upon to perform each task.

Risk Assessment

The dominant US approach calls for risk assessment to be guided by science – often described in statutes as the 'best available' science (see, e.g., Executive Order 12866, Regulatory Review, 58 Fed. Reg. 51735 (30 September 1993) (Clinton era Executive Order establishing principles for, among other things, risk assessment)). Thus understood, risk assessment calls for analysis of risk both as scientifically as possible – and, implicitly, as objectively as possible.

Peer review further bolsters an agency's scientific analysis and conclusions. Courts, in turn, afford great deference to agency decisions that rest upon scientific analysis. One effect of this approach is to reduce the likelihood that government will respond to people's perceptions of risk, as opposed to what is – at least according to the dominant view – scientifically and objectively established as risky.

Risk Management

After risk assessment is complete, the analysis shifts to the more policy-centered stage of risk management. The fact that risk management is policy-driven does not mean that there is an absence of general guidelines to which some risk managers might turn (or at least that some people believe ought to guide risk managers).

First, some statutes (as well as US presidential executive orders) call on agencies to employ cost–benefit analysis in promulgating regulations. In the case of a risk-based regulation, this would call upon the regulator to consider the benefit of the regulation (which on a typical metric would be the expected value of the risk that will be avoided by virtue of the regulation) against the societal costs that the regulation would impose. Thus, for example, it might be costly for industry to comply with an environmental regulation. That cost is imposed on society (and indeed might largely be passed along to consumers in the form of higher prices or fewer products). On the other hand, the regulation will reduce health risks. A cost–benefit analysis approach asks regulators to monetize those health benefits, and to compare them to the costs the regulation would impose. A strict cost–benefit analysis would direct that the regulation not be implemented unless the benefits outweigh the costs. A weaker form of cost–benefit analysis, recognizing the difficulties in both identifying all of a regulation's benefits and monetizing those benefits, allows for the implementation of regulations provided that the anticipated costs do not swamp the anticipated benefits.

Second, risk managers may undertake 'risk–risk analysis' (Graham and Wiener, 1997). Risk–risk analysis calls upon regulators to balance the risks that a proposed regulation will reduce against the risks to which regulation itself may give rise.

It rests upon the recognition that in general one does not simply succeed in reducing risk; instead, one may often in practice 'trade off' a reduction in one risk for an increase in another. For example, in deciding whether to engage in a high-speed car chase to try to apprehend a fugitive, the police (implicitly) balance the risk of allowing the fugitive to go free against the risk of injury to the police, the fugitive, and innocent people that might result from the high-speed car chase (Nash, 2012: pp. 214–215). Indeed, the police administration in advance might offer general directives to police officers as to how to proceed in general in such situations, and police officers faced with a high-speed chase might have to apply those directives and make quick risk–risk analytic trade-offs quickly and under considerable duress.

As another example, an environmental regulation might, if implemented, reduce health risks. However, complying with the regulation will be costly to industry, perhaps because it increases the cost of the production process. That additional cost might result in higher prices to consumers (which itself might impose risks on society if, for example, the product in question is medicine) or might result in the loss of jobs (which also might impose additional risks on society because unemployed individuals might have inferior, or no, health insurance).

The Precautionary Principle

What if risk assessors cannot establish a risk with any real degree of certainty? Perhaps, for example, scientific understanding is too primitive to provide a solid basis for analysis. Here, some policy advocates – perhaps especially in the area of environmental law – advert to the precautionary principle. In essence, the precautionary principle calls for a cautionary approach in the face of risk and uncertainty, although as the discussion elucidates below, there is considerable contention over what exactly the principle means and demands.

Commentators have pointed out the precautionary principle's lack of clarity. Indeed, commentators have identified no fewer than 19 interpretations of the principle (Wiener, 2002: p. 1513). For example, a weak form of the principle authorizes government regulation where there is uncertainty. A stronger form *mandates* regulation under conditions of uncertainty. A stronger form still imposes a burden of proof on one who wants to introduce into use a substance, the health consequences of which would be uncertain to show that the substance will not pose unduly large risks (see Geistfeld, 2001).

Commentators have also criticized the precautionary principle as inherently self-contradictory, and therefore indeterminate. The argument proceeds that the precautionary principle's logic applies not only to the risk that exists absent regulation but equally to the countervailing risk to which regulation may give rise – that is, to the risk that regulating false positives may create. Thus, taken to its logical extreme, the precautionary principle would 'bring valuable activities to a halt', and would 'be impossible' to apply broadly (Wiener, 2003: p. 224).

In the face of this lack of certainty and criticism, some commentators have endeavored to rescue the principle from ambiguity and contradiction by identifying a 'heartland' area of clear applicability for the principle. The precautionary principle's 'heartland' extends to situations where (1) 'the risks of harm are uncertain'; (2) 'harm might be irreversible and what is lost is irreplaceable'; and (3) 'the harm that might result would be catastrophic'. (see Posner, 2004: p. 148, calling for "a modest version of the precautionary principle" in cases of catastrophic risk that would "plac[e] a thumb on the cost side of the cost-benefit analysis"); Sunstein, 2006: p. 894 ("It is possible to combine a concern about catastrophe with a focus on irreversible harm in a way that generates an Irreversible and Catastrophic Harm Precautionary Principle."; Nash, 2008: pp. 502–504. Sometimes, however, the risks on both sides may meet these criteria, in which case the 'heartland' applicability of the principle will provide little guidance.

At the international level, the principle has attained quite a bit of caché. Its importance is evidenced by its inclusion in numerous international statements of environmental law and international and multilateral environmental treaties. Some assert that the precautionary principle has acquired the status of customary international law. The frequency with which the principle is quoted in documents, laws, and treaties of the EU and its member states illustrates that the EU in particular believes strongly in the principle (Nash, 2008: p. 499).

With some exceptions, US domestic environmental law has not embraced the precautionary principle. This does not mean, however, that the United States does not act in accordance with the principle in other areas. The question is more one of identifying those risks that each sovereign deems to be worth addressing with a precautionary approach. Thus, for example, while the EU and its member states may follow the strictures of the precautionary principle more closely and more often than the United States with respect to environmental risks, the United States adopts a more precautionary approach more than does the EU with respect to the risk of convicting the innocent and of risks posed by terrorism (Sunstein, 2007: pp. 515–516; Wiener, 2003: pp. 225–230).

Conclusion

Legal responses to risk depend heavily upon what goals society seeks to obtain. One approach is that the law should try to minimize the risks that people perceive to exist. To the extent that, as growing empirical evidence shows, culture strongly affects risk perceptions, this will mean that different societies – and indeed different segments within a society – may call upon the law to do different things.

Another possibility is to turn to experts to determine which risks exist, and how best to deal with those risks. Paradigmatic risk analysis takes that approach. More nuanced risk analysis also recognizes that it is natural, but also generally wrong, to think that one can simply reduce a risk and thus unilaterally increase certainty and safety. Instead, the better approach is to concede that reducing one risk likely increases another; the question then becomes how best to balance competing risks.

While particular laws may respond to different risks in different ways, overarching principles offer some guidance and may provide hope one day of greater harmony in risk regulation. The precautionary principle is one such example. The

precautionary principle is quite prominent on the international legal stage, and in some parts of the world – perhaps most notably the EU. Despite its high profile, the principle suffers from a lack of clear definition, and from claims that it often can be interpreted to offer inherently contradictory advice. Still, commentators work to identify an acceptable 'core' of the principle, believing that it may one day develop into a consistent, useful guidepost for appropriate risk regulation.

See also: Administrative Law: The United States and Beyond; Crime Policy in the United States; Criminal Procedure: Adversarial and Inquisitorial Legal Systems; European Union Law; Insurance and the Law; Law and Technology; Regulation and Administration; Risk Assessment in Social Work; Risk, Sociology and Politics of; Risk: Social Psychological Perspectives; Social Insurance: Legal Aspects; Torts in Law.

Bibliography

Biber, Eric, 2008. Two sides of the same coin: judicial review of administrative agency action and inaction. Virginia Environmental Law Journal 26, 461.

Ericson, Richard V., Doyle, Aaron, 2003. Risk and morality. In: Ericson, Richard V., Doyle, Aaron (Eds.), Risk and Morality, University of Toronto Press, Toronto, pp. 1–10.

Geistfeld, Mark, 2001. Implementing the precautionary principle. Environmental Law Reporter 31, 11326.

Graham, John D., Wiener, Jonathan B., 1997. Confronting risk tradeoffs. In: Graham, John D., Wiener, Jonathan B. (Eds.), Risk Vs. Risk: Tradeoffs in Protecting Health and the Environment, p. 1.

Kahan, Dan M., Braman, Donald, 2006. Cultural cognition and public policy. Yale Law and Policy Review 24, 147.

Knight, Frank H., 1921. Risk, Uncertainty, and Profit. Hart, Schaffner & Marx; Houghton Mifflin, Boston, MA.

Masur, Jonathan S., Nash, Jonathan R., 2010. The institutional dynamics of transition relief. New York University Law Review 85, 391.

Nash, Jonathan R., 2008. Standing and the precautionary principle. Columbia Law Review 108, 494.

Nash, Jonathan R., 2010. Null preemption. Notre Dame Law Review 85, 1015.

Nash, Jonathan R., 2012. The supreme court and the regulation of risk in criminal law enforcement. Boston University Law Review 92, 171.

Nash, Jonathan R., 2013. Standing's expected value. Michigan Law Review 111, 1283–1335.

Percival, Robert V., 1991. Checks without balance: executive office oversight of the environmental protection agency. Law and Contemporary Problems 54, 127.

Posner, Richard A., 2004. Catastrophe: Risk and Response. Oxford University Press, New York.

Sunstein, Cass R., 2006. Irreversible and catastrophic. Cornell Law Review 91, 841.

Sunstein, Cass R., 2007. On the divergent American reactions to terrorism and climate change. Columbia Law Review 107, 503.

Wiener, Jonathan B., 2002. Precaution in a multi-risk world. In: Paustenbach, Dennis D. (Ed.), Human and Ecological Risk Assessment: Theory and Practice, John Wiley & Sons, p. 1509.

Wiener, Jonathan B., 2003. Whose precaution after all? A comment on the comparison and evolution of risk regulatory systems. Duke Journal of Comparative and International Law 13, 207.

Wiener, Jonathan B., 2004. Convergence, divergence, and complexity in US and European risk regulation. In: Vig, Norman J., Faure, Michael G. (Eds.), Green Giants? Environmental Policies of the United States and the European Union, MIT Press, Cambridge, p. 73.

Law and Society

Lora M Levett and Adina M Thompson, University of Florida, Gainesville, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of law and society is characterized by a multidisciplinary approach to understanding the interaction between legal institutions and our everyday lives. Researchers may use a variety of theoretical approaches (e.g., conflict theory, feminist legal theory, functionalism, critical race theory) and methodologies (e.g., survey research, experimental methods, historical analyses, comparative approaches) to study the intersection of law and society. Many areas of study fall under the broad umbrella of law and society, including (but not limited to) exploring legal actors in the criminal justice system, status differences, attitudes about the law and examining assumptions underlying legal decisions.

Introduction and Overview

Law and society encompasses a variety of approaches to studying the interaction between legal systems and our society. Scholars who study law and society typically come from sociology, psychology, anthropology, law, economics, political science, and other fields. Thus, the study of law and society is an interdisciplinary, multifaceted endeavor, with no unifying theory or methodological approach, although research in law and society is generally empirically and theoretically driven (Vago, 2011).

Professional Organizations

There are a variety of professional associations that provide an academic and professional home for law and society research and researchers. Among them are the American Psychology-Law Society, the Society for Empirical Legal Studies, the American Law and Economics Association, the Association for Political and Legal Anthropology, and the flagship law and society organization, the Law and Society Association (LSA). LSA was founded in 1964 and currently has approximately 1600 members, an annual conference, journal (*The Law and Society Review*), and newsletter. However, law and society research may appear in a variety of journals, including the *Australian Journal of Law and Society*, the *Canadian Journal of Law and Society*, the *International Journal of the Sociology of Law*, the *Journal of Empirical Legal Studies*, the *Journal of Law and Society*, the *Journal of Law and Policy*, the *Journal of Law and Social Inquiry*, the *Journal of Social and Legal Studies*, the *Journal of Law and Human Behavior*, and *Law and Contemporary Problems*, among others (Freidrichs, 2012).

Within LSA, members can belong to a variety of collaborative research networks (CRNs), which are designed around a research area of interest to encourage international and interdisciplinary research collaboration. There are currently 44 CRNs in LSA, including CRNs on African Law and Society; Civil Justice and Disputing Behavior; Law and Health; Law and History; Lay Participation in Legal Systems; Legal Geography; Gender, Sexuality and Law; Islamic Law and Society; Teaching in Law and Society; and many others. For a full list of CRNs, descriptions, and contact information for the CRN coordinators, see www.lawandsociety.org.

Research

Law and society scholars study the law as it relates to other social institutions and entities. That is, law and society scholars ask research questions exploring how law affects society's social structures and systems, relationships between individuals and between individuals and institutions, and institutions within society. Conversely, law and society scholars also explore the reverse relationships: how our society's social institutions, structures, relationships, systems, legal actors, and individuals influence the creation of our legal system, law, and policies. Scholarship may focus around the functions and dysfunctions of law in society; the development and impact of law and lawmaking (judicial, legislative, and administrative lawmaking); the actions, roles, and decisions of various legal actors and groups (judges, attorneys, police officers, juries), describing legal systems and the role of legal systems in a larger society, the globalization of law, and the impact of policies and reforms on society.

Given the diversity in research interests, methodological approaches, and theoretical perspectives, the unifying theme is that law and society scholars recognize the importance of an interdisciplinary approach to understanding and studying the law (www.lawandsociety.org) and share the general assumption that the law has to be considered as a product of our collective historical, political, economic, cultural, and psychological contexts. In turn, the law also can impact each of these areas (Friedrichs, 2012).

We will first review a selection of functions of law that law and society scholars may study. Then, we will review major theoretical perspectives scholars use in studying law and society, followed by a sample of the research areas law and society scholars explore. We should note that a complete review of theory and research is not practical given the constraints of this article and the broad nature of the field, nor is it the intention of this article. Our discussion of theoretical perspectives and research areas is intended to provide a sample of the theory and research in law and society scholarship.

Functions of Law

Law and society scholars highlight many functions of law, and there is not general agreement in the field on a complete list of

functions of law or the importance of identified functions (Vago, 2011). There are both positive and negative functions of law (Friedrichs, 2012). From a positive viewpoint, the law maintains order, settles disputes, allows for social change or catharsis, promotes justice, promotes desirable social values, and administers practical needs to a society. From a negative viewpoint, the law can function to maintain the status quo (even if the status quo is outdated or inconsistent with a society's values), divide people, exploit people, promote conflict and violence, promote oppression and greed, or further the power of the state (Friedrichs, 2012).

The literature has focused mainly on the positive functions of law as a mechanism for social control, dispute settlement, and social change (Vago, 2011). As one of many mechanisms for maintaining social control, the law is a method used by society to maintain order and delineate which behavior is acceptable and unacceptable (Vago, 2011). The law is a formal means of social control, meaning that it has explicit rules for behavior, including sanctions for those who violate those rules and people who create, enforce, and interpret the rules (Vago, 2011).

As a mechanism for dispute settlement, the law provides a formal forum in which individuals or organizations, including governmental entities, can solve conflict in a formal, orderly manner. Resolving disputes in the legal system is a transition from the two parties resolving the dispute among themselves (e.g., by choosing to avoid the situation or informally compromise) to using a third party who is outside the original dispute to help resolve the conflict (Vago, 2011). Most disputes are resolved without the help of the legal system. Dispute resolution within the legal system typically involves resolving the legal issues in the conflict rather than addressing underlying causes, which does not necessarily reduce the tension between the parties (Vago, 2011).

The law also functions to aid and create social change. Social change refers to change in our behavior indicating that people engage in activities or relationships that are different from their previous activities (Vago, 2011). Social change can be a product of many different factors that may interact with each other, including law and legal culture, political climate, economic factors, technological innovation, and structural constraints, among others (Vago, 2011). The law can both produce social change and be an object of social change (Freidrechs, 2012). The law can have both an indirect and direct influence in producing social change. In having an indirect influence, the law can change the structure of our social institutions, which in turn shapes our society. For example, requiring citizens to attend school improved the quality of workforce. The improved workforce produced social change by increasing the industrialization of society (Vago, 2011). The law can also have a direct effect in producing social change. For example, enacting prohibition laws in the United States was intended to directly shape whether people drank alcohol (Vago, 2011). Many of the various functions of law were developed and are studied in the context of several of the theoretical perspectives in law and society.

Theoretical Perspectives

Law and society scholars tend to be influenced by a broad variety of theoretical perspectives, which is understandable given the interdisciplinary nature of the field and the diversity in the areas of study. Our review of theoretical perspectives will begin with a discussion of two ideal conceptions of society that provide a context for the discussion of theorists and contemporary theories and intellectual movements in law and society that follow.

Theoretical Dichotomy

Conflict and consensus are the two general, oppositional theoretical orientations of law and society scholarship. Both are idealized typologies of society (Vago, 2011). Conflict theory suggests that law is a weapon in social conflict (Turk, 1976) and is based on the notion that social change and conflict are ubiquitous, that all facets of society contribute to social change, and that some members of society are repressed or constrained by other members of society (Treviño, 2010). Conflict theory posits that members of interest groups will attempt to use the law to their advantage and to the disadvantage of members of opposing groups. Such conflict may be economic, political, moral, ideological, or religious. Power in society is expressed as either elitist conflict, whereby one group (for example, the economically advantaged) consistently remains in power, or group conflict, whereby two or more groups (for example, the pro-choice and pro-life groups) trade power over time. From a conflict perspective, law arises out of a desire to maintain power. Conflict theory is supported in the writings of scholars, such as Karl Marx, Max Weber, Richard Quinney, Donald Black, and Austin Turk, and in the tenants of the critical legal studies (CLS), critical race theory (CRT), and feminist legal theory (FLT) movements.

Consensus theory differs from conflict theory in that it views the various facets of society as a homeostatic system of interrelated parts. Law thrives when the majority of people agree on its execution. Building consensus therefore builds a healthy, cohesive society. Whereas conflict theory views the law as a weapon, consensus theory suggests that law, "serves to mitigate potential elements of conflict and to oil the machinery of social intercourse" (Parsons, 1962: p. 58). The law must have integrity so that it remains pertinent, and it must serve the interests of the people. If law fails to serve the people, the law will lose its legitimacy. Consensus theory is supported in the writings of scholars such as Emile Durkheim, Talcott Parsons, and Roscoe Pound and in the tenants of the functionalist movement.

Individual Theorists

Emile Durkheim (1858–1917), a French sociologist and consensus scholar, stated that sociology is the study of social facts. Social facts are methods of thinking, feeling, and acting that are coercive in nature (Treviño, 2010). Law is the preeminent social fact because it is enduring and because a lawless society is undesirable. Adherence to the law is cultivated through solidarity, which may be either mechanical or organic. Mechanical solidarity exists in small, homogenous societies. Solidarity exists because people are similar and share belief systems. Organic solidarity exists in larger, postindustrial, differentiated, and heterogeneous societies. It is a more evolved state whereby individuals rely on one another due to the division of labor (Treviño, 2010). Mechanic and organic solidarity

correspond with repressive and restitutive law. Repressive law, observed in societies organized based on mechanical solidarity, delivers sanctions designed to punish because of community outrage at having violated an agreed-upon standard. Restitutive law, visible in societies organized based on organic solidarity, is meant to provide conciliatory solutions rather than punitive sanctions.

Max Weber (1864–1920) was a German sociologist who looked to religious, political, and economic influences to define and explain the law. Law, Weber states, can be either rational or irrational. He distinguishes four types of legal thought – formal irrational, substantive irrational, substantive rational, and formal rational – and notes that formal-rational legal thought is unique to the Western legal system (Treviño, 2010). Weber defines a law as a convention that is subject to three criteria – external pressures to comply, pressures that involved coercion or force, and officials in place to apply that coercion or force. If one or more of these criteria are missing, the convention cannot be considered a law. An additional Weberian contribution to law and society is his three forms of political authority – charismatic authority, or that which is innate in a leader such as a prophet; traditional authority, or that which is inherited by a leader such as a king; and rational-legal authority such as authority derived from statute as in the case of an elected official (Treviño, 2010).

Donald Black (1941–present) argues that legal scholars confuse the science of studying the law with the policy implications of legal research and the effectiveness of the laws in question (Treviño, 2010). He states concisely that law is governmental social control (Black, 1976). Deviant behavior is conduct that is subject to that social control. The more social control applied to a behavior, the more deviant it is considered. Law, according to Black, is quantifiable and can be measured. Law exists in accusatory or remedial styles. Accusatory styles of law may be either penal or compensatory, and remedial styles of law may be either therapeutic or conciliatory. More than one style may exist in a single legal conflict. Black states that the amount of law in a society is inversely related to other forms of social control in that society (for example, religious or political control). A conflict theorist, Black suggests that discrimination is inherent in the law.

Karl Marx (1818–83) believed that the function of law was to uphold our economic system. He posited the term economic determinism, suggesting that the economic structure dictated the structure of the rest of society, including the legal system. Society, Marx suggested, is composed of two groups – the bourgeoisie, who are economically dominant and own the means of production, and the proletariat, who are economically subservient (Treviño, 2010). The bourgeoisie, he suggests, use the law to maintain their economic power over the proletariat. Marx predicts that once the proletariat is aware of their condition of subservience, they will rise up and overthrow the dominant class, causing the collapse of capitalism and creating a socialist society (Vago, 2011).

Roscoe Pound (1870–1964) suggests that the function of law is to harmonize and balance opposing interests in the name of social order (Pound, 1943). He recognizes three categories of interests – individual, public, and social – but identifies social interests as the most expedient way to evaluate an issue or idea. Law, therefore, must balance social interests. According to Pound, general security is the chief social interest, followed by security of social institutions, general morals, conservation of resources, progress, and other considerations (Pound, 1943).

Richard Quinney (1934–present) may be most famous for his text "The Social Reality of Crime" (2001). Crime, he suggests, results when actions are made illegal based on the desires of the ruling class. A conflict theorist, Quinney posits that,

1. crime is a definition of conduct;
2. crime is behavior that conflicts with the interests of society's powerful;
3. the powerful have the capacity to apply criminal definitions to actions;
4. those who act in manners that are contradictory to the desires of the powerful are more likely to be classified as criminal;
5. the larger the number of actions or behaviors that are classified as crime, the more crime will exist in society; and therefore
6. crime is the creation of criminal definitions, the application of those definitions, behaviors that conform to or contradict those definitions, and the creation of criminal conceptions (Quinney, 2001).

Contemporary Theories and Intellectual Movements in Law and Society

The functionalist approach in studying law and society is rooted in a consensus perspective (Barkan, 2009; Vago, 2011). The functionalist approach historically borrowed an analogy from biology to explore the role of law in society. In this analogy, the society is considered the organism as a whole and the systems within the society (e.g., the legal system) function to maintain that organism, similar to how the parts of the body function to maintain the organism (Barkan, 2009). In this perspective, one could examine both the structure of the society as a whole (i.e., the organism) in understanding how the systems in a society relate to one another and the function of the society, or the activities of each system within the society and the consequences of those activities for the society as a whole (Vago, 2011). Thus, to fully understand a society and how law functions within that society, one has to consider all the parts of the society and how those parts relate to and affect one another. According to this approach, the law (and other social systems) can have both manifest and latent functions. Manifest functions are those that are well understood by members of the society and the intended purpose of the social system (Vago, 2011). Latent functions are the unintended consequences of those manifest functions; they often are not formally recognized.

Criticisms of functionalism include the idea that functionalism is an oversimplification of the relationships between social systems and society as a whole (Vago, 2011). In addition, what might be functional for one group of people in society may be dysfunctional for another group, exposing a status difference between various groups in the population. However, functionalism continues to be one of the most popular viewpoints in studying law and society today (Vago, 2011).

The basic tenant of the CLS movement is that the law must be considered in light of the personal biases of the individual legal actors and the social context in which people operate (Vago, 2011). That is, because one can use precedent to argue each side of a case, the law becomes more of a product of the individual biases of the decision makers and the context of the individual case rather than a product of the precedents in a case (i.e., it is indeterminate; Barkan, 2009). Thus, even though law may seem to be neutral, it is not independent of our values, politics, economics, and social systems (Barkan, 2009; Vago, 2011). Thus, the law represents an ideology that maintains the status quo in society (Barkan, 2009). In addition to examining the basis of law, CLS scholars also examine legal education, as it contributes to the problems with law discussed above (Barkan, 2009). The CLS movement has been criticized for being controversial and cynical, and the ideas within the movement are still evolving; however, it provides a useful forum for examining the legal system (Vago, 2011).

FLT (or feminist jurisprudence) is rooted in the conflict perspective, and the source of the conflict in the theory is issues of gender, inequality, and the law. The law contributes to and reflects gender inequality in our society (Barkan, 2009). According to Vago (2011), there are three central, broad, and interrelated themes in feminist legal scholarship. The first theme addresses the male dominance in the legal profession and society as a whole and women's inequality that results from that dominance. The second theme is that the law itself reflects male ideas, ideals, and culture, and therefore, feminist legal theorists must create new law that is free of those characteristics that reinforce masculine ideals (Vago, 2011). For example, male culture may trivialize problems in which women are typically victims and shirk responsibility for those problems even though men are typically the perpetrators (e.g., sexual assault, sexual harassment, domestic violence; Vago, 2011). According to feminist legal theorists, the law reflects these masculine ideas about these types of incidents. The third broad theme reveals the gender biases inherent in the law itself, showing how the law is not a value-neutral institution.

Feminist legal theorists typically use three methods for exploring problems within the broader themes (Vago, 2011), and many feminist legal theorists use historical analysis within these methods (Barkan, 2009). The first method is asking the woman question, or exploring the law for gender implications and taking into account the experiences and perspectives of women (Vago, 2011). The second method is feminist practical reasoning, or identifying how women may think differently about the law compared to traditional legal reasoning and challenging those laws that claim they represent the voice of the community but in reality reflect masculine reasoning or culture (Vago, 2011). The third method is consciousness raising, or bringing to light the common experiences of women who have been adversely affected by legal principles, thereby challenging the legitimacy of those principles (Vago, 2011). Criticisms of FLT include that it is an oversimplification of the complex relationships in law and society that result in classes of people who are mistreated in the law (and that one has to consider class, race, and ethnic background in addition to gender; Vago, 2011). In addition, the literature is not clear on some issues, e.g., whether women should be treated similarly to men or whether women should be treated differently (but equally).

CRT is a movement in law and society that examines the intersection of race and ethnicity and the legal system using a variety of disciplinary perspectives (Barkan, 2009). As such, critical race scholars explore questions of oppression, discrimination, and inequality in the legal system and the legal profession (Vago, 2011). According to CRT, one cannot fully understand law without accounting for race (Barkan, 2009). CRT scholars acknowledge that racism is institutionalized in our current legal system and work to rectify the wrongs of racism in our society (Vago, 2011). The basic premise of CRT is that minorities or persons of color are oppressed in our current culture, and because of those oppressions, they are disadvantaged and experience the world differently than those who have not been oppressed (Vago, 2011). Some CRT scholars bring to light those experiences through narratives or storytelling to expose racism (Barkan, 2009). Others explore the structure of law as a source of institutionalized (and often unconscious) racism in society. Criticisms of CRT include that race is considered a macrolevel social construction, and the theory has not explored the construction of race on the interpersonal level (Vago, 2011). In addition, there is no general standard methodology or common tenants throughout CRT scholarship; however, this can be understood, given the theory is relatively new compared to other theories (Vago, 2011).

Areas of Study in Law and Society

Research in law and society is very broad, generally encompassing many disciplinary and methodological approaches. Law and society scholars may use time-series analyses that allow scholars to assess change in behavior over time or cross-sectional studies that allow researchers to examine differences in behavior or characteristics of different samples of the population, surveys, public opinion polls, ethnographic research, historical analysis, observational methods, survey analysis, or experiments (Friedrichs, 2012). It is beyond the scope of this article to provide a comprehensive review of all research in the area. Instead, our goal is to provide examples of areas of study and research under the wide umbrella of law and society. We will discuss legal actors, testing assumptions in the law, status differences in the context of law and society research, and examining attitudes about the law (Freidrichs, 2012).

The Legal Profession and Other Actors in the Legal System

Law and society scholars have devoted much scholarship toward researching the evolution, role, and decision-making processes of various legal actors in the criminal justice system, including attorneys, police officers, judges, and juries, among others (see Barkan, 2009; Vago, 2011 for reviews). For example, in one study, James Q. Wilson (1968) found that the policing agencies generally fell into one of three styles (although styles may overlap). In the watchman style, police emphasize the importance of maintaining order within the society and informally handle many violations in favor of keeping the peace. Watchman-style agencies tend to use the

most discretion in their duties relative to the other styles. The legalistic style can be considered the opposite of the watchman style, with officers using the least amount of discretion and treating every violation (even minor ones) as serious violations of the law. The service style combines the enforcement and order maintenance functions of the police, emphasizing community relations and using a moderate amount of discretion, separating more serious infractions from more minor offenses (Wilson, 1968).

Testing Assumptions in the Law

In making decisions about what behavior is acceptable or required, lawmakers may make assumptions about how people generally behave, make decisions, or will act in a given circumstance. These assumptions are not always based on science and can lead to negative outcomes. Law and society scholars have tested many of these assumptions in their research.

For example, researchers have examined the assumptions underlying the United States Supreme Court (USSC) decision in *Daubert v Merrell Dow Pharmaceuticals* (2003). In this decision, the USSC determined that judges would be the gatekeepers of scientific testimony in the courtroom, evaluating the science for relevance and reliability to determine admissibility. However, the USSC assumed judges were equipped to evaluate the reliability of science, an assumption that research has shown to be erroneous (see Kovera et al., 2002 for a review). Further, the USSC also referenced three safeguards (cross examination, judicial instruction on the burden of proof, and the presentation of contrary evidence, including opposing expert testimony) designed to help juries make good decisions about science in cases in which junk science was erroneously admitted. Research has examined these safeguards and found that the assumption that they will help jurors make a better decision about expert testimony is also erroneous (see Levett and Kovera, 2008 for an example). Thus, examining assumptions in the law may reveal opportunities to improve the effectiveness of the law by making it more consistent with how people actually behave.

Status Differences

Typically, law and society scholars who study status differences do so in the context of one of the theoretical perspectives discussed above. Scholars have explored the law in terms of differences in wealth, race or ethnicity, gender, power, or a host of other variables (Friedrichs, 2012). The scholarship typically examines how the law is a product of and enforces the inequality between groups of people, showing how law protects the interests of the more powerful group and oppresses the disadvantaged group.

One example of work examining status differences is Jeffrey Reiman and Paul Leighton's book *The Rich Get Richer and the Poor Get Prison: Ideology, Class, and Criminal Justice* (2009). Reiman and Leighton explore status differences among people of different classes in the criminal justice system. For example, the authors show how those things we most typically legally define as crime and treat the most seriously in the criminal justice system tend to be acts that are typically committed by the poor, and other crimes that could be considered just as serious (e.g., failing to make working conditions safe resulting in death and injury) are noncriminalized and typically perpetrated by more wealthy individuals (Reiman and Leighton, 2009).

Attitudes about the Law

Law and society scholars have devoted scholarship to studying how people think about the law, and in turn, how those attitudes affect their behavior. For example, in Tom Tyler's work, *Why People Obey the Law* (2006), he distinguishes between normative (what people believe is just) and instrumental (behavior that is a consequence of incentives or penalties) perspectives on why people follow the law. While much legislation is focused on creating law to affect behavior through an instrumental perspective, Tyler shows that a normative viewpoint may actually be more important. That is, if people see the law as just and appropriate, they will automatically feel motivated to follow the law, either because of personal morality or because they view the law and legal actors as legitimate (Tyler, 2006). In addition, if people perceive the process legal authorities follow in executing the law as fair (or the procedure as just), they will be more likely to see the law as legitimate and therefore comply with the law (Tyler, 2006). Thus, peoples' perceptions of the law can influence their propensity to follow the law.

Conclusion

As a field, law and society encompasses many areas of study, disciplinary approaches, and methods. Overall, the goal is to investigate how legal rules and systems affect our lives, and how society, in turn, can affect the formation and enforcement of those legal structures.

See also: Critical Legal Studies; Critical Race Theory; Discrimination and the Law; Disputes, Social Construction and Transformation of; Due Process in Law; Empirical Legal Studies; Expert Testimony; Feminist Legal Theory; Gender and the Law; Human Rights: Legal Aspects; Injustice: Legal Aspects; Islamic Law; Judges; Juries; Justice and Law; Law and Aging; Law and Economics: Empirical Dimensions; Law and Economics; Law and Everyday Life; Law and Geography; Law and Society: Development of the Field; Law as an Instrument of Social Change; Law, Sociology of; Law: Anthropological Aspects; Law: History of Its Relation to the Social Sciences; Law: The Socio-Legal Perspective; Lawyers: Social Organization of the Profession; Legal Culture and Legal Consciousness; Legal Issues: Public Opinion; Legal Process and Social Science: United States; Legal Scholarship; Litigation; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Natural Law; Negotiation and Bargaining: Role of Lawyers; Organizations and the Law; Paralawyers: Other Legal Occupations; Philosophy of Law; Plea Bargaining: A Comparative Analysis; Prosecution; Punishment: Social and Legal Aspects; Race and the Law; Remedies and Damages: Legal Aspects; Rights: Legal Aspects; Science and Law.

Bibliography

Barkan, S.E., 2009. Law and Society: An Introduction. Pearson, Prentice Hall, Upper Saddle River, NJ.

Black, D., 1976. The Behavior of Law. Academic Press, San Diego, CA.

Daubert, v Merrell Dow Pharmaceuticals, 1993. 509 U.S. 579.

Friedrichs, D.O., 2012. The Law in Our Lives, third ed. Oxford University Press.

Kovera, M.B., Russano, M.B., McAuliff, B.D., 2002. Assessment of the commonsense psychology underlying Daubert: legal decision makers' abilities to evaluate expert evidence in hostile work environment cases. Psychology, Public Policy, and Law 8, 180–200. www.lawandsociety.org (accessed 03.07.14.).

Levett, L.M., Kovera, M.B., 2008. The effectiveness of educating jurors about unreliable expert evidence using an opposing witness. Law and Human Behavior 32, 363–374.

Parsons, T., 1962. The law and social control. In: Evan, W.M. (Ed.), The Sociology of Law. The Free Press, New York, pp. 60–68.

Pound, R., 1943. A survey of social interests. In: Evan, W.M. (Ed.), The Sociology of Law. The Free Press, New York, pp. 69–82.

Quinney, R., 2001. The Social Reality of Crime. Transaction Publishers, NJ.

Reiman, J., Leighton, P., 2009. The Rich Get Richer and the Poor Get Prison: Ideology, Class, and Criminal Justice, tenth ed. Pearson, Upper Saddle River, NJ.

Trevino, A.J., 2010. The Sociology of Law: Classical and Contemporary Perspectives. Transaction Publishers, NJ.

Turk, A., 1976. Law as a weapon in social conflict. Social Problems 22 (3), 276–291.

Tyler, T.R., 2006. Why People Obey the Law. Princeton University Press, NJ.

Vago, S., 2011. Law and Society, tenth ed. Prentice Hall, NJ.

Wilson, J.Q., 1968. Varieties of Police Behavior. Harvard University Press, Cambridge, MA.

Law and Society: Development of the Field

Bryant G Garth, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article discusses the origins of the field of law and society (also called sociolegal studies), especially in the United States; relates it to 'legal realism' and to competition and cooperation between the fields of law and sociology; explores the symbiotic relationship of the field to the social welfare state; and then explains its more hostile relationship to the neoliberal state. The field grew very quickly in the 1960s in relation to the activist state, declined in relation to law and economics and the neoliberal state, and then has prospered again through globalization and the rise of interdisciplinary studies in law schools.

'Law and society' and 'sociolegal studies' refer to an interdisciplinary field represented by organizations such as the British Socio-Legal Studies Association and the US Law and Society Association, each of which developed out of a particular blending of law and social science that took place in the 1960s and 1970s in conjunction with the rise of the activist state. In the United States, a group of foundations came together to produce a field that began quite close to sociology but moved increasingly close to law. The power and prestige of this field ascended in the 1960s and 1970s, declined with the rise of economics as a challenge to the activist state, and then returned to prominence with globalization and the rise of interdisciplinary scholarship in US law schools.

Scholars who have sought to integrate law and social science have united under various banners, and the most important of these are 'law and society,' which began in the United States (Garth and Sterling, 1998), and 'sociolegal studies' (Campbell and Wiles, 1976), which began in Great Britain. More recently, 'empirical legal studies' in the United States is also used to describe those who combine social science and law.

There are scholars in many parts of the world who share these identities, but the field, even in Britain, owes much to origins in the United States. Its emergence in the United States relates to the fact that the legal profession has long held a dominant professional position in economic and state power, and social science disciplines in the United States developed very differently from their counterparts elsewhere. In the United States, the disciplines became more differentiated, more empirical, and more positivistic than in other settings. Those trained in the disciplines could challenge the hegemonic position of law in the United States, and this continuing challenge helped to produce the field of law and society. The relative success of this field in the United States made it attractive to others, including most notably the British and Canadians. The two largest organizations are the Law and Society Association (LSA) in the United States and the Socio-Legal Studies Association (SLSA) in Britain. The LSA has more than 1500 members, and the SLSA some 750. Both groups contain a diverse array of scholars, interests, and approaches. Common frames of reference include work on the legal profession, issues of gender and race, state regulation, courts and dispute resolution, and crime and criminal justice. These research topics define the mainstream of the field, which includes the full range of social science methods.

Recognizing the importance of the United States in defining the field of law and society is not inconsistent with seeing this field as part of an international phenomenon. First, there are many scholars in a number of countries, including Australia, India, Japan, Korea, and several countries of Europe and Latin America, who have been influenced strongly by, and who have contributed to, this field. The influence can be traced in part to the law and development movement in the 1960s and 1970s, which promoted educational interchange and a range of interdisciplinary work on the legal profession and the courts in particular. Notable examples include the work of Rogelio Perez Perdomo in Venezuela, Upendra Baxi and Rajiv Dhavan in India, and Setsuo Miyazawa in Japan. Much work in the law and society tradition continues to be produced by these individuals and others.

A second international influence came from the fact that participants in the law and development movement from the United States, such as Lawrence Friedman, Marc Galanter, and David Trubek, took their insights from abroad and used them to develop theories and models that could be applied to the United States. The best known of such approaches is Galanter's famous article on 'Why the haves come out ahead' (1974), which was in large part based on his experience in India.

More generally, a related but distinguishable field of 'legal sociology' took shape at roughly the same time that 'law and society' developed in the United States, and it continues to be important in Europe and elsewhere. The major institution is the Research Committee on the Sociology of Law, which is affiliated with the International Sociological Association and also meets with the US Law and Society Association every 5 years or so. One of the pioneers of legal sociology in the United States, William Evan, joined with Adam Podgorecki of Poland – then visiting in the United States – to take the initiative in 1962 to form the Research Committee. The first president was Renato Treves of Italy, who built the Research Committee as a distinct entity devoted to legal sociology. Again, while there have long been overlaps with law and society, the scholarship associated with the Research Committee is more 'European,' which can be defined as more theoretical and less empirical than that typically associated with law and society or sociolegal studies.

The development of 'law and society' in the United States occurred at almost the same time as the founding of the Research Committee, but the story begins with the legal realists, who used the growing prestige of social science and especially sociology in the 1920s to mount an attack against the formalism of the generation that controlled the legal academy at the time. Centered at Yale and Columbia Universities, the realists carried out empirical studies of the legal system and challenged the assumptions of those they labeled as 'formalists' – those who believed that the law was a self-contained system such that legal precedents actually dictated the results in legal cases. Among the prominent legal realists were Charles Clark, William O. Douglas, Jerome Frank, Underhill Moore, and Walter Wheeler Cook. Although the legal realist movement as such died by the end of the 1930s, its leaders had been absorbed into the mainstream of the legal profession, and the interest in social science had seemingly disappeared. The orientation of the realists, however, helped set the stage for later developments (see Legal Realism; Legal Formalism).

A gradual revival of the interest in law and social science after World War II generated the people and momentum to create the LSA in 1964. During this period, the Ford Foundation invested substantially in empirical projects that focused on the legal system. One major grant was to the newly established American Bar Foundation (ABF) to fund the Survey of the Administration of Criminal Justice, which was directed by Frank Remington of the University of Wisconsin. Another important grant was to the University of Chicago to support a variety of research projects, the best known of which was by Harry Kalven and Hans Zeisel on the American jury.

In the 1950s, one of the promoters of the legacy of legal realism, David Cavers, used his position as associate dean at Harvard Law School to influence a prominent alumnus, Walter E. Meyer, to provide in his will for the establishment of the Walter E. Meyer Research Institute of Law, which was to encourage law-trained researchers to use the tools of social science in their research. Once established in the late 1950s, the Institute made numerous grants over a 20-year period to help develop the field of law and society.

Even more than the Ford Foundation and the Meyer Institute, the Russell Sage Foundation's grants served to institutionalize these collaborative ventures in the 1960s. Prior to the time of the establishment of the LSA, itself promoted by the Russell Sage Foundation, the Foundation had already invested more than a million dollars in law and social science. The major investment began with a grant to the Center for the Study of Law and Society at the University of California, Berkeley, and other grants were made to establish centers at the University of Wisconsin (1962), Northwestern University (1964), and University of Denver (1964). Later in the 1960s, after the establishment of the LSA, the Meyer Institute and Russell Sage Foundation funded 6 years (1967–73) of summer institutes in Denver for legal academics in a program entitled Social Science Methods in Legal Education. The National Science Foundation program in law and social science began in 1972 and reinforced the idea that this set of topics comprised a distinctive field.

The 1960s and early 1970s represented a kind of golden age for law and social science. There was a strong relationship between what was happening in the production of knowledge and what was happening in the streets and in the civil rights movement outside of the academy. Social science provided a new professionalizing expertise that offered ways to manage the new social agenda. Sociological knowledge thus gained in value as a new generation helped to define a political agenda of concern with urbanization, race relations, poverty, and crime. This social science learning could be used to challenge the postwar complacency of law schools and the training and expertise of lawyers for governing the state.

Remaining followers of legal realism, including such figures as Cavers and Willard Hurst of the University of Wisconsin, used this rising prestige of the social sciences to renew their own attacks against legal traditionalism and formalism. As both a cause and a reflection of this challenge, a few philanthropic foundations made it their missions to build institutions that would fly the banner of this new sociological expertise. Their investments in learning and in building institutions succeeded in bringing a new expertise to law and to the activist state.

The period of relatively high prestige for scholars bridging law and social science was in retrospect relatively brief. The 1980s again brought a period of rapid social change, but social science generally, and 'law and society' in particular, declined in relative prestige. Indeed, just as law and society helped to build and legitimate the activist state (and the role of law in its construction), the competing movement of 'law and economics' provided much of the learning and legitimacy for the later turn away from social welfare and state activism.

The way that law and society linked with the activist state, and the tensions within the law and society field, can be seen in the stories of two of the Russell Sage centers – Berkeley and Wisconsin. The Russell Sage Foundation grant to establish the Law and Society Center in Berkeley in 1962 began a process that continued through the closing of the School of Criminology and the establishment of the Jurisprudence and Social Policy program at the law school in the late 1970s. Phillip Selznick was the key individual in developing the idea of the Center. Selznick's early work focused on organizations as institutions (Selznick, 1949), but he became increasingly interested in law and its relation to the normative lives of organizations. In 1969, Selznick et al. published *Law, Society, and Industrial Justice* (1969), written in collaboration with Philippe Nonet and Howard Vollmer and published by the Russell Sage Foundation. Drawing on his studies of institutions, Selznick and his collaborators examined the general institutional setting of labor–management relations, asking how 'emergent law' – including due process – could be brought to and realized in the industrial setting as 'industrial justice.' Paralleling Selznick's own move from institutionalism toward legal and even jurisprudential issues, the book asked "what it means to legalize an institution, that is, to infuse its mode of governance with the aspirations and constraints of a legal order" (Selznick et al., 1969, p. 8).

Others who joined the Center followed in Selznick's footsteps, advancing the study of law as an institution; these scholars included Sheldon Messinger, Jerome Skolnick, and Jerome Carlin. Within the law school, there was relatively limited but growing support for the Center and for sociolegal studies, in particular from Geoffrey Hazard and Caleb Foote. A new cohort of law professors, led by Sanford Kadish, began to move much closer to the Center later in the 1960s. Carlin

initially decided to focus his research on solo practitioners, and the result was a pioneering account of that world, entitled *Lawyers on Their Own* (Carlin, 1962). He moved to the Bureau of Applied Social Research at Columbia University, where he produced another major work, *Lawyers' Ethics* (Carlin, 1966). He then moved to Berkeley to work at the Center and the Sociology Department. Beginning in 1963 at the Center, his first project explored the relationship between the legal system and the class system. The project, which also involved Messinger and Jan Howard, resulted in the well-known article in the first issue of the *Law and Society Review* on 'Civil justice and the poor' (Carlin et al., 1966), and a book on the same topic. The conclusions emphasized the unequal access to law by the poor and helped make the case for the redistribution of legal services. It fitted very well with the 'war on poverty' and the ferment that was building around new legal services programs. The research connected closely to the social and legal activism of the time.

Also in Berkeley, but less oriented toward the Center, was Laura Nader, one of the few women involved in the early years of LSA and also a creator (or re-creator) of the subfield of legal anthropology. Nader's Harvard dissertation was based on fieldwork in Mexico with a focus on law cases. She joined the Berkeley Anthropology Department in 1960 with a specialty in the anthropology of law, which she then helped to develop. She organized the first major conference on law and anthropology, which resulted in a book that helped define this emerging subfield of legal anthropology. She also organized her students into the Berkeley Village Law Project, which focused on villages and 'disputing,' and became very influential in law and society as well as in legal anthropology. Legal anthropology continued to be a very live field into the 1970s, contributing to the emerging interest in alternative dispute resolution.

Eventually the Center moved into the law school, in part because of the Berkeley administration's decision to close the School of Criminology. What was left of the criminology school was merged in 1977 with the Center for Law and Society under the auspices of the new Jurisprudence and Social Policy (JSP) program. The social scientists moved to the law school. The law school, for the most part, remained dominated by the legal mainstream, but the legacy of the origins in sociology remained evident – even in the jurisprudential component of the JSP program.

Wisconsin, in contrast to Berkeley and the other centers, already had a thriving interdisciplinary community in the 1950s. It centered in the law school and was built on Willard Hurst and the legacy of legal realism. The Hurst contribution can be traced to Lloyd Garrison, the dean at Wisconsin Law School during the 1930s. At Wisconsin, Hurst worked enthusiastically with Garrison to develop first-year teaching materials for a course on 'law in society,' and he also invested substantially in the cohort of young professors in the law school. The group built an identity that helped to constitute law and society.

Following Hurst's lead, Stewart Macaulay also moved toward the sociology of law and the attack on legal formalism. Hurst also encouraged Lawrence Friedman, and many others, to write about law in the specific context of the state of Wisconsin, which led to the publication of *Contract Law in America* in 1965 (Friedman, 1965). Macaulay and Friedman teamed up to produce one of the early sets of materials on 'Law and the behavioral sciences' (Friedman and Macaulay, 1969). This Hurst-inspired group expanded in the 1960s with the law school at the center.

One of the most important individuals was Harry Ball, a sociologist recruited initially to work on the ABF Survey of the Administration of Criminal Justice. Not only was he a key person in the Russell Sage grant, but also he was one of the leaders with Robert Yegge (dean at the University of Denver College of Law) and Richard (Red) Schwartz (the first editor of *Law and Society Review*) in establishing the LSA in 1964. Along with Ball, Friedman, Hurst, and Macaulay, the law and society contingent grew to include Joel Grossman (Grossman, 1965; Grossman and Grossman, 1971), Joel Handler (Handler, 1967; Handler and Hollingsworth, 1971), Herbert Jacob (e.g., Jacob, 1965), Stanley Katz (Katz, 1971), Jack Ladinsky, Frank Remington, and others.

The Wisconsin tradition built around Hurst, which can be characterized as that of 'law as the dependent variable' or, perhaps more accurately, 'law in context,' provides one enduring approach in the field of law and society. It is characteristic of the work of Friedman and Macaulay, for example, and others from the same group such as Joel Handler. To take an example from the early days of the 'war on poverty,' Handler criticized the proposition, identified with Yale's Charles Reich, that the legalization of welfare and the provision of due process hearings would be sufficient to deliver concrete gains for welfare recipients. Criticizing the faith in legal formalism, he suggested that a careful examination of social context could make sure that the law would deliver better substance. In the writings of Hurst and other followers, the idea that law is a dependent variable was not inconsistent with idealism or an active role for law. Educated lawyers aware of social and economic contexts could make law a much more effective instrument for social progress.

Wisconsin was able to maintain a critical mass within the law school, nourished by the other departments of the university. It was able to sustain a law and society tradition that self-consciously offered a different model than that found in the elite law schools, which remained outside of the LSA orbit. The group in Wisconsin identified its approach with the LSA organization, with Friedman, Macaulay, Handler, Jacob, and Marc Galanter (who came later) all serving as presidents, and Joel Grossman serving also as editor of the *Review*.

This University of Wisconsin approach – built around law professors – both complemented and competed with the more sociological and jurisprudential orientation in Berkeley. Wisconsin's center was dominated by the law school and the personality of Hurst, while Berkeley's center was initially dominated by sociology – especially Selznick. The law professors at Wisconsin developed their approach by emphasizing in their struggles with the legal establishment that law was the dependent variable. This strategy allowed them to use social scientists and social scientific tools to attack legal formalism and the legal establishment on behalf of the concept of 'law in context.' The idea, especially in Hurst's works, was that the context was necessary to make the law more effective and better able to solve social problems.

The Berkeley sociologists, in contrast, emphasized jurisprudence and legal philosophy, which they – in particular Selznick – could see that sociology lacked. Scholars of institutions, as Selznick argued, needed to bring law to those institutions. The sociologists also could emphasize their superior training for social research. The law professors, they suggested, were incapable of making serious contributions to important theoretical questions of concern to social science, which was necessary to solve social problems. When the Berkeley sociologists and the Wisconsin lawyers found themselves on similar terrain, however, the opposition tended to turn into one between law and social science. Depending on what position one took in this competition, therefore, the sociologists at Berkeley could be seen as 'merely sociologists,' lacking any real legal expertise, or they could be characterized as 'real scholars,' in contrast to 'atheoretical' or dilettante lawyers.

The relative positions can be seen in the approaches to legalization and procedural due process. While Handler, from the perspective of law and context, based his criticism of law professors such as Charles Reich on their neglect of the social factors that would limit the impact that welfare rights and hearings would have on the lives of the poor (Handler, 1966), Selznick's *Law, Society, and Industrial Justice* (Selznick et al., 1969) emphasized due process as "the primary source of concepts and doctrines used in bringing the rule of law to new settings" (p. 274). The different approaches relate to the fact that Handler was initially using social science to challenge legal formalism, while Selznick was initially using law to show its relevance to the study of social institutions.

The same basic contrast between law and sociology was also evident in the approaches of the Walter Meyer Institute, which was closer to law, and the Russell Sage Foundation, which was closer to sociology. The LSA was initially the creation of people who, in terms of this relative opposition, were much closer to Russell Sage and to sociology. Yegge was the only law professor among the major founders, and the Russell Sage Foundation provided the initial funds. The contrasting structural positions, however, should not obscure the close proximity of these players on the emerging field. These were the lawyers closest to social science and the social scientists closest to lawyers. Even at the personal level, the protagonists were all friends and collaborators in building interest in the field of law and society. They shared a position against legal and sociological traditionalism, but they disagreed about how to define the new field of law and social science – and the relative positions of lawyers and social scientists within that field. They tended naturally to want to define the field in terms of their own expertises and approaches.

By the mid-1960s, this general field of law and social science had not only developed a number of institutions and adherents, but also served to define and build much of the expertise of the state as the war on poverty began and added further resources to the field. Many of the people intersected with the new set of state programs, and indeed the relative success of the field is reflected in the career successes of those who developed this field. At the time of the development of the LSA, in short, the expertise of law and social science – including but not limited to the LSA – was gaining legitimacy and moving rapidly into the state.

The field that was constructed and the LSA itself gradually tilted in favor of law. By the late 1960s, law had incorporated enough social science to stave off the social science challenge. This shift in the center of gravity of the field of law and social science generally was also evident in the 1970s, in the relative power within the LSA of law professors and in particular professors linked to Wisconsin. The presidents of the LSA were almost all law professors linked to Wisconsin. A later indicator of the shift in the center of gravity occurred when the Berkeley Center for Law and Society was brought into the domain of the law school. Put another way, much of the investment in the field of law and social science in the 1960s ultimately found its way into the LSA and the law schools.

The relative value of social science expertise declined in the 1980s. A know-how that served to help construct an activist state seemed to lose some of its relevance. Academic investment in economics by the mid- to late 1970s tended to pay richer legal career dividends than investment in the disciplines that were built more into the LSA. Just as bright and ambitious people were drawn to social science in the 1950s, many were drawn to economics in an era when inflation and the state were considered the great enemies of progress. Economics seemed to define the problems and the solutions for the 1980s just as sociology did for the 1960s.

Nevertheless, once established, the LSA developed its own organizational dynamic. A second generation brought up within the LSA came to define the scholarship around the LSA as a 'field' distinct from law and from the social science disciplines. The link of the LSA to the progressive politics of the 1960s also helped to give the LSA an enduring progressive political imprint.

The next group of leaders in the LSA followed an approach similar to that of the Wisconsin pioneers. Rather than emphasizing 'law and context' as an antidote to optimistic progressivism, however, the new generation – many educated at Wisconsin by the pioneers – took aim at the politically leftist legal scholars of the critical legal studies movement and suggested that they needed more social science to understand and substantiate their leftist critiques. The 'Amherst seminar,' which produced much of the LSA leadership in the 1980s (including Austin Sarat, Susan Silbey, Sally Merry, Christine Harrington, Barbara Yngvesson, Patricia Ewick, and Kristin Bumiller), showed how the 'constitutive' nature of law prevented the social change that had been sought in the 1960s.

As an ongoing organization with both established scholarly tools and an openness to those looking for an identity to challenge the legal mainstream, the LSA continues to thrive as an organization many decades after it was founded. The tensions and hierarchies present at its creation, however, continue to play a prominent role in shaping the organization and the field that it represents. The establishment of the Society for Empirical Legal Studies in 2005 is in part a reflection of those tensions – carving out a place for those emphasizing quantitative research of practical use to lawyers and courts.

Late in the 1990s, in addition, the field renewed the commitment to internationalization that characterized it in the 1960s and 1970s. There is a new law and development movement anchored in places such as the World Bank that adds momentum to the internationalization of the field. With unprecedented educational interchange both in the social

sciences and in law, in addition, the US-centered field is opening up to many scholars from outside the United States. Much of the growth of the LSA in recent years, for example, has come from its international membership. The 'collaborative research network' of the LSA that focuses on East Asia, for example, is sufficiently large to have its own conferences in Asia.

Globalization and the rise of interdisciplinary legal studies have brought renewed importance to the field of law and society, also called sociolegal studies. There are competing organizations that have somewhat different orientations, but the different groups take into account each other's work and approaches.

See also: Critical Legal Studies; Law and Development; Law and Economics; Law as Constitutive; Law as an Instrument of Social Change; Law: History of Its Relation to the Social Sciences; Legal Formalism; Legal Process and Social Science: United States; Legal Realism; Legal Scholarship.

Bibliography

Campbell, C.M., Wiles, P., 1976. The study of law and society in Britain. Law and Society Review 10, 547.
Carlin, J., 1962. Lawyers on Their Own. Rutgers University Press, New Brunswick, NJ.
Carlin, J., 1966. Lawyers' Ethics. Russell Sage Foundation, New York.
Carlin, J., Howard, J., Messinger, S., 1966. Civil justice and the poor: issues for sociological research. Law and Society Review 1, 9.
Friedman, L., 1965. Contract Law in America. University of Wisconsin Press, Madison, WI.
Friedman, L., Macaulay, S., 1969. Law and the Behavioral Sciences. Bobbs Merrill, Indianapolis, IN.
Galanter, M., 1974. Why the 'haves' come out ahead: speculations on the limits of legal change. Law and Society Review 9, 95.
Garth, B., Sterling, J., 1998. From legal realism to law and society: reshaping law for the last stages of the social activist state. Law and Society Review 32, 409.
Grossman, J., 1965. Lawyers and Judges: The ABA and the Politics of Judicial Selection. Wiley, New York.
Grossman, J., Grossman, M., 1971. Law and Change in Modern America. Goodyear Publishing, Pacific Palisades, CA.
Handler, J., 1966. Controlling official behavior in welfare administration. California Law Review 54, 479.
Handler, J., 1967. The Lawyer in His Community. University of Wisconsin Press, Madison, WI.
Handler, J., Hollingsworth, E.J., 1971. The Deserving Poor: A Study of Welfare Administration. Markham Publishing Co., Chicago, IL.
Jacob, H., 1965. Justice in America: Courts, Lawyers, and the Judicial Process. Little, Brown and Co., Boston, MA.
Kalven, H., Zeisel, H., 1966. The American Jury. Little Brown, Boston.
Katz, S., 1971. Colonial America: Essays in Politics and Social Development. Little, Brown and CO., Boston, MA.
Reich, C., 1964. The new property. Yale Law Journal 73, 733.
Selznick, P., 1949. TVA and the Grass Roots: A Study in the Sociology of Formal Organization. University of California Press, Berkeley, CA.
Selznick, P., Nonet, P., Vollmer, H., 1969. Law, Society, and Industrial Justice. Russell Sage Foundation, New York.

Law and Technology

Ruth L Okediji, William L. Prosser Professor of Law, University of Minnesota Law School, Minneapolis, MN, USA

Published by Elsevier Ltd.

Abstract

The conclusion of the Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPS Agreement) in 1994 established a minimum baseline for regulating access to technology and knowledge goods. This global regime of intellectual property rights has since been a contentious aspect of modern economic relations. It has been viewed by consumers around the world as a significant barrier to access to technology and associated knowledge goods, including stymieing opportunities for social and political engagement by citizens. Public disenchantment with intellectual property as a primary regulator of access to technology could be addressed by the explicit linkage of proprietary rights to a new welfare axis comprised of development aspirations, human rights norms, and liberty considerations. To the extent social norms that develop around new technologies facilitate positive returns recognized by these complementary legal regimes, intellectual property rights that are in tension with these regimes will likely continue to lose moral sway, making the future of the TRIPS Agreement far less stable and its minimum obligations more costly to enforce. Moreover, gaps in how formal law and social norms regulate technology are not easily captured by the rigid prescriptions of treaty provisions, and thus the TRIPS Agreement is far less capable of serving the important role of shaping contemporary approaches to access to technology. As technology continues to reach deep into the private lives of citizens, and to affect the capacity and trajectory of national development in less-advanced economies, the design and construction of formal laws around which technology is produced, disseminated, and used will have greater import if they purposefully accommodate other legal orders whose norms resonate powerfully in advancing stylized visions of societal progress and human well-being.

> [F]or the Progress of Science and the Useful Arts … Art. 1, sec. 8, cl. 8. U.S. Const.
>
> We've arranged a civilization in which the most crucial elements profoundly depend on science and technology.
>
> Carl Sagan

Introduction

Technology is fundamental to human welfare and social advancement, perhaps more so today where technological engagement is indispensable for participation in political, cultural, and economic processes. At the level of the nation-state, sustainable access to technological inputs has profoundly shaped the direction of national economic growth, establishing the baseline for global competition and fueling intense conflict over the design and optimal blend of policies that best assure dynamic innovation without unduly burdening social welfare. The task nationally has been daunting for even the most advanced countries given deeply conflicting ideals about the constituent elements of social welfare, or the more elusive term 'progress.' The national difficulty with constructing a cohesive framework within which conditions for the production of and access to technology, including associated cultural goods, can flourish suggests that the long-term prospects at the multilateral level are quite dim. Nevertheless, legal rules that establish the policy context in which new technologies are designed, deployed, and accessed are already indelibly embedded into the global economic system through the multilateral intellectual property (IP) regime, established by the Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPS Agreement).

The question whether the global IP regime can deliver welfare outcomes across countries of divergent institutional, economic, and technological capacities has elicited contentious international debates and, in some cases, dispute settlement by international tribunals. At one level, these battles reveal unresolved historical contests about the role of IP in economic development; at another, they suggest significant insecurity about the wisdom of private rights granted under national laws playing the role of technology regulator amid a complicated web of international norms and obligations. Consider, for example, that addressing global challenges such as climate change, global public health, and those reflected in the United Nation's ambitious Millennium Development Goals (MDGs), requires access to technology. Yet, in international fora where these issues are being addressed, the role of the IP system is either completely ignored (as in the case of the MDGs) or highly contentious (as with climate change and access to medicines). The sharp tension between the design of the multilateral IP rules and the need to facilitate optimal supplies of global public goods has been the subject of robust scholarly attention, but very little progress has occurred on the international policy front. The situation demands new approaches to constructing the relationship between law, technology, and society.

Toward a Coherent 'Welfare Axis'

Scholars have increasingly acknowledged that the IP regime sits within a broader political and normative framework. At least three major types of legal instruments mediate the relationship between law, technology, and social welfare imperatives: national constitutions, international economic arrangements, and the United Nations Charter on Human Rights. Each one manages the relationship between law and technology distinctly, with the advancement of human welfare ostensibly the only conceptual link between them. The three regimes are not independent, parallel means to a single end. Rather, each

represents a set of legal norms that interact and challenge the competing, and at times conflicting, premises upon which the global economic system rests, producing varying visions of the ideal relationship between law and technology. The globalization of IP through the TRIPS Agreement only adopts one of the distinct visions embodied in these regimes, namely the economic private rights imperative entrenched in trade policy. However, national constitutions and human rights regimes suggest that the economic narrative is incomplete and that an additional narrative of the relationship between law and technology is possible and necessary. This narrative views IP not only as a set of economic incentives to produce technology but also as a key player in the link between innovation and human welfare. The inclusion of IP in national constitutions, human rights instruments, and in a variety of economic agreements potentiates new social and political spaces in which economic, civil, and political rights can be reconciled within multiple multilateral frameworks. Across the various subject matter categories, the organizing norms of the IP system should be infected with global norms along a 'welfare axis' consisting of three main pillars: technology and human rights, technology and liberty, and technology and development.

Technology and Human Rights

Technology can be defined as innovation that pervades, alters, and transforms human potential and capabilities. Technology defines our human existence, both the limitations and the expressions. Technology encompasses both the process and outcome of applying knowledge to improve human existence. In the digital age, technology differs from the tool or product function it served when railroads revolutionized travel or the cotton gin transformed agriculture; instead, modern technology is both a means and an end in and of itself. Technology has become the principal medium of commerce, social interaction, communication, and political engagement. The United Nations recently declared Internet access to be a fundamental human right (UN Human Rights Council, 2011). The Internet, social media, and mobile phones have altered not only the means by which we communicate, but also redefined the format, frequency, and language of those communications. Many information technologies have risen to the level of essential facilities – key resources upon which social, political, and commercial activities depend (Marshall, 1999). One illustration is the role of social media as an instrument of political organization and change. Social media networks have been used to facilitate resistance to unpopular government policies, to mobilize networks of supporters of global social causes, and to galvanize uprisings and regime changes (Wolman, 2013). Technology has indisputably become an enabler of political and economic human rights and this reality must be reflected in the IP system.

Technology and Liberty

IP law frameworks are now the primary regulators of the private technologies that shape our public life and the public technologies that pervade our private spheres. The interaction of law and technology is increasingly mediated by exclusive rights granted to private actors, formally enforced by courts but also imposed through contractual arrangements or technology design choices that attempt further control on societal interaction. The growing intensity and pervasiveness of technology has engendered significant scrutiny and consternation about the appropriate form and content of legal rules affecting privacy, security, and freedom of expression. The political liberty interests of users are often undercut by the sheer breadth of control over the technical infrastructure that is essential to contemporary accounts of social welfare. From software code to social media, proprietors of information networks increasingly assume and exert control in ways less obvious but equally as troubling as government fiats. Constructive responses to the tension between liberty interests and technology regulation have been lacking in the IP policy space nationally (Cohen, 2012), and appear to be hobbling efforts unfolding within the international human rights framework (Land, 2012). In sum, the more exclusive rights of control we give IP owners in an age in which technology is a critical social good, the more likely there will be resistance to the unaccountable exercise of those rights in relation to other democratic interests.

Technology and Development

As a system of private rights in the products of human ingenuity, the IP system has long been at odds with global efforts to ensure economic growth and cultural development beyond the industrialized countries. While proscribing free riding, the rules of the TRIPS Agreement have had unintended consequences, most notably in framing a discourse about its negative effect on fundamental human needs. Technology is ubiquitous, yet access to technology is a mirage for the vast majority of the world's population. Despite the rapid rate of technological change, most global citizens live in poverty and many lack access to the most basic goods including water, health, and education. Access to technology may greatly enhance prospects for development. From stem cell cloning to mobile banking in rural Kenya, technology holds significant promise to extend and enrich human life, but only if the competing legal regimes can cohere meaningfully at the national and international levels.

These three organizing principles – human rights, liberty, and development – form an axis around which a new approach to global IP may be organized, in turn, catalyzing new approaches and attitudes toward pressing global challenges. This requires significant effort to transform the unyielding economic narrative in which the TRIPS Agreement is currently mired. To understand the incompleteness of the TRIPS Agreement as a comprehensive regulator of technology, the structural relationship between law and technology as mediated by the IP regime deserves brief attention in consideration of the welfare axis just articulated.

Global Economic Considerations, the TRIPS Agreement, and Domestic Welfare in the Design of IP Rules

The ideals that animate the terms and conditions of access to technology on a global scale are shaped by three legal regimes: IP, trade, and competition. Of the three, global IP rules have

been at the center of serious and protracted normative contests between countries, industries, and consumers over what the legal regulation of knowledge goods should accomplish and what values should inform the content of the governing rules. Between countries, these battles have remained stubbornly resistant to resolution, and instead have spread across diverse international fora and policy contexts (Helfer, 2004). The question of creating and managing global conditions that support technology development and use is further compounded by uncertainty over which structures and processes can best manage the confluence of social, political, economic, and institutional factors that shape conceptions of the role of technology in competing versions of the public good.

The legal rules countries adopt with respect to ownership, trade, and competition in technological goods are required to reflect the robust obligations of the TRIPS Agreement. The TRIPS Agreement, signed in 1994, established mandatory minimum standards of protection in all major categories of IP for all 158 member countries of the World Trade Organization (WTO). Most significantly, these obligations include extending IP rights on terms compatible with the principles of national treatment and most-favored nation status. The Agreement represented a significant upgrade from prior multilateral systems in terms of its comprehensive scope, the level of protection required, and, for the first time, the ability to compel compliance on pain of trade sanctions through the WTO dispute settlement process.

Despite this apparent international harmonization, the policy rationales underlying IP, competition, and trade are designed singularly to advance domestic welfare priorities. Laws and regulatory frameworks attendant to each subject are justified by explicit reference to *national* considerations for producing optimal conditions for innovation and creativity. Thus, despite strong rhetoric and a canonical narrative about the *global* welfare gains of harmonized IP laws, the international coordination of IP and associated legal regimes regulating technology is best understood or rationalized as an effort to limit the disruptive effects that free riding on public goods may have on the productive activity, and thus competitive advantage, of major technology exporters. In short, global IP rules have the effect of channeling the future competitive potential of developing countries away from technology-intensive sectors by making access to technology prohibitively expensive. Terms of global competition and economic growth are thus indirectly set through the design of global IP rights (Reichman, 2009), diluting the positive consequences of free trade, or otherwise limiting them to the technology exporters alone. The results are obvious: a longer path to economic development for less-advanced economies and a greater potential for dependency on foreign supplies of technical inputs at very high cost.

Nevertheless, it is clear that every country obtains technology by *some* combination of free riding, innovation, and trade. The dominant method at any given point in history is a function of, among other things, factor endowments, the size and quality of the labor pool, educational opportunities, and the terms or costs of access to technology. For example, critics of the IP regime established by the TRIPS Agreement were quick to point out that historically the United States and other industrialized countries enjoyed relatively low-cost access to technology by appropriating the inputs of foreign creators. In the same vein, ancient European city-states deliberately targeted skilled artisans, craftsmen, and innovators to import their technologies into the realm. These so called 'passports' – inducements to introduce new technologies into domestic markets outside the inventor's country – remain a justification for global IP, although less convincing today than once was the case. The reality is that such inducements alone have proven simply to intensify the structural conditions of underdevelopment by simultaneously raising the costs of access to technology and the provision of basic global public goods such as education and public health. It should be no surprise that the relationship between the IP regime and the demand for technology has become highly contested both domestically and internationally, exacerbated by rapid technological changes in information and communication technologies that constantly challenge the competitive capacity of major industries.

Global Welfare Concerns Shape the Legal Context of the Production of New Knowledge

But the TRIPS Agreement embodies more than just economic considerations. Or at least there are strong claims that it ought to. First, it is not self-evident that the economic incentives embedded in the TRIPS Agreement act to encourage innovation in the modern information economy. Even in the mature economies of developed countries, both the *presence* and *absence* of IP rights have galvanized new forms and expressions of innovative activity in the modern radically reconfigured economic environment. In this environment, global firms maintain diverse business models, championing some traditional IP ownership rights while simultaneously seeking to limit or jettison such rights where they threaten or disrupt activities that have become the mainstay of ordinary life for citizens in a globally networked information economy. It is also an environment in which new intermediaries provide much-needed information access points for citizens but whose relationship to IP owners and consumers is uneasy – especially with respect to liability for IP infringement. These new challenges, actors, kinds of economic behavior, and terms of sociopolitical engagement, all mediated by new technologies, point to innovation expressed around legal, cultural, and ethical considerations that demand a new construction of the TRIPS Agreement to accommodate values beyond its narrow economic logic.

Second, at an international level, the burden of legal rules that require compliance upon pain of trade sanctions has heightened social and political anxiety over the poverty-related challenges endured by the vast majority of the world's population. This constant anxiety has generated resistance to both the trade and IP regimes, destabilizing the normative assumptions of the public welfare vision around which (at least the domestic) IP rules have historically been organized. Although skepticism of strong IP rights has long been part of the multilateral IP domain, it has now become a fixture within the developed countries. Consumers everywhere have become the face and voice of a platform previously occupied solely by the technology-importing developing countries. The difficult

struggles over legal and technological controls on information flows resonate deeply in the native allegiance to personal liberty that occupies such a central role in Western democracies (Cohen, 2012). The collective impulse to reject incursions on personal liberty has converged in powerful ways: developing countries and their citizens, frustrated by their inability to enjoy the development path characterized by low-cost access to technology historically enjoyed by many developed countries have formed coalitions championing 'access to knowledge' and 'access to health' that serve as filters for assessing the welfare impact of global IP rules (Kapczynski, 2008).

As a result of these trends, the role of technology in global economic relations has become highly fractious, with no clear winners in the fight over the substantive legal norms that undergird international IP protection. The highly celebrated dispute settlement process of the TRIPS Agreement has had more symbolic effect than real; it has been less utilized (perhaps appropriately) against developing countries than originally expected, enforcement activities appear to be centered more on legislative action than on practical initiatives, and compliance with decisions has been suboptimal, particularly by the U.S. (Bird and Brown, 2012). Moreover, nationally, reform efforts and court decisions in leading technology-exporting countries suggest that the IP pendulum currently is not wholly aligned with the maximalist IP agenda of the domestic trade agencies. And among ordinary citizens, politically organized and technologically deployed social resistance to renewed efforts to increase the rights of IP owners has become part of the welfare motif in these same developed countries.

Human Welfare, Pervasive Technologies, and the Need for Legal Coherence

The ways in which technology is intimately integrated into every aspect of human life, from the most public to the most private spaces, mediating social, economic, and political interactions, have complicated the binary relationship between IP and access to technology. The ubiquity of technological choices directed at facilitating cultural interchange, political dialogue, and social expression has made it both doctrinally difficult and politically complex to simply view IP as a set of sacrosanct private rights insulated from the claims of other normative legal orders. The greatest challenges to the governance of knowledge goods and associated technological platforms stem *precisely* from the constraints that inhere in the role of the IP regime in facilitating effective and sustainable access to knowledge and the goods that embody it. For example, the tension between human welfare and IP at a global level has been particularly problematic because of the absence of a distinctive theory for global welfare outside the free trade context (Okediji, 1999). The advent of the TRIPS Agreement has broadened the responsibility for IP to address social and human welfare needs on global terms that have yet to be developed. A minimum function of the global IP regime should be to facilitate coherence among parallel legal regimes affected by the impact of technology – or the lack thereof – on the worldwide human welfare calculus.

The Complexity of Law and Technology Interactions

The lack of coherence among technology-dependent legal regimes stems partly from attempts to fit extremely complex interactions within simplistic, narrow models of behavior and interaction. Global IP policy should more accurately reflect the myriad ways that law and technology interact. These interactions are complex, reflecting both planned and unintended consequences and defining the organization, values, and possibilities of societies. Further, the interaction is bidirectional; technology influences the form and content of law and law influences the content and direction of technological progress. 'Law' includes formal, public codes, statutes, and regulations; private limitations imposed by technological architecture, contracts, and the market; and informal, public limitations imposed by social behaviors and norms. 'Technology' in this sense is defined as applied science or tools and techniques, as differentiated from 'basic' science, defined as fundamental research not designed to lead to a specific industrial application (Moses, 2007). Moreover in a broad sense, law and regulation are themselves a form of technology; they are the technology of civil ordering, a means of controlling inputs and outputs. So embedded is the notion of law as technology, and so intertwined are the functions, that policymakers often adopt the vocabulary of technology, seen in such references as 'policy tools' or 'regulatory toolkits.'

Since the explosion of the Internet, there has been a consensus among experts that technology design and use also function as a form of private law; the existing formal legal frameworks allow private technology owners and developers enormous latitude to shape the direction and form of technological progress, define conditions of access, and influence the architecture of social interaction and everyday life. For example, Apple smart phones running the operating software iOS are designed to access only Apple-approved content and certain cellular providers' networks. These limitations on use and access are built into the software code of the device; circumvention can only be achieved through a process known as 'jailbreaking,' a widely practiced method of unlocking the code limitations of somewhat dubious legality.

The widespread practice of jailbreaking despite its inconsistency with formal law illustrates the power of another form of 'law' – social norms. Social behaviors and norms exist in a more flexible and responsive framework than formal law, and are accordingly more rapidly molded and altered by technology. For example, the practice of jailbreaking phones arose in response to norms of sharing, personalization of content, and nondiscrimination on the basis of platform or type of device. Many online communities also self-regulate according to specific community norms; the online community Reddit, for example, is regulated by users who control the promotion of posts, stories, and comments through 'downvotes' and 'upvotes' in accordance with community norms that value novelty, inclusive dialogue, and brevity.

Of these three forms of law – formal law, the embedded private law of technology design and deployment, and social norms – social norms and private law are the most capable of quickly responding to technological change. Formal law, when it catches up, is either ratifying or dismantling an already existing relationship between social norms and technology,

suggesting that the TRIPS Agreement's focus on formal law undermines the powerful role of social norms and potentially leaves developing countries with fewer tools to respond to technological changes.

The Formal Law Lag

The TRIPS Agreement was negotiated and signed in the infancy of the information age: the full capacity of the Internet and mobile devices as economic, political, and cultural mediums of development and exchange had yet to be realized. Because the TRIPS Agreement was signed before the technological explosion and the development of a digital society, the legal regime it creates is obsolete and unresponsive to modern welfare concerns. The term 'cultural lag' describes this gap between the rate of change of material culture and nonmaterial culture (Ogburn, 1957). Material culture encompasses technological change and includes the physical equipment and the procedures for using and producing that equipment. Nonmaterial culture is comprised of moral, ethical, religious, and philosophical belief systems and values. Especially during rapid technological change, a culture lag can leave the nonmaterial culture without appropriate ethical tools by which to evaluate the design, use, and effects of technology.

There are several reasons that ethical tools lag behind technological development. Modern technological development occurs in insular and highly competitive communities (Marshall, 1999). In contrast to the idealized 'basic science' research community focused on publication and rapid dissemination of knowledge, IP regimes vital to technology firms place a premium on nondisclosure (Rai, 1999). Additionally, technological advances occur within tightly focused and controlled corporate structures largely isolated from public debate and disclosure over the technology being developed (Marshall, 1999).

Moreover, cultural lag is not purely linear, and different elements of material culture do not adapt and evolve at the same rate. Social norms and behaviors are more flexible and responsive than the more formal bodies responsible for public legal change, and are accordingly more rapidly molded and altered by technology than formal law. The resulting dissonance between formal law and social norms leads to a further 'law lag' within nonmaterial culture, where socially acceptable behavior or technology use is not reflected in the existing structure of formal law (Larsson, 2012). This lag is more pronounced today because of the rate of technological change and the comprehensive, global scope of communication made possible by the Internet and social media; social norms are globally consolidated and stabilized in ways that formal law simply cannot recreate and, instead, often seeks to delegitimize in the IP context.

Because of the flexibility and responsiveness of the social norm framework, social norms increasingly play the role of arbiter between technology and end users. Social norms do not pervade formal law as quickly as they pervade technology; technology is more responsive than formal law. By way of illustration, one view of the Google Books project to digitize and allow in-text searching of millions of books arguably is premised upon consumer expectations about the transformative nature of the Internet and the added value of digitized libraries. A 2013 lower court ruling holding that the scanning project fell within the 'fair use' exception in the United States copyright law could be framed as merely a formal confirmation of the already existing relationship between digitization technologies and experienced social norms. If an appeal by copyright owners is successful, it would begin a process to reinstate the explicit hierarchical order of property rights over decisions about how technology can be used and what the terms of such use might be.

The disconnect between the strong IP rights endorsed by the TRIPS Agreement and social norms is aptly demonstrated in the public debate and demonstrations that ultimately led to the defeat of the Stop Online Piracy Act (SOPA) and the Protect IP Act (PIPA) in the United States. Shaped by the ease of reproduction, variation, and distribution on the Internet, social norms of file sharing or file format shifting are seen as legitimate, while copyright law still defines those uses as infringing, contributing to a growing public perception of copyright enforcement as illegitimate. SOPA and PIPA, backed by copyright owner groups, sought to strengthen copyright enforcement by imposing affirmative requirements and liability on Internet intermediaries, such as service providers, search engines, and economic services like PayPal to identify and remove content that infringed copyrights. In response, more than 115 000 Web sites, including Wikipedia, Reddit, and Google 'blacked out' their home pages on 18 January 2012 to illustrate the potential effect of the proposals. As a result, an estimated 10 million people signed a petition against SOPA/PIPA, more than 8 million attempted phone calls, and 4 million e-mails were sent to Congress (Herman, 2012). This digital dissidence is remarkable not only because of its scope and impact, but also because of the abject failure of lawmakers to recognize the limits of IP laws in relation to social expectations of access to technology. The failure of formal lawmaking institutions to recognize how fundamentally out of touch existing IP frameworks are with new ideals of progress and the architecture of Internet technology remains one of the strong challenges to incorporating welfare concerns into interpretations of the TRIPS Agreement. Instead, as the law lag widens, the TRIPS Agreement remains a precedent for shifting regulatory authority away from public debate and formal legal bodies and toward private owner-mediated restrictions.

Reorienting IP Frameworks

Champions of the strong IP rights captured in the TRIPS Agreement reference the canonical bargains of IP rights as justification for the attendant negative externalities; IP rights are supposed to spur the creation of new products, knowledge, or creative works in exchange for legal protections. These bargains find theoretical backing in a narrative conceptualizing technology as a beneficial, autonomous force (Feng, 2000). This narrative also assigns value to certain exploitations and forms of technology, primarily widespread commercialization. However, the technology-centric narrative of TRIPS is not inescapable or necessary to spur innovation.

First, the conception of technology as value-neutral is neither accurate nor obligatory for a functioning, strong IP

framework. IP owners impose a form of private law by building limitations and controls into the design or architecture of technology. The value-driven aspects of technology are most obvious in software code. As Lawrence Lessig cogently explained, software code and the architecture of the Internet is a form of law and regulation (1999). Lessig's 'law of the code' is illustrated in content filtering software used to identify online content that infringes copyrights.

Second, the TRIPS Agreement embraces a narrow view of IP that assigns significance and recognition only to forms and uses of technology consistent with the developed world's definition of innovation and progress. However, it is increasingly clear that the categories and values we have chosen to recognize and elevate in the economic discourse are not ineluctable. As critical IP scholars have studiously argued, creativity in less-developed economies exists in some abundance. Forms of knowledge anchored in values distinctive to cultural institutions do flourish and are, ironically, misappropriated by 'innovators' from developed countries who exploit underinclusive IP frameworks. The axis around which contemporary IP relations revolve will unavoidably create a conflict between the protection that privileges innovation pathways unique to developed countries and the forms that creativity and innovation take in the developing world. The World Intellectual Property Organization (WIPO) Intergovernmental Committee for Genetic Resources, Traditional Knowledge and Folklore has been working on an international instrument designed to protect knowledge categories and resources traditionally excluded from IP frameworks. This is a vivid example of how new legal spaces for technology regulated through IP can be constructed. Within this WIPO process, the main proposal is for an international instrument to require patent or IP applicants to disclose sources of genetic resources and associated traditional knowledge where these have been accessed and used in the innovation process. A few developed countries are highly resistant to this proposal, sticking firmly to a conception of private rights and the patent system as the principal object of the human welfare calculus. As many of the developing countries have emphasized, however, the two knowledge systems are not irreconcilable within a broader framework of international law.

The Private Law of Technology

It is certainly improbable and even unfair to hold IP alone responsible for the fulfillment of human welfare aspirations, but it is also unwise to have IP resist or constrain the potential of other regimes directed at that end. With few exceptions, the TRIPS Agreement requires that formal legal limitations on technology be divorced from liberty and human rights policy considerations. As framed in the popular narrative, the TRIPS Agreement is indifferent to the evaluation or regulation of the ethical or socially responsible design, use, and distribution of the inventions in question, including essential inventions like genetically modified agricultural products, human genetic sequences, and information technologies. The inability of governmental entities to engage in policy consideration has placed technology owners in a position to fashion private law. Though framed as private rights, in reality, IP ownership is increasingly a right to govern by shaping the public space and exerting control over the deeply intimate choices of citizens.

IP owners fix the terms and conditions of access to technology in two main ways: by embedding access limitations within technology design and by setting the terms of deployment and distribution of essential technologies. The collective effect of this private law is a transition to a permission culture. Historically, innovation culture was postulated on an idea of 'free' culture – not as in 'no cost,' but rather as in free speech and free expression (Lessig, 2004). The opposite of free culture is a 'permission culture,' or a culture in which subsequent creators and inventors must seek permission in order to invent or create. This permission culture shift is driven not by public law or governmental regulation, but rather by the imposition of private controls in the regulatory spaces left open by gaps in formal regulation and by the expansion of IP protections without concurrent expansion of limitations. The permission culture of the private law of technology is not reflective of the interests of the majority of citizens, nor ultimately of the long-term economic interests of the private developer.

The shift to a permission culture controlled by private owners is of some concern because of the impact on essential liberty interests. New models of copyright enforcement require intermediaries, such as Internet service providers and search engines, to actively police for copyright infringement to escape liability (Graber, 2010). Many intermediaries have turned to automated content filtering technologies to police for likely infringement. These automated filtering technologies cannot cope with the inexact standards of the formal copyright law on the books, such as the many uses that fall within the 'fair use' exceptions to copyright infringement. Instead, the technology is overinclusive, acting upon and identifying a broader category of 'infringement' than identified in formal copyright law. In this way, the design or architecture of the software program, usually controlled by a private company, imposes constraints and grants rights different than the law on the books. The result is private, inconsistent, and often invisible intrusions upon the liberty interests of ordinary users fully sanctioned by both national laws and arguably mandated by the TRIPS Agreement.

The consequences of the shift to a permission culture reverberate in even the most fundamental areas of human rights, particularly health-care access. In the biotechnology field, the concentration of decision-making power in the hands of private owners can limit access to key treatments and diagnostics and inhibit future research and development. The patents at issue in the landmark U.S. Supreme Court case *Association for Molecular Pathology v. Myriad Genetics* covered genetic sequences and diagnostic tests associated with heightened risks of breast and ovarian cancer. Patents on these types of diagnostic tests are common, but many owners, recognizing the health and research implications, widely license the technology used in diagnostics. Myriad Genetics departed from this trend by refusing to license their diagnostic tests to other laboratories. This refusal to license allowed Myriad to acquire a monopoly, with all of the attendant consequences; tests were much more expensive, only covered by certain insurances, and limited to the most common mutations (excluding mutations more common in minorities). The American Civil Liberties Union challenged Myriad's patents on both patentability grounds and as violative of the First Amendment's guarantee of

free speech on behalf of researchers, genetic counselors, cancer survivors, breast cancer and women's health groups, and scientific associations. Although the Supreme Court held isolated genetic sequences to be unpatentable, patents on synthetic genetic sequences and diagnostic tests using those genetic sequences were sustained. Effectively, even after the decision, by virtue of owning the IP rights to the purified genetic sequences, Myriad is able to define the conditions of access to key health technology by imposing contractual conditions and other limits on the distribution of the test. Moreover, arguably the most valuable IP Myriad now has is its huge database of trade secret mutation information accumulated from all the years of performing the test, a database that will only be accessible, if at all, on the terms established by Myriad. In sum, IP rights engender resort to legal tools that foreclose public policy from influencing technology regulation, and these tools are beyond the purview of the TRIPS Agreement. Yet, as a form of regulation, the public welfare implications of these tools must be part of the social bargain inherent in the global IP regime.

The global IP system has long eschewed normative influences that might dilute its traditional emphasis on control and rent accumulation. The rapid evolution of information communication technologies has galvanized a much more sophisticated and growing public understanding of the extent to which the enjoyment of cultural goods and the exercise of political and social freedoms can be circumscribed by the exercise of IP rights. Put simply, the ubiquity of technology in the everyday lives of consumers strains the purely economic logic of the IP welfare calculus. This dissonance between how widely technology affects the exercise of political and social interests and the insistence on ever-increasing IP rights is largely behind contemporary cultural hostility to IP as the main policy instrument for regulating access to new knowledge goods and the technological platforms that ensure their proliferation. This hostility – both national and global – has also made policy discussions about the role of IP in promoting social welfare intractable and legal solutions politically unfeasible. To the extent IP rights justify the exercise of private power over all kinds of interactions with technology, it is unlikely to attract the legitimacy necessary to support the expansion of IP rights or improve prospects for better enforcement. The design and orientation of IP laws thus must be more accommodating of the values that define contemporary visions of human welfare as reflected in other legal regimes, such as human rights, in which private human ingenuity, economic development opportunities, and liberty interests can be nurtured and coexist.

Conclusion

Technology is deeply embedded in the routine and extraordinary moments of social interaction, economic organization, and political engagement. It seems implausible that the global IP regime can be meaningfully responsible for ensuring the fullest possible welfare gains associated with new technologies for consumers worldwide. It must therefore benefit from the multiple multilateral frameworks that also regulate the role of technology in facilitating human well-being. The global IP regime needs a welfare axis; it can and must cohere with development aspirations, strengthen liberty interests, and serve human rights goals. Nothing less will assuage the demand for a socially accountable system for promoting the production of technology and advancing human capacity to improve the material conditions that make human flourishing possible. TRIPS enforcement provisions notwithstanding, global consumers define what is lawful largely in reference to what is technologically possible, and not the other way round. Social practices thus inform and encourage technological production, use, and utility far more significantly in the information economy than exclusive economic incentives. Global IP policy must seek to narrow the gaps between formal law as established through the TRIPS Agreement, and informal social norms and technological controls that dictate consumer expectations of technology. In the end, such consumer expectations dictate the relationship between law and technology in ways much more powerful than the formal law on the books.

See also: Intellectual Property: Law in Context; International Law and Treaties; Law and Development; Law, Bioethics, and Biotechnology.

Bibliography

Agreement on Trade-Related Aspects of Intellectual Property Rights, 15 April 1994, Marrakesh agreement establishing the World Trade Organization, Annex 1C, 1869 U.N.T.S. 299, 33 I.L.M. 81(1994).

Association for Molecular Pathology v. Myriad Genetics, Inc., 133 S. Ct. 2107, 2117 (2013).

Bird, R.C., Brown, E., 2012. The protection of well-known foreign marks in the United States: potential global responses to domestic ambivalence. North Carolina Journal of International Law & Comparative Regulation 38, 1–41.

Cohen, J., 2012. Configuring the Networked Self. Yale University Press, New Haven.

Feng, P., 2000. Rethinking technology, revitalizing ethics: overcoming barriers to ethical design. Science and Engineering Ethics 6 (2), 207–220.

Graber, C.B., 2010. Internet Creativity, Communicative Freedom and a Constitutional Rights Theory Response to "Code Is Law." [Online] International Communications & Art Law Lucerne Working Paper No. 2010/03. Available at: http://www.unilu.ch/files/i-call_Working_Paper03_Graber_Internet_Creativity.pdf (accessed 26.02.14.).

Helfer, L.R., 2004. Regime shifting: the TRIPS agreement and new dynamics of international intellectual property lawmaking. Yale Journal of International Law 29 (1), 1–83.

Herman, B.D., 2012. A political history of DRM and related copyright debates, 1987–2012. Yale Journal of Law and Technology 14, 162–225.

Kapczynski, A., 2008. The access to knowledge mobilization and the new politics of intellectual property. Yale Law Journal 117, 804–885.

Land, M., 2012. Rebalancing TRIPS. Michigan Journal of International Law 33, 433–480.

Larsson, S., 2012. Conceptions in the code: what "the copyright wars" tell us about creativity, social change and normative conflicts in the digital society. Societal Studies 4 (3), 1009–1030.

Lessig, L., 1999. Law of the horse: what cyberlaw might teach. Harvard Law Review 113, 501–546.

Lessig, L., 2004. Free Culture: The Nature and Future of Creativity. Penguin Books, New York.

Marshall, K.P., 1999. Has technology introduced new ethical problems? Journal of Business Ethics 19 (1), 81–90.

Moses, L., 2007. Why have a theory of law and technological change? Minnesota Journal of Law, Science and Technology 8 (2), 589–606.

Ogburn, W.F., 1957. Cultural lag as a theory. Sociology & Social Research 41 (3), 167–174.

Okediji, R.L., 2006. The International Copyright System: Limitations, Exceptions, and Public Interest Considerations for Developing Countries. [Online] UNCTAD-ICTSD Project on IPRs and Sustainable Development Issue Paper No. 15. Available at: http://unctad.org/en/Docs/iteipc200610_en.pdf (accessed 26.02.14.).

Okediji, R.L., 1999. Copyright and public welfare in global perspective. Indiana Journal of Global Legal Studies 7 (1), 117–189.

Okediji, R.L., 2014. Public welfare and the international patent system. In: Okediji, R.L., Bagley, M. (Eds.), Patent Law in Global Perspective. Oxford University Press, New York, pp. 1–52.

Rai, A.K., 1999. Regulating scientific research. Northwestern University Law Review 94, 77–152.

Reichman, J.H., 2009. Intellectual property in the twenty-first century: will the developing countries lead or follow? Houston Law Review 46, 1115–1185.

United Nations Human Rights Council Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (16 May 2011) UN Doc. A/HRC/17/27.

Wolman, D., 2013. Facebook, Twitter help the Arab Spring blossom. Wired, [Online] 16 April. Available at: http://www.wired.com/magazine/2013/04/arabspring/ (accessed 26.02.14.).

Law and Terrorism in the US and Beyond

William C Banks, Syracuse University, Syracuse, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Development of a legal regime to counterterrorism is a modern phenomenon that is maturing in the twenty-first century. Prior to the 9/11 attacks, in most states terrorism was treated as criminal behavior, and the methods of deterring and sanctioning terrorism were drawn from the law enforcement experience. Counterterrorism law today includes distinct prevention elements, including investigations and detention authorities, discrete interrogation rules and practices, and the use of military force, among other elements. Counterterrorism law blends international and domestic laws, and includes human rights protections, special laws to govern armed conflicts, and several specific international treaties that respond to unique attributes of international terrorism.

Law and Terrorism in the United States and Beyond

Understanding the dynamics of terrorism and the law requires a clear picture of what terrorism is. Scholars have identified dozens of definitions of terrorism and a similar variety can be found in the laws of the United States and other nations and international bodies.

One definition that encompasses many of the widely accepted attributes of terrorism has been adopted by the US Department of State. It defines terrorism as "premeditated, politically motivated violence against noncombatant targets by subnational groups or clandestine agents, usually intended to influence an audience." The International Convention on the Suppression of the Financing of Terrorism, an international treaty adopted by the United Nations (UN) General Assembly in 1999 and ratified by over 130 countries, defines terrorism as "any…act intended to cause death or serious bodily injury to a civilian, or to any other person not taking an active part in hostilities in a situation of armed conflict when the purpose of such act, by its nature or context, is to intimidate a population, or to compel a government or an international organization to do or to abstain from doing an act." Finally, the US Department of Defense considers terrorism to be "the unlawful use, or threatened use, of force or violence against individuals or property to coerce and intimidate governments or societies, often to achieve political, religious, or ideological objectives."

Each of these definitions incorporates common aspects of terrorism. First, terrorism involves actual or threatened violent, premeditated actions. Second, the motivation behind terrorism – political, religious, or otherwise ideological – is distinct from ordinary crime. Third, terrorists use violence as a means of communication, getting the word out about their motivations and objectives. Thus, by using violent acts to convey political, religious, or ideological messages, terrorists seek an audience. That audience may be reached by the media coverage of the terrorist's issue, or through instilling fear in a population, which in turn pressures its government to recognize the terrorists' political agenda. Fourth, terrorists often seek to draw international attention to their causes.

Terrorism has ancient roots, where terrorists engaged in assassinations, fomented mass uprisings, or formed marauding bands to inflict harm on civilian populations. Terrorism evolved globally through changing tactics and the growing lethality and speed of weapons and other destructive instruments. In general, terrorists employed violence in discriminate and proportionate ways.

By the late 1960s, a new era of terrorism emerged. Terrorist groups began hostage taking for political purposes, often by hijacking airplanes. Three developments characterize the changing nature of terrorism during this time. First, terrorists began routinely to carry out attacks across international borders. For instance, it became common for airliners to be hijacked and diverted to another country for negotiations for the hostages' release. Second, the 'new terrorists' began targeting important political symbols, including the 1984 bombing of a hotel by Northern Ireland terrorists where British Prime Minister Margaret Thatcher was staying. Finally, terrorists learned how to take advantage of the media attention surrounding a terrorist event. For example, even though the attack by Palestinian terrorists during the 1972 Munich Summer Olympics ended in tactical failure, the worldwide media attention of the events drew attention to the Palestinian cause.

By the 1990s, the tendencies of terrorists to employ discriminate and proportionate violence in service of their goals was called into question. The number of incidents of international terrorism declined during the 1990s, but terrorists became more ruthless. The number of fatalities attributed to terrorist attacks grew dramatically during this time and into the 2000s. Experts only began to fully understand the nature of the threat of al-Qaeda and Islamic fundamentalist terrorism after the devastating attacks of 11 September 2001. Under the guise of religious rhetoric, Osama bin Laden declared war against America in 1996, calling on every Muslim to kill Americans. But in fact, al-Qaeda's objectives were political, seeking to expel the United States from Saudi Arabia and the Persian Gulf region. Al-Qaeda and other post-9/11 terrorist groups affiliated loosely with al-Qaeda depend on loose, network-based cells that operate with considerable autonomy, rather than a formal hierarchical arrangement. This model gives the group flexibility and resilience to counterterrorism strategies aimed at bringing down the organization.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86146-2

Terrorist Motivations and Operations

To institute and carry out preventive measures to counteract terrorist threats, it is helpful to understand the motivations, and organization and tactics of modern terrorists and their organizations.

Motivations

Terrorists defy easy classification. They can be male or female, young or old, and from a wide variety of ethnic, religious, and socioeconomic groups. Examples of this can be seen in the increasing number of female suicide bombers in the Middle East, the high levels of education of many leaders of terrorist networks, and the existence of groups across the world, including the Tamil Tigers, narcoterrorists in Latin America, the Irish Republican Army (IRA), and the Euskadi Ta Askatasuna (ETA) (Basque Homeland and Freedom). Furthermore, modern terrorism trends indicate a necessary level of sophistication, resources, and training far beyond terrorist capabilities of the 1970s and 1980s.

While some terrorist acts are historically a way to draw attention to political or ideological agendas that have been ignored by the established order, other activities are more difficult to classify. Motivations can range from highly personalized worldviews to pure anarchism. The typology of terrorist motivations can be broken down into three categories: state sponsored, non-state sponsored, and individual.

Historically, evidence indicates that the primary motivations for state-sponsored terrorism include political, religious, or ideological reasons, or a combination of these factors. This support is usually evidenced through financial and/or military resources. Examples include the former Libyan government's involvement in the attack on Pan Am Flight 103, and its indirect support of the IRA and other European and Middle Eastern terrorist groups; North Korea's terrorist-style attacks across Asia; and Iran and Syria's support of Hezbollah.

Motivations driving non-state-sponsored terrorism vary widely, and include political and ideological issues, religious beliefs, organized crime, and issue-specific concerns such as social issues including racial or ethnic purity, abortion, or the environment.

Subcategories within political and ideological issues include separatist movements and cults or sects. Separatist movements such as the ETA, Tamil Tigers, and IRA, resort to terrorist activities in an attempt to gain independence for a specific geographic area or category of individuals. In contrast, cults often prescribe to a messianic or apocalyptic vision that offers the possibility of salvation if a particular lifestyle is adhered to. Since the 1980s, several cults have engaged in terrorist acts, including the Rajneeshi cult in the United States and the Aum Shinrikyo cult in Japan. The Rajneeshi cult utilized salmonella and attempted to poison the local water supply in Oregon in an attempt to achieve their political objectives by preventing residents from voting in a local election. The leader of the Aum Shinrikyo cult aspired to dominate Japan, and eventually the world, and ordered his followers to carry out a large-scale chemical attack in the Tokyo subway system in 1995.

Religious and political motivations are often intertwined and can be hard to differentiate. Al-Qaeda, Hamas, and Hezbollah all fall into this category. While the groups invoke religious rhetoric, they also compose political and social movements within various countries. The most prominent organized crime groups include narcoterrorists and the mafia. South America and the Middle East have seen an increase in narcoterrorism, which is used to intimidate or threaten the government and can be used to finance violent attacks while the Russian mafia has proven effective at utilizing terrorist methods by permeating the government, transportation, and construction sectors.

Individual motivations are notoriously hard to pinpoint, ranging from mental illness, to personal grudges, to ideological or religious beliefs. Arguably, the most famous example of individual terrorism is seen in Timothy McVeigh's bombing of the Alfred P. Murrah Federal Building in Oklahoma City in 1995. In the most deadly act of domestic terrorism by a US citizen, McVeigh detonated a truck packed with an agricultural fertilizer and highly explosive motor fuel in front of the federal building, killing 168 people and injuring over 850. Evidence revealed McVeigh held strong antigovernment and neofascist beliefs, likely motivating the attack.

Organization and Tactics

Modern terrorist groups have mostly abandoned the traditional hierarchical leadership structure in favor of a flat organizational structure. This new structure is composed of semi-independent cells that make operational decisions autonomously from a central leader. This decentralized structure also allows many terrorist organizations to have cells that each focus on a particular region.

Terrorist tactics have perpetually evolved in the last half-century. During the Cold War, state-sponsored terrorism was prevalent, with both the United States and Soviet Union sponsoring proxy groups to overthrow governments that one or the other opposed. By the 1970s, non-state-sponsored terrorism became more prominent. The groups, such as the Palestine Liberation Organization (PLO), used attacks to gain media attention and public awareness of their cause. The tactic of causing mass casualties was not popular until the 1990s, with the increased use of suicide bombers. Concerns for the evolution of terrorist tactics over the next decade include the possible use of nuclear, biological, or chemical weapons.

The Pertinent Law

The legal regime for countering terrorism is in its formative stages, although it is maturing. For a long time, nations responded to terrorism as criminal behavior subject to law enforcement methods and processes. Indeed, law enforcement approaches remain central to countering terrorism, although in the modern era the criminal paradigm has expanded to focus on preventing terrorism in addition to bringing those who engage in terrorist acts to justice. As a result, terrorist threats must be prevented, not just detected and prosecuted like ordinary crimes. Prevention requires investment in a relatively open-ended collection of intelligence, leading to the blending of traditional law enforcement and national security intelligence operations. Meanwhile, the most visible response to the

9/11 attacks was labeled a 'war on terror' by the Bush Administration and many others. The use of military force was joined with traditional law enforcement, national security intelligence collection, and various economic and soft-power approaches in what has become a multifaceted counterterrorism regime worldwide.

The expansion of counterterrorism beyond the criminal law model to intelligence collection (including various forms of electronic surveillance), financial tools, and the use of force has complicated and enriched counterterrorism law. Particularly in the years since the 9/11 attacks, international law and several states' domestic laws have expanded and taken new shape as counterterrorism law matures. The following sections provide an overview of the most significant components of the emerging international and domestic law regimes.

International Law

The UN Charter and Geneva Conventions provide the foundation for the regulation of the use of armed force against terrorists. UN Charter provisions outline the *jus ad bellum*, the legal bases for the right to wage war, and the Geneva Conventions and other international treaties, judicial decisions, and customary law provide for the *jus in bello*, the bounds of lawful conduct after war has begun. The Geneva Conventions aim to protect the victims of armed conflict and their terms define much of international humanitarian law (IHL). Originally largely only applicable to international armed conflicts – those between nation-state signatories to the Conventions – the adoption of Additional Protocols I and II in 1977 extended the basic Geneva principles to a broader range of conflicts, including noninternational armed conflicts – those not of an international character occurring in the territory of a state party to the Conventions. While these conflicts can reach terrorist activities, due to the nature of the violence and the involvement of nonstate actors, the Conventions and Protocols have failed to provide clear guidance for the rules that apply to terrorists. The United States has not ratified either protocol, but applies them to the extent that they constitute customary international law.

Human rights law (HRL) is a species of international law that traditionally applies to protect all parties, both combatants and civilians, in times of peace. However, most scholars and international tribunals agree that it also serves to complement IHL during armed conflict. Created through international custom as well as treaties and conventions such as the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights, HRL protects a person's fundamental human rights against both their own government and other international actors who may attempt to violate those rights. These rights include the right to life, liberty, and security of person and the prohibition of cruel, inhuman, and degrading treatment. HRL thus questions the lawfulness of targeted killings and indefinite detention of terrorists, among other preventive techniques employed by states. Consequently, for example, outside of combat, UN directives supplement HRL principles and support the need for imminence and necessity and a lack of available alternatives in order to use deadly force by military personnel or for preventive detention to be allowed in the face of arguments that it deprives the alleged terrorist's right to basic due process.

Since 1963 and the rise of modern terrorism, the international community has adopted 14 major legal instruments governing responses to terrorism, including conventions and protocols. Among other things, these conventions criminalize the use of civilian aircrafts and ships as weapons, define and prohibit the taking of hostages, regulate the possession and use of nuclear weapons, and allow universal jurisdiction over terrorist bombings. In 2006, the UN member states also adopted the Global Counter-Terrorism Strategy, the result of a General Assembly Resolution and a Plan of Action. This instrument attempts to coordinate efforts to address the conditions conducive to the spread of terrorism, determine both how to prevent and combat terrorism and how to build States' capacity to do so, and to create measures to recognize human rights and the rule of law as the fundamental basis for the fight against terrorism.

In addition to the UN Charter, IHL, HRL, and the various conventions, treaties, and international agreements focused on countering terrorism, the counterterrorism practices of nation states have begun to create customary international counterterrorism law. Customary law requires consistent and recurring action by states, along with a general recognition by states that the actions taken are required by international law. Finally, domestic courts, military tribunals, and international courts make counterterrorism law as a by-product of their decisions in individual cases. Although the courts typically intend narrow rulings directly affecting only the parties before them, the principles embodied in the courts' holdings often influence the development of counterterrorism law going forward.

Law Enforcement Approaches to Terrorism

General

The law enforcement approach to terrorism derives from the same deterrence theory that forms the basis of criminal law in general. The risk of arrest and punishment for violating the law deters individuals from violating that law in the future. Questions linger about how effective deterrence actually is against terrorists motivated by the righteousness of their cause. Nonetheless, there are several costs and benefits associated with law enforcement methods of combating terrorism.

When a terrorist is arrested and prosecuted, like Ramzi Yousef, the mastermind of the 1993 World Trade Center bombing, there is the direct benefit that he will not strike again. His arrest and prosecution in our courts also has indirect benefits. The United States gains a moral high ground by adjudicating alleged terrorists through the same judicial process as any ordinary crime. Moreover, the arrest of one terrorist may also have the effect of deterring others who might follow in his footsteps.

However, there are serious drawbacks to the criminal prosecution of terrorists. In addition to the uncertain deterrence effect, there is the risk that prosecution will provoke retaliation in the form of new terrorist attacks. Second, the public nature of criminal trials may serve as a pulpit for prominent terrorist suspects to draw further attention to the terrorist's cause. Third, criminal prosecutions are typically

successful only against the operatives who carried out the attack, not the leaders or masterminds of the terrorist organization since those individuals are usually far removed from the operations themselves.

Additionally, there are inherent difficulties in criminal prosecutions in general. The 'beyond a reasonable doubt' standard of guilt sets a very high burden of proof on prosecutors for any crime. Also, Constitutional protections require prosecutors to divulge to the defendant the evidence it plans to use, including the identity of intelligence sources, how the information was collected, and any evidence that may be favorable to the defense. Since many terrorism cases involve classified information, the government does not want to turn over such sensitive information to suspected terrorists. Two final challenges are that some foreign governments refuse to cooperate with US prosecutions where the death penalty could result, and that obtaining jurisdiction over suspects outside the United States is often legally and practically complicated by unclear laws and by the complex political and diplomatic considerations foreign nations must weigh when deciding whether to cooperate with a US terrorism prosecution.

Relevant US Statutes

The use of law enforcement in the United States to counter-terrorism is a modern phenomenon. This is partly explained by the fact that few acts of terrorism occurred on the US soil until the 1990s.

State-sponsored terrorism in the 1990s prompted the United States to impose sanctions and embargoes through the president's emergency powers, further enabled by legislation. The Trading with the Enemy Act, a World War I era law, criminalized certain commercial activities with enemies of the United States during wartime. It also granted the president, the power to restrict a range of other economic transactions with enemies and allies. The Act was amended between the world wars to grant the president its authorities whenever an emergency arose. During the 1970s, Congress reined in the president's executive discretion and created new procedures for exercising emergency powers by modernizing those authorities in the International Emergency Economic Powers Act (IEEPA). The IEEPA allows the president to impose economic sanctions during emergencies, which include 'any unusual and extraordinary threat' from outside the United States to the national security, foreign policy, or economy of the United States.

Current criminal laws on terrorism in the United States grew from these earlier sources of executive discretion to punish state sponsors of terrorism. Still, there is no single or uniform crime of 'being a terrorist' or 'engaging in terrorist activities.' Instead, terrorism crimes have developed over time, usually created in response to a terrorist attack, outlawing specific types of terrorist activities.

The decade between 1975 and 1985 saw a dramatic increase in terrorist violence directed against the United States with 2500 terrorist acts directed at Americans or US interests abroad. In 1986, the Omnibus Diplomatic Security and Antiterrorism Act was the nation's first effort at comprehensive terrorism legislation. For the first time, this Act made it a crime to harm Americans abroad. The focus of the act was on international terrorism, defined as violent acts, or acts dangerous to human life, intended to intimidate a civilian population. Additionally, the Act aimed to bolster security for diplomatic missions and to increase maritime security. It also authorized the offering of rewards for tips leading to the arrest of a suspected terrorist, and created a means to compensate victims of terrorism.

The 1993 World Trade Center attack, the 1995 bombing of a US training center in Riyadh, Saudi Arabia, and the 1995 bombing of the Oklahoma City federal building prompted Congressional action for more comprehensive legislation. The Antiterrorism and Effective Death Penalty Act of 1996 (AEDPA) considerably expanded the number and scope of federal crimes relating to terrorism, and increased the penalties for some existing offenses. Significantly, the AEDPA demonstrated that the United States would use legal mechanisms not only to prosecute terrorist acts after they occur, but also to anticipate and prevent terrorist acts.

In part, the AEDPA authorizes a variety of financial deterrents beyond traditional criminal law mechanisms, such as freezing and forfeiture of assets, prohibiting financial transactions with state supports of terrorism, and designating terrorist groups as foreign terrorist organizations (FTOs). Once a foreign group is designated as an FTO, its members are denied entry to the United States and the group's assets in the United States are frozen.

The most widely used component of the AEDPA criminalizes provision of 'material support' to FTOs. Particularly, after the attacks of 11 September, 2001, material support crimes became a central tool in combating terrorism. However, the anticipatory nature of regulating or proscribing acts of support leading up to a terrorist attack runs the risk of overly aggressive enforcement, punishing innocent acts such as expressive conduct, which may be constitutionally protected. The courts and Congress have revised and updated the material support provisions several times since 1996, following constitutional challenges to convictions under the Act. While the material support provision offers new broad authority, the spike in prosecutions following 9/11 returned to pre-9/11 numbers within 5 years of the attack.

Within weeks following the 9/11 terrorist attacks, the US Congress enacted the USA PATRIOT Act (United and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism). Lengthy and complex, the legislation was rushed through Congress with little debate and thus lacks clarity and leaves many issues unresolved. While heavily criticized, especially by civil liberties groups, the Patriot Act made only modest substantive additions to criminal laws. In fact, despite its 342 pages, the Act only contained a small number of controversial changes in the law, and those changes were subject to sunset provisions requiring subsequent legislation in order to remain in effect.

In brief, the Patriot Act enhanced the investigative tools considered necessary for the prevention of terrorism and updated the technological means of keeping track of terrorist activities. Where the Patriot Act did add to criminal laws, it did so by expanding and clarifying some of the definitions of terrorist activity, by adding variations to existing crimes, and by adding a few new crimes, including terrorist acts against mass transit, harboring terrorists, and acts of 'domestic terrorism.'

Detention, Interrogation, and Prosecution

Antiterrorism tactics should be distinguished from counterterrorism approaches. Antiterrorism actions are defensive measures used to reduce the vulnerability of individuals and property to terrorist acts. In contrast, counterterrorism is a broader term and encompasses actions taken directly against terrorist networks and other approaches that indirectly seek to render global and regional environments inhospitable to terrorist networks.

Detention

In times of armed conflict, the Geneva Conventions and IHL provide the legal guidelines applicable for detaining captured enemy combatants and for assuring their prisoner of war (POW) status. However, the evolution of modern warfare to so often include nonstate terrorists has complicated the definition of enemy belligerents, thus making it more difficult to definitively classify a detainee and assign him or her a lawful status and the rights that go along with it. With the passage of Additional Protocol I, POW protections were limited to enemy belligerents who openly carry arms and comply with the laws and customs of war, including the wearing of fixed insignia and organization of their force through a chain of command. These changes, in conjunction with the reduction in the number of international armed conflicts and rise in conflicts involving nonstate armed groups, have allowed states to individually re-create their obligations toward enemy combatants and the statuses of these individuals. For instance, the United States refuses to offer POW status to any enemy belligerent unless the person has been captured during an international armed conflict and satisfies the Geneva Convention requirements. Given that suspected terrorist groups seldom meet these measures, the United States has declared itself bound only to minimal HRLs in regards to detention and interrogation.

The United States allows for the detention of suspected terrorists and individuals who potentially possess knowledge of a terrorist activity, prior to the commission of any crime. This preventive detention has raised multiple Constitutional issues, in particular Fourth Amendment search and seizure and Fifth Amendment due process rights challenges, but the practice has been largely upheld in the interests of national security. Other countries' detention methods have received similar criticism. For example, the United Kingdom allows suspected terrorists to be detained without charge for up to 28 days, despite critiques that the practice violates the European Convention of Human Rights, which requires any suspect to receive prompt notice of the charges against them.

Aside from the states' domestic laws, international law also provides detainees with certain rights. Under the Vienna Convention on Consular Relations, an arrested foreigner has the right to have his government notified of the arrest. The Universal Declaration of Human Rights prohibits arbitrary arrest or detention. The protection against arbitrary detention is also included in the International Covenant on Civil and Political Rights, which also provides the detainee with the right to appear before a court to determine the lawfulness of the detention.

Interrogation

Reliance on torture to extract confessions is at least as old as ancient Greece. In intelligence investigations that take place in the aftermath of a terrorist attack, investigators do not know in advance what the detainee knows that may be of value. It has been very difficult to set legal rules to guide this uncertain quest for information. If there is a human tendency to extract information by any means necessary, it may be that abusive interrogation is the inevitable result of any legal regime short of an absolute ban on coercive interrogation. When the grotesque images of abhorrent abuses of detainees in US custody at the Abu Ghraib prison in Iraq became prime time news in 2004, and follow-on investigations revealed coercive interrogations conducted at the Guantanamo Bay detention facility by US agents during the same period, investigators learned that the then-existing international and domestic laws provided inadequate protections for detainees and insufficient guidance for interrogators in the conflict against nonstate terrorist suspects and insurgents. Debates raged over the practice of waterboarding and the hypothetical ticking bomb scenario. Political and moral issues punctuated discussions of law reform.

International law has long prohibited torture, as a fundamental principle of customary international law, through the Universal Declaration of Human Rights, the International Convention on Civil and Political Rights, the Geneva Conventions, and the UN Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (Convention against Torture). While most states have adopted the Convention against Torture, its critical terms – 'torture,' 'cruel, inhuman, and degrading treatment' – are not precisely defined. Despite the reform of US and some other states' domestic laws since 2005 codifying the international law bans on torture and cruel, inhuman, and degrading treatment, some government officials and interrogation practitioners continue to argue that enhanced interrogation techniques are warranted in light of the potential terrorist threats. (For a thorough discussion of these issues, *see* Torture: Legal Perspectives.)

Preventive Prosecution

Particularly after 9/11, US prosecutors utilized the technique of preventive charging against suspected terrorists. Preventive charging involves prosecutions based on actual law violations, although those violations are often far removed from the actual terrorist threats officials are attempting to disrupt. This technique is sometimes referred to as the 'Al Capone approach' (the Prohibition era gangster of Chicago was responsible for murder, extortion, and innumerable other violent crimes, but was ultimately arrested and convicted for tax evasion), or the 'spitting on the sidewalk approach,' in which a suspect is prosecuted for violating an innocuous municipal ordinance. At first blush, preventive prosecution appears harmless. However, following 9/11, the strategy spurred controversy over allegations that the so-called PENTBOM investigation (the Federal Bureau of Investigation's 'Pentagon/Twin Towers Bombing Investigation') was implemented primarily against noncitizens, particularly Middle Eastern men.

Despite risks of abuse, preventive prosecution still necessitates an actual violation of law. When the suspicious person has

not committed a crime, an alternative prevention strategy exists in the United States and some other states that allows for the detention of someone deemed to be a 'material witness' to a criminal prosecution in order to assure that he or she will be available to testify. Upon an assertion by the government, a person who could offer testimony material to the criminal prosecution of a terrorist can be detained indefinitely, so long as a supervising judge periodically affirms the detention is justified.

Rendition/Extraordinary Rendition

Rendition involves the surrender of a person from one nation to another nation that has requested that person, usually for criminal prosecution. However, the forms of rendition can range from informal, ad hoc occurrences to a formalized process such as an extradition treaty between two or more states. An extreme form of rendition – extraordinary rendition – involves situations in which, for example, the United States transfers detainees to foreign countries in circumstances in which it is likely that the individuals will face torture or cruel, inhuman, or degrading treatment.

While traditional informal rendition is characteristically regulated by legal agreements between states, extradition generally involves the use of treaties. Therefore, extradition typically requires 'dual criminality,' that is, the crimes for which the requesting state seeks extradition for prosecution must also be crimes in the state that would extradite the suspect. Thus, in a terrorism context, extradition treaties are often part of broader formalized agreements between nations to cooperate in criminalizing, investigating, and prosecuting suspected terrorists.

Extraordinary rendition was employed by the United States and perhaps other states after the 9/11 attacks, at least in part to avoid the domestic laws limiting interrogators. By implication, the international laws that forbid torture and cruel, inhuman, and degrading treatment – the Convention against Torture and the Geneva Conventions, among others – were violated in these instances. For the victims, it has been difficult to obtain remedies in the domestic courts of the state perpetrators of the renditions. Some have had limited success in obtaining redress from other government institutions, and in one case from an international court.

International Cooperation

Since the threat of terrorism does not respect national borders, countering terrorism requires a set of transnational responses. International cooperation to combat terrorism involves different states and international organizations working in a wide range of areas. Diplomatic efforts seek to develop and strengthen international prohibitions of terrorist acts and to encourage other nations to sign on to the international treaties against terrorism. Through economic cooperation, nations aim to eliminate financing capabilities of terrorist organizations and to use economic pressure to punish states or nongovernmental actors that support terrorists. Nonetheless, international cooperation against terrorism can be problematic.

First, states have different perspectives on terrorism. Seemingly unending disagreement on how to define terrorism and on whether terrorist acts can be justified in some circumstances makes international consensus difficult. Second, the types of terrorist threats each nation faces are different and can lead to diverging views on which responses are appropriate. Third, the capacity of each state to respond to terrorism can be wide ranging, further affecting its calculus of appropriate responses.

In spite of the challenges, between 1963 and 1999, 12 international conventions were established, each addressing different aspects of terrorism. Because of ongoing disagreements, these conventions tend to outlaw specific terrorist acts, rather than banning terrorism altogether. For instance, following a rash of airplane hijackings in the 1960s, an international treaty was established in 1970 outlawing the unlawful seizure of aircrafts. Additionally, the UN Security Council has passed numerous resolutions imposing sanctions on state sponsors of terrorism and restricting the financing of terrorism. Throughout the 1980s and 1990s, it imposed sanctions on states including Sudan, Iran, Iraq, and Libya. In 1999 and 2000, the Security Council also passed resolutions demanding that the Taliban regime in Afghanistan surrender Osama bin Laden to international authorities for prosecution. Following the 9/11 attacks, the Security Council passed Resolution 1373, "which called on all states to deny financial or other support for terrorist groups, to deny them safe haven, to share information and to assist other states in tracking down terrorists, to criminalize in their domestic laws both active and passive assistance for terrorists, and to enforce these laws." In response to the discovery of Dr Abdul Qadeer Khan's illicit nuclear supply network to North Korea, Iran, and Libya in 2004, the UN Security Council passed Resolution 1540, criminalizing the spread of weapons of mass destruction, especially to terrorist groups. Both Resolution 1373 and Resolution 1540 were adopted under Chapter VII of the UN Charter, which makes compliance with their provisions binding on all states.

These international treaties function to set standards for domestic laws. They commit nations to harmonize their national laws with the international standards those treaties establish. Through this mechanism, the international community can work toward creating global norms. In addition to pressuring signatories to adapt domestic laws to conform to treaties, international conventions help shape global public opinion, particularly by promoting an international consensus that terrorism is illegitimate.

Military Action in Response to Terrorism

Military force can be used in several ways in responding to terrorism. As a deterrent or preemptive measure, military forces can debilitate physical and economic assets in an attempt to incapacitate a terrorist group's ability to attack or make the costs of an attack too high. Military force is also used in self-defense, sometimes resulting in questions of violations of state sovereignty when the defending state crosses sovereign borders to attack a terrorist group. While the UN Charter, Article 51, permits states to use self-defense in response to an armed attack, if the government of the state to be attacked is not

supporting the terrorist group in some way, a law violation and international condemnation may follow.

The use of military force to combat terrorism has been employed with growing frequency. However, while the United States and other Western states can use conventional military operations to combat traditional state enemies, military operations are more difficult when fighting terrorists. The military often cannot differentiate between civilians and terrorists, due to the lack of insignia and traditional military chain of command, not to mention terrorists' failure to obey the laws and customs of armed conflict. Furthermore, the deterrent value of responding to terrorism with military force can be difficult to assess in instances where the terrorists have committed to death, such as in suicide attacks.

The United States is emerging from major military operations in Afghanistan and Iraq. While the intervention in Afghanistan was a direct response to the 9/11 attacks and initially seen as a self-defense measure undertaken eventually with the permission of the Afghan government and originally primarily composed of US Special Forces personnel, the intervention in Iraq was initiated on more questionable legal grounds – preempting a supposed threat of the use of weapons of mass destruction by the Iraqi government – and consisted of a full-scale military operation. The wars in Iraq and Afghanistan in general and the conflict against al-Qaeda in particular highlight several challenges when combating terrorist groups through the use of military force. First, while the United States and most western states have the capability to defeat conventional forces, terrorist forces rely on the techniques of asymmetric warfare – fighting from civilian areas and taking shelter in civilian dwellings, places of business and worship, for example – that is much more difficult to combat in operational terms. The asymmetric tactics also make it difficult for the defending state to comply with the laws of war, and may engender more distrust among the affected civilian population, or even stiffen the resolve of civilians caught in the cross fire to take up the terrorists' cause.

Throughout what has become a kind of shadow war by North Atlantic Treaty Organization (NATO), the United States, and its allies against al-Qaeda and its affiliates, military operations occur outside the traditional battlefields of Afghanistan and occur in several states in the Middle East/North Africa region. Because the nonstate terrorists often operate from within weak, failing, or sympathetic states, the use of force by states defending against ongoing threats of terrorist attacks often is criticized as violating territorial sovereignty and constituting offensive aggression. The defending states respond by maintaining that self-defense may lawfully be executed in the sovereign territory of another state where that state has proven itself unable or unwilling to remove the terrorist threat. As this unresolved legal quagmire continues, state practice may come to recognize the lawfulness of this exception to the integrity of sovereign borders.

The use of lethal force against designated individual terrorists, particularly by missiles delivered by unmanned aerial vehicles (UAVs), has proven especially controversial. The use of targeted killing through the use of UAVs has become particularly appealing due to the minimal threat of harm to US forces (the UAVs are piloted from remote locations, often far away from any battle space) and the possibility of fewer civilian casualties than through the use of conventional forces (because of the relative precision in targeting that may occur by employing powerful cameras and imagery technologies). Like the shadow war generally, targeted killing by UAVs outside traditional battlefields risks violating state sovereignty and arguably generates as much or more opposition to the defending state as existed before the drone strikes killed their targets.

Finally, counterinsurgency (COIN) operations became popular in Iraq and then Afghanistan for a time, undertaken by NATO, the United States, and allied forces in an attempt to garner support for military operations against terrorists and insurgents from local populations and curtail the spread of terrorist influences within the country. These operations, including fixing roads, building schools, and providing medical services, are primarily conducted by special operations forces and other military groups due to their high level of danger. COIN operations may be conducted in a sort of legal netherworld between war and peace. At times, the military forces may be involved in combat operations as part of COIN, while at other times they may be building local infrastructure, cultivating community support, or even negotiating with local fighters. The law of COIN is in its infancy, and it blends elements of the laws of war (IHL), HRL, customary international law, and the operational law of the engaged military forces. The military activities have drawn criticism from nongovernmental organizations and aid groups, who argue that they are being put at risk by the 'militarization' of aid and that neutrality is needed to successfully protect civilians.

See also: Civil Liberties and Human Rights; Crime: Organized; Empirical Legal Studies; Human Rights, Anthropology of; Human Rights: Legal Aspects; Military History; Policing; Torture: Legal Perspectives; Women in Buddhism.

Bibliography

Banks, William C. (Ed.), 2011. New Battlefields/Old Laws: From the Hague Conventions to Asymmetric Warfare. Columbia University Press, New York.

Banks, William C. (Ed.), 2013. Counterinsurgency Law: New Directions in Asymmetric Warfare. Columbia University Press, New York.

Banks, William C., deNevers, Renee, Wallerstein, Mitchel B., 2008. Combating Terrorism: Strategies and Approaches. CQ Press, New York.

Bianchi, Andrea (Ed.), 2004. Enforcing International Law Norms against Terrorism. Hart Publishing, Oxford.

Carlton, David, Schaerf, Carolo (Eds.), 1975. International Terrorism and World Security. Croom Helm, London.

Dycus, Stephen, et al., 2011. Counterterrorism Law, second ed.

Gurulé, Jimmy, Corn, Geoffrey S., 2011. Principles of Counter-terrorism Law.

Hoffman, Bruce, 2006. Inside Terrorism, rev. ed. Columbia University Press, New York.

Ramraj, Victor, et al. (Eds.), 2012. Global Anti-terrorism Law and Policy, second ed. Cambridge University Press, Cambridge.

Solis, Gary D., 2010. The Law of Armed Conflict: International Humanitarian Law in War. Cambridge University Press, Cambridge.

Stern, Jessica, 1999. The Ultimate Terrorists. Harvard University Press, Cambridge, MA.

United Nations Global Counterterrorism Strategy, G.A. Res. 60/288, 2006.

U.N. General Assembly, 13 October 2010. The United Nations Global Counter-Terrorism Strategy. U.N. Doc. A/RES/64/297.

Law: Anthropological Aspects

Sally E Merry and Matthew C Canfield, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S.E. Merry, volume 12, pp. 8489–8492, © 2001, Elsevier Ltd.

Abstract

Anthropologists approach law as a form of normative ordering that includes rules and processes. The anthropology of law has evolved significantly from viewing law as a measure of civilization and savagery to understanding law as an emergent cultural feature and form of power. Contemporary anthropologists investigate law ethnographically, by looking at disputes in spaces that lack formal legal institutions and codified law, as well as in the institutional sites where law is produced. Plural legal orders, each invested with particular modes of authority, are seen as a feature of all societies. With the growth of globalization and transnational law, anthropology has extended its scope to explore the dialectical interaction between these multiscalar forms of law.

Anthropologists approach law from a comparative perspective and examine it in social context. 'Law' refers broadly to the rules, processes, and norms that regulate social life. Although there are frequently institutionalized systems for making decisions and enforcing them, legal anthropology does not restrict its attention to formal systems of courts or to the law of the nation-state. Anthropological approaches to law examine the legal ordering of villages, bands, and urban neighborhoods – places that usually lack specialized legal institutions and codified rules – as well as a wide range of institutional systems and technologies of law.

The anthropological study of law began from nineteenth-century evolutionary theories in which the form of law revealed the stage of social development. Twentieth-century legal anthropology examines processes of resolving disputes and maintaining social order on multiple scales. While interpretive approaches differ, law is understood as a form of power, invested with particular forms of authority. Recent scholarship has examined the relationship between law and culture, with special attention to the role of language. In examining law through its broader social and normative contexts, anthropologists have focused on the legal complexity produced by intersections among plural legal orders rooted in the community, the state, and international institutions and agents.

Historical Developments in the Anthropology of Law

The great nineteenth-century theorists of social evolution understood law as a measure of civilization and savagery. In *Ancient Law*, Sir Henry Maine (1861) theorized that with the transformation from kin-based to territorially based societies, law shifted its normative basis from status to contract. Similarly, sociological theorists such as Émile Durkheim and Marcel Mauss suggested that law emerged from an embedded form of repressive sanctions to an instrument arranged by specialized tribunals and functionaries that serves to maintain cooperation. These early theories identified law as a key facet of civilization, raising questions about the universality of law and the different forms it may take within distinct social and cultural contexts. As British social anthropologists pioneered the ethnographic method, they took up these questions and debated the very definition of law.

During the early twentieth century, two major schools of thought developed. One school followed from the work of Bronislaw Malinowski (1926), who defined law in functional terms, identifying it with a broad range of social processes that produce social conformity through social pressure and inducement. Through his fieldwork in the Trobriand Islands, Malinowski located law within the domain of everyday life, rather than the state or other institutions. He defined law in terms of the relationships of reciprocity that he argued were a fundamental mechanism for maintaining social control. With Malinowski's definition, every society has law. Critics of this view contend that defining law as social control is too broad and includes practices such as gossip, ostracism, and shaming that are not lawlike.

A second school took a more structural approach to define law in terms of institutions, requiring socially authoritative mechanisms for enforcing rules through the imposition of sanctions. Radcliffe-Brown pioneered this perspective, defining law as "social control through the systematic application of the force of politically organized society" (1933, p. 202). Under this definition, all societies have customs and sanctions, but not all societies have law. E.A. Hoebel defined law more broadly, but also within an institutional context, as "the legitimate use of physical coercion by a socially authorized agent" (1954, p. 26). In *The Cheyenne Way*, Hoebel and legal scholar Karl Llewellyn developed a case-method approach to look at 'trouble cases' (1941). This early collaboration demonstrated the importance of anthropological approaches for legal scholars and other social scientists. As progressive legal scholars associated with the legal realist school critiqued formalist or positivist jurisprudence – the then-dominant paradigm among legal scholars and practitioners – they turned to anthropologists to demonstrate the cultural effects of law. Critics of this view of law, however, argue that this rule-based perspective is based on Western conceptions of law (Comaroff and Roberts, 1981, pp. 6–7).

Between the 1930s and 1960s, anthropologists sought to describe the nature of rules of whole societies. For example, Isaac Schapera attempted to develop an entire casebook of

Tswana law based on his ethnographic research, while Leopold Pospisil sought to provide a more far-reaching universalistic definition of law based on his fieldwork with the Kapauku Papuans. Yet these scholars confronted the paradox that there were frequently discrepancies between these rules and everyday forms of behavior. Moreover, anthropologists debated the very nature of the comparative method?. While Max Gluckman argued that indigenous legal concepts processes could be translated into the terms of Western jurisprudence, Paul Bohannan countered that these terms could be understood only in reference to the total social and legal systems of which they were a part. This debate, which went to the heart of the definition of law and the nature of legal comparison, became increasingly sterile. By the 1970s, this debate was largely abandoned in favor of understanding law as a social process (see Moore, 1973; Nader, 1997).

Law and Dispute Processes

In the early 1970s, anthropologists turned to studying processes by which conflicts are handled and resolutions achieved (see Moore, 1978). Laura Nader and her students researched how litigants pursued disputes and how these disputes were resolved in a wide variety of societies (Nader and Todd, 1978). This approach used extended case study analysis and emphasized the perspective of litigants rather than judges. It broke dramatically with the earlier focus on rules and looked instead at the processes by which individuals pursued grievances. At the same time, it built on a tradition of analyzing extended case studies in the context of social relationships. For example, Max Gluckman argued that close-knit villages tended to have compromise settlement processes, while more loose-knit ones favored more adversarial processes (Gluckman, 1973). The focus on dispute processes remains a basic framework for the anthropology of law.

By the early 1980s, however, it became clear that a focus on the processes of resolving conflicts could not be separated from a consideration of rules and how they were formed. In their study of the Tswana, Comaroff and Roberts found that the meaning and relevance of rules are negotiated in disputes (1981, pp. 5–17). Arguments among disputants are framed by rules, but the content of rules is negotiated as disputes are argued and resolved (1981, p. 19). This more interactive understanding of rules and processes has proved valuable for examining the operation of law in communities as well as in more formal settings, such as courts. It led to an approach that conceptualized law as more deeply embedded within culture, as a system that was both constituted by and constructive of social meaning.

Law, Culture, and Language

The legal system of a society reflects its fundamental values and its normative order (Rosen, 1989). Geertz (1983) describes law as a form of 'local knowledge' that is 'a distinct way of imagining the real.' In developing an interpretive, comparative approach, Geertz argues that anthropological analyzes should approach law as a normative process of social representation. Law is not merely constitutive of social reality, but also reflects particular ideals of justice. To capture this dialectic, Geertz suggests that the comparative study of law should consist of the study of 'cultural translation.' In analyzing this process of translation, anthropologists have pointed out that law is essentially a language and that legal procedures are fundamentally linguistic.

Following a processual model of analyzing law, Sally Humphreys (1985) suggests that law can be studied as a form of discourse that is embedded in particular notions of power. Taking up this approach, scholars developed a focus on law and language that combines sociolinguistic analysis and law and society scholarship to look at the way that law creates meanings and exercises power through linguistic exchanges in courtrooms and other legal settings (Conley and O'Barr, 2005). This work provides a way to look at the detailed interactions that determine how the law exercises power. For example, many features of the kinds of talk that are persuasive in American courtrooms are more characteristic of male than female speakers, and of powerful rather than powerless speakers, so that male litigants in these courtrooms have an advantage (Conley and O'Barr, 2005). Analyzing the microdiscourses that take place in mediation centers, courtrooms, and other legal settings, scholars have applied insights drawn from linguistic anthropology, such as conversation analysis, to show how macrosociological relations of power are constituted through techniques such as controlling silences and suppressing arguments.

Legal language and institutions also offer avenues for resistance. In particular, the language of rights has proved a powerful framework for social justice movements, especially since the end of the Cold War. Rights talk has been central to labor movements, indigenous people's movements, the anti-apartheid movement in South Africa, and movements for women's rights, to name a few examples. Local groups around the world now draw on global conceptions of human rights for inspiration. Much contemporary social justice talk and activism take place within the domain of law. Anthropologists have questioned whether the proliferation of formal rights language offers a useful form of resistance in the diverse and complex struggles that individuals and groups are engaged in to contest culturally constituted inequalities in diverse regions of the world. Merry's (2006a) investigation of the Convention to End All Forms of Discrimination against Women focused on the ways that elites translate between the vernacular language of duties and responsibilities and the global language of human rights. Although human rights must be vernacularized to effectively change relationships of power within a given society, she argues, the language of human rights nevertheless "comes with a price. Human rights promote ideas of individual autonomy, quality, choice and secularism even when these ideas differ from prevailing cultural norms" (2006a, p. 4). This work has described the deeply embedded cultural and political conceptions within legal language as well as the limits of legal language as a form of social representation (Hastrup, 2003). As the language of law is deployed within different cultural milieus, anthropologists have attended to the ways that people wrestle with the legal categories through which they must make their claims.

Anthropologists have also drawn from linguistic anthropology to track the ways in which law serves as a particular ideology that is embedded within normative epistemologies. Elizabeth Mertz describes how the pedagogy of US law schools serves to socialize law students into a particular, 'culturally driven model of justice.' Students are systematically 'urged to pay attention more to abstract categories and legal (rather than social) contexts' that serve to cultivate the authority of textual analysis and legal form (2007, p. 5). Mertz shows while legal language and processes offer a creative system for handling disputes, they may also privilege particular cultural viewpoints over others. Similarly, Justin Richland (2008) describes how language ideologies often conflict within Hopi jurisprudence. While he too notes how legal language serves to socialize a particular population into the ideology of law, he also shows how these ideologies may be transformed over time through mutually constitutive interaction with divergent worldviews. Through close linguistic and interactional analysis within a Hopi courtroom, Richland examines the way in which Hopi judges and lawyers attempt to develop a Hopi form of jurisprudence based on 'tradition.' He suggests that Hopi law and ideas about tradition are constituted through a dialectical tension between Hopi and Anglo-American concepts of objectivity and forms of narrative.

Law and Power

The law also provides a rich vocabulary for describing and legitimating relationships of unequal power. It contributes in significant ways to the hegemony of a political system by engendering consent and fostering legitimacy for its exercise of power (see Nader, 1991). Since the 1980s, the anthropology of law has become concerned increasingly with the relationship between law and the exercise of power (see Starr and Collier, 1989). Inspired by Marxist and Foucauldian theories as well as the Critical Legal Studies Movement in law schools that argued that law reflects the interests of dominant groups rather than simply the logic of legal reasoning, anthropologists examined the way law supports relations of power in a wide range of social contexts (see Lazarus-Black and Hirsch, 1994). However, this research also demonstrated how law challenges dominant groups. It showed law to be a double-edged sword, expressing the rules and sanctioning authority of established political authorities and at the same time offering opportunities and justifications for resisting political authorities.

For example, social movements that call on civil rights, such as the US civil rights movement against racial discrimination in the 1950s and 1960s, challenge existing power hierarchies. Feminists have used the criminal law to challenge male privilege by demanding penal sanctions for those who use violence against women in intimate relationships. The discourse of human rights has been deployed by indigenous peoples, women, victims of war atrocities, prisoners of war, and many other groups to claim protection from state authorities. Thus, law usually reinforces existing power arrangements, but can provide an avenue of resistance through its institutions and the discourse of rights.

Attention to the way that legal forms and institutions exercise and legitimate power, while offering a mode of resistance, leads to concerns with the way that laws are produced and enforced as well as with the linguistic and social exchanges that take place in courtrooms, police stations, and lawyers' offices (e.g., Conley and O'Barr, 1990). Increasingly, anthropologists have examined the technologies and forms of law itself. Rosemary Coombe (1998) analyzes the regime of international property law to reveal the underlying conceptions of individualism and agency implicit within these laws. She documents how, as Canadian First Nations engage with the language of intellectual property laws to contest the 'cultural appropriation' of native knowledge and symbols, they also resist its underlying possessive individualism. Indigenous groups deploy a 'double-voiced rhetoric,' whereby "they empty tropes of a dominant language simultaneously engaging and subverting these metaphors through the character of the alternative claims they make into the voice of an authorial other" (1998, p. 242).

Another approach to understanding law and power integrates insights from Science and Technology Studies to examine the content, form, and agents involved in the production of legality as a technology and network. This approach does not focus directly on the relations of power that underlie law-making processes. Instead, it shows how technologies of law, in the form of doctrine, interpretive techniques, or dispute resolution tools, produce forms of knowledge that conceal more substantive, political issues. Both Bruno Latour (2010), in his study of the French Conseil d'État, and Annelise Riles (2001), in her study of Fijian activists negotiating human rights language for the United Nations Fourth World Conference on Women in 1995, seek to open up the 'black box' of law. This is Latour's term for complex techniques that scientists present in diagrams as taken for granted rather than present in detail. Latour and Riles both analyze the details of legal processes to show how technologies of lawmaking become tools to overcome political and epistemological limits. Whether studying the objectivity of judges or the use of 'bracketing' as a way to move past substantive disputes in the United Nations, each of these technologies produces knowledge that is translated into commonplace understandings. Similarly, Merry (2011) examines new tools of international governance, such as the use of quantitative indicators to measure qualitative categories such as 'the rule of law,' as technologies that transform political decisions into scientific, empirical knowledge (see also Davis et al., 2012).

Scholars have also looked to the way that law imposes new forms of power through constructing and transforming notions of space. Drawing on the work of cultural geographers, these studies recognize space as a key constituent of social life that plays an important role in the construction of social relationships and the formation of the political domain. Since law is the apparatus through which states define political territories, as well as their domestic and public spheres, it is a key form through which space is made (see Law and Geography). Benda-Beckmann et al. (2009) explore the relationship between competing spatial configurations of law within plural legal orders. By exploring the plural legal orders and their particular notions of space, they argue that "multiple legal constructions of space open up multiple arenas for the exercise of political authority, the localization of rights and

obligations, as well as the creation of social relationships and institutions that are characterized by different degrees of abstraction, different temporalities and moral connotations" (2009, p. 4).

Legal Consciousness and Identity

Anthropologists have investigated the relationship between law and power in the construction of personhood and subjectivity. Law constitutes social identities and understandings of social relationships through its regulations and through the spectacular as well as the mundane processes of arrest, trial, and mediation through which these regulations are enforced and failure to comply is punished. Even when laws do not impinge directly on social life, they exercise power by shaping individuals' legal consciousness. This refers to the way in which people see themselves defined by and within law, the rights they can assert, and their entitlement to seek legal remedies for grievances in studies of legal consciousness, anthropologists use disputes to investigate these matters. These studies approach disputing as "a contest over meanings in which the law provides one possible set of meanings. The process of disputing is one of quarreling over interpretations of social relationships and events" (Merry, 1990, p. 6). Studies of legal consciousness illuminate the way that everyday understandings of law become important factors in the development of 'community' and social relations. Merry's work on working-class, urban legal consciousness showed how law becomes a means for challenging hierarchies of power and asserting privacy in opposition to social conflicts or constraints engendered by other forms of informal ordering. Similarly, in a collaborative effort to investigate the role of law in three American towns, Carol Greenhouse et al. (1994) suggest that law is a social practice that constructs the meaning of 'community' by defining 'insiders' and 'outsiders.' Exploring the reasons for litigation within their three respective sites of study, the authors found that law was either interpreted as a sign of difference and 'outsider' status, or as a form of 'insider' status by those seeking to defend the 'community.' Hence, they use the disputing approach to demonstrate the important role that law plays in constituting notions of community.

More recent studies of legal consciousness by scholars in other disciplines explore how people perceive injustice, understand themselves as rights-bearing subjects, and develop their identity based on new rights. Ewick and Silbey (1998) describe the schemas, or positional stances, that individuals develop to understand their relationship to the law. Laura Beth-Nielsen's (2000) research examines the everyday attitudes and understandings of law as a way of dealing with sexual harassment. David Engel and Frank Munger (2003) analyze how the disability rights movement constructs identities around what it means to be disabled and when and why disabled people mobilize these rights. David Engel and Jaruwan Engel (2010) document a surprising shift in legal consciousness concerning personal injury cases in Thailand as ordinary people turn away from the law.

Anthropologists have also focused on the way in which legal ideologies constitute conceptions of subjectivity that shape identity claims and understandings of the self. Law participates in the construction of identity by soliciting specific expressions of difference. By constraining the types of claims that can be made, Collier et al. (1996) contend that the law sanctions particular types of identity making. Rather than recognizing all difference as socially produced, they suggest that legal mechanisms recognize only those claims that translate their identity into 'natural' categories. Thus, formal legal mechanisms do not recognize the types of difference that law itself has produced. This is a contradiction that anthropologists have documented as groups seek remedies for state and social discrimination by state courts. For example, James Clifford (1988) describes how the Mashpee struggled for recognition as a 'tribe' under the Non-Intercourse Act in the United States. In this case, the court became a site to adjudicate a group's collective sense of identity based on legal categories rather than the self-understanding of the community.

Land claims have proved a particularly important site for anthropologists to investigate the relationship between identity and law. Elizabeth Povinelli (2002) describes how aboriginal land claims in Australia require authentic indigenous subjects who are "called on to perform an authentic difference in exchange for the good feelings of the nation and the reparative legislation of the state" (2002, p. 6). Yet there is an inherent contradiction, a 'cunning of recognition,' in that as the state recognizes aboriginal people and difference more broadly, it defines what constitutes an appropriate form of difference. People and groups must therefore represent themselves through these categories in order to win recognition of their legal rights. This, the process of recognition, extends the reach of liberal legalism. Jan French's study of Brazil (2009) shows how recognized categories of indigenity alter the lived experiences of people who adopt legally recognized categories for identity, reshaping their social relationships and claims to land.

The intersections between law and identity are increasingly central issues in the anthropology of law. Law constrains and limits identity claims by granting or denying citizenship to people who travel across transnational boundaries. The complex relationship that ensues when the law seeks to erase past forms of identification and construct new relationships to the state via identity is the focus for Susan Coutin's (2000, 2003) and Barbara Yngvesson's (2010) works, but in different contexts. Coutin's research on Guatemalan and Salvadoran immigrants to the United States and Yngvesson's research on transnational adoption both explore the ways in which adoptees and immigrants must reconcile forms of identity no longer recognized by law. These authors show the power and authority of law as an epistemological regime that confers particular forms of 'truth' to individuals in the process of self-making. Not only does law define identity, as Yngvesson (2006) and Coutin (2000) and both together (Coutin and Yngvesson, 2008; Yngvesson and Coutin, 2008) demonstrate, but also, as borders shift, identities can change even as people remain in place. Tobias Kelly's (2004, 2006) analysis of residents of the West Bank in Palestine illustrates this phenomenon vividly. With the development of research on legal consciousness, the relationships among legal culture, legal consciousness, and legal identity are increasingly complicated and overlapping, requiring clear definitions of terms (Merry, 2012).

Legal Pluralism

In the contemporary world, societies typically have more than one legal system. Legal pluralism refers to situations of legal coexistence in a single social field (Merry, 1988). Initially described in colonial situations, in which the colonial power superimposed a European legal system over an existing indigenous system, legal pluralism is now understood as a fundamental characteristic of all legal systems. Early insights into legal pluralism within advanced capitalist states were drawn from Sally Falk Moore's (1978) description of semiautonomous social fields, domains of social life that are self-regulated but also subject to larger forms of regulation, such as commercial networks or criminal communities. Legal pluralism has since been used to understand the multiple layers of regulation developed out of both private and public forms of governance. For example, many institutions such as schools and corporations develop forms of private governance. Informal but rule-governed relationships are also characteristic of families, communities, and voluntary organizations, although there is considerable debate about whether such forms of normative ordering should be called law. The term 'interlegality' emphasizes the interactions among multiple legal orderings (De Sousa Santos, 1987).

Early descriptions of plural legal systems examined each in isolation, but more recent work emphasizes the way that plural systems shape one another. Typically, plural legal systems are shaped by differing relations of power. In colonial situations, the colonial authority often established one legal system for the subjugated and another for the colonial authorities (Moore, 1986; Merry, 2000). However, as anthropologists examine postcolonial and contemporary international forms of governance, they focus on interactions among legal systems and their impact on litigants and on each other. For example, Richard Wilson's study of the South African Truth and Reconciliation Commission argues that the plural legal systems in South Africa represent different forms of authority and approaches to dealing with past injuries. Taking a processual approach, he suggests that studies of legal pluralism should focus on the way that the state and other actors seek to legitimate legal orders within the broader context of state formation and resistance (Wilson, 2001).

Globalization increases legal pluralism in two ways. First, it fosters the circulation of legal forms and practices from one nation-state to another. Constitutions, codes, and legal institutions are commonly transplanted, much as they were during the earlier period of colonialism. These transplants introduce codes and procedures quite different from those of the receiving society, often with the intention of reform. Colonial transplants characteristically were phrased as furthering civilization, mid-twentieth-century transplants as modernizing institutions, and late twentieth-century ones as promoting democracy and the rule of law. Ugo Mattei and Laura Nader (2008) argue that in the post–Cold War era, the 'rule of law' has been promoted as a nonpolitical, nondistributive technology that lacks the institutional virtues of democracy promotion that were central to efforts of Third World development during the Cold War. As such, the export of rule of law, they argue, has been a proxy for privatization and neoliberalization.

The growing field of 'transitional justice,' which refers to the wide array of processes and formal mechanisms that states establish to deal with past mass human rights violations, constitutes another form of legal pluralism. Transitional justice processes seek to reconstitute the state and restore the rule of law as a necessary precondition for democratic transition. Over the last two decades, several formalized approaches to transitional justice have been created. These approaches include criminal prosecutions, truth commissions, reparations, lustration, and memorialization. In theory, these approaches are complementary. States may choose to apply multiple approaches at one time, or may use several approaches successively. These approaches are used in a wide variety of political environments ranging from military dictatorships to ethnic civil wars. The field of transitional justice has engendered a large cadre of professionals who do this work around the world.

Considerable anthropological research in the field of human rights and transitional justice focuses on the way these global justice systems operate in practice. Anthropologists tend to link local ethnographic research with examinations of the global institutions promoting human rights and international justice. Mark Goodale (2009) explores the history of anthropology's engagement with human rights, while Hinton and O'Neill (2009) offer a series of anthropological perspectives on the way that local communities in many parts of the world experience transitional justice in response to genocide. Harri Englund's (2006) study of the way that human rights are understood and used in Malawi provides valuable insights into the difficulties of translating and implementing human rights in situations of autocracy and political opposition. Kimberly Theidon (2012) examines the workings of the Peruvian Truth and Reconciliation Commission from the perspective of highland Peruvians who suffered from the violence of the period of struggles between the Shining Path and government soldiers. Kamari Clarke's (2009) analysis of the International Criminal Court takes a local as well as a global perspective on the kinds of justice that this new international regime provides, suggesting that its local practice differs from its global aspirations. Ethnographic studies of the meanings of human rights in local communities show that they are one discourse among many, and often not the dominant one. For example, Daniel Goldstein (2012) shows how concerns about security take precedence over human rights in a poor urban neighborhood in Bolivia.

Secondly, globalization increases legal pluralism by constructing a global legal order that exists outside state law. As international law and international organizations expand, legal pluralism provides a valuable analytical framework for understanding this broadening range of legal forms. The human rights system, made up of rules and sanctioning systems embedded in international institutions such as the United Nations, is a central facet of the new global legal order (*see* Human Rights, Anthropology of). Although the human rights system is a multilateral treaty regime built on a system of sovereign states, it is redefining the nature of state sovereignty itself. Pressure for compliance with human rights by states and transnational nongovernmental organizations is circumscribing state sovereignty to some extent. At the same time,

domestic actors such as indigenous peoples seek to constrain state sovereignty as they make claims for legal and political autonomy using international principles of self-determination and rights to protection of culture.

Anthropological research explores the processes by which human rights categories are constructed. For example, Tobias Kelly (2011) examines a series of settings in which legal institutions determine whether or not a person has been tortured as a way of understanding what torture means. He shows that the category is constructed through social processes rather than through a preexisting definition. Richard Wilson (2011) takes a similar approach in his analysis of the way in which international criminal tribunals generate historical accounts of war crimes and genocide by looking closely at the operation of the International Criminal Court.

Another form of transnational law that legal pluralism has been used to understand is the expanding field of international commercial law, which includes a wide range of mechanisms from formal treaties to informal law based on nonbinding agreements to corporate self-governance. Finally, the creation of new regional legal entities such as the European Union is also exploding the autonomy of local legal orders (Darian-Smith, 1999).

As anthropologists turn their attention to these transnational legal orders, they have expanded and further developed ethnographic methods to analyze intersecting legal orders on multiple scales through 'multisited' ethnography (Marcus, 1995; Merry, 2006b). Multisited ethnography, including sites that are vertically connected local and global spaces, enables anthropologists to explore the many ways in which both separate legal orders and the 'local' and 'global' are not opposing entities, but rather mutually constitutive social fields.

Conclusion

The anthropology of law changed dramatically over the twentieth century. While earlier theories saw law as a structural feature of a whole society and used it to classify societies, more recent anthropological studies examine dispute processing, the linguistic and cultural features of legal settings that shape the way power is exercised, and the extent to which legal categories and concepts are embedded in individual consciousness and local culture. At the same time, the notion of culture as an integrated, homogeneous, and consensual system has been replaced by models of law in society that focus on the multiple and contradictory nature of rules and procedures. Contests among these systems of power and symbolism account for social change. Some of the competing frameworks are the product of alternative systems of legal regulation developed through colonialism or globalization. Contemporary legal anthropology views law as the product of contestation about meanings and identities in social fields constituted by multiple legalities. It examines the way that law exercises power and sees this process as mediated by legal procedures, linguistic forms and exchanges, and cultural practices that shape legal settings and provide individuals with access to these settings.

See also: Colonialism, Anthropology of; Human Rights, Anthropology of; Law and Geography; Law and Society: Development of the Field; Legal Culture and Legal Consciousness; Science and Technology, Anthropology of.

Bibliography

Benda-Beckmann, F.V., Benda-Beckmann, K. von, Griffiths, Anne (Eds.), 2009. Spatializing Law: An Anthropological Geography of Law in Society. Ashgate Publishing Ltd.

Clarke, K.M., 2009. Fictions of Justice: The International Criminal Court and the Challenges of Legal Pluralism in Sub-Saharan Africa. Cambridge University Press.

Clifford, J., 1988. The Predicament of Culture: Twentieth-Century Ethnography, Literature, and Art. Harvard University Press.

Collier, J.F., 1973. Law and Social Change in Zinacantan. Stanford University Press.

Collier, J.F., Maurer, B., Suarez-Navaz, L., 1996. Sanctioned identities: legal constructions of modern personhood. Identities: Global Studies in Culture and Power 2, 1–27.

Comaroff, J.L., Roberts, S., 1981. Rules and Processes. University of Chicago Press.

Conley, J.M., O'Barr, W.M., 1990. Rules versus Relationships: The Ethnography of Legal Discourse. University of Chicago Press.

Conley, J.M., O'Barr, W.M., 2005. Just Words: Law, Language, and Power. University of Chicago Press.

Coombe, R.J., 1998. The Cultural Life of Intellectual Properties: Authorship, Appropriation, and the Law. Duke University Press, Durham.

Coutin, S., Yngvesson, B., 2008. Technologies of knowledge production: law, ethnography, and the limits of explanation. PoLAR: Political and Legal Anthropology Review 31, 1–7.

Coutin, S.B., 2000. Denationalization, inclusion, and exclusion: negotiating the boundaries of belonging. Indiana Journal of Global Legal Studies 7, 585–593.

Coutin, S.B., 2003. Legalizing Moves: Salvadoran Immigrants' Struggle for U.S. Residency. University of Michigan Press.

Darian-Smith, E., 1999. Bridging Divides: The Channel Tunnel and English Legal Identity in the New Europe. University of California Press.

Davis, K.E., Kingsbury, B., Merry, S.E., 2012. Indicators as a technology of global governance. Law and Society Review 46, 71–104.

Engel, D.M., Munger, F.W., 2003. Rights of Inclusion: Law and Identity in the Life Stories of Americans with Disabilities. University of Chicago Press.

Engel, D.M., Engel, Jaruwan S., 2010. Tort, Custom, and Karma: Globalization and Legal Consciousness in Thailand. Stanford Law Books.

Englund, H., 2006. Prisoners of Freedom: Human Rights and the African Poor. University of California Press.

Ewick, P., Silbey, S.S., 1998. The Common Place of Law. University of Chicago Press, Stories From Everyday Life.

French, J.H., 2009. Legalizing Identities: Becoming Black or Indian in Brazil's Northeast. University of North Carolina Press.

Geertz, C., 1983. Local Knowledge: Further Essays in Interpretive Anthropology. Basic Books.

Gluckman, M., 1973. The Judicial Process among the Barotse of Northern Rhodesia (Zambia). Manchester University Press.

Goldstein, D.M., 2012. Outlawed: Between Security and Rights in a Bolivian City. Duke University Press.

Goodale, M., 2009. Surrendering to Utopia: An Anthropology of Human Rights. Stanford University Press.

Greenhouse, C.J., Yngvesson, B., Engel, D.M., 1994. Law and Community in Three American Towns. Cornell University Press.

Hinton, A.L., O'Neill, K.L. (Eds.), 2009. Genocide: Truth, Memory, and Representation. Duke University Press.

Hastrup, K., 2003. Representing the common good: the limits of legal language. In: Anthropological Studies of Rights, Claims and Entitlements. Psychology Press.

Hoebel, E.A. ((1954) 2006). The Law of Primitive Man: A Study in Comparative Legal Dynamics. Harvard University Press.

Humphreys, S., 1985. Law as discourse. History and Anthropology 1, 239–264.

Kelly, T., 2004. Returning home? Law, violence, and displacement among West Bank Palestinians. PoLAR: Political and Legal Anthropology Review 27, 95–112.

Kelly, T., 2006. Law, Violence and Sovereignty among West Bank Palestinians. Cambridge University Press.

Kelly, T., 2011. This Side of Silence: Human Rights, Torture, and the Recognition of Cruelty. University of Pennsylvania Press.
Latour, B., 2010. The Making of Law: An Ethnography of the Conseil d'État. Polity.
Lazarus-Black, M., Hirsch, S.F. (Eds.), 1994. Contested States: Law, Hegemony and Resistance. Routledge.
Llewellyn, K.N., Hoebel, E.A., (1941(2000)). The Cheyenne Way: Conflict and Case Law in Primitive Jurisprudence. Wm S. Hein Publishing.
Maine, H.S., (1861(2007)). Ancient Law: Its Connection to the History of Early Society.
Malinowski, B., 1926. Crime and Custom in Savage Society. Rowman & Littlefield.
Marcus, G.E., 1995. Ethnography in/of the world system: the emergence of multi-sited ethnography. Annual Review of Anthropology 24, 95–117.
Mattei, U., Nader, L., 2008. Plunder: When the Rule of Law Is Illegal. John Wiley & Sons.
Merry, S.E., 1988. Legal pluralism. Law and Society Review 22, 869–896.
Merry, S.E., 1990. Getting Justice and Getting Even: Legal Consciousness among Working-Class Americans. University of Chicago Press, Chicago.
Merry, S.E., 2000. Colonizing Hawai'i: The Cultural Power of Law. Princeton University Press.
Merry, S.E., 2006a. Human Rights and Gender Violence. University of Chicago Press.
Merry, S.E., 2006b. Transnational human rights and local activism: mapping the middle. American Anthropologist 108, 38–51.
Merry, S.E., 2011. Measuring the world: indicators, human rights, and global governance. In: Damani, Partridge, Marina, Welker, Rebecca, Hardin (Eds.), Corporate Lives: New Perspectives on the Social Life of the Corporate Form. Current Anthropology, vol. 52. Wenner-Gren Symposium Series, pp. S83–S95 (Suppl. Issue 3).
Merry, S.E., 2012. What is legal culture? An anthropological perspective. Journal of Comparative Law 5 (2), 40–58.
Mertz, E., 2007. The Language of Law School: Learning to "Think Like a Lawyer". Oxford University Press.
Moore, S.F., 1973. Law and social change: the semi-autonomous social field as an appropriate subject of study. Law and Society Review 7, 719.
Moore, S.F., 1978. Law as Process: An Anthropological Approach. Routledge & Kegan Paul, London.
Moore, S.F., 1986. Social Facts and Fabrications: Customary Law on Kilimanjaro, 1880–1980. Cambridge University Press, Cambridge.
Nader, L., 1991. Harmony Ideology: Justice and Control in a Zapotec Mountain Village. Stanford University Press.
Nader, L., 1997. Law in Culture and Society. University of California Press.
Nader, L., Todd, H.F. (Eds.), 1978. The Disputing Process: Law in Ten Societies. Columbia University Press.
Nielsen, L.B., 2000. Situating legal consciousness: experiences and attitudes of ordinary citizens about law and street harassment. Law and Society Review 34, 1055.
Povinelli, E.A., 2002. The Cunning of Recognition: Indigenous Alterities and the Making of Australian Multiculturalism. Duke University Press.
Radcliffe-Brown, A., 1933. Law, primitive. In: Encyclopedia of the Social Sciences, vol. 9, pp. 202–206.
Richland, J.B., 2008. Arguing with Tradition: The Language of Law in Hopi Tribal Court. University of Chicago Press.
Riles, A., 2001. The Network Inside Out. University of Michigan Press.
Rosen, L., 1989. The Anthropology of Justice: Law as Culture in Islamic Society. Cambridge University Press.
Santos, B., de, S., 1987. Law: a map of misreading - toward a postmodern conception of law. Journal of Law and Society 14, 279.
Starr, J., Collier, J.F., 1989. History and Power in the Study of Law: New Directions in Legal Anthropology. Cornell University Press.
Theidon, K., 2012. Intimate Enemies: Violence and Reconciliation in Peru. University of Pennsylvania Press.
Wilson, R., 2001. The Politics of Truth and Reconciliation in South Africa: Legitimizing the Post-Apartheid State. Cambridge University Press, Cambridge, New York.
Wilson, R.A., 2011. Writing History in International Criminal Trials. Cambridge University Press.
Yngvesson, B., 2006. Backed by papers: undoing persons, histories, and return. American Ethnologist 33, 177–190.
Yngvesson, B., 2010. Belonging in an Adopted World: Race, Identity, and Transnational Adoption. University of Chicago Press.
Yngvesson, B., Coutin, S., 2008. Schrödinger's cat and the ethnography of law. PoLAR: Political and Legal Anthropology Review 31, 61–78.

Law as an Instrument of Social Change

Catherine R Albiston, University of California, Berkeley School of Law, Berkeley, CA, USA
Gwendolyn M Leachman, Williams Institute, UCLA School of Law, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by V. Gessner, volume 12, pp. 8492–8496, © 2001, Elsevier Ltd.

Abstract

To review law as an instrument of social change, one first has to address what one means by 'law' and by 'social change.' We can then examine the mechanisms through which law changes society. This article begins with methodological and definitional issues in studying law as an instrument of social change. It then reviews empirical findings about the mechanisms through which law changes society, focusing in particular on the debate about rights litigation as a social change strategy, the ways organizations respond to legal institutions and implement law, and the relationship between law and social movements.

Introduction

At least since *Brown v. Board of Education* ordered southern states in the United States to desegregate their schools, whether and how law operates as an instrument of social change has attracted intense scrutiny. This inquiry's deceptively simple formulation masks significant complexity in the concepts of law, social change, and the relationship between them. Because different empirical and epistemological traditions approach these concepts quite differently, this variation broadens the understanding of law and social change and fuels a rich and expanding literature in the social sciences. This brief article cannot address all the facets of this rich literature. The long tradition that examines law as a dependent variable – as a social institution that changes as society changes – will not be addressed here (see Friedman and Ladinsky, 1967; Chambliss, 1964). The role of law as an instrument in globalization, in expansion of capitalist markets, and in international development is quite important, but more suited to an in-depth exploration in its own right (*see* Globalization: Legal Aspects; Law and Development). Similarly, state actors can deploy law to repress resistance and thereby to suppress social change; although the authors touch on this point, this article cannot fully explore law's role in suppression and social control (see Barkan, 1984; Fernandez, 2008; *see* Control: Social; Nazi Law).

Even within these constraints, law as an instrument of social change is a complex, fascinating, and controversial subject. William Graham Sumner argued that "legislation cannot make mores," because laws that conflict with established social customs and norms cannot change them and are doomed to fail (Sumner, 1906: p. 77). Although prominent failures spring to mind – prohibition is one (Gusfield, 1986) – counterexamples are also easy to identify. These include the legal and social changes associated with the civil rights, women's rights, and disability rights movements; environmental protection legislation; gay and lesbian rights reforms; and even antismoking legislation (*see* Civil Rights Movements; Law and People with Disabilities; Environmental Movements; Gay and Lesbian Movements). Even though law sometimes faces backlash and resistance in the polity (Krieger, 2000; Klarman, 1994; Rosenberg, 1991), in many instances legal change precedes social change and shifts in public opinion, and these legal changes affect both social relations and social meaning. How have those changes come about? Answering this question requires a clear idea of what is meant by 'law' and 'social change.'

Defining Law

The concept of 'law as an instrument of social change' treats law as an independent variable that influences or causes changes in society. Research in this area conceptualizes law in many different ways. One approach is to focus on formal law, in the form of statutes, landmark judicial opinions, constitutions, and the like, to investigate how changes in formal law may bring about social change (see, e.g., Burstein and Edwards, 1994; Keck, 2009; Rosenberg, 1991). Much of this literature focuses on litigation, particularly impact litigation by social movements, and asks whether litigation is an effective strategy for bringing about social change (for a brief review, see Albiston, 2011). There is no consensus on the answer to this question.

A second approach follows the legal realist tradition (Fisher et al., 1993), expanding on the conception of 'law' to include the law in practice in courts, bureaucratic settings, and other settings (Blumberg, 1967; Feeley, 1992; Lipsky, 1980; Frohman, 1997; Ewick and Silbey, 1998). In this view, law includes not only the formal language of the statute, but also how that language is implemented, interpreted, and understood by institutional actors in the state or by individuals in everyday life. Formal law, in this conception, may have little social relevance because social and institutional processes affect how the law is translated into action. In addition, institutional relationships and incentives may produce unexpected and unintended consequences for legal reforms. In an early exemplar of this approach, Blumberg (1967) set out to examine how the US Supreme Court decisions in *Miranda* and *Gideon*, which provided new procedural protections and legal representation for criminal defendants, affected proceedings in American criminal courts. By studying

these legal changes on the ground, Blumberg discovered that long-standing relationships among prosecutors, judges, and defense counsel limited and distorted the impact of these decisions, pushed defendants into plea bargains, and prevented formal legal changes from doing much to improve criminal proceedings in the courts.

A third approach, often labeled the constitutive view, contends that 'law,' or 'legality,' is constructed through social interaction, wherein people mobilize legal concepts and language outside the courts to negotiate conflict or to interpret their experiences (Ewick and Silbey, 1998, 2003). The institutionalized legal meanings that emerge from these interactions become part of a broader cultural repertoire on which individuals may draw to understand and interpret social life, although legal meanings, once institutionalized, also constrain future interpretations (see Swidler, 1986; Sewell, 1992). In this view, law is defined to include customary practices that convey lawlike meanings, such as placing a chair in a newly shoveled parking place after a Boston snowstorm to mark 'personal property' even when no legally enforceable formal right exists (Ewick and Silbey, 1998). Constitutive scholars define law culturally, referencing the work of early legal anthropologists such as Malinowski, who conceptualized law "as a body of binding obligations, … kept in force by reciprocity and publicity inherent in the structure of society" (Malinowski, 1926: p. 58). This approach resists defining law in terms of Western institutions like courts and legislatures. It requires only that people organize their lives in accordance with implicit rules of behavior that create order within society and are recognized by everyone. Modern scholars, including those in the legal pluralist tradition, contend that law in this broader cultural sense often exists alongside formal legal rules, and may be more important than formal law in the ordering of social interactions (Merry, 1988; Fuller, 1994; Moore, 1973; Macaulay, 1963; Ellickson, 1986).

The way scholars define law can subtly shape their understandings of how law acts to produce social change. Some studies of law and social change take an instrumental approach, conceptualizing law as separate from society. According to this view, law acts upon society by providing remedies for harm, by punishing wrongdoing, by deterring unwanted conduct, by shifting moral values, and by providing a symbol around which to rally. This approach tends to assume that social actors have discrete policy goals that the law helps them attain. In contrast, constitutive approaches see social actors and their intentions as partially defined through legal concepts and constructs. For those who subscribe to this view, law is not separate from society but (at least partially) constitutive of it. Law creates and defines the concepts and categories of everyday life, such as private property, the corporation, the status of marriage, vagrancy, even disability (Sarat and Kearns, 1993). Comparative and historical studies reveal that seemingly natural categories such as these vary across legal regimes and national context (Stone, 1984; Ferree, 2003), and changes in these definitional legal concepts change society by redefining social categories and social meaning (Thompson, 1975). Nevertheless, the constitutive approach can still be seen as conceptualizing law as an agent of social change because law influences culture, even though culture also mediates law's social impact.

These definitional debates teach that institutions, actors on the ground, and widely shared cultural beliefs shape what law will mean in practice. Attending to these mediating factors is crucial to understanding law's effect on social change. The insight that law can be found not only in the courts and legislature, but also in bureaucratic organizations like workplaces and schools, in prosecutorial and police discretionary decisions, and in lawlike social customs and legal understandings that permeate everyday life, opens up many methodologies and locations for studying law as an instrument of social change.

Defining Social Change

Social change, the dependent variable in this inquiry, is similarly tricky to define. A common objective is to measure how society changes as the result of legal change, but there is no consensus about what constitutes social change. Social change resulting from law can include redistributing resources among social groups, reconfiguring social practices and interaction (disputing, collective action, and organizations), or reconstituting social meanings (definitions, constructs, values, and norms). Differences in empirical and epistemological approaches generate differences in conceptualizing social change; for example, studies may rely on quantitative or qualitative measures, and focus on direct or indirect changes resulting from law.

Direct vs Indirect Effects

Some approaches measure social change through the 'direct effects' of law, such as the material legal remedies ordered in judicial opinions. For example, Rosenberg (1991) found that *Brown v. Board of Education*, the landmark US Supreme Court decision ordering school desegregation in the American South, produced very little change in school integration patterns in the years immediately following the decision. Rosenberg's straightforward approach immediately drew criticism that choices about how to measure social change were highly subjective, and the choice of measurement heavily influenced the researcher's conclusions (Feely, 1993). For example, should the degree of social change be measured by comparing the results of litigation with the ideal outcome promised in the formal law (see, e.g., Rosenberg, 1991)? By comparing the results of litigation with the goals of the advocates, who brought the case to the Supreme Court (see, e.g., Feely, 1993)? By comparing the results of litigation with the status quo, had the case never been brought (see, e.g., Keck, 2009)? These are not just technical questions; they are deeply normative questions about what counts as meaningful social change.

Moreover, direct effects do not exhaust the possible social changes resulting from law. Rosenberg also considered some 'secondary' or 'indirect' effects of the Court's decision. Although he found no change in broad public opinion after *Brown* (Rosenberg, 1991), Rosenberg did find that *Brown* generated significant and concerted resistance from Southern supporters of segregation (Rosenberg, 1991: pp. 84–85), suggesting that rather than undermining segregation, *Brown* may have reinforced Southern commitment to it. Once again, critics

argued that evaluating these immediate changes did not exhaust the possible indirect effects of *Brown*. For example, the decision may have provided political cover that enabled legislative and executive action in the decade after *Brown*. It may have mobilized social movement protests against Jim Crow in the 1960s (Polletta, 2000). It may have empowered federal judges in the south, who had lifetime appointments, to rule for the plaintiffs in future cases (Klarman, 1994; McCann, 1993).

In a helpful discussion, Galanter (1983) lays out several ways in which courts and law might change society indirectly, through what he calls "general effects." General effects come in many forms, from using legal penalties to encourage potential law breakers to avoid crime, to changing people's rational, moral, or cognitive evaluations of a given social behavior. For example, criminalizing drunk driving stigmatized the practice and made it much less commonplace. Galanter conceptualizes this process as "enculturation," or the widespread shifting or intensifying of moral evaluations of prohibited conduct such as discrimination or smoking (see also Berkowitz and Walker, 1967). Galanter also discusses how litigation mobilizes collective political action by "provid[ing] symbols for rallying a group, broadcasting awareness of grievance[s] and dramatizing challenge[s] to the status quo" (Galanter, 1983: pp. 125–126; see also McCann, 1994; Barkan, 1980). In this view, social change should be understood to include not only material remedial gains for the litigants of a case, but also an emerging oppositional consciousness among activists and the public (Morris and Braine, 2001; Marshall, 2003).

Positivist vs Constitutive Approaches

Another issue in defining social change is the distinction between positivist versus constitutive conceptions of where to look for, and how to measure, social change. Positivist accounts tend to conceptualize social change in terms of discrete, easily measured dependent variables associated with changes in the law, such quantifiable measures of shifts in public opinion (Rosenberg, 1991), or changes in material circumstances (Klarman, 1994; Burstein and Edwards, 1994). Rosenberg's numerical data about integration of schools and shifts in public opinion are an example of this positivist approach. In contrast, constitutive accounts address how law shapes social relationships and conceptions of identity in everyday life (McCann, 1996; Engel and Munger, 2003). For example, the Americans with Disabilities Act (ADA) may produce more accessible buildings and Braille signs, changes that positivist accounts would readily recognize, but it also affects how one understands social relationships and social identities. The ADA changes employers' expectations and attitudes about the resources and services they should provide to impaired employees, constructs 'disabled person' and 'worker' as compatible identities, and changes impaired individuals' conceptions of themselves away from 'spoiled' identities (Goffman, 1963) toward a normalized and agentic conception of self (Albiston, 2010; Engel and Munger, 2003). Here, social change is located in changing meanings, identities, and expectations about social relationships, and not merely in the measurable number of newly added wheelchair ramps and Braille signs.

The way social change is conceptualized affects conclusions about the law's ability to produce it. Narrow measures of social change may yield pessimistic conclusions, whereas broader definitions of social change may produce more optimistic conclusions. Rosenberg's focus on the actual school integration and public opinion immediately following *Brown* provides a narrow snapshot of *Brown's* potential social impact, and his resulting assessment that *Brown* produced little social impact is therefore unsurprising. Broadening the frame of observation to include alternative measurements of social change – for example, expanding the time frame, seeking out the decision's other indirect consequences, or incorporating more qualitative data – may produce different results. Although Rosenberg's work has been a lightning rod for criticism, it has also sparked a rich literature about how to conceptualize social change. As a result, researchers have begun to explore multiple methods and locations for studying social change: quantitative measures of public opinion, changing material conditions, or organization characteristics; qualitative study of changing social meanings and the processes through which legal changes affect social movements and relationships; and ethnographic studies of changing practices and social meanings in particular social settings. Triangulating these measures, and understanding when and why they differ from one another, define the state of the field on this issue.

Mechanisms through Which Law Brings about Social Change

With definitional questions out of the way, one asks: What are the mechanisms through which law might bring about social change? What are the conditions under which law can change society, or fail to do so? Assessing how courts, and by extension law, can change society requires evaluating the resources and capacities of various audiences as well as the indigenous law-like conventions in social locations subject to law (Galanter, 1983). Law's instrumental effects on society are mediated by the actors who mobilize law, the institutions through which law is implemented, and the already-existing social order law seeks to change. Very few laws are self-enforcing; law must be mobilized to be effective. A victim of discrimination, for example, is unprotected under civil rights laws until he or she initiates legal action demanding that his or her rights be enforced. However, mobilizing the law often requires not only specialized knowledge about the law (or access to someone who has it), but also enough independent wealth to invest the time and money into pursuing a claim – unrealistic expectations for many if not most citizens. Furthermore, practical concerns, such as the need to maintain social relationships, as well as social norms can inhibit people from taking action to enforce legal rights (Bumiller, 1987; Morgan, 1999; Quinn, 2000). Obstacles like these may limit law's effectiveness for social change because courts generally decide only the legal questions that parties bring to them. To illustrate, the authors will discuss three foci of the law and social change literature: the rights debate about the utility of litigation for social change, law and social change in organizations, and how social movements mobilize law to pursue social change.

The Rights Debate

The Rights Debate refers to the debate among scholars about the utility of rights litigation as a strategy for social change. *Brown v. Board of Education* was a catalyst for studying these questions because the strategic litigation campaign that let up to *Brown* is the paradigmatic example of mobilizing law to change society. Several scholars tout how the NAACP's litigation strategy brought about social change even in the absence of significant political clout (Kluger, 1976). In this view, underrepresented communities that lack meaningful access to the polity nevertheless can change social policy through legal strategies (Zemans, 1983). This view is sometimes labeled 'liberal legalism': the idea that constitutional and legal boundaries constrain politics, and therefore legal claims provide a powerful mechanism for producing equality and social change.

Liberal legalism has come under attack from scholars skeptical of the utility of litigation for social change. For example, Scheingold (1974) coined the term 'myth of rights,' the recognition that rights are not self-executing but must be mobilized to be effective. Empirical research soon demonstrated that mobilization was the exception, not the rule (Miller and Sarat, 1981). From the myth of rights perspective, rights litigation undermines collective action by atomizing collective grievances into individual problems and narrowing grievances to only legally actionable claims (McCann, 1986; Scheingold, 1974). Empirical research suggests that litigation fails to produce lasting social change because courts lack institutional authority to implement radical reforms (Rosenberg, 1991), legal victories can be evaded or dismantled without a sustained movement toward reform (Handler, 1978), and powerful opponents retain strategic advantages even within a formally neutral legal system (Galanter, 1974; Albiston, 1999). Rights litigation can even be counterproductive because it legitimates a legal system that masks inequality by seeming to provide a remedy while producing little substantive change (Freeman, 1982, 1998; Tushnet, 1987).

Other scholars responded to this forceful critique of liberal legalism by advancing a more optimistic view based on the symbolic and strategic utility of rights litigation. In this view, rights litigation can change social meaning and give individuals symbolic recognition and personal dignity (Engel and Munger, 2003; Williams, 1991). Even if legal victories are hard to enforce and maintain, rights litigation also mobilizes social movements working outside the legal system to bring about change. For example, McCann, in his study of the US Equal Pay Act movement, finds that pay equity litigation crystallized an issue around which activists mobilized, attracted media attention and participants to the movement, and changed participants' conceptions of what might be possible (McCann, 1994). More instrumentally, the threat of rights litigation can provide leverage in negotiations and publicly embarrass opponents (McCann, 1994: pp. 140–153). Some of these mobilizing effects can occur even when litigation is unsuccessful (McCann, 1994; NeJaime, 2011). Thus, from this perspective, rights litigation affects more than legal outcomes because it can motivate activists to utilize additional avenues of change, such as protest or legislative advocacy (McCann, 1994; NeJaime, 2011).

This debate about rights and social change engendered a much broader understanding of the mechanisms through which law brings about social change, and an expansive view on the locations in which change might occur. Scholars have moved beyond a narrow focus on court-imposed legal remedies to consider movement-building effects, the development of oppositional consciousness, cover for legislative and executive branch actors to make significant changes, media attention and public agenda setting, the formation of collective identities, changes in everyday understandings and social interactions, and even changing perceptions of self-identity. These outcomes are highly interconnected, often triggering and reinforcing one another and driving mechanisms for even broader social reform.

A second important insight is that law may not have the same effects in all social locations. For example, research focusing on litigation in the courts suggests mixed outcomes. Some research suggests civil rights plaintiffs are relatively successful, while other research finds little material change following landmark court decisions (Rosenberg, 1991; Burstein and Monaghan, 1986; Burstein, 1991). Other research focused on everyday locations suggests that even informal rights mobilization can bring about change or even become a catalyst for collective action (Albiston, 2005, 2010). Similarly, some scholars find law positively constructs identity (Engel and Munger, 2003), whereas others find that claiming rights imposes too great a cost by requiring claimants to recognize an injury and identify as a victim (Bumiller, 1987; Quinn, 2000). Although this literature offers no easy conclusions about law as an instrument of social change, it does highlight the rich set of mechanisms and locations through which law may change society.

Organizational and Institutional Resistance

Organizational settings such as schools, prisons, or employers are major targets of legal reforms, but law does not penetrate organizations automatically, without cost or distortion. Neo-institutional organizational theory treats organizations as open systems that respond to their environment, including law, in complex ways that tend to produce homogeneity among organizations in the same organizational field (Meyer and Rowan, 1977). Organizations tend to be isomorphic, or similar, in their structure and practices because they operate under a shared system of rules – regulations, norms, and cognitive schemas – derived from a common social environment (DiMaggio and Powell, 1991). Legal institutions are part of that environment, and isomorphic processes are thus a mechanism through which law changes society, or at least organizations in society. Law facilitates 'coercive isomorphism' by imposing similar, sanction-backed legal requirements on organizations in the same field, such as the tax code rules governing the structure and activities of nonprofits (DiMaggio and Powell, 1983: p. 150). Isomorphic processes can also be 'mimetic,' when organizations in the same field respond to ambiguous legal requirements by adopting the practices of other organizations that they perceive to be successful (Edelman, 1990). A third mechanism, 'normative isomorphism,' occurs when professionals who have similar training, such as lawyers or human resource managers, diffuse

common practices through professional associations and then bring those practices back to the organizations of which they are a part (DiMaggio and Powell, 1991). Often legal values such as due process provide professionals normative guidance in determining which practices are legitimate and effective (Edelman, 1990).

Each of these processes can affect whether law will influence existing organizational processes or become marginalized to prevent interference with other organizational priorities. For example, in response to coercive legal rules, organizations may offer symbolic compliance in the form of policies and structures visible to actors outside the organization but only loosely coupled to actual practice within the organization (Edelman, 1992; Orton and Weick, 1990). Internally, organizations may interpret and implement legal mandates in ways consistent with managerial objectives but inconsistent with legal norms and mandates (Edelman, 1992; Albiston, 2005). When legal requirements are ambiguous, organizations may turn to others in their field to identify compliance strategies that will confer legitimacy, even if those strategies were largely determined by the organizations themselves and not the requirements of law (Edelman, 1990, 1992). These strategies then diffuse throughout organizational fields and become institutionalized as the meaning of compliance (Edelman et al., 2001, 2011). Normative isomorphism operates through professionals who learn common practices through professional training and associations and then bring those practices back to the organizations in which they work. Through this mechanism, professionals may help diffuse widespread but inaccurate myths that drive organizational behavior in a homogenous but irrational direction. For example, Edelman (1992) found that human resource professionals exaggerated the risk of employment discrimination suits compared to their actual incidence in the courts. Over time, organizational definitions of what constitutes compliance often find their way into legal doctrine because judges defer to widespread organizational practices as evidence of compliance without probing whether those practices have any tangible effects on life within the organization (Edelman et al., 2011).

These processes produce organizational structures and practices that respond to law but do little to bring about social change. Once these practices and norms become institutionalized, they become difficult to dislodge through subsequent legal change, creating barriers to social change (Kelly, 2010; Albiston, 2005). For example, Kelly (2010) found that, in the United States, noncompliance with the Family and Medical Leave Act tracked prior institutionalized practices adopted to comply with the Pregnancy Discrimination Act, even though the law had changed to allow fathers as well as mothers to take leave. Organizational norms and practices that grow up around social institutions rather than legal mandates can also be difficult to change. For example, Albiston (2005) found that workers faced gendered resistance to family leave consistent with the institutionalized male breadwinner/female homemaker stereotype embodied in standard work schedules, even though statutory leave rights were gender neutral.

When law does penetrate organizations, typically an authoritative agent or organizational actor is involved. Kalev et al. (2006) found, for example, that workplace diversity programs that designated specific organizational actors to take responsibility for compliance were most effective at increasing the proportions of African-Americans and white women in private sector management. Similarly, Albiston (2005, 2010) found that workers who knew their rights could mobilize law as an authoritative discourse to reinterpret the meaning of taking leave and pressure employers to comply. McCann (1994) found that union advocates mobilized the threat of pay equity lawsuits to gain leverage in negotiations with employers. But organizational actors face competing incentives. Managers rewarded for production and profit rather than legal compliance may view lawsuits as simply a cost of doing business, and therefore see noncompliance with legal mandates as a calculated risk in a complex business environment (Stone, 1975). For this reason, organizations do not seem to operate as unitary, rational actors when they respond to legal sanctions. Rather, organizations and organizational actors manage symbols of compliance and the risk of legal sanction in complex and nonintuitive ways.

Social Movements and Law

Although studies of law and social movements typically treat law as the dependent variable – the target of a movement's attempts at change – law can also affect how social movements define their cause, determine their agenda, and choose their strategies for social change. Perhaps the most fundamental way in which law influences social movements is by shaping the social construction of a movement's grievance, or the collective 'naming' of the social problem motivating the movement as unjust and subject to change (Felstiner et al., 1980). This process of collectively cultivating 'oppositional consciousness' is a crucial step in social movement formation (Piven and Cloward, 1977; Morris and Braine, 2001; see McAdam, 1982). Law factors into this process as a set of norms affecting people's understandings of their social world. The law's broad promise to protect people against 'discrimination' and 'inequality' provides a framework for perceiving and articulating injustice (Albiston, 2005; Marshall, 2003) and it encourages the sense that taking action will make a difference (McCann, 1994: p. 89). Legal constructs may also facilitate the perception of social grievances as collective harms. The legal construction of stigmatized identities such as 'criminal' or 'alien' can generate a common political identity among otherwise unassociated individuals who collectively opposed their legal stigmatization (Eskridge, 2001). Rights are another legal construct that movement actors draw on informally to highlight the movement's collective identity or collective experience of harm (Leachman, 2013). For example, women's movements in the United States have articulated collective rights claims in the past that emphasize women's shared vulnerability or common experience of societal discrimination (Pedriana, 2006; McCammon et al., 2007: p. 729; Ferree, 2003: p. 335). In these ways, legal constructions fuel perceptions of injustice and group identity, crucial elements of movement formation.

The legal mobilization literature also views law as a source of resources for social movement activists to mobilize, and these resources affect the movement's organization, development, and ability to survive (McCann, 1994). In this respect, legal mobilization literature parallels the sociological resource mobilization scholarship, which argues that social movements

arise and succeed when they are able to mobilize resources from sympathetic elites or from indigenous sources within the movement (Jenkins, 1983). Rather than focusing on the material resources that movements marshal, legal mobilization scholarship examines how law generates symbolic resources that movement actors use to motivate and rally constituents (Scheingold, 1974; McCann, 1994; Merry and Stern, 2005; Silverstein, 1996; Polletta, 2000; Coleman et al., 2005) or to generate public support for movement goals (Marshall, 2003; see also Jenness, 1999). Movement rhetoric is often rife with references to law, including not just rights claims, but also claims about citizenship, popular sovereignty, and judicial deference (Leachman, 2013). Such legal framing is often an effective way of packaging a movement's demands because the pervasiveness and cultural legitimacy of law make legal language resonate with a wide cross section of audiences (Snow and Benford, 1992; Pedriana, 2006: pp. 1727–1728; McCammon et al., 2007).

Social movement litigation also helps draw attention to injustice and put issues on the public agenda by attracting media coverage. Litigation strategies tend to attract more media coverage than other political tactics (McCann, 1994: p. 62; Silverstein, 1996: p. 198), directing public attention toward a movement's legal demands (see Levitsky, 2006). Movement actors can draw on those high-profile litigation campaigns to encourage donations in fund-raising efforts (Boutcher, 2005; NeJaime, 2011). Movement litigation also tends to attract the attention of political elites (Persily et al., 2007: p. 9), whom activists cultivate as powerful allies (Minkoff, 1999: pp. 1668–1669; Woods and Barclay, 2008) and use to leverage political concessions to movement demands (Handler, 1978: p. 92; Scheingold, 1974; Kolb, 2005). These studies illustrate the importance of linguistic and media resources both in their own right and as mediating mechanisms that generate additional material resources.

Third, the law shapes social movements' 'tactical repertoires,' the set of practices available to activists to accomplish movement goals (see Ennis, 1987). Law as a set of formal rules and procedures affects the availability of litigation as part of movements' tactical tool kit. Legal changes in the rules of standing opened up decades of environmental protection litigation that otherwise would not have been possible (Orren, 1976). Similarly, civil rights statutes provide a legal means to redress grievances about discrimination in employment and suppression of the franchise among minority communities (Albiston and Nielsen, 2007). Formal law may catalyze extralegal mobilization when a landmark judicial opinion generates a sense of hope or outrage (Andersen, 2005: p. 44; see Rosenberg, 1991: pp. 155–156). Litigation strategies may encourage a movement to undertake other forms of political action and grassroots organizing (McCann, 1994), for example, by providing a concrete focus or goal to demand (Boutcher, 2005; Coleman et al., 2005; McCann, 1994; Polletta, 2000).

Social movements that employ legal strategies may pay a price, however. Litigation is expensive, and may drain social movement resources away from potentially more effective strategies, such as protest or grassroots organizing (Rosenberg, 1991; Scheingold, 1974; see Hunt, 1990; but see Polletta, 2000). Although statutory individual rights provide a mechanism for addressing injustices such as discrimination, they tend to channel grievances into individualized litigation and discourage collective action (McCann, 1986; Scheingold, 1974). Litigation success, in part because of its visibility, may also mobilize a countermovement or cause the opposition to systematically resist legal reforms (Rosenberg, 1991; Klarman, 2005; Luker, 1985).

A price also may be exacted at the level of social meaning, identity, and the very definition of the cause. Using litigation as a social movement tactic can deradicalize the message and objectives of a movement, even if the litigation is successful. Existing legal rights and rules form discursive opportunity structures through which social movements may advance their causes. To access these opportunity structures, movement actors must shape their claims and demands to fit available legal templates. Because law is a relatively conservative institution, fitting claims to law's discursive opportunity structures risks taming the movement's demands and deradicalizing its agenda (Ferree, 2003). Furthermore, the law's ability to attract resources may support lawyers as movement leaders and make litigation strategies dominant, moving more confrontational strategies such as protest to the margin (see Haines, 1984). In turn, lawyers as elite actors can reshape movement messages and goals even when the broader movement's participants have different goals (Bell, 1976; Levitsky, 2006).

When identities are legally protected, litigation tactics tend to shape and define the identities around which movements mobilize, even if those legally expedient identities are not the ones activists or individuals would choose outside of law. Bumiller (1987) found, for example, that individuals who believed they had been discriminated against were reluctant to mobilize legal claims because they resisted taking on a victim identity required by discrimination law. Some gay and lesbian activists note how the pursuit of same-sex marriage has required plaintiffs to conform to a near-heteronormative family status to obtain legal rights, even though some movement activists would prefer to broaden the definition of who counts as family (Polikoff, 2008). Incremental social change, in this view, comes at the expense of more radical and far-reaching changes in how one understands identity and social relationships. Thus, when evaluating law as an instrument of social change, it is critical to consider which demands may have been marginalized or abandoned because a social movement relied on legal tactics rather than other mechanisms of social change (Albiston, 2011).

Conclusion

Sociolegal literature provides no definitive conclusions about the law's role as an instrument of social change, but it has provided researchers with guidelines for constructing a more nuanced understanding of the relationship between law and social change. The literature has moved beyond questions of whether law matters for social change toward investigating the ways in which, and the mechanisms through which, law influences social change. The research also has moved toward identifying multiple forms of social change, including material and symbolic changes, as well as direct and indirect changes, and toward theorizing the relationship among these forms.

Scholarship addresses the law's ability to generate social change in disparate social contexts and the various cultural and institutional factors that shape what law will mean in practice. These questions involve looking at the relationships among legal meanings, the social practices structured around those meanings, and the institutionalization of legal understandings that can promote radical change or retrenchment. The productive empirical work investigating these processes provides important theoretical frameworks for understanding how systematic factors may enhance or constrain the potential for social change through law.

See also: Civil Liberties and Human Rights; Civil Rights Movements; Civil Rights; Gay and Lesbian Movements; Human Rights, History of; Human Rights: Legal Aspects; Law, Mobilization of; Media and Social Movements; Movements, Social; Rights: Legal Aspects; Social Changes: Models; Social Movements and Political Violence; Social Movements, Geography of; Social Movements, History of: General; Social Movements: A Social Psychological Perspective.

Bibliography

Albiston, C., 1999. The rule of law and the litigation process: the paradox of losing by winning. Law & Society Review 33 (4), 869–910.

Albiston, C., 2005. Bargaining in the shadow of social institutions: competing discourses and social change in workplace mobilization of civil rights. Law & Society Review 39 (1), 11–50.

Albiston, C., 2010. Institutional Inequality and the Mobilization of the Family and Medical Leave Act: Rights on Leave. Cambridge University Press, Cambridge.

Albiston, C.R., 2011. The dark side of litigation as a social movement strategy. Iowa Law Review Bulletin 96, 61.

Albiston, C.R., Nielsen, L.B., 2007. The procedural attack on civil rights: the empirical reality of Buckhannon for the private attorney general. UCLA Law Review 54, 1087–1184.

Andersen, E.A., 2005. Out of the Closets and into the Courts. University of Michigan Press, Ann Arbor.

Barkan, S.E., 1980. Political trials and resource mobilization: towards an understanding of social movement litigation. Social Forces 58 (3), 944–961.

Barkan, S., 1984. Legal control of the southern civil rights movement. American Sociological Review 49 (4), 552–565.

Bell, D.A., 1976. Serving two masters: integration ideals and client interest in school desegregation litigation. Yale Law Journal 85, 470.

Berkowitz, L., Walker, N., 1967. Law and moral judgments. Sociometry 30 (4), 410–422.

Best, R., Edelman, L.B., Krieger, L., Eliason, S., 2011. Multiple disadvantages: an empirical test of intersectionality theory in EEO litigation. Law & Society Review 45, 991–1025.

Blumberg, A., 1967. The practice of law as a confidence game. Law & Society Review 1 (2), 15–39.

Boutcher, S.A., 2005. Making lemonade: turning adverse decisions into opportunities for mobilization. Amici 13, 8–13.

Bumiller, K., 1987. Victims in the shadow of the law: a critique of the model of legal protection. Signs 12 (3), 421–439.

Burstein, P., 1991. Legal mobilization as a social movement tactic: the struggle for equal employment opportunity. American Journal of Sociology 96 (5), 1201–1225.

Burstein, P., Edwards, M.E., 1994. The impact of employment discrimination litigation on racial disparity in earnings: evidence and unresolved issues. Law & Society Review 28 (1), 79–111.

Burstein, P., Monaghan, K., 1986. Equal employment opportunity and the mobilization of law. Law & Society Review 20 (3), 355–388.

Chambliss, W.J., 1964. A sociological analysis of the law of vagrancy. Social Problems 12, 67–77.

Coleman, C., Nee, L.D., Rubinowitz, L.S., 2005. Social movements and social-change litigation: synergy in the Montgomery bus protest. Law and Social Inquiry 30, 663.

DiMaggio, P.J., Powell, W.W., 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. American Sociological Review 48, 147–160.

DiMaggio, P.J., Powell, W.W., 1991. Introduction. In: Powell, W.W., DiMaggio, P.J. (Eds.), The New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago, pp. 1–38.

Edelman, L.B., 1990. Legal environments and organizational governance: the expansion of due process in the American workplace. American Journal of Sociology 95 (6), 1401–1440.

Edelman, L.B., 1992. Legal ambiguity and symbolic structures: organizational mediation of civil rights law. American Journal of Sociology 97, 1531.

Edelman, L.B., Fuller, S.R., et al., 2001. Diversity rhetoric and the managerialization of law. American Journal of Sociology 106 (6), 1589–1641.

Edelman, L.B., Krieger, L.H., et al., 2011. When organizations rule: judicial deference to institutionalized employment Structures. American Journal of Sociology 117 (3), 888–954.

Ellickson, R.C., 1986. Of coase and cattle: dispute resolution among neighbors in Shasta county. Stanford Law Review 38, 623.

Engel, D., Munger, F., 2003. Rights of Inclusion: Law and Identity in the Life Story of Americans with Disabilities. University of Chicago Press, Chicago.

Ennis, J., 1987. Fields of action: structure in movements' tactical repertoires. Sociological Forum 2, 520–533.

Eskridge Jr., W.N., 2001. Channeling: identity-based social movements and public law. University of Pennsylvania Law Review 150, 419–525.

Ewick, P., Silbey, S.S., 1998. The Common Place of Law: Stories from Everyday Life. University of Chicago Press, Chicago.

Ewick, P., Silbey, S., 2003. Narrating social structure: stories of resistance to legal authority. American Journal of Sociology 108 (6), 1328–1372.

Feeley, M., 1992. The Process is the Punishment. Russell Sage, New York.

Feely, M., 1993. Hollow hopes, flypaper and metaphors. Law and Social Inquiry 17, 745–760.

Felstiner, W.L.F., Abel, R.L., Sarat, A., 1980. The emergence and transformation of disputes: naming, blaming, claiming. Law & Society Review 15, 631–654.

Fernandez, L.A., 2008. Policing Dissent: Social Control and the Anti-globalization Movement. Rutgers University Press, New Brunswick, NJ.

Ferree, M.M., 2003. Resonance and radicalism: feminist framing in the abortion debates of the United States and Germany. American Journal of Sociology 109, 304.

Fisher III, William W., Horowitz, M.J., et al. (Eds.), 1993. American Legal Realism. Oxford University Press, Oxford.

Freeman, A., 1998. Antidiscrimination law from 1954 to 1989: uncertainty, contradiction, rationalization, denial. In: Kairys, D. (Ed.), The Politics of Law: A Progressive Critique. Basic Books, New York.

Freeman, A.D., 1982. Antidiscrimination law: a critical review. In: Kairys, D. (Ed.), The Politics of Law. Pantheon Books, New York.

Friedman, L.M., Ladinsky, J., 1967. Social change and the law of industrial accidents. Columbia Law Review 67 (1), 50–82.

Frohman, L., 1997. Convictability and discordant locales: reproducing race, class, and gender ideologies in prosecutorial decision making. Law & Society Review 31, 531.

Fuller, C., 1994. Legal anthropology, legal pluralism, and legal thought. Anthropology Today 10, 9.

Galanter, M., 1974. Why the haves come out ahead: speculations on the limits of legal change. Law & Society Review 9, 95–160.

Galanter, M., 1983. The Radiating Effects of Courts. Longman, New York.

Goffman, E., 1963. Stigma: Notes on the Management of Spoiled Identity. Prentice-Hall, Englewood Cliffs, NJ.

Gusfield, J.R., 1986. Symbolic Crusade: Status Politics and the American Temperance Movement. University of Illinois Press, Champaign, IL.

Haines, H.H., 1984. Black radicalization and the funding of civil rights: 1957–1970. Social Problems 32, 31–43.

Handler, J.F., 1978. Social Movements and the Legal System: A Theory of Law Reform and Social Change. Academic Press, New York.

Hunt, A., 1990. Rights and social movements: counter-hegemonic strategies. Journal of Law and Society 17, 309–328.

Jenkins, J.C., 1983. Resource mobilization theory and the study of social movements. Annual Review of Sociology 9, 527–553.

Jenness, V., 1999. Managing differences and making legislation: social movements and the racialization, sexualization, and gendering of federal hate crime law in the U.S., 1985–1998. Social Problems 46, 548–571.

Kalev, A., Dobbin, F., Kelly, E.L., 2006. Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. American Sociological Review 71, 589–617.

Keck, T.M., 2009. Beyond backlash: assessing the impact of judicial decisions on LGBT rights. Law & Society Review 43 (1), 151–185.

Kelly, E.L., 2010. Failure to update: an institutional perspective on noncompliance with the family & medical leave act. Law & Society Review 44 (1), 33–66.
Klarman, M., 2005. Brown and Lawrence (and Goodridge). Michigan Law Review 104, 431.
Klarman, M.J., 1994. How Brown changed race relations: the backlash thesis. Journal of American History 81 (1), 81–118.
Kluger, R., 1976. Simple Justice. Knopf, New York.
Kolb, F., 2005. The impact of transnational protest on social movement organizations: mass media and the making of ATTAC Germany. In: Tarrow, D.D.P.A.S. (Ed.), Transnational Protest and Global Activism, People, Passions, and Power. Rowman & Littlefield, Lanham, pp. 95–120.
Krieger, L., 2000. Afterword: socio-legal backlash. Berkeley Journal of Employment and Labor Law 21 (1), 475–519.
Leachman, G., 2013. A social constructionist approach to legal framing. Studies in Law, Politics, and Society 61, 25–59.
Levitsky, S., 2006. To lead with law: reassessing the influence of legal advocacy organizations in social movements. In: Sarat, A., Scheingold, S. (Eds.), Cause Lawyers and Social Movements. Stanford University Press, Stanford, CA, pp. 145–163.
Lipsky, M., 1980. Street-Level Bureaucracy. Russell Sage, New York.
Luker, K., 1985. Abortion and the Politics of Motherhood. University of California Press.
Macaulay, S., 1963. Non-contractual relations in business: a preliminary study. American Sociological Review 28, 55–68.
Malinowski, Bronislaw, 1926. Crime and Custom in Savage Society. Harcourt & Brace, New York.
Marshall, A.M., 2003. Injustice frames, legality, and the everyday construction of sexual harassment. Law and Social Inquiry 28 (3), 659–689.
McAdam, D., 1982. Political Process and the Development of Black Insurgency, 1930–1970. University of Chicago Press, Chicago.
McCammon, H.J., Muse, C.S., Newman, H.D., Terrell, T.M., 2007. Movement framing and discursive opportunity structures: the political successes of the U.S. women's jury movements. American Sociological Review 72, 725–749.
McCann, M., 1986. Taking Reform Seriously. Cornell University Press, Ithaca, NY.
McCann, M., 1993. Reform litigation on trial. Law and Social Inquiry 17, 715–743.
McCann, M., 1994. Rights at Work: Pay Equity Reform and the Politics of Legal Mobilization. Univeristy of Chicago Press, Chicago.
McCann, M., 1996. Causal versus constitutive explanations (or, on the difficulty of being so positive…). Law and Social Inquiry 21, 457–482.
Merry, S.E., 1988. Legal pluralism. Law & Society Review 22, 897.
Merry, S.E., Stern, R.E., 2005. The female inheritance movement in Hong Kong: theorizing the local/global interface. Current Anthropology 46, 387–409.
Meyer, J.W., Rowan, B., 1977. Institutionalized organizations: formal structure as myth and ceremony. American Journal of Sociology 83, 340–363.
Miller, R.E., Sarat, A., 1981. Grievances, claims and disputes: assessing the adversary culture. Law & Society Review 15, 525–566.
Minkoff, D.C., 1999. Bending with the wind: strategic change and adaptation by women's and racial minority organizations. The American Journal of Sociology 104, 1666–1703.
Moore, S.F., 1973. Law and social change: the semi-autonomous social field as an appropriate subject of study. Law & Society Review 7, 719.
Morgan, P.A., 1999. Risking relationships: understanding the litigation choices of sexually harassed women. Law and Society Review 33 (1), 67–91.
Morris, A., Braine, N., 2001. Social movements and oppositional consciousness. In: Mansbridge, J., Morris, A. (Eds.), Oppositional Consciousness: The Subjective Roots of Social Protest. University of Chicago Press, Chicago, pp. 20–37.
NeJaime, D., 2011. Winning through losing. Iowa Law Review 96, 941–1012.
Orren, K., 1976. Standing to Sue: Interest group conflict in the federal courts. American Political Science Review 70, 723–741.
Orton, J.D., Weick, K., 1990. Loosely coupled systems: a reconceptualization. The Academy of Management Review 15 (2), 203–223.
Pedriana, N., 2006. From protective to equal treatment: legal framing processes and transformation of the women's movement in the 1960s. The American Journal of Sociology 111, 1718.
Persily, N., Egan, P., Wallsten, K., 2007. Gay marriage, public opinion, and the courts. In: Public Law and Legal Theory Research Paper Series.
Piven, F.F., Cloward, R.A., 1977. Poor People's Movements: Why They Succeed, How They Fail. Pantheon Books, New York.
Polikoff, N.D., 2008. Beyond (Straight and Gay) Marriage: Valuing All Families under the Law. Beacon Press, Boston.
Polletta, F., 2000. The structural context of novel rights claims: southern civil rights organizing, 1961–1966. Law & Society Review 34 (2), 367–406.
Quinn, B.A., 2000. The paradox of complaining: law, humor, and harassment in the everday work world. Law and Social Inquiry 25 (4), 1151–1185.
Rosenberg, G.N., 1991. The Hollow Hope: Can Courts Bring about Social Change? University of Chicago Press, Chicago.
Sarat, A., Kearns, T.S., 1993. Beyond the great divide: forms of legal scholarship and everyday life. In: Sarat, A., Kearns, T.R. (Eds.), Law in Everyday Life. University of Michigan Press, Ann Arbor.
Scheingold, S.A., 1974. The Politics of Rights. Yale University Press, New Haven.
Sewell, W.H., 1992. A theory of structure: duality, agency, and transformation. American Journal of Sociology 98 (1), 1.
Silverstein, H., 1996. Unleashing Rights: Law, Meaning, and the Animal Rights Movement. University of Michigan Press, Ann Arbor.
Snow, D.A., Benford, R.D., 1992. Master frames and cycles of protest. In: Morris, A.D., Mueller, C.M. (Eds.), Frontiers in Social Movement Theory. Yale University Press, New Haven, pp. 133–155.
Stone, C.D., 1975. Where the Law Ends: The Social Control of Corporate Behavior. Harper & Row.
Stone, D.A., 1984. The Disabled State. Temple University Press, Philadelphia.
Sumner, W.G., 1906. Folkways. Dover, New York.
Swidler, A., 1986. Culture in action: symbols and strategies. American Sociological Review 51 (April), 273–286.
Thompson, E.P., 1975. Whigs and Hunters: The Origins of the Black Act. Pantheon Books, New York.
Tushnet, M., 1987. The NAACP's Legal Strategy against Segregated Education, 1925–1950. University of North Carolina Press, Chapel Hill.
Williams, P., 1991. The Alchemy of Race and Rights. Harvard University Press, Cambridge.
Woods, P.J., Barclay, S.W., 2008. Cause lawyers as legal innovators with and against the state: symbiosis or opposition? Studies in Law, Politics and Society 45, 203–231.
Zemans, F.K., 1983. Legal mobilization: the neglected role of the law in the political system. American Political Science Review 77, 690–703.

Law as Constitutive

Roger Cotterrell, University of London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Law as constitutive' refers to the idea that law helps to constitute the social world by shaping culture and social relations, and by helping to determine popular understandings and common sense. The idea contrasts with early approaches in sociology of law that typically saw law as itself largely controlled by wider social forces. To emphasize law's constitutive power is to recognize its social significance, but also to risk exaggerating this. Productive approaches will be ones that view law pluralistically as a facet of social experience, shaping and being shaped by other social phenomena, and integrated with them in complex and variable ways.

The idea of law as 'constitutive' is the idea that law not only regulates but also can create or determine social phenomena. According to this view, law shapes fundamental, moral, or political assumptions and general conceptions of the nature of social relations and social institutions. Jurists and some social theorists have long argued that law has a major role in defining basic social understandings of property, responsibility, and personality. The views of law's constitutive capabilities that have been developed in Marxist theory and in critical legal studies mainly stress law's conservative social power. They contrast with other sociological views of law as a relatively passive reflection of social forces, or as an instrument of government consciously aimed at producing social change. The major problem of the focus on law as constitutive is its tendency to exaggerate or assume law's independent social power, especially when law is treated exclusively as juridical regulation by the state. Wider 'pluralist' views of legal experience, however, offer possibilities for sophisticated analyses of law as both constituting and being constituted by other social phenomena.

The idea that law not merely regulates but also creates or forms social phenomena has been a powerful component of recent sociologically oriented approaches to the study of law. Yet it is an old idea that has resurfaced in social science and legal studies after having been marginalized through much of the modern development of sociology of law in the second half of the twentieth century (see Sociology of Law). In its recent forms, the idea of law as constitutive focuses particularly on law's power to shape fundamental, moral, or political assumptions, elements of culture, and general conceptions of the social. In this sense, law is seen as a mechanism shaping consciousness, establishing frameworks of thought and evaluation in terms of which social life is structured.

Classic Conceptions

Classic legal sociology, represented especially by the work of the Austrian jurist Eugen Ehrlich at the beginning of the twentieth century, confronted the constitutive role of law as an urgent political question: to what extent can the law of the nation-state define the nature of social relations in the face of powerful competing moral forces (Ehrlich, 2002). In a multicultural society (such as the disparate and fragmented Austro-Hungarian empire of Ehrlich's time) in which diverse religious and ethnic influences on social life remained strong and sometimes antagonistic to the values enshrined in state law, Ehrlich concluded that state law's constitutive power is limited and it must learn from, as much as seek to control, informal norms of social life (Hertogh, 2009). Many other theories in the formative era of modern legal sociology carried similar messages: of law either rooted in existing culture and established collective understandings or enfeebled by its separation from them. The idea is expressed in W.G. Sumner's (2002) view in 1906 that effective legislation must be consistent with social mores, in Roscoe Pound's (1917) insistence on the 'limits of effective legal action,' and in persistent echoes of Friedrich Carl von Savigny's (2002) claim in 1814 that law must be rooted in a "common consciousness" (*Volksgeist*) of the people, which it expresses but cannot change.

According to all such views, because law is strong when it represents popular understandings and weak when it runs counter to them, its power to constitute social relations is circumscribed. Law has a degree of independent social power only insofar as it avoids intruding on matters of belief or deep-rooted traditional practice; technically oriented law can be effective, while expressive (ideologically charged or value-infused) law is relatively weak unless the values and understandings it presupposes are already in place in social life.

Thus, an important agenda for sociology of law in the second half of the twentieth century became the exploration of conditions under which law's regulatory techniques could be used to bypass the assumed inertia of culture or to shape specific kinds of social relations in ways that might indirectly encourage more general social change. In an era (especially from the 1960s) in which urgent attempts were made to harness law to tasks of social engineering (e.g., in preventing racial and sexual discrimination), the old idea that law could only mirror established systems of values and beliefs symbolized a serious problem to be addressed. Legal sociology's practical task became to confront cultural obstacles that stood in the way of the successful use of law to induce social change. Law as an instrument of social control or social transformation might be able to operate instrumentally if appropriate regulatory techniques could be devised and conditions were favorable, but law remained limited by its cultural dependence, its general lack of power to constitute popular understandings, values, and beliefs.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86056-0

Classical social theory supported this view of law as culturally dependent only obliquely and partially. Émile Durkheim's sociology treats law as merely one dimension of a larger moral universe (see Durkheim, Emile (1858–1917)). Law neither constitutes nor is constituted by moral understandings and moral bonds but is an aspect of them; it is these bonds and understandings that constitute society (Cotterrell, 1999). Max Weber, associating modern law with a particular kind of rationality, sees law as running parallel in its development with other kinds of rational reordering of Western experience (Kronman, 1983), although his sociology remains ambiguous as to the nature of law's specific contribution to a more general modern 'rationalization' of life (see Weber, Max (1864–1920)). Karl Marx's conception of law as part of the superstructure of a mode of production (see Marx, Karl (1818–83)) is also ambiguous as to law's 'independent' constitutive role, but suggests that this role will, at best, be limited (Collins, 1982). Thus, while law may provide indispensable support (ideological and repressive) for a mode of production, it cannot, according to standard Marxist analyses, constitute such a mode or change it in any historically fundamental way. Nevertheless, law constitutes, in the sense of formalizing and defining, relations and understandings of contract and property. To that extent, it contributes powerfully to ideology in capitalist societies.

Modern Critical Perspectives

The Marxist-derived idea of law framing consciousness (by shaping ideas of responsibility, ownership, freedom, consent, agreement, causality, authority, etc.) has strongly influenced recent ideas about the constitutive power of law, especially as a result of its advocacy as a central tenet of the modern critical legal studies (CLS) movement. CLS, which developed in many Western countries and especially in the United States from the 1970s, had a major influence on destroying old ideas of law as a passive reflection of other social forces (see Critical Legal Studies). Its effect on sociology of law was to help to turn attention from questions about how law might be made more effective as an instrument of social change and toward studies of law's power to reinforce established social practices and institutions, primarily by influencing collective understandings. Thus, because of the continual invocation of basic legal concepts, especially of property and contract, in everyday regulatory practice, law, in the CLS view, contributes importantly to the 'reification' of concepts, that is, the turning of debatable ideas into settled, widespread, usually unchallenged understandings.

Property, for example, ceases to be a variable relationship and tends to become a virtually sacred private object (described in terms of its 'sanctity'). Early in the nineteenth century, Jeremy Bentham (2001), the founder of modern Anglo-American legal philosophy (see Bentham, Jeremy (1748–1832)) had insisted that 'property' is one of many 'fictitious entities' created by law (others include 'right,' 'duty,' 'power,' 'immunity,' 'privilege,' and 'liberty'), but in modern liberal societies property is commonly understood as antecedent to law and law is viewed as the protector of property, not its creator. CLS, indirectly following in the footsteps of Bentham and other pioneers such as the French jurist Emmanuel Lévy (Cotterrell, 2008), treated legal concepts and reasoning as having the power to create and sustain fundamental ways of thinking about social relations.

Marxist legal theory in its most sophisticated forms (e.g., Pashukanis, 1978) and modern critical legal theory emphasize that law shapes ideas of personality and subjectivity, which influence much wider currents of social thought. Thus, law defines who can be responsible for actions and who can engage in transactions that will be legally recognized. In particular societies at particular times, children, married women, mentally handicapped people, prisoners, and aliens of the state (foreigners) have been treated as wholly or partly lacking legal capacity (e.g., to own property directly). Thus, for certain purposes, law constitutes human capacities or denies them. Correspondingly, law routinely gives personality and capacity to nonhuman entities. In various societies, these have included religious idols and animals, but in modern law by far the most important example of the construction of nonhuman personality is the attribution of legal personality to the business corporation. This attribution constitutes the corporation as an entity entirely separate from the human individuals who work for it. In recent critical legal scholarship, the concern with law's constitution of personality has broadened to focus on the wider range of effects of law on the construction of identity.

Feminist theory has emphasized that law has typically constituted the human person in terms of male characteristics. In the past, this legal stance was often explicit, with the denial in legal doctrine of women's legal personality in certain contexts (e.g., as property owners or voters). Modern feminist theory has shown how legal thought and practice ignore differences between women's and men's experience, or covertly assimilate the former to the latter, applying a 'male standard' to the definition of all social relations, capacities, and responsibilities. It has been claimed that law constructs women (in its typical rhetoric and evaluations) as 'terrorized, sexualized and maternalised' (Frug, 1992). In a host of studies, critical sociolegal scholarship has also focused on the legal construction or denial of the social identities of minority groups (e.g., López, 1996) and, more generally, on law's power to define and exclude 'the other' in sharp opposition to an image of the 'normal' subject of law.

The effect of this accumulation of critical work is a widespread rejection among sociolegal scholars of the doubts about law's social importance that long haunted legal sociology. In their place is a confidence that legal ideas and practices are central matters for sociological study. The older legal sociology sometimes slipped into the self-defeating position of holding that law 'mirrored' society (thus implying not only law's lack of social scientific importance but also its 'nonsocial' identity as 'merely' technical regulation and so its substantial immunity from conceptualization as a social phenomenon in itself). The newer emphasis on law as constitutive asserts law's power to shape fundamental collective understandings and practices that define the nature of social life. Taken to extremes, this position can suggest that law constitutes the very idea of society, as opposed to society constituting law (Fitzpatrick, 1995). A further step might be to discard all reified conceptions of 'society' or even of 'the social' as a category (on the basis that the social picture is now too fluid and complex to allow such

encapsulations or to allow these concepts to retain analytical power). The aim then would be to focus specifically on the particular kinds of social conceptions (e.g., of identity, community, and culture) that legal doctrine itself proposes, assumes, or sustains. In a sense, such an approach entirely reverses the conception of law as mirror of society: the social (whether or not it remains possible to conceptualize this as any kind of totality) is seen as the construction of law.

As scholars and policy makers in Western societies are once again seeking ways to effectively regulate their economic and financial systems, partly as a consequence of acute economic crises, the question of how far law can shape the culture of business corporations and economic and financial networks has entered many recent agendas of research. Sociolegal scholars have suggested that law can operate not only through direct regulatory controls on institutions, procedures, and transactions but also by reinforcing templates of ethical behavior so as to help alter and shape attitudes within corporations and to influence business cultures. Some writers propose that such 'endogenous' effects of law can coexist with processes by which law also learns from business cultures (Edelman and Stryker, 2005). So if law can, to some extent, constitute culture indirectly, it may also be constituted partly by the culture it addresses. It might be said that it learns as it teaches.

Limitations of Constitutive Approaches to Law

A danger of constitutive approaches is that of greatly exaggerating law's power (how can we know that *law* does the constituting and, even if it does, under what conditions and constraints?). Jurists have long been tempted in this way, some even speculating that law largely determines morality by contributing to socialization (Olivecrona, 1939). Philosophers and jurists note law's 'word-magic,' the power of legal words as 'performatives' to bring about social effects, such as inaugurating marriages, creating personal statuses, or transferring property (Austin, 1982). But these formal effects raise rather than answer questions about law's social character. The risk is of attempting to analyze law in isolation from other social forces, treating it (as the earlier 'mirror' conceptions tended to do) as separate from other aspects of social life, with a causal relationship existing between law and society. This danger has been especially serious for CLS, which, being essentially a law school movement rather than a movement in social science, has tended to assume law's social significance and power (in conformity with the main trend of lawyers' thinking throughout modern times) rather than to seek to assess it systematically and empirically in a wider social context.

Equally, it is not always recognized sufficiently that law's constitutive power may vary greatly in different societies. In some societies, its power to shape ways in which politics is conducted (Balkin, 2011), or in which social movements are formed and their campaigns or collective strategies are pursued (Brigham, 1996), may be very clear, if complex in its effects. In other societies, different cultural values and traditions make law much less capable of constituting the form and character of social relations. Indeed, they might make it largely irrelevant.

It is also significant that one influential and sophisticated theoretical approach in contemporary legal sociology entirely rejects the idea of law as constitutive of society because it denies that law as a system can communicate with, and so influence, wider social life in any direct way (Luhmann, 2004). Autopoiesis theory sees modern law as a discourse that normatively constitutes (more specifically, produces, reproduces, and observes) only itself. One consequence is that its conceptions of agency, responsibility, personality, subjectivity, and so on are produced solely for its purposes and processes and have no necessary relation to anything existing outside these. Far from constituting economic, scientific, political, or cultural life, law produces decisions to which these other social systems may respond on their own terms. Autopoiesis theory denies all conceptions of law as a mirror of society. At the same time, its assertion of law's radical discursive autonomy deliberately provokes endless questions as to how law can engineer social effects that are other than intolerably unpredictable.

The Plurality of Legal Experience

Some recent sociolegal approaches have discarded both the idea of law as a mirror of society and that of law as a more or less independently constitutive force in social life, and replaced these with a view of law as social mediation (e.g., Mertz, 1994). Law thus appears as one way (or perhaps a cluster of ways, if it is not seen as a unitary discourse or practice) in which social experience is mediated, conceptualized, defined, and evaluated.

Such an approach presents its own difficulties as a sociological conception of law. If law mediates the social in some way, what enables it to do so, and from where does its power or authority to mediate derive? What is the nature of law if it is, in some way, separate from the social life that it mediates? The questions are ultimately no different from those that haunt the idea of law as constitutive (e.g., where does law derive its power or authority to constitute social life? And how, in what way, is law separated from social life in order to constitute it?).

In answering these questions, the temptation is often to dissolve one term into the other. Thus, law may be seen as, for all practical purposes, constituting the social so that the latter ceases to exist (at least for purposes of sociolegal analysis) as something independent of legal categorizations and evaluations. On this view, law may certainly be studied in social context, but its significance as an object of study requires no validation by social science. Alternatively, law can be seen merely as a mirror of social forces and social structure, meriting no special, independent sociological analysis. On this view, juristic concerns are socially insignificant unless reducible to general sociological categories and concepts.

A productive approach may be available to avoid both of these positions. Law is sometimes conceptualized sociologically as an aspect or field of social experience. In this sense, it is a component of social consciousness, a normative patterning of experience, or a framework of understandings or practices. As an aspect of social life, it is to be separated from other social phenomena only provisionally and contingently for purposes of study (Cotterrell, 2006). Important antecedents of this approach exist in legal sociology (Gurvitch, 1947), but they

have been obscured, to some extent, by approaches that treat law entirely as an instrument of the state. Viewed in this state-focused way, law's constitutive power appears problematic or mysterious given (1) the state's frequently noted remoteness from the local milieux it regulates, and (2) the increasing production of law by agencies other than the state.

Law may, however, be understood pluralistically as existing not only in the practices and procedures of state agencies, but also in the everyday circumstances, sites, and settings of a wide variety of social groups and networks (see Legal Pluralism). Legal experience, according to this view, is the subjective experience of individuals and groups confronting, in these diverse environments, 'local' issues of governmental regulation and social control. An approach along these lines will see law as constituting social life insofar as it is itself embedded in social life, in expectations about security, responsibility, and other matters. At the same time, state law (the legal practice of lawyers and courts serving the centralized state) needs to be not merely politically constituted, according to criteria of political legitimacy, but also (from a pluralist perspective) morally constituted by its relation to the social experience of the populations that it purports to regulate (Cotterrell, 2006: 34–44). From this perspective, the issue of law's moral authority, identified by Ehrlich a century ago, still remains crucial for modern law.

See also: Critical Legal Studies; Durkheim, Emile (1858–1917); Feminist Legal Theory; Law and Everyday Life; Law, Sociology of; Legal Pluralism; Legal Realism, Marx, Karl (1818–83); Pestalozzi, Johann Heinrich (1746–1827); Weber, Max (1864–1920).

Bibliography

Austin, J.L., 1982. How to Do Things with Words, second ed. Oxford University Press, Oxford, UK.

Balkin, J., 2011. Constitutional Redemption: Political Faith in an Unjust World. Harvard University Press, Cambridge, MA.

Bentham, J., 2001 [1864]. Principles of the Civil Code: Part 1 – Objects of the Civil Law. Reprinted in: Bentham, Selected Writings on Utilitarianism. Wordsworth, London.

Brigham, J., 1996. The Constitution of Interests: Beyond the Politics of Rights. New York University Press, New York.

Collins, H., 1982. Marxism and Law. Clarendon Press, Oxford.

Cotterrell, R., 1999. Émile Durkheim: Law in a Moral Domain. Edinburgh University Press, Edinburgh, UK.

Cotterrell, R., 2006. Law, Culture and Society: Legal Ideas in the Mirror of Social Theory. Ashgate, Burlington, VT and Aldershot, UK.

Cotterrell, R., 2008 [2004]. Emmanuel Lévy and Legal Studies: A View From Abroad. Reprinted in: Cotterrell, Living Law: Studies in Legal and Social Theory. Ashgate, Burlington, VT and Aldershot, UK.

Edelman, L.B., Stryker, R., 2005. A sociological approach to law and the economy. In: Smelser, N.J., Swedberg, R. (Eds.), Handbook of Economic Sociology, second ed.

Ehrlich, E., 2002 [1936]. Fundamental Principles of the Sociology of Law. Transaction, New Brunswick, NJ (reprint).

Fitzpatrick, P., 1995. Being social in socio-legal studies. Journal of Law & Society 22, 105–112.

Frug, M.J., 1992. Postmodern Legal Feminism. Routledge, New York.

Gurvitch, G., 1947. Sociology of Law. Routledge and Kegan Paul, London.

Hertogh, M. (Ed.), 2009. Living Law: Reconsidering Eugen Ehrlich. Hart, Oxford, and Portland, OR.

Kronman, A., 1983. Max Weber. Edward Arnold, London.

López, I.F.H., 1996. White by Law: The Legal Construction of Race. New York University Press, New York.

Luhmann, N., 2004. Law as a Social System. Oxford University Press, Oxford, UK.

Mertz, E., 1994. A new social constructionism for sociolegal studies. Law & Society Review 28, 1243–1265.

Olivecrona, K., 1939. Law as Fact, first ed. Einar Munksgaard, Copenhagen/Humphrey Milford, London.

Pashukanis, E.B., 1978. Law and Marxism: A General Theory. Ink Links, London.

Pound, R., 1917. The limits of effective legal action. International Journal of Ethics 27, 150–167.

Savigny, F C von, 2002 [1831]. Of the Vocation of Our Age for Legislation and Jurisprudence, second ed. Law Book Exchange, Union, NJ (reprint).

Sumner, W.G., 2002 [1906]. Folkways: A Study of the Sociological Importance of Usages, Customs, Mores and Morals. Dover, Mineola, NY (reprint).

Law, Autopoiesis in

Ralf Rogowski, University of Warwick, Coventry, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term autopoiesis is a key concept that has been adopted by modern sociological systems theory in explaining the nature of social systems as communication systems that are capable of reproducing themselves. Legal autopoiesis (self-reproduction) results from processes of hypercyclical self-constitution within the legal system. The modern legal system operates with reflexive forms of regulation that recognize the autopoietic nature of both the legal system and the regulated systems. Modern law is increasingly becoming global law and is engaged in double reflexivity by responding to processes of societal constitutionalization in various function systems and regimes in the world society.

Legal autopoiesis refers to the idea that the legal system presupposes and reproduces itself. Modern sociological systems theory employs the concept of autopoiesis to emphasize the self-constituting nature of social systems. In analyses of the legal system, it constitutes a hybrid form of a legal theory and a sociological account of law.

Autopoiesis and the Functionally Differentiated Society

Autopoiesis describes the capacity of an entity to reproduce itself. As a concept, it was first introduced in theoretical biology to explain cognition and the essence of life (see Maturana and Varela, 1980) and was then further developed in general systems theory (for example, von Förster, 1984). It has been widely applied in mathematics, in the study of cognition, and in studies of the nervous system as well as in information systems, cognitive science, and artificial intelligence (see Mingers, 1995).

In general systems theory, the following definition is widely accepted: autopoietic systems "are defined as unities and as networks of production of components that recursively, through their interactions, generate and realize the network that produces them and constitute, in the space in which they exist, the boundaries of the network as components that participate in the realization of the network" (Maturana, 1981, p. 21). Maturana's definition operates with a notion of recursive closure of the system. An autopoietic system is not only able to produce and eventually to change its structures, but is also capable of reproducing its components, and in particular its system-specific elements, through recursive self-reference of these elements. Thus, its self-reproduction relies exclusively on internal operations and is not dependent on the environment.

Niklas Luhmann (1927–98) (*see* Luhmann, Niklas (1927–98)) is responsible for introducing the concept of autopoiesis to social systems theory (Luhmann, 1986; see also Luhmann, 1995, 2012). His approach transcends the Parsonian version of sociological systems theory, which operates with an input–output model of the social system and its environment. A crucial element of Luhmann's concept is the idea of communications as the ultimate, indecomposable elements of an autopoietic social system. Social systems are capable of reproducing themselves because they are communication systems.

Autopoietic social theory distinguishes between society and functionally differentiated social systems. Society is seen as a first-order social system, whereas the legal system is viewed as a second-order social system or societal subsystem. Society as a first-order social system differs from second-order social systems insofar as it has no other social system as an environment; society's environment consists of natural and psychic systems, i.e., the consciousness of human beings. In fact, human beings do not form part of society or any social system but belong to their environment; the individuality and the consciousness of human beings are features of highly complex psychic systems that operate separately from social systems (Luhmann, 1995, Chapter 7).

Autopoiesis of the Legal System

In accordance with general systems theory, autopoietic legal theory assumes that law is a special communication system that reproduces itself in self-referential processes. Legal communications, such as legal norms, legal decisions, or legal doctrine, are central to the autopoiesis of the legal system (Teubner, 1993; Luhmann, 2004). In order to determine whether a communication is part of the system, the legal system operates with a specific binary code: law/nonlaw or legal/illegal (Luhmann, 1990).

However, the code itself is not sufficient for the creation of law. The code must be translated into legal programs. Furthermore, the creation of law is linked to handling expectations and in particular to dealing with disappointments of expectations. The legal system operates with a specific combination of normative closure and cognitive openness (Luhmann, 1987). It upholds the existing norms and procedures but it is also capable of learning in response to disappointments, and provides mechanisms for its own change (Luhmann, 2014). An important feature of autopoietic legal theory is the radicalization of the notion of legal autonomy. This means that law is conceived not just as being a separate entity in society but that the theory ultimately aims at explaining how this autonomy is the result of internal processes and is maintained over time through closure of the system and self-reference (Luhmann, 1982b). It is the peculiar character of autopoietic law that "only the

law can change the law. This is not to deny that the environment (in particular the political system) has an impact on the legal system. But the legal system reproduces itself through legal events and only by legal events" (Luhmann, 1990, p. 229). On this basis, the modern legal system becomes positive and reflexive.

A further consequence of radical legal autonomy is the emergence of a special legal understanding of the external world. Like other function systems, the autopoietic legal system develops an epistemology according to its own criteria (Teubner, 1989). It is this specific worldview that guides its internal operations and thereby influences legal development.

Legal Evolution and Structural Coupling

The legal system forms part of the functionally differentiated modern society (other function systems include among others the economic, the political, and the religious systems). Each system within society evolves separately along its own trajectory, with its own internal structures and processes determining its development. In explaining the transformation of the legal system into an autopoietic system, Gunther Teubner points to a 'hypercycle' by which internal processes turn law into a self-reproducing system (Teubner, 1993, Ch. 3).

However, the evolution of social systems is not determined exclusively by internal conditions. Specific external relations that derive from structural couplings also have an effect. Structural coupling allows only selective exchanges between the systems. Structural coupling produces irritations and perturbations inside the system that are implemented by the system through its network of operations into further operations (Luhmann, 2013/2014, Vol. 1, Ch. 3).

These perturbations do not actually cause changes; rather, they serve as triggers that may initiate internally controlled operations in operationally closed systems. Thus, structural coupling of social systems does not contribute directly to the reproduction of the systems. It is simply the form in which the system presupposes specific states or changes in its environment and relies on them.

The constitution provides an example of structural coupling between the legal and the political systems (Luhmann, 2004, Ch. 10). When constitutions emerged in the eighteenth century, they helped recombine the functionally differentiated political and legal systems. From an autopoietic perspective, constitutions were adopted in response to problems of self-reference, paradoxes, and asymmetries in the legal and the political systems of the respective countries. In the political system constitutions are the mechanisms of self-binding sovereign power. In the legal system it provides the latter with the power to determine the constitutionality of positive law.

Reflexive Law and Legal Regulation

The concept of reflexive law, which at the beginning arose from a debate with the theory of responsive law, accompanies autopoietic legal theory (Teubner, 1983). Gunther Teubner argued that the modern legal system is characterized not only by formal and substantive rationality but also, increasingly, by reflexive legal rationality. The reflexive law concept was originally intended to combine the social theories of Jürgen Habermas and Niklas Luhmann. However, as the theory of reflexive law became more integrated with autopoietic legal theory (Teubner, 1993), the influence of Habermas' theory decreased and Luhmann's systems theory became dominant.

The reflexive law concept emphasizes the limits of legal regulation and suggests a radical rethinking of the role of law as facilitator and enabler of self-regulation in other social systems. It analyzes the regulatory capacities of the legal system, both in relation to the regulation of other systems and in relation to the legal system's capacity to regulate itself. Law is confronted with a regulatory trilemma according to Teubner (1985). First, it encounters ineffectiveness when the regulated system simply does not respond to the regulatory attempt. Second, if law simply imposes its own solution, it risks legalizing the regulated system. Third, if law yields to the imperatives of the regulated object, it risks becoming. Teubner (1988) suggests that by facilitating self-regulation, law can overcome this trilemma.

There have been a number of attempts to use the reflexive law concept to analyze developments in particular areas of law (see Callies et al., 2009, Part III). For example, Collins (1999) sees modern contract law as a form of reflexive regulation. Reflexive environmental law (Orts, 1995) and reflexive labor law (Rogowski and Wilthagen, 1994; Rogowski, 2013) have been suggested as concepts that can provide new insights in developments of legal subdisciplines. There have also been efforts to apply it in criminal justice studies (Nobles and Schiff, 2000) and to combine reflexive law and postmodern legal theory in constitutional law (Ladeur, 1997; see also Thornhill, 2011).

Law in the World Society

An "unavoidable consequence of functional differentiation" (Luhmann, 1982a, p. 132) is that modern society becomes global. By concentrating on one function and excluding other functions, each subsystem gains a capacity to include all function-specific communications, unlimited by territorial boundaries. Society as a whole becomes the world society (Luhmann 1982a). However, the world society encounters new problems. It can no longer integrate its subsystems by common territorial frontiers. A new metacode emerges in the modern, functionally differentiated world society in the form of a binary code of inclusion and exclusion (Luhmann, 2013/2014, Vol. 2, pp. 16-27).

It is this global dimension and its implications with which the reflexive and autopoietic law concept fundamentally challenges traditional legal analyses. Autopoietic legal theory predicts that legal globalization is unavoidable and that a world legal system is emerging. Global law is plural and derives from other sources than international law, for example, international commercial arbitration (on *lex mercatoria* see Teubner, 1997). It develops largely without the support of states and reflects "the expression of deep contradictions between colliding sectors of a global society" (Teubner and Fischer-Lescano, 2004). These 'regime-collisions' result from the specific nature of the functionally differentiated world society in which an overarching mechanism of coordination is lacking.

Since function systems develop internal structures and regulate themselves in accordance with their autopoietic needs, global law has to become reflexive and consider itself as facilitator rather than authoritative regulator of function systems. Self-regulation of function systems in the world society can be been seen as societal constitutionalization (Teubner, 2004, 2010, 2011), which requires from global law 'double reflexivity' (Teubner, 2012, Ch. 4). Global law has to realize that constitutions are not just second-order legal norms but generate fundamental structures in other social systems with which the world legal system is structurally coupled (*see* Globalization: Legal Aspects).

See also: Evolution, Natural and Social: Philosophical Aspects; Evolutionary Sociology; Globalization: Legal Aspects; Law: Change and Evolution; Legal Systems, Classification of; Luhmann, Niklas (1927–98).

Bibliography

Callies, G.-P., Fischer-Lescano, A., Wielsch, D., Zumbansen, P. (Eds.), 2009. Soziologische Jurisprudenz – Festschrift für Gunther Teubner zum 65. Geburtstag. De Gruyter, Berlin, Germany.

Collins, H., 1999. Regulating Contracts. Oxford University Press, Oxford, UK.

Förster, H., 1984. Observing Systems. Intersystems Publications, Salinas, CA.

Ladeur, K.-H., 1997. Post-modern constitutional theory: a prospect for the self-organising society. Modern Law Review 60, 617–629.

Luhmann, N., 1982a. The world society as a social system. International Journal of General Systems 8, 131–138 (Reprinted in Luhmann, N., 1990. Essays on Self-reference. Columbia University Press, New York, pp. 175–90).

Luhmann, N., 1982b. The autonomy of the legal system. In: Luhmann, N. (Ed.), The Differentiation of Society. Columbia University Press, New York, pp. 122–137.

Luhmann, N., 1986. The autopoiesis of social systems. In: Geyer, F., Zouwen, J. (Eds.), Sociocybernetic Paradoxes. Sage, London (Reprinted in Luhmann, N., 1990. Essays on Self-reference. Columbia University Press, New York, pp. 1–20).

Luhmann, N., 1987. Closure and openness: on reality in the world of law. In: Teubner, G. (Ed.), Autopoietic Law. A New Approach to Law and Society. De Gruyter, Berlin, pp. 335–348.

Luhmann, N., 1990. The self-reproduction of law and its limits. In: Luhmann, N. (Ed.), Essays on Self-Reference. Columbia University Press, New York, pp. 227–245.

Luhmann, N., 1995. Social Systems (original German title Soziale Systeme 1984). Stanford University Press, Stanford, CA.

Luhmann, N., 2004. Law as a Social System (original German title Das Recht der Gesellschaft 1993). Oxford University Press, Oxford, UK.

Luhmann, N., 2012. Introduction to Systems Theory. Polity, Cambridge, UK.

Luhmann, N., 2013/2014. Theory of Society, 2 vols. Stanford University Press, Stanford, CA.

Luhmann, N., 2014. A Sociological Theory of Law. Routledge, Abingdon, UK and New York.

Maturana, H., 1981. Autopoiesis. In: Zeleny, M. (Ed.), Autopoiesis: A Theory of Living Organization. North-Holland, New York, pp. 21–33.

Maturana, H.R., Varela, F.J., 1980. Autopoiesis and Cognition: The Realization of the Living. Reidel, Dordrecht, The Netherlands.

Mingers, J., 1995. Self-Producing Systems: Implications and Applications of Autopoiesis. Plenum, New York.

Nobles, R., Schiff, D., 2000. Understanding Miscarriages of Justice: Law, the Media, and the Inevitability of Crisis. Oxford University Press, Oxford, UK.

Orts, E.W., 1995. Reflexive environmental law. Northwestern University Law Review 89, 1227–1340.

Rogowski, R., 2013. Reflexive Labour Law in the World Society. Edward Elgar, Cheltenham UK.

Rogowski, R., Wilthagen, T. (Eds.), 1994. Reflexive Labour Law. Studies in Industrial Relations and Employment Regulation. Kluwer, Deventer, The Netherlands.

Teubner, G., 1983. Substantive and reflexive elements in modern law. Law and Society Review 17, 239–285.

Teubner, G., 1985. After legal instrumentalism? Strategic models of post-regulatory law. In: Teubner, G. (Ed.), Dilemmas of Law in the Welfare State. De Gruyter, Berlin, Germany, pp. 217–241.

Teubner, G. (Ed.), 1988. Autopoietic Law. A New Approach to Law and Society. De Gruyter, Berlin, Germany.

Teubner, G., 1989. How the law thinks: toward a constructivist epistemology of law. Law and Society Review 23, 727–757.

Teubner, G., 1993. Law as an Autopoietic System. Blackwell, Oxford, UK (original German title Recht als autopoietisches System 1989).

Teubner, G. (Ed.), 1997. Global Law without a State. Dartmouth, Aldershot, UK.

Teubner, G., 2004. Global private regimes: neo-spontaneous law and dual constitution of autonomous sectors in world society. In: Ladeur, K.-H. (Ed.), Public Governance in the Age of Globalization. Ashgate, Aldershot, UK, pp. 71–87.

Teubner, G., 2010. Fragmented foundations: societal constitutionalism beyond the nation state. In: Dobner, P., Loughlin, M. (Eds.), The Twilight of Constitutionalism? Oxford University Press, Oxford, UK, pp. 327–341.

Teubner, G., 2011. Constitutionalizing polycontexturality. Social & Legal Studies 20 (2), 210–229.

Teubner, G., 2012. Constitutional Fragments: Societal Constitutionalism and Globalization. Oxford University Press, Oxford, UK.

Teubner, G., Fischer-Lescano, A., 2004. Regime-collisions: the vain search for legal unity in the fragmentation of global law. Michigan Journal of International Law 25, 999–1046.

Thornhill, C., 2011. A Sociology of Constitutions. Cambridge University Press, Cambridge, UK.

Law, Bioethics, and Biotechnology

Michele Goodwin, University of California – Irvine, Irvine, CA, USA
Allison M Whelan, University of Minnesota, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the emergence of bioethics as a legal discourse and the relationship between technology and bioethics. As a field of study, bioethics functions as a tool to analyze the multifaceted conundrums presented by modern medicine and a way by which those problems can be addressed. This article provides an overview of this broad field of study, beginning with the development of the Nuremberg Code, Declaration of Helsinki, and Belmont Report, and addressing the pivotal medical cases that ultimately spur legislative action, the drafting of codes of professional ethics, and the shaping of legal discourse. The article offers insights on several twentieth-century cases that inspired the development of bioethics in the United States and beyond and has since influenced the development of bioethics and technology throughout the world.

Introduction

According to Warren Thomas Reich, the term 'bioethics' emerged in 1971 – coined by Van Rensselaer Potter, one of the first scholars to advocate establishing a discipline of bioethics (Reich, 1994: p. 320). Potter explained bioethics as "a new discipline that would combine human value considerations with science in general and especially with the 'realities' of biological science in the Man/Earth relationship" (Potter, 1971: p. 1088). A few years later, Potter also endorsed a similar definition put forth by Dr. LeRoy Walters, Director of the Kennedy Institute Center for Bioethics (Potter, 1975). Walters described bioethics as "the study of social and ethical questions motivated by developments in the fields of biology and medicine, and includes subject matter drawn from the natural and social sciences and the humanities" (Potter, 1975: p. 2299). However, the professional standards and ethics framework known today as 'bioethics' began shortly after World War II in a movement to 'humanize' medical education and practice, largely in response to the atrocities carried out by Nazis on Holocaust victims (Pellegrino, 1999: p. 75; Annas, 2009).

Bioethics reflects the convergence of several discourses: law, theology, medical ethics, and philosophy, into a general framework that attempts to answer complex questions that frequently place the individual, medical providers, society, and the state in ethical conflict. Those who frame the field – bioethicists – identify thorny medical dilemmas and attempt to provide clear, normative recommendations to address the challenges of contemporary medical research and delivery. By extension, bioethicists attempt to answer unsettled ethical and moral dilemmas trapped within the intersections of law, medicine, and technology.

The dominant moral dilemmas that frame traditional bioethical concerns address medical experimentation on vulnerable subjects, physicians assisting their patients in suicide or euthanasia, scarce resource allocation and medical futility, human trials to develop drugs, organ and tissue donation, cloning, xenotransplantation, abortion, human enhancement, mandatory vaccination, and much more. The term 'bioethics' provides a lens, language, and guideposts to the study of medical ethics. It is worth noting, however, that medical experimentation is neither new nor exclusive to one country.

Cleopatra, for example, is alleged to have facilitated medical research on her handmaids by authorizing their impregnated wombs cut open for medical experimentation (Bolton, 1881: p. 4354). In 1796, Edward Jenner, an English physician, injected cowpox into James Phipps, a healthy 8-year-old boy (Reidel, 2005: p. 24). In the 1840s, an American gynecologist, Dr. Marion Sims, who is hailed as the 'father of gynecology,' conducted vaginal experiments on slave women without the aid of anesthesia (Kuhn McGregor, 1998: p. 44). In 1900, Walter Reed injected nearly 24 immigrant workers in Cuba with the agent for yellow fever (Cefrey, 2002: p. 29). In the late 1920s, several children in Lubeck, Germany, died from accidental exposure to a lethal tuberculosis vaccine (Coberly and Comstock, 2006: p. 76). In that same decade, Japanese military officials experimented with germ warfare, killing thousands. During the early twentieth century, governments often justified brutal experiments as necessary to protect national security or promote the public health (Harris, 2003). These two rationales remain potent.

Increasingly, economic concerns are included under the broader umbrella of bioethics. For example, should markets serve as an answer to the perceived organ shortage? Should governments rein in the modern-day outsourcing of gestational surrogacy? Bioethics finds its relevance in the persistent absence of formal law on contemporary ethical research problems. Some scholars claim that biotechnology doggedly outpaces law, stretching its reach far beyond the slow, lagging legislative process, which they argue forestalls rather than enables the development of coherent, rational controls over the direction of biotechnologies. In this context, legal rules result by default and happenstance rather than rational engagement.

Much like law and economics, feminist jurisprudence, and other 'new' areas of law, bioethics is an expanding discipline, gaining traction as an authoritative and even necessary method of evaluating medical–legal conflicts. This article tracks the trajectory of modern bioethics, assessing the rise of bioethics, which can in part be attributed to deeply entrenched formalism in the legal jurisprudence, and its turn toward professionalism in the medical and legal academies. It discusses how international tragedies gave rise to guiding documents and principles that ground contemporary bioethics (*autonomy*, *non-malfeasance*, *beneficence*, and *social justice*) and explains that as

technology forges ahead, law and the manufacture of legal rules often lags far behind. The field of bioethics often serves to fill in the gap, offering philosophical insights grounded in law, philosophy, and values to address health care dilemmas. In recent decades, bioethics has contributed to the intellectual ordering of modern medicine and law.

Early Beginnings: World War II, Nuremberg, and the Modern Bioethics Movement

Perhaps the oldest foundational tenet of bioethics can be traced to Hippocrates, the ancient Greek physician credited with the 'Hippocratic Oath,' a pledge commonly taken by new physicians and other health care providers to protect and hold patients' interests above their own. The core of the pledge reflects the Latin edict *primum non nocere*, 'first do no harm,' a moral injunction embraced by health care providers – even if its origins remain somewhat a mystery (Smith, 2005). To this day, the delivery of medicine presents moral, legal, and ethical challenges uneasily resolved by medicine alone. Among these are the demands for medical and scientific information and the use of human participants in nontherapeutic studies to acquire data. Frequently, mixed incentives dominate contemporary biotechnology and medical research practices, sometimes aligning the interests of patients and researchers and at other times subordinating patients' needs to scientists' priorities, such as patenting discoveries, achieving tenure, fame and wealth, gaining notoriety, and assisting national security through military experimentation.

Internationally, governments have justified nonconsensual medical experiments conducted on children, the mentally ill, prisoners, and soldiers based on national security. The collapse of Nazi Germany, the unearthing of medical experiments carried out against Chinese and Russian prisoners of the Japanese during World War II, and the subsequent trials at Nuremberg, where 'Third Reich' physicians and researchers revealed the mass horrors of their human experimentation and broader brutality in the quest for scientific knowledge mark the beginnings of contemporary bioethics. Documents, including photographs and witness statements uncovered after World War II, reveal grotesque research and experimentation conducted on vulnerable populations in China, Poland, and Germany. At a Chinese outpost, for example, the Imperial Japanese army erected a covert military installation where researchers "expose[d] prisoners to [sub-zero temperatures] to … freeze them and then try various techniques to literally dehydrate them, to see what was the most effective way of dealing with frost bite, so that that could be used in warfare as well, both to protect Japanese troupes as well as to effect the enemy" (Targeted Individuals Europe, 2011). At least 3000 died as a result of Japanese medical experimentation. Japanese officials attempted to justify experiments "to develop weapons of biological warfare, including plague, anthrax, cholera and a dozen other pathogens" based on national security interests (Kristof, 1995). In Germany, the Third Reich conducted similar atrocities against Jews and others. Their murderous experiments resulted in an international trial.

The Nuremberg Doctors' Trial, one of 13 trials at Nuremberg, began in 1946 and concluded 1 year later. The trials were conducted by the International Military Tribunal at Nuremberg and were presided over by judges from the four Allied powers (Annas, 2009). The Trials contributed to the articulation and establishment of well-recognized human rights principles in law and bioethics, namely that voluntary consent is an essential component of any medical experiment; any experiment should avoid subjecting the participants to unnecessary suffering; and research subjects must be at liberty to withdraw from human experimentation at any time (The Nuremberg Code, 1949).

Nuremberg Code

The Doctors' Trial ('Trial') stood apart because only judges from the United States presided over those prosecutions (Annas, 2009). At the Trial, prosecutors accused 23 physicians and scientists of performing "murderous and torturous experiments" in the Nazi concentration camps, such as high-altitude and freezing experiments, which resulted in the planned and expected deaths of research subjects (Annas, 2009: p. 21). Researchers who planned and conducted the experiments justified their research on nationalist-medical grounds, claiming the results could benefit German pilots and the broader, German common good (Annas, 2009). Prosecutors also pointed to more than 350 000 forced sterilizations conducted under the Nazi's eugenics regime. In response, Nazi doctors proclaimed that prosecutors were blinded by hypocrisy and exceptionalism, as the U.S. also engaged in eugenics by permitting the broad-scale sterilization of individuals considered socially or mentally unfit. As proof, Nazi doctors cited Justice Oliver Wendell Holmes' opinion in *Buck v. Bell*, a 1927 U.S. Supreme Court opinion sanctioning compulsory sterilization of the 'feeble-minded' (*Buck v. Bell*, 274 U.S. 200, 205 (1927)). Despite this creative defense, which cast light on U.S. involvement in eugenics, seven of the Nazi doctors were eventually sentenced to death by hanging (Annas, 2009).

From this historical perspective, the Nuremberg Trial ushered in modern bioethics. The field's first code of ethics, *The Principles of Nuremberg* (commonly referenced as the *Nuremberg Code*), emerged directly from this landmark prosecution (The Nuremberg Code, 1949). The Principles of Nuremberg, a ten-point statement governing human experimentation, recognized voluntary consent as a central tenet of human research and emphasized that research ought to benefit the social good and doctors ought to prioritize the avoidance of suffering. The 10 principles continue to serve as a framework for medical ethics:

1. Voluntary consent of the human subject is essential.
2. The experiment should be such as to yield fruitful results for the good of society, unprocurable by other methods or means of study, and not random and unnecessary in nature.
3. The experiment should be so designed and based on the results of animal experimentation and a knowledge of the natural history of the disease or other problem under study, that the anticipated results will justify the performance of the experiment.
4. The experiment should be so conducted as to avoid all unnecessary physical and mental suffering and injury.
5. No experiment should be conducted, where there is an a priori reason to believe that death or disabling injury will

occur; except, perhaps, in those experiments where the experimental physicians also serve as subjects.
6. The degree of risk to be taken should never exceed that determined by the humanitarian importance of the problem to be solved by the experiment.
7. Proper preparations should be made and adequate facilities provided to protect the experimental subject against even remote possibilities of injury, disability, or death.
8. The experiment should be conducted only by scientifically qualified persons. The highest degree of skill and care should be required through all stages of the experiment of those who conduct or engage in the experiment.
9. During the course of the experiment, the human subject should be at liberty to bring the experiment to an end, if he/she has reached the physical or mental state, where continuation of the experiment seemed to him/her to be impossible.
10. During the course of the experiment, the scientist in charge must be prepared to terminate the experiment at any stage, if he/she has probable cause to believe, in the exercise of the good faith, superior skill, and careful judgment required of him/her, that a continuation of the experiment is likely to result in injury, disability, or death to the experimental subject (The Nuremberg Code, 1949).

Subsequent documents, specifically the Declaration of Helsinki and the Belmont Report, further developed the bioethics cannon (as discussed below).

Despite near-global agreement that Nuremberg was the birthplace of modern bioethics, some scholars regard the origins of American bioethics as deeply rooted in a subsequent era, the 1960s, a time of heightened medical paternalism. If American bioethics was a phenomenon of the 1960s, the U.S. model of medical philosophy could be distinguished from that of the rogue terrors of Nazi Germany. Albert R. Jonsen, for example, describes the Nuremberg Code as a "beginning that would become bioethics," but places greater emphasis on later developments, such as the Declaration of Helsinki (Jonsen, 1998: p. 134). Similarly, David J. Rothman (1991: p. 62) argued that the Nuremberg Code 'might have served' as a model for American bioethics, but ultimately emphasizes the 1960s as the beginning of modern American bioethics. According to Rothman, the Nuremberg Code was not fully taken up in the United States because "the events described at Nuremberg were not perceived by researchers or commentators to be directly relevant to the American scene. The violations had been the work of Nazis, not doctors; the guilty parties were Hitler's henchmen, not scientists. … Madness, not medicine, was implicated at Nuremberg" (Rothman, 1991: pp. 62–63). Notwithstanding this account, the Nuremburg Code continues to have tremendous influence in the field of bioethics.

Declaration of Helsinki

In 1964, the World Medical Association issued the Declaration of Helsinki on human research (World Medical Association, 2008). Debates emerged about whether the declaration served as an alternative to the Nuremberg Code – which many physician-researchers considered 'overly rigid' because of its forceful edicts and focus on the health and well-being of research subjects – or as an enhancement to the prior document. In the end, the most salient aspects of the Nuremberg Code and the Declaration of Helsinki are quite similar, as both provide guidance to physicians engaged in human subjects research, emphasizing the need to prioritize the *individual's* health and well-being above all other goals.

A decade later, the gathering strength and relevance of a bioethics framework within the legal and medical academies culminated in The Belmont Report, drafted in 1979 by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research sponsored by the U.S. Department of Health, Education, and Welfare (DHEW). The Report proposed "to identify the basic ethical principles that should underlie the conduct of biomedical and behavioral research involving human subjects and to develop guidelines which should be followed to assure that such research is conducted in accordance with those principles" (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). Their efforts began in the wake of the discovery of the Tuskegee Syphilis Study.

Concretizing Medical Bioethics and Responding to Biotechnology

Paramount to the continued development and refinement of a bioethics agenda were two issues (1) researchers' disregard of the lofty aims of the Nuremburg and Helsinki declarations; and (2) the perception of law's failure to take seriously or incorporate the field's principles into practice. By the 1970s, the demands for a respected and legally sound bioethics agenda became clear due to global controversies involving human experimentation as well as the rapidly expanding fields of medicine, including organ transplantation, sophisticated pharmaceutical therapies, and improved biotechnologies. More than any other country, however, the U.S. advanced at the forefront of both biotechnologies and experimentation. Between the 1950s and 1970s, American physicians, scientists, and government agencies rapidly expanded medical and biotechnological knowledge through human experimentation. A cursory view of this time period provides a glimpse of the depth and range of unethical and illegal experimentation conducted by researchers.

1950: Dr. Joseph Stokes infects 200 women prisoners with viral hepatitis.
1952–74: Dr. Albert Kligman conducts skin experiments on inmates at Holmesbrook Prison.
1953–57: Uranium is injected into nearly a dozen patients at Massachusetts General Hospital.
1953–1960: Central Intelligence Agency conducts Lysergic acid diethylamide (LSD) experiments on hundreds of subjects.
1950s–70s: U.S. Army studies 2300 Seventh Day Adventist soldiers.
1960s: Pharmaceutical companies test 150 experimental drugs on prisoners in Pennsylvania (Hassner Sharav, n.d.).

It became apparent from medical experiments in developing nations, as well as the Tuskegee Syphilis Study in the U.S. (discussed below) and increased medical rationing of sophisticated technologies, that race, sex, and class served as

moderating forces in the acquisition of biotechnological knowledge and delivery of health care. In 1974, U.S. law governing human research finally came to fruition – in significant response to the infamous syphilis study. The National Research Act, enacted by the 93rd Congress, provided a federal response to human research conducted in the U.S. imposing guidelines for researchers and establishing a bioethics commission to advance further recommendations. The law created the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (National Commission), mandated voluntary, informed consent from all persons participating in studies funded through the DHEW, and required that all research funded through DHEW be reviewed by institutional review boards (IRBs). The end result was the 'Belmont Report,' which introduced what are now considered foundational bioethics principles: (1) respect for persons, (2) beneficence, and (3) justice (*The Belmont Report*, 1979: pp. 23193–23194). Today, the four main principles of bioethics are (1) Autonomy, (2) Nonmalfeasance, (3) Beneficence, and (4) Justice.

The National Commission served as a blueprint for President William Clinton's subsequent formation and launch of the National Bioethics Advisory Commission in 1995 and President George Bush's President's Council on Bioethics, established in 2001 (Protection of Human Research Subjects and Creation of National Bioethics Advisory Committee, 1995; Creation of the President's Council on Bioethics, 2001).

By the 1990s, European nations responded in three ways as follows: (1) by establishing the International Bioethics Committee in 1993, (2) by creating the 1997 Oviedo Convention on Human Rights and Biomedicine, and (3) by establishing the European Conference of National Ethics Committees (COMETH), which promotes public debate on biotechnologies and health sciences (Council of Europe, n.d.). The overarching goal for COMETH is to research interests against protecting human subjects. In recent years, several major European protocols have been established to address genetic testing for health purposes, cloning, transplantation, and biomedical research.

Cases That Shaped Bioethics

Bioethics is shaped by the previously discussed international documents along with a series of pivotal cases highlighting the tensions between medicine, autonomy, and the physician–patient relationship (Fox and Swazey, 2008: p. 215).

There is no shortage of cases, both in the U.S. and worldwide, that implicate bioethical principles, particularly as researchers pursue patents or develop new and lucrative drugs. Moreover, much of this research – particularly radiation experiments – were performed on powerless, vulnerable groups – African-Americans, mentally disabled children, indigenous populations, and the elderly. The cases discussed below do not serve as an exhaustive list. Rather these substantial medical ethics breaches offer insight into the relevance of bioethics, given that in each instance the research was passionately defended not only by the research teams, but by scientists in the field. These cases have greatly influenced the development of bioethics both in the United States and throughout the world.

Research on Women

Coercive Contraception

During the 1970s, in the wake of women's liberation movements occurring in some parts of the world, biotechnological human experiments involving unapproved contraceptives were taking place in some countries. In Zimbabwe, for example, studies using the drug Depo-Provera for contraceptive purposes attracted the attention of human rights groups (Kaler, 1998). Human rights activists alleged that the drug was used to control the population. They pointed to the fact that the drug was not approved for contraceptive use in the U.S. where it was manufactured, but was being used for that purpose in rural Zimbabwean communities. Reports emerged that the drug was administered to some women without their consent and sometimes without their knowledge (Rebecca Project for Human Rights, 2013; Sambisa, 1996). Nevertheless, in a paper published in the *Journal of Medical Ethics*, Professors Malcolm Potts and John Paxman (1984) wrote that despite the 'circumscribed' history of the drug's testing and distribution among poor women in developing nations, to remove it from the market would undermine the choice of women in developing countries to have broad access to various contraceptive medicines. In the wake of investigations, government officials in Zimbabwe banned the drug's use except for treating kidney and uterine cancer (Kaler, 1998).

Research on the Elderly

Jewish Chronic Disease Hospital Study (1963)

In 1963, research began at New York's Jewish Chronic Disease Hospital to study whether the body's ability to reject cancer cells was due to the cancer itself or to another malady, such as a compromised immune system (Katz, 1972). Previous studies suggested that healthy individuals' bodies could reject cancer cells promptly and researchers believed that 'debilitated patients' would similarly reject cancers, albeit at a substantially slower rate (SUNY Cortland Institutional Review Board, n.d.: pp. 4–5).

The study involved the intentional injection of live, foreign cancer cells into 22 senile patients hospitalized with various chronic conditions (SUNY Cortland Institutional Review Board, n.d.). Despite the establishment of the Nuremberg principles and sharp similarities to Nazi experimentation, the study advanced without review board approval or interventions on behalf of the patients. The research subjects were uninformed and misled as researchers convinced patients they would receive a "harmless substance which might cause a slight discomfort" (Morse, 1967: pp. 376–377). The researchers defended the study, claiming they had good reason to "predict that the cancer cells were going to be rejected" (SUNY Cortland Institutional Review Board, n.d.: p. 6). The Board of Regents of the State University of New York subsequently reviewed the study and investigated its lead researcher. The Board concluded that the doctors failed to submit the research protocols to the hospital's research committee and that the physicians responsible for the patients' clinical care were not consulted. The researchers were

found guilty of fraud and unprofessional conduct (SUNY Cortland Institutional Review Board, n.d.: p. 6).

Research on Children and the Mentally Ill

Willowbrook Hepatitis Study

For decades, children's bodies provided tremendous resources for human research, despite their inability to consent to medical trials. The Willowbrook hepatitis study, which began in the mid-1950s, is one such example. The was led by Dr. Saul Krugman and conducted at the Willowbrook State School in Staten Island, NY, a facility for mentally disabled children. Willowbrook remains "a potent symbol of unethical research," with some referring to it as the 'pediatric Tuskegee' (Robinson and Unruh, 2008: p. 80). As part of the study, children housed at Willowbrook were intentionally injected with hepatitis to determine the effectiveness of gamma globulin injections in protecting against hepatitis. The research was based on the premise that infection with a mild form of a viral agent can induce immunity. Krugman and others involved in the study justified the research, claiming that given the high prevalence of infection in the institution, it was practically inevitable that the children would become infected, regardless of whether they were enrolled in the study. Researchers concluded that results of the study could be extremely beneficial to society and to the institutionalized children. The doctors claimed the research could potentially control the 'endemic situation' at Willowbrook. And in fact, one of the results of the research was a reduction in the incidence of hepatitis among patients and employees by 80–85% (Robinson and Unruh, 2008: p. 82).

In the study's aftermath, the scientific community was divided. Some scholars claimed Krugman was unfairly vilified and that no real harm had materialized because the research subjects were already institutionalized and suffered far greater material harms than hepatitis infection. Walter M. Robinson and Brandon T. Unruh (2008) argued that the children's parents were not unacceptably coerced and that Krugman's research was motivated by the desire to help children. However, one of the study's most outspoken critics, Dr. Henry Beecher, an anesthesiologist at the Massachusetts General Hospital and professor at Harvard Medical School, published a paper in the *New England Journal of Medicine* entitled 'Ethics and Clinical Research,' citing the Willowbrook study as an egregious example where patients were not informed that they were subjects of an experiment (Beecher, 1966).

Pfizer Meningitis Study in Nigeria

In 1996, clinical researchers at Pfizer, Inc. (Pfizer), a large American drug company, learned of a meningitis outbreak in Nigeria's northern Kano State. By the end of this epidemic, there were at least 250 000 recorded cases of meningitis and 25 000 deaths. Pfizer, working alongside volunteer agencies (such as Doctors Without Borders) and Nigerian health officials, used the epidemic as an opportunity to conduct a randomized trial of Pfizer's antibiotic 'Trovan.' Pfizer's stated goal for this trial was to obtain "valuable empirical and clinical data on Trovan … [and to] pioneer[] a breakthrough treatment for the Third World" (Ezeome and Simon, 2010: p. 1). At the time of the trial, Nigeria was run by a military government with "one of the world's worst human rights corruption records," which meant Pfizer faced fewer regulatory hurdles in beginning and carrying out the trials (Stephens, 2000). Pfizer prepared and initiated the study in a mere 6 weeks, in contrast to the process in the U.S., which typically takes a year or more (Stephens, 2000).

Two hundred Nigerian children with meningitis were enrolled in the trial. And although the antibiotic was promising, it was also a 'potential blockbuster' for Pfizer, with Wall Street analysts predicting that Pfizer might earn 1 billion dollars if Trovan were approved for its many potential uses (Stephens, 2000). Throughout the trial, Pfizer did not follow various guidelines governing meningitis experiments, such as conducting a second spinal tap within a few days of beginning treatment. In all, 11 children died and many others continued to suffer meningitis-related symptoms including deafness, blindness, seizures, and loss of mobility. According to a Nigerian doctor, the patients did not understand or know they were in a research trial; "they just knew they were sick" (Stephens, 2000). And although Pfizer claims that local nurses explained the research to the children's families, there are no signed consent forms to prove its claims.

Trovan was eventually approved for sale in the U.S. but it was *not* approved for use in children and European regulators prohibited its sale completely, given discrepancies in the Nigerian test results as well as the drug's numerous side effects, including liver damage and death. Pfizer 'roared' in and out of Nigeria in 'a relative eye blink' – leaving 3 weeks after the study began. In 2011, the parents of four children who died during the trial were each awarded $175,000 (Smith, 2011). More compensation settlements were expected to follow from a $35 million fund created under the settlement between Pfizer and Kano State (McNeil, 2011).

Government-Sponsored Research

U.S. Sponsored Syphilis Studies in Tuskegee and Guatemala

In 1972, the most notorious human experiment conducted in the United States ended after four decades (Washington, 2006). Several hundred poor, African-American men infected with syphilis, a sexually transmitted disease caused by a bacterium and treatable with penicillin, served as the uninformed participants in a deadly 40-year study supervised by the U.S. Public Health Service. The men were told only that they suffered from a condition referred to as *bad blood* and that the state would provide free treatments for the illness, the men unwittingly surrendered to the behest of the researchers, providing doctors and government officials broad access to their flesh (Washington, 2006: p. 162).

The men's bodies became human canvases to be poked, prodded, and punctured with lumbar spinal taps. Indeed, the purpose of the study was neither to keep the men alive nor to find a cure for syphilis, but rather to study the effects of syphilis, particularly at death. For this reason, the participants were not provided penicillin – all were treated with pink pills of no medicinal value. In 1932, when the infamous "Study of Syphilis in the Untreated Negro Male" – euphemistically known as the *Tuskegee Study* – began, no research protocols existed in the United States to define the scope of human participation in medical experimentation (Washington, 2006: p. 157). Neither did federal or state laws guide medical experimentation, because such laws had never been enacted.

Ethical protocols remained at the discretion of doctors and their universities or hospitals. By 1947, penicillin emerged as the gold standard treatment for syphilis. However, researchers withheld this vital information from the infected men and their families. The Public Health Service doctors and a team of medical researchers denied the patients antibiotic treatments, withheld information, and prevented or obstructed their access to nearby treatment programs.

Similarly, in Guatemala from 1946 to 1948, the U.S. government, in collaboration with Guatemalan officials in the health ministries and military, infected prisoners, individuals with low IQs, and soldiers with syphilis and other sexually transmitted diseases. The Guatemalan experiments were conducted on 1500 men and women without informed consent, just like the men at Tuskegee. In Guatemala, however, the research subjects were later treated with antibiotics (Stein, 2010). Ironically, the antibiotic treatments were a test protocol and likely not meant as a proactive health measure.

In 1997, 55 years after the Tuskegee study commenced, President William Clinton issued an apology to the few remaining survivors whose crippling conditions made it remarkable that they somehow survived (Washington, 2006). In 2010, the U.S. issued an apology for its involvement in the Guatemalan syphilis study (Stein, 2010).

Human Plutonium Experiments

Human plutonium injection experiments were carried out during and after the Manhattan Project, which ran from 1942 to 1946 (Moss and Eckhardt, 1995). The experiments sought to develop diagnostic resources to "determine the uptake of plutonium in the body from the amount excreted in the urine and feces," an essential tool for protecting workers producing and fashioning plutonium metals for use in atomic bombs (Moss and Eckhardt, 1995: p. 177). The research subjects were injected with plutonium. Whether the patients injected with plutonium provided consent cannot be answered with certainty because all direct participants are deceased and documents that could shed light on this question are scattered, incomplete, or nonexistent. The evidence that does exist, however, makes it almost certain that patients were uninformed about being injected with plutonium. Indeed, until the end of World War II the word 'plutonium' was undisclosed (Moss and Eckhardt, 1995: pp. 213–214).

Cold War Experiments

From the end of World War II and well into the 1970s, the Cold War provided the impetus for government-funded and -conducted experiments. During that period, the Atomic Energy Commission, the Defense Department, the military services, the CIA, and other governmental entities exploited prisoners, drug addicts, mentally disabled patients, mentally ill patients, college students, and soldiers in experiments "to test the effects of everything from radiation, LSD and nerve gas to intense electric shocks and prolonged 'sensory deprivation'" (Budiansky et al., 1994). Rarely were these research subjects aware of the scale and scope of the research, risks involved, or even that they were part of an experiment. Collectively, these experiments are known as 'Project MKULTRA' (or 'MK-Ultra'). In 1976 and 1977, the United States Senate conducted investigations and held a joint committee hearing on MKULTRA (Project MKULTRA, 1977). According to the hearing report, the project sought to study and develop the covert use of biological and chemical materials for 'behavioral control' (Project MKULTRA, 1977: p. 2).

General Vernon A. Walters, Deputy Director of the CIA, testified in 1975 that the experiments "included covert drug tests on unwitting citizens at all social levels, high and low, native Americans and foreign" (Project MKULTRA, 1977: p. 2). Some of the experiments involved administering LSD to unknowing subjects. Researchers destroyed many of the documents related to these experiments. Nevertheless, the uncovered records "reveal a far more extensive series of experiments than had previously been thought," including involvement of more than 86 universities or institutions (Project MKULTRA, 1977: p. 3). Researchers and government officials involved with the Cold War studies were never prosecuted (Washington, 2006). Among those researchers, some continued to lead distinguished careers.

The United States was not alone, however, in conducting controversial and arguably unethical research experiments. Fearing that the Soviet Union was planning a chemical and biological attack, the British Government sought to study germ dispersion. To do so, scientists from Porton Down, the United Kingdom's government military science park, drove a Land Rover through local towns, spraying Zinc Cadmium Sulfide into the air in an attempt to simulate germ warfare. Although Porton Down's authorities argued that the Zinc Cadmium Sulfide posed no health danger, the residents were nevertheless unknowing and unconsenting research subjects. Porton Down's scientists also performed countless experiments on soldiers, who often agreed to be involved in research "in return for payment and a precious pass for 3 days leave." Many were not informed about the substance or purpose of the research and in 1953, serviceman Ronald Maddison died after taking part in a Sarin nerve gas experiment (British Broadcasting Company).

Prisoner Research

During the latter half of the twentieth century, researchers frequently sought the involvement of prisoners in medical research because the stability of prison life (controlled diet, ready availability of subjects for continued monitoring and follow-up) made incarcerated populations attractive research subjects (Office for Human Research Protections, 1993). However, ethicists and others concerned with the treatment of human research participants increasingly question using prisoners as research subjects, contending that the very fact of incarceration makes it difficult, if not impossible, for prisoners to give voluntary and informed consent. For example, prisoners were subjects in many of the MKULTRA and other government-sponsored experiments during World War II and the post-World War II era. In fact, after World War II, the United States was the only nation in the world continuing to legally use prisoners in clinical trials, with federal, pharmaceutical, and cosmetic companies funding a 30-year 'boom' in research involving prisoners (Washington, 2006). In an effort to address abuse in prison medical research, regulations and some federal agencies now require special considerations or protections for

research involving the incarcerated (Office for Human Research Protections, 1993).

Contemporary Challenges

One distinction in contemporary medical research that provides some distance from past ethical breaches concerns the role and rule of law. The firm establishment and broad acceptance of bioethical principles offers greater clarity at the macro level to universities, hospitals, and industry, as well as at the micro level, imposing rules and regulations on researchers. The establishment of firm research protocols has likely reformed modern medicine to some degree. Yet, despite mandatory research ethics training, the establishment of IRB processes, and increased government oversight throughout the world, challenges remain. And, much like egregious experiments of the past, research subjects most likely to be exploited are vulnerable by status: poor, uneducated, and of color. Pfizer's meningitis studies involving Nigerian children provide an important example.

Indeed, contemporary challenges in bioethics reflect the dynamics of past medical research: conflicts of interests, lack of meaningful enforcement, perverse incentives, and unfettered access to particularly vulnerable subjects. However, one major difference separating the early research experiments on vulnerable subjects from that of today is the use of tort law when codes of ethics are violated.

Legal settlements may also cross borders. For example, litigants in Nigeria sued Pfizer in U.S. court after their case was dismissed in Nigeria. Litigants argued that a fair trial could not be pursued in their homeland and successfully argued that Pfizer could be sued in U.S. courts for illegal conduct abroad (*Abdullahi v. Pfizer, Inc.*, 562 F.3d 163 (2d Cir. 2009)). Ultimately, Pfizer settled the claim (McNeil, 2011).

Conclusion

Contemporary bioethics involve a host of medical–legal concerns that extend beyond the conduct of medical researchers to include issues involving the right to refuse treatment, the constitutional 'right to die,' assisted suicide, and the balance between state interest and that of the individual. Bioethical challenges play out in the tensions between family members (illustrated by the conflicts over authority to remove life-sustaining support), in cases involving forced medical treatment, conflicts between pregnant women and fetuses, and medical futility, such as anencephalic babies kept on expensive, life-sustaining medical therapies. The scope of ethical concerns in medicine will continue to expand as technology advances. Most contemporary bioethical challenges and concerns reflect concerns from generations past, including individual autonomy, dignity, fiduciary obligations, justice, and informed consent. The field, though relatively nascent, has a long and complicated history and makes a distinct contribution among legal discourses.

See also: Biobanking: Ethical Issues; Bioethics in the Post-genomic Era; Bioethics: Philosophical Aspects; Bioethics; Common Rule: US Regulatory Protections for Human Research Participants; Direct-to-Consumer Personal Genetic Testing; Ethical Dilemmas, Research, and Access to Treatment; Ethics for Biomedical Research Involving Humans: International Guidelines; Eugenics as a Basis of Population Policy; Eugenics as an International Movement; Eugenics, History of; Euthanasia; Gender and Social Research; Genetic Screening for Disease; Informed Consent of Research Subjects; Insurance and the Law; Law and Risk; Placebo-Controlled Trials, Ethics of; Privacy of Individuals in Social Research: Confidentiality; Research Ethics, Cross-Cultural Dimensions of; Torts in Law.

Bibliography

Abdullahi v. Pfizer, Inc., 562 F.3d 163 (2d Cir. 2009).
Annas, George J., 2009. The legacy of the Nuremberg doctors' trial to American bioethics and human rights. Minnesota Journal of Law, Science, and Technology 10 (1), 19–40.
Beauchamp, Tom L., Childress, James F., 1979. Principles of Biomedical Ethics. Oxford University Press, Oxford.
Beecher, Henry K., 1966. Ethics and clinical research. New England Journal of Medicine 274 (24), 1354–1360.
Bolton, H. Carrington, 1881. The early practice of medicine by women. Scientific American Supplement XI (273), 4353–4355.
Bouvia v. Superior Court, 179 Cal. App.3d 1127 (Cal. Ct. App. 1986).
British Broadcasting Company. Porton Down—The Unwitting Victims. Available at: http://www.bbc.co.uk/insideout/west/series1/porton-down.shtml (accessed 17.03.14.).
Buck v. Bell, 274 U.S. 200 (1927).
Budiansky, Stephen, Goode, Erica E., Gest, T. (Eds.), 1 January 1994. The Cold War Experiments. U.S. News & World Report [online] Available at: http://www.usnews.com/usnews/news/articles/940124/archive_012286.htm (accessed 12.10.13.).
Cefrey, H., 2002. Yellow Fever. Rosen Publishing Group, Inc., New York.
Coberly, Jaqueline S., Comstock, George W., 2006. Epidemiology of tuberculosis. In: Raviglione, Mario C. (Ed.), Reichman and Hershfield's Tuberculosis, third ed. Informa Healthcare, New York.
Council of Europe, n.d. Bioethics: The Necessary Balance between Progress and Respect for Human Dignity. Available at: http://hub.coe.int/what-we-do/health/bioethics (accessed 10.04.14.).
Creation of the President's Council on Bioethics (28 November 2001), 66 Fed. Reg. 13237.
Cruzan v. Director, Missouri Department of Health, 497 U.S. 261 (1976).
Doran, Kamie, 30 November 2004. New York's HIV Experiment. BBC. Available at: http://news.bbc.co.uk/2/hi/programmes/this_world/4038375.stm (accessed 21.12.13.).
Ezeome, Emanuel R., Simon, Christian, 2010. Ethical problems in conducting research in acute epidemics: the Pfizer meningitis study in Nigeria as an illustration. Developing World Bioethics 10 (1), 1–10.
Farfel, Mark R., Chisholm, J. Julian, 1990. Health and environmental outcomes of traditional and modified practices for abatement of residential lead-based paint. American Journal of Public Health 80 (10), 1240–1245.
Fox, Renée C., Swazey, Judith P., 2008. Observing Bioethics. Oxford University Press, Oxford.
Grimes v. Kennedy Krieger Institute, 782 A.2d 807 (Maryland 2001).
Guinea Pig Kids, November, 30 2004. [TV documentary] BBC. Available at: http://www.guineapigkids.com/ (accessed 21.12.13.).
Harris, Sheldon H., 2003. Japanese biomedical experimentation during the World-War-II era. In: Beam, Thomas E., Sparacino, Linette R. (Eds.), Military medical ethics, vol. 2. Office of the Surgeon General at TMM Publications, Washington, DC.
Hassner Sharav, V., n.d. Human Experiments: a Chronology of Human Research. Available at: http://www.ahrp.org/history/chronology.php (accessed 08.04.14.).
Heller, Jean, 26 July 1974. Syphilis victims in U.S. study went untreated for 40 years. New York Times, 1.
Hornblum, Allen M., 1998. Acres of Skin. Routledge, New York.
In re A.C., 573 A.2d 1235 (D.C. App. 1990).
In re Quinlan, 355 A.2d 647 (N.J. 1976).
Jonsen, Albert R., 1998. The Birth of Bioethics. Oxford University Press, Oxford.

Kaler, Amy, 1998. A threat to the nation and a threat to the men: the banning of Depo-Provera in Zimbabwe, 1981. Journal of Southern African Studies 24 (2), 347–376.

Katz, Adolph, 27 February 1966. Prisoners volunteer to save lives. Philadelphia Bulletin.

Katz, Jay, 1972. Experimentation with Human Beings: The Authority of the Investigator, Subject, Professions, and State in the Human Experimentation Process. Russell Sage Foundation, New York.

Kristof, Nicholas D., 17 March 1995. Unmasking horror—a special report. Japan confronting gruesome war atrocity. New York Times. Available at: http://www.nytimes.com/1995/03/17/world/unmasking-horror-a-special-report-japan-confronting-gruesome-war-atrocity.html (accessed 08.04.14.).

Kuhn McGregor, D., 1998. From Midwives to Medicine: The Birth of American Gynecology. Rutgers University Press, New Brunswick.

Matter of Baby K,16 F.3d 590 (4th Cir. 1994).

McNeil, Donald G., 11 August 2011. Nigerians Receive First Payments for Children Who Died in 1996 Meningitis Drug Trial. New York Times. Available at: http://www.nytimes.com/2011/08/12/world/africa/12nigeria.html?_r=0 (accessed 10.04.14.).

Moss, William, Eckhardt, Roger, 1995. The human plutonium injection experiments. Los Alamos Science 23, 177–233.

Morse, Howard N., 1967. Legal implications of clinical investigation. William and Mary Law Review 8 (3), 359–386.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Available at: http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html (accessed 21.12.13.).

National Research Act, Public Law 1974, No. 93-348, 88 Stat. 342.

1949. The Nuremberg Code: Trials of War Criminals before the Nuremberg Military Tribunals under Control Council Law, vol. 2 (10). Available at: http://history.nih.gov/research/downloads/nuremberg.pdf (accessed 21.12.13.).

Office for Human Research Protections, 1993. Special classes of subjects. In: IRB Guidebook (Chapter 6). Available at: http://www.hhs.gov/ohrp/archive/irb/irb_chapter6ii.htm#g6 (accessed 21.12.13.).

Pellegrino, Edmund D., 1999. The origins and evolution of bioethics: some personal reflections. Kennedy Institute of Ethics Journal 9 (1), 73–88.

Potter, Van Rensselaer, 1971. Reflections: bioethics. BioScience 21 (21), 1088.

Potter, Van Rensselaer, 1975. Humility with responsibility—a bioethics for oncologists: presidential address. Cancer Research 35, 2297–2306.

Potts, Malcolm, Paxman, John M., 1984. Depo-Provera—ethical issues in testing and distribution. Journal of Medical Ethics 10 (1), 9–20.

Project MKULTRA, 1977. The CIA's program of research in behavioral modification: Joint hearing before the select committee on intelligence and the subcommittee on health and scientific research of the committee on human resources, 95th Cong. 1. Available at: http://www.nytimes.com/packages/pdf/national/13inmate_ProjectMKULTRA.pdf (accessed 21.12.13.).

Protection of Human Research Subjects and Creation of National Bioethics Advisory Committee, 3 October 1995. 66 Fed. Reg. 13237.

Quill, Timothy E., 2005. Terry Schiavo—a tragedy compounded. New England Journal of Medicine 352 (16), 1630–1633.

Rebecca Project for Human Rights, 2013. Depo-Provera: Deadly Reproductive Violence against Women. Available at: http://www.1037thebeat.com/wp-content/uploads/2013/06/DEPO-PROVERA-DEADLY-REPRODUCTIVE-VIOLENCE-Rebecca-Project-for-Human-Rights-June-2013-2.pdf (accessed 10.04.14).

Reich, Warren Thomas, 1994. The word "bioethics": its birth and the legacies of those who shaped it. Kennedy Institute of Ethics Journal 4 (4), 319–345.

Reidel, S., 2005. Edward Jenner and the history of smallpox vaccination. BUMC Proceedings 18 (1), 21–25.

Robinson, Walter M., Unruh, Brandon T., 2008. The hepatitis experiments at the Willowbrook state school. In: Emanuel, Ezekial J., et al. (Eds.), The Oxford Textbook for Clinical Research Ethics. Oxford University Press, Oxford (Chapter 7).

Rothman, David J., 1991. Strangers at the Bedside. Basic Books, New York.

Sambisa, William, 1996. Contraceptive Use Dynamics in Zimbabwe: Discontinuation, Switching, and Failure. Macro International Inc, Calverton, Maryland. Available at: http://dhsprogram.com/pubs/pdf/FA15/FA15.pdf (accessed 10.04.14).

Smith, Cedric M., 2005. Origin and uses of primum non nocere—above all, do no harm! Journal of Clinical Pharmacology 45 (4), 371–377.

Smith, David, 11 August 2011. Pfizer Pays Out to Nigerian Families of Meningitis Drug Trial Victims. The Guardian. Available at: http://www.theguardian.com/world/2011/aug/11/pfizer-nigeria-meningitis-drug-compensation (accessed 10.04.14.).

Stein, R., 1 October 2010. U.S. Apologizes for Newly Revealed Syphilis Experiments Done in Guatemala. Washington Post. Available at: http://www.washingtonpost.com/wp-dyn/content/article/2010/10/01/AR2010100104457.html (accessed 08.04.14.).

Stephens, Joe, 2000. Where Profits and Lives Hang in Balance. Washington Post, 17 December 2011. Available at: http://www.washingtonpost.com/wp-dyn/content/story/2008/10/01/ST2008100101390.html (accessed 10.04.14).

SUNY Cortland Institutional Review Board, n.d. Historical Perspectives and Their Significant Impact on the Development of Ethical Standards Governing Research using Human Participants.

Targeted Individuals Europe, 2011. Unit 731, Japanese Human Medical Experiments during the WW2. Available at: http://www.targeted-individuals-europe.com/?q=node/7 (accessed 08.04.14.).

Treaser, Joseph B., 1976. C.I.A.'s Files on LSD Death Found to Be Contradictory. New York Times, 1 January 1994. Available at: http://www.nytimes.com/1976/01/11/magazine/7601110LSON.html?_r=0 (accessed 21.12.13.).

Van Zyl, Mikki, de Gruchy, Jeanelle, Lapinsky, Shiela, Lewin, Simon, Reid, Graeme, 1999. The Aversion Project: Human Rights Abuses of Gays and Lesbians in the South African Defence Force by Health Workers during the Apartheid Era. Simply Said & Done, Cape Town, South Africa. Available at: http://www.mrc.ac.za/healthsystems/aversion.pdf (accessed 10.04.14.).

Washington, Harriet, 2006. Medical Apartheid: The Dark History of Medical Experimentation on Black Americans from Colonial Times to the Present. Doubleday, New York.

Washington v. Glucksberg, 521 U.S. 702 (1997).

Wilson, Robin Fretwell, 2009. Estate of *Gelsinger v. Trustees of University of Pennsylvania*: money, prestige, and conflicts of interest in human subjects research. In: Johnson, Sandra H., Krause, Joan H., Saver, Richard S., Wilson, Robin Fretwell (Eds.), Health Law and Bioethics: Cases in Context (Chapter 11).

World Medical Association, last amended 2008. Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. Available at: http://www.wma.net/en/30publications/10policies/b3/17c.pdf (accessed 21.12.13.).

Law: Change and Evolution

Lawrence M Friedman, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article focuses first on macroevolutionary change, looking at theories of long-term legal change involving whole systems, groups of systems, or law in general. Theorists discussed include Marx, Maine, Durkheim, Weber, Schwartz and Miller, Galanter, Selznick, and Watson. The second part is concerned with more general theories of legal change. It discusses the 'autonomy' of the legal system, typologies of legal change, the need to distinguish long-run change from short-run change, the differences made by democratic and totalitarian tendencies in society, and the role played by the legal system in social change.

Legal Evolution

There has been a good deal of speculation on the question whether legal systems 'evolve', that is, whether there is some kind of natural, automatic process or direction in which legal systems move and change over time. We might want to distinguish between macroevolution and microevolution. By macroevolution, I mean a theory that law in general, or whole systems or groups of systems, evolves in patterned ways. Microevolutionary theories would try to account for short-term changes in particular systems. Of course, there might be many gradations in between. Take, for example, the theory that common law doctrines evolve over time toward economic efficiency or cost reduction (Priest, 1977; Clark, 1981); the theory assumes a process, somewhat like Darwinian natural selection, that winnows out wasteful or irrational rules (for a discussion, see Elliott, 1985, pp. 62–71). A theory of this sort blurs the boundary between macro- and microevolution. This article will focus first on macroevolutionary change; then, briefly, it will consider more general theories of legal change.

A (macro) evolutionary theory of law would mean that legal systems grow or develop in a pattern of orderly stages. Legal systems would progress, in other words, from stage 1, then move on to stage 2, then 3, and so on. Few, if any, legal scholars believe that any such progression is inevitable. Practically speaking, historians and legal scholars think of legal evolution in a weaker sense: not all systems will reach the highest stage (whatever that might be), but if they do, they will pass through the same sorts of stages (Friedman, 1975).

The whole concept of evolution is, in a sense, distinctly modern, for the simple reason that traditional societies tended to regard their legal systems as essentially immutable. Law was divine, magical, or the product of natural reason, hence basically timeless and beyond human power. Change, if it occurred, might be (for example) some sort of dramatic and divine intervention. On the other hand, insofar as evolution implies 'progress,' it is as unfashionable an idea in contemporary times as is the idea of progress in general.

Any theory of legal evolution has to posit some source of evolutionary change, that is, some force or factor that impels legal systems to move along given paths. What might that source be? It is hard to generalize, but it seems intuitively right to suggest that changes in society drag legal systems along with them, so that stages of legal development will correspond to stages of social development. Most of the theorists discussed in this article fall into this category. In contrast, some people feel that law is, in essence, independent of society, although it may contain within itself some principle of growth or evolution, which unfolds gradually over time.

Such an idea is at least implicit in legal formalism, of both the common law and civil law varieties; development either occurs automatically or depends on the craftsmanship of judges or jurists; in either event, the law unfolds like the flower from the bud. Among modern authors, Alan Watson perhaps illustrates one version of this view: law, for Watson, is "largely autonomous and not shaped by societal needs"; legal institutions, to be sure, do not exist "without corresponding social institutions," but law itself "evolves from the legal tradition." Law is the "culture of the lawyers and … the lawmakers." (Watson, 2001, pp. 263, 264). Law might just possibly possess no special path; it might be changing in no special direction. A sophisticated – and plausible – version of this idea appears in Hutchinson (2005, p. 16); however, with regard to the common law, Hutchinson does state that its development was basically 'contingent upon social and historical circumstances,' and that 'any understanding of law must be deeply political,' that is, explained in terms of a particular time and a particular culture.

Many classical works in legal sociology and allied fields have at least touched on the subject of legal evolution, and most of them fall squarely within the first type mentioned here (Raiser, 1995, pp. 318–334; Elliott, 1985). Marx expressed an evolutionary theory, after all, in his views of the relationship between ownership of the means of production and the organization of society, and this theory had definite implications for the legal order. Sir Henry Maine, in his important book *Ancient Law* (1861), expressed the view that there was a definite pattern in the way legal systems grew and changed. Ancient law began in a patriarchal stage, in which the family, not the individual, was the unit of legal analysis; moreover, the father had basically the key rights and authority within the family. Birth order, gender, and age determined one's legal rights and duties. Society was based on status. The movement of 'progressive societies,' he stated, has gone in a 'uniform' direction: the individual replaces the

family; voluntary arrangements replace ascriptive ones; and legal systems, to use his famous phrase, evolved 'from Status to Contract.'

Implicit evolutionary theories also appeared in other classical theories of law and society. Durkheim, for example, drew a distinction between primitive and modern societies. In *The Division of Labor in Society*, Durkheim distinguished between 'mechanical' and 'organic' solidarity. In traditional societies (societies of 'mechanical' solidarity), rules and norms were held in common. Violations threatened social cohesion and had to be punished. The law was therefore mostly concerned with punishments. On the other hand, in modern society (characterized by 'organic' solidarity), there is a tremendous amount of complexity and interdependence, and legal rules are mostly civil and 'institutive.' In Durkheim's scheme, as in Maine's, there is strong emphasis on contract: societies as they enter the modern world are characterized more and more by voluntary agreements.

The most subtle, and influential, of the classic writings on legal evolution is Max Weber's (Rheinstein, 1954). Weber was interested in the question 'What makes modern law different from older and traditional systems?' The distinctive quality he saw was what he called rationality – essentially, the process of making and interpreting law on the basis of general principles and rules. Oracles, ordeals, lawmaking by charismatic leaders – these are 'irrational' forms of making or finding law. Modern legal systems in developed countries tend (according to Weber) to replace the irrational with the rational. In particular, modern legal systems are characterized by formal rationality, a 'peculiarly professional, legalistic, and abstract approach to law.' A formally rational system uses logic and deductive reasoning to derive results from rules, concepts, and principles. Irrational systems do not produce predictable results – results that can be controlled by the intellect; citizens therefore cannot rely on these processes. Hence modern law, the law of the modern capitalist state, has a natural preference for Weberian rationality.

Weber falls short of arguing that this progression, from the irrational to the rational, is inevitable or predetermined. His discussion of legal history cannot be summed up in this single formula, since Weber describes, in rich detail, at least four stages of development (Sterling and Moore, 1987). Charismatic legal revelation is succeeded by formally irrational 'kadi' justice – ad hoc decision making based on general social norms. Then comes a stage of substantive rationality; in this stage, a systematic body of rules is imposed on society, but less in terms of juristic reasoning than in terms of religious, ethical, or economic considerations. The stage of formal rationality comes last; in this stage, the jurists create a more or less self-contained, highly logical, and systematic body of law, in which abstract legal concepts are arranged in some definite and principled order. But this is merely a more or less idealized picture of the growth of law in the West. Weber does not claim that this is bound to happen elsewhere, and he never quite argues that the movement he describes is irreversible. In particular, modern law has tended to drift away from formal rationality toward what Weber called substantive rationality – law is made or found in accordance with principles and policies (economic, social, and cultural) that are not strictly legal.

Richard Schwartz and James C. Miller (1964; see Wimberley, 1973) tried to evaluate the concept of evolution on a systematic basis. They drew material from 51 different societies and tried to 'scale' the development of legal institutions. Their scales were in terms of legal institutions: the least advanced societies had no police, lawyers, or forms of mediation. Societies develop mediation first, then some form of policing, and a legal profession last of all. Most of the societies that Schwartz and Miller discussed were preliterate.

In a sense, Weber and Maine (although they wrote long before Schwartz and Miller) pick up where Schwartz and Miller leave off, since their main concern was with advanced societies. Galanter (1966), in an important essay, tried to define what makes a legal system 'modern,' which was, of course, Weber's chief interest as well. He suggests a list of 13 traits, which, in his view, set modern law apart from premodern law. The scheme, then, is implicitly evolutionary. For Galanter, modern law is bureaucratic, professional, rational, and universalistic; it is also 'amendable,' that is, it lacks 'sacred fixity.' Friedman (1994) returns to this theme and lists six traits that, in his view, together make up 'modern legal culture.' Modern law is committed to change, reform, and adaptation; it lacks the conception of timelessness of traditional legal systems. It is pervasive, ubiquitous, and frankly instrumental. At the same time, it does contain strong notions of fundamental rights. Rehbinder (2000) also lists a group of 'tendencies' that characterize contemporary law: a strain toward universalism (this was also one of Galanter's traits), but at the same time specialization, bureaucratization, and an increase in the sheer amount of legal material.

Most of these discussions of legal change focus on 'modernity,' and they are descriptive and explanatory. So, for example, Struck (2011) describes a progression from the 'archaic,' through the 'feudal,' to modern capitalism. But in some work, there is a strong normative strain. Selznick (1969), for example, was concerned, above all, with the evolution of 'legality.' In a later work, Philippe Nonet and Selznick discuss a tendency for laws to evolve toward 'responsiveness' (Nonet and Selznick, 1978). These ideas have influenced the work of Gunther Teubner (1983), who described legal systems as evolving from a kind of dry, rule-bound bureaucratic system to a more 'reflexive' one, in which the law, particularly within administrative systems, would be responsive to the needs and wants of the governed, and which would develop rules and procedures that would be just and adaptive. Teubner, like Nonet and Selznick, adds a distinctively normative flavor to the more descriptive tone of Weber. Teubner's views touched off a lively debate (see Blankenburg, 1984); some of the empirical issues raised in this line of work remain to be definitively examined.

Legal Change

The theory of legal change has to be dealt with rather briefly, not because it is less important than the theory of legal evolution, but precisely because legal change is pervasive, and a broad theory of legal change would be nothing more or less than a general theory of how and why the law develops, in both

the long and the short run. Moreover, practically any study in the field of sociolegal history is in fact a study of legal change and of the dynamics of legal change. (For example, studies of lobbying and interest groups are studies of the dynamics of legal change. So, too, is research on the impact of the industrial revolution, the sexual revolution, and every social factor imaginable.)

Still, constant change is one of the most salient aspects of modern legal systems. At every session, legislatures pass laws to amend or modify existing laws or even to embark on a new direction. The same is true of administrative agencies, city governments, and much of the work of courts: they all represent change. Indeed, the public expects change: candidates and parties, in modern societies, feel obliged to present a program, an agenda – in short, they promise change. (Dictatorships, of course, are under no such obligation.) All legal systems are dynamic; and it is part of the legal culture of today to expect them to *be* dynamic.

All modern legal systems, then, accept change as normal. This was not always the case. Many older legal systems (and preliterate systems) consider law as fundamentally immutable or, in any event, beyond the power of mere mortals. Change occurred nonetheless, but often in disguised form – through resort, for example, to legal fictions.

Theories of legal change – theories about what produces legal change – depend, in the first instance, on the answer to larger questions, very notably the question of the 'autonomy' of the legal system. A legal system, if it was fully autonomous, would operate through its own inner logic or program. It would be independent of the outside society. The opposite system would, on the contrary, faithfully follow, reflect, or mirror changes in the outside society. Needless to say, most scholars are not to be found at either pole of this continuum, but somewhere in between. Most orthodox legal scholars cluster closer to the autonomous end of the continuum; social scientists who study law cluster closer to the other end.

In discussing legal change, there are obviously many factors to be taken into account (for one general discussion, see Ferrari, 1997, Chap. 6). Not every legal system will change in the same way. This is true for both internal and external reasons. Societies undergoing rapid change will naturally process more legal changes than relatively static societies. Legal systems, particularly advanced ones, that legitimate change will contain rules, procedures, and institutions for making and processing change. Modern parliamentary democracies emphasize the formal role of legislatures in changing the law, although administrative agencies also have important rule-making power. Also, the role of courts as change agents is important in many if not most societies. Change also originates in customs, practices, and norms, which then influence the change-making institutions.

A major event (war, plague, or revolution), a major innovation (antibiotics, automobiles, or computers), or a major social change (democratization or sexual permissiveness) will inevitably lead to legal change, sometimes on a massive scale; but the exact form of the change is never precisely predictable. The 'outside' event or situation has its impact in a society that already has structures and cultural facts; those structures and cultures will bend and mediate the influences of the 'outside' forces prismatically. The 'outside' forces make an impact through effecting changes in the general legal culture – that is, in the pattern of demands, expectations, and values that concern the legal system – and legal change comes about because of this change in legal culture; it does not flow immediately and automatically from the 'outside' event or situation. Moreover, once a legal structure, doctrine, or institution is in place, it can exert its own force on the legal culture; it affects legal ideology, and it may come to seem natural or inevitable. The process of legal change, then, is dynamic and interactive, and it forms a never-ending chain of causes and effects.

In theory, one can distinguish four types of legal change, depending on the change's point of origin and final point of impact (Friedman, 1975): (1) some change originates inside the legal system, and also causes whatever impact it might have within that system; (2) some change begins inside the system, but in fact does have an impact on society; (3) some change originates outside the legal system, but affects only the legal system and ends there, with no further impact on society; and (4) some change originates outside the legal system and ultimately has some impact on society, that is, it is more or less effective.

The first two are examples of technical change, and include many projects of 'law reform' or codification, within legal systems – the technocratic work of jurists, which has little or no effect on the way that society actually functions. Codification often falls into this category. But sometimes, wittingly or unwittingly, what begins as purely technical change has unforeseen consequences in the larger society. That is, the technical change simply gives some group in society a convenient 'hook' or formal excuse to advance its goals.

A strong 'outside' force or movement should, in theory, produce an impact both on the law and on the society. But often some other social force, movement, or group fights hard to resist change, and the result is a kind of stalemate: a 'law' is passed, for example, after a bitter struggle between interest groups, but the 'losing' group has enough power to prevent effective enforcement.

The fourth type is, in many ways, the most interesting and pervasive. Some changes are quite revolutionary – for example, the type of wholesale change that took place after the Russian or the Chinese revolutions; or, contrariwise, that took place in Russia after the collapse of the Soviet Union; or the total extinction of the legal system of East Germany (Markovits, 1995). Other changes are small or incremental; they produce only a minor effect on some portion of the population.

The discussion in this article has mostly assumed change in formal, overt aspects of law. But legal systems are complex, and there are many instances where one aspect of the system changes while others remain (on the surface) unchanged. For example, the formal or official law might remain unchanged, but the operating system changes, perhaps quite drastically. One example would be divorce law and practice in the United States in the late nineteenth and early twentieth centuries. The official law of some states (e.g., New York) made divorce difficult and expensive; the law in action was completely different. Collusion and other devices made it possible for citizens of New York to get divorces that in theory they were not entitled to have. What kept this dual system alive? Apparently, it was a political situation that for various reasons

was deadlocked. Strong forces of 'morality' prevented the enactment of reform. The growing demand for divorce led reformers to seek other channels, which they did quite successfully. Legal systems very commonly change in this ragged and complex manner (Friedman, 2004; Grossman and Friedman, 2011).

One should also distinguish between long-run change and short-run change (and various gradations in between). In the long run, legal systems are almost infinitely plastic, and they change to fit the shape of the society in which they are embedded. A feudal system will generate a legal system that deals with institutions of feudal society and reflects its norms and structures; a capitalist, socialist, or tribal society will shift its laws to suit the existing system. But in the very short run, random, accidental, or 'internalist' factors may condition or shape forms of legal change. It is this fact in modern societies that promotes the illusion of legal autonomy – the image of a system with a tough and impermeable outer skin. The legal profession itself has to be reckoned with; this can be a powerful group, with a vested interest in resisting certain kinds of change.

The rate and forms of change also depend on whether the society is relatively open and democratic, or relatively closed and aristocratic or totalitarian. In an open society, 'public opinion' is an important factor in shaping the law. 'Public force' would be a more accurate way of putting the matter. 'Public opinion' as such – whatever it is that polls and surveys measure – does not bring about change, except insofar as it influences legislators or judges, or is translated into public action. Public opinion, in any event, is not simply what a poll suggests, or what the majority of the public thinks and feels. *Salience* is an important factor; it affects whether attitudes are translated into behavior. Moreover, in every society, no matter how 'democratic,' there are vast gaps between high and low, rich and poor, powerful and powerless. A wealthy corporation simply counts for more than, say, an impoverished and uneducated slum dweller. As an axiom of legal change, one can start with the notion that a legal rule, doctrine, practice, or institution will change if the social force exerted against it is greater than the force generated to preserve it.

This is a very abstract, and perhaps not very helpful, formulation. The interesting questions are more specific: in any given situation, what *are* the social forces that are in play, and what lies behind the exertion of social force? Moreover, it can be assumed that at any given time, the actual social force exerted on an issue is much less than the potential social force, since most people and institutions, even the powerful ones, most of the time, pay no attention to most legal issues and problems within a society. For this reason, scandals, incidents, and 'horror stories' are particularly important sources of legal change in open societies: they can arouse the sleeping giant and turn potential force into actual exerted force. Hence, the role of the media – which communicates these scandals, incidents, and 'horror stories' – is a particularly powerful, but insufficiently studied, catalyst for sociolegal change (see, for an excellent example, Haltom and McCann, 2004).

Particularly in common law systems, the role of courts in bringing about social change has been much discussed. To what extent, for example, did *Brown vs. Board of Education* (1954), the school desegregation case, make an actual impact on race relations in the United States? (For one view, see Rosenberg, 2008.) That the courts are an important institution in American society and play some role in bringing about social change cannot be denied; the question is how much. There is a small literature on the impact of judicial decisions, but it is hard to come to general conclusions; no two cases seem alike. Modern courts in many countries have the power of judicial review, which may have increased their potential for nudging society along certain paths. There are some signs of dramatic change in judicial culture in the Western world in general. Crusading judges in Italy led to the collapse of the party system; and an activist Spanish judge issued an order directed against General Augusto Pinochet, the former dictator of Chile, who had been virtually immune from prosecution before.

One factor that influences impact is the way that a legal decision and innovations in general are communicated to the relevant audience. We already considered the role of scandals and incidents, and the media's ability to dispense information in selective ways. One should also distinguish between immediate changes (e.g., on the litigants or on some particular group) and more indirect, long-term consequences. Many legal commands have effects that ripple out into society. Unfortunately, research on all these issues is fairly thin.

Studies of legal change are usually studies within one particular legal system; there are, however, some interesting studies that look at *comparative* legal change. More and more, legal systems are not isolated, since societies are themselves more and more interdependent. In a globalized world, legal systems arguably are converging; since the systems face common problems, they tend to generate common solutions (see Friedman, 2011). Not everyone agrees on this, however, and some scholars have argued that global trends have very different impacts on different societies (see, e.g., Engel and Engel, 2010, on Thailand). This is undoubtedly true, but the general lines of development seem reasonably clear.

See also: Durkheim, Émile (1858–1917); Marx, Karl (1818–83); Weber, Max (1864–1920).

Bibliography

Blankenburg, E., 1984. The poverty of evolutionism: a critique of Teubner's case for 'reflexive law'. Law and Society Review 18, 273.

Clark, Robert C., 1981. The interdisciplinary study of legal evolution. Yale Law Journal 90, 1238.

Elliott, E.D., 1985. The evolutionary tradition in jurisprudence. Columbia Law Review 85, 38.

Engel, D.M., Engel, J., 2010. Tort, Custom, and Karma: Globalization and Legal Consciousness in Thailand. Stanford University Press, Stanford, CA.

Ferrari, V., 1997. Lineamenti di Sociologia del Diritto: I. Azione Giuridica E Sistema Normative. Lalerza, Rome.

Friedman, L.M., 1975. The Legal System: A Social Science Perspective. Russell Sage Foundation, New York.

Friedman, L.M., 1994. Is there a modern legal culture? Ratio Juris 7, 117.

Friedman, L.M., 2004. Private Lives: Families, Individuals, and the Law. Harvard University Press, Cambridge, MA.

Friedman, L.M., 2011. The Human Rights Culture: A Study in History and Context. Quid Pro Press, New Orleans.
Galanter, M., 1966. The modernization of law. In: Weiner, M. (Ed.), Modernization. Basic Books, New York, p. 153.
Grossman, Joanna L., Friedman, L.M., 2011. Inside the Castle: Law and the Family in 20th Century America. Princeton University Press, Princeton, NJ.
Haltom, W., McCann, M., 2004. Distorting the Law: Politics, Media, and the Litigation Crisis. University of Chicago Press, Chicago, IL.
Hutchinson, A.C., 2005. Evolution and the Common Law. Cambridge University Press, Cambridge, MA.
Maine, H., 1861. Ancient Law. J. Murray, London.
Markovits, I., 1995. Imperfect Justice: An East-West German Diary. Oxford University Press, Oxford, NY.
Nonet, P., Selznick, P., 1978. Law and Society in Transition: Toward Responsive Law. Octagon Books, New York.
Priest, George L., 1977. The common law process and the selection of efficient rules. Journal of Legal Studies 6, 65.
Raiser, T., 1995. Das Lebende Recht: Rechtssoziologie in Deutsch-land. Nomos, Baden-Baden.
Rehbinder, M., 2000. Rechtssaziologie, fourth ed. Walter de Gruyter, Berlin.
Rheinstein, M. (Ed.), 1954. Max Weber on Law in Economy and Society. Harvard University Press, Cambridge, MA.
Rosenberg, G.N., 2008. The Hollow Hope: Can Courts Bring about Social Change?, second ed. University of Chicago Press, Chicago, IL.
Schwartz, R., Miller, J.C., 1964. Legal evolution and societal complexity. American Journal of Sociology 70, 159.
Selznick, P., 1969. Law, Society, and Industrial Justice.
Sterling, J.C., Moore, W.E., 1987. Weber's analysis of legal rationalization: a critique and constructive modification. Sociological Forum 2, 67.
Struck, G., 2011. Rechtssoziologie: Grundlagen und Strukturen. Nomos, Baden-Baden.
Teubner, G., 1983. Substantive and reflexive elements in modern law. Law and Society Review 17, 239.
Watson, A., 2001. The Evolution of Western Private Law. Johns Hopkins University Press, Baltimore, MD.
Wimberley, H., 1973. Legal evolution: one further step. American Journal of Sociology 79, 78.

Law: Economics of Its Public Enforcement

Alexander Volokh, Emory University School of Law, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A.M Polinsky, S. Shavell, volume 12, pp. 8510–8517, © 2001, Elsevier Ltd.

Abstract

This article summarizes the literature on the economics of public enforcement of law. First, it discusses the positive aspects of how criminals behave, both from a theoretical and empirical point of view, also touching on how enforcers behave. Next, it discusses the normative theory of how a public law enforcement should be designed within the dominant welfarist law and economics framework.

Introduction: From Beccaria to Becker

The economic theory of the public enforcement of law – a category that chiefly includes criminal law, though it is broad enough to include any law enforced by public authorities, for instance some regulatory law – has its *modern* genesis in the work of economist and Nobel laureate Gary Becker. But several of Becker's ideas were foreshadowed by earlier writers. "Lest the reader be repelled by the apparent novelty of an 'economic' framework for illegal behavior," Becker wrote in 1968, "let him recall that two important contributors to criminology during the eighteenth and nineteenth centuries, Beccaria and Bentham, explicitly applied an economic calculus."

Cesare Beccaria's *On Crimes and Punishments* (1986 [1764]), while by no means a 'law and economics' document, does show, in incipient form, some prominent ideas that would later be associated with the law and economics approach to public law enforcement. This makes sense, as Beccaria was influenced by late-eighteenth-century utilitarianism, which is the conceptual foundation for modern law and economics. Beccaria wrote that "pleasure and pain are the motives for action among sentient beings" (p. 15) and that "the true measure of crimes" is "the harm done to society" (p. 17). He wrote that deterrence is the proper goal of punishment (p. 23), that the harm of the punishment need only exceed the benefit to the criminal for deterrence to work (p. 46), and that inflicting any harm on the criminal beyond this minimum is undesirable (p. 23). Thus, he advocated that nonviolent crimes should be punished only by fines, provided the offender is able to pay (pp. 39–40). Again on deterrence grounds, he advocated that punishments be prompt (p. 36) and certain (p. 46). He advocated a gradation of penalties according to the severity of the crime, on marginal deterrence grounds (otherwise there would be no disincentive to commit more severe crimes) (p. 16), and in particular advocated lower penalties for attempts (p. 69). In fact, one of the reasons he advocated certainty of punishment was that this would allow punishments to be lower, which would make such a sliding scale more effective, as there are limits both to the endurance of the convicted defendant and to the ability of excessively cruel penalties to imprint themselves on the minds of the population to be deterred (p. 47).

Jeremy Bentham, as a founding father of utilitarianism, had a great deal more to say about punishment. Consider a sampling of his views from *An Introduction to the Principles of Morals and Legislation* (1907 [1789]: pp. 171–177). First, punishment might be "groundless" because "the act [is] not … mischievous upon the whole", for instance where there is consent, where "there is a certainty of an adequate compensation," or where the act produced a benefit that outweighed the harm. Second, punishment might be "inefficacious," as, for instance, where the penalty wasn't established or conveyed to the actor at the time of the act, where the actor is undeterrable (as when the actor is too young, insane, or intoxicated), or where the act was accidental. (Here we see clearly the potential conflict between a retributive view and the deterrence-focused view that dominates law and economics.) Third, punishment might be "unprofitable," as when the benefit of deterrence is outweighed by the pain of the prospective criminal at not being able to commit the act, the actual criminal's fear of being caught, the convicted defendant's pain in being punished, or the "pain of sympathy" caused to others. (Here we see the standard utilitarian view, controversial to nonutilitarians, that all people's happiness should be weighed equally, even the criminal's.) Fourth, punishment might be "needless," as where equivalent deterrence could be achieved through education or moral suasion. (Elsewhere (Bentham, 1830: p. 29), he writes that hanging a person "would be folly or cruelty" if hanging an effigy would produce "the same salutary impression of terror upon the minds of the people.")

For whatever reason, the development of a utilitarian theory of public enforcement of law was not particularly active for over a century after Beccaria and Bentham. And thus the world would have to wait until Becker's seminal 1968 article, *Crime and Punishment: An Economic Approach*, "a resurrection, modernization, and thereby", Becker hoped, "an improvement on these much earlier pioneering studies." Becker's approach is primarily normative, but he first develops a positive theory of how criminals behave – or perhaps one should say "how *people* behave," as a core component of his model is that criminals are fundamentally no different from anyone else. Everyone offends when the net benefits of offending exceed the net benefits of obeying the law. Moreover, everyone's utility is counted equally as part of social welfare, so the optimal approach is to set the particular structure of rewards and penalties (in particular, the level of fines and the probability of detection) to induce that level of offending that maximizes the difference between benefits and harms. Because

we take into account as a 'harm' the cost of apprehension and conviction, some underdeterrence is optimal; and because the social loss inherent in imprisonment is also a sort of social harm, it is optimal to use fines, which are virtually costless from a social point of view (as they are merely a reshuffling from one person's pocket to another's) whenever possible. Moreover, fines should relate to the marginal harm from the act and not to the wealth of the offender. On Becker's view, the optimal criminal law would essentially be a species of publicly enforced tort law. Space constraints prevent a fuller exposition of the ideas in Becker's article, which is worth multiple readings by those who are comfortable with differential calculus.

The remainder of this article discusses advances in the law and economic approach to the public enforcement of law since Becker. For more detailed surveys, the reader is referred to Eide et al. (2006), Polinsky and Shavell (2007), and Levitt and Miles (2007), to which this article is deeply indebted.

Law as an Incentive System: Positive Analysis

The previous section has summarized Becker's views on human behavior; the law and economics literature, by and large, has followed the same basic approach. The economic approach holds that everyone (except, perhaps, those with extremely severe mental disabilities) responds, in some degree, to the incentives created by legal rewards and punishments. Therefore, as a matter of *positive* analysis, the challenge is to empirically determine the magnitude of the behavioral response to different types and levels of punishment and other legal rules. And as a matter of *normative* analysis, optimal social design is a matter of fine-tuning the legal rules with these empirical results in mind.

The Distinctiveness of the Economic Approach

How does the positive side of the economic approach differ, if at all, from competing approaches in vogue among criminologists? Levitt and Miles (2007: pp. 458–462) summarize three different branches of the criminological literature and note similarities and differences with the economic approach.

One branch of the criminological literature focuses on the biological factors contributing to crime (Moffitt et al., 2011), from mental disorders and hormonal imbalances to IQ. The economic perspective does not deny the relevance of these factors: these factors can explain the baseline level of crime that exists in the absence of whatever incentives might be created by the probability and severity of punishment. But economists tend not to focus on the determinants of the baseline, instead placing greater stress on the incentives that affect the individual choice to deviate from the baseline. (The modern trend toward behavioral economics can be seen as an attempt to combine psychology and economics and incorporate cognitive limitations and 'bounded rationality' into the model of individual choice. See McAdams and Ulen, 2009.)

A second branch of the criminological literature "stems from Émile Durkheim's concept of 'anomie,' or an absence of clear norms to direct human behavior" (see, e.g., Merton, 1968). The idea of anomie predicts that people turn to crime when they are unable to realize their aspirations through legal means. This is, broadly speaking, consistent with the economic theory in that people with limited legal opportunities may realize greater utility through crime; but the important factor for the economist is the *relative* utility of following or violating the law. The economist would not predict that someone with limited legal opportunities would turn to crime unless the illegal opportunities were more attractive.

A third branch of the criminological literature 'considers the social processes through which criminality is learned or culturally transmitted' (see, e.g., Sutherland, 1956 [1937]). 'Criminality is thus not the behavior of the physically abnormal, but an activity learned from social interactions with others.' This approach is consistent with the economic model, though the economic model still insists that these factors influence criminality only through the familiar pathway of individual choice between legal and illegal activities. Thus, factors influencing criminality could be that one has learned better criminal techniques from the criminals in one's social network, that one assesses the probability of punishment by observing the frequency of punishment among one's colleagues, or that one feels a lesser social stigma of offending because illegal conduct is so prevalent within one's circle (see, e.g., Sah, 1991; Bar-Gill and Harel, 2001).

The difference between the economic and criminological approaches is thus mainly one of emphasis, but the differences in emphasis are significant enough that the two approaches lead to highly distinct modes of analysis in practice.

Theoretical Refinements

Post-Becker authors have extended the model of criminal behavior. Ehrlich (1973) treated the problem as one of time allocation: each person chooses how much time to allocate between legal activities and illegal activities. Within the time-allocation literature, one can also model leisure as a third potential use of one's time (Heineke, 1978; Wolpin, 1978; Schmidt and Witte, 1984).

Criminals' risk preferences have also been the subject of study. Risk preferences come in three flavors: risk-neutral, risk-averse, and risk-seeking. A risk-neutral person is indifferent between getting $10 and a 10% chance of getting $100. A risk-averse person dislikes risk and therefore strictly prefers the $10, while a risk-seeking person has the opposite attitude and prefers the uncertain outcome.

Becker had already considered the conventional wisdom that certainty of punishment has greater deterrent value than severity of punishment. He rephrased it more quantitatively, stating that an increase in probability of detection, coupled with a proportionately equal decrease in the penalty, has an increased deterrent effect. But if this is so, it must be because the greater certainty of apprehension (keeping the expected value of punishment constant) is unattractive to criminals. In other words, given two probability-and-penalty regimes with the same average punishment, where one is more certain and has lower penalties while the other is riskier and has higher penalties, criminals are more deterred by the more certain regime and would rather be subject to the riskier regime. But this attitude toward risk is exactly what we have just defined as 'risk seeking.'

The time-allocation authors listed above extended Becker's treatment of risk preference. Ehrlich (1973), for instance, considered all three types of risk preference; Heineke (1978) concentrated on risk aversion. The precise form of risk preference is important; in most models, risk-averse people offend less when sanctions rise, while results are ambiguous for risk seekers.

Block and Heineke (1975) relax the assumption that the actor only cares about the monetary consequences of obeying or disobeying the law; in their model, dollars earned by legal or illegal means could have different effects on utility, and so one can imagine people with a 'preference for honesty'.

The more complicated the model, the fewer unambiguous predictions one can get from theory alone (though the finding that the probability of arrest has a deterrent effect seems to be fairly robust across models). Choosing among competing theories thus requires empirical testing.

Empirical Results

When testing these models empirically, several challenges emerge.

Levitt and Miles (2007: pp. 462, 463) summarize three challenges. "First, the available data correspond only roughly to the conceptual model." The behavioral theory is individual – but data are aggregated, incomplete, and potentially manipulated by police agencies.

The second problem is 'the identification of causal relationships.' Suppose the deterrence theory is true, and visible increases in policing have a strongly negative effect on offending. But policing levels are not chosen randomly: they are chosen in response to some felt need that is strong enough to overcome the costs involved. Jurisdictions could adopt heavy policing in response to already-high levels of crime, and those policing levels could reduce crime somewhat but not completely. A cross-sectional analysis would then reveal a *positive* relationship between crime and policing, and a naïve researcher would wrongly reject the deterrence hypothesis. Controlling for the factors that lead to high crime could solve the problem, but the econometrician typically knows less than the people he studies, so omitted variable bias will always be a problem. Could not the problem be alleviated by a time-series analysis, which would show the high crime levels being reduced somewhat by heavy policing? Perhaps, but policing levels can also be adjusted in anticipation of exogenous changes in crime levels, so even time-series analysis risks confusing the cause for the effect. More recent econometric analyses tend to be more sophisticated at teasing out true causal relationships, for instance by taking advantage of exogenous policy shocks, such as court-ordered prisoner releases or election-pandering increases in policing (see generally Angrist and Pischke, 2010).

Third, it is difficult to 'distinguish between incapacitation and deterrence.' (When law and economics scholars talk about 'deterrence,' they usually mean 'general deterrence,' the tendency of a penalty to reduce offending among the population at large; 'specific deterrence,' the tendency of a penalty to reduce offending by a specific punished individual, usually is not discussed, except to the extent that incapacitation is considered a type of specific deterrence.) The effects of deterrence are *ex ante* and work even if the penalty is never applied; indeed, one consequence of *perfect* deterrence would be that the penalty is *never* applied, even though officials are always standing by, willing to apply it if necessary. The effects of incapacitation, on the other hand, are *ex post* and only come into play to the extent that the penalty (generally imprisonment) *is* applied. Crucially, the rational choice model that is the staple of law and economics depends on deterrence playing at least some role, so merely finding a negative relationship between imprisonment and crime (which could be entirely due to incapacitation) is insufficient to confirm law and economics models.

Cognizant of these empirical challenges, law and economics scholars have studied the effects of a number of aspects of public law enforcement.

- Policing levels appear to have a large negative effect on crime rates (Marvell and Moody, 1996; Levitt, 1997; Corman and Mocan, 2000; Di Tella and Schargrodsky, 2004).
- Imprisonment has negative effects on crime rates (Marvell and Moody, 1994; Levitt, 1996); some of these effects are due to deterrence and some due to incapacitation (see, e.g., Kessler and Levitt, 1999; Levitt, 1998). Severe prison conditions also have a deterrent effect (Katz et al., 2003).
- The effect of capital punishment continues to be hotly debated, with some papers finding a deterrent effect (Mocan and Gittings, 2003; Dezhbakhsh et al., 2003; Shepherd, 2004) and others taking a more skeptical approach (Donohue and Wolfers, 2006). The estimated effect seems highly sensitive to the estimation method used.
- The effect of gun laws on crime is also hotly debated (compare Lott and Mustard, 1997; Lott, 2010 with Donohue and Levitt, 2001; Ayres and Donohue, 2003).

Of course, whether strategies like more policing or imprisonment are optimal depends not only on their deterrent effect but also on their cost, so the argument that prison deters criminals is consistent with the argument that imprisonment rates are too high because imprisonment is too costly.

The Behavior of Enforcers

The behavior of law-enforcement officials, whether police, prosecutors, or regulatory agencies, is harder to pin down. The field of public choice is devoted to exploring the motivations of government actors, but with the diversity of organizational forms in government and the multitude of competing incentives, valid generalizations seem elusive. For regulatory agencies, for instance, various scholars have proposed the model of the budget-maximizing bureaucrat (Niskanen, 1971), the slack-maximizing bureaucrat (Wyckoff, 1990), and the risk-avoiding bureaucrat (Peltzman, 1973). These different models seem to have obviously different implications for public enforcement of law – the budget-maximizing bureaucrat might invest in too much enforcement, while the slack-maximizing bureaucrat might invest in too little, and the risk-avoiding bureaucrat might avoid high-risk, high-reward law-enforcement strategies. No doubt each of these models might be correct in a particular area.

That said, here are a few suggestive empirical results about the behavior of police officers and prosecutors.

One current area of debate is whether racial profiling by police officers is due to racial bias or is statistically rational (compare Knowles et al., 2001 with Antonovics and Knight, 2009; Barnes, 2005). Another is the effect of court-imposed criminal procedure rules, such as the exclusionary rule or the required *Miranda* warnings (see Atkins and Rubin, 2003 for the exclusionary rule, and compare Cassell and Fowles, 1998 with Donohue, 1998; Feeney, 2000 for *Miranda*).

Prosecutorial behavior has not been heavily studied by law and economics scholars. One question of interest is what prosecutors maximize. Early on, Landes (1974) hypothesized that prosecutors maximize expected sentence length; more recently, Glaeser et al. (2000) have argued that federal prosecutors maximize a combination of social welfare (which leads them to take on the 'worst' cases from a cost-of-crime perspective) and personal career concerns (which lead them to take on difficult, high-profile cases that increase their visibility and improve their skills).

The public-choice literature has also examined the incentives of various government actors, for instance public corrections officer unions or private prison firms, to lobby for stricter criminal law; see Volokh (2008) for one view of the subject.

The Design of Public Enforcement: Normative Analysis

Welfarism: The Dominant Normative Framework

The dominant normative paradigm in law and economics is a form of welfarism. The only facts about the world that are held to matter are individual utilities, though individual utilities may be aggregated in any number of different ways. Different forms of aggregation may give greater or lesser scope to distributive justice considerations, but nothing apart from welfare, such as fairness or justice, should have an independent effect on policymaking (Kaplow and Shavell, 2002; see Volokh, 2011b for one dissenting position).

How should individual welfare be considered? The least controversial position within law and economics is the 'Pareto principle': any policy that improves *everyone's* utilities over the status quo (a so-called Pareto improvement) should be adopted. (Moreover, the dominant position in economics is radically subjectivist, in that the affected person is the only judge of his own utility; a policy improves someone's utility if and only if he would choose a world with the policy over a world without it.) Such policies are of course hard to come by, though if a policy creates winners and losers, one can make such a policy Pareto-improving by transferring some of the gains to the losers, provided that, roughly speaking, 'the winners gain more than the losers lose.'

But what about policies where (as is usual) there is no Pareto improvement because the necessary transfers are not made? The dominant view in law and economics holds that such a policy should still go forward as long as 'the winners gain more than the losers lose.' This concept can be formalized in one of two ways. One formulation urges the adoption of any policy under which, if the policy were adopted, the winners would gain enough to be able to compensate the losers (a so-called 'Kaldor improvement' Kaldor, 1939), though, of course, the hypothetical compensation need never be paid. A related formulation looks to whether the losers would be able to compensate the winners so as to convince them to *not* adopt the policy. If they could convince the winners not to go forward, the policy should not be adopted (again, the hypothetical compensation need never be paid); otherwise, the policy is a 'Hicks improvement' (Hicks, 1939) and should go forward. The Kaldor and Hicks formulations thus differ, though they are identical when the policies do nothing more than change people's monetary wealth.

Simple Models of Law Enforcement

With this theoretical normative apparatus in hand, and taking as given the positive Becker-based model where people offend when the net benefits from offending exceed the net benefits from obeying the law, we may start with a simple model of law enforcement, derive normative results, and gradually work our way toward more complicated models.

No Deterrence of Desirable Acts

First, suppose an act causes some harm but also brings some benefit to its perpetrator. Then, the act should be committed if and only if the benefit exceeds the harm. Thus, hypothetically, if the act is detected with a 100% probability, the fine should not exceed the level of harm. For instance, suppose an act causing \$100 worth of harm were penalized by a fine of \$105. This fine would be too high, because it would wrongly deter the act even if it produced benefits of \$104 and thus added \$4 to social welfare.

Strict versus Fault-Based Approaches

Next, consider the choice between two models of law enforcement: the 'strict' model, under which *any* harmful act is penalized regardless of the benefits it produces, and the 'fault-based' model, under which such acts are penalized only if they reduce social welfare, i.e., only if their benefits do not exceed their harms. In the strict model, the \$100 fine is always imposed, even if the act produced a \$200 benefit and was therefore desirable; in the fault-based model, the \$100 fine is imposed only if the benefit to the actor is less than \$100. Either of these methods produces the same result – acts are deterred if the benefit is less than \$100 and not otherwise. (The strict and fault-based methods, as shown below, are no longer equivalent when incarceration is a possible penalty.)

Probabilistic Detection

Now, suppose the act might sometimes evade detection; for instance, midnight dumping of toxic chemicals might only be detected with a 10% chance. If this is the case, supposing again that the harm is \$100, a \$100 fine will be inadequate to adequately deter the activity. A \$1000 fine, on the other hand, yields a \$100 average penalty (10% of \$1000), so a \$1000 fine will be optimal if people are risk-neutral. At least when money is at issue, the usual assumption is that people are risk-averse, so this sort of \$100 average penalty has greater deterrent value than a \$100 certain penalty. With a 10% probability of detection, proper deterrence would then be achieved with a fine somewhat less than \$1000, say, for instance, \$800.

The Optimality of Maximal Fines?

So far, the probability of detection has been taken as given (first as 100%, then as 10%), and the only question has been the

level of the fine. In reality, though, the probability of detection can be set by policy: greater investments in police, monitors, inspectors, and the like can improve the probability of detection. So how should society jointly set the probability of detection and the level of the fine?

For simplicity, consider the case of risk-neutral actors. A $1000 fine imposed with a 10% probability is just as effective as a $100 fine imposed for certain. (Recall that this contradicts the principle that certainty of detection has a greater effect than the size of the penalty, holding the expected penalty constant. This principle does not hold true for risk-neutral actors, who only care about expected penalties, not about probabilities and penalties as such.) Though both probability-and-penalty regimes have equal deterrent value, the first option is actually strictly better from a social point of view, because maintaining a 10% probability of detection is cheaper, requiring much less investment in law enforcement. By the same logic, a $10 000 fine imposed with a 1% probability is still better, and a $100 000 fine imposed with a 0.1% probability is better yet. One can set the fine as high as the individual's wealth, say $1 000 000, bring the probability of detection down to 0.01%, and the average will still be equal to the amount of harm, $100 (Becker, 1968).

The Optimality of Underdeterrence when Detection Is Probabilistic

In fact, we can do even better. Suppose we reduced the probability of detection even a bit further, say, to .0099%. The average will be $99, so there will be some underdeterrence, as risk-neutral actors commit crimes where the harm is $100 for a private benefit of only $99.50. But these marginal crimes are ones where the harm to the victim is approximately equal to the gain for the actor, so the social loss is approximately zero. On the other hand, any reduction of the probability, such as from .01 to .0099%, results in a nontrivial reduction in the costs of policing, which is a social gain. Thus, some amount of underdeterrence is usually optimal (Polinsky and Shavell, 1984).

Risk-Aversion

Risk-averse actors are even more convenient for society, because the prospect of losing *their entire wealth*, even with a .01% probability, deters them more than $100. To maintain the same deterrent force, the fine can be kept at its maximal level, and the probability dropped further still, with even greater savings for society. However, optimal fines are not necessarily maximal in the risk-averse case. The very convenience of the risk-averse – that they can be deterred with a lower level of fines – is also what makes maximal fines troublesome from a welfarist perspective. Maximal fines with low probabilities impose substantial risk on criminals who have not yet been caught, which is costly for those who are risk-averse. Reducing the fine from its maximal level could, it is true, reduce deterrence (or call for increased police expenditures if the probability of detection were raised to compensate), but on the other hand it would have the advantage of reducing the psychic cost borne by risk-averse criminals fearing the loss of their entire wealth (Polinsky and Shavell, 1979).

The Normative Framework Revisited

One might quibble with this normative framework. Most obviously, the model counts the benefit of the crime to the criminal equally with the harm of the crime to the victim (Bentham, 1907 [1789]; Becker, 1968); moreover, the model counts the criminal's fear of an uncertain punishment as a negative, and as a reason to reduce punishments from the level that would otherwise be optimal. (And this is even before we have introduced the possibility of judicial error: these thought experiments have assumed an errorless enforcement system.) Stigler (1970) and Lewin and Trumbull (1990) have argued that the criminal's gain is illicit and should not be considered as part of social welfare. (Cooter and Ulen (2008) are neutral in this debate.)

But counting the criminal's gain seems to follow from the inexorable logic of welfarism. Consider a status quo world A, and an alternative world B. Each person's utility is slightly higher in B than in A; that is, each person, if asked to choose, would choose B over A. However, suppose that some (perhaps all?) of people's utility in B is due to ill-gotten gains. A social planner, taking the 'ethical path' and refusing to consider the ill-gotten gains, might only compare 'legitimate' utility; he might determine that some (all?) people's 'legitimate' utility in B is lower than in A; and on that basis, he might conclude that B is socially inferior to A. But, as each person (other than the planner) prefers B over A, choosing A over B would contradict the Pareto principle, which is a foundational principle of welfare economics. (This is the sort of argument made in Kaplow and Shavell (2002).)

Extending the Simple Model of Law Enforcement

Marginal Deterrence

The preceding analysis stresses the optimality of maximal fines; nonmaximal fines were only optimal for risk-averse actors, and that result derived from the utilitarian's solicitude for the utility of the criminal himself, who suffers inordinately when, having committed a crime, he fears losing his entire wealth. Another reason for having nonmaximal fines is the need for marginal deterrence. If fines are equal for all offenses, there is no disincentive to commit a more serious offense. Someone who commits a crime might as well compound the crime by shooting the witnesses, if doing so would reduce his probability of being caught. The need to deter these gradations of crime suggests that fines should not necessarily be maximal for less serious crimes, even if criminals are risk-neutral.

Penalties Other than Fines

One straightforward extension of the simple model would allow a choice between fines and imprisonment. Fines are assumed to be cheap: ideally, the convicted criminal simply writes a check. While it is true that fine collection can be somewhat costly, any such costs pale in comparison with the immense social cost of imprisonment, which includes the direct costs of running a prison system, the productivity costs of removing the offender from society (offset by possible incapacitation benefits, see below), and the possibly *dehabilitative* effects of prison, including the brutalization of the inmate, the deterioration of his social capital and the contraction of his postrelease legal opportunities, and the improvement of his criminal knowledge and contacts. Therefore, fines should be used first (ideally, up to the level of the criminal's wealth), and incarceration should only be used as a last resort. Empirically,

one actually observes some tendency to assess more fines on wealthy defendants and sentence them to shorter terms (see Waldfogel, 1995; Meade and Waldfogel, 1998), though perhaps the reason for this is not a desire to save on incarceration resources.

Moreover, when incarceration is a possibility, the 'strict' and 'fault-based' models discussed earlier are no longer equivalent, even though they have identical deterrence properties. Under the strict model, a harm-causing act is always penalized, regardless of whether the benefit exceeds harm; as a result, many desirable acts are punished. Under the fault-based model, a harm-causing act is not penalized unless the harm outweighs the benefit; ideally, if all undesirable acts are deterred, the only observed harm-causing acts are those for which the benefit exceeds the harm, so punishment is *never* actually imposed. When punishments are very costly, the fault-based model should be strongly preferred to the strict model, provided the obstacles to the implementation of the fault-based model (i.e., determining the benefit of the act to the perpetrator) are not prohibitive (Shavell, 1987b).

Goals Other than Deterrence

So far, the goal of law enforcement has generally been assumed to be deterrence (Beccaria, 1986 [1764]), but a full analysis should also consider other goals, several of which may sometimes be consistent with the welfarist framework.

The welfarist framework excludes retributivism as a goal in itself; whether a criminal 'deserves' a particular punishment is normatively irrelevant. Some people in society may have a subjective preference for retributivism (whether it's the victims' desire for vengeance or anyone else's abstract preference for people getting their just deserts), and satisfying this preference is an independent reason for implementing the punishments that retributivism would call for. Retributivist-looking punishments may thus be justified on utilitarian grounds, though the true basis for the punishment would not really be retributivism. (Kaplow and Shavell, 2002 call this a 'taste for justice.')

The welfarist framework is capable of handling incapacitation. As an initial matter, if all undesirable crimes (those where the benefit is less than the harm) are deterred, incapacitation is undesirable: any rational criminal who commits crimes under such a regime is committing desirable crimes, so his continued offending is optimal. Irrational criminals, including but not limited to the insane, may be insufficiently sensitive to deterrent incentives, so incapacitation may be optimal for them. There may be some tradeoff between incapacitation, which counsels releasing anyone whose harmfulness is less than the cost of his incarceration (such as the old), and deterrence, which may require draconian penalties to prevent the commission of the crimes themselves (Shavell, 1987a).

Rehabilitation may also be consistent with a welfarist framework, to the extent that it teaches self-control (i.e., makes criminals better able to respond to deterrent incentives), imparts job skills that will increase the returns to law-abiding behavior after release, or changes criminals' preferences in prosocial ways. As with incapacitation, there may be a similar tradeoff between rehabilitation, which may require non-brutal prison conditions (Chen and Shapiro, 2007), and deterrence, which may require draconian prison conditions (Katz et al., 2003; Bedard and Helland, 2004). Moreover, the effectiveness of existing programs is not guaranteed (see Volokh, 2011a; Donohue and Siegelman, 1998).

For both incapacitation and rehabilitation, any result should be understood as a special case of the general idea that, under a welfarist framework, it can theoretically be optimal to incapacitate or rehabilitate even those who have not (yet) offended, provided such people can be identified reliably enough. Moreover, any intervention to alter people's actions or preferences, e.g., to avoid discriminating or to conserve environmental resources or to enjoy cheaper food, may be optimal under a welfarist framework, whether or not criminal activity is involved.

Judicial Errors

So far, though the detection system has been described as possibly probabilistic, the judicial system has been assumed to be error-free. Errors are not difficult to incorporate into this framework. If an innocent person is 10% likely to be wrongly convicted (Type I error), the rewards for obeying the law decrease, which makes criminal behavior more likely. If a guilty person is 10% likely to be wrongly acquitted (Type II error), the rewards for violating the law increase, which again makes criminal behavior more likely. Thus, *both* kinds of errors dilute deterrence.

For instance, consider an act that causes \$1000 of harm and \$900 of private benefit (relative to a \$0 baseline if the act is not committed). Without judicial errors, and with a 100% probability of detection and a \$1000 fine, a rational person does not violate the law: the payoff for breaking the law is −\$100 (=\$900 − \$1000), relative to a \$0 payoff for obeying the law. With Type I error, the innocent person is 10% likely to have to pay \$1000, so the payoff to obeying the law falls from \$0 to −\$100: a rational person is now indifferent between legal and illegal behavior. With Type II error, the guilty person is 10% likely to avoid paying \$1000, so the payoff to breaking the law rises from −\$100 to \$0: a rational person now strictly prefers to commit the act. Each type of error reduces the difference between the rewards of lawful and unlawful activity (Png, 1986).

Perhaps surprisingly, the possibility of error does not *necessarily* cut in favor of milder punishments, as it might in a more justice-oriented account. Rather, the combined probability of error (20% in this case) acts the same as a reduced probability of detection (as though acts were 20% less likely to be detected). As discussed earlier, the optimal response to such a reduced probability is to *increase* fines (by scaling them up by approximately 20%), at least where actors are risk-neutral. If actors are risk-averse, then, as before, high penalties have the disadvantage that those who may be subject to the penalty suffer by having to bear a high risk of punishment. In the presence of error, this population is much higher than it would otherwise be, as even the law-abiding may accidentally be penalized. Therefore, lower punishments may be optimal in the presence of judicial error when actors are risk-averse (Block and Sidak, 1980).

Conclusion

From Beccaria and Bentham to Becker and beyond, scholars using the economic method have woven a complex theory of

public enforcement of law. Empirical scholars have verified that policing and imprisonment have deterrent effects on crime, though theorists have also warned that the costs of the system are relevant and thus that expensive sanctions like imprisonment should be used sparingly. The use of imprisonment is one area where a more thoroughgoing use of economics could radically change a nation's penal practices in a less punitive direction.

Acknowledgments

I am grateful to Joanna Shepherd Bailey, Kay L. Levine, and Jonathan R. Nash for their helpful comments.

See also: Behavioral Economics; Crime Policy in the United States; Criminal Procedure: Adversarial and Inquisitorial Legal Systems; Death Penalty; Deterrence Theory: Crime; Deterrence; Deterrent Effect of Police and Prisons; Empirical Legal Studies; Law and Economics: Empirical Dimensions; Law and Economics; Policing; Prevention of Crime and Delinquency; Prisons and Imprisonment; Prosecution; Public Choice Economics; Punishment: Social and Legal Aspects.

Bibliography

Angrist, Joshua D., Pischke, Jörn-Steffen, 2010. The credibility revolution in empirical economics: how better research is taking the con out of econometrics. Journal of Economic Perspectives 24 (2), 3–30.

Antonovics, Kate, Knight, Brian G., 2009. A new look at racial profiling: evidence from the Boston police department. Review of Economics and Statistics 91, 163–177.

Atkins, Raymond A., Rubin, Paul H., 2003. Effects of criminal procedure on crime rates: mapping out the consequences of the exclusionary rule. Journal of Law and Economics 46, 157–180.

Ayres, Ian, Donohue III, John J., 2003. Shooting down the "More guns, less crime" hypothesis. Stanford Law Review 55, 1193–1312.

Bar-Gill, Oren, Harel, Alon, 2001. Crime rates and expected sanctions: the economics of deterrence revisited. Journal of Legal Studies 30, 485–501.

Barnes, Katherine Y., 2005. Assessing the counterfactual: the efficacy of drug interdiction absent racial profiling. Duke Law Journal 54, 1089–1141.

Beccaria, Cesare, 1986 [1764]. On Crimes and Punishments (David Young, Trans.). Hackett Publishing Company, Indianapolis.

Becker, Gary S., 1968. Crime and punishment: an economic approach. Journal of Political Economy 76, 169–217.

Bedard, Kelly, Helland, Eric, 2004. The location of women's prisons and the deterrence effect of "Harder" time. International Review of Law and Economics 24, 147–167.

Bentham, Jeremy, 1830. The Rationale of Punishment. Robert Heward, London.

Bentham, Jeremy, 1907 [1789]. An Introduction to the Principles of Morals and Legislation. Clarendon Press, Oxford.

Block, Michael K., Heineke, J.M., 1975. A labor theoretic analysis of the criminal choice. American Economic Review 65, 314–325.

Block, Michael K., Sidak, Joseph Gregory, 1980. The cost of antitrust deterrence: why not hang a price fixer now and then? Georgetown Law Journal 68, 1131–1139.

Cassell, Paul G., Fowles, Richard, 1998. Handcuffing the cops? A thirty-year perspective on *Miranda*'s harmful effects on law enforcement. Stanford Law Review 50, 1055–1145.

Chen, M. Keith, Shapiro, Jesse M., 2007. Do harsher prison conditions reduce recidivism? A discontinuity-based approach. American Law and Economics Review 9, 1–29.

Cooter, Robert, Ulen, Thomas, 2008. Law & Economics, fifth ed. Pearson/Addison-Wesley, Boston.

Corman, Hope, Mocan, H. Naci, 2000. A time-series analysis of crime and drug use in New York city. American Economic Review 90, 584–604.

Dezhbakhsh, Hashem, et al., 2003. Does capital punishment have a deterrent effect? New evidence from postmoratorium panel data. American Law and Economics Review 5, 344–376.

Di Tella, Rafael, Schargrodsky, Ernesto, 2004. Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack. American Economic Review 94, 115–133.

Donohue III, John J., 1998. Did *Miranda* diminish police effectiveness? Stanford Law Review 50, 1147–1180.

Donohue III, John J., Levitt, Steven D., 2001. The impact of legalized abortion on crime. Quarterly Journal of Economics 116, 379–420.

Donohue III, John J., Siegelman, Peter, 1998. Allocating resources among prisons and social programs in the battle against crime. Journal of Legal Studies 27, 1–43.

Donohue III, John J., Wolfers, Justin, 2006. Uses and abuses of empirical evidence in the death penalty debate. Stanford Law Review 58, 791–846.

Ehrlich, Isaac, 1973. Participation in illegitimate activities: a theoretical and empirical investigation. Journal of Political Economy 81, 521–565.

Eide, Erling, et al., 2006. Economics of crime. Foundations and Trends in Microeconomics 2, 205–279.

Feeney, Floyd, 2000. Police clearances: a poor way to measure the impact of *Miranda* on the police. Rutgers Law Journal 32, 1–114.

Glaeser, Edward L., et al., 2000. What do prosecutors maximize? An analysis of the federalization of drug crimes. American Law and Economics Review 2, 259–290.

Heineke, J.M., 1978. Economic models of criminal behaviour: an overview. In: Heineke, J.M. (Ed.), Economic Models of Criminal Behaviour. North Holland, Amsterdam.

Hicks, John R., 1939. The foundations of welfare economics. The Economic Journal 49, 696–712.

Kaldor, Nicholas, 1939. Welfare propositions of economics and interpersonal comparisons of utility. The Economic Journal 49, 545–552.

Kaplow, Louis, Shavell, Steven, 2002. Fairness versus Welfare. Harvard University Press, Cambridge, Massachusetts, and London.

Katz, Lawrence, et al., 2003. Prison conditions, capital punishment, and deterrence. American Law and Economics Review 5, 318–343.

Kessler, Daniel, Levitt, Steven D., 1999. Using sentence enhancements to distinguish between deterrence and incapacitation. Journal of Law and Economics 42, 343–363.

Knowles, John, et al., 2001. Racial bias in motor vehicle searches: theory and evidence. Journal of Political Economy 109, 203–232.

Landes, William M., 1974. Legality and reality: some evidence on criminal procedure. Journal of Legal Studies 3, 287–337.

Levitt, Steven D., 1996. The effect of prison population size on crime rates: evidence from prison overcrowding litigation. Quarterly Journal of Economics 111, 319–325.

Levitt, Steven D., 1997. Using electoral cycles in police hiring to estimate the effect of police on crime. American Economic Review 87, 270–290.

Levitt, Steven D., 1998. Why do increased arrest rates appear to reduce crime: deterrence, incapacitation, or measurement error? Economic Inquiry 36, 353–372.

Levitt, Steven D., Miles, Thomas J., 2007. Empirical study of criminal punishment. In: Polinsky, A. Mitchell, Shavell, Steven (Eds.), Handbook of Law and Economics. North-Holland, Amsterdam, pp. 455–495.

Lewin, Jeff L., Trumbull, William N., 1990. The social value of crime? International Review of Law and Economics 10, 271–284.

Lott Jr., John R., 2010. More Guns, Less Crime, third ed. University of Chicago Press, Chicago.

Lott Jr., John R., Mustard, David B., 1997. Crime, deterrence, and right-to-carry concealed handguns. Journal of Legal Studies 26, 1–68.

Marvell, Thomas B., Moody Jr., Carlisle E., 1994. Prison population growth and crime reduction. Journal of Quantitative Criminology 10, 109–140.

Marvell, Thomas B., Moody Jr., Carlisle E., 1996. Specification problems, police levels, and crime rates. Criminology 34, 609–646.

McAdams, Richard H., Ulen, Thomas S., 2009. Behavioral criminal law and economics. In: Garoupa, Nuno (Ed.), Encyclopedia of Law and Economics, second ed. Edward Elgar, Cheltenham, UK, and Northampton, Massachusetts, pp. 403–436.

Meade, Jose, Waldfogel, Joel, 1998. Do Sentencing Guidelines Raise the Cost of Punishment? National Bureau of Economic Research. Working Paper No. 6361, January.

Merton, Robert K., 1968. Social Theory and Social Structure, enlarged ed. Free Press, New York.

Mocan, H. Naci, Gittings, R. Kaj, 2003. Getting off death row: commuted sentences and the deterrent effect of capital punishment. Journal of Law and Economics 46, 453–478.

Moffitt, Terrie E., Ross, Stephen, Raine, Adrian, 2011. Crime and biology. In: Wilson, James Q., Petersilia, Joan (Eds.), Crime and Public Policy, second ed. Oxford University Press, Oxford and New York, pp. 53–87.

Niskanen Jr., William A., 1971. Bureaucracy and Representative Government. Aldine-Atherton, Chicago.
Peltzman, Sam, 1973. An evaluation of consumer protection legislation: the 1962 drug amendments. Journal of Political Economy 81, 1049–1091.
Png, I.P.L., 1986. Optimal subsidies and damages in the presence of judicial error. International Review of Law and Economics 6, 101–105.
Polinsky, A. Mitchell, Shavell, Steven, 1979. The optimal tradeoff between the probability and magnitude of fines. American Economic Review 69, 880–891.
Polinsky, A. Mitchell, Shavell, Steven, 1984. The optimal use of fines and imprisonment. Journal of Public Economics 24, 89–99.
Polinsky, A. Mitchell, Shavell, Steven, 2007. The theory of public enforcement of law. In: Polinsky, A. Mitchell, Shavell, Steven (Eds.), Handbook of Law and Economics. North-Holland, Amsterdam, pp. 403–454.
Sah, Raaj K., 1991. Social osmosis and patterns of crime. Journal of Political Economy 99, 1272–1295.
Schmidt, Peter, Witte, Ann D., 1984. An Economic Analysis of Crime and Justice: Theory, Methods, and Applications. Academic Press, New York.
Shavell, Steven, 1987a. A model of optimal incapacitation. American Economic Review 77 (2), 107–110 (special issue).
Shavell, Steven, 1987b. The optimal use of nonmonetary sanctions as a deterrent. American Economic Review 77, 584–592.
Shepherd, Joanna M., 2004. Murders of passion, execution delays, and the deterrence of capital punishment. Journal of Legal Studies 33, 283–322.
Stigler, George J., 1970. The optimum enforcement of laws. Journal of Political Economy 78, 526–536.
Sutherland, Edwin Hardin, 1956 [1937]. The Professional Thief. Phoenix Press, Chicago.
Volokh, Alexander, 2008. Privatization and the law and economics of political advocacy. Stanford Law Review 60, 1197–1253.
Volokh, Alexander, 2011a. Do faith-based prisons work? Alabama Law Review 63, 43–95.
Volokh, Alexander, 2011b. Rationality or rationalism? The positive and normative flaws of cost-benefit analysis. Houston Law Review 48, 79–98.
Waldfogel, Joel, 1995. Are fines and prison terms used efficiently? Evidence on federal fraud offenders. Journal of Law and Economics 38, 107–139.
Wolpin, Kenneth I., 1978. An economic analysis of crime and punishment in England and Wales, 1894–1967. Journal of Political Economy 86, 815–840.
Wyckoff, Paul G., 1990. The simple analytics of slack-maximizing bureaucracy. Public Choice 67, 35–47.

Law Firms

David B Wilkins, Harvard Law School, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Large law firms patterned on the Anglo-American model can now be found in most major commercial centers around the world. Yet, at the very time that mode of the production of law is spreading around the world, the traditional 'Cravath System' upon which the Anglo-American model is based is increasingly under pressure, both by sophisticated clients demanding 'more for less,' and by a range of alternative providers offering competing services at the intersection of law, technology, project management, and other professional expertise. This article seeks to explain this paradox and posits some alternative futures for large law firms in the new 'global age of more for less.'

The majority of lawyers around the world continue to practice alone or in very small and relatively informal groupings (Abel and Lewis, 1988–1995). Nevertheless, there is now a strong trend toward the creation of clearly defined and bureaucratically organized law firms of increasing size and scope. This trend toward large law firms began in the United States during the first decades of the twentieth century (Henderson, 2011). By the close of that century, this 'American mode of the production of law' had spread to the United Kingdom and Western Europe (Trubek et al., 1994). In the first decades of the new millennium, law firms patterned on the Anglo-American Model – as well as a number of Anglo-American law firms – can now be found in virtually all of the important commercial centers in Asia, Latin America, Africa, and the states of the former Soviet Union (Liu, 2008; Flood, 2013; Krishnan, 2007; Wilkins, 2010). Indeed, even some 'plaintiffs'' lawyers – the lawyers who represent individuals in civil cases – are now forming law firms of increasing size and sophistication, particularly in Australia and the United Kingdom (Morrison and Grech, 2009).

Yet, at the very time that the Anglo-American model of the large law firm appears to be spreading around the world, these same institutions find themselves under increasing pressure in their home markets, particularly after the global financial crisis (GFC) that began in the United States in 2008 and eventually spread to slow growth in most of the world's major economies (Henderson, 2014). As a result, large law firms in the second decade of the twenty-first century stand at a paradoxical intersection in which this organizational form is both increasingly ubiquitous, and increasingly viewed as ubiquitously under threat.

This article attempts to unravel this paradox by first charting why large law firms have become the dominant form for organizing legal practice in a growing number of jurisdictions, and then analyzing the forces that are now putting pressure on key elements of the traditional law firm model. The article concludes by positing some alternative futures for the large law firm.

The Rise and Transformation of the Cravath Model

The New York law firm of Cravath Swaine & Moore is frequently credited with pioneering the model for the large law firm in the first decades of the twentieth century (Galanter and Palay, 1991; Swaine, 1946). Four features of the 'Cravath Model' distinguished this firm and its progeny from its nineteenth century American predecessors (Swaine, 1946). First, the *work* of Cravath Model firms consisted of providing a full range of services to corporate clients, as opposed to individuals, with a particular emphasis on transactional and commercial matters, as opposed to courtroom litigation. Second, the *lawyers* in these firms were hired directly out of the country's newly emerging law schools. Once at the firm, they were expected to stay for 6–10 years receiving training from senior lawyers on how to do the work of the firm's corporate clients. Third, at the conclusion of this probationary period, the best of these lawyers were *promoted to partnership* in an up-or-out system in which those not promoted were asked to leave the firm – and, in all likelihood, the large law firm sector altogether, since Cravath Model firms hired virtually no lateral partners or associates. Fourth, Cravath Model firms were *governed* as true partnerships, in which all partners shared equally in the firm's profits and losses (through a 'lockstep' compensation system), and were entitled to equal participation in firm decision-making.

The Cravath Model proved to be particularly well aligned to the market conditions of the developing corporate legal sector in the United States during the first half of the twentieth century. That market was characterized by a rapidly expanding economy, particularly after World War II, and a growing pool of law school graduates as legal education spread throughout the country (Galanter and Palay, 1991). Moreover, during this period very few U.S. companies had sophisticated in-house legal counsel, leading to significant information asymmetry between even the largest corporations and their law firm providers (Gilson, 1990). As a result, corporate clients established long-term and near-exclusive relationships with their principal outside law firms, which in turn provided these organizations with significant freedom to maintain the internal training, promotion, and governance practices of the Cravath Model. Regulatory restrictions and informal norms that prevented or discouraged market behavior, such as advertising or recruiting lawyers away from other firms, further reinforced a protected environment in which Cravath Model firms faced little competition in their home markets for either clients or talent (Galanter and Palay, 1991). By the

'golden age' of the large law firm in the 1960s, law firms organized on the Cravath Model had clearly become the industry standard in the United States, representing the most important clients, hiring the best law school graduates, and occupying the most prestigious and lucrative positions in the American legal market (Galanter and Palay, 1991; Smigel, 1969).

With two notable exceptions discussed below, between 1960 and 2008 demand for corporate legal services in the United States continued to grow at an exponential rate. This growth, in turn, produced an exponential growth in the size, geographic scope, and profitability of large U.S. law firms (Henderson, 2014; Henderson and Bierman, 2009; Galanter and Henderson, 2008). This period of unprecedented growth, however, has been accompanied by significant changes in each of the four core elements of the Cravath Model. First, the *work* of many U.S. law firms has become both broader and more specialized than the work typically performed by golden age law firms. It is broader, in that previously excluded work such as litigation, criminal defense (particularly with respect to white collar and corporate crime), and even plaintiffs work (often for corporate clients who are increasingly willing to file suit) are now an important part of the work of many large law firms (Heinz et al., 2005). At the same time, many law firms have abandoned the full-service model that typified their golden age counterparts and now concentrate on increasingly specialized areas of law. This trend is even more pronounced for partners – and even many associates – who spend virtually their entire careers concentrating on increasingly narrow subsections of transactional practice or litigation (Henderson, 2011).

Moreover, as U.S. law firms have grown in both size and geographic scope, many have abandoned the Cravath Model's traditional *hiring* and *promotion* practices. With respect to hiring, virtually all U.S. law firms now hire associates and even partners 'laterally' from other law firms (and to a lesser extent, government agencies and in-house legal departments), with several firms sourcing the majority of their lawyers in this way (Henderson and Bierman, 2009). Similarly, many U.S. law firms have abandoned their traditional up-or-out promotion policy, creating new categories of professionals such as 'permanent associates' and 'nonequity partners' (Henderson, 2006; Wilkins and Gulati, 1998).

Finally, the *governance* model of the prototypical U.S. law firm has also changed significantly. Although all U.S. firms are still formally organized as partnerships (a model, as discussed below, that is increasingly under pressure in many places around the world, most notably in the UK), few of these institutions retain the egalitarian decision-making structure and lockstep compensation system that formed the heart of the 'partnership model' of golden age large law firms. Instead, with the exception of a handful of top-performing New York law firms (including Cravath itself), most U.S. law firms are organized bureaucratically, with an increasing amount of decision-making authority concentrated in the hands of formal managers – many of whom are not lawyers – and employ an 'eat what you kill' compensation system that rewards partners on the basis of their individual contribution to firm profits (Wilkins, 2007a; Regan, 2004; Chambliss, 2009; Chambliss and Wilkins, 2002).

These modifications to the traditional Cravath Model have in turn affected the way in which large law firms have developed around the world.

Globalization of the Anglo-American Mode of Production of Law

The 'Magic Circle'

The United Kingdom was the first country to develop large law firms comparable to those found in the United States. In the 1940s, the largest English solicitor firms had fewer than a dozen partners. By the second decade of the twenty-first century, four of the UK's so-called 'magic circle' law firms – Freshfields, Linklaters, Allen & Overy, and Clifford Chance – rank among the largest, most profitable, and most globalized law firms in the world, with several other prominent U.K. firms arguably pursuing an even more ambitious global expansion strategy (Flood, 2013).

As U.K. firms grew, many were quick to adopt the major modifications to the Cravath Model that had already transformed U.S. law firms in the 1980s and 1990s. Indeed, by the first decade of the twenty-first century, many of the top English firms were even more specialized, more active in lateral hiring, and more bureaucratic than their U.S. counterparts (Flood, 2013). However unlike most American firms, the majority of large U.K. law firms still retain lockstep compensation as a core element of their structure. This variation in compensation systems has become a key element in the competition between U.S. and U.K. law firms, even as the overall Anglo-American mode of production of law has become increasingly dominant around the world (Flood, 2013).

Globalization of the Anglo-American Model

Spurred by rapid economic growth and the prospect of European integration, large and increasingly sophisticated law firms began to emerge in many Western European countries during the 1980s. These law firms were initially organized as hybrids between the Anglo-American model and the traditional European model, consisting of smaller and more hierarchical firms with strong ties to the academy and the state (Trubek et al., 1994). By the mid-1990s several of these hybrid European firms had grown to more than 100 lawyers, with the largest attempting to compete with the U.S. and the U.K. firms that were rapidly expanding into the European market. This competition, however, proved to be short-lived. Within a decade, many of the top European firms had merged with – or more accurately, been acquired by – Anglo-American firms (primarily from the UK), and those that remain independent have increasingly adopted the core features of the Anglo-American model (Dezalay and Garth, 2013).

A similar transformation has occurred in Japan and the Asian Tigers of Hong Kong, Singapore, Taiwan, and Korea – although regulatory barriers have meant that the Anglo-American model of the large law firm has spread more rapidly in these jurisdictions than has the presence of Anglo-American law firms (Suzuki, 2003). As in Western Europe, in the last decades of the twentieth century a handful of law firms grew rapidly in these jurisdictions to service the needs of

multinational companies investing in these markets, as well as the expanding domestic corporate sector. Like their European counterparts, these Asian large law firms adopted many of the core elements of the Anglo-American model (Suzuki, 2003).

Since 2000, the same pattern has emerged in the BRICS (Brazil, Russia, India, China, and South Africa) and other rising powers. In each of these jurisdictions, one can now find a growing number of large 'corporate' law firms, some of which with hundreds – and in the case of China thousands – of lawyers (Wilkins and Papa, 2013). The best of these law firms now compete aggressively with Anglo-American and European large law firms for both clients and talent in their home jurisdictions. A few have even opened branch offices in major Western commercial centers such as New York and London (Li and Liu, 2012). There continue to be many differences among these new entrants to the large law firm field and their more established Western peers. For example, many emerging market large law firms are still controlled by founding families or partners, whereas most Anglo-American firms transitioned away from family or founder control several decades ago (Wilkins, 2012; Wald, 2009). Notwithstanding these differences, however, most emerging market large law firms continue to use the Anglo-American model as their organizational template.

Theories of Growth

Scholars have offered several theories to explain both the growth and the structure of large law firms. These theories tend to cluster around the traditional poles of supply and demand. The most prominent of the supply-side theories are grounded in the internal dynamics of the original Cravath Model. For example, 'tournament theory' argues that large law firms are growing exponentially because the Cravath Model's traditional 'promotion to partnership tournament' creates an internal growth engine that requires firms to hire multiple associates for every new partner that they promote (Galanter and Palay, 1991). Similarly, 'portfolio theory,' and the related idea of 'reputational bonding' asserts that law firms must grow to diversify their service offerings to protect themselves against fluctuations in demand – e.g., a mergers and acquisition practice to take advantage of increasing transactional activity in a rising economy, and a bankruptcy and litigation practice to attract work when the corporate market is declining or restructuring (Burk and McGowan, 2011; Gilson and Mnookin, 1985). Demand-side theories, by contrast, argue that large law firms have grown as a result of the escalating demand for sophisticated corporate legal services (Thomas et al., 2001; Sander and Williams, 1989).

Both supply and demand theories have made a valuable contribution to our understanding of the growth and development of large law firms, particularly in the United States during the period before 1990. However, recent events in the U.S. legal market, and even more importantly, evidence about the growth and operation of large law firms around the world, underscore the limitations of each of these explanations. Thus, as Anglo-American firms have largely abandoned the Cravath Model's traditional up-or-out hiring and promotion practices, tournament theory has become a far less persuasive explanation for law firm growth, as even that theory's most prominent original exponent concedes (Galanter and Henderson, 2008). Instead, the overwhelming majority of U.S. and U.K. law firms have achieved most or all of their growth in the last two decades through a combination of lateral hiring (of individual lawyers or increasingly teams or whole practice groups), mergers with – or, more often, acquisitions of – other law firms (including an escalating number of international combinations), and the creation of new classes of lawyers who are neither associates nor partners, including contract lawyers, permanent associates, of-counsel lawyers, and nonequity partners (Aronson, 2007). This pattern of growth by lateral hiring, merger and/or acquisition is even more prevalent for law firms in emerging markets such as China (Liu and Wu, unpublished manuscript). Indeed, although firms in these jurisdictions are outwardly structured along the lines of the Cravath Model, it is not at all clear how many of them are running anything like a Cravath-style promotion to partnership tournament in which those who are promoted gain any significant equity or decision-making interest in the firm (Krishnan, 2013; Liu, 2008). Yet these same firms are among the fastest growing law firms in the world.

Nor are contemporary U.S. and U.K. law firms typically organized along the full-service model predicted by portfolio theorists. To the contrary, as indicated above, specialization rather than diversification has been the dominant trend, with many law firms shedding traditional countercyclical or noncyclical practices such as labor and employment law or trusts and estates in favor of concentrating on high-margin work such as mergers and acquisitions, securitization, and high stakes litigation – while other firms have rushed in to fill the void by building law firms that specialize in the areas that have been abandoned by other players.

Demand-side theories also do not adequately explain the current structure and practices of large law firms around the world – at least if demand is thought of as simply an external force independent of the market actions and incentives of law firms themselves. Escalating demand for corporate legal services around the world has clearly played a significant role in fueling the growth of large law firms. Consider, for example, the BRICS. Beginning in the 1990s, each of these jurisdictions liberalized their economies to a greater or lesser extent, pouring significant public and private resources into fostering the development of a vibrant corporate sector in which private, state-owned, and foreign companies were all given a role (Trubek et al., 2013). Predictably, this development generated the need for a host of new laws and legal institutions to manage the interface with the global economy and to regulate the growing domestic corporate sector (Papa and Wilkins, 2012). And just as predictably, this flood of new domestic and global regulation has in turned fueled the growth of a vibrant corporate legal sector capable of understanding, interpreting, and manipulating this new regulatory framework (Papa and Wilkins, 2012).

Nor is it surprising that the law firms that have developed in these jurisdictions to meet this rising demand resemble – at least superficially – the Anglo-American model. As 'economic globalization' has led Western multinationals to locate larger and more sophisticated operations in these jurisdictions, lawyers practicing within the BRICS have every incentive to

signal that they can provide the same level of service as the Anglo-American law firms that have followed their clients to the BRICS to capture this same work. The easiest way to do so is to appear to replicate as nearly as possible the industry standard of the Anglo-American Model (DiMaggio and Powell, 1983). The 'globalization of knowledge' reinforces this institutional isomorphism, as lawyers educated in the West who have received a subtle but powerful indoctrination in the superiority of the Anglo-American model of the large law firm seek to create institutions that at least appear to adhere to its tenets (Papa and Wilkins, 2012). Finally, the 'globalization of governance,' including the proliferation of multinational regulation in areas such as trade, antitrust, and anticorruption, has increased the need for both Western multinationals and the growing number of multinationals based in the emerging economies to obtain access to the kind of sophisticated, integrated, and efficient legal services that have come to be associated with Anglo-American large law firms operating under the Cravath Model (Papa and Wilkins, 2012). Collectively, these three forms of globalization have created a powerful demand for Anglo-American style large law firms (Wilkins and Papa, 2013).

Nevertheless, demand alone cannot explain the patterns of law firm growth around the world today. To begin, although demand in the BRICS and other emerging economies has continued to grow since the GFC, the demand for corporate legal services in the U.S. and other developed legal markets has been flat since 2010 – and significantly down from the historic high in 2007 (Henderson, 2014). And yet, the large law firm sector continues to grow, expanding by 4% in 2013 (Jones, 2014). Virtually all of this growth, however, has been as a result of mergers and acquisition, as law firms try to hold on to their share of demand by acquiring lawyers and law firms with existing client relationships (Jones, 2014). It is already clear that in many cases this strategy is unlikely to be successful. Indeed, some of the most spectacular law firm failures of the last few years have been the result of law firms that attempted to grow their way to increased profitability by rapid and large scale mergers and/or lateral hiring (Barrett, 2013; Lattman, 2012). And yet, there is a persistent sentiment among many law firms that if they do not grow – regardless of overall profitability – they will lose their 'star' partners to law firms that are growing (Henderson, 2014; Aronson, 2007).

Economic globalization is increasingly driving this perception. Much of the growth in the U.S. legal market since 2013, for example, is the result of a wave of international mergers (Jones, 2014). Much of the rest is due to the dramatic expansion in the number and size of foreign offices, particularly in China and other Asian markets (Silver et al., 2009). Yet, there is very little evidence that this foreign expansion has increased profitability (Stern and Li, unpublished manuscript). Even with respect to domestic mergers between law firms based in the US, the evidence is quite mixed about whether these combinations increase profitability (as opposed to gross revenues), particularly relative to the overall U.S. market due to the difficulty of integrating divergent law firm cultures and practices (Aronson, 2007). Similarly, 'star' lawyers who are recruited away from one law firm for the purpose of obtaining their 'book of business' of important client relationships – frequently at the cost of a significant premium – often leave within a few years for other firms that offer even more lucrative guarantees, without ever producing the promised increase in market share or profitability (Coates et al., 2011). Global expansion has arguably fared even worse, as a host of cultural, linguistic, and historical differences between Anglo-American firms and the law firms, lawyers, and clients they seek to serve in emerging markets have multiplied problems of postmerger integration and identifying and retaining 'star' talent. Indeed by many estimates, even in booming legal markets such as China the overwhelming majority of the branch offices of non-Chinese firms are only marginally profitable at best (Stern and Li, unpublished manuscript). A recent study of Chinese law firms suggests that rapid expansion of these firms by merger, acquisition, and foreign expansion may be following a similar course (Liu and Wu, unpublished manuscript). The premise of all of these combinations seems to be that growth will fuel demand rather than the other way around.

Although counterintuitive, a growth – or more accurately lawyer-led – theory of demand is not without precedent. Demand-side theories treat lawyers as passively responding to corporate clients' need for the services lawyers in large law firms provide. American lawyers, however, have always played a more active role in promoting the demand for their services than this model suggests. Thus, the lawyers who encouraged their corporate clients to use litigation as a tool for resolving business disputes are partly responsible for the increased demand for litigators (Galanter and Palay, 1991). Similarly, law firms such as Wachtell, Lipton, Rosen & Katz and Skadden, Arps, Meagher, Slate & Flom helped fuel the mergers and acquisition boom by developing new strategies for hostile takeovers (Caplan, 1993). It is this entrepreneurial style of lawyering that is spreading around the world.

The problem for lawyers in traditional large law firms is that this strategy may have worked too well. Precisely because Anglo-American law firms have been so successful in both responding to and helping to generate corporate demand, mimicking the key forms and practices of these law firms has become a critical form of symbolic capital for aspiring corporate lawyers around the world – regardless of the underlying economics (Liu and Wu, unpublished manuscript). The result is a proliferation of large law firms that look and feel increasingly alike, all with a symbolic imperative to grow, and all chasing a market where the key decision-makers are under increasing pressure to reduce the amount of money that they spend on corporate legal services (Wilkins, 2012). This dynamic, in turn, is putting increasing pressure on the Anglo-American model of the large law firm, ironically most urgently in the country of its birth.

The Death of Big Law?

The GFC had an immediate and dramatic effect on the market for corporate legal services. Demand for the kind of mergers and acquisitions and securitization work that had fueled so much of the growth in the global corporate legal services market declined significantly (Henderson, 2014). Law firms that relied heavily on these transactions were especially hard hit, with several going out of business. Moreover, unlike previous downturns in the legal market in 2001 and 1991, this

decline in transactional work was not counterbalanced by an increase in the demand for litigation, as companies sought out less expensive ways to resolve disputes arising from the GFC (Thomson Reuters, 2012). The net result was that by the end of 2009 many large U.S. law firms laid off significant numbers of associates and support personnel, with several also de-equitizing and/or laying off partners (Henderson, 2014). By 2010, law firms in the UK and Western Europe were following suit. And while the BRICS and other important emerging markets have fared better than the U.S. and other developed economies, growth in these countries has also slowed, as has the foreign investment from the West that was fueling so much of the corporate legal market in these jurisdictions (Li and Liu, 2012). Although economic growth has begun to rebound in the U.S. and most other major economies around the world, the legal market has grown more slowly, with hiring (particularly at the entry level), growth, and profitability all either flat or below prerecession levels (Henderson, 2014).

Both the size and duration of these changes have led some academics and more than a few professionals to conclude that we are now in the midst of a paradigm shift that will result in 'The Death of Big Law' – or even more apocalyptically 'The End of Lawyers' (Scheiber, 2013; Ribstein, 2010; Susskind, 2008). It is still far too early to tell whether anything this dramatic will take place – and most reasonable observers think it is unlikely that large law firms based on the Anglo-American model will soon disappear from the legal scene (Weiss, 2013). After all, projections about the death of the big law firm model have been made since the 1991 recession, and pundits have been predicting the demise of the billable hour for decades (Kronman, 1993). Nevertheless, there are good reasons to believe that a set of fundamental changes that have been accelerated and sharpened by the GFC are likely to have an important effect on the future of the large law firm.

Three such changes are especially important. First, throughout the economy – and increasingly in the market for corporate legal services – there has been a dramatic reduction in information asymmetry between buyers and sellers. This gap began closing when corporations started to hire sophisticated in-house legal departments in the 1980s, a trend that is now spreading around the world (Gilson, 1990; Wilkins, 2012). In recent years, however, in addition to their own knowledge of the legal marketplace these in-house lawyers can now access a significant amount of data about the price and performance of law firms (and the individual lawyers who work in these organizations) from a range of public sources, and, more importantly, a new industry of private consultants dedicated to helping companies reduce their expenditure on outside law firms. As a result, sophisticated general counsel have the ability to take legal work that law firms traditionally sold as an integrated whole – e.g., 'class action litigation,' or 'a $500 million acquisition' – and 'unbundle' it into its component parts – e.g., 'discovery,' 'pretrial motions,' and 'trial,' or 'due diligence,' 'asset pricing,' 'negotiation,' and 'document drafting' – and then spread those different aspects of the work across an increasingly global supply chain of legal producers (Regan and Heenan, 2010; Daly and Silver, 2007).

Second, this trend toward disaggregation has in turn spurred the development of a host of new producers that now aggressively compete for a portion of the work on this increasingly global legal supply chain. Thus, when sourcing 'legal work' general counsels can now choose among legal process outsourcing organizations located in India and other low-cost jurisdictions, electronic discovery companies, contract lawyers, and other temporary staffing solutions, the Big Four accountancy firms which now offer advice on litigation management and deal pricing, and, of course, a growing number of increasingly sophisticated law firms around the world (Rostain and Regan, 2014; Regan and Heenan, 2010). Many of these new competitors are organized as corporations, thus giving them easier access to capital and managerial resources than law firms (Weberstaedt, 2013). As a result, these newcomers are increasingly developing new products and services that further challenge the large law firms' business model (Hadfield, 2014). Traditionally, law firms have relied on a steady increase in the hourly rates that they charge corporate clients, and an equally steady increase in leverage (the percentage of associates, who do not share profits, to partners who do) to drive profitability. But as corporate consumers become more sophisticated about their legal needs, and have greater access to a global network of providers willing and able to provide these services, law firms have had a more difficult time utilizing either of these strategies. Instead, companies are demanding greater transparency and accountability from law firms about how they are doing the company's legal work – and requiring that the law firms price these services by their 'output,' as measured by the value to the company, as opposed to the traditional 'input' model based on the number of associate and partner hours it took to produce the work (Wilkins, 2012).

Finally, this pressure on the law firm's traditional business model is in turn exposing problems in their human capital model. Notwithstanding the current downturn in entry-level hiring, large law firms face increasing pressure to attract and retain the 'best' associates and partners. Demographic changes in the pool of entrants to the profession are making this competition even more intense. Law is increasingly a 'feminized' profession, where the majority – and in many countries, the overwhelming majority – of entrants are females (Michaelson, 2013). Yet the career path in large law firms is not only designed for a man, but a man who has a wife who does not work – a rarity in many societies (Wald, 2010). Indeed, the fact that law firms have had a poor record of integrating lawyers from all nontraditional backgrounds poses an important challenge to these organizations in a world where the talent pool is more diverse in every way – including a new generation of lawyers, many of whom have a different orientation toward work and family than their Baby Boom and Generation X predecessors (Furi-Perry, 2013; Wilkins, 2007b). The steep decline in U.S. law school applications and enrollment since 2010 will only sharpen the challenge that law firms will face in recruiting and retaining top talent in the coming years (Tamanaha, 2012).

Thus, as we approach the midway point of the second decade of the twenty-first century, the profits of Anglo-American model law firms around the world are being squeezed in three directions: vertically by increasingly sophisticated corporate clients demanding 'more for less'; horizontally by competition from other law firms and from a range of new competitors seeking to 'hollow out' the work done by law

firms, either by removing routine work from the bottom through technology and process engineering, or by capturing strategic work at the top by employing multidisciplinary expertise; and internally by a 'war for talent' which, in a human capital-intensive business, talent will inevitably win. How law firms respond to this triple squeeze will determine the future viability of the Anglo-American model of the large law firm.

The Future(s) of Large Law Firms in the Global Age of More for Less

Whether the changes outlined above herald the Death of Big Law, it is quite likely that many big law firms will not survive. Some will fail outright. But many more will be absorbed into larger entities as the pace of law firm mergers and acquisitions continues to accelerate. Moreover, as the recent merger between China's King & Wood and Australia's Mallesons Stephen Jaques and the UK's SJ Berwin underscores, much of this next wave of consolidation in Anglo-American model law firms will be led by firms with their origins outside of either England or the United States. It is unlikely that large law firms will ever reach the level of concentration that we see in other professional services such as accounting or investment banking. Although globalization rewards increasing size and scope, variations in law, language, and culture will continue to place a premium on local know-how – and "know-who" (Dezalay and Garth, 2011). The resulting "globalization" (Silver et al., 2009) will leave room for a range of law firms from globally integrated full-service providers to boutiques and national champions that seek to serve particular geographies, industries, or specialties. Nevertheless, the stratification among these diverse law firm providers is likely to increase, as the number of firms able to command premium pricing for customized 'bespoke' legal work shrinks as corporate clients drive to reduce their demand for this kind of high-end service (Susskind, 2013, 2008). The trend toward removing regulatory barriers that prevent law firms from competing across geographic boundaries – a trend that soon may be mandated by international agreements such as the General Agreement on Trade in Services – is likely to further stratify the relative winners and losers in the global large law firm market (DeStefano, 2012; Terry, 2010). Indeed, deregulation, technology, and the infusion of investor capital are likely to further the development of larger and more sophisticated law firms serving individual consumers as well (Hilborne, 2014). The fact that Jacoby & Meyers, one of the largest law firms in the United States, has filed suit to allow external investment in the American market underscores the dramatic changes that could reshape the law firm market in all sectors in the coming years (Glovin and Jeffrey, 2013).

How much space will ultimately remain for large law firms seeking to occupy the 'bespoke' niche that Cravath Model firms traditionally viewed as their exclusive domain will depend upon both the ability of new providers to produce 'disruptive innovations' that capture a greater share of high-end legal work, and the willingness of law firms to 'disrupt themselves' by absorbing these new innovations into their existing business models (Christensen, 1997). The key to success for both of these efforts will be the development of new and better metrics to measure the *quality* of legal services as well as the *full cost* (long-term as well as short-term, opportunity costs as well as production costs) of producing these outcomes. Advances in artificial intelligence and the use of 'big data' to predict future trends and performance will give both 'disrupters' and large law firm 'incumbents' the opportunity to develop such metrics. Those providers that harness these new tools most effectively and efficiently will determine how much of the nineteenth century Cravath Model that underlies the current Anglo-American mode of the production of law will survive into the middle decades of the twenty-first century.

See also: Globalization: Legal Aspects; In-House Counsel; Lawyers, Regulation of; Lawyers: Social Organization of the Profession; Legal Education; Organizational Emergence and Firm Formation.

Bibliography

Abel, R.L., Lewis, P.S.C. (Eds.), 1988–1995. Lawyers in Society, 4 vols. University of California Press, Berkeley, CA.

Aronson, B.E., 2007. Elite law firm mergers and reputational competition: is bigger really better? An international comparison. Vanderbilt Journal of Transnational Law 40, 763–831.

Barrett, P.M., 2 May 2013. Howrey's bankruptcy and big law firms' small future. Businessweek (accessed 15.06.14.). http://www.businessweek.com/articles/2013-05-02/howreys-bankruptcy-and-big-law-firms-small-future.

Burk, B.A., McGowan, D., 2011. Big but brittle: economic perspectives on the future of the law firm in the new economy. Columbia Business Law Review 1–117.

Caplan, L., 1993. Skadden: Power, Money, and the Rise of a Legal Empire. Farrar Straus Giroux, New York.

Chambliss, E., 2009. New sources of managerial authority in large law firms. Georgetown Journal of Legal Ethics Ethics 22, 63–95.

Chambliss, E., Wilkins, D.B., 2002. The emerging role of ethics advisors, general counsel, and other compliance specialists in large law firms. Arizona Law Review 44, 559–592.

Christensen, C.M., 1997. The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail. Harvard Business School Press, Boston, MA.

Coates, J.C., DeStefano, M., Nanda, A., Wilkins, D.B., 2011. Hiring teams, firms, and lawyers: evidence of the evolving relationships in the corporate legal market. Law & Social Inquiry 36, 999–1031.

Daly, M.C., Silver, C., 2007. Flattening the world of legal services? The ethical and liability minefields of offshoring legal and law-related services. Georgetown Journal of International Law 38, 401–449.

DeStefano, M., 2012. Nonlawyers influencing lawyers: too many cooks in the kitchen or stone soup? Fordham Law Review 80, 2791–2845.

Dezalay, Y., Garth, B. (Eds.), 2011. Lawyers and the Rule of Law in an Era of Globalization. Routledge, London.

Dezalay, Y., Garth, B., 2013. Elite European lawyers? The common market as new golden age or missed opportunity. In: Kauppi, N., Madsen, M.R. (Eds.), Transnational Power Elites: The European Complex in the Global Field of Power. Routledge.

DiMaggio, P.J., Powell, W.W., 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. American Sociological Review 48, 147–160.

Flood, J., 2013. Institutional bridging: how large law firms engage in globalization. Boston College Law Review 54, 1087–1121.

Furi-Perry, U., 2013. The Millennial Lawyer: Making the Most of Generational Differences in the Firm. American Bar Association, Washington, DC.

Galanter, M., Henderson, W.D., 2008. The elastic tournament: a second transformation of the big law firm. Stanford Law Review 60, 1867–1918.

Galanter, M., Palay, T., 1991. Tournament of Lawyers: The Transformation of the Big Law Firm. University of Chicago Press, Chicago IL.

Gilson, R.J., 1990. The devolution of the legal profession: a demand side perspective. Maryland Law Review 49, 869–916.

Gilson, R.J., Mnookin, R.H., 1985. Sharing among the human capitalists: an economic inquiry into the corporate law firm and how partners split profits. Stanford Law Review 37, 313–392.

Glovin, D., Jeffrey, D., 2013. Jacoby & Meyers Wins Round in Nonlawyer-investor Dispute. Bloomberg (accessed 18.06.14.). http://www.bloomberg.com/news/2013-01-09/jacoby-meyers-wins-round-in-nonlawyer-investor-dispute.html.

Grech, A., Morrison, K., 2009. Slater & Gordon: the listing experience. Georgetown Journal of Legal Ethics 22, 535–540.

Hadfield, G.K., 2014. The cost of law: promoting access to justice through the (un) corporate practice of law. International Review of Law and Economics 38, 43–63.

Heinz, J.P., Nelson, R.L., Sandefur, R.L., Laumann, E.O., 2005. Urban Lawyers: The New Social Structure of the Bar. University of Chicago Press, Chicago.

Henderson, W.D., 2014. From big law to lean law. International Review of Law and Economics 38, 5–16.

Henderson, W.D., 2011. Three generations of U.S. lawyers: generalists, specialists, project managers. Maryland Law Review 70, 373–389.

Henderson, W.D., 2006. An empirical study of single-tier versus two tier partnerships in the Am law 200. North Carolina Law Review 84, 1691–1750.

Henderson, W.D., Bierman, L., 2009. An empirical analysis of lateral lawyer trends from 2000–2007: the emerging equilibrium for corporate law firms. Georgetown Journal of Legal Ethics 22, 1395–1430.

Hilborne, N., 12 June 2014. ABSs capture a third of personal injury market, SRA research reveals. Legal Week. http://www.legalfutures.co.uk/latest-news/abss-capture-third-personal-injury-market-sra-research-reveals.

Jones, L., 4 June 2014. International law firm mergers bring rebound. National Law Journal (accessed 28.09.14.). http://www.nationallawjournal.com/law-firm-profiles-result/id=1401987883053/International-Law-Firm-Mergers-Bring-Rebound?back=TAL08.

Krishnan, J.K., 2013. Peel-off lawyers: legal professionals in India's corporate law firm sector. Socio-legal Review 9, 1–59.

Krishnan, J.K., 2007. Outsourcing and the globalizing legal profession. William and Mary Law Review 48, 2189–2246.

Kronman, A.T., 1993. The Lost Lawyer: Failing Ideals of the Legal Profession. Harvard University Press, Cambridge, MA.

Lattman, P., 28 May 2012. Dewey & LeBoeuf files for bankruptcy. The New York Times (accessed 15.06.14.). http://dealbook.nytimes.com/2012/05/28/dewey-leboeuf-files-for-bankruptcy/?_php=true&_type=blogs&_r=0.

Li, X., Liu, S., 2012. The learning process of globalization: how chinese law firms survived the financial crisis. Fordham Law Review 80, 2847–2866.

Liu, S., 2008. Globalization as boundary-blurring: international and local law firms in China's corporate law market. Law & Society Review 42, 771–804.

Liu, S., Wu, H., The Ecology of Organizational Growth: Chinese Law Firms in the Age of Globalization, unpublished manuscript.

Michaelson, E., 2013. Women in the legal profession, 1970-2010: a study of the global supply of lawyers. Indiana Journal of Global Legal Studies 20, 1071–1137.

Papa, M., Wilkins, D.B., 2012. Globalization, lawyers, and India: toward a theoretical synthesis of globalization studies and the sociology of the legal profession. International Journal of the Legal Profession 18, 175–209.

Regan, M.C., 2004. Eat what You Kill: The Fall of a Wall Street Lawyer. University of Michigan Press, Ann Arbor, MI.

Regan, M.C., Heenan, P.T., 2010. Supply chains and porous boundaries: the disaggregation of legal services. Fordham Law Review 78, 2137–2191.

Reuters Thomson, 2012. Peer Monitor Index Q3 2012 Executive Report (accessed 28.09.14.). https://peermonitor.thomsonreuters.com/ThomsonPeer/docs/PMI_2012_Q3.pdf.

Ribstein, L.E., 2010. The death of big law. Wisconsin Law Review 2010, 749–815.

Rostain, T., Regan, M.C., 2014. Confidence Games: Lawyers, Accountants, and the Tax Shelter Industry. MIT Press, Cambridge, MA.

Sander, R.H., Williams, E.D., 1989. Why are there so many lawyers? Perspectives on a turbulent market. Law & Social Inquiry 14, 431–479.

Scheiber, N., 21 July 2013. The last days of big law. The New Republic (accessed 17.06.14.). http://www.newrepublic.com/article/113941/big-law-firms-trouble-when-money-dries.

Silver, C., Phelan, N.D., Rabinowitz, M., 2009. Between diffusion and distinctiveness in globalization: U.S. Law firms go global. Georgetown Journal of Legal Ethics 22, 1431–1471.

Smigel, E.O., 1969. The Wall Street Lawyer: Professional Organization Man?, second ed. Indiana University Press, Bloomington & London.

Stern, R.E., Li, S., The Outpost Office: How International Law Firms Approach the China Market, unpublished manuscript.

Susskind, R., 2013. Tomorrow's Lawyers: An Introduction to Your Future. Oxford University Press, Oxford, UK.

Susskind, R., 2008. The End of Lawyers?: Rethinking the Nature of Legal Services. Oxford University Press, New York.

Suzuki, M., 2003. The protectionist bar against foreign lawyers in Japan, China, and Korea: domestic control in the face of internationalization. Columbia Journal of Asian Law 16, 385–407.

Swaine, R.T., 1946. The Cravath Firm and Its Predecessors, 1819–1947. Ad Press, New York.

Tamanaha, B.Z., 2012. Failing Law Schools. University of Chicago Press, Chicago.

Terry, L.S., 2010. From GATS to APEC: the impact of trade agreements on legal services. Akron Law Review 43, 875–984.

Thomas, R.S., Schwab, S.J., Hansen, R.G., 2001. Megafirms. North Carolina Law Review 80, 115–198.

Trubek, D.M., Coutinho, D.R., Schapiro, M.G., 2013. Toward a new law and development: new state activism in Brazil and the challenge for legal institutions. The World Bank Legal Review 4, 281–314.

Trubek, D.M., Dezalay, Y., Buchanan, R., Davis, J.R., 1994. Global restructuring and the law: studies of the internationalization of legal fields and the creation of transnational arenas. Case Western Reserve Law Review 44, 407–498.

Wald, E., 2009. The other legal profession and the orthodox view of the bar: the rise of Colorado's elite law firms. University of Colorado Law Review 80, 605–683.

Wald, E., 2010. Glass ceilings and dead ends: professional ideologies, gender stereotypes, and the future of women lawyers at large law firms. Fordham Law Review 78, 2245–2288.

Weberstaedt, J., 2013. English alternative business structures and the european single market. Humboldt University Berlin Working Paper (accessed 17.06.14.). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2333174.

Weiss, D.C., 25 July 2013. Is the death of BigLaw overhyped? ABA Journal (accessed 16.07.14.). http://www.abajournal.com/news/article/is_the_death_of_biglaw_overhyped1/.

Wilkins, D.B., 2012. Is the in-house counsel movement going global? A preliminary assessment of the role of internal counsel in emerging economies. Wisconsin Law Review 2012, 251–304.

Wilkins, D.B., 2010. Team of rivals?: toward a new model of the corporate attorney-client relationship. Fordham Law Review 78, 2067–2136.

Wilkins, D.B., 2007a. Partner Shmartner! EEOC v. Sidley Austin Brown & Wood. Harvard Law Review 120, 1264–1277.

Wilkins, D.B., 2007b. Valuing diversity: some cautionary lessons from the American experience. In: Empson, L. (Ed.), Managing the Modern Law Firm: New Challenges New Perspectives. Oxford University Press (2007).

Wilkins, D.B., Gulati, G.M., 1998. Reconceiving the tournament of lawyers: tracking, seeding, and information control in the internal labor markets of elite law firms. Virginia Law Review 84, 1581–1681.

Wilkins, D.B., Papa, M., 2013. The rise of the corporate legal elite in the BRICS: implications for global governance. Boston College Law Review LIV, 1149–1184.

Law: History of Its Relation to the Social Sciences

W Tim Murphy, London School of Economics, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The history of law's relation to the social sciences requires unpacking the multiple meanings in play on both sides of the relation. 'Law' as a general term covers the two very different decision-making processes of adjudication and legislation. 'Social science' encompasses a wide range of epistemic projects ranging from highly positivist engagements to highly interpretive ones. These social scientific projects have increasingly stripped themselves, in different ways, of the various legalisms that once framed them. At the same time, as they successfully establish themselves as autonomous epistemic practices or sources of scientificity, both legislative and adjudicative practice expand the horizons of relevant knowledge on which to draw for legal purposes.

This article first considers a number of problems of definition that arise in relation to the terms 'law' and 'legal science.' The implications of two broad images of law – law as legislation and law as adjudication – are explored in order to sharpen the focus of the question of the relationship between law and the social sciences. Three specific aspects of that relationship are then discussed: (1) the relationship between law, legal theory, and social theory; (2) the relationship of law to the interpretive and/or phenomenological branches of the social sciences; and (3) the relationship of law to the empirical or positivistic branches of the social sciences. In each case, the influence of law upon the development of the social sciences, and the limits of these influences, is considered.

Any attempt to discuss the history of the relationship between law and social science encounters a number of problems of definition. For the purposes of this contribution, it is most important to draw attention to some of the problems that can flow from the range of meanings attached to or associated with the term or idea of law. Deriving from this range of meanings, the variety of projects associated with the idea of a legal science can be problematic. As a result of both of these elements, when an attempt is made to formulate the boundaries of the question about the relation between law and the social sciences, indeterminacy results.

Two Images of Law

At its simplest, the idea of 'law,' at least in Occidental or even Indo-European culture, conjures up two broad images. One is of legislation or rule making. The other is of adjudication or judgment. Cognate distinctions, for example, between law and policy or law and politics, and principled statements of the supposedly modern requirements of a democratic order (e.g., the doctrine of the separation of powers), all play with or draw upon this much more deeply rooted distinction, tension, and ambiguity in the idea of law. And even this primordial distinction becomes more complicated with the emergence of a self-thematized modern era, in which the idea of law as rule making falls apart or splits into two broad patterns. One pattern is the idea of making highly general, abstract, decontextualized rules that establish the framework to govern the operations of society (the so-called European codes; for this, see van Caenegem, 1987); and the other is the idea, which persists in Anglo-American or 'common law' jurisdictions, that rule making is a matter of detailed, specific rule making targeting specific problems that come to the attention of the legislator (whether, in the traditional model, through deficiencies brought to light in the course of adjudication, or, in the modern model, because of problems demanding policy adjustments that come to light in, for example, governmental bureaucracies through the growth in knowledge of the shapes and characteristics of society and its pathologies). Some have argued that, at least in the common law world, the increasing volume and significance of legislation so overwhelm our idea of law that it comes to be projected back onto the activity of adjudication, which comes to be understood as a kind of inferior lawmaking, rather than a distinct process with its own specific forms of intelligibility (Simmonds, 1984). Since Bentham (and, to some degree, Montesquieu before him), the idea of a link between a science of legislation and a science of society has seemed obvious. More obscure to social science is the idea of adjudication as such a science. Paradoxically, perhaps, the view has long been generated from within the discipline of law that it is in the domain of adjudication or adjudication-related activities that a knowledge of society is both required and developed, and from this perspective acts of legislation are seen as arbitrary or at least as the product of political will rather than an understanding of society. This kind of lawyer-knows-best attitude carried over, especially in England, into the new encounter between adjudicators and social scientists, especially in the sphere of criminal justice and criminology.

Law and Social Science

It is necessary to highlight these divergent emphases because there persists an enthusiasm in many quarters, and by no means just among lawyers, to thematize the idea that law is the oldest social science, that law was and/or remains a 'science of society,' and, in some versions, that the modern social sciences represent merely a continuation of this science or, in some cases, a derogation from it, a diminution in what society knows about itself.

The question of the relation between law and social science and the history of this relation are further confused or complicated by many of the uncertainties involved in the

specification of what is or is not within the ambit of social science. Confining this enormous issue to a narrow focus for our present purpose, we need to distinguish three questions. The first concerns the relation between law and the empirical or positivistic branches of the social sciences. The second concerns its relation to the interpretive and/or phenomenological branches of the social sciences. The third, which cannot neatly be separated from the second, is about the relationship between law – or, to switch gears, legal theory – and social theory, which becomes relevant if we assume that there is some kind of connection or link between the conceptualizations of social theory and the problematizing that emerges in the diverse fields of the empirical social sciences.

It is perhaps unsurprising that the closest affinities and even influences between law and social science seem to emerge in the domain of social theory. It is increasingly acknowledged that the writings of some of the founding fathers of social theory, and (indirectly perhaps) of sociology, had a legal background and that their theoretical constructions and projects bear the imprint of legalistic or law-inspired concerns. If, however, this is true of both Durkheim and particularly Weber (Turner and Factor, 1994), then this observation does suggest that legal influences and origins could have led late-nineteenth-century aspiring social theorists to formulate very different questions and very different theories of society, assuming that the construction of 'social theory' was what they sought to achieve. A legal background, in other words, did not seem to lead to any particular kind of theoretical orientation. In something of the same spirit, one of the leading postwar texts in legal theory describes itself as an exercise in 'descriptive sociology' (Hart, 1961). This focus makes it possible to see strong continuities between law and (the theory underpinning) social science. Some contemporary political philosophers (Rose, 1984), lawyer/psychoanalysts (Goodrich, 1997), and historians of political thought (Kelley, 1990) have not hesitated to drive home the point, with the polemical claim that modern social theory masquerades as different or new, whereas the basic framework of Western social thought is already contained in the law and indeed the tradition of law already contains a more sophisticated understanding of the central issues than much modern social theory.

Alternatively, some influential theorists adopt what is more or less the opposite position (e.g., Foucault, 1976; Luhmann, 1981). On this view, there is a sharp break between an era of social thought in which law provides the basic conceptual scheme for understanding society, and the modern or late-modern era in which this vision of society is undermined or displaced by the growth of positivistic social sciences. Hacking (1990) encapsulates what is perhaps the central dimension that underlies this view of historical change in contrasting the role played by the concept of the 'normal' in the projects of Durkheim and Galton, respectively. Durkheim's notion of the normal is still legalistic; in Galton, it becomes entirely probabilistic. And in Hacking's view, it is Galton's version in the twentieth century that wins – in science, at least.

The second area is the relation between law and the interpretive social sciences. Here, the connections are in one sense much clearer and more visible, especially if we have the 'European' law-as-code model of law in mind. At first sight, this seems to be a terrain on which law, history, and interpretive social science meet in the sense that they are engaged in the same kind of endeavor. This is because each is engaged in an exercise of interpretation, and the paradigmatic interpretive situation is the 'presence' of a text that is in some sense 'other' to the person required to interpret it. In hermeneutical science, for example, the enterprise of interpretation from this starting point is characterized as precisely a bringing together, via the text of author/creator and reader/interpreter and as the quest for a fusion of horizons, a merging of perspectives that endeavors not to transcend temporality (as does analytical philosophy) but to make it work, productively, at both ends of its spectrum. It is apparent, viewed historically, that this style of science (whether it is 'scientific' is not a question for this contribution) owes much to law and just as much to theology, and to the study of texts in each discipline in the universities of medieval Europe (Smalley, 1983). Already in these disciplines, interpretive protocols had been formalized, and their shapes were largely carried forward into the emergence of more explicitly philosophical projects by the eighteenth century. Many have traced an unbroken genealogy from that point, in the writings of, for example, Boekh and later Schliermacher through to Gadamer (1965) and contemporary practitioners of this approach to 'truth' (Rorty, 1979). Much was made of 'meaning' in these narratives of the history of a partial discipline-under-construction, and therefore in many versions Weber or *Verstehende Soziologie* were annexed to the narrative as part of the same story. 'From the actor's point of view,' a slogan variously popularized (at least within the academic world) by Malinowski and Geertz (1973, 1988), could lay claim to being the heir to this perspective or 'methodology,' and claims that this involved a direct succession to the lawyer's perspective that, for example, '*actus non sit reus nisi mens sit rea*' (the doctrine of *mens rea*, which seemingly involves investigation of inner mental states and 'intentionality' and its degrees) possessed and still possess a certain plausibility.

But it is precisely at this point in the history of law and the social sciences that continuities may need to be balanced by discontinuities. First, the once-innovative proposition that social science can proceed on the basis that culture is a text has been much debated and criticized (e.g., Clifford and Marcus, 1986). Second, political and legal anthropology in particular have raised a number of questions about the tendency in earlier studies to posit a 'jural perspective' when investigating acephalous (lit. 'headless') societies, which presented a particular challenge because no visible structures of government or law in the familiar Western sense were available for outside observers to observe (which is not to say that they could not find them virtually present) (Comaroff and Roberts, 1981). Under the influence in part of some of this work, some historians came to question the top-down model of social integration and/or development that has had a habit of permeating social as well as political history.

The net result of these overlapping concerns has been to move away from or be suspicious of the deployment of a 'juridical' model in the interpretive social sciences, and this has served to highlight the difference in the concerns or preoccupations of lawyers and social scientists. For lawyers, the key question is what beliefs, assumptions, values, or intentions it is reasonable or rational to impute to actors or legal subjects. Therein lies the craft, one might say, of the law,

and there is nothing new about this. Social scientists making use of interpretive procedures are increasingly concerned with understanding or uncovering the logic of individual and social action, including motivations, aspirations, and intentions, and they use concepts like culture, horizon, context, and situation to help render these intelligible. These social scientific ambitions are not new, but a combination of background factors (e.g., postcolonialism) and increasing sophistication in formulating these components means that interpretation has lost what innocence it ever possessed, and that the kinds of protocols employed by lawyers whenever they need to concern themselves explicitly with actors rather than texts are increasingly divergent from the highly reflexive methods now required of social scientists. Indeed, legal self-understandings have come to look to these interpretive disciplines for some corrective to what are increasingly perceived to be deficiencies in epistemic capacity and practical deliverability within the legal world itself. The sharpest symptom of this (although appearances can sometimes deceive) is the rapid rise to prominence in policy making, especially in the Anglo-American common law world, regarding the adjudicative process of alternative dispute resolution, an approach to conflict resolution that has unambiguous roots in 1950s colonial and postcolonial anthropology. One example can perhaps serve as a sharp reminder of how far apart in some respects the preoccupations of the law and of modern interpretive social sciences have become. For the law, an axiom of prudence and indeed an operational condition of possibility has more or less forever been the postulate that '*ignorantia juris haud excusat*' (lit. 'ignorance of the law does not excuse'). Contrast this with the role that '*docta ignorantia*' (learned ignorance) played in Bourdieu's not-quite-Marxist sociology of domina-tion (Bourdieu, 1980). Axiomatically, the law disregards what people claim they do not know (based on the suspicion that such disclaimers are easy to make and dubious in the first place – a protocol rooted in skepticism, in other words). By contrast, Bourdieu perhaps led the contemporary field in focusing on a modern version of Marxist false consciousness, detached from economic determinism as such, but emphasizing the prison-house of cultural schemes that prevent people from knowing what they do. What for the law is something to be ignored is for the social sciences something to be explored. For the law, ignorance is sterile; for social sciences, it is productive.

Law and the Empirical Social Sciences

What is surely unambiguous is the relationship between law and the development of the empirical social sciences, especially to the extent that these sciences are underpinned by the burgeoning of statistical knowledge and growing sophistication in statistics-gathering methods and techniques. Yet this contains within itself new ambiguities. Specifically, at least at the level of images of truth and the supporting rhetoric or legitimation strategies available to science, it is noticeable how a number of writers from the side of science – including natural science – seek to invoke the image of the law court or some of its components – notably, the jury – as the central basis for the image of truth or of what it means to claim validity for scientific statements. Trial by your peers has much ideological resonance, especially in North America; but whether it is more than a cultural theme or idiom is a more intricate question that requires explicit if not detailed discussion here.

The metaphor of scientific truth as something resolved in a court of appeal in which the judges or jury decide following adversarial confrontation is a powerful and longstanding one in Western thought, albeit one that is by no means universal (see Lloyd, 1996; Needham, 1969). While it is not the same idea, it dovetails with the notion of truth as a matter of a gentlemen's agreement (Shapin and Schaffer, 1985; Shapin, 1994), which in turn rests and relies upon the presence of implicit paradigms that are not themselves usually open to inspection or discussion (Kuhn, 1970; Foucault, 1966). Yet all of this perhaps amounts to little more than the recognition that truth is contestable and the claim that scientific practice has developed – or has always possessed – procedures for resolving epistemic conflicts that arise in the natural course of scientific discovery and invention.

These metaphors of judgment (whether by judges or juries) draw upon the experience of the world of law but do not in reality replicate it. It is interesting to note that they appeal rather more to the common law adversarial style of legal argumentation in the courtroom than the so-called inquisitorial style of the European codified tradition. But whatever precise analogy is used, it runs the risk of obscuring the important and significant gulf that opens up in the modern world between legal science – or the practice of adjudication – and empirical social science.

The most significant here is revealed in the distinction between the level of the individual (for lawyers, a 'subject,' and for many social scientists, an 'actor') and the aggregate or collective ('society'). The law has been and remains concerned with disputes between individuals, with the attribution of liability to individuals, and with the language of individual responsibility. All of this requires a focus, much of the time, on individual mental states, on postulates about what these can be taken to be or how they can be sensibly understood – or on appeals to 'ordinary people' to use their own experience of life to decide what they believe or disbelieve in the claims about such matters made by the parties in court. Judges operate within a similar framework, although they bring to their decision making the added rigmarole of professional training and/or experience with previous kinds of problems or disputes in a forensic setting. In other words, legal schemes combine two types of individualism – the individual as decision maker relying on his or her experience and capacities to draw analogies from this reservoir of experience in order to make decisions, and the individual as the object of decision, as a person who can be assumed to be like or not like the person asked to make a judgment, in matters relating to responsibility for events or actions. The law, therefore, struggles with supraindividual entities like corporations, states, nations, and cultures. Can a corporation be made liable for manslaughter or culpable homicide? This individualism of the law extends beyond persons and into events. A typical legal scenario, for example in the field of tax avoidance, is the problem of deciding whether a chain of events is to be understood as a multiplicity of individual events or as

a composite ('quasi-individual') event. For lawyers, these are not simple questions.

Empirical social science, by contrast, is preoccupied, at its most subtle, with the often perplexing relationship between 'micromotives' and 'macrobehavior' (Schelling, 1978). Correspondingly, individualization of events is of much less (if any) interest than patterns, tendencies, logics, and trends. Critical to these perspectives are perhaps two modern capacities – measuring and modeling. And underpinning these new epistemic possibilities are modern notions of probability and of statistical thinking that diverge sharply from the traditional idea of probability still largely at work in the ideational schemes of the law (Hacking, 1975; Porter, 1986; Gigerenzer et al., 1989; Murphy, 1997).

These developments challenge legal schemes of reality. They create a reality through practices and procedures that are remote from the mechanisms through which law constructs 'its' truth. Surveys, data analysis and number crunching, the quantification of the qualitative (and the very emergence of an explicit distinction between these two, and the consequent problematization of the relationship between them), and probabilistic judgment – all these epistemic practices are quite different from the argumentative, judgmental, and commonsensical manner in which law decides what is true. Even before this became obvious, there were pressures upon law to acknowledge what was changing and to incorporate into itself the findings or the protocols of social science (most famously in the so-called Brandeis brief in the United States, in which statistical information was introduced into court as part of the case). In modern US and UK laws, for example, a concept of indirect discrimination has been introduced, which makes it unlawful to discriminate not just intentionally but also by applying to, for example, job applicants a requirement or condition with which members of the protected group – principally race or sex, but these definitions can be expected to expand over time through new legislation – are disproportionately unable or less able to comply. This kind of formulation inevitably invites the introduction of statistical data into the forensic setting. Other outputs of modern social and behavioral sciences increasingly find their way into such settings, notably in criminal law and family law. Social inquiry reports for offenders before sentencing, or all manner of inputs on child welfare, have come to play a role in the way that cases are disposed of in the areas of sentencing and child custody disputes. It is also increasingly the case that the output of the social sciences plays a role in the evaluation of current law, and generates critique, proposals for new legislation, and arguments for changes in the outcomes of judicial decision making.

These kinds of developments lead some observers and commentators to discern a historical process at work in which the legal sphere steadily succumbs to the influences of the social and behavioral sciences. Foucault's thesis about normalization and disciplinary power (Foucault, 1975), for example, can be read in this way (see also Donzelot, 1977). Others stress, rather, the rise of differentiation. Among many legal theories if not social theories, one of the most influential has been the function-systems theory of Niklas Luhmann (1984). At its simplest, what is at issue here is whether the 'worlds' of law and of social science are converging, merging, or pulling further apart. The answer is probably that both tendencies are simultaneously visible. Convergence, of course, now means the 'colonization' of law by social science, and this is most likely to occur in the area of individual and especially cognitive psychology. In the future, we can anticipate that genetics will play a greater role here. But pulling in the opposite direction are strong tendencies toward the assertion of legal autonomy, in the international as well as national and subnational spheres. As often in the previous history of law, legal thought often proceeds ahead of successful institutionalization, while without the latter the former remains stuck in the ebb and flow of opinion and fashion. However, to the extent that this tendency toward autonomy is becoming a 'reality,' it potentially means that the pre-social-scientific epistemic schemes of the law will succeed in extending their life and even consolidating their position into the future, disregarding the findings of social science on the grounds of increasingly abstract principle.

See also: Critical Legal Studies; Hermeneutics, History of; Law and Society: Development of the Field; Law as an Instrument of Social Change; Law: Anthropological Aspects; Legal Process and Social Science: United States; Social Science, The Idea of.

Bibliography

Bourdieu, P., 1980. Le Sens Pratique. Les Editions de Minuit, Paris (1990. The Logic of Practice. Polity, Cambridge, UK).

van Caenegem, R.C., 1987. Judges, Legislators and Professors: Chapters in European Legal History. Cambridge University Press, Cambridge, UK.

Clifford, J., Marcus, G.E. (Eds.), 1986. Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley, CA.

Comaroff, J., Roberts, S., 1981. Rules and Processes: The Cultural Logic of Disputes in an African Context. Chicago University Press, Chicago, IL.

Donzelot, J., 1977. La Police des Familles. Les Editions de Minuit, Paris (1979. The Policing of Families: Welfare Versus the State. Hutchinson, London).

Foucault, M., 1966. Les Mots et les Choses: Une Archéologie des Sciences Humaines. Editions Gallimard, Paris (1970. The Order of Things: An Archaeology of the Human Sciences. Tavistock, London).

Foucault, M., 1975. Surveiller et Punir: Naissance de la Prison. Editions Gallimard, Paris (1977. Discipline and Punish: The Birth of the Prison. Pantheon, New York).

Foucault, M., 1976. La Volonté de Savoir. In: Histoire de la Sexualité, vol. I. Editions Gallimard, Paris (1979. An Introduction. The History of Sexuality vol. I. Allen Lane, Harmondsworth, UK).

Gadamer, H.-G., 1965. Wahrheit und Methode, second ed. J. C. B. Mohr, Tübingen (1979. Truth and Method, second ed. Sheed and Ward, London).

Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.

Geertz, C., 1988. Works and Lives: The Anthropologist as Author. Stanford University Press, Stanford, CA.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Krüger, L., 1989. The Empire of Chance: How Probability Changed Science and Everyday Life. Cambridge University Press, Cambridge, UK.

Goodrich, P. (Ed.), 1997. Law and the Unconscious: A Legendre Reader. St. Martin's Press, New York.

Hacking, I., 1975. The Emergence of Probability. Cambridge University Press, Cambridge, UK.

Hacking, I., 1990. The Taming of Chance. Cambridge University Press, Cambridge, UK.

Hart, H.L.A., 1961. The Concept of Law. Clarendon Press, Oxford, UK.

Kelley, D.R., 1990. The Human Measure: Social Thought in the Western Legal Tradition. Harvard University Press, Cambridge, MA.

Kuhn, T.S., 1970. The Structure of Scientific Revolutions, second ed. University of Chicago Press, Chicago, IL.

Lloyd, G.E.R., 1996. Adversaries and Authorities: Investigations into Ancient Greek and Chinese Science. Cambridge University Press, Cambridge, UK.

Luhmann, N., 1981. Politische Theorie im Wohlfahrtsstaat. Olzog, Munich, Germany (1990. Political Theory in the Welfare State. Walter de Gruyter, Berlin).

Luhmann, N., 1984. Soziale Systeme: Grundriss Einer Allgemeinen Theorie. Suhrkamp, Frankfurt am Main, Germany (1995. Social Systems, Stanford University Press, Stanford, CA).
Murphy, W.T., 1997. The Oldest Social Science? Configurations of Law and Modernity. Oxford University Press, Oxford, UK.
Needham, J., 1969. The Great Titration: Science and Society in East and West. Allen and Unwin, London.
Porter, T.M., 1986. The Rise of Statistical Thinking 1820–1900. Princeton University Press, Princeton, NJ.
Rorty, R., 1979. Philosophy and the Mirror of Nature. Princeton University Press, Princeton, NJ.
Rose, G., 1984. Dialectic of Nihilism: Post-structuralism and Law. Blackwell, Oxford, UK.
Schelling, T.C., 1978. Micromotives and Macrobehavior. Norton, New York.
Shapin, S.A., 1994. Social History of Truth: Civility and Science in Seventeenth-Century England. Chicago University Press, Chicago, IL.
Shapin, S., Schaffer, S., 1985. Leviathan and the Air-Pump: Hobbes, Boyle and the Experimental Life. Princeton University Press, Princeton, NJ.
Simmonds, N.E., 1984. The Decline of Juridical Reason: Doctrine and Theory in the Legal Order. Manchester University Press, Manchester, UK.
Smalley, B., 1983. The Study of the Bible in the Middle Ages, third ed. Blackwell, Oxford, UK.
Turner, S.P., Factor, R.A., 1994. Max Weber: The Lawyer as Social Thinker. Routledge, London.

Law, Mobilization of

Calvin Morrill and Mayra Feddersen, University of California, Berkeley, CA, USA
Stephen Rushin, University of Illinois, Champaign, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E. Blankenburg, volume 12, pp. 8530–8534, © 2001, Elsevier Ltd.

Abstract

Legal mobilization refers to the social processes through which law is invoked in situations perceived as problematic. In liberal democracies, much of legal mobilization occurs in the activation of legal rights, emerging out of processes through which people transform the meaning of injurious situations and seek redress for them. Multiple contextual factors influence this process, including social hierarchy, interpersonal relationships, culture, and organizational and institutional (legal) structures. Collective legal mobilization is often aimed at social change with both instrumental (indirect/direct) and constitutive effects. Cutting-edge research has begun to examine legal mobilization across multiple contexts, including authoritarian regimes and transnational fields.

Legal mobilization refers to the social processes through which law is invoked in situations perceived as problematic. Such invocations can occur reactively as individuals or groups bring complaints to a legal system or by legal officials acting proactively (Black, 1972).

Individual-Level Mobilization

Research on individual-level mobilization arose out of the realization that North American and Western European legal systems are largely reactive and rights passive unless activated by citizens. To study this activation, scholars developed the 'transformative' approach to legal mobilization that focuses on how individuals make sense of situations they perceive to be normative breaches. The beginnings of such transformations typically involve 'initially amorphous and indistinct' social interaction in which it is unclear whether a normative infraction of any kind has occurred (Emerson, 2008: 485). Felstiner et al. (1980–1981) conceptualize these transformations as unfolding in unstable and incomplete stages, beginning with the 'naming' of an experience as a personal injury and/or rights violation; continuing to 'blaming' in which one party attributes fault to another party for an injury; and then moving to 'claiming' at which point the injured party voices their grievance to the party or parties believed to be responsible and seeks redress. In this formulation, if a claim is rejected, the process transforms into a 'dispute' and third-party involvement (e.g., by police or a court) becomes more likely.

In the United States, the only national survey of individual-level legal mobilization to use random sampling, the Civil Litigation Research Project (CLRP), found in the 1970s that only a small fraction of situations perceived as injurious evolve into disputes and an even smaller fraction, less than 5%, reach a trial stage. This aggregate natural history of legal mobilization in American society came to be conceptualized as the 'dispute pyramid,' with a broad base of recognized grievances and rights violations, yet dramatic attrition as one moves up through blaming, claiming, and finally to the dispute stage and legal intervention (Sarat and Miller, 1980–1981).

Despite the seemingly linear representations in both the 'naming, blaming, claiming' and 'dispute pyramid' approaches, neither model was intended to represent the process of legal mobilization as a completely orderly or linear process. Empirical work on legal mobilization underscores its nonlinearity. For example, a British survey documents how people cycle among self-help, consultations with citizens' advice bureaus, and the mobilization of legal officials. As aggrieved persons move through these contexts, their understandings of their disputes can change, oscillating back and forth along a continuum between more amorphous senses of trouble (akin to prenamed, anomalous social interaction) to more sharply defined grievances (Genn, 1999). Such indistinct and nonlinear processes also appear in Ellickson's (1991) study of how ranchers in Shasta County, California, settle disputes. When faced with property damage caused by cattle wandering across property lines, ranchers and others do not move in a linear process from attempts at informal remedies to claiming liabilities based on the laws of trespass. Instead, they resort to congeries of gossip, mild forms of physical retaliation (to the offending rancher's property), and negotiated attempts at remediation until an outcome is reached.

Although a steady stream of research has focused on individual-level legal mobilization in the United States and Europe, it is also useful to remember that anthropologists have studied legal mobilization as one form of social control across a broad array of cultures, including societies without state-based legal systems. Legal realist law professor Karl Llewellyn and anthropologist E. Adamson Hoebel's (1941) early classic, *The Cheyenne Way*, provided an important point of departure by introducing the 'trouble case' as a conceptual and methodological tool to study breakdowns in routine societal processes and responses to them via lawlike mobilization and other types of social control. Trouble cases are important, so Llewellyn and Hoebel argued, because they both reflect and help compose the sociocultural contexts in which they are embedded. Anthropologist Laura Nader built on Llewellyn and Hoebel's work to coin the concept of the 'disputing process' as an analytic window through which to peer cross-culturally into social situations when everyday social injustices are challenged, remedied, or reproduced by legal and other actions (Nader and

Todd, 1978). Resonant with CLRP and subsequent research, findings point to the mobilization of law or lawlike, third-party settlement agents, such as judges or tribunals, as relatively rare in human societies among all mechanisms for addressing social wrongs. Far more common are attempts at unilateral actions to repress or change normative breaches, dyadic negotiations, avoidance, or toleration (Black, 1993).

Taken together, these approaches and findings have led to a reconceptualization of individual-level legal mobilization in contemporary American contexts as a multidimensional process in which many kinds of actions can occur sequentially and simultaneously, including the possibility that people can enforce their rights extralegally. In their 'multidimensional process model' (MPM), Morrill et al. (2010) characterize legal mobilization as emerging out of a pool of perceived legal violations (which can have initially amorphous parameters and tenuous links to law) and comprising multiple modes of action: (1) formal legal action, such as filing a formal lawsuit, filing a formal complaint with a government agency (e.g., the United States Equal Employment Opportunity Commission or the European Commission on Human Rights), or contacting a lawyer; (2) quasi-legal action, such as using formal complaint procedures or internal dispute resolution forums (e.g., mediation, peer counseling) provided by an organization (e.g., a corporation, government agency, school, or church); (3) extralegal action, such as contacting the media, directly confronting a person verbally or physically, seeking support from a counselor or religious professional, avoiding the person, talking with peers or family members, or engaging in prayer; and (4) doing nothing ('lumping it'). **Figure 1** overlays the MPM (drawn from Morrill et al., 2010) on the outline of the classic dispute pyramid. With the exception of doing nothing, these responses are not mutually exclusive as individuals can pursue several strategies at once. Moreover, they may not unfold in temporally linear ways (e.g., from more formal to less formal or vice versa) given the social contexts in which they are embedded (Emerson, 2008). Instead, these strategies can operate as potentially reinforcing. Or they can be in tension with each other depending upon their salience and meaning in particular sociocultural and legal contexts. By conceiving of mobilization as a multidimensional process, this approach may reveal a broader range of options that people have for mobilizing their rights apart from law itself. This framework also can facilitate examining how people combine different options as they mobilize law and other mechanisms to pursue their rights. Ultimately, the MPM may enable a richer understanding of the roles that law plays in lives of individuals.

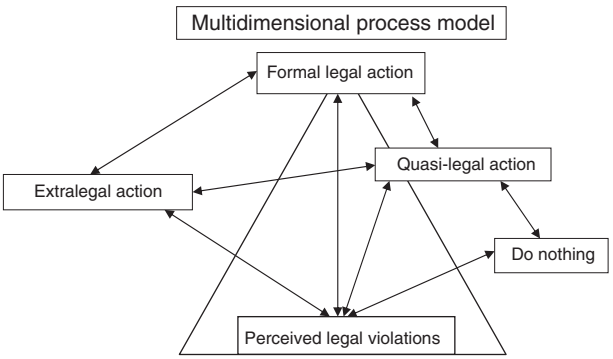

Figure 1 Multidimensional process model overlaid on the outline of the classic dispute pyramid. Morrill, C., Edelman, L.B., Tyson, K., Arum, R., 2010. Legal mobilization in schools: the paradox of rights and race among youth. Law and Society Review 44 (3-4), 651-694.

Contextual Factors Influencing Individual-Level Legal Mobilization

A growing body of research suggests that a variety of contextual factors based in social demographic, interpersonal, cultural, organizational, and institutional dynamics influence the individual mobilization of law. Various studies have found that individual attributes like race, gender, social class, and citizenship status affect the likelihood that a person will mobilize legal protections. For example, the high cost of litigation deters the poor from exercising legal protections. Additionally, interpersonal ties can influence whether and how people perceive normative infractions with social intimacy tending to channel claims away from disputing and third-party intervention. Legal consciousness – the cultural schemas that individuals draw on for their common sense understandings of law – may also alter a person's willingness to use legal remedies. And organizational structures and broader social institutions can impede or redirect individual attempts to mobilize the law.

Generally, members of socially or economically disadvantaged groups are often less likely to mobilize legal protections. Members of these groups also suffer from worse litigation outcomes. One of the earliest observations of this phenomenon came from Carlin et al. (1966), who found that although the poor were more likely to need legal services, they were actually less likely to mobilize their legal protections. The poor suffer more frequently from consumer fraud, illegal rent practices, and law enforcement abuses. But since the poor lack resources and often experience negative or hostile interactions with the law (particularly the police), they turn to law to solve problems less frequently than their wealthier counterparts. Later empirical studies verified these findings. For example, Mayhew and Reiss (1969) found that family income and property ownership strongly shape how and what people perceive to be violations of the law and predict an individual's proclivity to mobilize the law. This conclusion is not surprising. Given the high cost of litigation, individuals hoping to gain access to legal forums face high economic entrance barriers. This realization spurred the access to justice movement in the 1970s and 1980s, which in turn led to a proliferation of community-based walk-in legal clinics and alternative dispute resolution agencies. Contemporary studies expanded these findings by noting that other demographic characteristics, such as race, gender, and citizenship status affect the probability of legal mobilization. And in the last few decades, scholars have also observed that intersectionality – a person's overlapping membership in multiple demographic categories – uniquely disadvantages individuals in exercising their legal rights (Crenshaw, 1989; Best et al., 2011).

Apart from demography or material resources, relational closeness appears to 'suppress the invocation of legal rules in dispute processing and conflict management' (Morrill et al., 1998: 644). In socially intimate dyads, for example, those in

conflict are likely to have a future orientation to their relationship, which embeds the stakes and management of their conflicts in the interpersonal norms governing the relationship itself. Strong social ties in dyads thus lead away from third-party intervention unless dyadic remedial actions fail (Emerson, 2008).

Legal consciousness also influences whether a person will turn to the law to resolve a dispute. Individuals resolve conflict differently, due in part to their relative awareness of legal rights and understanding of the role of law in society. Ewick and Silbey (1998) have identified different categories of legal consciousness that predict how a given person will utilize the law in the event of a conflict. They predict that politically disenfranchised groups often distrust the law, and are thus unlikely to use the legal system to resolve grievances. Conversely, groups with power find the law accessible. These groups more commonly invoke legal protections in times of need. Thus, this theory may explain why certain racial minorities and socially disadvantaged groups are less likely to mobilize legal protections.

Abrego (2011) provides a compelling empirical investigation of Ewick and Silbey's approach to legal consciousness in the context of undocumented immigrants in the United States. In the US, all undocumented immigrants are technically violating the law by residing in the country and yet their legal consciousness varies quite a bit. Persons who migrated to the US without documentation during adulthood generally fear and are unwilling to mobilize the law, viewing the law as a source of possible deportation and punishment. They feel responsible for unlawfully entering the country and are more 'willing to accept their marginalized status in society' (Abrego, 2011: 363). By contrast, undocumented persons who migrated to the country as young children are less fearful. They have occupied legitimate spaces in the country by participating in public aspects of the society in ways parallel to documented persons (such as attending and graduating from public schools) and do not view themselves as marginal members of society or law as a source of fear, but instead as a source of unwarranted stigmatization. Abrego (2011: 363) concludes that while "stigma can certainly be a barrier to claims-making, the threshold to overcome it is relatively lower" than fear.

Yet another influence on individual-level legal mobilization can be found in formal social organizations, particularly large bureaucracies. To explain "why the haves come out ahead," Galanter (1974) originally drew the classic distinction between litigants who are 'one shotters' and 'repeat players'; the former typically being individual persons and the latter, bureaucratic organizations. Bureaucracies have distinct advantages compared to individuals in legal mobilization, including being able to anticipate 'repeated litigation' and 'pursue long-term interests' with specialists able to structure transactions in advantageous ways by lobbying or litigating to secure favorable laws (Galanter, 1974: 98). In the decades since Galanter (1974) made these observations, large organizations have increasingly 'internalized' the law by creating organizationally based dispute resolution procedures, expanding the prominence of in-house counsel in organizational decision making and disputing, and creating private police/security forces (Edelman and Suchman, 1999: 943). Although much of organizational conflict only rises to the level of disputing under particular organizational conditions, dispute resolution forums can channel both the ways that individuals perceive injurious situations and whether and how they mobilize the law (Morrill, 1995). By setting up dispute resolution forums, organizations in theory could provide employees with an affordable and efficient avenue to resolve disputes both with each other and with the organization. However, dispute resolution structures, especially when they are unmoored from third-party organizations, such as unions, who might balance power asymmetries for employees, run the risk of 'managerialization' in that they come to reflect the institutionalized interests of official, organizational authority (and entrenched powers in broader industries and fields of which they are a part) rather than employees who would be most likely to use them (Edelman, 1992). Indeed, an employee bringing a claim against their employer to an alternative forum might have to present their case to a judge or arbiter "who is a member of the organization's management team" (Edelman and Suchman, 1999: 966). Such forums, therefore, could ultimately suppress employee claims either out of fear of retaliation from the organization or simply because employees give up, believing such forums to be a sham.

Intertwined with the power of organizational structures to shape individual-level legal mobilization are social institutions. The concept 'social institution' refers to macrolevel social structures of meaning and organization. Examples of social institutions include the family, gender, race, or markets. Social institutions can influence both how people view and make sense of law (legal consciousness), as well as how they mobilize what they define as their rights. Marshall's (2005) study of legal mobilization in the wake of unwanted sexual attention in the workplace provides a poignant illustration. Managerial interests and institutionalized expectations of male–female relations shape policies intended to protect women from unwanted sexual attention. As a result, these policies often lead to managerial skepticism about the character and pattern of unwanted sexual attention. Women may anticipate this skepticism and develop strategies apart from legal or quasi-legal responses, including avoiding the person or persons they hold responsible and/or seeking emotional support to cope with the trouble. The interplay between organizational structures, social institutions, and practices thus enacts "a legal consciousness that reflects a narrow meaning for 'sexual harassment'," thereby discouraging legal mobilization (Marshall, 2005: 84). Similarly, Nielsen (2004) finds that the social institutions of gender and race pattern how individuals respond to remarks they deem as racist or sexist in public contexts. While all persons find offensive public remarks to be a serious social problem, white women and persons of color are more reluctant to turn to law than white men because of their skepticism that the law can be used to solve their problems and their fear that law 'would be used against them' (Nielsen, 2004: 2).

Other researchers find that legal mobilization is not only shaped by existing social institutions but is able to challenge them. Albiston (2010), for example, studied the role of social institutions on individual mobilization of legal rights under the Family and Medical Leave Act (FMLA). The FMLA provides up to 12 weeks of unpaid leave per year for workers in the United States who have family medical crises, but the meanings of the rights it confers and their mobilization by both men and women are conditioned by institutionalized workplace norms

about the meaning of work being a 'full-time, year-round endeavor' with the most common family configuration being the male breadwinner/stay-at-home housewife model (Albiston, 2010: 1, 5). These institutions can suppress rights mobilization by both men and women or lead to difficult negotiations with bosses about what is possible under the FMLA umbrella. Nonetheless, Albiston (2010) argues the mobilization of rights under the FMLA that can create opportunities for interrupting institutionalized expectations regarding family configurations and workplace norms as FMLA rights become legitimated and their mobilization normalized. Likewise, Levitsky (2008) argues that Americans remain reluctant to define health care as a 'right or entitlement of citizenship.' Instead, they imagine solutions to unmet health needs within existing institutionalized beliefs about families, markets, and the welfare state. At the same time, new needs compel individuals to articulate new claims, legal and otherwise, about health care thus challenging existing social arrangements but within the existing welfare state framework.

In the most ambitious theoretical project devoted largely to individual-level legal mobilization, Donald Black (1976) stitches together a broad array of the findings from cross-cultural research to develop propositions that predict in any social situation, which behaviors will be defined as illegal, who mobilizes the law, and who is likely to receive favorable outcomes from the invocation of law. These predictions are a function of the 'social structure' of cases, including their social inequality, social intimacy, cultural differentiation, organizational capacities, and link to extralegal normative orders. Unlike the transformations approach, Black explicitly eschews the subjective agency of individuals, instead treating legal mobilization as a purely behavioral dynamic. Moreover, he envisions law as a quantitative variable operationalized by the behavior of state officials such that, for example, "an arrest…is more law than if no such action occurs" (Black, 1980: 210). The theory is intended to apply to the mobilization of law in all historical periods and societies. Researchers have extended the theory in multiple ways to a variety of social and historical contexts.

Collective Legal Mobilization

Although the boundaries between individual-level and collective legal mobilization are sometimes fuzzy, this distinction, both theoretically and empirically, is useful. This boundary helps scholars more fully examine the assumptions and contradictory consequences of law as it relates to social change. Research on collective legal mobilization emerged out of the US civil rights era, particularly examining whether rights mobilized and legal victories gained through litigation strategies (e.g., the overturning of the 'separate but equal' doctrine), produced or can produce meaningful social change. Scheingold (1974) termed the fervent belief in the power of rights to produce social change, the 'myth of rights,' in order to underscore how foundational this belief is in the US, especially among disadvantaged social groups, despite equivocal evidence about the power of rights to produce change or even be mobilized. The myth of rights thus produces a politics of rights, becoming a valuable resource for social movements in the pursuit of social change.

Building in part off Scheingold (1974), Handler (1978) explicitly focuses on the tangible benefits instrumentally produced by movement-spurred litigation, identifying what he called *direct* and *indirect* effects. Movement litigation accomplishes direct effects when there is a definitive legal ruling that alters the law on the books in ways favorable to reform supporters and when those changes lead to sustained enforcement of the law in action. Movement litigation may have a limited capacity to achieve meaningful social change because such efforts do not consider the organizational and social contexts where such innovations would be implemented. Even if a movement succeeds in changing the laws on the books, for instance, it may fall short of changing the law in action due to a lack of administrative capacity, scarce resources, and/or countermobilization. Handler (1978) thus draws on the resource mobilization perspective in social movement theory (McCarthy and Zald, 1977) by calling attention to the importance of organizational resources in facilitating collective legal mobilization aimed at social reform. By contrast, the indirect effects of movement-spurred litigation leverage bargaining power for nonlitigation strategies, such as movement solidarity, publicity, and fundraising. Here again, resources play a key role but in a broader sense, as they fuel political strategy rather than only administrative enforcement. The indirect effects of collective legal mobilization can lead to imaginative extralegal tactics that differ from the narrower interests of the legal profession in changing the laws on the books (Handler, 1978: 222).

Conceptual schemes, sometimes explicitly or implicitly echoing Handler's (1978) instrumental perspective, continue to inform thinking about the consequences of collective legal mobilization in the United States. Rosenberg (1991), for example, argues that scholars regard courts as either 'constrained' or 'dynamic' in evaluating the direct effects of collective litigation to change national-level social practices across multiple bureaucratic and institutional settings. In the 'constrained court' view, various aspects of the American legal and political systems constrain courts from fulfilling the hopes believers place in them: boundedness to constitutional rights; a lack of judicial autonomy from other branches of government; and a paucity of 'tools' to implement law. Believers in the 'dynamic court' perspective claim that adherents to the 'constrained court' perspective 'underestimate' court effectiveness by ignoring how courts act as catalysts for social change as they 'dramatize' social issues (Rosenberg, 1991: 25). Rosenberg finds that the US Supreme Court has not directly instigated a single social change or indirectly motivated change through consciousness raising or other symbolic processes. Although the direct and indirect effects of movement litigation are difficult to measure owing to problems of appropriate comparison, operationalization, and temporality, in the end, Rosenberg (1991: 35) argues that only under conditions of extralegal incentives and costs (e.g., by the market) or when legal reforms offer plausible rationales (i.e., 'cover') for those 'willing to act' can collective litigation produce direct, significant social change.

In contrast to Handler (1978) and Rosenberg's (1991) instrumental approaches, constitutive approaches to collective legal mobilization view law not as an exogenous tool for social change or consciousness raising but rather as a cultural system of communicative practice and knowledge embedded

in society. The most well-known proponent of this approach is McCann (2004), who views law as complex, mixed, variable, and contingent upon social and historical context. From a constitutive vantage point, attention is devoted to how legal discourses and symbols intersect with and are expressive of broader ideological formations. Law becomes a way to structure social relations and categories, formulate moral visions for social life, and negotiate conflict. Constitutive approaches tap into the full array of social movement theory identifying how resources, political opportunities, and framing relate to collective legal mobilization (McAdam et al., 1996). Power figures prominently in these dynamics as institutional authorities can control or repress meanings, enforce boundaries, and use selective violence against those who violate official readings of the law.

McCann's (1994) study of the pay equity movement in American workplaces illustrates the constitutive approach. In his research, he found that rights are continually being negotiated, constructed, altered, and resisted depending upon where a movement is in its 'stage' of legal mobilization: movement building and emergence; identifying compelling formal changes in official policy; policy implementation; or informing subsequent struggles. Although collective legal mobilization can importantly raise consciousness in the initial stages of a movement, it can also pave the way for subsequent interpretations of law that can facilitate reform in unanticipated ways even in the wake of failed reform efforts. Beyond pay equity reform, the failed attempts to pass the Equal Rights Amendment (ERA) in the United States illustrate these processes. The ERA would have guaranteed equal rights for women. Although never passed as a piece of legislation, mobilization around it helped forge new constitutional interpretations that led to eventual recognition of equal protection doctrine prohibiting discrimination based upon sex (Siegal, 2006).

The recent US Supreme Court decision in the *Dukes v. Wal-Mart* (2011) gender discrimination case illustrates many of the complexities relevant to collective legal mobilization. The Court ruled that this class action case could not go forward on procedural grounds, specifically that "female employees would not be able to establish discrimination across all of Wal-Mart's 3400 stores." The Court also placed "great weight on the fact that 'Wal-Mart's announced policy forbids sex discrimination'" (Edelman et al., 2011: 889). In this sense, the Court deferred to the efficacy of organizational policies despite considerable evidence (again, in dispute) that those policies did not produce their ostensibly intended effects in Wal-Mart stores (Edelman et al., 2011). This case underscores how organizational power can thwart the capacity of law to effect social change – something that Handler (1978) long ago identified. Beyond Handler (1978), however, contemporary scholars point to the power of organizations to influence not only the implementation of legal reforms once they are decided, but also legal decision making itself. Moreover, *Dukes v. Wal-Mart* may carry constitutive effects as it symbolizes for many who face discrimination, the futility of participating in collective legal mobilization aimed at righting social injustice and will almost certainly influence subsequent legal decision making in this and related areas of law. Thus, the constitutive effects of law may cut both ways, paving the way for change and constraining reform.

Cutting-Edge Research on Legal Mobilization

Multiple generations of researchers have shed light on a number of features of individual-level and collective legal mobilization. At the individual level, there is a broad conceptual and empirical base with which to understand how people come to define problems as legal or rights violations, the conditions under which people attempt to invoke law or other normative mechanisms to pursue their rights, how disputing evolves, and what explains its outcomes. Researchers concerned with collective legal mobilization have focused on the group processes through which law can lead to social change, tracing back this relationship to foundational – albeit contradictory – ways of conceiving law, rights, and change. Sociolegal frameworks drawing on social movement theory provide great insight into the mechanisms of and constraints on law-related social change as well as different ways of assessing outcomes.

Despite these accomplishments, previous legal mobilization research is constrained in multiple ways. First, there has been little attempt to link theory and research on legal mobilization across the individual and collective levels of analysis. Second, previous research tends to focus on individual- and movement-initiated (reactive) legal mobilization in liberal democracies and in single societies. Less is known about the (proactive) repression of legal mobilization in liberal democracies, legal mobilization in authoritarian regimes, or how law is constructed and/or invoked transnationally. Finally, research on legal mobilization, particularly at the collective level, is marked by an emphasis on studies of single contexts or movements thus constraining the assessment of broader patterns. Cutting-edge research on legal mobilization addresses each of these constraints in turn.

Earl (2009) uses the 'crosscutting' concept of 'trouble' to argue that despite different conceptual languages there is significant overlap in the concerns of researchers studying individual-level and collective legal mobilization. At the most abstract level, legal mobilization researchers are interested in how people come to define trouble, assign responsibility for it, whether they turn to law to do something about it, and the outcomes of these efforts. These dynamics are captured at the individual level in Felstiner et al.'s (1980–1981) classic trilogy of 'naming, blaming, and claiming' and at the collective level in the framing processes through which social movement actors make sense of situations that can 'mobilize potential adherents and constituents' (Snow and Benford, 1988: 198). The meanings of law are often constructed in these processes and supply important frames that help diagnose what a problem is, suggest particular strategies and remedies, and signal legitimacy and compelling moral identities. Multilevel analysis might lead to understanding the conditions under which individual mobilization becomes collective and vice versa (i.e., the problem of demobilization) and the relationship of these processes to features of the legal system relevant to a particular normative domain (e.g., ambiguous rights and procedures).

The proactive mobilization of law leads back to the distinction between 'repressive' and 'responsive' law offered by Nonet and Selznick (1978). The former exists when a legal system gives 'short shrift to the interests of the governed' (Nonet and Selznick, 1978: 29). Responsive law, most often recognized in the ideals of liberal democracy, exists when

a legal system selectively adapts to the social needs of a society within the boundaries of enlightened civility and reason (1978: 74–113). In this sense, all law is repressive at least part of the time and most systems are responsive at some points in their lifetimes. When law only protects those interests that are deeply intertwined with existing power and privilege, it is a repressive system. With respect to the repression of legal mobilization in liberal democracies, some researchers have analyzed the policing of street protests, finding that the types of strategies used by police to control protest events vary considerably and that different types of strategies have chilling effects on subsequent protests and other forms of policing have little or facilitative effects (Earl and Soule, 2010). Legal mobilization also has been investigated in the context of prisons, often characterized as the most repressive of liberal-democratic institutions. In a counterintuitive study, Calavita and Jenness (2013) find almost three-quarters of randomly sampled, California prison inmates had filed internal grievances contesting poor medical care, bad food, lack of physical safety, unfair disciplinary action, unpredictable transfers, or theft. This finding stands in contrast to long-standing empirical findings (discussed above) that vulnerable, impoverished, and stigmatized populations often face considerable barriers to mobilizing the law. Calavita and Jenness (2013: 52) argue that the extensive quasi-legal mobilization by prisoners "must be seen within [the] context of prison…as a conspicuously *legal* [total institution] where explicit rules govern every aspect of behavior [emphasis in the original]." Inmates' high rates of mobilization thus reflect the law-drenched prison context in which they are held and the broader responsive law context of a liberal democracy. In these contexts, inmates mobilize the prison's rules and procedures against itself, opening up possible sites for challenge and change.

A venerable part of the tradition of examining legal mobilization outside liberal democracies explores the overturning of repressive legislative codes and the establishment of constitutional courts in the aftermath of the collapse of fascist, communist, and apartheid regimes. The establishment of human rights associated with transnational forums, such as the United Nations Human Rights Council, the International Criminal Court, or the European Convention on Human Rights, has opened a wave of legal mobilization that sometimes challenges but is typically dependent upon state-based codes of law (Shapiro and Stone Sweet, 2002).

As researchers have begun to study legal mobilization by citizens within existing authoritarian regimes, they have sometimes found such regimes to be surprisingly responsive to particular issues as long as invoking the law to solve social problems does not threaten the overall legitimacy of the central state. For example, Michelson's (2007) survey-based research in China suggests that members of households in the least economically developed areas of China turn to the centralized official justice system to solve economic, neighborhood, or personal injury grievances at greater rates than more economically developed areas. But family ties moderate this relationship – those households with stronger political connections to governmental officials are more likely to use the legal system. Thus, social class and social ties matter as one climbs the 'dispute pagoda' in China just as they matter as one climbs the 'dispute pyramid' in liberal democracies, but in different ways. At the same time, the Chinese state can be quite 'ambivalent' about the desirability of litigation, sending mixed messages to citizens and other interested parties, for example, in the case of the emerging field of environmental litigation. As Stern (2013: 4) observes, a 'medley' of actions and policies by the Chinese government simultaneously signals promotion of the 'rule of law' *yet* 'control' of courts, and economic growth *and* environmental protection even in the face of considerable environmental degradation. In other authoritarian contexts, the problematic character of 'rights' via collective mobilization becomes especially salient. Chua's (2012) work on collective mobilization by Lesbian, Gay, Bisexual, and Transgender (LGBT) activists in Singapore points to the alternative ways that activists mobilize law through 'pragmatic resistance' that avoids confrontational tactics and rights litigation, thus skirting repression by the state.

Aside from work on legal mobilization in authoritarian regimes, cutting-edge research also has studied legal mobilization in transnational contexts. In some ways, this work harkens back to anthropological studies of law that sought to understand responses to normative breaches in societies across a broad span of human cultures (e.g., Nader and Todd, 1978; Roberts, 1979). Newer traditions of research of this type have emerged in three sometimes overlapping, sometimes distinct strains. The first seeks to understand the construction of transnational legal orders by American and European lawyers and other elites, such as the field of commercial arbitration that governs much of the disputing among global commercial interests, especially large corporations (Dezalay and Garth, 1996). To borrow from Galanter (1974), such legal orders only serve repeat players and have zealously guarded and privileged boundaries that prevent all but the most powerful from entering. The second approach has focused on the growth of legal orders from below that challenges elite and neoliberal legal orders via the mobilization of national courts and transnational advocacy networks (Sousa Santos and Rodríguez-Garavito, 2005). These efforts slip back and forth between the legitimacy of law and the sometimes ambiguous, popular legitimacies and illegitimacies of extralegal mobilization. The third strain seeks to understand the interplay of legal orders proffered by global governance institutions and their reception in local contexts. Merry (2005), for example, argues in the context of human rights and the prevention of gender violence that for human rights regimes to be effective at the local level, they must be framed in ways that articulate with local conditions and state power. Lack of attention to the local contexts in which human rights or other forms of global legality are introduced invites disaster, such as in the Sudan where human rights not only failed to empower the 'war-ravaged poor' but also created the conditions for further social conflict and strife (Massoud, 2013).

As sociolegal scholars come to know more about the dynamics and consequences of legal mobilization, it becomes imperative that contemporary qualitative studies be supplemented with analyses of broader patterns. The national random survey that formed the basis for CLRP, now more than 30 years old, provides one such model for the study of individual-legal mobilization, which has inspired smaller-scale surveys in a number of institutional contexts and countries (e.g., Michelson, 2007; Morrill et al., 2010). Other projects,

such as the American Bar Foundation's Civil Justice Infrastructure Mapping Project, have begun painting a 'holistic portrait' of the access services available to deliver civil legal assistance, especially for low-income people (Sandefur and Smyth, 2011). Yet, there has not been a national-level survey of legal mobilization in the US that brings together both the legal mobilization and the services sides of the puzzle. Researchers have also begun to move beyond single cases or movements to compare how and why collective mobilization – both legal and extralegal – emerges in local communities and how the social, legal, and political conditions in those communities influence its outcomes (McAdam and Boudet, 2012). To continue the fruitful exploration of transnational fields of legality, multimethod approaches that can negotiate multiple levels of analysis also will be necessary. All these efforts, whether qualitative or quantitative, need to be informed importantly by engagement with the historical context in which legal mobilization unfolds; for it is only through careful attention to process, pattern, and context that we will come to know the full range of possibilities and constraints of legal mobilization.

See also: Civil Law; Civil Rights; Common Law; Courts and Adjudication; Critical Race Theory; Globalization: Legal Aspects; Justice, Access To; Law and Courts in Authoritarian States; Law and Everyday Life; Law as an Instrument of Social Change; Law, Sociology of; Law: The Socio-Legal Perspective; Legal Culture and Legal Consciousness; Legal Pluralism; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Organizations and the Law; Race and the Law; Rights: Legal Aspects.

Bibliography

Abrego, L., 2011. Legal consciousness of undocumented Latinos: fear and stigma as barriers to claims making for first and 1.5 generation immigrants. Law and Society Review 45 (2), 337–369.

Albiston, C., 2010. Institutional Inequality and the Mobilization of the Family and Medical Leave Act: Rights on Leave. Cambridge University Press, Cambridge, England.

Best, R.K., Edelman, L.B., Krieger, L.H., Eliason, S.R., 2011. Multiple disadvantages: an empirical test of intersectionality theory in EEO litigation. Law & Society Review 45 (4), 991–1025.

Black, D., 1972. The mobilization of law. Journal of Legal Studies 2 (1), 125–149.

Black, D., 1976. The Behavior of Law. Academic Press, New York.

Black, D., 1980. The Manners and Customs of the Police. Academic Press, New York.

Black, D., 1993. The elementary forms of conflict management. In: Black, D. (Ed.), The Social Structure of Right and Wrong. Academic Press, San Diego, CA, pp. 70–94.

Calavita, K., Jenness, V., 2013. Inside the pyramid of disputes: naming problems and filing grievances in California prisons. Social Problems 60 (1), 50–80.

Carlin, J., 1966. Lawyers' Ethics: A Survey of the New York City Bar. Russell Sage Foundation, New York.

Carlin, J.E., Carlin, H.J., Messinger, S.L., 1966. Civil justice and the poor: issues for sociological research. Law & Society Review 1 (1), 9–90.

Chua, L.J., 2012. Pragmatic resistance, law, and social movements in authoritarian states: the case of gay collective action in Singapore. Law & Society Review 46 (4), 713–748.

Crenshaw, K., 1989. Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics. University of Chicago Legal Forum 140, 139–167.

Dezalay, Y., Garth, B., 1996. Dealing in Virtue: International Commercial Arbitration and the Construction of a Transnational Legal Order. University of Chicago Press, Chicago, IL.

Earl, J., 2009. When bad things happen: toward a sociology of troubles. Sociology of Crime, Law, and Deviance 12, 231–254.

Earl, J., Soule, S.A., 2010. The impacts of repression: the effect of police presence and action on subsequent protest rates. Research in Social Movements, Conflicts, and Change 30, 75–113.

Edelman, L.B., 1992. Legal ambiguity and symbolic structures: organizational mediation of civil rights law. Journal of the American Society 97 (6), 1531–1576.

Edelman, L.B., Suchman, M., 1999. Why the 'haves' hold court: speculations on the organizational internalization of law. Law and Society Review 33 (4), 941–991.

Edelman, L.B., Krieger, L.H., Eliason, S.R., Albiston, C.R., Mellema, V., 2011. When organizations rule: judicial deference to institutionalized employment structures. Journal of the American Society 117 (3), 888–954.

Ellickson, R.C., 1991. Order without Law: How Neighbors Settle Disputes. Harvard University Press, Cambridge.

Emerson, R.M., 2008. Responding to roommate troubles: reconsidering informal dyadic control. Law and Society Review 42 (3), 483–512.

Ewick, P., Silbey, S.S., 1998. The Common Place of Law: Stories from Everyday Life. University of Chicago Press, Chicago, IL.

Felstiner, W.L.F., Abel, R.L., Sarat, A., 1980–1981. The emergence and transformation of disputes: naming, blaming, and claiming…. Law and Society Review 15 (3–4), 631–654.

Galanter, M., 1974. Why the 'haves' come out ahead: speculations on the limits of legal change. Law and Society Review 9 (1), 95–160.

Genn, H.G., 1999. Paths to Justice: What People Do and Think about Going to Law. Hart Publishing, Oxford, England.

Handler, J.F., 1978. Social Movements and the Legal System: A Theory of Law Reform and Social Change. Academic Press, New York.

Harris, D., Maclean, M., Genn, H., Lloyd Bostock, S., Fenn, P., Corfield, P., Brittan, Y., 1984. Compensation and Support for Illness and Injury. Clarendon Press, Oxford, England.

Levitsky, S., 2008. What rights? The construction of political claims to American health care entitlements. Law and Society Review 42 (3), 551–589.

Llewellyn, K.N., Hoebel, E.A., 1941. The Cheyenne Way: Conflict and Case Law in Primitive Jurisprudence. University of Oklahoma Press, Norman, OK.

Marshall, A., 2005. Confronting Sexual Harassment: Law and Politics of Everyday Life. Ashgate Publishing, Burlington, VT.

Massoud, M.F., 2013. Law's Fragile State: Colonial, Authoritarian, and Humanitarian Legacies in Sudan. Cambridge University Press, Cambridge, England.

Mayhew, L., Reiss Jr, A.J., 1969. The social organization of legal contacts. American Sociological Review 34 (3), 309–318.

McAdam, D., McCarthy, J.D., Zald, M.N., 1996. Introduction: opportunities, mobilizing structures, and framing processes – toward a synthetic, comparative perspective on social movements. In: McAdam, D., McCarthy, J.D., Zald, M.N. (Eds.), Comparative Perspectives on Social Movements: Political Opportunities, Mobilizing Structures, and Cultural Framings. Cambridge University Press, Cambridge, England, pp. 1–20.

McAdam, D., Boudet, H.S., 2012. Putting Social Movements in Their Place: Explaining Opposition to Energy Projects in the United States, 2000–2005. Cambridge University Press, Cambridge, England.

McCann, M.W., 1994. Rights at Work: Pay Equity Reform and the Politics of Legal Mobilization. University of Chicago Press, Chicago, IL.

McCann, M.W., 2004. Law and social movements. In: Sarat, A. (Ed.), The Blackwell Companion to Law and Society. Blackwell, Oxford, England, pp. 506–522.

McCarthy, J.D., Zald, M.N., 1977. Resource mobilization and social movements: a partial theory. Journal of American Society 82 (6), 1212–1241.

Merry, S.E., 2005. Human Rights & Gender Violence: Translating International Law into Local Justice. University of Chicago Press, Chicago, IL.

Michelson, E., 2007. Climbing the dispute pagoda: grievances and appeals to the official justice system in rural China. American Sociological Review 72 (3), 459–485.

Morrill, C., 1995. The Executive Way: Conflict Management in Corporations. University of Chicago Press, Chicago, IL.

Morrill, C., Johnson, M., Harrison, T., 1998. Voice and context in simulated everyday legal discourse: the influence of sex differences and social ties. Law and Society Review 32 (3), 639–665.

Morrill, C., Edelman, L.B., Tyson, K., Arum, R., 2010. Legal mobilization in schools: the paradox of rights and race among youth. Law and Society Review 44 (3–4), 651–694.

Nader, L., Todd Jr., H.F. (Eds.), 1978. The Disputing Process – Law in Ten Societies. Columbia University Press, New York.

Nielsen, L.B., 2004. License to Harass: Law, Hierarchy, and Offensive Public Speech. Princeton University Press, Princeton, NJ.

Nonet, P., Selznick, P., 1978. Law and Society in Transition: Toward Responsive Law. Harper, New York.

Roberts, S., 1979. Order and Custom: An Introduction to Legal Anthropology. St. Martins Press, New York.
Rosenberg, G.N., 1991. The Hollow Hope: Can Courts Bring about Social Change? University of Chicago Press, Chicago, IL.
Sandefur, R.L., Smyth, A.C., 2011. Access across America: First Report for the Civil Justice Infrastructure Mapping Project. American Bar Foundation, Chicago, IL.
Sarat, A., Miller, R.E., 1980–1981. Grievances, claims, and disputes: assessing the adversary culture. Law and Society Review 15 (3–4), 525–566.
Scheingold, S., 1974. The Politics of Rights: Lawyers, Public Policy and Political Change. University of Michigan Press, Ann Arbor, MI.
Shapiro, M., Stone Sweet, A., 2002. On Law, Politics, & Judicialization. Oxford University Press, Oxford, England.
Siegal, R., 2006. Constitutional culture, social movement conflict and constitutional change: the case of the de facto era. California Law Review 94 (5), 1323–1419.
Snow, D.A., Benford, R.D., 1988. Ideology, frame resonance, and participation mobilization. International Journal of Social Movement Research 1 (1), 197–217.
Sousa Santos, B., Rodríguez-Garavito, C. (Eds.), 2005. Law and Globalization from below: Towards a Cosmopolitan Legality. Cambridge University Press, Cambridge, England.
Stern, R.E., 2013. Environmental Litigation in China: A Study in Political Ambivalence. Cambridge University Press, Cambridge, England.

Law, Sociology of

Carol A Heimer, Northwestern University, Evanston, IL, USA; and American Bar Foundation, Chicago, IL, USA
Jaimie Morse, Northwestern University, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G. Rocher, volume 12, pp. 8544–8548, © 2001, Elsevier Ltd.

Abstract

Although early sociologists studied law as part of industrialization and modernization, a distinct subfield emerged after the mid-twentieth century. This article focuses on two key bodies of research in the sociology of law: (1) studies of the formal institutions of law and (2) studies of the circulation, reception, and consumption of law, broadly conceived, beyond formal institutions of law. Using the shorthand phrases 'law within the courts' and 'law beyond the courts', this article summarizes key questions and findings of this work, describes its influence on later studies of law and globalization, and assesses its value added for sociolegal studies.

Introduction

Although the implicit and explicit study of law can be found in the writings of many early sociologists, the demarcation of the sociology of law as a distinct subfield within sociology began largely after the mid-twentieth century. The Research Committee on the Sociology of Law of the International Sociological Association, created in 1962, and the Law and Society Association, founded in 1964, were vitally important in bringing together social scientists conducting research on law. Yet the gradual nature of field development is signaled by the fact the American Sociological Association Section on the Sociology of Law was formed only in 1992 (the Section on Crime, Law, and Deviance had been in existence for some time).

As seems to have occurred in other areas of sociology, the coalescing of this new subfield began with a focus on the core institutions and the main groups of people engaged with those institutions. Just as sociologists of medicine initially studied hospitals, doctors' offices, medical schools, doctors, nurses, and patients and the sociologists of science initially studied laboratories, scientists, and technicians, so the sociologists of law looked first at courts (at multiple levels), juries, judges, lawyers, paralegals, and clients. And as occurred in other maturing fields, sociologists gradually began to look beyond these core institutions, core actors, and core activities, thinking more broadly about how law and legal actors function outside courts and law offices. This article focuses especially on these two key bodies of research in the sociology of law: (1) studies of the formal institutions of law, conducted especially from the mid-1960s through the early 1980s (but continuing to the present) and (2) studies of the circulation, reception, and consumption of law, broadly conceived, beyond these formal institutional settings, conducted mainly since the early 1980s. Referring to these two bodies of research with the shorthand phrases 'law within the courts' and 'law beyond the courts', this article summarizes the key questions and findings of this work in Sections Law 'Within the Courts': Research on Formal Law, Legal Institutions, and Legal Actors and Law 'Beyond the Courts': Research on the Circulation, Reception, and Consumption of Law. The Section Globalization, Governance, and Transnational Legal Transfers describes investigations of the relationship between law and globalization, particularly in the late twentieth century. The Section Conclusions: Value Added by the Sociology of Law assesses the value added from research on the sociology of law.

Law 'Within the Courts': Research on Formal Law, Legal Institutions, and Legal Actors

The early beginnings of the sociology of law can be traced to the brilliant and still valuable work of Émile Durkheim, Max Weber, and Karl Marx. Primarily concerned with such issues as the growth of capitalism, the division of labor, or the exploitation of workers, these early sociologists were interested in law as one of the constellation of forces at work in industrialization and modernization. Although he shared these concerns with what law contributed to the organization of society, Talcott Parsons (1954) was also one of the first scholars to distinguish professions from other kinds of work and to emphasize the unique orientations and contributions of professionals such as lawyers and judges. This emphasis on legal actors themselves and the nature of their work helped propel the sociology of law into its next phase.

Animating the research on the sociology of law, then, were two sets of questions coming, on the one hand, from the legal academy and, on the other, from sociology. From the legal academy came questions about the gap between 'law on the books' and 'law in action', a discrepancy first noted by the legal realists of the early twentieth century (Pound, 1910). From sociology came questions about the organization of legal work, the careers of lawyers, and the capacity of law to redress injustice and mitigate inequality. Inspired by sociologists' questions about professions and legal scholars' questions about how law actually functions, sociologists – now beginning to identify themselves explicitly as sociologists of law – focused initially on legal sites, legal actors, and legal work. Over time, in a loose professional division of labor, some of these research areas came to be dominated by scholars from other fields. Jury studies and studies of procedural justice, for instance, have mostly been carried out by social psychologists; studies of court decisions

and law making have mostly been conducted by political scientists. Sociologists have continued to dominate, though, in studies of law firms, the legal profession, and the relationship between law and social inequality.

From a sociological perspective, the existence of the gap between law on the books and the practice of law is not particularly surprising. People sometimes prefer to do things informally rather than relying on formal legal tools such as contracts (Macaulay, 1963), and the law often is not implemented exactly as it was written. Notably, these findings could be anticipated in a discipline that has emphasized the discrepancies between formal organizations and the less formal workings of those same entities.

But although sociologists had examined the gap between plans and action and between formal and informal processes in a wide variety of settings, research on the gap in law has nevertheless been significant. Because law often works by 'double institutionalizing' (Bohanan, 1965) existing rules, laws often are supported and enforced by more than one institutional system; these extra resources might be expected to decrease the gap. Because law is often billed as the last hope for the weak and vulnerable, gaps may be especially consequential. Yet, at best, law seems only marginally better at narrowing the gap between prescription and implementation. But beyond demonstrating the universality of the gap, scholars have also documented and described variations in the nature of the gap (Gould and Barclay, 2012). In core legal institutions such as the civil and criminal justice systems, scholars have demonstrated that justice works differently for different kinds of people (Feeley, 1979; Provine, 2011). In essence, these studies join together observations about the workings of organizations and work groups with observations about social inequality: formal statements of rules are not good descriptions of how action unfolds, but as rules are implemented, those who have abundant resources are more likely than those with scanty resources to be able to make the rules work to their advantage. These findings are brought together in such summary statements as Black's (1976) argument that more law is applied to people lower in the social hierarchy, in Galanter's (1974) piece on the 'haves' and 'have nots', and in Lempert and Sanders' (1986) discussion of legal endowments (which can be read as an attempt to account for variations in the nature and magnitude of the gap). Among other things, the sociological literature on the gap shows that although a gap may be a problem for some, it can be an opportunity for others (from the point of view of institutionalists in sociology, official forms can provide cover). Moreover, the gap is in part created by the sometimes valorized, sometimes disparaged use of discretion, perhaps especially in the government agencies and regulatory bodies that translate and implement law (Lipsky, 1980).

In thinking about the relationship between inequality and law, sociologists of law have also looked closely at the configuration of law firms and at the careers of lawyers and judges. Parsons included lawyers among the professionals who were differentiated from other workers both by the importance of their contributions to the smooth functioning of society and by the norms that guided their work. The expectation was that they could be counted on to pursue the interests of their clients in a disinterested way and that external regulation could be relatively light because professionals would monitor their own and their colleagues' work, which in any case could not so easily be judged by others who lacked their training and expertise. Over the years, sociological research suggested that professionals, including lawyers, were not quite so disinterested as Parsons believed. A few indicators make the point: legal education is expensive and not easily accessible to any but an elite and although many law students enter law school with the intention of doing public interest law, far fewer have this aspiration by the time they graduate (Granfield, 1992; Schleef, 2005); minorities and women are less likely to be represented in the fraternity of lawyers, and when they do enter the profession, tend to have careers that are systematically different than those of their nonminority, male colleagues (Epstein, 1981; Kay and Gorman, 2008); legal services are costly and more available to the rich than the poor, reproducing and exacerbating existing inequalities (Sandefur, 2008); lawyers, like other professionals, prefer to work on esoteric problems (such as those raised in corporate or constitutional law) rather than mundane ones (such as those raised in family law or probate law (Sandefur, 2001)); lawyers who work on the problems of the rich and powerful or the problems of corporations are paid better than those who attend to individuals, especially poor ones (Sandefur, 2001; Heinz et al., 2005). If anything, the location of legal work in firms magnifies these differences. As research on law firms shows, legal work tends to be split into 'hemispheres' with one group of practitioners focused on individual clients and the other focused on organizations (Heinz and Laumann, 1982; Heinz et al., 2005).

Law 'Beyond the Courts': Research on the Circulation, Reception, and Consumption of Law

Beginning roughly in the 1980s, the sociology of law moved 'beyond the courts'. Where earlier work examined the practice of law within formal legal institutions and examined the role of lawyers and judges in the interpretation, application, and enforcement of formal law, later work took as its starting point understandings or contests over law as they emerge outside of formal legal institutions. Rather than accepting the internalist demarcations of the boundary between what counts as formal law and what does not, sociologists of law instead problematized the question of what comes to be understood as 'legal' and 'not legal'. They studied both the presence and absence of law in a variety of arenas to determine if, when, and how law did – and just as importantly, did not – appear in lay accounts of the disputes, grievances, and administrative transactions of daily life. Two developments stand out in this exceptionally rich vein of research on law 'beyond the courts'.

Lay Engagement, Legal Consciousness, and the Constitutive Perspective

In the spirit of the 'cultural turn' to the study of meanings and symbols that animated humanistic and social science scholarship in the 1980s, sociologists of law looked carefully at lay engagement with law (Silbey, 2005). They asked how lay understandings of law emerge and how legal institutions are able to retain their authority in the face of persistent gaps between legal doctrine and legal practice (Silbey and Bittner, 1982; Merry, 1985;

Ewick and Silbey, 1998; Nielsen, 2004). These studies of legal consciousness and the associated constitutive perspective emphasize the inseparability of the social and cultural in what comes to be understood as legal.

A central research question in this genre was how conflicts in everyday life may or may not come to be understood as legal disputes or as requiring legal intervention, encapsulated in the so-called dispute pyramid and associated 'naming, shaming, and blaming' rubric (Felstiner et al., 1980). Rather than treating law as 'black letter' law codified in formal legal doctrine and enforced by experts (lawyers) in formal legal institutions that exist at some remove from society, sociologists instead took a law *in* society approach to show that both legal and nonlegal actors create, interpret, and selectively use rules, regulations, and legal doctrine in particular historical, cultural, and political contexts. Taken from the vantage point of the everyday, what comes to be understood as within the purview of the law informs what law can (and cannot) do and, therefore, is a key factor mediating the observed gap between law on the books and law in action that animated scholarship on the workings of law in formal legal settings.

Interpretation, Implementation, and Enforcement: Institutionalization and Organizational Practices

As sociologists of law explored everyday uses of law, they also began to train attention on the mesolevel – the organizations, which are the central sites of adoption, adaptation, and resistance to new laws, rules, and regulations. Scholars have explored the role of law in regulating activity in a wide range of institutional spheres, including medicine (Heimer, 1999; Horowitz, 2012; Shapiro, 2012), science (Jasanoff, 1995; Silbey and Ewick, 2003), insurance (Heimer, 1985; Baker, 2010), and finance (Halliday and Carruthers, 2007), among others. The diffusion of law and legalistic modes of governance into these diverse settings have brought a good deal of extra compliance work, but have not often brought the leveling of the playing field or the decreases in injustice that Selznick and his colleagues (Selznick et al., 1969) might have anticipated.

This mesolevel research drew on neoinstitutionalists' research on the diffusion and institutionalization of new laws, regulations, and associated practices in professional fields and organizations. If enacted, do new laws and regulations achieve their intended effects? A key insight of neoinstitutional theorists is that the adoption of new policies or practices can be purely performative or ceremonial (Meyer and Rowan, 1977), a means to signal credibility, legitimacy, or superficial compliance with new regulations or norms. In fact, accepted professional norms or organizational practices can powerfully persist even as organizations seek to innovate, sometimes leading to 'institutional isomorphism' among a particular population of organizations with a shared purpose, function, or market (DiMaggio and Powell, 1983). This early research spurred the study of the decoupling of organizational policies from actual practices.

Of relevance to sociolegal inquiry is the cautionary message that there is no simple relationship between the adoption of new laws, policies, and regulations and changes in organizational practices. However, other studies have shown that initial symbolic adoption can later lead to instrumental effects, particularly if a profession or constituency develops to support and institutionalize a new policy or practice (Heimer, 1999; Dobbin, 2011). For example, even after the legal pressures for adoption of corporate affirmative action policies had waned, human resources personnel successfully made the case for a link between greater efficiency and greater workforce diversity (Dobbin, 2011). Organizational policies and practices also may influence common understandings of the problem the law was designed to address as well as the substance of the laws and regulations to which organizations are subject. For example, organizations in the United States anticipated employment discrimination lawsuits and preemptively created internal Equal Employment Opportunity Grievance Procedures. Because definitions of compliance were worked out in the back and forth between courts, regulatory agencies, and regulated organizations, these organizational policies ultimately served as a shield against corporate liability in court and effectively helped to constitute the nature and substantive content of laws designed to prevent and sanction discrimination in employment decisions (e.g., hiring and firing) and work conditions (e.g., sexual harassment) (Edelman et al., 1999, 2011). In effect, law is endogenous and organizations cocreate the meaning of the laws and regulations to which they are subject. If anything, organizations seem to have had the upper hand in this process of cocreation. Nelson and Bridges' (1999) painstaking study of organizations that were sued shows that courts all too often uncritically accepted employing organizations' market explanations for pay inequities, in effect legalizing gender-based wage inequality.

Seeking to explain legal change, scholars looked at the activities of the legal and nonlegal change agents who helped craft new laws, regulations, and organizational practices, both within and beyond courts. Institutional entrepreneurs created new organizational forms (DiMaggio, 1988); cultural entrepreneurs crafted new concepts to highlight existing problems (such as sexual harassment) and new rules and laws to reshape behavior (Saguy, 2003); and professionals functioned much like social movement activists by pushing for new organizational policy, structure, and practice within their areas of expertise (Davis et al., 2005). Because new organizational forms codify contested definitions of law and social problems, organizational settings offer sites in which scholars can learn how everyday decision making and interactions help produce higher-order social consequences, often but not always reproducing existing social inequalities.

Globalization, Governance, and Transnational Legal Transfers

More recently, sociologists of law have contributed to research on late twentieth-century processes of globalization. Of course, the global context has always shaped the research agenda of anthropologists of law and other sociolegal scholars studying legal pluralism, colonization, decolonization, and postcolonial legal reform. In contrast, sociologists of law became interested in global processes much later and have drawn on their disciplinary heritage to trace the exportation and transnational movement of legal professions, formal legal institutions, and legal regimes from the global North to the global South.

Sociologists of law have taken up studies of change in formal law, legal institutions, and the legal profession, both in comparative and global contexts. Actors, organizations, and professions make up transnational networks that facilitate the global diffusion of particular laws and legal regimes with their attendant legal and cultural norms. Similar to studies of the legal profession in studies of law 'within the courts' and change agents in studies of law 'beyond the courts', much work in this area initially focused on elite lawyers (Munger, 2012) who lead or challenge the adoption of new laws, regulations, or governance structures at the national level but who simultaneously influence and are influenced by developments at the transnational level (Dezalay and Garth, 1996, 2002, 2010). Later studies of domestic and transnational social movements have highlighted the role of nonelites and activists in legal and constitutional reform (Kay, 2011; Klug, 2000; Merry, 2006; Munger, 2008–09).

Legal sociologists have also carried their interest in understandings, uses, and impacts of law beyond courts and formal legal institutions into comparative and global research. Similar to studies of adoption and institutionalization of new rules within organizations based in the United Studies, studies of innovation and diffusion have animated scholarship on the globalization of law (Halliday and Osinsky, 2006) and public policy (Dobbin et al., 2007). Transnational diffusion inevitably involves resistance, adaptation, and innovation by domestic actors who often must do substantial work to make new laws, rules, norms, and technologies – often formulated in the global North – relevant in new environments – often in the global South (Klug, 2002; Merry, 2006; Heimer, 2013). For example, human rights activists 'vernacularize' purportedly universal norms, translating them into local discourses and using them to make more specific claims in their own countries (Merry, 2006). Such domestic actors exercise agency by resisting the imposition of foreign norms and simultaneously using these same norms strategically to further their own domestic agendas.

Just as sociologists of law demonstrated that legal and nonlegal processes enacted beyond the courts domestically often influenced processes enacted within national courts and other legal institutions, so scholars have documented similar feedback effects spanning national borders. In global bankruptcy regimes, for instance, Halliday and Carruthers (2007) find 'recursivity' at both national and transnational levels. Likewise, the 'hybridity' of new constitutions crafted after democratic transitions incorporate both domestic and international content and in turn create new exemplars for others to use (Klug, 2002).

Conclusions: Value Added by the Sociology of Law

As a subdiscipline, the sociology of law has fruitfully joined research on the central institutions of law with research that acknowledges the complexity of life outside these main sites of formal legal activity. But what special insights into sociolegal processes have come from these sociological studies? And, in turn, what new insights has the sociology of law brought to the discipline of sociology?

The sociology of law draws on sociology's longstanding study of complex organizations, professions, and social inequality to gain insight into what happens in courts, regulatory agencies, and law firms and the institutional spheres in which law is applied (e.g., hospitals, laboratories, insurance companies). Actors at work within these organizations, including lawyers, judges, and police, but also other nonlegal personnel (doctors, human resource personnel, administrators) charged with enforcing the law beyond the courts, variously cooperate, collaborate, resist, and adapt the law's requirements in ways that simultaneously apply, alter, and elaborate the law, thereby often influencing both the substantive content of formal law and the extent to which it achieves its intended effects. Clearly for law to do its work, it has to travel beyond the courts into other institutional settings. But when it arrives in those settings, it is deeply shaped by the day-to-day, practical constraints of organizations. Translating law into organizational routines fundamentally alters the meaning of law. Moreover, law in action is also deeply shaped by the fact that the professionals who do law-related work are not simply neutral instruments of the law, but are also people who care about compensation, prestige, challenging work, and the interests of their professions and their employers. And although law is sometimes billed as a tool to help the disadvantaged (and of course it is sometimes deployed that way), it often instead shores up inequalities. Actors with more resources or more experience using legal tools are more able to use law to their advantage. Examining law from sociological perspective thus gives us important insights about where gaps between law on books and law in action are especially prominent and why.

The sociology of law also informs core questions and concerns within the discipline of sociology. Whether training attention on the formal institutions of law (courts, police, law firms, regulatory agencies), or formal settings in which law is applied (hospitals, laboratories, corporations), work in this subfield has demonstrated that the importation of laws, rules, regulations, and legalistic ways of organizing social activity is deeply consequential. Clearly, law on the books differs from law in action with highly consequential implications for realizing justice and equity in society more broadly. Legal actors and nonlegal actors within and outside of formal legal institutions exercise discretion in the interpretation and application of formal law in a variety of domains; they thereby contribute to its uneven enforcement and participate in constructing the meaning and content of the law in practice. Individual lawyers, in aggregate, constitute a profession stratified based on differential prestige, training, occupation, and demographic characteristics much like any other profession, which in turn has implications for access to justice more broadly. Whether the diffusion of legalistic approaches has increased access to justice is debatable. What is clear, though, is that legalization has required organizations to add legally trained staff, create new positions and training programs for compliance workers, and establish new routines to track rules and demonstrate compliance (Heimer and Petty, 2010).

Taken together, work in this subfield suggests that formal institutions of law are sites in which broader inequalities can be challenged, but also reinforced, if not amplified. It also suggests that assessing the efficacy and impacts of new laws necessitates

looking outside the formal institutions of law to other institutional spheres that are central to the enforcement of legal mandates. Research in the sociology of law seeks to contextualize and explain the locally contingent and constructed nature of law, its meanings, and its enforcement. Based on insights derived through careful attention to practice and behavior both within and outside of the formal institutions of law, it demonstrates that there is no universal, neutral, or purely mechanical application of the law as it is written – thereby challenging popular ideas about how the law does or should work. In this way, close attention to law in practice reveals important gaps that fundamentally challenge some of the most foundational assumptions about the structure and application of law.

Extending the insights of the subfield to consider law in comparative and global contexts, sociologists of law continue to highlight the uneven distribution of legal resources and access to formal institutions of law, as well as the increasingly complex transfers that move law across both geographical and institutional boundaries, further complicating the insights developed largely in the context of the industrialized countries of North America and Western Europe. Unequal governance systems facilitate and condition the movement of laws, and norms move in tandem with resources from the global North to the global South (Silbey, 1997). Although South–South and South–North transfers do occur, they have less impact on governance and receive a good deal less attention. Like other 'globalized localisms' (Santos, 1995), law and legal norms have particular points of origin and serve the interests of particular power holders yet come to be described as global or universal once they diffuse transnationally. In this way, the unequal power relations that undergird many such transfers are obscured.

Many standardized tools and practices are described by their proponents as technically superior, neutral, and universal solutions. The claims on law's behalf are made especially strongly, suggesting that it is seen as the most universal of universal tools. Yet neutral, universal tools still have to be used to solve particular problems in particular contexts. It is in clarifying the importance of those details about differences among users, settings, and problems and how they are matched with the alleged universality of law that sociology of law makes its most important contribution.

See also: Disputes, Social Construction and Transformation of; Gender and the Law; Globalization: Legal Aspects; Injustice: Legal Aspects; Justice, Access To; Law and Everyday Life; Law and Society: Development of the Field; Law and Society; Law as Constitutive; Law as an Instrument of Social Change; Law, Mobilization of; Law: Anthropological Aspects; Law: History of Its Relation to the Social Sciences; Law: The Socio-Legal Perspective; Lawyers: Social Organization of the Profession; Legal Culture and Legal Consciousness; Legal Realism; Organizations and the Law; Postcolonial Law.

Bibliography

Baker, T., 2010. Insurance in sociolegal research. Annual Review of Law and Social Science 6, 433–447.

Black, D., 1976. The Behavior of Law. Yale University Press, New Haven, CT.

Bohanan, P., 1965. The differing realms of the law. American Anthropologist 67, 33–42.

Davis, G.F., McAdam, D., Scott, W.R., Zald, M.N., 2005. Social Movements and Organization Theory. Cambridge University Press, Cambridge, UK.

Dezalay, Y., Garth, B., 1996. Dealing in Virtue: International Commercial Arbitration and the Construction of a Transnational Legal Order. University of Chicago Press, Chicago, IL.

Dezalay, Y., Garth, B., 2002. The Internationalization of Palace Wars: Lawyers, Economists, and the Contest to Transform Latin American States. University of Chicago Press, Chicago, IL.

Dezalay, Y., Garth, B., 2010. Asian Legal Revivals: Lawyers in the Shadow of Empire. University of Chicago Press, Chicago, IL.

DiMaggio, P., 1988. Interest and agency in institutional theory. In: Zucker, L.G. (Ed.), Institutional Patterns and Organizations: Culture and Environment. Ballinger Publishing Company, Cambridge, MA, pp. 3–21.

DiMaggio, P.J., Powell, W.W., 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. American Sociological Review 48, 147–160.

Dobbin, F., 2011. Inventing Equal Opportunity. Princeton University Press, Princeton, NJ.

Dobbin, F., Simmons, B., Garrett, G., 2007. The global diffusion of public policies: social construction, coercion, competition, or learning? Annual Review of Sociology 33, 449–472.

Edelman, L.B., Uggen, C., Erlanger, H.S., 1999. The endogeneity of legal regulation: grievance procedures as rational myth. American Journal of Sociology 105, 406–454.

Edelman, L.B., Krieger, L.H., Eliason, S.R., Albiston, C.R., Mellema, V., 2011. When organizations rule: judicial deference to institutionalized employment structures. American Journal of Sociology 117, 888–954.

Epstein, C.F., 1981. Women in Law. Basic Books, New York.

Ewick, P., Silbey, S.S., 1998. The Common Place of Law: Stories from Everyday Life. University of Chicago Press, Chicago, IL.

Feeley, M.M., 1979. The Process is the Punishment: Handling Cases in a Lower Criminal Court. Russell Sage Foundation, New York.

Felstiner, W.L.F., Abel, R.L., Sarat, A., 1980. The emergence and transformation of disputes: naming, blaming, claiming. Law and Society Review 15, 631–654.

Galanter, M., 1974. Why the haves come out ahead: speculations on the limits of legal change. Law and Society Review 9, 95–160.

Gould, J.B., Barclay, S., 2012. Mind the gap: the place of gap studies in sociolegal scholarship. Annual Review of Law and Social Science 8, 323–335.

Granfield, R., 1992. Making Elite Lawyers: Visions of Law at Harvard and Beyond. Routledge, New York.

Halliday, T.C., Carruthers, B.G., 2007. The recursivity of law: global norm making and national lawmaking in the globalization of corporate insolvency regimes. American Journal of Sociology 112, 1135–1202.

Halliday, T.C., Osinsky, P., 2006. Globalization and law. Annual Review of Sociology 32, 447–470.

Heimer, C.A., 1985. Reactive Risk and Rational Action: Managing Moral Hazard in Insurance Contracts. University of California Press, Berkeley, CA.

Heimer, C.A., 1999. Competing institutions: law, medicine, and family in neonatal intensive care. Law and Society Review 33, 17–66.

Heimer, C.A., 2013. Resilience in the middle: contributions of regulated organizations to regulatory success. Annals of the American Academy of Political and Social Science 649, 139–156.

Heimer, C.A., Petty, J., 2010. Bureaucratic ethics: IRBs and the legal regulation of human subjects research. Annual Review of Law and Social Science 6, 601–626.

Heinz, J.P., Laumann, E.O., 1982. Chicago Lawyers: The Social Structure of the Bar. Russell Sage Foundation, New York.

Heinz, J.P., Nelson, R.L., Sandefur, R.L., Laumann, E.O., 2005. Urban Lawyers: The New Social Structure of the Bar. University of Chicago Press, Chicago, IL.

Horowitz, R., 2012. In the Public Interest: Medical Licensing and the Disciplinary Process. Rutgers University Press, New Brunswick, NJ.

Jasanoff, S., 1995. Science at the Bar: Law, Science, and Technology in America. Harvard University Press, Cambridge, MA.

Kay, F., Gorman, E., 2008. Women in the legal profession. Annual Review of Law and Social Science 4, 299–332.

Kay, T., 2011. Legal transnationalism: the relationship between transnational social movement building and international law. Law and Social Inquiry 36, 419–454.

Klug, H., 2000. Constituting Democracy: Law, Globalization, and South Africa's Political Reconstruction. Cambridge University Press, New York.

Klug, H., 2002. Hybrid(ity) rules: creating local law in a globalized world. In: Dezalay, Yves, Garth, Bryant G. (Eds.), Global Prescriptions: The Production, Exportation,

and Importation of a New Legal Orthodoxy. University of Michigan Press, Ann Arbor, MI, pp. 276–305.
Lempert, R.O., Sanders, J., 1986. An Invitation to Law and Social Science: Desert, Disputes, and Distribution. University of Pennsylvania Press, Philadelphia, PA.
Lipsky, M., 1980. Street-Level Bureaucracy: Dilemmas of the Individual in Public Services. Russell Sage Foundation, New York.
Macaulay, S., 1963. Non-contractual relations in business: a preliminary study. American Sociological Review 28, 55–76.
Merry, S.E., 1985. Concepts of law and justice among working class Americans. Legal Studies Forum 9, 59–71.
Merry, S.E., 2006. Human Rights and Gender Violence: Translating International Law into Local Justice. University of Chicago Press, Chicago, IL.
Meyer, J.W., Rowan, B., 1977. Institutional organizations: structure as myth and ceremony. American Journal of Sociology 83, 340–363.
Munger, F., 2008–09. Globalization, investing in law, and the careers of lawyers for social causes – taking on rights in Thailand. New York Law Review 53, 745–804.
Munger, F., 2012. Globalization through the lens of palace wars: what elite lawyers' careers can and cannot tell us about globalization of law. Law and Social Inquiry 37, 476–499.
Nelson, R.L., Bridges, W.P., 1999. Legalizing Gender Inequality: Markets and Unequal Pay for Women in America. Cambridge University Press, Cambridge, UK.
Nielsen, L.B., 2004. License to Harass: Law, Hierarchy, and Offensive Public Speech. Princeton University Press, Princeton, NJ.
Parsons, T., 1954. A sociologist looks at the legal profession. In: Essays in Sociological Theory, Revised Edition. Free Press, New York, pp. 370–385.
Provine, D.M., 2011. Race and inequality in the war on drugs. Annual Review of Law and Social Science 7, 41–60.
Pound, R., 1910. Law in books and law in action. American Law Review 44, 12–36.
Saguy, A.C., 2003. What is Sexual Harassment? University of California Press, Berkeley, CA.
Sandefur, R.L., 2001. Work and honor in the law: prestige and the division of lawyers' labor. American Sociological Review 66, 382–403.
Sandefur, R.L., 2008. Access to civil justice and race, class, and gender inequality. Annual Review of Sociology 34, 339–358.
Santos, B. de S., 1995. Toward a New Common Sense: Law, Science and Politics in the Paradigmatic Transition. Routledge, New York.
Schleef, D.J., 2005. Managing Elites: Socialization in Law and Business Schools. Rowman and Littlefield Publishers, Lanham, MD.
Selznick, P., Nonet, P., Volmer, H., 1969. Law, Society, and Industrial Justice. Russell Sage Foundation, New York.
Shapiro, S.P., 2012. Conflict of interest at the bedside: surrogate decision making at the end of life. In: Peters, A., Handschin, L. (Eds.), Conflict of Interest in Global, Public and Corporate Governance. Cambridge University Press, Cambridge, UK, pp. 334–354.
Silbey, S.S., 1997. Let them eat cake: globalization, postmodern colonialism, and the possibilities of justice. Law and Society Review 31, 207–236.
Silbey, S.S., 2005. After legal consciousness. Annual Review of Law and Social Science 1, 323–368.
Silbey, S.S., Bittner, E., 1982. The availability of law. Law and Policy 4, 399–434.
Silbey, S.S., Ewick, P., 2003. The architecture of authority: the place of law in the space of science. In: Sarat, A., Douglas, L., Umphrey, M. (Eds.), The Place of Law. University of Michigan Press, Ann Arbor, MI, pp. 75–108.

Relevant Websites

http://www.asanet.org/sections/law.cfm – The American Sociological Association Section on the Sociology of Law.
http://www.lawandsociety.org/ – The Law and Society Association (USA).
http://www.isa-sociology.org/rc12.htm – The Research Committee on the Sociology of Law of the International Sociological Association.
http://www.slsa.ac.uk/ – The Socio-Legal Studies Association (UK).

Law: The Socio-Legal Perspective

Lauren Edelman, University of California - Berkeley, Berkeley, CA, USA
Marc Galanter, University of Wisconsin - Madison, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The sociolegal perspective on law portrays law as dynamically intertwined with society, politics, markets, science and technology, culture, and other social institutions. In this view, law is seen as a set of socially institutionalized norms of varying degrees of formality, ranging from statutes and judicial decisions at the formal end of the spectrum to socially institutionalized norms and cultural understandings that become embedded in everyday life at the informal end of the spectrum. Law is envisioned as deeply embedded in society rather than as autonomous in relation to the citizens and organizations that it is designed to regulate. The sociolegal perspective tends to emphasize three broad dimensions along which law matters to social life: regulatory, procedural, and constitutive.

Introduction

The sociolegal perspective is our portmanteau term for a set of diverse but overlapping views of legal phenomena and their settings. It originally evolved in reaction to what was perceived as an overly formalistic emphasis in law schools and legal analysis on law as a rational set of edicts emanating from the state, applied by the courts, and followed by citizens. As part of the 'Law and Society Movement' of the mid-1960s, sociolegal scholars offered a broader and more socially and empirically grounded understanding of law. The sociolegal perspective understands law as much broader than the statutes, judicial decisions, and administrative regulations that are the primary focus of law schools and practicing lawyers, and as dynamically intertwined with society, politics, markets, science and technology, culture, and other social institutions. Generally, law is viewed as a set of socially institutionalized norms of varying degrees of formality, ranging from statutes and judicial decisions at the formal end of the spectrum to socially institutionalized norms and cultural understandings that become embedded in everyday life at the informal end of the spectrum. Law is envisioned as deeply embedded in society rather than as autonomous in relation to the citizens and organizations that it is designed to regulate. Law varies substantially across social contexts, moreover. Not only do formal laws differ across nations and subunits of nations, but the understandings and implementations of law vary across contexts even when formal laws look similar.

The boundaries of the sociolegal perspective are fluid and evolving, varying somewhat across disciplinary orientations. Political scientists tend to emphasize rules in the public sphere, focusing on judicial decision making (Epstein and Knight, 1998; Segal and Spaeth, 2002), as well as voting (Burch, 2011; Erikson, 1995), street-level bureaucrats in public agencies (Lipsky, 1980; Maynard-Moody and Musheno, 2012), the politics of criminal justice (Feeley, 1992); constitutions, constitutionalism, and constitutional courts (Elkins et al., 2009; Klug, 2000; Moustafa, 2003); and 'soft law' (Abbott and Snidal, 2000; Bothe, 1980; Chinkin, 1989; Gruchalla-Wesierski, 1984; Seidl-Hohenveldern, 1999; Tammes, 1983). Sociologists tend to emphasize legal meaning making by citizens and the subtle influence of 'legal consciousness' in everyday life (Abrego, 2011; Albiston, 2005; Engel and Munger, 2003; Ewick and Silbey, 1998; Merry, 1990; Nielsen, 2000; Sarat, 1990), governance processes in private organizations and other nonstate entities (Edelman et al., 1993; Edelman and Suchman, 1997; Edelman et al., 1999; Nelson and Bridges, 1999; Stryker, 2000; Suchman and Cahill, 1996), dispute resolution and barriers to the mobilization of justice (Felstiner et al., 1980; Michelson, 2007; Miller and Sarat, 1980; Morrill et al., 2010; Nielsen and Nelson, 2005; Albiston et al., 2014), the operation of power and inequality in legal processes and institutions (Curran, 1977; Galanter, 1974; Mayhew and Reiss, 1969) and in the legal profession (Galanter and Palay, 1994; Handler, 1967; Heinz et al., 2005; Rosenberg et al., 1993), and the limits of law in remedying inequalities on the basis of class, race, gender, and other social dimensions (Albiston, 2010; Best et al., 2011; Crenshaw, 1991; Frohmann and Mertz, 1994; Gómez, 2000, 2010; Gorman, 2005; Haney-López, 2007; Obasogie, 2013; Pager, 2007; Suk, 2010). Psychologists often study jury behavior (Diamond and Casper, 1992; Diamond and Rose, 2005) and perceptions of justice and legal institutions (Tyler, 1990; Tyler and Huo, 2002), whereas economists tend to focus on the efficiency of various formal legal rules and rule-making regimes (Becker, 1968; Cooter and Rubinfeld, 1989; Donahue and Heckman, 1991; Posner, 1975). Increasingly, humanities scholars have entered into the sociolegal arena and have brought attention to the operation of law through texts, social constructions, and meaning making outside of the legal realm (Constable, 2009; Lokaneeta, 2010; Reichman, 2009).

These disciplinary orientations do not have hard boundaries, however, and sociolegal scholarship is characterized by interdisciplinary innovation and cross-disciplinary attention to various sociolegal problems. New sociolegal problems in the areas of human rights, social movements, environment, technology, reproduction, globalization, and the creation and breakdown of markets generate new areas of sociolegal scholarship. The sociolegal perspective is a 'big tent' involving multiple epistemologies, methodologies, and disciplinary perspectives, unified primarily by its understanding of law as a set of norms or institutionalized rules.

Because of the breadth and lack of uniformity within the sociolegal perspective, any effort to characterize it will necessarily be imperfect and incomplete. There is a tension between

emphasizing founding principles and evolving understandings and debates. Nonetheless, we think that the sociolegal perspective tends to emphasize three broad dimensions along which law matters to social life: regulatory, procedural, and constitutive (Edelman and Suchman, 1997). The regulatory dimension involves studies or theories of law as a set of formal rules, informal norms, and principles underlying legal policies that are either meant to shape or control social behavior. Studies of the legal impact, compliance and noncompliance, and the role of extralegal normative orders generally emphasize the regulatory dimension of law. The procedural dimension involves studies or theories of law as a set of tools, practices, resources, or strategies that may be mobilized for creating, challenging, and resolving claims, disputes, and rule ambiguities; for seeking strategic advantages vis-à-vis other parties; for seeking to rationalize transactions; and the like. Studies of contracting behavior, dispute resolution, cause lawyering in litigation, and social movement efforts to enact favorable legal rules generally emphasize the procedural dimension of law. Finally, the constitutive dimension of law involves analyses of law as a set of symbols and forms that evoke and replicate meanings and understandings. Constitutive analyses generally study the role of law in everyday life including the ways in which legal categories and concepts structure social behavior, the way in which people understand the law and invoke it in their daily routines, and ways in which people mobilize or resist law (Ewick and Silbey, 1998).

The regulatory, procedural, and constitutive dimensions of law are not mutually exclusive, and, in fact, most social behavior involves aspects of more than one of these dimensions. To some extent, the regulatory and procedural dimensions of law were the focus of earlier sociolegal studies, which often examined the impact of particular legal rules; the 'gap' between 'law on the books' (i.e., formal law) and 'law in action' (i.e., law in its social context), or discretion in the application of formal legal rules. The constitutive dimension is more common in more recent sociolegal scholarship, which tends to examine the cultural facets of law, to deemphasize formal legal institutions, and to emphasize the ways in which law shapes social thought and social behavior not through rules and sanctions but rather through meaning making. The turn toward constitutive studies has caused a rethinking of the regulatory and procedural facets of law, leading to more attention to social construction and legal consciousness within analyses of the regulatory impact of law or the procedural uses of law.

To complicate matters further, the regulatory, procedural, and constitutive dimensions of law are loosely correlated with both disciplines and methodologies. Economists, psychologists, most political scientists, and some sociologists tend to focus on the regulatory and procedural dimensions of law using more traditional and more quantitative methodologies (sometimes referred to as 'normal science'), whereas anthropologists, some sociologists, and some political scientists, as well as some humanities scholars, especially recently, are more attentive to the constitutive dimension of law and tend to use qualitative empirical methods such as ethnography and interviewing or alternative methodologies such as discourse or textual analysis. Legal scholarship has become more interdisciplinary over time, as reflected in the work of scholars in law schools. Furthermore, many sociolegal scholars work across disciplinary and methodological boundaries, allowing greater understanding of the ways in which the regulatory, procedural, and constitutive facets of law are dynamically intertwined.

Substantive foci are also loosely correlated with these three dimensions. For example, studies of judicial behavior are primarily conducted by political scientists using quantitative methodological techniques. These studies examine the regulatory dimension of law to the extent that they focus on the substantive doctrine that judges devise or the impact of judicial attitudes or politics on judicial decisions, and they examine the procedural dimension of law to the extent that they focus on the impact of judges' strategic efforts to avoid being overturned or to mobilize legitimacy for the court. Studies of law and social movements are mostly conducted by sociologists, although some political scientists also work in this area, and tend to use qualitative methods. Many of these studies focus on the procedural dimension of law, that is, how movement activists use litigation and other strategies to achieve their substantive goals. Studies of legal consciousness are mostly conducted by sociologists and anthropologists, use almost exclusively qualitative methods, and highlight the constitutive dimensions of law.

Given the overlapping nature of the three dimensions we have outlined (both in terms of disciplinary and methodological orientation and substantive topics) and the chronological development of the field, there are many ways that the basic arguments of the sociolegal perspective could be presented. In the remainder of this article, we organize our discussion of the sociolegal conception of law by broad substantive themes in sociolegal analysis. Within our discussion of those themes, we discuss how, and the extent to which, scholarship on those topics emphasizes the regulatory, procedural, and constitutive aspects of law. We are also attentive to changes over time in the foci of sociolegal studies.

Early Themes

The themes we discuss in this section became prominent early in the articulation of the sociolegal perspective on law, that is, in the mid-1960s and 1970s, although these themes are far from absent in more recent analyses. These early analyses were primarily oriented toward the regulatory and procedural aspects of law, but many have been revisited in recent years with a more constitutive flare.

The Gap

Early law and society research focused on what is seen as a gap between 'law on the books' and 'law in action.' 'Law on the books' refers not simply to laws codified in statute books or judicial decisions, but more generally to the formal legal system and to assumptions about how that system operates or ought to operate. In contrast, 'law in action' refers to law in its social context, which is often determined through empirical study. Gap studies focus on both the regulatory and procedural facets of law. Attention to regulatory processes often takes the form of legal impact studies that examine whether law that seeks to regulate social behavior in fact produces compliance by organizations (Hawkins, 1984; Stone, 1975; Vaughan,

1985) or individuals (Tyler, 1990), and how law in fact operates in society given power, discretion, social roles, deviance, and other social realities (Albiston, 1999; Friedman, 1975; Galanter, 1974; Macaulay, 1963). Studies that emphasize the procedural dimension of law tend to identify how laws are used by various parties in ways that deviate from the formal model. For example, Macaulay's seminal article on contract law showed that businesses often preferred informal agreements to formal contracts and rarely invoked formal contractual remedies when contracts were breached (Macaulay, 1963). Merry's (1979) article on strategies of dispute management in an urban neighborhood shows how people file complaints in criminal courts as a way of harassing neighbors with whom they are having a dispute. Engel's (1984) work on personal injury dispute resolution shows how social norms affect the likelihood of using courts to resolve disputes of various types. He shows that in a small cohesive community, people are less likely to make use of personal injury litigation because it violates norms about personal responsibility. Among newcomers to the community, who were less likely to share these cultural ideas, personal injury litigation was more likely.

A good deal of law and society research examines the gap in the contexts of judges, courts, legislators, administrative agencies, lawyers, and police (*see*, e.g., Courts and Adjudication; Rights: Legal Aspects; Race and the Law; Civil Rights; Injustice: Legal Aspects). In general, the sociolegal perspective emphasizes the way in which these various actors and institutions adjust to their social contexts. Instead of a portrait in which judges are impartial and passive, legislators act in the public interest, laws are clear, administrative agencies enforce the law, courts and judges stand as impartial interpreters of law, lawyers act as adversarial advocates, police exercise no discretion, and people respond rationally and promptly to all violations of law, we see a world in which judges are influenced by their political backgrounds, attitudes, and strategic efforts to succeed at reelection and avoid being overturned by higher courts (Epstein and Knight, 2000; Segal and Spaeth, 2002); courts become policy makers (Feeley and Rubin, 1999); legislators make deals (Buchanan and Tullock, 1965); administrative agencies soften the formal law, mediate disputes, engage in 'soft enforcement' like education or negotiation, and are sometimes influenced by those they are charged with regulating (Gunningham, 1987; Hawkins, 1984; Scholz, 1991); and police make critical decisions about what constitutes crime and how people are charged (Muir, 1979). The focus on the gap has also led to inquiries into the conditions under which law can bring about social change (McCann, 1994; Rosenberg, 1991), the circumstances under which people obey law (Tyler, 1990), and the role of law in social control.

Although the idea of the gap has motivated some important inquiries into the workings of law in society, late-twentieth-century scholarship found the idea of the gap to be flawed theoretically. Abel (1973) published a strong critique of the idea that the gap is problematic, suggesting that there is no reason to think that legal rules and legal behavior would be consistent and that the preoccupation of social scientists with the gap distracts from the construction of a social theory of law. Feeley (1976) similarly argued that the characterization of the gap depends on an overly narrow conception of the formal law and inability to identify or measure the goals of that law. Because formal law is often quite ambiguous, in part because its creators had diverse interests and objectives, it is misleading to characterize the actual operation of law as a deviation from the clear goals of formal laws. Rather, law can only be understood as embedded in its social context, where it is both the product and the source of social practices.

Bargaining and Negotiation as Key Features of the Legal Order

Emphasizing the procedural facet of law, one of the early 'discoveries' of law and society research is the pervasiveness and centrality of bargaining throughout the legal world. Processes of authoritative decision making turn out in fact to revolve around negotiation. Ostensibly adjudicative processes become negotiative ones; decision makers clothed with arbitral powers are transformed into mediators. The negotiated outcome in the shadow of the law turns out to be the master pattern of disputing in American courts and administrative agencies.

Bargaining often takes place in the 'shadow of the law' (Mnookin and Kornhauser, 1979). The rules of formal law confer on the parties what Mnookin and Kornhauser call a 'bargaining endowment,' that is, a set of 'counters' to be used in bargaining between disputants (Ross, 1970). The bargaining endowment that the formal law bestows on parties derives not only from the substantive entitlements conferred by the legal rules but also from the adjectival rules that enable those entitlements to be vindicated. But rules are only one source of the endowment conferred by the legal process: The delay, cost, and uncertainty of eliciting a favorable determination also confer bargaining counters on the disputants. Delay, cost, and uncertainty may themselves be the product of rules, but they also result from such nonrule factors as the number, distribution, fees, and work patterns of courts and lawyers. The meaning of the endowment bestowed by the law is not fixed and invariable, but depends on the relative characteristics of the disputants: their preferences, negotiating skill, aversion to risk, ability to bear costs and delays, to respond to deadlines and emergencies, and so forth. A different mix of disputant capabilities may make a given endowment take on a very different significance.

Knowledge about the legal system is often incomplete or inaccurate, however, so that, social, business, and professional norms have a greater impact on how parties bargain than do their putative legal bargaining positions (Edelman et al., 1993; Ellickson, 1991; Galanter, 1986). Parties may be influenced less by what would happen in court than by how their neighbors or business partners or coworkers think they ought to behave. The preservation of relationships often trumps the assertion of legal entitlements. As a result, social reform laws – especially to the extent that they depend on victims to mobilize their rights – infrequently succeed in effecting major social change (*see* Law as an Instrument of Social Change). Recent attention to the procedural facet of law moves beyond the notions of bargaining and discretion to frame the actions of legal actors more in terms of power and moral agency than in terms of discretion (Maynard-Moody and Musheno, 2012).

The Distributive Tilt of the Legal Order

Although modern adjudication is infused with ideals of equality before the law, and the isolation of the dispute from

the identity, connections, and resources of the parties, the sociolegal perspective on law emphasizes the myriad ways in which these ideals are imperfectly realized when law is considered in social context. Early studies focused on the role of class in limiting access to justice (Carlin and Howard, 1965; Galanter, 1974; Handler et al., 1978).

Legal rights are not self-mobilizing and the process of mobilization is a key site in which social disadvantage rears its head. To realize the benefits of social reform legislation such as civil rights law or equal pay legislation, those whose rights are violated must mobilize the law. The sociolegal perspective emphasizes the extent to which those who have less social clout are also less likely to 'name' legal wrongs, to 'blame' a legally responsible party, and to 'claim' in a way that may lead to legal redress (Bumiller, 1987, 1988; Felstiner et al., 1980; Marshall, 2005a).

Those possessing wealth, power, and other resources can hire better lawyers, invest in marshaling proofs and arguments, use delay to their advantage, and adopt strategies that deflect unfavorable rules and reap the full advantage of law (Heinz et al., 2005; *see* Injustice: Legal Aspects; Equality and Inequality: Legal Aspects). The advantages of stronger parties who are recurrent users of the legal system are even more pronounced in nonjudicial forums such as legislatures and administrative agencies. So legal regimes that accord a central place to equality norms may at the same time enable "the 'haves' [to] come out ahead" (Albiston, 1999; Edelman and Suchman, 1999; Galanter, 1974; Hagan, 1990).

In response to this paradox, legal systems have incorporated various devices to overcome the disparity in legal resources (*see* Justice, Access to; Class Actions, Legal; Public Interest Law: The United States and Beyond). The achievements and deficiencies of these devices have been a focus of law and society research. Although the presence of lawyers tempers the disparities in legal resources, it does not overcome them, for legal services are typically purchased in markets where those with superior resources can attract more proficient practitioners. Apart from differences in quality and quantity, legal representation is infected with agency problems. Lawyers have goals and needs that do not align perfectly with those of their clients. It is not simply that financial incentives may induce lawyers to do too much or too little, but that even the lawyer who eschews financial gain may be deflected by appetite for fame, honor, or 'interesting' work. For example, the low monetary stakes and repetitiveness of the legal needs of the poor run against the powerful pull of craft satisfactions and the aspiration to heroic distinction.

Although early studies of the distributive tilt of law focused on the ways in which legal rules and procedures benefited the 'haves,' thus emphasizing the regulatory and procedural dimensions of law, more recent studies incorporate the constitutive perspective by focusing on ways in which cultural assumptions about race, gender, disability, national origin, and other dimensions of social stratification are reified and often exacerbated by the legal order (Albiston, 2010; Best et al., 2011; Engel, 1991; Gómez, 2012; Marshall, 2005a; Obasogie, 2013). Many of these studies emphasize the ways in which law helps to construct notions of racial or gender differences or the ways in which law tends to disregard social processes that offer advantages to whites and males. For example, the notion of 'intersectionality' calls attention to how antidiscrimination law presumes that discrimination occurs along only one dimension (e.g., race or sex or age), substantially disadvantaging plaintiffs who suffer multiple disadvantages such as race *and* age. (Best et al., 2011; Crenshaw, 1991). These works call attention to the meaning-making dimensions of law and how they interact with more structural forms of disadvantage.

Multiple Sites of Governance

Sociolegal analyses of the regulatory facet of law emphasize that the state is only one source of law and legal norms. In most modern societies, the law presents itself as unitary, but nonetheless consists of multiple systems of regulation that coexist and intertwine (*see* Legal Pluralism). Religious bodies, universities, private firms, trade groups, and associations of all kinds devise and broadcast norms, regulate their affairs, and impose sanctions – often by means of proceedings that to varying extents mimic, if imperfectly, those of state law. Some legal systems openly yoke together several bodies of law from different origins. For example, some ancient and modern societies have had a system of 'personal law' in which the incidence of particular rules depends on the religious or tribal identity of the parties. Many legal systems formally recognize alternative sets of rules. Hart (1961) pointed out that in addition to primary rules, the law contains various sets of secondary rules about the recognition and application of its own rules and the recognition and reinforcement of regulatory activity in other institutions.

Regulation is not a governmental monopoly; there are multiple arenas of normative enterprise, innovation, and interpretation. The state's law does not preside over a landscape barren of regulation, but over a thick tangle of rivals and companions. Its effects depend on the way that it interacts with the various sorts of indigenous ordering that surround it; the core official institutions themselves are the scene of persisting and pervasive local variation. These various departures from unity and uniformity have been studied under the rubrics of private government, indigenous law, semiautonomous social fields, and local legal culture. Pluralism is very much with us: The centralist view of a monolithic integrated legal order turns out not to be a description of modern law, but part of its ideology.

Although law is the official terrain of dispute resolution, the vast majority of disputes are handled outside of formal legal institutions. Felstiner et al. (1980) introduced the dispute emergence perspective, and Miller and Sarat (1980) introduced the notion of the 'dispute pyramid.' Both emphasize the limited likelihood that legal violations will be perceived or handled formally at all, and that only a small fraction of disputes will evolve into formal trials. Morrill et al. (2010) elaborate this model by presenting a multidimensional disputing model, suggesting that most disputes are handled through extralegal processes such as talking to a friend, family member, or clergy person; prayer; or contacting the media and that others are handled through quasi-legal processes such as organizational grievance procedures. Edelman and Suchman (1999) suggest that organizations are internalizing more aspects of legal regulation over time. The continuance of *de facto* pluralism is fostered by various ideological currents,

including movements to preserve and extend cultural diversity, promote local control, and deemphasize adjudication in favor of 'alternative dispute resolution.' To emphasize the many channels for dispute resolution outside of the formal legal system, Albiston et al. (2014) suggest that the dispute pyramid metaphor is misleading and that it should be reframed as a 'dispute tree' with many branches that represent various types of legal, quasi-legal, and extralegal forms of dispute resolution. The branches are differentially accessible, with differing types of flowers (symbolic results) and fruit (substantive result).

In sum, even though law regulates increasing areas of social life, most social regulation goes on within various institutional and relational settings; only a small and unrepresentative minority of disputes proceeds to a formal legal institution for resolution. To a large but perhaps diminishing extent, when formal legal institutions devise rules to regulate conduct in institutions and relationships, they refine and restate norms that already prevail in those settings (Bohannan, 1965); legal norms in turn transform and feed rather than displace or eradicate the regulatory activity indigenous to these settings.

More Recent Themes

The sociolegal perspective in more recent times has continued to address the regulatory and procedural aspects of law but increasingly emphasizes the constitutive aspects of law, or the way in which law helps to construct both social consciousness and social action. The themes identified in the previous section are still very salient to the sociolegal perspective, but more recent scholarship tends to focus more on meaning making and social construction, as well as the role of law in everyday life. Recent sociolegal approaches also moved from a primarily American focus to a focus that is more comparative and attentive to law around the world.

The Globalization of Law

The sociolegal perspective calls attention to the wide variety of legal systems around the globe and also to the diffusion of legal norms, materials, and devices from Western Europe, lately joined and surpassed by North America as a source of legal inspiration. Voluntary reception by settler societies, the imposition of colonial rule, and importations by independent regimes seeking a badge and instrument of advancement have been followed by a worldwide spread of constitutionalism and continued influence of Western legal models to promote new forms of economic activity stimulated by increases in global trade (*see* Postcolonial Law; Law and Development). Diffusion of Western models was accelerated by the demise of communism and the subsequent importation of Western experts to write constitutions and oversee the democratization of formerly communist regimes. Comparative analyses of constitutional borrowing, adaptation, and hybridization have become prominent in sociolegal scholarship (*see* Constitutionalism, Comparative).

Notwithstanding some convergence of formal legal doctrines and institutional structures, there remains broad variation in the amount and character of litigation, the number and roles of lawyers, and the styles of dispute resolution across nations and social contexts. Even when constitutions and statutes are imported directly from one country to another, their interpretation and significance reflect persisting differences in social, political, and legal settings.

Adapted institutions may combine with local culture to produce distinctively different patterns of use. For example, in Thailand, as Engel and Engel (2010) show, disruptive changes in social arrangements and ideology cause adopted legal institutions to assume very different meanings and to be used – and left unused – in a way quite at variance with their original embodiment.

Notwithstanding persisting differences, there has been an increase everywhere in the size of the legal plant. The number of lawyers has increased. This is partly an artifact of expanding systems of higher education. But even discounting for this, the number of lawyers has grown in relation to the population and in comparison to the size of the judiciary (Galanter, 2011). Increasingly, the upper strata of these lawyers are organized in large firms that have adopted the organizational form of the large 'promotion to partnership' firm originally developed in the United States in the early twentieth century (Galanter and Palay, 1994). The growth of such firms, which provide a wide range of services to corporate entities (and some rich individuals), accelerated in the later part of the twentieth century. The spread in recent decades of this style of law practice through local adoption and by overseas branching has accentuated a striking asymmetry in the spread of legal capability.

Firms of increasingly large size, made up of highly proficient lawyers, provide a wide range of services to corporate and government entities, while individuals receive a much narrower range of services from lawyers practicing in smaller units, frequently as sole practitioners (Heinz et al., 2005; Hadfield, 2000). Services to corporate and governmental clients include preventive planning services as well as the more familiar services provided after the occurrence of a problem or predicament (Hadfield, 2009).

The global spread of legal institutions displays a pronounced tilt toward American models of large law firms, "broad empowerment of private actors to assert legal rights" (Keleman and Sibbitt, 2004: 106), and a culture of adversarial legalism (Wiegand, 1996). "Economic liberalization … [has] undermined traditional, informal, opaque approaches to regulation and pressured governments to resort to more formal, transparent, adversarial, American-style processes" (Kelemen and Sibbitt, 2004: 131). While other features of American-style legalism spread rapidly, litigation lags and becomes a smaller province of the legal realm.

The Legalization of Society

Law has become more pervasive, both as a set of social structures and as a component of consciousness, and this process is likely to continue. The latter half of the twentieth century saw a steady increase in the amount and complexity of regulation; the frequency of litigation; the number, coordination, and productivity of lawyers; the number of legal actors and the resources they devote to legal activity; the amount of information about law and the velocity with which it circulates; and a concomitant rise in consciousness of law and awareness of

legal rights (Galanter, 1992; *see* Rights: Legal Aspects; Legalization).

An increasingly global economy together with new communications technologies have fueled the rise of legal activity, in areas such as intellectual property (*see* Property: Legal Aspects; Intellectual Property: Law in Context). The number of lawyers increased dramatically in many countries and became markedly more diverse (in terms of gender, ethnicity, and class origin) (Galanter, 2011; *see* Law Firms; Lawyers). Regulation has increasingly taken on the features of adjudication (Teubner, 1987). The role of courts has been enlarged by the spread of constitutionalism and judicial review, together with an increase in multitier regulation deriving from federalism or from participation in international institutions. Increasingly, courts are accepted as the final arbiters of many public issues (Cappelletti, 1985; Constitutional Courts).

In many quarters, these changes have engendered a recoil against law and regulation, not least among elites who are best protected by the legal order. Concerns about excessive legalization, 'juridification,' 'litigation explosion,' and the dominance of lawyers have led to the advocacy of deregulation, delegalization, and embrace of alternatives to traditional litigation. The term 'alternative dispute resolution,' virtually unheard of prior to 1970, has become part of the standard policy lexicon. With the rise of forums such as mediation, minitrials, and other private or court-annexed modes of dispute processing (*see* Mediation, Arbitration, and Alternative Dispute Resolution (ADR)), the locus of some legal activity has shifted away from adjudication and into nonpublic forums, but it remains doubtful that the tide of regulation has been reversed.

As discussed below, the prospect is that the legal system has an ever-expanding agenda. But the process of remedying injustice or simply assuaging grievances consumes resources and society gets an increment of justice instead of medicine or missiles. In addition to the costs of operating legal institutions and costs borne by the participants in the legal process, the use of law to address injustice imposes another kind of cost, as it accumulates into a thick layer of rules to be learned and followed, permissions to be obtained, forms to be filled out, and records to be kept, adding to expense and burdening activity. These costs of law are visible and palpable, while the benefits of legalization are less so: Victims may enjoy vindication or honor or a heightened sense of self-worth – goods that are not subject to the same zero-sum distributional rules as money. The legal process may generate public goods by crystallizing and broadcasting norms, preventing injustice by deterrence and education, and providing channels for interaction and coordination (e.g., driving on the right or left, as the case may be). The more successful such patterning is, the less visible it is.

The Intersection of Law and Organizations

The intersection of law and organizations entails the regulatory, procedural, and constitutive facets of law. The intersection of law and regulatory law occurs as organizations respond to ambiguous legal regulation by creating internal rules that on the one hand mimic formal legal rules but on the other infuse the law with managerial prerogatives and perspectives (Edelman et al., 2001). Over time, organizational responses to law tend to become widely accepted throughout organizational fields (Edelman et al., 1999; Sutton et al., 1994).

The intersection of organizations and procedural law occurs frequently as organizations engage in litigation as a way of influencing their markets and environments (Cheit and Gersen, 2000; Dunworth and Rogers, 1996; Priest and Klein, 1984), as they hire in-house counsel who help to specify the meaning of ambiguous legislation in terms of corporate interests (Chayes and Chayes, 1984; Nelson and Nielsen, 2000; Rosen, 2002), as they internalize dispute resolution by creating internal dispute resolution systems (Edelman and Cahill, 1998; Edelman et al., 1993; Edelman and Suchman, 1999; Talesh, 2009), and by using private police forces who answer to corporate rather than public concerns (Edelman and Suchman, 1999). The internalization of law makes organizations ever more powerful legal actors in the public realm as well, and they drive up the ante by dominating the dockets, playing for favorable statutes and judicial interpretations of those statutes, and employing strategies and experts that tend to increase the cost and complexity of the legal game.

The intersection of organizations and the constitutive realm of law occurs in several ways. First, organizations respond to ambiguity in their legal environments by creating a variety of 'symbolic structures' like antidiscrimination policies that symbolize attention to legal ideals but do not guarantee that organizational practices will further those ideals. Once in place, organizational structures become sites in which legal rights are constructed and often transformed. Edelman et al. (2001) show how legal constructs tend to be infused with managerial meaning and values over time, or 'managerialized.' Albiston (2005) shows how ideas about work and wages constrain employees' use of rights created under the Family and Medical Leave Act. Similarly, Marshall (2005b) shows how managerialized conceptions of law prevent employees from mobilizing their civil rights generally. Over time, compliance tends to be understood in terms of organizations' symbolic structures, irrespective of whether those structures further legal ideals. Thus, organizational responses to law tend to become institutionalized or taken for granted not only among organizations and actors within organizations but also among litigants, lawyers, and ultimately judges. This in turn produces 'legal endogeneity' or law that incorporates the ideas and perspectives of the social realms that it seeks to regulate (Edelman et al., 2011, 1999; *see* Organizations and the Law).

Social Movements and Law

Like organizations, social movements intersect with the regulatory, procedural, and constitutive facets of law. Social movements often seek to influence regulatory law by, for example, seeking stricter environmental regulation or more progressive stances on sexuality (Andersen, 2005). Social movements also consciously violate regulation when they engage in civil disobedience to protest nuclear energy, animal experimentation, or abortion provision. Social movements engage the procedural aspects of law when they litigate to advance a cause or seek injunctions to stop pollution (Sarat and Scheingold, 2006).

But social movements also help to shape the meaning of law, engaging the constitutive element of law, when they advance new 'frames' or ways of understanding the world by challenging well-accepted institutions (Benford and Snow, 2000). Notions of 'green construction,' 'reproductive justice,' and 'gay marriage,' for example, work their way into everyday language and understanding in ways that ultimately influence law. Social movements may empower individuals to engage in 'rights talk' to negotiate social benefits (Engel and Munger, 2003; Minow, 1987; Williams, 1991). Legal activists may use legal tactics as a tool for changing people's perceptions about stigmatized groups (Kostiner, 2003; Silverstein, 1996). In his analysis of the pay equity movement, for example, McCann (1994) shows how even failed attempts at legal change can serve to mobilize political action by shaping peoples understandings of social issues.

Recent sociolegal scholarship on social movements shows how law can provide a vocabulary for describing injustice (Albiston, 2005; Marshall, 2003), how law can mobilize social movement activity by shaping social identities (Diller, 2003; Engel and Munger, 2003; Yashar, 2005), and how legal change may prompt countermobilization (Goldberg-Hiller, 2002; Luker, 1984; Tushnet, 1987). Law also helps to shape activists' understandings of what types of claims are valid and what types of solutions are just (Edelman and Cahill, 1998; Ferree, 2003; Haney-López, 2000; Jenness, 1999). Recent approaches also call attention to the ways in which organizations, social movements, and law interact to shape understandings of social issues, rights, and the potential for change (Edelman et al., 2010).

Legal Consciousness, Rights Consciousness, and Legal Culture

The constitutive facet of law emphasizes meaning making and gives more attention than did the earlier perspectives to legal consciousness and legal culture, terms that are used to identify widely shared meanings either about law or that originate from law. Legal consciousness generally refers to how individuals think about law, whereas legal culture refers to aggregate understandings about law (Ewick and Silbey, 1998; Merry, 1990; Nelken, 1997; Sarat, 1990; Scheingold, 1974; *see* Legal Culture and Legal Consciousness; Law and Everyday Life).

Studies of legal consciousness and legal culture have called attention to the ways in which legal ideas influence every day social life in ways that might not previously have been understood as legal. Particular attention has been paid to rights consciousness, or the extent to which people recognize and mobilize their legal rights. Formal legal rights, for example, even when they are difficult to realize in court, may allow disenfranchised citizens to make claims against more powerful entities (Minow, 1987; Williams, 1991). Legal consciousness helps to explain when people mobilize the law and why they choose not to (Albiston, 2010; Bumiller, 1988; Marshall, 2005a; Nielsen, 2004), how people think about public welfare bureaucracies (Sarat, 1990), and when and how people encounter the law and why and how they resist law (Ewick and Silbey, 1992).

New and Developing Themes in the Sociolegal Perspective on Law

The sociolegal perspective on law continues to elaborate the regulatory, procedural, and constituent facets of law. Many of the new themes within this perspective are motivated by new societal challenges such as evolving technologies and technological possibilities (Berman, 2007; Elkin-Koren, 1996; Ross, 2002); the rise of a 'risk society' (Baker and Simon, 2002; Beck, 1992; Ericson and Haggerty, 1997; Giddens, 1999); global warming and the increasing number of natural disasters (Gunningham et al., 2003; Hartman and Squires, 2006); ethnic conflict and violence around the globe (Hagan and Rymond-Richmond, 2009; Ivković and Hagan, 2006); the transition from nations at war to the 'war against terror' (Hagan et al., 2008; Hirsch, 2008; Thacher, 2005); demographic movements through immigration and emigration, which raise new questions of citizenship (Boyle and Busse, 2006; Menjívar and Abrego, 2012; Wiles, 2007); the rise of global capitalism and corporate power (Carruthers and Halliday, 2006; Shamir, 2004); problems faced by new democracies seeking to craft constitutions (Elkins et al., 2009; Epstein et al., 2001; Ginsburg, 2003); the increasing role of nongovernmental organizations and other nonstate entities in rule making and meaning making; and the interplay between these entities and states in influencing international orders (Boyle and Preves, 2000; Davis et al., 2010; Massoud, 2011; Trubek et al., 2000).

The sociolegal perspective increasingly recognizes the nexus between a changing social world and changing notions of law and justice. The scope of potential legal intervention expands with these new challenges as well as with the growth of knowledge and the development of technology. The sociolegal perspective sees the realm of injustice as enlarged as events that were once seen as natural come to be seen as capable of being controlled by human institutions (Shklar, 1990). Hurricanes are misfortunes, but bungled relief efforts are injustices. The focus on the constitutive facet of law reveals that our consciousness of injustice increases as the potential for legal solutions increase.

The Sociolegal Perspective in Sum

This article necessarily gives short shrift to many topics within the sociolegal perspective on law. It is intended more to give a feel for the breadth of the perspective and its emphasis on understanding the interplay between law and its social context. From its early critique of legal formalism and emphasis on the gap between 'law on the books' and 'law in action,' the sociolegal perspective has evolved into a multidisciplinary and interdisciplinary analysis of law at every level of society. Studies of regulation focus not just on impact but also on the social inequalities at play in the construction of regulation, in its implementation, and in its interpretation. Studies of procedure focus not just on law as a strategy but also on the meaning making, moral judgments, and power plays that pervade every legal move. Studies of the constitutive of power of law now incorporate attention to regulation and procedure and focus on broader issues of legal consciousness, rights consciousness, social resistance, and the role of law in everyday life.

See also: Civil Rights; Class Actions, Legal; Constitutionalism, Comparative; Courts and Adjudication; Equality and Inequality: Legal Aspects; Indigenous Rights; Injustice: Legal Aspects; Justice, Access To; Law and Development; Law and Everyday Life; Law as an Instrument of Social Change; Lawyers: Social Organization of the Profession; Legal Culture and Legal Consciousness; Legal Pluralism; Organizations and the Law; Postcolonial Law; Public Interest Law: The United States and Beyond; Race and the Law; Rights: Legal Aspects.

Bibliography

Abbott, Kenneth W., Snidal, Duncan, 2000. Hard and soft law in international governance. International Organization 54 (3), 421–456.

Abel, Richard, 1973. Law books and books about law. Stanford Law Review 26, 175–228.

Abrego, Leisy J., 2011. Legal consciousness of undocumented Latinos: fear and stigma as barriers to claims-making for first- and 1.5-generation immigrants. Law & Society Review 45 (2), 337–370.

Albiston, Catherine R., 1999. The rule of law and the litigation process: the paradox of losing by winning. Law & Society Review 33 (4), 869–910.

Albiston, Catherine R., 2005. Bargaining in the shadow of social institutions: competing discourses and social change in workplace mobilization of civil rights. Law & Society Review 39 (1), 11–50.

Albiston, Catherine R., 2010. Institutional Inequality and the Mobilization of the Family and Medical Leave Act: Rights on Leave. Cambridge University Press, New York.

Albiston, Catherine R., Edelman, Lauren B., Milligan, Joy, 2014. The dispute tree and the (legal) forest. Annual Review of Law and Social Science 10.

Andersen, Ellen A., 2005. Out of the Closets and into the Courts: Legal Opportunity Structure and Gay Rights Litigation. University of Michigan Press, Ann Arbor.

Baker, Tom, Simon, Jonathan, 2002. Embracing Risk: The Changing Culture of Insurance and Responsibility. University of Chicago Press, Chicago.

Beck, Ulrich, 1992. Risk Society: Towards a New Modernity. Sage Publications, London.

Becker, Gary S., 1968. Crime and punishment: an economic approach. The Journal of Political Economy 76 (2), 169–217.

Benford, Robert D., Snow, David A., 2000. Framing processes and social movements: an overview and assessment. Annual Review of Sociology 26, 611–639.

Berman, Paul S. (Ed.), 2007. Law and Society Approaches to Cyberspace. Ashgate Publishing Company, Burlington, VT.

Best, Rachel K., Edelman, Lauren B., Krieger, Linda H., Eliason, Scott R., 2011. Multiple disadvantages: an empirical test of intersectionality theory in EEO litigation. Law & Society Review 45 (4), 991–1025.

Bohannan, Paul, 1965. The differing realms of the law. American Anthropologist 67 (6), 33–42.

Bothe, Michael, 1980. Legal and non-legal Norms–a meaningful distinction in international relations? Netherlands Yearbook of International Law 11, 65–95.

Boyle, Elizabeth H., Busse, Erika, 2006. Institutional vulnerablity and opportunity: immigration and America's "war on terror". Law & Social Inquiry 31 (4), 947–974.

Boyle, Elizabeth H., Preves, Sharon E., 2000. National politics as international process: the case of anti-female-genital-cutting laws. Law and Society Review 34 (3), 703–737.

Buchanan, James M., Tullock, Gordon, 1965. The Calculus of Consent: Logical Foundations of Constitutional Democracy. University of Michigan Press, Ann Arbor.

Bumiller, Kristin, 1987. Victims in the shadow of the law: a critique of the model of legal protection. Signs 12 (3), 421–434.

Bumiller, Kristin, 1988. The Civil Rights Society: The Social Construction of Victims. Johns Hopkins University Press, Baltimore, MD.

Burch, Traci, 2011. Turnout and party registration among criminal offenders in the 2008 general election. Law & Society Review 45 (2), 699–730.

Carlin, Jerome E., Howard, Jan, 1965. Legal representation and class Justice. UCLA Law Review 12, 381–437.

Carruthers, Bruce G., Halliday, Terence C., 2006. Negotiating globalization: global scripts and intermediation in the construction of Asian insolvency regimes. Law & Social Inquiry 31 (3), 521–584.

Cappelletti, Mauro, 1985. Repudiating Montesquieu? The expansion and legitimacy of constitutional justice. Catholic University Law Review 35 (1), 1–32.

Chayes, Abram, Chayes, Antonia H., 1984. Corporate counsel and the elite law firm. Stanford Law Review 37, 277.

Cheit, Ross E., Gersen, Jacob E., 2000. When businesses Sue Each other: an empirical study of state court litigation. Law & Social Inquiry 25 (3), 789–816.

Chinkin, Christine M., 1989. The challenge of soft law: development and change in international law. International and Comparative Law Quarterly 38 (4), 850–866.

Constable, Marianne, 2009. Just Silences: The Limits and Possibilities of Modern Law. Princeton University Press, Princeton.

Cooter, Robert D., Rubinfeld, Daniel L., 1989. Economic analysis of legal disputes and their resolution. Journal of Economic Literature 27 (3), 1067–1097.

Crenshaw, Kimberlé W., 1988. Race, reform, and retrenchment: transformation and legitimation in antidiscrimination law. Harvard Law Review 101 (7), 1331–1387.

Crenshaw, Kimberlé W., 1991. Mapping the Margins: intersectionality, identity politics, and violence against women of color. Stanford Law Review 43 (6), 1241–1299.

Curran, Barbara A., 1977. The Legal Needs of the Public: The Final Report of a National Survey, a Joint Undertaking by the American Bar Association Special Committee to Survey Legal Needs and the American Bar Foundation. American Bar Foundation, Chicago.

Davis, Kevin E., Kingsbury, Benedict, Merry, Sally Engle, 2010. Indicators as a technology of global governance. In: New York University Public Law and Legal Theory Working Papers; International Law and Justice Working Papers. Institute for International Law and Justice, New York University School of Law.

Dezalay, Yves, Garth, Bryant, 2001. The import and export of law and legal institutions: international strategies in national palace wars. In: Nelken, David, Feest, Johannes (Eds.), Adapting Legal Cultures. Hart Publishing, Oxford, pp. 241–255.

Diamond, Shari S., Casper, Jonathan D., 1992. Blindfolding the jury to verdict consequences: damages, experts, and the civil jury. Law & Society Review 26 (3), 513.

Diamond, Shari S., Rose, Mary R., 2005. Real juries. Annual Review of Law and Social Science 1, 255–284.

Diller, Matthew, 2003. Judicial backlash, the ADA, and the civil rights model of disability. In: Krieger, Linda H. (Ed.), Backlash against the ADA: Reinterpreting Disability Rights. University of Michigan Press, Ann Arbor.

Donahue, John J., Heckman, James, 1991. Continuous versus episodic change: the impact of civil rights policy on the economic status of blacks. Journal of Economic Literature 29 (4), 1603–1643.

Dunworth, Terence, Rogers, Joel, 1996. Corporations in court: big business litigation in U.S. Federal courts, 1971–1991. Law & Social Inquiry 21 (3), 497–592.

Edelman, Lauren B., Cahill, Mia, 1998. How law matters in disputing and dispute processing (Or, the contingency of legal matter in alternative dispute resolution). In: Bryant, G. Garth, Sarat, Austin (Eds.), How Law Matters. Northwestern University Press, Chicago, IL, pp. 15–44.

Edelman, Lauren B., Suchman, Mark C., 1997. The legal environments of organizations. Annual Review of Sociology 23, 479–515.

Edelman, Lauren B., Suchman, Mark C., 1999. When the "haves" hold court: speculations on the organizational internalization of law. Law & Society Review 33 (4), 941–991.

Edelman, Lauren B., Erlanger, Howard S., Lande, John, 1993. Internal dispute resolution: the transformation of civil rights in the workplace. Law & Society Review 27 (3), 497–534.

Edelman, Lauren B., Uggen, Christopher, Erlanger, Howard S., 1999. The endogeneity of legal regulation: grievance procedures as rational myth. American Journal of Sociology 105 (2), 406–454.

Edelman, Lauren B., Riggs Fuller, Sally, Mara-Drita, Iona, 2001. Diversity rhetoric and the managerialization of law. American Journal of Sociology 106 (6), 1589–1641.

Edelman, Lauren B., Leachman, Gwendolyn, McAdam, Doug, 2010. On law, organizations, and social movements. Annual Review of Law and Social Science 6, 653–685.

Edelman, Lauren B., Krieger, Linda H., Eliason, Scott R., Albiston, Catherine R., Mellema, Virginia, 2011. When organizations rule: judicial deference to institutionalized employment structures. American Journal of Sociology 117 (3), 888–954.

Elkin-Koren, Niva, 1996. Cyberlaw and social change: a democractic approach to copyright law in cyberspace. Cardozo Arts & Entertainment Law Journal 14, 215.

Elkins, Zachary, Ginsburg, Tom, Melton, James, 2009. The Endurance of National Constitutions. Cambridge University Press, Cambridge.

Ellickson, Robert C., 1991. Order without Law: How Neighbors Settle Disputes. Harvard University Press, Cambridge.

Engel, David M., 1984. The oven bird's song: insiders, outsiders, and personal injuries in an American community. Law & Society Review 18 (4), 551–582.

Engel, David M., 1991. Law, culture, and children with disabilities: educational rights and the construction of difference. Duke Law Journal 1, 166–205.

Engel, David M., Engel, Jaruwan S., 2010. Tort, Custom and Karma: Globalization and Legal Consciousness in Thailand. Stanford University Press, Stanford, CA.

Engel, David M., Munger, Frank W., 2003. Rights of Inclusion: Law and Identity in the Life Stories of Americans with Disabilities. University of Chicago Press, Chicago.

Epp, Charles R., 1998. The Rights Revolution: Lawyers, Activists, and Supreme Courts in Comparative Perspective. Cambridge University Press, Cambridge.

Epstein, Lee, Knight, Jack, 1998. The Choices Justices Make. CQ Press, Washington, DC.

Epstein, Lee, Knight, Jack, 2000. Toward a strategic revolution in judicial politics: a look back, a look ahead. Political Research Quarterly 53 (3), 625–661.

Epstein, Lee, Knight, Jack, Shvetsova, Olga, 2001. The role of constitutional courts in the establishment and maintenance of democratic systems of government. Law & Society Review 35 (1), 117–164.

Ericson, Richard V., Haggerty, Kevin D., 1997. Policing the Risk Society. Oxford University Press, Oxford.

Erikson, Robert S., 1995. State turnout and presidential voting. American Politics Quarterly 23, 387–396.

Ewick, Patricia, Silbey, Susan S., 1992. Conformity, contestation, and resistance: an account of legal consciousness. New England Law Review 26, 20.

Ewick, Patricia, Silbey, Susan S., 1998. The Common Place of Law: Stories from Everyday Life. University of Chicago Press, Chicago.

Feeley, Malcolm M., 1976. The concept of laws in social science: a critique and notes on an expanded view. Law & Society Review 10 (4), 497–523.

Feeley, Malcolm M., 1992. The Process Is the Punishment: Handling Cases in a Lower Criminal Court. Russell Sage Foundation, New York.

Feeley, Malcolm M., Rubin, Edward L., 1999. Judicial Policy Making and the Modern State: How the Courts Reformed America's Prisons. Cambridge University Press, Cambridge.

Feldman, Eric A., 2000. The Ritual of Rights in Japan: Law, Society, and Health Policy. Cambridge University Press, Cambridge.

Felstiner, William L.F., Abel, Richard L., Sarat, Austin, 1980. The emergence and transformation of disputes: naming, blaming, claiming…. Law & Society Review 15, 3–4.

Ferree, Myra M., 2003. Resonance and radicalism: feminist framing in the abortion debates of the United States and Germany. American Journal of Sociology 109 (2), 304–344.

Friedman, Lawrence M., 1975. The Legal System: A Social Science Perspective. Russell Sage Foundation, New York.

Frohmann, Lisa, Mertz, Elizabeth, 1994. Legal reform and social construction: violence, gender, and the law. Law & Social Inquiry 19 (4), 829–852.

Galanter, Marc, 1974. Why the haves come out ahead: speculations on the limits of legal change. Law & Society Review 9 (1), 95–160.

Galanter, Marc, 1986. Adjudication, litigation, and related phenomena. In: Lipson, Leon, Wheeler, Stanton (Eds.), Law and the Social Sciences. Russel Sage Foundation, New York, pp. 151–257.

Galanter, Marc, 1992. Law abounding: legalisation around the North Atlantic. The Modern Law Review 55 (1), 1–24.

Galanter, Marc, 2005. Planet of the APs: reflections on the scale of law and its users. Buffalo Law Review 53, 1369–1417.

Galanter, Marc, 2011. More lawyers than people: the global multiplication of legal professionals. In: Cummings, Scott L. (Ed.), The Paradox of Professionalism: Lawyers and the Possibility of Justice. Cambridge University Press, Cambridge, pp. 68–89.

Galanter, Marc, Palay, Thomas, 1994. Tournament of Lawyers: The Transformation of the Big Law Firm. University of Chicago Press, Chicago.

Giddens, Anthony, 1999. Risk and responsibility. The Modern Law Review 62 (1), 1–10.

Ginsburg, Tom, 2003. Judicial Review in New Democracies: Constitutional Courts in Asian Cases. Cambridge University Press, Cambridge.

Goldberg-Hiller, Jonathan, 2002. The Limits to Union: Same-Sex Marriage and the Politics of Civil Rights. University of Michigan Press, Ann Arbor.

Gómez, Laura E., 2000. Race, colonialism, and criminal law: Mexicans and the American criminal justice system in territorial New Mexico. Law and Society Review 34 (4), 1129–1202.

Gómez, Laura E., 2010. Understanding law and race as mutually constitutive: an invitation to explore an emerging field. Annual Review of Law and Social Science 6, 487–505.

Gómez, Laura E., 2012. Looking for race in all the wrong places. Law & Society Review 46 (2), 221–245.

Gorman, Elizabeth H., 2005. Gender stereotypes, same-gender preferences, and organizational variation in the hiring of women: evidence from law firms. American Sociological Review 70 (4), 702–728.

Gruchalla-Wesierski, Tadeusz, 1984. A framework for understanding soft law. McGill Law Journal 30, 37.

Gunningham, Neil, 1987. Negotiated non-compliance: a case study of regulatory failure. Law & Policy 9 (1), 69–95.

Gunningham, Neil A., Kagan, Robert A., Thornton, Dorothy, 2003. Shades of Green: Business, Regulation, and Environment. Stanford University Press, Stanford, CA.

Hadfield, Gillian K., 2000. The price of law: how the market for lawyers distorts the justice system. Michigan Law Review 98, 953–1006.

Hadfield, Gillian K., 2009. Higher demand, lower supply? A comparative assessment of the legal resource landscape for ordinary Americans. Fordham, Urban Law Journal 37, 129–156.

Hagan, John, 1990. The gender stratification of income inequality among lawyers. Social Forces 68 (3), 835–855.

Hagan, John, Rymond-Richmond, Wenona, 2009. Darfur and the Crime of Genocide. Cambridge University Press, Cambridge.

Hagan, John, Ferrales, Gabrielle, Jasso, Guillermina, 2008. How law rules: torture, terror, and the normative judgments of Iraqi judges. Law & Society Review 42 (3), 605–644.

Handler, Joel F., 1967. The Lawyer and His Community: The Practicing Bar in a Middle-Sized City. University of Wisconsin Press, Madison.

Handler, Joel F., Hollingsworth, Ellen J., Erlanger, Howard S., 1978. Lawyers and the Pursuit of Legal Rights. Academic Press, New York.

Haney-López, Ian, 2007. "A nation of minorities": race, ethnicity, and reactionary colorblindness. Stanford Law Review 59 (4), 985–1063.

Haney-López, Ian F., 2000. Institutional racism: judicial conduct and a new theory of racial discrimination. Yale Law Journal 109 (8), 1717–1884.

Hart, H.L.A., 1961. The Concept of Law. Oxford University Press, Oxford.

Hartman, Chester, Squires, Gregory D. (Eds.), 2006. There Is No Such Thing as a Natural Disaster: Race, Class, and Hurricane Katrina. Routledge, New York.

Hawkins, Keith, 1984. Environment and Enforcement: Regulation and the Social Definition of Pollution. Oxford University Press, Oxford.

Heinz, John P., Nelson, Robert L., Sandefur, Rebecca L., Laumann, Edward O., 2005. Urban Lawyers: The New Social Structure of the Bar. University of Chicago Press, Chicago.

Heyer, Katharina C., 2000. From special needs to equal rights: Japanese disability law. Asian-Pacific Law & Policy Journal 1, 6.

Heyer, Katharina C., 2002. The ADA on the road: disability rights in Germany. Law & Social Inquiry 27 (4), 723–762.

Hirsch, Susan F., 2008. Fear and accountability at the end of an era. Law & Society Review 42 (3), 591–604.

Ivković, Sanja K., Hagan, John, 2006. The politics of punishment and the siege of sarajevo: toward a conflict theory of perceived international (in)justice. Law & Society Review 40 (2), 369–410.

Jenness, Valerie, 1999. Managing differences and making legislation: social movements and the racialization, sexualization, and gendering of federal hate crime law in the US, 1985–1998. Social Problems 46, 548.

Kelemen, R. Daniel, Sibbitt, Eric S., 2004. The globalization of American law. International Organization 58 (Winter), 103–136.

Klug, Heinz, 2000. Constituting Democracy: Law, Globalism and South Africa's Political Reconstruction. Cambridge University Press, Cambridge.

Kostiner, Idit, 2003. Evaluating legality: toward a cultural approach to the study of law and social change. Law & Society Review 37 (2), 323–368.

Lipsky, Michael, 1980. Street-Level Bureaucracy: Dilemmas of the Individual in Public Services. Russell Sage Foundation, New York.

Lokaneeta, Jinee, 2010. A rose by another name: legal definitions, sanitized terms, and imagery of torture in 24. Law, Culture and the Humanities 6 (2), 245–273.

Luker, Kristin, 1984. Abortion and the Politics of Motherhood. University of California Press, Berkeley.

Macaulay, Stewart, 1963. Non-contractual relations in business: a preliminary study. American Sociological Review 28 (1), 55–67.

Marshall, Anna-Maria, 2003. Injustice frames, legality, and the everyday construction of sexual harassment. Law & Social Inquiry 28 (3), 659–689.

Marshall, Anna-Maria, 2005a. Confronting Sexual Harassment: The Law and Politics of Everyday Life. Ashgate Publishing, Burlington, VA.

Marshall, Anna-Maria, 2005b. Idle rights: employees' rights consciousness and the construction of sexual harassment policies. Law & Society Review 39 (1), 83–124.

Massoud, Mark F., 2011. Do victims of war need international law? Human rights education programs in authoritarian Sudan. Law & Society Review 45 (1), 1–32.

Mayhew Jr., Leon, R., Albert J., 1969. The social organization of legal contacts. American Sociological Review 34 (3), 309–318.
Maynard-Moody, Steven, Musheno, Michael, 2012. Social equities and inequities in practice: street-level workers as agents and pragmatists. Public Administration Review 72 (s1), 16–23.
McCann, Michael W., 1994. Rights at Work: Pay Equity Reform and the Politics of Legal Mobilization. University of Chicago Press, Chicago.
Menjívar, Cecilia, Abrego, Leisy J., 2012. Legal violence: immigration law and the lives of Central American immigrants. American Journal of Sociology 117 (5), 1380–1421.
Merry, Sally E., 1979. Going to court: strategies of dispute management in an American urban neighborhood. Law and Society Review 13 (4), 891–925.
Merry, Sally E., 1990. Getting Justice and Getting Even: Legal Consciousness among Working-Class Americans. University of Chicago Press, Chicago.
Michelson, Ethan, 2007. Climbing the dispute pagoda: grievances and appeals to the official justice system in rural China. American Sociological Review 72 (3), 459–485.
Miller, Richard E., Sarat, Austin, 1980. Grievances, claims, and disputes: assessing the adversary culture. Law & Society Review 15 (3/4), 525–566.
Minow, Martha, 1987. Interpreting rights: an essay for Robert cover. The Yale Law Journal 96 (8), 1860–1915.
Mnookin, Robert H., Kornhauser, Lewis, 1979. Bargaining in the shadow of the law: the case of divorce. The Yale Law Journal 88 (5), 950–997.
Morrill, Calvin, Edelman, Lauren B., Tyson, Karolyn, Arum, Richard, 2010. Legal mobilization in schools: the paradox of rights and race among youth. Law & Society Review 44 (3–4), 651–694.
Moustafa, Tamir, 2003. Law versus the state: the judicialization of politics in Egypt. Law & Social Inquiry 28 (4), 883–930.
Muir, William K., 1979. Police: Street Corner Politicians. University of Chicago Press, Chicago.
Nelken, David, 1997. Comparing Legal Cultures. Dartmouth Pub Co, Hanover.
Nelson, Robert L., Bridges, William, 1999. Legalizing Gender Inequality: Courts, Markets and Unequal Pay for Women. Cambridge University Press, Cambridge.
Nelson, Robert L., Nielsen, Laura B., 2000. Cops, counsel, and entrepreneurs: constructing the role of inside counsel in large corporations. Law & Society Review 34 (2), 457–494.
Nielsen, Laura B., 2000. Situating legal consciousness: experiences and attitudes of ordinary citizens about law and street harassment. Law & Society Review 34 (4), 1055–1090.
Nielsen, Laura B., 2004. License to Harass: Law, Hierarchy, and Offensive Public Speech. Princeton University Press, Princeton.
Nielsen, Laura B., Nelson, Robert L., 2005. Rights realized? An empirical analysis of employment discrimination litigation as a claiming system. Wisconsin Law Review, 663–711.
Obasogie, Osagie K., 2013. Blinded by Sight: Seeing Race through the Eyes of the Blind. Stanford University Press, Stanford, CA.
Pager, Devah, 2007. Marked: Race, Crime, and Finding Work in an Era of Mass Incarceration. University of Chicago Press, Chicago.
Posner, Richard A., 1975. The economic approach to law. Texas Law Review 53, 757–782.
Priest, George L., Klein, Benjamin, 1984. The selection of disputes for litigation. The Journal of Legal Studies 13 (1), 1–55.
Reichman, Ravit, 2009. The Affective Life of Law: Legal Modernism and the Literary Imagination. Stanford University Press, Stanford.
Rosen, Robert Eli, 2002. We're all consultants now: how changes in client organizational strategies influence change in the organization of corporate legal services. Arizona Law Review 44, 637.
Rosenberg, Gerald N., 1991. The Hollow Hope: Can Courts Bring about Social Change? University of Chicago Press, Chicago.
Rosenberg, Janet, Perlstadt, Harry, Phillips, William R.F., 1993. Now that we are here: discrimination, disparagement, and harassment at work and the experience of women lawyers. Gender & Society 7 (3), 415–433.
Ross, H. Laurence, 1970. Settled Out of Court: The Social Process of Insurance Claims Adjustment. Aldine Publishing, Hawthorne, NY.
Ross, Richard J., 2002. Communications revolutions and legal culture: an elusive relationship. Law & Social Inquiry 27 (3), 637–684.
Sarat, Austin, 1990. The law is all over: power, resistance and the legal consciousness of the welfare poor. Yale Journal of Law & the Humanities 2, 343.
Sarat, Austin, Scheingold, Stuart A. (Eds.), 2006. Cause Lawyers and Social Movements. Stanford University Press, Stanford.
Sarat, Austin, Scheingold, Stuart A. (Eds.), 2001. Cause Lawyering and the State in a Global Era. Oxford University Press, Oxford.
Scheingold, Stuart A., 1974. The Politics of Rights: Lawyers, Public Policy and Social Change. Yale University Press, New Haven.
Scholz, John T., 1991. Cooperative regulatory enforcement and the politics of administrative effectiveness. The American Political Science Review 85 (1), 115–136.
Segal, Jeffrey A., Spaeth, Harold J., 2002. The Supreme Court and the Attitudinal Model Revisited. Cambridge University Press, Cambridge.
Seidl-Hohenveldern, Ignaz, 1999. International Economic Law. Kluwer Law International, Dordrecht.
Shamir, Ronen, 2004. Between self-regulation and the alien tort claims act: on the contested concept of corporate social responsibility. Law & Society Review 38 (4), 635–664.
Shklar, Judith N., 1990. The Faces of Injustice. Cambridge University Press, Cambridge.
Silverstein, Helena, 1996. Unleashing Rights: Law, Meaning, and the Animal Rights Movement. University of Michigan Press, Ann Arbor.
Stone, Christopher D., 1975. Where the Law Ends: The Social Control of Corporate Behavior. Harper & Row, New York.
Stryker, Robin, 2000. Legitimacy processes as institutional politics: implications for theory and research in the sociology of organizations. Research in the Sociology of Organizations 17, 179–223.
Suchman, Mark C., Cahill, Mia L., 1996. The hired gun as facilitator: lawyers and the suppression of business disputes in Silicon Valley. Law & Social Inquiry 21 (3), 679–712.
Suk, Julie C., 2010. Are gender stereotypes bad for women? Rethinking antidiscrimination law and work-family conflict. Columbia Law Review 110 (1), 1–69.
Sutton, John R., Dobbin, Frank, Meyer, John W., Scott, W. Richard, 1994. The legalization of the workplace. American Journal of Sociology 99 (4), 944–971.
Talesh, Shauhin A., 2009. The privatization of public legal rights: how manufacturers construct the meaning of consumer law. Law & Society Review 43 (3), 527–562.
Tammes, A.J.P., 1983. Soft law. In: Radke Arbor, E. (Ed.), Essays on International & Comparative Law. Springer Netherlands, Houten, pp. 187–195.
Teubner, Gunther, 1987. Autopoietic Law - A New Approach to Law and Society. Walter de Gruyter & Co., Berlin.
Thacher, David, 2005. The local role in homeland security. Law & Society Review 39 (3), 635–676.
Trubek, David M., Mosher, Jim, Rothstein, Jeffrey S., 2000. Transnationalism in the regulation of labor relations: international regimes and transnational advocacy networks. Law & Social Inquiry 25 (4), 1187–1211.
Tushnet, Mark V., 1987. The NAACP's Legal Strategy against Segregated Education, 1925–1950. University of North Carolina Press, Chapel Hill.
Tyler, Tom R., 1990. Why People Obey the Law. Princeton University Press, Princeton.
Tyler, Tom R., Huo, Yuen J., 2002. Trust in the Law: Encouraging Public Cooperation with the Police and Courts. Russell Sage Foundation, New York.
Vaughan, Diane, 1985. Controlling Unlawful Organizational Behavior: Social Structure and Corporate Misconduct. University of Chicago Press, Chicago.
Wiegand, Wolfgang, 1996. Americanization of law: reception or convergence. In: Friedman, Lawrence M., Scheiber, Harry N. (Eds.), Legal Culture and the Legal Profession. Westview Press, Boulder, CO, pp. 137–152.
Wiles, Ellen, 2007. Headscarves, human rights, and harmonious multicultural society: implications of the French Ban for interpretations of equality. Law & Society Review 41 (3), 699–736.
Williams, Patricia J., 1991. The Alchemy of Race and Rights: Diary of a Law Professor. Harvard University Press, Cambridge.
Yashar, Deborah J., 2005. Contesting Citizenship in Latin America: The Rise of Indigenous Movements and the Postliberal Challenge. Cambridge University Press, Cambridge.

Laws, Conflict of

Symeon C Symeonides, Willamette University College of Law, Salem, OR, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Conflict of laws is the body of law through which each state (or country) seeks to resolve international or interstate private-law disputes between persons or entities other than states as such. A dispute is international or interstate if one or more of its constituent elements, such as the events that give rise to the dispute, the location of its object, or the parties' nationality, citizenship, domicile, residence, or other affiliation, are connected with more than one country or state. Conflicts law consists of three parts: (1) jurisdiction concerns which of the involved states' courts or other organs will adjudicate the dispute; (2) choice of law concerns whether the merits of the dispute will be resolved under the substantive law of the state of adjudication or instead under the law of another involved state; and (3) judgment-recognition concerns the requirements under which the courts of one state will recognize and enforce a judgment rendered in another state.

Definitions

Conflict of Laws

The term 'conflict of laws' describes generally the body of law that seeks to provide solutions to international or interstate legal disputes between private or public persons or entities other than countries or states as such. A dispute is considered international, interstate, or multistate if one or more of its constituent elements is connected with more than one country or state. For brevity's sake, the word 'state' is used hereafter to denote any country or any territorial subdivision of such country, such as a state or province, that has its own system of private law. The law of conflict of laws deals with both international (country-to-country or country-to-state) and intranational (state-to-state) cases, so long as they have one or more foreign elements. These elements may relate to the parties' nationality, citizenship, domicile, residence, or other affiliation with another state; to the location of the events that give rise to the dispute; or to the location of the object of the dispute. Examples of such cases include contractual disputes between citizens of different countries or domiciliaries or residents of different states; torts resulting from conduct occurring in one state and causing injury in another state or involving domiciliaries of different states; or property disputes between residents of one state regarding assets situated in another state.

Private International Law

The term 'conflict of laws' was coined in Europe by the Dutch jurist Ulricus Huber (1624–94) and is the term primarily used in the United States, Canada, and more recently in England. In the rest of the world, the corresponding term is 'private international law,' which was coined in the United States by Justice Joseph Story (1779–1845). These two terms reflect different and also debatable assumptions about the nature, scope, and function of this subject.

The term conflict of laws seems to assume that (1) in all multistate cases, each involved state has an active or passive desire, claim, or 'interest' to apply its own law; (2) these claims 'conflict' in the sense of always pulling in opposite directions; and (3) there exists an impartial mechanism of hierarchically superior authority for refereeing these conflicts. These assumptions have been questioned throughout history. For example, even accepting the propriety of using anthropomorphic terms to describe state objectives, one can question whether (or why) a state would be interested in the outcome of disputes between private persons. Similarly, one can question whether such a state interest is implicated in each case connected with that state, or whether the application of that state's law indeed effectuates that interest.

The term private international law seems less loaded and more descriptive, but also more idealistic than the term conflict of laws. The word international describes an important attribute of the disputes that fall within the scope of this subject – they are international (or interstate) in that they have contacts with more than one state. The word private serves a dual function: (1) it confines the scope of this subject to private-law disputes, namely disputes between private persons other than a state in the exercise of governmental authority; and (2) it distinguishes this subject from public international law, also known as the law of nations, which regulates the conduct of sovereign states, analogous entities, and international organizations at the international level (*see* International Law and Treaties).

However, the word private also allows an inference that the disputes that comprise this subject implicate only the interests of the disputants and not the interests of their respective home states or the states that have other pertinent contacts with the dispute. Many critics have questioned this inference, especially during the last half century. Similarly, the word international allows the inference that this law emanates from a supranational source. Reality is much different. Aside from a few international conventions that avoid or resolve conflicts through substantive or conflicts rules, international law provides little guidance on the subject. Nevertheless, the word international reflects the initial internationalist aspirations of this field, especially in the nineteenth century. The classic school of private international law taught that, in entering this field, national lawmakers should act as surrogates of a nonexistent international legislature. They should act unselfishly, impartially, and evenhandedly, treating equally foreign and domestic laws, as well as foreign and domestic litigants. They should aim for international harmony and uniformity by adopting only those rules that would be capable of 'internationalization' through their adoption by

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86064-X

other nations. How much, if any, of this initial idealism survives today is very much debatable.

The Sources of Conflicts Law

Thus, despite the appellation international, private international law or conflicts law is essentially national law in the sense that each country decides for itself how to resolve multistate private-law disputes, subject to relatively mild restraints imposed by public international law or suggested by international comity. From the perspective of federal countries, national law usually means federal law. Indeed, in most federal countries, the resolution of both international and intranational conflicts of laws falls within the competence of the central law–making body. However, in the United States, the resolution of interstate conflicts as well as conflicts between state law and foreign law (as opposed to conflicts between federal law and foreign law) is primarily left to the individual states, despite the existence of a federal constitutional power to federalize the entire field.

The Source of Conflicts of Laws

Legal Diversity

Conflicts of laws occur because each state has its own system of private law and because human activity often is not confined to a single state. If either one of these conditions is missing, conflicts do not arise. One who lives in complete isolation and avoids contact with nonlocal residents and foreign products need not be concerned with conflicts of laws. Similarly, one whose contacts with nonlocals are confined to residents of states that have identical laws would not need to be concerned with problems caused by legal diversity. Needless to say, human activity transcends national borders, and countries do have different laws. For example, the contracts law of France differs from that of England, which in turn differs in some respects from that of Scotland, and the torts law of the Mexican province of Sinaloa differs from that of the State of Arizona, which in turn differs from that of California. Indeed, differences exist not only between countries that belong to different legal traditions, such as France and England which belong to civil-law and common-law traditions, respectively (*see* Legal Systems, Classification of), but also between countries that partake of the same legal tradition, such as France and Italy, or England and the United States, respectively. Moreover, such differences also exist as between or among the component units of a single country, such as the states of the United States, or the provinces of Mexico or Canada, or in the United Kingdom between England and Scotland (*see* Legal Pluralism).

Interpersonal or Intertemporal Conflicts Law

This legal diversity is usually coextensive with a distinct territory that is administered by a governmental unit with authority over that territory. In some instances, however, the existence of a distinct body of law, such as family law, the law of personal status, or the law governing indigenous tribes, is not attached to a particular territory but rather to a particular ethnic or religious group, regardless of territory. Conflicts among these laws are referred to as interpersonal or interreligious conflicts and are not discussed here. Also not discussed here are the so-called intertemporal conflicts of laws, namely conflicts that arise when a law purports to regulate, or is sought to be applied, to events or transactions that occurred before its enactment (retroactive application).

The Three Parts of Conflicts Law

While many problems of conflict of laws are resolved through negotiation, mediation, or arbitration, a good number of them end up in litigation. In planning for this eventuality, or when forced to confront it, parties involved in multistate activity must keep in mind three major questions: (1) Where can or should litigation be initiated? (2) Which law will the court apply? (3) Where can the resulting judgment be enforced? These three questions correspond to the three consecutive phases that comprise the process of judicial resolution of most conflicts problems: (1) jurisdiction, (2) choice of law, and (3) recognition and enforcement of judgments. In the United States and other common-law systems, these are also the three major divisions of the law of conflict of laws. In contrast, in civil-law systems, jurisdiction and judgment-recognition form part of international procedural law rather than of private international law, which deals primarily with choice of law and secondarily with the law of nationality and citizenship and the condition of aliens.

The jurisdictional question and the choice-of-law question are analytically distinct questions, although in many cases they are practically interrelated. Indeed, the two questions are oriented in different directions and seek to satisfy different ends. The jurisdictional question focuses on the relationship between the defendant and the state of litigation (forum state) and asks whether that relationship is sufficiently close as to justify subjecting the defendant to litigation in that state and utilizing that state's judicial resources to adjudicate the particular dispute. The choice-of-law question focuses on both parties and their dispute and seeks to identify that state, be it the forum state or another state, whose relationship to the dispute is such as to render most appropriate the application of its law to the merits of the dispute. In many cases, the same relationship that forms the basis of a state's jurisdiction to adjudicate the case can also be the basis for applying that state's law to the merits. This, however, is by no means true in all cases.

As said previously, the law of conflict of laws does not deal with public-law disputes, namely, disputes involving a state in the exercise of governmental authority, such as in the enforcement of its criminal laws. Indeed, in such disputes, all three of the above questions are merged into one – the question of jurisdiction. When a state has jurisdiction to try a criminal case, that state applies its own law and enforces the resulting judgment. It does not apply the criminal laws or, strictly speaking, enforce the criminal judgments of another state. Cases in which more than one state has concurrent jurisdiction, such as cases involving transborder crimes, are addressed and resolved as jurisdictional conflicts – for example, through extradition – rather than as choice-of-law conflicts.

Jurisdiction

In its broadest sense, jurisdiction is the power of a decision-maker to resolve a dispute in a binding manner. When, as in the context pertinent here, the decision-maker is a court or other tribunal, the dispute is resolved through adjudication and the power to do so is called judicial or adjudicatory jurisdiction. It can be contrasted with legislative or prescriptive jurisdiction, which is the power to enunciate norms governing the merits of the dispute.

The source of judicial jurisdiction is the law that constitutes the particular court and thus it is a national or state law rather than a supranational or suprastate law. However, in some countries, the exercise of jurisdiction is restricted by international treaties. In some federal systems like the United States, the exercise of jurisdiction by the constituent states must conform to certain outer limits imposed by federal law, primarily the due process clause of the federal constitution. A similarly uniform regime of jurisdiction (and judgment-recognition) exists within the European Union (but only with regard to defendants domiciled within the Union) after the adoption of the Brussels I Regulation (previously the Brussels Convention). Although these superarching sources impose a certain degree of uniformity on the affected states, the jurisdictional laws of the various countries or states continue to differ widely in both nomenclature and substance.

Thus, each state delineates for itself the bases on which its courts are authorized to adjudicate cases with nonlocal elements (jurisdiction in the international sense) and then allocates internally that authority among the various geographical subdivisions or judicial districts within that state (internal or local jurisdiction or venue) and among its various general or specialized courts, tribunals, or other adjudicating agencies in each locality (competence or subject-matter jurisdiction). To render a valid judgment, a court must, in the absence of acquiescence by the defendant, satisfy all three types of jurisdiction.

In the United States, for historical reasons, a state court's interstate or international jurisdiction is divided into personal jurisdiction (*in personam*) and jurisdiction over things (*in rem*). Personal jurisdiction, which is the power to adjudicate claims against a defendant and to bind him or her personally, is further subdivided into general and specific jurisdiction. General jurisdiction exists when the defendant is domiciled, incorporated, has its principal place of business, or conducts 'continuous and systematic' business in the forum state. Historically, the service of process on the defendant within the territory of the court was also a basis of general jurisdiction, but the constitutionality of this basis is now questionable, at least when it is not supported by other defendant contacts. When a court has general jurisdiction, the court may adjudicate any claims against the defendant, even claims unrelated to the defendant's contacts with, or activity in, that state. In contrast, when a court has only specific jurisdiction, the court may adjudicate only claims against the defendant that arise from the defendant's contacts with, or activity in, the forum state.

Jurisdiction over things is based on the presence of a thing within the territory of the forum state and is subdivided into jurisdiction *in rem* and *quasi in rem*. When the lawsuit seeks to determine the parties' rights in, or relating to, that thing, the court has jurisdiction *in rem* and may adjudicate those rights even if it does not have personal jurisdiction over the defendant. In contrast, when the lawsuit seeks to vindicate claims against the defendant that are unrelated to that thing, the court is said to have *quasi in rem* jurisdiction. Since 1977, this basis of jurisdiction is – in the absence of other defendant contacts – constitutionally insufficient unless the claim has been previously reduced to a judgment rendered by another court that had jurisdiction over the defendant.

In other countries, jurisdictional bases are in some respects broader and in other respects narrower than those just described. For example, some European countries theoretically allow the exercise of what might be called *quasi in rem* general jurisdiction, namely, the adjudication of claims against a defendant who owns property situated in the forum state, even if the claims are unrelated to the property or vastly exceed its value. (The Brussels I Regulation prohibits the use of this basis against defendants domiciled within the European Union.) On the other hand, while most countries agree that the defendant's domicile, incorporation, seat, or principal place of business in the forum state is a basis for general jurisdiction, they reject general jurisdiction based on either service of process alone or 'doing business' within the forum state.

Beyond this, most countries and states agree that the defendant's actual or imputed consent is a sufficient basis of jurisdiction even when jurisdiction is otherwise lacking, that contractual claims can be adjudicated in the place where the contract was to be performed, and that claims arising from torts can be adjudicated in the place where the wrong was committed or the injury foreseeably occurred.

Because of the differences between the jurisdictional bases recognized by the various states, but also because, even when they do not disagree, these states may authorize more than one jurisdictional basis for certain categories of cases, a given case may often be subject to the concurrent jurisdiction of more than one state. In such a case, the plaintiff, who always has the choice of where to file the lawsuit, may 'shop' for the most advantageous forum, a phenomenon known as 'forum shopping.' One of the mechanisms devised by common-law systems to discourage this phenomenon is the doctrine of forum *non conveniens*, which allows a court to decline to adjudicate the lawsuit if, despite the existence of jurisdiction, litigation in that state would be seriously inconvenient and provided that another more convenient forum is available to the plaintiff. Another mechanism that is sanctioned in varying degrees by most systems but is available only when the disputants are parties to a preexisting contractual relationship, is a 'choice-of-forum' clause, namely a clause by which the parties agree that disputes arising from that relationship will be litigated in the courts of the agreed-upon state.

Choice of Law

When a court exercises its jurisdiction to adjudicate a case with foreign elements, the court applies its own procedural law to the conduct of the proceedings before it. With regard to the merits of the case, however, the court may or may not apply its own substantive law. This is the choice-of-law question and it is answered by the forum's choice-of-law rules, which may be established legislatively (as in most civil-law systems) or

through judicial precedent (as in most common-law systems, including those of the United States). These rules may point either to the law of the forum state or to the law of another state, depending, inter alia, on each state's pertinent contacts with the case. For example, in tort cases, these rules may point to the state where the tort was committed (*lex loci delicti*) or the injury occurred (*lex loci damni*), in contract cases to the state where the contract was made (*lex loci contractus*), and, in cases involving immovable property, to the state where the property is situated (*lex rei sitae*).

The above rules are typical of what is usually referred to as a traditional choice-of-law system. For much of the twentieth century, these rules and others like them were more or less uniformly followed in most countries. These rules place a high premium on providing certainty and predictability in the choice of the governing law, leaving little room for judicial discretion. As the above examples illustrate, these rules give no preference to the forum state, and indeed they aspire to produce international or interstate uniformity, that is, make it more likely that each multistate case will be governed by the same law regardless of where the case is litigated. The choice of the governing law is based not on the content of the laws of the involved states but rather on the territorial or other contacts of those states to the case at hand. Once a state is found to have the predesignated contact, e.g., the place of the injury, that state's law is applied almost automatically and – save for a few limited exceptions – regardless of its content, its underlying policy, or the substantive quality of the result that such application produces. Indeed, the objective of the traditional choice-of-law process is not supposed to be to ensure a substantively just result (*material justice*), but rather to ensure the application of the spatially appropriate law (*conflicts justice*).

Traditional choice-of-law rules like the ones described above continue to be followed in many countries, including about a dozen states of the United States. However, in the 1960s, a movement emerged in the United States which has been characterized as a conflicts revolution, and which appeared to reject both the rules and the goals of the traditional choice-of-law process. At least in tort and contract conflicts (which are the most numerous), preconceived choice-of-law rules were abandoned in favor of various approaches, which, although different in many respects, reject the notion of preselecting the applicable law on the basis of a single territorial contact and instead leave the selection to be made by the court on a case-by-case basis. The choice is to be based on multiple contacts and factors, such as the content of the conflicting substantive laws and their underlying policies, the presumed interests or claims of the involved states to apply their respective laws, and the substantive quality of the result that the chosen law will produce in the individual case.

Parallel movements have also appeared in other countries, though nowhere with the same intensity. For example, European systems now appear more willing than in the past to entrust courts with a certain degree of flexibility in choosing the applicable law. Although most of these systems still rely on statutory choice-of-law rules, the rules enacted in the last four decades are 'softer' than those of the preceding era. For example, some of these rules base the choice of the applicable law not on a single territorial contact but rather on multiple contacts, usually described by the phrase 'strongest' or 'closest connection,' while other rules contain clauses that authorize courts to deviate from the predesignated law in appropriate circumstances, depending on the needs of the particular case. Some of these features can be seen in the European Union's 'Rome I' and 'Rome II' Regulations, which provide for the law governing contractual and noncontractual obligations, respectively.

As a result of the above movements, the choice-of-law process at the beginning of the twenty-first century is more flexible, less predictable, and more ethnocentric than it was only one generation earlier. The possibility that the forum state will apply its own law to most multistate cases adjudicated by its courts is now greater than in much of the previous century, especially in the United States. In turn, this increases the possibility that a given case will be governed by a different law depending on where it is litigated, and this in turn reinforces the incentive for forum shopping described earlier.

Yet, while national conflicts systems have been gradually diverging in most respects, they have also been converging in at least one important respect. Most of them now recognize the principle of party autonomy, which grants contracting parties the power to agree in advance on the law that will govern disputes arising from their relationship. Although this is an ancient principle whose origin can be traced as far back as ancient Greece, it has had a checkered history through the centuries. By the second half of the twentieth century, however, party autonomy acquired the status of a universal principle and is now sanctioned in one form or another by most countries in the world. In fact, some of these countries have extended this principle beyond ordinary contracts and have made it available to matrimonial agreements, other family-law agreements, and even unilateral juridical acts such as testaments.

The use of a choice-of-law clause, especially when combined with a choice-of-forum clause or an arbitration clause, can provide parties with a modicum of predictability. This is a good antidote to the uncertainty generated by the movements described above. This antidote, however, can only work in some cases. For example, the principle of party autonomy is not available in cases such as torts, in which the disputants are not parties to a preexisting relationship. Second, while this principle is recognized in most countries, it is also subject to several limitations and exceptions that differ from country to country. Thus, the possibility remains that choice-of-law clauses can be enforced in some countries but not in others, along with the possibility of a different outcome depending on where the case is litigated.

Recognition and Enforcement of Foreign Judgments

One of the many differences between local and multistate cases is the increased possibility that in multistate cases a judgment may have to be enforced in a state other than the one in which it was rendered. This will be the case if the defeated litigant (judgment debtor) does not voluntarily comply with the judgment and does not have sufficient assets in the forum state against which the successful litigant (judgment creditor) can execute the judgment. In such a case,

the creditor will seek to have the judgment recognized and enforced in another state that has jurisdiction over the debtor and in which the debtor has sufficient assets. This section discusses the requirements under which the second state will recognize such a judgment.

As between states of the United States, these requirements are prescribed by federal law, primarily the Full Faith and Credit Clause of the Constitution. As interpreted by the Supreme Court, this clause requires each state to give to a judgment of a sister state the same effect that the judgment had in the state of rendition, provided that the judgment is 'final,' it is 'on the merits,' and it was rendered by a court that had jurisdiction under its own law which must conform with federal due process standards. Once these requirements are met, the judgment must be recognized and may not be reviewed on the merits, even if it applied the wrong law or reached a result that contravenes the public policy of the second state. The reason for this strict regime is the national policy of finality embodied in the doctrine of *res judicata*, which seeks to conserve judicial resources and protect party expectations by prohibiting relitigation of disputes that have been finally decided. In 1996, the Defense of Marriage Act (DOMA) introduced the only exception to this regime by providing that "(n)o State shall be required to give effect to any … judicial proceeding in any other State … respecting a relationship between persons of the same sex that is treated as a marriage under the laws of such other State … or a right or claim arising from such a relationship."

The Full Faith and Credit Clause does not apply to the judgments of foreign countries and thus American courts are not constitutionally compelled to enforce such judgments. Nevertheless, as a matter of comity, American courts recognize foreign-country judgments that meet the requirements for recognizing a sister-state judgment, but also reserve the right to refuse recognition on certain grounds that are not available with regard to sister-state judgments. For example, recognition may be refused if the judgment does not satisfy American jurisdictional standards, even if it does satisfy the foreign-law standards; if the judgment is one for taxes or penalties; or if it was rendered in a system that does not provide impartial tribunals or does not satisfy certain minimum standards of procedural fairness, or contravenes the public policy of the recognizing court.

In many other countries, foreign judgments are recognized under similar but somewhat less hospitable standards. For example, in some countries, recognition depends on whether the rendering court had applied the law that would have been chosen under the choice-of-law rules of the recognizing court or had reached a substantially equivalent result. In other countries, recognition is usually denied on public policy grounds to judgments awarding 'excessive' sums of money, especially American judgments imposing punitive damages. Finally, some countries condition recognition on reciprocity, that is, on whether the country of rendition would recognize similar judgments from the second country, while other countries refuse recognition outright in the absence of a treaty requiring such recognition.

One such treaty was the Brussels/Lugano Convention on Jurisdiction and the Recognition and Enforcement of Judgments among the member states of the European Union and the European Free Trade Association. This Convention established a comprehensive regime for jurisdiction and judgment-recognition that approximates in important respects the regime that exists within the United States. With regard to defendants domiciled in a member state, the Convention defined the permissible bases of jurisdiction and then provided that a judgment rendered by a court that satisfies those bases must be recognized in the other member states without review of the merits and subject to only limited exceptions such as public policy, lack of proper and timely notice, or conflict with a prior judgment entitled to recognition under the Convention or a judgment of the recognizing state. In 2001, the Brussels Convention was converted into a Regulation (known as 'Brussels I'), which is directly binding on 26 of the 27 EU countries, and, in 2007, the Lugano Convention was accordingly renewed.

In 2005, a Convention on Choice of Court Agreements was negotiated under the auspices of the Hague Conference of Private International Law. The Convention provides for the recognition of judgments rendered by a court chosen by the parties under a choice-of-court agreement. The Convention has been signed by the European Union and the United States but it is not yet in force.

See also: International Law and Treaties; Legal Systems: Private.

Bibliography

Audit, B., 2010. Droit International Privé (Private International Law), sixth ed. Economica, Paris.
Batiffol, H., Lagarde, P., 1993. Traité de Droit International Privé (Treatise on Private International Law), eighth ed. Librairie générale de droit, Paris.
Bucher, A., Bonomi, A., 2004. Droit International Privé (Private International Law), second ed. Helbing & Lichtenhahn, Bale, Switzerland.
Castel, J.G., 1997. Canadian Conflict of Laws, fourth ed. Butterworths, Toronto, ON, Canada.
Cheshire, G., North, P., Fawcett, J., 2008. Private International Law, fourteenth ed. Oxford, Oxford, UK.
Davies, M., Bell, A., Bereton, P., 2010. Nygh's Conflict of Laws in Australia. Lexis-Nexis.
Dicey, A., Morris, J., Collins, L., 2006. Conflict of Laws, fourteenth ed. Collins Sweet & Maxwell, London.
Esplugues Mota, C., Iglesias Buhigues, J., 2011. Derecho Internacional Privado (Private International Law), fifth ed. Tirant Lo Blanch, Valencia, Spain.
Felix, R., Whitten, R., 2011. American Conflicts Law, sixth ed. Transnational, Ardsley, New York, NY.
Hay, P., Borchers, P., Symeonides, S., 2010. Conflict of Laws, fifth ed. West, St. Paul, MN.
Kegel, G., Schurig, K., 2004. Internationales Privatrecht (Private International Law), ninth ed. C.H. Beck, Munich, Germany.
Loussouarn, Y., Bourel, P., Vareilles-Somières, P., 2007. Droit International Privé (Private International Law), ninth ed. Dalloz, Paris.
McClean, D., Beevers, K., 2009. Morris (on) the Conflict of Laws, seventh ed. Collins Sweet & Maxwell, London.
Rigaux, F., Fallon, M., 2005. Droit International Privé (Private International Law), third ed. Larcier, Brussels, Belgium.
Stone, P., 2010. EU Private International Law, second ed. Elgar, Cheltenham, UK.
Symeonides, S., 2006. The American Choice-of-Law Revolution: Past, Present and Future. Martinus Nijhoff, Leiden, The Netherlands.
Symeonides, S., 2008. American Private International Law. Wolters Kluwer, The Netherlands.
Verschraegen, B., 2012. Internationales Privatrecht (Private International Law). Manz, Wien.
Verschraegen, B. (Ed.), 2001–2012, International Encyclopaedia of Laws: Private International Law, vols 1–4. Kluwer Law International, The Netherlands.
Weintraub, R., 2010. Commentary on the Conflict of Laws, sixth ed. Thomson Reuters/Foundation Press, New York, NY.

Lawyers, Regulation of

Laurel S Terry, Penn State Dickinson School of Law, Carlisle, PA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E. Chambliss, volume 12, pp. 8559–8564, © 2001, Elsevier Ltd.

Abstract

This article explains that the topic of 'regulation of lawyers' can be difficult to research because of jurisdictional differences in the meaning of the term lawyer and because of differences in the sources and types of lawyer regulation and differences in the reserved activities for which one needs to be a licensed lawyer. Although there is no centralized source that identifies lawyer regulators in different jurisdictions or the content of that regulation, this article identifies resources that can be useful when researching lawyer regulation. The article also identifies current issues and changes that have occurred in lawyer regulation.

Introduction

This article focuses on the contemporary regulation of lawyers (see Section Definitions). For useful perspectives about the goals and intention of lawyer regulation and for historical information, see the articles on Lawyers, Professionalization/Professions in History, Legal Education, and Law Firms.

Definitions

Before addressing the regulation of lawyers, there are several definitional issues one must resolve. What does one mean by the term 'lawyer'? What does one mean by 'regulation'? What time period of regulation is one referring to? Should one refer to the 'regulation of legal services' rather than 'regulation of lawyers?'

According to the "1998 Background Note on Legal Services" by the World Trade Organization (WTO) Secretariat, experts have identified the following as the main legal families: Romano-Germanic law (which usually referred to as civil law), common law, socialist law, Hindu law, Muslim Law, laws of the Far East, and black Africa and Malagasy law. The University of Ottawa's Juriglobe website identifies the type of legal system used in most countries in the world, including whether the country has a mixed legal system. This article focuses primarily on the regulation of lawyers in civil law and common-law systems. Socialist law has largely disappeared and that some countries in Africa and Asia have common law or civil law systems.

Comparing the regulation of lawyers from different jurisdictions requires a shared definition of 'lawyer.' Yet the Anglo-American common law concept of 'lawyer' has no counterpart in civil law countries because most civil law countries do not have a unified legal profession. Instead, civil law countries recognize multiple legal occupations, such as judge, prosecutor, civil servant, and private practitioner, with private practitioners typically constituting a relatively small proportion of the total. Furthermore, in civil law countries, the list of 'legal' occupations includes jobs that in common law countries would be performed by those without legal training, such as notaries (everywhere), police chiefs (in Brazil and Norway), and bailiffs and process servers (in France). None of these groups sees itself as part of a unified legal 'profession'; typically, each group has its own occupational associations and licensing criteria (Abel, 1988). In some jurisdictions, such as Belgium and certain other jurisdictions in Europe, legal professionals who provide legal services solely for their employers are not regarded as independent and therefore are not entitled to professional privilege nor are they entitled to register as a licensed (private practitioner) lawyer (Fish, 2004; Terry, 2011). For systems with a divided rather than a unified legal profession, this article focuses primarily on lawyers in private practice, rather than prosecutors, in-house counsel, judges, civil servants, or others. For additional information, *see* Lawyers: Social Organization of the Profession; Prosecution; In-House Counsel.

Comparing the regulation of lawyers from different jurisdictions also requires an understanding that there are significant variations among jurisdictions with respect to the activities that are reserved exclusively to 'lawyers' (as that term is defined previously). Some jurisdictions, such as Finland and Sweden, have no reserved activities for lawyers, whereas other jurisdictions, such as Italy, limit reserved activities to courtroom appearances, and still other jurisdictions have a much broader monopoly for their lawyers (Paterson et al., 2003). The United States and Germany are examples of jurisdictions that have a broad monopoly that includes transactional work as well as appearance before a court (ABA CPR, Task Force on the Model Definition of the Practice of Law; Paterson et al., 2003). Some jurisdictions, such as England and Wales, reserve specified litigation and limited transactional work, such as property conveyance, but they have a broad space in which those who are not licensed solicitors or barristers may operate (U.K. Legal Services Act, Section 12 and Schedule 2). If someone who is not authorized engages in activities reserved to lawyers, they sometimes are said to be engaging in the 'unauthorized practice of law.' It is worth noting that there also can be significant differences of opinion within a jurisdiction as to whether a particular activity is reserved to lawyers, as has happened in India, the United States, and elsewhere (Krishnan, 2010; Rhode, 1981). Thus, to fully understand a particular jurisdiction's regulation of lawyers, it is useful to be familiar with the scope of activities reserved to lawyers in that jurisdiction and whether there are any controversies with respect to that issue.

Turning to the question of 'what is regulation,' one must understand that many different entities and many different

sources of law and norms affect lawyer conduct and serve a regulatory function (Wilkins, 1992). These include formal regulators, such as legislatures, law societies, courts, and frontline regulatory entities. Lawyers are subject to less formal or official sources of regulation by malpractice insurance carriers, risk management and ethics advisors, clients, the news media, judges, professional organizations, nongovernmental organizations (such as the American Law Institute), custom, and peer pressure (Chambliss and Wilkins, 2002; Davis, 1996; Whelan and Ziv, 2012). Governmental and nongovernmental international entities, such as the WTO, the Financial Action Task Force, the International Bar Association (IBA), and the *Union Internationale des Avocats*, may operate as soft-law regulators (Terry, 2008a). Although there are many different types of regulators and many different sources of regulation, when this article refers to 'lawyer regulation,' its focus is on formal sources of regulation. As will be noted, there are significant jurisdictional variations in the identity of the formal regulators and in the number of formal regulators in each jurisdiction.

The next definitional issue is whether the term 'lawyer regulation' is intended to encompass regulation regarding: entry into the profession, the rules governing a lawyer's conduct (e.g., rules governing competence, confidentiality, conflicts of interest, fees, and advertising), and the conditions under which one is sanctioned or expelled from the profession. Although commentators sometimes use the term 'lawyer regulation' as a short-hand reference to ethics codes and thus the second kind of regulation, in many jurisdictions 'regulation of lawyers' refers to all three time periods in the life cycle of a lawyer (Terry et al., 2012a). For example, in the Canadian provinces, the same entity may have the primary responsibility for such regulation of lawyers during all three stages of the lawyer's life cycle (Federation of Law Societies of Canada (FLSC), National Regulatory Initiatives). In other jurisdictions, such as US jurisdictions, there may be a different frontline regulator for the entry stage (e.g., a Board of Bar Examiners), for the conduct stage (e.g., the state's highest court), and for the discipline or end stage of a lawyer's life cycle (e.g., an Office of Disciplinary Counsel) (Terry et al., 2012a). In this article, 'regulation of lawyers' refers to regulation of all three time periods.

The final definitional issue is an observation that the title of this article – regulation of lawyers – implicitly assumes that regulation is of lawyers rather than regulation of legal services. As commentators have noted (Susskind, 2008), globalization and technology have made it easier to disaggregate and commodify all or parts of legal work and for nonlawyers to offer services previously considered to be legal services reserved to lawyers. Accordingly, some lawyer regulators now are explicitly asking whether they should regulate lawyers or legal services (Terry, 2013). This article focuses on the regulation of lawyers, but highlights this issue of whether regulation should focus on services or providers.

Current State of Knowledge about Lawyer Regulation

The study of lawyer regulation is a relatively young academic field that builds on prior work. Historically, lawyer regulation and legal ethics tended to be taught by practicing lawyers or judges, rather than legal academics, or it was included in on-the-job training or as part of an apprenticeship (Moliterno, 1991). Some of the earliest and best academic literature was grounded in the fields of history and sociology (Abbott, 1988; Abel and Lewis, 1988–89, Abel, 2008; Gordon 1988, Heinz and Laumann, 1982; Rueschemeyer, 1986).

In the United States, this field grew significantly in the 1970s and 1980s. In 1969, the American Bar Association (ABA) adopted its Code of Professional Responsibility, and in 1974, in the aftermath of the Watergate scandal in which a number of lawyers were involved, the ABA first required that legal ethics be taught in ABA-accredited law schools (Clark, 2000). This ABA action was influential because most US states require attendance at an ABA-accredited law school as a condition of taking a bar exam and becoming a lawyer (National Conference of Bar Examiners (NCBE) Comprehensive Guide to Bar Admissions, 2013). One consequence of the ABA's actions was a burgeoning cadre of academics who specialized in legal ethics codes and other rules related to the regulation of the legal profession. As this field evolved, scholars began to call for a rethinking of the traditional regulation paradigm. Rather than focusing only – or primarily – on professional self-regulation by way of ethics codes, scholars called for increased attention to other sources of regulation, such as judicial, legislative, and administrative action; malpractice; liability and risk management; and constitutional and international law (Cone, 1996; Davis, 1996; Freedman, 1990; Luban, 1988; Morgan, 2010; Pearce and Wald, 2012; Schneyer, 1991, 2008; Wilkins, 1992). Thus, in the United States, the field of study that used to be called 'legal ethics' increasingly is referred to more broadly as 'the law governing lawyers' (American Law Institute, Restatement, 2000). In 2009, the Professional Responsibility Section of the Association of American Law Schools published the results of a survey regarding the scope of US professional responsibility/legal ethics courses. There were 105 responses from 77 different law schools; survey questions addressed topics that included when the course was taught, the number of units, the type of instructor, the topics addressed in the course, assessment methods, course philosophy, and advice for new teachers (AALS, 2009).

Elsewhere in the world, there has been growing academic interest in the topic of lawyer regulation. Canada has a number of scholars, books, and articles devoted to this field (Cotter et al., 2012; Dodek, 2008; Woolley, 2011). In 2012, Canadian scholars established the Canadian Association for Legal Ethics. Australia, New Zealand, and the United Kingdom are home to a number of academics who specialize in lawyer regulation; they have written books and journal articles and have issued reports commissioned by regulators (Boon and Levin, 2008; Dare and Wendel, 2010; Faulconbridge et al., 2012; Fleming and Shackel, 2013; Griffiths-Baker, 2002; Nicolson and Webb, 2000; Moorehead et al., 2003; Mortensen et al., 2010; Parker and Evans, 2007; Parker and Aitken, 2011; Sommerlad et al., 2010). Some scholars study lawyer regulation in Asia (Liu, 2008; Michelson and Liu, 2010; Myerson, 1995).

Lawyer regulation also has been studied in Europe (Henssler and Prütting, 2010; Andrasi, 2013; Leubsdorf, 2001).

Although it is still relatively uncommon for European academics outside of the United Kingdom to specialize in lawyer regulation, the number of academics who do so is growing, and there are academic institutes that specialize in this topic (e.g., the University of Cologne, Institut für Anwaltsrecht). The adoption of the 1977 and 1985 European Union Directives on lawyer mobility and the 1995 WTO General Agreement on Trade in Services undoubtedly will contribute to increase scholarly interest in the topic of lawyer regulation.

The first International Legal Ethics Conference (ILEC) was held in 2004 in England; it provided a forum for academics from around the world who are interested in legal ethics and the regulation of lawyers to gather together. ILEC V was held in Banff, Canada, in 2012; its speakers and attendees came from more than 14 countries located in Africa, Asia, Australia, Europe, and North and South America. In 2010, during ILEC IV, international scholars agreed to create the International Association of Legal Ethics (IALE). The membership of IALE and the ILEC conferences demonstrate that there are a growing number of scholars from around the world who study lawyer regulation. The conference programs are available online and provide a useful resource to locate some of the most active academics in this field.

Despite the significant increase of academic professionals who study the lawyer regulation systems in their jurisdictions, the current state of cross-cultural knowledge about lawyer regulation is relatively weak. Unlike many fields, there is no single source or database to which one can look to find out which kind(s) of lawyers exist in a particular jurisdiction, the entity or entities responsible for regulating each type of lawyer in that jurisdiction, or the content of that jurisdiction's lawyer regulation. This situation is starting to change, and the final section of this article lists a number of resources, some of which were created recently, to which one can look to find out more about lawyer regulation in other countries. These information aggregators include entities commonly known by their acronyms: the ABA, Asia Pacific Economic Cooperation (APEC), Council of Bars and Law Societies of Europe (CCBE), Conference of Regulatory Officers (CORO), FLSC, IBA, Law Council of Australia, Legal Services Board (LSB), NCBE, National Organization of Bar Counsel (NOBC), Solicitors Regulation Authority (SRA), and the WTO.

Sources and Methods of Lawyer Regulation

As noted in Section Definitions, this article focuses on formal methods of lawyer regulation. Jurisdictions vary significantly with respect to whether regulation occurs on a national or subnational basis, the types of entities that engage in lawyer regulation, the methods of regulation they use, their sources of authority, and the number of regulatory layers that exists.

In a few jurisdictions, such as those in the United States, the primary source of lawyer regulation comes from the judicial branch of government, rather than a legislative act. Because US lawyers primarily are regulated on a subnational basis, it is the highest court in each US state that holds primarily responsible for regulating lawyers (Green and Zacharias, 2009). In many jurisdictions, however, an overarching legislative act establishes a regulatory framework, with a secondary level of regulation provided by a frontline regulator that issues its own regulations. In some jurisdictions, such as Canada, Japan, and others, these frontline regulators may be law societies or bar associations in which practicing lawyers participate. In other jurisdictions, such as the SRA in England and Wales, there is a coregulatory system in which oversight is assigned to bodies that might have limited lawyer participation. The resources listed in the final section include entities that have aggregated information about the identity of lawyer regulators.

Currently, one of the main differences among lawyer regulators is the degree to which they use a principles-based system of lawyer regulation, as opposed to a rules-based system. For example, the U.K. SRA has adopted an outcomes-focused regulatory approach that relies on principles, rather than rules (U.K. SRA, Outcomes Focused Regulation). To the extent that a jurisdiction uses this approach, it may use fewer of the regulatory instruments that are described in the subsequent paragraphs, or it may have less detailed regulatory instruments. Another important difference in regulatory methods is the extent to which a regulator relies on ex ante versus ex post regulation of lawyers (Terry et al., 2012a). New South Wales, Australia's ex ante approach to lawyer regulation undoubtedly has received the most publicity (Parker et al., 2010).

Jurisdictions use many different types of regulatory instruments, and not all jurisdictions use all regulatory instruments (Terry et al., 2012a). As noted, lawyer regulation sometimes is found in legislation. This legislation may be profession specific, such as the legal profession acts found in Canada, Australia, or Germany. Additionally or alternatively, legislation applicable to lawyers may be found in legislative codes that have a broader application, such as bankruptcy statutes, criminal statutes directed against money laundering, or criminal statutes that require lawyer confidentiality. Examples of the first kind of legislation include the US bankruptcy rules applicable to lawyers counseling debtors; examples of the second kind of legislation are the UK rules on money laundering, and examples of the third kind are German laws requiring lawyer confidentiality (Hellwig, 2009; Leubsdorf, 2009; Terry, 2008a).

Regulations constitute a second type of regulatory instrument; they often are adopted pursuant to a legislative delegation of power. For example, in Germany, the regulatory bar association called the *Bundesrechtsanwaltskammer* (Federal Bar Association) has adopted binding regulations (the *Berufsordnung*) for the legal profession (*Rechstanwälte*) pursuant to the authority granted to it by federal legislation known as *Bundesrechtsanwaltsordnung* or BRAO (Terry et al., 2012a). The US Securities and Exchange Commission has adopted regulations applicable to lawyers pursuant to the authority granted to it by the US Congress under the Sarbanes-Oxley law (Leubsdorf, 2009).

A third source of lawyer regulation are lawyer conduct rules or principles, which sometimes are known as ethics rules, ethical principles, codes of conduct, rules of professional responsibility, or rules of professional conduct. In some cases, the primary regulatory authority, such as a US state supreme court, will adopt these codes of conduct. In other cases, several

levels of delegation may be involved before one reaches the entity responsible for adopting these types of provisions. For example, the 2007 U.K. Legal Services Act (U.K. Act) created the LSB, which has approved the SRA as a frontline regulator for solicitors in England and Wales. In 2011, the SRA adopted a new *Handbook for Solicitors*; this handbook includes a Code of Conduct that took effect on 6 October 2011 and that uses outcomes-focused or principles-based system of regulation.

A fourth instrument of lawyer regulation comes in the form of guidelines, guidance, comments, or commentary, which may or may not be binding. For example, the Law Society of Upper Canada has included 'commentary' in its Code of Conduct, and a number of US jurisdictions have adopted the comments to the ABA Model Rules of Professional Conduct and treat them as interpretative tools. Conversely, other US states have refused to officially adopt the Comments although they might reprint them (ABA CPR State Adoption of Comments, Law Society of Upper Canada Rules of Professional Conduct). Not all jurisdictions will use this regulatory instrument.

A fifth type of regulation can come in the form of regulations applicable to the entity that regulates the legal profession. For example, a law society or regulator might be subject to bylaws or other limitations, in addition to the legislation that created the entity. The Law Society of Upper Canada, for example, must comply with both its bylaws and the Ontario Legal Profession Act that created the Law Society (Terry et al., 2012a).

A sixth source of regulation is the jurisdiction's constitution, treaty, or foundational documents. Lawyer regulations (of any sort) that are inconsistent with the jurisdiction's foundational law may be struck down. This has happened in the United States and in the European Union (ABA Lawyer Advertising Cases; CCBE Judgments; Knake, 2010).

Of these six sources of regulation, the ethics code of a particular jurisdiction may be the most visible and most often the subject of discussion and comparison (ILEC Conferences, IBA Regulation webpage.) But when researching lawyer regulation, be sure to look for all of these different sources of regulation, even though not all instruments will be used in all jurisdictions.

As a final observation on the sources and methods of lawyer regulation, it is worth noting that it is common to hear lawyers around the world describe themselves as members of a self-regulated profession and to have academic professionals comment on this phenomenon (Rhode and Woolley, 2012; Zacharias, 2009; Ziv, 2009). Despite this rhetoric, there are few if any countries in which lawyers are completely self-regulated without any oversight, direction, or limits from other sources, such as the executive, legislative, or judicial branches of government. Nor is there necessarily any shared cross-cultural understanding of what is meant by the term 'self-regulation' (Terry et al., 2012a). Thus, when claims of lawyer 'self-regulation' are made, they should be examined to determine exactly what is meant.

Changes in Focus and Emphasis over Time

There has been a change in the focus and emphasis of lawyer regulation over time. Scholars also have called for increased attention to the organizational contexts in which lawyers practice, and the effects of organizational structure and culture on lawyers' individual conduct (Chambliss, 2000, 2006; Evans et al., 2008; Galanter and Palay, 1991; Galanter and Henderson, 2008; Nelson, 1988; Wilkins, 1998). Given that lawyers increasingly practice and are socialized within firms, both regulators and scholars have begun to focus on how such firms are managed and what firm-level regulatory controls, if any, should be required (Schneyer, 1991). The emergence of new organizational forms, such as multinational and multidisciplinary partnerships and alternative business structures (ABS), has only underscored the interest in developing a strategy for firm-level or 'entity' regulation (Chambliss, 2000; Chambliss and Wilkins, 2003).

Although regulation in many jurisdictions is still limited to the regulation of 'lawyers' (i.e., individuals), there has been a marked increase in the number of regulators that focus on regulation of the *entity* within which the lawyer works. For example, although entity disciplinary regulation generally is not available in the United States, other than through the liability system, New York and New Jersey have the authority to discipline law firms as well as individual lawyers. Several Canadian provinces authorize discipline of law firms, and other provinces are considering this concept. In several Australian states, if a lawyer chooses to practice in the form of an incorporated legal practice, that entity is subject to additional regulation. The most dramatic change is the adoption of an entity-based system of lawyer regulation by the U.K. SRA (Terry et al., 2012a).

In addition to the increased interest in entity regulation, there has been a change over time about the topics that seem to be the subject of lawyer regulation, particularly in those regulatory instruments that govern the middle time period and that often are referred to as 'legal ethics codes.' As exemplified by the ABA Canons of Professional Ethics and the CCBE Declaration of Perugia, it was common for older codes of ethics to be written as aspirational documents, rather than regulatory documents intended for enforcement (Schneyer, 2008; Terry, 1993). Many older ethics codes contained provisions that now are considered to involve matters of etiquette or that were self-protectionist or simply were platitudes. Although some of these provisions still remain, the current codes of conduct are much more likely to have provisions that are intended to be enforced as regulatory provisions. Most current sources of lawyer regulation will have rules governing lawyer competence, confidentiality, conflicts of interest, and perhaps additional issues, such as fees or advertising.

Although only a small fraction of all lawyers practice in large, global firms, transnational practice is one of the fastest growing sectors in the legal services market (Abel, 1994; Aronson, 2007; Cone, 1996; Dezalay and Garth, 1996; Flood, 1996; Silver et al., 2009; Terry, 2008b; Trubek et al., 1994; WTO, 1998, 2010). This rapid growth of transnational practice has put pressure on regulators to 'recognize' the qualifications of foreign lawyers and grant them practice rights (ABA, Section on Legal Education and Admissions to the Bar, Special Report, 2009). In general, however, local regulation has not kept pace with market demands (Flood, 2012). Although some jurisdictions have relaxed restrictions on foreign lawyers in recent years, many continue to impose

significant barriers to cross-border practice, and there continues to be significant jurisdictional variation (WTO, 1998, 2010). Lawyers have responded to the complexities of local regulation in part by noncompliance with the rules (Abel, 1994; Krishnan, 2010), and many scholars and regulators urge deregulation or more appropriate regulation (Goebel, 2000; Silver, 2003). In the coming years, this promises to be an important arena for regulatory reform.

Given the growth in transnational legal practice, there also has been a growing recognition of the need for lawyers from different jurisdictions to have a shared set of ethical rules when interacting with one another (Hellwig, 2009). The CCBE developed a Code of Conduct that applies to activities or individuals who cross EU borders (Terry, 1993). The IBA issued its General Principles of the Legal Profession in 2006. The International Law Association adopted in 2010 The Hague Principles on Ethical Standards for Counsel Appearing before International Courts and Tribunals in an effort to have lawyers from different cultures who appear before international courts and tribunals use the same rules. These represent an effort to identify principles in common (or to establish an agreed-upon approach in which it is impossible to identify principles in common). These cross-cultural conversations about lawyer regulation are likely to continue in the future (Terry, 2005).

The inclusion of services, including legal services, in the WTO's General Agreement on Trade in Services (GATS) has greatly expanded the universe of those interested in lawyer regulation. The GATS took effect in 1995 and was the first time that services, as opposed to goods, were included in a global trade agreement. The WTO has more than 155 member states, and the GATS meant that governments needed to decide what additional commitments, if any, they were willing to make with respect to their regulation of foreign lawyers. One consequence of the GATS is that it changed the way in which legal services are discussed. A standard terminology now is used to describe the methods by which legal services are delivered (i.e., GATS Modes 1–4) (Terry, 2010). There is also more standardization in the vocabulary used to describe the regulation of lawyers 'foreign' to one's jurisdiction (host jurisdiction, home jurisdiction; fly-in, fly-out (or FIFO); foreign legal consultants (FLCs); limited licenses; association rights; and skills transfer). This growing shared vocabulary has been useful in the efforts of entities, such as the WTO, APEC, and the IBA, to aggregate data about lawyer regulatory policies (Terry, 2010, 2013; WTO Report, 2010; APEC Legal Services Inventory, the forthcoming IBA Survey).

Another change in lawyer regulation is the increased interest by government antitrust (competition) authorities in the content of lawyer regulation. The US Supreme Court has struck down advertising bans and minimum fee schedules as violative of the antitrust laws. The European Court of Justice has considered antitrust challenges to fee schedules, and rules limiting the business forms in which lawyers may operate. The Organization for Economic Cooperation and Development, the European Commission, and the Australian, UK, and Canadian Competition authorities, among others, have issued reports about competition in the legal profession. These competition reports focus on topics that include lawyer regulatory provisions limiting advertising, setting minimum fees, evaluating whether entry into the profession is too restrictive, discussing the scope of the lawyers' monopoly or reserved activities, and setting rules limiting the business structures in which lawyers may operate (Terry, 2009). Some significant regulatory reforms have followed these reports.

Two recent changes in lawyer regulation may have a dramatic impact on lawyer regulation throughout the world. In 2007, the Australian law firm Slater and Gordon became the first law firm to be publicly traded on a stock exchange. In 2012, the U.K. SRA began issuing ABS licenses, which allow outside investment and ownership of law firms. The latter development is especially likely to be influential because of the number of law firms located in London and the number of countries in which London-based firms are located (Terry, 2013).

The 2007 UK changes were the result of a process that began with the U.K. Office of Fair Trading and culminated in the 2007 Legal Services Act, which dramatically reshaped lawyer regulation in the United Kingdom; one of the goals of the UK revisions was to increase access to justice by making legal services more accessible. One noteworthy aspect of the UK changes is the increased analysis of lawyer regulation from those who are 'outside' the traditional lawyer regulatory community (Black, 2010; Decker and Yarrow, 2010; U.K. Legal Services Board, Research). This phenomenon also has occurred outside the United Kingdom (Hadfield, 2008; Ribstein, 1998).

Globalization and technology, along with this increased competition focus, have created pressures on lawyer regulatory systems around the world (ABA Commission on Ethics 20/20; Gillers, 2012). Questions have been raised about the 'who-what-when-where-why-and-how' of lawyer regulation (Terry et al., 2012a). Should lawyer regulation reside primarily with the profession or should it be part of a coregulatory system, or should it be something else entirely (Hadfield, 2008; Rhode and Woolley, 2012)? Who or what should be the object of regulation: legal services or providers? Should regulators try to develop an ex ante system of lawyer regulation that encourages or requires lawyers to take proactive actions, such as continuing education or self-assessments as a way of preventing client mistakes? There is great interest in this issue in light of research showing a one-third reduction in client complaints after imposition of Australia's ex ante system of lawyer regulation (Parker et al., 2010). Where should regulation occur in an era in which law is practiced both transnationally and virtually, but regulation is tied to geographic political entities (Flood, 2011)? What should be the regulatory objectives of lawyer regulation (Terry et al., 2012b)? Finally, what methods should be used to achieve the desired regulatory goals and minimize unintended consequences (Regan et al., 2008; Rhode and Woolley, 2012)? Some believe that the next 20 years are likely to see greater transformation in how the world's legal professions are organized and ply their services than was true for any comparable period in history (Gillers, 2012).

Current and Emerging Directions in Theory and Research

By examining lawyer regulation resources and recurring academic conferences, one can identify current and emerging

schools of thought and directions in scholarly research. As noted earlier, the first ILEC took place in 2004 in England. The most recent ILEC conference was the ILEC V, held in 2012 in Banff, Canada. Banff presenters were asked to identify which of four 'streams' their presentation was suited for: Culture, Ethics, and Society; Empirical Approaches to Legal Ethics; Philosophy and Legal Ethics; or Regulation of the Profession(s)/Ethics and Education (see http://www.ucalgary.ca/ilec5/ilec_streams).

In the United States, those interested in lawyer regulation often must choose between two different conferences that are held annually at similar times. The ABA Center for Professional Responsibility hosts an annual National Conference on Professional Responsibility, which usually is attended by US and Canadian academics, regulators, and practicing lawyers and tends to focus on North American doctrinal developments and risk management. The Law and Society Association, an international scholarly association based in the United States, also holds annual conferences, often at a similar time. For a number of years, there have been coordinated panels on legal services and lawyers at the annual Law and Society Association Conference. The 2012 conference had panel sessions on a wide range of topics related to lawyers and judges. Although others have held periodic conferences devoted to various lawyer regulation topics, no other jurisdiction or entity seems to have annual conferences devoted to lawyer regulation.

Some of the leading specialty journals include the *Georgetown Journal of Legal Ethics* and the *Journal of the Professional Lawyer* in the United States and the peer-reviewed UK journals entitled *Legal Ethics* and the *International Journal of the Legal Profession*. A number of US institutions have centers that focus on the legal profession and those institutions have hosted conferences that focus on the legal profession (e.g., Harvard, Georgetown, Stanford, Fordham, Hofstra, Michigan State, South Texas, and Akron). As noted earlier, there have now been five International Conferences on Legal Ethics. The presentations and papers from these conferences usually are available online and sometimes are published in Symposium issues (Symposium, 1996). Some academic institutions have published papers from practitioner-oriented conferences on lawyer regulation (Terry, 2008c). These resources are among the most efficient way to locate leading academics and current and emerging directions in theory and research.

Methodological Problems

One of the methodological problems facing researchers and stakeholders is the difficulty in locating and sharing information about lawyer regulation in other countries. A second methodological problem is the difficulty one has in making cross-cultural or cross-professional comparisons given the definitional issues noted earlier. A third methodological problem is access to the data necessary to conduct empirical work (Silver, 2010;Chambliss, 2012). There have been some excellent qualitative and quantitative studies and some useful collaborations among scholars and regulators (American Bar Foundation After the J.D. Publications; Fortney, 1996, 2009; Parker and Aitken, 2011; Parker et al., 2010; Parker et al., 2009; Parker and Haller, 2010; Regan, 2004; Sommerlad et al., 2010). Nevertheless, many academics lack access to data that would allow them to engage in critical empirical analysis.

Links to Databases and Software

In addition to the domestic sources and authors, several resources aggregate useful information about the regulation of lawyers. The number of these resources seems to be growing.

With respect to US lawyer regulation, information about US admissions regulators is found on the webpage of the NCBE and the ABA Section of Legal Education Admissions to the Bar. Extensive information about US state ethics rules, including the degree to which a state has varied from the ABA Model Rules of Professional Conduct, has been aggregated on the Policy Implementation webpage of the ABA Center for Professional Responsibility. Information about US lawyer discipline is found on the webpages of the NOBC and the ABA Center for Professional Responsibility Standing Committee on Professional Discipline. The American Law Institute has issued a Restatement of the Law Governing Lawyers, which is exceptionally helpful to those studying the regulation of lawyers.

The FLSC aggregates regulatory information for Canada. Its website includes information about model FLSC policies, and includes links to provincial and territorial law societies and their regulatory provisions.

The CCBE maintains an active website with extensive material about lawyer regulation in Europe. When examining an issue or before preparing a position paper, the CCBE often will survey its members about lawyer regulation in their jurisdiction with respect to a particular issue. Most of these CCBE surveys and position papers may be accessed through the relevant link on the CCBE Committees webpage. The CCBE Discipline Committee webpage has links to a list of disciplinary contact points in CCBE member states and a summary of each jurisdiction's disciplinary process and to another page that summarizes European Court of Justice cases related to lawyer regulation. The CCBE Deontology Committee webpage includes links to national codes of conduct and to reports summarizing EU privilege and confidentiality rules and to the 2004 Fish Report on EU Member State Treatment of In-house Counsel.

The Law Council of Australia describes itself as the "peak national representative body of the Australian legal profession." In addition to the individual legal practitioners it represents, the Law Council has 'constituent members' that include the state and territorial bar associations and law societies. Its website collects information on lawyer regulatory policies. Additional information about lawyer regulatory policies is available on the CORO website, which is an umbrella organization of lawyer regulators in Australia and New Zealand.

It is increasingly common to have international or regional entities collect lawyer regulation information from different countries. The WTO currently consists of more than 155 members with broad geographic representation. The WTO website allows one to go online to determine whether a particular WTO member made legal services 'commitments.' If it has,

then that country may have lawyer regulation that is more liberal than its commitments, but it may not have regulation of foreign providers that is less liberal than its legal services commitments (Terry, 2010). These WTO Schedules can provide useful starting data about lawyer regulation (WTO Services Database).

The Legal Services Initiative of the APEC collected lawyer regulatory information from many of the APEC member economies and posted it online. Its webpage includes both a high-level overview and detailed jurisdictional information, including contact information for lawyer regulators.

The Bar Issues Commission of the IBA has begun posting links to the legal ethics codes and other rules on its webpage entitled 'IBA, Documents on the Regulation of the Legal Profession.' (The IBA has more than 200 bar association members and wide geographic representation.) The IBA also has published a useful book on privilege and confidentiality (Greenwald and Russenberger, 2012). Its Anti-Money Laundering Forum summarizes money-laundering legislation applicable to lawyers International Bar Association. Anti-Money Laundering Forum. The IBA's 2013 release of the results of a report it commissioned on lawyer regulation should be available online.

Several additional resources are worth noting. The U.K. LSB, which was created as part of the 2007 U.K. Legal Services Act, has commissioned a number of research reports that sometimes include useful comparative information on lawyer regulation. The consultations issued by the SRA also have useful information U.K. Solicitors Regulation Authority, Consultations. In September 2012, the SRA sponsored the first International Conference of Legal Regulators; the materials from this conference are posted online and include a Directory of information about the regulatory system provided by that country's regulators. The second meeting of this group will take place in San Francisco, California, in August 2013. In other fields, such as antitrust, securities, and labor, international networks or regulators have served as a useful source of information (Terry, 2012). The ABA Guide to International Bar Admissions (Dombrow and Matos, 2012) includes useful information on lawyer regulation. Finally, the IBA has commissioned a survey of lawyer regulation. If this IBA survey is completed and posted online, it will be a useful resource regarding lawyer regulation.

See also: Ethical Codes, Professional: Business Codes; In-House Counsel; Legal Education; Legal Process and Social Science: United States; Professions and Professionalization, History of; Regulatory Agencies; Social Science Professions and Professionalization.

Bibliography

Abbott, A., 1988. The System of Professions: An Essay on the Expert Division of Labor. University of Chicago Press, Chicago.

Abel, R.L., 1988. Lawyers in the civil law world. In: Abel, R.L., Lewis, P.S.C. (Eds.), Lawyers in Society: The Civil Law World. University of California Press, California.

Abel, R.L., 1994. Transnational law practice. Case Western Reserve Law Review 44, 737–825.

Abel, R.L., 2008. Lawyers in the Dock: Learning from Attorney Disciplinary Proceedings. Oxford University Press, New York.

Abel, R.L., Lewis, P.S.C. (Eds.), 1988–1989. Lawyers in Society, vol. 1. The Common Law World, vol. 2. The Civil Law World, vol. 3. Comparative Theories. University of California Press, California.

American Bar Association, Cases on Lawyer Advertising. Available at: http://www.americanbar.org/groups/professional_responsibility/resources/professionalism/professionalism_ethics_in_lawyer_advertising/clearinghouse.html.

American Bar Association Center for Professional Responsibility, Commission on Ethics 20/20. Available at: http://www.americanbar.org/groups/professional_responsibility/aba_commission_on_ethics_20_20.html.

American Bar Association Center for Professional Responsibility, Policy Implementation. Available at: http://www.americanbar.org/groups/professional_responsibility/policy.html.

American Bar Association CPR Policy Implementation Committee, State Adoption of the Aba Model Rules of Professional Conduct and Comments (May 23, 2011). Available at: http://www.americanbar.org/content/dam/aba/migrated/cpr/pic/comments.authcheckdam.pdf.

American Bar Association Center for Professional Responsibility. Standing Committee on Professional Discipline. Available at: http://www.americanbar.org/groups/professional_responsibility/committees_commissions/disciplinecommittee.html.

American Bar Association Center for Professional Responsibility, Task Force on the Model Definition of the Practice of Law. Available at: http://www.americanbar.org/groups/professional_responsibility/task_force_model_definition_practice_law.html.

American Bar Association. Section on Legal Education and Admissions to the Bar. Available at: http://www.americanbar.org/groups/legal_education.html (includes links to admissions-related information).

American Bar Association, Section on Legal Education and Admissions to the Bar, 2009. Report of the Special Committee on International Issues. Available at: http://www.americanbar.org/content/dam/aba/administrative/legal_education_and_admissions_to_the_bar/council_reports_and_resolutions/20090715_international_issues_report.authcheckdam.pdf.

American Bar Foundation. After the J.D. and Related Publications. Available at: http://www.americanbarfoundation.org/publications/AftertheJD/AJD_Publications.html.

American Law Institute, 2000. Restatement of the Law: The Law Governing Lawyers: As Adopted and Promulgated by the American Law Institute at Washington, DC, May 12, 1998. American Law Institute Publishers.

Andrasi, G., 2013. Teaching Legal Ethics: What Can We [in Hungary] Learn from UK Standards and Practices?. SSRN Abstract 2203356.

Aronson, B.E., 2007. Elite firm mergers and reputational competition: is bigger really better? An international comparison. Vanderbilt Journal of Transnational Law 40, 763–831.

Asia Pacific Economic Cooperation (APEC). The Legal Services Initiative. Available at: http://www.legalservices.apec.org/index.html.

Association of American Law Schools (AALS) Section of Professional Responsibility, 2009. A Survey of Professional Responsibility Courses at American Law Schools in 2009. Available at: http://www.legalethicsforum.com/files/pr-survey-results-final.pdf.

Black, J., 2010. Risk based regulation: a short introduction for the Legal Services Board. In: Roy, A. (Ed.), The Future of Legal Services: Emergent Thinking. Available at: http://www.legalservicesboard.org.uk/news_publications/press_releases/2010/pdf/14_june_conference_papers_master.pdf.

Boon, A., Levin, J., 2008. The Ethics and Conduct of Lawyers in England and Wales. Hart Publishing, Oxford.

Canadian Association for Legal Ethics. Available at: http://ethicsincanada.com/.

Chambliss, E., 2000. Professional responsibility: lawyers, a case study. Fordham Law Review 69, 817–857.

Chambliss, E., 2006. The professionalization of law firm in-house counsel. North Carolina Law Review 84, 1515–1575.

Chambliss, E., 2012. Whose ethics? The benchmark problem in legal ethics research. In: Levin, L., Mather, L. (Eds.), Lawyers in Practice: Ethical Decision Making in Context. University of Chicago Press.

Chambliss, E., Wilkins, D.B., 2002. The emerging role of ethics advisors, general counsel, and other compliance specialists in large law firms. Arizona Law Review 44, 559–592.

Chambliss, E., Wilkins, D.B., 2003. A new framework for law firm discipline. Georgetown Journal of Legal Ethics 16, 335–351.

Clark, K., 2000. The legacy of watergate for legal ethics instruction. Hastings Law Journal 51, 673–682.

Cone, S.M., 1996. International Trade in Legal Services: Regulation of Lawyers and Forms in Global Practice. Little Brown, Boston.

Cotter, B., Devlin, R., Law, J., Woolley, A., 2012. Lawyers' Ethics and Professional Regulation, second ed. LexisNexis, Toronto.

Conference of Regulatory Officers (CORO) (of Australia and New Zealand). Available at: http://www.coro.com.au/.

Council of Bars and Law Societies (CCBE). CCBE Committees and Working Groups. Available at: http://www.ccbe.org/index.php?id=25&L=0.

Council of Bars and Law Societies (CCBE). Deontology Committee (includes links to National Codes of Conduct (of CCBE Members) and to useful reports on privilege and the status of in-house counsel). Available at: http://www.ccbe.org/index.php?id=94&id_comite=7&L=0.

Council of Bars and Law Societies (CCBE). Discipline Working Group. Available at: http://www.ccbe.org/index.php?id=94&id_comite=24&L=0.

Council of Bars and Law Societies (CCBE). Judgments of the Court of Justice of the European Communities, Judgments and Orders of the Court of First Instance Concerning the Legal Profession (12.04.2011). Available at: http://www.ccbe.org/fileadmin/user_upload/NTCdocument/Court_of_Justice_of_1_1302595800.pdf.

Council of Bars and Law Societies (CCBE). Summary of Disciplinary Proceedings and Contact Points in the EU and EEA Member States (Last Update March 2011). Available at: http://www.ccbe.org/fileadmin/user_upload/NTCdocument/Table_discipline__Ma1_1335781934.pdf.

Council of Bars and Law Societies (CCBE). Training Committee. Available at: http://www.ccbe.org/index.php?id=94&id_comite=13&L=0.

Dare, T., Wendel, W.B. (Eds.), 2010. Professional Ethics and Personal Integrity. Cambridge Scholars Press, Newcastle upon Tyne.

Davis, A.E., 1996. Professional liability insurers as regulators of law practice. Fordham Law Review 65, 209–232.

Decker, C., Yarrow, G., 2010. Understanding the Economic Rationale for Legal Services Regulation: A Report for the Legal Services Board. Available at: http://www.legalservicesboard.org.uk/news_publications/latest_news/pdf/economic_rationale_for_Legal_Services_Regulation_Final.pdf.

Dezalay, Y., Garth, B., 1996. Dealing in Virtue: International Arbitration and the Construction of a Transnational Legal Order. University of Chicago Press, Chicago.

Dodek, A.M., 2008. Canadian legal ethics: ready for the twenty-first century at last. Osgoode Hall Law Journal 46, 1–49.

Dombrow, R.W., Matos, N.A. (Eds.), 2012. The ABA Guide to International Bar Admissions. American Bar Association Publishing, Chicago.

Evans, A.H., Haller, L.R., Le Mire, S., Mortensen, R.G., Parker, C.E., 2008. The ethical infrastructure of legal practice in large law firms: values, policy and behavior. The University of New South Wales Law Journal 31 (1), 158–188.

Faulconbridge, J., Muzio, D., Cook, A., 2012. Learning to be a lawyer in transnational law firms: communities of practice, institutions and identity regulation. Global Networks 12 (1), 48–70.

Federation of Law Societies of Canada. National Regulatory Initiatives (and items linked from this page). Available at: http://www.flsc.ca/en/national-standards/.

Fish, J., 2004. Regulated Legal Professionals and Professional Privilege within the European Union, the European Economic Area and Switzerland, and Certain Other European Jurisdictions: A Report by John Fish, Former President of the CCBE and Solicitor, Dublin. Available at: http://www.ccbe.org/fileadmin/user_upload/NTCdocument/fish_report_enpdf1_1184145269.pdf.

Fleming, D., Shackel, R. (Eds.), forthcoming. Australian Legal Profession. Thomson Reuters.

Flood, J., 1996. Megalawyering in the global order: the cultural, social and economic transformation of global legal practice. International Journal of the Legal Profession 3 (2/3), 169–214.

Flood, J., 2011. The re-landscaping of the legal profession: large law firms and professional re-regulation. Current Sociology 59 (4), 507–529.

Flood, J., 2012. Transnational lawyering: clients, ethics and regulation. In: Levin, L., Mather, L. (Eds.), Lawyers in Practice: Ethical Decision Making in Context. University of Chicago Press, Chicago.

Fortney, S., 1996. Are law firm Partners Islands unto themselves? An empirical study of law firm peer review and culture. Georgetown Journal on Legal Ethics 10, 271–316.

Fortney, S., 2009. Tales of two regimes for regulating limited liability law firms in the U.S. and Australia: client protection and risk management lessons. Legal Ethics 11, 230–240.

Freedman, M.H., 1990. Understanding Lawyers' Ethics. Matthew Bender, New York.

Galanter, M., Henderson, W.D., 2008. The elastic tournament: the second transformation of the big law firm. Stanford Law Review 60, 1867–1929.

Galanter, M., Palay, T., 1991. Tournament of Lawyers: The Transformation of the Big Law Firm. University of Chicago Press, Chicago.

Gillers, S., 2012. A profession, if you can keep it: how information technology and fading borders are reshaping the law marketplace and what we should do about it. Hastings Law Journal 63, 953–1022.

Goebel, R.J., 2000. The liberalization of interstate legal practice in the European Union: lessons for the United States? International Lawyer 34, 307.

Gordon, R.W., 1988. The independence of lawyers. Boston University Law Review 68, 1–83.

Green, B.A., Zacharias, F.C., 2009. Rationalizing judicial regulation of lawyers. Ohio State Law Journal 70, 73–139.

Greenwald, D., Russenberger, M., 2012. Privilege and Confidentiality: An International Handbook, second ed. International Bar Association Publications, Haywards Heath.

Griffiths-Baker, J., 2002. Serving Two Masters: Conflicts of Interest in the Modern Law Firm. Hart Publishing, Oxford.

Hadfield, G.K., 2008. Legal barriers to innovation: the growing economic cost of professional control over corporate legal markets. Stanford Law Review 60, 1689–1732.

Heinz, J., Laumann, E., 1982. Chicago Lawyers: The Social Structure of the Bar. Russell Sage Foundation, New York.

Hellwig, H.J., 2009. At the intersection of legal ethics and Globalization: International conflicts of law in lawyer regulation. Penn State International Law Review 27, 395–401.

Henssler, M., Prütting, H., 2010. Kommentar zur Bundesrechtsanwaltsordnung. C.H. Beck-Verlag, München, 3. Aufl. 2010, (zugleich als Autor Kommentierung der §§ 4–17, 43a, 46, 48–49a, 59c – m BRAO sowie von Vorschriften der BORA und des PartGG).

International Association of Legal Ethics. Available at: http://www.stanford.edu/group/lawlibrary/cgi-bin/iaole/wordpress/.

International Bar Association, Documents on the Regulation of the Legal Profession. Available at: http://www.ibanet.org/documents_on_the_regulation_of_the_legal_profession.aspx.

International Bar Association. Anti-Money Laundering Forum. Available at: http://www.anti-moneylaundering.org/globalchart.aspx.

International Conference of Legal Regulators. Available at: http://www.international-conference-of-legal-regulators.org/the-london-conference/.

International Legal Ethics Conference (ILEC) Programs (2004–2012). Available at: http://www.stanford.edu/group/lawlibrary/cgi-bin/iaole/wordpress/wp-content/uploads/2011/11/ILEC_1_Programme.pdf.

Knake, R., 2010. The supreme court's increased attention to the law of lawyering: mere coincidence or something more? American University Law Review 59, 1499–1574.

Krishnan, J., 2010. Globetrotting law firms. Georgetown Journal of Legal Ethics 23, 57–99.

Law Council of Australia. Available at: http://www.lawcouncil.asn.au/.

Law Society of Upper Canada, Rules of professional conduct (Adopted by Convocation June 22, 2000, Amendments Current to January 24, 2013). Available at: http://www.lsuc.on.ca/with.aspx?id=671.

Leubsdorf, J., 2001. Man in His Original Dignity: Legal Ethics in France. Ashgate, Burlington, VT.

Leubsdorf, J., 2009. Legal ethics falls apart. Buffalo Law Review 57, 959–1055.

Liu, S., 2008. Globalization as boundary-blurring: international and local law firms in China's corporate law market. Law & Society Review 42 (4), 771–804.

Luban, D., 1988. Lawyers and Justice: An Ethical Study. Princeton University Press, Princeton, NJ.

Michelson, E., Liu, S., 2010. What do Chinese lawyers want? political values and legal practice. In: Li, Cheng (Ed.), China's Emerging Middle Class: Beyond Economic Transformation. Brookings Institution Press, Washington, DC, pp. 310–333.

Moliterno, J.E., 1991. An analysis of ethics teaching in law schools: replacing lost benefits of the apprentice system in the academic atmosphere. University of Cincinnati Law Review 60, 83–134.

Moorhead, R., Sherr, A., Paterson, A., 2003. What clients know: client perspectives and legal competence. International Journal of the Legal Profession 10 (1), 5–36.

Morgan, T., 2010. The Vanishing American Lawyer. Oxford University Press, New York.

Mortensen, R., Tranter, K., Bartlett, F., Corbin, L., Robertson, M., 2010. Reaffirming Legal Ethics: Taking Stock and New Ideas. Routledge, Oxford.

Myerson, T.S., 1995. The Japanese system. In: Daly, M.C., Goebel, R.J. (Eds.), Rights, Liability and Ethics in International Legal Practice. Transnational Juris Publications, Irvington-on-Hudson, New York.

National Organization of Bar Counsel. Available at: http://www.nobc.org/ (research drop down tab includes links to U.S. state disciplinary authorities and rules, ethics rules, and other items).

National Conference of Bar Examiners (NCBE). Available at: http://www.ncbex.org/ (includes directory of U.S. state bar admissions offices and related materials).

National Conference of Bar Examiners (NCBE). Comprehensive Guide to Bar Admissions 2013. Available at: http://www.ncbex.org/assets/media_files/Comp-Guide/CompGuide.pdf.

Nelson, R.N., 1988. Partners with Power: Social Transformation of the Large Law Firm. University of California Press, Berkeley, CA.

Nicolson, D., Webb, J., 2000. Professional Legal Ethics: Critical Interrogations. Oxford University Press, New York.

Parker, C.E., Evans, A.H., 2007. Inside Lawyers' Ethics. Cambridge University Press, Cambridge.
Parker, C., Rosen, R., Nielsen, V., 2009. The two faces of lawyers: professional ethics and business compliance with regulation. Georgetown Journal of Legal Ethics 22, 201–248.
Parker, C.E., Gordon, T., Mark, S., 2010. Regulating law firms ethics management: an empirical assessment of an innovation in regulation of the legal profession in New South Wales. Journal of Law and Society 37, 466–500.
Parker, C.E., Haller, L., 2010. Inside running: internal complaints management practice and regulation in the legal profession. Monash University Law Review 36, 217–249.
Parker, C.E., Aitken, L., 2011. The Queensland "workplace culture check": learning from reflection on ethics inside law firms. Georgetown Journal of Legal Ethics 24, 399–441.
Paterson, I., Fink, M., Ogus, A., et al., 2003. Research report: economic impact of regulation in the field of liberal professions. In: Different Member States: Regulation of Professional Services: Regulation of Professional Services: Study for the European Commission, DG Competition: Final Report—Part 1—Main Report, 1–128, at 47, Table 3–7 Legal Services (Lawyers): Scope of Activities. Available at: http://ec.europa.eu/competition/sectors/professional_services/studies/prof_services_ihs_part_1.pdf.
Pearce, R.G., Wald, E., 2012. Rethinking lawyer regulation: how a Relational approach would Improve professional rules and Roles. Michigan State Law Review 2012, 513.
Regan Jr, M.C., 2004. Eat what You Kill: the Fall of a Wall Street Lawyer. University of Michigan Press, Ann Arbor, Mich.
Regan Jr., M.C., MacEwen, B., Ribstein, L.E., 2008. Law firms, ethics, and Equity Capital. Georgetown Journal of Legal Ethics 21, 61–94.
Ribstein, L.E., 1998. Ethical rules, agency costs, and law firm structure. Virginia Law Review 84, 1707–1759.
Rhode, D.L., 1981. Policing the professional monopoly: a constitutional and empirical analysis of unauthorized practice Prohibitions. Stanford Law Review 34, 1–112.
Rhode, D.L., Woolley, A., 2012. Comparative perspectives on lawyer regulation: an agenda for reform in the United States and Canada. Fordham Law Review 80, 2761–2790.
Rueschemeyer, D., 1986. Comparing legal professions cross-nationally: from a professions-centered to a state-centered approach. American Bar Foundation Research Journal, 415–446.
Schneyer, T., 1991. Professional discipline for law firms? Cornell Law Review 77, 1–46.
Schneyer, T., 2008. How things have changed: contrasting the regulatory environments of the canons and the model rules. Journal of the Professional Lawyer 2008, 161–187.
Silver, C., 2003. Regulatory mismatch in the international market for legal services. Northwestern Journal of International Law and Business 23 (3), 487–550.
Silver, C., 2010. What we don't know can hurt us: the need for empirical research in regulating lawyers and legal services in the global economy. Akron Law Review 43, 1009–1079.
Silver, C., De Bruin Phelan, N., Rabinowitz, M., 2009. Between diffusion and distinctiveness in globalization: U.S. law firms go glocal. Georgetown Journal of Legal Ethics 22, 1431–1471.
Sommerlad, H., Webley, L., Muzio, D., Tomlinson, J., Duff, L., 2010. Diversity in the Legal Profession in England and Wales: A Qualitative Study of Barriers and Individual Choices. Legal Services Board, London. Available at: http://www.legalservicesboard.org.uk/what_we_do/Research/Publications/pdf/lsb_diversity_in_the_legal_profession_final_rev.pdf.
Susskind, R., 2008. The End of Lawyers? Rethinking the Nature of Legal Services. Oxford University Press, New York.
Symposium, 1996. Institutional choices in the regulation of lawyers. Fordham Law Review 65, 33–463.
Terry, L.S., 1993. An introduction to the European community's legal ethics code part I: an analysis of the CCBE code of conduct. Georgetown Journal of Legal Ethics 7, 1–87.
Terry, L.S., 2005. U.S. legal ethics: the coming of age of global and comparative perspectives. Washington University Global Studies Law Review 4, 463–553.
Terry, L.S., 2008a. The future regulation of the legal profession: the impact of treating the legal profession as 'service providers'. Journal of the Professional Lawyer 2008, 189–211.
Terry, L.S., 2008b. The legal world is flat: globalization and its effects on lawyers practicing in non-global law firms. Northwestern Journal of International Law and Business 28 (3), 527–559.
Terry, L.S., 2008c. Foreword: 2008 global legal practice symposium. Penn State International Law Review 27, 269–289.
Terry, L.S., 2009. The European Commission Project regarding competition in professional services. Northwestern Journal of International Law and Business 29, 1–117.
Terry, L.S., 2010. From GATS to APEC: the impact of trade agreements on legal services. Akron Law Review 43, 875–984.
Terry, L.S., 2011. Introductory note to the court of justice of the European Union: the Axzo nobel EU attorney-client privilege case. International Legal Materials 50, 1–4.
Terry, L.S., 2012. Infrastructure, the rule of law, and the need for a global umbrella organization for lawyer regulators. Michigan State Law Review 2012, 735–768.
Terry, L.S., 2013. Global and Canadian trends in lawyer regulation. University of Saskatchewan Law Review 76, 145–184.
Terry, L.S., Mark, S., Gordon, T.A., 2012a. Trends and challenges in lawyer regulation: the impact of globalization and technology. Fordham Law Review 80, 2661–2684.
Terry, L.S., Mark, S., Gordon, T.A., 2012b. Adopting regulatory objectives for the legal profession. Fordham Law Review 80, 2685–2760.
Trubek, D., Dezalay, Y., Buchanan, R., Davis, J., 1994. Global restructuring and the law: studies of the internationalization of legal fields and the creation of transnational arenas. Case Western Reserve Law Review 44 (2), 407–498.
University of Ottawa, Juriglobe. Available at: http://www.juriglobe.ca/eng/index.php.
U.K., 2007. Legal Services U.K. Act. Available at: http://www.legislation.gov.uk/ukpga/2007/29/contents.
U.K. Legal Services Board, Research. Available at: https://research.legalservicesboard.org.uk/.
U.K. Solicitors Regulation Authority, Consultations. Available at: http://www.sra.org.uk/consultations/.
U.K. Solicitors Regulation Authority, Outcomes Focused Regulation. Available at: http://www.sra.org.uk/solicitors/freedom-in-practice/outcomes-focused-regulation.page.
Whelan, C.J., Ziv, N., 2012. Privatizing professionalism: client control of lawyers' ethics. Fordham Law Review 80, 2577–2613.
Wilkins, D.B., 1992. Who should regulate lawyers? Harvard Law Review 105, 801–887.
Wilkins, D.B., 1998. Everyday practice is the troubling case: confronting context in legal ethics. In: Sarat, A., Constable, M., Engel, D., Hans, V., Lawrence, S. (Eds.), Everyday Practice and Trouble Cases. Northwestern University Press, Evanston, IL.
Woolley, A., 2011. Understanding Lawyers' Ethics in Canada. LexisNexis, Toronto.
WTO Council for Trade in Services, 2010. Legal Services – Background Note by the Secretariat. S/C/W/318.
WTO Council for Trade in Services, 1998. Legal Services, Background Note by the Secretariat. S/C/W/43.
WTO Services Database Output. Predefined Reports on Uruguay Round Legal Services Commitments. These are no longer available on the WTO webpage, but copies are available at: http://www.americanbar.org/groups/professional_responsibility/policy/gats_international_agreements/uruguay.html.
Zacharias, F.C., 2009. The myth of self-regulation. Minnesota Law Review 93, 1147–1190.
Ziv, N., 2009. The lawyer's role in a contemporary democracy, tensions between various conceptions of the lawyer's role, regulation of israeli lawyers: from professional autonomy to multi-institutional regulation. Fordham Law Review 77, 1763–1794.

Lawyers: Social Organization of the Profession

Fiona M Kay, Queen's University, Kingston, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Abel, volume 12, pp. 8553–8559, © 2001, Elsevier Ltd.

Abstract

The social organization of the legal profession varies from country to country and across systems of law. Although the majority of lawyers work in private practice, many lawyers work in other settings such as government, corporations, education, and law clinics. In recent years, dramatic changes have taken place, including a tremendous growth in the size of the legal profession and increasing demographic diversity. The nature of legal work has also changed, with a smaller proportion of lawyers working independently and more lawyers employed in law firms and other organizations. A rise in corporate legal work has led to the growth of large law firms with branch and international offices. These changes present new challenges for professional ethics, protection of jurisdictions, and self-regulation of the legal profession.

The social organization of the legal profession assumes distinct forms across various systems of law, across nations, and even within nations. From jurisdiction to jurisdiction, the profession is organized differently according to titles and roles, education and training requirements, and the settings in which legal professionals work. Yet across jurisdictions, members of the legal profession, as the experts of law, play a central role in facilitating the aims of law in terms of social integration, regulation of behavior, and social change.

The legal profession shares several core elements with other professions (such as medicine, accounting, and engineering). The first is that lawyers are holders of expert knowledge. This body of knowledge is composed partly of formal, abstract principles (Abbott, 1988). Their training equips lawyers to interpret the law and to facilitate various transactions, from business contracts to intellectual property rights, and from criminal prosecutions to estates planning. The control over a body of knowledge leads legal professionals to enjoy a comparatively high level of technical autonomy in their work. This is reflected in the relative independence of lawyers' work objectives from larger organizational imperatives. This autonomy is also seen in systems of self-regulation, governance and licensing, and in admission requirements of the legal profession. Another hallmark of professions is a normative orientation toward the service of others. This service orientation is often instilled in law schools, where students volunteer hours at legal clinics and then continue in their professional life efforts to ensure access to justice, including helping those in need through pro bono (e.g., free and voluntary) legal services. A final characteristic of law as a profession concerns status: lawyers are typically of high status in society and are well remunerated (Goode, 1961).

In recent years, dramatic changes have taken place in the world of law practice, changes that challenge this traditional model of the profession. The legal profession has grown exponentially in size over the last 30 years. There has also been a tremendous rise in very large law firms handling corporate legal work. The nature of that work is increasingly transnational, with many big firms operating branch and overseas offices. Within firms, work is increasingly structured by departments and subdivided into specializations of fields of law. Moreover, one has seen alternative career tracks emerge, supplementing or supplanting the 'up or out' conventional promotion model of law firms. These changes have impacted significantly the social organization of the legal profession, leading to greater fragmentation and intensification of hierarchy within the profession. As a result, the hallmark characteristics of law as a profession are also undergoing a transformation. This article reviews research on the changing social organization of the legal profession with respect to work roles, education and training, sectors within the profession, career paths, demographic change, and contemporary challenges facing the legal profession.

Systems of Law, Professional Roles, and Titles

Differences in the legal profession, specifically with reference to education and core knowledge, are largely based in a country's system of law. Civil law systems are characterized by a vast network of codified laws; one example is France, where Napoleon introduced a national code in 1804 (other civil law-based countries include Spain, Italy, Germany, Brazil, Turkey, Mexico, Chili, and Venezuela) (Deflem, 2008; Pérez-Perdomo, 2006). Common law countries (e.g., Ireland, the United States, the United Kingdom, South Africa, and Australia) typically draw on the British legal system and rely heavily on case law, based on precedent established through earlier court decisions. In some countries, hybrid models of law exist. In Canada, for example, a history of colonization by England and France resulted in civil and common law systems infusing different regions of the country. Even in countries where a single system of law exists, there is increasingly an infusion of the other system; jurisprudence is becoming more relevant in civil law countries, and common law countries are experiencing increasing levels of codification. Nevertheless, these two systems of law persist in terms of legal education: civil law tradition law schools encourage the study of codes and statutes, while common law tradition law schools emphasize analysis of court cases (Deflem, 2008). Other combinations also exist, such as countries with religious laws (e.g., Saudi Arabia, Sudan, and Yemen based on Islamic law)

and pluralist models (e.g., Egypt, Syria, and Morocco with Islamic law and civil law, and Nigeria and Pakistan with English common law and some Islamic law applied).

The titles and roles of legal professionals also appear differently in common and civil law nations. In Great Britain, a distinction exists between solicitors and barristers. Solicitors are legal advisors to clients, while barristers, the smaller and more elite professional group, are trial lawyers in court. Some common law countries, notably the United States and Canada, do not make a distinction between solicitor and barrister. In these countries, lawyers operate as barristers and solicitors, although they may classify their work as primarily that of a solicitor or as a barrister or litigator. Notaries exist as a separate occupation, although in restricted numbers, without the requirement of a law degree, and their responsibilities are largely limited to administering oaths and attesting documents (Brockman, 1999).

There are further variations in the words used to describe lawyers. For example, the term 'attorney' is more commonly used in the United States than in Canada to refer to a lawyer or litigator. In civil law countries, variations in titles and roles take a different form. In contemporary France, *avocet* encompasses *avoué* (pleading cases in court), *conseil juridique* (legal advisor in business), and solicitor work (e.g., administrative law, wills and estates, and real estate). Meanwhile, in the province of Quebec, Canada, following the French civil law tradition, the *avocat(e)* and *notaire* represent two distinct traditions of law practice. Both professional groups attend law school (*notaires* for an additional year to earn a diploma in notarial law). *Notaires* may express opinions in all areas of law. However, unlike *avocats*, *notaires* are public officials, required to exercise neutrality and provide advice to all parties. The system of exclusive jurisdictions in Quebec specifies that *notaries* traditionally practice in the areas of real estate, wills and estates, marriages and family mediation (Kay, 2009); litigation (criminal law, civil litigation, divorce, and so on) and advocacy are the realm of *avocats*, and only *avocats* can be appointed judges (Kay and Brockman, 2000).

Education, Licensing, and Practice Settings

Educational training of legal professionals also varies across countries. Historically, legal professionals trained through a system of apprenticeship in law offices. Over the last century, university law programs became the primary platform for educating students of law. In many countries, law can be studied as an academic discipline in college or university, following secondary schooling (e.g., Spain, Italy, and Portugal), while in other countries (e.g., the United States and Canada) a college degree is normally required prior to entrance to professional law schools offering a JD (Juris Doctor) or LLB (Bachelor of Laws) degree. Many countries require an apprenticeship period after law school. In civil law regions such as France and Quebec, Canada, this training period includes practical professional training at an *École Barreau* and supervision in a law practice setting. Some common law countries (e.g., Canada, the United Kingdom, and Ireland) require a similar period that lasts several months, followed by a bar admissions examination (Kay, 2009). After admission to the bar (or 'swearing in' ceremonies, as is the case with English solicitors and French *notaires*), law graduates embark on careers across a broad array of work settings.

The majority of law graduates enter private practice, working in law firms with other lawyers or on their own as solo practitioners. In the United States, 75% of lawyers work in private practice. About half of these lawyers work as solo practitioners and half in firms, with the majority of lawyers employed in smaller firm sizes (20% of lawyers work in firms of 10 or fewer lawyers, 18% in firms between 11 and 50 lawyers, and 20% in firms larger than 50 lawyers) (ABA, 2012). Although research focuses on large law firms, the reality is most law firms in the United States consist of just 2–5 lawyers (76%), followed by firms of 6–10 lawyers (13%). Only 11% of law firms are composed of more than 10 lawyers. And while the majority of law graduates work in private practice, sizeable numbers work in government (8%), private industry (8%), the judiciary (3%), education (1%), legal aid/public defenders (1%), and private associations (1%) (ABA, 2012).

The Core of the Profession: Private Practice

The Two Hemispheres Thesis

The legal profession is frequently depicted as bifurcated, divided between more elite, highly remunerated lawyers serving wealthy corporate clients and general practitioners, who are less handsomely paid, work in smaller offices, and serve individual clients. For example, in their classic study of Chicago lawyers, Heinz and Laumann (1982) demarcated two kinds of law practice: elite lawyers who represent large corporations and those who represent individuals. Lawyers within each sector differed systematically in terms of the prestige of law schools they attended, career histories, likelihood of engaging in litigation, forums of litigation, social and political values, and social networks (Heinz and Laumann, 1982). The organizational differences within the hemispheres shaped professional opportunities and remuneration.

Twenty years later, Heinz and colleagues returned to study Chicago's bar, comparing 1975 and 1995 cross-sectional surveys of the city's lawyers. In their book, *Urban Lawyers: The New Social Structure of the Bar* (2005), Heinz et al. revealed that the contemporary organization of work is subdivided into smaller, more highly specialized clusters that are less visibly separated by the distinction between corporate and personal client types. Furthermore, they observed a shift from client-based to transaction-based work groups within private practice. They also challenged the idea of hemispheres, particularly the notion of a globe spliced into equal halves. They observed that increasing differentiation of lawyers' roles over the 20 years has resulted in fortified internal boundaries that are 'more well-defined and difficult to cross' (Heinz et al., 2005: p. 317). The hemispheres themselves also changed. Work for corporate clients has become a much larger part of the profession than work for individuals or small businesses. Fields are now more distinct, more clearly separated. Specialization has increased markedly in most fields, especially in the corporate sector. The incomes of partners and associates in major law firms have become lucrative,

while the earnings of lawyers working in solo practices, small firms, and government employment have declined (Heinz et al., 2005).

Upward Mobility: The Partnership Tournament of Law Firms

Not only has the overall structure of the legal profession changed in recent decades, but mobility routes within the profession have also changed. This is particularly the case in law firms. By the 1960s, most elite law firms in the United States adopted a structure pioneered by Cravath, Swaine, and Moore decades earlier. The 'Cravath system' divided firms into two groups: associates, who came directly from law school (or judicial clerkship) and received an annual salary, and partners, who were more experienced lawyers and shared in the profits of the firm (Milton, 2006). The more senior lawyers had more work than they could manage and therefore recruited highly qualified but inexperienced graduates to work on the firm's cases under the supervision of senior lawyers. After an extended probationary period of typically 5–8 years, marked by increasing responsibility, the most proficient associates would be invited into the partnership (the ownership circle of the firm). Associates outnumbered partners and only a select few would advance to the partnership. Those not offered partnership would leave, to be replaced by new recruits. This competition provided incentives for young lawyers to work long hours and for older lawyers to train new recruits without fear they would leave and take clients with them. This model was widely adopted in the United States and became a model embraced in many countries (Henderson and Galanter, 2008).

The ratio of associates to partners, known as 'leverage,' became an important predictor of a firm's profitability. Rising numbers of salaried associates often resulted in enhanced financial returns for partners (Bruck and Canter, 2008). But beginning in the 1960s, law firms began to transform. Businesses faced increasing government regulation through employment law, civil rights, intellectual property, product liability, and antitrust and corporate contracts (Galanter and Palay, 1991). There was an unprecedented surge in demand for corporate legal services. In the 1970s, firms underwent further changes. Clients and lawyers both became more mobile. Firms became more commercial and market oriented, embracing marketing and professional management (Henderson and Galanter, 2008). Starting in 2000, the tournament shifted again. The traditional up-or-out principle was liberalized to include retention of 'permanent associates' and appointment of salaried or nonequity partners (Gorman, 1999). There emerged a class of nontournament lawyers, such as counsel, staff lawyers, staff associates, contract lawyers, and lawyers at outsourced locations. Partners in major law firms, formerly treated as having life tenure, were now sometimes expelled from the partnership. Replacements were recruited from other firms and clients' loyalty extended to lawyers rather than to firms (Coates et al., 2011; Henderson and Bierman, 2009; Heinz et al., 2001). Henderson and Galanter (2008) refer to this new universe as the 'elastic tournament,' because it involves a stretching of the tournament to include partners who work longer hours, accept unequal financial shares, fear losing their share of firm profits, and worry about being pushed to retire early.

Contemporary Changes: From Golden Era to Global Reach

The mid-twentieth century is often depicted as the 'golden age' of professional work. In this era, the earnings and social status of professionals (including lawyers) was much higher than almost all occupations (Gorman and Sandefur, 2011: p. 277). The 1970s and 1980s brought a period of transformative changes. There was a dramatic increase in the overall number of lawyers, and in the number of very large law firms dealing with corporate legal work. There were also fluctuations that occurred in tandem with economic cycles. For example, when the economy stalled in the 1990s, corporate law firms reduced the hiring of new lawyers and many firms terminated excess lawyers (Heinz et al., 2001). The growth of corporate sector practice and the decline in the amount of personal and small business legal work was paralleled by a corresponding realignment of the organizational contexts in which lawyers work. Legal work became increasingly dominated by large organizations as employers of lawyers – these employers included large law firms, corporations, and government. Solo practitioners, once the majority in private practice (64% of all lawyers in the United States worked as solo practitioners in 1960) (Heinz et al., 1998), now represent slightly less than half of private practitioners (ABF, 2012).

Several areas of law practice flourished during the 1980s and 1990s. New specialties gained prominence, such as sports, entertainment, cyber, and health law. Changes in legislation regulating business activity generated commercial legal demands. Welfare legislation, environmental protection, land rights, and antidiscrimination laws further increased demand for legal services (Anleu, 1992). Alongside this growth in substantive areas of law, new types of practice such as legal aid and community legal centers emerged (Weisbrot, 1985).

Another major change was the system of billing clients. Fixed fee schedules gave way to a system of hourly billing, and prohibitions on advertising were eroded in several countries (including Australia and the United States, for example) (Anleu, 1992). The expectations for annual billable hours by lawyers working in law firms rose dramatically in the 1990s (Fourtney, 2005; Richmond, 2000), with a decline after 2007, likely due to the slowdown in the legal economy starting in 2008. Billable hours appear to be on the rise again (NALP, 2012).

During this period, law firms grew dramatically in size and developed distinct practice areas (Nelson, 1988). Departments emerged, defined by substantive expertise such as tax, litigation, real estate, and mergers and acquisitions. Typically, within departments, work was allocated by managers, often assisted by a second level of supervisors (Flood, 2011; Heinz et al., 1998). During the last two decades, large law firms have become increasingly corporatized with the inclusion of human resources, marketing directors, public relations, risk analysis and compliance counsel, finance, and technology departments (Henderson and Galanter, 2008).

These shifts are often described as signaling a trend toward commercialization and 'bureaucratization of the bar' (Heinz et al., 1998: p. 770). Law firms increasingly make use of alternative business structures and management systems for ensuring provision of legal services in compliance with

professional ethical obligations (Parker et al., 2010). As a consequence, autonomy at the organizational level remains considerable, while autonomy for individual lawyers has declined (Gorman and Sandefur, 2011). Some scholars have expressed concern that the changes are evidence of a loss of control over work goals, content, and performance (Abel, 1986). Meanwhile, other scholars find little evidence of managerial priorities subsuming professional ones. Ackroyd and Muzio (2007: p. 731), for example, claim that changes in law firms in England and Wales remain focused on a 'refurbishment' of professional ideals. In a similar vein, Faulconbridge and Muzio (2008) contend that law is undergoing a new form of professionalism, one they term 'organizational professionalism,' whereby traditional principles and objectives are secured through law firm organizational systems.

The massive growth and bureaucratization of large law firms has introduced new obstacles. The rise of large law firms, with broad geographic platforms and highly specialized and profitable practice groups and heavily bureaucratized corporate departments, challenges the task of maintaining shared cultural values and long-term vision (Henderson and Bierman, 2009). Massive growth has been accompanied by rising billable hour expectations, resistance of clients to pay for associate training, more lateral hiring, alarming levels of associate attrition, and a decline in firm culture and community. Under these conditions, law firms may exhibit remarkably little initiative to pursue noneconomic objectives, such as racial and gender diversity or the training and mentoring of new lawyers (Henderson and Galanter, 2008). New challenges are emerging as an incoming generation of lawyers expresses dissatisfaction at the career paths offered within large law firms and clients protest rising billing rates (Nelman, 2010). These tensions may give rise to innovative new forms of legal service delivery. For example, Henderson and Bierman suggest that this dynamic "props open a door to a new wave of boutique and regional law firms with lower cost-structure and a more innovative, cooperative, and client-centric ethos" (2009: p. 1429).

Globalization and International Law Firms

Growth in firm size has in some ways been dwarfed by the emerging global reach of contemporary corporate law firms. While almost all law firms were once locally based, by the 1980s and 1990s it became common for law firms to have offices in several major cities in the United States and abroad (Segal-Horn and Dean, 2011). With ongoing investment of American and British capital into previously untapped markets (South America, Eastern Europe, and Asia, for example), globalization has created a host of new profit-making and client opportunities for those law firms from the United States and Britain that branch beyond domestic borders. There are now American and British law firms in Hong Kong, Shanghai, and Beijing, for example. Curiously in India, where multinational corporations and international accounting agencies and thousands of businesses from abroad actively operate, foreign law firms are barred from practicing (Krishnan, 2010). The emergence of large, often global, professional service firms poses a challenge for traditional professional regulatory regimes (Flood, 2011). Attempts by transnational law firms to import English and American understandings of legal practice to other institutional contexts have led to conflict, negotiation and adaptation, and hybridization in best practices (Faulconbridge et al., 2012).

Global law firms encounter even more complex challenges in China. Since 1992, the year the Chinese government formally permitted foreign law firms to establish representative offices in mainland China, more than 200 law firms have rushed to this lucrative market. The majority of these law firms come from the United States, Britain, France, Germany, Japan, and Australia. The two dominant global law firm clusters are headquartered and parented in the United States and the United Kingdom. However, foreign law firms are barred from practicing Chinese law and from employing licensed Chinese lawyers (Liu, 2008).

At the same time, the legal profession in China has undergone vast changes in recent years. Michelson (2007) highlights how depoliticization of the legal profession (through reform aimed at severing ties with government in the 1990s) resulted in lawyers flocking back to the official system in search of political protection. Some scholars critique a theory of market transition and emphasize the continued role of state-exercised power. For example, Liu (2008) coined the term 'the logic of fragmentation,' to describe the way five service fields related to the legal profession (rural dispute resolution, grassroots legal services, foreign-related legal services, corporate and government legal services, and criminal defense) compete with and intermingle with each other. Liu (2008) claims that state power guided these changes in the Chinese legal profession since reform. Other scholars identify recent powerful social changes that have shaped Chinese legal practice. For example, Xueyao and Jinhua (2013) document Chinese citizens' growing demands on the legal profession since the nation has experienced urbanization, the spread of higher education, and social modernization across fields. These changes have led to increased autonomy of the legal profession, a rise in the number of full-time lawyers, an expanding role played by lawyers in public life, a mounting volume of nonlitigatious cases that require specialization, and emerging differences in the organization of the profession across regions (Xueyao and Jinhua, 2013).

Broader Contours of the Profession

Many lawyers work in small firms or independently as solo practitioners. For most people, these are the lawyers they approach for legal matters. And while large law firms dominate in the urban core of big cities (and in the scholarship about legal profession trends), many sole practitioners and small firms have developed a vibrant clientele base in the suburbs surrounding cities (Seron, 1996). Some small firms exist as franchise law firms offering standardized services to a relatively low-income clientele (Van Hoy, 1997). Other lawyers set up their offices dedicated to particular areas of law, such as divorce and family law (Howieson, 2011). Lawyers also work in numerous contexts besides the standard private practice setting. For instance, today many large, complex organizations employ in-house legal counsel to

manage their legal affairs (Park et al., 2011). In addition, many lawyers work in government, in law clinics and legal aid services, and in international settings such as the World Trade Center or The Hague International Criminal Tribunal.

Demographic Diversity: Gender and Racial Minorities

For many years, the legal profession was composed of an exclusively male and largely majority racial group. As legal barriers to the entry of women fell in the late nineteenth and early twentieth centuries, a gradual trickle of women entered the profession, although access to certain realms of law practice remained restricted (Mossman, 2006). For example, in France, in the interwar years (1930s), women law graduates were denied access to the highly corporatist legal positions of notaries (*notaires*), solicitors (*avoués* and *agréés*), and court clerks (*greffiers*) (Fette, 2007). Changes in legal education facilitated increased demographic diversity in the profession. The move from an apprentice model to law schools in the early twentieth century in the United States, for example, enabled the training of more prospective lawyers. As law schools adopted more universalistic criteria, including standardized tests, access to legal education became more gender neutral. In addition, legal changes emanating from the civil rights movement of the 1960s mandated equal access to higher education for historically excluded groups. As a result, the legal profession came to reflect growing numbers of minorities and women.

Women's representation in the legal profession now far surpasses the small numbers of the 1970s; women have made inroads into all sectors of practice, including elite law firm partnerships, the judiciary, and governance of the bar (Kay and Gorman, 2008). Women represent 30% of the US legal profession (ABA, 2012) and make up 47% of US law graduates (ABA, 2013). However, a significant amount of gender inequality persists. Overall, women are more likely to be found in less remunerative and less prestigious practice settings. In the United States, for example, women are more likely to work in solo practice, government, education, legal aid, and law clinics. There also remains some gender imbalance across substantive practice areas. For instance, women remain underrepresented in law firms. Women are also vastly underrepresented among law firm partners (Kay and Gorman, 2008; Kay and Hagan, 1998). Outside of law firms, women appear to lag behind in senior-level positions in government, public-interest organizations, in-house corporate counsel offices, and in law schools (Hull and Nelson, 2000). Meanwhile, women are overrepresented among those leaving the practice of law (Kay, 1997). Gender inequality and women's attrition from law have been attributed to law firms' inflexible work arrangements and excessive time demands (Walsh, 2012) and to gender bias that limits women's career progress (Wallace and Kay, 2012).

Today's legal profession is also more diverse in terms of race and ethnicity – in 2010, 88% of US lawyers were white, nearly 5% African-Americans, less than 4% were Latino, and about 3% Asian-Americans (U.S. Census, 2010) – yet inequities persist. Racial minority groups are disproportionately concentrated in certain types of practices (Heinz et al., 1998). In the United States, for example, relative to whites, minorities are less likely to be made partners within firms (Chambliss and Uggen, 2000) and those who are invited to the partnership are more likely to be made nonequity partners (Wilkins, 1999). There is also evidence of a significant differential in earnings among law firm lawyers based on race (Ornstein, 2010). Race not only correlates with reduced promotions and lower incomes, but there is also evidence that racial minorities tend to have higher representation in firms with particular characteristics. For example, racial minorities are better represented in offices that are larger and affiliated with larger firms, offices of firms with higher revenues, offices with greater associate-partner leverage, and in branch offices rather than principal offices (Gorman and Kay, 2010).

Explanations for minorities' low rates of retention and promotions in law firms emphasize employers' conscious or unconscious bias rather than individual lawyers' choices and qualifications. Law firm partners may be less likely to select minorities in hiring and promotions due to cognitive bias favoring whites. The cognitive bias, often based on racial stereotypes or in-group favoritism, may lead organizational leaders to perceive minorities as less competent and to hold minorities to stricter standards in order to demonstrate their abilities (Wilkins and Gulati, 1996). This problem may be even greater for minority women. For example, in García-López's (2008) study, female Mexican-American lawyers perceived that their contributions were undervalued within firms and that they were assigned responsibilities beneath their capabilities. As a result, minority associates may be less likely to receive the developmental opportunities that are essential to acquire skills needed for success in law firms (Kay et al., 2009). This disadvantage may compound over time and across firm hierarchies (Kay and Gorman, 2012). As Wilkins (1999) demonstrated through his interviews with African-American partners at major law firms, African-American partners face obstacles in recruiting new clients, obtaining internal referrals for work from existing firm clients, and procuring the services of talented associates.

Consequences and Challenges for the Legal Profession

Professional Ethics and Regulation

The legal profession, similar to medicine and engineering, adheres to a model of self-regulation. Under this model, government has granted the professional community of lawyers a high degree of autonomy to organize their occupation in return for the profession's commitment to assure quality of services and to place public interest above their own (Flood, 2011). The legal profession has established means to inculcate ethical standards in their new members, including through formal professional education (e.g., law school and bar admission courses), apprenticeships, formal examination for admission to the bar, continuing legal education programs, and workplace socialization. Professional associations, such as bar associations and law societies, have established formal systems to monitor and discipline the behaviors of members through such sanctions as suspensions and disbarment.

Dramatic changes in the world of legal work in recent years may have eroded the efficacy of this model of self-regulation. Internally, the legal profession has become more fragmented and less cohesive. The growth in specialization within the profession means specialists in particular legal fields may identify more closely with their community of specialists than with the profession as a whole (Mather and Levin, 2012). Large corporate law firms have expanded globally to meet transnational demand and this global reach may mean lawyers working overseas experience weaker involvement in their local or national communities (Gorman, in press; Wilkins and Papa, 2013). Law firms have to comply with their headquarters' regulatory system and that of the host country where overseas offices are based. This creates a need for regulation that transcends borders (Flood, 2011). Furthermore, although large law firms and corporations that employ lawyers impose their own competing constraints on professional conduct, changing conditions of work threaten the traditional structures of law firms that foster a community of legal professionals necessary for informal monitoring and accountability (Gorman and Sandefur, 2011).

Jurisdictional Disputes and Professional Monopoly

The legal profession's market for services has grown in recent years with the expansion of global and corporate transactions. At the same time, established jurisdictions of the legal profession have been eroded by other groups providing legal advice, including accountants, realtors, managerial consultants, mediators, notary publics, and paralegals (Ackroyd and Muzio, 2007). Many tasks traditionally performed by lawyers are now assigned to paraprofessionals, such as paralegals. The relationship with these other occupations is at times complimentary (as is the case of paralegals employed by solo practitioners or law firms) and at other times, competitive (e.g., where paralegals provide a less costly alternative for legal services). In some countries, like Canada, paralegals have been absorbed into the organizational structure of law societies – the governing body of lawyers. Law societies now monitor educational standards, licensing, and discipline of paralegals and lawyers. Intraprofessional competition also occurs in mixed jurisdictions among legal professionals, as is the case in Quebec, Canada, where *notaries* and *avocats* compete to steer law reform, recruit law students, and expand jurisdictions of practice (Kay, 2009). Yet despite some encroachment on services by other professional groups and internal jurisdictional disputes, lawyers continue to maintain dominance as the experts of legal knowledge.

Today the legal profession remains a highly distinguished profession, one that is organized around rigorous educational requirements and professional training. Lawyers provide legal advice and representation, working across a diverse range of settings, including private practice, government service, private industry, legal clinics, education, judiciary, and corporate counsel. The legal profession has undergone vast changes in recent years in terms of a dramatic growth in numbers, increased demographic diversity, expanded corporate and transnational work, larger law firms, and changing career ladders. Nonetheless, the legal profession continues to enjoy considerable autonomy – to oversee education, accreditation, monitoring, and discipline of its members – as well as a resilient monopoly over legal services.

See also: Civil Law; Common Law; Equality and Inequality: Legal Aspects; Gender and the Law; Globalization: Legal Aspects; Islamic Law; Law Firms; Lawyers, Regulation of; Legal Systems, Classification of; Paralawyers: Other Legal Occupations; Political Lawyering; Race and the Law.

Bibliography

Abbott, A., 1988. The System of Professions: An Essay on the Division of Expert Labor. University of Chicago Press, Chicago.

Abel, R.L., 1986. The transformation of the American legal profession. Law & Society Review 20, 7–17.

Ackroyd, S., Muzio, D., 2007. The reconstructed professional firm: explaining change in English legal practices. Organizational Studies 28 (5), 729–747.

American Bar Association, 2012. The Lawyer Statistical Report, 2012 ed.

American Bar Association, 2013. Section of Legal Education and Admission to the Bar. http://www.americanbar.org/groups/legal education/resources/statistics.html (accessed 13.09.09.).

Anleu, S.L.R., 1992. The legal profession in the United States and Australia: deprofessionalization or reorganization? Work and Occupations 19 (2), 184–204.

Brockman, J., 1999. 'A cold-blooded effort to bolster up the legal profession': the battle between lawyers and notaries in British Columbia, 1871–1930. Social History 32, 209–235.

Bruck, A., Canter, A., 2008. Supply, demand, and the changing economics of large law firms. Stanford Law Review 60, 2087–2130.

Chambliss, E., Uggen, C., 2000. Men and women of elite law firms: reevaluating Kanter's legacy. Law and Social Inquiry 25, 41–68.

Coates, J.C., DeStafano, M.M., Nanda, A., Wilkins, D.B., 2011. Hiring teams, firms, and lawyers: evidence of the evolving relationships in the corporate legal market. Law & Social Inquiry 36 (4), 999–1031.

Deflem, M., 2008. Sociology of Law: Visions of a Scholarly Tradition. Cambridge University Press, Cambridge.

Faulconbridge, J.R., Muzio, D., 2008. Organizational professionalism in globalizing law firms. Work, Employment & Society 22 (1), 7–25.

Faulconbridge, J.R., Muzio, D., Cook, A., 2012. Institutional legacies in TNCs and their management through training academies: the case of transnational law firms in Italy. Global Networks 12 (1), 48–70.

Fette, J., 2007. Pride and prejudice in the professions: women doctors and lawyers in Third Republic France. Journal of Women's History 19 (3), 60–86.

Flood, J., 2011. The re-landscaping of the legal profession: large law firms and professional re-regulation. Current Sociology 59, 507–529.

Fourtney, S.S., 2005. The billable hours derby: empirical data on the problems and pressure points. Fordham Urban Law Journal 33, 171–192.

Galanter, M., Palay, T., 1991. Tournament of Lawyers: The Transformation of the Big Law Firm. The University of Chicago Press, Chicago.

García-López, G., 2008. Nunca te toman en cuenta [They never take you into account]: the challenges of inclusion and strategies for success of Chicana attorneys. Gender and Society 22 (5), 590–612.

Goode, W.J., 1961. The librarian: from occupation to profession. The Library Quarterly 31, 306–320.

Gorman, E.H., 1999. Moving away from "up or out": determinants of permanent employment in law firms. Law & Society Review 33 (3), 637–666.

Gorman, E.H. Professional self-regulation in law and accounting: does the traditional model still hold in the United States and Canada? Sociology Compass, in press.

Gorman, E.H., Kay, F.M., 2010. Racial and ethnic minority representation in large U.S. law firms. Studies in Law, Politics, and Society 52, 211–238.

Gorman, E.H., Sandefur, R.L., 2011. 'Golden Age,' quiescence, and revival: how the sociology of professions became the study of knowledge-based work. Work and Occupations 38 (3), 275–302.

Heinz, J.P., Laumann, E.O., 1982. Chicago Lawyers: The Social Structure of the Bar. Russell Sage Foundation and American Bar Foundation, New York.

Heinz, J.P., Laumann, E.O., Nelson, R.L., Michelson, E., 1998. The changing character of lawyers' work: Chicago in 1975 and 1995. Law & Society Review 32 (4), 751–775.

Heinz, J.P., Nelson, R.L., Laumann, E.O., 2001. The scale of justice: observations of the transformation of urban law practice. Annual Review of Sociology 27, 337–362.

Heinz, J.P., Nelson, R.L., Sandefur, R.L., Laumann, E.O., 2005. Urban Lawyers: The New Social Structure of the Bar. The University of Chicago Press, Chicago.
Henderson, W.D., Bierman, L., 2009. An empirical analysis of lateral lawyer trends from 2000 to 2007: the emerging equilibrium for corporate law firms. The Georgetown Journal of Legal Ethics 22, 1395–1430.
Henderson, W.D., Galanter, M., 2008. The elastic tournament: the second transformation of the big law firm. Stanford Law Review 60, 1867–1930.
Howieson, J., 2011. The professional culture of Australian family lawyers: pathways to constructive change. International Journal of Law, Policy and the Family 25 (1), 71–99.
Hull, K.E., Nelson, R.L., 2000. Assimilation, choice or constraint? Testing theories of gender differences in the careers of lawyers. Social Forces 79 (1), 229–264.
Kay, F.M., 1997. Flight from law: a competing risks model of departures from law firms. Law & Society Review 31 (2), 301–335.
Kay, F.M., 2009. Intra-professional competition and earnings inequalities across a professional chasm: the case of the Québec legal profession. Law & Society Review 43 (4), 901–938.
Kay, F.M., Brockman, J., 2000. Barriers to gender equality in the Canadian legal establishment. Feminist Legal Studies 8, 169–198.
Kay, F.M., Hagan, J., 1998. Raising the bar: the gender stratification of law firm capitalization. American Sociological Review 63 (5), 728–743.
Kay, F., Gorman, E., 2008. Women in the legal profession. Annual Review of Law and Social Sciences 4, 299–332.
Kay, F.M., Gorman, E., 2012. Developmental practices, organizational culture, and minority representation in organizational leadership: the case of partners in large U.S. law firms. The Annals of the American Academy of Political and Social Sciences 639, 91–113.
Kay, F.M., Hagan, J., Parker, P., 2009. Principals in practice: the importance of mentorship in the early stages of career development. Law & Policy 31 (1), 69–110.
Krishnan, J.K., 2010. Globetrotting law firms. The Georgetown Journal of Legal Ethics 23, 57–102.
Liu, S., 2008. Globalization as boundary-blurring: international and local law firms in China's corporate law market. Law & Society Review 42 (4), 771–804.
Mather, L., Levin, L., 2012. Why context matters. In: Levin, L., Mather, L. (Eds.), Lawyers in Practice: Ethical Decision Making in Context. University of Chicago Press, Chicago, pp. 1–24.
Michelson, E., 2007. Lawyers, political embeddedness, and institutional continuity in China's transition from socialism. American Journal of Sociology 113 (2), 352–414.
Milton, C.R., 2006. Eat What You Kill: The Fall of a Wall Street Lawyers. University of Michigan Press, Ann Arbor.
Mossman, M.J., 2006. The First Women Lawyers: A Comparative Study of Gender, Law and the Legal Professions. Hart Publishing, Oxford.
National Association for Law Placement, 2012. Number of Associate Hours Worked Increases at Largest Firms. NALP Bulletin. http://www.nalp.org/billable_hours_feb2012 (accessed 02.10.13.).
Nelman, J., 2010. A little trust can go a long way toward saving the billable hour. Georgetown Journal of Legal Ethics 23, 717–732.
Nelson, R., 1988. Partners with Power: The Social Transformation of the Large Law Firm. University of California Press, Berkeley.
Ornstein, M., 2010. Racialization and Gender of Lawyers in Ontario. A Report for the Law Society of Upper Canada. http://www.lsuc.on.ca/media/convoct04_ornstein.pdf (accessed 17.09.13.).
Park, S., Sine, W.D., Tolbert, P.S., 2011. Professions, organizations, and institutions: tenure systems in colleges and universities. Work and Occupations 38 (3), 340–371.
Parker, C., Gordon, T., Mark, S., 2010. Regulating law firm ethics management: an empirical assessment of an innovation in regulation of the legal profession in New South Wales. Journal of Law and Society 37 (3), 466–500.
Pérez-Perdomo, R., 2006. Latin American Lawyers: A Historical Introduction. Stanford University Press, Stanford.
Richmond, D.R., 2000. The new law firm economy, billable hours, and professional responsibility. Hofstra Law Review 29, 207–234.
Segal-Horn, S., Dean, A., 2011. The rise of super-elite law firms: towards global strategies. The Service Industries Journal 31 (2), 195–213.
Seron, C., 1996. The Business of Practicing Law: The Work Lives of Solo and Small-Firm Attorneys. Temple University Press, Philadelphia.
U.S. Census, 2010. Bureau of the Census.
Van Hoy, J., 1997. Franchise Law Firms and the Transformation of Personal Legal Services. Quorum Books, Westport.
Wallace, J.E., Kay, F.M., 2012. Tokenism, organizational segregation, and coworker relations in law firms. Social Problems 59 (3), 389–410.
Walsh, J., 2012. Not worth the sacrifice? Women's aspirations and career progress in law firms. Gender, Work and Organization 19 (5), 508–531.
Weisbrot, D., 1985. The changing face of Australian legal practice. Australian Quarterly 58, 426–439.
Wilkins, D., 1999. Partners without power? A preliminary look at black partners in corporate law firms. Journal of the Institute for the Study of Legal Ethics 2, 15–48.
Wilkins, D., Gulati, M., 1996. Why are there so few black lawyers in corporate law firms? An institutional analysis. California Law Review 84, 493–625.
Wilkins, D.B., Papa, M., 2013. The rise of the corporate legal elite in the BRICS: implications for global governance. Boston College Law Review 54 (3), 1149–1184.
Xueyao, L., Jinhua, C., 2013. Structural constraints on legal change: Chinese lawyers in the interaction between the state, the market and society. Social Sciences in China 34 (1), 58–77.

Relevant Websites

http://www.americanbar.org/aba.html – American Bar Association.
http://www.barcouncil.org.uk/ – The Bar Council, Representing Barristers in England and Wales.
http://www.barreau.qc.ca/pdf/publications/brochure-profession-avocat.pdf – le Barreau du Québec.
http://www.cba.org/ – Canadian Bar Association.
http://www.lawsociety.org.uk/ – The Law Society of England and Wales, Representing Solicitors in England and Wales.
http://www.nalp.org/ – The National Association for Law Placement (US).
http://www.wsba.org/Legal-Community/Minority-Bar-Associations/National-Minority-Bar-Association-Directory – National Minority Bar Association Directory (US).

Lazarsfeld, Paul Felix (1901–76)

Christian Fleck, University of Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Paul F. Lazarsfeld is best known for his contributions to methodology, in particular to the quantitative side. However, he possessed a much broader portfolio that enabled him to become one of the leading figures of social sciences in the decades after the end of WWII. His Viennese upbringing stocked him with cultural capital and his mingling in the Social Democratic movement of his hometown convinced him that the world could be made a better place, although this *weltanschauung* may later have shifted to changing the smaller world around him. His major impacts on social sciences of the second half of the twentieth century have been (1) the creation of research institutes devoted to empirical social research, funded by outside sources and offering students a training site; (2) inventing and refining techniques of data collection and analysis; and (3) enriching the usability and applicability of the sociological body of knowledge in the fields of socio-psychological consequences of long-term unemployment, mass media, and audience communication, and the then underdeveloped field of the sociology of social sciences.

Not all scholars possess the same type of academic portfolio, nor do they follow identical paths to become recognized and exhibit influence. Self-confident and humble inhabitants of the academic world pursue observable strategies – even hermits have one – to make their names and accomplishments known to others. Usually academics of all branches favor written communication, publish books or articles, or try to be present in person at as many professional gatherings as possible, exhibiting what even the vernacular now refers to by a sociological-sounding term: networking. In the past, only a minority explored alternatives to the traditional mode of communicating scholarly deeds. One twentieth century social scientist was particularly inventive in making his vision of social research known to others; his name is Paul Felix Lazarsfeld (PFL for short in the remaining text). He created institutions and invented a new label for his preferred version of research: 'empirical social research.' The establishment of stable patterns of collaboration allowed him to disseminate his vision of the craft to others by recruiting disciples to work on his ideas, and he was successful even in persuading people of equal status to join in such endeavors. A portrait of this "institution man," a label he proposed for people like himself (Lazarsfeld, 1969: 302), has to pay tribute to several facets of his performance and personality but should not sidestep the more conventional products, his writings. I will start with presenting PFL's core convictions with regard to the craft of social research, then discuss his methodological contributions, followed by an overview of his substantive work, and end with a gloss on his biography and the contexts in which he was able to develop all these accomplishments.

Lazarsfeld's Vision

At the core of PFL's convictions is the belief that the explication of what ordinary researchers do will help to improve future work in social sciences. He might have picked up the technique of *explication de texte*, or close reading, during his early visits to France, or perhaps his call for explication merely exhibits the pattern of multiple discoveries. In any case, this attitude became a characteristic element of his work habit and philosophy over a period of half a century. Something of a corollary to this is his understanding of the aims and merits of methodology: "I read somewhere a quotation from the English poet Wordsworth who said that poetry is emotion recollected in tranquility […] Since then I have always stressed that methodology is intuition reconstructed in tranquility" (Lazarsfeld to Hans Zeisel, 5 October 1967, as quoted in: Pasanella (1994): 22). Also he phrased his vision of the task of the methodologist elsewhere only slightly differently: "The sociologist studies man in society: the methodologist studies the sociologist at work" (Lazarsfeld, 1959: 171). Besides writing on methodology, he devoted even more energy to disseminating his insights and the findings to others, for which he used different paths and suggested even more than he used.

A consequence of his reconstructive methodological efforts was that PFL became aware of lacunae and shortcomings in the conduct of other social researchers, and to fill these gaps he regularly invented new techniques and encouraged others to do the same. Alongside the explicator we see therefore the constructor and individual who experimented regularly to improve the toolbox of social researchers. Disentangling these two sides is nearly impossible and not what is called for here, besides the fact that in looking at PFL one needs to be aware of both these sides of him.

Whether particular traits of a personality belong to the individual's self-image, or could be analyzed separately, or should be seen as the causal factor behind one's utterances do not need to be resolved here. But PFL's managerialism, activism, and restlessness – or what one could call a nearly permanent eagerness to join in and make suggestions – dramatically influenced his successes and also contributed to his failures. PFL always was a *homo politicus*, someone who wanted to intervene into the world around him and who probably devoted more energy to keeping his businesses running than to disengaging from the world at least temporarily to fine-tune a manuscript. At the same time, PFL was highly aware of the role of published words and regularly

pleaded with editors and publishers regarding his overdue submissions by proclaiming that he would not become famous for adhering to deadlines.

Lazarsfeld's Substantive Contributions to Social Research

As indicated, PFL's methodological work consists of three parts: first, inventions of new techniques; second, reconsiderations of the work done by others from the point of view of the implicit rules that made those works possible; and third, didactic means to help others to profit from predecessors.

It is easiest to start with the third point. At the age of 28, PFL published his first textbook *Statistisches Praktikum für Psychologen und Lehrer* (Lazarsfeld, 1929) which offered the students of psychology at the University of Vienna, Austria basic knowledge of statistics. In that work PFL explained simple features at some length, for example, how to transform absolute numbers into percentages and how to calculate a correlation coefficient. After migrating to New York, PFL contributed chapters to a textbook-like compendium for market researchers, *The Technique of Marketing Research* (1937), presenting insights about the design of questionnaires from a psychological perspective and elaborating on techniques of classification. During the next two decades, PFL wrote a huge number of papers and assembled his own together with papers from others in the highly influential *The Language of Social Research: A Reader in the Methodology of Social Research*, edited together with Morris Rosenberg and published with a leading publishing house for social sciences at that time, The Free Press, in 1955. He downplayed the number of papers that were his own by attributing some of them to Elias Smith, his pseudonym. While the reader's title was telling, only its subtitle provided a real rationale for the volume from the editors: by offering best practices the individual creative scholar could get a better starting background for her endeavors. In 1972, *Continuities in the Language of Social Research* expanded the scope of contributions even further, but by then the market for research technique textbooks was not as receptive as in the middle of the 1950s when the Lazarsfeld-Rosenberg reader was used nearly everywhere social research was taught.

Providing teaching material was one side of PFL's mission to transform empirical social research from an amateur activity into a cumulative scientific effort. From the very start in Vienna, he regularly brought together groups of collaborators, mostly subordinates working on his projects, to offer them instruction, develop new ideas and applications, and refine present ones. In addition to his social democratic background that favored egalitarianism, he was convinced that if one wants to become a good empirical social researcher one needs to go through a period of apprenticeship (Stehr, 1982: 154). There were several more or less well-defined steps one had to climb before one could reasonably start doing her own project. Even ancillaries such as secretaries, endlessly typing draft after draft of manuscripts, and *Strichler*, people who count questionnaires' answers by hand in the form of stripes on paper, could rise to higher levels of prosperity (one of PFL's first American secretaries, Rose K. Goldsen, eventually became a professor of sociology at Cornell University). Later on, when the Bureau of Applied Social Research (BASR) became affiliated with Columbia University, graduate students could start their careers at the BASR as a coder, calculator, or research assistant, use its data and infrastructure to write a dissertation, and then be hired as an affiliate of the 'Bureau,' as it was called. Those who successfully completed third party financed projects remained for a while or went on in their academic careers. PFL's attempts to institutionalize this form of instruction beyond the space of his 'firm' however failed.

The list of publications from PFL does not show any direction. Contrary to major sociological theoreticians, his oeuvre does not evolve like a program incubated step by step. Looking for a label for PFL's style of methodological thinking one could characterize it as offering solutions on demand. Several observers and witnesses confirmed that PFL always saw interesting connections between variables, caught evolving problems out of field notes, and detected spurious correlations much quicker than anyone else around the table. Some of these serendipitous problems he solved overnight, so to speak, others remained in his head and in his notes for very long periods. Let us for comparison look at one smaller revelation and one of the long-lasting methodological struggles PFL fought.

During the years PFL directed the Princeton Radio Research Project (1937–40), he invented, without calling it by the name, secondary analysis. He was able to do so because the financial backer of the whole project, the Rockefeller Foundation, did not press him to collect data first hand. The foundation officers were interested in understanding the then new mass communication device, the radio, and in particular its effects on the audience. At the very beginning the following question was put in front of the research team: "do those who listened to educational programs change their opinions afterward or not?" Ingeniously PFL persuaded market researchers to hand their data over to him and promised to deliver them surplus value in exchange. These data could not be analyzed in today's style of meta-analysis because the Hollerith machines that were used severely restricted reuse of punched cards, and data stacking techniques were not yet available. Still, PFL identified, in each study, crucial variables, cross-tabulated them, and then compared the results from the different studies. For instance, he was able to show that those who held an opinion beforehand overwhelmingly held that they did not care to intentionally listen to educational or political programs, but a third of those listening did reform their opinion afterward, be it by changing from one party to another, or from nonvoting to voting, and so on. PFL presented these insights first in a small paper (originally published as Lazarsfeld, 1939; reprinted in Lazarsfeld, 2011: pp. 401–416) where he inserted new variables step-by-step and came to revealing conclusions. Even more telling than the results is the long-lasting effect of the insight it provided. One could easily draw the invention of the panel study design back to these small observations about changes during a campaign. During the 1940 presidential campaign which resulted in the re-election of Franklin D. Roosevelt for the third time, PFL first used repeated interviews of the same sample of respondents, the panel design (Lazarsfeld et al., 1944).

Not all problems could be satisfactorily solved in similar ways, PFL worried. Take as an illustration his struggle finding out why people do what they do. Very early in his career he contemplated the occupational choices of young Austrian workers (Lazarsfeld, 1975a). Under the influence of his then mentor Karl Bühler he thought about the interplay between 'outer' and 'inner' factors and called the acting of one upon another a 'motive.' Unsatisfied with this pseudo-solution to his big problem PFL some years later, probably under the influence of those American sociologists who later became known as symbolic interactionists, distinguished between 'inner motives' and 'mechanism,' which direct an individual, and the outer 'situation,' whereby only the encounter of these two independent forces cause an 'action' (Lazarsfeld and Kornhauser, 1955). Some years later, he again returned to this topic when he tried to understand why people vote as they do. Now under the spell of quite another intellectual environment he theorized that a citizen, labeled an 'organism,' stands under the influence of on the one hand a stimulus, for example, a political campaign, that she recognizes selectively, and on the other hand a set of personal dispositions. This aspect of the formation of the voting process is the concern of psychology, whereas sociology studies the 'implementation' of all these causes, which results in a particular vote as the response to all the foregoing influences (Berelson et al., [1954]1986: 278). Still not really satisfied with this solution to the big problem, which he once jokingly called the "methodological equivalent of socialist voting and buying of soap" (Lazarsfeld, 1969: 279), he finally persuaded the editor of the 1968 edition of the *International Encyclopedia of the Social Sciences*, his longtime collaborator and later biographer David Sills, to reserve an entry for his 'theory': reason analysis (Kadushin, 1968). It does not diminish his legacy if one adds that this particular part of his oeuvre was not highly consequential.

Nothing could be further from the point than to come to the conclusion that PFL's attempts should be seen as failures. Just the opposite, along the way he found serendipitously other interesting things and even invented a machine: the famous Program Analyzer that documents positive versus negative evaluative responses by listeners, which was the result of his collaboration with Frank Stanton from CBS. During his whole career PFL codified data collection techniques, improved the phrasing of questions, and developed data analysis tools and routines.

Usually PFL is put into the quantitative camp of the social sciences which is correct only with some qualifications. The majority of his written contributions to methodology belong more or less to the quantitative branch, but from the very beginning until his last days PFL thought seriously about nonquantitative research strategies. Although he once argued prominently that qualitative research should be restricted to the exploratory steps of any research project (Barton and Lazarsfeld, 1955; reprinted in Lazarsfeld, 1993: 210–217), he conceded elsewhere to qualitative approaches a more independent value. First, PFL was never an advocate of random sample large-scale survey research. Instead he preferred smaller and pointed samples – as in all the famous voting studies under his direction: Erie County, Ohio; Decatur, Illinois; and Elmira, New York – but he was willing to go one step further by using case study designs from the very beginning onward.

Second, he assigned an independent role to qualitative analysis whenever quantitative data analysis resulted in anomalies it could not resolve statistically. PFL's plea for what he labeled 'deviant case analysis' did not get as much attention as others of his suggestions but it remained on the table since its first announcement.

Third, he even prepared a separate statement on methodology immediately after the study on unemployment was published in 1933. This study's subtitle classified the investigation as belonging to sociography, never a widely used name. The methodology manuscript remained unpublished but PFL recycled its main message when he wrote his memoir 34 years later. Quoting from the then still unpublished paper (now included in Lazarsfeld, 2011: 243–265) he proposed five rules for data collection:

> a. For any phenomenon one should have objective observations as well as introspective reports.
> b. Case studies should be properly combined with statistical information.
> c. Contemporary information should be supplemented by information on earlier phases of whatever is being studied.
> d. One should combine 'Natural and experimental data,' meaning questionnaires and solicited reports and unobtrusive measures.
>
> Lazarsfeld, 1969: p. 282

Given his educational background in mathematics it is no wonder that PFL tried to utilize his competences, but it seems that he largely abstained from pursuing this path, focusing instead on more substantive thinking. Whatever he might have contributed to the new specialty of mathematical sociology – obviously, he helped establish it organizationally – PFL remained an empiricist. While some of his collaborators and disciples later joined the newly formed rational choice camp in sociology, one could argue, probably convincingly, that PFL would have continued proposing an empirical theory of action (see Lazarsfeld, 2011) instead of jumping into the highly abstract mode of theorizing to be found in rational choice papers.

Substantive Contributions by PFL

PFL is very often reduced to his role as a methodologist, sometimes even more narrowly seen as a quantitativist; and he is not to be acquitted from contributing to this caricature, himself, for example, when he referred to this stereotype in jokes he made or quoted, not realizing that the audience might not get his irony (e.g., Lazarsfeld, 1972: xvi; Stehr, 1982). However, the list of substantive contributions to several special branches of sociology and neighboring disciplines is impressive, and the fact that decades after his death some of them are still quoted is impressive in itself. A study of socio-psychological consequences of long-lasting unemployment made him known to social scientists for the first time when in 1933 *Die Arbeitslosen von Marienthal* (The Unemployed of Marienthal) came out in print. Its first edition did not list any authors' names on the cover but indicated that the study had been done by the *Wirtschaftspsychologische Forschungsstelle*; the authors' Jewish-sounding names had been removed to secure

the study's publication at that turbulent time (see my introduction to the reprint for more details, Jahoda et al., [1933] 2002). During his first years in New York, PFL wrote some shorter papers on this topic and influenced through verbal communication Robert and Helen Lynd who redesigned their follow-up study *Middletown in Transition* (Lynd and Lynd, 1937) under the influence of *Marienthal*. Marie Jahoda, PFL's Marienthal study collaborator and first wife, exhibited a continuing interest in unemployment; she returned to this topic in the 1970s when the level of unemployment again was growing, whereas PFL stopped studying unemployment himself and contributed to this topic later on only indirectly as the dissertation supervisor of Mirra Komarovsky ([1940]1971).

PFL contributed several key ideas to mass media and audience research, and the studies done at the BASR formed what he called the 'Columbia tradition' (because he disliked 'school' as a designation). According to this tradition the effects of mass media are not as dramatic as educated lay people feared. As a consequence PFL did not buy into the hidden persuader message of Vance Packard, which influenced the upper stratum of middle class people in the whole Western world, and he refrained from joining the chorus of those lamenting the spread of mass society. The main point of disagreement between the Columbia perspective and the, in itself highly diverse, critical approaches can be formulated as the difference between causal attribution (media cause something) and second-tier effects (media support particular attitude formation). Critical theorists of mass media's influence argue that media itself, starting with the movie theater, radio, then television, and more recently the Internet, cause alienation, passivity, isolation, and so on, whereas PFL and his collaborators assigned to the media a more differentiated ensemble of functions (Katz, 1987). Obviously, mass media confer status to whatever is covered by them. Someone's appearance on a television show makes this person known to a huge number of viewers, and the same is true for political issues and other impersonal topics covered by the media. This way mass media legitimize the status of what they have chosen to present to their audience. Secondly, the mass media assure social norms. Recall that particular deviant behavior was banned from being broadcast up to very recent times, and 'problematic content' is still banned in countries like the People's Republic of China. During PFL's lifetime, this function was also enforced in advanced Western societies, and one can only speculate what he would say about recent developments that have nearly broken down any barriers by airing the weirdest forms of behavior. Nevertheless, the media suitability of only particular types of politicians, for example, could still be seen as proof of this point. Very much in agreement with any critical theorists, the Columbia duo – PFL and Robert K. Merton – identified as a third feature the narcotizing dysfunction of the media, something the two wrote about as early as 1948 but could be found in more recent studies on the decline of civility, weakening of social bonds, and so on (Lazarsfeld and Merton, 1948).

The most consequential and most often cited concepts PFL coined originated at the crossroad between audience and voting research. Studies about the formation of political decisions revealed telling insights. The two-step-flow of communication might be the most prominent. This was first mentioned, in a characteristically Lazarsfeldian way, in the closing remarks of *People's Choice*. There PFL hinted at more research that should be done, for which he asked for future money, and mentioned in particular the dissemination of political messages through opinion leaders and the micro details of the change of one's political opinion (Lazarsfeld et al., 1944). The notion of a two-step-flow derived from the community study he conducted where researchers first met 'opinion leaders,' and obtained people's reaction toward 'cross pressure' before Election Day, however did not see a direct and unequivocal impact of the media. The 'crystallizers,' – individuals who did not change their preferences during the campaign, reacted toward the media messages completely differently from those who followed the media for entertainment reasons, not to mention those who made use of the news to persuade members of their social networks for whom to vote.

Seminar discussions at the Bureau helped to develop another pair of concepts. BASR's associate director Merton distilled around that time, from different empirical material, the opposition of the local and the cosmopolitan (Merton, 1968: pp. 441–474). According to PFL and his colleagues, political messages could not be delivered to voters directly but only through the help of the cosmopolitan opinion leaders. Political parties, candidates, and anyone who wants to distribute their messages to audiences need the help of intermediators who enforce the message through their talking to other people. In *Personal Influence* PFL with the help of Elihu Katz refined and confirmed the two-step flow mechanism further (Katz and Lazarsfeld, 1955; for a more recent reevaluation see Simonson, 2006).

A third section of PFL's substantive contributions consists of his writings on the history of empirical social research, the present status of the field, and the utilization of applied social research. His concern with the past could be traced back at least to *Marienthal* when PFL suggested to his friend Hans Zeisel to round out the book with a historical sketch of sociography. *Academic Mind* (Lazarsfeld and Thielens, 1958), a monograph on the 'social scientist in a time of crisis' according to the subtitle, can be seen as the beginning of a 'sociology of the social sciences' as a distinct feature of PFL's portfolio. It also contains several of his own papers on the roots of the style of research he committed himself to. Particularly worthy of mention here is his rediscovery of Adolphe Quételet – PFL became Quételet Professor of Social Science at Columbia in 1962 – as a forerunner of quantitative studies, as also his commissioning of a couple of dissertations devoted to early stages of empirical social research in the United States, Germany, and France. Besides the rediscovery of a forgotten quantifier and the investigation of developments at the national level, his publications also contain investigations of the role of social sciences, and empirical social research in particular, in contemporary times. During the last decade-and-a-half of his life PFL devoted some of his curiosity to this topic and, again, he tried to persuade others to join him, this time less successfully than in his earlier attempts (Lazarsfeld et al., 1967, 1975).

Academic Mind demonstrates PFL's capabilities and convictions in more than one way. First of all, it took some courage to take over a study about the consequences of the hysteria

orchestrated by the US Senator Joseph McCarthy in the midst of the 1950s. Secondly, the design of the data collection is unique: PFL was able not only to hire two public opinion research agencies to administer the interviews but also commissioned David Riesman afterward to execute a restudy of both the interviewers and the interviewees to prove the validity of the whole study. The report of the findings is of the finest PFL and his collaborators could deliver. Cleverly constructed indices, measuring 'worry' and 'caution,' result in a combined index of apprehension. This measure varies according to characteristics of individuals and also with those of the place where one works: 'permissiveness' on the one side and 'quality of the college' on the other. If one has to select an exemplar for illustrating the BASR style of conducting empirical social research, this book could serve above all others. It is indeed surprising that when PFL came under fire for his 'abstracted empiricism' Mills [1972]1959 he did not pull out this book for defense, but instead reprinted the accusation (Lazarsfeld, 1972: 428–440).

The Personal Equation

In his 1969 'Memoir' PFL listed as a third factor to disentangle the roots of his research style, what he called the 'personal equation' – the two others were the ideological component and intellectual climate. Whether his introspective explanation was pertinent is not to be discussed here. But it is clear that in his case biography did shape his oeuvre. There is no book length biography of PFL at hand, but besides his autobiographical essay and some interviews he gave, we do have several biographical portraits providing the cornerstones of his trajectory (Sills, 1979, 1987; Pollak, 1980; Sills, 1987; Morrison, 1998). Born in 1901 into an assimilated Jewish upper middle class family in Vienna, young Paul was more affected by his mother than his reserved father Robert, a lawyer. Sofie Lazarsfeld (1881–1976) not only held a 'salon' in the family's inner district apartment, but she also earned recognition as an individual psychologist, as the dissidents from Freudianism under the leadership of Alfred Adler called their approach, and as the author of a counseling book *Woman's Experience of the Male* (1931). Due to the social network of his mother, PFL became acquainted with several leading intellectuals of the time. Friedrich Adler, no relation to Alfred, the lifelong lover of Sofie, impressed PFL both intellectually and politically. Because of him, PFL started studying mathematics, and after Adler's assassination of the Habsburg prime minister in 1916 PFL visited him in jail and took part in demonstrations outside court. PFL's early years were marked primarily by his political activism. Being what today in the United States would be called a 'community organizer' brought young PFL in close contact with workers and activists from the Social Democratic party. He organized summer camps for working class children and young adults where he also gave lessons. Party affiliations brought him to France for a while, and during his student years he also participated in a political cabaret and played the viola at the home of the Jahodas, the family of his first wife. His Jewishness was highly visible, but his affiliation with Judaism did not last long. When the Austrian Republic was founded in 1918 people could renounce their religion and become *religionslos* ('nonreligious'), and PFL immediately joined the group of nonbelievers. His Jewish background would have made it difficult to climb up in the party; and although this would not have been completely impossible, a regular career in the Austrian academic world was banned for leftist Jews. PFL, therefore, never held any official position at the University of Vienna but belonged to the circle around the psychologists Charlotte and Karl Bühler from their arrival in Vienna in 1924 until his move to New York in 1933. Intellectually, his perspective consisted of an Austro-Marxist view which functioned for him as a substitute for any formal education in social sciences, a diverse familiarity with both the academic and the extramural wing of psychology, and a competence in mathematics which enabled him to do statistical analyses at a much more sophisticated level than most of the other people around him doing social research. Due to the intellectual climate in interwar Vienna, he also picked up some of the ideas which later became known as the Vienna Circle of neopositivism.

PFL came to the United States first on a fellowship that the Rockefeller Foundation offered to 'bright young men,' as it was expressed back then. At the end of the 2-year (1933–35) fellowship, the still young man – recently divorced, jobless in his European hometown, which also had been taken over by a right wing authoritarian political movement, but possessing a considerable amount of cultural capital – took the riskier path and returned from his native Vienna to New York to look for a job. Thanks to Robert Lynd, PFL was able to secure a meager job in New Jersey, acting as a counsel for young unemployed people and being paid himself from New Deal schemes. He survived the first two years on the fringe of the academic world and got a chance in 1937 to run a well-funded research project, which became known as the Princeton Radio Research Project. From then onward PFL managed to hide himself inside institutions he himself had established. His first attempt, back in Vienna, where he founded a research unit called the *Wirtschaftspsychologische Forschungsstelle* in 1931, was a failure, economically. It never earned enough money to pay the bills, but it was intellectually stimulating enough to produce what is still the only world-renowned empirical social research book written by Austrians, the *Marienthal* study. In Newark, PFL managed to raise enough money so that his small unit could survive. After the first Rockefeller contract came to an end in 1940 PFL became a devoted and successful 'foundation beggar.'

In 1941, Columbia University hired two new sociology professors, PFL at the level of associate and Merton as assistant professor, because the department could not come to a consensus on what specialization they should prefer. Instead of fighting each other the two men who did not know each other before started a marvelous collaboration and also became close friends; both told the story first to a journalist and later wrote about it from their respective points of view (Hunt, 1961; Lazarsfeld, 1975; Merton, 1998). PFL remained in Columbia until he reached retirement age in 1969. Probably disappointed by Columbia not offering him emeritus status, he commuted during the last years of his life to the University of Pittsburgh, where he had been appointed Distinguished Professor of Social Sciences. As a consultant for the Ford Foundation and other

such organizations he traveled regularly to Europe, where he arranged consultancies or visiting professorships for himself in Oslo, Paris, and Vienna. At the Bureau, in particular, from whose directorship PFL resigned as early as in 1949 but where he remained an associate director for the rest of his life, he assembled a distinct group of disciples around him. Out of this microenvironment several famous studies emerged which did not fall into PFL's own field of specialization but whose authors had gotten much inspiration from him. To mention just a few: Lipset's *Union Democracy* (Lipset et al., 1956), Lerner's *Passing Tradition* (Lerner, 1963), and Coleman et al.'s early network study on the diffusion of knowledge within the medical profession (Coleman et al., 1966). However, it seems that the Bureau had its best time before the 1960s, at which time a new mood conquered the hearts and minds of young social scientists. PFL's ingenuity at getting funds from every corner of the moneyed world – the publisher of low-quality magazines McFadden, Philip Morris, or the Office of Naval Research – was suddenly seen from the opposite perspective, and he was accused of 'following the money.'

At least one generation lies between his death and the present, so PFL has become a figure from history. As such it would be worthwhile to see a full-fledged biography of him come out, one offering an interpretation of some of the sides of him that have been difficult to understand: his contradictory feelings of inferiority, marginality, superiority, and his professional self-confidence; why he saw his migration to New York more like an exile than an early instance of being a global player; his generosity and his bullying that some characterized as running a 'machine'; or his lack of awareness of the potentialities for misunderstandings as, for example, in the case of selecting the name of his own style of doing research, namely 'administrative research.'

While waiting for such a book it is worth studying his writings, at least to learn how good empirical social research has been administered.

See also: Celebrity; Communication Research and Media Studies, History of; Communication, Twostep Flow of; Deprivation: Relative; Frankfurt School: Institute for Social Research; Functionalism, History of; Gatekeepers in Social Science; Media Events; Merton, Robert K (1910–2003); Mills, Charles Wright (1916–62); Public Opinion: Social Attitudes; Quetelet, Adolphe (1796–1874); Rational Choice Theory in Sociology; Sociology, History of; Vienna Circle: Logical Empiricism; Voting, Sociology of.

Bibliography

Barton, A., Lazarsfeld, P.F., 1955. Some functions of qualitative analysis in social research. Frankfurter Beiträge zur Soziologie 1, 321–361. Partly reprinted in Lazarsfeld 1993, 210–217.

Berelson, B., Lazarsfeld, P.F., McPhee, W.N., 1986[1954]. Voting: A Study of Opinion Formation in a Presidential Campaign. University of Chicago Press, Chicago.

Coleman, J.S., Katz, E., Menzel, H., 1966. Medical Innovation: Diffusion of a Medical Drug Among Doctors. Bobbs-Merrill, Indianapolis.

Hunt, M.M., 28 January 1961. Profiles: how does it come to be so? The New Yorker 39–63.

Jahoda, M., Lazarsfeld, P.F., Zeisel, H., 2002[1933]. Marienthal: The Sociography of an Unemployed Community. With a new introduction by C. Fleck. Transaction Publishers, New Brunswick.

Kadushin, C., 1968. Reason analysis. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences. Macmillan, New York, pp. 338–343.

Katz, E., 1987. Communications research since Lazarsfeld. Public Opinion Quarterly 51, S25–S45.

Katz, E., Lazarsfeld, P.F., 1955. Personal Influence: The Part Played by People in the Flow of Mass Communications. Free Press, Glencoe.

Komarovsky, M., 1971[1940]. The Unemployed Man and His Family: The Effect of Unemployment upon the Status of the Man in Fifty-nine Families. Octagon Books, New York.

Lazarsfeld, P.F., 1929. Statistisches Praktikum für Psychologen und Lehrer. G. Fischer, Jena.

Lazarsfeld, P.F., 1939. The change of opinion during a political discussion. Journal of Applied Psychology 23, 131–147 reprinted in Lazarsfeld 2011: 401–416.

Lazarsfeld, P.F., 1959. Problems in methodology. In: Merton, R.K., Broom, L., Cottrell Jr., L.S. (Eds.), Sociology Today: Problems and Prospects. Basic Books, New York, pp. 39–78.

Lazarsfeld, P.F., 1969. An episode in the history of social research: a memoir. In: Fleming, D., Bailyn, B. (Eds.), The Intellectual Migration: Europe and America, 1930–1960. Belknap Press of Harvard University Press, Cambridge, pp. 270–337.

Lazarsfeld, P.F., 1972. Qualitative Analysis: Historical and Critical Essays. Allyn & Bacon, Boston.

Lazarsfeld, P.F., 1975a[1929]. Jugend und Beruf: Kritik und Material. Arno Press, New York.

Lazarsfeld, P.F., 1975. Working with Merton. In: Coser, L.A. (Ed.), The Idea of Social Structure: Papers in Honor of Robert K. Merton. Harcourt Brace Jovanovich, New York, pp. 35–66.

Lazarsfeld, P.F., 1993. On Social Research and its Language. University of Chicago Press, Chicago. With an introduction by R. Boudon.

Lazarsfeld, P.F., 2011. An Empirical Theory of Social Action: Collected Writings. Bardwell Press, Oxford.

Lazarsfeld, P.F., Gaudet, H., Berelson, B., 1944. The People's Choice: How the Voter Makes up His Mind in a Presidential Campaign. Duell, Sloan and Pearce, New York.

Lazarsfeld, P.F., Kornhauser, A., 1955[1935]. The analyis of consumer actions. In: Lazarsfeld, Rosenberg (Eds.), pp. 392–404.

Lazarsfeld, P.F., Merton, R.K., 1948. Mass communication, popular taste, and organized social action. In: Bryson, L. (Ed.), Communication of Ideas. Harper, New York, pp. 95–118.

Lazarsfeld, P.F., Rosenberg, M. (Eds.), 1955. The Language of Social Research: A Reader in the Methodology of Social Research. Free Press, Glencoe.

Lazarsfeld, P.F., Thielens Jr., W., 1958. Academic Mind: Social Scientists in a Time of Crisis. Free Press, Glencoe.

Lazarsfeld, P.F., Sewell, W.H., Wilensky, H.L. (Eds.), 1967. The Uses of Sociology. Basic Books, New York.

Lazarsfeld, P.F., Pasanella, A.K., Rosenberg, M. (Eds.), 1972. Continuities in the Language of Social Research. Free Press, New York.

Lazarsfeld, P.F., Reitz, J.G., Pasanella, A.K., 1975. An Introduction to Applied Sociology. Elsevier, New York.

Lerner, D., 1963. The Passing of Traditional Society: Modernizing the Middle East. Free Press, Glencoe.

Lipset, S.M., Trow, M., Coleman, J.S., 1956. Union democracy: The Internal Politics of the International Typographical Union. Free Press, Glencoe.

Lynd, R.S., Lynd, H.M., 1937. Middletown in Transition: A Study in Cultural Conflicts. Harcourt, Brace and Company, New York.

Merton, R.K., 1968. Social Theory and Social Structure. Free Press, New York.

Merton, R.K., 1998. Working with Lazarsfeld: notes and contexts. In: Lautman, J., Lécuyer, B.-P. (Eds.), Paul Lazarsfeld (1901–1976). La sociologie de Vienne à New York. L'Harmattan, Paris, pp. 163–211.

Mills, C.W., 1972[1959]. Abstracted empiricism. In: Lazarsfeld (Ed.), pp. 428–440.

Morrison, D.E., 1998. The Search for a Method: Focus Groups and the Development of Mass Communication Research. University of Luton Press, Luton.

Pasanella, A.K., 1994. The Mind Traveller: A Guide to Paul F. Lazarsfeld's Communication Research Papers. Columbia University, New York.

Pollak, M., 1980. Paul F. Lazarsfeld: a socio-intellectual biography. Knowledge 2, 157–177.

Sills, D.L., 1979. Lazarsfeld, Paul F. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences. Biographical Supplement. Free Press, New York, pp. 411–427.

Sills, D.L., 1987. Paul F. Lazarsfeld 1901–1976. A biographical memoir. Biographical Memoirs 56, 251–282.

Simonson, P. (Ed.), 2006. Politics, Social Networks, and the History of Mass Communications Research: Rereading Personal Influence, vol. 608. The Annals of the American Academy of Political and Social Science.

Stehr, N., 1982. A conversation with Paul F. Lazarsfeld. American Sociologist 17 (3), 150–155.

Leader–Member Exchange Theory

Berrin Erdogan and Talya N Bauer, Portland State University, Portland, OR, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Leader–member exchange (LMX) theory is a relationship-based, dyadic theory of leadership. According to this theory, leadership resides in the quality of the exchange relationship developed between leaders and their followers. High quality exchanges are characterized by trust, liking, and mutual respect, and the nature of the relationship quality has implications for job-related well-being and effectiveness of employees. This article summarizes what LMX is, how it is measured, and the best practices in the study of LMX. Studies conducted on its antecedents, and past research related to its consequences in organizational settings are reviewed. In addition, we review issues related to multiple relationships such as differentiation and relative LMX. Finally, we conclude the article by identifying and summarizing key themes and questions for future research.

Foundations of Leader–Member Exchange Theory

Leader–member exchange (LMX) theory is a relationship-based, dyadic theory of leadership. Unlike behavioral leadership theories that focus on what leaders do, such as transformational, authentic, servant, or empowering leadership theories, LMX theory rests firmly on the assumption that leaders influence employees in their group (referred to as members) through the quality of the relationships they develop with them. A high quality relationship is characterized by trust, liking, professional respect, and loyalty (Liden and Maslyn, 1998). One of the early findings of the LMX theory is that, leaders develop relationships of varying quality with their subordinates and such differentiation characterizes a wide majority of the work groups studied (Liden and Graen, 1980). The theory inspired and served as the basis of over 600 journal articles as of December 2013, with its popularity in leadership studies steadily increasing over the years.

Researchers tend to rely on a social exchange-based rationale to explain the nature of the relationship between LMX quality and its outcomes. High-quality relationships are characterized by the exchange of valued resources. In these relationships, leaders provide support, developmental opportunities, mentoring, and other benefits to the employee. The provision of such resources results in a motivation to reciprocate to the leader on the part of members, by demonstrating behaviors such as loyalty and higher levels of voluntary behaviors. In other words, feeling obligation and high levels of commitment to the supervisor are often thought of as the link between high LMX quality and promanager and sometimes proorganizational behaviors. Furthermore, the degree to which the employees regard their leader's promises to be fulfilled is a link between LMX quality and outcomes. For example, Dulac et al. (2008) showed that psychological contract violation was a mediator of the relationship between LMX quality and outcomes including trust and turnover intentions.

When the nomological network of leader behaviors (such as transformational or ethical leadership) and outcomes (such as commitment and effectiveness) is considered, LMX quality is thought of as a mediator (e.g., Walumbwa et al., 2011; Wang et al., 2005). This is because the manner in which leaders behave toward members is an indicator of how supportive, trustworthy, and loyal the leader is toward members, which feeds into the relationship quality positively or negatively. At the same time, studies linking leader behaviors to LMX quality tend not to adopt a longitudinal design, nor involve newly forming relationships. Thus, it could alternatively be argued that those in a high-quality exchange will perceive their leaders as more transformational, ethical, authentic, and less abusive and undermining.

Measurement of Exchange Quality

Most empirical research on LMX theory utilizes one of two measures. The LMX-7 measure (Scandura and Graen, 1984) is unidimensional and consists of seven items. The LMX-Multidimensional (Liden and Maslyn, 1998) consists of 12 items capturing four dimensions with three items each (affect, loyalty, contribution, and professional respect). The multidimensional measure consists of highly correlated dimensions falling under a second-order factor and many researchers choose to collapse the dimensions. This choice is typically driven by the research question at hand. Meta-analytic results suggest that type of measure is not an influence over observed relationships (Dulebohn et al., 2012).

Another common tendency within LMX studies is to measure LMX quality from the perspective of the employee. According to a review article by Hiller et al. (2011), in the period between 1985 and 2009, 83% of all LMX studies captured the employee perspective. When LMX is measured from the perspective of members *and* leaders, correlations tend to be modest. Sin et al. (2009) reported a meta-analytic correlation of 0.37 between the two. Furthermore, the correlation tends to be smaller during the early stages of relationship development, and the overlap increases as time goes by (Nahrgang et al., 2009). The lack of agreement could be explained by a number of different mechanisms. First, when employees and managers are each asked the degree to which they like, respect, and feel loyalty to the other party, it is natural that the way each party regards the other could be different. In other words, parties to the exchange are not necessarily reporting an objective reality: Instead they are reporting their thoughts about the other as an exchange partner. Because these

perceptions will influence their own attitudes and behaviors, they are each valid for the respondent. Second, each person may have different levels of success in meeting the other's expectations in their dyad. Third, leaders may be less likely to report a low-quality exchange with a team member, given social desirability concerns. Finally, the measurement approach may be responsible for some of the low observed correlations. If employees are asked to report the degree to which their manager likes, respects, and values them as opposed to the degree to which they like, respect, and value their manager, the relationship may emerge as stronger. Because of the divergence in perspectives, it seems that the employee perspective emerged as the *de facto* norm in measuring LMX quality in empirical studies. As a result, we know more about the consequences of employee perceptions of LMX quality as opposed to manager perceptions. Investigations into the implications of convergence between employee and manager perspectives are still in their infancy.

Research methodology tends to make a difference in LMX studies. Major studies of LMX tend to involve collecting data from intact teams where each member reporting to the same manager is invited to participate in the study. When researchers reach out to managers and members independently of each other, every member and leader is given the opportunity to participate in the study. In contrast, in studies using ad hoc sampling methods, researchers invite employees to participate in the study and employees are asked to give a corresponding survey to their managers. Alternatively, some studies rely on asking managers to distribute surveys to a few of the employees reporting to them. Meta-analytic evidence shows that in studies utilizing ad hoc sampling, the convergence between employee and manager perspectives of LMX quality is higher (Sin et al., 2009). Sampling bias is a possible explanation, and is, in our view, a weaker design, as it cannot rule out faking of data. At the very least, in such designs surveys are more likely to be distributed to those who have a higher quality exchange with the leader.

Development of High-Quality Relationships: LMX Antecedents

Figure 1 summarizes the antecedents and consequences we will cover in this article. While there are potentially additional constructs we could have included, we chose to focus on the most commonly studied and most potentially impactful constructs we uncovered in our literature review. One such question is: *How are high-quality exchanges developed?* Answering this question effectively requires studying leader–member dyads as they are in the formation stage of the relationship. Such studies have been rare, but informative. As an early example of such research, Liden et al. (1993) studied 166 new university employees and their supervisors at four time periods. This study showed that contrary to researchers' expectations, demographic similarity or leader-rated performance had little influence over LMX quality. Instead, the important predictors emerged to be perceived similarity, member expectations of the leader, and leader expectations of the member. In other words, similarity–attraction phenomenon seems to play an important role in LMX development.

'Testing' of the exchange partner is thought to be at the heart of LMX development. When the relationship begins, trust develops (or does not develop) as a result of a mutual testing process (Dienesch and Liden, 1986). Employee's satisfactory responses to the testing efforts of leaders result in the development of trust on the part of the leader. Bauer and Green (1996) examined this process in a sample of new employees and managers. Graduating students at a university filled out surveys, once before, and twice after the start of their new jobs. This study showed that early positive performance of employees was reciprocated with further delegation of responsibilities on the part of managers that resulted in higher levels of performance. The result of this trust development process was high LMX quality. In other words, member performance and competence appear to be important predictors in the LMX development process as these are helpful behaviors in establishing trustworthiness.

In a more recent study, Nahrgang et al. (2009) examined MBA student led undergraduate teams, surveying them multiple times in the relationship development process. This study showed that early predictors of LMX quality were different from later predictors. In the earliest stages, member personality traits (extraversion and agreeableness) were predictors of LMX. In later stages, member performance emerged as a more important predictor. It seems that factors that are more influential in the first few days are different from factors that are more influential as weeks pass and members

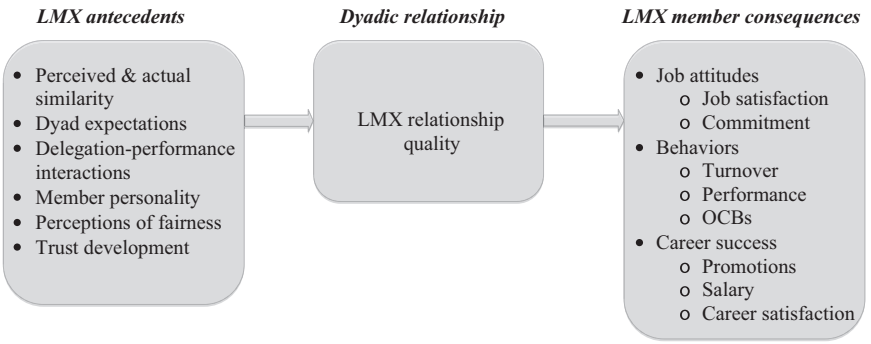

Figure 1 Summary of the antecedents and consequences of LMX relationships. LMX, leader–member exchange; OCBs, organizational citizenship behaviors.

have the opportunity to display their competence and effectiveness.

In addition to member performance and similarity to leaders, member personality has been frequently examined as a predictor in cross-sectional work. Meta-analytic results (Dulebohn et al., 2012) revealed that following member competence and perceived similarity, member positive affectivity and the locus of control are the characteristics with the strongest correlations to LMX quality. Furthermore, goal orientation has been explored as an antecedent. Mastery orientation, which refers to the degree to which a person is interested in acquiring new skills, improving and learning, has been shown to be positively related to LMX quality, whereas performance orientation, which refers to the degree to which the person is preoccupied with looking like a high performer and being evaluated well, has been negatively related (Janssen and van Yperen, 2004).

Leader–member actual similarity (as opposed to perceived similarity) has been explored as a predictor, and only recently researchers started identifying patterns of relationships. Zhang et al. (2012) showed, using polynomial regression, that leader and member similarity in proactive personality was related to LMX quality. Furthermore, the direction of the dissimilarity also mattered: When leaders had higher proactive personality compared to members, dissimilarity was much more detrimental.

While performance and competence are important currencies of exchange helping build trust, newcomers react positively to manager behaviors that communicate manager support of the newcomer. For example, in a study of newcomers in telemarketing organizations, Sluss and Thompson (2012) found that how supervisors socialized newcomers in the form of providing advice, role modeling, and guidance were associated with higher LMX quality.

Among leader behaviors associated with LMX quality, perhaps the most consistent is leader fairness. Masterson et al. (2000) showed that interactional justice, or a sense of treatment with dignity and respect was associated with LMX quality. This study and others (e.g., Erdogan et al., 2006) contended that leader fairness would relate to LMX quality whereas organizational fairness such as procedural justice would relate to the quality of the exchange relationship with the organization in general. Recent meta-analytic evidence is not supportive of this distinction: Colquitt et al. (2013) showed that supervisor-focused and organization-focused justice perceptions did not differ in their strength of relationship with LMX quality. In other words, even when researchers make an effort to measure organizational justice perceptions, it is likely that supervisors are regarded as at least partially responsible for the observed state of affairs, resulting in higher LMX.

Once established, how stable are LMX relationships? This is an underinvestigated but important research question. Longitudinal research (e.g., Liden et al., 1993) indicates that LMX relationships start being developed very early in the relationship (within the first days and weeks) and show remarkable stability (Liden et al., 1993). One potential reason for this is that once a relationship is categorized as high or low trust, individuals interpret later events and actions using that lens and the relationship does not change dramatically. Shapiro et al. (2011) explored the effects of leader transgressions on subsequent relationship quality and the likelihood of members to engage in punitive actions following such transgressions. The results showed that leaders in high-quality exchanges with their members had substantial levels of idiosyncrasy credits, leading to lesser likelihood of punitive action toward leaders who engage in potentially trust-violating behaviors. High-quality relations may afford dyad members higher capacity to forgive their exchange partner.

Consequences of High-Quality Relationships

Job attitudes are among the most frequently studied outcomes of LMX quality. Meta-analytic evidence (Dulebohn et al., 2012; Gerstner and Day, 1997) showed that job satisfaction, organizational commitment, and turnover intentions are consistently correlated with LMX quality. The relationship between LMX quality and actual turnover has been more inconsistent, with simple correlations close to zero (Gerstner and Day, 1997). A contextual approach has proven helpful in understanding this relationship. For example, Bauer et al. (2006) showed that in a sample of new high-level pharmaceutical executives, the LMX to turnover relationship was contingent on the extraversion level of the new executive, with LMX mattering only for the turnover behavior of employees who were relatively lower in extraversion.

High LMX members also demonstrate more desirable behaviors. The relationship between LMX and behaviors such as task performance and citizenship behaviors has been well documented (Dulebohn et al., 2012; Gerstner and Day, 1997; Ilies et al., 2007). Furthermore, high LMX members demonstrate behaviors that are desirable within the particular organizational context. For example, Hofmann et al. (2003) showed that LMX quality was positively related to safety citizenship behaviors (such as being proactive about removing safety hazards) when safety climate in the organization is high. In other words, high LMX members tend to 'fit the mold' and demonstrate behaviors that are desirable within that context.

LMX quality has been associated with career success. Wayne et al. (1999) showed relations with supervisor ratings of promotability, actual salary progression, and career satisfaction for those who have a high-quality LMX. High LMX members tend to be more successful. This is probably due to at least two mechanisms. High LMX members tend to have higher levels of performance ratings, and given that these ratings are obtained from supervisors, such a relationship is not surprising. Furthermore, high LMX members may be 'sponsored' to a greater degree by their managers. By introducing the member to influential others, giving guidance about how to advance, or informing high LMX members of impactful and visible opportunities within the organization, managers may play an important role in the career success of high LMX members. Of course, such sponsorship is also contingent on the degree to which the leader is in a position to provide these benefits. For example, Sparrowe and Liden (2005) showed that LMX quality was positively related to member emergence as a central actor in the work group advice network. However, this relationship was contingent on the degree to which the leader was central in the advice network among leaders.

LMX quality has effects that go beyond the leader–member dyad. For example, researchers have explored the relationship between the leader's LMX quality with his/her own boss, and its implications for the employees the leader supervises. The relationship between leaders and their own bosses is referred to as the leader–leader exchange (LLX) and it has been demonstrated that leaders with high LLX are in a better position to develop high-quality exchanges with their subordinates (Venkataramani et al., 2010). Leaders with high LLX have also been shown to provide higher levels of empowerment for their own members (Zhou et al., 2012). This could be a result of learning where leaders observe how to build effective relations from their own managers and model the same behaviors to their subordinates, or it could be because high LLX puts managers in a better position to provide valuable resources to their own lower-level exchange partners.

When Do High-Quality Relationships Matter the Most?

LMX seems to be more important, more influential, and more impactful for some members as opposed to others. Factors that increase the importance of LMX quality include introversion (Bauer et al., 2006), and low levels of person–organization fit or value congruence (Erdogan et al., 2004). These moderators suggest that some employees may need and value the benefits associated with high-quality LMXs more than others. The absence of high LMX quality can also be dealt with through structural mechanisms. For example, one of the key benefits associated with high LMX is high empowerment (Dulebohn et al., 2012). Perhaps for this reason, research examining negotiated special deals between employees and organizations (termed idiosyncratic deals, or i-deals) has shown that such deals were most strongly related to desirable organizational behaviors for employees with low LMX quality. This is likely because high LMX members are given more flexibility and autonomy, and structural arrangements make more difference for employees who are not in a position to receive these benefits.

Furthermore, some leaders are simply more powerful, more resourceful, and better positioned, allowing them to provide more meaningful benefits to their high LMX members. In other words, not all LMX relationships are created equal because no two dyadic partners are the same. When people feel that the leader embodies the organization (due to being in a high-level position or being in a position of authority), LMX quality is more positively related to attachment to the organization (Eisenberger et al., 2010). Similarly, when managers perceive greater levels of support from the organization (Erdogan and Enders, 2007), the effects of LMX quality for employee job attitudes and performance are greater.

Is LMX quality more important in some cultures compared to others? Studies of LMX have been conducted around the world, including collectivistic and power-distant cultures. Rockstuhl et al. (2012) conducted a meta-analysis comparing the pattern of relations between LMX, job attitudes, and behaviors around the world. Their results suggested that in horizontal-individualistic (i.e., Western) cultures, the relations between LMX, citizenship behaviors, trust in leader, and justice perceptions were stronger compared to that in vertical-collectivist (e.g., Asian) cultures. At the same time, there were no significant differences among correlations between LMX, commitment, and job performance. Even when such differences existed, they were modest. For example, the largest difference in correlations was between LMX and distributive justice, which was 0.46 vs 0.30. While the difference is statically significant, even in collectivistic cultures LMX seems to be meaningfully associated with and explain substantial variation in employee attitudes and behaviors.

Emerging Issues: LMX Differentiation and Relative LMX

In the recent years, the focus of LMX theory started including the implications of differentiated exchanges. In fact, the distinctiveness of LMX theory from other leader behavior theories resides in its recognition that leadership varies across dyads, and within the same work group there will be variation in the quality of exchanges managers develop with their different employees. However, the majority of LMX studies in the past four decades focused on understanding antecedents, consequences, and moderators of these relationships at the individual level.

Recently, LMX differentiation started becoming the focus of LMX studies. LMX differentiation refers to the degree to which the quality of LMXs within the work group varies. Oftentimes, LMX differentiation is operationalized using the standard deviation of LMXs within the work group, with the provision that the response rate within the group is high (60% is a frequently used cutoff point for response rates; for example, Liden et al., 2006). Liden et al. (2006) showed that LMX differentiation is positively related to group performance to the degree to which members report high task interdependence and their LMX median in the group is low. The argument Liden et al. made was that LMX differentiation may be beneficial to members given that it will ensure an appropriate role differentiation. At the same time, subsequently scholars recognized the potential downsides of LMX differentiation in the form of poorer-quality job attitudes, higher levels of conflict, and lower well-being (Hooper and Martin, 2008).

The importance of LMX differentiation studies is the understanding that employees are aware of each others' LMX quality with the manager. Further, dissimilarity in LMX quality within the work group has negative consequences for members. For example, Sherony and Green (2002) showed that among pairs of coworkers, the interaction of the two LMXs was related to the quality of the relationship between coworkers. In other words, when LMX quality of one coworker diverged from the other, coworker–coworker exchanges were negatively affected. By the same token, variation within the work group is thought to be associated with negative outcomes for members as well as groups, because LMX differentiation tends to indicate that some members do not have a high-quality exchange with the leader, potentially resulting in envy, differences in resource distributions, even potential for feelings of favoritism. Low LMX differentiation is a more inclusive leadership style (provided that the mean level of LMX is high within the group) and such inclusiveness contributes to retention, particularly in diverse work groups (Nishii and Mayer, 2009).

It is important to note that the presence of LMX differentiation increases the salience of fairness concerns. Erdogan and Bauer (2010) showed that the effects of LMX differentiation on work attitudes, coworker relations, and retention within the work group were contingent on the fairness climate that existed in the group such that, LMX differentiation had negative effects on outcomes only when fairness climate was low.

Studies on LMX differentiation highlight the importance of employee awareness of how LMXs are distributed within the group. Employees react not only to their own relationship quality, but also to their coworkers' relationship quality, and distribution of LMXs matter. Taking this idea a step further, scholars also started investigating social comparison processes directly, by introducing the concept of relative LMX. Relative LMX is a statistical computation of the degree to which a person's LMX quality is higher or lower than the team's LMX average. Utilizing a social comparison approach, scholars argued that having a higher-quality exchange compared to one's team members is a source of satisfaction. Controlling for one's LMX quality, relative LMX is positively related to self-efficacy, performance, citizenship behaviors (Hu and Liden, 2013), and psychological contract fulfillment (Henderson et al., 2008).

In addition to examining relative LMX operationalized as the difference between focal person's LMX score from the group mean, researchers developed a perceptual measure of relative LMX, directly asking individuals to compare their own relationship quality to the other relations the leader develops with team members. Vidyarthi et al. (2010) found that this perception (labeled as LMX social comparisons) explained variation in performance and citizenship behaviors controlling for LMX and relative LMX, again indicating the importance of LMX distribution on member behaviors.

It seems that such interpersonal comparisons among coworkers necessitate explicit attention to the context of a focal LMX relationship. Other LMXs, including a leader's LMX quality with a superior, one's coworkers' LMX, the nature of LMX distribution within the work group and how much diversity exists in LMX quality, how one's LMX compares to one's team members (in actuality and perceptually) are all increasingly well studied constructs gaining research attention. We predict that these areas will witness further research activity.

Future Research Opportunities and Directions

Investigations of the nomological network of LMX quality are numerous and this is a mature field of investigation. Yet, there are still research avenues that are important to investigate. One issue is the evolving nature of organizations. LMX theory originated in the 1970s, at a time organizations were characterized by tall hierarchies, unity of command, and authority concentrated more at the top. Today, while such organizations continue to exist, there are more novel and contemporary structures under which managers and employees develop relationships. For example, in many contemporary organizations, employees may report to more than one manager, whereas LMX theory is based on the assumption that each member has one, clearly identifiable manager who controls resources valued by the member.

Recent research (Vidyarthi et al., 2014) has examined LMX relationships in such a context and showed that in a sample of information technology consultants reporting to two managers, convergence of the quality of these relationships was associated with more positive outcomes. The authors contended that each LMX relationship would serve as a comparison point for the other relationship, evoking social comparison processes. In other words, similar to the comparisons employees engage in with their coworkers' LMXs, it seems that they also compare the multiple exchanges they have with different leaders in their work lives. Such findings indicate that LMX theory would benefit from an extension and testing of the theory in settings that are different from the traditional organizational forms. As organizations introduce matrix structures where members report to multiple leaders for finite periods of time, or when they eliminate managers by introducing lattice organizations where there are no assigned leaders, the utility of the theory remains unclear. Extension of LMX theory to contemporary organizational structures is an important future direction.

A second research direction relates to an examination of LMX quality in relation to coworker relationships. We know that LMX quality is associated with positive job attitudes and behaviors. However, we know significantly less about when and why coworkers experience envy or jealousy, or feel negatively toward high LMX members. Are high LMX members more sought after among coworkers because they are a liaison to the manager? Or are they shunned and avoided because they are regarded as teacher's pets? When does each happen? Erdogan et al. (in press) showed that the degree to which LMX quality is positively associated with one aspect of coworker relations (advice network centrality) is contingent on the degree to which high LMX members also demonstrate helpfulness and discretion. Systematic investigation of the effects of LMX quality on coworker emotions, behaviors, and reactions to the focal person is a noteworthy area of research. The importance of this topic is also evidenced by the fact that the most recent meta-analysis in LMX theory (Dulebohn et al., 2012) includes relationships of LMX to a large number of outcomes, but any indicators of coworker relationships is curiously missing, which likely indicates the small number of studies examining LMX quality in relation to coworker relations.

Finally, an interesting, but underinvestigated, research topic is whether high LMX quality is uniformly desirable or whether it may have a dark side for the dyad members or the organization. The vast majority of the literature focuses on how high LMX quality contributes to individually and organizationally desirable outcomes such as job attitudes, behaviors, and work-related well-being. Yet, is it possible that there are potential downsides to having a high-quality exchange with a manager? Under what situations such negative repercussions would emerge? The only example of a potential 'dark side' to LMX was revealed by Ballinger et al. (2010). In their investigation of employees working in veterinary hospitals, they showed that while LMX quality was negatively related to actual turnover, following leader turnover high LMX members demonstrated increased probability to leave. In other words, one interesting and important downside of high LMX relationships is that when they leave, they may take those close to them with them.

Identifying and understanding any potential downsides to LMX quality remain important to fully understand how this theory can be utilized to increase leadership effectiveness in the workplace.

Conclusion

In conclusion, while research on LMX has entered a mature phase where much is known about its measurement, antecedents, boundary conditions, and consequences, much also remains left to uncover. In this article, we summarized what is known but also highlighted avenues for future research. These include future understanding of how LMX relationships develop and the boundary conditions for relationship development, how LMX is measured, how relative LMX affects what we know, as well as the key future research themes of the changing nature of work in terms of content and organizational structures, the influence of the social network of relationships, as well as the dark side of LMX. Given these and other potential research questions, we are excited about the future of LMX as a vibrant research area.

See also: Authentic Leadership; Job Satisfaction; Organizational Commitment; Psychological Contracts; Supervision, Abusive.

Bibliography

Ballinger, G.A., Lehman, D.W., Schoorman, F.D., 2010. Leader-member exchange and turnover before and after succession events. Organizational Behavior and Human Decision Processes 113, 25–36.

Bauer, T.N., Erdogan, B., Liden, R.C., Wayne, S.J., 2006. A longitudinal study of the moderating role of extraversion: leader-member exchange, performance, and turnover during new executive development. Journal of Applied Psychology 91, 298–310.

Bauer, T.N., Green, S.G., 1996. Development of leader-member exchange: a longitudinal test. Academy of Management Journal 39, 1538–1567.

Colquitt, J.A., Scott, B.A., Rodell, J.B., Long, D.M., Zapata, C.P., Conlon, D.E., Wesson, M., 2013. Justice at the millennium, a decade later: a meta-analytic test of social exchange and affect-based perspectives. Journal of Applied Psychology 98, 199–236.

Dienesch, R.M., Liden, R.C., 1986. Leader-member exchange model of leadership: a critique and further development. Academy of Management Review 11, 618–634.

Dulac, T., Coyle-Shapiro, J.A.M., Henderson, D.J., Wayne, S.J., 2008. Not all responses to breach are the same: the interconnection of social exchange and psychological contract processes in organization. Academy of Management Journal 51, 1079–1098.

Dulebohn, J.H., Bommer, W.H., Liden, R.C., Brouer, R.L., Ferris, G.R., 2012. A meta-analysis of antecedents and consequences of leader-member exchange: integrating the past with an eye toward the future. Journal of Management 38, 1715–1759.

Eisenberger, R., Karagonlar, G., Stinglhamber, F., Neves, P., Becker, T.E., Gonzalez-Morales, M.G., Steiger-Mueller, M., 2010. Leader-member exchange and affective organizational commitment: the contribution of supervisor's organizational embodiment. Journal of Applied Psychology 95, 1085–1103.

Erdogan, B., Bauer, T.N., 2010. Differentiated leader-member exchanges (LMX): the buffering role of justice climate. Journal of Applied Psychology 95, 1104–1120.

Erdogan, B., Bauer, T.N., Walter, J. Deeds that help and words that hurt: helping and gossip as moderators of the relationship between LMX and advice network centrality. Personnel Psychology, in press.

Erdogan, B., Enders, J., 2007. Support from the top: supervisors' perceived organizational support as a moderator of leader-member exchange to satisfaction and performance relationships. Journal of Applied Psychology 92, 321–330.

Erdogan, B., Kraimer, M.L., Liden, R.C., 2004. Work value congruence and intrinsic career success: the compensatory roles of leader-member exchange and perceived organizational support. Personnel Psychology 57, 305–332.

Erdogan, B., Liden, R.C., Kraimer, M.L., 2006. Justice and leader-member exchange: the moderating role of organizational culture. Academy of Management Journal 49, 395–406.

Gerstner, C.R., Day, D.V., 1997. Meta-analytic review of leader–member exchange theory: correlates and construct issues. Journal of Applied Psychology 82, 827–844.

Henderson, D.J., Wayne, S.J., Shore, L.M., Bommer, W.H., Tetrick, L.E., 2008. Leader-member exchange: differentiation, and psychological contract fulfillment: a multilevel examination. Journal of Applied Psychology 93, 1208–1219.

Hiller, N.J., DeChurch, L.A., Murase, T., Doty, D., 2011. Searching for outcomes of leadership: a 25-year review. Journal of Management 37, 1137–1177.

Hofmann, D.A., Morgeson, F.P., Gerras, S.J., 2003. Climate as a moderator of the relationship between leader-member exchange and content specific citizenship: safety climate as an exemplar. Journal of Applied Psychology 88, 170–178.

Hooper, D.T., Martin, R., 2008. Beyond personal leader-member exchange (LMX) quality: the effects of perceived LMX variability on employee reactions. Leadership Quarterly 19, 20–30.

Hu, J., Liden, R.C., 2013. Relative leader-member exchange within team contexts: how and when social comparison impacts individual effectiveness. Personnel Psychology 66, 127–172.

Ilies, R., Nahrgang, J.D., Morgeson, F.P., 2007. Leader-member exchange and citizenship behaviors: a meta-analysis. Journal of Applied Psychology 92, 269–277.

Janssen, O., van Yperen, N.W., 2004. Employees' goal orientations, the quality of leader-member exchange, and the outcomes of job performance and job satisfaction. Academy of Management Journal 47, 368–384.

Liden, R.C., Erdogan, B., Wayne, S.J., Sparrowe, R.T., 2006. Leader-member exchange, differentiation, and task interdependence: implications for individual and group performance. Journal of Organizational Behavior 27, 723–746.

Liden, R.C., Graen, G., 1980. Generalizability of the vertical dyad linkage model of leadership. Academy of Management Journal 23, 451–465.

Liden, R.C., Maslyn, J.M., 1998. Multidimensionality of leader-member exchange: an empirical assessment through scale development. Journal of Management 24, 43–72.

Liden, R.C., Wayne, S.J., Stilwell, D., 1993. A longitudinal study on the early development of leader-member exchanges. Journal of Applied Psychology 78, 662–674.

Masterson, S.S., Lewis, K., Goldman, B.M., Taylor, M.S., 2000. Integrating justice and social exchange: the differing effects of fair procedures and treatment on work relationships. Academy of Management Journal 43, 738–748.

Nahrgang, J.D., Morgeson, F.P., Ilies, R., 2009. The development of leader-member exchanges: exploring how personality and performance influence leader and member relationships over time. Organizational Behavior and Human Decision Processes 108, 256–266.

Nishii, L.H., Mayer, D.M., 2009. Do inclusive leaders help to reduce turnover in diverse groups? The moderating role of leader-member exchange in the diversity to turnover relationship. Journal of Applied Psychology 94, 1412–1426.

Rockstuhl, T., Dulebohn, J.H., Ang, S., Shore, L.M., 2012. Leader-member exchange (LMX) and culture: a meta-analysis of correlates of LMX across 23 countries. Journal of Applied Psychology 97, 1097–1130.

Scandura, T.A., Graen, G.B., 1984. Moderating effects of initial leader-member exchange status on the effects of a leadership intervention. Journal of Applied Psychology 69, 428–436.

Shapiro, D.L., Boss, A.D., Salas, S., Tangirala, S., Von Glinow, M.A., 2011. When are transgressing leaders punitively judged? An empirical test. Journal of Applied Psychology 96, 412–422.

Sherony, K.M., Green, S.G., 2002. Coworker exchange: relationships between coworkers, leader-member exchange, and work attitudes. Journal of Applied Psychology 87, 542–548.

Sin, H.P., Nahrgang, J.D., Morgeson, F.P., 2009. Understanding why they don't see eye to eye: an examination of leader-member exchange (LMX) agreement. Journal of Applied Psychology 94, 1048–1057.

Sluss, D.M., Thompson, B.S., 2012. Socializing the newcomer: the mediating role of leader-member exchange. Organizational Behavior and Human Decision Processes 119, 114–125.

Sparrowe, R.T., Liden, R.C., 2005. Two routes to influence: integrating leader-member exchange and network perspectives. Administrative Science Quarterly 50, 505–535.

Venkataramani, V., Green, S.G., Schleicher, D.J., 2010. Well-connected leaders: the impact of leaders' social network ties on LMX and members' work attitudes. Journal of Applied Psychology 95, 1071–1084.

Vidyarthi, P., Erdogan, B., Anand, S., Liden, R.C., Chaudhry, A., 2014. One member, two leaders: extending leader-member exchange theory to a dual leadership context. Journal of Applied Psychology 99, 468–483.

Vidyarthi, P.R., Liden, R.C., Anand, S., Erdogan, B., Ghosh, S., 2010. Where do I stand? Examining the effects of leader-member exchange social comparison on employee work behaviors. Journal of Applied Psychology 95, 849–861.

Walumbwa, F.O., Mayer, D.M., Wang, P., Wang, H., Workman, K., Christensen, A.L., 2011. Linking ethical leadership to employee performance: the roles of leader-member exchange, self-efficacy, and organizational identification. Organizational Behavior and Human Decision Processes 115, 204–213.

Wang, H., Law, K.S., Hackett, R.D., Wang, D., Chen, Z.X., 2005. Leader-member exchange as a mediator of the relationship between transformational leadership and followers' performance and organizational citizenship behavior. Academy of Management Journal 48, 420–432.

Wayne, S.J., Liden, R.C., Kraimer, M.L., Graf, I.K., 1999. The role of human capital, motivation and supervisor sponsorship in predicting career success. Journal of Organizational Behavior 20, 577–595.

Zhang, Z., Wang, M., Shi, J., 2012. Leader-follower congruence in proactive personality and work outcomes: the mediating role of leader-member exchange. Academy of Management Journal 55, 111–130.

Zhou, L., Wang, M., Chen, G., Shi, J., 2012. Supervisors' upward exchange relationships and subordinate outcomes: testing the multilevel mediation role of empowerment. Journal of Applied Psychology 97, 668–680.

Leadership

S Alexander Haslam, School of Psychology, University of Queensland, QLD, Australia
Stephen D Reicher, School of Psychology, University of St. Andrews, Scotland, UK
Michael J Platow, Research School of Psychology, Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Leadership is the process through which one or more people influence other group members in a way that motivates them to contribute to the achievement of group goals. As such, it is a group process that is demonstrated through the followership of others. Over time, the field of leadership research has moved through several distinct phases. Initially, a classical approach viewed leadership as a reflection of particular individuals' special qualities. This was then supplanted by a contextual approach which supplemented this analysis with considerations of social and organizational context. More recently, an identity approach has emphasized the importance of the psychological bond between leaders and followers. This progression represents a response to four key observations about the leadership process: (1) different forms of leadership are required in different contexts, (2) followers' perceptions of leaders are critical, (3) leaders and followers motivate and influence each other, and (4) in the process of developing the power to change the world, leaders and followers are themselves transformed.

Leadership is the process of influencing people in a way that motivates them to contribute to the achievement of group goals (e.g., Haslam et al., 2011; Hollander, 1985). As such, it is a central feature of effective organizations and societies, and the focus of intense academic and public debate for over 2000 years. Testament to this, the British Library in London holds over 12 000 books that have leadership in their title (including 35 that are simply entitled *Leadership*). This interest arises from the fact that in fields as diverse as politics and religion, science and technology, art and literature, sport and adventure, industry and business, leadership is commonly seen as the process by which means people are marshaled to contribute to the collective projects that ultimately become the stuff of history.

Definitional Issues

The above definition of leadership contains at least four important elements that are important to recognize. First, leadership is a process not a property. Accordingly, it is not something that a person possesses, but rather something that he or she *does*. Second, leadership is not something that a person does on his or her own. It necessarily involves *other* people – most particularly, those over whom influence is exerted. Third, it follows from this that ultimate proof of leadership is not found in the actions of leaders (e.g., their speeches, their policy, their vision), but in the *followership* of those they influence (Gibb, 1947; Haslam et al., 2011; Hollander, 1995). Fourth, it is important to distinguish leadership from a range of other processes with which it is commonly confused. Leadership is not primarily a matter of power, coercion, or resource management (Turner, 2005). Rather, because it is about influence, it is about winning others over so that they *want to do* what is being asked of them and do so willingly not grudgingly, with enthusiasm rather than rancor.

Bearing these points in mind, the literature on leadership is primarily concerned with the question of what it is that allows the plans and vision of an individual to be translated into the actions of a group. What is it that turns one person's wishes into a mission that directs the work of dozens, thousands, or millions of other people?

As we argue in our recent book, *The New Psychology of Leadership* (Haslam et al., 2011), over time researchers have tended to answer this question in one of three broad ways. Those who follow a *classical approach* generally frame their answers in terms of the qualities that particular individuals have (or do not have). Building upon this, adherents of a *contextual approach* tend to supplement such analysis with a consideration of features of social and organizational context that promote (or else compromise) the effectiveness of individual leaders. And finally proponents of an *identity* approach see leadership as a group process that centers on a social–psychological bond between leaders and followers. While the first of these approaches can be considered 'old,' the second is more 'contemporary,' and the third is relatively 'new' (Haslam et al., 2011).

The Old Psychology of Leadership: A Classical Approach

According to Plato (380BC/1993), leaders are those small number of people who are born with a cluster of traits that set them apart from the general population – for example, quickness of learning, courage, and broadness of vision. Although it was largely discursive and anecdotal, Plato's analysis provided a narrative framework that has informed most leadership research in the subsequent two-and-a-half millennia – a literature that can be sampled liberally in the popular texts that proliferate in airport bookstores. Its popularity, however, was cemented in the nineteenth century through the writings of the Scottish essayist Thomas Carlyle who declared "the history of what man has accomplished in this world, is at bottom the History of the Great Men who have worked here" (1840: p. 5). Carlyle's 'great man' analysis

focused his readers' attention firmly on the psychology of individual (male) leaders and argued that their distinctive and exceptional qualities qualified them not only for responsibility and high office, but also for widespread admiration and respect. More recently, it is also the exceptional nature of such 'stuff' that is seen to justify the exorbitant salaries that are routinely awarded to executive leaders.

What, though, is the stuff of leadership? In an early influential review, Stogdill (1948) identified 29 personality dimensions that various researchers had associated with successful leadership. Within these, he singled out five factors for special attention: capacity, achievement, responsibility, participation, and status. In a later review, Mann (1959) looked at the predictive power of over 500 different personality measures that he organized into seven meaningful clusters: intelligence, adjustment, extroversion, sensitivity, masculinity, conservatism, and dominance. The primary observation in both reviews was that personal attributes are generally rather unreliable predictors of leadership. Nevertheless, two attributes have proved to have enduring appeal for researchers and commentators alike – charisma and intelligence.

Leadership as Charisma

Although it was an important concept in earlier writings (particularly, early Christian commentaries; Esler, 2003), the term charisma is generally seen to have been popularized by the German social and economic theorist Max Weber. For Weber (1921/1946) *charisma* was "a certain quality of an individual personality by which [a leader] is set apart from ordinary men and treated as endowed with superhuman, or at least specifically exceptional powers or qualities," (p. 359). Inspired by this, in recent years, neo-Weberians have treated charisma as an aspect of leadership that hinges on an individual's capacity to articulate a group vision, to recruit others to his or her cause, and to develop close and strong relationships with group members (e.g., Bass and Riggio, 2006; Burns, 1978).

Consistent with this analysis, research provides fairly consistent evidence that leaders who are both visionary and empathic are more successful than those who are not. In particular, findings from major studies in the 1940s showed that leaders who *initiate structure* (e.g., by devising plans that flesh out their vision) and show *consideration* toward followers tend to be more successful than those who do not (Fleishman and Peters, 1962). Research in this tradition also indicates, however, that followers' *perceptions* are critical in this process. This speaks to Weber's observation that it is followers who confer charisma upon the leader and who, on this basis, prove willing to commit their energies to his or her cause. At the same time, though, there is also evidence that such attributions vary considerably with features of the social context. Among other things, this is because evaluators seem generally to see leaders of in-groups to be far more charismatic than those who lead out-groups (Platow et al., 2006).

Leadership as Intelligence

Despite the fact that the construct of charisma therefore seems quite 'slippery,' one might imagine that – as a result of its proud psychometric heritage – *intelligence* would prove to deliver greater predictive traction. Indeed, a key reason why this has been an important research focus is that in most reviews this emerges as the best single predictor of leader success (e.g., Mann, 1959). However, measures of leader intelligence such as IQ typically account for only a very small amount of variance in leader success (around 3–5%), whereas again *perceived* intelligence accounts for far more (around 30%; Judge et al., 2004). Moreover, it also appears that analytic intelligence (of the form measured by IQ tests) is often less predictive of leaders' success than their creative intelligence and emotional intelligence (defined as "the ability to understand and manage moods and emotions in the self and others"; George, 2000). In the end, however, as with charisma, these various forms of intelligence prove extremely hard to pin down (Antonakis et al., 2009), and whether or not a person is perceived to have them varies markedly as a function of who is doing the perceiving and when. Accordingly, over time, such observations have led researchers to place increasing emphasis on the importance of *social context* in the leadership process.

The Contemporary Psychology of Leadership: A Contextual Approach

In response to the limited predictive power of approaches that focus only on the character of the leader in isolation, most contemporary researchers advocate *contextual* approaches to leadership that take account of the situations in which leaders operate. Radical approaches of this form suggest that context is everything, and that the character of the individual counts for nothing (e.g., as proposed by Zimbardo in his analysis of the Stanford Prison Experiment; Zimbardo, 2007). However, largely because these models envisage leaders as having little or no role to play in the leadership process, such analyses have garnered little support among leadership theorists and practitioners. It is therefore far more common for researchers to endorse *contingency models* in which context is seen to moderate but not entirely suppress the contribution of the individual leader.

Leadership as Contingency

The contingency approach is defined by the assertion that leadership arises from a perfect match between the individual and the circumstances of the group that he or she leads (Gibb, 1958). There are a very large number of such theories, but Fred Fiedler's (1964) *least preferred co-worker* (LPC) model is probably the most influential (at least in academic circles). This argues that leaders can be distinguished in terms of their disposition toward their LPC and that context can be differentiated in terms of (1) the quality of leader–member relations, (2) the degree to which the task is structured, and (3) the leader's position power. High-LPC leaders (those with a positive view of their LPC) are predicted to be the most effective in conditions where either (1) relations are good, structure is high, and the leader's position is strong, or (2) relations are bad, structure is low, and the leader's position is either weak or strong. Stated crudely, task-oriented, or 'hard' (i.e., low-LPC) leaders are predicted to do well when conditions are all very

favorable or all very unfavorable, but relationship-oriented or 'soft' (i.e., high-LPC) leaders are predicted to do well when conditions are more mixed.

Contingency theories vary considerably in the attributes of the leader and the characteristics of the situation that they consider important. Nonetheless, they remain the most widely embraced class of leadership theory (Fiedler and House, 1994). In particular, they are a popular component of management and personal development courses that typically seek first to classify individuals as having a particular leadership style and then to train them to identify (or create) situations in which this style will be effective.

Yet, despite being very popular, there are some serious empirical and theoretical problems with such models. In the first instance, they can be criticized for being reductionist in their attempts to distil a complex whole down to a limited subset of component parts that are seen as fixed and as having no capacity to shape each other. For example, these approaches do not allow for the possibility that the quality of leader–member relations might have a bearing on a leader's power (and vice versa) or that perceptions of task structure might be a determinant of leader–member relations (and vice versa). Such objections speak to the fact that contingency models typically treat the interaction between person and context as *mechanical* rather than *dynamic* (Reynolds et al., 2010).

A somewhat different problem is that while contingency models, like Fiedler's, recognize the importance of leaders' relationship with followers, those followers are themselves not subjected to any meaningful form of psychological analysis. Might not their perceptions of the leadership process differ from those of the leader, and might not this have a bearing on their followership? As with classical approaches, a core problem with contingency models is thus that they treat rank-and-file group members as passive consumers of leadership rather than active agents (Reicher et al., 2005). Beyond these theoretical concerns, however, an even more basic problem is that empirical support for such models is mixed at best – and tends to become weaker as tests move from the laboratory into the field.

Leadership as Transaction

Disenchantment with basic contingency models is reflected in two distinct theoretical movements that have grown in popularity in recent decades. The first of these is the *transactional approach*. This argues that any analysis of leadership needs to be complemented by an examination of the role that followers play in validating and empowering any leader. Most closely associated with the work of Edwin Hollander (1964), this approach argues that leadership emerges from a process of *social exchange* whereby followers work to enact a leader's commands only when they believe that the leader is doing something for them in return. More specifically, it is suggested that followers will only respond constructively to a leader's creative ideas if that leader has built up 'idiosyncrasy credits' based on demonstrated service to the group. As support for this analysis, research points to evidence that followers are more likely to support a leader if he or she is elected rather than appointed, and if he or she emerges from within the group rather than being imposed on it from outside.

Much of Hollander's own work has involved exploring the relevance of these ideas to leadership in the world at large (e.g., Hollander, 1995). In particular, he notes that large disparities in the benefits that leaders and followers receive for their work will tend to trigger a sense of inequity that ultimately detracts from group performance. Aside from an impressive array of empirical studies that bear this point out, this analysis can also be seen to provide a compelling explanation for failures of big business (e.g., as seen in the 2008 Global Financial Crisis) – where unfettered enthusiasm for the ideology of the 'great man' leads to overconfidence in the abilities of a few highly paid leaders and disregard for the legitimate concerns of followers.

More recently, though, empirical work on the transactional dimensions of leadership has also been taken forward by proponents of *leader–member exchange (LMX) theory* (e.g., Graen and Uhl-Bien, 1995). A key argument here is that low-quality LMX relationships (which center on standard contractual obligations) tend to lead to negative group outcomes, whereas high-quality relationships (in which both parties actively promote the interests of each other) enhance outcomes not only for those who share that relationship (e.g., the specific leader and follower) but also for the broader entity of which they are part (e.g., the group or organization as a whole).

A large body of empirical work is consistent with these arguments – showing that when leaders and followers are motivated to help each other out, this tends to have positive consequences both for them and their group (Ilies et al., 2007). Exchange-based approaches also have intuitive appeal because explaining leadership with recourse to a cost–benefit analysis makes sense in terms of theories of economic exchange that have widespread currency elsewhere in psychology (and in the social sciences more generally, particularly sociology and economics). Nevertheless, the concepts of 'cost' and 'benefit' prove to be very elastic and can be difficult (if not impossible) to specify independently of the vantage point from which they are assessed. One followers' meat is another's poison, and indeed what the same follower considers meat in one context, he or she will consider poison in another.

Finally, again, another objection to this approach is that it reduces a higher-order experience to a relatively mechanical cost–benefit analysis. Whereas plenty of texts attest to the joy of experiencing great leadership, *The Joy of LMX* seems an unlikely title. In part this is because it treats followership as a question of "What is in it for me?," when such questions seem remote from the phenomenology of successful leadership. Indeed, there is evidence that asking such questions often leads to group disintegration rather than group success (Tyler and Blader, 2000).

Leadership as Transformation

Recognition of the fact that leadership often appears to rely on leaders and followers rising above their personal self-interest has led to a reemphasis of the role that leader charisma plays in facilitating acts of citizenship, loyalty, and service to the group. This concern is central to models of *transformational leadership*, as evidenced, for example, by the work of James MacGregor Burns (1978). Burns prefaced his work with a stark assessment of the failings of approaches that see

leadership as being primarily about power and resource management rather than influence. True leadership, he contended, arises from *working with* followers and is about much more than simply satisfying their wants and needs in exchange for support. In particular, he suggested that leadership necessarily moves beyond a contractual arrangement whereby people act in particular ways because they feel *obliged* to and, instead, engages with higher-level sensibilities that lead people to pursue a particular course of action because they *want* to, and because they feel that what they are doing is *right*.

Transformational theory expands on these ideas by drawing on motivational and developmental theories which assert that human development involves people's progression from lower-level understandings of themselves and their world (dictated by relatively base urges for things like money) through to more sophisticated higher-level understandings (underpinned by loftier concerns for things like self-actualization and self-esteem). For Burns, a key feature of successful leadership is that it helps people progress up such hierarchies, thereby allowing them to scale greater psychological and moral heights.

Nevertheless, while transformational theory suggests that leadership involves processes of mutual respect and shared perspective, it is worth noting that the psychological theorizing on which it draws assumes that the highest state of motivation and morality is characterized by individual autonomy (Ellemers et al., 2004; Haslam, 2001). This emphasis on the individual is also apparent in practice where, as with classical approaches, measures of transformational leadership (in particular, the *Multifactor Leadership Questionnaire* (MLQ); Bass and Avolio, 1997) focus on the abilities and qualities of the individual leader.

One unique feature of the MLQ is that it is completed not only by individual leaders but also by their superiors, peers, and subordinates (as well as outsiders – e.g., clients). This process of *360 degree feedback* is seen to provide a more valid assessment of the extent to which a leader is charismatic, intelligent, inspirational, and considerate. Indeed, a clear advantage of such procedures over those used by proponents of classical approaches is that they recognize that these characteristics are ultimately conferred by followers rather than simply possessed by leaders themselves (i.e., as argued by Lord and Maher, 1991; Weber, 1921/1946).

Again, though, as with advocates of classical models of old, those who adopt this approach (especially practitioners) often treat these characteristics as fixed and personal rather than dynamic and social. The approach also provides limited insight into the processes that lead to a given leader being seen as having transformational qualities. For example, there appear to be few personal characteristics that people like Mahatma Gandhi, Adolf Hitler, Princess Diana, Nelson Mandela, Margaret Thatcher, and Saint Paul have in common, and which separate them from less charismatic mortals. As we will clarify further below, what they do have in common is the fact that they envisioned, and became emblematic of, particular forms of progress for groups that, at particular points in history, met with some 'success.' Indeed, this latter observation has led researchers like James Meindl (1993) to argue that leadership has much less to do with the character of leaders than with the favorability of the general social environment within which they operate and the romantic attributions to leadership that followers make under these circumstances (for a more radical critique of this form, see Gemmill and Oakley, 1992).

In sum, then, work on transformational leadership makes a strong case for the importance of coming to terms with a leader's transformational capacity, but it sheds relatively little light on the psychological processes that underpin this capacity. It is here that we can turn to insights from the social identity tradition. For here, three decades of research have focused very much on the social psychology of motivation and influence, and on the capacity for this to be transformed by shared group membership.

The New Psychology of Leadership: An Identity Approach

Although it has clear limitations, work in classical and contingency traditions is instructive in a number of ways – most obviously in pointing to the fact that an adequate theory of leadership needs to explain a number of disparate features of this process. At core, these relate to the following four observations:

1. Different forms of leadership are required in different contexts (Fiedler and House, 1994).
2. Followers' perceptions of leaders are all-important (Lord and Maher, 1991), but vary as a function of social context.
3. Leaders and followers have a dynamic capacity to motivate and influence each other (Ellemers et al., 2004; Reicher et al., 2005).
4. In the process of developing the power to change the world, leaders and followers are themselves transformed (Burns, 1978; Turner, 2005).

In striving to develop an analysis that might account for these observations, two further points are worth making that are commonly overlooked. The first is that leaders are never just leaders in the abstract. They are always leaders *of* some specific group or collective – a country, a political party, a sporting team, and so on. By the same token, second, their followers do not come from anywhere. Potentially at least, they too are members of the same group. Leaders and followers are therefore bound together by their being part of – and by their sense that they are part of – a common group.

This idea that leadership centers on individuals' conception of themselves as group members is at the heart of work in the social identity tradition (e.g., see Haslam, 2001; van Knippenberg and Hogg, 2003; Turner, 2005). Social identity refers to a person's sense that he or she is member of a particular social group (e.g., a nation, an organization) and that this group membership is important and meaningful (Tajfel and Turner, 1979). Social identity thus allows people to define the self as 'we' and 'us' (rather than just 'I' and 'me') and to use this is a basis not only for perception but for action (Turner, 1982).

When it comes to leadership, work that has explored the dynamics of social identity makes at least four key contributions to our understanding of this process (Haslam et al., 2011).

First, it shows that when people define themselves in terms of a given group membership, they are motivated to see that in-group as positively distinct from other out-groups, and that here what counts for an individual is not his or her personal fate but the standing of their group as whole (Tajfel and Turner, 1979). Second, it shows that different forms of group behavior stem from different definitions of this social identity (Reicher et al., 2005). In other words, how we behave as a group depends on 'who we think we are.' Third, it shows that social identities are always bound up with social context – depending, among other things, on where we are and who we compare ourselves with (Oakes et al., 1994). Fourth, it shows that social identity is the necessary basis for collective behavior (Turner, 1982). It is only when we define ourselves as members of a common group that we can act as a group.

Leadership as Identity Representation

All of the forgoing points have important implications for the analysis of leadership. Indeed, just as John Turner (1982) asserted that social identity is what makes group behavior possible, so too one can argue that it is social identity that makes both leadership and followership possible (Haslam, 2001; Turner and Haslam, 2001).

This point can be spelled out further by noting that when people define themselves in terms of a given social identity (e.g., as 'us Australians') they seek both to discover what being a member of that group entails (e.g., liking sport, disliking 'tall poppies') and then to act in ways that accord with this (e.g., going to sports matches, deriding high-flyers). But who do we turn to for information about these things? The obvious answer is fellow in-group members. Indeed, one potent way of thinking about leaders is to see them as precisely such people: individuals who are perceived to be qualified to provide information about us and our place in the world by virtue of the fact that they are representative – or, in more technical language, *prototypical* (Turner, 2005) – of groups with which we identify.

These ideas have been supported by a large body of research. In particular, experimental research has shown that leaders who are prototypical of a perceiver's in-group exert more influence over them than those who are nonprototypical (Hogg and van Knippenberg, 2004). Indeed, a consistent finding in this work is that the capacity for a leader to shape the attitudes and behavior of others – in particular, by encouraging original and creative acts of followership – increases to the extent that he or she is seen to be 'one of us' (Haslam and Platow, 2001). Experiments show that it is only when a leader has a history of representing and advancing the interests of a contextually defined in-group, that followers prove willing to 'go the extra mile' in order to turn help translate the leader's vision for the group into reality.

Research by Michael Hogg also shows the influence of prototypical leaders is attributable to their prototypicality rather than the fact that they conform to particular leadership stereotypes (Hogg and van Knippenberg, 2004). In other words, it matters more that a leader looks like 'one of us' than that he or she looks like a 'typical' leader. Indeed, research shows that when stereotypical leader qualities such as trustworthiness, fairness, and charisma predict who emerges as a leader (and who does not), this is because, in the context in which they are studied (i.e., for the particular group in question) these are qualities that are prototypical of the in-group (Haslam et al., 2011).

Extending this point, the programmatic work of Michael Platow also shows that stereotypical leadership qualities like trustworthiness, fairness, and charisma are actually the *consequence* of in-group prototypicality (Platow et al., 2003). It is therefore possible to refine Weber's (1921/1946) claim that leaders' charisma is conferred by followers by observing that this only occurs to the extent that the leaders in question are representative of 'us' (Platow et al., 2006). Refining Meindl's (1993) insights, it also appears that the extent to which followers romantically (mis)attribute group success to the superior quality of leaders is similarly contingent on those leaders' in-group prototypicality (Haslam et al., 2011).

Importantly, too, to the extent that a group member is seen as in-group prototypical, he or she is given greater latitude to display creativity by moving the group in new directions – directions that might otherwise be seen as inappropriate, objectionable, or disloyal (Platow et al., 2006). Yet, by the same token, if aspects of leadership (including the process of appointment) serve to break leaders' ties to their group, then this will tend to undermine their capacity to direct the group effectively. This claim is consistent with evidence that group processes and structures which set leaders apart from the group they are trying to lead tend to undermine both follower support and overall group performance (Hollander, 1985).

Leadership as Identity Realization

The above arguments suggest that a leader's capacity to influence followers flows from his or her capacity to represent a group membership that they both share. But shared identity is not something that is fixed and given. It is also something that can be cultivated (and, for that matter, destroyed). Appreciation of this point leads to the claim that precisely because social identity constitutes such a powerful social force, then anyone who is interested in shaping and changing the world needs to be interested in defining social identity. As Steve Reicher and Nick Hopkins argue (Reicher et al., 2005; Reicher and Hopkins, 1996) leaders thus need to be *entrepreneurs of identity* who work with followers to craft a sense of who they are and what they want to be in ways that make the case for their own relative in-group prototypicality.

One extensive body of work that supports these ideas has involved detailed examination of politicians' speeches to explore the processes through which social identity shapes collective action. One pertinent example is provided by analysis of speeches by leaders of British political parties (Margaret Thatcher and Neil Kinnock) to their respective party conferences during the British miners' strike of 1984–85 (Reicher and Hopkins, 1996). This showed how these speeches served to construe events in such a way that (1) the leader's party could be seen as representative of a positively defined in-group encompassing (almost) the entire population and (2) the leader's policies were consonant with the definition of that in-group identity.

Work of this form shows how identity definition mobilizes followers around a collective vision of identity. Yet however

compelling a vision, and whatever its ability to mobilize in the short-term, vision and rhetoric alone are not enough to sustain followers' faith in social identity in the longer term. For social identity is only useful to the extent that it allows group members to envision and then *create* a better future. Accordingly, if collective mobilization fails to translate the definition of identity into consonant forms of material reality, then that definition – and those leaders who advance it – will fall by the wayside. By contrast, where mobilization does succeed in creating realities that reflect a given definition of identity, then that definition will enjoy increasing support and the leader will receive considerable acclaim. In these terms, then, the X-factor that Gandhi, Hitler, Diana, Mandela, Thatcher, and Paul had in common was that they were effective *identity impresarios* who worked hard to develop structures that allowed the idea of 'us' to be lived out and translated into material change in the world (Haslam et al., 2011).

Conclusion

The forgoing analysis leads us to the conclusion that leadership is a group process that centers on an identity-based relationship between leaders and followers wherein leaders gain effectiveness through an ability to represent and realize the meaning and aspirations of the group. On the one hand, this acts as a constraint upon leaders. Leaders cannot say anything or get followers to do anything. They are reliant on their ability to persuade followers of their prototypicality and normativity, and this in turn depends upon features of social context. But on the other hand, it is social identity that enables leaders to energize people with their vision, and to recruit the agency of followers in order to transform both their self-understanding and the world they inhabit (Haslam et al., 2011; Reicher et al., 2005).

Leaders, followers, and situations are not static elements that exist and operate independently of each other. Rather, they interact to shape each other. Moreover, it is through this dynamic interaction – and the process of galvanizing a collection of disparate 'I's into a powerful world-changing 'we' – that the transformational power of leadership is generated. Ironically, too, it is also in this process that the seductive myths of leadership come to the fore and prove to be its undoing. For once they have changed the world by mobilizing the power of the group, there is great temptation (and considerable inducement) for leaders to fall back on classical models which lead them to imagine that this was all their own doing. Airport bookshelves heave under the weight of such hubris and beckon us to our ruin.

See also: Collective Behavior, Social Psychology of; Group Processes, Social Psychology of; Identity and Identification, Social Psychology of; Self-Categorization Theory; Social Identity in Social Psychology; Social Psychology.

Bibliography

Antonakis, J., Ashkanasy, N.M., Dasborough, M., 2009. Does leadership need emotional intelligence? Leadership Quarterly 20, 247–261.

Bass, B.M., Avolio, B.J., 1997. Full Range Leadership Development: Manual for the Multifactor Leadership Questionnaire. Mind Garden, Redwood City, CA.

Bass, B., Riggio, R.E., 2006. Transformational Leadership, second ed. Taylor & Francis, New York.

Burns, J.M., 1978. Leadership. New York.

Carlyle, T., 1840. Heroes and Hero Worship. Harrap, London.

Ellemers, N., De Gilder, D., Haslam, S.A., 2004. Motivating individuals and groups at work: a social identity perspective on leadership and group performance. Academy of Management Review 29, 459–478.

Esler, P.F., 2003. Conflict and identity in Romans: the social setting of Paul's letter. Fortress Press, Minneapolis.

Fiedler, F.E., 1964. A contingency model of leader effectiveness. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 1. Academic Press, New York, pp. 149–190.

Fiedler, F.E., House, R.J., 1994. Leadership theory and research: a report of progress. In: Cooper, C.L., Robertson, I.T. (Eds.), Key Reviews in Managerial Psychology. Wiley, New York, pp. 97–116.

Fleishman, E.A., Peters, D.A., 1962. Interpersonal values, leadership attitudes, and managerial success. Personnel Psychology 15, 43–56.

Gemmill, G., Oakley, J., 1992. Leadership: an alienating social myth? Human Relations 45, 113–129.

George, J.M., 2000. Emotions and leadership: the role of emotional intelligence. Human Relations 53, 1027–1055.

Gibb, C.A., 1947. The principles and traits of leadership. Journal of Abnormal and Social Psychology 42, 267–284.

Gibb, C.A., 1958. An interactional view of the emergence of leadership. Australian Journal of Psychology 10, 101–110.

Graen, G.B., Uhl-Bien, M., 1995. Relationship-based approach to leadership: development of leader-member exchange (LMX) theory of leadership over 25 years: applying a multi-level multi-domain perspective. Leadership Quarterly 6, 219–247.

Haslam, S.A., 2001. Psychology in Organizations: The Social Identity Approach. Sage, London.

Haslam, S.A., Platow, M.J., 2001. The link between leadership and followership: how affirming a social identity translates vision into action. Personality and Social Psychology Bulletin 27, 1469–1479.

Haslam, S.A., Reicher, S.D., Platow, M.J., 2011. The New Psychology of Leadership: Identity, Influence and Power. Psychology Press, London and New York.

Hogg, M.A., van Knippenberg, D., 2004. Social identity and leadership processes in groups. Advances in Experimental Social Psychology 35, 1–52.

Hollander, E.P., 1964. Leaders, Groups, and Influence. New York.

Hollander, E.P., 1985. Leadership and power. In: Lindzey, G., Aronson, E. (Eds.), The Handbook of Social Psychology, third ed. Random House, New York, pp. 485–537.

Hollander, E.P., 1995. Organizational leadership and followership. In: Collett, P., Furnham, A. (Eds.), Social Psychology at Work: Essays in Honour of Michael Argyle. Routledge, London, pp. 69–87.

Ilies, R., Nahrgang, J.D., Morgeson, F.P., 2007. Leader-member exchange and citizenship behaviors: a meta-analysis. Journal of Applied Psychology 92, 269–277.

Judge, T.A., Colbert, A.E., Ilies, R., 2004. Intelligence and leadership: a quantitative review and rest of theoretical propositions. Journal of Applied Psychology 89, 542–552.

van Knippenberg, D., Hogg, M.A. (Eds.), 2003. Leadership and Power: Identity Processes in Groups and Organizations. Sage, London.

Lord, R.G., Maher, K.J., 1991. Leadership and Information Processing: Linking Perceptions and Performance. Unwin Hyman, London.

Mann, R.D., 1959. A review of the relationship between personality and performance in small groups. Psychological Bulletin 56, 241–270.

Meindl, J.R., 1993. Reinventing leadership: a radical, social psychological approach. In: Murnigham, J.K. (Ed.), Social Psychology in Organizations: Advances in Theory and Research. Engelwood Cliffs, NJ, pp. 89–118.

Oakes, P.J., Haslam, S.A., Turner, J.C., 1994. Stereotyping and Social Reality. Blackwell, Oxford.

Plato, 380BC/1993. The Republic. Oxford.

Platow, M.J., Haslam, S.A., Foddy, M., Grace, D.M., 2003. Leadership as the outcome of self-categorization processes. In: van Knippenberg, D., Hogg, M.A. (Eds.), Leadership and Power: Identity Processes in Groups and Organizations. Sage, London, pp. 34–47.

Platow, M.J., van Knippenberg, D., Haslam, S.A., van Knippenberg, B., Spears, R., 2006. A special gift we bestow on you for being representative of us: considering leader charisma from a self-categorization perspective. British Journal of Social Psychology 45, 303–320.

Reicher, S.D., Hopkins, N., 1996. Self-category constructions in political rhetoric: an analysis of Thatcher's and Kinnock's speeches concerning the British miners' strike (1984–5). European Journal of Social Psychology 26, 353–372.

Reicher, S.D., Haslam, S.A., Hopkins, N., 2005. Social identity and the dynamics of leadership: leaders and followers as collaborative agents in the transformation of social reality. Leadership Quarterly 16, 547–568.

Reynolds, K.J., Turner, J.C., Branscombe, N.R., Mavor, K.I., Bizumic, B., Subašić, E., 2010. Interactionism in personality and social psychology: an integrated approach to understanding the mind and behaviour. European Journal of Personality 24, 458–482.

Stogdill, R.M., 1948. Personality factors associated with leadership: a survey of the literature. Journal of Psychology 25, 35–71.

Tajfel, H., Turner, J.C., 1979. An integrative theory of intergroup conflict. In: Austin, W.G., Worchel, S. (Eds.), The Social Psychology of Intergroup Relations. Brooks/Cole, Monterey, CA, pp. 33–47.

Turner, J.C., 1982. Towards a cognitive redefinition of the social group. In: Tajfel, H. (Ed.), Social Identity and Intergroup Relations. Cambridge University Press, Cambridge, pp. 15–40.

Turner, J.C., 2005. Examining the nature of power: a three-process theory. European Journal of Social Psychology 35, 1–22.

Turner, J.C., Haslam, S.A., 2001. Social identity, organizations and leadership. In: Turner, M.E. (Ed.), Groups at Work: Advances in Theory and Research, pp. 25–65. Hillsdale, NJ.

Tyler, T.R., Blader, S.L., 2000. Cooperation in Groups: Procedural Justice, Social Identity, and Behavioral Engagement. Psychology Press, Philadelphia, PA.

Weber, M., 1921/1946. The sociology of charismatic authority. In: Gerth, H.H., Milles, C.W. (Trans. and Eds.), Max Weber: Essays in Sociology. Oxford University Press, New York, pp. 245–252.

Zimbardo, P., 2007. The Lucifer Effect: How Good People Turn Evil. Random House, London.

Leadership Education: Theory and Practice

Robert Emmerling, Basak Canboy, Ricard Serlavos, and Joan MB Foguet, Ramon Llull University – ESADE Business School, Leadership Development Research Centre, Sant Cugat (Barcelona), Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Leadership theory has evolved from an emphasis on traits and behavioral styles, to the more interactive theories associated with contingency theory, transactional and transformational leadership, and the emerging areas of servant leadership and shared leadership. The implication of evolving leadership theory for the education and development of leaders is explored as well as the methods commonly associated with developing leaders. Methodological issues related to research on leadership education and emerging areas related to developing leaders are also discussed.

Introduction

Developing effective leaders is one of society's most critical tasks, and one that directly affects the well-being, productivity, and sustainability of the organizations and societies in which we live. As such, it is important to understand how leaders can be best prepared to effectively deal with current and future challenges. Business, government, military, educational institutions, and not-for-profit organizations all require effective leaders to function well and serve the needs of their various constituencies. Today's modern organizations are characterized by rapid change and increasing complexity, which makes the need for effective leadership even more acute. Thousands of studies on leadership, various definitions of leadership, and theories of leadership development have made the field of leadership studies a complex one (Bass, 2008; Yukl, 2012).

According to a survey of Human Resource (HR) leaders, the number one problem for HR directors is identifying and developing the leadership talent needed for growth and expansion of their respective organizations (Fegley, 2006). Organizations have made leadership development a top priority in terms of education and training. In fact, 2012 saw an estimated increase of 14% in spending on leadership development in the United States, bringing annual spending to more than 13 000 000 000 dollars (O' Leonard and Loew, 2012). However, even with the substantial amount of money and resources allocated to leadership development, organizations and stakeholders often complain about the lack of effective leaders as well as the efficacy of leadership development efforts to bring about meaningful improvement in leadership capability. Moreover, the consequences of ineffective leadership reverberate across all spheres of society, leading many to question the motives, ethics, and capability of today's leaders. High-profile scandals involving business leaders, government officials, and leaders in various other domains has put a renewed emphasis on ethics and corporate social responsibility and has generally revitalized the debate on how best to educate and develop leaders (Ghoshal, 2005).

Evolving Definitions and Theories of Leadership

According to Yukl (2012), "Leadership is the process of influencing others to understand and agree about what needs to be done and how to do it, and the process of facilitating individual and collective efforts to accomplish shared objectives." Bennis (2009) defined leadership as "… a function of knowing yourself, having a vision that is well communicated, building trust among colleagues, and taking effective action to realize your own leadership potential." While definitions of leadership vary among authors and scholars, all tend to agree that the process of leadership involves the ability to influence, motivate, and enable others toward achieving goals important to the group(s) to which they are members. Even though agreement might be found on this general definition, how to conceptualize leaders and how they go about the leadership role has evolved over time. These evolving conceptualizations have had direct implications for how leadership education has been framed and implemented.

Interest in more scientific leadership research started in the early part of the twentieth century with trait and 'great man' theories of leadership (Borgatta et al., 1954), which typically viewed leadership as an inborn quality of individuals. Early trait theories of leadership assumed effective leaders can be defined by a set of traits and abilities that they possess that nonleaders or ineffective leaders do not. More recent meta-analytic studies of personality have found extraversion, openness to experience, conscientiousness, and neuroticism/emotional stability to be related to leader emergence and leadership effectiveness, with extraversion being the factor with the most consistent relationship (Judge et al., 2002; Bono and Judge, 2004). This is most likely due to the tendency of extroverts to convey more positive emotions and optimism as well as their tendency to be energized by group settings. Meta-analytic studies of cognitive intelligence (Judge et al., 2004) have also shown a correlation with leadership, but in the words of the authors, "… results suggest that the relationship between intelligence and leadership is considerably lower than previously thought." Even when personality traits and intelligence are taken into account, what seems clear is that understanding leadership requires additional constructs to provide a more complete picture of leadership effectiveness.

Research on leadership effectiveness conducted at Ohio State University in the 1940s (Hemphill and Coons, 1950) and Michigan State University in the 1950s (Katz et al., 1950) began a research tradition often referred to as the behavioral approach to leadership which, instead of focusing on traits, focused on the specific behaviors associated with leadership. Key findings

from the Ohio State University studies were the dimensions of 'consideration' and 'initiating structure.' Consideration is the degree to which the leader is concerned with the welfare of group members and 'initiating structure' is the leader's concern for roles, actions, and how to organize activities so that tasks are accomplished by the groups they lead. Studies at Michigan State University in the 1950s had similar findings but also identified a construct labeled participative leadership. This represented an extension of leadership theory by viewing leadership as more of a process of mutual influence by which the leader seeks out and values the input and contribution of followers. Occurring in the historical context of the growing influence of behaviorism in psychology more generally, these theories differed from trait theories in that they sought to understand the specific behaviors of effective leaders; behaviors which could then be taught to current or prospective leaders.

As theory and research progressed, a move to incorporate characteristics of followers, and the environment in which the leader is embedded, started to become more salient. The 1960s and 1970s found interest emerging in contingency theories of leadership, which held that there was no one 'right' way to lead, but that leadership effectiveness depends on matching leaders to situations in which they were likely, because of their natural dispositions, to be successful. The contingency theory of leadership by Fiedler (1967) posited that leaders will excel if their leadership style matches the dimensions of the situation. However, a leader's style was not viewed by Fiedler as particularly malleable or adaptable. An alternative contingency approach to leadership can be seen in the situational leadership model (Hersey and Blanchard, 1969), which posits that leaders can be taught to diagnose specific aspects of the leadership situation (e.g., follower needs or readiness to engage in specific tasks) and modify their leadership behavior to a delegating, supporting, coaching, or directing style based on the situation. Like behavioral theories, skills such as assessing the needs of followers, as well as the specific behaviors associated with specific styles, represent knowledge and skills that could be transferred to leaders through training.

The 1970s saw the rise of transactional theories of leadership, which held that leaders motivate employees by providing sought after rewards for members to engage in behavior consistent with the leader's agenda and threats or punishments for poor performance. More recent research and theory have focused on transformational leadership, which is a leadership style defined by a dynamic interchange between leader and followers in which each challenges the other to higher aspirations. This leadership style tends to be characterized as a more visionary and charismatic style less reliant on the reciprocity characteristic of transactional forms of leadership. In this way, the leader appeals to higher level needs, aspirations, and expectations of followers. Transformational leadership has been operationalized into (1) idealized influence, in which the leader is seen as a role model, is authentic, trustworthy, and highly credible; (2) inspirational motivation, in which the leader provides a shared vision meaningful to followers and works to simplify complex ideas and problems; (3) individualized consideration, in which the leader appreciates the individuality of followers while treating them equally and developing them to higher levels of potential; and (4) intellectual stimulation, in which the leader encourages followers to question their old way of doing things and to break with the past. These facets of the transformational leadership model can be measured with the Multifactor Leadership Questionnaire (Bass and Avolio, 2000) and provides a specific framework of leadership capabilities that can be measured and taught.

The leader's motivational profile has also been put forward as a useful construct to better understand what drives specific behavioral patterns. The social motives of achievement, affiliation, and power, originally conceptualized by McClelland as on-going needs, wants, or concerns that drive patterns of thoughts and behaviors, have also been linked to leadership (McClelland and Boyatzis, 1982). The social motives are believed to operate at the unconscious level, meaning that most people likely have very little insight into their motive profile. The position that leadership behavior, like other forms of behavior, is partially determined and driven by forces outside of conscious awareness has been held by other leadership scholars as well (cf Kets de Vries, 2006). Research on the need for achievement, defined as a need to surpass a standard of excellence and need for power, defined as the need to have impact on others, have both been linked to effective management/leadership (McClelland and Boyatzis, 1982). The affiliation motive, defined as a need to establish, maintain, or restore harmonious relationship with others, was found to be lower than both achievement and power motives in more effective leaders.

Building on a theory of social motives, a theory of competencies, now referred to as the behavioral approach to emotional intelligence (EI) (Boyatzis, 2009), began to emerge from applied research on distinguishing characteristics of average vs effective leaders (Spencer and Spencer, 1993). The social motives are viewed as driving the frequency and intrinsic satisfaction derived from engaging in specific leadership behaviors associated with a given competency. While the social motives can be viewed as a trait, arousal of a given motive by factors in the environment may increase the frequency of specific behavioral manifestations of the motive. For example, when a leader's power motive is aroused, he or she is more likely to demonstrate behaviors associated with competencies such as influence or inspirational leadership. Also influencing, behavioral expression of specific competencies are the leader's values and self-image. The competency-based approach has implications for developing leaders by pointing to a more complex interplay between intentional behaviors driven by forces below conscious control and moderated by personal values and the leader's self-image. This research line contributed to a growing realization that developing one's capacity as a leader required a more complex approach than simply imparting knowledge, skills, and abilities. For leadership development to be truly transformational, it requires that leaders explore their identity as a leader, their values, and their future goals and aspirations.

The transformation of motives and personality traits of leaders has been challenged by many as these constructs have traditionally been viewed as more fixed and static in adults. This has given rise to questions about the efficacy of attempting to change these constructs through leadership development initiatives targeted at adults. The belief that leaders are born is one, which has stimulated several studies that have attempted to better understand the role of genetics in understanding both

personality and leadership. While personality traits have a significant genetic component, behavioral-genetic research on complex behavioral traits leaves much of the variability in personality traits unexplained. A review of this research by Plomin and Daniels (1987) concluded that it is rare for any study to find genetic evidence to account for more than half of the variance in complex behavioral traits. Genetic research aimed at predicting who ends up in leadership roles does not only demonstrate a significant relationship between genetics and leader emergence (De Neve et al., 2013), but also makes clear that other factors are needed to more fully explain leader emergence. Moreover, recent research also provides evidence that personality traits continue to change in adulthood and that these changes may be quite substantial and consequential (Roberts and Mroczek, 2008). The limited explanatory power of genetic explanations of leadership, combined with questions about the fixed nature of personality traits, calls into questions about the deterministic notions of who can and cannot become a leader.

While genetic and trait-based explanations may provide some insight into leader emergence, leader effectiveness once in the leadership role might be better explained by a leader's developmental history and experiences. Ligon et al. (2008) researched the life histories of 120 outstanding historical leaders and reported that more positive and constructive leaders had experienced life events that had firmly established positive personal beliefs and values early in their life span. The opposite was found for more destructive or selfish leaders, which tended to have experienced more negative events early in their development. The experience of such events likely influence one's narrative identity, defined as an internalized and evolving life story from which individuals seek to give meaning and purpose to significant life experiences (McAdams and Olson, 2010). The fact that narratives evolve overtime, and are subject to modification through different experiences and subsequent interpretation provides support for many leadership development practices. Understanding the leader's life in context and the practice of encouraging leaders to focus on the construction of a future vision, help leaders better understand their life context, and shape their future development plans. The importance of a personal vision is integral to many contemporary theories and practices related to leadership development.

Leadership theories reviewed up to this point tend to assume a more hierarchical, or vertical model of leadership. However, the realities of modern organizations have prompted us to view leadership as a shared or reciprocal process among team members. Pearce and Conger (2003) define shared leadership as "a dynamic, interactive influence process among individuals in groups for which the objective is to lead one another to the achievement of group or organizational goals or both. This influence process often involves peer, or lateral, influence and at other times involves upward or downward hierarchical influence." The framing of leadership as shared, in the context of changing authority dynamics, requires a fundamental shift in how leaders think about their role. Preparing leaders to operate within less formal hierarchies in which they may need to play the role of leader, peer, and subordinate all in the same meeting requires developing leaders with flexible influence and collaborative skills. Additional implications of shared leadership theories for leadership education include the need for action learning that requires teams to work on actual or simulated problems/projects without assigning a formal leader. Training, simulations, or developmental assignments that require leaders to assume multiple roles and to exercise influence in situations where they do not have formal authority should prove useful in developing the skills necessary to work effectively in situations, which require shared leadership.

A competing leadership theory is servant leadership, which was originally coined by Greenleaf (1991) but has been reinterpreted and expanded since its original conceptualization. A literature review by Russell and Stone (2002) found that servant leadership can be understood in terms of functional attributes of servant leaders, which include having a vision, being honest, trustworthy, service oriented, a role model, demonstrating appreciation of others' service, and empowerment. In relation to abilities and competencies, servant leaders are characterized as good communicators and listeners, credible, competent, encouraging of others, teachers, and delegators. In contrast to other theories of leadership, which tend to assume that the leader is primarily focused on achieving organizational goals, servant leaders are assumed to prioritize the needs of followers, with organizational objectives taking a secondary role. The conceptualization of servant leadership suggests that development of the associated attributes and competencies could be accomplished through service-learning projects that encourage learners to address issues outside their immediate organizational environment. Servant leadership makes an explicit link between leadership, ethics, and prosocial behavior, and defines a leader's contribution as measured by their impact on consumers, employees, communities, and a more widely defined network of stakeholders. This way of defining leadership corresponds well with contemporary notions of corporate social responsibility, which have now become more visible in business school curriculums. The increased use of service-learning programs both in university-based programs and other organizations throughout the world can partially be attributed to a rise in the popularity of servant leadership as a theory of leadership (**Table 1**).

Can Leadership Be Taught and Developed?

While theories of leadership vary, and tend to emphasize different knowledge, skills, abilities, and values, programs designed to enhance a leader's effectiveness also differ significantly in terms of structure, content, and approach. Given the wide swath of programs that fall within the leadership education paradigm, it should not be surprising that the return on time and money invested in leadership development can vary dramatically (Burke and Day, 1986; Collins and Holton, 2004; Avolio et al., 2009, 2010). While a comprehensive and accepted theory related to leadership development remains elusive, compelling evidence suggests that leaders can be educated to be more effective.

Research on job performance has demonstrated that job performance varies as a function of job complexity such that the more complex the job, the more performance variability there will be among individuals holding that job (Hunter et al., 1990). Leaders often have to deal with multiple levels of complexity and

Table 1 Summary of leadership theories

Leadership theory	*Time period*	*Implications for leadership education*	*Selected references*
'Great man' theories	Early twentieth century	Leadership cannot be developed because it was seen as an inborn quality	Borgatta et al. (1954)
Behavioral approach to leadership as a process of mutual influence	Mid-twentieth century	Focus on teaching leaders specific behaviors associated with leadership	Hemphill and Coons (1950) Katz et al. (1950)
Contingency theories including situational leadership model	1960s and 1970s	Focus on teaching leaders to diagnose specific situations and follower needs in order to use adequate leadership style	Fiedler (1967) Hersey and Blanchard (1969)
Transactional leadership theory	1970s	Teaching leaders to understand what motivates employees in order to provide adequate rewards or punishments	Bass (2008)
Transformational leadership theory	1980s and onward	Focus on teaching leaders how to be a role model, inspire with a shared vision, appreciate individuality of followers, and stimulate them intellectually	Bass and Avolio (2000)
Shared leadership theories	1990s and onward	Less-defined hierarchies require leaders to develop flexible influence and collaborative skills. Action learning offers an adequate framework as leaders usually work in teams on actual problems without formally assigned leadership roles	Pearce and Conger (2003)
Servant leadership	1990s and onward	Leadership education involves service-learning projects, which encourage learners to address issues outside their immediate organizational environment with a focus on the needs of employees, consumers, communities, and other stakeholders	Greenleaf (1991) Stone et al. (2004)
Social motives and competencies	1970s and onward	Intrinsic social motives impact the frequency of specific behaviors associated with different competencies. Leadership education has to allow leaders to explore their identities, values, and future goals and aspirations	Boyatzis (1982) McClelland (1987) Spencer and Spencer (1993)
Individual developmental history and experiences	Twenty-first century	Leadership development practices focus on one's understanding of life in context, interpreting life narratives and modifying them through positive experiences as well as constructing a future vision	Ligon et al. (2008) McAdams and Olson (2010)

lead others through ambiguous situations fraught with varying levels of risk, making leadership roles highly variable in terms of performance. Given the large variability in leadership performance, understanding the specific leadership competencies that account for variance in measurable organizational outcomes can help to focus leadership development efforts on specific competencies linked to specific organizational outcomes identified through applied research on high performing leaders (Ryan et al., 2009). An evaluation of a leadership development program, which was customized to develop specific competencies, identified through applied research, demonstrated favorable return on investment when business unit profitabilities of trained leaders were compared with a control group of untrained leaders in the same organization (Spencer, 2001). Moreover, the potential for training effects of leadership development to 'cascade' down through the organization also provides a mechanism by which the Return on Investment (ROI) of training leaders may potentially be magnified.

While the potential for significant, measurable impact from leadership development is great, systematic investigations of leadership development interventions are somewhat rare in the literature as are theories of leadership development. Nonetheless, several research studies and program evaluations have shown that leadership development interventions can have a positive effect on the attitudes, behaviors, and performance of leaders and their followers (Avolio et al., 2010). The first formal meta-analysis of managerial and leadership development initiatives conducted by Burke and Day (1986) analyzed studies published from 1952 to 1982 and found such interventions to be moderately effective. Collins and Holton (2004) extended and updated the previous study by conducting a meta-analysis on 83 intervention studies published between 1982 and 2001 and found moderate to strong effect sizes. Another meta-analysis also showed that the majority of leadership development initiatives both in laboratory and field settings have a positive impact, but this effect varies dramatically based on the type of intervention and the specific theory, which guided the design and implementation of the program (Avolio et al., 2009).

While research has been generally supportive of the positive effects of leadership development initiatives, some significant methodological issues have to be considered. Evaluations of leadership development programs typically focus on one or two specific leadership tasks or skills, for example, communication methods such as discussing performance problems constructively, motivation, or participative problem-solving techniques. Given the complexities associated with the leadership role, it is unlikely that these types of isolated interventions match the desire of organizational stakeholders to develop more capable leaders. Moreover, research that evaluates leadership development tends to be more focused on immediate or short-term outcomes, yet those sponsoring and participating in leadership development initiatives likely expect the effects of such interventions to be more long term.

Though not unique to leadership theory and practice, the wide range of leadership definitions poses a significant methodological challenge to research on leadership education. To test any substantive hypothesis related to leadership, the meanings of variables involved must be specified in observational terms and a method established for how they can be measured. However, as Bass (2008) points out, "There are almost as many definitions of leadership as there are persons who have attempted to define the concept." As a consequence, generalization about the effectiveness of leadership development becomes problematic, as does the implication of leadership development research for practice. The multiple settings and different populations, which regularly engage in leadership education also introduces unique factors that limit the development of a general theory of leadership development.

Venues traditionally associated with educating leaders have been business schools and universities. While interest in leadership education within university settings has been growing, compelling evidence of the effectiveness of such university-based programs has been elusive, causing many to question the utility of such programs. A notable exception is the competency-based MBA program at the Weatherhead School of Management (WSOM) at Case Western Reserve University (Boyatzis et al., 1994). Longitudinal studies at WSOM have demonstrated that MBA students can improve the emotional, social, and cognitive intelligence competencies that distinguish outstanding leaders and that improvements can be maintained for several years after program completion. The program at WSOM along with 14 other programs were identified as model programs by the Consortium for Research on Emotional Intelligence in Organizations (www.eiconsortium.org) due to evaluation data, which demonstrated that these programs had a measurable impact on social and emotional competencies and/or organizational outcomes. Based on model program evaluation, expert panels and applied research, 22 guidelines for best practice were established and now provide a comprehensive framework for developing social and EI competencies in adults (Cherniss et al., 1998).

While universities would seem a natural choice of venue for the education of leaders, critics both outside and inside academia have questioned this assumption. Mintzberg (2004) in his influential book *Managers not MBAs* argued that several factors work against the development of effective managers in traditional MBA programs and that managing is a practice that cannot be taught in a traditional classroom setting. He argues that teaching theories of management to nonpracticing managers is ineffective as they have a limited amount of relevant experience that can be related to what is being taught. Training in management and leadership should be limited to practicing managers and efforts made to leverage the knowledge and experience of those managers and share with them theories, which help them make sense of their current reality. Sharing of experience, self-reflection, and the necessity of application and experimentation of management theories back on the job lead to the transfer of theories into the practice of effective management and leadership.

Theory and Practices Related to Developing Leaders

Early practice related to developing leaders relied on the state of theory at the time. As such, earlier attempts to develop leadership relied more heavily on trying to teach theory and specific behaviors thought to be associated with effective leadership. Perhaps the biggest shift in educating leaders is to reframe

the process of leadership education from isolated training events to a learning process with various stages and that involves multiple training and development strategies in addition to traditional classroom-based training. Issues related to the formation of the person's identity as a leader, self-awareness, self-exploration, developmental readiness, motivation to change, the role of the leader's life experiences, and social environment have all begun to take a more central role in theory and practice related to developing leaders.

An example of viewing leadership development as a process can be seen in the more recent emphasis on understanding, and in some cases trying to enhance, a leader's readiness to engage in and benefit from leadership education. Developmental readiness is related to concepts previously identified in the training literature such as 'learner readiness' (Holton et al., 2000) or 'trainability' (Noe and Schmitt, 1986) but differs somewhat from these constructs in that developmental readiness goes beyond readiness to acquire more basic knowledge and skills and is related to a willingness to expand one's overall capacity and develop aspects of one's identity and social and emotional competencies associated with leadership. A review of the literature in clinical psychology suggests that the readiness of the individual to engage in therapy is likely more important than either the therapist or specific therapeutic technique employed (Avolio and Hannah, 2008). Given the transformative nature of many leadership development initiatives, such programs require individuals be ready to engage in a process that is essentially about a 'metamorphosis in actions, habits, or competencies associated with leadership effectiveness' (Boyatzis, 2008).

While individual level constructs are key to developmental readiness, the organizational climate must also be perceived as psychologically safe and supportive to facilitating transfer of training. Research on leadership training dating back to the 1950s (Fleishman, 1953) demonstrated that trainees that were trained to adopt a more supportive leadership style were only able to transfer this style to their role if their own supervisor had a similar leadership style.

Intentional change theory (ICT) (Boyatzis, 2006) explicitly addresses issues related to increasing motivation to engage in the developmental process as well as aspects of the social context in which leadership development occurs. In ICT, development is seen as an integrated process by which first individuals develop a clear longer term vision of their life and the leadership role(s) they would like to aspire to. This future vision of one's ideal self is then compared against one's current state in terms of competencies, values, and traits. This current state is referred to as the real self. Aspects of the real self are often assessed through tools such as multirater feedback, psychometric assessments of personality variables and values, or assessment center methodologies and simulations. Comparing one's future ideal self with the current real self motivates leaders to work to reduce the difference between the two selves and thus bring themselves more in line with their ideal self. Development of a detailed vision of one's ideal self is of critical importance as it helps to provide an emotionally engaging, positively framed ideal to which the leader can aspire. Without a clear and emotionally engaging future vision, motivation to develop will be limited and the ability of the learner to persist in the face of obstacles and setbacks related to their development will be reduced.

According to ICT, once the ideal and real selves have been defined, only then does one start to set learning goals and to work on a learning plan, which could include traditional skill-building courses (e.g., presentation skills), developmental assignments, which require a person display and practice specific competencies, individual coaching, and/or mentoring. With the plan completed, often with the help of a coach or mentor, the leader begins to implement their learning plan and begins to practice and experiment with new behaviors and competencies. In ICT, development is seen as a circular process in which the leader again returns to the question of their personal vision. Meaning that as one develops greater capacity and achieves one's developmental goals, thus bringing one closer to one's ideal self as articulated in their personal vision, a revised vision is often needed given the individual's changing role, career stage or life circumstances. Also central to ICT is the role of resonant relationships characterized by trust, support, and a positive emotional connection, which serve to provide encouragement, feedback, and advice to aid the leader in the pursuit of their development goals. Consistent with contemporary thinking on adult learning, the specific development goals are achieved primarily through experiential learning techniques.

The experiential learning theory of Kolb (1984) builds on earlier theories, which emphasized the role of experience in learning. Experiential learning theory defines learning as "the process whereby knowledge is created through the transformation of experience. Knowledge results from the combination of grasping and transforming experience" (Kolb, 1984). In this way, cognitive, affective (emotional), and behavioral systems are all engaged as part of an initial experience, which are later reflected on and incorporated in the form of new knowledge, skills, and abilities. Researchers from the Center for Creative Leadership have suggested that 70% of leadership development is accomplished through experience on the job (e.g., developmental assignments), 20% is accomplished through the social effects of mentoring, coaching, and shadowing role models that demonstrate specific leadership competencies and 10% through attending formal courses or reading about specific leadership topics (Lombardo and Eichinger, 2000). The emphasis on multiple learning methodologies offers the learner a diverse experience and is generally consistent with current thinking in leadership education about the need to provide training using multiple methods to appeal to people with different learning styles.

Developmental assignments expose leaders to new and novel experiences that can work to enhance personal development and build leadership capacity. Often developmental assignments broaden the current role of a person within the same organization. Such assignments can also be a valuable method to develop leadership competencies through collaboration in practical projects outside one's current role. Service learning can be seen as a specific type of developmental assignment, and is often applied in educational settings with the objective of developing specific leadership skills and competencies. Such assignments commonly revolve around activities that address human and community needs,

promoting ethical and socially responsible behavior in participants (Dumas, 2002). Though more common in academic settings, companies have started to discover these types of collaborations as meaningful opportunities to enhance leadership skills within their in-company development programs. The 'Ulysses Program' has leaders participate in an 8-week collaboration project with social entrepreneurs, companies, or NGOs. Combined with individual assessment, coaching, and team building, this international field experience allows participants to develop in concepts like responsible mind-set, ethical literacy, cultural intelligence, global mind-set, self-development, and community building (Pless et al., 2011). Service-learning assignments in higher education and corporate settings have been employed, in part, as a response to calls for leadership education interventions, which specifically seek to develop business ethics and corporate social responsibility along with other leadership competencies.

The use of leadership coaches in conjunction with leadership development programs has expanded dramatically in the last 20 years. The Center for Creative Leadership (Ting and Hart, 2004) defines coaching as a formal one-on-one relationship, "in which the coachee and coach collaborate to assess and understand the coachee and his or her leadership developmental tasks, to challenge current constraints while exploring new possibilities, and to ensure accountability and support for reaching goals and sustaining development." Leadership coaching tends to be an individualized process developed to meet the unique needs, characteristics, and experiences of the coachee taking into account the needs of the organization in which the coachee is a member. Leaders engage in coaching to address a wide variety of needs, which may include a desire to increase self-awareness, develop specific leadership competencies, career transition issues, or to seek insight and perspectives on current or future challenges. The diversity of coaching approaches, and issues addressed by coaches in relation to leadership development, has made research into the coaching process and outcomes difficult. As a result, empirical research on coaching outcomes is currently limited to a handful of studies, many of which suffer serious methodological issues. Like other areas of leadership education, coaching is an area that could benefit from additional research to better understand how it impacts specific facets of leadership and leads to valued outcomes.

Formal and informal mentoring programs have also become more prevalent over the last 20 years and represent a more customized and context sensitive approach to educating and developing leaders. In contrast to coaching, where the relationship is built with an external professional, mentoring is based on a relationship linked to a more senior person within the same organization but outside the direct reporting line (McCauley and Douglas, 1998). Mentors help their mentees by supporting them with organizational understanding and creating a network that will be useful to the mentee's career advancement. The opportunity to observe and seek feedback and practical advice on current and future leadership challenges makes mentoring a useful form of relational learning for leaders. Moreover, the nature of learning in mentoring relationships is two way, meaning that both enter into the relationship expecting to be both expert and learner, to give and receive feedback, and enable and be enabled (Kram and Cherniss, 2001).

Leaders operating in today's more globalized environment are required to deal with a level of complexity and interconnection not previously seen. International and intercultural work has become the norm for large companies as well as a growing number of small- and medium-sized enterprises. Advances in technology has facilitated collaboration across international borders and heightened the need for leaders to develop their cultural intelligence. In terms of leadership education in university settings, this has often meant the opening of satellite extensions in different countries, international partnerships, and research collaborations intended to expose students and leaders to a diversity of cultural perspectives. The structure of many university-based programs now encourages students to work in cross-cultural project teams. More traditional approaches to developing the cultural intelligence of leaders have been to educate leaders about different cultural values and norms from the work of anthropologists and cross-cultural psychologists.

Emerging Trends in Leadership Education

Recent research and theory from psychology, neuroscience, organizational behavior, and corporate social responsibility are currently influencing trends in leadership education. The positive psychology movement and strength-based approaches to leadership have been emerging and exerting an influence on the practice of leadership education. In contrast to the 'disease model of human nature' that had dominated the last century of theory and practice within psychology, positive psychology seeks to understand what allows people to flourish and thrive. The role of positive emotions on learning, developing, and relationship building also represents a contribution of this emerging area to leadership education. An example of this can be seen in Fredrickson's (2001) broaden-and-build theory, which assumes that a key function of positive emotions is to encourage the broadening and building of social networks and open people to the possibility of developing new skills. According to Fredrickson, positive affective states could provide leaders with more positive psychological resources that they can draw on to support development. A trend to focus on a leader's current strengths and how these might be leveraged and used to overcome or compensate for weaknesses provides a different lens through which to view leadership development.

Theory related to EI has also begun to exert an influence in leadership theory and education. EI has been defined by Mayer et al. (2004) as "… the capacity to reason about emotions, and of emotions to enhance thinking as well as abilities to accurately perceive emotions, to access and generate emotions so as to assist thought, to understand emotions and emotional knowledge, and to reflectively regulate emotions so as to promote emotional and intellectual growth." Building on this definition and linking EI to a theory of work performance, Boyatzis (2009) defined a behavioral approach to EI as "… an ability to recognize, understand, and use emotional information about oneself that leads to or causes effective or superior performance." Given that emotions are a critical factor in

creating and sustaining motivation that makes the intelligent use of emotions an important leadership task. Competency in the area of empathy and emotional management should help leaders develop and communicate their visions and help them establish an emotional bond with followers. Moreover, the intelligent use of emotions also seems important to allow leaders to cope with the potentially distressing emotions associated with personal change, as well as permitting the leader to stay open to and learn from the emotional experiences inherent in more experiential methods of leadership development.

Emotion-related processes have historically been absent from most theories and models of leadership. The integration of emotional processes into leadership and management theory has provided a more complete picture of how leaders go about motivating and engaging those around them. Key findings from affective neuroscience have also begun to influence theory and practice related to leadership development. How leaders affect the emotions of followers and how emotional process works to build rapport and resonant relationships with followers is being informed by neuroscience. The concept of resonance has its neurophysiological underpinnings in neurons in the brain referred to as mirror cells (Rizzolatti et al., 2001). These cells are widely dispersed in areas of the brain and lead us to mimic and mirror the feelings and behaviors of others. The process of emotional contagion, characterized as processes that allow the sharing or transferring of emotions from one individual to another, often occurs without conscious knowledge (Barsade and Gibson, 2010). Research on leaders and emotional contagion has found the leader's positive mood, had a subsequent influence on group coordination and effort (Sy et al., 2005). Neurological research has also begun to understand the dynamic neurological underpinnings of concepts such as transformational leadership mechanisms that may underpin the leadership qualities of individuals (Balthazard et al., 2012). Moral reasoning and decision making has become an interesting area for affective neuroscience. Though research on neuroscience and leadership is in the early stages, this research could potentially see neurological assessment used to compliment more traditional leadership assessment for use in leadership development.

E-leadership and virtual teams is also emerging as an important issue for both research and practice related to leadership as the nature of virtual work teams continues to expand and evolve. Survey research has shown that areas that tend to be addressed related to the leadership of virtual teams includes leading virtual team meetings, coaching and mentoring team members virtually, developing trust in a virtual environment, as well as how to monitor performance and deal with performance issues virtually (Rosen et al., 2006; Malhotra et al., 2007). While research is increasing in this area, few program evaluation studies on the effects of specific training content or methodologies have been conducted. However, research has shown that virtual teams that perceived themselves to have more of a shared leadership model, vs a more traditional hierarchical model, had higher levels of performance (Hoch and Kozlowski, 2012). As technology continues to advance and globalization continues, it is likely that leading virtual teams will become even more relevant for leaders in the future.

The topic of leadership and leadership education will continue to hold the attention of both scholars and the general public for the foreseeable future. Determining the conditions and practices that must exist to meaningfully enhance leadership capability should continue to drive research on leadership education. What is certain is that that the landscape in which leaders operate will continue to change, yet the need for leaders to effectively motivate and inspire others will not.

Acknowledgments

The authors of this article are funded by MICINN (Spanish Ministry of Science and Innovation) via the project EDU2010/15250 and some have additionally received financial support from SUR del ECO de la Generalitat de Catalunya (Catalan government).

See also: Affect-Regulation Motivation; Behavioral Theories of Organization; Charisma, Social Aspects of; Cultural Intelligence and Competencies; Education and Learning: Lifespan Perspectives; Emotional Intelligence and Competencies; Emotions and Work; Globalization in Education; Leadership in Organizations, Sociology of; Leadership; Organizational Climate; Organizations: Authority and Power; Personality, Trait Models of; Power, Politics, and Influence in Organizations; Social Intelligence and Competencies; Teamwork and Team Performance Measurement.

Bibliography

Avolio, B.J., Avey, J.B., Quisenberry, D., 2010. Estimating return on leadership development investment. Leadership Quarterly 21, 633–644.

Avolio, B.J., Hannah, S.T., 2008. Developmental readiness: accelerating leader development. Consulting Psychology Journal: Practice and Research 60, 331–347.

Avolio, B.J., Reichard, R.J., Hannah, S.T., Walumbwa, F.O., Chan, A., 2009. A meta-analytic review of leadership impact research: experimental and quasi-experimental studies. The Leadership Quarterly 20, 764–784.

Balthazard, P.A., Waldman, D.A., Thatcher, R.W., Hannah, S.T., 2012. Differentiating transformational and non-transformational leaders on the basis of neurological imaging. The Leadership Quarterly 23, 244–258.

Barsade, S.G., Gibson, D.E., 2010. Why Does Affect Matter in Organizations? [References]. Readings in organizational behavior Routledge/Taylor & Francis Group; US, New York, NY.

Bass, B.M., 2008. Handbook of Leadership: Theory, Research, and Managerial Applications, fourth ed. Free Press, New York.

Bass, B.M., Avolio, B.J., 2000. MLQ Multifactor Leadership Questionnaire. Mind Garden, Redwood City.

Bass, B.M., Bass, R., 2008. Handbook of Leadership: Theory, Research, and Application. Free Press, New York.

Bennis, W., 2009. On Becoming a Leader, fourth ed. Basic Books, New York.

Bono, J.E., Judge, T.A., 2004. Personality and transformational and transactional leadership: a meta-analysis. Journal of Applied Psychology 89, 901–910.

Borgatta, E.F., Bales, R.F., Couch, A.S., 1954. Some findings relevant to the great man theory of leadership. American Sociological Review 19, 755–759.

Boyatzis, R., 1982. The Competent Manager: A Model for Effective Performance. John Wiley & Sons, New York.

Boyatzis, R.E., 2006. An overview of intentional change from a complexity perspective. Journal of Management Development 25, 607–623.

Boyatzis, R.E., 2008. Leadership development from a complexity perspective. Consulting Psychology Journal: Practice and Research 60, 298–313.

Boyatzis, R.E., 2009. Competencies as a behavioral approach to emotional intelligence. Journal of Management Development 28, 749–770.

Boyatzis, R.E., Cowen, S.S., Kolb, D.A., 1994. Innovation in Professional Education: Steps on a Journey from Teaching to Learning. Jossey-Bass, San Francisco, CA.
Burke, M.J., Day, R.R., 1986. A cumulative study of the effectiveness of managerial training. Journal of Applied Psychology 71, 232–245.
Cherniss, C., Goleman, D., Emmerling, R.J., Cowan, K., Adler, M., 1998. Bringing Emotional Intelligence to the Workplace. Consortium for Research on Emotional Intelligence in Organizations: Rutgers University, New Brunswick, NJ.
Collins, D.B., Holton, E.F., 2004. The effectiveness of managerial leadership development programs: a meta-analysis of studies from 1982 to 2001. Human Resource Development Quarterly 15, 217–248.
Daloz Parks, S., 2005. Leadership Can Be Taught. Harvard Business School Press, Boston.
De Neve, J., Mikhaylov, S., Dawes, C.T., Christakis, N.A., 2013. Born to lead? A twin and genetic association study of leadership role occupancy. The Leadership Quarterly 24, 45–60.
Dumas, C., 2002. Community-based service-learning: does it have a role in management education? International Journal of Value-based Management 15, 249–264.
Fegley, S., 2006. SHRM 2006 Strategic HR Management Survey Report. Society for Human Resource Management, Alexandria, VA.
Fiedler, E., 1967. A Theory of Leadership Effectiveness. McGraw-Hill, New York.
Fleishman, E.A., 1953. Leadership climate, human relations training, and supervisory behavior. Personnel Psychology 6, 205–222.
Fredrickson, B.L., 2001. The role of positive emotions in positive psychology: the broaden-and-build theory of positive emotions. American Psychologist 56, 218–226.
Ghoshal, S., 2005. Bad management theories are destroying good management practices. Academy of Management Learning & Education 4, 75–91.
Greenleaf, R.K., 1991. The Servant as Leader. The Robert K. Greenleaf Center, Indianapolis, IN.
Hemphill, J.K., Coons, A.E., 1950. Leader Behavior Description. Personnel Research Board, Ohio State University, Columbus, Ohio.
Hersey, P., Blanchard, K.H., 1969. Life cycle theory of leadership. Training & Development Journal 23, 26–34.
Hoch, J.E., Kozlowski, S.W., Dec 2012. Leading virtual teams: hierarchical leadership, structural supports, and shared team leadership. Journal of Applied Psychology. No Pagination Specified.
Holton III, E.F., Bates, R.A., Ruona, W.E., 2000. Development of a generalized learning transfer system inventory. Human Resource Development Quarterly 11, 333–360.
Hunter, J.E., Schmidt, F.L., Judiesch, M.K., 1990. Individual differences in output variability as a function of job complexity. Journal of Applied Psychology 75, 28–42.
Judge, T.A., Bono, J.E., Ilies, R., Gerhardt, M.W., 2002. Personality and leadership: a qualitative and quantitative review. Journal of Applied Psychology 87, 765–780.
Judge, T.A., Colbert, A.E., Ilies, R., 2004. Intelligence and leadership: a quantitative review and test of theoretical propositions. Journal of Applied Psychology 89, 542–552.
Katz, D., Maccoby, N., Morse, N., 1950. Productivity, Supervision, and Morale in an Office Situation. Institute for Social Research, Ann Arbor, MI.
Kets de Vries, M.F.R., 2006. The Leader on the Couch: A Clinical Approach to Changing People and Organizations. Jossey-Bass, San Francisco, CA.
Kolb, D.A., 1984. Experiential Learning: Experience as the Source of Learning and Development. Prentice-Hall, Englewood Cliffs, NJ.
Kram, K.E., Cherniss, C., 2001. Developing emotional competence through relationships at work. In: Cherniss, C., Goleman, D. (Eds.), The Emotionally Intelligent Workplace. Jossey-Bass, San Francisco, CA.
Ligon, G.S., Hunter, S.T., Mumford, M.D., 2008. Development of outstanding leadership: a life narrative approach. Leadership Quarterly 19, 312–334.
Lombardo, M.M., Eichinger, R.W., 2000. The Career Architect Development Planner, third ed. Lominger Limited, Minneapolis, MN.
Malhotra, A., Majchrzak, A., Rosen, B., 2007. Leading virtual teams. The Academy of Management Perspectives 21, 60–70.
Mayer, J.D., Salovey, P., Caruso, D.R., 2004. Emotional intelligence: theory, findings, and implications. Psychological Inquiry 15, 197–215.
Mcadams, D.P., Olson, B.D., 2010. Personality development: continuity and change over the life course. Annual Review of Psychology 61, 517–542.
Mccauley, C.D., Douglas, C.A., 1998. Developmental relationships. In: Mccauley, C.D., Moxley, E., Van Velsor, E. (Eds.), The Center for Creative Leadership Handbook of Leadership Development. Jossey-Bass, San Francisco, CA.
Mcclelland, D.C., 1987. Human Motivation. Cambridge University Press, New York.
Mcclelland, D.C., Boyatzis, R.E., 1982. Leadership motive pattern and long-term success in management. Journal of Applied Psychology 67, 737–743.
McKee, A., Boyatzis, R.E., Johnston, F., 2008. Becoming a Resonant Leader: Develop Your Emotional Intelligence, Renew Your Relationships, Sustain Your Effectiveness. Harvard Business School Press, Boston, MA.
Mintzberg, H., 2004. Managers, not MBAs: A Hard Look at the Soft Practice of Managing and Management Development. Berrett-Koehler Publishers, San Francisco, CA.
Noe, R.A., Schmitt, N., 1986. The influence of trainee attitudes on training effectiveness: test of a model. Personnel Psychology 39, 497–523.
O' Leonard, K., Loew, L., 2012. Leadership Development Factbook 2012: Benchmarks and Trends in US Leadership Development. Bersin & Associates, LLC, Oakland, CA.
Pearce, C.L., Conger, J.A., 2003. Shared Leadership: Reframing the Hows and Whys of Leadership. Sage, Thousand Oaks, CA.
Pless, N.M., Maak, T., Stahl, G.K., 2011. Developing responsible global leaders through international service-learning programs: the ulysses experience. Academy of Management Learning & Education 10, 237–260.
Plomin, R., Daniels, D., 1987. Why are children in the same family so different from one another? Behavioral and Brain Sciences 10, 1–16.
Rizzolatti, G., Fogassi, L., Gallese, V., 2001. Neurophysiological mechanisms underlying the understanding and imitation of action. Nature Reviews Neuroscience 2, 661–669.
Roberts, B.W., Mroczek, D., 2008. Personality trait change in adulthood. Current Directions in Psychological Science 17, 31–35.
Rosen, B., Furst, S., Blackburn, R., 2006. Training for virtual teams: an investigation of current practices and future needs. Human Resource Management 45, 229–247.
Russell, R.F., Stone, A.G., 2002. A review of servant leadership attributes: developing a practical model. Leadership and Organization Development Journal 23, 145–157.
Ryan, G., Emmerling, R.J., Spencer, L.M., 2009. Distinguishing high-performing European executives: the role of emotional, social and cognitive competencies. Journal of Management Development 28, 859–875.
Spencer, L.M., 2001. The economic value of emotional intelligence competencies and EIC-based HR programs. In: Cherniss, C., Goleman, D. (Eds.), The Emotionally Intelligent Workplace: How to Select for, Measure, and Improve Emotional Intelligence in Individuals, Groups, and Organizations. Jossey-Bass, San Francisco, CA.
Spencer, L.M., Spencer, S., 1993. Competence at Work: Models for Superior Performance. John Wiley & Sons, New York.
Stone, A., Russell, R., Patterson, K., 2004. Transformational versus Servant Leadership: a difference in leader focus. Leadership & Organization Development Journal 25, 349–361.
Sy, T., Cote, S., Saavedra, R., 2005. The contagious leader: impact of the leader's mood on the mood of group members, group affective tone, and group processes. Journal of Applied Psychology 90, 295–305.
Ting, S., Hart, E.W., 2004. Formal coaching. In: Mccauley, C.D., Van Velsor, E. (Eds.), The Center for Creative Leadership Handbook of Leadership Development. John Wiley & Sons, San Francisco, CA.
Van Velsor, E., McCauley, C.D., Ruderman, R.N., 2010. The Center for Creative Leadership Handbook of Leadership Development. Jossey-Bass, San Francisco.
Yukl, G., 2012. Leadership in Organizations, eighth ed. Prentic Hall, Englewood Cliffs, NJ.

Relevant Websites

www.ccl.org – Center for Creative Leadership.
www.eiconsortium.org – Consortium for Research on Emotional Intelligence in Organizations.
http://www.esade.edu/research-webs/eng/glead – Leadership Development Research Centre – ESADE.

Leadership in Organizations, Sociology of

Manfred Kets de Vries and Alicia Cheak-Baillargeon, INSEAD Global Leadership Centre, Fontainebleau Cedex, France

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.F.R. Kets de Vries, Volume 13, pp. 8573–8578, © 2001, Elsevier Ltd.

Abstract

This article updates an entry made in the first edition of the Encyclopedia (2001). It maintains the same structure by opening with a discussion on the definitional confusion among researchers of leadership – a confusion grounded in the fact that leadership can be looked at either a property (the position of the 'personalists') or a process (the position of their opponents, the 'situationists'). A number of influential leadership theories are examined, including theories that emphasize the importance of traits, behaviors, contingency, attribution, and symbolism. Charismatic and transformational leaderships are reviewed. Attention is given to the importance of the clinical paradigm in leadership research. Finally, we revisit developments in leadership studies since the first edition of the Encyclopedia a decade ago and propose emerging areas of research.

A Definitional Confusion

The Anglo-Saxon etymological origin of the words lead, leader, and leadership is *laed*, which stands for 'path' or 'road.' The verb *laeden* means 'to travel.' Thus a leader is one who shows fellow travelers the way by walking ahead. This metaphor of the leader as helmsman is still very much on the mark. Unfortunately, the clarity of leadership's etymology is rarely matched with clarity of meaning. When we plunge into the organizational literature on leadership, we quickly become lost in a labyrinth: there are endless definitions, countless articles, and never ending polemics. Papers, books, and articles claiming to delineate leadership proliferate, yet their conclusions can be confusing and even conflicting.

The most recent handbooks on leadership (Bass and Bass, 2008; Bryman et al., 2011) demonstrate the richness of leadership studies in their multiplicity of perspectives (social, psychological, historical, political, cultural, and even military) and approaches (theoretical, empirical, interdisciplinary, and policy-centered). Among the more popular are descriptions in terms of traits, behavior, relationships, and follower perceptions. Prevalent themes in the last decade include personality-based approaches, contingency theories, transformational leadership, leader-followership, innovation and creativity, the role of emotions, and the shadow side of leadership. More recently, with the financial crisis that began in 2008 and the poignancy of a number of high-profile corporate scandals, the nature and integrity of leadership practices have also come under the spotlight. The number of academic journals partially or fully devoted to the study of leadership continues to increase (*Leadership Quarterly*, *Leadership*, *Journal of Leadership Studies*, *International Journal of Leadership Studies*, *Journal of Leadership and Organizational Studies*, to name a prominent few) reflecting the diverse and creative discourse on leadership study, as have the number of practitioner or commercial books on leadership, indicating popular interest in the subject.

We continue to see a continued movement away from laboratory experiments, observations of leaderless groups, or the activities of lower level supervisors toward what leaders at a higher level are doing in the context of their work environment. In *Harvard Business Review's 10 Must Reads on Leadership* (2011), all of which include case studies with CEOs and top executives, the opinion on what makes a great leader may vary, but the authors generally agree that leadership is not a gift inherent to a chosen few, but something which can be cultivated. This includes fine-tuning one's emotional intelligence, implementing key leadership behaviors, adopting effective strategies for adapting to crisis and leading through change, the ability to find meaning and to learn from extremely difficult events, and practicing authenticity.

While much progress has been made, the proliferation of studies has only generated more questions. The observations and profiles of actual leaders show a theatrical gamut of personalities, attributes, strengths, and weaknesses, with different contexts calling for different leadership styles. A review by Yammarino et al. (2005) cited in Crossan and Mazutis (2008) identified at minimum 17 different leadership theories, providing effectiveness remedies ranging from classical approaches to more contemporary forms such as charismatic and transformational leadership. However, this review precludes more recent streams of leadership studies in areas such as strategic leadership or shared leadership or more positive forms such as authentic, spiritual, ethical, or responsible leadership. While we continue to broaden our understanding on leadership, great leadership, while consisting of many teachable components, remains illusive.

A Proliferation of Theories

Broadly speaking, two extreme positions can be identified in leadership research. On one side of the spectrum are the 'personalists' – researchers who argue that specific personality variables determine leadership effectiveness. On the other side of the spectrum are the 'situationists' – those who deny the influence of individual differences and attribute all variations in leadership effectiveness to environmental constraints. While personalists view leaders as heroic helmsmen, in control of whatever situation they find themselves, situationists turn leaders into figureheads – puppets manipulated by the forces of the environment.

These opposing positions set the stage for a cornucopia of theories, each backed by strong defenders. We can find 'great man' theories, trait theories, situational theories, psychoanalytic theories, political theories, humanistic theories, cognitive theories, leader-role theories, reinforced change theories, path-goal theories, contingency theories, multiple linkage theories, vertical dyad linkage theories, exchange theories, behavioral theories, and attribution theories. This lack of apparent convergence has caused some scholars to abandon the subject altogether.

One of the problems in dealing with the subject of leadership is that it can be looked at as both a property and a process. As a property, leadership is seen as a set of characteristics – role behaviors and personality attributes – that make certain people more effective in attaining a set of goals. As a process, it is seen as an effort by a leader, drawing on various bases of power, to influence members of a group to direct their activities toward a common goal. Taking the property–process debate as a point of departure, let us turn first to the property perspective.

Order Out of Chaos

The trait theory (or great man theory) offered one of the earliest conceptual ways of looking at leadership. This theory holds that there is one best way to lead and that deeply seated personality variables allow certain people to master that best way. According to this theory, there are a number of universal characteristics of personality that determine a leader's effectiveness, without regard to behavior in a given situation. Because leadership is viewed as a set of relatively stable and enduring personal traits or physical properties, specific personality characteristics distinguish effective from ineffective leaders.

The initial search for these universal traits applicable to any setting was not overly successful, however. The results were conflicting, with methodological problems in research design cited as the major reason. Disappointed by the results of these studies, many scholars interested in leadership abandoned this line of research altogether, turning to other approaches to leadership. In the early 1990s, however, after a long hiatus in trait research (and with the help of better measuring techniques) a revival of trait theory was observed. Those studies went beyond the simplistic, atomistic approach of previous trait studies and identified a number of personality characteristics that consistently emerged, differentiating leaders from nonleaders – dimensions of character that can be mapped into the Big Five model of personality structure (Hogan et al., 1994). These various dimensions can be described in terms of surgency (a broad term that embraces competitiveness, achievement orientation, self-assuredness, and dominance), agreeableness, conscientiousness, emotional stability, and intelligence (including emotional intelligence). In addition, these studies regularly listed factors such as physical energy and extraversion.

Another group of leadership scholars espouses a behavioral theory. Distinguishing between the technical actions of a leader and the human actions, these behaviorists emphasize a set of observable role behaviors rather than traits. Like trait theorists with their individual characteristics, these scholars see certain role behaviors as being universal – that is, as producing leadership effectiveness regardless of the setting. Generally, however, the constructs employed by these scholars have been too rudimentary. This approach often looks at behaviors via dimensions such as consideration vs initiation, or task orientation vs relationship orientation (Bales, 1958). Another popular typology with a behavioral slant on leadership contrasts autocratic and democratic approaches (Tannenbaum and Schmidt, 1958).

From this approach to leadership behavior, we move on the continuum to contingency theory. Instead of taking the position that leadership traits or behaviors are applicable to any situation, those supporting the contingency theory claim that the emergence of any one style is contingent on the environment in which the leader is operating. According to this point of view, the most effective leader is the one who is able to adapt his or her actions depending on the situation. In the model of one of the most prominent advocates of this point of view, the effectiveness of task- or relationship-oriented leaders depends on the favorableness of the situation as defined by the power of their position, the task structure, and the quality of the leader–member relationship (Fiedler, 1967).

The Importance of Attribution to Leadership

Another group of scholars espouses the attribution theory of leadership. According to these situationists, leadership is not a viable scientific construct; it is a mere label given to behavior. Only people's inferences about and reactions to leaders are viable (Calder, 1977). Because individuals have an inherent need to explain events that surround them, they assume that certain types of behaviors and actions can be attributed to the leader. Thus leadership is a perceptional issue, an illusion: individuals infer causation from observed behavior. The knowledge of the outcome causes individuals to attribute certain qualities to a leader.

This more situational point of view has been reinforced by a number of scholars of leadership who doubt whether leaders affect organizational performance. Advocates of this line of thinking contend that there are powerful external forces that shape organizational activities. Each leader is embedded in a social system – a system in which other actors not only have expectations regarding appropriate behavior but also make efforts to modify the leader's behavior – that places serious constraints on leader behavior. Leadership turns into a set of myths encouraging social construction of meaning, creating an illusionary causality (Pfeffer and Salancik, 1978).

The symbolic role of leadership has been further explored by other leadership scholars (Meindl et al., 1985). In an extension of the attributional school of thought, these researchers call attention to the 'romanticized' conception of leadership. Ironically (given the attributional school's situational roots), these scholars suggest that leaders can play an important role through the manipulation of symbols in the management of meaning – activities that can be highly effective in influencing others. Using the vehicle of symbolism, many advocates of the situational point of view have been inching toward an interactionist approach to leadership

studies, positioning the leader/led relationship clearly at central stage.

Leadership as a Charismatic Process

Like the attributional school of thought, the contingency approach has also set the stage for a more relational slant on the study of leadership. Believing a leader cannot be studied meaningfully in isolation from his or her surroundings, this approach views leadership as an interactive process among the leader, the followers, and the situation. This orientation, transcending earlier, more naive approaches to leadership behavior, is an important step forward.

The problem with many relational theories in the past is that their point of convergence was too narrow. Dimensions such as initiation vs consideration, social orientation vs task orientation, and autocracy vs democracy are overly simplistic in describing leadership in its context. Furthermore, these earlier relational studies focused far too much on exclusive superior–subordinate relationships, ignoring (or slighting) the various constituencies of the leader: the industry environment, the national culture, and the culture that characterizes the organization.

This shortcoming has opened the door for a fresh look at charismatic leadership. Another impetus for this line of research has been the prevalence of a business climate of uncertainty and unpredictability – a breeding ground for the emergence of charismatic leadership. In our competitive, global world, where the transformation and revitalization of organizations holds a central position, the leader is increasingly seen as a crucial agent of change.

The new focus, then, is on the inspirational role of leaders. Researchers are turning to the study of leaders who by force of their personality have an extraordinary effect on their followers. The challenge for leaders of organizations becomes how to affect the mind-set of the organizational participants through value creation, through influencing the organization's culture, and through building commitment to the organization's mission, objectives, and strategies to obtain well-above-average organizational performance.

The first person to take up this new challenge was political scientist MacGregor Burns (Burns, 1978). In his writing, he makes a distinction between transactional and transformational leadership. While transactional leadership can best be viewed as a mundane contractual exchange based on self-interest (often described in the literature as the manager's role), transformational leadership seeks to satisfy the higher needs of followers – to engage in a process of mutual stimulation and elevation whereby followers will transcend their own self-interests for the good of the group (Bennis and Nanus, 1985; Conger and Kanungo, 1998).

A number of researchers have built on Burns' notion of transformational leadership, using observed behavior of leaders to break the concept down into various components, in an effort to broaden early charismatic conceptualizations. For example, Bass and Avolio (1993), who view charisma as a subset of transformational leadership, list four behavioral components in the context of transformational leadership: (1) charisma or idealized influence, (2) inspiration, (3) intellectual stimulation, and (4) individualized consideration. According to them, charisma alone is insufficient to put in place a successful transformation process. Shamir et al. (1993) contend that charismatic leadership affects followers' self-concepts and has motivational consequences due to (1) changing follower perceptions of the task that has to be accomplished, (2) offering an attractive vision of the future, (3) creating a group identity, and (4) heightening individual and collective feelings of self-efficacy.

These various offshoots of a focus on the inspirational role of leadership contribute to a rich description of what the leadership mystique is all about. Researchers who view leadership as a charismatic or transformational process give proper attention to the contextual and cultural dimensions that are part and parcel of leadership dynamics. They are sensitive to the impact of the environment on leaders and on their behavior. Furthermore, they reject narrow instrumentalism in favor of a perspective whereby the leader is seen as the transformational agent of change. Other scholars have made the point, however, that the transactional role of leadership should not be ignored. They suggest that the most effective leaders take on two roles: a charismatic role (consisting of envisioning, empowering, and energizing) and an architectural role (designing the organization, setting up structures, and formulating control and reward systems (Kets de Vries and Florent-Treacy, 1999)).

Some scholars of leadership argue, however, that in spite of the new, richer color given to leadership research, additional steps need to be taken to deepen our understanding of the leader's relational interchanges (Martin, 2007; Kets de Vries and Balazs, 2011). And the challenge is formidable. In spite of the various rational ways in which researchers attempt to deconstruct leadership and charisma, charismatic leadership is not rational in the traditional sense of the word. By its very nature, it is unstable, in that it exploits what can be interpreted as irrational processes. We need now to find ways to explore the forces that transcend rationality. Critics also argue that the study design of many researchers evaluating inspirational leaderships treats all leaders and all followers as amorphous, interchangeable groups of people; in other words, they fail to attend to differences in personality style. To rectify these shortcomings in leadership research, deeper insight into people's desires, wishes, and needs is needed, and that insight can be provided by a clinical focus.

The Clinical Paradigm

The clinical orientation to leadership research uses findings from psychoanalysis, cognitive theory, developmental psychology, and family systems theory to arrive at a richer understanding of personality and leadership. In the deconstruction of the dynamics of leadership, this orientation looks to the triangle of mental life consisting of emotion, cognition, and behavior. While in other approaches to leadership, the focus is generally on cognition and behavior, in the clinical approach emotions enter the equation. Research on how people alter has revealed that cognition alone does not create change; cognition needs to be complemented by emotion. The clinical paradigm also factors in unconscious processes. It presupposes that the personality of a top executive influences the strategy, corporate culture, and even structure of his or her

organization to such an extent that often organizations cannot perform successfully if no attention is given to a leader's intrapsychic world.

The clinical orientation toward the study of leadership has helped achieve greater understanding of the leader–follower interchange. Research into the dyadic relationships created by leaders (and acquiesced to by followers) suggests that failure in leadership can also be considered failure in 'followership': just as influence moves down from the leader to followers, so also does it move up. Understanding the impact of transferential processes such as 'mirroring' and 'idealizing' – processes characterized by confusion of time and place between the leader and the led – helps researchers clarify otherwise inexplicable phenomena.

The clinical paradigm, with its belief that every thought and action has a reason, sheds new light on irrational behavior in organizations. Processes such as projection, projective identification, splitting, collective regression, identification with the aggressor, *folie à deux*, the fear of success, scapegoating, narcissism, vindictiveness, and containment – along with other elements of the clinical paradigm – can help researchers better understand the leader–follower exchange. The clinical paradigm can provide insights into the dynamics of group behavior and the role of the leader; illuminate male–female differences in leadership through the analysis of fantasies around gender; and provide an understanding of the positive and negative effects leaders have on the corporate culture, structure, and decision-making processes (Kets de Vries, 2006; Kets de Vries and Balazs, 2011). By looking at a leader's inner theater, scholars can better appreciate the reasons why that leader derails and his or her company fails.

Recent Developments

Although, at first step, venturing into the domain of leadership research may seem like walking on quicksand, this brief overview demonstrates the considerable advances that have been made over the last decades. Due to promising new research directions, especially those working toward a convergence of situationist and personalist positions, the prevailing attitude of disillusionment with leadership studies is experiencing a turnaround. After years of homogeneity in leadership research, the past two decades have seen a significant expansion in terms of theoretical frameworks and approaches. Most researchers of leadership now perceive the importance of a relational, interactionist point of view that looks at actual leaders in their 'natural' setting. Moreover, it is no longer difficult – given the rapid changes of our era – to convince researchers and practitioners of the relevance of the transformational side of leadership.

Since leadership is so closely tied to its context, the following represent a number of developments in the field as shaped by trends in contemporary organizational landscape:

1. *Cross-cultural leadership*: As global mergers, acquisitions, and strategic alliances become increasingly common, more attention needs to be given to the cross-cultural dimensions of leadership. When organizations join together in a partnership or merge to form a new company, it's very difficult to bring the strengths and values of both organizations together and to keep them alive in the new entity. Moreover, we have to take into account that leader effectiveness is contextual, and that success in one culture may not be successful in another. This is the reason why so many M&As fail. Hence, global or cross-cultural leadership is now becoming an important quality of a leader. Osland et al. (2006) define global leadership as "a process of influencing the thoughts, attitudes, and behaviors of a global community to work together synergistically toward a common vision and common goals" (p. 204). Such leaders are able to adjust to and lead effectively in different environments and to be able to communicate and work with diverse groups of people and partners. They have to be flexible and open to numerous diverse experiences. They have to have the acumen to decipher differences and idiosyncrasies of another culture and to know how to build off its strengths. However, questions remain as to what extent does effective leadership behavior vary from culture to culture, to identify the required competencies for leaders who run global organizations, and to foresee what are the future competencies he or she will need as organizations become more flexible, collaborative, and boundaryless (Martin, 2007).
2. *Shared, collective leadership*: The ability to create high-performance teams, in which leadership and decision making is shared, is also essential in present-day organizations. Leaders are confronted with new configurations such as virtual team leadership, distributed shared team leadership, and leadership within multiteam systems. In many organizations, teams have replaced executives acting independently and many teams exist for the specific purpose of handling what used to be traditional executive functions. Moreover, leadership is increasingly structured as a constellation of different individuals possessing different, but complementary roles (Kets de Vries, 2008). As such, team leadership requires people who have a solid dose of emotional intelligence who can get the best out of people; who unite their constituents around a common cause, and connect with them as human beings; and who help their people perform beyond expectations. Such leaders need to be able to create collaborative, cooperative teams of people who work together to prioritize activities and apply the organization's resources to reaching desired goals (Martin, 2007). While leading in a team can be one of the most fulfilling experiences of ones professional life, conversely it can also be one of the most frustrating and demotivating. Nothing is worse than being part of a dysfunctional, conflict-ridden team. Managing the darker side of team dynamics is also an important role of a leader (Kets de Vries, 2006; Kets de Vries and Balazs, 2011; Rosenbach et al., 2012). Questions remain as to how to capitalize on the synergy from diverse teams, how to create a shared reality among different members, and what can be done to 'stretch' and get the best out of the people in the team, and how a culture of shared leadership can be created throughout the organization?
3. *Moral leadership*: With the prevalence of corporate scandals in the last decade, governance issues are also taking the spotlight, with themes such as responsibility,

accountability, and ethical and socially responsible leadership. Corporate governance refers to the set of systems, principles, and processes by which a company is governed. They provide the guidelines as to how the company can be directed or controlled such that it can fulfill its goals and objectives in a manner that adds to the value of the company and is also beneficial for all stakeholders in the long term. But governance is part of the bigger issue of moral leadership, and one which attaches itself to the integrity and credibility of the leader and how he or she serves as the symbols to express the values that hold the society or organization together (Gardner, 1965). These are the enlightened and responsible leaders who are able to lift "people out of the petty preoccupations, carry them above the conflicts that tear a society apart and unite them in the pursuit of objectives worthy of their best efforts" (Gardner, 1965: p. 295). However, critical questions remain as to how society or organizations are developing current and future leaders to essentially rise above the collective and to take on the bigger questions of what the society or organization they are leading stands for, where they are heading, where they want to go, and what are the major forces and trends determining their future and to have the conviction, confidence, and courage to assume that role.

Emerging Areas of Research

Although leadership research has come a long way, more work needs to be done. One key factor in the selection of topics that require further investigation should be relevance of a topic to leadership's various constituencies. But 'relevance' needs to be broadly construed: the subject of leadership should be seen as applicable not only to a few highly exceptional individuals at the top of the organization but to a much broader audience.

Because of the challenges growing out of the exponential rate of change in this age of transformation, a number of issues feature prominently on the research agenda of the future.

1. *Adaptive leadership*: Organizations are now facing challenges at a rate unprecedented to previous times, with companies having to operate on a global level, keeping up with shifting competition bases, increased accountability from stakeholders, and the need for innovation and reinvention in order to stay ahead of the game. Leaders have to learn new ways of operating and mobilize their people throughout the organization to change its mind-set and behaviors in order to thrive in new business environments (Heifetz and Laurie, 1997). Leaders have the challenge of breaking long-standing behavioral patterns of their own, provide leadership that involves others, and to help organizational members to change their own behavior, while also managing the stress and uncertainty of change. They have to be able to create an environment in order to allow people to collaborate and work more interdependently, to build important relationships across boundaries, to lead with a more flexible style, and to be open and adaptable to new ideas (Martin, 2007). However, how does one go about to create a cultural change within the organization, and given the increased reliance of organizations on creativity and innovation, what can leaders do to stimulate and sustain this process?
2. *Authentic leadership*: Authenticity has its roots in Greek philosophy "know thyself" (Plato) and has seen variations over time "to thine own self be true" (Shakespeare). To guide others effectively, leaders must know themselves, who they are, and why they do what they do and to live up to their values. Authenticity suggests that individuals need to have a complete view of themselves and this includes taking into account both their strengths and weaknesses. Those who fail to consider their irrational and dark sides are like captains who blindly plow their ships into a field of icebergs: the greatest danger is hidden below the surface. Avolio et al. (2004) define authentic leaders as those who are "deeply aware of how they think and behave and are perceived by others as being aware of their own and others' values/moral perspectives, knowledge, and strengths; aware of the context in which they operate; and who are confident, hopeful, optimistic, resilient, and of high moral character" (p. 4). In confronting, with honesty, ones inner theatre, an individual is then free to discover and express his or her true self. Moreover, authenticity reflects back onto others. Through increased self-awareness, self-regulation, and positive modeling, authentic leaders can foster the development of authenticity in followers and provide them with a sense of purpose and self-determination. In turn, followers' authenticity contributes to their well-being and the attainment of sustainable and veritable performance (George et al., 2007). However, the link between authentic leadership and sustained organizational performance needs to be clarified. A leader may be authentic in his or her actions, but this might not necessarily lead to better organizational performance, nor be transmitted to followers. What are the factors which facilitate this transmission? In what ways can authenticity have a positive effect on organizational performance and can these positive effects be transmitted to followers? Another interesting issue is the mitigating role of culture on the transmission of authentic behavior from leaders to followers all across the organization. Links can also be made to help executives, through group coaching sessions, to ask the basic questions of who they are, what they believe in, and why they do what they do and in doing so, discover their authentic side. Group coaching has proven to be highly effective in creating authentizotic organizations, places where people are authentic, feel at their best – and give their best (Kets de Vries, 2001).
3. *Group coaching for leadership development*: Leadership coaching, which most commonly takes the form of one-on-one interactions between an executive and a coach, has changed the way many progressive organizations view professional and personal growth and development. It is a specific form of intervention that can be carried out strategically with individuals, teams, or an entire organization. While one-on-one coaching certainly has its benefits, leadership group (or team) coaching – in essence, an experiential training ground for learning to function as a high-performing team – is a great antidote to organizational silo formation and thinking, and a very effective way to help leaders become

more adept at sensing the hidden psychodynamic undercurrents that influence team behavior. Its aim is to direct a group of people toward a specific, mutually determined goal and to accelerate organizational performance by providing focus and awareness; it fosters strategic agility. Leadership group coaching sessions provide a safe space for honest and open explorations and confrontations wherein teams challenge and reassess their assumptions about themselves and others and in doing so, understand why they behave the way they do and why the team as a whole behaves the way it does. They undergo a cohesive experience, bringing the team closer together, not only in terms of resolving conflict and achieving mutual understanding, but also increasing shared accountability and renewed commitment (Kets de Vries et al., 2007). When instilled as part of its leadership development culture, team-oriented coaching cultures are like networked webs in the organization, connecting people laterally in the same departments, across departments, between teams, and up and down the hierarchy. However, more attention has to be paid to differentiating the quality and effects of the many leadership coaching programs currently available to determine those that have viable effects on sustained organizational performance and to measure such effects once executives take it out of the group coaching sessions and into their organizational life.

Leaders fulfill many different roles in people's imagination. They are catalysts of change; they are symbols; they are objects of identification; and they are scapegoats when things go wrong. Leaders are also prone to hubris. As Napoleon (an expert on the topic of hubris as well as leadership) once said, "Glory is fleeting, but obscurity lasts forever." All leaders are susceptible to the darker sides of power. The most effective leaders, however, are the ones who know how to balance action with reflection by using self-insight as a restraining force when the sirens of power are beckoning.

See also: Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Charisma and Charismatic; Charisma, Social Aspects of; Corporate Governance; Elites: Sociological Aspects; Leadership; Organizational Behavior, Psychology of; Organizational Climate; Organizational Decision Making; Organizations and the Law; People in Organizations; Weber, Max (1864–1920).

Bibliography

Avolio, B., Luthans, F., Walumbwa, F.O., 2004. Authentic Leadership: Theory-Building for Veritable Sustained Performance. Working Paper. Gallup Leadership Institute, University of Nebraska, Lincoln.

Bales, R.F., 1958. Task roles and social roles in problem-solving groups. In: Maccoby, E.E., Newcomb, T.M., Hartley, E.L. (Eds.), Readings in Social Psychology, third ed. Holt, New York.

Bass, B.M., Avolio, B.J., 1993. Transformational Leadership: A Response to Critiques in Leadership Theory and Research Perspectives and Directions. Academic Press, New York.

Bass, B.M., Bass, R. (Eds.), 2008. The Bass Handbook of Leadership: Theory, Research, and Managerial Applications, fourth ed. Free Press, New York.

Bass, B.M., Avolio, B.J., 1993. Transformational Leadership: A Response to Critiques in Leadership Theory and Research Perspectives and Directions. Academic Press, New York.

Bennis, W., Nanus, B., 1985. Leaders: The Strategies for Taking Charge, first ed. Harper & Row, New York.

Bryman, A., Collinson, D., Grint, K., Jackson, B., Uhl-Bien, M. (Eds.), 2011. The Sage Handbook of Leadership. Sage Publications Ltd, London.

Burns, J.M., 1978. Leadership, first ed. Harper & Row, New York.

Calder, B.J., 1977. An attribution theory of leadership. In: Straw, B.M., Salancik, G.R. (Eds.), New Directions in Organizational Behavior. St. Clair Press, Chicago, pp. 179–204.

Conger, J.A., Kanungo, R.N., 1998. Charismatic Leadership in Organizations. Sage, Thousand Oaks, CA.

Crossan, M., Mazutis, D., 2008. Transcendent leadership. In: Rosenbach, W.E., Taylor, R.L., Youndt, M.A. (Eds.), Contemporary Issues in Leadership, seventh ed. Westview Press, Boulder, CO, pp. 51–68.

Fiedler, F.E., 1967. A Theory of Leadership Effectiveness. McGraw-Hill, New York.

Gardner, J.W., 1965. The antileadership vaccine. In: Rosenbach, W.E., Taylor, R.L., Youndt, M.A. (Eds.), Contemporary Issues in Leadership, seventh ed. Westview Press, Boulder, CO, pp. 287–295.

George, B., Sims, P., McLean, A.N., Meyer, D., 2007. Discovering your authentic leadership. In: HBR's 10 Must Reads on Leadership. Harvard Business School Publishing Corporation, Boston, MA, pp. 163–177.

HBR's 10 Must Reads on Leadership, 2011. Harvard Business School Publishing Corporation, Boston, MA.

Heifetz, R.A., Laurie, D.L., 1997. The work of leadership. In: HBR's 10 Must Reads on Leadership. Harvard Business School Publishing Corporation, Boston, MA, pp. 57–78.

Hogan, R., Curphy, G.J., Hogan, J., 1994. What we know about leadership-effectiveness and personality. American Psychologist 49 (6), 493–504.

Kets de Vries, M., 2006. The Leadership Mystique: Leading Behavior in the Human Enterprise, second ed. FT/Prentice Hall, London.

Kets de Vries, M., 2008. Decoding the team conundrum: the eight roles executives play. Organizational Dynamics 36 (1), 28–44.

Kets de Vries, M., Balazs, K., 2011. The shadow side of leadership. In: Bryman, A., Collinson, D., Grint, K., Jackson, B., Uhl-Bien, M. (Eds.), The Sage Handbook of Leadership. Sage Publications Ltd, London, pp. 380–392.

Kets de Vries, Manfred, F.R, 2001. Creating authentizotic organizations: well-functioning individuals in vibrant companies. Human Relations 54 (1), 101–111.

Kets de Vries, M., Korotov, K., Florent-Treacy, E., 2007. Coach and Couch: The Psychology of Making Better Leaders. Palgrave Macmillan, Hampshire, UK.

Kets de Vries, Manfred, F.R., Florent-Treacy, E., 1999. The New Global Leaders: Richard Brandson, Percy Barnevik, and David Simon and the Making of the International Corporation, first ed. Jossey-Bass, San Francisco.

Martin, A., 2007. The Changing Nature of Leadership. Centre for Creative Leadership.

Meindl, J.R., Ehrlich, S.B., Dukerich, J.M., 1985. The romance of leadership. Administrative Science Quarterly 30, 78–102.

Osland, J., Bird, A., Osland, A., Mendenhall, M., 2006. Developing global leadership capabilities and global mindset: a review. In: Stahl, G., Bjorkman, I. (Eds.), Handbook of Research in International Human Resource Management, pp. 197–222.

Pfeffer, J., Salancik, G.R., 1978. The External Control of Organizations: A Resource Dependence Perspective. Harper & Row, New York.

Rosenbach, W.E., Taylor, R.L., Youndt, M.A., 2012. Contemporary Issues in Leadership, seventh ed. Westview Press, Boulder, CO.

Shamir, B., House, R.J., Arthur, M.B., 1993. The motivational effects of charismatic leadership: a self-concept based theory. Organization Science 44, 577–594.

Tannenbaum, R., Schmidt, W.H., 1958. How to choose a leadership pattern. Harvard Business Review 36, 95–101.

Yammarino, F., Dionne, S., Chun, J., Dansereau, F., 2005. Leadership and levels of analysis: a state of the art review. The Leadership Quarterly 16 (3), 79–919.

Learning and Instruction: Social-Cognitive Perspectives

Felice Carugati, Alma Mater Studiorum Università di Bologna, Bologna, Italy
Anne-Nelly Perret-Clermont, Université de Neuchâtel, Neuchâtel, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, we try to expand the lenses classically used in social psychology of development, and in particular, in the post-Piagetian tradition, to recent contributions of social and cognitive dynamics in development and learning. Psychological development has to be redefined as involving socially framed, culturally mediated, and interpersonally negotiated processes, and the dynamic relation between the person, others, objects, and instruments that are reconfigured through teaching–learning activities. The units of analysis, besides the traditional focus on the individual and/or isolated cognitive event, also include nowadays peer interaction and partners' roles, dialogical processes, argumentation, and specific institutional features of human practices, as illustrated through experimental social psychology. According to this general framework, learning and thinking appear more clearly as the collaborative result of autonomous minds confronting viewpoints and cultural artifacts (tools, semiotic mediations, tasks, division of roles) and trying to manage differences, feedbacks, and conflicts to pursue their activities. Moving from one activity to another, and from one space to another (pretest, joint activity, posttest), children have to reorganize their understanding, their language, and the organization of their social interactions.

From Individual to Collective Levels: The Social and the Psychological in Intellectual Development

In discussions about learning in the twenty-first century, there is a need for a better understanding of processes at the individual as well as at the collective level. At the same time, theoretical frameworks for learning have undergone several paradigm shifts from 'cognitive revolution' to more recent, various sociocultural approaches that have an increasingly greater impact in research and in the wider discussion as well. The units of analysis, besides the traditional focus on the individual and/or isolated cognitive event, nowadays include peer interaction, dialogicality, argumentation, roles of partner, and various institutional features of human practices. Another major issue is the status of knowledge in information society that, according to Engeström and Middleton (1996: p. 2), can be analyzed either by agency-driven microsociology-without-history or by historically relevant macrosociology-without-agency. A third way might be sociocultural approaches where human agency is a starting point and micro- and macro-cultural dynamics are included as far as they become relevant for understanding the activities participants perform *in situ* and over time.

In fact, the fruitfulness of interplay between 'the social' and 'the cognitive' was recommended a long time ago by Lévy-Bruhl who, according to Luria (1976), was among the first scholars to point out the qualitative feature of primitive thought and to treat logical processes as products of historical development.

As Moscovici (1990) suggests, there are at least two ways of approaching the interplay between 'the social' and 'the cognitive' (i.e., between social and cognitive or developmental psychology) or, in other terms, how deeply social relations and social changes are accompanied by profound transformations of thought, with regard to the appropriation of cognitive instruments both in children and in adults.

The 'Bartlett way' recommends the comparison of representations and beliefs of groups situated at different levels of evolution. In particular, Bartlett (1928) reminds us that the best introduction to the understanding of the social dynamics of cognitive constructions is to compare the novices and the experts within their own field or culture and not to compare the primitive and the ordinary member of a modern social group. Thus, the main aim of cognitive social psychology must be to carefully describe the modern and the normal in their milieu and in relation to topics that are significant for their common life.

A complementary contribution to this way can be found in Cattaneo's work (1864) about the crucial role interpersonal and group conflicts play in producing mental phenomena and mental organization.

The 'Vygostkian way,' which is also that of Piaget, passes through the child as the equivalent of a simpler or primitive and younger culture, and follows the child's development. This way therefore postulates that the social nature of human cognition appears in the individual's interiorization of social experiences. Progressively, as this process develops by a sort of causal spiral between social experiences and new cognitive tools, we can grasp the operations through which the people's consciousness is articulated with that of their partners and with the culture of which they are a part of.

Following these two different traditions, it is recommended that research in development and learning become an anthropology of the modern world, while developmental and learning psychology can be seen as the two tools for documenting the ontogenesis of the modern child.

After being separated almost to the point of losing all contact with each other, learning and developmental psychology are beginning to return to a shared line of research, which was the very starting point of developmental psychology, as Baldwin (1913) maintained: According to him, the individual is not socialized by society, but society individualizes itself in different individuals; or in other terms, society expresses itself continuously through each single individual.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92035-X

Shifting the Paradigms: From the Individual to the Societal

Studies on learning have been undergoing gradual paradigm shifts, from the overemphasis of the biopsychological functioning of the individual learner, where failure is seen as a sign of incompetence, to taking into account other social factors, either in a perspective inspired by Lewin's notion of *psychological field* or in a more macro-sociological approach. Notably, the latter has called attention to major social problems surrounding the discrepancies in learning performances in school pupils according to their social position (socioeconomic background, gender, ethnicity), educational methods and their hidden curriculum, the streaming structure of school institutions, etc., and differences between the same subjects' performances in and out of schools. At the psychosocial level, evidence has pointed to field effects on performances. Since Orne's pioneering work in 1962, many studies have shown that the assessment of the subjects' competencies is affected by social factors such as the linguistic, motivational, and attributional factors, as well as with identity characteristics of the subjects, teachers' expectations, individual and public dimensions of the situation, and the climate of cooperation or competition. The rather controversial results of these studies (see below) have called for a more integrative theoretical framework, articulating the various levels of analysis from the individual to the societal, in order to take into account the ways in which these factors interact, relying upon precise descriptions of the networks of sociocognitive processes within which subjects think, behave, and learn (e.g., in the laboratory, at home, at school, on the street) and observations of the impact these experiences have on individual and collective cognition.

This has led to a new shift from determinism to constructivism, and from product-oriented to process-oriented studies. Studies have focused on the interdependency between successful formal learning in schools and family socialization, and closer attention has been paid to the teacher's role in defining precise expectations and giving adequate feedback to the students, on the one hand, and to the communication modes and patterns elicited in the learning and testing situations on the other.

Learning as a Sociocognitive Activity: The Essential Role of Partners

It has been established that interaction, in order to be instructive, does not always have to be between a child (or novice) and an adult (or teacher/expert), but it can be just as effective if it takes place between equally novice peers, given a certain number of conditions and for certain tasks. But this called for detailed studies on the influence of social interaction on the appropriation of cognitive tools. Three generation waves of research could be documented since the early 1970s (Carugati and Gilly, 1993; Psaltis et al., 2009). Recovering the importance of conflict both in social and developmental psychology, the notion of *sociocognitive* conflict (Doise et al., 1975; Mugny et al., 1984) was introduced in order to stress the effect of the simultaneous confrontation of different individual perspectives or viewpoints during social interaction that necessitates and gives rise to their integration within a new cognitive organization. In the same vein, the notion of *social marking* (i.e., the representations of social rules pertinent to the cognitive rule for correctly solving the task – conservations of space and respect for the asymmetry of social positions between child and adult) allows showing the positive influence of societal dynamics on cognitive progress, even in the absence of direct interaction with a partner. The importance of social relations and norms as a symbolic order giving meaning to the everyday practices at school becomes apparent here, and is manipulated in order to study how symbolic order mastered by children might be effective in building cognitive tools.

The Contribution of Experimental Social Psychology of Cognitive Instruments

Developmental social psychologists have stressed the social nature of conflict. They propose that interindividual conflicts are essential for the stimulation of cognitive development. These conflicts are sociocognitive and positive for individual cognitive progress under the conditions that they involve confrontations with a partner proposing a different solution to the same problem. Some conditions are described in terms of both prerequisites and of constructive/obstacle outcomes to the confrontations.

Two levels of prerequisites are illustrated by Mugny et al. (1984). These concern cognitive competence (some basic cognitive organization is required) and social competence (some abilities to communicate and to interpret information).

Constructive confrontations are promoted by arranging tasks in such a way that different viewpoints are possible and can be confronted, even when an incorrect response is proposed by a partner (Mugny et al. (1984), *em*). Another useful strategy is to encourage controversy in procon issues while stressing the cooperative context (Johnson and Johnson, 2002).

Obstacles to constructive confrontations are involved when the dynamics of communication conflict do not prevent both partners from solving the problem through a relational compromise: Compliance with the more assertive partner, social comparison, competition/avoiding competition, threatening partner's competence, activating demonstration of performance instead of mastery.

Summing up, ample research has shown that the effect of sociocognitive conflict depends highly on the situation in which it takes place (Buchs et al., 2004).

Conflict could lead low-competent participants to an improvement in their reasoning strategies, unless a competence-threatening competition was introduced. Along the same line, it was shown that conflict between high-competent participants led to trying to defend and protect one's own self-competence and rejecting the other's viewpoint – with disruptive effects on reasoning – unless competition was hindered (Buchs and Butera, 2009). Other research demonstrated that students could learn from a disagreement with an expert (Carugati and Mugny, 1985; Quiamzade, 2007). However, if the latter makes them feel incompetent, this results in a superficial compliance without deep learning. Another line of research has demonstrated that in a context enhancing mastery goals (i.e., improving one's knowledge, to learn), conflict was beneficial for learning whereas in a context enhancing performance goals (i.e., demonstrating one's ability), it was not (Darnon et al., 2007).

What, in these situations, changes the effects of a conflict? According to some authors, these situations are characterized by the fact that they do or do not threaten self-competence (Buchs and Butera, 2009). For these authors, two types of conflicts can be distinguished: In a context in which self-competence is not threatened, the conflict can be *epistemic*, namely, it can be focused on the question of knowledge. This kind of conflict leads individuals into rethinking the issue, examining the validity of different propositions, and trying to understand the problem. This can result in high-level cognitive processes and improvement in learning. However, if self-competence is under threat, individuals focus their attention on the question of self-competence protection instead of learning, and the benefits of the conflict are lost. Authors qualified this type of conflict as a *relational* conflict, thus echoing previous research with children (Mugny et al., 1984).

In conclusion, the significance of social psychological contributions to learning and instruction and in a wider sense, to educational research appears fully grounded, even in the framework of social influence processes as a grid, to approach the complexity of educational dynamics. There is evidence that social influence *stricto sensu* is relevant for such dynamics. Students, but also teachers, are continuously faced with social interactions in which a given source tries to convince them to change their beliefs, attitudes, or behavior. The empirical works consistently show that the outcome of such encounters largely depends on a regulation that is sociocognitive or relational in nature, leading to learning and development in some cases, or to resistance in others.

It is also useful to define social influence processes *lato sensu*. This allows considering a larger variety of processes through which the social context modifies dynamics relevant to learning instruction and educational settings. For instance, the negative (but also positive) impact gender stereotypes may have on performance is exemplary to this respect, as are the multiple and reciprocal influences between teachers and pupils or students in the classroom. In all these cases, pupils, university students, and teachers regulate their motives, beliefs, attitudes, and reasoning as a function of relevant social knowledge or information that is made salient in the context.

What is also clear is that the most meaningful knowledge that steers such regulations in one direction or another is the one that is relevant to self-competence. This is partly due to the fact that the social hierarchy is at least partially legitimized by a hierarchy of competence. Those who are more competent have a basis from which they can argue that they should be on top of others. In turn, this may explain why negatively stereotyped gender/groups may feel an identity threat that undermines their performance or learning. Also, scholastic judgments are guided by social utility, and social utility is determined by the demonstration of both intellectual or academic success and conformity to social norms (e.g., internality). In all cases, identity stakes are aroused by the social salience of the competence level expressed by performance.

Social Interaction as Dynamic and Dialogical Processes

Recent advances in this area have drawn specific attention to the dialogical dimensions of thinking skills. Given that neither thinking nor context is a static reality, each affects the other via its authors' and participants' initiatives and interpretations. Joint actions and thought sharing (via language and semiotic mediations) are dynamic processes likely to be productive for knowledge creation and learning in the individual and in groups. Dynamic processes do origin on verbal interactions and collaboration. Tartas et al. (2010) contribute to summarize the studies aimed at clarifying an unresolved issue from the study of Grossen et al. (1993), i.e., the apparent failure of adult training to generate progress in the posttest of adult-trained children. (The design was the same as Grossen et al. (1993). One hundred children took part, 46 of whom formed the control condition (pre- and posttest sessions: phase I and IV). Two other experimental conditions were organized. Condition 1: The novices went through phase I (individual pretest), II (adult training), III (joint activity), and IV (individual posttest). During phase III, they were asked to perform a joint activity with a novice (not knowing that the partner was a novice). Phase IV was a posttest similar to phase I. Condition 2: Children were not trained by the adult. They only went to the phases I, III, and IV. In the joint activity, the dyads always comprised a novice and an expert: The expert could either be an expert diagnosed as such at the pretest (condition 2) or a child who has become an expert via adult training (condition 1). In phase II, a third of novices were trained by an adult in a way that put them in an active position with opportunities to externalize their spontaneous thinking. The adult adjusted to the child when bridging academic concepts in the conversations and suggesting more advanced strategies. The intention was to enable the child to appropriate a specific kind of talk about the task (Kohs cubes) via granting agency to the learner trying to adjust the support to the specificity of the problems encountered in doing the task). The overall results are interesting (Tartas et al., 2004). From a cognitive viewpoint, social interaction with a more advanced peer was as efficient as a condition/sequence consisting of child-centered adult training followed by interaction with a novice peer. From a social viewpoint, the authors observed that this horizontal relationship between peers offered the novices opportunities for progress.

The results show that there were no major differences across the experimental and control conditions, except for the adult-trained children: They changed level significantly more than other groups. Theoretically stated, it has been shown that there are two kinds of competencies possibly acquired by children even after the same adult training: surface or deep. Deep competencies are acquired when children do not simply internalize the semiotic means offered by the adult, but when they engage in a more profound appropriation of the same means through thinking at the higher level.

Children diversely interpret the social role allowed by the experimenter. Some children are doing things their own way, for instance, autonomous but not very social. They have a limited understanding of what the experimenter means by 'working together.' Some other children seem to have understood that they had to learn to role-play what the expert-adult had done during phase II and to behave as such in front of their peer. However, some children misunderstood the adult's intents and they played the role of unqualified observers, and not the role of qualified observer and helper that the adult had taken.

The relationship to the peer was also diversely understood by children. Some limited themselves to expressing their own viewpoint while others looked at their peer partner's production and commented on it. They participated in a real dialogue with the peer, they started gaining resources that allowed them an opportunity to try out during the peer activity what they had learned with the adult, eventually then master it and reuse it successfully at the individual posttest phase. The instructions given by the adult were the same, but the 'thinking spaces' children constructed were quite different, probably because in front of this 'Brunerian' adult who was scaffolding their activity (enrolling them in the task, alleviating difficulties, expliciting verbally the names of the elements and the actions undertaken, etc.) the children had understood that their duty was to make explicit their viewpoints and strategies.

To summarize, learning and thinking appear more clearly as a collaborative result of autonomous minds confronting viewpoints and cultural artifacts (tools, semiotic mediations, tasks, division of roles) and trying to manage differences, feedbacks, and conflicts to pursue their activities. Moving from one activity to another, from one space to another (pretest, joint activity, posttest), children have to reorganize their understanding, their language, and the organization of their social interactions.

A Specific Type of Social Interaction: Argumentation

Argumentation is of interest to researchers who are concerned with the social and cognitive processes that promote learning. However, learning, argumentation, and learning by arguing raise theoretical and methodological questions: How and when do learning processes develop in argumentation? Is it the case for all subjects? How does one design effective argumentative activities? How can the argumentative efforts of pupils be sustained? What are the psychological issues involved when arguing with others? How can what the learners produce be analyzed and evaluated? The argumentative activity requires specific intellectual and social skills, and it is often emotional and demanding. Introducing argumentative activities in educational settings is not yet common. It requires attention at different levels. The complex argumentation skills must be given opportunities to develop in the growing child. At the interpersonal level, argumentation means confronting other people's perspectives. People often avoid these kinds of situations, which they tend to perceive as a risk to the self and to the relationship. At the institutional level, argumentative activities are sometimes considered time consuming when curricula are already overloaded. These activities require special social skills from the teachers, as well as ad hoc teacher training and assessment practices. At the cultural level, argumentation means the acceptance that social harmony is not threatened by the expression of a plurality of opinions; that assertions have to be backed up; that authority is not sufficient; and that discussions are permitted even when relationships are asymmetrical. As a result of this complexity, it is not possible for teachers to just improvise argumentation-based learning activities in the classroom. Precise design and adaptive management are needed.

An interesting line of research is the development of argumentation theories in the contemporary epistemological scene. Rigotti and Greco Morasso (2009) consider, in particular, the pragma-dialectical approach for its focus on the theoretical kernel of the discipline and for systematically eliciting, from this, the connected methodological implications. The key notion of argument is specified by comparing it to the apparently near notion of demonstrative proof. Analogies and differences are brought to light, and the rather fuzzy but challenging and fundamental notion of reasonableness is identified as denoting the main value at stake in argumentative interactions. The authors propose a model of argumentative intervention in which argumentation is conceived as a particular type of communicative interaction. The model aims both at producing and at analyzing/evaluating argumentative interventions. The fundamental claim is that assuring the quality of argumentation implies contributing to a healthy social consensus and promoting cultural development at the individual and collective levels.

In the volume edited by Muller Mirza and Perret-Clermont (2009), argumentation is examined as a psychosocial practice embedded in institutional, historical, and cultural contexts. Argumentation occurs when the conversation flow is disrupted by a disagreement, a question, or an alternative hypothesis. It is not easy to develop this peculiar communication, as it entails complex issues at the personal and interpersonal levels. Even though they are, in reality, interwoven, several dimensions are distinguished. At the cognitive and individual level, the questions include the following: What are the cognitive prerequisites for engaging in an argumentative interaction? How does the development of argumentative skills take place in children? Beyond the individual level the authors take into consideration other dimensions that are important, such as the relational and dialogical aspects of argumentation, the status of the partners, and the characteristic of the audience. Developmental, social, and sociocultural approaches in psychology are thus convened in order to construct a better understanding of this complex practice.

Schwarz (2009) provides multiple perspectives on the intricate relationship between argumentation and learning. Different approaches to learning impinge on the way argumentation is conceived: as a powerful vehicle for reaching shared understanding, as a set of skills pertaining to critical reasoning, or as a tool for social positioning. Each perspective has harvested empirical studies that have stressed the importance of argumentation in learning. In spite of the pluralistic stance adopted, Schwarz attempts to draw connections between the findings obtained in the different subject areas and perspectives: In mathematics, studies are presented that show deep gaps between argumentation and proof; in science, experimental studies are reviewed to examine whether and how argumentation promotes conceptual change; in history, Schwarz considers the role of argumentation in challenging narratives and in claiming a position; and lastly, the article describes the new wave that characterizes civic education programs geared toward the instillation of argumentative practices in democratic citizenship.

Argumentation may be a useful tool at different levels of educational systems. According to the claim that education is a dialogic process in which both the talk between teachers and

learners and the talk among learners have important roles to play, Mercer (2009) describes some classroom-based research in the primary school which has enabled teachers to encourage the development of children's use of spoken language for thinking and arguing effectively together, providing empirical support for the relationship among thought, language, and social activity, as claimed by the Russian psychologist Lev Vygotskij.

At the level of university students, Andriessen (2009) presents one case of using interactive media for supporting collaborative argumentation. The discussion is descriptive, focusing on the scenario and the tools that are used and on examples of actual discussions by students. Some basic mechanisms of employing argumentation are illustrated by students using computer tools (chat, forums, graphical tools) for producing an argumentative essay.

By drawing upon existing theoretical and empirical resources to discuss the successes and difficulties encountered in trying to introduce or sustain argumentative activities in learning settings, this approach contributes to the promotion of a large program of research. Considering argumentation as a key activity at the heart of many developmental processes, in individuals and in society, the study of argumentation opens the way to a deeper reconsideration of teacher training, curricula, and also of the nature of human knowledge and its potential advancements.

Classroom Learning and Teacher's Role

One requirement of group tasks at primary school level is that they are controversial; that is, amenable to different perspectives. This has been confirmed in research by Cohen (1994) that compares controversial with noncontroversial tasks. However, granted the possibilities provided by controversy, further considerations are also relevant. First, research suggests that all group members must believe that both their own *and* their partners' contributions are important. According to Meyers (1997), individuals exert less effort in groups when they understand their work as not embedded in institutional, historical, and cultural contexts.

By contrast, when students perceive their contributions to be original and significant, they continue to participate even if their work remains anonymous. In addition, individuals sometimes reduce their efforts to match the level at which they believe other group members are contributing. Thus, children may withdraw their participation to avoid the possibility of being exploited by 'social loafers' within the group (Liden et al., 2004). Second, it is accepted that tasks should be inherently group based and not amenable to completion by individuals working independently What is needed is a task that requires resources (information, knowledge, heuristic problem-solving strategies, materials, skills) that no single individual possesses, so that no single individual is likely to solve the problem or accomplish the task objectives without at least some input from others (Cheong, 2010). Third, tasks should be challenging relative to children's current level of understanding (Ofsted, 2012). With routine tasks, children tend to engage in fairly low-level thinking and interactions (Cohen, 1994). Indeed, social loafing may increase (and motivation decrease) when tasks are too easy (Liden et al., 2004).

While the need for controversy, challenge, mutual value, and group basis can be taken for granted, there is disagreement over the significance of group rewards. Slavin has argued repeatedly (e.g., 1990) that achievement is enhanced when group members are rewarded as a group. He believes that group goals and collective outcomes provide incentives for children to help each other and to encourage each other to put forth maximum effort. However, Slavin's research has been discussed on several issues (Cohen, 1994; Bridges, 2008). It is accepted that reward can help some students, but it is not effective for all. Moreover, according to Cohen (1994) offering rewards on a competitive basis, although effective in increasing motivation of team members to work together, may have negative effects on intergroup relations, more specifically on the perceptions that team members have toward other teams, or in other terms, may induce a social comparison at the intergroup/interindividual levels and even a stereotype threat with their negative impact.

According to Howe and Mercer (2007), research into the nuances of task design is in its infancy at present, and as it develops it will undoubtedly help to clarify why classroom dialogue remains unproductive and how this can be addressed. However, few scholars believe that the answer lies purely with the specifics of task. Over 20 years ago, Wells (1986) drew attention to difficulties with the overall *climate* of British primary schools, arguing that the *normative environment* for talk in most classrooms is not compatible with children's active and extended engagement in using language to construct knowledge.

This characterization of the classroom environment for talk is also one that emerges from more recent work by Alexander (2006). According to Alexander, classroom discourse is overwhelmingly monologic in form, as teachers typically offer children opportunities for making only brief responses to their questions.

It is interesting that Wells was writing at a time when British primary education was still heavily influenced by Plowden philosophy (cf The Plowden Report, 1967), while Alexander was addressing a more curriculum- and assessment-led context. It appears that the forces that oppose exploratory talk transcend policy changes. However, the monologic climate of most classrooms does not fully explain why, when left to work together, children only rarely use talk of an 'exploratory' kind. As demonstrated many years ago (Edwards and Mercer, 1987), the norms or ground rules for generating particular functional ways of using language in primary school – spoken or written – are rarely made explicit. It is often simply assumed that children will pick these sorts of things up as they go along (taken for granted). But while picking up the ground rules and 'fitting in' in a superficial way with the norms of classroom life may be relatively easy, this may mask children's lack of understanding about what they are expected to do in educational activities and why they should do this. What is expected in terms of behavior may be accepted without really being understood. The distinction between structures for classroom management (e.g., lining up in pairs or sitting rather than kneeling on chairs) and structures that develop learning (e.g., listening to a partner or asking a question) may not be

apparent to children. Indeed, even when the aim of talk is made explicit – talk together to decide; discuss this in your groups – teachers rarely make explicit what kind of interactions they expect to take place or discuss with their pupils' ways of using talk to engage productively in joint activities. There may be no real understanding, on the part of at least some children, of how to talk together or for what purpose. Many children may not appreciate the significance and educational importance of their talk with one another. Moreover, in the monologic environments described by Alexander (2006), they may frequently assume that the implicit ground rules in play in the classroom are such that teachers want the right answers rather than discussion.

Research has shown that, under certain conditions, interaction with peers helps children's learning and development. The expression of different views, such as alternative explanations or possible solutions to problems among children working together, seems to be particularly useful in stimulating learning and development, and it does not seem necessarily to matter if those differences are always resolved (or resolved productively) through discussion. Children can also develop important communicative skills through interaction with their peers, which again they would not learn by only taking part in conversations with adults. But, paradoxically, observational studies have shown that collaborative talk in classrooms is often unproductive and inequitable. There is no reason to believe that even young primary school children are incapable of using exploratory talk (as they have been observed to do this in informal contexts), but they often seem not to use their skills during classroom activities.

Some studies have suggested that the design of tasks can promote the use of exploratory talk, and hence more productive interaction; but changing task design does not seem to be sufficient in itself. Others suggest that the quality of collaboration can be improved if attention is given to developing an atmosphere of trust and mutual respect. There is evidence from intervention studies that this does lead to improved interaction and to improved learning, but the complexities of children's social histories suggest that it may not be sufficient. Other intervention studies support the view that the quality of interaction is significantly improved if children are helped to become more aware of how they use language as a tool for thinking together and taught some specific strategies for carrying on productive discussions. That research has recorded gains not only in children's use of exploratory talk and improved collaboration, but also in their individual skills in reasoning and academic attainment. Overall, it seems that children's social histories affect how they engage in collaborative learning activities, and those histories will embody their social learning both in and out of school (Howe and Mercer, 2007).

Although collaborative interactions and discussions are potentially very valuable for children's learning and development, that potential may only be realized if children are given structured guidance by their teachers on how to make the most of the opportunities that those activities offer.

In fact, evidence has amassed over the past 20 years to indicate that group work among children can support knowledge and understanding in elementary school (science classes). Group work has been promoted in many countries as a key component of teaching. Its relevance to science education is, for instance, a recurring theme of the teachers' professional journals. Nevertheless, as Howe (2009) points out, group work will never be sufficient to deliver the curriculum of a specific school subject; expert guidance is also essential. On the other hand, expert intervention can often undermine the value that group work is known to possess. One response to this debated point is to conclude that group work is trouble and valuable as well.

An approach by Howe (2009) revolves around the dynamics of *consensus/agreement* as a way to achieve positive cognitive outcome during group interaction. Several decades after the introduction of the theory of sociocognitive conflict (Doise et al., 1975; Mugny et al., 1984) it is acknowledged in an indirect way the role of divergence of initial opinions, as a starting point for a simultaneous comparison of different individual perspectives or viewpoints during social interaction that might escalate to their integration within a new cognitive organization. The key issue in Howe's research is that it requires children to discuss and achieve (or not) consensus about relevant factors during the first stage of experiments (shadow size, rate of cooling), as a prelude to subjecting consensual vs 'ideas people have' positions to empirical test during the second phase with the expert guidance. Besides the richness of the results, it seems worth noting that achieving consensus creates receptivity to expert guidance. But a further questions are raised: It remains uncertain whether by virtue of addressing consensual positions, expert guidance is seen as serving children's needs rather than having authority over these needs, and even if this did happen, whether this was relevant for supporting children and children's outcomes, both/either at the group and/or individual level.

Summing up, when studies highlighting the challenge posed by expert guidance in group-based work with children about science subjects have been reviewed (Webb, 2009), they uncover multiple dimensions of the expert's role in fostering beneficial group dialogue, including preparing students for collaborative work, forming groups, structuring the group-work task, and influencing student interaction through teachers' discourse with small groups and with the class. Common threads through the research are the importance of students explaining their thinking, and teacher strategies and practices that may promote student elaboration of ideas. In conclusion, teachers have many roles to perform when using small-group work in the classroom, including preparing students for collaborative work, making decisions about the group-work task and the composition of groups, making decisions about the structure and requirements of group work, monitoring groups' functioning and intervening when necessary, and helping groups reflect on and evaluate their progress.

See also: Apprenticeships and School Learning: General Considerations; Co-constructivism in Educational Theory and Practice; Cognitive Development: Mathematics Learning and Instruction; Cooperative Learning in Schools; Learning Theories and Educational Paradigms; Learning to Learn; Piaget's Theory of Human Development and Education; Situated Learning (Learning In Situ); Vygotsky's Theory of Human Development and New Approaches to Education.

Bibliography

Alexander, R., 2006. Towards Dialogic Teaching: Rethinking Classroom Talk. Dialogos, Cambridge.

Andriessen, J.E.B., 2009. Argumentation in higher education: examples of actual practices with argumentation tools. In: Muller Mirza, N., Perret-Clermont, A.-N. (Eds.), Argumentation and Education: Theoretical Foundations and Practices. Springer, Dordrecht Heidelberg London New York, pp. 195–214.

Baldwin, J.M., 1913. From John Locke to the Present Time. History of Psychology, vol. 2. Watts, London.

Bartlett, F., 1928. Psychology and Primitive Culture. MacMillan, New York.

Bridges, D., 2008. Evidence-based reform in education: a response to Robert Slavin. European Educational Research Journal 7 (1), 129–133. http://dx.doi.org/10.2304/eerj.2008.7.1.129.

Buchs, C., Butera, F., 2009. Is a partner's competence threatening during dyadic cooperative work? It depends on resource interdependence. European Journal of Psychology Education 24 (2), 145–154.

Buchs, C., Butera, F., Mugny, G., Darnon, C., 2004. Conflict elaboration and cognitive outcomes. Theory into Practice 42, 23–30.

Carugati, F., 2004. Learning and thinking in adolescence and youth: how to inhabit new provinces of meaning? In: Perret-Clermont, A.N., Pontecorvo, C., Resnick, L., Zittoun, T., Burge, B. (Eds.), Joining Society: Social Interaction and Learning in Adolescence and Youth. Cambridge University Press, Cambridge, pp. 119–140.

Carugati, F., Gilly, M., 1993. The multiple sides of the same tool: cognitive development as a matter of social construction of meaning. In: Carugati, F., Gilly, M. (Eds.), Everyday Life, Social Meanings and the Social Construction of Cognitive Functioning. European Journal of Psychology of Education (special issue) VIII (4) pp. 345–354.

Carugati, F., Mugny, G., 1985. La théorie du conflit socio-cognitif. In: Mugny, G. (Ed.), Psychologie sociale du développement cognitif. Peter Lang, Berne, pp. 57–70.

Cattaneo, C., 1864. Dell'antitesi come metodo di psicologia sociale. Il Politecnico, XX, 262–270.

Cheong, C., 2010. From group-based learning to: a metacognitive approach to project-based group supervision. Informing Science: The International Journal of an Emerging Transdiscipline 13, 73–86.

Cohen, E.G., 1994. Restructuring the classroom: conditions for productive small groups. Review of Educational Research 64, 1–35.

Darnon, C., Harackiewicz, J.M., Butera, F., Mugny, G., Quiamzade, A., 2007. Performance-approach and performance-avoidance goals: when uncertainty makes a difference. Journal of Personality and Social Psychology B 33, 813–827.

Doise, W., Mugny, G., Perret-Clermont, A.-N., 1975. Social interaction and the development of cognitive operations. European Journal of Social Psychology 5, 367–383.

Edwards, D., Mercer, N., 1987. Common Knowledge: The Development of Understanding in the Classroom. Methuen/Routledge, London.

Engeström, Y., Middleton, D. (Eds.), 1996. Cognition and Communication at Work. Cambridge University Press, Cambridge.

Grossen, M., Liengme Bessire, M.-J., Iannaccone, A., Perret-Clermont, A.-N., 1993. Modes d'acquisition de l'expertise et interactions sociales entre enfants (Modes of acquiring expertise and social interaction among children). Rapport de recherche du projet 'Perception de l'expertise et interactions sociales chez l'enfant' (Report and documents from the research project 'Perception of expertise and social interaction among children'). Institut de Psychologie de l'Université de Neuchâtel. n. 2.

Howe, C., 2009. The role of the teacher in the transformation of knowledge in classroom interaction. In: Schwarz, B., Dreyfus, T., Hershkowitz, R. (Eds.), Transformation of Knowledge through Classroom Interaction. Routledge Earli, London and New York.

Howe, C., Mercer, N., 2007. Children's Social Development, Peer Interaction and Classroom Learning. Interim reports, Research Survey 2/1b. University of Cambridge. http://www.primaryreview.org.uk/.

Johnson, D.W., Johnson, R.T., 2002. Social interdependence theory and university instruction. Theory into practice. Swiss Journal of Psychology 61, 119–129.

Liden, R.C., Wayne, S.J., Jaworski, R.A., Bennet, N., 2004. Social loafing: a field investigation. Journal of Management 30 (2), 285–304.

Luria, A., 1976. Cognitive Development. Harvard University Press, Cambridge, MA.

Mercer, N., 2009. Developing argumentation: lessons learned in the primary school. In: Muller Mirza, N., Perret-Clermont, A.-N. (Eds.), Argumentation and Education: Theoretical Foundations and Practices. Springer, Dordrecht Heidelberg London New York, pp. 177–194.

Meyers, S.A., 1997. Increasing student participation and productivity on small-group activities for psychology classes. Teaching of Psychology 24, 105–115.

Moscovici, S., 1990. Social psychology and developmental psychology: extending the conversation. In: Duveen, G., Lloyd, B. (Eds.), Social Representations and the Development of Knowledge. Cambridge University Press, Cambridge, pp. 164–185.

Mugny, G., De Paolis, P., Carugati, F., 1984. Social regulations in cognitive development. In: Doise, W., Palmonari, A. (Eds.), Social Interaction in Individual Development. Cambridge University Press Cambridge, pp. 127–146.

Muller Mirza, N., Perret-Clermont, A.-N., 2009. Psychosocial processes in argumentation. In: Muller Mirza, N., Perret-Clermont, A.-N. (Eds.), Argumentation and Education: Theoretical Foundations and Practices. Springer, Dordrecht Heidelberg London New York, pp. 67–90.

Ofsted, (Office for Standards in Education), 2012. Challenging Tasks Matched to Pupils' Learning Needs. http://www.fromgoodtooutstanding.com/2012/02/ofsted-2012-challenging-tasks-matched-to-pupils-learning-needs.

Orne, M.T., 1962. On the social psychology of the psychological experiment: with particular reference to demands characteristics and their implications. American Psychologist 17, 776–783.

Psaltis, C., Duveen, G., Perret-Clermont, A.-N., 2009. The social and the psychological: structure and context in intellectual development. Human Development 52, 291–312.

Quiamzade, A., 2007. Imitation and performance in confrontation between competent peers: the role of representation of the task. European Journal of Psychology of Education 22 (3), 243–258.

Rigotti, E., Greco Morasso, S., 2009. Argumentation as an object of interest and as a social and cultural resource. In: Muller Mirza, N., Perret-Clermont, A.-N. (Eds.), Argumentation and Education: Theoretical Foundations and Practices. Springer, Dordrecht, Heidelberg, London, New York, pp. 9–66.

Schwarz, B.B., 2009. Argumentation in learning. In: Muller Mirza, N., Perret-Clermont, A.-N. (Eds.), Argumentation and Education: Theoretical Foundations and Practices. Springer, Dordrecht, Heidelberg, London, New York, pp. 91–126.

Slavin, R.E., 1990. Cooperative Learning: Theory, Research and Practice. Allyn and Bacon, Needham Heights.

Tartas, V., Perret-Clermont, A.-N., Marro, P., Grossen, M., 2004. Interactions sociales et appropriation des stratégies par l'enfant pour résoudre un problème: quelles méthodes? Bulletin de Psychologie. Numéro special "Développement et fonctionnement: perspective historico-culturelle" 469 (57), 111–115.

The Plowden Report, 1967. The Plowden Report. Children and Their Primary Schools. A Report of the Central Advisory Council for Education (England). Her Majesty's Stationery Office, London.

Webb, N.M., 2009. The teacher's role in promoting collaborative dialogue in the classroom. British Journal of Educational Psychology 79, 1–28.

Wells, G., 1986. The Meaning Makers. Hodder and Stoughton, London.

Learning and Memory, Neural Basis of

Hans J Markowitsch and Angelica Staniloiu, University of Bielefeld, Bielefeld, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Learning is described as a universal attribute of animals that is accompanied by diverse changes in the nervous system. On the brain level, memory is seen as being represented in distributed neural networks. On the behavioral level, memory is seen as divided along a time- and a content-based axis (short-term, working, long-term memory; procedural memory, priming, perceptual and semantic memory (knowledge system), and episodic-autobiographical memory). Neurological and psychiatric disease conditions can be used to determine brain regions that act as bottleneck structures in information processing.

Learning and memory are universal attributes of the animal kingdom and consequently express themselves in hugely differing neural systems from planarians to human beings. It is nevertheless possible to point to some common mechanisms by which information seems to be acquired, stored, retained, and retrieved by the nervous system. Among the mechanisms are changes in gene expression, synaptic plasticity, long-term potentiation or depression, synchronization of neuronal firing patterns, protein synthesis, and the formation and remodeling of neural nets. Within the mammalian nervous system, the time-dependent nature of memory and its splitting into time- and contents-related subsystems (short-term and working memory; long-term memory; priming, a higher probability of recognizing previously perceived information; procedural, mechanical, or motor-based skills; perceptual, the recognition of stimuli related to familiarity judgments; semantic (knowledge system), the representation of context-free facts; and episodic-autobiographical memory, the conjunction of subjective time, autonoetic consciousness, and the experiencing self; cf **Figure 1**) engage different neural networks and require processes of information distribution and convergence ('bottleneck structures') that modulate, filter, and associate information to varying degrees (Brand and Markowitsch, 2003). Consequently, various contributions from molecular to behavioral and cognitive neuroscience have increased our understanding of the neural bases of memory. Its representation in the brain, nevertheless, remains largely hypothetical and speculative.

Basic Learning- and Memory-Related Neural Changes

Much of our current understanding of basic memory-related neuronal changes and modifications stems from the pioneering research of Eric Kandel and his coworkers (reviewed in Kandel, 2009) who largely used an invertebrate, the marine mollusc *Aplysia*, for their studies. Initially, they investigated simple memory mechanisms such as sensitization and habituation and found that there are both short- and long-term forms hereof and that the physiological correlates of the short-term forms are facilitated neuronal responses (increases in the magnitude of the excitatory postsynaptic potentials formed between sensory and motor neurons), and that of the long-term forms (repeated stimulus applications) required new ribonucleic acid and protein synthesis, involving the growth of new synaptic connections between sensory and motor neurons. Later, they extended their work to gene expression and transcription factors that provided the genetic switch for memory consolidation and storage. A specific protein, named CREB, which stands for cAMP-responsive element binding protein, became a new standard label within this kind of research. cAMP means cyclic-adenosine monophosphate, a substance important for providing energy to cells. Other workers from both the invertebrate and vertebrate side demonstrated, for example, that adenylyl cyclase is an important enzyme for short-term memory (Zars et al., 2000) and long-term potentiation (LTP) as well as long-term depression (LTD). These are important mechanisms for memory consolidation (Tronson and Taylor, 2007). There is evidence that LTP and LTD lead to changes in the morphology of nerve cells, particularly to the so-called synaptic plasticity (Kullmann and Lamsa, 2007) and to changes in dendritic spine density (Segal, 2005).

The crucial questions are whether these findings can be directly transmitted to human beings and whether they hold for different kinds of memories (Markowitsch and Staniloiu, 2012a). Here is less safe ground from current research and the focus lies more strongly on metabolic changes that can be related to brain circuitries (e.g., Sharot et al., 2007).

Memory Representation within the Mammalian Nervous System

Amnesias (*see* Amnesia: General) in humans occur after a wide range of disease conditions (see Panel 2 in Markowitsch and Staniloiu, 2012a). This fact alone implies that information processing in the mammalian brain is not controlled by a single center but, instead, by distributed networks. Furthermore, memory is also dependent on a large number of sensory, perceptual, attentive, emotional, and motivational processes each with its own anatomical substrates. The investigation of brain-damaged individuals consequently is of only limited predictive value for the action of the normal brain, as it demonstrates only what the brain cannot do without a particular region (sometimes, a brain with a damaged region is compared with a chain in which a link is missing; this implies that with the lesioning approach the functions of the damaged area are, in a way, never attainable, as the respective region is missing). Therefore, the complementary approach of studying the intact brain with neuromonitoring (electroencephalography, evoked

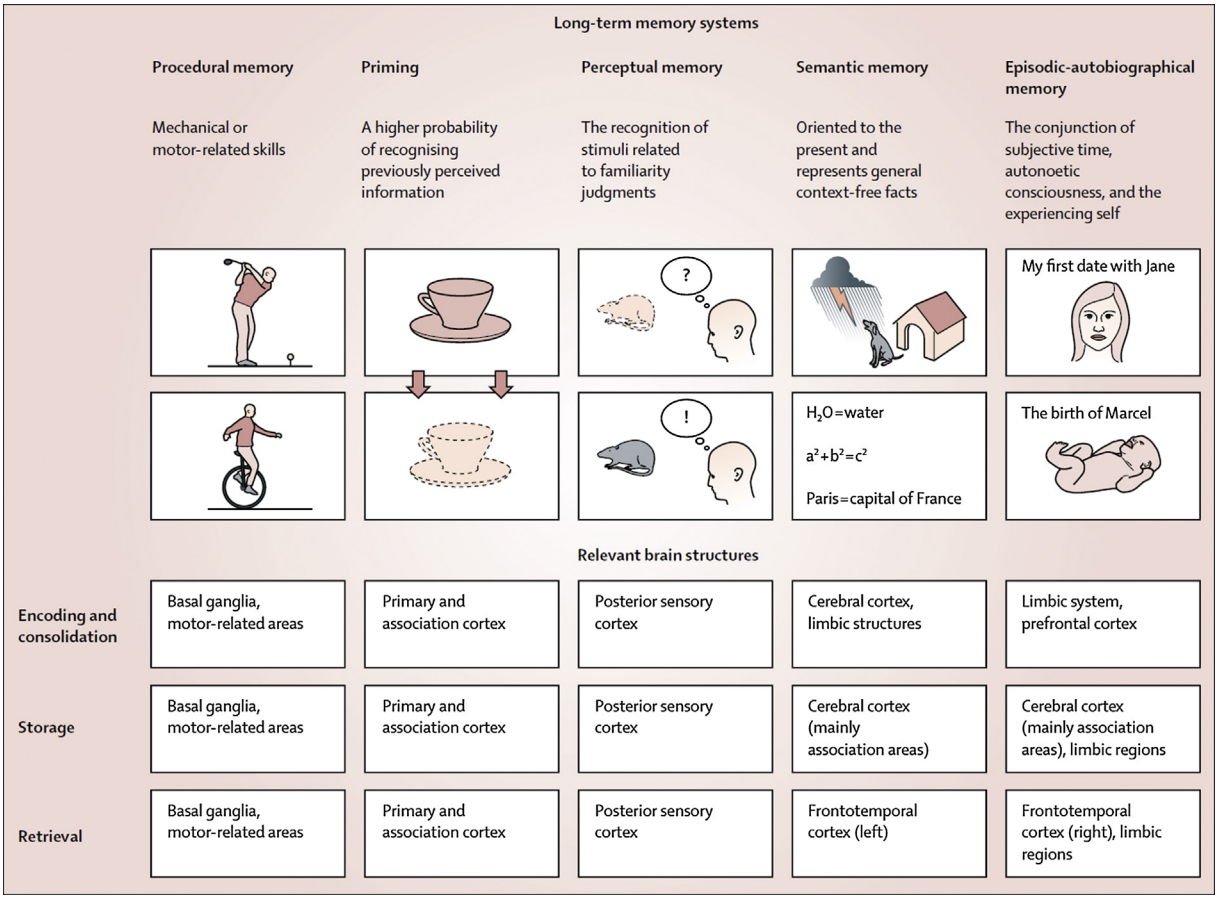

Figure 1 The five long-term memory systems and their presumed brain bases. Reproduced from **Figure 1** of Markowitsch and Staniloiu, 2012a; with permission.

potentials, magnetoencephalography) and functional brain imaging methods (positron emission tomography and especially functional magnetic resonance imaging and its variants such as diffusion tensor imaging) provides the necessary completion for inferences on the action of the brain.

In spite of the present-day range of methodological approaches for studying learning and memory representations in the brain, the old dichotomy between localizationists and antilocalizationists remains and can be perceived as one of ideology. The localizationists have a more mosaiclike view of information representation, they speak of 'centers' and 'modules'; the antilocalizationists prefer a Gestalt-like view and emphasize interactions and networks. A bridge between the two can be built by assuming a mixture of hierarchical and parallel processes of information representation (Markowitsch, 1984).

Memory Systems and Their Functional and Anatomical Correlates

The study of mammalian and in particular of human memory has to acknowledge the existence of several memory systems and their different neural representations. Aside from short-term memory, there are at least five major long-term memory systems: (1) procedural memory, (2) priming, (3) perceptual memory, (4) the knowledge system (or semantic memory), and (5) episodic-autobiographical memory (see **Figure 1**). Procedural memory is usually skill based; examples are riding a car, skiing, or playing the trumpet – that is, nonconscious, automatic processing of information. Priming refers to the fact that nonconscious presentation of a stimulus may lead to a higher probability of reidentification of it or of a similar stimulus at a later point in time. Both these systems are usually regarded as processing information implicitly or anoetically. The perceptual and semantic memory/knowledge systems are considered to be noetic and as processed mainly explicitly. Perceptual memory refers to familiarity judgments that are or can occur on a presemantic level, while semantic memory (the knowledge system) refers to general facts (world knowledge) of the kind: 'Madrid is the capital of Spain' or '12^2 is 144.' While the knowledge system consists of context-free facts, the episodic-autobiographical memory system represents context- (time-, locus-) embedded information. These episodes allow a mental time travel. Usually, episodic-autobiographical memories require a synchronous action of emotion- and fact-processing regions/networks in the brain (Markowitsch and Staniloiu, 2012a,b). One of the important distinctions between semantic and episodic-autobiographical memory is therefore that the latter allows remembering the episode and

the former simply to know the fact ("remember-know distinction"; Denkova et al., 2006). Episodic-autobiographical memories are usually processed autonoetically, that is, by reflecting on oneself.

Short-Term and Working Memory

The concept of working memory derived from the earlier one of short-term (and long-term) memories. While, however, short-term memory is seen as the online holding of information for an initial time period after first confrontation with it, working memory refers to a multicomponent, active processing of information, which also includes the transmission of already long-term stored information in a temporary buffer prior to retrieval, and which is composed of a controlling central executive system and a number of subsidiary slave systems (Baddeley, 2002). Numerous reports employing different methods from single unit recording, analyzing the consequences of brain damage to functional brain imaging, have reported relations between the prefrontal cortex and working memory (e.g., McNab and Klingberg, 2008). Some reports found that damage to left parietal regions impairs the online processing of information (cf Panel 3 in Markowitsch and Staniloiu, 2012a). Consequently, a frontoparietal network is seen as engaged in the control of attention and in processes of online working with memory.

Procedural Memory and Priming

The two principal memory systems that function independent of conscious reflection are procedural memory and priming. There are two basic forms of priming, perceptual and conceptual – perceptual priming refers to the direct match between the presently and the previously perceived stimuli, while conceptual priming refers to the same category or class of the present and the previously perceived stimuli. The distinction between these two forms is important as conceptual priming apparently is more complexly represented in the brain, recruiting more or other neural structures than perceptual priming. Consequently, conceptual priming deteriorates in age-related brain diseases and possibly even in patients with limbic system-related amnesias, while perceptual priming does not (or much less so). Limbic system is a conglomerate of cortical regions, nuclei, and fiber tracts extending from the cerebral cortex to the midbrain. Their main functions are related to the processing of emotions and memory. Conceptual priming has been related to prefrontal and even partly to medial temporal regions, while perceptual priming has much stronger affinities to unimodal areas of the posterior neocortex.

The procedural memory system is closely related to motor regions of the brain – to the basal ganglia, to portions of the frontal lobe, particularly of the premotor and motor cortex, and to cerebellar structures (some authors question, however, a specific role of cerebellar structures in procedural memory). Therefore, diseases such as Parkinson disease are taken as models – at least during the initial disease process – for selective procedural memory deficits (e.g., Cohen and Pourcher, 2007).

While not much is known about the specific brain regions involved from information encoding to information retrieval, there is some evidence that principally the same structures are activated during all stages from encoding over storage to retrieval (Cohen and Pourcher, 2007, however, found that encoding may function in a group of patients with Parkinson disease, while consolidation of procedural memories is impaired).

Brain Regions Relevant to Perceptual Memory, the Knowledge System, and Episodic-Autobiographical Memory

Perceptual memory as a noetically ('consciously') processed system has been introduced more recently than the other four memory systems. It is considered to be processed by polymodal association cortices of the posterior half of the brain. Episodic-autobiographical memory and the knowledge system are sometimes combined and then named declarative memory (Manns and Eichenbaum, 2006). Both systems have in common that they – at least during the stages of memory encoding and consolidation – rely, among others, on structures of the limbic system (Markowitsch, 1999), the episodic-autobiographical memory system most likely in addition on prefrontal structures. For retrieval of information, there seems to be a distinct lateralization; it appears that primarily right-hemispheric regions are engaged in episodic-autobiographical and primarily left-hemispheric ones in semantic-memory/knowledge system retrieval (LaBar and Cabeza, 2006; Markowitsch and Staniloiu, 2012a).

The structures of the limbic system have long been discussed as important for transferring information from short- to long-term memory. More recently, components of this system have been combined under the headings 'medial temporal lobe'; 'medial diencephalic' and 'basal forebrain memory system' and have been named bottlenecks for information transmission from short- to long-term memory (Markowitsch and Staniloiu, 2012a). Bilateral damage to structures within these systems usually leads to anterograde amnesia; that is, to an inability to store new memories long term (on a conscious level) (Markowitsch and Staniloiu, 2012a).

For the first system, the hippocampal formation was (and still is) seen as its hub; for the second, medial thalamic nuclei and the hypothalamic mammillary bodies are core structures and Korsakoff syndrome is one of the most well-known diseases associated with it. The basal forebrain system comprises cholinergic nuclei (basal nucleus of Meynert, diagonal band of Broca, and septal nuclei) and surrounding portions of the orbitofrontal cortex. In addition to various degrees of amnesia, confabulatory tendencies are prominent among patients with damage in this region (Borsutzky et al., 2010). Behavioral distinctions between the three systems are given in Table 1 of Markowitsch and Staniloiu (2012a).

There is another division of the limbic system that receives support from a number of sources; namely, the distinction between two interacting circuits: (1) the medial (or Papez) circuit centered around the hippocampus and (2) the basolateral limbic circuit (amygdaloid circuit). While Papez (cited in Markowitsch, 1999) considered his proposed circuit (hippocampal formation → fornix → mammillary bodies → mammillothalamic tract → anterior

thalamus → thalamocortical peduncles → cingulate cortex → cingulum → hippocampal formation) as principally engaged in the analysis of emotions, subsequent work showed that it plays a critical role for the transfer of information into long-term memory. The basolateral circuit is more closely related to emotional processing, and is also relevant for encoding the emotional valence of experiences. It includes the amygdala, the mediodorsal nucleus of the thalamus and associated regions, largely situated in the basal forebrain and in its extensions to anterior insular and temporopolar cortical areas plus interconnecting fibers such as the ventral amygdalofugal pathway, the anterior thalamic peduncle, and the diagonal band (Figure 2).

Bilateral damage to structures within the medial circuit is usually followed by severe memory disturbances – a fact that led researchers to speculate that amnesia might arise from an interruption of interconnections within this circuit and that the structures of this circuit are necessary for consolidation and for the transfer from short- to long-term storage. Bilateral damage to structures within the basolateral limbic circuit, on the other hand, seems to degrade memory processes by disconnecting the emotional from the mnestic properties of a given set of stimuli. For example, bilateral damage to the amygdaloid region may induce the Klüver–Bucy syndrome characterized by hyperorality, hyperphagia, agnosia, amnesia, hypersexuality, tameness/placidity, and hypermetamorphosis (Markowitsch and Staniloiu, 2011, 2012b). Since information encoding is, especially for episodic-autobiographical information, dependent on emotional content, damage to the amygdala is likely to have detrimental effects on memory. This indeed was shown in patients with Urbach–Wiethe disease (Siebert et al., 2003); such patients manifest a selective bilateral damage (mineralization) of the amygdalae. Two such patients were presented stories containing emotionally charged items and asked to recall them. They were significantly impaired because they could not show the preferential encoding of emotionally laden vs neutral items (Cahill et al., 1995).

Generally, bilateral damage to structures (including fibers such as the fornix; Staniloiu and Markowitsch, 2012) leads to severe and lasting memory disturbances. In addition, there may be disturbances of conscious reflection and self-reflection and executive functions; also, anosognosia and mood disorders may be present (Markowitsch and Staniloiu, 2012a).

Encoding and Consolidation

While encoding refers to the more early stage of information acquisition, there are different views on the beginning and end of the consolidation phase. Sometimes, it is seen as directly succeeding the encoding phase; sometimes it is viewed as a process occurring after information has already been stored long term. As both encoding and consolidation of episodic-autobiographical and semantic memories engage structures of the limbic system, they can be perceived as closely related.

There are nevertheless single cases of patients with amnesias of various origins showing a wide range of unstable memory processing periods so that it has to be concluded that there are several states and stages of memory consolidation, which depend on complex anatomicobehavioral interdependencies (Squire, 2007; Tronson and Taylor, 2007). A comparison with repeated attempts of producing a complex painting may be apt. Still, medial temporal, basal forebrain, and diencephalic regions contribute substantially to both the stages of encoding and consolidating by triggering, accompanying, and assisting the steps from initial binding to the final assignment. Especially the hippocampus has been proposed to index (previously encoded) information and assign it to specific (neo-)cortical loci. Amygdaloid and septal

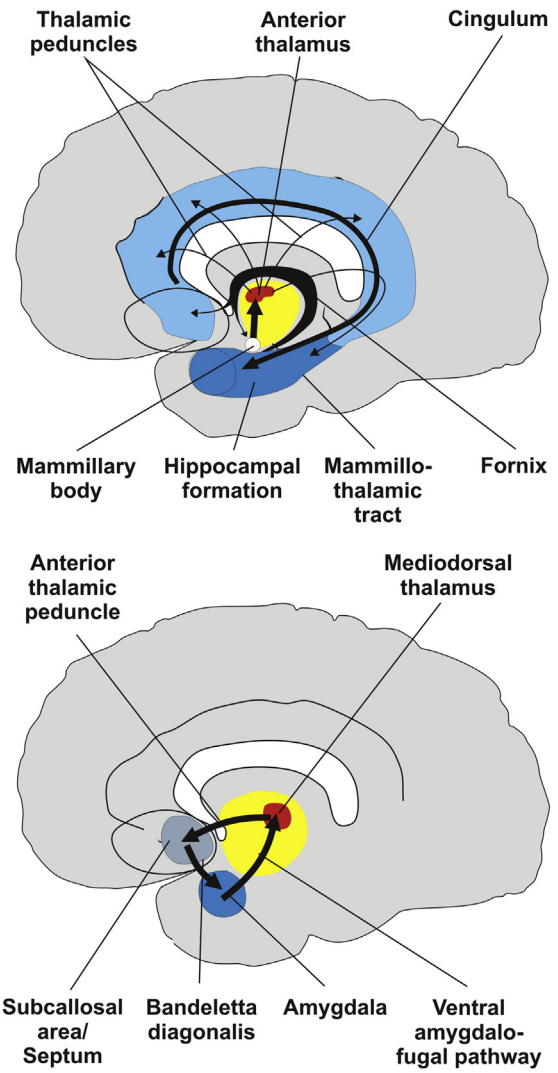

Figure 2 The medial (Papez) (top) and the basolateral limbic circuits (bottom). The Papez circuit is composed of the hippocampal formation, mammillary bodies, anterior thalamus, and (conventionally) the cingulate cortex. These regions are interconnected via the fornix, mammillothalamic tract, thalamocortical peduncles, and the cingulum. The components of the basolateral limbic circuit are the amygdala, the thalamic mediodorsal nucleus, and the subcallosal area (basal forebrain); here, interconnections are the ventral amygdalofugal pathway, the inferior thalamic peduncle, and the bandeletta diagonalis. It is assumed that both circuits are relevant especially for encoding of information associated with the episodic–autobiographical and knowledge memory systems and that the medial circuit is more strongly engaged in the cognitive and the basolateral one in the affective side of information processing.

regions may evaluate and filter for importance – that is, they screen the biological and social significance of perceived stimuli.

Aside from the more 'classic' limbic regions, functional brain imaging results have pointed to a strong role of prefrontal cortical areas in both encoding and retrieving of information (Brand et al., 2009; Dickerson et al., 2007; Kroll et al., 1997).

Storage of Information

The ways in which information is – frequently life-long – stored in the brain is one of the biggest riddles of the neurosciences. Most scientists agree that the nervous system contains or forms the medium for storage, the more concrete mechanisms vary between species – invertebrates with their more 'rudimentary' nervous system may indeed have adopted other representations than mammals. Genetic and biochemical processes may nevertheless be common denominators (Kandel, 2009). For mammalian nervous systems, information may very likely be distributed along widespread cerebral cortical networks with or without auxiliary engagement of subcortical limbic structures (Markowitsch and Staniloiu, 2012a). The expanded cortical association areas with their large number of neurons and neuronal connections are primary candidates for information representation. Especially for emotionally colored memories – such as biographical episodes – the cortical nets may engage and trigger affect-coding limbic regions for their rerepresentation or 'ecphory' (Markowitsch and Staniloiu, 2012a). (The term ecphory describes the process by which retrieval cues interact with stored information so that an image or a representation of the information in question appears.)

There are, however, a few scientists who are of the opinion that most of our memories are represented outside our nervous system (e.g., Romijn, 1997; Sheldrake, 1988). Romijn (1997) considers a submanifest order of being as the principal storage place of memories. He is of the opinion that the whole universal in its spatial–temporal configuration would be permanently present in the submanifest order of being. Of this, a small portion would continuously materialize and thereby become available for psychic processing.

Retrieval of Information

The conventional view is that retrieval activates similar (namely, mainly limbic) brain structures as does encoding. This view may, however, be false or at least only partly correct: as mentioned above, the prefrontal cortex is a primary region for information retrieval. Converging evidence from studying brain-damaged patients with neuropsychological and normal subjects with functional imaging methods points to the necessity of (inferolateral) frontal and anterolateral temporal cortical regions for triggering the networklike stored information for successful retrieval (Fink et al., 1996) (**Figure 3**). There is evidence that emotionally laden episodic-autobiographical events are preferentially retrieved via frontotemporal structures of the right hemisphere, while general facts are preferentially retrieved by corresponding structures from the left hemisphere (LaBar and Cabeza, 2006; Markowitsch and Staniloiu, 2012a). It should be noted that the frontotemporal regions are bidirectionally connected via the uncinate fascicle, which seems to grow even in adulthood (Lebel et al., 2008) and seems to be affected in psychopathy and in maltreated children leading to an overgeneral memory effect in them (Staniloiu and Markowitsch, 2012).

While a few authors even assume the temporal cortex to be the actual storehouse of information, here the view is favored

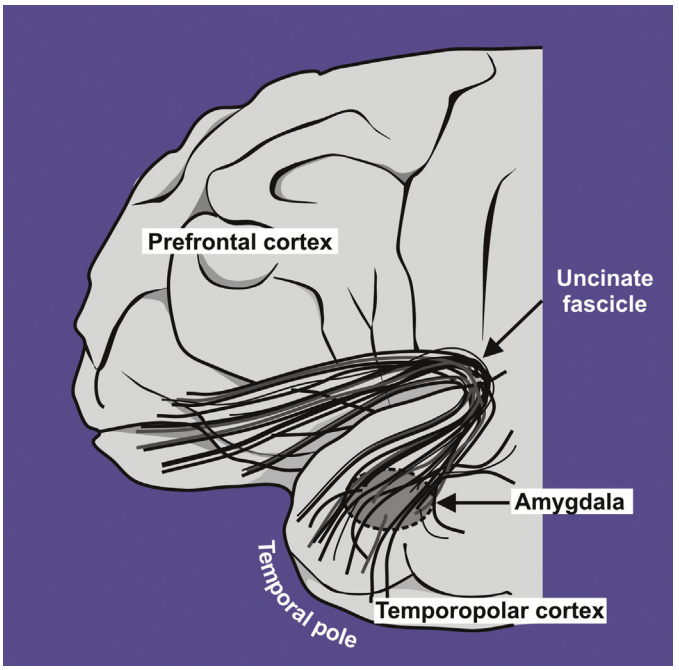

Figure 3 The regions of the temporal pole and the inferior lateral prefrontal cortex, which are interconnected via the uncinate fascicle. This regional constellation of the right hemisphere is assumed to control the retrieval of episodic-autobiographical memories, and the left-sided one that of the knowledge system.

that its anterior portion provides the most direct access to the cortical storage network in the posterior half of the brain, while the prefrontal cortex is responsible for providing the effort necessary for a successful overall neural activation. Limbic structures of the medial diencephalon and the medial temporal lobe may have supporting roles in context-rich information retrieval (Markowitsch and Staniloiu, 2012a).

Retrograde amnesia – the failure of ecphory – may occur as a consequence of a progressively decaying cortical network – as is the case in patients with cortical dementia (e.g., Alzheimer disease) or, sometimes, after a condition of severe hypoxia (e.g., after a heart attack and coma state). In this group of patients, the information is usually lost permanently. When the damage is restricted to the hippocampus and related structures in the medial temporal lobes, the retrograde amnesia may be temporally limited (covering a period of weeks in rodents, up to years or even decades in humans, leaving remote memories intact), suggesting to several authors that the hippocampus may play a role in the initial formation of new memories, and may somehow contribute to the gradual integration of new memories into neocortical memory circuits (Squire and Bayley, 2007); other authors assume a general role of hippocampal structures in episodic-autobiographical retrieval (Rekkas and Constable, 2005).

Alternatively, retrograde amnesia may be the consequence of a disrupted access to the net. Here, the expression 'mnestic block syndrome' has been coined to describe this phenomenon (Markowitsch, 2002), which can be found both after manifest brain damage in the frontotemporal cortex and after so-called functional amnesia, which is seen in patients with dissociative forms of amnesia. Such functional amnesias especially appear as a consequence of stress and trauma experiences. In this second category of patients, the mnestic block is usually selective, affecting primarily emotion-laden personal episodes, and it is assumed that the information – the engram – is in principle still in the cortical nets.

The Learning and Memory Networks – Conclusions

Learning and memory engage a wide range of brain regions – nevertheless, a conclusion that it is the total mass of the nervous system that codes information for storage would be a false one. Instead, the principles of convergence and divergence of information transmission hold throughout the nervous system and there are both modality- and material-specific channels for information encoding, storage, and even retrieval and there are bits of information that are rich in affect and consciously reflected (and consequently require wide neural nets in which individual structures act in synchrony), while others are processed only subliminally and may never gain conscious representation (Markowitsch and Staniloiu, 2012a). Furthermore, information is not represented in a static manner within the brain – mood, the environment, focused attention, internal associated cues, hormones, and many additional factors influence, modify, and rearrange our knowledge. The phenomena of state-dependent learning and retrieval and of false memories are examples for the susceptibility and vulnerability of our memories (Kühnel et al., 2008). The concept of plasticity has gained considerable support from invertebrate to human research and also the concept of lifelong learning, proposed by the social sciences, demonstrates that information representation is much less hardwired and rigid than formulated by traditional frameworks (Kim and Linden, 2007; Leal-Galicia et al., 2008).

See also: Amnesia: General; Amnesia: Psychogenic; Autobiographical Memory; Declarative Memory, Neural Basis of; Episodic Memory; Functional Brain Imaging of Language Processes; Hippocampus and Related Structures; Hypothalamus; Limbic System; Short-Term Memory: Psychological and Neural Aspects; Synaptic Mechanisms and Protein Synthesis in Memory; Temporal Lobe; Working Memory, Neural Basis of.

Bibliography

Baddeley, A.D., 2002. Is working memory still working? European Psychologist 7, 85–97.
Borsutzky, S., Fujiwara, E., Brand, M., Markowitsch, H.J., 2010. Susceptibility to false memories in patients with ACoA aneurysm. Neuropsychologia 48, 2811–2823.
Brand, M., Eggers, C., Reinhold, N., Fujiwara, E., Kessler, J., Heiss, W.-D., Markowitsch, H.J., 2009. Functional brain imaging in fourteen patients with dissociative amnesia reveals right inferolateral prefrontal hypometabolism. Psychiatry Research: Neuroimaging Section 174, 32–39.
Brand, M., Markowitsch, H.J., 2003. The principle of bottleneck structures. In: Kluwe, R.H., Lüer, G., Rösler, F. (Eds.), Principles of Learning and Memory. Birkhäuser, Basel, pp. 171–184.
Cahill, L., Babinsky, R., Markowitsch, H.J., McGaugh, J.L., 1995. The amygdala and emotional memory. Nature 377, 295–296.
Cohen, H., Pourcher, E., 2007. Intact encoding, impaired consolidation in procedural learning in Parkinson's disease. Experimental Brain Research 179 (4), 703–708.
Denkova, E., Botzung, A., Manning, L., 2006. Neural correlates of remembering/knowing famous people: an event-related fMRI study. Neuropsychologia 44 (14), 2783–2791.
Dickerson, B.C., Miller, S.L., Greve, D.N., Dale, A.M., Albert, M.S., Schacter, D.L., Sperling, R.A., 2007. Prefrontal–hippocampal–fusiform activity during encoding predicts intraindividual differences in free recall ability: an event-related functional-anatomic MRI study. Hippocampus 17 (11), 1060–1070.
Fink, G.R., Markowitsch, H.J., Reinkemeier, M., Bruckbauer, T., Kessler, J., Heiss, W.-D., 1996. Cerebral representation of one's own past: neural networks involved in autobiographical memory. Journal of Neuroscience 16, 4275–4282.
Kandel, E.R., 2009. The biology of memory: a 40 year perspective. Journal of Neuroscience 29, 12748–12756.
Kim, S.J., Linden, D.J., 2007. Ubiquitous plasticity and memory storage. Neuron 56 (4), 582–592.
Kroll, N., Markowitsch, H.J., Knight, R., von Cramon, D.Y., 1997. Retrieval of old memories – the temporo-frontal hypothesis. Brain 120, 1377–1399.
Kühnel, S., Woermann, F.G., Mertens, M., Markowitsch, H.J., 2008. Brain activations during correct and false recognitions of visual stimuli. Implications for eyewitness decisions on an fMRI study using a film paradigm. Brain Imaging and Behavior 2, 163–176.
Kullmann, D.M., Lamsa, K.P., 2007. Long-term synaptic plasticity in hippocampal interneurons. Nature Reviews in Neuroscience 8 (9), 687–699.
LaBar, K.S., Cabeza, R., 2006. Cognitive neuroscience of emotional memory. Nature Reviews in Neuroscience 7, 54–64.
Leal-Galicia, P., Castaneda-Bueno, M., Quiroz-Baez, R., Arias, C., 2008. Long-term exposure to environmental enrichment since youth prevents recognition memory decline and increases synaptic plasticity markers in aging. Neurobiology of Learning and Memory 90 (3), 511–518.
Lebel, C., Walker, L., Leemans, A., Philips, L., Beaulieu, C., 2008. Microstructural maturation of the human brain from childhood to adulthood. NeuroImage 40, 1044–1055.
Manns, J.R., Eichenbaum, H., 2006. Evolution of declarative memory. Hippocampus 16 (9), 795–808.

Markowitsch, H.J., 1984. Can amnesia be caused by damage of a single brain structure? Cortex 20, 27–45.
Markowitsch, H.J., 1999. Limbic system. In: Wilson, R., Keil, F. (Eds.), The MIT Encyclopedia of the Cognitive Sciences. MIT Press, Cambridge, MA, pp. 472–475.
Markowitsch, H.J., 2002. Functional retrograde amnesia – mnestic block syndrome. Cortex 38, 651–654.
Markowitsch, H.J., Staniloiu, A., 2011. Amygdala in action: relaying biological and social significance to autobiographic memory. Neuropsychologia 49, 718–733.
Markowitsch, H.J., Staniloiu, A., 2012a. Amnesic disorders. Lancet. http://dx.doi.org/10.1016/S0140–6736(11)61304-4. Advance online publication.
Markowitsch, H.J., Staniloiu, A., 2012b. The contribution of the amygdala for etablishing and maintaining an autonomous self and autobiographical memory. In: Yilmazer-Hanke, D. (Ed.), Insights into the Amygdala: Structure, Function and Implications for Disorders. Nova Science Publishers, Hauppauge, NY, pp. 277–318.
McNab, F., Klingberg, T., 2008. Prefrontal cortex and basal ganglia control access to working memory. Nature Neuroscience 11 (1), 103–107.
Rekkas, P.V., Constable, R.T., 2005. Evidence that autobiographic memory retrieval does not become independent of the hippocampus: an fMRI study contrasting very recent with remote events. Journal of Cognitive Neuroscience 17 (12), 1950–1961.
Romijn, H., 1997. About the origin of consciousness. A new, multidisciplinary perspective on the relationship between brain and mind. Proceedings van de Koninklijke Nederlandse Akademie van de Wetenschappen 100, 181–267.
Segal, M., 2005. Dendritic spines and long-term plasticity. Nature Reviews in Neuroscience 6, 277–284.
Sharot, T., Martorella, E.A., Delgado, M.R., Phelps, E.A., 2007. How personal experience modulates the neural circuitry of memories of September 11. Proceedings of the National Academy of Sciences of the United States of America 104 (1), 389–394.
Sheldrake, R., 1988. The Presence of the Past: Morphic Resonance and the Habits of Nature, first ed. Times Books, NY.
Siebert, M., Markowitsch, H.J., Bartel, P., 2003. Amygdala, affect, and cognition: evidence from ten patients with Urbach–Wiethe disease. Brain 126, 2627–2637.
Squire, L.R., 2007. Neuroscience. Rapid consolidation. Science 316 (5821), 57–58.
Squire, L.R., Bayley, P.J., 2007. The neuroscience of remote memory. Current Opinion in Neurobiology 17 (2), 185–196.
Staniloiu, A., Markowitsch, H.J., 2012. The splitting of the brain: a reorientation towards fiber tracts damage in amnesia. In: Schäfer, A.J., Müller, J. (Eds.), Brain Damage: Causes, Management, and Prognosis. Nova Science Publishers, Hauppauge, NY, pp. 41–70.
Tronson, N.C., Taylor, J.R., 2007. Molecular mechanisms of memory reconsolidation. Nature Reviews in Neuroscience 8, 262–275.
Zars, T., Fischer, M., Schulz, R., Heisenberg, M., 2000. Localization of a short-term memory in *Drosophila*. Science 288, 672–675.

Learning Disability and Social Work

Robyn Munford, School of Health and Social Services, Massey University, Palmerston North, New Zealand
Garth Bennie, Ministry of Education, Palmerston North, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article focuses on learning disability and social work. It outlines key debates in the field including philosophical debates, policy, and practice. Social work practice in this field has experienced significant change as models of service delivery have been transformed over time including a major change, the move from institutional to community care. Alongside of these structural changes and developments in service delivery were changes in the philosophical approaches that informed practice, for example, the emergence of key ideas such as normalization and the social model of disability, that presented challenges to thinking about disability and service delivery including social work. Central to social work practice in this field is work with families and some of the key aspects of this work are discussed in this article.

Introduction

This article focuses on learning disability and social work. It explores some of the philosophical and theoretical perspectives that have informed practice and highlights the history of these frameworks and the changes to perspectives over time. A significant development in the field of learning disability was the change from custodial care in large institutions to community care and this major change in service delivery resulted in new approaches to supporting people with learning disabilities and their families. These approaches to support and the factors informing social work practice, including the issues that may arise in practice, will also be discussed. The article concludes with a discussion of collaborative practice which provides a framework for practice that enables social workers to work alongside people with learning disabilities to promote inclusion in community life.

Definitions

The term mental handicap has been replaced by terms such as learning disability. In some countries the word intellectual disability is used as is the phrase: people with intellectual disabilities (for further discussion on the genesis of the term learning disability see the website: http://www.learningdisabilities.org.uk). In this article the term learning disability will be used. The UK Department of Health (2001) in their policy and practice documents uses the following definition:

Learning disability is described as a:

- significantly reduced ability to understand new or complex information and to learn new skills,
- reduced ability to cope independently which starts before adulthood with lasting effects on development.

In this article the phrase disabled people will also be used as this is also an accepted terminology and consistently used by disabled people themselves.

Philosophical and Theoretical Perspectives

A number of philosophical frameworks have informed the provision of support and services to people with learning disabilities. A key influence on the lives of people with learning disabilities has been the adoption of the principles of normalization which has informed service provision since the 1970s (Munford, 2012). The decade of the 1970s was a watershed year for developments in support and service provision; the key proponents of normalization, Nirje and Wolfensberger, argued for an approach that challenged the practices of institutionalization and supported people to lead culturally valued lives within their communities. In many countries around the world such perspectives were the catalyst for changes in policy and practice that led to the closure of large institutions and encouraged the move from custodial care to care in the community. Here the challenge was to support people with learning disabilities to engage in culturally normative experiences (Sullivan and Munford, 1998, 2005). Promoting this movement from institutional to community care was the advocacy by parents to have their family members living with them or in community settings; this resulted in a focus in the last few decades on developing community services that are designed to include people with learning disabilities in 'everyday' activities and to support them to fully participate in community life.

While normalization principles have been the predominant perspective informing service provision for people with learning disabilities theoretical and philosophical frameworks such as the social model of disability have also had a key role in defining the experience of disability (Oliver, 1990). Central to these ideas is understanding disability as a set of experiences rather than as a pathological condition requiring medical intervention. Impairment is viewed as a natural part of the diversity that constitutes the human condition (Munford and Bennie, 2009). The social model emerged as a dominant perspective in the 1960s as a response to the medical model (Sullivan and Munford, 1998, 2005) which had its origins in the rise of medical science, economic rationalism, and social Darwinism (Munford and Bennie, 2009: p. 210). The emphasis

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28118-X

here is on disability as an individual problem; disabled people are viewed as the victims of personal circumstance and tragedy and the response to this focuses on diagnosis, prevention, and personal adjustment. As Munford and Bennie (2009: p. 210) argue, the "medical model had a pervasive influence"; this has been well-illustrated by the large-scale institutionalization of disabled people that occurred throughout the nineteenth and twentieth centuries. Criticisms of the medical model gave rise to new perspectives and thinking about the experience of disability. The social model had a profound influence and changed the approach to service provision (Sullivan and Munford, 1998, 2005). The social model regards social reality as a product of social interactions and disability is seen to be constructed through the interactions between disabled and nondisabled people. The construction of disability is one of deviance and deficit where disabled people have stigmatized identities (Hallahan, 2010; Munford and Bennie, 2009; O'Brien and Sullivan, 2005). The social model provides an analysis of the situations of disabled people and challenges the way in which their lives have been constructed by others. Central to this is reframing the interpretation of impairment and supporting disabled people to create meaningful lives.

Under the rubric of a social model of disability the term 'disability' describes the negative experiences that people with impairments encounter in a 'disabling society' for example, marginalization, lack of opportunity, and overt discrimination (Munford and Bennie, 2009). Such an approach requires a reconceptualization of the issues facing people with disabilities as issues of citizenship and participation in the 'everyday' experiences of community life. The ideas that have constituted a social model of disability have been extended into a political model of disability in order to more adequately describe the role of social structures in defining the experience of disability. Under this approach the relationship between impairment, the construction of disability, and the social conditions (such as unemployment, poor housing and unemployment) is explored (Munford and Bennie, 2009; Stainton et al., 2010). The political model has its origins in the critical tradition and in post-modernist thinking (Munford and Sanders, 2008) and promotes a view of the world as one of irreconcilable differences. These differences arise out of power relations and conflicts between those who control resources and administer power and those who are subject to these practices.

A political model of disability seeks to understand the nature of differences and conflicts and asserts that these can only be resolved through fundamental changes to the structure, organization, and dominant values of society (Munford and Bennie, 2009: p. 211). As with the social model, a political model defines disability as something people experience rather than something people have; rather than being viewed as the consequence of impairment itself, disability is regarded as a consequence of living with impairment in a disabling society. This society is organized in such a way that it fails to take account of impairment (Oliver, 1990). Within this perspective the problem is located not within the individual, or even between individuals, but in a social milieu where structures and dominant values make disabled people the collective victims of an unknowing and individualistic society (Munford and Bennie, 2009: p. 211). Disability is therefore to be understood as a political problem, not just a medical or social problem and within this perspective disabled people have a place within which to assert their human and political rights (Sullivan and Munford, 1998, 2005).

While the political model has advanced thinking around the experiences of disability and provided a way forward in terms of political action, writers in the field also remind us not to forget about what happens to individual 'bodies' and to think more deeply about the nature of impairment and its impact on lived experience (Reeve, 2002; Stevens, 2008; Williams, 1999). Reeve (2002) argues for a perspective that incorporates a psycho-emotional dimension of disability that acknowledges disability as a social problem but that also explores "the role played by impairment and illness in the life experience of disabled people" (p. 494). Here embodiment and the lived experience of the body are central and in using Foucauldian concepts such as the 'technologies of power.' Reeve (2002: p. 495) argues that psycho-emotional dimensions of disability broaden the analysis to include what disabled people can 'be' as well as 'do.' One key 'technology of power' is that of the 'gaze' where disabled people are observed and assessed and constantly come under the scrutiny of others. For people living with learning disabilities being subject to the 'gaze' of others who determine what they can and cannot do becomes part of their daily lived experience; for example, this is often the experience of parents with learning disabilities (Booth and Booth, 2003; Llewellyn and McConnell, 2002). Reeve (2002: p. 504) suggests that "acknowledging the personal experiences of living with impairment and disability is politically unifying rather than divisive because it enables recognition of the range of disability experiences, and how these are mediated by impairment and other interesting identities such as gender and age."

Williams (1999) also argues for a broadening of the social model of disability by rewriting the body back into disability politics; such an approach counteracts the tendency to portray a "homogeneity of interest" (p. 810) that masks the diversity of the experiences of disabled people. He argues that to understand the diverse and different experiences of disabled people we have to understand the experiences of "real impaired bodies" (p. 811). This understanding provides insights into the struggles for personal coherency within adversity and builds a political understanding that contains the seeds of resistance and emancipation (p. 811). This view of disability is as an 'emergent' property which results from an interplay between physiology, structure, and sociocultural systems that represent a 'dynamic' relationship between individuals and societal contexts. It is in this way that 'a critical view of agency' is retained including the "transformatory potential of embodied social praxis" (p. 813). These perspectives have much to offer the political and social change projects for people with learning disabilities; such approaches provide a theoretical and practice framework that incorporates understanding about the daily struggles emerging from impairment alongside an understanding of the context of social life that either restricts or expands opportunities. As Hallahan (2010: p. 127) suggests, a full account of 'real' lived experiences requires a nuanced approach that works with multiple meanings and experience of disability including inter alia: embodiment and what impairment means in the daily lives of disabled people; the

oppression experienced within service systems that have functioned to exclude people from participation; issues that arise from the response of others to the impairment; and underlying all of these experiences, "the desire to be honoured, respected, and supported alongside [other] … citizens."

All of the various perspectives have, over time, influenced policy frameworks and service provision and it is to a discussion of these that we now turn.

Policy Frameworks and Service Provision

In many countries around the globe the interests of disabled people are now firmly placed on policy agendas and much thought is given to how services should be delivered (Munford, 2012). People with learning disabilities had been silenced, often living in segregated communities including large institutions that often discouraged regular contact with their families and kept them hidden from society. Over the last few decades there has been a major change in the lives of disabled people who are now visible in our communities; legislative change and policy reforms have crystallized the issues associated with the move from segregated settings to community care. Disabled people and their support networks have challenged the policies and services that have functioned to exclude rather than involve them in the daily lived experiences of their communities in spheres such as education, work, parenting, and recreation (Munford and Bennie, 2009: p. 213).

Strong and sustained advocacy by disabled people and their support networks has facilitated access to resources and services in the community. However, service provision for people with learning disabilities is of variable quality and in many countries appropriate, and high quality services (that advance human rights and participation in community life) are simply not available. While there may be a commitment to supporting people with learning disabilities to have access to appropriate services, issues over entitlement can limit access and exclude some individuals (Sullivan and Munford, 2005). The assessment procedures used to determine access to services may be focused on tightly circumscribed needs rather than on individual rights given that a service system must deliver what is economically feasible. In such conditions restrictions on resources often mean that entitlements based on rights are denied and needs are not met (Munford and Bennie, 2009). However, despite these restrictions alternative options for living and working have opened up new possibilities for people with learning disabilities. New technologies such as computer-assisted communication have enhanced participation in education and provided for better access to information that has also been a catalyst for people to be better informed of their rights and entitlements.

People with disabilities continue to demand "their rights as full and equal citizens" (Stainton et al., 2010: p. 1) and these rights are enshrined in international legislation such as the UN Convention on the Rights of Persons with Disabilities. Such recognition in law secures rights and "demands for control over how, when, and by whom" (p. 1) support is provided. For people with learning disabilities progress can be assessed across the key domains of living environments, personal lifestyles, education, and employment. While advances have been made in each of these areas many challenges remain. For example, with regard to living environments, while there is now an expectation that disabled people will live in services that are located within neighborhoods that enable them access to the ordinary daily activities that are available to other citizens, some of these services are still segregated and do not facilitate full participation in community life. In many countries in the education field, there is now an expectation that people with learning disabilities do not need to be educated in special schools but can participate in mainstream educational opportunities, that is, in schools in their communities. This aspirational goal of an inclusive education guides practice in education in many countries around the globe. However, the reality can be different because given the competition for scarce educational resources the experience of people with learning disabilities in mainstream settings can be one of alienation and marginalization. This means that they are excluded from the ordinary daily activities that other pupils take for granted. In these settings social workers often work alongside families and people with learning disabilities to become strong advocates for full participation. These experiences in the school system also play out in the employment arena, where again, the aspirational goal of full engagement in employment in all settings cannot be realized because of either scarce resources, or, because of a lack of commitment to provide the support that is required for disabled people to enter and sustain involvement in the labor market. Key challenges also arise in the transition from school to work where people with learning disabilities may have been well supported in school environments but this support has not been transferred into work settings. With regard to choices about personal lifestyles, these too, may be restricted as in many countries there is still not entitlement to the full spectrum of experience that constitutes human life, such as sexuality and caring for others including becoming parents (Booth and Booth, 2003; Llewellyn and McConnell, 2002; Munford et al., 2008a,b; Ward, 2011). As Hallahan (2010: p. 127) suggests, disability is an evolving and complex situation where policy and social practices will determine the opportunities open to people. Social workers can have a key role in opening opportunities taking on key support roles that position them as 'allies in emancipation' (O'Brien and Sullivan, 2005).

The Social Work Role – Key Elements and Challenges

Social workers are in a unique position to support people with learning disabilities to fully participate in community life. Working across systems and at the individual, community, and policy level social workers are well-placed to build a deep understanding of the issues confronting disabled people in their daily lives. Moreover, social workers can also resist roles that require them to act as gatekeepers to resource and services, but position them as agents of change in disabled people's interests (Hallahan, 2010). Central to this role is developing an analysis of how services function and whether they promote meaningful opportunities and possibilities for valued identities. A critical analysis enables social workers to work on the cutting-edge and to challenge restrictive service

practices; for example, putting the spotlight on community-based residential services that have become entrenched in models that restrict choices and isolate people with learning disabilities in community enclaves. Similarly, as Munford and Bennie (2009: p. 215), argue, there may be a need to challenge the policies that result in the isolation of work and day activities in segregated environments such as sheltered workshops. Here, there is often little connection between services and the communities in which they are located and as a consequence "people's lives may become compartmentalised, bureaucratised and isolated" (p. 215). Moreover, under such restrictive practices, the daily lived experience of people with learning disabilities become constructed in terms of service systems rather than experience being framed by people "merely wanting to be ordinary citizens participating in ordinary activities and having ordinary lives" (p. 215).

Given the challenges outlined above, a key task for social workers is that of advocacy. As members of teams social workers often work alongside the people with learning disabilities to support them to access community resources and to be supported in mainstream environments, such as schools and work places. Social workers may also be required to work at the macro-level to highlight the effects of policy and service decisions and to identify gaps in services. This work is informed by their understanding of the perspectives on disability outlined above. From a psychosocial perspective this requires working with individuals to understand the support required to enable people with learning disabilities to fully participate including becoming partners and parents and living and working in community settings. Working within a political model, social workers would extend this understanding of lived experience at the individual level and critically examine why people may be marginalized in their communities and excluded from participation. For example, in the employment domain social workers may help the person access resources that would help them find a job and then facilitate practical support in 'real' work contexts. The social worker may also work at the political level challenging the systems that prevent direct entry into the labor market and may invoke human rights legislation such as the UN Convention on the Rights of Persons with Disabilities to challenge discrimination (Munford and Bennie, 2009: p. 216).

Alongside direct support work with people with learning disabilities social workers will also work with families including parents with learning disabilities. Social workers support families to access services and resources and also provide a valued role in assisting families to make sense of the discrimination they often experience. This may involve the social worker standing alongside the family as they challenge service systems and community networks that prevent rather than encourage participation. For example, the social worker may use the inclusive education policies in a particular education jurisdiction that underline the rights of people with learning disabilities to be included in mainstream schools, to support the family to actually make this happen. The social worker may have an on-going role with the school in that they will not only support the family to achieve the right to enroll their child in the local school, but they may also be available to work through any emerging issues that may restrict involvement of the person in all school activities – academic, social, recreational, and extracurricular activities.

The example cited above, demonstrates the importance of social workers building a deep understanding of policy, legislation, and services, while they work alongside families and people with learning disabilities to realize their dreams of full participation in community life. Historically the medical model of disability would have promoted a view of the child with a learning disability as suffering deficits and the best response to this would be to institutionalize this child (Munford and Bennie, 2009: p. 216). We have seen from the discussion above, that emerging theoretical and practice perspectives have changed the way in which disability is constructed. These perspectives align with other approaches that promote inclusive practices and involvement of disabled people in community life. A key perspective informing social work with families who support people with learning disabilities is strengths-based approaches. Such approaches encourage social workers to move from defining the issues families face as 'deficits' to defining these as challenges that can be addressed with the right supports, networks, and services. This approach aligns with those in the disability field that call for a reframing of the issues facing disabled people as issues of citizenship, participation, opportunity, and support (Munford, 2012).

Strengths perspectives encourage social workers to think differently about family life and to understand that a focus on 'deficits' and on 'problems' can mask the 'multiple positions' families may occupy (Munford and Sanders, 2008). This approach does not downplay the challenges and daily struggles that families and people with learning disabilities confront, but directs attention to understanding the diversity of family life and to determining how families can be more effectively supported. Underpinning strengths approaches is the belief that people with learning disabilities have a right to an ordinary life and that it is unhelpful to perceive those who face enduring challenges as suffering human beings living tragic and sad lives that need to be 'fixed' before they can fully participate in their communities. The principles inherent in strengths approaches encourage social workers to work differently in their 'professional' roles, holding the notion of 'expert' lightly so that the knowledge and strengths of families and disabled people direct the support relationship (Munford, 2012). This approach is in contrast to the role social workers had when services were informed by the medical model. In this period social workers were more likely to take on roles that required them to independently make decisions about service delivery and access to resources without necessarily consulting with those would be affected by these decisions. In contrast to this approach social workers, under the rubric of approaches such as strength-based practice, are encouraged to work collaboratively with families and disabled people. This includes finding ways to affirm the support relationship and its important role in enabling people with learning disabilities to live in their local communities (Munford, 2012).

Central to these new approaches of service delivery and social work practice with people with learning disabilities and their families is to ensure that the disabled person remains at the center of support relationships and directs decision making. In other words, decisions are not made on their behalf and without their knowledge. They direct what decisions need

to be made and also the decision-making processes; here the social worker has a key role in ensuring that all of the information required to make these decisions is available in an accessible form and that options are realistic and meaningful. The goal in such practice is to support people to find positive and valued identities and this requires a commitment on behalf of the social worker to understand the barriers to social inclusion and then to work with the person to overcome these (Munford, 2012).

A significant component of social work support is enabling family members to fully participate in decision-making processes and to accept the disabled person's right to direct these processes. This can be difficult for families who have had to take on strong advocacy roles for their family members and who have trusted service providers who have failed to deliver on their promises of providing responsive and empowering services. This is often the experience for mothers who have often taken on the major care roles for their child/adult and have often been challenged by professionals that they have become too involved and overprotective. Taking a back seat can be difficult for these families who have faced many challenges and this requires social workers to be skilled in relationship management so that all voices in the support relationship are heard and respected (Munford and Sanders, 2008). For disabled people who are parents themselves the challenges can be intensified as their parenting is often scrutinized and judged by others (Booth and Booth, 2003; Llewellyn and McConnell, 2002; Munford et al., 2008a). To this effect disabled people remind us that effective support relationships are based on interdependence and reciprocity (Munford et al., 2008b; Ward, 2011). This is a direct challenge to the ways in which disabled people have been defined as passive recipients of care where they have had things done to them rather than with them. As Ward (2011: p. 178) argues, "care [support] is a democratising process" and that "the binaries that position people as dependent or independent, carer and cared for" need to be disrupted. What is required is an understanding of care [support] relationships as mutual and reciprocal and this orientation to support opens up new spaces for participation and for promoting disabled people as active citizens (p. 178).

The call from disabled people to be part of support relationships that are based on mutuality and reciprocity has led to some considered and deep thinking about what factors are central in building successful support relationships. The research suggests that to be effective support relationships should be responsive and flexible and that the ultimate goal is to build an interdependent network of support within communities. At the core of this network of support are collaborative partnerships between people with learning disabilities, their family members, and social workers and other support people. Collaborative practice grows from the recognition that when people have power and control over their circumstances they are more likely to find positive solutions to their issues and challenges (Munford, 2012). A collaborative orientation to support emphasizes joint agenda setting and identification of shared goals that take account of the needs and rights of all those involved. Key to this approach is recognizing and respecting what people bring to the relationship; alliances are formed in order to access services and these alliances are respectful of differing meaning systems. Moreover in keeping with the principles of strengths approaches social workers adopting these practices work to understand family life and the daily lived experiences of disabled people in order to understand what has worked in the past and to then use this knowledge to find solutions to current issues. Taking a collaborative approach means that there is considered thinking on issues and attentiveness to opportunity and possibility; within this approach alternative strategies are generated so that there is different thinking about services so that the most can be made of available resources.

Key principles informing collaborative social work practice are:

- The person is at the center of the support partnership and determines the relationship with social workers, other professionals, and service systems. Time is taken by social workers to understand the daily lived experiences of the disabled person including their support networks. An example of this in practice involves social workers harnessing the person's networks in order to facilitate full participation in community life; for example, connecting the person with the local neighborhood activities (O'Brien and Sullivan, 2005).
- Social workers also take the time to understand the meaning systems of people with learning disabilities including building an understanding of the way in which context shapes their daily lived experiences. This involves understanding how political, social, economic, religious, and cultural factors influence and shape daily lives and includes an analysis of how these have restricted or opened up possibilities for the disabled person (Munford and Sanders, 2008).
- In building successful collaborative support partnerships social workers need to understand their key roles and the values, knowledge, and skills they bring to these partnerships. Central to this is their value orientation which requires them to interrogate their own beliefs about participation, citizenship, and social justice. It also requires them to have a theoretical understanding of the perspectives that have informed the provision of services to people with learning disabilities combined with a knowledge of the policies that determine how services are organized and resources are allocated. Their knowledge and skills will incorporate understanding about the psychosocial experiences of living with impairment and includes an ability to work at the individual level. They may also be required to work at the wider community level in advocacy roles and also participate in social change initiatives alongside people with learning disabilities (Sullivan and Munford, 2005).
- Social workers adopting collaborative approaches need to critically examine their own spheres of influence and whether they are able to be allies alongside disabled people in their struggles to achieve full participation. Social workers may find themselves in conflicted positions where they will be required to challenge other professionals in order to facilitate access to resources for disabled people. Understanding policy and service allocation processes and the impact of these on practice are important elements in being able to adopt collaborative approaches; social workers need to have clarity around their own positions in order to be effective. They need to strive to be genuine and authentic

with people with learning disabilities including being honest about the roles they are able to take on.

- Social workers can have a key role in advocating for integrated services; such services embrace collaborative practice and encourage cooperation and the sharing of knowledge and resources. Within integrated approaches services are prepared to engage differently with their communities and to work in partnership with other services in order to ensure that service provision is responsive and flexible. These services have a commitment to finding innovative solutions to intractable issues by working together to confront the barriers that prevent people with learning disabilities gaining access to resources and opportunities for participation. Such services work in close partnership with disabled people in order to receive ongoing feedback about their performance and the role they can have in more effectively facilitating access to resources (Munford and Bennie, 2009).
- Building on the above, a key element of collaborative practice is sharing knowledge and experience. Here there is recognition and acceptance of the diverse experience and learning that informs support relationships. At the heart of effective support relationships are respectful learning relationships. This idea gives effect to the notions of reciprocity, mutuality, and interdependence which create connections and solidarity between disabled people and others, such as social workers. This idea of shared learning invokes the notion of interdependence; it may require a change in thinking on behalf of social workers who need to position themselves as learners as they join with disabled people to learn about what is possible in terms of service delivery and participation in community life (Munford, 2012; Ward, 2011).
- Collaborative practice requires a strong value base that is built on a commitment to being future-focused and to thinking about what is possible for people with learning disabilities. It requires social workers to keep to the forefront of their practice ideas about citizenship and what constitutes inclusion and full participation in community life. This requires ongoing critical reflection on practice; this reflection should be an integral component of practice as it enables social workers to think deeply about their practice and to always remain open to new possibilities. Importantly, reflection on practice reminds social workers to reflect on their role in the support relationship and to ask themselves whether in their daily encounters with people with learning disabilities they are 'allies in emancipation' (O'Brien and Sullivan, 2005).

Conclusion

This article has explored social work practice and the experiences of people with learning disabilities. It began by outlining the definitions currently in use in this field of practice and then moved to a consideration of philosophical and theoretical perspectives. It explored how these perspectives had changed over time and that alongside these changes to thinking about the experience of disability were changes in policy and service delivery. Policy agendas around the globe have determined how services for people with a learning disability are to be provided. Both the policies and the resultant services have undergone major changes including the movement of disabled people from large isolated institutions into community settings where they have been encouraged to fully participate in community life. Social work practice has been influenced by changes to the theoretical perspectives, policy frameworks, and service delivery models; at various times social workers have been required to be gatekeepers of resources while at other times they have taken on support roles and have become involved in advocacy work and social change initiatives. The section on the social work role discussed the key elements of this role and also highlighted the challenges faced by social workers as they sought to form collaborative partnerships with people with learning disabilities and their families. This section explored the perspective of strengths-based practice which has had a major influence on social work practice with disabled people and their families. The article concluded with a discussion on the key principles that inform collaborative practice; this orientation to practice enables social workers to work alongside people with learning disabilities to achieve citizenship and full participation in community life.

See also: Disability and Disabling Practices.

Bibliography

Booth, T., Booth, W., 2003. In the frame: photovoice and mothers with learning difficulties. Disability and Society 18 (4), 431–442.

Department of Health, 2001. Valuing People: A New Strategy for Learning Disability in the 21st Century. Department of Health, United Kingdom.

Hallahan, L., 2010. Legitimising social work disability policy and practice: pain or praxis? Australian Social Work 63 (1), 117–132.

Llewellyn, G., McConnell, D., 2002. Mothers with learning difficulties and their support networks. Journal of Intellectual Disability Research 46 (1), 17–34.

Munford, R., 2012. Supporting Families: Building Strengths and Resilience. Keynote presentation at the ASID conference, 9 November, Wellington, New Zealand. http://www.asid-2012.org.nz/.

Munford, R., Bennie, G., 2009. Social work and disability. In: Connolly, M., Harms, L. (Eds.), Social Work: Contexts and Practice, second ed. Oxford University Press, Australia, pp. 209–221.

Munford, R., Sanders, J., 2008. Assessment of families. In: Rowe, W., Rapp-Paglicci, L.A. (Eds.), Comprehensive Handbook of Social Work and Social Welfare: Social Work Practice. W. John Wiley & Sons Inc., USA, pp. 399–422.

Munford, R., Sanders, J., Mirfin-Veitch, B., Condor, J., 2008a. Looking inside the bag of tools: creating research encounters with parents with an intellectual disability. Disability and Society 23 (4), 337–447.

Munford, R., Sanders, J., Mirfin-Veitch, B., Condor, J., 2008b. Ethics and research: searching for ethical practice in research. Ethics and Social Welfare 2 (1), 50–66.

O'Brien, P., Sullivan, M., 2005. Allies in Emancipation: Shifting from Providing Support to Being of Support. Thomson, Australia.

Oliver, M., 1990. The Politics of Disablement. Macmillan, London.

Reeve, D., 2002. Negotiating psycho-emotional dimensions of disability and their influence on identity constructions. Disability and Society 17 (5), 493–508.

Stainton, T., Chenoweth, L., Bigby, C., 2010. Social work and disability: an uneasy relationship. Australian Social Work 63 (1), 1–3.

Stevens, A.R.A., 2008. Social models of disability and social work in the twenty-first century. Ethics and Social Welfare 2 (2), 197–202.

Sullivan, M., Munford, R., 2005. Disability and support: the interface between disability theory and support: an individual challenge. In: O'Brien, P., Sullivan, M. (Eds.), Allies in Emancipation: Shifting from Providing Support to Being of Support. Thomson, Australia, pp. 19–33.

Sullivan, M., Munford, R., 1998. The articulation of theory and practice: critique and resistance in Aotearoa New Zealand. Disability and Society 13 (2), 183–198.

Ward, N., 2011. Care ethics and carers with learning disabilities: a challenge to dependence and parternalism. Ethics and Social Welfare 5 (2), 168–180.

Williams, S.J., 1999. Is anybody there? Critical realism, chronic illness and the disability debate. Sociology of Health and Illness 21 (6), 797–819.

Learning: Mathematical Learning Theory, History of

Sandy Lovie, University of Liverpool, Liverpool, UK
Pat Lovie, Keele University, Keele, Staffordshire, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by W.K. Estes, volume 14, pp. 9391–9393, © 2001, Elsevier Ltd.

Abstract

This article covers the role that mathematical modeling and associated graphical representation have had in the study of animal and human learning. Starting with the nineteenth century, the account moves rapidly to post–World War II developments in the field, ending with the major contribution of W.K. Estes and his influential theory of stimulus sampling.

Historical and Conceptual Arguments, and Contexts

The topic of this article can be construed in either a narrow or a broad historical fashion: the latter approach is taken here since a key worker in the field, W.K. Estes, has already provided a more focused and insider's view in the first edition of this *Encyclopedia* (Estes, 2001). It is also the case that the history of recent experimental psychology scarcely exists as an academic subject, particularly the roles of mathematical modeling and certain branches of statistics in its development. Consequently, this article should be viewed as little more than a modest start in accounting for what is by now a considerable and expanding area of psychology.

This broader approach also includes some observations (where appropriate) on the key relationships that have always existed in modern psychology between mathematical models and their varied forms of representation and evaluation, including graphical representations or plots that consist of original data *and* fitted representations/models of the data. Thus, the intention of this article is to suggest that mathematical psychology is a complex argumentation that interweaves mathematical developments with tailored graphical representations whose purpose is to persuade the reader to accept a specific interpretation of the results of the joint and interacting mathematical and graphical works.

Moreover, what seems to have happened in practice is that the apparent superiority of mathematical argumentation over graphical methods is often reversed, with the graphical frequently bearing the important results and conclusions; this is particularly the case with the work of C.L. Hull, as will be seen later when the rhetorical role of one approach against another is discussed. It could also be argued that in the case of two important players, Bush and Mosteller, at least one of their reasons for providing plots as well as mathematical analyses was pedagogical in nature since their 1955 book is clearly designed and laid out to be as much a teaching resource as an exposition of their approach.

The contention here is that such representations, which mix data and models, are the key to grasping the history of experimental psychology's attempt to achieve what has generally been conceived of as a scientific understanding of such fundamental operations as learning and memory. In other words, the mathematization of psychology was usually judged by insiders to be one of the cornerstones of a scientific psychology. Of course, most psychologists, being in the ever expanding and conceptually ambiguous areas that now make up their subject, might well balk at this interpretation. Nonetheless, the small pool of mathematical insiders would probably concede that while the wider psychological community has to deal as best it can with highly intractable material, such as complex problems of mental illness, eventually even this will succumb to the former's mathematical wiles.

Mathematics and Psychology in Practice

We begin with something of a paradox: for most psychologists, whether the earliest self-conscious practitioners of this 'art and sullen craft' in the nineteenth century, or today's computer literate and neurologically savvy experts, their only encounter with mathematics would be through the statistical analyses of their data. Now, for those mathematicians who developed this field, psychology was only one of many areas to which they applied their efforts. An early pioneer of statistics, and one of the most important and influential mathematicians who worked within it, was R.A. Fisher, whose principal area of application was agricultural research. Yet psychologists throughout the world daily make use of the method that he termed the *analysis of variance* (ANOVA). Moreover, ANOVA (and related techniques such as MANOVA, or *multivariate* ANOVA) have been construed by statisticians since the 1940s as examples of model building and testing analogous to the processes covered in this account.

On the other hand, model building in psychology as an explicit process attracts only a tiny handful of mathematically sophisticated psychologists and sympathetic statisticians. The situation is, therefore, very different from that of physics, where theoreticians and experimenters play a more complementary game. Psychology has never been as conceptually coherent and as focused as physics, although at different times and different places in its history it has often wished to be so viewed. Thus, in recounting the development of the mathematization of a tiny area of psychology, the reader should be aware that what they are seeing is only one out of a myriad ways of thinking about and doing psychology. And, of course, since insiders, like any other victors, usually write the history, so accounts of the past are determined by those who practice the arts of today. A relevant example here is the nineteenth-century German philosopher G. Fechner, whose simple law ($S = k \log P$) underlies modern

psychophysics (here S is Sensation, P is a Stimulus value, and k a fitting constant). Fechner's purpose in developing this simple relationship was to find support for his philosophically grounded conjecture that measurable actions or judgments by the human soul could be directly expressed in external, physically defined, and measurable aspects of the world (apart from the simple log transform of this latter aspect). Here, a philosophical/metaphysical concept (*monism*, as distinct from *dualism*, which argues that the soul is separate from the world) is being expressed and tested in an experimental and mathematical form. Yet the history of modern psychophysics rarely, if ever, acknowledges the original philosophical basis of the work and routinely applies such methods to many areas both inside and outside psychology. Thus while Fechner had some form of intellectual justification for his mathematical work (here the need to demonstrate a simple isomorphism between the organism and the world), the attraction of psychophysics more generally was of its utilitarianism and consequently of its lack of any particular intellectual commitment or history.

The point is that mathematics (and its supporting offshoots and surrogates such as data displays and graphs) is, for the experimenter and theorist alike, at its most useful and persuasive when it is linked to some form of relevant and accepted reality. This does not mean that coupling a mathematical operation or element to aspects of the observable world endows it with an unquestioned authority or completeness, but it does provide a certain increased power to represent and manipulate certain relevant aspects of the world, as will be seen. It also encourages the user of such models and representations to predict, and perhaps (for followers of the neo-behaviorist B.F. Skinner) control, an organism's behavior.

The latter part of the nineteenth century saw H. Ebbinghaus's work on memory (which is the first area covered by Estes, 2001). Ebbinghaus was interested, among other things, in memory loss over time and used himself as an experimental subject (1885/1913). The decline over time in his ability to hold three-letter nonsense syllables in his memory was expressed as an equation fitted to a version of the data, in the same way that Fechner had done (and the physiologist E. Weber, who was one of Fechner's inspirations, had done even earlier). Fitting a curve to his data would nowadays be referred to as *empirical curve fitting*.

In Ebbinghaus's hand-fitted equation $b = 100k/((\log t)^c + k)$, b is a retained memory score, t is the time between learning and testing, while k and c are constants (but, see below). Although he does not report what nowadays would be referred to as 'goodness of fit measures,' he does comment on whether the deviations of the observed from the calculated values fall within the limits of what was termed at the time 'probable error.'

Strangely, as Estes (2001) points out, although Ebbinghaus had put considerable effort into arriving at his equation and the empirical estimates of its parameters, he was reluctant to attribute much significance to the result 'at the present time' (1885/1913, Ch. 7, end of section 29). In other words, Ebbinghaus fitted a mathematically defined curve to his data, though without, somewhat paradoxically, seeing the potential of the exercise. But what if Ebbinghaus had tried to plot out the data and his fitted curve? This would have allowed him to appreciate aspects of the results that were not easily discernable from the raw data alone. Eyeballing the plot, therefore, immediately reveals several points of interest: in particular, the amount remembered b, as a percentage of correct recall, tested at increasing time intervals after the initial learning does not show a straight line reduction with time, rather it is a complex curve that rapidly descends from the initial testing time to 2 days after this point, with the curve staying almost flat up to the final time of 31 days after the first test. The score at 2 days is about 30%, while at 31 days it is about 25%. Notice here that Ebbinghaus must have systematically tried out values for the two constants t and k so as to maximize the fit of the curve to the data.

At this stage, of course, the formula has only empirical validity, in that it fits the data in hand, but one could begin *post hoc* to think of other experiments that take into account or build upon the results of this initial analysis. From such a standpoint one could also start to think about the meaning of the phenomena, or at least a more systematic exploration of the situation and the collection of experimental regularities. However, what was ideally needed was a psychologically/substantively grounded system that would generate numerical outcomes and their plots; in other words, the requirement would be to employ a rational process of matching plot to data via a theory, rather than the ad hoc one of empirical curve fitting after the data had been safely gathered in.

This curve-fitting approach then inspired an increasing number of workers, particularly in the United States, to experiment with trying out a variety of curves to a range of memory recall results. However, even where a more recent player in this story claimed to be offering a theory-based curve-fitting exercise rather than an empirical one, it is still the case that basic processes in his scientific system involved fitting empirical or even subjectively produced curves as a fundamental part of his conceptual and methodological work. This person is the learning theorist C.L. Hull, whose research from the 1930s until his death in 1952 figures prominently in Estes's earlier contribution to this *Encyclopedia*. According to Estes, "Hull (1943) set forth a series of axioms for a general theory of conditioning and elementary learning" (Estes, 2001, p. 9392). What Estes concentrates on in his brief account of Hull's work are the putative outcomes or purposes of the work, while omitting the grand, all-embracing sweep of Hull's theoretical efforts, whose final outcome was intended to be (to echo the title of his final work) *A Behavior System* (Hull, 1952).

Hull's system consisted of a complex of interacting processes and stages in determining the acquisition and storage of a range of material, all teased out using a variety of experimental tasks and routines. What Hull used extensively as a method of representation, argumentation, and insight was a large array of fitted curves (see particularly Hull, 1951). However, as was implied earlier, his a priori selection of and devotion to exponential functions in particular and the rather interesting way in which they could be augmented and conceptually enhanced show that at least with this foundational aspect of his work Hull was more empirical than rational where one can detect no less than six graph fitting stages between raw data and finished plot (see Lovie, 1992, p. 102). Lovie also analyzes, in some detail, a similar scaling and curve-fitting exercise undertaken by the social psychologist C.I. Hovland, who was, in the late 1930s, a much-valued graduate student of Hull (Lovie, 1992, pp. 68–70).

The outcome of these exercises by Hull and Hovland was the successful fitting of one form or other of an exponential function. Now, in many ways such a function has an obvious and direct appeal to those interested in learning, for instance, as the dependent variable such as time or trials progresses, so the organism learns the material making up the task fairly quickly (say, learning a list of nonsense syllables or running a maze, and eventually mastering the task, so the error rate flattens out at zero or the running time reduces to a fixed minimum). If some measure such as errors is plotted cumulatively, the curve climbs at first but then finally flattens out and asymptotes at some fixed value. An exponential function looks a priori plausible here, since most tasks can be attained eventually, and most people/organisms experience a rising success score with most problems as they apply themselves to the task (providing that the problem is soluble).

Hull was quite happy to proceed in this ad hoc manner, perhaps because he could also control the curve-fitting routine, thus achieving agreement between data and the mathematical models generating the curves. It is also the case that goodness-of-fit tests were increasingly available at this time, but Hull still chose to eye-fit his curves, perhaps because he had only a tiny number of data points available, which would have ruled out current goodness-of-fit techniques that were notoriously data hungry, but it also could be argued that the use of such tests would have removed Hull's control over the fitting process and the interpretation of its outcomes. Finally, it is also true that Hull was happy to extrapolate the extent of his data fitting many times further from the values of his limited data set, while he also claimed a symmetrical extension of such curves over experimental values not even tested by his student Hovland (see Lovie, 1992, p. 100). A similar ambiguity over the use of goodness-of-fit tests is also evident in later and more mathematically sophisticated contributions by Bush and Mosteller (1955).

Finally, Hull was fiercely attacked by S. Koch in 1954 in what was an otherwise positively toned, multiauthored volume entitled *Modern Learning Theory*. Hull, whose influence was still strong until at least the end of the 1950s, was equally fiercely defended against Koch's onslaught by F.A. Logan in a 1956 review of the book. In addition to referring positively to Hull in 1959, Estes, who clearly viewed him as his intellectual predecessor, concluded his 2001 essay by pointing to Hull's ideas as the foundation for modern cognitive psychology. He also praised Hull in the same article as contributing positively to recent interpretations of Pavlovian conditioning.

Our penultimate study is of Bush and Mosteller's (1955) book *Stochastic Models for Learning*. The book itself is divided into two unequal sections, with the larger initial part devoted to the necessary mathematical ideas that would then allow the reader to make more sense of the second, where some five data sets were exposed to lengthy mathematical analyses. Thus modern mathematical and statistical ideas, such as matrix methods, parameter estimation and the properties of modern estimators, mathematical operators, sequences, Markov chains, discrete mathematics, and moment-generating functions, are all explored in elegant and (for anyone prepared to work hard enough) accessible treatments of the material.

Above all else, however, the book rigorously pursued the probabilistic or stochastic approach to behavior, which explicitly accepted the fundamental uncertainty of the world but also allowed one to understand such a place provided certain assumptions were made about it, for example, that successive responses were independent (or, more generally, that any state was dependent only upon the previous one, material learnt or not learnt, for instance), or that it did not matter in what order a specified state was achieved (so-called path independence assumption). This approach differentiated it from Hull's theorizing, for example, where the dependent variable was usually the response alone, not the *probability* of the response.

Unfortunately, most of empirical studies subjected to such heavy weight analyses seem far too trivial and out of the mainstream of the experimental psychology of the period to be of much interest to the average experimental psychologist, with only Solomon and Wynne's classic study (1953) on avoidance conditioning in dogs representative of mainstream experimental psychology (and perhaps the disappointing analysis of rat running times from a 1954 study by Weinstock). Even here, however, Bush and Mosteller deliberately chose to analyze simple, binary-coded avoidance responses (*avoid, not avoid*), a measure that Solomon and Wynne themselves had rejected as almost meaningless, since in practice it had yielded little, if any, useful response variance (see Lovie and Lovie, 2004, for a more detailed analysis).

Consequently, what we have here is still an (admittedly) nuanced and mathematically sophisticated form of curve fitting, since there is little attempt to derive the various formulae from theoretical psychological considerations except of the most obvious or banal kind, for example, dogs in the two-chamber shock avoidance task will quickly learn to jump from the shock chamber into the safe one when the unconditioned stimulus sounds. Further, the major criterion of success used by them was (rather ambiguously) both an unstated measure of the degree of fit, or a simple eyeballing of the plots of theory and data, with an invitation to see how close the agreement is visually (Bush and Mosteller, 1955, pp. 250–251). Some of the other applications yielded tractable goodness-of-fit tests, but others did not so simple eyeballing had to do. It is also worth noting that the least satisfactory analysis was that for the Weinstock data since their model was unable to capture the tails of the run time data.

But were there any psychological theories at the time able to suggest anything more than mean differences between experimental levels, or at best the overall form of the learning curves, for example, rising fairly steeply with practice, then leveling off as the material was learnt to some criterion? In other words, was there a group of experimental psychologists comparable to the physicists who could work productively with their theoreticians? The answer then and, probably still, now is 'no.' Perhaps the productive interaction as between theoretical and experimental physicists is an inappropriate paradigm for psychology. Certainly, the mathematical skills needed to understand and benefit from Bush and Mosteller's text are not trivial; perhaps they mistook either the mathematical skills necessary or the amount of training and commitment needed before such mathematical psychologists were thick on the ground. Either way, the book (which appeared as part of the prestigious series *Wiley Publications in Statistics*, though not, notice, in psychology) did not achieve any major recognition

within psychology, even though it is both stimulating and clearly written as was the case with all Mosteller's works. It was also obviously structured as a teaching aid, but it is less clear for what audience; perhaps Bush and Mosteller were hoping to create such a teaching area. The problem is that it was mainly written by an outsider, Mosteller, who had considerable mathematical skills but less commitment to psychology.

There is also the problem that was raised in the rather neglected final chapter of Bush and Mosteller. Here they note that since any reasonably sized data set can yield an enormous number of descriptive statistics, the problem becomes how to make sense of these summary numbers, which means that the psychologist must have some theory-based criteria as to what measured outcome is relevant to that theory and what is not. One further issue that arises here is that if there were theories in psychology that were expressible as numerical summaries (or parameter estimates) that had been arrived at independently of the data set of interest, then one would be able unambiguously to test any viable mathematical hypotheses using that data set. However, if the estimated numbers were both based on the present data set and used to test out hypotheses about this data, then one cannot assert the independence of these conclusions; thus a different set might yield a different conclusion, and there would be no way of saying which was the better result. Even Bush and Mosteller's own worked examples could be said to violate this rule, so conclusions drawn from their model/data were ambiguous, at least using their test of independence of conclusions (see Lovie and Lovie, 2004). It is therefore not too surprising that Mosteller at least quickly busied himself with statistical projects outside psychology.

Our final researcher is W.K. Estes. He, unlike Mosteller, was trained as a psychologist, at one time serving as a research associate of Skinner. Although he had published work on animal learning, the most influential phase of Estes's work was begun in 1950 when he produced the first of many papers outlining and developing what he later termed *stimulus sampling theory* (SST). Like the parallel work by Bush and Mosteller, this offered a probabilistic account of conditioning and learning where this was conceived within the learning theories of its time, thus it is possible to see echoes of I. Pavlov, E.R. Guthrie, Hull, and, of course, Skinner in his work, who together offered a complex dish made up of simple contingency (Guthrie and Pavlov), blended with some leavening of reinforcement and probabilistic responding (Hull and Skinner). Although some commentators (e.g., Cox, 1997) have limited his sources of influence to Guthrie, while, as noted earlier, Estes himself also cited the foundational role of Hull, who had emphasized the importance of reinforcement in learning.

SST can best be described as a probabilistic and dynamic recasting of basic stimulus–response learning theory. Thus, on a particular trial a stimulus is sampled at random and a single response (also randomly sampled) is associated with it. On subsequent trials these connections made between specific stimuli and responses may or may not persist, thus eventually generating a probability distribution over all such pairs over trials (the assumption here being that both stimuli and responses are finite in number, and hence the number of possible pairings is finite as well). Finally, a random number of such pairings are made on any one trial, a number that can also change at random from trial to trial. Put this way, SST, because of its highly abstract nature, was not an approach that could be directly or easily tested on particular learning situations.

However, so broad an approach can potentially be generalized to many situations, which is what happened over time. Estes also explored various possible features of such differing realizations, for example, the effects on SST of learning in a conditioning experiment or the differential impact of using simple or compound stimuli (Estes and Burke, 1953). The effects of (spontaneous) stimulus variation on the mathematical description of a range of learning phenomena such as spontaneous recovery, regression, and the effects of trial spacing can all be captured simply within a generalized SST framework (Estes, 1955a, 1955b). Suddenly, the one approach provided by SST meant that significant areas of classical learning theory could in principle be rescued from the qualitative speculations of Skinner and Guthrie, or the quasi-quantitative ones of Hull. Of course, to accept SST one had also to agree that a modeling approach yields more scientific information than the more traditional, informal, and qualitative ones, but Estes's wish for psychology to emulate physics makes it clear on whose side his sympathies lay (Estes, 1959). Note too that the stimulus for the approach adopted by both Estes, and Bush and Mosteller, came directly from modern, post-Einsteinian physics, with its conceptual foundations firmly planted in an uncertain and probabilistic worldview. (A question that arises at this point is whether the world is intrinsically uncertain, or if it is our knowledge of the world that is incomplete, and thus only apparently uncertain.)

The problem in the longer term, however, was that while SST made a moderately successful attempt at outflanking just about every alternative approach in classical learning theory, it was not so general as to be able to do the same for the cognitive juggernaut that was, even in the late 1950s, beginning to build up a ferocious head of steam. So SST went the way of the dodo, rather like Bush and Mosteller's efforts, but not for the same reasons. Thus, while the latter espoused no particular substantive approach to psychology, Estes, in the end, backed the wrong elderly horse (see, however, Bower, 1966, p. 379, and also Bower and Hilgard, 1981, p. 251, while acknowledging the decreasing impact of SST, nevertheless put a slightly more positive gloss on its survival).

In the End

Estes's extended coverage of SST in 1959 represents essentially the last major outing of the system as a main influence on either theoretical or experimental psychology. What lay before psychology was the growth of cognitive approaches and the computer analogy. This offered so many actual and potential riches to psychologists, and fitted in so nicely with the (non-behaviorist) direction of its future path, that it was taken up with enthusiasm by many young researchers, particularly in the United States and the United Kingdom. There was also a conscious effort by these young workers, drawing on the notion of a scientific revolution put forward by T.S. Kuhn in 1962, to argue that cognitive psychology represented just such a revolution, and hence any behaviorist ideas could be rejected out of hand as lying on the wrong side of the

conceptual gulf (see Lachman et al., 1979, for a mature assertion of this Kuhnian position).

Estes's continued commitment to behaviorism generally (as embodied in his system of SST) and to Hull in particular, has largely consigned his work to the past. Mathematical psychology, however, has continued as a minority taste, though its survival may well be dependent on how far it keeps itself relevant to developments in psychology at large. Indeed, there is no significant evidence of the symmetrical two-way traffic between theorists and experimenters in psychology that can be found in physics. Bush and Mosteller's conception died almost as soon as it was born because they ignored this precept, while Estes, who lasted a little longer, still lost out as the conceptual wind changed. Further, Skinner, who invented the most successful and long-lived form of behaviorism yet seen, eschewed any form of mathematical modeling other than the plotting of cumulative records that 'speak for themselves,' as he might have said.

Finally, one stumbling block restricting the translation of mathematics into psychology was the technical limits to the subject, that is, the issue of 'tractability,' for example, the problem of finding closed, finite solutions to the various mathematical functions investigated. It is likely that Bush and Mosteller's assumption of path independence, for example, had to be made because, at the time, the mathematical problems raised by any alternative propositions could not be solved with the mathematical insights or tools available to them. Any systematic branch of learning faces both technical and conceptual limitations that help shape what can and cannot be claimed.

See also: Mathematical Psychology: History; Mathematical Psychology; Memory Models: Quantitative; Model Testing and Selection, Theory of; Psychology, History of (Twentieth Century); Stochastic Models.

Bibliography

Bower, G.H., 1966. Mathematical learning theory. In: Hilgard, E.R., Bower, G.H. (Eds.), Theories of Learning, third ed. Prentice-Hall, Englewood Cliffs, NJ, pp. 334–380.

Bower, G.H., Hilgard, E.R., 1981. Theories of Learning, fifth ed. Prentice-Hall, Englewood Cliffs, NJ.

Bush, R.R., Mosteller, F., 1955. Stochastic Models for Learning. Wiley, New York.

Cox, C., 1997. Estes, William K. In: Sheehy, N.P., Chapman, A.J., Conroy, W.A. (Eds.), Biographical Dictionary of Psychology. Routledge, London and New York, pp. 174–175.

Ebbinghaus, H., 1885/1913. Memory: A Contribution to Experimental Psychology (H.A. Ruger, C.E. Bussenius, Trans.) (accessed 19.12.12). http://psychclassics.yorku.ca/Ebbinghaus/index.htm.

Estes, W.K., 1950. Toward a statistical theory of learning. Psychological Review 57, 94–107.

Estes, W.K., 1955a. Statistical theory of spontaneous recovery and regression. Psychological Review 62, 145–154.

Estes, W.K., 1955b. Statistical theory of distributional phenomena in learning. Psychological Review 62, 369–377.

Estes, W.K., 1959. The statistical approach to learning theory. In: Koch, S. (Ed.). Psychology: A Study of a Science. Study 1. Conceptual and Systematic, vol. 2. McGraw-Hill, New York, pp. 380–491.

Estes, W.K., 2001. Mathematical learning theory, history of. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of Social and Behavioral Science. Elsevier, Oxford, pp. 9391–9393.

Estes, W.K., Burke, C.J., 1953. A theory of stimulus variability in learning. Psychological Review 60, 276–286.

Hull, C.L., 1951. Essentials of Behavior. Appleton-Century, New York.

Hull, C.L., 1952. A Behavior System. Yale University Press, New Haven, CT.

Koch, S., 1954. Clark L. Hull. In: Estes, W.K., Koch, S., MacCorquodale, K., Meehl, P.E., Mueller Jr., C.G., Schoenfeld, W.N., Verplanck, W.S. (Eds.), Modern Learning Theory: A Critical Analysis of Five Examples. Appleton-Century-Crofts, New York, pp. 1–176.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.

Lachman, R., Lachman, J.L., Butterfield, E.C., 1979. Cognitive Psychology and Information Processing: An Introduction. Lawrence Erlbaum, Hillsdale, NJ.

Logan, F.A., 1956. Koch's analysis of Hull. Contemporary Psychology 1, 3–5.

Lovie, A.D., 1992. Context and Commitment: A New Psychology of Science. Harvester Wheatsheaf, London.

Lovie, S., Lovie, P., 2004. A privileged and exemplar resource: traumatic avoidance learning and the early triumph of mathematical psychology. History of Psychology 7, 248–264.

Solomon, R.L., Wynne, L.C., 1953. Traumatic avoidance learning: acquisition in normal dogs. Psychological Monographs 67 (4). (Whole No. 354).

Weinstock, S., 1954. Resistance to extinction of a running response following partial reinforcement under widely spaced trials. Journal of Comparative and Physiological Psychology 47, 318–322.

Learning: Organizational

Silvia Gherardi, University of Trento, Via Verdi, Trento, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Organizational learning is a relatively recent metaphor for the organization that matches two concepts – learning and organization – and enables exploration of the organization as if it were endowed with a stock of knowledge, skills, and expertise. A short history of the concept will illustrate its development in organization studies. During the 2000s, not only did attention to processes and temporality increase, but also an epistemology of becoming appeared where the boundaries between learning and knowing, order and disorder, organization and organizing became conflated. Knowing took the place of knowledge, and instead of considering knowledge as an object or a resource, the concept of knowing makes it possible to see it as a collective knowledgeable doing that is situated in working practices.

Introduction

Organizational learning is a metaphor for the organization that matches two concepts – learning and organization – and enables exploration of the organization as if it were endowed with a stock of knowledge, skills, and expertise. It enables exploration of the organization as if it were a subject that learns, deals with knowledge, and reflects on experiences.

A metaphor works by matching what is distant with what is close; that is, similar features with dissimilar ones; and it is a cognitive tool that develops creativity and social imagination and reveals the importance of language and symbols in the construction of reality and in the formulation of theory. In this way, cognitive transpositions come into operation that make it possible to imagine and to talk about the object in question (the organization) as if it were another already-known object (learning).

Therefore, the study of organizations as if they were systems of knowing and learning may be of assistance in the design of knowledge-intensive organizations and of networks of organizational practices that create and transfer knowledge. At a theoretical level, the representation of learning in organizations may stand as a valid theoretical alternative to the rational organization because it replaces the classical concept of the organization that acts according to the principles of rational choice with that of the organization that acts according to principles of experimentation, of trial and error, of success and failure, and of discovery and invention. Nevertheless, we shall see how in the historical development of the field, several streams of research have arisen from controversies on how to understand what constitutes knowledge and why it should be studied (to manage it or to understand it).

The following sections outline the history of the concept of organizational learning, the divergences within the field of learning and knowing as they grew in the 1990s, and further developments in the field.

A Brief History of the Concept of Organizational Learning

The first mentions of the concept can be traced back to March and Simon (1958), but it was for long peripheral to organizational theory. March has always alluded to learning and to the link between learning and decision-making processes, although only few of his works are specifically devoted to this topic. This inquiry, which was sociological, continued with Cyert and March (1963) and conceived learning as changes in organizational routines that affect future behavior.

The study generally recognized as pioneering the field is by Argyris and Schön (1974), and it illustrates how researchers interested in knowledge have examined specific organizational phenomena in terms of learning defined in a variety of ways:

1. *As learning curves*, a concept that rests on the principle that the man-hours required to produce a unit of output decrease in proportion to the experience accumulated. Productive processes display characteristic learning curves in relation to a set of organizational factors.
2. *As behavioral change through trial and error*. Learning is a self-regulating process that detects errors and corrects them or, if preferred, checks actions against their outcomes and makes appropriate adjustments to subsequent actions.
3. *As adaptation to the environment*, and therefore as the capacity to 'fit in,' to adjust to external stimuli that must be discovered opportunely and rapidly. Learning is therefore synonymous with structural change and acts as a signal of successful evolution.
4. *As economy of search*. The economizing of cognitive energies in problem solving has been demonstrated within this stream of research. New solutions are first sought in familiar settings and then, because the costs of the search increase proportionately, in progressively more unfamiliar ones.
5. *As the modification of mental maps*. Since Argyris and Schön (1974) pointed out the gap between what organizations say they do (exposed theory) and what they actually do (theory-in-use), learning has resided in the ability to bring explanations and the ideas guiding what is done into alignment with the organizational changes introduced.

Since its early theorization in the 1970s, the concept of organizational learning has related to the management of organizational knowledge in relation to change in cognition and beliefs or in behavior and action. It is argued that one of the main shortcomings of the concept of organizational learning is that the traditional psychological models of learning

(based on stimulus-response theory) have been transferred acritically to the study of organizational learning, and with them a realist ontology. An authoritative commentator has remarked that "people in organization theory began to talk about learning … just about the time psychologists began to desert the concept" (Weick, 1991).

The growth of the field began in the 1980s, and for most scholars of the time, organizational learning and its cognate concept the learning organization were linked to change in organizational knowledge as a function of experience. Although the two terms (organizational learning and the learning organization) were broadly accepted, there was little consensus on their meaning, and their ambiguity served to separate two communities of scholars pursuing different research interests (Easterby-Smith et al., 1998).

Two bodies of literature developed according to whether the aim was analytic and descriptive or action oriented and prescriptive. The former focused on 'How does an organization learn?' The latter focused on 'How should an organization learn?' Peter Senge (1990), to whom the term learning organization is attributed, worked around two main questions: How can we build organizations in which continuous learning occurs? And what kind of person can best lead the learning organization?

From the late 1980s to the beginning of the 2000s, a comingling of several streams of research occurred (Argote, 2011) and a focus on organizational processes captured the attention of a large number of scholars. For the latter, the concept of organizational learning related to strategies and mechanisms to develop and exploit the learning capabilities of employees (Senge, 1990) in ways so that organizational performances improve.

The organizational processes that attracted the most attention were as follows:

1. *Knowledge creating*, which relates to the activity that occurs inside organizations to generate new knowledge (Huber, 1991) and directs attention to how knowledge is co-created (Antonacopoulou, 2009) and how creativity may affect the elaboration of experience. The study of knowledge creation subsequently linked with organizational creativity and creative industries (Kazanjian and Drazin, 2012), while the following two subprocesses initially interested more scholars (Argote, 2011).
2. *Knowledge retaining*, the process by which organizational knowledge persists through time (Kieser and Koch, 2008) or decays. Especial interest developed around transactive memory systems (Hollingshead, 2001).
3. *Knowledge transfer*, the process that analyzes the mechanisms whereby knowledge may or may not transit from one organization to another (Hansen, 1999), and in relation to competitive advantage and strategic behavior (Teece et al., 1997).

We may say that, historically, the relation between organization and learning has been explored in terms of change, under the influence of psychological models of learning transferred to organizations, and in terms of an intrinsic quality of organizational performance, be it an identity as in the learning organization or in organizational routines as a source of competitive advantage. The polysemy of the term 'organizational learning' may explain the richness and diversity of the studies conducted under its aegis.

The Growth of the Field and the Widening Divergences within It

The growth of the field of studies centered on the organization-as-learning metaphor may be understood in several ways, both internal to the ambiguity of the concept that explains its plasticity, and in relation to the expansion of the knowledge society, economy, and technological innovation. While an attempt was made to unify the field within a well-defined framework (Crossan et al., 1999), and while much debate was conducted on what organizational learning 'is,' new research directions emerged from the proliferation of definitions and ways to frame the object of research and the changing context in which it was studied.

In a scenario of rapid economic and technological changes, knowledge came to be seen as an invisible asset, and many consultants and firms caught on to the commercial significance of organizational learning and the managerial interest in bringing knowledge under control at the service of production or innovation. The speed of technological change means that there is constant pressure on firms to reduce the time from the conception of new products to their launch, and communications innovations mean that companies, and the knowledge workers within them, may have to adapt radically and in unforeseen directions. As a consequence, firms cannot rely on established practices: they need to learn new ways to do things. The advance of globalization is an additional source of pressure. It began with the opening up of labor markets in Asia in the early 1980s, and in Central Europe in the late 1980s, and it has accelerated with the establishment of global financial markets and global consumer markets fueled by the spending power of new middle classes in emerging economies such as India and China. In order to compete on a global scale, companies must radically increase their scales of operation in order to keep up with the evolution of new technologies and businesses anywhere in the world.

The preceding factors have generated a widely held view that the most important contributor to corporate competitiveness in the future will be the ability of each organization to learn faster than its rivals. The exploitative ethos is particularly evident in the economics of knowledge and knowledge management, whose starting point has been identification of knowledge as a production factor distinct from the traditional ones of capital, labor, and land. The resource-based theory of the firm has conceptualized knowledge as 'core competencies' or 'core capabilities,' naturalizing the relationship by means of the metaphor of the tree of knowledge: the trunk and major limbs are core products, and the root system is the core competence (Prahalad and Hamel, 1990). The reification of knowledge has grown more overt with the 'objectified transferable commodity' envisaged by the knowledge management approach, which treats knowledge as practically synonymous with information created, disseminated, and embedded in products, services, and systems.

In a way similar to the controversy between organizational learning and the learning organization, a new controversy arose

in the 1990s between organizational knowledge and knowledge management. Whilst the former was more focused on understanding how knowledge is created through the transformation of tacit and explicit knowledge, the latter focused more on managing learning and knowledge. Whereas Nonaka and Takeuchi's (1995) book became the reference point for both stating and criticizing this conception of knowledge transformation within an academic milieu, in the field of consultancy the idea of managing knowledge was put forward in relation to the potentialities of information technology (IT) systems and the reuse of knowledge.

The dilemma – as March (1996) expressed it – is between exploration and exploitation in the use of knowledge. The same dilemma is also reflected in the way knowledge is conceived and situated in the relationship between the knowing subject and the object of knowledge. In order to illustrate a different direction that the field of study took after the 2000s, we need to go back to Weick and Westley's (1996: 440) definition of the concept of organizational learning as an *oxymoron*: 'organizing and learning are essentially antithetical processes,' since organizing is linked to processes of ordering while learning recalls disorganizing and increasing variety. How can the two processes happen at the same time?

From Organizational Learning to Knowing in Practice and Becoming

The semantic shift that involved the increasing substitution of nouns with verbs in organization studies from 1996 onward (Clegg and Hardy, 1996) was accompanied by a different epistemology with a different way of looking at organizational phenomena. Instead of organization, many scholars preferred to look at organizing and, instead of knowledge, to investigate knowing. This line of inquiry into the nature of knowledge and its changing status in contemporary society is the continuation of the epistemological debate that Cook and Brown (1999) represented in oppositional terms as an epistemology of possession, constructed around the vision of knowledge as an object, and an epistemology of practice, where practices are the sites of knowing, learning, and organizing.

During the 2000s, not only did attention to processes and temporality increase, but also an epistemology of becoming appeared where the boundaries between learning and knowing, order and disorder, organization and organizing became conflated. Knowing took the place of knowledge, and instead of considering knowledge as an object or a resource, the concept of knowing makes it possible to see it as a collective knowledgeable doing that is situated in working practices.

The relation among learning, becoming, and organizing has been explored in order to illustrate how the three concepts can be conceived as verbs that perform themselves. They can be 'understood as an 'intelligent life in between' in that it oscillates between complexification and simplification, de- and re-construction, de- and re-territorialization' (Clegg et al., 2005: 158). Learning is therefore considered as taking place in between order and chaos, as the oxymoron that organizational learning illustrates. Within an epistemology of becoming, organizations are not treated as separate entities or accomplished events, but as enactments, as unfolding processes (Tsoukas and Chia, 2002).

During the same years, and within similar debates on knowledge, the so-called practice turn (Schatzki et al., 2001) contributed to a resumption of the concept of practice in organizational theory; and from the field of organizational learning, which rediscovered the concept, came a new solution to an old problem. Practice became the third way between a mentalistic and cognitive model of learning in organizations and the commodification of knowledge performed by knowledge management. Knowing in practice is a way to study knowledge as an activity and to look at it as it happens. Its main questions are "How do people 'do' knowledge as a collective endeavor?" and "How is knowledge transformed when it is being used?"

From the beginning of the 1990s onward, the stream of research put in motion by a social learning perspective (Brandi and Elkjaer, 2011) continued through the success of the concept of communities of practice (Wenger, 1998). Its subversion in favor of the idea of the practices of a community formed a 'bandwagon of practice-based studies' (Corradi et al., 2010) around the ideas that practice is the site of knowing and that learning takes place within sociomaterial relations (Orlikowski, 2002).

Practice is the concept that allows the processes of 'knowing' at work and in organizing to be articulated as historical, material, and indeterminate processes. In practice-based theorizing, different traditions in the study of learning converge without losing their identity: situated learning theory offers the concept of community of practice, actor–network theory brings the idea of knowledge grounded in materiality, activity theory offers a pragmatic theory of action, and workplace studies, with their ethnographic study of knowledgeable communities, bring a research method that is anti-individualistic.

In regard to future directions for research on organizational learning, there are as many directions as there are lines of inquiry. Nevertheless, the second edition of Easterby-Smith and Lyles's handbook contains an interesting observation on the lack of research on actual learning processes and knowledge, since these have been mainly considered as dependent variables influenced by other phenomena. This is because, according to Easterby-Smith and Lyles (2011: 15), "we do not know the constructs that influence knowledge or learning utilization," and because real-time learning poses numerous methodological questions, we can expect that the field will continue to grow and host further controversies.

See also: Information and Knowledge, Organizational.

Bibliography

Antonacopoulou, E.P., 2009. Impact and scholarship: unlearning and practicing to co-create actionable knowledge. Management Learning 40 (4), 421–430.

Argote, L., 2011. Organizational learning research: past, present and future. Management Learning 42 (4), 439–446.

Argyris, C., Schön, D.A., 1974. Theory in Practice. Jossey-Bass, San Francisco.

Brandi, U., Elkjaer, B., 2011. Organizational learning viewed from a social learning perspective. In: Easterby-Smith, M., Lyles, M. (Eds.), Handbook of Organizational Learning and Knowledge Management. Wiley, Chippenham, pp. 23–42.

Clegg, S.R., Hardy, C., 1996. Organizations, organization and organizing. In: Clegg, S.R., et al. (Eds.), Handbook of Organization Studies. Sage, London.

Clegg, S.R., Kornberger, M., Rhodes, C., 2005. Learning/becoming/organizing. Organization 12 (2), 147–167.

Cook, S., Brown, J.S., 1999. Bridging epistemologies: the generative dance between organizational knowledge and organizational knowing. Organization Science 10 (4), 381–400.

Corradi, G., Gherardi, S., Verzelloni, L., 2010. Through the practice lens: where is the bandwagon of practice-based studies heading? Management Learning 41 (3), 265–283.

Crossan, M.M., Lane, H.W., White, R.E., 1999. An organizational learning framework: from intuition to institution. Academy of Management Review 24 (3), 522–537.

Cyert, R.M., March, J.C., 1963. A Behavioral Theory of the Firm. Prentice-Hall, Englewood Cliffs, NJ.

Easterby-Smith, M., Lyles, M., 2011. The evolving field of organizational learning and knowledge management. In: Easterby-Smith, M., Lyles, M. (Eds.), Handbook of Organizational Learning and Knowledge Management. Wiley, Chippenham, pp. 1–20.

Easterby-Smith, M., Snell, R., Gherardi, S., 1998. Organizational learning and learning: diverging communities of practice? Management Learning 29 (3), 259–272.

Hansen, M., 1999. The search-transfer problem: the role of weak ties in sharing knowledge across organizational subunits. Administrative Science Quarterly 44 (1), 82–112.

Hollingshead, A., 2001. Cognitive interdependence and convergent expectations in transitive memory. Journal of Personality and Social Psychology 81 (6), 1080–1089.

Huber, G., 1991. Organizational learning: the contributing processes and the literatures. Organization Science 2 (1), 88–117.

Kazanjian, R.K., Drazin, R., 2012. Organizational learning, knowledge management and creativity. In: Munford, M. (Ed.), Handbook of Organizational Creativity. Elsevier, London, pp. 547–568.

Kieser, A., Koch, U., 2008. Bounded rationality and organizational learning based on rule changes. Management Learning 39 (3), 329–347.

March, J.G., 1996. Exploration and exploitation in organizational learning. In: Cohen, M., Sproull, L. (Eds.), Organizational Learning. Sage, Thousand Oaks, CA, pp. 101–123.

March, J.C., Simon, H.R., 1958. Organizations. Wiley, New York.

Nonaka, I., Takeuchi, H., 1995. The Knowledge-creating Company. Oxford University Press, Oxford.

Orlikowski, W.J., 2002. Knowing in practice: enacting a collective capability in distributed organizing. Organization Science 13, 249–273.

Prahalad, C.K., Hamel, G., 1990. The core competence of the corporation. May–June. Harvard Business Review, 79–91.

Schatzki, T.R., Knorr-Cetina, K., von Savigny, E. (Eds.), 2001. The Practice Turn in Contemporary Theory. Routledge, London.

Senge, P., 1990. The Fifth Discipline: The Art and Practice of the Learning Organization. Doubleday, NY.

Teece, D., Pisano, G., Shuen, A., 1997. Dynamic capabilities and strategic management. Strategic Management Journal 18 (7), 509–533.

Tsoukas, H., Chia, R., 2002. On organizational becoming: rethinking organizational change. Organization Science 13 (5), 567–582.

Weick, K., 1991. The nontraditional quality of organizational learning. Organization Science 2 (1), 116–124.

Weick, K.E., Westley, F., 1996. Organizational learning: affirming an oxymoron. In: Clegg, S.R., Hardy, C., Nord, W. (Eds.), Handbook of Organization Studies. Sage, London, pp. 440–458.

Wenger, E., 1998. Communities of Practice: Learning, Meaning, and Identity. Cambridge University Press, Cambridge, UK.

Learning Theories and Educational Paradigms

Robert D Tennyson and Ashley Volk, College of Education and Human Development, University of Minnesota, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article is an introductory review of learning philosophies and theories directly associated with human learning. We present an introduction to educational foundations followed by major sections on the two foundational areas of learning philosophy and learning theory. The focus of this article is on educational foundations; the goal is to help the individual understand the foundational relationships between learning philosophy and learning theory. The purpose is to set the underlying principles by which learning decisions are made.

Introduction

This article provides a review of the learning philosophies and theories directly associated with student learning. Our goal is to present sufficient information in the subject areas of learning philosophy and learning theory so that individuals can begin employing theories of learning without first having to acquire formal and disciplined study in educational foundations. However, it is not our intention to replace or to negate the importance of formal study in educational foundations. Rather it is our hope that the educator will seek out additional study in this important domain of instructional theory. That is, this article should provide a means for educators to formalize their personal educational foundations in the areas of learning philosophy and learning theory.

We first present an introduction to educational foundations followed by major sections on the two foundational areas of learning philosophy and learning theory.

The focus of this article is on educational foundations; the goal is to help the individual understand the foundational relationships between learning philosophy and learning theory. The purpose is to set the underlying principles by which teaching process decisions are made.

The *Why* and *How* of Learning

Philosophy provides the *why* for learning theories. From that philosophical foundation or underlying belief of human learning, a learning theory describes the *how* of the cognitive system. This description of learning is typically a model that shows the basic components employed in human cognition. For example, behavioral models describe the how of the system without mentioning internal components or processes. The explanation of *why* in behavioral models range from physical needs (e.g., hungry, pain, pleasure) to placement of reinforcers to increase probability of desired behavior. In contrast are cognitive models of learning that focus on the internal variables and mechanisms to describe the *how* of cognition. The *why* is more elusive in cognitive models because of the internal nature of human behavioral control. There is the implied assumption that internal motivation factors influence progress and effort in learning. Each of these positions is elaborated in this article.

Instructional theories are prescriptive models that identify the variables and conditions that are necessary to improve learning. The variables and conditions defined in instructional theory are in large part the products of their underlying learning philosophy and learning theory. The robustness of an instructional theory therefore is in direct relationship to the scope and depth of the learning theory. The predominance of behavioral theories since the 1940s has provided a rich set of instructional theories for classroom learning. For example, the works of Skinner in transferring learning theory to instructional theory set the tone for instructional models that still dominate classroom practice. The growth of cognitive psychology and constructivism is now having an effect on educational paradigms.

Learning Philosophy

To fully understand contemporary educational paradigms, it is important to trace its roots through the rich philosophical traditions embedded in psychological history. Although the modern tradition of psychology is typically credited with a brief history, its ancient antecedents can be traced to the earliest inquiring minds of philosophers and theologians who sought answers and explanations to the workings of the human mind. Precursor inquiry into such themes as human motivation, thinking and memory, perception, dreams, and behavior established a vital continuity between these earliest thinkers and the scientific inquiry of contemporary psychological theories. Therefore, the distinction between these ancient antecedents and modern educational paradigms lies not in the kinds of questions asked, but in the methods used to answer such inquiry.

Self, Nature, Nurture, and Society

The range of explanations that philosophers have focused on in terms of human learning includes four basic constructs: self, nature, nurture, and society. Each of these constructs represents contrasting concepts for explaining human learning. However, most practical explanations of learning typically involve two or more of these four constructs.

Self

Self refers to the concept that humans have direct responsibility and regulation for the conduct of their lives: Learning is the direct responsibility of the individual. Experience is an important component of the self-regulation aspect of the human because the individual can use that experiential knowledge to make increasingly complex decisions. The concept of *self* is highly influenced by effort and progress in learning. Life is not deterministic or controlled by forces outside the self. The *self* takes on the responsibility of direction in learning, and the effort and affects necessary to learn. An important component of the *self* is the development of personal criteria for problem solving and decision making. This personal criterion arms the individual with the means for self-observation and judging. Regulation is a matter of self-monitoring and self-imposed control over external and internal factors.

Nature

Nature is the amount of control and influence that an individual's inheritance has over learning and behavior. The extreme position is that the person's DNA regulates behavior through such variables as motivation, personality, and intellect rather than environmental factors. Proponents of the *nature* explanation of learning have highly impressive empirical data to support their reliance on genetic functions to regulate learning. Depending on the specific theory, proponents of nature place the control of genes from a low of 50% to a high of above 75%. Cross-cultural studies show findings that identical twins separated at birth exhibit similar behaviors that are not due to their respective environments. In comparison to *self*, control and regulation are manifested within the constraints of the genetic variability of each individual. That is, the individual lives within the borders established at time of conception.

Nurture

In contrast to *nature* or *self* as the controller and regulator of learning, is the concept that the environment determines through conditioning what a person learns and how they behave. This *nurture* explanation is based on the influences coming from the learner's environment. For the most part, proponents of the *nurture* concept of learning maintain that behaviors are a result of environmental conditions. Given the all-inclusive role of the environment in learning, nurture proponents' favor a highly structured environment governed by constant manipulation by superiors. In other words, change the environment and you change the behavior. Learning occurs through direct manipulation of external conditioning of responses through variables such as reinforcements and rewards. In other words, others deliberately shape learning in either overt or covert controls of the individual's environment. At birth the individual is a blank slate without either predisposed genetic constraints or internal means of self-control.

Society

The fourth basic construct in explaining human learning is *society*. *Society* is a twofold construct: internal-self correspondence and external collaboration. In the latter part, cultural influences are highly regarded as shaping the direction of the individuals learning path. In the former part, the individual engages in internal cognitive discourse. The roles of internal correspondence and external collaboration explain that individual learning is a product of the interactions of these two concepts. In a sense, society features a combination of self, nature, and nurture. The debate between this construct and the other three is on degree of influences. For example, nurture implies a direct manipulation of conditioning paradigms whereas society is less manipulative and controlling; it is more self-observing of others' behaviors but being to able to be internally selective. That is, the individual is within the environment but is aware of what to take from it.

The following subsections review the evolution of philosophical thought. The goal of the presentation is to help the reader in forming their individual learning philosophy foundation. At the conclusion of this section on educational philosophy, the reader should attempt to put in writing their individual philosophy of learning. The value of a formal learning philosophy is that it is used to improve validity and consistency in the ID decision-making process.

From the earliest of inceptions forward, the waxing and waning in popularity of numerous established philosophical positions has shaped modern psychology. Deeply embedded in the *zeitgeist* or spirit of the times, the prevailing philosophical position was often upstaged when new traditions, out of protest or direct revolt, attempted to improve upon the shortcomings of the then popular position. In this manner, new philosophical positions challenged the zeitgeist and gained momentum from the established position by proposing new concepts and inquiry techniques that remained popular until a new revolution took place. Frequently, a scientific discovery, brilliant for its time, was rejected by the prevailing position only to be accepted years or decades later when the zeitgeist complimented such methods. Modern psychology, a reflection of contemporary zeitgeist, consists not of a single tradition that boasts the total truth, but a rich and complex incorporation of all that has gone before. Following next in this section is a review of the growth in philosophical explanations of the *why* in human learning.

Dualism

The origins of Western learning philosophy can be traced to Plato (427–347 BC), a disciple of Socrates, who was the first Western thinker to distinguish three aspects of the human soul that contemporaries correspond with intellect, emotion, and will. Furthermore, this philosophy, termed *dualism*, maintained a separation of the mind and body; suggesting that intellect or reason was connected with the spirit and was untouched by the behavior exhibited by the individual. Originating with Plato, this dualistic theme provides the material for much controversy throughout the history of psychology. Aristotle (384–323 BC), a student of Plato, further distinguished various psychological functions and suggested that the soul might be divided into cognitive functions and emotional and moral functions. It was not until Quintillion, however (35–95 AD) that the existence of unique variations in individual ability were suggested,

though structured study of individual differences evolved much later in history.

Reasoning

Early Christian scholars augmented the mind–body dualism proposed by the early Greek philosophers of the mind synonymous with the concept of a soul independent of physical causes. Aquinas (1225–74), a Catholic theologian, expanded on the concepts introduced by Plato by further subdividing the functions of the mind to include sensation, perception, memory and reproductive imagination, and reasoning and creative imagination. Philosophical inquiry into reason was in harmony with Western Christendom and its views of faith and study of the soul into the middle ages. At the dawn of the modern age, however, science began to challenge many of these traditional beliefs. Breaking from the traditionally acceptable truth that the earth was fixed at the center of the universe, Galileo Galilei (1564–1642) further suggested that all that existed was nothing more than particles of matter in motion. These challenges catalyzed the movement toward science and reason becoming the authority, freeing scientific inquiry from the rigid doctrines that had dominated for centuries.

Compromising the initial challenges of Galileo, French philosopher, mathematician, and scientist Descartes (1596–1650) accepted the dualistic philosophy suggesting that both mind and body are essential to the study of the totality of human behavior. Descartes further believed that humans are born with all knowledge and that the relationship between mind and body was one of intimate mutual interaction. The mind possessed only the function of thought and all other processes were functions of the body, though mind, and body influenced one another. This interactionist position established the foundation for a more extreme philosophical view emerging in England in the eighteenth century, *Empiricism*.

British Empiricism

Agreeing with the dualistic theory of Descartes, a group of British philosophers accepted the position that the mind is the seat of all knowledge, that the proper subject for the study of psychology is the mind of the human, and that the acceptable method of inquiry is empirical introspection. These British empiricists opposed the nature theory of innate ideas suggesting that all knowledge comes through the senses and that the mind grows through progressive accumulation of sensory experience.

The major proponent of the empiricist school was Locke (1632–1704), who initiated the *nature/nurture* debate by suggesting the concept of *tabula rasa* that the mind of a human is, at birth, a clean and empty slate on which experience would write. In harmony with the mechanistic concepts derived from physics, Locke suggested that knowledge reduced to its simplest particle consist of ideas that, like atoms, combine to form more complex concepts.

Law of Association

Locke's immediate successor, Berkeley (1685–1753) filled an obvious void of empiricism by introducing the concept of associations that provided the mortar that connects simple ideas into mental elements. Hume (1711–76), who established laws of the mental universe of which the *law of association* of ideas was one, further advanced this associationist theory within British empiricism. In addition, Hume differentiated the concepts of impressions and ideas.

Credited as the founder of associationism as a formal school of thought, Hartley (1705–57) collated and clarified the previous work of the empirical philosophers while additionally suggesting that the law of association was one of continuity. Carrying these concepts to a further extreme, Mill (1773–1836) applied a more pure concept of mechanism to associationist thought by proposing that the mind is merely a machine, totally passive, and void of will. This extreme position in arguing for nature influenced the later emergence of contemporary behaviorism.

Positivism

Concurrent with the emergence of empiricism as the dominant philosophical position in the mid-nineteenth century, two less powerful forces established themselves as influential to the emerging traditions. *Positivism*, a theme popular in Europe, proposed that facts be accepted only if objectively observed through the methods of science. Materialism, which supported the theory of positivism, focused on the physical aspect of mental phenomenon. As these philosophical positions interacted to shape a more empirical and mechanistic approach to the study of the human mind, psychology finally separated in the late nineteenth century from its philosophical traditions to emerge as a disciplined science.

Experimental Physiology

During the time that associationism dominated philosophical thought focusing on the importance of the senses; a group of German scientists, boldly embedded in *experimental physiology* were applying an empirical approach to describe how the senses worked. Helmholtz (1821–94), a German physicist, applied empirical methods to study the speed of nerve impulses, and to develop theories of hearing and vision. Weber (1795–1878), experimenting in threshold perception, established the first quantitative law of psychology with his work in just-perceivable differences. Fechner (1801–87), a German physiologist, philosopher, and psychologist, developed an interest in sensations while experimenting with light, resulting in the emergence of the field of psychophysics. Additionally, Fechner applied his philosophical ideas to relate the mental and physical worlds quantitatively. These applications of the experimental approach to the study of the mind merged with the prevailing disposition of empiricism to establish the foundations for the emerging philosophical position that was to dominate the next generation in psychology.

Structuralism

Continuing in the empirical physiological traditions firmly established in Germany, Wundt (1832–1920), considered the founder of modern psychology, established the first systematic school of thought in psychology later titled by Titchener, his

successor, as *structuralism*. Reflecting the zeitgeist in Europe, the rationalists additionally influenced structuralism, specifically the work of Galton (1822–1911), who held that man is born with all knowledge and through our senses, and these concepts are integrated and refined. Galton proceeded in the scientific spirit to investigate the genetics of intelligence and invent the first test to measure this realm. Galton believed that intelligence consisted of a general ability and several special abilities, thus expanding the earlier empiricist position to include innate attributes such as creativity of humans.

Wundt combined the rationalist, empiricist, and scientific spirit to advance the structuralist position that soon dominated the new emerging field of psychology. Believing in the creative synthesis of the mind, Wundt became deeply interested in the sensory process, propelling him toward a formal study of psychology. For the next 46 years, Wundt refined the structuralist position from his first established psychology laboratory in Leipzig. From there, Wundt became a prolific writer and teacher of the world's leading psychologists.

Experimental Psychology

Ebbinghaus (1850–1909), intrigued by the experimental advances of the structuralists, specifically explored the higher mental process of memory through a method utilizing nonsense syllables. By exploring the complexity of a task and cognitively relating it to assessing of higher mental processes, Ebbinghaus became one of the first psychologists to explore learning under experimental conditions.

Advancing the work of Wundt was his devout student Titchener (1867–1927) who translated the writings of Wundt into English, formally bringing structuralism to America. As head of the department of psychology at Cornell University, Titchener continued to follow in the footsteps of Wundt, invoking the principle of psychophysical parallelism, a form of dualism that suggested a parallelism between the mind and body, affecting one another though not interactive. Structuralism further advanced the concepts of sensory states, images, and affective states of consciousness. Titchener's work gave structuralism an American flavor different from its German roots. Unfortunately his unwillingness to remain open to new advances in psychology limited the structuralist point of view in North America and later led to its collapse as a formal school of thought. Propelling from the weaknesses imbedded in the propositions of structuralism was a new school of thought that reflected the zeitgeist of advancing times.

Functionalism

Emerging out of opposition to structuralism, a strongly pragmatic and practical form of new psychology known as *functionalism* dominated toward the end of the nineteenth century. Thriving in the more accepting climate of the North American scene, functionalism enjoyed a position of greater importance in America than it did in Europe. Opposing the structural claims of Wundt and Titchener who studied conscious contents, the functionalists emphasized mental acts or processes as the proper subject of study in psychology. Gaining momentum from the earlier work of Darwin (1809–82) whose theory of human evolution emphasized the importance of the mind in terms of its usefulness in adapting to the environment provided the material for functionalism. These works combined with the earlier mentioned research by Galton in mental inheritance, introduced to psychology the possibility of scientific methods applicable to both animal and humans and the use of techniques other than experimental introspection.

William James, considered by many as the greatest American psychologist, developed the first American theory of human learning – *functionalism*. This theory was in part based upon the work of Spencer (1820–1910), who held a utopian view of social Darwinism, which suggested that man and society without imposed order, would naturally evolve toward perfection. James (1842–1910), while an instructor in physiology at Harvard, established the first psychology laboratory in America where he advanced the work of functionalism with his theory of *stream of consciousness*. James advocated that mental process be studied as such and not as static pieces of consciousness. He further maintained that the purpose of consciousness must have some biological utility in order for its continued survival.

Continuing in the functional spirit, Hall (1844–1924) applied evolutionary theory particularly to the study of children, founding the child study movement in America. Cattell (1860–1944) augmented the practical aspect of functionalism in application to his work on reaction time and individual differences by initiating the mental test movement. Cattell became an ambassador of the functional school, promoting applied psychology throughout America. John Dewey (1859–1952) attacks the earlier philosophies of molecularism (analysis of small units) and reductionism (reduce elements to their smallest components) calling instead for a more functional psychology. Assuming leadership of the functional school during his appointment at the University of Chicago, Angell (1869–1949) formalized this philosophy into a formal school of psychology.

The work of the functionalists catalyzed much advancement in psychology through both direct and indirect influence due to its emphasis on a functional psychology versus one of structure. Improvement was seen in the area of research methodology, mental tests, and objective analysis of behavior. In addition, functionalism allowed for the study of children, the mentally retarded and the insane and introduced increasing research in animal psychology. Functionalism became the dominant school of thought in American psychology by the early 1900s and, though not formally dominant, remains an important orientation of contemporary American psychology. In the usual progression of history, the early 1900s saw a direct revolt against the dominant schools of structuralism and functionalism. The philosophy of the new challengers was later to become the dominant school of thought in American psychology.

Behaviorism

The progression of psychology toward applied experimental methodology, pragmatism, and emphasis on function combined with the acceptance of animal psychology and physiological influences set the scene for the behavioral revolution. Early work in conditioning by Russian physiologists provided much of the material for American behavioral

psychology. Sechenov (1829–1905), known as the father of Russian physiology, suggested that experimental approaches used in physiology should be applied to psychology. Further, Sechenov was also noted for his work in the role that reflexes and learning play in behavior. Bechterev (1857–1927) credited for developing reflexology, advanced the position that *external* behavior is the only valid consideration in scientific experimental inquiry. Following in these traditions, the Russian physiologist, Pavlov (1849–1936) showed that certain reflexes could be conditioned (Pavlov, 1927). Through his work in glandular secretions and motor movement, Pavlov initially included reference to mental states in his research with animals but in his later writings, excluded such references resulting in a purely objective stance. Most influential to American behaviorism was his research resulting in classical or Pavlovian conditioning, the process in which an initially neutral stimulus, called the *conditioned stimulus*, is repeatedly paired with a reinforcer or unconditioned stimulus, replacing the initial response, referred to as the unconditioned response with a conditioned response.

Applying these concepts in America, Watson (1878–1958) purposely intended to establish a new school of psychology in America, one that was to apply a purely objective and experimental approach to the study of behavior, both animal and human. Rejecting the teachings of the functionalist school, Watson suggested that the goal of psychology was to *predict* and *control* behavior, and he discarded all reference to consciousness. Watson was particularly influenced by the work of Thorndike (1874–1949), whose work advanced the development of animal psychology that culminated in his theories of trial and error learning. Reflecting on these advances, Watson applied the method of observation to comparative psychology in his work on fear responses using children and animals. Although pragmatic in practice, Watson's *objective* behaviorism was foundational to the work of neobehaviorists that further advanced behaviorism to its position of dominance in American psychology.

Neobehaviorism

Beginning in the 1930s and progressing into the 1960s was a second phase of behaviorism known as *neobehaviorism*. Proponents advocated explanatory systems and encouraged increased precision through common experimental methods and terminology. Applying postulates from both molercularism and Gestalt theory (which will be discussed further), Tolman (1866–1959) introduced the concept of *purposive* behaviorism that utilized a systematic treatment of the data in his research in learning and cognitive complexity based on studies done with animals in mazes. Employing a more theoretical than empirical approach, Guthrie (1886–1959) established *contiguous* conditioning based on an instrumental conditioning model that suggested that learning occurs through the pairing of a stimulus and a response, with reinforcement not necessary for learning to occur. Hull (1884–1952) applied concepts of mathematics to learning resulting in his *mathematico-deductive* theory of learning based on classical conditioning that emphasized habit strength and drive. Hull's theory was quite complex and usually stated in mathematical formulations. The most famous of the neo-behaviorists was Skinner (1904–90), developer of *operant* conditioning and its schedules of reinforcement. Skinner proposed that if the occurrence of an operant, or emitted response were followed by a reinforcing stimulus, the rate of responding would increase. Skinner's theory of operant conditioning has been of vital influence to the field of instructional development.

Gestalt

Parallel to behaviorism though independent from the American behaviorists, was a movement in Germany known as *Gestalt* psychology. Not unlike behaviorism, Gestalt emerged in opposition to both structuralism and functionalism, though in contrast to behaviorism, Gestalt psychology admitted the existence of consciousness and advocated the study of such as a whole entity rather than reducing it to elemental analysis. Based on the work of Kant (1724–1804), the Gestalt position focused on unity of perception and insight in learning. The founder of the school of Gestalt psychology, Wertheimer, together with his students Köffka (1886–1941) and Köhler (1887–1967), engaged in a direct crusade against Wundtian elementalism. Wertheimer's work in perceptual organization suggested that objects are perceived in the same unified manner as motion is perceived and that these perceptions are organized by a series of rules.

With the rise of anti-intellectual sentiment invoked by the rise to power of the Nazi political party, the leaders of Gestalt psychology fled from Germany reducing the movement to one of limited significance. As Gestalt psychologists migrated to the United States, they experienced some difficulty gaining popularity due to the strong behavioral sentiment dominating American psychology by the mid-1930s, particularly in light of the inherent language barrier. Strongly imbedded in behavioral tradition, American psychologists criticized Gestalt psychologists as employing poor experimental procedures with inadequate control of variables and weakly supported assumptions. Regardless, American psychologist Lewin (1890–1947) attempted to add a psychology of motivation to the existing Gestalt philosophy resulting in his *field* theory, a system that adds both motivation and social factors to the principles of Gestalt psychology.

Psychoanalysis

Although it did not strive to become a form of pure science, as did the other philosophical positions previously discussed, psychoanalysis emerged at the turn of the century as a theory applied to clinical psychotherapy. Psychoanalysis, founded by Freud (1856–1939), deviated from the other psychological schools in method, goals, and subject. Regardless of this lack of connection with the other schools of thought, psychoanalysis is mentioned here due to its profound influence on modern psychology. Psychoanalysis studied individuals in clinical observation, speculating on the unconscious, and applying these results to disturbed individuals in individual therapy. The obvious criticisms of psychoanalysis by the dominant schools of psychology were in its unscientific and nonempirical methods of data collection and inability to quantify its data.

Humanism

Emerging in opposition to psychoanalysis and behaviorism, the humanistic movement is titled the third force in modern American psychology. Growing out of the work of Maslow (1908–70), a personality theorist, and Rogers (1902–86), a psychotherapist, humanism attempts to draw psychology back to inclusion of human processes by advocating subjective experience, freedom of choice, individual uniqueness, and striving toward individual human potential within the broader frame of society. As it influences contemporary psychology, the humanistic movement has been particularly influential in the field of psychotherapy.

Constructivism

Constructivism is a philosophical branch of psychology that maintains that what an individual learns and understands is constructed by the individual. Educational constructivists differ in the degree to which they ascribe knowledge construction to the student. On one hand, certain constructivists view mental models as reflective of external realities, whereas, on the other hand, some constructivists see no independent reality external to the mental model of the individual. Furthermore, constructivist philosophy emphasizes the role of social interactions (e.g., student to student, teacher to student, and student to media) as influencing what students' learn. For educational purposes, most constructivists agree on the following attributes for a learning environment: (1) that students are active in constructing their own knowledge and (2) that social interactions are important to knowledge construction.

Cognitive Psychology

The final school of philosophical thought influencing the field of educational paradigms is cognitive psychology, the scientific study of mental events. Emerging as an empirical science in the late nineteenth century, many of the psychologists mentioned of the late 1800s were actually cognitive psychologists as they focused on mental events, including Wundt and James. Experiencing limited popularity during the domination of behaviorism, cognitive psychology was spurred on by the outbreak of World War II that employed the expertise of psychologists in the selection and training of military personnel. Successful at influencing such complex cognitive tasks, a group of psychologists became interested in mental structures and processes initiating the rebirth of cognitive psychology.

Facilitating the further growth of cognitive psychology were the areas of information science and specific studies in linguistics, both of which attempted to study mental structures and knowledge representation. Perhaps the most influential event in the advancement of cognitive psychology was the computer, which provided for objective measure of these processes. Cognitive psychology specifically addresses the importance of knowing, understanding, reasoning, remembering, and perception. We will further elaborate on cognitive psychology in the next section on learning theory.

Learning Theory

In this section we present an overview of learning theories that form the core knowledge for this competency area in educational foundations. The reader is encouraged to seek out additional literature in learning theories to further elaborate their individual core knowledge. The information presented in this section should arm the novice with enough core knowledge to make an initial statement regarding their personal learning theory.

Behavioral Psychology

Companioned with Freud's psychoanalytic theory and Maslow's humanistic approach, behaviorism has been titled the third dominant force to shape modern psychology. Although behaviorism is comprised of several heterogeneous subgroups, Skinner's radical behaviorism has endured the longest period of time, surviving for more than half a century. Furthermore, Skinner's work is seen as being the most important contribution to psychology during the post-World War II period. Despite its stature within American psychology, the past few decades have been witnessed a growing dissatisfaction with radical behaviorism while interest in psychological theories that acknowledge mediational processes, such as cognitive psychology, have been on the rise. However, behavioral theory offers much to the understanding of human learning and the means by which instruction can be designed.

Basic Assumptions of Behaviorism

The importance of behavioral theory is the fine toning of its basic underlining principles. For the most part, these assumptions have a rich tradition in the quantitative nature of behaviorism. The foremost assumption is that principles of learning apply equally to different behaviors and to different species of animals. This is a fundamental premise of the basic behavioral research paradigm. That is, it is entirely possible to study behavior in animals and generalize the findings to humans. A second assumption is that behavior can be studied most objectively when the focus of study is on stimuli and responses. A result of this assumption is that internal cognition is excluded from scientific study. All organisms, including humans are born with no inherited knowledge or experiences. Everything is learned. Learning is a change in behavior resulting from environmental events. Thus, learning can be described as *conditioning* resulting from a relationship between stimuli and responses. Finally, the most useful theories of learning tend to be parsimonious in content.

Background

Behaviorism developed at the start of the twentieth century as a reaction to the functional paradigm. Functionalism was America's first contribution to a theory of learning and represented an application of Darwin's theory of evolution to animal behavior. Based on that theory, functionalism explained that behavior was a result of the animal adapting to environmental variables and conditions. From this perspective, functionalism tried to explain observable behavior by studying

mentally derived rationales. This methodology used introspection as the means for discovering internal mental causes of behavior. Two early leaders in the behavioral revolution were Edward Thorndike and John Watson.

Edward Thorndike

The father of educational psychology is often bestowed upon Edward Thorndike because of his desire to bridge the gap between his laboratory research and findings and application into the classroom. Thorndike's theory of learning meets a behavioral tenet that theories be parsimonious. His classroom instructional strategies were contained within two laws of learning. The first, *law of effect*, states that responses to a situation that are followed by satisfaction are strengthened; responses that are followed by discomfort are weakened. This law set the tone for the behavioral paradigm that positive reinforcers are more effective than negative. Correct behaviors were rewarding while incorrect behaviors failed to elicit positive responses. In other words, punishment was not considered a useful means for learning. The second, *law of exercise*, states that stimulus–response connections that are repeated are strengthened; stimulus–response connections that are not used are weakened. Thorndike based an important part of classroom learning on the concept of drill and practice. Worksheets are still an integral component of classroom learning.

John Watson

While a student at the University of Chicago, John Watson, the founder of behavioral psychology, became dissatisfied with this dual methodology that considered both overt and introspective activity, particularly criticizing the failure of researchers to agree on data obtained through supposed observations of the conscious (mental) states of animals. Instead he proposed that psychology restrict itself only to the study of overt behavior, examining observable activities. His presidential address at the American Psychological Association in 1913 is often used as the formal start of behavioral psychology. In keeping with the parsimonious nature of behavioral theory, Watson proposed from his research studies two laws of learning. The first, *law of frequency*, states that the more frequently a stimulus and response occur in association with each other, the stronger that S–R condition would become. The second, *law of recency*, states that a response that has most recently occurred after a particular stimulus is the response most likely to be associated with that stimulus. In contrast to Thorndike's laws that were aimed for the classroom, Watson moved his laws into the realm of advertising. This came about because of his having to leave the academic community for New York City in the early 1920s. He was quite successful in applying his laws to the relatively new field of advertising.

Respondent Conditioning

What followed the work of Thorndike and Watson was a field of psychology acknowledging two basic forms for studying animal behavior. The first, *respondent conditioning* (also labeled as Type S – emphasis on the stimulus), is represented by observing how animals can be conditioned to behave by controlling the association between a stimulus and response (Pavlov, 1927). **Table 1** illustrates a situation in which a desired association is developed by introducing a *conditioned stimulus* (CS, a unique sound) with an existing association (in this case between giving food to an animal and the salivation process). This paradigm assumes that the CS will replace an *unconditional stimulus* (UCS, food pellet) because of the close proximity to the *unconditioned response* (UCR, salivation). The example shows that eventually when the CS is introduced, the response will follow paired, in turn, with the UCS. In time, the UCS is removed totally, and the association is between the CS and the *conditioned response* (CR). This form of behaviorism is still a major research area in contemporary psychology.

Table 1 Respondent conditioning procedure

Stage	*Stimulus*	*Response*
1	UCS (food pellet)	UCR (salivation)
2	CS (unique sound), the UCS (food pellet)	UCR (salivation)
3	CS (unique sound)	CR (salivation)

Operant Conditioning

The second behavioral format of animal study was Skinner's extension of Thorndike's law of effect: *operant* (or instrumental) *conditioning*. Although Skinner acknowledged the existence of cognitions, he did not consider them germane to understanding human behavior. In fact, Skinner believed that learning occurred in similar ways for both humans and infrahumans, that is, through operant conditioning (Skinner, 1938). In its simplest form, operant conditioning has been expressed as an *S–R* model where the *S* represents the operant or behavior that is emitted by the organism; the *R* represents the response. The operant response has the following characteristics: it must be spontaneous, meaning that it is already in the organism's response repertoire; it must be autonomous; it must be measurable; and, it must occur in a consistent predictable manner under the influence of a particular controlling variable.

In contrast to Type S conditioning, operant or Type R conditioning emphasizes the response aspect. Skinner explained Type R conditioning and extinction as follows:

> If the occurrence of an operant is followed by presentation of a reinforcing stimulus, the strength is increased . . . . If the occurrence of an operant already strengthened through conditioning is not followed by reinforcing stimulus, the strength is decreased.
>
> Skinner, 1938: p.21

There is also an optional antecedent stimulus event frequently called the discriminative stimulus (S^D) that precedes the operant response and sets the occasion for responding. The discriminative stimulus is a specific physical stimulus in whose presence the operant behavior is reinforced (R^R) (**Figure 1**). Learning occurs when these behaviors become strengthened by a reinforcing stimulus. Simple behaviors can be chained to form more complex behaviors. The reinforcing stimulus has

$$S^D \rightarrow R \rightarrow R^R$$

Figure 1 Basic operant model of conditioning.

several characteristics that have differential effects on operant responding, including amount, quantity, magnitude, or size of stimulus; immediacy of delivery; and schedule of reinforcement administered, namely fixed, variable, ratio, or interval. For example, if reinforcement was provided for some desired behavior when a light was turned on but was not given when the light was off, the subject would eventually learn to emit the desired behavior only when the light was on. The light, when turned on, would have become a discriminative stimulus indicating that reinforcement was forthcoming if the desired behavior was emitted in its presence.

Operant responding can be strengthened or increased via positive and negative reinforcement. In positive reinforcement, a valued stimulus is presented after a response that, in turn, increases the probability that the behavior will reoccur. In negative reinforcement, an adverse stimulus is removed after a response that also increases the probability that the behavior will reoccur. Operant responding is weakened through two types of punishment (positive and negative). In positive punishment, a negatively valued stimulus is presented after a response that, in turn, decreases the probability that the behavior will return. In negative punishment (omission), a valued stimulus is removed after a response, which, like positive punishment, will decrease the probability of reoccurrence of behavior.

Edwin Guthrie

In contrast to the concept of reinforcement, especially as applied in the form of rewards, as the means to improve the probability of learning taking place, is the single occurrence law of Edwin Guthrie. Simply, Guthrie maintained that a stimulus that is followed by a particular response would, upon its recurrence tend to be followed by the same response again. This S–R connection gains its full strength on one trial. As such, Guthrie totally rejected the role of reward in learning. Guthrie, a contemporary of Skinner, did not popularize his theory in the practice of education because of the intuitive nature of humans to accept the concept of rewards as a *why* for learning.

Modern Trends in Behavioral Theory

In the 1980s, behaviorism continued to survive more in application than it did in theory development. Although cognitive psychology had become the dominant theoretical force in psychology in the 1980s, behavioral approaches could be found in settings such as clinical psychology, special education, industry, and the military. However, even in these bastions of behaviorism there was some evidence that, at least theoretically, behaviorists were beginning to acknowledge the potential importance of mediational influences on behavior.

Contemporary behavioral theory maintains a strong emphasis on behavior. These include the concepts of *active* respondents (e.g., talking, writing, experimenting and demonstrating) and *assessment* of learning (e.g., actual observation of students' behaviors changing as a result of instruction (e.g., higher test scores, improved performance, more appropriate social interaction skills, or better study habits)). Drill and practice, repetition of stimulus–response interactions, continues in education – especially in adult literacy, where the instructional strategy is read, read, read! Likewise, there are updates in Guthrie's theory for breaking habits (make a new response to the same old stimulus). Three useful methods are as follows:

- Exhaustion method. Continue an S–R activity until the individual is too tired to continue.
- Threshold method. Start a stimulus task below the intensity level. Gradually increase intensity so that the individual still does not respond. For example, in test anxiety present less demanding items that are actually enjoyable. In other words, move the instruction from easy to hard.
- Incompatible stimulus method. Present stimulus when the response cannot occur.

For example, in cooperative learning, substitute cooperation for competition. Rewards (reinforcement) for desirable behavior, timing, magnitude, and consistency of reinforcement, are still viewed to affect the rate at which new behaviors are learned.

However, not all behaviorists became more liberal in the 1990s. In his article, *The Shame of American Education*, Skinner (1984) lamented that the technology for effective teaching was currently available but was not being used. He viewed the trend toward cognitive psychology as being in the wrong direction. Indeed, he stated that cognitive psychology was not a revolt against behaviorism but rather a retreat. Skinner criticized cognitive psychology in that the promises to significantly improve education made a quarter of a century ago had not come to pass. He also contended that in the area of teaching language, linguists and supporters of cognitive science had not produced anything as effective as programmed instruction. Cognitive psychology had gained credibility due to its connections with brain research. However, Skinner noted that cognitive psychologists were a long way from developing useful instructional practices based on such research.

Cognitive Psychology

Behavioral psychology in its various forms holds to two primary principles of learning. The first principle is the roles of environmental conditions (stimuli) in establishing a framework from which learning is initiated. The second is the occurrence of observable behaviors (responses) resulting from the environment. Developing concurrently with behavioral theories were two alternatives to the why and how of learning: social learning theory and cognitivism. The concern of cognitive theories with knowledge structures, metacognitive strategies for problem solving, and integration of new and existing knowledge structures by the learner is leading to a number of educational paradigm changes. These changes include: development of new theoretical bases for design of learning environments intended to facilitate exploration and reinforce context; expansion of the role of learner control in instructional management; enriched dialog between the learner and the learning environment to support individual learner differences

to facilitate social integration and enhance motivation; and a redefinition of the structural requirements for simulations and games.

Background: Social Learning Theory

Social learning theory found earlier proponents for application in the classroom than did cognitive theory because of the focus in the social theories on ways in which students learn from interacting with one another. The main concept in social learning theory is that learning occurrences by individuals observing others, especially one's peers. Additionally, social theory blends behavioral concepts of reinforcement and punishment with cognitive concepts of awareness and expectations. A reciprocal causation between environmental conditions and cognitive variables explain the how of learning from this perspective.

The work of Albert Bandura (1969, 1986) spans much of the rich literature in social learning theory. Bandura's model of social learning theory revolves around the individual having the opportunity to model a best example of a desired behavior. Four variables within the modeling paradigm are necessary for learning to occur. These four are attention, retention, motor reproduction, and motivation. Attention actives the individual's focus on the presentation of the behavior to be learned. This variable establishes the role of the environment as the primary source of information to be learned. Retention, on the other hand, evokes a cognitive condition in which the individual recalls the desired behavior. For learning to have a lengthy duration, retention must occur. The reciprocal causation to retention is the behavioral variable of performing an observable behavior (or motor reproduction). This third variable requires the individual to assume the necessary role conditions of the desired behavior. Simply imaging a behavior in a mental model is insufficient for learning. Finally, learning a behavior is largely dependent on the effort or motivation that the individual puts into the task. Motivation in this model is the responsibility of the individual.

The reciprocal relationship between cognitive variables and environmental conditions is the within the affective domain. For example, self-efficacy, defined as a situational attitude, reflects on the person's belief toward possible task success. High self-efficacy would imply that a person feels that they can achieve success in a given situation. On the other hand, low self-efficacy would necessarily inhibit a goal of success. Another affective domain variable that strongly shows the cognitive role in learning is the concept of self-regulation. This variable is central to the cognitive theories of constructivism. For self-regulation to develop, the individual needs to establish personal criteria by which to make decisions and choices. Other affective/cognitive variables in social learning theory are self-observation and judging, self-monitoring, and self-imposed stimulus control.

Background: Cognitive Theory

In an outline of cognitive psychology history, Tennyson and Elmore (1997) write that cognitive psychology grew out of a history of behaviorism (e.g., Skinner) and psychoanalysis (e.g., Freud). These two divergent approaches focused on the individual; additionally, the measurement tools developed in cognitive psychology have also been focused on the learner. To the extent that educators understand the learner, they can develop more effective instruction, but cognitive psychology has not been primarily concerned with that. Effective instruction is an implementation of research. That is, experimental psychologists conducting basic research in human learning have not traditionally been responsible for diffusion and dissemination of their findings into the various fields of education.

Cognitive psychology separated from the behavioral tradition because of the dissatisfaction in the 1960s with the 'thoughtless' approach to human learning offered by behaviorism (Neisser, 1967). The focus of cognitive psychology theory development and research was on the processes of how people perceive, interpret, remember, and otherwise think about the environmental events they experience. These newer developments actually extended work begun several decades earlier. For example, Edward Tolman (1932) viewed learning as occurring within a context (molar level) and without reinforcement. And, even earlier, Gestalt psychologist wrote about the importance of organizational processes in perception (the whole is more than the sum of its parts), in learning (law of Prägnanz – simple explanations), and problem solving (restructuring and insight).

Jean Piaget's developmental theory (1928) became increasingly popular in schooling environments during the 1970s and still forms much of the foundational theory for cognitive theory, especially in the area constructivism. Briefly, Piaget held that learners are active processors of information and the resulting knowledge in memory can be described in terms of structures that change with development (scheme). Furthermore, cognitive development occurs with interaction between physical and social environments.

In Piagetian terms, there is two cognitive processes people use to interact: assimilation and accommodation. Assimilation accounts for the acquisition of new knowledge while accommodation is mental process of accepting new knowledge with the old. This latter process attempts to establish an equilibration between new (equilibrium) and accepting of new with the old (disequilibrium). All of this mental functioning occurs over four stages of human development. An important component of Piaget's theory is the philosophy that people are intrinsically motivated.

One final early learning theorist whose effect on contemporary cognitive theory occurred decades after his death is the Russian psychologist, Lev Vygotsky's. Most of Vygotsky's research in developmental cognition was done in the 1920s, with his interest in the concept that mental processes begin as social activities and later, as they become internalized, can be transferred and generalized to other situations and activities. Of much interest to adult educators is Vygotsky's theory that learning is more efficient if done while interacting with more competent persons. This can be accomplished if the teacher, for example, can assess both the learner's actual developmental level and his/her level of potential development. The gap between these two levels is labeled the zone of proximal development. A recent

translation (Vygotsky, 1997) offers a good look at this important theory of learning.

Information Processing Models

A primary application of computer techniques in describing learning was the development of information processing (IP) models (Simon and Newell, 1964). The basic components of the IP model of learning (**Figure 2**) include the following components: sensory receptors, perception, short-term (working) memory, and long-term memory. Contemporary IP models, unlike the behavioral models, have two primary sources of knowledge acquisition: external and internal. External information enters the cognitive system through the standard sensory mechanisms while internal information is constructed as a result of the output–input relationship between the three system components. External behavior is exhibited through the output of the sensory receptor component.

Sensory Receptors Component

External information enters and behavior exits the cognitive system through the sensory receptors. Basically, these sensory receptors are the learner's ears and mouth (auditory) and eyes (visual). The information in this register decays rapidly and is easily interrupted. External stimuli include those aspects of instructional development referred to as delivery systems; such as, text materials, visuals, audio sources, graphics, illustrations, drawings, etc.

Perception Component

Information coming from the external receptors or from internal knowledge sources passes through the perception component which performs the function of being aware of and evaluating the potential value and worth of that information and knowledge, respectively. The perception component can be viewed as a filtering device for the cognitive system. Overtime, the perception component performs this filtering task in an autonomous or subconscious matter; all within a short time period, 0.5–2.0 s. For example, when a person enters a new course for the first time, they may view the course as too easy, thus putting in a minimum amount of effort or they many view the course as too difficult, again, putting in minimum effort. The learner is evaluating the situation in the perception component prior to actual cognitive activity. Perception is often a precognition mental operation.

Short-Term (Working) Memory Component

There is considerable debate as to the divisions or architecture of memory, but in a broad sense, the exact details are not important to the instructional implications of the various theories. In general, there is agreement that memory comes in two forms, a store for previously learned information and a store for information that is currently being processed. This latter form, short-term (working) memory, is defined with these salient aspects: (1) it is limited in storage capacity and time (approximately 20 s); (2) items in working memory are subject to manipulations such as rehearsal, comparison, or matching and reordering by the processes that operate in short-term memory; and (3) items are selected for inclusion in short-term memory either by some consciously active process or by automatic action of well developed processes in such activities as reading, processing of verbal discourse, imagery evoking processes, etc.

Long-Term Memory Component

The acquisition of knowledge and the means to employ and elaborate knowledge occur within the storage and retrieval subsystems of the long-term memory component. Within the storage subsystem information is encoded into the knowledge base according to various representations (i.e., declarative and procedural knowledge), while the retrieval subsystem uses cognitive abilities to employ knowledge (i.e., differentiation and integration).

The storage subsystem is where coded knowledge is assimilated into a learner's existing knowledge base. A knowledge base can be described as an associative network of concepts (or schemas) that varies per individual according to amount, organization, and accessibility of its knowledge. Amount refers to the actual volume of knowledge coded in memory, while organization implies the structural connections and associations of that knowledge, and accessibility referring to the processes used in servicing the knowledge base. The latter two forms of knowledge are those that separate an expert from a novice. That is, a large amount of knowledge is not the key to expert performance, but rather the ability to both find and employ knowledge appropriately.

Within long-term memory there are two types of knowledge: declarative and procedural. Declarative knowledge implies an awareness of knowledge and refers to the *knowing that*, for example, the student understands that underlining keywords in a text will help recall. Procedural knowledge implies a *knowing how* to use given facts, concepts, rules, and principles. For example, the student knows how to perform an analysis of variance in a statistical test.

The values of declarative and procedural knowledge are their employment in the service of recall, recognition, and problem solving. Cognitive abilities in the storage subsystem can be distinguished as either domain-dependent (cognitive

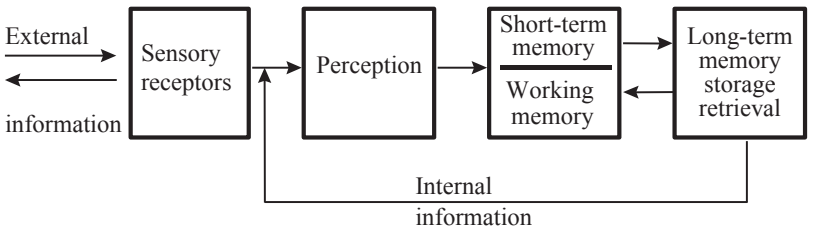

Figure 2 Illustration of a generic information processing model.

skill) or domain-independent (cognitive strategy). In the cognitive literature, skill is often referred to as a specific application of a strategy within a given situation whereas strategy has come to imply a generic ability that can be applied across domains. Both skills and strategies include two primary cognitive operations: differentiation and integration (Schröder, 1971). Differentiation is defined as the twofold mental activity to understand a given situation and to apply appropriate contextual criteria (i.e., the standards, situational appropriateness, and/or values) by which to selectively retrieve specific knowledge from storage. Integration is the ability to elaborate or restructure existing knowledge in the service of previously unencountered problem situations.

A number of learning theories dealing with a cognitive approach to learning have been developed over the last several decades. We will review five of these theories in this article: schema theory, advance organizers, assimilation theory, discovery learning, and creativity.

Schema Theory

Cognitive theorists, departing from the past focus on behaviorism, have emphasized the importance of schemata in learning new information. A schema can be defined as representing a structural framework into which new knowledge is instantiated. While there is not one *schema theory*, there is a wide range of cognitive theories which include the concept of schemata and which emphasize cognitive structure.

Early Schema Theory

Bartlett (1932) was one of the first psychologists to use the term *schema* in relationship to learning theory. His general theory was that remembering is schematically determined. He postulated that, "Circumstances that arouse memory orientations set up an attitude that is primarily toward a particular schematic organization" (p. 24). He theorized that the construction or reconstruction of recalled knowledge is always within this organization. Bartlett's definition of schemata involved a general perception of life; thus, a schema governs a person's perception of new or recalled information. Bartlett contended that schemata influence memory by interacting with incoming stimulus content in an effort to make the content meaningful. The schemata strongly influence the perception of the information and this, in turn, determines what is retained.

Cognitive View

Ausubel (1980), representing a more contemporary cognitive view and in contrast with Bartlett's definition of schema which is more attitudinal and affective, maintains that this interpretive process is not perceptual but cognitive. Additionally, in Ausubel's view, the emphasis should not be on remembering, for many of the effects of schemata are present in the initial assimilation of the information. Accordingly, the existing cognitive structure at the time of learning is, perhaps, the most important factor influencing the learning process. If the cognitive structure is clear and well organized, clear concepts can be formed and meaningful learning can take place. If, however, a given cognitive structure is disorganized and ambiguous, retention is limited and meaningful learning is unlikely. The ease and speed of learning is, therefore, dependent upon the ease with which new information can be placed into an existing schema. Thus, the single most important aspect in facilitating learning is strengthening the relevant cognitive structure. Ausubel further describes schemata as providing *ideational scaffolding*, containing slots that can be instantiated with particular cases. Cognitive information, such as declarative and procedural knowledge, must be used to instantiate these slots.

Contemporary View

Current learning theories view schemata as cognitive structures that serve to organize information into meaningful concepts. Developed cognitive structures aid in learning because new information can be incorporated more readily and the association with established concepts extends the structure to the new information. The number of schemata present is also significant, because overlapping structures (structures containing common information *slots*) provide additional opportunities to use information and can aid in recall by strengthening semantic networks. For example, without a schema, learning is slow and uncertain. While this is rare, a more common situation is the contrast between a well-constructed schema and a poorly constructed one. In a disorganized schema, the slots in the framework can remain empty, or information can be distorted before being instantiated, if a fit is otherwise difficult. In other words, schemata provide a structural framework, like the steel frame for a building provides a framework, for the walls, floors, and windows that follow. This framework must form a *cognitive bridge* between what the student already knows and the new information. The absence of such a conceptual framework may contribute to the rapid loss of information shortly after a test is taken and may also account for frequent trial-and-error approaches.

Whenever someone has learned something, a schema has been produced. The knowledge structures acquired even for simple tasks can be quite involved. Additionally, it is important to characterize the schematic knowledge available to students. Since events recalled are influenced by the currently activated schemata, it is necessary to have some idea of the current schemata and to modify it, if necessary, to aid in the retention of appropriate material. For example, in a story about a restaurant, it is necessary to have some knowledge about restaurants in general, to be able to infer the information necessary to make sense out of the story. This general knowledge is a schema, and without this context the material is not likely to take on any meaning, even after numerous repetitions.

Learning often involves making inferences from incomplete information. When presented with new information, people make a constructive hypothesis, which is often remembered better than the information, and includes many inferences, based on what they hear. Remembering is often based on a reconstructive hypothesis which involves reconstructing the details from a general idea of what was presented, somewhat like problem solving; that is, determining what is likely to have happened based on the information or general idea remembered. If the information is consistent with the context and expectations recall will be accurate; if however, it is not, either recall will be low or reconstruction

will be erroneous, based on a context not consistent with the actual details. This leads to the assumption that a suitable context (schema) is necessary for understanding and reconstructing new information and those students with more contextual information will be better at assimilating new information.

Advance Organizers

An important concept in cognitive learning theories is the use of advance organizers. Closely associated with schema theory, advance organizers help to establish a relevant cognitive structure to aid in meaningful learning. Both quantitative and qualitative effects of repetition in learning accomplish this. Repetition influences the quantity learned because with each repetition, more information is added to memory. Repetition increases the quality of learning because the student is able to build a conceptual framework, using an *assimilative encoding strategy* to build conceptual relationships and reorganize information into a whole. In this sense, advance organizers perform a function similar to repetition, improving problem solving and overall learning (qualitative) but not the learning of verbatim information (quantitative). This is consistent with general schema theory where advance organizers aid in the formation of a conceptual framework, reducing the need for repetition.

Assimilation Theory

In presenting cognitive theory, recall that much of the behaviorist research, leading to a stimulus–response (S–R) theory of learning, was done on animals, who do not form concepts, and in human rote learning. Since concept learning is an important part of human learning, behaviorally based research is not sufficient in describing human learning.

In cognitive learning theory, there is a continuum from rote learning to meaningful learning. In rote learning, new information is placed in cognitive structure in an arbitrary manner. In meaningful learning, new information is linked to relevant, existing concepts in cognitive structures (or mental models). The structural core is the interaction of this new knowledge with the preexisting knowledge, often modifying both the new and preexisting information. With meaningful learning, old and new information is assimilated and formed into a more structured cognitive organization.

Assimilation encoding theory is a three-stage model that involves actively integrating new information with existing knowledge and measuring results in terms of breadth rather than amount retained. The three stages are: receipt of information from the outside world; availability of relevant anchoring knowledge in long-term memory; and, transfer of anchoring knowledge from long-term memory to working memory and active integration of that knowledge with incoming information during learning.

This theory predicts that advance organizers, given before learning, can facilitate the availability of anchoring knowledge and the integration of this existing knowledge with incoming information. There are several reasons, according to Assimilation Encoding Theory, why advance organizers may succeed or fail to influence learning. If the material to be learned lacks an overall structure, then an effective advance organizer cannot be constructed. If the advance organizer does not provide an adequate assimilative context, then integration of new and existing information may not occur. If the learner already possesses relevant experience and knowledge or a well-developed cognitive structure (schema), the use of advance organizers may have no benefit.

Discovery Learning

In the early 1970s Egan and Greeno (1973) tested the effectiveness of discovery learning with learning by rule. The group using the discovery method showed a wider discrepancy in performance than students using the rule method. This was because several cognitive skills necessary for learning by discovery were found to be less essential to learning by rule. The findings also indicated that discovery leads to the structural integration of previously known concepts, while the outcome of learning by rule is the addition of new structures. High ability students performed about as well using either method, while low ability students learned more effectively by rule, because they lacked the cognitive structures (schemata) necessary to form new structures.

Creativity

In the educational community there has been a strong focus on the need to improve creativity. Making people better thinkers and problem solvers has been emphasized as an educational goal since John Dewey's (1932) writings on progressive education. These themes continue to appear as national concerns due to the dissatisfaction of educators with current curricular goals and instructional methods used in the schools and in training.

Creativity is defined as having the three attributes of novelty and value: The first is personal (i.e., the individual recognizes the uniqueness of the created knowledge); the second is peer group (i.e., recognition by experts within a given domain); and, the third is societal (i.e., recognition by the general members of the given community).

The work of Torrance and his associates (e.g., Torrance and Ball, 1984) has developed instruments to measure creativity through the number of ideas developed for specific situations and tasks. This implies that creativity is the internal creation of knowledge not previously part of the individual's knowledge base. Another feature of measuring creativity is the inclusion of the individual's ability to establish a criterion for ideas created. That is, not only is the individual creating ideas, but they are also defining the means by which their ideas can be judged. For example, in preparing for an experiment in a physics class, the student needs to define the possible value that results might provide not just the fact that they put some elements together.

Tennyson and Breuer (1997) explain creativity using cognitive complexity theory of differentiation and integration. To these authors, creativity is the concurrent employment of cognitive complexity with the process of internal acquisition of knowledge, thus a twofold cognitive ability: (1) creating knowledge in order to solve a given problem

proposed from the external environment and (2) creating a problem and then the necessary knowledge to propose a solution(s). Integral to the creation of the problem and knowledge is the creation of the criteria that can then be used to determine the possible effectiveness of the solution proposal. Criterion is defined as the perceived value or worth employed by the student in the creating process. That is, one must also consider creating the criteria concurrently with creating the problem and/or knowledge that is consistently employed with a high level of reliability and productivity.

Summary

In conclusion, we would like to summarize the main components and ideas of cognitive theory. This summary can be compared with the summary on behaviorism to contrast similarities and differences between these two main bodies of learning theories.

- Some learning processes may be unique to human beings.
- Cognitive processes are the focus of research and study.
- Objective, systematic observations of people's behavior should be the focus of scientific inquiry; however, inferences about unobservable cognitive processes can and should be drawn from such behavior.
- Individuals are actively involved in the learning process. Learners control their own learning.
- Learning involves the formation of cognitive associations that are not necessarily reflected in overt behavior changes.
- Knowledge is organized (knowledge, beliefs, attitudes, and emotions are associated and interconnected).
- Learning is a process of relating new information to previously learned information.

The distinction between learning philosophy and learning theory is often confusing if viewed from an abstract position. However, by understanding a learning philosophy, the task of employing learning theory becomes much clearer. Therefore, we suggest that the reader prepare a working statement on their personal learning philosophy prior to selecting and employing learning theory. In this process, the learning philosophy statement can be reviewed and refined if necessary to balance the validity between the two areas.

Bibliography

Ausubel, D.P., 1969. A cognitive theory of school learning. Psychology in the Schools 6, 331–335.

Ausubel, D.P., 1980. Schemata, cognitive structure, and advance organizers: a reply to Anderson, Spiro, and Anderson. American Educational Research Journal 17, 400–404.

Bandura, A., 1969. Principles of Behavior Modification. Holt, Rinehart & Winston, New York.

Bandura, A., 1986. Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice Hall, Englewood Cliffs, NJ.

Bartlett, F.C., 1932. Remembering. Cambridge University Press, London.

Bruner, J.S., 1964. Study of Thinking. Wiley, New York.

Carroll, J.B., 1993. Human Cognitive Abilities. Cambridge University Press, New York.

Dewey, J., 1910. How We Think. D.C. Heath, Boston.

Dewey, J., 1932. Human Nature and Conduct. Holt, New York.

Egan, D.E., Greeno, J.G., 1973. Acquiring cognitive structure by discovery and rule learning. Journal of Educational Psychology 64, 85–97.

Gagné, R.M., 1985. The Conditions of Learning, fourth ed. Holt, Rinehart, & Winston, New York.

Herrenstein, R.J., 1977. The evolution of behaviorism. American Psychologist 32, 593–603.

Mager, R., 1962. Instructional Behavioral Objectives. Fearon Press, San Francisco.

Maslow, A., 1962. Toward a Psychology of Being. Van Nostrand, Princeton, NJ.

McKeachie, W.J., 1974a. The decline and fall of the laws of learning. Educational Researcher 3, 7–11.

Molenda, M., 1997. Historical and philosophical foundations of instructional design: a North American view. In: Tennyson, R.D., Schott, F., Seel, N., Dijkstra, S. (Eds.), Instructional Design: International Perspectives. Theory and Research, vol. I. Erlbaum, Mahwah, NJ, pp. 41–54.

Neisser, U., 1967. Cognitive Psychology. Appleton-Century-Crofts, New York.

Pavlov, I.P., 1927. Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex (G.V. Anrep, Trans.). Oxford University Press, London.

Piaget, J., 1928. Judgment and Reasoning in the Child (M. Warden, Trans.). Harcourt, Brace, New York.

Piaget, J., 1952. The Origins of Intelligence in Children. International Universities Press, New York.

Rogers, C., 1961. On Becoming a Person. Houghton-Mifflin, Boston.

Simon, H.A., Newell, A., 1964. Information processing in computer and man. American Scientist 52, 281–300.

Schröder, H.M., 1971. Conceptual complexity and personality organization. In: Schroder, H.M., Suedfeld, P. (Eds.), Personality Theory and Information Processing. Ronald Press, New York, pp. 128–146.

Skinner, B.F., 1938. The Behavior of Organisms. Appleton-Century-Crofts, New York.

Skinner, B.F., 1954. The science of learning and the art of teaching. Harvard Educational Review 24, 86–97.

Skinner, B.F., 1957. Verbal Behavior. Appleton-Century-Crofts, New York.

Skinner, B.F., 1958. Teaching machines. Science 128, 969–977.

Skinner, B.F., 1968. The Technology of Teaching. Meredith, New York.

Skinner, B.F., 1984. The shame of American education. American Psychologist 39, 947–954.

Tennyson, R.D., Breuer, K., 1997. Psychological foundations for instructional design theory. In: Tennyson, R.D., Schott, F., Seels, N., Dijkstra, S. (Eds.), Instructional Design: International Perspectives. Theory and Research, vol. I. Erlbaum, Mahwah, NJ, pp. 113–133.

Tennyson, R.D., Elmore, R.L., 1997. Learning theory foundations for instructional design. In: Tennyson, R.D., Schott, F., Seel, N., Dijkstra, S. (Eds.), Instructional Design: International Perspectives. Theory and Research, vol. I. Erlbaum, Mahwah, NJ, pp. 55–78.

Tennyson, R.D., Elmore, R.L., Snyder, L., 1992. Advancements in instructional design theory: contributions from cognitive science and educational technology. Educational Technology Research and Development 40, 9–22.

Tennyson, R.D., Thurlow, R., Breuer, K., 1987. Problem-oriented simulations to develop and improve higher-order thinking strategies. Computers in Human Behavior 3, 151–165.

Thorndike, E., 1913. The Psychology of Learning: Educational Psychology, vol. 2. Teachers College Press, New York.

Tolman, E.C., 1932. Purposive Behavior in Animals and Men. Century, New York.

Torrance, E.P., Ball, O.E., 1984. Torrance test of creative thinking. Scholastic Testing Service, New York.

Vygotsky, L.S., 1997. Educational Psychology. St. Lucie Press, Boca Raton, FL.

Learning to Learn

Claire E Weinstein, Jaimie M Krause, and Nancy Stano, The University of Texas at Austin, Austin, TX, USA
Taylor W Acee, Texas State University – San Marcos, San Marcos, TX, USA
M Krause Jaimie and Nancy Stano, The University of Texas at Austin, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C.E. Weinstein, volume 13, pp. 8620–8623, © 2001, Elsevier Ltd.

Abstract

Research has shown that learners can intentionally improve how effectively and efficiently they construct knowledge and develop skills. Teaching students how to learn is critical to helping them learn and use content knowledge in any subject area. It can also help students develop more effective and efficient learning routines and adapt and transfer their learning processes to new contexts. In this article, we overview the progression of research on learning to learn, summarize the Model of Strategic Learning, and describe learning to learn assessments and interventions that can be used to help students become more strategic and self-regulated learners.

Learning to Learn

Learning to learn involves a complex and dynamic interaction of cognitive, metacognitive, motivational, affective, and behavioral processes selected to enhance the probability of reaching a goal for knowledge acquisition or application. The quest for understanding and harnessing an individual's ability to identify, reflect upon, improve, and control his or her learning processes and outcomes began in ancient times with the study of mnemonics (memory aids) and Socratic teaching methods. However, it is only since approximately 1975 that incredible progress in understanding learning to learn and the methods that can be used to improve it has occurred (Weinstein and Acee, 2013; Weinstein et al., 2011). This article will present a definition of learning to learn followed by a brief history of the development of the study of learning to learn processes. A model summarizing much of the current research in this area will then be presented as well as a summary of the results of current research examining the nature and effectiveness of methods for improving learning abilities and skills. The article concludes with a discussion of suggestions for future research and conceptual development in the area of learning to learn.

Learning to Learn Strategies

Learning to learn strategies include any thoughts, behaviors, beliefs, or emotions that facilitate the acquisition, understanding, or later application and transfer of new knowledge and skills in different performance contexts. They range from active rehearsal to help remember word lists, to the use of elaboration and organization to encode, integrate, and later recall or apply knowledge across several content areas (Weinstein and Mayer, 1986). Learning to learn strategies help generate meaning for the new information that is to be learned. Repeating the names of the planets in order, organizing the discoveries of the great explorers by creating timelines and mind maps, comparing and contrasting the causes of World War I and World War II are all examples of learning strategies. They all are designed to help the learner generate meaning and store the new information in memory in a manner that will facilitate integration with related knowledge and increase the probability of later recall and use, particularly in transfer contexts.

A learning strategy is also a plan for orchestrating cognitive and motivational resources to help reach a learning goal (Weinstein and Acee, 2013; Weinstein et al., 2011). Learning to learn strategies have several characteristics in common. First, they are goal directed. Learning strategies are used to help meet a standard of performance or to reach a learning goal. Second, learning strategies are intentionally invoked, which implies at least some level of active selection. The selection of one or more of these strategies is determined by a number of factors, such as a student's prior experience with the strategy, his or her prior experience with similar learning tasks, his or her ability to deal with distractions, and the student's commitment to his or her goals. Third, cognitive learning strategies are effortful; they require time and often involve using multiple, highly interactive steps. Because of the effort required, a student must be motivated to initiate and maintain strategy use (e.g., *see* Motivation, Learning, and Instruction). In addition, the student must believe that the strategy will be effective and that he or she can be successful using the strategy. Finally, cognitive learning strategies are not universally applicable – they are situation specific to a greater or lesser degree (Weinstein and Acee, 2013; Weinstein et al., 2011). In general, learning strategies have differing domains of applicability and usefulness. For example, time management strategies can be applied in a wide variety of contexts ranging from reading a chapter in a physics textbook to planning a study schedule for the week before an exam. Other strategies are more limited in their generalizability such as using an information processing strategy for solving algebraic equations with one unknown variable. In addition, the student's goals, the task requirements, the context, and other factors all interact to help determine which strategy may be best. To be successful in selecting and using a strategy, a student must understand under what circumstances a given strategy is or is not appropriate. The following section will discuss the evolution of learning to learn as a field of study and a source of

applications for many students experiencing difficulty in an educational setting.

Historical Overview

In ancient times the great Greek orators were known for giving long speeches that sometimes lasted more than 2 or 3 h. They managed this task through the use of mnemonic or memory devices that helped them to remember both the ideas and the order of the ideas for their presentations. The most famous of these techniques is the Method of Loci. Basically, using the Method of Loci involves remembering a series of locations in order. This could be the buildings and landmarks the orator passed on the way to the arena or the rooms in a large villa. As the ideas were generated for the speech, the orator would 'store' an image of the idea at a specific location. When the time came to give the speech, the ideas would be recalled as the speaker 'went' from location to location.

For years the study of mnemonics lay dormant until interest renewed in the second half of the twentieth century. Psychology has gone through major upheavals in the twentieth century. The rejection of philosophy and introspection as primary tools for the study of psychology led to the rise of behaviorism and the rejection of 'mental' phenomena. In the rush to be scientific in developing theories and methods, researchers lost their focus on the mind and cognitive processes. It was not until the late 1960s and the 1970s that studies of perception, memory, and cognitive processes once again gained prominence (*see* Educational Learning Theory; Learning Theories and Educational Paradigms; Zimmerman and Schunk, 2003). Some of the earliest studies of memory reexamined the classical mnemonic systems in an attempt to identify and understand the cognitive processes involved in creating and using them. In the 1970s this literature was enriched by the work of Flavell and others on what he called 'metacognition' (Flavell, 1979). Metacognition refers to thinking about thinking. It focuses on knowledge, reflection, and analyses of how an individual thinks and learns (*see* Metacognitive Development: Educational Implications; Hacker et al., 2009). Building on both of these themes, researchers increasingly turned their attention to cognition, the variables that impact cognition, and the degree to which cognition can be influenced through educational interventions. This is the research that led most directly to the study of learning to learn.

Types of Learning to Learn Strategies

Claire Ellen Weinstein and Richard E. Mayer (1986) summarized much of the research and findings in 'learning to learn' and developed a taxonomy to describe the six major categories of learning strategies. The following is a summary of the categories in the taxonomy.

Rehearsal Strategies for Basic Learning Tasks

Many instructional tasks require simple recall. The acquisition of basic knowledge, however, is often just a first step in the creation of a more extensive and integrated knowledge base in an area. An example of a task in this category is memorizing the order of the colors in the spectrum. A useful strategy for this type of learning might be to repeat, in sequential order, the names of the colors.

Rehearsal Strategies for Complex Learning Tasks

Learning tasks in this category are more demanding and tend to involve knowledge and skills extending beyond the superficial learning of lists or unrelated bits of information. Examples of strategies in this category would include highlighting class notes or copying key ideas from the class readings. Like the rehearsal strategies for simple tasks, these methods are often most appropriate in the early stages of building a knowledge base in an area.

Elaboration Strategies for Basic Learning Tasks

Meaningful learning often involves building bridges between what is already known and what a student is trying to learn. Elaboration strategies help the student build these bridges by using previous knowledge or experiences to add to what he or she is trying to learn as a way to make the new information more meaningful and memorable. Examples of strategies in this area include generating a mental image of a scene described in a novel and relating a scientific principle to everyday experience. The facilitating effect of using these strategies is usually attributed to both the processing that is involved in creating the elaboration and the product itself.

Elaboration Strategies for Complex Learning Tasks

Strategies in this category help the learner to be more active in building bridges between prior knowledge or experience and more complex tasks. Examples of strategies in this category include paraphrasing and summarizing, applying a problem-solving strategy to a new problem, creating analogies, and using everyday experience to help understand a new concept.

Organization Strategies for Basic Learning Tasks

Organization strategies involve translating or transforming information into another form and creating some sort of scheme to provide structure for this new way of characterizing the information. Examples of organization strategies for basic tasks include grouping foreign language vocabulary words by their parts of speech and grouping the battles of World War II by their geographic location. Like elaboration strategies, the facilitating effect of organization strategies is usually attributed to both the processing that is involved in creating the organization and the product itself.

Organization Strategies for Complex Learning Tasks

Complex tasks can often be made more meaningful and more manageable by using organization strategies. Examples of strategies in this category include creating a hierarchy of main ideas to use in writing a term paper, generating a force diagram in physics, or creating an outline for the class notes from

a section of a course. Like all of the strategies in the elaboration and organization categories, it is both the process and the outcome that appear to help in both knowledge acquisition and later recall and performance.

Model of Strategic Learning

As research in 'learning to learn' expanded in the 1980s and 1990s, it became clear that a number of different factors contribute to and influence both choice and type of strategy used by a student (*see* Self-Regulated Learning; Weinstein and Acee, 2013; Zimmerman, 2008). For example, it is not enough for a student to know about different types of learning strategies. He or she must also know how to use them and under what conditions it is appropriate to use them. Another factor that affects the strategies selected and used by a student is the social contexts of both the educational setting, in general, and the individual classroom. There are also support variables that affect a student's attitudes and beliefs about learning and his or her motivation to succeed. These variables help to generate a context in which effective learning can take place. For example, a student must be clear about his or her long-term and short-term academic goals in order to use them to aid his or her motivation to initiate and maintain use of effective learning strategies.

In the past researchers and educational program developers have usually focused on one or a subset of topics within learning to learn, such as cognitive elaboration strategies or student motivation. Current work is purposefully examining the interaction among two or more components related to the acquisition and use of learning strategies (see Pintrich, 2004; Schunk et al., 2008; Weinstein and Acee, 2013; Weinstein et al., 2011; Zimmerman and Schunk, 2011). This change is a result of the increasing understanding of the nature of student learning and school achievement at all educational levels. Researchers have found that it is the interaction among varying factors that results in successful learning and transfer of new knowledge and skills. The components that seem to have the greatest impact on a student's acquisition and use of learning strategies are summarized in a model developed by Weinstein and her colleagues at the Cognitive Learning Strategies Project located at the University of Texas at Austin in the USA (see **Figure 1**).

This model is an extension of the earlier model developed by Weinstein and Mayer (1986) and has been described in other publications, including Weinstein and Acee (2013), Weinstein et al. (2011), and Weinstein et al. (2000). It focuses on variables impacting strategic learning: learning that is goal driven. Briefly, the core of the model is the learner's long-term learning goals and self-system variables. Goals are considered to be critical components that provide direction for self-regulating thoughts and actions and for generating and sustaining the motivation necessary to carry out those thoughts and behaviors. A hallmark of a strategic learner is that he or she sets specific and challenging, yet realistic,

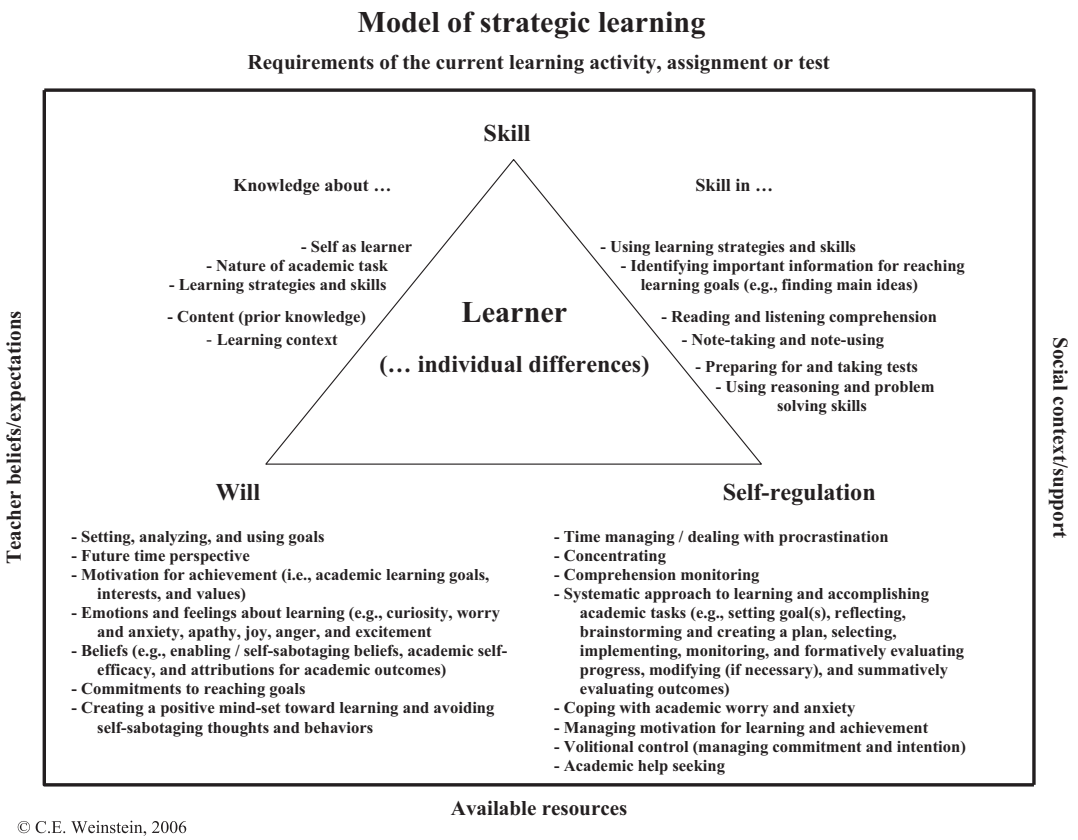

Figure 1 Model of Strategic Learning.

learning goals and actively pursues them. These learning goals are outcomes to be achieved in and of themselves, but are also subsumed within larger and much longer-term goals and thereby have a future utility value. However, setting specific learning goals is necessary but not sufficient to ensure successful learning outcomes. A strategic learner must also have the skills, motivation, and self-regulatory thoughts, beliefs, and behaviors necessary to successfully pursue such goals. The core of the model is surrounded by these three components: (1) the learner's 'skill' levels in relation to the learning task at hand; (2) the learner's 'will' or motivation to accomplish the desired outcome; and (3) the learner's 'self-regulating thoughts, beliefs, and actions' (sometimes referred to as executive control processes in earlier literature). It is the properties that emerge from the interactions among the learner and these components that are at the core of 'learning to learn.' For any given learning task, the student must take into account variables from each of these components.

Assessment of Learning to Learn

A preponderance of the literature in learning to learn focuses on creating conceptual and instructional foundations for possible interventions to help students become more strategic and self-regulated learners. However, in order to develop interventions it is necessary to be able to assess where students' existing strengths lie and the areas in which they need to improve. Diagnostic–prescriptive measures of students' current strengths and weaknesses in strategic and self-regulated learning provide (1) an initial snapshot of students' status as a strategic learner; (2) a tool to help students become more aware of their own strengths and areas needing improvement; (3) a basis for prescribing interventions and individualizing within more complex interventions (such as a semester-long course in learning to learn); (4) a pre/postintervention to both demonstrate student progress and identify areas needing further development; and (5) an evaluation tool to help determine the effectiveness of interventions for aggregate, subgroup, and individual students.

Many different methods and tools have been developed to help assess learning to learn. These include using talk-alouds, checklists, and questionnaires (Weinstein and Acee, 2013; Zimmerman, 2008). Talk-alouds can be concurrent or retrospective. Basically, students are asked to verbalize their thoughts as they are completing a learning task or studying for a class. Although talk-aloud assessment methods are useful for identifying different ways in which students approach learning tasks, they are difficult to use as diagnostic–prescriptive assessments of students' learning approaches in general because they usually focus on one particular activity. Additionally, the act of reporting thinking processes in concurrent talk-alouds interferes with the very processes being studied, and a retrospective self-report about specific tasks is subject to the reconstructive additions and modifications of memory. Checklists are another way to help identify students' strategic learning strengths and areas needing improvement. Using checklists often involves creating lists of different strategies for one or more types of learning tasks and asking students to check the strategies they use. One problem with checklist assessment methods is that students do not have the option of indicating the degree to which they used or relied upon particular strategies. Furthermore, students often check off more strategies than they actually use, which often results in higher correlations with social desirability than performance.

Perhaps a relatively more effective approach to creating diagnostic–prescriptive learning to learn assessments has been to use questionnaires that capitalize on some of the strengths of talk-alouds and checklists while reducing some of the pitfalls. However, it is important to note that since we are talking about cognitive, metacognitive, emotional, affective, and behavioral thoughts and actions, these are still self-report measures and are susceptible to students' level of self-knowledge and accurate reporting of that knowledge (e.g., responding in a socially desirable manner) as well as some of the problems discussed for talk-alouds and checklists. There are two leading measures of learning to learn processes and skills: the *Motivated Strategies for Learning Questionnaire* (MSLQ; Pintrich et al., 1991) and the *Learning and Study Strategies Inventory* (LASSI; Weinstein et al., 2002). Both questionnaire assessments address different areas of strategic learning but the MSLQ places greater emphasis on specific theoretical components of motivation and the LASSI is more broad spectrum including topics ranging from motivation to comprehension monitoring, generating and using study aids, and preparing for exams. The LASSI is a widely used instrument for learning to learn interventions and will be discussed in more detail as an example of a measure of strategic and self-regulated learning.

"The LASSI is a 10-scale, 80-item assessment of students' use of learning and study strategies related to skill, will, and self-regulation components of the Model of Strategic Learning" (Weinstein and Palmer, 2002). Students respond using a Likert-type scale ranging from 1 to 5, where 1 is not at all typical of me, 2 is not very typical of me, 3 is somewhat typical of me, 4 is fairly typical of me, and 5 is very much typical of me. The assessment items do not ask students about specific instances when they studied or tried to learn something but rather what is typical of them in different learning contexts, topics, and task types. The 10 scales on the LASSI are Anxiety, Attitude, Concentration, Information Processing, Motivation, Selecting Main Ideas, Self-Testing, Study Aids, Test Strategies, and Time Management. The factors measured by the LASSI have been found to contribute significantly to success in college and be amenable to improvement through educational interventions (see Weinstein and Acee, 2013). This is important for a diagnostic/prescriptive assessment. A number of variables that relate to learning success and that could be measured, such as certain types of personality factors, are not amenable to educational interventions in a high school or college setting. Although they can be quite diagnostic, they are not prescriptive in that they do not help identify areas of remediation or enhancement that can be targeted in an educational intervention. With regard to reliability, the Cronbach's alpha coefficients for all 10 scales range from a low of 0.73 to a high of 0.89 (Weinstein and Palmer, 2002). It is interesting to note that Weinstein and her colleagues developed a metacognition scale for the LASSI but it correlated at 0.7 and above with every other LASSI scale, suggesting that metacognition is foundational to many elements of strategic and self-regulated learning and

cannot be teased apart easily. However, the scale that focuses on self-testing does address a subset of metacognitive processes referred to as comprehension monitoring – checking our understanding during class or study sessions to see if we are really understanding the material.

The LASSI can be used as "… (a) a diagnostic measure to help identify areas in which students could benefit most from educational interventions; (b) a basis for planning individual prescriptions for both remediation and enrichment; (c) a means for instructors to use for examining individual students' scores and class trends to help make decisions about instruction, assignments, etc.; (d) an evaluation tool to assess the degree of success of strategic learning intervention courses or programs; and (e) an advising/counseling tool" to direct students to helpful resources or place them in a learning to learn class (Weinstein and Acee, 2013). Students can generate declarative (knowing about a variety of learning strategies), procedural (knowing how to effectively and efficiently use a variety of learning strategies), and conditional (knowing when and under what circumstances it may be useful to use particular learning strategies) knowledge and skills in each of the categories of the LASSI scales. Each form of knowledge can help students reach their learning goals more effectively and efficiently across a variety of learning environments. Students also need to know how to pick and choose learning strategies from among the various elements within the skill, will, and self-regulation components of the MSL to help them reach specific learning goals and objectives. For example, students experiencing high levels of anxiety about an upcoming math exam might need to use strategies to reduce their anxiety before they can effectively use information processing strategies to study for the exam. The next sections will highlight the individual LASSI scales and the component of the MSL that they are most related to conceptually. This discussion of the LASSI is taken directly from the *User's Manual* with permission from the authors and publisher (Weinstein and Palmer, 2002).

LASSI Scales Most Related to the Skill Component of Strategic Learning

The LASSI scales most related to the skill component of strategic learning are Information Processing, Selecting Main Ideas, and Test Strategies. These scales examine students' learning strategies, skills, and thought processes related to identifying, acquiring, and constructing meaning for important new information, ideas, and procedures, and how they prepare for and demonstrate their new knowledge on tests or other evaluative procedures.

The Information Processing Scale assesses how well students can use imagery, verbal and visual elaboration, organization strategies, and reasoning processes as learning strategies to help learn new information and skills and to build bridges between what they already know and what they are trying to learn and remember. Do students try to summarize or paraphrase their class reading assignments? Do they try to relate what is being presented in class to their prior knowledge? The Selecting Main Ideas Scale assesses students' skill at identifying important information for further study from less important information and supporting details. Can students identify the key points in a lecture? Can they decide what is important to underline in a textbook? The Test Strategies Scale assesses students' use of both test preparation and test taking strategies. Do they know how to study for tests in different types of courses? Do students review their answers to essay questions?

LASSI Scales Most Related to the Will Component of Strategic Learning

The LASSI scales related to the will component of strategic learning are Anxiety, Attitude, and Motivation. These scales measure the degree to which students worry about their academic performance, their receptivity to learning new information, their attitudes and interest in college, and their diligence, self-discipline, and willingness to exert the effort necessary to successfully complete academic requirements. The Anxiety Scale assesses the degree to which students worry about school and their academic performance. Do students worry so much that it is hard for them to concentrate? Are they easily discouraged by low grades? The Attitude Scale assesses students' attitudes and interests in college and achieving academic success. How clear are students about their own educational goals? Is school really important or worthwhile to them? The Motivation Scale assesses students' diligence, self-discipline, and willingness to exert the effort necessary to successfully complete academic requirements. Do they stay up to date in class assignments? Do students easily 'give up' in difficult classes?

LASSI Scales Most Related to the Self-regulation Component of Strategic Learning

The LASSI scales related to the self-regulation component of strategic learning are Concentration, Self-Testing, Study Aids, and Time Management. These scales measure how students manage, or self-regulate and control, the whole learning process through using their time effectively, focusing their attention and maintaining their concentration over time, checking to see if they have met the learning demands for a class, an assignment, or a test, and using study supports such as review sessions, tutors, or special features of a textbook. The Concentration Scale assesses students' ability to direct and maintain their attention on academic tasks. Are students easily distracted? Can they direct their attention to school tasks? The Self-Testing Scale assesses students' use of reviewing and comprehension monitoring techniques to determine their level of understanding of the information or task to be learned. Do the students review before a test? Do they stop periodically while reading to review the content? The Study Aids Scale assesses students' creation and use of support techniques, materials, or resources to help them learn and remember new information. Do students complete practice exercises? Do they create or use organizational aids? The Time Management Scale assesses students' use of time management principles and tactics for academic tasks. Are students well organized? Do they anticipate scheduling problems?

Once students and educators know the strengths and weaknesses of their students' learning to learn skills and strategies, the next step is to develop, adapt, use, and evaluate

interventions designed to address students' weaknesses and build on their strengths. In the next section, types of interventions will be described with a focus on a semester-long intervention for students who are particularly weak in a number of areas in the Model of Strategic Learning (MSL) and measured by the LASSI.

Learning to Learn Interventions

A number of different types of academic prescriptions and applications have been developed to enhance students' learning to learn skills, motivation, affect, and behavior (see Hodges and Agee, 2009). These research-based interventions have been primarily developed for college or university students but are currently being expanded to focus on K-12 education as well, particularly for high school students. Part of the difficulty teaching learning to learn strategies to younger students is the developmental progression in thinking that takes place as children age. Many of the strategies that have been studied require a level of cognitive ability beyond the scope of younger children. However, some researchers are directing their attention to precursors of more sophisticated adult learning to learn skills and processes, and this is an exciting area for both future research and application (Kaya and Kablan, 2013; Kitsantas et al., 2009).

Interventions developed to enhance college students' learning to learn strategies include workshops and short courses, tutoring, supplemental instruction (SI), and extended interventions (e.g., semester-long learning to learn courses, paired courses, and summer bridge programs).

Workshops and Short Courses

Workshops and short courses will usually target specific strategies and skills, such as goal setting, time management, elaborative note taking, learning from text, and comprehension monitoring. Their duration varies but usually ranges from 1 to 8 h. These interventions are often delivered by learning center staff, counselors, or content-area instructors with a background in, or related to, educational psychology. In some cases, these interventions are integrated within content-area courses. Although these interventions can be helpful, they are not as powerful as more extensive interventions (like a semester-long learning to learn course) because the content of the workshop is generally isolated from other core elements of strategic learning that may interact with the process being taught. For example, a workshop teaching reading strategies for mathematics textbooks may not be as effective if part of the problem is that the student has no goals or interest in learning mathematics. One of the first things this student would need to do is work on his or her motivation, for example, by learning and implementing goal setting and interest enhancement strategies.

Tutoring and SI

Tutoring is another delivery mode that has been researched as a means to help students develop and expand their repertoire of learning to learn strategies. Tutoring usually occurs in a one-to-one or small group setting and either focuses directly on teaching learning to learn or embeds the instruction in the help given students with the content they are having trouble understanding. It is a targeted intervention usually restricted to one content area, such as biology or psychology, so teaching students how to transfer what they are learning about how to learn to other courses is not a priority. Tutors can be students who excelled in the target course in a previous semester and are trained how to be an effective tutor, paraprofessionals who undergo a formal training program, or professional student support personnel who undergo more extensive training in tutoring and higher education. SI is a more elaborate and extended form of tutoring and can take a variety of forms. In its most basic form SI is an academic support program that uses peer-assisted supplementary class sessions to enhance student performance and retention in classes that historically have high failure and/or low retention rates. The peer leaders are usually students who succeeded in the course previously and they help current students by providing regular review sessions, getting students to work collaboratively with others in the class, discuss readings, and prepare for tests. Some SI leaders also incorporate how to study the content area while they help clarify and explain the material. SI can be an option in a course, required, or, in multiple-section courses, it can be a component of a subset of the sections. It has proved to be highly effective, particularly for high-risk courses. Although SI is an extended intervention that takes place across an entire semester, it focuses on a particular subject matter. It is a highly successful intervention but limited in its generalizability to other courses in different content areas. Of course, this is not the goal for SI. For researchers more interested in the acquisition of a range of learning to learn skills and strategies that can be applied to a variety of content areas and a number of different academic tasks, more extended and extensive interventions have been developed and investigated.

Extended Leaning to Learn Interventions

Extended learning to learn interventions include (1) semester-long courses in learning to learn (often called strategic learning, learning strategies, or learning frameworks courses); (2) paired courses, where a learning to learn course is paired with the same content-area course and the learning to learn strategies and skills being learned can be applied and reinforced in the content-area course; (3) summer bridge programs that incorporate an intensive and often condensed learning to learn component; and (4) implementing a metacurriculum that involves content-area instructors teaching students how to learn the content of their courses in addition to teaching the content itself.

Semester-Long Learning to Learn Courses

Semester-long learning to learn courses focus on a wide variety of learning strategies and skills from all of the components of the MSL (see Weinstein and Acee, 2013; Weinstein et al., 2000). In the United States, these courses are found in all types of tertiary education but are most often housed in 2-year academic and vocational training settings. Extended

interventions are designed to help students improve in areas where they are not strong (as measured by an instrument such as the LASSI) and excel in areas where they have existing strengths. Although these courses do not focus on any one specific academic subject area, they do use materials from a variety of majors, subject courses, and academic tasks to help students transfer their new knowledge and skills to their other concurrent and future classes. These courses can be enormously successful. For example, the effectiveness of a strategic learning course taught at the University of Texas at Austin that has been developed by Weinstein and her associates is evidenced by highly significant increases in general academic performance, reading comprehension, and retention to graduation. It is important to note that this is not simply a result of increased motivation by students who take the course. In fact, most students are forced into the course and both waiting list control groups and statistical control groups evidence much lower performance and retention.

Paired Courses

Paired courses is a related intervention where students participate in a learning to learn course as well as another academic course where the instructor gives the students opportunities to practice using their new knowledge and skills. The two instructors work together to integrate the two curricula and sometimes this includes sharing information across the semester on student progress. Usually, the paired course is one where students have traditionally had difficulty succeeding. This type of extended intervention capitalizes on the SI model but includes instruction in a wider variety of learning skills and strategies that can be used in other courses as well.

Summer Bridge Programs

Summer bridge programs are one other form of extended intervention. These can involve either a stand-alone course in learning to learn or paired courses. They are usually targeted to students who are identified to be at risk for low achievement, dropping out, or failure. The idea is to provide helpful strategies and skills before the students enter tertiary education. Often, summer bridge programs will also focus on adapting to the demands of college academically, personally, and socially to help students develop a realistic view of what college will be like for them, the resources available to help them, and the academic demands and rigor that will be required.

Current Research, Methodological Issues, and Directions for Future Research

Much of the current research in learning to learn has focused on improving student learning outcomes, primarily in college or tertiary settings. Researchers have demonstrated the effectiveness of teaching learning to learn strategies to students, often with quite dramatic results (see Hodges and Agee, 2009). These results have also been robust across content domains, individual classes, and years of college.

Given the importance of learning to learn for student achievement and retention to graduation, additional research is needed to address some of the methodological and conceptual issues in this area. First, there is a need for the development of consistent terms for the various aspects of learning to learn. Currently, each individual or group of researchers uses somewhat different terms to often describe very similar ideas and processes. Second, there is a need for the further development of models of learning to learn at the global level and at more task-specific levels. Third, more knowledge is needed about the nature of transfer of 'learning to learn' strategies. Fourth, measurement of 'learning to learn' strategies and processes needs to be refined. Current instruments like the *Learning and Study Strategies Questionnaire* and the *Motivated Strategies for Learning Questionnaire* are good general diagnostic screening measures, but more precise measurements of the components of 'learning to learn' and how they interact in a given traditional, blended, or online learning context are needed. Fifth, additional developmental work to create instructional paradigms to teach learning to learn in traditional, blended, and online learning contexts is needed.

'Learning to learn' is crucial for success in any learning context. As society moves toward a world where lifelong learning is a fundamental reality, the importance of understanding and using learning to learn strategies will only increase.

See also: Cognitive Development: Mathematics Learning and Instruction; Cognitive Styles and Learning Styles; Educational Learning Theory; Mastery Learning; School Learning for Transfer; Self-Regulated Learning.

Bibliography

Flavell, J.H., 1979. Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. American Psychologist 34 (10), 906–911.

Hacker, D.J., Dunlosky, J., Graesser, A.C. (Eds.), 2009. Handbook of Metacognition in Education. Taylor & Francis, New York.

Hodges, R., Agee, K.S., 2009. Program management. In: Flippo, R.F., Caverly, D.C. (Eds.), Handbook of College Reading and Study Strategy Research, second ed. Routledge, New York, pp. 351–378.

Kaya, S., Kablan, Z., 2013. Assessing the relationship between learning strategies and science achievement at the primary school level. Journal of Baltic Science Education 12 (4), 525–534.

Kitsantas, A., Steen, S., Huie, F., 2009. The role of self-regulated strategies and goal orientation in predicting achievement of elementary school children. International Electronic Journal of Elementary Education 2 (1), 65–81.

Pintrich, P.R., 2004. A conceptual framework for assessing motivation and self-regulated learning in college students. Educational Psychology Review 16 (4), 385–407.

Pintrich, P.R., Smith, D.A., Garcia, T., McKeachie, W.J., 1991. A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ) (Technical Report No. 91-B-004). National, Ann Arbor, MI.

Schunk, D.H., Pintrich, P.R., Meece, J.L., 2008. Motivation in Education: Theory, Research, and Applications, third ed. Prentice Hall, Upper Saddle River, NJ.

Svinicki, M., McKeachie, W.J., 2011. Teaching Tips. Strategies, Research, and Theory for College and University Teachers, thirteenth ed. Wadsworth, Cengage Learning, Belmont, CA.

Weinstein, C.E., Acee, T.W., 2013. Helping college students become more strategic and self-regulated learners. In: Bembenutty, H., Cleary, T.J., Kitsantas, A. (Eds.), Applications of Self-regulated Learning across Diverse Disciplines: A Tribute to Barry J. Zimmerman. Information Age Publishing, Charlotte, NC, pp. 197–236.

Weinstein, C.E., Acee, T.W., Jung, J.H., 2011. Self-regulation and learning strategies. New Directions for Teaching & Learning 2011 (126), 45–53. http://dx.doi.org/10.1002/tl.443.

Weinstein, C.E., Mayer, R.E., 1986. The teaching of learning strategies. In: Wittrock, M. (Ed.), Handbook of Research on Teaching, third ed. Macmillan, New York, pp. 315–327.

Weinstein, C.E., Husman, J., Dierking, D.R., 2000. Self-regulation interventions with a focus on learning strategies. In: Boekaerts, M., Pintrich, P.R., Zeidner, M. (Eds.), Handbook of Self-regulation. Academic Press, San Diego, CA, pp. 728–747.

Weinstein, C.E., Palmer, D.R., 2002. User's Manual for Those Administering the Learning and Study Strategies Inventory, second ed. H & H Publishing, Clearwater, FL.

Weinstein, C.E., Palmer, D.R., Schulte, A., 2002. The Learning and Study Strategies Inventory, second ed. H & H Publishing, Clearwater, FL.

Zimmerman, B.J., 2008. Investigating self-regulation and motivation: historical background, methodological developments, and future prospects. American Educational Research Journal 45 (1), 166–183. http://dx.doi.org/10.3102/0002831207312909.

Zimmerman, B.J., Schunk, D.H. (Eds.), 2011. Handbook of Self-regulation of Learning and Performance. Taylor & Francis, New York.

Zimmerman, B.J., Schunk, D.H. (Eds.), 2003. Educational Psychology: A Century of Contributions. Erlbaum, Mahwah, NJ.

Relevant Websites

http://www.crla.net/ – College Reading & Learning Association.
http://www.hhpublishing.com/ – H&H Publishing.
http://www.howtostudy.org/ – How to Study.
https://teachingcenter.ufl.edu/ – University of Florida Teaching Center.
http://www.nade.net/ – National Association for Developmental Education.
http://www.sc.edu/fye/ – University of South Carolina National Resource Center.
http://www.studygs.net/index.htm – Study Guides and Strategies.

Legal Bases of Social Work

Michael Preston-Shoot, University of Bedfordshire, England

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores the relationship between law and social work. It navigates the importance of the relationship and the similarities and differences between the objectives of legal and social work practitioners. It critically reviews transnational conventions and how individual countries have legislated to safeguard and promote the welfare of children and adults at risk. It reflects on how far nation-states uphold human rights and the rule of law and evaluates how social work has been positioned and has performed in different legislative contexts.

Introduction

Why should social work concern itself with law and with the legal frameworks for practice? Why is applying law to social work identified as a key component of qualifying education across countries with either developed or developing social welfare provision, for example in Europe (Adams et al., 2001), Australia (Swain and Rice, 2009), Israel (Doron et al., 2010), and India (Menoa, 2002)?

Perhaps, most obviously, social work practice does not operate in a vacuum. It is located in a context, a country's history, culture, and politics (Tice and Long, 2009), which combine to shape its development of legal rules and its response to people's needs. Thus, social work resides within a macro context with degrees of debate about social policy. It engages both in a micro-context when it interacts with individuals and their families and communities, and in a mezzo- or organizational context when it participates in a process by which legal rules and social policy intentions are translated into operational procedures.

Less obviously but no less critically, social workers are themselves citizens whose lives, along with those with whom they work, are shaped significantly by what is (not) legislated, the content of and omissions from the legal rules. Their lives will be shaped by how legislation is interpreted and implemented through local procedures and practices, and by how authority is exercised by decision-makers. As citizens and as professional practitioners, social workers will be profoundly affected by the values which the legal rules uphold and promote (Preston-Shoot et al., 2001).

Moreover, social work has positioned law centre-stage. With human rights and social justice as principled commitments, and with promoting individual and social change to enhance people's well-being as a core focus (IFSW/IASSW, 2001), social work engages both ethics and law. However, despite its commitment to challenging national governments and social welfare organizations when policies and practices are oppressive, inhumane, and unethical (Hölscher and Berhane, 2008), its leadership on and advocacy for human rights and social change has not been unequivocal (Laird, 2007; Healy, 2008; Preston-Shoot, 2010).

Equally, social work education and practice has been slow to develop guidance on how practitioners and managers should respond when legal rules and professional ethical standards collide rather than converge. Social workers and legal practitioners have arguably also been slow to appreciate the contribution that each other's professionalism can make to their practice (Dickens, 2005; Zifcak, 2009).

Accordingly, this contribution locates social work within national legal rules and transnational conventions prior to identifying critical questions for practitioners, managers, and educators. These questions consider how countries approach the development of social work provision, and the promotion of equality and human rights. They explore critically how countries respond to social issues, including the balance struck between the needs of children and adults at risk with available resources. The contribution closes by reviewing the impact of national and transnational principles and legal rules on people's quality of life, and of social work as an independent, critical voice and practice on behalf of those who face disadvantage.

Transnational Declarations

Perhaps because national governments acknowledge the need to act collectively to guarantee human rights, the number of supranational conventions is growing. Some relate to specific continents, for example the European Convention on Human Rights and Fundamental Freedoms (ECHR) (1950), the European Social Charter (1961), the African Charter on Human and People's Rights (1968), the African Charter on the Rights and Welfare of the Child (1999), and the American Declaration on Human Rights (1969). Others are global in ambition, such as the Universal Declaration of Human Rights (1948), UN Convention on the Rights of the Child (UNCRC) (1989), the International Covenant on Economic, Social and Cultural Rights (1966), the International Convention on the Elimination of all forms of Racial Discrimination (1969), the Convention on the Elimination of all forms of Discrimination against Women (1979), the UN Convention on the Rights of Persons with Disabilities (2006), and the UN Standard Minimum Rules for the Administration of Juvenile Justice (1985).

Key principles, aligning closely with social work values, emphasize fair and humane treatment, social inclusion, promotion of health and well-being, equality, protection, dignity, and social justice. They advance a rights-based discourse and represent a powerful advocacy and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28050-1

empowerment tool for social groups and the social welfare professionals working with them (Doron and Apter, 2010). However, their reach and impact are nullified if ratification by individual countries does not subsequently translate into implementation of their principles and standards in national laws and policies (Nicklett and Perron, 2010). Explanations vary for failing to legislate nationally to embed the conventions in state jurisdictions but include a preference to express human rights through cultural and religious traditions (Skegg, 2005), resource austerity and opposition from government departments or third sector organizations (Tang, 2003).

The rights contained in transnational declarations remain soft (Tang, 2003) and limited in their effectiveness not just when they are not embedded in national jurisdictions but also when their recommendations and safeguards are nonbinding. Without sanctions for noncompliance (Mishra, 2005), the rights risk being declaratory and symbolic rather than a force for meaningful social change (Doron and Apter, 2010). Thus, despite the UN Standard Minimum Rules for the Administration of Juvenile Justice, not all states have created a separate system from adult criminal justice for young people or raised the age from which children can be prosecuted for offences. Similarly, the African Charter on Human and People's Rights creates a commission to promote such rights as to life, liberty, education, and the best attainable physical and mental health, but not a court capable of providing redress to individuals. Moreover, even when bodies such as the International Court of Justice are created, less than a third of states represented in the United Nations accept such jurisdiction (Bingham, 2011), while the UK government is probably not alone in behaving lukewarmly toward the European Court of Human Rights depending on the outcome of its judgments.

The UK's approach to unaccompanied asylum-seeking minors, the age of criminal responsibility for children, and corporal punishment, falls outside the UNCRC, for which it has been sharply criticized but has remained apparently unmoved (Braye and Preston-Shoot, 2009). This illustrates how adherence to UNCRC principles by national governments has been piecemeal and half-hearted (Nicklett and Perron, 2010). By contrast, unfavorable judgments in the European Court of Human Rights, which enforces the rights contained in the ECHR, have led to changes in legislation in the UK and other signatory states, with a direct impact on social work practice (e.g., Gaskin v UK [1986] 12 EHRR 36; Z v UK [2001] 2 FLR 612). Here then is an example of how individual citizens can hold national governments and their agents accountable for acts of abuse, harm, omission, and discrimination.

The picture of incomplete protection is reinforced by gaps within the framework of human rights, including the absence of an international convention on the rights of older people (Doron and Apter, 2010), and the failure to articulate what levels of social rights, or norms and standards of social provision, are feasible when considered against the resource capacity of individual nation-states (Mishra, 2005).

Holding Governments Accountable

With its espoused commitment to human rights and social justice (IFSW/IASSW, 2001), social work might be expected to engage with holding local and national governments accountable for how people's needs are met and their rights protected. Where countries as diverse as South Africa (Sachs, 2009) and the Czech Republic (Chytil and Popelková, 2000) have constitutions, close scrutiny reveals what rights and freedoms are upheld and the ease with which citizens may access constitutional courts. Where countries have developed or developing systems of human rights and the rule of law, as in the UK (Bingham, 2011) and Mongolia (Namdaldagva et al., 2010), detailed monitoring is required of how legal rules codify individual needs and rights, accord power to state organizations, and facilitate citizens to hold such agencies accountable for abuse of power through access to legal remedies. As Bingham (2011) argues, the price for complacency can be despotic democracy or elected dictatorship.

The Forum for a New World Governance (2011) offers one assessment of how individual countries perform on five domains, namely peace and security (including approach to refugees, asylum seekers, and violent crime), rule of law (including independence and effectiveness of the judiciary), human rights and participation (covering respect for civil rights and promotion of equality), sustainable development (embracing employment protection), and human development (including quality of life and promotion of well-being). On overall rankings Scandinavian countries occupy the first five places, with other results including New Zealand (6), Australia (9), Canada (12), France (15), the UK (18), the US (32), Israel (49), Croatia (50), South Africa (68), Mongolia (75) Thailand (85), and Vietnam (120).

This assessment contextualizes reports from social work and legal practitioners of abuses of power. These include the abrogation of civil liberties and human rights in the UK and the US (Bingham, 2011), the systematic violation of civil rights in Eritrea (Hölscher and Berhane, 2008), variable responses to crimes motivated by hate (Pollack, 2009), and the violation of human rights and social justice in West African states by sexual abuse, neglect of disabled children, child labor, and trafficking (Sossou and Yogtiba, 2009). In the UK, public authorities and legal systems use human rights positively to protect individuals and communities. Nonetheless, human rights could be protected more fully (EHRC, 2012), evidence uncovering failures to protect the rights of children and disabled people, abuse of vulnerable people in institutional care, and degrading treatment toward victims of human trafficking and unaccompanied asylum-seeking minors.

The index, in evaluating how countries promote the rule of law and human rights, captures how far national jurisdictions embrace social and economic rights alongside civil and political rights (Skegg, 2005; Sachs, 2009). It clearly reveals that ratification of international and regional conventions, and passing of national legislation, is no guarantee that children and adults at risk will be protected, perhaps because of a lack of state infrastructure, the influence of cultural norms, and poverty of resources (Sossou and Yogtiba, 2009). Even where a legislative and procedural perspective emphasizes dignity, respect, and inclusion, in countries such as Australia, Scandinavia, and the UK, it is not just resources which create a gulf between policy intentions and practice outcomes, but also professional attitudes whereby social work staff and agencies

fail to advocate for, and protect service users (Johnson and Marriott, 2009; Preston-Shoot, 2010).

National Social Welfare Legislation

Social work is located within country-specific legal rules which express a nation-state's social policy. Social work is therefore shaped by and should comment on the balance that is struck between empowerment and protection of individuals, families, and communities; voluntary engagement with or enforced control of people at risk, such as those who misuse drugs and alcohol; and upholding individual autonomy or acting in someone's best interests (Fasol and Fraccaroli, 2001). Legislative activity on social welfare issues may be reluctant or enthusiastic, partial or ambitious, ambivalent or assertive on human rights and equality, evidence informed or ideologically driven, and motivated by religious convictions or secular orientations. Legal rules may benefit or disenfranchise people with diverse social problems and will shape the relationship between public service providers and independent welfare or nongovernmental organizations (Klug, 2001). Social work is well placed to comment on the degree to which needs are met by current national legislation and welfare systems (Sasaki, 2010).

Social work acquires powers and duties, roles, and tasks, from legislation, for example in the UK and Finland (Hämäläinen and Niemalä, 2000). However, social workers should act collectively in their national context to comment upon the degree to which those in need or at risk have legal rights or entitlements to assistance (Erath et al., 2000) through a state's constitution or rule of law. Thus, how far are legal rules framed around notions of need, who is (not) deserving of assistance, or principles of social cohesion, citizenship, inclusion, and civil rights (Feu, 2000; Jovelin and Tully, 2000)? Critique is possible of how legal rules and social policy articulate or ignore the assistance required by those in need due to older age, poor health, poverty, homelessness, mental distress or incapacity, unemployment, and child abuse or neglect (Feu, 2000; Erath et al., 2000; Chytil and Popelková, 2000; Hämäläinen and Niemalä, 2000). Social work advocacy may be important as different nation-states, for example in Asia, struggle to frame a legislative response to rapid growing social, environmental, and economic challenges, such as racial discrimination (Oommen, 2002); drug abuse; population and ecological pressures; hate crime, child sexual exploitation, and human trafficking (Thi Thai Lan et al., 2010; Zaviršek and Herath, 2010; Namdaldagva et al., 2010). Similarly, detailed challenge may be necessary as states debate whether to extend beyond civil and political rights toward social, economic, and environmental rights, and as legal rules struggle with questions of distributive justice and relational justice – with how goods and services are made available and with how power is shared and meaningful participation guaranteed to citizens (Tice and Long, 2009).

Thus, for example, child care and child protection legislation has been enacted and, to varying degrees, given statutory roles to social workers, in West African states (Sossou and Yogtiba, 2009), Vietnam (Thi Thai Lan et al., 2010), Taiwan (Lin and Wang, 2010), Japan (Sasaki, 2010), Costa Rica (Tice and Long, 2009), and Nepal (Nikku, 2010). Equally, concern exists about the lack of protection for children at risk, for example in Malawi and Mongolia (Tice and Long, 2009) and in China (Scannapieco and Batchelor, 2012) where legislative support is limited and perhaps an impediment to the development of a public child welfare infrastructure. The position in respect of countries enacting legislation on child trafficking, and then giving social work a prominent role, is variable.

Similar variability can be found with respect to disabled people, older people, gender equality, and domestic violence, with positive legislative developments, for example, in Japan and Taiwan (Sasaki, 2010; Lin and Wang, 2010) but concerns about the adequacy of legal frameworks to protect older and disabled people in Portugal (Carvalho, 2012), Kenya, Ghana, and Zimbabwe (Mba, 2007), and across African states and the UK (Krug et al., 2002; Daichman, 2005; Ozanne, 2009; Braye et al., 2012).

However, social work's critique must also extend to consider whether the legal rules are accessible, intelligible, clear, and predictable (Bingham, 2011). Adult protection research in the UK has found that social workers and other professionals experience considerable difficulty in understanding the legislative requirements, for example on the interface between human rights and the sharing of confidential information, and in negotiating the boundaries between an individual's right to autonomy in their private and family life and their mental capacity (Perkins et al., 2007; Pinkney et al., 2008). The volume and complexity of the legislation does not necessarily serve the rule of law (Bingham, 2011) or facilitate the meeting of their needs.

Three cautionary notes should be sounded here. The first is about expecting too much of the law alone. Sachs (2009) advises skepticism of the law's claims but avoidance of cynicism about its possibilities. Put another way, helping people meet their needs and overcome disadvantage and oppression requires law and social work values, knowledge, and skills (Braye and Preston-Shoot, 1990). The second is that social work has not consistently engaged in legislative debates about social questions (van der Laan, 2000). The third is that the level of a country's economic development is a blunt predictor of the degree to which it commits to the development of the rule of law, human rights, and the protection of children and adults at risk (Nicklett and Perron, 2010; Forum for a New World Governance, 2011). Although less industrialized countries generally have lower levels of legislative activity than middle-income countries, the pattern is not uniform, as the 2011 governance index reveals for countries like Costa Rica and Namibia.

Implementing Legal Provisions

Previous sections charted the development of the law in theory. Here the focus is on the law in-between (Jenness and Grattet, 2005) – whether and how legal provisions are translated into operational procedures by organizations responsible for the discharge of statutory powers and duties. This is because legislation may be applied in ways not foreseen or intended by legislators (van der Laan, 2001). Lack of legal and organizational infrastructure, poverty of resources, poor planning, professional cultures, excessive workloads, lack of coordination

and personal attitudes among the workforce may explain this phenomenon (Braye and Preston-Shoot, 2009; Nicklett and Perron, 2010).

Implementation problems include mandatory reporting in Cyprus of domestic violence (Panayiotopoulos, 2011), the failure to protect children from being trafficked and enforce provisions for treating young offenders separately from adults in Nepal and the Philippines (CLRRD, 2002; Boele, 2005), and the availability of direct payments for adults requiring community care support in the UK (Braye and Preston-Shoot, 2009). In China, Scannapieco and Batchelor (2012) argue that child abuse is insufficiently addressed, with maltreatment inadequately defined in government policy, and often tolerated as a family matter.

Then there is the law in practice, how social workers, social welfare organizations, and other professionals implement legislative provisions. Examples of institutional abuse, where children have not been protected, for example in the UK, Sweden, Malawi, and Romania (Tice and Long, 2009), illustrate the challenges involved in ensuring that those responsible for safeguarding and promoting the welfare of young people know what the legal rules expect of them and have the values, knowledge, and skills to meet their professional responsibilities. Sometimes social workers have exposed these abuses and experienced considerable hostility from their organizational employers as a result; sometimes, however, concerns have not been escalated about known abuses and practice failings (Kline and Preston-Shoot, 2012). A similar picture emerges regarding adults in need or at risk (Preston-Shoot, 2010).

Regulating Social Work

Whether social work exercises vigilance and advocacy concerning how social welfare legislation is developed and implemented will depend, partly, on if it is recognized and supported nationally. Currently, a minority of African countries have social work education programs (Sossou and Yogtiba, 2009). In Asia, a variable picture emerges, with governments yet to formally recognize social work in Nepal (Nikku, 2010) and Vietnam (Thi Thai Lan et al., 2010) but Taiwan and Japan have systems of certification and examination in order for people to practise as social workers (Lin and Wang, 2010; Sasaki, 2010). In China, social work education programs have proliferated but, as a profession, social work has yet to become established, with its role and status recognized (Liu et al., 2012).

The position on social work regulation also varies. Across Europe, some states have developed or are evolving systems of regulation and protected title (Hämäläinen and Niemalä, 2000; Shardlow, 2001). In one review (Weiss-Gal and Welbourne, 2008), the UK, Spain, and South Africa had licensing procedures but Sweden, Chile, Hungary, Mexico, and India did not. Such regulation may reinforce notions of professionalism and accountability but does not guarantee ethical or lawful behavior (Preston-Shoot, 2010). While all the countries surveyed had distinctive social work education programs in higher education institutions, not all had a national ethical code and social workers enjoyed variable levels of prestige and status which impacted on their ability to exert influence on policy development. Similarly, social work has not established professional associations across all nation-states, which equally impacts on its ability to exert influence. Moreover, where such associations do exist, dependence on government funding or exposure to political pressure may weaken their autonomy and advocacy (Yu, 2006).

Curriculum

Whether social workers can engage in policy development and advocacy surrounding the legal rules, as well as in case-based tasks, will depend on the content of their qualifying education and continuing professional development. Law is increasingly being taught within social work education, for example in India (Menoa, 2002), Nepal (Nikku, 2010), Vietnam (Thi Thai Lan et al., 2010), and Taiwan (Lin and Wang, 2010). Law is well established in social work curricula in the UK, the US, Israel, and Australia (Braye et al., 2005; Doron et al., 2010; Swain and Rice, 2009).

Evidence about the outcomes of teaching and learning law in social work education is not extensive. However, there is some empirical evidence to suggest that social work students emerge from their qualifying studies more positive about using the law in practice and with greater confidence in their legal knowledge and practice skills (Doron et al., 2010; Preston-Shoot and McKimm, 2012a).

However, in a crowded curriculum, commentators from different countries have expressed concerns that insufficient attention is given to the relationship between law and social work, the socioeconomic and cultural factors that create inequality and disregard for human rights, and the development of knowledge and skills for working with children and adults at risk (Menoa, 2002; Sossou and Yogtiba, 2009; Nikku, 2010; Zaviršek and Herath, 2010; Preston-Shoot and McKimm, 2012b; Spajic-Vrkas, 2012). Social workers may also feel underprepared for actively engaging in political and social struggles, and for deploying advocacy skills in search of law reform (Sabater, 2001; Healy, 2007). Much, then, depends on whether the taught academic curriculum offers learning that focuses not just on the technical content of the legal rules (doing things right) but also on knowledge and skills required to negotiate the interface between law and professional values (doing right things) and between law, social work, and human rights (right thinking) (Braye and Preston-Shoot, 2009). Similarly, much also depends on whether what social workers learn in practice settings supports or undermines what they have been taught (Preston-Shoot and McKimm, 2012b).

One key element in the curriculum is learning about the similarities between law and social work in terms of values, objectives of securing people's rights and best interests, scrutiny of evidence, standards for decision-making and the exercise of authority, use of research, and advocacy (Hämäläinen and Niemalä, 2000; Preston-Shoot et al., 2001; Doron et al., 2010; Zifcak, 2009). Another key element is to develop a critique of both the legal rules and the systems that implement them, coupled with the development of skills to manage situations where law and professional values collide rather than converge, and where legal rules challenge traditional cultural norms (Healy, 2007; Preston-Shoot, 2010). Both elements need greater emphasis in curricula offered by social work education.

Concluding Discussion

Social workers, and their employing organizations, do make positive differences for children and adults at risk, often in challenging circumstances. They have challenged dehumanizing practices and empowered individuals and communities (Noyoo, 2004; Skegg, 2005). However, social work has sometimes failed to promote human rights and social justice, and to enhance the well-being of individuals, their families, and communities (IFSW/IASSW, 2001). This might be because in particular countries the legal rules and welfare frameworks do not exist or are under development, or do not define clearly what are the roles and tasks which social workers may assume. However, there are other possible explanations, including the tendency of social work to shy away from policy issues and from work with the most disadvantaged groups (Noyoo, 2004; Laird, 2007), and location within political situations which require considerable personal resilience and courage to practise according to the IFSW/IASSW statement (Yu, 2006; Healy, 2008; Hölscher and Berhane, 2008). Equally, social workers' professional autonomy may be circumscribed by the conditions of their employment (Weiss-Gal and Welbourne, 2008; Kline and Preston-Shoot, 2012). Their practice might be constrained by excessive bureaucracy and workloads (Chytil and Popelková, 2000; Oommen, 2002; Braye and Preston-Shoot, 2009; Panayiotopoulos, 2011) and/or by resource shortages (Hämäläinen, and Niemalä, 2000; Jovelin and Tully, 2000; Preston-Shoot, 2010).

Social workers must retain ways of exercising their independent voice despite these powerful organizational, structural, cultural, and political constraints, especially when employed by agencies which are reluctant to challenge received governmental wisdom and to uphold people's rights, dignity, and entitlements (Oommen, 2002; Preston-Shoot, 2010; Zaviršek and Herath, 2010). In such an environment the legal rules as expressed nationally and transnationally may provide important support but alone will be insufficient to prevent poor or abusive practice and to ensure the highest standards of ethical, research-informed, knowledge-driven, and skilled social work.

See also: Child Protection: International Issues; Child Protection; Children and Law; Civil Liberties and Human Rights; Critical Reflexivity and Social Work Practice; Empirical Legal Studies; Human Rights and Social Work; Human Rights, Anthropology of; Human Rights, History of; Indigenous Rights; Regulation and Administration; Risk and Uncertainty in the Social Sciences: Implications for Social Work Theory and Practice; Social Action and Social Justice; Social Justice and Criminal Justice ; Social Justice and Social Inequalities; Social Work Ethics.

Bibliography

Adams, A., Erath, P., Shardlow, S. (Eds.), 2001. Key Themes in European Social Work: Theory, Practice, Perspectives. Russell House Publishing, Lyme Regis.

Bingham, T., 2011. The Rule of Law. Penguin Books, London.

Boele, F., 2005. Rights of the Child in the Philippines. OMCT, Geneva.

Braye, S., Preston-Shoot, M., 1990. On teaching and applying the law in social work: it is not that simple. British Journal of Social Work 20 (4), 333–353.

Braye, S., Preston-Shoot, M., 2009. Practising Social Work Law, third ed. Palgrave Macmillan, Basingstoke.

Braye, S., Preston-Shoot, M., Cull, L.-A., Johns, R., Roche, J., 2005. Knowledge Review – Teaching, Learning and Assessment of Law in Social Work Education. Social Care Institute for Excellence, London.

Braye, S., Orr, D., Preston-Shoot, M., 2012. The governance of adult safeguarding: findings from research. Journal of Adult Protection 14 (2), 55–72.

Carvalho, M., July 2012. Social Work and Violence in the Elderly. Challenges to Promote the Rights and Dignity (A Case Study). Paper given at the IASSW, ICSW and IFSW 2nd Joint World Conference on Social Work and Social Development: Action and Impact. Stockholm.

Centre for Legal Research and Resource Development, 2002. Baseline Survey on Criminal Justice in Nepal. CLRRD, Bhaktapur.

Chytil, O., Popelková, R., 2000. Social policy and social work in the Czech Republic. In: Adams, A., Erath, P., Shardlow, S. (Eds.), Fundamentals of Social Work in Selected European Countries: Historical and Political Context, Present Theory, Practice, Perspectives. Russell House Publishing, Lyme Regis.

Daichman, L.S., 2005. Elder abuse in developing nations. In: Johnson, M.L. (Ed.), Cambridge Handbook of Age and Ageing. Cambridge University Press, Cambridge, pp. 323–331.

Dickens, J., 2005. Being "the epitome of reason": the challenges for lawyers and social workers in child care proceedings. International Journal of Law, Policy and the Family 19 (1), 73–101.

Doron, I., Apter, I., 2010. The debate around the need for an international convention on the rights of older persons. The Gerontologist 50 (5), 586–593.

Doron, I., Karpel, M., Or-Chen, K., 2010. Social workers' attitudes to the law: an Israeli perspective. International Journal of Social Welfare 19 (1), 95–103.

EHRC (Equality and Human Rights Commission), 2012. How Fair is Britain? Equality and Human Rights Commission, Manchester.

Erath, P., Klug, W., Sing, H., 2000. Social policy, social security and social work in Germany. In: Adams, A., Erath, P., Shardlow, S. (Eds.), Fundamentals of Social Work in Selected European Countries: Historical and Political Context, Present Theory, Practice, Perspectives. Russell House Publishing, Lyme Regis, pp. 49–64.

Fasol, R., Fraccaroli, F., 2001. Social alarm and public reassurance: the Italian example. In: Adams, A., Erath, P., Shardlow, S. (Eds.), Key Themes in European Social Work: Theory, Practice, Perspectives. Russell House Publishing, Lyme Regis, pp. 77–84.

Feu, M., 2000. The development of social work in Spain. In: Adams, A., Erath, P., Shardlow, S. (Eds.), Fundamentals of Social Work in Selected European Countries: Historical and Political Context, Present Theory, Practice, Perspectives. Russell House Publishing, Lyme Regis, pp. 103–118.

Forum for a New World Governance, 2011. World Governance Index. Why Should World Governance Be Evaluated and for what Purpose? Charles Léopold Mayer Foundation for the Progress of Humankind, Paris.

Hämäläinen, J., Niemalä, P., 2000. Social policy and social work in Finland. In: Adams, A., Erath, P., Shardlow, S. (Eds.), Fundamentals of Social Work in Selected European Countries: Historical and Political Context, Present Theory, Practice, Perspectives. Russell House Publishing, Lyme Regis.

Healy, L., 2007. Universalism and cultural relativism in social work ethics. International Social Work 50 (1), 11–26.

Healy, L., 2008. Exploring the history of social work as a human rights profession. International Social Work 51 (6), 735–748.

Hölscher, D., Berhane, S., 2008. Reflections on human rights and professional solidarity: a case study of Eritrea. International Social Work 51 (3), 311–323.

IFSW/IASSW, 2001. International Definition of Social Work. International Association of Schools of Social Work and the International Federation of Social Workers, Copenhagen.

Jenness, V., Grattet, R., 2005. The law-in-between: the effects of organizational perviousness on the policing of hate crime. Social Problems 52 (3), 337–359.

Johnson, K., Marriott, A., 2009. The neverending story: deinstitutionalisation and people with intellectual disability. In: Swain, P., Rice, S. (Eds.), In the Shadow of the Law. The Legal Context of Social Work Practice, third ed. The Federation Press, Sydney, pp. 34–36.

Jovelin, E., Tully, E., 2000. Social protection and social work in France. In: Adams, A., Erath, P., Shardlow, S. (Eds.), Fundamentals of Social Work in Selected European Countries: Historical and Political Context, Present Theory, Practice, Perspectives. Russell House Publishing, Lyme Regis, pp. 37–48.

Kline, R., Preston-Shoot, M., 2012. Professional Accountability in Social Care and Health: Challenging Unacceptable Practice and Its Management. Sage/Learning Matters, London.

Klug, W., 2001. Social work and the third sector: the example of the German welfare state. In: Adams, A., Erath, P., Shardlow, S. (Eds.), Key Themes in European Social Work: Theory, Practice, Perspectives. Russell House Publishing, Lyme Regis.

Krug, E., Dahlberg, L., Mercy, J., Zwi, A., Lozan, R. (Eds.), 2002. World Report on Violence and Health. World Health Organisation, Geneva.
Laird, S., 2007. The application of African practice models to the social problems of emergent democracies. In: Borrmann, S., Klassen, M., Spatscheck, C. (Eds.), International Social Work. Social Problems, Cultural Issues and Social Work Education. Barbara Budrich Publishers, Opladen.
Lin, W.-I., Wang, Y.-T., 2010. What does professionalization mean? Tracing the trajectory of social work education in Taiwan. Social Work Education 29 (8), 869–881.
Liu, Y., Lam, C.-M., Yan, M.-C., 2012. A challenged professional identity: the struggles of new workers in China. China Journal of Social Work 5 (3), 189–200.
Mba, C., 2007. Elder abuse in parts of Africa and the way forward. Gerontechnology 6 (4), 230–235.
Menoa, N., 2002. Law and social work: an emerging partnership? In: Visvesvaran, P. (Ed.), Social Work Today. Present Realities and Future Prospects. Madras School of Social Work, Chennai.
Mishra, R., 2005. Social rights as human rights. Globalizing social protection. International Social Work 48 (1), 9–20.
Namdaldagva, O.-E., Myagmarjav, S., Burnette, D., 2010. Professional social work education in Mongolia: achievements, lessons learned and future directions. Social Work Education 29 (8), 882–895.
Nicklett, E., Perron, B., 2010. Laws and policies to support the wellbeing of children: an international analysis. International Journal of Social Welfare 19 (1), 3–7.
Nikku, B.R., 2010. Social work education in Nepal: major opportunities and abundant challenges. Social Work Education 29 (8), 818–830.
Noyoo, N., 2004. Human rights and social work in a transforming society: South Africa. International Social Work 47 (3), 359–369.
Oommen, T., 2002. Social work and human rights: for a feasible linkage. In: Visvesvaran, P. (Ed.), Social Work Today. Present Realities and Future Prospects. Madras School of Social Work, Chennai.
Ozanne, E., 2009. Elder abuse and the law. In: Swain, P., Rice, S. (Eds.), In the Shadow of the Law. The Legal Context of Social Work Practice, third ed. The Federation Press, Sydney.
Panayiotopoulos, C., 2011. Mandatory reporting of domestic violence cases in Cyprus; barriers to the effectiveness of mandatory reporting and issues for future practice. European Journal of Social Work 14 (3), 379–402.
Perkins, N., Penhale, B., Reid, D., Pinkney, L., 2007. Partnership means protection? Perceptions of the effectiveness of multi-agency working and the regulatory framework within adult protection in England and Wales. Journal of Adult Protection 9 (3), 9–23.
Pinkney, L., Penhale, B., Manthorpe, J., Perkins, N., Reid, D., Hussein, S., 2008. Voices from the frontline: social work practitioners' perceptions of multi-agency working in adult protection in England and Wales. Journal of Adult Protection 10 (4), 12–24.
Pollack, D., 2009. Hate crimes and social work: an international perspective. International Social Work 52 (3), 409–415.
Preston-Shoot, M., 2010. On the evidence for viruses in social work systems: law, ethics and practice. European Journal of Social Work 13 (4), 465–482.
Preston-Shoot, M., McKimm, J., 2012a. Perceptions of readiness for legally literate practice: a longitudinal study of social work student views. Social Work Education 31 (8), 1071–1089.
Preston-Shoot, M., McKimm, J., 2012b. Tutor and student experiences of teaching and learning law in UK social work education. Social Work Education 31 (7), 896–913.
Preston-Shoot, M., Roberts, G., Vernon, S., 2001. Values in social work law: strained relations or sustaining relationships? Journal of Social Welfare and Family Law 23 (1), 1–22.
Sabater, J., 2001. Social work education and training; the Spanish experience. In: Adams, A., Erath, P., Shardlow, S. (Eds.), Key Themes in European Social Work: Theory, Practice, Perspectives. Russell House Publishing, Lyme Regis, pp. 85–95.
Sachs, A., 2009. The Strange Alchemy of Life and Law. Oxford University Press, Oxford.
Sasaki, A., 2010. Social work education in Japan: future challenges. Social Work Education 29 (8), 855–868.
Scannapieco, M., Batchelor, J., 2012. The importance of the social work degree in working with families and children: implications for China's growing social work profession. China Journal of Social Work 5 (3), 237–256.
Shardlow, S., 2001. Ethical aspects of social work – a common set of values and the regulation of practice: England an example. In: Adams, A., Erath, P., Shardlow, S. (Eds.), Key Themes in European Social Work: Theory, Practice, Perspectives. Russell House Publishing, Lyme Regis, pp. 39–46.
Skegg, A.-M., 2005. Human rights and social work. A Western imposition or empowerment to the people? International Social Work 48 (5), 667–672.
Sossou, M.-A., Yogtiba, J., 2009. Abuse of children in West Africa: implications for social work education and practice. British Journal of Social Work 39 (7), 1218–1234.
Spajic-Vrkas, V., July 2012. Human Rights and Aging in Two Contexts: Research on the Elderly Citizens of the City of Zagreb. Paper given at the IASSW, ICSW and IFSW 2nd Joint World Conference on Social Work and Social Development: Action and Impact. Stockholm.
Swain, P., Rice, S. (Eds.), 2009. In the Shadow of the Law. The Legal Context of Social Work Practice, third ed. The Federation Press, Sydney.
Tang, K.-L., 2003. Implementing the United Nations convention on the rights of the child. The Canadian experience. International Social Work 46 (3), 277–288.
Thi Thai Lan, N., Hugman, R., Briscoe, C., 2010. Moving towards an 'indigenous' social work education in Vietnam. Social Work Education 29 (8), 843–854.
Tice, C., Long, D. (Eds.), 2009. International Social Work Policy and Practice. Wiley, New Jersey.
van der Laan, G., 2000. Social work in the Netherland. In: Adams, A., Erath, P., Shardlow, S. (Eds.), Fundamentals of Social Work in Selected European Countries: Historical and Political Context, Present Theory, Practice, Perspectives. Russell House Publishing, Lyme Regis.
van der Laan, G., 2001. Social work and economic policy: the Netherlands an example. In: Adams, A., Erath, P., Shardlow, S. (Eds.), Key Themes in European Social Work: Theory, Practice, Perspectives. Russell House Publishing, Lyme Regis.
Weiss-Gal, I., Welbourne, P., 2008. The professionalisation of social work. International Journal of Social Welfare 17 (4), 281–290.
Yu, N., 2006. Interrogating social work: Philippine social work and human rights under martial law. International Journal of Social Welfare 15, 257–263.
Zaviršek, D., Herath, S., 2010. "I want to have my future, I have a dialogue": social work in Sri Lanka between neo-capitalism and human rights. Social Work Education 29 (8), 831–842.
Zifcak, S., 2009. Towards a reconciliation of legal and social work practice. In: Swain, P., Rice, S. (Eds.), In the Shadow of the Law. The Legal Context of Social Work Practice, third ed. The Federation Press, Sydney.

Legal Culture and Legal Consciousness

Susan S Silbey, School of Humanities, Arts, and Social Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Legal culture and legal consciousness are used analytically to identify the meanings of law that circulate in social relations, referring to what people do as well as say about law. The prefix *legal* characterizes an aspect of the general culture that is associated with law, legal institutions, legal actors, and their behaviors. Culture is used to refer to the system of communicative signs within human groups. Legal *culture* refers to an aggregate (macro or group) level phenomenon; legal *consciousness* usually refers to micro-level social action, specifically the ways in which individuals mobilize, invent, and interpret legal meanings and signs.

The phrases 'legal culture' and 'legal consciousness' are used analytically to identify the meanings of law that circulate in social relations. The prefix *legal* characterizes an aspect of the general culture that is associated with law, legal institutions, legal actors, and their behaviors. Legal *culture* refers to an aggregate (macro or group) level phenomenon; legal *consciousness* usually refers to micro-level social action, specifically the ways in which individuals mobilize, invent, and interpret legal meanings and signs.

Culture and consciousness are hotly debated and contested constructs created by scholars for diverse research purposes; the adoption of these terms specifically in legal scholarship has also been fraught with controversy (Nelken, 2012). The referents of these terms are unstable, both theoretically and empirically. This may be due to a long history in social science as well as common parlance, but it may also be, as Raymond Williams writes, because "culture is one of the two or three most complicated words in the English language" (Williams, 1983: p. 87). Some confusion derives from intermingling two common meanings of culture: as an analytic tool of social analysis referring to the communicative semiotic aspects of human action and as empirically observable ways of doing and thinking that are associated with a particular group or society. In the latter use, describing the distinctive customs, opinions, and practices of a group or society, the term is often used in the plural, as in the legal cultures of Japan and China or in reference to African or Latin cultures. In the former analytic sense, the word is used in the singular, as in legal culture, or the culture of academia. These distinctions need to be clearly specified in using these terms, lest confusion proliferate.

Despite often theoretically obstruse debate, many scholars find the concepts of legal culture and consciousness hard to avoid when they address aspects of law not confined to formal legal materials (such as doctrines, statutes, cases, constitutions) or offices and institutions (such as lawyers, judges, courts). Legal culture and consciousness are terms used to emphasize, for analytic purposes, ways in which formal legal institutions and everyday social relations intersect and share cognitive and linguistic resources.

The Concept of Culture in General

The concept of culture as an aspect of social life has been invoked in numerous ways. Referring primarily to learned behavior as distinct from that which is given by nature or biology, culture has been used to designate everything that is humanly produced (habits, beliefs, arts, and artifacts) and passed from one generation to another. In this formulation, culture is distinguished from nature and also distinguishes one society from another. A narrower conception of culture refers to those human institutions specifically devoted to the production of signs and meanings. In this more limited usage, cultural institutions include, for example, art, music, theater, fashion, literature, religion, media, and education. While the first definition is overly broad including just about all of human life, the second is too specific: the meanings produced and circulating through other institutions and 'noncultural' spheres of life are ignored or devalued.

Contemporary cultural analyses have moved beyond these extreme conceptions of culture as either everything human or as only that which calls itself culture to conceive of culture as a system of symbols and meanings and their associated social practices. This formulation of culture treats as inseparable signs and performances, meanings, and actions. It highlights by abstracting "the meaningful aspect of human action out of the flow of concrete interactions … [by disentangling], for the purpose of analysis, the semiotic influences on action from the other sorts of influences – demographic, geographical, biological, technological, economic, and so on – that they are necessarily mixed with in any concrete sequence of behavior"(Sewell, 2005: p. 160). This attention to linguistic and symbolic aspects of social action derives from work in anthropology (Levi-Strauss, 1963; Schneider, 1976; Turner, 1969; Geertz, 1973), but rejects any notion that culture is uniform, static, or ubiquitously shared that may have been suggested in those works. Moreover, this conception of culture goes beyond a focus on language alone.

By drawing from sociological theories of action and practice (Bourdieu, 1977, 1990; Giddens, 1984), culture is conceived as an arena or field of practical activity, competition, and struggle. Culture is not only a system of signs but is performed daily in all social interactions and settings. Swidler's (1986) formulation of culture as a 'repertoire' of 'strategies of action' is widely adopted as a succinct formulation of this conception. Here, culture is not a coherent, logical, and autonomous system of symbols but a diverse collection of semiotic resources that are deployed in the performance of action. Variation and conflict concerning the meaning and use of these symbols and resources is likely and expected. Although many of these

cultural resources are discrete, local, and intended for specific purposes, nonetheless, it is possible to observe patterns so that we are able to speak of a culture or cultural system.

"System and practice are complementary concepts: each presupposes the other … The employment of a symbol," Sewell (2005: p. 164) writes, "can be expected to accomplish a particular goal only because symbols have more or less determinate meanings – meanings specified by their systematically structured relations to other symbols. Hence practice implies a system. But it is equally true that the system has no existence apart from the succession of practices that instantiate, reproduce, or – most interestingly – transform it. Hence a system implies practice. System and practice constitute an indissoluble duality or dialectic: the important theoretical question is thus not whether culture should be conceptualized as practice or as a system of symbols and meanings, but how to conceptualize the articulation of system and practice."

The Concept of Legal Culture Introduced and Debated

Writing in 1975, Lawrence Friedman introduced the concept *legal* culture as a means of emphasizing the fact that law was best understood and described as a system, a product of social forces and itself a conduit of those same forces. Although law was, and is, commonly understood as "a set of rules or norms, written or unwritten, about right and wrong behavior, duties and rights," (1975: p. 2) according to Friedman, this conventional notion attributed too much independence and efficacy to the law on the books and acknowledged too little the power and predictability of law in action. To advance a social scientific study of law in action, Friedman adopted the model of a system – a set of structures that process inputs (demands and resources) from an environment to which it sends its outputs (functions) in an ongoing recursive feedback loop (e.g., *see* Luhmann, Niklas (1927–98)). He identified three central components of the legal system: (1) the social and legal forces that, in some way, press in and make 'the law'; (2) 'the law' itself – structures and rules; and (3) the impact of law on behavior in the outside world. "Where the law comes from and what it accomplishes – the first and third terms – are essentially the *social* study of law," he wrote (Friedman, 1975: p. 3). This relatively simple image spawned a robust scholarly community, in part because of the nuance and complexity left unsaid.

Although the social scientific study of law was, with minor but notable exceptions, marginalized in American law schools and social science departments, Friedman drew from a tradition with strong European roots. In jurisprudence, for example, Friedrich Karl von Savigny (1975[1831]), a nineteenth-century German statesman and jurist had described law as one of the most important expressions of the 'spirit of a people' (Volksgeist), a continuous thread in an evolving culture. Oliver Wendell Holmes (1881), the American judge and jurist writing from a very different, distinctly realist and pragmatic rather than romantic perspective, expressed a similar conviction that the law was a great anthropological document. Such notions of the cultural character and significance of law challenged positivist jurisprudence that defined law as an autonomous system of officially sanctioned and logically derived rules and procedures (e.g., *see* Legal Positivism).

Classical works of European sociology also placed law at the center of social life rather than at the margins. Emile Durkheim (1933[1902]), for example, argued that law had become an embodiment of the 'collective conscience' in societies with advanced divisions of labor, where interdependence and reciprocity prevailed. For Durkheim, "law is the example par excellence of the social fact. It is a visible symbol of all that is essentially social" (Hunt, 1978: p. 65). For Weber, modernity was characterized by the rise of formal legal and logical rationality and bureaucratic administration. This modern movement toward formal legal rationality generated tensions and contradictions, according to Weber (1947, 1954) between demands for predictability and equally valued demands for substantive justice. Both Durkheim and Weber offered similar analyses of the forms and functions of law as both expressions of broader social formations, in particular the transformations toward modernity, and as channels for developing social sensibilities, interests, and actions (*see* Durkheim, Émile (1858–1917); Weber, Max (1864–1920)).

Friedman chose the phrase legal culture to name the "social forces … constantly at work on the law," "those parts of general culture – customs, opinions, ways of doing and thinking – that bend social forces toward or away from the law" (1975: p. 15). As an analytic term, legal culture emphasized the role of taken-for-granted and familiar actions that operated on and within the interactions of the legal system and its environment. As a descriptive term, it identified a number of related phenomena: public knowledge of and attitudes toward the legal system as well as patterns of behavior with respect to the legal system. These included judgments about the law's fairness, legitimacy, and utility. To the extent that patterns of attitudes and behaviors are discernible within a population and vary from one group or state to another, it was possible, he said, to speak of the legal culture(s) of groups, organizations, or states (1975: p. 194). As an example of variations within legal cultures, Friedman distinguished the internal legal culture of professionals working *in* the system from the external legal culture of citizens interacting *with* the system. As the "ideas, values, expectations, and attitudes toward law and legal institutions, which some public or some part of the public holds," legal culture was meant to name a range of phenomena that would be, in principle, measurable (Friedman, 1997: p. 34). Although Friedman never elaborately theorized his concept of legal culture, reformulated it several times in different texts, and promoted it as an orienting concept rather than a set of quantifiable indicators, he remained convinced that the concept was useful as a way of "lining up a range of phenomena into one very general category" (1997: p. 33).

Following its introduction, researchers began using the concept of legal culture in a range of empirical projects including studies of children's knowledge of and attitudes to law (Tapp and Levine, 1974), rights consciousness among Americans (Scheingold, 1974), the practices of criminal courts (Nardulli et al., 1988; Eisenstein et al., 1988; but see Kritzer and Zemans, 1993), comparative analysis of different groups, and nation-states (Kidder and Hostetler, 1990; Sanders and Hamilton, 1992; Bierbrauer, 1994; Barzilai, 1997; French, 1998; Chanock, 2001). For those who attempt to measure variations in legal culture, the indicators include such diverse

phenomena as litigation rates and institutional infrastructures (Blankenburg, 1994, 1997); public opinion surveys (Hertogh, 2012); crime rates (Haley, 1978; Miyazawa, 1987), or assumptions about who should use the legal system and when.

Predictably, debates have arisen among researchers who have attempted to use the concept in empirical projects (Nelken, 1997, 2012). The most persistent divide and heated dispute seems to be among those who use culture as an analytic concept within a more developed sociological theory and those who view legal culture as concrete, measurable phenomena. Those who attempt to use the concept as a focus for comparative research more often move toward measurement and a more limited concept, although exceptions exist (Kurkchiyan, 2012). For some of these researchers, when the concept of legal culture is used with insufficient specificity, the distinction between all of culture and legal culture is unclear, and what constitutes the legal seems too often assumed and undefined (Blankenburg, 1997). Some researchers insist that legal culture is that which is produced and studied most effectively among professional legal actors, while others insist that such a narrow definition belies the theoretical utility of the concept of legal culture as a way of marking the inescapable connection between law and everyday life, just that feedback loop that Friedman posited in his notion of a legal system.

Cotterrell has produced one of the most sustained critiques of the concept (1997). He claims that "everything about law's institutions and conceptual character needs to be understood in relation to the social conditions which have given rise to it. In this sense law is indeed an expression of culture" (1984: p. 26). Nonetheless, Cotterrell is unwilling to accept a concept of legal culture that makes it indistinguishable from other forms of social control or normative ordering. Somewhere between a "thoroughgoing legal pluralism" (p. 42) in which "law can be distinguished from other social norms only in vague terms" and a too narrow, too simple conception of state law, Cotterrell seeks a middle ground that recognizes the cultural influences on and from law but yet retains a recognition of the distinctiveness of legal forms and doctrines. He urges a notion of legal ideology as an alternative concept that emphasizes the essential connection to legal doctrine, although this too is actively debated by other legal scholars (Nelken, 1997, 2012; Friedman, 1997). Silbey (2012) illustrates the utility of the concept of legal culture, and its association with formal legal materials rather than simply common social norms through a close reading of a particular urban practice: placing space holders in shoveled out parking spots on snowy city streets. This practice is performed and understood by city dwellers only because it draws upon and is debated within widely circulating interpretations of historic and contemporary concepts found in legal doctrine.

To some extent, the debates among socio-legal scholars reproduce the controversies among anthropologists and sociologists concerning the concept of culture generally. The most important issues are less empirical than theoretical because the measurement problems that fuel debate derive from the theoretical arguments: how is legal culture evident and measurable and yet diffuse and abstract; what is the relative importance of causal explanation as against description and interpretive understanding in producing a social scientific study of law; how central is formal legal doctrine in understanding participation in, support for, and consequences of law. Contemporary research on legal consciousness has dealt directly with these questions.

Legal Consciousness

If research on legal culture focuses attention on the myriad ways in which law exists within society generally, not just in formal legal institutions and the work of official legal actors, the study of legal consciousness traces the ways in which law is experienced and interpreted by specific individuals as they engage, avoid, or resist the law and legal meanings. This empirical attention to popular understandings of law reformulates some of the theoretical debates expressed in the study of legal culture.

Legal Consciousness as Epiphenomena and Structural Residues

Some scholars regard legal consciousness as a by-product of the operations of social structures rather than the formative agent in shaping structures and history. Structural anthropology, for instance, describes social actors located in complex webs of patterned social relationships that determine their perceptions and actions. Similarly, some Marxist structuralism treats ideas, including cultural symbols and narratives, as a superstructural residue of material conditions that serve the interests of elites. Following this perspective, legal consciousness is an epiphenomenon because a particular social and economic structure is understood to produce a corresponding legal order with appropriately determined legal subjects.

In most structural accounts, individuals seem to lack agency, portrayed as the bearers rather than authors of social relations. Work in this tradition often describes how the needs of capitalist production and reproduction mold legal behavior and consciousness. Studies focus on the production and practice of law, its accommodation to class interests, and the inequities that result. Some research from this structural perspective complicates the proposed relationship between law and structure by suggesting that the legal order develops in response to conflicts and inconsistencies generated by the capitalist mode of production rather than as a direct instrument of particular class interests. "To legitimize the inconsistencies and irrationalities born of the contradictions of the economy, the legal order constitutes myths, creates institutions of repression, and tries to harmonize exploitation with freedom, expropriation with choice, inherently unequal contractual agreements with an ideology of free will" (Chambliss and Seidman, 1982: p. 70).

An alternative view within the structural tradition looks at legal consciousness as one of the ways in which organizations and institutions produce the means of authorizing, sustaining, and reproducing themselves. By focusing on the legitimating functions of law, research describes the ways in which law helps people see their worlds, private and public, as both natural and right. Balbus (1973), for example, argues that certain features of liberal law, such as the highly prized claims to formal equality and what will be described below as procedural justice, serve to buttress and legitimate the

inequality of the existing economic order. The formal equality instantiated in due process rights provides "a stable and apparently neutral framework from which bourgeois class interests in accumulation and profit maximization can flourish" because due process and formal equality help convince "the 'propertyless' that they have the legal right and, hence, the real opportunity of rising into the bourgeoisie" (Balbus, 1973: p. 6). Some of this literature has focused on 'false consciousness,' or the inability of subjects, especially less educated members of the working class, to perceive their true interests or recognize opposing interests. This body of research harkens back to the contradictions in modern law, mentioned above and identified by Weber, Durkheim, and Marx. With subtle variations, these structural accounts cast law and legal consciousness as products rather than producers of social relations (e.g., *see* Structuralism; Marxist Theory in Law).

However, there are some structural and Marxist theorists who do not treat cultural symbols as superstructural residues and persons without agency. Althusser (1971), for example, suggested that cultural symbols, or ideologies, are themselves material practices and are relatively autonomous; hence culture and ideology are more than just residues. In this framework, the distinction between structure and consciousness is overdrawn. Rather than by-products of material practices, culture and consciousness actively participate in the production of material practices and social realities. During the past three or four decades, a burgeoning community of scholars working within the structural tradition "have carved out a conception of culture as a realm of creativity that escapes from the otherwise pervasive determination of social action by economic or social structures" (Sewell, 2005: p. 159). E.P. Thompson's account of the creation and enforcement of the 1723 Black Act introducing capital punishments for a variety of offenses in which feudal tenants and propertyless classes challenged the closing of English common lands, *Whigs and Hunters*, depicts the rule of law emerging from bitter struggles over access to hunting, contested common rights, as well as high crime and espionage. Rather than a simple mask for the rule of a particular class, Thompson describes the law emerging from the active struggles between the opposing sides. [It is not] "to say that the rulers had need of law, in order to oppress the ruled, while those who were ruled had need of none. What was often at issue was not property, supported by law, against no-property; it was alternative definitions of property-rights: for the landowner, enclosure – for the cottager, common rights; for the forest officialdom, 'preserved grounds' for the deer; for the foresters, the right to take turfs. For as long as it remained possible, the ruled – if they could find a purse and a lawyer – would actually fight for their rights by means of law; occasionally the copyholders, resting upon the precedents of sixteenth century law, could actually win a case. When it ceased to be possible to continue to fight at law, men still felt a sense of legal wrong: the propertied had obtained their power by illegitimate means" (Thompson, 1975: p. 261).

Legal Consciousness as Individual Attitudes and Opinions

In another common formulation, the role of individual agency is centrally assumed, if not always explicitly articulated and theorized. Here, scholars conceptualize consciousness as the ideas and attitudes of individuals that, when taken together, determine the form and texture of social life. For the most part, these studies are not distinctly different from studies of legal culture relying on surveys of individual attitudes, opinions, and actions. A legacy of the liberal tradition in political and legal theory, this conception of consciousness (and culture) suggests that social groups of all sizes and types (e.g., families, peer groups, corporations, and communities) emerge out of the aggregated actions of individuals but do not specify the mechanisms (processes or structures) of aggregation. According to this approach, sometimes referred to as methodological individualism, "political society is … an association of self determining individuals who concert their will and collect their power in the state for mutually self-interested ends" (Wolff, 1969: p. 5). Here, individual consciousness consists of both reason and desire. According to liberal ideology, desire, which also remains unexplained for the most part, is "the moving, active, or primary part of the self … What distinguishes men [sic] from one another is not that they understand the world differently, but that they desire different things, even when they share the same understanding" (Unger 1981[1975]: pp. 39,40). In common parlance, this is often referred to as expressions of different values, or in economics utilities, those values assumed to be innate differences, nature not nurture.

Relying on this attitudinal conception of legal consciousness, researchers have attempted to document the variation in beliefs, attitudes, and actions of individuals as a means of explaining the shape of legal institutions and practices, assessing the degree to which the rule of law is part of the socialization and habits of a population (Stalans and Kinsey, 1994), and predicting the emergence of stable liberal legal regimes within states undergoing transitions from authoritarian or colonial rule (e.g., Gibson and Gouws, 1997). Lind and Tyler (1988), for example, have documented Americans' attachment to ideas of fairness and due process and to what they call procedural justice (cf Thibaut and Walker, 1975). The research suggests that people commonly evaluate their legal experiences in terms of the processes and forms of interaction rather than the outcomes of those interactions. Attitudes about the law correlate strongly with judgments about the fairness of the procedures used by legal authorities rather than with substantive principles, winning or losing in that process. Lind and Tyler describe deep, broad-based normative consensus among Americans (e.g., *see* Compliance and Legal Authority).

In contrast, a study of legal attitudes among Europeans observes considerable variability in citizens' commitments to, and alienation from, law and legal values rather than consensus (Gibson and Caldeira, 1997). For example, respondents in Greece report high alienation from legal institutions, limited support for the rule of law, and a slight tendency to value order over liberty, while Danish respondents report very low alienation and strong support for liberty and the rule of law. Although Gibson and Caldeira find international variation among the European states, they also find strong variation within nation-states. Studies in nations transitioning from communist or other forms of authoritarian rule to democracy also report notable intra-nation as well as international variation, some of which is attributed to different stages of cognitive development, variations in past experience

and familiarity with legal processes, and generally with national cultures and styles (Kourilsky-Augeven, 1997).

Constitutive Theories of Legal Culture and Consciousness

Constitutive analyses work to resolve the debates concerning determinism, structure, and agency in studies of culture and legal consciousness. Research from a constitutive perspective emphasizes the role of consciousness or cultural practice as communicating factors between individual agency and social structure rather than expressions of one or the other. Defining legal consciousness as participation in the processes of social construction, research has focused on the ways in which local concrete action accumulates into systemic institutions and structures (Ewick and Silbey, 1998). In other words, research adopting a constitutive perspective works to show how the signs, symbols, and invocations relating to law form a mutually reinforcing system so that the articulated opinions or interpretations make sense to speakers and audience by reference to commonly circulating meanings and relations to other signs – the system or structure of culture. Similarly, the legal culture or system of signs does not exist independently of these individual invocations and articulated opinions and performed actions. "In legal consciousness research, the push and pull of legal ideas lie at the heart of modern explanations of the texture of law in our everyday existence. On one side is the pull of law in constructing and constraining individual action and decisions. The law imposes constraints in the form of elaborate regulations, codes … On the other side, individuals' own interpretations of law provide the push, shaping the new versions of legality. In this approach, law is not passive, but rather a dynamic force. By enacting legality in daily life, ordinary people give flesh and meaning to what is otherwise an 'abstract but binding form'" (Marshall and Barclay, 2003: pp. 617–618, citing Habermas, 1996: p. 318). "By inviting legal scholarship to focus on everyday life, rather than on legal doctrine," scholars "seek to bring into view, if not give primacy to, the lively normative resources of the everyday. These, no doubt, are resources powerfully shaped by law; but they are also resources on which law itself deeply depends" (Sarat and Kearns, 1993: p. 56).

Working within a Gramscian framework (e.g., *see* Gramsci, Antonio), these analysts describe how the taken-for-granted aspects of social relations, including the legal aspects, are produced, enacted, and reproduced. They document situations in which local processes recursively reproduce macro social structures and institutions and, at the same time, provide openings for creativity in reshaping those structures (e.g., Hebdige, 1979; Hall and Jefferson, 1975). In particular these constitutive/cultural analyses of legal consciousness describe the processes by which law contributes to the articulation of meanings and values in daily life (Henry, 1983). Attention is directed to the local contests over signification within diverse and competing discourses that take place in most aspects of ordinary life including families (Merry, 1990; Yngvesson, 1993), religious communities (Greenhouse, 1986), medical care (Heimer, 1999), engineering (Espeland, 1998), gender, race, or age discrimination (Bumiller, 1988), managing handicapped identities (Engel, 1991; Engel and Munger, 2003), or poverty and public welfare (Sarat, 1990; Munger, 1999) (e.g., *see* Law and Everyday Life). In these analyses, researchers observe both the orchestration of the local contest and the systematic outcome, in this way mediating micro and macro perspectives and identifying the mechanisms for aggregating individual actions and preferences. Within this framework, consciousness is understood to be part of a reciprocal process in which the meanings given by individuals to their world become patterned, stabilized, and objectified. These meanings, once institutionalized, become part of the material and discursive systems that limit and constrain future meaning-making. In this constitutive conception of legal consciousness, law does more than reflect or encode what is otherwise normatively constructed. Law enables, as well as constrains the possibilities of social interaction. "The commonplace operation of law in daily life makes us all legal agents insofar as we actively make law, even when no legal agent is involved" (Ewick and Silbey, 1998: p. 20).

In one study, Conley and O'Barr observed litigants in small claims courts to determine how ordinary "people identify and analyze legal problems, how they decide when and in what form to bring a problem to the legal system, and how they respond to the demands that the system makes of them" (1990: p. ix). The authors observed two distinct discursive styles. One style is rule oriented, adopted primarily by experienced players who see law as system for assessing and distributing responsibility; this style is valued by the judges and court personnel. The second style focuses on and emphasizes relationships, sequences of events, and the temporal unfolding of a narrative; this style is preferred by minority and group litigants as well as women and poorer citizens. Although having one's day in court is an American ritual, highly prized as an essential human right, many citizens end up frustrated, the researchers claim, because legal authorities prefer and often demand something other than a heart-rending narration of pain or an impassioned moral claim.

In a study of the legal consciousness of ordinary Americans, Ewick and Silbey (1998) synthesize diverse results of earlier studies with new data. They construct three accounts of law that encompass the range of cultural materials with which people produce and experience legality as a structure of social action. Drawing upon different cultural images (e.g., a bureaucracy, a game, pragmatic coping strategies such as 'making do'), each story describes a familiar way of acting and thinking with respect to the law. The stories incorporate alternative normative bases for legal authority, varying constraints that define action as legal, varying resources for legal capacities, and varying temporal and spatial locations for law. In one account, the law is remote, impartial, and objective, something to be invoked for solemn and collective purposes that transcend the messiness and partiality of individual lives. The law itself resides in times and spaces separate from everyday places and, while enacted by legal functionaries, exists apart from the words or deeds of particular persons. In a second account, legality is understood to be a game of skill, resource, and negotiation where persons can legitimately seek their own interests against others. Law in this rendering appears as a defined arena for strategic interactions, sometimes

engaged playfully and sometimes with dead seriousness, but always simultaneously alongside and within everyday life. A third story describes the law as an arbitrary power against which people feel virtually incapacitated. Often the only way to deflect legal power is to employ various subterfuges and evasions. These minor forms of resistance typically leave the law unchallenged and unchanged. Their employment and effectiveness is premised, however, upon a potentially subversive recognition of the structure and organization of legality in everyday life.

Ewick and Silbey demonstrate the connection between the micro phenomenon of legal consciousness and the macro institutions of law by showing how the multiple forms of legal consciousness expressed in the three stories of law, together as an ensemble, sustain legality as a durable structure of social action. Although the dual depictions of law as God-like (remote, transcendent, objective and magisterial) and game-like (legitimate, rule-bound, and resource-dependent pursuit of self-interest) seem to challenge one another, they are actually complementary. Although each view emphasizes different normative values and provides a different account of the social organization of law, as an ensemble they cover the range of conventional experiences of legality. Any particular experience can find expression within the heterogeneity of the whole. The law is neither rendered irrelevant for everyday life (by virtue of being remote) nor subsumed by it (by virtue of being familiar).

The schemas Ewick and Silbey identified organized the varieties of legal consciousness that had been described in other research. They confirmed the observations of deep normative consensus with regard to norms of procedural fairness (e.g., Lind and Tyler, 1988) while also observing tactical, sometimes cynical, sometimes earnest, employment of law's devices for a variety of personal or organizational interests that so much of the literature describes. They also documented the resistance to legal processes that Bumiller (1988), Greenhouse (1986), Sarat (1990), had been describing in a wide range of social fields and actions. The combination of extensive data collection (for generalizability) with intensive conversations (for rich textual accounts) was able to reproduce many of the results of the existing literature, at the same time developing a theoretical synthesis. More recent studies have challenged some of the methods (Garcia-Villegas, 2003), but have also confirmed the description of legality as multiply stranded cultural schemas (Cowan, 2004; Fleury-Steiner, 2003, 2004; Hoffman, 2003; Hull, 2003; Kostiner, 2003; Marshall, 2003, 2005; Nielsen, 2000, 2004; Pelisse, 2004; Quinn, 2000; Sagay, 2003; Steiner, 2001).

These constitutive studies reject the dualisms in recurrent debates about the relative roles of structure and agency in shaping the world and focus instead on the role of consciousness (participation in social interaction) in (re) producing the social world (e.g., *see* Law as Constitutive). Moreover, these constitutive analyses of culture and consciousness specify how myriad discrete actions accumulate into institutional forms through processes of reification in which the contingent human production of social life is effaced by attending to its continuous, durable features. This attention to systematic aspects makes the social interactions and practices appear to be independent of just that human action of which it is composed. Thus, studies of legal culture and consciousness try to emphasize the daily, locally repeated enactment or practice of social structure. Some cultural accounts, however, go so far as to reject explicitly any distinction between structure and agency, insisting that the analytic distinction itself contributes to the ongoing processes of reification constituting social reality. Finally, in these constitutive studies of legal culture and consciousness, law is understood to be part of a complex totality in which "it constitutes as well is constituted, shapes as well as is shaped" (Kairys, 1982: p. 5).

Recent Advances in Research on Legal Consciousness

Contemporary studies of legal consciousness trace the understandings of law that circulate through social relations attempting to show how consciousness is "shaped by class, education, geography, and occupational position" (Cooper, 1995: p. 510). Researchers attempt to map the variations in legal consciousness by associating distinct interpretations and actions with different sociological and demographic markers or organizational locations. The legal consciousness of people who share social characteristics of identity or experience is expected to influence their interpretations of institutions and processes including the law (Marshall and Barclay, 2003, Nielsen, 2004; Engel and Munger, 2003). Almost all the research describes the double bind of experiencing constraints in the attempt to enact human agency; the studies specifically emphasize the frustrations and contradictions of empowering oneself by deploying the law's powers. Although dozens of studies confirm that the legal consciousness of ordinary citizens is not a unitary phenomenon, few studies have been able to produce compelling associations between legal consciousness and "particular types of laws, particular social hierarchies, and the experiences of different groups with the law" (Nielsen, 2000: pp. 1055, 1086). Silbey (2005) identifies the limits of the research agenda in misunderstandings of contemporary popular cultural production, an overreliance on surface indicators, and a confusion between theoretical concepts and empirical indicators.

Research seeking to capture the authentic voices of ordinary people belies the suffusion of everyday life by professional, marketed cultural production. Poll results have become a form of politically manipulated cultural representation; rather than a snapshot of the public at any moment in time, public opinion polling has become a means of representing the public (to the public) as supportive of the interests and policies of particular elites. Importantly, because legal culture and consciousness were developed as concepts to analyze the law without using the law's own assumptions and language, we need to recognize that the ability to seduce the public through mediated messages has been a direct product of social scientific knowledge about human desire and cognition. In the effort to construct an account of the ways in which law suffuses and perhaps dominates everyday life, it would be foolish to deny real, experiential, material differences in social spaces and lives, extraordinary variations in life expectations and conditions. Nonetheless, the use of the concepts of legal culture and consciousness are used by scholars not to document whether the conditions of our lives

vary, but whether the cultural terms with which we understand and communicate, and with which we constitute our lives, can be correlated with legal concepts and doctrines, as well as social locations.

See also: Civil Liberties and Human Rights; Cultural Psychology; Cultural Sociology, History of; Culture: Contemporary Views; Disputes, Social Construction and Transformation of; Human Rights: Legal Aspects; Law and Everyday Life; Law and Society: Development of the Field; Law and Society; Law as Constitutive; Law, Sociology of; Law: Anthropological Aspects; Law: The Socio-Legal Perspective; Legal Process and Social Science: United States; Norms; Public Opinion: Social Attitudes; Social Structure: History of the Concept; Social Structure; Subculture, Sociology of.

Bibliography

Althusser, L., 1971. Ideology and ideological state apparatuses. In: Lenin and Philosophy. New Left Books, London.

Balbus, I., 1973. The Dialectics of Legal Repression: Black Rebels before the American Criminal Courts. Russell Sage Foundation, New York.

Barzilai, G., 1997. Between the rule of law and the laws of the ruler: the supreme court in Israeli legal culture. International Social Science Journal 49 (152), 143–150, 2.

Bertaux, D., 1997. Les Facteurs de Variation de la Conscience Individuelle du Droit. In: Kourilsky–Augeven, C. (Ed.), Socialisation Juridique et Conscience du Droi Reseau Europeen Droit et Societe, vol. 2, pp. 141–146.

Bierbrauer, G., 1994. Toward an understanding of legal culture: variations in individualism and collectivism between Kurds, Lebanese, and Germans. Law and Society Review 28 (2), 243–264.

Blankenburg, E., 1994. The infrastructure of legal culture in Holland and West Germany. Law and Society Review 28 (4), 789–809.

Blankenburg, E., 1997. Civil litigation rates as indicators for legal cultures. In: Nelken, D. (Ed.), Comparing Legal Cultures. Dartmouth Publishing Company, Brookfield, VT, pp. 41–68.

Bourdieu, P., 1977. Outline of a Theory of Practice (R. Nice, Trans.). Cambridge University Press, New York.

Bourdieu, P., 1990. The Logic of Practice (R. Nice, Trans.). Stanford University Press, Stanford, CA.

Bumiller, K., 1988. The Civil Rights Society: The Social Construction of Victims. Johns Hopkins University Press, Baltimore.

Chanock, M., 2001. The Making of South African Legal Culture 1902–1936: Fear, Favour and Prejudice. Cambridge University Press, Cambridge.

Chambliss, W., Seidman, R., 1982. Law, Order, and Power. Addison-Wesley, Reading.

Conley, J.M., O'Barr, W.M., 1990. Rules versus Relationships: The Ethnography of Legal Discourse. University of Chicago Press, Chicago.

Cooper, D., 1995. Local government legal consciousness in the shadow of juridification. Journal of Law and Society 22 (4), 506–526.

Cowan, D., 2004. Legal consciousness: some observations. Modern Law Review 67 (6), 928–958.

Cotterrell, R., 1992[1984]. The Sociology of Law: An Introduction, second ed. Butterworth, London.

Cotterrell, R., 1997. The concept of legal culture. In: Nelken, D. (Ed.), Comparing Legal Cultures. Dartmouth Publishing Company, Brookfield, VT, pp. 13–32.

Durkheim, E., 1933[1902]. The Division of Labor in Society. Macmillan, New York.

Eisenstein, J., Flemming, R.B., Nardulli, P.F., 1988. The Contours of Justice: Communities and Their Courts. Little, Brown, Boston.

Engel, D.M., 1991. Law, culture, and children with disabilities: educational rights and; the construction of difference. Duke Law Journal 1, 166–205.

Engel, D.M., Munger, F.W., 2003. Rights of Inclusion: Law and Identity in the Life Stories of Americans with Disabilities. University of Chicago Press, Chicago.

Espeland, W.N., 1998. The Struggle for Water: Rationality, and Identity in the American; South West. University of Chicago Press, Chicago.

Ewick, P., Silbey, S., 1998. The Common Place of Law: Stories from Everyday Life. University of Chicago Press, Chicago.

Fleury-Steiner, B., 2003. Before or against the law? Citizens' legal beliefs and expectations as death penalty jurors. Studies in Law, Politics and Society 27, 115–137.

Fleury-Steiner, B., 2004. Jurors' Stories of Death. University of Michigan Press, Ann Arbor.

French, J.D., 1998. Drowning in laws but starving (for justice?): Brazilian labor law and the workers' quest to realize the imaginary. Political Power and Social Theory 12, 181–218.

Friedman, L.M., 1975. The Legal System: A Social Science Perspective. Russell Sage Foundation, New York.

Friedman, L.M., 1997. The concept of legal culture: a reply. In: Nelken, D. (Ed.), Comparing; Legal Cultures. Dartmouth Publishing Company, Brookfield, VT, pp. 33–40.

Garcia-Villegas, M., 2003. Symbolic power without symbolic violence? Critical comments on legal consciousness studies in USA. Droit et Société 53, 137–162.

Geertz, C., 1973. The Interpretation of Cultures. Basic, New York.

Geertz, C., 1983. Local Knowledge: Further Essays in Interpretive Anthropology. Basic Books, New York.

Gibson, J.L., Gouws, A., 1997. Support for the rule of law in the emerging South African democracy. International Social Science Journal 49 (152), 173–191.

Gibson, J.L., Caldeira, G.A., 1997. The legal cultures of Europe. Law and Society Review 30 (1), 55–85.

Giddens, A., 1984. The Constitution of Society: Outline of the Theory of Structuration. University of California Press, Berkeley & Los Angeles, CA.

Gilliom, J., 2001. Overseers of the Poor. University of Chicago Press, Chicago.

Greenhouse, C.J., 1986. Praying for Justice: Faith, Order, and Community in an American Town. Cornell University Press, Ithaca, New York.

Habermas, J., 1996. Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy (William Rehg, Trans.). MIT Press, Cambridge, MA.

Haley, J.O., 1978. The myth of the reluctant litigant. Journal of Japanese Studies 4, 359–390.

Hall, S., Jefferson, T., 1975. Resistance through Rituals: Youth Subcultures in Post-war Britain. Macmillan, London.

Hebdige, D., 1979. Subculture, the Meaning of Style. Methuen, London & New York.

Heimer, C., Stassen, L., 1998. For the Sake of the Children. University of Chicago Press, Chicago.

Heimer, C., 1999. Competing institutions: law, medicience and family in neo-natal intensive care. Law & Society Review 33 (1), 17–66.

Henry, S., 1983. Private Justice. Routledge & Kegan Paul, London.

Hertogh, M., 2012. The curious case of Dutch legal culture: a reassessment of survey evidence. Journal of Comparative Law 2, 146–168.

Hoffman, E.A., 2003. Legal consciousness and dispute resolution: different disputing behavior at two similar taxicab companies. Law & Social Inquiry 28, 691–718.

Hull, K.E., 2003. The cultural power of law and the cultural enactment of legality: the case of same sex marriage. Law & Social Inquiry 28 (3), 629–658.

Holmes, O.W., 1881. The Common Law. Little, Brown, and Company, Boston.

Hunt, A., 1978. The Sociological Movement in Law. Macmillan, London.

Kairys, D. (Ed.), 1990[1982]. The Politics of Law. Pantheon Books, New York.

Kennedy, D., 1980. Toward an Historical Understanding of Legal Consciousness.

Kidder, R.L., Hostetler, J.A., 1990. Managing ideologies: harmony as ideology in Amish and Japanese societies. Law and Society Review 24, 895–922.

Kostiner, I., 2003. Evaluating legality: toward a cultural approach to the study of law and social change. Law & Society Review 37 (2), 323.

Kourilsky-Augeven, C., 1997. Socialisation Juridique et Modele Culturel. In: Kourilsky-Augeven, C. (Ed.), Socialisation Juridique et Conscience du Droit Reseau Europeen Droit et Societe, vol. 2, pp. 11–32.

Kritzer, H.M., Zemans, F.K., 1993. Local legal culture and the control of litigation. Law and Society Review 27 (3), 535–557.

Kurkchiyan, M., 2012. Comparing legal cultures: three models of court for small civil cases. In: Nelken, D. (Ed.), Using Legal Culture. Simmonds & Hill Publishing, London, pp. 218–250.

Levi-Strauss, C., 1963. Structural Anthropology. Basic, New York.

Lind, A.E., Tyler, T.R., 1988. The Social Psychology of Procedural Justice. Plenum Press, New York.

Marshall, A., 2003. Injustice frames, legality, and the everyday construction of sexual harassment. Law & Social Inquiry 28 (3), 659–690.

Marshall, A., 2005. Idle rights: employees' rights consciousness and the construction of sexual harassment policies. Law & Society Review 39 (1), 83–124.

Marshall, A.M., Barclay, S., 2003. Their own words: how ordinary people construct the legal world. In Law and Social Inquiry 28 (3), 617–628.

Merry, S.E., 1990. Getting Justice and Getting Even: Legal Consciousness among Working-Class Americans. University of Chicago Press, Chicago.

Miyazawa, S., 1987. Taking Kawashima seriously: a review of Japanese research on Japanese legal consciousness and disputing behavior. Law and Society Review 21, 219–241.
Munger, F., 1999. Immanence and identity: understanding poverty through; law and society research. Law & Society Review 32 (4), 931–967.
Nardulli, P.F., Eisenstein, J., Flemming, R.B., 1988. The Tenor of Justice: Criminal Courts and the Guilty Plea Process. University of Illinois Press, Urbana.
Nelken, D., 1997. Comparing Legal Cultures. Dartmouth Publishing Company, Brookfield, VT.
Nelken, D., 2012. Using Legal Culture. Wildy, Simmonds & Hill Publishing, London.
Nielsen, L.B., 2000. Situating legal consciousness: experiences and attitudes of ordinary citizens about law and street harassment. Law & Society Review 34 (4), 1055–1090.
Nielsen, L.B., 2004. License to Harass: Law, Hierarchy and Offensive Public Speech. Princeton University of Press, Pinceton, NJ.
Pelisse, J., 2004. From negotiation to implementation: a study of the reduction of working time in France (1998–2000). Time Society 13 (2/3), 221–244.
Quinn, B.A., 2000. The paradox of complaining: law, humor, and harassment in the everyday work world. Law & Social Inquiry 25, 1151–1175.
Sagay, A.C., 2003. What is Sexual Harassment: From Capital Hill to the Sorbonne. Univ. Calif. Press, Berkeley.
Sanders, J., Hamilton, V.L., 1992. Legal cultures and punishment repertories in Japan, Russia, and the United States. Law and Society Review 26 (1), 117–138.
Sarat, A., 1990. "…The law is all over": power, resistance, and the legal consciousness of the welfare poor. Yale Journal of Law and Humanities 2 (2), 343–379.
Sarat, A., Kearns, T., 1993. Law in Everyday Life. University of Michigan, Press, Ann Arbor.
von Savigny, F.K., 1975[1831]. Of the Vocation of Our Age for Legislation and Jurisprudence (A. Hayward, Trans.). Arno Press, NY.
Scheingold, S.A., 1974. The Politics of Rights: Lawyers, Public Policy, and Political Change. Yale University Press, New Haven, CT.
Schneider, D.M., 1976. Notes toward a theory of culture. In: Basso, K.H., Selby, H.A. (Eds.), Meaning in Anthropology. University of New Mexico; Press, Albuquerque, NM, pp. 197–220.
Sewell Jr., W.H., 2005. The concept(s) of culture. In: The Logics of History. University of Chicago Press, Chicago, pp. 152–174.
Silbey, S.S., December 2005. After legal consciousnss. Annual Review of Law and Social Science 1, 323–368.
Silbey, S.S., 2012. J. Locke, op. cit.: invocations of law on snowy streets. In: Nelken, D. (Ed.), Using Legal Culture. Wildy, Simmonds & Hill Publishing, London, pp. 120–152.
Stalans, L.J., Kinsey, K.A., 1994. Self-presentation and legal socialization in society: available messages about personal tax audits. Law and Society Review 28 (4), 859.
Steiner, B., 2001. The consciousness of crime and punishment: reflections on identity politics and lawmaking in the war on drugs. Studies in Law, Politics and Society 23, 185–212.
Swidler, A., 1986. Culture and action: symbols and strategies. American Sociological Review 51, 273–286.
Tapp, J.L., Levine, F.L., 1974. Legal socialization: strategies for an ethical legality. Stanford Law Review 27, 1–72.
Thibaut, J., Walker, L., 1975. Procedural Justice: A Psychological Analyis. L. Erlbaum Associates, New York.
Thompson, E.P., 1975. Whigs and Hunters. Random House, New York.
Turner, V.W., 1969. The Ritual Process: Structure and Anti-Structure. Aldine Publishing, Chicago.
Unger, R.M., 1981[1975]. Knowledge and Politics. Free Press, NY.
Weber, M., 1947. The Theory of Social and Economic Organization (A.M. Henderson, T. Parsons, Trans.). Oxford University Press, NY.
Weber, M., 1954. On Law in Economy and Society (M. Rheinstein, Trans.). Simon and Schuster, NY.
Williams, R., 1983[1976]. Keywords: A Vocabulary of Culture and Society, rev. ed. Oxford University Press, NY.
Wolff, R.P., 1969. Beyond tolerance. In: Wolff, R.P., Moore Jr., B., Marcuse, H. (Eds.), A Critique of Pure Tolerance. Beacon Press, Boston, pp. 3–52.
Yngvesson, B., 1993. Virtuous Citizens, Disruptive Subjects: Order and Complaint in a New England Court. Routledge, NY.

Legal Education

Kim Economides, Flinders Law School, Flinders University, Adelaide, SA, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines a specific form of education designed to serve both professional formation and liberal education, namely legal education. There are five sections: the first section discusses the origin and historical evolution of legal education; the next section describes the unique goals of legal education and competing interests of different stakeholders; the subsequent section presents an analysis of formative influences and drivers of change common to modern legal professions; this is followed by a section on educational continuum covering an explanation of the different phases of legal education; and the article concludes with the latest developments and influences shaping the future, namely managerialism, audit, diversity, and access.

Legal education, in any systematic sense, is a relatively new phenomenon in most jurisdictions. In ancient Greece and Rome there was some formal training in advocacy (moral and legal) supporting the establishment of a professional class of person with legal expertise but this class was not a profession as such, and early forms of legal education were delivered primarily through observation and training acquired by serving as an apprentice to a legal consultant (*jurisprudent*) rather than through formal academic study. Legal texts were for many centuries quite rare that meant that the art of rhetoric inevitably formed the 'core' of early legal education. Gradually, however, as texts on both law and custom became more common, and records started to be kept of court decisions, this dominant oral tradition was supplanted with a new written tradition created by full-time law teachers dedicated to establishing and communicating a literature on, about, and for law. Whether this literature was written to serve the goals of the academy or of the professional bodies has never been clear and to the present day the ultimate goals of legal education have been strongly contested by its various stakeholders.

The Establishment of Organized Legal Education

Legal education may have started out in classical times by asking universal questions linking legal rules, morals, and justice and only much later were these elements separated, with some discarded altogether. Very early legalists in the ancient world were certainly not parochial in their outlook and did not draw sharp distinctions between law and morality, nor did they fail to consider the relationship with other professions and disciplines. If one draws a distinction between legal *education* and *training* it could be said that with the passage of time the latter has steadily displaced the former with the result that, until recently, the boundaries of legal thought have been shrinking and the legal mind has become more introverted and absorbed with technical ephemeral rules rather than fundamental values. Roman law, canon law, and international law – the *ius commune* – was the focal point of legal study in the medieval universities on the continent of Europe and these legal sources formed the basis of a common transnational legal language facilitating trade between commercial classes. The English common law was an exception to this trend and although to some extent influenced by this *ius commune*, its impact in England was greatly reduced by the early development of an indigenous and sophisticated legal culture relatively insulated from Roman law.

The subsequent focus on national legal systems correlates with the rise of nation-states and liberal legal professions. Legal education in medieval England was provided by the Inns of Court that organized highly vocational apprenticeships involving both shadowing practitioners and practical exercises – and of course dining – within a closed, and almost monastic, community (Brand, 1999; Baker, 1990). The historic impetus for legal education therefore came from a divided profession that bequeathed a persistent legacy differentiating English legal education from university-based legal education found elsewhere. With the advent of printing, legal texts started to supplant, but not replace, these informal methods of socialization. University education in English law did not commence until Sir William Blackstone started to give his famous lectures at Oxford in the 1750s, but these were not so much for fledgling lawyers following a degree course as stand-alone lectures for would-be justices of the peace and gentlemen. It was not until the nineteenth century, with pioneers such as John Austin, Jeremy Bentham, and Sir Henry Maine, that university legal education was established at University College London and beyond, including the Council of Legal Education that was set up in 1852 (see Cocks, 1983, 1988). The Society of Legal Scholars, along with other bodies, further promoted the resources and status of law teaching and law teachers throughout the twentieth century in the United Kingdom and Ireland (Cownie and Cocks, 2009), while in the United States legal education is generally seen to have begun with the foundation of Harvard Law School in 1817, with subsequent development assisted and monitored by the Association of American Law Schools (Sutherland, 1967; Stevens, 1983). In England a golden age in which legal scholarship seriously engaged with larger issues of politics, policy, history, and society, such as found north of the border with the Scottish Enlightenment, had to wait for the great reformer Jeremy Bentham and, a century later, a scholarly and US-influenced 'renaissance' – associated with interdisciplinary movements such as law and society, critical legal studies, feminist legal studies – that ultimately claimed victory with the contested assertion "we are all realists and contextualists now" (Sugarman, 2011). At the

start of the twenty-first century, aided by the Internet and occupying virtual space, new supranational associations have emerged alongside more established regional bodies such as the Australasian Law Teachers Association (1977) and the European Law Faculties Association (1995): today the International Association of Law Schools (2005) and the International Association of Legal Ethics (2010) also promote legal education and scholarship across regional and global stages and over time could supersede the role of national bodies that to date have concentrated on more parochial interests of academic lawyers and practitioners. It is possible, following these latest institutional initiatives, we shall witness a return to the original, more universal questions that prevailed at the inception of legal education.

The Aims of Legal Education

Toward the close of the last century, several national reviews of legal education were set up to report on the state of legal education (Ormrod, 1971; Arthurs, 1984; Pearce, 1987; MacCrate, 1992; Franken, 1995; ACLEC, 1996). From these national reviews it is apparent that throughout the world legal education has increasingly diverse aims serving concurrently both academic and professional goals (Cownie, 2010). The precise mix balancing scholarship and practice varies between, and within, particular jurisdictions. And even within particular law schools heated debates about theory, substance, and method are found. Scholarly approaches have become pluralistic and fall into a number of subcategories: doctrinal ('black letter'); sociolegal; contextual; critical; jurisprudential; interdisciplinary ('law and …'); comparative; clinical; and skills oriented (see Twining, 1994). At the most general level almost all would agree that legal education should introduce students to what it means to "think like a lawyer" (Ormrod, 1971, p. 102). Law is an intellectual discipline with distinctive concepts, institutions, and methods of reasoning, and a primary responsibility of all law schools is to create, develop, and transmit this 'core' knowledge. In England and Wales, as elsewhere, the professional bodies take an active interest in defining this core but their involvement is often resented, or resisted, by the academy that asserts jurisdiction over the academic or 'initial' stage (see Section The 'Initial' Stage). The ACLEC Report sought to balance the interests of various stakeholders by concentrating on articulating an 'outcome statement' which specified the broad skills, knowledge, and values law students should possess by the time they graduate without dictating to law teachers precisely how their students should be taught (ACLEC, 1996, p. 72). Debates over the 'core' continue (e.g., Bradney, 2003) and a major evidence-based review of legal education and training requirements is currently under way in England and Wales, sponsored by the professional bodies but carried out by academics, and due to report in December 2012. This latest review aims "to ensure that the future system of legal education and training will be effective and efficient in preparing legal service providers to meet the needs of consumers" (LETR, 2012).

Interdisciplinary approaches, initially outside or on the periphery of the curriculum, have become increasingly part of the mainstream but the risk now is that of curriculum overload and superficial learning. Anglo-American legal education has embraced 'sociolegal studies' and the study of 'law and society,' but law schools in continental Europe have for a long time made nonlaw subjects, such as legal history, economics, philosophy, or sociology, compulsory for undergraduate law students. In the civil law, world legal education continues to be dominated by the study of codes that can be traced back to Roman law. Codes provide the focal point for formal instruction that tends to discourage or displace critical reflection but also within the common law systems doctrinal analysis still prevails although it is now generally accepted that law students should also be exposed to ideas about the context of law. Australia and New Zealand adopt yet another model which involves the majority of law students enrolling on 'double degrees' that provide more than a 'taste' of other disciplines; they actually ground law students in a wide range of other academic disciplines so that by the time of graduation they can think not only 'like a lawyer' but are skilled in, rather than knowledgeable about, reasoning associated with other sciences and humanities.

By the time of their graduation, law students are expected to have laid the foundation for a law-related career but not necessarily one within the legal profession. Although many, if no longer a majority of, law students in England and Wales still aspire to qualify as either a solicitor or barrister, this has never been true in continental Europe, the United States, or Japan where the emphasis has always been on producing lawyers who will disperse into a range of occupations in the public and private sectors both inside and outside the legal profession. Increasingly it is being recognized that law schools not only prepare students for a career in law but rather they enable people who can handle legal sources to cope with change in a wide variety of work environments which could be described as the 'legal services industry.' 'Globalization,' new regulatory frameworks and technology, and interdisciplinary tendencies common to all modern professions are transforming this industry (Flood, 2011). Legal education continues to teach students how to 'think like a lawyer,' but today this thinking requires much more than a passive understanding of primary sources, principles, and concepts and must include a critical awareness of what is happening outside the world of law. Because so much law has a built-in obsolescence, it is vital that law students possess the skills and confidence to interpret new rules and situations (Economides and Smallcombe, 1991, p. 23). University legal education must therefore, inter alia, provide a solid foundation for training at the vocational stage, and what happens subsequently during continuing professional development as well as a wider contextual framework for understanding the uncertainty that lies ahead.

Trends in Legal Education

In demographic terms university legal education has been steadily expanding in most jurisdictions after the Second World War creating a mismatch between the supply of law graduates and their demand from prospective employers, at least those from within the legal profession. Most lawyers today are middle-class university graduates although they are not

necessarily law graduates, male, or white. Methods of delivering legal education are constantly evolving but a university degree, which acts as a qualifying examination recognized by the profession, has replaced apprenticeship that for centuries was the normal entry route to a legal career. Women have become increasingly visible in law schools and the profession, and this numerical presence is also challenging traditional male-dominated attitudes toward both the law and legal work. Within common law jurisdictions it is possible to enter the profession having studied another degree, and indeed some law firms express a clear preference for those with nonlaw or mixed degrees (Economides, 1999, p. 400). The professions have always sought to control entry, partly to maintain standards and protect what they see as a finite amount of legal work that can be shared (Abel, 1999). On the other hand, because universities (and to some extent government policy on higher education) decide who enters law school – and in most countries these have been expanding in both size and number – law students continue to flood the market. In Japan, which produces over 40 000 law graduates a year, recent reforms have increased the total number of places on the professional training course to still only around 1000 places. What some may describe as the overproduction of law graduates who fail to gain full professional qualifications has also brought about new forms of legal service delivery through so-called 'paralegals,' while at the same time encouraging experimentation with new access policies opening up entry to mature students and those from poorer backgrounds and ethnic minorities.

The increase in the numbers of those entering law school has not always been matched by a corresponding increase in resources resulting in adverse staff–student ratios or an imbalance in the social and ethnic composition of the law student population. In order to cope with this expansion in numbers, new teaching methods have evolved whereby students participate in more independent, active learning and teachers are under continual pressure to explore the potential of computer-assisted learning (CAL). Distance learning in legal education remains uncommon, although is not unknown, and could have significant implications not only for citizens' access to legal education but also for citizens' access to legal services more generally. The use of simulated learning environments, particularly at the vocational stage, is also spreading and these new forms of pedagogy hold considerable potential when it comes to introducing future lawyers to the risks and realities of legal practice. Progress with the notion of "Community legal education" (Cooper and Trubek, 1997, p. 233) could be accelerated with increasing access to the Internet and the development of expert systems in law which threaten the power base of the legal profession and its exclusive control over legal knowledge and expertise (Susskind, 1998, 2008).

Law schools tend to deploy a variety of methods to deliver teaching and national traditions and available resources determine these. Perhaps as a consequence of increased communication and awareness of different approaches there is convergence resulting from the exchange of ideas and experimentation. The casebook approach pioneered by Christopher Columbus Langdell at Harvard over a century ago and the 'Socratic method' in which students explain the logic of legal argument through a question and answer session conducted in a large classroom have been exported to other common law countries. Most undergraduate law teaching is delivered through a combination of formal lectures backed up by tutorials, workshops, and/or seminars. Postgraduate legal studies are more library and seminar based. Class sizes are what distinguish the common law from the civilian legal systems, with the latter typically requiring lecturers in core subjects to perform before audiences that can reach thousands rather than mere hundreds. Common law systems select students prior to entry while in many civil law systems law students are sifted at the end of their first year. Teaching legal skills has now become an important part of the informal law curriculum in Anglo-American law schools that have moved beyond mooting (mock trials) to include client-interviewing and clinical approaches that sometimes involve advising or representing live clients (Evans, 2011). However, the fact that European Union law – the new *ius commune* – is now part of the mainstream of the curriculum means that both civilian and common law methodologies are understood by law students throughout Europe (East and West) and that European law students can observe in the jurisprudence of courts in Luxembourg and Strasbourg the synthesis of these hitherto contrasting legal methodologies.

A further recent trend in legal education common to many modern law schools concerns the revival of interest in the ethics and morality of law. In the United States, the Watergate scandal led to compulsory courses in law schools in professional responsibility while in many continental legal systems 'deontology' is taught but usually in a highly formal uncritical manner which does not depart far from the code of professional conduct. For most of the past two centuries positivism has prevailed in the teaching of law with law students being taught to focus almost exclusively on studying legal rules divorced from their moral, political, or social context. Contextual and critical tendencies in legal scholarship challenged legal formalism since the 1970s, but more recently there have been attempts to (re)introduce into the undergraduate curriculum ethical perspectives in order to enrich and deepen the education of law students (Economides, 1998; Boon, 2010). By exposing law students to ethics and values, the aim is to promote a more responsible and less cynical approach to both legal study and practice and instill a sense of professionalism (Sullivan et al., 2007). Evidence from several jurisdictions in both the common law and civilian legal systems suggests that the modern law school inculcates egotistical cynicism rather than an altruistic sense of public duty or any strong commitment to justice (Cooper and Trubek, 1997, 1999; Economides, 1998, 1999; Junqueira, 1999; Erlanger et al., 1996; Granfield, 1992). The need to strengthen the teaching of legal ethics arises not only at the initial stage of legal education but is also necessary at other stages of the educational continuum (Economides and Rogers, 2009).

The Educational Continuum

Legal education should be seen as part of 'lifelong learning' and need not be confined to what happens inside law schools or preparation for legal practice. The term 'law-related education' (LRE) usually refers to legal education in schools as

opposed to 'public legal education' (PLE) for adults, outside of school (Cassidy, 2000; Plenet). The fact that law schools are expected to deliver lawyers ready for practice at or soon after graduation, often, results in narrow, cramped, and overcrowded curricula with an unhealthy obsession with examination grades. This obsession with grades (which determine entry and status within the job market) arguably undermines the educational experience at any given point along the educational continuum and substitutes competitiveness for collegiality (Thornton, 2012). It also inhibits intellectual freedom by forcing the law student to memorize rules rather than explore ideas, with the result that legal education tends toward 'training for hierarchy' (Kennedy, 1982). While the education of lawyers clearly falls into distinct phases, it is perhaps preferable to regard the acquisition of legal skills and knowledge as part of ongoing processes that are continually refined throughout a legal career, and in this way reduce some of the pressure and expectations invested in particular *rites de passage* such as graduation and professional qualification. Ideally each stage should prepare the student for what follows next without being too prescriptive about the content or methods deployed at any particular stage.

The 'Initial' Stage

Legal education normally commences with an undergraduate degree although in some countries law may be taught in secondary schools along with instruction in civics or citizenship. In the United States most law schools require graduate entry and this is also the case in India and Pakistan. The undergraduate law degree – in which the foundations of legal knowledge are laid – is thus the normal initial or academic stage of legal education although most countries permit a measure of flexibility with regard to entry (and exit) to professional qualification. Postgraduate degrees serve to prepare for specialist areas of legal work, a career in academic law or cleanse an undistinguished first degree, and here students may follow a taught Master's course or read for a higher degree leading to an MPhil or PhD.

In the Anglo-American system, and many other countries in the common law world, students are expected to pass examinations in four or five subjects at the end of each year. It is also increasingly common for law schools themselves to be ranked according to their research and teaching quality, such rankings appearing in league tables organized and publicized by government or leading newspapers. The US News and World Report, for example, grades law schools in the United States, and in the United Kingdom the Higher Education Funding Council (HEFCE) administers the assessment of research quality through the Research Excellence Framework (REF), the results of which are fed into league tables organized by national newspapers. Examinations also feature in many continental European countries but here the university law degree invariably takes much longer, and in Germany the norm is 5 years (although this currently is under review) and concludes with a single comprehensive examination (the First State Examination) at the end of those 5 years (Korioth, 2006–07). Other European countries combine the system of yearly examinations in each course with a comprehensive examination at the end of the 5-year period of study (Lonbay, 1990, 2010a,b). One of the costs of the examination system is compartmentalized thinking in which students are unable to think across subject boundaries and for this reason it has become common to introduce general topics such as 'Public Law' and 'Private Law' which are spread over 2 years. Apart from the written examination most law schools offer students the opportunity to write a long essay or dissertation, sometimes in lieu of written examination, and continental law schools in Italy and France make use of oral examinations and a thesis to be written in the students' final year. In English law schools the *viva voce* examination is usually confined to resolving problematic or borderline results but increasing use is made of various forms of continual assessment. Students can sometimes also gain credit for working on law reviews, but this remains quite rare outside the United States.

The 'Vocational' Stage

In England and Wales, lawyers qualify as barristers or solicitors by taking either the Bar Vocational Course (BVC) or the Legal Practice Course (LPC) (see Economides and Smallcombe, 1991). Both courses have introduced a new skills element now taken at various franchised institutions including some university law schools. Simulated learning in virtual learning environments is being pioneered for vocational courses in both the United Kingdom and Australia (Maharg, 2007). Following the Courts and Legal Services Act 1990, which blurred the distinction between barristers and solicitors by allowing certain of the latter to gain higher courts rights of audience, the arguments in favor of common training for both branches have become much stronger (see ACLEC, 1996: Chapter 5; Lonbay, 2012, pp. 484–485). Currently barristers' training is focused on the four Inns of Court (Lincoln's Inn, Inner Temple, Middle Temple, and Gray's Inn) and falls under the jurisdiction of the Bar Standards Board (BSB), while solicitors' training is now governed by the Solicitors Regulation Authority (SRA), with the new Legal Services Board (LSB) overseeing the work of these frontline regulators following the Legal Services Act 2007. University graduates may obtain partial or total exemption from the first part, depending on their degrees, provided these satisfy the Joint Announcement on Qualifying Law Degrees. To qualify as a barrister, one must pass the BVC and serve a period of apprenticeship (pupilage) that involves 6 months in chambers attached to a barrister. The pupil barrister cannot practice unsupervised until she or he has been a pupil for 1 year. A major obstacle for students coming from poorer backgrounds has been finding a place in chambers but the profession has initiated reforms to try and assist the fair distribution of the limited training places that become available. In order to qualify as a solicitor one must pass the LPC and get a training contract in which one serves an apprenticeship as a trainee solicitor, normally attached to a firm of solicitors, for 2 years. Scotland and Ireland have similar arrangements although both jurisdictions have progressed farther with the notion of common training. Judicial training, which usually follows a successful career at the Bar, is now overseen by the Judicial Studies Board while employed lawyers, for example in the Crown Prosecution Service, are subject to further training in-house. Vocational legal education in recent years has been dominated by the large city firms which have invested heavily in in-house

training (and toyed with the idea of setting up their own vocational course provider), while many small firms in the provinces and rural areas, even though willing, have been unable to afford to take on trainee solicitors. Common training exists in the United States in that passing the state bar examination (which does vary) both grants admission to the Bar and qualifies one for all categories of legal work.

In civil law countries specialization takes place following graduation. In Japan, Spain, and France there are special national schools to train judges, entry to which is highly competitive and based on performance in an examination, while in Germany and Scandinavia law graduates intending to pursue a judicial career also take further vocational examinations and are then attached to a court while serving a probationary period at the end of which tenure for life is granted. Those wishing to qualify as lawyers follow training geared to particular forms of legal practice. In France, the choice is among qualifying as an advocate, an *avoué*, or a notary. In Germany, law graduates wishing to practice follow a program of practical training as a *Referenda* in diverse legal settings across the public and private sectors at the end of which they must pass the state bar examination (*Assessorexamen*) (see further Lonbay, 1990).

Continuing Professional Development

Most, if not all, modern legal professions either encourage or offer practitioners some form of post-qualification practical training in order to refresh their knowledge of substantive law while informing them of procedural and other developments that concern the business and management side of legal practice. Increasingly, however, Continuing Professional Development (CPD) is compulsory. England and Wales have led the field: the Law Society was the first to introduce mandatory CPD in 1985 and the Bar followed in 1997 (see ACLEC, 1997). In the United States CPD, there known as Continuing Legal Education (CLE), is now mandatory in the vast majority of states with some, for example, New York and Washington State, going further to require ethics as a subset of overall CLE credits earned within a reporting period (American Bar Association, 2012). Similarly in Australia, all states now require general mandatory CPD, with some such as Victoria and Queensland establishing a CPD core that must include ethics, advocacy, or practice management. New Zealand is following suit by making CPD mandatory from 2012. Over the past decade, continental Europe has also been coming on board steadily with 18 European countries now requiring CPD (CCBE, 2011). All 13 of the remaining European states that are yet to implement a mandatory regime offer various continuing training opportunities organized through professional associations, with Austria, Cyprus, Hungary, Latvia, and Malta all actively contemplating introducing mandatory CPD.

The Future of Legal Education

Future trends are difficult to discern, but it seems probable that legal education will become more international and virtual in character with law libraries possibly disappearing altogether as technology and online legal resources are more common, not only among law students but also members of the wider public. Law schools, as with legal professions, are increasingly subjected to managerialism and quality audits that question the traditional ideals of a liberal legal education. Graduate attributes that embrace skills needed in the workplace are challenging if not displacing the traditional role of law schools in developing legal reasoning through communicating the art of 'thinking like a lawyer.' As legal work evolves in a rapidly changing market for legal services so also are the expectations placed upon legal education which will need to produce minds that are reflexive, flexible, and innovative. The modern law school will continue its core functions of educating lawyers and legal scholars and producing research that creates the knowledge base for this education. But increasingly, a third function – outreach – is likely to manifest itself as law schools invest in the dissemination of legal knowledge and skill that serve the wider community (Coper, 2010). In conclusion, it may be predicted that legal education will become increasingly 'common,' and in at least three distinct ways. First, within jurisdictions, following current regulatory reform of the legal services market that blurs both inter- and intraprofessional boundaries, elements of advocacy and ethics training might be shared by trainee solicitors and barristers that eventually may lead to a complete fusion of vocational training and, in England and Wales, a merger of the LPC and BVC. Secondly, at federal and transnational levels, we are likely to witness greater convergence in the structures if not substance of legal educational frameworks so that, within Europe, North America, and Australasia, common approaches and prescriptions are adopted that mean legal standards and qualifications shift from state to regional, if not global, levels. Where one has trained is no longer going to be so closely tied to where one practices law as 'common training' among like-minded jurisdictions disregards geography. Finally, if current outreach initiatives meet with any success, legal education will become far more common in the sense of 'ordinary,' no longer the exclusive preserve of a closed professional elite or the middle classes, but an ever more inclusive and empowering experience that truly serves the legal needs of ordinary citizens.

See also: Comparative Research in Education: IEA Studies; Computer-Assisted Instruction; Education: Skills Training; Higher Education Management; Higher Education Market; Lifelong Learning and Its Support with New Technologies; Moral Development and Education; Moral Panics; Socialization and Education: Theoretical Perspectives.

Bibliography

Abel, R., 1999. The politics of professionalism: the transformation of English lawyers at the end of the twentieth century. Legal Ethics 2 (2), 131–147.

Advisory Committee on Legal Education and Conduct (ACLEC), 1996. First Report on Legal Education and Training. London.

Advisory Committee on Legal Education and Conduct (ACLEC), 1997. Continuing Professional Development for Solicitors and Barristers: A Second Report on Legal Education and Training. London.

American Bar Association, 2012. Mandatory Continuing Legal Education. At: http://www.americanbar.org/publications_cle/mandatory_cle.html.

Arthurs, H., 1984. Law and Learning. Report to the Social Sciences and Humanities Research Council of Canada by the Consultative Group on Research and Education in Law. Social Sciences and Humanities Research Council, Ottawa.
Australasian Law Teachers Association (ALTA), 1977. http://www.alta.edu.au/.
Baker, J.H., 1990. The Third University of England: The Inns of Court and the Common-Law Tradition. Selden Society Lecture, Seldon Soc, London.
Bradney, A., 2003. Conversations, Choices, and Chances: The Liberal Law School in the Twenty-First Century. Hart Publishing, Oxford.
Boon, A., 2010. Legal Ethics at the Initial Stage: A Model Curriculum. The Law Society, London. At: http://www.lawsociety.org.uk/new/documents/2011/Model-ethics-syllabus-Nov2010.pdf.
Brand, P., 1999. Legal education in England before the Inns of Court. In: Bush, J.A., Wijffels, A.A. (Eds.), Learning the Law: Teaching and the Transmission of Law in England, 1150–1900. Hambledon, London, pp. 51–84.
CCBE, December 6, 2011. Continuing Training in the CCBE Full Member Countries: Summary, 1/7.
Cassidy, W., 2000. Law-related education: promoting awareness, participation, and action. In: Goldstein, T., Selby, D. (Eds.), Weaving Connections: Educating for Peace, Social and Environmental Justice. Sumach Press, Toronto.
Cocks, R., 1983. Foundations of the Modern Bar. Sweet & Maxwell, London.
Cocks, R., 1988. Sir Henry Maine. A Study in Victorian Jurisprudence. Cambridge University Press, Cambridge.
Cooper, J., Trubek, L. (Eds.), 1997. Educating for Justice: Social Values and Legal Education. Dartmouth.
Cooper, J., Trubek, L. (Eds.), 1999. Educating for Justice: Legal Education, Legal Practice and the Community. Dartmouth.
Coper, M., 2010. Educating lawyers for what? Reshaping the idea of law school. Penn State International Law Review 29 (1), 25–39.
Cownie, F., Cocks, R., 2009. A Great and Noble Occupation!: The History of the Society of Legal Scholars. Hart Publishing, Oxford.
Cownie, F. (Ed.), 2010. Stakeholders in the Law School. Hart Publishing, Oxford.
Economides, K., Smallcombe, J., 1991. Preparatory Skills Training for Trainee Solicitors Law Society Research Study No.7. The Law Society, London.
Economides, K. (Ed.), 1998. Ethical Challenges to Legal Education and Conduct. Hart Publishing, Oxford.
Economides, K., 1999. Learning the law of lawyering. In: Freeman, M.D.A. (Ed.), Current Legal Problems, vol. 52. Oxford University Press, Oxford, pp. 392–418.
Economides, K., Rogers, J., 2009. Preparatory Ethics Training for Future Solicitors. The Law Society, London. At: http://www.lawsociety.org.uk/influencinglaw/policyinresponse/view=article.law?DOCUMENTID=419357.
Erlanger, H., Epp, C., Cahill, M., Haines, K., 1996. Law student idealism and job choice: some new data on an old question. Law & Society Review 30 (4), 851.
European Law Faculties Association (ELFA), 1995. http://elfa-afde.eu/legal.aspx.
Evans, A., 2011. Normative Attractions to Law and Their Recipe for Accountability and Self-assessment in Justice Education. In: Bloch, F. (Ed.), The Global Clinical Movement. OUP, New York. Chapter 25.
Flood, J., 2011. Legal Education in the Global Context: Challenges from Globalization, Technology and Changes in Government Regulation. U. of Westminster School of Law Research Paper No. 11–16. Available at SSRN: http://ssrn.com/abstract=1906687.
Franken, H., 1995. Een eigen richting voor het recht…Eindrapport van de Verkenningscommissie Rechtsgeleerdheid (A Sense of Direction…The Future of Legal Research in the Netherlands) Final Report of the Legal Education Review Commission, Leiden,.
Granfield, R., 1992. Making Elite Lawyers: Visions of Law at Harvard and Beyond. Routledge, New York.
International Association of Law Schools, 2005. http://www.ialsnet.org/.
International Association of Legal Ethics, 2010. http://iaole.org/.
Junqueira, E.B., 1999. Faculdades de Direito ou Fábricas de Ilusões? (Faculties of Law or Factories of Illusion?). Letra Capital/IDES, Rio de Janeiro.
Kennedy, D., 1982. Legal education and the reproduction of hierarchy. Journal of Legal Education 32, 591.
Korioth, S., 2006–07. Legal education in Germany today. Wisconsin International Law Journal 24 (1), 85.
Legal Education and Training Review (LETR), 2012. http://letr.org.uk/.
Lonbay, J., 1990. Training Lawyers in the European Community. The Law Society, London.
Lonbay, J., November 2010a. The education, licensing, and training of lawyers in the European union, part II: the emerging common qualifications regime and its implications for admissions in Europe. The Bar Examiner 25–35.
Lonbay, J., 2010b. Assessing the European market for legal services: developments in the free movement of lawyers in the European union. Fordham International Law Journal 33 (6), 1629–1669.
Lonbay, J., 2012. The changing regulatory environment affecting the education and training of Europe's lawyers. Journal of Legal Education 61, 479.
MacCrate, R., 1992. Legal Education and Professional Development – An Educational Continuum. Report of the Task Force on Law Schools and Profession: Narrowing the Gap. American Bar Association, Chicago.
Maharg, P., 2007. Transforming Legal Education: Learning and Teaching the Law in the Early Twenty-First Century. Ashgate, Aldershot.
Ormrod Report, 1971. Report of the Committee on Legal Education. HMSO Cmnd.4595, London.
Pearce, D., 1987. Australian Law Schools: A Discipline Assessment for the Commonwealth Tertiary Commission. Australian Publishing Service, Canberra. Public Legal Education Network (Plenet) at: www.plenet.org.uk.
Stevens, R., 1983. Law School: Education in America from the 1850s to the 1980s. The University of North Carolina Press, Chapel Hill and London.
Sugarman, D., 2011. A special relationship? American influences on English legal education, c.1870–1965. International Journal of the Legal Profession 18, 1–2, 7–57.
Sullivan, W.A., et al., 2007. Educating lawyers: preparation for the profession of law. At: http://www.carnegiefoundation.org/sites/default/files/publications/elibrary_pdf.
Susskind, R., 1998. The Future of Law. Facing the Challenges of Information Technology. Oxford University Press, Oxford.
Susskind, R., 2008. The End of Lawyers?: Rethinking the Nature of Legal Services. Oxford University Press, Oxford.
Sutherland, A.E., 1967. The Law at Harvard. A History of Ideas and Men, 1817–1967. Belknap Press, Harvard, Cambridge, MA.
Thornton, M., 2012. The Case of Law Routledge. Privatising the Public University, London.
Twining, W., 1994. Blackstone's Tower: The English Law School. Sweet & Maxwell, London.

Legal Formalism

Daniel A Farber, School of Law, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Kennedy, volume 13, pp. 8634–8638, © 2001, Elsevier Ltd.

Abstract

Legal formalists seek certainty and clarity in law. Formalism can take several jurisprudential forms: a preference for bright-line rules over more flexible legal standards, belief that legal decisions never require recourse to nonlegal materials, or belief that every legal decision has a single right answer. In the United States, formalism was a dominant view between the Civil War and the New Deal, but then disappeared until the 1980s, when it reemerged in the hands of conservative jurists advocating originalism in constitutional law and textualism in statutory interpretation.

Legal formalism is a familiar idea to lawyers, judges, and legal historians. Yet providing a rigorous definition of legal formalism is difficult, and on examination it turns out to have several different dimensions. If nothing else, the term connotes an effort to reduce or eliminate the role of policy judgments in judicial decision making, but this effort can take multiple forms. The history of legal formalism in the United States also is more complex than suggested by a traditional story of post-Civil War rise and post-New Deal fall. Among other complications to this story, formalism has had a revival of sorts in recent years, particularly in the hands of conservative lawyers and judges. Although this article focuses on the robust debates over legal formalism within the United States, similar developments have occurred in other common law jurisdictions.

The author begins by considering the meaning of legal formalism. Like 'judicial activism' and 'judicial restraint,' it is a term often used for praise or condemnation, but its exact meaning can be elusive (Kennedy, 2001). Formalism remains something like the official ideology in civil code countries, but has a much more varied history in the United States (Merryman and Perez-Perdomo, 2007). The history of formalism in American law falls into two distinct periods. One period straddles the turn of the twentieth century, running from the end of post-Civil War Reconstruction in the 1870s to the New Deal in the 1930s, and the other beginning in the final quarter of the twentieth century and continuing (so far) into the first quarter of the twenty-first. These two formalist periods are discussed in separate sections below. Finally, the author considers the somewhat complex relationship between formalism, judicial behavior, and political ideology.

What Is Legal Formalism?

Legal formalism takes different forms, making the concept multidimensional (Schauer, 1988; Bix, 1999). It may focus on the utility of clear rules as opposed to more amorphous legal standards such as reasonableness. Another variant of formalism attempts to protect the authority of legal texts against efforts to broaden the interpretative context or import ideas about fairness or social policy into interpretation. A third variant of formalism emphasizes the aspiration of the legal system to closure, making it possible to decide cases without adverting to consideration outside of the law itself (Kennedy, 2001). These three types of formalism require separate discussion.

The first type of formalism can be called rule formalism. Legal scholars distinguish between rules that provide bright-lines versus more flexible standards (Kennedy, 1976). The standard example is a numerical speed limit versus a prohibition on driving unreasonably fast. Rules have obvious advantages in terms of providing clear notice and objective standards for decision makers (Schauer, 1988, 1991). Rule formalism states a preference for rules over standards. A disadvantage of rules, however, is that they can never precisely fit the circumstances of each case. For example, without some kind of standard-based exception, a numerical speed limit would make it illegal to travel faster to the hospital in a medical emergency. Standards may provide less notice but have the advantage of being better tailored to produce fair outcomes in individual cases.

A second type of formalism emphasizes the authoritative status of governing legal texts (Scalia and Garner, 2012). Again, there are clear advantages and disadvantages. Texts provide a confined field for legal interpretation, as opposed to approaches allowing consideration of a broader range of legal sources such as legislative history or later judicial interpretations. This may provide clearer notice of legal mandates and might limit the range of disagreement compared to approaches that feature a broader range of sources or reliance on public policy or equitable considerations. The disadvantages are that the text may be misleading if read in isolation from the surrounding context and may lead to undesirable decisions if social policy and equity are excluded from consideration (Llewellyn, 1960).

A third type of formalism emphasizes the autonomy of law, ruling out any considerations that are not based in identifiably 'legal' materials, and suggesting that in principle every legal problem has a right answer. The central idea is that law is a closed system. This idea may be associated with proponents of some forms of legal positivism but is rejected by other legal positivists who see a role for policy considerations in resolving gaps in existing law (Hart, 1994). Moreover, some nonpositivist thinkers also view law as a closed system (Dworkin, 1978).

These three types of formalism are often attractive to the same thinkers, but they are logically distinct. For instance, although the other versions of formalism may share their proponents' aspiration to limit the influence of nonlegal factors, support for rule formalism and textualism may be based on pragmatic justifications rather than a belief in formalism as a jurisprudential position. A rule formalist might believe that the content of the rules should be determined entirely on the basis of policy considerations. Conversely, a jurisprudential formalist could believe that relevant legal materials are not limited to particular texts such as statutes or Constitutions, that texts must be interpreted on the basis of context, and that law often takes the form of standards whose application can only be determined by considering the facts of individual cases.

The ideal type of formalist would combine all these approaches, believing that law is (or should be) a closed system of rules that can be understood and applied through interpretation of authoritative legal texts. Civil code systems in Europe and elsewhere aspire to satisfy all three types of formalism, although perhaps with less success than civil lawyers might like to admit (Merryman and Perez-Perdomo, 2007). As seen in the following section, there was also a period in American law during which, at least according to some legal historians, this ideal type of formalism held sway.

The Rise and Fall of Classical Formalism

By most accounts, the heyday of American legal formalism was the long period of industrialization beginning with the end of Reconstruction (1876) and ending with the New Deal (1933–36). Historians view judicial opinions prior to this era as being more pragmatic and more willing to advert to policy considerations rather than rooting decisions entirely in authoritative legal texts (Gilmore, 1977; Llewellyn, 1960). After the Civil War, however, there was a move toward more rigorous reliance on texts and deduction from general legal principles.

In the classic period of legal formalism, courts disclaimed agency in their rulings, sometimes portraying their roles in more mechanical terms as measuring a situation against legal yardsticks. The role of law was seen as policing clear lines of demarcation among various law-making bodies and between them and private citizens (Kennedy, 1980). Nineteenth-century formalism is often thought to have been allied with business interests, providing protection against regulatory efforts or potential ameliorative changes in the common law (Grey, 1983). But historians generally paint a more complex picture of the relationship between nineteenth-century law and business interests (Friedman, 1973).

Nineteenth-century formalism also saw the law as complete, in the sense of supplying the necessary grounds for deciding legal questions without reference to external systems of morality or policy (Gilmore, 1977). For instance, a leading exponent of formalism famously declared issues of practicality irrelevant to contract law (Grey, 1983). Formalists aspired to a kind of geometric logic in which general premises would generate specific conclusions through deduction (Grey, 1983), although it is not clear whether any formalists actually believed that in practice the law attained this level of logical rigor.

As with many historical periods, the ability to perceive the nature of legal thought during this classical period is obscured by the efforts of succeeding periods to portray negatively prior thinking in order to make the case for new approaches. A full understanding of law from the period of approximately 1870–1930 would surely be more complex than the picture sketched above.

The turn of the twentieth century saw the beginning of a reaction to formalism, starting with work by Oliver Wendell Holmes and culminating in the movement known as legal realism (Pound, 1917; Cardozo, 1921; Gilmore, 1977). Legal realists took a variety of positions but shared a disdain for the formalist view of law as an impersonal neutral process (Kalman, 2011). Instead, they stressed that judicial decisions were intimately related to the judge's values and worldview (Frank, 1949). Following World War II, legal scholars were repelled by the abuse of the legal system under totalitarian regimes and began to place renewed emphasis on the rule of law (Bix, 1999). The result was a domesticated form of legal realism known as the legal process school, which conceived of law as a purposive endeavor requiring sensitivity to context and flexibility in interpretation. Legal process theory in its turn was challenged by the Law and Economics movement, which shared the legal process interest in social policy but called upon the use of economics to resolve policy issues (Posner, 1990). At roughly the same time, the Critical Legal Studies movement revived legal realism with an admixture of postmodernist theory (Kennedy, 1976; Gordon, 1984). What Critical Legal Studies and Law and Economics shared was a rejection of the idea of law as an autonomous enterprise that could be detached from broader policy disputes. In that sense, they were both the antithesis of formalism.

The Formalist Revival

In the post-World War II era, formalism appeared a thing of the past, with the term acquiring a pejorative connotation (Friedman, 1973). Critical Legal Studies and the Law and Economics movements seemed to mark a definitive rejection of the aspiration to policy-free legal decision making. Just when legal formalism seemed extinct in American legal thought, however, it began a revival.

This revival was partly based in the legal academy but received important support from a cohort of lawyers and judges associated with the Reagan Administration (Avery and McLaughlin, 2013). As discussed later, there is no necessary logical connection between formalism and conservative politics, but much of the support for reviving formalism in the United States did come from conservative quarters, paralleling the favorable reception of formalism among conservative jurists in the earlier formalist period.

The revival of formalism was largely but not entirely directed toward public law. Late twentieth-century formalists called for a rejection of evolutionary visions of constitutional law and a return to the understandings of the original framers of the Constitution of the United States. In statutory

interpretation, they called for strict reliance on the text of statutes, undiluted by attention to legislative history or other indicators of legislative intent. Formalists also addressed private law, however, arguing in favor of clear rules to bound tort liability and reasserting the importance of written contract terms as against commercial context or policy considerations.

Originalism in Constitutional Law

Perhaps the most important focal point of the formalist revival, however, was constitutional law. Conservatives rebelled against the concept of a 'living Constitution' and instead demanded a return to the original understanding of the Constitution by its Framers (Scalia, 1998). Originalism rested on a cluster of arguments. First, originalists argued, the Constitution's authority derived from its adoption by a supermajority, so that any interpretation contrary to their views lacked this legitimizing foundation. On the other hand, when judges carried out the views of this supermajority, they could claim a democratic pedigree even for rulings contrary to the views of current majorities (Bork, 1997).

Second, according to originalists, interpretation based on the original understanding was consistent with the methods used to interpret other legal documents such as contracts and wills (Scalia and Garner, 2012). Thus, originalism was faithful to the nature of the Constitution as a legally operative rather than purely political document. It is also, according to originalists, consistent with the 'writeness' of the Constitution, which implies a desire to fix rules rather than leaving them open for revision under the guise of interpretation (Scalia and Garner, 2012).

Third, originalists claimed that their approach would provide a principled basis for litigation, eliminating or at least sharply limiting judicial discretion. Other approaches were too formless to constrain judges from simply imposing their own policy views at the expense of the democratic process. In the view of originalists, such judicial policy making happened all too often especially in cases involving the rights of criminal defendants, religious freedom, abortion, and federal power. They viewed these decisions as lacking any foundation in historical evidence of the Framers' understanding of the Constitution (Bork, 1997).

Originalism initially took the form of reliance on the intentions of the Framers (Bork, 1997). Although this approach still has its adherents, it confronted significant criticisms because of its tension with formalist premises. In formal terms, it is the text of the Constitution that is legally operative, rather than mental states of the Framers. Moreover, the intentions of the various Framers may have been different and even inconsistent with each other. Reliance on expectations about how the Constitution would be applied was an especially slippery basis for interpretation, since those expectations may have been based on factual assumptions that might not hold today or that current society might have viewed as false even then (Berman, 2009).

In response to these concerns, many originalists switched their focus from expectations or intentions to understandings or 'semantic meaning' (Scalia and Garner, 2012; Berman, 2009). Put another way, they de-emphasized the authors of the Constitution in favor of its audience, seeking to identify Constitutional meaning with the interpretation of a reasonable reader of the time. This move to audience understandings has the advantage of operationalizing originalism and providing a recipe of sorts for its application. But it also opens an array of questions about the background knowledge of the reasonable reader and the methods of interpretation that the reader would have applied. Typical questions might be whether the reasonable reader was familiar with Blackstone or Locke, whether the reader should be presumed to apply methods of legal interpretation current at the time, whether the reader would be familiar with documents such as the Federal Papers, and whether the reader had clearly formulated views of constitutional structure and purpose beyond those apparent in the text itself.

Regardless of its particular form, originalism received a barrage of criticism. Critics argued that originalism failed as an effort to root judicial review in the prior democratic adoption of the Constitution. To begin with, critics contended, the adoption of the Constitution itself had suspect democratic credentials. At the time the Constitution or the major Reconstruction amendments were adopted, women did not vote, and blacks were rarely able to do so. Property qualifications were standard in the earlier period as well. And it remains unclear, even apart from these considerations, that a restriction on the current democratic majority can gain legitimacy solely from the facts of its enactment by an earlier, long-dead majority – the well-known 'dead hand' issue (Berman, 2009; Farber and Sherry, 2001).

Another line of attack focused on the originalist contention that their methodology simply applied usual methods of legal interpretation to the Constitution. Some critics questioned this description of legal interpretation. In particular, they questioned whether originalism corresponded to the methods of interpretation used by courts during the period when the Constitution was adopted. Others observed that the Constitution is not the typical legal document and might call for different methods of interpretation (Rakove, 1997).

Critics also attacked the originalist claim of reducing judicial discretion and thereby depoliticizing constitutional law. Historians were startled by the implication that historical events have readily ascertainable interpretations or that contemporary meanings of constitutional texts were clear cut. They also questioned whether lawyers and judges possessed the training and historical knowledge needed for expert judgment about long-ago periods of time. Meanwhile, critics also argued that originalism is indeterminate on the critical question of the level of generality; for instance, in a case involving racial segregation, it is unclear whether judges should be seeking the original understanding about whether segregated schools violate the equal protection clause or the original understanding of the general concept of equality embedded in the clause (Berman, 2009; Farber and Sherry, 2001).

Finally, critics argued that if originalism were consistently applied, it would have unacceptable consequences. Modern Supreme Court decisions banned racial segregation, provided legal protection to advocacy by dissidents, limited discrimination against women, and allowed federal regulation of matters such as employment discrimination, environmental pollution, and organized crime. Critics portrayed originalism as a threat to reimpose an archaic legal order to the detriment of

racial and gender equality and of civil liberties (Farber and Sherry, 2004).

Textualism in Statutory Interpretation

Paralleling the effort to establish originalism in Constitutional law was a campaign for textualism as a basis for statutory interpretation. By the 1970s, the standard judicial approach to statutory interpretation involved consideration of statutory text, purpose, and legislative history (Frickey et al., 2006). Textualists argued with particular force against reliance on legislative history, and for reliance on the statutory text alone. They questioned the credibility of legislative history materials as reflections of the views of legislators. Rather than representing the intent of the legislators, textualists argued that legislative history materials reflected only the views of committees, individual prominent legislators, or even staff members. Consequently, reliance on legislative history in effect delegated power to control interpretation of a statute to a subset of the legislative branch, violating separation of powers principles (Scalia, 1998).

Textualists also argued that recourse to legislative history and speculation about legislative purposes misunderstood the nature of the legislative process. They contended that legislation is inevitably the product of compromise, especially in the US system where so many potential veto points have to be surmounted and minorities thus have multiple opportunities to kill legislation. Such compromises might not be apparent in the legislative history and might indeed be deliberately hidden from publicity (Manning, 2005).

As with originalism, supporters of textualism argued that their approach would constrain judicial discretion and help depoliticize statutory interpretation. Sprawling legislative histories of major legislation provide too many opportunities for judges to cherry-pick snippets supporting their desired outcome. Purposes can be characterized in different ways, making attributions of purpose inherently manipulable. In contrast, the statutory text is a delimited object meant to be understood by the legal community, with any potential ambiguities resolvable by well-known canons of interpretation such as interpretation of ambiguous criminal statutes in favor of the accused (Scalia, 1998; Scalia and Garner, 2012).

Like originalism, textualism gave rise to fierce scholarly debate. Defenders of the use of legislative history argued that it is more credible than textualists admit, because there are likely to be sanctions against committees or sponsors who misrepresent the content of legislation. In addition, many legislators rely on committee reports rather than directly accessing statutory language, so the committee report may be a better gauge of the intentions of Congress. In addition, critics of textualism argue that reliance on legislative history in statutory interpretation does not involve delegation of law-making power to subunits of the legislative branch because legislative history is used to illuminate binding legal texts rather than having any independent force of law (Frickey et al., 2006).

Critics also suggest that textualists oversimplify the legislative process. Positive political theory suggests that legislators are far from chaotic, making it less implausible to attribute purposes to legislation. Moreover, although some legislative provisions are compromises between competing interest groups, others may reflect coherent legislative purposes sought by a unified, dominant coalition. Careful attention to the legislative record combined with an understanding of the key hurdles for particular bills may allow identification of the key actors and reconstruction of their understandings of the legislation (Rodriguez and Weingast, 2003).

Critics of textualism also argue that during the framing period, courts employed flexible methods of statutory interpretation employing ideas of statutory purpose and equity. They point to examples of eighteenth-century courts interpreting statutes to cover significantly less than their texts would suggest, or in some cases significantly more (Eskridge, 2001). Textualists respond that this form of interpretation predated a better understanding of the system of separation of powers inaugurated by the Constitution, and that courts soon began to employ less flexible methods of interpretation (Manning, 2001).

Like critics of originalism, critics of textualism also argue that its determinacy is substantially overrated. In interpreting a statute, a textualist must choose a segment of text to interpret (the whole statute, the narrow provision at issue, or something in between). The textualist also has a choice of different resources to use in interpreting language, including multiple dictionaries that may have varied definitions of terms. In construing ambiguities, the textualist has recourse to a stable of interpretative canons that often conflict in their implications in particular cases (Frickey et al., 2006). Thus, textualism may not be successful in cabining discretion, at least if the critics are correct.

Formalism and Precedent

Formalists employing originalism or textualism also must confront another problem. In many cases, a court does not operate on a clean slate, but instead confronts one or more precedential opinions, which often will not rests on these formalist theories. The formalist judge may then be faced with following precedents that the judge views as incorrect or rejecting wholesale an established body of precedent. If the judge decides not to overturn precedent, the question arises of how broadly to interpret the precedent. Few formalist judges are willing to completely ignore precedent. But accepting nonformalist past rulings can also result in a patchwork of rulings based on inconsistent grounds (Farber and Sherry, 2008).

The Formalist Revival in Private Law

Formalism has also had a revival in private law. In contract law, for example, post-World War II doctrine and scholarship stressed contextual interpretation in order to give effect to the intentions of the parties, including broad consideration of trade usage, past dealings between the parties, and the negotiations over the contract. Courts were called upon to fill contractual gaps on the basis of reasonableness (Llewellyn, 1931). Since the 1980s, however, there has been a countermovement among scholars to give more weight to the text of the agreement and to limit consideration of context or determinations of reasonableness. This

countermovement is based on doubts about whether courts can accurately determine the intentions of the parties and on the desirability of providing certainty regarding the meaning of commercial agreements (Schwartz and Scott, 2010).

Similarly, in tort law, post-World War II scholars and judges tended to view torts as a mechanism to deal with the social costs of accidents. This perspective supported a broadening of liability and a move away from reliance on clear rules toward more general standards based on reasonableness. More recently, however, there has been a renewed effort to provide clear limits on tort liability by delimiting a defined set of potential victims who are recipients of the duty of due care (Goldberg, 2010).

Rules versus Standards

The rise of originalism and textualism was paralleled by a more general defense of rules as against standards by scholars, including now-Justice Antonin Scalia (Scalia, 1989; Alexander and Sherwin, 2001). The argument stresses certain functions of law that may be best achieved by rules. First, to the extent that laws are intended to guide behavior, rules may give clearer guidance than standards about society's expectations. The implicit comparison is between a standard and a single clearly stated rule. Translating the substance of a standard into rules, however, might require a complicated code that may be opaque to many individuals.

A second formalist argument for rules is that they better serve the dispute-resolution purpose of law (Alexander and Sherwin, 2001). If a rule is clear, all parties should be able to agree on the resolution of the dispute. On the other hand, a standard will be much more subject to dispute. Worse, application of the standard may depend on normative judgments based on concepts such as reasonableness or good faith, and the parties are likely to have different moral judgments about the situations (Scalia, 1989; Weinrib, 1988).

Finally, to the extent society aspires to the rule of law rather than the rule of persons, rules may be more readily subject to objective interpretation rather than depending on the perspective and values of the judge. It is for this reason that some formalists have suggested that the rule of law and the law of rules are identical.

The Two Eras of Formalism

There is an obvious kinship between the two eras of formalism in their goals and in at least some of their methods. Modern formalists, however, do not commonly portray themselves as seeking to return to a lost golden era of late nineteenth-century law. The reason may be that the legal realists so successfully lambasted classical formalism as to discredit it as a source of inspiration.

Like classical legal formalism, the revival of formalism in American legal thought was linked with conservative ideology, in particular with the conservative revival of the Reagan years (Greene, 2009; Avery and McLaughlin, 2013). This historical connection with conservative politics raises the question whether formalism and conservatism share a deep conceptual connection. It also raises the question of whether formalism in fact constraints judges or merely provides a rhetorical framework for decisions that are driven by other factors. The author considers those issues below.

Formalism, Ideology, and Judicial Behavior

Despite the correlation between formalist and conservative views, the relationship is surely not universal. For example, Richard Posner, a leading conservative scholar and jurist, is profoundly antiformalist (Posner, 1990). On the other hand, some progressive legal scholars such as Jack Balkin have embraced originalism (Balkin, 2011), and the most notable contemporary liberal legal philosopher endorsed a version of the formalist view of legal closure (Dworkin, 1978).

These examples indicate that the link between formalist approaches and specific legal outcomes may not be powerful. If both conservatives and liberals can embrace originalism without any sense of discomfort, this suggests that originalism may not be experienced as requiring changes of position regarding issues such as affirmative action, abortion, or states' rights. If this is correct, it may have negative implications for the formalist aspiration to constrain decision making and reduce the influence of 'nonlegal' factors such as political ideology or judicial background.

The success of formalism as a normative system is best assessed through empirical evidence. There is a limited but growing empirical literature in this area. The extant evidence, discussed in this section, reinforces doubts about the potential for formalist methodologies to shape outcomes in judicial decision making.

Despite all the debates over textualism and statutory interpretation, scholars know relatively little about the practical impact of various theories of interpretation. A good test would involve presenting randomly selected cases to two judges who are similar in most other respects but have sharply different theories of interpretation. A natural experiment is provided by a study of two judges on the same court, Judges Posner and Easterbrook, who at the time of the study had sat together on over 800 reported decisions by three-judge panels of the United States Court of Appeals for the Seventh Circuit (Farber, 2001). Posner is a leading advocate of pragmatism, while Easterbrook is equally well known for his support of formalism. In other respects, the two judges are very similar: both are conservative, both were professors at the University of Chicago Law School, and both are highly cited. As it turns out, the two judges voted differently in only 1% of these roughly 800 cases. This was substantially below the average level of disagreement among Seventh Circuit judges, and was also below the average rate for courts of appeals generally. Moreover, a careful examination of the statutory cases in which the two judges disagreed confirms that their theoretical disputes were not outcome determinative. Because Posner and Easterbrook have shown particularly serious interests in theories of interpretation (compared to most judges), the apparently slight level of impact on outcomes seems especially surprising. While far from definitive, this study suggests strongly that any link between theories of interpretation and judicial decisions

may be tenuous. If true, that would have negative implications for the aspiration of textualists to constrain judicial behavior, and instead would support more legal realist views that emphasize judicial ideology and worldview.

Other studies also suggest that formalist methods have at most modest effects on judicial outcomes. Using data derived from briefs filed by litigants over 8 years, researchers tested the influence of arguments about text and intent on the decisions of US Supreme Court Justices (Howard and Segal, 2002). They found that judges supported textual or intentional arguments when they were made by parties in their own ideological camp but not otherwise. Their analysis indicated that ideology was a more powerful influence on decision making than adherence to originalism. Questions about the constraining effect of originalism were also raised by a study of a federal appellate court, which found that reliance on originalism by liberal judges actually increased the likelihood of a liberal result (Czarnezki and Benesh, 2009). A more recent study of originalism suggested that it may also be unsuccessful in constraining judicial behavior. Although originalism may figure increasingly in judicial opinions, the study suggested that substantial flexibility is provided by the complexities of the historical record and by the methodological options available within originalist theories (Cross, 2013). This flexibility allows judges to deploy originalism, whether as a rationalization or as a sincere justification, on more than one side of controversial issues such as Second Amendment protection for guns.

In terms of the closure thesis, the advent of modern databases of judicial opinions has allowed some systematic study of the impact of the formalist revival on the types of materials referenced by judges. The evidence to date indicates that citations to formalist sources may have increased but that courts continue to rely on a wide variety of sources beyond statutory and constitutional text or historical documents such as the Federalist Papers (Zeppos, 1992; Krishnakumar, 2010; Schacter, 1998).

In terms of rules versus standards, empirical evidence seems to be lacking on whether in application formal legal mandates based on rules provide more predictability or a better public understanding of legal requirements than mandates based on standards. The answer is less obvious than it might seem, in part because what appears to be a rule (it is illegal to drive over the speed limit) may in application be a standard (drivers will be ticketed if they drive too much over the speed limit).

Legal formalism tends to be a subject on which legal philosophers, historians, and judges take sides. The subject would benefit from more dispassionate investigation, and in particular from more empirical investigation of formalist and antiformalist claims. Despite what some observers might have thought a half century ago, however, there is no sign that formalism is likely to disappear or that the vigorous debates over its validity are likely to fade.

See also: Critical Legal Studies; Law and Society: Development of the Field; Legal Realism; Legal Reasoning and Argumentation.

Bibliography

Alexander, L., Sherwin, E., 2001. The Rule of Rules: Morality, Rules, and the Dilemmas of Law. Duke University Press, Durham, NC.

Avery, M., McLaughlin, D., 2013. The Federalist Society: How Conservatives Took the Law Back from Liberals. Vanderbilt University Press, Nashville, TN.

Balkin, J., 2011. Living Constitutionalism. Harvard University Press, Cambridge, MA.

Berman, M., 2009. Originalism is Bunk. New York University Law Review 84, 1–78.

Bix, B., 1999. Jurisprudence: Theory and Context. Carolina Academic Press, Durham, NC.

Bork, R., 1997. The Tempting of America. Free Press.

Cardozo, B., 1921. The Nature of the Judicial Process. Yale University Press, New Haven, CT.

Cross, F., 2013. The Failed Promise of Originalism. Stanford University Press, Palo Alto, CA.

Czarnezki, J., Benesh, S., 2009. The ideology of legal interpretation. Journal of Law & Policy 29, 113–132.

Dworkin, R., 1978. Taking Rights Seriously. Harvard University Press, Cambridge, MA.

Eskridge, W., 2001. All about words: early understandings of the "Judicial power" in statutory interpretation, 1776–1806. Columbia Law Review 101, 190–306.

Farber, D., 2001. Theories of statutory interpretation matter? Northwestern University Law Review 94, 1409–1444.

Farber, D., Sherry, S., 2001. Desperately Seeking Certainty: The Misguided Quest for Constitutional Foundations. University of Chicago Press, Chicago, IL.

Farber, D., Sherry, S., 2008. Judgment Calls: Politics and Principle in Constitutional Law. Oxford University Press, New York, NY.

Frank, J., 1949. Law and the Modern Mind. Coward-McCann, New York.

Frickey, P., Eskridge, W., Garrett, E., 2006. Legislation and Statutory Interpretation, second ed. Foundation Press, Eagan, MN.

Friedman, L., 1973. A History of American Law. Simon and Schuster, New York.

Gilmore, G., 1977. The Ages of American Law. Yale Univeristy Press, New Haven, CT.

Goldberg, J., 2010. Torts as wrong. Texas Law Review 88, 917–981.

Gordon, R., 1984. Critical legal histories. Stanford Law Review 36, 57–125.

Greene, J., 2009. On the origins of originalism. Texas Law Review 88, 1–96.

Grey, T.C., 1983. Langdell's orthodoxy. University of Pittsburgh Law Review 45, 1–53.

Hart, H.L.A., 1994. The Concept of Law. Clarendon, Oxford, UK.

Howard, R., Segal, J., 2002. An original look at originalism. Law and Society Review 36, 113–138.

Kalman, L., 2011. Legal Realism at Yale, 1927–1960. North Carolina University Press.

Kennedy, D., 1976. Form and substance in private law adjudication. Harvard Law Review 89, 1685–1778.

Kennedy, D., 1980. Toward an historical understanding of legal consciousness: the case of classical legal thought in America, 1850–1940. Research in Law and Sociology 3, 3–24.

Kennedy, D., 2001. Legal formalism. In: The International Encyclopedia of the Social and Behavioral Sciences 8634. Pergamon Press, Oxford, U.K.

Krishnakumar, A., 2010. Statutory interpretation in the Roberts Court's first era. An Empirical and Doctrinal Analysis Hastings Law Journal 62, 221–289.

Llewellyn, K., 1931. What price contract?—An essay in perspective. Yale Law Journal 40, 704–751.

Llewellyn, K., 1960. The Common Law Tradition – Deciding Appeals. William S. Hein & Co. Aspen Books, Denver, CO (1996 reprint by Vintage).

Manning, J., 2005. Textualism and legislative intent. Virginia Law Review 91, 419–450.

Manning, J., 2001. Textualism and the equity of the statute. Columbia Law Review 101, 1–141.

Merryman, J., Perez-Perdomo, R., 2007. The Civil Law Tradition: An Introduction to the Legal Systems of Europe and Latin America, third ed. Stanford University Press, Palo Alto, CA.

Posner, R., 1990. The Problems of Jurisprudence. Harvard University Press, Cambridge, MA.

Pound, R., 1917. The end of law as developed in juristic thought II: the nineteenth century. Harvard Law Review 30, 201–246.

Rakove, J., 1997. Original Meanings: Politics and Ideas in the Making of the Constitution. Vintage Books, New York.

Rodriguez, D., Weingast, B., 2003. The positive political theory of legislative history: new perspectives on the 1964 Civil Rights Act and its interpretation. University of Pennsylvania Law Review 151, 1417–1521.

Scalia, A., 1989. The rule of law as a law of rules. University of Chicago Law Review 56, 1175–1188.

Scalia, A., 1998. A Matter of Interpretation: Federal Courts and the Law. Princeton University Press, Princeton, NJ.

Scalia, A., Garner, B., 2012. Reading Law: The Interpretation of Legal Texts. Thomson Reuter, Eagan, MN.
Schacter, J., 1998. The confounding common law originalism in recent Supreme Court statutory Interpretation: implications for the legislative history debate and beyond. Stanford Law Review 51, 1–66.
Schauer, F., 1988. Formalism. Yale Law Journal 97, 509–548.
Schauer, F., 1991. Playing by the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and in Life. Oxford University Press, New York.
Schwartz, A., Scott, R., 2010. Yale Law Journal 119, 926–963.
Weinrib, E.J., 1988. Legal formalism: on the immanent rationality of law. Yale Law Journal 97, 949–1016.
Zeppos, N., 1992. The use of authority in statutory interpretation: an empirical analysis. Texas Law Review 70, 1073–1143.

Legal Issues: Public Opinion

Robert J MacCoun, University of California, Berkeley, CA, USA
Karin D Martin, John Jay College of Criminal Justice, CUNY, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R. MacCoun, volume 13, pp. 8641–8646, © 2001, Elsevier Ltd.

Abstract

This article reviews trends in public opinion toward laws, legal authorities, and the functioning of the legal system, and the psychological and sociological bases for those opinions. Section Public Evaluations of the Legal System: An Overview examines the general favorability of public evaluations of the legal system. Section Sources and Mechanisms of Influence on Public Views examines mechanisms of influence on public opinion. Section Underlying Dimensions of Public Evaluations of the Legal System documents the highly multidimensional nature of citizen evaluations of legal issues.

This article reviews trends in public opinion toward laws, legal authorities, and the functioning of the legal system, and the psychological and sociological bases for those opinions. Section Public Evaluations of the Legal System: An Overview examines the general favorability of public evaluations of the legal system. Citizens in many nations tend to rate the police favorably, but there are notable exceptions, both within nations (e.g., African Americans in the United States) and across nations (e.g., East Germany in the 1980s). Judges are viewed favorably, but in some nations there is a persistent view that criminal sentences are too lenient. Section Sources and Mechanisms of Influence on Public Views examines mechanisms of influence on public opinion. Direct experience with legal authorities is important, but because such experience is fairly limited for many citizens, the mass media play a significant and frequently distortive role in shaping public views. Cognitive mechanisms, such as hindsight bias, can also skew public perceptions and evaluations. Section Underlying Dimensions of Public Evaluations of the Legal System documents the highly multidimensional nature of citizen evaluations of legal issues. These dimensions include outcome satisfaction, distributive justice, procedural justice, retributive justice, fear of crime, fear of legal sanctions, and legitimacy. Research has identified distinct criteria that influence each dimension.

Two guiding principles of Western democratic thought are that effective government requires the consent of the governed, and public consent requires the fair and just rule of law. Thus, research on public opinion toward the law and the legal system has preoccupied mainstream social science for much of its history, reappearing in various guises, including the 'free press vs fair trial' debate, seemingly paradoxical links between actual crime and fear of crime, public disdain for 'judicial leniency,' and theories of perceptual deterrence.

This short review will of necessity be selective in its coverage. With respect to substantive legal issues, it will emphasize criminal justice, providing only passing references to tort litigation and no coverage of public opinion in the domains of contract law, family law, intellectual property law, and so on. With respect to methodology, it will emphasize quantitative public opinion surveys, statistical modeling of field data, and laboratory experiments, neglecting the rich but idiosyncratic literature using qualitative interviews and qualitative analyses of textual sources. Finally, the article – like the English-language empirical literature – gives disproportionate emphasis to North American and British research, although some lines of cross-cultural research are noted.

Public Evaluations of the Legal System: An Overview

Evaluating the Police

In the aggregate, Americans and Europeans evaluate the police favorably. A majority of Americans report that the police have high or very high honesty and ethical standards (Bureau of Justice Statistics, 2011). Such confidence is not to be taken for granted. For example, in 2011, most European Union citizens expressed a high level of trust in the police (between 60 and 90%), while the average was much lower in nine Eastern European countries (e.g., around 40% for the Czech Republic, Romania, Slovenia, Slovakia, and Bulgaria; Eurobarometer©, European Union, 2012). Within the United States, African Americans consistently report more negative views of police conduct than the white majority; they are about twice as likely to report low confidence in the police or to express the view that the police have low ethical standards.

Evaluating Attorneys

Most adults know, or at least recognize, a few of the multitude of 'lawyer jokes' long in circulation (see Galanter, 1998). We rarely tell jokes about physicians, whom we hold in high regard, or molesters, whom we loathe. This suggests that we are ambivalent about lawyers; we decry their ruthless mercenary adversarialism – except when they are representing us in a conflict. Thibaut and Walker (1975) asked Americans to imagine various hypothetical conflicts and various mechanisms for resolving them – including two-party mediation, third-party investigation and resolution by a neutral inquisitor, and a third-party decision based on investigation and arguments presented by advocates for each side. Importantly, the descriptions were stripped of labels like 'lawyer' or 'judge' and other overt references to actual legal systems.

A strong majority identified the adversarial system as the fairest mechanism. Subsequent studies (see Lind and Tyler, 1988) have replicated this finding in European nations with inquisitorial systems, suggesting that it is not an artifact of American socialization or familiarity with adversarial justice. However, respondents in Asian countries tend to divide their support between the adversarial model and two-party mediation.

Evaluating the Courts

Americans – of all ethnicities – are more than three times as likely to rate the honesty and ethical standards of judges as high or very high (47%) than low or very low (14%) (Bureau of Justice Statistics, 2011). But this high regard does not translate into satisfaction with judges' decisions. For several decades in the twentieth century, American public opinion polls showed consistent majority support for the view that the courts are too lenient in their treatment of criminal defendants (Bureau of Justice Statistics, 2011). Interestingly, African Americans are only slightly less likely to endorse this view, and the gap between racial groups has been closing.

Taken at face value, these statistics imply that the courts are either 'out of touch' with public views or intentionally ignoring them. There is some basis for the latter interpretation, of course; Western systems of rights and protections evolved in large part to protect individuals from group tyranny. But a number of clever experiments reveal that these citizen complaints are largely based on distorted perceptions of the justice system and its caseload (see Roberts and Stalans, 1998). Citizens who view courts as lenient in the abstract have been found to recommend criminal sentences that fall *below* statutory mandatory minimum sentences or the discretionary sentences of judges. And citizens' views in actual cases become significantly more lenient when they read lengthy trial descriptions rather than short trial summaries.

While there is consistent evidence that Americans hold much more favorable opinions of the jury system than of the courts more generally (Harris Poll, 2008; Hans, 1993) and specifically prefer trial by jury to trial by judge (MacCoun and Tyler, 1998), there is evidence that race strongly influences this preference. Rose et al. (2008) find that non-Hispanic whites have a strong and consistent preference for juries, African Americans and 'English-dominant' Hispanics have a less pronounced preference for juries, and Spanish speakers and noncitizens have a clear preference for judges.

Sources and Mechanisms of Influence on Public Views

Direct Experience

The procedural justice literature documents the importance of direct personal experience in shaping the public's views of law and legal authorities. Most citizens have had direct contact with the police, some have had contact with lawyers, fewer have participated in trials, and only a minority has direct experience with the penal system. There is evidence that the custodial aspect of the criminal justice system has a negative effect on political participation for those who do come into contact with it, even controlling for criminal propensity or socioeconomic status (Weaver and Lerman, 2010).

For many citizens, jury service is their major (or only) direct experience with the court system. Diamond (1993) has reviewed survey evidence on the effects of jury service on public attitudes, finding that a majority of jurors provide positive evaluations of the jury system, and of the legal system, after their service. Most jurors say their attitudes became more favorable than they were before their service, and attitudes are more favorable among those who are selected for a jury than among those who report but are not seated.

Mass Media Effects

Citizens without direct contact with the legal system must base their views on other sources. But even those with direct experience are routinely bombarded with indirect, vicarious information. There is an enormous literature demonstrating that the mass media provide and promote distorted views of not only crime but also the criminal justice process and the outcomes of the justice system (e.g., Roberts and Stalans, 1998; Warr, 1994). The presumed causal chain is ACTUAL EVENTS → MEDIA FILTERING AND DISTORTION → BIASED PUBLIC BELIEFS AND OPINIONS, with these public views feeding back to the media via consumer choices.

The first part of the chain is well supported. Numerous content analyses over several decades show a consistent pattern: media coverage provides a highly skewed picture of crime and the legal system, and this is true of newspapers, television news, 'reality-based' police shows, and fictional portrayals on television and in movies. In the criminal domain, reported offenses are disproportionately violent, defendants are disproportionately African American, police officers are disproportionately European American, victims are disproportionately female, and offenses disproportionately involve strangers (see Warr, 1994). In the personal injury domain, reported lawsuits are much more likely than the typical or average lawsuit to result in a jury trial and to produce a plaintiff victory, while reported state court jury awards are 14 times larger than the average award and 34 times the size of the median award (Bailis and MacCoun, 1996).

There is also evidence for the second half of the causal chain. The prejudicial effects of pretrial publicity on criminal juror judgments are well documented (see Steblay et al., 1999). And numerous studies indicate that citizens' views of court leniency are shaped by the mass media (see Warr, 1994). Citizens tend to overestimate the severity of typical crimes and criminals (relative to official statistics), and these misperceptions reliably predict the propensity to see the courts as excessively lenient; citizens given newspaper accounts of trials supported harsher criminal sentences than citizens who read actual court documents from the same cases (see Roberts and Stalans, 1998).

Hindsight Bias

Hindsight bias is the *ex post* tendency to overestimate the *ex ante* likelihood of an outcome, relative to what one would have actually guessed before the event. Because most legal judgments are made *ex post*, they are vulnerable to this bias, as

documented in a variety of experimental studies (Rachlinski, 1998). Hindsight bias also influences citizens' *ex post* reactions to newsworthy legal decisions. Trial evidence appears more incriminating to people who believe a defendant has been convicted than for those who believe the defendant was acquitted (Bodenhausen, 1990).

A pernicious example of this phenomenon was the elite and popular reaction to the acquittal of actor–athlete O.J. Simpson at his criminal homicide trial in the United States. Because nine of the 12 jurors were African American, a large number of commentators, including the lead prosecutor, 'explained' the verdict as a case of 'racial nullification' – the notion that African Americans are more lenient when judging other African Americans. In an October 1995 Gallup poll, 48% of whites vs 18% of African Americans said they had 'less confidence' that 'jurors can reach a verdict in a trial without letting their racial attitudes affect their judgment' (Moore and Saad, 1995). In fact, the jury literature provides little support for 'racial nullification' as a general phenomenon (Kerr et al., 1995). The reasons for the acquittal will never be known with certainty, but it seems clear that pronouncements of 'racial nullification' reflected hindsight bias. Despite widespread knowledge of the composition of the jury, surveys during the trial showed that few trial attorneys and citizens expected an acquittal; indeed, the *Sporting Index* reported that most American gamblers were betting on a conviction (*Los Angeles Times*, 1995).

Underlying Dimensions of Public Evaluations of the Legal System

The history of theory and research on lay justice judgments involves a continuous branching process in which simple, unidimensional models have given way to more complex multidimensional models (see Tyler et al., 1997).

Outcome Satisfaction, Relative Deprivation, and Distributive Justice

Economics, behaviorist psychology, the social exchange tradition of sociology, and the public choice tradition of political science share a common emphasis on the explanatory power of outcomes and incentives. The mass media and the legal literature tend to perpetuate the view that outcomes – especially monetary outcomes – drive legal behavior, legal judgments, and evaluations of the legal system. An early refinement of this view was relative deprivation theory (see Tyler et al., 1997) – the notion that what matters to citizens are relative outcomes (mine vs yours or theirs) rather than absolute outcomes. As a predictive theory, the relative deprivation approach has had only mixed success, in large part because the theory *overpredicts* resentment and rebellion. One problem is that the theory fails to predict which of many possible comparison standards the citizen will use; citizens do not invariably choose the source that provides the most invidious comparisons. Another weakness is the failure to predict whether outcome discrepancies will be seen as justified or not.

Descriptive (rather than normative) theories of distributive justice address this latter gap. In the 1970s, the most influential model was equity theory (e.g., Walster and Walster, 1975), which models justice judgments as a function of the relative ratio of inputs to outcomes for each actor in a relationship. Indeed, in reactions to job-related conflicts, citizens in Eastern and Central Europe (Bulgaria, Hungary, Poland, and Russia) are at least as likely to apply this equity standard as are citizens in traditional Western market democracies (France, Spain, and the United States) (Cohn et al., 2000). But it has become clear that this model works less well in nonmarket contexts and relationships (Fiske, 1992). Across various settings, studies have found support for allocation by equality, by need, or by more complex multidimensional decision rules (see Mellers and Baron, 1993). In addition, judgments of distributive justice (and their behavioral consequences) vary by social level of analysis; for example, perceiving that one's group has received unfair outcomes is more likely than purely individual deprivation to motivate anger and political action (see Tyler et al., 1997).

Citizens' judgments about monetary outcomes – punitive damage awards, child support awards, and regulatory fines – can appear incoherently volatile, but there is more coherence than meets the eye. Citizens tend to agree on relative magnitudes across case characteristics (i.e., the slopes of the predictor variables) but vary considerably in their sense of absolute magnitudes (i.e., the y-intercept terms). Much of the latter variation is explainable by differences in distributive justice beliefs, but some is due to the inherent difficulty of mapping subjective judgments onto the dollar metric (Ellman et al., 2012).

Procedural Justice

Outcomes are clearly linked to public evaluations of justice, but there is growing evidence that scholars have overestimated the importance of this linkage. Thibaut and Walker's (1975) influential experimental research program provided considerable evidence that the processes by which outcomes are reached matter profoundly to citizens. Since the 1970s, this 'fair process effect' has been replicated across a wide variety of nations (in North America, Europe, and Asia), domains (criminal justice settings, tort litigation, private organizations, and the political sphere), and dependent variables (satisfaction with the legal system, acceptance of third-party decisions, compliance with legal authorities, and decisions to file lawsuits) (see Lind and Tyler, 1988; MacCoun, 2005; Tyler et al., 1997). In a surprising number of studies, procedural justice judgments emerge as stronger independent predictors than either objective or subjective measures of the outcomes that citizens received.

Thibaut and Walker (1975) contended that procedures matter to citizens because fair procedures produce fair outcomes. The key parameters of their model were decision control and process control; the latter is important because it is rarely possible to resolve conflicts without relinquishing some decision control. The predominant determinant of process control in the Thibaut–Walker program was 'voice' – the opportunity to control the presentation of one's arguments to a third-party decision maker. Even among tort litigants, popularly portrayed by the media as greedy plaintiffs and stingy defendants, the 'opportunity to tell my story' is the

most commonly cited goal of both sides. Indeed, this voice motive helps to explain why alternative dispute resolution (ADR) mechanisms have generally failed to streamline court trial dockets; because ADR provides a relatively cheap way to 'tell my story,' it diverts more cases from private settlement than from trial (see MacCoun, 2005).

Thibaut and Walker's control-based account of voice effects has fared less well. Lind and Tyler (1988) have provided the most empirically successful account of voice effects. Their group value theory argues that voice effects are 'relational' rather than instrumental; they reflect citizens' motivation to achieve a favorable and secure social identity. In numerous studies, relational concerns – the sense that authorities acted with benevolent and impartial motives, and that one was treated with dignity and respect – emerge as more important determinants than either outcome evaluations or perceptions of control (Lind and Tyler, 1988; Tyler, 1990; Tyler et al., 1997). Within the United States, these relational factors are found in judgments across gender, ethnic, educational, and occupational categories, and these findings have been replicated in Europe and even in more collectivist Asian societies (Tyler et al., 1997).

There are moderators and boundary conditions on the model. Procedural justice effects are weaker for citizens who identify strongly with a subordinate group but weakly with a superordinate group (e.g., minority 'separatists' as opposed to 'assimilationists'). Another is that procedures matter more when citizens lack clear objective criteria for evaluating outcomes (van den Bos et al., 1997).

Nevertheless, fair process effects are so robust that scholars fear they pose a threat of 'false consciousness' – the notion that authorities can use the appearance of fair procedure (dignity, respect, and voice) as an inexpensive way to co-opt citizens and to distract them from outcomes that by normative criteria might be considered substantively unfair or biased (e.g., MacCoun et al., 1992). For example, an insurance executive requested a meeting with two procedural justice researchers to ask how he might increase the formality of meetings between clients and insurance adjustors, in order to reduce the rate of contested claims.

Retributive Justice

It is important to note that citizens' departures from pure economic self-interest are not merely warm, sentimental, and benevolent. The behavioral game theory literature has provided ample evidence that people will often forgo personal gain in a spiteful attempt to harm others, and it is tempting to interpret the strong retributive streak in American public opinion in a similar way. From the 1980s through the 2000s, the United States experienced a dramatic escalation in the size of its prison and jail populations (Glaze, 2011), especially relative to those of other industrialized nations. Between 1985 and 1996, the prison population more than doubled, yet the percentage of Americans who felt that their courts were 'not harsh enough' only dropped from 84 to 78% (Bureau of Justice Statistics, 2011).

Public misperception of crime rates is surely a partial explanation, but several lines of evidence implicate retributive motives, rather than an instrumental attempt at crime control. For example, many citizens overtly endorse a deterrence rationale for the death penalty, believing that 'it will prevent crimes.' Yet most do not change their views when asked how they would feel if there were unequivocal evidence that execution provided no marginal deterrence above and beyond life imprisonment (Ellsworth and Gross, 1994). Moreover, there is evidence that even those who endorse deterrence apply a 'just deserts' framework of proportionality to justify punishment – suggesting that people believe crimes like murder deserve the most severe punishment, but not necessarily the death penalty (Carlsmith et al., 2002) (*see* Death Penalty).

Fear of Crime

It is a cliché among academicians that fear of crime bears no relation to actual crime rates. In fact, between 1973 and 1987, American fear ratings moved in tandem with the violent crime rate (Warr, 1994). But perceptions of crime trends are not very accurate; a growing minority (from 4% in 1993 to 35% in 1998) recognize that US crime fell in the late 1990s, but since then, an increasing number (from 34% in 2000 to 68% in 2011) contend that 'there is more crime in the US than a year ago' (Bureau of Justice Statistics, 2011).

In many surveys, the elderly and women express the most fear of crime; these patterns are called 'paradoxical' (because young males are the group at greatest risk of victimization) but only under the peculiar notion that fear should solely reflect relative statistical risk. It appears that these patterns reflect not only differences in the way women and the elderly conceptualize the question (Ferraro, 1995; Warr, 1994), but also the influence of media depictions that exaggerate the incidence of crime (Warr, 2000) and the proportion of crimes in which the victims are white females (Chiricos et al., 1997). That fear of crime is an expedient tool for garnering political support further complicates the issue (Warr, 2000; Chevigny, 2003).

But the fear literature is plagued by inconsistent findings, perhaps due to a frequent reliance on crude, single-item 'fear' measures (e.g., 'How safe do you feel being alone in your neighborhood at night?') that may not measure the experience of fear (i.e., events in the amygdala and autonomic nervous system) and that surely tap other constructs, like perceived risk, attitudes toward the police and the courts, attitudes toward stereotypic offender groups (youth, males, and ethnic minorities), and so on. For example, Quillian and Pager (2001) found that the percentage of black males in a neighborhood was significantly associated with the perception of crime rates – even controlling for the actual crime rate and other neighborhood characteristics. Local racial composition is also a significant influence on the perception of crime victimization risk (Quillian and Pager, 2010).

Fear of Legal Sanctions

The general public tends to exaggerate the risks of arrest and punishment for many crimes, relative to actual risks and to the more accurate risk perceptions of active criminal offenders; indeed, citizens in states that have decriminalized marijuana are only slightly less likely to believe that possession of small quantities is punishable by incarceration (MacCoun et al., 2009). These facts are consistent with either a deterrence

interpretation (a perceived legal risk deters crime) or a learning interpretation (offenders learn that legal risks are low). But quasi-experimental studies of legal interventions and panel self-report surveys suggest that the certainty of sanctions does deter crime, and that the effects of sanction severity are weak and unreliable (Paternoster, 1987) (*see* Deterrence).

Legitimacy

In a series of experiments, Robinson and Darley (1997) identified numerous discrepancies between the Model Penal Code and lay intuitions about legal liability. They hypothesize that congruence between lay views and the formal law enhances both legitimacy and compliance. Tyler's (1990) panel survey of Chicago citizens suggested that those who perceived laws and legal authorities to be fair were less likely to report having committed criminal offenses when surveyed at a later point in time. A reanalysis of data from the Milwaukee Domestic Violence Experiment found that the perceived unfairness of police interventions was a stronger predictor of subsequent domestic violence than was the initial decision to arrest or not to arrest the offender (Paternoster et al., 1997).

Conclusion

Public opinion research on legal issues has benefitted enormously from the increasing use of advanced multivariate analyses, panel surveys tracking individuals' views over time, and factorial experiments. But there are far too few cross-national studies. The bulk of the English language literature describes the views of North American and British citizens in the late twentieth century. Thus, we still know far too little about how people judge laws and legal authorities outside of common law systems in wealthy, advanced industrial nations. We do not know whether these Anglo-American views reflect fundamental psychological principles or simply the particular context in which they were expressed.

See also: Attitudes, Political and Public Opinion; Criminal Justice, Sociology of; Death Penalty; Disputes, Social Construction and Transformation of; Juries; Justice and Law; Justice, Conflicts, and the Justice of Conflict Resolution; Justice: Philosophical Aspects; Justice: Social Psychological Perspectives.

Bibliography

Bailis, D.S., MacCoun, R.J., 1996. Estimating liability risks with the media as your guide: a content analysis of media coverage of tort litigation. Law and Human Behavior 20, 419–429.

Bodenhausen, G.V., 1990. Second-guessing the jury: stereotypic and hindsight biases in perceptions of court cases. Journal of Applied Social Psychology 20, 1112–1121.

Bureau of Justice Statistics, 2011. Sourcebook of Criminal Justice Statistics Online, 2011. Table 2.0016.2010; Table 2.47.2011; Table 2.35.2011; Table 6.28.28.2011. http://www.albany.edu/sourcebook/.

Carlsmith, K.M., Darley, J.M., Robinson, P.H., 2002. Why do we punish? Deterrence and just deserts as motives for punishment. Journal of Personality and Social Psychology 83, 284–299.

Chevigny, P., 2003. The populism of fear: politics of crime in the Americas. Punishment & Society 51, 77–96.

Chiricos, T., Eschholz, S., Gertz, M., 1997. Crime news and fear of crime: toward an identification of audience effects. Social Problems 44, 342–357.

Cohn, E.S., White, S.O., Sanders, J., 2000. Distributive and procedural justice in seven nations. Law and Human Behavior 24, 553–579.

Diamond, S.S., 1993. What jurors think: expectations and reactions of citizens who serve as jurors. In: Litan, R.E. (Ed.), Verdict: Assessing the Civil Jury System. Brookings Institution, Washington, DC.

Ellman, I.M., Braver, S.L., MacCoun, R.J., 2012. Abstract principles and concrete cases in intuitive lawmaking. Online first version. Law and Human Behavior. http://wwwspringerlinkcom/content/t2611582v5335337/fulltextpdf.

Ellsworth, P.C., Gross, S.R., 1994. Hardening of the attitudes: Americans' views on the death penalty. Journal of Social Issues 50, 19–52.

Eurobarometer©; European Union, 1995–2012. http://ec.europa.eu/public_opinion/cf/showchart_column.cfm?keyID=2197&nationID=11,1,27,28,17,2,16,18,13,32,6,3,4,22,33,7,8,20,21,9,23,24,12,19,29,26,25,5,14,10,30,15,&startdate=2010.11&enddate=2010.11 (last accessed July 2013.).

Ferraro, K.F., 1995. Fear of Crime: Interpreting Victimization Risk. State University of New York Press, Albany, NY.

Fiske, A.P., 1992. The four elementary forms of sociality: framework for a unified theory of social relations. Psychology Review 99, 689–723.

Galanter, M., 1998. The faces of mistrust: the image of lawyers in public opinion, jokes, and political discourse. University of Cincinnati Law Review 66, 805–845.

Glaze, L.E., 2011. Correctional Population in the United States, 2010. U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, NCJ 236319.

Hans, V.P., 1993. Attitudes toward the civil jury: a crisis of confidence? In: Litan, R.E. (Ed.), Verdict: Assessing the Civil Jury System. Brookings Institution, Washington, DC.

The Harris Poll® #9, January 21, 2008. Regina A. Corso, Director, The Harris Poll, Harris Interactive.

Kerr, N.L., Hymes, R.W., Anderson, A.B., Weathers, J.E., 1995. Defendant–juror similarity and mock juror judgments. Law and Human Behavior 19, 545–567.

Lind, E.A., Tyler, T.R., 1988. The Social Psychology of Procedural Justice. Plenum Press, New York.

Los Angeles Times, 25 October 1995. They Bet on It. Los Angeles Times. A6.

MacCoun, R.J., 2005. Voice, control, and belonging: the double-edged sword of procedural fairness. Annual Review of Law and Social Science 1, 171–201.

MacCoun, R.J., Tyler, T.R., 1988. The basis of citizens' perceptions of the criminal jury: Procedural fairness, accuracy, and efficiency. Law and Human Behavior 12, 333–352.

MacCoun, R.J., Lind, E.A., Tyler, T.R., 1992. Alternative dispute resolution in trial and appellate courts. In: Kagehiro, D.K., Laufer, W.S. (Eds.), The Handbook of Psychology and Law. Springer Verlag, New York.

MacCoun, R., Pacula, R.L., Reuter, P., Chriqui, J., Harris, K., 2009. Do citizens know whether they live in a decriminalization state? State marijuana laws and perceptions. Review of Law and Economics 5, 347–371.

Mellers, B.A., Baron, J. (Eds.), 1993. Psychological Perspectives on Justice. Cambridge University Press, New York.

Moore, D.W., Saad, L., 2–8 October 1995. No immediate signs that Simpson trial intensified racial animosity. Gallup Poll Monthly.

Paternoster, R., 1987. The deterrent effect of the perceived certainty and severity of punishment: a review of the evidence and issues. Justice Quarterly 4, 173–217.

Paternoster, R., Brame, R., Bachman, R., Sherman, L.W., 1997. Do fair procedures matter? The effect of procedural justice on spouse assault. Law and Society Review 31, 163–204.

Quillian, L., Pager, D., 2001. Black neighbors, higher crime? The role of racial stereotypes in evaluations of neighborhood crime. American Journal of Sociology 107, 717–767.

Quillian, L., Pager, D., 2010. Estimating risk: stereotype amplification and the perceived risk of criminal victimization. Social Psychology Quarterly 731, 79–104.

Rachlinski, J.J., 1998. A positive psychological theory of judging in hindsight. University of Chicago Law Review 65, 571–625.

Roberts, J.V., Stalans, L.J., 1998. Crime, criminal justice, and public opinion. In: Tonry, M. (Ed.), The Handbook of Crime and Punishment. Oxford University Press, New York.

Robinson, P.H., Darley, J.M., 1997. The utility of desert. Northwestern University Law Review 91, 453–499.

Rose, M.R., Ellison, C., Diamond, S.S., 2008. Preferences for juries over judges across racial and ethnic groups. Social Science Quarterly 89, 372–391.

Steblay, N.M., Besirevic, J., Fulero, S.M., Jimenez-Lorente, B., 1999. The effects of pretrial publicity on juror verdicts: a meta-analytic review. Law and Human Behavior 23, 219–235.

Thibaut, J., Walker, L., 1975. Procedural Justice. Erlbaum Associates, Hillsdale, NJ.

Tyler, T.R., 1990. Why People Obey the Law. Yale University Press, New Haven, CT.

Tyler, T.R., Boeckmann, R.J., Smith, H.J., Huo, Y.J., 1997. Social Justice in a Diverse Society. Westview Press, Boulder, CO.

van den Bos, K., Lind, E.A., Vermunt, R., Wilke, H.A.M., 1997. How do I judge my outcome when I do not know the outcome of others? The psychology of the fair process effect. Journal of Personality and Social Psychology 72, 1034–1046.

Walster, E., Walster, W.G., 1975. Equity and social justice. Journal of Social Issues 31, 21–43.

Warr, Mark, 2000. Fear of crime in the United States: avenues for research and policy. Criminal Justice 4, 451–489.

Warr, M.J., 1994. Public perceptions and reactions to violent offending and victimization. In: Reiss, A., Roth, J.A. (Eds.), Understanding and Preventing Violence. Consequences and Control, vol. 4. National Research Council, Washington, DC.

Weaver, V., Lerman, A., 2010. Political consequences of the carceral state. American Political Science Review 104, 817–833.

Legal Personality

Gregory A Mark, Rutgers-Newark School of Law, NJ, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8646–8650, © 2001, Elsevier Ltd.

Abstract

Legal personality is the compendium of rights and duties that enable an actor to function in a legal system. A legal actor may be a human being, an organized entity, or even an object, or an office. Legal personalities need not all have the same content. The existence and content of any given legal personality may be a matter of both theoretical and historical controversy. At one extreme some theorists hold that legal personality inheres in certain individuals and entities. At the other extreme theorists maintain that large personality is a matter of a grant of personality from the sovereign authority. In practice all systems reflect aspects of both theories. While the debate over legal personality has deep roots in Western culture, the most prominent example in recent centuries of disputed legal personality is that of the business corporation. The history of the legal personality of the business corporation demonstrates not only the tension between the theoretical approaches to legal personality but also the ways in which the debate reflects larger politics, instrumental needs within a culture, and the malleability of jurisprudential concepts.

In order to be able to participate in any legal or political system an actor must have a legal personality, that is, a set of rights and duties recognized by that system. In theory a legal system may be premised on the belief that certain actors have an inherent legal personality, or that legal personality is entirely a matter of the grant of rights and placement of duties by the sovereign, or that some actors have inherent legal personalities and some are granted a personality. In practice, however, every legal system both recognizes rights and duties in certain actors and insists that certain legal personalities are a matter of positive law. As a historical matter, the content of an actor's legal personality has waxed and waned depending on the legal system in question, instrumental questions about the utility of legal personality, and cultural presuppositions about who and what might have the benefit of rights or the burden of duties. Especially because a legal actor may be neither human nor animate, as when a corporate body has a legal personality, discussions of both the legitimacy and the content of the actor's legal personality may become quite controversial. Thus, the jurisprudential basis for an actor's legal personality is a matter of theoretical and historical concern.

Every legal system must decide which individuals and collectivities, whether formally or informally constituted, or anything else, have legal personalities and what the content of those legal personalities might be. Some legal systems assume that certain actors have rights within the system based on inherent features of the actors, such as their humanity. Every legal system also grants certain actors not assumed to be able to act within the system the right, sometimes limited, to pursue their legal aims. This ascription of rights and duties upon an entity not presumed to be able to act within the system is its legal personality. Some legal systems, notably those premised on natural law (*see* Natural Law), but certainly not limited to such systems, have recognized that some actors have inherent legal personalities. In systems of positive law (*see* Legal Positivism), however, no quantum of rights inheres in any actor, at least in theory. Nonetheless, the history of legal personality is replete with examples of legal systems recognizing inherent rights in individuals, groups, and entities, whether their right to participate in the system is presupposed or granted as a matter of positive law.

The most prominent twentieth-century example of an entity granted legal personality is the business corporation, though the history and theory of legal personality long antedate the business corporation. Religious, charitable, and governmental bodies are among the entities that have historically been granted legal personalities, especially insofar as legal systems have recognized distinctions between the rights of an office as distinct from the rights of the person occupying the office. Moreover, in legal systems characterized by hierarchies of actors not all human beings have been assumed to have full legal personalities. Therefore, women (*see* Gender and the Law), children (*see* Children and Law), slaves (*see* Slaves and Slavery, History of), and foreigners, among others, have had to have their legal personality defined. Recently, the possibility of ascribing legal personality to animals and other entities has entered jurisprudential debates, largely in the hope that in doing so the legal system would necessarily find that they have rights that might be vindicated.

Legal Personality, Inherent Rights, and Positive Law

Two theoretical difficulties attend the concept of legal personality. The first is the class or classes of entities, which have personality in the eyes of the law. The second is whether legal personality inheres in any entity and what the content of inherent legal personality might be. Systems premised in natural law of some variety, for example, assume both that individual human beings have an irreducible quantum of legal personality and that the content of that personality provides each individual with a certain level of protection and dignity. Such systems need not, however, treat all human beings as having equivalent legal personality. An aristocrat and a slave may each possess a legal personality, but they need not be equal. Nor do political systems have to limit their vision to an

individualist conception of personality. Theories of organic polities, for example, those premised in republican (*see* Republicanism: Philosophical Aspects) or communitarian ideologies, maintain that group identity is fundamental to social organization and should be recognized at law. Even societies with a largely individualist politics recognize the legitimacy of some group activity. Legal personality may therefore be presumed to inhere in collectivities within society, such as tribes, governments, churches, and even offices, such as governmental or religious offices. The difficulty in any such system is, of course, that human beings are required to discover both who and what might have a legal personality and then to discover the content of that personality.

Systems of positive law appear to solve the problem of legal personality present in systems, which maintain beliefs in inherent legal personality. Under positive law legal personality does not inhere in anything or anybody but is rather the result of a grant by the sovereign authority. Hence, legal personality is a fiction created by the sovereign for instrumental purposes. The sovereign may decide to grant the right to act within the legal system to a person, to a class of persons, or to groups of people, constituted either formally or informally within the social structure. Indeed, the sovereign may grant the right to act to inanimate objects or to entities, which are themselves the creation of the sovereign, such as corporate or subsidiary political entities. The difficulty presented by systems of positive law is, however, one of rights consciousness within a culture. That is, cultural conceptions of the dignity of human beings, either individually or collectively, the naturalness of human associations, or even conceptions of the inherent rights of, for example, animals, may war with the sovereign's willingness to grant legal personality, no matter how minimal.

In practice virtually no system is either purely one in which legal personality inheres in people or things or in which the sovereign jealously guards the prerogative of the grant of personality. Systems premised on an understanding of inherent legal personality find themselves constantly revising the classes of individuals and entities covered and constantly discovering new aspects of their legal personalities. Moreover, sovereigns within such systems also indulge in the power to grant legal personality. Systems of positive law regularly discover that they must recognize aspects of legal personality that are not spelled out within the grant of personality. Couched sometimes as inferences from the grant and sometimes as necessary components of any grant of personality, these redefinitions of legal personality depend upon changing cultural, especially changing political and economic, circumstances for their content, even though the content of the changes must be recognized by some arm of the sovereign, for example, the courts. The history of legal personality demonstrates, moreover, that conceptions of legal personality have not neatly evolved from concepts of inherent personality to positive grants of personality. Instead, concepts of legal personality both depend upon and in turn influence surrounding circumstance.

Legal Personality in Jurisprudential History

The conflict between inherent and fictive conceptions of legal personality has a long history. Among the first instances in which the distinction between inherent and fictive legal personalities arose in Western culture came during the thirteenth century during the papacy of Innocent IV. The Pope, attempting to consolidate the power of the Church and to harmonize society under its auspices, used the concept to delegitimate the authority of institutions inferior to the Church. As the Church was the agent of God, all collective human activity was subordinate to it. Thus, other institutions were inferior to the Church and owed their existence not to their independent authority but existed at the behest of the Church or by its grace. They were, therefore, legal fictions. The Church was not alone in its attempt to subordinate other human institutions by fictionalizing their existence.

The early modern state and its successors have also tried to centralize power by subordinating other human institutions in the name of sovereignty. For example, guilds could only be legitimate to the extent that they acknowledged their derivative status. Similarly, corporate bodies could maintain corporate status only by virtue of charters. Under the Bubble Act collective corporate activity in England could be prosecuted unless done under charter, though such prosecutions were evidently quite rare. In the United States the power of cities, towns, and other local entities derives from the individual states. Indeed, even the sovereign authority of the states and the other sovereignty recognized within the United States, that of the Indian nations, have their own tortured histories. The reason they are tortured, of course, is that modern political consciousness accepts Innocent's positivist precept, that authority to act within a legal system depends on a grant from the sovereign authority, turning the concept of a sovereign existing within another sovereign into a first rank problem of political logic.

By way of contrast, the concept that legal personality is inherent stems both from aspects of natural law and political conceptions of an organic society. The religious traditions that informed natural law suggested that individuals possessed a minimum of dignity that others, the sovereign included, were bound to respect. Failure to respect human dignity imperiled one's soul. The history of the law of slavery is, for example, in many ways dependent on the distinction between the recognition of a slave's human dignity and the consideration of the slave's status as mere property. Parallel considerations animating organic politics legitimate recognition of groups and group activity. Some political theorists have long argued that social groups and collective activity are fundamental aspects of human life and therefore the recognition by a legal system of collective identity and the legal right of collective activity is not a function of sovereign grace but rather a function of the human condition. Such recognition has taken a variety of forms throughout history. Recognition of social and political class divided parliamentary bodies in England and France, whereas, at least in the pre-Civil War United States, one's loyalty to a state, defined formally as a geographic matter and informally as a matter of culture and politics, in part grounded the legitimacy of the federal system. Belief in the inherent legitimacy of group activity has informed spirited intellectual and political activity, such as Frederic Maitland's late nineteenth- and early twentieth-century desire to undermine certain individualist premises in the common law and to personify at law labor unions in an attempt to protect concerted activity by the working class in England.

In modern jurisprudence, however, no form of collective activity has caused more ink to be spilled in the debate over the nature of legal personality than the business corporation. Nowhere was the debate over the meaning of corporate personality as keenly debated in such a concentrated period as in the United States from the end of the second decade of the nineteenth century through the first third of the twentieth century. No mere academic dispute, the debate over corporate personality was framed by the twin considerations of society's power to promote and regulate economic activity on the one hand and the nature of collective activity, in this case entrepreneurial activity, on the other.

While English and American discussions of corporate personality existed before the nineteenth century, both in the common law and elsewhere, the first definitive exploration of the American position came in the Supreme Court's decision in *Dartmouth College vs Woodward*, 17 US 518 (1819). In that case Chief Justice John Marshall wrote that a corporation "is an artificial being, invisible, intangible, … possess[ing] only those properties which the charter of its creation confers upon it, either expressly, or as incidental to its very existence. These are such as are supposed best calculated to effect the object for which it was created." The declaration of the Chief Justice appeared to put the United States squarely in the camp of positive law insofar as corporate personality was concerned. Considering that elsewhere in American jurisprudence the legal personality of white male property-owning citizens was assumed, it is, however, quite clear that neither a belief in inherent legal personality nor a purely positivist vision animated overall American legal thought. Indeed, even within the realm of corporate personality both the case generally and Marshall's own opinion left room for both visions.

Despite the clear import of Marshall's language, the case held that the sovereign could not unilaterally alter the charter of its own creation when that creation was a private entity, which the college was. Even within the realm of positive law, therefore, Innocent's vision was subtly altered. The sovereign could, of course, reserve the power to alter charters it granted, and the chartering states quickly proceeded in that direction by legislative act. Nevertheless, even without recourse to the legislature, Marshall had left open the potential expansion of corporate personality by recognition of the necessary incidents of corporate existence.

Dartmouth itself was, of course, an educational and not a business institution. In the United States, however, the expansion of the economy and the entrepreneurial fervor of the citizenry meant that individuals looking for vehicles under which to organize their capital for business purposes quickly adopted the corporate form as their preferred mode. States, eager to accommodate, granted corporate charters in ever greater numbers. As the corporate form grew to be the natural mode for organizing capital, Marshall's supposition that corporations were 'artificial being[s]' grew ever more suspect. Rather than beings created by the state, they seemed creations of entrepreneurs. These creations, moreover, did not have simply the rights granted by the sovereign, but had whatever properties necessary to accomplish their object. Business existed in environments that required properties of great flexibility in order to operate and to achieve their object. They needed, among many other things for example, the right to own property, both real and personal, including intellectual property; they needed the right to operate across jurisdictional lines in order to better exploit markets for their products; they needed to be able to exploit resources in different jurisdictions, be they natural resources, pools of labor, sources of power or transport, or even pockets of capital. In addition to finding the elementary aspects of legal personality, such as the right to sue and to be sued, courts began to find other rights incidental to corporate existence. What is equally interesting is the tension between the legislature, which had the right to set the terms of a corporation's legal personality, and the courts, upon which fell the interpretive burden of discovering rights, which were incidental and through their discovery legitimating those rights. To be sure, legislatures often resisted granting certain rights and American courts moved in fits and starts in recognizing incidental rights, but the direction was clear. Marshall's qualification was, therefore, one of the ways in which the American jurisprudential image of the corporation reversed itself.

Well before the end of the nineteenth century, corporations asserted and courts recognized their changing legal personality. Though the object was clear, the protection of the property interests of the owners of corporations, the corporation's malleable legal personality was deployed as a jurisprudential weapon. At the time of the Dartmouth College judgment, corporations were rare, the state's imprimatur was thought to be useful in aggregating capital for risky and large ventures, and legislative consideration of corporate charters was real. A fictional understanding of the corporation's legal personality made sense, both to the state and to the individuals who sought a charter. By the last quarter of the nineteenth century, however, none of those conditions obtained. Corporations, especially business corporations, were common. Organizing entrepreneurial activity under corporate auspices was increasingly the norm. Capital markets had advanced to the point where the state's grant was largely irrelevant to the capacity to obtain capital and, as important, the state's stamp of approval was no longer felt to be necessary. Regulation of the behavior of corporations through their charters was largely nominal and the granting of charters was becoming an administrative rather than a legislative process. Indeed, the politics of regulation, always present, grew more contentious as the century progressed, especially as corporate activity made state and local regulation problematic. Regulation of behavior, regardless of legal form, increasingly became the norm and owners of corporations, in turn, sought to protect their property through new conceptions of corporate legal personality.

American legal treatises concerned with corporate law at first edged in the direction of abandoning the conception of the corporation as an entity with a legal personality of its own and instead made a brief attempt to see the corporation as the aggregation of the property of individuals and protected as would be the individual property of those individuals. Decades of jurisprudence dealing with corporate personality, the separation of owner and entity powerfully reinforced by the limited liability of corporate shareholders, and the willingness of courts to find an ever greater sphere of corporate rights suggested a different path. By the end of the century American jurisprudence, while still paying verbal homage to Marshall's understanding, had, in fact, begun to assume that many aspects

of a corporation's legal personality were inherent. Ironically, the American jurisprudence of corporate personality, which proved to be politically conservative, fed on many aspects of progressive jurisprudence. Maitland's writings on personification of the corporation, for example, though designed to help the aspirations of English labor as part of a larger jurisprudential vision, abetted an understanding of the naturalness of corporate rights in the United States. Indeed, the use of the term legal person to describe the corporation carried with it strong implications of a legal personality the equivalent of other persons.

In American jurisprudence the debate over the personification of the corporation exhausted itself in the years immediately prior to the World War II. In part the business corporation had achieved a legal personality sufficient for its needs and the needs of its owners. Further dispute on the reality and inherent content of a corporation's legal personality seemed irrelevant. Corporations existed and their legal rights and duties gave them sufficient flexibility to operate, flourish, and protect themselves. Simultaneously, attacks on general conceptions of rights by legal realists and others in favor of concepts of legal interests that society ought to recognize turned out to be just as hospitable to corporations in an environment that valued economic productivity more highly than much else. Nonetheless, the instrumental value of asserting an inherent legal personality was far from lost. In the last 1980s, for example, the United States Supreme Court recognized free speech rights of corporate bodies premised on the First Amendment of the Constitution, going so far as to characterize as 'extreme' the position that corporations have only the rights granted to them by states.

Legal personality, as the history of corporate personality demonstrates, is a jurisprudential tool with an independent significance. Without a legal personality, no person, entity, or thing may actively participate in a legal system. Once having obtained a legal personality, either by the assertion of an inherent personality and the recognition of that personality by the appropriate political bodies, or by the grant of personality from the sovereign, the content of the personality is plastic but the capacity to participate, even if limited, is clear. By right of participation, the avenue to enlarging the content of that personality opens itself up to assertions of right and instrumental value.

See also: Children and Law; Equity in Law; Gender and the Law; Slaves and Slavery, History of.

Bibliography

Dewey, J., 1926. The historic background of corporate legal personality. Yale Law Journal 35, 655–673.

Gierke, O., 1900. Political Theories of the Middle Ages. Cambridge University Press, Cambridge, UK.

Hallis, F., 1930. Corporate Personality. Oxford University Press, London.

Hurst, J., 1970. The Legitimacy of the Business Corporation in the Law of the United States. University Press of Virginia, Charlottesville, VA.

Laski, H., 1916. The personality of associations. Harvard Law Review 29, 404–426.

Maitland, F., 1911. Moral Personality and Legal Personality. The Collected Papers of Frederic William Maitland. Cambridge University Press, Cambridge, UK.

Mark, G., 1987. The personification of the business corporation in American Law. University of Chicago Law Review 54, 1441–1483.

Nekam, A., 1938. The Personality Conception of the Legal Entity. Harvard University Press, Cambridge, MA.

Vinogradoff, P., 1924. Juridical persons. Columbia Law Review 24, 594–604.

Wolff, M., 1938. On the nature of legal persons. Law Quarterly Review 54, 494–521.

Legal Pluralism

John Griffiths, University of Groningen, Groningen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Legal pluralism' refers to the coexistence of more than one regulatory order in a society. The term can be used in two ways. Juridical legal pluralism refers to the recognition by state law of the 'customary law' of particular groups, as in the case of colonial legal pluralism. Empirical legal pluralism refers to the situation in which the behavior of an actor is in fact subject to more than one set of rules. This article deals with the history of the empirical concept of legal pluralism, theoretical controversy it has engendered, and its use in various areas of empirical research.

The concept of legal pluralism refers in the most general sense to the "coexistence of different normative orders within one socio-political space" (Benda-Beckmann, 1997: p. 1). While recognition of the phenomenon has a longer history (e.g., Ehrlich, 1913; Gurvitch, 1935), its modern name only became common currency in the late 1960s and early 1970s: *legal pluralism* (Schiller, 1965; Hooker, 1975); *rechtspluralismus* (Benda-Beckmann, 1970), *pluralisme juridique* (Gilissen, 1971).

The concept plays a key role in modern sociology and anthropology of law and, to a lesser degree, in related disciplines such as the economic analysis of law (e.g., Ellickson, 1991) and (legal) history (e.g., Thompson, 1991; Arthurs, 1985). It is the most important contribution that anthropology of law has made to legal theory, demanding attention to unresolved questions concerning the concept of 'law,' the sources of law, and the relationship of 'law' to the nation-state (e.g., Ingber, 1971; Twining, 2009; Woodman, 1996). It can also be used in legal analysis, for example, in comparative law and conflict of laws (e.g., Michaels, 2009), in connection with the role of 'customary law' in legal development (e.g., Sheleff, 1999) and of the 'cultural defense' in criminal law (e.g., Renteln, 2004). In recent years, it has become a feature of studies of globalization and a new *lex mercatoria* ('global legal pluralism' – Michaels, 2009).

Juridical and Empirical Legal Pluralism

The concept of legal pluralism has two related but distinct meanings, depending on whether it is being used for purposes of legal analysis or of social scientific description and empirical theory (Griffiths, 1986). 'Juridical' legal pluralism refers to a possible characteristic of formal 'law' (of the state or some similar institution): a way in which a legal order that is in principle unitary can deal with cultural and other sorts of diversity in its population by providing (or recognizing) distinct rules and procedures for specific groups – indigenous peoples, ethnic and religious groups, local communities, merchants and other economic groups, etc. (Vanderlinden, 1971). In the case of juridical legal pluralism, there are different legal rules for different actors and situations, but (in principle) only one legal rule for a given actor in a particular situation. As F. von Benda-Beckmann (1997: pp. 1–2) points out, it was largely lawyers – scholars of Indonesian 'adat law' (of whom the most important is Vollenhoven, 1918–1933), and of African law (e.g., Allott, 1960, 1970; Morris and Read, 1972) – who laid the foundation for the later work of other lawyers (e.g., Vanderlinden, 1971; Hooker, 1975) – who in turn brought the term 'legal pluralism' into currency and made it available for anthropologists and others to use, albeit in a quite different sense.

That quite different sense is 'empirical' legal pluralism, which refers to a social state of affairs in which the behavior of an actor is subject to different and possibly conflicting behavioral expectations (e.g., when the rules that obtain among the participants in a particular economic activity differ from those of state law). Twining (2010) refers to this conception of legal pluralism as 'social fact' legal pluralism. Legal pluralism in this sense is not normative but descriptive. Seen as 'people's law,' it can be the subject of social romanticism, but there is nothing necessarily 'good' about it. As K. von Benda-Beckmann (2001: p. 40) has observed: "Taking legal pluralism seriously … is not the same as endorsing every [local] rule …. To take it seriously is to acknowledge that it is there, that it affects people's behaviour, and that it also affects the way legislation is implemented."

The remainder of this article is largely devoted to legal pluralism in the sense most relevant for the social sciences: the empirical sense.

History of the Concept

Legal pluralism has presumably always existed in all societies of any complexity. But recognition of the phenomenon as something deserving of specific attention required the rise of the modern nation-state and the notion that the state is the source of all law, properly so-called, and that law prevails over all other forms of regulation (see Ingber, 1971; see for the history of the concept of legal pluralism, Bergh, 1986; Rouland, 1988; Woodman, 1998; Michaels, 2009).

When Western European expansion began to take the form of colonization, this 'Westphalian' conception of law (Twining, 2009) confronted the existence of indigenous peoples with lawlike rules and institutions that could not simply be ignored. Largely for reasons of administrative convenience (e.g., 'indirect rule') and the 'ethical' concerns of enlightened lawyers like van Vollenhoven (1918–1933), the modern

conception of juridical legal pluralism as a form of governance was born. Colonial law incorporated by reference (some parts of) the preexisting 'laws and customs' of subjected peoples, rejecting others as 'repugnant' to morality and public order. Colonial legal pluralism brought forth a considerable body of law (legislation, case law, and legal doctrine) and an extensive legal literature (e.g., Morris and Read, 1972; Hooker, 1975).

The intellectual roots of the empirical conception of legal pluralism now prevalent in sociology and anthropology of law can be traced back to Ehrlich (1913) who insisted that the 'rules for behavior' that actually govern social interaction are the norms that constitute the 'inner order of the associations' of which a society consists – its 'living law' – whereas the law of the state consists of 'rules for decisions' addressed to courts and other legal decision makers, and often quite remote from everyday social practice. In his day, Ehrlich had little influence on either legal theory or sociology and anthropology. Gurvitch's (1935) pluralist theory of the sources of positive law – similar to Ehrlich in its insistence on "l'existence d'innombrables centres générateurs du droit, de foyers autonomes du droit [such as unions and factories, churches, decentralized public services, and so forth]" – met a similar fate.

But after the Second World War and the demise of Western colonialism, the gradual emancipation of anthropology from its focus on the colonial situation led to a fresh start in which 'legal pluralism' served as the rallying cry for a new approach to the study of law in society. Vanderlinden's essay '*Le pluralisme juridique: essai de synthèse*' marked the transition from preoccupation with the relationship between colonial law and local customary law to the realization that legal pluralism in the empirical sense is also characteristic of modern, industrialized societies. Vanderlinden first defined legal pluralism as "L'existence, au sein d'une société déterminée, de mécanismes juridiques différents s'appliquant à des situations identiques" (1971: p. 19). He later argued that it can better be defined with reference to the situation of the actor whose behavior is "governed by multiple and various regulatory orders" (1989: p. 153).

The conceptual apparatus required for further theoretical and empirical development was not complete with an encompassing conception of 'law' and an actor-centered conception of 'legal pluralism.' Discussions of the subject tended to assume that the entities that produce and enforce 'law' are 'corporate groups' such as the state, the Roman Catholic Church, business corporations, or politically organized 'tribes.' It was Moore (1973) who argued that the social locus of 'law' should be defined not in organizational but in processual terms: the generation and maintenance of rules of behavior. The social entities that do this she called 'semiautonomous social fields' and, however cumbersome the name she gave them (compare Homans', 1950, for an equivalent term, 'human group'), Moore's conception has proven highly influential in the subsequent history of the empirical conception of legal pluralism (cf Merry, 1988: p. 878).

Armed with such ideas about the nature of 'law in society,' legal pluralism came to seem ubiquitous: careful empirical study revealed it in the business world (Moore, 1973; Macaulay, 1963; Bernstein, 1992), in industrial and bureaucratic organizations (Macaulay, 1986), among rural neighbors (Ellickson, 1991), and so forth. Recently, the supranational arena (the European Union, international organizations, international trade, and economic globalization) has come to be recognized as an important locus of legal pluralism (see Michaels, 2009; Snyder, 1999).

In the 1970s, scholarly interest in legal pluralism became institutionalized. The most important international scholarly journal devoted to the field – the *Journal of Legal Pluralism* – was founded in 1969 as *African Law Studies*, reflecting in its name the still dominant influence of the juridical approach to (post) colonial law. By 1981, the empirical approach had become dominant, and the name was changed to reflect this. In 1978, the Commission on Folk Law and Legal Pluralism of the International Union of Anthropological and Ethnological Sciences was founded. It organizes regular international symposia, publishes books and proceedings, and since 2013 is the publisher (in cooperation with Taylor & Francis) of the *Journal of Legal Pluralism*.

The whole range of approaches to be found in the social sciences have their champions in the literature on legal pluralism: for example, feminism (e.g., Griffiths, 1997), Marxism (e.g., Fitzpatrick, 1980; Snyder, 1981b), and systems theory (e.g., Teubner, 1997). Most writing on legal pluralism has been done by Western scholars, but Chiba (1989) is a notable exception.

Theoretical Controversy

The juridical conception of legal pluralism – denoting a particular arrangement within a legal order (in particular, state law) – is generally accepted and has never been the subject of significant disagreement. The empirical conception, by contrast, has given rise to some controversy, in particular concerning the question whether 'law' can be said to exist separately from the state (or whether state 'law' and other 'laws' are fundamentally different) and whether it is possible to define 'law' in a purely empirical way (that is, without normative assumptions). See Tamanaha (1993, 2000), Roberts (1998), Woodman (1998), F von Benda-Beckmann (2002); Twining (2010). J. Griffiths (2005) has argued that the 'taxonomic' approach to the definition of law (as one type of the genus 'social control') is the root of the problem of an empirical concept of law and proposes to conceptualize 'law,' for social scientific purposes, not as a distinct type, but rather as a dimension of continuous variation in the degree of differentiation in social control.

Use of the Empirical Concept in Theory and Research

Legal Eclecticism

Colonial legal pluralism was characterized by the 'choice of law' approach: either state law or customary law applied to a given case, and if customary law was applicable, then it was either the law of group A or that of group B. Scholars working with the empirical conception of legal pluralism soon came to appreciate that actors with access to multiple sources of law construct relationships, conduct disputes, and form complex social institutions such as 'family,' 'property,' and 'international trade' eclectically, fashioning a patchwork composed of

bits and pieces of local and mercantile law, religious law, state law, and international law. Thus, property and inheritance relationships in Minangkabau combine elements of Indonesian law, local adat law, and Islamic law (cf Benda-Beckman, 1979) and the social security arrangements that people in developing countries construct for themselves include elements of state law and of customary and religious laws, and if they migrate to a Western country its formal social security arrangements get included in the mixture (Benda-Beckman et al., 1988). Similar legal eclecticism can be observed in the examples of legal pluralism in 'modern' and 'Western' settings mentioned in Section History of the Concept above.

Dispute Processes (Litigation and Adjudication)

One of the most active areas of empirical research and theory in which legal pluralism has played a central role concerns dispute processes. Legal pluralism is an important factor in understanding what people have disputes about, how disputes emerge in social life, what choices are made about how to deal with them, how they progress, what ultimately happens with them, and how the disposition of a particular dispute affects the way other people deal with similar problems. At all stages of a dispute, the existence of alternative norms, procedures, and institutions profoundly influences what happens. Exemplary studies are Santos (1977) on informal law and disputing in a Brazilian squatter settlement, Abel (1979) on patterns of court use in a situation of legal pluralism, Todd (1978) on disputing relationships and social status in a Bavarian village, Ellickson (1991) on cattle-trespass disputes between neighbors in rural California. K. von Benda-Beckmann (1984) has shown how 'forum-shopping' litigants make differential use of state and nonstate laws in different phases of a dispute, so that the execution of a court judgment, for example, may be regulated in practice by nonstate norms.

Abel (1974), Galanter (1981), and J. Griffiths (1983) have integrated the perspective of legal pluralism into general formulations of litigation theory.

Social Working

Research in both Western and non-Western settings has shown repeatedly that legal pluralism and the eclectic use of plural legal resources can be key to understanding the effects of the rules of state law (see, e.g., Collier, 1976). Until recently, for example, proponents of 'law and development' tended to see legal pluralism as an obstacle to efforts of the state to promote economic development. However, the common idea that the relationship between local law and state law is that the former generally frustrates the effectiveness of the latter is too simple (cf Tamanaha et al., 2012). As Moore (1973: p. 721) observed long ago, "the various processes that make informally generated rules effective are often also the immediate forces that dictate the mode of compliance or non-compliance to state-made legal rules" (for a striking example see Kagan and Skolnick, 1993 for the case of rules regulating smoking). State law often adapts and builds upon forms of regulation that have earlier emerged in, and are enforced by, local social fields. The Dutch experience with the regulation of medical practice surrounding death (Griffiths, 2000) is a good example of the process by which such self-regulation (in this case, by the medical profession) can be adopted as 'law' by legal institutions, but remain dependent on the social field of doctors for its interpretation and application in medical practice.

J. Griffiths has integrated the perspective of legal pluralism into a general theory of the 'social working' of legal rules, of which the key proposition is that "[I]t is not the legislator who determines when a legal rule will be followed in social interaction. It is society that determines when and to what extent it is regulated by law. And it does this in a highly self-regulated way" (Griffiths, 2003: p. 66).

Common Property

In a famous article dealing with the problem of the 'population explosion,' Hardin (1968) invoked the 'tragedy of the commons,' the supposed failure of the old institution of village commons to prevent overgrazing and consequent land degradation, although it would have been in the interest of everyone to limit grazing to the carrying capacity of the land. Hardin's Hobbesian argument was sound; collective goods such as common property, clean air, or a stable population cannot be maintained without regulation. But his reliance on the example of the village commons was misplaced, since in that case use-rights were recognized and effectively maintained over long periods. His mistake lay in confusing common property with 'open (that is, unregulated) access' and overlooking nonstate regulation and enforcement (cf Feeny et al., 1990).

In recent years, much attention has been paid to the maintenance, by nonstate regulation of access, of collective goods such as irrigation systems (e.g., Roth, 2011), hunting and fishery resources (e.g., Ginkel, 1989; Bavinck, 1996), and so forth (see generally Acheson, 2011).

Interaction of State and Nonstate Laws and Institutions

A fifth area of interest concerns the ways in which a situation of legal pluralism affects the various legal orders involved. A well-known example is the 'creation of customary law': the fact that when local law is 'recognized' by state law (as in the case of colonial legal pluralism), it must be ascertained and applied by noncustomary tribunals. In the process, a great deal of transformation takes place, so that the 'customary law' applied in state courts may have only a distant relationship to its supposed sources (Snyder, 1981a; Woodman, 1988). Although it has received less attention, the opposite transformation – the 'creation of state law' in 'indigenous' interactions – seems to be at least as important (cf Benda-Beckmann, 1989; Griffiths, 1999).

A related phenomenon is the effect on local political authority when it is incorporated into a (post)colonial system of government. Such incorporation brings with it notorious errors in the identification of indigenous authorities, the witting or unwitting attribution to them of various nonindigenous powers, the manipulation of indigenous processes of succession to high office, and so forth. The fundamental bases of legitimacy tend to be corrupted and indigenous mechanisms of control over political authority eroded. This is a central

theme in a number of studies of local political authority (e.g., Holleman, 1969; Owusu, 1970).

On the other hand, the relationship between state law and other 'regulatory orders' (to use Vanderlinden's expression) is not necessarily just one of reciprocal confusion and frustration. The literature on effective 'governance' (e.g., of fisheries and coastal systems) increasingly recognizes the potential importance of 'interactive governance,' in which the state enhances governability by working in tandem rather than at cross-purposes with other sources of binding rules (see Jentoft, 2011).

In South Africa, where section 211 of the Constitution recognizes customary law and 'traditional leadership' so long as they are consistent with the Constitution and with legislation that specifically deals with customary law, recent years have seen fascinating legal developments in the way the state and the courts deal with customary law. In the *Shilubana* case (2008 (9) BCLR 914 (CC)), the South African Constitutional Court held that the authorities of a 'traditional community' are empowered to make changes in its customary law, in particular to bring it into conformity with the requirements of the Constitution (the case involved a change in the eligibility rules for succession to the position of Chief, so that a woman could succeed). *En passant* the Court observed that the common law rules for identifying legally binding 'custom' do not apply to customary law in South Africa; customary law is 'living law' of the community concerned and current practice is more important to the legal status of an alleged rule than its historical pedigree. The Court's judgment refers extensively to earlier decisions.

Surveys

Useful surveys of the literature on legal pluralism include Schiller (1965), Merry (1988), Rouland (1988), Woodman (1998), Michaels (2009), and Twining (2010).

See also: Conflict and Conflict Resolution, Social Psychology of; Cultural Relativism, Anthropology of; Disputes, Social Construction and Transformation of; East Asian Legal Traditions; Federalism; Law and Development; Law and Society: Development of the Field; Law as an Instrument of Social Change; Law, Mobilization of; Laws, Conflict of; Legal Culture and Legal Consciousness; Legal Positivism; Legal Systems, Classification of; Legal Systems: Private; Legal Traditions of Southeast Asia; Legalization: The Expansion of Law's Domain; Lex Mercatoria; Litigation; Middle Eastern and North African Legal Traditions; Pluralism and Tolerance; Pluralism; Relativism: Philosophical Aspects; Remedies and Damages: Legal Aspects; Sub-Saharan African Legal Traditions; Value Pluralism.

Bibliography

Abel, R., 1974. A comparative theory of dispute institutions in society. Law and Society Review 8, 217–347.

Abel, R., 1979. Western courts in non-western settings: patterns of court use in colonial and neo-colonial Africa. In: Burman, S., Harrell-Bond, B. (Eds.), The Imposition of Law. Academic Press, New York, pp. 167–200.

Acheson, J., 2011. Ostrom for anthropologists. International Journal of the Commons 5, 319–339.

Allott, A., 1960, 1970. Essays in African Law and New Essays in African Law. Butterworth, London.

Arthurs, H., 1985. Without the Law. Administrative Justice and Legal Pluralism in Nineteenth-Century England. University of Toronto Press, Toronto.

Bavinck, M., 1996. Fisher regulations along the Coromandel coast: a case of collective control of common pool resources. Marine Policy 20, 475–482.

Benda-Beckmann, F. von, 1970. Rechtspluralismus in Malawi. Weltforum Verlag, Munich (published in English in 2007: Legal Pluralism in Malawi, Kachere Series, Zomba, Malawi).

Benda-Beckmann, F. von, 1979. Continuity and Change in the Maintenance of Property Relationships through Time in Minangkabau, West Sumatra. Martinus Nijhoff, The Hague, Netherlands. Property in Social Continuity.

Benda-Beckmann, F. von, 1989. Scapegoat and magic charm: law in development theory and practice. Journal of Legal Pluralism 28, 129–148.

Benda-Beckmann, F. von, 1997. Citizens, strangers and indigenous peoples: conceptual politics and legal pluralism. Law and Anthropology 9, 1–42.

Benda-Beckmann, F. von, 2002. Who's afraid of legal pluralism? Journal of Legal Pluralism 47, 37–82.

Benda-Beckmann, F. von, et al., 1988. Between Kinship and the State: Social Security and Law in Developing Countries. Foris, Dordrecht, The Netherlands.

Benda-Beckmann, K. von, 1984. The Broken Stairways to Consensus: Village Justice and State Courts in Minangkabau. ICG, Dordrecht, Netherlands.

Benda-Beckmann, K. von, 2001. Legal pluralism. Tai Culture: International Review of Tai Cultural Studies 7 (1 & 2), 18–40.

Bergh, G. van den, 1986. The concept of folk law in historical context: a brief outline. In: von Benda-Beckmann, K., Strijbosch, F. (Eds.), Anthropology of Law in the Netherlands: Essays on Legal Pluralism. Foris, Dordrecht, The Netherlands, pp. 67–89.

Bernstein, L., 1992. Opting out of the legal system: extralegal contractual relations in the diamond industry. Journal of Legal Studies 21, 115–157.

Chiba, M., 1989. Legal Pluralism: Toward a General Theory through Japanese Legal Culture. Tokai University Press, Tokyo.

Collier, J., 1976. Political leadership and legal change in Zinacantan. Law and Society Review 11, 131–163.

Ehrlich, E., 1913. Grundlegung der Soziologie des Rechts. Duncker & Humblot, Munich (fourth ed. 1989, Rehbinder, M. Ed., Dunker & Humblot, Berlin; English translation 1936, Fundamental Principles of the Sociology of Law, Harvard University Press, Cambridge, MA).

Ellickson, R., 1991. Order Without Law. How Neighbors Settle Disputes. Harvard University Press, Cambridge, MA.

Fitzpatrick, P., 1980. Law and State in Papua New Guinea. Academic Press, London.

Feeny, D., et al., 1990. The tragedy of the commons: twenty-two years later. Human Ecology 18, 1–19.

Galanter, M., 1981. Justice in many rooms: courts, private ordering and indigenous law. Journal of Legal Pluralism 19, 1–47.

Gilissen, J. (Ed.), 1971. Le Pluralism Juridique. Editions de l'Université de Bruxelles, Brussels.

Ginkel, R. van, 1989. Plunderers into planters: Zeeland oystermen and the enclosure of the marine commons. In: Boissevain, J., Verrips, J. (Eds.), Dutch Dilemmas: Anthropologists Look at the Netherlands. Van Gorcum, Assen, Netherlands, pp. 89–105.

Griffiths, A., 1997. In the Shadow of Marriage: Gender and Justice in an African Community. University of Chicago Press, Chicago.

Griffiths, J., 1983. The general theory of litigation—a first step. Zeitschrift für Rechtssoziologie 5, 145–201.

Griffiths, J., 1986. What is legal pluralism? Journal of Legal Pluralism 24, 1–50.

Griffiths, J., 1999. Legal knowledge and the social working of law: the case of euthanasia. In: van Schooten, H. (Ed.), Semiotics and Legislation. Jurisprudential, Institutional and Sociological Perspectives. Deborah Charles Publications, Liverpool, pp. 81–108.

Griffiths, J., 2000. Self-regulation by the medical profession of medical behavior that potentially shortens life. In: Krabbendam, J., ten Napel, H.-M. (Eds.), Regulating Morality: A Comparison of the Role of the State in Mastering the Mores in the Netherlands and the United States. Maklu, Antwerp, pp. 173–190.

Griffiths, J., 2003. The social working of legal rules. Journal of Legal Pluralism 48, 1–84.

Griffiths, J., 2005. The idea of sociology of law and its relation to law and to sociology. In: Freedman, M. (Ed.), Law and Sociology. Oxford University Press, Oxford, pp. 49–68.

Gurvitch, G., 1935. L'Expérience juridique et la philosophie pluraliste du droit. In: Pedone, A. (Ed.), Librarie de la Cour d'Appel et de l'Orde des Avocats. Paris.

Hardin, G., 1968. The tragedy of the commons. Science 162, 1234–1248.

Holleman, J., 1969. Chief, Council and Commissioner. Van Gorcum, Assen, Netherlands.
Hooker, M., 1975. Legal Pluralism: An Introduction to Colonial and Neo-Colonial Laws. Clarendon Press, Oxford.
Ingber, L., 1971. Le pluralisme juridique dans l'œuvre des philosophes du droit. In: Gilissen, J. (Ed.), Le Pluralisme Juridique. Editions de l'Université de Bruxelles, Brussels, pp. 57–84.
Jentoft, S., 2011. Legal pluralism and the governability of fisheries and coastal systems. Journal of Legal Pluralism 64, 149–172.
Kagan, R., Skolnick, J., 1993. Banning smoking: compliance without enforcement. In: Rubin, R., Sugarman, S. (Eds.), Smoking Policy: Law, Politics and Culture. Oxford University Press, Oxford, pp. 69–94.
Macaulay, S., 1963. Non-contractual relations in business: a preliminary study. American Sociological Review 28, 55.
Macaulay, S., 1986. Private government. In: Lipson, L., Wheeler, S. (Eds.), Law and the Social Sciences. Russell Sage Foundation, New York, pp. 445–518.
Merry, S., 1988. Legal pluralism. Law and Society Review 22, 869–896.
Michaels, R., 2009. Global legal pluralism. Annual Review of Law and Social Science 2009, 243–262.
Moore, S., 1973. Law and social change: the semi-autonomous social field as an appropriate subject of study. Law and Society Review 7, 719–746.
Morris, H., Read, J., 1972. Indirect Rule and the Search for Justice: Essays in East African Legal History. Oxford University Press, Oxford.
Owusu, M., 1970. The Uses and Abuses of Political Power. A Case Study of Continuity and Change in the Politics of Ghana. University of Chicago Press, Chicago.
Renteln, A., 2004. The Cultural Defense. Oxford University Press, New York.
Roberts, S., 1998. Against legal pluralism. Some reflections on the contemporary enlargement of the legal domain. Journal of Legal Pluralism 42, 95–106.
Roth, D., 2011. The subak in diaspora: Balinese farmers and the subak in South Sulawesi. Human Ecology 39, 55–68.
Rouland, N., 1988. Anthropologie Juridique. Presses Universitaires de France, Paris (Planel, P., Legal Anthropology. The Athlone Press, London, 1994, Trans.).
Santos, B. de Souza, 1977. The law of the oppressed: the construction and reproduction of legality in Pasagarda. Law and Society Review 12, 5–126.
Schiller, A., 1965. Law. In: Lystad, R. (Ed.), The African World: A Survey of Social Research. Praeger, New York, pp. 165–198, 512–524.
Sheleff, L., 1999. The Future of Tradition: Customary Law, Common Law and Legal Pluralism. Frank Cass, London and Portland, Oregon.
Snyder, F., 1981a. Colonialism and legal form: the creation of 'customary law' in Senegal. Journal of Legal Pluralism 19, 49–90.
Snyder, F., 1981b. Capitalism and Legal Change: An African Transformation. Academic Press, New York.
Snyder, F., 1999. Governing economic globalization: global legal pluralism and European law. European Law Journal 5, 343–374.
Tamanaha, B., 1993. The folly of the social scientific concept of legal pluralism. Journal of Law and Society 20, 192–217.
Tamanaha, B., 2000. A non-essentialist version of legal pluralism. Journal of Law and Society 27, 296–321.
Tamanaha, B., Sage, C., Woolcock, M., 2012. Legal Pluralism and Development, Scholars and Practitioners in Dialogue. Cambridge University Press, Cambridge.
Teubner, G., 1997. Global Bukowina: legal pluralism in the world society. In: Teubner, G. (Ed.), Global Law Without a State. Dartmouth, Aldershot, UK, pp. 3–28.
Thompson, E., 1991. Customs in Common. Merlin Press, London.
Todd, H., 1978. Litigious marginals. Character and disputing in a Bavarian village. In: Nader, L., Todd, H. (Eds.), The Disputing Process—Law in Ten Societies. Columbia University Press, New York, pp. 86–121.
Twining, W., 2009. General Jurisprudence. Understanding Law from a Global Perspective. Cambridge University Press, Cambridge.
Twining, W., 2010. Normative and legal pluralism: a global perspective. Duke Journal of Comparative and International Law 20, 473–517.
Vanderlinden, J., 1971. Le pluralism juridique: essai de synthèse. In: Gilissen, J. (Ed.), Le Pluralisme Juridique. Université de Bruxelles, Brussels, pp. 19–56.
Vanderlinden, J., 1989. Return to legal pluralism: twenty years later. Journal of Legal Pluralism 28, 149–157.
Vollenhoven, C. van, 1918–1933. Het Adatrecht Van Nederlands-indië. Brill, Leiden (Translated and edited by Holleman, J., 1981. Van Vollenhoven on Indonesian Adat Law. Martinus Nijhoff, the Hague).
Woodman, G., 1988. How state courts create customary law in Ghana and Nigeria. In: Morse, B., Woodman, G. (Eds.), Indigenous Law and the State. Foris, Dordrecht, Netherlands, pp. 181–220.
Woodman, G., 1996. Legal pluralism and the search for justice. Journal of African Law 40, 152–167.
Woodman, G., 1998. Ideological combat and social observation: recent debate about legal pluralism. Journal of Legal Pluralism 42, 21–59.

Relevant Websites

www.commission-on-legal-pluralism.com – 'Commission on Legal Pluralism'.
www.jlp.bham.ac.uk – 'Journal of Legal Pluralism' [full text of back issues].

Legal Positivism

Brian Z Tamanaha, Washington University in St. Louis, St. Louis, MO, USA

Published by Elsevier Ltd.

Abstract

Legal positivism is the label attached to a widely accepted, although hotly disputed, theory about the nature of law. This article describes how legal positivism developed by way of contrast to natural law theories of law. The core elements of the legal positivist view are then set out, with a discussion of the main disputes among legal positivists and the main criticisms of legal positivism. The final section discusses the relationship between legal positivism and social scientific approaches to law.

Legal positivism is the label attached to a widely accepted, although hotly disputed, theory about the nature of law. Within jurisprudence, legal positivism is most often discussed by way of contrast to natural law theories of law. The contest between these rival traditions has long dominated the discussion in legal philosophy, although there are signs that the debate may have run its course. Because legal positivism originated as a reaction to natural law theories about law, the best way to understand legal positivism is through this debate. The following section will describe the nature of the debate; the section thereafter will set out the core elements of legal positivism; and the final section will discuss the reaction to, and significance of, legal positivism, from the standpoint of social scientific approaches to law.

The Debate between Legal Positivism and Natural Law

Ironically, Hobbes, who is widely credited to be the early progenitor of legal positivism, built his positivistic theory of law on a natural law base (Bobbio, 1993; Fuller, 1940). According to Hobbes, the necessity for peace and order is the fundamental precept of natural law. Given that the state of nature is, in his immortal phrase, "solitary, poore, nasty, brutish, and short", people join together and give over law-making and enforcing power to a Sovereign for the purpose of maintaining peace and order (Hobbes, 1946: Chaps. XIV, XV). Thereafter, law consists of the commands of the Sovereign. Although the Sovereign is to be guided by reason in the establishment of law, even unjust commands are valid laws; to suggest otherwise would invite the disorder that will result if people were allowed to apply their own judgment to evaluate the law. Under Hobbes' formulation, while natural law provides the initial foundation for positive law, once that foundation is established, natural law becomes irrelevant. It does not provide a continuing standard or test for the validity of the law (*see* Natural Law; Justice and Law; Injustice: Legal Aspects).

This conclusion raises the issue that has long set legal positivists and natural law theorists at odds. Thomas Aquinas, citing Augustine, asserted that "A law that is not just, seems to be no law at all" (Q. 96, A. 4, Obj. 3). "Consequently every human law has just so much of the nature of law, as it is derived from the law of nature. But if in any point it deflects from the law of nature, it is no longer a law but a perversion of law" (Q. 95, A. 2, Obj. 4). A similar statement was made by William Blackstone. "[N]o human laws are of any validity, if contrary to this [law of nature]" (Boorstin, 1996: p. 49). More recently, Lon Fuller asserted that there is a minimum moral element to law, without which it is "not simply bad law, but not law at all" (Boorstin, 1999: p. 197).

Legal positivists are adamantly opposed to this position. Austin, who produced the first widely influential articulation of the legal positivist position, wrote a scathing response to Blackstone. "Now, to say that human laws which conflict with the Divine law are not binding, that is to say, are not laws, is to talk stark nonsense. The most pernicious laws, and therefore those which are most opposed to the will of God, have been and continually are enforced as laws by judicial tribunals" (Austin, 1995: p. 158). "The existence of law is one thing; its merit or demerit is another." What law *is*, and what law *ought* to be, legal positivists insist, are separate questions. This is known as the separation thesis.

Aside from the general confusion it engenders, the failure to maintain the separation between *is* and *ought* raises two particular political concerns for legal positivists (Hart, 1958: pp. 597–598). The first concern involves a person who concludes, based on his or her understanding of natural law, that a particular immoral law is invalid and, therefore, can be freely disregarded. An example of this is the radical antiabortionist who invokes natural law to justify shooting doctors who perform abortions on the grounds that laws allowing abortion are invalid because they condone the killing of innocent lives. The second concern is that the identification of law with morality will lead people to be less critical of the law, their vigilance softened by the seductive assumption that the law *is* what it *ought* to be. Many legal positivists believe that the fact that a norm has valid legal status does not of itself give rise to an obligation to abide by the law (Schauer, 1996). Keeping apart law and morality in this fashion allows us to more critically evaluate the law from a moral standpoint. These two concerns, while not contradictory, point in opposite directions: the first worries about people choosing to disregard the law on the grounds that it is invalid, while the second worries about people slavishly following the law because they assume it is necessarily morally worthy.

The second concern, more so than the first, animates many legal positivists in their dispute with natural law theorists, and their concern is not a purely abstract one. In the aftermath of World War II, some legal theorists blamed the legal positivist

separation of law and morality for the failure of the German legal establishment to resist atrocities committed by the Nazi government through law (Hart, 1958: pp. 615–621; Rommen, 1998). The view that 'law is law' regardless of its immorality, according to Gustav Radbruch, led the German legal profession to be completely subservient to state legal power. The legal positivist response – besides pointing out the lack of empirical support for the asserted causal connection – is that the separation of law from morality has the effect of *reducing* the prestige of law, thereby exposing it to greater moral critique. Under this view, the fact that a pronouncement is valid law carries little weight, and an immoral law should be resisted. If the German legal establishment failed to resist immoral laws, therefore, it was not the fault of legal positivism.

The longstanding debate between legal positivists and natural law theorists is a narrow one. Legal positivists argue that a law can count as valid law, even when it is immoral. Nevertheless, they immediately assert that a valid law or legal system can and should be evaluated from a moral standpoint, and not considered obligatory if found wanting. Many natural law theorists, on their part, have asserted that consistency with natural law is a condition of the validity of positive law. An immoral law is, therefore, not 'law' at all, and carries no obligation (except perhaps when a greater good might be served by adherence to a bad law). However, anything that qualifies as law pursuant to this understanding automatically generates an obligation to comply because it has already met a moral test. Whichever direction one approaches from, in the end, both sides agree that an immoral positive law generally does not give rise to an obligation to obey.

After generations of disagreement, there are a few signs of a rapprochement. John Finnis, the most influential modern exponent of natural law theory, argued that Aquinas was actually an early legal positivist, in the sense that he recognized that immoral laws were still in fact 'laws,' albeit deviant or deficient forms of law (Finnis, 1980, 1996). A leading legal positivist, Neil MacCormick (1992), has acknowledged that there are certain respects in which law or legality contains moral qualities, although this does not vitiate the separation thesis. If natural law theorists, like Finnis, are willing to acknowledge that an immoral law is still 'law' in positive law terms – a point that a number of natural law theorists have not actually contested – the original basis for the dispute dissipates (although a different opposition arises in relation to Dworkin, addressed shortly).

To be sure, not everyone agrees that the debate is over, but the bulk of the discussion of late has shifted to other matters, away from the issue of legal validity. One aspect of this classical debate remains unresolved, however. Although it has attracted relatively little attention, it is of continuing relevance and bears mention. The primary focus of the debate has been on whether immoral laws are invalid; the reverse question this raises is whether natural law principles have the status of law even if not affirmatively recognized by positive law. If the answer is yes, then the Nuremberg Trials, and more recently the trials of the German border guards prosecuted for shooting people attempting to escape former East Germany, involved the application of preexisting natural law principles with direct legal effect. If the answer is no, then these trials can be condemned as impositions of power by the ultimate victors in a political battle. Furthermore, if the answer is no, a strong argument can be made that these trials themselves violated the natural law prohibition against the *ex post facto* imposition of law, criminalizing and punishing behavior that was not illegal at the time it was committed. The issue of the legitimacy and legality of these kinds of actions, with serious unresolved moral as well as practical implications, will continue to arise as long as there are attempts to subsequently hold government officials legally responsible for conduct that was not illegal under the positive law when they engaged in their actions.

The Core Elements of Legal Positivism

As with every theoretical tradition, there are many varieties of legal positivism. Of late, as the dispute between legal positivism and natural law has moderated, internal divisions within legal positivism have taken center stage in the theoretical discussion. The nuances of the internal debate need not be explored. It takes place within shared parameters that characterize most legal positivists, and these parameters are what matter. There have been two dominant versions of legal positivism in the twentieth century, the first by Hans Kelsen and the second by H.L.A. Hart. The essential difference between the theories lies in their orientation (Bix, 1996: Chap. 6). Kelsen constructed an abstract system of norms (Kelsen, 1967), whereas Hart built on social practices. The focus in this discussion will be on Hart's version, which has proven more influential (at least outside the European Continent), but a few words will be said about Kelsen's approach.

Kelsen's 'pure theory of law' focused exclusively on law as a system of norms – separate from issues of justice (natural law) and from how people actually behave in connection with law (sociological jurisprudence) (Kelsen, 1961). Law is a normative system that contains its own rules for the creation and application of norms, and prescribes coercive sanctions for violations of its norms. It is a hierarchical system in which higher norms govern the content and creation of lower norms, and unity of the legal order rests upon a basic norm or *Grundnorm*. Initially, Kelsen asserted that the basic norm that grounds the unity of the system must be *presupposed* – it cannot be enacted because that would require a higher authority, which threatens an infinite regress – but in later writings, he suggested that the idea of the basic norm is a *fiction* (Paulson, 1992). His theory merged the law and the state: "One of the distinctive results of the pure theory of law is its recognition that the coercive order which constitutes the political community we call a state, is a legal order" (Kelsen, 1961: p. 64). Although Kelsen has devoted supporters, the daunting complexity of his pure theory of law, combined with questions about whether his late shift in view gutted what made his theory unique, has limited the broader influence of his version of legal positivism.

Hart's *The Concept of Law* (1961) is the single most important legal positivist text, and probably the most influential book in the field of jurisprudence in the second half of the twentieth century. Hart's project was to produce "a theory of what law is which is both general and descriptive" (Hart, 1994: p. 239). His theory was 'general' in the sense that he intended

to clarify the nature of law as a social phenomenon found in many different cultures and societies, by way of contrast to a 'particular' jurisprudence, which is focused on a specific legal system. His text was 'descriptive' in the sense that he eschewed evaluative analysis, limiting himself to describing practices and clarifying concepts related to law. He characterized his standpoint as that of an external observer describing the ways participants engage in and understand law from an internal point of view. To emphasize his orientation, Hart described his book as "an essay in descriptive sociology" (Hart, 1961: p. v).

According to Hart, a legal system consists of the union of what he called primary and secondary rules. Primary rules are the rules that govern relations within society. Secondary rules are the rules legal officials follow to identify, apply, and change the primary rules. Hart labeled the core secondary rules the 'rules of recognition,' "which specify some feature or features possession of which by a suggested rule is taken as conclusive affirmative indication that it is a rule of the group to be supported by the social pressure it exerts" (Hart, 1961: p. 92). Norms created in accordance with the rules of recognition thereby obtain legal status. According to Hart, the emergence of secondary rules is what separates prelegal (or primitive) societies from societies with law. Although simple societies with only primary rules can exist, they are static and inefficient in responding to new situations or bringing about the desired change. The availability of secondary rules cures these defects.

In his dual division, secondary rules identify primary rules; the remaining question is how to identify secondary rules. Hart's answer to this – in contrast to Kelsen's *Grundnorm* as the ultimate source of all other norms – is that the rules of recognition can be identified by observing the *actual social practices* of legal officials when they identify law. The existence and nature of the rules of recognition in a given legal system is thus 'a matter of fact.' Law is whatever legal officials recognize as law in their social practices, with no content restrictions on what that might be.

Hart concluded that there are two minimum conditions necessary and sufficient for the existence of a legal system. The first condition is that the primary rules must be generally *obeyed* by the populace (it is not essential that the populace normatively embrace the primary rules, although that is ideally the case). It does not make much sense to say that a legal system exists if the bulk of the populace generally disregards it. The second condition is that the legal officials normatively accept the secondary rules as critical standards for their behavior. Hart added this requirement to distinguish law from mere habits of obedience, and from the 'Gunman Writ Law,' which was the impression left by John Austin's command theory of law. Hart felt that the coercive view of law failed to recognize that law contains a normative aspect. This normative aspect, according to Hart, can be recognized if one takes into consideration the internal point of view of legal officials. Seen from the internal perspective, legal officials *accept* the secondary rules as "common standards of official behavior and appraise critically their own and each other's deviations as lapses" (Hart, 1961: p. 113). Consistent with his nonevaluative stance, Hart specifies that officials must accept the secondary rules without suggesting whether this acceptance is merited (which can only be determined by examining particular legal regimes). When the primary rules are generally obeyed by the populace and the secondary rules are accepted by the legal officials, a legal system will exist.

With its simplicity, clarity, elegance, and scope, Hart's reconstruction revitalized legal positivism and spawned an enormous response, in support and against. Most legal positivists (in the Hartian tradition) adhere to two core theses about the nature of law: the Separation Thesis and the Social Sources Thesis. The Separation Thesis holds that there is no necessary relation between law and morality (Coleman, 1982, 1996). The genesis of this thesis has already been identified in the earlier described insistence of legal positivists that the fact of legality alone does not carry with it the imprimatur of moral rectitude. It is important to see that this is a limited, negative thesis, a denial of the natural law position, rather than an affirmative statement of the typical relationship between law and morality. Legal positivists do not contend that there are no connections between law and morality, only that there is no *necessary* connection. They recognize that the "law of every modern state shows at a thousand points the influence of both the accepted social morality and wider moral ideals" (Hart, 1961: p. 199). This influence comes in a myriad of ways. Law and legal norms often reflect (or claim to reflect) the morality of their community. Legal norms often expressly refer to and incorporate moral principles like justice, fairness, reasonableness, and due process. Law often has positive moral consequences, especially in maintaining order in society. Law also participates in the moral education of members of the community. Recognition of these important connections between law and morality is consistent with the Separation Thesis. Stated at the most abstract level, the Separation Thesis claims only that a legal system can be imagined to exist in which morality is not a condition of legality. At a more practical level, the Separation Thesis insists that it is not necessarily the case that a moral norm automatically becomes a legal norm, and that it is not necessarily the case that an immoral law is devoid of legal validity.

The Social Sources Thesis holds that what is or is not valid law in a given system is a social fact, as determined in accordance with the rules of recognition. By observing the practices of the legal officials within a legal system, it will be possible to identify the criteria applied by the legal officials to determine valid law. In many legal systems, these practices include treating legislative enactments and judicial precedents as valid sources of law.

Although legal positivists share these two theses, there is a substantial internal divergence among legal positivists about what they stand for and allow or disallow. Much of the internal disagreement has been generated by efforts by Hart and other legal positivists to respond to the sustained critique of legal positivism by Ronald Dworkin. This complex internal and external debate cannot be recounted here, although its core themes can be articulated summarily.

Dworkin has attacked Hart at several different levels and points (Dworkin, 1977, 1985, 1986). He denies the very possibility or value of general, purely descriptive jurisprudence. Dworkin insists that jurisprudence itself is an aspect or extension of the law of a particular legal system and must, therefore, be conducted from the particular, internal perspective, which is inherently evaluative in orientation. Dworkin denies the Separation Thesis by arguing that law and morality are

inextricably intertwined, with moral principles providing an integral backdrop for the interpretation and application of legal rules. Law incorporates a large body of the community's moral principles; with the inclusion of these background principles, he claims, law always provides a single correct answer to legal problems. Dworkin further argues that the social practice aspect of Hart's rules of recognition fails to accurately describe our own legal system, which expressly incorporates moral norms (like fairness) in a manner that renders the identification of the primary rules dependent on difficult and controversial questions about the truth of moral norms. He identifies two specific problems with Hart's notion of rules of recognition. First, rules of recognition are incapable of identifying all of the many and varied moral norms and principles incorporated into the law. Second, Dworkin insists that legal positivists, consistent with the requirements of their own theory of law, cannot accommodate the incorporation of such norms and principles for this reason: making law contingent upon the resolution of complex moral questions (as Dworkin says does occur) would defeat legal positivists' claims that the identification and meaning of legal rules can be determined in ways that avoid uncertainty and provide guidance to the populace.

This is an abbreviated recitation of Dworkin's criticisms of Hart and legal positivism, but it suffices for these purposes. Some of the issues separating Hart and Dworkin can be resolved, or at least avoided, by recognizing that Hart was focused on the legal system as a whole, while Dworkin focused primarily on judging, and that general and particular jurisprudence are two different enterprises which need not necessarily conflict. The most significant controversy relates to the extent to which rules of recognition can refer to open-ended moral questions to determine the content of law (rather than just social practices). Hart insists that there are no restrictions on rules of recognition that prohibit a particular legal system from turning the identification of law on the resolution of moral issues, as long as that is what the actual social practice in a given system requires. Hart labels his view 'soft positivism' (or 'inclusive' legal positivism) (Hart, 1994: pp. 250–254; Coleman, 1998). Against Dworkin's claim that this defeats the point of legal positivism, Hart (1994: p. 251) responds that positivism tolerates a certain degree of uncertainty, and resort to moral standards would not necessarily result in an excessive amount of uncertainty.

In theory, inclusive legal positivism would still be a correct description of a legal system that has only a single rule of recognition which says this: do justice (Coleman, 1998). Taken this far, it is no longer evident, as with the debate between natural law and legal positivism, that anything of great moment rests upon the resolution of the debate between inclusive legal positivists and Dworkin – for both allow the substantial incorporation of moral principles (albeit construed differently). Contrary to Hart and other inclusive legal positivists, however, there are 'hard' or 'exclusive' legal positivists, led by Joseph Raz, who do not concede Dworkin's position about the inclusion of open questions of morality within law (Raz, 1979). Now much of the debate is an internal one between inclusive and exclusive legal positivists (Bix, 1999; Leiter, 1998; Coleman, 1998). Although hard positivists recognize that the law often refers to such norms, they assert that at that moment the law runs out and the judge's decision is based on extralegal considerations. The moral considerations to be resolved are not a part of the preexisting law before the issue is decided, although they can thereafter become law if recognized and treated as legal precedent (consistent with the applicable rules of recognition). Hard positivists believe it is essential to take this position if legal positivism is to satisfy its claims that the function of law is to provide guidance to the public on the rules of conduct, and if law is to claim independent authority, apart from morality.

Once again, it would appear that the issue that divides the disputants is a relatively narrow one. Neither side denies that often the law refers to moral norms (like justice and fairness) which turn on the resolution of issues of substance or content. At that point the inclusive legal positivist says the judge is still doing law, while the hard legal positivist says that the law has run out. This is not to suggest that the debate is insignificant. When considering these issues, legal philosophers are attempting to come to an understanding of the nature of law. As with several of the debates discussed above, the resolution will not be found in resort to the facts, for these are conceptual or analytical disagreements over the best way to understand law from a theoretical standpoint.

Legal Positivism and Social Science

Legal positivism generally, and Hart's theory in particular, has an ambivalent relationship with the social sciences. Even what might seem like an obvious connection – use of the label 'positivism' – is not free of ambiguity. The term positivism literally refers to the fact that law is 'posited,' a human social creation. Some legal theorists assert that there is no direct connection between legal positivism and scientific positivism (Patterson, 1953: p. 15; Posner, 1990: p. 20); others suggest that the connection lies in the shared insistence on the distinction between *is* and *ought* (Schauer, 1996: p. 49). Still others insist that there is a strong connection, grounded in a shared empiricist narrowing of reality, the need for certainty (Rommen, 1998: pp. 109–113), and a shared opposition to theological and metaphysical modes of thought (Lee, 1989: pp. 16–18). Hart himself invoked social science. As noted earlier, he described his book as an essay in descriptive sociology, although legal philosophers have since distanced themselves from this claim. Furthermore, in his discussion of the internal aspect of rules, Hart acknowledged Peter Winch's *The Idea of a Social Science* (1958), and he has been lauded for importing the hermeneutic perspective from social science into jurisprudence.

Hart's theory has been criticized from certain social science quarters. A standard criticism is that Hart made a host of putatively descriptive sociological claims about law without having first done the requisite sociological investigation (Cotterrell, 1989: pp. 94–106). An anthropological critique of Hart focuses on the fact that his concept of law was abstracted from state legal systems, which leads to the conclusion that certain societies – those lacking in secondary rules – did not have law, suggesting that they are primitive or less civilized (Roberts, 1979: pp. 23–25). This smacks of ethnocentric privileging of Western legal systems. Social scientists who

hold the belief that "no society is without law" (Moore, 1978: p. 215) must reject Hart's concept of law. A third criticism, aimed at legal positivism more generally, is that legal positivist views tend to mask the social reality that there are a plurality of 'legal' systems operative in society, not just the state legal system that legal positivism emphasizes (Griffiths, 1986).

Despite these criticisms, Hart's concept of law has had a significant, if mostly underground, influence in social scientific approaches to the concept of law. Paul Bohannan's widely followed anthropological definition of law as the reinstitutionalization of norms bears a strong resemblance to Hart's notion of primary and secondary rules (Bohannan, 1965). Marc Galanter, one of the leading theorists of legal pluralism, adopts Hart's concept of law to distinguish legal from nonlegal forms of social regulation (Galanter, 1981: pp. 18–19, n. 26). Since it was abstracted from the state model, Hart's concept of law is attractive because it comports with general intuitions about law in the modern world. Whichever concept of law one formulates, it must be derived from a preconception about what law is (Tamanaha, 1993, 1995).

Although analytical jurisprudents have become excessively abstract of late, the legal positivist ambition to construct a general and descriptive approach to law is one the social sciences should embrace (Tamanaha, 2012). Law and society scholars have been building upon strains of legal positivism to take up issues like legal pluralism and the spread and penetration of law through globalization (Twining, 2009). Social scientists can conduct the sociological studies Hart failed to complete, using his model of law as the union of primary and secondary rules to guide their research, without necessarily accepting the separate propositions that this is the best or only way to characterize and understand law. With adjustments, a sociological or 'sociolegal' version of positivism has the potential to provide information about the relationship between law and society and to establish a bridge between jurisprudence and the social scientific study of law (Tamanaha, 2001).

See also: Logical Positivism and Logical Empiricism; Natural Law; Positivism, History of; Positivism, Sociological; Rule of Law (and Rechtsstaat).

Bibliography

Acquinas, T., 1987. Treatise on Law. Regnery Gateway, Washington, DC.
Austin, J., 1995. The Province of Jurisprudence Determined. Cambridge University Press, Cambridge, UK.
Bix, B., 1996. Jurisprudence: Theory and Context. Westview Press, Boulder, CO.
Bix, B., 1999. Patrolling the boundaries: inclusive legal positivism and the nature of the jurisprudential debate. Canadian Journal of Law and Jurisprudence 12, 17–32.
Bobbio, N., 1993. Thomas Hobbes and the Natural Law Tradition. University of Chicago Press, Chicago.
Bohannan, P., 1965. The differing realms of the law. American Anthropologist 67, 33–42.
Boorstin, D.J., 1996. The Mysterious Science of Law. University of Chicago Press, Chicago.
Coleman, J.L., 1982. Negative and positive positivism. Journal of Legal Studies 11, 139–164.
Coleman, J.L., 1996. Authority and reason. In: George, R.P. (Ed.), The Autonomy of Law: Essays on Legal Positivism. Oxford University Press, Oxford, UK.
Coleman, J.L., 1998. Incorporationism, conventionality, and the practical difference thesis. Legal Theory 4, 381–425.
Cotterrell, R., 1989. The Politics of Jurisprudence. Butterworth, London.
Dworkin, R.M., 1977. Taking Rights Seriously. Duckworth, London.
Dworkin, R.M., 1985. A Matter of Principle. Belknap Press of Harvard University Press, Cambridge, MA.
Dworkin, R.M., 1986. Law's Empire. Belknap Press of Harvard University Press, Cambridge, MA.
Finnis, J., 1980. Natural Law and Natural Rights. Clarendon Press, Oxford, UK.
Finnis, J., 1996. The truth in legal positivism. In: George, R.P. (Ed.), The Autonomy of Law: Essays on Legal Positivism. Oxford University Press, Oxford, UK, pp. 195–214.
Fuller, L., 1940. The Law in Quest of Itself. Foundation Press, Chicago.
Fuller, L., 1969. The Morality of Law. Yale University Press, New Haven, CT.
Galanter, M., 1981. Justice in many rooms: Courts, private ordering, and indigenous law. Journal of Legal Pluralism 19, 1–47.
Griffiths, J., 1986. What is legal pluralism? Journal of Legal Pluralism 24, 1–55.
Hart, H.L.A., 1958. Positivism and the separation of law and morals. Harvard Law Review 71, 593–629.
Hart, H.L.A., 1961. The Concept of Law. Clarendon Press, Oxford, UK.
Hart, H.L.A., 1994. The Concept of Law, second ed. Clarendon Press, Oxford, UK.
Hobbes, T., 1946. Leviathan. Blackwell, Oxford, UK.
Kelsen, H., 1961. The pure theory of law and analytical jurisprudence. Harvard Law Review 55, 44–70.
Kelsen, H., 1967. The Pure Theory of Law. University of California Press, Berkeley, CA.
Lee, K., 1989. The Positivist Science of Law. Avebury, Aldershot.
Leiter, B., 1998. Realism, hard positivism, and conceptual analysis. Legal Theory 4, 533–547.
MacCormick, N., 1992. Natural law and the separation of law and morals. In: George, R. (Ed.), Natural Law Theory. Oxford University Press, Oxford, UK.
Moore, S.F., 1978. Law as Process: An Anthropological Approach. Routledge and Kegan Paul, Boston.
Patterson, E.W., 1953. Jurisprudence: Men and Ideas of the Law, first printed ed. Foundation Press, Brooklyn, NY.
Paulson, S.L., 1992. Kelseon's legal theory: the final round. Oxford Journal of Legal Studies 12, 265–274.
Posner, R.A., 1990. The Problems of Jurisprudence. Harvard University Press, Cambridge, MA.
Raz, J., 1979. The Authority of Law: Essays on Law and Morality. Oxford University Press, Oxford, UK.
Roberts, S., 1979. Order and Dispute: An Introduction to Legal Anthropology. St. Martin's Press, New York.
Rommen, H., 1998. The Natural Law: A Study in Legal and Social History and Philosophy. Liberty Fund, Indianapolis.
Schauer, F., 1996. Positivism as pariah. In: George, R.P. (Ed.), The Autonomy of Law: Essays in Legal Positivism. Oxford University Press, Oxford, UK.
Tamanaha, B.Z., 1993. The folly of the 'social scientific' concept of legal pluralism. Journal of Law and Society 20, 192–217.
Tamanaha, B.Z., 1995. An analytical map of social scientific approaches to the concept of law. Oxford Journal of Legal Studies 15, 501–535.
Tamanaha, B.Z., 2001. Sociolegal positivism and a general jurisprudence. Oxford Journal of Legal Studies 21, 1–32.
Tamanaha, B.Z., 2012. What is general jurisprudence? A critique of universalistic claims by philosophical concepts of law. Transnational Legal Theory 2, 287–308.
Twining, W., 2009. General Jurisprudence: Understanding Law from a Global Perspective. Cambridge University Press, Cambridge, UK.
Winch, P., 1958. The Idea of a Social Science and Its Relation to Philosophy. Humanities Press, Atlantic Highlands, NJ.

Legal Process and Social Science: United States

John Monahan, University of Virginia School of Law, Charlottesville, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The use of social science in American courts is rooted in the legal realist movement of the early twentieth century. However, the standard for the admissibility of scientific evidence was not established until the Supreme Court's decision in Daubert v. Merrell Dow Pharmaceuticals in 1993. Social science evidence has three uses in court: to assist judges in making general law, to assist juries in determining specific facts, and to provide juries with general context for determining specific facts. At the beginning of the twenty-first century, efforts to educate judges in social science methods are burgeoning.

Introduction

There are two complementary orientations in the field of 'law and social science' in the United States. The first, usually called the 'law and society' approach, addresses theories of how law comes into being and changes in a culture. The second, referred to as the 'social science in law' approach, deals with more applied questions of how courts and juries make use of social science information in legal cases. This article is concerned with the latter orientation. It posits that social science has three uses in American courts: to assist juries in determining specific facts, to assist judges in making general law, and to provide juries with a general context for determining specific facts. At the beginning of the twenty-first century, rapid developments in the law of scientific evidence are taking place, and efforts to educate judges and lawyers in scientific methodology are burgeoning.

Jurisprudential Roots of Social Science Evidence in American Law

Since the rise of legal realism early in the twentieth century, American courts have, to varying extents, been open to using social science research when that research could help resolve an empirical issue that arose in litigation. Social science research was first urged upon a court in 1908 by Louis Brandeis in Muller v. Oregon (208 U.S. 412), a landmark case dealing with the constitutionality of social welfare legislation limiting to 10 h workday of any woman or girl employed in a factory or laundry. Brandeis assembled a substantial body of social science research tending to show the debilitating effect on women and girls of working long hours, and he presented this material to the US Supreme Court in a brief defending Oregon's statute. In the opinion in the Muller case upholding the legislation, after referring to the social science materials in a footnote, the Court stated that although they 'may not be, technically speaking, authorities,' the studies would nonetheless receive 'judicial cognizance.'

Modern Standards of Admissibility for Social Science Evidence

The US Supreme Court has articulated detailed standards to regulate the introduction of scientific evidence in court. In the 1993 case of Daubert v. Merrell Dow Pharmaceuticals, Inc. (509 U.S. 579), the Court held that judges should not simply defer to members of a scientific discipline when deciding whether to admit expert testimony. Trial judges are to be evidentiary gatekeepers. They have the responsibility to assess for themselves the validity of the methods underlying expert evidence. With Daubert, a judge's responsibility changed from merely assessing the degree of consensus in a field – the 'general acceptance' test articulated in the 1923 case of Frye v. United States (54 App. D.C. 46) – to evaluating the research methods and the analyses that lay behind expert opinion. In 2011, the Federal Rule of Evidence (Rule 702) dealing with testimony by expert witnesses was rewritten as follows to reflect Daubert:

> A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:
>
> (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
> (b) the testimony is based on sufficient facts or data;
> (c) the testimony is the product of reliable principles and methods; and
> (d) the expert has reliably applied the principles and methods to the facts of the case.

Daubert is now the accepted test of admissibility for scientific expert testimony not only in all federal courts but also in the courts of many states. Together with its subsequent rulings in 1997 in General Electric Co. v. Joiner (522 U.S. 136) that an appellate court owes deference to a trial court's decisions on the admissibility of scientific evidence, and in 1999 in Kumho Tire v. Carmichael (526 U.S. 137) that a court's gatekeeping responsibilities extend to all expert testimony, not just to expert testimony by scientists, the Supreme Court's decision in Daubert shifted the judicial paradigm for dealing with expert evidence.

Three Uses of Social Science in Law

Although the use of what came to be known as 'Brandeis briefs' became common in the years after Muller, legal commentators

were hard pressed to explain an apparent anomaly: How was it that Brandeis could present research to an appellate court in a written brief, when 'facts' were supposed to be introduced at the trial level by the oral testimony of witnesses? This conundrum persisted until Kenneth Culp Davis, in a seminal article published in 1942, distinguished two types of 'fact.' The first type, which he called adjudicative fact, referred to facts that were used to decide questions of interest only to the specific parties to a lawsuit, such as whether a particular traffic light was red or green. The second type, which he called legislative fact, referred to facts that were used by courts to help decide broad questions of law or policy, as Brandeis had used research to help decide the constitutionality of social welfare legislation. Davis's distinction soon became widely accepted and now forms the traditional scheme used by courts and commentators to describe the judicial uses of social science. In addition, there are strong indications that a third use of social science in American law – combining aspects of both legislative and adjudicative facts – has developed.

Social Science Used to Make Law

Many of the uses of social science research to support the creation of new law, or a change in existing law – what Davis (1942) called 'legislative facts' and what others have conceptualized as 'social authority' (Monahan and Walker, 1986) – have been in the area of discrimination. For example, the first modern use of social science in racial discrimination cases occurred with the citation in Brown v. Board of Education (347 U.S. 483 (1954)) to the famous 'doll study,' finding that African-American children in segregated schools preferred to play with a white doll rather than a black doll. Courts still often cite social science research in racial discrimination cases, but now it is often cited to oppose affirmative action on the basis of race. For example, in 2007, in the case of Parents Involved in Community Schools v. Seattle School District No. 1 (551 U.S. 701), a more conservative US Supreme Court held that a school board's act of voluntarily taking race into account to achieve racial balance violated the US Constitution. Chief Justice John Roberts wrote for the 5-4 majority that "The way to stop discrimination on the basis of race is to stop discriminating on the basis of race." In his concurring opinion, Justice Clarence Thomas cited 10 social science studies and literature reviews. The dissent by Justice Stephen Breyer cited another 10 social science studies and literature reviews to support the contrary conclusion that "to invalidate the plans under review is to threaten the promise of Brown."

Unsurprisingly, the law of gender discrimination has changed much more than the law of racial discrimination in recent decades, and is still highly unsettled. In the 1996 case of United States v. Virginia (518 U.S. 515), the US Supreme Court overturned a lower court finding that "a substantial body of contemporary scholarship and research supports the proposition that … females and males characteristically learn differently." The Supreme Court held that the justification for classifications based on gender 'must not rely on overbroad generalizations about the different talents, capacities, or preferences of males and females,' and that denying qualified women the right to enroll in a state-supported military academy violated the Constitution. Whether gender segregation in public elementary schools (e.g., all-girl mathematics classes) similarly violates the Constitution, however, has yet to be decided by the Supreme Court.

The death penalty is another area in which social science has been used to make law that has been transformed in recent years. In 1976, for example, Gregg v. Georgia (428 U.S. 153) cited many social science studies on the deterrent value of the death penalty, only to conclude that 'the results simply have been inconclusive' and that deterrence is 'a complex factual issue the resolution of which properly rests with the legislatures.' Since Gregg, major cases have found that 'death-qualified juries' are not so inherently biased toward the prosecution that they violate the fair cross-section of the community and the impartiality requirements of the Constitution (Lockhart v. McCree, 476 U.S. 162 (1986)), and that the administration of the death penalty does not violate the Equal Protection Clause of the Constitution, even if the death penalty is disproportionately imposed on murderers whose victims are white (McCleskey v. Kemp, 481 U.S. 279 (1987)). Each of these cases contained extensive citations to social science research, usually by the dissenting justices. Other Supreme Court decisions, however, hold that it does violate the Constitution to execute an offender who was younger than 18 years old when he or she committed a capital crime (Roper v. Simmons, 543 U.S. 551 (2005)), and that the Constitution likewise prohibits imposing the death penalty on a person convicted of child rape in which the child was not killed (Kennedy v. Louisiana, 554 U.S. 407 (2008)). Both of these cases also contained extensive citations to social science research. But in these decisions, the citations to social science research were usually by justices in the majority.

Jury size is yet another area on which social science research has been brought to bear. In Ballew v. Georgia (435 U.S. 223 (1978)), the Supreme Court – citing a great deal of social science research – declared Georgia's five-person criminal juries to be below the constitutional minimum. Thirty years later, a distinguished group of academics from law and the social sciences filed an amicus brief with the Supreme Court arguing that Ballew combined "erroneous empirical factfinding with mistaken history, producing an incoherent statement of the law which should be troubling to legal theorists of various perspectives—originalists, empiricists, and doctrinalists alike." The authors of the brief urged the Court to reconsider their decisions on jury size. The Court, however, declined to do so (Gonzalez v. Florida, 555 U.S. 1056 (2008)).

Social Science Used to Determine Facts

Adjudicative facts – also called 'social facts' – are facts that apply only to the particular parties before the court. They are used to determine (or 'adjudicate') what happened in a specific case, and not for some larger purpose, such as to argue that the law should be changed. One of the most frequent uses of social science research as adjudicative fact involves trademarks. The Lanham Trademark Act (15 USCA Sec. 1127) states that a new trademark will not be registered if it so resembles

a trademark already registered to another person "as to be likely … to cause confusion." A person who sells a product that is likely to cause confusion with an already trademarked product is liable for trademark infringement. Social science research in the form of surveys of consumers or potential consumers to ascertain the degree of confusion between products has been admitted in court for many decades. Initially, such evidence was often successfully challenged as contravening the prohibition against 'hearsay' evidence, since the respondents to the surveys were not present in court to testify. In Zippo Manufacturing Co. v. Rogers Imports (216 F. Supp. 670 (1963)), however, the hearsay objection was definitively laid to rest: "The weight of case authority, the consensus of legal writers, and reasoned policy considerations all indicate that the hearsay rule should not bar the admission of properly conducted public surveys."

The use of empirical research to determine consumer confusion in trademark law is well illustrated by Pharmacia Corp. v. Alcon Laboratories (2002). Pharmacia had registered the trademark Xalatan as its name for a prescription medication to treat the eye disease glaucoma. Pharmacia sued Alcon Laboratories for using the name Travatan for the glaucoma medication that Alcon produced, alleging that consumers would be confused between the two names. Alcon introduced Professor Shari Seidman Diamond as an expert witness. She testified to having conducted a true experiment in which ophthalmologists and optometrists (who, the court found, "make the ultimate determination as to which medications pharmacists will dispense and end-users—patients—will receive") were questioned after having been shown advertisements for Travatan or for a control medication. The court found that Diamond's research 'demonstrates persuasively a confusion rate of 1.5%,' and that such a low rate of confusion did not justify a preliminary injunction prohibiting Alcon from using the name Travatan.

The use of empirical methods to determine whether consumers are confused between two products has become so routine that the failure of a trademark owner to conduct a study often gives rise to an adverse inference by courts: if no survey is introduced, the likely reason is that there is no confusion to be found. In-person surveys of consumer confusion have been widely admitted in trademark cases since Zippo in 1963. Web-based trademark surveys, unsurprisingly, have been admitted in court only recently, but clearly seem the wave of the future.

The key to understanding how courts deal with social science research used to determine an adjudicative fact is that, unlike 'legislative' facts, adjudicative facts are susceptible to proof by the usual rules of evidence. Evaluating the research for the purpose of admissibility is limited to determining whether it is material, whether it has any probative value, and whether its admission violates any other evidentiary rule. The ultimate weight to be placed on the social science evidence is a matter for the fact finder.

Social Science Used to Provide Context

Most of the judicial uses of social sciences fall into one of the two categories originally described by Davis. In recent decades, however, courts have begun to confront uses of social science research that do not conform to the established classification. There are increasing indications that a third use of social science in law has emerged in which 'general' social science research is used to provide a context for deciding specific factual issues in a case – a use that was proposed by Walker and Monahan (1987) and termed 'social frameworks.' One can organize courts' use of social frameworks as applied to criminal defendants in a temporal sequence: the identification of people who will act illegally in the future, of people who are presently acting illegally, and of people who have acted illegally in the past.

The use of social science research to identify future legal actors is manifest in cases dealing with whether risk assessment tools developed by social scientists could be used to assist in making bail or parole decisions. Not only is there no longer a question as to whether denial of bail can constitutionally be predicated on violence risk assessments – as the Supreme Court stated in United States v. Salerno (481 U.S. 739 (1987)), "there is nothing inherently unattainable about a prediction of future criminal conduct" – but empirically based risk assessments are now routinely used in making bail decisions in many jurisdictions. Likewise, in the context of parole decision making, the tools developed by social scientists to assess the risk of recidivism are now being used in many states.

The use of social science research to identify people presently in the process of violating the law is exemplified by the approval of the social science-based 'skyjacker profile' to identify people in the process of commandeering an aircraft. Not surprisingly, such uses of social science in the national security context have been greatly expanded in light of the events of 11 September 2001.

The uses of social science research to identify people who have violated the law in the past – including 'cultural defense' cases; cases involving defendants who argue that they are battered women or battered children, or are suffering from combat-induced posttraumatic stress disorder; and cases in which the prosecution argues that the complainant exhibited symptoms of rape trauma – are now very commonplace.

Perhaps the most frequently litigated example of a social framework is when it is applied not to a criminal defendant but to an eyewitness to a crime. In State v. Chapple (660 P.2d 1208 (1983)), for example, the only evidence that the defendant had committed murder was the word of two eyewitnesses. At trial, the expert testimony of a research psychologist was offered by the defendant to counter the testimony of the eyewitnesses. In a proffer of the evidence, the psychologist described published studies on factors such as the speed with which memory decays over time, the effects of stress on eyewitness accuracy, and the relationship between the confidence of a witness and the accuracy of identification. The trial court held that the research on the accuracy of eyewitness identification was not a 'proper subject' for expert testimony and excluded the expert. However, the Arizona Supreme Court reversed and remanded the case for a new trial, stating that "there were a number of substantive issues of ultimate fact on which the expert's testimony would have been of significant assistance."

As an indication of how far some American courts have come in informing juries about social science research on the vagaries of eyewitness identification, consider the 2011 decision of the Supreme Court of New Jersey in the case of State v. Henderson (27 A.3d 872). Henderson argued that an eyewitness mistakenly identified him as an accomplice to a murder. The court stated:

> We appointed a Special Master to evaluate scientific and other evidence about eyewitness identifications. The Special Master presided over a hearing that probed testimony by seven experts and produced more than 2,000 pages of transcripts along with hundreds of scientific studies. Study after study revealed a troubling lack of reliability in eyewitness identifications. From social science research to the review of actual police lineups, from laboratory experiments to DNA exonerations, the record proves that the possibility of mistaken identification is real. Indeed, it is now widely known that eyewitness misidentification is the leading cause of wrongful convictions across the country.

The court concluded that an extensive array of instructions to the jury on the factors known to compromise the accuracy of eyewitness identifications should be given in appropriate cases, and in addition that 'there will be times when expert testimony will benefit the trier of fact,' and therefore expert testimony on eyewitness identification should be permitted as well.

In neither of these eyewitness cases was social science research being used to provide legislative facts. No creation or modification of a rule of law was sought. Rather, in both cases the parties offering the research accepted the applicable legal rules and sought to show that the research would help the jury to decide the specific factual issues being litigated. Yet in neither of these eyewitness cases was social science research being used to provide adjudicative facts either. The parties to the cases were not involved in the research at all: the experts relied on research studies published before the events that gave rise to the litigation took place.

The way that social science was used in Chapple and Henderson, however, reflects the defining characteristics of both legislative fact and adjudicative fact. In each case, the research demonstrated the critical component of legislative fact – generality. Just as the research used in Brown v. Board of Education (1954) focused on the benefits of racially integrating all public schools, the research used in Chapple and Henderson focused on 'general factors' affecting all eyewitnesses. In each case, the research used also possessed the critical component of adjudicative fact – specificity. Just as the research used in Pharmacia Corp. v. Alcon Laboratories (2002) spoke only to whether consumers were confused between two given products, the studies in Chapple and Henderson were used only to challenge the testimony of particular eyewitnesses to a particular crime.

The research used in these cases and many like them is thus neither wholly legislative nor wholly adjudicative fact but has the essential elements of both of the conventional categories. As mentioned in this article, Walker and Monahan (1987) have proposed social frameworks to denote the use of general conclusions from social science research to determine factual issues in a specific case.

Future Directions

At the beginning of the twenty-first century, social science in the legal process is being transformed by the rapid rise in the availability of material to train lawyers and judges in the uses of scientific evidence, including social science evidence (Simon, 2012). Shortly after Daubert, the Federal Judicial Center published a volume titled *Reference Manual on Scientific Evidence* that in 2011 went into its third edition. In addition, Faigman et al. (2011) published a comprehensive, five-volume, frequently updated work titled *Modern Scientific Evidence: The Law and Science of Expert Testimony*. A law school casebook titled *Social Science in Law: Cases and Materials* (Monahan and Walker, 2010) reviews the field and is now in its seventh edition. Works educating social scientists about the law also abound (Dvoskin et al., 2012; Skeem et al., 2009).

In the past decade, it has become difficult to find a Supreme Court decision implicating an empirical question in which the majority or the dissent did not cite social science research – a practice now as common among conservative as among liberal justices. Amicus briefs replete with references from the social sciences have become the rule rather than the exception in major litigation. Debates about relying on social science to create general contexts for resolving issues specific to a case have gone from whether to permit this procedure to how best to present empirical information to the fact finder.

Finally, a landmark report by the National Research Council, Strengthening Forensic Science in the United States: A Path Forward (2009), while focusing on biological science evidence such as DNA and fingerprints, may have significant implications for improving the quality of social science evidence admitted in court as well. In this regard, many see the future of social science in law as inextricably bound with recent and rapid developments in the biological sciences, particularly in the neurosciences. This may well be the case. But that future is not here yet (Morse, 2011), and may be quite some time in arriving.

See also: Death Penalty; Discrimination and the Law; Expert Testimony; Juries; Justice and Law; Law and Society: Development of the Field; Legal Realism; Science and Law; Scientific Evidence: Legal Aspects.

Bibliography

Davis, K.C., 1942. An approach to problems of evidence in the administrative process. Harvard Law Review 55, 364–425.

Dvoskin, J., Skeem, J., Novaco, R., Douglas, K. (Eds.), 2012. Using Social Science to Reduce Violent Offending. Oxford University Press, New York.

Faigman, D., Blumenthal, J., Cheng, E., Mnookin, J., Murphy, E., Sanders, J., 2011. Modern Scientific Evidence: The Law and Science of Expert Testimony, vols. 1–5. West, Eagan, MN.

Federal Judicial Center, 2011. Reference Manual on Scientific Evidence, third ed. West Group, St. Paul, MN.

Monahan, J., Walker, L., 1986. Social authority: obtaining evaluating and establishing social science in law. University of Pennsylvania Law Review 134, 477–517.

Monahan, J., Walker, L., 2010. Social Science in Law: Cases and Materials.

Morse, S., 2011. Lost in translation?: an essay on law and neuroscience. In: Freeman, M. (Ed.), 2011. Law and Neuroscience: Current Legal Issues, vol. 13. Oxford University Press, New York, pp. 529–562.
National Research Council, 2009. Strengthening Forensic Science in the United States: A Path Forward.
Simon, D., 2012. In Doubt. Harvard University Press, Cambridge, MA.
Skeem, J., Douglas, K., Lilienfeld (Eds.), 2009. Psychological Science in the Courtroom: Consensus and Controversy. Guilford Press, New York.
Walker, L., Monahan, J., 1987. Social frameworks: a new use of social science in law. Virginia Law Review 73, 559–598.

Relevant Websites

www.socialscienceinlaw.com – Social Science in Law.
http://www.ap-ls.org/ – American Psychology-Law Society.
http://www.lawschool.cornell.edu/sels/ – Society for Empirical Legal Studies.

Legal Realism

John Henry Schlegel, State University of New York, Buffalo, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Legal realism was a movement in American legal thought between World War I and World War II. The American Legal Realists were legal academics who shared a politics that was liberal, though not socialist, and four overlapping interests. The first was in what Pound called 'the law in action,' where the Realists pursued studies of state and federal courts, banks, and parking and traffic enforcement. The second interest was in criticizing the doctrinal results of legal formalism (*see* Legal Formalism) where they worked in areas such as antitrust, commercial and corporate law, conflict of laws, evidence, and torts. The third interest was in judicial decision making, where the Realists argued that the inability of the doctrinal formulation of legal rules to yield determinate answers in concrete cases established the subjectivity of judicial decisions. The fourth was in legal education where they supported the reorganization of the course materials of legal study in terms of the functions performed by legal institutions and encouraged a shift from the justification of law in terms of derivation from supposedly logical principles to one based on assumed knowledge of social conditions to which law is applied in the pursuit of supposedly agreed-upon social policy.

'Legal Realism' is the name generally given to a movement in American legal thought between World War I and World War II. However, any attempt to identify precisely the intellectual content of the Realist movement is made difficult by the fact that its salient tenets, and thus its meaning, were widely contested during its heyday and are still widely contested.

Difficulties in Defining 'Legal Realism'

The Realist movement has been seen as an outgrowth of the Sociological Jurisprudence of Rudolf von Jhering as expounded by Roscoe Pound (1911–1912), as well as a reaction to Pound's work. In addition, many have understood Realism to be an outgrowth of the thought of Oliver Wendell Holmes, Jr. (1897), although it is not clear that such an identification is warranted, given Holmes' relentless pessimism and lack of faith in democracy. Also, while all agree that Realism was a reaction to legal formalism, sometimes referred to as 'conceptualism' – the notion that legal rules can be correctly deduced from a proper understanding of the nature of a limited set of important, narrowly legal premises – there nevertheless is disagreement as to whether any significant American legal thinker held such a position and thus what an opposition to formalism might signify (*see* Legal Formalism). The most that can be said is that all scholars agree that some portion of the activities of individuals commonly called Realists was directed at critique of the doctrinal results of what these scholars pejoratively labeled legal formalism, at understanding judicial decision-making, at empirical research in law, at reform of legal education, and at participation in the activities of the New Deal administration of Franklin D. Roosevelt.

There is, however, nothing close to agreement among scholars about the relative importance of these various activities for understanding Realism; moreover, the weighing of relative importance in part determines who is included in a list of the Realists. Thus, while it is likely that most scholars would agree that Thurman Arnold, Charles E. Clark, Walter Wheeler Cook, William O. Douglas, Jerome N. Frank, Walton H. Hamilton, Karl N. Llewellyn, Underhill Moore, Herman Oliphant, and Hessel Yntema belong on the list, the inclusion or exclusion of men such as Adolph A. Berle, Walter Bingham, Felix Cohen, Morris R. Cohen, Arthur L. Corbin, John Dewey, Leon Green, Robert L. Hale, Robert M. Hutchins, Edwin Patterson, Thomas Reed Powell, Edward S. Robinson, Max Radin, Fred Rodell, Wesley Sturges, and Leon Tulin, all plausible figures, will change one's understanding of Realism, as well as be dictated by that understanding. What follows, therefore, is only one of several possible expositions of American Legal Realism. Other expositions can be found in the works of Duxbury (1995), Horwitz (1992), Hull (1997), Kalman (1986), Purcell (1973), Rumble (1968), Schlegel (1995), Stevens (1983), Twining (1973), and Verdun-Jones (1979).

A Contestable Definition

'Legal Realism' is best understood as the reaction of a group of individuals within the American legal academy to (1) the collapse of the Progressive movement after World War I, (2) the decisions of the U.S. Supreme Court invalidating economic regulation on grounds either of substantive due process or of interference with vested contractual rights, and (3) the general shift of the United States to the political right in the Harding, Coolidge, and Hoover administrations. The important words in this definition are 'within the American legal academy,' for the movement largely took place at three schools – Columbia, Johns Hopkins, and Yale. Obviously, this definition does not succinctly describe the political orientation of the Realist movement. On this topic, the best that can be said is that, taken as a whole, the Realists were hardly radicals. It would be astonishing to learn belatedly that any of these men were members of the Communist Party, and it is doubtful whether any even voted Socialist. Many served in the Roosevelt Administration, but doing so was an indication of socialist leanings only to the most reactionary members of the business elite and their Republican Party allies. Thus, Realists are better seen as mainstream liberals caught in a very conservative time.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86076-6

The relationship of the Realist movement to the Progressive movement can be seen both positively and negatively. Positively, there was the similarity in the political causes supported (organized labor), and opposed (big business). Also, the empirical research engaged in by many of the Realist scholars was similar to research undertaken by Progressive reformers (although not progressive law professors) before World War I. Indeed, both groups shared an 'all men of good will' ethics, as is evident from the notion underlying their empirical research that once the facts about a social problem are known, then reform will follow. In this sense, neither group was committed to what we would now see as an ideological view of law that would suggest that exposure of the 'facts' about a social practice might bring embarrassment (and thus denial and resistance to change). For both the Progressives and the Realists, evil existed in the world, but it existed in individuals, not in classes or other affinity groups.

The Realists' relationship to the Progressive movement can be seen negatively in the reaction of the members of the movement to Pound as dean of the Harvard Law School. Pound was a mainstay of the legal wing of Progressive reform movements before World War I. His condemnation of 'Mechanical Jurisprudence' (Pound, 1908) was a rallying cry for Progressives, as was his exposition of 'Sociological Jurisprudence' (Pound, 1911–1912). Pound's jurisprudence was designed to be a response to the first wave of antiprogressive constitutional decision-making by the Supreme Court. It seemingly encouraged empirical research by emphasizing the difference between "the law on the books and the law in action" (Pound, 1910), in order to urge courts to avoid a formal, mechanical application of law and instead to recognize, balance, and thus adjust the interests of the parties in the light of contemporary social and economic conditions.

Yet, for the Realists, Pound was also part of the problem that their activities were designed to address. For them, all of Pound's talk about the gap between the law on the books and the law in action and about the need for balancing of interests had never translated into the activity of exploring these gaps or identifying the interests in question through empirical research, or into actually suggesting how particular interests might be balanced. Moreover, Pound's law school was little changed from the school of Langdell and Ames, seen by Realist scholars to be among the arch formalists. It reflected none of what they saw as the implications of Progressivism and Sociological Jurisprudence for legal education – a revision of category systems, an emphasis on factual investigation, and explicit work on social policy.

What Realists Did

What the above definition pointedly lacks is some explicit statement of what the Realists believed, and therefore what Realism is. This omission is intentional. Realism is what Realists did, and it is only with that aphorism in mind that one can make sense of the disparate materials that were published by what must be seen as a network of friends, not a coherent intellectual position. Start, then, at the beginning.

Although some pre–World War I work by Bingham (1912) has a particularly Realist cast, Realism largely started at Columbia with pieces by Cook and Moore, who respectively argued for a more empirically based understanding of the real, as opposed to verbal, activities of appellate courts in the field of conflict of laws and of the real, as opposed to verbal, activities that made up the institutional practices that the law defined, such as property and contract. From their work, one can see that Legal Realism is not even a distant relative of philosophical realism, the doctrine that the concepts of daily life are ontologically real, but, as Leiter (1997) has suggested, is much closer to naturalism, a commitment to see and describe the realities of daily life scientifically, that is, without idealization or the avoidance of the ugly. However, in Legal Realism's concern with democracy and the activities of the middle class, there is some affinity with literary and artistic realism as well.

Curiously, the group of Realists who taught at Columbia, including Oliphant, Llewellyn, and Douglas, was not particularly interested in constitutional law, as might have been expected given the group's affinities to Progressivism. Instead they were interested in corporate and commercial laws. These men argued that, because of antiquated ideas about judicial decision making, in areas of law related to business, just as in constitutional law, the law was lagging behind changes in economic life, particularly the growth of the large corporation employing thousands of workers and the change in methods of production and distribution of goods that had broken older patterns of finance and trade. This focus had much in common with the Institutional Economics of John Commons, Richard T. Ely, and Thorstein Veblen that emphasized the way that the institutional (by which they meant regular) patterns of economic behavior determined the growth or decline of industries, attitudes toward labor, and practices in the pricing of goods. Thus, Moore worked on the law of banking, Oliphant on labor and antitrust, Llewellyn on the sale of goods, and Douglas on corporations and commercial bankruptcy.

Even more curiously, the empirical research that was implied by the work at Columbia was not undertaken there, but rather at Yale and Johns Hopkins, mostly in the years after the Columbia law faculty blew up over the results of a dean search. At Yale, pioneering empirical work, superintended by Clark, Douglas, and Moore, focused on the activities of the federal courts in civil, criminal, and bankruptcy cases; on the activities of bankers in handling negotiable instruments; on the operation of the tort system in auto accidents; on the effect of parking regulations in altering the behavior of drivers; and on the legal needs of citizens of modest means. At Johns Hopkins, equally pioneering work, superintended by Oliphant, Yntema, and their colleague at Columbia, Leon C. Marshall, was done on the operation of state courts in civil, criminal, divorce, and justice of the peace cases.

Much of the rest of the scholarship of the Realists, though not empirical in the narrowest sense, was still within artistic and literary naturalism's desire to describe the realities of daily activities. Cook's (1942) continuing work on what courts were really doing in conflict of laws doctrine and Green's (1927, 1930) parallel work on torts doctrine fit this mold, as does the works of Arnold (1935) and Robinson (1935) on the symbolic uses of law, Llewellyn's work on commercial practice (1936, 1937, 1939a, 1939b), Hutchins and Slesinger's (1928, 1929) work on the psychological underpinnings of evidence law, and Berle and Means' (1932) work on corporate governance. All of

this work, and most of the more narrowly empirical work as well, was directed at critiquing legal, particularly judge-made rules, but Realism is more famous for its related, but more frontal assault on contemporary understandings of judicial decision-making.

Realists attacked formalist understandings of judicial decision-making in three ways: logically, psychologically, and sociologically. The logical attack argued that legal concepts that the formalists, especially the formalists on the Supreme Court, saw as necessarily determining decisions were not in fact determinative. Here, Realists relied on Wesley N. Hohfeld, who argued that legal concepts inappropriately treated as unitary various discrete legal relationships, any one of which could not be implied from the existence of another; on Judge Benjamin N. Cardozo, who asserted that there were four different methods for deciding a case and so allowed the judge the 'sovereign prerogative of choice'; on Cook, who claimed that legal rules 'hunt in pairs,' so that for every rule there is an exception or counterrule whose application would lead to a different result; and on Llewellyn, who eventually demonstrated that there is a great range of acceptable techniques for dealing with a legal precedent. Together, these observations suggested that legal concepts did not produce necessary results by the use of logic alone, but rather permitted judges a choice of results, a choice that was made surreptitiously, as it were, without explicitly considering the function that the rules played in society or the policy behind those rules.

The psychological attack argued that judicial opinions, traditionally presented in a seemingly syllogistic form, hid a decision that was reached in very different ways. Here, Realists relied on the work of another judge, Joseph Hutchinson, who argued that most decisions were decided on the basis of a 'hunch' about the appropriate result, and of Frank, who argued strenuously that a judicial opinion was but a rationalization of a decision reached in another way, that formalism was motivated by an infantile search for paternal authority, and that a truly adult judge would recognize and expose the many factors, particularly the idiosyncrasies of his response to the facts proven at trial, that led to the choice that was made in an individual case. Together these observations suggested that judicial decisions obscured the real factors that operated to generate results in particular cases behind a web of false necessity.

The sociological attack, which was more muted than either of the other two, attempted to identify the factors at work behind the facade of the judicial opinion. Here, the most common argument implicitly drew on the work of William J. Ogburn, who argued that the structure of social understandings of the world commonly lagged behind the practices that these understandings purported to describe. Versions of this concept of 'cultural lag' were offered by the Realists to explain how it was that judges used an 'outdated' economic theory to understand labor conditions or commercial transactions, an 'outdated' psychological theory to understand witness testimony, or an 'outdated' political theory to limit governmental action to address the conditions of modern life. It was here that the Realists' jurisprudence, as that word is traditionally used, met up with their interest in empirical research. Faced with the facts gathered from empirical research, it was assumed that judges and other legal actors would come to abandon their outdated theories.

The Realists' remedy for the inadequacies of formalism was functionalism. The problems with formalism were seen to be the lack of relationship between the concepts and categories of the law and the 'realities' of modern life. Closer attention to these realities would allow one to focus the law's attention on institutions such as the organized activities of businessmen in a trade association, and on the functions that these activities represented, and so see to it that the law addressed directly the problems in that trade as a matter of effective and appropriate social policy. Such an analysis would disclose just when the law needed to be updated because of its lag behind social practice and also would expose circumstances where law was dysfunctional, or where a body of doctrine was ineffective, as, for example, when Arnold argued that the antitrust laws provided the comfort of a myth that law was doing something to address public fears about big business, while ignoring the anticompetitive practices that made for monopoly.

For all the talk about functionalism as an antidote to formalism as a jurisprudence, indeed for all that Realism is taken to be a jurisprudence, it is important to understand that much of the impetus behind realism was educational. The Realists were mostly law professors for whom the question of how to present the materials of law study was a daily concern. It is with respect to legal education that functionalism and empirical research might have been expected to have some real bite. Functionalism raised questions about what were the appropriate categories under which to group the legal materials, and a study of the entire curriculum at Columbia was undertaken to examine how the curriculum could be organized more functionally. Similarly, empirical research was designed to answer the questions about the efficacy of the institutions and rules of law in the aid of understanding what social policy was appropriate in each functionally defined area.

Although almost no part of the curriculum was rearranged functionally (e.g., the course in trusts is still taught despite the absolutely withering critique that Arnold leveled against it), and though law professors do little more empirical research today than was the case before World War I, it is in the area of education that Legal Realism had its greatest impact. The critique of formalism utterly changed the way that law was justified in the law school classroom. From a formalism based on assertedly logical derivation from the nature of a given legal category, the justification of legal rules has shifted to a realism based on assumed knowledge of social conditions to which law is applied in the pursuit of supposedly agreed upon social policy. The law professors now talk the language of the early twentieth century political science professors, as well as of the bureaucrats in Washington and in scattered state capitals. Whether this is an advance is for others to say (*see* Legal Education).

Realism Today

The continuing interest of scholars in Realism, a movement that had lost its energy by the time of the entry of the United States into World War II, is one of its most interesting aspects. Recently Barzun revisited Frank's *Law and The Modern Mind* and Fenster examined Arnold's understanding of the administrative state. The Law and Economics movement and Critical Legal Studies

movement both have claimed to be the legitimate descendants of Realism, despite these movements' different political orientations and their different understandings of what Realism was about. The Law and Society movement has done so as well, again with own understanding of what Realism was about, though arguably with a quite similar political orientation. The effort to protect the law from Legal Realism is of long standing (see Purcell, 1973; Duxbury, 1995); recently Tamanaha (2010) has added to that effort. From different perspectives, de Been (2008) has endeavored to protect Realism from Critical Legal Studies and Leiter (2007), from both Critical Legal Studies and Law and Society. Exactly why American Legal Realism is such a prized ancestor, as well as such a threat to law, is a matter that needs to be addressed in the literature.

See also: Law and Society: Development of the Field; Legal Education; Legal Formalism; Legal Positivism; Legal Scholarship.

Bibliography

Arnold, T., 1935. The Symbols of Government. Yale University Press, New Haven.
Berle, A., Means, G., 1932. Modern Corporation and Private Property. Commerce Clearing House, New York.
Bingham, J., 1912. What is the law? Michigan Law Review 11, 1–25, 109–121.
Cook, W., 1942. The Logical and Legal Basis of the Conflict of Laws. Harvard University Press, Cambridge.
deBeen, W., 2008. Legal Realism Regained: Saving Realism from Critical Acclaim. Stanford University Press, Stanford, CA.
Duxbury, N., 1995. Patterns of American Jurisprudence. Clarendon Press, Oxford, UK.
Green, L., 1927. Rationale of Proximate Cause. Vernon Law Book Co, Kansas City, MO.
Green, L., 1930. Judge and Jury. Vernon Law Book Co, Kansas City, MO.
Holmes Jr., O.W., 1897. The path of the law. Harvard Law Review 10, 457–478.
Horwitz, M.J., 1992. The Transformation of American Law 1870–1960: The Crisis of Legal Orthodoxy. Oxford University Press, New York.
Hull, N.E.H., 1997. Roscoe Pound and Karl Llewellyn: Searching for an American Jurisprudence. University of Chicago Press, Chicago.
Hutchins, R., Slesinger, D., 1928. Some observations on the law of evidence – memory. Harvard Law Review 41, 860–893.
Hutchins, R., Slesinger, D., 1929. Some observations on the law of evidence – state of mind to prove and act. Yale Law Journal 38, 283–298.
Kalman, L., 1986. Legal Realism at Yale 1927–1960. University of North Carolina Press, Chapel Hill, NC.
Leiter, B., 1997. Rethinking legal realism: toward a naturalized jurisprudence. Texas Law Review 76, 267–315.
Leiter, B., 2007. Naturalizing Jurisprudence: Essays on American Legal Realism and Naturalism in Legal Philosophy. Oxford University Press, New York.
Llewellyn, K., 1936. On warranty of quality and society (pt. 1). Columbia Law Review 36, 699–744.
Llewellyn, K., 1937. On warranty of quality and society (pt. 2). Columbia Law Review 37, 341–409.
Llewellyn, K., 1939a. Across sales on horseback. Harvard Law Review 52, 725–746.
Llewellyn, K., 1939b. The first struggle to unhorse sales. Harvard Law Review 52, 874–904.
Pound, R., 1908. Mechanical jurisprudence. Columbia Law Review 8, 101–116.
Pound, R., 1910. The law on the books and the law in action. American Law Review 44, 12–36.
Pound, R., 1911–1912. The scope and purpose of sociological jurisprudence. Harvard Law Review 24, 591–619; 25: 140–168, 489–516.
Purcell Jr., E.A., 1973. The Crisis of Democratic Theory: Scientific Naturalism and the Problem of Value. University of Kentucky Press, Lexington, KY.
Robinson, E., 1935. Law and the Lawyers. Macmillan, New York.
Rumble Jr., W.E., 1968. American Legal Realism: Skepticism, Reform and the Judicial Process. Cornell University Press, Ithaca, NY.
Schlegel, J.H., 1995. American Legal Realism and Empirical Social Science. University of North Carolina Press, Chapel Hill, NC.
Stevens, R., 1983. Law School: Legal Education in America from the 1850s to the 1980s. University of North Carolina Press, Chapel Hill, NC.
Tamanaha, B., 2010. Beyond the Formalist-Realist Divide: The Role of Politics in Judging. Princeton University Press, Princeton, NJ.
Twining, W., 1973. Karl Llewellyn and the Realist Movement. Weidenfeld and Nicholson, London.
Verdun-Jones, S., 1979. Cook, Oliphant and Yntema. Dalhousie Law Journal 5, 3–44.

Legal Reasoning and Argumentation

Geoffrey Samuel, Kent Law School, UK; Law School, Sciences-Po, Paris

© 2015 Elsevier Ltd. All rights reserved.

Abstract

What is legal reasoning and argumentation? This article sets out to investigate this question both from a diachronic and a synchronic perspective. However one difficulty to be faced is the formulation of an appropriate epistemological framework. If one adopts a diachronic approach one can certainly appreciate what legal reasoning has been but it proves difficult to predict where it is going. Having started out as a descriptive-inductive technique, reasoning arrived in the nineteenth century as a formal deductive-axiomatic process. But, following the abandonment of such formalism in the twentieth century, what stage now follows the axiomatic? If one takes a synchronic approach various different models emerge in turn provoking conflict and debate about the nature and acceptability of these different models of reasoning.

Introductory Remarks

Legal reasoning and legal argumentation, from a practitioner's point of view, is at one level relatively easy to describe and to define: it is the reasoning used by lawyers to solve a legal problem, to advise a client or to justify a legal decision. Some would say that it is just ordinary reasoning used in a legal context (see e.g., Alexander and Sherwin, 2008). If one sees legal knowledge – as most lawyers seemingly do – as consisting of rules and principles (Susskind, 1987: 78–79), then legal reasoning is the reasoning attached to the application of these norms to specific problems. At another level, however, legal reasoning and argumentation can be seen to be extremely complex. One reason for this is that *all* method and reasoning in the sciences and social sciences is complex (see Law: History of Its Relation to the Social Sciences); the knowledge that emerges from the social science reasoning processes depends upon the schemes of intelligibility employed, the inferential reasoning techniques applied, the level of observation adopted, the theory brought into play and the paradigm orientations that inform these schemes and reasoning techniques (Desjeux, 2004; Berthelot, 2001). Legal reasoning might seem more straightforward than reasoning in, say, sociology, given that it would appear to be just a matter of applying rules to facts, but such a view is simplistic. Moreover, legal reasoning, strictly speaking, needs to be distinguished from legal argumentation: broadly speaking, one has the aim of both arriving at and justifying a decision while the other aims to convince.

Methodological and Theoretical Approaches

In fact one of the problems with legal reasoning is that it is capable of being approached from a variety of disciplinary and theoretical perspectives, each of which gives a rather different account. This may, of course, be true of reasoning in other areas as well; yet with respect to law it is more acute for what might be called a paradigm reason. Law, like theology, is subject to a particular paradigm orientation which might be labeled the authority paradigm (Samuel, 2008, 2009). Lawyers work with texts that can be criticized but never questioned in terms of their authority. A law is a law, and so the scope of reasoning about laws is always going to be limited in comparison to work in other, what might be called enquiry paradigm, disciplines. Consequently legal reasoning viewed from inside the authority paradigm – what might be called the internal view – could well be different from such reasoning viewed from outside the paradigm, that is to say, an external view, in that the epistemological models might well be different.

One of the challenges therefore is to determine the approach to be adopted. Does one look at what legal reasoning has been (diachronic) or does one look at what it is now, that is to say legal reasoning as a timeless set of techniques situated in the here and now (synchronic)? Does one adopt an approach internal to the law – that is, from the viewpoint of a lawyer and (or) judge – or does one adopt an approach external to the law, for example, as a social scientist or a philosopher? If one adopts an internal view, does one envisage legal reasoning to be based on a model of rules, of rights, of policy considerations, of remedies, of causes of action? If one adopts an external view, does one employ a causal, structural, functional, hermeneutical, individualist actional, or a dialectical scheme of intelligibility (see Berthelot, 1990: 62–83)? Moreover how are the standard techniques of inferential reasoning – induction, deduction, abduction, analogy, and metaphor – to be accommodated within these various approaches?

In addition underpinning these models and techniques are questions of theoretical approaches to law. Broadly speaking one can, traditionally, talk of three different schools of legal theory, namely natural law, positivism, and realism (see Legal Positivism; Legal Realism; and Natural Law). How might these theory positions affect reasoning and argumentation? Another distinction is between realism and formalism (Twining and Miers, 2010: 114–116; and see articles Legal Realism and Legal Formalism). How might this dichotomy impact upon reasoning? Epistemological issues bring their own distinctions and dichotomies (see Atias, 1985, 1994, 2002). Is law, and its reasoning, any different in essence from scientific reasoning? For example, is there a fundamental distinction to be made between explanation (scientific reasoning) and understanding (social science reasoning) (see Makkreel, 2006)? Might one wish to distinguish between axiomatic and casuistic reasoning (see Jonsen and Toulmin, 1988)? Has legal reasoning passed through distinct stages? Thinking about legal reasoning and argumentation will necessarily require an examination of, and reflection upon, all of these different models, techniques, schemes, and paradigm orientations.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86159-0

The epistemological issues remain acute even when one examines legal reasoning and argumentation from a diachronic viewpoint. Now it might be tempting to think that the 'what legal reasoning has been' question has formed the object of considerable research. Yet this is not really the case, although the situation is changing (see e.g., Gordley, 2013). There is certainly much literature examining the different schools of civilian jurists and their methods, and this literature is of importance (see recently Champeil-Desplats, 2014). But the problem is that legal reasoning as the subject of research and analysis requires an epistemological framework through which the old jurists can be understood. As has been suggested, several different frameworks can be adopted. One could adopt, for example, a dichotomy that goes back to Aristotle, namely the one between axiomatic (formal) and casuistic (practical) reasoning. "In formal arguments," write Jonsen and Toulmin, "particular conclusions are deduced from ('entailed by') the initial axioms or universal principles that are the apex of the argument" (1988: 34), whereas practical arguments "draw on the outcomes of previous experience, carrying over the procedures used to resolve earlier problems and reapplying them in new problematic situations" (1988: 35). This model might be further developed by applying Robert Blanché's diachronic model to explain the stages through which scientific reasoning typically passes: it starts out from a descriptive stage passing through an inductive and then deductive stage before arriving at an axiomatic stage (Blanché, 1983: 65).

This Blanché model might work well enough for scientific reasoning, but with respect to the social sciences it could well break down (although it does work in part for law). And so another possibility is to analyze reasoning through the six schemes of intelligibility proposed by Jean-Michel Berthelot. These are the causal, structural, functional, hermeneutic, actional, and dialectical schemes, each of which is capable of combining with one or more of the other schemes depending upon the program or paradigm of the researcher (Berthelot, 1990: 62–83). Berthelot himself summarized the schemes as follows; there is:

> the *causal* scheme (if x, then y or $y = f(x)$); the *functional* scheme (S→X→S, where one phenomenon X is analyzed from the position of its function – X→S – in a given system); the *structural* scheme (where X results from a system founded, like language, on disjunctive rules, A or not A); the *hermeneutical* scheme (where X is the symptom, the expression of an underlying signification to be discovered through interpretation); the *actional* scheme (where X is the outcome, within a given space, of intentional actions); finally, the *dialectical* scheme (where X is the necessary outcome of the development of internal contradictions within a system).
>
> Berthelot, 2001a: 484

The importance of these schemes is that they provide the basic forms of epistemological engagement with social reality; they are the means by which this reality can be 'read.' They can – or some of them can – also be used in the engagement with a text. A structural reading can, for example, be contrasted with a functional interpretation.

There are other epistemological possibilities of course – one might focus for example on the concepts developed by the Romans and their subsequent history – but often the reasoning attached to the development and application of these concepts takes one back to schemes of intelligibility. History also reveals the importance of interpretation given the central role of texts in law (see e.g., Maclean, 1992). Again, as we have suggested, interpretative methods lead one back to schemes of intelligibility; structuralism, functionalism, dialectical techniques, and so on, when applied to the understanding of a text, can be behind differing interpretations (see further Samuel, 2009a).

Above and below schemes of intelligibility are other important epistemological considerations. Above the various schemes there are paradigm orientations within which the activity of reasoning takes place. When one examines the various historical schools of jurists these orientations are not always the same, although, as we have suggested the authority paradigm is largely dominant. Below the schemes are the modes of reasoning themselves. These modes or techniques of reasoning are associated with the idea of inference (Masquelet, 2006: 76); that is the operation by which one moves from one descriptive situation, proposition, or other given to another descriptive situation, proposition, or other given by means of an intellectual process. The four traditional modes of inferential reasoning are induction, deduction, abduction, and analogy (Masquelet, 2006: 75–93).

Early Civilians

All of this 'theorizing' will explain why the starting point for a historical examination of legal reasoning in Europe is not to be determined by time but by paradigm. If one was to look at legal reasoning in terms of time, the starting point would obviously be Roman law in Ancient Rome (see Stein, 1999). Yet such an approach is possibly not that helpful if one wishes to understand contemporary reasoning methods. Arguably the starting point should be the medieval jurists who worked, from the eleventh to the fifteenth century in Italy and elsewhere, on the rediscovered Roman texts that had been compiled and preserved by the Emperor Justinian in the sixth century AD. One starts, then, not at the beginning so to speak, but at a half-way point from where one can look both backward to the Roman jurists and forward to the subsequent civilians and common lawyers.

As James Gordley observes, the medieval jurist "was one who, by virtue of his specialized training, knew the Roman texts and could cite them to support his opinions" for law "was based, as much of it still is, on the exposition of authoritative texts" (Gordley, 2013: 32). A number of points need to be stressed about the medieval jurists who developed what became known in Europe as the *ius commune*. First, these jurists regarded the rediscovered Roman texts as completely authoritative statements not of 'Roman law' but of 'the law' (Stein, 1999: 46). They were not historians (thus they regarded every text as having equal authority), but lawyers whose job was both to explain the law to students in the newly established law faculties and to apply the texts to the practical problems of their time (Stein, 1999: 47). The method of explanation was exegetic; they explained the texts through margin notes (glosses) and, later, commentaries on the different parts of the Roman materials (*Corpus Iuris Civilis*) (Gordley, 2013: 28–51).

These glosses and commentaries themselves subsequently became authoritative (Jones, 1940: 15–21).

Second, although the Roman texts, and the practical problems discussed by the Roman jurists, were large and extensive, they did not cover everything and so the medieval jurists were faced with the problem of applying these texts to new situations. They did this in a largely casuistic way; they started out at the level of fact – persons and things – and developed Roman law notions using induction and analogy. Thus they might generalize a text by way of induction so that the rule (*regula*) could be applied to the new situation or they might take a particular factual problem discussed in the Roman texts and apply it by way of analogy (Gordley, 2013: 33–45). Equally they limited the application of a text or a rule by the use of distinctions and dialectical oppositions (Ullmann, 1975: 87). In doing this the later medieval jurists – the Commentators – fashioned what might be termed an algorithmic approach to legal reasoning and problem-solving (*aut* ... *aut*: either … or).

Third, the medieval jurists approached law and problem-solving in a way that today would be described as a highly 'positivistic.' The *modus arguendi* in law was limited entirely to the texts of the *ius commune* (Roman law). All arguments and reasons had to be based on a text from, and only from, the *Corpus Iuris Civilis*. In turn what the *Corpus* offered the jurist was a complete body of interrelating concepts, institutions, and technical language which had been sufficiently developed by the Roman jurists to serve the needs of the sophisticated commercial society that was Rome in the beginning centuries AD (Stein, 1999: 16–20).

Looking back from the medieval jurists to Roman law itself what the Roman jurists offered the European renaissance, through the preserved texts, was a firm foundation upon which a particularized legal reasoning could be developed. The system of concepts framed – in the *Institutes of Justinian* – around the three institutions of persons (*personae*), things (*res*), and actions (*actiones*) was the basic framework (Stein, 1999: 18–20). Between these institutions the jurists fashioned a series of relations between person and thing and person and person. The key relations between person and thing were ownership (*dominium*, *proprietas*), possession (*possessio*), and rights (*iura*) in another's property; and between person and person there was the obligation subsequently defined as a 'legal chain' (*vinculum iuris*). Attaching to these relations were other descriptive – or at least quasidescriptive – notions such as 'damage,' 'fault,' (*dolus*, *culpa*) and 'interest' (see generally Pichonnaz, 2008). Here, then, was a series of legal concepts and categories for a model of 'social reality' that could be manipulated to produce its own 'legal reality.'

Accordingly, and following on from this institutional model, the Roman jurists offered a sophisticated vocabulary and taxonomy that enabled the later jurists not only to reason in a 'positivistic' way but to develop this vocabulary and taxonomy to a more sophisticated level. And so, for instance, while there is a range of contemporary areas of legal studies such as constitutional, administrative, corporate, and human rights law that were completely unknown to the Romans, the foundations of these legal subjects can nevertheless be traced back – often via the medieval jurists – to the *Corpus Iuris Civilis* (see Mestre, 1985; Tierney, 1982; Canning, 1987).

In addition the texts reveal a series of reasoning techniques that became standard methods in the Western legal tradition. The jurists not just describe the various legal actions, legal texts, and legal rules (*regulae*) but interpret, illustrate, and test them through the posing of factual problems. In fact the jurists did more than this. They used these factual situations to develop the law itself by moving outward from one factual situation to another often via the question 'what if' (*quid enim si* …?) or some similar opening expression. In other words their casuistic techniques – which also included induction, distinction, and division – provided a means in themselves by which Roman legal reasoning could be employed by jurists in medieval and modern Europe (Kelley, 1990: 64–66). James Gordley compares this technique against the methods used by a philosopher, scientist, or social scientist all of whom define concepts abstractly and then work out their implications (Gordley, 2013: 8).

Later Civilians

When one looks forward from the medieval jurists a very different reasoning model gradually emerges. The change came in the sixteenth century with a school of French jurists known as the Humanists who developed a new approach toward the Roman texts (see generally Kelley, 1970). The humanists sought, on the basis of Roman classical law, to establish law as a science through which it could be explained clearly and systematically (Stein, 1999: 79–82). They wanted to cut through all the old 'pedantic' glosses and casuistic reasoning of the past – to get back to the original Roman texts and intentions of the authors who wrote them – so that the law could be perceived as a coherent whole. The emphasis henceforth was no longer to be on the particularities of the case and its analysis through dialectical distinctions and analogies but on what might be called the 'architecture' of the legal system – on how it fitted together as an interrelating systematic whole (Thireau, 2001: 232–238; Villey, 2006: 454–487).

Despite the fact that the humanist project had, at the time, little influence on legal practice or the methods employed by practitioners – the *Mos Italicus* (late medieval Italian methods) continued to dominate practice right into the seventeenth century (Jones, 1940: 34–35) – the longer term effect of this humanist revolution on legal reasoning proved immense. During the seventeenth and eighteenth centuries attention turned away from the kind of casuistic methods that characterized the Roman and medieval jurists toward a deductive and then axiomatic approach to reasoning. The authority of texts, which was the basic epistemological paradigm of the medieval age, gave way to a new paradigm, namely the authority of coherence. Thanks to jurists who were also mathematicians, such as Wilhelm Leibniz (1646–1716) and Christian Wolff (1679–1754), legal reasoning became associated with geometry (*mos geometricus*); solutions to legal cases were, accordingly, to be solved uniquely by the use of deductive logic (Gordley, 2013: 165–194; Champeil-Desplats, 2014: 60–63). The result was a complete change in the nature and style of law books. The dialectical method of

division, distinction, induction, and analogy, rooted in practical case examples, was replaced by a highly systematized set of ever more abstract legal propositions that no longer followed the order set out in the main source of Roman law itself, namely the *Digest*. They adopted the style and structure which characterized the *Institutes*. Thus, in France, the jurist Jean Domat (1625–96) rearranged Roman law in its 'natural order' so that the Roman law was to be learned and applied, not through the reading and application of the original texts, but through a set of principles. The great advantage that this textbook reform brought to law – an advantage for practitioners as well as students – was that the law (in effect Roman law) became very much easier to read and absorb (Gordley, 2013: 145).

Moreover, before the nineteenth century, these principles or 'axioms' (*primae notiones*) were seen as part of a system of transnational natural law (*ius naturale*) and this resulted in a reasoning process founded on the premise that there are no gaps in the law. Even if the domestic law (*ius proprium*) proved wanting, there would always be the higher principles (*ius naturale*), discoverable and applicable through 'right reasoning' (*recta ratio*), which would provide a solution. The idea that there is a 'right answer' – associated with the twentieth century jurist Ronald Dworkin – to every hard case is, therefore, no new idea. Even with the coming of positivism and monism in the nineteenth century this axiomatic conceptualism was by no means abandoned. Indeed it was consolidated (Gordley, 2013: 195–204, 212–215). The *ius naturale* was replaced by the idea of *scientia iuris*, the thesis that the authoritative legal texts themselves – now the only valid source of law – were underpinned by a set of concepts from which legal solutions would be logically deduced by legal scientists (Champeil-Desplats, 2014: 71–93).

The humanist jurists changed legal perception in another way as well. In addition to emphasizing the importance of system they reflected the effect of the nominalist revolution that had been provoked by the medieval philosophers. These thinkers held the view that abstractions had no existence in the world of reality and were thus only names (*nomina*); there were, for example, no such things as 'forests' only trees (Jones, 1940: 169–170; and see generally Villey, 2006). The result was that the Roman expression *ius* took on a new, more individualistic meaning. Now, in classical Roman law, the term *ius* probably gave expression to the idea of a legal relation rather than some kind of individual 'right' and certainly when it came to problem solving the key institution was the *actio* (Jolowicz, 1957: 77). This notion of a right was subsequently to become a fundamental legal reasoning concept because it could act both as a focal point for legal thinking – legal disputes are about determining the 'rights' of the parties – and as a vehicle for importing philosophical, social, and (or) political ideas into law.

After the sixteenth century, therefore, legal reasoning could be framed not only around the interpretation and application of a rule (*regula*) but equally around the existence or nonexistence either of a remedy (*actio*) or of a right (*ius*). In other words several models of reasoning possibilities in law were now established. In particular one could view law objectively as a system of rules (*regulae*) or subjectively as a system of rights (*iura*).

Common Law Reaction

By the beginning of the twentieth century, then, the dominant form of legal reasoning in the civil law tradition was the axiomatic model, these 'axioms' finally being given expression in the great civil codes of France (1804) and Germany (1900). Legal solutions were determined through the use of deductive logic. However this formalism was to be challenged by the American common lawyers, who turned away from conceptual structuralism and hermeneutics to a different scheme of intelligibility, namely the functional approach (see e.g., Cohen, 1935). Yet to appreciate the impact of American Realism on common law thinking something obviously needs to be said about what legal reasoning has been in the common law tradition (see Common Law). And, as with the civilians, the most useful starting point is probably not a linear one determined by time; it is more useful to start midway and then to look both backward and forward.

According to Sir John Baker, the key period in the development of the English common law – at least in terms of legal thought and reasoning – was the sixteenth century (Baker, 2003). It was during this period that there began to develop a more sophisticated view of English legal history which, of course, involved a sense of it not being one involving Roman law or humanist learning. There was no blueprint like the *Corpus Iuris Civilis*; what had been built up over the previous centuries were, primarily, cases and these had been organized around a system of writs, or forms of action, that owed nothing to the Roman *actiones*. Looking backward from the sixteenth century there was not really a common law in the sense of a body of rules (*regulae*) or rights (*iura*) because the major decision-making institution was the jury, consisting of nonlawyers who did not give reasons for their decisions. As for judges, they were little more than umpires comparable to a sports referee (Baker, 2003: 43). Indeed even the Court of Chancery, whose procedure was so very different from the one that governed the three courts of common law, "behaved more like a jury than a court of law" (Baker, 2003: 42; and see article Equity in Law). There was no textbook law of personal property, contract, and delict (tort) because cases came to court through formal actions (or writs) such as debt or trespass and decisions were taken by juries largely on a mixture of fact and law. In the sixteenth century the position changed. Juries became less central and were gradually restricted to deciding questions of fact leaving judges to decide questions of law. The scene was set for the development of an English law – and of course an English form of legal reasoning and argumentation – founded on the decisions of the royal courts.

Looking forward from the sixteenth century, the position at "the beginning of the eighteenth century" was that "most common lawyers did not seek to put their law into rational form" because "they saw the law as a form of reasoning working within a set of remedies" (Lobban, 1991: 9). English lawyers may or may not have been interested in civilian learning, but there was certainly a suspicion of the deductive method. If there was a legal 'science,' it was a science of pleading. As Michael Lobban explains, this procedural attitude gave rise to a form of legal reasoning "which allowed lawyers to look outside the law for solutions to legal problems, once the

terms of the legal dispute had been set" (1991: 80). Thus, says the author, there were two types of reasoning; there was reasoning within the law (which was a matter of applying precedents by way of analogy) and reasoning outside the law (which "drew on a vast range of 'sources', invoking natural law, justice, political philosophy, political economy, and convenience"). If there was one driving force underpinning English legal reasoning at the end of the eighteenth century it was the desire "in judges' minds … to provide the legal answer which best served the needs of society, so that the common law was based less on concepts of 'justice' or 'natural law' and more on expedience and public policy" (1991: 90).

One might say, then, that the functional challenge – in the form of policy reasoning – is quite deeply embedded in the history of common law reasoning. Nevertheless the system of forms of action "was displaced in the nineteenth century when common lawyers became concerned not with the results that the courts were reaching … but with explaining these results" (Gordley, 2013: 27). Accordingly, with the abolition of the forms of action, effectively in 1852, the common lawyers turned to civil law for their categories and their theories (see Samuel, 2010: 8–9). A textbook tradition and the emergence of full time law faculties at universities that set out the law as a series of rules (*regulae*) arising either from precedent or from statute started to replace the old law of actions thinking; and once the idea of law as a set of coherent rules established itself English legal reasoning shifted toward a more syllogistic approach.

Realism was a reaction against this deductive approach but, as Gordley points out, it did not as such provide an alternative reasoning method (2013: 281). Nevertheless it did open up legal thinking to interdisciplinary alternatives that drew on ideas and concepts from the social sciences more generally (see Singer, 1988; Law: History of Its Relation to the Social Sciences). The results were that a range of approaches emerged, many of which either attached law to some other discipline or put the emphasis on concepts such as social interest and (or) policy (see Law and Economics: Empirical Dimensions). This move away from a strict structural conceptualism with its emphasis on logical deduction now characterizes much legal reasoning in the common law world (Lloyd and Freeman, 2008: 999). However it would be a mistake to think that this was a move away from 'scientific' method. What realist thinking permitted – or some kinds if realist thinking permitted – was the possibility of applying a causal scheme of intelligibility to legal argumentation and reasoning. A legal decision could be justified on the basis of, for example, promoting a particular social interest objective or some economic result (cause and effect). In turn a discipline like economics could 'explain' legal decisions both functionally and causally (Posner, 1983). At any rate, what realism has contributed to legal reasoning is a willingness on behalf of many common law judges to have recourse to policy as a form of argumentation and justification.

Legal Argumentation

This policy reasoning approach did not go unchallenged. Ronald Dworkin (1931–2013) forcefully asserted that cases should be decided on the basis not of policy but of rights (see Rights: Legal Aspects). He argued that the rights thesis gives a more satisfactory explanation than policy of how judges use precedent; rights can be justified within a theory of law whereas policy arguments cannot since the latter belong to the operation of the political process (Dworkin, 1977: 82–86). This thesis in turn is based upon a distinction between arguments of principle and arguments of policy.

This movement away from a functional scheme of intelligibility entailed a return to hermeneutics and to structuralism. Accordingly, to take the latter scheme first, what the ideal judge must do is to "construct a scheme of abstract and concrete principles that provides a coherent justification for all common law precedents and, so far as these are to be justified on principle, constitutional and statutory provisions as well" (Dworkin, 1977: 116–117). In addition, Dworkin insisted, legal reasoning and argument is about integrity; it is about "the best constructive interpretation of past legal decisions" (Dworkin, 1986: 262). Indeed law itself "is an interpretive concept"; and so judges "should decide what the law is by interpreting the practice of other judges deciding what law is" (1986: 410). Dworkin did not, however, see this constructionism as "an algorithm for the courtroom" (1986: 412). Rather it was a literary process. Thus the philosopher compared this hermeneutical process to the writing of a chain novel, that is to say "an enterprise [where] a group of novelists write a novel seriatim; each novelist in the chain interprets the chapters he has been given in order to write a new chapter, which is then added to what the next novelist receives, and so on." Each novelist is under an obligation, when writing his or her chapter, "to make the novel being constructed the best it can be, and the complexity of this task models the complexity of deciding a hard case under law as integrity" (1986: 229).

This is a very serious, and imaginative, attempt to give an account of legal reasoning, at least within the common law tradition. However it is not free from difficulty. The first and main problem is with the distinction between a proposition of principle and one of policy. Neil MacCormick observed that the word 'policy' is "a hideously inexact word in legal discourse" and, in fact, that when "we raise the question whether a given policy goal is desirable or not, we are raising a question of principle" (MacCormick, 1978: 263). Stephen Waddams has equally shown that from a historical point of view the distinction between principle and policy – at least in the area of the law of contract – is untenable (Waddams, 2011).

A second problem was perceived by Paul Ricœur. The application of a legal rule is a very complex operation where, said Ricœur, the interpretation of facts and that of the norm are mutually interrelated in a conditional way and that the facts are capable of being interpreted in different manners (1995: 180). Precedent is founded on the idea of similar cases, but when are cases similar? This is not just a question of textual interpretation; it is equally an issue of deciding when facts are similar and dissimilar. And, as Stephen Waddams noted, facts "may be stated at countless levels of particularity, and legal issues and legal rules may be formulated at countless levels of generality" (Waddams, 2003: 14). The importance of these remarks cannot be overstated. The idea that legal reasoning and argumentation is reasoning with respect to rules and texts is misleading. How a lawyer constructs the facts can be just as important in the argumentation, decision making, and justification processes

as the way she interprets a rule or a series of cases (Atias, 1985: 129; and see Samuel, 2003).

Models of Legal Reasoning

History, then, has bequeathed a number of different possible models which coexist, even if this coexistence is uneasy at times. Indeed the models often overlap and some are the result of a particular judicial culture. This history, as has been seen, is a long one characterized by different schools of jurists who, in the civil law tradition, brought to bear on the texts of Roman law different methods and techniques. Even within the common law tradition, so different in many ways from continental jurisprudence, the influence of Roman law has not been absent.

Rule Model

By far the most important model for legal reasoning is what might be called the rule model. The epistemological basis of this model is that knowledge of law is a matter of knowing rules (*regulae iuris*: D.50.17) – including principles (rules at higher levels of abstraction) – and thus applying the law is a matter of applying these statutory or precedent rules (and principles) to the factual litigation problems that present themselves. As the French Code of Procedure puts it, the "judge shall decide the litigation in accordance with the rules of law which are applicable to it" (article 12). The normal process by which this application takes place is through the syllogism (Perrot, 2010: 181–182; but cf Lasser, 2004). The rule is the major premise, the facts the minor premise and when the conditional elements of the rule are matched to the facts the solution logically follows (Champeil-Desplats, 2014: 80). This approach, as has been seen, is associated with the civilian rationalists whose writings began to dominate legal thinking in Europe from the sixteenth century onward; but it is also present in the common law (see *Lupton* v *FA & AB* [1972] AC 634, 658–659).

However an automatic inferential process is possible only when the conditions contained in the rule are unambiguous in the face of the facts; if there is ambiguity the syllogism is insufficient and what is required in addition is an interpretative process. A rule stating that vehicles are prohibited in the park can probably be syllogistically applied if a person enters on her motorbike, but the rule cannot be so easily applied to a child who enters on his small tricycle (see further Twining and Miers, 2010: 344–348).

Interpretation Model

This problem of uncertainty in language leads one toward another model of which (along with the rule model) is of primary importance given that most cases – even in the common law world – involve the application of a statutory text. It is unlikely that many disputes would reach litigation stage if there was not an interpretation problem. The kind of techniques brought to bear on the text depend upon the position of the observer; a lawyer taking a view from inside the discipline – and thus inside the authority paradigm – is likely to see the reasoning in terms of established legal rules of interpretation (see Twining and Miers, 2010: 230–267). A social scientist observing interpretation from outside the law discipline might well apply a rather different epistemological framework; how texts are interpreted often depends upon the scheme of intelligibility applied. A structural approach ('offer' has the same meaning throughout the law: see *Fisher* v *Bell* [1961] 1 QB 394) could well result in a different interpretation from a functional one (public interest or policy). A dialectical scheme founded on a *distinctio* (a 'lake' is defined in relation to it not being a 'pond') could give a different result from a hermeneutical one which seeks understanding of the legislator's intention (*mens legislatoris*).

Viewed from a historical perspective, the interpretation model has a long history, the starting point of which is the relevant material in the Roman sources (Jolowicz, 1957: 6–20). What emerges from the scholarship of the medieval and Enlightenment jurists is a series of dialectical oppositions between the words of a text (*verba*), the spirit of the law (*ratio iuris*), and the intention of the legislator (*mens legislatoris*) (Maclean, 1992). The interpretative model of law is, then, not just a sophisticated reworking of the rule model; it is a legal reasoning theory that injects into law a dynamic aspect. The judge, thought Dworkin, is an author in some huge project that is analogous to a never ending chain novel.

Rights Model

Dworkin equally emphasized that judges should concern themselves only with rights and not with policy arguments. This rights model has a tradition that, in terms of the word *ius*, goes back to Roman law. Yet "[t]he word 'right' is used in a variety of different senses, both popular and jurisprudential" and thus suffers from an inherent ambiguity (Lord Oliver in *In re KD* (*a minor*) [1988] 1 AC 806, 825). There have, of course, been attempts to provide clarity. For example while some have tried to create an abstract conceptual model from which rights could be inferred or deduced as a matter of logic, other writers have attempted to define rights, bottom up so to speak, in relation to empirical notions such as social interests. A right is a particular social interest that the law has endowed as such thus elevating it above other, conflicting, social interests that the law does not treat as rights. All of these definitions are not without their difficulties, but the rights model has been developed by a generation of younger jurists working in the area of private law (see e.g., Robertson and Wu, 2009). According to these jurists, reasoning in law is about determining and balancing rights that individuals have with respect to each other and involves distinguishing rights, for example, from interests, from liberties, and from privileges.

Policy Model

As has been mentioned, Dworkin developed his rights thesis in opposition to policy reasoning. This latter model is functional (or instrumental) in its orientation in that it looks to the social, economic, or (and) political purpose of a rule or concept in order to justify a particular interpretation and decision. In this model legal reasoning is about justifying decisions in terms not of logic or recourse to some abstract conceptual structure of

rights but of social or economic goals. According to one English appeal judge:

> In previous times, when faced with a new problem, the judges have not openly asked themselves the question: what is the best policy for the law to adopt? But the question has always been there in the background. It has been concealed behind such questions as: Was the defendant under a duty to the plaintiff? Was the relationship between them sufficiently proximate? Was the injury direct or indirect? Was it foreseeable, or not? Was it too remote? And so forth.
>
> Lord Denning MR in *Dutton* v *Bognor Regis UDC* [1972] 1 QB 373, 397

This approach has been taken to a more sophisticated level by the law and economics school which considers that legal reasoning should largely be founded on a mixture of the functional and causal schemes of intelligibility (see Law and Economics: Empirical Dimensions). Decisions are to be justified by reference to economic goals. Yet one problem with this functional model is that judges often do not support their policy arguments with any serious economic or social science research and this is what leaves them open to attack from jurists who are more formalist in their approaches.

Remedies Model

Another approach that stands in contrast to the rights model is one that sees legal reasoning as a matter of categorization of facts in relation to the availability of legal actions. It has been well expressed by Sir Nicolas Browne-Wilkinson who observed that in "the pragmatic way in which English law has developed, a man's legal rights are in fact those which are protected by a cause of action." He continued by saying that it "is not in accordance ... with the principles of English law to analyze rights as being something separate from the remedy given to the individual." Consequently, "in the ordinary case to establish a legal or equitable right you have to show that all the necessary elements of the cause of action are either present or threatened" (*Kingdom of Spain* v *Christie, Manson & Woods Ltd* [1986] 1 WLR 1120, 1129). As for a 'cause of action,' it is "simply a factual situation the existence of which entitles one person to obtain from the court a remedy against another person" (*Letang* v *Cooper* [1965] 1 QB 232, 242).

If the rights model is characterized by the expression *ubi ius ibi remedium* (where there is a right there is a remedy), the remedy model finds expression in a reversal of this maxim: *ubi remedium ibi ius*. It is a model that seems to apply both to early Roman law and to early English law and tends to generate a form of thinking where the emphasis is on the factual situation. One examines the facts to see if there is a 'cause' (see D.9.2.52.2). The method of reasoning associated with this model is one of analogy: do the facts of a potential litigation match the factual situation envisaged by the cause of action and remedy?

Concluding Remarks

These models are not exhaustive and, moreover, they largely represent a viewpoint from inside the authority paradigm. They are models that might well be discussed in an introductory work on legal reasoning aimed at law students and one imagines that the judiciary would find little of it threatening. One reason for this 'internalism' is that the foundations, and to some extent the limits, of the debate have been determined by the Roman jurists and the Latin concepts and categories that they employed. Expressions such as *ius*, *regulae iuris*, *actiones*, *verba*, *ratio iuris*, *scientia iuris*, *ius naturale*, *interesse*, *utilitas*, and so on have acted as focal points for schemes of reasoning and a good many modern theories can be traced back to these notions via the medieval jurists (Jones, 1940). As has been seen, the common lawyers were not immune to this continental theorizing. This is not to suggest, of course, that any Roman jurist, if miraculously transported into the future in a time machine, would have understood any of the post sixteenth century literature on Roman law. But these modern jurists, at least until recently, had a different epistemological outlook than either the Romans or the medieval thinkers; they were engaged in a search for intellectual coherence since this would help endow law with a scientific validity. Just as theologians once used systems thinking and its coherence to affirm the Christian message (Riaudel, 2007; see article Science and Religion), so jurists turned to the same methodological approach to affirm the (claimed) scientific message – the authority – inherent in the Roman texts. A systems approach also – and most importantly – merged the diachronic with the synchronic. It thus held (seemingly) enormous promise for rationalist jurists.

If one steps outside of the authority paradigm, then perhaps a different set of models would be more appropriate. A social science epistemologist would, perhaps, prefer to explain legal reasoning through the six schemes of intelligibility identified by Jean-Michel Berthelot. All the same, this theorist was himself rather dismissive about law's contribution to social science epistemology; what he called the 'science of texts' either defines itself by a specific object approached in interdisciplinary fashion or makes use of a hermeneutical program that is not its own (2001: 12). According to Berthelot, then, law has little as a discipline to contribute to epistemology. Perhaps this present survey will confirm this opinion or, in contrast, it could identify a particular body of reasoning techniques that creates its own institutional reality. This is a closed and formalized world of *personae*, *res*, *actiones*, *iura*, and *regulae* in which a range of schemes of intelligibility and inferential reasoning techniques – induction, deduction, and analogy – function just as in other areas of knowledge. Accordingly, analyzed into its constituent parts, there is, one might say, nothing special about legal reasoning. Taken as a formalized whole, however, there might be more than the sum of its parts. A client involved with a dispute will expect from the professional lawyer a prediction as to how a court might decide the case or, if not that, what are his or her 'rights.' For "[e]lementary justice or ... the need for legal certainty demands that the rules by which a citizen is to be bound should be ascertainable by him (or, more realistically, by a competent lawyer advising him)" (Lord Diplock in *Fothergill* v *Monarch Airlines* [1981] AC 251, 279).

In giving a reasoned response the 'competent lawyer' might not apply the interpretative integrity demanded by Dworkin or the deductive rigor desired by a nineteenth century German jurist. The lawyer might not even turn to economics or some other discipline to supply a fully functional response. But the lawyer will, probably, apply a mixture of hermeneutical,

structural, functional, and or dialectical techniques, together with the standard forms of inference (induction, deduction, and analogy, depending on the source of the laws applicable), to the facts as described by the client and reconstructed by the lawyer. And the lawyer will do this with reference to a formalized model of persons, things, and remedies between which flow a range of concepts and relations themselves organized by a taxonomy that has been determined as much by history – in particular the Romans – as by rationality.

While legal reasoning could be described as managing uncertainty, it ideologically cannot easily admit this. Justice demands certainty. Law sets its own limits within an authority paradigm that bolsters this authority through a belief in coherence and thus predictability (see e.g., Robertson, 2009). Legal reasoning is constantly under disciplinary pressure that only conceals its fragmented epistemological foundation. It functions 'as if' there is a coherent model of rules and concepts 'out there' so to speak; and it is this consensual model that acts as the framework for a set of reasoning techniques by a group of 'disciplined' professionals. There may be disagreement about the elements of the model (rules, rights, remedies, or whatever) but there is consent that there is a model 'out there' that governs. It is arguably all a fiction (Vaihinger, 1924, 2013), but that is another story.

See also: Common Law; Courts and Adjudication; Disputes, Social Construction and Transformation of; Equity in Law; Human Sciences, History of; Judges; Law and Economics: Empirical Dimensions; Law as Constitutive; Law: History of Its Relation to the Social Sciences; Legal Culture and Legal Consciousness; Legal Formalism; Legal Positivism; Legal Realism; Legal Reasoning: Computational Models; Legal Scholarship; Natural Law; Rights: Legal Aspects; Science and Law; Science and Religion; Truth and Credibility in Science.

Bibliography

Alexander, I., Sherwin, E., 2008. Demystifying Legal Reasoning. Cambridge University Press.
Amselek, P. (Ed.), 1995. Interprétation en droit. Bruylant.
Atias, C., 1985. Épistémologie Juridique. Presses Universitaires de France.
Atias, C., 1994. Épistémologie du droit. Presses Universitaires de France.
Atias, C., 2002. Épistémologie Juridique. Dalloz.
Baker, J., 2003. The Oxford History of the Laws of England: Volume VI 1483–1558. Oxford University Press.
Bergel, J.-L., 2012. Théorie générale du droit, fifth ed. Dalloz.
Berthelot, J.-M., 1990. L'intelligence du social. Presses Universitaires de France.
Berthelot, J.-M. (Ed.), 2001. Épistémologie des sciences sociales. Presses Universitaires de France.
Berthelot, J.-M., 2001a. In: Berthelot (Ed.), Programmes, Paradigmes, Disciplines: pluralité et unité des sciences sociales, p. 457.
Birks, P. (Ed.), 1997. The Classification of Obligations. Oxford University press.
Birks, P., 1997a. In: Birks (Ed.), Definition and Division: A Meditation on Institutes 3.13, p. 1.
Blanché, R., 1983. L'épistémologie, third ed. Presses Universitaires de France.
Canning, J., 1987. The Political Thought of Baldus de Ubaldis. Cambridge University Press.
Champeil-Desplats, V., 2014. Méthodologies du droit et des sciences du droit. Dalloz.
Cohen, F., 1935. Transcendental nonsense and the functional approach. Columbia Law Review 35, 809.
Desjeux, D., 2004. Les Sciences Sociales. Presses Universitaires de France.
Dworkin, R., 1977. Taking Rights Seriously. Duckworth.
Dworkin, R., 1985. A Matter of Principle. Oxford University Press.
Dworkin, R., 1986. Law's Empire. Fontana.
Goltzberg, S., 2013. Chaïm Perelman : L'argumentation Juridique. Michalon.
Gordley, J., 2013. The Jurists: A Critical Histor. Oxford University Press.
Jestaz, P., Jamin, C., 2004. La Doctrine. Dalloz.
Jolowicz, H., 1957. Roman Foundations of Modern Law. Oxford University Press.
Jones, J.W., 1940. Historical Introduction to the Theory of Law. Oxford University Press.
Jonsen, A.R., Toulmin, S., 1988. The Abuse of Casuistry: A History of Moral Reasoning. University of California Press.
Kelley, D., 1970. Foundations of Modern Historical Scholarship. Columbia University Press.
Kelley, D., 1990. The Human Measure: Social Thought in the Western Legal Tradition. Harvard University Press.
Lacoste, J.-Y. (Ed.), 2007. Dictionnaire critique de théologie, second ed. Presses Universitaires de France.
Lasser, M., 2004. Judicial Deliberations: A Comparative Analysis of Judicial Transparency and Legitimacy. Oxford University Press.
Llewellyn, K., 1951. The Bramble Bush. Oceana.
Lobban, M., 1991. The Common Law and English Jurisprudence 1760–1850. Oxford University Press.
Lloyd, D. (Lord), Freeman, M., 2008. Lloyd's Introduction to Jurisprudence, eighth ed. Sweet & Maxwell.
MacCormick, N., 1978. Legal Reasoning and Legal Theory. Oxford University Press.
Maclean, I., 1992. Interpretation and Meaning in the Renaissance. Cambridge University Press.
Makkreel, R., 2006. In: Mesure, Savidan (Eds.), Expliquer et comprendre, p. 441.
Masquelet, A.C., 2006. Le Raisonnement Médical. Presses Universitaires de France.
Mestre, J.-L., 1985. Introduction historique au droit administratif français. Presses Universitaires de France.
Mesure, S., Savidan, P. (Eds.), 2006. Le dictionnaire des sciences humaines. Presses Universitaires de France.
Perelman, C., 1979. Logique Juridique: Nouvelle Rhétorique, second ed. Dalloz.
Perrot, R., 2010. Institutions Judiciaires, fourteenth ed. Montchrestien.
Piaget, J., 1988. L'épistémologie Génétique, fourth ed. Presses Universitaires de France.
Pichonnaz, P., 2008. Les fondements romains du droit privé. Schulthess.
Posner, R., 1983. The Economics of Justice. Harvard University Press.
Riaudel, O., 2007. In: Lacoste (Ed.), Systématique (Théologie), p. 1367.
Ricœur, P., 1995. In: Amselek (Ed.), Le problème de la liberté de l'interprète en herméneutique générale et en herméneutique juridique, p. 177.
Robertson, A., 2009. In: Robertson, Wu (Eds.), Constraints on Policy-based Reasoning in Private Law, p. 261.
Robertson, A., Wu, T.H. (Eds.), 2009. The Goals of Private Law. Hart.
Samuel, G., 2003. Epistemology and Method in Law. Ashgate.
Samuel, G., 2008. Is law really a social science? A view from comparative law. Cambridge Law Journal, 288.
Samuel, G., 2009. Interdisciplinarity and the authority paradigm: should law be taken seriously by scientists and social scientists? Journal of Law and Society 36, 431.
Samuel, G., 2009a. Can legal reasoning be demystified? Legal Studies 29, 181.
Samuel, G., 2010. Law of Obligations. Edward Elgar.
Samuel, G., 2013. A Short Introduction to the Common Law. Edward Elgar.
Singer, J., 1988. Legal realism now. California Law Review 76, 465.
Stein, P., 1966. Regulae Iuris: From Juristic Rules to Legal Maxims. Edinburgh University Press.
Stein, P., 1980. Legal Evolution: The Story of an Idea. Cambridge University Press.
Stein, P., 1999. Roman Law in European History. Cambridge.
Susskind, R., 1987. Expert Systems in Law. Oxford University Press.
Thireau, J.-L., 2001. Introduction historique au droit. Flammarion.
Tierney, B., 1982. Religion, Law, and the Growth of Constitutional Thought 1150–1650. Cambridge University Press.
Twining, W., Miers, D., 2010. How to Do Things with Rules, fifth ed. Cambridge University Press.
Ullmann, W., 1975. Law and Politics in the Middle Ages. Sources of History.
Vaihinger, H., 1924. The Philosophy of "As if" (C.K. Ogden, Trans.). Routledge & Keegan Paul.
Vaihinger, H., 2013. La philosophie du comme si (C. Bouriau, Trans.), second ed. Éditions Kimé.
Valade, B., 2001. In: Berthelot (Ed.), De l'explication dans les sciences sociales: holisme et individualisme, p. 357.
Van Hoecke, M. (Ed.), 2011. Methodologies of Legal Research: Which Kind of Method for Which Kind of Discipline?. Hart.
Van Hoecke, M., 2011a. Legal Doctrine: Which Method(s)? In: Van Hoecke (Ed.), p. 1.
Villey, M., 2006. La formation de la pensée juridique moderne. Presses Universitaires de France, Quadrige.
Waddams, S., 2003. Dimensions of Private Law: Categories and Concepts in Anglo-american Legal Reasoning. Cambridge University Press.
Waddams, S., 2011. Principle and Policy in Contract Law: Competing or Complementary Concepts? Cambridge University Press.

Legal Reasoning: Computational Models

Henry Prakken, Utrecht University, Utrecht, The Netherlands; and University of Groningen, Groningen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews computational models of legal reasoning as they are being developed in the field of artificial intelligence and law. A theoretical aim of these models is to understand legal reasoning by simulating it in a computer program, while a practical aim is to study how advanced information technology can aid legal practice. First logical deduction is discussed as a necessary but insufficient component of realistic models of legal reasoning. Then models of defeasible legal reasoning and legal argumentation are discussed, which focus on the generation and comparison of reasons or arguments for and against legal claims. Computational models of legal interpretation emphasize the interplay between rules and cases and the role of principles, purposes, and values. Computational models of legal proof account for the uncertainty in legal proof in three alternative ways: with Bayesian probability theory, argumentation, and narrative. Finally, procedural models of legal reasoning are based on the idea that the quality of a legal decision partly depends on how it was reached.

Introduction

This article reviews computational models of legal reasoning as they are being developed in the field of artificial intelligence (AI) and law. (Some parts of the text of this article are adapted from Prakken and Sartor (2002) or Prakken and Sartor (submitted).) Computational models of legal reasoning have been developed for two main purposes. A theoretical aim is to understand legal reasoning by simulating it in a computer program, while a practical aim is to study how advanced information technology can aid legal practice. Both types of research have had considerable success. Computational models of legal argument have advanced the understanding of legal reasoning, while legal knowledge-based systems are in routine use in areas of legal administrative decision making in several countries.

Some may fear that using computational methods in the law enforces mechanical law application, without regard for social consequences or for principles of fairness and justice. However, this is a misconception. Computational approaches require clarity of meaning and explicitness of structure, but this should not be confused with mechanical law application or with the view that the only valid source of the law is legal rules. Although some early systems were open to this criticism, there is nothing in a computational approach that requires judges to mechanically apply rules without looking at the consequences of applying the rule. On the contrary, computational models exist of arguing why an applicable rule should not be applied to a case or why a nonapplicable rule should be broadened to make it apply. Likewise, there is nothing in a computational approach that insists on the uniqueness or completeness of interpretations of the law. On the contrary, computational models exist of arguing for alternative interpretations and about their relative merits. Finally, computational models of legal reasoning exist that acknowledge cases, principles, or opinions in society about justice and fairness as valid sources of the law.

Models of legal reasoning must be diverse. A legal case has various aspects, each with its own modes of reasoning: determining the facts, classifying the facts under legal concepts or conditions, and deriving legal consequences from the thus classified facts. When determining the facts, the modes of reasoning are often probabilistic and may involve reasoning about causation and about mental attitudes such as intent. Classifying the facts under legal concepts involves interpretation. Here the prevailing modes of reasoning are analogy, appeals to precedent or policy, and the balancing of interests. Finally, when deriving legal consequences from the classified facts, the main modes of reasoning are deductive but leave room for exceptions to rules, either statutory or based on principle and purpose, and have room for conflict based on the general hierarchy of legal systems, where rules derive from different sources. Models of legal reasoning also depend on the task that is modeled, for example, advice, advocacy, or decision making. This article will primarily focus on the context of a court proceeding, where a judge has to adjudicate a conflict between two adversaries.

About any computational model of reasoning three questions can be asked:

1. What is the underlying theory of reasoning?
2. Can the knowledge required by the reasoning be represented in the computer?
3. Can the computer reason efficiently with this knowledge?

This article will primarily focus on the first question but the other two will not be totally ignored. The third question is addressed in other areas of AI, and the resulting techniques can be used in AI and law when available. The problem addressed by the second question has so far been the bottleneck, hindering scaling computational models of legal reasoning up to practical use. Where it has been solved (either by explicit representation or by automatically learning the knowledge from data), feasible and successful applications have resulted.

Legal Reasoning as Deduction

A simple but naive model of legal reasoning sees it as logically deducing legal consequences from a precisely stated body of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86161-9

facts and legal rules. On this account, once a legal text and a body of facts have been clearly represented in a logical language, the valid inferences are determined by the meaning of the representations and so techniques of automated deduction apply. However, this mechanical approach leaves out most of what is important in legal reasoning. To start with, the law is not just a conceptual or axiomatic system but has social objectives and social effects, which must be taken into account when applying the law. Moreover, legislators can never fully predict in which circumstances the law has to be applied, so legislation has to be formulated in general and abstract terms, such as 'duty of care,' 'misuse of trade secrets,' or 'intent,' and qualified with general exception categories, such as 'self-defense,' 'force majeure,' or 'unreasonable.' Such concepts and exceptions must be interpreted in concrete cases, a process which creates room for doubt and disagreement. This is reinforced by the fact that legal cases often involve conflicting interests of opposing parties. The prosecution in a criminal case wants the accused convicted while the accused wants to be acquitted. The plaintiff in a civil-law suit wants to be awarded compensation for damages, while the defendant wants to avoid having to pay. The tax authority in a tax case wants to receive as much tax as possible, while the taxpayer wants to pay as little as possible. Both aspects of the law, that is, the tension between the general terms of the law and the particulars of a case, and the adversarial nature of legal procedures, cause legal reasoning to go beyond the meaning of the legal rules. It involves appeals to precedent, principle, policy, and purpose, as well as the consideration of reasons for and against drawing certain conclusions. For these reasons law relies not just on deduction but on argument.

Nevertheless, deduction is still an essential ingredient in legal reasoning. To start with, the principle of legal certainty requires that the meaning of the law is not arbitrarily set aside. And when arguing that the law should be set aside in a certain case because applying it would have undesirable effects, the effects of applying the law must be deduced. Also, contemplating alternative interpretations of a legal concept involves deducing how the various interpretations affect the outcome of a case. Moreover, deductive techniques have been practically successful, especially in the application of knowledge-based systems in large-scale processing of administrative law, such as social benefit law. Such systems apply computational representations of legislation to the facts as interpreted by the human user. The use of such systems has been proved to greatly reduce two major sources of errors in the processing of social benefit applications by 'street-level bureaucrats': their incomplete knowledge of the relevant regulations and their inability to handle the often complex conditions of the regulations (Svensson, 2002). These systems leave it to the user (the official with the authority to make a decision) to decide whether to accept the system's recommendation or to deviate from it on nonstatutory grounds. Thus these systems do not enforce mechanical application of the law, at least as long as the correctness of their recommendations is not taken for granted. We can conclude that, even though legal reasoning does not equate with logical deduction, the study of the deductive aspects of legal reasoning is important for both theoretical and practical reasons.

Much research on these aspects exist and we can discuss only some of it here. A seminal study of the logical structure of legislation was Sergot et al.'s (1986) modeling of the British Nationality Act as a logic program, which showed how logical programming (and logic more generally) enables intuitively appealing representations that can be directly deployed to generate automatic inferences.

Here is how a legislative rule from the British Nationality Act is modeled as a logical clause:

x is a British citizen.
IF x was born in the United Kingdom AND x was born on date y
AND y is after or on commencement
AND x has a parent who qualifies under 1.1 on date y

Sergot et al. also showed how further clauses can be represented that determine the conditions under which the conditions in the legal rules hold. This work has greatly influenced subsequent research on knowledge-based systems for processing legislation.

Other important work is on the logical characterization of the various legal positions or relations that may exist between normative agents as regards an action or state of affairs, in terms of their rights, obligations, and permissions. This work enables the expression of subtle but important distinctions, such as whether a patient's normative position with regard to access to his or her medical data, besides the patient's permission to access the data, does or does not include the medical staff being forbidden to impede the patient's access. In the first case, the patient has the right to access his or her medical data; in the second case he or she is merely permitted to do so. Sergot (2001) presents a computational method for exhaustively specifying and computing the alternative normative positions of two parties concerning an action. Such methods can be useful, for example, for drafting or disambiguating security or access policies.

Among other work on deductive aspects of legal reasoning we mention methods for representing systems of legal concepts (Bench-Capon and Coenen, 1992) and methods for representing change in legal systems, normative positions, or legal relations over time (Rotolo and Governatori, 2009).

Exceptions and Rule Hierarchies

Any attempt to computationally model legal regulations soon meets two very common structural features of legal regulations that provide obstacles for deductive techniques: the separation of general rules and exceptions, and the use of hierarchies over legislative sources to resolve conflicts between different regulations. With exceptions the problem is that the law often makes legal effects dependent on the assumption that they do not occur as long as there is no evidence for their occurrence. For example, Section 3:32 of the Dutch civil code states that persons have the capacity to perform legal acts (this means, for instance, that they can engage in contracts or sell their property), unless the law provides otherwise. One place in which the law provides otherwise is Section 1:234 of the Dutch civil code, which states that minors have the capacity to perform legal acts if and only if they have consent from their legal representative. If we take a person named John and then consider these two

rules, it follows that John has the capacity to perform legal acts. However, this inference is presumptive or defeasible in that it can be invalidated by further information. For example, if evidence is provided that John is a minor, then the conclusion that he can perform legal acts is invalidated unless evidence is also provided that John has a consent of his legal representative (usually one of his parents).

The making and retracting of such defeasible inferences goes beyond standard deductive models of reasoning. One way in which AI and law has tackled this problem is by the use of so-called nonmonotonic logics. Such logics, which are a result of general AI research, allow the making of inferences subject to evidence to the contrary, and the retraction of such inferences when evidence to the contrary comes in. These logics can also model reasoning with hierarchies. Models of deduction have no room for conflict resolution or hierarchies: when a body of information is logically inconsistent, then standard deductive methods simply come to a halt. In contrast, nonmonotonic logics have been designed in which inconsistencies can be resolved by defining preference relations on rules, and these preferences can be used to express legal hierarchies. This too makes the reasoning defeasible, since new information may make a conflicting rule of higher rank applicable that overrules a previously applicable rule with lower rank. Nonmonotonic logics have been shown useful in modeling legislative rule-exception structures and legislative hierarchies (Prakken and Sartor, 1996), and in modeling legal presumptions and notions of burdens of proof (Prakken and Sartor, 2009; Gordon and Walton, 2009).

Although nonmonotonic techniques technically deviate from deductive logic, their spirit is still the same, namely, of deriving consequences from clear and unambiguous representations of legal rules and facts. The logical consequences of a representation are still clearly defined. While technically most nonmonotonic logics allow for alternative conclusion sets (which capture cases in which conflicts cannot be resolved), in legal practice statutory rule-exception structures and legislative hierarchies still usually yield unambiguous outcomes. More often, conflicts arise not from competing norms but from the variety of ways in which they can be interpreted. A real challenge for deductive accounts of legal reasoning is the gap between the general legal language and the particulars of a case. Because of this gap, disagreement can arise, and it will arise because of the conflicts of interests between the parties.

Rule-Based Accounts of Legal Interpretation

At first sight, it might be thought that disagreements about interpretation are resolved in concrete cases by courts, so that additional interpretation rules can be found in case law. If different courts disagree on an interpretation, such disagreements could be represented with nonmonotonic techniques for handling conflicting rules. This was indeed the approach taken by Gardner (1987), who designed a program for so-called 'issue spotting' in law school contract law exam problems. Given an input case, the task of the program was to determine which legal questions involved were easy and which were hard, and to solve the easy ones. If all the questions were found easy, the program reported the case as clear, otherwise as hard. The system contained domain knowledge of three different types: legal rules, common-sense rules (e.g., on the interpretation of utterances like "Will you supply …?"), and rules extracted from cases. The program considered a question as hard if either 'the rules run out,' or different rules or cases point at different solutions, without there being any reason to prefer one over the other. Before a case was reported as hard, conflicting alternatives were compared to check whether one is preferred over the other. For example, case law sets aside conflicting legal rules or common-sense interpretations of legal concepts. Gardner's program was reconstructed by Gordon (1991) with techniques from nonmonotonic logic. Another approach is Verheij et al.'s Reason-Based Logic (1998), which models reasoning as the weighing of reasons for and against a conclusion, where rules are just one source of reasons, along with, for example, precedents or principles. Thus they could, like Gardner, model a statutory rule being set aside not by a conflicting statutory rule but by other considerations.

A problem with rule-based approaches is that rules derived from precedents are often specific to the case, so a new case will rarely exactly match the precedent, and techniques for handling conflicting rules fall short. Instead, reasoning forms are called for in which case law rules can be refined and modified. Here factors and reasons play an important role, and analogies between cases are drawn and criticized.

Interpreting Legal Concepts with Cases

What are the reasoning forms "when the rules run out"? They can be found in their most explicit form in common-law jurisdictions, in which judicial precedents can be legally binding beyond the decided case, so that court decisions legally constrain the decision in new cases. This leads to reasoning forms where legal rules are formulated by courts in the context of particular cases and are constantly refined and modified to fit new circumstances that were not taken into account in earlier decisions. These reasoning forms can to a lesser extent also be found in civil-law jurisdictions, since interpretations of the law by higher civil-law courts, even though strictly speaking not binding beyond the decided case, still tend to be followed by lower courts.

Perhaps the earliest work in AI and law on legal interpretation was the TAXMAN II project of McCarty (1977, 1995). According to McCarty (1995: p. 285), "The task for a lawyer or a judge in a 'hard case' is to construct a theory of the disputed rules that produces the desired legal result, and then to persuade the relevant audience that this theory is preferable to any theories offered by an opponent." In McCarty's approach, legal concepts have three components: firstly, a (possibly empty) set of necessary conditions for the concept's applicability; secondly, a set of instances ('exemplars') of the concept; and finally, a set of rules for transforming one case into another, particularly for relating 'prototypical' exemplars to 'deformations.' According to McCarty, the way lawyers typically argue about application of a concept to a new case is by finding a plausible sequence of transformations which maps a prototype, possibly via other cases, onto the new case. McCarty's particular modeling of this approach has not found many followers but his general claim that legal interpretation is theory construction has been very influential.

Much AI and law work on the interpretation of legal concepts centers around the notion of a factor, an idea going back to the HYPO system of Ashley (1990) and the CATO system of Aleven (2003). Factors are abstractions of fact patterns that favor (pro factors) or oppose (con factors) a conclusion. Factors are thus in an intermediate position between the specific facts of a case and the legal predicates to which such facts may be relevant. For example, in CATO, which like HYPO argues about misuse of trade secrets, some factors pro misuse are that a nondisclosure agreement was signed, that the plaintiff had made efforts to maintain secrecy and that the copied product was unique; and some factors con misuse are that disclosures were made by the plaintiff in negotiations and that the information was reverse-engineerable.

The HYPO and CATO systems are meant to model how lawyers make use of past decisions when arguing a case. They do not compute an 'outcome' or 'winner' of a dispute; instead they are meant to generate debates as they could take place between 'good' lawyers. HYPO generates disputes between a plaintiff and a defendant of a legal claim concerning misuse of a trade secret. Each move conforms to certain rules for analogizing and distinguishing precedents. These rules determine for each side, which are the best cases to cite initially, or in response to the counterparty's move, and how the counterparty's cases can be distinguished. According to the theory described by Ashley (1989) and implemented in CATO, a case is represented as a set of factors for a decision and a set of factors against that decision, plus the decision that resolves the conflict between the competing factors. A case is citable for a side if it has the decision wished by that side and shares with the current fact situation (CFS) at least one factor that favors that decision. Thus citable cases do not have to exactly match the CFS, which is a way of coping with the case-specific nature of case law decisions. A citation can be countered by a counterexample, that is, by producing a citable case that has the opposite outcome. A citation may also be countered by distinguishing, that is, by indicating a factor in the CFS that is absent in the cited precedent and that supports the opposite outcome, or a factor in the precedent that is missing in the CFS and that supports the outcome of the cited case. HYPO also allows for multivalued factors, called dimensions, to vary the degree to which a factor promotes a certain outcome. For example, a dimension is the number of people to which a trade secret has been disclosed, or the extent to which security measures were taken. A Boolean factor is then a specific value of a dimension. Dimensions allow an additional way to distinguish a precedent, namely, on a shared pro-decision factor that more strongly favors the decision in the precedent than in the CFS.

The HYPO system was embedded in the CABARET system of Skalak and Rissland (1992), which modeled the use of precedents to confirm or contest the application of a rule. The system implemented a hierarchical argumentation model in terms of strategies, moves, and primitives. A strategy is a broad characterization of how one should argue, given one's particular viewpoint and dialectical situation. A move is a way to carry out the strategy, while a primitive is a way to implement a move. For example, when one wants to apply a rule, and not all of the rule's conditions are satisfied, then a possible strategy is to broaden the rule. This strategy can be implemented with a move that argues, using an analogized precedent, that the missing condition is not really necessary. This move can in turn be implemented with HYPO's ways to analogize cases. Similarly, CABARET also permits arguments that a rule, which prima facie appears to cover the case, should not be applied to it. Here the strategy is discrediting a rule and the move may consist of analogizing a case in which the rule's conditions were met but the rule was not applied. Again the move can be implemented with HYPO's ways to analogize cases.

Aleven (2003) developed HYPO (without dimensions) into the CATO system for teaching case-based argumentation skills to law students, also in the trade secrets domain. CATO's main new component was a 'factor hierarchy,' which expresses expert knowledge about the relations between the various factors: more concrete factors are classified according to whether they are a reason for or against the more abstract factors to which they are linked; links are given a strength (weak or strong), which can be used to solve certain conflicts. Thus the factor hierarchy can be used to explain why a certain decision was taken, which in turn facilitates debate on the relevance of differences between cases.

For instance, the hierarchy positively links the factor 'security measures taken' to the more abstract concept 'efforts to maintain secrecy.' Now if a precedent contains the first factor but the CFS lacks it, then not only can a citation of the precedent be distinguished on the absence of 'security measures taken' but also this distinction can be emphasized by saying that thus no efforts were made to maintain secrecy. However, if the CFS also contains a factor 'agreed not to disclose information,' then the factor hierarchy enables downplaying this distinction, since it also positively links this factor to 'efforts to maintain secrecy': the party that cited the precedent can say that in the current case, just as in the precedent, efforts were made to maintain secrecy.

Perhaps the most elaborate representation of cases was produced in Branting's (1999) GREBE system in the domain of industrial injury, where cases were represented as semantic networks. The program matched portions of the network for the new case with parts of the networks of precedents, to identify appropriate analogies.

Various logical accounts of factor-based reasoning with cases have been given. Here the nonmonotonic logics that were originally used to model reasoning with exceptions and legal hierarchies turned out to be useful for this task also. Most of this work applies argumentation-based nonmonotonic logic, in which rules can be repeatedly applied to construct arguments, and in which conclusions are drawn by comparing alternative arguments for or against conclusions. (But some accounts are formulated in terms of weighing sets of reasons for or against a conclusion.) The key idea in these approaches is that existing case decisions give rise to conflicting rules (or conflicting sets of reasons) plus a preference expressing how the court resolved this conflict: In the notation of Prakken and Sartor (1998):

$$r_1 \qquad \text{Pro factors} \Rightarrow \text{Decision}$$

$$r_2 \qquad \text{Con factors} \Rightarrow \text{Not Decision}$$

$$r_1 > r_2$$

The priority expresses the court's decision that the pro factors in the body of rule r_1 together outweigh the con factors in the body of rule r_2. This approach allows for 'a fortiori' reasoning, in which adding factors to a pro-decision rule or removing factors from a con-decision rule does not affect the rule priority. It allows for analogical uses of a pro-decision rule by deleting pro factors, which is called 'broadening a rule.' A broadened rule does not inherit the priority relations of the rule from which it is obtained. So if rule r_1 in our schematic example is broadened to rule r_1' by deleting one of the pro-factors in rule r_1, then one cannot conclude from the priority $r_1 > r_2$ in the precedent that $r_1' > r_2$, since the deleted factor might have been essential in preferring r_1 over r_2. So a decision argued for with a broadened rule is not (defeasibly) implied by a body of precedents.

An interesting question is under what conditions such a decision is still allowed by the body of precedents. This question is addressed by Horty (2011). Consider the Keeble case from the common law of property, part of a well-known series of cases on ownership of wild animals that are being chased. In Keeble a pond owner placed a duck decoy in his pond with the intention to sell the caught ducks for a living. Defendant used a gun to scare away the ducks, for no other reason than to damage plaintiff's business. Here the court held for plaintiff. The pro-plaintiff factors were that plaintiff was hunting for a living (*PlLiving*) and was hunting on his own land (*OwnLand*). The single pro-defendant factor was that the animals were not yet caught (*Not Caught*). The issue is whether plaintiff became the owner of the ducks (Owner). In the notation of Prakken and Sartor (1998), we have

Keeble
k_1: PlLiving, OwnLand $\Rightarrow$ Owner
k_2: Not Caught $\Rightarrow$ Not Owner
$k_1 > k_2$

In another precedent, Young, both plaintiff and defendant were fishermen fishing in the open sea. Just before plaintiff closed his net, defendant came in and caught the fishes with his own net. Here not only the plaintiff but also the defendant was hunting for a living (*DefLiving*). Then we have

Young
γ_1: PlLiving $\Rightarrow$ Owner
γ_2: Not Caught, DefLiving $\Rightarrow$ Not Owner

To decide *Young* for plaintiff, the required priority $\gamma_1 > \gamma_2$ cannot be based on the precedent, since γ_1 lacks one antecedent of k_1, and also since γ_2 adds a con factor to k_2. However, deciding Young in accordance with Keeble is still allowed by Keeble, since that leaves the decision in the precedent unaffected. Horty calls this following the precedent.

This is different if the case base also contains a second precedent, which is almost like Keeble except that the defendant also hunted for a living and in which the defendant won.

Precedent 2
p_1: PlLiving, OwnLand $\Rightarrow$ Owner
p_2: Not Caught, DefLiving $\Rightarrow$ Not Owner
$p_2 > p1$

The priority $p_2 > p_1$ then implies $\gamma_2 > \gamma_1$, since γ_2 has the same con-decision factors as p_2 while γ_1 lacks one pro-decision factor of p_1. But $\gamma_2 > \gamma_1$ is inconsistent with $\gamma_1 > \gamma_2$, so deciding *Young* as *Keeble* is not allowed by the extended case base, since it would amount to overruling the second precedent.

In Horty's approach not all deviations from a precedent are overrulings. Suppose that precedent 2 is not in the case base but is a new case, so its decision is not yet known. Then $p_2 > p_1$ is consistent with $k_1 > k_2$, so deciding the new case differently than in the precedent is allowed by the case base, since it leaves all decisions in precedents unaffected. Horty here says that deciding con the original decision in the new case *distinguishes* the precedent. Horty has thus given precise logical formalizations of the important common-law notions of following, distinguishing, and overruling a precedent.

These approaches have been extended by allowing for chains of reasoning from more concrete to more abstract factors, so that detailed reasoning about how the factors relate to a conclusion can be modeled (cf CATO's factor hierarchy). For example, Prakken and Sartor (1998), using an idea of Loui and Norman (1995), represent cases in general as multistep arguments for and against conclusions, where rule priorities are expressed for each pair of conflicting rules, and where any rule or priority of a precedent can be used in a new case. This makes it possible to argue in various ways that in a new case some of the arguments pro the decision cannot be constructed or new arguments con the decision can be constructed, so that the new case must have a different outcome (cf Loui and Norman, 1995).

The work on chains of reasoning and dialectical structures does not explain why the most abstract factors are factors pro or con a decision. Berman and Hafner (1993) argued that often a factor can be said to favor a decision by virtue of the purposes served or values promoted by taking that decision because of the factor. A choice in case of conflicting factors is then explained in terms of a preference ordering on the purposes, or values, promoted or demoted by the decisions suggested by the factors. Cases can then be compared in terms of the values at stake rather than on the factors they contain.

Bench-Capon (2002) computationally modeled this approach and illustrated it with the series of cases from American property law on ownership of wild animals that are being chased. For example, plaintiff in *Keeble* could argue that people should be protected when pursuing their livelihood, since society benefits from their activities. He could also argue that he was hunting on his own land, so that the value of protection of property is another reason why he should win. Defendant in *Keeble* could argue that since plaintiff had not yet caught the ducks, he had no right to the ducks, since if such rights depended on who first saw the animals, there would be no clear criterion and the courts would be flooded with cases. Thus defendant argues that deciding for him promotes the value of legal certainty. Since plaintiff won in *Keeble*, we can, on this interpretation of the case, say that the court found that the combination of the values of protecting property and protecting the pursuit of livelihood outweighs the single value of legal certainty. This value preference can be cited in new cases where the same values are at stake (but note that in *Young* the value of protecting the pursuit of livelihood is cancelled out by the fact that in *Young* the defendant was also pursuing his livelihood).

A limitation of this approach is that often the extent to which a value is promoted or demoted is also important. Even if a new case contains the same values as a precedent, the degrees of promotion or demotion may be different, which can give rise to arguments against the citation. This could explain why in legal practice almost no generally valid hierarchies of values are accepted: the point is that different degrees of promotion or demotion give rise to different value preferences in different cases.

Bench-Capon and Sartor (2003) embedded Bench-Capon's (2002) approach in a method for constructing theories that explain a given set of cases. They modeled theory construction as an adversarial process, where both sides take turns to modify the theory so that it explains the current case in the way they want. The process starts with a set of statements about which factors promote which values and a set of cases represented in terms of factors and an outcome. Then the theory is constructed by creating rules plus rule preferences derived from value preferences. Chorley and Bench-Capon (2005) study various ways to automate this process.

The attention paid to the role of value and purpose led to accounts of legal interpretation as a decision problem, namely, as a choice between alternative interpretations on the basis of the likely consequences of these interpretations in terms of promoting and demoting values. Thus current AI models of argumentation for decision making can be applied. This approach in fact regards legal reasoning as a form of what philosophers call practical reasoning. It compares arguments about what decision to take in terms of the various positive and negative effects of the decision alternatives on goals and values, in terms of the relative preferences between these goals or values, and sometimes also in terms of the extent to which they are promoted or demoted. This line of research has especially been pursued by Bench-Capon and Atkinson and colleagues, for example, in Atkinson et al. (2005), applying it both to legal interpretation and to legislative and policy debates.

A practical-reasoning account of legal interpretation can be applied to model the problem of a court deciding whether to follow or distinguish a precedent when a new factor arises (assuming that distinguishing is consistent with the body of precedents). For example, deciding *Young* for the defendant can be reconstructed as follows. First practical-reasoning arguments for following or distinguishing *Keeble are stated.* Then the argument for distinguishing *Keeble* is preferred over the argument for following *Keeble*, on the ground that following *Keeble* only promotes the value of protecting livelihood while distinguishing *Keeble* in addition promotes the value of legal certainty.

The work reviewed in this section does not amount to a full theory of interpretative legal argument. One thing hardly addressed in the above models is the step from facts to factors. Further study is needed of how legal reasoning actually bridges the gap between the concrete facts and the more abstract factors in a case.

Establishing the Facts of a Case

While legal education and scholarship mostly focuses on reasoning with and about the law, in practice most cases are decided on the facts, so insight as to how facts can be proven is crucial for legal practice. AI and law research has addressed two main questions: which model of rational proof can best be applied to the law, and what is the logical nature of important legal evidential constructs like burdens of proof and presumptions? We address these issues in turn.

Models of Legal Proof

Theoretical models of rational legal proof are generally of three kinds: probabilistic, story-based, or argument-based. All three approaches acknowledge that evidence cannot provide watertight support for a factual claim but always leaves room for doubt and uncertainty, but they account for this in different ways. Probabilistic approaches express uncertainty in terms of numerical probabilities attached to hypotheses given the evidence. Often a Bayesian approach is taken, nowadays more and more with Bayesian networks (Fenton and Neil, 2011). Probabilistic approaches are by no means uncontroversial. One objection is that in legal cases the required numbers are usually not available, either because there are no reliable statistics, or because experts or judges are unable or reluctant to provide estimates of probabilities. Another objection is that probability theory imposes a standard of rationality that cannot be attained in practice, so that its application would lead to more instead of fewer errors. To overcome these and other claimed limitations of probabilistic models, argumentation-based and story-based models have been proposed.

Story-based approaches go back to the work of the psychologists Bennett and Feldman (1981), who observed that the way judges and prosecutors tend to make factual judgments is not by probabilistic or logical reasoning but by constructing and comparing alternative plausible stories about what might have happened. In these approaches, the story that best explains the evidence must, if it does so to a sufficient degree, be adopted as true. Room for doubt is provided since some as yet unknown story might be the true one or since new evidence might make another of the known stories the best one. Thagard (2004) sketches how this approach can be computationally modeled as inference to the best explanation.

Both Bayesian and story-based approaches reason from hypotheses to the evidence in that, to assess alternative hypotheses, they model how likely the evidence is under the various hypotheses. In contrast, argumentation-based approaches reason from the evidence to the hypothesis, by stepwise building evidential arguments from the available evidence to the hypotheses. Room for doubt is provided since additional evidence might give rise to new arguments that defeat the currently undefeated arguments. These approaches go back to Wigmore's (1931) charting method of legal evidence, with which alternative arguments from evidence to hypotheses can be graphically displayed and thus sources of doubt in these arguments can be revealed. An important source of doubt is the empirical generalizations needed to justify the various steps in an evidential argument, such as that witnesses usually speak the truth. Generally, three sources of doubt in witness testimonies are considered: the witness has reason to lie, the witness has flawed memory, or the witness could not reliably observe, hear, feel, or smell

what he testified to. Bayesian and story-based approaches account for this by considering four hypotheses: the witness statement is correct, the witness lied, the witness misremembered, or the witness observed incorrectly. They would then estimate how likely the fact that the witness made his or her statement is under these alternative hypotheses. By contrast, argumentation approaches would defeasibly infer from the witness statement that the accused did what the witness testified to, and they would account for sources of doubt as giving rise to counterarguments.

Bex et al. (2003) model the Wigmorean approach in terms of logical frameworks for argumentation. Evidential reasoning proceeds by applying various argument schemes for evidential reasoning to the evidence, such as schemes for perception, memory, induction, applying generalizations, reasoning with testimonies, and temporal persistence. The so-called critical questions of these schemes specify exceptions to these reasoning forms.

Argumentation approaches are good for modeling how evidence can be related to hypotheses and for revealing sources of doubt in evidential arguments, but less good for providing a clear overview of masses of evidence. This is better done in the story-based approach, which formulates scenarios with a central action made plausible by the context. To combine the strengths of both approaches, Bex (2011) proposed a hybrid theory of legal evidential reasoning. The idea is that stories or scenarios are connected events while evidence for these events and support for their connections is provided by argumentation with argument schemes. Then Bex formulated various criteria for the internal plausibility and coherence of a story and for how well it is supported by the evidence.

Burdens of Proof and Presumptions

Legal proof is not a scientific problem but takes place under legal and practical constraints. The parties involved in a proceeding have limited technical and financial resources, while a decision has to be reached within reasonable time and in a fair way. Legal systems have developed ways to deal with these constraints, such as presumptions and burdens of proof. Research has been done on modeling these notions with techniques from nonmonotonic logic and argumentation (Prakken and Sartor, 2009; Gordon and Walton, 2009). Briefly, the idea is that a burden of proof for a claim is fulfilled if at the end of a proceeding the claim is acceptable according to the argumentation logic applied to the then available evidence. Some care has to be taken here to model that in the law burdens of proof for different sub-issues (including both burdens of production and burdens of persuasion) can be distributed over the parties. Another challenge is to account for the fact that different kinds of legal issues can have different standards of proof. For example, in common-law jurisdictions claims must in criminal cases be proven 'beyond reasonable doubt' while in civil cases usually proof 'on the balance of probabilities' suffices. In logical accounts defeat between arguments is an all-or-nothing matter, but proponents of probabilistic approaches might say that this fails to respect that defeat between evidential arguments is a matter of degree. Gordon and Walton (2009) and Prakken and Sartor (2011) study ways to reconcile the gradual nature of proof standards with the all-or-nothing nature of defeat between arguments in argumentation logics.

Legal presumptions obligate a fact finder to draw a particular inference from a proved fact. Typical examples are a presumption that the one who possesses an object in good faith is the owner of the object, or a presumption that when a pedestrian or cyclist is injured in a collision with a car, the accident was the driver's fault. Some presumptions are rebuttable while others are irrebuttable. Prakken and Sartor (2009) argue that rebuttable presumptions can be logically interpreted as defeasible rules, so that reasoning with rebuttable presumptions can be modeled with a nonmonotonic logic.

Legal Reasoning and Legal Procedure

Legal reasoning usually takes place in the context of a dispute between adversaries, within a prescribed legal procedure. This makes the setting inherently dynamic and multiparty. The facts and theories are not given at the start of a case, but the adversaries advance their points of view and provide their evidence at various stages, and they accept, reject, or challenge their opponent's claims when these are made. The adjudicator can allocate burdens of proof and rule about admissibility of evidence or arguments during a proceeding, before he or she decides the dispute in the end. All these things mean that the quality of a legal decision not only depends on its grounds but also on how it was reached. It is therefore relevant to study how legal procedures contribute to the quality of a decision. In short, legal procedures should promote rational, fair, and effective dispute resolution.

AI and law research has studied these issues by embedding argumentation models like the ones discussed above in dialogue models of legal procedures. The procedures modeled are not actual but artificial ones, often simplified compared to actual legal procedures. The point of this is partly analytical: by formalizing and studying simple artificial legal procedures, insight can be obtained as to the properties of actual legal procedures. But this also has a normative point. Desirable properties for legal procedures can be precisely formulated, that is, properties that make procedures fair and/or effective, and whether these procedures have these properties can be formally or computationally investigated. For example, a fairness property is that if the participants' joint knowledge bases logically (according to an argumentation logic) imply a certain outcome, then the procedure allows a dialogue in which this outcome is reached. Insights thus obtained about artificial procedures can then be used in (re-)designing actual legal procedures.

An influential computational model of a legal procedure was Gordon's (1994) Pleadings Game, which formalized a normative model of pleading, founded on first principles. It was meant to identify the issues to be decided at trial, given what the parties had claimed, conceded, challenged, and denied in the pleadings phase and what (defeasibly) follows from it. Other games (usually two-party) define the outcome in terms of whether the adversaries in the end agree on the main issue, while Prakken (2008) defines a three-party game where in the end an adjudicator (e.g., a judge or a jury) decides the dispute.

The dynamic and multiparty setting of legal procedures also raises issues of strategy and choice. One of the few

investigations of these issues in a legal context is by Riveret et al. (2008), who apply a combination of game theory and argumentation logic to the problem of determining optimal strategies in debates with an adjudicator. In such debates, the opposing parties must estimate how likely it is that the premises of their arguments will be accepted by the adjudicator. Moreover, they may have preferences over the outcome of a debate, so that optimal strategies are determined by two factors: the probability of acceptance of their arguments' premises by the adjudicator and the costs/benefits of such arguments.

Conclusion

Computational models of legal reasoning aim to respect that the central notion in legal reasoning is not deduction but argument. Deduction has its place in legal reasoning, but only as part of a larger model of constructing, attacking, and comparing arguments. Moreover, computational models of legal reasoning aim to respect that rules are not the only source of legal knowledge: the roles of cases, principles, purposes, and values should not be ignored, and these models stress the importance of dynamics, procedure, and multiparty interaction. While most work has addressed legal interpretation and normative determinations, some work is addressing legal proof. Here too, deduction is just one of the tools, as part of either probabilistic, argumentation-based, or narrative approaches. A main concern here is developing models of legal proof that are rationally well founded but respect the practical and cognitive constraints faced by the participants in court proceedings.

See also: Bayesian Theory, History, Applications, and Contemporary Directions; Decision and Choice: Bounded Rationality; Deductive Reasoning Systems; Knowledge Representation; Legal Reasoning and Argumentation; Logic Models; Problem Solving and Reasoning: Case-based.

Bibliography

Aleven, V., 2003. Using background knowledge in case-based legal reasoning: a computational model and an intelligent learning environment. Artificial Intelligence 150, 183–237.

Ashley, K., 1989. Toward a computational theory of arguing with precedents: accomodating multiple interpretations of cases. In: Proceedings of the Second International Conference on Artificial Intelligence and Law. ACM Press, New York, pp. 39–102.

Ashley, K., 1990. Modeling Legal Argument: Reasoning with Cases and Hypotheticals. MIT Press, Cambridge, MA.

Atkinson, K., Bench-Capon, T., McBurney, P., 2005. Arguing about cases as practical reasoning. In: Proceedings of the Tenth International Conference on Artificial Intelligence and Law. ACM Press, New York, pp. 35–44.

Bench-Capon, T., 2002. The missing link revisited: the role of teleology in representing legal argument. Artificial Intelligence and Law 10, 79–94.

Bench-Capon, T., Coenen, F., 1992. Isomorphism and legal knowledge based systems. Artificial Intelligence and Law 1, 65–86.

Bench-Capon, T., Sartor, G., 2003. A model of legal reasoning with cases incorporating theories and values. Artificial Intelligence 150, 97–143.

Bennett, W., Feldman, M., 1981. Reconstructing Reality in the Courtroom: Justice and Judgment in American Culture. Methuen-Tavistock, London.

Berman, D., Hafner, C., 1993. Representing teleological structure in case-based legal reasoning: the missing link. In: Proceedings of the Fourth International Conference on Artificial Intelligence and Law. ACM Press, New York, pp. 50–59.

Bex, F., 2011. Arguments, Stories and Criminal Evidence: A Formal Hybrid Theory. Springer, Dordrecht.

Bex, F., Prakken, H., Reed, C., Walton, D., 2003. Towards a formal account of reasoning about evidence: argumentation schemes and generalisations. Artificial Intelligence and Law 12, 125–165.

Branting, L., 1999. Reasoning with Rules and Precedents: A Computational Model of Legal Analysis. Kluwer Academic Publishers, Dordrecht/Boston/London.

Chorley, A., Bench-Capon, T., 2005. An empirical investigation of reasoning with legal cases through theory construction and application. Artificial Intelligence and Law 13, 323–371.

Fenton, N., Neil, M., 2011. Avoiding legal fallacies in practice using Bayesian networks. Australian Journal of Legal Philosophy 36, 114–151.

Gardner, A., 1987. An Artificial Intelligence Approach to Legal Reasoning. MIT Press, Cambridge, MA.

Gordon, T., 1991. An abductive theory of legal issues. International Journal of Man-Machine Studies 35, 95–118.

Gordon, T., 1994. The Pleadings Game: an exercise in computational dialectics. Artificial Intelligence and Law 2, 239–292.

Gordon, T., Walton, D., 2009. Proof burdens and standards. In: Rahwan, I., Simari, G. (Eds.), Argumentation in Artificial Intelligence. Springer, Berlin, pp. 239–258.

Horty, J., 2011. Rules and reasons in the theory of precedent. Legal Theory 17, 1–33.

Loui, R., Norman, J., 1995. Rationales and argument moves. Artificial Intelligence and Law 3, 159–189.

McCarty, L., 1977. Reflections on TAXMAN: an experiment in artificial intelligence and legal reasoning. Harvard Law Review 90, 89–116.

McCarty, L., 1995. An implementation of Eisner v. Macomber. In: Proceedings of the Fifth International Conference on Artificial Intelligence and Law. ACM Press, New York, pp. 276–286.

Prakken, H., 2008. A formal model of adjudication dialogues. Artificial Intelligence and Law 16, 305–328.

Prakken, H., Sartor, G., 1996. A dialectical model of assessing conflicting arguments in legal reasoning. Artificial Intelligence and Law 4, 331–368.

Prakken, H., Sartor, G., 1998. Modelling reasoning with precedents in a formal dialogue game. Artificial Intelligence and Law 6, 231–287.

Prakken, H., Sartor, G., 2002. The role of logic in computational models of legal argument: a critical survey. In: Kakas, A., Sadri, F. (Eds.), Computational Logic: Logic Programming and beyond. Essays in Honour of Robert A. Kowalski, Part II. Springer Lecture Notes in Computer Science 2048, Berlin, pp. 342–380.

Prakken, H., Sartor, G., 2009. A logical analysis of burdens of proof. In: Kaptein, H., Prakken, H., Verheij, B. (Eds.), Legal Evidence and Proof: Statistics, Stories, Logic. Ashgate Publishing, Farnham, pp. 223–253.

Prakken, H., Sartor, G., 2011. On modelling burdens and standards of proof in structured argumentation. In: Atkinson, K. (Ed.), Legal Knowledge and Information Systems. JURIX 2011: The Twenty-Fourth Annual Conference. IOS Press, Amsterdam, pp. 83–92.

Prakken, H., Sartor, G. Law and Logic: A Review, submitted.

Riveret, R., Prakken, H., Rotolo, A., Sartor, G., 2008. Heuristics in argumentation: a game-theoretical investigation. In: Besnard, P., Doutre, S., Hunter, A. (Eds.), Computational Models of Argument. Proceedings of COMMA 2008. IOS Press, Amsterdam, pp. 324–335.

Rotolo, A., Governatori, G., 2009. Changing legal systems: legal abrogations and annulments in defeasible logic. Logic Journal of IGPL 18, 157–194.

Sergot, M., 2001. A computational theory of normative positions. ACM Transactions on Computational Logic 2, 581–622.

Sergot, M., Sadri, F., Kowalski, R., Kriwaczek, F., Hammond, P., Cory, H., 1986. The British Nationality Act as a logic program. Communications of the ACM 29, 370–386.

Skalak, D., Rissland, E., 1992. Arguments and cases. an inevitable intertwining. Artificial Intelligence and Law 1, 3–44.

Svensson, J., 2002. The use of legal expert systems in administrative decision making. In: Grönlund, A. (Ed.), Electronic Government: Design, Applications and Management. Idea Group Publishing, London, pp. 151–169.

Thagard, P., 2004. Causal inference in legal decision making: Explanatory coherence vs. Bayesian networks. Applied Artificial Intelligence 18, 231–249.

Verheij, B., Hage, J., van der Herik, H., 1998. An integrated view on rules and principles. Artificial Intelligence and Law 6, 3–26.

Wigmore, J., 1931. The Principles of Judicial Proof, second ed. Little, Brown and Company, Boston.

Legal Scholarship

Edward Rubin, Vanderbilt University, Nashville, TN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The author sketches the history of legal scholarship and the relation of law to other academic disciplines. First, jurisprudence and doctrinal studies between 1900 and 1970 are reviewed, beginning with legal formalism; moving to its principal critiques, legal realism and legal positivism; and concluding with the rehabilitation of law as a coherent body of doctrine. Between 1970 and 1985, further attacks on the concept of law came from the Law and Economics Movement and from the Critical Legal Studies Movement. Developments in the late twentieth and early twenty-first centuries include post-Chicago School economics, 'outsider scholarship' (e.g., feminist and critical race theory), a revised form of 'standard' legal scholarship, and the growth of sociolegal and empirical legal studies.

Legal scholarship can be distinguished from scholarship about law, such as the history of law or the sociology of law, by virtue of its internal perspective (Dworkin, 1986; Hart, 1961). This means that it treats law as a set of norms that can be analyzed according to their internal coherence and moral significance, rather than as one component of a social system whose contours and content are determined by the same forces that affect other aspects of the society. Within the boundaries of this internal approach, there has traditionally been a division between jurisprudence, which concerns the theory of law, and doctrinal studies, which concern specific legal provisions. In recent years, however, both the distinction between internal and external perspectives and that between jurisprudence and doctrinal studies have been significantly eroded by intellectual developments (Abel, 1973) such as law and economics, critical legal studies, critical race theory, feminist legal theory, gay legal studies, sociolegal studies, and empirical legal studies. These developments have deployed the methodologies of other disciplines in the analysis of the internal structure and moral significance of law. At the same time, they have made the study of particular legal doctrines more theoretical and rendered legal theory less unified and more subject specific. While the distinctions between legal scholarship and scholarship about law, and between jurisprudence and doctrinal studies, remain useful, they now refer to bodies of scholarship with generalized family resemblances, rather than distinct methodologies with clearly defined boundaries.

Jurisprudence and Doctrinal Studies between 1900 and 1970

The Attack on Formalism

In the early years of the twentieth century, doctrinal studies in both the Continental and the Anglo-American worlds were dominated by an approach that can be described as legal formalism, or conceptualism (*see* Legal Formalism). Its essential premise was that the law is an autonomous system, based on general principles, that is accessible to reason and that effectively guides judges when they decided contested cases (Schauer, 1988). Formalism did not necessarily advance the natural law idea that this system represented transcendental truth (*see* Natural Law), nor the naïve idea that judicial decisions could be derived from general principles by mechanistic or syllogistic logic (Tamanaha, 2010), but it did claim that law was animated by a generalized conceptual structure that transcended particular political or policy positions. In Continental Europe, the source of this autonomous system was regarded as each nation's legal code; in Great Britain and the United States, it was the common law. By virtue of the claim that law is autonomous, and the rather close focus on judicial decisions, both versions of formalism were somewhat insulated from major jurisprudential movements such as Marxism and sociological jurisprudence.

The attack on formalism came from two sources. The first, generally described as legal realism, asserted that judicial decisions and legal concepts are indeterminate or contextual, that they are the product of social attitudes rather than reasoned analysis, and that they are subordinated to the political process rather than transcending it (*see* Legal Realism). American legal realism emphasized the incoherence of general legal principles, the indeterminacy of judicial decisions, and the dependence of these decisions on the particular views of the judge. Continental realism, which was centered in Scandinavia, emphasized the contextual nature of legal concepts and their dependence upon more general social attitudes. A related line of attack was sociological jurisprudence, which attempted to develop an alternative conception of law as determined by, and contributing to, more general social forces (Ehrlich, 1975). This was connected with the work of anthropologists and sociologists who treated law as an aspect of social systems, particularly those without formal legal institutions. A number of leading American realists saw the sociological approach as an alternative to doctrine, and a few tried to study legal behavior as a sociological phenomenon (Llewellyn and Hoebel, 1941; see Schlegel, 1995).

Apart from the legal realist and sociological approach, the dominant theme in early twentieth-century jurisprudence was legal positivism (*see* Legal Positivism). According to Kelsen, law is the set of threats promulgated by public officials whose authority is established by a (fictitious) single, agreed-upon principle, or *Grundnorm* (Kelsen, 1967) As such, it is completely separate from morality; a particular normative statement is a law if, and only if, it has been duly promulgated by

a recognized public official, or by a private party who has been delegated to act by such a recognized official. Legal positivism was motivated partially by larger philosophic trends (such as logical positivism) and partially by the effort to refute natural law jurisprudence. While its connection with doctrinal studies is not self-evident, it seems consistent with American realists' rejection of general common law, that is, the idea that legal rules can be derived from general principles and need not correspond to the decisional law of any specific jurisdiction, and with the Scandinavian realists' insistence that there are no overarching legal concepts, but only concepts specific to particular societies.

The Rehabilitation of Law

The years following World War II saw renewed interest in the concept of law in both jurisprudence and doctrinal studies. In jurisprudence, scholars began searching for a way to reinfuse some sense of social morality into the concept of law, so that one would not be required to apply that term to the rules promulgated by a generally immoral regime such as Nazi Germany. This led to a revival of natural law (Finnis, 1980), but the appeal of this work was limited by its continuing religious overtones (*see* Natural Law). More influential was Fuller's effort to articulate a purely procedural version of natural law (Fuller, 1964), which asserted that any declarations that one would be willing to regard as law must have certain universal characteristics such as generality, constancy, and nonretroactivity. Still more influential was Hart's recharacterization of legal positivism in light of Wittgenstein's linguistic theory (Hart, 1961). Law, in Hart's view, is a complex set of social rules, consisting of 'primary' rules that not only impose sanctions but also create a sense of obligation among citizens, and 'secondary rules' that state the manner in which public officials impose the primary rules, develop new rules, and empower action by private parties that will be enforced by public authorities.

Doctrinal scholarship rejected legal realism more definitively than jurisprudence rejected legal positivism. In the United States, the legal process school conceded that the legal system was subordinate to politics, thus acknowledging positivism's essential claim, but insisted that the concept of law retained an important role within this more limited sphere of action. Fuller argued that reasoned decision making according to rules established in advance is the best way to adjudicate disputes between parties, although a poor way to formulate new social policy (Fuller, 1978). Other members of this school explored the rationales by which the courts could use the US Constitution to evaluate, and, if necessary, strike down legislation, without substituting its political judgment for that of the legislature (Bickel, 1962; Wechsler, 1959). Central to this inquiry was the belief that doctrine possessed an inherent logic or coherence, so that judges could reach generally accepted conclusions by means of reasoned argument. An important corollary was that public officials, particularly judges, generally acted in the public interest, and that their decisions harmed social welfare only when they strayed beyond the limits of their institutional competence.

In Europe, theories of judicial decision making focused on statutory interpretation, which generally (although not in Britain) meant the interpretation of a comprehensive civil or criminal code. Many scholars asserted, or assumed, that judges could reach generally accepted conclusions by means of reasoned argument if they followed and extrapolated the inherent logic of the code (MacCormick and Summers, 1991). As in the case of the American legal process school, this work was premised on the idea that doctrine was coherent within a politically defined area of proper operation. Because it concerned interpretation of a legal code, rather than common law or a vaguely worded constitution, it was more closely allied to legal positivism than its American equivalent. Both the American and European versions of this postwar approach were sufficiently focused upon law's internal rules to minimize the influence of social science studies of law.

The Critique of Law, 1970–85

Law and Economics

The attack on this rehabilitated concept of law was initiated primarily in the United States, and came from two sources that were associated with opposite ends of the American political spectrum. The first was the Law and Economics Movement, which was derived from the Chicago School of microeconomics and shared its basic premise that all social action is a rational effort to maximize the actor's material self-interest (*see* Law and economics). According to this approach, legislators are motivated solely by their desire to be reelected (assuming that effective sanctions exist against outright bribery); thus, the laws they produce have no inherent logic, and no necessary relation to the public interest. While these laws might correspond to the preferences of voting citizens, they are more likely to reflect the distorting influence of special interest groups that can organize and raise funds more effectively than the general public (Olson, 1965). A judge who is interpreting such laws cannot be guided by their internal logic; all he or she can do is enforce the bargain that the interest groups reached, or interpret the law narrowly in deference to public values (Easterbrook, 1983; Macey, 1986). Posner originally asserted that the common law, as opposed to statutory law, represented a coherent body of doctrine that implemented the social policy of efficiency (Posner, 1972). Other law and economics writers soon abandoned this implausible contention and devoted their efforts to analyzing the self-interested motivations that lead private parties to engage in or to avoid litigation, and that lead lawyers to encourage or discourage them from doing so, to settle cases, or to adopt specific trial strategies (Priest and Klein, 1984). This work suggests that common law, like statutory law, lacks any conceptual coherence, and is merely the product of economic forces.

In addition to its analysis of doctrine, the Law and Economics Movement advances the jurisprudential position that the only proper purpose of law, or of government in general, is to achieve economic efficiency, that is, to maximize the nation's wealth (Posner, 1972). When combined with the economic analysis of law, this leads to the conclusion that regulatory law is generally undesirable and that the social control of economic relations should be left to the market. The argument seems more political than legal, in that it addresses the purpose of law rather than its nature, and it is clearly related to the conservative politics of most of the early

law and economics scholars. But it also represents a theory of law, in that it advances a general rule that identifies improper laws and improper roles for legal actors.

Critical Legal Studies

The Critical Legal Studies Movement was derived, albeit rather loosely, from the critical theory of the Frankfurt School. It revived the realist claim that law, both statutory and decisional, is not an internally coherent system but a collection of unrelated rules. Unlike legal realism, however, critical legal studies asserted that the reason for this lack of coherence is that law is a device by which the political elite maintains its hegemony, or social control (Horwitz, 1977; Unger, 1975). It is a particularly effective device, moreover, because the widely accepted claim that law is coherent or logical masks its class-oriented character; in effect, it creates a false consciousness among the disadvantaged that leads them to accept their subordinated status (*see* Critical Legal Studies). With respect to doctrine, critical legal studies scholars sought to disprove the claim that legal decisions are logical or coherent, and thereby to reveal the hegemonic purposes of these decisions. In undertaking this analysis, they went beyond the legal realists' demonstration of the internal inconsistencies of doctrine, invoking various aspects of Continental thought such as deconstruction and postmodernism to demonstrate that any claim to linguistic coherence is essentially fraudulent.

This analysis of doctrine is linked to the jurisprudential position that the social dominance of the elite, and its use of law to secure its position, is profoundly immoral. A just society achieves equality among its members – if not complete equality, then certainly a much higher degree of equality than contemporary society displays. Thus, modern society needs to be emancipated from its existing social structures and from ideologies, including law, which maintain those structures and preach their legitimacy to the disadvantaged. Again, this is essentially a political argument, with an obvious connection to the progressive or radical politics of critical legal studies scholars. But again, it constitutes a general theory of law by distinguishing between the improper, hegemonic purposes of law and its proper, emancipatory purposes.

Despite their obvious differences in political orientation, law and economics and critical legal studies rest on closely related insights and exercise similar impacts on the character of legal scholarship. Both movements renew the realist attack on law's doctrinal coherence, and both go beyond realism by grounding this attack on a political analysis. In both cases, moreover, that analysis asserts that law is produced by political and economic elites for the purposes of securing their positions. By thus attributing the incoherence of doctrine to political causes, both law and economics and critical legal studies join doctrinal analysis to larger jurisprudential questions in a manner that was not true for legal formalism, legal realism, or the legal process school. The two movements are further allied in their heavy dependence upon nonlegal methodologies. For the first time, two large groups of legal scholars were prepared to abandon the idea that law is an autonomous or quasi-autonomous system that provides its own analytic tools, and look to other disciplines for the main elements of their analytic framework (Rubin, 1988).

The Topography of Legal Scholarship in the Early Twenty-First Century

Law and Economics

Law and economics remains a robust tradition within legal scholarship, and much work continues to be done according to the Chicago School model that assumes rational, self-interested actors and smoothly functioning markets. But the criticisms of this approach by legal scholars, combined with the internal evolution of the field, have led to an increasing body of post-Chicago Law and Economics scholarship. One theme of this work, reflecting developments in economics itself (Williamson, 1985), is that institutional arrangements exercise powerful effects in structuring behavior. Since legal actors generally function in an institutional context, this has proved to be a productive line of inquiry. Positive political theory and game theory, which treat institutions themselves as rational actors, have also been utilized in current legal scholarship to take institutional factors into account (Cooter, 2000; Georgetown Law Journal, 1992; Rodriguez and Weingast, 2003). A second, and related, theme is that internalized norms exercise powerful effects that conflict with the self-interest, or at least the immediate self-interest, of the actor. Some scholars argue that legal rules can be regarded as internalized norms that reflect varying degrees of long-term rational self-interest (University of Pennsylvania Law Review, 1996); others assert that particular communities, such as merchants in a given trade (Bernstein, 1996; Cooter, 1996), or farmers in a given area (Ellickson, 1991), engage in heavily norm-determined behavior that strongly affects the impact of official law on these communities. The combined effect of focusing on institutions and on norms has been to modify, without eliminating, the external perspective that law and economics generally adopts. Law is not being treated as presumptively coherent, but its internal structure is being given a role that followers of the Chicago School tend to deny.

Still another influence on law and economics has been the attack on human rationality by cognitive psychologists (Kahneman et al., 1982). This work, which is heavily empirical, has revealed that people, at least in laboratory settings, are irrationally optimistic, irrationally attached to the status quo, and irrationally influenced by the way that choices are presented. Legal scholars have applied these results to a variety of settings, such as jury decisions, contracting behavior, and policy making, and have also begun conducting their own empirical research to study behavior that is relevant to specific legal situations (Gunningham et al., 2003; Jolls et al., 1998; Hans, 2000; O'Connell, 2008; Vanderbilt Law Review, 1998). Thus, the effort to develop an account of human behavior that is more realistic and more modulated than the rational actor model of law and economics has led to the extensive incorporation of a second discipline, psychology, into legal scholarship.

Outsider Scholarship

Critical legal studies has been succeeded by a series of movements that tend to refer to themselves as Outsider Scholarship.

They include critical race theory, feminist theory, and gay legal studies (*see* Critical Race Theory; Feminist Legal Theory; Sexual Orientation and the Law). While critical legal studies remains a viable approach that continues to generate a certain amount of legal scholarship, its role within the academy has been largely replaced by these latter movements, which share its critical approach but contest its particular conclusions. Outsider Scholarship, like critical legal studies, asserts that law is an instrumentality of hierarchy and social domination, but it rejects economic class analysis as an explanation for this phenomenon. Instead, it argues that more specific prejudices are involved, namely, the prejudice of the racial majority against minorities like African-Americans, Latinos, and Native Americans, the prejudice of men against women, and the prejudice of heterosexuals against gays and lesbians (Bell, 1987; MacKinnon, 1989; Eskridge, 1994). These biases may be explicit or implicit (Greenwald and Krieger, 2006; Levinson and Smith, 2012).

One result of this approach is to create a critical scholarship with a more direct connection to American politics; while class has never been a salient political issue in the United States, race, gender, and sexual orientation are all subjects of active political movements. Another, and not unrelated result, has been to decrease the breadth of the critical attack on law. Critical legal studies tended to regard law in its entirety as an oppressive device that hides its true effect behind a facade of objectivity. Outsider scholarship, because it concerns social phenomena that are less comprehensive than class, views law as oppressive because it has denied outsiders the benefits it grants to the dominant group. Its prescription, therefore, is not to abandon or even transform law, but to provide law's benefits to the outsiders, fulfilling the promise of equality that it has thus far denied.

The effect of this outsider scholarship is thus to modify the external perspective of its predecessor. Law is still viewed with suspicion, but its role as a system with internal values is affirmed. For example, instead of arguing that rights are illusory, or an impediment to genuine equality because they can only be validated through litigation, outsider scholarship tends to argue that the disadvantaged cannot afford to wait for such comprehensive transformations, but should use the concept of rights to achieve a more limited but more accessible equality within the existing system (Williams, 1991).

Outsider scholarship shares the critical legal studies view that legal concepts are social constructions that serve specific purposes, often those of dominant elites, and that these concepts are radically indeterminate. But it has not relied as heavily on the indeterminacy argument. As many scholars have pointed out, the argument is itself too indeterminate, since it lacks the criteria to judge how determinate the law could be, and how much determinacy is desirable (Kress, 1989). In addition, most legal scholars recognize that the indeterminacy argument does not necessarily lead to the political conclusions that critical legal studies advanced. Thus, outsider scholarship analyzes legal rules in fairly traditional terms. Its methodological innovation has been its extensive use of narrative, or storytelling, as a means of communicating the personal experience of oppression that racial minorities, women, or gays and lesbians have experienced (e.g., Bell, 1987; Williams, 1991). While this has a literary quality (when done well, at least), it is not related to literary criticism or to the methodology of any other discipline. Rather, it is a technique that is intended to appeal to the values of the reader within the framework of legal scholarship and the legal system.

Standard Legal Scholarship (Reformulated)

Throughout the period when law and economics and critical legal studies were articulating their critique of law and of the legal process school, standard legal scholarship continued to be written. At that time, and at present, it continues to constitute the largest proportion of the legal scholarship being produced. It is standard not only because of its quantity, but also because it is not explicitly identified with any particular school of thought or methodology. In fact, this scholarship can be loosely regarded as a continuation of the legal process school. Like legal process, it acknowledges that law is subordinate to politics, but treats legal rules as a body of doctrine that has enough coherence to generate analytic standards. In addition, it shares the legal process view that not only the participants in the legal system, particularly judges, but also legislators and administrators, are public-oriented enough so that scholarly prescriptions for improving their decisions can be addressed to them (e.g., Choper, 1980; Dworkin, 1986; Ely, 1980). The standard scholarship that is currently being produced differs from its predecessors because it has at least partially absorbed the lessons of both law and economics and critical legal studies, although without accepting their critique of law's legitimacy (Rubin, 1996).

Most legal scholars now acknowledge that viewing legal actors as public oriented is problematic. They recognize that these actors are sometimes motivated by their material self-interest and sometimes intent on preserving existing hierarchies and class relations. Thus, they do not see the legal system as a realm of reasoned, public-oriented decision making that transcends politics, as legal formalism did, nor as a realm of reasoned, public-oriented decision making that is circumscribed by politics but insulated from it, as legal process did. Rather, politics both sets the limits of the legal system and intercalates itself into particular legal decisions. But standard scholarship rejects the notion that decision makers are motivated by any uniform value system, whether it is self-interest, prejudice, or public orientation; the prevailing attitude seems to be that some degree of conscientiousness can be assumed, but that the actual motivations must be sorted out in particular situations.

Methodologically, standard scholarship has assimilated the idea that law is not a quasi-autonomous system that provides its own analytic tools. Scholars regularly rely on informal law and economics or public choice insights in analyzing the behavior of public officials, and, to a lesser extent, on the insights of critical theory. More importantly, these scholars regularly incorporate data from other fields, such as political science, sociology, economics, and psychology, in framing recommendations. The criterion for stating that a particular legal decision, whether by a judge, legislator, or administrator, is incorrect, or could be improved, is less likely to be the internal coherence of legal doctrine and more likely to be some judgment about social welfare. Such judgments will frequently involve empirical information generated by another discipline and will almost always acknowledge,

implicitly if not explicitly, that such information is relevant to the conclusion.

Contemporary Sociolegal Studies and Empirical Legal Studies

These developments in law and economics, outsider scholarship, and standard legal scholarship represent a process, extending over the entire course of the twentieth century and into the first decades of the twenty-first century, by which the distinctions between internal and external scholarship, and between jurisprudence and doctrinal studies, have been partially effaced. This process has important implications for the prevailing conception of legal knowledge, implications that can be illustrated by considering a fourth aspect of contemporary scholarship, namely, sociolegal studies, empirical legal studies, and new legal realism. The strength of these approaches, and their importance for the future, is their consistency with modern law and economics, outsider scholarship, and the reformulated version of standard legal scholarship. In fact, empirical legal studies can be used to support and advance any of these approaches, in addition to serving as a methodology of its own.

The sociology of law, as developed by nineteenth-century scholars such as Maine and Savigny, was originally a jurisprudential position, articulated in opposition to the universalist claims of both natural law and the Napoleonic Code. Its basic insight was that law is the product of particular cultures. With the growth of academic sociology and anthropology, this became translated into detailed studies of legal behavior that were not only more specific, but also more external to the legal system that they studied. Rather than focusing on the internal coherence of the system, or the experience of its participants, sociolegal work treated law as a phenomenon to be determined by examining the behavior of relevant actors. Legal scholars typically ignored such studies because they regarded law as a self-contained, meaningful system that generated its own criteria of analysis.

In response to the attack on the entire concept of law by law and economics and critical legal studies, most legal scholars now recognize that law possesses only a provisional coherence. Its value is no longer determined by its internal logic, but by its ability to achieve political and social goals. For most scholars, however, law is not simply a set of unrelated provisions that achieve these goals, but a system with both institutional and conceptual features. To have legal knowledge, one must be familiar with the features of this system; to use this knowledge, one must make predictions, or prescriptions, about the way this system achieves independently determined goals. Thus, data from fields such as economics, sociology, political science, anthropology, and psychology become directly relevant to legal scholarship (Edelman, 2001). Some of these data come from work that is not specifically concerned with law, such as organization theory or cognitive psychology, but a good deal come from modern sociolegal studies, where law is the social scientist's explicit concern. But this work, however useful, does not displace legal scholarship. The legal scholar still has a crucial role in determining how such data will be used by the institutional and conceptual elements of the legal system (Nourse and Shaffer, 2009). In fact, legal scholars have recently begun to reemphasize these conceptual elements, arguing that, within a richly described social context, law possesses a wide range of symbolic and constitutive roles (Habermas, 1996; Lessig, 1995; Luhmann, 1995) (*see* Law, Autopoiesis in; Law as Constitutive). Sociolegal scholars have come to recognize this as an empirical fact, and have thus taken advantage of legal scholarship's greater receptivity to social science data by engaging more fully with the internal features of the legal system. The result is a partial effacement of the distinction between external and internal scholarship, without its complete elimination.

The challenge to the concept of law, the multiplicity of responses, and the greater methodological range of the resulting scholarship has also partially effaced the distinction between jurisprudence and doctrinal studies. Even when they are engaged in the detailed analysis of doctrine, scholars now realize that they must often confront basic questions about the nature and purpose of law, and the nature of legal knowledge. Such questions can be ignored, of course, in the interest of simplicity, but the work will be open to challenge by those who question its basic premises. Again, the distinction has not been entirely eliminated; as a practical matter, many scholars content themselves with an analysis of doctrine within one or another subdiscipline of legal scholarship and let others fight the more theoretical battles on their behalf. But most recognize that the status of their work as legal knowledge depends on theoretical arguments in a more immediate way than was true a century before.

These developments can be illustrated by considering some of the principal topics of modern sociolegal studies. One notable social science approach to judicial decision making in the early postwar era was the judicial attitude studies that treated court decisions as a pure reflection of the judge's ideology, uninfluenced by legal doctrine (Schubert, 1959; Ulmer, 1960). While work of this sort continues to be produced, often by the same authors, more recent social science studies of judicial decisions tend to emphasize the interplay between ideology and doctrine (Geyh, 2011; Shapiro, 1981), a process that enables legal scholars to utilize their findings and extend them into jurisprudential areas (Feeley and Rubin, 1999; Kennedy, 1997; Tamanaha, 1997). This work is sometimes described as New Legal Realism (Farber, 2001; Miles and Sunstein, 2008; Nourse and Shaffer, 2009).

A second distinctive strand during the early postwar period emphasized the discontinuity between the law on the books and the law in action. The seminal work in the field argued that contract law, as reflected in the decided cases that were the exclusive focus of most law school courses, did not determine contractual practices; rather, the relationship between the parties and their underlying cultural understandings were the primary determinants (Macaulay, 1963; MacNeil, 1980). Modern sociolegal studies often recharacterize this issue in terms of the implementation process, which involves an analysis of the interplay between statutory or regulatory language and the behavior of the private parties subject to the statutes or regulations in question (Bardach and Kagan, 1982; Coglianese, 1997; Weisberg, 1992). For example, a branch of public, or regulatory, law called New Public Governance questions the effectiveness of command and control regulation, which is based on the idea that law consists of a series of written commands, and recommends

more flexible approach based on this idea that regulation is a dynamic relationship aimed at producing real-world results (Lobel, 2004; Nourse and Shaffer, 2009; Sabel and Simon, 2004). Once again, such an approach yields jurisprudential implications more readily than its predecessor, since it replaces an inevitable abyss between written and practiced law with a dynamic interchange where the practice suggests a reconsideration of the legal structure (Ayres and Braithwaite, 1992; Kagan, 2000).

As these perspectives have become more common, they have been generalized into a mode of legal scholarship described as empirical legal studies. From an institutional perspective, this approach has flourished as leading law schools have hired entry level faculty who have doctoral degrees in social science fields such as sociology, political science, anthropology, and, of course, economics. The result has been a new generation of legal scholars who are trained – and inclined – to obtain and analyze data in an increasingly sophisticated way, often through the use of statistical techniques (George, 2006; Heise, 2002). Substantively, the ideas of New Legal Realism and law in action have combined to produce an extensive scholarly inquiry into the external, or real world, operation of law. In addition to studying the attitudes of judges and the effects of law on private actors, empirical legal scholars investigate patterns of legislation and regulation by administrative agencies, the behavior of public officials, interactions between political events and legal developments, and the effects of social movements on the legal system (Rubin, 2001; Suchman and Mertz, 2010; Trubek, 1984). This expanded scope of inquiry contributes to the continuing attenuation of the distinction between external and internal scholarship. Moreover, as the scope of empirical inquiry expands, and the general approach begins to influence pedagogy as well as scholarship, judges are gradually becoming more interested in and more capable of using empirical studies to obtain information relevant to their decisions, or to shape their decisions with empirical effects in mind. Thus, modern scholarship contributes to a process in which the legal system itself internalizes an external perspective on law, thereby enabling scholars to continue shrinking the distinction between these two perspectives.

See also: Critical Legal Studies; Empirical Legal Studies; Law and Economics: Empirical Dimensions; Law and Society: Development of the Field; Law, Sociology of; Law: History of Its Relation to the Social Sciences; Legal Education; Legal Formalism; Legal Pluralism; Legal Positivism; Legal Realism.

Bibliography

Abel, R., 1973. Law books and books about law. Stanford Law Review 26, 175.
Ayres, I., Braithwaite, J., 1992. Responsive Regulation. Oxford University Press, New York.
Bardach, E., Kagan, R., 1982. Going by the Book. Temple University Press, Philadelphia, PA.
Bell, D., 1987. And We Are Not Saved: The Elusive Quest for Racial Justice. Basic Books, New York.
Bernstein, L., 1996. Merchant law in a merchant court: rethinking the code's search for immanent business norms. University of Pennsylvania Law Review 144, 1765.
Bickel, A., 1962. The Least Dangerous Branch. Bobbs-Merrill, Indianapolis, IN.
Choper, J., 1980. Judicial Review and the National Political Process. University of Chicago Press, Chicago.
Coglianese, C., 1997. Assessing consensus: the promise and performance of negotiated rulemaking. Duke Law Journal 46, 1255–1349.
Cooter, R., 1996. Decentralized law for a complex economy: the structural approach to adjudicating the new law merchant. University of Pennsylvania Law Review 144, 1643.
Cooter, R., 2000. The Strategic Constitution. Princeton University Press, Princeton, NJ.
Dworkin, R., 1986. Law's Empire. Belknap Press, Cambridge, MA.
Easterbrook, F., 1983. Statute's domains. University of Chicago Law Review 50, 533.
Edelman, L.B., 2001. Legality and the endogeneity of law. In: Kagan, R.A., Krygier, M., Winston, K. (Eds.), Legality and Community: On the Intellectual Legacy of Philip Selznick. Institute of Governmental Studies Press/Rowman & Littlefield, Berkeley, CA.
Ehrlich, E., 1975. The Fundamental Principles of the Sociology of Law. Russell & Russell, New York.
Ellickson, R., 1991. Order Without Law. Harvard University Press, Cambridge, MA.
Ely, J., 1980. Democracy and Distrust. Harvard University Press, Cambridge, MA.
Eskridge, W., 1994. Gay legal narratives. Stanford Law Review 46, 607.
Farber, D., 2001. Toward a new legal realism. University of Chicago Law Review 68, 279.
Feeley, M., Rubin, E., 1999. Judicial Policy Making and the Modern State: How the Courts Reformed America's Prisons. Cambridge University Press, New York.
Finnis, J., 1980. Natural Law and Natural Rights. Clarendon Press, Oxford, UK.
Fuller, L., 1964. The Morality of Law. Yale University Press, New Haven, CT.
Fuller, L., 1978. Forms and limits of adjudication. Harvard Law Review 92, 353.
George, Tracey, 2006. An empirical study of empirical legal scholarship: the top law schools. Indiana Law Journal 81, 141.
Geyh, C., 2011. What's Law Got to Do With It? What Judges Do, Why They Do It, and What's at Stake. Stanford University Press, Stanford CA.
Georgetown Law Journal, 1992. Symposium: positive political theory and public law. Georgetown Law Journal 80, 457.
Greenwald, Anthony G., Krieger, Linda Hamilton, 2006. Implicit bias: scientific foundations. California Law Review 94, 945.
Gunningham, Neil A., Kagan, Robert A., Thompson, Dorothy, 2003. Shades of Green: Business, Regulation, and Environment. Stanford University Press, Stanford, CA.
Habermas, J., 1996. Between Facts and Norms (W. Rehg, Trans.). MIT Press, Cambridge, MA.
Hans, Valerie P., 2000. Business on Trial: The Civil Jury and Corporate Responsibility. Yale University Press, New Haven, CT.
Hart, H.L.A., 1961. The Concept of Law. Clarendon Press, Oxford, UK.
Heise, M., 2002. The past, present and future of empirical legal scholarship: judicial decision making and the new empiricism. University of Illinois Law Review, 819.
Horwitz, M., 1977. The Transformation of American Law. Harvard University Press, Cambridge, MA.
Jolls, C., Sunstein, C., Thaler, R., 1998. A behavioral approach to law and economics. Stanford Law Review 50, 1471.
Kagan, R., 2000. The consequences of adversarial legalism. In: Kagan, R., Axelrad, L. (Eds.), Regulatory Encounters, pp. 372–413.
Kahneman, D., et al., 1982. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, New York.
Kelsen, H., 1967. Pure Theory of Law (M. Knight, Trans.). University of California Press, Berkeley, CA.
Kennedy, D., 1997. A Critique of Adjudication: Fin de Siecle. Harvard University Press, Cambridge, MA.
Kress, K., 1989. Legal indeterminacy. California Law Review 77, 283–337.
Lessig, L., 1995. The regulation of social meaning. University of Chicago Law Review 62, 943.
Levinson, Justin D., Smith, Robert J. (Eds.), 2012. Implicit Racial Bias across the Law. Cambridge University Press.
Llewellyn, K., Hoebel, E., 1941. The Cheyenne Way: Conflict and Case Law in Primitive Jurisprudence.
Lobel, O., 2004. The renew deal: the fall of regulation and the rise of governance in contemporary legal thought. Minnesota Law Review 89, 342.
Luhmann, N., 1995. Social Systems (J. Bednarz, Trans.). Stanford University Press, Stanford, CA.
Macaulay, S., 1963. Non-contractual relations in business: a preliminary study. American Sociological Review 28, 55.

MacCormick, N., Summers, R. (Eds.), 1991. Interpreting Statutes. Dartmouth, Aldershot, UK.
Macey, J., 1986. Promoting public-regarding legislation through statutory interpretation: an interest group model. Columbia Law Review 86, 223.
MacKinnon, C., 1989. Toward a Feminist Theory of the State. Harvard University Press, Cambridge, MA.
MacNeil, I., 1980. The New Social Contract. Yale University Press, New Haven, CT.
Miles, T., Sunstein, C., 2008. The new legal realism. University of Chicago Law Review 75, 761.
Nourse, V., Shaffer, G., 2009. Varieties of new legal realism: can a new world order prompt a new legal theory? Cornell Law Review 95, 61.
O'Connell, Ann Joseph, 2008. Political cycles of rulemaking: an empirical portrait of the modern administrative state. Virginia Law Review 94, 889–986.
Olson, M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.
Posner, R., 1972. Economic Analysis of Law. Little, Brown, Boston.
Priest, G., Klein, B., 1984. The selection of disputes for litigation. Journal of Legal Studies 13, 1.
Rodriguez, D., Weingast, B., 2003. The positive political theory of legislative history: new perspectives on the 1964 civil rights act and its interpretation. University of Pennsylvania Law Review 151, 1417.
Rubin, E., 1988. The practice and discourse of legal scholarship. Michigan Law Review 86, 1835.
Rubin, E., 1996. The new legal process, the synthesis of discourse, and the microanalysis of institutions. Harvard Law Review 109, 1393–1438.
Rubin, E., 2001. Passing through the door: social movement literature and legal scholarship. University of Pennsylvania Law Review 150, 1.
Sabel, C., Simon, W., 2004. Destabilization rights: how public law litigation succeeds. Harvard Law Review 117, 1015.
Schauer, F., 1988. Formalism. Yale Law Journal 97, 501–548.
Schlegel, J.H., 1995. American Legal Realism and Empirical Social Science. North Carolina Press, Chapel Hill, NC.
Schubert, G., 1959. Quantitative Analysis of Judicial Behavior. Free Press, Glencoe, IL.
Shapiro, M., 1981. Courts: A Comparative and Political Analysis. Chicago University Press, Chicago.
Suchman, M., Mertz, E., 2010. Toward a new legal empiricism: empirical legal studies and new legal realism. Annual Review of Law and Social Science 6, 555.
Tamanaha, B., 1997. Realistic Socio-legal Theory. Clarendon Press, Oxford, UK.
Tamanaha, B., 2010. Beyond the Formalist-Realist Divide: The Role of Politics in Judging. Princeton University Press, Princeton, NJ.
Trubek, D., 1984. Where the action is: critical legal studies and empiricism. Stanford Law Review 36, 575.
Ulmer, S., 1960. The analysis of behavior patterns in the Supreme Court of the United States. Journal of Policy 22, 429.
Unger, R., 1975. Knowledge and Politics. Free Press, New York.
University of Pennsylvania Law Review, 1996. Symposium: law, economics and norms. University of Pennsylvania Law Review 144, 1643.
Vanderbilt Law Review, 1998. The legal implications of psychology: human behavior, behavioral economics and the law. Vanderbilt Law Review 51, 1495–1788.
Wechsler, H., 1959. Toward neutral principles of constitutional law. Harvard Law Review 73, 1.
Weisberg, R., 1992. Criminal law, criminology, and the small world of legal scholars. University of Colorado Law Review 63, 521.
Williams, P., 1991. The Alchemy of Race and Rights. Harvard University Press, Cambridge, MA.
Williamson, O., 1985. The Economic Institutions of Capitalism. Free Press, New York.

Legal Systems, Classification of

David S Clark, Willamette University College of Law, Salem, OR, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Law' and 'legal systems' as concepts are closely related. Since there have been many philosophies and definitions of law, ideas about legal systems have been similarly diverse. A system involves regular interactions among elements that together make up an entity with boundaries. Thus, lawyers, judges, legislators, administrators, the police, and legal scholars all work with particular rules in regularized ways that involve cultural expectations about their roles and the legal institutions and processes with which they interact. This view of a legal system is greater than only the rules themselves. Classification suggests comparison. Comparison here contemplates more than one legal system and the search for similarities and differences among legal systems.

'Law' and 'legal systems' as concepts are closely related. Just as there have been many philosophies and definitions of law, ideas about legal systems have been similarly diverse. A legal system may refer to the rules or institutions of a tribe, city, nation, or international order, or to the natural rules for humankind itself. Moreover, a legal system involves regular interactions among elements that together make up an entity with boundaries. Thus, lawyers, judges, legislators, administrators, the police, and legal scholars all work with particular rules in regularized ways that involve cultural expectations about their roles and the legal institutions and processes with which they interact. This view of a legal system is greater than only the rules themselves, although some people conflate law or even a legal system with legal rules.

Classification of legal systems suggests comparison, inviting the search for similarities and differences among legal systems. Approaches to comparing legal systems or parts of legal systems often involve a broad view for the discipline of comparative law (Clark, 2012). This article is a description of that process of classification and comparison, the varying interest in such matters among civilizations, the approaches that individuals have used, and current theories, purposes, and criticism of legal system classification and comparison.

De-emphasis on Classification in Classical Legal Traditions

Athens

Ancient Greeks made the earliest recorded efforts to explicitly classify legal systems based on their interest in political structure, philosophical speculation, and the comparative study of law. For instance, Plato (429–348 BCE) in the *Laws* discussed the rules of several Greek and other *poleis* (city-states) in formulating his ideal code and legal institutions for Magnesia. He showed how differing constitutional forms yielded either a good state or a bad state. In the *Republic*, he sought to ground law and politics in a model state with a transcendental metaphysics that elite Guardians would implement with an educated vision of the Form of the Good. This approach to achieving justice resonated with subsequent natural law jurists, who would have the basis to compare an actual system of legal rules with a utopian system. Karl Marx (1818–83) used Plato's class-based society in his own theory of law, where each ruling class he analyzed applied law – the product of economic forces – to control society's lower classes.

Aristotle (384–322 BCE) examined many legal structures in his *Politics* before settling on the three preferred categories of kingship, aristocracy, and constitutional government (*politeia*) with their deviant siblings of tyranny, oligarchy, and democracy. He rejected Platonic transcendentalism as well as the Sophists' relativism and skepticism. Aristotle's approach was empirical, rooted in how humans actually govern themselves. He found virtue in the moderation (golden middle) of a legislator, judge, or legal rule. Aristotle's successor, Theophrastus (372–287 BCE), examined comparative commercial law (Bodenheimer, 1974, pp. 6–14, 79–80).

Rome

In Rome, jurists' interest in foreign law was more practical, particularly with the creation of the office of *praetor peregrinus* in 242 BCE. This praetor supervised jurisdiction over disputes involving non-Romans. It was here that the mixture of a less formalistic Roman *ius civile*, Greek law, and other foreign legal norms led to an internationalized *ius gentium*. Roman political authorities allowed distinct ethnic groups in the provinces to rule themselves in most matters, but Roman jurists did not have a philosophical interest in studying and classifying foreign legal systems since they believed in the superiority of their own law. Cicero (106–43 BCE) illustrated this view. An exception that makes the point was the fourth-century *Collatio legum Mosaicarum et Romanarum*, in which excerpts from Roman jurists were paired with the laws of Moses either to improve the stature of Christianity or to demonstrate that Rome had implemented Mosaic law (Clark, 2012).

China

The experience with law in Chinese history is completely different. For most of the past two and a half millennia, Chinese philosophers and political leaders have de-emphasized the use of law in society and therefore found no need to make comparisons with legal rules and institutions in

foreign places. Like the Romans, the Chinese believed in the superiority of their own approach to ordering society and maintaining peace. This separation from foreign legal sources ended in the mid-nineteenth century, when European governments (and the United States) forced Chinese emperors during the Qing Dynasty into unequal treaties granting Western powers and their citizens' special rights, including in Chinese territory.

In place of law, the Chinese in East Asia during the Zhou Dynasty developed the idea that human nature is inherently good, social relations are part of the natural order, and virtuous conduct can be achieved through education and deference to the example of community rulers. Confucius (K'ung fu-tze, 551–479 BCE) refined these ethical tenets to serve also as the basis for government and politics. Subsequent philosophers, such as Dong Zhongshu (179–104 BCE), further elaborated a form of Confucianism emphasizing cosmic harmony.

The ideal person within this ideology is one learned in *li* – rites or norms of proper behavior – who follows the rules modestly and instinctively, thereby repressing his or her own self-interest for the greater good of maintaining harmony. It is obvious that such a community would not need a large legal apparatus with its own set of rules to control human conduct, maintain social order, and facilitate justice. Order and justice are subsumed within harmony. This underlay Chinese ideology until the twentieth century and informs national legal cultures up to the present in those legal systems that China has influenced.

Before Confucianism prevailed as the spiritual and philosophical basis of the Chinese state during the Han Dynasty (206 BCE–220 CE), the political reality of the 'Warring States' Period (481–221 BCE) involved continued armed conflict among rival territorial princes. This reality – so unlike Confucian society – supported the emergence of the Legalist School, with philosophers such as Han Fei Zi (280–233 BCE). Legalists held that humans, in essence, were selfish, looking out for their own interests. When education could not induce proper behavior, the state must punish individuals – both to control them and to set an example for others – with legal rules (*fa*) and processes.

Emperors issued a series of codes that dealt almost completely with administrative and criminal matters. The Tang Code (624 CE) is the earliest that has survived intact. Even when the issue appears to be family or succession law, the rules treat those subjects in an administrative (such as taxation) or penal context. The codes recognize the Confucian notion that legislation is a necessary evil, required only when the social order is seriously disrupted or the state is threatened. The Confucian influence on the penal law appears in the gradation of penalties based on the social status of the criminal and victim as much as the actor's intent or surrounding circumstances. Beyond China, Confucianism has been a major influence on the legal systems of Japan, Korea, Taiwan, and Vietnam.

As far as private law was concerned, the teachings of Confucianism emphasized modes of dispute resolution outside courts. Emperors sent imperial magistrates – men educated in the Confucian literature – to the provinces to carry out administrative and judicial duties and to encourage obedience to Confucian virtue. Disputes should be settled within the smallest unit with authority, whether the family, kin group, village, or commercial guild. Conciliators – with prestige due to age or rank – used the standards of *li* to seek compromise. Community pressure supported that solution. Reconciliation was justice. Only when the conciliators failed would a disputant take the matter to a provincial magistrate, suffering censure for that decision, and experiencing the delay, expense, and corruption of the state courts (Glenn, 2010, pp. 319–360; Menski, 2006, pp. 493–569; Merryman et al., 2010, pp. 553–573).

India and Hindu Law

As in China with Confucianism, Hinduism in India is an ancient system of thought, with religious, philosophical, and social underpinnings, that has served as a guiding force in society to control human conduct. The aim of Hinduism is to provide the individual with a moral compass to guide virtue and piety. Karma from good deeds in this life will permit the transmigration of the soul to a better existence in the next life, perhaps to a higher caste or, ultimately, the soul's release as a higher spiritual being. There is no record of efforts among classical Hindu legal scholars to classify or compare legal systems, but distinct schools of law in India developed in the eleventh and twelfth centuries that might have stimulated such interest.

Hindu law has passed through several periods of reconfiguration from Vedic times (1500 BCE) to the classical phase (500 BCE–1100 CE), the postclassical era, and English influence in India through the East India Company in the seventeenth century. Unlike Islam, scholars do not largely derive Hindu law from ancient religious scripture. Hindus do not need to hold specific religious beliefs. Rather, Hindu law is a kind of common law, continuously developed through customary practices and the historical records of Brahmin scholars until British officials took a greater interest in Hindu law after 1772. Since then, an Anglo-Hindu case law developed together with British statutory intervention in India, which today legal scholars distinguish from premodern Hindu law.

Two points are relevant here. First, Hindu law was less connected to central government activities than the situation with Confucianism in China. India did not have a similar series of imperial codes or imperial magistrates such as those in China. Second, Hindu law was more diverse than Chinese law with a large number of local variations. It also had a richer private law. It emphasized the practice of plurality and relative justice with little interest in uniformity of law. What the two systems shared was a de-emphasis of law in society compared to Western traditions. Hindu law today applies beyond India to Nepal and parts of Africa and Southeast Asia.

The ancient Hindu texts (in Sanskrit) related to law emphasize *dharma*, the obligation of every person to do the right thing at all times – to take the virtuous path. The pluralism of Hindu law lies in the diverse implementation of this principle in the infinite sociocultural circumstances of Indian life, by historical period, region, caste status, gender, age, and so on. The tension with modern Western law is obvious. Hindu law teaches that fixed rules might cause injustice. The endless distinctions treat every individual as separate units, linked all the same by a common conceptual bond in a macrocosmic order (*rita*) or secular truth (*satya*).

Dharma, the appropriate action, must consider all the circumstances with a view to promote the common good. In one sense, the individual is the ultimate agent to determine the 'law' in any particular situation, consequently reducing the role for the state as lawmaker (Glenn, 2010, pp. 288–318; Menski, 2006, pp. 193–278; Zweigert and Kötz, 1998, pp. 313–319).

Islamic Law

Islam, including its law, developed in the Arabian Peninsula in the seventh century CE. It considered Muhammad (570–632) the last in a series of prophets for the monotheistic faith of Abraham, Moses, and Jesus. Islamic law (shari'a) encompasses the system of duties incumbent upon a Muslim by virtue of his or her religious belief. It is the expression of Allah's command for the Muslim community, the revealed word of God (the Qur'an). Spread by conquest and trade, Islam has had a proselytizing appeal, since it recognized no barriers of race or class.

Shari'a is, as is canon law for Roman Catholics or Halachah (Jewish law) for Jews, the law of a religious community. Canon law prior to 1500 had almost universal applicability in Europe, since the Christian church had all the attributes of government (jurisdictions) it shared with secular European rulers. Jurists consider canon law as an early example of 'modern' in the Western sense since it borrowed some of the sophisticated legal norms and lawmaking and interpreting techniques of Roman law. Together with rediscovered and elaborated Roman law, canon law constituted the *jus commune* of Europe. Canon law declined in importance with the rise of statism and nationalism in Europe.

Shari'a today dominates in the Middle East, most of the north half of Africa, Southwest Asia, Malaysia, and Indonesia. It is also influential in other parts of the world; globally, Islam has more than 1 billion believers. Islamic law is a total belief system, since it regulates a person's relationship not only with others and with the state, but also with God. Shari'a is a comprehensive normative system that embraces both private and public conduct. It thus includes ritualistic practices as well as a system of legal rights and duties.

Fundamental to Islamic law is the belief, largely abandoned in Western secular law, that law is an expression of divine will. With the death of Muhammad in 632, communication of the divine will to man ceased. Unlike Western law, which jurists expect to evolve in response to major changes in society, the dominant Muslim jurisprudence states that it should not be society that frames the law, but rather the divine law that molds and controls society. The 'unchanging character of Islamic law' created interesting problems of interpretation for the scholars and operators of Islamic law, and differing views concerning the appropriate solution to such problems account for some of the differences between 'schools' of Islamic law. The major Sunni schools, established by the eleventh century, are Hanafi (Turkey, west Mediterranean, central Asia, Egypt, Afghanistan, Pakistan, India, and Bangladesh), Maliki (north and central Africa), Shafi'i (east Africa, Malaysia, and Indonesia), and Hanbali (the Arabian Peninsula). About 10% of believers are Shi'a, most of whom live in Iran, Iraq, Pakistan, India, Yemen, and Turkey.

These Sunni schools and their difference with Shi'a jurists – the basis for legal system classification – provided the opportunity to develop comparative law, but there is little evidence that this occurred. As with the split of Western Christendom into Roman Catholics and Protestants, the divergence in religious law most often led to hostility and the attempt to exclude or conquer the nonconforming believers. Nevertheless, there were many examples in classical Islam of tolerance for Jews and Christians to live with their own laws.

With the expansion of Western colonization in the nineteenth century, Islamic law declined in importance. In the twenty-first century, however, there is renewed debate in Muslim countries about the place for classical Islam in modern society. Its importance is obvious, and comparatists are not ignoring its relevance for understanding the legal systems of much of Africa and Asia. The utility of humanistic and social science methods is apparent (Glenn, 2010, pp. 181–236; Menski, 2006, pp. 279–379; see Rosen, 1999).

Influence of the Classical Legal Traditions

From the classical modes of thinking about law, its classification, and its place in society, the only one to leave a lasting legacy for modern comparative law was Roman law. Of all the great ancient civilizations, the thousand-year history of Rome as it developed in Western Europe first accommodated foreign law within the republic through the institutional framework of the *praetor peregrinus*. This praetor's annual edict, regulating jurisdiction in disputes between foreigners, and later between citizens and foreigners, together with its juristic interpretation comprised a pragmatic response to the reality of conquest and trade. During the empire, jurists interpreted the legal norms that developed within this institution as *ius gentium*, the law of peoples. Romans thought of this as one version of a universal law in comparison to the other universal, less practical, natural law, or the *ius civile* peculiar to Roman citizens. The Romans borrowed and absorbed some Greek thinking about law, which saw, in addition, a second life during the Renaissance as university law masters rediscovered Aristotle and Plato.

An important difference between Roman law and the other classical laws was that Roman law did not have a strong ideological or religious basis for its existence and authority. Jurists had differentiated it from religion, morality, or philosophy. The important elements of Roman law were pragmatic, with private jurists developing most legal norms in a highly decentralized fashion that they incorporated into classical Roman law, eventually embodied in Emperor Justinian's *Corpus juris civilis* (529–534). This compilation served as the basis for university law study and influence on canon law in Europe after the twelfth century. The law faculties that existed in a large number of European universities, teaching and studying Roman and canon laws, served as the natural place to develop interest in comparative law after the Enlightenment forces of statism and nationalism began to support national law as their alternative in the eighteenth century.

Moreover, while the Chinese, Hindu-Indian, and Islamic legal systems continue in force up to the present, and therefore governments supporting those systems have a strong interest in maintaining their mythic authority, such is not the case for Roman law. In the eighteenth and nineteenth centuries, Roman law often served as the alternative to national law in utilizing comparative law methods. This was as true for Sir William

Blackstone (1723–80) at Oxford in his thesis that English law was as rational as Roman law as it was for John Adams (1735–1826) in arguing against English law as applied in the American colonies. The major exception to this was the new German empire, which embraced Roman law as part of the *usus modernus pandectarum* (Clark, 2012).

Modern Classification and Comparison of Legal Systems

Early Modern Classification by European Jurists

In the early modern era, some natural law jurists such as Hugo Grotius (Hugo de Groot, 1583–1645), Samuel Pufendorf (1632–94), and Baron Charles-Louis de Montesquieu (1689–1755) used comparative law methods to provide empirical backing for natural law principles. Others such as Blackstone had the additional objective of providing prestige for their national law: his four-volume *Commentaries on the Laws of England* (1765–69) emphasized similarities among English common law, Roman law, and natural law.

Montesquieu's extensive speculation about the relation of law to natural and social forces and explicit classification of legal systems emphasized the distinct manners (*moeurs*) of a particular people in a particular era. His *Lettres Persanes* (1721), published anonymously, used the device of two Persians traveling in Europe who satirized the customs of French and European societies. They pointed to the abuses and cruelties of criminal justice systems and argued for a rational theory of punishment adjusted to the settled sensibilities of the people.

Montesquieu's relativist view of humans' ability to achieve justice was further developed in *De l'esprit des lois* (1748), which drew upon causative factors such as climate, soil, population size, people's morals and customs, religion, and commerce. He classified governments along with the appropriate principle for their constitutive societies. Therefore, democracy required civic virtue; aristocracy, moderation against the people; monarchy, honor for the law and intermediate institutions; and despotism, fear to maintain order. He found different mechanisms for lawmaking and adjudication in the first three types of legal systems, but pointed out that law was not necessary for despotism.

Montesquieu had an important impact on others, including Cesare Beccaria (1738–94) and, in the nineteenth century, Alexis de Tocqueville and Émile Durkheim. Beccaria shared Montesquieu's concern with injustice and cruelty in criminal law and procedure in Europe. Beccaria's *Dei delitti e delle pene* (1764), soon translated into English, French, and German, led to restrictions on the use of torture, judge-interpreted crime, and punishment that exceeded a proportionality standard measured by social harm (Clark, 2012).

Nineteenth-Century Classifications

By the nineteenth century, many European jurists and other scholars were developing comparative sociolegal studies. A few of the best known were Léon Duguit (1859–1928), Eugen Ehrlich (1862–1922), Otto von Gierke (1841–1921), Eugenio María de Hostos (1839–1903), Cesare Lombroso (1835–1909), Gustav von Schmoller (1838–1917), William Graham Sumner (1840–1910), Gabriel de Tarde (1843–1904), and Alexis de Tocqueville (1805–59), along with the American Lewis Henry Morgan (1818–81). In Europe, theorists, often at universities, had secured the field of political economy as one that examined the usefulness of political and legal institutions and policies at the level of the state to develop economic production, exchange, distribution, and wealth. Some of those working in this tradition, such as Adam Smith (1723–90), were comparative and explicitly dealt with legal institutions and processes.

Legal anthropology at this time emphasized evolutionary studies, but was also often comparative. European scholars were fascinated with ancient and 'primitive' societies, which by their nature they would have to compare with contemporary European legal systems to provide meaning for the reader. Henry Maine is an obvious illustration, but so too were the studies of Johann Bachofen (1815–87), John McLennan (1827–81), and Lewis Morgan. Although modern anthropologists find most of this research inadequate due to its armchair methods, the emphasis on cultural and historical context remains important. The shift in methodology in the twentieth century to intensive ethnographic research in the subjects' culture and language led to less comparative work, although there have been a few exceptions (Goodale and Mertz, 2007, pp. 68–70).

Further sophistication in comparative law and society came with attempts to classify law and legal systems within two or more categories. Nineteenth-century evolutionary theorists developed elaborate classifications of local and national legal systems. Law in these schemes was part of the larger social system that changed over time. For instance, Sir Henry Maine (1822–88), interested in the striking parallels between the development of English common law and classical Roman law, posited in *Ancient Law* (1861) stages of social development common to different peoples that could be correlated with particular instruments of legal growth. Jurists could characterize legal systems by whether social change occurred primarily by legal fictions, equity or natural law, or conscious creative legislation.

Karl Marx, schooled in Roman law as well as history and philosophy, accepted the reigning concepts of his time about the nature of humans and society. After he added the study of economics, he developed (along with Friedrich Engels, 1820–95) a conception of law that was materialist. Law, in *Zur Kritik der politischen Ökonomie* (1859), assumed a pivotal role in his base–superstructure model of social evolution. In essence, Marx's works supported a theory of law as an instrument of social power, which itself was a reflection of economic relations. In turn, it was people's gradual conquest of nature, a matter of technology, which determined economic relations. Marxist historical materialism comprehends society as structured by the material conditions at any given time. It identified five successive stages in the evolution of these material conditions in Europe, each with a different role for law. These stages were primitive communism, slavery, feudalism, capitalism, and socialism – and perhaps, one day, communism without the state or law.

Émile Durkheim (1858–1917), the first university sociologist recognized as such in France, had wide-ranging interests that included law. In his first book, *De la Division du Travail*

Social (1893), Durkheim described the evolution of societies from mechanical solidarity to organic solidarity. He argued that forms of law provide an index of the type of social solidarity that exists. Repressive law, concerned with punishing offenses, represents mechanical solidarity or the type of social cohesion based on people's shared beliefs and personal ties. A limited degree of religious or cultural diversity could exist in such a society. By comparison, restitutive law, which restores the status quo if social relations are disturbed, represents organic solidarity or cohesion based on functional interdependence. Here, there is substantial division of labor, and comprehensive contract, tort, and commercial law mediates among diverse individual and social interests. In Durkheim's later writing, he emphasized the sacredness of abstract individuals, a religion of humanity to replace the traditional religion found in repressive-law regimes. This was not an egoistic individualism, but a moral individualism that would enhance human dignity and autonomy.

Twentieth-Century Classifications

Anthropologists had frequently observed that some societies operate without government but have social order and law-like institutions. In modern nations, there might be many levels or discrete groups in societies (with their own 'legal' systems), so it became appropriate to speak of legal pluralism. This emphasis on legal pluralism in the twentieth century replaced the earlier prominence of evolutionary analysis. Comparing legal systems then depends on the characteristics of different societies or their subgroups. Leopold Pospíšil, for example, described this situation as a patterned mosaic of subgroups that belong to certain identifiable types with different memberships. Each subgroup owes its existence to a legal system that regulates the behavior of its members. These legal systems then form a hierarchy reflecting the degree of inclusiveness of the corresponding subgroups that one categorizes by legal level, such as the family, lineage, community, region, or nation. People are simultaneously members of several subgroups of different inclusiveness, which may lead to contradictory legal loyalties (1971, pp. 106–126).

Durkheim had written, "Comparative sociology is not a particular branch of sociology; it is sociology itself" (1938, p. 139). This seems self-evident in that the gold standard of the scientific method in the social sciences rests on the comparison of experimentally allocated groups. For sociology of law – even more so for some other social sciences related to law – the challenges of attaining this standard are great. There are statistical ways to simulate scientific comparison using the data that scholars and institutions produce, including from experiments and quasi-experiments, but the task is still difficult.

Max Weber (1864–1920) approached the problem differently. He used many of the same historical sources as legal evolutionists, but in *Wirtschaft und Gesellschaft* (1922) constructed ideal types to facilitate the comparison of actual social and legal systems in trying to explain the rise of industrial capitalism. His four-cell typology for lawmaking and law application considered the rationality or irrationality of legal thought as well as whether legal norms and decision making were highly differentiated from religion, ideology, or emotion. He argued that the category of formal (autonomous) rationality, with a consistent body of general legal norms, provided the predictability facilitating modern capitalism.

Lawrence Friedman in the 1970s argued that scholars could classify legal systems according to their internal legal culture, which includes the attitudes and values of judges, lawyers, and other legal actors toward a task such as lawmaking or adjudication that relies on reasoning or some other basis of legitimacy. He considered two variables. First, when a legal system requires reasoning premised on legal propositions, it is a closed system; when propositions may be nonlegal, the system is open. Second, some systems regard legal propositions as fixed and timeless; others accept change as normal.

Using these two variables, Friedman derived four types of reasoning associated with law. First, closed systems, such as those based on sacred law, formally reject the idea of change. What change does occur must be based on legal fictions. Second, a legal science system such as that in Germany is closed but permits change by simply adding legal rules or deriving new legal principles. Third, traditional or customary law systems do not accept explicit change but are open to new norms based on shifting behavioral patterns. Fourth, a system permitting change by looking outside the stock of legal propositions includes what Weber called substantive rationality. This could be revolutionary legality, such as that in socialist Russia, or welfare legality, which looks to ethical imperatives or utilitarian principles (1977, pp. 76–86).

The comparative politics revolution in the United States during the 1950s shifted law and political science from a field that tended to emphasize public law and describe the legal institutions of foreign governments to one that explained political behavior in structural and functional terms. Gabriel Almond and Bingham Powell, for instance, used the developmental variable of cultural secularization and structural differentiation together with the low or high level of political subsystem autonomy associated with interest groups and political parties to classify political systems. Modern systems could be premobilized, mobilizing, or penetrative and, within each of these categories, either authoritarian or democratic (1978).

By the 1980s and 1990s, an emphasis on empirical and comparative methods to analyze government institutions and processes and the ways in which people relate to them had spread beyond the United States to Europe and elsewhere. Globalization was one of the forces stimulating research on judicial behavior, juries, courts and their processes, criminal justice systems (overlapping criminology), administrative law, dispute transformation and resolution, constitutional courts, and civil and human rights.

Recent Theory and Critique of Legal System Classifications

A useful classification of legal systems should make the criteria of division clear and allow scholars to assimilate, simplify, and comprehend the great wealth of legal minutiae existing in individual systems. The dominant approach to classifying national legal systems in contemporary comparative law is to identify aggregate families, styles, or traditions. Beyond satisfying the aim of simplification, this idea might permit one to

understand more clearly pluralistic legal systems with an official, dominant legal tradition but also including significant religious or ethnic communities lacking a sovereign political organization or indigenous communities with their own law. In the last few decades, recognition of the importance of legal pluralism within national legal systems and consideration of the system's structural and cultural aspects have opened classifications to influence from disciplines related to law.

Legal Families: The Influence of René David

Comparative lawyers have disagreed about the criteria to use for classification, but the results have been surprisingly similar, differing primarily in the number of categories emerging. René David (1906–90) in 1950 used two criteria that were further refined in his 1964 *Les grands systèmes de droit contemporains*, which saw 11 editions through 2003 and translation into English and other languages (David, 1950, 1964; David and Brierley, 1985). David believed that the primary criterion should be the legal system's ideology; its philosophical, political, or economic principles; its social objectives; and the place of law within the social order. A secondary criterion should be a system's legal technique, the law's conceptual structure, and the theory of sources of law. Based on these guides, David found three prominent legal families: Romano-Germanic, common law, and socialist law. Some elements from one or more of these families have influenced all national legal systems. In addition, he identified lesser legal systems, such as those for religious law (Muslim and Hindu) and for East Asia and Sub-Saharan Africa (Pargendler, 2012). David's typology substantially influenced the organization of the *International Encyclopedia of Comparative Law*, a 17-volume undertaking that was the most ambitious of the twentieth century. He edited the second volume, entitled *The Legal Systems of the World: Their Comparison and Unification* (1974).

Legal Traditions

David and Brierley argued that it "is a superficial and indeed false view to see law as being composed of the totality of [legal] rules. … [I]t is linked to a view of the social order itself which determines the way in which the law is applied and shapes the very function of law in that society" (David and Brierley, 1985, p. 19). John Merryman proposed a similar idea for legal traditions, with the same three rated as most influential: civil law, common law, and socialist law. Merryman asserted that a legal tradition is not a set of rules: "Rather it is a set of deeply rooted, historically conditioned attitudes about the nature of law, about the role of law in the society and the polity, about the proper organization and operation of a legal system, and about the way law is or should be made, applied, studied, perfected, and taught. It puts the legal system into cultural perspective" (1969, p. 2). This notion of legal traditions or families still dominates the teaching of comparative law in most of the world.

Legal Styles

Similar to David's revised formulation of legal families and Merryman's legal traditions, Konrad Zweigert (1911–96) wrote about his theory of legal circles, or *Rechtskreisen* (1961). He looked at the styles (*Stile*) of legal systems to divide them into legal spheres or families. He analyzed earlier attempts at classification as too one-dimensional, turning on primarily one criterion. Zweigert, as updated by Hein Kötz (1998), found five factors to be crucial to classify the style of a legal system: (1) historical background and development, (2) predominant mode of legal thought, (3) distinctive legal doctrine or institutions, (4) sources of law, and (5) ideology. Using these factors, Zweigert and Kötz settled on eight categories: Romanistic, Germanic, Nordic, common law, People's Republic of China (replacing the earlier socialist style), Japanese, Islamic, and Hindu.

Critique of Current Classifications

There are several problems with these classification schemes, which their proponents often recognized. First, one intends a classification for a specific purpose, which may make it unsuitable for another purpose. Thus, Zweigert and Kötz intended their divisions to serve as chapters in their comparative law treatise for students and jurists. They omitted an African legal style because they argued it is still in the process of emerging.

In addition, scholars base almost all comparative law classifications on a private law emphasis, which weakens their utility for public law comparisons. For instance, a study of the judicial review of legislation as an aspect of constitutionalism could put England and France in one category but Germany and the United States in another, confounding the civil law–common law dichotomy. Mirjan Damaška rejected the usual classification of legal families in his comparative study of procedure and the administration of justice. He formed a Weberian four-cell model that considered the structure of state authority on one axis (hierarchy or coordinate) and the type of proceeding on the other (conflict solving or policy implementing). Illustrations included the active judge in continental European civil procedure in the conflict-solving hierarchy cell, English and US civil and criminal procedures in the conflict-solving coordinate cell, Soviet civil and criminal procedures in the policy-implementing hierarchy cell, and US public interest litigation in the policy-implementing coordinate cell (1986).

Second, no matter how logical a classification, there will always be hybrid legal systems that do not fit easily within one category or another. Well-known hybrids include, for example, Scotland, Louisiana, Puerto Rico, Québec, Israel, South Africa, Sri Lanka, and the Philippines, which share civil law and common law elements. The People's Republic of China has both East Asian and socialist elements, India has Hindu law for family and inheritance matters but common law for obligations and commercial issues, and Arabian countries have Islamic law for family and inheritance issues but French-influenced law for the commercial sector. Some comparatists assert that all national legal systems are hybrid, thus undercutting the value of classification (Glenn, 2010).

Third, related to the problem of hybrid systems is the situation of legal pluralism, associated with the personality principle for choice of law, in which an official, dominant legal tradition coexists with an indigenous or religious community

that by permission of the state or *de facto* regulates itself according to its own norms. A cultural or anthropological approach is useful for examining this condition (Glenn, 2010).

Fourth, Ugo Mattei suggested a new tripartite classification to reduce the Eurocentrism of earlier approaches. His families depend on the dominant pattern of legal decision making in a system, which he characterized as professional law, political law, or traditional law. In the professional law family (including Western European civil law nations and the common law United Kingdom, Oceania, United States, and Canada), the idea of law predominates, autonomous from politics and religion. The political law family includes ex-socialist and developing countries (Eastern Europe, Sub-Saharan Africa, and Latin America) in which the notion of autonomous law is not strong enough to prevent politics from supplanting legal rules for important legal decisions. Finally, the traditional law family consists of legal systems that may have absorbed substantial elements from the civil law or common law (Asia and North Africa), but in which the dominant culture controlling legal decision making is rooted in Confucianism, Hinduism, or Islam (1997).

Finally, some comparatists have recently engaged the 'legal origins' thesis developed by economists. This approach classifies countries into either the civil law or common law category, and it utilizes empirical data to support two assertions. First, the structure and processes of legal institutions are associated with the origins of the legal system, either civil law or common law. To isolate the salient features of these two traditions (and avoid the unique features of the societies in which they arose), researchers examined those nations subject to European colonization, assuming that the Europeans imposed the transplant with little regard for indigenous factors. Second, the nature of legal institutions will affect economic growth. This is similar to Weber's attempt to explain the rise of industrial capitalism. For some scholars, this framework is descriptive; for others, it is evaluative. The World Bank, for instance, has adopted the latter perspective; it now collects relevant legal data for its series of *Doing Business* national reports. Those in the evaluative camp argue that common law nations perform better overall in economic growth than civil law nations. This outcome proceeds through two channels. For the political channel, judges in the common law are more independent than in the civil law. Thus, the government has less influence on market operations. In the adaptability channel, the case law of common law countries is more flexible and able to adapt to changing social conditions than the legislative texts of the civil law (Michaels, 2009).

Conclusion

Comparative law and society is a challenge to the dominant legal theory prevailing today in most countries among politicians and legal professionals who often ignore cross-cultural legal classifications. The clearest challenge comes from legal anthropologists and others who emphasize legal pluralism as a fundamental reality of social life. Legal pluralism contests the four premises of established theory: nationalism, statism, monism, and positivism.

Nationalism as it relates to law took hold in its modern form in two major waves. The first occurred after the American and French revolutions and dramatically reshaped the legal landscape of Europe and the Americas in the nineteenth century. The second wave happened after World War II in conjunction with the United Nations' mandate for decolonization. Since 1945, under the principle of self-determination, more than 80 former colonies have gained their independence. Some of these new political units were the result of splitting or reformulating former legal structures, and others emerged from consolidation. However, all emphasized what was special about the people within them and different from other people. This singularity, tied to citizenship, could coincide with language, ethnicity, culture (or history), religion, climate, geography, or other features. Nationality demanded symbols to further unity: a flag, a song, a seal, the pledge of temporal allegiance, perhaps a bird, tree, flower, or motto, and possibly a currency or state religion.

Statism suggests that the state should have a monopoly over sovereignty and its implementation through jurisdiction. In traditional Enlightenment terms, this includes jurisdiction to legislate, administer, and adjudicate. Statism supports nationalism and is hostile toward legal pluralism. The latter maintains that individuals may have multiple obligations or loyalties or in essence live outside official state authority. Legal pluralism characterized all legal systems prior to the rise of nationalism. It reflected the more complicated legal universe in which people lived, especially salient during colonialism.

Monism in law is similarly hostile to legal pluralism. Monism holds that there exists the possibility of a single, internally coherent, legal system. Thus, it is supportive of both nationalism and statism so long as the level of concern is the state. However, in the modern nation-state system of independent (and equal) sovereign countries, how can law, binding upon a particular state, exist beyond that state? How can there be international law? The universal monist response, as in France, is that the state will incorporate international legal norms and jurisdiction into its national legal system without exception. Consequently, the state is in reality a subunit of the larger international legal system. The more frequent national monist response, as in the United States (with its supreme Constitution) and even more strikingly in the United Kingdom (with its parliamentary supremacy), is essentially dualist. There exist two distinct legal systems: one national, and the other international. The former will usually incorporate valid international law into its state system as is politically convenient, but in the case of a direct conflict between an international norm and the supreme national law, the latter prevails.

Finally, there is positivism. Positivism in law is the philosophical (or jurisprudential) view that purported norms or institutions that the state does not create or recognize as 'law' are not law. This is the only or dominant theory taught to those who become legal professionals in most of the world. It excludes a meaningful role for natural law, historical jurisprudence, or sociological jurisprudence, even in the face of strong religious influence, ethnic and cultural cohesion, or community institutions and processes such as juries.

The strength of this dominant legal theory and its four elements works against the relevance of comparative law in the

Weltanschauung of most legal professionals. Their legal education is oriented toward national law, perhaps a bit of international law, and little foreign or comparative law. This generalization has less validity today in European countries that are members of the European Union, but even among those nations, some governments make little effort to implement proper education about foreign legal systems. In the rest of the world, nationalism in law is a strong force.

Nevertheless, there have been developments since the 1990s that are encouraging for the understanding of foreign legal systems and the study of comparative law. Economic and cultural globalization is one such obvious development. Small segments of legal professionals in most countries occupy themselves with servicing and dealing with the effects of globalization. A larger group, including judges and government administrators, have occasion to handle foreign and comparative law issues. This involves both transnational transaction work and transnational litigation and arbitration. A few law schools have even promoted themselves as global institutions that are aware and supportive of this international legal integration.

Another encouraging trend is the increased interest among legal scholars and researchers in social science disciplines about legal norms, professionals, institutions, processes, and culture. Most of this work is 'academic,' which in most societies means that it will have little or no impact on the people who do most law jobs. Nonetheless, it offers promise that some small percentage of people will come to appreciate the variety and benefit of legal pluralism. Just as biodiversity is a measure of the health of ecosystems, legal diversity is recognition of the variability and richness of human culture (Clark, 2012).

Bibliography

Almond, Gabriel A., Powell Jr., G. Bingham, 1978. Comparative Politics: System, Process, and Policy, second ed. Little, Brown, Boston, MA.

Bodenheimer, Edgar, 1974. Jurisprudence: The Philosophy and Method of the Law, rev. ed. Harvard University Press, Cambridge.

Clark, David S., 2012. History of comparative law and society. In: Clark, David S. (Ed.), Comparative Law and Society. Edward Elgar Publishing, Cheltenham, UK, pp. 1–37.

Damaška, Mirjan R., 1986. The Faces of Justice and State Authority: A Comparative Approach to the Legal Process. Yale University Press, New Haven, CT.

David, René, 1950. Traité élémentaire de droit civil comparé: Introduction à l'étude des droits étrangers et à la méthode comparative. Librairie Générale de Droit et de Jurisprudence, Paris.

David, René, 1964. Les grands systèmes de droit contemporains. Dalloz, Paris.

The legal systems of the world: their comparison and unification. In: David, René (Ed.), 1974. International Encyclopedia of Comparative Law, vol. 2. J.C.B. Mohr, Tübingen.

David, René, Brierley, John E.C., 1985. Major Legal Systems in the World Today: An Introduction to the Comparative Study of Law, third ed. Stevens & Sons, London.

Durkheim, Émile, 1938. The Rules of Sociological Method (Sarah A. Solovay and John H. Mueller, Trans.), eighth ed. University of Chicago Press, Chicago, IL.

Friedman, Lawrence M., 1977. Law and Society: An Introduction. Prentice Hall, Englewood Cliffs, NJ.

Glenn, H. Patrick, 2010. Legal Traditions of the World: Sustainable Diversity in Law, fourth ed. Oxford University Press, New York.

Goodale, Mark, Mertz, Elizabeth, 2007. Anthropology and law. In: Clark, David S. (Ed.), 2007. Encyclopedia of Law and Society: American and Global Perspectives, 1. Sage Publications, Thousand Oaks, CA, pp. 68–77.

Mattei, Ugo, 1997. Three patterns of law: taxonomy and change in the world's legal systems. American Journal of Comparative Law 45, 5–44.

Menski, Werner, 2006. Comparative Law in a Global Context: The Legal Systems of Asia and Africa, second ed. Cambridge University Press, Cambridge.

Merryman, John Henry, 1969. The Civil Law Tradition: An Introduction to the Legal Systems of Western Europe and Latin America. Stanford University Press, Stanford, CA.

Merryman, John Henry, Clark, David S., Haley, John Owen, 2010. Comparative Law: Historical Development of the Civil Law Tradition in Europe, Latin America, and East Asia. Matthew Bender & Co, New Providence, NJ.

Michaels, Ralf, 2009. Comparative law by numbers? Legal origins thesis, *Doing Business* reports, and the silence of traditional comparative law. American Journal of Comparative Law 57, 765–795.

Pargendler, Mariana, 2012. The rise and decline of legal families. American Journal of Comparative Law 60, 1043–1074.

Pospíšil, Leopold, 1971. Anthropology of Law: A Comparative Theory. Harper & Row, New York.

Rosen, Lawrence, 1999. The Justice of Islam: Comparative Perspectives on Islamic Law and Society. Oxford University Press, New York.

Zweigert, Konrad, 1961. Zur Lehre von den Rechtskreisen. In: Nadelmann, Kurt H., et al. (Eds.), XXth Century Comparative and Conflicts Law: Legal Essays in Honor of Hessel E. Yntema. A.W. Sythoff, Leyden, the Netherlands, pp. 42–55.

Zweigert, Konrad, Kötz, Hein, 1998. Introduction to Comparative Law (Tony Weir Trans.), third ed. Oxford University Press, Oxford.

Legal Systems: Private

Stuart Henry, San Diego State University, San Diego, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Private legal systems are nonstate means for obtaining compliance with the norms or rules of a social order. They comprise the numerous forms of social control that are part of group or organizational life, but that are not formally part of state law. Private legal systems are ubiquitous; more people are directly governed by them than they are by state law. Private legal systems have been analyzed from the perspective of legal pluralism, but research has challenged this approach as inadequate to explain the interwoven relationship between formal law and private legal systems, suggesting that they each exist in a mutually constitutive relationship.

Private legal systems are nonstate institutional means for obtaining compliance with the norms or rules of a social order. They comprise the numerous forms of social control that are part of group or organizational life, but that are not formally part of state law, even though state law has increasingly come to depend on them. They include, for example, the practices of disciplinary bodies, boards, and councils of nonprofit organizations, nongovernment organizations or private businesses or corporations, the tribunals of professional societies or trade associations and labor unions, and the judicial committees of universities and colleges. They also include the peer sanctioning practiced by members of relatively amorphous voluntary associations, such as self-help and mutual aid groups. They now extend to the self-regulatory practices of virtual or online societies of Internet users (Kokswijk, 2010). They also include the private arbitration industry and the restorative justice movement, particularly in schools and in victim-offender and peer-mediation practices. Finally, private legal systems include private policing and private prisons, which exist in a complementary relationship with public systems of criminal justice, expanding its scope but also developing their own semiautonomy (Simmons, 2007). It is in this last form they can be subcontracted by state systems or can be created or contracted by private corporations to exercise social control on behalf of specific economic interests. At the same time, global legal pluralism requires acknowledgment at least of the semiautonomy of nonstate normative orders in creating law: "Globalization reminds us that the state is constrained not only by other states and supranational organizations but also by non-state organizations (e.g., NGOs), communities (e.g., religious groups), and powerful private players (e.g., multinational corporations). All these actors, in one way or other, play roles in the globalizing world that were traditionally reserved for the state. One of these roles might be the role of lawmaker. This idea is articulated in the theory of global legal pluralism, which posits the existence of a plurality of legal orders created both by states and by nonstate communities" (Michaels, 2005: p. 1211).

Mechanisms of Control

Normative orders are the systems of rules and social and behavioral expectations governing a social group or community. Private legal systems bring about conformity to the normative expectations of their collective settings through one of three mechanisms for administering either positive or negative sanctions: (1) coercion, the threat or use of negative sanctions that reduces a person's social standing, either economically, socially, or psychologically; (2) persuasion or ideological control that uses positive inducements to conformity; and (3) suggestion that reminds people of their obligations, responsibilities, and duties in relation to others.

In some private or nonprofit forms, such as the judicial bodies of universities or professional associations, the administration of justice is elaborate and formalized. It may have considerable history and established precedent, and its sanctions may involve suspension, expulsion, fines, loss of privileges, 'probation' counseling, or suspension of licensure. For example, a university judicial body may judge students who violate a campus's drug or alcohol policy to be in serious violation of their agreed-upon rights and responsibilities, and after conducting an investigation and holding a hearing, it may decide to expel the students from the university, to suspend them from all university programs, or to put them on academic probation and require mandatory attendance at drug and alcohol counseling at the students' own expense. In other cases, such as the community or neighborhood association or the local meeting of Weight Watchers, the private administration of justice may be minimal, implicit, or nonexistent, comprising simply of the exercise of sanctions such as ostracism, shaming, gossip, or ridicule. For discussion of related concepts readers should *see*: Legal Pluralism; Laws, Conflict of.

Significance

Nonstate private legal systems and their administration of justice are important because more people are directly governed by and controlled through them on a daily basis than they are by state law, which remains relatively distant from most people's day-to-day life (e.g., in 2009 approximately 3.1% of the US population were subject to legal sanctions of: incarceration in prison or jail, probation, or parole). However, while state law and its various agencies and operations are publicly accountable, private systems of justice often escape accountability. While numerous refinements and endless energy are invested in public systems of law to protect the

individual's right to due process, to guard against excessive or abusive use of state power, and to minimize arbitrary decisions, rarely are such limitations on potential abuse exercised over the day-to-day working of private legal systems and their forms of justice. Insofar as the ideal of reducing arbitrariness is the hallmark of legality, this is 'a variable achievement,' made more so by the existence of a semiautonomous quasi-legal set of private legal systems (Selznick, 1969). A classic illustration of this is found in the numerous private systems that govern employee dismissal that, although subject to state antidiscrimination laws, need not afford employees due process because the existence of 'at will' employment mitigates against 'just cause' employment termination through due process hearings (Henry, 1994). Likewise, as the Internet and virtual world become arenas of increased social presence, they also become spaces subject to new forms of social control through private legal systems in the form of selfregulation (Kokswijk, 2010). Finally, there has been increasing recognition that instead of law being the exclusive province of the state, "Law is created mainly by non-state communities in the peripheries, not by invoking an authority, but by using the code of legal/illegal" (Michaels, 2005: p. 1213; Teubner, 1997).

Private Legal Systems as Types of Law

Consistent with Weberian sociolegal analysis, private legal systems are types of law insofar as they are guaranteed by the probability that physical or psychological coercion to bring about conformity to the order, or to avenge violation, is applied by a staff of people holding themselves especially ready for that purpose. Yet Philip Selznick (1969) argued that if law is seen only in terms of coercive rules, its moral idealism, its progressive striving to reduce arbitrariness in the administration of justice, is overlooked. For Selznick (1969: p. 7) law is 'generic' in that it is "endemic in all institutions that rely for social control on formal authority and rule making." This means that the norm of legality is not confined to the state but occurs in every group that develops rational rules to administer justice. However, while this approach accommodates the concept of private legal systems, it restricts legality to the formal rational system and discounts the informal, which may still deploy legal authority over those subject to its control.

Georges Gurvitch (1895–1965) was a Russian-born French sociologist who specialized in the sociology of knowledge. His sociology of law built on the works of Max Weber, Eugene Ehrlich, and others to demonstrate that law emanated from social relations and ranged in its degree of organization from highly organized and written in advance to spontaneous and unorganized. He also saw law ranging from formal to informal, and thus he founded a sophisticated analysis of legal pluralism. Building on Gurvitch's (1947) more inclusive analysis, all legal systems, whether rational or irrational, can be considered as existing in two planes. On the horizontal plane is the range of legal systems differentiated by the scale of organizational complexity within which they are constituted and through which they primarily function. This horizontal continuum ranges from the global to the group levels, with a variety of intermediate organizational fields including society, community, and organization. Within any organizational field there are many different types of organization and many types of groups. Moreover, each type of organization may exist in a wider context, but each also has a degree of separation from its encompassing social form, based on its own semiautonomous existence in relation to the whole, of which it is a part. Then, for analytical purposes, consider the degree of formality of the legal system in a vertical plane, as ranging from formal to informal. For Gurvitch, the formal is characterized by being organized, written, fixed, and planned in advance, whereas the informal is unorganized, flexible, spontaneous, and intuitive (see **Table 1**). Clearly it is not necessary for informal systems of private law to be spontaneous, since they can be established over time and build up considerable institutional history. **Table 1** also shows the relationship between different levels of social order ranging from macrolevel (global, societal) through mesolevel (organizational, communal) to microlevel (group) (Ellickson, 1991; Kokswijk, 2010), their related private legal systems embodying primary philosophies of sanctioning and the types of sanctions administered by these various state and private legal systems.

This schema reveals that informal social control is not the exclusive control mechanism of the group. While developed organizational forms, such as the state, appear to be dominated by formal social control (such as criminal law), within law's organization, and reflective of its constitutive groups, there exist numerous levels of informal private justice working within, and sometimes against, the explicit principles of its wider system. Examples would include police discretion and plea bargaining. Conversely, formal social control is not the exclusive mechanism of the state or formal organization. While levels of organizational development often preclude small groups from having highly formal private legal systems, they sometimes develop rudimentary formal controls (Henry, 1983). Thus, each type of collectivity has depths of formal and informal operation. For example, an organization such as a manufacturing company is governed by the regulations of formal state law, but also has its own nonstate private legal system for administering internal discipline. Insofar as this relies on written formal rules, procedures, and penalties, it would be considered a type of formal private law. However, within the same company there may also be a range of groups of employees of different professions, skills or trades, different unions, work teams and social cliques, and so forth. Each may exercise its own private collective control over its members. Arguably the most effective system of control stems from multiple kinds of social control, where each level of operation (from formal to informal) complements the others in bringing about conformity. Indeed, rather than using only one type of legal system, a collectivity typically uses a combination of types and levels of legal systems simultaneously (Henry, 1983).

Models of Justice Operative within Private Legal Systems

In addition to the types of private justice systems and their levels of organization and formality, it has been increasingly

Table 1 Law, private legal systems, and levels of formality

	Global	*Societal*	*Communal*	*Organizational*	*Group*
Type of legal system	International law	Different national state legal systems; different kinds of state law	Private legal systems; norms and controls established by communities	Private legal systems; conflict rules and controls established by organizations	Private group control
Formal control agent (third party control)	International law on human rights; law on genocide. Treaty	State law, US statutory law; US common law, e.g., criminal law	Neighborhood rules and ordinances affecting housing, noise, pollution	Rules on gross misconduct, cheating, employee theft, absenteeism, property damage	Social norms of group governing behavior of its members
Informal control (examples)	Norms governing smuggling	Police discretion; plea bargaining	Neighborhood norms on street crime gay rights racial tolerance	Social norms of indulgence; levels of acceptable violation	Norms of employee work team; norms of street drug gangs
Primary philosophies of sanctioning	Conciliation, Incapacitation	Incapacitation, deterrence, retribution (just desserts), rehabilitation, treatment	General deterrence, reparation, restitution, restoration	Incapacitation, deterrence	Just desserts, incapacitation
Sanction/response	Cooperative enforcement: exclusion, economic sanctions	State enforcement/ sanctions: death, prison, probation, fines	Communal discipline, sanctions: fines, public shaming	Organizational discipline, sanctions: expulsion, demotion, fines	Group self-help (collective aggression, shaming, negotiation)

recognized that systems of justice operate according to different sets of principles, driven by different ideologies, to achieve different justice outcomes. One of the first to address this was Herbert Packer (1968), who distinguished between a crime control model of justice (concerned with protecting the public from victimization by suppressing rule violators) and a due process model of justice (concerned with protecting individual rights). Since Packer, Michael King (1981) introduced an additional four models of criminal justice and Peter Kraska (2004) added several others, which he describes as "criminal justice orientations." In addition to (1) the crime control and (2) the due process models are (3) the medical/social pathology model, in which the administration of justice diagnoses the cause or multiple causes that led to a rule violation, whether individual or social, and prescribe a treatment solution; (4) the bureaucratic/administrative model, which is aimed at preventing rule violation through surveillance, behavior analysis, early detection and the efficient and cost-effective management of risk through the management of offenders; (5) the status passage/constructionist model, which sees justice systems as institutions designed to transform an accused's public identity or social status through impression management using denunciation, public degradation and shaming; and (6) the power-domination model, wherein justice is seen as a myth promoted to provide legitimation for dominant economic and political elite, and to deflect attention from issues of class conflict. Importantly, these same sets of models of justice are manifest as competing interpretations for the operation of different private legal systems.

The Interrelationship between Private Legal Systems and State Law

Although private legal systems, by definition, exist outside the state, they have a semiautonomous relationship with state law (Moore, 1978); they are subject to its broad framework yet within these constraints are free to create and enforce their own rules. That they exercise control consistent with the broad requirements of the state means that they can reaffirm state law, and so function as state surrogates. As such, private legal systems are a subset of the more generic notion of social control, which refers to broad societal, community, or organizational level processes such as socialization, education, social welfare, and so forth that bring about conformity to societal social order. In this broader sense, private legal systems are a reaffirmation of the relations of order, part of the ongoing construction of a collectivity's social reality. In this sense private legal systems merge with ongoing daily interaction as an integral part of the constitution of group process to develop the appearance of stability, predictability, and order. This point is at the heart of Ehrlich's (1913) observation that 'living law' is that which dominates life itself, even though it has not been posited in legal propositions.

Legal pluralism (Pospisil, 1971), founded in Ehrlich's (1913) concept of 'living law,' is the idea that there exist multiple sources of law and that the state need not be the most important source of law. An untold number of laws are created by groups, communities, organizations, societies, and global entities. So the semiautonomy of private legal systems allows them to sometimes assume a contradictory and subversive role in relation to the state. This variety of legal systems may exist in

harmony or conflict, reinforcing or undermining the social control of other private legal systems and even the formal law of the state, such that they may condone and even approve of behaviors that the overarching state system would condemn. For example, consider the toleration of some level of pilfering (theft) by employees in a food manufacturing company or expense overclaims by executives at an energy supplier, both of which are seen as 'perks' of the job.

Any consideration of nonstate private legal systems, therefore, must also involve an examination of the theory of 'legal pluralism' and must address the varying interrelationships between the different types of social control and between their varying levels of formality. Further, insofar as private legal systems exist in an ambivalent relationship to state law, sometimes supporting it, at other times contradicting it (and maybe even undermining it), consideration must also be given to how these forms relate to, and even constitute, the totality of control in a society.

The Limits of Traditional Legal Pluralism

Recognition of the significance of nonstate forms of law has led sociolegal theorists to revise their thinking about the law and society relationship. The suggestion that state law and its institutions for administering justice are but one form of social control has been a core feature of legal pluralism from the foundations of theorizing about law. However, the legal pluralist tradition has focused on the character, source, and hierarchical place of nonstate normative orders in relation to state law. Until the mid-1970s, much of their argument explored the relationship between indigenous tribal custom and European colonial law (Pospisil, 1971). This tradition saw indigenous orders as autonomous, independent, yet subordinate to colonial law. By the mid-1980s, the emphasis had shifted to noncolonial societies but retained its concern with the hierarchical power relations between dominant and subordinate law forms. Thus, many critical theorists take 'informal justice institutions' as the subordinate normative order and examine how they serve the ideological function of blurring the power of the state so that it appears to be a benign part of the social fabric. They have shown that this ideological subordination is accomplished by the co-optation and exploitation of a human desire for informal, localized, community justice, and that the episodic tendency toward an 'informal,' decentralized state control serves a dual legitimating and net-widening control function for the state (Cohen, 1985). However, these criticisms seem more applicable to the growth of state-sponsored dispute settlement institutions, community policing, and restorative justice (Simmons, 2007) than to explaining the state's interrelationship with other, established, private legal systems, which have largely been ignored. During the late-1980s, attention turned to the contribution that the social relations of these nonstate systems of law make to the constitution of state law.

Private Legal Systems and Constitutive Legal Theory

Critical contributions to the legal pluralist debate argued that recognition of the interpenetrating role of nonstate normative orders requires a new conception of law in society. This has been variously referred to as the 'relational theory of law,' 'constitutive theory,' 'postmodernist' or 'deconstructionist' theory, and 'discursive theory' of law.

Constitutive theory (Moore, 1978; Fitzpatrick, 1984; Henry and Milovanovic, 1996) explores how law or legal relations are generated within nonlegal social relations, and how legal relations and social relations are interrelated. Thus, instead of assuming that state law is the hub of social control whose spokes radiate as unidirectional pathways of influence to other social and normative orders, constitutive theory examines forms and mechanisms whereby legal relations are penetrated by extralegal social relations. A constitutive approach examines both the presence and the source of other social forms (see Hunt, 1993).

It is not, then, just that law is created by classes or interest groups, whether instrumentally or symbolically, to maintain or increase their power. Rather, some of the relations of these groups, particularly their rules and procedures are, and indeed become, the relations of law, as relations of law similarly become social relations. Thus "what constitutes 'the law' in any specific context will depend upon which legal networks … intersect in that context, how these orders are mobilized, and how they interact" (O'Malley, 1991: p. 172).

Peter Fitzpatrick (1984, 1992) explored the notion of interrelationships between law and the larger social matrix within which it is set, developing the concept of 'integral plurality.' He argues that the reason state law is, in part, shaped by the plurality of other social forms, while these forms are simultaneously being shaped by it, is because 'elements of law are elements of other forms and vice versa' (Fitzpatrick, 1984). Fitzpatrick (1988, 1992) also highlights that the interrelations between state law and nonstate normative orders form new entities, as a common discourse and set of practices is worked out between participating arenas of power. He says that in the synopsis of existing rules and practices, the participating networks retain their own relative autonomy, but corresponding structures are formed that merge selected elements of the component networks into an emergent whole that becomes the new law. Pat O'Malley describes these attempts at legal synthesis as 'synoptic projects' (O'Malley, 1991: pp. 172–173), which are most likely to occur where changing conditions, such as the emergence of obstacles to the continued effectiveness of existing arrangements, make their continued operation problematic. A similar misreading occurs, says Teubner (1992, 1997), when the organization reincorporates legal rules developed and refined in disciplinary proceedings and makes use of them to restructure its organizational decision making (*see* Law and Everyday Life; Legal Culture and Legal Consciousness).

Although much has been written about how law shapes social forms and about how interests shape law, until the mid-1980s little research had been done on the way that semi-autonomous nonstate private legal systems interrelate with and constitute state law while also being constituted by it. Neglected is the extent to which their agency contributions, as carriers for the private normative systems, transform that law in the process of constituting and reflecting it. A critical research question becomes how private legal systems, such as professional self-regulatory bodies, workplace disciplinary practices, and the variety of other nonstate forms of social control, are constitutive of state law.

In considering the theoretical significance of constitutive legal theory to understanding the relationship between law and private legal systems that govern normative orders, it is worth stating the obvious. Law, and particularly statutory law, is often a deliberate effort to incorporate and synthesize custom and practice, norms, and rules prevailing in other settings into a general legal framework. Legal scholars project that this will likely continue: "By their very nature, states likely will continue what they do now: incorporate, defer to, or delegate normative orders, but not accept them as law. This may seem narrow-minded or blind. Yet this blindness is the consequence of the blind spot in the conception of the state" (Michaels, 2005: p. 1249). Michaels argues that the self-identity of the state is such that it has to transcend the normative authority of these private legal systems in order to maintain its own authority and superiority. The point, however, is that by recognizing the relative autonomy of these nonstate systems of private law, and simultaneously recognizing their ability to interrelate with state law, constitutive theory allows for the possibility that law creation will have an unpredictable outcome. In this process the eventual product is always different from what its drafters intended. This is so, precisely because of the relative autonomy of nonstate private legal systems.

Scholarly writings between 1980 and 2000 often took a somewhat romantic view of the role of nonstate law, one that overstated these orders' semiautonomy. Sharafi (2008) points out that during this period nonstate law came to be seen as a form of 'popular justice' or 'people's law' that was present in all societies, not just colonial or postcolonial ones, where it was first studied. However, recent scholarship has shifted toward reengaging the colonial perspective through a legal pluralism that recognizes cultural diversity as importing elements of minority or other cultures' normative systems into the Western legal model through such mechanisms as 'the cultural defense.' This perspective acknowledges that different cultural values and traditions govern normative behavior and should reduce the consequences for law violation in state courts. Others, notably feminist legal scholars, have criticized the cultural defense version of legal pluralism, arguing that it decriminalizes violence against minority women in the name of acknowledging the validity of nonstate law.

See also: Justice: Social Psychological Perspectives; Legal Systems, Classification of; Lex Mercatoria; Social Justice and Social Inequalities.

Bibliography

Cohen, S., 1985. Visions of Social Control. Polity Press, Oxford, UK.
Ehrlich, E., 1913. Fundamental Principles of the Sociology of Law. Harvard University Press, Cambridge, MA.
Ellickson, R.C., 1991. Order without Law: How Neighbors Settle Disputes. Harvard University Press, Cambridge, MA.
Fitzpatrick, P., 1984. Law and societies. Osgoode Hall Law Journal 22, 115–138.
Fitzpatrick, P., 1988. The rise and rise of informalism. In: Matthews, R. (Ed.), Informal Justice? Sage, London.
Fitzpatrick, P., 1992. The Mythology of Modern Law. Routledge, London.
Gurvitch, G., 1947. The Sociology of Law. Routledge and Kegan Paul, London.
Henry, S., 1983. Private Justice: Toward and Integrated Sociology of Law. Routledge and Kegan Paul, London.
Henry, S., 1994. Employee dismissal: justice at work. The Annals of the American Academy of Political and Social Sciences 536.
Henry, S., Milovanovic, D., 1996. Constitutive Criminology: Beyond Postmodernism. Sage, London.
Hunt, A., 1993. Explorations in Law and Society: Toward a Constitutive Theory of Law. Routledge, London.
King, M., 1981. The Framework of Criminal Justice. Croom Helm, London.
Kokswijk, J., 2010. Social control in online society – advantages of self-regulation on the Internet, 2010 International Conference on Cyberworlds, pp. 239–246.
Kraska, P., 2004. Theorizing Criminal Justice: Eight Essential Orientations. Waveland Press, Long Grove, IL.
Michaels, R., 2005. The re-statement of non-state law: the state, choice of law, and the challenge from global legal pluralism. The Wayne Law Review 51, 1209–1259.
Moore, S.F., 1978. Law as Process. Routledge and Kegan Paul, London.
O'Malley, P., 1991. Legal networks and domestic security. Law, Politics and Society 111, 170–190.
Packer, H.L., 1968. Two models of the criminal justice process. In: Cole, G.F., Gertz, M.G., Bunger, A. (Eds.), The Criminal Justice System: Politics and Policies, ninth ed. Wadsworth, Belmont, CA.
Pospisil, L., 1971. Anthropology of Law. Harper and Row, New York.
Selznick, P., 1969. Law, Society and Industrial Justice. Russell Sage Foundation, New York.
Sharafi, M., 2008. Justice in many rooms since Galanter: de-romanticizing legal pluralism through the cultural defence. Law and Contemporary Problems 71, 139–146.
Simmons, V., 2007. Private criminal justice. Wake Forest Law Review 42, 911–990.
Teubner, G., 1992. The two faces of Janus: rethinking legal pluralism. Cardozo Law Review 13, 1443–1462.
Teubner, G., 1997. Breaking frames: the global interplay of legal and social systems. American Journal of Comparative Law 45, 149–169.

Legal Traditions of Southeast Asia

Andrew J Harding, National University of Singapore, Singapore

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Just as pluralism is the mark of Southeast Asia societies, legal pluralism defines much of Southeast Asia's laws and legal traditions. Indigenous legal traditions were based on customs and religion, and government conformed to ideas of monarchy. Colonialism brought in European law, which still forms the framework of law in the region, but it conflicted with local legal cultures and traditions. Nationalism brought in authoritarian laws rather than a reversion to indigenous legal traditions. Development has, however, led to some convergence of public and private laws.

This article endeavors to describe the legal traditions of Southeast Asia. There are obvious difficulties in identifying the meaning of 'legal tradition' and even of 'Southeast Asia,' as well as in justifying their conjunction in this article. Generalizations of any kind about law in such a diverse region as Southeast Asia, which is one of most diverse regions of the world, are in fact hard to come by, unless legal pluralism itself (evident in many parts of the world and even arguably in all legal systems) may count as a defining set of norms. M.B. Hooker's groundbreaking work on legal pluralism significantly takes Southeast Asia as its field of inquiry (Hooker, 1975). The same author's monumental two-volume *Laws of South East Asia* (Hooker, 1986, 1988; also Hooker, 1978) is similarly imbued with legal pluralism and provides the scholarly bedrock for this subject. The concept of legal pluralism in the sense of formal or official legal pluralism (see below) defines Southeast Asian legal systems, not in the sense of being a theory consistently adopted across the region as a matter of public policy or legal theory, but in the sense of being the medium in which law necessarily operates at all levels given the diversity of the societies in question.

Southeast Asia

The term 'Southeast Asia' itself was invented by the British navy as a means of dividing command during World War II. There was, and still is, little to define the region in terms of law, and even in terms of nonlegal factors, such as ethnicity, religion, political system, or trajectory of economic development (Cambridge History of Southeast Asia, 1992). Nowadays one can at least point to a binding fact of geopolitics, namely the existence since 1967 of the Association of South East Asian Nations (ASEAN), which from the late 1990s has included the 10 southern Asian continental states that do not fall under 'South Asia' or 'Greater China.' This list includes Cambodia, Laos, Myanmar, Vietnam, Indonesia, Malaysia, Philippines, Singapore, Thailand, and Brunei. (Some might also include Timor Leste, which obtained independence from Indonesia in 2002, and which is likely to join ASEAN.) These countries perceive that they share economic and strategic interests, and together they represent nearly one twelfth of the world's population. The existence of ASEAN has not thus far created any discernible tendency for the region's legal systems to converge (ASEAN Law Association, 1995), although the increase in intraregional trade resulting from ASEAN cooperation creates the possibility of some harmonization, at least in the field of economic law; and the creation of a human rights mechanism following the adoption of the ASEAN Charter in 2008 (Durbach et al., 2009; Tan Hsien-li, 2011) may ultimately produce some commonality in approaches in that area too.

Given that ASEAN is a creation of the last half century we struggle to find commonalities in remoter history. Empires and kingdoms came and went – the Sri Vijaya Empire in the eighth to twelfth centuries; the Khmer Empire in the ninth to fifteenth centuries; the Majapahit empire in the thirteenth to fourteenth centuries; the Malacca Empire in the fifteenth century. Individual kingdoms waxed and waned in power – Vietnam, Siam, and Burma, for example – but no single power has ever held sway over all or even most of Southeast Asia or left a distinctive cultural or legal mark on the region (Hall, 1981). Even during the years of colonial rule the region was divided between Britain, France, the Netherlands, Spain, and Portugal, with Siam (called Thailand from 1939) remaining independent. The experience of colonialism was common to the region but its effects tended to divide rather than unite it. This is especially true in the field of legal traditions (Hooker, 1988).

Legal Pluralism and Legal Traditions

Southeast Asia is not therefore a region that associates itself naturally with any particular legal tradition or traditions, unlike, say, the Islamic world or continental Europe. Indeed, apart from the customary laws of the region's many indigenous peoples, all of Southeast Asia's legal traditions are imported ones, having their origins outside the region; there are in fact very few of the world's legal traditions that are not represented in some way in Southeast Asia. It is a region in which there will be found a bewildering plurality of legal traditions in confusing relationships with each other and with state law. Thus within the geographical area described by each of these states lie complex webs of plural ordering around ethnicity and religion (Harding, 2001).

We are forced to confront squarely here the term 'legal tradition.' In Southeast Asia legal traditions in practice exist within or outside the formal legal system. To a large extent the

region's states practice formal legal pluralism in that the state recognizes forms of law other than the law created by the state itself or continued in effect by the state, for example, when the colonial period came to an end (Hooker, 1975). This might seem to be a contradiction in terms, but for the fact that these other forms of legal ordering predate modern formal state law and have been incorporated or corralled within it – not always, but surprisingly often. Beyond this it is a problem for both 'legal tradition' and 'legal pluralism' that there are many forms of legal ordering that exist independently of formal state law, and there is no obvious criterion whereby one would include or exclude them from consideration as 'Southeast Asian legal traditions.' This in turns raises the perennial problem of how to define law itself. A state-centered definition would draw a line at formal legal pluralism; it is often argued, however, that in Southeast Asia, despite its extensive formal legal pluralism, this is arbitrary and misleading (von Benda-Beckmann, 2002). A community-centered definition of law would include customs and practices even where these are not endorsed by formal state law (Griffiths, 1986); here it may be argued that 'law' loses any clear meaning, shading into potentially any form of social ordering or social practice (Tamanaha, 1993). The term 'legal tradition' itself suggests a body of knowledge handed down between generations rather than simply found in a library or some authoritative governmental source; this does not imply a notion of simply how people behave, but a notion of behavior according to some received normative framework. While an expansive view of 'legal tradition' will be attractive to Southeast Asia experts, it creates the enormous empirical difficulty of verifying which legal traditions may be included. For present and purely pragmatic purposes we will take state law to represent 'law,' even though the official recognition or nonrecognition of some legal traditions in Southeast Asia is in many respects quite arbitrary or controversial (see Legal Pluralism for further discussion of the ambiguity of this term). In adopting this pragmatic definition we have to recognize nevertheless that there is an important other side to law that it would be foolhardy to ignore if one wished either to understand, to apply or to reform state law in Southeast Asian countries; or indeed to take the variety of *unofficial* customary law seriously.

Precolonial Law

Before the intervention of colonial powers from the fall of Malacca to the Portuguese in 1514, law in Southeast Asia was based either on customs, on religion, or on codes, and often on all three. Buddhist legal traditions prevailed in continental Southeast Asia (Huxley, 1996), and were used along with royal decrees and village customary dispute settlement systems. Hinduism had influence across much of archipelagic Southeast Asia, replacing Buddhism to some extent in Indonesia, and even to this day in Bali. Islam began to be influential from the early fifteenth century, especially in the Malay world of peninsular Malaya, Indonesia, and Mindanao in the southern Philippines. Muslim communities are, however, spread across all of Southeast Asia, and Islamic legal traditions are formally recognized not just in Malaysia and Indonesia, but also in Myanmar, Philippines, Singapore, and Thailand (Hooker, 1986).

Malay customary law or *adat* is by far the most typical, distinctive, and widespread form of customary law in Southeast Asia. It mingled with Islamic law in what Hooker calls the Malay legal world of Indonesia and Malaysia, and has survived both colonial rule and postcolonial nationalist modernization. *Adat* varied somewhat from place to place, but conforms to conventional notions of customary law: it was handed down as an oral tradition; often took the form of easily recalled proverbs, homilies, or brief lines of poetry, often containing analogies to the natural world; and applied generally at the village level. It dealt not just with social and moral issues (family and criminal law in modern parlance) but also with property, agriculture, the environment, and even constitutional issues. Its typical mode of enforcement was via criminal or social sanctions (Hooker, 1970). The famous Dutch legal scholar Cornelis Van Vollenhoven argued strongly and successfully in favor of allowing the Indonesians to be governed by their own *adat* instead of Dutch law (Fasseur, 1992). He founded the *adatrecht* school at Leiden University, in which Dutch and Indonesian scholars mapped out 19 regional varieties of *adat* in the Dutch East Indies (Holleman, 1981; Burns, 2004). A fundamental distinction within these varieties was that between *adat temenggong* and *adat perpatih*; the latter, applicable mainly in the Minangkabau areas of West Sumatra (Indonesia) and Negri Sembilan (Malaya) excited much interest with its matrilineal system of succession to property and its female-centered laws. Nowadays *adat* still prevails as an aspect of law applicable in the religious courts of Malaysia, Indonesia, and the Philippines, and as the main form of law applicable to the 'native' or indigenous peoples of Borneo. The Malaysian state of Sarawak, for example, applies an almost complete system of formal legal pluralism by applying *adat* in the Native Courts, *shari'a* in the Syariah Courts, and common law in the civil courts, which also apply the customary law of nonnative groups, such as the Chinese (Hooker, 1980).

It would, however, be erroneous to think that codes and statutes played no part in precolonial law. There are many examples in Southeast Asia of attempts to codify the law. These should be distinguished from codes in the European sense of attempts to reform as well as to consolidate the law in a given area. Traditional law in Southeast Asia did not distinguish fields of law in same way that European law consistently did: notably they did not distinguish criminal, public, family, and property law. Attempts at codification were designed to consolidate the entire law or at least large portions of it. Preeminent examples are the *Undang–Undang Melaka* and the *Undang–Undang Laut Melaka* (Laws of Malacca, fifteenth century) (Liaw, 1976); Siam's 1805 Law of the Three Seals (Harding, 2008); and the codes of the Le dynasty in Vietnam (fifteenth to eighteenth centuries). The Law of the Three Seals is of particular interest in terms of codification. A response to the perceived decline in legal veracity resulting from the destruction of Siam's legal records in the sack of Ayutthaya by the Burmese in 1767, it attempted to restate the entire law of Siam in terms of the Buddhist Dhammasat along with royal decrees of general application. It was not initially published, indicating its function as a reliable guide for officials and judges rather

than a statement of the rights of Siamese subjects (Hooker, 1986; Vickery, 1996).

Colonial Law

The advent of European colonialism from the early sixteenth century onward, and especially in the nineteenth century, resulted in the gradual imposition of European laws as the general law in all of Southeast Asia. In some cases such as Burma, Tonkin (Vietnam), Philippines, and Cambodia, the colonial power actually overthrew or displaced the traditional monarchies and the legal traditions that were attached to them. Even so, these legal traditions were retained to the extent that they represented exceptions to the general law, especially for personal law purposes. The case of Burma was particularly unfortunate, as Burma had in many ways the most advanced legal system in Southeast Asia prior to 1885 (Huxley, 2001) when the monarchy was overthrown by the British in the Third Anglo–Burmese War. Elsewhere, such as in British Malaya, parts of the Dutch East Indies, and the French Vietnamese protectorate of Annam where the Nguyen dynasty reigned until 1945, traditional rulers were retained in a system of indirect rule, and European laws were only introduced piecemeal (Ta Van Tai, 2004–2005). As the European powers moved from establishing trading posts to exerting political control over hinterlands during the mid- to late-nineteenth century, they increasingly imposed their own laws. There were, however, variations in terms of the geographical spread of these laws, the people to whom they applied, and the legal subjects they covered. In general personal laws, being based on the fundamental social values or religions of the Asian peoples under colonial subjection, remained untouched, although legal structures under colonial control (courts, judges, legislation) were put in place to administer them. This of course tended to alter the effect and the dynamism of the local legal traditions. Property and commercial law, being essential to the economic purposes of colonial governments, and criminal laws, designed to maintain social order and the power of the colonial government, generally and increasingly conformed to the law of the colonial power. Migrant peoples, such as the Chinese and those of South Asian heritage, were generally allowed to have their own law, at least initially; eventually even these were gradually absorbed under the aegis of the colonial law (Lev, 2001a).

An interestingly contrasted approach is revealed as between British and Dutch legal policy regarding legal pluralism. The British introduced English law as the general law in the Straits Settlements of Penang, Malacca, and Singapore (Phang, 1990), but they did not enforce it in a monolithic way. Instead, they modified it in its application to the subject peoples, sometimes even recognizing full exceptions (such as Islamic personal law for Muslims), on the ground that to do otherwise would be oppressive, as well as courting rebellion. The Dutch in the Dutch East Indies, on the other hand, adopted a policy of recognizing 'law populations'; under this theory each people having its own law would be dealt with according to that law (Lukito, 2013). Over time, the tendency of the English common law and pragmatic legislative measures to make exceptions, and the tendency of the Dutch system to reduce the exceptions by common legislation, narrowed the difference, in practical application at least, between these starkly different theories. Only in relation to the Dutch East Indies was there an intense debate as to the correct ethical position for a colonial power (Fasseur, 1992). In the British-held territories the Indian precedent was followed; in other words, in due course the codified common law of British India was adopted (the Penal Code, the Criminal and Civil Procedure Codes, the Contracts Act, the Evidence Act, and so forth) (Tan Poh-ling, 1997).

Thus by the late nineteenth or early twentieth century European laws became the rule rather than the exception, governing most areas of law, or at least framing their administration. The Dutch maintained a system of religious courts (Lev, 1972), while the British had, by treaties, reserved to the Rulers of the Malay states jurisdiction over matters affecting Islam and *adat*. As a result of these colonial legal developments, without exception the Southeast Asian states even now base their legal system on a European framework (Harding, 2001) conforming to a variety of the civil law or else of the English common law: French law in Vietnam, Cambodia and Laos; Dutch law in Indonesia; Spanish law in the Philippines; Portuguese law in Timor Leste; the English common law in Singapore, Malaysia, Brunei, and Myanmar. The Philippines, having been under American rule for the first half of the twentieth century, has a hybrid system, with some areas (such as commercial and constitutional law) resembling the law of the United States. Malaysia, Indonesia, and Brunei maintain a bifurcated legal system as between the ordinary courts applying the common law, and the religious courts applying the *shari'a* law to Muslims in personal law matters. In Singapore the Syariah Court remains a subordinate court within the main legal architecture. In Malaysia a constitutional amendment in 1988 separated the Syariah Courts from the 'civil' or ordinary courts by providing that the latter could not exercise the jurisdiction of the former (Federal Constitution, Article 121(1A): Harding, 2012; Moustafa, 2013).

An exception here in several respects is Siam, which, although never colonized, came under colonial influence if only by virtue of being under the whip of extraterritoriality treaties concluded with Western powers and Japan in the latter half of the nineteenth century (Hooker, 1986), and the witnessing of neighboring Burma's fate. These treaties provided for citizens of the country in question to be treated under their own law while on Siamese territory, and expressly linked the rescinding of this arrangement to the enactment of 'modern codes of law' in Siam. King Chulalongkorn the Great (1868–1910) and his immediate successors achieved a transformation of Siamese law as complete as that imposed by any colonial power elsewhere in Asia. Over a half century from about 1885 to 1935, Siam adopted codes based eclectically on civilian systems with some influence from the previous law based on the Buddhist *dhammasat* and the common law. By the time the Civil and Commercial Code was completed in 1935, the extraterritoriality treaties had as a result been entirely revoked. Siam's model here was to follow Japan's voluntary espousal of the civil law rather than India or Vietnam's enslavement to a single foreign legal system. Unlike Japan, Siam was mainly influenced by French and Belgian, rather than German, jurists, having earlier flirted with the adoption of the common law, which proved too

problematic in its practical application to be an appropriate model. Following this complete transformation of Siam's legal codes, the Law of the Three Seals was consigned to legal history (Harding, 2008).

Postcolonial Law in the Period of Nationalism

The struggle against colonialism took different forms and affected the vectors of national legal systems across Southeast Asia. In the common law countries (Burma, Malaya, Singapore, Brunei), as well as in the Philippines, following India's independence in 1947, independence was gained without much difficulty, and the common law was assumed to represent an inevitable form of legal continuity. In Indonesia and Indo-China, however, independence from the Netherlands and France, respectively, was attained only after fierce, and in Vietnam's case, prolonged, warfare. The continuation of colonial laws in these latter countries represented a temporary expedient rather than a solution. Vietnam, initially divided (1954–76) between the capitalist South and the communist North, was torn apart by a devastating proxy war, the North becoming ultimately victorious by 1975 and bringing all of Vietnam, and Laos too, under socialist law based on that of the Soviet Union (Gillespie and Nicholson, 2005). Cambodia under the Khmer Rouge (1975–79) saw its laws and legal professions, indeed even legality as such, tragically and completely obliterated by genocide, and the road back from the ground zero of 1979 (legally and in general) has been extremely difficult.

Under the influence of the legal scholar Supomo, Indonesia adopted a constitution in 1945 (that is still in force) that embodied an integralist view of the state that was hostile to individual rights and the rule of law; it also rejected as highly divisive the notion that the state should be based on Islam (Lev, 2001b). Nationalism did not, however, succeed in bringing Indonesians under anything resembling the revolutionary law (*hukum revolusi*) desired by President Sukarno (1945–66). Rather it secured the temporary eclipse (1963–66) of the Dutch civil code, which during the New Order (*Orde Baru*) period of President Suharto (1966–98) ultimately survived only to be hollowed out by successive new laws on, for example, land and marriage (Agrarian Law, 1960; Marriage Law, 1974) (Lev, 2001c). Personal laws are divided as between those for Muslims, based on Islamic law and *adat*, and those for non-Muslims. The 1998 economic crisis, which affected all of Southeast Asia, brought an end to the authoritarianism of Suharto's New Order and the commencement of a new period of reform (*reformasi*), marked by extensive legal reforms and constitutional development led by four major amendments (1999–2002) to the 1945 Constitution that altered its character profoundly (Butt and Lindsey, 2012). The *reformasi* period has transformed Indonesian law once again as democracy replaced authoritarian government, and the rule of law (*negara hukum*) replaced nationalist laws such as the Judiciary Law 1970, which brought the judiciary under government control (Lev, 2001b).

A slightly similar pattern may be seen in the Philippines, where postwar independence ultimately led to the dictatorship of President Ferdinand Marcos and rule by emergency decree under an increasingly manipulated constitution during 1972–86. The ousting of Marcos by 'people power' in 1986 led to a new constitution in 1987, securing a vibrant if sometimes violent and unstable democracy. Some elements of Philippines law, such as the absence of divorce law and antiabortion law, can be traced to its largely Catholic population. The old Spanish Civil Code is still in force, but the American common law is influential in some areas, notably constitutional and commercial law (Santos, 2000).

An extreme case of this now-familiar legal trajectory is Burma, which became independent in 1947. As was the case with other Southeast Asian states, Burma did not attempt to reintroduce the traditional legal system, just as it did not reintroduce the monarchy. Indeed its laws were simply in effect the codified laws of British India, of which it had been a major province between 1886 and 1937. In 1962 the coup of General Ne Win resulted in a gradual demolition of the system left by the British and the 1948 Constitution. Judicial independence and the doctrine of precedent were done away with in favor of government policy as a prevailing norm under the 'Burmese path to socialism' (Myint Zan, 2000). Legislation was introduced as Revolutionary Council decrees. The Anglo-Indian codes remained but were not always used in practice. In 2011, after half a century of military rule, Myanmar (as it has been called since 1989) began to open to the outside and reform its legal system, posing anew the issue of how to reestablish legality and constitutionalism in a country plagued by interethnic wars and intercommunal strife. The persistence and progress of these reforms remain in doubt, as does the nature of Burmese legal traditions (Crouch and Lindsey, 2014)

Asian Laws, Asian Values

Underlying many of these developments lies the issue of the fundamental nature of law in Southeast Asia. To what extent is it or can it be indigenous, or conform to patterns markedly different from those of its former colonial powers? To what extent also does it or can it embrace international norms of the rule of law, constitutionalism, human rights, and property? Is it in fact resistant to the foreign legal transplants to which it has been repeatedly subjected?

The diversity of the region makes it difficult even to explore general answers to these questions. As we have seen, the region has not used the independence of the postwar era to return to its legal roots in precolonial law. Rather it has attempted to build national legal traditions on a base of colonial law – we have seen in Indonesia's case how contradictory this can be. Whatever the merits or otherwise of traditional law, no sustained argument has succeeded in restoring even some limited elements of precolonial law. Indeed in many areas, even in personal law, the law has moved further away from precolonial law and from legal pluralism, for example, with the establishment of monogamy, consent, female equality, and divorce grounds in marriage law. *Adat*, it is true, has staged something of a comeback in Indonesia, and Islamic revival has prompted a demand across parts of the region for the correction of

the displacement of Islamic law by the colonial powers. In neither Malaysia nor Indonesia, however, has Islamic law been embraced as the general law, nor has the state taken on a genuinely and expressly Islamic character. In Vietnam socialist law has been somewhat eclipsed by the recognition of private property and liberalization in many areas of law (Sidel, 2010; Gillespie and Nicholson, 2005).

During the 1990s attention was given to the notion of 'Asian values' with specific reference to the autonomy of Asian states to develop their own laws without necessarily conforming to Western laws or international norms (Bell, 2000). Part of the argument about Asian values was that Asian societies embraced a form of communalism or family values that was seen as hostile to what was perceived as extreme Western liberalism and individualism that tended to dissolve social bonds. This argument was mainly advanced by political leaders in Singapore, Malaysia, and Indonesia, but had resonances in other Southeast Asian countries. This development was a direct response to advances in international human rights, and as expressed in the Bangkok Declaration on Human Rights 1993 (Thio, 1999). The Asian values argument did not reject human rights as such (indeed all Southeast constitutions acknowledge human rights in some form), and several Southeast Asian states (Indonesia, Malaysia, Myanmar, Philippines, and Thailand) even have a national human rights institution (e.g., Whiting, 2003). However, proponents of the Asian values argument refused to accept that all human rights as defined by Western powers in the evolving international system were totally applicable in Asian societies; or that human rights were indivisible or to be interpreted according to Western normative values. The main problem with this thesis is that, given the diversity of Asian value systems, it has proved in practice difficult to find common ground between these various traditions or between different types of legal consciousness. It also appeared that the argument was advanced in a self-serving manner by Asian leaders anxious to buttress their power by appeal to a common authoritarian ideology, which sits uncomfortably with advancing democratization. Nonetheless, one can observe that, while in practice Southeast Asian states do on the whole assert their autonomy with regard to human rights (for example, with regard to laws on national security), it is also true that human rights are being increasingly recognized either in constitutional law and in actual practice (for example, with regard to religious freedom, ethnic minorities, gender, and gay rights).

The Future of Law in the Region

From this brief account of Southeast Asia's legal traditions we can see that "it is appropriate to think of law in South East Asia geologically, as a series of layers each of which overlays the previous layers without actually replacing them, so that in places, due to tectonic shifts, the lower layers are still visible, although not perfectly distinguishable from each other" (Harding, 2001: p. 205). The issue for Southeast Asians is whether this accommodation of legal pluralism is a fundamental problem for the unification of legal traditions under one system based on the rule of law and legal equality, or whether it is a pragmatic or even principled balancing of legal traditions under a system that recognizes the differences between them. The surprising fact is that the legal systems of the region have in practice achieved the degree of syncretism that they have. Legal pluralism continues to present difficulties (for example, citizenship and customary law in Thailand, mixed marriages in Indonesia, and single-parent conversion of children in Malaysia). However, the larger issue is rather the credibility and performance of legal institutions per se: the independence of the judiciary; the rule of law in one-party or dominant-party states; the observance of human rights and civil liberties; corruption in the judiciary, the police, the bureaucracy, and the electoral system. Across Southeast Asia in the early twenty-first century it is therefore increasingly through *constitutional* debates that the nature of law and the relevance of legal traditions are being addressed. While it appeared, in the midst of the Asian values debate, that Southeast Asian states were perhaps largely settling into semiauthoritarian systems accommodating legal pluralism, events since then have suggested otherwise. The rapid growth of civil society; people power in the Philippines from 1986; *reformasi* in Indonesia from 1998; the emergence of a two-party system and legal reforms in Malaysia from 2008; the establishment of constitutional government in Cambodia in 1993; constitutional reforms in Thailand from 1992; the progressive espousal of the rule of law as official doctrine in Vietnam since 1987; the opening of Myanmar since 2010; the creation of an ASEAN human rights mechanism in 2011 – all of these are taking the region toward a different but uncertain future. What does seem clear is that debates as to the relevance of legal traditions are now concerned not so much with the accommodation of legal pluralism and the applicability of colonial law in an era of nationalism, but rather with the nature of citizenship and diffusion of justice in rapidly but unevenly developing economies and urbanizing societies.

See also: Human Rights: Legal Aspects; Legal Culture and Legal Consciousness; Legal Pluralism; Postcolonial Law; Socialist Law.

Bibliography

ASEAN Law Association, 1995. ASEAN Legal Systems. Butterworths, Singapore.
Basic Agrarian Law, No. 5 of 1960, Republic of Indonesia.
Bell, D.A., 2000. East Meets West: Human Rights and Democracy in East Asia. Princeton University Press, Princeton especially (Chapter 3).
Black, A., Bell, G. (Eds.), 2011. Law and Legal Institutions of Asia: Traditions, Adaptations and Innovations. Cambridge University Press, Cambridge.
Burns, P., 2004. The Leiden Legacy: Concepts of Law in Indonesia. KITLV, The Hague.
Butt, S., Lindsey, T., 2012. The Constitution of Indonesia: A Contextual Analysis. Hart Publishing, Oxford (Chapter 1).
von Benda-Beckmann, F., 2002. Who's afraid of legal pluralism? Journal of Legal Pluralism 47, 38.
von Benda-Beckmann, F., 1992. Symbiosis of indigenous and western law in Africa and Asia: an essay in legal pluralism. In: Mommsen, W.J., de Moor, J.A. (Eds.), European Expansion and Law: The Encounter of European and Indigenous Laws in 19th and 20th Century Africa and Asia. Berg Publishers, Oxford, pp. 307–325.
Cambridge History of Southeast Asia, 1992. The Writing of Southeast Asian History, vol. 1:1. Cambridge University Press, Cambridge (Chapter 1).

Chiba, M., 1986. Asian Indigenous Law in Interaction with Received Law. KPI, London/New York.
Crouch, M., Lindsey, T. (Eds.), 2014. Law, Society and Transition in Myanmar. Hart Publishing, Oxford.
Durbach, A., Renshaw, C., Byrnes, A., 2009. "A tongue but no teeth?": the emergence of a regional human rights mechanism in the Asia-Pacific region. Sydney Law Review 31, 211.
Engel, D.M., 1975. Law and Kingship in Thailand during the Reign of King Chulalongkorn. University of Michigan, Center for South and Southeast Asian Studies, Ann Arbor, MI.
Engel, D.M., 1978. Code and Custom in a Thai Provincial Court: The Interaction of Formal and Informal Systems of Justice. University of Arizona Press, Tucson, AZ.
Engel, D., Engel, J., 2010. Tort, Custom, and Karma: Globalization and Legal Consciousness in Thailand. Stanford University Press, Stanford, CA.
Fang, Liaw Yock, 1976. Undang-undang Melaka (The Laws of Malacca). Martinus Nijhoff, The Hague.
Fasseur, C., 1992. Colonial dilemma: Van Vollenhoven and the struggle between adat law and western law in Indonesia. In: Mommsen, W.J., de Moor, J.A. (Eds.), European Expansion and Law: The Encounter of European and Indigenous Law in 19th- and 20th-century Africa and Asia. Oxford University Press, Oxford.
Gillespie, J., Nicholson, P. (Eds.), 2005. Asian Socialism and Legal Change: The Dynamics of Vietnamese and Chinese Reform. Asia-Pacific Press, Canberra.
Griffiths, J., 1986. What is legal pluralism? Journal of Legal Pluralism 24, 1.
Hall, D.G.E., 1981. A History of South East Asia, fourth ed. Basingstoke, Macmillan.
Harding, A.J., 2001. Comparative law and legal transplantation in South East Asia: making sense of the"Nomic Din" (Chapter 9). In: Nelken, D., Feest, J. (Eds.), Adapting Legal Cultures. Hart Publishing, Oxford.
Harding, A.J., 2002. Global doctrine and local knowledge: law in South East Asia. International and Comparative Law Quarterly 51, 25.
Harding, A.J., 2008. The eclipse of the astrologers: King Mongkut, his successors and the reformation of law in Thailand. In: Nicholson, P., Biddulph, S. (Eds.), Examining Practice, Interrogating Theory: Comparative Legal Studies in Asia. Martinus Nijhoff, Leiden.
Harding, A.J., 2012. The Constitution of Malaysia: A Contextual Analysis. Hart Publishing, Oxford.
Harding, A., Leyland, P., 2011. The Constitutional System of Thailand: A Contextual Analysis. Hart Publishing, Oxford.
Holleman, J.F. (Ed.), 1981. Van Vollenhoven on Indonesian Adat Law. KITLV, The Hague.
Hooker, M.B. (Ed.), 1970. Readings in Malay Adat Laws. Singapore University Press, Singapore.
Hooker, M.B., 1975. Legal Pluralism: An Introduction to Colonial and Neo-colonial Laws. Oxford University Press, Oxford.
Hooker, M.B., 1978. A Concise Legal History of South East Asia. Oxford University Press, Oxford.
Hooker, M.B., 1980. Native Law in Sabah and Sarawak, Singapore, Malayan Law Journal.
Hooker, M.B. (Ed.), 1986. Laws of South East Asia, vol. i. Butterworths, Singapore.
Hooker, M.B. (Ed.), 1988. Laws of South East Asia, vol. ii. Butterworths, Singapore.
Hsien-li, T., 2011. The ASEAN Intergovernmental Commission on Human Rights: Institutionalising Human Rights in Southeast Asia. Cambridge University Press, Cambridge.
Huxley, A. (Ed.), 1996. Thai Law, Buddhist Law: Essays on the Legal History of Thailand, Laos and Burma. White Orchid Press, Chiang Mai.
Huxley, A., 2001. Pre-colonial Burmese Law: Conical Hat and Shoulder Bag. IIAS Newsletter Online 25. http://www.iias.nl/iiasn/25/theme/25T7.html.
Kasemsup, P., 1986. Reception of Law in Thailand – A Buddhist Society. In: Chiba, M. (Ed.), Asian Indigenous Law, KPI, New York, pp. 267–299.
Lev, D.S., 1972. Islamic Courts in Indonesia: A Study in the Political Bases of Legal Institutions. University of California Press, Berkeley, CA.
Lev, D.S., 2001a. Colonial law and the genesis of the Indonesian state. In: Lev, D.S. (Ed.), Legal Evolution and Political Authority in Indonesia: Selected Essays. Kluwer Law International, The Hague.
Lev, D.S., 2001b. Judicial unification in post-colonial Indonesia. In: Lev, D.S. (Ed.), Legal Evolution and Political Authority in Indonesia: Selected Essays. Kluwer Law International, The Hague.
Lev, D.S., 2001c. The lady and the banyan tree: civil law chance in Indonesia. In: Lev, D.S. (Ed.), Legal Evolution and Political Authority in Indonesia: Selected Essays. Kluwer Law International, The Hague.
Lukito, R., 2013. Legal Pluralism in Indonesia: Bridging the Unbridgeable. Routledge, Abingdon.
Marriage Law, No. 1 of 1974, Republic of Indonesia.
Moustafa, T., 2013. Judging in God's name: state power, secularism, and the politics of Islamic law in Malaysia. Oxford Journal of Law and Religion 1.
Phang, Andrew, B.L., 1990. The Development of Singapore Law. Butterworths, Singapore.
Santos, S.M., 2000. Common law elements in the Philippine mixed legal system. Asian Law 2 (1), 34.
Sidel, M., 2010. Law and Society in Vietnam: The Transition from Socialism in Comparative Perspective. Cambridge University Press, Cambridge.
Tamanaha, B.Z., 1993. The folly of the 'social scientific' concept of legal pluralism. Journal of Law and Society 20, 192.
Tan, Poh-ling (Ed.), 1997. Asian Legal Systems: Law, Society and Pluralism in East Asia. Butterworths, Sydney.
Thio, Li-ann, 1999. Implementing human rights in ASEAN countries: promises to keep and miles to go before i sleep. Yale Human Rights & Development Law Journal 2 (1).
Van Tai, Ta, 2004–2005. Buddhism and human rights in traditional Vietnam. Review of Vietnamese Studies. http://www.vietnamesestudies.org/uploads/4/5/8/7/4587788/tavantaibuddhism_and_human_rights.pdf.
Vickery, M., 1996. The constitution of Ayutthaya: the three seals code. In: Huxley, A. (Ed.), Thai Law, Buddhist Law: Essays on the Legal History of Thailand, Laos and Burma. White Orchid Press, Chiang Mai.
Whiting, A., 2003. Situating Suhakam: human rights debates and Malaysia's national human rights commission. Stanford Journal of International Law 39, 59.
Zan, M., 2000. Judicial independence in Burma: no march backwards towards the past. Asia-Pacific Law and Policy Journal 5, 1–39.

Legalization: The Expansion of Law's Domain

Jethro K Lieberman, New York Law School, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

As human societies have become more complex, they have seen increasing legalization – understood here as the tendency of placing problems under the control of the legal system. Legalization has become a dominant factor in industrially developed nations since the nineteenth century, as law has spread in an effort to control complex relations that have emerged with developments in technology, communications, travel, and the commercial market. Among its consequences are increases in rates of litigation and the development of law-based institutions, even in the international arena, to control harmful conduct by individuals, organizations, and the state.

Legalization is a Janus-faced concept that betokens both freedom to act and subjection to law. In the first sense, it means being freed by law to act in ways formerly forbidden – for example, the repeal of Prohibition by the US Constitution's Twenty-first Amendment that legalized drinking, or the 'decriminalization' of so-called victimless crimes such as prostitution and drug use in some nations. Legalization in this sense is also exemplified in legislative acts that free worshipers from the straits of official religion. In the second sense (for which other terms have been suggested, e.g., 'juridification' (Silverstein, 2009: pp. 1–11)), legalization signifies transferring control of an action or behavior to the legal system. Acts of legalization may certainly embrace both senses of the term – for example, the legal recognition of same-sex marriage serves, on the one hand, to provide status and benefits formerly unavailable, while on the other hand, it subjects the couple to legal requirements of support that were formerly undertaken on a voluntary basis. This article focuses exclusively on the second meaning of legalization, the tendency of a jurisdiction to place problems under the control of the legal system.

At its most general, legalization describes two types of lawmaking. In the first, the legislature moves a private relationship or action to the public sphere by banning, requiring, or regulating it (although, in a common-law system, the regulation can often come from the courts) – for example, requiring employers to compensate their employees for injuries on the job. A second type of lawmaking subjects public political players (legislators, other elected officials, or state policy makers) to the orders of courts – for example, holding legislative and other political districts in which elections are held to articulated apportionment rules. In a hierarchical system such as American federalism, a 'higher' level of government may impose legal accountability on jurisdictions that fail to enforce their own legal requirements strictly enough. Analogous international institutions, through the United Nations (UN) and various treaty obligations, impose, albeit still imperfectly, human rights norms that can lead to prosecutions of public officials who commit genocide or violate human rights, even if those actions are authorized by their national law.

Legalization Generally

Taken over long periods, legalization as a process has been part of the development of all but the smallest of human societies. Placing responsibility for the control of many aspects of human behavior in the legal system is central to civilization itself. In the long run, legalization replaces the unfettered decision of rulers, the bonds of custom, and religious authority with the rule of law, applying predictable and stable legal rules to the citizenry equally and avoiding the imposition of arbitrary power. A familiar example is the growth of the common law in England over several centuries, as legal principles and policies absorbed rules and traditions of various institutions that once held sway over most aspects of life, including family, work, property, religion, and political relations.

But legalization can also be understood in a narrower sense; the spread of law and lawlike institutions need not be synonymous with the rule of law or constitutional order. Throughout the twentieth century, communist societies and certain others (most notably fascist Germany) dispensed with constitutionalism in fact while pretending to it in theory, even as legal decrees and rules infiltrated numerous activities, including the economy and social relations, which had been much freer of control under the prior regime.

Nor does legalization necessarily imply any particular content to laws, which may be as liberal or repressive as the regimes that promulgate them. For example, laws regulating travel across political borders, which were instituted as early as the sixteenth century, have varied considerably over the centuries, and from country to country, in the strictness of the compulsory passport system that most nations have created. Except during the American Civil War, no passports were required to enter or stay in the United States, and this situation lasted until World War I. Quite the opposite prevailed in France, beginning in the 1790s, and until the middle of the nineteenth century in much of Europe. Likewise, as a means of regulating internal travel, many countries adopted internal passport requirements, while others, notably the United States, have maintained steadfastly a constitutional right to travel without hindrance or papers.

In modern times, the process of legalization can be discerned most clearly in three primary developments: (1) the increasing legal regulation of activities and behaviors previously unregulated or regulated by other means; (2) the growth of private rights of action by which individuals may press legal claims against other individuals and the state; and (3) the spread of legal relations into the international sphere. The first two developments are closely connected, most especially in the United States.

Increasing Legal Regulation

In 1881, Oliver Wendell Holmes Jr, the future US Supreme Court justice, declared in his classic book *The Common Law* that "the general principle of our law is that loss from accident must lie where it falls, and this principle is not affected by the fact that a human being is the instrument of misfortune" (Holmes, 1963: p. 76). But at that very time, American law was undergoing a sea change; the law of negligence would transform many accidents into legally compensable events, and the courts were pulling down a barricade of rules against defendants' liability that they had constructed laboriously during much of the nineteenth century. The growth in American law closely followed the rise of industrialization, the explosion of technological development, and the spread of the market economy. Not all increases in the number of statutes and judicial decisions can be counted as legalization, since refinement and improvement in law is not necessarily the same as the spread of law and legal relations to new areas. Nevertheless, much of the outpouring of new law was part of the legalization of the country, a fitful, although eventually sweeping, movement toward 'total redress' (Lieberman, 1981: p. 31) or 'total justice' (Friedman, 1985). Throughout the twentieth century, the state endeavored to refashion the theory of harm (Lieberman, 2012) and to elaborate innovative ways of deterring and redressing the manifold injuries that came to be recognized as products of changing economic and social relations. This legalization of harm has unfolded in four principal ways.

First, from the 1880s to the time of this writing, the legislatures of the US states and the US Congress have enacted a tidal wave of statutes, dwarfing all previous eras of legislative enactment in all other countries. Since this is not a history of lawmaking, the list that follows is not intended as a chronology but rather as a snapshot of the kinds of activities that were enveloped by law during the twentieth century: labor and employee relations (including working conditions, pay, health, and safety), environmental pollution, consumer concerns (including banking and credit, housing, product safety, and advertising), the market mechanism (including antitrust, corporate structure, securities regulation, and the like), communications, elections, transportation, education, discrimination (by a host of institutions, public and private, on a variety of bases, including race, sex, ethnicity, religion, and disability), family relations (including marriage and divorce, child custody, and welfare), public assistance (including unemployment and health care insurance, welfare benefits, and old-age security), and personal relations (including sexual relations and procreation, and violence against women), to name only the most prominent of the areas that have come under legal sway.

Second, legalization has been enshrined in administrative bureaucracies created to manage a host of private relations. Although some statutes spell out bright-line rules (for instance, setting the drinking age at 21 years, or placing a ceiling on interest rates that may be charged), legislatures routinely delegate not merely enforcement power to administrative agencies to carry out major regulatory programs but also the very power to define the law. One measure of the scope of administrative law in the United States is contained in the pages of the *Federal Register*, the US government's official digest of proposed and promulgated federal regulations; it runs to around 60 000 pages annually. The impulse to legalize has thus created a vast bureaucracy and undercut the rule of law as understood by one of its leading proponents, A.V. Dicey, who, speaking of the British legal system in the nineteenth century, said "there can be with us nothing really corresponding to the 'administrative law'" (Dicey, 1959: p. 203).

Third, alongside the legislatures and administrative agencies, courts themselves have multiplied legal rights. The common-law understanding of harmful activity has expanded to encompass, for example, strict liability in tort, so that manufacturing operations that were once legally unapproachable in the absence of statutes have been a regular fare at the judicial feast. Likewise, the courts in the United States have expanded vastly the panoply of public rights, particularly in the arena of the Fourteenth Amendment's command that the states provide due process and equal protection of the laws.

Fourth, as a direct consequence of the preceding three developments, individual litigants have gained the right to take their multiplying grievances to court, as discussed in the next section.

Growth of Private Rights of Action

The widely observed penchant of Americans to sue springs from a rich tradition. As long ago as 1835, Alexis de Tocqueville observed, in an often quoted passage in *Democracy in America*, that "there is hardly a political question in the United States which does not sooner or later turn into a judicial one" (de Tocqueville, 1966: p. 248). Nevertheless, during the last quarter of the twentieth century, it became the conventional wisdom of many observers that the United States was suffering from something new: a 'litigation explosion' (Olson, 1991). Although the term is obscure, and such statistics as are available are subject to different interpretations (Galanter, 1983, 1993, 1996, 1998), the types of actions that plaintiffs may bring to court were greatly extended during the twentieth century. Many activities once entrusted to other domains – for example, politics – or left wholly unregulated and open to the whims of individuals, have become objects of legal oversight, in many instances prompting courts to assume administrative supervision of schools, prisons, and other major institutions not for the moment but for decades, a regime that has been termed 'democracy by decree' (Schoenbrod and Sandler, 2003).

Perhaps the most notorious example of the trend toward legalizing what was once purely political was the investigation of US presidents and high-ranking executive officials by so-called independent counsel. Although in form a public prosecutor, the independent counsel was in effect a combination of

a private lawyer and a client, since the executive branch was precluded from effective oversight of the prosecution. The independent counsel movement had its origins in the Watergate scandal in 1973, when the US attorney general authorized the appointment of a private lawyer as a special prosecutor and promised to keep the government's hands off the investigation. President Nixon's discharging of Archibald Cox, the first Watergate special prosecutor, led to a federal statute providing for the appointment of independent counsel by the federal court. (The law was allowed to lapse in 1999.)

A highly publicized instance of legalization of institutional behavior long thought to be private and outside the sphere even of political institutions was prompted by the worldwide pedophilia scandal in the Catholic Church. The revelations that thousands of priests had sexually abused children while their superiors shielded them from consequences led prosecutors in the United States and some other countries to enforce existing laws that had long been ignored and prompted legislators to enact new laws requiring evidence of such crimes to be reported to public authorities. Since the late 1990s, more than 3000 civil suits have been filed in the United States alone, leading to payouts totaling more than $1 billion and to the bankruptcy of several Catholic dioceses.

Other notable examples: The overthrow of racial segregation laws in the United States primarily began with cases filed in individual courtrooms in the 1950s. Apportionment of legislative districts, once solely a matter for political resolution (even in the face of constitutional commands widely disregarded by the legislators themselves), has become a cottage industry of litigation in the United States since the 1960s. In the 1970s, manufacturing and other practices were subject to sweeping oversight in court efforts to clean up the environment. Sexual harassment, a still evolving concept, became justiciable as the result of statutes and court decisions in the 1980s, overturning a centuries-long practice of remitting women to their private suffering in the face of it. In the 1990s, the tobacco industry was forced to pay hundreds of billions of dollars in health-related claims, largely through settlements in the wake of legal claims, and a similar but so far unavailing effort to control the distribution and sale of guns was mounted in the late 1990s. At the beginning of the twenty-first century, animal rights activists are attempting to prevent medical experimentation on animals by seeking to persuade courts to declare animals not merely property but possessors of rights.

Less generally discussed are the legal devices that serve as a spur to further legalization. In the Sherman Antitrust Act, for example, the US Congress authorized private individuals to sue for restraints of trade and monopolization by business interests, and the law permitted winning plaintiffs to recover both treble (triple) damages and their lawyers' fees from losing defendants. Other devices include contingent fees in personal injury suits; the class action, which permits the consolidation of small claims arising from the same event into a single suit, thus giving the plaintiffs' lawyers a financial incentive to proceed with the case; and the relaxation of so-called standing rules that determine who the proper party to file a suit is.

Legalization through private rights of action is not limited to the lawsuit itself. Understanding that lawsuits are always possible has led some people to fashion personal relationships as legal ones. Two examples from the marriage context are illustrative. Consider first the widespread use of prenuptial contracts allocating assets in the event of divorce, even before the marriage license is signed. More legalized still, apparently "some celebrities have asked wedding guests to sign nondisclosure agreements as a condition of attending a ceremony that might attract the interest of the tab[loid]s" (Beam, 1999: p. 64).

The consequence of growing legalization may alter conduct in ways that do not touch directly on legal relationships. For example, people may be more cautious in a law-saturated age about expressing opinions, even about matters with which they are professionally involved. During a Senate confirmation hearing, for example, US President Obama's 2013 nominee to head the Central Intelligence Agency refused to say whether he thought waterboarding constituted torture; although he called it 'reprehensible,' "I am not a lawyer, … and I can't address that question," he insisted (Terkel, 2013).

Growth of the Legal Profession

One measure of legalization sometimes propounded is the gross number of lawyers or lawyers per capita. In the United States, lawyers numbered about 285 000 in 1960 and more than quadrupled, to about 1.3 million, in 2013; per capita, the ratio rose from one lawyer for every 631 Americans in 1960 to one lawyer for every 243 in 2013 (however, perhaps only half the total are actually in practice). By comparison, in Japan in 2013, lawyers number around 30 000 (although the meaning of 'lawyer' is different and the comparison is not exact), for a per capita ratio of one lawyer for every 4233 Japanese. Whether these differences are meaningful, and if so, in what way, remains controversial. The number of lawyers may be explained at least as convincingly by the number and nature of transactions and interpersonal contacts as by some tendency to greater legalization. Quantifying legalization is a research task that remains to be undertaken.

Legalization in and of the International Arena

Historically, legalization has been an internal affair of each nation. Relations between nations have been governed by diplomatic and political process and, failing that, by war. But following World War II, the prosecutions of Nazi war criminals laid the groundwork for the principle that government officials are not legally free to do whatever they wish as leaders of sovereign nations, and they may not always claim immunity when finally caught. The Nuremberg trials established the concept that war crimes and crimes against humanity are justiciable. In modern times, the shield of diplomatic and sovereign-state immunity may be penetrated by the courts under certain circumstances.

In 1945, the UN established the International Court of Justice (ICJ) as its judicial arm. Located in The Hague, it is a tribunal in which member states may adjudicate claims against each other. ICJ rulings can be vetoed by any of the five permanent members of the UN Security Council, of which the United States is one. Despite obvious problems and a modest track record, it represents the ideal of submitting disputes to legal rather than political or military authority.

Although large numbers of war crimes since World War II have doubtlessly gone unpunished (only one major Vietnam war crimes case, concerning the massacre at My Lai, was prosecuted in the United States), by the 1990s a new sentiment led to the creation of two temporary International Criminal Tribunals, also situated in The Hague, with jurisdiction to investigate and prosecute acts of genocide and other war crimes in the former Yugoslavia and Rwanda. In 1999, then-President Slobodan Milosevic of Serbia became the first sitting president to be indicted by an international tribunal (along with other high officials of the regime), and in 2000 the International Criminal Tribunal began a series of prosecutions that garnered international attention. A similar temporary international tribunal, the Special Court for Sierra Leone, prosecuted Charles G. Taylor, former president of Liberia, who in 2012 became the first former head of state since Nuremberg to be convicted by an international court of war crimes and crimes against humanity; a special appeals panel upheld his conviction and a 50-year sentence in 2013 (Simons and Cowell, 2013).

Following the success of the early temporary tribunals, a permanent international tribunal, the International Criminal Court, was established in 2002 in The Hague under the so-called Rome Statute, a multilateral treaty to which more than 120 countries are parties (but not the United States). It is authorized to prosecute cases of genocide, crimes against humanity, war crimes, and the crime of aggression but only when individual nations are unable or unwilling to do so. In its first decade, it has acted in a handful of cases against leaders and military officers of eight African nations.

Domestic courts may also be willing to try such offenses. Perhaps a portent of things to come, in 2013 for the first time a former head of state was tried and convicted in his own country on charges of genocide and crimes against humanity; Gen. Efrain Rios Montt of Guatemala was sentenced to 80 years in prison by a Guatemalan court for directing the army during the 1980s to massacre Mayan families and wipe out their villages.

Other portents include successful civil suits against the estate of Ferdinand Marcos, the Philippine dictator, for human rights violations, and the 1998 British proceedings against former Chilean dictator Augusto Pinochet. He was arrested in London on a warrant issued by Spain, charging him with numerous civil rights violations during his 17-year rule in Chile, including murder, torture, and kidnapping. In the United States, the courts have been amenable increasingly to civil suits brought by victims of terrorism against their attackers. In 1998, for example, US federal courts awarded more than $300 million in two suits against Iran for its involvement in supporting terrorists involved in killing and kidnapping, and in 2000 Iran was ordered to pay $300 million in damages to one of the hostages taken in Teheran in 1980. Also in 2000, thousands of Filipino slave laborers conscripted by the Japanese Army during World War II filed a class action under California law against Japanese companies or their successors that forced them to labor in mines and plantations. In that same year, Congress enacted legislation enabling victims to collect damage awards from nations that sponsor terrorism, when those nations have assets in the United States.

These and other developments have led one political theorist to conclude that a 'dualistic world order' is in the making, with the development of international legal institutions that are semiautonomous from the world's sovereign nation-states (Cohen, 2012).

Legalization Outside the United States

Although the trend toward legalization is more advanced in the United States than in other countries, there is evidence that the direction of legal change is toward more law, not less; the increase in law is occurring in both industrialized Western nations and less developed nations worldwide, for the same reasons that the rule of law emerged in the West. A globalized economy demands stability, and globalized communications demonstrate to peoples everywhere the gap between stable and unstable legal systems. Not every nation fits this abstraction; despite its emerging markets, communist China maintains a strong political hold on its courts, just as religiously fundamentalist countries prefer clerical decrees to secular legal rules. In the West, what few comparative studies exist suggest that litigation rates in certain European countries, adjusted for differences in costs, are not appreciably different from rates in the United States (Galanter, 1992: p. 10). Two articles from the *New York Times*, chosen at random, suggest that the impulse toward legalization may be growing in radically different societies. A Chinese citizen sued Canon, Inc., a Japanese company, for $12 million for mental distress, because a computer disk promoting Canon's ink-jet printers said that China, Hong Kong, and Taiwan are separate countries, thus allegedly undermining China's sovereignty (Smith, 2000). In 2012, in a decision that alarmed scientists throughout the world, an Italian court convicted seven seismologists of manslaughter and sentenced them to 6 years in prison and ordered them to pay more than $10 million in damages for failing to predict an earthquake that killed 300 people in the city of L'Aquila in 2009 (Povoledo and Fountain, 2012). The case remains on appeal.

Some percentage of increased legalization may be attributable to the vast changes in global communications in the early twenty-first century, permitting instantaneous publication of and easy access to news that can pressure governments to act. The story of a woman in New Delhi who in December 2012 was brutally raped and murdered on a bus by a gang of thugs became an overnight sensation around the world. The news shook India, provoked protests throughout the country, where rape cases have been largely ignored, and forced the government to initiate a series of legal changes, including creation of a fast-track rape court (Barry and Sharma, 2013).

Continuing Trends

Without clear measures of legalization, it is impossible to predict at what pace it will continue, but there is no reason to suppose that the force of legalization will diminish. Increasingly it is the solution that political leaders favor to confront new hazards and challenges, and to reduce risks of all kinds. In 2008, for example, New York City mandated disclosure of calories on the menus of certain classes of restaurants in the hope of combating an epidemic of obesity; other cities and a few states have enacted such laws too, and a similar provision was included in the US federal health care act in 2010. Likewise,

increasing numbers of American states, as well as dozens of countries, now ban drivers from talking and texting on cell phones. Legalization is not limited to challenges posed by diet or technology. Rapid increase in international business transactions has led to many new laws, treaties, and conventions to regulate international trade. First ratified by 11 governments in 1988, the UN Convention on Contracts for the International Sale of Goods had 80 adherents by 2013. Nearly 80% of the world's nations, although not the United States, have ratified the UN Convention on the Law of the Sea, but even the United States accepts much of this regime regulating legal rights in ocean navigation and resources as binding international law. As nations, organizations, and even individuals around the world interact more often and more closely, it seems inescapable that legal regimes will be pressed into smoothing out the resulting collisions.

See also: International Law and Treaties; Law and Terrorism; Law as an Instrument of Social Change; Litigation; Nazi Law; Regulation and Administration; Regulatory Agencies; Rights: Legal Aspects; Socialist Law; War Crimes Tribunals, Crimes against Humanity and Transitional Justice.

Bibliography

Barry, E., Sharma, B., 13 September 2013. Death sentences set in gang rape that shook India. New York Times, A1.

Beam, A., August 1999. Tabloid law. The Atlantic Monthly, 55–68.

Cohen, J., 2012. Globalization and Sovereignty: Rethinking Legality, Legitimacy, and Constitutionalism. Cambridge University Press, New York.

de Tocqueville, A., 1966. In: Mayer, J.P., Lerner, M. (Eds.), Democracy in America. Harper & Row, New York.

Dicey, A.V., 1959. The Law of the Constitution, tenth ed. Macmillan, London.

Friedman, L., 1985. Total Justice. Russell Sage Foundation, New York.

Galanter, M., 1983. Reading the landscape of disputes: what we know and don't know (and think we know) about our allegedly contentious and litigious society. University of California at Los Angeles Law Review 31, 4–71.

Galanter, M., January 1992. Law abounding: legalisation around the North Atlantic. The Modern Law Review 55, 1–24.

Galanter, M., 1993. News from nowhere: the debased debate on civil justice. Denver University Law Review 71, 77–113.

Galanter, M., 1996. Real world torts: an antidote to anecdote. Maryland Law Review 55, 1093–1160.

Galanter, M., 1998. An oil strike in hell: contemporary legends about the civil justice system. Arizona Law Review 40, 717–752.

Holmes Jr., O.W., 1963. In: Howe, M. de W. (Ed.), The Common Law. Little, Brown, Boston.

Lieberman, J.K., 1981. The Litigious Society. Basic Books, New York.

Lieberman, J.K., 2012. Liberalism Undressed. Oxford University Press, New York.

Olson, W., 1991. The Litigation Explosion. Dutton, New York.

Povoledo, E., Fountain, H., 23 October 2012. Italy orders jail terms for 7 who didn't warn of deadly earthquake. New York Times, A4.

Schoenbrod, D., Sandler, R., 2003. Democracy by Decree: What Happens When Courts Run Government. Yale University Press, New Haven.

Silverstein, G., 2009. Law's Allure: How Law Shapes, Constrains, Saves, and Kills Politics. Cambridge University Press, New York.

Simons, M., Cowell, A., 27 September 2013. 50-year sentence upheld for ex-president of Liberia. New York Times, A3.

Smith, C.R., 7 August 2000. Miffed Chinese sue Japan companies. New York Times, A6.

Terkel, A., 2013. John Brennan Refuses to Say Whether Waterboarding Is Torture. Huffington Post. www.HuffingtonPost.com/02/07/2013.

Legitimacy, Sociology of

Cathryn Johnson and Lesley Watson, Emory University, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by B. Badie, volume 13, pp. 8706–8709, © 2001, Elsevier Ltd.

Abstract

Scholars who study legitimacy processes examine, for example, how new patterns of behavior, beliefs, and structures become legitimated, how the legitimacy of groups' status hierarchies emerges, and how authorities acquire legitimacy in organizations. Scholars also investigate the consequences of legitimacy. The stability provided by legitimated structures is the basis for much coordination and cooperation. Yet, legitimacy processes also perpetuate inefficient practices and inequalities, as well as allow advantages to high-status individuals and groups, limit responses to procedural and distributive injustices, and undermine collective identity and action among groups targeted by discrimination.

Definition of Legitimacy

The early twentieth century sociologist, Max Weber, argued that even though individuals may not always hold the same beliefs, norms, or values, their behavior still falls in line with the cultural beliefs and rules that they presume are accepted by most others. Because individuals perceive that other people support and accept this social order, the order seems like a valid, objective feature of social life. As a result, individuals act in accord with that order, whether they agree or disagree (Weber, 1978). This Weberian formulation offers a central insight for a working definition of legitimacy. Legitimacy is the taken-for-granted support of an aspect of social life (e.g., an action, a procedure, an individual, a position, or a structure of positions) by a real or an implied audience. It is fundamentally a collective process, and depends on the implied presence of a social audience that is assumed to accept the framework of beliefs, norms, and values. Finally, it has both a cognitive dimension, whereby an aspect of social life is perceived as valid and objective, and a normative dimension where this taken-for-granted social object is perceived as what is right (see Johnson et al., 2006 for a discussion of similarities in definition within social psychological and organizational literature, see Deephouse and Suchman, 2008 for discussion of the concept of legitimacy in organizational institutionalism, see Bobocel et al., 2010 for discussion of legitimacy and justice in psychology).

An example of legitimacy is seen in the interaction between referees and coaches. When a referee orders an unruly coach to leave the soccer field, it is typically not necessary for the referee to persuade the coach to leave. The coach may be very upset, throw his or her hat on the ground, or kick the dirt, but will ultimately comply with the referee. Compliance, obeying even when one privately disagrees, in this case occurs because both the referee and coach operate within a larger social system, with its set of roles, rules, and beliefs that go beyond any individual. This larger system is accepted and approved of by a majority of the social audience, including the players, the coaches, the fans, and the administrators of that particular league. There is, then, a general consensus that the referee has the right, in his or her role, to throw a disruptive coach out of the game, and that the coach must obey, even if he or she personally disagrees. Imagine if there was no legitimate order in soccer; chaos would ensue with every disagreement and conflict. These same sorts of legitimated roles and rules apply in a range of social settings including, but not limited to, courts, airports, and the workplace. In each of these settings, legitimacy acts as a sort of glue that holds together larger social systems in which people operate and interact with one another on a daily basis.

Many things can be legitimated. For example, a particular act such as a manager promoting a subordinate, an individual in a position (e.g., floor supervisor), a status hierarchy in a group, a practice or policy in an organization (i.e., family leave), a structure of positions in groups or organizations, or intergroup status relations in a society become legitimated through a social process in which people assume that other people in general accept the object for what it is and, often over time, for what it should be. Questions of legitimacy, then, repeatedly arise in studies pertaining to political and organizational structures, status relations in and between groups, and inequality.

Some scholars who study legitimacy processes examine how new patterns of behavior, beliefs, and structures become legitimated. Others study legitimacy processes in groups and focus on how the legitimacy of groups' status hierarchies emerges, and how authorities acquire legitimacy in organizations. Scholars also investigate the consequences of the legitimacy, for example, of status hierarchies in groups and status relations among groups, as well as the conditions under which these aspects of social life become inefficient and perpetuate inequality within groups, organizations, and society. They also investigate the consequences of legitimacy of authorities for interaction in organizations.

Legitimacy as a Social Process: Four Stages

Legitimacy theories examine how new patterns of behavior, beliefs, and structures become widely accepted into the broader cultural framework so that they become taken-for-granted aspects of social life. When comparing legitimacy research in sociological social psychology with institutional work in organizations, scholars reveal four stages in the legitimation of new social objects: innovation, local validation, diffusion, and general validation (Johnson et al., 2006).

In the first stage, a social innovation is created to address actors' needs, purposes, or goals in a local situation. One example could be the development of new procedures to meet federal policies in schools. Often, new social objects are created in reaction to changes in structural conditions, such as newly gained access to or the distribution of valued resources. For instance, a school may develop new procedures in order to obtain federal grant money for special projects. Responses to structural conditions create strategic interests or contingent events for actors in local situations. Clearly, some new objects never acquire legitimacy (e.g., subscription financing for commercial radio in the United States see Leblebici et al., 1991), whereas others do successfully become part of the broader cultural landscape, readily available and acceptable to other actors outside of the initial local context. In addition, sometimes the same innovation arises in several local contexts simultaneously.

In the second stage, the innovation is accepted locally. In order for this to occur, the local actors must construe the innovation as consistent with and linked to the existing, widely accepted cultural framework of beliefs, norms, and values (Walker, 2004). At the local level, where the innovation is initiated, actors may explicitly justify this link by providing accounts or justifications of how the innovation meets the needs and goals in a way that is clearly consonant with the existing cultural landscape. Or the innovation may gain local consensus simply by not being implicitly or explicitly challenged. Continuing with the example of education policies, a school committee may develop modules that fulfill the requirements of the federal grant, and justify these changes by showing how the modules fit within the current school curriculum. As result of either explicit justifications or being implicitly accepted because it is not challenged by actors, the innovation acquires local validation. A new prototype or cultural schema is adopted as the acceptable way of doing things to meet local needs and/or goals.

Once this new prototype or schema is locally validated, in the third stage, it may be diffused into other new, local situations. For instance, other schools in the district or state may adopt similar educational modules in order to secure grant funding, rather than developing their own modules to suit the needs of their students and curriculum. Because the innovation is now interpreted as a valid social fact, actors in other local situations are more likely to adopt it as mere fact. The diffusion process is partly a result of actors' expectation that people in other local contexts accept this new prototype. Importantly, actors are more likely to have this expectation when the innovation is perceived to meet goals/purposes that are consonant with already widely accepted goals (e.g., efficiency) (Walker, 2004). The innovation then becomes a useful and even necessary cultural schema for how and why people do things. As the innovation spreads, its adoption into new contexts needs less explicit justification than when it was adopted in the first local context. In addition, often the initially created innovation did indeed address some immediate goal or problem, but it may eventually become less useful and efficient in later adopter contexts, yet still be perceived as an acceptable way of doing things. For example, another theme of institutional scholarship focuses on how early adopters of an innovation are driven by technical concerns, whereas later adopters can be driven by emulation.

Over time, as a result of this diffusion process across local contexts, actors eventually assume that most people believe that the innovation is acceptable. As a result, the innovation acquires widespread acceptance, becoming part of society's shared culture. Initially validated either in one local context or in a few spontaneously, the innovation becomes generally validated, that is, becomes part of the status quo, and is used to frame the future behavior of actors across many contexts. For example, schools may continue to teach the same modules for years after the grant funding has run out, based on the notion that 'we always teach this.' Although there is certainly room for new prototypes to appear that challenge the original innovation, once a prototype/cultural schema becomes generally validated, it is not easily replaced compared to innovation that only remain locally validated.

Johnson et al. (2006) illustrate this four-stage process by drawing upon institutional theories in organizations to illustrate how two organizational forms, the for-profit corporate form and the nonprofit form in the arts, emerged in local contexts and then became part of the larger cultural framework (e.g., DiMaggio, 1992; Roy, 1997). They also examine how newly emerging status beliefs may become widely accepted as part of the broader cultural framework. Status beliefs are cultural schemas about which categories of individuals are thought to be worthy or unworthy in society. Drawing upon status construction theory (Ridgeway, 1991, 2011), they examine how a new type of worthy/unworthy actor in society may emerge from the transformation of a mere difference of categories of people (i.e., difference not attached to status value) into a widely accepted status characteristic (where two or more categories of a characteristic are differentially valued in a society). In this case, some categories of people (e.g., whites and men) are perceived as more worthy and competent than other categories of people (e.g., blacks and women) in general. They show how these new social objects, organizational forms and status beliefs examined at two different levels of analysis, have similar paths in terms of their legitimation processes.

Emergence of Legitimacy of Status Structures in Groups

Once status beliefs emerge in societies where a mere difference in categories of people transforms into a widely accepted status characteristic, these status beliefs drive the formation of status hierarchies in groups, as well as the legitimation process of these hierarchies. Status hierarchies develop in groups (e.g., committees, task forces, and gangs), where some members are seen as more worthy and esteemed than other members. Status structures of groups are rank-ordered relationships between group members where some members have low status and others have high status in the group. Individuals who have higher status in society tend to have higher status in face-to-face groups, and there are widespread cultural beliefs that indicate this pattern. In the United States, for example, men, whites, and middle/upper class individuals tend to have higher status and are perceived as more competent and

worthy in general than lower status groups like women, blacks/Latinos, or poor/working class individuals. Because of this link between states of social characteristics and general competence beliefs, these members are more likely to offer their opinions and be more influential in decision making. As a result, they are also afforded high status in specific groups, and their opinions are more likely adopted than opinions of other group members. A status structure thus forms in which these members have higher status than other members who possess the less valued state of the characteristic (Correll and Ridgeway, 2003). How does this status structure become accepted and supported, either implicitly or explicitly, by the group members, where they believe that the high-status members should have high status in the group?

Legitimacy theories argue that low- and high-status members in a group expect that those who possess highly valued states of characteristics will also occupy highly valued positions within the group because this is typically what they perceive around them. In other words, individuals with highly valued status characteristics are also likely to hold high-status positions in groups within occupational, economic, political, and religious structures in general (Correll and Ridgeway, 2003). For example, white men are typically seen in top positions in political, educational, and economic institutions (e.g., Chief Executive Officers of Fortune 500 US companies and Fortune Global 500 companies, brokers on Wall Street, and senators in the United States). In addition, clergy, rabbi, and other religious leaders in religious institutions are still predominately men.

Consequently, when individuals advantaged in terms of status characteristics become high-status members within a group, low- and high-status members tend to react to this as if this ranking should have happened this way because they, in fact, expected it. Thus, low-status members enact deferential behaviors such as giving esteem, respect, and honor (communicated verbally or nonverbally) toward high-status members. This interaction creates a process where everyone believes that everyone else supports who is more worthy and exercises more influence. Members believe that other members support the status structure. If no one challenges this pattern of deference, members will continue to act as if this should happen, and the hierarchy becomes implicitly legitimate.

Emergence of Legitimacy of Individuals in Authority Positions

Scholars also examine how individuals in authority positions acquire legitimacy. Authority refers to the ability and right of an actor (e.g., an individual, an organization, or a governmental body) to issue orders to others. The actor with authority has rights vested in the position to direct, regulate, and evaluate the behavior of others, within limits. Individuals in authority positions have a range of directives that are considered appropriate, such as determining work assignments, pay, and evaluating performance.

Studies show that individuals in authority positions are more likely to be legitimated when their appointments are based on qualifications and past achievements and are designated by someone from the top of the authority structure (i.e., when they are supported from above or authorized) (Zelditch and Walker, 1984). They are also more likely to be accepted as the authority when they are supported by their subordinates and peers (i.e., endorsed).

Authorities acquire support from subordinates, in part, in two fundamental ways in which they interact with their subordinates: (1) the use of fair procedures when making decisions and (2) the benevolent use of their power. Authorities are perceived to use fair procedures when they make decisions that are seen as unbiased, respectful, consistent across individuals, and they take into account subordinates' views. An authority's use of fair procedures and treatment ensures that subordinates feel respected within their group, which in turn increases subordinates' feelings of self-worth (Tyler, 2001). Many scholars assume that individuals are motivated by and desire positive social identities from their standing in a group and the value of their group. The use of fair procedures by authorities signals to the subordinates that they are respected within their group. Therefore, when an authority uses fair procedures to make decisions, the legitimacy of that authority increases.

Authorities also gain endorsement when they provide resources to their subordinates that benefit their welfare (Blau, 1964). Authorities often have opportunities to contribute to their subordinates' welfare by distributing rewards that support them in being successful in their jobs. For example, they have access to valuable knowledge, skills, training, and strategic information that is useful to subordinates. They may also offer guidance, assistance, and advice to enhance and facilitate subordinates' work. When authorities provide rewards that contribute to the collective interest of subordinates, joint obligations between the authority and the subordinates arise (Blau, 1964; Hegtvedt and Johnson, 2009). Repeated successful exchanges (exchanges of rewards and cooperation) between authorities and their subordinates are likely to stimulate perceptions of trust and fairness and feelings of social obligation (Lawler, 2006; Molm, 2006). Taken together, using fair procedures in decision making and distributing valued rewards fairly among subordinates contribute to a collective sense of legitimacy. That is, subordinates perceive that other subordinates support the authority (i.e., endorsement) and this, in turn, perpetuates the authority's legitimacy.

Additionally, authorities can also acquire legitimacy based on their interactions with and ties to other more powerful and legitimate members of the organization, as well as their peers. Legitimacy (and power) comes from breadth and depth of these social connections, especially those that are outside one's immediate group (Smith-Lovin, 2007). For example, sponsors or top managers can support leaders in their positions, allow them to bypass red tape, provide them with inside information, and give them status purely by association. In addition, alliances with peers allow for direct exchanges of resources such as favors, information, and challenging and visible work assignments.

Consequences of Legitimacy of Authority

Legitimated social structures, procedures, and actions tend to remain stable. In the case of status hierarchies or authority

structures, in some contexts, the legitimacy of authority relationships fosters stability, coordination, and cooperation in the interactions between members of the organization (Dornbusch and Scott, 1975; Zelditch, 2006). The benefits of legitimacy often have downsides, however. Just as legitimacy helps to maintain order and stability, and prevent chaos in some situations, in other situations legitimacy upholds inefficient practices and contributes to the perpetuation of inequality within and among groups, organizations, and societies. Furthermore, while the benefits of legitimacy are greater for women and minorities, these individuals often have the most difficulty becoming legitimated. These consequences are taken in turn.

Some positive outcomes of legitimacy processes, in certain situations, are solidarity, cooperation, and productivity. Subordinates who perceive their superiors as legitimate authorities are more compliant and deferential, and they also perceive legitimate authorities as more effective and influential. These authority figures have a broader range of directives they may give to subordinates, assuming that these directives fall within the scope of their authority.

Legitimacy also fosters organizational stability because in conflict situations, legitimated authorities are more likely to elicit compliance by their subordinates, as opposed to authorities who are not legitimated who may have subordinates form coalitions or go above their heads to resolve the conflict (Johnson et al., 2000; Johnson et al., 2006). In addition, if members do not defer to high-status members in the group, they face the threat of informal sanctions from other group members. As a result, compliance with high-status members is expected and, in fact, is what typically occurs. This pattern of behavior, then, maintains the status quo (Correll and Ridgeway, 2003). While the ability of legitimated authorities to avoid conflict in this manner may lead to better outcomes for those who have high status and greater cooperation within the organization, this also points to key problems caused by legitimacy: it may aid in the persistence of inefficient practices and inequality.

Legitimated authority structures have staying power, in part, because when individuals perceive a system is supported by others they are more compliant and engage in fewer observable efforts to change it. In fact, when a system is perceived to be authorized and endorsed, its stability and ability to elicit compliance persist even in the absence of individual propriety (i.e., personal support for the system). Their persistence is related to the perception of collective support, rather than the appropriateness or effectiveness of the structure. This is not to say that individuals do not recognize practices as inefficient or unfair. Rather, the issue is the pressure to change the system is countered by pressures of legitimacy that negate attempts to change the structure (Zelditch, 2006). These pressures include the appearance of consensus, making people think that the legitimated structure or practice is just 'the way things are,' and a system of sanctions for noncompliance. Sanctions may be formal or informal, ranging from actions such as negative evaluations from supervisors and stunted pay growth to having one's ideas dismissed publicly or social exclusion. In the face of seemingly widely supported authority, possible repercussions for noncompliance may appear too costly based on perceived prospects for success in changing the system.

An example of compliance even in the absence of propriety can be found in airport security. While individuals may be opposed to the procedures in airport security, or consider them an ineffective means of ensuring public safety, they are unlikely to act out against this system. Even when airline passengers are personally inconvenienced or perhaps feel violated by security procedures, they are likely to assume that other people agree this system is necessary to ensure their safety and these procedures are the most effective way to achieve that goal. Furthermore, going against the grain would likely have a range of negative consequences, formal and informal. Informal consequences could include upsetting or delaying other passengers by refusing to remove personal items or be screened, but formal consequences could include issues such as having personal belongings confiscated or being denied permission to fly.

On the positive side, the aforementioned benefits of legitimacy are greater for women and minorities in authority positions, because they are less likely to receive cooperation and deference when they are not legitimated. For instance, when female leaders receive institutionalized support and authorization, they have more influence than if they were assigned to leadership roles randomly or by task ability. These female leaders do not achieve the same influence levels as male leaders who were selected based on task ability, but it remains that disadvantaged leaders can attain influence comparable to other more advantaged actors when they have institutional support (Lucas, 2003).

Despite the benefits of legitimacy for disadvantaged groups like minorities and women, however, individuals in these groups often have the most difficulty acquiring legitimacy. Barriers to legitimacy are often rooted in double standards women and minorities face in regard to proving competency (Foschi, 2000). Gender and racial stereotypes associated with assumptions about competence make it more difficult for women and minorities to become legitimated in their positions. Yet, having legitimacy affords high-status group members the right to use dominance in groups to ensure compliance (Ridgeway et al., 1994). Given that women and minorities are less likely to acquire legitimacy in groups, they are not expected or allowed to behave in dominant ways. If they do, they are often perceived to be overly directive. Women and minorities, then, are more likely to be perceived as assertive and strong willed than men and whites who employ the same behaviors and are considered competent (Eagly and Carli, 2007). These findings show that low-status group members are more limited in the range of behaviors they can effectively engage in, even if they are placed in high-ranking positions. In addition, people who occupy positions of power without being legitimated also focus more efforts on attending to negative stereotypic attributes and information about their subordinates (Rodriguez-Bailon et al., 2000), presumably in order to justify their position and maintain the existing social structure. Thus, legitimacy plays a role in maintaining stereotypes.

Scholars also consider the effects of legitimacy when existing authority structures combine, as in the case of organizational mergers (Terry and O'Brien, 2001). They find that when

the organizations of varied status levels merge, and their status differences are legitimated, members from the low-status group are more likely to identify as members of the new organization, to view the new organization as a collective identity for members of both prior organizations, and to experience less threat and distress from the merger. Members of the higher status group, however, are less likely to see both groups as members of one common organization, and experience more distress during the merger. This pattern may emerge because high-status members are concerned that the merger may provide opportunities to challenge their status.

Consequences of Legitimacy of Deviance, Resource Distributions, and Discrimination

Scholars also focus on the consequences of the legitimation of deviance, unjust distributions of resources, and discrimination. While deviance is, by definition, behavior that violates the norms of a collective, deviant behavior may also be legitimated in certain contexts by smaller groups. For example, when people perceive that their peers support deviant behavior they are more likely to cheat, even if cheating behavior provides the same self-serving benefits in the presence or absence of endorsement from peers (Younts, 2008).

Scholars examining the perceptions of injustice to distributions of resources consider how the legitimacy of the outcome distribution may affect these perceptions. When resources are allocated in an unfair way, individuals are less likely to perceive the outcome as unjust if the distribution is legitimated (Hegtvedt and Johnson, 2000). Under these circumstances, disadvantaged actors will experience less severe negative emotions, and thus are less likely to respond behaviorally. When unjust distributions are highly legitimated, actors will respond to the injustice psychologically rather than behaviorally. Similarly, in organizations where promotional procedures are legitimated from above and below, workers are less likely to question the fairness of the rewards they receive from the company. In fact, legitimacy proves more important than self-interest in influencing perceptions of justice in these cases (Mueller and Landsman, 2004).

Social processes like discrimination can also be legitimated. For individuals who are members of groups targeted by discrimination, the effects of the discrimination vary based on whether or not individuals perceive the discrimination is legitimate. Group members who perceive the discrimination is illegitimate are less likely to experience self-directed anger, which provides a buffering effect between the discrimination and an individual's self-esteem. Alternatively, those group members who perceive the discrimination as legitimate are more likely to experience self-directed anger, exacerbating the effect of the discrimination. In these cases, the individual internalizes the idea that his or her group membership is a reasonable basis for the negative interaction (Hansen and Sassenberg, 2011). When discrimination is legitimated, individuals are less likely to identify with the targeted group, and therefore less likely to act collectively (Jetten et al., 2011). The inability to act collectively has clear implications for maintaining the status quo.

In summary, legitimacy of social structures and procedures has some positive outcomes in the form of maintaining social order and efficiency. The stability provided by legitimated structures is the basis for much of daily life. Unfortunately, legitimacy processes also perpetuate inefficient practices and inequalities by providing more support and allowing more agency for individuals with status advantages, limiting responses to procedural and distributive injustices, and undermining collective identity and action among groups targeted by discrimination.

Delegitimation

Legitimacy may be a driving force in maintaining the status quo, for better or worse, but the stability and consistency of legitimated social objects is not absolute. The legitimacy of problematic structures can be and is at times challenged, and this provides the opportunity for new status hierarchies, cultural schemas, practices, and procedures to emerge. For example, challenges to status hierarchies may arise both externally and internally. If contradictions emerge in the evaluation of a high-status member, for example, that actor may stand to be delegitimated in his or her status position (Berger et al., 1998). Also, a leader may enjoy authorization and endorsement within the organization, but if an authority figure from outside the organization gives a negative evaluation of the leader and his or her work, this calls into question the leader's fitness for the position internally as well.

Scholars focusing on intergroup relations examine how intergroup competition may bring about social change (Hogg, 2006). Under certain conditions, when status-disadvantaged groups believe their position is illegitimate or unstable, they may conceive of a different social order and engage in intergroup competition to bring about change in their status. Research conducted with East Germans who perceived their group as inferior to West Germans during the reunification highlights how legitimacy and stability affect intergroup competition (Mummendey et al., 1999). This study shows that separatists, who see the status differences between East and West Germans as stable and illegitimate, and view the boundaries as impermeable, are distinct from assimilators who see the status differences as unstable and legitimate, and view the boundaries as permeable. They find that separatists identify more strongly as East German, and are less inclined to leave their own group and identify with the recategorization as Germans. Additionally, separatists display more social competition in attempts to elevate the status of East Germans than do assimilators. These challenges to the legitimacy of a status hierarchy may not be sufficient to delegitimate the status hierarchy independently, but they serve as chinks in the armor of legitimacy, providing opportunities for new structures that have the potential to be more efficient and equitable.

The understanding of the process of delegitimation of a social object is limited, but it may be that this delegitimation process begins with a change in the structural conditions that currently underlie a social object, even if these conditions are different from the original ones at the early stage of innovation. Even so, does the delegitimation of a social object require being

replaced by a new social object? For example, must new organizational forms and practices be created and legitimated for old forms and practices to be perceived as inappropriate or inefficient? Can a social object lose widespread acceptance by an audience without anything to replace it? Finally, taken-for-granted notions of gender, race, sexual orientation, and religion may disappear, but can they disappear without being replaced by other taken-for-granted ideas that privilege some groups over others? These and many other questions remain, but the study of legitimacy as a social process, and the antecedents and consequences of legitimacy, also remains essential in understanding social life.

See also: Authority, Social Theories of; Groups, Sociology of; Organizations, Sociology of; Racial Relations; Social Psychology, Sociological; Status and Role, Social Psychology of; Status and Role: Structural Aspects.

Bibliography

Berger, J., Ridgeway, C.L., Fisek, M.H., Norman, R.Z., 1998. The legitimation and delegitimation of power and prestige orders. American Sociological Review 63, 379–405.

Blau, P.M., 1964. Exchange and Power in Social Life. John Wiley, New York.

Bobocel, D.R., Kay, A.C., Zanna, M.P., Olson, J.M., 2010. The Psychology of Justice and Legitimacy. Psychology Press of Taylor & Francis Group, New York.

Correll, S.J., Ridgeway, C.L., 2003. Expectation states theory. In: Delamater, J. (Ed.), Handbook of Social Psychology. Kluwer Acad./Plenum, New York, pp. 29–51.

Deephouse, D.L., Suchman, M., 2008. Legitimacy in organizational institutionalism. In: Greenwood, R., Oliver, C., Sahlin-Andersson, K., Suddaby, R. (Eds.), The Sage Handbook of Organizational Institutionalism. Sage Pub, Thousand Oaks, CA, pp. 49–77.

DiMaggio, P., 1992. Cultural boundaries and structural change: the extension of the high culture model to theater, opera, and the dance, 1900–1940. In: Lamont, M., Fournier, M. (Eds.), Cultivating Differences: Symbolic Boundaries and the Making of Inequality. Univ. Chicago Press, Chicago.

Dornbusch, S.M., Scott, W.R., 1975. Evaluation and the Exercise of Authority. Jossey-Bass, San Francisco.

Eagly, A.H., Carli, L., 2007. Through the Labyrinth: The Truth About How Women Become Leaders. Harvard Bus. Press, Cambridge, MA.

Foschi, M., 2000. Double standards for competence: theory and research. Annual Review of Sociology 26, 21–42.

Hansen, N., Sassenberg, K., 2011. Exploring the self-directed anger of the stigmatized: the interplay between perceived legitimacy and social identification. Group Process and Intergroup Relations 14, 807–818.

Hegtvedt, K.A., Johnson, C., 2000. Justice beyond the individual: a future with legitimation. Social Psychology Quarterly 63, 298–311.

Hegtvedt, K.A., Johnson, C., 2009. Power and justice: towards and understanding of legitimacy. American Behavioral Scientist 53, 376–399.

Hogg, M.A., 2006. Social identity theory. In: Burke, P.J. (Ed.), Contemporary Social Theories. Stanford Univ. Press, Stanford, CA, pp. 111–136.

Jetten, J., Schmitt, M.T., Branscombe, N.R., Garza, A.A., Mewse, A.J., 2011. Group commitment in the face of discrimination: the role of legitimacy appraisals. European Journal of Social Psychology 41, 116–126.

Johnson, C., Dowd, T.J., Ridgeway, C., 2006. Legitimacy as a social process. Annual Review of Sociology 32, 53–78.

Johnson, C., Ford, R., Kaufman, J., 2000. Emotional reactions to conflict: do dependence and legitimacy Matter? Social Forces 79, 107–137.

Lawler, E.J., 2006. The affect theory of social exchange. In: Burke, P.J. (Ed.), Contemporary Social Theories. Stanford Univ. Press, Stanford, CA, pp. 244–268.

Leblebici, H., Salancik, G.R., Copay, A., King, T., 1991. Institutional change and the transformation of interorganizational fields: an organizational history of the US radio broadcasting industry. Administrative Science Quarterly 36, 333–363.

Lucas, J.W., 2003. Status processes and the institutionalization of women as leaders. American Sociological Review 68, 464–480.

Molm, L.D., 2006. The social exchange framework. In: Burke, P.J. (Ed.), Contemporary Social Theories. Stanford Univ. Press, Stanford, CA, pp. 24–45.

Mueller, C.W., Landsman, M.J., 2004. Legitimacy and justice perceptions. Social Psychology Quarterly 67, 189–202.

Mummendey, A., Klink, A., Mielke, R., Wenzel, M., Blanz, M., 1999. Socio-structural characteristics of intergroup relations and identity management strategies: results from a field study in East Germany. European Journal of Social Psychology 29, 259–285.

Ridgeway, C.L., 1991. The social construction of status value: gender and other nominal characteristics. Social Forces 70, 367–386.

Ridgeway, C.L., 2011. Framed by Gender: How Gender Inequality Persists in the Modern World. Oxford Univ. Press, Oxford, UK.

Ridgeway, C.L., Johnson, C., Diekema, D., 1994. External status, legitimacy, and compliance in male and female groups. Social Forces 72, 1051–1077.

Rodriguez-Bailon, R., Moya, M., Yzerbyt, V., 2000. Why do superiors attend to negative stereotypic information about their subordinates? Effects of power and legitimacy on social perception. European Journal of Social Psychology 30, 651–671.

Roy, W.G., 1997. Socializing Capital: The Rise of the Large Industrial Corporation in America. Princeton Univ. Press, Princeton, NJ.

Smith-Lovin, L., 2007. The strength of weak identities: social structural sources of self, situation and emotional experience. Social Psychology Quarterly.

Terry, D.J., O'Brien, A.T., 2001. Status, legitimacy, and ingroup bias in the context of an organizational merger. Group Process and Intergroup Relations 4, 271–289.

Tyler, T.R., 2001. A psychological perspective on the legitimacy of institutions and authorities. In: Jost, J., Major, B. (Eds.), The Psychology of Legitimacy: Emerging Perspectives on Ideology, Justice, and Intergroup Relations. Cambridge Univ. Press, Cambridge, UK, pp. 416–436.

Walker, H.A., 2004. Beyond power and domination: legitimacy and formal organizations. Research in the Sociology of Organizations 22, 239–271.

Weber, M., 1978(1924). In: Roth, G., Wittich, C. (Eds.), Economy and Society, vols. I and II. Univ. Calif. Press, Berkeley.

Younts, C.W., 2008. Status, endorsement and the legitimacy of deviance. Social Forces 87, 561–590.

Zeldtich Jr., M., 2006. Legitimacy theory. In: Burke, P.J. (Ed.), Contemporary Social Theories. Stanford Univ. Press, Stanford, CA, pp. 324–374.

Zelditch Jr., M., Walker, H.A., 1984. Legitimacy and the stability of authority. Advances in Group Process 1, 1–25.

Leisure Activities Choices among Adolescents

Maria K Pavlova and Rainer K Silbereisen, Center for Applied Developmental Science (CADS), Friedrich Schiller University Jena, Jena, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.K. Silbereisen, volume 1, pp. 122–125, © 2001, Elsevier Ltd.

Abstract

In this article, we provide descriptive time-use data on adolescent leisure. Next, we link normative developmental tasks or concerns of adolescence to the context of leisure. We then outline other, nondevelopmental factors of adolescents' leisure participation, such as external constraints, individual motivation, and personality differences. Finally, we review the methods used to assess leisure participation and motivation in adolescence, touch upon the implications of leisure for adolescent health and development, and outline possible avenues of intervention.

Leisure is typically understood by researchers as freely chosen and enjoyable activities that are pursued during one's discretionary time (Kleiber, 1999). Adolescence (here loosely defined as the second decade of life) is a developmental period when growing autonomy fosters (and is fostered by) independent decision making, making the study of leisure choices particularly relevant. However, beyond individual choices, sociologists, psychologists, and researchers in tourism and recreation have studied a broader spectrum of reasons or causes underlying participation in particular leisure activities. In this article, we summarize these diverse perspectives. First, we describe time-use data on adolescents' leisure behaviors. Second, we classify leisure activities in terms of the common psychological functions of leisure. Third, we link normative developmental tasks or concerns of adolescence to the context of leisure. Fourth, we summarize potential nondevelopmental influences on adolescents' leisure participation. Finally, we review the methods used to assess leisure participation and motivation in adolescence, the implications of leisure for adolescent health and development, and possible avenues of intervention. As the relevant literature is predominantly based on Western, mostly U.S. populations, this article represents these populations better than the rest of the world, but we introduce a cross-cultural perspective wherever possible.

Leisure Time Use in Adolescence

Leisure time available to adolescents varies across regions of the world, due to differences in school enrollment, in the hours spent in school, and in adolescent involvement in productive work (Larson and Verma, 1999). Working adolescents from nonindustrial, unschooled populations report having 4–7 free hours per day; across European nations, adolescents report having 5.5–7.5 free hours per day; in East Asian postindustrial countries where school achievement is important (Nishino and Larson, 2003), adolescents have only 4–5.5 free hours per day (estimates averaged across school days and weekends; more recent data are not available for these countries). By far, U.S. adolescents enjoy the largest amount of free time: Older adolescents have approximately 8 free hours per day, whereas older Canadian adolescents report having 6 free hours per day.

Most Prevalent Activities

Where TV and other mass media are widely accessible, electronic media use (passive activities, such as watching TV and listening to music, as well as more mentally challenging undertakings, such as Internet surfing, video games, and communicating with others) is the most popular leisure activity among adolescents: In the US, for example, a Kaiser Family Foundation study revealed 4 h per day spent watching TV, 3 h spent using computer, 3 h of music listening, 2 h of talking and texting via cell phones, and 0.5 h of reading. About a quarter of this time involves multitasking – using more than one medium at a time; thus, pure media consumption is less than suggested by these estimates.

A sizable portion of adolescent free time is spent in public places, such as parks, cafes, movie theaters, and shopping malls, and at somebody else's home. Visiting, hanging out, shopping, and dating, all forms of socializing, mostly with peers, account for 1–1.5 free hours per day among adolescents in the US, Canada, and the UK. However, in Islamic and East Asian societies, where adults impose restrictions on adolescent peer interactions and do not encourage seemingly purposeless pastime, considerably less time is spent socializing with peers, especially other-sex peers.

Sports and other forms of physical activity are also popular, accounting for more than 1 h of free time in the US and Norway, 0.5–1 h in Europe, Canada, and Australia, and less than 0.5 h in East Asian countries. Higher levels of physical activity are common in rural, nonindustrial populations because of ample opportunities to work or play outdoors.

In urban regions and postindustrial countries, adolescents can take part in structured extracurricular activities such as organized sports, school clubs, youth organizations, performing arts opportunities, and volunteering (Mahoney et al., 2005). Despite the variety of options and the fact that adolescents can engage in more than one extracurricular activity at a time, such activities take less than 0.5 h of daily time even in

the US, where extracurricular programing is most developed, partly because not all adolescents participate in such programs. In less affluent regions or in societies where adolescents are supposed to spend the better part of their free time with the family, extracurricular activities are uncommon.

Various productive and intellectual activities, such as performing arts, playing chess, gardening, and cooking, may be pursued individually (or at least not in structured settings) as hobbies. Time-use estimates are difficult to obtain here because of the unsupervised and unsystematic nature of many such activities, but available evidence suggests that they take a relatively small proportion of adolescents' time on average. For instance, time playing a musical instrument ranges from 5 to 30 min per school day in the US and Europe. Finally, although it is disputable whether religious activities, praying in particular, should be regarded as leisure or rather as obligatory activities, they constitute a salient time-use category in religious populations, accounting for nearly 10 min of adolescent time daily in the US and Canada and more than 1 h per day in Indonesia.

A sizable proportion of adolescents' waking hours is spent alone (20–25% in rural regions and 30–35% in urban regions). These estimates include time spent on homework and self-care, and perhaps on what could be considered leisure activities. Time alone is most often spent in one's own bedroom (if one has a private bedroom), although solitary walks are also quite common where the neighborhood is safe and attractive natural settings are available.

Activity Portfolios

Studies conducted in the US typically yield two opposite patterns of activity participation: Highly involved adolescents report being engaged in many activity domains (doing sports and taking part in other extracurricular activities, socializing with peers, reading, using electronic media, etc.) and sometimes also working for pay. In contrast, uninvolved adolescents are usually not employed and report predominantly unstructured and rather passive leisure activities, such as watching TV, hanging out with friends, and sometimes pursuing individual hobbies. The highly involved group includes 11% to more than 50% of adolescents, depending on the number of activities taken into account. The uninvolved group makes up 12–25% of adolescents studied. Between these two extremes, a large number of adolescents focus on one or two activity domains. Across the board, socializing with peers is very common, with only a very small minority of adolescents reporting little or no peer interaction.

Demographic Differences

Because of their typically heavier domestic workload and greater diligence at school in many countries, female adolescents report having less discretionary time than male adolescents. Female adolescents also tend to engage in different leisure pursuits than their male peers do. Across societies, female adolescents spend less time than their male peers participating in sports, playing video games, and watching TV, but more time reading and engaging in performing arts. In Western societies with little restriction on peer interactions, females devote more time than males to socializing with peers. In contrast, in Islamic and East Asian countries, female adolescents spend less time with peers than males.

In the US, marked racial and ethnic differences also exist: Latino and African-American adolescents, who are often economically disadvantaged, are less likely to participate in extracurricular activities than European-American adolescents and spend more time watching TV. African-American adolescents are particularly less involved in sports and spend more free time with family members and less time with peers than European-American adolescents.

As regards age differences, in Western samples, there is a decline in TV viewing from early to late adolescence in favor of music listening and talking on the phone and a decrease in sports and other organized activity participation and time spent with the family in favor of unsupervised time spent alone and with peers.

Leisure Types and Functions

Various leisure types have been described. One dichotomy is between active and passive leisure. Active leisure involves some challenge, combines effort and enjoyment, and thereby brings opportunities for learning and skill building, be it social, musical, motor, or any other skills and competencies. Passive leisure encompasses a broad range of nonchallenging activities, from watching entertainment programs on TV to idling. Its main functions are pleasure, relaxation (often in the context of coping with stress; Nishino and Larson, 2003), and filling free time. A more narrow distinction, used in the public health literature, is between physically active and sedentary leisure.

Stebbins (2007) divided leisure into casual, serious, and project-based activities: Serious leisure activities are skill-based, require considerable effort and persistence, and provide durable benefits, such as self-actualization, self-expression, and self-enrichment, as well as identification with a group of like-minded people and social recognition. Project-based leisure includes short-term, moderately complicated, creative undertakings that require planning, effort, and skill but are not associated with a long-term commitment. Finally, casual leisure includes pleasurable activities that require little special training to enjoy them. Apart from its purely hedonic and recreational functions, it fosters the development and maintenance of interpersonal relationships, as many social interactions occur in the context of casual leisure (e.g., a family dinner).

A conceptually similar distinction has been made between structured and unstructured leisure during childhood and adolescence. Structured leisure occurs in organized settings, such as schools or sports clubs, is supervised by adults, is performed in a group of youth, and otherwise fulfills the criteria for active leisure. It is therefore similar to serious and project-based leisure, with an emphasis on adult supervision and guidance that vary from quite directive to more subtle as when an adult acts as a facilitator and youth are allowed to make decisions and plan their activities themselves (Mahoney et al., 2005). Unstructured leisure may involve both active (e.g., hobbies) and passive (e.g., hanging out) leisure undertakings and may take place both with and without adult supervision;

its distinctive feature is a lack of an agenda and organization, analogous to casual leisure within Stebbins's classification.

Adolescent Development and Leisure

Adolescent leisure activities are both a manifestation of emergent developmental tasks and concerns and a context that offers opportunities for these tasks and concerns to be resolved (Kleiber, 1999). In this section, we discuss both of these.

Adolescent Brain

Adolescents are notorious for their propensity to engage in risky behaviors (e.g., unprotected sex, substance abuse, and drunk driving), especially in the context of unstructured and unsupervised leisure in a company of peers. Recent research in developmental neurobiology suggests that risk taking in human adolescents may be linked to brain maturation during this period. More specifically, whereas the cognitive control system (associated with cortical brain regions, e.g., prefrontal cortex) matures gradually from childhood into young adulthood, more rapid changes in the incentive processing system (associated with limbic subcortical regions, e.g., nucleus accumbens) occur during adolescence, which entail heightened reactivity to rewards, especially social rewards. As a result, the adolescent cognitive control system, although more mature than that in children, is not yet fully developed and may not adequately inhibit impulsive responses to immediate rewards. At the behavioral level, this nonlinear developmental dynamics may lead to adolescents performing well on logical reasoning and cognitive estimation of risks but still making risky choices in emotionally laden, potentially rewarding situations that are most likely to occur during peer interactions. If such biological factors do, to some extent, influence adolescents' leisure preferences, this may be especially true in societies in which adolescents have a great deal of leisure time and relatively low levels of adult supervision.

Identity Development

Erik Erikson considered adolescence as a critical period in the formation of personal and social identities, which occurs through exploration of different options and subsequent commitment to some of them. The context of leisure offers ideal opportunities for identity exploration, as leisure presumes more freedom and fewer obligations than school or paid work. Alan Waterman suggested that in the process of identity formation, some adolescents may follow the discovery metaphor, attempting to find their 'true selves.' They seek to engage in 'personally expressive' activities that fit with their hidden potentials and are experienced as both challenging and enjoyable (cf the concept of flow below). Indeed, several empirical studies suggest that sustained involvement in structured or serious leisure (e.g., sport, volunteering) is associated with a more mature identity and with personal growth experiences. In addition, voluntary solitude and quiet contemplation may be used strategically by adolescents to find their 'true selves.' Interestingly, solitude and privacy are increasingly desired by adolescents, perhaps allowing them to arrange a space (e.g., private bedroom) that reflects their unique personalities.

In contrast, the creation metaphor implies that there is no 'true self' to discover but that adolescents adaptively construct their selves, fitting their personae to the available options and expectations of others. Identity creation requires more extensive exploration and experimentation with possible selves, as well as more feedback from others, than identity discovery. Again, leisure settings provide excellent venues to create various potential identities. The creation metaphor may help us understand why many adolescents try to adopt a particular image (e.g., with respect to clothes, musical taste, and favorite movies), abruptly change their images, are fascinated with interactive media, and sometimes enjoy playing with different identities in anonymous chat rooms or in online role-playing games.

But is either the discovery or the creation of an identity good for adolescents? It depends on the breadth of exploration and on the social desirability of choices made. In fact, the quest for identity can be associated with maladaptive leisure choices (Kleiber, 1999). For instance, there is some evidence that adolescents with low self-concept clarity and poor social skills may overuse the Internet, probably in an attempt to overcome these difficulties. Moreover, members of youth gangs have a strong sense of social identity, and their delinquent activities are both challenging and enjoyable to them. Similarly, overinvestment in and overidentification with a single activity (e.g., fandom) may satisfy one's need for identity but limit personal development. Thus, having access to diverse leisure opportunities that are both enjoyable and socially desirable is vital to normative identity development in adolescence.

Parents

Throughout childhood, parents serve as primary role models concerning leisure activities; many adolescents continue to pursue leisure interests that are similar to those of their parents. Joint leisure activities also remain quite common. Drawing on their beliefs about their children's abilities and needs, parents may encourage or discourage participation in certain activities and provide corresponding facilities. Parents may also influence leisure choices of their offspring indirectly, through their general parenting style. In particular, authoritative parenting, which combines warmth, autonomy support, structure, and discipline, fosters the development of social competence and responsibility and thereby promotes healthy leisure choices in adolescents. In contrast, when adolescents feel controlled and mistrusted or neglected and unguided, they may be drawn toward passive leisure and unstructured, unsupervised leisure in the company of deviant peers.

Peers and Youth Subcultures

As peers and peer groups gain in importance (particularly in Western societies), adolescents choose leisure activities that make them popular among peers (e.g., skateboarding), are performed by friends and romantically attractive peers, and/or give them opportunities to socialize with peers without

adult supervision, whereas adult-supervised activities, even interesting and useful ones, gradually lose their appeal. Peer-focused activities can have both positive and negative consequences, depending on the values of the peer group. In addition, adolescents may avoid activities where they expect or experience peer rejection and victimization (e.g., overweight teenagers often withdraw from organized sports).

Adolescents are exposed to and attracted by broader youth subcultures or lifestyles such as those exemplified by high school 'crowds' (e.g., brains, jocks, and Goths). Youth subcultures are often associated with a distinct outward appearance (especially clothing), jargon, musical style, and leisure preferences. Belonging to a particular subculture therefore defines a whole range of behaviors. Initially, it was thought that subcultures emerged among working-class youth in response to social problems, as a symbolic means of protest. Now that youth subcultures and lifestyles have become commercialized and widely popularized in media, they are rather regarded as a way to celebrate one's distinctness (as a group and as an individual) and to have fun.

Changing Relational Concerns

John Coleman and Leo Hendry's focal theory linked age-graded relational concerns with leisure transitions during adolescence. According to Hendry, adult-organized leisure activities of early adolescence are displaced by casual (unstructured) leisure pursuits in middle adolescence, activities that mainly consist of socializing with peers and reflect developmental concerns about peer acceptance and rejection. In late adolescence, casual leisure gives way to commercialized leisure (e.g., visiting pubs, discotheques, sport clubs, or wellness centers), which adolescents use to affirm their maturity and independence from parents. Although empirical studies yield mixed support for this particular sequence of events (in affluent societies, commercialized leisure currently dominates even among younger children), the focal theory underscores the importance of changing developmental concerns in leisure activities choice. In a similar vein, Silbereisen et al. (1992) showed that adolescents' choice of public or private leisure contexts depends on their progress in romantic friendships. For instance, adolescents hoping for a relationship prefer public hangout places, where they can meet potential romantic partners, but once a relationship has been initiated, couples retreat to more private leisure settings.

Autonomy Development

Achieving an optimal level of psychosocial autonomy from both parents and peers is a major developmental task of adolescence. Leisure becomes truly leisure only when individuals can freely choose their activities; therefore the context of free time becomes an important arena (and sometimes a battlefield) of autonomy development (Kleiber, 1999). Whereas unstructured and unsupervised peer time serves as an escape from adult supervision, time spent alone helps in keeping a distance from peers as well. By experimenting with and combining different leisure options, adolescents gradually learn to balance the conflicting expectations of parents and peers with their own emerging interests.

Gender Role Development

Socialization into masculine and feminine gender roles and the formation of gender role identity begin in early childhood. Gender-typed behaviors (e.g., playing with dolls or with cars) first emerge in the family context and are then reinforced in gender-segregated peer groups throughout childhood and adolescence. Although puberty is sometimes characterized as a period of gender roles intensification, in general, gender differences in adolescent leisure activities reflect a continuation rather than an abrupt increase in gendered behavior patterns. Furthermore, large interindividual differences exist in the extent to which gender role stereotypes are acquired and followed throughout childhood and adolescence. Last but not least, gender typing is a cultural phenomenon; for instance, Muslim cultures hold very strict notions about the gender appropriateness of particular activities, whereas gender-related expectations are much looser in contemporary Western societies.

Understanding Individual Differences in Leisure

In this section, we introduce major explanations of leisure preferences and participation that apply to adults and adolescents alike. Some of these theoretical accounts offer a developmental perspective, some do not. We will illustrate them with examples from adolescent research.

The Ecology of Leisure: Constraints–Facilitators Framework

In leisure research, the hierarchical model of leisure constraints (Crawford et al., 1991) is commonly used to explain variations in leisure participation. Constraints are objective or perceived factors that inhibit formation of leisure preferences, prevent individuals from enjoying a particular kind of leisure, or hinder them from participation altogether. Intrapersonal constraints represent individual barriers to leisure participation, such as a lack of interest in a particular activity, a belief that one's skills and abilities are insufficient, a lack of knowledge about available leisure opportunities, and health conditions (e.g., disability). Interpersonal constraints explain the lack of participation at the level of relationships and group attitudes, such as a lack of company and disapproval from significant others. Structural constraints include institutional and economic barriers, such as a lack of time and money, poor availability of leisure facilities, unsafe neighborhood, and legal restrictions. Finally, sociocultural constraints, such as restrictions on women's participation in certain leisure activities, have recently been suggested as an additional level of analysis. The constraints model has been criticized for its implicit assumption that nonparticipation is always bad and for its failure to consider the role of resources and other positive factors eliciting motivation, promoting formation of leisure preferences, and encouraging leisure participation, factors that have been called facilitators to leisure.

In research based on the constraints model, adolescents report negotiating constraints to leisure at all levels except for the sociocultural, a level of which they seem unaware. Typical examples include disinterest and perceived lack of skills (intrapersonal); parents imposing restrictions on youth's

leisure activities and friends being uninterested in certain activities (interpersonal); and inadequate recreational facilities, financial constraints, limited accessibility of attractive leisure sites, and dangerous neighborhoods (structural). Adolescents' subjective perceptions aside, structural and sociocultural constraints, such as race, gender, and socioeconomic divisions in adolescent leisure participation, are well documented. For instance, structured extracurricular activities are not available in many communities and are most accessible to white, middle-class, American adolescents (Mahoney et al., 2005).

Quality of Experience and Leisure Motivation

If individuals pursue free time activities they truly enjoy, they are believed to be intrinsically motivated. Scholars distinguish between activities that bring hedonic enjoyment (i.e., pleasure and relaxation; typical activities are socializing and watching movies) and those that bring eudaimonic enjoyment (i.e., self-expression and personal growth; typical activities are volunteering, sports, and performing arts). Although both types of activities are needed to live a 'full life,' individuals differ in how much value they place on hedonic and eudaimonic enjoyment. Mihály Csíkszentmihályi and colleagues conducted landmark research on flow experiences (i.e., a match between a high challenge of the activity and a high skill of the performer, which generates the feelings of intense involvement, alertness, and happiness). These studies showed that some people are more likely than others to seek out such activities and to experience flow ('autotelic personality').

However, individuals sometimes perform free time activities they do not enjoy much. For instance, many common leisure activities, such as watching TV, are associated with rather negative emotional states (apathy). Edward Deci and Richard Ryan's self-determination theory (SDT) helps understand how individuals become involved in unenjoyable activities. According to SDT, wherever people are not intrinsically motivated, they may still perform an activity for a variety of extrinsic reasons. SDT describes four levels of goal internalization ranging from purely external control by others to integrated regulation, where an individual voluntarily pursues the activity and recognizes it as important to the self. Finally, individuals may perform an activity for no particular reasons (amotivation).

According to SDT, both extrinsically motivated and amotivated behaviors may occur when individuals' basic psychological needs for autonomy, competence, and relatedness are thwarted; such activities are usually experienced as boring or too demanding. For instance, the evaluative and competitive aspects of sports participation may undermine some individuals' intrinsic enjoyment of sports and exercise, leading to dropout among many adolescents. In contrast, coaches who support autonomy (i.e., give freedom of choice and focus on mastery instead of competition) can facilitate the participants' intrinsic motivation for sports and continued involvement. Although adolescents may become involved in a given activity, such as community service, for extrinsic reasons (e.g., pressure from parents or peers), under the right conditions (i.e., if participation in the activity satisfies their psychological needs), they may recognize the importance of this activity for themselves or even discover intrinsic enjoyment in it.

Some, but not all, adolescents seem to be especially prone to boredom (Hunter and Csikszentmihalyi, 2003). Complaints of being bored, both at school and in free time, can become more frequent as young people progress from early to late adolescence. As predicted by SDT, boredom is associated with feelings of controlled motivation (based on poorly internalized goals) and amotivation. As adolescence is the period when individuals should begin to exercise more autonomy in their daily life, chronic boredom in the context of leisure may signal that this developmental task is not being supported. Researchers have offered several explanations for adolescent amotivation and boredom, including excessive adult control and a lack of skills to organize one's leisure activities in a satisfying and productive way. It appears that adolescents still need a lot of support from adults in organizing their leisure (e.g., structured activities are typically experienced as interesting), but this support should be provided in an unobtrusive way so that the need for autonomy is not thwarted. Although cultures differ in the amount of autonomy granted to adolescents by their parents, the basic link between autonomy in activity choices and activity enjoyment has been replicated across cultures.

Drawing on SDT, Robert Vallerand and colleagues distinguished between harmonious and obsessive passion for activities. Passion is opposite to amotivation and is characterized by high investment, valuing, and liking of the activity. Whereas harmoniously passionate individuals feel free to perform or not to perform their favorite activity, those obsessively passionate often feel urged or obliged to perform the activity, which entails less enjoyment and a rigid persistence in pursuing it. Both types of passion develop over time if individuals invest much time and effort in a given activity, highly value it, and strongly identify with it. Furthermore, harmonious passion is more likely to develop in the context of autonomy support, whereas obsessive passion emerges in a controlling context where there is little freedom of choice and approval by others depends on persistence and performance. Children's and parents' beliefs that one should give up on other pursuits in order to excel in a chosen activity also foster the development of obsessive rather than harmonious passion (Mageau et al., 2009).

Expectancies for Success and Leisure Motivation

Participation in achievement-related and skill-based types of leisure, such as sports, artistic activities, and career volunteering, is especially contingent on one's expectations to succeed. In his social cognitive theory of motivation, Albert Bandura distinguished between two kinds of expectancies: outcome expectations (beliefs that certain behaviors can lead to desired outcomes) and self-efficacy beliefs (individuals' judgments concerning their ability to organize and perform a given behavior to produce a given outcome, even in the presence of obstacles). Self-efficacy beliefs are based on one's own past experiences, on the feedback received from others, and on the observations of others' successes or failures on similar tasks. In the leisure context, higher self-efficacy regarding sports and physical activity is related to higher physical activity involvement among adolescents and moderates the negative effects of

other factors, such as perceived low accessibility of recreational facilities. High self-efficacy is also required to abstain from unhealthy leisure choices, for instance, higher self-efficacy in resisting peer pressure is related to a lower likelihood of substance abuse in adolescence.

An integrative model of achievement-related motivation and its developmental precursors was suggested by Jacquelynne Eccles and colleagues. In their expectancy–value model, activity choices are predicted by expectancies for success and subjective task value, which combines intrinsic enjoyment of the activity with its subjective utility, its centrality to one's identity and core goals, and cost. Further, through a chain of psychological and social mediators, expectancy–value beliefs are shaped by past experiences and by beliefs and behaviors of socializing agents (e.g., parents), who are themselves influenced by a broader socioeconomic and cultural context.

To take an example, because of gender role stereotypes, parents recognize and foster athletic aptitudes in their sons more readily than in their daughters, and in this way girls grow to be less confident about their athletic performance and less interested in sports activities. Furthermore, several studies (e.g., Simpkins et al., 2012) showed that parental beliefs and behaviors are even stronger predictors of children's and adolescents' skill-based leisure activities, such as sports, instrumental music, and out-of-school computer activities, than of their academic activities, probably because the former are not as well incorporated into the school curriculum and require more parental inputs. Children's and adolescents' own expectancies for success and task values predict such leisure activities. Findings from longitudinal studies suggest that both expectancies for success and subjective task values in achievement-related activities, such as sports, decline steadily from childhood through adolescence, in part because achievement-related leisure becomes increasingly competitive as children grow up. As denied entry onto a team is likely to be experienced as a personal failure, rejected adolescents downgrade the importance of such leisure pursuits and drop out of such competitive activities. Thus, the expectancy–value model concurs with SDT in that the attitudes and behaviors of socializing agents and the framing of structured activities (e.g., as competitive or collaborative) are important determinants of adolescents' leisure motivation.

Personality and Leisure

Although situations and contexts are important determinants of leisure motivation, stable individual differences in leisure preferences also exist. First, individuals differ in the content of activities that appeal to them. The very definitions of some personality traits include preferences for leisure activities. Unsurprisingly, extroverted individuals prefer social activities and, to a lesser extent, sports, whereas introverted individuals are more drawn toward the arts. High openness to experience is related to the preference for artistic and cultural activities. High neuroticism is an obstacle to leisure participation and enjoyment in general, but especially with regard to competitive sports, which require high emotional stability. High sensation seeking is associated with a range of leisure activities, such as adventurous and dangerous sports, outdoor activities, and illegal leisure (e.g., drug use and drunk driving). Boredom proneness is related to sensation seeking, which explains why frequent boredom forebodes the involvement in risky behaviors, particularly among adolescents. In contrast, a resilient and prosocial personality predicts involvement in volunteering and community service. Finally, specific individual abilities predict engagement in corresponding aptitude-based activities, such as sports, playing music, and intellectual activities; accordingly, such leisure pursuits also yield relatively high heritability estimates.

Second, individuals differ in their ability to experience intrinsic enjoyment of leisure and in the type of enjoyment they look for (hedonic vs eudaimonic). By adolescence, the striking difference between chronically interested and chronically bored individuals becomes apparent. Interested adolescents report higher and more stable self-esteem, a more internal locus of control, and greater optimism than their bored counterparts (Hunter and Csikszentmihalyi, 2003). In adults, proneness for flow is related to high conscientiousness and low neuroticism. Although temperamental bases of motivational orientations may exist, socializing influences are also key. For example, authoritative parenting and parental role modeling in childhood predict flow experiences and eudaimonic versus hedonic pursuits in adolescents and young adults (Huta, 2012).

Assessing Leisure Preferences and Participation in Adolescence

Methods of assessing adolescent leisure participation or nonparticipation differ depending on the research purpose. In particular, behavioral indicators capture what adolescents actually do in their free time, measures of leisure preferences or interests reflect what they would like to do, measures of constraints to participation investigate why they are not involved in a given leisure activity, and measures of underlying motivation reveal why they do what they do. Behavioral indicators are usually time-use data collected through survey questions (i.e., respondents estimate how much time they typically spend on a given activity), time diaries (i.e., over the course of several days, respondents note what they are doing or recall what they were doing on the previous day), or experience sampling techniques (i.e., for a number of days, respondents receive random signals from a mobile device prompting them to note immediately what they have been doing).

Measures of leisure preferences typically sample a wide range of leisure activities and ask adolescents to rate their liking/disliking of them, similarly to vocational interest scales. For example, Hansen's 250-item Leisure Interest Questionnaire consists of 20 scales, such as travel, culinary pursuits, and computer activities, and can be administered to late adolescents. King et al.'s Preferences for Activities in Children scale is a shorter instrument designed for children and adolescents with disabilities, which consists of 55 items divided into two activity domains (formal and informal) and five activity types (recreational, active physical, social, skill-based, and self-improvement). It complements another instrument measuring actual participation, Children's Assessment of Participation and Enjoyment; both are applicable to individuals without disabilities.

Measures of perceived constraints to leisure are based on Crawford et al.'s hierarchical model and assess intrapersonal, interpersonal, and structural constraints. Instruments assessing leisure motivation are usually tailored to specific theories of motivation. For instance, the 20-item Free Time Motivation Scale for Adolescents draws on SDT to assess five types of motivation ranging from amotivation to intrinsic motivation. The Passion Scale adapted for children and adolescents consists of 11 items and is used to distinguish between harmonious and obsessive passion for activities. Expectancy–value scales based on Eccles et al.'s theory measure expectancies for success and task values associated with specific achievement-based activities. Finally, qualitative methods for the assessment of leisure participation and motivation, such as focus groups, are also increasingly used.

Consequences of Leisure and Ways to Intervene

What adolescents do in their free time has short- and long-term implications for their health and psychosocial development. Participation in high-quality structured after-school programs is often enjoyable, predicts higher academic success, and provides developmental experiences in a variety of domains, such as identity work, initiative, emotional regulation, teamwork and social skills, positive relationships, and social capital (Feldman Farb and Matjasko, 2012; Mahoney et al., 2005). Programs where adults provide indirect guidance and adolescents feel in control of their activities facilitate the development of agency skills, for instance, strategic thinking. However, not all structured programs are of high quality, and involvement in low-quality programs may engender negative experiences leading to dropout, which should not be considered a negative outcome in this case (Rose-Krasnor, 2009). In turn, unstructured and unsupervised peer time, which is a leisure option highly attractive to Western adolescents, provides opportunities for self-exploration, autonomy development, and development of social competence, but also carries risks, such as substance use, unprotected sex, and delinquency. That is, just like structured activities, unsupervised peer interactions may be of different quality. Likewise, media-based leisure pursuits, such as watching TV, Internet surfing, and playing video games, may have both positive and negative effects on health and development depending on their particular content and the amount of time devoted to them.

Numerous programs have been developed to prevent risky behaviors in adolescence, such as substance use and unprotected sex, but few have focused on promoting healthy, developmentally beneficial choices and behaviors in the context of free time. A noteworthy exception is the TimeWise program developed by Linda Caldwell and colleagues (Caldwell et al., 2004), which aims at fostering general leisure skills and optimal motivation. TimeWise is partly based on SDT, consists of six curricular lessons delivered in Grade 7 classrooms in the US, and has already yielded positive short-term effects. Recently, it has been applied in South Africa as part of a larger prevention program.

Yet another approach to intervening with adolescents' leisure is to encourage their participation in specific desirable activities, most commonly sports and physical exercise. School-based interventions to promote physical activity among adolescents have rather short-term effects (to produce sustained effects, an intervention should last at least several months) and have limited generalizability to leisure contexts. More comprehensive interventions, for instance, those involving parents, tend to be more effective.

Conclusion

Adolescent leisure choices are influenced by a multitude of factors, many of which, such as structural facilities, financial constraints, and the attitudes of socializing agents (e.g., parents, teachers, and peers), adolescents cannot control. Nevertheless, to benefit from leisure experiences, adolescents need to exercise control – to be free to choose and plan their leisure, to balance different types of leisure activities, and to refrain from risky leisure pursuits. This requires organization and self-regulation skills as well as the ability to enjoy leisure. Where such skills and abilities are lacking, it is not the fault of adolescents themselves but rather the indication that their socializing agents lack these same skills and abilities or do not create a supportive and challenging environment optimal for psychosocial development. Future researchers would therefore be well advised to develop comprehensive interventions fostering leisure skills and optimal leisure experiences in children and adolescents, for example, by including corresponding components in family and community intervention programs.

See also: Academic Achievement Motivation, Development of; Eudaemonism; Flow in Motivational Psychology; Interest, Psychology of; Leisure, Sociology of; Media Effects on Children; Motivation in Youth Sport and Physical Activity: Developmental Perspectives; Motivational Development, Systems Theory of; Self and Identity Development During Adolescence across Cultures; Self-Determination Theory; Self-Efficacy; Social and Emotional Development in the Context of the Family; Time Use and Gender; Youth Culture, Sociology of.

Bibliography

Caldwell, L.L., Baldwin, C.K., Walls, T., Smith, E., 2004. Preliminary effects of a leisure education program to promote healthy use of free time among middle school adolescents. Journal of Leisure Research 36, 310–335.

Crawford, D.W., Jackson, E.L., Godbey, G., 1991. A hierarchical model of leisure constraints. Leisure Sciences 13, 309–320.

Feldman Farb, A., Matjasko, J.L., 2012. Recent advances in research on school-based extracurricular activities and adolescent development. Developmental Review 32, 1–48.

Hunter, J.P., Csikszentmihalyi, M., 2003. The positive psychology of interested adolescents. Journal of Youth and Adolescence 32, 27–35.

Huta, V., 2012. Linking people's pursuit of eudaimonia and hedonia with characteristics of their parents: parenting styles, verbally endorsed values, and role modeling. Journal of Happiness Studies 13, 47–61.

Kleiber, D.A., 1999. Leisure Experience and Human Development: A Dialectical Interpretation. Basic Books, New York.

Larson, R.W., Verma, S., 1999. How children and adolescents spend time across the world: work, play, and developmental opportunities. Psychological Bulletin 125, 701–736.

Mageau, G.A., Vallerand, R.J., Charest, J., Salvy, S.-J., Lacaille, N., Bouffard, T., Koestner, R., 2009. On the development of harmonious and obsessive passion: the role of autonomy support, activity specialization, and identification with the activity. Journal of Personality 77, 601–645.

Mahoney, J.L., Larson, R.W., Eccles, J.S. (Eds.), 2005. Organized Activities as Contexts of Development: Extracurricular Activities, After-School and Community Programs. Lawrence Erlbaum Associates, Mahwah, NJ.

Nishino, H.J., Larson, R., 2003. Japanese adolescents' free time: *Juku*, *Bukatsu*, and government efforts to create more meaningful leisure. New Directions for Child and Adolescent Development 99, 23–35.

Rose-Krasnor, L., 2009. Future directions in youth involvement research. Social Development 18, 497–509.

Silbereisen, R.K., Noack, P., von Eye, A., 1992. Adolescents' development of romantic friendship and change in favorite leisure contexts. Journal of Adolescent Research 7, 80–93.

Simpkins, S.D., Fredricks, J.A., Eccles, J.S., 2012. Charting the Eccles' expectancy-value model from mothers' beliefs in childhood to youths' activities in adolescence. Developmental Psychology 48, 1019–1032.

Stebbins, R.A., 2007. Serious Leisure: A Perspective for Our Time. Transaction Publishers, New Brunswick, NJ.

Steinberg, S., Parmar, P., Richard, B. (Eds.), 2006. Contemporary Youth Culture: An International Encyclopedia. Greenwood Press, Westport, CT.

Leisure and Cultural Consumption: The European Perspective

Philippe Coulangeon, Observatoire Sociologique du changement (OSC), SciencesPo/CNRS, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Leisure and cultural consumption are today considered key components in a person's – or a nation's – quality of life. Despite the historic decline in the hours spent at paid labor, current trends in leisure time are unclear. Recent developments suggest a possible halt in the road toward the leisure society and a reversal of the leisure/status gradient. However, the sheer availability of leisure time is not the whole story. Scarcity in leisure time can be offset by a more intensive use of it. As discussed in this article, such use relies on financial and cultural resources which are associated with education and, to some extent, with nationality

Leisure is increasingly considered a key component of quality of life (Stiglitz et al., 2010). Leisure is classically defined as "all activities that we cannot pay somebody else to do for us and we do not really have to do at all if we do not wish to" (Burda et al., 2007: p. 1). In other words, leisure corresponds to the portion of time freed from the necessity of labor, whether paid or unpaid. Looked at this way, leisure is equivalent to resting time. But this very general perspective fails to grasp some of leisure's qualitative aspects: How people spend their free time matters, too. Therefore, researchers interested in leisure usually focus on specific topics such as sport or cultural participation.

Hereafter, leisure is thus chiefly considered in relation to cultural participation and consumption. Leisure and cultural consumption are also considered from the perspective of social inequalities in access to recreational and cultural facilities. Because many European countries devote a significant amount of public funding to culture, cultural participation is also a public policy issue. And the main concern of policy makers in that domain is precisely to ensure fair access to cultural and recreational goods.

General Trends

Trends in leisure and cultural participation can be studied in various ways. First, one can rely on expenditure data. But expenditure alone can be misleading insofar as it reduces the value of leisure activities to their market price and conceals all recreational and cultural activities that do not entail any market transaction. In addition, as the data on expenditure mainly relate to households, they cannot assess individual practices. For these reasons, it is necessary to supplement expenditure data with information on the amount of time people spend on leisure and recreational activities and with data on cultural participation.

Trends in Cultural and Recreational Expenditures

Cultural and recreational expenditures consist of a wide variety of goods and services: books, newspapers, cinema, museums; TV and radio taxes; goods used in amateur activities, such as drawing materials and musical instruments; audiovisual, photographic, and information processing equipment; games and toys; plants and flowers; package holidays; pets and related products. In 2005, cultural and recreational expenditure accounted on average for approximately 9% of total household expenditure in 27 European countries, ranging from 3% in Bulgaria to 13% in Sweden (see **Figure 1**).

The ranking of European countries displays a clear South/North gradient and, to a lesser extent, an East/West one. A first group of countries, where the share of cultural and recreational expenditure accounts for more than 10% of total expenditure of households, includes all Scandinavian countries together with Finland, Germany, the Czech Republic, the Netherlands, and the United Kingdom. In contrast, cultural and recreational expenditures account for less than 5% of household expenditure in Lithuania, Romania, Greece, and Bulgaria. Variations across countries reflect differences in wealth, price structures, and availability of cultural and recreational amenities. Availability, in turn, is partly dependent on the level of public funding for leisure and cultural infrastructures.

But these disparities also reveal variations in intrinsic preference for leisure. This is shown in **Figure 2**, which plots household total expenditure against share of cultural and recreational expenditure in this total. The crossing of these two dimensions defines four groups of countries. The first consists of ones where both household total expenditure and share of the recreational and cultural expenses are above the EU27 mean: Belgium, Germany, Ireland, Malta, Austria, the Netherlands, and the United Kingdom. All of them are quite wealthy countries, supporting the idea that the share of total expenditure devoted to leisure and culture is primarily linked to earnings and wealth. A second group – Bulgaria, Lithuania, Romania, Estonia, Slovakia, Latvia, Poland, Hungary, and Portugal – consists of countries where both household total expenditure and share of the recreational and cultural expenses are below the EU27 mean. As these countries are on the whole poorer than those of the first group, this second configuration also supports the hypothesis that cultural and recreational expenditure share is mainly due to global wealth.

The other two groups are not completely supportive of this pure wealth effect, however. The third group, reduced to one country – the Czech Republic – is characterized by the highly improbable combination of above-average share of leisure and total expenditure below the European mean. Interestingly, the final group consists of countries where a high level of

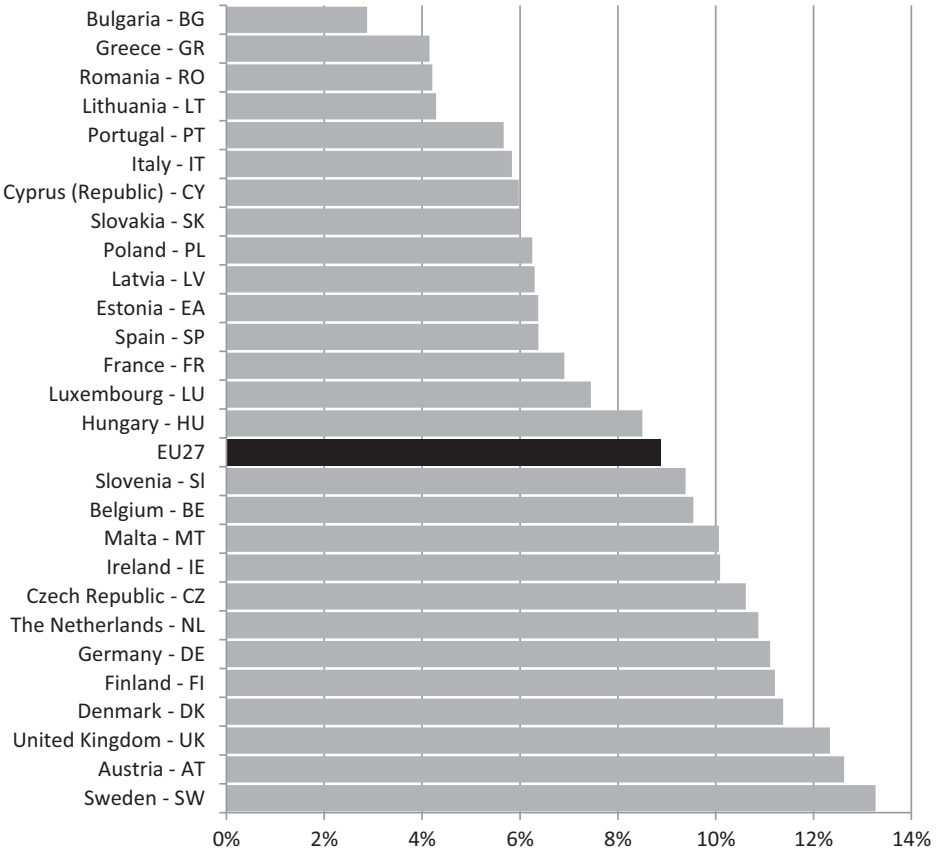

Figure 1 Share of recreation and culture in households expenditures in 2005 in EU27. Source: Eurostat, Household Budget Survey 2005.

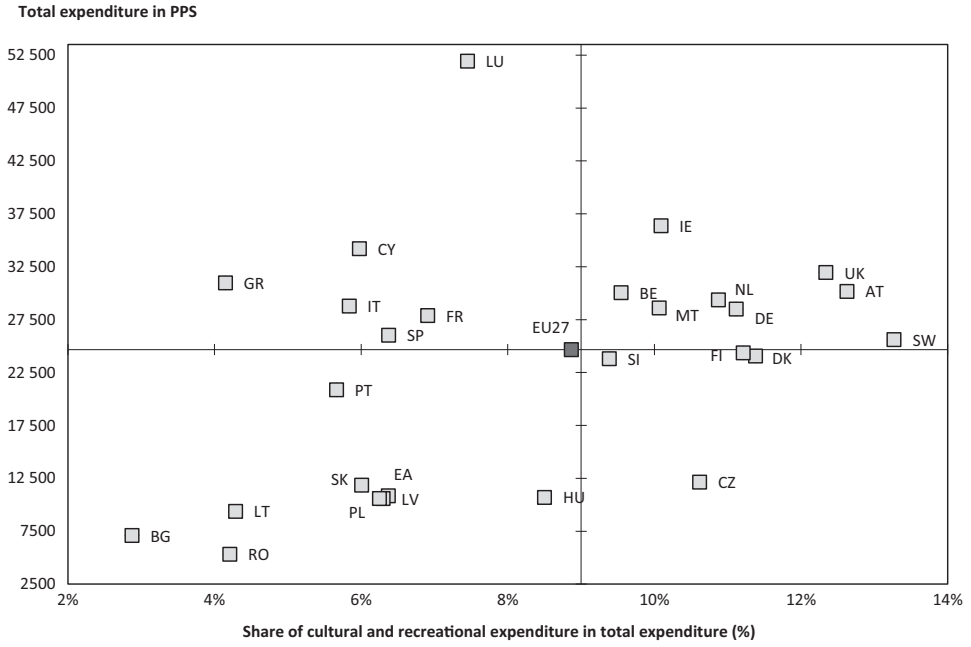

Figure 2 Total household expenditure and share of cultural and recreational expenditure in 2005 – EU27. Note: households final consumption expenditure is measured in pps (purchasing power standard) Source: Eurostat, Household Budget Survey 2005.

household expenditure coexists with a rather low share dedicated to culture and leisure: France, Spain, Italy, Greece, Cyprus, and Luxembourg. These two configurations, especially the latter, illustrate the limits of cross-national comparisons in cultural and recreational expenditure. To the extent that a substantial part of cultural and recreational amenities, although highly variable from one country to another, is publicly funded, cross-national comparisons are biased by the coexistence of market and nonmarket goods. In that sense, a relatively low level of private cultural and recreational expenditure is not necessarily synonymous with a small commitment to culture. For this reason, the household expenditure approach must be supplemented by leisure time budget and cultural participation statistics.

Trends in Leisure Time

Technically speaking, leisure might be considered 'residual time': the time remaining after paid and unpaid work (household chores), child and personal care have been subtracted from the total amount of waking time per day. According to this criterion, however, many activities are ambiguous. Is the time devoted to reading stories to a child a matter of leisure or a matter of child care? Is time devoted to cook a meal for friends a matter of domestic chores or recreational time? To what extent can the time devoted to repairing a car, decorating a house, or doing arts and crafts at home be included in leisure time?

To explicitly take into account the ambiguous nature of such activities, they are sometimes labeled 'semi-leisure' activities. Relying on this extended notion of leisure time, three major tendencies can be highlighted in Europe in recent decades: (1) a slight decrease in time devoted to leisure, at least in some countries; (2) an increasing variability of the distribution of time allocated to leisure; and (3) a strong heterogeneity in the cultural contents of leisure.

A Pause on the Road toward the 'Leisure Society'

At the beginning of the 1960s, the French sociologist Joffre Dumazedier prophesied the coming of a 'leisure society' (Dumazedier, 1962). From the end of the nineteenth century to the third-quarter of the twentieth century, the continuous decline in the number of hours of paid work observed in many Western countries supported Dumazedier's prediction. Since the late 1970s, however, in many countries, the trend has not continued (Zuzanek et al., 1998; Gershuny, 2000; Chenu and Herpin, 2002). Since then, European countries have followed slightly different trends. In some (e.g., Sweden and the United Kingdom), the annual number of paid work hours started to increase at the beginning of the 1980s, whereas in others (e.g., Denmark and, to a lesser extent, France), this reversal occurred later and was less pronounced (**Figure 3**).

The same holds true when leisure is defined as the time that remains after subtracting paid and unpaid work. Since the end of the 1970s, countries such as Norway, the Netherlands, and France experienced a significant decline in the leisure time of men. The same is true for women in Norway, the Netherlands, and the United Kingdom (Gimenez-Nadal and Sevilla, 2012). As argued by Gershuny (2009), the work/leisure balance has evolved differently for men and women. In nearly all countries, men experienced an increase in the proportion of unpaid relative to paid work, whereas the opposite was true for women. Changes in leisure time may reflect these shifts between paid and unpaid labor. Furthermore, the average amount of leisure time is highly variable form one country to another. According to the Eurostat Time Use Survey, leisure time in 2000 varied from a minimum of 4:13 a day in Lithuania to a maximum of 5:46 in Norway (**Figure 4**).

According to Gershuny (2009), this trend (which of course may be only temporary) might reflect a deeper change in the social meaning of leisure. In the industrial society, leisure was the privilege of the dominant classes who owned the means of production and exploited the workforce of nonowners. At that time, as asserted by Veblen (1899) in his leisure class theory, idleness was a correlate of wealth. But the emulation principle that governed social dynamics made leisure universally desirable: People tried to emulate the lifestyle of those just above them in the hierarchy of social prestige and power. Combined with the labor productivity gains that went along with the development of capital accumulation during the first three-quarters of the twentieth century, this emulation principle led to the observed decline in the duration of paid work and the associated increase in leisure time.

In contemporary Western countries, however, it is not clear whether these social dynamics are still in play. Modern capitalist companies increasingly separate ownership and management. The upshot is that for the dominant social classes, human capital has become more important than

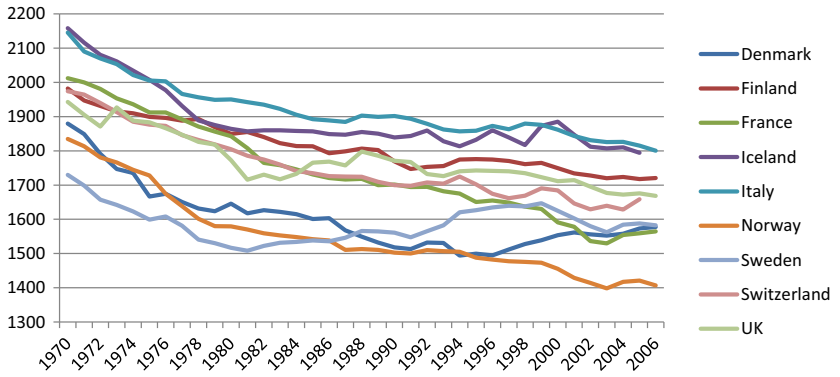

Figure 3 Annual hours worked by the total employed population in selected OECD countries. Source: OECD, Employment Outlook 2006.

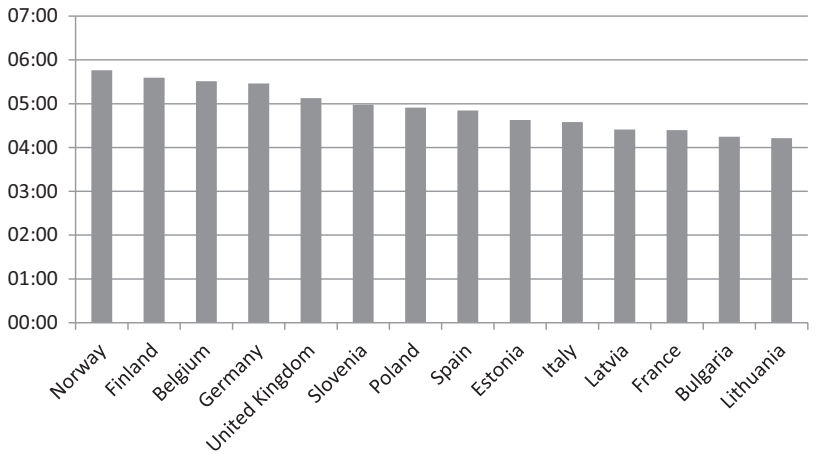

Figure 4 Leisure time in 2000 in 14 European countries. Source: Eurostat, Time Use Survey 2000.

capital as such. Income and power increasingly derive not from ownership of capital (e.g., factories) but from highly qualified, highly paid jobs (Gershuny, 2009). In short, people at the top are working longer hours, thereby realizing the counter-prophecy of a harried dominant class (Linder, 1970). The trends in leisure inequalities noted above are partially supportive of this scenario.

The Reversal of the Leisure/Status Gradient

Another meaningful evolution observed in recent decades in the West in general, and in Europe in particular, is the increasingly unequal distribution of leisure since the 1970s (Gimenez-Nadal and Sevilla, 2012). This increase in leisure inequality is closely but negatively linked to the increasing inequality in earnings that occurred in the same period. Indeed, those who have benefited the most from the former are those who have suffered the most from the latter. In an increasingly knowledge-based economy, high earnings predominantly accrue to the most educated people. In other words, high incomes function as an incentive to increase the amount of paid work and reduce leisure time. By contrast, less-educated people work fewer hours, earn less, and have more leisure time (Gimenez-Nadal and Sevilla, 2012). This results in a reversal of the leisure/status gradient: whereas in the past the wealthiest people had more leisure time than others, now they have less (Gershuny, 2009).

This is not to say, though, that increased leisure time for the less wealthy is adequate compensation for their economic downgrading. One could argue that the wealthiest and most educated people compensate for the scarcity of their leisure time by using it intensively, made possible by the economic and cultural resources available to them (Coulangeon et al., 2002; Degenne et al., 2002; Gronau and Hamermesh, 2008).

It might also be noted that leisure time in Europe is highly gendered. As shown in **Figure 5**, across Europe men have more leisure than women (about 38 min more on average). The cross-national variations display a rather clear north/south and west/east gradient. Gender differences are minimal in Norway (4 min) and maximal in Poland (56 min) and Italy (79 min).

The Cultural Paradox of Free Time

Inequality in leisure is not exclusively a matter of quantity; it is also – and primarily – a matter of quality. In other words, what people do with their free time, whatever its amount, matters. This can be illustrated by the total time spent on TV and TV as share of total leisure time according to level of education. As

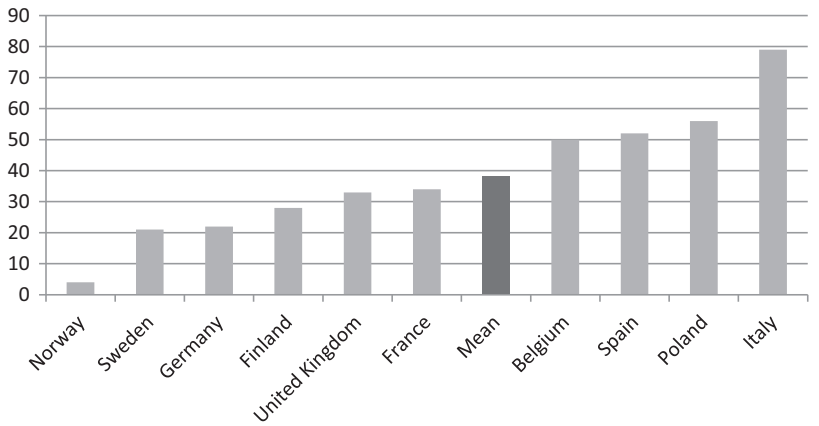

Figure 5 Gender differences (minutes per day) in leisure time, showing more leisure time for men, 2006. Source: OECD estimates, based on national and multinational time use surveys.

shown in Table 1, the average daily time devoted to TV ranges from 1:49 in Germany to 2:27 in Bulgaria. Across all nine countries shown in Table 1, TV accounts for roughly 40–60% of total leisure time, ranging from 39% in Spain to 58% in Bulgaria.

With the exception of Slovenia and Bulgaria, daily time devoted to TV is negatively related to level of education. This negative relation is particularly strong in the United Kingdom and France, where the least educated watched TV for 1:15 and 1:20, respectively, more than the most educated. In these countries especially, TV watching appears as a particularly illegitimate practice for the culturally well-to-do (Lahire, 2004).

It is also noticeable that in all countries time spent on reading increases with education (Table 2). More generally, it has been shown elsewhere that TV watching is the only cultural practice whose intensity is negatively correlated with all other cultural practices (Coulangeon and Lemel, 2009).

In the nine countries shown in Table 2, daily reading time ranges from 15 min in Spain to 46 min in Finland. Time devoted to reading is strongly and positively related to education. But even among the most educated people of the most reading countries, such as Finland or Estonia, reading never represents more than 19% of total leisure time.

More generally, it can be demonstrated that the cultural content of leisure time tends to be negatively correlated with quantity. On average, people who have the most leisure time have the fewest number of cultural practices (Coulangeon et al., 2002). Among the most educated people and among the upper classes, intensity and efficiency of the use of time seems to partially compensate for its relative scarcity. In that sense, access to more active and culturally intensive practices seems to be a matter of financial and cognitive resources rather than a matter of time as such. Furthermore, many recreational practices take place over a longer timescale that cannot be adequately grasped on a daily

Table 1 Time spent watching TV (h:min per day) by level of education, 2000

	Bulgaria	*Germany*	*Estonia*	*Spain*	*France*	*Lithuania*	*Slovenia*	*Finland*	*UK*
Mean time spent on TV									
Level 1	2:05	2:22	2:52	2:25	2:37	2:34	1:31	2:46	3:03
Level 2	2:29	2:15	2:43	1:55	2:03	2:51	1:53	2:20	2:35
Level 3–4	2:30	1:53	2:16	1:34	1:54	2:17	2:00	2:08	2:14
Level 5B	2:27	1:49	1:59	1:29	1:27	1:52	1:42	1:57	2:02
Level 5A	2:25	1:28	1:49	1:21	1:17	1:48	1:34	1:50	1:48
Mean	**2:27**	**1:49**	**2:16**	**1:53**	**2:01**	**2:16**	**1:57**	**2:13**	**2:23**
Share of TV in total leisure time									
Level 1	48%	41%	49%	44%	53%	49%	30%	42%	50%
Level 2	59%	38%	51%	41%	46%	57%	39%	41%	50%
Level 3–4	60%	34%	51%	35%	45%	56%	40%	40%	47%
Level 5B	57%	34%	49%	33%	38%	52%	31%	38%	41%
Level 5A	54%	28%	43%	30%	32%	45%	33%	35%	38%
Mean	**58%**	**33%**	**49%**	**39%**	**46%**	**54%**	**39%**	**40%**	**46%**

Note: Level 1: Primary education or first stage of basic education.
Level 2: Lower secondary or second stage of basic education.
Level 3 and 4: Upper secondary and postsecondary nontertiary education.
Level 5B: First stage of tertiary education, programs which are practically oriented and occupationally specific.
Level 5A: First stage of tertiary education, tertiary programs that are theoretically based/research preparatory or giving access to professions with high skill requirements.
Source: Eurostat, Time Use Survey 2000.

Table 2 Time spent reading (h:min per day) by level of education, 2000

	Bulgaria	*Germany*	*Estonia*	*Spain*	*France*	*Lithuania*	*Slovenia*	*Finland*	*UK*
Mean time spent on reading									
Level 1	0:03	0:20	0:32	0:09	0:22	0:18	0:12	0:49	0:31
Level 2	0:09	0:33	0:35	0:12	0:25	0:20	0:14	0:39	0:22
Level 3–4	0:19	0:35	0:35	0:16	0:18	0:22	0:23	0:41	0:20
Level 5B	0:28	0:38	0:38	0:24	0:25	0:22	0:39	0:48	0:29
Level 5A	0:36	0:43	0:45	0:30	0:34	0:32	0:36	0:59	0:32
Mean	**0:18**	**0:38**	**0:37**	**0:15**	**0:23**	**0:23**	**0:23**	**0:46**	**0:26**
Share of reading in total leisure time									
Level 1	1%	6%	9%	3%	7%	6%	4%	12%	9%
Level 2	4%	9%	11%	4%	9%	7%	5%	12%	7%
Level 3–4	8%	11%	13%	6%	7%	9%	8%	13%	7%
Level 5B	11%	12%	16%	9%	11%	10%	12%	16%	10%
Level 5A	13%	14%	18%	11%	14%	13%	12%	19%	11%
Mean	**7%**	**12%**	**13%**	**5%**	**9%**	**9%**	**8%**	**14%**	**8%**

Source: Eurostat, Time Use Survey 2000.

basis. For this reason, beyond the time–budget approach, data on cultural and recreational practices may provide further insight.

Trends in Cultural Participation

Cultural facilities have different rates of attendance. As shown in Figure 6, about 50% of Europeans visit historical sites or attend the cinema at least once a year, whereas other cultural facilities are less frequently attended.

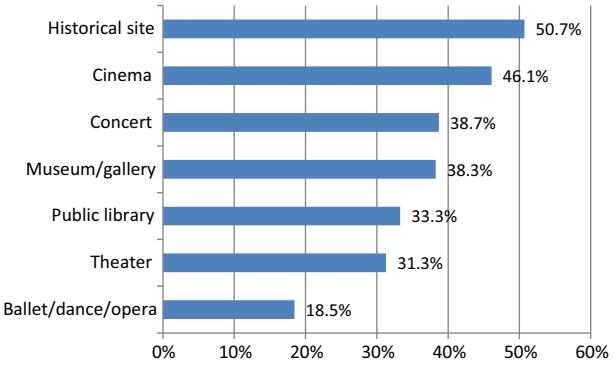

Figure 6 Attendance at cultural facilities, EU27, 2013. Note: proportion of respondents who have been at least one time in each of the facilities in the year preceding the survey. Source :Eurostat, Eurobarometer 79.2, 2013.

Beyond this general statement, cross-national comparison displays huge discrepancies across Europe. For example, on average 18.5% of Europeans attend a ballet, dance performance, or opera at least once a year, but, as shown in Table 3, the rate varies from 7.5% in Portugal to 34.3% in Sweden.

Again, western and northern European countries differ from eastern and southern ones. One might suspect that these variations may be due to differences in the availability of cultural facilities, which is itself highly dependent on degree of urbanization. However, as Table 3 shows, when the data are examined separately for rural area, middle town and large town, the ranking of the countries remains much the same. Differences between countries are not reducible to their level of urbanization, therefore, but reflect dissimilarities in global wealth and supply of cultural amenities. In that sense, it is particularly noticeable that, except for Estonia, the countries where the contrast between rural and urban areas is the strongest (as shown by the odds ratio in the right-hand column) are also those with the smallest rate of attendance overall.

The contrasts between European nations remain consistent when we turn to more common cultural practices, such as movie going (Table 4). The ranking is similar, as is the

Table 3 Ballet, dance performance, or opera attendance, 2013

	Rural area	*Middle town*	*Large town*	*Mean*	*OR*
Portugal	5.65	9.92	6.51	**7.5**	1.2
Greece	5.42	14.72	11.07	**10.1**	2.2
Poland	6.6	7.87	17.63	**10.1**	3.0
Romania	5.42	12.26	18.89	**11.4**	4.1
Bulgaria	4.35	11.69	15.59	**11.6**	4.1
Hungary	7.65	10.82	15.93	**11.6**	2.3
Cyprus (Republic)	2.9	22.16	9.34	**12.3**	3.4
Slovakia	12.12	13.73	23.13	**14.3**	2.2
Slovenia	11.47	17.02	18.4	**14.9**	1.7
Spain	12.64	18.5	17.03	**15.1**	1.4
Malta	14.72	15.87	25	**16.2**	1.9
Czech Republic	16.56	14.86	20.73	**16.9**	1.3
Finland	13.61	16.95	23.86	**17.1**	2.0
Italy	14.29	18.11	21.13	**18.1**	1.6
EU27	**14.43**	**18.66**	**23.19**	**18.5**	**1.8**
Austria	16.36	19.24	21.97	**19.1**	1.4
Germany – West	17.59	18.03	24.54	**19.4**	1.5
Ireland	14.37	17.53	25.2	**19.4**	2.0
United Kingdom	21.96	20.22	21.57	**21.1**	1.0
Belgium	18.2	22.87	25.53	**21.5**	1.5
Germany – East	15.38	18.29	32.78	**22.5**	2.7
Lithuania	17.32	18.49	34.98	**23.4**	2.6
The Netherlands	20.49	23.37	28.63	**23.5**	1.6
Latvia	19.24	20.06	33.23	**24**	2.1
Estonia	14.44	20.5	40.38	**24.4**	4.0
France	23.25	23.46	30.88	**24.9**	1.5
Denmark	19.42	26.71	32.23	**26.1**	2.0
Luxembourg	30.45	28.26	34.55	**29.9**	1.2
Sweden	28.46	33.41	40.8	**34.3**	1.7

Note: The figures above represent the percentage of respondents who attended at least one ballet, dance performance, or opera in the year preceding the survey.
OR = odds ratio (for explanation, see footnote 1).
Source: Eurostat, Eurobarometer 79.2, 2013.

Table 4 Cinema attendance, 2013

(%)	*Rural area*	*Middle town*	*Large town*	*Mean*	*OR*
Romania	10.6	20.7	31.6	**19.8**	2.1
Portugal	15.5	33.6	31.6	**25.9**	1.9
Bulgaria	13.8	13.7	39.5	**26.2**	2.2
Hungary	22.9	29.5	40.9	**31.5**	1.5
Cyprus (Republic)	19.6	41.1	34.6	**32.9**	2.0
Slovakia	31.7	32.7	48.5	**34.4**	1.1
Greece	25.3	35.6	40.2	**35.3**	1.6
Poland	26.7	31.8	55.0	**36.3**	1.6
Slovenia	34.4	44.4	43.6	**39.9**	1.3
Lithuania	25.2	38.1	54.5	**40.1**	2.0
Malta	37.1	48.4	39.1	**40.2**	1.1
Estonia	27.8	40.1	57.1	**40.8**	1.8
Latvia	30.9	43.0	51.6	**41.3**	1.6
Finland	34.7	40.7	55.7	**41.5**	1.3
EU27	**38.0**	**47.9**	**53.7**	**46.1**	**1.4**
Germany – East	41.9	38.6	60.0	**46.4**	1.2
Spain	40.5	54.2	55.9	**47.2**	1.3
Czech Republic	42.7	48.1	52.4	**47.6**	1.2
Belgium	48.6	52.1	53.7	**51.0**	1.1
Italy	47.0	52.9	61.5	**53.7**	1.3
Germany – West	52.2	57.9	59.3	**56.3**	1.2
United Kingdom	47.8	58.7	61.1	**56.7**	1.4
Austria	52.0	61.2	65.6	**59.2**	1.3
Luxembourg	59.1	60.0	61.8	**59.8**	1.0
France	54.8	63.0	68.6	**61.0**	1.3
Ireland	52.1	63.2	69.2	**61.8**	1.5
The Netherlands	58.1	67.1	71.0	**64.4**	1.3
Sweden	56.5	67.0	78.3	**67.7**	1.6
Denmark	66.2	69.8	77.7	**70.7**	1.2

Note: The figures above represent the percentage of respondents who attended the cinema at least once in the year preceding the survey.
OR = odds ratio (see footnote 1).
Source: Eurostat, Eurobarometer 79.2, 2013.

contrast between western and northern Europe, on the one hand, and southern and eastern Europe, on the other. The most significant exceptions are Estonia and Latvia, which rank much higher on ballet, dance performance, or opera than on cinema. In addition, the differences between urban and rural areas are much smaller for cinema than for ballet, dance, and opera.

All in all, cross-national differences in access to cultural amenities are quite similar in the realm of mass culture and in the elite culture domain. Differences between countries appear to be more a matter of global affluence than cultural legitimacy, in Bourdieu's sense of the notion (Bourdieu, 1984[1979]).

Across Europe, access to cultural facilities displays huge inequalities with respect to social class, education, age, and gender. Relative to other groups, professionals display a higher commitment to cultural participation, as shown by the values of the cultural participation index reproduced in Table 5. The countries' ranking on this index is consistent with previous results, contrasting once again the same subgroups. It is also again noticeable that the relative distance between groups, as measured by the ratio of the professionals' score to the manual workers' score, is highest in countries with the lowest average scores, such as Portugal and Hungary, and lowest in countries with the highest average scores, such as Sweden, Finland, and Denmark. These three last countries are also known to be more egalitarian in many other respects (income, wealth, health, education). The variations in the magnitude of the between-group differences in cultural participation thus suggest that culture is a stronger social marker in unequal societies than in more egalitarian ones (Hjellbrekke and Korsnes, 2014). These variations might also be due to the impact of educational and cultural policies.

Finally, it is worth noting that the cultural participation scores of the professionals are much more homogeneous across nations than the scores of other groups, as shown by the relative standard deviations in the bottom row of Table 5. Note that these values are lowest for the professionals, highest for farmers and fisherman, and twice as high for manual workers than for professionals. European elites appear more culturally homogeneous than the rest of the population.

Table 5 Cultural Participation Index by occupation, 2013

	FARM	*SSE*	*PROF*	*MID*	*NMWkrs*	*MWkrs*	*Total*	*Ratio PROF/MWkrs*
Portugal	3.1	3.3	8.1	7.2	4.8	2.3	3.7	3.6
Greece	1.0	3.0	8.0	6.3	5.0	2.5	4.1	3.3
Romania	1.7	5.0	8.5	4.2	5.1	2.9	4.3	2.9
Cyprus (Republic)	0.8	3.6	7.1	5.6	4.6	3.0	4.5	2.4
Hungary	2.2	6.7	11.1	7.0	5.9	3.0	4.8	3.7
Poland	2.0	5.9	8.6	8.8	5.9	3.3	5.2	2.6
Bulgaria	1.5	4.6	8.5	6.8	6.0	3.1	5.3	2.7
Italy	1.8	5.0	9.2	8.3	6.2	3.7	5.6	2.5
Slovakia	7.5	6.3	9.1	8.7	6.9	4.0	6.5	2.3
Malta	6.5	6.5	11.3	9.0	7.6	4.6	6.5	2.5
Spain	4.5	6.3	10.4	11.1	8.2	5.4	6.8	1.9
Czech Republic	7.6	6.2	9.3	10.5	7.4	4.8	7.4	2.0
EU27	**3.8**	**6.2**	**11.0**	**10.3**	**8.1**	**4.7**	**7.5**	**2.3**
Austria	4.2	8.0	10.5	9.6	7.5	5.0	7.5	2.1
Belgium	4.8	6.2	11.6	9.9	8.5	4.2	7.5	2.8
Germany –West	2.8	7.0	10.8	8.8	7.8	4.7	7.9	2.3
Germany – East	10.3	3.9	11.1	9.9	7.9	4.8	8.0	2.3
Lithuania	7.0	8.5	10.9	9.6	8.4	5.4	8.0	2.0
Slovenia	4.7	9.1	11.9	9.8	8.3	5.5	8.1	2.2
Ireland	9.1	10.7	11.1	10.6	7.7	5.9	8.2	1.9
France	6.4	9.2	13.5	10.7	8.2	5.4	8.7	2.5
Latvia	9.0	9.1	10.7	10.8	8.8	6.1	8.7	1.8
Finland	3.9	8.1	9.4	11.1	9.7	7.0	8.8	1.3
Luxembourg	6.4	9.3	12.1	10.6	9.2	5.9	9.1	2.0
United Kingdom	5.7	9.6	11.5	10.8	10.0	6.7	9.2	1.7
Estonia	7.0	10.0	12.0	12.3	9.9	6.8	9.4	1.8
The Netherlands	9.1	8.3	12.5	11.4	10.9	7.0	10.7	1.8
Denmark	7.5	10.3	13.6	13.6	11.2	8.3	11.1	1.6
Sweden	7.7	9.9	13.7	14.3	12.4	9.6	12.8	1.4
RSD	0.54	0.32	0.17	0.23	0.25	0.35	0.30	

Note: The relative standard deviation coefficient (RSD) corresponds to the standard deviation divided by the mean of the values of the cultural index across the 27 nations.
FARM: Farmers and fishermen.
SSE: Small self-employed.
PROF: Professionals, business proprietors, owners, employed professionals, top managers.
MID: Middle managers, supervisors.
NMWkrs: Nonmanual routine workers.
MWkrs: Manual routine workers.
Source: Eurostat, Eurobarometer 79.2, 2013.

Summary and Conclusions

Regarding leisure and cultural consumption in Europe, three main tendencies deserve to be underlined. First, time budget surveys suggest a possible halt on the road toward the so-called leisure society. Second, and as shown repeatedly in this article, variations in leisure and cultural consumption vary significantly from one country to another. Third, leisure time remains subjected to significant inequalities, especially in terms of class, education, and gender. The pattern of inequalities is, however, complex. On average, the most advantaged in leisure time are the most disadvantaged in economic, educational, and cultural resources. As demonstrated by Marie Jahoda, Paul F. Lazarsfeld, and Hans Zeisel about the Marienthal unemployed (Jahoda et al., 1971[1932]) – and dramatically illustrated today in some European countries grappling with massive unemployment – overabundant free time is not necessarily synonymous with true leisure. Leisure time cannot be an end in itself if people cannot afford to enjoy it.

See also: Cultural Participation, Trends in; Cultural Policy Regimes in Western Europe; Formal Methods of Cultural Analysis; Globalization and World Culture; Leisure and Cultural Consumption: US Perspective; Social Inequality in Cultural Consumption Patterns.

Bibliography

Bourdieu, P., 1984. Distinction: A Social Critique of the Judgment of Taste. Harvard University Press, Cambridge, MA (Original work published 1979).

Burda, M., Hamermesh, D.S., Weil, P., 2007. Total Work, Gender and Social Norms. National Bureau of Economic Research.

Chenu, A., Herpin, N., 2002. Une pause dans la marche vers la civilisation des loisirs? [A break in the march towards the leisure civilization?]. Économie et statistique 352 (1), 15–37.

Coulangeon, P., Lemel, Y., 2009. The homology thesis: distinction revisited. In: Robson, K., Sanders, C. (Eds.), Quantifying Theory: Pierre Bourdieu. Springer, pp. 47–60.

Coulangeon, P., Menger, P.-M., Roharik, I., 2002. Les loisirs des actifs: un reflet de la stratification sociale [Workers' leisure time: a reflection of social stratification]. Économie et statistique 352 (1), 39–55.

Degenne, A., Lebeaux, M.-O., Marry, C., 2002. Les usages du temps: cumuls d'activités et rythmes de vie [How time is put to use: accumulating activities and lifestyles]. Économie et statistique 352 (1), 81–99.

Dumazedier, J., 1962. Vers une civilisation du loisir? [Towards a Society of Leisure?]. Le Seuil, Paris.

Esping-Andersen, G., 1990. The Three Worlds of Welfare Capitalism. Polity Press, Cambridg UK.

Gershuny, J., 2000. Changing Times: Work and Leisure in Postindustrial Society. Oxford University Press.

Gershuny, J., 2009. Veblen in reverse: evidence from the multinational time-use archive. Social Indicators Research 93 (1), 37–45.

Gimenez-Nadal, J.I., Sevilla, A., 2012. Trends in time allocation: a cross-country analysis. European Economic Review 56 (6), 1338–1359.

Gronau, R., Hamermesh, D.S., 2008. The demand for variety: a household production perspective. The Review of Economics and Statistics 90 (3), 562–572.

Hjellbrekke, J., Korsnes, O., 2014. Cultural distinctions in an egalitarian society. In: Coulangeon, P., Duval, J. (Eds.), The Routledge Companion to Bourdieu's 'Distinction'. Routledge, Oxford.

Jahoda, M., Lazarsfeld, P., Zeisel, H., 1971[1932]. Marienthal: The Sociography of an Unemployed Community. Aldine, Chicago.

Lahire, B., 2004. La culture des individus: dissonances culturelles et distinction de soi [The Culture of individuals. Cultural dissonance and self distinction]. La découverte, Paris.

Linder, S.B., 1970. The Harried Leisure Class. Columbia University Press, New York.

Stiglitz, J.E., Sen, A., Fitoussi, J.-P., 2010. Report by the Commission on the Measurement of Economic Performance and Social Progress. Commission on the Measurement of Economic Performance and Social Progress, Paris.

Veblen, T., 1899. The Theory of the Leisure Class: An Economic Study of Institutions. Macmillan, New York.

Zuzanek, J., Beckers, T., Peters, P., 1998. The 'harried leisure class' revisited: Dutch and Canadian trends in the use of time from the 1970s to the 1990s. Leisure Studies 17 (1), 1–19.

Leisure and Cultural Consumption: US Perspective

John P Robinson, University of Maryland, College Park, MD, USA
Timothy Triplett, Urban Institute, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.P. Robinson, volume 13, pp. 8709–8714, © 2001, Elsevier Ltd.

Abstract

Leisure is classically defined as 'freedom from the necessity of labor' – a state of mind that is distinguished from free time, which can be quantified using time diaries. US national time-diary figures over the last 50 years show declines in the hours spent at paid and unpaid work for those aged 18–64, leading to about 38.5 h a week of free time for activities like socializing, fitness activity, and mass media use. Almost half of this free time is devoted to television, suggesting Americans are unprepared to use this important resource wisely. Reinforcing these concerns are arts participation surveys showing a recent decline in these activities, which take up less than 1% of free time. Social class differences implicit in Veblen's 'leisure class' may be reversing, although signs of 'conspicuous consumption' clearly remain. More positive subjective indicators indicate no increases in Americans feeling rushed, being fatigued, or feeling unhappy. Nonetheless, there remains the need to develop more sensitive indicators of the quality of American leisure.

Most definitions of leisure treat it as activity (or inactivity) opposite to or the counterpoint of work. Leisure can be considered as a reward for work, as the purpose for work, or as 're-creation' to recover from work. In a postmodern society in which opposites are increasingly blended, we see attempts by individuals and institutions both to make their work more leisurely (e.g., taking yoga or exercise breaks, celebrating birthdays) and to make leisure more work-like. Thus, in their work time, people dress more casually, take sabbaticals, or attend retreats; while in their 'harried leisure' away from work they 'work out', act as purchasing agents for their households, carry online devices into the wilderness or master new work skills.

A classic definition of leisure speaks of it as 'freedom from the necessity of labor'. In the economic sense of that definition, America has become a leisure society *par excellence*, with unprecedented proportions of its retired citizens no longer needing to earn money from their work labors to obtain adequate provisions of food, clothing, and shelter (not to deny the significant minorities who remain below the poverty line and need to work to make ends meet).

Free Time vs Leisure

One political philosopher who foresaw many of these developments was Sebastian de Grazia (1962), who lamented how capitalistic democracies had overcome the 'tyranny of man' only to submit to the 'tyranny of the clock' (as exemplified by the gift of gold watches to retiring employees who least needed them). DeGrazia made a sharp distinction between free time, which can be quantified, and leisure, which cannot. In ancient societies, he saw the purest form of leisure emerging when individuals pursued activities (contemplation, composing music or poetry) from a personal sense of commitment, quality, and fulfillment. In modern capitalistic societies, DeGrazia considered such leisure as almost invisible, even as free time has continued to increase.

Perhaps the closest modern equivalent to classic leisure is evident in Csikszentmihalyi's (1990) studies of 'flow', except that here individuals seek optimal experience and joy in more active pursuits, like skydiving or extreme sports. Two ironic features of flow episodes are that (1) the sense of time in a time-oriented society largely disappears and (2) most flow episodes he cites occur during work hours rather than free time. Like leisure, flow remains unquantified, since it is not clear in time-use studies when such profound cultural experiences occur.

Measurement of Work and Free Time

Until the last decades, most measures of work and free time have been based on reductions from survey 'estimates' made by workers of their work hours (e.g., "How many hours did you work last week?"), rather than on more detailed accounts that use on-site observation, electronic pagers, or time diaries. The great value of these diary accounts is that workers report on *all* their daily activities, and not just their work time. In accounting for time in time diaries, for example, respondents are less prone to problems of memory loss, self-projection, and double counting of time than when they make time estimates.

The surprising implication from the detailed diary data in Figure 1, then, is that Americans during their prime working years (ages 18–64) now have slightly more free time than in the 1950s or 1960s (Robinson, 2009; Aguiar and Hurst, 2009).

To some extent, that increased free time has come about because of demographic or structural changes in the workforce, in that today's workers:

1. enter the workforce later in life, as they receive more years of college education;
2. may retire or work a reduced schedule at age 62 or earlier;
3. live longer and thus can enjoy more free time during their retirement years.

The end result is that, when aggregated across a person's lifetime, their free time hours will notably outweigh their time at the work (Ausubel and Grubler, 1994). Nonetheless, there remains the question of why Americans continue to work more hours than they need to, and whether they are incapable of appreciating

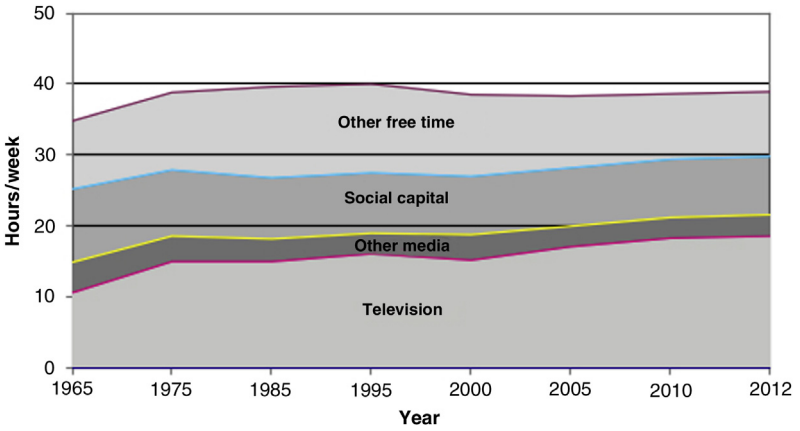

Figure 1 1965–2012 Time-diary trends in weekly hours of free time (age 18–64). Robinson, J., Godbey, G., 1999. Time for Life. Penn State Press, State College, PA; ATUS-X.

leisure, as Reisman (1958) and de Grazia (1962) worried a half-century ago. The finding that more than 60% of the free time of retired people goes to watching TV provides further reinforcement for their concern (Robinson and Caporaso, 2009).

Similar arguments arise from examination of the free-time activities reported by Americans during the hours they are not working. As shown in **Figure 1**, close to 45% of US free time is spent watching TV content that is generally *undemanding* and/or uncritical of material lifestyles or environmental degradation. **Figure 1** shows that their 16+ television hours per week dwarf the combined 3–4 h spent reading, being online, or using other media. Further concern has arisen over the decreased time spent on 'social capital' free-time activities that provide the important social 'glue' that allows societies to flourish (Putnam, 2000). Perhaps the most optimistic trend in these diary accounts is that fitness activity has more than doubled since the 1960s (as part of other 'healthier' life-style changes, such as decreased smoking or red meat consumption).

The dominance of TV in free time is often taken as a sign of a waste of this precious resource by social critics (Mander, 1977; Postman, 1987), or as reducing participation in social capital activities (Putnam, 2000). Viewers themselves rate TV as unchallenging and boring in relation to other daily activities (Kubey and Czikszentmihalyi, 1991) as unenjoyable 'in general' and as the first activity they would give up if more urgent daily needs came up (Robinson and Godbey, 1999). However, these respondents rate TV much more positively in their 'real-time' diaries than in general (Robinson and Martin, 2008, 2009). Moreover, it has been argued that its content has become more sophisticated and challenging in recent years (Johnson, 2004), the greater viewing of serious programs like *60 Minutes* and cable TV movies is further evidence that TV can be an uplifting cultural experience.

Demographic Correlates of Free Time

Who, then, has more of the resource of free time, which at least affords the opportunity of leisure?

Gender

While diary studies through 1995 showed American women aged 18–64 generally equal with men in their free time, studies since then show women with 3–5 h less free time than men, with employed women having about 4 h less than employed men and more than 10 h less than nonemployed women. Thus this gender inequality has been increasing over time, unlike the more equal gender reported by Goldschmidt-Claremont and Pagnossin-Aligisakis (1992).

Age

As might be expected, free time is U-shaped across the lifecycle, reaching more than 40 h a week for teens and young adults, dropping to 30 h or less for those adults raising families during the peak career years of their 30s and 40s, then rising to 40 h again in the 'empty nest' years – and approaching 50 h in their retirement years past age 55.

The main reason for this curve is not age per se but the various role obligations that peak in middle age – namely work, marriage, and children. The more hours of work, obviously the less free time; however, as noted above, there is far from a 1:1 correspondence. Moreover, workers who put in long hours are actually more active in the arts, in fitness activity, in civic participation, and in sexual activity (Robinson and Godbey, 1999; Putnam, 2000). Getting married usually means less free time as well, as does having children, so that married women with children have up to 10 h less free time a week than single childless women of equivalent work status.

Status

While social status may be a powerful predictor of one's power and resources to access more sophisticated forms of leisure, it has recently become associated with having less free time to enjoy such pursuits. That is because more of them are employed and those who work do work longer hours. In addition there is the greater enjoyment of work (and other leisurely features) among higher status workers,

and the need for lower status workers to work longer hours to offset their lower wages. The rich 'leisure class' of which Veblen (1899) wrote so critically over a century ago seems to have given way to rich people who work 80–90 h workweeks today. Indeed, working long hours has become something of a status symbol in its own right (Mack, 2005; Gershuny, 2009), and Hochschild (1997) has documented how certain workers can treat work as a consumption activity.

Ecological Factors

Surprisingly little difference in free time can be found by region, and in (presumably less hectic) rural areas; nor are many notable differences found in different seasons of the year.

International Comparisons

The conclusions drawn above refer mainly to studies conducted in the USA. Many of them, however, are replicated in other Western societies. Using data from the 30+ country archive of time-diary studies at the University of Oxford, for example, Gershuny (2009) found average cross-national gains in free time since the 1960s, much as Bittman (1998) found earlier. These gains largely mirrored those reported by Robinson and Godbey (1999) and by Aguiar and Hurst (2009) for the US. Also as in the US, Gershuny found more work hours for the higher status respondents across time, as work hours for the working class declined. This led to a scenario he described as "Veblen in Reverse" in that the former 'leisure class' were working longer hours as their new badge of honor. Again, the extent that such workers experience more leisure or pleasure during that work time suggests the need for more qualitative research.

These combined multinational free-time figures for 23 European and other mainly Western countries are shown in **Table 1** from Fisher and Robinson (2011). Here, average TV viewing times in other countries (mainly 10–14 h per week) can be seen as 2–8 h lower than in the US, even while they have about the same amount (35–40 h) of total free time. In percentage of free time availability, the US 44% figure in the last column of **Table 1** can be seen as notably higher than in most Western European countries (where it tends to be less than 30%, dipping to less than 20% in Holland). It is close the rates in Bulgaria, Japan, and the Baltic countries, although residents of these countries tend to have less free time available to watch.

Cultural Consumption

Arts Participation

Table 2 summarizes recent trends from a set of US benchmark data on 'higher culture' or 'cultural uplifting' free-time events that has been monitored regularly since 1982 by the Survey of the Public's Participation in the Arts (SPPA), conducted by the National Endowment for the Arts. It can be seen that in 2012 the proportions attending these arts events in the previous year varied from 2% for opera to 15% for musical theater, but that that attendance for each of them has declined since 1992. There are wide social status differences in attending these SPPA arts events, with those of higher education and income being up to five times or more likely to attend one of these benchmark events than those of lower education and income (with

Table 1 Differences in free-time hours and activities done in free time across 23 countries

(In hours per week)	*TV*	*IT*	*Read/audio*	*Social life*	*Organ/church*	*Fitness*	*Other*	*=All free-time hours*	*%TV*
US	17.0	+1.2	+2.2	+11.9	+3.4	+2.0	+1.4	=38.4 h	44
Canada	13.5	1.0	2.3	13.7	2.0	3.1	5.7	41.4	33
Australia	*13.5*	*0.4*	*5.4*	*8.5*	*1.3*	*2.8*	*7.5*	*39.4*	*34*
UK	15.6	1.2	3.2	12.7	1.5	3.2	3.5	40.9	38
Netherlands	8.1	1.8	7.7	17.2	3.2	1.8	2.9	42.7	19
Belgium	15.4	2.6	3.0	15.0	0.8	3.1	4.4	44.3	35
France	13.2	0.6	2.6	13.3	1.3	3.0	3.9	37.9	35
Spain	12.0	1.1	2.0	11.5	1.4	5.2	4.1	37.3	32
Italy	10.6	0.7	2.4	13.4	1.8	3.6	4.9	37.4	28
Germany	*12.1*	*2.0*	*4.5*	*14.6*	*2.2*	*3.5*	*4.6*	*43.5*	*28*
Norway	12.6	1.3	4.4	18.0	1.5	3.9	3.6	45.3	28
Sweden	11.9	1.4	3.8	13.2	1.6	4.0	4.9	40.8	29
Finland	*14.7*	*0.9*	*5.8*	*11.9*	*2.0*	*4.3*	*5*	*44.6*	*33*
Estonia	15.4	0.4	4.8	6.9	1.8	2.7	3.6	35.6	43
Lithuania	15.3	0.8	3.1	7.1	1.9	2.3	2.5	33.0	46
Latvia	13.8	0.5	3.3	8.5	1.4	3.4	3.6	34.5	40
Poland	15.3	1.1	3.5	10.1	2.9	3.2	2.8	38.9	39
Slovenia	13.2	0.7	3.1	11.8	1.4	4.1	4.8	39.1	34
Bulgaria	*16.6*	*0.1*	*2.5*	*8.9*	*1.1*	*3.0*	*2.6*	*34.8*	*48*
Turkey	*13.8*	*NA*	*1.8*	*18.0*	*4.4*	*0.8*	*5.9*	*44.7*	*31*
Korea	14.6	1.5	2.0	12.8	0.5	2.9	3.1	37.4	39
Japan	*13.9*	*1.2*	*2.9*	*5.3*	*0.6*	*2.1*	*5.1*	*31.1*	*45*
Brazil	13.3	0.5	1.5	12.5	3.2	1.6	8.7	41.3	32

From Fisher and Robinson, 2009.

Table 2 Trends in arts participation: 1982–2012

	% Attend a live performance of					*% Using mass media*	
	1982	*1992*	*2002*	*2008*	*2012*	*Radio/TV 2012*	*Internet 2012*
Jazz	9.6	10.6	10.8	7.8	8.1	9.9	5
Classical	13	12.5	11.6	9.3	8.8	11.3	5.2
Opera	3	3.2	3.2	2.1	2.1	4	1.8
Musical	18.6	17.4	17.1	16.7	15.2	6	2.6
Play	11.9	13.5	12.3	9.4	8.3		
Ballet	4.2	4.7	3.9	2.9	2.7	4.7	1.8
Other dance	NA	7.1	6.3	5.2	5.6		
ANY BENCHMARK	39	40.7	38.4	34.6	33.3		
Art museum	22.1	26.7	26.5	22.7	21		
Literature	56.4	54.2	46.6	50.2	46.9	4.3	4.9

Source: Survey of Public Participation in the Arts (SPPA, 2012).

education by far being the greater predictor). However, these recent notable *declines* in arts participation occurred despite the dramatic increase in public general education levels (Robinson, 2012). As summarized at the bottom of Table 2 (and the middle of Figure 2 below), the proportion attending any of these benchmark arts events declined from 41 in 1992 to 33% in 2012, and that is further reflected in parallel declines in attending art museums and reading serious forms of literature, like novels or poetry.

The 2012 figures on the right of Table 2 show that these arts participation levels in live attendance can be matched or exceeded by including arts exposure via radio or TV, the Internet or other new media. That reflects how these arts activities continue to be engaged in far more as spectators than as performers/producers (although the 1997 SPPA found more than 8% of American adults personally playing or singing music at home, but only about 1% in a public performance).

At the same time, the ATUS time diaries indicate that less than 1% of Americans' free time is devoted to all such Table 2 arts participation activities, too small to be detectable in the 'social capital' block in Figure 1 from ATUS. Even the Table 2 levels of arts participation in SPPA, total less than five yearly episodes (or less than 15 annual hours) of arts/culture participation per year. However, these SPPA data do show that the arts are engaged in more by individuals who have a long working week or are heavily involved in other activities, suggesting a 'Newtonian' pattern of behavior in which 'Bodies in motion stay in motion' (a corollary to Parkinson's Law (1962) that 'If you want something done, give it to a busy person').

Figure 2 compares the Table 2 SPPA declines in benchmark arts participation with declines in other leisure activities over these three decades. Even more notable downward trends are found for gardening (from 60 to 39%) and attending sports events (from 48 to 30%) than for arts events. At the same time, the time-diary data in the ATUS and earlier diary studies in Figure 1 do reveal at lease a doubling of exercise and other fitness activity since 1965 (Robinson and Godbey, 1999).

Figure 3 shows the trends in Americans' ranking of their music preferences in the SPPA between 1982 and 2012, one of the few surveys to continuously monitor something of the

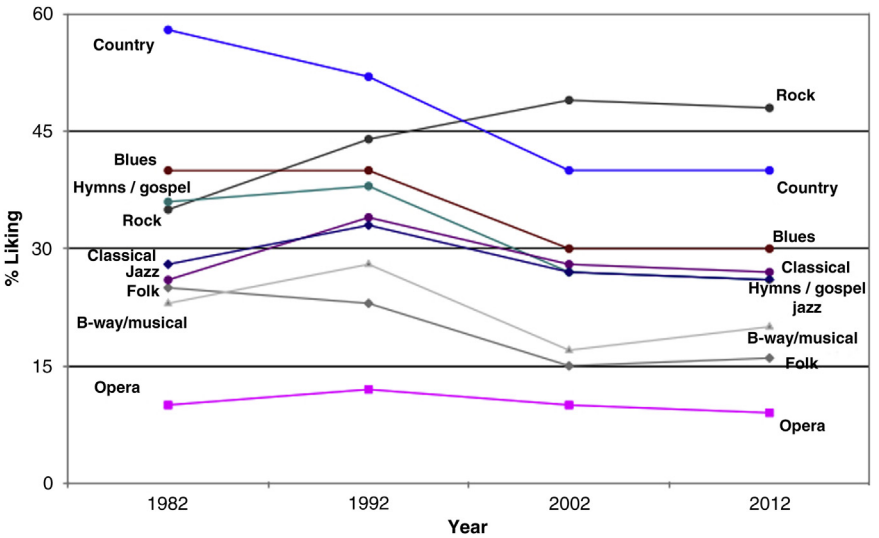

Figure 2 1982–2012 Trends in % of Americans liking different music (age 18+). SPPA, 2012.

'quality' of US cultural consumption. It can be seen that there have been some marked changes in these music tastes over that 30-year period. Whereas country-Western music was rated as most liked in 1982 (when nearly 60% of Americans said they liked it), that dropped nearly 20 points to 40% liking it in 2008. That put country music well behind the 48% of the 2012 public saying they liked rock (or classic rock) music, the new most liked of the music genres presented to 2012 SPPA respondents. In terms of when asked to name their *single* most favored music, rock music was also highest at 26%, with country music at 17% (and down only from the 20% ranking country music highest in 1982).

Classical music, a more elitist or 'high-brow' of the SPPA music genres, stayed about the same 26–28% liking level as it had in 1982 – staying at about on its par with jazz, blues, and hymns/gospel music. Liking of opera (perhaps the most elitist genre) remained at the bottom of liked music, but again close to its 10% liking level in 1982.

While there are significant education differences in these rankings, respondent education does little to explain the decline in liking of country music – which happened in all education groups (in contrast to classical music which showed remarkable stability by education).

Conspicuous Consumption

Veblen (1899) first argued that Americans at the time engaged in conspicuous consumption as their way of expressing their culture and values. Sullivan and Gershuny (2004) have more recently argued it is inconspicuous consumption. One answer may lie in the argument that the costs of the 'survival market basket' keep increasing as community norms and expectations about possessions keep rising. Another is the increased media emphasis on the consumption of goods and services, as characterized in Schor's (1999) 'Overspent American' and by Ritzer's (1999) depiction of 'Cathedrals of Consumption' – the ever-more elegant shopping malls, vacation cruises, and home entertainment systems on which Americans spend their work incomes. The hours invested in these goods and locales suggest that the 'culture' being consumed with their work dollars is rather superficial, redundant, and spiritually uninspiring in contrast to the potentials of classic leisure.

Subjective Aspects of Free Time

Since leisure has been argued to be a fundamentally subjective state, it should require more subjective measures to reflect it. Four types of subjective measures that have been proposed as reflective of greater leisure or quality of life include the following.

Greater Enjoyment

In terms of overall happiness, life satisfaction, or quality of life; there has been little change in the roughly 33% who say they are 'very happy' since 1972 (although there has been about a 2% increase in saying 'not very happy' since 2008). This does not seem to reflect the small increase in free time found in Table 1, if one might expect happiness to increase with more free time.

When asked to rate the enjoyment levels of separate different activities in their ATUS time diaries, respondents generally rate their free-time activities higher than housework, paid work, or other 'productive' or 'necessary' activities, but with several exceptions. These subjective well-being ratings among ATUS respondents indicate that the most positively rated diary activities involve social interaction – in reading and playing with children, in volunteering activities, in religious practice. Lower enjoyment is found for routine housework and work activities, with medical care and job searching being least enjoyed (Robinson, 2014). Many prominent free-time activities rate below average in enjoyment, including TV, communicating by mail and other media, and even relaxing.

However, earlier surveys found that when asked simply to rate them on a general side-by-side basis, most workers said they enjoyed their work and free-time activities equally (Robinson and Godbey, 1999). Saturday and Sunday activities, when free time is 50% higher per day than during the week, are rated as no more enjoyable than weekday activities, further supporting the indistinguishability of work and free time.

Less Time Pressure or Fatigue

Also despite the increased free time in Figure 1, only slightly more people now describe their lives as 'always rushed' than they did 45 years ago, and as having less discretionary 'time on their hands'. Both of these less pressured perceptions, however, are associated with notably higher levels of personal happiness (Robinson, 2012). Indeed, almost half respondents who said they almost never felt rushed *and* almost never had unwanted time on their hands said they were 'very happy' compared to about 25% of the rest of the public. In separate surveys, more workers continue to report that they would give up a day's pay to get a day off from work, and far more Americans say they would prefer to work longer hours in general – but presumably mainly to earn extra income. Fewer than 25% of workers in 2009 said they were 'very tired' at the end of their work day – much the same as found in 1975. More educated workers are less likely to feel tired, as are male workers in general.

Appreciation of Quality

The SPPA surveys described above do not find increasing proportions of Americans saying that they enjoy more 'sophisticated' forms of music, like classical music or jazz. However, there is anecdotal evidence as well of increased complexity of TV programs (Johnson, 2005) and consumption of more 'gourmet' foods and beverages. One example is in patterns of beer consumption, in which American microbreweries have taken the lead in reinvigorating classic and creating innovative new beer styles. These craft brewers now reaching close to 10% of the beer market, a small percent but double that of a decade earlier.

Time Estimates

Further clues to how Americans feel about their time and leisure emerge from comparing their *estimated* activity times with their presumably more accurate time-diary times. Paid work activities continue to be overestimated but far less than

for housework and other family care activities or volunteering (by 50–100%). In contrast, sleep is *underestimated* by an hour a day (Robinson and Michelson, 2010), while free time is underestimated by 50% or more (about 18 estimated weekly hours vs. closer to 40 diary hours); the GSS estimate question on TV time, surprisingly, is accurately estimated at about 3 h per day. Thus, the lives that Americans think they are leading seem to reflect Weber's 'Protestant ethic' more than the activities respondents report in their 'real-time' diaries.

Summary and Conclusions

While the Figure 1 time-diary trend indicates that Americans of working age still have slightly more free time than they did 50 years ago, Table 2 and Figure 2 from the SPPA suggest declines in the proportions of Americans participating in the arts or in other more serious forms of leisure. GSS data also indicate that these declines in 'serious' have occurred during a time when the college education levels of the US public have improved significantly. Moreover, this increase in public education has not produced a population with improved vocabulary skills according to GSS data (Robinson, 2010). At the same time, Figure 3 shows no notable decrease in the proportions of the Americans who claim to appreciate classical, opera, or jazz music.

Table 1 suggests that the American way of life and leisure seems rather similar to that in other Western countries, in creating more free time and thus more opportunity for leisure, although our hours and proportions of free time devoted to TV have increased. However, significant proportions of Americans still describe their lives as hectic and remain committed to working longer hours than may be necessary, presumably to keep up with the culture of consumerism. Work may provide more quality of life for people than is commonly believed, even if it does not contribute to the ideals of leisure identified by philosophers and leisure scholars.

What is needed to advance our understanding and appreciation of leisure are more systematic qualitative studies of the meaning and benefits of various activities in people's lives (much as begun by Marsden and Swingle, 1994). There are several insightful individual studies of this type (e.g., Viditch and Bensman, 1958; Bellah et al., 1985; Hochschild, 1997; Darrah et al., 2007), but they too rarely connect with one another, address issues raised by quantitative research, or are conducted with representative samples. The meaning of free time and leisure for typical Americans could be the ideal focus for studies using the 'representative ethnography' approach described in Robinson and Meadow (1982).

What is it, then, that would represent uses of free time to satisfy criteria for leisure, or the improved quality of life implied by the term 'leisure'? Various philosophers and social critics offer alternative criteria and rare unanimity, but one might emphasize the following features or experiences:

(a) a (nonwork) skill or talent for its own sake;
(b) a more discriminating palate or set of tastes;
(c) a personal cosmology or philosophy of life;
(d) spiritual regeneration or insight;
(e) stronger ties to some larger community;
(f) greater self-actualization; and
(g) elements of play, curiosity, and wonder.

What seems to be called for at this point, then, is greater attention to operationalization of a set of reliable and valid

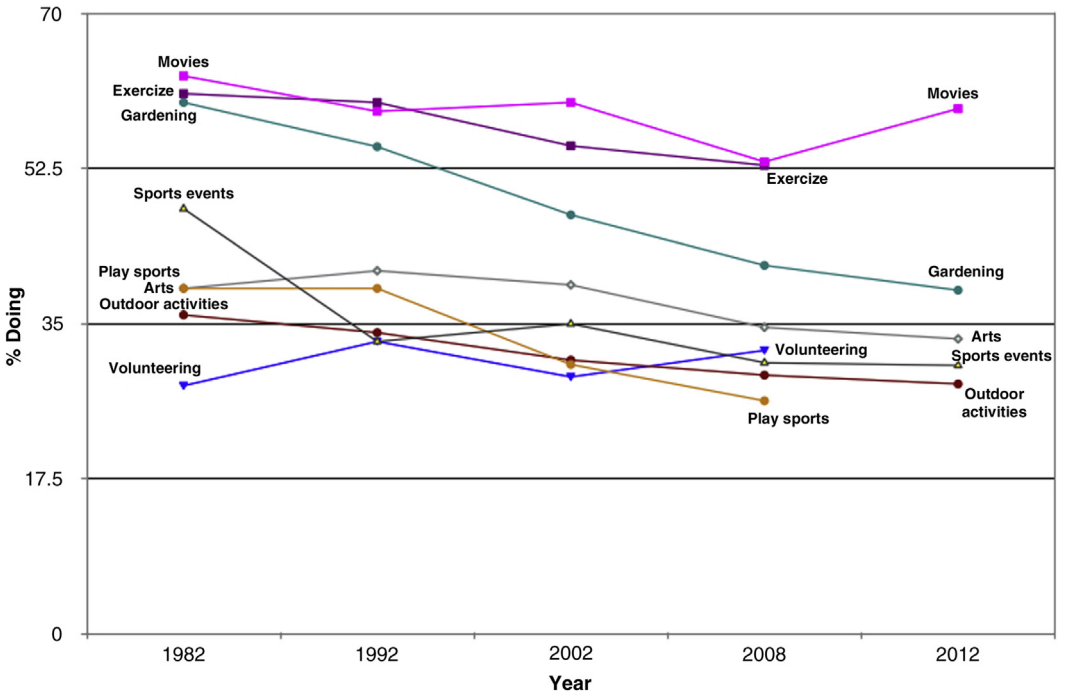

Figure 3 1982–2012 Trends in leisure and arts activities (age 18+). SPPA, 2012.

'leisure indicators' based on such concepts that could be regularly monitored to reflect whether progress is being made on the nonmaterial, as well as the material, quality of the public's leisure, *and* work time.

See also: Art and Culture, Economics of; Consumer Psychology; Consumption, Sociology of; Cultural Participation, Trends In; Culture Policy Regimes in China; Entertainment; Leisure and Cultural Consumption: The European Perspective; Leisure, Sociology of; Time Use Research Methods.

Bibliography

Aguiar, M., Hurst, E., 2009. A summary of trends in American time allocation. Soc. Indicators Res. 92, 70–80.

Ausubel, J., Grubler, A., 1994. Working Less and Living Longer: Long Term Trends in Working Time and Time Budgets. Working Paper 94–99. International Institute for Applied Systems Analysis, Laxenberg, Austria.

Bellah, R., Madsen, R., Sullivan, W., Swidler, A., Tipton, S., 1985. Habits of the Heart. University of California Press, Berkeley, CA.

Bittman, M., 1998. The land of the long lost weekend: trends in free time among working age Australians, 1974–1992. Society and Leisure 21 (2), 353–379.

Csikszentmihalyi, M., 1991. Flow: The Psychology of Optimal Experience. Harper Collins, New York.

Darrah, C., Freeman, J., English-Lueck, J., 2007. Busier Than Ever! Why American Families Can't Slow Down. Stanford University Press, Stanford, CA.

Fisher, K., Robinson, J., 2011. Daily life in 23 countries. Social Indicators Research 101, 295–304.

Gershuny, J., 2009. Veblen in reverse. Social Indicators Research 93, 37–45.

Goldschmidt-Claremont, C., Pagnossin-Aligisakis, E., 1992. Measures of unrecorded activities in fourteen countries. In: Human Development Report, vol. 20. United Nations Development Office, New York.

de Grazia, S., 1962. Of Time Work and Leisure. Twentieth Century Fund, New York.

Hochschild, A., 1997. The Time Bind. Henry Holt, New York.

Johnson, S., 2005. Everything Bad Is Good for You. Riverhead, New York.

Kubey, R., Czikszentmihalyi, 1991. Television and the Quality of Life. Earlbaum, New York.

Mack, A. (Ed.), 2005. Busyness. Social Research, 72 (2).

Mander, J., 1977. Four Arguments for the Elimination of Television. Morrow, New York.

Marsden, P.V., Swingle, J.F., 1994. Conceptualizing and measuring culture in surveys: values, strategies and symbols. Poetics 22 (4), 263–289.

Postman, N., 1987. Amusing Ourselves to Death. Penguin, New York.

Putnam, R., 2000. Bowling Alone. Simon and Schuster, New York.

Reisman, D., 1958. Leisure and work in post-industrial society. In: Meyersohn, R., Larabee, E. (Eds.), Mass Leisure. Free Press, Glencoe, IL, pp. 363–388.

Ritzer, G., 1999. Enchanting a Disenchanted World. Pine Forge Press, Thousand Oaks, CA.

Robinson, J., 2009. Social change in America: 1965–2005. Soc. Indicators Res. 93, 40–50.

Robinson, J., October 2010. Sex, arts and verbal abilities: three further indicators of how American life has not improved. Social Indicators Research 99 (1), 1–12.

Robinson, J., 2012. Americans less rushed, but no happier: 1965–2010 trends in subjective time and happiness. Social Indicators Research 113 (3), 1091–1104 (online November 2012).

Robinson, J., 2013. Happiness is being rushed just enough. Scientific American (online Feburary 19).

Robinson, J., 2014. As We (Still) Like it: Socializing, Religion, Kids Remain our Favorite Daily Activities. Working paper No. PWP-MPRC-2013-024. Maryland Population Research Center (accessed 31.1.14.). http://papers.ccpr.ucla.edu/papers/PWP-MPRC-2013-024/PWP-MPRC-2013-024.pdf.

Robinson, J., Caporaso, A., December 2009. Senioritis in repose. Electronic International Journal of Time Use Research (eIJTUR) 6 (2), 306–313.

Robinson, J., Godbey, G., 1999. Time for Life. Penn State Press, State College, PA.

Robinson, J., Martin, S., 2008. What do happy people do? Social Indicators Research 89, 565–571.

Robinson, J., Martin, S., September 2009. Of time and television. The Annals of Social and Political Science (The End of Television) 265, 74–87.

Robinson, J., Meadow, R., 1982. Polls Apart. Seven Locks Press, Cabin John, MD.

Robinson, J., Michelson, W., November 2010. Sleep as a victim of the time crunch – a multinational analysis. Electronic International Journal of Time-Use Research 7 (1), 61–72.

Robinson, J., Martin, S., Glorieux, I., Mirren, J., June 2011. The overestimated workweek revisited. Monthly Labor Review, 1–13.

Schor, J., 1991. The Overworked American. Basic Books, New York.

Schor, J., 1999. The Overspent American. Basic Books, New York.

Sullivan, O., Gershuny, J., March 2004. Inconspicuous consumption. Journal of Consumer Culture 4 (1), 79–100.

Veblen, T., 1899. Theory of the Leisure Class. Heubsch, New York.

Viditch, A., Bensman, J., 1958. Small Town in Mass Society. Princeton University Press, Princeton, NJ.

Zill, N., Winglee, M., 1990. Who Reads Literature? Seven Locks Press, Cabin John, MD.

Relevant Websites

ATUS03t07.

ATUS-X.

GOOGLE.

NORC/gss.

http://arts.gov/news/2013/national-endowment-arts-presents-highlights-2012-survey-public-participation-arts – SPPA 2012 Highlights report: 2012 Survey of Public Participation in the Arts September 26, 2013.

TIMEUSE.org.

Leisure, Sociology of

Ken Roberts, University of Liverpool, Liverpool, England, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by N. Woolsey Biggart, volume 13, pp. 8717–8720, © Elsevier Ltd.

Abstract

This article reviews the topics that have been investigated, the research methods that have been employed, the main theoretical perspectives, agreed conclusions, and issues that remain unresolved since this subdiscipline was formed in the 1960s. We see that solid advances in knowledge have been accomplished through research into the relationships between the uses of leisure and occupations, gender roles, and life course stages. Meanwhile, grand claims about the role of leisure in society have yet to be developed into testable theories about the roles of particular kinds of leisure under specified circumstances. Also, the sociology of leisure has yet to embrace the likelihood that the western version is not necessarily the sole form of modern leisure.

Theoretical Approaches

History and Status within Sociology

The sociology of leisure was created as a subdiscipline in the 1960s, as indicated by the formation of the International Sociological Association's Research Committee 13 (RC13) in 1970. Sociologists had studied leisure previously (see in particular De Grazia, 1962; Lundberg et al., 1934; Veblen, 1899), but usually as an offshoot of studies of workers in particular occupations or in family and neighborhood research. From these beginnings, right through to the present day, this subdiscipline has encountered difficulties in defining its boundaries and claiming possession of its field. There are several reasons for this.

Leisure Studies

At the same time, (in the 1960s), when the sociology of leisure was being institutionalized, universities across North America were creating a subject called 'leisure studies.' Initially the undergraduate and postgraduate programs were preparing students for careers in, and faculty were conducting investigations relevant to, parks and other public recreation services; but the curricula and the research interests of the faculty spanned the whole of leisure. From the 1970s onward, leisure studies departments and courses spread throughout universities in the UK, Australia, and New Zealand, and then, albeit far less densely, throughout the rest of Europe, Asia, and Latin America. A consequence is that everywhere in the world, sociologists who study leisure have been as likely, probably more likely, to be based in leisure studies rather than squarely in sociology departments; and within leisure studies, the boundaries between sociological and other contributions have never been clear.

Challenges from within Sociology

Within its own discipline the sociology of leisure has always faced rival claims on its field. Initially the main challenge was from cultural studies, especially the genre that examined popular cultures. Media sociology has always positioned itself outside the sociology of leisure, and likewise the sociology of the arts. Youth researchers have been more likely to investigate youth *cultures* than youth *leisure*. Another strong challenge during the last 30 years has been from the sociology of consumption.

Fragmentation within

In some respects, the sociology of leisure has fallen victim to its own success in reaching a size where it has been possible for specialists in sport and tourism to organize their own conferences, form scholarly associations, and launch their own journals. These sources of fragmentation affect all disciplines but leisure may be especially prone to be affected by it. Students, teachers, and researchers are far more likely to be enthusiastic, even passionate, about a particular art form or sport than about leisure in general. Moreover, the term 'leisure industries' is not a composite category. Providers organize themselves into tourism, TV, publishing, computer games, etc.

The sociology of leisure has needed, and needs perpetually, to demonstrate that there is value added by scanning the entire field of leisure. It needs to show that inserting an intermediate level of analysis is preferable to leaping straight from data on sport or tourism to theories about society at large. It also needs to show that leisure is a concept with strengths that complement those of consumption and culture; and that sociologists, against leisure scholars from other disciplines, have something extra, something distinctively sociological and valuable, that they can contribute.

Findings

Work and Leisure

This was originally the main research issue in the sociology of leisure. It became so simultaneously, in three different ways. First, *modern* leisure was treated as a product of the modernization of work. Second, a long-term growth of leisure was seen as an outcome of the shrinkage of work time. Third, uses of leisure offered additional evidence about the 'long arm of the job.'

Modern Work, Modern Leisure

Most sociologists have treated 'modern' leisure as a historical outcome – a product of the broader division of labor associated with modernization – rather than a socio-cultural universal construct. There is continuing debate about whether a leisure concept can be usefully applied to the free time, amusements, and pastimes of premodern eras, and whether these societies had a different kind of leisure. One view is that modern leisure is a debased version of the purer leisure enjoyed in ancient Greece, albeit just by a minority of free men (see De Grazia, 1962).

That said, it has been agreed that during the industrial revolution, paid work was relocated into offices, mines, and factories. In industry employees worked at set times fixed by the clock, under work-specific authority, and often amid work-specific relationships with colleagues. Modernized industrial work was organized rationally – to minimize unit costs and to maximize output per unit of resources invested. Customary ways were swept aside, if necessary. Work was certainly not organized to maximize job satisfaction. Thus work became a part of life, and leisure, while not occupying the whole of nonwork time, was located within it.

Work Time and Leisure Time

The pioneers of the sociology of leisure in the 1960s were inspired, above all else, by the historical growth of leisure time. They knew that standards of living were rising; they wrote during the first decades of so-called mass affluence. They knew that the benefits of economic growth were being taken partly in the form of higher real earnings and higher levels of consumer spending, but they were more impressed by the extent to which working time had been rolled back. Since the industrial revolution, the length of a normal work day had been reduced from 12 h to 10 h, and then to 8 h. The weekend had expanded from one day to one-and-a-half days, and then to two full days. The annual holiday entitlement had grown. The normal working life had been trimmed at both ends. Thus in the 1960s Joffre Dumazedier (1967) was able to proclaim a great historical inversion – the typical worker was spending more time at leisure than at work. There appeared to be no reason why these trends should not continue.

Some sociologists were cautious. Harold Wilenski (1963) noted that most of the extra free time created since the early decades of industrialism had enlarged the size of groups outside the paid workforce (mainly the young and the retired); that the decline in hours of paid work had affected only manual employees; and that their gains had achieved no more than winning back the free time that was lost during industrialization. Staffan Linder (1970), an economist, noted that leisure time was increasing more slowly than consumer spending power, and forecast the growth of a 'harried leisure class' whose main problem would be 'finding the time,' Geoff Godbey (1975) claimed that in the United States it was not leisure, but a condition that he described as 'anti-leisure' that was growing – time-pressured compulsive activity and consumption. These warning voices have proved remarkably prescient.

The terms of the work time/leisure time issue changed abruptly in 1991.We can date this precisely because the catalyst was the publication of Juliette Schor's book, *The Overworked American*. Schor, another economist, claimed that in America, working time was lengthening, that Americans were working too long for their own good, pressured by greedy employers and the 'addictive power of consumption,' Schor's book ignited a search for other countries where working time was lengthening (see Zuzanek et al., 1998). Up to now there is no country where a general lengthening of work time has been confirmed (as opposed to alleged). In the United States, Schor's claim has not been corroborated by time budget evidence (see Robinson and Godbey, 1999). However, by the end of the 1990s there was agreement that the earlier shortening of work schedules had ended, probably in the 1970s; this consensus has not prevented 'work–life balance' becoming a public and a thoroughly researched issue. Although there are 'time pioneers' (see Horning et al., 1995) who have deliberately opted to downshift to enjoy slower and more frugal lifestyles, up to now these pioneers have not been trendsetters.

The Long Arm of the Job

This phrase underlines the extent to which people's jobs control not only their lives at work but also what they do in their 'own' time. From the very beginnings of leisure research it has been noted repeatedly that the middle classes 'do more,' the main exception today being television viewing.

However, in the 1950s and 1960s, rather more attention was paid to how particular occupations tended to foster distinctive uses of leisure. This was observed among deep-sea trawlermen (Tunstall, 1962), coal miners (Dennis et al., 1956), assembly line workers (Friedmann, 1961), and so-called organization men (Whyte, 1957). Spillover and compensatory work–leisure relationships were distinguished. Stanley Parker (1971) relabeled and expanded these concepts into an extension, neutrality, and opposition typology of work–leisure relationships, arguing that extension was most likely when people found their jobs interesting, when they identified with their work roles, and with work colleagues also. In contrast, opposition was likely when people found their jobs disagreeable and celebrated the end of work shifts by doing things that were enjoyed largely because they were different.

In the mid-1990s, initially in North America, researchers discovered that the middle classes (managers and professional-grade employees) were typically leisure omnivores, characterized by the wide variety of their leisure tastes and activities (Erikson, 1996; Peterson and Kern, 1996). Subsequently middle class omnivores have been spotted all over the world (see, for example, Sintas and Alvarez, 2002). The middle classes not only dine out more frequently than the working classes but do so in a wider variety of catering establishments (Warde et al., 1999). The middle classes now produce most rock and pop as well as classical musicians. One explanation of this phenomenon is the scale of upward social mobility into expanding middle-class occupations (see Eijck, 1999).

Unemployment

There was a flurry of studies into the leisure of the unemployed in the 1980s, when mass unemployment had returned to the Western world after the period now recalled as the '30 glorious years' – the post-World War II decades of full employment and

steady economic growth. The initial questions posed about the leisure of the unemployed were soon answered. We learnt quickly that the unemployed tend to 'do less' for a variety of reasons – financial, loss of work-based social relationships, and loss of status (reluctance to expose spoiled identities to the public gaze). Does participation in leisure activities help to maintain the well-being and morale of the unemployed? Yes, when levels of leisure activity are maintained (which is exceptional), but even then, not at the levels of well-being that are normal within the employed population (see Glyptis, 1989; Havitz et al., 2004).

Leisure Work

Higher levels of leisure spending and leisure activity have led to the expansion of leisure industries, which have become a main, if not the main, source of new jobs in countries all over the world.

Leisure industries include some of the so-called creative industries, whose main asset is intellectual property rights (for example, computer software, musical and other kinds of 'texts'). Some leisure industries have glamor roles (sports and screen stars). Employees can be attracted into mundane jobs by their proximity to glamor, but overall the leisure industries contain many mundane occupations. Many of the jobs are seasonal (especially in tourism), part-time, and involve being on duty at unsocial hours (jobs in the nighttime economy, for example). Employees are often required to perform aesthetic labor (see Warhurst and Nickson, 2007), where the appearance and personality of the employee are part of the service rendered to customers.

Gender and Leisure

During the 1980s this topic replaced work as the leading issue in the sociology of leisure. The second wave of feminism forced 'gender' to move up the research agenda. Feminists' initial complaints were that leisure research had neglected gender in general and women's leisure in particular. These criticisms were immediately accepted and it is now 20 years since it was possible to complain that gender or women's leisure are neglected. The original questions raised about leisure and gender have now been answered, and new issues have taken their place.

Women and Patriarchy

An initial claim was that, in leisure, women were the disadvantaged sex due to a combination of time and money constraints and heavier socio-cultural regulation than experienced by men (see Deem, 1986; Green et al., 1990; Henderson et al., 1989). Women's leisure time disadvantages arise from the double shift, with housework and child care remaining primarily women's work (see Bittman and Wajcman, 1999). As regards money, it is accepted that women have been and are still disadvantaged. Men are more likely to be in employment, earning their own money. Women in employment generally earn less than men. Household income is not always shared out equally between males and females. Women are less likely than men to be able to separate their 'own funds' from household funds (Pahl, 1990). Socio-cultural regulation makes it more improper for women than for men to consume alcohol fulsomely and become intoxicated; to play strenuous, aggressive, competitive sports; to go alone to cinemas and other leisure places; and to enjoy leisure in mixed-sex company when not accompanied by their sexual partners. However, researchers have noted that times are changing. In Western countries the differences between men's and women's uses of leisure have narrowed.

Cross-cutting the debate about the extent to which the leisure of one sex has been and remains disadvantaged, there have been claims that women's leisure is qualitatively different to men's leisure. It has been claimed that mainstream leisure research has proceeded unwittingly with a masculine concept of leisure, and that, as a result, women's leisure has been misrepresented. The main gender differences are said to be, first, that men associate leisure with particular times (when they are not at work) whereas women tend to identify leisure with particular kinds of experiences, activities, and being with particular people. Second, it is alleged that men tend to appraise their leisure quantitatively (how much time they have and how much they do in this time) whereas women's appraisals are in terms of the quality of their experiences (see Gregory, 1982; Lenskyj, 1988).

Men and Masculinity

Men's studies require that masculinity (just like femininity) be treated as a problematic social construct, following which we recognize that there can be different versions of masculinity (maybe specific to social classes or ethnic groups), and that some men may find their masculine scripts just as constraining as femininity is for some women.

As yet there are no signs of men collectively rejecting all forms of masculinity. Rather, we now have abundant evidence of men using leisure to retain conventional masculine identities in inhospitable historical contexts. Deindustrialization has destroyed swathes of jobs that required toughness and endurance. The advent of greater independence for women has deprived males of the traditional breadwinner role. One male response has been to use leisure to maintain conventional masculine identities (see Blackshaw, 2003).

Sexualities

Most leisure sites are places where sexual roles and identities are celebrated and enacted. Anyone who enters a mainstream leisure facility is likely to be treated as a sexualized subject, and to feel able to treat others similarly. Women in particular are likely to be subject to sought or unwanted attention.

Sociology's interest has been, first, in the use of leisure spaces to develop and enjoy sexual partnerships, and second, in how the operation of heteronormativity marginalizes people with other sexual identities – gays and lesbians, and presumably pedophiles also. Breaching heteronormativity still invites ridicule or worse in sports clubs and, indeed, in most leisure places (see Wellard, 2002). A response in sociology has been the development of 'Queer' studies, in this case viewing leisure sites from the perspectives of nonheteros (see Cauldwell, 2006). A response in the wider western societies has been for gays, lesbians, and bisexuals to create their own leisure spaces – bars, villages, sport clubs, etc (see Skeggs, 1999).

The Life Course

Leisure and Age

Sociological work on leisure and life course really began with the Rapoports' pioneering study in London in the 1970s (Rapoport and Rapoport, 1975). These investigators observed how the family life cycle created different preoccupations and interests at successive life stages, which interacted with employment careers to produce characteristic uses of leisure. Shortly afterward in the United States, Estes and Wilenski (1978) identified the 'life cycle squeeze.' This refers to the squeeze on adults' time and money when they embark on the formation of new households and families. Family life events are still the main junctures where leisure patterns unfreeze and are reconstituted (Gershuny, 2003).

In recent years, there has been an upsurge of interest in leisure in later life; a response to increased longevity, the supersized baby boomer cohorts approaching and entering later life, and evidence (see also below) that leisure activities are especially beneficial for the well-being of older people, for whom 'satisfiers' connected with work and family roles are less likely to operate.

Leisure Biographies

Some of the most thought-provoking research findings have been from studies that have traced individuals' leisure biographies, sometimes with retrospective data (there can be recall problems) and sometimes with panel methods. These studies have revealed powerful long-term continuities in leisure interests and activities. People who play sport and hold office in voluntary associations during late adulthood, were typically playing these same roles (though not necessarily in the same sports or associations) when they were young adults (McGuire et al., 1987; Mihalik et al., 1989; Roberts et al., 1991; Scott and Willits, 1989). Childhood and youth are the prime life stages when individuals form leisure interests and acquire basic skills. In some individuals these leisure interests and skills may lapse, in which case they are less likely to be resumed than in the case of those who have remained continuously involved and active.

Age, Cohort, and Generation

Current and earlier cohorts of older people were children at a time when opportunities to form leisure interests and to practise such activities were far fewer than today. So the low rates of leisure participation in current older age groups may not be straight-forward age effects but could be a cohort effect, due to the poverty of childhood leisure socialization years ago.

The baby boomer cohorts who are now approaching or entering retirement are the first cohorts to have grown up in postscarcity societies. The Woopies (well-off older people) who remain highly leisure active, are a minority within current cohorts but will become more numerous if the baby boomers turn out to be a 'change generation' that remains willing to take on debt and, if necessary, to spend capital that is tied up in their dwellings, in order to sustain active leisure for as long as they are physically able.

Leisure and Well-Being

In leisure studies (rather than just the sociology of leisure) the leading theories in recent years have been about whether, and if so how, leisure can improve personal well-being. We now know how leisure can be a source of 'optimal experience,' what Csikszentmihalyi (1990) calls 'flow.' We know about the special satisfactions that arise from serious leisure (Stebbins, 1992). We also know that all uses of leisure raise levels of self-reported life satisfaction provided the leisure is active (takes people out of their homes), social rather than solitary, and structured so that the behavior can become routine, making personal motivation on each occasion unnecessary. These same uses of leisure also improve physical health, but here physically active leisure is especially beneficial (see Isao-Ahola and Mannell, 2004).

Leisure and Status Attainment

Sociologists have investigated the role of leisure, if any, in lifetime status attainment. There are never-ending leisure projects to save youth judged to be at risk, excluded, or impeded by multiple disadvantages. Although many of these projects are monitored and can usually produce enough signs of positive outcomes to secure continued funding, the social science verdict is that the evidence remains inconclusive (see Coalter, 2007; Collins and Kay, 2003).

The Societal Role of Leisure

Parsonian functionalism gives leisure a latency role – allowing pentup emotions and other otherwise suppressed drives in personality systems to be expressed (Parsons, 1951). The Frankfurt School's critique of the culture industry claimed that its role was to pacify a potentially rebellious working class (see Bottomore, 1984). More recent neo-Marxist critiques of the consumer society deploy a similar argument about consumer culture's soothing balm (Baudrillard, 1998).

Joffre Dumazedier (1967), the first president of RC13, had different ideas. As leisure grew in scale it was expected to play a stronger role in society. Specifically, Dumazedier treated leisure as a source of values (choosing and doing things for intrinsic satisfaction) that he expected to invade other domains like work places and families. Subsequently these ideas have been reformulated with 'identity' replacing 'values' as the central concept. Postmodern conditions are said to have weakened modern structures and their associated roles (occupational and gender roles, for example), whereas people are able to use consumer roles to tell themselves and others who they are.

Leisure and Comparative Sociology

The comparative sociology of leisure remains in a prolonged infancy. The earliest and the most recent comparative studies have simply drawn together accounts of leisure in their countries by country-based authors, writing to a common template (Cushman et al., 2005; Szalai, 1972). The European Time Use Survey (see MacInnes, 2006) has made progress in harmonizing data, but only for a limited number of European countries, and the evidence is purely quantitative.

Quantitative comparisons have found very similar uses of leisure in all modern societies; the impressive amounts of time spent watching TV, for example. These comparisons have also

found similar social class, gender, and age differences in the distribution of leisure time and participation in different leisure activities.

Since the 1960s the sociology of leisure has been developed mainly by western scholars, and those based in other parts of the world have been more likely to adopt than to challenge Western sociology's definitions and classifications of leisure activities. However, leisure is now being studied by sociologists in all parts of the world: in Eastern Europe, Africa, the Middle-East, Latin America, and especially in the Asian countries that are becoming major global economic centers (see Amaral, 2008; Eglite, 2007; Jain, 2007; Qiyan Wang, 2007; Saad, 2007; Tsai, 2008). It is noteworthy that the cultural specifics of Western leisure are currently being highlighted by scholars from, or writing about, leisure in India, Japan, and China (Fox and Klaiber, 2006; Ma Huidi and Liu Er, 2009; Iwasaki et al., 2007).

Assessment, Prospects, Future Directions

The sociology of leisure was established as a collective enterprise at a time when major new technologies of the early- and mid-twentieth century – radio, movies, recorded music, television, the motor car, and air travel – had already changed or had begun to change, people's lives in first world countries. The relevant changes were occurring at a time when paid work schedules were contracting and disposable incomes (not required for essentials) were growing.

The present-day sociology of leisure has far less to say about the future than was the case in the 1960s and 1970s. This is despite the arrival of another generation of new technologies – the information and communication technologies – that have delivered multichannel television, personal computers, mobile phones-plus, and the internet. It is also despite confident predictions of climate change that will be only ameliorated by changes in how we produce and consume, and knowledge that the planetary resources on which current first world lifestyles depend will one day be exhausted. Sociology in general, not just the sociology of leisure, has become cautious. We know that many future forecasts made in the mid-twentieth century have been confounded.

Sociologists of leisure know that their subdiscipline needs theoretical renewal. Building squarely on the foundations already laid cannot be sufficient, given that these foundations have always been subject to criticism (see, in particular, Rojek, 1995, 2009). It is likely that some of the most exciting, paradigm-shattering sociological work on leisure will forthwith be conducted outside the first world countries where the subdiscipline was born. Are the new market economies of Eastern Europe, Latin America, Africa, and Asia still catching up; or are late-development effects propelling these countries ahead in certain respects? Leisure is a site where issues of much wider interest within and beyond sociology can now be explored.

See also: Consumption, Sociology of; Culture, Sociology of; Sport, Sociology of; Work, Sociology of; Youth Culture, Sociology of.

Bibliography

Amaral, S.C.F., 2008. Public leisure policies in Porto Alegre, Brazil: from representative democracy to participant democracy. World Leisure Journal 50 (2), 127–137.

Baudrillard, J., 1998. The Consumer Society. Sage, London.

Bittman, M., Wajcman, J., 1999. The Rush Hour: The Quality of Leisure Time and Gender Equity. SPRC Discussion Paper 97. University of New South Wales.

Blackshaw, T., 2003. Leisure Life: Myth, Masculinity and Modernity. Routledge, London.

Bottomore, T., 1984. The Frankfurt School. Tavistock, London.

Cauldwell, J. (Ed.), 2006. Sport, Sexualities and Queer Theory. Routledge, London.

Coalter, F., 2007. A Wider Role for Sport: Who's Keeping the Score? Routledge, London.

Collins, M.F., Kay, T., 2003. Sport and Social Exclusion. Routledge, London.

Csikszentmihalyi, M., 1990. Flow: The Psychology of Optimal Experience. Harper and Row, New York.

Cushman, G., Veal, A.J., Zuzanek, J. (Eds.), 2005. Free Time and Leisure Participation: International Perspectives. CABI, Wallingford.

De Grazia, S., 1962. Of Time, Work and Leisure. Twentieth Century Fund, New York.

Deem, R., 1986. All Work and No Play? Open University Press, Milton Keynes.

Dennis, N., Henriques, F., Slaughter, C., 1956. Coal Is Our Life. Eyre and Spottiswoode, London.

Dumazedier, J., 1967. Towards a Society of Leisure. Free Press, New York.

Eglite, P., 2007. Leisure choice and its fulfillment in Latvia. World Leisure Journal 49 (1), 24–29.

Eijck, K van, 1999. Socialisation, education and lifestyle: how social mobility increases the cultural heterogeneity of status groups. Poetics 26, 309–328.

Erikson, B.H., 1996. Culture, class and connections. American Journal of Sociology 102, 217–251.

Estes, R.J., Wilenski, H., 1978. Life-Cycle Squeeze and the Morale Curve. Institute of Industrial Relations, Reprint 422. University of California, Berkeley.

Fox, K.M., Klaiber, E., 2006. Listening for a leisure remix. Leisure Sciences 28, 411–430.

Friedmann, G., 1961. The Anatomy of Work. Heinemann, London.

Gershuny, J., 2003. Time through the Lifecourse, in the Family. ISER Working Paper 2003-3. University of Essex, Colchester.

Glyptis, S., 1989. Leisure and Unemployment. Open University Press, Milton Keynes.

Godbey, G., 1975. Anti-leisure and public recreation policy. In: Parker, S., et al. (Eds.), Sport and Leisure in Contemporary Society. Leisure Studies Association, London.

Green, E., Hebron, S., Woodward, D., 1990. Women's Leisure, What Leisure? Macmillan, London.

Gregory, S., 1982. Women among others: another view. Leisure Studies 1, 47–52.

Havitz, M.E., Morden, P.A., Samdahl, D.M., 2004. The Diverse Worlds of Unemployed Adults: Consequences for Leisure, Lifestyle and Well-being. Wilfrid Laurier University Press, Waterloo.

Henderson, K., Bialeschki, M.D., Shaw, S.C., Freysinger, V.J., 1989. A Leisure of One's Own. Venture, Pennsylvania.

Horning, K.H., Gerhard, A., Michailow, M., 1995. Time Pioneers: Flexible Working Time and New Lifestyles. Polity Press, Cambridge.

Huidi, Ma, Er, Liu, 2009. Social transformation: the value of traditional leisure culture of China revisited. World Leisure Journal 51, 3–13.

Isao-Ahola, S.E., Mannell, R.C., 2004. Leisure and health. In: Haworth, J.T., Veal, A.J. (Eds.), Work and Leisure. Routledge, London, pp. 184–199.

Iwasaki, Y., Nishino, H., Onda, T., Bowling, C., 2007. Leisure research in a global world: time to reverse the western dominance in leisure research. Leisure Sciences 29, 113–117.

Jain, R., 2007. Access to leisure: impact of forces of liberalisation, privatisation and globalisation on socially disadvantaged groups in India. World Leisure Journal 49 (1), 15–23.

Lenskyj, H., 1988. Measured time, women, sport and leisure. Leisure Studies 7, 233–240.

Linder, S., 1970. The Harried Leisure Class. Columbia University Press, New York.

Lundberg, G.A., Komarovsky, M., McInerny, M.A., 1934. Leisure: A Suburban Study. Columbia University Press, New York.

MacInnes, J., 2006. Work-life balance in Europe: a response to the baby bust or reward for the baby boomers. European Societies 8, 223–249.

McGuire, F.A., Dottavio, F.D., O'Leary, J.T., 1987. The relationship of early life experiences to later life leisure involvement. Leisure Sciences 9, 251–257.

Mihalik, B.J., O'Leary, J.T., Maguire, F.A., Dottavio, F.D., 1989. Sports involvement across the life span: expansion and contraction of sports activities. Research Quarterly for Exercise and Sport 60, 396–398.
Pahl, J., 1990. Household spending, personal spending and the control of money in marriage. Sociology 24, 119–138.
Parker, S., 1971. The Future of Work and Leisure. MacGibbon and Kee, London.
Parsons, T., 1951. The Social System. Routledge, London.
Peterson, R.A., Kern, R.M., 1996. Changing highbrow taste: from snob to omnivore. American Sociological Review 61, 900–907.
Rapoport, R., Rapoport, R.N., 1975. Leisure and the Family Life-Cycle. Routledge, London.
Roberts, K., Minten, J.H., Chadwick, C., Lamb, K.L., Brodie, D.A., 1991. Sporting lives: a case study of leisure careers. Society and Leisure 14, 261–284.
Robinson, J.P., Godbey, G., 1999. Time for Life: The Surprising Ways Americans Use Their Time, second ed. Pennsylvania State University Press, Penn State.
Rojek, C., 1995. Decentring Leisure. Sage, London.
Rojek, C., 2009. The Labour of Leisure. Sage, London.
Saad, N.M., 2007. Egyptian Moslem mothers and their leisure patterns. World Leisure Journal 49 (1), 44–51.
Schor, J.B., 1991. The Overworked American. Basic Books, New York.
Scott, D., Willits, F.K., 1989. Adolescent and adult leisure patterns: a 37 year follow-up study. Leisure Sciences 11, 323–335.
Sintas, J.L., Alvarez, E.G., 2002. Omnivores show up again: the segmentation of cultural consumers in Spanish social space. European Sociological Review 18, 353–368.
Skeggs, B., 1999. Matter out of place: visibility and sexuality in leisure spaces. Leisure Studies, 213–232.
Stebbins, R.A., 1992. Amateurs, Professionals and Serious Leisure. McGill-Queens University Press, Montreal.
Szalai, A. (Ed.), 1972. The Use of Time. Mouton, The Hague.
Tsai, C.-T.L., 2008. Gender and leisure: influencing school PE policy in Taiwan. World Leisure Journal 50 (4), 275–284.
Tunstall, J., 1962. The Fishermen. MacGibbon and Kee, London.
Veblen, T., 1899/1970. The Theory of the Leisure Class. Allen and Unwin, London.
Wang, Qiyan, 2007. A comparison of Chinese and Japanese leisure modes. World Leisure Journal 49 (3), 142–154.
Warde, A., Martens, L., Olsen, W., 1999. Consumption and the problem of variety: cultural omnivorousness, social distinction and dining out. Sociology 33, 105–127.
Warhurst, C., Nickson, D., 2007. Employee experience of aesthetic labour in retail and hospitality. Work, Employment and Society 21, 103–120.
Wellard, I., 2002. Men, sport, body performance and the maintenance of "exclusive masculinity". Leisure Studies 21, 235–247.
Whyte, W.H., 1957. The Organization Man. Anchor Books, New York.
Wilenski, H.L., 1963. The uneven distribution of leisure: the impact of economic growth on "free time". In: Smigel, E.O. (Ed.), Work and Leisure. College and University Press, New Haven, pp. 21–40.
Zuzanek, J., Beckers, T., Peters, P., 1998. The harried leisure class revisited: a cross-national and longitudinal perspective. Dutch and Canadian trends in the use of time: from the 1970s to the 1990s. Leisure Studies 17, 1–19.

Lesbian and Gay Activism in Latin America

Jordi Díez, University of Guelph, Guelph, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Lesbian and gay activism in Latin America exhibits features that manifest themselves in other parts of the world, it possesses a variety of specificities that are the result of colonialism, indigenous characteristics, and the influence of global processes. Chief among these are fluid understandings of sexuality, the historical omnipresence of one religion and, more recently, the adoption of more contemporary conceptualizations of sexual citizenship. While movement characteristics vary across Latin America, in recent years, several gay and lesbian movements have assumed a collaborative approach with governments and have been very successful in attaining a variety of policy objectives, ranging from antidiscrimination legislation to same-sex relationship recognition.

Introduction

Latin America's sociopolitical development has historically been influenced by exogenous forces which, blended with indigenous elements, have produced distinct regional characteristics. Colonialism, and the region's dependent integration into the global economic system, have encountered strong local traditions that have created social and political hybridities, making the region difficult to classify: while its cultures possess numerous Western traits, they also exhibit important indigenous characteristics. Lesbian and gay activism in Latin America has emerged from such hybrid contexts and its characteristics are the result of global processes, fluid understandings of sexuality, the historical omnipresence of one religion, and, more recently, the adoption of more contemporary conceptualizations of sexual citizenship.

Regional Understandings of Sexuality

Latin American understandings of sexuality have historically been fluid. The conquest of Latin America by Iberians during the sixteenth century saw an attempt to impose very rigid European definitions of gender and sexuality, which revolved around ideas of patriarchy, sex, reproduction, and gender relations guided by natural law principles. These European definitions date back to the writings of Saint Thomas Aquinas. Thomist philosophy, adopted by the Catholic Church during the Second Spanish Scholasticism, supported a binary conceptualization of gender roles and viewed sexual relations as having as primary goal, the procreation of the species. Same-sex relations therefore violated natural law and 'sodomy' was made the worst type of sin (the heinous crime or *pecado nefando*). The proscription of homosexuality became Church policy in medieval Spain, punishable by death, by the *fuero real*. The colonial enterprise attempted to impose these ideas on the new conquered lands of the Americas by Spanish conquistadors. In 1497, 5 years after the conquest, the Spanish Kings reaffirmed the *fuero real* and upheld the death penalty for the commission of 'heinous crime against nature' (Nesvig, 2001). The Church's official position on homosexuality became colonial policy and it regulated same-sex relations for centuries. Starting with the prosecuting of the first 'sodomite' by Mexican Inquisition in 1524, same-sex relations were systematically persecuted by civilian and church authorities.

However, the colonial enterprise attempted to impose these views on a continent in which inhabitants possessed widely varied understandings of sexuality. While that exists limited research on pre-Columbian sexuality, the one there suggests that some societies did not adhere to uniform norms regulating sexuality that followed dichotomized male–female gender roles based on sexual activity guided by reproduction. In some societies, such as the Mexica, the Zapotec, and the Nahua, hereditary lineage among the political elite was an important mechanism for the maintenance of power, and marriage between a man and a woman played a central role in the process (Sousa, 1997; Buckhart, 1997). In these cases, binary gender systems built around procreation appear to have been important. However, in what were highly stratified societies, evidence suggests that, outside the ruling classes, the existence of nondichotomized gender systems among indigenous peoples was common (Williams, 1986). In effect, scholars have pointed to the existence of a variety of gender, sexual, and social roles that did not conform to Western understandings of dual male–female gender systems and to varied ideas of what constituted appropriate sexual behavior.

Colonialism saw a forceful attempt to impose natural law ideas surrounding sexuality, but fluidity in understanding sexuality has survived. Unlike experiences elsewhere, processes of state formation in Latin America were characterized by top-down dynamics, which were not completely successful in transforming preexisting social relations. This is most obviously observed in the perseverance of indigenous cultures and languages in many parts of the region: the newly created states were not able to reach down to all areas, and semimarginal cultural pockets out of the reach of the state have survived. In many of these culturally autonomous pockets, natural law principles simply did not take hold. A good example is perhaps the *muxe* of Oaxaca, Mexico. Muxes are individuals who identify themselves as a third gender and who practise sexual relations with individuals of different sex (Stephen, 2002). Muxes are accepted by their society.

Fluidity has also remained in nonmarginal societies. The particularities of state formation in Latin America have resulted

in a gap between expected social norms and their compliance. This attitude is well encapsulated by the Latin American adage of *obedezco pero no cumplo* (I obey but do not comply with). While it is commonly associated with the dualism that exists between legal formalism and compliance with the law (lax adherence to the rules), the adage also captures a generalized attitude regarding sexual practices. Despite the separation of Church and State that most countries instituted, with varying degrees, during the nineteenth century, the Church continued to play an important role in dictating and upholding social mores. However, adherence to these norms has for the most part been limited to the public sphere, and this includes sexual practices. While homosexuality has been, and continues to be, stigmatized, the criminal codes of the new republics, which were modeled after the Napoleonic code of 1804, did not criminalize consensual same-sex relationships (with very few exceptions, such as Chile). Unlike some western European countries and their colonies elsewhere, the creation of the category of 'homosexual' in the 1870s did not come about with an explicit legal regulation. The result has been an important dualism between what is socially expected in the public sphere and what individuals practise in private. As Mario Pecheny has argued, invisibility and discretion have been expected of Latin American homosexuals and social interaction around homosexuality follows a double moral standard that condemns homosexual practices in public. Yet it tolerates them provided they are hidden from view (2004). As a result, in Latin America the organization of peoples' lives and the construction of identities around categorized sexualities have historically not been common and many individuals who may engage in same-sex practices do not define themselves as 'gay.' Contemporary Latin American lesbian and gay activism thus emerged from within these broader conceptualizations of strong distinctions between public expression and private sexual behavior.

Early Activism

The dualism that existed between the public and the private realms allowed homosexuals to carve up intimate socializing spaces in Latin America's bigger cities in the 1940s and 1950s. Activities such as the *jornais cariocas*, in Rio de Janeiro, associations made up of *felipitos* in Bogotá, and the lesbian events of *fiesteras*, in Buenos Aires, allowed for such socialization. However, these activities did not have political objectives. Political activism did not emerge until the late 1960s and 1970s in the Argentina, Brazil, and Mexico within a context of broader sociopolitical processes both domestic and international. On the domestic front, the bigger countries began a process of industrialization during the postwar years, which contributed to the expansion of a middle class, growing urbanization, higher levels of education, and increased secularization. A new generation of Latin Americans, one that was more critical of established understandings of social and sexual relations, began to adopt more liberal ideas regarding gender relations and sexuality. This generation began to challenge norms regarding contraception, patriarchy, and traditional values. In several countries of the region multiple countercultural groups emerged. At the same time, liberation theology, which called for a new interpretation of Christian scriptures to bring about social change, was adopted by numerous young Catholics. It spread throughout the region and motivated various sectors of society to mobilize as it rejected passive acceptance of perceived injustices. A new generation of Latin American youths became therefore a lot more receptive to political events taking place abroad as part of a cycle of protest that swept the world by the late 1960s, which expressed itself most notably through the 1968 French student movement, the demonstrations against the Vietnam War and Prague's Spring. This context allowed for the emergence of women's, African-American, indigenous and student movements throughout the region.

Latin American Lesbian, Gay, Transsexual and Bisexual (LGTB) activism emerged within these larger sociopolitical processes (Brown, 2002; Díez, 2011). By the late 1960s, Argentine and Mexican homosexuals decided to organize to discuss collectively their personal experiences with sexuality and the social stigmatization of homosexuality. The first organizations were thus formed. While the Stonewall riots of 1969 of New York are widely viewed as catalyst for the emergence of the gay and lesbian movement in the United States, and according to some, in the world, some Latin American homosexuals began to organize before then. In effect, the first gay organization in Latin America, *Nuestro Mundo* (Our World), was founded in 1967 in Argentina, 2 years before those summer riots. Our World was followed by the formation of the Homosexual Liberation Fronts in 1971 in both Mexico and Argentina. Inspired by the development of the first gay liberationist groups in the United States and some western European countries, and the literature produced by the US organization the *Black Panthers*, these early groups began to articulate a discourse based on the need to liberate homosexuals from the oppression created by social stigmatization.

These organizations were formed mostly by leftist students, intellectuals, and activists from union movements and had a Marxist bent. They were for the most part urban, middle-class individuals who adopted antiestablishment positions and called for an end to imperialist oppression from the global North. An important characteristic of these early groups was the uneasy relationship they had with the institutional Latin American left given its traditional understandings of sexuality. Many of the first activists were not welcome by established leftist movements and some were actually expelled from Communist parties, which had official positions declaring homosexuality to be a bourgeois practice. In spite of their isolation, Latin American gay activism has since then taken a decidedly leftist position on politics. The same was not true for relationships that many activists formed with other social movements. A salient aspect of early activism in Latin America was the close relationships that activists formed with women's groups early on. Gay and lesbian activists were heavily influenced by second-wave feminism and forged close relationships with women's organizations. Steeped in feminist debates, some activists included a broader perspective in their early struggles and argued against patriarchal oppression and for gender equality. Given that the important role students and intellectuals played in early activism, activists were steeped in the theoretical discussions regarding homosexuality that gathered pace at this time in advanced democracies.

In countries in which gay activists organized and mobilized, they managed to ignite public debates regarding sexuality in the early 1970s. Through the publication and distribution of

bulletins, flyers, and magazines, and in some cases television interviews, they were able to begin national discussions regarding the ethics and medicalization of homosexuality. However, it is important to note that gay mobilization at this time was limited to the capital cities of some of the largest countries (Argentina and Mexico). In most of the region it would take years, and in most cases decades, for gays and lesbians to decide to mobilize.

Identity Formation

The evolution of gay and lesbian activism after the first spurts of mobilization varied across countries and it was largely shaped by larger political processes. In some countries, such as Argentina and Chile, the mid-1970s saw the overthrow of civilian governments by military juntas, which abruptly ended any possibility for mobilization. Activism therefore came to a sudden end as the military dictatorships undertook campaigns to 'cleanse' societies from subversives, which included homosexuals. The level of repression varied from country to country. In some cases, the military rulers launched explicit campaigns to exterminate homosexuals. As part of preparatory activities for the Soccer World Cup, which was held in Argentina in 1978, the military regime established a moral brigade, part of the federal police, with the objective of 'cleansing the streets' of homosexuals so as to 'not perturb decent people' (Rapisardi and Modarelli, 2001). Some accounts point the killing of up to 400 homosexuals during Argentina's phase of state terror (1976–83) (Jáuregui, 1987). In cases such as Chile, military rulers did not establish explicit extermination campaigns, but the regimes overall plans to suppress any form of social and political activity made it impossible for gays and lesbians to mobilize. Many homosexuals were forced to into exile. In countries, in which military repression was harshest, it would take until the end of military rule for gays and lesbians to start organizing again.

In cases in which authoritarian rule was not as repressive, such as Brazil and Mexico, activism strengthened despite unfavorable circumstances and, by the end of the 1970s, it 'came out of the closet.' In both countries, 1978 marked a turning point. Following the intensification of protests against the authoritarian regime, which saw strong student mobilization and workers' strikes in the automobile sector in São Paulo in 1977, the two first Brazilian organizations were formed: *Lampão da Esquina* (corner lantern) in São Paolo (as an alternative newspaper) and *Somos* (we are) in Rio de Janeiro. Two years later, women split from these organizations and founded the Autonomous Lesbian Feminist Group. In Mexico, the Homosexual Liberation Front operated mostly underground since its formation in 1971. However, 1978 saw the formation of three organizations, which took to the streets: the Revolutionary Homosexual Front (*Frente Homosexual de Acción Revolucionaria*, FHAR), Lambda (*Grupo Lambda de Liberación Homosexual*), and Okiabeth. On 26 July 1978, members of FHAR participated in a march commemorating the Cuban Revolution and on 2 October 1978, members of the other two groups joined another demonstration to commemorate the tenth anniversary of the 1968 government repression, which killed hundreds of students. These were the first in a series of actions that would allow these incipient movements to attain important visibility in subsequent years. In democratic Colombia, some gay organizations were also established in the 1970s. León Zuleta and Manuel Velandia formed the organization *Movement for Gay Liberation* in Bogotá. The majority of these groups asserted a liberationist discourse and adopted common slogans, such as "there is no political freedom without sexual freedom" and "the personal is political." The main objectives of these organizations were to attract attention, generate visibility, and demonstrate that homosexuality was not a medical condition. In many instances, their calls were addressed to police forces demanding the end of harassment and repression. In 1979, Mexican activists held the first gay march in Latin America, which was held in Mexico City.

Similar to what occurred elsewhere at the time, divisions among activists set in soon after the formation of many of these organizations. A wedge issue was the extent to which feminist vindications were espoused. For some groups, such as Mexico's Lambda, sexual liberation also involved a fight for gender equality and its membership articulated a vision steeped in feminist arguments. Many of its members were in fact women. Others, such as FHAR, rejected feminism and argued that, because the problémtaique of gay men was distinct from women's, their struggles could not be fought united. Another divisive issue regarded the relationship that activists thought had to be established with the state. Given the origins of activism, radical and anarchist currents permeated some sectors, which refused to work with state institutions. These groups called for more fundamental societal changes. Others, often referred as the 'reformists,' took a completely different approach and decided to work closely with political parties. In Mexico and Brazil, these reformists decided to enter the electoral field and worked together with leftist parties, which fielded the first openly gay candidates for elections in 1982. While in Mexico, reformists coalesced around one political party (the Revolutionary Workers Party) for the congressional and presidential elections held that year, in Brazil they approached candidate and parties across the political spectrum for gubernatorial and legislative elections (De la Dehesa, 2010).

It is during this time that Latin American gay and lesbian activism began to adopt characteristics of gay movements elsewhere and assumed a frame based on identity politics. Because the leaders of most of these organizations were university students and intellectuals, they were influenced by global debates of the time. The writings of philosophers such as Michel Foucault, whose *History of Sexuality* was translated into Spanish in the early 1980s, were fundamental in their understandings of sexuality and the elaboration of discourse that challenged the prevailing ethics of sexuality. The term homosexual is thus replaced with a postliberationist term 'gay' by some activists. Foucault's work, among that of others, helped many activists to deconstruct contemporary ideas of sexuality and to articulate arguments that attempted to demonstrate that same-sex relations were natural forms of sexual expression and bases upon which to build legitimate identities. In addition to an identity-based discourse, Latin American gay activism began thus to adopt traits and symbols of their northern counterparts, such as notion of being 'proud' of one's sexuality and the rainbow flag. For these 'assimilationist' groups, the struggle was

not capitalism but rather to improve people's perception of homosexuality and to fight stigmatization.

The adoption of a 'gay identity' by Latin American activism in the mid-1980s was not generalized, however. For some of the more radical sectors, the assumption of an identity politics frame amounted to a capitulation to the capitalist system and an abdication of the commitment to push for wider social change. It also remained limited to the larger urban sectors. While some gay organizations sprung in other cities, such as Guadalajara in Mexico and Rosario in Argentina, activism continued to be concentrated in the main cities and did not reach much beyond. In effect, some anthropological work points to the failure of these frames to reach homosexuals outside the big metropolises. For example, Guillermo Nuñez Noriega's work shows that, in northern Mexico, men are able to distinguish between their sexual attractions and practices and their daily lives and men that have sex with men do not always see themselves as homosexual, nor do they attempt to build a lifestyle around their sexual practices. That is to say, their sexual preferences are not the main trait around which their lives are built. In certain places, then, the Latin American historical tradition of separating the public expressions from private practices appears not to have adopted a gay identity frame. The formation of gay identities also emerged differently than in some parts of the global North: whereas the 1980s saw the establishment of gay neighborhoods in many of the larger cities in the United States, this did not happen in Latin America. Commercial establishments dedicated to allow gay men and lesbians socialization spaces sprouted during the 1980s and 1990s, and numerous bars, restaurants, and clubs opened, even in authoritarian Chile. But gay ghettos were not established.

The Influence of HIV/AIDS

The onset of the HIV/AIDS crisis in the mid-1980s had important effects on gay and lesbian activism in Latin America. Upon the report of the first cases, a moral panic emerged and conservative sectors, led by the Catholic Church leadership, mounted accusatory campaigns blaming the unnatural behavior of homosexuals for the spread of the 'gay cancer' or 'pink plague' (Díez, 2011). These campaigns had a destabilizing effect on some activists as, without much knowledge about the virus, they were forced to turn inward to look for answers. In several countries, the spread of the pandemic had a negative effect on the leadership of many organizations as numerous male leaders succumbed to the disease. In Mexico, for example, activism lost strength and lost public visibility.

In several countries, the spread of the disease significantly shaped activism and changed the relationship it had maintained with the state in important ways, marking a new phase in its evolution. While, prior to the onset of crisis activists had established a relatively confrontational approach with the governments demanding the end of police harassment and repression, they turned to a more collaborative relationship with governments. As some scholars have pointed out, the HIV/AIDS crisis in Latin America was used by many activists as an opportunity to establish a more institutionalized relationship with the state (Pecheny, 2003). Never before had questions regarding homosexuality been as visible as it became an issue of public health. Numerous organizations almost entirely devoted to the running of prevention programs were therefore established by activists, organizations that began to receive significant financial resources from governments and international organizations, such as UN AIDS. In cases such as Brazil, cooperation in the fight against HIV/AIDS established a new model of interaction whereby nongovernmental organizations became allies of the state in the formulation and implementation of prevention programs. Groups such as *Grupo Gay da Bahia* and *Triângulo Rosa* assumed significant visibility as their resources and influence increased. In some countries, the new collaborative relationships contributed to the formation of important national and international networks among government officials, practitioners from international organizations, and activists that have lasted for decades and which have been fundamental in the expansion of health care services to HIV-positive individuals (Torres-Ruiz, 2011).

Another important change regards the articulation of a rights-based discourse. In several countries of the region, HIV/AIDS allowed many activists to begin to frame their demands as an issue of rights. As the diseased ravaged on, (mostly) gay men began to mobilize around the concept of the right to health care and to transform a victimization discourse into one that placed the responsibility for the containment of the disease on the state.

The ability of gay activists to induce policy response varied across the region. In cases such as Argentina, activists managed to force the government to pass legislation fairly early on (1991) to fight the epidemic, which became central for the distribution of antiretroviral drugs when they became available in the mid-1990s. In Brazil and Mexico, legislative change in the mid-1990s made the drugs universally available to those infected. In other cases, such as Chile, it would not be until 2001 that the government would enact legislation, but even then it did not adopt a universality principle (Contardo, 2011).

Globalization and Economic Liberalism

Latin American gay and lesbian activism assumed unprecedented visibility and vigor by the end of the 1990s. As many of the region's democracies began to consolidate, and as new opportunities for mobilization arose, innumerable organizations began to proliferate throughout the region and, for the first time, activism in the smaller, less urbanized, countries became visible. By 2000, all countries in the region, including authoritarian Cuba, had some form of gay and lesbian mobilization.

A salient characteristic of the latest phase of this new activism was the adoption of a global discourse around the idea of 'sexual diversity.' Whereas demands had previously focused on the end to repression and access to health care, activism began to adopt the global discourse of equal citizenship rights for a variety of nonconforming sexual identities and expressions. Activism therefore expanded to include numerous groups, which had been invisible, such as transgender individuals. Organized transgender activism assumed public attention for the first time in several countries of the region and very active and well-organized associations were formed, such as Argentina's *Asociación por la Identidad Trasvesti Transexual* and

the *Asociación de Trasvestis Transexuales y Transgenérico*. The adoption of a new discourse was partly influenced by the splash queer theory made within larger theoretical debates around sexuality. In many countries of the region, some university gender programs expanded their curricula to include queer studies and, by the late 1990s, numerous activists and academics began to deploy the discursive tools the new theoretical approach offered.

It was also influenced by global debates on multiculturalism. During the 1990s, discussions regarding the protection of cultural minorities gathered pace and countries around the world enacted a variety of measures to protect ethnic, indigenous, and cultural minorities (Kymlicka, 2007). In Latin America, these international debates coincided with the strengthening of indigenous mobilization, which sought to renegotiate contemporary understandings of national identity and, which demanded the recognition of the multiethnic character of these societies (Yashar, 2005). The diversity called for by queer theorists therefore found a receptive audience in Latin America. Queer theory took hold among many activists in the region and, given the region's rather fluid understandings of sexuality, it resonated deeply in countries that are highly culturally diverse, such as Mexico (Monsiváis, 2004). The categorization of identities based on separate (medical-legal) sexualities lost force. Hybridity continued to characterize Latin American gay and lesbian activism.

Economic liberalization has also had an impact on activism. The region's move away from authoritarian rule coincided with the adoption of market-friendly economic programs, which liberalized markets and reduced the role of states. As market forces were allowed a freer rein, activism witnessed a certain degree of commercialization. In the main capitals of the region, as unprecedented freedoms began to be enjoyed, a new generation of gays and lesbians began to emulate consumerist lifestyles marketed in advanced democracies, especially the United States. Debates among activists on the marketization of gay identities ensued. Strong criticisms, introduced to activism by intellectuals, such as Néstor Perlongher in Argentina, Herbert Daniel and Leila Miccolis in Brazil, and Carlos Monsiváis in Mexico, were formulated about the emergence of a new commercialized identity in some quarters, which was characterized by white, middle-class men. In a part of the world with the highest inequalities between rich and poor, the new class-based identities were seen as highly exclusionary for it was unattainable for the majority of the population.

State-Directed Activism

Since the early 2000s, certain currents within gay and lesbian activism in Latin America have focused most of the attention to the push for policy reform and have, therefore, states as their sites of their struggles. Because, unlike other regions, same-sex relations were not criminalized in most Latin American countries, activism has not had to fight for the decriminalization of sodomy. As police repression and harassment diminished many activists instead turned their attention to the push for legislation to protect sexual minorities from discrimination. More recently, they have also begun to demand the recognition of same-sex relations and the expansion of benefits.

Success in the attainment of policy reform had varied widely across the region. Antidiscrimination on sexual orientation legislation has been introduced in several jurisdictions in the region and in some cases, such as Mexico, enshrined into the constitutions. In regard to same-sex relationship recognition, civil unions are recognized in numerous countries, such as Brazil, Uruguay, Argentina, Mexico, and Colombia, while gay marriage was enacted in Argentina in 2010 and is accessible to all gays and lesbians in Mexico who contract marriage in Mexico City, where it was also legalized in 2010. However, in other countries, activists have not been as successful and the push for reform continues. The expansion of rights has also included other sexual and gender minorities (Encarnación, 2011). In several countries (Mexico, Argentina, Brazil, and Uruguay), legislation has been introduced to allow transgender individuals to change their identity legally. In 2012, Argentina approved the most advanced gender identity in the world: it completely demedicalizes transgenderism, making a change of gender a mere bureaucratic procedure.

The factors behind the ability of Latin American gay and lesbian activists to attain such unprecedented policy changes evidently vary across the region. Scholarship on these developments has identified a variety of factors, which have been central, such as: cross-movement and international alliances in pushing for policy reform; access to financial resources, which has allowed them to institutionalize and professionalize their activities; the establishment of extensive networks of allies, inside and outside the state, which have supported their efforts; a decrease in homophobia levels across the region; and the framing of their demands for fuller conceptions of sexual citizenship, which have resonated in countries in which process of democratization continues to unfold (Corrales and Pecheny, 2010; Pierceson et al., 2012; Díez, 2013).

See also: Gay and Lesbian Movements; Sexual Minorities and Sexual Citizenship; Sexual Orientation: Historical and Social Construction.

Bibliography

Brown, S., 2002. 'Con discriminación y represión no hay democracia': the lesbian and gay movement in Argentina. Latin American Perspectives 29, 119–138.

Buckhart, L., 1997. Mexica women on the home front: housework and religion in Aztec Mexico. In: Schroeder, Susan, Wood, Stephanie, Kaskett, Robert (Eds.), Indian Women of Early Mexico. University of Oklahoma Press, Norman, pp. 25–43.

Contardo, Ó., 2011. Raro: Una historia gay de Chile. Planeta, Santiago.

Corrales, J., Pecheny, M., 2010. Introduction: the comparative politics of sexuality in Latin America. In: Corrales, Javier, Pecheny, Mario (Eds.), The Politics of Sexuality in Latin America. The University of Pittsburgh Press, Pittsburgh, PA, pp. 1–30.

De la Dehesa, R., 2010. Queering the Public Space in Mexico and Brazil: Sexual Rights Movements in Emerging Democracies. Duke University Press, University Park, PA.

Díez, J., mayo–agosto 2011. La trayectoria política del movimiento lésbico-gayen México. Estudios Sociológicos 29 (86), 687–712.

Díez, J., February 2013. Explaining policy outcomes: the adoption of same-sex unions in Buenos aires and Mexico City. Comparative Political Studies 46 (2), 1–24.

Encarnación, O., 2011. Latin America's gay rights revolution. Journal of Democracy 22 (2), 104–118.

Jáuregui, C., 1987. La homosexualidad en la Argentina. Terso, Buenos Aires.

Kymlicka, W., 2007. Multicultural Odysseys: Navigating the New International Politics of Diversity. Oxford University Press, Oxford.

Monsiváis, C., 2004. La emergencia de la Diversidad: las comunidades marginales y sus batallas por la visibilidad. Debate Feminista XV (29), 187–205.

Nesvig, M., 2001. The complicated terrain of Latin American homosexuality. Hispanic American Historical Review 81, 3–4.

Núñez Noriega, G., 2000. Sexo entre Varones: Poder y resistencia en el campo sexual. Universidad Nacional Autónoma de México, México.

Pecheny, M., 2003. Sexual orientation, AIDS, and human rights in Argentina: the paradox of social advance amid health crisis. In: Eckstein, Susan, Wickham-Crowley., Timothy P. (Eds.), Struggles for Social Rights in Latin America. Routledge, London and New York, pp. 253–271.

Pecheny, M., 2004. Lógicas de acción colectiva de los movimientos por los derechos sexuales: Un análisis con aires abstractos de experiencias bein concretas. In: Cáceres, Carlos, Frasca, Tim, Pecheny, Mario, Treto Jr., Veriano (Eds.), Ciudadanía sexual en América Latina: Abriendo el debate. Universidad Peruana Cayetano Heredia, Lima, pp. 203–215.

Pierceson, J., Adriana, P.-C., Shawn, S., 2012. Same-Sex Marriage in Latin America. Lexington Books, Lanham, MD.

Rapisardi, F., Modarelli, A., 2001. Fiestas, baños y exilio en la última dictadura. Sudamericana, Buenos Aires.

Sousa, L.M., 1997. In: Schroeder, Wood, Kaskett (Eds.), Women and Crime in Colonial Oaxaca: Evidence of Complementary Gender Roles in Mixtec and Zapotec Societies, pp. 199–216.

Stephen, L., 2002. Sexualities and genders in Zapotec Oaxaca. Latin American Perspectives 29 (2), 41–59.

Torres-Ruiz, A., 2011. HIV/AIDS and sexual minorities in Mexico: a globalized struggle for the protection of human rights. Latin American Research Review 46 (1), 30–53.

Williams, W., 1986. The Spirit and the Flesh: Sexual Diversity in American Indian Culture. Beacon Press, Boston.

Yashar, D., 2005. Contesting Citizenship in Latin America: The Rise of Indigenous Movements and the Postliberal Challenge. Cambridge University Press, Cambridge.

Lesbian, Gay, Bisexual and Transgender Aging

Sue Westwood, University of Keele, Keele, Staffordshire, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Older lesbian, gay, bisexual, and transgenders (LGBTs) share many of the concerns of younger LGBT, exacerbated by aging. They also share many of the concerns of older heterosexual people, magnified through the lens of gender and sexuality. The accumulative effects of discrimination and marginalization can have an impact on well-being among many LGBT in older age, although for some strong social support networks can act as buffers and promote resilience. They are also echoed in health and social care provision for older people, which is often heteronormative at best and prejudiced at worst. Many older LGBT face being unwillingly 'outed' through care needs and/or having to go back into the closet to avoid discrimination. There is a need to address ageism in the LGBT community and sexuality and gender discrimination among age-related services.

Introduction

Until very recently, there has been a dearth of knowledge about older lesbian, gay, bisexual, and transgender (LGBT) people. Studies of lesbian, gay, bisexual (LGB) aging in previous decades, coming primarily from the United States, have tended to privilege the experiences of middle class older gay men and underrepresent the experiences of older lesbians, bisexual people, and all LGB people in the upper age ranges (Grossman, 2008). The study of trans aging is an even more recently emerging field (Cook-Daniels, 2006). The term 'trans' encompasses both transgender, i.e., people who transgress gender norms for a variety of reasons, and transsexual, i.e., people who live their lives as being of a gender different from that which they were assigned at birth. Some may also identify as lesbian, gay, or bisexual, others identifying as heterosexual. With the expanding number and visibility of trans individuals, and in particular older people undergoing gender reassignment, trans identities are becoming more of a later life phenomenon.

More recently, there has been a growing body of literature on LGB and LGBT aging Witten and Eyler (2012) and a growing awareness that LGBT aging both magnifies older-age concerns, and also adds new dimensions to them. LGBT aging needs to be understood in context, including the history of oppression experienced by the most older LGBT people across their lives. There are many countries now where there is full or at least partial legal recognition and protection for LGBT people, and, in turn, their aging issues are more visible and amenable to study. However, there remain many countries in the world where same-sex relationships and gender transgressions are unlawful, even punishable by death. There are also other countries or regions, particularly those where orthodox religion plays a strong role in cultural life, where, even if not illegal, LGBT people are socially marginalized and at risk of prejudice and discrimination (AGE Platform Europe and ILGA Europe, 2012). In these variously unsafe spaces, LGBT existence itself is hard to study, because of its need to remain clandestine, and the study of LBGT aging has hardly got off the ground. As a consequence, the studies and understandings of LGBT aging that will be reported here focus primarily on research from those countries where there is greater visibility, i.e., the United Kingdom, Australia, Canada, and the United States.

LGBT aging also needs to be understood in the context of variable experiences across the life course. 'Older old' LGBT people, i.e., those in their 80s and 90s, even in those countries where it is now reasonably safe to be open to a certain extent, have typically lived most of their lives under legal and social censure, and may have been 'in the closet' (i.e., have hidden their identities) for most of their adult lives. By contrast, 'younger old' LGBT people, i.e., those in their 60s and early 70s, may have been out, or at least partially out, for most of their adult lives, in those countries where it is relatively safer to do so. All have been impacted by discrimination in varying ways, which, in turn, influences well-being in later life and concerns relating to accessing formal health and social care provision for older people.

In later life, older LGBT people share many of the same concerns as older heterosexual people (Stonewall, 2011). Those concerns are, however, magnified through the respective lens of sexuality and gender nonconformity and the cumulative effects of living in an environment of discrimination. Older LGBT people also share many of the concerns of younger LGBT people. These concerns are, in turn, mediated by older age and by ageism within the LGBT community. In fact, the notion of an LGBT 'community' can be misleading, there being comparatively little interaction between older and younger LGBT people and between older lesbians and older gay men (Pugh, 2002). Additionally, older bisexual people often experience issues of visibility and hostility in both heterosexual and LGBT contexts (Dworkin, 2006) and older trans people often experience transphobia particularly among (older) LGBT groups (Auldridge et al., 2010). Consequently, older-age issues are experienced very differently by lesbians, gay men, bi women, and men, and by trans individuals, as well as additionally layered through intersections with gender, class, race/ethnicity, religion, region, and culture.

Aging and Relationships

Older LGBT Relationships

Older LGB are more likely than their heterosexual peers to be single and live alone, to have fractured relations with biological family and to be childless (Stonewall, 2011). This is for

a number of reasons: the stigma and social marginalization, particularly historical, attached to LGB sexualities have resulted in many traditional families not being accepting of their LGB members. For many older gay men, their cohort was decimated by the AIDS crisis of the 1980s, when many gay men died, and so there are actually far fewer older gay men still alive now as a result (Rosenfeld et al., 2012). For older LGB people, social networking was very often more problematic in previous decades when many led clandestine lives, making, forming long-term associations more difficult. Many who did form long-term dyadic partnerships did so in secrecy and isolation and when one partner dies, the surviving partner is often left isolated and alone (Almack et al., 2010). Additionally, the reproductive technologies now available to individuals and same-sex couples to enable them to have children were not available to earlier generations, and at the same time social stigma made fostering and adoption also a difficult, if not impossible, path to parenthood, denying older LGB access to support from children and grandchildren in later life. The exception to this is those individuals who have formed same-sex relationships later in life, having lived the majority of their lives as heterosexuals (Cronin and King, 2013).

On the positive side of adapting to discrimination, many older lesbians and gay men have 'families of choice' (Weeks et al., 2001) networks, i.e., social support networks of very close friends, including exlovers and sometimes-lovers. However, these are often same-generational, meaning they can all need increased support at the same time. Strong social support networks act as buffers to the more harmful effects of accumulated experiences of discrimination across the life course, both reducing risk and enhancing protective factors in older LGBT (Fredriksen-Goldsen et al., 2013a, 2103b). Facilitating formal social support networks for older LGBT people can also serve to enhance resilience (Knocker et al., 2012).

For those older LGBT individuals who are in same-sex relationships, partnership recognition can play a key role in their material, financial, and emotional well-being, as will be explored in the following section. For other older LGBT people, nonconjugal relationships play an important role in later life, and the extent to which they are legally recognized, and whether that recognition is desirable, is explored in the subsequent section.

Partnership Recognition

In those countries where same-sex partnership recognition is available on a par with heterosexual marriage, older LGB couples (including LGB identifying trans individuals) in same-sex relationships are able to access the same legal rights (and responsibilities) as heterosexual couples. In terms of finances, this includes legal recognition in relation to pensions, benefits, taxes, and inheritance rights. In terms of authority, it includes: access to a partner in health and social care facilities; rights to participate in decision making about a partner's health and social care; inclusion (where possible) in a partner's end-of-life plan; and, after a partner has died, full participation in their funeral arrangements. In terms of social networks, it also means being recognized by legal and financial organizations as a key player in a loved one's life, being recognized as the most significant person by health and social care staff, and being recognized as a bereaved spouse after a partner has died, and afforded all the informal and formal social support that a heterosexual widow or widower might receive.

By contrast, in those (many) countries where limited, or no, same-sex partnership recognition is available, few or any of these rights and privileges are enjoyed, and this can have devastating consequences in older age (Knauer, 2009). This can include material and financial losses, such as lack of welfare benefits or pension entitlements and even loss of one's home upon a partner's death (if it is inherited by biological family in cases of intestacy). It can mean being denied access to a partner in hospital or a residential care facility, even not being told where a partner has been placed at times; it can involve not only being excluded from medical decision making and end-of-life-planning, but being prevented from being with a lifelong partner when they are dying. After a death, the bereaved partner who is not recognized in law and/or society can be excluded from their loved one's funeral. It can involve a lack of recognition, support, and social validation for a bereaved partner, leading to 'disenfranchised grief,' which can result in depression and other mental health problems (Almack et al., 2010). Overall, the less legal recognition available to same-sex couples, the more these potential compound problems associated with older age are for older LGBT who are in partnerships.

Recognition of Nonconjugal Relationships

For many older LGBT nonconjugal relationships, 'families of choice' (Weeks et al., 2001) are of central importance. Current law and social policy, even in those countries where there are LGBT rights, privilege the conjugal couple, biological and filial relationships, leaving friendship on the margins of regulatory recognition (Westwood, 2013). While in those countries where there is freedom of testation (i.e., an individual can determine who inherits their estate), LGBT individuals can and do mobilize wills to ensure their families of friends inherit their estates if they wish them to. However, this is not possible in those countries where there is no freedom of testation. Even where it is possible, the economic resources required for adequate legal testation are not equally available to all LGBT's (Heaphy, 2009), and where there are disputes, or in cases of intestacy, the courts are likely to take a heteronormative perspective on the deceased's likely testatory wishes. Moreover, older LGBT people often use powers of attorney to ensure their LGBT networks can represent them, but there is variable recognition and respect for this in differing socio-legal contexts (Westwood, 2013). What this means is that while in some countries families of friends will be included in health and social care contexts, in terms of access, being provided with information, and being included in decision making, in other countries they will not. Worst-case scenario, very close friends may be denied contact with one another even at the point of death, particularly if biological family object (Knauer, 2009).

Because of the voluntaristic nature of friendship, there are mixed views about its recognition and regulation in law. However many people wish to see flexible forms of relationship recognition (Westwood, 2013). These would in particular benefit older LGBT whose significant relationships go beyond the conjugal family form. However, for a significant number of older LGBT people, loneliness and isolation, and having very

few people in their social network, is a more pressing issue. This will be explored in the following section.

Health and Well-Being

Effects of Discrimination

The overall health and well-being of older LGBT people needs to be understood in the context of lifelong exposure to direct and indirect discrimination. Older LGBT people are highly likely to have experienced the cumulative effects of 'minority stress,' i.e., lifelong navigation of minority identities and associated stigma and social marginalization (Kuyper and Fokemma, 2010). This in turn has impacted their physical and mental health both directly, and indirectly, i.e., in terms of lifestyle factors which may predispose them to ill-health (e.g., use of alcohol, obesity, etc.). Additionally, when older LGBT people do need to access services they may be reluctant to do so because of concerns about heteronormativity, homophobia, and biphobia among health and social care providers. This in turn can delay or impair the benefits they might gain from those services (Stonewall, 2011; McNeil et al., 2012).

Physical Health

Very little is as yet known about the specific health needs of older LGBT people. Some extrapolations can be made from health care research relating to younger LGBT people, and there is a small amount of older-age-related research beginning to emerge. In terms of physical health, older lesbians and bisexual women are less likely to engage with health screening for women because of concerns about prejudice and discrimination. This means that lesbians and bisexual women are more likely to be diagnosed later with women's health problems, including breast cancer and cervical cancer, and as a result have less positive outcomes in the treatment of those conditions (Hunt and Fish, 2009). There is no reason to think that older lesbians and bisexual women are any less likely to follow this trend, and additionally some may have lifestyle issues (e.g., more likely to drink alcohol, smoke, be obese) which predispose them to increased health risks (Musingarimi, 2008).

Older gay men are often marginalized from physical health care for men in general and feel that there is an overemphasis on sexuality and health in terms of gay men's health care, to the detriment of other health issues in later life (Clover, 2006). Moreover, much of that sexual health work is geared up toward younger gay men even though high rates of risky sexual behaviors are found among older gay and bisexual men, suggesting the need for more targeted, age appropriate, health education (Klein, 2012). Due to improvements in medical treatment, many more gay and bisexual men are aging with HIV, yet at the same time their aging cohort has been decimated due to the period when the disease was out of control and many of their peers had died. As a result, they are likely to require additional support in terms of adjusting to aging with HIV (Rosenfeld et al., 2012).

There is still much to be learned about trans health issues (Fredriksen-Goldsen et al., 2013a). Transsexual individuals who have undergone gender reassignment may have associated complex health needs (Cook-Daniels, 2006). Long-term hormonal treatments can lead to both health benefits and health risks. A male-to-female trans woman may also be vulnerable to certain health risks associated with a man's body (e.g., prostate cancer). Similarly, a female-to-male trans man may be at slight risk of health problems associated with a woman's body (e.g., breast cancer, even if the breasts have been removed). Prosthetics (e.g., testicular implants, penile prosthetics, breast implants, and artificial vaginas) may wear out and require replacement and some hormonal treatments can clash with treatments for certain medical conditions.

Mental Health and Well-Being

Lesbians and gay men are at higher risk of mental health problems than the general population (Fredriksen-Goldsen et al., 2013b). This particularly applies to those in the upper age ranges and even more so to older gay men (D'Augelli and Grossman, 2001). Key factors are both the cumulative effects of 'minority stress' and increased likelihood of loneliness and isolation (Kuyper and Fokemma, 2010). Loneliness and isolation impact upon emotional and mental well-being, and are, in particular, risk factors for depression. For gay men in particular, very often their social networks have revolved around commercial resources (e.g., club, restaurants, bars, and pubs) commonly available to gay men. These resources are increasingly youth-orientated and not inclusive of older gays, thus older gay men are more likely to lose access to the resource-based social support network; this is less likely to occur among older lesbians since their relationships are more likely to be based on noncommercial resources (Knocker et al., 2012).

With increasing social acceptance in some countries, growing numbers of previously heterosexually married individuals, particularly women, are forming same-sex relationships in later life. Accessing informal LGBT social networks may be a challenge, as they have not built up links via those networks earlier in their lives (Cronin and King, 2013). Thus while on the one hand, because they are more likely to have children and grandchildren, they enjoy greater intergenerational support, on the other, they may feel marginalized from LGBT communities, even those which do have spaces for older people.

Trans older people also report a mixed experience in terms of mental health and well-being (Fredriksen-Goldsen et al., 2013a). Some are socially well-connected and, particularly for those transsexuals who have transitioned in later life, report an increasing sense of well-being associated with increasing congruence with their gender. However, there are also strong indications that among older trans people there are increased risks of long-term depression compared with both the general population and with LGB people, while at the same time an avoidance of psychiatric services because of the conflation of trans identities and lifestyles with psychiatric disorders (Fredriksen-Goldsen et al., 2011). Again, social connection and social support play a key role in older trans individuals' well-being.

Health, Housing, and Social Care

In terms of health and social care provision (e.g., medical care, home care, community care, day care, residential care, mental

health services, dementia care, and end-of-life care) research indicates that these services are based on an assumption that older people are asexual or heterosexual and as such are underprepared to meet the needs of older LGBT people and may hold prejudiced attitudes toward them (Ward et al., 2011). As prejudice based on sexual orientation is most prevalent among the oldest older in the general population (Valentine and McDonald, 2004), older LGBT are particularly vulnerable to exposure to prejudice among fellow services users who comprise that age group. As a result, older LGBT people are concerned about prejudice, discrimination, bullying, and harassment in care contexts, not only from staff but also fellow service users (Stonewall, 2011; AGE Platform Europe and ILGA Europe, 2012).

Many lesbians and gay may have heightened sensitivities due to having been treated when younger as mentally ill, criminal, and/or sinful (River and Ward, 2012). Older LGB people who have lived hidden lives, are worried about being exposed by their own or a loved one's care needs. By contrast, those who have lived more open lives, are concerned about having to hide their sexual identities in later life in order to stay safe (Stonewall, 2011). Many trans people are concerned about transphobia, and for those who cross-dress, the lack of space and opportunity in care homes to do so (Auldridge et al., 2010). Transsexual people are concerned about how carers will react when providing personal care, e.g., if the trans person's body does not align with her or his gender performance (e.g., some trans women still have a penis, need to shave, etc.) (Age UK, 2010).

Many older LGBT people would benefit from access to sheltered housing. However, research suggests it is not geared up to their needs (Auldridge et al., 2012). The Stonewall survey found that three in five older LGB did not feel confident that housing providers would be able to understand and meet their needs (Stonewall, 2011). As a consequence, many want LGBT/lesbian- or gay-specific housing. There are only a small number of such projects at present, in the United States, Spain, and Australia (Carr and Ross, 2013). What is clear is that there should be a menu of housing choices for older LGBT people.

There is an urgent need to improve both housing and care services for older LGBT people. While there are some emerging specialist community support projects for older LGBT (e.g., Opening Doors in the UK, Knocker et al., 2012, and SAGE, in the United States, Kling and Kimmel, 2006) they are few and far between. Such support needs to be far more widespread and easily accessible.

Summary

Older LGBT people have spent most of their lives in an atmosphere of discrimination. For some living in those parts of the world where they live in fear and hiding, aging only serves to exacerbate vulnerability and marginalization that they have long experienced. Even among those living in more liberal parts of the world, the majority of their lives have been informed by an environment of discrimination and its effects on their current physical, mental, and social well-being. Among those who have seen tremendous social change in recent years, and growing acceptance and legal/social inclusion, there is an increasing expectation of equality in service provision, which is not yet met in age-related health housing and social care contexts. It also informs the services, which are available to them in later life, and their reluctance to engage with those services as a result. There is a need for choice in provision, with both LGBT-friendly mainstream services *and* LGBT-specific services on offer. There is also a need to address ageism within and among LGBT individuals and collectives, and to encourage greater reciprocal intergenerational support between younger and older LGBT individuals.

See also: Aging and Health in Old Age; Friendship Experiences of Lesbian, Gay, Bisexual, and Transgender Individuals; Gay and Lesbian Movements; Gender: Gendering of Categories; Heterosexism and Homophobia; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Lesbians: Social and Economic Situation; Medicine and Health for Sexual Minorities; Non-Heterosexual British Asian Men: Social Representations, Identity and Social Relations; Queer Geography; Sexual Minority Families of Choice and Family Networks: We Are Family; Sexual Orientation: Historical and Social Construction; Social Movements: A Social Psychological Perspective; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

AGE Platform Europe, ILGA Europe, 2012. Equality for Older Lesbian, Gay, Bisexual, Trans and Intersex People in Europe. Age Platform Europe and ILGA-Europe, Brussels.

Age UK, 2010. Transgender Aging in Later Life. Age UK, London.

Almack, K., et al., 2010. Exploring the impact of sexual orientation on experiences and concerns about end of life care and on bereavement for lesbian, gay and bisexual elders. Sociology 44 (5), 908–924.

Auldridge, A., et al., 2010. Improving the Lives of Transgender Older Adults: Recommendations for Policy and Practice. Services and Advocacy for GLBT Elders and National Center for Transgender Equality, New York.

Carr, S., Ross, P., 2013. Assessing Current and Future Housing and Support Options for Older LGB People. Joseph Rowntree Foundation, York.

Clover, D., 2006. Overcoming barriers for older gay men in the use of health services: a qualitative study of growing older, sexuality and health. Health Education Journal 65, 41–52.

Cook-Daniels, L., 2006. Trans aging. In: Kimmel, D., et al. (Eds.), Lesbian, Gay, Bisexual and Transgender Aging. Columbia University Press, New York, pp. 20–35.

Cronin, A., King, A., 2013. Only connect? Older lesbian, gay and bisexual (LGB) adults and social capital. Aging and Society 1 (1), 1–22.

D'Augelli, A., Grossman, A., 2001. Disclosure of sexual orientation, victimization and mental health among lesbian, gay and bisexual older adults. Journal of Interpersonal Violence 16 (10), 1008–1027.

Dworkin, S.H., 2006. The aging bisexual: the invisible of the invisible minority. In: Kimmel, D., et al. (Eds.), Lesbian, Gay, Bisexual and Transgender Aging. Columbia University Press, New York, pp. 36–52.

Fredriksen-Goldsen, K., et al., 2011. The Aging and Health Report: Disparities and Resilience Among Lesbian, Gay, Bisexual, and Transgender Older Adults. Sage, New York.

Fredriksen-Goldsen, K.I., et al., 2013a. Physical and mental health of transgender older adults: an at-risk and underserved population. The Gerontologist. http://dx.doi.org/10.1093/geront/gnt021. First published online: 27 March 2013.

Fredriksen-Goldsen, K.I., et al., 2013b. The physical and mental health of lesbian, gay male, and bisexual (LGB) older adults: the role of key health indicators and risk and protective factors. The Gerontologist 53 (4), 664–675.

Grossman, A., 2008. Conducting research among older lesbian, gay and bisexual adults. Journal of Gay and Lesbian Social Services 20, 51–67.

Heaphy, B., 2009. The storied, complex lives of older GLBT adults choice and its limits in older lesbian and gay narratives of relational life. Journal of GLBT Family Studies 5, 119–138.
Hunt, R., Fish, J., 2009. Prescription for Change: Lesbian and Bisexual Women's Health Check 2008. Stonewall, London.
Klein, H., 2012. A comparison of HIV risk practices among unprotected sex-seeking older and younger men who have sex with other men. The Aging Male 15 (3), 124–133.
Kling, E., Kimmel, D., 2006. SAGE: New York City's Pioneer Organization for LGBT Elders. In: Kimmel, D., Rose, T., David, S. (Eds.), Lesbian, Gay. Bisexual and Transgender Aging. Columbia University Press, New York, pp. 265–276.
Knauer, N., 2009. LGBT elder law: toward equity in aging. Harvard Journal of Law and Gender 32, 301–358.
Knocker, S., et al., 2012. Opening doors and opening minds: sharing one Project's experience of successful community engagement. In: Ward, R., et al. (Eds.), Lesbian, Gay, Bisexual and Transgender Aging: Biographical Approaches for Inclusive Care and Support. Jessica Kingsley, London and Philadelphia, pp. 150–264.
Kuyper, L., Fokemma, T., 2010. Loneliness among older lesbian, gay, and bisexual adults: the role of minority stress. Archives of Sexual Behaviour 39, 1171–1180.
McNeil, J., et al., 2012. Trans Mental Health and Emotional Wellbeing Study 2012. The Scottish Transgender Alliance, Edinburgh.
Musingarimi, P., 2008. Health Issues Affecting Older Gay, Lesbian and Bisexual People in the UK: A Policy Brief. The International Longevity Centre - UK (ILCUK), London.
Pugh, S., 2002. The forgotten: a community without a generation. In: Richardson, D., Seidman, S. (Eds.), Handbook of Lesbian and Gay Studies. Sage, London, pp. 161–181.
River, L., Ward, R., 2012. Polari's Life Story: Learning from Work with Older LGBT People. In: Ward, R., Rivers, I., Sutherland, M. (Eds.), Lesbian, Gay, Bisexual and Transgender Ageing: Biographical Approaches for Inclusive Care and Support. Jessica Kingsley, London and Philadelphia, pp. 135–149.
Rosenfeld, D., et al., 2012. Out of the closet and into the trenches: gay male baby boomers, aging, and HIV/AIDS. Gerontologist 52 (2), 255–264.
Stonewall, 2011. Lesbian, Gay Bisexual People in Later Life. Stonewall, London.
Valentine, G., McDonald, I., 2004. Understanding Prejudice: Attitudes towards Minorities. Stonewall, London.
Ward, Richard, et al., 2011. Don't Look Back? Improving Health and Social Care Service Delivery for Older LGB Users. Equality and Human Rights Commission, London.
Weeks, J., et al. (Eds.), 2001. Same Sex Intimacies: Families of Choice and Other Life Experiments. Routledge, London.
Westwood, S., 2013. 'My friends are my family': an argument about the limitations of contemporary law's recognition of relationships in later life. Journal of Social Welfare and Family Law, 1–17 (ahead-of-print).
Witten, T., Eyler, A.E. (Eds.), 2012. Gay, Lesbian, Bisexual and Transgender Aging: Challenges in Research, Practice and Policy. Johns Hopkins University Press, Baltimore, MA.

Relevant Websites

http://www.crag.surrey.ac.uk/esrc-seminar-series-minding-the-knowledge-gaps/ – ESRC seminar series: Older lesbian, gay, bisexual and trans people. 'Minding the knowledge gaps' (UK).
http://stumaddux.com/GEN_SILENT.html – GENSILENT (video documentary about LGBT older people).
http://grai.org.au/ – GRAI – Gay, Lesbian, Bisexual, Trans and Intersex Retirement Association (Australia).
http://www.lgbtagingproject.org/ – LGBT Aging Project (USA).
http://www.lgbtagingcenter.org/ – National Resource Center on LGBT Aging (USA).
http://www.openingdoorslondon.org.uk/ – Opening Doors London (UK).
http://www.sageusa.org/ – Services & Advocacy for GLBT Elders (SAGE) (USA).

Lesbian, Gay, Bisexual and Trans-sexual Butch-Femme Subcultures

Julia A Puckett and Heidi M Levitt, The University of Massachusetts, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The terms butch and femme refer to two gender identities that are rooted in sexual minority women's communities. Butch and femme can be descriptions of gender expressions as well as gender identities. Within this article, the origins of these identities are reviewed, as well as the ways in which they have changed over time. Research on butch and femme identities, relationships, and experiences of discrimination is reviewed as is methodological issues related to this research.

For many decades, popular cultures have stereotyped lesbian women as being masculine in appearance and mannerisms. Lesbians were assumed to be imitating men or desiring to be them and this understanding was used to condemn lesbians as 'not real women.' This charge ignored the fact that there was much more variation in gender expression among lesbians than was suggested by this stereotype. In addition femme women, who were not masculine in their gender presentation, were often not seen as real lesbians. This article describes the construction, history, and variations in butch–femme subcultures within lesbian/bisexual communities, but first there is a need to establish how terminology will be used.

When discussing the meaning of gender within the lesbian/bisexual community, a few definitions are helpful as many of the key terms are commonly conflated in writing, research, and everyday language. *Sex* refers to whether someone is male, female, or intersex biologically, whereas *gender* refers to socially constructed clusters of traits that tend to be associated with differences in sex, such as masculinity and femininity. In addition, *gender expression* describes the sets of symbols and signifiers (e.g., symbols of masculinity such as a suit or short-hair) that come to be associated with genders in a given culture (e.g., Butler, 1990). Butch and femme are two gender identities within lesbian/bisexual communities. Generally speaking, butch is used to refer to sexual minority women who are more masculine in their gender expression and femme is used to refer to sexual minority women who are more feminine in their gender expression. The meanings of these identities, however, may shift when considering different time periods, locations, and communities at hand as described here.

History and the Evolution of Butch–Femme Communities

Within the gay, lesbian, bisexual, and transgender communities, different subcultures and identities have evolved at different times in relation to political and cultural social changes (e.g., Faderman, 1991; Levitt, 2006). Butch and femme gender identities came to be claimed by some lesbian and bisexual women starting in the 1930s and became more common in the 1940s (Kennedy and Davis, 1993) as societal shifts in women's gender expression were taking place. At the beginning of this period, there were more restrictive gender roles for all women than there are today and it was uncommon for women to violate traditional gender norms for clothing, such as by wearing trousers (Kennedy and Davis, 1993). While sexual minority women were often not visible in these earlier times, the 1940s gave rise to a number of establishments and bars that catered to working-class lesbian and bisexual women, and so a sense of community and connections between sexual minority women developed, which contributed to the growth of butch and femme communities during that time (Burch, 1998; Kennedy and Davis, 1993). Through these communities, butch and femme gender identities gained in recognition at the same time as images that promoted lesbian visibility were popularized. Though there was increasing lesbian visibility, the social context of lesbians and bisexual women during this period dictated to a large extent the way that they socialized. For instance, private house parties were more common at the time for upper-class lesbians and bisexual women and in the Black lesbian/bisexual community.

The period during and after World War II was marked by greater independence for women and provided them with more freedom in their gender expression (Kennedy and Davis, 1993). For example, it became more socially acceptable for women to wear pants instead of skirts or dresses. Although butch women typically still had to wear more feminine clothing for work because of social norms, they dressed as butch when out at lesbian bars. Butch women emphasized wearing clothing such as suits, ties, vests, and starched shirts, whereas femme women emphasized wearing fashionable clothing, makeup, and heels. In addition, while cutting one's hair short has often been a significant marker in the representation of butch women, in the 1940s there was greater fear that this would result in being fired from one's job which meant that few women were able to cut their hair short (Kennedy and Davis, 1993).

Just as mainstream gender expressions shift across generations, butch and femme women's gender expressions have responded to the cultural images of their time. For example, these women adopted masculine and feminine gender expressions that were similar to then-popular media figures (e.g., James Dean, Marilyn Monroe). Their gender identities were signified in their appearance, mannerisms, and behavior, including the types of clothing they wore, their haircuts, and their interpersonal dynamics, which included their sexual relationship roles (Kennedy and Davis, 1993; Kraus, 1996). Throughout the 1950s, butch women increasingly dressed in clothing similar to working-class male garb (e.g., sweaters, T-shirts, suits; Kennedy and Davis, 1993). This appearance

became a hallmark of sexual minority status for butch women and came at the cost of reduced job security, which was an issue femme women did not face as often (e.g., Nestle, 1992).

These gendered identities and associated gender expressions created a new subculture of lesbians and bisexual women who were organized to resist the heterosexism in the mainstream heterosexual culture (Faderman, 1991; Kennedy and Davis, 1993). For butch women, their gender expression was read by others within the broader community as signifying their identity as lesbian, while also challenging the gender norms often assigned to females (Kraus, 1996). Femme women's identities as lesbians often became visible when in the company of butch women and their decisions to partner with butch lesbians challenged the idea that lesbians had to be masculine (Nestle, 1992; Kraus, 1996). The public visibility of these lesbian and bisexual women led to dangers such as police harassment and brutality (Kennedy and Davis, 1993), and to having their names listed in newspapers when lesbian bars were subjected to police raids – which could lead to public shame and social rejection (Faderman, 1991). Within the community, butch and femme women often supported one another in resisting this marginalization and would both verbally and physically defend one another (Kennedy and Davis, 1993).

Although not all sexual minority women adopted a butch or femme identity, most did because these were the popular gender aesthetics at the time and so social and dating dynamics were organized in relation to these identities. These dynamics broadened the gender expressions permissible for women and also acted to create public visibility for sexual minority women. Although this pairing of masculine and feminine genders might resemble heterosexuality at a glance, historians have highlighted that butch–femme dynamics challenged the more traditional images of women and feminine gender that were typically marked by passivity (Kennedy and Davis, 1993; Kraus, 1996).

There were many ways in which these gender identities differed from heterosexual genders and the types of privilege they entailed were markedly different. Butch women often had difficulty in gaining or maintaining employment, because their gender expressions readily identified them as lesbian, so often their femme partners were the more financially successful member of a couple (Kennedy and Davis, 1993; Nestle, 1992). Also, butch women were not raised with male privilege and so the adoption of signs of masculinity had a different meaning than for men, for whom social power was often natural and unconscious. For instance, if femme women allowed their butch partners to pay for meals (even when they gave them the money first), it was a decision they made to affirm their partners' butch gender within a profoundly heterosexist context that subjected butch lesbians to great discrimination, rather than a passive acceptance of benevolent sexism. To provide another example of how these gender identities had distinct features, sex for butch women was focused on satisfying femme women and there was social pressure for butch women to be 'stone' – meaning that butch women would gain physical pleasure themselves only via tribadism and refrain from nudity or other types of sexual contact (Kraus, 1996). This sexual interaction allowed both partners to be pleased while affirming their internal sense of their gender (Levitt and Hiestand, 2005).

Just as in many contexts with gender norms, in past times it was easier to be accepted and form relationships if lesbians and bisexual women were read within butch or femme gender identities, thus some women felt marginalized if they did not fit neatly into the categories of butch and femme (O'Sullivan, 1999). While women that did not identify as butch or femme felt marginalized at times, femme women have also reported feeling marginalized within lesbian spaces although this may have varied across racial and ethnic communities. For example, while butch women often were positioned as leaders within White lesbian communities, femme women played a more central role within the Black lesbian community (Kennedy and Davis, 1993).

According to Burch (1998), around the 1960s, androgynous gender expression became more common among lesbians because this expression was seen as more egalitarian within the second wave of feminism that promoted equality by challenging gender differences. Lesbian feminists came to see butch–femme relationships as imitations of heterosexuality. Butch women were thought to be seeking male privilege and femme women were thought to be supporting the sexist objectification and treatment of women. Although portions of the butch and femme communities vocally took issue with this interpretation of their gender expressions as patriarchal (Faderman, 1992), as this thinking came to dominate many lesbian communities in the 1970s, butch and femme women were seen as suspect and many butch/femme women either changed their gender expressions or went underground. These changes, however, were not equal across communities, butch and femme identities remained more common within racial and ethnic minority and working-class sexual minority communities (Crawley, 2001; Faderman, 1992; Moore, 2006).

Butch and femme gender identities began to be reclaimed in the 1980s and 1990s, but there were new understandings of these gender identities. Although some continued to think that these expressions were not egalitarian and degraded women, others believed that lesbians should embrace and explore their genders and sexuality (Burch, 1998; Levitt and Hiestand, 2004). After the second wave of feminism, gender itself was thought to be more fluid and, similarly, sexuality and gender expression within the lesbian community became more flexible. As Burch (1998) states, "a woman might be both butch and femme, alternately or in combination" (p. 362). In addition, butch and femme were no longer single identities and instead, there was a pluralism of butchness and femmeness (e.g., stone butch, baby butch, aggressive femme, etc.) that greater diversified the gendered presentations and partnering of sexual minority women (Faderman, 1992). Also, in some communities, such as the Black lesbian community, new gender identities were claimed as well. Instead of using identifiers such as butch, Black lesbian women may be more likely to claim identities as aggressive or studs (Moore, 2006).

Current Understandings of Butch and Femme Identities

Writings about butch and femme identities have flourished over the last two decades, including personal narratives (e.g., Newman, 1995), theoretical essays (e.g., Munt, 1998), and

research studies (both qualitative and quantitative; e.g., Levitt and Hiestand, 2004; Levitt et al., 2012). Current writings on the topic of lesbian genders have taken a variety of approaches, with some seeing butch and femme genders as having a biological basis (e.g., Pearcey et al., 1996; Singh et al., 1999), some describing these genders as rooted in performance (e.g., Butler, 1990), and others citing the intersection of biological and social forces (e.g., Levitt, 2006).

Research on Butch Identity

Qualitative research on contemporary butch and femme identities can be helpful in developing an understanding of these women's experiences of their own gender identities. Findings from a couple of these studies are summarized here to describe some common developmental experiences (Levitt and Hiestand, 2004; Hiestand and Levitt, 2005). In their qualitative analyses of butch lesbian's stories about their gender identities, the researchers found that butch participants viewed their identities as essential to their sense of self-beginning at an early age. These women reported being butch before they had a word for it. They described themselves as having a masculine energy and essence, one that had shared characteristics with men even though they still identified as women. These women reported having a sense of being different from most girls that could result in their feeling isolated at times. Also, many of the women reported that others picked up on this difference and called them names or teased them, especially once junior high school commenced and they no longer passed as tomboys. Often puberty was experienced as a stressful time as their boyish appearances changed and their bodies seemed to be betraying their internal sense of self. Some of the participants reported pressures to conform to a feminine gender expression and to feign interest in dating men.

They reported eventually entering lesbian communities where they found interpersonal acceptance and a space to explore their sexual orientation. Other butch-identified women provided models of acceptance and pride for their gender, and sometimes mentored the participants as they learned the social meanings associated with butch identity. Also, femme women showed them that their butch gender expression was attractive and sexually desirable – which was a transformative idea for some. Community stereotypes about butch gender could be restricting at times, although many of the participants reported embracing both aspects of their gender that were supported by a butch stereotype along with aspects that were not seen as intrinsically butch.

Research on Femme Identity

Similarly, qualitative interviews of femme lesbians can provide an understanding of the types of experiences associated with this gender. Levitt et al. (2003) asked femme interviewees to describe their process of developing their gender identities. The women interviewed reported that they saw their femme gender identity as a core part of themselves. Their development of a femme identity was complicated because of what this identity meant for them when navigating heterosexual and lesbian spaces. Through their youths, most participants reported that they did not match the stereotypical images of lesbians (i.e., butch or androgynous women), which sometimes prevented them from seeing themselves as lesbians until later in life. Also, these stereotypes taught them that real lesbians should be attracted to feminine women, which did not fit for them either since most femme women reported to be attracted to more masculine women. However, when entering into lesbian communities, they were able to embrace their sexual minority identities. Also, they could further the development of a desirable butch aesthetic in others that could overcome the negative societal view of butch or masculine women. Some femme interviewees, however, reported experiencing some marginalization from within lesbian communities, and that their sexual minority identities were sometimes questioned because of their gender expressions.

Also, the participants reported having difficulty being recognized as lesbian by other lesbians and thus often needing to verbally disclose their sexual minority identities. Although they experienced this as challenging at times, they also found that their gender expression could permit them a different type of activism. While butch women created more lesbian visibility, the femme women thought that heterosexist people might become more comfortable with them initially because of their feminine appearance and then might be more receptive to learning about their lesbian identities. These themes from research on femme identity echo personal narratives that have also been written by femme lesbians (e.g., Coyote and Sharman, 2011; Harris and Crocker, 1997; Newman, 1995; Rose and Camilleri, 2002).

Intersectionality and Butch/Femme Identities

While this account centers upon the changing of butch–femme communities in the United States, it should be noted that similar identities and communities have existed in other regions as well. For example, the T-Po community in Taiwan has been described as a butch–femme community that began in the 1960s (Chao, 2000). Similar genders have been developed in places such as China (e.g., Zheng and Zheng, 2013), Israel (e.g., Luzzatto and Gvion, 2004), and Latin America (Esquibel, 2003) – often with unique twists and meanings associated with these genders.

Most research exploring butch and femme identities has focused on the experiences of White sexual minority women. More recently, research has started to examine the experiences of racial and ethnic minority women. For example, within the Black lesbian community, there are a variety of identity labels that are not represented within White lesbian discourse, such as 'stud' and 'aggressive,' which categorize women who may be labeled as butch within White lesbian communities (Wilson, 2009). In addition, Black lesbians may claim labels anywhere along the gender spectrum, such as hard stud (similar to a stone butch within White lesbian culture), or may reject these gender identity labels (Wilson, 2009). While research is starting to examine the intersection of race and lesbian identity, this is also being done within the media. For example, the documentary U People (Demetrius and Walidah, 2009) explores gender identities within a Black lesbian community in New York.

Turning to sexual orientation itself, preliminary research in a large national survey of sexual minority women addressed the intersection of sexual orientation and butch–femme genders.

In that research, femme women were much more likely than butch women to identify as bisexual, while butch women were found to be almost exclusively lesbian identified (Levitt and Bridges, 2007). Additionally, butch women appeared to be more attracted to feminine women while femme-identified women reported a stronger preference for their partners to be more masculine.

Historically, social class has been related to butch–femme identities within lesbian communities, with these two identities more common among working-class women (Kennedy and Davis, 1993). In addition, butch and femme cultures originated in bars that catered to working-class lesbians (Faderman, 1991). More recent research has found that social classes, including income and education, are predictors of whether women claim identities as butch and femme but also have found increased diversity of class status than reported in these earlier generations (Weber, 1996).

Relationships within Butch–Femme Subcultures

Historically, butch and femme women have been expected to partner with each other (e.g., Faderman, 1991). While butch and femme women have been expected to desire one another, some women desire women of the same gender identity as themselves (e.g., butch–butch relationships) and not all women within the sexual minority community identify as butch or femme (Levitt and Hiestand, 2005). Defining these women's gender based on their partners alone can marginalize women attracted to similarly gendered others (i.e., a butch woman attracted to other butch women; Hemmings, 1999; Inness and Lloyd, 1995).

In a qualitative study, Levitt and Hiestand (2005) examined sexuality within butch and femme lesbians. Through the grounded theory analysis, results revealed that participants' romantic relationships often were structured around valuing their gender differences. Butch women were empowered in their ability to seduce and be desired by femme women. Knowing the social marginalization that butch women encountered, femme women could demonstrate their support for their partners' gender identity by allowing them to take on the interpersonal mannerisms traditionally associated with heterosexual men. Playing with their differences via flirtation served to validate each other's gender identities that were marginalized in different ways by mainstream or lesbian communities while increasing their reciprocal desire. For femme women, being read as lesbian when with butch women was experienced as positive and was in contrast with their sexual minority identities usually being misperceived as heterosexual within heterosexual spaces (Eves, 2004).

While there have been other studies on partner preferences in butch and femme women, most have suffered from methodological limitations. For example, many have relied on using personal ads as data. These studies have shown that women may prefer other women who describe themselves as more feminine, regardless of whether they identify as butch or femme (Bailey et al., 1997; Smith and Stillman, 2000). Even so, there are various issues that impede interpretation of these results, including the potential of having biased samples. As Smith and Stillman point out, butch women are more visible as sexual minorities because of their gender expression, whereas femme women often are read as being heterosexual. This misreading may result in women who desire femmes using personal ads as a way to meet partners because of the difficulty in identifying femme sexual minority women. It also may result in inaccurate interpretations that sexual minority women desire only feminine or femme women. Other studies have forced participants to identify as either butch or femme (e.g., Bassett et al., 2001) and neglected the diversity of gender present in sexual minority communities.

Discrimination and Psychological Research on Butch–Femme Women

Research has shown that individuals who conform less to the expected male and female gender norms are generally viewed more negatively by others and are less accepted (Horn, 2007). Along with being viewed more negatively, butch women are more likely to encounter heterosexist events, such as discrimination and harassment, than femme-identified women (Lehavot and Simoni, 2011; Levitt et al., 2012). Even though sexual minority women with more masculine gender expressions or those who have butch identities experience more heterosexism than feminine or femme women, some research has found that butch women do not have worse mental health outcomes (Levitt et al., 2012). This resiliency may be the result of butch women's stronger connections to the lesbian community and their being more out, which might buffer the negative impact of heterosexism on mental health (Levitt et al., 2012).

In contrast, some research shows that there are differences between mental health for butch and femme women, but that the higher rates of experiencing heterosexism explain these relationships (Lehavot and Simoni, 2011; Rosario et al., 2008). While femme women may encounter less heterosexist events, they may encounter minority stress in other ways, including having higher levels of internalized heterosexism that may result in negative mental health outcomes (Lehavot and Simoni, 2011). More research is needed to understand these different findings on the impact of heterosexism on mental health and well-being in butch- and femme-identified women, as well as the potential buffering impact of claiming these identities and belonging to butch–femme communities.

Methodological Issues Related to Butch–Femme Research

One of the major methodological issues related to research on butch and femme identities is how these identities are ascertained. Some research relies on participants' self-identification, while other research has had participants rate their gender expression on a continuum from butch to femme or from masculine to feminine. For research that uses a continuum, there is also diversity in the ways this is approached, with some assessing gender on one scale with masculinity and femininity on separate ends (e.g., Levitt et al.,

2012) or separate scales measuring butchness and femmeness (e.g., Singh et al., 1999). More recently, scales are beginning to be developed to provide a multifaceted measurement of gender within sexual minority women instead of relying on single items (e.g., Lehavot et al., 2011).

Another major limitation of previous work is that, in large, it has focused on the experiences of White sexual minority women. This focus neglects the experiences of sexual minority women of color. More research is needed to understand fully the ways ethnic and racial minority women experience their gender expression and identities as butch or femme (as well as other gender identities that may be more common within their ethnic or racial communities). Also, with the emergence of transgender identities, the positioning of butch and femme identities has become more complex – with some writers seeing butch as a transgender identity (e.g., Feinberg, 1993) and others positioning butch and transmen in contrast to one another (e.g., Halberstam, 1998; Hale, 1998).

Conclusion

Within lesbian history, butch and femme identities have played an important role as they were the first lesbian gender identities that became publicly visible. These identities have continued to evolve over time, with the continual changing of gender roles in society. In addition, these communities allowed women to resist heterosexist oppression in the 1940s and 1950s and created more visibility for sexual minority women. The actions and resistance of these early generations left a legacy of community support that has bolstered community organizing in modern lesbian communities.

See also: Feminist Theory: Radical Lesbian; Gay and Lesbian Movements; Gay, Lesbian, and Bisexual Youth; Heterosexism and Homophobia; Lesbian, Gay, Bisexual and Trans-sexual Drag Culture; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; School Environment for LGBTQ/Sexual Minority Youth; Sexual Minority Families of Choice and Family Networks: We Are Family; Sexual Orientation: Historical and Social Construction.

Bibliography

Bailey, J.M., Kim, P.Y., Hills, A., Linsenmeier, A.W., 1997. Butch, femme, or straight acting? Partner preferences of gay men and lesbians. Journal of Personality and Social Psychology 73, 960–973.

Bassett, J., Pearcey, S., Dabbs, J.M., 2001. Jealousy and partner preference among butch and femme lesbians. Psychology, Evolution & Gender 3, 155–185.

Brown, N.R., 2009. "I'm in transition too": sexual identity renegotiation in sexual-minority women's relationships with transsexual men. International Journal of Sexual Health 21, 61–77.

Burch, B., 1998. Lesbian sexuality/female sexuality: searching for sexual subjectivity. Psychoanalytic Review 85, 349–372.

Butler, J., 1990. Gender Trouble: Feminism and the Subversion of Identity. Routledge, New York.

Chao, A., 2000. Global metaphors and local strategies in the construction of Taiwan's lesbian identities. Culture, Health, & Sexuality 2, 377–390.

Coyote, I.E., Sharman, Z. (Eds.), 2011. Persistence: All Ways Butch and Femme. Arsenal Pulp Press, Vancouver, Canada.

Crawley, S.L., 2001. Are butch and fem working-class and antifeminist? Gender & Society 15, 175–196.

Demetrius, O., Walidah, H., (Producers), 2009. U People [Film], New York City.

Esquibel, C.R., 2003. Shameless histories: Chicana lesbian fictions talking race/talking sex. In: Torres, L., Pertusa, I. (Eds.), Tortilleras: Hispanic and U.S. Latina Lesbian Expression. Temple University Press, Philadelphia, PA, pp. 258–276.

Eves, A., 2004. Queer theory, butch/femme identities and lesbian space. Sexuality 7, 480–496.

Faderman, L., 1991. Odd Girls and Twilight Lovers: A History of Lesbian Life in Twentieth-Century America. Columbia University Press, New York.

Faderman, L., 1992. The return of butch and femme: a phenomenon is lesbian sexuality of the 1980s and 1990s. Journal of the History of Sexuality 2, 578–596.

Feinberg, L., 1993. Stone Butch Blues. Firebrand Books, San Francisco, CA.

Feinberg, L., 1996. Transgender Warriors: Making History from Joan of Arc to Dennis Rodman. Beacon Press, Boston, MA.

Halberstam, J., 1998. Transgender butch: butch/FTM border wars and the masculine continuum. GLQ 4, 287–310.

Hale, C.J., 1998. Consuming the living, des(re)membering the dead in the butch/Ftm borderlands. GLQ: A Journal of Lesbian and Gay Studies 4, 311–348.

Harris, L., Crocker, E. (Eds.), 1997. Femme: Feminists Lesbians & Bad Girls. Routledge, New York.

Hemmings, C., 1999. Out of sight, out of mind? Theorizing femme narrative. Sexualities 2, 451–464.

Hiestand, K.R., Levitt, H.M., 2005. Butch identity development: the formation of an authentic gender. Feminism & Psychology 15, 61–85.

Horn, S.S., 2007. Adolescents' acceptance of same-sex peers based on sexual orientation and gender expression. Journal of Youth and Adolescence 36, 363–371.

Inness, S.A., Lloyd, M., 1995. "G.I. Joes in Barbie land": recontextualizing butch in twentieth-century lesbian culture. NWSA Journal 7, 1–23.

Kennedy, E.L., Davis, M.D., 1993. Boots of Leather, Slippers of Gold: The History of a Lesbian Community. Penguin, New York.

Krakauer, I.D., 2002. The impact of group membership on lesbians' physical appearance. Journal of Lesbian Studies 6, 31–43.

Kraus, N., 1996. Desire work, performativity, and the structuring of a community: butch/fem relations of the 1940s and 1950s. Frontier 17, 30–56.

Lehavot, K., King, K.M., Simoni, J.M., 2011. Development and validation of a gender expression measure among sexual minority women. Psychology of Women Quarterly 35, 381–400.

Lehavot, K., Simoni, J.M., 2011. The impact of minority stress on mental health and substance abuse among sexual minority women. Journal of Consulting and Clinical Psychology 79, 159–170.

Levitt, H.M., 2006. Butch, femme, bear and leatherman: a programmatic exploration of gender identities within gay and lesbian subcultures. In: Kam-Shing, Y.I.P. (Ed.), Psychology of Gender Identity: An International Perspective. Nova Science, Hauppauge, NY, pp. 105–121.

Levitt, H.M., Bridges, S.K., 2007. Gender expression in bisexual women: therapeutic issues and considerations. In: Firestein, B. (Ed.), Handbook of Counseling and Psychotherapy with Bisexual Clients. Columbia University Press, New York, pp. 297–311.

Levitt, H.M., Gerrish, E.A., Hiestand, K.R., 2003. The misunderstood gender: a model of modern femme identity. Sex Roles 48, 99–113.

Levitt, H.M., Hiestand, K.R., 2004. A quest for authenticity: contemporary butch gender. Sex Roles 50, 605–621.

Levitt, H.M., Hiestand, K.R., 2005. Gender within lesbian sexuality: butch and femme perspectives. Journal of Constructivist Psychology 18, 39–51.

Levitt, H.M., Puckett, J.A., Ippolito, M., Horne, S.G., 2012. Sexual minority women's gender identity and expression: challenges and supports. Journal of Lesbian Studies 16, 153–176.

Luzzatto, D., Gvion, L., 2004. Feminine but not femme: the dual lesbian body. Journal of Homosexuality 48, 43–77.

Moore, M.R., 2006. Lipstick or timberlands? Meanings of gender presentation in Black lesbian communities. Signs: Journal of Women in Culture and Society 32, 113–139.

Munt, S.R. (Ed.), 1998. Butch/Femme: Inside Lesbian Gender. Cassell Wellington House, Herndon, VA.

Nestle, J., 1992. The Persistent Desire: A Femme-Butch Reader. Alyson Publications, Boston, MA.

Newman, L. (Ed.), 1995. The Femme Mystique. Alyson Publication, Inc., Los Angeles, CA.

O'Sullivan, S., 1999. I don't want you anymore: butch/femme disappointments. Sexualities 2, 465–473.

Pearcey, S.M., Docherty, K.J., Dabbs Jr., J.M., 1996. Testosterone and sex role identification in lesbian couples. Physiological Behavior 60, 1033–1035.

Rosario, M., Schrimshaw, E.W., Hunter, J., 2008. Butch/femme differences in substance use and abuse among young lesbian and bisexual women: examination and potential explanations. Substance Use & Misuse 43, 1002–1015.

Rose, C.B., Camilleri, A. (Eds.), 2002. Brazen Femme: Queering Femininity. Arsenal Pulp Press, Vancouver, Canada.

Singh, D., Vidaurri, M., Zambarano, R.J., Dabbs, J.M., 1999. Lesbian erotic role identification: behavioral, morphological, and hormonal correlates. Journal of Personality and Social Psychology 76, 1035–1049.

Smith, C.A., Stillman, S., 2000. Do butch and femme still attract? The Gay & Lesbian Review, 17–18.

Walters, K.L., Evans-Campbell, T., Simoni, J.M., Ronquillo, T., Bhuyan, R., 2006. "My spirit in my heart": identity experiences and challenges among American Indian two-spirit women. Journal of Lesbian Studies 10, 125–149.

Weber, J.C., 1996. Social class as a correlate of gender identity among lesbian women. Sex Roles 35, 271–280.

Wilson, B.D.M., 2009. Black lesbian gender and sexual culture: celebration and resistance. Culture, Health, & Sexuality 11, 297–313.

Zheng, L., Zheng, Y., 2013. Butch-femme identity and empathizing-systemizing cognitive traits in Chinese lesbians and bisexual women. Personality and Individual Differences 54, 951–956.

Lesbian, Gay, Bisexual and Trans-sexual Drag Culture

Shae D Miller and Verta Taylor, University of California, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Drag queens and drag kings have received a great deal of attention for their roles in the lesbian, gay, bisexual, and transgender (LGBT) community and movement. Their sexual and gender identities vary but most are gay men, lesbians, or transgender people who perform identities on the stage, which both depict and defy conventional binary gender and sexual categories. Drag queens and kings are distinct identity categories that have become integral to the contemporary understanding of gay and lesbian politics and communities. Recent scholarship raises questions about the political potential of drag. This body of research suggests, nevertheless, that drag performances are fundamental to the creation and maintenance of LGBT identity and community.

Introduction: Drag Queens, Drag Kings, and Faux Queens

At the most basic level, drag queens and drag kings are men, women, and transgender people who perform femininity, masculinity, or something in between. These categories, however, are not stable and encompass a wide range of bodies and enactments of gender. Typically, drag queens are conceived as male-bodied individuals who perform femininity, while drag kings are thought of as female-bodied people who perform masculinity. But not all men who dress as women or women who dress as men identify as drag queens or drag kings. Other categories include: cross-dressers (formerly known as transvestites, who are generally straight men who wear women's clothing for nonperformance related reasons); butch women (including some lesbians); effeminate men (including some gay men); preoperative transsexuals and transgendered people who display and embrace a gender identity at odds with their assigned sex category. (For more information, see Rupp and Taylor, 2003; Rupp et al., 2010).

There is a strong link between drag and gay identity. Despite the fact that not all drag kings and drag queens identify as gay and lesbian, most of them do. Some drag queens, known as 'tittie queens,' acquire breasts through either hormone treatment or implants but have no desire to live as women in their daily lives. While some drag queens incorporate transgender identities into their biographical narratives, most are male-bodied, self-identified, gay men, whose feminine personas are restricted to the stage. Drag kings are more likely than drag queens to embrace gender nonconformity as part of their sexual identity (Ayoup and Podmore, 2002). Rather than referring to themselves as lesbians, many drag kings account for their gender anomalies by identifying themselves as sexually 'queer' – an umbrella term for the nonheteronormative status. Although some drag kings are transgender men who adopt masculine drag personas onstage, most are female-bodied people who self-identify either as gender queer or as women, as well as either lesbian or sexually queer. Faux queens, also referred to as 'bio queens,' are generally female-bodied, woman-identified individuals, who perform femininity onstage and may be lesbian, queer, or straight.

The Early History of Drag

The meaning and substance of drag shifts drastically across cultural and historical contexts. Predating drag queens were the 'mollies' or 'queens' of the late seventeenth century in England. These individuals comprised a subculture of men who used feminine attire and mannerisms to express their same-sex sexual desires. In the early nineteenth century, however, this subculture was distinguished from the 'dame comedians' – men who performed as old women to elicit laughter. During the mid- to late nineteenth century, the term 'drag' was first used to describe men who wore women's clothing not only to elicit laughter but also to perform glamorous female roles in serious theater productions. This legitimated form of female impersonation was first made possible, in part due to the cultural and historical context in which these men were performing; at the time, women were barred from involvement in the theater and men therefore played both male and female roles on the stage. Use of the term 'drag queen' to describe female impersonators began in the 1930s and was generally linked with the presumed homosexuality of the men who were performing (for more information, see Rupp, 2001).

Drag expanded beyond the theater and began to play a central role in bar cultures in the US throughout the 1920s and 1930s. During the 1920s and early 1930s, Prohibition era speakeasies used drag performances to attract mixed (homosexual and heterosexual) crowds of middle class and working class audiences in many cities and in Harlem, drew mixed audiences of African Americans and 'Whites.' During this time, public drag balls offered homosexual men the opportunity to dress in women's clothing and dance with other men in mixed heterosexual and homosexual crowds. Prohibition era drag cultures expanded during the first few financially tenuous years of the Great Depression, as business owners struggled to solicit paying customers by featuring female impersonators in their clubs. This resulted in a short-lived 'pansy craze' in New York City's Times Square, as patrons flocked to see these shows.

But the pansy craze of the early Depression years was cut short when prohibition was repealed in 1933. The norms of law-abiding citizenry were reinstated and the government established regulatory control over liquor-serving establishments through agencies like the Liquor Control Board in New

York. New regulations included the ban on homosexual behavior in reputable establishments, which forced homosexuals and drag queens out of mainstream bars. Ironically, this exclusion led to the establishment of the first bars catering to homosexuals, which have subsequently become integral spaces for the development and cohesion of same-sex oriented subcultures and communities (for more information, see Chauncey, 1994).

Although generally relegated to same-sex oriented spaces following the 1930s, drag made a sanctioned appearance in the United States and Canadian militaries during World War II. Drag performances by male soldiers were an essential aspect of morale building among men, who served for long periods of time without female companionship. Although there is written history to suggest that some of the drag performers were so convincing that they warranted the romantic attentions of other soldiers, the specter of homosexuality never haunted these performers until women were recruited into the armed forces during the Second World War and took over the role of female entertainers. Although some performers were discharged for psychological reasons in Canada, it is unclear whether these reasons were related to accusations or instances of homosexuality. Moreover, there is no evidence that male performers were sanctioned in any way for their roles as female impersonators and some pursued drag after their enrollment in the armed forces (for more information, see Halladay, 2004).

Following the war, drag continued to flourish in the United States by catering to mostly straight audiences. For example, heterosexual tourists flocked to the reputably salacious city of San Francisco beginning in the late 1930s, to watch drag shows in bars that catered to homosexuals. Still, the postwar environment took its toll on drag performances. Wearing drag on the street was still illegal in many cities and female impersonators had to carry cabaret cards to prove they were performers. The range and expertise of female impersonators shifted again in the mid-twentieth century, when financial issues led to the use of prerecorded music and lip-synching in drag shows, allowing amateur performers to enter the drag scene in greater numbers.

Drag in the LGBT Community

Drag queens, as we currently conceive of them, achieved notoriety in the United States during the 1960s and 1970s. As police continued to harass patrons at gay bars, drag queens became explicitly political, using drag performances to raise public awareness of gay identity and support for gay rights. Although they played a key role in the gay liberation movement throughout the 1970s, drag queens were marginalized within the gay and lesbian community throughout the 1970s and 1980s, because their gender transgressions contradicted the claims to normalcy made by gay men and lesbians. As a result, drag queens and other gender rebels like butch lesbians and radical faeries took a pivotal role in the struggle against gender conformity in the gay liberation movement. This is when the visibility of drag queens as a public and integral force of the gay and lesbian movement increased. However, as transgender visibility increased in the 1980s and 1990s, so too did the stigmatization of drag queens, who were now seen in an even more deviant light (for more information, see McNeal, 1999; Berkowitz et al., 2007; Berkowitz and Belgrave, 2010.)

Even though drag queens have continued to face stigmatization based on gender nonconformity, drag can also be an empowering medium. Drag queens who perform for pleasure are more likely to experience stigma and marginalization than those who perform professionally. As drag queens become more established, they often earn high levels of prestige in the contexts where they perform – primarily gay bars and pride celebrations. However, the desirability of drag queens as entertainers rarely translates into their personal and romantic lives, as gay men often actively reject the idea of dating a drag queen. Thus drag queen performances can be interpreted as reflecting and challenging the "ambivalence of being gay and male in the United States," because they represent a moment where gay male femininity is "not stigmatized and ashamed, but in control and assertive, retaliating against a hegemonic straight world" (McNeal, 1999: p. 347).

Drag offers some performers a medium through which they can explore the nuances of their own sexual and gender identities (for more information, see Shapiro, 2007; Taylor and Rupp, 2004; Patterson, 2002). Drag queens and kings often explain that their drag personas and performances represent exaggerated aspects of their personalities that are hidden in their daily lives. Although drag and transgender identities are not the same thing, some transgender women and men use drag performances to express aspects of their gender identities, particularly in the process of exploring their own gender identities. Both queens and kings have expressed strong connections between drag personas and their personal sexual and gender identities. Some describe their entry into drag as being part and parcel of their own exploration of homosexuality and gender nonconformity. Drag kings in particular often explore their gender identities through performance, although many drag kings are butch women to begin with. As a result, questions of transgender identification are sometimes central to drag performers own identities. More often, however, drag performers, particularly drag queens, tend to deny this connection as a result of the stigma associated with being transgendered. Many drag kings and queens identify with the gender that they were assigned at birth, using drag specifically as a performative technique that is specific to their stage lives. It is common, however, for drag queens to move back and forth between the use of male and female names and pronouns, referring to themselves as 'she' even when they are out of drag.

Because of their roles as entertainers, drag queens and drag kings have opportunities to expand their community connections through performances at gay and lesbian functions. Drag communities can also foster a sense of belonging for individuals. Drag king troupes offer a nurturing environment for masculine women and transgender men, who face hostility and violence in the outside world. In some cases drag communities develop into fully formed family support structures, as is the case with 'ballroom drag.' Ballroom drag, a cultural phenomenon that originated in New York over 50 years ago, encompasses a variety of forms of drag that challenge categories of gender, race, and class (for more information, see Bailey, 2009). Members of the 'ballroom culture' are most often Black and Latino youth, who have been rejected from their communities of origin and subsequently form 'houses' with

one another, through which they sponsor competitive ball events. Each house has a 'mother' (and sometimes a 'father') who takes care of and trains the 'children' in the performance of drag. Drag mothers who teach the tricks of the trade, generally socialize drag queens into their identities. These familial structures provide a nurturing environment for queer people of color, primarily gay men and transgender women.

The Political Potential of Drag

During the 1969 Stonewall riots in New York, which sparked the modern gay liberation movement, drag queens and transgender women were among the most militant protesters against police discrimination and brutality toward gay communities. Earlier in San Francisco, drag queens and transgender women had already been actively engaged in resistance against homophobia in their neighborhoods; but much of this has gone unrecognized by scholars. However, the political role of drag queens and drag kings in gay and lesbian communities has warranted increased scholarly attention, with the rise of queer theory and the growing field of gender studies since the 1990s.

The primary assertion of queer theorists is that drag problematizes the understanding of gender as naturally attached to particular bodies and that the socially constructed nature of gender is revealed through drag (for more information, see Butler, 1990; Rupp and Taylor, 2003). Drag performers can also disrupt heteronormative notions of sexuality. When drag queens perform sexualized femininity for straight audiences, the desire elicited from straight men can destabilize heteronormativity. This is especially true when these same queens disrupt their performances of femininity by calling attention to the male bodies that are present underneath the sequins. This form of drag, known as 'camp,' reveals femininity as a performance.

Although drag has clear potential to disrupt gender and sexual norms, it can also be read as reifying gender stereotypes. Particularly in the context of their retained male privilege, drag queens are interpreted by some people as being misogynistic in their representations of femininity. This is highly prevalent in the case of camp, where drag queens exaggerate femininity to a comedic scale. Since the 1990s, questions have emerged regarding the transgressive potential of drag kings who exemplify the superfluous attachment of masculinity to male bodies. However, unlike drag queen camp, there are limitations to the enactment of parody in drag king performances, due to the understated nature of masculinity in the US culture. It can be difficult to caricature a gender, which is so dominant that it seems culturally invisible. At the same time, the taboo on parodying masculinity can be an incredibly effective form of resistance, since it is socially less acceptable than ridiculing or making fun of femininity.

Despite gendered differences in the ways that drag queens and drag kings perform gender, they both undermine normative distinctions and ideas about gender and sexuality. Drag queens are essentially male-bodied and this affords them some privilege compared with drag kings who, although they are performing masculinity, are usually female-bodied. Drag queens use this authority to engage antagonistically with audience members, particularly straight men. Sometimes they exert superiority over women by claiming that they are better at 'doing femininity' than real women; but drag queens also engage playfully with heterosexual women with whom they have amicable rapport. More often, they attempt to arouse desires in members of the audience that are outside of their expressed sexual identities. Through the arousal of desire across lines of gender and sexuality, drag queens seek to blur and destabilize distinctions between heterosexuality and homosexuality (for more information, see Rupp et al., 2010; Taylor and Rupp, 2006; Rupp and Taylor, 2003).

Drag kings, on the other hand, often challenge sexual and gender norms through the explicit enactment of multigendered desires on the stage. Particularly in drag troupes where the kings emphasize group numbers, the performers interact with one another to execute scenes that contain instances of gay male sexualities, butch/femme sexualities, and butch/butch sexualities. Through such performances, drag kings offer a relatable space for both performers and audience members to explore multiple embodiments of gender and desire, representing the varied and complex sexual identities and dynamics that characterize contemporary lesbian communities.

Scholars have raised questions about the ways that drag performances might also denaturalize race, class, and sexual categories (for more information, see Rhyne, 2004; Muñoz, 1999; Volcano and Halberstam, 1999). Intersectionality is a theory that acknowledges that race, class, sexuality, and gender cannot be understood as a collection of isolated categories, because they intersect to create unique experiences and identities (Crenshaw, 1989; Collins, 2000). When drag queens and drag kings of color incorporate aspects of their racial and ethnic cultures into performances, they make visible their intersectional identities on the stage and challenge 'taken for granted' notions of gender that are informed by white norms. Through these performances, drag artists illuminate the ways that gender and race intersect to create different meanings associated with masculinity and femininity for people of color (For more information, see Muñoz, 1999). White queens also illuminate the ways in which gender norms are racialized to privilege 'whiteness.' This is particularly true in the case of the 'camp aesthetic,' where white queens perform femininity in ways that illuminate the socially constructed nature of proper and constrained, white, middle class, femininity through exaggerated performances of hypersexual, excessive, and trashy white femininity.

Kings tend to more explicitly address racialized performances of masculinity in their discussions of social justice. Still, the masculinities performed by drag kings are often middle class and racialized extremes of masculinity, leaving middle class white masculinities more or less unchallenged. Still, drag kings have the potential to embody an intersectional approach to queer politics on stage. The different performative approaches taken by white kings and kings of color express differences in their investments in queer and racial politics. For example, Dre, a black king who performs black masculinity onstage, does so as homage to alternate forms of black masculinity, challenging racist discourses about hypersexualized and aggressive black men. Black kings and other kings of color are less likely to make an outright mockery of masculinity than white kings. At the same time, when white kings perform white

masculinity, they often do so in ways that exaggerate whiteness and the ways in which it is tied to particular forms of dominant masculinity (for more information, see Crowley and Teach, 2002; Pauliny, 2002; Volcano and Halberstam, 1999).

The politics of drag can also move beyond the stage, to the context of everyday life. For example, Diane Torr's 'Man for a Day Workshops' take drag to the streets, as she trains lesbians, straight women, gay men, and transgender men in the subtle nuances of masculine embodiment (Torr and Bottoms, 2010). Through this process, she unmasks the 'taken-for-granted' power of masculinity and the male-bodied and demonstrates that literally anyone can do it. Drag kings, who must learn to adopt the understated embodiments of men on the stage or who are already masculine or butch, are often mistaken for men when out of drag. Conversely, drag queens are not generally mistaken for women when they are out of costume; and many of them are quite masculine when not in drag. But this does not mean that they do not bring the politics of drag into the streets. In Berlin, *Tunten* are men who dress in drag to challenge normative gender expectations through the exhibition of mundane and trashy femininities both on and off stage. *Tunten* are part of a stigmatized subcultural community that engages in explicitly expressed political activity, contrary to their glamorous drag queen counterparts whose impersonations of celebrities warrant them a great deal of prestige in the contemporary gay community of Berlin, which emphasizes cosmopolitan consumerism and individuality (Balzer, 2004).

Conclusion

Many drag performers view their own performances as explicitly political, while others see their performances as mere entertainment. As we have argued, the political potential of drag is highly contingent upon place and space, as well as on how the audiences receive performances. Despite these caveats, drag can be a politically potent medium when performed for both homosexual and heterosexual audiences. On stage, drag kings and drag queens are contextually normal rather than deviant, allowing them to engage with heterosexual and homosexual audiences and both men and women, from a space that is temporarily privileged rather than marginalized. In short, for both drag queens and kings, the political intentions and outcomes of their performances are highly contextual.

See also: Gay and Lesbian Movements; Gay and Lesbian Tourism; Gay, Lesbian, and Bisexual Youth; Gender: Gendering of Categories; Heterosexism and Homophobia; Lesbian, Gay, Bisexual and Trans-sexual Butch-Femme Subcultures; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Queer Anarchism; Queer Theory and Intersectionality; Sexual Minority Families of Choice and Family Networks: We Are Family; Sexual Orientation: Historical and Social Construction; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Ayoup, C.A., Podmore, J., 2002. Making kings. Journal of Homosexuality 43 (3/4), 51–74.
Bailey, M., 2009. Performance as intervention: ballroom culture and the politics of HIV/AIDS in detroit. Souls: A Critical Journal of Black Politics, Culture and Society 11 (3), 253–274.
Balzer, C., 2004. The beauty and the beast: reflections about the socio-historical and subcultural context of drag queens and "tunten" in Berlin. Journal of Homosexuality 46 (3/4), 55–71.
Berkowitz, D., Belgrave, L.L., 2010. She works hard for the money: drag queens and the management of their contradictory status of celebrity and marginality. Journal of Contemporary Ethnography 39 (2), 159–186.
Berkowitz, D., Belgrave, L.L., Halberstein, R.A., 2007. The interaction of drag queens and gay men in public and private spaces. Journal of Homosexuality 52 (3/4), 11–34.
Butler, J., 1990. Gender Trouble: Feminism and the Subversion of Identity. Routledge, New York.
Chauncey, G., 1994. Gay New York: Gender, Urban Culture, and the Making of the Gay Male World, 1890–1940. BasicBooks, New York.
Collins, P.H., 2000. Black Feminist Politics: Knowledge, Politics and the Politics of Empowerment. Routledge, New York.
Crenshaw, K., 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. University of Chicago Legal Forum, pp. 139–167.
Crowley, V., Teach, D., 2002. Drag kings "down under": an archive and introspective of a few aussie blokes. Journal of Homosexuality 43 (3/4), 285–309.
Halladay, L., 2004. A lovely war: male to female cross-dressing and Canadian military entertainment in World War II. Journal of Homosexuality 46 (3/4), 19–34.
McNeal, K.E., 1999. Behind the make-up: gender ambivalence and the double-bind of gay selfhood in drag performance. Ethos 27, 344–378.
Muñoz, J.E., 1999. Disidentifications: Queers of Color and the Performance of Politics. University of Minnesota Press, Minneapolis.
Patterson, J.L., 2002. Capital drag: kinging in Washington, DC. Journal of Homosexuality 43 (3/4), 99–124.
Pauliny, T., 2002. Erotic arguments and persuasive acts: discourses of desire and the rhetoric of female-to-male drag. Journal of Homosexuality 43, 221–249.
Rhyne, R., 2004. Racializing white drag. Journal of Homosexuality 46 (3/4), 181–194.
Rupp, L.J., 2001. Toward a global history of same-sex sexuality. Journal of the History of Sexuality 10 (2), 287–302.
Rupp, L.J., Taylor, V., 2003. Drag Queens at the 801 Cabaret. University of Chicago Press, Chicago.
Rupp, L.J., Taylor, V., Shapiro, E.I., 2010. Drag kings and drag queens: the difference gender makes. Sexualities 13 (3), 275–294.
Shapiro, E., 2007. Drag Kinging and the transformation of gender identities. Gender and Society 21 (2), 250–271.
Taylor, V., Rupp, L.J., 2004. Chicks with dicks, men in dresses: what it means to be a drag queen. Journal of Homosexuality 46 (3/4), 113–133.
Taylor, V., Rupp, L.J., 2006. Learning from drag queens. Contexts: Understanding People in Their Social Worlds 5 (2), 12–17.
Torr, D., Bottoms, S., 2010. Sex, Drag, and Male Roles: Investigating Gender as Performance. University of Michigan Press, Ann Arbor.
Volcano, D.L., Halberstam, J., 1999. The Drag King Book. Serpent's Tail, London.

Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview

Don Barrett, California State University San Marcos, San Marcos, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides a discussion about the origin and goals of the Lesbian, Gay, Bisexual, and Transgender (LGBT) and sexual minority section of the encyclopedia, the importance of terminology that is inclusive of all of those who are same-sex involved for achieving those goals, and the need for an international perspective on these issues. This article also includes a description of the creation of the section, and discussion of four themes that appear across the articles: social construction of sexual minority categories; patterns of cultural discrimination; discrimination: consequences and responses; and future directions.

Scholarship on the lives of those who are sexually active with same-sex others has increased tremendously since the rapid expansion and visibility of the modern gay movement in the 1970s. For many of those years, the main focus of the scholarship has been on the middle-class white gay male Western populations and communities that were accessible, limited by discrimination, and likely to benefit positively from research. Same-sex sexual activity, however, is not something that occurs only in Western cultures or only among those who, other than their sexuality, would be among the country's politically and socially dominant population. Given the lower availability of scholarship on those same-sex active minorities who are less visible, a driving goal in the creation of this section in the encyclopedia was to provide users with an opportunity to view and compare scholarship on the experiences of the less visible sexual minorities with the scholarship on those who have been most studied.

To reinforce the importance of giving weight to the experiences of the less visible same-sex involved, this overview uses the term 'sexual minority' as a general term and reserves the use of the terms 'gay' or 'LGBT' for specific instances. Though public discourse and much of scholarship uses the terms 'gay' or 'LGBT' to refer to people who are same-sex involved, this article and the articles in this section highlight the extent that those terms hide important differences in both same-sex involvement and in societal responses to such involvement. The articles thus illustrate the value in using approaches that are not restricted by a term laden with multiple and confusing social meanings. By often questioning the assumptions placed on the use of 'gay' or similar common terms, and by drawing attention to experiences internationally, the articles in this section highlight the breadth of sexual minority experiences, the social limitations placed upon them, and their reactions to those limits.

Creation of the LGBT and Sexual Minority Section

While it would have been desirable to have articles from scholars from diverse countries, even the process of finding available authors reflects the influence of Western scholarship on sexual minority research. Probably due to greater acceptance of sexual minority scholarship in Western cultures, quite often even the authors who were studying sexual minorities in non-Western countries were affiliated with a university in a Western country. Due to this situation, the majority of authors are from four English language Western countries (Australia, Canada, the United Kingdom, or the United States) though there are additional authors from China, South Africa, and Turkey. Despite the Western and English-language affiliations of the authors, the articles in the section range farther afield covering sexual minorities in South Africa, the Maghreb region of North Africa, Turkey, China, Latin America, Germany, and Russia, as well as Indian and Pakistani immigrants living in the United Kingdom and Latinos in the United States. Additionally all authors, whenever possible, addressed international and social status (e.g., class, race, ethnicity, indigenous status, age, gender, religion) variations within their topics.

To further capture a variety of perspectives, recruitment of authors also focused on gaining a variety of academic perspectives. Though a majority (65%) of the authors (including coauthors) are sociologists or psychologists, three have interdisciplinary degrees, two are historians, two political scientists, two legal scholars, two geographers, and one each French literature and Russian studies.

A predominance of scholarship on same-sex involved males, however, was unavoidable. Male sexual expression is typically a defining characteristic in the shape of cultures, thus there is more interest in the challenges brought to cultures by male same-sex sexuality resulting in a literature on same-sex sexuality that is dominated by exploration of male–male sexuality. While this section attempts to address that male dominance by encouraging authors to address issues encountered by women and transsexuals, overall the section reflects the dominance of work on males within the literature.

These articles supply a rich and varied contextualization of sexual minorities and, when considered overall, provide global perspectives on four themes (presented below). Encyclopedia users are encouraged to use these themes, as well as other themes they might identify, for an examination of the international commonalities and crosscurrents that occur in the scholarship on sexual minorities, in the lived experiences of sexual minorities, and in the societal and governmental responses to the needs of sexual minorities.

Social Construction of Sexual Minority Categories

Key to understanding the lives of sexual minorities is a very basic question of whether a culture assumes that human sexual behavior should be considered to fall into distinct categories. It is customary in most contemporary Western cultures, and in many other cultures, to have socially constructed distinctions between sexual behavior that is same-sex (homosexual) or opposite-sex (heterosexual), and to assume that the distinction is based on some sort of natural division. However, the scholarship in this section indicates that such divisions are, by no means, based on some sort of natural division.

A well-accepted explanation for the construction of categories is that the cultural division of sexuality into homosexual and heterosexual, and the stigmatization of homosexuality, coincides with the nineteenth century advent of a stage of economic development (particularly industrialization) that results in changes in the functions of families as work units, and in an increase in geographic separation that results in individuals having greater flexibility for exploring and expressing sexuality.

However, not all cultures are at the same stages of development, not all social groups within a culture are affected similarly by development, and other cultural factors may influence how sexual division are constructed. One noted exception to the assumption that constructs of homosexuality and heterosexuality are related to economic development is the situation for women where their typical lower social status results in the treatment of their sex lives as unimportant except to the extent it complements male sexual needs and social reproductive needs; thus the lack of attention given by many cultures to classifying as distinct the same-sex sexuality of women. In a somewhat opposing direction, scholarship also shows that not all same-sex active men are considered to be distinctly different from opposite-sex active men; in some cultures only men who take on stereotypically feminine passive sexual roles are labeled and stigmatized as sexual minorities, men who assume traditionally masculine assertive roles are treated as being no different from men who are primarily heterosexual. Variation is also seen regarding the application of sexual categories to social minority populations (e.g., racial, ethnic, indigenous minorities) where minority populations are often viewed by the dominant culture as having either a hyper- and unfocused sexuality or as being asexual, both of which results in the dominant culture considering sexual categories to be irrelevant for such social minorities. One final, serious, conundrum to consider when questioning the cultural construction of sexual differentiation is presented by increasing research suggesting that biology may, in fact, play a part in influencing whether sexual desires are same-sex or opposite-sex oriented.

Patterns of Cultural Discrimination

As the articles in the LGBT and sexual minority section indicate, despite increasing acceptance of sexual minorities in many (primarily Western) cultures, hostility to, and discrimination against, sexual minorities remains strong in particular regions or population segments in those Western cultures, and is almost universal in some non-Western cultures. The international perspective is particularly helpful here in indicating the mix of reasons used to justify discrimination.

One relatively common justification for discrimination is the assertion that the open expression of sexual orientation (self-identifying as gay) results in the contamination of a nation's sense of identity. This national-identity threat perspective is evident in the differences between political parties in more accepting Western cultures, though the successes of the gay-identified political movement have diluted such effects; the national-identity threat perspective is much stronger in many economically less-developed areas where gay culture is seen as an undesirable import from Western cultures. Religious beliefs are also very often seen as justification for discrimination; such justifications are essentially inseparable from national-identity threat justification in cultures where a religion is part of the national-identity, though religion appears to play less of a separate role in other cultures. Ethnic minority immigrant groups also often use national-identity threat or religious arguments from their country of origin as justifications for discriminating against members of their own groups' adoption of gay identities.

Closely linked to the national-identity threat paradigm is a concern based on the stereotypes of sexual minorities as being unable to perform the various roles that are typically assigned by gender within a culture. This assumption that sexual minorities will dilute gender standards within a culture is usually based on assumptions that sexual minority males will be insufficiently masculine to perform the roles expected of males, including concerns about military service and concerns about producing offspring that will support future generations. Though gender-performance standards for women are less often cited as standards for sexual minority discrimination, close examination of patriarchy in the articles indicates how deeply ingrained female gender-performance standards are with the cultural definitions of masculinity that are used to justify sexual minority discrimination.

Discrimination: Consequences and Responses

As would be expected, the sexual minority scholarship shows that the long histories of discrimination have had a negative impact on the health and well-being of sexual minorities in general, and specifically on that of more vulnerable populations such as youth, seniors, and women. Sexual minorities have organized politically in response to these specific negative effects, as well as for broader demands of an end to stigmatization and freedom of sexual and emotional expression.

The political activities of sexual minorities have resulted in substantial gains in their social acceptance in Western cultures, as well as in vital gains in recognition of their needs in non-Western cultures. As the scholarship indicates, often such gains are associated with the development of actual physical neighborhoods that include businesses, social and political activities, health service providers, and residential areas where sexual minorities can live in relative comfort and feel that their sexuality is validated by others. Accompanying such development of actual communities has also been an increase in the number and types of employers that have specific policies of

acceptance of sexual minorities, as well as the development and implementation of governmental policies that guarantee the rights of full citizenship (including parenting) for sexual minorities. Clearly scholarship such as that used in this section has also benefited from this organizing.

Future Directions

Even with the very successful organizing for the rights of sexual minorities, the benefits of that organizing are not universally accessible. The sexual minority scholarship suggests that the benefits of this political and social organizing are often most apparent in cosmopolitan cities, but even there the benefits are often not accessible to indigenous, ethnic, racial, or working-class sexual minorities.

What, then, to do? Questions are raised among the scholars regarding whether the existing political activities have actually been too limited by focusing on rights within the context of Western, majority-based assumptions about sexuality, identity, power (particularly patriarchal power), and citizenship, resulting in too much of a focus on the ability to be accepted into the standards of the dominant population and too little on the ability to develop alternative approaches to sexuality and self. Scholarship that covers the history of sexual minority politics indicates that this tension between acceptance by a culture that is predominantly and normatively heterosexual, and the ability to express sexual and emotional needs that do not easily fit within those norms, is historically reoccurring.

It is often now argued that we may be entering a 'postgay' world where politics, actions, and lives based on identification with a differentiated, same-sex, 'gay' sexuality, no longer matter. Considering the historical and continuing limits on same-sex sexual expression encountered by significant social groups in the West and even more so outside the West, the end of sexual politics that is suggested by the term 'postgay' appears to be premature – though the labeling of that sexual politics as 'gay' politics may change.

See also: Clinical Psychology, Psychiatry and Homosexuality; Feminist Theory: Radical Lesbian; Friendship Experiences of Lesbian, Gay, Bisexual, and Transgender Individuals; Gay Men's Emancipation Movement in Germany: 1897–1933; Gay and Bisexual Latinos in US Gay Culture; Gay and Lesbian Movements; Gay and Lesbian Tourism; Gay, Lesbian, and Bisexual Youth; Gays, Lesbians, and Sports: No Longer on the Sideline; Heterosexism and Homophobia; Lesbian and Gay Activism in Latin America; Lesbian, Gay, Bisexual and Trans-sexual Butch-Femme Subcultures; Lesbian, Gay, Bisexual and Trans-sexual Drag Culture; Lesbian, Gay, Bisexual and Trans-sexual: Commercial Sex; Lesbian, Gay, Bisexual and Trans-gender Aging; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Lesbians: Social and Economic Situation; Non-Heterosexual British Asian Men: Social Representations, Identity and Social Relations; Queer Anarchism; Queer Anthropology; Queer Geography; Queer Theory and Intersectionality; Same-Sex Sexualities in Turkey; School Environment for LGBTQ/Sexual Minority Youth; Sexual Minorities and Sexual Citizenship; Sexual Minorities in China; Sexual Minorities in Modern Russia; Sexual Minorities in South Africa; Sexual Minorities in the Maghreb; Sexual Minority Families of Choice and Family Networks: We Are Family; Sexual Orientation: Biological Influences; Sexual Orientation: Historical and Social Construction; Transgender, Transexual, and Gender Variant Individuals.

Lesbian, Gay, Bisexual and Trans-sexual: Commercial Sex

Jason H Prior and Penny Crofts, University of Technology Sydney, Broadway, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Forms of commercial sex extend beyond prostitution to include a broad range of businesses that produce sexual goods and services; many of these cater to a broader range of sexual predilections, and include Lesbian, Gay, Bisexual, Transgender, and Intersex (LGBTI) consumers among their clientele. This article provides an overview of the shifting attitudes toward and culture of LGBTI commercial sex, and how the placement and access to LGBTI commercial sex within Western jurisdictions cannot be considered a simple response to patterns of supply and demand, but rather as the outcome of complex interactions of moral codes, legal structures, and other forms of regulations.

Introduction

Sex has long been a commercial concern. Despite the frequent focus on the 'prostitution' form of commercial sex, commercial sex has long existed in diverse forms, and in recent decades, it has increasingly been described using the rubrics of *sex industry* and *adult entertainment*. Forms of commercial sex extend beyond prostitution to include businesses that produce sexual goods and services for sale through bars, restaurants, cabarets, clubs, brothels, discotheques, saunas, massage parlors, sex shops with private booths, hotels, flats, dungeons for bondage and domination, Internet sites, cinemas, and anywhere else that sex is offered for sale ranging in frequency from occasionally through to commercialized sexual festivals. Commercial sex also includes the production of goods and services including erotic phonelines, escort services, films and videos, souvenirs, toys, clothes, equipment, and live and 'virtual' spectacles via web cameras. Although most commercial sex is for heterosexual consumption, many of these venues cater to a broader range of sexual predilections and tastes, including Lesbian, Gay, Bisexual, Transgender and Intersex (LGBTI) consumers among their clientele.

The social actors involved in LGBTI commercial sex, like in the broader sex industry, are diverse including not only those who sell sex directly and their customers but also business owners and investors, independent contractors and nonsexual employees (waiters, cashiers, guards, drivers, accountants, lawyers, and doctors), and facilitators of business processes travel agents, guides, estate agents, matrimonial agents, newspaper and magazine editors, as well as Internet entrepreneurs. Not all of those involved in LGBTI commercial sex are themselves LGBTI; there is a degree of sexual fluidity between those who take part in LGBTI commercial sex and those who participate in the broader sex industry, with 'gay-for-pay' by pornographic stars or sex workers who self-identify as heterosexual and heterosexual prostitution by self-identified lesbians occurring. Additionally, the regulation of LGBTI commercial sex occurs at the intersection of discourses about homosexuality *and* commercialization of sex; thus policy makers and legal agents are also involved.

The Culture of LGBTI Commercial Sex

In recent decades, parallel processes of economic and social mainstreaming have brought LGBTI forms of sex commerce out of the shadows and into the formal economy, with new technologies (e.g., the Internet and mobile telephony, see Zook, 2003) creating diverse opportunities for direct and indirect sexual commerce for LGBTI consumers. We have seen the rise of LGBTI brands, such as the international newspaper *The Advocate*, and of adult content on the Internet and on pay-to-view television, suggesting a significant shift from the margins to the center as all forms of sexuality become more openly discussed (Attwood, 2006). This increased mainstreaming has a clear spatial relationship to the rising visibility of sex premises, in particular for gay men, men who have sex with men, and bisexual men, within many Western cities. Venues such as exotic dance clubs, sex cinemas, sex clubs, and gay and/or lesbian bathhouses are now visible in many city centers and entertainment districts, widening the range of entertainment choices for LGBTI consumers. Likewise, adult retailers and shops are branding products specifically for LGBTI consumers, while in some jurisdictions, specific escort agencies, brothels, and erotic massage parlors are now regarded as legitimate land uses and cater specifically to LGBTI consumers (Prior and Crofts, 2011).

Although LGBTI commercial sex has never been a solely urban phenomenon, in recent decades it has been most visible in specific areas of urban towns and cities; districts that often center on a cluster of commercial establishments such as gay bars and bookstores, restaurants, cabarets, clubs, brothels, discotheques, saunas, massage parlors, sex shops, hotels, flats, and dungeons for bondage and domination; as well as cinemas that cater specially to the needs of LGBTI consumers (Nash, 2006). Sometimes, art exhibitions and theatrical plays appear to temporarily form part of the industry and many of the sites and forms mentioned also promote and sell nonsexual products and services. Some of these precincts, such as Oxford Street, Sydney, Australia, are often associated with larger scale collective events such as lesbian and gay pride festivals. These LGBTI precincts, festivals, and the LGBTI goods and services that they produce are increasingly being incorporated into western cities' place marketing and business strategies. Research suggests that the underlying political and economic processes that created these commercial areas have been driven by largely urban-based, predominantly white, male-dominated gay social and political movements that have taken their own alternative codings of space 'out of the closet' and into the public sphere, but usually within racist, sexist, and procapitalist discourses

(Knopp, 1992; Brown, 2009). Their most obvious impact has been the proliferation of visible (but disproportionately white, male, and middle-class) gay commercial, residential, and leisure spaces. This has resulted in vibrant gay commercial and entertainment scenes as well as the pink economies of cities such as Amsterdam, London, San Francisco, and Sydney.

Conversely, others have noted how lesbians also asserted their rights to commercial space, although sometimes more quietly than gay men. Researchers have suggested that lower incomes and a higher likelihood of child custody shifts lesbian residential location choice toward less expensive areas with more attractive facilities for children. Lesbian clusters have also existed in places like Erskineville, Sydney, and in smaller cities like Northampton, Massachusetts (Forsyth, 1997). Lesbians appear to be particularly sensitive to the dominance of markers of heterosexuality (e.g., advertisements and public announcements referring to heterosexual couples) in urban space (Valentine, 1993) and, unlike gay-male-dominated areas, often do not create visibly marked commercial territory in the built environment. This subtlety has often meant that lesbian neighborhoods are not as visible as overtly gay (male) areas (Podmore, 2001). For example, Rooke (2007) suggested that many lesbian bars exist not as fixed places, but as temporary (i.e., rotating) women's nights in various locations, allowing communities of like-minded women to blend into the built environment.

These apparent differences in approaches to the development of commercial opportunities for lesbians and gay men can be understood as developing on the complex social totality and fundamental frame of reference for Western culture that emerged in the nineteenth century, characterized by dualistic thinking and cross-cutting as well as mutually constitutive sets of social relations including those of race, class, gender, and sexuality, which saw cities characterized by very traditional gender-based spatial divisions of labor dominantly coded as heterosexual and imaged and experienced in terms of public and private spheres of existence (Sedgwick, 1994).

Increasing research has provided insight not only into the way in which LGBTI sex commerce has evolved but also into the way in which such commercial sex has played a role in the formation of LGBTI identities, practices, and behaviors (Brodsky, 1993). Within this context, LGBTI commercial sex is presented as an evolving culture that adapts to significant shifts in LGBTI sex cultures, such as the arrival of acquired immunodeficiency syndrome. Such adaptations have included the reinvention of sexual goods and environments to support modified sexual behavior, such as the development by the proprietors of commercial sex premises of jack off rooms to replace backrooms as a low-tech yet more embodying response to safe sex. Other adaptations include the development of video and printed pornography, phone sex lines, bars featuring erotic dancers and strippers, and computer-networked bulletin boards. Thus LGBTI sex commerce shifts and evolves over time and place, with some forms becoming unpalatable for gay culture, while others emerge as appropriate. Along these lines, the assimilation of gays has resulted in the demise of LGBTI neighborhoods within many Western cities, with studies revealing how some former LGBTI neighborhoods have become less tolerant of LGBTI people and the commercial sex businesses that anchor these LGBTI communities (Doan and Higgins, 2011). Apparently in response to these changes, we are seeing the rise of new forms of LGBTI commercial sex, part of which involves the Internet and mobile phone.

Attitudes to LGBTI Commercial Sex

Attitudes to commercial LGBTI sex reflect a complex confluence of discourses about commercialization of sex and about LGBTI identities and practices that vary across and within Western jurisdictions. These attitudes depend on a range of complex and often conflicting discourses, including religious, psychiatric/psychological, legal, popular, economic, and political ideas and practices, which guide which sexualities are considered good, moral, healthy, normal, and virtuous and which are not. Emerging research has pointed out that these conflicting discourses are not stable and are contingent. Commercial sex often enjoys only a precarious position within many Western jurisdictions, despite the evident demand and supply of sex for sale in Western cities, past and present. Since the 1970s, social research has argued that these discourses have tended to privilege a certain portion of human sexuality, heterosexual, monogamous relationships focused around 'love' and the family (Rubin, 1993). By reducing sex to a commercial transaction rather then one based on 'love' (although the two are not mutually exclusive), commercial sex has historically been regarded as harboring sexual and moral disorder. Here, the influence of religious groups has often been considerable, with the antipornography movement drawing sustenance from fundamentalist worldviews that emphasize the dangers associated with 'an open marketplace of morals' associated with the reduction of sex to a commodity that can be bought and sold (Hayes et al., 2011). Given that LGBTI identities and practices have also often been viewed as sexual and moral disorder, LGBTI commercial sex becomes regarded as doubly deviant – breaching multiple assumed codes of sexuality and love.

In part due to the double deviance, the operation of LGBTI commercial sex beyond the margins of privileged human sexuality has led to opportunities to operate in ways that are outside the purview of regulation, with resulting venues created for specifically LGBTI audiences. At the same time, increases in discourses of legitimacy and rights have provided rationales for access to and location of commercial sexuality, resulting in freedom of choice in terms of LGBTI consumerism. In recent decades, due to the activities of feminist, gay, lesbian, and other movements arguing for legal change, the strength of sex worker advocacy groups, and a liberal authoritarianism that increasingly regards the sex industry and its clients as able to take some responsibility for self-governance, LGBTI commercial sex has begun to move from spaces of transient ownership (e.g., in public parks and wasteland, pseudoindustrial districts, and areas in transition) to more visible and commercial sites (Prior and Crofts, 2011).

These shifting and competing discourses have meant that in recent decades some sexualities at the borders of acceptability within some Western jurisdictions, such as some forms of commercial sex, sexual practices, and identities such as homosexuality, bisexuality, transgender, and intersex, have edged across the line of acceptability and been rewarded with greater levels of freedom (Rubin, 1993). For example, the

emergence of prominent areas of gay and lesbian commerce within some Western cities has been identified as evidence of this shift (Binnie and Skeggs, 2004). However, with greater visibility, there is also greater regulation.

Even though attitude surveys indicate that the public is becoming generally more liberal in its opinion of commercial sex and toward LGBTI within a range of Western jurisdictions (Tibbetts and Blankenship, 1999), the increased commercialization of LGBTI sexuality within the media, sex premises, and cities has not been universally welcomed. Venues and practices that were previously beyond the margins of the law are now situated within legal frameworks and must satisfy legal duties and responsibilities (Prior and Crofts, 2011). For example, sex premises including gay and/or lesbian bathhouses remain capable of exciting considerable opposition (Prior, 2008). When confronted with commercial sex in their own backyard, many residents and business owners display a strong 'Not in My Back Yard' attitude, with residents and businesses demonstrating over their neighborhoods' contamination by commercial sex. Such campaigns of opposition are particularly pronounced when sex premises are proposed in areas with little prior experience of the sex industries, such as suburban and residential areas (Hubbard, 2009). Often, considerable social conflict is associated with the perception that commercial sex is taking over large areas of Western cities. Within these contexts, concerns are often expressed for the relationship between commercial sex and a broader range of potential criminogenic effects, with customers and workers alike accused of criminality ranging from the mundane (public urination and illegal parking) and opportunist (drug dealing and theft) through to predatory acts of sexual violence (international organized crime, illegal immigration, and sex trafficking) (Edwards, 2011). In many cases of opposition, it appears that these anxieties coalesce with the objections of religious groups and antipornography campaigners who regard LGBTI commercial sex as a form of *moral* pollution (Hanna, 2011).

Furthermore, LGBTI communities often express concern for the way in which their sexual practices and identities are being commercialized. As Binnie and Skeggs (2004) highlight, the increased visibility and marketing of gay consumer precincts in recent decades, such as the Manchester's gay village, results in the use of those precincts by nongays. In particular, the consumption of gay culture by straight-identified women has been contested by some gay men who fear a heterosexual invasion and thus a loss, a reheterosexualization, of 'their' space. Similarly, recent pride festivities in Sydney have come under critical scrutiny for their increasing commercial focus and the apparent replacement of politics with hedonism.

The Regulation of LGBTI Commercial Sex

The regulation of LGBTI commercial sex expresses discursive constructions of homosexuality and the commercialization of sex. The location and access of LGBTI commercial sex within Western jurisdictions cannot be considered a simple response to patterns of supply and demand; rather it is the outcome of complex interactions of moral codes, legal structures, and other forms of regulations. Within an increasing number of Western jurisdictions, the private sexual conduct of consenting LGBTI adults is rarely considered in need of regulation. However, the same cannot be said of businesses that profit through the sale of sexual services, goods, and imagery, including but not limited to LGBTI products. Much, but not all, of the regulation of commercial sex has been developed with a focus on heterosexual commercial sex, with an assumption of the need to protect women from exploitation by men. This regulation seeks to police and monitor these sites, goods, and imagery, at the same time as restricting access for those who might be offended by their sights or sounds or to protect those deemed vulnerable to 'corruption.' In most jurisdictions, there has been a clear anxiety to make such spaces off-limits to minors, as well as distance them from specific cultural and religious groups who object to commercial sex on moral grounds.

Regulating the diverse forms of LGBTI services, goods, and imagery presents a number of legal and logistical challenges to those who wish to regulate them. As a result, many forms slip through the net, subject to no specific forms of regulation. Nonetheless, most Western jurisdictions have adopted laws and policies that aim to confine sexual acts, goods, and nudity to specific spaces, and many forms of sexual display are subject to obscenity acts, which attempt to limit the circulation of materials with sexual content in both public and private spaces. While some have seen the rise of access to LGBTI sex goods, sex services, and pornographic imagery through the Internet as part of the ongoing sexualization of culture (Zook, 2003), it has presented new challenges to those Western jurisdictions who seek to control access to such commerce. Within a range of Western jurisdictions, the desire to protect 'family values' has thus encouraged the introduction of a rich and diverse range of legislation intended to control the content of and access to, for example, pornographic materials, with legislators constantly seeking to refine the boundaries between what is suitable for general consumption, what only (consenting) adults can access, or what is prohibited to all.

The regulation of LGBTI sexual commerce on the Internet is part of a long history of attempts to regulate spaces of LGBTI-related sexual commerce, including sex services, sex shops, adult bookshops, cinemas, gay bathhouses, and other premises where LGBTI sexual content is displayed or exchanged in Western jurisdictions. Such regulation often proceeds from the assumption, as discussed earlier, that commercialization of LGBTI sex promotes crime, nuisance, and disorder (Prior and Crofts, 2011). The extent to which commercial sex should be treated as potentially 'harmful' is hindered by the paucity of evidence about the nature and extent of its impacts. For example, currently, there is a reliance on a small number of US studies on 'secondary effects' that draw markedly different conclusions about the impacts of sex businesses in different cities (e.g., Linz et al., 2004). Regardless of whether decisions regarding harm are based on evidence or perception, the idea that LGBTI sexual commerce is more deviant than heterosexual sexual commerce has meant that LGBTI sex commerce has traditionally been subject to moral campaigns of prohibition, or subject to being socially and spatially marginalized.

Historically, much of the regulation of commercial sex has been driven by a focus on heterosexuality. This has had at times unexpected consequences on LGBTI commercial sex. It has also led to strange bedfellows – with radical feminists concerned about the exploitation of women by men, hooking up with

conservatives concerned about the commercialization of sex and homosexuality (West, 1987). For example, radical feminists Dworkin and MacKinnon (1988) developed an Antipornography Ordinance premised upon the harm of pornography to women by men. The ordinance was overturned as a violation of the First Amendment in the United States, but was influential in Canada. Although the ordinance was aimed at protecting women from men, there is evidence that gay and lesbian images were particularly targeted – displacing the concern about harm to women with a more conservative agenda of concern about LGBTI. Why the extra concern about LGBTI sexual commerce? LGBTI commercial sex for LGBTI audiences challenges dominant constructions of femininity and masculinity by displacing the heterosexual norm (Stychin, 1992), bringing into question the radical feminist analyses of women as 'fuckees' and as victims (Cornell, 1995). An example of such a challenge can be found in the actions at lesbian strip clubs such as *Gurlesque* in Sydney, Australia. *Gurlesque* was formed with the idea of unpacking patriarchal heterosexual preconceptions of the perfect female body, offering opportunities for professional and nonprofessionals alike to construct and express lesbian sexuality.

While criminal law has long been used as a mechanism to regulate LGBTI sexualities and their commercial forms, social research has increasingly highlighted how obscenity, licensing, zoning, planning, and licensing law are becoming more important in controlling the visibility and location of LGBTI commercial sex within a range of Western jurisdictions (Doan and Higgins, 2011). These mechanisms play a significant role in the process of sociospatial ordering within these diverse jurisdictions. Indeed, numerous studies have shown that the interpretation and enactment of these various mechanisms through the work of planners, licensing officers, environmental health officers, and police officers produce a 'patchwork' of local norms, marking conduct as either appropriate or transgressive depending on where it occurs (see Forsyth, 2001; Prior, 2008). The consequences of this for LGBTI businesses can be far reaching, with hotels, clubs, gay and lesbian bathhouses, and shops targeting LGBTI consumers deemed inappropriate in many residential spaces. For example, Doan and Higgins (2011) provide plentiful evidence that planning officers seek to promote neighborhood renewal by displacing LGBTI businesses to make areas more appealing to heterosexual families. Furthermore, irrespective of their function, many LGBTI-focused facilities be they for commercial sex purposes or not, are classified as commercial sex premises, and hence subject to the same types of zoning ordinances, planning instruments, licensing, criminal laws, that seek to locate sexually oriented businesses and sex premises away from 'family' areas, education facilities, and other 'sensitive' populations and land uses (Ryder, 2004). Regulations can also impact on LGBTI practices. For example, the aforementioned lesbian strip club *Gurlesque* had difficulties finding a venue that was licensed for full female nudity due to laws that were constructed to protect women from the male gaze. These rules seemed somewhat farcical when applied to lesbians stripping for lesbians.

The ongoing regulation of LGBTI commercial sex can be understood as sustaining broader economic and social processes that are unrelated to LGBTI culture. For instance, it has been argued that maintaining a separation between 'family-friendly' spaces and spaces of commercial sex (LGBTI oriented or not) helps construct a divide between good and bad femininity. Lasker (1998) argues that zoning protects 'good women with family values' from commercial sex land uses by "zoning such uses as far away from single-family, detached home neighborhoods as possible, and into neighborhoods that lower income women call home." Accordingly, the association of commercial sex with immorality and neighborhood nuisance appears an extremely effective strategy for reasserting the sexual values that lie at the heart of the nation–state and – at the same time – displacing commercial sex from valuable city center sites. On this basis, it is possible to interpret ongoing attempts to regulate commercial sex as simultaneous strategies of *capital accumulation* (i.e., enhancing the value of specific tracts of land or communities) and *social reproduction* (i.e., normalizing assumptions that pornography should not be consumed by minors or visible in the public realm). This is evidenced in recent zero tolerance approaches to commercial sex in potentially lucrative real estate locations, with zoning laws and police powers being used to displace commercial sex, ostensibly to protect local communities but often with the intention of prompting the gentrification of areas blighted by sex businesses (Papayanis, 2000).

This increased regulation of commercial sex, including LGBTI commercial sex premises, has contradictory outcomes. While such regulation reproduces gendered assumptions, the containment and isolation of spaces of adult entertainment conversely enhance the commercial value of erotic goods and services by granting them scarcity. Indeed, history suggests that isolating commercial sexuality within marginal locations does not serve to devalue it. Rather, the opposite is true: with spatial marginalization, LGBTI commercial sex is brought within the ambit of a *restricted economy* that hoards desire to capitalist ends. Thus the segregation of family spaces from commercial spaces serves to valorize both the family (as an idealized sociosexual relation) and certain forms of LGBTI commercial sex (as an illicit but sought-after commodity) (Sanchez, 2004).

See also: Feminist Theory: Radical Lesbian; Gay and Lesbian Movements; Gay and Lesbian Tourism; Heterosexism and Homophobia; Lesbians: Social and Economic Situation; Mass Media, Representations In; Prostitution and Sex Work; Queer Geography; Residential Segregation: Geographic Aspects; Sexual Minority Families of Choice and Family Networks: We Are Family; Sexual Orientation: Historical and Social Construction; Sexuality and Space; Social Movements: A Social Psychological Perspective; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Attwood, F., 2006. Sexed up: theorizing the sexualization of culture. Sexualities 9 (1), 77–94.

Binnie, J., Skeggs, B., 2004. Cosmopolitan knowledge and the production and consumption of sexualized space: Manchester's gay village. Sociological Review 52, 39–61.

Brodsky, J.I., 1993. The mineshaft: a retrospective ethnography. Journal of Homosexuality 24 (3), 233–251.

Brown, G., 2009. Thinking beyond homonormativity: performative explorations of diverse gay economies. Environment and Planning A 41 (6), 1496–1510.

Cornell, D., 1995. The Imaginary Domain: Abortion, Pornography and Sexual Harassment. Routledge, New York.
Doan, P., Higgins, H., 2011. The demise of queer space? Resurgent gentrification and LGBT neighbourhoods. Journal of Planning Education and Research 31 (1), 6–25.
Dworkin, A., MacKinnon, K., 1988. Pornography and Civil Rights: A New Day for Women's Equality. Organizing Against Pornography, Minneapolis.
Edwards, M., 2011. Gender, social disorganization theory, and the locations of sexually oriented businesses. Deviant Behavior 31, 135–158.
Forsyth, A., 1997. Out in the valley. International Journal of Urban and Regional Research 21 (1), 36–60.
Forsyth, A., 2001. Sexuality and space: nonconformist populations and planning practice. Journal of Planning Literature 15, 338–358.
Hanna, J.L., 2011. Toxic strip clubs: the intersection of religion, law and fantasy. Theology and Sexuality 16 (1), 19–58.
Hayes, S., Carpenter, S., Dwyer, A., 2011. Sex, Crime and Morality. Routledge, London.
Hubbard, P., 2009. Opposing striptopia: the embattled spaces of adult entertainment. Sexualities 12 (6), 721–745.
Kennedy, E., 1997. Telling tales: oral history and the construction of pre-Stonewall lesbian history. In: Duberman, M. (Ed.), A Queer World. New York University Press, New York, pp. 181–198.
Knopp, L., 1992. Sexuality and the spatial dynamics of capitalism. Environment and Planning D: Society and Space 10 (6), 651–669.
Lasker, S., 1998. Sex and the city: zoning pornography peddlers and live nude shows. UCLA Law Review 49, 1139–1185.
Linz, D., Paul, B., Land, K.C., Ezell, M.E., Williams, J.R., 2004. An examination of the assumption that adult businesses are associated with crime in surrounding areas: a secondary effects study in Charlotte, North Carolina. Law and Society Review 38 (1), 69–104.
Nash, C., 2006. Toronto's gay village (1969–1982): plotting the politics of gay identity. The Canadian Geographer 50 (1), 1–16.
Papayanis, M., 2000. Sex and the revanchist city: zoning out pornography in New York. Environment and Planning D: Society and Space 18, 341–354.
Podmore, J., 2001. Lesbians in the crowd: gender, sexuality and visibility along Montreal's Boulevard St-Laurent. Gender, Place and Culture 8 (4), 333–355.
Prior, J.H., 2008. Planning for sex in the city: urban governance, planning and the placement of sex industry premises in inner Sydney. Australian Geographer 39 (3), 339–352.
Prior, J.H., Crofts, P., 2011. Queerying urban governance: the emergence of sex industry premises into the planned city. In: Doan, P. (Ed.), Queerying Planning. Ashgate Publishing Ltd, New York, pp. 185–208.
Rooke, A., 2007. Navigating embodied lesbian cultural space: towards a lesbian habitus. Space and Culture 10 (2), 63–87.
Rubin, G., 1993. Rethinking sex: notes for a radical theory of a politics of sexuality. In: Abelove, H. (Ed.), The Lesbian and Gay Studies Reader. Routledge, London, pp. 3–44.
Ryder, A., 2004. The changing nature of adult entertainment districts: between a rock and a hard place or going from strength to strength? Urban Studies 41, 1659–1686.
Sanchez, L., 2004. The global erotic subject, the ban, the prostitute-free zone. Environment and Planning D: Society and Space 22, 861–883.
Sedgwick, E.K., 1994. Epistemology of the Closet. Penguin, London.
Stychin, C., 1992. Exploring the limits: feminist and the legal regulation of gay male pornography. Vermont Law Review 16, 857.
Tibbetts, S.G., Blankenship, M., 1999. Explaining citizens' attitudes toward pornography: differential effects of predictors across levels of geographic proximity to outlet sources. Justice Quarterly 16 (4), 735–756.
Valentine, G., 1993. (Hetero)sexing space: lesbian perceptions and experience of everyday spaces. Environment and Planning D: Society and Space 11, 395–413.
West, R., 1987. The feminist-conservative anti-pornography alliance and the 1986 Attorney General's Commission on pornography report. American Bar Foundation Research Journal 12, 681.
Zook, M., 2003. Underground globalization: mapping the spaces of flow of the internet adult industry. Environment and Planning A 35, 1261–1286.

Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures

Francisco I Surace and Heidi M Levitt, The University of Massachusetts Boston, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The lesbian, gay, bisexual, transgender, queer (LGBTQ) community often has been reduced to a homogenous group. However, multiple LGBTQ subcultures have developed across time. The leathermen and bears are two subcultures that arose out of a critique of the mainstream gay men's masculinity. The leathermen movement of the 1960s developed an aesthetic based upon the masculinity portrayed in media representations of bikers of that period, with a focus upon leather attire along with dominant and submissive sexual scripts. In contrast, the bears of the 1980s formed a masculinity that idealized naturalism, with a heavy set and hirsute appearance along with nurturing sexual practices. Both groups began by hosting invitation-only parties but became visible parts of the LGBTQ community through grassroots organization and using the community organization techniques of the civil rights movement.

Leathermen and Bears

The gay community, while sometimes seen as a single homogenous group united by a shared same-sex sexual orientation, contains multiple subcultures and identities. Bears and Leathermen are two of these identities that have an intertwined history but have evolved into their own subcultures. These two subcultures are defined by their focus on redefining the enactment of gay masculinity and by the development of large networks of gay men who claim identities based on membership in the subculture. Both subcultures were influenced by the social issues of their time in the United States, borrowing organizational techniques from the civil rights and human potential movement. Both subcultures have produced complex symbolic understandings and constructions of masculinity, highly structured social organizations, and elaborate practices.

Leathermen

History of the Leathermen

The birth of the leathermen movement can be traced back to the mid-1950s in the backdrop of the biker clubs (Rubin, 2004). These clubs held a camaraderie and thrill that drew in some of the gay veterans from World War II whose experiences in the armed forces had instilled them with the values of order and discipline. These values combined with the aesthetics of the bikers soon developed into the social codes and sexual roles of leathermen (Baldwin and Bean, 1993).

Leather bars, which were key locations for participating in the subculture, began opening in major cities like New York, Los Angeles, and Chicago in the early 1950s. Although San Francisco was considered to be late in the leather scene, it became a focal point of this movement – opening its first bar in 1958, which was said to have closed soon thereafter. However, in 1962 San Francisco saw the opening of the Tool Box followed by three other leather bars – forging three blocks of Folsom Street into what was considered 'leather square'. Additionally, leather paraphernalia that was used in the subculture became accessible via mail companies by the end of the 1960s.

The leathermen primarily utilized leather garb to accentuate their masculinity. Their assertion of masculinity was in sharp contrast to the mainstream representation of gay men as effeminate. When gay men's gender was presented as feminine it often was the subject of humor as well as heterosexist hostility. Leathermen defined themselves in contrast to this media image as well as to mainstream gay men, to whom they referred as 'clones'. The radical masculinity that the leathermen claimed could be seen as having a number of meanings (e.g., Mosher et al., 2006). It could be understood as a self-protective reaction in response to their gay identities being socially maligned as feminine within a heterosexist context. Their masculinity could be seen as a source of pride and as a way of outperforming heterosexual masculinity. At the same time, the valuing of masculinity has been associated with an internalization of sexism and a disparaging of the feminine.

The gay leather culture had close links with an increasing sadomasochistic (S/M) culture (described further below) that facilitated their national presence during the 1970s when multiple leather and S/M groups began to form throughout the United States, including the Eulenspiegel Society in 1971 and Society of Janus 1974 (Rubin, 2004). These two organizations were political in nature and were composed primarily of heterosexual men and women, but remained gay friendly, which helped move both the S/M community and leathermen forward. The 1970s were also marked by a growth in S/M events and parties (Rubin, 2004). For instance, in 1975 the Chicago Hellfire Club (CHC) hosted the first of a special invitation-only annual party called the Inferno. Commercial clubs were opened later such as Mineshaft in 1976, an internationally renowned afterhours sex club in New York City. The Catacomb opened that same year in San Francisco, a primarily leather gay male club that was also invitation-only. Later on the Catacombs also held a separate night for women and had mixed sex and mixed sexual orientation parties. By 1978, Samois, the first lesbian S/M organization was created, but was disbanded in 1982 (Rubin, 2004).

In time the leather community gained more attention and integration within the mainstream gay community (Stein, 2004). They became involved in two organizations that

were created in New York City during the 1980s, Gay Male S/M Activists (GMSMA) and the Lesbian Sex Mafia (LSM). Unlike other gay leather and S/M groups of the time across the United States such as the CHC and the 15 Association in San Francisco, GMSMA, and LSM were open groups and their members did not need sponsorship to join, which increased their numbers dramatically. In 1982, the GMSMA also helped further unite the leather community of New York City by creating the Leather Pride Night Committee, a coalition of local leather organizations. It was during this decade, however, that the leather community was hit hard by HIV/AIDS and with that sense of loss and devastation, leathermen became active in efforts to support AIDS victims, including founding the AIDS Emergency Fund (Levine et al., 1997). With their involvement in AIDS, they were increasingly considered to be part of the mainstream gay community. Seats were reserved for men and women representatives of the leather community on the steering committee for the 1987 March on Washington for Gay and Lesbian Rights, and by 1991, leathermen and corresponding leatherwomen had become a major presence in all pride marches around from New York, San Francisco, and Los Angeles.

Current State of the Leathermen Movement

One of the major events for the leather and S/M communities is the Folsom Street Fair, held every year in San Francisco and considered to be the Mecca of both the leather and S/M movement (http://www.folsomstreetevents.org/). The street fair was created in 1984 in response to challenges to the South of Market leather and S/M communities by both AIDS and urban development. Its aim was to create a fund-raising, political activism, and general entertainment for leathermen and for people who practice S/M. It succeeded very well, becoming an annual event that currently brings 400 000 people together from around the nation and world. The fair also has become a space for educating those interested in S/M by providing workshops and demonstrations.

The Folsom Street Fair inspired the creation of Folsom Street East and Folsom Europe. Folsom Street East (not officially affiliated with the Folsom Street Fair organization), started in 1997, has been held in New York, and is considered to be the largest S/M block party of the east coast. Its future, however, is uncertain since construction in the fair area resulted in the cancellation of the 2013 fair (http://www.folsomstreeteast.org/tdance2013). Folsom Europe, organized by the Folsom Street Fair organization, started in 2004 and is held in Berlin, drawing 25 000 to its 2011 event (http://www.folsomeurope.info/). In addition to these events, leathermen continue to have a strong presence in gay pride and gay cultural events.

Leathermen: Gender and Sexuality

Leathermen's sexuality closely overlaps with S/M sexual practices. The practice of S/M has a long history; Richard von Krafft-Ebing (1998) was the first person to introduce the 'sadism' and 'masochism' into medical terminology, which at the time were not related to pleasure or pain but rather were theories of control. In general, sex theoreticians have postulated psychological mechanisms driving its practice (Rubin, 2004). In his 1940 classic study of American sexual behavior, Kinsey recorded the number of individuals who participated in these sexual acts and his research somewhat normalized S/M sexuality, though at the time those who practiced these sex acts were thought to be living out their fantasies in secrecy and shame.

Due to its association with S/M, the leather community has long found itself outside of both the mainstream gay and heterosexual culture, as it became known for its different approach to gender and sexuality. In their interview-based research on leather masculinity, Mosher et al. (2006) found that leathermen saw this marginality as a central to their identity. To these men the leather community provided a sense of connection and camaraderie that resulted in greater self-acceptance and positive self-concept. They reported not fitting into the greater gay community, which they saw as holding superficial and classist values in contrast to the leather community that emphasized loyalty, egalitarianism, nondiscrimination, and acceptance. To offset the leather communities' marginality, multiple webs and networks were formed nationally to create a sense of 'brotherhood' among leathermen that extends beyond local groups.

Two important signifiers of the leathermen identity are their gender expression and sexual practices (Mosher et al., 2006). The overall presentation and appearance of leathermen is one of toughness and severity. Leathermen wear clothes made of leather such as vest, chaps, and cuffs, to accentuate their physiques as well as render them visible as members of this community. In addition, placing handkerchiefs or bandanas in back pockets have been used to signal sexual role and practice preferences where different colors signify desired sexual practices, and where pocket placement signifies top, bottom, or versatile positions. The use of these methods for symbolic communication has been referred to as the handkerchief code or hanky code, and was seen in the United States and abroad most often by leathermen and some casual sex seekers. Their leather garb and equipment is not only used as a form of communication, however, but it can acquire emotional significance through different sexual rites and can mark the process of becoming a leatherman. In addition, special leathermen flags are used to unite the community; these flags are made up of black, blue, and white stripes with a heart at the upper left quadrant.

Leathermen's sexuality is connected with their sense of self and identity, and the ways in which they structure their personal relationships by denoting dominance or submission as well as status within their community. According to interviews with leathermen (Mosher et al., 2006), their sexuality requires trust and transparent communication about sexual desires, decisions, and consensus about participation in S/M. This communication style is thought to result in a deeper attunement to sexual partners. A man in their study conveyed:

> There has to be a lot of trust involved in leather sex…you have to communicate and really know what you are talking about. I guess in a lot of situations I didn't really put that much thought into it – just somebody you meet on the street and you are going to have sex, you can probably walk away knowing that it's not going to be anything you can't handle. But in leather sex you never know. You have to talk about it first…. [Stating] your limits and what you are into and what you are not into…. After a while you learn to ask the right questions.
> (Interviewee 05; Mosher et al., 2006: pp. 112–113)

While leathermen borrow from the S/M community they use different terminology, for example using 'Leather Daddy' and 'Leather Boy' as ways to mark the power differential in the relationship (Mosher et al., 2006). The lines between leather sex and S/M are also blurry with less emphasis in bondage and pain, and more attention to leather symbols of masculinity. As in S/M in general, roles are negotiated and consensually adopted by partners in an exchange of mutual nurturance, communicated by subservience from the submissive and control by the dominant partner.

While submission, vulnerability, and nurturance are not traditional values associated with masculinity, the leathermen community codified these qualities into markers of masculinity through their sexual and social practices. For instance, submission in leather sex is not thought to be weak position, but rather a display of strength and commitment. In the same token, dominance is not seen as a stronger position given its dependence on the willingness of a submissive partner as well its ability to provide nurturance. In both positions strength and vulnerability are commingled (for classic texts on leather culture, etiquettes, and histories, see Townsend (1972), Mains (2002), and Baldwin and Bean (1993)).

Critiques of the Leathermen: Feminism, Culture, and Class

S/M culture, and by association leathermen, has been criticized by radical feminists who see the use of domination as both patriarchal and by extension unethical and nonconsensual (Deckha, 2011). Feminist critiques of S/M often focus on women and highlight the perpetuation of patriarchy in S/M by endorsing S/M's negotiated enactment of economic, psychological, and sexual abuse. The use of this abuse is argued to be evidence of objectification of women that is believed to be internalized by lesbians from a sexist society where pleasure is derived from the pain of women (Hopkins, 1994). However, this critique may be extended to gay men, who are also raised in a patriarchal and heterosexist society, and who are seen as deriving pleasure from inciting pain in submissive men. Moreover, these feminist theorists suggest that it is a naïve belief that individuals can consent in S/M. They raise concerns that S/M may be evidence of an internalized patriarchy that renders the abuse of women as permissible. The extent to which the experiences of leathermen are relevant to these concerns is debatable as their sexuality is enacted within male community and at times with careful attention to ensuring consensual practices (e.g., Henkin and Holiday, 1996).

Another critique comes from within the leathermen community itself. In the past, psychiatry had used the S/M practices of leathermen to further the notion that homosexuality was pathological; in response the community made changes to reduce psychiatric stigma (Harris, 1997). In this process, leathermen sex shifted away from an original power differential based upon the 'old guard' codes, the leather practices of the early 1950s, said to be based on discipline and order used in the military (Baldwin and Bean, 1993). In the old guard codes the slave was expected to provide pleasure to the master, currently these roles are said to have switched where master is responsible for the pleasure of the slave. It is argued that S/M became a practice of self-realization in line with the human potential movement resulting in an emasculation and feminization of S/M (Harris, 1997).

An additional critique from recent research has highlighted heightened risks for HIV within this community. Leathermen have been found to be 61% more likely to be HIV-positive when compared to nonleathermen (e.g., Moskowitz et al., 2011). The decreased condom use in this community was seen as one of the risk factors.

Bears

History of the Bear Movement

The beginnings of the Bear community and identity are hotly debated and are usually placed somewhere between the 1960s and 1980s (Wright, 1997). The bear movement was rooted in two existing cultures of the time, leathermen of the 1960s, and Girth and Mirth of the 1970s, usually sharing their spaces (Hennen, 2005). The term has been traced back to 1966 with Satyr, a Los Angeles biker club, mentioning the beginnings of a new bear club, but self-identification as bear was more common during the 1970s (Wright, 1997). The street usage of the term 'bear' was initially popularized in Dallas, Texas, by a group of men who proclaimed themselves lovers of 'Papa Bear'. Later on the term was used in an article titled 'Who's who in the Zoo: A Glossary of Gay Animals' written by George Mazzei that was published in the *Advocate* on 26 July 1979. By the 1980s the term 'bear' was used in San Francisco, New York, and Toronto to describe sections of the gay community.

The formal organization of the bear movement was set in 1986 in the middle of the AIDS pandemic and at a time when the leather community was becoming more difficult to access. At that point, men who called themselves bears had begun to form underground pockets of community and those communities, along with the appearance of paraphernalia related to bear lifestyles, started gathering greater attention. The bear community provided an alternate image of gay masculinity to the leather community. Like leathermen, this masculinity was in reaction to the stereotypes of gay men as feminine but they also accepted new members without the need for the rituals or training procedures that were common in leather subcultures – making it a more accessible community. At the time, some have viewed the bear body, full-bodied and hairy, as a symbolic defense against the images of thin gay men being shown as AIDs patients experiencing wasting syndrome (Hennen, 2005). Additionally, bears presented themselves in a manner that reflects an idealization of what are believed to be 'natural' aspects of masculinity such as strength, vigor, and virility.

During this time, computer bulletin boards and listservs were created making communication between individuals more accessible. These technological advances allowed bears to communicate about gatherings, locate community, and become involved in community even when living in nonurban settings. Bears adapted the practice of their fellow leathermen who wore handkerchiefs in their back pockets to make visible their sexual preferences, but instead bears wore Teddy Bears, as totemic representation of the natural and aggressive yet warm and nurturing aspects of being bears (Graves, 2007). In addition, the Teddy Bear was a way to not only attain visibility but to protest the reduction of gay identity to sexual acts and

instead emphasize their preference for cuddling and emotional intimacy.

The bear movement was influenced by the civil rights activity of the time, and borrowed from the early gay liberation movement that requested gay men to 'come out' in the name of visibility. The bear community asked for a second coming out of sorts – a proclamation of the bear identity (Wright, 1997). In this process, men engaged in public displays of the bear identity by adopting a bear appearance and joining groups that held similar views about being gay and gay masculinity.

In July 1987, San Francisco saw the first private invitation-only party called 'Bear Hug' created with an intention to promote an atmosphere of socializing, along with erotic and safe sex practices in midst of the AIDS scare (Wright, 1997). The party became popular and helped further unite the community that frequented them. In addition, *BEAR Magazine* began to circulate in San Francisco. In 1989, across the nation other bear communities began forming to such as the 'Bear Paws of Iowa' who created the first bear club. San Francisco also saw the opening of its first official bear bar, the Lone Star Saloon. The bar was revered as the Mecca of Beardom.

Wright (1997) described that in the 1990s the bear community entered what he termed the formalization phase. By 1994, 40 bear clubs had formed throughout the nation that helped generate a national understanding of what it means to be a bear. These clubs brought with them gay pride representation as well as bear contests, often based within bear clubs, and the International Mr Bear Contest. In addition 1994 saw the birth of new competing bear magazines such as *American Bear*. In 1991, *The Bear Cult: Photographs by Chris Nelson* (Nelson and Lucie-Smith, 1992) captured the image of bears as masculine men who sported beards and ranged in social class and physique.

What Is a Bear? Gender and Sexuality

The literature on the bears is still in its nascent stage, and much of the work is owed to Les Wright (1997) who edited two books, *The Bear Book* and *The Bear Book II*, outlining the history and evolution of the group. As he claims in the introductory chapter 'in this book bears speak for themselves' (Wright, 1997) and articulate their lived experience as means to create a working language. Most other research about the bear community that comes from the social sciences is similarly based upon interviews of bears, such as using grounded theory (Manley et al., 2007) and ethnography (Hennen, 2005).

Bear culture challenges both mainstream gay culture assumptions of what it means to be a gay man as well as the heterosexual notions of masculinity (Manley et al., 2007). While there is no one way to be a bear, and debate continues as to what constitutes a bear, different physical, psychological, and social criteria have been postulated throughout bear history. Overall, what seems to be central for bears is the valuing of naturalism and the social construction of an aesthetic that represented this value. The value of the natural man was embodied in a heavier build and being bearded, less maintained appearance in place of youthful clean-faced trim (mainstream gay culture), or masculine muscular exacting (leather culture) aesthetics. Many people who identified as bears praise the image of actor Steve Reeves in his role as a gladiator in 1950s, a look described as a rugged and mature masculinity.

Physically, two major features have been proposed as 'essential' to being a bear: (1) facial hair (beards, mustaches, and goatees) and (2) general hirsuteness. Facial hair is deeply intertwined with the bear self-identification and identity proclamation. A third criterion, a heavy set (burly or stocky) physique, has shifted over time within the bear community, with controversy dating back to the creation of *American Bear* magazine developed in the early 1990s. These magazines have displayed images of the bear ideal, with *American Bear* representing more men that had a stocky and heavy set build, and *BEAR Magazine* displaying an idea that was more in line with either 'muscle bears', and skinnier 'otters', where both maintained the first two criteria of facial and body hair. As part of the naturalism value, body acceptance and physical comfort were eroticized and a blue-collar or working-class attire was common (Hennen, 2005). Bears sporting ripped jeans, baseball caps, T-shirts, and flannel shirts typified this look.

In contrast to the focus on appearance, others within the bear community have emphasized the psychological dimensions and contend that being a bear is about a relaxed masculine attitude (Manley et al., 2007). Bear social networks are characterized by bears as having an easy-going outlook, camaraderie, and inclusivity, and are often contrasted with what is described by bears as the shallowness, pettiness, competition, and exclusivity of mainstream gay culture. In accordance with the emphasis on the natural, bears are considered to be specifically accepting of maturation and aging and include older members as active parts of their communities.

In addition, while the actual bear animal is thought of as dominating, such is not the case with bear masculinity since top (sexual insertive) or bottom (sexual receptive) positions are usually negotiated within a particular relationship (Manley et al., 2007). Sexual roles are sometimes defined in traditional familial terms with a 'daddy role' defined by assertiveness, nurturance, and guidance, with younger or less experienced men sometimes referred to as 'cub'.

Qualities of nurturance, affection, and emotional expression, which have been more commonly associated with femininity in Western cultures, have been redefined as aspects of bear masculinity. Bear social gatherings are marked by greetings with 'Bear hugs', and a sense of camaraderie is demonstrated between all members of the community through body acceptance (Manley et al., 2007; Hennen, 2005). In addition, bears rescript phallic-centric sexuality by masculinizing caressing and cuddling and eroticizing the whole body. This enactment of bear masculinity challenges essentialist notions of gender.

Socially, a bear identity provides men with a different set of values and aesthetics and a different enactment of masculinity. Yet, their masculine presentation also results in a dual invisibility where they may not be recognized or accepted by other gay men and are imperceptible as gay to heterosexuals. While this might provide assimilation and easy movement between mainstream heterosexual culture and bear culture, it also creates conflict for some who wish to be acknowledged as gay across these social spheres.

Many bears participate in local, national, and international events where they are able to meet other bears, have sexual

encounters, and develop relationships. These events not only help solidify bear identity, but include the opportunity for sexual expression, in what they call play parties. Overall, sexuality among bears has a wide and varied range with some opting for a one partner closed relationship and others for open or polyamorous relationships or for casual sex. In research, bears report that bears are both more sexually active and have more enduring relationships than other gay men (Manley et al., 2007). One participant in this study explained:

> I have seen bear couples in open relationships that have stayed together for many years, including my own…I don't see that same sort of relationship longevity in other groups within the gay community … [the longevity] may be due to more honesty in the relationship and the fact that there's no sneaking around…everything is out in the open.
>
> (Participant #2; Manley et al., 2007: p. 102)

The emphasis on intimacy and communication was thought to be characteristic of the bear community. It should be noted, however, that bears have been identified as being more likely to engage in unprotected anal intercourse in comparison to the mainstream gay culture (Willoughby et al., 2008).

Critiques of the Bear Movement: Feminism, Culture, and Class

The relationship of the bear community to the working class is complicated (Rofes, 1997). While bears aesthetics include dressing like working-class men, not all bears actually have blue-collar jobs – a long-standing critique of the movement's tendency to eroticize the bodies of working-class men and use this imagery as a symbolic place holder for masculinity that their everyday vocation does not provide (Hennen, 2005). After analyzing multiple articles from *American Bear* magazine that depicted working-class men while in contact with middle-class men, Rofes (1997) highlighted how working-class men were not only positioned as more masculine, but also viewed with disgust given the type of labor that they were scripted to perform. Therefore, he concluded that there might be an element of reaffirming class privilege present when middle- and upper-class men are drawn to bear spaces.

Race and class intersections are even more complicated in the bear community because the representations by bears after all are dominated by White working-class men. Few bears of color are present in bear media, despite the bear community's interest in diversifying their ranks (Hennen, 2005). In the book *Modern Bear* (Smith and Bale, 2012), a chapter was dedicated to the black bear (i.e., Black men) and another to the panda bear (i.e., Asian/Asian-American men), providing some representation of bears of color. It should be noted, though, that White bears are not associated with a specific bear type (e.g., polar bear), thus identification of bears of color might still remain controversial given the long and dehumanizing history of associating races with specific animals.

Current State of the Bear Movement

The bear community provides a sense of belonging and a place for gathering and connecting with other like men, as well as shelter from the marginalization these men might experience by both heterosexual and mainstream gay culture because of their sexual orientations and gender expressions, respectively (Manley et al., 2007). Currently, the bear community has grown, organized, and expanded to a global scale. In 2013, the Web site *Resources for Bears* (http://www.resourcesforbears.com) listed 60 active clubs in cities across the United States, 6 in Canada, 14 in Europe, 5 in Central America, and 6 in New Zealand/Australia. These bear clubs are open to Bears and 'their admirers' expanding the accessibility and membership base. In addition, bear visibility has expanded through the use of 'Bearphernalia' such as mugs, shirts, and bumper stickers displaying the colors of the bear flag: white and brown striped with a bear paw print. The visibility of the community has increased further via movies such as *Bear City* (Langway, 2010) and *Bear City 2* (Langway, 2012), documentaries *Bear Run* (Hunt, 2009) and *Bear Nation* (Ingram, 2011), and popular books such as *The Bear Book* (Wright, 1997), *The Bear Book II* (Wright, 2001), *Bears on Bears* (Suresha, 2009), and *Guide for the Modern Bear* (Smith and Bale, 2012).

Mr Bear contests and bear events, called 'bear runs' and 'bear weekends', have also become staples of bear culture that maintain the community. Competitors are judged on how much they have a bear 'presence'. In Massachusetts the Massbearz vote for Mr Boston Bear, in Orlando, Florida, for Mr Bear Bash, and in Kentucky men compete to be Mr Kentucky Bourbon Bear. In addition, many associations in the nation throw weekend-long parties once a year such as the Northeast Bear Weekend in Provincetown, MA, the SF Bear weekend in San Francisco, CA, and the Urban Bear Weekend in New York City. The Web site www.Bear411.com has become a major source for information for individuals interested in bear culture as well as those who may be looking for to meet other bears or to becoming involved in bear community. In addition, modern forms of communication have brought new ways for bears to connect such as using applications for smartphones like Growlr and Scruff that use employ GPS location and time-based technologies to identify other the distance of other bears who may be using the same application nearby.

Conclusion

It is evident that while the leather and bear community developed at different points in history, they both share important elements in their definition as a distinct subculture and identity within the gay community. For instance, they both have developed critiques of mainstream gay masculinity and new definitions of what it means to be a gay man via masculine gender expression. Both groups developed special spaces for meetings such as bars and dungeons, unique symbols such as flags, dress, and body types, and community events such as the street fairs and expos to unite their respective communities. The cultures formed by leathermen and bears extend beyond the generation of new gender expressions to redefine and reconstruct sexual practices. Both communities emphasize the role of trust, honesty, and nurturance as well as rethinking notions of monogamy and promiscuity. Yet another similarity between these groups is their marginality from both the heterosexual and gay mainstream resulting in a double exclusion that might

have led to openness toward those who would like to join their community. Both communities also are marked by an atmosphere of close camaraderie with their members. Another unfortunate commonality is that the risk of HIV appears to be elevated in both groups.

At the same time these two groups are marked by differences in gender and sexual expressions. Leathermen construct a masculinity that is an image of strength and power that places itself in sharp contrast to the feminine stereotypes of gay men. While bears' construction of masculinity includes an acceptance of what is thought to be a 'natural' physique and sets itself apart from both feminine images of gay men and muscle-bound leathermen, their rescripting of nurturance as masculine has been shaped in response to their perceptions of the leather culture as focusing on a harsh masculinity and of gay mainstream culture as emphasizing body ideals.

Leathermen and bears not only demonstrate the possibility of creating highly organized groups within an already marginalized culture, but also highlight how gender expression may evolve as reactions to discrepancies between internal experiences of selfhood and assumptions and stereotypes at play within dominant cultures (Levitt, 2006). These discrepancies can lead to sets of values (e.g., nurturance, naturalism) that may become embodied, materialized, and collectively reproduced in signs, symbols, and practices that become intrinsic to each culture. Ultimately, the genders created can have powerful social and personal implications, affecting identity and selfhood, the social and interpersonal spheres, and the erotic and aesthetic cultures within these communities.

See also: Feminist Theory: Radical Lesbian; Gay and Lesbian Movements; Gay and Lesbian Tourism; Gay, Lesbian, and Bisexual Youth; Gender: Gendering of Categories; Heterosexism and Homophobia; Lesbian, Gay, Bisexual and Trans-sexual Butch-Femme Subcultures; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Masculinity, Scripts of; Queer Geography; Sexual Orientation: Historical and Social Construction; Sexuality and Space.

Bibliography

Baldwin, G., Bean, J.W., 1993. Ties that Bind: The SM/Leather/Fetish Erotic Style: Issues, Commentaries and Advice. Daedalus Publishing, Los Angeles, CA.

Deckha, M., 2011. Pain as culture: a postcolonial feminist approach to S/M and women's agency. Sexualities 14 (2), 129–150.

Freud, S., 1953. Three essays on the theory of sexuality. In: The Standard Edition of the Complete Psychological Works of Sigmund Freud. (1901–1905): A Case of Hysteria, Three Essays on Sexuality and Other Works, vol. VII. The Hogarth Press, London, England, pp. 123–246.

Gayle, R., 2004. The catacombs: a temple of the butthole. In: Thompson, M. (Ed.), Leatherfolk: Radical Sex, People, Politics, and Practice. Daedalus Publishing, Los Angeles, CA, pp. 119–140.

Graves, D., 2007. The experience of being a 'bear': a phenomenological study of an American gay subculture. Dissertation Abstracts International 68.

Harris, D., 1997. Metamorphosis of the modern dungeon. The Harvard Gay & Lesbian Review 4 (2), 29.

Henkin, Holiday, 1996. Consensual Sadomasochism: How to Talk about It and How to Do It Safely. Deadalus Publishing, Los Angeles, CA.

Hennen, P., 2005. Bear bodies, bear masculinity: recuperation, resistance, or retreat? Gender & Society 19 (1), 25–43.

Hopkins, P.D., 1994. Rethinking sadomasochism: feminism, interpretation, and simulation. Hypatia 9 (1), 116–141.

Hunt, D. (Producer/Director), 2009. Bear Run: Celebrating the Bear Community. Motion Picture, United States.

Ingram, M. (Producer/Director), 2011. Bear Nation. Motion Picture, United States.

von Krafft-Ebing, R., 1998. Psychopathia Sexualis. Arcade Publishing, New York, NY.

Langway, D. (Producer/Director), 2010. BearCity. Motion Picture, United States.

Langway, D. (Producer/Director), 2012. BearCity 2. Motion Picture, United States.

Levine, M.P., Iardi, P.M., Gagnon, J.H., 1997. In Changing Times: Gay Men and Lesbians Encounter HIV/AIDS. The University of Chicago Press, Chicago, IL.

Levitt, H.M., 2006. Butch, femme, bear and leatherman: a programmatic exploration of gender identities within gay and lesbian subcultures. In: Kam-shing, Y.I.P. (Ed.), Psychology of Gender Identity: An International and Comprehensive Perspective. Nova Science, Hauppauge NY, pp. 105–121.

Mains, G., 2002. Urban Aboriginals: A Celebration of Leathersexuality. Daedalus Publishing, Los Angeles, CA.

Manley, E., Levitt, H., Mosher, C., 2007. Understanding the bear movement in gay male culture. Journal of Homosexuality 53 (4), 89–112.

Mosher, C.M., Levitt, H.M., Manley, E., 2006. Layers of leather: the identity formation of leathermen as a process of transforming meanings of masculinity. Journal of Homosexuality 51 (3), 93–123.

Moskowitz, D.A., Seal, D.W., Rintamaki, L., Reiger, G., 2011. HIV in the leather community: rates and risk-related behaviors. AIDS Behavior 15, 557–564. http://dx.doi.org/10.1007/s10461-009-9636-9.

Nelson, C., Lucie-Smith, E., 1992. The Bear Cult: Photographs by Chris Nelson. Editions Aubrey Walter, Swaffham.

Rofes, E., 1997. Academics as bears: thoughts on middle-class eroticization of workingmen's bodies. In: Wright, L. (Ed.), The Bear Book: Readings in the History and Evolution of a Gay Male Subculture. The Haworth Press, Bringhamton, NY, pp. 89–99.

Rubin, G., 2004. S/M's the catacombs: a temple of the butthole. In: Thompson, M. (Ed.), Leatherfolk: Radical Sex, People, Politics, and Practice. Daedalus Publishing, Los Angeles, CA, pp. 119–140.

Smith, T., Bale, C., 2012. Guide for the Modern Bear. Pixelita Press, Townsend, WA.

Stein, D., 2004. S/M's copernican revolution: from a closed world to the infinite universe. In: Thompson, M. (Ed.), Leatherfolk: Radical Sex, People, Politics, and Practice. Daedalus Publishing, Los Angeles, CA, pp. 119–140.

Suresha, R., 2009. Bears on Bears: Interviews and Discussions. Bear Bones Books, Birmingham, AL.

Townsend, L., 1972. The Leatherman's Handbook. Other Traveller, New York.

Willoughby, B.L., Lai, B.S., Doty, N.D., Mackey, E.R., Malik, N.M., 2008. Peer crowd affiliations of adult gay men: linkages with health risk behaviors. Psychology of Men & Masculinity 9 (4), 235–247.

Wright, L., 1997. The Bear Book: Readings in the History and Evolution of a Gay Male Subculture. The Haworth Press, Bringhamton, NY.

Wright, L., 2001. The Bear Book II: Further Readings in the History and Evolution of a Gay Male Subculture. The Haworth Press, Bringhamton, NY.

Lesbians: Social and Economic Situation

MV Lee Badgett, Center for Public Policy & Administration, University of Massachusetts Amherst, Amherst, MA, USA
Laura Durso, Center for American Progress, Washington, DC, USA
Alyssa Schneebaum, Vienna University of Economics and Business, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M.V.L. Badgett, volume 13, pp. 8725–8729, © 2001, Elsevier Ltd.

Abstract

Lesbians, or women who primarily have sexual and emotional relationships with other women, face a distinctive set of social, legal, and cultural challenges and stigmas worldwide. With new sources of data about sexual orientation, social scientists have identified features of lesbians' social and economic situation. An overview of this body of research reveals that lesbians generally face economic and social disadvantage, both in the workplace and in organizing their family life. However, an international trend toward greater legal recognition of lesbian families and protections against workplace discrimination suggests that the legal and social context for lesbians will improve.

Legal and Normative Challenges

In most countries, lesbians' economic and social conditions are determined within a social, legal, and cultural context ranging from situations of formal legal equality to contexts that are hostile to the existence of lesbians (ILGA, 2014; Rosenbloom, 1996). In some countries, same-sex sexual practices between women are punishable by prison terms or corporal punishment. States also actively oppress lesbians through arrests and police harassment, discrimination in government employment, exclusion from military service, and censorship of lesbian voices and organizations.

In many societies, norms about gender roles and sexuality stigmatize women who form primary and intimate relationships with other women, and lesbians are often denied the protection of law against the social, physical, and economic abuse that accompanies social stigma. Some evidence of the existence of norms against lesbianism (or homosexuality more generally) comes from public opinion surveys from a variety of countries on every continent (see Rosenbloom, 1996; Badgett, 2009). Reports from 30 countries provide examples of physical violence against lesbians, including rape, murder, forced conversion therapies, incarceration, and suicide (Rosenbloom, 1996).

In most countries in the world, lesbians cannot marry their same-sex partners, which is the method by which heterosexual women formalize and gain state and public support of their relationships. In addition, although other types of legal recognition have been extended to same-sex couples, the legal frameworks of these statuses differ across countries and jurisdictions, potentially limiting the ability of lesbians to support their families. Lesbians often lack legal ties to each other that assign inheritance rights, allocate property after a breakup, provide access to income security programs, and make couples eligible for spousal employment benefits, for example. The psychological and social value of marriage in most societies is also, therefore, denied to lesbians and can exacerbate social exclusion.

Social stigma can have wide-ranging consequences beyond legal inequality, such as being cut off from family support, being denied housing or access to important public resources, and facing discrimination in employment. While civil rights laws in most countries ban many forms of sex discrimination and offer recourse when discrimination occurs, civil rights laws rarely protect lesbians from discrimination based on sexual orientation. Firing or refusing to hire a lesbian because of her sexual orientation is legal in most of the world.

However, the legal recognition of same-sex couples has been rapidly changing, and at least one country on every continent but Asia gives substantial legal rights to same-sex couples. As of 2014, lesbians are able to marry in Argentina, Belgium, Brazil, Canada, Denmark, France, Iceland, Luxembourg, the Netherlands, New Zealand, Norway, Portugal, Spain, South Africa, Sweden, the United Kingdom, and Uruguay. Other countries have established alternative forms of legal recognition at the federal and/or city level, such as registered partnerships, and some countries grant some legal rights and obligations to cohabiting same-sex couples. In the United States as of mid-2014, nineteen states as well as the District of Columbia offer marriage rights to same-sex couples, two additional states offer civil unions or broad domestic partnerships, and one offers limited partnership recognition.

Protections against employment discrimination have expanded greatly in recent decades. In 2000 the European Commission issued a directive requiring member states to have laws against employment discrimination based on sexual orientation. At the time, only eight member countries had such laws, but the other countries have since added such protections. Most South American countries have protections in at least some areas of the country. In the United States as of 2014, 21 states had laws banning sexual orientation discrimination in employment, but proposed legislation at the federal level has not been enacted. Some research has shown the U.S. laws to be effective in reducing discrimination against gay men, but the research has not shown an effect for lesbians.

Finally, lesbians are subject to the same economic and social challenges as are all women. Research in several countries continues to show that lesbians earn less than gay and heterosexual men, just as heterosexual women do.

Methodological Issues in Studying Lesbians

Since the 1970s, researchers have begun to study the economic and social position of lesbians. Research on lesbians must first

address the tricky issue of defining 'lesbian.' Studies of sexuality in the United States reveal the lack of concordance between several possible definitions (e.g., Laumann et al., 1994). Women who are attracted sexually to other women will not necessarily act on those attractions. Women who have had sex with women may not refer to themselves as 'lesbians.' Researchers have focused generally their attention on women who have sexual relationships with other women or who identify themselves as lesbian, depending on the historical period or on the culturally appropriate and legally relevant definition.

Research on lesbians in surveys is challenging but has been improving. The social stigma attached to being a lesbian leads some lesbians to keep their sexual identities and intimate lives hidden. Researchers have addressed this situation by using 'convenience samples' drawn from bars or organizations frequented by lesbians, or through 'snowball sampling' methods to recruit study participants through personal networks of other lesbians. These nonprobability sampling strategies do not allow researchers to generalize a study's findings to the broader population of lesbians, unfortunately, since respondents might differ in important ways from the typical lesbian.

In the 1990s, some countries started counting unmarried couples in censuses and other population-based data, and some of those surveys also allowed for the identification of same-sex couples, creating opportunities in some countries to study women who lived in domestic relationships with another woman. Also, Canada, the United Kingdom, and the United States began to include sexual orientation questions on government and private surveys. These probability samples generated higher quality samples for studying the social and economic situation of lesbians. However, many samples are relatively small, allowing for little analysis of subpopulations of lesbians by race, ethnicity, or class, for example. Furthermore, samples of lesbian couples might not be representative of all lesbians. Lesbians in couples might have different demographic and psychosocial characteristics than single lesbians. To the extent that a difference between the groups of partnered and unpartnered lesbians exists, studies which rely on a sample of lesbians from just one group would not be able to generalize their results to the entire lesbian population. (For an excellent survey of these methodological issues, see Institute of Medicine (IOM), 2011.)

As a result of these issues, existing research on lesbians' economic and social conditions is not consistently available across countries and must be interpreted carefully. Generalizations about all lesbians based on a nonrandom sample are inappropriate, and even generalizations from probability samples might be inappropriate if disclosure of sexual orientation on a survey is biased in some way.

In addition to the methodological challenges, research on lesbians has been hampered by the stigma associated with such research. Some political scientists and sociologists in the United States and lesbian researchers from 14 countries, for instance, have reported difficulties in doing research on lesbian- and gay-related topics, such as job discrimination, professional isolation, and lack of funding (see Solarz, 1999).

Despite these roadblocks, researchers are making progress in describing and understanding the lives of lesbians. This research has been spurred on by social scientists' recognition that the study of lesbians (and gay men) can shed light on larger social and economic dynamics (e.g., Blumstein and Schwartz, 1983) and by the need for knowledge created by political debates arising from an international movement for gay and lesbian rights. A growing cadre of social scientists in sociology, anthropology, psychology, economics, history, political science, and other fields has analyzed lesbians' situation with respect to employment and other economic issues and to family life.

The Position of Lesbians and Their Families

In the U.S. women comprise more than half of the population of adults identifying as lesbian, gay, bisexual, or transgender (LGBT), with younger women (18–29 years of age) nearly twice as likely as men to identify themselves as a member of this community (Gates and Newport, 2012). These data indicate that women who form primary dyadic relationships with other women make up significant proportions of the LGBT community, including both lesbians and bisexual women. Additional demographic and cross-cultural research has demonstrated the racial and ethnic diversity of the lesbian community.

Lesbians are a part of civic life, just as heterosexual women, although there are some differences between the groups. One study from the United States (Herek et al., 2010) reported on demographics and social characteristics of lesbians, gay men, and bisexual men and women, showing that lesbians (along with gay men and bisexuals) are less politically conservative than the general population. In their sample, two-thirds of lesbians identified themselves as politically liberal, with around 80% identifying as Democrats. Lesbians were also more likely than heterosexual women to have served in the military.

Like their heterosexual counterparts, lesbians seek romantic partners who share their values and interests, and meet their partners through friends, workplace connections, or social events (Peplau and Fingerhut, 2007). However, the boundaries between lesbian friendships and romantic relationships may be less defined, presenting challenges to both initiating and dissolving intimate relationships. Compared to heterosexual women, lesbians are more likely to maintain friendships and engage in more frequent social contacts with former romantic partners (Peplau and Fingerhut, 2007). Lesbians are also more likely to be in an interracial relationship, compared to heterosexual women in different-sex relationships.

In some cases, lesbians chose to formalize their relationships by entering into a legally recognized union, such as a marriage, civil union, or domestic partnership, while others choose to take part in religious or personal ceremonies without seeking state recognition of the partnership. Individuals outside the primary relationship, such as 'families of choice' (Weston, 1991), may be considered part of the woman's familial network based on emotional and sometimes material bonds.

Many lesbian and gay couples have opted for legal recognition when it is available. One study estimated that 47% of American same-sex couples who live in states that offer some form of relationship recognition have legalized their relationship at some point in time, and female couples are more likely

to register or marry in the U.S. than are male couples (Badgett and Herman, 2011). Rates of marriage or registered partnership are lower in European countries, however (Badgett, 2009).

Many lesbian families include children. In Sweden, 34% of lesbian couples entering registered partnerships were raising children, as were 24% of Norwegian lesbian couples entering registered partnerships (Andersson et al., 2006). In the Netherlands, 18% of lesbian couples are raising children (Steenhof and Harmsen, 2003). Data from probability samples of the United States indicate that between 20 and 49% of lesbians have had children or have children in their households, and more than one-third of lesbians without children reported wanting to have them in one survey. Though a significant number of these children are often from past heterosexual relationships, recent data indicate that this phenomenon may be decreasing, in favor of other types of family formation, including adoption and the use of assisted reproductive technologies (Gates, 2011). Research related to the development of children raised by lesbian mothers indicates that these youth are well cared for and demonstrate similar levels of psychological and social adjustment as children raised by different-sex parents (see review by Patterson, 2009).

Studies of the inner workings of lesbian families find some similarities in the types of tasks and pressures faced by those families compared to families led by heterosexual couples. Lesbians in families raise children, divvy up housework, take jobs to support their families, and care for the family's adult members. As with heterosexual families, lesbians appear to arrive at family decisions through bargaining and appeals to norms, individual preferences, individual skills, and efficiency. In addition, lesbian couples are equally as likely as heterosexual and gay male couples to report being satisfied with their relationships and sexual lives, and use similar problem-solving strategies to address concerns within the partnership.

Although the processes within families show some similarities to heterosexual families, other contextual factors are quite different and result in marked differences. Studies of nonrandom samples of lesbian and heterosexual couples in the US and Britain show that lesbian couples have more equal distributions of housework and labor market work than do heterosexuals, and that their power relations are less affected by which person has higher earnings than is true of heterosexual or gay male couples (Blumstein and Schwartz, 1983; Kurdek, 1995; Dunne, 1997; Rothblum, 2009). Additional research using nonprobability samples indicates that same-sex couples show less negative affect toward each other, exhibit less fear or tension, and express greater affection, humor, and joy with their partners, compared to heterosexual married couples (Gottman et al., 2003, as cited in Rothblum, 2009).

However, while lesbians (and gay men) have reported to researchers that they prefer a more equal division of labor and have claimed to uphold these ideals themselves (Blumstein and Schwartz, 1983; Kurdek, 2007), some ethnographic research in California and New York has suggested that some lesbian couples display a more pronounced division of labor in their household work (Carrington, 1999; Moore, 2008). Other empirical research finds that fewer lesbian (and gay) couples divide labor market participation differentially when there are no children present, but the introduction of children into the household makes the patterns of the division of labor in lesbian (and gay male) couples more similar to those found in straight couples (Antecol and Steinberger, 2011; Giddings et al., 2014).

Social scientists have offered varying explanations for these patterns, with the hypotheses rooted in different theories about the division of family labor. Lesbians might have different and more egalitarian norms related to the contributions of partners and the economic role of women. Lesbians' expectations about likely labor market earnings for each partner might be more equal than those for married opposite-sex couples, reducing the advantage of having one partner do all the housework and the other partner earn all the money income. Lesbians do not have equal access to marriage, an institution that encourages stricter specialization of household labor. Or two lesbians' bargaining power might be more equal as a result of more equal earnings potential. All of those hypotheses would predict more egalitarian divisions of labor within the household, and additional research will be necessary to identify explanations that are most appropriate (Blumstein and Schwartz, 1983; Badgett, 2001).

The Economic Situation of Lesbians

The economic situation of lesbians is somewhat paradoxical. Although lesbians face the challenge of workplace discrimination based on sexual orientation (which is legal in most parts of the world and in most states in the United States), evidence suggests that lesbians' earnings are at least comparable to heterosexual women's.

Supporting the argument that the economic situation of lesbians is poorer than that of heterosexual women, many lesbians in the United States and the UK report on surveys that they have either been discriminated against or harassed at work because of their sexual orientation, or that they hide their sexual orientation at work in an effort to avoid discrimination. Social scientists have demonstrated that discrimination against lesbians in hiring is more than a suspicion. Results from so-called 'correspondence studies,' in which researchers send job applications with fictional résumés from potential employers who differ only in their sexual orientation, reveal that employers in Austria (Weichselbaumer, 2003) and in Greece (Drydakis, 2009) prefer heterosexual job applicants to equally qualified lesbians, even when the only thing that differs is the sexual orientation of the two applicants. The Austrian study used job applicants who varied in gender presentation as well as sexual orientation, and found that employers showed little difference in preferences for masculine vs feminine job applicants, but had a strong preference for heterosexual over lesbian applicants.

Beyond hiring discrimination, unequal and unfair treatment in the workplace can make a lesbian's job more difficult and less pleasant to do, and the resulting stress could lower a lesbian's productivity and therefore affect her future earnings. Moreover, the earnings of lesbian couples are lower than the earnings of many straight or gay male couples; since women generally earn less than men, the income of a couple with two women is on average lower than the income of a couple which includes a man. Lesbians therefore do not face economic challenges based only on their sexual orientation, but additionally based on their gender.

Conversely, other research on the earnings of lesbians finds that lesbians' earnings are comparable to or higher than heterosexual women's. Studies using large representative samples from the US, the UK, Canada, Australia, Sweden, and the Netherlands find that there is either no statistically significant difference in the wages of lesbians and straight women with similar education levels and working experience, or that lesbians earn higher wages than straight women (Klawitter, 2011). No study has found a negative relationship between wages and sexual orientation for women, although data to test for wage differences by sexual orientation are available only in a handful of rich countries. These studies rely on a variety of data sources, and the lesbians represented in them differ in important ways. In some surveys used in these studies, the lesbians under investigation are women who live with a same-sex partner. In other studies, lesbians are women who say that they identify as such. In others, lesbians are identified as women who have had any or mainly same-sex sexual partners. The identification of lesbians in these studies is therefore quite variable. Despite that fact, each of the studies finds that lesbians earn at least as much as straight women. Klawitter's (2011) meta-analysis of the lesbian earnings studies shows that the largest 'lesbian wage premium' is found in studies that use sexual behavior as the identifying characteristic of lesbians.

In the US studies which identify lesbians as those women living with a same-sex partner, lesbian wages are shown to be higher than straight women's. However, it is important to note that despite their higher earnings, the probability that a couple in the US will be in poverty is higher for lesbian (and gay) couples than straight couples (Albelda et al., 2008). Lesbian couples in the US are also less likely than heterosexual couples to own a home (Leppel, 2007), although they do hold more money in savings (Negrusa and Orefice, 2010.).

There are several explanations for the paradox that lesbians face discrimination in the labor market but do not have significantly lower earnings, and these have been tested in the US, the only country where large, representative datasets are available for this research. Both partnered and single lesbians have higher educational attainment than straight women (Black et al., 2003, 2007), a fact that contributes not only to higher wages but also to increased economic and social security. Further, lesbians in couples are more likely to work in occupations and industries that are dominated by men and that are higher paying (Antecol et al., 2008). In this sense, lesbians benefit from nonconformity to female gender roles. Lesbians might also benefit from employers' expectations that lesbians would be better workers than straight women; employers might believe, for example, that lesbians would have more aggressive and competitive personalities than straight women (traits that are desirable in a worker), or that lesbians are less likely to leave the labor force because of pregnancy or in order to handle childcare or family responsibilities.

Indeed, partnered lesbians have much higher rates of participation in the labor force than straight women, both in terms of choosing to work outside the home at all and in terms of the number of hours per week and weeks per year spent working (Leppel, 2009; Antecol and Steinberger, 2011). However, there are two important similarities in the labor supplies of straight couples and lesbian couples: in both cases, the person with higher yearly earnings in the couple works more hours per year than her or his partner, and the effect of having children on the labor supply of the higher earner is weaker than the effect on the lower earner (Antecol and Steinberger, 2011). In other words, in both lesbian and straight couples, comparative earnings within a couple play an important role in determining the working patterns of the people in the couple. These similarities between straight and lesbian couples hint toward the possibility of a division of labor in lesbian couples with children that may parallel that in heterosexual couples: one person does more of the household and childcare work, while the other person earns more of the income (Schneebaum, 2013).

Lesbian Health and Life Course Issues

Research indicates that different social issues emerge for lesbians across their life span. Youth who identify as lesbian, or who have romantic relationships with women, may experience family rejection or bullying and victimization among peers, all of which have been associated with poor health outcomes and poor academic achievement among lesbian, gay, and bisexual youth (IOM, 2011). However, many young lesbians are highly resilient and can effectively cope with these challenges (Saewyc, 2011).

Older lesbians face similar challenges as do heterosexual women, such as becoming caregivers for loved ones and discussing end of life care with their partners. However when compared to gay men, lesbians have a greater fear of outliving their income as they age and feel less prepared for retirement (MetLife Mature Market Institute, 2006). Indeed, probability surveys in the United States indicate that lesbians over the age of 65 who are in same-sex relationships are less likely to be eligible for and participate in an employer-sponsored retirement plan, have less income than married different-sex couples, and are more likely to be in poverty (Albelda et al., 2008; Goldberg, 2009). As a result, older lesbian couples are more likely to use public assistance programs (Goldberg, 2009).

Though subject to many of the methodological challenges addressed above, research largely conducted in the U.S. on lesbian health suggests that lesbians are at greater risk for a number of poor health outcomes, including cardiovascular disease, breast cancer, polycystic ovarian syndrome, and mental health concerns compared to heterosexual women (IOM, 2011). These outcomes are likely influenced by health behavior-related differences between lesbians and heterosexual women, including that lesbians are more likely to be overweight, to use tobacco products, and to consume alcohol (IOM, 2011). There is also evidence that lesbians do not seek preventive health care, such as receiving Pap tests or mammograms, as frequently as heterosexual women (IOM, 2011). These outcomes may be due in part to experiencing discrimination in the health care environment and/or fears of disclosing sexual orientation to providers (IOM, 2011). In the United States, lesbians in couples are less likely to have health insurance than are heterosexual married women, which might play a role in health disparities.

Research also suggests that the health of lesbians (and other sexual minorities) is affected by the sociocultural environment in which they live. For example, in the United States, negative

mental health outcomes are more frequently observed among lesbians, gay men, and bisexuals living in states that have constitutionally limited marriage to only different-sex couples, as well as in states that do not have laws protecting sexual minorities from hate crimes and employment discrimination (Hatzenbuehler, 2010). Discrimination in employment, lower wages, and a greater likelihood of being in poverty may all additionally contribute to lesbians not having health insurance coverage or not being able to pay for needed medical care.

Conclusion

While the social and economic position of lesbians around the world has improved in the last several decades, lesbians still face significant challenges to full equality with straight women. In much of the world, lesbian sexual activity is forbidden, if not heavily stigmatized and a cause of severe harassment. Lesbians face poorer health outcomes than straight women, cannot marry their same-sex partners in most countries, and face legally sanctioned discrimination in the workplace. Despite the extensive challenges discussed above, lesbians and their allies have made significant progress in obtaining social and legal equality based on sexual orientation, and lesbians and their allies have made important strides in demanding and obtaining social and economic equality for lesbians.

See also: Friendship Experiences of Lesbian, Gay, Bisexual, and Transgender Individuals; Gay and Lesbian Movements; Lesbian, Gay, Bisexual and Trans-sexual Butch-Femme Subcultures; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Lesbian, Gay, Bisexual and Transgender Aging; Medicine and Health for Sexual Minorities; Sexual Minorities in China; Sexual Minorities in South Africa; Sexual Minorities in the Maghreb; Sexual Minority Families of Choice and Family Networks: We Are Family; Sexual Orientation: Historical and Social Construction.

Bibliography

Albelda, Randy, Badgett, MV Lee, Schneebaum, Alyssa, Gates, Gary J., 2008. Poverty in the Lesbian, Gay, and Bisexual Community. The Williams Institute, Los Angeles.

Andersson, Gunnar, Noack, Turid, Seierstad, Ane, Weedon-Fekjær, Harald, 2006. The demographics of same-sex marriages in Norway and Sweden. Demography 43, 79–98.

Antecol, Heather, Steinberger, Michael D., 2011. Labor supply differences between married heterosexual women and partnered lesbians: A semi-parametric approach. Economic Inquiry 10, 1465–7295.

Badgett, MV Lee, 1995. The wage effects of sexual orientation discrimination. Industrial and Labor Relations Review 48, 726–739.

Badgett, MV Lee, 2001. Money, Myths, and Change: The Economic Lives of Lesbians and Gay Men. The University of Chicago Press, Chicago.

Badgett, MV Lee, 2009. When Gay People Get Married. New York University Press, New York.

Badgett, MV Lee, Herman, Jody, 2011. Patterns of Relationship Recognition by Same-sex Couples in the United States. Williams Institute, Los Angeles.

Black, Dan A., Makar, Hoda R., Sanders, Seth G., Taylor, Lowell J., 2003. The earnings effects of sexual orientation. Industrial and Labor Relations Review 56, 449–469.

Black, Dan A., Sanders, Seth G., Taylor, Lowell J., 2007. The economics of lesbian and gay families. The Journal of Economic Perspectives 21, 53–70.

Blumstein, Phillip, Schwartz, Pepper, 1983. American Couples: Money, Work, Sex. William Morrow & Co, New York.

Carrington, Christopher, 1999. No Place Like Home: Family Life among Lesbians and Gay Men. University of Chicago Press, Chicago.

Drydakis, Nick, 2009. Sexual orientation discrimination in the labour market. Labour Economics 16, 364–372.

Dunne, Gillian A., 1997. Lesbian Lifestyles: Women's Work and the Politics of Sexuality. University of Toronto Press, Toronto, ON.

Gates, Gary J., 2011. Family formation and raising children among same-sex couples. National Council on Family Relations Report FF51, F1–F5.

Gates, Gary J., Newport, Frank, 2012. Gallup Special Report: The U.S. Adult LGBT Population. http://www.gallup.com/poll/158066/special-report-adults-identify-lgbt.aspx (accessed 15.08.14.).

Gay and Lesbian Equality Network (GLEN) and NEXUS Research Cooperative, 1996. Poverty: Lesbians & Gay Men, the Economic and Social Effects of Discrimination. Combat Poverty Agency, Dublin, Eire.

Giddings, Lisa, Nunley, John M., Schneebaum, Alyssa, Zietz, Joachim, 2014. Birth cohort and the specialization gap between same-sex and different-sex couples. Demography 51, 509–534.

Goldberg, Naomi G., 2009. The Impact of Inequality for Same-sex Partners in Employer-sponsored Retirement Plans. The Williams Institute, Los Angeles.

Gottman, John M., Levenson, Robert W., Gross, James, Frederickson, Barbara L., McCoy, Kim, Rosenthal, Leah, Ruef, Anna, Yoshimoto, Dan, 2003. Correlates of gay and lesbian couples' relationship satisfaction and relationship dissolution. Journal of Homosexuality 45, 23–43.

Hatzenbuehler, Mark L., 2010. Social factors as determinants of mental health disparities in LGB populations: implications for public policy. Social Issues and Policy Review 4, 31–62.

Herek, Gregory M., Norton, Aaron T., Allen, Thomas J., Sims, Charles L., 2010. Demographic, psychological, and social characteristics of self-identified lesbian, gay, and bisexual adults in a US probability sample. Sexuality Research and Social Policy 7, 176–200.

Institute of Medicine, 2011. The Health of Lesbian, Gay, Bisexual, and Transgender People: Building a Foundation for Better Understanding. The National Academies Press, Washington, DC.

International Lesbian, Gay, Bisexual, Transgender and Intersex Association (ILGA), 2014. www.igla.org. (accessed 15.08.14.).

Klawitter, Marieka M., 2011. Meta-analysis of the Effects of Sexual Orientation on Earnings. Unpublished manuscript. Accessed at: http://evans.uw.edu/sites/default/files/public/EvansWorkingPaper-2011-08.pdf (accessed 15.08.14.).

Klawitter, Marieka M., Flatt, Victor, 1998. The effects of state and local antidiscrimination policies for sexual orientation. Journal of Policy Analysis and Management 17, 658–686.

Kurdek, Lawrence A., 1995. Lesbian and gay couples. In: D'Augelli, Anthony R., Patterson, Charlotte J. (Eds.), Lesbian, Gay, and Bisexual Identities over the Lifespan: Psychological Perspectives. Oxford University Press, New York, pp. 243–261.

Kurdek, Lawrence A., 2007. The allocation of household labor by partners in gay and lesbian couples. Journal of Family Issues 28, 132–148.

Laumann, Edward O., Gagnon, John H., Michael, Robert T., Michaels, Stuart, 1994. The Social Organization of Sexuality: Sexual Practices in the United States. University of Chicago Press, Chicago.

Leppel, Karen, 2007. Home-ownership among opposite- and same-sex couples in the US. Feminist Economics 13, 1–30.

Leppel, Karen, 2009. Labour force status and sexual orientation. Economica 76, 197–207.

Lewin, Ellen, 1998. Recognizing Ourselves: Ceremonies of Lesbian and Gay Commitment. Columbia University Press, New York.

MetLife Mature Market Institute, 2006. Out and Aging: The MetLife Study of Lesbian and Gay Baby Boomers. MetLife Mature Market Institute in Conjunction with the Lesbian and Gay Aging Issues Network of the American Society on Aging and Zogby International. Retrieved from: http://www.sageusa.org/uploads/OutandAging.pdf.

Moore, Mignon R., 2008. Gendered power relations among women: a study of household decision making in black, lesbian stepfamilies. American Sociological Review 73, 335–356.

Negrusa, Brighita, Oreffice, Sonia, 2010. Sexual orientation and household savings: do homosexual couples save more? Discussion paper series, Forschungsinstitut zur Zukunft der Arbeit, No. 4961. http://www.econstor.eu/bitstream/10419/36940/1/627212670.pdf (accessed 15.08.14.).

Patterson, C.J., 2009. Lesbian and gay parents and their children: a social science perspective. In: Hope, D.A. (Ed.), Contemporary Perspectives on Lesbian, Gay, and Bisexual Identities. Springer, New York, pp. 141–182.

Peplau Letita, A., Fingerhut, Adam, 2007. The closet relationships of lesbians and gay men. Annual Review of Psychology 58, 405–424.

Rosenbloom, Rachel (Ed.), 1996. Unspoken Rules: Sexual Orientation and Women's Human Rights. Cassell, London.

Rothblum, Esther D., 2009. An overview of same-sex couples in relation ships: a research area still at sea. In: Hope, Deborah A. (Ed.), Contemporary Perspectives on Lesbian, Gay, and Bisexual Identities, Nebraska Symposium on Motivation, vol. 54. Springer, New York.

Saewyc, Elizabeth M., 2011. Research on adolescent sexual orientation: development, health disparities, stigma, and resilience. Journal of Research on Adolescence 21, 256–272.

Schneebaum, Alyssa, 2013. Motherhood and the lesbian wage premium. University of Massachusetts-Amherst Department of Economics Working Paper 2013-04. Available at https://www.umass.edu/economics/publications/2013-04.pdf.

Solarz, Andrea L. (Ed.), 1999. Lesbian Health: Current Assessment and Directions for the Future. National Academy Press, Washington, DC.

Steenhof, Liesbeth, Harmsen, Carel, 2003. Same-Sex Couples in the Netherlands Statistics Netherlands. http://www.cbs.nl/NR/rdonlyres/74975167-2503-43A0-8821-66F88DA2B6B2/0/samesexcouples.pdf (last accessed 08.16.12.).

Weichselbaumer, Doris, 2003. Sexual orientation discrimination in hiring. Labour Economics 10, 629–642.

Weston, Kath, 1991. Families We Choose: Lesbians, Gays, Kinship. Columbia University Press, New York.

Levels of Analysis in Social Psychology

Willem Doise, Faculté de Psychologie et des Sciences de l'Éducation, Université de Genève, Genève, Switzerland
Joaquim Pires Valentim, Faculdade de Psicologia e de Ciências da Educação, Universidade de Coimbra, Coimbra, Portugal

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article presents a recent extension of the levels of explanation model initially presented by Doise (1982a, 1986). This extension now also includes a neurological level and an intersocietal level. The latter proved to be useful for explaining recent developments in the study of universal rights. The system/metasystem model initially proposed by Moscovici (1961, 2008) is also further extended. Its application to the judicial realm leads to new interpretations of Milgram's obedience experiments and of Zimbardo's Stanford Prison Experiment. Several studies also show the intervention of metasystems in attitudes toward social solidarity. Finally, complete sets of social psychological research articles in Italian and Portuguese journals may be exhaustively coded when using principles derived from the above models.

Levels of Analysis

The Four Levels Model

On different occasions and also in former dictionary contributions, Doise (1997a, 1997b) has already dealt with the diversity of explanations in social psychology. On the basis of previous extensive studies (Doise, 1982a, 1986), four levels of explanation were distinguished in the fields of the social psychological study of intelligence, of social influence, and of intergroup relations. These were the intraindividual, the interindividual/situational, the positional, and the ideological levels.

Intraindividual

This level of analysis refers to research dealing with the way individuals organize their perceptions of the social environment and the way they behave toward this environment. In these models, the interaction between individuals and social environment is not the direct focus of analysis. The object of analysis here is the mechanisms that, at the individual level, allow people to organize their experiences. In social psychology, classical research about perception of complex stimuli are also typical cases of this level of analysis, for example, research using the models of cognitive balance (Heider, 1946; Cartwright and Harary, 1956) or cognitive dissonance (Festinger, 1957). Also, the process of categorization (Tajfel and Wilkes, 1963) is focused on the modalities according to which an individual organizes his/her experience about the social environment.

Paradoxically, the first level's explanations do not seem to imply social explanations as they are focused on the cognitive and motivational processes individuals adopt in organizing their experiences. However, they are social to the extent that they necessarily introduce components of explanations that we locate at the other three levels.

Interindividual

The second level of explanations involves the intervention of interindividual and situational processes. Often at this level individuals are considered as interchangeable and it is their interaction systems that offer the explanations typical of this level of analysis. At this level of explanation, cognitive processes are, for instance, studied as embedded in, or even generated by, different types of interindividual coordinations. The research of communication structures (Bavelas, 1951) and attribution theory (Kelley, 1967) are good examples of studies situated at this level of analysis.

Positional

The third level of analysis concerns the different positions occupied by individuals and social categories in a given societal framework. For instance, concerning explanations of learning and other modes of cognitive functioning at this level, there is an intervention of different conceptions that social groups have in relation to the role of the school system in a given society. The study of the effects of different social positions in interaction was already present in one of the first experiments on social attribution (Thibaut and Riecken, 1955). In this study, an interindividual relation (one's success to persuade another) was articulated with differences in status preexisting to the experimental situation. This level of explanation is also important in the study of relationships concerning intra- and intergroup differentiation. Actually the articulation of intraindividual level (more or less intracategorial variability in the individual cognitive apprehension) and the positional level (group status or prestige) allows researchers to integrate contradictory results about the out-group homogeneity effect and to go a step further in their conceptualizations of this domain (Devos et al., 1996; Lorenzi-Cioldi and Doise, 1990; Valentim, 2008).

Ideological

The fourth level of analysis concerns ideologies, belief systems, social representations and norms, characteristics of a given society. Their role is to maintain the functioning of a given society notwithstanding its many social differentiations, oppositions, and divisions. The studies of Lerner (1980) on the belief in a just world, and the role of the prestige of science in the interpretation of Milgram (1974) results on the obedience to authority studies are examples of this level of analysis.

Assessing Explanations with the Four Levels of Analysis (1980)

An extensive study (Doise, 1980) of the first seven volumes of the *European Journal of Social Psychology* (EJSP) led to the conclusion that the four different levels of analysis were useful for assessing

explanations by an important sample of authors. The inspection of this corpus of published texts confirmed that indeed most explanations used in the EJSP could be situated at four levels. Notwithstanding the fact that levels 1 and 2 – respectively situated at the individual and interindividual level – accounted together for more than 70% of all explanations offered, other explanations in terms of societal characteristics were also used. Furthermore, when combinations of explanations of different levels occurred, and when level 3 – status and group differences – and level 4 – general beliefs and values – were used, they were often aimed at explaining variations of processes described at level 1 or 2, whereas the inverse pattern rarely occurred. Reductionism, i.e., recurring to individual or interindividual processes for explaining more collective phenomena, was rarely found. Concerning the differences between so-called American and European social psychology, it should be admitted that the four levels were used on both sides of the Atlantic, although the individual and interindividual levels may have been more favored by leading American social psychologists than they were by their European peers.

Contemporary Enrichments and Extensions of the Model

Research developments in social psychology since 1982 led Doise (2011) to revisit the four levels of analysis and to add two more levels, one on each 'side' of the spectrum: an intersocietal level and a neurological level.

Intersocietal Level

Doise (2002) tried to extend the levels of analysis model at an intersocietal level. In this epoch of globalization, humans of different origins and societies become aware of their interdependence and in these relationships, they initiate symbolic representations, social norms, and contractual principles, which often remain implicit. When they enter into a relationship with other persons, they know that the fate of participants in such a relation will be affected by their interaction, in some measure, within certain limits, and at a certain cost. Normative representations exist on what these effects and costs should be. Multiple forms of interdependence exist, characterized by all sorts of differences in status, purpose, interdependency, and formality. Various models of acceptable relationships, prototypes of fair and just relationships, principles of contracts that govern relationships exist. Such models of relationships intervene in multiple interactions; they are culturally but also cross-culturally defined and their application is sometimes guaranteed by institutions. In this context, Doise et al. (1999) developed an international research program on the social representations of principles of universal rights. More recently, colleagues in Lausanne, Spini et al. (2008) extended this research to the study of attitudes and beliefs in war-torn societies concerning basic principles of humanitarian law. This research is not just about societal values and beliefs, it implies the existence of a kind of supraordinate realm of relationships that tie together societies in a common normative framework.

Neurological Level

Since the original publication of the *Levels of analysis* book, other research groups have extended the spectrum of analyses at the opposite end (Doise, 2011). Many researchers now assess variations in neurological cerebral functions while social relations evolve, and some of them do not hesitate to recur to the terms of *social brain* or *social neuroscience*. For example, Amodio (2008) did this in the field of intergroup relations and Harris and Fiske (2009) in the important societal issue of dehumanization. Studies like these are important in furthering the construction of more precise distinctions between characteristics of sociopsychological functioning that seem to involve different neurological networks in the brain. However, the first integrative reports on studies using neural imagery do not redefine preexisting models of sociopsychological processes; on the contrary, they use these models to derive their working hypotheses on the intervention of specific neurological functions, for instance, when more affective or more cognitive aspects of a social experience are involved.

Intersocietal studies and studies of brain functioning do not make obsolete conclusions about variables and phenomena studied at the four levels described by Doise (1982a, 1986). They draw our attention to the important fact that societal values and beliefs are also shaped in an intersocietal context and that in order to be actualized at an individual level, specific neurological structures of the brain have to function. These new studies may enrich our comprehension of preexisting models; they do not invalidate them.

Articulation of Different Levels

The Concept of Articulation

While different kinds of explanation are to be distinguished in social psychology, this does not mean that in the reality of social psychological studies they are actualized separately. Actually, from the beginning of the studies of these levels, it was shown that different levels of explanation often intervened in the study of the same phenomenon. This finding gave raise to the notion of articulation of analyses and to the idea that the articulation of different levels of analysis provides a specific object of study for social psychology. The nature of each articulation is defined by the explanations that it puts in correspondence. Since they are of a great variety, so also are their articulations. "Articulating levels of explanation by hypothesizing the intervention of factors at other levels prompts the researchers to better describe a process conceptualized at one of the levels, while making explicit that this functioning presupposes conditions involving other levels of analysis as well" (Doise, 2011: p. 12).

Other Illustrations of the Articulation: The System and Metasystem Model

Already in 1982, Doise (1982b) pointed out another way of evidencing articulations of explanations based on the concepts of system and metasystem introduced by Serge Moscovici in his *princeps* study on social representations (1961, 2008; *see* Social Representations). Using interview studies, analyzing newspaper articles and survey results, he researched opinions, attitudes, stereotypes, and other kinds of beliefs that French people held about psychoanalysis more than 50 years ago. This led him (Moscovici, 2008: p. 167) to draw "major implications at the level of the working of the intellect. We can see that two cognitive systems are at work in the reflexive effort characteristic of science, philosophy or any form of thought whose goal is the 'apprehension' of categories. The first is an operational

system that works with associations, inclusions, discriminations and deductions; the second is a sort of metasystem that reworks the material produced by the first. The same is true of natural thought, but there is one difference. The metasystem or the relations that constitute it are usually and primordially normative relations. We have, in other words, ordinary operational relations on the one hand, and normative relations that check, test and direct them on the other. Normative values and principles are, by definition, organized."

The metasystem functions in social regulations; in different domains of adult thinking, the organizing principles of the metasystem may also explain intraindividual variations. When, for instance, the same individual acts as a scientist, or on another occasion as a politician, he/she may change his/her way of arguing about the same topic while adopting in one situation reasoning principles that would not be acceptable in another. On different occasions, individuals are part of different metasystems. It is of interest to note that a contemporary French sociologist, Alexis Ferrand (2011) independently came to a similar conclusion.

The development of an individual's cognitive skills or cognitive competences does "not oblige us to believe that these operations will be applied to any and every content. Once they have mastered their physical and ideological universe the child and adolescents are far from arriving at a general use of their intellectual tools. Society does not demand it from them" (Moscovici, 2008: p. 189).

Hence an important task for social psychologists is to study the links between cognitive functioning and social context, between cognitive operations and social regulations, in order to answer the question: "Which social regulations activate which systems of cognitive functioning in which specific context?" (Doise, 1990: p. 115). This means that we are necessarily dealing with the articulation between the level of cognitive functioning and a metasystem (the social regulations) that can be considered at different levels (interindividual, positional, ideological, and intersocietal).

Metasystems can reinvade the experimental setting in a way not anticipated by the experimenter. This was, for instance, the case in Zimbardo's Stanford Prison Experiment (Zimbardo, 1989). However, for our present concern it is more important to comment on the links that Zimbardo (2007: chapter 15) draws between the Stanford Prison Experiment and the Abu Ghraib scandal in Iraq. For explaining the Abu Ghraib incidents as a social psychologist, and also as a citizen, he had to take into account the larger societal context and this was highly important. It showed the limits of Zimbardo's own Prison Experiment, where an intervention of social representations about the prison system in the USA was not even envisaged: it was as if prisons had their social functioning independently of the larger social context.

Zimbardo does not use in his book the terms of system and metasystem but these concepts could be used for explaining what happened in situations such as the Abu Ghraib prison, where competent authorities had created an unlawful situation in which a relevant judicial metasystem ought not to be activated at the right time. Zimbardo's effort to situate the Abu Ghraib scandal in its larger context is fully supported by more extended documentation on the situation by Gourevitch and Morris (2008) based on interviews with the prison guards.

National and international social systems create judicial systems at their own level. Once they are functioning, they may also become kinds of metasystems intervening in different social settings. Such an intervention of a judicial metasystem was illustrated in experiences carried out in the Netherlands by Meeus and Raaijmakers (1995), who very carefully constructed an experimental paradigm for replicating essential aspects of the famous experimental paradigm that Stanley Milgram (1974) created in order to study obedience to authority. Like Milgram, the Dutch colleagues verified that the large majority of participants in their experiments did not hesitate to violate fundamental rights of another person in order to comply with orders given by an academic authority. In their case, the basic right that was violated was the right for an unemployed person to be hired for a job.

These authors introduced new experimental conditions in their studies in which the rate of obedient behavior was drastically lowered in comparison with the standard conditions of their research and of Milgram's research. In these new conditions, subjects were required to sign a statement in which they accepted full judicial responsibility for what could happen when the unemployed person involved might want to sue the academic authorities for what had happened to him/her in the experiment. The well-known obedience to authority collapsed when the possible intervention of a judicial metasystem was evoked. Often fatalistic views on the readiness of people to engage themselves in all kind of reprehensible acts refer to the Milgram experiments. Meeus and Raaijmakers' new experimental conditions about judicial responsibility showed that the independent functioning of judicial institutions may also be part of a social environment and that its mere evocation can lead persons to disobey authorities who want them to victimize others. Maybe autonomy to function at an individual level in some conditions requires institutions to function at a more societal level.

In a book chapter on normative social psychology, Valentim and Doise (2008) insist on the importance of distinguishing a description of an actual observed reality and a description of the normative ideas people have about what, according to them, this reality should be or could have been. In our opinion, social psychology research, and maybe especially experimental social psychological research, is often guided by a kind of fatalistic worldview, a belief that social reality as it is, is necessarily as it is. In other realms of thinking about social reality, distinctions such as the one bearing on a matter of fact and a matter of law often suggest that a given situation should have been, or could have been, different from the actual situation. It is in the light of such distinctions that we try to counter the harsh criticism that Zimbardo (2006, 2007: chapter 11) addresses to the so-called BBC Prison Experiment by Haslam and Reicher (Haslam and Reicher, 2012; Reicher and Haslam, 2006). Basically, he argues that it did not reproduce a real life situation.

Indeed, the BBC prison differed in important aspects from the Stanford Prison Experiment. One of these differences concerned the guard/prisoner relationship; participants were, for instance, informed that the prisoners who behaved well could become guards. These changes apparently blurred role differences and went against current social representations of the prison system. In the BBC experiment, a kind of virtual world was presented that differed from the actual world, a metasystem

was activated, different from the one that regulated the behavior in the Stanford Prison Experiment.

We can broaden perspectives on the articulation between systems and metasystems with further illustrations namely in the domain of studies in attitudes toward social solidarity. The research by Clémence et al. (1994) on *Solidarités sociales en Suisse* was presented by Emiliani and Palmonari (2009) in the first volume of *Paradigmi delle rappresentazioni sociali* (pp. 197–198). Let us remember here that according to our colleagues, basically two metasystems orient people's conceptions on solidarity in Switzerland: one considering society as a more or less harmonious set of social relations based on individual qualities as attitudes of altruism, and the other of a more conflicting nature, characterized by a vision of society opposing individual and collective interests that have to be coordinated.

Are these not two basically different ways of defining one's identity as a Swiss citizen? The first vision furthers a more homogenous vision of society with less danger and insecurity, and which does not involve an important investment of society in a redistribution of resources. The opposite is true for the other conception. The harmonious vision also leads to a more meritocratic view: everyone according to his/her merits, with some possible temporarily limited intervention of public funds in case of hardship, accidents, and so on. Quite different is a vision based on the assumption that inequalities exist between groups and categories of society, implying that public authorities should intervene in order to redress situations of social injustice related to their group memberships.

Hence for Swiss people at least two basic ways of defining one's own national identity and the identity individuals adhere to explain to some extent the readiness of their holders to allow public funding in favor of members of such categories as asylum-seekers, young and old people, handicapped people, drug addicts, and other marginalized people.

The existence of these links was investigated by Clémence et al. and assessed for younger and older respondents, and also by using experimental methods for enhancing the saliency of either basic conception.

Actually, none of the studies that we have presented mentioned the notions of levels of analysis and of their articulation. However, we think that the authors of these studies evoked successively explanatory principles from different levels and it is precisely the articulation of these principles that allows for a better comprehension of the phenomena studied.

Assessing the System and Metasystem Model

As mentioned before, the complete set of the first seven volumes of the EJSP was used to test the levels of analysis model. In order to test the system and metasystem model, the complete set of the five volumes 2006–10 of *Psicologia Sociale* was also analyzed. Let us briefly summarize the main conclusions of the study reported by Doise (2012). A total of 40 systems having their origin *outside* the individual were invoked as explanatory variables. These systems were related to social category memberships ($n = 16$); to family relationships, other proximity relations, or educational institutions ($n = 12$); to structures of relationships and interactions between individuals ($n = 7$); and to political issues ($n = 5$). Only 19 systems were considered as more *internal* to the individual: individual organizations of cognitions or emotions ($n = 11$), individual cognitive or emotional organization directly related to the perpetration or avoidance of some actions ($n = 8$).

Valentim et al. (2013) have done the same kind of analysis in three Portuguese journals covering the period between 2006 and 2012. They used the social psychology articles published in *Análise Psicológica*, *Psicologia*, and *Psychologica*. In the 74 articles analyzed, there is also a clear predominance (73.8%) of the 'systems having their origin *outside* the individual,' and the great majority of these systems are related to social categorizations and/or social identity systems. Systems considered more internal to the individual (cognitions and motivations associated with actions) are only a small part of the total of explanatory systems (26.2%).

Hence the conclusion of these studies is as follows. At the first glance, the omnipresence of systems of variables, under different forms, in the articles analyzed, seems to become a sort of 'lapalissade.' In fact, the evidence gathered could be considered trivial if we did not look in more depth. Of course, we do not advocate such a suspension of interpretative effort.

Certainly, much of the research published in the journals analyzed, using experimental, quasi-experimental, questionnaire-based, or other empirical methods and even theoretical articles, involves analyses that also appeal to other systems while recurring to analyses of social positions and of differentiation among social entities.

If the analyses that surpass the individual or the interindividual level abound in Italian and Portuguese social psychology studies, it is because the only metasystem that guides and regulates the research endeavors of social psychologists is not the one of a scientificity that would be blind to the dynamics of societal and intergroup relations. In fact, we consider that several metasystems are functioning within the general system of social sciences. They result from multiple modalities of overlapping motivations of research in individuals and teams, and in the nature of societal and intersocietal problems. It is clear that scientificity is a general rule but the application of such a standard of scientificity to more specific issues may also result from other social dynamics that are part of a scientist's motivation. The definition of an object of study is not only dependent on, and does not just translate a state of established knowledge. It may also arise from a basic desire to better comprehend the social dynamics in which we participate. Within these dynamics, the general conceptions about the relative importance of social units like families, educational systems, regions, political ideologies, nations, and cultures, have been for a long time already the subject of various discussions that now can be reformulated in terms of scientific debate heuristics.

Conclusion

Results of 30 or more years of enormous developments in social psychology research invited us to revisit the levels of analysis grid proposed in 1982 by the first author of this article. The four levels of analysis in social psychology (intraindividual, interindividual, positional, and ideological) proposed at that time could be expanded nowadays with two more levels for a more accurate articulation of explanations: the social brain and the intersocietal studies.

Furthermore, our investigations about explanations in social psychology were initiated without explicit references to

Moscovici's system and metasystem theory in the princeps study on social representations. Here we extend that aspect of social representations theory beyond its initial framework. The model proved its heuristic value for one of the main tasks of social psychological research: the articulation of explanations from different levels of analysis.

See also: Group Processes, Social Psychology of; Indigenous Psychology; Intergroup Relations; Justice: Social Psychological Perspectives; Obedience: Social Psychological Perspectives; Political Psychology; Representations, Social Psychology of; Social Categorization; Social Identity in Social Psychology; Social Psychological Theory, History of; Social Psychology; Tyranny.

Bibliography

Amodio, D.M., 2008. The social neuroscience of intergroup relations. European Review of Social Psychology 19, 1–54.

Bavelas, A., 1951. Réseaux de communications au sein de groups placés dans des conditions expérimentales de travail. In: Lasswell, H.D., Lerner, D. (Eds.), Les "sciences da la politique" aux Etats-Unis. A. Colin, Paris.

Cartwright, D., Harary, F., 1956. Structural balance: a generalization of Heider's theory. Psychological Review 63, 277–293.

Clémence, A., Egloff, M., Gardiol, N., Gobet, P., 1994. Solidarités sociales en Suisse. Réalités Sociales, Lausanne.

Devos, T., Comby, L., Deschamps, J.-C., 1996. Asymmetries in judgements of ingroup and outgroup variability. In: Stroebe, W., Hewstone, M. (Eds.), European Review of Social Psychology, vol. 7. John Wiley & Sons, Chichester, pp. 95–144.

Doise, W., 1980. Levels of explanation in the European Journal of Social Psychology. European Journal of Social Psychology 10, 213–231.

Doise, W., 1982a. L'explication en psychologie sociale. Presses Universitaires de France, Paris.

Doise, W., 1982b. A mudança em psicologia social. In: Knoke, A., Figueiredo, E., Tajfel, H., Leyens, J.-P., Jesuíno, J.C., et al. (Eds.), Mudança social e psicologia social. Livros Horizonte, Lisboa, pp. 41–65.

Doise, W., 1986. Levels of Explanation in Social Psychology. Cambridge University Press, Cambridge.

Doise, W., 1990. Les représentations sociales. In: Ghiglione, R., Bonnet, C., Richard, J.-F. (Eds.), Traité de psychologie cognitive, Cognition, Représentation, Communication, Tome 3. Dunod, Paris, pp. 111–174.

Doise, W., 1997a. Psicologia sociale. In: Enciclopedia delle scienze sociali, vol. 7. Instituto della Enciclopedia Italiana, Torino, pp. 189–198.

Doise, W., 1997b. Sociale (psychologie). In: Dictionnaire fondamental de la psychologie, vol. 2. Larousse, Paris, pp. 1203–1208.

Doise, W., 2002. Human Rights as Social Representations. Routledge, London.

Doise, W., 2011. The homecoming of society in social psychology. In: Valentim, J.P. (Ed.), Societal Approaches in Social Psychology. Peter Lang, Berne, pp. 9–34.

Doise, W., 2012. Sistemi e metasistemi negli articoli di ricerca di *Psicologia Sociale* (2006–2010). Psicologia Sociale 7 (1), 149–158.

Doise, W., Spini, D., Clémence, A., 1999. Human rights studied as social representations in a cross-national context. European Journal of Social Psychology 29, 1–29.

Emiliani, F., Palmonari, A., 2009. La scuola di Ginevra. In: Palmonari, A., Emiliani, F. (Eds.), Paradigmi Delle Rappresentazioni Sociali. Mulino, Bologna, Italy, pp. 177–209.

Ferrand, A., 2011. Appartenances Multiples. Opinion Plurielle. Presses Universitaires du Septentrion, Villeneuve d'Ascq.

Festinger, L., 1957. A Theory of Cognitive Dissonance. Row, Peterson, Evanston, Illinois.

Gourevitch, P., Morris, E., 2008. The Ballad of Abu Ghraib. Penguin Books, New York.

Harris, L.T., Fiske, S.T., 2009. Social neuroscience evidence for dehumanised perception. European Review of Social Psychology 20, 192–231.

Haslam, A., Reicher, S., 2012. When prisoners take over the prison: a social psychology of resistance. Personality and Social Psychology Review 16, 154–179.

Heider, F., 1946. Attitudes and cognitive organization. Journal of Psychology 21, 107–121.

Kelley, H.H., 1967. Attribution theory in social psychology. In: Levine, D. (Ed.), Nebraska Symposium on Motivation, vol. 15. University of Nebraska Press, Lincoln.

Lerner, M.J., 1980. The Belief in a Just World: A Fundamental Delusion. Plenum, New York.

Lorenzi-Cioldi, F., Doise, W., 1990. Levels of analysis and social identity. In: Abrams, D., Hogg, M.A. (Eds.), Social Identity Theory: Constructive and Critical Advances. Springer-Verlag, New York, pp. 71–88.

Meeus, W.H.J., Raaijmakers, Q.A.W., 1995. Obedience in modern society: the Utrecht studies. Journal of Social Issues 51 (3), 155–175.

Milgram, S., 1974. Soumission à l'autorité. Calmann-Lévy, Paris.

Moscovici, S., 1961. La psychanalyse, son image et son public. Presses Universitaires de France, Paris.

Moscovici, S., 2008. Psychoanalysis: Its Image and Its Public. Polity Press, Cambridge.

Moscovici, S., Marková, I., 2006. The Making of Modern Social Psychology: The Hidden Story of How an International Social Science Was Created. Polity Press, Oxford.

Reicher, S.D., Haslam, S.A., 2006. Rethinking the psychology of tyranny: the BBC Prison Study. British Journal of Social Psychology 45, 1–40.

Spini, D., Elcheroth, G., Fasel, R., 2008. The impact of group norms and generalization of risks across groups on judgments of war behavior. Political Psychology 29 (6), 919–941.

Tajfel, H., Wilkes, A.L., 1963. Classification and quantitative judgement. British Journal of Psychology 54, 101–114.

Thibaut, J., Riecken, H.W., 1955. Some determinants and consequences of the perception of social causality. Journal of Personality 24, 113–133.

Valentim, J.P., 2008. Identidade pessoal e social: entre a semelhança e a diferença. Psychologica 47, 109–123.

Valentim, J.P., Doise, W., 2008. De um estado de facto a um estado de direito: elementos de psicologia social normativa. In: Fonseca, A.C. (Ed.), Psicologia e justiça. Nova Almedina, Coimbra, pp. 37–77.

Valentim, J.P., Forte, T., Figueiredo, A., 2013. Psychosocial explanations in research articles in psychology: the Portuguese case and comments to Torres & Neves. Estudos de Psicologia 18 (1), 13–16.

Zimbardo, P., 1989. Quiet Rage: The Stanford Prison Study Video. Stanford University, Stanford, CA.

Zimbardo, P., 2006. On rethinking the psychology of tyranny: the BBC prison study. British Journal of Social Psychology 45, 47–53.

Zimbardo, P., 2007. The Lucifer Effect. Understanding How Good People Turn Evil. Random House, New York.

Levi-Strauss, Claude (1908–2009)

Dominique Bouchet, University of Southern Denmark, Denmark; and School of Creative Arts, James Cook University, Townsville, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Before Lévi-Strauss, anthropology was merely empirical and mostly interested in identifying different cultures' contextual framework. Lévi-Strauss is not just interested in understanding and explaining different cultures from within. He wants to figure out how humans think fundamentally and from there to understand cultural diversity. Lévi-Strauss is the first anthropologist to connect cultural diversity with the unity of humans, that is to say, with human nature. All humans do not think the same, neither do they unite in the same way, but they compose in the same way. They have the same propensity to create representations and institutions, the same desire to bring things into order. The human mind universally processes fundamental intellectual operations upon the perceived world. Those operations build on what the human senses and mind make possible.

Introduction

Claude Lévi-Strauss is one of the very few thinkers who shed new light on humankind and the human's ways to relate to the world. Today, it is almost impossible to reflect upon humankind, society, culture, nature, and the global world without referring to what Lévi-Strauss taught us. He achieved this contribution just after the very darkest moments in humanity's history; that of Nazism and the Second World War. He introduced a complete break with an anthropological tradition that had gone astray and contributed to find a new humanism instead of one that had been soiled by racism. The democratic world had to recover after Nazism. Anthropology had a responsibility in what had occurred. Thus, the French Anthropological Society (*La Société d'Anthropologie de Paris*), which had been established in 1859, had drifted toward a biological and racial discourse. This discourse had already been ousted by the Durkheimian sociological school, which claimed the necessity of studying man as a social being rather than as a biological being. Lévi-Strauss took over the project and illustrated how the symbolic opens a gap never to be closed again between the biological and the social. Based on an understanding of the specificity of the symbolic, he promoted respect for the diversity of cultures and called for the reconciliation of humankind and nature.

Lévi-Strauss, who had Jewish origins fled from France in 1941, like so many other French intellectuals, to take refuge in New York where the *École Libre des Hautes Études* was soon to be founded. There, he taught until 1994 and met Roman Jakobson who introduced him to the principles of modern linguistics, which inspired him to look anew into the way humans originally began to make sense of their world.

Exoticism in Brazil and Revelation in New York

Lévi-Strauss was born in 1908 in an open-minded but not very well-off French family that included several painters and musicians who often took him to the theater, the Louvre, and the Opera. After a secular education and the reading of many novels, he studied law and philosophy. Beginning in 1932, he briefly taught philosophy in two French high schools and soon realized that it was too repetitive a task for him. In 1933, he had just discovered ethnology. He was fascinated by Robert Lowie's book *Primitive Society* published in 1919. In the fall of 1934, he was advised to apply for a newly created sociology professorship at the University of São Paulo. This Brazilian sojourn, which would last until 1939, allowed him to lead two ethnographic expeditions. During the first one in 1936, in one of the very few Bororo villages that had not been influenced by Christian missionaries, he observed a population that had kept its customs and almost entirely dedicated daily life to an incredibly complex religious system that impacted all domains of social life. Marriages, clans, games, dances, costumes, weapons, all clearly related to the religious system. At this point, Lévi-Strauss wrote his first ethnological paper and gained recognition from Marcel Mauss and Paul Rivet who were then the French masters of the field. It is as an ethnologist that in 1938 he set out on an expedition through the Mato Grosso plateau up to the Amazon basin to meet the Nambikwara and Tupi. He was looking for societies, he only found human beings in despair. All the groups he met were in decay. As a result of repeated persecutions, they lived in utter destitution. They had to renounce their ways and lived along the routes where telegraph lines were being installed to beg in order to survive. Lévi-Strauss realized that the civilization developed in Europe was not good for poor countries. Later, in 1955, in his seminal so beautifully written *Tristes Tropiques*, he wrote: "The first thing we see as we travel around the world is our own filth slung onto the face of mankind" (Lévi-Strauss, 2011(1955): 38). Lévi-Strauss never went back to ethnographic work. He had come to the conclusion that social reality now best could be grasped by theoretical work based on already existing observations made before the state of decomposition was reached.

Soon after coming back to France in 1939, he was assigned as liaison officer in a war soon to be momentarily lost. It is thanks to his reputation as an ethnographer that he was invited by the ethnologists, Robert Lowie and Alfred Métraux, to the *New School for Social Research* in New York, thus being able to

escape persecution himself. His flat in Paris had already been looted and all the notes from his first fieldwork stolen.

In New York, he will meet with many a French intellectual, and many prominent American anthropologists: Franz Boas, Alfred Kroeber, Ralph Linton, Ruth Benedict, Margaret Mead, etc. The field is much older in America than in France. It is well established. Also, there is a long tradition and experience in field research. For a whole century a tremendous amount of information has been gathered about all the details of the Indian everyday life. Lévi-Strauss spends many hours in the New York Public Library perusing the vast literature gathered about the systems of kinship. Before leaving France he had just read in the *Annales Sociologiques*, Marcel Granet's 'Matrimonial Categories and Relations of Proximity in Ancient China.' In it, Granet was shedding light on seemingly very diverse rules of kinship by pointing out a limited number of principles integrated in those rules.

At the *New School for Social Research* he met Alexandre Koyré, the Russian born professor of religion history from the Parisian *École Pratique des Hautes Études*, who introduced him to the Russian linguist Roman Jakobson. Thanks to Jakobson, he discovered the principles of modern linguistics and phonology that were to play a very decisive role in his conception and definition of structure. With Jakobson and Troubetzkoy, phonology had moved from the study of conscious linguistic phenomena to their unconscious infrastructure. The smallest units of speech (the phonemes) do not carry an intrinsic meaning. Still, they contribute to communication. They are combined unconsciously by those who speak in a systematic way that can be studied. The important thing is not so much the sound units as such, but the relations between them. What matters is how they make a difference. The system is made of distinctions and oppositions.

A Fundamentally Symbolic Taboo

It is this very idea that will inspire Lévi-Strauss to study the different traditions of kinship as systems, the principles of which remain unconscious to those who enact them year after year. For humans to be able to communicate through language, ways of making distinctions have to be present in the human mind. They are present in all languages over the world. For kinship relations to exist and be maintained, there need be mental operations bringing into play some fundamental structures of the human mind present in all human beings of all times and of any society. As phonemes, kinship relations/terms are signification elements to be combined in a system.

Back in Paris, Lévi-Strauss was to submit his doctoral thesis *The Elementary Structures of Kinship* in 1948, which gave completely new insights into issues that were puzzling the field since the beginning (Lévi-Strauss, 1977(1949)). Rules of kinship stipulating who to marry or not was one of anthropology's major headaches. Most anthropologists, confronted with the contradictory nature of the diverse explanations of the prohibition of incest had left it aside, in much the same way that linguistics had renounced trying to account for the origin of language. Lévi-Strauss explained that the incest taboo could not be grasped as a psychological or biological phenomenon but had to be thought of, not as a negative prohibition, but as a positive rule of exchange instituting the social: The initiative of making unions is no longer at the sole initiative of natural relations. It is removed from the realm of nature and entrusted to the social group and ensures the group's existence as a group. It is such an imposition of rules that brings about culture as it replaces chance by organization. The prohibition of incest prevents the social group from spilling into "a multitude of families, which forming so many closed systems or sealed monads which no pre-established harmony could prevent from proliferating or from coming into conflict" (Lévi-Strauss, 1977(1949): 479). "The first logical end of the incest prohibition is 'to freeze' women within the family, so that their distribution, or the competition for them, is within the group, and under group and not private control" (Lévi-Strauss, 1977(1949): 45). Not only does the prohibition of incest institute society as the order of culture, but also as the realm of alliance. According to Lévi-Strauss, a culture is a system of differences, that is to say a set of terms and relations governed by rules. (But, as we will see, it is not *a Ding an sich*: The reality of each cultural group is always relational.)

What matters is not so much the highly varied forms of matrimonial but the universal character of incest prohibition. Under these varied forms, this founding norm of culture is to be found in all societies presented as a natural fact. For Lévi-Strauss, this indicates that something fundamental happened when the human species began to think and communicate in such a *symbolic* way. Lévi-Strauss argues that it is the advent of the *symbolic order* that constituted the advent of a truly human society. He speaks of a *symbolic origin of society* whereas Émile Durkheim wanted to explain the symbolic by the social. He also goes against Marcel Mauss whom he considers to be his master: "Mauss still thinks it possible to develop a sociological theory of symbolism, whereas it is obvious that what is needed is a symbolic origin of society" (Lévi-Strauss, 1987(1950): 21). Therefore, sociology cannot explain the genesis of symbolic thought, but has to take it for granted in man. He stresses that all social activity and production are indeed symbolic. However, symbolism is the very condition of the formation of any human society. It is coextensive to the very idea of institution. It is what makes possible all the different systems of differences – that can be called cultures – we experience in the various societies we now know of. They could not exist, if it was not for the symbolic system made possible by the human brain when language emerged. For Lévi-Strauss, "No social phenomenon may be explained, and the existence of culture itself is unintelligible, if symbolism is not set up as an *a priori* requirement of sociological thought" (Lévi-Strauss, 1945: 517–518). According to Lévi-Strauss, symbolism is a constructed order governed by a rule. Its arrangement of terms makes up a system. The specificity of a symbolic system is not to signify or illustrate; it is to operate. In other words, it is not just an arrangement; it makes it possible to perform something and to relate to many things. The symbolic function cannot be separated from a social system of reciprocal recognition and accepted conventions. A symbolic operation consists in grasping anything endowed with meaning within a system of relations to make sense of it in an operational way. "Any culture can be considered as combinations of symbolic systems" (Lévi-Strauss, 1987(1950): 16). The symbolic systems in which cultures originate are unconscious, and that there is

something to be combined in a symbolic system is due to the fact that the human brain ended up being able to organize its specific perception and interpretations in specific ways.

According to Lévi-Strauss, societies are rooted in a foundational belief. Cultures and societies are creations brought into being by the thought of their existence; they are not something that has an independent intrinsic existence independent of what mankind may think. Money is an institution; a river is not. In order for money to exist, people have to agree that it can exist and be used by them. All institutions are like that. They are artifacts, not things. For a society to exist, people have to agree that it does and to make it alive.

The relation between the evolution of the brain and the emergence of the symbolic would later be confirmed by studies of human evolution that showed the relation between the evolution of the brain and that of techniques. At the time of Durkheim the social was often conceived in relation to techniques. A few decades earlier, it was still common to assume that technical intelligence preceded and helped develop language. Today after so many scientists – aided with new technologies and new collective ideas – studied the evolution of tools, skull sepultures, and parietal art, the most convincing explanation is that both developed simultaneously and that art, spirituality, and rituals, such as funerals, could not have existed if it were not for the human language (Gibson and Ingold, 1995; Gibson, 1996; Langaney, 2000(1998): 26). The twofold structured symbolic language that humans use to transmit, address, and discuss meanings and significations is much more complex than animal languages that cannot be used to assign and combine significations as is permitted by human language.

Myths and Structures

In 1950, Claude Lévi-Strauss becomes professor of religious science at the *École Pratique des Hautes Études* where he is to study and teach comparative religious systems of peoples without inscribing whose way of life has remained the closest to what it was in the most distant past. This leads Lévi-Strauss to the study of myths to which he will apply the same principles he used in the study of kinship with the same ambition: To figure out the common logic at stake behind the seemingly very different cultural creations. He will continue applying the basic principle from linguistics that no symbol exists in isolation; that symbols always belong to symbolic systems. His structural hypothesis is again that behind the diversity of the many myths there exists rules to be discovered. No myth, no element of any myth can be understood in itself. Like the rules of kinship, myths, totems, masks can only be understood if one tries to look into what is said and not said, what is there and not there, and how the different elements relate to each other. To construct a mythology, Lévi-Strauss will employ his method of looking into differences, inversions, oppositions, and disparities.

It is a 'structural' method but it is not so much that a structure actually exists; and this structure does not have much in common with the former models inspired by architecture – from which, for instance, the outdated Marxist model of basis and superstructure came from – or biology. It is more like an astronomic configuration but with more complexity. (In fact, Lévi-Strauss once explained how he was inspired when watching a dandelion in bloom.) For Lévi-Strauss, a structure does not have much in common with a biological entity, it is a tailor-made model constructed in order to reveal the hidden active 'underlying' organization of an observed phenomenon. It is the mode of organization of a sensory realm that turns it into a symbolic system. And, one can neither observe a symbolic system or a language-like one does a biological entity. Therefore, Lévi-Strauss did not find Radcliffe-Brown's comparison between organic structure and social structure appropriate (Lévi-Strauss, 1972(1958): 303). For him, "the term 'social structure' has nothing to do with empirical reality but with models which are built after it" (Lévi-Strauss, 1972(1958): 279).

It is known that myths play a common role. They have a function.

> "What peoples without writing ask of myths, what humanity as a whole has asked of them for the hundreds of thousands of years – perhaps millions of years – of human history, is to explain the order of the world that surrounds us and the structure of the society into which one is born. The aim of myths is to demonstrate the soundness of the world as a whole and that of the particular society [82–83] to which one belongs, to inspire the absolute confidence that they will remain the same as when they were created at the beginning of time"
>
> (Lévi-Strauss, 2013(2011): 82–83)

Beyond this fact, Lévi-Strauss wants to reveal the logical framework of the stories. Thus, it is not so much the stories as such that matter. One again has to look into the different and similar relations between the components, to study the relations the series of components maintain/have with one another. The kind of thinking that myths convey as a whole differs from what is expressed story wise in each myth respectively.

Lévi-Strauss chose to study the gathered corpus of the Amerindian myths. He finds out that the many different Indian cultures developed very particular myths, which at the same time are remarkably homologous with respect to each other. It is so, Lévi-Strauss tells us, because they are the fruit of a single human mind. It is because there are some fundamental, not categories, but differentiations. That is to say, there are fundamental building bricks prior to conceptualization. His thesis is that myths represent the first and most profound expression of the human mind. Myths can be said to be consubstantial with the mind. The same goes for language in its constitutive form, which is why the way myth and language operate have so much in common. The human mind, or rather a mechanism proper to the human mind, generates the structure of myths, as well as the structure of language. Lévi-Strauss confronts us with a multiplicity of levels that strangely enough only make sense in their relations and transformations. This huge network of permutation and transfer works on its own. No human transmitting the myths from generation to generation is actually conscious of the mechanism. No one is aware of what goes on and of how it works. Lévi-Strauss points out that it is as if the myths think by themselves. Somewhat, myth as such does the talking.

Thus, the concept of myth in Lévi-Strauss's work does not so much refer to a specific kind of story relating to the origins of the world, people, languages, rituals, and such things. No, for Lévi-Strauss, myth is a fundamental and original category of the human mind, to which a specific way of thinking corresponds, namely mythic thought, which proceeds by successive dichotomies and oppositions. It is because of this universality of mind, that one can identify redundant themes and similar fundamental oppositions in the stories told in all different cultures. By studying the appearance, disappearance, and replacement of certain themes in favor of others, Lévi-Strauss did show that myths make use of oppositions between sensitive categories promoted to logic, a logic that matters as least as much as 'reality.'

Lévi-Strauss's corpus of myth is made of more than 800 myths taken from dozens of different Amerindian tribes across the two Americas. Thus, for instance, in the first of the five volumes of *Mythologiques*, he looks into how the 'raw-cooked' dichotomy – which relate to the opposition between nature and culture – operates in various myths (Lévi-Strauss, 1986(1964)). Mythic thought, he explains, reduces reality by means of a specific binary logic of oppositions, and it is so because its own rule of functioning has to build upon such a kind of binary code. This is why one finds similar ways of arranging the world into myths in so many different cultures. This is why myth and thought can be said to be independent of contingent historical conditions.

In his seminal introduction to the work of Lévi-Strauss, Marcel Hénaff, stresses the importance of the mechanisms proper to the human mind in the generation of the structures of the human myths, and why according to Lévi-Strauss one should not look for a social origin of symbolism, but rather for a symbolic origin of the social: "The intelligibility that there is in institutions and organizations comes not from the social as such but from the fact that the collective systems has the required content, and it is because it has that content that the social is possible" (Hénaff, 1998(1991): 119).

This of course does not mean that myths are transcendental in nature; to the contrary. Because this mind and this binary logic come from the materiality of the human brain, "myths signify the mind that evolves them by making use of the world of which it is itself a part" (Lévi-Strauss, 1986(1964): 341).

Structuralism, Not Determinisn

Lévi-Strauss tells us that "structural arrangements are not a pure product of mental operations" (Lévi-Strauss, 1985(1983): 117). He emphasizes the importance of the senses: "The eye does not merely photograph the outside world. It rather encodes its formal characteristics. These characteristics consist less in the sensuous qualities of things than in the relationships which prevail between them" (Lévi-Strauss, 1973: 20). He notes that atoms, molecules, cells, organisms, and the sense organs also function structurally. He points out that the mind continues to work out structurally a material that it has received structured. And that "it can only do so if the mind, the body to which the mind belongs, and the things perceived by body and mind are part and parcel of one and the same reality" (Lévi-Strauss, 1985(1983): 117).

Organisms do rely on internal interactions and do interact. It does not imply only one world to be possible. Likewise, Lévi-Strauss pointing at the binary logic emerging from the materiality of the human brain does not imply an absolute determinism. He knows that societies evolved differently and could have evolved differently. Brain proposes, society disposes. He is more interested at pointing out the fact that, "however variable the rules may appear when we compare societies, they include constant elements, which are interpreted differently or which differ only in their application" (Lévi-Strauss, 1978(1973): 70).

Lévi-Strauss does not intend to deduct all further social development from the origins of language and culture. He just shows that equipped with symbolic language humans are related to their world, unconsciously making use of distinctions and oppositions related to their sensory abilities. He also explains that such building bricks of sense making will always be in use in human minds and cultures even though the whole system and every subsystem obviously ended up mingling into something very complex. As Maurice Godelier points out:

> By revealing the existence of mental constraints and forms of sociability common to all humanity, Levi-Strauss shows that the cultural and social diversity in time and place, does not exclude but rather presupposes the universality of core structures of the human mind. Therefore, absolute relativism does not make sense because, beyond what distinguishes and separates cultures and forms of society, the same mental structures and patterns of sociability remain present and active in all societies and at all times. Thus each society is a unique combination of universal and specific structures, which do not oppose each other. They don't because the universal structures are the condition of existence of the specific ones and also the distinctive form they take as social realities.
>
> Godelier, 2013: 62; my translation

Too many introductions to Lévi-Strauss work tend to reduce his structuralism to a determinism ignorant of historical factors. In fact, Lévi-Strauss stresses that history from the very beginning introduced intrinsic modifying elements into the underlying original, never fully completed symbolic system. He points out that "a society is always spatial-temporal given, and therefore subject to the impact of other societies and of earlier states of its own development" (Lévi-Strauss, 1987(1950): 17–19). He emphasizes that even in hypothetic isolated societies, "the different systems of symbols which all combine to constitute the culture or civilisation would never become reducible to one another" (Lévi-Strauss, 1987(1950) #24987: 17). The translation of one system into another happens at different paces and within a network. In other words, the translation is "conditioned by the introduction of constants with values external to the two systems" (Lévi-Strauss, 1987(1950): 17).

Cultural Diversity and the Unity of Humanity

According to Lévi-Strauss, "anthropology should not be reduced to the study of relics to be sought close to home or far away. What matters above all is not the archaism of these forms of life but their differences from one another and from those that have become our own" (Lévi-Strauss, 1978(1973): 29). The purpose of anthropology is not so much to discover the

unique characteristics of societies but in what they differ from one another. "As in linguistics, it is the *discontinuities* which constitute the true subject matter of anthropology" (Lévi-Strauss, 1972(1958): 318).

Lévi-Strauss has "tried to show that the real contribution of cultures does not consist in a list of their particular inventions but in the *contrastive features* which exist between them" (Lévi-Strauss, 1978(1952): 358). Lévi-Strauss points out that "the diversity of human cultures (…) is less a function of the isolation of groups than of the relationships which unite them" (Lévi-Strauss, 1978(1952): 328). He stresses the importance of cultural diversity for the many cultures and the whole civilization, which, as he sees it, consists of the coexistence of cultures. World civilization, he says, "could not be anything on the world scale except the coalition of cultures, each preserving its originality" (Lévi-Strauss, 1978(1973): 358).

Thus Lévi-Strauss has always stressed the inanity and vanity of condescension as well as contempt. His seminal UNESCO lecture from 1952 (Lévi-Strauss, 1978(1973)) is one of the greatest monuments ever erected not only against racism of any kind (biological and cultural) but also against arrogance. In this lecture entitled *Race and History*, he analyzed the notion of progress and showed that no objective criteria could be provided for a comparison and evaluation of human societies. Cultural progress is not due to the superiority of a specific society but to each culture's aptitude in establishing mutual exchanges with the others. Just like humans and other living beings do, cultures rely on each other. Lévi-Strauss often stressed that the dominating Western civilization would not have turned up to be what it is today if it had not been confronted by other cultures that made its members think differently. In Europe, the Crusades, Marco Polo's long visit to China, the Ottoman invasions, the progress of geographical exploration, fed cultural inspiration, and anthropological reflection (Lévi-Strauss, 2013(2011): 7–9). It is worth noticing that the Renaissance rediscovered forgotten notions and methods in ancient literature. However, more important still, "it realized the means of putting its own culture in perspective – by confronting contemporary concepts with those of other times and other places" (Lévi-Strauss, 1978(1973): 272); an intellectual method Lévi-Strauss called the "technique of estrangement" (Lévi-Strauss, 1978(1973): 272). Already then, it was known that "no fraction of humanity could aspire to understand itself without reference to all the other human beings" (Lévi-Strauss, 1978(1973): 272), that "no civilization [could] define itself if it does not have at its disposal some other civilization for comparison" (Lévi-Strauss, 1978(1973): 272).

Nevertheless, humanism was then still restricted in many ways. The types of cultures known were geographically limited. The classic civilizations had disappeared. The interest in the Middle and Far East applied only to their most scholarly and refined productions. Appropriate anthropological methodologies were still lacking to study civilizations without written documents 'discovered' in the Americas. Last but not least, classical humanism was also restricted in its beneficiaries and linked to industrial and commercial interests. Little by little, ethnological knowledge of other customs that astonish us, shock us, or repel us called ourselves into question, and invited us "to temper our misplaced vanity, to respect other ways of life" (Lévi-Strauss, 1978(1973): 36–37). Constructed in dialogue with different scientific traditions, anthropology revealed that "every custom, every belief, however shocking or irrational it may appear to us when we compare it to our own, is part of a system whose internal balance has been established over the course of centuries" (Lévi-Strauss, 1978(1973): 44). Furthermore, Lévi-Strauss means to have shown "that one cannot eliminate an element from the whole without running the risk of destroying all the rest" (Lévi-Strauss, 1978(1973): 44).

Since his visit to the Bororos, Nambikwaras, and Tupis in the late thirties, Lévi-Strauss has been very much concerned with a lack of understanding: that of the diversity and unity of mankind. Not only does it have disastrous consequences for the so-called 'primitive' people whom he clearly showed were at least as intelligent, clever, and skilled as we are, but in a different way, it also threatens our species and our planet with extinction. Economic growth and globalization might lead to societal blight. The diversity, so vital to any form of life, decreases. Lévi-Strauss is concerned about the biosphere, but also about cultural diversity. He tells us that those two problems are related. As cultures exchange, their differences diminish and might come to the point of disappearing and there will not be much more to exchange. This might weaken humanity as it has to cope with the challenges of a decrease in biological diversity. So, "both the absence and the excess of communication have their dangers" (Lévi-Strauss, 1958: 149; Lévi-Strauss, 1961(1952): 206–207). Thus, one of the most prominent fighters against racism of all time dares criticize the dangers of a naïve and blind form antiracism that does not take into account the lessons of social anthropology. Value relativism and multiculturalism have their limits. He stresses that cultures can exhibit greater or lesser affinity with other cultures – a fact of history and a normal behavior that should not be confused with racism (Lévi-Strauss, 1958: 150; Lévi-Strauss, 1961(1952): 209). Here is what he has to say on this:

> Humanity must learn (…) that all true creation implies a certain deafness to the appeal of other values, even going so far as to reject them if not denying them altogether. For one cannot fully enjoy the other, identify with him, and yet at the same time remain different. When integral communication with the other is achieved completely, it sooner or later spells doom for both his and my creativity. The great creative eras were those in which communication had become adequate for mutual stimulation by remote partners, yet was not so frequent or so rapid as to endanger the indispensable obstacles between individuals and groups or to reduce them to the point where overly facile exchanges might equalize and nullify their diversity.
>
> Lévi-Strauss, 1985(1983): 24

Of course, an oversimplified understanding of this point provided an opportunity for a few politically right-thinking people to feel some self-satisfaction. Lévi-Strauss's contribution is relevant when discussing cultural relativism, an issue that has gained importance lately. Lévi-Strauss clearly shows that all and each human being is never totally imprisoned within his or her own cultural bubble. By showing this, Lévi-Strauss denies the very possibility of cognitive relativism. Individually and collectively, you are bound to evaluate and compare. Not wanting to take a look outside or trying to render your bubble opaque implies values and comparisons. Thus,

cultural and cognitive relativism are in fact self-refuting. They are also almost antiscientific as they tend to render anthropological or historical work meaningless.

More generally, it is somewhat paradoxical that not so few intellectuals criticized Lévi-Strauss for things he neither did or said. Choosing to see in his structuralism a negation of the subject and in his critique of the excess of communication a new form of intolerance is misleading. His focusing on the principles at work in social systems does not imply an ignorance of the importance of individual reflection and engagement. Lévi-Strauss greatly contributed to ethics, politics, and tolerance. He seriously and sincerely dared to ask all the questions he thought ought to be asked in relation with the knowledge and challenges of his time. He criticized a form of philosophy (like that of Sartre) that "cuts itself off from scientific knowledge which it despises, as well as from human reality, whose historical perspectives and anthropological dimensions it disregards, in order to arrange a closed and private little world for itself" (Lévi-Strauss, 1981(1971): 640). According to him, too much of philosophy was still reflecting in ways that did not take into account the prodigious developments of science and the alarming development of global society. Still, as he turned his back on philosophy to dedicate himself to social anthropology, he ended up contributing to it. Of course many of his theories can be criticized. As an example, regarding his theory of the elementary structures of kinship, Lévi-Strauss indicated that the reason men ruled the exchanges of women was a natural one (which does not imply that culture cannot develop against this original stand). It is most likely that there were social reasons for this: women had to cope with the fact that men had the monopoly of violence. One could of course call this a natural difference. Nevertheless, out of this necessity of sexual collaboration and uneven distribution of strength, heavy tools, weapons, and violence, one can identify a principle other – and maybe even more fundamental – than that of the rules of kinship. In these sharing of roles and power that are most likely to have preceded symbolic language, one might find a fundamental principle at the very basis of all social life, and one that does not have to proceed directly from the structure of the mind: the principle of reciprocity, of sharing. Basic relations of exchange are also experienced in nonhuman societies. For example, the behavior of chimpanzees and bonobos are now mentioned as examples of how humans might gradually and simultaneously have developed their brain, language, relations, and tools resulting in the ability to postpone their actions and exchanges. Thus, as Godelier points out: "it is not from the unconscious that the principle of reciprocity arose out of, but from 'our way of being in the world'" (Godelier, 2013: 125). This implicit obligation of reciprocity, Godelier says, "is immanent to the fact that no one can ever exist without society existing and that no individual can ever develop her or his hidden potentials without society's permission" (Godelier, 2013: 125). Society provides or refuses the means.

And there are other relevant and constructive critiques that can and have been made of the work of Lévi-Strauss. Most of them somehow are in the same vein as his: trying to unveil human nature and the nature of human relations. What is most striking about Lévi-Strauss's theoretical models is the powerfulness of the footbridges and springboards they constituted. It is "his ability to theorise complex phenomena in such a way as to bring out their relevance for an understanding of other kinds of phenomena" (Wiseman, 2007: 17).

One of the main characteristics of Claude Lévi-Strauss was his ability to ask new questions and combine information in a new way. His work will continue to inspire many a scientist from many a discipline. Also, through his work, the almost forgotten wisdom of societies practically eradicated still provides us with ideas that might help us relate to ourselves and our planet.

> The human groups thus designated [as primitive] are for the most part those that differ from our own by virtues of the absence of writing and mechanization. But we must not forget a few primordial truths about them: these societies provide the only model for understanding how human beings lived together over a historical period undoubtedly corresponding to 99 percent of the total duration of the collective life of humanity. Their existence lasted until recent times (from a geological perspective) and covered three-quarters of the surface of the inhabited earth.
>
> What these societies offer, therefore, are not really lessons about the phases of our distant past. They illustrate rather a general situation, a common denominator of the human condition. Seen from that standpoint, it is the high civilizations of the West and East that constitute exceptions.
>
> Lévi-Strauss, 2013(2011) #24723: 13–14

"Lévi-Strauss explained the vital importance for all human beings – and particularly us – that cultural groups may carry on sending us the message that the historically dominant way of being human today is but one among many" (Lenclud, 2008: 26). He also showed that shared representations evolve and sometimes, for internal and cross-cultural reasons, people end up having to reorganize their system of thought. Today, we may have entered a state of crisis at the global level because of the accelerating pace of globalization.

Not so long ago these representations were not necessarily made explicit and were generally taken for granted: they manifested themselves in institutions that were accepted and made sense. Nowadays, as one commentator of Lévi-Strauss points out: "it is almost as though each cultural group has to constantly build and rebuild its own representations – usually in terms of communication – in order for its ongoing existence – one that is different from all other groups – to be recognised" (Lenclud, 2008: 21, translated from the French).

See also: Anthropology, History of; Anthropology: Overview; Boas, Franz (1858–1942); Durkheim, Emile (1858–1917); Jakobson, Roman (1896–1982); Mauss, Marcel (1872–1950); Symbolism in Anthropology.

Bibliography

Gibson, Kathleen R., Ingold, Tim (Eds.), 1995. Tools, Language and Cognition in Human Evolution. Cambridge University Press, Cambridge, UK.

Gibson, Kathleen R., 1996. The ontogeny and evolution of the brain, cognition, and language. In: Lock, A., Peters, C.R. (Eds.), Handbook of Human Symbolic Evolution. Clarendon Press, Oxford, pp. 407–431.

Godelier, Maurice, 2013. Lévi-Strauss. Seuil, Paris.

Hénaff, Marcel, 1998(1991). Claude Lévi-Strauss and the Making of Structural Anthropology (Mary Baker, Trans.). University of Minnesota Press, Minnepolis and London.

Langaney, André, 2000(1998). La Conquête Du Territoire. In: Langaney, A., Clottes, J., Guilaine, J., Simonnet, D. (Eds.), La Plus Belle Histoire De L'homme. Seuil, Paris, pp. 19–66 (Collection Points 779).

Lenclud, Gérard, 2008. Claude Lévi-Strauss Aujourd'hui. Études dy CEFRES (Centre français de recherche en sciences sociales). USR 3138 CNRS-MAEE N. 12. Novembre 2008: 1–27.

Lévi-Strauss, Claude, 1945. French sociology. In: Gurvitch, G., Moore, W.E. (Eds.), Twentieth Century Sociology. The Philosophical Library, New York, pp. 503–537.

Lévi-Strauss, Claude, 1958. Race and History. UNESCO, Paris.

Lévi-Strauss, Claude, 1961(1952). Race Et Histoire. Gonthier, Paris.

Lévi-Strauss, Claude, 1972(1958) (Claire Jacobson, Brooke Grundfest Schoepf, Trans.). Structural Anthropology, vol. 1. Penguin, London.

Lévi-Strauss, Claude, 1973. Structuralism and ecology. Social Science Information 12 (7), 7–23.

Lévi-Strauss, Claude, 1977(1949). The Elementary Structures of Kinship. Beacon Press, Boston.

Lévi-Strauss, Claude, 1978(1952). Race and history. In: Lévi-Strauss, C. (Ed.), Structural Anthropology, vol. 2. Penguin, Harmondsworth, pp. 323–362.

Lévi-Strauss, Claude, 1978(1973). Structural Anthropology, vol. 2. Penguin, Harmondsworth.

Lévi-Strauss, Claude, 1981(1971). The Naked Man (Translated from the French by John and Doreen Weightman). In: Introduction to a Science of Mythology, vol. 4. Jonathan Cape, London.

Lévi-Strauss, Claude, 1982(1975). The Way of the Masks. The University of Washington Press, Seattle.

Lévi-Strauss, Claude, 1985(1983). The View from Afar (Joachim Neugroschel, Phoebe Hoss, Trans.). Basic Books, New York.

Lévi-Strauss, Claude, 1986(1964). The Raw and the Cooked. Introduction to a Science of Mythology. Penguin Books, Harmondsworth, England.

Lévi-Strauss, Claude, 1987(1950). Introduction to the Work of Marcel Mauss (Felicity Baker, Trans.). Routledge and Kegan, London.

Lévi-Strauss, Claude, 2011(1955). Tristes Tropiques (John and Doreen Weightman, Trans.). Penguin, Harmondsworth.

Lévi-Strauss, Claude, 2013(2011). Anthropology Confronts the Problems of the Modern World (Jane Marie Todd, Trans.). The Belknap Press of Harvard University Press, Cambridge, Massachusetts and London, England.

Rousseau, Jean-Jacques, 1986(1754). On the Origin of Language. University of Chicago Press, Chicago.

Wiseman, Boris, 2007. Lévi-Strauss, Anthropology and Aesthetics. Cambridge University Press, Cambridge.

Lewin, Kurt (1890–1947)

Carl-Friedrich Graumann, University of Heidelberg, Heidelberg, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8737–8740, © 2001, Elsevier Ltd.

Abstract

First, the personal and scientific biography of Kurt Lewin is sketched, from his early education in Posen and Berlin, and two semesters (of medicine and biology) in Freiburg and Munich, until the time when Berlin and its university became Lewin's permanent residence, and, where interrupted by the war of 1914 to 1918, he embarked on the first stage of his career, which soon gained international recognition. This recognition turned out to be vital when in 1933 Lewin, as a Jewish scientist, had to leave Germany and found asylum in the USA until his early death in 1947. The academic stations of his US career were Cornell, Iowa, and the MIT. In conclusion, this article reports and assesses Lewin's major research topics, the development of his field theory, and his innovative research style.

Biography

Kurt Lewin was born on 9 September 1890 in Mogilno, a small town in the (then Prussian) province of Posen, now Poland. He was the second of the three children of Leopold and Recha Lewin who owned a general store and a small farm. Kurt was sent to a *Gymnasium*, a secondary school, in Posen until, in 1905, the whole family moved to Berlin, as did many ascending Jewish families living in the eastern provinces of the Reich and in Eastern Europe.

In Berlin, Kurt Lewin attended the humanistic *Kaiserin-Augusta-Gymnasium*. In 1909 he took up university studies of medicine and biology in Freiburg and Munich, but in 1910 he was back in Berlin to take courses in philosophy and science. His special interest soon became philosophy of science, in which field the (neo-Kantian) philosopher Ernst Cassirer and the psychologist Carl Stumpf were most influential with regard to Lewin's intellectual development. Stumpf, the director of the Psychological Institute and the patron of what was to become the Berlin school of gestalt psychology (*see* Köhler, Wolfgang (1887–1967)), became Lewin's doctoral supervisor.

Since in 1914 Lewin volunteered for the army and soon did service on the Western Front, he was awarded his doctoral degree as a soldier in 1916. While on furlough from the front in 1917, he married Maria Landsberg, a fellow student from prewar days. The same year, 1917, was also the year of Lewin's first two publications, which, in their dissimilarity, indicate the scope of Lewin's interests: a paper on mental activities in the inhibition of processes of volition, based on his doctoral thesis, and an almost phenomenological paper on the 'war landscape,' foreshadowing Lewin's later involvement in ecological problems.

Also predictive for Lewin's further intellectual as well as social development were his extracurricular activities and associations as a student, as from early on he had engaged himself in socialist student groups which took an active interest in the social and, mainly, educational problems of workers in agriculture and industry. From this enduring interest resulted two early publications in 'applied psychology,' one on rationalization in agriculture, one on the socialization of the Taylor system. Many later publications followed as a result of his and other Jewish psychologists' forced emigration to the USA.

Before the war ended Lewin was severely wounded and hospitalized. Decorated with the Iron Cross and promoted to the rank of lieutenant, he was discharged from military service.

In the following years, Lewin, who had returned to the Psychological Institute of Berlin University, and received his *Habilitation* in 1921, became assistant in the new department for applied psychology, a position which he held until his emigration in 1933. In 1927 he was appointed *außerordentlicher Professor* – a title, not a tenured position.

Although in close spatial and intellectual proximity with Köhler's group of gestalt psychologists, Lewin, who concentrated on studies of will, affect, and action, soon gathered a group of young talented students around him, 'whose background was more often than not female, Jewish, and eastern European' (Danziger, 1990: p. 173). Some of Lewin's Berlin women students gained international recognition by their association with Lewin, and some shared his later fate as refugees. But it is fair to say that Lewin's rapidly growing scientific reputation was also due to the works of his doctoral students, which Lewin represented at conventions and in journals, preferably in his own series in the journal *Psychologische Forschung*.

An important date in Lewin's vita was 1929. In this year Lewin, who had separated and divorced from his wife Maria, remarried. His second wife, Gertrud, had been a friend of the family for several years. Professionally, 1929 was the year for which he received an invitation to attend the ninth International Congress of Psychology at Yale. This was shortly after J.F. Brown, an American who had studied with Lewin, had published a paper on 'The methods of Professor Lewin' in the 1929 *Psychological Review*, which had attracted the attention of US psychologists to the work of Lewin's Berlin group. In Yale Lewin impressed his audience with a short paper on environmental forces (read in German) and an equally short, but self-made motion picture of an 18-month-old girl trying to sit on a stone for the first time in her life – one of the many films that Lewin found useful for the graphic representation of the life space (see below). According to G.W. Allport, who had witnessed Lewin's performance in Yale, 'This ingenious film was decisive in forcing some American psychologists to revise their own theories of the nature of intelligent behavior and of learning' (quoted in Marrow, 1969: pp. 50f.). It was his first

encounter with a US audience that resulted in Lewin being invited as visiting professor by Stanford University in 1932.

It turned out that after this stay in Stanford Lewin would only briefly return to Berlin, as in January 1933 the Nazis seized power in Germany. On his return trip from California via Japan and the USSR Lewin decided that he and his family could not live in Nazi Germany. Fortunately, shortly after his return he received a cable informing him that Cornell was inviting him to join the faculty staff, which he gladly accepted – just in time.

From 1933 to 1935 Lewin then worked at the Cornell School of Home Economics. In 1935 he was offered a position at the Iowa University Child Welfare Research Station. For 8 years he worked in Iowa as a professor of child psychology. It was in these years that Lewin taught excellent students who became his disciples and, later, colleagues; for example, Barker, Bavelas, Cartwright, Festinger, French, Lippitt, White, and Zander. It was also during this period that Lewin's research interests shifted toward social psychological problems. In 1937 he became one of the founders of the Society for the Psychological Study of Social Issues, whose president he became in 1941.

This social engagement reflects Lewin's increasing concern with the interaction between research and practice which led to his conception of 'action research,' and to the idea and foundation of the Research Center for Group Dynamics at MIT in 1945, whose main research task was to be 'the development of scientific methods studying and changing group life and the development of concepts and theories of Group Dynamics' (quoted in Patnoe, 1988: p. 9).

This institution framed the final period of Lewin's life. In 1947, in the middle of a rather hectic spell of activity, he died of a heart attack, on February 12 in Boston, at the age of 56. In the same year the Research Center moved to the University of Michigan, Ann Arbor, where it still exists as a prominent place of social research.

Research Topics and Fields

Social interaction, mainly in small face-to-face groups, was such a driving force in Lewin's professional life that it is difficult, often impossible, to single out and identify his strictly individual contribution to what the Berlin or Cornell or Iowa or MIT group produced. Conversely, psychological knowledge today owes a great deal to names such as Barker, Cartwright, Dembo, Festinger, French, Hoppe, Karsten, Kelley, Lippitt, Ovsiankina, Pepitone, Thibaut, Zeigarnik (an alphabetical, hence arbitrary, order of students, and colleagues who worked with Lewin). From their oeuvre, whether published individually or as coauthors, whether during Lewin's lifetime or thereafter, it would be difficult to do an auctorial analysis of variance. But it would be futile, too, since Lewin, at least, believed firmly in the creativity of dialogical thinking. Hence, Lewin is not only an author's name, it is also a brand name for products of the interaction between Kurt Lewin and his associates.

If one is looking for a name designating the unity of products from Lewin's groups the choice, after all, would be 'field theory,' which was also Lewin's final preference although for many years the favorite was 'topology.' It has been stated frequently but is useful to repeat that 'field theory is probably best characterized as a method: namely a method of analyzing causal relations and of building scientific constructs' (Lewin, 1997).

The conceptual tools of this metatheory and methodology (rather than theory and method) are found mainly in two monographs (Lewin, 1936, 1938) and two anthologies (Lewin, 1997, 1999). Lewin adopted the philosophical foundation of Ernst Cassirer's (1953) distinction between 'substance' and 'function,' his theory of a historical transition from 'thing-concepts' to 'relation- (or function-) concepts' and from an abstraction- to a construction-mode of concept formation. Lewin applied this conception to psychology in his 1931 paper on the transition from an Aristotelian to a Galileian mode of thought (in Lewin, 1935). For him psychology deals with relations or functions (rather than thing-like entities) which are basically dynamic, i.e., forces.

In accordance with gestalt theoretical principles, every individual psychological instance (experience or behavior) is the result of an interaction of several features of a given whole or situation; ultimately, each behavior is determined by the total situation in which it occurs. For Lewin the total situation was the 'life space' organized by interdependent forces. In other words, the life space is the 'totality of facts which determine the behavior (B) of an individual at a certain moment. The life space (L) represents the totality of possible events. The life space includes the person (P) and the environment (E). '$B = f(L) = f(P, E)$' (Lewin, 1936: p. 216).

It is this formula, which binds person and environment inextricably together, which has become a kind of shorthand for field theory: It is important to note that owing to the principle of interdependence, neither person nor environment is granted independence. The field, which is structured into regions, paths, and barriers, is basically a field of forces (tendencies to change or to resist change). Some regions are attractive to the person (they have a positive 'valence'). Others are repulsive (they have a negative 'valence'). The person is also considered to be a differentiated region with similar structural and dynamic properties.

These formal characteristics of the life space may be sufficient to explain why Lewinian psychology, whether of learning, of memory, of decision, of action, or of social interaction and relations, is intrinsically dynamic and deals with tension systems (as in anger, aggression, regression, substitution, flight into irreality, etc.), with forces affecting the level of aspiration, conflict, decisions, pressure, resistance, and with styles of leadership or the social-emotional atmosphere of small groups. It was mainly the latter type of studies, their designation (of 'authoritarian' vs 'democratic' vs 'laissez-faire' styles of leadership), and, last but not least, the fact that they were published in the early war years, that attracted wide and lively public interest, curiosity, and critique. From these studies, in which things 'happened' that were not designed to happen (such as 'laissez-faire' in lieu of a 'democratic' style), moving toward the use of research for policies of social change, what Lewin has called 'action research,' was but a small, though consequent and innovative step. It was final confirmation for the interdependence that Lewin had postulated between theory and practice, basic research and applied science.

Assessment

A critical assessment of what in the first five decades after Lewin's death has variously been called the Lewin tradition, legacy, or heritage is difficult because this is a multidimensional and multiperspectival task.

It is multidimensional because of the many research fields in which Lewin was engaged, beginning with the general experimental psychology of learning, memory, affect, motivation, and action, through child and personality psychology, to social, mainly small-group psychology, education, and social change, as well as, from the start, metatheory and methodology. It is multiperspectival since, depending on which experts or resources are consulted at any given time, Lewin's influence on present psychology will be differently evaluated.

It may be easy to neglect the temporal or epochal factor. While, in the first two decades after his death, Lewin's or 'Lewinian' theory could rightly be named a 'contemporary systematic framework,' and 'field theory' occupied sizeable chapters in handbooks of social psychology, this situation had begun to change in the 1980s to such an extent that quite a few of the centennial retrospectives due in 1990 were celebrations with reservations (cf Graumann, 1992; Schönpflug, 1992; Wheelan et al., 1990). Field theory was no longer one of the grand theories of (social) psychology (if indeed there were any great ones left). Topology and vector psychology were going out of fashion. Progress in mathematical psychology no longer included topology and hodology as Lewin (1936) had envisaged and practiced them. If Lewin identified himself and has been identified with topology and vector psychology, his significance for contemporary psychology has noticeably decreased. None of his influential disciples, who all remember Lewin drawing 'bathtubs and eggs' in class, has seriously pursued topology as a theoretical and methodological system. But, as outlined above, many of the 'psychological concepts' listed in Lewin's 'Topology' of 1936 have not only survived but become common psychological knowledge.

What is more important for Lewin's heritage and his presence in contemporary psychology is the diversity with which his *oeuvre* has been handed down by his disciples and followers. To name but two traditions which may both rightly be called Lewinian, but in fact have little in common, there is, on the one hand, the prominent group of experimental social psychologists, going back to Lewin's Iowa period (1935–44) and the beginnings of the MIT Research Center for Group Dynamics through Lewin's death in 1947. On the other hand, there is the tradition in which group dynamics and action research have been seen essentially as means of changing human relations, individuals, groups, and societies. The journal *Human Relations,* founded by Lewin in cooperation with the Tavistock Clinic in London, has become an outlet for publications on applied group dynamics from which so-called 'sensitivity training' and resulting 'encounter groups' have gained the broadest notoriety by growing into a movement (cf Back, 1973).

But even if one stays within Lewinian experimental social psychology, the variance of derivations from Lewin's original ideas is considerable. If one looks for the common element, most of Lewin's senior disciples testify that it was Lewin's 'style' and the 'atmosphere' he created around him – a thoroughly Lewinian remark since 'Lewinian style included an appreciation and sensitivity to context – the interplay of person with environment' (Patnoe, 1988: p. 88; cf Graumann, 1992). What Patnoe (1988: p. 15) concluded from the testimonies of Lewin's contemporaries, namely that 'Lewin created his own microculture,' has been taken by historiographers of psychology as an enduring characteristic of Lewin's research style. Mitchell Ash (1995), who as a historian of gestalt psychology (and of other èmigré psychologists after 1933) is interested in the role of milieu changes, states that Lewin, who had already established 'a scientific microculture around him' in Berlin (Ash in Schönpflug, 1992: p. 20), accomplished the same in Iowa. The (sometimes very intense) interpersonal interaction in Lewin's research group was not just different from the methods of the (other) gestalt theorists (Ash, 1995: p. 272). For Kurt Danziger, who has specialized in the history of psychological research, Lewin developed an investigative practice that 'departed fundamentally from the natural scientific model of experimentation' (Danziger, 1990: p. 175). Considering that Lewin's Berlin experiments were not in social but in personality psychology, it is a striking innovation that even here (and in principle) the 'investigative situation is recognized as intrinsically social, and in research on personality it must form part of the object of investigation' (Danziger, 1990: p. 175).

In anticipation of and in accordance with Lewin's later theorizing, the experimental subject is not a decontextualized individual, but a person-in-a-situation, and an important interacting part of the situation is the person of the experimenter. It is here that the Galileian thought and postulate which gave birth to field theory have become instrumental in an innovative, psychologically autonomous research practice – a legacy of Lewin's that psychologists still have to appropriate.

See also: Behavior Analysis, Applied; Contextual Studies: Methodology; Heider, Fritz (1896–1988); Interactionism, Symbolic; Köhler, Wolfgang (1887–1967); Psychology, History of (Early Period); Psychology, History of (Twentieth Century); Social Psychological Theory, History of; Social Psychology: Research Methods; Social Psychology.

Bibliography

Ash, M., 1995. Gestalt Psychology in German Culture, 1890–1969: Holism and the Quest for Objectivity. Cambridge University Press, Cambridge, UK.

Back, K.W., 1973. Beyond Words: The Story of Sensitivity Training and the Encounter Movement. Penguin, Baltimore, MD.

Cassirer, E., 1953. Substance and Function. Dover, New, York.

Danziger, K., 1990. Constructing the Subject. Historical Origins of Psychological Research. Cambridge University Press, Cambridge, UK.

Graumann, C.F., 1992. Lewin 1990. In: Schönpflug, W. (Ed.), Kurt Lewin. Leben, Werk und Umfeld. Lang, Frankfurt, Germany, pp. 223–230.

Lewin, K., 1935. A Dynamic Theory of Personality, first ed. McGraw-Hill, New York.

Lewin, K., 1936. Principles of Topological Psychology, first ed. McGraw-Hill, New York.

Lewin, K., 1938. The Conceptual Representation and the Measurement of Psychological Forces. Duke University Press, Durham, NC.

Lewin, K., 1981. In: Graumann, C.F. (Ed.), Kurt Lewin Werkausgabe. Huber, Switzerland.
Lewin, K., 1997. Resolving Social Conflicts and Field Theory in Social Science. American Psychological Association, Washington, DC.
Lewin, K., 1999. In: Gold, M. (Ed.), The Complete Social Scientist, first ed. American Psychological Association, Washington, DC, pp. 321–326.
Marrow, A.J., 1969. The Practical Theorist. The Life and the Work of Kurt Lewin. Basic Books, New York.
Patnoe, S., 1988. A Narrative History of Experimental Social Psychology: The Lewin Tradition. Springer Verlag, New York.
Schönpflug, W. (Ed.), 1992. Kurt Lewin. Leben, Werk und Umfeld. Lang, Frankfurt, Germany.
Stivers, E., Wheelan, S.A. (Eds.), 1986. The Lewin Legacy-Field Theory in Current Practice. Springer Verlag, Berlin.
Wheelan, S.A., Pepitone, E.A., Abt, V. (Eds.), 1990. Advances in Field Theory. Sage, London.

Lex Mercatoria

Christine Windbichler, Humboldt Universität zu Berlin, Berlin, Germany

Published by Elsevier Ltd.

Abstract

The Latin term *lex mercatoria* denotes transnational self-regulation among merchants at least as early as the Middle Ages. Legal scholars disagree about whether historical *lex mercatoria* constituted a legal system. Nowadays, transnational private governance is practiced and expanded. Modern formulating agencies like International Institute for the Unification of Private Law and UN Commission on International Trade Law work out systematic restatements, principles, and internationally standardized trade terms to further worldwide commercial transactions. Governmental and nongovernmental organizations formulate the so-called soft law, that is, guidelines and best practices, to bring about international norms in many areas where no global legislation is available. Historically and currently, parties in international business disputes avoid state courts and prefer arbitration. Arbitration tribunals apply *lex mercatoria*, the theoretical dispute about the qualification of the rules as law notwithstanding. In a global context, the private, self-governing *lex mercatoria* approach is indispensable. Legal scholars and social scientists today are exploring the legitimacy of rules developed by private organizations and the interplay of law and private ordering.

Lex mercatoria, often translated as 'law merchant,' does not really fall within the scope of law in the sense of a government-based legal system. The origins of these kinds of rules are private, autonomous, international, a-national, or transnational. *Lex mercatoria* is not devised by state courts or a state legislative body. Its practical impact, however, has been undisputed since ancient Roman times. Modern *lex mercatoria* is discussed under the auspices of continuing globalization and post-national law.

Historical Development and Current Definitions

In Roman law, *lex mercatoria* was part of *ius gentium* (law of nations), that is, a kind of international law more flexible than the rather rigid Roman *ius civile* (civil law) (*see* Civil Law; International Law and Treaties). *Lex mercatoria* encompassed customary law in international trade. Historically, it is the major root of the consensual contract. However, it is disputed whether Roman *lex mercatoria* as part of *ius gentium* offers a precedent for medieval *lex mercatoria* as a legal system apart from national legal systems (De Ly, 1992).

The medieval *ius commune* was based on Roman *ius civile*. For centuries it had provided a legal framework generally applied in Europe. Legal scholars cultivated a common understanding based on Justinian's *Corpus Iuris*, despite the division of medieval Europe into many more distinct political entities than today, all with their various local laws. In the same vein, medieval *lex mercatoria*, a cosmopolitan mercantile law, developed from transnational customs and practices. For example, merchants copied deeds used for similar transactions. Disputes were settled by market tribunals or special national courts like the English admiralty courts. The English *Carta Mercatoria* of 1303 provided certain foreign merchants with a fast dispute settlement mechanism and application of law merchant (as opposed to feudal law or city law; De Ly, 1992). The English admiralty courts relied on the best sources available, be they domestic or foreign, in elaborating a worldwide maritime law as a subset of law merchant (Juenger, 1995).

In the seventeenth century, English Common Law courts began to apply law merchant by incorporating customary rules developed by traders and used in arbitration or special courts. 'Law merchant' was defined as

> a number of usages, each of which exist among merchants and persons engaged in mercantile transactions, not only in one particular country, but throughout the civilized world, and each of which has acquired notoriety, not only amongst those persons, but also in the mercantile world at large, that courts of this country will take judicial notice of it. (Lethulier's Case (1692) 2 Salk 443 (per Holt, CJ))

This merchant law was less rigid and less formal than the common law but accepted by common law courts, which consequently developed into a forum for commercial litigation. Merchant law with its informal and international roots became common law. The change of label gradually also took away its international character; merchant law became national commercial law. Later codifications of commercial law, for example, the English Sale of Goods Act of 1893 and the American Uniform Commercial Code, are to a large extent restatements of law merchant principles as represented in common law. Today, the term 'merchant law' is often used to mean 'national common law relating to commercial transactions.'

Modern usage understands *lex mercatoria* as an arrangement of standardized clauses and general legal principles; for example, the principle of good faith, applying to international trade and commerce. To set it off from law merchant incorporated in national common law, some authors prefer 'new' or 'modern *lex mercatoria*' or '*new law merchant*' and similar terminology when referring to modern (informal) codifications of transnational commercial law (Hellwege, 2012).

A number of intergovernmental and nongovernmental organizations developed standardized clauses and rules based on extensive comparative studies. Often enough, these rules served as a model for national lawmaking and international conventions. An outstanding example of a formulating agency

is the International Institute for the Unification of Private Law (*Institut international pour l'unification du droit privé*, UNIDROIT), set up in 1926 as an auxiliary organ of the League of Nations. UNIDROIT is now an independent intergovernmental organization with 63 member states. Its purpose is to study the needs and methods for modernizing, harmonizing, and coordinating private and, in particular, commercial law and to formulate uniform legal instruments, principles, and rules. Many of UNIDROIT's drafts have resulted in international conventions, such as the Convention on a Uniform Law on the Formation of Contracts for the International Sale of Goods, which was adopted at a diplomatic conference of the member states in The Hague in 1964. The UN Commission on International Trade Law (UNCITRAL), formed by the UN General Assembly in 1966, also has contract law topics on its agenda. A prominent example is the Vienna Convention on the International Sale of Goods. The International Chamber of Commerce (ICC), founded in 1919 and now represented in 130 countries, provides INCOTERMS® (International Commercial Terms), an important set of clauses in worldwide use last updated in 2010.

Lex mercatoria also refers to unwritten or, at least, not systematically collected principles and clauses that reflect international trade usage or custom (Goode, 1997). Such rules are applied by arbitration tribunals; for example, the ICC International Court of Arbitration, established in 1923. The ICC publishes arbitral awards but does not disclose parties and other sensitive details. Legal practitioners as well as scholars compile and discuss modern *lex mercatoria* based on such arbitral awards. Such reasoned compilations gradually replace unwritten custom. UNIDROIT has prepared a well-researched collection of contract law used in international trade, the *UNIDROIT Principles of International Commercial Contracts*, first published in 1994. These principles intend to spell out rules that are common to many existing legal systems and accommodate the requirements of international trade. The UNIDROIT Principles do not claim to have binding force. They are intended to be applied in practice only by means of their persuasive value. Similarly, in many other fields of law, general principles, guidelines, or best practices are set forth by professional bodies, trade associations, interest groups, legal scholars, and governmental or intergovernmental agencies. Examples are International Financial Reporting Standards (IFRS) of the International Accounting Standards Board, an institution of the independent, not-for-profit private sector IFRS-Foundation. In some instances, like the UNIDROIT project, the objective is to establish an unbiased and well-balanced restatement; in other instances, a political or manipulating impact is intended.

Main Characteristics and Practical Impact of Modern *Lex Mercatoria*

Based on historic development and present-day definitions, the modern *lex mercatoria* can be characterized as transnational, a-national, or even post-national. In the absence of a 'world legislator,' international trade and commerce have developed functional rules based on common practice. Modern ideas of law from the beginning of the nineteenth century assumed that all law must emanate from the sovereign; therefore, the concept of an independent *lex mercatoria* declined (Dalhuisen, 2007). However, parties to a transnational contract tend to avoid national courts and domestic laws that would present imponderables in procedure and substance. In a globalized world, internationally active professionals develop and rely on well-established commercial clauses and norms that enable contracting parties to overcome language barriers and biases. Business people and their lawyers draw from the condensed and standardized experience in their special fields of business. Such standardized usage can be summarized as modern *lex mercatoria*. However, parties or their counsel writing a transnational contract rarely invoke *lex mercatoria* as such. They rather make use of specific clauses and principles. Therefore, the term '*lex mercatoria*' is not actually used in international trade but rather denotes an abstraction.

Usually, agreements of arbitration or other forms of dispute resolution that avoid state courts include references to *lex mercatoria* clauses and principles. Arbitration rules themselves are often part of *lex mercatoria*, too, such as UNCITRAL's, the World Bank International Center for the Settlement of Investment Disputes', or ICC's rules on international commercial arbitration. In the case of *Scherk v. Alberto Culver Co* (417 U.S. 506), the US Supreme Court pointed out the purpose of such arbitration agreements:

> A contractual provision specifying in advance the form in which disputes shall be litigated and the law to be applied is … an almost indispensable pre-condition to achievement of the orderliness and predictability essential to any international transaction. Furthermore, such a provision obviates the danger that a dispute under the agreement might be submitted to a forum hostile to the interests of one of the parties or unfamiliar with the problem area involved.

This case from 1974 is in line with the UN Convention on the Recognition and Enforcement of Foreign Arbitral Awards from 1958, which has been ratified by 148 countries.

Theoretical Analysis

In legal theory, it is disputed whether *lex mercatoria* is 'law.' Proponents of the theory of *lex mercatoria* as an independent body of law (mainly French, English, Dutch, and German scholars) argue on the basis of a functional or sociological concept of law (Stein, 1995). In this line of thinking, an autonomous body of law governing international commerce can exist without a source in state power: "*Il existe un droit dont la source est ailleurs que dans l'Etat*" (Bonell, 1997; Carbonneau, 1990; also Dalhuisen, 2007; Dezalay, 1995; Kahn, 1961; Osman, 1992). However, it would be considered a third category between national and international law.

Other theorists, especially legal positivists, who define law as based on the power of the (nation) state, deny the existence of an autonomous *lex mercatoria*. They argue that the lack of judicial enforceability cannot be substituted by private autonomy, as even contracts need national law to be put into effect. Constitutionalism, positivism, and the concept of the sovereign state are irreconcilable with rule making by private

governance. Additionally, a contract is only conceivable within a legal order; there is no *contrat sans loi* (contract without law).

Furthermore, it is argued that *lex mercatoria* is much too vague, difficult to ascertain, and incomplete to qualify as a legal system. A more recent academic dispute challenges the notion that medieval *lex mercatoria* existed as a uniformly and universally adopted system (Kaden, 2012; Michaels, 2007, 2012). *Lex mercatoria* is called a 'myth,' its proponents 'medievalists' or 'mercatorists'. The larger context of this controversy is the debate about whether private ordering in a globalized world can work independently from government involvement. Hence, the issue of whether medieval *lex mercatoria* was a legal system turns mostly on one's definition of 'system,' which is likely to be subject to sociological overtones (Teubner, 1997; *see* 03183; Law, Autopoiesis in; Legal Systems, Classification of; Luhmann, Niklas (1927–98)).

This academic dispute – whether medieval law merchant is the archetypal sophisticated legal system that private groups can create when not impeded by the intermeddling of the state – is of only limited practical impact. No theoretical question arises when model laws or other products of formulating agencies have been incorporated into national law. Courts apply the resulting statutes, and parties to a contract can refer to such statutes in choice of law clauses. In other cases, when a state is not a signatory or the respective set of clauses is a private convention, courts test the validity of *lex mercatoria* clauses as standard business conditions according to national law. Any international recognition of such clauses gives them a certain dignity and adds to their acceptance. Choice of law clauses that call for the application of *lex mercatoria* as such are rare; the objection of vagueness would then be well argued. Parties to international contracts tend to be specific. Sets of clauses develop and find their way into domestic contracts, too. Standardized private arrangements even dominate financial instruments, especially derivatives, in the financial markets. Terms of stock options, for instance, follow distinct patterns as 'European' or 'American' depending on the dates on which the options may be exercised. Such patterns are designed by custom and practice, not by law.

Arbitration clauses are prevalent in international trade (*see* International Arbitration). An arbitration tribunal is free to apply *lex mercatoria* principles and reach its decision on the equities of the case. The validity of arbitration clauses is generally recognized by state courts, especially if the clause inserted is in accordance with the "Rules of Conciliation and Arbitration" of the ICC, the UNCITRAL rules on international commercial arbitration, or a similarly established procedure (Lando, 1994; Lehmann, 2004).

A different perspective focuses specifically on the nonbinding nature of certain norms. The so-called soft law, such as guidelines, recommendations, codes of conduct, best practices, and so forth, has gained ground in the last few decades. The practical impact and even legal relevance of nonbinding principles can be considerable. The OECD Guidelines for Multinational Enterprises, for example, provide voluntary principles and standards for responsible business conduct consistent with applicable laws (Horn, 1980). Similarly, the OECD Principles of Corporate Governance assist member and nonmember governments in their efforts to evaluate and improve the legal, institutional, and regulatory framework for corporate governance in their countries and to offer guidance and suggestions for stock exchanges, investors, corporations, and other parties that have a role in the process of developing good corporate governance.

Governments and international institutions may discuss sensitive policy issues more freely and have fewer reservations when they enter into agreements that are nonbinding (Dryden, 1999). From the perspective of private parties, guidelines, recommendations, and best practices leave private autonomy formally untouched despite their persuasive authority. If not properly explained, deviation from widely recognized standards can result in loss of reputation with respect to business partners or investors. Compliance with soft law can serve as safe harbor, such as by helping a defendant fend off a claim of negligence. The way to deal with accepted best practices is to 'comply or explain.' Some binding laws honor nonbinding codes and require a declaration of compliance or reasoned noncompliance (examples can be found in the US-American Sarbanes–Oxley-Act, 2002, a law reacting to numerous financial scandals). Noncompliance with informal standards does not lead to legal sanctions but noncompliance with disclosure rules possibly will. Construction of technical legal concepts as, for instance, the business judgment rule, may well draw upon best practices.

Perspectives

Globalization of commerce, information, and other areas of life brings about challenges similar to what ancient Roman merchants and English tradesmen faced and met by establishing *lex mercatoria*. In global commerce, the realities of the borderless world of cyberspace and open financial markets run headlong into geographically delimited national jurisdictions of sovereign states. The attainments of constitutionalism, the nation-state, and the rule of law are limited in a global market. For example, in the environment of electronic communication, private regulation is gradually absorbed or superseded by national law and international conventions. In sports, too, world championships and similar events follow elaborate procedures on the basis of rules that are not law in a formal sense. A worldwide network of sports associations has undisputed jurisdiction over the rules of the game. 'Private ordering' is gaining ground much in the same way *lex mercatoria* established itself as a legal, commercial, or sociological reality. Given the extreme difficulties of global regulation, effective market-driven, self-regulatory mechanisms, assisted by formulating agencies, are possible and desirable. State-based legislation and self-regulation are not incompatible but complementary (Dalhuisen, 2007; Dryden, 1999).

The controversy over whether medieval *lex mercatoria* constituted a system, that is, was internally uniform and externally autonomous, or rather an archipelago of well-established rules in a fluid sea of individual contracts and customs, is theoretical and should not divert us from more pressing issues. One issue is the legitimacy of a rule that is not conferred by contract when it comes to third-party

interests; another is undue influence by biased formulating agencies (Bachmann, 2006). Further research is also needed in the interpenetration of private ordering, binding, and nonbinding law. The adoption of commercial clauses or rules as nonmandatory state law and disclosure requirements for nonbinding practices are examples of such complements.

The history of *lex mercatoria* shows that transnational, nongovernmental regimes can be functional. Modern *lex mercatoria* harmonizes specific fields of law, sets standards in the form of codes of conduct, and provides self-regulation in the absence of a world legislator.

See also: Administrative Law: The United States and Beyond; Business Law; Common Law; Contracts: Legal Perspectives; Functionalism, History of; Globalization: Legal Aspects; International Arbitration; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; Law and Development; Law, Autopoiesis In; Law, Sociology of; Laws, Conflict of; Legal Positivism; Legal Systems, Classification of; Legal Systems: Private; Luhmann, Niklas (1927–98); Organizations and the Law.

Bibliography

Bachmann, G., 2006. Private Ordnung: Grundlagen Ziviler Regelsetzung. Mohr Siebeck, Tübingen, Germany.

Bonell, M.J., 1997. An International Restatement of Contract Law, second ed. Transitional Publishers, Irvington-on-Hudson, NY.

Carbonneau, T. (Ed.), 1990. Lex Mercatoria and Arbitration. Transnational Juris Publications, Dobbs Ferry, NY.

Dalhuisen, J., 2007. Transnational and Comparative Commercial, Financial and Trade Law. Hart Publishing, Oxford and Portland, OR.

De Ly, F., 1992. International Business Law and Lex Mercatoria. North-Holland, Amsterdam.

Dezalay, B., 1995. Merchants of law as moral entrepreneurs. Law & Society Review 29, 12–27.

Dryden, J., 1999. The Work of the OECD on Electronic Commerce. r0.unctad.org/ecommerce/event_docs/colombo_oecd.pdf.

Goode, 1997. Usage and its reception in transnational commercial law. International and Comparative Law Quarterly 46, 1.

Hellwege, P., 2012. Lex mercatoria. In: Basedow, J., Hopt, K.J., Zimmermann, R. (Eds.), Max Planck Encyclopedia of European Private Law, vol. II. Oxford University Press, Oxford, UK, pp. 1086–1089.

Horn, N., 1980. Codes of conduct for MNEs and transnational Lex mercatoria: an international process of learning and law making. In: Horn, N. (Ed.), Legal Problems of Codes of Conduct for Multinational Enterprises. Kluwer, Deventer, The Netherlands. IASB-material. http://www.ifrs.org. ICC-material. http://www.iccwbo.org.

Juenger, 1995. American conflicts scholarship and the new law merchant. Vanderbilt Journal of Transnational Law 28, 487.

Kaden, E., 2012. The myth of the customary law merchant. Texas Law Review 90, 1153.

Kahn, P., 1961. La Vente Commerciale Internationale. Paris.

Lando, O., 1994. Assessing the role of the UNIDROIT principles in the harmonization of arbitration law. Tulane Journal of International and Comparative Law 3, 129.

Lehmann, M., 2004. A Plea for a transnational approach to arbitrability in arbitral practice. Columbia Journal of Transnational Law 42, 753.

Michaels, R., 2007. The true lex mercatoria: law beyond the state. Indiana Journal of Global Legal Studies 14, 447.

Michaels, R., 2012. Legal mediavalism in lex mercatoria scholarship. Texas Law Review 90, 259.

Osman, F., 1992. Les principes généraux de la lex mercatoria. Contribution à; l'étude d'un ordre juridique anational. Libr. générale de droit et de jurisprudence, Paris.

Stein, U., 1995. Lex Mercatoria: Realität und Theorie. V. Klostermann, Frankfurt am Main, Germany.

Teubner, G. (Ed.), 1997. Global Law without a State. Dartmouth, Aldershot, UK. UNIDROIT-material. http://www.unidroit.org.

Lexical Consolidation

Jakke Tamminen, Royal Holloway, University of London, Egham, Surrey, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

An extensive body of evidence has emerged over the recent years suggesting that representations of newly learned words continue to develop over the hours and days following learning. This process of lexical consolidation is particularly critical for the integration of new words in the mental lexicon. Work using neuroimaging and sleep methods suggests that this is due to the operation of complementary learning systems in the medial temporal lobe and the neocortex, with the former rapidly generating discrete episodic representations, and the latter slowly integrating the new words with existing information during off-line periods such as sleep.

Adding a new word into your existing store of words, the 'mental lexicon,' seems intuitively like a straightforward task. You need to find out what the new word sounds like, how it is spelled, and what it means. Once you have these pieces of information, you are ready to use the word and to understand it when it is encountered in print or in speech. However, a systematic research effort over the past decade or so has begun to map the many cognitive operations involved in word learning and has revealed some surprising findings. One of the major realizations has been that it is not sufficient to ask what you know about a word. It is also necessary to ask whether the new word has been connected with other components of the language comprehension and production system, as we know familiar words are. This distinction was lucidly made by Leach and Samuel (2007) who termed these processes 'lexical configuration' and 'lexical engagement.' Lexical configuration refers to knowledge of the factual information about a word, such as its spelling, phonology, and meaning. Lexical engagement on the other hand refers to the interaction of the word with respect to other lexical or sublexical units, and measures the degree to which a word has been integrated in the mental lexicon. Recent research into adult word learning has shown that our intuition about word learning being fast and relatively straightforward is more or less accurate relative to lexical configuration: people can indeed memorize new words quickly even with little training. Interestingly, at the same time our intuition is wrong: evidence of lexical engagement is most often seen only when the learner is tested some hours or days after training. It is hypothesized that this delay in the emergence of lexical engagement reflects off-line memory consolidation, a process occurring over time in the absence of further exposure to the new stimuli during which newly acquired memories gradually become stronger, integrated with existing knowledge, and resistant to interference. Here I will describe evidence for lexical consolidation in word learning. I will then outline a theoretical view of the cognitive and neural mechanisms that mediate lexical consolidation, based on the influential complementary learning systems (CLS) accounts of memory (McClelland et al., 1995; Davis and Gaskell, 2009). It is important to note that most of the studies that address the issue of consolidation in word learning come from adult populations, but the emerging evidence suggesting that the same principles apply in children will also be reviewed.

Behavioral Evidence for Lexical Consolidation

Consolidation appears to be critical for lexical engagement. One example of lexical engagement is seen during the process of recognizing a spoken word. As the spoken signal gradually unfolds, listeners entertain several hypotheses about the identity of the word (Marslen-Wilson, 1987). Hearing the first few sounds of a word will activate a cohort of competing word candidates matching the incoming signal. As the ensuing lexical competition continues and as more of the signal is heard, the number of matching candidates is reduced until only one remains. The point at which only one word matches the signal is known as the uniqueness point. The uniqueness point affects the word's recognition time: words with an early uniqueness point are recognized faster than words with a late uniqueness point. Lexical competition is a good example of lexical engagement, an instance of interaction between lexical units and hence a potentially reliable test of lexical integration of novel words.

Gaskell and Dumay (2003) exposed participants in a phoneme monitoring task to meaningless spoken novel words, and sought to examine whether these novel words engage in lexical competition with existing words. To this end, the novel words were derived from existing base words with relatively early uniqueness points (e.g., *cathedruke* derived from *cathedral*). Lexical competition effects were tested by measuring lexical decision and pause detection times to the base words. Both tasks are thought to index lexical activity of the lexical competitors. The authors found that being exposed to *cathedruke* slowed down response times to *cathedral*, not immediately after exposure, but only in a delayed test session several days later. Explicit form knowledge, a measure of lexical configuration, was tested in a two-alternative forced choice recognition memory task requiring discrimination between trained novel words and similar sounding untrained novel words. Highly accurate performance was found immediately after training, with only a slight improvement with time. The lexical decision and pause detection data showed that newly acquired words engage in lexical competition with existing words only after a period of consolidation while the forced choice data showed that the form of the novel words had been acquired to a high degree immediately after training. Together these data suggest that

novel words integrate in the mental lexicon after consolidation. The time course of lexical integration was further narrowed down by Dumay et al. (2004) who showed that lexical competition effects can be observed as soon as 24 h after exposure. Tamminen and Gaskell (2008) retested eight months after initial exposure and still found lexical competition effects in a lexical decision task. This was interpreted as evidence against an argument that the lexical representations created in these experiments might be only transient and not as well established as words learned outside the laboratory environment.

It is plausible that the effects of lexical consolidation in these word learning studies could be an artifact of repeated testing following training. Perhaps the lexical competition effects emerged as a function of increasing practice with the test task? This potential confound was eliminated in a study reported by Davis et al. (2009) who modified the testing procedure so that only one testing session was required. Participants learned one set of novel words on day 1, and another set on the following day. Testing took place immediately after the second training session on day 2. Hence one set of novel words had been learned on the previous day, and had a chance to be consolidated prior to testing. As expected, the consolidated set showed lexical competition effects despite the fact that only one testing session occurred. The unconsolidated set did not show competition effects.

These lexical competition effects are not limited to learning new spoken words. Many models of written word recognition postulate a similar mechanism of competition where orthographic neighbors, such as *banish* and *vanish*, inhibit each other's activation (McClelland and Rumelhart, 1981). Bowers et al. (2005) taught participants novel words which, if lexically integrated, should become competitors with existing 'hermit' words, that is, words with no prior neighbors (e.g., *banana* and the novel word *banara*). Novel words showed weak inhibitory effects immediately after training and strong effects one day later. This finding replicates the lexical consolidation effect with weak immediate effects and strong effects one day later, with no additional training in between (although cf. Qiao et al., 2009).

The data reviewed so far are usually interpreted as demonstrating the generation and consolidation of a new lexical entry. However, this interpretation is complicated by the fact that in all of these studies the novel words have been close variants of existing words (e.g., *cathedruke* vs *cathedral*). This means that it is unclear whether consolidation in these studies operates on new lexical representations. It might be the case that what is being learned and consolidated is either a new variant of an existing lexical entry or an episodic memory trace of the novel word. In the latter case, lexical competition effects might arise after consolidation due to the novel episodic memory trace being activated by exposure to the closely related real word (e.g., *cathedral*), and this activation diverts attentional resources which may cause slowing of responses to the base words. These alternative accounts were addressed by Dumay and Gaskell (2012) who trained participants on novel spoken words that fully embed the base word, such as *lirmucktoze*. Here the familiar base word *muck* is far less noticeable than in the onset overlapping novel words discussed so far. Competition effects were found one day and week after training but not immediately after training, thus replicating the lexical consolidation effects obtained in earlier studies. Dumay and Gaskell (2012) tested the episodic account with a different task of lexical engagement: word spotting. In word spotting, participants are asked to spot spoken words such as *mess* in carrier sequences such as *domess*. The typical finding is that if the carrier sequence is compatible with a familiar word, then word spotting times are slowed down (e.g., *domess* being compatible with *domestic*), and this reflects the inhibition of the short embedded word (*mess*) by its longer competitor (*domestic*). Applied to word learning, the logic is that spotting *muck* in *lirmuckt* should be slower if participants have previously learned *lirmucktoze*. If lexical consolidation is needed for lexical engagement, this inhibitory effect should emerge only after a delay. The episodic account makes the opposite prediction. If what is being consolidated is an episodic memory trace, then word spotting times should speed up over time as a strengthening episodic trace should help phonetic decoding of the carriers. The data however showed facilitation immediately after training, which turned into inhibition after consolidation, supporting the view that initial episodic lexical representations become integrated in the lexicon over a period of consolidation. Also, a two-alternative forced choice task and a free recall task both showed improved performance after consolidation.

Integrating novel word forms with familiar word forms in the mental lexicon is not the only type of integration newly learned words must undergo, as there are other types of linguistic knowledge that the words need to be integrated with. One important example is semantic knowledge. Integration of a novel word in existing semantic stores of knowledge should manifest as the novel word affecting the activation of other semantically related words or concepts. One example of such interactions comes from experiments where printed words can slow naming times of semantically related pictures when the two are presented simultaneously or with a very short stimulus onset asynchrony (i.e., picture–word interference). Clay et al. (2007) showed that presenting a novel word simultaneously with a picture, slowed picture naming times relative to presenting an untrained nonword, and this effect was observed immediately after training. This general picture–word interference effect however does not imply that the meaning of the novel word has been learned, as the same effect is observed when any real word is presented with a picture, regardless of the semantic relationship between the items. Crucially, presenting a novel word simultaneously with a semantically related picture slowed naming times relative to presenting an unrelated novel word. This effect was not observed immediately after training, but only a week later. These data suggest that a lexical representation can be created very quickly (the source of general picture–word interference). However, generation of a semantic representation able to exert picture–word interference required a period of consolidation, suggesting that lexical consolidation is necessary for integrating novel words with existing semantic networks.

Tamminen and Gaskell (2013) addressed the same issue using semantic priming. In semantic priming the processing of a word (e.g., *doctor*) is facilitated by previous exposure to

a semantically related word (e.g., *nurse*). Such effects are usually interpreted as reflecting automatic spreading of semantic activation amongst related words (McNamara, 2005). Therefore, Tamminen and Gaskell (2013) reasoned that if a novel word becomes integrated in existing networks of semantic knowledge, they should come to prime semantically related existing words. This was found to be the case, both when the novel word prime was presented long enough to be consciously accessed, and when it was presented very briefly and masked so that participants had little conscious awareness of it. This masking procedure is important because it implies that the spread of activation was automatic rather than influenced by response strategies. The priming effect did not emerge immediately after training on the novel words, but only after a week's consolidation opportunity. Together these semantic priming data and the picture–word interference provide strong evidence that lexical consolidation does not operate only on representations of word forms, but also in their integration with semantic knowledge.

We have seen that lexical consolidation is necessary for integrating new words with existing lexical knowledge of word forms and meanings. What about the influence of novel words on phonological knowledge? Snoeren et al. (2009) demonstrated compensation for place assimilation with newly learned spoken words. Place assimilation refers to the finding that in continuous speech certain word-final consonants change according to the properties of the first consonant of the following word (e.g., the /n/ in *lean bacon* is often assimilated with the following /b/, resulting in *leam bacon*). Listeners however perceive the assimilated phoneme as an instance of the canonical form, in other words they compensate for assimilation. This compensation is typically not found in assimilated nonwords, suggesting that the effect is lexically driven. In Snoeren et al. (2009), participants learned novel words ending in /t/ or /n/ (e.g., *decibot*), and were later exposed to a test phase where the assimilated form of the words were presented in a spoken sentence (e.g., *the decibo[p] behaved badly*) in a phoneme categorization task. Participants did indeed show compensation with the newly learned words but not with untrained nonwords, suggesting that there is significant top-down lexical involvement in perception of assimilation, and that only trained novel words had acquired enough of a lexical status to allow for compensation for assimilation to operate. This was the case for words tested immediately after training and one day after training. However, phoneme categorization responses at test were faster for the words learned the day before testing, suggesting that consolidation facilitated access to the phonological form of the novel words.

Lindsay et al. (2012) used the Ganong effect as a tool for measuring the degree with which novel words can bias phonemic categorization. The Ganong effect refers to the finding that in a phoneme categorization task listeners interpret an ambiguous phoneme that could equally well be /d/ or /t/ (henceforth /?/) as more /d/-like when presented in *?ash* than in *?ask*, because *dash* is a familiar word while *dask* is not (Ganong, 1980). If newly learned words can exert a similar lexical influence on phonemic categorization, this would suggest that they enjoy a similar lexical status to familiar words. Lindsay et al. (2012) trained participants on novel words that ended in a /d/ (e.g., *confald*) or did not end in a /d/ but instead had an untrained regular past tense inflected form that would end in /d/ (e.g., *confal* whose past tense inflected form would be *confalled*). At test, participants carried out a phoneme categorization task on *confal?*. The results showed an immediate bias toward interpreting /?/ as /d/ in these novel words, showing consolidation-independent lexical influences in phonological processing, just like in Snoeren et al. (2009). However, the size of this effect increased over a week of consolidation. In addition, phoneme categorization response times to test items where the final phoneme was not ambiguous (e.g., *confald*) became faster over the week, suggesting that access to novel word phonology is facilitated by consolidation.

Neuroimaging Evidence for Lexical Consolidation

The consolidation effects probably reflect a change over time in how novel words are represented, and therefore effects of consolidation should also be detectable in how these words are represented in the brain. Davis et al. (2009) used functional magnetic resonance imaging (fMRI) to discover the neural structures mediating processing of consolidated novel words (learned the day before scanning) and unconsolidated novel words (learned shortly before scanning). Here the primary measure of lexical integration was the neural response to consolidated novel words, unconsolidated novel words, untrained novel words, and real words. An elevated response to unconsolidated and untrained novel words compared with real words was found in a bilateral region of the superior temporal gyrus, and also in bilateral motor cortex, supplementary motor area, and cerebellum. Interestingly, this difference was not found between the consolidated novel words and real words, suggesting that only consolidated and not unconsolidated novel words were mediated by the same neural structures as real words. When a contrast was made between consolidated, unconsolidated, and untrained novel words, a region of interest encompassing the hippocampus showed higher activation to untrained than unconsolidated words in the first scanning run. This suggested that the untrained novel words, which had never been heard before, engaged the hippocampus upon their first exposure. This activity declined in the following two scanning runs, and became similar to activity elicited by trained novel words. Magnitude of hippocampal activity on the first run was also positively correlated with later recognition memory of the novel words in a two-alternative forced choice task.

A similar pattern of hippocampal activity was reported by Breitenstein et al. (2005). Participants learned novel names for familiar objects while brain activity was monitored using fMRI. As training progressed, hippocampal response to the novel words decreased. This hippocampal disengagement was mirrored by increasing activity in the inferior parietal lobe as a function of training. The authors suggested that this is the neocortical site for permanent phonological storage of novel words. Hippocampal activity during training also correlated with behavioral measures of novel word knowledge. Participants who showed less decrease of hippocampal activity over the course of training had better learning

outcome at test, and improved more during training. Both the Davis et al. (2009) and the Breitenstein et al. (2005) studies suggest that the hippocampus is important in the initial encoding of novel words, but that other areas of the brain take over longer-term representation over the period of lexical consolidation, a prediction made by the CLS accounts of memory. I now turn to a more detailed description of the CLS accounts, which provide more information on the processes that are involved in consolidation, and make further predictions about its role in word learning.

Complementary Learning Systems

Lexical consolidation appears to allow the strengthening of newly learned lexical representations, and their integration with lexical–semantic networks of existing knowledge. The neuroimaging evidence shows that this change is associated with a shift from early hippocampal representation of new words to neocortical representation. These findings fit elegantly with predictions made by CLS accounts of memory (Alvarez and Squire, 1994; McClelland et al., 1995).

CLS models postulate two learning systems with different rates of learning, different structures of representation, and a transfer process between the two systems. The neocortex is responsible for permanent memory storage, and information here is represented in a form of overlapping representations. This overlap allows the information to be integrated with existing knowledge and generalization in learning to take place. However, there are several reasons why the learning rate needs to be slow in a system like this. McClelland et al. (1995) argued that slow learning, interleaved with reactivation of existing related memories, is necessary to ensure that the existing knowledge structure is not disturbed by the incoming information. This is a particularly important lesson from computational modeling, where massed learning of new information has been shown to lead to 'catastrophic interference,' where the new information can overwrite the old information. Alvarez and Squire (1994) further motivate a slow learning rate by pointing out that memories in the neocortex are represented over geographically disparate areas with slowly emerging connectivity. In these models, consolidation occurs in the form of gradual strengthening of the neocortical trace, during a reinstatement process where the memory trace is repeatedly reactivated over time. Reinstatement can occur with direct exposure to the stimulus, or with explicit recall of the stimulus. However, much of the time new memories are generated in response to events that occur only once, not providing the necessary reinstatement. Hence the CLS models argue that reinstatement takes place off-line as well, including during sleep.

Because the initial neocortical memory trace is too weak to activate fully on its own in the absence of external stimulation or in response to partial stimulation, a second system is needed to 'bind' the neocortical representation together until it has become strong enough through consolidation. The CLS models suggest that this system is provided by the medial temporal lobe, including the hippocampus. Unlike the neocortex, the hippocampus allows fast generation of sparse, non-overlapping representations. The sparseness of the hippocampal representations circumvents the catastrophic interference problem while allowing fast learning. Furthermore, since the connections between the hippocampus and neocortex can be modified quickly, the hippocampal and the emerging neocortical representations are linked immediately. This means that even partial activation of the emerging neocortical representation can be boosted with the support from the hippocampus to activate the complete representation. As the neocortical trace gains in strength with consolidation, the hippocampal support will eventually become superfluous, and the neocortical memory will become independent of the hippocampus.

Lexical Consolidation during Sleep

One of the predictions made by the CLS accounts is that much of memory consolidation should occur during sleep, as sleep provides an interference-free time for replay of newly acquired memories (Diekelmann and Born, 2010). Dumay and Gaskell (2007) asked whether lexical competition effects in novel words emerge after a consolidation opportunity consisting mostly of sleep or wake. Participants learned novel spoken words either in the evening (sleep group) or in the morning (wake group), and were tested immediately and after 12 h. Note that only the sleep group who learned the words in the evening got to sleep before the second test. The results showed that only the sleep group showed lexical competition effects after the 12 h, and that free recall performance also increased during sleep but not during wake. When the wake group was tested for a third time, after a night of sleep, their performance converged with that of the sleep group.

Recall that CLS accounts suggest that consolidation is driven by hippocampal replay of new memories, allowing their redistribution to neocortical stores. Tamminen et al. (2010) sought evidence of this in lexical consolidation during sleep. Participants learned novel spoken words in the evening and then spent the night in the laboratory under polysomnographic (PSG) monitoring. PSG allows the researcher to record the amount of time participants spend in different stages of sleep, as well as the identification of other sleep-related neural events. Lexical competition effects emerged during the night of sleep and were associated with sleep spindle activity. Sleep spindles are brief bursts of high frequency activation seen in nonrapid eye movement sleep. They are highly relevant to consolidation processes, as it has been suggested that they provide a marker of hippocampal–neocortical dialogue (Diekelmann and Born, 2010). This study provides evidence that the shift from hippocampal to neocortical representation of new words observed by Davis et al. (2009) occurs during sleep and is mediated by hippocampal reactivation, as predicted by CLS accounts.

Sleep-associated consolidation is relevant also for learning new words in a foreign language. Gais et al. (2006) asked English speaking participants to memorize English–German word pairs. Recall of the German words was tested immediately and again 24 or 36 h later. Half of the participants were trained in the evening, so they slept immediately after learning. The other half was trained in the morning, so training was followed by a normal day of wakefulness prior to a normal night of

sleep. Participants who slept shortly after training forgot significantly fewer German words than participants who were trained in the morning. In another experiment, half of the participants proceeded to have a normal night of sleep after learning, while the other half were sleep deprived for the duration of the night following learning. Recall test after a recovery night showed that the sleep group forgot significantly fewer words. These data show that sleep shortly after learning protected the new words from forgetting, possibly by rendering the new lexical representation more robust.

Lexical Consolidation during Wake

While sleep may provide the optimal brain state for memory consolidation, consolidation can also occur during wake (Mednick et al., 2011). Under the CLS view, consolidation during off-line states is needed due to the requirement of the slow-learning neocortex to have new information presented spread over time and interleaved with activation of existing related information. These conditions are needed to avoid disturbing existing knowledge. If learning during wake meets these conditions, then consolidation should occur even without sleep. Lindsay and Gaskell (2013) trained participants on novel spoken words either in one training session (massed learning) or in several sessions spread over the course of a day (spaced learning). In addition, the base words from which the novel words were derived from (e.g., *cathedral*, from which the novel word *cathedruke* was derived) were also presented over the course of the day so that novel word training was interleaved with activation of the relevant existing information. Lexical integration occurred by the end of the day (in both massed and spaced learning), but only if interleaving was allowed. If no interleaving was incorporated into the schedule, integration was found only after a night of sleep. These data support the assertion that one of the purposes of consolidation is to allow the integration of new information with existing information in an interleaved manner. Training that meets this requirement rarely occurs in real life, and therefore sleep may provide an off-line alternative.

Szmalec et al. (2012) have also argued that properties of the training regime determine whether sleep is necessary for lexical consolidation. They used the Hebb repetition paradigm to train participants on novel words that overlap with existing words. Here participants were visually presented with sequences of nine syllables for immediate serial recall, with one sequence repeated every third trial (i.e., the Hebb sequence). The Hebb sequence in this study consisted of three novel words. In the test phase, the integration of these novel words into the mental lexicon was tested using the pause detection and lexical decision tasks, following Gaskell and Dumay (2003). No evidence of lexical competition immediately after completion of the Hebb repetition task was found, but the effect emerged 12 h later, both when the 12 h consisted of sleep or wake. This again demonstrates the need for lexical consolidation to take place before lexical integration can be observed, but suggests that this consolidation can occur during wake. The authors argue that sleep is necessary only when the training results in weaker memory traces and when participants are explicitly asked to learn the new words. In the Hebb repetition paradigm participants were not explicitly aware that they would be tested on the novel words contained in the Hebb sequences, and the paradigm involved repeated testing during the training session, possibly resulting in a stronger memory trace for the novel words than in the phoneme monitoring task used in the studies that have reported effects of sleep on spoken novel word learning (Dumay and Gaskell, 2007; Tamminen et al., 2010).

Lexical Consolidation in Children

Most of the studies demonstrating lexical consolidation have been carried out on adults. What role, if any, does lexical consolidation play in young children? It is well known that children as young as 2-years-old are able to learn new words remarkably quickly as a result of minimal exposure to the word and its meaning. This phenomenon of 'fast-mapping' (Carey and Bartlett, 1978) suggests that perhaps the need for consolidation seen in adults is less prominent in children. To explore this issue, Henderson et al. (2012) trained 7–12 year olds novel spoken words and tested lexical competition effects immediately and after a consolidation opportunity consisting of sleep or wake, following the design of Dumay and Gaskell (2007). Just like adults, the children showed no evidence of newly learned words engaging in lexical competition immediately after training. The effect emerged after 12 h, but only in the group that slept during that time. A two-alternative forced choice recognition memory task and a cued recall task also showed improvements in performance that were specific to sleep, as reported in adults (Tamminen et al., 2010). Converging data were reported by Brown et al. (2012) who found that in 7- and 12-year-olds cued recall of newly learned words is improved over a 24-h period following learning (and including nocturnal sleep), but not over a short period of 3–4 h that also did not include sleep. Together these data suggest that word learning in children is not complete immediately after exposure, and that lexical consolidation operates in children much like it does in adults. Furthermore, sleep appears to be the optimal brain state for lexical consolidation to occur for children as well as adults.

Conclusions

The evidence described here suggests that lexical consolidation operates on several levels. At one level, consolidation facilitates access to novel word representations as seen in gradually improving performance in recognition and recall tasks. At another level, consolidation is necessary for integrating novel lexical representations with different levels of linguistic knowledge, ranging from low level phonological knowledge to lexical and semantic knowledge. Sleep appears to be the optimal state for this lexical consolidation to occur, and this is likely to be due to hippocampal–neocortical interaction taking place during sleep. Neuroimaging evidence supports this view by implicating the hippocampus in the early stages of word learning prior to consolidation, while neocortical areas seem to mediate representation of consolidated novel words. These

findings are consistent with CLS accounts of memory, and seem to apply both to adult and child populations.

While this article has focused on the consolidation of new lexical representations, it is worth noting in closing that new research has begun to identify other aspects of language learning that require memory consolidation. For example, Tamminen et al. (2012) trained participants on new affixes (e.g., *-nule*) attached to familiar word stems (e.g., *sleepnule*). Participants showed evidence of having learned the new affix immediately after training when it was attached to the same stem as during training, but such evidence was seen only after consolidation if the affix was attached to a stem not seen during training (e.g., *jugnule*), suggesting that newly learned affixes become context-independent only after consolidation. Similar evidence of linguistic generalization was shown by Gomez et al. (2006) who exposed infants to an artificial language governed by a syntactic rule. Only infants who took a nap after exposure were able to apply the rule to new words, suggesting that sleep-associated consolidation is critical in generalization. These and the examples described earlier demonstrate how language learning provides a uniquely rich environment for uncovering the mechanisms underlying memory consolidation. Our understanding of both domains will benefit as interest in this work continues to grow.

See also: Cognitive Neuroscience; Hippocampus and Related Structures; Implicit Memory; Language Acquisition; Learning and Memory, Neural Basis of; Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects; Lexical Semantics; Lexicon; Psycholinguistics: Overview; Recognition; Semantic Memory; Sleep and Memory; Sleep: Neural Systems; Speech Perception; Visual Word Recognition, Neurocognitive Psychology of; Vocabulary Acquisition.

Bibliography

Alvarez, R., Squire, L.R., 1994. Memory consolidation and the medial temporal lobe: a simple network model. Proceedings of National Academy of Sciences (USA) 91, 7041–7045.

Bowers, J.S., Davis, C.J., Hanley, D.A., 2005. Interfering neighbours: the impact of novel word learning on the identification of visually similar words. Cognition 97, B45–B54.

Breitenstein, C., Jansen, A., Deppe, M., Foerster, A.-F., Sommer, J., Wolbers, T., Knecht, S., 2005. Hippocampus activity differentiates good from poor learners of a novel lexicon. NeuroImage 25, 958–968.

Brown, H., Weighall, A., Henderson, L.M., Gaskell, M.G., 2012. Enhanced recognition and recall of new words in 7- and 12-year old children following a period of offline consolidation. Journal of Experimental Child Psychology 112, 56–72.

Carey, S., Bartlett, E., 1978. Acquiring a single new word. Proceedings of the Stanford Child Language Conference 15, 17–29.

Clay, F., Bowers, J.S., Davis, C.J., Hanley, D.A., 2007. Teaching adults new words: the role of practice and consolidation. Journal of Experimental Psychology: Learning, Memory, and Cognition 33, 970–976.

Davis, M.H., Di Betta, A.M., Macdonald, M.J.E., Gaskell, M.G., 2009. Learning and consolidation of novel spoken words. Journal of Cognitive Neuroscience 21, 803–820.

Davis, M.H., Gaskell, M.G., 2009. A complementary systems account of word learning: neural and behavioural evidence. Philosophical Transactions of the Royal Society 364, 3773–3800.

Diekelmann, S., Born, J., 2010. The memory function of sleep. Nature Reviews Neuroscience 11, 114–126.

Dumay, N., Gaskell, M.G., 2007. Sleep-associated changes in the mental representation of spoken words. Psychological Science 18, 35–39.

Dumay, N., Gaskell, M.G., 2012. Overnight lexical consolidation revealed by speech segmentation. Cognition 123, 119–132.

Dumay, N., Gaskell, M.G., Feng, X., 2004. A day in the life of a spoken word. In: Forbus, K., Gentner, D., Regier, T. (Eds.), Proceedings of the 26th Annual Conference of the Cognitive Science Society. Erlbaum, Mahwah, NJ, pp. 339–344.

Gais, S., Lucas, B., Born, J., 2006. Sleep after learning aids memory recall. Learning & Memory 13, 259–262.

Ganong, W.F., 1980. Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception and Performance 6, 110–125.

Gaskell, M.G., Dumay, N., 2003. Lexical competition and the acquisition of novel words. Cognition 89, 105–132.

Gomez, R.L., Bootzin, R.R., Nadel, L., 2006. Naps promote abstraction in language-learning infants. Psychological Science 17, 670–674.

Henderson, L.M., Weighall, A., Brown, H., Gaskell, M.G., 2012. Vocabulary acquisition is associated with sleep in children. Developmental Science 15, 674–687.

Leach, L., Samuel, A., 2007. Lexical configuration and lexical engagement: when adults learn new words. Cognitive Psychology 55, 306–353.

Lindsay, S., Gaskell, M.G., 2013. Lexical integration of novel words without sleep. Journal of Experimental Psychology: Learning. Memory & Cognition 39, 608–622.

Lindsay, S., Sedin, L.M., Gaskell, M.G., 2012. Acquiring novel words and their past tenses: evidence from lexical effects on phonetic categorisation. Journal of Memory and Language 66, 210–225.

Marslen-Wilson, W., 1987. Functional parallelism in spoken word recognition. Cognition 25, 71–102.

McClelland, J.L., McNaughton, B.L., O'Reilly, R.C., 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychological Review 102, 419–457.

McClelland, J.L., Rumelhart, D.E., 1981. An interactive activation model of context effects in letter perception: part 1. An account of basic findings. Psychological Review 88, 375–407.

McNamara, T.P., 2005. Semantic Priming: Perspectives from Memory and Word Recognition. Psychology Press, New York.

Mednick, S.C., Cai, D.J., Shuman, T., Anagnostaras, S., Wixted, J.T., 2011. An opportunistic theory of cellular and systems consolidation. Trends in Neurosciences 34, 504–514.

Qiao, X., Forster, K., Witzel, N., 2009. Is banara really a word? Cognition 113, 254–257.

Snoeren, N.D., Gaskell, M.G., Di Betta, A.M., 2009. The perception of assimilation in newly learned novel words. Journal of Experimental Psychology: Learning, Memory, and Cognition 35, 542–549.

Szmalec, A., Page, M.P.A., Duyck, W., 2012. The development of long-term lexical representations through Hebb repetition learning. Journal of Memory and Language 67, 342–354.

Tamminen, J., Davis, M.H., Merkx, M., Rastle, K., 2012. The role of memory consolidation in generalisation of new linguistic information. Cognition 125, 107–112.

Tamminen, J., Gaskell, M.G., 2008. Newly learned spoken words show long term lexical competition effects. Quarterly Journal of Experimental Psychology 61, 361–371.

Tamminen, J., Gaskell, M.G., 2013. Novel word integration in the mental lexicon: Evidence from unmasked and masked semantic priming. Quarterly Journal of Experimental Psychology 66, 1001–1025.

Tamminen, J., Payne, J.D., Stickgold, R., Wamsley, E.J., Gaskell, M.G., 2010. Sleep spindle activity is associated with the integration of new memories and existing knowledge. Journal of Neuroscience 30, 14356–14360.

Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects

Matthew A Lambon Ralph, School of Psychological Sciences, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This entry reviews the psychological and neural aspects of the mental lexicon and word knowledge. This topic is covered in six main sections. The first summarizes the various empirical techniques – experimental psychology, neuropsychology, transcranial magnetic stimulation, functional imaging, and computational modeling – which are used to investigate word knowledge and how that is represented in the brain. This is followed by a description of the traditional conception of the lexicon including an overview of Morton's influential logogen theory. The third section summarizes the data, which have been used to argue that word knowledge is supported by a number of interactive yet functionally distinct lexicons. These traditional views of the mental lexicon are contrasted in the fourth section with a different approach that has arisen from a form of computational models (e.g., the Seidenberg and McClelland word reading and recognition model), which are able to extract word knowledge without having a 'lexicon.' The penultimate section reviews how lexical knowledge interacts with the remaining language system to give rise to specific language activities such as speech production, reading aloud, and verb inflection. The final section considers how future work may ultimately inform us about the neural basis of word knowledge.

> He doth not neglect the profit of Lexicons (wherein all sayings and speeches are numbered)
>
> E. Topsell, The historie of foure-footed beastes, 1607.

If one looks in any psychology or psycholinguistics textbook for a definition of the mental *lexicon*, the superficial description given does not differ to one of the first uses of the word quoted above from the Oxford English Dictionary. Most modern researchers would agree that the 'lexicon' refers to our knowledge of the words in our vocabulary. There is, however, a very broad range of opinion if one asks the next obvious questions: what is this knowledge and how is it represented in the brain?

As with many areas of higher brain function, studies of the mental lexicon generally follow one of two levels of explanation: descriptive and computational. There are currently no theories of word knowledge at the level of individual neurons. Descriptive explanations try to outline the nature of the lexicon in detail, attempting to establish both what information is represented by the lexicon and how this information is accessed. Computational explanations provide more explicit descriptions of word knowledge and lexical processes either through formal mathematical treatments or in the form of computer-based models that take a specific input and transform it to the appropriate output.

Studies on the lexicon can be grouped into four categories, each employing a number of different research methodologies:

1. *Neurologically intact participants*: This type of research relies principally on reaction time and to a lesser extent on accuracy, as the critical dependent measures. Participants are typically required to give some form of response to spoken or written words (word recognition) or to give a spoken, single name (word production). A large number of different tasks are employed in each of these modalities. These include lexical decision (deciding whether a spoken or written stimulus is a real word or not), word monitoring (listening to spoken input and indicating when a target word has been presented), and picture naming and word naming (single word reading aloud). All these techniques aim to capture variations in reaction times in terms of differences between words of different types and/or words with different psycholinguistic characteristics. Alternatively, the experimenter aims to manipulate a baseline reaction time through different presentation conditions (for example, the experimenter may investigate the effect of semantic priming – presenting a target word soon after a semantically related item).
2. *Permanent or transient brain dysfunction*: The critical assumption underpinning neuropsychological research is that different aspects of language or, in this case, word knowledge are supported by different areas of the brain (concentrated almost entirely within the left, dominant hemisphere). Therefore, by studying patients with damage to one of these areas, it should be possible to reveal the underlying characteristics of how the human brain supports language, including word knowledge. Unlike studies with neurologically intact participants, most neuropsychological research relies on accuracy rather than reaction time measures. Studies are conducted on groups of patients (comparing them to control participants and/or different types of patient) or individual cases (Shallice, 1988). A number of recent studies have also used a technique known as case-series methodology in which patients are selected on the basis of having the same type of deficit but at different levels of severity (Schwartz and Dell, 2011). More recently, transcranial magnetic stimulation (TMS) has been used to test the effects of transient, localized cortical dysfunction on task performance in neurologically intact participants. The technique uses an electromagnetic field to stimulate the cortex underlying the coil, and the effect on task performance is measured (Walsh and Cowey, 2000). With the correct design of coil (e.g., a figure-of-eight), it is possible to apply a focused magnetic field and thus influence a targeted, small cortical region. As a consequence, it is possible to investigate the role of target areas and also the timecourse of

their function. The method is, therefore, a complement to neuropsychological studies and provides a mechanism by which the functional relevance of regions identified by functional neuroimaging can be assessed.

3. *Functional neuroimaging*: For those techniques that concentrate more on spatial than temporal resolution, the underlying assumption for functional neuroimaging is the same as for neuropsychological studies with aphasic patients. In this way, functional magnetic resonance imaging (fMRI) or positron emission tomography (PET) compare words that vary according to a specific characteristic either using a parametric, subtraction methodology (looking for areas of the brain that are more active for one set than another) or a correlational technique (looking for areas in the brain where changes in activation are correlated with a changing characteristic of the presented words). Other imaging techniques are able to investigate how the timecourse of activation compares for words of different type. Such methods include event-related potentials and magnetoencephalography.
4. *Computational modeling*: This is a formal technique designed to instantiate a proposed computational description of the processes underlying a specific set of behaviors. Most computational models do this by specifying an arrangement of simple processing units (known as a *network architecture*). The architecture describes how simulated activation can potentially be passed between these units. Information is represented either in localist form (the activation of a single unit corresponds to a specific piece of information – e.g., a word) or in a distributed manner (a number of pieces of information are represented by different patterns of activation across the same set of units; therefore, the activation of a particular unit is not unique to a specific datum). These computational models also differ along another dimension. In some models, the strength of the connection between units (that governs how much activation is passed between them) is specified explicitly by the modeler. Other networks use mathematical algorithms such that the model can extract its own set of weights to performance optimally. In doing so, these latter types of network can also be used to compare patterns of development as well as mature performance between the model and human subjects. Both types of model can be used to simulate disordered behavior (e.g., through brain injury) by various means including the removal of units and/or their connections, or by adding noise to the activation of units.

These four methods have given rise to a vast literature on the nature of the lexicon and its interaction with the rest of the language system. Some of the major themes arising from this work are summarized below.

The Traditional Conception of a Lexicon

Many theories of lexical representation are not that different from Topsell's definition from 1607: the lexicon is a list of all the words in a person's vocabulary. By some form of 'lookup' function, all or part (in the case of models with multiple lexicons, see below) of our knowledge about a word is retrieved. An early functional and mathematical description of this process was given by Morton with his influential logogen theory (1969), which was originally applied to explain word recognition but was extended later to address spoken and written recognition and production. The key elements to this approach are as follows. Each word is represented by a separate unit, or logogen, which gathers evidence for the presence of that word in the input (in the form of sensory input for recognition or semantic memory for production). Each logogen does this relatively independently of the others. A word is recognized or produced when there is sufficient positive evidence for that word in the input (e.g., the letters in the written input match those for the target word). At this point, the logogen makes other information about that word available – such as the word's meaning or its pronunciation. With reference to signal detection theory, Morton also included the notion of a 'threshold,' which governs how much positive evidence is required before the logogen makes a response. In this way, word frequency effects in recognition and production are described in terms of different thresholds for different logogens (high-frequency words have lower thresholds and thus require less positive evidence to respond).

How Many Lexicons Are There?

If one considers the literate adult, the sum total of the acquired knowledge can be classified into three forms: the meaning of words; the sound of each word; and the written form or orthography. With reference to the traditional view of the lexicon (see above), some researchers have argued that these three forms of information are represented in separate lexicons: lexical semantics, a speech lexicon, and an orthographic lexicon. Although most models have interconnections between these different lexicons (at some level they need to combine information to reflect all aspects of word knowledge), the notion is that they are discrete entities at both a functional and a neural level. Functionally, each lexicon can then have its own dedicated form of representation and associated processes. At the neural level (for both neuropsychological and neuroimaging studies) this means there are distinct regions in the brain that can be damaged/activated relatively independently of each other. The cognitive neuropsychology literature, in particular, has also encouraged a further division in the mental lexicon – for input and output processes. The notion here is that there might be four separate lexicons to cover the four combinations of recognition and production for spoken and written words (for an overview, see Ellis and Young, 1988). The same division is not proposed for meaning, which in a sense would be nonsensical anyway (for effective communication, the meaning driving the speaker/writer's output must be the same as that decoded by the listener/reader). The need to have separate input and output representations is still contested, however. Some neuropsychological data would seem to favor this position – for example, there are patients with word-finding difficulties (anomia) who do not have trouble understanding words (word deafness), and other patients with the opposite pattern. However, more studies are needed to rule out the possibility that the patterns reflect modality-specific nonlexical deficits or the fact that expression is always more vulnerable to impairment than receptive abilities.

Two separate strands of research have argued for another lexicon – the semantic lexicon or lemmas, which mediate between meaning and other language processes and representations (neuropsychology – e.g., Butterworth et al., 1984; speech production – e.g., Levelt, 1992). In speech production, Levelt (1992) has argued for a two-stage process within lexical access. First, the meaning of the intended utterance is used to select a single abstract lexical form, a lemma (that also codes for syntactical markings such as gender and number), and then this in turn activates a lexeme, which makes the sound of the word available in terms of the constituent phonemes, their arrangement, and associated stress pattern. There is, however, a general concern with this direction of theorizing in that a proliferation of multiple lexicons means the potential for redundancy of representations is much higher, and tends to make the theories unfalsifiable.

An Alternative Conception of the Nature of Lexical Knowledge?

A quite different approach has been adopted by some parallel distributed processing (PDP) computational models in which there is knowledge of the words in our vocabulary but not a lexicon – at least, in the traditional sense of that word. A landmark publication in this respect was the word reading and recognition model of Seidenberg and McClelland (1989). In this framework word knowledge is represented as distributed patterns of activation across three 'primary' systems corresponding to orthography, phonology, and semantic memory. Through a gradual learning process, in which the weights of the connections within and between these representational layers are adjusted, models of this type extract word 'knowledge' from the vocabulary they are trained on. Note that this is quite different from the traditional metaphor of looking up an entry in the dictionary or Morton's logogen theory. The Seidenberg and McClelland model does not have a list of all the known words (i.e., a lexicon) so the terms 'access' and 'retrieval' are redundant. Instead the model takes one form of input (e.g., the written word) and then computes the appropriate representation reflecting the meaning and/or the pronunciation. In this way the model can be said to have word knowledge but not a lexicon. This is further reinforced by the fact that the model processes novel words (nonwords that, by definition, cannot be in a lexicon) through the very same set of units corresponding to the written or spoken form. Another potential allure of this approach to word knowledge is that the formulation may help when future research (see below) attempts to bridge the gap between brain and behavior. The neural basis for the primary systems of semantics, phonology, and orthography is also much more obvious. It is possible, for example, that orthography reflects a stable set of representations extracted through visual perception (Behrmann et al., 1998), phonology arises from the interplay of auditory and motor interactions (Plaut and Kello, 1999), and semantic memory, or conceptual knowledge, reflects high-level abstraction based on all forms of verbal and nonverbal experience, perhaps supported primarily by regions within the temporal lobes (Lambon Ralph et al., 2010; Patterson et al., 2007).

This alternative approach has also been used to address some of the data that have led researchers, using a traditional view of the lexicon, to posit the need for multiple lexicons (see above). For example, patients with both poor comprehension and production of words (whether spoken or written) are described from a traditional viewpoint in terms of damage to a semantic lexicon that mediates between conceptual knowledge and the language system (Butterworth et al., 1984). The same data are adequately captured by PDP computational principles in terms of the nature of the mapping between representations for words vs pictures/objects (Lambon Ralph et al., 2001; Rogers et al., 2004). The fact that there is an arbitrary mapping between phonology/orthography and conceptual knowledge, unlike the mapping for pictures or objects, means that verbal comprehension and production are highly vulnerable to damage to the system that represents the meaning of both words and pictures.

How Does Word Knowledge Give Rise to Specific Language Activities?

Within the repertoire of language activities, there are some that are assumed to depend directly on word knowledge or lexical information. For example, understanding words or producing words is assumed to proceed with reference to lexical information. So, as described above, within a logogen framework speech production occurs by the target meaning acting as an input to a set of logogens, which selects the correct word and then makes the spoken form available. There is also a direct use of word knowledge in PDP computational models as well – the knowledge extracted through learning the mapping between semantics and phonology is used to produce the correct spoken form in speech production. Other language activities are assumed to depend upon a combination of lexical knowledge and other nonlexical processes. So, for example, Levelt (1992) proposes that speech production is based upon a 'formulator' that specifies the syntactical frame for a sentence into which lexical representations (lemmas) and their associated knowledge are fitted. Two other well-known examples are reading aloud and the past tense transformation of English verbs. Researchers studying the process of reading aloud have classified words into three basic types: (1) most words, known as regular or consistent words, are pronounced in line with their spellings (examples of these are YARD and MINT); (2) a smaller but still substantial number of irregular or inconsistent words deviate from the typical spelling-to-sound translations (examples of these are YACHT and PINT); (3) we are also able to read new or novel words without extra training (e.g., EMAIL or WUG). There is a very similar classification for describing the past tense of an English verb. (1) There are many regular verbs, all of which are turned into the past tense by the addition of -ED (e.g., *talk* → *talked* or *need* → *needed*), and a smaller number of irregular verbs which mark the past tense in another way (e.g., *speak* → *spoke* or *think* → *thought*). We are also able to produce the past tense for new or novel verbs (e.g., *email* → *emailed*, *wug* → *wugged*), again without any specific training on these novel forms. The traditional theories of verb inflection (the dual-mechanism account – Pinker, 1999) and reading aloud (the dual-route theory – Coltheart et al., 2001) have both used this tripartite division of word type to argue that each language activity is underpinned by two distinctly different processes. The two theories are sufficiently similar that they can

be summarized together. Irregular words/verbs (e.g., PINT → 'pint' for reading aloud and 'ring' → 'brought' for verb tense) are said to be translated by reference to a lexicon – or indeed, in the case of reading, two lexicons: one orthographic and the other phonological – that stores the whole specific word form for every irregular item in the language. The ability to read or inflect novel forms is explained in these theories by positing a task-specific, generative process encapsulated by a set of rules: three rules for the three allomorphs required for the inflection of regular verbs (/t/,/d/,/@d/) and an extensive series of rules for the translation of letters/graphemes into phonemes. Neuropsychological investigations of impairments to both reading aloud and past tense production have highlighted two contrasting patterns that have been taken to support these dual-mechanism theories (some patients have specific difficulties with irregular forms and others with novel forms: see Ullman et al. (1997) for verb morphology; an overview of acquired dyslexia can be found in Ellis and Young (1988)).

Again, PDP computational models have taken a different approach (verb morphology (Joanisse and Seidenberg, 1999); reading aloud (Plaut et al., 1996)). In these frameworks word knowledge is not reflected in a lexicon and thus the strict division between the lexicon and the generative rules does not apply. Instead the interaction between and within the primary systems (semantics, phonology, and orthography) not only gives rise to word knowledge, the models are also capable of transforming one form of representation into another (e.g., reading is orthography → phonology) for all three word types (irregular, regular, and novel). Critically, however, the division of labor among the primary representations is not constant for the different word types, and dissociations in performance can occur if one of the primary systems is damaged (for an overview of this perspective on acquired dyslexias, see Patterson and Lambon Ralph, 1999). This overarching hypothesis has been tested and explored in both neuropsychological and TMS investigations. The core, central role of semantic information in supporting a range of verbal and nonverbal activities has been demonstrated through detailed investigations of patients with semantic dementia who have a selective yet progressive degradation of multimodal conceptual knowledge as a result of atrophy focused on the ventrolateral anterior temporal lobes (Patterson et al., 2006). These studies are important because they show that the same, singular semantic deficit leads to impaired performance across a range of diverse language activities (including surface dyslexia, poor past tense production of irregular verbs, surface dysgraphia, pronounced anomia, etc.). Under the more traditional lexicon-based accounts, this set of co-occurring deficits would require the unlikely situation of simultaneous yet selective impairments of the lexical representations in all domains. The importance of semantic representations in supporting 'lexical knowledge' has also been demonstrated in neurologically intact participants following TMS of the lateral anterior temporal lobe (a region that exhibits significant atrophy in semantic dementia patients (Holland and Lambon Ralph, 2010; Pobric et al., 2007)).

Beyond semantic representations, some studies have also explored the important role that phonological and visual 'primary' systems play in verbal activities. These include demonstrations that generalized phonological impairments lead to a specific form of acquired dyslexia and dysgraphia in which performance is more accurate for words than nonwords (Crisp and Lambon Ralph, 2006; Patterson and Marcel, 1992; Rapcsak et al., 2009). Likewise, some researchers have argued that pure alexia (a selective disorder of word recognition that spares spelling – 'alexia without agraphia) is caused by a more general visual impairment (Behrmann et al., 1998; Woodhead et al., 2011).

Future Directions

Research on the psychological aspects of word knowledge will, no doubt, continue in the future. One potential area for additional research must be on the neural aspects of word knowledge. Clues about the neural basis of word knowledge can be gleaned, in part, from relating the neuropsychological impairments of patients back to the underlying distribution of brain damage. In addition, functional neuroimaging can be used to explore which brain areas become more active when word knowledge is required. Interestingly, although some well-known early functional neuroimaging studies searched for the neural bases of spoken and written lexicons (Howard et al., 1992), most contemporary investigations are focused on specific language activities (for an excellent and thorough review of the development of language-related functional neuroimaging and its findings, see Price, 2012). Likewise, there is potential to use TMS to explore the necessity of specific brain regions in various aspects of word knowledge. While there have been various explorations of semantics and phonology using TMS (e.g., Devlin et al., 2003; Hartwigsen et al., 2010; Pobric et al., 2007), this line of research is still in its infancy and it has the potential to reveal much finer aspects of word knowledge and its neural representation. Finally, it should be noted that even if we have a range of techniques to explore the behavioral and neural correlates of word knowledge, we will still need formal methods for exploring how behavior arises from these neural mechanisms – i.e., to bridge between brain and mind. Indeed, functional neuroimaging has already demonstrated that many language activities are not supported by a single brain region but rather by a network of brain areas (Price, 2012). Clearly then, there is a need to understand not only the neurocomputation occurring in each region but also their connectivity and network-based activity. One possibility in this domain is for computational models of language to embrace increasing amounts of neuroanatomy by utilizing this information to constrain the architecture of the models. Although only few in number, the handful of implemented models that have taken this approach show great promise to provide a formal bridge between the neural and cognitive aspects of word knowledge (Lambon Ralph et al., 2001; Plaut, 2002; Ueno et al., 2011).

Bibliography

Behrmann, M., Plaut, D.C., Nelson, J., 1998. A literature review and new data supporting an interactive account of letter-by-letter reading. Cognitive Neuropsychology 15, 7–52.

Butterworth, B., Howard, D., McLoughlin, P., 1984. The semantic deficit in aphasia: the relationship between semantic errors in auditory comprehension and picture naming. Neuropsychologia 22 (4), 409–426.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., Ziegler, J., 2001. DRC: a dual route cascaded model of visual word recognition and reading aloud. Psychological Review 108, 204–256.

Crisp, J., Lambon Ralph, M.A., 2006. The continuum between phonological and deep dyslexia: the keys are in phonology and semantics. Journal of Cognitive Neuroscience 18, 348–362.

Devlin, J.T., Matthews, P.M., Rushworth, M.F.S., 2003. Semantic processing in the left inferior prefrontal cortex: a combined functional magnetic resonance imaging and transcranial magnetic stimulation study. Journal of Cognitive Neuroscience 15 (1), 71–84.

Ellis, A.W., Young, A.W., 1988. Human Cognitive Neuropsychology. Lawrence Erlbaum Associates Ltd, London.

Hartwigsen, G., Baumgaertner, A., Price, C.J., Koehnke, M., Ulmer, S., Siebner, H.R., 2010. Phonological decisions require both the left and right supramarginal gyri. Proceedings of the National Academy of Sciences of the United States of America 107 (38), 16494–16499.

Holland, R., Lambon Ralph, M.A., 2010. The anterior temporal lobe semantic hub is a part of the language neural network: selective disruption of irregular past tense verbs by rTMS. Cerebral Cortex 20 (12), 2771–2775.

Howard, D., Patterson, K., Wise, R., Brown, W.D., Friston, K., Weiller, C., et al., 1992. The cortical localization of the lexicons: positron emission tomography evidence. Brain 115 (6), 1769–1782.

Joanisse, M.F., Seidenberg, M.S., 1999. Impairments in verb morphology after brain injury: a connectionist model. Proceedings of the National Academy of Sciences of the United States of America 96, 7592–7597.

Lambon Ralph, M.A., McClelland, J.L., Patterson, K., Galton, C.J., Hodges, J.R., 2001. No right to speak? The relationship between object naming and semantic impairment: neuropsychological evidence and a computational model. Journal of Cognitive Neuroscience 13, 341–356.

Lambon Ralph, M.A., Sage, K., Jones, R.W., Mayberry, E.J., 2010. Coherent concepts are computed in the anterior temporal lobes. Proceedings of the National Academy of Sciences of the United States of America 107 (6), 2717–2722.

Levelt, W.J.M., 1992. Accessing words in speech production: stages, processes and representations. Cognition 42, 1–22.

Morton, J., 1969. The interaction of information in word recognition. Psychological Review 76, 340–354.

Patterson, K., Lambon Ralph, M.A., 1999. Selective disorders of reading? Current Opinion in Neurobiology 9, 235–239.

Patterson, K., Lambon Ralph, M.A., Jefferies, E., Woollams, A., Jones, R.W., Hodges, J.R., et al., 2006. "Presemantic" cognition in semantic dementia: six deficits in search of an explanation. Journal of Cognitive Neuroscience 18 (2), 169–183.

Patterson, K., Marcel, A., 1992. Phonological ALEXIA or PHONOLOGICAL alexia? In: Alegria, J., Holender, D., Juncas de Morais, J., Radeau, M. (Eds.), Analytic Approaches to Human Cognition. North-Holland, Amsterdam.

Patterson, K., Nestor, P.J., Rogers, T.T., 2007. Where do you know what you know? The representation of semantic knowledge in the human brain. Nature Reviews Neuroscience 8 (12), 976–987.

Pinker, S., 1999. Words and Rules: The Ingredients of Language. Weidenfeld & Nicolson, London.

Plaut, D.C., 2002. Graded modality-specific specialization in semantics: a computational account of optic aphasia. Cognitive Neuropsychology 19, 603–639.

Plaut, D.C., Kello, C.T., 1999. The emergence of phonology from the interplay of speech comprehension and production: a distributed connectionist approach. In: MacWhinney, B. (Ed.), The Emergence of Language. Lawrence Erlbaum Associated, Inc., Mahweh, NJ.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S., Patterson, K., 1996. Understanding normal and impaired word reading: computational principles in quasi-regular domains. Psychological Review 103 (1), 56–115.

Pobric, G.G., Jefferies, E., Lambon Ralph, M.A., 2007. Anterior temporal lobes mediate semantic representation: mimicking semantic dementia by using rTMS in normal participants. Proceedings of the National Academy of Sciences of the United States of America 104, 20137–20141.

Price, C.J., 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. NeuroImage 62 (2), 816–847.

Rapcsak, S.Z., Beeson, P.M., Henry, M.L., Leyden, A., Kim, E., Rising, K., et al., 2009. Phonological dyslexia and dysgraphia: cognitive mechanisms and neural substrates. Cortex 45 (5), 575–591.

Rogers, T.T., Lambon Ralph, M.A., Garrard, P., Bozeat, S., McClelland, J.L., Hodges, J.R., et al., 2004. The structure and deterioration of semantic memory: a neuropsychological and computational investigation. Psychological Review 111, 205–235.

Schwartz, M.F., Dell, G.S., 2011. Case series investigations in cognitive neuropsychology. Cognitive Neuropsychology 27 (6), 477–494.

Seidenberg, M., McClelland, J.L., 1989. A distributed, developmental model of word recognition and naming. Psychological Review 96, 523–568.

Shallice, T., 1988. Specialisation within the semantic system. Cognitive Neuropsychology 5, 133–142.

Ueno, T., Saito, S., Rogers, T.T., Lambon Ralph, M.A., 2011. Lichtheim 2: synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. Neuron 72 (2), 385–396.

Ullman, M.T., Corkin, S., Coppola, M., Hickok, G., Growdon, J.H., Koroshetz, W.J., et al., 1997. A neural dissociation within language: evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. Journal of Cognitive Neuroscience 9 (2), 266–276.

Walsh, V., Cowey, A., 2000. Transcranial magnetic stimulation and cognitive neuroscience. Nature Reviews Neuroscience 1 (1), 73–79.

Woodhead, Z.V.J., Wise, R.J.S., Sereno, M., Leech, R., 2011. Dissociation of sensitivity to spatial frequency in word and face preferential areas of the fusiform gyrus. Cerebral Cortex 21 (10), 2307–2312.

Lexical Semantics

Dirk Geeraerts, Faculty of Arts, KU Leuven, Leuven, Belgium

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D.A. Cruse, volume 13, pp. 8758–8764, © 2001, Elsevier Ltd.

Abstract

The scope of linguistic lexical semantics is described from a theoretical and descriptive point of view: an overview of the main theoretical approaches that have succeeded each other in the history of lexical semantics (prestructuralist historical semantics, structuralist semantics, cognitive semantics) provides a framework for illustrating the various descriptive topics and models that were introduced by the successive theories.

Lexical semantics is the study of word meaning. To get a good idea of the scope of lexical semantics (and the present article), a number of demarcational distinctions need to be introduced. First, linguistic semantics is obviously not restricted to lexical semantics: for a broader view of meaning in language, including the crucial areas of grammatical and discursive semantics, see the article Semantics. Second, although the label 'lexical semantics' may suggest otherwise, morphological semantics (the study of the meaning of morphemes and the way in which they combine into words) is usually considered a separate field from lexical semantics proper. Morphological semantics is not covered by the present article. Third, the study of the meaning of words does not exhaust the linguistic study of words: topics like lexical variation across language varieties, borrowing, phraseology, or the origin of words are distinct from the study of word meaning, but like lexical semantics, they form part of the broader field of lexicology. Again, lexicology at large (and lexicography, as an applied discipline) lies beyond the scope of the present chapter; see the article 53059 for more.

The article is organized in a historical rather than thematic way, because the various topics treated in lexical semantics are very much tied up with the different theoretical approaches that have dominated the development of lexical semantics. For an entry-level text on lexical semantics, see Murphy (2010) or the chapters devoted to word meaning in Saeed (2009), Riemer (2010), and Cruse (2010). For a more extensive and detailed overview of the main historical and contemporary trends of research in lexical semantics, see Geeraerts (2010).

Prestructuralist Historical Semantics

The prestructuralist period (ranging from the middle of the nineteenth century up to the 1930s) was the heyday of historical semantics, in the sense that the study of meaning change reigned supreme within semantics. The main theoretical achievement of prestructuralist historical semantics consists of various classifications of types of semantic change, coupled with considerable attention for psychological processes as the explanatory background of changes: the general mechanisms of change included in the classifications were generally considered to be based on the associative patterns of thought of the human mind. Important figures are Paul (1920) (first edition 1880), Bréal (1897), and Stern (1931).

The most common classification of semantic changes emerging from this tradition includes four basic types: specialization, generalization, metaphor, and metonymy. In the case of semantic *specialization*, the new meaning is a restriction of the old meaning: the new meaning is a subcase of the old. In the case of semantic *generalization*, the reverse holds: the old meaning is a subcase of the new. Classical examples of specialization are *corn* (originally a cover term for all kinds of grain, now specialized to 'wheat' in England, to 'oats' in Scotland, and to 'maize' in the United States), *starve* (moving from 'to die' to 'to die of hunger'), and *queen* (originally 'wife, woman,' now restricted to 'king's wife, or female sovereign'). Examples of generalization are *moon* (primarily the earth's satellite, but extended to any planet's satellite) and French *arriver* (which originally meant 'to reach the river's shore, to embank,' but which now signifies 'to reach a destination' in general). There is a lot of terminological variation in connection with specialization and generalization. 'Restriction' and 'narrowing' of meaning equal 'specialization,' while 'extension,' 'schematization,' and 'broadening' of meaning equal 'generalization.' Also, the meanings involved can be said to entertain relations of taxonomical subordination or superordination: in a taxonomy (a treelike hierarchical classification) of concepts, the specialized meaning is subordinate with regard to the original one, whereas the generalized meaning is superordinate with regard to the original.

Like specialization and generalization, it is convenient and customary to introduce metaphor and metonymy together, even though the relationship is not as close as with the former pair. *Metaphor* is then said to be based on a relationship of similarity between the old and the new reading, and *metonymy* on a relationship of contiguity. Current computer terminology yields examples of both types. The desktop of your computer screen, for instance, is not the same as the desktop of your office desk – except that in both cases, it is the space (a literal space in one case, a virtual one in the other) where you position a number of items that you regularly use or that urgently need attention. The computer desktop, in other words, is not literally a desktop in the original sense, but it has a functional similarity with the original: the computer reading is a metaphorical extension of the original office furniture reading. Functional similarities also underlie metaphorical expressions like *bookmark*,

clipboard, *file*, *folder*, *cut*, and *paste*. *Mouse*, on the other hand, is also metaphorically motivated, but here the metaphorical similarity involves shape rather than function. But now consider a statement to the effect that your desktop will keep you busy for the next 2 weeks, or that you ask aloud where your mouse has gone when you are trying to locate the pointer on the screen. In such cases, *desktop* and *mouse* are used metonymically. In the former case, it's not the virtual space as such that is relevant, but the items that are stored there. In the latter case, it's not the mouse as such (the thing that you hold in your hand) that you refer to, but the pointer on the screen that is operated by the mouse. The desktop and the stored items, or the mouse and the pointer, have a relationship of real-world connectedness that is usually captured by the notion 'contiguity.' When, for instance, one drinks a whole bottle, it is not the bottle but merely its contents that are consumed: *bottle* can be used to refer to a certain type of container, and the (spatially contiguous) contents of that container. When lexical semanticians state that metonymical changes are based on contiguity, contiguity should not be understood in a narrow sense as referring to spatial proximity only, but more broadly as a general term for various associations in the spatial, temporal, or causal domain.

With the shift toward a structuralist approach that occurred around 1930, lexical semantics switched from a preference for diachronic studies to a preference for synchronic studies. However, the poststructuralist cognitive approach (see the Section Cognitive Semantics below) provides a new impetus for historical lexical semantics (see, e.g., Geeraerts, 1997; Winters et al., 2010).

Structuralist Semantics

Inspired by the Saussurean conception of language, structural semantics originated as a reaction against prestructural historical semantics. The origins of structural semantics are customarily attributed to Trier (1931), but while Trier's monograph may indeed be the first major descriptive work in structural semantics, the first theoretical and methodological definition of the new approach is to be found in Weisgerber (1927), a polemical article that criticized historical linguistics on three points. First and foremost, because the vocabulary of a language is not simply an unstructured set of separate items, and because the meaning of a linguistic sign is determined by its position in the linguistic structures in which it takes part, the proper subject matter of semantics is not the atomistic changes of word meanings that historical semantics concentrated on, but the semantic structure of the language that demarcates the meanings of individual words with regard to each other. Second, because that structure is a linguistic rather than a psychological phenomenon, linguistic meanings should not be studied from a psychological perspective, but from a purely linguistic one. And third, because semantic change has to be redefined as change in semantic structures, synchronic semantics methodologically precedes diachronic semantics: the synchronic structures have to be studied before their changes can be considered. The realization of this attempt to develop a synchronic, nonpsychological, structural theory of semantics depends on the way in which the notion of semantic structure is conceived. In actual practice, there are mainly three distinct definitions of semantic structure that have been employed by structuralist semanticians. More particularly, three distinct kinds of structural relations among lexical items have been singled out as the proper methodological basis of lexical semantics. First, there is the relationship of semantic similarity that lies at the basis of semantic field analysis and componential analysis. Second, there are unanalyzed lexical relations such as synonymy, antonymy, and hyponymy. Third, syntagmatic lexical relations lie at the basis of a distributional approach to semantics.

Lexical Fields and Componential Analysis

A lexical field is a set of semantically related lexical items whose meanings are mutually interdependent. The single most influential study in the history of lexical field theory is Trier's (1931) monograph, in which he presents a theoretical formulation of the field approach and investigates how the terminology for mental properties evolves from Old High German up to the beginning of the thirteenth century. Theoretically, Trier emphasizes the structuralist insight that only a mutual demarcation of the words under consideration can provide a decisive answer regarding their exact value. Words should not be considered in isolation, but in their relationship to semantically related words: demarcation is always a demarcation relative to other words. Trier illustrates the idea with the image of a mosaic: the substance of human knowledge, the contents of cognition is divided by language into a number of adjoining small areas, in the way in which a mosaic divides two-dimensional space by means of contiguous mosaic stones.

While different conceptions of the notion 'lexical field' were suggested after Trier's initial formulation, the most important development is the emergence of componential analysis as a technique for formalizing the semantic relationships between the items in a field: once you have demarcated a lexical field, the internal relations within the field will have to be described in more detail. It is not sufficient to say that the items in the field are in mutual opposition – these oppositions will have to be identified and defined. Componential analysis is a method for describing such oppositions that takes its inspiration from structuralist phonology: just like phonemes are described structurally by their position on a set of contrastive dimensions, words may be characterized on the basis of the dimensions that structure a lexical field. Componential analysis provides a descriptive model for semantic content, based on the assumption that meanings can be described on the basis of a restricted set of conceptual building blocks – the semantic 'components' or 'features.' Componential analysis was developed in the second half of the 1950s and the beginning of the 1960s by European as well as American linguists, largely independently of each other. Its major impact would come from its incorporation into generative grammar: the publication of Katz and Fodor (1963) marked a theoretical migration of lexical semantics from a structuralist to a generativist framework. (This is a line of development that we will not pursue here, however, because it leads away from lexical semantics to grammatical and formal semantics.)

For a brief illustration of the principles of componential analysis, we may refer to the work of Pottier. Pottier (1964) provides an example of a componential semantic analysis in his description of a field consisting, among others, of the terms

siège, *pouf*, *tabouret*, *chaise*, *fauteuil*, and *canapé* (a subfield of the field of furniture terms in French). The word which acts as a superordinate to the field under consideration is *siège*, 'seating equipment with legs.' If we use the dimensions s1 'for seating,' s2 'for one person,' s3 'with legs,' s4 'with back,' s5 'with armrests,' s6 'of rigid material,' then *chaise* 'chair' can be componentially defined as (+s1, +s2, +s3, +s4, −s5, +s6), and *canapé* 'sofa' as (+s1, −s2, +s3, +s4, +s5, +s6), and so on.

Relational Semantics

Like componential analysis, relational semantics as introduced by Lyons (1963) develops the idea of describing the structural relations among related words, but restricts the theoretical vocabulary that may be used in such a description. In a componential analysis, there is no restriction on the features that enter into the description, and as in Pottier's example, these features are often of a 'real world' kind: they name properties of the things referred to rather than properties of the meanings as such. But structuralism is interested in the structure of the language rather than the structure of the world, and so it may want to use a different type of descriptive apparatus, one that is more purely linguistic. Relational semantics looks for such an apparatus in the form of sense relations like synonymy (identity of meaning) and antonymy (oppositeness of meaning): the fact that *aunt* and *uncle* refer to the same genealogical generation is a fact about the world, but the fact that *black* and *white* are opposites is a fact about words and language. Instead of deriving statements about the synonymy or antonymy of a word (and, in general, statements about the meaning relations it entertains) from a separate and independent description of the word's meaning, the meaning of the word could be defined as the total set of meaning relations in which it participates. A traditional (or perhaps naive) conception of synonymy would for instance describe the meaning of both *quickly* and *speedily* as 'in a fast way, not taking up much time,' and then conclude to the synonymy of both terms on the basis of the identity in their content description. Lyons by contrast deliberately eschews such content descriptions, and equates the meaning of a word like *quickly* with the synonymy relation it has with *speedily*, plus any other relations of that kind.

In the actual practice of relational semantics, 'relations of that kind' specifically include – next to synonymy and antonymy – relations of hyponymy (or subordination) and hyperonymy (or superordination), which are both based on taxonomical inclusion. The major research line in relational semantics involves the refinement and extension of this initial set of relations. The most prominent contribution to this endeavor after Lyons is found in Cruse (1986). Murphy (2003) is a thoroughly documented critical overview of the relational research tradition.

Distributional Semantics

Lexical fields as originally conceived are based on paradigmatic relations of similarity. One extension of the field approach, then, consists of taking a syntagmatic point of view. There are in fact two ways in which a syntagmatic analysis was suggested to be relevant for structuralist semantics. First, words may have specific combinatorial features which it would be natural to include in a field analysis. A verb like *to comb*, for instance, selects direct objects that refer to hair, or hairlike things, or objects covered with hair. Describing that selectional preference should be part of the semantic description of *to comb*. For a considerable period in the development of structural linguistics, these syntagmatic affinities received less attention than the paradigmatic relations, but in the 1950s and 1960s, the concept surfaced under different names in structuralist and generativist semantics. Firth (1957) for instance introduced the (now widely used) term *collocation*.

The second way in which the syntagmatics of lexical items could play a role in lexical field analysis is more radical than the mere incorporation of lexical combinatorics into the notion of lexical field: if the environments in which a word occurs could be used to establish its meaning, structuralist semantics could receive a firm methodological basis. The general approach of a distributionalist method is summarized by Firth's dictum: "You shall know a word by the company it keeps." A similar assumption is expressed by the 'distributional hypothesis' as formulated by Harris (1954): words that occur in the same contexts tend to have similar meanings. In the final decades of the twentieth century, major advances in the distributional approach to semantics were achieved by applying a distributional way of meaning analysis to large text corpora. Sinclair, a pioneer of the approach, developed his ideas (see Sinclair, 1991) through his work on the *Collins Cobuild English Language Dictionary*, for which a 20 million word corpus of contemporary English was compiled.

The distributional analysis in the corpus-based approach is not restricted to constituents and syntactic classes, but it takes into account the actual words in the context of which the target word appears. Firth remarked that part of the 'meaning' of *cows* can be indicated by such collocations as *They are milking the cows*, *Cows give milk*. This observation is taken as a methodological starting point: the words co-occurring with another one help to identify the properties of the word under scrutiny. An example (taken from Stubbs, 2002) may illustrate the basic idea. A classic example of homonymy in English is the item *bank*, which is either a financial institution or an area of sloping ground, specifically the raised ground on the side of the river or underneath a shallow layer of water. The sets of words that these two exemplars of *bank* normally occur with hardly overlap. Looking at compounds on the one hand, and on the other hand at co-occurring items within a few words to the left or right of *bank*, Stubbs comes up with the following lists:

bank account, bank balance, bank robbery, piggybank cashier, deposit, financial, money, overdraft, pay, steal

sand bank, canal bank, river bank, the South Bank, the Left Bank, Dogger bank, Rockall Bank, Icelandic Banks cave, cod, fish, float, headland, sailing, sea, water

The entities in the environment of the two homonyms appear to differentiate efficiently and effectively between the two meanings, and in that sense, a systematic analysis of the co-occurring items would appear to be an excellent methodological ground for lexical-semantic analysis.

In Sinclair's original conception, a collocational analysis is basically a heuristic device to support the lexicographer's

manual work. A further step in the development of the distributional approach was taken through the application of statistics, as when Church and Hanks (1990), working in the context of Sinclair's CoBuild project, introduced the Pointwise Mutual Information index as a statistical method for establishing the relevance of a collocation. Once a statistical measure of association of this kind was introduced, further possibilities opened up for quantifying the distributional approach; for instance, a whole variety of association measures has been suggested and researched. Also, the statistical turn in thinking about contextual distributions allowed for a rapprochement with the field of information retrieval and Natural Language Processing where so-called *word space models* constitute an advanced form of distributional corpus analysis, applied to problems like word sense disambiguation and synonym extraction (see Agirre and Edmonds, 2006). In this methodological form, the theoretical link between the distributional approach and structuralist semantics has all but disappeared: the method is very much a theoretically neutral one.

Cognitive Semantics

Compared to prestructuralist semantics, structuralism constitutes a move toward a more purely linguistic type of lexical semantics, focusing on the linguistic system rather than the psychological background or the contextual flexibility of meaning. With the poststructuralist emergence of cognitive semantics, the pendulum swings back to a position in which the distinction between semantics and pragmatics is not a major issue, in which language is seen in the context of cognition at large, and in which language use is as much a focus of enquiry as the language system. Cognitive semantics emerged in the 1980s as part of cognitive linguistics, a loosely structured theoretical movement that opposed the autonomy of grammar and the marginal position of semantics in the generativist theory of language. Important contributions to lexical semantics include prototype theory, conceptual metaphor theory, frame semantics, and the emergence of usage-based onomasiology.

Prototype Effects and Radial Sets

The prototype-based conception of categorization originated in the mid-1970s with Rosch's psycholinguistic research into the internal structure of categories. Rosch concluded that the tendency to define categories in a rigid way clashes with the actual psychological situation. Linguistic categories do not have sharply delimited borderlines. Instead of clear demarcations between equally important conceptual areas, one finds marginal areas between categories that are only unambiguously defined in their focal points. This observation was taken over and elaborated by cognitive semanticists. Specifically, it was applied not just to the internal structure of a single word meaning, but also to the structure of polysemous words, i.e., to the relationship between the various meanings of a word. A popular descriptive model for charting that relationship is the radial set model, which graphically represents the way in which less central meanings branch out from the prototypical, core reading. Prototype theory, arguably the most influential contribution of cognitive linguistics to lexical semantics, receives a separate treatment in the article Prototype Theory in Linguistics.

Conceptual Metaphor and Metonymy

The interest in the internal structure of lexical items, as illustrated and embodied by the research covered in the previous subsection, automatically entails an interest in the semantic relations that interconnect the various readings of an item. Within a radial set for instance, the elements are not only connected through the fact that a less prototypical sense derives from a more central one, but also through a specific mechanism of semantic extension: similarity, or metaphor, or metonymy – basically, any of the mechanisms of meaning change that constituted the focus of historical semantics as discussed in the Section Prestructuralist Historical Semantics above.

Metaphor, in particular, constitutes a major area of investigation for cognitive semantics. The major impetus here came from Lakoff and Johnson (1980), a book that worked like an eye-opener for a new generation of linguists. In the linguistic climate of the 1970s, dominated by the formal framework of generative grammar, semantics seemed a peripheral issue, but *Metaphors We Live By*, more perhaps than the other foundational publications in cognitive semantics, was instrumental in putting semantics back on the research agenda.

Conceptual metaphor theory, the approach introduced by Lakoff, includes two basic ideas: first, the view that metaphor is a cognitive phenomenon rather than a purely lexical one; second, the view that metaphor should be analyzed as a mapping between two domains. To illustrate the first point, metaphor comes in patterns that transcend the individual lexical item. A typical example is the following.

LOVE IS A JOURNEY

Look how far we've come. We are at a crossroads. We'll just have to go our separate ways. We cannot turn back now. We are stuck. This relationship is a dead-end street. I don't think this relationship is going anywhere. It's been a long, bumpy road. We have gotten off the track.

The second pillar of conceptual metaphor theory is the analysis of the mappings inherent in metaphorical patterns. Metaphors conceptualize a target domain in terms of the source domain, and such a mapping takes the form of an alignment between aspects of the source and target. For LOVE IS A JOURNEY, for instance, the following correspondences hold.

Source	*Target*
The travelers	The lovers
The means of transport	The relationship itself
The journey	The development of the relationship
The obstacles encountered	The difficulties experienced
Decisions about which way to go	Choices about what to do
The destination of the journey	The goals of the relationship

The appeal of Conceptual Metaphor Theory engendered a revival of the interest in metonymy. In Lakoff and Johnson (1980), metonymy figured next to metaphor as one of the conceptual mechanisms behind the semantic structure of language. That clearly should not come as a surprise: an approach that is interested in the semantic mechanisms behind language use and linguistic structures is likely to rediscover the traditional mechanisms of semantic extension. Lakoff and Johnson list a number of metonymic patterns that might have been taken straightforwardly from a prestructuralist treatise on semantic change of the type that was discussed in the Section Prestructuralist Historical Semantics above. Lakoff and Johnson emphasize that such metonymic patterns are conceptual and not purely linguistic, in much the same way that metaphorical concepts are. In the first place, metonymical concepts allow us to think of one thing in terms of its relation to something else. In that sense, we can distinguish a source and target in the description of metonymy just like we can for metaphors. In the second place, metonymies are systematic in the sense that they form patterns that apply to more than just an individual lexical item.

Thus, the *bottle* example mentioned above exhibits the name of a container (source) being used for its contents (target), a pattern that can be abbreviated as 'container for contents.' Making use of this abbreviated notation, other common types of metonymy are the following: 'a spatial location for what is located there' (*the whole theater was in tears*); 'a period of time for what happens in that period, for the people that live then, or for what it produces' (*the nineteenth century had a nationalist approach to politics*); 'a material for the product made from it' (*a cork*); 'the origin for what originates from it' (*astrakhan, champagne, emmental*); 'an activity or event for its consequences' (when the *blow* you have received hurts, it is not the activity of your adversary that is painful, but the physical effects that it has on your body); 'an attribute for the entity that possesses the attribute' (*majesty* does not only refer to 'royal dignity or status,' but also to the sovereign himself); and of course 'part for whole' (*a hired hand*). The relations can often work in the other direction as well. *To fill up the car*, for instance, illustrates a type 'whole for part': it's obviously only a part of the car that gets filled. For the current state of affairs in metonymy research from a cognitive semantic point of view, see Benczes et al. (2011).

Frame Semantics

The models presented in the previous two subsections illustrate two important features of cognitive semantics: the interest for the contextual flexibility of meaning (as reflected in prototype-based models for describing polysemy and semantic nuances), and the tight links between language and thinking in general (as illustrated by conceptual metaphor and metonymy research). A third feature of cognitive semantics that needs to be mentioned in this respect is its interest in the way our knowledge of the world is organized in larger 'chunks of knowledge,' and how these interact with language. The most articulate model in this respect is Fillmore's frame theory (Fillmore and Atkins, 1992). Frame theory is specifically interested in the way in which language may be used to perspectivize an underlying conceptualization of the world: it's not just that we see the world in terms of conceptual models, but those models may be verbalized in different ways. Each different way of bringing a conceptual model to expression so to speak adds another layer of meaning: the models themselves are meaningful ways of thinking about the world, but the way we express the models while talking adds perspective. This overall starting-point of Fillmorean frame theory leads to a description on two levels. On the one hand, a description of the referential situation or event consists of an identification of the relevant elements and entities and the conceptual role they play in the situation or event. On the other hand, the more purely linguistic part of the analysis indicates how certain expressions and grammatical patterns highlight aspects of that situation or event.

To illustrate, we may have a look at the standard example of frame theory, the COMMERCIAL TRANSACTION frame. The commercial transaction frame involves words like *buy* and *sell*. The commercial transaction frame can be characterized informally by a scenario in which one person gets control or possession of something from a second person, as a result of a mutual agreement through which the first person gives the second person a sum of money. Background knowledge involved in this scenario includes an understanding of ownership relations, a money economy, and commercial contracts. The categories that are needed for describing the lexical meanings of the verbs linked to the commercial transaction scene include Buyer, Seller, Goods, and Money as basic categories. Verbs like *buy* and *sell* then each encode a certain perspective on the commercial transaction scene by highlighting specific elements of the scene. In the case of *buy*, for instance, the buyer appears in the participant role of the agent, for instance as the subject of the (active) sentence. In active sentences, the goods then appear as the direct object; the seller and the money appear in prepositional phrases: *Paloma bought a book from Teresa for €30*. In the case of *sell* on the other hand, it is the seller that appears in the participant role of the agent: *Teresa sold a book to Paloma for €30*.

Basic Levels and Onomasiological Salience

Possibly the major innovation of the prototype model of categorization is to give salience a place in the description of semasiological structure: next to the qualitative relations among the elements in a semasiological structure (like metaphor and metonymy), a quantifiable center–periphery relationship is introduced as part of the architecture. But the concept of salience can also be applied to the onomasiological domain. (Semasiology looks at the relationship between words and meaning with the word as starting point: it is basically interested in the polysemy of words. Onomasiology takes the converse perspective: given a concept to be expressed or a thing to be categorized, what options does a language offer, and how are the choices made?)

The initial step in the introduction of onomasiological salience is the *basic-level hypothesis*. The hypothesis is based on the ethnolinguistic observation that folk classifications of biological domains usually conform to a general organizational principle, in the sense that they consist of five or six taxonomical levels (Berlin, 1978). The basic-level hypothesis embodies a notion of onomasiological salience because it is a hypothesis about alternative categorizations of referents: if a particular referent (a particular piece of clothing) can be alternatively

categorized as a garment, a skirt, or a wraparound skirt, the choice will be preferentially made for the basic-level category 'skirt.' But differences of onomasiological preference also occur *among* categories on the same level in a taxonomical hierarchy. If a particular referent can be alternatively categorized as a wraparound skirt or a miniskirt, there could just as well be a preferential choice: when you encounter something that is both a wraparound skirt and a miniskirt, the most natural way of naming that referent in a neutral context would probably be 'miniskirt.' If, then, we have to reckon with intralevel differences of salience next to interlevel differences, the concept of onomasiological salience has to be generalized in such a way that it relates to individual categories at any level of the hierarchy.

This notion of *generalized onomasiological salience* was first introduced in Geeraerts et al. (1994). Using corpus materials, this study established that the choice for one lexical item rather than the other as the name for a given referent is determined by the semasiological salience of the referent (i.e., the degree of prototypicality of the referent with regard to the semasiological structure of the category), by the overall onomasiological salience of the category represented by the expression, and by contextual features of a classical sociolinguistic and geographical nature, involving the competition between different language varieties. By zooming in on the latter type of factor, a further refinement of the notion of onomasiological salience is introduced, in the form of the distinction between *conceptual* and *formal* onomasiological variation. Whereas conceptual onomasiological variation involves the choice of different conceptual categories for a referent (like the examples presented so far), formal onomasiological variation merely involves the use of different synonymous names for the same conceptual category. The names *jeans* and *trousers* for denim leisure wear trousers constitute an instance of conceptual variation, for they represent categories at different taxonomical levels. *Jeans* and *denims*, however, represent no more than different (but synonymous) names for the same denotational category.

It is further observed that all types of lexical variation may be subject to contextual, 'lectal' variation. Different dialects may use words in different meanings (semasiological variation). Experts are more likely to use specific, technical terms than laymen (conceptual onomasiological variation). And denotationally synonymous expressions may have different sociolinguistic distributions (formal onomasiological variation). This recognition then leads to an upsurge of sociolexicological studies, as part of cognitive sociolinguistics, i.e., the study of language variation from a cognitive linguistic point of view: see Kristiansen and Dirven (2008).

Overview and Outlook

The development of linguistic lexical semantics appears to be characterized by a succession of distinct theories. Crucially, each of the competing theoretical frameworks has contributed to the descriptive expansion of lexical semantics, i.e., each of them has drawn the attention to specific phenomena. Focusing on the major phenomena, these contributions successively include the links between the various senses of words in prestructuralist historical semantics, the semantic relationships within the vocabulary in the structuralist era, and the importance of semasiological and onomasiological salience effects in cognitive semantics. Regardless of the theoretical oppositions, these phenomena all belong to the descriptive scope of current lexical semantics: the emergence of new points of attention has not made the older topics irrelevant. To give just one example, prototype theory may trigger a reconsideration of what it means for two words to be synonymous, but the notion of synonymy remains a serious topic of investigation for lexical semantics.

Next to the theoretical shifts and the descriptive expansion, a methodological evolution characterizes the historical path of lexical semantics: a structuralist focus on syntagmatic relations, boosted by the growing availability of text corpora, has led to a distributional method of corpus-based semantics. Together with the theoretical and descriptive innovations introduced by cognitive semantics, this new quantitative and computational method is likely to shape the evolution of lexical semantics in the following years.

See also: Etymology; Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects; Lexicon; Parts of Speech; Second Language Acquisition; Semantics; Visual Word Recognition, Neurocognitive Psychology of.

Bibliography

Agirre, Eneko, Edmonds, Philip, 2006. Word Sense Disambiguation. Algorithms and Applications. Springer, Berlin.
Benczes, Réka, Barcelona, Antonio, Mendoza Ibáñez, Francisco Ruiz de (Eds.), 2011. Defining Metonymy in Cognitive Linguistics. Towards a Consensus View. John Benjamins Publishing Company, Amsterdam.
Berlin, Brent, 1978. Ethnobiological classification. In: Rosch, Eleanor, Lloyd, Barbara B. (Eds.), Cognition and Categorization. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 9–26.
Bréal, Michel, 1897. Essai de sémantique. Science des significations. Hachette, Paris.
Church, Kenneth, Hanks, Patrick, 1990. Word association norms, mutual information, and lexicography. Computational Linguistics 16, 22–29.
Cruse, D. Alan, 1986. Lexical Semantics. Cambridge University Press, Cambridge.
Cruse, D. Alan, 2010. Meaning in Language. An Introduction to Semantics and Pragmatics, third ed. Oxford University Press, Oxford.
Fillmore, Charles J., Atkins, B.T. Sue, 1992. Toward a frame-based lexicon: the semantics of 'risk' and its neighbors. In: Lehrer, Adrienne, Kittay, Eva Feder (Eds.), Frames, Fields and Contrasts. New Essays in Semantic and Lexical Organization. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 75–102.
Firth, John R., 1957. Papers in Linguistics 1934–51. Oxford University Press, Oxford.
Geeraerts, Dirk, 1997. Diachronic Prototype Semantics. A Contribution to Historical Lexicology. Clarendon Press, Oxford.
Geeraerts, Dirk, 2010. Theories of Lexical Semantics. Oxford University Press, Oxford.
Geeraerts, Dirk, Grondelaers, Stefan, Bakema, Peter, 1994. The Structure of Lexical Variation. Meaning, Naming, and Context. Mouton de Gruyter, Berlin.
Harris, Zellig, 1954. Distributional structure. Word 10, 146–162.
Katz, Jerrold J., Fodor, Jerry A., 1963. The structure of a semantic theory. Language 39, 170–210.
Kristiansen, Gitte, Dirven, René (Eds.), 2008. Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems. Mouton de Gruyter, Berlin/New York.
Lakoff, George, Johnson, Mark, 1980. Metaphors We Live By. University of Chicago Press, Chicago.
Lyons, John, 1963. Structural Semantics. Blackwell, Oxford.
Murphy, M. Lynne, 2003. Semantic Relations and the Lexicon. Antonymy, Synonymy, and Other Paradigms. Cambridge University Press, Cambridge.
Murphy, M. Lynne, 2010. Lexical Meaning. Cambridge University Press, Cambridge.

Paul, Hermann, 1920. Prinzipien der Sprachgeschichte, fifth ed. Max Niemeyer Verlag, Halle.
Pottier, Bernard, 1964. Vers une sémantique moderne. Travaux de linguistique et de littérature 2, 107–137.
Riemer, Nick, 2010. Introducing Semantics. Cambridge University Press, Cambridge.
Saeed, John I., 2009. Semantics, third ed. Wiley-Blackwell, Malden.
Sinclair, John M., 1991. Corpus, Concordance, Collocation. Oxford University Press, Oxford.
Stern, Gustaf, 1931. Meaning and Change of Meaning, with Special Reference to the English Language. Elanders Boktryckeri Aktiebolag, Göteborg.
Stubbs, Michael, 2002. Words and Phrases. Corpus Studies of Lexical Semantics. Blackwell, Oxford.
Trier, Jost, 1931. Der deutsche Wortschatz im Sinnbezirk des Verstandes. Die Geschichte eines sprachlichen Feldes I. Von den Anfängen bis zum Beginn des 13. Jhdts. Winter, Heidelberg.
Weisgerber, Leo, 1927. Die Bedeutungslehre – Ein Irrweg der Sprachwissenschaft? Germanisch-Romanische Monatsschrift 15, 161–183.
Winters, Margaret E., Tissari, Heli, Allan, Kathryn (Eds.), 2010. Historical Cognitive Linguistics. De Gruyter Mouton, Berlin/New York.

Lexicology and Lexicography

Wolfgang Klein, Max Planck Institut für Psycholinguistik, Nijmegen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Lexicology is the scientific investigation of the lexicon of a language, including, for example, its historical development, its social stratification, its quantitative composition, or the way in which some thematic area is encoded. Lexicography, the oldest subdiscipline of linguistics, deals with the compilation of dictionaries. There are many types of dictionaries, depending mainly on which lexical units are included, and which of their properties – such as sound, spelling, grammatical features, meaning, etymology, and others – are described. Two technological innovations are crucial to the presentation of lexical information – the art of printing, which led to the familiar printed dictionary, and the computer, which makes it possible to build up and to exploit large data sources and to develop complex digital lexical systems.

Introduction

Each language has a lexicon and a grammar, i.e., a set of elementary expressions and a set of rules according to which complex expressions are constructed from simpler ones. Some of these rules form complex words; others operate beyond the boundaries of the word, thus producing phrases and sentences. These distinctions, familiar from the days of the Greek grammarians, are not always clear-cut, for at least two reasons. First, the notion of 'word' is not very well defined. Second, there are complex expressions, whose meaning is more or less predictable from the meaning of its components, whereas this is not true for other complex expressions. The former are said to be 'compositional,' whereas the latter are 'lexicalized'; slightly different terms to characterize this opposition are 'productive' vs 'idiomatic,' and 'free' vs 'fixed'; in each case, the distinction is gradual. Lexicalization is rarely observed for inflected words (a possible exception are 'participles' such as *crooked* in *a crooked street*), but very frequent for compound words, such as *landlord* or (*to*) *withdraw*, or phrases such as *to kick the bucket*, which has a compositional as well as a lexicalized reading. Do lexicalized expressions belong to the lexicon of a language or to its grammar? There is no straightforward answer; their form is complex and rule based, their meaning is not. Therefore, it is useful to take the term 'lexicon' in a somewhat broader sense; it contains all elementary expressions (= lexicon in the narrower sense) as well as those expressions which are compound in form but not accordingly in meaning (*see* Lexicon). The scientific investigation of the lexicon in this sense is usually called lexicology; it includes, for example, the historical development of the lexicon, its social stratification, its quantitative composition, or the way in which some subfield is encoded in lexical items (e.g., 'terminology of hunting,' 'verbs of movement'). Lexicography, by contrast, deals with the compilation of dictionaries. There is considerable overlap between both disciplines, and in fact, not all authors make such a terminological distinction.

The Lexicon

The lexicon of a language is stored primarily in the head of its speakers, and for most of the history of mankind, it was only stored there. We do not know what form the 'mental lexicon' has (*see* Psycholinguistics: Overview). There is agreement, however, that it consists of individual lexical units which are somehow interrelated to each other. There is no generally accepted term for lexical units. The familiar term 'word' is both too broad and too narrow; one would not want to consider *goes* as a lexical unit, although it is a word, whereas expressions such as (*to*) *knock out* or *red herring* are lexical units but consist of several words. Other terms occasionally found are 'lexeme,' 'lemma,' or 'lexical entry,' but since these are also used in other ways, it is probably best to speak of lexical units.

It is important to distinguish between a lexical unit and the way in which it is named. The word *house* in a dictionary, followed by all sorts of explanations, is not the lexical unit – it is a name for such a unit. The lexical unit itself is a bundle of various types of properties. These include:

1. phonological properties, which characterize how the lexical unit is pronounced; they include sounds, syllabic structure, lexical accent and, in some languages, lexical tone;
2. graphematic properties, which characterize how the lexical unit is written (*see* Spelling, Psychology of);
3. morphosyntactic properties, which characterize how the unit can become part of more complex expressions; typically, they concern word class, inflectional paradigm, and government relations;
4. semantic properties, which concern the 'lexical meaning' of the unit, i.e., the contribution which it makes to the meaning of the construction in which it occurs.

Some of these properties may be absent. This is most obvious for graphematic properties, since not all languages are written. There are a few lexical units without lexical meaning, such as the expletive *there* in English. Many linguists also stipulate 'zero elements,' i.e., units with morphosyntactic and semantic properties but without phonological properties (such as 'empty pronouns'); but these are normally treated in the grammar rather than in the lexicon.

Whereas these four types of properties are the defining characteristics of a lexical unit, other information may be associated with it, for example, its etymology, its frequency of usage, its semantic counterpart in other languages, or encyclopedic knowledge (thus, it is one thing to know the

meaning of *bread* and a different thing to know various sorts of bread, how it is made, its price, its role in the history of mankind, etc.).

The lexical units of a lexicon are in many ways interrelated. They may share some phonological properties (for example, they may rhyme with each other), they may belong to the same inflectional paradigm, they may have the opposite meaning ('antonyms,' such as *black* and *white*), approximately the same meaning ('synonyms,' such as *to begin* and *to start*), or when complex in form they may follow the same construction pattern. Lexicological research is often oriented toward these interrelations, whereas lexicography tends to give more weight to the lexical unit in itself. In general, there is much more lexicographical than lexicological work (for a survey of the latter, see Cruse et al., 2005); in fact, if there is any piece of linguistic description for some language, it is probably an elementary bilingual dictionary. The depth of this work varies massively not only across languages, but also with respect to the particular lexical properties. Whereas the phonological, graphematic, and morphosyntactic features of the lexicon in Latin, English, French, and some dozen other languages with a comparable research tradition are fairly well described, there is no theoretically and empirically satisfactory analysis of the semantics of the lexicon for any language whatsoever. This has three interrelated reasons. First, there is no well-defined descriptive language which would allow the researcher to represent the meaning of some lexical unit, be it simple or compound; the most common practice is still to paraphrase it by an expression of the same language. Second, there is no reliable and easily applicable method of determining the lexical meaning of some unit; the most common way is to look at a number of occurrences in ongoing text and to try to understand what it means. Third, the relation between a particular form and a particular meaning is hardly ever straightforward; this is strikingly illustrated by a look at what even a medium-sized English dictionary has to say about the meaning of, for example, *on*, *sound*, *cast*, or (*to*) *put*. As a rule, there is not just one lexical meaning, but a whole array of uses, which are more or less related to each other. This is not merely a practical problem for the lexicographer; it also casts some doubt on the very notion of 'lexical unit' itself (*see* Lexical Semantics).

Making Dictionaries

Lexicographers often consider their work to be more of an art or a craft than a science (see Svensén, 2009). This does not preclude a solid scientific basis, but it reflects the fact that their concrete work depends largely on practical skills such as being 'a good definer,' on the one hand; and that it is to a great extent determined by practical, often commercial, concerns, on the other. Dictionaries are made for users, and they are intended to serve specific purposes. Their compilation requires a number of practical decisions.

Which Lexical Units Are Included?

Languages are neither well defined nor uniform entities; they change over time, and they vary with factors such as place, social class, or area talked about. A great deal of this variation is lexical. It is not possible nor would it be reasonable to cover this wealth in a single dictionary. Large dictionaries contain several hundred thousands of 'entries'; since idiomatic expressions are usually listed under one of their components (such as *to kick the bucket* under (*to*) *kick*), they contain many more lexical units, perhaps up to one million. But, even so, they are by no means exhaustive. The second edition of the *Deutsches Wörterbuch* (see Section Shortcomings), the largest dictionary of German, covers less than 25% of the lexical units found in the sources, and these sources are quite restricted themselves.

Which Lexical Properties Are Described?

Just as it is impossible to include all lexical units of a language in a dictionary, it is neither possible nor desirable to aim at a full description of those which are included. Since a dictionary is traditionally a printed book, the graphematic properties of the unit (its 'spelling') are automatically given. Among the other defining properties, meaning is traditionally considered to be most important. Samuel Johnson's dictionary from 1755 (see Section Shortcomings) defines 'dictionary' as "A book containing the words of any language in alphabetical order, with explanations of their meaning." But Johnson also noted which syllable carries the main stress, and he gave some grammatical hints. In general, however, information on phonological properties was rare up to the end of the nineteenth century, and information on grammatical properties is usually still very restricted in nonspecialized dictionaries. But there are, of course, also dictionaries which specifically address these properties as well as some of the nondefining properties associated with a lexical entry, such as its origin (etymological dictionary) or, above all, its equivalent in other languages ('bilingual dictionary').

What Is the Description Based Upon?

Usually, two types of sources are distinguished: 'primary sources' are samples of text in which the unit is used, 'secondary sources' refer to prior work of other lexicographers (and lexicologists). In fact, there is a third source, normally not mentioned in the theory of lexicography: This is the lexicographer's own knowledge of the language to be described, including his or her views on what is 'good' language. In practice, the bulk of a new dictionary is based on older dictionaries. This is always immoral and often illegal, if these are simply copied; but on the other hand, it would be stupid and arrogant to ignore the achievements of earlier lexicographers.

How Is the Information Presented?

A dictionary consists of lexical entries arranged in some conventional order. Normally, an entry combines several lexical units under a single 'headword'; thus, all lexical units which include the word *put* may be listed under this headword, forming a kind of nest with an often very complex microstructure. We are used to alphabetically ordered dictionaries; but there are other possibilities, for example, by thematic

groups or by first appearance in written documents. Languages without alphabetic writing require different principles; in Chinese, for example, entries are usually arranged by subcomponents of the entire character and by the number of strokes.

These four questions can be answered in very different ways, resulting in very different types of dictionaries (see the survey in Hausmann et al., 1991: 968–1573).

History

The first lexicographic documents are lists of Sumerian words (up to 1400) with their Akkadian equivalents, written in cuneiform script on clay tablets about 4700 years ago. The practice of compiling such word lists was continued throughout Antiquity and the Middle Ages; thus, the oldest document in German, the *Abrogans* (written around 765), is an inventory of some Latin words with explanations in German. Usually, these 'glossaries' did not aim at a full account of the lexicon; they simply brought together a number of words which, for one reason or another, were felt to be 'difficult,' and explained them either by a more familiar word in the same language or by a translation. Words were ordered alphabetically, by theme, or not at all. But there are also more systematic attempts, such as the *Catholicon*, a mixture of encyclopedia and dictionary which, compiled around 1250, was the first printed lexical work in Europe (Mainz 1460).

In the sixteenth century, two developments led to major changes. The first of these was the invention of printing by Gutenberg. By 1500, virtually all classical authors were available in print, thus offering a solid basis for systematic lexical accounts of Latin and Greek, such as Calepinus' *Dictionarium* (Dictionary) (1502), soon to be followed by two early masterpieces: Robert Etienne's *Dictionarium seu Latinae Linguae Thesaurus* (*Dictionary or Thesaurus of the Latin Language*) (Paris 1531) and Henri Etienne's *Thesaurus Graecae Linguae* (*Thesaurus of the Greek Language*) (Paris 1572). The second major development was the slow but steady rise of national languages. Since early Italian, French, English, or German were hardly codified, a major aim of the first dictionaries in these languages was to give them clear norms. In some countries, national Academies were founded to this end. The outcome were dictionaries with a strongly normative, often puristic, stance, such as the *Vocabulario degli Academici della Crusca* (*Vocabulary of the Members of the Academia della Crusca*) (Venice 1612), the *Dictionnaire de l'Académie Française* (*Dictionary of the French Academy*) (Paris 1694), and the *Diccionario de autoridades publicado por la Real Academia Espanola* (*Dictionary of the Authorities published by the Royal Spanish Academy*) (1726–1739). The bulk of lexicographic work, however, was always done by enterprising publishers and engaged individuals, such as Dr Samuel Johnson. Helped by six assistants, he produced *A Dictionary of the English Language* (London 1755), the first scholarly description of the English vocabulary, in less than 8 years. It surpassed all its predecessors, including Bailey's *Dictionarium Britannicum* (*British Dictionary*) from 1736, which Johnson took as his point of departure, by the systematic use of quotations, taken from the 'best writers,' and by his brilliant, sometimes somewhat extravagant, definitions (not everybody would dare to characterize patriotism as 'the last refuge of a scoundrel').

The rise of historical–comparative linguistics in the early nineteenth century led to an enormous increase in grammatical and lexical knowledge. The first dictionary which tried to cover this knowledge was the *Deutsches Wörterbuch* (*German Dictionary*) by Jacob Grimm and (to a lesser extent) his brother Wilhelm. Its first fascicle appeared in 1852, after about 10 years of preparatory work, in which the Grimms were helped by about 100 scholars providing excerpts ("covering my desk like snowflakes," Jacob Grimm). At that time, it was already clear that the original plan of six to seven volumes, to be finished within 10–12 years, was unrealistic. The Grimms finished only letters A to (most of) F, and the final folio volume (of altogether 32) appeared in 1961. This long duration, as well as the varying talents and preferences of the contributors, has led to many inconsistencies; some entries got out of balance (no less than 60 folio pages are devoted to the single word *Geist* (mind, spirit, ghost)); still, it is an incommensurable source of lexical information.

The work of the Grimms inspired a number of similar ventures, such as Emile Littré's masterly *Dictionnaire de la langue française* (*Dictionary of the French Language*) (1863–73), which is much shorter, but also much more consistent; Matthias de Vries and his numerous successors' voluminous *Woordenboek der Nederlandsche Taal* (*Dictionary of the Dutch Language*) (1864–1998), and finally *A New English Dictionary on a Historical Basis* (1884–1928), generally referred to as the *Oxford English Dictionary* (OED). It was initiated in 1857 by the philologist and churchman Richard Trench; in 1860, members of the Philological Society started to collect excerpts; in 1879, the Clarendon Press appointed James Murray as the Principal Editor. The first fascicle appeared in 1882, and the whole work was completed in 1928, 13 years after Murray's death. More than 200 scholars were involved in its production, more than 2000 people are known to have contributed excerpts. The OED is not without flaws, even in its revised edition, which appeared in 1989 in print and in 1992 on CD-ROM; but among all attempts to describe the lexicon of a language, it comes closest to falsify what Dr Johnson stated in the preface to his own dictionary: "Every other author may aspire to praise; the lexicographer can only hope to escape reproach." (For a comprehensive survey of lexicographic work across languages, see Hausmann et al., 1991: 1679–2710, 2949–3119.)

Shortcomings

At the end of the twentieth century, the art of making dictionaries had reached a remarkable degree of perfection. But two reservations are in order here. First, this perfection only holds for about 5% of the about 7000 languages of the world; for all others, we only have only more or less fragmentary and often unreliable accounts of their lexical repertoire – if any. Second, even the best dictionaries have individual as well as systematic shortcomings. Lexicographers are human beings, and thus, there are individual errors, poor definitions, or arguable decisions that make their achievements imperfect and popular targets of criticism. But there are

also deficiencies that result from the very nature of the printed book itself. These include:

1. Size restrictions. In a comprehensive corpus of present-day German, we find about five million lexical units (not counting dialects); the largest dictionary of present-day German describes about 170 000 lexical units – in 10 volumes, thus less than 5% of the words that are actually used.
2. Inadequacies of description. As a rule, the information given in a dictionary grossly underspecifies the various properties of a lexical item. Phonological properties are characterized by a phonetic script, such as the International Phonetic Alphabet (IPA); but from such a transcription, most users can only guess what a word really sounds like. The grammatical properties of a word are normally much more complex than what can be described by simple indications such as 'noun' or 'transitive verb.' Hardly anyone understands which familiar English words are targeted at by the meaning descriptions "Before the time in question, beforehand; by now; as early or as soon as this" or "diversion, amusement, sport; also, boisterous, jocularity or gaiety, drollery. Also, a source or cause of amusement or pleasure." (This is how the OED – the best dictionary of the world – describes the main readings of *already* and *fun*, respectively.)
3. Exclusion of overarching lexical properties. The vocabulary of a language is not just an aggregation of individual words, it is a highly complex structure, defined by the many relations which obtain between the phonological, grammatical, and semantical properties of all words (*see* Lexicon). A traditional dictionary with its focus on individual lexical units can hardly account for this aspect of a lexicon.

Since these deficiencies reflect inherent limits of a printed book – and this is what we normally understand a dictionary to be – it is practically impossible to overcome them within the limits of that format. Other formats must be developed to compile and to present lexical information. It is the advent of digital methods which begins to make this possible.

Digital Lexicography

Over the last decades, digital methods began to change, and in fact to overthrow, traditional lexicography. Five main lines in this development can be distinguished. First, it was necessary to create specific dictionaries for various tasks of natural language processing, such as machine translation, man–machine dialogue systems, or automatic text analysis. These 'dictionaries' are sometimes based on a machine-readable version of an existing printed dictionary. But the lexical information is stored on the computer, and it is given by some formal representation, rather than in a natural language. In the event, this led to completely new ways to describe the meaning of lexical units: the classical paraphrase in a few words was replaced by a netlike structure, in which various elements which play a role for that word are connected by semantic relations such as 'is a,' 'is a part of,' 'entails,' etc. The best known of these systems are WordNet (developed in Princeton) and FrameNet (developed in Berkeley).

Second, there is an increasing number of databases, mostly accessible via the Internet, which provide computer-generated information about various lexical properties of a language, for example, the frequency of words (or parts of words such as graphemes and morphemes) or of word co-occurrences. Good examples are the databases at the Brigham Young University (http://view.byu.edu) for English, the *Centre National de resources textuelles et lexicales* (National Center for textual and lexical resources) (www.cnrtl.fr) for French or *Wortschatz* (Vocabulary) (http://corpora.informatik.uni-leipzig.de), which provides statistical information about word frequencies and collocations in about 230 languages. Such databases often exploit huge corpora, sometimes in the range of billions of words. But these corpora are mostly not carefully compiled but based on material that happens to be available in digital form (like recent newspapers or Wikipedia texts), and, more importantly, their analysis is based on bare written word forms, without looking at the meaning. Thus, homographs such as *bear* and *bear* or *left* and *left* are considered as the same lexical item; this may limit their reliability for lexicology and lexicography.

Third, existing printed dictionaries were transferred to the computer and adapted to the purposes of the common user. Such a transfer offers several advantages: Search is faster and more exhaustive; the information is easier to revise and to update; and it is possible to add information not available in book format, for example, spoken sound instead of phonetic transcriptions. There are an increasing number of mostly free web sites which contain links to or even host dozens of – mostly bilingual – dictionaries, such as www.lexilogos.com, dictionary.cambridge.org, or www.dict.cc. They offer a cheap, fast, and easy source of lexical information, whose quality, though, varies considerably. Revisions and amendments, if there are any, are largely in the hands of volunteers.

Fourth, computers were increasingly used as tools in the *production* of a new dictionary. Rather than having a number of people read through books and newspapers and make excerpts of all occurrences which look interesting, it is now possible to compile huge text corpora that cover all varieties of a language, to scan these texts for all occurrences of words or word combinations, to sort these occurrences by various criteria, to link them to other occurrences, to add as much context as needed, etc. (Atkins and Rundell, 2008; *see* Corpus Linguistics). The first OED was based on about five million excerpts, mostly handwritten on paper slips. A computer can easily process corpora of several hundred million words, i.e., several hundred million occurrences; new sources can rapidly be added. This allows a much broader and much more representative coverage of a lexicon than ever. But electronic corpora only provide the raw material; it still awaits lexical analysis. This analysis can be facilitated by computer tools, also; but no computer can tell us what a word means in a particular context. If only 1 min is devoted to each occurrence in a one billion word corpus, it would take 100 lexicographers 100 years to go through it. This means that printed dictionaries can never reflect the wealth of information accessible in large corpora, because they are outdated when the lexicographers come to the end of their analysis. Therefore, the most recent version of the OED is no longer a printed book but a regularly updated and expanded computer

dictionary that is only searchable over the Internet (www.oed.com).

The fifth and last development is complex *digital lexical systems* which integrate various sources of information. They consist of (1) a computer-accessible and expandable corpus, (2) a set of tools, which allow, for example, not only the search for certain items but also statistical analysis or the determination of the first occurrence, or just look up of a number of occurrences that illustrate the use of an unknown word, and (3) a selective but steadily proceeding lexical analysis of the corpus by experienced lexicographers; in that respect, they are in line with the established lexicographical tradition. According to the need of the users and the means that are available, this analysis may vary in depth and coverage. Thus, it is possible to add spoken forms in various dialects, information about word classes, or the semantic analysis of some subset of lexical units, say all prepositions or all morphologically simple verbs. Similarly, translation equivalents can be added. Unlike printed dictionaries, such a lexical retrieval system, such as the *Digitale Wörterbuch der Deutschen Sprache* (*Digital Dictionary of the German Language*) (www.dwds.de), will never come to an end, it is steady work in progress to which many can contribute and which will give us a deeper and broader understanding of the lexicon than any other method.

See also: Corpus Linguistics; Encyclopedias, Handbooks, and Dictionaries; Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects; Lexical Semantics; Lexicon.

Bibliography

Atkins, Rundell, 2008. The Oxford Guide to Practical Lexicography. Oxford University Press, Oxford.

Cruse, D. Alan, Hundsnurscher, Franz, Job, Michael, Lutzeier, P.R. (Eds.), 2005. Lexikology – Lexicology. de Gruyter, Berlin.

Hausmann, Franz Joseph, Reichmann, Oskar, Wiegand, Herbert Ernst, Zgusta, Ladislav (Eds.), 1991. Wörterbücher – Dictionaries – Dictionnaires. De Gruyter, Berlin.

Jackson, Howard, 2013. The Bloomsbury Companion to Lexicography. Bloomsbury, London.

McArthur, Tom, 1986. Worlds of Reference. Lexicography, Learning and Language from the Clay Tablet to the Computer. Cambridge University Press, Cambridge, UK.

Svensén, Bo, 2009. A Handbook of Lexicography: The Theory and Practice of Dictionary-Making. Cambridge University Press, Cambridge.

Lexicon

James Pustejovsky, Brandes University, Waltham, MA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8768–8774, © 2001, Elsevier Ltd.

Abstract

The lexicon is standardly viewed as a listing of all the morphemes of a language, with information indicating how each morpheme behaves in the components of grammar involving phonology, syntax, and semantics.

The Notion of Lexicon

The lexicon is standardly viewed as a listing of all the morphemes of a language, with information indicating how each morpheme behaves in the components of grammar involving phonology, syntax, and semantics. In no small part, the shape and character of grammar is determined by what the lexicon contains for these other grammatical devices. Nevertheless, both historically and conventionally, the lexicon has been seen as the passive module in the system of grammar.

More recently, the model of the lexicon has undergone significant revision and maturation. In particular, two trends have driven the architectural concerns of lexical researchers: (1) a tighter integration of compositional operations of syntax and semantics with the lexical information structures that bear them; and (2) a serious concern with how lexical types reflect the underlying ontological commitments of the grammar. In the process, the field has moved toward addressing more encompassing problems in linguistic theory, such as those below:

1. How can we explain the polymorphic nature of language?
2. How can we capture the creative use of words in novel contexts?
3. How can semantic types predictably map to syntactic representations?
4. What are the 'atoms' of lexical knowledge, if they exist at all?

In this article, we first review the conventional view of the lexicon and then contrast this with the theories of lexical information that have emerged since around 1990.

By all accounts, the conventional model of the lexicon is that of a database of words, ready to act in the service of more dynamic components of the grammar. This view has its origins squarely in the generative tradition (Chomsky, 1955) and has been an increasingly integral part of the concept of the lexicon ever since. While the 'Aspects' model of selectional features restricted the relation of selection to that between lexical items, work by Jackendoff (1972) and McCawley (1968), showed that selectional restrictions must be available to computations at the level of derived semantic representation rather than at deep structure. But where did this view come from? In order to understand both the classical model of the lexicon as a database and the current models of lexically encoded grammatical information, it is necessary to appreciate the structuralist distinction between 'syntagmatic processes' and 'paradigmatic systems' in language. The lexicon has emerged as the focal point communicating between these two components, and can be seen as a hook which links the information at these two levels. One can go further still and view the elements of the lexicon as not just the building blocks for the more active components of the grammar, but also as actively engaging the building principles themselves.

While syntagmatic processes refer to the influence of horizontal elements on a word or phrase, paradigmatic systems refer to vertical substitutions in a phrasal structure. Syntagmatics evolved into the theory of abstract syntax while paradigmatics was all but abandoned in generative linguistics. In an early discussion of syntagmatic dependencies, Hjelmslev (1943) uses the term 'selection' explicitly in the modern sense and notes the importance of integrating paradigmatic systems with the syntagmatic processes they participate in. For Hjelmslev, there are two possible types of relations that can exist between elements in a syntagmatic process: 'interdependence' and 'determination,' the latter of which is related to the notion of selectional restriction as developed by Chomsky (1965). As Cruse (1986) notes "One reason that selectional restrictions were not integrated into mechanisms of grammatical selection and description in the 1970s and 1980s is that, if they are imposed correctly, the grammar is forced to model two computations:

1. The entailment relations between selectional restrictions as features must be modeled formally, in order to contribute to the computation of a syntactic description;
2. The manner in which selectional features or constraints contribute to the determination of the meaning of expressions must be enriched in order to exploit these very features."

Recently, with the convergence of several areas in linguistics (lexical semantics, computational lexicons, and type theories) several models for the determination of selection have emerged which actively integrate these central syntagmatic processes into the grammar, by making explicit reference to the paradigmatic systems which allow for grammatical constructions to be partially determined by selection. Examples of this approach are Generative Lexicon Theory (Bouillon and Busa 2001; Pustejovsky, 1995), and to a certain extent, Construction Grammar (Goldberg, 1995), Combinatory Categorical Grammar (CCG) (Steedman, 1997), and Head-driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994).

These recent theoretical developments have led to a new direction in lexical design. Rather than restricting the scope of

the lexicon to that of a passive database, current frameworks have rearchitected the relationship between syntactic and semantic representations and the underlying recurrence relations that generate them. These developments have helped to characterize the approaches to lexical design in terms of a hierarchy of semantic expressiveness. There are at least three such classes of lexical description, which are defined (cf Pustejovsky, 1995 for discussion):

1. Sense Enumerative Lexicons: lexical items have a single type and meaning, and ambiguity is treated by multiple listings of words.
2. Polymorphic Lexicons: lexical items are active objects, contributing to the determination of 'meaning in context' under well-defined constraints.
3. Unrestricted Sense Lexicons: meanings of lexical items are determined mostly by context and conventional use. Few, if any, restrictions are imposed on how a word may refer.

Although there have been proponents for each class of lexical description defined here, the most promising direction seems to be a careful and formal elucidation of the polymorphic lexicons, and this will form the basis of our subsequent discussion of both the structure and content of lexical entries.

The Structure of a Lexical Entry

As mentioned in Section The Notion of Lexicon, it is generally agreed that there are three components to a lexical item: phonological, syntactic, and semantic information. In this entry, we mainly focus on the manner in which syntactic and semantic representations are encoded in the lexical entry.

There are two types of syntactic knowledge associated with a lexical item: its 'category' and its 'subcategory.' The former includes traditional classifications of both the major categories, such as noun, verb, adjective, adverb, and preposition, as well as the minor categories, such as adverbs, conjunction, quantifier elements, and determiners (*see* Parts of Speech).

Knowledge of the subcategory of a lexical item is typically information that differentiates categories into distinct, distributional classes. This sort of information may be usefully separated into two types, 'contextual features' and 'inherent features.' The former are features that may be defined in terms of the contexts in which a given lexical entry may occur. Subcategorization information marks the local syntactic context for a word. It is this information that ensures that the verb 'devour,' for example, is always transitive in English, requiring a direct object; the lexical entry encodes this requirement with a subcategorization feature specifying that an Noun Phrase (NP) appear to its right. Another type of context encoding is collocational information, where patterns that are not fully productive in the grammar can be tagged. For example, the adjective 'heavy' as applied to 'drinker' and 'smoker' is collocational and not freely productive in the language (Mel'čuk 1988). 'Inherent features,' on the other hand, are properties of lexical entries that are not easily reduced to a contextual definition, but rather refer to the ontological typing of an entity. These include such features as count/mass (e.g., 'pebble' vs 'water,' abstract, animate, human, physical, and so on).

Semantic information can also be separated into two categories: 'base semantic typing' and 'selectional typing.' While the former identifies the semantic class that a lexical item belongs to (such as entity, event, property), the latter class specifies the semantic features of arguments and adjuncts to the lexical item.

Word Classes and Typing Information

There are two major approaches to classifying lexical items by their type: syntactic and semantic. (Another influential tradition for verb classification, which we will not discuss in detail here, is a more descriptive approach to word classes, where membership is defined on the basis of grammatical behavior and verbal valency alternation, such as that elaborated on and compiled in Levin, 1993).

One obvious way to organize lexical knowledge, be it syntactic or semantic, is by means of lexical inheritance mechanisms. In fact, much recent work has focused on how to provide shared data structures for syntactic and morphological knowledge (Flickinger et al., 1985). Evans and Gazdar (1990) provide a formal characterization of how to perform inferences in a language for multiple and default inheritance of linguistic knowledge. The language developed for that purpose, DATR, uses value terminated attribute trees to encode lexical information. Briscoe et al. (1993) describe a rich system of types for allowing default mechanisms into lexical type descriptions.

Type structures can express the inheritance of syntactic features (Sanfilippo, 1993) as well as the relationship between more conventional taxonomic information, such as that shown in Section Argument Structure (cf Copestake and Briscoe, 1992; Pustejovsky, 1995; Pustejovsky and Boguraev, 1993) and **Figure 1**.

Given a semi-lattice of types such as this, the lexical items in the language can be associated with a much richer system of differentiated semantic classes. Verbs may also be structured in hierarchical relations, as done in HPSG, Lexical Functional Grammar (LFG), and other lexical frameworks (cf Pollard and Sag, 1994; Alsina, 1992; Koenig and Davis, 1999).

Argument Structure

Once the base syntactic and semantic typing for a lexical item has been specified, its subcategorization and selectional information must be encoded in some form. There are two major techniques for representing this type of knowledge:

1. associate 'named roles' with the arguments to the lexical item;

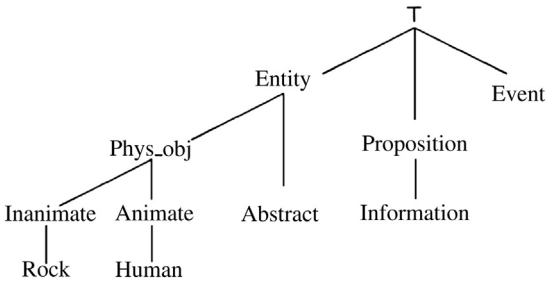

Figure 1 Fragment of a type hierarchy.

2. associate a logical decomposition with the lexical item; meanings of arguments are determined by how the structural properties of the representation are interpreted (cf Hale and Keyser, 1993; Levin and Rappaport, 1995; Pustejovsky, 1995).

One influential way of encoding selectional behavior is the theory of thematic relations (cf Gruber, 1976; Jackendoff, 1972). Thematic relations are typically defined as partial semantic functions of the event being denoted by the verb or noun, and behave according to a predefined calculus of roles relations (e.g., Carlson, 1984; Dowty, 1989). For example, semantic roles such as agent, theme, and goal can be used to partially determine the meaning of a predicate, when they are associated with the grammatical arguments to a verb.

The theory of argument structure as developed by Williams (1981), Grimshaw (1990), and others can be seen as a move toward a more minimalist description of semantic differentiation in the verb's list of parameters. The argument structure for a word can be seen as the simplest specification of its semantics, indicating the number and type of parameters associated with the lexical item as a predicate. For example, the verb 'build' can be represented as a predicate taking two arguments, while the verb 'give' takes three arguments.

1.
   a. build ($\underline{x}$, y)
   b. give ($\underline{x}$, y, z)

What originally began as the simple listing of the parameters or arguments associated with a predicate has developed into a sophisticated view of the way arguments are mapped onto syntactic expressions. Williams' (1981) distinction between 'external' (the underlined arguments) and 'internal' arguments and Grimshaw's proposal for a hierarchically structured representation (cf Grimshaw, 1990) provide us with the basic syntax for one aspect of a word's meaning. Similar remarks hold for the argument list structure in HPSG (Pollard and Sag, 1994) and LFG (Bresnan, 1994).

The interaction of a structural argument list and a rich system of types, such as that presented, provides a mechanism for semantic selection that overcomes the difficulties mentioned in Section The Structure of a Lexical Entry. The most direct impact of semantic type systems on syntactic subcategorization can be seen with the analysis of a simple example.

2.
   a. The man/the rock fell.
   b. The man/*the rock died.

Returning to the example in (2), consider how the selectional distinction for the feature [+/− animacy] is modeled. For the purpose of illustration, the arguments of a verb will be represented in a list structure, where each argument is identified as being typed with a specific value.

3.
   a. $\begin{bmatrix} \text{fall} \\ \text{ARGSTR} = [\text{ARG}_1 = \text{x} : \text{phys_obj}] \end{bmatrix}$
   b. $\begin{bmatrix} \text{die} \\ \text{ARGSTR} = [\text{ARG}_1 = \text{x} : \text{animate}] \end{bmatrix}$

In the sentences in (2), it is intuitively clear how rocks cannot die and men can but it is still not obvious how this inference is computed, given what we would assume are the types associated with the nouns 'rock' and 'man' respectively. What accomplishes this computation is a rule of subtyping, Θ, that allows the type associated with the noun 'man' (i.e., human) to also be accepted as the type animate, which is what the predicate 'die' requires of its argument as stated in (3b)

4. Θ [human $\sqsubseteq$ animate]: human $\rightarrow$ animate

The rule Θ applies, since the concept human is subtyped under animate in the type hierarchy. Parallel considerations rule out the noun 'rock' as a legitimate argument to 'die' since it is not subtyped under animate. Hence, one of the concerns given for how syntagmatic processes can systematically keep track of which 'selectional features' are entailed and which are not is partially addressed by such lattice traversal rules as the one presented here.

Selection can also be employed to solve the problem of polymorphism, where the same lexical item can appear in multiple syntactic contexts, as illustrated in (5) below.

5.
   a. Mary began to read the novel.
   b. Mary began reading the novel.
   c. Mary began the novel.

Gerative Lexicon Theory handles such examples by the use of semantic selection (cf Pustejovsky, 1995) and canonical syntactic mapping rules, specifying how semantic types correspond to syntactic expressions. Here, the verb 'begin' is typed as taking an event description as its internal argument, begin (Ind, $\mathscr{E}$), which can be realized in one of the three syntactic forms shown. This illustrates how a generative system of type operations can account for polymorphic behavior of selection in the syntax.

Decomposition and Event Structure

The second major approach to the specification of lexical knowledge is that taken by decompositional theories. Most of this research has focused on the lexical specification for verbs, their arguments, and their syntactic behavior. Some recent work, however, has been done on noun semantics as well (Busa, 1996). In this section, we examine the motivations for lexical decomposition in linguistic theory, and the various proposals that have emerged on how to encode lexical knowledge as structured forms. We then relate this to the manner in which verbs refer to events, since this directly impacts the nature of the lexical decomposition structure.

Since Davidson (1967), events have played an increasingly important role in the determination of verb meaning. While early researchers on decompositional models (Dowty, 1979; McCawley, 1968) made no ontological commitments to events in the semantics for verbs, a new synthesis has emerged in recent years which attempts to model verb meanings as complex predicative structures with rich event structures (cf Hale and Keyser, 1993; Pustejovsky, 1991). This research has developed the idea that the meaning of a verb can be analyzed into a structured representation of the event that the verb

designates, and has furthermore contributed to the realization that verbs may have complex, internal event structures. Recent work has converged on the view that complex events are structured into an inner and an outer event, where the outer event is associated with causation and agency, and the inner event is associated with telicity (completion) and change of state. Under this view, a canonical accomplishment predicate as in 'John sliced the bread,' for example, can be represented as composed of an inner and an outer event. The inner event is the telic event in which the bread undergoes a change of state, and the outer event is the event in which John acts agentively (to do whatever is involved in the act of slicing). Since the outer event causes the inner one, it is associated lexically with causation.

Although there is a long tradition of analyzing causation as a relation between two events in the philosophical and psychological literature in contemporary models of natural language semantics this idea has only recently been adopted. For example, Carter (1976), one of the earlier researchers in this area, represents the meaning of the verb 'darken' as follows:

6. x CAUSE ((y BE DARK) CHANGE))

The predicate CAUSE is represented as a relation between a causer argument x and an inner expression involving a change of state in the argument y, although there is an intuition that the cause relation involves a causer and an event. Carter does not make this commitment explicitly. Levin and Rappaport (1995) follow a similar strategy, with a CAUSE predicate relating a causer argument and an inner expression involving a change of state in the argument y. The change of state is represented with the predicate BECOME:

7. wipe the floor clean:
   x CAUSE [y BECOME (AT) z] BY [x 'wipe' y]]
   x CAUSE [floor BECOME (AT) clean] BY [x 'wipe' floor]

The work of Levin and Rappaport, building on Jackendoff's Lexical Conceptual Structures, has been influential in articulating the internal structure of verb meanings (see Levin and Rappaport, 1995).

Jackendoff (1990) develops an extensive system of what he calls 'Conceptual Representations' which parallel the syntactic representations of sentences of natural language. These employ a set of canonical predicates including CAUSE, GO, TO, and ON, and canonical elements including Thing, Path, and Event. These approaches represent verb meaning by decomposing the predicate into more basic predicates. This work owes obvious debt to the innovative work within generative semantics, as illustrated by McCawley's (1968), analysis of the verb 'kill.' Recent versions of lexical representations inspired by generative semantics can be seen in the Lexical Relational Structures of Hale and Keyser 1993, where syntactic tree structures are employed to capture the same elements of causation and change of state as in the representations of Carter, Levin and Rapoport, Jackendoff, and Dowty.

Pustejovsky (1988, 1991) extends the decompositional approach presented in Dowty (1979) by explicitly reifying the events and subevents in the predicative expressions. Unlike Dowty's treatment of lexical semantics, where the decompositional calculus builds on prepositional or predicative units (as previously discussed), a 'syntax of event structure' makes explicit reference to quantified events as part of the word meaning. Pustejovsky further introduces a tree structure to represent the temporal ordering and dominance constraints on an event and its subevents. For example, a predicate such as 'build' is associated with a complex event such as that shown in (8):

8. $[_{\text{transition}}[e_1\text{: PROCESS}][e_2\text{: STATE}]]$

The process consists of the building activity itself, while the state represents the result of there being the object built. Grimshaw (1990) adopts this theory in her work on argument structure, where complex events such as 'break' are given a similar representation. In such structures, the process consists of what x does to cause the breaking, and the state is the resultant state of the broken item. The process corresponds to the outer causing event as discussed, and the state corresponds in part to the inner change of state event. Both Pustejovsky and Grimshaw differ from the authors above in assuming a specific level of representation for event structure, distinct from the representation of other lexical properties. Furthermore, they follow Higginbotham (1985) in adopting an explicit reference to the event place in the verbal semantics.

Qualia Structure

Thus far, we have focused on the lexical information associated with verb entries. All of the major categories, however, are encoded with syntactic and semantic feature structures that determine their constructional behavior and subsequent meaning at logical form. How this is accomplished, of course, varies from theory to theory.

In Generative Lexicon Theory, it is assumed that word meaning is structured on the basis of four generative factors, or 'qualia roles,' that capture how humans understand objects and relations in the world and provide the minimal explanation for the linguistic behavior of lexical items (these are inspired in large part by Moravcsik's (1975, 1990) interpretation of Aristotelian *aitia*).

formal: the basic category that distinguishes the object within a larger domain;
constitutive: the relation between an object and its constituent parts;
telic: its purpose and function;
argentive: factors involved in the object's origin or 'coming into being.'

Qualia structure is at the core of the generative properties of the lexicon, since it provides a general strategy for creating new types. For example, consider the properties of nouns such as 'rock' and 'chair.' These nouns can be distinguished on the basis of semantic criteria which classify them in terms of general categories such as natural_kind, artifact_object. Although very useful, this is not sufficient to discriminate semantic types in a way that also accounts for their grammatical behavior. A crucial distinction between 'rock' and 'chair' concerns the properties which differentiate natural_kinds from 'artifacts': functionality plays a crucial role in the process of individuation of artifacts, but not of natural kinds. This is reflected in grammatical behavior, whereby 'a good chair,' or 'enjoy the chair' are well-formed expressions reflecting the specific purpose for

which an artifact is designed, but 'good rock' or 'enjoy a rock' are semantically ill-formed since for 'rock' the functionality (i.e., TELIC) is undefined. Exceptions exist when new concepts are referred to, such as when the object is construed relative to a specific activity, such as in 'The climber enjoyed that rock'; 'rock' itself takes on a new meaning, by virtue of having telicity associated with it, and this is accomplished by integration with the semantics of the subject NP. Although 'chair' and 'rock' are both physical_object, they differ in their mode of coming into being (i.e., AGENTIVE): artifacts are man-made, 'rocks' develop in nature. Similarly, a concept such as 'food' or 'cookie' has a physical manifestation or denotation, but also a functional grounding, pertaining to the relation of 'eating.' These apparently contradictory aspects of a category are orthogonally represented by the qualia structure for that concept, which provides a coherent structuring for different dimensions of meaning.

For relations, the qualia act in a similar capacity to thematic relations, but where the individual qualia are possibly associated with entire event descriptions, and not just individuals. For discussion, see Pustejovsky (1995).

Consequence of Lexical Design

Considering the potential range of information that can be represented lexically results in a reconceptualization of what a lexicon is, so that the very design of the grammar is significantly impacted. Furthermore, our current understanding of psychological and computational properties of language processing suggests that the resources available for lexical storage and access are considerably higher than originally imagined by early grammatical theorists. The consequence is that, what had originally been accomplished in syntax, because of the combinatoric properties inherent in production rules, can be handled by the lexicon itself. The various approaches to lexical encoding can be analyzed in terms of two parameters.

1. Precompiling the information into lexical items forms;
2. Computing or generating new forms or senses during the compositional process.

Typically, idioms are presented as examples of precompiled lexical entries. But some theories have adopted this idea as fundamental for the entire compositional operation. The best example of this is CCG (Steedman, 1997). CCG has recently been articulated in enough detail to handle most of the major linguistic phenomena using a library of precompiled lexical types, together with the combinatoric rules of categorical syntax. If the grammar utilizes representations with such nonlocal dependencies, then there must be additional mechanisms for unifying these representations; these are provided in the form of function composition rules and lexical rules (Schabes et al., 1988).

Lexical rules have been invoked in HPSG, as well, to explain the relationship between the various senses for lexical items, from grinding and packaging operations (such as that relating the animal and food senses of 'chicken' and 'lamb,' to the relation between logically polysemous items, such as 'book' information and physical object), and 'lecture' (information and event) (cf Copestake and Briscoe, 1992). In Generative Lexicon such relations are explicitly represented in context (cf Johnston, 1995). It is very likely, however, that language makes use of both types of devices, namely complex types such as dot objects, as well as the application of lexical rules. Regardless, both types of devices must be seriously constrained by the grammar in order not to overgenerate unwanted forms and interpretations.

Considered independently of the issue of precompiled vs generated forms and senses, there is no question that the mental lexicon is large, containing arguably up to 400 000 lexical entries. This is based on fairly conservative estimates of speaker competence with active and passive vocabularies. For example, an average speaker lexicon might contain at least 5000 distinct verbs, 30 000 distinct nominal forms, and over 5000 adjectives. Combine this with an additional 10 000 compound forms and at least 300 000 distinct proper names. Obviously, the psychological (and hence computational) demands of these classes are quite distinct. There are two dimensions that can help us distinguish these classes: (1) the degree of combinatoric (functional) complexity of the lexical item; and (2) whether the lexical item is part of active or passive lexical knowledge. Most closed class items, for example, are functionally complex, as are many open class verbs and relational nouns. The majority of the open class items will also involve a fair amount of information regarding combinatoric possibilities. The class of names, however, is unique in that, although it is by far the largest class of lexical items, it is the least demanding in terms of computational resources.

In conclusion we see that the lexicon is neither a mere listing of morphemes in the language, nor a database of items passively waiting in the service of grammatical processes. Rather, the lexicon is a dynamic and active system of grammar, incorporating as well as dictating essential components of syntactic and semantic composition and interpretation.

See also: Generative Grammar; Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects; Lexical Semantics; Lexicology and Lexicography; Syntax–Semantics Interface.

Bibliography

Alsina, A., 1992. On the argument structure of causatives. Linguistic Inquiry 23 (4), 517–555.
Baker, M., 1988. Incorporation: A Theory of Grammatical Function Changing. University of Chicago Press, Chicago.
Bresnan, J., 1994. Locative inversion and the architecture of universal grammar. Language 70 (1), 2–31.
Boguraev, B., Briscoe, E., 1989. Computational Lexicography for Natural Language Processing. Longman, Harlow and London.
Boguraev, B., Pustejovsky, J., 1996. Corpus Processing for Lexical Acquisition. Bradford Books/MIT Press, Cambridge, MA.
Bouillon, P., Busa, F., 2001. The Syntax of Word Meaning. Cambridge University Press, Cambridge.
Briscoe, T., de Paiva, V., Copestake, A. (Eds.), 1993. Inheritance, Defaults, and the Lexicon. Cambridge University Press, Cambridge, UK.
Busa, F., 1996. Compositionality and the Semantics of Nominals (Ph.D. Dissertation). Brandeis University.

Carlson, G., 1984. Thematic roles and their role in semantic interpretation. Linguistics 22, 259–279.
Chomsky, N., 1955. The Logical Structure of Linguistic Theory. University of Chicago Press, Chicago.
Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.
Copestake, A., Briscoe, E., 1992. Lexical operations in a unification-based framework. In: Pustejovsky, J., Bergler, S. (Eds.), Lexical Semantics and Knowledge Representation. Springer Verlag, New York.
Cruse, A., 1986. Lexical Semantics. Cambridge University Press.
Davis, A.R., Koenig, J.P., 2000. Linking as constraints on word classes in a hierarchical lexicon. Language 76 (1).
Dowty, D.R., 1979. Word Meaning and Montague Grammar. D. Reidel, Dordrecht, The Netherlands.
Dowty, D.R., 1989. On the semantic content of the notion 'thematic role'. In: Chierchia, G., Partee, B., Turner, R. (Eds.), Properties, Types, and Meaning, vol. II. Dordrecht.
Dowty, D., 1991. Thematic proto-roles and argument selection. Language 67, 547–619.
Flickinger, D., Pollard, C., Wasow, T., 1985. Structure-sharing in lexical representation. In: Proceedings of 23rd Annual Meeting of the ACL. Chicago, IL, pp. 262–267.
Goldberg, A.E., 1995. Constructions: A Construction Grammar Approach to Argument Structure. University of Chicago Press, Chicago.
Grimshaw, J., 1979. Complement selection and the lexicon. Linguistic Inquiry 10, 279–326.
Grimshaw, J., 1990. Argument Structure. MIT Press, Cambridge, MA.
Gruber, J.S., 1976. Lexical Structures in Syntax and Semantics. North Holland, Amsterdam.
Guthrie, L., Pustejovsky, J., Wilks, Y., Slator, B., 1996. The role of lexicons in natural language processing. Communications of the ACM 39 (1).
Hale, K., Keyser, J., 1993. On argument structure and the lexical expression of syntactic relations. In: Hale, K., Keyser, J. (Eds.), The View from Building 20. MIT Press, Cambridge, MA.
Halle, M., Bresnan, J., Miller, G. (Eds.), 1978. Linguistic Theory and Psychological Reality. MIT Press, Cambridge, MA.
Higginbotham, J.T., 1985. On sematics. Linguistics Inquiry 16, 547–593.
Hjelmslev, L., 1961[1943]. Prolegomena to a Theory of Language (F. Whitfield, Trans.). University of Wisconsin Press, Madison.
Ingria, R., 1986. Lexical Information for Parsing System: Points of Convergence and Divergence. Automating the Lexicon. Marina di Grosseto, Italy.
Ingria, R., Bogurae, V.B., Pustejovsky, J., 1992. Dictionary/Lexicon. In: Shapiro, S. (Ed.), Encyclopedia of Artificial Intelligence, second ed. Wiley, New York.
Jackendoff, R., 1972. Semantic Interpretation in Generative Grammar. MIT Press, Cambridge, MA.
Jackendoff, R., 1990. Semantic Structures. MIT Press, Cambridge, MA.
Johnston, R., 1995. Semantic underspecification and lexical types: capturing polysemy without lexical rules. In: Proceedings of ACQUILEX Workshop on Lexical Rules, August 9–11, 1995. Cambridgeshire, UK.
Koenig, J.-P., Davis, A., 1999. Linking as constraints on word classes in a hierarchical lexicon. In: Rordoni, V. (Ed.), Proceedings of the ESSLLS – 99 Workshop on Lexical Semantics and Linking in Constraint-based Theories, pp. 1–45.
Levin, B., 1993. Towards a Lexical Organization of English Verbs. University of Chicago Press, Chicago.
Levin, B., Rappaport, M.H., 1995. Unaccusativity: At the Syntax–Semantics Interface. MIT Press, Cambridge, MA.
Lyons, J., 1968. Introduction to Theoretical Linguistics. Cambridge University Press, Cambridge, UK.
McCawley, J., 1968. Lexical insertion in a transformational grammar without deep structure. In: Proceedings of the Chicago Linguistic Society 4.
Mel'čuk, I.A., 1988. Semantic description of lexical units in an explanatory combinatorial dictionary: basic principles and heuristic criteria. International Journal of Lexicography 1, 165–188.
Miller, G., 1990. WordNet: an on-line lexical database. International Journal of Lexicography 3, 235–312.
Miller, G., 1991. The Science of Words. Scientific American Library.
Moravcsik, J.M., 1975. Aitia as generative factor in Aristotle's philosophy. Dialogue 14, 622–636.
Moravcsik, J.M., 1990. Thought and Language. Routledge, London.
Pinker, S., 1989. Learnability and Cognition: The Acquisition of Argument Structure. MIT Press, Cambridge, MA.
Pollard, C., Sag, I., 1994. Head-driven Phrase Structure Grammar. University of Chicago Press and Stanford CSLI, Chicago.
Pustejovsky, J., 1988. The Geometry of events. In: Tenny, C. (Ed.), Studies in Generative Approaches to Aspect. Lexicon Project Working Papers, 24. MIT, Cambridge, MA.
Pustejovsky, J., 1991. The syntax of event structure. Cognition 41, 47–81.
Pustejovsky, J., 1992. Lexical semantics. In: Shapiro, S. (Ed.), Encyclopedia of Artificial Intelligence, second ed. Wiley, New York.
Pustejovsky, J., 1995. The Generative Lexicon. MIT Press, Cambridge, MA.
Pustejovsky, J., Boguraev, P., 1993. Lexical knowledge representation and natural language processing. Artificial Intelligence 63, 193–223.
Sanfilippo, A., 1993. LKB encoding of lexical knowledge. In: Briscoe, T., de Paiva, V., Copestake, A. (Eds.), Inheritance, Defaults, and the Lexicon. Cambridge University Press, Cambridge, UK.
Schabes, Y., Abeille, A., Joshi, A., 1988. Parsing strategies with lexicalized grammars. In: Proceedings of the 12th International Conference on Computational Linguistics, Budapest, Hungary.
Steedman, M., 1997. Surface Structure Interpretation. MIT Press, Cambridge, MA.
Talmy, L., 1985. Lexicalization patterns: semantic structure in lexical forms. In: Shopen, T. (Ed.), Language Typology and Syntactic Description 3: Grammatical Categories and the Lexicon. Cambridge University Press, Cambridge, pp. 57–149.
Weinreich, U., 1972. Explorations in semantic theory. Mouton, The Hague, The Netherlands.
Williams, E., 1981. Argument structure and morphology. Linguistic Review 1, 81–114.

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 14

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser—Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 14

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 14 *xliii*

VOLUME 14

Liaison and Consultation Psychiatry 1
Albert Diefenbacher and R Burian

Liberal Christianity 8
Linda Woodhead

Liberalism, Anthropology of 11
Naomi Schiller

Liberalism: Historical Aspects 18
Dieter Langewiesche

Liberalism: Impact on Social Science 22
Richard P Bellamy

Liberalism: Political Doctrine and Impact on Social Science 26
Richard Bellamy

Liberation Theology in Social Work 32
Carolina Muñoz-Guzmán

Libertarianism 39
Chandran Kukathas

Libraries 45
Andrea J Copeland

Life Course and Event History Analysis 51
Hans-Peter Blossfeld and Gwendolin J Blossfeld

Life Course Criminology 59
Marvin D Krohn, Molly Buchanan, and Andrea Davis

Life Course Epidemiology 67
Gita D Mishra, Diana Kuh, and Yoav Ben-Shlomo

Life Course in History 76
JR Gillis

Life Course: Sociological Aspects 80
Robert Crosnoe and Glen H Elder Jr

Life Management, Developmental Psychology of 85
AM Freund

Life Table 89
Jan M Hoem

Lifelong Learning and Its Support with New Technologies 93
Marco Kalz

Lifespan Development, Theory of 100
Brandilynn Villarreal and Jutta Heckhausen

Lifespan Development: Evolutionary Perspectives 108
Heidi Keller

Life-style, History of the Concept 112
Johanna Muckenhuber

Likelihood in Statistics 116
Anthony WF Edwards

Likelihoods that Eliminate Nuisance Parameters 120
Brunero Liseo

Limbic System 125
Allison M Auchter and Marie H Monfils

Liminality 131
Hazel Andrews and Les Roberts

Limits to Growth 138
Ugo Bardi

Linear Hypothesis 144
Ronald Christensen

Linear Hypothesis: Fallacies and Interpretive Problems (Simpson's Paradox) 150
DV Lindley

Linear Hypothesis: Regression (Basics) 153
Sanford Weisberg

Linear Hypothesis: Regression (Graphics) 157
R Dennis Cook

Linear Mixed Models 162
Garrett M Fitzmaurice and Nan M Laird

Linguistic Anthropology 169
Marcel Danesi

Linguistic Approaches to Agrammatism 176
Alan Beretta

Linguistic Fieldwork 183
Andrea L Berez

Linguistic Presupposition 190
Rob A van der Sandt

Linguistic Synesthesia 193
Sean A Day

Linguistic Turn 199
C Fynsk

Linguistic Turn and Discourse Analysis in History 202
John E Toews

Linguistic Typology 208
Masayoshi Shibatani

Linguistics, History of 215
EF Konrad Koerner

Linguistics: Incorporation 220
Elke Nowak

Linguistics: Overview 224
David Crystal

Linked Lives 230
Maja Djundeva

Literacy and Illiteracy, History of 236
Istvan G Tóth

Literacy and Orality 241
Wolfgang Raible

Literacy Education 246
Maureen Walsh

Literary and Philosophical Utopian Thought 251
Adrienne Redd and Tsvi Bisk

Lithics and Archaeology 260
Jacques Pelegrin

Litigation 265
Herbert M Kritzer

Llewellyn, Karl Nickerson (1893–1962) 276
Brian Leiter

Local Economic Development 279
Christian M Rogerson

Location, Absolute and Relative 284
Harvey J Miller

Location Theory 287
Cees Gorter and Peter Nijkamp

Logic and Linguistics 293
Thomas Ede Zimmermann

Logic Models 299
Joy A Frechtling

Logical and Hypothetical Reasoning in Adolescence, Development of 306
Eric Amsel and David Moshman

Logical Positivism and Logical Empiricism 313
Herbert Keuth

Logics for Knowledge Representation 319
Bernhard Nebel

Log-Linear and Marginal Models 322
Tamás Rudas

Longitudinal Analyses of Sexual Development through Early Adulthood 328
Joseph Gonzales and Emilio Ferrer

Longitudinal Causal Inference 340
Miguel A Hernán and James M Robins

Longitudinal Data 345
Burton H Singer

Longitudinal Data: Event–History Analysis in Discrete Time 351
Elja Arjas

Longitudinal Research: Panel Retention 354
Jennifer A Parsons

Long-Term Potentiation and Long-Term Depression 358
Tammy L Ivanco

Lorenz, Konrad (1903–89) 366
Wolfgang M Schleidt

Lotka, Alfred James (1880–1949) 372
Nathan Keyfitz

Love and Intimacy, Psychology of 375
Panos Kordoutis

Luhmann, Niklas (1927–98) 382
Rudolf Stichweh

Luhmann's *Social Systems*: Meaning, Autopoiesis, and Interpenetration 390
Kenneth C Bausch

Luria, Aleksandr Romanovich (1902–77) 396
Michael Cole and Xavier E Cagigas

Macroeconomics, History of From 1933 to Present 400
Kevin D Hoover

Macroeconomics, History of up to 1933 406
Robert W Dimand

Macrosociology-Microsociology 414
Gianluca Manzo

Mafia, Anthropology of 422
Anton Blok

Mafias, Criminology of 429
Gavin Slade

Magic, Anthropology of 435
Vassos Argyrou

Magnetic Resonance and Molecular Imaging in Psychiatry 439
Oliver D Howes

Majoritarianism and Majority Rule 446
Douglas W Rae

Majorization and Stochastic Orders 449
Harry Joe

Malinowski, Bronislaw (1884–1942) 453
Michael W Young

Malthus, Thomas Robert (1766–1834) 457
Jacques Dupâquier

Managed Care and Psychotherapy in the United States 462
Linda D Weaver and Clarissa C Marques

Management Accounting 466
Alnoor Bhimani

Management: General 471
Alberto Martinelli

Managerial and Organizational Cognition 479
Gerard P Hodgkinson

Mania 484
Neel Halder

Mann, Michael (1942–) 490
Ralph Schroeder

Mannheim, Karl (1893–1947) 496
Volker Meja

Manual Based Treatments 500
Johanna Thompson-Hollands and David H Barlow

Māori Social Work 505
Oliver Helena Fiona Te Momo

Marital Quality and Health 512
Deborah Carr and Tetyana Pudrovska

Marital Status, Divorce, and Child Development 518
Colleen M Carr and Sharlene A Wolchik

Maritime Networks, History of 525
Gordon Boyce

Markedness in Linguistics 533
Edwin Battistella

Market and Nonmarket Allocation 538
Debra Satz

Market Geographies 544
Jean-Claude Thill

Market Structure and Performance 549
Paul B Ellickson

Marketing Strategies 555
Enrico Valdani and Alessandro Arbore

Markets: Anthropological Aspects 559
Theodore Bestor

Markets: Artistic and Cultural 564
Victoria D Alexander

Markov Chain Monte Carlo Methods 569
Peter Müller

Markov Decision Processes 573
Michael L Littman

Markov Models and Social Analysis 576
Lynne Billard

Markov Processes in Learning Spaces 584
Jean-Claude Falmagne

Maroons in Anthropology 591
Richard Price

Marr, David (1945–80) 596
Shimon Edelman and Lucia M Vaina

Marriage 599
Perveez Mody

Marriage and the Dual-Career Family 605
Sara Raley

Marriage Systems, Evolution of 611
Laura Fortunato

Marshall, Alfred (1842–1924) 620
John K Whitaker

Marshall, Thomas Humphrey (1893–1981) 626
Yasemin N Soysal

Martingales 630
René L Schilling

Maruyama, Masao (1914–96) 635
Andrew E Barshay

Marx, Karl (1818–83) 642
Terrell Carver

Marxism in Contemporary Sociology 649
Matt Vidal

Marxism–Leninism: The Ideology of Twentieth-Century Communism 656
William J Morgan

Marxist Approaches to Social Work 663
Tom Vickers

Marxist Archaeology 670
Antonio Gilman

Marxist Geography 674
Richard Peet

Masculinity and Social Work Practice 679
Bob Pease

Masculinity, Scripts of 685
Elizabeth Nagel, Rachel Kalish, and Michael Kimmel

Mass Communication: Normative Frameworks 690
John D Peters

Mass Imprisonment and Its Consequences 696
Holly Foster and John Hagan

Mass Killings and Genocide, Psychology of 702
Ervin Staub

Mass Media and Cultural Identity 707
Philip Schlesinger

Mass Media and Politics 712
Walt Borges

Mass Media and Sports 719
David Rowe

Mass Media: Introduction and Schools of Thought 726
Peter Dahlgren

Mass Media, Political Economy of 733
Graham Murdock

Mass Media, Representations in 738
Robert Entman

Mass Society: History of the Concept 743
Salvador Giner

Mass Trauma: Psychopathological Effects across the Life Span 747
Frank Neuner, Claudia Catani, and Verena Ertl

Mastery Learning 752
Thomas R Guskey

Material Flow Analysis 760
Heinz Schandl and Anke Schaffartzik

Materiality and Culture 765
Stuart McLean

Mathematical Concepts during Early Childhood Across Cultures, Development of 772
Inge Schwank and Elisabeth Schwank

Mathematical Models in Geography 785
Sergio J Rey

Mathematical Models in Philosophy of Science 791
Zoltan Domotor

Mathematical Psychology 800
Jean-Claude Falmagne and Michael D Lee

Mathematical Psychology: History 808
William H Batchelder

Mathematics Education 816
Lieven Verschaffel, Fien Depaepe, and Erik De Corte

Matrifocality 822
Raymond T Smith

Mauss, Marcel (1872–1950) 825
Victor Karady

McDonaldization 828
George Ritzer and Chih-Chin Chen

Mead, George Herbert (1863–1931) 831
Daniel R Huebner

Mead, Margaret (1901–78) 837
Margaret M Caffrey

Meaning and Rule Following 841
Hans-Johann Glock

Measurement Error Models 850
Alicia L Carriquiry

Measurement, Representational Theory of 856
Reinhard Suck

Measurement Theory: Conjoint 862
Peter Fishburn

Measurement Theory: Conjoint Analysis Applications 865
Jordan J Louviere

Measurement Theory: History and Philosophy 868
Joel Michell

Mechanism-Based Causal Analysis 873
Carly R Knight

CONTRIBUTORS TO VOLUME 14

Victoria D. Alexander
University of Surrey, Guildford, UK

Eric Amsel
Weber State University, Ogden, UT, USA

Hazel Andrews
Liverpool John Moores University, Liverpool, UK

Alessandro Arbore
Bocconi University, Milan, Italy

Vassos Argyrou
University of Hull, Hull, UK

Elja Arjas
University of Helsinki, Helsinki, Finland

Allison M. Auchter
University of Texas at Austin, Austin, TX, USA

Ugo Bardi
Università di Firenze, Firenze, Italy

David H. Barlow
Boston University, Boston, MA, USA

Andrew E. Barshay
University of California Berkeley, Berkeley, CA, USA

William H. Batchelder
University of California, Irvine, CA, USA

Edwin Battistella
Southern Oregon University, Ashland, OR, USA

Kenneth C. Bausch
Institute for 21st Century Agoras, Cincinnati, OH, USA

Richard P. Bellamy
University of Reading, Reading, UK

Richard Bellamy
University College London (UCL), London University, London, UK; and European University Institute, Florence, Italy

Yoav Ben-Shlomo
School of Social and Community Medicine, University of Bristol, Bristol, UK

Alan Beretta
Michigan State University, East Lansing, MI, USA

Andrea L. Berez
University of Hawai'i at Mānoa, Honolulu, HI, USA

Theodore Bestor
Cornell University, Ithaca, NY, USA

Alnoor Bhimani
London School of Economics, London, UK

Lynne Billard
University of Georgia, Athens, GA, USA

Tsvi Bisk
Center for Strategic Futurist Thinking, Ziv Group, Ltd, Kfar Saba, Israel

Anton Blok
University of Amsterdam, Amsterdam, The Netherlands

Gwendolin J. Blossfeld
Nuffield College, Oxford, UK

Hans-Peter Blossfeld
European University Institute, Fiesole (FI), Italy

Walt Borges
University of North Texas — Dallas, Dallas, TX, USA

Gordon Boyce
University of Newcastle, Callaghan, NSW, Australia

Molly Buchanan
University of Florida, Gainesville, FL, USA

R. Burian
Charité - Universitätsmedizin Berlin, Berlin, Germany; and Evangelisches Krankenhaus Königin Elisabeth Herzberge, Berlin, Germany

Margaret M. Caffrey
University of Memphis, Memphis, TN, USA

Xavier E. Cagigas
University of California, Los Angeles, CA, USA

Colleen M. Carr
Arizona State University, Phoenix, AZ, USA

Deborah Carr
Institute for Health, Health Care Policy & Aging Research, Rutgers University, New Brunswick, NJ, USA

Alicia L. Carriquiry
Iowa State University, Ames, IA, USA

Terrell Carver
University of Bristol, Bristol, UK

Claudia Catani
Bielefeld University, Bielefeld, Germany

Chih-Chin Chen
University of Maryland, College Park, MD, USA

Ronald Christensen
University of New Mexico, Albuquerque, NM, USA

Michael Cole
University of California, La Jolla, CA, USA

R. Dennis Cook
University of Minnesota, St. Paul, MN, USA

Andrea J. Copeland
School of Informatics and Computing, Indiana University, Indianapolis, IN, USA

Robert Crosnoe
Population Research Center, University of Texas at Austin, Austin, TX, USA

David Crystal
University of Wales, Bangor, UK

Peter Dahlgren
Lund University, Lund, Sweden

Marcel Danesi
Victoria College, University of Toronto, Toronto, ON, Canada

Andrea Davis
University of Florida, Gainesville, FL, USA

Sean A. Day
Trident Technical College, Charleston, SC, USA

Erik De Corte
Center for Instructional Psychology and Technology, Leuven, Belgium

Fien Depaepe
Center for Educational Policy, Innovation and Teacher Training, Leuven, Belgium

Albert Diefenbacher
Charité - Universitätsmedizin Berlin, Berlin, Germany; and Evangelisches Krankenhaus Königin Elisabeth Herzberge, Berlin, Germany

Robert W. Dimand
Brock University, St. Catharines, ON, Canada

Maja Djundeva
University of Groningen, Groningen, Netherlands

Zoltan Domotor
University of Pennsylvania, Philadelphia, PA, USA

Jacques Dupâquier
EHESS, Paris, France

Shimon Edelman
Cornell University, Ithaca, NY, USA

Anthony W.F. Edwards
Cambridge University, Cambridge, UK

Glen H. Elder
Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Paul B. Ellickson
University of Rochester, Rochester, NY, USA

Robert Entman
North Carolina State University, Raleigh, NC, USA

Verena Ertl
Bielefeld University, Bielefeld, Germany

Jean-Claude Falmagne
University of California, Irvine, CA, USA

Emilio Ferrer
University of California, Davis, Davis, CA, USA

Peter Fishburn
AT&T Laboratories, Florham Park, NJ, USA

Garrett M. Fitzmaurice
Harvard School of Public Health, Boston, MA, USA

Laura Fortunato
Institute of Cognitive and Evolutionary Anthropology, University of Oxford, Oxford, UK; and Santa Fe Institute, Santa Fe, NM, USA

Holly Foster
Texas A&M University, College Station, TX, USA

Joy A. Frechtling
Westat, Rockville, MD, USA

A.M. Freund
Max Planck Institute for Human Development, Berlin, Germany

C. Fynsk
Binghamton University, Binghamton, NY, USA

J.R. Gillis
Rutgers University, New Brunswick, NJ, USA

Antonio Gilman
Encino, CA, USA

Salvador Giner
University of Barcelona, Barcelona, Spain

Hans-Johann Glock
Institute of Philosophy, University of Zurich, Zürich, Switzerland

Joseph Gonzales
University of California, Davis, Davis, CA, USA

Cees Gorter†
VU University, Amsterdam, The Netherlands

Thomas R. Guskey
College of Education, University of Kentucky, Lexington, KY, USA

John Hagan
Northwestern University, Evanston, IL, USA; and American Bar Foundation, Chicago, IL, USA

Neel Halder
Alpha Hospitals, Bury, UK

Jutta Heckhausen
University of California Irvine, Irvine, CA, USA

Miguel A. Hernán
Harvard School of Public Health, Boston, MA, USA

Gerard P. Hodgkinson
Warwick Business School, University of Warwick, Coventry, UK

Jan M. Hoem
The Max Planck Institute for Demographic Research, Rostock, Germany

Kevin D. Hoover
Duke University, Durham, NC, USA

Oliver D. Howes
MRC Clinical Sciences Centre, Imperial College-Hammersmith Hospital Campus, London, UK; and Institute of Psychiatry, King's College London, London, UK

Daniel R. Huebner
The University of North Carolina at Greensboro, Greensboro, NC, USA

Tammy L. Ivanco
University of Manitoba, Winnipeg, MB, Canada

Harry Joe
University of British Columbia, Vancouver, BC, Canada

Rachel Kalish
Stony Brook University, Stony Brook, NY, USA

Marco Kalz
Open University of the Netherlands, Heerlen, The Netherlands

Victor Karady
CNRS, Paris, France

Heidi Keller
School of Human Sciences, Osnabrück University, Osnabrück, Germany

Herbert Keuth
Eberhard Karls Universität, Tübingen, Germany

Nathan Keyfitz†
Harvard University, Cambridge, MA, USA

Michael Kimmel
Stony Brook University, Stony Brook, NY, USA

Carly R. Knight
Harvard University, Cambridge, MA, USA

E.F. Konrad Koerner
Zentrum für Allgemeine Sprachwissenschaft (ZAS), Berlin, Germany

Panos Kordoutis
Panteion University of Social and Political Sciences, Athens, Greece

Herbert M. Kritzer
University of Minnesota Law School, Minneapolis, MN, USA

Marvin D. Krohn
University of Florida, Gainesville, FL, USA

Diana Kuh
MRC Unit for Lifelong Health and Ageing, London, UK

Chandran Kukathas
University of New South Wales, Campbell, NSW, Australia

Nan M. Laird
Harvard School of Public Health, Boston, MA, USA

Dieter Langewiesche
Universitat Tubingen, Tubingen, Germany

Michael D. Lee
University of California, Irvine, CA, USA

† Deceased.

Brian Leiter
University of Texas, Austin, TX, USA

D.V. Lindley
Minehead, UK

Brunero Liseo
MEMOTEF, Sapienza Università di Roma, Viale Castro Laurenziano, Roma, Italy

Michael L. Littman
Brown University, Providence, RI, USA

Jordan J. Louviere
University of Sydney, NSW, Australia

Gianluca Manzo
GEMASS—CNRS and University Paris-Sorbonne, Paris, France

Clarissa C. Marques
Magellan Health Services, Columbia, MD, USA

Alberto Martinelli
University of Milan, Milan, Italy

Stuart McLean
University of Minnesota, Minneapolis, MN, USA

Volker Meja
Memorial University of Newfoundland, St. John's, NL, Canada

Joel Michell
University of Sydney, Sydney, NSW, Australia

Harvey J. Miller
The Ohio State University, Columbus, OH, USA

Gita D. Mishra
School of Population Health, The University of Queensland, Herston, QLD, Australia

Perveez Mody
King's College, Cambridge, UK

Oliver Helena Fiona Te Momo
The School of Māori Studies, Knowledge & Education (AKE), Massey University, Auckland, New Zealand

Marie H. Monfils
University of Texas at Austin, Austin, TX, USA

William J. Morgan
University of Nottingham, Nottingham, UK

David Moshman
University of Nebraska—Lincoln, Lincoln, NE, USA

Johanna Muckenhuber
University of Graz, Graz, Austria

Peter Müller
Duke University, Durham, NC, USA

Carolina Muñoz-Guzmán
Pontificia Universidad Católica de Chile, Santiago, Chile

Graham Murdock
Loughborough University, Loughborough, UK

Elizabeth Nagel
Stony Brook University, Stony Brook, NY, USA

Bernhard Nebel
Institut für Informatik, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany

Frank Neuner
Bielefeld University, Bielefeld, Germany

Peter Nijkamp
VU University, Amsterdam, The Netherlands

Elke Nowak
Berlin, Germany

Jennifer A. Parsons
University of Illinois at Chicago, Chicago, IL, USA

Bob Pease
Deakin University, Geelong, VIC, Australia

Richard Peet
Graduate School of Geography, Clark University, Worcester, MA, USA

Jacques Pelegrin
Allée de l'Université, NANTERRE cedex, France

John D. Peters
University of Iowa, Iowa City, IA, USA

Richard Price
College of William & Marry, Williamsburg, VA, USA

Tetyana Pudrovska
Population Research Center, University of Texas-Austin, Austin, TX, USA

Douglas W. Rae
Yale University, New Haven, CT, USA

Wolfgang Raible
Freiburg, Germany

Sara Raley
McDaniel College, Westminster, MD, USA

Adrienne Redd
Schools for Sustainability, Philadelphia, PA, USA

Sergio J. Rey
GeoDa Center for Geospatial Analysis and Computation, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ, USA

George Ritzer
University of Maryland, College Park, MD, USA

Les Roberts
University of Liverpool, Liverpool, UK

James M. Robins
Harvard School of Public Health, Boston, MA, USA

Christian M. Rogerson
University of Johannesburg, Johannesburg, South Africa

David Rowe
Institute for Culture and Society, University of Western Sydney, Penrith, NSW, Australia

Tamás Rudas
Faculty of Social Sciences, Eötvös Loránd University, Budapest, Hungary

Debra Satz
Stanford University, Stanford, CA, USA

Anke Schaffartzik
Institute of Social Ecology, Alpen-Adria University, Vienna, Austria

Heinz Schandl
Commonwealth Scientific and Industrial Research Organization, Canberra, ACT, Australia

Naomi Schiller
Temple University, Philadelphia, PA, USA

René L. Schilling
Institut für Mathematische Stochastik, Technische Universität Dresden, Dresden, Germany

Wolfgang M. Schleidt
University of Vienna, Vienna, Austria

Philip Schlesinger
University of Stirling, Stirling, UK

Ralph Schroeder
Oxford Internet Institute, University of Oxford, Oxford, UK

Elisabeth Schwank
Institute for Cognitive Mathematics, University of Osnabrueck, Osnabrueck, Germany

Inge Schwank
Faculty of Mathematics and Natural Sciences, University of Cologne, Cologne, Germany

Masayoshi Shibatani
Rice University, Houston, TX, USA

Burton H. Singer
Princeton University, Princeton, NJ, USA

Gavin Slade
Centre for Criminology and Socio-Legal Studies, University of Toronto, Toronto, ON, Canada

Raymond T. Smith
University of Chicago, Chicago, IL, USA

Yasemin N. Soysal
University of Essex, Colchester, UK

Ervin Staub
University of Massachusetts, Amherst, MA, USA

Rudolf Stichweh
Forum Internationale Wissenschaft, University of Bonn, Bonn, Germany

Reinhard Suck
Universitat Osnabrueck, Osnabrueck, Austria

Jean-Claude Thill
University of North Carolina at Charlotte, Charlotte, NC, USA

Johanna Thompson-Hollands
Boston University, Boston, MA, USA

John E. Toews
University of Washington, Seattle, WA, USA

Istvan G. Tóth
Hungarian Academy of Sciences, Budapest, Hungary

Lucia M. Vaina
Harvard Medical School, Boston University, Boston, MA, USA

Enrico Valdani
Bocconi University, Milan, Italy

Rob A. van der Sandt
University of Nijmegen, Nijmegen, The Netherlands

Lieven Verschaffel
Center for Instructional Psychology and Technology, Leuven, Belgium

Tom Vickers
Northumbria University, Newcastle, UK

Matt Vidal
King's College London, London, UK

Brandilynn Villarreal
University of California Irvine, Irvine, CA, USA

Maureen Walsh
Australian Catholic University, and University of Sydney, NSW, Australia

Linda D. Weaver
Magellan Health Services, Columbia, MD, USA

Sanford Weisberg
University of Minnesota, St Paul, MN, USA

John K. Whitaker
University of Virginia, Charlottesville, VA, USA

Sharlene A. Wolchik
Arizona State University, Phoenix, AZ, USA

Linda Woodhead
Lancaster University, Lancaster, UK

Michael W. Young
Australian National University, Canberra, ACT, Australia

Thomas Ede Zimmermann
Goethe University Frankfurt, Frankfurt, Germany

Liaison and Consultation Psychiatry

Albert Diefenbacher and R Burian, Charité - Universitätsmedizin Berlin, Berlin, Germany; and Evangelisches Krankenhaus Königin Elisabeth Herzberge, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article outlines historical and conceptual aspects, as well as the scope of consultation–liaison psychiatric services in primary and hospital care settings, and in the care of residents with psychiatric disorders in nursing homes. Community studies show that members of the general population who suffer from a physical illness are at increased risk of suffering additionally from a psychiatric disorder. These figures increase in patients seeking help from primary care doctors, and are highest in patients treated in medical–surgical wards of general hospitals. Patients in general hospitals show high rates of comorbid psychiatric disorders, especially organic psychoses, e.g., delirium in dementia, alcohol and other addictions, somatoform and affective disorders including adjustment disorders. Studies show a significant association between psychiatric comorbidity and more complicated illness courses, with longer lengths of stay in hospital and more frequent readmissions, leading to poorer outcomes and increased costs in the treatment of such patients. Inpatients suffering from psychiatric comorbidity have longer lengths of hospital stay than patients with the same severity of the underlying or concomitant physical disease, who are not in addition psychiatrically ill. The high rates of psychiatric comorbidity in physically ill general-hospital inpatients, primary care patients, and nursing home residents are the focus of consultation–liaison psychiatry. Psychiatric care leads to improved management of patients with medical–psychiatric comorbidity, and a reduction of unnecessary costs inflicted on health care systems through underdiagnosis and undertreatment of psychiatric comorbidity. This holds true especially for patients presenting with medically unexplained symptoms not due to organic disease. In order to improve course and outcome of patients with medical–psychiatric comorbidity, psychiatrists have to be involved in their care, as in most of the Western industrialized countries this became possible by the integration of psychiatric departments in general hospitals. General hospitals present an important pathway to specialist psychiatric care for such patients, who will often be seen for the first time by a psychiatrist after having been admitted for a physical disease or for physical symptoms that cannot be explained by an underlying physical disease. Educational and training developments should lead to more elaborate consultation–liaison psychiatric curricula within specialist psychiatric training. Far more, training programs are needed to increase nonpsychiatric physicians' and medical staffs' knowledge of and experience in diagnosis and treatment of psychiatric issues, in order to improve care of patients with somatopsychic disorders in primary and hospital care settings. The development of guidelines for consultation–liaison psychiatric practice and the emergence of joint psychiatric and medical approaches in the care of patients with physical diseases and psychiatric comorbidity, especially within primary care, are expected in years to come.

Consultation–liaison psychiatry delineates the work of psychiatrists with patients who are in medical treatment primarily for somatic reasons in nonpsychiatric departments of a general hospital, or in outpatient care of a nonpsychiatric physician. More specifically, consultation psychiatry describes a practice where psychiatrists see patients treated by other nonpsychiatric physicians only when called upon to give advice about diagnostic and treatment issues to consultees, who in turn are supposed to follow through on these recommendations in the treatment of their patients. In liaison psychiatry, however, psychiatrists are involved more directly in the treatment of physically ill patients with psychiatric comorbidity, assuming more responsibility in joint patient care (e.g., having the prerogative of seeing patients admitted to a pain clinic or an oncological ward without being specifically endorsed to do so in every given case), as well as starting and following through psychiatric or psychotherapeutic treatment. As in everyday clinical care, boundaries between consultation and liaison approaches are not clear-cut, a worldwide consensus has developed to use the phrase consultation–liaison psychiatry (or c–l psychiatry), with the notable exception of the UK, where liaison psychiatry is the only term used for both kinds of approaches. In the United States, c–l psychiatry has been approved by the American Board of Medical Specialties as a subspecialty of psychiatry, referred to as 'Psychosomatic Medicine' (Angelino and Lyketsos, 2011).

Psychiatric Comorbidity in Physically Ill Patients

The need for c–l psychiatric services stems from the high comorbidity of psychiatric illnesses in patients suffering from physical diseases.

Psychiatric Comorbidity in the General Hospital

It is estimated that between 30 and 50% of physically ill patients in general hospitals show comorbid psychiatric disorders, especially organic psychoses (*ICD-10* F0), psychological and behavioral disturbances due to psychotropic substances (*ICD-10* F1), and acute stress reactions (*ICD-10* F4), with percentages of these diagnoses ranging as high as 16, 8 and 10% in medical, and 17, 8 and 13% in surgical patients, respectively (Diefenbacher et al., 2011). Special patient populations – such as those in organ transplantation programs, intensive care units, or oncological patients – show an even higher psychiatric comorbidity. Patients with head and neck cancer, for instance, have been reported to have persistent

depression up to 40% (Hodgkiss and Mascarenhas, 2012). The presence of any psychiatric comorbidity does not always call for immediate psychological treatment at the same time as the underlying or concomitant acute physical disease, which has led to hospital admission, is being cared for by physicians or surgeons. Nevertheless, conservative estimates point to a figure of some 10% of general-hospital inpatients who should receive psychiatric care alongside somatic care during the index episode.

While the evaluation of patients committed to Accident and Emergency Departments after attempting to harm themselves is an important task for psychiatrists, and will often be performed by c–l services, their care is only one part of a larger picture of care for patients suffering from medical–psychiatric comorbidity, with the latter impacting negatively on course and outcome of the former. A prerequisite for psychiatric services to achieve an impact on this group of patients is the integration of psychiatry into general hospitals. Thus, in the United States, where the foundation of psychiatric departments as integrated parts of general hospitals started as early as the 1920s, c–l psychiatry has a long tradition, whereas in the Federal Republic of Germany, for instance, where state mental hospitals have given way to general hospital psychiatric departments only since the end of the 1970s, c–l psychiatry started to gain shape only toward the end of the twentieth century. Overall, international interest in c–l psychiatry increased outside the United States during the 1980s and 1990s with special emphasis in the UK, the Netherlands, Australia, and New Zealand, where c–l psychiatric sections were founded within national psychiatric organizations. More recently, in other European countries, e.g., Switzerland, and on a European level, as represented by the European Association of Psychiatry, sections for c–l psychiatry have been founded (Lobo et al., 2007; Georgescu et al., 2009).

The history of c–l psychiatry shows that in order to achieve acceptance of psychiatric care for patients with medical–psychiatric comorbidity, regular availability of psychiatrists in general hospitals is very helpful, as this provides the opportunity to demonstrate the added value of psychiatric diagnosis and treatment to other medical disciplines in the care of comorbid patients.

Psychiatric Comorbidity in the Primary Care Sector

While c–l psychiatry developed primarily in the general hospital setting, there was growing emphasis worldwide on diagnosis and treatment of psychiatric comorbidity of patients in the primary care sector toward the end of the twentieth century. Results from the *National Institute of Mental Health Epidemiological Catchment Area Program* show that about 16% of the community-dwelling adult population fulfills criteria for at least one psychiatric disorder. This figure increases to at least 21–26% of primary care patients. Up to 90% of patients with depression and anxiety disorders are estimated to be treated solely in primary care (Archer et al., 2012). In most cases, these patients present with physical symptoms that are not caused by physical disease (*see* Somatoform Disorders). The proportion of undetected and untreated psychiatric (co)morbidity in this patient group is high (Goldberg, 2010).

Psychiatric Morbidity in Nursing Homes

The percentage of the population older than 65 years of age is increasing. It is estimated that 20–50% of these people will eventually live in nursing homes. The very high prevalence of psychiatric-related illness among those in nursing homes is documented in a number of studies, showing that up to 70% of nursing home residents may meet diagnostic criteria for dementia, and 90% may have at least one form of behavioral disturbance, such as agitation, delusions, or depression. Depression is largely undetected and undertreated in this population, despite the fact that the majority of cases treated, e.g., with antidepressants, will recover from a depressive episode. Only recently, there has been growing awareness that people with learning disability show a significantly higher percentage of psychiatric comorbidity than the average population. While attempts are made to achieve a close to normal living situation for a large part of this clientele, people with severe mental retardation – which is often complicated not only by psychiatric comorbidity, but also by other physical diseases such as epilepsy – will receive adequate care mostly in nursing home settings (Martinez-Leal et al., 2011).

In sum, there has been increasing awareness of hidden mental comorbidity in the general hospital, primary care, and nursing home sectors and of the need for specialist psychiatric treatment in a significant percentage of patients or residents being treated or living in those settings.

Effects of Psychiatric Comorbidity in Physically Ill Patients

Studies show a significant association between psychiatric comorbidity and more complicated illness courses, both for patients in the primary care sector and for general-hospital inpatients. Patients with somatization disorder use nine times the health care services of the general population, panic disorder patients have 10 times the number of visits to hospital emergency rooms than nonanxious patients, and 70% will see more than 10 physicians before an accurate diagnosis is made. Medically ill patients with depression have a poorer prognosis than nondepressed patients with equal somatic morbidity. For example, postmyocardial infarction patients with depression are at significantly greater risk of death over a period of 6 months following the infarction, and depressed stroke patients show poorer results in rehabilitation programs with more persistent disability than nondepressed stroke patients. It has been shown that a small percentage of patients are responsible for the use of a major portion of health care, with, for example, 10% of patients of a large health maintenance organization in the United States consuming approximately one-third of all primary care visits, 52% of all speciality visits, and 40% of in-hospital days. Fifty percent of those high utilizers of health care showed some form of psychiatric comorbidity, e.g., 20.2% somatization disorder, 33.6% generalized anxiety disorder or panic disorder, and 40% major depression and dysthymia. The finding of high psychosomatic morbidity in frequent health care utilizers was confirmed in several studies (e.g., Ferrari et al., 2008).

In general-hospital inpatients, it has been shown that medically ill patients with psychiatric comorbidity show longer lengths of stay in hospitals, more frequent readmissions, and

an overall increase in costs for treatment as compared to patients without psychiatric comorbidity. Adjustment for other factors associated with longer length of stay and rehospitalizations, such as severity of underlying medical disease, did not reduce the significance of the negative impact exerted by psychiatric comorbidity (Diefenbacher, 2000). The elderly are exposed to special risks, with about 10% of elderly patients admitted to a general hospital suffering from delirium. About 30% of elderly medical inpatients have dementia, which makes more complex management necessary, with problems arising from, for example, diminished compliance with medication, and in following through with rehabilitation programs (Teodorczuk et al., 2010). With regard to limited financial health care resources, this problem is more and more recognized in the medical setting and models of care have been developed, which include improvement of prevention, early detection and, management of psychosomatic comorbidity in the elderly. These models are often nurse-led and involve psychiatric support by multiprofessional c–l teams in stepped care approaches (Gurlit et al., 2013).

The Practice of Consultation and Liaison Psychiatry

The scope of c–l psychiatric services spans primary and hospital care settings, the care of elderly residents in nursing homes, and of clients with learning disability and accompanying psychiatric disorders.

C–L Psychiatric Services in the General Hospital

Despite the high percentage of psychiatric comorbidity in the general hospital, referrals to psychiatric c–l services amount at most to only a small number of physically ill inpatients, approaching an average number of 1–2% of general-hospital inpatients worldwide. Usually, the majority of patients is referred by internal medicine departments, with a large percentage being older than 65 years of age. Despite all differences in service delivery that are due to local necessities in the provision of care for given populations, c–l psychiatrists most often diagnose delirious states and other disorders of cognition, followed by depressive syndromes and adjustment disorders, fitting well to the above mentioned prevalence figures of psychiatric comorbidity in general-hospital patients. In contrast, despite their obvious frequency, patients with alcohol-use disorders are referred much less by their treating physicians. The low rate of referrals to psychiatric c–l services has led to the assumption that nonpsychiatric physicians do not recognize psychiatric disorders sufficiently, or, if they do, do not think of psychiatric interventions as being helpful to their patients. The concept of liaison psychiatry properly takes its vantage point from this discrepancy: other than the usual psychiatric consultation with its mainly emergency psychiatric character that by default is only able to care for the tip of the iceberg of psychiatric comorbidity, i.e., patients having passed the filter of recognition of a nonpsychiatric physician. Liaison psychiatric approaches aim at the denominator of psychiatric comorbidity, triggering specialist psychiatric care, e.g., by screening patients on hospital admission, or identifying problem behavior in joint rounds with nonpsychiatric physicians. More recent c–l developments show the advantages of multiprofessional c–l teams including psychiatrists, as well as mental health nurses (MHN), social workers, and clinical psychologists. In these models, MHN play a leading role not only in caring for medically ill patients with psychiatric comorbidity, but also in providing a role model and peer-to-peer training for medical staff (Burian et al., 2014; Sharrock et al., 2008). Thus, liaison psychiatric approaches try to go well beyond the 'firefighter function' of typical psychiatric consultation, aiming at changing referral patterns of nonpsychiatric departments, e.g., enabling physicians to deal with and/or prevent the development of delirious states, or increase referral rates of depressed patients or patients with alcohol-use disorders by demonstrating the efficacy of even time-limited psychiatric interventions (RCPhys and RCPsych, 2003).

Both in consultation and, to a somewhat lesser extent, in liaison psychiatric approaches, psychiatrists have to be aware of the stages a psychiatric consultation will typically go through. In the initial phase and due to uncertainty about how to deal with a patient, the consultant's referral diagnosis might be rather imprecise. In such cases, it must be the task of the c–l psychiatrist to help in working out what is in fact the problem. The ensuing intervention must then be, other than in a typical psychiatric office situation, directed not only toward the patients, but also take into consideration the whole system, including the ward doctor, nurses, and relatives. Otherwise, c–l psychiatric interventions, when performed in isolation and/or out of context, might be unsuccessful (Diefenbacher, 2000).

C–L Psychiatric Services in Primary Care and in Nursing Homes

Along with the growing interest in psychiatric (co)morbidity in primary care patients and nursing home residents, international c–l psychiatry shows a remarkable shift from the general hospital setting to collaborative care practices in the primary care setting, and, though still in its beginnings, to liaison psychiatric services with nursing home populations. For primary care settings, studies show that inclusion of a nurse-therapist and a psychiatrist in primary care treatment teams, e.g., on the basis of regularly performed nurse-based visits as well as consultations by a psychiatrist in a stepwise approach, leads to a marked improvement in the symptoms of patients with somatoform disorders or depression compared to provision of services by primary care physicians alone (van der Feltz-Cornelis et al., 2010). For nursing home populations, the need to consider psychopharmacological as well as, and more importantly, psychosocial interventions such as supportive environmental therapies, recreational therapies, and so on, in combined treatment plans tailored to specific residents and specific nursing home settings, stress the need for psychiatrists to oversee the implementation of such integrated approaches. Hence, liaison models with more intense involvement of psychiatrists working with other nonpsychiatric physicians and psychologists, as well as administrative and nursing staff in addition to care for residents and their relatives will be developed in the near future. The latter holds especially true for people with learning disability, as diagnostic and treatment approaches in this clientele will require special expertise not necessarily familiar to primary care physicians, or even psychiatrists themselves (Dagnan, 2007). Especially in patients

who are not able to communicate verbally, unexplained behavioral changes or challenging behavior may reflect an underlying psychiatric disorder, which in turn may be treatable, as in the case of hostile behavior in demented patients with dysphoric depression that may well respond to pharmacological treatment with, for example, selective serotonin reuptake inhibitors. C–l psychiatrists working in such settings are ideally placed to teach nursing staff and primary care physicians that expert assessment of behavioral and psychological symptoms in nonverbal patients may lead to more specific management and hence better patient care. Thus a patient with dementia whose delusional experiences lead to bizarre behavior should be treated differently from a demented patient who suffers from pain, or a depressed nonverbal patient who refuses aggressively to accept food. Of note, in the nursing home setting, but also due to demographic changes with a further increase of frail elderly patients in the general hospital, geriatricians also play an important and ever increasing part in the care for this population, and c–l psychiatrists should seek to establish joint approaches with geriatric services, when available.

Despite the increasing interest in models of collaborative care in the outpatient, as well as nursing home sectors, it has to be kept in mind, that, at least at this point in time, many patients with somatopsychic comorbidity are seen for the first time at all by a psychiatric specialist only after having been admitted to a general hospital. Thus, psychiatrists as well as planners of health services will have to bear in mind that general hospitals may be referred to as 'virtual psychiatric sectors.' This underlines the importance of the general hospital as a pathway to specialist psychiatric care.

Evaluation of C–L Psychiatric Interventions

Psychiatric care leads to improved management of medically ill patients with psychiatric comorbidity. This is the rationale of the practice of c–l psychiatry, whose benefits can be demonstrated within a value matrix, taking into account treatment outcomes in medically ill patients with psychiatric comorbidity, as well as economic savings by adding specialist psychiatric treatment to standard medical treatment. Cost-benefit issues of consultation psychiatry service delivery have been demonstrated initially in general-hospital inpatients, with primarily US studies showing a trend to reduced length of hospital stay with corresponding reductions in cost of treatment in medical patients after a psychiatric liaison service was inaugurated, compared with control groups without such an intervention program. However, as other studies failed to show such an effect, the overall conclusion was that this research approach had to be judged as inconclusive in the face of a sharp decrease in lengths of hospital stays worldwide due to secular trends in shifting the provision of health care into the outpatient sector, and the inauguration of diagnosis related groups (DRG)-based payment systems, in order to reduce the escalation of overall health care costs (Diefenbacher and Strain, 2002). Consequently, research on the efficacy of psychiatric intervention strategies in patients with psychiatric (co)morbidity in the primary care sector has increased. Studies showed that integrated c–l psychiatric approaches, but not those that only offered just screening strategies to identify patients with psychiatric disorders, have led to improved outcomes in patient care. In the inpatient setting, models of integrated medical and psychiatric care also are being proposed. It could be shown that multiprofessional liaison approaches lead to reduced incidence and early detection of delirium especially in older patients in the general hospital. These interventions were mostly nurse-based and train medical staff on internal-surgical wards to manage patients with elevated risk for delirium. Such approaches may lead to better outcomes and prevention of prolonged hospital treatment due to complications of delirium (Leentjens et al., 2011).

Even in outpatient primary care settings, best results are achieved with multiprofessional and integrated care approaches that make all professionals involved in patients therapy work collaboratively together. According to structured treatment plans and follow-up schedules, trained nurses who apply psychosocial interventions to the patients play a major role. Psychiatrists are involved as specialists who supervise the treatment process and apply patient visits in a stepped care plan only to those who do not get better in the baseline treatment provided by patient's general practitioners (GPs) and trained nurses. Collaborative care models have been developed for depression, anxiety, and somatoform disorders. These approaches show benefits for patients' well-being, and also cost-efficacy (Archer et al., 2012; Ekers et al., 2013). Far more, collaborative care models also have an impact on the somatic morbidity. For instance, heart failure patients who were treated for depression in a collaborative care approach had significantly less adverse coronary events than patients with the same comorbidity who were treated in usual care (Davidson et al., 2010).

Future Developments in C–L Psychiatry

Future developments will see an increased emphasis on the education of both psychiatrists and primary care practitioners in collaborating in the treatment of medically and psychiatrically ill patients, in the development of models of integrated care, and the need to develop and implement guidelines and standards of care in quality assurance programs for patients suffering from somatic and psychiatric disorders alike. Collaborative care models in primary care as well as multiprofessional nurse-based services in general-hospital inpatients settings that have been proven as successful and cost saving, should now be rolled out in daily clinical practice.

Education

The education of physicians working at the interface of medical–surgical and psychiatric medicine has to be twofold. Psychiatrists have to learn how to treat medically ill patients with psychiatric comorbidity. Psychiatric training cannot remain restricted to gaining experience predominantly with patients with schizophrenia or manic-depressive illness, but has to take into account the needs of patients suffering from medical illnesses and comorbid psychiatric illness, for example, depressive patients suffering from cardiac disease or stroke or kidney failure, who require somatic as well as psychological treatment. To achieve this goal, c–l psychiatric curricula are

being developed within specialist psychiatric training, e.g., in Europe and the United States. Psychiatrists will have to become proficient in mastering diagnostic and treatment problems arising from somatopsychic comorbidity, e.g., address problems of diagnosing depressive disorders in the medically ill, and adjusting psychological as well as psychopharmacological treatment issues for patients suffering from both medical and psychiatric disorders.

On the other hand, c–l psychiatrists will have to add to the education and training of primary care physicians and medical staff. Training programs have been developed to broaden nonpsychiatric physicians' and nurses' knowledge of and experience in diagnosis, referral, and treatment of patients with psychiatric (co)morbidity, in order to improve the care of patients with somatopsychic disorders in primary and hospital care settings. Psychiatrists, as well as nonpsychiatric physicians, will have to adopt behavioral medicine treatment strategies for patients with, for example, somatization disorders and chronic pain syndromes, and, most importantly, will have to learn treatment strategies other than psychotropic drug prescribing strategies such as cognitive behavioral treatment for patients with adjustment disorders, anxiety disorders, and the milder depressive disorders (the so-called subthreshold disorders), as these disorders make up the bulk of psychiatric comorbidity treated in the primary care sector. Psychopharmacological treatment has to be applied with caution. Recent studies showed that their effect on depressive and anxious symptoms is depending on symptom severity at baseline. Mild and moderate depressed patients benefit from placebo as good as from medication. These findings indicate that non-pharmacological interventions can be applied, so patients will be not exposed to the risk of adverse drug reactions (Fournier et al., 2009).

Guidelines and Quality Assurance

As the awareness for an improved psychological care of medical and surgical patients increased toward the end of the twentieth century, one step has been the development of guidelines in order to define standards of care. In the UK, conjoint working parties of physicians, surgeons, and psychiatrists (e.g., Royal Colleges of Physicians, and Psychiatrists, 2003) published documents with practical suggestions regarding methods of identifying potentially harmful psychological components of patients' illnesses, and recommendations for the provision of care as well as for referral for psychiatric or psychological help when appropriate. As only a minority of patients will need psychiatric care in addition to medical–surgical treatment, these guidelines underline that normally physicians and surgeons should be able to make some assessment of psychological factors as integrated part of their daily routine. The guidelines propose simple screening mechanisms in order to improve detection of psychological disorders in the general hospital, for example, the *CAGE Questionnaire* for screening for alcohol misuse or the *Confusion Assessment Method (CAM)* for the early detection of delirium, to be used by medical–surgical physicians and nursing staff alike. Overall, the implementation of guidelines is thought to help in a number of areas, such as in reducing the number of (unnecessary) investigations performed for physical symptoms that result from underlying stress, as, for example, in patients with somatization, help in relieving symptoms of stress and improving the quality of life of patients with serious physical illness.

It is assumed that the implementation of guidelines within quality assurance programs will lead eventually to improved patient care, although guidelines alone do not work unless there is sufficient training and a clear protocol with scheduled follow-ups (NICE, 2010; Young and George, 2003). However, measures against which to gauge c–l psychiatric interventions have yet to be established. Given the low referral rate for psychiatric consultation as compared with the high prevalence of medical–psychiatric comorbidity, and the need for specific psychiatric treatment in general-hospital inpatients, as a first step to improving care, the C–l Psychiatric Section of the Royal Australian and New Zealand College of Psychiatrists proposes to define the benchmark for psychiatric referrals to be 5% of all admissions to medical–surgical wards of a general hospital. If this percentage is not reached, an audit should be performed to establish whether there are identifiable reasons for not providing psychiatric support in an adequate number of patients in medical–surgical departments of a given general hospital.

Research in C–L Psychiatry

Basic requirement for research in c–l psychiatry is the identification of high-risk patient groups, such as those suffering from a significant amount of psychiatric comorbidity. Failure to identify such high-risk groups, along with the overstated assumption that just one psychiatric consultation might lead to improved patient outcomes, may have been the basis of the ambiguous results of general-hospital inpatient intervention studies performed in the United States in the 1980s. C–l research in the primary care setting, on the other hand, has demonstrated that restriction to a high-risk group with identifiable psychiatric comorbidity, e.g., high utilizers in primary care suffering from a psychiatric disorder, such as somatoform disorder, may demonstrate that a psychiatric intervention can lead to improved patient outcomes, if the intervention is intense enough. There is good evidence that collaborative care models as described above reduce health care utilization in primary care patients treated for depression and somatoform disorders (van der Feltz-Cornelis et al., 2010).

Problems that need to be solved in order to improve care of patients with medical–psychiatric comorbidity are:

1. To develop a more precise description of the phenomenology of psychiatric illnesses comorbid with particular physical illnesses, e.g., as in the case of the differential diagnosis of demoralization versus depression in patients severe and fatal medical disorders like cancer (Price and Hotopf, 2012).
2. To further study efficacy and effectiveness of psychotherapeutic and psychopharmacological treatment strategies in patients with medical–psychiatric comorbidity. Meta-analyses point out that multiprofessional psychiatric treatment, especially when delivered in collaboration with patients attending somatic doctor or GP, is as effective in depressive patients with medical comorbidity as in those without (Archer et al., 2012).

3. To further study service delivery for migrant patients and ethnic minorities. Globalization demands adaptation of care models to the specific needs of these patients. Migration is known to be a risk factor for psychiatric morbidity and migrants often face barriers for access to health care and mental health care. Far more, the misinterpretation of culturally different presentations of mental health problems, e.g., bodily and pain symptoms as signs for underlying depression, may lead to underdiagnosis of psychiatric disorders in such patients in primary care or general hospitals.
4. To develop new models of conjoint somatic and psychological care within general-hospital inpatient settings that must take steps to be prepared for an increasing number of frail elderly patients who suffer from dementia and will react with behavioral disturbances or even the development of delirium due to medical interventions and/or the, at present, not overtly supportive environment they are put into without having time to adapt themselves (cf Section Integration of Medical and Psychiatric Services) (Bartels and Naslund, 2013).

The effectiveness of different models of care for medical patients with psychiatric comorbidity with regard to different outcome measures such as clinical improvement, quality of life, or cost effectiveness, will require further investigation, especially in inpatient settings for different patient populations, e.g., acute versus chronic care patients, with regard to differential effects of more traditional consultation or liaison approaches versus collaborative care as described above.

Integration of Medical and Psychiatric Services

A model for integrated medical and psychiatric treatment programs within the general hospital were the so-called medical–psychiatric units that were developed primarily in the United States during the 1990s, but they remained controversial in terms of cost-efficacy and did not achieve a significant impact within the departments of somatic medicine, as they were mostly integrated within psychiatric departments. But as the need for specialized care for high-risk patients, for example, delirious patients and somatically ill patients with dementia and behavioral disturbances who cannot be managed on medical–surgical wards, or suicidal patients in need of acute medical treatment that cannot be provided on a psychiatric ward, is increasing, there are attempts to put such ideas of conjoint care on a broader basis. It has been proposed recently to move on and develop the so-called Complexity Intervention Units, beyond departmental boundaries and outside psychiatric departments (Kathol et al., 2009). Such conjoint models of care in the inpatient setting should lead to improved patient outcome while saving costs, as compared with traditional models of sequential care with referrals from psychiatric to medical departments and vice versa. Apart from the development of pharmacological treatment strategies to prevent the development of postoperative delirium that is increasingly advocated (Teslyar et al., 2013), recent developments favor mobile approaches in which multiprofessional and mostly nurse-based teams support the attending teams on the somatic wards in the management of complex patients with high degrees of medical–psychiatric comorbidity, such as 'delirium teams' or 'self-harm teams.' These teams shall also take responsibility for the continuity of care once the patients are discharged from hospital back into their community.

Epilogue

Readers interested in widening their knowledge of the topics dealt with in this article are referred to textbooks of consultation–liaison psychiatry and psychosomatic medicine (e.g., Levenson, 2011; Creed et al., 2011). Journals like 'Psychosomatics,' 'General Hospital Society,' and the 'Journal of Psychosomatic Research' are dedicated to the field of medical–psychiatric comorbidity. The homepages of national and international psychiatric and psychosomatic associations are frequently updated to deliver information on congresses, symposia, and training workshops in c–l psychiatry and psychosomatic medicine (www.europsy.net, www.eapm.eu.com, www.apm.org).

See also: Geriatric Psychiatry; Hospitals and Health: Psychosocial Aspects; Psychosomatic Medicine; Somatoform Disorders.

Bibliography

Angelino, A., Lyketsos, C.G., 2011. Training in psychosomatic medicine: a psychiatric subspecialty recognized in the United States by the American board of medical specialties. Journal of Psychosomatic Research 71, 431–432.

Archer, J., Bower, P., Gilbody, S., Lovell, K., Richards, D., Gask, L., Dickens, C., Coventry, P., 2012. Collaborative care for depression and anxiety problems. Cochrane Database of Systematic Reviews (10). http://dx.doi.org/10.1002/14651858.CD006525.pub2. Art. No.: CD006525.

Bartels, S.J., Naslund, J.A., 2013. The underside of the Silver Tsunami – older adults and mental health care. New England Journal of Medicine 368, 493–496.

Burian, R., Protheroe, D., Grunow, R., Diefenbacher, A., 2014. Establishing a nurse-based psychiatric CL service in the accident and emergency department of a general hospital in Germany. Nervenarzt. http://dx.doi.org/10.1007/s00115-014-4069-8.

Creed, F., Hennigsen, P., Fink, P. (Eds.), 2011. Medically Unexplained Symptoms, Somatization and Bodily Distress. Cambridge University Press, Cambridge.

Dagnan, D., 2007. Psychosocial interventions for people with intellectual disabilities and mental ill-health. Current Opinion in Psychiatry 20, 456–460.

Davidson, K.W., Rieckmann, N., Clemow, L., et al., 2010. Enhanced depression care for patients with acute coronary syndrome and persistent depressive symptoms: coronary psychosocial evaluation studies randomized controlled trial. Archives of Internal Medicine 170 (7), 600–608.

Diefenbacher, A., 2000. Consultation and liaison psychiatry. In: Helmchen, H., Henn, F., Lauter, H., Sartorius, N. (Eds.), Contemporary Psychiatry, General Psychiatry, vol. 1, Part 2. Springer, Berlin, pp. 253–267.

Diefenbacher, A., Strain, J.J., 2002. Consultation-liaison psychiatry: stability and change over a 10-year-period. General Hospital Psychiatry 24 (4), 249–256.

Diefenbacher, A., Burian, R., Klesse, C., Härter, M., 2011. Konsiliar- und Liaisondienst für psychische Erkrankungen. In: Berger, M. (Ed.), Psychische Erkrankungen Klinik und Therapie, fourth ed. Verlag, München, Urban und Fischer.

Ekers, D., Murphy, R., Archer, J., et al., 2013. Nurse-delivered collaborative care for depression and long-term physical conditions: a systematic review and meta-analysis. Journal of Affective Disorders. http://dx.doi.org/10.1016/j.jad.2013.02.032.

Ferrari, S., Galeazzi, G.M., Mackinnon, A., Rigatelli, M., 2008. Frequent attenders in primary care: impact of medical, psychiatric and psychosomatic diagnoses. Psychotherapy and Psychosomatics 77 (5), 306–314.

Fournier, J.C., DeRubeis, R.J., Shelton, R.C., et al., 2009. Prediction of response to medication and cognitive therapy in the treatment of moderate to severe depression. Journal of Consulting & Clinical Psychology 77 (4), 775–787.

van der Feltz-Cornelis, C.K., Van Os, T.W.D., Van Marwjik, H.W.J., Leentjens, A.F.G., 2010. Effect of psychiatric consultation models in primary care. A systematic review and meta-analyses of randomized clinical trials. Journal of Psychosomatic Research 68, 521–533.

Georgescu, D. (Chair), Cimilli, C., Garyfallos, G. (Working Group of UEMS European Board of Psychiatry), 2009. Report of the European board of psychiatry: consultation liaison psychiatry in Europe. http://www.uemspsychiatry.org/board/reports/2009-Oct-CL%20Psychiatry%20in%20Europe.pdf (accessed 25.01.13.).

Goldberg, D., 2010. The detection and treatment of depression in the physically ill. World Psychiatry 9, 16–20.

Gurlit, S., Thiesemann, R., Wolff, B., Brommer, J., Gogol, M., 2013. Caring for People with Dementia in General Hospitals: An Education Curriculum from the Alzheimer's Society of Lower Saxony. Germany. Z Gerontol Geriatr 46 (3), 222–225.

Hodgkiss, A., Mascarenhas, S., 2012. Psycho-oncology. In: Guthrie, E., Rao, S., Temple, M. (Eds.), Seminars in Liaison Psychiatry, second ed. Royal College of Psychiatrists Publications, London, pp. 186–204.

Kathol, R.G., Kunkel, E.J.S., Weiner, J.S., et al., 2009. Psychiatrists for medically complex patients: bringing value at the physical health and mental health/substance-use disorder interface. Psychosomatics 50, 93–107.

Leentjens, A.F.G., Rundell, J.R., Rummans, T., et al., 2011. Delirium: an evidence-based medicine (EBM) monograph for psychosomatic medicine practice, commissioned by the Academy of Psychosomatic Medicine (APM) and the European Association of Consultation and Liaison Psychiatry and Psychosomatics. Journal of Psychosomatic Research 73, 149–152.

Levenson, J. (Ed.), 2011. Psychosomatic Medicine Psychiatric Care of the Medically Ill. The American Psychiatric Publishing, second ed. Washington, DC; and London.

Lobo, A., Lozano, M., Diefenbacher, A., 2007. Psychosomatic psychiatry: a European view. Journal of Psychiatry 21 (2), 65–80.

Martinez-Leal, R., Salvador-Carulla, L., Linehan, C., et al., 2011. The impact of living arrangements and deinstitutionalization in the health status of persons with intellectual disability in Europe. Journal of Intellectual Disability Research 55, 858–872.

NICE – National Institute of Clinical Excellence, 2010. Clinical Guideline 103 Delirium: Diagnosis, Prevention and Management. http://www.nice.org.uk/nicemedia/live/13060/49909/49909.pdf.

Parsonage, M., Fossey, M., 2012. Economic Evaluation of a Liaison Psychiatry Service. Mental Health Network, NHS Confederation. http://www.centreformentalhealth.org.uk/pdfs/liaison_psychiatry_in_the_modern_NHS_2012.pdf (last accessed 08.04.13.).

Price, A., Hotopf, N., 2012. Palliative care psychiatry. In: Guthrie, E., Rao, S., Temple, M. (Eds.), Seminars in Liaison Psychiatry, second ed. Royal College of Psychiatrists Publications, London, pp. 186–204.

Royal Colleges of Physicians and Psychiatrists, 2003. Joint working party report: the psychological care of medical patients: a practical guide, 2. Royal College of Psychiatrists Council Report No. CR 108, London.

Sharrock, J., Bryant, J., McNamarra, P., Forster, J., Happell, B., 2008. Exploratory study of mental health consultation-liaison nursing in Australia: part 1 demographics and role characteristics. International Journal of Mental Health Nursing 7 (5), 180–188.

Teodorczuk, A., Welfare, M., Corbett, S., Mukaetova-Ladinska, E., 2010. Developing effective educational approaches for liaison old age psychiatry teams: a literature review of the learning needs of hospital staff in relation to managing the confused older patient. International Psychogeriatrics 22 (6), 874–885.

Teslyar, P., Stock, V.M., Wilk, C.M., et al., 2013. Prophylaxis with antipsychotic medication reduces the risk of post-operative delirium in elderly patients: a meta-analysis. Psychosomatics 54, 124–131.

Young, L.J., George, G., 2003. Do guidelines improve the process and outcomes of care in delirium? Age Ageing 32, 525–528.

Liberal Christianity

Linda Woodhead, Lancaster University, Lancaster, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 3, pp. 1767–1770, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

Liberal Christianity arose in the seventeenth and eighteenth centuries in response to the Enlightenment and modernism. The progressive force of liberal Christianity concerned itself with rationalism and individual liberty, in response to authoritarian fundamentalism and antimodernism. While earlier treatments argued that liberal Christianity was unsustainable, contemporary developments have shown that it remains a major feature of twenty-first century Christianity and that it enjoys considerable influence throughout the West.

'Liberal' is the designation given to a broad trajectory in Christianity by sympathizers and critics alike. In some of its manifestations it is also referred to as 'modernism.' Both terms draw attention to a defining characteristic: a dissatisfaction with earlier forms of religion and a concern to replace them with less restrictive alternatives more open to accommodating secular society. While it has rarely led to the establishment of new churches, liberalism has had a significant impact within all the mainline Christian denominations in the West.

Understood in this broad sense, liberal Christianity can be seen to date back even before the Renaissance to the fifteenth century, when the term 'modern' was first invoked in relation to reforming movements in theology and spirituality like the 'Via Moderna' and the 'Devotio Moderna.' Since then Western Christianity has been challenged repeatedly by individuals and movements who have championed the freedom of the individual Christian from the authority of institutional religion. Yet it is in the modern period that liberal Christianity has become most prominent and has played its most important role in shaping modernity itself.

By the end of the twentieth century, something of a consensus had developed among sociologists of religion that liberal Christianity was in inexorable decline. This consensus coincided with a period when theological liberalism had fallen out of fashion; however, this view has been challenged in the new millennium as conservative Christianity has been shown to be less ascendant than was assumed, and many liberal religious groups have seen considerable growth or stability (Sherkat, 2014). The dismissive view of liberal Christianity results from three critical oversights: (1) it overlooks the internal variety of liberal Christianity; (2) it ignores evidence of its continuing vitality on the ground; and (3) it fails to recognize that liberal Christianity remains well adapted to many of the socioeconomic formations and cultural trends of the modern world.

Varieties of Liberalism

Rationalist

One of the most important ways in which liberal Christianity is implicated with the rise of modernity is through its central role in the Enlightenment of the mid-seventeenth to mid-eighteenth centuries. In their different ways, such figures as the Deists, Thomas Jefferson, Tom Paine, Rene Descartes, John Locke, and Immanuel Kant were all concerned with the liberal reform of religion. They helped shape a liberal Christianity, which was characterized by (1) hostility to 'traditional' religion, conceived as superstitious, heteronomous, and divisive; (2) confidence in human reason and the primacy of the sovereign individual; (3) activist optimism about the possibilities of human and social improvement; (4) belief in the harmonious unity of all true ('natural') religion; and (5) high valuation of freedom. Rationalist liberalism tended to be primarily an intellectual movement, though it was often related closely to political radicalism. At the institutional level, it gave rise to the Unitarian and Universalist churches, whose rejection of first the divinity of Jesus and later of deism could classify them as a new religious movement (Sherkat, 2014).

Romantic

While many sociologists identify liberal Christianity with its rationalist forms, there are other important varieties. By the late eighteenth century, for example, a Romantic liberalism had begun to exercise an important cultural and religious influence. In America, this was best represented by Transcendentalism, and in Germany by the theology of Friedrich Schleiermacher. Romantic liberalism inherited rationalist liberalism's belief that the individual rather than the institution was the locus of true religion ("my mind is my church," as Tom Paine had put it), but stressed the authority of feeling, imagination, experience, and self-consciousness rather than reason. This stress on the importance of individual experience of God has been carried through into the twentieth century in the work of theologians like Rudolf Bultmann and Paul Tillich.

Ethical and Social

Kant had located religious authority in moral reason. This 'ethicization' of the Christian religion was taken further in Germany by theologians like Albrecht Ritschl, Wilhelm Herrmann, and Adolf von Harnack, and in America by theologians like Horace Bushnell and Walter Rauschenbusch. The last of these, the leader of the Social Gospel movement, was responsible for interpreting Christianity as a force for social reform in industrial society.

Liberation and Feminist

In the twenty-first century, the influence of liberal Christianity continues to be felt in new reform movements within the churches and theology, most notably in some important varieties of Liberation theology (a product of Latin America) and Feminist Theology (a product of North America and Western Europe). In both, we find a characteristic emphasis on human liberation and the authority of experience, combined with a thoroughgoing criticism of existing structures of power and domination in both the churches and wider society, and a bias toward the oppressed and marginal. The importance of liberation theologies has been reinforced by current developments in the Catholic Church, particularly under Pope Francis.

The Development of Modern Liberalism

Liberalism has had the greatest impact within mainline Protestant churches, and has been particularly influential in countries where Protestantism has been dominant. Most commentators consider its heyday to have been from the mid-nineteenth century through the 1920s. Liberalism flourished because it was able to meet the new challenges posed by the forces of modernization. Its critical stance toward Christian tradition enabled it to assimilate the rise of historical method and its application to the Bible. Equally, its long-held belief in progress and the capacity of human reason to fathom the mysteries of God and the world helped it embrace the discoveries of modern science, including evolution. What is more, its libertarianism and individualism enabled it to support the interests of the new middle classes and to play a legitimating role in relation to democracy and the modern state. Far from being threatened by rapid modernization in the nineteenth century, liberal Christianity tended to view itself as the religious and moral engine of social progress. Nowhere was this truer than in the United States. In Catholic countries, the picture was very different. There, led by an increasingly defensive Rome, the church explicitly repudiated the errors of 'modernism' (which included political and religious liberalism), and allied itself with the forces of reaction. In 1907, Pope Pius X condemned those Catholic 'modernists' who had embraced aspects of theological liberalism, and introduced an antimodernist oath for the clergy which effectively ended all attempts to develop a liberal Catholicism. It was not until the Second Vatican Council of 1962–65 that a moderate liberalism gained official sanction. Since then there has been an explosion of liberal thought in the Catholic Church, despite Pope John Paul II's and Pope Benedict XVI's attempts to curb its influence. Indeed, the installation of Pope Francis marks a renewed liberalism in the Catholic Church. The work of theologians like Hans Kung, as well as the rise of Liberation Theology and Feminist Theology, has been particularly notable.

The twentieth century has brought unprecedented challenges for liberal Christianity. These include the rise of conservative evangelical and fundamentalist Christianity and, at the theological level, the rise of neoorthodoxy. At the same time, many elements of the liberal creed – including belief in progress, reason, 'humanity,' and the ideal of religious unity – have been shaken both inside and outside the churches. This together with shrinking attendance in the mainline Protestant churches most influenced by liberalism has led many commentators to conclude that it is in terminal decline. What this conclusion overlooks, however, is that the mainline denominations still account for the vast majority of Christians (especially if one includes the Catholic Church), and that many within these denominations (including the Catholic Church) continue to embrace some version of liberal Christianity. What is more, liberalism remains extremely well adapted to societies that show no real sign of diminishing their commitment to freedom, equality, democracy, and a liberal individualism reinforced by the institutions of a free-market economy.

Liberalism on the Ground

Dean Kelley's book *Why Conservative Churches Are Growing* (1972) captured a new mood in the sociology of religion. Where secularization theory had previously encouraged the conclusion that the more traditional and antimodern forms of religion would be those that would decline most rapidly, the growth of conservative and sectarian bodies alongside evidence of the decline of more liberal denominations led to a change of mind after the 1960s. In the late twentieth century, sociological theory was quickly mobilized to explain this change, and to reinforce the prediction of further liberal decline.

Much of the case for liberal decline focuses on the American example. It is clear from survey data in the United States, like the General Social Survey, that liberal Protestant denominations such as the Presbyterians, Episcopalians, and the United Church of Christ did indeed decline among younger generations, and this was also true for many moderate Protestant groups like the United Methodists, Lutherans, and Disciples of Christ. By contrast, more conservative denominations like the Southern Baptist Convention and the Seventh-day Adventists appeared to grow in numbers.

In contrast to the claims of supply-side 'rational choice' theorists (e.g., Iannaccone, Finke, and Stark), however, the decreasing proportion of Americans identifying with liberal and moderate Protestant churches was not accompanied by a surge in affiliation with antimodernist sectarian groups. Indeed, the largest sectarian denomination, the Southern Baptist Convention, is in decline, and other sectarian groups are stable, not growing (Sherkat, 2014). The Catholic Church remains stable in spite of its liberalism, and the growing Latino Catholic population may further spur its embrace of liberal Christianity. While critics of liberalism enjoy pointing to the growth of the Mormons, they remain a small denomination, which now has very low rates of retention among the young. Further, the Unitarian Universalists are enjoying considerable growth from switching – higher than that found among Mormons (Sherkat, 2014). Indeed, the primary reason for higher levels of decline in liberal and moderate denominations is that their members have lower fertility rates (Sherkat, 2014).

It is also necessary to exercise caution when relying on data relating to levels of denominational attendance and affiliation alone in assessing the state of liberal Christianity in the churches. The main problem with this method is that it is blind to complex patterns of belief and commitment within churches and denominations. For example, most mainline churches

today – both Catholic and Protestant – contain both liberals and conservatives within their congregations, and it is increasingly hard to label whole denominations 'liberal' or 'conservative.' What is more, there is some important evidence that even more conservative forms of Catholic and evangelical Christianity are increasingly permeable to the influence of liberalism. Thus, James Davison Hunter's extensive research among college-age Evangelicals in American uncovered evidence of an increasing liberalization of their belief, values, and practice (Hunter, 1987).

Ammerman (1997) who studied 23 representative congregations across the United States and surveyed almost 2000 individuals discovered that respondents fell into three categories: liberal or 'Golden Rule' Christians (51%), evangelicals (29%), and social activists (19%). On the basis of her research, Ammerman challenges the assumption that religious liberalism is a spent force, and that liberal religiosity is a paler reflection of conservative Christianity. The Golden Rule Christianity she describes is characterized by an emphasis on the primacy of good deeds motivated by love, care, and compassion, and by a belief in the importance of religious tolerance.

Liberal Christians remain in positions of dominance throughout the West. The sectarian Protestantism amplified by Kelley (1972) and others had no influence on the character of Christianity in Europe, though it has profoundly affected African Christianity. The antimodernist impulse of sectarian groups hinders their members' ability to succeed in broader society, and liberal Christians continue to dominate educationally, in politics, and in income and wealth attainment (Keister and Sherkat, 2014). Indeed, even in the few examples of sectarian Protestants attaining positions of power, most, like US Presidents Jimmy Carter and Bill Clinton, are liberal Christians. Growing ethnic diversity also bodes well for liberal Christianity, since social justice orientations fostered by liberal Christianity are attractive to ethnic Christian minorities, particularly in the United States. Perhaps most importantly, the liberal Christianity has embraced individual liberty on issues of sexuality and gay and lesbian rights, in contrast to sectarian Christianity, and the increasing embrace of sexual freedom and civil rights for gays and lesbians makes these groups more attractive to younger generations (Sherkat, 2014).

Sociological Interpretations

At least three clusters of explanations have been offered by sociologists of religion to account for the rise of liberal Christianity in modern times. It has been explained as (1) an accommodation or even a capitulation to modernity (Peter Berger), (2) a natural outgrowth of Protestantism and, in particular, of the liberal Protestantism's emphasis on personal subjective conviction (Ernst Troeltsch), and (3) a means by which the clergy, their social status undermined by modernity, have attempted to protest and regain a social role (Jeffrey Hadden). Woodhead and Heelas (2000) have also drawn attention to liberal Christianity's compatibility with modern socioeconomic formations and wider cultural trends such as the turn to the self.

Sociologists have also developed theories to account for liberal Christianity's apparent decline. Kelley explained this by drawing a contrast with conservative religion. Where the conservative Christianity was 'strict' and 'challenging,' the liberal Christianity was the opposite. As such, he argued, it was unable to generate or sustain commitment, consensus, or strong community. Peter Berger offered a more rigorous version of this explanation by arguing that plausibility is a function of unanimity. The strong, unified communities that characterize conservative religion are better able to sustain plausibility than are the more diffuse and less disciplined communities of liberalism. Meanwhile, sociologists like Talcott Parsons and David Martin have also argued that liberalism is a victim of its own success: because its beliefs and values are so close to those of the wider culture, it is no longer able to sustain a distinctive identity or to hold or attract adherents.

Much of this theoretical work depends on a contrast drawn between liberal and conservative Christianity. It is also important to note another boundary: that between liberal Christianity and radical or 'alternative' forms of spirituality. For while the fate of liberal Christianity is bound up with that of conservative Christianity, it is also bound up with that of new forms of religiosity like the New Age. In some ways the latter seems to represent an intensification of key liberal themes like individualism and freedom, but without liberal Christianity's continuing commitment to some form of institutional church. It remains to be seen whether the apparent growth of such religiosity will in the end have the effect of strengthening or of weakening liberalism.

See also: Civil Religion; Liberalism, Anthropology of; Liberalism: Historical Aspects; Rationality in Society; Reformation and Confessionalization; Religion and Politics in the United States; Religion, Sociology of; Religion: Evolution and Development; Women's Religiosity.

Bibliography

Ammerman, N.T., 1997. Congregation and Community. Rutgers University Press, New Brunswick, NJ.
Hunter, J.D., 1987. Evangelicalism. The Coming Generation. University of Chicago Press, Chicago.
Hutchison, W.R. (Ed.), 1968. American Protestant Thought in the Liberal Era. University Press of America, Lanham, MD.
Hutchison, W.R., 1976. The Modernist Impulse in American Protestantism. Harvard University Press, Cambridge, MA.
Keister, L.A., Sherkat, D.E., 2014. Religion and Inequality in America. Cambridge University Press, Cambridge.
Kelley, D.M., 1972. Why Conservative Churches Are Growing. Harper & Row, New York.
Michaelsen, R.S., Roof, W.C. (Eds.), 1986. Liberal Protestantism: Realities and Possibilities. Pilgrim Press, New York.
Miller, D.E., 1981. The Case for Liberal Christianity, first ed. Harper & Row, San Francisco.
Reardon, B.M. (Ed.), 1968. Liberal Protestantism. Stanford University Press, Stanford, CA.
Roof, W.C., 1978. Community and Commitment. Religious Plausibility in a Liberal Protestant Church. Elsevier, New York.
Roof, W.C., McKinney, W., 1987. American Mainline Religion. Its Changing Shape and Future. Rutgers University Press, New Brunswick, NJ.
Sherkat, D.E., 2014. Changing Faith: The Dynamics and Consequences of Americans' Shifting Religious Identities. NYU Press, New York.
Woodhead, L.J.P., Heelas, P.L., 2000. Religion in Modern Times. Blackwell, Malden, MA.

Liberalism, Anthropology of

Naomi Schiller, Temple University, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Over the past 15 years, anthropologists have increasingly turned their attention to liberalism. A contemporary anthropology of liberalism concentrates on how liberalism shapes anthropological knowledge practices and how liberal norms are lived and inhabited. Ethnographic work analyzes how liberal values are encountered, appropriated, rejected, and reworked by people who claim liberalism as an autochthonous tradition and by those who have been historically cast as liberalism's 'others.'

Over the past 15 years, anthropologists have increasingly turned their attention to liberalism, a set of political, philosophical, and economic doctrines, policies, and practices with roots in seventeenth- and eighteenth-century Enlightenment thought. In growing recognition that liberalism persists as "the most dominant ethic of our age" (Mahmood, 2007: p.148), which "polices the boundaries of the speakable and the unspeakable today" (Boyer and Yurchak, 2010: 180), anthropologists examine how various forms of liberalism frame social science epistemology as well as how liberal norms shape the conditions of possibility for human sociality globally. Recent ethnographic work analyzes how liberal values are encountered, appropriated, rejected, and reworked by people who claim liberalism as an autochthonous tradition and by those who have been historically cast as liberalism's 'others.' With the profusion of scholarly attention to the contours and impacts of neoliberalism, it has become ever more necessary to understand liberalism's history and wide ranging contemporary force.

Defining Liberalism

Given the pervasive hegemony of liberal norms, forging a cultural analysis of liberalism is no small task. Liberalism is deeply intertwined with projects of modernity, colonialism, capitalism, and democracy, making the work of parsing liberalism's unique characteristics or referents difficult if not impossible. To conceptualize liberalism in anthropological terms requires viewing the tradition of liberalism as 'heteroglossic' meaning that it is "under constant negotiation, reformation and critique through the very vicissitudes of everyday life" (Coleman and Golub, 2008: p.258; Goodale, 2008). Liberalism is best defined, then, as an ever-changing project rather than a specific historical period, narrow political philosophy, or type of society.

For many students, a discussion of liberalism is plagued by a basic confusion about what liberal or liberalism connotes. In the contemporary period 'liberal' sometimes indexes a progressive political position in support of a social welfare state and other state intervention. Yet, in many contexts, 'liberal' signifies a conservative advocate of free markets. Conflicting beliefs recognized as liberal values can, in fact, be traced to the same body of work of scholars and philosophers who have been categorized retrospectively as the founders of liberalism, such as John Locke and Adam Smith.

Liberal thought is generally predicated on a belief in the unilineal progress of humankind, made possible through the robust application of reason. Liberalism consists of manifold traditions, some dedicated to individual liberty, others more concerned with social equality. Liberal ideals were first formulated in western Europe in opposition to feudal society with the aim of redefining the limits of political authority. Proponents opposed the arbitrary power of the King and the privileges granted by birth to the nobility. The development of liberal thought involved the formulation of a new notion of the individual. The sovereign, according to early liberal thinkers, needed to be subordinate to the liberty and rights of the individual.

According to early and more recent proponents, liberal governments must be powerful enough to secure individual liberty but limited enough that it cannot impose restraints on individuals. The philosophical basis on the liberal state is a doctrine of natural rights that all 'men' possess fundamental rights (life, liberty, security, and happiness) on which the state must not tread. Whether individual autonomy is the aim of liberal governance that needs to be produced through state intervention or a naturally occurring condition of humanity has long been a key fault line within liberal doctrine (Hindess, 1993).

A rights-based state, according to liberal doctrine, is meant to employ constitutional mechanisms to guard against the misuse of power, respect due process, and guarantee equal protection against the tyranny of the majority. Liberal government proposes to manage a balance between order and liberty by identifying domains that are purportedly 'outside' politics, such as the market and civil society. The state is obligated not to limit the autonomy of these supposedly naturally existing autonomous spheres. Contemporary proponents such as John Rawls (1993) argue that political liberalism is uniquely able to foster pluralism by creating conditions that can tolerate conflicting moral, religious, or philosophical beliefs while avoiding comprehensive ideas or unitary visions about 'the good.'

Rather than formulate a comprehensive analysis of liberal political reason and its genealogy, anthropology has sought to document and analyze the everyday effects, lived contradictions, impasses, and potentialities of liberal approaches to the individual, law, the state, reason, difference, property, and freedom (among other concepts). Ethnographic analyses of how people encounter liberal concepts illustrate that liberalism is far from

neutral in regards to notions of the good, imposing particular norms, morals, and values (Comaroff and Comaroff, 2004; Gledhill, 1997) as well as a belief in a transcendental rational subject whose freedom is linked to an individual will to act autonomously (Mahmood, 2004, 2007). A contemporary anthropology of liberalism concentrates first and foremost on how liberalism shapes anthropological knowledge practices and how liberal norms are lived and inhabited.

Precursors

Recent anthropological explorations of everyday experiences of liberalism have been driven by two key developments. Foucault's scrutiny of liberalism as a rationality of rule greatly influenced the direction of much anthropological scholarship. Foucault approached liberalism not as an ideology or a philosophy but instead as a way to craft governance and practice freedom and discipline (Foucault, 2009; Rose, 1996). Rather than accepting at face value claims that liberalism seeks to defend freedom and distribute power, building on Foucault and poststructuralist scholars, anthropologists turned to examine how liberal thought organizes, constitutes, and defines power and freedom. Moreover, beginning in the 1990s, anthropologists began to devote extensive attention to the impacts of the post-Cold War narrative that liberal capitalist democracy had proven itself as the only viable form of human organization and to the application of this world view in the form of neoliberal policy and ideology. The growing influence of neoliberal thought has created a pressing need to understand this most recent variant of liberalism within a broader framework.

Anthropology's engagement with liberalism, however, begins much earlier, as anthropology is itself a product of liberal enlightenment thought. While this article focuses on the anthropology of liberalism – the self-conscious engagement by anthropologists of the meaning and uptake of liberal logics in peoples lives – one must also consider, to paraphrase Joan Vincent (1990), the liberalism of anthropology. Like all social science disciplines, anthropology is both a creation and critic of liberalism in its different historical iterations: from the Enlightenment, colonial and anticolonial projects, 'scientific' racism and antiracist scholarship, struggles over the development of the welfare state in the United States and Western Europe, the Cold War, the rise of neoliberal ideology, and the recent 'War on Terror.' Early anthropology built on and developed ideas and commitments of western European enlightenment thinkers, including 'community,' 'civil society,' 'contract,' 'secularism,' 'cosmopolitan' and principles of equality, universal humanity, individual rights, progress, private property, and reason (Vincent, 2009). Early and mid-twentieth-century anthropology embraced certain liberal precepts and contributed to the distinction and, hence, the formation of different civilizational orders and peoples according to liberal metanarratives of progress and modernization (Ferguson, 1999; Asad, 1973).

While anthropologists were complicit in the exclusion of non-Europeans, women, and workers from nineteenth- and twentieth-century liberal and colonialist claims for who was considered a reasonable and equal subject capable of self-rule, anthropological critique and knowledge has also long been mobilized to counter such exclusion. Anthropology has struggled over the internal tensions and multivalent possibilities of liberal ideals, observable in Franz Boas' antiracist cultural relativism and debates of the late 1940s over the Universal Declaration of Human Rights, when Melville Herskovits condemned the claim that European derived values were universally applicable. American anthropology advocated respect for pluralism and tolerance of different cultures. Anthropology's relativist stance fortified a homogenous and bounded notion of culture, which further strengthened the conceptual divide between the purportedly normative rational individual of the West, and liberalism's 'others,' the allegedly politically inadequate subject of the non-West. Contemporary scholarship on human rights, multiculturalism, and cosmopolitanism reviewed below, continues to debate the questions of liberal relativism and universalism first raised by mid-twentieth-century anthropology.

Marxist and feminist anthropology, which gained traction and legitimacy in 1960s and 1970s, generally dismissed liberalism as a cover for capitalism. Influenced by Marx's skepticism for liberal or 'bourgeois' political economy, advocates of this turn questioned the concept of the atomized autonomous individual or culture group and called for more direct engagements of anthropologists in politics as well as acknowledgment of global interconnection within relationships of unequal power (Wolf, 1982).

More recently, anthropologists have rejected the Cold War's projection of liberalism and socialism as diametrically opposed, building on scholars who understand these movements as deeply intertwined utopian projects, both of which grow out of the Enlightenment, and have long been defined in relationship to each other (Buck-Morss, 2000). Concepts or analytics that developed under late socialism, such as forms of political parody, prove to be useful tools to denaturalize the logics of recent manifestations of liberalism or, what anthropologists have recently begin to call 'late' liberal politics (Boyer and Yurchak, 2010).

Liberalism's Diaspora, Moral–Ethical Aspirations, and Global Force

Contemporary anthropological scholarship explores how liberal ideas become embedded in disparate contexts and have been a resource for marginalized peoples to critique Western domination in the colonial and postcolonial period. Liberalism has been and continues to be a product of the confluence between the colonizer and the colonized. Under the duress of anticolonial struggles and social movement claims for inclusion and democratization, liberal ideals have shifted over time (Goodale, 2008; Mahmood, 2004; Povinelli, 2006). Anthropologists document the tension inherent in systems of liberal democracy that allows for the disempowered or oppressed to draw on liberal principles of equality and liberty in order to make claims on a system that has excluded them often on the basis of these same principles. Mark Goodale argues, for example, that rather than 'conscripts' of liberal modernity, as David Scott might argue (2004), marginalized peoples are 'alchemists' (2008: p. 170).

Yet, even as liberalism's ideals are pressured to embrace an expanding view of who 'counts,' liberal normative assumptions continue to exclude certain "projects and moral–ethical aspirations" (Mahmood, 2007: p. 149). Anthropologists uncover the moral sensibility that remains hidden in liberal commitments to universal recognition and rationality. Particular moral injunctions often deemed pretheoretical are embedded in liberal legal structures, making them far more imperceptible than in socialist contexts, where the connection between politics and moral positions have been explicitly cultivated and expressed by state actors (Roudakova, 2009). Povinelli (2002: p. 9) points out that liberalism is at an impasse when critical rational knowledge that is supposed to form the basis for reasonable decision making under liberalism does not align with a liberal moral sensibility.

A central concern for many anthropologists exploring liberalism's contemporary force is how, in the name of humanitarianism, liberal values are used to justify intervention in sovereign states that have been judged as forfeiting their right to self-determination because they have violated individual rights. The division between normative liberal subjects and inadequate or culturally repugnant 'others' continues to support projects of governmental and neoimperial intervention, most recently observable in the US-led 'War on Terror' (Abu-Lughod, 2002). This divide takes place using liberalism's key concepts as a measure of a 'Eurocentric index of accomplishment' (Comaroff and Comaroff, 1999: p. 17). Recent anthropological scholarship on liberalism, which I review below, focuses on exploring how ideas of individualism, freedom, human rights, state, civil society, freedom of expression, and multiculturalism are mobilized and experienced, and what they accomplish.

The Individual, Agency, and Freedom

A notion of the individual as a self-authorizing subject and the primary social unit is central to liberal thought. Liberal ideals of freedom and agency center on the individual. Anthropologists subject this concept of the individual to immanent critique, asserting that the autonomous individual subject is a normative historical construct and cultural artifact.

For decades, ethnographic work has focused on the development or local meanings of personhood. Marilyn Strathern's work in Melanesia, for example, refuted the universal applicability of the idea of the individual finding that Melanesians construed themselves not as authors of their internal selves, but as multiply constituted (1988: p. 119). More recent work has turned from refuting the universalism of the liberal category of the individual to examining how liberal subjectivity – a sense of individual self-authorship – is constituted in everyday life. In Lila Abu-Lughod's ethnography of television viewing in Egypt, she illustrates how state-produced televised melodramas encourage the constitution of the modern liberal subject, one that views the self as autonomous and self-activating (Abu-Lughod, 2005).

Feminist anthropologists have long noted that proponents of liberalism posit the liberal subject as universal, neutral, and abstract when, in fact, the individual autonomous individual is presumed historically to be a male head of household responsible for dependent women and children. Recent appraisals of feminist anthropology take previous scholarship to task, however, for reproducing liberal assumptions about individual agency. Saba Mahmood finds that poststructuralist feminist scholarship produced by anthropologists and others continues to insist that human beings have an innate individual will for freedom that is expressed when people assert their agency by challenging social norms (Mahmood, 2004: p. 5). In Mahmood's work on the women's piety movement in Egypt, she argues instead that "agentival capacity is entailed not only in those acts that resist norms but also in the multiple ways in which one inhabit norms" (Mahmood, 2004: p. 15).

Building on and reappraising poststructuralist critiques, a growing body of ethnographic work alongside Mahmood's challenges a liberal definition of freedom as being opposed to or liberated from norms or power. Anthropologists reveal how a liberal concept of freedom is mobilized as "a means of governance" (Englund, 2006: p. 9) and as part of a strategy to rule through the "regulated and accountable choices of autonomous agents" rather than through social institutions or the state (Rose, 1996: 160). Anthropologists provide empirical evidence of how liberal governance produces purportedly 'free' individuals through institutions such as schools, media outlets, human rights organizations, development initiatives, courts, and other sites to create self-regulating subjects. Liberal freedom, anthropologists and others argue, is particularly a mode of exercising power over ourselves and others (Rose, 1999).

Liberals lean heavily on the binary concepts of individual freedom and social constraint to differentiate civilizational orders and peoples, and to govern love, access to rights, sociality, and the body (Povinelli, 2006). Elizabeth Povinelli, a prolific anthropologist of liberalism, asks:

> Why does it matter, and to whom, that certain communities and their economic, ritual, and social practices be apprehended as 'free' and others as 'constrained'? What work does this dichotomy do in defining the social and political exceptionality of liberal forms of constitutional democracies? How does this dichotomy allow democratic institutions to exercise disciplinary force as if they were advancing freedom and protecting the good? (2006: p. 99)

In Povinelli's analysis of disparate groups in the liberal settler societies of Australia and the United States she goes to great lengths to answer these questions. She reveals how indigenous Australians are deemed trapped by their 'genealogical inheritance' or traditions, while queer radical faeries in the United States are caught within the same liberal logic but face different conditions of recognition of their rights given their purported individual freedom, agency, and lack of tradition. Being understood as 'free' or 'constrained' shapes the possibility of gaining recognition and rights. In a different context, Amahl Bishara uncovers how Israeli state forces deprive Palestinian journalists of rights by regarding them as hopelessly constrained or biased by their customs and history and thus lacking the ability for rational deliberation (Bishara, 2013). The liberal binary of freedom and constraint, anthropologists reveal, is crucial to legitimizing the liberal state's selective distribution of rights and resources at local and national levels and shapes neoimperial politics on a global scale.

In ethnographic inquiries into the lived experience of liberal ideals of individual freedom in the United States, anthropologists have examined how computer hackers negotiate and critique

liberal commitments to individual autonomy and property (Coleman and Golub, 2008). Hackers and free and open software developers draw on liberal ideals of self-reliance and individual liberty even while free software seems to violate the logic and principles of private ownership and individual autonomy (Coleman and Golub, 2008; Kelty, 2008). Free and open software proponents connect individual autonomy to the use of property, but as Coleman and Golub point out, "the use and enjoyment is enabled through sharing," not individual property rights (2008: p. 268). They create a "rival liberal morality" and translate it into a legitimate legal register (Coleman, 2009: p. 448). This ethnographic work uncovers the processes through which traditional liberal ideals persist and are transformed.

Civil Society and the State

Over the past few decades, anthropologists have exposed the way that liberal conceptions of 'state' and 'civil society' have become part of common sense ways of viewing political practice and space, both in mainstream accounts and in much social science literature (Ferguson, 2006). These concepts have multiple histories. The concept of civil society can be traced, for example, to eighteenth-century Scottish Enlightenment thinkers, and was subsequently reworked by a series of scholars from Hegel to Marx to Gramsci. The more recent appropriation and popularity of 'civil society' in the 1980s and 1990s was due to a confluence of forces: shifting world politics at the end of the Cold War, the conjunction of progressive grassroots organizing and neoliberal efforts, and the political economy of translation and academic publishing.

In most variants of liberalism, state and society are ideally separate and autonomous spheres. The state is seen as a set of external constraints on individual freedom, necessary for an ordered social existence but a threat to liberty. In order to study ethnographically how the state is understood and represented, anthropologists examine banal contact with bureaucracies and state institutions, the science and art of governance, and fantasies about the state. Akhil Gupta's groundbreaking ethnography (1995) uncovered how the boundaries between state and society are, in fact, blurred in everyday practice. Rather than a monolith set apart from everyday interactions and situated social relations, the state is an ensemble of local practices that include what are often considered civil society institutions. Commonplace bureaucratic practices as well as spectacular displays of power help produce an understanding of states as spatially above society and encompassing the local (Ferguson and Gupta, 2002; Coronil, 1997). Ethnographers assert that the supposed separation between state and civil society, so central to liberal thought, is produced by the way that governance is enacted and represented. The anthropology of liberalism thus challenges the actual existence of autonomous spheres of civil society and the state.

Liberal Law, Human Rights, Universalism, and Vernacularization

Human rights law (a branch of liberal law) and the possibility of its local vernacularization has been an enduring preoccupation of anthropologists since the 1940s. Ethnographic attention to human rights practices, organizations, and discourse illustrates how human rights discourse works as a "liberal vanguard of the contemporary imaginary" (Goodale, 2008: p. 127). Anthropologists have observed that politicians, scholars, and activists increasingly turned to a language of human rights to replace the grand narrative of liberal democratic freedom versus communist tyranny, which animated the Cold War in the west.

Anthropologists have 'provincialized' human rights discourse, highlighting its roots in European philosophy that promotes a particular individualized framework for social justice dependent on liberal notions of autonomy, equality, choice, and secularism (Merry, 2006; Wilson, 1997). Human rights discourse demands a sense of selfhood autonomous from family, religion, and community. This self is imagined to be protected by the liberal state, the central arbiter of justice, rather than community or kin. The ensemble of international nongovernmental and governmental agencies and organizations dedicated to advocating for universal human rights often seek to transform subjectivities to conform with modern liberal notions of the individual rational subject in ways reminiscent of colonial states.

Anthropology explores the ongoing internal contradiction within liberalism that simultaneously facilitates universalist and relativist claims for human rights, both of which can put to progressive or conservative ends and rely on essentialist notions of humanity or culture. As Tejani notes, "Relativism may respect essential difference, but first it must enumerate the essence of difference. Likewise, universalism may argue toward the essential commonality of human subjects, but first it must locate the essence of that commonality. In either case, it is clear, essentialism is a necessary tool for the relativist and universalist renderings of cultural reality" (Tejani, 2004: p. 335). Cultural relativism, as Abu-Lughod notes (2002), is an improvement on ethnocentrism and racism, but often overlooks the fact that the forms of life we see around the world are not pristine examples of essential human ways of being but rather the products of a history of intervention and interaction. A critical cultural relativism acknowledges power differentials within and across groups and that it may be 'too late not to interfere' in a given context when we acknowledge a history of intervention (Abu-Lughod, 2002: p. 786).

While anthropologists reject the universal liberal subjectivity and normative morality at the heart of human rights discourse, many continue to advocate for the rights and needs of marginalized populations. Yet the conceptual ground on which to fight for rights and resources is tricky terrain. While human rights discourse can be used to impose an untenable universalism or can reify essentialisms, many anthropologists search for a recognition of commonalities cross-culturally that can be used as a basis to fight for social justice unwedded from liberalism's secular humanism and essentialisms. Some scholars have begun to forge a framework of 'critical cosmopolitanism' as a normative basis of shared human commensurability not dependent on liberal norms or metanarratives (Werbner, 2008; Glick Schiller et al., 2011).

Acknowledging the far-reaching spread of human rights discourse across the globe, anthropologists have turned their attention to analyzing human rights discourse in action. Liberal human rights ideals are incorporated into domestic law, people

are encouraged to complain about problems in human rights terms, judiciaries are educated about international treaties and standards, and governments are monitored and pressured to be accountable. Rather than assume that concepts such as 'women's rights' have a knowable essential meaning or actual ontology, anthropologists trace the uptake and movement of these ideas, exploring what they are used to accomplish (Abu-Lughod, 2010). Anthropologists ask, how are liberal logics of rights locally appropriated, rejected, or hybridized? What are the outcomes of efforts at vernacularization?

Anthropologists contend that the contemporary human rights system is far from a coherent or unified approach, but is rather a disparate collection of mechanisms, policies, and declarations appropriated by subaltern groups, not simply imposed on them (Lazarus-Black, 2006). Yet, given how liberalism is highly inscribed in the language of law, access to legal literacy proves fundamental to challenge and transform liberal regimes (Coleman, 2009; Comaroff and Comaroff, 2004). Skills for technological and legal understanding are often quite compatible, making well-educated highly literate people such as computer hackers and free software writers – who Kelty (2008: p. 19) calls "superalterns" – especially equipped to challenge existing approaches to liberal law.

Ethnographic inquiries into the possibility that people can vernacularize or hybridize human rights law underscores the actual limits of liberal claims to openness and tolerance. Human rights professionals, anthropologists find, often position tradition or religion as an impediment to the cultivation of reason and equality necessary to secure universal rights. Even when local human rights advocates attempt to frame human rights issues according to local histories, the framing process is often restricted by commitments to liberal notions of justice, individual rights, autonomy, choice, and equality. Anthropologists have documented, moreover, how in the face of criticism of what human rights practitioners understand as local practices, groups often resort to affirming an essentialist cultural nationalism or religious fundamentalism.

Anthropologists have also recently explored what happens when people seek to legitimize their vocation or social movement within a liberal imaginary of human rights discourse even while their practices and experiences are structured by different understandings of politics and justice (Hasty, 2005; Roudakova, 2009; Schiller, 2011). Goodale, for example, finds that when marginalized Bolivians turn to a liberal human rights discourse grounded in rights to private property and contract law they ultimately restrict their ability to advance communal notions of equality and access to resources (2008: p. 152). Jennifer Hasty uncovers how Ghanaian journalists turn to essentialized notions of cultural difference in order to explain why their practice falls short of Western ideals of autonomy.

Despite trepidations concerning the often negative outcomes of attempts at vernacularization of human rights discourse, many anthropologists remain committed to a global liberal legal framework for its potential to secure rights and reparations.

Liberal Multiculturalism

Building on anthropological insight about human rights discourse, universalism, and relativism, scholars have recently explored the rise of liberal multicularism. Anthropologists note that states have created official policies of cultural recognition as a way to contain anticolonial and new social movements' demands for social justice. Multicultural policies and ideologies seek to sidestep the problems that capitalism and (post)colonialisms raise for defining national identities (Povinelli, 2011; Comaroff and Comaroff, 2003; Povinelli, 2002). These policies aim to valorize and accept cultural differences and identities without destabilizing the framework of liberal justice, which does not address socioeconomic inequality. Anthropologists explore the limits of the liberal multicultural embrace of pluralism and tolerance in the face of radical alterity, while pointing to the ways that multicultural discourse can also reinforce concepts of essentialized cultural and religious difference.

Liberal multiculturalism, anthropologists argue, both creates and contains difference. Difference, which is seen as an essence rather than a product of social relations, is marked and granted legal protection only if it does not seriously counter liberal norms for what is morally acceptable and legally tenable. Recognition of indigenous customary law in Australia, as Povinelli notes, is punctuated with an invisible disclaimer "provided [they] … are not so repugnant" (2002: p. 12). Minority subjects are required to provide evidence of enough difference so that liberal governments can demonstrate their tolerance and respect of multiplicity, but without asserting values that actually shake liberal legal or moral foundations.

Liberal multiculturalism often works as a form of domination by encouraging minority subjects to assert an essential authentic self. Cultural 'others' must demonstrate or perform a certain degree of alterity to merit recognition or entitlements. Nevertheless, indigenous subjects find ways to creatively engage liberal discourses of multiculturalism to fulfill needs. Liberal multiculturalism's stifling of local complexities, in other words, is not the whole story. John and Jean Comaroff maintain that in the 'policultural' postcolony of South Africa "a 'living constitution,' tolerant of everyday ambiguity is being forged in the space of strategic engagement opened up by the alternative languages and cultures of legality" (Comaroff and Comaroff, 2003: p. 245). While the formal prescriptions of the liberal state restrict the acceptance of multiple epistemologies or authorities, in practice a vernacular praxis embraces the ambiguous and multiple grounds of morality and reason.

Challenges to liberal claims of tolerance take place not only in postcolonial nations, but in former colonial powers. France's attempts to ban religious symbols from public institutions, for example, present anthropologists rich terrain to explore the malleability of liberal pluralism. Tejani (2004) explores how migrants in the European Union attempt to 'repatriate' liberal ideals of tolerance in ways that challenge historic parameters for national identity and sovereignty.

Understanding the *Neo* of Neoliberalism, the *Late* of Late Liberalism

Neoliberalism in not simply a revival of classical liberal political economy that advocates for a small state and a free market, but

represents instead the imposition of values of privatization, market freedom, and atomized individual responsibility to *all* spheres of human life. As Kingfisher and Maskovsky note (2008), many anthropologists approach neoliberalism through a materialist lens that highlights new concentration of class power and capitalist reorganization since the 1970s (Harvey, 2005), some embrace a poststructuralist perspective that emphasizes particular forms of governance and subjectification (Rose, 1996), while others highlight the specificities and contingencies that neoliberal ideas allow for citizenship (Ong, 2006).

Given the fervent critical attention devoted to the rise of neoliberalism, its imposition and the backlash against it, anthropologists and others have begun to insist on the importance of situating neoliberalism with the *longue durée* of liberal ideals and practices. Grasping the philosophical and political underpinnings of neoliberal ideas is important to overcome overly conceptual and totalizing approaches that endow neoliberalism with an agency all its own (Kingfisher and Maskovsky, 2008). Situating neoliberalism allows scholars to draw on the history of debates about liberalism's internal inconsistencies, potentialities, and lived impacts in order to better evaluate neoliberalisms contours and identify with specificity how people reshape social worlds using neoliberal policies and ideas.

Neoliberalism is one aspect of what anthropologists have begun to term 'late liberalism,' but does not represent the totality of how liberal norms are asserted and inhabited in the contemporary period. Recent ethnographic research highlights the characteristics of late liberalism, which include the effects of anti- and postcolonial and new social movements on how difference is treated in liberal states (Povinelli, 2011: p.103), the substitution of class or party affiliation with essentialized identities that champion culture rather than politics (Comaroff, 2011), as well as the 'hypernormalization' of authoritative political discourse that leads to modes of irony and overidentification (Boyer and Yurchak, 2010). 'Late' liberalism is not meant to mark a time period, but to signal the recent changes to liberal forms of governance and contemporary conditions of capitalism.

The recent concerns of anthropology of liberalism reviewed above fruitfully intersect with new directions in related fields of geography, sociology, political science, and literature.

See also: Colonialism, Anthropology of; Ethnocentrism; Human Rights, Anthropology of; Liberalism: Impact on Social Science.

Bibliography

Abu-Lughod, Lila, 2002. Do Muslim women really need saving? American Anthropologist 104, 783–790.
Abu-Lughod, Lila, 2005. Dramas of Nationhood. University of Chicago Press, Chicago.
Abu-Lughod, Lila, 2010. The active social life of "Muslim women's rights." Journal of Middle East Women's Studies 6, 1–45.
Asad, Talal, 1973. Anthropology and the Colonial Encounter. Ithaca Press, London.
Bishara, Amahl, 2013. Back Stories: U.S. News Production and Palestinian Politics. Stanford University Press, Stanford.
Boyer, Dominic, Yurchak, Alexei, 2010. American stiob: or, what late-socialist aesthetics of parody reveal about contemporary political culture in the west. Cultural Anthropology 25, 179–221.
Buck-Morss, Susan, 2000. Dreamworld and Catastrophy. MIT Press, Boston.
Coleman, Gabriella, 2009. Code is speech: legal tinkering, expertise, and protest among free and open software developers. Cultural Anthropology 24, 420–454.
Coleman, Gabriella, Golub, Alex, 2008. Hacker practice: moral genres and the cultural articulation of liberalism. Anthropological Theory 8, 255–277.
Comaroff, Jean, 2011. Populism and late liberalism a special affinity? The Annals of the American Academy of Political and Social Science 637, 99–111.
Comaroff, Jean, Comaroff, John. L., 2003. Reflections on liberalism, policulturalism, and ID-ology. Social Identities 9, 445–473.
Comaroff, John L., Comaroff, Jean, 1999. Introduction. In: Comaroff, John L., Comaroff, Jean (Eds.), Civil Society and the Political Imagination in Africa. The University of Chicago Press, Chicago, pp. 1–43.
Comaroff, John L., Comaroff, Jean, 2004. Criminal justice, cultural justice. American Ethnologist 31, 188–204.
Coronil, Fernando, 1997. The Magical State. University of Chicago Press, Chicago.
Englund, Harri, 2006. Prisoners of Freedom. University of California Press, Berkeley.
Ferguson, James, 1999. Expectations of Modernity. University of California Press, Berkeley.
Ferguson, James, 2006. Global Shadows: Africa in the Neoliberal World Order. Duke University Press, Durham.
Ferguson, James, Gupta, Akhil, 2002. Spatializing states: toward an ethnography of neoliberal governmentality. American Ethnologist 29, 981–1002.
Foucault, Michel, 2009. Security, Territory, Population: Lectures at the College de France 1977–1978. Picador, New York.
Gledhill, John, 1997. Liberalism, socio-economic rights and the politics of identity. In: Wilson, Richard (Ed.), Human Rights, Culture and Context. Pluto Press, London, pp. 70–110.
Glick Schiller, Nina, Darieva, Tsypylma, Gruner-Domic, Sandra, 2011. Defining cosmopolitan sociability in a transnational age. Ethnic and Racial Studies 34, 399–418.
Goodale, Mark, 2008. Dilemmas of Modernity: Bolivian Encounters with Law and Liberalism. Stanford University Press, Stanford.
Gupta, Akhil, 1995. Blurred boundaries: the discourse of corruption, the culture of politics, and the imagined state. American Ethnologist 22, 375–402.
Harvey, David, 2005. A Brief History of Neoliberalism. Oxford University Press, Oxford.
Hasty, Jennifer, 2005. The Press and Political Culture in Ghana. Indiana University Press, Indiana.
Hindess, Barry, 1993. Liberalism, socialism and democracy. In: Barry, Andrew, Osborne, Thomas, Rose, Nikolas (Eds.), Foucault and Political Reason. University of Chicago Press, Chicago, pp. 65–80.
Kelty, Chris, 2008. Two Bits: The Cultural Significance of Free Software. Duke University Press, Durham.
Kingfisher, Catherine, Maskovsky, Jeffrey, 2008. Introduction. Critique of Anthropology 28, 115–126.
Lazarus-Black, Mindie, 2006. The politics of place: practice, process, and kinship in domestic violence courts. Human Organization 65, 140–155.
Mahmood, Saba, 2004. Politics of Piety: The Islamic Revival and the Feminist Subject. Princeton University Press, Princeton.
Mahmood, Saba, 2007. Ten. In: Shaikh, Nermeen (Ed.), The Present as History. Columbia University Press, New York, pp. 148–171.
Merry, Sally, 2006. Human Rights and Gender Violence. University of Chicago Press, Chicago.
Ong, Aihwa, 2006. Neoliberalism as Exception. Duke University Press, Durham.
Povinelli, Elizabeth A., 2002. The Cunning of Recognition. Duke University Press, Durham.
Povinelli, Elizabeth A., 2006. The Empire of Love. Duke University Press, Durham.
Povinelli, Elizabeth A., 2011. Economies of Abandonment. Duke University Press, Durham.
Rawls, John, 1993. Political Liberalism. Columbia University Press, New York.
Rose, Nikolas, 1996. Governing "advanced" liberal democracies. In: Barry, Andrew, Osborne, Thomas, Rose, Nikolas (Eds.), Foucault and Political Reason. University of Chicago Press, Chicago, pp. 37–64.
Rose, Nikolas, 1999. Powers of Freedom. Cambridge University Press, Cambridge.
Roudakova, Natalia, 2009. Journalism as "prostitution". Political Communication 26, 412–429.
Schiller, Naomi, 2011. Liberal and Bolivarian regimes of truth. Transforming Anthropology 19, 35–42.
Scott, David, 2004. Conscripts of Modernity. Duke University Press, Durham.

Strathern, Marilyn, 1988. The Gender of the Gift. University of California Press, Berkeley.
Tejani, Riaz, 2004. Liberalism repatriated. Anthropological Quarterly 77, 331–338.
Vincent, Joan, 1990. Anthropology and Politics. University of Arizona Press, Tucson.
Vincent, Joah (Ed.), 2009. The Anthropology of Politics. Blackwell Publishers, Massachusetts.
Werbner, Pnina, 2008. Anthropology and the New Cosmopolitanism. Berg Publishers, Oxford.
Wilson, Richard, 1997. Human rights, culture and context: an introduction. In: Wilson, Richard (Ed.), Human Rights, Culture and Context. Pluto, London, pp. 1–27.
Wolf, Eric, 1982. Europe and the People without History. California University Press, Berkeley.

Liberalism: Historical Aspects

Dieter Langewiesche, Universitat Tubingen, Tubingen, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8792–8797, © 2001, Elsevier Ltd.

Abstract

Liberalism is a sociopolitical idea about political order. It aims at creating a civil society, but adjusts its concrete demands to the existing situation of society and the state. This readiness to adjust has made liberalism willing to compromise as well as open to development. For this reason, the history of liberalism as an idea and as political practice can be adequately understood only as a continuing struggle about interpretation, but as controversial as the definition of liberalism has always been, liberalism has always aimed to make responsible citizenship possible. This is true for all societies, even if the paths liberalism has taken have differed in Europe, the USA, and other parts of the world. Europe is where liberalism originated and formed its liberal ideas of political order. For this reason, the article will concentrate on European liberalism. It will stress four aspects: (1) liberalism as an idea of political order; (2) liberal ideas of social and economic order; (3) liberalism and the nation; (4) liberal organization and milieus.

Liberalism is a sociopolitical idea about political order that has contributed substantially, since the late eighteenth century, to the modern age. It aims at creating a civil society, but adjusts its concrete demands to the existing situation of society and the state. This readiness to adjust has made liberalism willing to compromise as well as open to development. The core principles of liberalism endure, but the details of what it understands as civil society have changed in the course of history and have always been debated among liberals. For this reason, the history of liberalism as an idea and as political practice can be adequately understood only as a continuing struggle about interpretation, but as controversial as the definition of liberalism has always been, liberalism has always aimed to make responsible citizenship possible. This is true for all societies, even if the paths liberalism has taken have differed in Europe, the USA, and other parts of the world. Europe is where liberalism originated and formed its liberal ideas of political order. For this reason, we will concentrate on European liberalism.

Since its beginnings, liberalism has demanded three kinds of citizenship rights: (1) equality before the law, guaranteed by the rule of law; (2) equal opportunities for political participation, made possible by the right to vote and to participate in public life, and (3) the provision of the basic prerequisites for opportunities in life. The demand for equality before the law was made and fulfilled earliest. Then came political equality, though in practice liberals, in contrast to democrats, long favored giving different weights to the votes of members of different classes. Whether and to what extent the state should guarantee social rights has remained controversial among liberals to this day.

Liberalism as an Idea of Political Order

Liberal thinking centers on the individual, whose freedom of decision is to be defended against state and social constraints. All demands for individual freedom can thus be considered part of the liberal tradition. This would make liberalism as old as human memory, but a liberal doctrine of the state and society first arose in seventeenth century England. Its most influential thinker was John Locke (*Two Treatises of Government*, 1690). The European Enlightenment's attempts to limit Absolutism and to enable 'man's exit from self caused lack of self-responsibility' (Immanuel Kant) and finally the impetus of the American and French Revolutions (human rights, popular sovereignty, constitutionality) are further important stages of liberal thinking.

As political terms, 'liberalism' and 'liberal' were coined in Spain and did not establish themselves until after 1800. The nineteenth century was then the century of liberalism. No other sociopolitical idea of political order changed the state and society as radically. With the idea of the citizen, liberalism sketched a program for the future that was egalitarian in principle and that was directed against everything that limited individual freedom. At that time, this revolutionary program, which liberals nevertheless wanted to realize evolutionarily in step-by-step reforms, was directed primarily against absolutism and the rule of the nobility, and thus also against class and denominational privileges.

The liberals' central political demand was for a constitutional state. It was to guarantee the continuity of law, secure individual property, and enable the citizen to participate in the state. The liberals saw the absolutist state as their antagonist and wanted to tame it with constitutional law. The individual was to have free access to the public political arena and, as voters, exert influence on the state's decision making process, but the state was not to be stripped of its executive powers, only its functionings to be liberalized. The liberals did not call for a weak 'night watchman' state, but for freedom in and through the state. Liberalism's specific political program varied substantially from country to country and over time. What critics condemned as programmatic weakness or even as a betrayal of principles was precisely the strength of the liberal movement. Thanks to its flexibility, it was able to adjust its idea of political order to very disparate conditions of action. Liberals argued everywhere for the parliamentary form of government, but they also accepted compromises to avoid endangering the stability of state and social order.

Among the large European countries, the liberals' will to establish the parliamentarian system of government was

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62089-5

weakest in autocratic Russia and in preWorld War I, Germany. Liberals did not insist upon a particular form of state, but their ideal was parliamentary monarchy; they wanted to preserve the monarchy as a reserve constitution for states of emergency. This sharply distinguished them from the democrats, whose ideal was the republic. The European liberals of the nineteenth century were willing to accept the republican form of state (first in France and Italy), but, since the French Revolution of 1789, most liberals saw the republic as burdened with the odium of social revolution and the annihilation of civil society. The liberals' fear of social revolution put them at a distance from the idea of the republic, which could prevail only through revolution wherever monarchical states existed. In republican North America, this was not an issue. Limiting the power of the state and securing private property were the foci of American property-oriented, individualistic liberalism, which emphasized self-help and self-organization. Franklin D. Roosevelt's New Deal liberalism created the USA's first instruments for state intervention in society and economic policy.

In practice, the liberal idea of political order focusing on a society with egalitarian citizenship was compatible with inequality. The liberals of the nineteenth century wanted immediate implementation only of equality before the law; political equality could wait. They saw freedom of opinion and the right to vote as the most important means of apportioning participatory rights in the state and the community. By the early twentieth century, liberals had accepted democratic equal voting rights, but they never advocated or demanded the establishment of such rights. Rather, the majority of liberals accepted them only when they could no longer be avoided. The core of their idea of civil society was the economically self-sufficient, educated individual.

Liberalism offered an educational program for those who had not yet achieved this social status; the individual himself was to be responsible for its future fulfillment. This program was directed against a society that apportioned different political rights in accordance with membership in specific groups; but in the century of liberalism, the struggle against the historical and traditional society of privileges did not yet aim at the 'mass democracy' of the twentieth century. When the latter appeared on the horizon in the second half of the nineteenth century, liberalism had to change fundamentally if it was to survive party-political competition for votes. European liberalism was by and large able to cope with this adaptation until the beginning of World War I.

Emancipation to become a citizen with full rights not only developed at different paces for different classes, but also for the two sexes. Only a few nineteenth-century liberals could imagine woman as a citizen; one of these was John Stuart Mill (1859) *On Liberty and Other Essays*, an exceptional liberal figure precisely in the radicality of his emancipatory demand for gender parity. Not until the twentieth century did the majority of average liberals come to accept and recognize woman as having equal political rights. Until then, they used the idea of a naturally given eternal inequality of the sexes to avoid the consequences of their egalitarian model of law. Their constitutional idea of civil society was egalitarian, because it was based on the principle of contractual freedom for all. But they did not want to see this principle applied to the family.

Their refusal to politically equate man and woman turned out to have great consequences for their model of society, because they saw the family as the foundation of the state and civil society. In the long run, the liberals could not suppress the contradiction between their political model of an egalitarian civil society and their attempt to construe the family as fundamentally gender-hierarchical. The egalitarian philosophy of liberalism forged an intellectual weapon against the practice of liberal politics, which tolerated or even affirmed certain forms of political inequality. The golden age of liberalism ended with World War I, although its chances for development seemed more favorable than ever before. After all, the war ended in a victory for the Western democracies (USA, Great Britain, and France), which were regarded as the primary representatives of liberalism, while the monarchies of Central Europe and Russia did not survive the war.

Political liberalism did not sweep the field after the fall of these anti-liberal bulwarks. Instead, the signature of the new epoch was rivalry and finally struggle among the liberal-democratic and the new Bolshevist and fascist ideologies, movements, and states. The National Socialist dictatorship, in whose parlance 'liberalistic' was perverted into a life-endangering, denunciatory pejorative, finally threatened to extinguish forever the central idea of liberalism, the dignity of the individual, first in Germany and then in Europe's conquered states. The disintegration of liberal parties, including in Great Britain, the mother country of liberalism, continued after World War I and even after World War II, but at the same time, the constitutional orders and individual behavioral norms in the Western world have been characterized by fundamental liberal values more than ever before. Their validity is no longer tied to the strength of liberal parties. Liberalism's power to form parties appears to decline when its political ideas prevail and are recognized even by opposing political parties.

The collapse of the Soviet empire seems to confirm this observation. The successor states adopted liberal principles, or what they took to be liberal principles, in their new political and economic orders, but this triumph of liberalism over its communist antagonist has at the same time triggered a new international debate over the central principles of a liberal system of state and society (Ackermann, 1992). This debate focuses on the old problem of whether a liberal society can trust solely in the regulative miracle of the market or if it should strive to create a state that, through a 'combination of citizenship rights and chances for welfare' (p. 196), empowers all people 'to take part as citizens in the life of the society' and 'to enjoy the achievements of their time' (Dahrendorf, 1987, pp. 196, 219).

Ideas of Social and Economic Order

Historical liberalism arose in the pre-industrial age. It strove for a civil, rather than bourgeois society, in both its ideas of social and economic order as well as in its model of the state. With the idea of the citizen, European liberalism sketched a vision of the future of a society of citizens without great property-ownership differences that would have contradicted the liberal ideal of a middle-class society with a broad distribution of assets (Schapiro, 1958). This socially harmonious guiding idea was

directed against the class/feudal order, but did not initially field an industrial view of the world against it. Liberalism comprised a plurality of social ideas that, depending on a country's state of social development, ranged from agrarian models of the future (as in Hungary, Spain, and Italy) to industrial capitalistic models (as in the USA, Great Britain, Belgium, and, in part, France and Germany). The liberals demanded social and economic 'progress,' but not an industrial-capitalistic class society. Only after the mid-nineteenth century did pre-industrial liberal ideas of order pale in favor of industrial capitalistic ones.

This shift, pushed forward by the wave of revolutions of 1848, marked the transition from early liberalism to liberalism in the age of industrialization. Only now, when dynamic industrialization ended the pauperism of the first half of the century and sociopolitical control mechanisms seemed renouncable, did political liberalism fuse with *laissez-faire* economics. The main current of liberalism now became more bourgeois than ever before, without narrowing to a bourgeois economic liberalism, even in this phase of Manchester Liberalism. Even now, the liberals were not characterized by convinced affirmation of a free, unregulated market. The liberal demand for free trade among states did not rule out a willingness to accept and call for regulation of the market within the state.

Since the last third of the nineteenth century, the trend to eliminate social-welfare aspects of ideas of societal order in favor of *laissez-faire* liberalism was tempered by a renaissance of social-liberal guiding ideas. Like early liberalism, the New Liberalism, an international phenomenon, displayed a marked social responsibility, but it assumed the conditions of an industrial society in which the freedom of the individual would be ensured by a policy of social welfare. Liberals had originally contributed to the political climate that, since the 1880s, gave birth to social security laws, first in Germany and then in rapid succession in other countries. This launched a development that led to the modern type of social welfare state. In the United States, it was not until Roosevelt's New Deal that political theory allowed for the state's substantial involvement, in contrast to the traditional liberalism of interest groups (Lowi, 1979).

No less important was community liberalism, which played a key role in the building of infrastructure in the enormously expanding cities, a precondition of urban life. This community welfare policy ranged from improved hygiene through municipal provision of gas and electricity to the new parks, museums, theaters, and mass education facilities. The achievements of community liberalism show how intensely liberals participated in the communal protection of individual life chances when, in the face of social and economic developments, individual provisions were no longer adequate or possible for increasing segments of the population.

But the individual, not the collective, was still the focus of liberal norms and politics. The conviction that the individual must take priority over the collective distinguishes even the decidedly social liberalism of the late nineteenth and twentieth centuries from the other progressive movements, such as socialism. But the dwindling numbers of liberal voters since the early twentieth century, teaches us that the variant offered by the liberal parties as a mediation between collective welfare security and individual freedom to decide no longer had the persuasive power that had characterized the liberal guiding ideas of the nineteenth century.

Liberalism and the Nation

Beginning in the late eighteenth century, the demands for popular sovereignty and national self-determination became two sides of the same coin. Only states that were recognized as national states could prevail. Liberals took a leading role in making the nation the new supreme principle of legitimacy. Like liberalism, nationalism developed a model opposed to the existing order. Both demands aimed at a society ordered in accordance with egalitarian values, constituted as a state and employing a collective, and thus equally egalitarian concept of sovereignty. This was why nationalism historically arose as, and can still present itself as an ideology of liberation and a movement of emancipation, despite all nationalistic atrocities. Anyone who wanted to ally with nationalism had to accept this basic egalitarian stance. Liberals did so. In the nineteenth century, they became the spokesmen of national movements, formulated their goals, and provided their elites.

This symbiosis of liberalism and nationalism strengthened and shaped both. Liberalism's influence expanded far beyond its narrower primary bourgeois clientele, and nationalism developed a reform program promising more social as well as political democracy. The vision of a national civil society that expands citizenship rights into the social realm and promises to allow all to share in the resources produced by a society were visions of the future common to liberalism and nationalism. These promises of participation for all, regardless of social status, are what made these movements so unrivaled in attractiveness. Liberalism was one of the creators of this vision that, here too, uncoupled itself from liberalism. Socialists and conservatives both adopted it.

The task of mediating between the central liberal principle of the responsible individual and the collective claims of the nation was nowhere more urgent and difficult than in the multinational Habsburg monarchy (Judson, 1996). The attempt to de-escalate national conflicts through constitutional guarantees of the right to national cultural individuality failed, but it indicated a path that would be taken with the help of the League of Nations after World War I in Europe's new multinational nation-states. This second attempt failed as well, but fundamental principles based in the guiding liberal idea were formulated: the will of the individual must be central, and no collective has the right to disregard this individual will. Making cultural assimilation the precondition for national minorities to enjoy equal citizenship rights in a society with a different national majority contradicts the central convictions of historical liberalism as it had developed since the late eighteenth century, but liberals have often violated this principle in political practice.

Organization and Milieus

Even before the emergence of lastingly organized modern parties, liberals tried to penetrate the public sphere with their

ideas and to influence state politics. Liberalism had the most favorable conditions for influence where parliamentary institutions already existed.

Secret organizations with revolutionary intentions, on the other hand, were not part of the liberal arsenal for action, because liberal politics did not aim for revolution, but rather for institutionalized, constitutionally regulated cooperation in state politics. Liberals sought to avoid revolutions or to transform them as rapidly as possible into parliamentarily legitimated reforms and thus to put them in a legal framework.

After parliaments became more significant and male voting rights expanded, political movements that sought a voice in state politics could no longer renounce party organizations. Beginning in the second half of the nineteenth century, the liberals, too, formed parties, though this was more difficult for them than for their competitors on the right and left. The guiding liberal idea of the responsible individual, who would determine the 'common good' in free discussion, continuously acted as a certain obstacle to hierarchically organized decision-making procedures. Among the liberals, the party apparatus generally had less importance than the parliamentary groups and local dignitaries.

The guiding image of civil society that has been associated with liberalism since its beginnings is also reflected in its social profile: its focus is on citizens, not the bourgeoisie. The social origins of liberal spokesmen displayed a broad spectrum, especially in the early phase. Nowhere was it limited to the bourgeoisie. Even in Great Britain, the pioneer of industrialization, from 1859 to 1874, only about 21% of the liberal delegates to the House of Commons were entrepreneurs. In Italy and Spain, liberalism found a bulwark among the nobility, and in Hungary the nobility dominated completely. In France, Germany, and Belgium, the bourgeoisie took a more prominent place in leading liberal circles. Everywhere here, liberal opinion-makers were highly respected men from respected families. They stood at a distance from tradition, but did not break with it. This membership in 'good society' was an important precondition for the power of liberal ideas to prevail even in those states that denied liberals access to state offices. A strictly bourgeois liberalism could not have achieved this.

Since the second half of the nineteenth century, the social 'catchment field' of liberal parties narrowed. They became more bourgeois. Two lines of conflict increasingly limited the liberal milieu:

1. One of the liberal parties' heaviest losses was that the workers, whose number greatly increased during the process of industrialization, turned away from them. This happened earliest in Germany (beginning in the 1860s), while in Great Britain the first workers' party still worked together with the liberals in elections (Lib–Lab coalition), but the British liberals, too, lost their worker votes when the Labor Party rapidly grew after World War I.
2. The secularity of the state was always a central liberal credo. In political practice, this goal could be tied to religious tolerance, but also to a struggle against Church competencies and claims. Intense conflicts between the state and the Church broke out in many countries on the European continent in the second half of the nineteenth century. This greatly limited the extent of liberal milieus. Catholics loyal to the Church generally turned away from liberal parties, which they saw as representatives of laicism, but the situation was different in Great Britain. Until World War I, the Irish question tied Catholic voters to the liberals, whose strongest support came from nonconformist congregations.

See also: Attitudes, Political and Civic Culture; Citizenship, Historical Development of; Citizenship: Political; Civil Society and the Public Sphere: History of the Concept; Civil Society, Concept and History of; Egalitarianism; Equality and Inequality: Legal Aspects; Equality: Philosophical Aspects; Freedom: Political; Liberal Christianity; Liberalism: Impact on Social Science; Nationalism, Historical Aspects of: The West; Nationalism: General.

Bibliography

Ackermann, B., 1992. The Future of Liberal Revolution. Yale University Press, New Haven, CT.
Arblaster, A., 1984. The Rise and Decline of Western Liberalism. Blackwell, Oxford, UK.
Bramsted, E.K., Melhuish, K.-J. (Eds.), 1978. Western Liberalism. A History of Documents from Locke to Croce. Longman, London/New York.
Dahrendorf, R., 1987. Fragmente eines neuen Liberalismus. Deutsche Verlags-Anstalt Stuttgart, Germany.
Freeden, M., 1982. The New Liberalism. Clarendon Press, Oxford, UK.
Gall, L., Koch, R. (Eds.), 1981. Der europäische Liberalismus im 19. Jahrhundert. vols. 5. Suhrkamp, Frankfurt am Main, Germany.
Hartz, L., 1955. The Liberal Tradition in America. Harcourt, Brace and World, New York.
Judson, P.M., 1996. Exclusive Revolutionaries. Liberal Politics, Social Experience and National Identity in the Austrian Empire, 1848–1914. University of Michigan Press, Ann Arbor, MI.
Langewiesche, D., 2000. Liberalism in Germany. Princeton University Press, Princeton, NJ.
Langewiesche, D., 1988. Liberalism im 19. Jahrhundert. Deutschland im europäischen Vergleich. Vadenhoeck and Ruprecht, Göttingen, Germany.
Langewiesche, D., 1993. Liberalism and the middle classes in nineteenth-century Europe: A comparative study. In: Kocka, J., Mitchell, A. (Eds.), Bourgeois Society in Nineteenth-Century Europe. Berg, Oxford, UK, pp. 40–69.
Lowi, T.J., 1979. The End of Liberalism. The Second Republic of United States, second ed. Norton, New York.
Marshall, T.H., 1950. Citizenship and Social Class. Cambridge University Press, Cambridge, UK.
Ruggiero, G., de, 1959. The History of European Liberalism. Beacon Press, Boston.
Schapiro, J.S., 1958. Liberalism. Its Meaning and History. Van Nostrand, New York.
Sheehan, J.J., 1978. German Liberalism in the Nineteenth Century. University of Chicago Press, Chicago.

Liberalism: Impact on Social Science

Richard P Bellamy, University of Reading, Reading, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8797–8801, © 2001, Elsevier Ltd.

Abstract

A doctrine with roots in eighteenth century thought, liberalism emerged in the nineteenth century as the prevailing ideology of Western capitalist societies and democracies. Philosophically, liberalism consists of a commitment to the ideals of equality, liberty, individuality, and rationality. However, liberals have divided over their social and political implications. Some liberals have believed these values to be inscribed within the very fabric of modern societies, reflecting the innovative technologies, social differentiation and free movement of labor, goods, and capital characteristic of the commercial and industrial age and its aftermath. Others have thought the links between the two are fortuitous and contingent, and that these and other related social changes, such as the growth of bureaucracy, potentially threaten liberalism. Politically united against traditional and hierarchical societies, liberals have differed, therefore, over whether liberalism requires various kinds of political support, such as the state regulation of markets, or merely the absence of any interference by either the state or others to the spontaneous and autonomous social activities of individuals. Meanwhile, critics of liberalism have either blamed it for all contemporary social ills or regarded it as a transitory phase to be replaced by superior forms of social organization.

A doctrine with roots in eighteenth century thought, liberalism emerged in the nineteenth century as the prevailing ideology of Western capitalist societies and democracies. Though much transformed by the social and political developments of the twentieth century and seriously challenged by the various movements associated with them, especially fascism and socialism in the first half of the century and feminism, ecologism, and multiculturalism in the second, liberalism has retained its dominant place. It combines philosophical, social, and political elements. Philosophically, liberalism consists of a commitment to the ideals of equality, liberty, individuality, and rationality. However, liberals have combined these concepts in numerous and often opposed configurations, which reflect different social assumptions. Many liberals have believed these values to be inscribed within the very fabric of modern societies, reflecting the innovative technologies, social differentiation, and free movement of labor, goods, and capital characteristic of the commercial and industrial age and its aftermath. Others have thought the links between the two are fortuitous and contingent, and that these and other related social changes, such as the growth of bureaucracy, potentially threaten liberalism. Politically united against traditional and hierarchical societies, liberals have differed, therefore, over whether liberalism requires various kinds of political support, such as the state regulation of markets, or merely the absence of any interference by either the state or others to the spontaneous and autonomous social activities of individuals. Meanwhile, critics of liberalism have either blamed it for all contemporary social ills, or regarded it as a transitory phase to be replaced by superior forms of social organization. Liberalism has featured in social science in correspondingly divergent ways: as the 'natural' value system of modern societies, as a historically and culturally specific phase of social development, as either pernicious or expressive of basic human goods.

Liberalism Defined

For analytical purposes it is convenient to break liberalism down into its philosophical, social, and political components, though these form a package that different liberals have interpreted in diverse ways.

Liberal Philosophy

At a philosophical level, liberals seek the maximal equal liberty possible for all individuals to rationally and freely choose and realize their various projects and preferences without interfering with each other. This position draws together the concepts of equality, liberty, individuality, and rationality in a distinctive way. Liberals value equality in the sense of denying that anyone is the natural subordinate of others. This view does not entail regarding people as the same, merely that everyone has equal moral worth. Instead of seeking equality of outcome like certain socialists, liberals simply desire that all should have an equal opportunity to deploy what talents they do possess on the same basis as everyone else. Their attachment to liberty follows from this interpretation of equality and its implicit endorsement of the moral primacy of the individual. Among moral equals, individuals should be allowed to pursue their own good in their own way to the extent that is compatible with a like liberty for others. Finally, as rationalists, liberals believe that individuals should behave in a coherent and consistent manner and that decisions in the public domain should be open to critical scrutiny and amendment and capable of attaining the reasoned assent of those they affect.

Though liberals share a moral ethos and broadly agree on the conceptual relationships between liberalism's core values, they offer very different interpretations of them reflecting a wide variety of epistemological and ontological perspectives. Thus, liberals include methodological individualists such as Adam Smith or Herbert Spencer and holists such as Emile Durkheim, materialists such as John Locke and idealists such as T.H. Green or Benedetto Croce, utilitarians such as J.S. Mill and natural rights theorists such as Tom Paine, abstract universalists such as Kant, and historicist relativists such as Herder, and so on. Moreover, their main practical differences stem less from these considerations than from contrasting social visions and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03204-9

political judgements. Though individual thinkers naturally relate these last to their particular metaphysical and methodological views, the connections between them are far from clear cut. Similar theoretical perspectives often generate quite different practical stances, while thinkers of quite diverse theoretical positions may often be in substantial practical agreement.

Liberal Society

Historically, the most significant social influence on the formation of liberalism was the passage from feudalism to capitalism from the sixteenth to the nineteenth century. This fueled the liberal attack on ascribed status and its commitment to equality of opportunity. The closed feudal order of inherited rights, with its hierarchical system of aristocratic privileges, was to give way to an open and egalitarian capitalist order, in which an individual's social position and success supposedly mirrored his or her ability and effort – a way of life they felt was best realized in a free market economy.

Some contemporary liberals and many of liberalism's critics regard the ethos resulting from such social origins as essentially materialist, instrumental, and egoistic. Traditionally, however, liberals have been extremely wary of such attitudes and have sought to avoid them. The liberal conception of economic behavior emerged from the experience of the entrepreneurs, traders, shopkeepers, and artisans of early capitalism, and invokes the image of an idealized market order in which rewards are closely linked to effort. This linkage was codified in the labor theory of value and, in the British and North American context at least, drew additional support from the religious beliefs of certain Protestant sects that encouraged industriousness as a mark of salvation. The market also acts as a discovery mechanism, with competition and the profit motive weeding out bad practices and poor products whilst stimulating the search for new processes and commodities, and so promoting the progress of society.

Liberals have disagreed, however, over how far a policy of *laissez-faire* can secure these advantages, and doubts have grown as capitalism has taken on a more corporate form. Most have accepted the need for some regulation to protect the market against itself and maintain its freedom – such as anti-trust and monopolies legislation. Others have sought to distinguish between the economic and social disadvantages that individuals are responsible for and those they are not, supporting quite extensive unemployment insurance and state-supported training schemes, for example, to get the unintentionally out of work back into jobs. Finally, liberals have always acknowledged certain public goods will be badly provided or even neglected by market mechanisms, even if some have believed state provision could employ market elements such as competitive tendering – defense and certain infrastructural goods, such as light-houses, were the classic cases. Clean air, education, and a health service serve as more recent and contentious examples.

Liberal Politics

The political face of liberalism follows on from the philosophical and social elements outlined above but was additionally influenced by the wars of religion and the rise of modern science in the sixteenth and seventeenth centuries, which fueled calls for religious toleration and the related belief in rationalism and the moral equality of individuals. Along with the shift from social structures based on status to contractual relations between civic equals, these commitments led to the demand that politics reflect liberal principles. The state, like society as a whole, came to be seen as a voluntary association held together by the mutual consent of its members rather than the ties of deference to social superiors – a conception of political order summed up in the theory of the Social Contract.

The basic components of a liberal political system were established in the wake of the British, American, and French revolutions of 1688, 1776, and 1789, respectively. All three gave rise to new constitutional settlements and bills of rights. Two criteria informed these arrangements. First, laws were to be framed and applied impartially and universally. There were to be no special exemptions for particular groups, such as clerics and nobles, as in the past. Second, the laws were to guarantee to the greatest extent possible the equal right of each individual to pursue their own plan of life. For the majority of liberals, the most basic rights in this respect were traditionally the rights to private property and freedom of belief. These rights were essential to their understanding of the virtues of the market mechanism and of toleration, which they regarded as the embodiments of the new ethos.

At the heart of liberal politics is a distinction between the state and civil society (*see* State and Society: History of the Concept). The purpose of the state is to facilitate social interaction rather than to substitute for individual initiative through state control of social institutions. The latter is the mark of dictatorships in liberal eyes. However, social liberals nevertheless believe economic intervention and state provision of certain social goods can have an enabling role, though market liberals contend such steps undermine liberalism. Liberals are also often said to favor a related distinction between the public and private sphere, allowing behavior in the latter they would ban or restrict in the former. However, this distinction has never come out cleanly. To the extent liberals do distinguish between the two it is as a corollary of their views of the state's role in regulating society. The marriage contract, for example, offers a quintessential liberal device of a public character with a clear impact on private behavior. Just as liberals of different hues dispute the degree to which workers are free to sell their labor and the extent to which a valid contract implies certain standards of pay and conditions, so they have also debated whether and in what ways the marriage contract constrains the behavior of couples and grants women, in particular, rights to common property and protection against rape and abuse.

We have already disputed the view of liberalism as promoting possessive individualism in the socioeconomic sphere. A related objection contends that liberals advocate subjectivism, skepticism, and relativism in the political sphere, with all views treated as little more than personal opinions. However, the liberal insistence on the individual's right to follow his or her own path does not entail regarding all ways of life as equally valuable. Drawing on the experimental method in modern science and the Protestant defense of freedom of conscience, liberals contend that truth and morality emerge

through individuals being allowed to call into question accepted doctrines and try out new and occasionally eccentric or potentially offensive pursuits. Ideas have to be tested for their validity while personal virtue only comes from our taking responsibility for our lives. For the state or any group to paternalistically impose its views on the rest of society risks falling into dogmatic error, demoralizes the population, and produces stagnation. Individuals must be able to make mistakes so that they can learn from them and emerge morally and materially improved.

Nevertheless, liberals have seen no need to tolerate the 'intolerable,' defined as views that are incompatible with treating others as moral equals. Some critics have regarded this position as inconsistent with liberalism's claims to openness. Yet certain limits are clearly implicit in liberalism and would otherwise render it incoherent. Where exactly these boundaries lie and their legitimacy and justifiability are more contentious matters, however. Liberalism has been accused of a gendered, imperialist, and Euro-centric bias, and historically this has often been the case, albeit to differing degrees and with certain important if qualified exceptions, such as J.S. Mill's attack on the *Subjection of Women*. How far gender and multiculturalism remain a blind spot is less clear. Many contemporary liberals maintain that multicultural, gay, and feminist demands for recognition and respect, like other appeals against unjust discrimination, implicitly invoke liberal norms of equality and tolerance. They have simply revealed these principles to have a broader application than certain past liberals had appreciated.

The complexities of the liberal position emerge when we consider their ambivalent view of democracy. Liberals have recognized that democracy offers a means for holding governments to account, obliging policies to be justified and improved through critical debate, and allowing politicians to be peacefully removed when they prove incompetent, biased, or corrupt. However, past liberals sought to limit political citizenship to those with the intellectual attainment and economic resources they believed were required for 'independent' rational decision making – criteria that usually had the effect of ruling out workers and women. Even so, liberals worried about the 'tyranny of the majority' as represented by an irrational and dependent mass, whose votes could be bought by unscrupulous populists willing to pander to their prejudices and interests. As a result, liberals have tended to insist on the need for constitutional safeguards to protect liberal values from being overturned by democratic decisions.

Liberalism and Modern Society

Social scientists have seen liberalism as either an inherent and defining characteristic of modernity or as threatened by it. Amongst the first camp, moreover, there are those who see it as symptomatic of modernity's ills rather than its advantages, identifying liberalism with a possessive individualism that is not only antisocial but antisociological as well.

Liberalism as the Ethos of Modernity

The view of liberalism as the ethos of modernity was best expressed by British social scientists such as Adam Smith, Herbert Spencer, J.S. Mill, T.H. Green, and L.T. Hobhouse, and to a lesser extent French thinkers such as Benjamin Constant, Francois Guizot, and Emile Durkheim. For these thinkers, liberalism's philosophical commitment to expanding equal individual liberty was underwritten by a theory of social evolution whereby potential conflicts between different liberties would be harmonized. This latter thesis idealized the market relations between small-scale entrepreneurs characteristic of early capitalism. These were taken as a model of a meritocratic society of self-reliant and responsible citizens, who freely contract with each other for mutual advantage. Liberal politics was conceived as the superstructural products of these social and economic changes. Freedom, reason, morality, and progress came in this way to be identified, with an optimistic account of social development supporting an implicitly ethical naturalist philosophical belief in the compatibility of different forms of self-realization.

Adam Smith's doctrine of the 'invisible hand' of the market and Herbert Spencer's characterization of the 'industrial' form of organization as realizing the 'Law of Equal Freedom' offer classic examples of this position. The first a product of the commercial society of the eighteenth century, the second of the early industrial age of the 1840s, by the late nineteenth century the concentration of capital and the rise of organized labor had made liberals rather less sanguine about the inevitability of their vision. To the dismay of Spencer and other adherents of the earlier 'classical' liberalism, liberals began to envisage a more proactive role for the state in sustaining favorable social conditions. T.H. Green and L.T. Hobhouse, for instance, saw policies such as factory hours regulation and unemployment insurance as mechanisms for promoting freedom of contract and freeing individuals from certain structural constraints. Importantly, though, both saw the state as reducing the mutual interferences between individuals and fostering autonomy. In a similar manner, Durkheim argued that the liberal 'cult of the individual' was inherent to industrial societies and that the modern division of labor fostered an organic solidarity between autonomous agents. However, he had to confront various commercial crises, bankruptcies, and conflicts between capital and labor that suggested a less harmonious picture. He attributed these disturbances to contemporary societies being in a 'transitional phase' in which people had yet to fully assimilate liberal social norms. He too saw a role for the state in fostering these attitudes, with various 'intermediary institutions' serving to inform state policies as to social needs and act as a counter weight to state oppression of the individual (*see* Social Question: Impact on Social Thought).

Critics of liberalism sometimes portray this supposed relationship between liberalism and modernity in negative terms. They associate liberalism with a naturally instrumental and individualist model of human behavior, a disposition they believe is encouraged by the liberal conception of the market and democracy. They also claim it produces a view of social science as following the physical sciences by elaborating general laws of social development on the basis of various constant human passions. Some commentators have even argued that it was only radical and conservative opponents of liberalism who appreciated the distinctiveness of the human sciences by noting, respectively, how human behavior was modified by social structures and by culture, and the capacity of

individuals to interpret and evaluate circumstances in different ways. At best, liberalism results in a positivistic bourgeois ideology that ignores the impact on human thought and action of material conditions and productive forces, historically evolving communal traditions and conventions, and the possibility for reflective criticism. Thus liberalism is associated with certain behaviorist versions of political science and economics but not with sociology or anthropology. These arguments result from a highly selective and tendentious reading of the liberal tradition whereby, for example, De Tocqueville becomes a 'conservative,' Durkheim either a 'conservative' or a 'socialist,' and Herbert Spencer, read as an advocate of egoistic possessive individualism, becomes the exemplar of liberal social science. The liberal tradition, however, is much more diverse than this caricature, both methodologically and substantively. The vast majority of liberals distinguished respect for 'individuality' from 'egoism' and even positivist behaviorists, such as J.S. Mill, appreciated that autonomy only flourished given certain material conditions, cultural values, and economic structures and played a part in its turn in how these were fashioned in the future. This awareness of the social and historically contingent character of liberalism is no more evident than in those liberals who saw it as being undermined by modernity.

Liberalism as Threatened by Modernity

The threat posed to liberalism by elements of modernity was especially evident in countries such as Italy and Germany, where a modernizing economy and society went along with fragile and incomplete liberal regimes which ultimately succumbed to fascism. The challenge these developments posed for liberal social science is powerfully illustrated by the 'frustrated' liberalism of the Italian Vilfredo Pareto and the 'disenchanted' liberalism of Weber. Though commonly categorized as sociologists, their work ranges over the whole gamut of the social sciences, covering economics, politics, and, in Pareto's case, social psychology. Pareto was in many respects a sophisticated exponent of the instrumentally rational actor model of liberal individualism and was initially a strong advocate of both *laissez-faire* economics and a libertarian and democratic politics. However, he became convinced that what he called 'logico-experimental' thinking was the exception rather than the rule. Instead, most conduct was 'nonlogical' and arose from certain common psychological dispositions of either a conservative or more inventive kind. Politics simply consisted of the manipulation of these psychic states, a process that had been greatly aided by contemporary mass democracy. The electorate was too large and poorly informed and the decisions needed too numerous and diverse for rational debate to be possible. Democracy simply allowed populist leaders to legitimize their rule via emotive appeals to a popular will that was largely of their own making. He even briefly supported fascism as a possible way of breaking free from this system and establishing a liberal market order from above.

Weber, by contrast, offered a more historically and sociologically nuanced account of similar phenomena. He disputed that economic development per se fostered liberalism or that liberal values could be justified a priori. Liberalism's apparent basis in natural law had been historically contingent, reflecting the peculiar social and political experiences of early modern Europe outlined above. Not only was this constellation of factors unlikely to reappear, but even within established liberal regimes liberalism was threatened by the spread of economic instrumentalism, bureaucratic rationality, and corporate organization resulting from the further development of capitalism. The chances of liberal democracy being established in societies such as Russia, which had not benefited from the earlier favorable historical experiences, were minimal, therefore. Quite different cultural values were likely to shape the modernizing processes of such societies, as indeed has occurred throughout Asia where capitalism has by and large taken a nonliberal form. Meanwhile in already existing liberal societies, the autonomous action characteristic of liberalism was possible only for the political leaders and entrepreneurs who controlled and gave direction to the various political and economic administrative structures that govern our lives (*see* Modernity).

Most contemporary liberal thinkers regard Weber as too pessimistic and many believe liberal values can be justified on grounds distinct from their historical and social origins. However, the possible disjunction between late modernity and liberalism, on the one hand, and the search for liberal foundations that appeal to reasons that ought to be accessible to all and so compatible with a variety of different cultures and ideals, on the other, remain the prime challenges a viable liberal theory has to meet.

See also: American Revolution, The; Civil Liberties and Human Rights; Civil Society, Concept and History of; Conservatism: Historical Aspects; Dependency Theory; Freedom: Political; French Revolution, The; Liberalism: Historical Aspects; Liberalism: Political Doctrine and Impact on Social Science; Mill, John Stuart (1806–73); Modernity: Anthropological Aspects; Modernity: History of the Concept; Modernity; Smith, Adam (1723–90); Sociological Traditions.

Bibliography

Arblaster, A., 1984. The Rise and Decline of Western Liberalism. B Blackwell, Oxford, UK.
Bellamy, R.P., 1992. Liberalism and Modern Society. Polity Press, Cambridge, UK.
Bellamy, R.P., 2000. Rethinking Liberalism. Continuum, London and New York.
De Ruggiero, G., 1927. The History of European Liberalism. Oxford University Press, Oxford, UK.
Gray, J., 1986. Liberalism. Open University Press, Milton Keynes, UK.
Seidman, S., 1983. Liberalism and the Origins of European Social Theory. B Blackwell, Oxford, UK.

Liberalism: Political Doctrine and Impact on Social Science

Richard Bellamy, University College London (UCL), London University, London, UK; and European University Institute, Florence, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A doctrine with roots in eighteenth-century thought, liberalism emerged in the nineteenth century as the prevailing ideology of Western capitalist societies and democracies. Philosophically, liberalism consists of a commitment to the ideals of equality, liberty, individuality, and rationality. However, liberals have divided over their social and political implications. Some liberals have believed these values to be inscribed within the very fabric of modern societies, reflecting the innovative technologies, social differentiation, and free movement of labor, goods, and capital characteristic of the commercial and industrial age and its aftermath. Others have thought the links between the two are fortuitous and contingent, and that these and other related social changes, such as the growth of bureaucracy, potentially threaten liberalism. Politically united against traditional and hierarchical societies, liberals have differed, therefore, over whether liberalism requires various kinds of political support, such as the state regulation of markets, or merely the absence of any interference by either the state or others to the spontaneous and autonomous social activities of individuals. Meanwhile, critics of liberalism have either blamed it for all contemporary social ills or regarded it as a transitory phase to be replaced by superior forms of social organization. However, contemporary liberals seem increasingly critics themselves of the inequality and abuse of rights found in modern societies, although how far their proposals can be made a reality remains uncertain.

A doctrine with roots in eighteenth-century thought, liberalism emerged in the nineteenth century as the prevailing ideology of Western capitalist societies and democracies. Though much transformed by the social and political developments of the twentieth century and seriously challenged by the various movements associated with them, especially fascism and socialism in the first half of the century and feminism, ecologism, and multiculturalism in the second half, liberalism has retained its dominant place. It combines philosophical, social, and political elements. Philosophically, liberalism consists of a commitment to the ideals of equality, liberty, individuality, and rationality. However, liberals have combined these concepts in numerous and often opposed configurations, which reflect different social assumptions. Many liberals have believed these values to be inscribed within the very fabric of modern societies, reflecting the innovative technologies, social differentiation, and free movement of labor, goods, and capital characteristic of the commercial and industrial age and its aftermath. Others have thought that the links between the two are fortuitous and contingent, and that these and other related social changes, such as the growth of bureaucracy, potentially threaten liberalism. Politically united against traditional and hierarchical societies, liberals have differed, therefore, over whether liberalism requires various kinds of political support, such as the state regulation of markets, or merely the absence of any interference by either the state or others to the spontaneous and autonomous social activities of individuals. Meanwhile, critics of liberalism have either blamed it for all contemporary social ills or regarded it as a transitory phase to be replaced by superior forms of social organization. Liberalism has featured in social science in correspondingly divergent ways: as the 'natural' value system of modern societies, as a historically and culturally specific phase of social development, and as either pernicious or expressive of basic human goods.

Liberalism Defined

For analytical purposes, it is convenient to break liberalism down into its philosophical, social, and political components, though these form a package that different liberals have interpreted in diverse ways.

Liberal Philosophy

At a philosophical level, liberals seek the maximal equal liberty possible for all individuals to rationally and freely choose and realize their various projects and preferences without interfering with each other. This position draws together the concepts of equality, liberty, individuality, and rationality in a distinctive way. Liberals value equality in the sense of denying that anyone is the natural subordinate of others. This view does not entail regarding people as the same, but merely that everyone has equal moral worth. Instead of seeking equality of outcome like certain socialists, liberals simply desire that all should have an equal opportunity to deploy what talents they do possess on the same basis as everyone else. Their attachment to liberty follows from this interpretation of equality and its implicit endorsement of the moral primacy of the individual. Among moral equals, individuals should be allowed to pursue their own good in their own way to the extent that is compatible with a like liberty for others. Finally, as rationalists, liberals believe that individuals should behave in a coherent and consistent manner and that decisions in the public domain should be open to critical scrutiny and amendment and capable of attaining the reasoned assent of those they affect.

Though liberals share a moral ethos and broadly agree on the conceptual relationships among liberalism's core values, they offer very different interpretations of them reflecting a wide variety of epistemological and ontological perspectives. Thus, liberals include methodological individualists such as

 http://dx.doi.org/10.1016/B978-0-08-097086-8.93070-8

Adam Smith or Herbert Spencer and holists such as Emile Durkheim, materialists such as John Locke and idealists such as T.H. Green or Benedetto Croce, utilitarians such as J.S. Mill and natural rights theorists such as Tom Paine, abstract universalists such as Kant and historicist relativists such as Herder, and so on. Moreover, their main practical differences stem less from these considerations than from contrasting social visions and political judgments. Though individual thinkers naturally relate these last to their particular metaphysical and methodological views, the connections between them are far from clear-cut. Similar theoretical perspectives often generate quite different practical stances, while thinkers of quite diverse theoretical positions may often be in substantial practical agreement.

Liberal Society

Historically, the most significant social influence on the formation of liberalism was the passage from feudalism to capitalism from the sixteenth to the nineteenth century. This fueled the liberal attack on ascribed status and its commitment to equality of opportunity. The closed feudal order of inherited rights, with its hierarchical system of aristocratic privileges, was to give way to an open and egalitarian capitalist order in which an individual's social position and success supposedly mirrored his or her ability and effort – a way of life they felt was best realized in a free market economy.

Some contemporary liberals and many of liberalism's critics regard the ethos resulting from such social origins as essentially materialist, instrumental, and egoistic. Traditionally, however, liberals have been extremely wary of such attitudes and have sought to avoid them. The liberal conception of economic behavior emerged from the experience of the entrepreneurs, traders, shopkeepers, and artisans of early capitalism and invokes the image of an idealized market order in which rewards are closely linked to effort. This linkage was codified in the labor theory of value and, in the British and North American context at least, drew additional support from the religious beliefs of certain Protestant sects that encouraged industriousness as a mark of salvation. The market also acts as a discovery mechanism, with competition and the profit motive weeding out bad practices and poor products while stimulating the search for new processes and commodities, and so promoting the progress of society.

Liberals have disagreed, however, over how far a policy of *laissez-faire* can secure these advantages and doubts have grown as capitalism has taken on a more corporate form. Most have accepted the need for some regulation to protect the market against itself and maintain its freedom – such as antitrust and monopolies legislation. Others have sought to distinguish between the economic and social disadvantages that individuals are responsible for and those they are not, supporting quite extensive unemployment insurance and state-supported training schemes, for example, to get the unintentionally out of work back into jobs. Finally, liberals have always acknowledged certain public goods will be badly provided or even neglected by market mechanisms, even if some have believed state provision could employ market elements such as competitive tendering – defense and certain infrastructural goods, such as lighthouses, were the classic cases. Clean air, education, and a health service serve as more recent and contentious examples.

Liberal Politics

The political face of liberalism follows on from the philosophical and social elements outlined above, but was additionally influenced by the wars of religion and the rise of modern science in the sixteenth and seventeenth centuries, which fueled calls for religious toleration and the related belief in rationalism and the moral equality of individuals. Along with the shift from social structures based on status to contractual relations among civic equals, these commitments led to the demand that politics reflects liberal principles. The state, like society as a whole, came to be seen as a voluntary association held together by the mutual consent of its members rather than the ties of deference to social superiors – a conception of political order summed up in the theory of the Social Contract.

The basic components of a liberal political system were established in the wake of the British, American, and French revolutions of 1688, 1776, and 1789, respectively. All three gave rise to new constitutional settlements and bills of rights. Two criteria informed these arrangements. First, laws were to be framed and applied impartially and universally. There were to be no special exemptions for particular groups, such as clerics and nobles, as in the past. Second, the laws were to guarantee to the greatest extent possible the equal right of each individual to pursue their own plan of life. For the majority of liberals, the most basic rights in this respect were traditionally the rights to private property and freedom of belief. These rights were essential to their understanding of the virtues of the market mechanism and of toleration, which they regarded as the embodiments of the new ethos.

At the heart of liberal politics is a distinction between the state and civil society (*see* State and Society: History of the Concept). The purpose of the state is to facilitate social interaction rather than to substitute for individual initiative through state control of social institutions. The latter is the mark of dictatorships in liberal eyes. However, social liberals nevertheless believe that economic intervention and state provision of certain social goods can have an enabling role, though market liberals contend that such steps undermine liberalism. Liberals are also often said to favor a related distinction between the public and private sphere, allowing behavior in the latter they would ban or restrict in the former. However, this distinction has never come out cleanly. To the extent liberals do distinguish between the two, it is as a corollary of their views of the state's role in regulating society. The marriage contract, for example, offers a quintessential liberal device of a public character with a clear impact on private behavior. Just as liberals of different hues dispute the degree to which workers are free to sell their labor and the extent to which a valid contract implies certain standards of pay and conditions, so they have also debated whether and in what ways the marriage contract constrains the behavior of couples and grants women, in particular, rights to common property and protection against rape and abuse.

We have already disputed the view of liberalism as promoting possessive individualism in the socioeconomic sphere. A related objection contends that liberals advocate

subjectivism, skepticism, and relativism in the political sphere, with all views treated as little more than personal opinions. However, the liberal insistence on the individual's right to follow his or her own path does not entail regarding all ways of life as equally valuable. Drawing on the experimental method in modern science and the Protestant defense of freedom of conscience, liberals contend that truth and morality emerge through individuals being allowed to call into question accepted doctrines and try out new and occasionally eccentric or potentially offensive pursuits. Ideas have to be tested for their validity, while personal virtue only comes from our taking responsibility for our lives. For the state or any group to paternalistically impose its views on the rest of society risks falling into dogmatic error, demoralizes the population, and produces stagnation. Individuals must be able to make mistakes so that they can learn from them and emerge morally and materially improved.

Nevertheless, liberals have seen no need to tolerate the 'intolerable,' defined as views that are incompatible with treating others as moral equals. Some critics have regarded this position as inconsistent with liberalism's claims to openness. Yet, certain limits are clearly implicit in liberalism and would otherwise render it incoherent. However, where exactly these boundaries lie and their legitimacy and justifiability are more contentious matters. Liberalism has been accused of a gendered, imperialist, and Eurocentric bias, and historically this has often been the case, albeit to differing degrees and with certain important if qualified exceptions such as J.S. Mill's attack on the *Subjection of Women*. How far gender and multiculturalism remain a blind spot is less clear. Many contemporary liberals maintain that multicultural, gay, and feminist demands for recognition and respect, like other appeals against unjust discrimination, implicitly invoke liberal norms of equality and tolerance. They have simply revealed these principles to have a broader application than certain past liberals had appreciated.

The complexities of the liberal position emerge when we consider their ambivalent view of democracy. Liberals have recognized that democracy offers a means for holding governments to account, obliging policies to be justified and improved through critical debate, and allowing politicians to be peacefully removed when they prove incompetent, biased, or corrupt. However, past liberals sought to limit political citizenship to those with the intellectual attainment and economic resources they believed were required for 'independent' rational decision-making – criteria that usually had the effect of ruling out workers and women. Even so, liberals worried about the 'tyranny of the majority' as represented by an irrational and dependent mass, whose votes could be bought by unscrupulous populists willing to pander to their prejudices and interests. As a result, liberals have tended to insist on the need for constitutional safeguards to protect liberal values from being overturned by democratic decisions.

Liberalism and Modern Society

Social scientists have seen liberalism either as an inherent and defining characteristic of modernity or as a threat to it. Among the first camp, moreover, there are those who see it as symptomatic of modernity's ills rather than its advantages, identifying liberalism with a possessive individualism that is not only antisocial but antisociological as well.

Liberalism as the Ethos of Modernity

The view of liberalism as the ethos of modernity was best expressed by British social scientists such as Adam Smith, Herbert Spencer, J.S. Mill, T.H. Green, and L.T. Hobhouse, and to a lesser extent French thinkers such as Benjamin Constant, Francois Guizot, and Emile Durkheim. For these thinkers, liberalism's philosophical commitment to expanding equal individual liberty was underwritten by a theory of social evolution whereby potential conflicts between different liberties would be harmonized. This latter thesis idealized the market relations between small-scale entrepreneurs characteristic of early capitalism. These were taken as a model of a meritocratic society of self-reliant and responsible citizens, who freely contract with each other for mutual advantage. Liberal politics was conceived as the superstructural products of these social and economic changes. Freedom, reason, morality, and progress came in this way to be identified, with an optimistic account of social development supporting an implicitly ethical naturalist philosophical belief in the compatibility of different forms of self-realization.

Adam Smith's doctrine of the 'invisible hand' of the market and Herbert Spencer's characterization of the 'industrial' form of organization as realizing the 'law of equal freedom' offer classic examples of this position. The first was a product of the commercial society of the eighteenth century, the second of the early industrial age of the 1840s. By the late nineteenth century, the concentration of capital and the rise of organized labor had made liberals rather less sanguine about the inevitability of their vision. To the dismay of Spencer and other adherents of the earlier 'classical' liberalism, liberals began to envisage a more proactive role for the state in sustaining favorable social conditions. T.H. Green and L.T. Hobhouse, for instance, saw policies such as factory hours regulation and unemployment insurance as mechanisms for promoting freedom of contract and freeing individuals from certain structural constraints. Importantly though, both saw the state as reducing the mutual interferences between individuals and fostering autonomy. In a similar manner, Durkheim argued that the liberal 'cult of the individual' was inherent to industrial societies and that the modern division of labor fostered an organic solidarity between autonomous agents. However, he had to confront various commercial crises, bankruptcies, and conflicts between capital and labor, which suggested a less harmonious picture. He attributed these disturbances to contemporary societies being in a 'transitional phase' in which people had yet to fully assimilate liberal social norms. He too saw a role for the state in fostering these attitudes, with various 'intermediary institutions' serving to inform state policies as to social needs and act as a counterweight to state oppression of the individual (*see* Social Question: Impact on Social Thought).

Critics of liberalism sometimes portray this supposed relationship between liberalism and modernity in negative terms. They associate liberalism with a naturally instrumental and individualist model of human behavior, a disposition they believe is encouraged by the liberal conception of the market

and democracy. They also claim that it produces a view of social science as following the physical sciences by elaborating general laws of social development on the basis of various constant human passions. Some commentators have even argued that it were only radical and conservative opponents of liberalism who appreciated the distinctiveness of the human sciences by noting, respectively, how human behavior was modified by social structures and culture, and the capacity of individuals to interpret and evaluate circumstances in different ways. At best, liberalism results in a positivistic bourgeois ideology that ignores the impact on human thought and action of material conditions and productive forces, historically evolving communal traditions and conventions, and the possibility for reflective criticism. Thus, liberalism is associated with certain behaviorist versions of political science and economics, but not with sociology or anthropology. These arguments result from a highly selective and tendentious reading of the liberal tradition whereby, for example, De Tocqueville becomes a 'conservative,' Durkheim either a 'conservative' or a 'socialist,' and Herbert Spencer, read as an advocate of egoistic possessive individualism, becomes the exemplar of liberal social science. The liberal tradition, however, is much more diverse than this caricature, both methodologically and substantively. The vast majority of liberals distinguished respect for 'individuality' from 'egoism' and even positivist behaviorists, such as J.S. Mill, appreciated that autonomy only flourished given certain material conditions, cultural values, and economic structures and played a part in its turn in how these were fashioned in the future. This awareness of the social and historically contingent character of liberalism is no more evident than in those liberals who saw it as being undermined by modernity.

Liberalism as Threatened by Modernity

The threat posed to liberalism by elements of modernity was especially evident in countries such as Italy and Germany, where a modernizing economy and society went along with fragile and incomplete liberal regimes which ultimately succumbed to fascism. The challenge these developments posed for liberal social science is powerfully illustrated by the 'frustrated' liberalism of the Italian Vilfredo Pareto and the 'disenchanted' liberalism of Weber. Though commonly categorized as sociologists, their work ranges over the whole gamut of the social sciences, covering economics, politics, and, in Pareto's case, social psychology. Pareto was in many respects a sophisticated exponent of the instrumentally rational actor model of liberal individualism and was initially a strong advocate of both *laissez-faire* economics and a libertarian and democratic politics. However, he became convinced that what he called 'logicoexperimental' thinking was the exception rather than the rule. Instead, most conduct was 'nonlogical' and arose from certain common psychological dispositions of either a conservative or more inventive kind. Politics simply consisted of the manipulation of these psychic states, a process that had been greatly aided by contemporary mass democracy. The electorate was too large and poorly informed and the decisions needed too numerous and diverse one for rational debate to be possible. Democracy simply allowed populist leaders to legitimize their rule via emotive appeals to a popular will that was largely of their own making. He even briefly supported fascism as a possible way of breaking free from this system and establishing a liberal market order from above.

Weber, by contrast, offered a more historically and sociologically nuanced account of similar phenomena. He disputed that economic development per se fostered liberalism or that liberal values could be justified a priori. Liberalism's apparent basis in natural law had been historically contingent, reflecting the peculiar social and political experiences of early modern Europe outlined above. Not only was this constellation of factors unlikely to reappear but even within established liberal regimes, liberalism was threatened by the spread of economic instrumentalism, bureaucratic rationality, and corporate organization resulting from the further development of capitalism. The chances of liberal democracy being established in societies such as Russia, which had not benefited from the earlier favorable historical experiences, were minimal, therefore. Quite different cultural values were likely to shape the modernizing processes of such societies, as indeed has occurred throughout Asia, where capitalism has by and large taken a nonliberal form. Meanwhile, in already existing liberal societies, the autonomous action characteristic of liberalism was possible only for the political leaders and entrepreneurs who controlled and gave direction to the various political and economic administrative structures that govern our lives (*see* Modernity).

Most contemporary liberal thinkers regard Weber as too pessimistic and many believe liberal values can be justified on grounds distinct from their historical and social origins. However, the possible disjunction between late modernity and liberalism, on the one hand, and the search for liberal foundations that appeal to reasons that ought to be accessible to all and so compatible with a variety of different cultures and ideals, on the other, remain the prime challenges a viable liberal theory has to meet.

Contemporary Liberalism

Contemporary liberal philosophers have sought to overcome the Weberian challenge by seeking to divorce liberal principles from controversial ethical, epistemological, and ontological viewpoints. They argue that the chief advantage of liberal theories of the market, democracy, and justice resides in letting people holding different moral and material priorities and preferences live peaceably and profitably together. However, in spite of this agreed approach, these theorists split along the main ideological divisions of the contemporary world. The gradual electoral drift of liberals to social democracy, on the one hand, and to conservatism, on the other, has been reflected in a parallel division in contemporary liberal philosophy between social democratic and New Right, neo-liberal theorizing. I shall briefly examine an example of liberal thinking from both these camps.

Liberalism between Social Democracy and Neo-Liberalism

The major exponent of contemporary social democratic liberalism is John Rawls (1921–2002), whose *A Theory of Justice* (1972) is probably the most important work of political philosophy written in English since World War II. Rawls uses

the classical liberal device of a social contract to arrive at two principles of justice, which he believes will serve to regulate the relations among individuals within modern societies. He asks us to imagine which principles we would choose to adopt if we knew neither what sort of person we were nor the precise circumstances of the society we might be born into. All we know are certain general facts about human psychology, sociology, economics, and politics.

Rawls believes that this thought experiment, which he elaborates in great detail, will produce a theory of justice that everyone will accept as legitimate because it is untainted by any preconceptions about the worth or lack of it of particular beliefs or ways of life. Rather, we shall be drawn to appreciate the value of showing equal concern and respect to all people, regardless of their views. For, faced with the radical uncertainty of what Rawls calls the 'original position,' we will seek to arrive at principles of justice that are as fair as possible to all persons, whatever their talents or ideals might be. Indeed fairness, rather than truth or goodness, is the only acceptable criterion of modern justice for Rawls.

What then are Rawls' two principles and what kind of policies do they require for their realization? According to the first principle, the state must provide the most extensive system of equal liberties possible. All individuals must enjoy the greatest degree of liberty consistent with the enjoyment of a like liberty by everyone else. According to the second principle, there must be equal opportunity to obtain positions and status in society, regardless of color, gender, or creed. To help achieve this goal, each person is to have an equal share of wealth *except* when an unequal distribution would improve the welfare of the least well-off in society, by providing incentives that encourage people to work harder and produce more in ways that raise the living standards of everyone. Even so, a progressive taxation policy would be necessary if the trickle-down effect failed to produce any benefits from this greater affluence for those at the bottom of society. Like earlier social liberals, such as Hobhouse or Durkheim, therefore, Rawls tries to combine the liberal values of liberty and equality with a concern for social justice tempered, in characteristic social-liberal fashion, by the requirements of efficiency. Unlike earlier social liberals, however, he has sought to dissociate his argument for these policies from the traditional liberal ethos and metaphysics.

Three years after Rawls' book appeared, Robert Nozick (1938–2002) published *Anarchy, State and Utopia* (1974), defending the market and a minimal state and criticizing any form of welfarist policy on libertarian grounds. He argues that any attempt to redistribute resources or to intervene in the economy involves an illegitimate attempt to use people to further some social purpose with which they may not agree. In his view, the advantage of the market is that it reflects no particular conception of the good, the resulting distributions being merely the unplanned consequence of numerous exchanges among individuals. So long as nobody is deliberately and physically coerced into exchanging goods or selling his or her labor to another and there are no legal barriers discriminating against who can apply for jobs or what can be sold, then the liberal criteria of liberty and equality have been satisfied and the market system can be judged as fair. The inequalities that result from its operation follow no strict pattern. Those who do well cannot be said in any absolute sense necessarily to deserve or merit their success – they have merely been fortunate to be born to rich rather than poor parents or to have had talents that proved to be marketable. Their entitlement to their wealth rests simply on the right not to be interfered with by others, for what we have obtained through transactions which honor the right of each and every individual not to be coerced is rightfully ours. As he puts it, governments have no legitimate grounds for meddling with capitalistic acts between consenting adults.

Just as Rawlsean social liberalism has been taken up by social democrats, so New Right theories such as Nozick's, which offer a neoclassical liberal defense of the free market, have been taken up by conservatives. Indeed, his argument does contain a strong conservative bias, since it rests on an explicitly historical view of justice based on the pattern of distribution that evolves through market exchanges over time. However, Nozick never claims to be a conservative and others with similar views, such as Hayek, explicitly reject the label. As Roger Scruton has pointed out in his book *The Meaning of Conservatism*, Nozick's right-based argument is in direct opposition to the Old Right emphasis on "the web of obligations by which citizens are bound to each other and the state." Moreover, although his theory accords best with *laissez-faire* capitalist society, he draws on the anarchist and Utopian traditions to suggest that it could be compatible with any form of social organization provided it was voluntarily entered into. In common with other contemporary liberal thinkers, the central commitment of his theory is to a pluralist framework which provides scope for people to live according to diverse conceptions of the good.

How convincing are these two versions of liberalism? Once we leave the Rawlsean 'original position,' for example, and discover who and where we are, what would motivate us to adhere to the two principles if they appeared to conflict with our most important ideals or interests? Surely, only someone who had swallowed a good dose of liberalism in the first place would feel at all inclined to distance themselves from their cherished beliefs and social and material attachments in the way Rawls' theory requires? Similarly, Nozick offers no reasons for why we have certain rights rather than others, while his defense of Locke's 'labor mixing' theory of property ownership rests on a dubious metaphysics. As so-called left-libertarians, such as Hillel Steiner and Philippe van Parijis, have noted, if there is no predetermined distribution of land and no reason to believe that – all things being equal – some people deserve more property than others, it makes sense to think of resources as commonly owned. There is no reason to think that the first use gives an entitlement to ownership. If so, land remains in common and those who use it owe the rest a 'rent.' Nor is it at all clear that either theory will be pluralist in outcome. In other words, neither theory seems to work unless it invokes some of the ethos of earlier versions of liberalism. To the extent that this ethos has proved inadequate, then so are these versions of liberalism.

Liberal Ideals in a Nonideal Society

For all their philosophical sophistication, such theories seem politically and sociologically naive, greatly exaggerating the

possible independence and opportunities available to individuals within complex societies. The division of labor, international markets, the organizational structures of modern economies – from large-scale bureaucratic states to multinational corporations and financial and political institutions such as the IMF and the EU – these and other factors make theories that take as their starting point the autonomous individual agent, capable of making his or her way in the world through talent, effort, and voluntary agreements with other individuals, appear unrealistic. As a result, liberal philosophy has shifted from offering an idealization of the modern social reality to providing an ideal critique of that reality.

Rawls' last two works that followed up *A Theory of Justice – Political Liberalism* (1993) and *The Law of Peoples* (1999) – in many respects mark the transition between these two forms of liberal theorizing. In the first work, he made plain that his arguments were designed for liberal democratic states and sought to outline their underlying logic and the public attitudes and reasoning that citizens and public officials needed to adopt for them to work legitimately. He also suggested that justice should be seen primarily as political and secondarily as social and economic – a means for fairly resolving the many disagreements citizens have about how to shape their collective life. In the second work, he offered a sketch of a theory of international justice that controversially both took the state as the main actor in international affairs and sought to accommodate 'decent' but nonliberal democratic regimes.

These works have been heavily criticized by many who admired *A Theory of Justice* and had regarded it as a theory of universal and global scope and that applied to individuals rather than to political communities. Rawls' response was that he had aimed at a realistic Utopia – a vision that seeks ideal principles for a nonideal world, not least because of the morally and rationally imperfect people who inhabit it. Nevertheless, some believe these works made too many concessions to the status quo, giving too much leeway to cultural differences and social inequalities in particular. They contend that liberal ideals should produce changes in the world rather than simply understanding it, and have sought to deploy his arguments to advocate major reforms to the state system to address global poverty and climate change, among other issues. However, others believe his realism did not go far enough. They suggest that a realistic liberalism takes the inevitability of political conflict as its core, and that a global egalitarianism has to accommodate cultural and social differences, especially when embodied within existing self-governing political communities.

Conclusion

In certain respects, liberalism has become a hegemonic political language. Even those who do not espouse liberal values have a tendency to invoke them by appealing to toleration, equality, or rights when seeking a hearing for their views. Yet, the pervasiveness of liberal discourse also proves problematic. On the one hand, it produces a plethora of liberalisms so that it becomes difficult to identify a coherent and generally accepted liberal program. On the other hand, liberalism can seem either as an ideal at odds with the realities of contemporary societies, and as such either unrealized or possibly unrealizable, or as providing current realities with either a justified or spurious legitimation. Liberals and their critics can be found espousing all of these positions: indeed, even the same individual may voice elements of each of them with regard to different issues. Yet, the continuing relevance and interest of liberalism rests on its dual character as both the ideal and the reality of modernity, at once its best defense and critique. To be convincing, it has to be able both to understand and to change contemporary societies – the challenge confronting liberal theorists is to show that it can be both real enough to do the first and ideal enough to achieve the second.

See also: Civil Liberties and Human Rights; Civil Society, Concept and History of; Conservatism: Historical Aspects; Dependency Theory; Freedom: Political; French Revolution, The; Individualism versus Collectivism: Philosophical Aspects; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Liberalism: Historical Aspects; Mill, John Stuart (1806–73); Modernity: Anthropological Aspects; Modernity; Smith, Adam (1723–90).

Bibliography

Arblaster, A., 1984. The Rise and Decline of Western Liberalism. B Blackwell, Oxford, UK.
Bellamy, R., 1992. Liberalism and Modern Society. Polity Press, Cambridge, UK.
Bellamy, R., 1999. Liberalism and Pluralism: Towards a Politics of Compromise. Routledge, London.
Bellamy, R., 2000. Rethinking Liberalism. Continuum, London and New York.
De Ruggiero, G., 1927. The History of European Liberalism. Oxford University Press, Oxford, UK.
Gray, J., 1986. Liberalism. Open University Press, Milton Keynes, UK.
Gray, J., 2000. The Two Faces of Liberalism. Polity Press, Cambridge.
Seidman, S., 1983. Liberalism and the Origins of European Social Theory. B. Blackwell, Oxford, UK.

Liberation Theology in Social Work

Carolina Muñoz-Guzmán, Pontificia Universidad Católica de Chile, Santiago, Chile

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The notion of empowerment has been a common outcome of both liberation theology and anti-oppressive social work and provided models for practice. These liberating movements rooted in the 1960s represented a challenge for traditional practices within the Catholic Church and social work practice, respectively. The concept of hope crosses both perspectives relating hopelessness and powerlessness, in a positive causal relation whose end point is the cessation of oppression. The aims of liberation theology and anti-oppressive social work are to promote emancipation by identifying the structural conditions wherein the roots of oppression lie.

Liberation Theology and Social Work: Anti-Oppressive Concerns Are Common Grounds

The notion of empowerment has been a common outcome of both liberation theology and anti-oppressive social work and provided an important element in models for practice. These liberating movements became rooted in Latin America, during the decade of 1960s and represented a challenge for traditional practices within the Catholic Church in the first case, and within the social work practice in the second. Both models have had an enduring legacy in developing forms of social involvement and political engagement.

The concept of hope encompasses both perspectives and has established a significant relation between hopelessness and powerlessness, especially in promoting in a positive causal relation in ending oppression. The aims of liberation theology and anti-oppressive social work are related to promoting emancipation and identifying the structural conditions wherein the roots of oppression are embedded. Therefore, both perspectives conduct a process of conscientization that increases awareness about the condition of victims in a social, economic, and political system that produces oppression. The process of conscientization requires collective support to realize that oppression is not an individual problem but a shared one and to set up strategic agendas promoting social change. The related concepts of hope, empowerment, and liberation should be kept in mind when studying liberation theology and social work.

Liberation Theology as Practical Hope for the Oppressed

Liberation theology meant practical hope for Latin America's poor communities seeking to overcome human hardships, because emancipation could be realized in reality through liberation and not by expecting a Messiah (Scherer-Warren, 1990). Traditionally in Latin America, the Catholic Church had failed to implement its role of liberating oppressed peoples, primarily because it became affiliated to those in power. However, with the growing importance and defense of social equality and equity in the 1960s and 1970s, the Church began to advocate for social justice for marginalized or oppressed peoples (Noelliste, 1995; Smith, 1991); as Levine stated "once seen as a cultural reservoir for apathy and fatalistic resignation, religion, above all the Catholicism of popular classes, now commonly appears synonymous with solidarity and resistance to injustice and repression in cases other-wise as different as Brazil, Chile, or El Salvador" (1988: p. 242).

The emergence of liberation theology was a reaction to the politics and economy of the late 1960s and early 1970s, which included repressive regimes and vast capitalist expansion in Latin America (Noelliste, 1995; Smith, 1991; Levine, 1988; Boff, 1980). The Second Vatican Council ended by concluding with a socially oriented doctrine, especially for people in what was then called the Third World who were defined as the proletariat of contemporary humanity. This change in the Catholic Church's focus and newfound interest in social justice brought many new people into the Church, and became very popular within the sociopolitical context of the Latin America of the time. The new view has been defined as based on a critical sociostructural analysis, anchored on Marxist categories of class, conflict, and exploitation where an explicit effort at analysis aimed to understand the social context through the eyes of poor people and reducing distances between the Church and average believers (Levine, 1988: p. 243).

The social context then prevailing in Latin America was one of the general crisis affecting the economical, social, political, and ideological spheres. The region was becoming an object of development for industrialized countries, and Latin American peoples became concerned with the levels of exploitation and marginalization that they were living within because development was being achieved at the expense of poor people while benefiting traditional elites and partners of international companies (Smith, 1991; Boff, 1980). There was awareness about the contradictions produced by political and economic policies. On one side, there was capitalist development, and on the other, there were several populist measures involving marginalized groups that had suffered failures similar to those of the Alliance for Progress (1961–69). These capitalist initiatives impoverished a high proportion of the Latin American population who became doubly affected by tragedy through the repression exercised under the emergency powers of the military dictatorships that had assumed control in many countries (Dodson, 1979; Boff, 1980; Gutierrez, 1984; Silva, 2009).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28051-3

An awareness of their conditions gave rise to social discontent. With the support of pastoral organizations from universities and grassroots organizations, these people became engaged in the implementation of other strategies aimed at raising higher levels of consciousness about inequality and oppression. The results were the development of the Basic Educational Movements and Christian grassroots communities that were composed of lay people and nuns. These movements were strongly influenced by the Brazilian educator, Paulo Freire (1970) who advocated the development of critical consciousness to facilitate the process of liberation. The Second Vatican Council (1962–65) gave institutional support to these movements that increasingly faced rejection from their official governments. Consequently, some of these became militarized to fight social movements and keep control of government. The Second Vatican Council also contributed to reflections about a faith-based liberation praxis that searched for and validated social justice, participation, dignity, and fraternity.

Classical writers of liberation theology, like Boff (1978) and Gutierrez (1984), when analyzing the Latin American situation, made the distinction between 'developing countries,' 'countries in development,' and 'developed countries.' These definitions favored the last category and stated a country would be considered a developed country only when it reached a specific income per capita, degree of consumerism, a determined number of vehicles, hospitals, schools, libraries, life expectancy, and other variables that defined it as a developed country (Boff, 1978). Thus, it would only be possible to be considered a developed country by adequately conforming to the culture and expectations of the Western countries then dominating the world. Underdeveloped countries were categorized as dependent countries. These countries had started to incorporate the presence and expectations of the dominant countries both internally and externally in the hopes of becoming a developed country. However, by allowing the domination or guidance of developed countries, they became dependent upon those, and thus laid the grounds for further oppression and exploitation (Gunder-Frank, 1969).

These types of relations explained that dependency theory became one of the underlying explanations of the subjugation experienced by people in Latin America who were living under an imposed dependency emanating from the center poles of development in relation to the peripheral ones. These ideas generated controversy and broke the notion of interdependency among countries, when the oppressive character of the labor relations in Latin America were exposed and revealed the suffering of its workers and how this impeded real social progress and transformation. Furthermore, there was an acknowledgment of the causal relations between opulence and poverty. Oppression represented an historical condition for Latin American peoples, and liberation theology became a legitimate vehicle through which channel the liberation demands of oppressed populations (Boff, 1980; Ellacuría, 1989; Gutierrez, 1990). There have been some critics to this notion of dependency, including those considering it reductionist because it turned people into homogeneous passive victims. Dodson (1979: pp. 208–209) has claimed that dependency theory constituted an overreaction by most the theologians of liberation because their awareness of dependency theory was limited, reducing it to a simplified and homogeneous understanding of Latin American realities when the region is characterized by great complexity and diversity in its socioeconomic and political conditions.

At that time, the Encyclical, Popularum Progressio (1967) was especially directed toward Latin America to promote the overcoming of injustice and marginalization. In this scenario, the Episcopal Council of Latin America was created. Several meetings of Latin American Bishops were celebrated, one of them taking place in Medellín, Colombia (1968). This meeting reinforced the process of radicalization of the Latin American Church, a process that culminated with the third meeting of Latin American Bishops in Puebla, Mexico (1979) which also kept its strong commitment with a Catholic Church that engaged with oppressed peoples.

Liberation theology focused not only on critically analyzing reality, but also on transforming it. Transforming reality was made possible through the methodological originality of this theology, namely, its claims for transformative action through praxis. This praxis should advance toward salvation understood as a process of integral liberation including the economy (liberation from starving) and politic (liberation from marginalization); culture (liberation from analphabetism or illiteracy and ignorance) and pedagogy (liberation from depersonalized dependency); and religious (liberation from traditional understanding of sin and rejection of the Historical plan of God) (Silva, 2009: pp. 157–158; Alberigo and Jossua, 1987: p. 92).

The Bishops meetings produced clarity about the meaning of liberation theology and its methodology which was composed of three fundamental steps. The first was to ascertain a reflexive practice about reality (moment to see); the second was to enlighten these practices with a theological thinking (moment of judgment); and finally the moment to set up the tracks for pastoral action (moment of action). Under these movements, the Second Vatican Council appeared enlightened by the transformation of Latin America and not the reverse.

The social and political effervescence produced by this consciousness among oppressed and marginalized peoples ended in 1968 when the initiative was assumed by several military governments in Latin America, most of them highly repressive. Many pastoral communities suffered repression, torture, and prison. Repressive measures were also visited upon secular people, as well as nuns and priests, like the case of Monsignor Romero in El Salvador, who suffered a tragic demise, because he spread liberation beliefs and practices to the highest levels of the Church hierarchy (Katra, 2012: p. 37). Similar situations occurred in other countries because ruling oligarchies were threatened by the liberation church. Their responses became increasingly punitive and were sometimes expressed as "obscene repression like the one exerted against the progressive Church in Guatemala, with the martyrdom of priests, nuns, and lay workers; the forced exile of the Maryknoll and Jesuit orders; the official support for rival Protestant groups; and the de facto disappearance of pastoral agents from some regions" (Katra, 2012: p. 37). This situation gave a new perspective to the theology of liberation and the theology of Cautiverio, concept that was understood as confinement, captivity, which emerged as a new moment of reflection wherein the interest of national security disguised the defense of capitalism (Berryman, 1987).

The Theology of Cautiverio emerged as many dictatorships were installed throughout Latin America: Brazil (1964), Bolivia (1971), Uruguay (1973), Chile (1973), and Argentina (1976). Previous military governments turned toward right-wing ideologies, like Peru (1975) and Ecuador (1976), and repressive governments continued in power in most parts of Central America, only Mexico, Colombia, Venezuela, and Costa Rica stayed under formal democratic mechanisms, and even these governments used hard measures of repression (Silva, 2009). Whereas prior to the oppressive regimes, the Catholic Church and its members could publicly fight inequalities, after the military coups in these countries, they had to fight and try to change things from with the surveillance of the dictatorship and its repressive regulations. Cautiverio was related to liberation theology; the difference, however, is that its focus lay on dependence instead of freedom. "The Theology of Cautiverio differentiates between being free from and being free for; being free from represents a confined situation, whereas being free for assumes having control over the situation" (Boff, 1980: p. 119). Practicing under a military dictatorship made the latter option extremely difficult.

Therefore, the Theology of Cautiverio was not only social, political, and economical, it was "the embodiment of a deeper deviation of man that reaches the fundamental sense of being and living as understood by [the] power-conquest-domination [nexus], engrained in oppression, repression and a global[ly] confined regime. It is history told by those who triumphed by force, of those on top, of those who were able to assume the decisive power over others. Nobody talks anymore about lost hope of those humbled and assassinated" by the brutality of their own rulers (Boff, 1980: p. 179).

According to Boff (1984), Gutiérrez (1970), and Scherer-Warren (1990) to face this crisis, only five lines of theological-pastoral actions were provided by the Church and promoted in the different meetings, from Medellín (1968) to (PUEBLA, 1979: p. 1220).

Preferred Action for Poor People

There was a crucial change when the Catholic Church no longer had an interest in blaming poor people. It had become a Church with the poor, thereby implying a more active role for poor people in defining the pastoral action of the Church and assuming that the condition of poverty was caused by structural factors based on economic, social, and political arrangements. "This change will not be truly honest. At least it goes along to a change of personal and collective mind-sets about the ideal of a human dignity and human happiness, all of which dispose people favorably toward conversion" (PUEBLA, 1979: p. 1152).

Integral Liberation

This notion overcame the idea of liberation being restricted to the spiritual sphere and includes a wider understanding that is a practical, political, and social action to enhance the space of freedom among poor populations.

Foundational Christian Communities (Known as Ecclesiastic Grassroots Communities)

The people, individually and collectively, build a plural society, and the Church acknowledges that they are agents of a process of general conscientization about common social responsibilities and as communities that bring the voice of people into transforming the Church (PUEBLA, 1979: p. 1220).

Human Rights as God's Rights

Human rights were to be protected by the Catholic Church, especially at a historical moment when people's movements were politically restricted and their rights were violated.

Preferred Action for Young People

Young people represented more than half of the total of Latin American population at that time. The Catholic Church trusted these young and saw in them its future hope; Latin American young people had the true potential for present and future evangelization because they could be used to mobilize the social corps and particularly the ecclesial corps and renew its structures and populations (PUEBLA, 1979: p. 1186).

The methods used to generate awareness among the oppressed peoples of Latin America (poor people, women, children and young people, black people, and indigenous peoples) included the action of bringing together people who suffered the same type of oppression to form a group identity, encourage the rediscovery of their dignity, and stimulate their confidence in change (Scherer-Warren, 1990).

Liberation theology was an answer to modern culture, a revision that acknowledged the main victories of modern thought: consciousness and through autonomy, religious freedom, secularization, and the historicity of humanity, which means "a desire to read the signs of the times, valuing and learning from the world in which the Church is intimately and inextricably entangle" (Levine, 1988: p. 244) and therefore, in search of the meaning and direction of historical change.

The Second Vatican Council allowed the vision of the Catholic Church as part of the world and not separated from it, which meant the search for recovering the dignity of the human person and transforming oppressed peoples into agents of their own history, process that happens through the interpretation of everyday life and not through the obedience to pre-existing orders. This understanding gave support to a plurality of interpretations enriching political and social action and producing results in the apostolic work of the Church and in the commitment to civil organization for human rights, as well as in the engagement in actions aimed at promoting radical changes. Therefore, liberation theology offered a hope in the power of collective organization to promote change for a better society, becoming a basis for social action and political commitment.

A contribution from liberation theology has been the consideration of the insight into reality brought by people; it shows a hermeneutical position that appreciates the "value of ordinary experience as a guide to reflection and action" (Levine, 1988: p. 254), where the method for inquiry values an interpretive discipline, framed by the particular historical

context being explored. Therein lies a fundamental commonality between liberation theology and the understanding of social work as a profession in search of an anti-oppressive world.

Anti-Oppressive Social Work: Fighting the Roots and Effects of Social Oppression

Social work in the 1960s questioned its professional practice as one profoundly influenced by society's economic and political forces, and marked by apathy in response to inequities caused by structural sources of oppression over excluded groups in society. These critics gave origin to Structural Social Work in Canada (Lundy, 2004; Moreau, 1989; Mullaly, 2007), and in Britain, it began as Radical Social Work (Bailey and Brake, 1975) which brought social workers in close proximity to service users through action-research and attempts to form alliances with them and to develop new forms of practice. Although this quickly moved to Marxist Social Work (Corrigan and Leonard, 1978), this approach was found lacking for ignoring crucial social divisions such as 'race' and gender and gave rise to alternative theories and paradigms for practice. Crucial among these were Antiracist Social Work (Dominelli, 1988, 1997, 2008), Feminist Social Work (Dominelli and McLeod, 1989; Dominelli, 2002b), and Anti-oppressive Practice (Dominelli, 2002a). In Australia, these critiques were elaborated further through postmodernism and what became termed as Critical Social Work (Fook, 2002).

Structural social work, like radical social work and its successors, was conceived of social work within a critical tradition concerned with the broad socioeconomic and political dimensions of society that disrupted egalitarian relations among individuals and communities, and aimed to promote the emancipation of oppressed populations and fight structural barriers that affected the living conditions of service users. Structural and radical social work rejected traditional professional views and definitions of practice. Similar to the demand of liberation theology in regard to eliminating its top-down and hierarchical relations with people, and becalming involved in more dialogical and collaborative relationships, structural social work assumed that similar actions would help the process of advancing institutional and structural change within social services (Moreau, 1989).

The antiracist, feminist, anti-oppressive, and critical approaches to social work introduced a dominant concern with ensuring the realization of social justice and human rights as an integral part of their critique of structural barriers to equality and in providing the foundation for paradigms of practice that were led by service users' participation in what was developed. Thus, these approaches appealed to "social work's value base" (Dominelli, 2002a: p. 4) and sought to link personal troubles to public issues as a way of undermining structural barriers to liberation for all marginalized and disenfranchised groups. Thus, all these perspectives shared an understanding of the relation between individual action and social structures and the impact each one has on each other, and demanded a permanent critical self-examination that allowed advancement toward change that reshaped mainstream social services and probation provisions (Dominelli, 2002a).

Radical social work as proposed by Bailey and Brake (1975: pp. 9–10) acknowledged the need to understand the position of oppressed people in the context of the social and economic structures they lived in, and like liberation theology was based on a socialist perspective that challenged the traditional professional perspective that supported ruling-class hegemony to argue that social work should assist people to understand their alienation in terms of their oppression and build up their self-esteem (Bailey and Brake, 1975: pp. 9–10).

Corrigan and Leonard (1978) tried to promote Marxist social work and emphasized that social workers and service users should avoid experiencing their social world from an individualized perspective. Instead, they argued for collective action based on solidarity between social workers and service users though group action. It sought to build on the existing richness of the existing working-class movement by bringing together community political activity with collective class action through the working class movement including trade unionism to achieve growth. This emphasis on a collective understanding and support is a feature it shares with liberation theology.

Dominelli (1988, 1997, 1997b, 2008), one of the foremost educators and theoreticians in British social work, was instrumental in developing radical social work in new directions. This culminated in *Green Social Work* which introduced environmental justice as an integral part of social justice within the framework of citizenship-based entitlements and human rights. This had been preceded by *Anti-oppressive Social Work* (2002a), which built on the first engagement with an *Anti-racist Social Work* perspective, which sought to advance toward "social justice and its capacity to challenge exiting social relations through strategic collaborations with other like-minded professionals" (1997b: p. 7). This challenge for a collective and reflective action in the profession, evolved through the time, giving an account of globalization through which Dominelli (2002a, 2004) called for the enhancement social justice and well-being within the complexities of the racialization processes that focused not only on service delivery, but also on the very labor process that affected social workers themselves and their professional labor. Additionally, the theme of social justice and equalitarianism was promoted through an analysis of gender relations and the discussion of feminist social work (Dominelli and McLeod, 1989; Dominelli, 2002b). These also developed research and scholarship in the arena of feminist social work that concerned itself not only with the well-being of women, but also that of men and children, themes that others subsequently studied further. This corpus of work defined social problems as socially constructed and rooted in political practice rather than abstract theory with little relevance to the realities of life as experienced by ordinary people and marginalized groups, insights essential for emancipatory social work (Dominelli, 1997).

In search for a more complex concept of power, Dominelli (2002a) challenged binary thinking, and through anti-oppressive social work created a body of knowledge that replaced the dyad of empowerment or oppression with a continuum of power relations that moved from oppression to anti-oppression to non-oppression in complex and fluid ways that were negotiated by all those involved in particular interactions, rather than being predetermined, thus leading to

major advances in contemporary social work. Anti-oppressive practice considered the daily experience of oppression in which people lived as framed by social divisions and structural inequalities, including the oppression generated through social work practice itself. Anti-oppressive practices searched for "a more egalitarian value system concerned with reducing the deleterious effects of structural inequalities upon people's lives; a methodology focusing on both process and outcome; and a way of structuring relationships between individuals that aims to empower users by reducing the negative effects of social hierarchies on their interaction and the work they do together" (Dominelli, 1996: p. 3). Thus, anti-oppressive social work seeks to empower people and engage their consciousness and capacity to exercise agency through their interaction with others, ensuring that power was not simply the exercise of control over others, but a "complex phenomenon which is constantly negotiated and renegotiated between social actors" (Dominelli and Gollins, 1997: p. 396).

Finally, Fook (2002) and Healy (2000) espoused critical social work in Australia, highlighting advances in theorizing the concept of power with a complexity of influences that shape power relations and reduce the potential of social work to exercise power on behalf of powerless people and the importance of empowering disadvantaged people through their own agency rather than disempowering them through social labeling.

In considering reality as socially constructed and embedded in power relations between different groups of people, anti-oppressive and critical approaches to social work share a similar emphasis, with liberation theology, and promote practice that enhances the value of everyday experiences and utilizes the insight of ordinary people to critically review and critique the forms of social work practice to which they are exposed. This progressive advance of social work in identifying inputs, processes, methodologies, social relations, and outcomes to find interpretations of social reality affecting peoples' lives in a holistic manner approximates the praxis proposed by liberation theology in terms of promoting a reflexive practice about reality, enriched by a moment of judgment exerted by the oppressed population itself and finally reaching the moment of setting up action, an action that has been concerned with the fundamental questions proposed by Dominelli (1996) about what, where, when, why, who, and how social work could proceed with the aim of bringing under analysis the different elements of oppression, and promoting organizational, societal, and personal change.

Latin American social work was not indifferent to these reflections, assuming a critical perspective of capitalist societies, impelled by the movement involving the re-conceptualization of social work (Faleiros, 2011; Netto, 1976), based on a Marxist analysis about capitalism, searching for transformations of social structures, critically analyzing social work's daily practices, and questioning its servile disposition toward dominant social structures.

The contribution that anti-oppressive social work gives to the discipline is the notion of empowerment as a deliberated process whose protagonist is the person gaining power and taking over control in his or her life. This new role allows people greater access to social resources that not only favor their personal goals, but also their collective aims. Anti-oppressive practice seeks to raise critical awareness and for oppressed communities to understand structural inequalities and obstacles, contribute in the development of their capacities and social capital to increase their control and influence in society, and improve their access to social resources and their participation in social decision making.

The challenges identified by anti-oppressive social work have been acknowledged by the International Association of Schools of Social Work (IASSW) and the International Federation of Social Workers (IFSW), as both organizations have expressed as one the main concerns of the social work profession "the liberation of people to enhance well-being" (IFSW, 2002). This commitment implies a permanent reflexive practice for social workers as they engage with the constraints professionals face in relation to their different and competing loyalties when playing dual roles as helper and controller, and permanent institutional demands, frameworks, and situations that require highly ethical awareness in and responses to professional interventions (Strier, 2007).

Structural social work, as a seed for critical action in the profession, and anti-oppressive social work as an empowerment challenge for oppressed communities, had highlighted and questioned the structural origins of service users' problems and called for radical social change. Anti-oppressive social work theory is also based on critical analyses of practice relations, promoting a critical practice theory that demanded social justice in practice and policy. The achievement of these aims requires defeating the structures of oppression that are reproduced through ideological processes, power relations, and routine practices, and underpin radical social change as within sustained, inclusive, and participatory models of action. This explains Fook's claim that consciousness of a social context informed by the experience of service users becomes one of the earliest principles of a critical social work (2002). Therefore, anti-oppressive social work drives its forces toward making visible sociopolitical and economic contexts that affect individuals and develop awareness about their collective nature, as well as helping in reaching a wider consciousness and critical thinking as a way of empowering marginalized people (Lundy, 2004).

Empowerment, anti-oppression, and critical awareness emerged from social work's development as well as from liberation theology as these searched for an emancipation process at the service of oppressed people. Both movements embrace a similar mission, which is the fight against the roots and effects of social oppression. Additionally, both movements demanded political and social models oriented toward liberation and social collaboration.

Social work and liberation theology shared the emancipatory perspective as common ground, to change the structures and procedures in relating to oppressed peoples. The Catholic Church shifted from being for the poor when the liberation perspective compelled it to being with the poor. Social work's anti-oppressive practice was required to make visible the power that users have, instead of defining them a priori as disempowered. Both approaches challenged established nostrums and declared that the promised land is on earth, the question of difference was about the steps needed to transform these

challenges from a starting point of recognizing oppression to reaching an arrival point of liberation.

Challenges of a Liberating Perspective

Despite the enduring legacies of liberation theology and anti-oppressive social work, there remain conventional forces in the Catholic Church and mainstream social work contributing to the hegemony of oppressive structures including racism, classism, patriarchy, and sexism which maintain conformance toward the culture and expectations of the dominant social system.

Social work can play a significant role in resisting the interests of traditional and conservative groups that pursue their own interests against the most disadvantaged groups. Its impact could be strengthened through research guided by the principles of the anti-oppressive movements studied here: hope, empowerment, and liberation. If social work research legitimates the incidence of participatory and action research, it would value the outcomes found in studies where under-represented populations have a central role influencing the inquiry process. As Strier (2007: p. 25) proposed for anti-oppressive research, policy formation would be less controlled by those who have the power and authority 'when participants are acknowledged as experts of their own lives and capable of self-determination' and the outcomes used to benefit these participants that include formative processes of collective unity that recognize a common identity, and advance toward a more egalitarian society. One prerequisite of this process is a transformation of the power relations between researchers and participants, evolving toward horizontal relationships, as the way to understand oppressive social relationships (Strier, 2007). Such understanding will help to reduce alienation produced by the silence and invisibility suffered by excluded social groups.

Strier (2007) proposed restoring the liberating vocation of social work beyond practice by making research a core dimension of challenging dominant research patterns that can become tools for oppression. This requires the capture of silenced service user voices, the assumption of a critical stance toward social policy that emphasizes a restricted vision of what constitutes evidence-based practice and valued only as effective outcomes (Strier, 2007). In conclusion, critical research, interested in promoting liberation from oppression, hope, and empowerment needs to consider partnership, not paternalistic relationships between researchers and research subjects drawn from marginalized groups. Its results are especially relevant for Latin American societies, where the prevalence of hierarchical relations based on class distinctions act against the engagement of individuals as active agents rather than passive recipients in social action as well as in the research processes. The option for an inclusive research practice requires a defensive strategy to beat the struggle of being opposed to regular models of power and abusive relations, because the empowerment of civil society promised in Latin America is by earthly salvation through theology of liberation which had not translated easily into real power at the grassroots levels. This gives urgency for new impulses whereby to realize the material hopes and counter disillusionment among ordinary people.

See also: Catholicism; Christianity in Latin America; Dependency Theory; Marxist Approaches to Social Work; Oppression; Poverty, Geography of; Poverty, Sociology of; Religion: Mobilization and Power; Religion: Morality and Social Control; Social Action and Social Justice; Social Justice and Social Inequalities; Social Work Theory.

Bibliography

Alberigo, G., Jossua, J.P., 1987. La recepción del Vaticano II. Ediciones Cristiandad, Madrid.
Bailey, R., Brake, M., 1975. Radical Social Work. Pantheon Books, London.
Berryman, P., 1987. Liberation Theology. The Essential Facts about the Revolutionary Movement in Latin America and beyond. Pantheon Books, New York.
Boff, L., 1978. Gracia y liberación del hombre: experencia y doctrina de la gracia. Cristiandad, Madrid.
Boff, L., 1980. Teología del cautiverio y de la liberación. Ediciones Paulinas, Madrid.
Boff, L., 1984. Desde el lugar del pobre. Ediciones Paulinas, Bogotá, Colombia.
Corrigan, P., Leonard, P., 1978. Social Work Practice under Capitalism: A Marxist Approach. Macmillan, London.
Dodson, M., May 1979. Liberation theology and Christian radicalism in contemporary Latin America. Journal of Latin American Studies 11 (1), 203–222.
Dominelli, L., 1988. Anti-racist Social Work, first ed. BASW Macmillan, London.
Dominelli, L., 1996. Deprofessionalizing social work: anti-oppressive practice, competencies and postmodernism. British Journal of Social Work 26 (2), 153–175.
Dominelli, L., 1997a. Sociology for Social Work. Macmillan, London.
Dominelli, L., 1997b. Anti-racist Social Work, second ed. Palgrave Macmillan, London.
Dominelli, L., 2002a. Anti-oppressive Social Work Theory and Practice. Palgrave Macmillan, London.
Dominelli, L., 2002b. Feminist Social Work Theory and Practice. Palgrave Macmillan, London.
Dominelli, L., 2004. Social Work: Theory and Practice for a Changing Profession. Polity Press, Cambridge.
Dominelli, L., 2008. Anti-racist Social Work, third ed. BASW Palgrave Macmillan, London.
Dominelli, L., Gollins, T., 1997. Men, power and caring relationships. The Sociological Review 45 (3), 396–415.
Dominelli, L., McLeod, E., 1989. Feminist Social Work. Macmillan Education, London.
Ellacuría, I., 1999. Fe y justicia. Desclée de Brouwer, Bilbao.
Faleiros, V., 2011. Desafíos del trabajo social frente a las desigualdades. Emancipação, Ponta Grossa 11 (1), 117–128.
Fook, J., 2002. Social Work, Critical Theory and Practice. Sage, London.
Freire, P., 1973. Education for Critical Consciousness. Seabury Press, New York.
Gunder-Frank, A., 1969. Capitalism and Under-Development in Latin-America. Monthly Review Press, London.
Gutiérrez, G., 1970. Apuntes para una Teología de Liberación. CEP, Lima.
Gutiérrez, G., 1984. Beber en su propia pozo. El itinerario espiritual de un pueblo. Editorial Sigueme, Salamanca.
Gutiérrez, G., 1990. Teología de la liberación: perspectivas. CEP, Lima.
Healy, K., 2000. Social Work Practices. Contemporary Perspectives on Change. Sage, London.
IFSW, 2002. Definition of Social Work (in 16 languages). Author, Berne, Switzerland.
Katra, B., 2012. The sign of the times: the changing faces of liberation theology. Analytic Teaching and Philosophical Praxis 33 (2), 33–45.
Levine, D., Marzo 1988. Assessing the impacts of liberation theology in Latin America. The Review of Politics 50 (02), 241–263.
Lundy, C., 2004. Social Work and Social Justice. Broadview Press, Peterborough, Ontario.
Moreau, M., 1989. Empowerment through a Structural Approach to Social Work. A Report from Practice. National Welfare Grants Program, Health and Welfare Canada, Ottawa.
Mullaly, B., 2007. The New Structural Social Work, third ed. Oxford University Press, Toronto.

Netto, J.P., 1976. La crisis del proceso de reconceptualización del servicio social. In: Alayón, N. (Ed.), Desafío al servicio social. ¿Crisis de la reconceptualización?, vol. 20. de Colección Desarrollo social Humanitas, Argentina, pp. 85–105.

Noelliste, D., July 1995. Towards a theology of theological education. Evangelical Review of Theology 19 (3), 299.

PUEBLA, 1979. III Conferencia General del Episcopado Latinoamericano. http://www.celam.org/conferencias/Documento_Conclusivo_Puebla.pdf (accessed 15.08.13.).

Scherer-Warren, I., 1990. Rediscovering our dignity: an appraisal of the utopia of liberation in Latin America. International Sociology 5, 11.

Silva, S., 2009. La Teología de la Liberación. Teología Y Vida L, 93–116.

Smith, C., 1991. The Emergence of Liberation Theology: Radical Religion and Social Movement. University of Chicago Press.

Strier, R., July 2007. Anti-oppressive research in social work: a preliminary definition. British Journal of Social Work 37 (5), 857–871.

Libertarianism

Chandran Kukathas, University of New South Wales, Campbell, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8801–8807, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Libertarianism is a largely American political movement of the twentieth century. It is also a political ideology whose main commitment is to individual liberty. Libertarianism is a species of liberalism – one which is highly skeptical of – the authority of the state, and which, at the extreme, rejects its legitimacy altogether. Two ideas are fundamental to libertarianism. First, each individual is entitled to live as he or she thinks right, provided he or she does not try to force others to live in the same fashion. Second, each individual has the right to acquire and keep property. If government is necessary, it is only to uphold these fundamental rights. Libertarians have generally been defenders of the market order, though their philosophical arguments have varied. Among their most important representatives have been Robert Nozick, Ayn Rand, Murray Rothbard, and F.A. Hayek. There is also a variant identified as left-wing libertarianism, found notably in the writings of Hillel Steiner.

Predominantly an American political movement of the twentieth century, libertarianism is a political doctrine whose deepest commitment is to liberty. Libertarians believe that the freedom of the individual is the most important value that should guide political reflection and that the good society is one in which that freedom is respected. Yet libertarians disagree about what a commitment to individual liberty entails politically and about the basis of this commitment. To understand libertarianism, then, it is necessary to come to terms with the variety within its tradition and with its relationship with other political ideas.

Libertarians approach liberty by asking what the legitimate functions of the state are. Broadly, they offer two kinds of answers. The libertarian anarchist answer is that no state is legitimate; for minimal state libertarians, however, the state is legitimate to the extent that its activities are confined to a limited range of functions, including national defense, the protection of individual rights, and the enforcement of contracts. Even within these alternatives, however, further divisions are to be found. Libertarians differ over the proper scope of government but also over the foundations of individual freedom. Any consideration of libertarianism, then, would have to account not only for the fundamental principles or commitments which give it shape and for the differences between minimalist and anarchist libertarianisms, but also for the varying doctrines advanced by libertarians from Robert Nozick to Murray Rothbard.

Fundamentals of Libertarianism

Origins

Libertarianism is a species of liberalism. Liberalism is a tradition of political thought which understands the good society as one in which individuals are free to pursue their own ends rather than one which enforces one particular understanding of what constitutes the good life. Central to liberalism is a commitment to individual liberty. For the liberal, liberty is a good which ought to be honored and protected; and doing so requires recognizing that each and every individual is entitled to a domain within which he or she may act without interference from others. Liberalism, however, describes a broad tradition which includes thinkers as diverse as John Locke in the seventeenth century and John Rawls in the twentieth century.

Libertarianism is that side of liberalism which emphasizes the importance of individual freedom from the state and which is generally wary of the operation of government. It is skeptical about power but especially about governmental power. Modern libertarians see themselves in this regard as the heirs of a long tradition in European thought, which champions the rights of the individual against the state. While they insist that this tradition stretches back into medieval times (when kings could not create law or legislate but only find or declare law,) it is to writers of the seventeenth century that they most commonly turn when looking for the first expressions of libertarian thinking. In particular, it is the ideas of the Levellers – John Lilburne, William Walwyn, and Richard Overton – and later of John Locke, which provide the inspiration for much of libertarianism.

Libertarians take two fundamental ideas from these thinkers. The first is that each individual is entitled to live as he or she chooses, provided he or she does not try to force others to live as he or she thinks right. The second is that each individual is entitled to acquire and keep property. If government is necessary, it is to protect these rights by securing individuals against those who would invade or aggress against their property and their persons. But any government that itself violates these rights by coercing them or taking their property without their consent, is tyrannical and may rightly be resisted.

Libertarianism has its origins in these ideas and its heroes are thinkers who have, in different ways, asserted the rights and liberty of the individual against the state. However, libertarianism as an intellectual and political movement did not emerge until after World War I in the United States. It arose largely as a response by some intellectuals to the growth of government and the rise of socialism. The most important writers who criticized socialism, central planning, and the growth of government were the Austrian economists Ludwig von Mises and F.A. Hayek who, throughout the 1930s and the 1940s,

debated the advocates of central planning over the feasibility of socialism. Although based in Europe, their work received a more enthusiastic reception in the United States among intellectuals hostile to the politics and economics of the New Deal. These included H.L. Mencken, Albert Jay Nock, Garet Garret, Rose Wilder Lane, Isabel Paterson, and Henry Hazlitt – journalists, writers, and economists who wrote warning of the dangers to constitutionalism posed by the growth of the state, touting the merits of free-market economics, and defending the virtues of American individualism.

It was in the 1950s that the term 'libertarianism' first came into use to describe this political stance, when Leonard Read, who established the Foundation for Economic Education, began calling himself a libertarian. While the term was rejected by writers like Ayn Rand and her followers, it came into use in the 1960s; and in 1971 the Libertarian Party was founded.

Principles

Libertarians share two sets of convictions. First, all individuals are entitled to live as they choose, free from interference by other persons or by the state. Second, in the absence of deliberate design by government or the state, order will prevail: granted the freedom to contract and exchange, markets will coordinate the production and distribution of goods better than any other available institutional mechanism. Yet, while most libertarians see these two convictions as consistent and mutually supportive, they differ over which is the fundamental basis of libertarian thinking. For Mises the commitment to individual freedom rests on straightforwardly utilitarian considerations: capitalism produces greater human welfare than socialism or the mixed economy. Yet for Murray Rothbard, the rights of the individual are defensible independently from any consequentialist argument.

Both of these convictions are intimately associated with libertarian thinking; however, it is the first which is more important philosophically, since it better identifies what is theoretically distinctive about libertarianism. For libertarians are particularly concerned to defend the freedom of individuals to live without being subject to aggression by others – against their persons or the property they may rightly acquire. An early statement of the first aspect of this claim was offered by the Leveller Richard Overton, in *An Arrow Against All Tyrants*: "To every individual in nature is given an individual property by nature, not to be invaded or usurped by any: for everyone as he is himself, so he hath a self-propriety, else could he not be himself, and on this no second may presume to deprive any of, without manifest violation and affront to the very principles of nature …" (Overton, 1646: p. 68). This is the thesis of self-ownership. And for libertarians, out of this thesis comes the contention that individual freedom includes the freedom to own private property. The most famous formulation of this claim is John Locke's in his *Second Treatise of Government*: "Though the earth, and all inferior creatures be common to all men, yet every man has a property in his own person. This nobody has any right to but himself. The labor of his body, and the work of his hands, we may say, are properly his. Whatsoever then he removes out of the state that nature hath provided, and left it in, he hath mixed his labor with, and joined to it something that is his own, and thereby makes it his property."

The centrality to libertarianism of self-ownership and the corollary right to own property, has led many to conclude that, perversely, the fundamental value for libertarians is not 'liberty' but the right to private property (including, but by no means only, one's property in one's person). All individuals must enjoy liberty; but only that liberty that is consistent with a respect for the property rights of others.

This understanding of individual rights is the basis of the libertarian view of the justification and proper role of the state. If individuals own themselves and have a right not to be aggressed against, no government is legitimate save that which enjoys the consent of the people. And the purpose of government is nothing more or less than protecting individuals against those who would violate their rights by invading their persons or their property. Libertarianism, in this regard, is profoundly antipaternalistic. It is no part of the purpose of government to improve people or society or even to provide for their welfare. It should not busy itself with human needs or deserts. Its only concern is individual rights.

The Market

Although its philosophical principles are what make libertarianism distinctive, the emphasis its adherents place on free markets and the virtues of market society is no less important. Libertarians defend capitalism and the market order. For many libertarians, capitalism is simply that form of economic organization which grows out of the freedom of individuals to exchange private property. For others, market society should be defended because it has important virtues of its own.

For thinkers like Hayek and Mises, the defense of the market was crucial to the defense of freedom because they saw the challenge to freedom as coming from modern socialism. Socialism, seemingly being put into practice in the former Soviet Union and gaining intellectual adherents in the West, criticized the market for being, at once, insensitive to the poor, who fared badly under capitalism, and less efficient than a properly planned economy. Markets were wasteful in production and inequitable in distribution. Socialism's attack on the market made the defense of the market necessary.

In the 1930s, Hayek and Mises devoted their energies to showing the impossibility of socialism understood as a form of economic organization which eschewed markets. In the absence of the signals sent out by market prices, economic coordination could not happen and central planners could not acquire the knowledge needed to effect such a coordination; for that knowledge was dispersed throughout society and often available temporarily and locally. Coordination without markets would produce shortages of some goods and an excess of others: supply and demand would never be in step. Hayek further argued that the deeper danger of socialist central planning was not economic but political. In order to pursue the goal of production without markets, planners would have to resort increasingly to coercion to curb people's natural proclivities to trade. Central planning required the creation of monopolies; and this would require suppressing those whose activities, intentionally or not, competed with monopoly suppliers. In the long run, Hayek argued, this posed a threat to political freedom, for it would undermine the rule of law by granting to some, privileges under the law.

This argument resonated particularly with libertarians in the USA in the 1950s. Here it was deployed less against socialism than against the expanding regulatory state. Government intervention in economic life, libertarians claimed, was neither necessary nor benign. More often than not, it upheld privileges – notably those enjoyed by producers and professionals seeking protection from competition which might challenge their monopolies. The most notable proponent of this argument in popular form was Milton Friedman, notably in his book, *Capitalism and Freedom* (Friedman, 1962). However, libertarians, unlike conservatives, also challenged the defensibility of government intervention, not only in the economy but also in other areas of society. Friedman and others took issue with the draft, the legal prohibition of drug use, licensing, and various other forms of compulsion by the state.

Another important target of the libertarian challenge was the welfare state. The state was not needed to provide what markets clearly could: health care, disability and unemployment insurance, and pensions, among other things. Government, libertarians argued, would, like any monopoly supplier, provide such services at greater cost and lower quality than would be found under competitive conditions. In part they objected to the limiting of individual liberty: compulsory retirement insurance, for example, compelled people to invest old age savings with the state, which provided only one of a range of possible retirement contracts – and not a particularly favorable one at that. The reply that such institutions were needed to ensure that people were protected should they fail to make provision for themselves, was rejected as paternalistic.

Nonetheless, libertarians have always disagreed over whether or not it is part of government's role to provide a welfare safety net. Hayek maintained that government needed to supply some forms of welfare because markets and charity would not be able to meet the demand. Friedman argued for a negative income tax to ensure that no one dropped below some minimum threshold. Insofar as libertarians have accepted the need for a public policy response to such problems as welfare, however, they have sought to remove the discretionary power of governments, which they think more likely to be captured by particular interests than to become the guardian of the general good.

Varieties of Libertarianism

While libertarians share certain general commitments or attitudes, they often differ both in the specific content and in the justification of their philosophical positions. The post-World War II period has seen the development of numerous attempts to state the case for, and the basis of, libertarianism.

Nozick's Minimal State Libertarianism

The most important mainstream libertarian is Robert Nozick, whose 1974 book, *Anarchy, State and Utopia* (Nozick, 1974), has been widely discussed and criticized. Beginning with the presupposition that "Individuals have rights, and there are things no person or group may do to them (without violating their rights)"; and the conviction that "So strong and far-reaching are these rights that they raise the question of what, if anything, the state and its officials may do," Nozick set out to examine, from a libertarian point of view, three of the fundamental questions of political philosophy: Can the state be justified?; What principles of justice must govern such a state?; and What would a good society look like if it ran along such principles?

Nozick's first concern is the repudiation of anarchism. He is a minimal state libertarian and the task of the first part of his book is to show how such a state can, and must necessarily, arise out of a prepolitical state without violating individuals' rights. For Nozick, individuals have rights to their persons and their property; and these rights act as 'side-constraints' on the actions of others, forbidding such others to act in any way that violates rights – even if violating rights would improve the general welfare or, indeed, help reduce the number of rights violations. Rights are absolute; they are also fundamental: rights are good in themselves and are not derived from some other, more basic, good. For this reason, rights cannot be weighed in the balance with other, competing goods.

In these circumstances, how could a state ever be justified? Nozick tries to show that without rights being violated, a minimal state would emerge out of a state of nature in a four-stage process. First, individuals would form mutual protection associations in that condition, so a number of protection agencies would appear. Second, one agency or federation of agencies would come to dominate any given area – through its superior economic efficiency and without any violation of rights. Third, the dominant agency will rightly be entitled to protect its clients against independent agencies because individuals have the right to protect themselves. Finally, however, the dominant agency will have to extend its protection to those independents by way of compensation for the violation of their (and their clients') rights. Emerging out of all this is a minimal state – one which claims a monopoly on the use of force, while offering protection to all.

What remains doubtful in all this is whether such a state is indeed consistent with libertarian principles, since it must deny persons the right to enforce their rights against others and require them to pay taxes. And Nozick has duly been criticized for this, particularly by libertarian anarchists. No less controversial, however, is Nozick's insistence that only a minimal state is justifiable: it is no part of the state's right or duty to do more than offer protection. Nozick's theory of distributive justice explains why.

According to Nozick, justice is a matter of how people have come to have the goods they have rather than a matter of the pattern of distribution of people's holdings. Unlike the theory of John Rawls, Nozick's theory is historical rather than structural. There are three principles of justice in holdings: a principle of justice in the acquisition of property; a principle of justice in transfer; and a principle of justice in rectification of past injustice. According to this view, if property is justly acquired and justly transferred, the resulting distribution of property is, necessarily, just – however small or great the resulting differences of wealth or income among people.

What, however, would make an original appropriation of property just? Can the first person to come across unowned property simply acquire it all? In Nozick's theory, original acquisition is constrained by the 'Lockean proviso' that each is allowed to appropriate as much as he or she desires, provided

"enough and as good [is] left in common for others." The importance of this proviso lies not only in what it forbids by way of future appropriation, however, but also in its implications for all existing holdings. How a person may transfer property will also be governed by the Lockean proviso; and an earlier violation of the proviso may cast doubt on the worth of a particular property right.

Nozick's insistence on the importance of the proviso is consistent with his view of justice as historical rather than structural. His theory is as much an attack on patterned or end-state theories as a defense of his own 'entitlement' theory. End-state theories think of justice in terms of some particular distributive outcome, with which principles of distribution must conform. For Nozick, this is incompatible with a proper understanding of justice and of rights. But he has a further objection: the attempt to secure preferred distributive patterns necessarily requires interference with the voluntary activities of individuals. Liberty upsets patterns, so liberty will have to be compromised if patterns are to be preserved.

Nozick's theory of justice has been vigorously criticized, however, by those who reject his libertarian starting point as well as by others who want to resist his conclusions. The thesis of self-ownership and Nozick's account of its implications, have been subject to searching examination by G.A. Cohen, in particular.

Rand's Objectivism

While Nozick has dominated academic philosophy, however, no libertarian has reached a wider audience than the Russian-born American writer Ayn Rand. In a series of novels, most famously *The Fountainhead* and *Atlas Shrugged* (Rand, 1943, 1957), essays, and newsletter articles, she offered a vigorous defense of capitalism and articulated a comprehensive philosophy – 'Objectivism' – which claimed to set individual rights on a firm epistemological and metaphysical basis.

Capitalism, according to Rand, "is a social system based on the recognition of individual rights, including property rights, in which all property is privately owned." On this account, capitalism is understood, or at least defined, not as an economic system but as a social arrangement consistent with the principles of ethics. Capitalist society is one which respects individual rights. The question, however, is: What rights do individuals have? and Why do they have these (and not any other) rights? Rand defines the rights an individual has by saying that "for every individual, a right is the moral sanction of a *positive* – of his freedom to act on his own judgment, for his own goals, by his own *voluntary*, uncoerced choice. As to his neighbors, his rights impose no obligations on them except of a *negative* kind: to abstain from violating his rights." Rights are thus freedoms to act but not assurances of anything more than forbearance on the part of others. An important corollary of this, however, according to Rand, is that the individual has a right to acquire property, since human beings need property if they are to survive. The basis of individual rights, she argues, is the right to life. "The right to life is the source of all rights – and the right to property is their only implementation. Without property rights, no other rights are possible. Since man has to sustain his life by his own effort, the man who has no right to the product of his effort has no means to sustain his life." But none of this means that an individual is entitled to any particular share of property. All it means is that no one may rightly be prevented from seeking to acquire property noncoercively and that no one should be prevented from using as he or she wishes, the property he or she has rightly acquired.

Rand, in her extensive writings, attempted to defend this stance by showing that it followed from a proper understanding of humans' nature and a grasp of the value of life. This philosophy is called 'objectivism.' Beginning with assumptions about the existence of an "objective reality," the centralilty of life as a value, and the importance of reason for human survival, she claims to have shown that the only ethically defensible social system is capitalism. Even within libertarianism, these arguments have proven controversial. Nozick, for example, finds them unconvincing. But there has been no shortage of philosophers, such as Douglas Den Uyl and Douglas Rasmussen, for whom the construction and defense of a Randian position continues to be an important task.

Although Rand, unlike Nozick, receives little attention from mainstream academic philosophers, the extent of her influence and the interest of her thought should not be underestimated. Not only do her books continue to reach a large popular audience, making her one of the best-selling authors of the twentieth century but she has also exerted a significant influence on academic philosophers, economists, and policy-makers (the most famous being Alan Greenspan, Chairman of the Federal Reserve Bank and co-contributor to Rand's *Capitalism: The Unknown Ideal* (Rand, 1967)).

Murray Rothbard's Libertarian Anarchism

For Ayn Rand, a proper understanding of human rights could yield only a conception of government whose role was limited to the protection of individual rights. Even then, taxation would have to be voluntary; for government, as policeman, could not violate individuals' rights by taking by force what individuals were not prepared to give up. For many libertarians, however, individual rights are in no way compatible with the existence of the state. The most important libertarian defender of the anarchist position was Murray Rothbard.

Rothbard has been libertarianism's great polymath: an economist in the Austrian tradition, he was also a historian and social philosopher, as well as a pamphleteer of the highest order. His works ranged from a treatise on economics, *Man, Economy and State*, to a history of *America's Great Depression*, to a philosophical examination of *The Ethics of Liberty* (Rothbard, 1970, 1972, 1982).

According to Rothbard, the 'crucial axiom' of the libertarian creed is this: "no man or group of men have the right to aggress against the person or property of anyone else." Aggression is defined as "the initiation of the use or threat of physical violence against the person or property of someone else." Central to libertarianism, then, is "the absolute right to private property of every man: first, in his own body, and second, in the previously unused natural resources which he first transforms by his labor." In Rothbard's view, these two axioms, "the right of self-ownership and the right to 'homestead,' establishes the complete set of principles of the libertarian system."

In common with Nozick and Rand, Rothbard holds that individuals have rights to do as they wish, so long as they do not invade the rights of others. Thus, individuals should be free to use or dispose of their property as they see fit and live their lives – as well or as badly – as they wish, provided they do not violate anyone's rights. However, Rothbard also maintains that the existence of the state is incompatible with libertarianism, for the state, by definition, must infringe these basic rights. Nozick's 'immaculate conception' of the state, he argues, is unsuccessful both because the logic of its derivation is flawed and because, even if it could be shown that a state could arise without the violation of rights, this would not serve to legitimize that state (nor, *a fortiori*, any existing state). Nozick, however, has at least offered a philosophical argument attempting to rebut anarchism. Rand, on the other hand, according to Rothbard, does little more than evade the difficulty posed by her own libertarian principles, for her claim that there is a role for government and for her allegiance to the American state. Rand's patriotic defense of America during the Vietnam War was untenable for Rothbard, given the fact that military conscription – is an unambiguous violation of the libertarian principle of nonaggression and its prohibition of involuntary servitude.

For Rothbard, minimal state libertarianism must either abandon libertarianism's core commitment to nonaggression or lapse into incoherence. None of the extant nonanarchist forms of libertarianism have come to terms with this problem, since none have been able to explain how the most fundamental need of the state – to raise revenue by taxation – could be accounted defensible in libertarian terms. More than any other libertarian, Rothbard has tried to push the principles of libertarianism to their logical conclusion; to show that if individuals have rights, no form of authority which might invade their persons or take their property without their consent, is morally legitimate.

Hillel Steiner's Left-Wing Libertarianism

In rejecting as flawed the libertarianisms of Nozick and Rand, Rothbard assumes that individuals not only have property in their persons but also can acquire just title to property in the external world. Indeed, for Rothbard, the right to acquire property is not constrained by anything like a Lockean proviso that enough and as good be left over for others. Should all land, for example, be appropriated by those who come to it first, leaving nothing for the johnny-come-latelies, that is, simply, the way of the world. Thus he rejects the idea, associated in the nineteenth century with Henry George, that the earth itself cannot be owned, even if individuals can own the product of their labor. If the original land is nature- or God-given, he argues, then so are all the people's talents, health, and beauty; if land must be owned only by society, so must everything else. Rothbard therefore rejects any idea that we start by assuming that everyone has an equal claim to land or to the earth or its resources.

Yet this is itself a highly controversial issue. And some libertarians, notably Hillel Steiner in his *An Essay on Rights* (Steiner, 1994), have argued that the correct starting point is an assumption of an equal right of all to the earth and its resources. Persons have, on this account, two original rights: to self-ownership and to an equal share of initially unowned things. Each 'person' is entitled to ownership of himself, for to be owned by others is to be a slave; and so only 'things' may be owned by others. But no one is entitled to appropriate every unowned thing. Indeed, no one is entitled to appropriate more than the Lockean proviso allows. In a world of fully appropriated things, "each person's original right to an equal portion of initially unowned things amounts to a right to an equal share of their total *value*." Moreover, and correlatively, any person's possession of a just title to any thing, encumbers him or her with a duty to pay every person an equal share of its value. Since there is, Steiner argues, no moral right of inheritance, – for only existing (and not past or future persons) have rights – the upshot of all this is that a just distribution of things must reflect our original rights, and not the accidents of history or geography. Our just entitlements do not depend upon where, or to whom, we were born. The implication of this libertarian view is that the recognition of libertarian rights is fully consistent with, and indeed requires, a significant measure of property redistribution among persons, and that means all persons, regardless of nationality.

From such libertarian premises, some, like Philippe Van Parijs, who would not describe themselves as libertarians ideologically, go on to argue that justice requires recognizing that everyone is entitled to a form of unconditional basic income. As he points out in *Real Freedom for All*, the core problem for libertarians is to find a nonarbitrary criterion of original appropriation. Homesteading or an application of the principle of 'first come, first served,' is not, in this view morally defensible; for it would be a serious error to suggest that anyone could be free living under a regime in which he or she has no property rights, except in his or her person, since all property has already been appropriated and bequeathed. This left-wing libertarianism takes much more seriously the importance of equality as a moral consideration, just as the other variants of libertarianism reject its significance altogether as a substantive value.

Libertarianism as Ideology

Politically, libertarianism is often identified as a 'right-wing' ideology but, in fact, its position on the ideological spectrum is problematic. While some libertarians, like Rand, describe themselves as opponents of 'the left' or as right-wing, others, like Rothbard, see themselves as having much more in common with the left as opponents of the state – which they tend to see as the captive or instrument of big business. Moreover, most libertarians try to distinguish (if not dissociate) themselves from conservatives, since they reject the conservative's interest in the regulation of personal morality. To the extent that they align themselves with any other ideological viewpoint, libertarians are inclined to describe themselves as the heirs of classical liberalism. Libertarians as diverse as Hayek, Rothbard, and Steiner have been ready to describe themselves as classical liberals.

In public affairs, however, libertarianism has found itself allied, more often than not, with the conservative side of politics. Over the short history of modern libertarianism, some

conservatives have tried to transform it by incorporating its commitment to liberty but rejecting its hostility to authority and tradition. A notable advocate of such a 'fusionist' view was the American conservative Frank Meyer. To date, however, libertarianism has resisted such attempts. It remains a more radical political philosophy than conservatism can countenance.

Editor's Note

Popular and scholarly attention to libertarian political philosophies in the United States has increased since the original publication of this article, largely because of the Presidential candidacies of Ron Paul and his son Rand Paul. Recent works that will be of interest to readers include: David Boaz, 1997; Peter Vallentyne, 2012; Manuel Vargas, 2014; Serena Olsaretti, 2013; Eric Schansberg, 2002; Alfred R. Mele, 2013.

See also: Civil Liberties and Human Rights; Freedom: Political; Liberalism: Impact on Social Science; Republicanism: Philosophical Aspects.

Bibliography

Barry, N.P., 1986. On Classical Liberalism and Libertarianism. Macmillan, London.
Boaz, D. (Ed.), 1997a. The Libertarian Reader: Classic and Contemporary Writings from Lao-Tzu to Milton Friedman. Free Press, New York.
Boaz, D., 1997b. Libertarianism: A Primer. Free Press, New York.
Boaz, D., 1997. Libertarianism: A Primer. Free Press, New York. An updated edition was released in 2013.
Cohen, G.A., 1995. Self-ownership, Freedom and Equality. Cambridge University Press, Cambridge, UK.
Eric Schansberg, D., 2002. Common ground between the philosophies of Christianity and libertarianism. Journal of Markets & Morality 5 (2), 439–457.
Friedman, M., 1962. Capitalism and Freedom. University of Chicago Press, Chicago.
Hayek, F.A., 1960. The Constitution of Liberty. Routledge and Kegan Paul, London.
Hospers, J., 1971. Libertarianism. Nash, Los Angeles.
Mele, Alfred R., 2013. Philosophy and Phenomenological Research 87 (1), 72–92.
Narveson, J., 1988. The Libertarian Idea. Temple University Press, Philadelphia.
Nozick, R., 1974. Anarchy, State and Utopia. Blackwell, Oxford, UK.
Overton, R., 1646. An Arrow against All Tyrants. In: Aylmer, G.E. (Ed.), The Levellers in the English Revolution. Thames and Hudson, London, 1975.
Olsaretti, S., 2013. Rescuing justice and equality from libertarianism. Economics and Philosophy 29 (1), 43–63.
Rand, A., 1943. The Fountainhead. Bobbs-Merrill, New York.
Rand, A., 1957. Atlas Shrugged. Random House, New York.
Rand, A., 1967. Capitalism: The Unknown Ideal. Signet, New York.
Rothbard, M.N., 1970. Man, Economy, and State: A Treatise of Economic Principles. Nash, Los Angeles.
Rothbard, M.N., 1972. America's Great Depression. Sheed and Ward, Kansas City, KS.
Rothbard, M.N., 1978. For a New Liberty: The Libertarian Manifesto, second ed. Macmillan, New York.
Rothbard, M.N., 1982. The Ethics of Liberty. Humanities Press, Atlantic Highlands, NJ.
Steiner, H., 1994. An Essay on Rights. Blackwell, Oxford.
Vallentyne, P., 2012. Libertarianism and the justice of a basic income. Basic Income Studies 6 (2), 1–11.
Vargas, M., 2014. The Runeberg problem: theism, libertarianism, and motivated reasoning. In: Timpe, K., Speak, D. (Eds.), Libertarianism and Free Will: The Interplay of Religious Belief and Free Will. Oxford University Press, New York.
Wolff, J., Nozick, R., 1991. Property, Justice and the Minimal State. Polity Press, Cambridge, UK.

Libraries

Andrea J Copeland, School of Informatics and Computing, Indiana University, Indianapolis, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C. Pawley, volume 13, pp. 8807–8810, © 2001, Elsevier Ltd.

Abstract

There are many types of libraries, each with their purpose within the communities they serve. While differences define each type of library, commonalities exist among all. The various types of libraries are briefly described and then in greater detail defined in terms of their relationships with finances, information, individuals, communities, and place.

Introduction

Since their beginning, libraries have collected, organized, archived, and made accessible the intellectual output of humankind. Dating back to 2600 BCE in Sumer, the earliest known libraries were found to have clay tablets with cuneiform script documenting commercial inventory and transactions. Today's libraries contain a diversity of formats covering a wealth of subject areas. There are many types of libraries, each with their purpose within the communities they serve. While differences define each type of library, commonalities exist among all. The various types of libraries are briefly described and then in greater detail defined in terms of their relationships with finances, information, individuals, communities, and place.

Types of Libraries

Libraries are defined foremost in terms of the community served and also with regards to their physical structures, types of information stored or accessed, and funding sources.

Academic, university, or college libraries are designed to support the teaching and research needs of the faculty, staff, and students affiliated with academic institutions. Most research libraries are supported by universities, although some are public or governmental institutions. Research libraries are designed to be comprehensive and exhaustive in their coverage of subjects and acquisitions are permanent. Physical libraries are designed with particular subsets of a university's population in mind, such as undergraduate, law, medicine, and science. In the last 20 years, new acquisitions have trended toward a digital format, especially in the case of journals and more recently, books. As more and more university libraries move content online, the physical space of the library is being allocated for communal purposes, such as social and meeting spaces and computer workstations (Dewey, 2008). Public university libraries are largely funded through tax funds and student tuition. Private universities are supported by tuition and endowments. Academic libraries have faced budget difficulties owing to the rising cost of journal subscriptions and as a result many have saved money through collaboration and participation in open access publication initiatives (Bosch et al., 2011).

Corporate or special libraries are found in a variety of business settings: law firms, hospitals, news publications and broadcast companies, and pharmaceutical companies. These types of libraries serve specialized clientele and as a result have collections that have great depth but little breadth. Currently, for most special libraries, materials are acquired electronically. The library is considered a support service and as a result the physical space is small; often it is an office within a larger physical plant of the corporation. Librarians working in corporate environments often possess knowledge specific to the domain and work alongside other members of the organization. Often librarians will manage the company's intellectual capital and as a result fill knowledge management roles as well; principally, codifying knowledge to develop information retrieval systems and taxonomies specific to the corporate enterprise. Corporate libraries are supported by company profits and when companies experience loss, the library budget is likely to be one of the first cut as it is considered a support service (Corcoran, 2002). That having been said, studies from Kuwait, New Zealand, South Africa, and the United States indicate that jobs in corporate libraries and in knowledge management initiatives continue to grow.

Various levels and agencies of government have libraries. Many countries have national libraries, such as Nigeria, South Africa, Portugal, France, Germany, Scotland, Great Britain, China, and India, with the United States being an exception. National libraries serve the function of a depository of the national publications (both print and electronic), provision of reference, bibliography, preservation, and lending services; preservation and promotion of the national cultural heritage; and leadership in literacy campaigns (IFLA, 2010). Government agencies will support libraries to manage information produced by that agency as well as the information needed to support the research and information needs of the agency. For example, the National Aeronautics and Space Administration (NASA) and the European Space Agency (ESA) have libraries that support science and engineering research. Government information is moving online as well as government libraries. The Government Printing Office in the United States distributes more than 95% of its information in a digital format (Chadwick et al., 2012). Government libraries are funded by taxes and therefore must provide access to citizens as well as agency members. Their dependence on taxes for funding also means that their vitality waxes and wanes with the economy. For example, in the United States, the Environmental Protection Agency (EPA) closed all of their physical library locations in 2006 and moved access to EPA resources online in a money-saving effort. However, the outcry from their community of

users was so great that the physical libraries reopened within 2 years.

Public libraries are open to all members of a community. By design they are intended to serve the educational and entertainment needs of community members from cradle to grave. As a result, materials available through public libraries range from picture books to scholarly journals. The most heavily used public library collections today are comprised of children's materials and popular reading for adults. Most public library acquisitions are still in print format, but in recent years many public libraries have seen an increase in digital acquisitions, particularly electronic books. In many communities, public libraries are the only free source of access to the Internet and are instrumental in addressing problems associated with the digital divide (Jaeger et al., 2012). A public library is typically a system of one large central library and several smaller branch libraries serving neighborhoods within the system's service area. In general, libraries are funded in varying degrees by public and private sector funds. Public libraries are typically funded by tax dollars, although in certain countries small membership fees are also charged, such as the Netherlands, Alberta (Canada), Zimbabwe, and France. Funding for public libraries has been increasingly threatened in recent decades. The global economic crisis of 2007–11 brought closure to many public library branches (American Library Association, 2012). However, in many instances, public protest and willingness to fund local libraries through special tax initiatives kept doors from closing.

School libraries support the information needs of students and faculty of primary and secondary schools. A school library is typically one room within the school's facility. Materials are both print and digital with some schools opting for an all-digital environment. School libraries were early adopters of multimedia materials and as a result were initially called media centers. School libraries support the classroom curriculum and the leisure reading of students. As educators, school librarians collaborate with classroom teachers to support reading and development of information literacy skills of students. Funding for public school libraries is most often taxed based, and like public libraries, school libraries have experienced a reduction in services and resources provided in recent years, including the services of professional librarians. Studies have shown that students who have access to school librarians outperform those that do not on tests related to reading development.

Digital Libraries

The influence of digital information has changed the perception of libraries and libraries themselves. Many myths surround technology and information; for example, it is commonly thought that individuals no longer need to consult information intermediaries, travel to a library, or seek a publisher's permission to publish (Brown and Duguid, 2002). As the lines between authors, publishers, and distributors of information blur, there has been much speculation about libraries going the way of the travel agency and the newspaper. In many cases, technology has fundamentally changed the ways in which information is accessed and published; however, there are still for-profit industries thriving on the fact that not all digital information is free. Digital libraries are conduits to accessing both proprietary and freely available digital information.

A digital library has been defined in a diversity of ways depending on the perspective, user or designer. Digital libraries are purposefully constructed of electronic records which are collected and organized to meet the information needs of a specific community of users. Further, these libraries have technical and functional capabilities that support searching for, using, and creating information (Borgman, 1999). Digital libraries are international in scope and are usually created and maintained by university libraries and academic departments, public libraries, or government agencies (Witten, 2003).

Digital collections are often the electronic resources and services that exist alongside a physical library. Library users can use digital libraries and collections from their homes and offices. Technology allows library services and resources to be delivered seamlessly to individuals allowing for personalization of use within the user's own information environment. The distance between the individual and the physical library has been removed and access to information has been decentralized. The virtual or digital library becomes one information resource among many, whereas the physical library remains a beacon of knowledge within its community.

Finances

On the whole, libraries are nonprofit organizations and do not produce an information product but provide services that support access to information. However, libraries are in the position of doing business with organizations that are for-profit. Publishers, journal aggregators, book distributors, and authors need to turn profits in order to survive. The growth of the information economy and the influences of technology on information access have made the economic challenges faced by libraries even more acute. In the cycle of publication, libraries have traditionally filled the role of providing long-term access by preserving published works and making them intellectually accessible through organizational systems. The current system for funding libraries speaks to the good that society believes libraries possess and to the fact that individuals cannot produce the same level of organization and access to these materials less expensively on their own. Hence, the need for formal organization of information has economic justification. Information may be readily accessible and affordable to the individual but it is the long-term access and organization of that information that individuals cannot afford.

The decentralization of information has led some to believe that libraries' role in information access will lessen as information intermediaries are no longer necessary. However, it is unlikely that publishers or individuals will want to or are in the position to take on the roles of preservation and long-term access. The recent struggle over access to electronic books between public libraries and book publishers indicates that publishers may feel that a new paradigm is upon us (Besen and Kirby, 2012). Libraries have the right to loan materials indefinitely given the first-sale doctrine. This level of accessibility to readers reduces the level of profit for publishers and authors. The laws that applied to physical print-based materials are no longer applicable to electronic information. Individual and

public rights and privileges associated with electronic information are still being worked out in the courts (Electronic Frontier Foundation, 2012).

Further, libraries more often than not pay for licenses to access digital information rather than own it outright. Licensing agreements can inhibit lending practices and long-term access. Through collaboration and advances in the technical infrastructures of libraries, librarians have been successful in negotiating perpetual access and unhindered lending privileges to digital information (Lipinski, 2012). The library as both a learning and memory organization will keep libraries from simply becoming an office that manages licenses and contracts. And as learning and memory are fundamental to human existence, libraries will always be a part of the human experience.

Information

Libraries are defined, in part, in relationship to information. Information has been defined in a variety of ways. Information as process is the act of being informed. Information as knowledge is that which reduces uncertainty. Information as thing refers to objects such as documents and data (Buckland, 1991). Documents are stored or made accessible by libraries for the purpose of facilitating the exchange of information as knowledge between the author and the reader.

Libraries have evolved from a primary relationship of owning and storing information to one that now includes providing access to documents without necessarily either owning or storing the documents. The degree to which ownership is privileged over access varies mostly depending on type of library and the extent to which technology drives document formats.

Selection, organization, and preservation are three major functions that define a library's relationship to information as thing, particularly information that is owned by the library.

Based on the professional values of librarians, libraries provide diverse viewpoints on any topic included in a library's collection. Objectivity and neutrality in the development of collections supports the notion that libraries are a place for all members of a particular community. As mentioned earlier, libraries make selections based on the information needs of their community members. Selections are prioritized for acquisition based on those needs and budgetary constraints. Judgments are made as libraries cannot afford to purchase or physically hold every book relevant to its community's needs.

Traditionally, libraries filled information requests not met by their collections through interlibrary loan. Interlibrary loan is supported by member libraries that are willing to lend books, journal articles, and other media to other libraries in the network, who then in turn lend those items to their users. The process has become less labor and time intensive given the influence of the Internet. Requests for journal articles are filled within minutes or hours, while physical books travel more slowly in real time; however, even this is changing. An example would be HathiTrust, a collaborative large-scale digitization project resulting in a shared repository with nearly 6 million books accessible online. Digitized and networked information have created a level of cooperative collection-building previously not thought possible.

Organizational systems, cataloging, classifying, and indexing provide long-term intellectual access to materials that are owned by libraries (Svenonius, 2000). Traditionally, these systems have taken considerable time and effort on the part of librarians to implement; each document needing to be examined, interpreted, and described for cataloging, classifying, and indexing. These three systems along with bibliography allow for comprehensive access to a library's collection and the contents of that collection (Mann, 1994). These organizational systems are necessary, for without the means to provide long-term intellectual access there is little need to provide for long-term physical or digital preservation.

Organizational systems are changing to take advantage of enhanced access capabilities possible in a networked information environment. Recent developments to cataloging and cataloging records include the use of entity relationships, metadata, and user-generated subject tags. All of these changes allow for greater interoperability of cataloging records, liberating them from the confines of a single collection or institution (Gardner, 2012). The shift in importance to access of digital information is transforming organizational systems which traditionally provided static, structured, and pre-coordinated access to documents into a system that includes dynamic search processes. Keyword searching has become the norm foremost while the use of controlled vocabularies, like the Library of Congress Subject Headings, is left to the domain of information professions. The ephemeral nature of keyword searching is a conceptual match for the immediacy that networked access to digital information provides. However, for long-term access, formal systems of bibliographic control are essential.

As mentioned earlier, long-term intellectual access is the first step toward preservation. There is an understanding that libraries, like museums and archives, are memory organizations created in part to preserve the documentation that represents our collective past. There is the expectation of preservation of physical and digital information alike. The science of preserving physical materials is well established with film and sound recordings presenting the greatest challenges in analog preservation (Ross, 1993). Digital preservation is an evolving science with emphasis on collaboration, open source formats, technical metadata, redundant systems, and migration strategies (Caplan, 2008).

Libraries face additional challenges to preservation of digital information, especially that which is accessed rather than owned. This type of arrangement provides for use but not necessarily for long-term access. In addition to licensing access to content, libraries are also licensing the search interface, delivery platform, and the hosting space on a server. Libraries can negotiate perpetual access to digital information but a technical infrastructure of support must also exist in order to ingest that content into their collections. Smaller libraries with limited resources and staff expertise are likely not able to ingest large quantities of digital content.

Collaboration among libraries and other agencies concerned with long-term access to the cultural record is necessary for digital preservation. The Lots of Copies Keep Stuff Safe

(LOCKSS) Program and HathiTrust are examples of two successful collaborations. LOCKSS is designed to provide libraries and publishers with perpetual access to electronic subscriptions to journals and books. Supported by numerous universities, Google, the Internet Archive, and Microsoft, HathiTrust is a digital preservation repository comprised of works in the public domain as well as copyrighted works. While beneficial to researchers, the inclusion of copyrighted works resulted in legal challenges to HathiTrust. These collaborative efforts indicate the continued importance of libraries to provide information through ownership. The desire for long-term, equitable access to the cultural record will be realized through a preservation plan that is based on the cooperation of networks, libraries, publishers, and document creators. As a result, libraries will become increasingly social entities – entities that are as much defined by their communities of users as by their communities of collaborators.

Individuals

Libraries are also defined by their relationships with individuals. Libraries are sources of connection, education, entertainment, and inspiration for individuals. Libraries have moved from a static one-way relationship between document and user to a system of information processes that facilitate knowledge building. Libraries provide the opportunity for the transfer and exchange of information as process and as knowledge. Through libraries, individuals connect with librarians and other users and with ideas expressed in documents. Individuals can engage with processes that facilitate interacting and innovating through these connections.

Libraries educate individuals through access to librarians and documents. Libraries have always supported individual learning both formal and informal. The public library is often referred to as the university of the people as it offers endless resources for lifelong learning. Librarians assist individuals with information seeking and help them find the resources they need. These interactions occur in the library and online using e-mail or chat services and with varying intensity and duration depending on the nature of the information need. Increasingly, individuals turn to libraries and librarians to learn how to navigate the vast and rapidly growing information landscape in which libraries are only one resource for information among many. This is seen in an increase in information literacy skills instruction in all types of libraries (Breivik et al., 1998). Through this type of instruction, librarians help individuals develop the skills necessary to find, evaluate, and use information effectively.

Libraries have a special relationship with a particular sort of individual: authors. Libraries provide the resources to generate the creation of new knowledge and new works. Unlike other agencies involved with the production of published works, libraries are committed to providing long-term access to those works. In this way, they play a role in the life cycle of knowledge and the creation of new works. Authors look to the memories and cultural context preserved and organized by libraries to generate new ideas. Sir Isaac Newton once said that new insight is obtained by standing on the shoulders of giants; those giants are the authors whose works live on in libraries.

Community

The relationship between a library, regardless of type, and its community of users is paramount to understanding the nature of any particular library. The information needs of a library's community of users determine the resources and services offered by that library. The relationship between library and community is reciprocal. The library influences the lives of its community members. Until recently, historians mostly examined the influence of libraries on the individual, community, and society rather than the influences of individual readers on the development of the library (Wiegand, 2011). Libraries are now often assessed in terms of how well they are meeting their communities' needs through the evaluation of community demographics, usage statistics, and surveying users for feedback on satisfaction related to resources and services (Westbrook, 2001).

Libraries facilitate connections within communities among community members and organizations. These connections are believed to contribute to the development of social capital within a community. Social capital describes the networks and connections among individuals within a community that foster social trust and cooperation for the benefit of all (Putnam, 2000). Neighborhoods with a high level of social capital are more likely to have heavily utilized libraries; both entities thrive together.

Within a community, libraries can be an information node among many, such as bookstores, museums, archives, government agencies, and retail stores. Libraries are also considered a channel for information along with the Internet, family members, and experts. In their capacities as nodes and channels, they serve as distribution points for information. Access to government information provides a clear example of this, particularly in times of crisis. In recent history, following natural disasters in Chile, Haiti, Japan, New Zealand, and the United States, public libraries were significant in the recovery processes in the areas affected. Individuals sought out libraries to connect online with family members and to gain access to government information and other resources to aid in their personal recovery.

As mentioned previously, governments are moving much of their information to an online-only format. Access to this information is essential for government to function, yet government cannot require or expect individuals to afford the technology needed to access digital information. With the move to online distribution venues, public libraries and librarians have become indispensable as information intermediaries for a variety of government agencies. In many communities, the public library is the only source of free access to online information. While access to government information has been decentralized through the use of the Internet, which ultimately increases access, physical access to a device is still required. Digital information requires an infrastructure of support; the information must be rendered before it can be utilized. The necessary devices and networks must be located in close proximity to the citizenry. Public libraries provide a ready solution to this problem. Also, librarians are available to help individuals navigate the technical and intellectual aspects of finding and using government information.

Communities are documented and represented through library collections. In the largest sense, libraries document communities of scholars and authors. Libraries of all types represent their communities through the documentation and preservation of intellectual and cultural output. In academic libraries, institutional repositories are being built to collect and preserve the digital works of faculty and students (Lynch, 2003). Likewise, corporate and government libraries are building repositories to preserve and to provide access to the intellectual output of their constituents. These efforts benefit the institutions as a whole as these collections encourage collaboration among researchers and make accessible information for reuse. While public libraries are creating digital collections of unique materials, they have been less inclined to build community repositories to collect the intellectual and creative output of its community members. However, public libraries have the potential to play a significant role in documenting and preserving the social history of their communities (Copeland and Barreau, 2011).

Place

Libraries as place are often central to community life. In the words of historian Shelby Foote, "A university is just a group of buildings gathered around a library." Libraries are the intellectual and social centers of academic life. Community spaces and technologies that facilitate group work have been added to the physical structures of academic libraries. Public libraries have been referred to as third places, with first and second place being home and work, respectively. A third place is a neutral place where individuals from all backgrounds from a community meet to converse with neighbors in friendly manner – a place to be connected to others around ideas not related to work or family (Oldenburg, 1999). Libraries provide community members with a neutral yet vital environment to engage in a host of human activities.

Libraries are monuments. Many large cities around the world have libraries that are architectural feats and deserve admiration for their design alone, regardless of the collections contained within. Sendai Mediatheque, Japan; National Library, France; Villanueva Public Library, Colombia; and José Vasconcelos Library, Mexico, are all examples of buildings notable for their aesthetics. The grandeur of these libraries reflects their value to society and the value of what is contained therein. The buildings are as symbolic as they are functioning agencies of information. Libraries have their place among the cultural landscape along with museums, cathedrals, and theaters. The *Book of Kells*, an illustrated manuscript of the Gospels authored by Celtic Monks circa 800 CE, on permanent display at the Trinity College Library in Dublin, is often referred to as Ireland's finest national treasure. This one book attracts over a half million visitors each year. Libraries document the accomplishments of humankind, and by doing so celebrate us all.

As of 2002, approximately 92% of all new information created each year is stored on magnetic media (Lyman, 2003). The predominance of digital information has significant and long-lasting implications for libraries' collections and physical spaces. The digital information age has just begun and the long-term implications of this major shift in communication are yet to be fully understood. While organizing and preserving digital information for access and use is a cultural imperative, the need for libraries to maintain physical collections is just as important. Collectively, libraries have spent countless sums of money on physical materials documenting hundreds of years of human accomplishments. These materials are valuable as information and as artifact. In many instances, through digitization, libraries are creating digital collections of unique materials from their collections. Libraries are making content accessible that would otherwise not be available without physical use of a library. In this way, online access extends the presence of the physical library and integrates its resources with the greater web of knowledge.

See also: Museums; Publishing: Academic and University Presses.

Bibliography

American Library Association, 9 April 2012. 2012 State of America's Libraries Report. Retrieved from: http://www.ala.org/news/mediapresscenter/americaslibraries. Document ID: 4b062fa7-764d-f004-c17d-d2ecb3466c74.

Besen, S., Kirby, S., September 2012. E-Books and Libraries: An Economic Perspective. Report to the American Library Association. http://www.ala.org/offices/sites/ala.org.offices/files/content/oitp/publications/booksstudies/ebooks_libraries_economic_perspective.pdf.

Borgman, C., 1999. What are digital libraries? Competing visions. Information Processing and Management 37, 227–243.

Bosch, S., Henderson, K., Klusendorf, H., 2011. Under pressure, times are changing. Library Journal 136 (8), 30–34.

Breivik, P., Hancock, V., Senn, J.A., March 1998. A Progress Report on Information Literacy: An Update on the American Library Association Presidential Committee on Information Literacy: Final Report. Retrieved from: http://www.ala.org/acrl/publications/whitepapers/progressreport.

Brown, J.S., Duguid, P., 2002. The Social Life of Information. Harvard Business Review Press, Cambridge, MA.

Buckland, M., 1991. Information as thing. Journal for the American Society of Information Science 42 (5), 351–360.

Caplan, P., 2008. The preservation of digital materials. Library Technology Reports 44 (2).

Chadwick, C., Pilato, R., Conge, M., Rubin, R., Shaffer, G., 2012. The future of the FDLP in public libraries. Public Libraries 52 (4), 40–46.

Copeland, A., Barreau, D., 2011. Helping people to manage and share their digital information: a role for public libraries. Library Trends 59 (4), 637–649.

Corcoran, M., 2002. How to survive and thrive in the new economy: follow the money. Online 26 (3), 76–77.

Dewey, B., 2008. Social, intellectual, and cultural spaces: creating compelling library environments for the digital age. Journal of Library Administration 48 (1), 85–94.

Electronic Frontier Foundation, 2012. https://www.eff.org/.

Gardner, S., 2012. Cresting toward the sea change: literature review of cataloging an classification 2009–10. Library Resources & Technical Services 56 (2), 64–79.

International Federation of Library Associations and Institutions (IFLA), 22 November 2010. About the National Libraries Section. Retrieved from: http://www.ifla.org/about-the-national-libraries-section.

Jaeger, P.T., Bertot, J., Thompson, K.M., Katz, S.M., DeCoster, E.J., 2012. The intersection of public policy and public access: digital divides, digital literacy, digital inclusion, and public libraries. Public Library Quarterly 31 (1), 1–20.

Lipinski, T., 2012. The Librarian's Legal Companion for Licensing Information Resources and Services. Neal Schuman Publishers, Inc., New York.

Lyman, P., 2003. How Much Information? http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/.

Lynch, C., 2003. Institutional repositories: essential infrastructure for scholarship in the digital age. portal: Libraries and the Academy 3 (2), 327–336.

Mann, T., 1994. Library Research Models: A Guide to Classification, Cataloging, and Computers. Oxford University Press, New York.

Oldenburg, R., 1999. The Great Good Place: Cafes, Coffee Shops, Bookstores, Bars, Hair Salons, and Other Hangouts at the Heart of a Community. Marlowe & Company, Washington, DC.

Putnam, R., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon & Schuster, New York.

Ross, H., 1993. Preservation in Libraries: Principles, Strategies, and Practices for Librarians. Bowker-Saur, New York.

Svenonius, E., 2000. The Intellectual Foundation of Information Organization. The MIT Press, Cambridge, MA.

Westbrook, L., 2001. Identifying and Analyzing User Needs: A Complete Handbook and Ready-to-Use Assessment Workbook with Disk. Neal-Schuman Publishers, New York.

Wiegand, W., 2011. Main Street Public Library: Community Places and Reading Spaces in the Rural Heartland, 1876–1956. University of Iowa Press, Chicago.

Witten, I., 2003. Example of practical digital libraries. D-Lib Magazine 9 (3). http://www.dlib.org/dlib/march03/witten/03witten.html.

Life Course and Event History Analysis

Hans-Peter Blossfeld, European University Institute, Fiesole (FI), Italy
Gwendolin J Blossfeld, Nuffield College, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Life course analysis has become a very active research field. This article summarizes the main research questions of the life course paradigm and its most important theoretical concepts. It also reports some recent research findings of life course studies. Then, this article introduces the terminology and models of discrete and continuous event history analysis. A concrete event history analysis example is given in order to demonstrate the strength of life course studies.

Introduction

Life course analysis has become a very active research field in the social sciences these last decades for good reasons. Large retrospective life history and prospective panel data sets are now available in most modern societies. Well-known examples are the National Longitudinal Surveys (NLS66, NLS79, NLS97, NLSW) or the Panel Study of Income Dynamics (PSID) in the United States, the Life History Study, the Socio-Economic Panel Study (SOEP), or the National Educational Panel Study (NEPS) in Germany, the National Cohort Studies (NSHD, NCDS, BCS70, MCS), the British Household Panel Study (BHPS), and its successor, the study Understanding Society, in the United Kingdom, or the Household, Income and Labour Dynamics in Australia (HILDA) Survey. Most of these data sets are nationally representative for the population or some selected birth cohorts. They provide longitudinal data on intergenerational relationships, educational histories, job trajectories, family careers, fertility events, changes in health status in the life course, and often also data on the life courses of other family and household members such as partners or spouses. Life course data explicitly recognize the dynamic nature of social roles and circumstances as men and women move through their life paths, the interdependence of lives and life choices, the situational imperatives confronting actors in various phases of the life course, and the accumulation of advantages and disadvantages experienced by the individual over the life course (Elder et al., 2003; Crosnoe and Elder, this volume). The dynamic analysis of life course data is a rewarding but formidable challenge. It has been made possible by the development of statistical methods such as event history and panel analysis. These methods do not only allow studying changes over time but also offer better opportunities for a more appropriate time-related causal analysis (Blossfeld et al., 2007).

This article focuses on the close interrelationship of life course and event history analysis. Event history analysis is the statistical study of processes that are characterized in the following general way: (1) there is a population of individuals, each moving among a finite (usually small) number of life course states; (2) these changes (or events) may occur at any point in time; and (3) there are time-constant and/or time-varying factors influencing the timing of these events. Examples are workers who move between unemployment and employment; men and women who enter into consensual unions or marriages; people who are mobile between different regions or nation-states, and so on. This article first summarizes the main research questions of the life course paradigm and its most important theoretical concepts. It will also report some recent research findings of life course studies. Then, this article introduces the terminology and models of discrete and continuous event history analysis. Finally, a concrete event history analysis example is given in order to demonstrate the strength of life course studies.

Life Course Analysis

In the last four decades, there have been great advances in the study of lives over time, and they extend across various disciplines such as sociology, developmental psychology, demography, or medical sciences (Carr, 2008). We limit ourselves to the literature in the social sciences (Elder and Giele, 2009). Here we identify at least five typical research questions of life course researchers.

Studying the Endogenous Logic of Trajectories in the Life Course

A first aim of life course researchers is to study the endogenous (causal) logic of trajectories over the life course. In sociology, the concept of trajectory is often used to refer to an individual's sequence of roles in one domain of life such as the education, work, or family career. The conceptual tool of trajectory encompasses both sequences of different qualitative states (e.g., being single, married, or divorced in a family career) and continuous increases or decreases in quantitative characteristics (e.g., such as income or competence trajectories). In this article, we limit our discussion to life course studies that record and analyze the sequence of qualitative states occupied by individuals over the life course and the timing of changes among these states. For example, an event history of job mobility consists of more or less detailed information about each of the jobs and the exact beginning and ending dates of each job.

Data on trajectories are often collected retrospectively in so-called life history studies. Collecting retrospective data is cheaper than panel studies and has the advantage that they code the life course data into one framework of codes and meaning. But retrospective studies also suffer from several

limitations (see Blossfeld et al., 2007). In particular, data concerning motivational, attitudinal, cognitive, or affective states are difficult (or even impossible) to collect retrospectively because the respondents can hardly recall the timing of changes in these states accurately. Also the retrospective collection of behavioral data has a high potential for bias because of its reliance on autobiographic memory. To reduce these kind of methodological problems, modern panel studies (e.g., the PSID in the United States, the BHPS and the Understanding Society study in the United Kingdom, or the SOEP and the NEPS in Germany) are using a mixed data collection design which combines traditional panel information (i.e., multiple observations on the same individuals at discrete panel waves) with retrospectively collected event history data for the period before the first panel wave and between the successive panel waves. This combined design also allows linking information about an individual's prospective life planning (e.g., career or family planning) collected in earlier panel waves, with life course behavior recorded in later panel waves.

Trajectories of individuals often depend strongly on the structural context and thus draw theoretical attention to societal traditions and norms, social institutions, and organizational structures shaping the careers of individuals. For example, institutions such as the educational system define the times of possible school transitions in the life course and structure the school options available to individuals at certain points in their careers. For example, Shavit and Blossfeld (1993) have found declining effects of family background on successive educational transitions over the educational career for a broad selection of modern societies. It is, however, still an open question, whether the empirical pattern of declining social background effects across successive educational school transitions is (1) a life course effect because the students get increasingly independent from their families of origin, or (2) a selection effect since at each transition only the most competent students from lower social classes make it to the next level while also mediocre students from the upper social classes enter the higher level (Cameron and Heckman, 1998). Thus, as long as it is not possible to control for the unobserved competence development of individuals in educational transition studies, this issue cannot be decided.

Life course researchers have been particularly interested to study whether and to which extent events and states at earlier stages of a trajectory have consequences for its later outcomes (Mayer and Tuma, 1990). Dannefer (1987) introduced the so-called Matthew effect into the literature on the life course. The Matthew effect means that small initial inequalities in a trajectory become magnified over the life span. For example, in modern societies there seems to be a general logic in educational careers that the 'already educated get even more education in the next step' while the 'poorly educated at the same time get less,' so that the difference in education between the two groups gradually increases. The Matthew effect is sometimes also referred to as the cumulative disadvantage/advantage hypothesis (O'Rand and Henretta, 1999). It offers a cumulative explanation of how intracohort inequality is increased over the life course in many life domains. The life course literature also suggests a 'status maintenance' and a 'status leveling' hypothesis. The first one contends that initial inequalities are carried along as individuals move through their life course. The second one points to the possibility of a narrowing of the inequality gap (O'Rand and Henretta, 1999).

Studying the Timing of Transitions in the Life Course

Transitions in the life course often represent discrete changes in states, for example starting a new job or getting married. The timing of these transitions in the life course is often characterized by inertia effects. Inertia is the increasing resistance of an individual to change its position with increasing duration in a state (duration dependence). For example, it is well-known that with increasing job-specific experience the job mobility rate of an individual is generally declining or with increasing time of living in a particular place, the rate of spatial mobility is generally decreasing. The mechanisms responsible for this life course patterns are often cumulative investments into job- or firm-specific human capital or into local 'social capital' (one's social ties) which make it increasingly costly for the individual to become a mover. Sometimes the time dependence in a state has a nonmonotonic shape. For example, it is well-known that the rate of entry into first motherhood has a nonmonotonic pattern across age in modern societies. As women's age increases, the rate of entry into motherhood initially rises, reaches a peak, and then decreases. This bell-shaped rate is often explained by two competing mechanisms: (1) the increasing readiness of individuals to enter into motherhood and (2) the probability of the individual to meet a single partner of the opposite sex. At younger ages, there are many unmarried singles on the marriage market in the relevant age range. Then, some of these young men and women marry as they are getting older. With increasing age, however, it is also increasingly difficult to meet an unmarried individual of the opposite sex at the relevant age band. Thus, the marriage rate is at first increasing with an increasing readiness, reaching a peak, and then decreasing with an increasingly difficult marriage market (Blossfeld and Huinink, 1991).

Studying the Effects of Early and Late Life Course Transitions

Life transitions often are associated with societal age norms that create expectations when a life course transition ought to occur in an individual's life (Carr, 2008). They provide a kind of social clock or timetable for important life events (such as the age of completing education, entering the labor market, or having a first child, etc.). With regard to certain life transitions, individuals might be 'early,' 'on-time,' or 'late.' Age norms, connected with social rewards and punishments, exert pressure on individuals to hasten or delay a certain transition and thus encourage age conformity of life course transitions.

Studying the Sequence of Multiple Life Course Transitions

Life course studies indicate that normative sequences exist with regard to various life course transitions. For example, Hogan (1981) argued that as a result of general societal norms, individuals move from school to work and then to marriage as well as parenthood. In order to satisfy this sequencing norm, young adults will be less likely to enter

parenthood before they have finished schooling. In addition, individuals enrolled in education may consider themselves not economically ready for marriage and motherhood. Thus, the completion of education is expected to count as an important prerequisite for entering into parenthood. Indeed, many life course studies document that women's increasing delays of entry into first marriage and motherhood are to a large extent connected with their longer educational participation (Blossfeld and Huinink, 1991).

Studying the Effects of Parallel and Interdependent Processes on the Life Course

From a causal-analytical point of view, the dynamic study of parallel and interdependent processes is one of the most important advances of life course research (Blossfeld et al., 2007). These parallel and interdependent processes can operate (1) at the level of individual's different domains of life (e.g., one may ask how an individual's job career influences his/her family trajectory; see Blossfeld et al., 2005); (2) there may be parallel processes at the level of some few individuals interacting with each other (e.g., one might study the effect of husband's career events on his wife's labor force participation); (3) there may be parallel processes at the intermediate level (e.g., one can analyze how organizational growth of a firm influences career advancement or how changing household structure determines women's labor force participation; see Blossfeld and Hofmeister, 2006); (4) there may be parallel processes at the macro level (e.g., one may be interested in the effect of changes in the business cycle on family formation or individual career advancement); and (5) there may be any combination of such processes of type (1)–(4). For example, in the study of life course, cohort, and period effects, processes at different levels can be measured and included simultaneously in a multilevel analysis (Blossfeld, 1986).

Interdependent Trajectories at the Level of the Individual

At the level of the individual, events in one trajectory are often presumed to influence the transitions in other domains of life and vice versa (e.g., Blossfeld et al., 2007). For example, the economic theory of the family expects that women's rising educational attainment levels and better career resources affect their rate of entry into first marriage and first motherhood (opportunity cost hypothesis). However, life course analyses demonstrated that changes in the level of educational attainment or changes in career resources had little or no effect on women's marital timing. The higher-educated women only marry later because they stay longer in school and therefore delay entry into marriage (sequence norm). Yet, there is an effect of women's educational and career investments on entry into first birth. Because women, and not men, still take primary responsibility for child care in most modern societies. Women therefore work less or even interrupt their careers when a child is born. Therefore, women who have accumulated a high stock of human capital in their life course tend to gradually postpone or even avoid the birth of a first child. This effect is especially pronounced in traditional family systems, where insufficient child care facilities are available, such as those in Southern Europe (Blossfeld, 1995).

Interdependencies in the Process of Spouse Selection (Educational Homogamy)

In most modern societies the participation in higher education has increased dramatically in recent decades – and more so for women than for men. Life course researchers therefore have been especially interested to explore the role of the educational system as a marriage market and to analyze the changes in educational assortative mating induced by educational expansion (Blossfeld, 2009). In their longitudinal analyses they reconstructed the processes of spouse selection in the life courses of single men and women step by step and compared the results across birth cohorts and various countries (Blossfeld and Timm, 2003). These life course studies tried to take into account the dynamic opportunity structures to meet potential partners in the various parts of the educational system. One finding of these longitudinal analyses has been that with increasing duration in the educational system the rate of homogamous marriage is indeed rising. Increasing rates of educational homogamy across cohorts, however, means that social networks are closing to outsiders and that the opportunities of the generation of children are gradually segmenting.

Interdependencies of the Careers of Spouses

Life course researchers have been particularly interested to study linked lives (Elder et al., 2003). For example, they analyzed the effect of husbands' careers on their wives' labor force participation. A comparative analysis of spouses in 13 countries provided evidence that despite substantial improvements in wives' educational attainment and career opportunities, the change in gender specialization within dual-earner couples has been rather modest (Blossfeld and Drobnič, 2001). Gender role change within the family to a large extent has been asymmetric almost everywhere: Housework and child care primarily remained 'women's work,' and husband's participation in housework and child care has not increased substantially (Baxter et al., 2013). The result is often a double burden on women (Evans and Baxter, 2013).

Regardless of wife's job status, husbands in general pursue a life-long continuous employment career. Even in dual-earner marriages, men still tend to define themselves as primary breadwinners, even in cases where their female partners have a similar or even higher occupational status (Blossfeld and Drobnič, 2001). Thus, the career patterns of (married) men are to a large extent independent of career patterns of their partners. Such independence generally does not apply when (married) women's employment trajectories are analyzed in dependence of their husbands. Male partners tend to exhibit a significant influence on their female partners' career patterns in many countries (Blossfeld and Drobnič, 2001). For example, Germany stands out as a country with a particularly traditional division of labor in married couples. In this country, the analysis shows that high career resources of the husband can even suppress wife's career potentials and drive her out of the labor market.

Interdependencies at an Intermediate Level

At the intermediate level, life course researchers might examine how the changing household or organizational structure of a firm determines women's labor force participation in different countries. For example, in modern societies there is

very often a complex association between the age of the youngest child and female labor force participation (Blossfeld and Hakim, 1997; Blossfeld and Drobnič, 2001). The common interpretation is that the care of young children tends to keep mothers at home in conservative welfare states. However, it is also plausible that the lack of jobs encourages women to enter into a stay-at-home marriage. Event history analysis is able to shed light into this complex interdependence.

Dependencies of Life Courses on Processes at the Macro and Global Level

At the macro level of society, life course researchers have been interested, for instance, in the effect of changes in the business cycle on family formation (e.g., Blossfeld and Huinink, 1991). In these studies often effects at different levels are jointly examined. For instance, in the study of life course, cohort, and period effects, processes at the individual level (accumulation of career resources) have been jointly analyzed with two kinds of processes at the macro level: (1) variations in economic entry conditions specific for certain (birth, marriage, etc.) cohorts (cohort effects); and, (2) changes in particular historical conditions affecting all cohorts in the same way (period effects) (Blossfeld, 1986).

More recent studies have also demonstrated that similar processes at the global level have diverse effects on individual life courses in various countries (Mayer, 1997). In all modern countries, the effect of globalization is 'filtered' by domestic institutions and shaping the life course transitions in country-specific ways (Blossfeld, 2009). The globalization impact is most important at the transition from youth to adulthood (e.g., at entry into the labor market and the phase of family formation, Blossfeld et al., 2005). In particular the growing economic and temporal uncertainties lead young people to postpone or even forgo family formation (Blossfeld et al., 2005). Established men in their midlife career, however, are remarkably unaffected by globalization (Blossfeld et al., 2006a). The kind of labor market participation and the different welfare regimes are very important, if the impact of globalization on women's midlife transitions between employment and family are considered in modern societies (Blossfeld and Hofmeister, 2006). Finally, the effects of globalization on late job careers of men and women and their transition into retirement in modern societies are shaped by different pension systems (Blossfeld et al., 2006b).

Event History Analysis

Event history analysis provides an appropriate statistical framework for investigating the various dimensions of life course processes just described.

Basic Terminology

Event history analysis studies transitions across a set of discrete life course states, including the length of time intervals between entry to and exit from these states. The basic analytical framework is a state space and a time axis. The choice of the time axis or clock (e.g., age, experience, marriage duration, etc.) used in the life course analysis must be based on theoretical considerations. An episode, spell, waiting time, or duration–terms that are used interchangeably in the literature–is the time span an individual spends in a specific state. The states are discrete and usually small in number (e.g., single, married, divorced, remarried). The definition of a set of possible states, called the state space, is also dependent on substantive considerations. Thus, a careful, theoretically driven choice of the time axis and design of state space are crucial for a successful life course analysis. These issues are often serious sources of misspecification (Blossfeld et al., 2007).

The most restricted event history model is based on a process with only a single episode and two states (one origin and one destination state) (see Blossfeld et al., 2007: 38pp.). An example may be the duration (or episode) of first marriage until the end of the marriage (for whatever reason). In this case, each individual who entered into first marriage (origin state) started an episode, which could be terminated by a transition to the destination state 'not married anymore.' If more than one destination state exists, these models are called multistate models. They are also referred to as models with competing events or competing risks. For example, the first marriage might be terminated by the event 'death' or the event 'divorce.' If there are repeated events over the life course, these models are called multiepisode models. For example, if we analyze not only first marriages but all marriages of individuals over the life course at once. The individual then moves repeatedly between different states. Thus, in event history analysis, we have often a sample of $i = 1,\ldots,N$ multistate-multiepisode data. A complete description of the data is given by

$$(u_i, m_i, o_i, d_i, s_i, t_i, x_i) \;\; i = 1, \ldots, N$$

where u_i is the identification number of the individual; m_i is the serial number of the episode; o_i is the origin state, the state held during the episode until the ending time of the episode; d_i is the destination state defined as the state reached at the ending time of the episode; s_i and t_i are the starting and ending times, respectively. In addition, there is a covariate vector x_i with time-constant and/or time-changing factors associated with the episode. We always assume that the starting and ending times are coded such that the difference $t_i - s_i$ is the duration of the episode, which is positive and greater than zero.

Censoring

Observations of event histories are often censored. Censoring occurs when the information about the duration in the origin state is incompletely recorded. If the length of time an individual has already spent in the origin state is unknown, the episode is censored on the left. This kind of censoring should be avoided because it is not easy to handle in methodological terms (see Blossfeld et al., 2007: 39pp.). Most life course studies are designed in a way that the processes are observed right from the beginning. Right-censoring on the other hand, is very common. This type of censoring typically occurs in life course studies at the time of the retrospective interview, or in panel studies at the time of the last panel wave. For example, if one is interested in entry into first motherhood, women who are still in their fertile years and have not had a first baby until the interview, have a right-censored episode. This means that

these women might get a first baby later, but we do not know. Because the timing of the interview is normally independent of the timing of the substantive processes under study, this type of right censoring is unproblematic. It can easily be handled with event history methods.

The Dependent Variable

Event history models can be formulated in continuous-time (Blossfeld et al., 2007) or discrete-time (Mills, 2011; Yamaguchi, 1991; Vermunt, 1997). For each point in time (continuous-time models) or for each time interval (discrete-time models), they predict future levels or changes of the transition rate of the dependent process on the basis of states and/or events of other processes in the past. The central concept of event history analysis is the transition rate. Because of the various origins of event history analysis in the different disciplines, the transition rate is also called the hazard rate, intensity rate, failure rate, transition intensity, risk function, or mortality rate. The transition rate describes in detail how the dependent process evolves over time.

If time can be considered to be (approximately) continuous, i.e., when events at least in principle can happen at any point in time, the transition rate $r(t)$ can be interpreted as the propensity (or intensity) to change from an origin state to a destination state, at time t (Blossfeld et al., 2007: p. 36):

$$r(t) = \lim_{t' \to t} \Pr(t \leq T < t' | T \geq t)/(t' - t) \text{ with } t < t'$$

It is important to note that the propensity $r(t)$ is defined in relation to a risk set ($T \geq t$) at t, i.e., the set of units that still can experience the event because they have not yet had the event before t. In other words, the transition rate is a conditional density function, that is, the density function $f(t)$ divided by the survivor function $G(t)$

$$r(t) = f(t)/G(t)$$

where $f(t) = \lim_{t' \to t} \Pr(t \leq T < t')/(t' - t)$ with $t < t'$ and $G(t) = \Pr(T \geq t)$.

If events can only happen within fixed time-intervals, the continuous time axis is arbitrarily split into a series of time intervals $\tau_0 < \tau_1 < \tau_2 < \ldots < \tau_q$, with $\tau_0 = 0$. The number of the time interval then becomes a discrete random variable $T^* = t \Leftrightarrow T^* \in (\tau_{t-1}, \tau_t)$, with $t = 1, \ldots, q$. T^* denotes the time interval, where an event happens. The discrete-time transition rate is then a (conditional) probability,

$$r^*(t) = \Pr(T^* = t | T^* \geq t) \text{ with } 0 \leq r^*(t) \leq 1$$

$r^*(t)$ is the probability that the dependent variable changes from an origin state to a destination state in time interval t under the condition that the event did not yet happen until the beginning of that interval. In the discrete-time model the transition rate $r^*(t)$ is therefore a conditional probability function, where the probability function $f^*(t)$ is divided by the survivor function $G^*(t)$

$$r^*(t) = f^*(t)/G^*(t)$$

where

$$f^*(t) = \Pr(T^* = t)$$

and

$$G^*(t) = \Pr(T^* \geq t)$$

For small time intervals (e.g., when life histories are recorded on the basis of months) continuous-time models can be approximated by discrete-time models (Blossfeld et al., 2007: p. 37). The estimation results of equivalent time-continuous and time-discrete event history models $\Pr(T^* = t)$ then yield very similar coefficients.

Statistical Models

The central idea in event history analysis is to make the continuous-time transition rate $r(t)$ or the discrete-time transition rate $r^*(t)$ dependent on concepts of time (e.g., duration t) and on a set of time-constant x or time-dependent $x(t)$ covariates:

$$r(t) = g[t, x, x(t)] \text{ or } r^*(t) = g[t, x, x(t)]$$

The causal interpretation of the transition rate requires that we take the temporal order in which the processes evolve very seriously. In the continuous-time mode, at any given point in time t, the transition rate $r(t)$ can be made dependent on conditions that happened to occur in the past (i.e., before t), but not on what is the case at t or in the future after t. Equivalently, in the discrete-time model, in any given time interval t, the discrete-time transition rate $r^*(t)$ can be made dependent on conditions that happened to occur before the beginning of interval t, but not on what is the case in time interval t or after time interval t.

There are several possibilities to specify the functional relationship $g(.)$ between covariates and the transition rate in continuous-time event history analysis (see Blossfeld et al., 2007). First, the 'exponential model,' which normally serves as a kind of baseline model, assumes that the transition rate can vary with different constellations of time-constant covariates, so that the rates are time-constant, too (see Blossfeld et al., 2007: 87pp.)

$$r(t) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n) = \exp(X\beta)$$

where β_0 is a constant, X is the row vector of covariates, and β is a corresponding column vector of coefficients.

In most life course applications of transition rate models, the assumption that the forces of change are constant over time is, however, not justified. This is particularly true, if we are interested in causal relationships which relate changes (or events) in some (explaining) processes to changes (or events) in a dependent life course process. In our view, the most important step forward in event history analysis has been to explicitly measure and include time-dependent covariates in transition rate models. In such cases, covariates can change their values over process time. Time-dependent covariates can be qualitative or quantitative, and may stay constant for finite periods of time or change continuously (see Blossfeld et al., 2007: p. 128). Time-dependent covariates can be included (1) by using a piecewise constant exponential model, (2) by applying the method of episode splitting in parametric or semiparametric transition rate models, and (3) by specifying the distributional form of the time-dependence and directly estimating its parameters using maximum likelihood method.

A first simple and very useful generalization of the 'exponential model' is the so-called 'piecewise constant exponential model.' It allows the transition rate to vary across fixed time periods with period-constant or period-specific effects of covariates (see Blossfeld et al., 2007: 116pp.). The piecewise constant model is particularly helpful when researchers are not in a position to measure and include important causal factors (time-dependent covariates) explicitly or when they do not have a clear idea about the form of the time-dependence of the process, after controlling for important covariates. The basic idea of the piecewise constant exponential model is to split the duration into L time periods.

$$I_t = \{t|\tau_1 < t \leq \tau_{l+1}\} \quad 1 = 1, \ldots, L$$

based on arbitrary split points on the time axis $0 = \tau_1 < \tau_2 < \ldots < \tau_L$, with $\tau_L = \infty$.

The transition rate can then vary over the $l = 1, \ldots, L$ time periods based on the changing period-specific constants β_l (baseline hazard rate)

$$r(t) = \exp\{\beta_1 + X\beta\} \quad \text{if } t \,\varepsilon\, I_1$$

where X is a (row) vector of covariates, and β is an associated vector of coefficients. Note that in this model there is no additional constant in X.

Since causal relationships relate changes (or events) in explaining processes to changes (or events) in a dependent life course process, it is very important to be able to represent the changes of the causal forces in event history models directly. An easy way to do this is to include time-varying covariates $x(t)$ in the 'exponential model' via the method of episode splitting.

$$\begin{aligned} r(t) &= \exp\big[\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \beta_{n+1} x_{n+1}(t) \\ &\quad + \beta_{n+2} x_{n+2}(t) + \ldots + \beta_{n+p} x_{n+p}\big] \\ &= \exp[X(t)\beta] \end{aligned} \qquad [1]$$

where β_1 is the regression constant ($x_1 = 1$). The transformation of the person-spell file into a person-period file through episode splitting is described in detail in Blossfeld et al. (2007: 135pp.).

Another model is the so-called Cox (1972) model, where the baseline hazard rate $h(t)$ is left unspecified. Thus, it is only possible to estimate the effects of the time-constant or/and time-varying covariates, controlling for an unknown baseline hazard rate. This model is also called semiparametric model, since only part of the rate function is parametrically specified, or partial likelihood model, because only part of the likelihood function is maximized (see Blossfeld et al., 2007: 216pp.),

$$\begin{aligned} r(t) &= h(t)\exp\big[\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \beta_{n+1} x_{n+1}(t) \\ &\quad + \beta_{n+2} x_{n+2}(t) + \ldots + \beta_{n+p} x_{n+p}(t)\big] \\ &= h(t)\exp\big[X(t)\beta\big] \end{aligned}$$

Note that in this model there is again no regression constant in $X(t)$.

However, in some applications in life course research, substantive theory or previous empirical research may suggest a specific shape of time or duration dependence of the transition rate. However, time itself is no causal factor. Rather measures of time may serve as proxies for time-changing causal factors that could not be observed directly (see Blossfeld et al., 2007: 182pp.). For example, duration might serve as a proxy for 'the changing amount of marriage-specific investments' in divorce studies. There are different parametric models which are based on specific shapes of the time-dependence. For example, the Gompertz(-Makeham) and Weibull models are able to specify monotonically increasing or monotonically decreasing shapes of the hazard rate. Further parametric models such as the Sickle, the Log-logistic, and the Log-normal models allow to estimate a (at first increasing and then decreasing) nonmonotonic time-dependency (see for more details Blossfeld et al., 2007: 204pp.).

Finally, an important problem of event history analysis is also the issue of unobserved heterogeneity. In this case, the transition rate that is estimated for a population can be the result (a mixture) of quite different transition rates in the subpopulations (see Blossfeld et al., 2007: 247pp.). There have been several proposals to deal with unobserved heterogeneity in time-continuous transition rate models. The basic idea is to incorporate an 'error term' into the model specification. For example, the continuous-time transition rate can be made dependent on an exponential model, the observed (time-dependent) covariates $X(t)$ and a stochastic error term ν, which is e.g., Gamma distributed (see Blossfeld et al., 2007: 256pp.)

$$r(t|\nu) = \exp[X(t)\beta]\nu \quad \text{with } \nu \geq 0 \qquad [2]$$

A likelihood ratio test can then be used to test, whether the transition rate models with and without this unobserved heterogeneity term differ.

In the case of discrete-time models, the estimation of covariate parameters is normally achieved by preparing a period-person data file and then estimating a simple logit model:

$$r^*(t) = \frac{\exp(\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n)\exp[\beta(t)]}{1 + \exp(\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n)\exp[\beta(t)]} \qquad [3]$$

or

$$r^*(t) = \frac{ab(t)}{1 + ab(t)} \qquad \begin{aligned} &\text{with } a = \exp(\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n) \\ &\text{and } b(t) = \exp[\beta(t)] \end{aligned}$$

If $b(t) = \exp(\beta_0)$, the logit model estimates an exponential model, if $b(t) = \exp(\beta_0 + \beta_1 t)$ it estimates a Gompertz model, if $b(t) = \exp(\beta_0 + \beta_1 \ln t)$ it estimates a Weibull model. A piecewise constant model is estimated for $b(t) = \exp(\beta 1)$ if $t \,\varepsilon\, I_1$ where changing period-specific constants $\beta 1$ can vary over the l = 1,…, L time periods.

Application Example

In order to demonstrate the strength of life course analysis with event history models in a substantive context, we estimate the rate of dissolution of first marriage for East and West German women. Using a subsample of retrospective life histories of 3957 married women born in East and West Germany from the NEPS, we estimate a period-specific discrete-time event history logit model with time-constant and time-varying covariates. The NEPS collects dates of transitions and events on a monthly basis. For our analysis, we define for each woman a spell starting at entry into first marriage and ending at the time of its dissolution. For women, who do not experience a marital dissolution until the time of the interview, we censor the spells

on the right. The spells are also right-censored for women whose husband passed away. This leads to a person-oriented spell data set, where each married woman in the sample has exactly one record of data. We have then transformed this person-level data set into a person-period data set using the method of episode splitting (Blossfeld et al., 2007), in which each woman has multiple records – one for each month. Such a person-period data set allows an easy integration of time-varying covariates because the covariates can change their values across each of the person-month records.

In our longitudinal analysis, we are using the following covariates:

1. *Duration of marriage* (time-dependent covariate). We include dummy variables distinguishing eight periods of marriage duration;
2. *Place of birth* (time-constant covariate). We use the dummy variable 'East Germany' to identify women born in East Germany, women born in West Germany are the reference category;
3. *Historical periods* (time-dependent covariates). To distinguish the periods before and after unification in our analysis, we use the dummy variable 'Period after 1990.' The reference category is the 'Period until 1990.' We include also an interaction dummy variable 'East*Period after 1990' to estimate the effect of the 'German unification' on East German women;
4. *Husband's educational attainment* (time-constant covariate). In order to model husband's highest educational attainment, we distinguish seven educational degrees. We then attach the average number of years that are necessary to achieve these degrees;
5. *Wife's educational attainment* (time-dependent covariate). Each time when a woman attains a higher level of education, the educational attainment level is adjusted. We distinguish seven educational attainment levels and express each degree as the average number of years necessary to achieve it;
6. *Partner's relative levels of education* (time-dependent covariates). In order to model the relative resources of partners, we introduce two dummy variables indicating whether a woman has a higher or lower level of education than her partner (ref. category: both partners have the same educational attainment level);
7. *Premarital states* (time-constant covariates). We introduce three dummy variables to control if women cohabited with their husband at the time of marriage (ref. category: women who did not cohabit with their husband at the time of marriage), were pregnant at the time of marriage (ref. category: women who were not pregnant at the time of marriage), or have had a child before marriage (ref. category: women who had no child before marriage);
8. *Age at first marriage* (time-constant covariates). To include the age at first marriage into our analysis we distinguish five age categories;
9. *Age of the youngest child* (time-dependent covariate). We include the age of the youngest child by distinguishing seven time-dependent dummy variables.

Table 1 Estimation results of a discrete-time event history model for the dissolution of women's first marriage in Germany

Covariate	*Coefficient*	
Duration (in years)		
0–1	−9.286	c
2–3	−7.328	c
4–5	−6.957	c
6–7	−7.149	c
8–10	−7.349	c
11–15	−7.700	c
16–20	−7.675	c
20+	−7.950	c
Place of birth		
East Germany	0.214	
Historical Period		
Period after 1990	−0.142	
East*Period after 1990	−0.562	b
Education		
Husband's education	−0.095	c
Wife's education	0.096	c
Partner's relative levels of education		
Woman higher	0.378	c
Both equal (ref.)		
Woman lower	0.268	a
Premarital states		
Cohabitation at time of marriage	0.511	c
Pregnancy at the time of marriage	−0.052	
Birth of a child before marriage	0.237	a
Age at first marriage		
Under 20 (ref.)		
20–22	−0.250	a
23–25	−0.484	c
26–28	−0.771	c
>28	−0.800	c
Age youngest child (years)		
No child (ref.)		
<1	−0.222	a
1–2	0.259	
3–5	0.701	c
6–10	0.645	c
11–14	0.806	c
>14	0.516	b
Log likelihood	−5902.277	
Number of events	742	
Number of subepisodes	959 259	
Chi^2	36140.81	
Degrees of freedom	28	

[a] $p < 0.05$.
[b] $p < 0.01$.
[c] $p < 0.001$.
$n = 3957$.
Source: Own computations based on the NEPS data.

Table 1 shows the results of our discrete event history analysis of women's dissolution of first marriage in Germany including time-varying and time-constant variables. In order to test whether the duration of the first marriage has an effect on divorce, we include information on the duration of first marriage into our analysis. The results show a nonmonotonic bell-shaped pattern of the duration of marriage on divorce. In other words, the rate of dissolution increases up to the fifth year and then decreases.

To analyze the divorce rate in East and West Germany before and after German unification, we include dummy variables to control the place of birth (East vs West Germany), historical periods (before and after German unification), and

an interaction term 'East*Period after 1990.' The results show that there is a declining divorce rate in East Germany after the German unification because of the high level of uncertainty in the transition process from socialism to capitalism.

Table 1 also shows that a higher educational attainment level of the husband stabilizes the marriage, while a higher educational attainment level of the wife has a destabilizing effect. If we compare the relative resources of husband and wife, we find that only homogamous marriages are more stable.

If we include information on premarital cohabitation, childbirth before marriage and pregnancy at the time of marriage into our model, our results show that premarital cohabitation and childbirth before marriage have a destabilizing effect on marriages. Being pregnant at the time of marriage has no effect. We also include information on the age at first marriage. The result shows that marriages are more stable the older the woman is at entry into first marriage. Finally, we include information whether a couple has children and the age of the youngest child into our example analysis. Couples who have a baby younger than 1 year have the lowest divorce rate. The divorce rate then increases with the age of the youngest child and rises above the divorce rate of couples without children. If we control for unobserved heterogeneity, the results of this analysis do not change.

See also: Age Stratification; Age, Sociology of; Aging, Theories of; Assortative Mating in the Marriage Market; Capitalism: Global; Cohabitation: United States; Demographic Models; Demographic Techniques: LEXIS Diagram; Divorce, Sociology of; Family Theory: Economics of Childbearing; Family Theory: Economics of Marriage and Divorce; Fertility Change: Quantum and Tempo; Fertility Theory; Fertility of Single and Cohabiting Women; Labor, Division of; Life Course: Sociological Aspects; Life Table; Lifespan Development, Theory of; Lifespan Development: Evolutionary Perspectives; Linked Lives; Marriage and the Dual-Career Family; Multistate Transition Models in Demography; Period and Cohort Analysis in Demography; Retirement and Encore Adulthood: The New Later Life Course; Social Stratification; Transition to Adulthood.

Bibliography

Baxter, J., Hewitt, B., Haynes, M., Western, M., 2013. Pathways through the life course: the effect of relationship and parenthood transition on domestic labour. In: Evans, A., Baxter, J. (Eds.), Negotiating the Life Course. Stability and Change in Life Pathways. Springer, Dordrecht, pp. 145–159.

Blossfeld, H.-P., 2009. Educational assortative marriage in comparative perspective. Annual Review of Sociology 35, 513–530.

Blossfeld, H.-P., 1995. The New Role of Women. Family Formation in Modern Societies. Westview Press, Boulder, CO.

Blossfeld, H.-P., 1986. Career opportunities in the Federal Republic of Germany: a dynamic approach to the study of life course, cohort, and period effects. European Sociological Review 2, 208–225.

Blossfeld, H.-P., Buchholz, S., Hofäcker, D., 2006b. Globalization, Uncertainty and Late Careers in Society. Routledge, London and New York.

Blossfeld, H.-P., Drobnič, S., 2001. Careers of Couples in Contemporary Societies. From Male Breadwinner to Dual-Earner Families. Oxford University Press, Oxford.

Blossfeld, H.-P., Golsch, K., Rohwer, G., 2007. Event History Analysis with Stata. Lawrence Erlbaum, Mawah, NJ.

Blossfeld, H.-P., Hakim, C., 1997. Between Equalization and Marginalization. Women Working Part-Time in Europe and the United States of America. Oxford University Press, New York and Oxford.

Buchholz, S., Hofäcker, D., Mills, M., Blossfeld, H.-P., Kurz, K., Hofmeister, H., 2009. Life courses in the globalization process: the development of social inequalities in modern societies. A summary of findings from the GLOBALIFE project. European Sociological Review 25, 53–72.

Blossfeld, H.-P., Hofmeister, H., 2006. Globalization, Uncertainty and Women's Careers in International Comparison. Edward Elgar, Cheltenham, UK and Northampton, MA, USA.

Blossfeld, H.-P., Huinink, J., 1991. Human capital investments or norms of role transition? How women's schooling and career affect the process of family formation. American Journal of Sociology 97, 143–168.

Blossfeld, H.-P., Mills, M., Klijzing, E., Kurz, K., 2005. Globalization, Uncertainty and Youth in Society. Routledge, London and New York.

Blossfeld, H.-P., Mills, M., Bernardi, F., 2006a. Globalization, Uncertainty and Men's Careers in International Comparison. Edward Elgar, Cheltenham, UK and Northampton, MA,USA.

Blossfeld, H.-P., Timm, A., 2003. Who Marries Whom? Educational Systems as Marriage Markets in Modern Societies. Kluwer Academic Publishers, Dordrecht. A Comparison of Thirteen Countries, European Studies of Population.

Cameron, S.V., Heckman, J.J., 1998. Life cycle schooling and dynamic selection bias. models and evidence for five cohorts of American males. Journal of Political Economy 106, 262–333.

Carr, D., 2008. Encyclopedia of the Life Course and Human Development, 3 vols. Gale, Detroit.

Cox, D.R., 1972. Partial likelihood. Biometrika 62, 269–276.

Dannefer, D., 1987. Aging as intercohort differentiation. Accentuation, the Matthew Effect, and the life course. Sociological Forum 2, 211–236.

Elder, G.H., Kirkpatrick, J.M., Crosnoe, R., 2003. The emergence and development of life course theory. In: Mortimer, J.T., Shanahan, M.J. (Eds.), Handbook of the Life Course. Springer, New York, pp. 3–19.

Elder, G.H., Giele, J.Z., 2009. Life course studies: an evolving field. In: Elder, G.H., Giele, J.Z. (Eds.), The Craft of Life Course Research. The Guilford Press, New York, pp. 1–24.

Evans, A., Baxter, J., 2013. Negotiating the Life Course. Stability and Change in Life Pathways. Springer, Dordrecht.

Hogan, D.P., 1981. Transitions and Social Change: The Early Lives of American Men. Academic Press, New York.

Mayer, K.U., 1997. Notes on a comparative political economy of life courses. Comparative Social Research 16, 203–226.

Mayer, K.U., Tuma, N.B., 1990. Event History Analysis in Life Course Research. University of Wisconsin Press, Madison.

Mills, M., 2011. Introducing Survival and Event History Analysis. Sage, London.

O'Rand, A.M., Henretta, J.C., 1999. Age and Inequality: Diverse Pathways through Later Life. Westview Press, Boulder, CO.

Shavit, Y., Blossfeld, H.-P., 1993. Persistent Inequality. Changing Educational Attainment in Thirteen Countries. Westview Press, Boulder, CO.

Vermunt, J.K., 1997. Log-linear Models for Event Histories. Sage, Newbury Park, CA.

Yamaguchi, K., 1991. Event History Analysis. Sage, Newbury Park, CA.

Life Course Criminology

Marvin D Krohn, Molly Buchanan, and Andrea Davis, University of Florida, Gainesville, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The life-course perspective on crime recognizes it is essential that we view the causes and consequences of criminal behavior throughout individuals' developmental stages. In this article, we review the general conceptual foundation of the life-course perspective. We illustrate the use of this perspective by briefly reviewing four specific theories. Analytical tools to examine intraindividual trajectories are also reviewed. We conclude by suggesting future conceptual and methodological directions to advance the life-course perspective.

Theories of crime were primarily developed to explain the onset, prevalence, and frequency of adolescent delinquency. To an extent, this made pragmatic sense because official data revealed that, for the most part, the onset of arrests did not occur until the early adolescent years and reached the peak in prevalence at ages 15 through 17 years. However, the recognition that there was a small group of 'career offenders' who disproportionally contributed to the prevalence of serious crime and were likely to continue their offending into the adult years broadened our theoretical and research foci to incorporate what happened with these offenders as they transitioned from adolescence to young adulthood. Very quickly, we came to the conclusion that those who disproportionally contributed to the crime problem by committing frequent and serious criminal behaviors and had criminal careers of longer duration than less serious offenders were also precocious in their criminality, beginning their offending at younger ages. Hence, it became critical that we extend our focus to the childhood years. To do so would require a conceptual framework that addressed the full span of a person's life. Such a vantage point would allow us to not only assess individuals' criminal histories but also explore other aspects of individuals' lives and assess how those aspects both impacted and were impacted by their criminal behavior. That conceptual framework has come to be known as the life-course approach.

The life-course approach provides a framework within which we can examine how people develop through the various stages of their lives. In doing so, the focus is on both stability and changes that occur in behavior and attitudes, the causes and consequences of that stability or change, and the interrelationships among different trajectories or pathways that characterize the life course. As such, rather than focusing on interindividual differences to discern patterns that may reflect the causes of behavior, the life-course approach emphasizes intraindividual stability and change to determine how both selection and socialization factors impact the life course. Within this broad conceptual approach, separate theoretical statements have been offered. While differing on some specifics of what they posit as important in describing and understanding human development, all of these theories are commonly concerned with the interweaving of trajectories representing different arenas of one's life and the important changes or transitions that occur during the life course.

In this article, we examine the conceptual bases of a life-course approach. We select four theoretical perspectives to review as illustrations of how life-course concepts have been used to account for the causes and consequences of crime. We also examine some of the unique analytical tools that have been developed and used to examine intraindividual change. This article concludes with suggestions for areas for future development.

The Conceptual Basis of the Life-Course Approach

The life-course approach to the study of human behavior conceptualizes the life course as a sequence of age-graded stages. The course through which a person traverses these stages within a domain of behavior is referred to as a trajectory (Elder, 1985). People traverse multiple trajectories such as educational, family, and work trajectories. In addition, behaviors such as criminal involvement can also be seen as trajectories. Importantly, the life-course perspective emphasizes the interconnectedness among trajectories. That is, the perspective is particularly interested in how the movement along one trajectory affects the path of another trajectory.

As people move along a particular trajectory, they will make transitions. A transition is a life event that is embedded within a trajectory and often results in "changes in state that are more or less abrupt" (Elder, 1985: pp. 31–32). For example, graduation from high school can be considered a transition in one's educational trajectory that may not only influence the course of that trajectory but may also influence what happens in the person's work life and family relationships. Related to the concept of transitions is the term 'turning points.' Turning points refer to particularly significant changes in the direction of one's life course (Elder, 1985). Getting married is a transition that has been conceptualized as a potential turning point in the criminal behavior trajectory (Sampson and Laub, 1993). That is, marriage will often deflect a person who is on trajectory of continuing crime away from problematic behavior and more toward conforming behavior.

The timing and sequential order of transitions can be consequential. There are cultural expectations for the appropriate timing of transitions, which, when violated, may cause problems for the person. We are expected to stay in school at

least through high school graduation. Quitting school prior to that time may have serious ramifications for the life course. Other off-time transitions such as having a child during one's teenage years often impact one's educational and work-related trajectories (Krohn et al., 1997).

The above review of key concepts in the life-course approach provides a foundation for further exploring the theoretical perspectives directed at explaining criminal behavior from a life-course perspective and the methodologies that have been developed to analyze these issues.

Life-Course Theories of Crime

The key concepts and general orientation of the life-course approach have been incorporated into several theories focusing specifically on the causes and consequences of criminal behavior (see Farrington, 2005). In this section, four theories that have had an important impact on the field are briefly described. Any explanation of what transpires across the life course is typically going to be complex; therefore, these theories cannot be presented fully in this brief article. Rather our emphasis is on what distinguishes one theory from the others.

Moffitt's Dual Taxonomy Theory

Drawing from concepts in both developmental and neuropsychology, Terrie Moffitt introduced a dual taxonomic theory to explain the age–crime curve (Moffitt, 1993). This developmental theory asserted that indications of persistent antisocial behavior could be detected during childhood. What distinguishes this theory from other general or developmental models is that it proposed that not all offenders are the same and not all antisocial behaviors are caused by the same explanatory factors. More specifically, the theory identifies two unique 'offending groups,' or trajectories, that can be differentiated based on two factors: the age when conduct problems begin and the trajectories of those problems (Moffitt, 1993). The two offending groups are referred to as 'early starters,' or life-course persistent offenders (LCPs), and 'late starters,' or adolescence-limited offenders (ALs).

There are significant differences between LCPs and ALs. Within the late-starter group, antisocial behavior typically begins and ends during adolescence (hence the group name of ALs); thus, their delinquency career is brief (Moffitt, 2003). Moreover, adolescence-limited offending is said to be both situational and temporary. Moffitt suggests that a maturity gap – the time period after puberty when biological maturation has occurred, but youths do not yet have access to 'adult' privileges and resources – might contribute to the normalization of antisocial behavior. During this maturity gap, ALs may mimic the behaviors of LCPs in an attempt to show independence from parents or accelerate social maturity (Moffitt, 2003). But, because their early developmental years were prosocial and they do not have the neuropsychological deficits characteristic of LCPs, most ALs desist from crime and adopt a prosocial lifestyle upon entering young adulthood. By young adulthood, individuals typically have gained more independence from parents, leaving little reason for them to further mimic LCP behaviors.

Contrary to ALs who desist from delinquent behaviors as they gain independence, LCPs are characterized by continuity of antisocial behavior. That is, their delinquency careers begin in early childhood, persist into adulthood, and are stable over the life course. According to Moffitt, LCPs often have neuropsychological deficits that impede their verbal development. In addition, LCPs may have early emotional and behavioral problems such as bad temperament or hyperactivity (Moffitt, 1993). These early deficits may adversely influence parenting quality, as dealing with such children can often be difficult. Moreover, the inheritance of many of these antisocial qualities has been hypothesized, further suggesting that biological parents of LCPs may be poor candidates for effectively handling and coping with their children's problem behaviors. Moffitt uses the term 'cumulative continuity' to suggest that early problems in the family can lead to continuing problems in school and other arenas in which LCPs interact. In addition, 'contemporary continuity' or the effects of deficits continue to be felt throughout different development stages. These high-risk neuropsychological and environmental interactions during the formative first two decades of life eventually develop into disordered personalities characterized by antisocial behaviors that carry on into midlife (Devers, 2011).

Sampson and Laub's Age-Graded Theory

Growing out of Hirschi's (1969) bonding theory, Sampson and Laub (1990) introduced the age-graded theory of informal social control to developmental criminology. This theory aimed to move away from what the authors believed were limitations of prior life-course research, including the adolescence focus and viewing desistance in a reductionist 'criminal to noncriminal' framework. Instead, Sampson and Laub viewed crime from childhood to adulthood, using individuals' differences and social ties to explain their initiation and desistence from criminal behavior. Although they recognized continuity in deviant behavior, they theorized that high-quality social ties in adulthood could deflect criminal behavior trajectories toward conforming behaviors in adulthood (Sampson and Laub, 1990).

The age-graded model of informal social control predicts that individuals with more social capital, high-quality marital bonds, and secure employment are more likely to desist from criminal behavior. Thus, the basic premise of this model is that, in spite of the continuity of criminal behavior, turning points, such as getting married, can change an individual's criminal trajectory and result in desistance from crime (Laub and Sampson, 1993). This age-graded model recognizes the influence of childhood behaviors while also highlighting the importance of major life transitions, or turning points, in explaining criminality over the life span.

Sampson and Laub (1992) delineated different stages of the life course that help provide context for their theory. From early childhood up to age of 18 years, structural problems such as low family socioeconomic status or parental divorce, in addition to individual differences in temperament, are predicted to influence whether juveniles become delinquent and how they

form bonds in adulthood. During adolescence specifically, unhealthy relationships with parents, attachment to delinquent peers, and low levels of attachment to school are all expected to lead to juvenile delinquency. Ages 18 through 25 yeas represent the stage at which deviant activity can greatly threaten the development of informal social bonds, with possible incarceration and its subsequent stigma, making attachment to prosocial bonds even more difficult in adulthood. Therefore, career criminality may develop from ages 25–45 years due to individuals' failure to reestablish prosocial attachments. In contrast, individuals who successfully develop prosocial attachments, whether through stable employment, family bonds, or high-quality marriages, will no longer commit crime because of these informal mechanisms of social control (Devers, 2011).

Thornberry and Krohn's Interactional Theory

Terence Thornberry (1987) was one of the first criminologists to emphasize the importance of assessing how potential predictors of delinquency and crime might operate differently at different stages of the life course. He began by presenting three models depicting the causal processes of crime during early, middle, and late adolescence, respectfully. He elaborated on a basic social control model (Hirschi, 1969) by incorporating concepts from both social learning theory and social structural perspectives. In doing so, he emphasized three aspects of his theory: (1) the need to examine how the variables included in interactional theory may vary by stage of development, (2) the incorporation of the construct of structural disadvantage as antecedent to the elements of the social bond, and (3) the recognition that some of the variables may be reciprocally related to one another.

In part, because of his continuing work on the Rochester Youth Development Study (RYDS), a longitudinal panel analysis of high-risk adolescent youths, which eventually developed into both a follow-up into the adult years and a study of the children of the original target subjects, Thornberry further developed his theory (Thornberry and Krohn, 2001, 2005), transforming it into an explanation that covered the life course.

In addition to the three building blocks of interactional theory listed above, Thornberry and Krohn (2005) added the premise of proportionality of causal influences. In recognizing that delinquency and crime had multiple causes and that not all causes had to occur in order for a particular outcome to result, Thornberry and Krohn focused on the magnitude of a causal force as being critical for producing an outcome. The concept of proportionality would then suggest that "as the magnitude of the causal force increases, the person's involvement in crime a) becomes more likely and b) increases in severity." The concept of proportionality has implications for how their perspective differs from Moffitt's dual taxonomy theory. According to Thornberry and Krohn (2005), there are not two types of offenders, but rather the range in the magnitude of causal influences results in some offenders beginning at an earlier age, committing more serious and frequent crime, and having a longer career than other offenders.

Taking into consideration how multiple causal factors play potentially different roles at different stages of development results in a somewhat complex theory that can account for the onset, duration, and seriousness of criminal behaviors. It is not possible to adequately describe the full range of the theory in this article, so we will concentrate on the onset of delinquency in order to illustrate some of the dimensions of the theory.

Thornberry and Krohn (2005) hypothesize that the early onset of antisocial factors is a result of individuals' characteristics, problematic parenting, and positions in the social structure. The early manifestation of antisocial behaviors during toddlerhood and preschool years only results if the interaction of these three areas is of great magnitude (proportionality). The uncoupling of these three deficits weakens the causal mechanisms and is likely to delay the onset of problematic behavior until the social networks widen when the child goes to school. Now whatever deficits the child has are more likely to emerge, resulting in the onset of antisocial behavior. School and peer factors become increasingly important as the child ages. Early participation in antisocial behavior, of course, impacts performance in school and the choice of friends as well (bidirectional causality). Youths who have strong parental and school support will delay the onset of any delinquent behaviors until the adolescent years. Then the quest for independence, a necessary prerequisite for the successful transition into adulthood, may cause stress in the lives of some youths, creating more reliance on peers and, at least for some, engagement in delinquent behavior.

Thornberry and Krohn (2005) go on to account for both the continuity in delinquent behavior and the potential for change in those behaviors. The importance of the magnitude of deficits and the impact of prior behaviors on subsequent life chances and continuing criminal behavior are emphasized. For example, in describing continuity of criminal behavior among those who had an early onset, Thornberry and Krohn (2005: p. 198) state: "In brief, we hypothesize that individuals who initiate antisocial behavior at very young ages are more likely than average to persist because the causal factors are likely to remain in place and because early involvement in antisocial behavior generates cumulative and cascading consequences in the person's life course." For those whose deficits are uncoupled, change toward more conforming behaviors is more likely.

Farrington's Integrated Cognitive Antisocial Potential Theory

David Farrington's (2003) research using the Cambridge Study in Delinquent Development identified numerous risk factors of delinquent behavior and the consequences of such behavior throughout the life course of the study's respondents. Based on the voluminous findings from that study, Farrington developed (1992) and then revised (2005) his integrated cognitive antisocial potential (ICAP) theory.

The key construct in his theory is antisocial potential (AP). AP is the potential individuals have to commit antisocial behaviors. Farrington further distinguishes between long-term AP and short-term AP. Long-term AP is due to some combination of impulsiveness, strain, socialization processes, and life

events, while short-term AP depends on situational factors (Farrington, 2005: p. 76).

Farrington (2005) also identifies a set of risk factors that predicts the levels of both long-term and short-term AP. The factors suggested to predict long-term AP include those typically identified in developmental life-course theories related to social structural position, parenting, socialization, life events, and individual factors. Short-term AP is dependent on emotive states (bored, angry), the influence of male peers, and opportunities to commit crime.

The commission of a crime is dependent on the coming together of long-term and/or short-term AP with cognitive processes. Cognitive processes include a constellation of thoughts a person might have regarding the costs and rewards of behaviors and the probabilities of achieving desired outcomes. In turn, the benefits of committing a behavior are somewhat dependent on anticipated approval from parents, peers, and partners. Farrington recognizes that the different combinations of long-term AP, short-term AP, and various cognitive processes may differ for different types of people but does not speculate on those dynamics.

The ICAP theory is a compilation of Farrington's research findings that focused primarily on establishing the risk factors for delinquent and criminal behavior. The causal processes that may explain why those factors present risk for criminal behavior have not been presented in his theory. Further development of those interconnections may provide a more thorough explanation of crime across the life course.

Methodological Advances and the Life-Course Approach

As outlined above, a main premise of the life-course approach is age-graded stages of development across various life domains. These stages can be further conceptualized as trajectories or long-term developmental patterns throughout the life course (Elder, 1975). A compendium of longitudinal data sets has emerged in recent years along with subsequent advances in methodological techniques. Researchers now have the enhanced ability to statistically 'look back' and simultaneously model respondents' developmental trajectories while considering the impact of theoretically relevant covariates. The following section describes a few of these methodological advances, focusing especially on semiparametric group-based trajectory modeling (GBTM) (Nagin, 2005).

Historical Background

During the criminal–career–age–crime curve debates of the 1980s, Blumstein et al. (1988) contended that offenders' patterns of criminal continuity and discontinuity were not static but, rather, were dynamic and varied over time. As such, these scholars posited that the classic, aggregate age–crime curve masked important differences in the prevalence and incidence of criminal involvement. This position was strongly debated by Gottfredson and Hirschi (1988), who argued in support of the static nature of offending behavior. Moreover, as proponents of the classic age–crime curve, Gottfredson and Hirschi (1986) posited that the pattern revealed at the aggregate level would be similar to patterns revealed from alternative individual-level analyses. Said differently, the authors argued that *all* individuals followed the rise, peak, and fall pattern seen in the aggregate. However, the significance of individuals' different levels of crime participation and frequency of criminal involvement became increasingly recognized with groundbreaking longitudinal analyses using data from the Philadelphia birth cohort study (see Wolfgang et al., 1972). These findings revealed that, over time, a *small* group of offenders was responsible for the *most* serious offenses. More importantly, these longitudinal results revealed patterns and facets of criminal careers not previously apparent from cross-sectional analyses.

The life-course approach has generated an increasingly accessible database and methodological techniques with which to examine theoretical implications. We are no longer limited to identifying and interpreting correlations between behaviors and later outcomes. Rather, exploration into *how* behaviors are connected throughout the life course and their influence on the continuity or change of antisocial behaviors can now be more accurately assessed.

Methodological Techniques Popular to the Life-Course Approach

Increased use of longitudinal designs and analyses has resulted in more accurate estimation of individuals' behavioral patterns (Sampson and Laub, 1992). Constrained to the statistical techniques available, early longitudinal researchers ran the risk of potentially 'underfitting' complex data sets and leaving certain developmental patterns undiscovered. However, recent advances in the methodological tools available allow researchers to better capture the connectedness of behaviors over time.

Three specific methods have been chosen for comparison in this article: hierarchical linear models (HLM), latent growth curve models (LGCM), and GBTM. All the three methods have been used to achieve a "common analytical objective" (Nagin and Odgers, 2010b: p. 114) of identifying shapes of developmental trajectories as functions of age or time and assessing the variability of identified trajectories within a population (Nagin, 1999; Nagin and Odgers, 2010b). Deciding which technique is best suited for analyses is typically guided by the theoretical framework, the research question at hand, and, especially, the assumptions made by each of the three methods (Curran and Bollen, 2001). This section gives an overview of important differences between these three models, ultimately providing a more thorough review of GBTM and its advantages for developing and modifying life-course theories.

Oft mentioned as alternatives to GBTM (Nagin and Tremblay, 2005b), HLM or LGCM offer multilevel approaches for modeling repeated measures to assess individuals' change over time. The two methods allow models of individuals' unobserved (latent) patterns, or trajectories, to be estimated using either manifest, observed measures in HLM (Raudenbush et al., 2004) or latent constructs, as is the case

with LGCM (Loehlin, 2004). Likewise, both methods assess differences in within-person, or intraindividual change over time, with outcomes revealing mean developmental patterns within a population as well as individuals' variability from these average patterns (Preacher et al., 2008).

HLM analyses assess variables across different levels. Initially, unconditional models are estimated to separate the fixed and random effects and provide both intra- and interindividual variance estimates. Level 1 variables address within-person measures that can vary over time, while Level 2 measures reflect independent variables that can be used to predict or explain variability in within-person changes between individuals (Raudenbush and Bryk, 2002). The unconditional model estimates are used to calculate intraclass correlations (ICCs). The ICCs are interpreted as differences attributable to the Level 2, *inter*(between)-individual measures. Further analyses include regression of Level 1 factors onto Level 2 variables. Such analyses result in cross-level interactions between the levels. These interaction terms are used to evaluate differences in individuals' developmental trajectories as a function of their Level 2 covariates. In sum, HLM allow for assessing the influence of between-person differences on within-person change over time (Raudenbush and Bryk, 2002).

Despite many similarities to HLM, LGCM focus on latent, unobserved variables, or growth factors, that may shape and influence the population's mean and variances for an outcome (Bollen and Curran, 2006). Like in HLM, initially estimated unconditional models provide the shape of developmental trajectories using the population mean growth and different rates of change. Instead of estimating variables by levels, LGCM measure growth factors at multiple times then relate those factors to other influential covariates (Bollen and Curran, 2006). Overall, LGCM have been touted as being more flexible than HLM, specifically better handling measurement error, reducing bias, and detecting main effects (Bollen and Curran, 2006; Muthen and Asparouhov, 2011). On the other hand, HLM have been found to provide more accurate estimates of cross-level interactions.

While basic HLM or LGCM share many similarities, more distinct differences exist between them and GBTM models. These differences are especially important when the empirical goal is identifying distinct, unique developmental patterns within a population. First, both HLM and LGCM assume that developmental patterns are distributed continuously. Such a distribution suggests that all individuals within a population move along the same, general development pathways, with their individual trajectories increasing or decreasing over time but following the same general shape or pattern (Nagin and Tremblay, 2005a). GBTM takes a very different approach, allowing for situations where estimated developmental pathways may not 'fit' or follow one common shape or growth pattern (Nagin and Tremblay, 2005a). GBTM's different approach to modeling developmental patterns – assuming individuals within a population can be stratified into homogenous clusters that may follow differently shaped pathways – makes the method a prime technique for allowing unique, distinct developmental trajectories to emerge from a population (Nagin, 1999, 2005). Another strength of GBTM is the decreased subjectivity by which researchers identify clusters of individuals on similar developmental trajectories. While HLM and LGCM employ a more ad hoc style of classifying groups *prior* to estimating developmental trajectories, GBTM provides statistical output for further assessment *after* the identification of trajectory groups (Nagin, 1999, 2005). These posterior probability statistics can be used to guide more precise classifications of groups and are an important advantage to be aware of when considering the best technique to identify distinct intraindividual developmental trajectories.

GBTM and Life-Course Theories

With the continuity of human development central to life-course theories, it is important to use analytical techniques that can account for continuous unfolding of behaviors and within-person changes in criminal involvement over time (Bushway et al., 2003; Nagin, 2005; Thornberry and Matsuda, 2011). As such, the use of GBTM in life-course criminology is strengthened by its allowance for variations in rates at which individuals move along developmental trajectories as well as permitting individuals' trajectories to differ by developmental stages, times, or ages (Krohn et al., 2013). Furthermore, by identifying rather than assuming patterns of antisocial development within a population, GBTM has been an invaluable tool in answering the types of questions asked by life-course criminologists (Nagin and Odgers, 2010a).

As an example, Moffitt's (1993) dual taxonomy model suggests two clear, potential developmental offending paths, with individuals' age of onset as the integral consideration for classifying them as ALs or LCPs. Alternatively, other life-course theories, such as Thornberry and Krohn's (2005) interactional theory, do not model such strict classifications of offending groups. Using GBTM to address empirical questions stemming from these life-course theories, researchers have found support for some theoretical contentions but also revealed unexpected developmental patterns of antisocial behavior (i.e., offending groups beyond ALs and LCPs) (Farrington, 2005; Nagin and Tremblay, 2005a). These studies have certainly established variation of individuals' offending patterns over time, showing differences by age of onset, severity of antisocial behavior (i.e., frequency, prevalence), and timing of desistence (Krohn et al., 2013; Piquero, 2008). A comprehensive review of research using different longitudinal data to estimate patterns of offending behavior over time reveals an average of four offending groups (Piquero, 2008).

For illustrative purposes, **Figure 1** presents the results from one study that has estimated respondents' (aged 13–22 years) offending trajectories. Using data from the RYDS, Bushway et al. (2003) find a best-fitting model of seven distinct trajectory groups. These trajectories reveal unexpected patterns, for example, one noteworthy group of 'late bloomers' (group P3). Distinguishing late bloomers from other offending trajectories is their unexpected age of onset when older, as well as their increasing severity of antisocial behaviors (Bushway et al., 2003; Krohn et al., 2013). For a certain period, late bloomers are even shown to surpass the

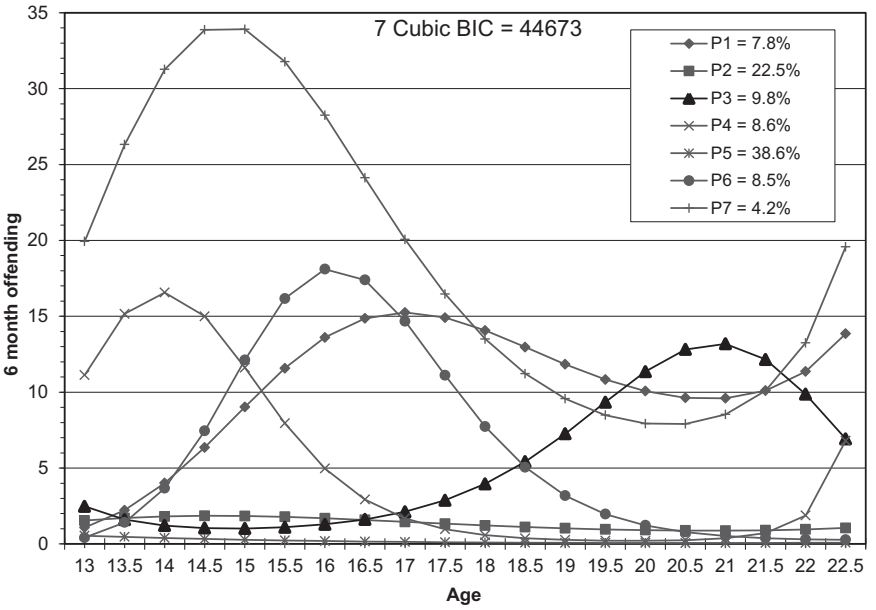

Figure 1 Rochester Youth Development Study offending trajectories estimated with GBTM (Bushway, S.D., Thornberry, T.P., Krohn, M.D., 2003. Desistance as a developmental process: a comparison of static and dynamic approaches. Journal of Quantitative Criminology 19, 129–153.).

offending patterns displayed by high-level chronics (Bushway et al., 2003; Krohn et al., 2013).

The unearthing of this late-bloomer group using GBTM encourages theoretical explanation and advancement, with scholars already offering reasons for late-bloomer groups. Recent efforts include suggestions that the late-bloomer trajectory is a reflection of social changes – specifically the timing of certain developmental events (i.e., careers, marriage, parenthood) coming at a later age than they did in prior decades (Mata and van Duhlmen, 2012).

Important Considerations for GBTM

In sum, because no extant theory clearly or accurately outlines how many developmental trajectories should exist in a population, GBTM is an asset for scholars attempting to identify such patterns and further develop theoretical propositions. The following are key considerations when applying GBTM. First, groups are not to be considered as literal representations; they are estimates and are theory driven. Developmental trajectory groups are latent strata or clusters of individuals with similarities. These latent strata groups are sometimes referred to as 'statistical fictions,' suggesting the groups *do* exist in the best-fitting model of the data (i.e., there *is* an LCP group of offenders), but the best-fitting model is still an estimate of the *true* population parameters (Nagin and Tremblay, 2005c). Secondly, once trajectories are identified, posterior probabilities can also be used to assess precision of group membership. These posterior probabilities can also be used to classify groups and allow for better description of the characteristics, or profiles, of clusters following different developmental patterns. In addition to these key features, research focusing on GBTM or more general latent class models notes that such analyses are sensitive to the data being utilized, with longer follow-up lengths (i.e., more waves of data) and environmental exposure (i.e., incarcerated versus in the community) impacting trajectories' shapes (Eggleston et al., 2004; Sampson and Laub, 1993). As such, GBTM and related techniques are increasingly combined with other statistical techniques such as propensity score matching (e.g., see Haviland et al., 2007; Ward et al., 2013) to better assess causal relationships or dual trajectory estimation to assess contemporaneous, or comorbid events over the life course.

Future Directions

The recognition that to fully understand the onset, duration, seriousness, continuity, and desistance of crime requires a full examination of what occurs throughout the life course is a relatively recent insight. The life-course approach to the study of crime is still in its early developmental stage. This brief overview of major concepts, theories, and analytical tools used in the life-course perspective touches on only some of the major themes of the approach and, frankly, fails to provide an appreciation for the complexity of examining crime across the life course. While there has been progress in beginning to explore issues related to the trajectory of criminal behavior, there is still much to be done before a true understanding of both the continuity and change in crime and the effects it has on other life-course trajectories can be accomplished. In the remaining paragraphs, we offer a few suggestions for future work on crime across the life course.

To date, much of the conceptual work on crime across the life course has focused on the degree to which there is continuity versus change in criminal behavior. For example, Moffitt's

dual taxonomy approach emphasizes continuity of crime for LCPs. Factors such as neurological deficits set in motion a sequence of consequences that result in child conduct disorders leading to juvenile delinquency and continuing crime into adulthood. Sampson and Laub's age-graded theory emphasizes the role played by adult social bonds in changing the trajectory of crime. Oftentimes, the debate on the degree of overlap between continuity and change is couched in terms of whether selection factors (e.g., neuropsychological deficits) or socialization factors (e.g., marriage) are more important in determining individual trajectories of crime. Such debate has led to an overexaggeration of the differences between these positions. While this discussion may have been necessary for the emerging life-course perspective, it is now superseded by the recognition that there are both continuity and change in the patterns of crime and that both selection and socialization are important components of such patterns. To advance our understanding of the life-course perspective, the first step is to rephrase the conceptual question from 'Whether there is more stability or change in crimes?' to 'Under what conditions do people continue on their criminal or noncriminal trajectory and under what conditions are there changes in the initial pattern?'

Relatedly, we need to identify similarities across approaches and the conceptualizations contained in different perspectives, as well as explore consistencies in findings across theoretically directed research. David Farrington (2005) has paved the way for these efforts by requesting scholars representing major life-course perspectives address how their theories answer his proposed set of 24 questions concerning issues such as the onset, duration, desistence, and the scope of the theory. Farrington further provides a summary chapter that could be considered a template for identifying the overlap in life-course theories. The next step is to examine the findings on the shared propositions.

Much of the examination of factors that affect changes in crime across the life course has focused on mediating processes. Essentially this approach considers the sequential causal processes that lead to continuity or change in criminal behavior. This, of course, is an essential approach if we are to understand the life course of criminal behavior. However, predicated on the risk and protective factor approach, we are increasingly beginning to see analyses that focus on moderating factors (Krohn et al., 2014; Lizotte et al., 2013). Moderating factors are those that interact with causal variables to either reduce or enhance covariates' impact on criminal behaviors. One potentially valuable contribution of this approach is its focus on factors that could be manipulated to reduce the probability of continuing criminality for those most at risk.

Also, although many life-course theories recognize that genetic factors play a role in accounting for the trajectory of crime, the specifics of that role and relative importance have not been fully explored. In part, this is because criminologists have just begun to seriously consider the role of genetics in their theoretical models. One reason for this is the lack of longitudinal data sets that include molecular genetic data. Fortunately, this is changing as collection of these types of data is becoming more frequent. Shanahan and Hofer (2005) suggest that the gene–environment interaction may manifest in two distinct ways. The most common way this interaction is conceptualized is with the environment playing a moderating role in the relationship between genetic factors and crime. That is, a positive social environment will impact individuals' propensity for crime. Such propensities for crime may be genetically coded. Alternatively, genetic factors may act as a trigger, enhancing the problematic effects of stressful life conditions. Krohn et al. (2013) discuss how contextual triggers may operate in the explanation of the late-bloomer group found in some studies (e.g., see Bushway et al., 2003). Future data collection efforts increasingly need to include collection of molecular genetic data, while theoretical perspectives need to provide insight into the role genetic factors may play in the crime over the life course.

Methodologically, we continue to see the development of analytical tools allowing for the examination of the intra-individual change. One advancement in trajectory modeling that shows promise for examining the relationship among different trajectories is dual trajectory modeling. Dual trajectory modeling allows for two separate outcomes to be modeled via repeated measures over time (Nagin, 2005; Nagin and Odgers, 2010b). Dual trajectory modeling has gained traction not only due to its ability to model co-occurring comorbid outcomes, but also due to the straightforwardness and explanatory power of resulting statistical output (Nagin and Odgers, 2010b). Such output relays information about the unique trajectory groups for both measures: probabilities of group memberships for each estimated trajectory group as well as probabilities connecting trajectory group memberships across outcomes (Nagin, 2005; Nagin and Odgers, 2010b). It is these probabilities relating group memberships across outcomes that have distinguished dual trajectory modeling from other longitudinal methods, with further analyses allowing for examination of variations according to individual-level factors (Nagin and Odgers, 2010b).

In his 2011 presidential address to the American Society of Criminology, Frank Cullen stated, "life-course criminology (LCC) now is criminology" (Cullen, 2011: p. 310). He further noted the current limited utility of what he characterized as adolescence-limited criminology. Indeed, he called for continuing emergence and further development of the life-course paradigm. This article provided an overview of the general life-course approach, reviewed four of the prominent life-course theories, and examined alternative analytical tools for examining crime, and its causes and consequences over the life course.

See also: Control Theories and Crime; Criminology: Development and Current Status of the Discipline; Quantitative Methods and Crime; Social Learning Theory and Crime.

Bibliography

Blumstein, A., Cohen, J., Farrington, D.P., 1988. Criminal career research: its value for criminology. Criminology 26, 1–35.

Bollen, K.A., Curran, P.J., 2006. Latent Curve Models: A Structural Equation Approach. Wiley, New York.

Bushway, S.D., Thornberry, T.P., Krohn, M.D., 2003. Desistance as a developmental process: a comparison of static and dynamic approaches. Journal of Quantitative Criminology 19, 129–153.

Cullen, F.T., 2011. Beyond adolescence-limited criminology: choosing our future – the American society of criminology 2010 Sutherland address. Criminology 49, 287–330.
Curran, P.J., Bollen, K.A., 2001. The best of both worlds: combining autoregressive and latent curve models. In: Collins, L.M., Sayer, A.G. (Eds.), New Methods for the Analysis of Change. American Psychological Association, Washington, DC, pp. 107–135.
Devers, L., 2011. Desistence and Developmental Life Course Theories: Research Summary. Under Contract No. GS-10F-0114L, Order No. 2008-F_08151 for Bureau of Justice Statistics. CSR Incorporated, Arlington, VA.
Eggleston, E.P., Laub, J.H., Sampson, R.J., 2004. Methodological sensitivities to latent class analysis of long-term criminal trajectories. Journal of Quantitative Criminology 20, 1–26.
Elder, G.H., 1975. Age differentiation and the life course. Annual Review of Sociology 1, 165–190.
Elder, G.H., 1985. Perspectives on the life course. In: Elder, G.H. (Ed.), Life Course Dynamics: Trajectories and Transitions, 1968–1980. Cornell University Press, Ithaca, NY, pp. 23–49.
Farrington, D.P., 1992. Criminal career research in the United Kingdom. British Journal of Criminology 32, 521–536.
Farrington, D.P., 2003. Developmental and life-course criminology: key theoretical and empirical issues – the 2002 Sutherland award address. Criminology 41, 221–255.
Farrington, D.P., 2005. Integrated Developmental Life-Course Theories of Offending. Transaction Publishers, New Brunswick.
Gottfredson, M., Hirschi, T., 1988. Science, public policy, and the career paradigm. Criminology 26, 37–55.
Haviland, A.M., Nagin, D.S., Rosenbaum, P.R., 2007. Combining propensity score matching and group-based trajectory analysis in an observational study. Psychological Methods 12, 247–267.
Hirschi, T., 1969. Causes of Delinquency. University of California Press, Berkeley, CA.
Krohn, M.D., Lizotte, A.J., Perez, C., 1997. The interrelationship between substance use and precocious transitions to adult statuses. Journal of Health and Social Behavior 38, 87–103.
Krohn, M.D., Gibson, C.L., Thornberry, T.P., 2013. Under the protective bud the bloom awaits: a review of theory and research on adult onset and late blooming offenders. In: Gibson, C.L., Krohn, M.D. (Eds.), Handbook of Life-Course Criminology. Springer Verlag, New York, pp. 183–201.
Krohn, M.D., Lizotte, A.J., Bushway, S.D., Schmidt, N.M., Phillips, M.D., 2014. Shelter during the storm: a search for factors that protect at risk adolescents from violence. Crime and Delinquency 60, 381–403.
Laub, J.H., Sampson, R.J., 1993. Turning points in the life course: why change matters to the study of crime. Criminology 313, 301–325.
Lizotte, A.J., Phillips, M., Krohn, M.D., Thornberry, T.P., Bushway, S.D., Schmidt, N.M., 2013. Like parent like child? The role of delayed childrearing in breaking the link between parent's offending and their children's antisocial behavior. Justice Quarterly. http://dx.doi.org/10.1080/07418825.2013.771206.
Loehlin, J.C., 2004. Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis, fourth ed. Taylor Francis Group, New York.
Mata, A.D., van Duhlmen, M.H., 2012. Adult-onset antisocial behavior trajectories: associations with adolescent family processes and emerging adulthood functioning. Journal of Interpersonal Violence 27, 177–193.
Moffitt, T.E., 1993. Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy. Psychological Review 100, 674–701.
Moffitt, T.E., 2003. Life-course persistent and adolescence-limited antisocial behavior: a 10-year research review and a research agenda. In: Lahey, B.B., Moffitt, T.E., Caspi, A. (Eds.), Causes of Conduct Disorder and Juvenile Delinquency. Guilford Press, New York, pp. 49–75.
Muthen, B., Asparouhov, T., 2011. Beyond multilevel regression modeling: multilevel analysis in a general latent variable framework. In: Leckie, G. (Ed.), Handbook of Advanced Multilevel Analysis. Routledge, New York, pp. 15–40.
Nagin, D.S., 2005. Group-Based Modeling of Development. Harvard University Press, Cambridge, MA.
Nagin, D.S., 1999. Analyzing developmental trajectories: a semiparametric, group-based approach. Psychological Methods 42, 139–157.
Nagin, D.S., Odgers, C.L., 2010a. Group-based trajectory modeling nearly two decades later. Journal of Quantitative Criminology 26, 445–453.
Nagin, D.S., Odgers, C.L., 2010b. Group-based trajectory modeling in clinical research. Annual Review of Clinical Psychology 6, 109–138.
Nagin, D.S., Tremblay, R., 2005a. What has been learned from group-based trajectory modeling? Examples from physical aggression and other problem behaviors. The Annals of the American Academy of Political and Social Science 602, 82–117.
Nagin, D.S., Tremblay, R., 2005b. Developmental trajectory groups: fact or a useful statistical fiction? Criminology 43, 873–899.
Nagin, D.S., Tremblay, R., 2005c. Further reflections on modeling and analyzing developmental trajectories: a response to Maughan and Raudenbush. The Annals of the American Academy of Political and Social Science 602, 145–154.
Piquero, A.R., 2008. Taking stock of developmental trajectories of criminal activity over the life course. In: Liberman, A.M. (Ed.), The Long View of Crime: A Synthesis of Longitudinal Research. Springer, New York, pp. 23–78.
Preacher, K.J., Wichman, A.L., Briggs, N.E., MacCallum, R.C., 2008. Latent Growth Curve Modeling. Sage, Thousand Oaks, CA.
Raudenbush, S.W., Bryk, A.S., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods, second ed. Sage, Newbury Park, CA.
Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., Congdon, R.T., 2004. HLM 6: Hierarchical Linear and Nonlinear Modeling. Scientific Software International, Chicago.
Sampson, R.J., Laub, J.H., 1990. Crime and deviance over the life course: the salience of adult social bonds. American Sociological Review 555, 609–627.
Sampson, R.J., Laub, J.H., 1992. Crime and deviance in the life course. Annual Review of Sociology 18, 63–84.
Sampson, R.J., Laub, J.H., 1993. Crime in the Making: Pathways and Turning Points through Life. Harvard University Press, Cambridge, MA.
Shanahan, M., Hofer, S., 2005. Social context in gene-environment interactions: retrospect and prospect. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 60, 65–76.
Thornberry, T.P., 1987. Toward an interactional theory of delinquency. Criminology 25, 863–891.
Thornberry, T.P., Krohn, M.D., 2001. The development of delinquency: an interactional perspective. In: White, S.O. (Ed.), Handbook of Youth and Justice. Plenum, New York, pp. 289–305.
Thornberry, T.P., Krohn, M.D., 2005. Applying interactional theory to the explanation of continuity and change in antisocial behavior. In: Farrington, D.P. (Ed.), Integrated Developmental and Life-Course Theories of Offending: Advances in Criminological Theory. Transaction Publishers, New Brunswick, NJ, pp. 183–210.
Thornberry, T.P., Matsuda, M., 2011. Why Do Late Bloomers Wait? An Examination of Factors that Delay the Onset of Offending. Presented at the Stockholm Criminology Symposium. Stockholm, Sweden.
Ward, J., Gibson, C.L., Krohn, M.D., 2013. The effects of police contact on trajectories of violence: a group-based, propensity score matching analysis. Journal of Interpersonal Violence 29, 440–475.
Wolfgang, M., Figlio, R., Sellin, T., 1972. Delinquency in a Birth Cohort. University of Chicago Press, Chicago.

Life Course Epidemiology

Gita D Mishra, School of Population Health, The University of Queensland, Herston, QLD, Australia
Diana Kuh, MRC Unit for Lifelong Health and Ageing, London, UK
Yoav Ben-Shlomo, School of Social and Community Medicine, University of Bristol, Bristol, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Life course epidemiology examines biological, behavioral, and social pathways that link exposures during gestation, childhood, adolescence, and adult life, and across generations, to influence health and health inequalities in later life. This article traces the origins of the field from its focus on early life factors and adult chronic diseases to a wider investigation of how the trajectory of exposures can act across life to influence disease etiology. It describes how simple life course models are applied to elaborate the independent, cumulative, and interactive effects of exposures. It concludes with recent developments and future directions for life course studies.

Introduction

With the aging demographic profile of many countries, research interest has increasingly shifted toward understanding disease etiology – especially for chronic diseases – and the long-term effects of health behaviors and lifestyle from a whole life perspective. Recent decades have seen the corresponding emergence of life course epidemiology as a coherent discipline within the health sciences. Life course epidemiology can be defined as the study of the long-term effects on adult health and health-related outcomes of biological (including genetic), environmental and social exposures during gestation, childhood, adolescence, earlier in adult life, and across generations (Kuh et al., 2003). This inherently multidisciplinary approach emphasizes the role of timing of exposures and understanding how both risk and protective factors act independently, cumulatively, and interactively across life. The overarching aim of the life course approach is to produce findings that not only can provide insights on the underlying biological, behavioral, or psychosocial pathways operating from the earliest stages of life, but also have implications for public health policy in terms of the optimal timing and targeting of preventive health strategies throughout life.

Historical Background

The core notions of life course epidemiology, regarding the potential of exposures at various life stages – especially early life – for latent and long-term consequences on health, have deep historical origins (Kuh et al., 2003). Yet it was in the nineteenth century, with the burgeoning growth of industrial cities, that we saw a growing awareness of overcrowding, poor sanitary conditions, pollution, and a host of other public health challenges. By the early twentieth century, the perceived decline in the fitness and vitality of the general population was placed at the forefront of the public health debate (Kuh et al., 2003). New maternal and infant welfare services were being funded, which sparked debate over whether fitness was primarily genetic (in which case genetically unfit infants would be kept alive by these services, potentially leading to a degeneration in the population's overall health) or could be encouraged and shaped by early environments and 'maternal efficiency' (which would ultimately improve national fitness) (Pearson, 1912; Paton and Findlay, 1926). While this provides an early example of a life course approach framing the discourse of population health at the time, epidemiological evidence to support either of the propositions was limited (Kuh and Davey Smith, 2004).

In the subsequent decades, interest shifted away from early life influences as postwar cohort studies of middle-aged men focused on adult chronic diseases and identified their more proximal biological risk factors, such as high blood pressure (Kuh and Davey Smith, 2004; Ben-Shlomo et al., 2013). As a result, a life course perspective was apparent only sporadically in epidemiological research. For example, in 1934, Kermack, McKendrick, and McKinlay analyzed British mortality rates from 1845 until the 1930s, and described what are now termed 'birth cohort effects' (Kermack et al., 2001; Kuh et al., 2003). They observed that "the actual calendar year is of relatively little importance in determining the improvement in the specific death-rates" (Kermack et al., 2001: p. 679). Instead they identified that a key determinant was the birth year of the generation or group of individuals under consideration. Exemplifying their life course perspective, the paper concluded that "the important factor from the point of view of the health of the individual during his whole life is his environment up to the age of say 15 years" (Kermack et al., 2001: p. 683). Furthermore, in 1941 Ciocco et al. reported a comparison of the physical examinations of school children with their acceptance for or rejection from military duty 15 years later, based on another physical examination (Ciocco et al., 1941). They found that men who were rejected showed signs of physical infirmity in childhood, with the authors emphasizing in their conclusion that disease onset can be triggered by the aggregate effects of many conditions and/or incidents that occur as early as infancy (Kuh et al., 2003).

Research that would today be considered under the umbrella of life course epidemiology gathered momentum in the 1970s, with for instance work by Forsdahl linking county infant mortality rates to coronary heart disease in later life (Forsdahl, 1977). Forsdahl suggested that poverty in childhood and adolescence followed by later life prosperity might explain the variability in observed mortality from cardiovascular diseases

seen in Norwegian counties. This idea was subsequently popularized and expanded by a group of researchers, under the leadership of David Barker, who concentrated on the prenatal environment and fetal development. The Barker group proposed that the cause of many chronic diseases could be traced back to a critical period in embryological development, during which time the fetus is subject to the influence of maternal nutrition during pregnancy and prepregnancy – a process referred to as 'fetal programming' (Barker, 1998). This body of work, known initially as the 'Barker' hypothesis, stood in contrast to the popular research approach of the time on the impact of adult behaviors. In this somewhat polarized context, a multidisciplinary group of epidemiologists and social scientists from the United Kingdom coined the term 'life course epidemiology,' which instead offered a conceptual framework that could encompass both approaches. The programming hypothesis itself eventually evolved into a more generalized developmental model that extended into the postnatal and childhood period (Gluckman and Hanson, 2004).

The modern revival of life course epidemiology was also ushered in by the progressive establishment of population-based birth cohort studies (Power et al., 2013) and the imaginative and opportunistic use of historical records enabling one to construct a retrospective cohort study. One of the earliest studies to be established was the Medical Research Council National Survey of Health and Development, otherwise known as the 1946 British birth cohort study (Kuh et al., 2011). Originally, this study had been set up to investigate declining fertility, but over time it was extended and strengthened with repeated measures in regular surveys, combined with clinic-based data collection and biological samples. The duration of birth cohort and other life course studies with long-term follow-up also posed unique methodological challenges, such as the retention of cohort members over decades, and implementing robust and consistent survey methods (Pickles et al., 2007; Hardy et al., 2014). As the array of cohort studies matured, however, the wealth of longitudinal data available provided increasing empirical evidence for a multiplicity of links between both the prenatal and postnatal environment and developmental characteristics that influenced adult health, function, disease, and aging. With this catalyst, life course epidemiology began to evolve as a coherent research discipline that could usefully engage with divergent fields including chronic disease epidemiology, developmental biology, epigenetics, neurodevelopment, human development, and social demography (Halfon et al., 2014). This has been accompanied by the ongoing development of a conceptual and theoretical framework for this multidisciplinary research (Ben-Shlomo and Kuh, 2002; Kuh et al., 2003; Mishra et al., 2009). As a result, life course epidemiology is increasingly recognized in the current research environment for its potential to guide health policy and practice (Kuh, in press), with its particular emphasis on evidence that can support the appropriate timing and targeting of interventions.

Life Course Models

Central to the theoretical foundations of life course epidemiology has been the adoption of a series of simple models that form a conceptual framework for studying exposures and subsequent health outcomes using data from life course studies. A description is provided of each of the life course models – *critical period*, *sensitive periods*, and *accumulation models* plus their variants (Kuh et al., 2003; Ben-Shlomo and Kuh, 2002). In most cases these models have been adapted from other disciplines rather than originating within life course epidemiology, but the distinctive aspect of the approach is their delineation as an ensemble to cover the various types of relationships. Thus the strength lies in providing a framework for hypothesizing a specific causal pathway that can be tested against the available data and for combining individual models to form descriptions of far more complicated processes operating through the lifespan. As will be explained, however, none of this obviates the need for a systematic analysis and testing of models and careful interpretation of their results.

Critical Period Model

As the name suggests, the critical period model (**Figure 1**, top) refers to a limited time window within which an exposure is linked with irreversible changes in developmental mechanisms of biological subsystems – typically during a particularly vulnerable stage in life such as the early stages of development – that have adverse or protective effects for disease outcomes in later life (Ben-Shlomo and Kuh, 2002). The introduction has already mentioned the concept of 'fetal programming,' which is the preeminent example of a critical period model. Outside this critical period, no excess disease risk is associated with the exposure.

Sensitive Period Model

This life course model may be interpreted as an extension of the critical period model, in the sense that the effects of an exposure are no longer just limited to specific time window, but rather the strength of the link between exposure and disease risk attenuates outside a sensitive period (**Figure 1**, bottom). In other words, while an exposure may have an effect on outcomes at various periods in life, there is also a period whereby the exposures are particularly important. Whereas critical period models are primarily associated with disease risk that arises from interruptions to the development of particular biological subsystems, sensitive periods are more common in behavioral development that occurs primarily during childhood to early adulthood (Ben-Shlomo and Kuh, 2002).

Critical Period/Sensitive Periods Model with Later Modification

In practice these basic models are typically expanded to include the possibility that the effects during the critical or sensitive period, such as early life, may be subsequently modified by *other unrelated exposures* in later life, and thereby either enhancing or decreasing the risk of adverse outcomes in later life (**Figure 2**). For instance, extremely low birth weight (<800 g) has been linked with the low level of fitness in early adolescence (Rogers et al., 2005). This relationship can be better understood, however, in terms of the impact of premature birth on the development of the motor system, that is then *modified by* (here increased by) the influence of an inactive

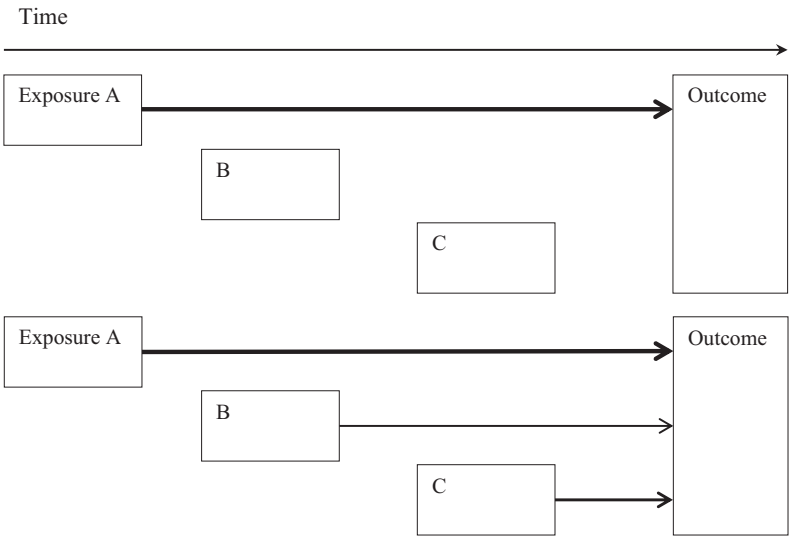

Figure 1 Critical (top) and sensitive period (bottom) life course models.

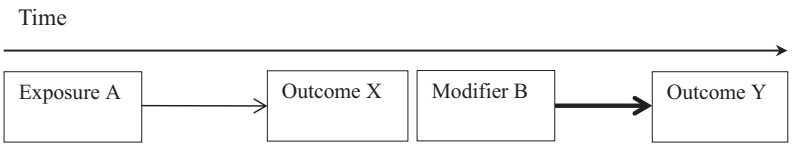

Figure 2 Critical/Sensitive period model with later modification.

lifestyle, so that it is the interaction of these two factors that forms the association with fitness level in adolescence (Whitfield and Grunau, 2006).

Accumulation Model – Independent Exposures and Direct Effects

Even if not explicitly presented as such, the accumulation model represents one of the most widely applied models and takes a number of variants. For independent exposures it refers to a simple additive model corresponding to the number of different exposures, each contributing an additional effect on the risk of the outcome (Figure 3, top). One of the key questions for debate in applying this and other models concerns whether the exposures can be treated as independent, or potentially as a cluster of exposures that result from another single factor or 'shared risk factor' (Figure 3, bottom). For instance, low socioeconomic position (SEP) can be interpreted as a common cause for a range of potential adverse exposures, since it may be related to living in an area where burglary is far more likely to occur, or to have an occupation with less job security.

Accumulation Model – Duration or Dosage of a Single Exposure

Another form of the accumulation model deals with a single exposure over a period of time, and treats the risk as accumulating according to the length of exposure or the dosage received (Figure 4). For instance, as a hypothetical illustration the health risk associated with living in a polluted environment might be considered as proportional to the number of years spent in that environment and may also include an index for the severity of the pollution. Another example might be the health risks associated with the years spent in a sedentary lifestyle, or those linked with different levels of physical (in) activity over the same time period. This simple approach is useful, as often there is only limited information about the exposure, but can also be potentially misleading without careful consideration of the potential mechanisms involved. The issue is discussed in more detail in section on health behaviors and life course trajectories below, using an example of the risks associated with smoking.

Accumulation Model – Correlated Exposures or the Chain of Risk Model

In setting out an accumulation model, the notion of independent exposures needs to be carefully examined. Often exposures are not independent but correlated as a series of related events each affecting the likelihood of the next. There are two versions particularly worth noting: the first exposure may have both an independent effect and an influence that sets an individual on a particular trajectory (Figure 5, top). An example of this scenario is the relationship between smoking and coronary thrombosis. Smoking accelerates the process of hardening and narrowing arteries, thus the risk of blood clots is directly increased. A second independent risk factor for clots is an excess of plasma fibrinogen, which is significantly decreased by exercise (Imhof and Koenig, 2001). However, smoking may

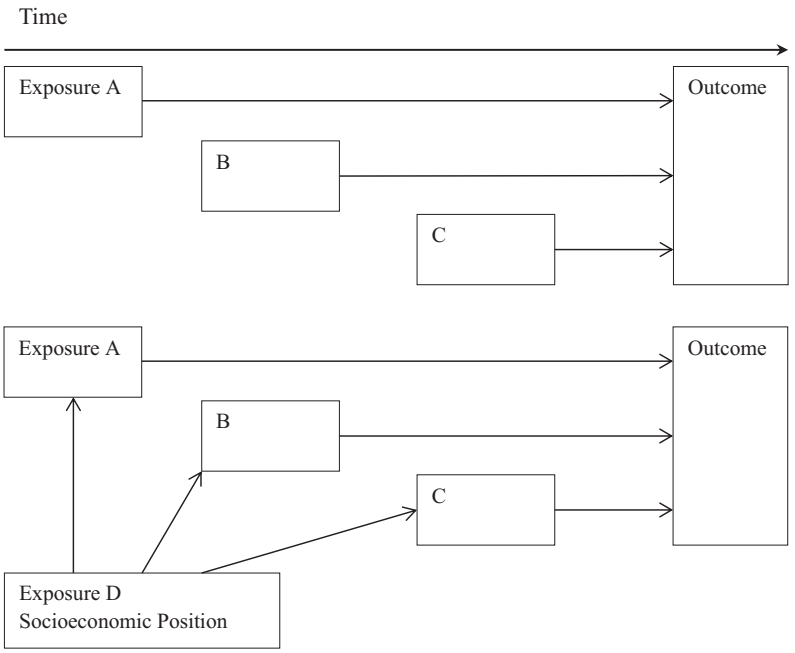

Figure 3 Accumulation life course models with independent exposures (top) and exposures with 'shared risk factors' (bottom), such as socioeconomic position.

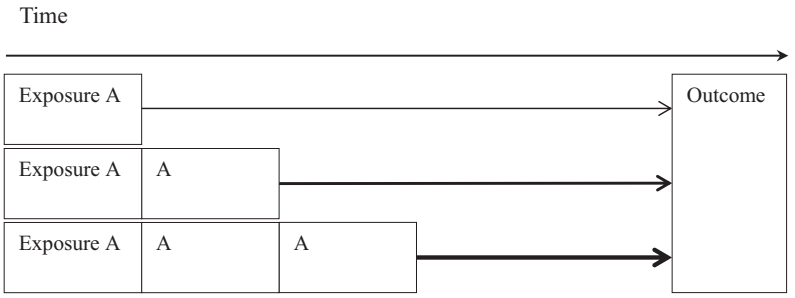

Figure 4 Accumulation life course models with same exposure and where the effect increases with the duration or the dosage of the exposure.

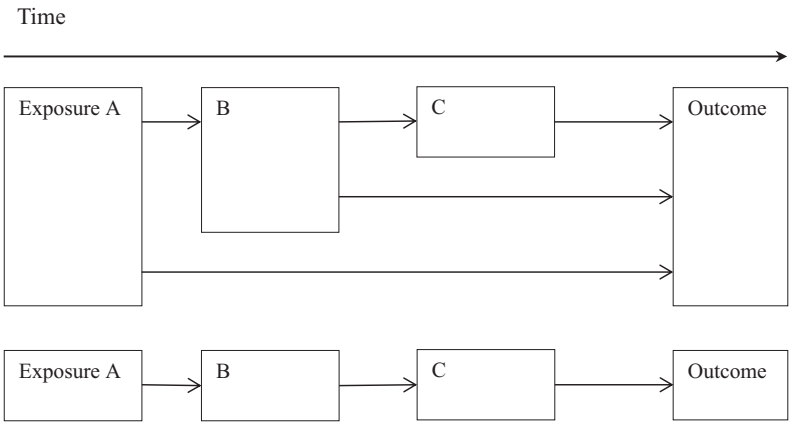

Figure 5 Accumulation life course models for independent exposures and correlated effects (top) and correlated exposures (bottom) that form a chain of risk, sometimes ending with a 'trigger exposure' (exposure C) without which the effect on the outcome would not occur.

reduce one's aerobic capacity due to respiratory symptoms, which might decrease exercise behavior and therefore levels of plasma fibrinogen are increased. Both smoking and a lack of exercise independently increase the risk of coronary thrombosis, but smoking also decreases exercise behavior.

The second scenario is sometimes referred to as a 'chain of risk' where the series of earlier events creates a situation or condition that has a predisposition for an adverse outcome to occur, but it is the final exposure that is required for an increased risk of the outcome to occur. The latter scenario is referred to as a 'chain of risk with a trigger' (Figure 5, bottom).

Life Course Models in Practice

Each of these basic life course models can be combined to form a broader set of hypotheses for the causal pathways operating in a particular group. Indeed, given the multiplicity of possible and related exposures that can occur through the life course, it is unlikely to be the case that a single simple model will suffice. Examples of augmented models have already been given, such as the use of modifying factors applied to critical and sensitive models. One of the simplest examples that combine models is where the basic accumulation model is overlaid with the chain of risk model, i.e., where correlated exposures have both direct and indirect effects on the outcome.

If the same exposure is considered, such as SEP of an individual for a series of time periods through life, then the basic critical period, sensitive period, and accumulation models can be formally interpreted as special cases of a more general model. So rather than being qualitatively different, the defining characteristics of the models is described by the constraints applied to weightings for the relationship between the exposure and outcome. (This can be seen visually if Figures 1 and 2 are reexamined, with attention paid to the thickness of each arrow linking exposure to outcome.) Thus for the critical period the weighting is such that the effect of an exposure is derived from one period and exposures at other periods have no contribution, whereas for the accumulation model the strength of the associations are constrained to being the same for each period. The sensitive period model represents the most unconstrained form, since it allows the association to vary from one period to another.

Last, the advantage of these models lies in the way they encourage the consideration of exposures, even at the conceptual level, in terms of *a trajectory* or pathway through life and even across generations. Rather than just a health outcome in later life, there is often a related trajectory for metric of health, such as physical capability, through life. From infancy, through childhood and early adulthood, there may be a series of stages of rapid development as marked by, for instance, aspects of physical capability like grip strength, that typically lead to a peak in adulthood, and followed by a decline in midlife and beyond (Cooper et al., 2014). At each time period exposures such as diet and physical activity, may play a different role in contributing to growth and hence to the peak of physical capability or slowing the rate of its decline over subsequent decades. Furthermore, the direction of any association should be examined rather than assumed; for instance a decline in physical capability may result in a decline in the quality of diet (by it being more difficult to shop for fresh foods), leading to further adverse outcomes in later life. If there are sufficient repeated measures, one of the advantages of life course studies lies in providing insights on causal mechanisms, where data can be evaluated before and after an exposure or health event. Although far from unqualified evidence for inferring causality, findings can nevertheless contribute to either support or refute specific causal pathways.

Health Behaviors and Life Course Trajectories

This ensemble of life course models provides a formal approach for setting out to identify specific mechanisms operating at different time periods in life. Combining models also permits the progressive construction of a more comprehensive model for the effect of exposures across the life course on health outcomes. To achieve this in practice, however, it is important to understand some of the challenges involved in unraveling the different life course models and the role of trajectories of factors.

Without an understanding of the trajectories of exposure, the relationship between exposures at one time point to another can easily result in a potentially misleading interpretation of findings. To illustrate, take a simple cumulative model for a hypothetical example of the trajectory of smoking among women (Ben-Shlomo et al., 2013). Suppose an initial analysis of the data identifies that the length of duration with smoking behavior is associated with an increase in the breast cancer risk (Figure 4). Rather than just duration of the smoking behavior, we might then quantify the smoking exposure in terms of the number of pack-years (that is the number of packs of cigarettes smoked per day multiplied by the number of years of smoking). If the findings again showed a 'dose–response' effect, this might be taken as a clear evidence for a biological mechanism that incorporated a steady accumulation of damage to parts of the breast tissue and hence corresponding increased risk of the disease. In policy terms such evidence might suggest the use of strategies that aim to reduce smoking at any time through life. In a naive interpretation, the evidence would not suggest to priorities targeting of health measures at any particular time point.

Now suppose in further analysis we were to compare the subgroup of women with 20–30 pack-years, but who started smoking in puberty and found their risk was higher than those with the same pack-years but who commenced their smoking behavior later in life (Figure 6). Such findings would suggest a sensitive period model and a possible biological mechanism whereby tobacco exposure during puberty – a time of rapidly dividing breast tissue – presents an additional cancer risk. At least one study has found that initiation of smoking in puberty may be related to a greater risk of breast cancer after taking smoking duration into account (Band et al., 2002). Another perspective is to consider the trajectory of smoking behavior: in a cohort study that investigated the relationship of timing of when smoking behavior commenced among both men and women with the number of pack years over the lifespan, it was found that those who started smoking by the age of 16 were at increased risk of becoming a lifelong smoker (Clennell et al., 2008). In contrast with the interpretation from findings relying on the simple accumulation model above, both of these

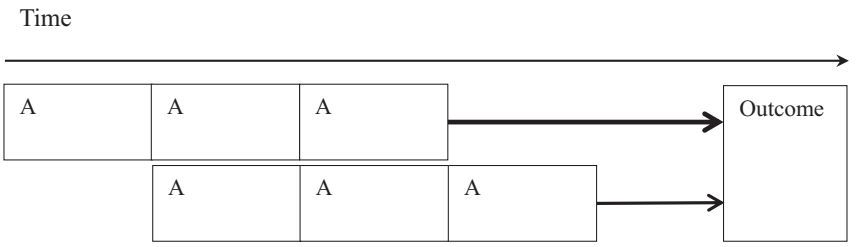

Figure 6 Exposure of the same length, but with different effects on the outcome according to when the exposure commenced.

results from additional analysis would strongly support preventive health policies that prioritized discouraging the commencement and reduction of smoking among very young women.

Socioeconomic Mobility and Life Course Model Selection

One of the main research areas of social epidemiology concerns the impact of SEP over the life course – typically signified by education, occupation, or income level. The conventional approach has been to assume an accumulation model, where the effects in later life are proportional to the time spent in a particular SEP category, and with evidence linking low SEP with a swathe of adverse health outcomes (Marmot and Bell, 2012). Further to this is the investigation of social mobility, which occurs when an individual moves from a lower to a higher SEP, or vice versa – though the latter is less often the case. Social mobility can be described as intragenerational, when an individual moves between SEP within a lifetime, or more typically as intergenerational, whereby SEP changes across generations (e.g., from parental SEP during the participant's childhood to the participant's adult SEP).

Social mobility represents an area of considerable interest from a sociological and policy perspective, since it can act as a marker of significant social change, such as improved access to higher qualifications. Yet the interpretation of the effects of SEP and social mobility requires careful consideration both in terms of interpreting the trajectory involved, including across generations, comparison between age cohorts, and the hypothesized mechanism at work. For instance, higher education levels in a young cohort may just reflect broader secular change rather than an improvement in social mobility, per se. Thus secondary education in a young cohort may not confer the same social advantage that it had in previous generations and therefore would not necessarily be expected to be associated with the same health benefits. Identifying SEP and hence social mobility for different population sectors can also be problematic, such as women in older cohorts when work force participation was lower (and then resorting to using data on the SEP of the spouse).

Another aspect to consider is *social causation*, which highlights the pathways in which SEP affects health. For instance, low SEP is linked with an inadequate diet which could be the factor leading to poor health. Similarly *social selection* refers to the possibility of reverse causation, where poor health may lead to low SEP (by for instance limiting the occupation level of the individual). Last, data on SEP through life may cover unequal time periods – sometimes only two data points, with one in childhood and one in adulthood – and hence correspond to different lengths of exposure. This would not be captured in an index used for a simple accumulation model that just scored SEP across life according to the number of data points in a particular level.

Hallqvist et al. (2004) raised the fundamental question of the appropriateness of assuming a priori that SEP should be treated according to the accumulation model. He concluded, however, that unraveling the eight trajectories for SEP that corresponded to the different life course models – critical period, sensitive period, and accumulation model, was intractable. In two subsequent papers, Mishra et al. (2009, 2013) set out an alternative formal approach that overcomes the difficulties by framing each of the life course models as special cases of the more general or *full model*, and differing only by the weighting given to the SEP measured at each time period (and their interaction terms). Thus the critical period model gives full weighting to just one time period, whereas the accumulation model sets equal weights to each time period, as was outlined in the previous section. Essentially each model can then be tested statistically against the full model, to see which provides the best fit. If the basic life course model differs significantly (in statistical terms) from the associations identified using the full model, then introducing the constraints of the basic model has lost information that is encapsulated in the full model. If there is no statistical difference, then the basic life course model being tested captures the relationships as well as (again, in statistical terms) the full model.

These brief examples serve to illustrate how life course models on the one hand provide a way of articulating potential mechanisms or for providing evidence to support specific policy options, such as interventions in childhood or early adulthood compared with other times in life. But on the other, without thoughtful consideration of the appropriate research hypotheses and systematic approach to their analysis, their use as evidence for the support for a specific mechanism or of one policy strategy over another would not be justified. In summary, proceed with due caution.

Current Challenges and Future Directions

We have presented life course epidemiology as an approach that aims to move beyond simple associations of risk factors with health outcomes – findings that have been the historical mainstay of case control and cohort studies – to provide instead a framework for a far more overarching approach and synthesis of ideas for understanding human health.

For example, Halfon et al. (2014) have proposed a Life Course Health Development (LCHD) model where human health develops in response to multiple nested and dynamic genetic, biological, behavioral, social, and economic contexts: during critical and sensitive periods of development, risk and protective factors affect the development of biobehavioral regulatory system; the wider context affects the health at individual and population levels through the impact and timing of biological, cultural, and historic factors (Ben-Shlomo et al., 2013). For epidemiology to address such a complex web and sequence of influences, to examine pathways that optimize health outcomes across the lifespan, life course studies have needed to continue to evolve and expand dramatically.

Multicohort Studies

Age cohort and period effects are an inevitable challenge for many life course studies. For instance, the childhood diet of participants in the 1946 Birth cohort study was subject to the unique restrictions of postwar food rationing in the UK (Wadsworth et al., 2006). Thus findings from that study related to dietary effects on subsequent disease risk may have limited relevance for subsequent generations with entirely different dietary intakes. Though, this 'natural experiment' may aid researchers to control for confounding, as the normal variability in dietary intake by socioeconomic status would have been attenuated given these special circumstances. Additionally, many social conditions and population characteristics have changed even in recent decades. Beyond systematic reviews and meta-analysis, researchers have begun to combine data at the individual level across multiple cohort studies. Specifically, findings from such multicohort studies that demonstrate systematic similarities or differences across cohorts, derived from distinct generations, diverse populations, or cultures, can provide evidence that is more powerful and detailed than those from a single study. For instance, it may be possible to identify the impact of a public policy – such as interventions in early childhood health and education – from the time of its initiation, based on comparison between results from a cohort where this occurred and those from other age cohorts or cohorts from other populations where the policy had not been implemented. Methodological difficulties typically arise with combining data from multicohort studies, such as harmonization of the survey instruments used, which can limit the detail and type of comparisons that can be drawn. Nevertheless an example of an interdisciplinary research collaboration of this type that has been established is the Healthy aging across the life course (HALCyon) research collaboration. This project uses the power of nine UK cohort studies to investigate life course influences on physical and cognitive capability, social and psychological well-being, and underlying biology (www.halcyon.ac.uk; Kuh et al., 2012).

Record Linkage

Often a wealth of information already exists on study participants as part of the administration of health and community services. If the ethical issues are fully addressed – and this issue should be considered paramount – the information from record linkage can represent an immense opportunity to add value to existing cohort studies. For instance, with the participants' permission and appropriate ethics approval, the possibility may exist to link survey data with medical or school records. At the most fundamental level this may allow a cohort study with baseline data in adulthood to be extended as a life course study through the inclusion of sufficient early life data, such as birth weight, anthropometric data to capture childhood growth trajectories and childhood development indices. There is also potential to link with medical and health service use data that would provide further benefits, for instance in quantifying the impact of life events like prolonged unemployment or the impact of early schooling (Baker et al., 2007) on subsequent health service use are potentially invaluable for guiding preventive health strategies. Combined with health survey data from a baseline, these records may reveal evidence for the long-term effects of a widespread treatment in the population, such as the use of antidepressants or hormone replacement therapy. Any such effects may only be implicated in a specific subsample of the population or at a particular time of life, for instance during adolescence. To reiterate, the ethical questions are crucial and need to be addressed, especially the informed consent from participants and the maintenance of their anonymity in any research and findings. Nevertheless protocols are already standard practice in life course studies that can address privacy issues with secure matching of individual records across databases, with researchers only dealing with anonymised data.

Advances in genomic technologies has seen growing interest in the potential for genetic and epigenetic research by gathering biological samples – if they have not previously been done so for other aspects of the study – from existing cohorts for detailed genetic analysis. This means that mature cohort studies that primarily consist of repeat survey data can be reexamined in light of genetic information, such as in the predisposition for certain conditions among individuals, combined with specific environmental exposures that result in a greatly increased risk for that condition becoming manifest.

Family-Based Studies

Family-based studies represent another important emerging direction for life course epidemiology. Rather than focus on just the individual cohort member, family-based studies are designed to collect data on more than one family member (Lawlor and Mishra, 2009). The rational is the familial context which provides multifaceted influences that are fundamental to the developmental trajectory of an individual. Throughout childhood and into adolescence, parents and siblings bear strong influence on one's social and psychological life experience. Furthermore, family-based designs provide a unique opportunity for disentangling genetic, environmental, and social factors, opening up research questions that go beyond simple associations into multilevel analyses. For instance with data on siblings, comparisons within and between different family members can begin to provide evidence as the basis for inferring causality (in conjunction with plausible

biological or other mechanisms) (Lawlor and Mishra, 2009). Family-based studies also present an opportunity for exploring the role of timing of exposures across the life course, since particular family members can have varying effects on one's health at different points through life.

The family-based design is not, however, without methodological challenges (Frisell et al., 2012) due to potential bias from nonshared confounders and greater attenuation of associations with random measurement error. When comparing intergenerational exposures and outcomes or across populations, it is imperative to take into account the changing nature of the family unit and the potential impact on the family environment that such changes might incur. For example, over the last few decades the proportion of working mothers and single-parent families has increased greatly. This might be reflected socioeconomically or have various other social and environmental impacts for an individual (e.g., more time spent in day care). Complex family structures where two single parents and their offspring join to form one household are not uncommon. Family-based studies also pose various other practical and methodological challenges. One example is dealing with proxy responses where one individual provides information on another family member, such as a child responding to questions about their siblings. While in some age groups, the use of proxy responses may be entirely appropriate, such as maternal responses on their infant child, they may also be unreliable, as has been shown in the reporting of disabilities (Todorov and Kirchner, 2000).

Nevertheless, family-based studies represent an essential way forward for life course epidemiology given the increasing evidence for the key role of the nurturant qualities of the environments where children grow up (Maggi et al., 2010). This interest in family-based studies as part of life course epidemiology fits into the broader trend of augmenting existing cohort studies, with direct data collection on family members or through record linkage of registry data. Some recent cohort studies have been specifically designed to address family associations from the outset, such as the Generation Scotland study that comprises of 50 000 individuals within sibling and parent–offspring groups (Smith et al., 2013). The China Children and Families Cohort Study provides another example, with 300 000 families recruited in some cases even prior to conception of the prospective main participant (Brown et al., 2007). New and future family-based studies need to be undertaken by researchers drawing on experience from existing cohort research that have provided methodological rigor and best practice over time. In due course this should permit cross-cohort comparisons at the international level, where the life course approach could be applied to investigate the impact of exposures and disease mechanisms across generations and populations with disparate cultural, socioeconomic, and environmental conditions.

Conclusions

In summary, the life course approach can be seen as having two essential and distinct features of modern epidemiology: it represents a synthesis of ideas from diverse disciplines and a framework for multidisciplinary research, while it also incorporates a set of life course models to provide tools for testing specific hypotheses and causal pathways. This combination of both a broad perspective and detailed focus has facilitated and is reflected in the spectrum of ideas from fetal programming hypothesis to rich theoretical models, such as exemplified by the Life Course Health Development model.

The contribution of life course epidemiology for research translation, to provide insights for causal pathways or to strengthen the evidence base for better targeting and timing of health interventions, is tempered by the need for careful selection of life course models. Methods and guidelines for ensuring that researchers adopt best practice in their analysis and interpretation of findings are at the cusp of recent advances in the field and require much further development.

Yet life course epidemiology is relatively well positioned for the inevitable rise in the use of 'big data' as part of the augmentation of both new family-based studies and existing cohort studies that goes far beyond the traditional use of survey data collection and clinical measures. It provides an established framework that encapsulates an intergenerational perspective and lifespan trajectory, with a way to address the multiplicity of biogenetic and social relationships at work and that lie at the crux of epigenetic research. It is unsurprising, therefore, that in just a few short decades life course epidemiology has emerged as an integral to both current and future directions of public health research.

See also: Adolescent Health and Health Behaviors; Epigenetics in Social and Behavioral Sciences; Human Development and Health; Social Deprivation and Health; Social Epidemiology; Socioeconomic Status and Health.

Bibliography

Baker, J.L., Olsen, L.W., Sorensen, T.I., 2007. Childhood body mass index and the risk of coronary heart disease in adulthood. New England Journal of Medicine 357, 2329–2337.

Band, P.R., Le, N.D., Fang, R., Deschamps, M., 2002. Carcinogenic and endocrine disrupting effects of cigarette smoke and risk of breast cancer. Lancet 360, 1044–1049.

Barker, D.J.P., 1998. Mothers, Babies and Health in Later Life, second ed. Churchill Livingstone, Edinburgh.

Ben-Shlomo, Y., Kuh, D., 2002. A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. International Journal of Epidemiology 31, 285–293.

Ben-Shlomo, Y., Mishra, G.D., Kuh, D., 2013. Life course epidemiology. In: Ahrens, W., Peugot, I. (Eds.), Handbook of Epidemiology, second ed. Springer, Bremen.

Brown, R.C., Dwyer, T., Kasten, C., Krotoski, D., Li, Z., Linet, M.S., Olsen, J., Scheidt, P., Winn, D.M., 2007. Cohort profile: the International Childhood Cancer Cohort Consortium (I4C). International Journal of Epidemiology 36, 724–730.

Ciocco, A., Klein, H., Palmer, C.E., 1941. Child health and the selective service physical standards. Public Health Reports (1896–1970), 2365–2375.

Clennell, S., Kuh, D., Guralnik, J.M., Patel, K.V., Mishra, G.D., 2008. Characterisation of smoking behaviour across the life course and its impact on decline in lung function and all-cause mortality: evidence from a British birth cohort. Journal of Epidemiology and Community Health 62, 1051–1056.

Cooper, R., Hardy, R., Aihie Sayer, A., Kuh, D., 2014. A life course approach to physical capability. In: Kuh, D., Cooper, R., Hardy, R., Richards, M., Ben-Shlomo, Y. (Eds.), A Life Course Approach to Healthy Ageing. Oxford University Press, Oxford.

Forsdahl, A., 1977. Are poor living conditions in childhood and adolescence an important risk factor for arteriosclerotic heart disease? British Journal of Preventive and Social Medicine 31, 91–95.

Frisell, T., Oberg, S., Kuja-Halkola, R., Sjolander, A., 2012. Sibling comparison designs: bias from non-shared confounders and measurement error. Epidemiology 23, 713–720.

Gluckman, P.D., Hanson, M.A., 2004. Developmental origins of disease paradigm: a mechanistic and evolutionary perspective. Pediatrics Research 56, 311–317.

Halfon, N., Larson, K., Lu, M., Tullis, E., Russ, S., 2014. Life course health development: past, present and future. Maternal and Child Health Journal 18 (2), 344–365.

Hallqvist, J., Lynch, J., Bartley, M., Lang, T., Blane, D., 2004. Can we disentangle life course processes of accumulation, critical period and social mobility? An analysis of disadvantaged socio-economic positions and myocardial infarction in the Stockholm Heart Epidemiology Program. Social Science and Medicine 58, 1555–1562.

Hardy, R., Muniz-Terrera, G., Hofer, S., 2014. Design of life course studies of healthy ageing. In: Kuh, D., Cooper, R., Hardy, R., Richards, M., Ben-Shlomo, Y. (Eds.), A Life Course Approach to Healthy Ageing. Oxford University Press, Oxford.

Imhof, A., Koenig, W., 2001. Exercise and thrombosis. Cardiology Clinics 19, 389–400.

Kermack, W., Mckendrick, A., Mckinlay, P., 2001. Death-rates in Great Britain and Sweden. Some general regularities and their significance. International Journal of Epidemiology 30, 678–683.

Kuh, D., Ben-Shlomo, Y., Lynch, J., Hallqvist, J., Power, C., 2003. Life course epidemiology. Journal of Epidemiology and Community Health 57, 778.

Kuh, D., Cooper, R., Richards, M., Gale, C., Von Zglinicki, T., Guralnik, J., 2012. A life course approach to healthy ageing: the HALCyon programme. Public Health 126, 193–195.

Kuh, D., Davey Smith, G., 2004. A life course and adult chronic disease: an historical perspective with particular reference to coronary heart disease. In: Kuh, D., Ben-Shlomo, Y. (Eds.), A Life Course Approach to Chronic Disease Epidemiology. Oxford University Press, Oxford, pp. 15–40.

Kuh, D., Davey Smith, G., 1993. When is mortality risk determined? Historical insights into a current debate. Social History of Medicine 6, 101–123.

Kuh, D., Pierce, M., Adams, J., Deanfield, J., Ekelund, U., Friberg, P., Ghosh, A.K., Harwood, N., Hughes, A., Macfarlane, P.W., Mishra, G., Pellerin, D., Wong, A., Stephen, A.M., Richards, M., Hardy, R., 2011. Cohort profile: updating the cohort profile for the MRC National Survey of Health and Development: a new clinic-based data collection for ageing research. International Journal of Epidemiology 40, e1–e9.

Kuh, D., Ben-Shlomo, Y., Tilling, K., Hardy R. Life course epidemiology and analysis. In: Detels, R., Muniz-Terrera, G., Abdool Karim, Q., Chuan, Tan. (Eds.), Oxford Textbook of Public Health, sixth ed. Oxford University Press, Oxford, in press.

Lawlor, D.A., Mishra, G.D., 2009. Family Matters: Designing, Analysing and Understanding Family Based Studies in Life Course Epidemiology. Oxford University Press, Oxford.

Maggi, S., Irwin, L.J., Siddiqi, A., Hertzman, C., 2010. The social determinants of early child development: an overview. Journal of Paediatrics and Child Health 46, 627–635.

Marmot, M., Bell, R., 2012. Fair society, healthy lives. Public Health 126 (Suppl. 1), S4–S10.

Mishra, G., Nitsch, D., Black, S., De Stavola, B., Kuh, D., Hardy, R., 2009. A structured approach to modelling the effects of binary exposure variables over the life course. International Journal of Epidemiology 38, 528–537.

Mishra, G.D., Chiesa, F., Goodman, A., De Stavola, B., Koupil, I., 2013. Socio-economic position over the life course and all-cause, and circulatory diseases mortality at age 50–87 years: results from a Swedish birth cohort. European Journal of Epidemiology 28, 139–147.

Paton, D.N., Findlay, L., 1926. Poverty, Nutrition & Growth: Studies of Child Life in Cities and Rural Districts of Scotland. HM Stationery Office, London.

Pearson, K., 1912. The intensity of natural selection in man. Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character 85, 469–476.

Pickles, A., Maughan, B., Wadsworth, M., 2007. Epidemiological Methods in Life Course Research. Oxford University Press, Oxford.

Power, C., Kuh, D., Morton, S., 2013. From developmental origins of adult disease to life course research on adult disease and aging: insights from birth cohort studies. Annual Review Public Health 34, 27–28.

Rogers, M., Fay, T.B., Whitfield, M.F., Tomlinson, J., Grunau, R.E., 2005. Aerobic capacity, strength, flexibility, and activity level in unimpaired extremely low birth weight ($<$or $=$ 800 g) survivors at 17 years of age compared with term-born control subjects. Pediatrics 116, e58–e65.

Smith, B.H., Campbell, A., Linksted, P., Fitzpatrick, B., Jackson, C., Kerr, S.M., Deary, I.J., Macintyre, D.J., Campbell, H., Mcgilchrist, M., Hocking, L.J., Wisely, L., Ford, I., Lindsay, R.S., Morton, R., Palmer, C.N., Dominiczak, A.F., Porteous, D.J., Morris, A.D., 2013. Cohort profile: generation Scotland: Scottish Family Health Study (GS: SFHS). The study, its participants and their potential for genetic research on health and illness. International Journal of Epidemiology 42, 689–700.

Todorov, A., Kirchner, C., 2000. Bias in proxies' reports of disability: data from the National Health Interview Survey on disability. American Journal of Public Health 90, 1248.

Wadsworth, M., Kuh, D., Richards, M., Hardy, R., 2006. Cohort profile: the 1946 national birth cohort (MRC national survey of health and development). International Journal of Epidemiology 35, 49–54.

Whitfield, M.F., Grunau, R.E., 2006. Teenagers born at extremely low birth weight. Paediatrics and Child Health 11, 275–277.

Life Course in History

JR Gillis, Rutgers University, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8814–8817, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

There is an ever-growing body of literature concerned with the life course and life cycle, inspired in part by Philippe Ariès' *Centuries of Childhood* (first English language edition, 1965). The article surveys the central approaches within this historical literature, paying particular attention to the nuances of class, gender, and ethnicity. In addition the article examines the fluidity of age and life course categories prior to the nineteenth century and the tightening of these definitions under the impact of industrialization, nation-state formation, and the rise of positivism in the natural and social sciences. This approach is now being questioned, with recent work restoring a sense of contingency and agency within new interpretative frameworks.

This article will review the changing historical literature on the life course and life cycle from the 1960s to the present. It will begin with the impact of Philippe Ariès' *Centuries of Childhood* (first English language edition, 1965) on the historical study of childhood and then move to the work that has been on youth, middle age, and older age groups since that time. Because the most important developments in the field have occurred in Europe and North America, the piece will necessarily focus on the Western world, using non-Western examples only for comparative purposes. The article will pay attention to the nuances of class, gender, and ethnicity that have been recently introduced into the literature. In addition to reviewing the literature, the article will provide a brief overview of current understandings of the life course from early modern to modern times. It will discuss the fluidity of age and life course categories prior to the nineteenth century as background to understanding the tightening of these definitions under the impact of industrialization, nation-state formation, and the rise of positivism in the natural and social sciences. It will show how in the last 20 years these categories have come to be questioned, and how new understandings of the life course and aging are now emerging in the context of the demographic change, the restructuring the nation-state, and the emergence of new interpretative frameworks. The piece will conclude with a sense that historical studies of the life course and life cycle are moving in a humanistic direction, restoring a sense of contingency and agency that was apparent in the first generation of historical treatments of the subject without entirely displacing more positivist notions.

The history of the life course, of age categories, and of aging is essential to understanding how any society organizes itself over time. The notion of life cycle was first used by social science in the 1930s, but has now been superseded by the more flexible notion of life course. Historians had long been interested in the aging of individuals, but were slower to appreciate the social dimension. Today, however, notions of life course and aging are as essential to historical practice as they are to the social and behavioral sciences.

Foundations of the Field

Representations of the ages of man had previously attracted the interest of historians of art and religion, but it was not until the 1960s that the life course became a focus of social and cultural history. Beginning with Philippe Ariès' *Centuries of Childhood* (1962), there began a veritable avalanche of studies which have explored the history of every dimension of the Western life course from the Greeks to the present. Today there is a substantial body of literature on the history of childhood, youth, old age, and death, focused mainly on Europe and North America, but also touching on the non-Western world.

Ariès was concerned mainly with representations of aging, but his successors were influenced by the social sciences to direct their attention to behavior. The first generation of new work in the 1960s and 1970s was heavily influenced by historical demography, psychology, biology, and quantitative sociology, seeking to define general patterns of historical development and understand the structures and functions of age groups.

Today, the influence of positivist notions of human development are less evident. In the most recent studies of the life course it is the variability and instability of the life course and age categories that have gained the most attention.

Development of the Field

The historical study of the life course began with childhood, moved subsequently to youth, old age, and death, and, most recently, has focused attention on midlife. This sequence is less the effect of research design than historical contingency. The youth movements of the 1960s accounted for the interest in the history of adolescence; the evident importance of older age groups as a consequence of the longevity revolution focused interest on them from the 1970s onwards. Death and dying became important topics in the same period, while the recent work on mid life reflects the uncertainties experienced by adults in recent decades. What began as a field concerned mainly with the middle-class male life course has now become much more sensitive to gender and ethnic differences, yet another reflection of the ways in which the studies in this area reflect contemporary concerns.

The study of childhood has undergone dramatic changes since it was introduced by Ariès. His notion that there was no concept of the child before the early modern period was immediately challenged by medieval and Renaissance

historians who produced linguistic, artistic, and medical evidence of concern with children in their respective periods. The idea that children were subject in premodern times to an indifference bordering on neglect has also been modified in recent decades, so that we now have a much more nuanced view of earlier eras of child rearing and childhood, but one that preserves the essence of Ariès' original insight into the malleability of notions of childhood over the centuries. We have come to understand that practices such as abandonment, which may seem cruel to us, made sense in other social and cultural settings.

Ariès' insight into the mutability of age categories also influenced the first generation of studies of youth. It became clear that in premodern Europe youth was more a function of social context than biological age. Numerical age was less important than work and marital status in defining the boundary between youth and adulthood. It was not until the later nineteenth century that entry and exit into this age category became standardized, and then mainly for middle-class males who constituted the first group to be called 'adolescents.'

It was not until the twentieth century that everyone was seen as being entitled to both a childhood and an adolescence as these were defined by the emerging medical and psychological sciences. As the understanding of childhood and youth became ever more universal and precise, there developed a notion of a normal life course, leading through adulthood to old age. By following the proper rules of aging, individuals could expect to arrive at what Erik Erikson, the major theorist of the linear life course, deemed a coherent understanding of themselves and life more generally. Failure to move through the prescribed stages of life in a timely manner was often portrayed as a cause of personal and social disorder.

By the 1980s, attention had become focused on old age and death. As more and more people began to live longer, healthier, more active lives, many stereotypes of older people began to be challenged. Old age could no longer be equated with poor health, debility, and passivity. These changes contributed to new interest in old age in earlier periods, and a more nuanced view of aging and 'retirement' in the past. It turned out that the modern period has no monopoly on notions of good aging or, for that matter, of the good death.

Even adulthood, that part of the life course that had previously been neglected by historians because it seemed so universal and unproblematic, has come in for its share of attention. Awareness of midlife crises, first among men and later among women, prompted a number of recent studies focused on the changing nature of adulthood over time. Studies of fatherhood and motherhood demonstrate just how historically variable this dimension of the life course has also been, and how changing definitions of masculinity and femininity, and of sexual orientation, have affected notions of adulthood over time.

Premodern and Modern Life Courses

Historical work from 1960 to 2000 has complicated our understanding of the life course and aging in the present as well as in the past. It is now generally agreed that prior to the modern era, age boundaries were more fluid and the life course less rigidly segmented. The identity of premodern peoples was determined less by how old they were numerically than by the status or place they occupied in the social order. In a society where heading a household conferred adulthood regardless of age, a property-less man or woman would remain a 'boy' or 'girl' all his or her life. Schooling was not age graded and there was no set age of retirement.

The nature of the agricultural economy, the absence of centralized institutions, and high levels of mortality and fertility all contributed to this condition. So too did a prescientific culture, which had no notion of universal quantifiable laws of nature, but was prepared to accept sudden, nonlinear qualitative change as quite normal.

All this changed in the course of the nineteenth century. For the first time numerical age came to define status for a growing part of the population. In a process that Martin Kohli has called 'chronologization,' age became the criterion for schooling, entry into work, and military service. By the end of the century, it was also associated with entitlements such as retirement and welfare benefits.

Chronologization proceeded fastest among urban populations, affecting middle-class men first, but by the early twentieth century it was shaping the lives of a majority of the European and North American populations. Not only was aging standardized, but it was normalized, producing an acute sense of being too old or too young for certain categories of activities. The life course was organized into a linear sequence consisting of childhood/youth, adulthood, and old age, each with its appropriately gendered roles and expectations. By the mid-twentieth century, norms of successful aging held sway throughout Western society and were being exported to the non-Western world as integral to what it meant to be modern.

In retrospect, we can see that chronologization was the product of an urban industrialized society concerned to maximize productivity through the rationalization of time, including age. The newly developed nation-state was equally interested in maximizing the resources of its citizens by organizing them into age cohorts. These efforts were reinforced by the new biological, medical, and psychological sciences, which took over the responsibility for assigning meaning to age and aging from religion. By the beginning of the twentieth century, each age had its own body of expertise. Childhood was now covered by the newly invented field of pediatrics; adolescent psychology was recognized; and a new discipline of gerontology would soon develop. In our own time, death has become yet another area of expertise, thanatology.

The process of chronologization reached its apogee in the post-World War II period, but by the 1980s some were foreseeing the arrival of the 'age-irrelevant society.' Since 1980 there has been a noticeable dechronologization of the Western life course. Not only has numerical age ceased to be such a powerful determinant of status, but the boundaries between ages have blurred and new age categories such as the 'young–old' or what Peter Laslett called the 'Third Age.'

The timing and sequencing of the life course have become less uniform and normative. The ages of entry into and leaving work and schooling show greater variation. The same is true of marriage and parenting. Given new reproductive technologies, the biological clock is no longer as determinative of women's life course. There is no longer consensus on what constitutes

a child, a youth, or an old person. Adulthood, previously that part of the life course that could be taken for granted, has also become problematic.

All this has led to a reconsideration of the linearity of the life course. In the late twentieth century, the institutions that once enforced age norms no longer did so. Schools, employers, and the state all now acknowledge, even encourage, flexibility with respect to age. More concerned with consumption than with production, Western industrial society has encouraged both younger and older age groups to enter into roles in the marketplace previously reserved for adults. Children and the youth have become major consumers, while older persons are encouraged to lead active lives in the consumer society.

The nation-state has also relaxed its age criteria. Its high-tech armies are no longer composed exclusively of young men, but now include not only women but a much wider range of age groups. Today boundaries between age groups blur. The legal distinctions between childhood and adulthood have become less clear. Legal ages of drinking, driving, and voting have all been revised in recent decades.

While most people still maintain some normative notion of the life course, this gets less reinforcement than it once did from the medical and behavioral sciences. Recent studies have challenged the predictive power of numerical age in everything from learning capacity to physical fitness. It is no longer clear that life follows a uniform, predictable sequence. Gender differences and variation by class and ethnicity have complicated the contemporary picture of aging to the point that generalizations become increasingly suspect.

Today, age is seen as more subjective, more contingent. We seem to be returning to a situation similar to that which existed before the process of chronologization began. But now it is not one's status that determine's one's age, but one's own effort. We are told that how we age is up to us, dependent on how we eat, exercise, and medicate ourselves. Notions of successful aging and the good death were present in earlier periods, but were largely defined by religion. Today, religion competes with the prescriptions of science and medicine. As recent studies show, factors of class, gender, generation, and ethnicity still have a powerful effect on the way we think about age. However subjective age categories have become they do not exist outside history or culture.

Current State of the Field and Future Directions of Research

Current interest in the history of the life course reflects contemporary uncertainty about aging and the desire to know more about its contingent nature. The history of the life course and aging is no longer so influenced by the behavioral sciences as it once was, while, at the same time, psychologists and sociologists have become more aware of the historicity of their subject. Psychologists now question the notion of a linear life course and the assumption of universal laws of development that underlay it. It is now generally acknowledged that gender, class, race, and cultural background have enormous effect on the life course and life chances.

The relationship between history and the behavioral sciences has been profoundly altered. There is now much greater appreciation of the variability not only of conceptions of age, but of age relations themselves. Aging has ceased to be seen as some invariable law of nature, to be discovered and obeyed. It is no longer a problem to be formulaically solved, but a challenge that confronts every generation and to which there is no one answer, only an ongoing dialogue with ourselves and others about the meaning of age.

No period in history has ever achieved agreement on the question of when life begins and ends, or how the life span should be divided or sequenced. Furthermore, our conceptions of life are not limited to the actual life span. We have the capacity to imagine life before birth and after death, and conceptions like the fetus as 'unborn child' or notions of life after death have very powerful effects on the way we lead our lives.

Over time the perceived number of ages of man has varied enormously. Some cultures have assigned the highest value to the oldest ages; others have valorized childhood or youth over age. The so-called 'prime of life' has been equally variable over time and space. Even within a single society in a given period, people have thought about the life course in very different ways. Religion, which has traditionally been the source of ideas about age and aging, has always been a point of division. Catholics traditionally assigned meaning to particular moments in life, such as the deathbed, while Protestantism insisted that every moment was equally significant in the eyes of God.

Today, intrareligious battles over the meaning of the life course pale in comparison with the struggles between religion and science over the definition of the beginning and ending of life.

Different class groups have also been at odds when thinking about the life course. And there are obvious differences in the ways that different ethnic groups think about age and aging. Gender is also a source of variation. The notion of the linear life course has been closely associated with a certain kind of middle-class masculinity, which places great value on the development of a career. Until recently, few women have had access to careers, their lives having been more contingent on family responsibilities. They have tended to have a more flexible notion of the life course and to be more adaptable than men. In the late twentieth century, when so many people lived beyond all previous societal expectations, this kind of flexibility proved to be a great asset.

It is clear from recent research that the life course is as much the product of culture as of biology. Life is not so much a script we follow, as one we as individuals and as societies write as we go along. In premodern times that script was a religious one. In the nineteenth century, the natural and behavioral sciences took over its authorship, producing a grand narrative of the life course sufficiently consistent with the requirements that the dominant economic and political institutions endowed it with until the fourth quarter of the twentieth century.

Today, the traditional naturalistic grand narrative no longer suffices. The global economy demands flexibility; the nation-state ceases to police age boundaries and moves to divest itself of the entitlements (public schooling, mandatory retirement, social security) that were once associated with age. Individuals are encouraged to become more flexible and

self-reliant, charged with the responsibility not only for defining but narrating their own life courses. This is reflected in the explosion of personal memoirs and family histories in the last few decades. In the absence of grand narratives of the life course, everyone seems bent on telling his or her own story.

This new development represents both a freedom and a burden to individuals, and a challenge to experts in the area of life course studies. It requires a new sensitivity to what people are telling us about their lives, new methods that go beyond the old quantitative measures of age to assess subjective experience of aging. It also calls for new theories of aging that take account of the interaction of cultural and biological factors. Historians as well as social scientists now acknowledge that we inhabit not only a physical world but also a symbolic world, and that it is the interplay between the two that determines the history as well as the sociology of the life course and aging.

Historians as well as behavioral scientists have begun to acknowledge that the sources of data they have traditionally relied on to measure aging, mainly physical and behavioral data, tells us only part of what we need to know. They are beginning to explore further the stories that people have told themselves about the aging process and how different subcultures have imagined the life course. For it is through these that we grasp the ways in which people not only think about age but think with age, and how the meanings that are constructed through and around the life course affect the way we live. For we not only live with certain historically defined notions of life course and aging, but live by them. These constitute the source of the meaning as well as the structure of any social system.

Historical studies of the life course and aging are beginning to turn to those humanistic disciplines such as literature and art history traditionally devoted to close analysis of language, narrative, and images. This does not mean that they are discarding the alliances with the behavioral and social sciences that have been so productive to this point, but rather that all the disciplines concerned with the life course and aging have become more aware of the need to examine their epistemological and ontological assumptions.

See also: Aging and Health in Old Age; Aging, Theories of; Childhood: Anthropological Aspects; Family and Kinship, History of; Family as Institution; Households in Anthropology; Kinship in Anthropology; Midlife Psychological Development; Motivation: Life Course and Sociological Perspectives; Youth Culture, Anthropology of; Youth Culture, Sociology of.

Bibliography

Ariès, P., 1960. L'enfant et la vie familiale sous L'Ancien Régime. Plon, Paris.
Ariès, P., 1962. Centuries of Childhood: A Social History of Family Life. Knopf, New York.
Ariès, P., 1981. The Hour of Our Death. Knopf, New York.
Ariès, P., Duby, G., 1987–91. A History of Private Life. Belknap Press of Harvard University Press, Cambridge, MA.
Botelho, L.A., Ottaway, S.R., 2008. The History of Old Age in England, 1600–1800. Pickering & Chatto, London.
Botelho, L.A., Thane, P., 2001. Women and Ageing in British Society since 1500. Pearson Education, Harlow, England.
Chudacoff, H., 1981. How Old Are You? Age Consciousness in American Culture. Princeton University Press, Princeton, NJ.
Cole, T.R., 1992. The Journey of Life: A Cultural History of Aging in America. Cambridge University Press, Cambridge, UK.
Demos, J., 1986. Past, Present, and Personal: The Family and Life Course in American History. Oxford University Press, New York.
Erikson, E., 1963. Childhood and Society, second ed. Norton, New York.
Gillis, J., 1974. Youth and History: Tradition and Change in European Age Relations, 1770—present. Academic, New York.
Hareven, T.K. (Ed.), 1996. Aging and Generational Relations in the Life Course: A Historical and Cross-Cultural Perspective. de Gruyter, New York.
Johnson, P., Thane, P., 1998. Old Age from Antiquity to Post-modernity. Routledge, London.
Kingston, J., 2011. Japan in Transformation, 1945–2010. Longman, Harlow, England.
Klein, H.S., 2004. A Population History of the United States. Cambridge University Press, Cambridge, UK.
Kohli, M., 1985. Die Institutionalisierung des Lebenslaufes. Vierteljahresschrift fuer Soziologie und Sozialpsychologie 1, 1–29.
Laslett, P., 1991. A Fresh Map of Life: The Emergence of the Third Age. Harvard University Press, Cambridge, MA.
Modell, J., 1989. Into One's Own: From Youth to Adulthood in the United States. University of California Press, Berkeley, CA.
Pollock, L.A., 1983. Forgotten Children: Parent–Child Relations from 1500–1900. Cambridge University Press, Cambridge, UK.
Winslow, M., 2001. War, Resettlement, Rooting and Ageing: An Oral History Study of Polish Émigrés in Britain.

Life Course: Sociological Aspects

Robert Crosnoe, Population Research Center, University of Texas at Austin, Austin, TX, USA
Glen H Elder Jr., Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G.H. Elder Jr., volume 13, pp. 8817–8821, © 2001, Elsevier Ltd.

Abstract

The study of individual lives has long been a major focus of social science, but approaches to this field of study have evolved considerably over the last century. Increasingly, the pathways that individuals take through life have been contextualized within the broader social, economic, and cultural changes of society as well as long-term population change and shorter-term historical events. The concept of the life course captures this contextualized nature of human lives, and life course theory presents an orienting frame for studying lives this way. This article gives a basic overview of the life course as both concept and theoretical paradigm.

Studying Individual Lives

As a concept, the life course refers to the socially structured sequences of events and roles that organize and direct the life trajectories of individuals. The term life course also refers to a general theoretical orientation, a framework for studying individual lives by facilitating the questions that need to be asked and guiding how the answers to these questions can be pursued. This perspective, life course theory, emphasizes the dynamic mutual influences between the age-graded sequences of the life course and unfolding patterns of development, adaptation, and experience within economic, social, cultural, and historical contexts. This theory fuses together the insights and approaches of multiple disciplines – especially sociology, psychology, history, and demography – and draws on a data sources that are longitudinal and multi-level to link the individual and society in concrete ways.

Just as human lives are historically embedded, so too is the scientific study of human lives. The clear momentous effects of major events of the early twentieth century (e.g., depressions and wars) on adults and children proved to be highly influential to the formulation and development of life course theory, aided by the emergence of new longitudinal panel studies. As the latter half of the twentieth century played out, population change (e.g., the increasingly old age structure of American society, rising immigration, and differential fertility by race and social class) emphasized the need to examine the full span of life as well as the substantial diversity in lives. Contemporary social change, such as the global economic restructuring that has reshaped work, education, and family roles and opportunities as well as the loosening of norms about gender and sexuality, have injected new ways of thinking about role sequences, age-related expectations of behavior, and the wide array of contexts that shape the basic direction of life trajectories. Over time, research and theory on the life course has grown more refined to capture the increasing complexity of lives in the modern world.

To give an overview of the life course as both a concept and theory, this article explains its basic components (e.g., trajectory, transition, and turning point), including a discussion on related terms that are often used interchangeably despite their distinct meanings (e.g., life span, life cycle, and life history). After describing the emergence of life course theory and its influence on research on the life course, this article turns to five fundamental principles of the theory – lifelong development, timing, linked lives, human agency, and historical time and place – before closing with new directions of research that can work from and build on these principles on the way forward.

Basic Concepts

As already noted, the life course is a dynamic sequence of age-graded events and social roles that are embedded in an evolving social structure. This concept is the centerpiece of life course theory, which is concerned with the trajectories of individuals – the pathways that they take through time – and how these trajectories reflect and reshape the social expectations, norms, and values about individual action and behavior (Elder and Shanahan, 2006). Trajectories play out within specific domains of life, so that an individual life can be thought of as a tapestry of trajectories of education, work, marriage, civic engagement, and other experiences representing engagement with major societal institutions as well as psychological and behavioral trajectories, such as cognitive functioning, mental health, and criminality. Trajectories within a specific domain are often referred to as 'careers,' such as the educational careers of children traversing the K-12 elementary system or the criminal careers of adult men going back to their adolescence.

Embedded within long-term trajectories are transitions, which are the concrete short-term changes in role, status, state, or setting that give trajectories their distinctive form and meaning. Frequently discussed examples of transitions include leaving home, entering the labor force, marrying, and retiring but also more biological events such as going through puberty or entering menopause. Transitions can be quite complex, and many can be internally divided into multiple micro-transitions that link together to produce a major life change. For example, an individual marital trajectory might involve several transitions involving multiple partners, as an individual breaks up with a partner, spends time alone, and re-partners with someone anew. At the same time, the single breakup within

that individual marital trajectory might be decomposed into periods of alienation, separation, divorce, and recovery (Cherlin, 2009).

Importantly, transitional periods are when a major change in life trajectories (i.e., a redirection of individual pathways) are most likely to occur. When such changes are anchored in transitions, those transitions can be labeled as turning points in the life course. Trajectories of crime offer several insights into how turning points work. One of the strongest predictors of whether long-time criminals desist from their criminal activity is whether they marry – a mechanism of social control with power to break antisocial associations (Warr, 2002). As for turning points outside the personal realm, unexpected historical events can also sharply alter criminal trajectories. In one striking example, the recidivism rates of new parolees in New Orleans diverged sharply before and after Hurricane Katrina because that cataclysmic storm powerfully disrupted criminal networks (Kirk, 2009). Moving to another domain, the schooling careers of US children – which can be thought of as long sequences of curricular statuses – illustrate how more institutional transitions can be turning points in individual trajectories. For most youth, the transition from middle school into high school is experienced negatively (e.g., as a source of stress and disruption) or neutrally, but this transition can be quite positive for students who had done poorly socially or academically in their middle schools, as long as they enter their new high schools with few of their middle school classmates. In this way, a quirk of district-directed feeder patterns can allow some struggling youth to turn their educational careers around (Langenkamp, 2010).

These basic concepts can be applied to different levels of the life course; in other words, the normative life course of a society or the specific life course of an individual in that society. First, institutionalized pathways may be established on the population level by laws and social policies of the state or firm, such as when the presence or absence of family leave policies influence how women combine work and parenthood or whether they transition from one role to another. Second, the individual life course is worked out at the micro-level in terms of choices and lines of action, such as when women make decisions about the timing of fertility based on their own expectations about work and current work circumstances. Third, the developmental trajectory of the individual is influenced by biological, cognitive, and other personal characteristics, such as when personality or intellectual traits facilitate or interfere with women's balancing of work and parenting activities (Bianchi, 2000).

For example, school-to-work patterns in industrialized societies involve young adults exiting the educational system, entering the labor force, and then iteratively moving through positions and jobs. These patterns are highly structured by larger social and institutional systems, such as the organization of firms and industries, policies that regulate schools and the labor market, and fluctuations in economic climate. Economic work has demonstrated that young adults who graduate from college during economic downturns have truncated earnings trajectories over subsequent decades because their initial job mismatches mean that they start the cumulative career-building process at a disadvantage (vs if they entered the labor force during another era) and then have to play 'catch up' (Kahn, 2010). On the other side of school-to-work patterns, early disparities in cognitive and academic skills, which reflect both environmental and genetic influences, lead to disadvantages in curricular placements and other schooling positions. These disadvantages compound from year to year and ultimately lead to larger disparities in end-of-school outcomes that then shape future work prospects (Entwisle et al., 2005).

Other concepts are related to the life course, some closely enough so that they are often *used interchangeably with the life course*. Similar to a life trajectory, life cycle describes a sequence of life events from birth to death within any domain, although it more accurately refers to stages of parenthood over the life course – the birth of a first child to the departure of children from the home to the childbearing of the next generation and the transition into grandparenthood. The life cycle is repeated from one generation to the next and links together generations. Similarly, life history typically refers to a lifetime chronology of events and activities as well as social psychological processes that occur within and between them. Life history interview techniques can be retrospective and lead respondents through their own personal histories with a recounting and dating of major transitions, or they can be prospective and focus on the present and look forward to the future (Giele and Elder, 1998). Finally, life span is a term more common in psychology that refers to the temporal frame of study – how much of life is being considered. A life span study extends across a substantial period and links behaviors in at least two stages of life, if not more. As such, life span work stands in contrast to much of the developmental and aging literature, which often focuses on one period or stage at a time. All three of these terms stress the value of studying how individual lives unfold over time and, as such, capture the basic spirit of the life course approach.

The Development of Life Course Theory

The 1960s and 1970s proved to be the seminal years in the emergence of life course ideas. A major source of inspiration in this development was the presence of pioneering longitudinal data collections with children who had grown up decades earlier, samples that had potential for analysis beyond simply following unfolding developmental patterns (see Bruckner and Mayer, 1998). The Oakland Growth Study and Berkeley Guidance Study, which were housed at the University of California–Berkeley, and the Stanford-Terman Study at Stanford University are prominent examples. They had existed for many years and had been studied extensively, but they also had untapped potential, including their long-running span into adulthood and their contiguity to major historical events. As such, they offered new opportunities to study lives as dynamic, socially structured, and historically embedded and to develop a deeper conceptualization than growth-oriented ontogeny.

One foundational advance of this early period was the formulation of life span concepts of development, including plasticity (i.e., the potential for change), agency (i.e., self-direction and control of one's own experiences and settings), and the mutual feedback between individual and context (Lerner et al., 2010). In psychology, these advances led to an emphasis on inter-individual diversity and intra-individual continuity. In sociology, they led to the conceptualization of

some of the core concepts of life course theory and the articulation of processes of cumulative disadvantage and advantage. Life-span thinking on human development occurred initially with little attention to a well-established role-theoretical perspective on human lives, dating back at least to Thomas and Znaniecki's (1918–20) *The Polish Peasant in Europe and America*. Such perspectives provided a way of thinking about socialization, generational succession, and social networks into the 1960s. The central concept at the time was life cycle; it depicted life organization in terms of generational succession through kin relations. Yet, one area in need of further development was timing issues of role entry, exit, and duration that needed to be more explicitly linked to temporality and positioned within historical time. The work of Neugarten (1968) on age norms and age-graded roles facilitated such development. This work documented the variability in life when individuals assumed major social roles and packaged them together. Role timing, sequences, and configurations continue to be influential in the social and behavioral sciences, especially in relation to the evolving transition from adolescence to adulthood in USA and other western countries (Settersten and Ray, 2010), and the decoupling of marriage and parenthood (Edin and Kefalas, 2005).

Another foundational advance during the 1960s and 1970s was the increased understanding of individual experiences in the context of particular historical contexts and events, including Ryder's (1965) conceptualization of cohort as a tool for studying social change. In general, cohort refers to groupings of individuals according to some temporal dimension, such as entry into a system or year of birth. A birth cohort demarcates individuals who experience historical events at the one age from those who experience them at younger or older ages. Cohort effects, therefore, occur when the experience of some historical event differs across cohorts; in other words, according to how old individuals are when it happens. When the Cultural Revolution occurred in China, for example, many adolescents and young adults were 'sent down' to rural areas to be resocialized, but older citizens were not. Being sent down then had lasting effects on the lives of those cohorts that had no real counterpoint in older segments of the population (Zhou and Hou, 1999). Cohorts often differ in norms and values and define themselves in contrast to older and younger cohorts, representing different contexts for coming of age and then aging. For instance, age-period-cohort analyses of substance use have revealed that apparent historical and developmental trends in marijuana use actually reflect the progression of the baby boomer cohorts transitioning into, through, and out of their prime drug-using years. Younger cohorts have turned to other drugs, leading to changes in the overall population distribution of substance use (Miech and Koester, 2012).

These different streams of research and theory were exemplified by the work of Elder on cohorts of children growing up during the Great Depression and traversing adulthood in the context of World War II, work that proved to be seminal to life course theory (see Elder, 1999). Elder began by exploring issues of generational change, socialization, and social roles but soon turned to the analytical and conceptual meaning of age as a means of linking family and individual experience to historical change and identifying age-graded trajectories across the life course. His research explored Ryder's ideas about cohorts by comparing the effects of drastic income loss during the Great Depression on the life experiences of the Oakland cohort members (born 1920–21) with those of a younger Berkeley cohort. As an illustration of cohort effects, the younger boys were more adversely affected by family hardship than the older boys. As an illustration of the value of tracking role sequences and life trajectories, these differences faded in subsequent years as the youth moved through the education and into the labor force, entered military service, and formed families.

From this seminal research, life course theory has continued to be constructed and refined over decades, so that it is now a general framework for studying lives in context that influences research on multiple topics (e.g., education, families, health, and immigration) in multiple disciplines (e.g., sociology, economics, and public health). This theoretical development has both inspired and been influenced by significant advances in methodologies for studying lives dynamically (e.g., growth curve analyses and sequence analysis) and within context (e.g., multilevel modeling and social network studies), the proliferation of long-term longitudinal population studies for public use (e.g., the US Panel Study of Income Dynamics and its numerous European counterparts, such as the Socioeconomic Panel Study in Germany; the US National Longitudinal Study of Youth and the British cohort studies), and the increased emphasis on mixing quantitative and qualitative approaches (Elder et al., 2003).

Principles of Life Course Theory

Life course theory is defined by a set of core principles. The first principle is that *human development is a lifelong process*. In short, individual behavior cannot be fully understood by focusing solely on the specific life stage in question. Where someone is at any one point reflects the legacy of past experience as well as anticipations of the future. For example, overweight youth who become thin adults often have mental health profiles more similar to overweight adults than thin adults, as identities developed in the context of the stigmatization of their weight early in life prove to be difficult to leave behind even when weight has been lost for years (Carr and Jaffe, 2012).

A second principle of life course theory is the importance of *timing*. People and their situations change across the life course, raising the probability that influences also differ by stage of life. At the same time, the meaning and implications of particular life transitions can change with age, with transitions being on or off time and timetables being compressed or slow. Although timing is often illustrated by age-related differences in the experience of role transitions like the entry into grandparenthood or exit from the labor force, biosocial research offers new ways of thinking about timing. For example, genetic influences on problem behavior can switch from being expressed in conduct disorder in childhood to substance use in adolescence and young adulthood as young people move outside the direct control of parents and navigate broader social networks (Dick et al., 2006). As another example, going through puberty earlier than the norm can have negative effects on girls by selecting them into older peer groups and sexualizing them before they are psychologically mature (Mendle et al., 2010).

Finally, the development of socio-emotional areas of the brain temporarily outpaces the development of cognitive capacities in the brain, leading to sensation-seeking and reward-oriented behaviors that often pose physical dangers to young people. The potential impact of peer influences on youth, therefore, depends on how far along they are on this age-related process of brain development (Steinberg, 2008).

Linked lives is the third principle, stressing how the circumstances and experiences of one individual can affect other people to whom they are connected. Although linked lives refers to a wide variety of social and personal relationships, this concept most often is discussed within the context of families. As an illustration, the partnership histories of mothers shapes the behavioral and educational trajectories of their children, with students having more trouble navigating the stratified, cumulative curricula of high schools when their mothers have had much more turnover in their romantic unions. The mechanisms linking lives across generations in this way involve not only the resources that mothers can muster to support their children's education but also their own engagement in parenting (Cavanagh and Fomby, 2012). Going in the other direction, the health and well-being of aging parents can be deeply distressed by the long-term burdens of caring for disabled or unhealthy adult children (Seltzer et al., 2011).

As for the fourth principle, namely *human agency*, the life course is shaped by the choices individuals take and the actions they make, even as the availability of choices and actions is constrained by broader social forces. In educational research, ample evidence suggests that the most important determinant of academic positions in secondary education is the demonstrated ability and performance (e.g., test scores), but also that, even within ability/performance levels, students from more privileged backgrounds manage to obtain more advantaged positions by making more informed and strategic decisions within the general boundaries set by their academic histories (Jackson, 2013). On the other hand, biosocial work suggests that genetic influences are influential determinants of behavior in situations of strong social control (e.g., well-organized communities with dense social ties) but that behavior reflects greater personal agency in situations of low social control (Guo et al., 2008).

All four principles up to this point play out within *historical time and place* – the fifth and final principle. Especially in societies undergoing rapid change, a different birth year exposes individuals to vastly different worlds, with their associated constraints and options. For example, Silbereisen's (2005) work on transitioning to adulthood revealed the powerful influence of reunification of Germany as compared to the years before or after. As another example, pregnant Arab–American women had more negative birth outcomes if they were pregnant in the aftermath of the 9/11 attacks rather than before (Lauderdale, 2006). Some aspects of historical time and place, period effects, are less context-specific but instead encompass influences of social change that are relatively uniform across successive birth cohorts. Divorce is more likely and less stigmatized today than it was decades ago, which alters the experience of divorcing. Similarly, coming out as gay or lesbian is a different experience now than even a few years ago.

Challenges and New Directions

Today, the study of the life course as the link between individuals and society is a prominent field of research activity that connects multiple disciplines, both aided by and encouraging data collection on lives as they unfold within contexts over long periods of time. The development of the field – and of life course theory in particular – is expected to continue, and opportunities for advancing the field are numerous.

First, the task of linking early and later experiences in the life course remains a challenge, even with the explosion of longitudinal data. Studies have traced the persistent effect of early events and experiences to later outcomes, but the chain of linking mechanisms is largely uncharted.

Second, advances in how to capture life experiences, circumstances, and behaviors in more dynamic ways have been significant and similar significant advances in how to link dynamic trajectories across domains (e.g., education, family, and health) are needed. Methodologies such as parallel growth mixture modeling, which produce categories of individuals based on configurations of different trajectories, is an example of how such advances can be made.

Third, biological as well as social events are clearly linked to life patterns, and the biosocial examples mentioned in this article suggest that the interplay of the two is increasingly studied. Such approaches are a growth area in life course research.

Lastly, these future directions are all subsumed within the general need in this field of study for cross-cultural comparisons, both within and between societies. European life course research has led the way with this approach (see Blossfeld, 2009). Such comparisons are the most effective way for considering the social and historical embeddedness of human lives.

See also: Age Structure; Aging, Theories of; Biography and Society; Childhood and Adolescence: Developmental Assets; Life Course in History; Lifespan Development, Theory of; Linked Lives; Transition to Adulthood.

Bibliography

Bianchi, S.M., 2000. Maternal employment and time with children: dramatic change or surprising continuity? Demography 37, 401–414.

Blossfeld, H.-P., 2009. Comparative life course research: a cross-national and longitudinal perspective. In: Elder Jr., Glen H., Giele, Janet Z. (Eds.), The Craft of Life Course Research. Guilford Press, New York, pp. 280–308.

Bruckner, E., Mayer, K.-U., 1998. Collecting life history data: experiences from the Germany life history study. In: Giele, Janet Z., Elder Jr., Glen H. (Eds.), Methods of Life Course Research: Qualitative and Quantitative Approaches. Sage Publications, London, pp. 152–182.

Carr, D., Jaffe, K., 2012. The psychological consequences of weight change trajectories: evidence from quantitative and qualitative data. Economics & Human Biology 10 (4), 419–430.

Cavanagh, S.E., Fomby, P., 2012. School context, family instability, and the academic careers of adolescents: the role of family instability within schools. Sociology of Education 85, 81–97.

Cherlin, A., 2009. The Marriage-Go-Round. Random House, New York.

Dick, D.M., Bierut, L., Hinrichs, A., Fox, L., Bucholz, K.K., Kramer, J., Kuperman, S., Hesselbrock, V., Schuckit, M., Almasy, L., Tischfield, J., Porjesz, B., Begleiter, H., Numberger, J., Xuei, X., Edenberg, H.J., Foroud, T., 2006. The role of GABRA2 in risk for conduct disorder and alcohol and drug dependence across developmental stages. Behavior Genetics 36, 577–590.

Edin, K., Kefalas, M., 2005. Promises I Can Keep: Why Poor Women Put Motherhood before Marriage. University of California Press, Berkeley.

Elder Jr., G.H., Johnson, M., Crosnoe, R., 2003. The emergence and development of life course theory. In: Mortimer, J., Shanahan, M. (Eds.), Handbook of the Life Course. Plenum, New York, pp. 3–22.

Elder Jr., G.H., Shanahan, M., 2006. The life course and human development. In: Handbook of Child Psychology. Wiley, New York.

Elder Jr., G.H., 1999. Children of the Great Depression: Social Change in Life Experience, twenty-fifth anniversary ed. Westview Press, Boulder.

Entwisle, D.R., Alexander, K.L., Olson, L.S., 2005. First grade and educational attainment by age 22: a new story. American Journal of Sociology 110, 1458–1502.

Guo, G., Roettger, M.E., Cai, T., 2008. The integration of genetic propensities into social-control models of delinquency and violence among male youths. American Sociological Review 73, 543–568.

Jackson, M., 2013. Determined to Succeed? Performance, Choice and Education. Stanford University Press, Palo Alto, CA.

Kahn, L.B., 2010. The long-term labor market consequences of graduating from college in a bad economy. Labour Economics 17, 303–316.

Kirk, D.S., 2009. A natural experiment on residential change and recidivism: lessons from Hurricane Katrina. American Sociological Review 74, 484–505.

Langenkamp, A.G., 2010. Academic vulnerability and resilience during the transition to high school: the role of social relationships and district context. Sociology of Education 83, 1–19.

Lauderdale, D.S., 2006. Birth outcomes for Arabic-named women in California before and after September 11. Demography 43, 185–201.

Lerner, R., Lewin-Bizan, S., Alberts, A., 2010. Concepts and theories of human development: historical and contemporary dimensions. In: Bornstein, M., Lamb, M. (Eds.), Developmental Science: An Advanced Textbook, sixth ed. Erlbaum, Mahwah, NJ, pp. 3–49.

Mendle, J., Harden, K.P., Brooks-Gunn, J., Graber, J.A., 2010. Development's tortoise and hare: pubertal timing, pubertal tempo, and depressive symptoms in boys and girls. Developmental Psychology 46, 1341–1353.

Miech, R., Koester, S., 2012. Trends in US, past-year marijuana use from 1985 to 2009: an age–period–cohort analysis. Drug and Alcohol Dependence 124, 259–267.

Neugarten, B.L., 1968. Middle Age and Aging: A Reader in Social Psychology. University of Chicago Press, Chicago.

Ryder, N.B., 1965. The cohort as a concept in the study of social change. American Sociological Review 30, 843–861.

Seltzer, M.M., Floyd, F., Song, J., Greenberg, J., Hong, J., 2011. Midlife and aging parents of adults with intellectual and developmental disabilities: impacts of lifelong parenting. American Journal on Intellectual and Developmental Disabilities 116, 479–499.

Settersten, R., Ray, B., 2010. Not Quite Adults: Why 20-Somethings Are Choosing a Slower Path to Adulthood, and Why It's Good for Everyone. Random House, New York.

Silbereisen, R., 2005. Social change and human development: experiences from German unification. International Journal of Behavioural Development 29, 2–13.

Steinberg, L.D., 2008. A social neuroscience perspective on adolescent risk taking. Developmental Review 28, 78–106.

Thomas, W.I., Znaniecki, F., 1918–20. The Polish Peasant in Europe and America, vols. 1–2. University of Chicago Press, Chicago.

Warr, E.M., 2002. Companions in Crime: Social Aspects of Criminal Conduct. Cambridge University, Cambridge.

Zhou, X., Hou, L.R., 1999. Children of the Cultural Revolution: the state and the life course in the People's Republic of China. American Sociological Review 64, 12–36.

Life Management, Developmental Psychology of

AM Freund, Max Planck Institute for Human Development, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8827–8832, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

How do people manage their lives? Is it possible to identify general processes of life management across the entire lifespan? After a brief historical introduction to the relatively new concept of life management, these questions are addressed by discussing three central approaches to investigating life management in developmental psychology. These three complementary approaches are (1) self-regulation models of life management, (2) models focusing on age-related expectations as guidelines for managing one's life across the lifespan, and (3) models stressing the role of personal goals as building blocks for life management. These three approaches are not mutually exclusive but complementary and converge in stressing the importance of motivational processes and goals for life management across the lifespan.

Brief Historical Perspective on the Concept of Life Management

The term life management is not frequently used in the psychological literature. One of the reasons might be that the active role of the individual in managing his or her life has only entered psychological models after the 'cognitive revolution.' In the following section, a brief historical overview of the emergence of the concept of life management is given, concluding with a tentative definition of the term 'life management.'

Emergence of the Recognition of the Active Role of the Individual in Shaping One's Life

To introduce the notion of 'life management' into the description and explanation of human development implies the assumption that people actively contribute to their developmental pathways and are not entirely under the control of external stimuli or biological processes. Although hardly controversial nowadays, it took a revolution – the 'cognitive revolution' – to introduce this notion into psychology in the 1960s that was dominated by behaviorism of the Watsonian or Skinnerian type. The importance of self-regulation for describing and predicting human behavior was stimulated by research by Mischel (1968). Their research has convincingly shown that situational, external stimuli are but one factor influencing behavior and that, when attempting to understand human behavior, it is necessary to take sources of behavior regulation into account that lie within the individual (i.e., self-regulatory processes).

Developmental psychology has long been dominated by the view that development is best described either in terms of biologically determined maturation (i.e., unfolding of a biological program or blueprint) or as the result of a specific learning history in the sense of classic and operant conditioning (Kazdin, 2012). The latter view is exemplified in Watson's famous statement that if one were to give him a dozen babies he could – by arranging the environmental conditions – produce whatever adults one would wish to see (Watson, 1925). Watson also noted that he was overstating the role of the environment but he wanted to juxtapose the introspective approach to psychology focusing on inner states such as feelings and thinking with the behaviorist position. Although the interaction of biological and environmental factors for development has also been stressed early on, for instance by Piaget, the proactive role of the individual in creating his or her development has only been recognized fairly recently (Baum, 2005). In this perspective, development is viewed as the co-constructive interplay of biology, culture/society, and the individual (Heckhausen, 2006). On the one hand, individuals proactively shape their environment and place themselves into certain contexts. On the other hand, individuals react to their environment by continuously adapting and reacting to their physical, cultural, and social environment.

One important way in which individuals play an active role in their development is by choosing, committing to, and pursuing a set of (life-)goals such as wanting to become a teacher or deciding to have a child. The kind of goals a person develops and commits to is, again, embedded in sociocultural, biological, and phylogenetic factors. The less behavior is under control of these factors, the more degrees of freedom a person has to choose what personal goals he or she wants to pursue and how. In other words, the less societal or biological influences structure and determine which, when, and how to pursue goals, the more influential are self-regulatory processes for life management.

An Attempt to Define *'Life Management'*

The term 'life management' is only rarely used in the psychological literature. When searching the currently most comprehensive database for publications in psychology (PsychInfo 11/2000) for 'life management,' only 72 publications are listed for the past century, mostly referring to coping with stress in various life-domains such as work, family, or health. In this context, 'life management' is mostly used more or less synonymously with 'coping.' Taking a developmental perspective, however, a focus on coping appears too narrow. I propose to use the term life management in a broader sense to refer to how

people actively take part in and shape their developmental trajectories across the lifespan.

Central Approaches to Life Management in Developmental Psychology

Currently, there is no unified model of life management in psychology. Instead, there are a number of approaches to life management that can loosely be clustered into three groups: (1) self-regulation models, (2) models focusing on age-related expectations, and (3) models stressing the role of personal goals. Despite their differences, these approaches have in common that they stress motivational, goal-related processes as central processes of life management across the lifespan.

Self-Regulation Theories

In contrast to 'life management,' self-regulation is a popular concept with more than 1300 entries listed for the past decade in the PsychInfo database (PsychInfo 11/2000). Mostly, self-regulation is used as an explanatory construct intervening between situational/external factors and individual behavior. In these publications, the term self-regulation refers to a variety of different processes located within the individual regulating behavior, such as social comparison, perceived control, emotional control, delay of gratification, or self-efficacy. Despite the heterogeneity of processes that are subsumed under self-regulation, they have one function in common: they are all concerned with internal processes regulating thoughts, feelings, or actions. Very broadly, self-regulation can be defined as 'the extent to which people influence, modify, or control their own behavior' according to goals or standards (Baumeister and Heatherton, 1996, p. 1; Muraven, 2000).

What goals and standards do people adopt? How do people initiate, execute, and end goal-related actions? Feedback-loop models of self-regulation attempt to answer these questions by referring to cybernetic control processes. Wrosch (e.g., 2003) applied cybernetic models to personal goals. Personal goals can be defined as states a person wishes to achieve, avoid, or maintain. In this context, self-regulation comprises the following three aspects. (1) *Setting of goals and standards*: the setting of goals and standards creates a negative discrepancy of the actual and the desired state that the organism tries to overcome by employing goal-relevant means. (2) *Monitoring*: the comparison of the actual and desired state indicates whether the right means are chosen for achieving one's goals/standards or whether other means need to be employed. (3) *Operation* (goal-pursuit): the initiation, maintenance, and ending of actions for attaining one's goals/standards. There is a large empirical basis supporting the importance of each of these processes for describing goal-directed behavior (see Wrosch et al., 2003, for an overview).

One of the features that make this model particularly attractive is that it elaborates general processes of self-regulation that are not content specific and that can be applied equally well to goals of varying levels of abstractness, including short-term goals (e.g., wanting to dine at a nice restaurant) and to long-term goals (e.g., wanting to become the owner of a nice restaurant). Taking a developmental approach to life management, however, this model has one severe limitation: it starts when goals or standards are already set, and does not explain the process of goal-setting itself. How do people develop and select the goals that direct their behavior? This issue is central for the concept of developmental tasks and age-related expectations that will be discussed in the following section.

Age-Related Expectations as Guidelines for Managing One's Life Across the Lifespan

The question, what kinds of goals a person selects at a certain age, is tightly linked with theories about normative development. These theories suggest that individual development is embedded in an age-sequential order of developmental tasks. In contrast to the very general feedback-loop model of self-regulation, took a highly content-specific approach to life management. According to Erikson's developmental model, there is a sequence of eight developmental stages covering the entire lifespan, each of which is characterized by a specific theme or, in Erikson's terms a 'psycho-social crisis.' According to Erikson, progression to a higher stage can only occur after having successfully solved the previous crises. Examples for such crises are 'identity versus identity diffusion' in adolescence and 'generativity versus stagnation' in middle adulthood. In this framework, life management can be viewed as successfully coping with a sequence of salient themes in each life-phase.

Taking a similar approach to life management, Havighurst introduced the concept of 'developmental tasks' referring to 'a task which arises at or about a certain period in the life of the individual, successful achievement of which leads to his happiness and to success with later tasks, while failure leads to unhappiness in the individual, disapproval by society, and difficulty with later tasks' (Havighurst, 1972, p. 2). Havighurst proposes a set of about eight developmental tasks for each of the six phases into which he divides the lifespan from 'infancy' to 'later maturity.' Examples of developmental tasks are 'starting a family' for young adulthood, and 'adjustment to retirement' for old age.

The notion of a universal and unidirectional hierarchy of psychosocial crises or developmental tasks has been criticized because it neglects variability between cultures, between individuals, and within individuals (i.e., there is no uniform development across domains of functioning). There is high agreement, however, that within most societies there exist age-related expectations that have evolved from and reflect biological, societal, and historical influences. Such age-related expectations provide information about a timetable of being 'on-time' or 'off-time' (Heckhausen, 2006; Cain, 2009), and a frame of reference regarding the kinds of goals that should be pursued at a given age.

Age-related expectations reflect in part age-graded social opportunity structures (i.e., the amount of resources, support, and reinforcement a society provides for pursuing certain goals for people of a specific age). Complementarily, certain goals are more difficult to strive for when deviating from age-related expectations because necessary resources and support might be less accessible or even negatively sanctioned. In other words, age-related expectations contain information about the likelihood of goal attainment for a given point in the lifespan.

Motivational psychology has shown that the expectation to succeed is one of the key determinants for selecting a goal. In this sense, age-related expectations function as a guideline for adaptive goal selection because they indicate which goals should best be pursued at which point in the lifespan, and which goals should better be abandoned. Empirical evidence supports the notion that individuals select personal goals that are congruent with the age-related structure of the lifespan.

Taken together, age-related expectations provide a reference standard regarding the kinds of goals that should be pursued at a given point in the lifespan. Moreover, as they reflect age-related opportunity structures, age-related expectations help to guide an individual towards goals for which resources are available and that, therefore, have a higher likelihood of being achieved than 'off time' goals. In this sense, age-related expectations are important guidelines for life management across the lifespan.

Personal Goals as Building Blocks for Life Management Throughout the Lifespan and into Old Age

Action-theoretical approaches view goals as central building blocks of life management because goals structure and direct attention, emotions, and behaviors into particular pathways (Schmuck et al., 2001). Recently, a number of models have stressed the importance of goal-related processes for developmental regulation and life management. In the author's view, the currently most prominent models are the model of selection, optimization, and compensation (SOC-model), the model of optimization in primary and secondary control (OPS-model), and the model of assimilative and accommodative coping. In this section, how these complementary models – that all have received good empirical support – conceptualize life management is briefly presented.

The SOC-model (Baltes & Smith, 2003) posits that, throughout the lifespan, individuals continually seek to successfully manage their lives through the orchestration of three processes of developmental regulation: selection, optimization, and compensation.

The importance of selection of goals is based on the assumption that, throughout the lifespan, biological, social, and individual opportunities and constraints specify a range of alternative domains of functioning. The number of options, usually exceeding the amount of internal and external resources available to an individual, needs to be reduced by selecting a subset of these domains on which to focus one's resources. Selection (i.e., developing, elaborating, committing to goals) gives direction to development by guiding and focusing resources on certain domains of functioning (specialization). Moreover, goals organize behavior into action sequences and reduce the complexity of a given situation as they guide attention and behavior.

According to the SOC-model, by delineating the range of goals on which to focus one's resources, selection is the first important step for successfully managing one's life. To actually achieve higher levels of functioning in the selected domains (goals), however, optimization needs to take place. Optimization denotes the process of acquiring, refining, coordinating, and applying the means or resources that are necessary to reach the goals envisioned. The importance of optimization is illustrated in the research on expertise. This literature has demonstrated the central role of the deliberate practice of skills, encompassing the acquisition of new skills, the refinement of existing skills, and the integration and automatization of skills for achieving peak-performance. The process of optimization addresses the growth-aspect of development.

The third component of the SOC-model is compensation (see also De Frias et al., 2003). Compensation is aimed at counteracting losses in previously available goal-relevant means and involves substitution of lost means by activating unused means or by acquiring new means. The objective is to maintain functioning in the face of losses of means. A second form of managing losses is to restructure one's goals by focusing resources on the most important goals, adapting goal standards, or substituting no longer achievable goals (loss-based selection). The SOC-model posits that loss-based selection is an important process of life management as it allows one to focus or redirect resources adaptively when other means for the maintenance of functioning substituting the loss (compensation) are not available or would be invested at the expense of more promising goals.

According to the SOC-model, although all of these processes have positive developmental functions, it is their orchestration that contributes mostly to adaptive life management in the face of changing developmental opportunities, demands, constraints, and personal needs.

Building on the SOC-model and focusing on control, Heckhausen (2006) propose in their model of optimization in primary and secondary control (OPS-model) two modes of control jointly contributing to life management across the lifespan, primary and secondary control. According to the OPS-model, people have an innate need to control their world and try to maximize their control potential. Primary control encompasses instrumental efforts to modify the environment according to one's goals. Secondary control is employed when primary control strategies directed at the external world are not available or fail. Secondary control focuses on internal aspects such as changing one's goals, engaging in self-protective attributions, or social comparisons. Heckhausen distinguish between selective and compensatory forms of primary and secondary control strategies. Selective primary control refers to the investment of means for goal achievement. Compensatory primary control denotes the employment of external resources (e.g., help of others) or the acquisition of alternative means when available internal resources are insufficient.

Selective secondary control refers to enhanced selectivity of resource investment into existing goals. Compensatory secondary control denotes cognitive reframing of goals such as devaluation of unobtainable goals. According to the OPS-model, life management is only optimized when the four types of control are orchestrated according to an individual's developmental ecology.

A similar view proposed people use two complementary forms of coping to achieve a match between the actual and the desired developmental state, assimilation, and accommodation. Assimilation denotes an actional mode of coping that is characterized by tenaciously pursuing one's goals by modifying the environment according to personal goals. Accommodation

refers to a nonintentional, unconscious mode of coping, namely flexible goal adjustment. Accommodation comprises changing, downgrading, or discarding personal goals or lowering one's level of aspirations to match the environment. People will first use assimilative strategies to close the mismatch between the actual state and personal goals. When assimilative attempts fail, a gradual shift to the accommodative mode occurs, first by investing additional goal-related activities, then by engaging external aids or support systems, and, finally, by adjusting standards or even disengaging from goals (e.g., Brandtstädter et al., 2002).

The three models reported above converge in the importance they place on goals and motivational processes as building blocks of life management. Moreover, these models suggest that key processes of life management across the lifespan lie in the adaptation to developmental ecologies by opportunity-oriented selection, focused investment of resources, and compensation of losses. Age-related changes in the availability and efficiency of resources necessitate the flexible use of goal related strategies.

Starting out to be highly positive in childhood through early adulthood, the ratio of gains to losses becomes less favorable with age. Thus, with increasing age, individuals need to allocate more resources into the maintenance of functioning and compensation of losses than into growth. Life management, following these models, encompasses optimizing developmental potential and simultaneously compensating for losses.

Outlook

Different lines of research inform us that one important way in which individuals manage their lives is by developing, committing to, and pursuing a set of (life-) goals. What goals a person selects and pursues is embedded in sociocultural, biological, and phylogenetic factors.

An interesting question awaiting empirical investigation is the relative importance of these factors across the lifespan. During childhood, adolescence, and young adulthood, biological constraints and social expectations regarding the kinds of goals to be pursued are relatively strong. In late adulthood and old age, there are fewer normative age-related expectations. This relatively greater social freedom might strengthen the impact of personal goals for life management. At the same time, in old age resources needed for shaping one's environment according to personal goals decline. Therefore, it seems particularly crucial to focus on goals that are both personally very important and for which resources are still available. Future investigations of the interaction of the various factors influencing the selection, pursuit, and maintenance of goals across the lifespan might increase our understanding of how people manage their lives.

See also: Aging, Theories of; Childhood and Adolescence: Developmental Assets; Life Course: Sociological Aspects; Lifespan Development, Theory of; Personality and Adaptive Behaviors; Self-Regulation in Adulthood.

Bibliography

Baltes, Paul B., Smith, Jacqui, 2003. New frontiers in the future of aging: From successful aging of the young old to the dilemmas of the fourth age. Gerontology 49 (2), 123–135.

Baum, William M., 2005. Understanding behaviorism: Behavior, culture, and evolution.

Baumeister, R.F., Heatherton, T.F., 1996. Self-regulation failure: an overview. Psychological Inquiry 7, 1–15.

Brandtstädter, Jochen, Rothermund, Klaus, 2002. The life-course dynamics of goal pursuit and goal adjustment: A two-process framework. Developmental Review 22 (1), 117–150.

Cain, Leonard D., 2009. Life course and social structure 31.

De Frias, Cindy M., Dixon, Roger A., Bäckman, Lars, 2003. Use of memory compensation strategies is related to psychosocial and health indicators. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 58 (1), P12–P22.

Havighurst, Robert J., 1972. Developmental Tasks and Education, third ed. Davis McKay, New York.

Heckhausen, Jutta, 2006. Developmental regulation in adulthood: Age-normative and sociostructural constraints as adaptive challenges. Cambridge University Press.

Kazdin, Alan E., 2012. Behavior modification in applied settings. Waveland Press.

Mischel, Walter, 1968. Personality and Assessment. Wiley, New York.

Muraven, Mark, Baumeister, Roy F., 2000. Self-regulation and depletion of limited resources: Does self-control resemble a muscle? Psychological bulletin 126 (2), 247.

Schmuck, Peter, Sheldon, Kennon M., 2001. "Life goals and well-being." Toward a Positive Psychology of Human Striving. Hogrefe, Seattle.

Watson, John B., 1925. Behaviorism. Norton, New York.

Wrosch, Carsten, et al., 2003. The importance of goal disengagement in adaptive self-regulation: When giving up is beneficial. Self and Identity 2 (1), 1–20.

Young, Mary E., 2002. From early child development to human development: Investing in our children's future. World Bank Publications.

Life Table

Jan M Hoem, The Max Planck Institute for Demographic Research, Rostock, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, modern statistical language is used to explain the basic notions of the oldest tool of demographic analysis, namely the life table. Following a long tradition in which demographers seek to establish sensible results out of a suboptimal data set, a real-life example is used to describe how one might produce a simple life table in the presence of digital preference (age heaping, grouped time reporting). An extension to multiple-decrement (or competing-risk) life tables is also provided.

The life table is perhaps the oldest tool developed for the analysis of survival patterns in human and other populations. Its roots go back at least to the famous book on the English Bills of Mortality that John Graunt published in 1662, a book that is sometimes cited as the beginning of systematic statistical science. Life-table techniques are described in detail in most introductory textbooks on the methods of actuarial statistics, biostatistics, demography, and epidemiology (see, e.g., Chiang, 1984; Elandt-Johnson and Johnson, 1980; Manton and Stallard, 1984; Hinde, 1998; Preston et al., 2001). In this article the basic notions of life-table construction are explained with a focus on issues that are sometimes underemphasized in other accounts. More of the underlying mathematical theory can be found in Hoem (1998) and in the textbooks. For the history of the topic, consult Seal (1977), Smith and Keyfitz (1977), and Dupâquier (1996).

A Cohort Life Table

Classical presentations of life-table techniques typically start with data in a format like that of the first four columns in **Table 1**. These particular data come from the Eritrean Demographic and Health Survey of 1995 and show the survival up through their fourth birthday of 7098 girls whose births and deaths were recorded in the survey. This data set is different from normal textbook examples in two ways. First, the data are given for intervals of single-month segments instead of the usual single-year intervals. Second, the present data are subject to strong heaping in the reported age at death for the girls that died. In part this is due to a typical inherent inaccuracy of age reporting in societies like the data source at the time of data collection, and in part due to the fact that (probably for that very reason) only integer years of age attained was asked except for very young ages at death. By contrast, an individual child's age when observation ended (age at censoring) could be computed accurately to the month, for in these data a child's record was censored only because the child's parent was interviewed; the age of a live child could then be calculated from its date of birth. This unconventional data set is used in order to illuminate how a number of practical issues that confront a demographer can be handled in a first analysis of mortality. A more complete investigation would subsequently use intensity-regression methods from event-history analysis (*see* Event History Analysis: Applications), as was done by Woldemicael (1999).

If there had not been any heaping in the reported ages at death, one would calculate $R_x = S_x - \frac{1}{2}(D_x + W_x)$ as an approximate number of person-months of exposure at age x. Here S_x is the number of children who survive to exact age x months, D_x is the number of deaths recorded between exact ages x and x+1 months, and W_x is the corresponding number of records that were censored during this interval. The approximation consists in assigning deaths and interviews to in the middle of the month on average. It has become known as the 'actuarial method.' The next step would be to compute occurrence/exposure rates D_x/R_x (*see* Demographic Techniques: LEXIS Diagram). The age heaping is abundantly evident in the plot of the 'raw' death rates (converted to 'deaths per 1000 person-years') in **Table 1**.

Such age heaping must have been caused by a shift of the real age at death to a reported near-by 'round' anchor age like 3 months, 6 months, or multiples of 12 months. The deaths in some interval $[x_0-a,\ x_0+b]$ around each anchor age x_0 have largely been assigned to the anchor age instead of to each age in the interval, but the total number of deaths $\sum_{t=-a}^{b} D_{x_0+t}$ in such an interval should be largely correct. Similarly, the total exposures $\sum_{t=-a}^{b} R_{x_0+t}$ in the interval should be about right, even though the individual term in the latter sum may be wide off the mark. Under these circumstances it is more sensible to compute death rates for groups of reported ages and list them in the row for the corresponding anchor age. By this reasoning,

$$m_3 = \sum_{x=2}^{4} D_x / \sum_{x=2}^{4} R_x$$

is entered for age 3 months,

$$m_6 = \sum_{x=5}^{7} D_x / \sum_{x=5}^{7} R_x$$

is entered for age 6 months, and so on, as indicated in column 7 of **Table 1**. For the intervening ages x, rates m_x have been computed by linear interpolation. The corresponding curve of rates is given with the label 'grouped' in **Table 1**.

The rest of the life-table computations are based on the following simple mathematical theory. Each individual child i has a lifetime T_i with some distribution function $F(t) = P\{T_i \le t\}$ (with $F(0) = 0$), a probability density $f(t) = \mathrm{d}F(t)/\mathrm{d}t$,

Table 1 Life-table computations for Eritrean girls aged 0 through 4 years

Age x in months	*Observed survivors*	*Deaths* D_x	*Censored* W_x	*Exposures* R_x	*Raw rate* D_x/R_x *per 1000*	*Age grouping (months)*	*Grouped rates* m_x	*Death probab.* $1000q_x$	*Life-table survivors* l_x
0	7098	208	15	6986.5	357.26	0	357.26[a]	29.33	100 000
1	6875	48	39	6831.5	84.32	1	84.32[a]	7.00	97 067
2	6788	36	39	6750.5	64.00		70.01[b]	5.82	96 387
3	6713	43	36	6673.5	77.32	2–4	55.70[a]	4.63	95 826
4	6634	14	29	6612.5	25.41		56.00[b]	4.66	95 383
5	6591	21	22	6569.5	38.36		56.30[b]	4.68	94 939
6	6548	50	42	6502	92.28	5–7	56.60[a]	4.71	94 494
7	6456	21	23	6434	39.17		54.98[b]	4.57	94 050
8	6412	26	31	6383.5	48.88		53.35[b]	4.44	93 620
9	6355	19	26	6332.5	36.00		51.73[b]	4.30	93 204
10	6310	16	23	6290.5	30.52		50.11[b]	4.17	92 803
11	6271	10	37	6247.5	19.21		48.48[b]	4.03	92 417
12	6224	144	30	6137	281.57	8–17	46.86[a]	3.90	92 044
13	6050	6	34	6030	11.94		45.45[b]	3.78	91 685
14	6010	7	31	5991	14.02		44.05[b]	3.66	91 339
15	5972	3	15	5963	6.04		42.64[b]	3.55	91 004
16	5954	5	35	5934	10.11		41.24[b]	3.43	90 681
17	5914	3	31	5897	6.10		39.83[b]	3.31	90 370
18	5880	25	34	5850.5	16.40		38.43[b]	3.20	90 071
19	5821	5	24	5806.5	10.33		37.02[b]	3.08	89 783
20	5792	2	28	5777	4.15		35.62[b]	2.96	89 506
21	5762	0	16	5754	0.00		34.21[b]	2.85	89 241
22	5746	0	29	5731.5	0.00		32.81[b]	2.73	88 987
23	5717	2	18	5707	4.21		31.40[b]	2.61	88 744
24	5697	148	25	5610.5	316.55	18–30	30.00[a]	2.50	88 512
25	5524	0	34	5507	0.00		29.28[b]	2.44	88 291
26	5490	0	32	5474	0.00		28.57[b]	2.38	88 076
27	5458	0	31	5442.5	0.00		27.85[b]	2.32	87 866
28	5427	0	27	5413.5	0.00		27.14[b]	2.26	87 663
29	5400	0	35	5382.5	0.00		26.42[b]	2.20	87 465
30	5365	0	33	5348.5	0.00		25.71[b]	2.14	87 272
31	5332	0	39	5312.5	0.00		24.99[b]	2.08	87 086
32	5293	0	28	5279	0.00		24.27[b]	2.02	86 904
33	5265	0	36	5247	0.00		23.56[b]	1.96	86 729
34	5229	0	24	5217	0.00		22.84[b]	1.90	86 559
35	5205	0	28	5191	0.00		22.13[b]	1.84	86 394
36	5177	109	22	5111.5	255.89	31–42	21.41[a]	1.78	86 235
37	5046	0	38	5027	0.00		20.44[b]	1.70	86 081
38	5008	0	20	4998	0.00		19.48[b]	1.62	85 935
39	4988	0	32	4972	0.00		18.51[b]	1.54	85 795
40	4956	0	31	4940.5	0.00		17.54[b]	1.46	85 663
41	4925	0	30	4910	0.00		16.57[b]	1.38	85 538
42	4895	0	30	4880	0.00		15.60[b]	1.30	85 420
43	4865	0	24	4853	0.00		14.63[b]	1.22	85 309
44	4841	0	26	4828	0.00		13.66[b]	1.14	85 205
45	4815	0	18	4806	0.00		12.70[b]	1.06	85 108
46	4797	0	20	4787	0.00		11.73[b]	0.98	85 018
47	4777	0	16	4769	0.00		10.76[b]	0.90	84 935
48[c]	4761	46	20	4728	116.75	43–54	9.79[a]	0.82	84 859

[a]Computed as (sum of deaths)/(sum of exposures) for age groups indicated. (Used for rows where age groups are indicated.)
[b]Computed by linear interpolation.
[c]Additional rows for ages 49–60 months have been deleted.

and a 'force of mortality' or 'death intensity' $\mu(t) = f(t)/\{1 - F(t)\}$. The various lifetimes T_i are independent of each other and they have the same distribution $F(t)$. Simple integration gives $F(x) = 1 - \exp\{-\int_0^x \mu(t)\mathrm{d}t\}$. The death probability is.

$$q_x = P\{T_i \le x + 1 | T_i > x\} = \{F(x+1) - F(x)\}/F(x)$$

from which is derived

$$q_x = 1 - \exp\left\{-\int_0^1 \mu(x+t)\mathrm{d}t\right\} \qquad [1]$$

If one approximates the function $\mu(x+t)$ by the constant m_x for $0<t<1$ (the assumption of piecewise constancy), then (1) gives $q_x = 1 - e^{-m_x}$, which has been used to compute the next-to-last column in Table 1. Finally, a life-table 'survival function' is defined as $\ell(x) = \ell(0)\{1 - F(x)\}$, with $l(0) = 100\,000$. This is the expected number of survivors to exact age x out of an original cohort of $\ell(0)$ initial individuals. Simple manipulation with the definition of q_x gives

$$\ell(x+1) = \ell(x)(1 - q_x), \quad \text{for } x = 0, 1, \ldots \qquad [2]$$

This is a practical recursive formula by which we have computed the final column in Table 1. From the final entry in the $\ell(x)$ column in Table 1 we see that more than 15% of the original cohort of girl babies would die before age 4 (years) according to our data, because $1 - 0.848\,59 = 0.151\,41$.

The mean number of months lived between birth and age n is

$$\mathring{e}_{0:\overline{n}|} = E\min(T_i, n) = \int_0^n \ell(t)\mathrm{d}t/\ell(0) \qquad [3]$$

If one partitions the integral here into a sum of n terms $\int_{x-1}^{x} \ell(t)\mathrm{d}t$ for $x = 1, 2, \ldots, n$, use the trapezoidal rule of numerical integration to see that this integral is approximately equal to $\frac{1}{2}\{\ell(x-1) + \ell(x)\}$, and collect terms suitably, we get the numerical approximation

$$\mathring{e}_{0:\overline{n}|} \simeq \left\{ \sum_{x=0}^{n} \ell(x) - \frac{1}{2} - \frac{1}{2}\ell(n) \right\} / \ell(0) \qquad [4]$$

The girls in the Eritrean data set lost 5 months of life on average before they reached their fourth birthday, for according to eqn [4] with $n = 48$ months they experienced a mean lifetime of 43 months.

Other Life-Table Procedures

Life-table construction consists in the estimation of parameters and tabulation of functions like those above from empirical data. The data can be for age at death to individuals, as in the illustrative example, but they can also be observations of duration until recovery from an illness, of intervals between births, of time until breakdown of some piece of machinery, or of any other positive duration variable. In general, a T_i is the duration until occurrence of some event that ends individual survival in a given status. The method of estimation depends on the character of the data. If accurate (ungrouped) individual-level data are available, then the Kaplan–Meier estimator (*see* Event History Analysis: Applications) can be used to estimate $\ell(x)$ for all relevant x, and estimates of the other life-table functions can then be computed subsequently. Alternatively a segment of the Nelson–Aalen estimator can be used to estimate each integral $\int_0^1 \mu(x+1)\mathrm{d}t$.

Equation [1] can then be used to estimate q_x for each integer x, and the rest of the computations follow suit. Sometimes the force of mortality is represented by some function $h(x;\theta)$, where θ is a vector of parameters. For instance, actuaries often use the classical Gompertz–Makeham function $h(x; a, b, c) = a + bc^x$ for the force of mortality in their life tables (*see* 'Demographic Models; Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems).

From any given schedule of death probabilities $q_0, q_1, q_2, \ldots$, the $\ell(x)$ table is easily computed using the recursive eqn [2]. Much of the effort in life-table construction is, therefore, concentrated on providing a schedule $\{q_x\}$.

So far, this account has assumed that the data come from a group of independent individuals who have all been observed in parallel, essentially a cohort that is followed from a significant starting point (namely from birth in our empirical example) and that is diminished over time due to 'decrements' ('attrition') caused by the risk in question and also subject to reduction due to censoring. The empirical example displayed data only for the first 4 years of life of such a cohort. One would have had to follow the cohort until the end of the cohort's total life to produce a 'complete cohort life table.' It is more common to compute a q_x schedule from data collected for the members of a population during a limited time period and to use the mechanics of life-table construction to produce a 'period life table' from the q_x values. If real mortality patterns are tied to cohorts, individuals who live at widely differing ages in the observational period do not normally have the same risk structure, and the period taken by the life table to reflect the patterns of a 'synthetic (fictitious) cohort' exposed to the risks of the period at their various ages.

Whichever way the life-table survivor column has been produced, eqn [4] can always be used to calculate mean numbers of time-units lived under the risk for which the life-table was constructed. If the time unit is a year and n is chosen so large that no one lives more than n years (symbolized by letting $n = \infty$ and $\ell(\infty) = 0$), $\mathring{e}_{0:\overline{\infty}|}$ is called the 'life expectancy' of the life table and is normally denoted $\mathring{e}_0$. The subscript 0 indicates that the calculation is made for a newborn individual. Similar computations can be made at any age x and one gets expected remaining lifetimes of the form $\mathring{e}_{x:\overline{n}|} = \int_0^n \ell(x+t)\mathrm{d}t/\ell(x)$, with approximation formulas similar to eqn [4].

Multiple-Decrement Tables

When suitable data are available, the risk of death (or force of mortality) may be partitioned according to mutually exclusive causes of death. Let $\mu_k(t)$ be death intensity or force of mortality for cause k at age t, and let the total force of mortality be $\mu(t) = \sum_k \mu_k(t)$. Then the probability that an individual who is alive at age x will die from cause k before age $x+1$ is

$$q_x^{(k)} = \int_0^1 {}_tp_x \mu_k(x+t)\mathrm{d}t \qquad [5]$$

where ${}_tp_x$ is the probability of surviving to age $x+t$ for an individual who is alive at age x. For given risk intensities $\{\mu_k(\cdot)\}$, $q_x^{(k)}$ can be computed by numerical integration in eqn [5]. A corresponding column of $q_x^{(k)}$ values may then be added to the life table for each k, and other cause-specific life-table

functions may also be computed. Note that $q_x^{(k)}$ is influenced by risk intensities other than $\mu_k(\cdot)$, for

$$ {}_tp_x = \exp\left\{ -\sum_j \int_0^t \mu_j(x+s)\mathrm{d}s \right\} $$

where the exponent depends on all $\mu_j(\cdot)$, including those for $j \neq k$.

Several further life-table functions can be defined by formal reduction or elimination of one or more of the risk intensities in formulas like eqn [5]. In particular, a 'single-decrement life table' can be computed for each cause k in order to show what the life table would look like in the hypothetical case where this cause was the only one operating in the study population and where it did so with the risk function estimated from the data. The purpose is to see the 'pure' effect of the risk in question on a fictitious cohort without interference from other causes. (The decrement probabilities and other features of the single-decrement life table do not depend on any other intensities than the one for which the table is computed.) This does not mean that one should believe that the total risk intensity can actually be reduced to the risk in focus or that this risk operates independently of other causes.

Most single-decrement life-table functions have straightforward interpretations. If, for instance, $\ell_k(x)$ denotes the corresponding survival function, then $\ell_k(x)/\ell_k(0)$ is the probability of surviving to age x in the fictitious table. Suppose, for example, that the study population is a group of unpartnered individuals subject to the competing risks of marital and nonmarital union formation as well as to the risk of death. One can then compute a single-decrement life table based on the risk of transition into a nonmarital union alone and find the probability that an individual would end up in such a union by age 50 years, say, if marriage were no alternative and death did not operate, even though in reality marriages continue to be formed by group members and unpartnered individuals continue to die. In such a table one may get $\ell_k(\infty) > 0$ because the event never occurs to some individuals, even if it operates alone. (We will all die but some individuals never enter a nonmarital partnership even if they live to age 100 years.)

Life expectancies $\mathring{e}_{15:\overline{n}|}$ computed in the three-decrement life table would be interpreted as the mean number of years (say) lived in the unpartnered state between ages 15 and $15+n$ years. In the single-decrement life table for nonmarital-union formation, one possibility is to compute the mean age of entry into a nonmarital union only for those who do enter such a union by age n, that is, the conditional mean.

$$ \int_0^n \ell_k(x)\mathrm{d}x / \{\ell_k(0) - \ell_k(n)\} \qquad [6] $$

Conversely one can compute a 'cause-deleted life table' by eliminating one or more of the cause intensities in the formulas. To 'remove' cause k, introduce $\mu_{-k}(t) = \mu(t) - \mu_k(t)$ and construct a 'normal' life table by replacing $\mu(t)$ by $\mu_{-k}(t)$ everywhere. This would show what the life table would look like if it were possible to eliminate cause k without changing the risk of any other cause. It could be used for example to see how partnering probabilities would appear if the partnering process were undisturbed by mortality. One would then remove the force of mortality from the life-table formulas and operate only with the sum $\mu_{-k}(t)$ of the intensities of marital and nonmarital union formation in the computations. Note that this is again a purely hypothetical construct; in fact it is counterfactual for mortality cannot of course be completely removed.

Life expectancies can be calculated for a cause-deleted life table as for any other life table. If $\ell_{-k}(x)$ is the table's survival function and $\ell_{-k}(\infty) = 0$, then a formula for the cause-eliminated life expectancy is $\mathring{e}_0^{(-k)} = \int_0^\infty \ell_{-k}(x)\mathrm{d}x/\ell_{-k}(0)$. The classical applications were to cause-of-death removal, and the difference $\mathring{e}_0^{(-k)} - \mathring{e}_0$ was interpreted as the (fictitious) gain in life expectancy produced by removal of cause k. For example, Westergaard (1907) found that according to the English life tables of 1881–90, life expectancy for females would increase by 2.7 years if tuberculosis deaths could be eliminated, by a year if cancer deaths could be removed, and by 1.1 years if deaths due to diarrhea and dysentery could be deleted separately. The effects were not additive, and if all three cause-of-death groups could be removed, the gain would be 5.6 years and not 4.8. The life expectancy for females in the original table was $\mathring{e}_0 = 47.2$ years.

See also: Age Structure; Demographic Models; Demographic Techniques: LEXIS Diagram; Event History Analysis: Applications; Motivation: Life Course and Sociological Perspectives.

Bibliography

Chiang, C.L., 1984. The Life Table and Its Applications. Krieger, Malabar, FL.

Dupâquier, J., 1996. L'invention de la table de mortalité. Presses universitaires de France, Paris.

Elandt-Johnson, R.C., Johnson, N.L., 1980. Survival Models and Data Analysis. Wiley, New York.

Graunt, J., 1662. Natural and Political Observations Made upon the Bills of Mortality. Roycroft, London.

Hinde, A., 1998. Demographic Methods. Arnold, London.

Hoem, J.M., 1998. The life table. In: Armitage, P., Colton, T. (Eds.), Encyclopedia of Biostatistics. Wiley, Chichester, pp. 2235–2239.

Hoem, J.M., 2001. Demographic analysis: Probabilistic approach. In: Neil J. Smelser, Paul B. Baltes (Eds.), International Encyclopedia of the Social and Behavioral Sciences, first ed., vol. 5, pp. 3428–3432.

Manton, K.G., Stallard, E., 1984. Recent Trends in Mortality Analysis. Academic Press, Orlando.

Preston, S.H., Heuveline, P., Guillot, M., 2001. Demography: Measuring and Modeling Populations. Blackwell, Oxford, UK.

Seal, H., 1977. Studies in history of probability and statistics, 35: multiple decrements or competing risks. Biometrika 64 (3), 429–439.

Smith, D., Keyfitz, N., 1977. Mathematical Demography: Selected Papers. Springer-Verlag, Heidelberg, Germany.

Westergaard, H., 1907. The horoscope of the population in the twentieth century. In: Compte-Rendu de la XI Session de l'Institut international de Statistique á Copenhague du 26 au 31 aout 1907.

Woldemicael, G., 1999. Infant and Child Mortality in Eritrea: Levels, Trends, and Determinants (Ph.D. thesis). Stockholm University.

Lifelong Learning and Its Support with New Technologies

Marco Kalz, Open University of the Netherlands, Heerlen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G. Fischer, volume 13, pp. 8836–8840, © 2001, Elsevier Ltd.

Abstract

This article provides an overview of the use of new technologies for lifelong learning. While in the past learning technologies were mostly provided by educational institutions to support a specific lifetime or shorter learning episodes, nowadays more personal technologies are used for lifelong learning to support self-organized learning. Four important developments are introduced in this article, namely, open learner models and learning analytics, learning networks and networked learning, open educational resources and practices, and last but not least, mobile and contextualized learning. The state of the art in these research fields is summarized and the future potential and requirements for lifelong learning are highlighted.

Introduction

Lifelong learning is a concept that integrates a wide variety of learning formats and is usually seen as education throughout the lifetime. Lifelong learning activities can happen in a formal context (organised by an educational institution), it can happen outside an educational institution (nonformal), or it can happen accidentally without previous planning (informal). Any purposeful learning activity undertaken on an ongoing basis targeted at the increase of knowledge, skills, and competences can thus be seen as lifelong learning. Longworth (2003) stresses the importance of lifelong learning for the twenty-first century because of global demographics, environmental imperatives, the pervasive access to information due to new technologies, and the speed of innovation in science and technology. Due to these factors it becomes, more and more, a necessity to not only equip people with knowledge in educational institutions but to also prepare them to update their knowledge, skills, and competences and take responsibility for learning throughout their lives. The author identifies the dismantling of barriers to lifelong learning as an important action point to be addressed by research and developmental activities and by society as a whole. Among others the author identifies the following barriers to lifelong learning:

- Poor family culture of learning
- Lack of finance to participate in lifelong learning
- Learning providers not geared to the needs of learners
- Poor information services for attracting people to learning
- Distance from the educational provision
- Lack of facilities to study at home
- Perception that the benefits system discourages learning

These identified barriers are also confirmed by recent statistics from the Labor Force Survey of the European Commission (Eurostat, 2012). This survey shows that in Europe between 2006 and 2011, there was a decrease in participation in lifelong learning activities. Participants in this survey mention access, time, place, and lack of personalization as barriers to involvement in lifelong learning activities. This survey shows that there is a gap between the societal relevance of lifelong learning and the daily practices of professionals. An overall fragmentation can be added as an additional barrier to lifelong learning.

While learning technologies have been traditionally offered by educational institutions with a focus on supporting short learning episodes, little attention has been given to support the complex array of contexts, life spans, and the individual characteristics of learners. Koper and Tattersall (2004) argue that in the past, the timescales and the episodic and multi-institutional nature of lifelong learning were not reflected in mainstream learning technologies.

In the field of technology-enhanced learning (TEL), a number of research directions have been developed in the last few years that help to address the problems lifelong learners are facing today. In this article I introduce the current state-of-the-art of new technologies for lifelong learning. For this purpose four research topics are introduced that contribute to the dismantling of barriers for lifelong learning. These new technologies are open learner models and learning analytics; learning networks and networked learning; open educational resources (OER) and practices; and last but not least, mobile and contextualized learning. In the following paragraphs, I introduce the current state-of-the-art in these research topics and discuss their contribution in addressing barriers for lifelong learning. Finally, I provide an outlook on future research and development as well as the implications for policy frameworks for lifelong learning.

Open Learner Models and Learning Analytics

One of the challenges in the field of TEL is the modeling and recognition of the activities and contexts of learners. Since lifelong learners might constantly change their learning context, location, goals, learning environments, and also learning technologies, the recognition of the current situation of the learner with the goal to personalize and adapt the learning environment is not only a challenging but also a promising research direction. A lifelong learner might start his day during travel, with the reading of a work-related textbook on his tablet computer, continue during work with a discussion of a specific problem in a professional social network, and join in the evening an online master class about a new topic in

which he would like to build expertise. These short learning episodes during 1 day are a representative picture of lifelong learning as a whole. Learners are active in different learning contexts, in different learning formats, and with different learning technologies. The analysis, construction, and recognition of the learner context have been treated traditionally under the topic of learner models from the perspective of adaptive hypermedia (AH) (Brusilovsky et al., 2007). Based on an analysis of the activities and contexts of learners, algorithms have been developed that are able to predict the behavior of a learner, to provide guidance for the learning process, or to personalize the presentation of learning content. But there is one important limitation in this work. The data collection and rationale governing these data work well in closed contexts like a dedicated electronic learning environment (e.g., learning management system/course management system). But there are limitations when it comes to integrating data from different learning contexts as presented above. This problem has been discussed as the 'open corpus problem' (Brusilovsky and Henze, 2007). Open corpus AH do not work on the closed set of resources known at the design time but rather, work on the assumption that the learning resources and learning context constantly change and evolve.

To address this problem and to enable personalization, several initiatives have been started for lifelong learner modeling. A learner model is the sum of all knowledge that a software system has about a learner. This model gets constantly updated during learning activities and should thus reflect the current state of knowledge of the learners. While traditional learner models were closed and only used by the technological infrastructure, recently open learner models for the use of the learner have been proposed and evaluated. The current state of such a learner model is presented to the user and there are different ideas about the benefits and the management of such an open learner model. Bull and Kay (2010) describe independent open learner models as being fully controlled by learners instead of learner models controlled by a system or in a cooperative manner. This approach addresses the personalization barrier of lifelong learning on a technological level and gives learners control about their digital representation in a learner model.

Kay (2008) takes this approach one step further and proposes to link and aggregate learner models from different contexts into long-term learning models that integrate the different contexts of lifelong learners. Besides the issue of collecting and storing these data, she also proposes to integrate representations of the lifelong learner models into the working contexts of lifelong learners. This direction is also explored in a relatively young research topic in the field of TEL called 'Learning Analytics' (Greller and Drachsler, 2012). Learning analytics exploit the potential of available large data sets (about learners, courses, behavior, etc.) to provide feedback in the form of meaningful visualizations to different stakeholders (learner, teacher, organization). While several initiatives in the field remain in the traditional boundaries of educational segments or institutions, some authors propose a more open approach to learning analytics that is potentially useful in the context of lifelong learning (Buckingham Shum and Fergusson, 2012). Approaches in which learners' activities are monitored and collected in a number of different TEL environments have the special potential to help learners to connect different learning contexts (Romero-Zaldivar et al., 2012). Thus, open learner models and learning analytics represent new technological solutions that can help learners to connect their data from learning activities in different learning environments and contexts, and contribute to the dismantling of the personalization barrier of lifelong learning.

Current research focuses on the potential of open learner models and learning analytics to increase awareness about the learning process itself and its effects on meta-cognition. On the other hand, the evaluation contexts chosen mainly focus on the context of educational institutions, and the transfer of results to an authentic lifelong learning context cannot be done without limitations. One limitation for both approaches presented here is a relatively high visual literacy requirement in order to make use of the data representation provided. Therefore, some of the issues in open learner models and learning analytics are related to advancement of visualization approaches for learners and learning activity data.

Learning Networks and Networked Learning

The networked society and the potential to use network effects for learning and competence development have motivated new initiatives to apply this potential to the context of lifelong learning. Koper et al. (2005b) propose the following model of a learning network: "An ensemble of actors, institutions, and learning resources, which are mutually connected through and supported by, information and communication technologies in such a way that the network self-organizes and thus gives rise to effective lifelong learning." **Figure 1** shows this conceptual model (Koper and Tatersall, 2004).

The model consists at its core of collections of learning activities around a specific domain. Learners in this network are connected via their shared learning objective (called the target) and they can optionally receive assessment and certification. The authors derive several requirements for the technical development of learning networks for lifelong learning. These requirements are as follows:

- Learners should be responsible for their own learning and they should be able to conduct learning activities in different contexts at the same time. Thus, the learners should be put at the center of the stage and should be able to take self-directed action.
- The focus of the learning network should be to enable learning and not to support instruction.
- Learners should be enabled to connect formal, nonformal and informal learning activities.
- Learning networks should allow access to learning resources and activities from different providers.

This model has been translated into an architectural model to allow its technical implementation. While initially peer-to-peer technologies were used for implementation (Koper et al., 2005), later this model was integrated based on the service-oriented architecture paradigm. A number of technical services have been developed to support learners in learning

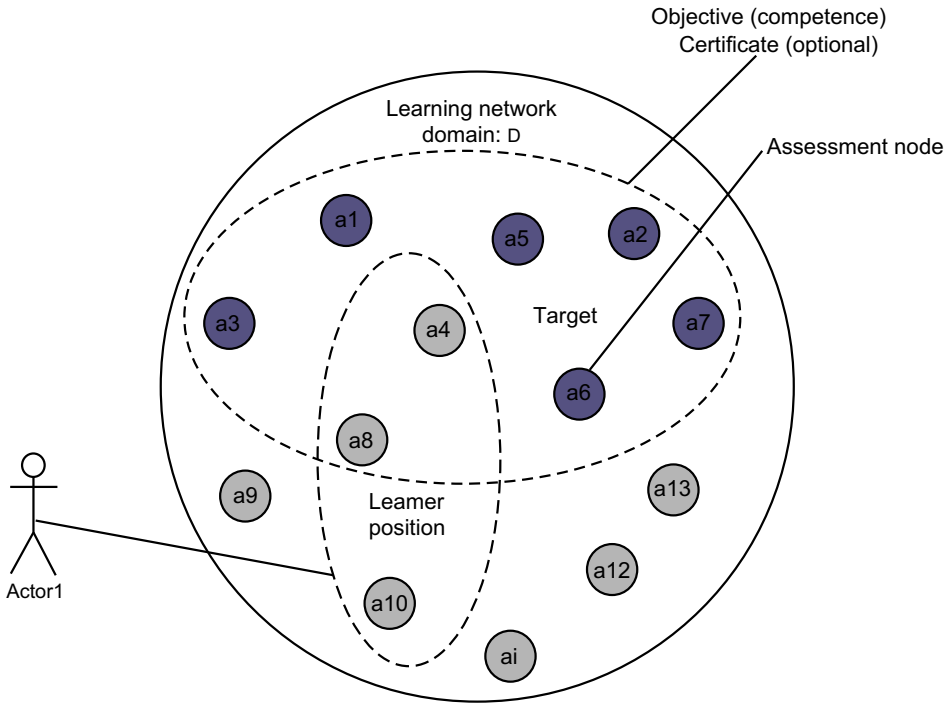

Figure 1 Model of a learning network.

networks. The individual routes of learners through this learning network are called 'learning paths.' To avoid unnecessary repetition and to support personalization, a positioning service has been developed (Kalz et al., 2007). This service has been built on the basis of accreditation for prior learning and has used language technologies to approximate the learner's position, based on a similarity analysis between the learner portfolio and the resources in the learning network. Based on the position of a learner in this network and the emerging behavior of other members in this network, the next 'best learning' activities can be recommended (Janssen et al., 2007). Drachsler et al. (2008) discuss the complexity and requirements of using recommender systems in learning networks. Recommender systems are a well-known approach for the acquisition of products or media items for users; but to use recommender systems in learning networks, a more complex set of requirements is required, due to interdependencies of learning resources and learning activities. The model of learning networks has been implemented, tested, and extended by a number of technical services in the European project 'TENCompetence' (Koper and Specht, 2008), afterward used as technical infrastructure for innovative educational projects in different countries.

Alternatively, other authors have proposed the concept of networked learning. Networked learning is more loosely defined in comparison with the learning networks approach. Networked learning is regarded as learning in which information and communication technology is used to enable connections between learners, learners and tutors, the learner and learning resources, and the learning community as a whole (Goodyear, 2001). Recently more attention has been given to the role of online social networking platforms and their potential for learning and competence development. Veletsianos and Navarrete (2012) provide a review of the literature on the use of social networking sites for learning and present a case study to analyze the experiences of students in courses delivered via social networking sites. The study showed that learning in social networks can help to decrease problems of isolation and lack of support in networked learning contexts. Due to the widespread adoption of social software and social media, their potential for lifelong learning has been explored from the perspective of self-directed learning. Dabbagh and Kitsantas (2012) propose the use of personal learning environments in combination with social media to allow the learner three type of activities that are important for self-organized learning. In their framework, the phases of personal information management, social interaction and collaboration, and information aggregation and management play a central role.

The approach of learning networks and networked learning has been applied in a number of contexts. These case studies have mostly been implemented in the professional development context, to allow professionals a flexible and self-directed access to learning resources and experts in the profession. A limitation for learning in learning networks and networked learning is that a relatively high level of self-organization and at the same time technology literacy is needed to be able to organize one's learning process based on a combination of tools.

To sum it up, learning networks, networked learning, and approaches based on the use of social networking platforms, social software, and social media, offer new contexts for lifelong learners to learn, regardless of time and place, and to be active in a social context while focusing on self-organized

learning. These new technologies contribute thus to the dismantling of the barriers of time, place, and personalization for lifelong learning but also need a relatively high level of technology literacy and self-organization competences. They have the potential to address the fragmentation problem of lifelong learning; but on the other hand, the networked approach, by and large, does not lead to centralization but to a complex array of tools and resources.

OER and Open Education

One of the developments that specifically contributed to decrease the accessibility barrier of lifelong learning is the trend to publish learning resources as OER. While traditionally learning resources have not been made available to the public, several initiatives have been started in the last 10 years that have focused on making learning resources publicly available. These initiatives have partly been motivated by the need to increase student recruitment, to maximize access to learning resources, to help disadvantaged communities of learners, and finally to give to the public what it has paid for. The last argument stems from a wider discussion about open access for scientific work financed by public resources.

The Organisation for Economic Cooperation and Development (OECD) Center for Educational Research and Innovation argues that OER are governmental means to support lifelong learning, representing a cost-effective way to ensure access to a diverse set of learning opportunities (OECD, 2007). The OER initiatives have been conducted in three different cycles. Initially, some large institutions of higher education, like MIT, started to publish learning resources publicly on the Internet, without charging access fees. Several other institutions followed this example. The vast amount of resources thus available on different platforms, motivated a second stage of development, in which the access and search of domain-specific, OER was targeted. Several technical standards have been developed to ensure that learning resources are accessible to end users (Klemke et al., 2010). A third stage has changed the focus again from the resources to the learning process as a whole, to enable open educational practices (Kalz et al., 2008). While the first and second stages have primarily focused on making isolated and reusable learning resources available and accessible, the third stage was motivated by the need of learners for an educational framing and the embedding of the learning resources in a wider learning process and context. Geser (2007) has stressed in the roadmap deliverable of the Open eLearning Content Observatory Services project, the need for new educational practices that go beyond OER use and reuse. The embedding of OER in competence-focused, collaborative knowledge-acquisition scenarios is sketched as a promising future doorway to let open educational practices emerge. Recent research and development projects have not only made OER available, but also offered new technical services to integrate these learning resources into individual competence-development trajectories or have enabled the reuse of learning resources with social media (e.g., Kalz et al., 2010). Thus OER can potentially be used in self-directed learning scenarios, as described in earlier paragraphs.

In general, open universities around the world have a long tradition of offering open learning opportunities to diverse audiences. But the Internet and OER have enabled new actors to be active in the field of open distance learning. More recently the concept of 'Massive Open Online Courses' (MOOCs) has been intensively discussed. This new format is a continuation of OER initiatives but with a focus on enabling open educational practices. Daniel (2012) discussed this new educational format critically and he stressed the adaptation of the original model of MOOCs. While the initial idea of MOOCs was centered on networked learning and cocreation of knowledge, recently several campus-based institutions have taken the initiative to implement their own MOOC concept. In this case the MOOC concept has been implemented based on an instructional paradigm with a focus on knowledge transmission via recorded lectures.

The scholarly discussion about this format is at a very early stage. Nonetheless, the speed of adoption of MOOCs has again contributed to the diversification of learning resources available to the public without access costs. While the concepts of OER and MOOCs both focus on enabling open access to learning resources and learning activities, there are potentials and limitations for both concepts. The idea of OER was not only to provide open access but also to allow adaptation of the published material under a licensing scheme like Creative Commons. This licensing scheme enables producers of digital resources to implement a fine-grained regulation of the reuse and adaptation allowed for their resources. This has enabled other teachers and learners to use existing material instead of developing something from scratch. Through the reuse goal, the resources published were mostly as context-neutral as possible, to allow adaptation for different contexts. The assumption for reuse of OER was that these resources are found, adapted, and integrated into a new learning context. The situation for MOOCs is exactly the opposite: While both formats allow open access, the concept of reuse available in OER is not given in the MOOC context. Instead, the focus of MOOCs is rather on cocreation of knowledge without a focus on learning content; or the implementation realizes a classical lecture-based model that follows a classical instructional paradigm without any reuse of resources.

To sum it up, OER and open education have addressed the accessibility barrier and the financial barrier for lifelong learning. A number of initiatives have shifted from open access and reuse of resources to enable new open educational practices. While the scholarly discussion about the effects of these new approaches is not yet mature, these initiatives increase the diversification of opportunities for lifelong learning.

Mobile and Contextualized Learning

The pervasive use of mobile devices on a global scale and in nearly all contexts of life has contributed to a growing interest in mobile and contextualized learning scenarios. While mobile learning was regarded in its early years from a technocentric perspective, later the focus shifted more to the mobility of the learner (Traxler, 2009). Mobile technologies are available whenever learning and learning needs occur to people. Thus it enables people to connect their fragmented learning

experiences to their long-term learning goals. This offers the opportunity to connect the unconnected learning contexts of lifelong learners; mobile and contextualized learning scenarios can bridge the work context and the learning context. This bridging function of mobile learning has been described by Wong and Looi (2011) as an agenda for research into seamless learning. A seam is always present when there is a change in the context of the learner. The following seams that mobile and contextualized learning should address have been identified as subjects for research:

- Encompassing formal and informal learning
- Encompassing personalized and social learning
- Learning across time
- Learning across locations
- Ubiquitous knowledge access
- Encompassing physical and digital worlds
- Combined use of multiple device types
- Switching between multiple learning tasks
- Knowledge synthesis
- Encompassing multiple pedagogical or learning activity models

These issues have also an impact on how lifelong learning is organized and conducted today and many seams of this research agenda are directly related to the barriers identified earlier.

Sharples (2000) argues that lifelong learning is always situated, can focus on different learning activities, and can happen in different learning contexts. Tools that support lifelong learners should therefore be highly portable, individual, unobtrusive, available, adaptable, persistent, useful, and intuitive. He discusses the convergent development of personal technologies and changing notions of lifelong learning. **Table 1** presents the convergent development of new communication technologies and lifelong learning (Sharples, 2000).

In a recent meta-analysis of mobile learning, it was shown that most studies in the field are conducted in higher educational institutions and elementary schools (Wu et al., 2012). This means that the lifelong learning context is not yet in the focus of attention, despite the convergence of requirements for lifelong learning and the development of new personal technologies.

To enable the vision of seamless learning experiences, the core task is to identify the key parameters that influence the learning processes and the technical consequences of developing mobile solutions to support a seamless learning experience. Besides appropriate educational content, contextualization, personalization, interaction, awareness, and reflection are the main inputs required to realize seamless learning experiences. Recently, nonformal, informal learning contexts and authentic problems in the mobile and contextualized learning field have been explored. The focus on mobile devices is also shifting to a whole group of personal technologies that could potentially be used in the lifelong learning context (Specht et al., 2012).

To sum it up, mobile and contextualized learning and the use of personal technologies have the potential to contribute to the dismantling of the fragmentation of lifelong learning activities. They support the mobility of learner and contribute to time- and place-independent access to learning resources and activities. Several case studies from disadvantaged learning communities show that they also have the potential to also address the lack of study facilities at home.

Table 1 The match of new communication and information technology to lifelong learning

Lifelong learning	*New technology*
Individualized	Personal
Learner centered	User centered
Situated	Mobile
Collaborative	Networked
Ubiquitous	Ubiquitous
Lifelong	Durable

Reproduced from Sharples, M., 2000. The design of personal mobile technologies for lifelong learning. Computers & Education 34 (3), 177–193.

Connecting the Dots of Technologies for Lifelong Learning

The following table maps the new technologies introduced with the most important barriers (**Table 2**).

Although these developments all have individual value for addressing the barriers for lifelong learning, serious improvements for many individuals will only be visible if these innovations are connected. At the moment, these topics and technological developments are still studied and developed relatively independently from each other – sometimes even in different scientific communities.

Besides these different problem areas, there is also a time perspective in the challenges presenting for technological support of lifelong learning activities. In the third age of life (also called 'older adulthood'), new requirements for learning can arise through specific health conditions. On the one hand, these requirements can be addressed by following so-called web accessibility guidelines; on the other hand, the target group of learners in later lifetime has different cognitive abilities and attitudes toward the use of new technologies

Table 2 Mapping of lifelong learning barriers and new technologies

Barrier for lifelong learning	*New technology*
Lack of finance	Open educational resources/open educational practices
Lack of personalization	Modeling of lifelong learning/open learner models and mobile and contextualized learning
Time/place	Learning networks/networked learning and mobile learning
Lack of facilities to study at home	Mobile and contextualized learning
Fragmentation	Mobile and contextualized learning/learning networks and networked learning
Health/age	Open learner models and accessibility tools

(Czaja et al., 2006). Personalization approaches for this target group have to take into account these specific requirements. Therefore, recently new initiatives have been started with a focus on providing learning opportunities for elderly people. These initiatives not only focus on the acquisition of new skills or knowledge but there is also a trend toward developing learning technologies to maintain cognitive abilities.

But these connections can only emerge if there are policy frameworks that support and foster lifelong learning for large parts of society and encourage the development of new business models for companies that support learners in lifelong learning or in learning activities that are accessible and compliant with requirements of lifelong learning. One of the largest framework programs to foster the advancement of lifelong learning is the Lifelong Learning Program developed by the European Commission. This framework needs to be combined with national initiatives to balance the societal relevance of lifelong learning and the participation in lifelong learning activities by a large part of the society.

Conclusions and Outlook

In this article, I have provided an overview of recent developments in the use of new technologies for lifelong learning purposes. These new technologies help to dismantle barriers to lifelong learning. An overall challenge for the development and evaluation of technology for lifelong learning is the evaluation context: While the context of lifelong learning is described as being episodic and multi-institutional, most evaluation contexts lack this important feature; thus scientific evidence is collected in the context of educational institutions and in a relatively short time span. In the future, more field studies and long-term evaluation scenarios, which focus on mobility problems and seamless learning experiences of lifelong learners in authentic contexts, are needed.

Another challenge to the use of new technologies for lifelong learning is the level of self-organization and self-directedness required to use these technologies to organize one's own learning. While educational institutions have an emphasis on knowledge transfer, more emphasis needs to be placed on the use of new technologies for self-directed learning. This issue is seen as one of the key skills for our society but educational institutions have not yet adapted to prepare people not only for a lives as professionals, but also for lives as lifelong learners.

See also: Computer-Assisted Instruction; Distance Education; Education and Learning: Lifespan Perspectives; Open Education: Learning and Teaching; Tangible User Interfaces in Learning and Education.

Bibliography

Brusilovsky, P., Henze, N., 2007. Open corpus adaptive educational hypermedia. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (Eds.), The Adaptive Web. Springer, Berlin, pp. 671–696.

Brusilovsky, P., Kobsa, A., Nejdl, W., 2007. The Adaptive Web: Methods and Strategies of Web Personalization, vol. 4321. Springer-Verlag, New York.

Buckingham Shum, S., Ferguson, R., 2012. Social learning analytics. Educational Technology & Society 15 (3), 3–26.

Bull, S., Kay, J., 2010. Open learner models. In: Nkambou, R., Bordeau, J., Miziguchi, R. (Eds.), Advances in Intelligent Tutoring Systems. Springer, Heidelberg, pp. 318–338.

Czaja, S.J., Charness, N., Fisk, A.D., Hertzog, C., Nair, S.N., Rogers, W.A., Sharit, J., 2006. Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). Psychology and Aging 21 (2), 333.

Dabbagh, N., Kitsantas, A., 2012. Personal learning environments, social media, and self-regulated learning: a natural formula for connecting formal and informal learning. The Internet and Higher Education 15 (1), 3–8.

Daniel, J., 2012. Making sense of MOOCs: musings in a maze of myth, paradox and possibility. Journal of Interactive Media in Education 3. http://www-jime.open.ac.uk/article/2012-18/pdf (accessed 15.05.13.).

Drachsler, H., Hummel, H.G., Koper, R., 2008. Personal recommender systems for learners in lifelong learning networks: the requirements, techniques and model. International Journal of Learning Technology 3 (4), 404–423.

Eurostat, 2012. Lifelong Learning Statistics – Statistics Explained (2013/6/5). http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Lifelong_learning_statistics# (accessed 03.06.13.).

Geser, G., 2007. Open Educational Practices and Resources: OLCOS Roadmap 2012. Open eLearning Content Observatory Services, Salzburg.

Goodyear, P., 2001. Effective Networked Learning in Higher Education: Notes and Guidelines. Project report. Networked Learning in Higher Education Project (JCALT). Deliverable 9. http://csalt.lancs.ac.uk/jisc/docs/D10%20Final%20report%20Volume%201.doc (accessed 23.05.13.).

Greller, W., Drachsler, H., 2012. Translating learning into numbers: a generic framework for learning analytics. Educational Technology & Society 15 (3), 42–57.

Janssen, J., Tattersall, C., Waterink, W., Van den Berg, B., Van Es, R., Bolman, C., Koper, R., 2007. Self-organising navigational support in lifelong learning: how predecessors can lead the way. Computers & Education 49 (3), 781–793.

Kalz, M., Drachsler, H., Van Bruggen, J., Hummel, H., Koper, R., 2008. Wayfinding services for open educational practices. International Journal of Emerging Technologies in Learning 3 (2), 1–6.

Kalz, M., Specht, M., Nadolski, R., Bastiaens, Y., Leirs, N., Pawlowski, J., 2010. OpenScout: competence based management education with community-improved open educational resources. In: Halley, S., Birch, C., Tempelaar, D., McCuddy, M., Hernández Nanclares, N., Reeb-Gruber, S., Gijselaers, W., Rienties, B., Nelissen, E. (Eds.), Proceedings of the 17th EDINEB Conference. Crossing Borders in Education and Work-Based Learning. FEBA ERD Press, Maastricht, The Netherlands, pp. 137–146.

Kalz, M., Van Bruggen, J., Rusmann, El, Giesbers, B., Koper, R., 2007. Positioning of learners in learning networks with content-analysis, metadata and ontologies. Interactive Learning Environments 15, 191–200.

Kay, J., 2008. Lifelong learner modeling for lifelong personalized pervasive learning. IEEE Transactions on Learning Technologies 1 (4), 215–228.

Klemke, R., Ternier, S., Kalz, M., Specht, M., 2010. Implementing infrastructures for managing learning objects. British Journal of Educational Technology 41 (6), 873–882.

Koper, R., Specht, M., 2008. Ten-competence: life-long competence development and learning. In: Cicilia, M.-A. (Ed.), Competencies in Organizational e-learning: Concepts and Tools. IGI-Global, Hershey, pp. 234–252.

Koper, R., Tattersall, C., 2004. New directions for lifelong learning using network technologies. British Journal of Educational Technology 35 (6), 689–700.

Koper, R., Giesbers, B., Van Rosmalen, P., Sloep, P., Van Bruggen, J., Tattersall, C., Vogten, H., Brouns, F., 2005a. A design model for lifelong learning networks. Interactive Learning Environments 13 (1–2), 71–92.

Koper, E.J.R., Rusman, E., Sloep, P., 2005b. Effective learning networks. Lifelong Learning in Europe 1, 18–27.

Longworth, N., 2003. Lifelong Learning in Action: Transforming Education in the 21st Century. Routledge, London.

OECD, 2007. Giving knowledge for free: the emergence of open educational resources. In: OECD, (Ed.), Knowledge Creation Diffusion Utilization, vol. 263. OECD Publishing, France, pp. 1–153. http://www.oecd.org/dataoecd/35/7/38654317.pdf (accessed 20.05.13.).

Romero-Zaldivar, V.A., Pardo, A., Burgos, D., Delgado Kloos, C., 2012. Monitoring student progress using virtual appliances: a case study. Computers & Education 58 (4), 1058–1067.

Sharples, M., 2000. The design of personal mobile technologies for lifelong learning. Computers & Education 34 (3), 177–193.
Specht, M., Börner, D., Tabuenca, B., Ternier, S., De Vries, F., Kalz, M., Drachsler, H., Schmitz, B., 2012. RTST Trend Report: Lead Theme Contextualisation. Deliverable 1.7 of STELLAR Network of Excellence. Heerlen, The Netherlands.
Traxler, J., 2009. Learning in a mobile age. International Journal of Mobile and Blended Learning 1 (1), 1–12.
Veletsianos, G., Navarrete, C., 2012. Online social networks as formal learning environments: learner experiences and activities. The International Review of Research in Open and Distance Learning 13 (1), 144–166.
Wong, L.-H., Looi, C.-K., 2011. What seams do we remove in mobile assisted seamless learning? A critical review of the literature. Computers & Education 57 (4), 2364–2381.
Wu, W.-H., Jim Wu, Y.-C., Chen, C.-Y., Kao, H.-Y., Lin, C.-H., Huang, S.-H., 2012. Review of trends from mobile learning studies: a meta-analysis. Computers & Education 59 (2), 817–827.

Relevant Websites

http://creativecommons.org – Creativecommons.
http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Lifelong_learning_statistics – Eurostat Lifelong Learning Statistics.
http://ec.europa.eu/education/lifelong-learning-programme/ – Lifelong Learning Program European Commission.

Lifespan Development, Theory of

Brandilynn Villarreal and Jutta Heckhausen, University of California Irvine, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by U.M. Staudinger, volume 13, pp. 8844–8848, © 2001, Elsevier Ltd.

Abstract

Life span developmental psychology is the study of individual development from conception until death. The hallmark of this approach is the coconstruction of biology, society, and the individual to produce development. Individuals actively regulate their development according to opportunities and constraints in their social ecology. Life span development theories address the impressive ability of individuals to optimize their development across the life span through the use of three central processes of developmental regulation: goal engagement, goal disengagement, and metaregulation.

Life span developmental psychology (or life span psychology) is the study of individual development (ontogenesis) from conception until death (Baltes, 1987, 1997). The life span view of human development capitalizes on the capacity of individuals to successfully adapt to the environment and optimize development across the life span. This article provides a historical background of the field as well as key propositions, major theories, current methods, and future research areas.

Historical Background

The beginnings of the field of developmental psychology date back more than 200 years ago when Johann Nicolaus Tetens (1736–1807) published a two-volume work on human nature and its development in 1777 (Baltes et al., 2006). This work covered the life span of human development and included several of what we now regard as key tenets of life span development theory. In the first few decades of the 1900s, Charlotte Bühler (1893–1974) picked up this thread and contributed groundbreaking conceptual and empirical work about the development of children, adolescents, and older adults (e.g., Bühler, 1933).

Despite a long history in Germany, what emerged as developmental psychology in North America and other areas of Europe around 1900 was distinct from the vision of Tetens and his predecessors. Influenced by biology, evolution, and genetics via Darwinism at the turn of the century, early North American developmental psychology was dominated by conceptions of maturation and growth with an almost exclusive focus on child and adolescent psychology and development.

Modern life span developmental psychology emerged in the late 1960s through the work of Hayne Reese, K. Warner Schaie, Paul B. and Margret Baltes, John R. Nesselroade, Richard M. Lerner, and their colleagues (Baltes, 1979b; Baltes et al., 1980; Lerner et al., 1983). In the 1970s and 1980s, life span psychology as a field was established with a series of conferences on the subject held at the University of West Virginia (Goulet and Baltes, 1970) and the publication of multiple volumes on life span research (e.g., Baltes, 1987; Baltes and Brim, 1984; Baltes et al., 1986).

Prior to the establishment of the field, several factors combined to fuel and popularize the life span perspective in the 1960s and 1970s (Baltes et al., 2006). Interest in the field arose from a rapidly aging population and the emergence of gerontology as a discipline. Additionally, participants from large-scale longitudinal studies of child development in the 1920s and 1930s were approaching adulthood and midlife, making it possible to study systematic developmental change across the life span.

At around the same time that life span psychology was gaining popularity as a subdiscipline of developmental psychology, life course sociological perspectives became prominent through the work of developmental psychologist Neugarten (1969), educational scientist Havighurst (1948, 1973), and life course sociologist Elder (1974, 1985) (*see* Life Course: Sociological Aspects). These scholars were pioneers in creating an interdisciplinary framework for research about life span development and the life course. Life span developmental psychology and life course sociology developed parallel to one another, both being heavily influenced by developmental systems theories (*see* Human Development, Theories of). Both fields emphasize the individual as an active agent in his or her development and the dynamic, bidirectional interactions between the individual and the environment over the life span. Not surprisingly, the main difference between the two fields is that life span psychology emphasizes the mechanisms and processes of individual behavioral and mental development, whereas life course sociology focuses on sociostructural and social-institutional conditions for shaping individual life courses.

Metatheoretical Propositions

Several key propositions serve as the foundation for life span development theories (Baltes, 1979a, 1987; Lerner, 1983). Collectively, these principles define the life span perspective and separate it from earlier conceptions of human development. One of the most fundamental of these propositions is that *development is lifelong*. Historically, development was thought to be, by-and-large, complete after adolescence. Current conceptualizations of life span psychology recognize that all developmental periods are important to understanding human development (*see* Human Development, Theories of). Individuals continue to develop and change into old age.

In addition to being lifelong, development is *multidirectional* such that every developmental stage experiences growth and decline, gains and losses. In midlife, for example, adults experience a loss in the efficiency of multitasking (e.g., Kemper et al., 2003; Lindenberger et al., 2000), yet make substantial gains in competency and productivity (Lachman et al., 1994). Even age phases that involve physical and mental decline, such as old age, can witness growth in certain domains of functioning (e.g., language competency and wisdom) (*see* Retirement and Encore Adulthood: The New Later Life Course). In the same vein, periods of rapid growth, like infancy, are prone to losses in certain domains. For example, infants of 4–6 months of age are considered 'universal learners' because they can discriminate between the phonetic differences of all languages (Eimas, 1978; Werker et al., 1981) and between the facial characteristics of even nonhuman primates (Pascalis et al., 2002). However, at around 10–12 months of age, infants experience perceptual narrowing; they can no longer perceive differences in similar phonemes in foreign languages (Cheour et al., 1998; Kuhl et al., 1992; *see* First Language Acquisition: Cross-linguistic). On the one hand, infants gain a more specialized and expert grasp of their own language. On the other hand, they lose the ability to recognize certain sounds in the broad spectrum of speech perception. In this developmental advance, gains are facilitated by losses. Perceptual losses have also been documented in the perception of musical rhythm (Hannon and Trehub, 2005) and cross-species facial recognition.

A characteristic going hand-in-hand with multidirectionality is *multidimensionality* of development. Both gains and losses occur in different domains of functioning and combine to influence overall development. Changes occur broadly in biological, cognitive, and sociocultural domains and more specifically in areas such as memory, perception, motivation, problem solving, emotion, and relating to other people. An example of multidimensional development comes from the psychometric theory of intelligence, which distinguishes two dimensions of intelligence: fluid and crystallized (Cattell, 1971; Horn, 1970, 1989) (*see* Intelligence: Central Conceptions and Psychometric Models). Fluid intelligence refers to the mechanics of the mind: basic information processing tasks that use memory and logic as well as one's ability to problem solve and mentally rotate objects. Crystallized intelligence, on the other hand, is made up of pragmatic functions or cultural knowledge, such as facts, vocabulary, rules, and procedures. These two components of intelligence show distinct developmental trajectories. Fluid intelligence shows steep gains in childhood and adolescence, followed by slow but steady declines in early adulthood and beyond. Crystallized intelligence, by contrast, remains fairly stable throughout adulthood and does not show decline in old age, at least not until very advanced age.

An important criterion for optimal development and aging throughout the life span is to *maximize gains and minimize losses*. The ratio between gains and losses in functioning changes with time such that cumulative losses eventually surpass cumulative gains. In a study by Heckhausen et al. (1989), individuals of all ages perceived and expected significantly more psychological gains than losses during most of adulthood, but toward old age, particularly after 70 years of age, expected losses outnumbered expected gains (see **Figure 1**).

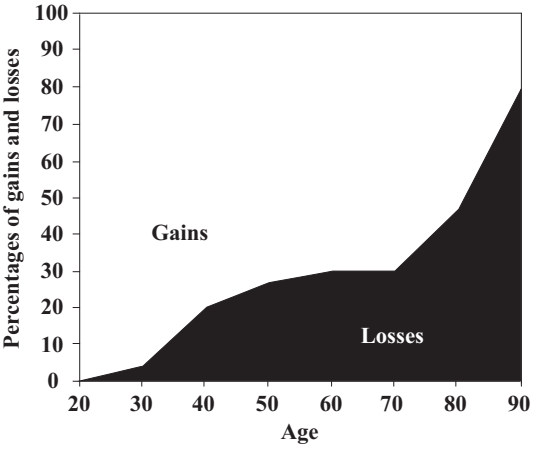

Figure 1 Expectation of fewer gains and greater losses across adulthood. Source: from 'Gains and losses in development throughout adulthood as perceived by different adult age groups' by Heckhausen, J., Dixon, R.A., Baltes, P.B., 1989. Developmental Psychology 25, 109–121.

Perhaps the most important characteristic of life span development is *plasticity*. Plasticity is within-person variability or an individual's capacity for change. Plasticity allows individuals to generate new behavior and adapt to changing biological and societal expectations or challenges throughout the life span. Although humans display an impressive amount of plasticity and capacity for developmental change, there are limits to plasticity (Baltes, 1987; Lerner, 1984). Plasticity is especially challenged after critical or sensitive periods of development. If environmental input or stimulation is not received during this time, the developmental phenomenon is not likely to occur. In the area of language development, for example, mastering the grammar of a new language is extremely difficult if the language is learned after late adolescence (Oyama, 1976).

In speech perception, although the capacity for plasticity declines as development progresses, change can still occur under certain conditions. In the case of cross-species facial recognition, plasticity can be maintained through familiarity and practice (*see* Face Recognition, Psychological and Neural Aspects). For example, 9-month-old infants who had familiarity with monkey faces through picture books were able to discriminate equally well between human faces and primate faces, an ability that is typically lost by this time (Pascalis et al., 2002; *see* Brain Maturation and Plasticity During Infancy and Adulthood). Interestingly, this type of exposure to monkey faces did not work with adults, demonstrating a permanent loss of plasticity.

Researchers are particularly interested in the limits of plasticity for domains that show declines, such as cognitive functioning (*see* Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences). For example, it is well documented that older adults experience age-related declines in memory and fluid intelligence (Salthouse, 2004; Schaie, 1994) (*see* Cognitive Aging; Aging and Memory). To test plasticity, researchers teach older adults a new set of skills (e.g., the method of loci

memory technique), give them extensive practice using these skills, and assess performance under optimized conditions (e.g., unlimited time to respond) (Baltes, 1993; Willis and Schaie, 1994). The results of studies like these suggest that age-related declines in fluid intelligence are reversible with training, at least for everyday cognitive tasks without time pressure. For example, in a series of training experiments, older adults reached the IQ performance levels of younger adults, although transfer to additional tasks was limited (Willis and Schaie, 1994; Woodruff-Pak, 1993).

When time constraints are placed on everyday tasks and when performance is assessed at the limits of capacity, conclusions about the reversibility of age-related declines are not as optimistic. The 'testing-the-limits' paradigm allows researchers to assess the scope of cognitive plasticity and uncover performance asymptotes of different age groups (Kliegl and Baltes, 1987; Lindenberger and Baltes, 1995). In these studies, younger adults often perform superior to older adults despite extensive training. For example, in a study by Lindenberger et al. (2000), younger, middle-aged, and older adults practiced walking along a complicated path and memorized a list of words at the same time. In all performance measures, older adults incurred increased costs when multitasking. Thus, at the limits of performance, older adults show robust cognitive declines that cannot be reversed with training. In everyday functioning, however, these limits are rarely reached.

The final key characteristic of development is that it is influenced by *contextual factors* (Baltes, 1987). Change occurs in different contexts and domains of development, and is sensitive to universal, individual, and cultural variations. Individuals are embedded in multilevel contexts that include families, neighborhoods, schools, churches, cities, and countries (Bronfenbrenner, 1977, 1979). Each of these contexts is influenced by historical, economic, social, and cultural factors, which are also changing over time. Contexts are dynamic and their interaction with the individual is essential for understanding developmental change.

Influences on Development

One of the objectives of life span developmental theory is to outline the structure and sequence of development as it unfolds over the life span (Baltes et al., 2006). In order to do this, we must understand the factors that influence human development and how they change over time. Human development is a product of the interaction of three factors: biology, culture (or society), and the individual agent. This view of development as the coconstruction of biology, society, and the individual is the hallmark of a modern life span developmental approach. Previous conceptualizations noted two, often dichotomous, influences on development: nature (biology) and nurture (the environment) (Lerner, 2002). The previous entry of Lifespan Developmental Theory similarly specified biology and culture as "the main influences shaping human development" (Staudinger, 2001: p. 8846).

Researchers have long acknowledged the important role of *biology* in guiding development, both growth and decline, through genetic inheritance, physical maturation, physical aging, hormones, temperament, and other mechanisms. Biological factors set timetables for developmental milestones that influence the course of development. For example, women experience menopause between 39 and 59 years of age (median age = 50–52) (Avis, 1999; Rossi, 2004). Because of the rapid decline of estrogen production in the ovaries, women can no longer bear children after menopause. As a result, women typically pursue developmental goals related to childbearing when there are biological opportunities to do so, and disengage from these goals when they are no longer attainable (after menopause or infertility) (Heckhausen et al., 2001).

Society influences development through its social structure (e.g., social inequality and segregation of social strata), age-graded cultural timetables, and social institutions (e.g., educational system and regulations of promotions and retirement) (*see* Age Stratification; Social Stratification). Key contextual contributors to development include the neighborhood or community in which the individual lives, the type of parenting received, and stressful life events such as poverty or sociopolitical upheaval. Similar to biological influences, societal factors also set timetables for development. Individuals of a particular society are exposed to an agreed-upon, age-graded sequencing of major life events. For example, individuals expect to experience significant life events, such as graduating from school or college, marrying, having their first child, and retiring, at certain points in their lives (Heckhausen, 1999). These guidelines are responsible for the common life paths of individuals in a given society as well as age-appropriate behavior, roles, and responsibilities. Finally, social institutions like the education and legal system play a large role in structuring the life course. Minimum age requirements and mandatory retirement laws for certain professions and in certain countries determine when an individual can participate in the labor force. Thus, society and culture structure the opportunities to pursue developmental goals across the life span.

Current conceptualizations of life span psychology emphasize the role of the *individual* as a coproducer of development (Lerner and Busch-Rossnagel, 1981) and active agent of change (Brandtstädter, 1999; Heckhausen, 1999). Thus, individuals are also producers of their own development. Individuals play both an active and passive role in regulating development across the life span (Lerner and Busch-Rossnagel, 1981). For example, *passive* physical characteristics such as being tall or physically attractive can contribute to an individual's popularity and even lead to greater career opportunities and income (Deaton and Arora, 2009). Being an extrovert, which is a fairly innate aspect of one's temperament, can lead to a larger social network and with it, additional opportunities and resources. Individuals *actively* influence their development primarily through the selection and pursuit of developmental goals (Heckhausen, 1999). By choosing to invest in particular developmental paths or life domains over others, individuals optimize the fit between their preferences and the environment. Individuals shape their social ecology through the friends they make, the colleges they choose, and the careers they pursue. They play an active role in optimizing their development by striving to control the environment, seizing opportunities as they become available, learning from others, and exploring new behaviors and their effects.

Thus far, we have described the three influences of development and how they change over the life span. We will now describe three patterns of influence that define development across the life span: history-graded, age-graded, and nonnormative life events (Baltes et al., 1980). These patterns of influence contribute to commonalities among individuals, interindividual differences, and intraindividual change (or plasticity) (Lerner, 1984). History-graded events are experienced by all individuals living at the same time and lead to differences between birth cohorts. They can include environmental or man-made catastrophes such as natural disasters, wars, or economic recessions. Alternatively, history-graded events can encompass the general zeitgeist of the era (e.g., the present state of social institutions and laws, the general quality of health, or the advancement of technology). Age-graded events are universal for members of a given age group and contribute to commonalities experienced across the life span. For example, all members of an age group experience certain biological and maturational milestones such as learning to talk and walk. Similarly, members of an age group experience societal expectations and pressures in the form of developmental tasks (Havighurst, 1948, 1973). Individuals can expect to participate in life tasks depending on their age, such as attending school in childhood and entering the workforce in adolescence or adulthood. Nonnormative life events are idiosyncratic but can exert a powerful influence on development, such as a genetic abnormality. The influences of development and their patterns and layers combine to produce an action field of development. It is only through the interaction of all these influences that we gain a complete understanding of human development across the life span.

Fundamental Theories of Individual Agency in Human Development

The key metatheoretical principles proposed by Baltes (1987) and other early life span developmentalists were broad and open enough to allow the emergence of several developmental models, each with their own theoretical concepts and empirical phenomena (Heckhausen, 2003). Models for specific domains of functioning gave way to broad, process-oriented theories of developmental regulation, the three most prominent of which being selection, optimization, and compensation (SOC) proposed by Paul Baltes and Margaret Baltes and their colleagues (Baltes, 1987, 1997; Baltes and Baltes, 1990); the dual-process model of assimilation and accommodation proposed by Brandtstädter (1989, 2009); and the motivational theory of life span development (MTD) proposed by Heckhausen et al. (2010).

As a group, the theories of life span development aim to describe, explain, predict, and optimize human development across the life span. Similar to developmental systems theories, these theories stress the important role of individuals and dynamic person–context relations in influencing human development. Individuals are able to regulate their development according to opportunities and constraints in their social ecology. They actively seek to adjust the use of developmental regulation strategies to optimize development across the life span.

Selection, Optimization, and Compensation

The SOC Model (Baltes and Baltes, 1990; Baltes et al., 1998; Freund and Baltes, 2002) provides a metatheoretical framework for understanding three processes of developmental regulation across the life span: goal selection, goal pursuit, and goal maintenance. These processes are sustained through the use of SOC strategies. A key aspect of this model is that individuals engage in these processes in a coordinated manner to extract the most from their environments and achieve adaptive developmental outcomes.

Elective selection involves the construction of a goal hierarchy and commitment to a set of goals and preferred activities. Individuals focus on goals that are most important to them for their particular life stage. These goals change throughout the life course according to opportunities and constraints. For example, an older individual may select goals and optimize gains in domains that he or she can excel in, such as volunteering, rather than domains that are no longer possible, such as becoming a world-class athlete. Optimization refers to the investment of resources in the selected activity or goal. For example, pursuing a goal may require acquiring new skills. Compensation occurs when individuals experience a shortcoming, like age-related decline. In these cases, individuals invest more resources and use compensatory strategies to maintain an adaptive level of functioning and achieve their goals. An individual may rely more on culture or redistribute available resources to areas of need. Concert pianist Arthur Rubinstein who still performed at 80 years of age demonstrates all three aspects of SOC in adaptive functioning (Baltes and Baltes, 1990). As he got older, Arthur maintained a high level of performance by selecting fewer pieces to play, practicing these pieces more, and using certain 'tricks' to make up for a decline in motoric speed, such as purposely slowing down before fast and difficult segments. Finally, the SOC model specifies adaptive loss-based selection strategies that individuals use when functional loss has occurred. For example, individuals can adjust current standards and goal hierarchies to facilitate future goal pursuits.

Dual-Process Model of Assimilation and Accommodation

While the SOC model views the fundamental function of developmental regulation as the maximization of gains and minimization of losses, the dual-process model emphasizes the need to preserve self-consistency across the life span (Haase et al., 2013). As an action theory of development, Brandtstädter's (1989, 2009) model postulates the self-concept as the organizing and guiding force behind development. Development is regulated via a self-organizing process to minimize discrepancies between desired and actual goal states. Individuals seek to reduce self-discrepancies and emotional distress through the use of two developmental regulatory processes: assimilation and accommodation.

Assimilation refers to an individual's active and intentional efforts to attain desired developmental goals and outcomes by changing the environment. For example, an individual may strive tenaciously to achieve a valuable goal, such as a successful career or happy family. Assimilation is most adaptive when opportunities to pursue the goal, resources, and goal importance are high. Accommodation, on the other hand, is most adaptive

when resources are scarce and the goal is less important and easily substituted. Accommodation involves subconscious cognitive processes that reduce the discrepancy between the actual and desired goal states. For example, an individual may nonintentionally reappraise his or her actual goal state in a more positive light. Conversely, the individual may also reappraise the desired state as less positive. Accommodation strategies are particularly salient when challenges emerge or goals can no longer be attained. In this case, individuals should use accommodative cognitions to flexibly adjust goals, reevaluate the self, or find meaning in failure.

Assimilation and accommodation processes have the same goal: to reduce discrepancy and distress in the self. However, they operate in opposing ways, such that the use of one process inhibits the other (Brandtstädter and Rothermund, 2002). While assimilative processes emphasize the role of intentionality in forming goals, values, beliefs, and volitions, accommodative processes are mostly beyond conscious awareness.

Motivational Theory of Life Span Development

The MTD is based on the proposition that, throughout life, humans strive to maximize their control of the environment and their own development (Heckhausen et al., 2010). Individuals regulate their development by engaging with appropriate developmental goals and disengaging from goals which have become unproductive. In doing this, individuals make use of the enormous potential for plasticity and change across the life span, as well as take into consideration the age-graded constraints on plasticity (e.g., biological, sociocultural, and age-normative limitations). It is through this interaction between plasticity and constraints that individuals select specific developmental goals to pursue. Individuals select the most adaptive and age-appropriate developmental goals using a higher order self-regulatory process called optimization (Heckhausen, 1999; Heckhausen and Schulz, 1993). The optimization of goal choice, which differs from optimization in the SOC model, is a higher order regulatory strategy that guides individuals' engagement with and disengagement from various developmental goals.

The specific heuristics of the optimization principle help individuals determine the ideal time to pursue a goal, taking into account both the opportunities available to the individual as well as consequences and trade-offs for pursuing one goal over another. For example, goal pursuit is only adaptive if sufficient opportunities in the developmental ecology exist to facilitate its attainment. Thus, goal pursuit should be congruent with age-graded opportunity structures and developmental timetables. Equally important when choosing developmental goals is to consider the broader context and impact of pursuing the goal for overall functioning and long-term development. Pursuing the goal should not negatively impact the attainment of concurrent or future goals. At the same time, the individual should maintain a minimum diversity of goals to guard against developmental dead ends if the goal is not attained or goal pursuit is no longer possible. Thus, these heuristics work together to support goal engagement as opportunities in the environment arise, and goal disengagement as opportunities fade away. For example, a woman of childbearing age who wishes to have a child should pursue this goal when the biological opportunities to do so are favorable. In addition, she should consider the consequences of having a child on other important domains, such as career, and keep in mind that investing too heavily in any one domain may have negative consequences for long-term development (Heckhausen et al., 2001).

The action-phase model of development proposed by Heckhausen (1999) outlines the motivational and self-regulatory strategies used to regulate development throughout the life span, as it is organized into cycles of goal selection, goal engagement, and goal disengagement. These processes are especially important when challenges and difficulties arise in goal pursuit. Goal engagement involves various control strategies that augment an individual's behavioral and motivational resources. Behavioral resources include the investment of effort, time, strategies, and persistence in attaining the goal. Individuals can also seek out additional external sources of support if needed. Motivational resources involve the level of commitment and importance tied to the goal as well as the ability to stay focused on the goal despite attractive alternatives. If the goal has a deadline, such as the 'biological clock' of childbearing for women, the opportunities to successfully attain the goal decrease with time, creating a sense of urgency (Heckhausen et al., 2001). Thus, individuals increase the intensity of striving for the goal as the deadline nears since there is only a limited amount of time to attain the goal. Indeed, research shows that individuals who tailor their control striving to the controllability of outcomes and engage with goals when control opportunities are plentiful, as well as disengage from goals when controllability is low, achieve better developmental outcomes, enjoy better psychological and physical health, and live longer lives (e.g., Hall et al., 2010; Heckhausen et al., 2001; Wrosch and Heckhausen, 1999; Wrosch et al., 2007).

Goal disengagement occurs when the goal is no longer attainable; either the goal has been successfully completed or it can no longer be pursued for a number of reasons. For example, the developmental deadline may have passed. In this case, whether or not the goal state was reached, it is adaptive to disengage from the goal. If the goal was successfully met, the individual can begin a new goal cycle of selection and engagement. If the goal was not attained, individuals may choose to adjust the goal or pursue another goal altogether. Compensatory strategies associated with goal disengagement work to protect motivational resources and self-esteem after failure so the individual can continue to pursue new goals. Individuals who continue to pursue goals after they are no longer attainable have worse mental and physical health than those who disengage from the goal. This relationship holds for goals in various domains, such as childbearing, romantic relationships, and health (Heckhausen et al., 2001; Wrosch and Heckhausen, 1999; Wrosch et al., 2002).

Altogether, it is through these various control strivings that individuals optimize goal pursuit and their development. Individuals differ in the degree to which they adjust their goal engagement and disengagement to changing opportunities in the environment. Those who use adaptive strategies at appropriate times as well as transition quickly and effortlessly to different phases in the action cycle will attain better developmental outcomes and optimize their mental and physical health.

Methods

What has set the field of life span psychology apart from other subdisciplines of psychology is its emphasis on developmental change across the life span. Only by following individuals and their interactions with the environment over time is it possible to discern patterns of change. The most important research design to test propositions of life span theories is a longitudinal design (*see* Longitudinal Data). This type of study allows researchers to capture real-life processes that unravel over the course of the individual's life. Longitudinal studies assess interindividual patterns of intraindividual change, while cross-sectional studies are limited to interindividual differences and cannot address age-related change. Ideally, longitudinal studies are prospective in nature with a large and representative sample to sufficiently track developmental change. Multiple cohorts of participants can be added, as in cross-sequential designs, to help rule out confounding cohort effects.

Several important developmental phenomena can be assessed with longitudinal studies. One is the study of life transitions, such as entering or exiting the workforce. Studying life events is important because as particularly stressful periods in one's life, they have the potential to set individuals on different developmental trajectories. For any life event, there is heterogeneity in reactions that vary across different domains of functioning. People change in different directions and at different rates. Since data are collected across a span of years or decades, researchers can disentangle the impact of earlier life events and conditions on later developmental outcomes, such as health and well-being. Even if the event is a nonnormative one, such as acquiring a chronic disease, a prospective study will guarantee that a subset of individuals experience the life event if the sample is sufficiently large and representative.

Although longitudinal designs are necessary to study developmental change (see Propensity Scoring), they need to be combined with the most up-to-date methods of analysis (*see* Life Course and Event History Analysis). Analyses must go above and beyond average levels of change (i.e., between-person differences) to produce information about individual levels of change. One current method of analyses is multiphase growth curves (see Growth Curve Analysis; Statistical Analysis: Multilevel Methods). These models capture responses to life events, but also account for individual difference characteristics and risk or protective factors that were collected at earlier assessments (Gerstorf et al., 2010; McArdle and Nesselroade, 2003; Ram and Grimm, 2007; Singer and Willett, 2003). Thus, researchers are able to uncover the psychological and biological mechanisms underlying differential levels of change and development. Researchers can also investigate the cumulative effects of these variables and their influence on successful development.

In combination with longitudinal designs, propensity score matching can be used to increase our understanding of developmental change. Propensity score matching simulates an experimental design by investigating important covariates and matching participants on these variables to create a 'control' group and a 'treatment' group. Propensity scores of each group can then be used in a logistic regression to examine the differential likelihood of outcomes by group (i.e., whether or not the individual experienced a particular life event) (Stuart, 2010; Thoemmes and Kim, 2011). This methodology allows for more confident causal statements than traditional quasiexperimental designs. Finally, longitudinal design would benefit from including a microlongitudinal study or 'measurement burst' within the larger study. An intense series of measurement points can reveal the underlying mechanisms and processes associated with developmental change, such as how day-to-day events influence long-term outcomes (Infurna et al., under review).

Additional methodologies are useful in understanding the basic processes involved in plasticity, or the ability to change throughout the life span. For example, the plasticity of cognitive decline is assessed with a testing-the-limits experimental paradigm (Kliegl and Baltes, 1987). This type of short-term longitudinal study typically uses training in an effort to reverse cognitive decline among older individuals. The baseline capacity of individuals without training reveals everyday differences in cognitive functioning. Baseline reserve capacity refers to performance on tasks when conditions for learning are optimized, but no training has occurred. Finally, developmental reserve capacity refers to individuals' maximum performance attained with extensive training that has elevated performance to an asymptote. The potential for plasticity can be seen if baseline reserve and developmental reserve capacities are greater than baseline capacities.

Areas for Future Research and Conclusions

As a still emerging field in psychology, life span psychology has a bright and interdisciplinary future. Three similar, yet distinct, theories currently dominate the field of life span psychology: the SOC model, the dual-process model, and the MTD. Because the theories are conceptually and functionally similar, Haase et al. (2013) proposed a theoretical and empirical integration: a unified theory of life span development with three central processes of adaptive developmental regulation: goal engagement, goal disengagement, and metaregulation. The SOC model and the dual-process model both contain constructs similar to goal engagement and disengagement, as proposed by the MTD. Goal engagement is similar to Brandtstädter's (1989) assimilation and the SOC model's elective selection, while goal disengagement is similar to Brandtstädter's (1989) accommodation and the SOC model's loss-based selection. The MTD is the only theory that recognizes a higher order developmental regulation process: optimization. More research is needed on the validity of the optimization construct within varying contexts and populations. By its nature, optimization is difficult to assess via self-report measures and would benefit from a combination of scale-based and open-formatted interviews.

Theories of life span development share the goal of using regulation strategies to optimize development across the life span. However, current theories disagree on what should be the ultimate criteria of successful or adaptive development (Heckhausen, 2003). The dual-process model does not operationalize successful development, making empirical tests of the model difficult. The SOC model gives several definitions of adaptiveness in developmental regulation, such as the maximization of gains and the minimization of losses or attaining desired outcomes and avoiding undesired outcomes, which value both subjective and objective indicators of success

(Baltes, 1997). While subjective criteria incorporate the personal goals and cultural context of individuals, they are also subject to rationalization biases and the possibility that an individual may accomplish little by objective standards but be highly satisfied. Although the SOC model argues against a universal criterion of adaptive development, there is a general consensus regarding what it means to be successful in different cultures. That is, high levels of functioning are desired in cognitive and intellectual performance, physical mobility (or absence of disability), affective functioning, and social relations (Heckhausen, 2003). The criterion for successful development in the MTD is to maximize the control potential of the individual in valued domains across cultures. In moving forward, the field would benefit from adopting criteria for successful development that is explicit, nonrelativistic, and objectifiable.

The majority of research in life span development gathers data through self-report measures. This type of assessment works well when investigating the explicit and effortful intentional processes of goal engagement and disengagement, but falls short in measuring processes that are beyond conscious awareness. Intentional processes are those that deliberately manipulate volitional commitment, such as making the decision to disengage from a goal. There are several problems inherent in self-report measures, such as the possibility of selective and biased informational processing and recall (Brandtstädter, 2000). Conversly, subintentional processes include various secondary control processes designed to cope with loss and protect resources. Examples include shifting priorities or increasing cognitions that avert self-blame. One of the ways these processes are measured is through an incidental memory paradigm (Heckhausen et al., 2001). Because secondary control processes are associated with well-being and adaptive development, it is important to accurately tap into their subintentional nature with measures that go beyond self-report.

According to the MTD, discrete and radical shifts between goal selection, goal engagement, and goal disengagement are most adaptive, but more research is needed to confirm this hypothesis (Heckhausen et al., 2010). It is not clear which specific processes are involved when individuals transition from one phase of the action-phase model to another. In the real world, there is much variability in how people transition and this may have important consequences for long-term developmental outcomes.

Finally, additional research is needed to investigate the interaction between biological and sociocultural constraints and individual characteristics. Elias (1969) described a dynamic relationship between societal regulation of behavior and individual regulation of behavior. As one increases the other decreases, so when societal constraints are few or nonspecific, the individual must invest more in regulating his or her development by seeking to control the environment. And conversely, when societal constraints are more severe, the individual does not need to internalize rules and norms. This way, a not-too-much-and-not-too-little level of regulatory constraint is maintained across societal change. In the context of ubiquitous international migration, individuals need to adjust their life span-regulatory striving when migrating between countries with different levels of constraints in social mobility.

At its core, development proceeds as a transaction between the individual and his or her social ecology. Theories of life span developmental psychology showcase the impressive ability of individuals to optimize their development across the life span. As coproducers of development, individuals construct an action field that takes into account changing life opportunities and constraints in biology and society to select the most adaptive developmental goals. They maximize the likelihood of goal attainment through motivational and self-regulatory strategies, and use additional strategies to adjust for setbacks or compensate for age-related declines in functioning. The potential for developmental change and plasticity across the life span is great.

See also: Age Stratification; Aging Mind: Facets and Levels of Analysis; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Face Recognition, Psychological and Neural Aspects; Human Development, Theories of; Infancy and Childhood: Emotional Development; Life Course and Event History Analysis; Life Course: Sociological Aspects; Neural Plasticity; Retirement and Encore Adulthood: The New Later Life Course.

Bibliography

Avis, N.E., 1999. Women's health at midlife. In: Willis, S.L., Reid, J.D. (Eds.), Life in the Middle: Psychological and Social Development in Middle Age. Academic Press, San Diego, pp. 105–147.

Baltes, P.B., 1979a. Life span developmental psychology: some converging observations on history and theory. In: Baltes, P.B., Brim Jr., O.G. (Eds.), Life Span Development and Behavior, vol. 2. Academic Press, New York, pp. 255–279.

Baltes, P.B., 1979b. On the potential and limits of child development: life span developmental perspectives. Newsletter of Society for Research in Child Development 1–4.

Baltes, P.B., 1987. Theoretical propositions of life-span developmental psychology: on the dynamics between growth and decline. Developmental Psychology 23, 611–626.

Baltes, P.B., 1993. The aging mind: potential and limits. Gerontologist 33, 580–594.

Baltes, P.B., 1997. On the incomplete architecture of human ontogeny: selection, optimization, and compensation as foundation of developmental theory. American Psychologist 52, 366–380.

Baltes, P.B., Baltes, M.M., 1990. Psychological perspectives on successful aging: the model of selective optimization with compensation. In: Baltes, P.B., Baltes, M.M. (Eds.), Successful Aging: Perspectives from the Behavioral Sciences. Cambridge University Press, New York, pp. 1–34.

Baltes, P.B., Brim Jr., O.G. (Eds.), 1984. Life-span Development and Behavior, vol. 6. Academic Press, New York.

Baltes, P.B., Featherman, D.L., Lerner, R.M. (Eds.), 1986. Life-span Development and Behavior, vol. 7. Erlbaum, Hillsdale, NJ.

Baltes, P.B., Lindenberger, U., Staudinger, U.M., 2006. Life-span theory in developmental psychology. In: Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology: Theoretical Models of Human Development, sixth ed., vol. 1. Wiley, New York, pp. 569–664.

Baltes, P.B., Lindenberger, U., Staudinger, U.M., 1998. Life-span theory in developmental psychology. In: Lerner, R.M. (Ed.), Handbook of Child Psychology: Theoretical Models of Human Development, fifth ed., vol. 1. Wiley, New York, pp. 1029–1143.

Baltes, P.B., Reese, H.W., Lipsitt, L.P., 1980. Life-span developmental psychology. Annual Review of Psychology 31, 65–110.

Brandtstädter, J., 1989. Personal self-regulation of development: cross-sequential analyses of development-related control beliefs and emotions. Developmental Psychology 25, 96–108.

Brandtstädter, J., 1999. The self in action and development: cultural, biosocial, and ontogenetic bases of intentional self-development. In: Brandtstädter, J., Lerner, R.M. (Eds.), Action and Self-development: Theory and Research through the Life-span. Sage, Thousand Oaks, CA, pp. 37–65.

Brandtstädter, J., 2000. Emotion, cognition and control: limits of intentionality. In: Perrig, W.J., Grob, A. (Eds.), Control of Human Behavior, Mental Processes, and Consciousness: Essays in Honor of the 60th Birthday of August Flammer. Erlbaum, Mahwah, NJ.

Brandtstädter, J., 2009. Goal pursuit and goal adjustment: self-regulation and intentional self-development in changing developmental contexts. Advances in Life Course Research 14, 52–62.

Brandtstädter, J., Rothermund, K., 2002. The life-course dynamics of goal pursuit and goal adjustment: a two-process framework. Developmental Review 22, 117–150.
Bronfenbrenner, U., 1977. Toward an experimental ecology of human development. American Psychologist 32, 513–532.
Bronfenbrenner, U., 1979. The Ecology of Human Development: Experiments by Nature and Design. Harvard University Press, Cambridge, MA.
Bühler, C., 1933. Der Menschliche Lebenslauf Als Psychologisches Problem [The Course of Human Life as a Psychological Problem]. Hirzel, Leipzig.
Cattell, R.B., 1971. Abilities: Their Structure, Growth, and Action. Houghton Mifflin, Boston.
Cheour, M., et al., 1998. Development of language-specific phoneme representations in the infant brain. Nature Neuroscience 1, 351–353.
Deaton, A., Arora, R., 2009. Life at the top: the benefits of height. Economics & Human Biology 7, 133–136.
Eimas, P.D., 1978. Developmental aspects of speech perception. In: Held, R., Leibowitz, H.W., Teuber, H.L. (Eds.), Handbook of Sensory Physiology, vol. 8. Springer, Berlin.
Elder Jr., G.H., 1974. Children of the Great Depression. University of Chicago Press, Chicago.
Elder Jr., G.H. (Ed.), 1985. Life Course Dynamics: Trajectories and Transitions 1968–1980. Cornell University Press, Ithaca, NY.
Elias, N., 1969. The History of Manners. The Civilizing Process, vol. 1. Blackwell, Oxford.
Freund, A.M., Baltes, P.B., 2002. Life-management strategies of selection, optimization, and compensation: measurement by self-report and construct validity. Journal of Personality and Social Psychology 82, 642–662.
Gerstorf, D., et al., 2010. Late-life decline in well-being across adulthood in Germany, the United Kingdom, and the United States: something is seriously wrong at the end of life. Psychology of Aging 25, 477–485.
Goulet, L.R., Baltes, P.B. (Eds.), 1970. Life Span Developmental Psychology: Research and Theory. Academic Press, New York.
Haase, C., Heckhausen, J., Wrosch, C., 2013. Developmental regulation across the life span: toward a new synthesis. Developmental Psychology 49, 964–972.
Hall, N.C., Chipperfield, J.G., Heckhausen, J., Perry, R.P., 2010. Control striving in older adults with serious health problems: a 9-year longitudinal study of survival, health, and well-being. Psychology and Aging 25, 432–445.
Hannon, E.E., Trehub, S.E., 2005. Metrical categories in infancy and adulthood. Psychological Science 16, 48–55.
Havighurst, R.J., 1948. Developmental Tasks and Education. Davis McKay, New York.
Havighurst, R.J., 1973. History of developmental psychology: socialization and personality development through the life span. In: Baltes, P.B., Schaie, K.W. (Eds.), Life Span Developmental Psychology. Academic Press, New York, pp. 3–24.
Heckhausen, J., 1999. Developmental Regulation in Adulthood: Age-Normative and Sociostructural Constraints as Adaptive Challenges. Cambridge University Press, Cambridge, England.
Heckhausen, J., 2003. The future of lifespan psychology: perspectives from control theory. In: Staudinger, U.M., Lindenberger, U. (Eds.), Understanding Human Development: Dialogues with Lifespan Psychology. Kluwer Academic, Dordrecht, The Netherlands, pp. 383–400.
Heckhausen, J., Dixon, R.A., Baltes, P.B., 1989. Gains and losses in development throughout adulthood as perceived by different adult age groups. Developmental Psychology 25, 109–121.
Heckhausen, J., Schulz, R., 1993. Optimization by selection and compensation: balancing primary and secondary control in life span development. International Journal of Behavioral Development 16, 287–303.
Heckhausen, J., Wrosch, C., Fleeson, W., 2001. Developmental regulation before and after a developmental deadline: the sample case of "biological clock" for childbearing. Psychology and Aging 16, 400–413.
Heckhausen, J., Wrosch, C., Schulz, R., 2010. A motivational theory of life-span development. Psychological Review 117, 32–60.
Horn, J.L., 1970. Organization of data on life span development of human abilities. In: Goulet, L.R., Baltes, P.B. (Eds.), Life Span Developmental Psychology: Research and Theory. Academic Press, New York, pp. 423–466.
Horn, J.L., 1989. Model of intelligence. In: Linn, R.L. (Ed.), Intelligence: Measurement, Theory, and Public Policy. University of Illinois Press, Urbana, pp. 29–73.
Infurna, F.J., Gerstorf, D., Ram, N., Heckhausen, J. Analytic strategies for the study of adaptation to major life events: making the most of large-scale longitudinal surveys. In: Blossfeld, H.P., von Maurice, J. Skopek, J., (Eds.), Methodological Issues of Longitudinal Surveys: The Example of the National Educational Panel Study, under review.
Kemper, S., Herman, R.E., Lian, C.H.T., 2003. The costs of doing two things at once for young and older adults: talking while walking, finger tapping, and ignoring speech or noise. Psychology and Aging 18, 181–192.
Kliegl, R., Baltes, P.B., 1987. Theory-guided analysis of mechanisms of development and aging mechanisms through testing-the-limits and research on expertise. In: Schooler, C., Schaie, K.W. (Eds.), Cognitive Functioning and Social Structure over the Life Course. Ablex, Norwood, NJ, pp. 95–119.
Kuhl, P.K., Williams, K.A., Lacerda, F., Stevens, K.N., Lindblom, B., 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. Science 255, 606–608.
Lachman, M.E., Lewkowicz, C., Marcus, A., Peng, Y., 1994. Images of midlife development among young, middle-aged, and older adults. Journal of Adult Development 1, 201–211.
Lerner, R.M. (Ed.), 1983. Developmental Psychology: Historical and Philosophical Perspectives. Erlbaum, Hillsdale, NJ.
Lerner, R.M., 1984. On the Nature of Human Plasticity. Cambridge University Press, New York.
Lerner, R.M., 2002. Concepts and Theories of Human Development, third ed. Erlbaum, Mahwah, NJ.
Lerner, R.M., Busch-Rossnagel, N. (Eds.), 1981. Individuals as Producers of Their Development: A Life Span Perspective. Academic Press, New York.
Lerner, R.M., Hultsch, D.F., Dixon, R.A., 1983. Contextualism and the character of developmental psychology in the 1970s. Annals of the New York Academy of Sciences 412, 101–128.
Lindenberger, U., Baltes, P.B., 1995. Testing-the-limits and experimental simulation: two methods to explicate the role of learning in development. Human Development 38, 349–360.
Lindenberger, U., Marsiske, M., Baltes, P.B., 2000. Memorizing while walking: increase in dual-task costs from young adulthood to old age. Psychology and Aging 15, 417–436.
McArdle, J.J., Nesselroade, J.R., 2003. Growth curve analysis in contemporary psychological research. In: Schinka, J.A., Velicer, W.F. (Eds.), Handbook of Psychology: Research Methods in Psychology, vol. 2. John Wiley & Sons, New York, pp. 447–480.
Neugarten, B.L., 1969. Continuities and discontinuities of psychological issues into adult life. Human Development 12, 121–130.
Oyama, S., 1976. A sensitive period for the acquisition on of a nonnative phonological system. Journal of Psycholinguistic Research 5, 261–283.
Pascalis, O., de Haan, M., Nelson, C.A., 2002. Is face processing species-specific during the first year of life? Science 296, 1321–1323.
Ram, N., Grimm, K., 2007. Using simple and complex growth models to articulate developmental change: matching theory to method. International Journal of Behavioral Development 31, 303–316.
Rossi, A.S., 2004. The menopausal transition and aging processes. In: Brim, O.G., Ryff, C.D., Kessler, R. (Eds.), How Healthy Are We: A National Study of Well-Being in Midlife. University of Chicago Press, Chicago, pp. 550–575.
Salthouse, T., 2004. What and when of cognitive aging. Current Directions in Psychological Science 13, 140–144.
Schaie, K.W., 1994. The course of adult intellectual development. American Psychologist 49, 304–313.
Singer, J.D., Willett, J.B., 2003. Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. Oxford University Press, Oxford.
Staudinger, U.M., 2001. Life reflection: a social-cognitive analysis of life review. Review of General Psychology 5, 148–160.
Stuart, E., 2010. Matching methods for causal inference: a review and a look forward. Statistical Science 25, 1–21.
Thoemmes, F.J., Kim, E.S., 2011. Systematic review of propensity score methods for social sciences. Multivariate Behavioral Research 46, 90–118.
Werker, J.F., Gilbert, J.H.V., Humphrey, K., Tees, R.C., 1981. Developmental aspects of cross-language speech perception. Child Development 52, 349–355.
Willis, S.L., Schaie, K.W., 1994. Cognitive training in the normal elderly. In: Forette, F., Christen, Y., Boller, F. (Eds.), Plasticité Cérébrale et Stimulation Cognitive [Cerebral Plasticity and Cognitive Stimulation]. Fondation Nationale de Gérontologie. France, Paris, pp. 91–113.
Woodruff-Pak, D.S., 1993. Neural plasticity as a substrate for cognitive adaptation in adulthood and aging. In: Cerella, J.E., Rybash, J.M., Hoyer, W.E., Commons, M.L. (Eds.), Adult Information Processing: Limits on Loss. Academic Press, New York, pp. 13–35.
Wrosch, C., Heckhausen, J., 1999. Control processes before and after passing a developmental deadline: activation and deactivation of intimate relationship goals. Journal of Personality and Social Psychology 77, 415–427.
Wrosch, C., Miller, G., Scheier, M.F., de Pontet, S., 2007. Giving up on unattainable goals: benefits for health? Personality and Social Psychology Bulletin 33, 251–265.
Wrosch, C., Schulz, R., Heckhausen, J., 2002. Health stresses and depressive symptomatology in the elderly: the importance of health engagement control strategies. Health Psychology 21, 340–348.

Lifespan Development: Evolutionary Perspectives

Heidi Keller, School of Human Sciences, Osnabrück University, Osnabrück, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H. Keller, volume 13, pp. 8840–8844, © 2001, Elsevier Ltd.

Abstract

From an evolutionary perspective, reproductive success can be understood as the driving force for specific developmental pathways, which present coherent answers to contextual demands. The sequence and structure of developmental stages across the life span constitute reproductive styles that promise best inclusive fitness within the respective context. Although the prolongation of the life span in humans compared with all others species is certainly a result of selective forces, senescence, i.e., biological aging, seems not to be shaped by natural selection, but rather a by-product of earlier neutral or beneficial gene expression.

Evolutionary Assumptions

Evolutionary theorizing considers striving for optimal reproductive success as the ultimate goal, also, of human development. This assumption specifies that not only the proximate regulation of physical and biological systems, but also the social behavior and complex psychology of human beings follow a reproductive logic. Reproductive success is expressed in terms of inclusive fitness, which comprises the individual's own procreation (Darwinian fitness) and the procreation of relatives with whom the individual shares genes (indirect fitness), thus establishing the principle of kin selection (Hamilton, 1964). Therefore, the evolutionary perspective focuses on the gene as a unit of analysis. However, "… it is not the naked gene, that is exposed to selective forces directly" (Mayr, 1994: p. 206), but rather the "realized animal, that lives or dies, breeds or helps the relatives" (Daly and Wilson, 1983: p. 32). Although individuals are considered as 'vehicles' of their genes (Dawkins, 1976), the importance of the phenotypical individual for determining reproductive success has recently been (re)emphasized (Laland et al., 2000).

However, this position has been challenged since some time, claiming that the group selection is also favored from an evolutionary perspective. In an article in Brain and Behavioral and Brain Science (2014), Smaldino argues that the level of group organization has to be taken into consideration since properties of particular human groups may represent evolutionary advantages over other human groups.

According to Smaldino's (2014) group-level traits, as characterized by the structured organization of differentiated individuals, represent a mode of selection that is different from traditional group selection approaches (multilevel selection theory). The implied focus on cooperative group activities (compared to a focus on competition that is implied by the assumption that genes drive selfish interests (genetic egoism perspective)) is compatible with the cooperative nature and the shared intentionality in social actions that very young human infants already display, e.g., emphasized by Michael Tomasello (1999) (for a critique of this group selectionist approach see Pinker, 2012, in his Edge essay, The false allure of group selection, http://edge.org/conversation/the-false-allure-of-group-selection) and the interesting comments by leading scholars in the field (the reality club). It will be a matter of empirical proofs to test hypotheses from these competing theoretical points of view.

It is, however, commonsense that it is not the adult individual who represents the product of evolution; rather, the whole life span with the different developmental phases is shaped by selective forces with the eventual exception of senescence, which itself seems not to be a product of selection. Aging is a concomitant of sexually reproducing individuals. The inevitability of death due to environmental stressors during human history has resulted in a decrease of individuals with increasing age, so that the power of natural selection also decreased (Medawar, 1952).

Life history theory (Alexander, 1990) regards the developmental course of the individual life span as the result of a trade-off between investment into one's own growth and development (somatic effort) and the investment into reproduction, comprising the functional systems of mating and parenting (reproductive effort; Chisholm, 1996). Somatic efforts accumulate reproductive potential whereas reproductive efforts decrease the residual reproductive value. Evolution has thus shaped humans into reproductive strategists who constantly have to make 'decisions' about the best allocation of investment (Voland, 1998; Shenk, 2011). The heuristics of reproduction are contingent upon environmental demands, which comprise material and ecological forces as well as social complexity (see Dunbar, 1996). Especially, the evolution of intelligence as a necessary prerequisite for exploring environmental resources and regulating the social behavior of expanding groups favored learning as a second asset of adaptation. Yet learning is not random, but follows (genetically imposed) biases and constraints ('central tendencies,' MacDonald, 1997; 'epigenetic rules,' Lumsden and Wilson, 1981) to acquiring specific environmental information during specific phases of development allowing easy learning at these points in time (Boyd and Richerson, 1985).

Psychological development across the life span can simultaneously be interpreted as a structured sequence of developmental tasks in order to improve social and cognitive skills as well as the formation of a reproductive style. Evolutionary socialization models have described the ontogeny of individuals' reproductive styles as a meaningful passage through the

developmental tasks of childhood and adolescence. In particular, the resource situation and family climate during early childhood are considered to be predictive of attachment formation. The quality of attachment in turn influences psychological adjustment during late childhood. Psychological development also influences neurophysiological and maturational processes, leading to the somatic consequences of defining the beginning of puberty. The beginning of puberty channels the reproductive style; especially with regard to the age at which first giving birth, the spacing, and number of children with the respective implications for parental investment (Belsky et al., 1991; Keller, 2007, 1996). Empirical studies from different parts of the world have supported the view of different developmental trajectories as specified by these models (Keller and Kärtner, 2013). The new perspective of evolutionary theorizing is thus the assumption that neurophysiological and psychological regulations of behavioral development (proximate causes) are influenced by adaptive, i.e., shaped through the history of humankind, mechanisms that serve the ultimate function of optimizing reproductive success. The postreproductive phase is therefore not subject to the same degree of selection pressure as the prereproductive and reproductive life phases (Baltes, 1997; Geary and Bjorklund, 2000). Accordingly, old age is not a distinct phase in stage conceptualizations of the life history approach (Bogin, 1997).

It also remains an open debate in evolutionary approaches, which are the principles that guide these trade-offs. Especially, discussions about the morality of humans and conscience raise questions about the prevalence of a trade-off concerning reproduction or a genuine social orientation (Voland, 1998).

In the following sections, the functional significance of the different developmental phases for reproductive success will be analyzed. The developmental phases of infancy, childhood, adolescence, adulthood, and old age will be briefly characterized with respect to their possible evolutionary roots. Old age is included in this analysis, since the assumption of structural continuity implies that earlier experiences shape later ones, so that the quality of old age could be regarded as a consequence of the earlier reproductive life (Keller, 1991).

Developmental Phases across Life Span

Infancy

The extreme helplessness ('altriciality') of the human newborn can be considered as part of a reproductive strategy to becoming a better adult ('better adult hypothesis,' Alexander, 1990), since it allows the baby to invest all available resources in growth and development. This view is in contrast with the common interpretation of infants' helplessness as a consequence of a physiological preterm birth, due to hominid brain growth (Prechtl, 1984). Early development is shaped by the investment that parents allocate to each child individually, depending on their own and the child's reproductive values (e.g., age of parent, social support, sex, and birth position of child). Early experiences form the basis of the psychological adaptation to the respective physical and social environment in prompting phenotypic development. Accordingly, infants adopt internal representations of social relationships (*see* Integrated Attachment Theory; Attachment during Infancy and Early Childhood: Cultural Variations, Development of) and a primary conception of the self, focusing on the expectations of reciprocity as the balance of give and take that can be expected in future social relationships.

Childhood

At about 2 years of age, a developmental transition with a new focus on the extension of the physical and social range of children's activities can be observed in many cultural contexts. This marks the beginning of children's apprenticeship that involves the adoption of promising strategies for acquiring resources in their particular environment. Mediated through the culturally primed beliefs and values ('ethnotheories') of their social environmental and social competence through participatory learning as well as by adopting and creating developmental niches ('niche picking'), which include the implicit and explicit reflection of sex differences as concomitants of adult reproductive life.

Adolescence

Adolescence constitutes the transition from childhood to adulthood. The beginning of adolescence is universally defined by entry into puberty. The duration of adolescence, however, differs substantially across cultures, varying from transitional rituals during a few weeks to a stage that last several years, depending mainly on economic constraints and educational provisions (see Schlegel and Barry, 1991). The attainment of sexual maturity is a biological marker variable with a hereditary component and is contingent upon ecological factors, such as nutritional status and physical condition, as well as previous socialization experiences, especially the security and stability of the early home environment. At the same time, this marker variable is believed to be predictive of future reproductive decisions (e.g., age at first sexual intercourse and at first childbirth, duration of interbirth intervals). Although similar developmental trajectories for males and females have been empirically identified, sex-specific reproductive strategies with their respective costs make it likely that the environmental factors influence female fecundity more strongly than male. Therefore, and owing to the greater accessibility of menarche as an empirical measure, female developmental trajectories are overrepresented in the literature. Psychosocial stress in the family of origin, especially marital conflict and father's absence, has been documented as decreasing the age of menarche in different contextual systems. Particularly, the divorce of parents during a child's first 5 years of life could reliably be related to early menarche (Chasiotis et al., 1998). The perception of this environmental scenario may support the development of a more quantitative-oriented reproductive style with lesser parental investment whereas marital stability could be considered as part of a qualitative style with higher maternal and paternal investment. Thus, low-ecological stress (i.e., good nutritional status and a balanced physical condition) as well as high-psychosocial stress (e.g., divorce of parents before the age of 5 years) can result in earlier maturation, whereas high-ecological stress and/or low-psychosocial stress can lead to later maturation. There is also

evidence that prenatal factors influence the age of menarche, e.g., being born smaller accelerates menarche (Gluckman and Hanson, 2006). The main reproductive function of the variability of sexual maturity is to reduce the costs of premature reproduction. If an individual 'waits' for sexual maturity, the resources can be allocated to growth, which will increase the reproductive value at maturity. On the other hand, the delay might increase the probability of surviving until the reproductive age (Charnov, 1989).

The onset of reproduction varies with the duration of adolescence across cultural contexts. Especially in traditional rural societies, reproduction begins soon after sexual maturity and seems to be more quantitatively oriented, whereas reproduction in urban-educated societies is delayed until individuals complete education and seems to be more qualitatively oriented. It is a novel phenomenon that there is a mismatch in timing between the age of menarche (earlier) and the achievement of psychosocial maturity (later) (Gluckman and Hanson, 2006). This mismatch may be indicative of another sign of separation between genetic and cultural evolution, which poses further challenges to the understanding of the evolutionary bases of behavioral development.

Adulthood

In evolutionary terms, adulthood is the phase of reproduction with the task to resolve various allocation conflicts. First, the conflict between current and future reproduction (i.e., investment in growth or reproduction) has to be addressed, culminating in the decision of when to start reproduction, and the related issue of nepotistic effort, i.e., investing in one's own offspring or temporarily (or even permanently) adopting the role of a 'helper in the nest' (Voland, 1998). The second allocation conflict concerns the question of investment in quality or quantity of offsprings, which determines the number and spacing of children. Since these decisions not only have to be made generally as an expression of reproductive style, but also individually for each offspring separately, parent–child conflict is an inevitable consequence (Trivers, 1974). Survival and psychological adaptation are facilitated when not only maternal, but also paternal investment is spent on the infant. It is assumed that partner bonding and family cohesion evolved as psychological mechanisms to facilitate the development of 'better adults.' This view seems to confirm the relationship between father's absence and (increased) infant mortality in traditional societies, e.g., among the Ache Indians in South America (Hill and Hurtado, 1996).

The determinants of reproductive success differ for males and females. Females per se make a higher investment owing to intrauterine pregnancy and nursing, supporting a qualitative parenting style with higher investment in fewer offspring. They have to decide whether they want to continue to look for a better mate or start reproducing. The prototypical male strategy, on the other hand, should be more quantitative, that is, to invest less in more offspring, due mainly to the fact that it is difficult for a male to know precisely whether he is the father of an infant ('paternity insecurity'). Moreover, males show greater variability in their reproductive rates, from having no children to having many children. Men have to decide between the 'cad versus dad' strategies, i.e., maximizing fertilization or parental investment (Voland, 1998). The trade-off that individuals have to accept therefore entails the balance of mating and parenting efforts. Studies of partner selection have documented nevertheless universal preferences for the two sexes: females prefer partners with characteristics indicative of economic success, whereas males prefer females with characteristics that indicate reproductive capacity (youth, health). In addition, there is preference by both males and females for mates who are similar to themselves in a variety of characteristics ('assortative mating,' Wolf and Figueredo, 2011). Parental investment constitutes the intergenerational link through which the offspring experiences the socialization style that the parents have developed as a result of their own past as well as current experiences. In the case of human beings, not only time and energy as parameters of investment decisions have to be taken into consideration, but also the psychology of the phenotypic adult (Keller, 2000).

Old Age

Old age covers the postgenerative phase of the life span. Yet, it has been proven meaningful to differentiate a third young-old age span form about 60–75 years where the vital capacity has improved to compensate for about 5 years due to cultural evolution, from a fourth age span from around 80 to 100 years (cf Baltes, 1997), which has emerged due to increasing life expectancy, but where senescence has predominantly detrimental effects. These might originate as by-products from possibly two processes: (1) selective neutrality of the genome with no effect at early life and deleterious effects at later life and (2) antagonistic pleiotropic gene action that supports reproductive success during the reproductive life stages (Martin et al., 1996). One detrimental factor that has been identified across species may be metabolic oxidative damage operating in macromolecules. On the other hand, it has been argued that the extension of the life span serves individual fitness by supporting the development of the child's and grandchild's generation (Johow and Voland, 2013). Although this 'grandmother' hypothesis (Peccei, 1995) is controversial, researchers nevertheless report that being a grandparent increases life satisfaction in old age. The physical deterioration especially of the oldest-old age requires extrasocietal cultural resources, which at the same time become more difficult to be exploited effectively. Accordingly, the regulation and compensation of losses become a major mechanism for ontogenetic processes (Baltes, 1991). Relating interindividual differences in regulatory competencies to the earlier reproductive life histories might be a challenge for life span psychology.

Outlook

Although essential questions such as the interplay of genetic and social influences on phenotypic development remain largely open, the evolutionary view on the human life span offers a stimulating perspective on many controversial issues of psychological development, e.g., the nature of sex differences or the normative character of secure attachment relationships. During their ontogenetic development, individuals, based on

universal genetic programs (e.g., for language acquisition, partner preferences, parenting styles), acquire the contextually relevant ecological, social and cultural information (e.g., mortality rate of their population, economic constraints, cultural values) both intentionally and intuitively to promote their psychological development so that it is adaptive for the reproductive style that promises best inclusive fitness in the given situation. Old age challenges the cultural nature of humans in a very special way, since trade-offs are no longer associated with epigenetic rules or central tendencies. Humans therefore have to use the accumulated knowledge of the previous life stages in order to redefine and redirect the allocation of resources and balance universally evolved behavioral heuristics, ontogenetically acquired styles, and situational regulations, making each life history unique. The question of assumed ultimate causes, "how do these behavioral regulations or these motivational systems contribute to reproductive success," might allow us to reframe empirical findings and adapt research designs accordingly. Evolutionary approaches face many challenges due to rapidly changing environments and the concomitant separation of genetic and cultural evolution. The declining fertility with increasing resources is such a challenge (see Mace, 2000). For the future discourses, multidisciplinary perspectives will be crucial.

See also: Attachment During Infancy and Early Childhood: Cultural Variations, Development of; Cognition, Evolution of; Emotion, Evolution of; Emotions, History of; Evolution and Language: Overview; Human Cognition, Evolution of; Integrated Attachment Theory; Intelligence: History of the Concept; Lifespan Development, Theory of; Midlife Psychological Development.

Bibliography

Alexander, R.D., 1990. Epigenetic rules and Darwinian algorithms – the adaptive study of learning and development. Ethology and Sociobiology 11, 241–303.

Baltes, P.B., 1991. The many faces of human aging: towards a psychological culture of old age. Psychological Medicine 21, 837–854.

Baltes, P.B., 1997. On the incomplete architecture of human ontogeny: selection, optimization, and compensation as foundation of developmental theory. American Psychologist 52, 366–380.

Belsky, J., Steinberg, L., Draper, P., 1991. Further reflections on an evolutionary theory of socialization. Child Development 62, 682–685.

Bogin, B., 1997. Evolutionary hypotheses for human childhood. Yearbook of Physical Anthropology 40, 63–89.

Boyd, R., Richerson, P.J., 1985. Culture and Evolutionary Process. University of Chicago Press, Chicago.

Charnov, E.L., 1989. Natural selection on the age of maturity in shrimp. Evolutionary Ecology 3, 236–239.

Chasiotis, A., Scheffer, D., Restemeier, R., Keller, H., 1998. Intergenerational context discontinuity affects the onset of puberty: a comparison of parent-child dyads in West and East Germany. Human Nature 9 (3), 321–339.

Chisholm, J.S., 1996. The evolutionary ecology of attachment organization. Human Nature 7 (1), 1–38.

Daly, M., Wilson, M., 1983. Sex, Evolution and Behavior, second ed. PWS Publication, Boston.

Dawkins, R., 1976. The Selfish Gene. Oxford University Press, Oxford.

Dunbar, R., 1996. Grooming, Gossip and the Evolution of Language. Faber & Faber, London.

Geary, D.C., Bjorklund, D.F., 2000. Evolutionary developmental psychology. Child Development 71 (1), 57–65.

Gluckman, P.D., Hanson, M.A., 2006. Evolution, development and time of puberty. Trends in Endocrinology & Metabolism 17 (1), 7–12.

Hamilton, W., 1964. The genetical evolution of social behaviour (I + II). Journal of Theoretical Biology 7, 1–52.

Hill, K., Hurtado, A.M., 1996. Ache Life History. The Ecology and Demography of a Foraging People. Walter de Gruyter, New York.

Johow, J., Voland, E., Willführ, K.P., 2013. Reproductive strategies in female post-generative life. In: Fisher, M.L., Garcia, J.R., Sokol Chang, R. (Eds.), Evolution's Empress – Darwinian Perspective on the Nature of Women. Oxford University Press, Oxford, pp. 243–259.

Keller, H., 1991. A perspective on continuity in infant development. In: Lamb, M.E., Keller, H. (Eds.), Infant Development: Perspectives from German Speaking Countries. Erlbaum, Hillsdale, pp. 135–150.

Keller, H., 1996. Evolutionary approaches. In: Berry, J.W., Poortinga, Y.H., Pandey, J. (Eds.), Handbook of Cross-Cultural Psychology, Theory and Method, vol. 1. Allyn & Bacon, Boston, pp. 215–304.

Keller, H., 2000. Human parent-child relationships from an evolutionary perspective. American Behavioral Scientist 43 (6), 957–969.

Keller, H., 2007. Cultures of Infancy. Erlbaum, Mahwah, NJ.

Keller, H., Kärtner, J., 2013. Development – the culture-specific solution of universal developmental tasks. In: Gelfand, M.L., Chiu, C.-Y., Hong, Y.Y. (Eds.), Advances in Culture and Psychology, vol. 3. Oxford University Press, Oxford, pp. 63–116.

Laland, K.N., Odling-Smee, F.J., Feldman, M.W., 2000. Niche construction, biological evolution, and cultural change. Behavioral and Brain Sciences 23, 131–146.

Lumsden, C.J., Wilson, E.O., 1981. Genes, Mind and Culture: The Coevolutionary Process. Harvard University Press, Cambridge.

MacDonald, K.B., 1997. Life history theory and human reproductive behavior. Environmental/contextual influences and heritable variation. Human Nature 8 (4), 327–359.

Mace, R., 2000. Evolutionary ecology of human life history. Animal Behavior 59, 1–10.

Martin, G.M., Austad, S.N., Johnson, T.E., 1996. Genetic analysis of aging: role of oxidative damage and environmental stresses. Nature Genetics 13, 25–34.

Mayr, E., 1994. Evolution – Grundfragen und Mißverständnisse (Evolution – basic questions and misunderstandings). Ethik und Sozialwissenschaften 5 (2), 203–209.

Medawar, P.B., 1952. Unsolved Problems of Biology. Lewis, London.

Peccei, J., 1995. The origin and evolution of menopause: the altriciality-lifespan hypothesis. Ethology and Sociobiology 16, 425–449.

Pinker, S., 2012. The False Allure of Group Selection. http://edge.org/conversation/the-false-allure-of-group-selection (accessed 29.04.13.).

Prechtl, H., 1984. Continuity of Neural Functions from Prenatal to Postnatal Life. Spastics. International Medical Publications, London.

Shenk, M.K., 2011. Parental decisions – how much to invest in your offspring. In: Frey, U.J., Störmer, C., Willführ, K.P. (Eds.), Essential Building Blocks of Human Nature. Springer, Berlin & Heidelberg, pp. 17–38.

Schlegel, A., Barry III, H., 1991. Adolescence. Maxwell Macmillan International, New York.

Smaldino, P.E., 2014. The cultural evolution of emergent group-level traits. Behavioral and Brain Sciences 37, 243–254. http://dx.doi.org/10.1017/S0140525X13001544.

Tomasello, M., 1999. The Cultural Origins of Human Cognition. Harvard University Press, Cambridge.

Trivers, R.L., 1974. Parent-offspring conflict. American Zoologist 14, 249–264.

Voland, E., 1998. Evolutionary ecology of human reproduction. Annual Review of Anthropology 27, 347–374.

Wolf, P.S.A., Figueredo, A.J., 2011. Fecundity, offspring longevity, and assortative mating: parametric tradeoffs in sexual and life history strategy. Biodemography and Social Biology 57 (2), 171–183. http://dx.doi.org/10.1080/19485565.2011.614569.

Life-style, History of the Concept

Johanna Muckenhuber, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term lifestyle is widely used in sociology and many other subdisciplines of the social sciences. Alfred Adler and Max Weber can be named as the first who developed the term and the concept.

Lifestyle can be defined as a set of behaviors, preferences, and practices, which are present in daily life.

In modern social theory, Beck and Bourdieu further developed the concept from an agency perspective (Beck) and from a more structural perspective (Bourdieu).

The term lifestyle is used in many empirical studies in different subdisciplines of the social sciences, but in a very broad way, in many cases without reference to the underlying theoretical assumptions.

The term lifestyle gained importance during the last decades. Having its origins in sociological and psychological theory, it was further developed, operationalized, included in empirical research, and finally distributed to many different disciplines. The aim of this article is to describe the origins of the term lifestyle and to describe how it is used in different disciplines of the social and behavioral sciences, sociology, health sciences, and leisure studies, in particular.

Definition

The concept of lifestyle was developed first by the German sociologist Max Weber (see Weber, 1968, 1972) and by the Austrian psychologist and psychoanalyst Alfred Adler (see Adler, 1924, 1933).

The Oxford English Dictionary (OED) (OED, 2014) offers several definitions of lifestyle. According to it lifestyle in general today can be defined as "a style or way of living (associated with an individual person, a society, etc.); esp. the characteristic manner in which a person lives (or chooses to live) his or her life." In Adlerian psychology, it is "a pattern of reactions and behavior that is established in childhood and remains characteristic of an individual" (OED, 2014). In addition, the OED refers to compounds of lifestyle "relating to the way in which one lives (or chooses to live) one's life, esp. with regard to quality of life, as lifestyle advice, lifestyle change, lifestyle factor, lifestyle group, lifestyle option, etc." and in the specific case of marketing "relating to a publication, a product, advertising, etc., which is designed to appeal to consumers by association with a lifestyle regarded as desirable, glamorous, or attractive, as lifestyle advertising, lifestyle brand, lifestyle magazine, lifestyle marketing, etc." Finally in related to health, OED refers to lifestyle as "lifestyle drug n. a drug prescribed to treat a condition that is not necessarily serious or life-threatening but that has a significant impact on the quality of life" (OED, 2014).

The term lifestyle was and is used in different ways. One important aspect of differentiation is the actor who is addressed. Some authors refer to individuals with a certain lifestyle, some authors refer to groups of persons (which are at least partly stable over time), and other authors refer to a certain class and its lifestyle. The class is not a group with individuals who know each other. Individuals of the class have a set of properties in common (see Ansbacher, 1967: p. 11).

The individual lifestyle is defined as a kind of stable characteristics and even more, most actions of the individuals are interpreted as to fit in a relatively consistent manner to their lifestyle.

The definition of the group lifestyle differs according to the authors. Some authors refer to the dynamics and specific constellations of a group when writing about group lifestyle. Others like Parsons (see Parsons and Bales, 1955) refer more to classificatory systems. Parsons, for example, argues that the husbands' and fathers' occupation is central in determining the lifestyle and in this context the status of a family.

Concepts of lifestyle, which belong to a certain class, are often used to describe and to analyze status groups and occupational groups, but cultures and subcultures as well. Thereby they refer to the concept of class but go beyond class. Max Weber was one of the sociologists who used the term lifestyle in this way. In addition, these definitions can be found in more recent theories. In this article, they are discussed in the context of sociological theory and research and in the context of the use of the term lifestyle in other disciplines later on.

Lifestyle by Alfred Adler

Alfred Adler as psychologist and psychoanalyst focused on the individual dimension of lifestyle. He argued that individuals' actions are purposive, goal directed, throughout consistent, and unique for each individual. Therefore following Adler, all actions are ultimately subjectively determined.

Beginning with 1926, Adler started to use the term style of life. In 1929, he officially abandoned the term life plan to use lifestyle instead (see Ansbacher, 1967). In the development of the concept, Adler seems to be influenced by Folkert Wilken who worked together with him. According to Ansbacher, Adler describes lifestyle as follows: "In considering the structure of a personality, the chief difficulty is that its unity, its particular style of life and goal, is not built upon objective reality, but upon the subjective view the individual takes of the facts of life" (Ansbacher, 1956). But,

nevertheless, pointing out the importance of the concept of unity, Adler argued that it is also important to take into consideration the context of the individuals when analyzing their lifestyle.

Following Adler, the lifestyle develops in the first years of the life of a child. Adler pointed out the importance of early childhood. He argues that the particular style of life even of a child "is not built upon objective reality, but upon the subjective view the individual takes upon the facts of life" (Adler in Ansbacher, 1967: p. 191).

Adler as a psychotherapist was mainly concerned with personality change. According to him, psychotherapy needs to investigate the patient's "present field of action and then proceed to his general movement through life, to his lifestyle" (Dreikurs, 1961: p. 89). In this context, he writes about the life plan, which guides the persons through their life. The life plan covers the idea of a goal in each individual's life and the idea of a coherence of all partial movements during the life course (Adler, 1924: p. 4). Adler was very close to some modern sociological theories when writing that everybody has a kind of subconscious idea about himself, what he or she should do in, and with his/her life without really understanding and knowing what this is (Adler, 1933).

But it was Max Weber who developed the concept of lifestyle in sociology, more or less at the same point in time as Alfred Adler developed his concept.

Lifestyle and 'Lebensführung' by Max Weber

Max Weber in contrast to Adler focused on a collective dimension of lifestyle. He differentiated between 'Lebensführung,' 'Lebenschancen,' and 'Lebensstil.' 'Lebensführung' can be literally translated as life conduct and refers to choice and self-direction in a person's behavior. Therefore, according to Weber, life conduct is an element of choice within his overall concept of lifestyles. The second basic component of the lifestyle is life chances (Lebenschancen), which deal with aspects that cannot be changed voluntarily by the persons. Life chances therefore cover the idea of the importance of structure for agency. These life chances include the probability that a person can realize his or her choices in life and the importance the economic situation of the individuals have on their life conduct. Weber pointed out that "the possibility of status-specific life conduct is of course in part economically conditioned" ("Denn die Möglichkeit, ständischer Lebensführung pflegt naturgemäss ökonomisch mitbedingt zu sein") (Weber, 1972: p. 537). 'Lebensstil' combines both Lebensführung and Lebenschancen. In the process of translation to English, the term 'Lebensführung' and the term 'Lebensstil' were mistranslated according to Abel and Cockerham (Abel and Cockerham, 1993) because both were translated with the same English term 'lifestyle.'

According to Ansbacher, Max Weber used the term lifestyle before Adler. Weber referred to a collective phenomenon and its meaning was quite similar to that of subculture (Ansbacher, 1967: p. 196). Differences in lifestyles are mainly explained by educational differences and Weber points out, that also the status of a person is expressed by a specific style of life (see Weber, 1968).

Lifestyle in the Social Sciences

Sociological Theory and the Concept of Lifestyle

In sociology, the concept of lifestyle is associated with the description of groups, subcultures, occupational or educational groups, and is mostly related to a certain status of a group.

In sociology, the concept of lifestyle is argued to be closely related to the position of the individuals in the sociostructural environment like the economic resources and the level of education.

Thorstein Veblen can be cited as an American sociologist who developed a concept of the lifestyle in the late-nineteenth century. In his theory of the leisure class, he develops the idea that individuals who belong to the upper class do not only own their riches, but that for them it is also important to show their wealth. Veblen develops the idea of a kind of conspicuous consumption. He argued that this conspicuous consumption is important for the upper class in order to keep their status (Bögenhold, 2003; Veblen, 1899). It has been argued that in modern societies this kind of consumption and demonstration of the wealth has changed. Partly this can be refuted, for example, when regarding the ostentatious demonstration of wealth in yachts. But in other aspects, such a change can be observed where wealth is not easily identifiable for everybody but when hidden symbols identify wealth and the belonging to the subgroup of the wealthy just for those who know about these symbols. This knowledge about the symbols of distinction is closely related to Pierre Bourdieu's idea of distinction. Veblen as well as Bourdieu follows the idea that the economic class is strongly related to certain practices and sets of preferences, which constitute the lifestyle of subgroups (Trigg, 2001).

Modern sociological approaches to lifestyle can be separated into those who argue that class and position in social structure are important factors determining individuals' lifestyles, and individualistic approaches who deny the role of class and structural influences.

Pierre Bourdieu is one of the authors who points out the importance of class and social structure. In Bourdieu's theory about social inequality, the concept of lifestyle is closely related to his concept of the habitus. Bourdieu's interest was to develop a theory, which builds a link between structuralism and the main ideas of agency theory. He argued against the dichotomy of both. His concept of the habitus is seen as a missing link, which can explain the pathways how social structure influences human agency. Bourdieu argues that individuals do not dispose of different amounts of economic capital, which can be measured in terms of monetary income and economic possessions but, that persons also dispose of social and cultural capital. Social capital is defined as the "the aggregate of the actual or potential resources an individual has access to" (Bourdieu, 2005: p. 51). These resources depend on the belonging to certain groups. They can be based on material or symbolic exchange relationships and can be more or less institutionalized. In other words, the resources can be based on subjective feelings like recognition, respect, and friendship, or on institutionalized guaranties like legal claims (Bourdieu, 2005: 1:63ff). When writing about cultural capital, Bourdieu refers to the individuals' level of education but also to the knowledge of individuals about cultural and intellectual traditions. This knowledge is crucial in order to develop practices

and to behave in a way, which fits to the class they belong to and, which fits the expectations linked to their position in the social field. According to Bourdieu, individuals are situated in a social field, which is strongly influenced by their economic, social, and cultural capital. This position in the social field has a strong influence on the habitus of a person. The habitus in turn produces a certain lifestyle. Its manifestations can be seen in certain cultural practices (Fuchs-Heinritz et al., 1994: p. 394). These cultural practices comprise, for example, what the persons read, their food preferences, their furnishing, and their preferences for music styles etc. Therefore following Bourdieu the lifestyle is a very important category of distinction in the social interplay between individuals and groups.

Ulrich Beck follows an approach quite different to Bourdieu's theory. Beck is a representative of the individualization theory. He argues that in developed societies we can find rising levels of income, education, social security, and mobility. As a consequence individuals can dispose of more resources than in former times. This in turn leads to a broader and still broadening range of opportunities of individual action. According to this theory, life can be shaped and organized more and more by the individuals. This has two consequences. On the one hand, the persons have more freedom to make decisions and on the other hand, they have the pressure to decide for their own lives with consequences for their present and their future opportunities. Therefore theorists of individualization argue that individuals nowadays feel more anxieties to fail. According to Beck and others, one of the consequences of this situation is that individuals do not develop more and more different forms and styles of life, but in contrast they orient themselves toward similar subgroups and adopt lifestyles, norms, and values of these groups to a large extent. Summing up, it can be stated that following theories of individualization, lifestyles can be chosen by the individuals and are not determined by the position in the social structure, but the choice of a certain lifestyle is influenced by the personality and the subgroup a person belongs to and wants to belong to (Hradil and Spellerberg, 2011).

When comparing structuralist theories like Bourdieu's with his concept of the habitus and theories following the concept of individualization, it is obvious that they differ in some important aspects and lead to different assumptions regarding modern lifestyles (Hradil and Spellerberg, 2011).

1. While Beck's theory of individualization assumes conscious decisions for certain lifestyles, Bourdieu's theory of the habitus assumes more unconscious processes, which have specific practices and lifestyles as a consequence.
2. Theories of individualization assume that lifestyles are a matter of choice and these choices are influenced mainly by individual preferences. In contrast, Bourdieu's theory assumes that the habitus and with this the lifestyle is determined (or if not determined so influenced to a great extent) by material resources an individual disposes of.
3. When following theories of individualization lifestyles should get more diverse over the time, they should be independent from class and social structure, and should be changeable. More structuralist theories, like Bourdieu's, lead to the assumption of relatively stable lifestyles, which differentiate between individuals according to their position in the social structure.

Both theoretical currents have many followers. But empirical findings are as diverse and ambiguous as the theories are. Therefore no decision in favor of one of the two theories can be taken.

Lifestyle in Different Disciplines of the Social Sciences

In different disciplines of the social sciences, the concept of lifestyle gained importance in the last decades and still gains importance. As it is not possible to give an overview on the use of the term lifestyle in all disciplines of the social sciences in the following paragraphs, it will be depicted exemplarily for health sciences and for the field of leisure studies.

In Health Sciences

In health sciences, the term lifestyle has been used for some decades now. In particular, public health studies and applied sciences in health promotion use the concept. In contrast to most sociological research, many studies in health sciences are strongly empirically oriented. Regarding the theoretical background there is a split in health sciences. Some research in social medicine assumes that social structure strongly influences health behavior and health. This research shows and showed that the socioeconomic status of a person (SES) is an important determinant of health. Other research refers to agency theories and points out the capacity of individuals to choose their behavior. Lifestyles relevant for health comprehend behavior like smoking habits, diet, exercise, alcohol consumption, etc. These lifestyles can either be seen as a consequence of structural conditions or as a consequence of rational choice. Past health research focused on biological and psychosomatic as well as structural explanations for health. But in late modernity, health has been interpreted more and more as an achievement, which can be reached by the individuals when following certain rules of behavior, when adopting a healthy lifestyle. This change of a paradigm regarding causes of health and ill-health has to be seen in the context of changing disease patterns, modernity, and an adjustment in social identity. While during past centuries acute diseases constituted the major burden for health, during the last century chronic diseases gained importance and now account strongest for mortality. These chronic diseases can be influenced to a large degree by lifestyle in contrast to acute diseases. Modernity and changing social identities have as a consequence that theories gained importance, which postulate that individuals are agents of their lives who are free to choose their lifestyles (Beck, 1992).

Regarding the evolution of the term lifestyle in health sciences we can see that the meaning changed very much when compared to the way the term was initially conceptualized by Weber and Adler. Both argued that the lifestyle is related to a certain position within the social system, to a certain class and status. Most research in health sciences now postulates free choice of lifestyle regardless income, level of education, gender, or ethnicity. Structural influences on lifestyle are denied (see Cockerham, 2005). As a consequence, most recommendations for public health action are about public health interventions to reduce risk behavior like smoking and to enforce health behavior like exercise and healthy diet. These public health interventions often fail. They reach individuals with high SES

(high income and high education) who already pursue a healthy lifestyle, but they fail to reach low-SES individuals who maintain their unhealthy lifestyle.

More recent approaches referring to Weber and Bourdieu are important for health lifestyle research. These approaches include the interplay between life chances and life choices, the interplay between structure and agency in their research on reproduction of and on possibilities to change health lifestyles (Cockerham, 2005). Some research on health lifestyle includes the idea of the importance of social capital for health. It has been shown that social capital is an important determinant of health but the corresponding theoretical approaches differ strongly. Most studies refer to Robert Putnam (Putnam, 2000) and Ichiro Kawachi (Kawachi et al., 2008) who are representatives of agency theory. As an element of structure they refer to affiliations to concrete groups, whereas Bourdieu refers to the belonging to a class, to individuals who do not know each other, but have certain properties in common and access to the same resources. This difference between group and class is one of the main differences between the different definitions of social capital.

Even if it has to be stated that still most research on health lifestyles is strongly committed to agency theory, some research (see for example Cockerham, 2005) argues that living conditions, socialization, and experience as well as life chances in general should be studied in their interplay with life choices in order to analyze health lifestyles. The concept of the habitus as dispositions for life choices, actions, and practices of daily life seems to gain importance in health lifestyle studies.

Lifestyle in Leisure and Market Sciences

Research on lifestyle in leisure studies is mostly empirical and few studies refer to social theory as it is the case in health sciences and most other social science disciplines. Most lifestyle typologies in leisure and market research are constructed by means of quantitative methods like factor- and cluster analyses.

In marketing research, lifestyle is an important topic since the 1960s. The interest in marketing research is mostly oriented on variation in buying behavior with the aim to reach better results in marketing. In the 1960s, marketing research found that there was more variation in buying behavior within social classes than between them. Studies resulted in typologies like the VALS (Values, Attitudes and Lifestyles) typology of nine American lifestyle groups. Typologies like this one describe groups of distinctive consumption patterns and leisure behavior.

Other approaches in leisure and lifestyle research point out the importance of space and geographical location of the persons. Studies on suburban lifestyle have to be named in this context. Spatial approaches on lifestyle are concerned to build typologies including housing conditions, socioeconomic and sociodemographic factors, and their influence on leisure and consumption behaviors. The underlying assumption of spatial approaches is a spatial embeddedness of these behaviors and the idea that they differ between persons and groups living at different locations.

The different approaches to research on lifestyle in leisure and market studies have in common, that most of them do not focus on or not even take into account class and status position of the persons and that they do not include theoretical assumptions on social structure (see Veil, 2000).

Conclusion

Concluding it can be stated that the term lifestyle gained importance since it has been conceptualized by Weber and Adler, and it still gains of importance in many different disciplines of the social sciences like sport sciences, sustainability studies, the field of health studies, and sociology. But even if the term is used, its meaning differs between and within the disciplines. Therefore it can be argued, that the term spreads but with a broad meaning, which in some contexts seems to be very far away from its initial meaning.

See also: Habitus: History of a Concept; Individuals in Society: History of the Concept; Social Structure: History of the Concept.

Bibliography

Abel, Thomas, Cockerham, William C., 1993. Critical remarks on the mistranslation of Weber's, class, status, party. The Sociological Quarterly 34 (3), 551–556.

Adler, Alfred, 1924. Praxis und Theorie der Individualpschologie, second Aufl. Bergmann, München.

Adler, Alfred, 1933. Der Sinn des Lebens. Fischer, Frankfurt am Main.

Ansbacher, Heinz L., 1956. The Individual Psychology of Alfred Adler. Basic Books, New York.

Ansbacher, Heinz L., 1967. Life style: a historical and systematic review. Journal of Individual Psychology 23 (2), 191–212.

Beck, Ulrich, 1992. Risk Society: Towards a New Modernity. Übersetzt von Mark Ritter. Sage, London.

Bögenhold, Dieter, 2003. Social inequality and the sociology of life style. Material and cultural aspects of social stratification. American Journal of Economics and Sociology 60 (4) , 829–847.

Bourdieu, Pierre, 2005. Die verborgenen Mechanismen der Macht. In: Schriften zu Politik und Kultur, Bd. 1. VSA-Verlag, Hamburg.

Cockerham, William C., 2005. Health lifestyle theory and the convergence of agency and structure. Journal of Health and Social Behavior 46, 51–67.

Dreikurs, Rudolf, 1961. The Adlerian approach to therapy. In: Stein, M.I. (Ed.), Contemporary Psychotherapies. Free Press of Glencoe, New York.

Fuchs-Heinritz, Werner, Lautmann, Rüdiger, Rammstedt, Otthein, Wienold, Hanns (Eds.), 1994. Lexikon zur Soziologie, third völlig neu bearbeitete und erweiterte Auflage. Westdeutscher Verlag, Opladen.

Hamilton, R.F., 1964. The behavior and values of skilled workers. In: Shostak, A.B., Gomberg, W. (Eds.), Blue-Collar World, pp. 42–57. Prentice-Hall, Englewood Cliffs, NJ.

Hradil, Stefan, Spellerberg, Annette, 2011. Lebensstile und soziale Ungleichheit. Gesellschaft, Wirtschaft, Politik (GWP) (1), 51–62.

Kawachi, I., Subramanian, S.V., Kim, Daniel (Eds.), 2008. Social Capital and Health. Springer, New York.

Miller, S.M., Riessman, F., 1964. The working-class subculture: a new view. In: Shostak, A.B., Gomberg, W. (Eds.), Blue-Collar. World, pp. 24–36. Prentice-Hall, Englewood Cliffs, NJ.

OED, 2014. Oxford English Dictionary. http://www.oed.com/view/Entry/108129?redirectedFrom=life+style#eid (19.08.14.).

Parsons, T., Bales, R.F., 1955. Family, Socialization and Interaction Process. Free Press, Glencoe.

Putnam, R., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon and Schuster, New York.

Trigg, Andrew B., 2001. Veblen, Bourdieu, and conspicuous consumption. Journal of Economic Issues 35 (1), 99–115.

Veblen, Thorstein, 1899. The Theory of the Leisure Class. New York.

Veil, A.J., 2000. Leisure and lifestyle. A review and annotated bibliography. Leisure and Lifestyle. www.business.uts.edu.au/lst/research/research_papers.html.

Weber, Max, 1968. In: Roth, Guenther, Wittich, Claus (Eds.), Economy and Society. University of California Press, Berkley and Los Angeles.

Weber, Max, 1972. Wirtschaft und Gesellschaft (Economy and Society). Mohr, Tübingen.

Wilensky, H.L., 1964. Mass society and mass culture: interdependence or independence? American Sociological Review 29, 173–197.

Likelihood in Statistics

Anthony WF Edwards, Cambridge University, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 13, pp. 8854–8858, © 2001, Elsevier Ltd., with an updated Bibliography supplied by the Editor in Chief.

Abstract

A statistical model for phenomena in the social sciences is a mathematical construct that associates a 'probability' with each of the possible outcomes. If the data are discrete, such as the numbers of people falling into various classes, the model will be a discrete probability distribution, but if the data consist of measurements or other numbers that may take any values in a continuum, the model will be a continuous probability distribution. When two different models, or perhaps two variants of the same model differing only in the value of some adjustable parameter(s), are to be compared as explanations for the same observed outcome, the probability of obtaining this particular outcome can be calculated for each and is then known as the 'likelihood' for the model or parameter value(s) given the data. The concept of likelihood, which was introduced by R.A. Fisher, plays an important role in all approaches to statistical inference, in particular allowing many of the advantages perceived for Bayesian methods without embracing their controversial prior probabilities.

A statistical model for phenomena in the sciences or social sciences is a mathematical construct, which associates a *probability* with each of the possible outcomes. If the data are discrete, such as the numbers of people falling into various classes, the model will be a discrete probability distribution, but if the data consist of measurements or other numbers that may take any values in a continuum, the model will be a continuous probability distribution. When two different models, or perhaps two variants of the same model differing only in the value of some adjustable parameter(s), are to be compared as explanations for the same observed outcome, the probability of obtaining this particular outcome can be calculated for each and is then known as the *likelihood* for the model or parameter value(s) given the data. This article enlarges on the definition of likelihood, discusses its properties and uses, and relates it to other approaches to statistical inference.

Probability and Likelihood Distinguished

Probabilities and likelihood are easily (and frequently) confused, and it is for this reason that in 1921 R.A. Fisher introduced the new word: "What we can find from a sample is the likelihood of any particular value of [the parameter], if we define the likelihood as a quantity proportional to the probability that, from a population having that particular value, the [observed sample] should be obtained. So defined, probability and likelihood are quantities of an entirely different nature" (*see* Fisher, Ronald A. (1890–1962)).

The first difference to be noted is that the variable quantity in a likelihood statement is the hypothesis (a word that conveniently covers both the case of a model and of particular parameter values in a single model), the outcome being that actually observed, in contrast to a probability statement, which refers to a variety of outcomes, the hypothesis being assumed and fixed. Thus a manufacturer of dice may reasonably assert that the outcomes 1, 2, 3, 4, 5, 6 of a throw each have 'probability' 1/6 on the hypothesis that his dice are well-balanced, while an inspector of casinos charged with testing a particular die will wish to compute the 'likelihood' for various hypotheses about these probabilities on the basis of data from actual tosses.

The second difference arises directly from the first. If all the outcomes of a statistical model are considered their total probability will be 1, since one of them must occur and they are mutually exclusive; but since in general hypotheses are not exhaustive – one can usually think of another one – it is not to be expected that the sum of likelihoods has any particular meaning, and indeed there is no addition law for likelihoods corresponding to the addition law for probabilities. It follows that only 'relative' likelihoods are informative, which is the reason for Fisher's use of the word 'proportional' in his original definition.

The Likelihood Function

The most important application of likelihood is in parametric statistical models. Consider the simplest binomial example, such as that of the distribution of the number of boys r in families of size n (an example which has played an important role in the development of statistical theory since the early eighteenth century). The probability of getting exactly r boys will be given by the binomial distribution indexed by a parameter p, the probability of a male birth. Denote this probability of r boys by $\mathrm{P}(r|p)$, n being assumed fixed and of no statistical interest. Then we write

$$\mathrm{L}(p||r) \propto \mathrm{P}(r|p)$$

for the likelihood of p given the particular value r, the double vertical line || being used to indicate that the likelihood of p is not conditional on r in the technical probability sense. In this binomial example, $\mathrm{L}(p||r)$ is a continuous function of the parameter p and is known as the 'likelihood function.' When only two hypotheses are compared, such as two particular values of p in the present example, the ratio of their likelihoods is known as the 'likelihood ratio.'

The value of p that maximizes $\mathrm{L}(p||r)$ for an observed r is known as the 'maximum-likelihood estimate' of p and is

denoted by p; expressed in general form as a function of r, it is known as the 'maximum-likelihood estimator.' Since the pioneering work of Fisher in the 1920s, it has been known that maximum-likelihood estimators possess certain desirable properties under repeated sampling (consistency and asymptotic efficiency, and in an important class of model sufficiency and full efficiency). For this reason, they have come to occupy a central position in repeated-sampling (or 'frequentist') theories of statistical inference (*see* Estimation: Point and Interval). However, partly as a reaction to the unsatisfactory features, which repeated-sampling theories display when used as theories of evidence, coupled with a reluctance to embrace the full-blown Bayesian theory of statistical inference, likelihood is increasingly seen as a fundamental concept enabling hypotheses and parameter values to be compared directly.

The Likelihood Principle

The basic notion, championed by Fisher as early as 1912 while still an undergraduate at Cambridge but now known to have been occasionally suggested by other writers even earlier, is that the likelihood ratio for two hypotheses or parameter values is to be interpreted as the degree to which the data support the one hypothesis against the other. Thus a likelihood ratio of 1 corresponds to indifference between the hypotheses on the basis of the evidence in the data, while the maximum-likelihood value of a parameter is regarded as the 'best-supported value,' other values being ranked by their lesser likelihood accordingly. This was formalized as the *Law of Likelihood* by Ian Hacking (1965). Fisher's final advocacy of the direct use of likelihood will be found in his last book *Statistical Methods and Scientific Inference* (1956).

Such an approach, unsupported by any appeal to repeated-sampling criteria, is ultimately dependent on the primitive notion that the best hypothesis or parameter value on the evidence of the data is the one which would explain what has in fact been observed with the highest probability. The strong intuitive appeal of this can be captured by recognizing that it is the value which would lead, on repeated sampling, to a precise repeat of the data with the least expected delay. In this sense it offers the best statistical explanation of the data.

In addition to specifying that relative likelihoods measure degrees of support, the likelihood approach requires us to accept that the likelihood function or ratio contains all the information we can extract from the data about the hypotheses in question on the assumption of the specified statistical model – the so-called 'Likelihood Principle.' It is important to include the qualification requiring the specification of the model, first because the adoption of a different model might prove necessary later, and second because in some cases the structure of the model enables inferences to be made in terms of 'fiducial probability' which, though dependent on the likelihood, are stronger, possessing repeated sampling properties, which enable confidence intervals to be constructed (*see* Fiducial and Structural Statistical Inference).

Though it would be odd to accept the 'Law of Likelihood' and not the 'Likelihood Principle,' Bayesians necessarily accept the Principle but not the Law, for although the likelihood is an intrinsic component of Bayes' Theorem, Bayesians deny that a likelihood function or ratio has any meaning in isolation. For those who accept both the Law and the Principle, it is convenient to express the two together as:

> *The Likelihood Axiom:* Within the framework of a statistical model, all the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data, and the likelihood ratio is to be interpreted as the degree to which the data support the one hypothesis against the other.
>
> (Edwards, 1972)

The likelihood approach has many advantages apart from its intuitive appeal. It is straightforward to apply because, the likelihood function is usually simple to obtain analytically or easy to compute and display. It leads directly to the important theoretical concept of 'sufficiency' according to which the function of the data that is the argument of the likelihood function itself carries the information. This reduction of the data is often a simple 'statistic,' such as the sample mean (*see* Statistical Sufficiency). Moreover, the approach illuminates many of the controversies surrounding repeated-sampling theories of inference, especially those concerned with ancillarity and conditioning. Birnbaum (1962) argued that it was possible to derive the 'Likelihood Principle' from the concepts of sufficiency and conditionality, but to most people the Principle itself seems the more primitive concept and the fact that it leads to notions of sufficiency and conditioning seems an added reason for accepting it.

Likelihoods are multiplicative over independent data sets referring to the same hypotheses or parameters, facilitating the combination of information. For this reason 'log-likelihood' is often preferred because information is then combined by addition. In the field of genetics, where likelihood theory is widely applied, the log-likelihood with the logarithms taken to the base 10 is known as a LOD, but for general use natural logarithms to the base e are to be preferred, in which case log-likelihood is sometimes called 'support.' Most importantly, the likelihood approach is compatible with Bayesian statistical inference in the sense that the posterior Bayes distribution for a parameter is, by Bayes' Theorem, found by multiplying the prior distribution by the likelihood function. Thus when, in accordance with Bayesian principles, a parameter can itself be given a probability distribution (and this assumption is the Achilles' heel of Bayesian inference) all the information the data contain about the parameter is transmitted via the likelihood function in accordance with the 'Likelihood Principle.' It is indeed difficult to see why the medium through which such information is conveyed should depend on the purely external question of whether the parameter may be considered to have a probability distribution, and this is another powerful argument in favor of the Principle itself.

Parameter Estimation

In the case of a single parameter the likelihood function or the log-likelihood function may easily be drawn, and, if it is

unimodal, limits may be assigned to the parameter, analogous to the confidence limits of repeated-sampling theory. Calling the log-likelihood the 'support,' m-unit support limits are the two-parameter values astride the maximum at which the support is m units less than at the maximum. For the simplest case of estimating the mean of a normal distribution of known variance the 2-unit support limits correspond closely to the 95% confidence limits which are at ± 1.96 standard errors. In this normal case the support function is quadratic and may therefore be characterized completely by the maximum-likelihood estimate and the curvature at the maximum (the reciprocal of the radius of curvature) which is defined as the 'observed information.' Comparable definitions apply in multiparameter cases, leading to the concept of an m-unit support region and an observed information matrix. In cases in which the support function is not even approximately quadratic the above approach may still be applied if a suitable transformation of the parameter space can be found.

The representation of a support function for more than two parameters naturally encounters the usual difficulties associated with the visualization of high-dimensioned spaces, and a variety of methods have been suggested to circumvent the problem. It will often be the case that information is sought about some subset of the parameters, the others being considered to be nuisance parameters of no particular interest. In fortunate cases it may be possible to restructure the model so that the nuisance parameters are eliminated, and in all cases in which the support function is quadratic (or approximately so) the dimensions corresponding to the nuisance parameters can simply be ignored. Several other approaches are in use to eliminate nuisance parameters. Marginal likelihoods rely on finding some function of the data that does not depend on them; notable examples involve the normal distribution, where a marginal likelihood for the variance can be found from the distribution of the sample variance, which is independent of the mean, and a marginal likelihood for the mean can similarly be found using the t-distribution. Profile likelihoods, also called maximum relative likelihood, are found by replacing the nuisance parameters by their maximum-likelihood estimates at each value of the parameters of interest. It is easy to visualize from the case of two parameters why this is called a profile likelihood.

Naturally, a solution can always be found by strengthening the model through adopting particular values for the nuisance parameters, just as a Bayesian solution using integrated likelihoods can always be found by adopting a prior distribution for them and integrating them out, but such assumptions do not command wide assent. When a marginal likelihood solution has been found it may correspond to a Bayesian integrated likelihood for some choice of prior, and such priors are called neutral priors, to distinguish them from so-called uninformative priors for which no comparable justification exists. However, in the last analysis there is no logical reason why nuisance parameters should be other than a nuisance, and procedures for mitigating the nuisance must be regarded as expedients.

Relation to Significance Testing

All the common repeated-sampling tests of significance have their analogues in likelihood theory, and in the case of the normal model it may seem that only the terminology has changed. At first sight an exception seems to be the χ^2 goodness of fit test, where no alternative hypothesis is implied. However, this is deceptive, and a careful analysis shows that there is an implied alternative hypothesis that allows the variances of the underlying normal model to depart from their multinomial values. In this way the paradox of small values of χ^2 being interpreted as meaning that the model is 'too good' is exposed, for in reality they mean that the model is not good enough and that one with a more appropriate variance structure will have a higher likelihood. Likelihood ratio tests are based on the distribution under repeated-sampling of the likelihood ratio and are therefore not part of likelihood theory.

When likelihood arguments are applied to models with continuous sample spaces it may be necessary to take into account the approximation involved in representing data, which are necessarily discrete, by a continuous model. Neglect of this can lead to the existence of singularities in the likelihood function or other artifacts that a more careful analysis will obviate.

It is often argued that in comparing two models by means of a likelihood ratio, allowance should be made for any difference in the number of parameters by establishing a 'rate of exchange' between an additional parameter and the increase in log-likelihood expected. The attractive phrase 'Occam's bonus' has been suggested for such an allowance (Edwards, 1969). However, the proposal seems only to have a place in a repeated-sampling view of statistical inference, where a bonus such as that suggested by Akaike's information criterion is sometimes canvassed.

The major application of likelihood theory so far has been in human genetics, where log-likelihood functions are regularly drawn for recombination fractions (linkage values) (see Ott, 1991), but even there a reluctance to abandon significance testing altogether has led to a mixed approach. Other examples, especially from medical fields, will be found in the books cited below.

History and Further Reading

Although historically the development of a likelihood approach to statistical inference was almost entirely due to R.A. Fisher, it is interesting to recall that the Neyman–Pearson approach to hypothesis testing derives ultimately from a remark of 'Student's' (W. S. Gosset) in a letter to E. S. Pearson in 1926 that "if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability … you will be very much more inclined to consider that the original hypothesis is not true," a direct likelihood statement (quoted in McMullen and Pearson, 1939). Indeed, it has been remarked that "Just as support [log-likelihood] is Bayesian inference without the priors, so it turns out to be Neyman–Pearson inference without the 'errors'" (Edwards, 1972).

The literature on likelihood is gradually growing as an increasing number of statisticians become concerned at the inappropriate use of significance levels, confidence intervals and other repeated-sampling criteria to represent evidence. The movement is most advanced in biostatistics as may be seen from books such as Clayton and Hills (1993) and Royall (1997), but general texts such as Lindsey (1995) and Sprott

(2000) exist as well. Among older books Cox and Hinkley (1974) contains much that is relevant to likelihood, while Edwards (1972, 1992) was the first book to advocate a purely likelihood approach, and is rich in relevant quotations from Fisher's writings. The history of likelihood is treated by Edwards (1974; reprinted in Edwards, 1992).

See also: Probabilistic Thinking; Probability and Chance: Philosophical Aspects.

Bibliography

Berger, J.O., Wolpert, R.L., 1988. The Likelihood Principle. Institute of Mathematical Statistics, Hayward, CA.

Birnbaum, A., 1962. On the foundations of statistical inference. Journal of the American Statistical Association 57, 269–326.

Clayton, D.G., Hills, M., 1993. Statistical Models in Epidemiology. Oxford University Press, Oxford, UK.

Cox, D.R., Hinkley, D.V., 1974. Theoretical Statistics. Chapman and Hall, London.

Edwards, J.H., 1969. In: Morton, N.E. (Ed.), Computer Applications in Genetics. University of Hawaii Press, Honolulu, HI.

Edwards, A.W.F., 1972. Likelihood. Cambridge University Press, Cambridge, UK.

Edwards, A.W.F., 1974. The history of likelihood. International Statistical Review 42, 9–15.

Edwards, A.W.F., 1992. Likelihood. Johns Hopkins University Press, Baltimore, MD.

Fisher, R.A., 1912. On an absolute criterion for fitting frequency curves. Messenger of Mathematics 41, 155–160.

Fisher, R.A., 1921. On the 'probable error' of a coefficient of correlation deduced from a small sample. Metron 1 (4), 3–32.

Fisher, R.A., 1956. Statistical Methods and Scientific Inference. Oliver and Boyd, Edinburgh, UK.

Hacking, I., 1965. Logic of Statistical Inference. Cambridge University Press, Cambridge, UK.

Held, Leonhard, Bové, Daniel Sabanés, 2014. Frequentist Properties of the Likelihood. Springer, pp. 79–122 in their book Applied Statistical Inference.

Lee, Peter M., 2012. Bayesian Statistics: An Introduction. Wiley.

Lindsey, J.K., 1995. Introductory Statistics: A Modelling Approach. Clarendon Press, Oxford, UK.

McMullen, L., Pearson, E.S., 1939. William Sealy Gosset, 1876–1937. Biometrika 30, 205–250.

Ott, J., 1991. Analysis of Human Genetic Linkage. Johns Hopkins University Press, Baltimore, MD.

Royall, R., 1997. Statistical Evidence: A Likelihood Paradigm. Chapman and Hall, London.

Sprott, D.A., 2000. Statistical Inference in Science. Springer, New York.

Varin, Cristiano, Reid, Nancy Margaret, Firth, D. (David), 2011. An overview of composite likelihood methods. Statistica Sinica 21 (1), 5–42.

Likelihoods that Eliminate Nuisance Parameters

Brunero Liseo, MEMOTEF, Sapienza Università di Roma, Viale Castro Laurenziano, Roma, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Elimination of nuisance parameters is a central problem in statistical inference, and has been formally studied in virtually all approaches to inference. Given a statistical model, a researcher tries to make inferences about an unknown state of nature ψ (assumed to lie within some set Ψ of possible states of nature) upon observing the value $Y = y$ of some random vector $Y = \{Y_1, ..., Y_n\}$ whose probability distribution is completely determined by ψ. If Y has a density function $f(y|\psi)$, strong arguments, generally known as "The Likelihood Principle" suggest that inference about ψ ought to depend upon Y only through the *likelihood function* $L(\psi) = f(y|\psi)$, which is to be viewed as a function of ψ for the given y. Rarely is the entire parameter ψ of interest to the analyst. It is common to select a parameterization $\psi = (\theta, \lambda)$ of the statistical model in a way that simplifies the study of the "parameter of interest," here denoted θ, while collecting any remaining parameter specification into a "nuisance parameter" λ. In this article, we review some of the ways that have been used or proposed for eliminating the nuisance parameter λ from the analysis to achieve some sort of "likelihood" $L^*(\theta)$ for the parameter of interest.

Introduction

Consider a given statistical model, that is a triad $(\mathcal{Y}, \mathcal{F}, \Psi)$, where $\mathcal{Y}$ is the sampling space, Ψ is the parameter space, and we assume $\Psi \subset \mathbb{R}^d$, for some integer d, and

$$\mathcal{F} = \{f(y;\psi), \psi \in \Psi, y \in \mathcal{Y}\}$$

is a family of probability distributions defined on $\mathcal{Y}$. In general, one is interested in a low dimensional function θ of the parameter vector ψ, say $\theta = \theta(\psi) \in \mathbb{R}^k$, with $k < d$. Modern parametric statistical theories, at least the approaches based on likelihood and Bayesian theories, aim at constructing a likelihood function which depends on θ only. If such a likelihood exists it will represent the basis upon which one can construct inferential procedures for the parameter of interest θ. In this article, we will review the most popular approaches to obtain a likelihood function which is free of the nuisance parameters λ. It is well known that a likelihood function is defined, as a function of the parameters, up to a proportionality constant. In the following, we will deliberately ignore such constant, and the likelihood function expression will be always preceded by the symbol $\propto$.

Historically, the seminal papers addressing the problem are Kalbfleisch and Sprott (1970) and Basu (1977). In the former paper, two important examples of "partial" likelihoods were introduced, namely the *marginal* and *conditional* likelihoods.

Marginal Likelihood

If, in our model, there exists a statistics $T = t(y)$ whose distribution only depends on θ, we can use this distribution as a marginal sampling model. In practice, if it is possible to write the density of y as

$$f(y;\theta,\lambda) = f_1(t;\theta) f_2(y|t;\theta,\lambda)$$

we may ignore the fact that f_2 depends on θ and use f_1 as a marginal likelihood for θ.

Example 1

Let $(Y_1, \ldots, Y_n)$ be n Gaussian random variable with mean λ and variance θ. The complete likelihood function is

$$L(\theta,\lambda) \propto \frac{1}{\theta^{n/2}} \exp\left(-\frac{1}{2\theta}\sum_{i=1}^{n}(y_i-\lambda)^2\right)$$
$$\propto \frac{1}{\theta^{n/2}} \exp\left(-\frac{n}{2\theta}\left[S^2 + (\bar{y}-\lambda)^2\right]\right)$$

Set $Z = nS^2/\theta = \sum_{i=1}^{n}(y_i - \bar{y})^2/\theta$. It is known that Z has a Chi-squared distribution with $n-1$ degrees of freedom, that is

$$f_Z(z) = \frac{z^{\frac{n-1}{2}-1} e^{-\frac{z}{2}}}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})}$$

Using a simple linear transformation argument, one can see that the distribution of $\theta Z = nS^2$ is

$$f_{nS^2}(nS^2;\theta) = \frac{1}{\theta}\left(\frac{nS^2}{\theta}\right)^{\frac{n-1}{2}-1} \exp\left(-\frac{nS^2}{2\theta}\right) \frac{1}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})}$$

So, one can use this density as a marginal likelihood for θ: after eliminating the unnecessary constants, one obtains

$$L_M(\theta;\mathbf{y}) \propto \frac{1}{\theta^{(n-1)/2}} \exp\left(-\frac{n}{2\theta}S^2\right)$$

In general, it is easy to find a marginal likelihood for statistical models with a group structure (Pace and Salvan, 1997).

Conditional Likelihood

In a related way, it is possible to define a conditional likelihood based on the distribution of some statistics whose conditional distribution is free of λ. If the model has a sufficient statistics (T,S), the likelihood function can be written as

$$L(\theta,\lambda;\mathbf{y}) = f_1(t|s,\theta,\lambda)\ f_2(s;\theta,\lambda)$$

If it happens that the term f_1 does not depend on λ, the distribution of $T|S$ may be used as a conditional likelihood

for θ. The efficiency of this approach depends on the amount of information which is lost by ignoring the term f_2 in the above expression.

Example 2

A capture-recapture experiment. A software code has N faults, with N unknown. Two reviewers independently check the code. Each reviewer detects each fault – independently of each other – with probability p. Let N_1 and N_2 be the number of faults detected by the two reviewers, and let M be the number of faults detected by both of them. Here the parameter of interest is $\theta = N$ and the nuisance parameter is $\lambda = p$. It is easy to see that N_1 and N_2 are independent binomial random variables with parameters (N, p) and the conditional distribution of $(M|N_1 = n_1, N_2 = n_2, N, p)$ is hypergeometric, that is

$$\Pr(M = m|N_1 = n_1, N_2 = n_2) = \binom{n_1}{m}\binom{N-n_1}{n_2-m}\Big/\binom{N}{n_2} \qquad [1]$$

Since the last expression does not depend on p, it can be used as a conditional likelihood for N. Conditional likelihoods are easy to determine if the sampling model belongs to the exponential family (Pace and Salvan, 1997).

Profile Likelihood and Its Ramifications

Apart from the simple group structure or the exponential family, it is rare to be able to derive a conditional or a marginal likelihood. A more general and popular approach is to substitute the unknown nuisance parameter with its conditional (on θ) maximum likelihood estimate. Thus we obtain the so-called *profile* likelihood

$$\widehat{L}(\theta) = L\left(\theta, \widehat{\lambda}_\theta\right) \qquad [2]$$

with $\widehat{\lambda}_\theta = \arg\max_\lambda L(\theta, \lambda)$. The main advantages of using a profile likelihood are listed below:

- The maximum profile likelihood estimate $\widehat{\theta}$ will coincide with the one obtained from the joint likelihood $L(\theta, \lambda)$.
- $\widehat{L}(\theta)$ is easy to compute, and the inverse of its observed Fisher information is equal to the lower right corner of the inverse of the observed Fisher information computed with the true likelihood.
- It is invariant with respect to reparameterizations of the nuisance parameter λ.
- Test and confidence regions based on the profile likelihood can be approximately computed.

The main drawbacks are as follows:

- The profile score vector has not zero mean.
- Replacing an unknown parameter with a point estimate may dramatically overestimate the precision of the inferential procedures. This danger is particularly serious when the nuisance parameter has large dimension.

Example 1 (continued)

The conditional maximum likelihood estimate for λ is of course the sample mean $\widehat{\lambda}_\theta = \bar{y}$, which does not depend on θ. The profile likelihood is then

$$\widehat{L}(\theta) \propto \frac{1}{\theta^{n/2}} \exp\left(-\frac{n}{2\theta}S^2\right) = \theta^{-1/2} L_M(\theta)$$

Example 2 (continued)

In the capture-recapture example, the joint likelihood is

$$L(N,p) \propto \binom{N}{n_1}\binom{N}{n_2} p^{n_1+n_2}(1-p)^{2N-n_1-n_2} \frac{\binom{n_1}{m}\binom{N-n_1}{n_2-m}}{\binom{N}{n_2}}$$
$$\propto \frac{N!}{(N-r)!} p^k (1-p)^{2N-k}$$

where $r = n_1 + n_2 - m$ is the total number of detected faults and $k = n_1 + n_2$. Then $\widehat{p}_N = \frac{k}{2N}$, and the profile likelihood is

$$\widehat{L}(N) \propto \frac{N!}{(N-r)!}\frac{(2N-k)^{2N-k}}{(2N)^{2N}} = \frac{N!}{(N-r)!}\frac{1}{(2N)^k}\left(1-\frac{k}{2N}\right)^{2N-k} \qquad [3]$$

The profile approach of replacing λ by its conditional maximum likelihood estimator, $\widehat{\lambda}_\theta$, would appear to be dangerous in that it ignores the uncertainty in λ. Example 1 demonstrates a very mild version of this problem; the profile likelihood has one more 'degree of freedom' than it should, given the replacement of λ by $\bar{y}$. Indeed, many of the modifications to profile likelihood that have been advanced have, as one of their primary goals, adjustments to account for nuisance parameter uncertainty. It is important to emphasize that such modifications can be crucial, especially because 'raw' likelihood procedures will typically be anti-conservative, erring on the side of suggesting more accuracy than is actually warranted.

Modifications of the Profile Likelihood

Many proposals have been done to overcome the limitations of the profile likelihoods. Most of them results in a multiplicative correction factor to $\widehat{L}$ although their derivations can be very different in spirit. Here we briefly discuss the most important and general proposals. See, for example, Fraser and Reid (2003) for more details.

Conditional Approximated Likelihood

Suppose that the nuisance parameter λ can be re-expressed into a new parameter $\xi = g(\lambda, \theta)$ such that θ and ξ are orthogonal, that is the Fisher expected information in the (θ, ξ) parameterization is block-diagonal. This is always possible when θ is scalar. In this case one can use an approximation of the conditional distribution of the data, given the conditional maximum likelihood estimator of ξ, say $\widehat{\xi}_\theta$. So we can define the *conditional approximated likelihood* or *conditional profile likelihood* (CPL) as (Cox and Reid, 1987)

$$L_{\mathrm{CPL}}(\theta) \propto \widehat{L}(\theta)\left|j_{\xi\xi}\left(\theta, \widehat{\xi}_\theta\right)\right|^{-\frac{1}{2}} \qquad [4]$$

Formula [4] is then a modification of [2] with the correction factor given by the square root of the inverse of observed Fisher matrix for the nuisance parameter, evaluated at $\xi = \widehat{\xi}_\theta$. The correction factor is a penalty factor accounting for the uncertainty on the estimation of ξ. The necessity of finding an orthogonal parameterization explains that the conditional profile likelihood is not invariant with respect to reparameterizations of the nuisance parameter.

Example 1 (continued)
The nuisance parameter λ is already orthogonal to θ; an easy calculation shows that $j_{\xi\xi}(\theta, \widehat{\xi}_\theta) = n/\theta$ so

$$L_{\mathrm{CPL}}(\theta) \propto \theta^{\frac{1}{2}}\widehat{L}(\theta) = L_M(\theta)$$

Example 2 (continued)
Since N is an integer parameter, the Fisher information quantity is not readily available here, and we cannot compute a $L_{\mathrm{CPL}}(N)$ in this example. There are proposals to circumvent this limitation: Lindsay and Roeder (1987) propose to use the linear difference score for models with integer parameters.

Modified Profile Likelihood

A limitation of L_{CPL} is the need of finding an orthogonal parameterization. This might be impossible when θ is a vector. To overcome this difficulty, a different modification to the profile likelihood has been proposed: although it can be introduced in several different ways (see, for example, Barndorff-Nielsen (1983) and Pace and Salvan (1997)), it is easily interpreted as a 'multiplicative' correction of the conditional profile likelihood. Its expression is

$$L_{\mathrm{MPL}}(\theta) \propto \widehat{L}(\theta)M(\theta) \propto \widehat{L}(\theta)\left|j_{\lambda\lambda}\left(\theta, \widehat{\lambda}_\theta\right)\right|^{-\frac{1}{2}}\left|\frac{\partial\widehat{\lambda}_\theta}{\partial\widehat{\lambda}}\right|^{-1} \qquad [5]$$

The last term, the derivative of the conditional maximum likelihood estimator with respect to the unconditional one, accounts for nonorthogonality of the parameters. The correction factor $M(\theta)$ may be written in several different ways: see, Severini (2000) for details.

Example 1 (continued)
Here the parameters are orthogonal, so $\widehat{\lambda}_\theta = \widehat{\lambda}$ and the $L_{\mathrm{MPL}}(\theta)$ coincides with $L_{\mathrm{CPL}}(\theta)$.

Integrated Likelihood

Let $q(\lambda;\ \theta)$ be a nonnegative weight function defined on Λ, possibly depending on θ, such that $\int_\Lambda |q(\lambda;\theta)|\mathrm{d}\lambda < +\infty$. An integrated likelihood function for the parameter of interest θ is then

$$\tilde{L}_q(\theta) \propto \int_\Lambda f(y;\theta,\lambda)q(\lambda;\theta)\mathrm{d}\lambda = f_q(y;\theta)$$

It is intuitive to see that a Bayesian would select $q(\lambda;\ \theta)$ as his/her conditional prior distribution for λ given θ, say $\pi(\lambda|\theta)$. As long as $\pi(\lambda|\theta)$ is a proper distribution this would correspond to the use of the marginal model $f_q(y;\ \theta)$: in this sense $L_q(\theta)$ is a genuine likelihood function. If π is improper, as it is often the case in the default noninformative approaches, this is not guaranteed and care must be taken in the choice of π. In practice, possible choices for $\pi(\lambda|\theta)$ are as follows:

- A subjectively chosen prior: this is advisable in the presence of genuine prior information about λ. When $\pi(\lambda|\theta)$ is naturally a proper conditional density, certain of the concerns with interpreting and using the resulting integrated likelihood disappear. In particular, the resulting integrated likelihood can unambiguously be interpreted as a likelihood and can be combined with any proper prior density for θ to produce a true posterior. Today it is popular for computational reasons to use 'vague' proper priors (often 'vague' conjugate priors). Unfortunately, the use of such does not really provide protection against the concerns that arise with use of improper conditional priors; if a difficulty would arise in using an improper conditional prior, the same difficulty would manifest itself through great sensitivity to the degree of 'vagueness' chosen.
- The uniform choice of $\pi(\lambda|\theta) = 1$ has already been mentioned. It is still the most commonly used default conditional prior and is an attractive choice when nothing else is available. There are, however, well-documented difficulties with the uniform prior, perhaps the most well known being its lack of invariance to reparameterization. Using the uniform prior for a positive parameter λ will yield a different integrated likelihood than using the uniform prior for its logarithm $\lambda^* = \log\lambda$, for example.
- A more sophisticated objective conditional prior, namely the *conditional reference prior*. The derivation of this conditional prior is a slight modification of the more popular reference prior approach proposed by Berger and Bernardo (1992) and it is discussed in detail in Berger et al. (1999).

Example 1 (continued)

In this case the nuisance parameter λ is the location parameter. Virtually every noninformative Bayesian approach would suggest to use an improper flat prior, that is $\pi(\lambda|\theta) \propto 1$. The resulting integrated likelihood is

$$\begin{aligned}\tilde{L}(\theta) &= \int_{-\infty}^{\infty} \frac{1}{\theta^{n/2}}\exp\left(-\frac{n}{2\theta}\left[S^2 + (\bar{y}-\lambda)^2\right]\right)\mathrm{d}\lambda \\ &= \frac{1}{\theta^{n/2}}\exp\left(-\frac{n}{2\theta}S^2\right)\int_{-\infty}^{\infty}\exp\left(-\frac{n}{2\theta}(\bar{y}-\lambda)^2\right)\mathrm{d}\lambda\end{aligned}$$

The last integrand is proportional to a Gaussian density so the value of the integral is $\sqrt{2\pi\theta/n}$. Then

$$\tilde{L}(\theta) \propto \frac{1}{\theta^{(n-1)/2}}\exp\left(-\frac{n}{2\theta}S^2\right) = L_M(\theta)$$

One could also use his/her own prior information and produce a subjective integrated likelihood. For example, one can assume that, a priori

$$\lambda|\theta \sim N(0, \theta/k)$$

This assumption is equivalent to say that the information provided by the prior corresponds to an extra observation whose precision is k times greater than the precision of the actual observations; of course k must be subjective elicited. In this case, one has

$$\tilde{L}(\theta) \propto \frac{1}{\theta^{n/2}} \exp\left(-\frac{n}{2\theta}S^2\right) \frac{1}{\sqrt{\theta}} \int_{-\infty}^{\infty} \exp\left(-\frac{n}{2\theta}(\lambda - \bar{y})^2\right) \times \exp\left(-\frac{k}{2\theta}\lambda^2\right) \mathrm{d}\lambda$$
$$\propto \frac{1}{\theta^{n/2}} \exp\left(-\frac{n}{2\theta}S^2\right) \exp\left(-\frac{nk}{2(n+k)\theta}\bar{y}^2\right)$$

Notice that, as $k \to 0$, $\tilde{L}(\theta)$ will converge point-wise to the profile likelihood $\widehat{L}(\theta)$, because of the 'extra' observation, which adds a degree of freedom.

Example 2 (continued)

Here a reasonable choice for the conditional prior for the nuisance parameter p is a symmetric Beta(a, a) density, that is

$$\pi(p|N) \propto p^{a-1}(1-p)^{a-1}$$

Simple algebra shows that in this case the integrated likelihood is

$$\tilde{L}_a(N) \propto \frac{N!}{(N-r)!} \frac{\Gamma(2N-k+a)}{\Gamma(2N+2a)} \qquad [6]$$

Notice that, in this example, the conditional uniform prior and the conditional reference prior are special cases of the above Beta prior. The former corresponds to $a = 1$ and the latter to $a = 0.5$. Using a Stirling approximation, that is $\Gamma(t+1) \approx 2\pi e^{-t} t^{t+1/2}$, we can see that for large N

$$\tilde{L}_a(N) \approx \frac{N!}{(N-r)!} \frac{1}{(2N+2a-1)^{k+a}} \left(1 - \frac{k+a}{2N+2a-1}\right)^{2N-k+a-\frac{1}{2}}$$

which, for $a = 1/2$, becomes

$$\tilde{L}_a(N) \approx \frac{N!}{(N-r)!} \frac{1}{(2N)^{k+\frac{1}{2}}} \left(1 - \frac{k+\frac{1}{2}}{2N}\right)^{2N-k}$$

qualitatively similar to the profile likelihood eqn [3].

Generally speaking, the most popular objective prior in Bayesian analysis, namely the Jeffreys' prior is not well suited for producing integrated likelihoods. The Jeffreys' prior approach, in fact, will produce a prior $\pi^J(\theta, \lambda)$ which is proportional to the square root of the determinant of the expected Fisher information matrix and this prior is not able to distinguish between the parameter of interest and the nuisance parameters; then it is technically difficult to separate $\pi^J(\lambda|\theta)$ from $\pi^J(\theta)$, since the Jeffreys' prior is usually improper, that is, its integral over the parameter space is not a finite number. The resolution of this difficulty is at the core of the development of the conditional reference prior approach in Berger et al. (1999).

Some Particular Models

In the previous section, we have described the most popular approaches to the elimination of nuisance parameters and we have illustrated them through very simple examples. Modern statistics, however, deals with more complicated models and new problems arise in this context. Consider our Example 2: we have assumed that there were only two reviewers and the probability of detecting a fault was the same for both of them. We can imagine that there might be more than two reviewers with different abilities. With k reviewers, our model would have a k-dimensional nuisance parameter $\tilde{p} = (p_1, \ldots, p_k)$ and it would be interesting to note how the previous methods would change in this setup. In the profile likelihood approach one has to estimate k different parameters now, and the amount of "hidden" uncertainty under the procedure may become very serious. With the integrated likelihood approach one needs to elicit k different conditional priors, or a k-dimensional prior, over $\tilde{p}$. This maybe a problem for large k although from a different perspective, it may represent an advantage, in those cases where one knows that the reviewers have different skills indeed! It is interesting to note that the conditional likelihood, in this particular situation, will only show minor changes; for $k = 2$ and $p_1 \neq p_2$, it would be exactly equal to eqn [1].

We conclude this review with a discussion on methods of eliminating nuisance parameters in more complex models. A nonexhaustive list of problems where the elimination of nuisance parameters is still a vivid research problem must include:

- Statistical models with many nuisance parameters
- Semiparametric problems.

Models with many nuisance parameters refer to a general class of models where subsets of observations share common parameters. This setup is known under different names in different fields of application: latent structure in social and medical statistics, stratum parameters (Sartori, 2003) in biostatistics, incidental parameters in econometrics. Consider again Example 2 and assume now that the probability of detecting an error depends on the error itself; in that case we would have a model with N (unknown) incidental parameters and some structure among them should be assumed in order to make valid inferences. In general, problems where the number of nuisance parameters increases with the sample size are challenging because the usual asymptotic theory does not hold. We will only consider here the historical and famous problem of many Normal means (Neyman and Scott, 1948). Consider n different sets of k i.i.d. Normal random variables with common variance θ and different means λ_j, that is

$$Y_{j1}, Y_{j2}, \ldots, Y_{jk} \overset{iid}{\sim} N\left(\lambda_j, \theta\right), \quad j = 1, \ldots, n$$

Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$. The likelihood function is then

$$L(\theta, \boldsymbol{\lambda}) \propto \frac{1}{\theta^{nk/2}} \exp\left(-\frac{k-1}{2\theta} \sum_{j=1}^{n} S_j^2 + k(\bar{y} - \lambda_j)^2\right)$$

where S_j^2 is the usual unbiased estimate for the variance within a single sample. Since $\widehat{\lambda}_\theta = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n)$, the profile likelihood is

$$\widehat{L}(\theta) \propto \frac{1}{\theta^{nk/2}} \exp\left(-\frac{k-1}{2\theta} \sum_{j=1}^{n} S_j^2\right)$$

which is maximized at $\widehat{\theta} = (k-1)\sum S_j^2/(nk)$. Since the S_j^2's are i.i.d. r.v.'s with mean θ, the law of large numbers suggest

that $\widehat{\theta}$ is not even a consistent estimator since, as $n \to \infty$, $\widehat{\theta} \to (k-1)\theta/k$. The bad behavior of the $\widehat{L}(\theta)$ can be corrected either by using a modified profile argument or by integrating out $\boldsymbol{\lambda}$ with respect to a flat prior, which is incidentally the conditional reference prior for this problem, $\pi(\lambda) \propto 1$. Both of them introduce a correction factor of $\theta^{n/2}$ which leads to a consistent estimator of θ.

Loosely speaking, a semiparametric model is a model where the parameter space Ψ can be written as $\Theta \times \Lambda$, with $\Theta \subset \mathbb{R}^d$ for some integer d, while Λ is infinite dimensional. If θ is the parameter of interest, then inference on this model is a problem of elimination of the infinite dimensional nuisance parameters λ. This kind of models is getting very popular in applied statistics since today the increased availability of data makes it difficult to select a particular statistical model and, on the other hand, there is still a given quantity θ we are interested in.

Murphy and van der Vaart (2000) establish a first-order asymptotic likelihood theory for general classes of semiparametric models. Their approach is based on the construction of a least-favorable parametric family and then the usual asymptotic results are considered with respect to this family. In their framework the profile likelihood inherits many of its asymptotic properties for the parametric case. A detailed discussion of these issues is beyond the scope of this article. We limit ourselves to the discussion of a relatively simple, although meaningful, example of the construction of a semiparametric likelihood for the median of an unknown distribution (Liseo, 2005).

Example 3

Given $\theta \in \mathbb{R}$ and for a given continuous probability density $\lambda(\cdot)$, defined on $\mathbb{R}$ and having zero median, suppose we observe n i.i.d. random variables $Y_i = \theta + \varepsilon_i$, $i = 1, \ldots, n$, with $\varepsilon_i \sim \lambda(\cdot)$. We are interested in θ and λ is the infinite dimensional nuisance parameter. Consider the set $\mathcal{F}$ of all continuous, zero-median, probability density functions. This problem is very simple to state and it has been discussed in several papers: see Liseo (2005) for references. The infinite dimensionality of the nuisance parameter is a challenge for the techniques we have illustrated above. Non-Bayesian inference is usually performed via the construction of some ad-hoc pseudo-likelihood for θ. The original idea dates back to Jeffreys (1961), p. 211; let $K(\theta)$ be the number of observations in the sample which are not greater than θ. Then, for fixed θ, $K(\theta) \sim \mathrm{Bin}(n, 0.5)$ and a possible pseudo-likelihood for θ is

$$L_{\mathrm{PS}}(\theta) \propto \binom{n}{k(\theta)}$$

Notice that this likelihood is based on the distribution of K, which is not a statistic in the strict sense, since it depends on θ, so L_{PS} is not a genuine likelihood.

Alternatively one can maximize $L(\theta, \lambda)$ over the set $\mathcal{F}$, for fixed θ, thus obtaining the profile likelihood

$$\widehat{L}\left(\theta\right) \propto k(\theta)^{-k(\theta)}[n - k(\theta)]^{-[n-k(\theta)]}$$

Both the above likelihood functions have the drawback that they do not vanish for large values of $|\theta|$ and that makes hard their use in testing and confidence regions. A Bayesian analysis of the problem can be done by assuming a Polya tree prior for $\lambda(\cdot)$. Ghosal et al. (1999) show that, under some constraints on the choice of the parameters of the Polya tree, the posterior of θ will be consistent for a large class of true symmetric densities. However, this approach is not objective, since the final answer will depend on the hyperparameters of the Polya tree prior.

An alternative and general approach for computing an approximated integrated likelihood function is based on a novel use of ABC, namely Approximate Bayesian Computation. This approach is illustrated in Grazian and Liseo (in press) with several examples.

See also: Bayesian Statistics.

Bibliography

Barndorff-Nielsen, O., 1983. On a formula for the distribution of the maximum likelihood estimator. Biometrika 70 (2), 343–365.

Basu, D., 1977. On the elimination of nuisance parameters. Journal of the American Statistical Association 72 (358), 355–366.

Berger, J.O., Bernardo, J.M., 1992. Ordered group reference priors with applications to a multinomial problem. Biometrika 79, 25–37.

Berger, J.O., Liseo, B., Wolpert, R.L., 1999. Integrating likelihood methods for eliminating nuisance parameters. Statistical Science 14, 1–28.

Bjørnstad, J., 1996. On the generalization of the likelihood function and the likelihood principle. Journal of the American Statistical Association 91, 791–806.

Berger, J.O., Wolpert, R.L., 1988. Lecture Notes-Monograph Series. The Likelihood Principle, vol. 6. IMS, Hayward, CA, USA.

Cox, D.R., Reid, N., 1987. Parameter orthogonality and approximate conditional inference. Journal of the Royal Statistical Society, B 49 (1), 1–39.

Fraser, D.A.S., Reid, N. Likelihood Inference in the Presence of Nuisance Parameters. Phystat2003, SLAC. Available at: http://arxiv.org/pdf/physics/0312079.pdf.

Ghosal, S., Ghosh, J., Ramamoorthi, R., 1999. Consistent semiparametric bayesian inference about a location parameter. Journal of Statistical Planning and Inference 77, 181–193.

Grazian, C., Liseo, B. Approximate integrated likelihood via ABC methods. Statistics and Its Interface, in press, http://arxiv.org/pdf/1403.0387v1.pdf.

Jeffreys, H., 1961. Theory of Probability, second ed. Clarendon Press, Oxford.

Kalbfleisch, J.D., Sprott, D.A., 1970. Application of likelihood methods to models involving large numbers of parameters. Journal of the Royal Statistical Society, B 32, 175–208.

Lindsay, B.G., Roeder, K., 1987. A unified treatment of integer parameter models. Journal of the American Statistical Association 82 (399), 758–764.

Liseo, B., 2005. The elimination of nuisance parameters. In: Bayesian Thinking: Modelling and Computation. Handbook of Statistics, vol. 25. Elsevier/North-Holland, Amsterdam, pp. 193–219.

Murphy, S., Van der Vaart, A., 2000. On profile likelihood. Journal of the American Statistical Association 95 (450), 449–465.

Neyman, J., Scott, E.L., 1948. Consistent estimates based on partially consistent observations. Econometrica 16, 1–32.

Pace, L., Salvan, A., 1997. Principles of Statistical Inference. From a Neo-Fisherian Perspective. World Scientific Publishing Co. Inc, River Edge, NJ.

Sartori, N., 2003. Modified profile likelihoods in models with stratum nuisance parameters. Biometrika 90 (3), 533–549.

Severini, T.A., 2007. Integrated likelihood functions for non-Bayesian inference. Biometrika 94 (3), 529–542.

Severini, T.A., 2000. Likelihood Methods in Statistics. Oxford University Press, Oxford.

Limbic System

Allison M Auchter and Marie H Monfils, University of Texas at Austin, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.L. Isaacson, volume 13, pp. 8858–8862, © 2001, Elsevier Ltd.

Abstract

While the definition of the limbic system has evolved since Paul Broca first described the distinct set of cortical structures on the medial surface of the cerebrum, the major structures have remained fairly consistent. Here we describe the major anatomical structures most commonly included in the limbic system and explore their major behavioral and cognitive functions. In addition, we propose and alternative way of conceptualizing the limbic system as more than just the primary system for emotional processing.

Introduction

In order to characterize the functionality of the highly complex organ serving as the commander in chief of our nervous system, scientists often divide the brain into separate systems – clearly defined anatomically distinct regions and pathways that function together to interpret inputs. Commonly recognized sensory systems include those of vision, hearing, somatic sensation, taste, and olfaction. Beginning around 1930, scientists began to argue that there also exists a system for processing the experience of emotion. This argument was supported largely by the integration of Paul Broca's work in the 1870s and James Papez' in the 1930s. Broca (1878) noted that all mammals possess a group of cortical areas on the medial surface of the cerebrum that are distinctly different from the surrounding cortex. He named this collection of cortical areas the limbic lobe, after the Latin word limbus, which means 'border' or 'fringe.' Later, Papez (1937) proposed the idea of an 'emotional system' that involves several interconnected structures on the medial surface of the brain, many of which overlap with Broca's limbic lobe. Paul MacLean (1952) popularized the term 'limbic system' to describe the group of limbic and cortical structures that govern the experience and expression of emotion.

Experimental work supports the involvement of the limbic system in emotion processing; however, the limbic structures also play a role in a number of other functions, including learning and memory. As scientists have continued to develop new anatomical, physiological, and behavioral approaches to investigate the brain in both clinical and research settings, many cortical regions, nuclear groups, and tracts that are structurally and/or functionally interconnected with limbic regions have come to be considered part of this network, broadening the concept of the limbic system even further. For this reason, there is much disagreement in the scientific community about which structures should be included in the limbic system, and consequentially, what its overarching role might be in mental functions and behavior. In this article, we will (1) describe the anatomical brain regions *most commonly* associated with the limbic system, (2) explore important behavioral and cognitive functions of these structures, and (3) propose an alternative way of conceptualizing the limbic system.

Anatomy of the Limbic System

It is important to have an anatomical understanding of the major structures and tracts of the limbic system before we delve into its functions. In this section, we will provide a basic anatomical overview, however, the connections and interactions between these areas and other brain regions are much more extensive than the scope of this article. We will discuss the major nuclear groups, limbic cortex, and fiber tracts that form the structural foundation of the limbic system. A schematic representation of these structures and connections is shown in Figure 1.

Nuclear Groups and Major Structures

Amygdala

Perhaps the most featured structure in recent research on the limbic system is the amygdala. This almond-shaped formation on the medial temporal lobe sits immediately anterior to the hippocampi and is often divided into three nuclei: the basolateral nucleus, the corticomedial nucleus, and the central nucleus. Due to its distinct functions, the basolateral nucleus can be further subdivided into the basal and lateral amygdala. One unique feature of the amygdala is that it receives afferents from neocortex in each lobe of the brain and each sensory system, as well as the hippocampal and cingulate gyri (Young et al., 1994). Additionally, the interconnections between the amygdala nuclei allow for the integration of information from each sensory system. Its extensive connectivity makes the amygdala ideally suited for valence-laden modulation of memory and emotional processing.

Hippocampal Formation and Associated Structures

The hippocampal formation is a folded structure in the medial temporal lobe consisting of the hippocampus and subiculum. Additionally, there are three cortical regions ventral to the hippocampus that are highly interconnected with the hippocampal formation: the entorhinal cortex, the perirhinal cortex, and the parahippocampal gyrus. Like the amygdala, these regions receive inputs from all sensory modalities, allowing for the integration of complex sensory information. The entorhinal cortex is the main source of input to the hippocampal formation, sending information to the hippocampus through

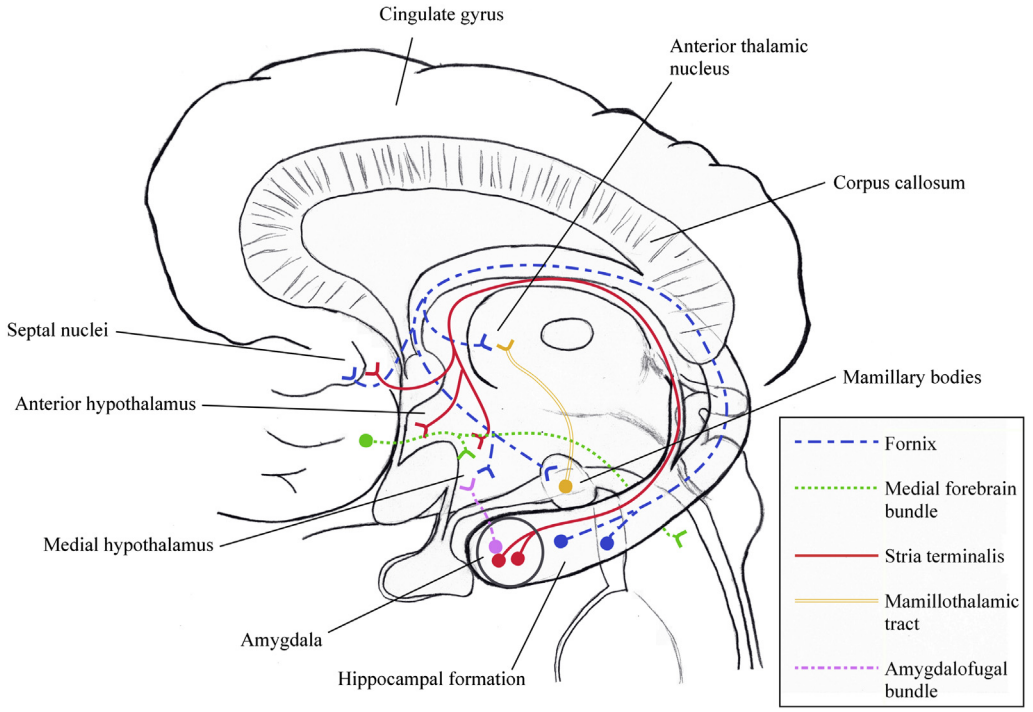

Figure 1 Major structures and connections of the limbic system.

a bundle of axons called the perforant path. From the hippocampus, information is carried in a C-shaped bundle of fibers called the fornix to the hypothalamus and mamillary bodies. Research implicates the hippocampus in consolidation of declarative and spatial memories (Squire, 1992).

Hypothalamus and Mamillary Bodies

The hypothalamus is a diencephalic structure surrounding the third ventricle below the thalamus. Due to its extensive reciprocal connections with almost every major part of the central nervous system including the cerebral cortex, hippocampus, amygdala, brain stem, and spinal cord, it is responsible for a wide variety of regulatory functions. In the context of the limbic system, multisynaptic cerebral pathways relay processed information to the hypothalamus, which regulates various kinds of emotional responses accompanied by physiological arousal through the autonomic nervous system (ANS) (Powley and Laughton, 1981). The mamillary bodies are located at the caudal end of the hypothalamus, but they have different functional responsibilities than the other hypothalamic nuclei. Namely, they are not as responsible for autonomic and endocrine functions, but instead serve more as relay nuclei for information coming from the medial temporal lobes and traveling to the anterior thalamus and eventually the cingulate cortex (Irle and Markowitsch, 1982).

Septal Nuclei and Nucleus Accumbens

The septal nuclei are a set of structures inferior to the rostrum of the corpus callosum and anterior to the lamina terminalis. These nuclei have a wide array of reciprocal projections with many brain structures, such as the hippocampus, amygdala, hypothalamus, midbrain, habenula, cingulate gyrus, thalamus, and olfactory bulbs (Raisman, 1966). Along with the nucleus accumbens, which is a part of the ventral striatum, the septal nuclei play a role in reward and reinforcement (Zahm, 2000).

Limbic Cortex

The human neocortex is typically organized into six distinct layers. In Paul Broca's original observations, he noted that all mammals possess a group of cortical areas on the medial surface of the brain that do not follow the normal neocortical organization, but instead show a simpler cellular organization with only three cortical layers. It is this deep band of phylogenetically older tissue that Broca originally named 'the great limbic lobe.' This cortex is now referred to as paralimbic cortex due to its proximity to other limbic structures and includes the cingulate gyrus, subgenual cortex, dentate gyrus, and parahippocampal gyrus. The anterior portion of the parahippocampal gyrus is occupied mostly by entorhinal cortex, a region closely related to the hippocampus. The remainder of the anterior parahippocampal gyrus is part of piriform cortex, which also includes the cortical amygdala and uncus.

As research on memory and emotion has progressed, some neocortical regions have been included in the limbic system. One of these regions is the orbitofrontal cortex, mainly due to its role in emotion and reward (Kringelbach, 2005). This region of the prefrontal cortex has direct and indirect (via the thalamus) connections with the hippocampus and associated neocortex, the amygdala, and the hypothalamus.

Fiber Tracts

Medial Forebrain Bundle

The medial forebrain bundle is perhaps the most well-known hypothalamic fiber tract. Unlike some neural pathways, this tract does not form a well-defined 'bundle' as the name suggests, but instead the term is used to designate a heterogeneous group of axons that run through the lateral hypothalamus. Through descending fibers, this bundle can directly influence preganglionic sympathetic and parasympathetic nuclei in the brain stem and spinal cord. Additionally, some of these fibers connect the amygdaloid body with the lateral hypothalamus and other forebrain regions, allowing the forebrain and limbic regions to modulate both somatic and visceral activity (Powley and Laughton, 1981). This is one example of how information about the body's physical state of arousal is integrated and modulated through limbic structures and connections.

Fornix

The fornix is a large fiber bundle that travels anteriorly through the thalamus, eventually terminating in the medial and lateral parts of the hypothalamus, mamillary bodies, septal nuclei, and nucleus accumbens. Both the hippocampus and amygdala have extensive projections to the hypothalamus through the fornix. Many neurons from the subiculum of the hippocampal formation travel through the fornix and eventually terminate in the mamillary bodies. Efferents from the mamillary bodies travel through the mamillothalamic tract to the anterior thalamus.

Stria Terminalis and Ventral Amygdalofugal Bundle

The stria terminalis is a projection of reciprocal fibers that connect the amygdaloid body and the medial hypothalamus. It follows a similar trajectory as the fornix, traveling dorsally behind and above the thalamus. In the region of the anterior commissure, the stria terminalis divides and distributes fibers to the bed nucleus of the stria terminalis, the medial hypothalamus, and other areas of the basal forebrain. This tract allows the amygdala to regulate hypothalamic functions. Like the stria terminalis, the ventral amygdalofugal bundle also arises from neurons in the amygdala and terminates in the hypothalamus. These tracts between the amygdala and the hypothalamus essentially allow the amygdala to indirectly regulate the endocrine and autonomic components in the expression of emotional responses.

Functions of the Limbic System

The more we explore the neural circuitry driving emotion, cognition, and behavior, the more we discover the extensive role of limbic system structures in these processes. Here, we will explore some of the specific roles of 'limbic system' networks, which include the structures described above, in emotion, learning and memory, attention and social processing.

Emotional Memory

Klüver and Bucy (1937) were among the first to demonstrate the role of temporal lobe structures, particularly the amygdala, in emotion when they demonstrated that medial temporal lobe lesions in rhesus monkeys resulted in a range of odd behaviors, including approaching feared objects and exhibiting an array of unusual sexual behaviors. The most dramatic emotional change in the monkeys with bilateral temporal lobectomies was a marked decrease in fear and aggression, implicating temporal lobe structures in regulating emotional responses. Since then, huge strides have been made in uncovering the role of the amygdala and other associated brain regions in mediating emotional influences on attention, perception, and regulation of emotional responses (Phelps, 2006).

When discussing emotion in the context of the limbic system, there are two phenomena in emotional memory that should be addressed. First, there is the process of associating positive or negative emotional dispositions toward previously neutral stimuli. Much of the focus of research on this emotional association learning has been on the neural mechanisms of fear (Davis, 1992; LeDoux, 2000; Fanselow and Gale, 2003), so we will limit our exploration of the neural mechanisms of emotion to this circuit. Second, there is the widely observed phenomenon whereby memories for events that resulted in emotional arousal tend to be more accurate and longer lasting than memories of neutral events (Berlyne, 1969; Heuer and Reisberg, 1992; Kleinsmith and Kaplan, 1963). This mechanism for memory storage is evolutionarily adaptive, as emotional memories are generally the ones that an organism needs to remember the most. Interestingly, the amygdala is implicated in both of these situations.

In order to survive and thrive in an environment, it is imperative that an organism be able to form affective associations between sensory information and situation. This is particularly true for learned fear. Through socialization or painful experience, an organism must learn to avoid certain behaviors to escape/avoid harm. These fear associations are mediated by the amygdala independently of the events in which the disposition was acquired (Kim and Fanselow, 1992; Phillips and LeDoux, 1992). In other words, normal emotional conditioning can occur even if declarative memory for the learning situation is impaired (e.g., through hippocampal damage). In a fearful situation, sensory information reaches the basolateral nucleus of the amygdala, and the signal is then relayed to the central nucleus. Efferents from the central nucleus project to the hypothalamus and other subcortical areas controlling a broad range of fear-related responses, including activation of the sympathetic nervous system and motor behavior (e.g., increases in heart rate, blood pressure, perspiration, startle response somatomotor immobility, and ultrasonic vocalizations) (for reviews, see Sarter and Markowitsch, 1985; Kapp et al., 1992; McGaugh et al., 1992). After these associations have been established, if the organism later finds itself in a situation resembling the one in which the fear association was created, then it will demonstrate these fear-related responses, even in the absence of the fearful stimulus itself.

To address the assertion that memories tied to emotional events are stronger and more vivid than neutral memories, again the amygdala comes into play. The role(s) of the amygdala in modulating the consolidation of memories of emotionally arousing experiences has been extensively studied over the past several decades. These processes are beyond the

scope of this article, but are detailed in other recent reviews (McGaugh, 2004; Phelps, 2006; Schultz, 2006; Shors, 2006). In summary, when an event happens that is emotionally arousing, epinephrine and glucocorticoids are released by the adrenal glands in response to increased activity in the ANS. This leads to an increase in norepinephrine and consequentially an increase in activity in the amygdala. Evidence from a large number of animal studies suggests that the amygdala's connections with the striatum, hippocampus, neocortex, and nucleus basalis (an indirect path that leads to innervation of much of the cortex) allow it to mediate memory consolidation in other brain regions in response to an emotional event. In other words, the consolidation of what would simply be a declarative memory is enhanced in situations that evoke strong emotions. While the involvement of the amygdala in creating stimuli associations is somewhat independent, this particular phenomenon works in conjunction with the next network that will be discussed: the hippocampal memory system.

The Hippocampal Memory System

Current evidence suggests that there are multiple memory systems that are governed by distinct and interacting neural substrates. These memory systems compete, collaborate, and/or, work in parallel to support and drive behavior. We discussed the amygdala's role in emotional memory in the previous section. Here, we will focus the other major memory system that is closely linked to the limbic system: the hippocampal memory system.

The hippocampal memory system involves the hippocampal formation and regions of the cerebral cortex that serve both as a source of information to the hippocampus and parahippocampal gyrus and as targets for hippocampal projections. Research on the involvement of the hippocampus in memory was sparked by an epilepsy patient in the 1950s named Henry Molaison, known famously as H.M. In order to control his seizures, a neurosurgeon named William Scoville removed a substantial amount of tissue from the medial aspect of Molaison's temporal lobes. While the surgery alleviated his seizures, Molaison was left with severe anterograde amnesia and showed almost no capacity for new episodic learning (Scoville and Milner, 1957). Molaison's memory deficits were first characterized by Brenda Milner, who performed a battery neuropsychological test on him. Ultimately, she revealed that while Molaison showed profound memory deficits, his intellectual and perceptual abilities were generally normal, he was able to hold small amounts of information as long as he was actively rehearsing the information, his childhood memories were intact and he had the ability to acquire new motor skill learning (Milner, 1962). This was some of the first evidence that neural mechanisms underlying procedural and declarative, as well as short-term and long-term memory, are not identical.

Our understanding of memory systems has advanced greatly since the H.M. case. There is now a large body of literature that describes how damage to the hippocampus and parahippocampal regions results in the inability of the brain to consolidate new declarative memories, while perceptual, motor, and intellectual functions remain largely intact. For a long time, based on the observation from human lesion studies, it was thought that the key structures that lead to a deficit in declarative memory consolidation were the hippocampus and amygdala. However, research now shows that selective lesions to the amygdala and hippocampus can have little or no effect on recognition memory and may only produce mild amnesia (Mishkin, 1978). The most severe memory deficits actually result from damage to the perirhinal cortex, which is an area of cortex where information from all sensory modalities converges before it is relayed to the hippocampus (Zola-Morgan et al., 1989). Hippocampal damage may also result in retrograde memory loss, but this effect is time-limited (Kim and Fanselow, 1992). In other words, the hippocampal formation is involved in memory consolidation and reconsolidation for a certain period of time, but once the memory has been represented in the brain for a prolonged period of time, memory retrieval can be supported by other areas of the brain. Scientists have hypothesized distinct roles for the hippocampus and associated cortices in learning and memory, but thus far, there is no general consensus about the distinction.

One popular line of research that aims to form a causal link between the hippocampus and memory formation stems from the discovery of long-term potentiation (LTP) by Terje Lømo in 1966 in Oslo, Norway, laboratory of Per Andersen (Lømo, 1966). There, Lømo and his collaborator Timothy Bliss conducted a series of neurophysiological experiments in both anesthetized and unanesthetized rabbits in which they were able to observe rapid and persistent increases in synaptic strength in postsynaptic neurons of the dentate gyrus induced by a brief burst of stimulation in presynaptic neurons in the perforant path (Bliss and Lømo, 1973; Bliss and Gardner-Medwin, 1973). Subsequent studies have shown that there are many mechanisms for the induction of LTP in many additional areas of the brain (e.g., cerebellum, neocortical regions, and the amygdala). The tremendous experimental interest in LTP is driven by the working hypothesis that this form of synaptic plasticity may participate in encoding and retrieval of information. Hebbian forms of LTP exhibit associativity, which supports the postulate that LTP is a neural correlate of learning and memory. However, while evidence is accumulating for the causal link between LTP and memory (e.g., Martin et al., 2000; Whitlock et al., 2006), the task has proven to be rather difficult.

The hippocampal formation is also involved in spatial memory and navigation. This idea stemmed from rodent research in the 1970s by John O'Keefe and his colleagues that described neurons in the hippocampus that showed activity in response to the rat's location in its environment (O'Keefe and Dostrovsky, 1971). Since then, place responses in rodents have been investigated in hundreds of experiments, and while there is still no general consensus on the exact mechanisms of spatial navigation or whether these place cells exist in the human brain, there is now almost universal agreement that spatial coding plays a large role in hippocampal function.

Lesions in areas of the brain outside the temporal lobes that receive projections from the hippocampus through the fornix can also lead to memory deficits. Recall from the anatomy section that neurons from the hippocampal formation project to the hypothalamus and mamillary bodies through the fornix, from the mamillary bodies to the anterior nucleus of the thalamus through the mamillothalamic tract, and then from

the thalamus to the cingulate cortex. Evidence suggests that lesions to any of these structures may leads to mild or severe memory deficits, depending on the size and location of the lesions (Aggleton et al., 1995; Ennaceur et al., 1997; Kapur et al., 1996).

Additional Roles of Limbic System Structures

As we have discussed above, in order to address questions of how complex psychological functions are produced by specific neural circuits, a series of neural 'systems' have been constructed to describe specific functions. However, at a certain level of behavioral complexity, it becomes difficult to create clear definitions between these systems. We have already discussed the extensive role of limbic system structures in emotion, learning, and memory. As behaviors and cognitions become increasingly complex, emotion and memory become critical components of more complex processes such as reward, motivation, and social processing. For example, in the absence of immediate goals, an organism must be able to use the memory of past experiences and associations to predict the likelihood of an occurrence. This classical or instrumental learning reinforces certain behaviors, and – in conjunction with the nigrostriatal and mesolimbic dopamine systems – is crucial in mediating the processes of reward and motivation (Wise, 2004).

The limbic system, specifically the amygdala, is also involved in social processing. In fact, one of the first studies of human amygdala lesions showed that bilateral amygdala damage can lead to impairments in the processing of faces and other social signals (Jacobson, 1986). Later lesion and functional neuroimaging studies confirm the importance of the amygdala in processing emotional facial expressions, especially fear (Adolphs et al., 1994; Anderson et al., 2003; Breiter et al., 1996; Morris et al., 1996). Additional evidence suggests that the amygdala's response to facial expressions can be conditioned (Morris et al., 1998), implicating that the amygdala's role in fear conditioning and memory consolidation may play a role in this type of social processing.

The previous examples are only a few ways in which the limbic system is involved in complex cognitions and behaviors. Because we use emotion, learning, and memory so extensively, the roles of the limbic system are essentially limitless.

Moving Forward

While the notion of the limbic system has spurred a variety of important scientific experiments aiming to uncover the neural correlates of emotional behavior, learning, and memory, the limbic system remains an enigma to many scientists. As reflected above, one major problem in the conceptualization of the limbic system is its inherent lack of clearly defined boundaries between limbic and nonlimbic forebrain regions. While the limbic system was first described as the emotional system of the brain, recent research on the functions of limbic structures has led to an understanding that the system is much more complex and elusive than originally construed. Though limbic system structures do play a large role in emotional processing, they also have many other functional identities. As such, the limbic system should not be conceptualized as a functionally complete unit. In fact, no single functional grouping of the brain should be considered independent. It should never be forgotten that the brain is constantly using its countless networks of billions of neurons to work together to accomplish specific goals. While the unclear boundaries of the limbic system have created problems for scientists, the use of the term 'limbic system' may not be completely obsolete, as long as it is understood that any area of the brain may have some degree of limbic involvement (as it may have involvement in any brain system) at certain times or in certain situations. The role of the limbic system adapts to the needs of the organism at any given time, and the degree to which different brain regions are involved with the limbic system also adjust to ensure that all of the brain's systems operate together as a harmonious whole.

See also: Behavioral and Cognitive Neuroscience Memory: General Issues Motivational/Emotional Psychology

Bibliography

Adolphs, R., Tranel, D., Damasio, H., Damasio, A., 1994. Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. Nature 372 (6507), 669–672.

Aggleton, J.P., Neave, N., Nagle, S., Hunt, P.R., 1995. A comparison of the effects of anterior thalamic, mamillary body and fornix lesions on reinforced spatial alternation. Behavioural Brain Research 68 (1), 91–101.

Anderson, A.K., Christoff, K., Panitz, D., De Rosa, E., Gabrieli, J.D., 2003. Neural correlates of the automatic processing of threat facial signals. The Journal of Neuroscience 23 (13), 5627–5633.

Berlyne, D.E., 1969. Arousal, reward and learning. Annals of the New York Academy of Sciences 159 (3), 1059–1070.

Broca, P., 1878. Anatomie comparee des circonvolutions cerebrales. Le grand lobe limbique et la scissure limbique dans la serie des mammifers. Review of Anthropology 1 (Ser. 2), 385–498.

Bliss, T.V., Lømo, T., 1973. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. The Journal of Physiology 232 (2), 331–356.

Bliss, T.V., Gardner-Medwin, A.R., 1973. Long-lasting potentiation of synaptic transmission in the dentate area of the unanaesthetized rabbit following stimulation of the perforant path. The Journal of Physiology 232 (2), 357.

Breiter, H.C., Etcoff, N.L., Whalen, P.J., Kennedy, W.A., Rauch, S.L., Buckner, R.L., Rosen, B.R., 1996. Response and habituation of the human amygdala during visual processing of facial expression. Neuron 17 (5), 875–888.

Davis, M., 1992. The role of the amygdala in fear and anxiety. Annual Review of Neuroscience 15 (1), 353–375.

Ennaceur, A., Neave, N., Aggleton, J.P., 1997. Spontaneous object recognition and object location memory in rats: the effects of lesions in the cingulate cortices, the medial prefrontal cortex, the cingulum bundle and the fornix. Experimental Brain Research 113 (3), 509–519.

Fanselow, M.S., Gale, G.D., 2003. The amygdala, fear, and memory. Annals of the New York Academy of Sciences 985 (1), 125–134.

Heuer, F., Reisberg, D., 1992. Emotion, arousal, and memory for detail. The Handbook of Emotion and Memory: Research and Theory, 151–180.

Irle, E., Markowitsch, H.J., 1982. Connections of the hippocampal formation, mamillary bodies, anterior thalamus and cingulate cortex. Experimental Brain Research 47 (1), 79–94.

Isaacson, R.L., 2001. Limbic system. International Encyclopedia of the Social and Behavioral Sciences 64 (10), 8858–8862.

Jacobson, R., 1986. Disorders of facial recognition, social behaviour and affect after combined bilateral amygdalotomy and subcaudate tractotomy – a clinical and experimental study. Psychological Medicine 16 (02), 439–450.

Kapp, B.S., Whalen, P.J., Supple, W.F., Pascoe, J.P., 1992. Amygdaloid Contributions to Conditioned Arousal and Sensory Information Processing.

Kapur, N., Thompson, S., Cook, P., Lang, D., Brice, J., 1996. Anterograde but not retrograde memory loss following combined mammillary body and medial thalamic lesions. Neuropsychologia 34 (1), 1–8.

Kim, J.J., Fanselow, M.S., 1992. Modality-specific retrograde amnesia of fear. Science 256 (5057), 675–677.

Kleinsmith, L.J., Kaplan, S., 1963. Paired-associate learning as a function of arousal and interpolated interval. Journal of Experimental Psychology 65 (2), 190.

Klüver, H., Bucy, P.C., 1937. "Psychic blindness" and other symptoms following bilateral temporal lobectomy in Rhesus monkeys. American Journal of Physiology 119, 352–353.

Kringelbach, M.L., 2005. The human orbitofrontal cortex: linking reward to hedonic experience. Nature Reviews Neuroscience 6 (9), 691–702.

LeDoux, J.E., 2000. Emotion circuits in the brain. Annual Review of Neuroscience 23 (1), 155–184.

Lømo, T., 1966. Frequency potentiation of excitatory synaptic activity in the dentate area of the hippocampal formation. Acta Physiologica Scandinavica 68 (Suppl. 277), 128.

MacLean, P.D., 1952. Some psychiatric implications of physiological studies on frontotemporal portion of limbic system (visceral brain). Electroencephalography & Clinical Neurophysiology 4 (4), 407–418.

Martin, K.C., Barad, M., Kandel, E.R., 2000. Local protein synthesis and its role in synapse-specific plasticity. Current Opinion in Neurobiology 10 (5), 587–592.

McGaugh, J.L., Introini-Collison, I.B., Cahill, L., Kim, M., Liang, K.C., 1992. Involvement of the Amygdala in Neuromodulatory Influences on Memory Storage.

McGaugh, J.L., 2004. The amygdala modulates the consolidation of memories of emotionally arousing experiences. Annual Review of Neuroscience 27, 1–28.

Milner, B., 1962. Les troubles de la memoire accompagnant des lesions hippocampiques bilaterales. Physiologie de l'hippocampe, 257–272.

Mishkin, M., 1978. Memory in monkeys severely impaired by combined but not by separate removal of amygdala and hippocampus. Nature 273, 297–298.

Morris, J.S., Frith, C.D., Perrett, D.I., Rowland, D., Young, A.W., Calder, A.J., Dolan, R.J., 1996. A differential neural response in the human amygdala to fearful and happy facial expressions. Nature 383 (6603), 812–815.

Morris, J.S., Öhman, A., Dolan, R.J., 1998. Conscious and unconscious emotional learning in the human amygdala. Nature 393 (6684), 467–470.

O'Keefe, J., Dostrovsky, J., 1971. The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. Brain Research 34, 171–175.

Papez, J.W., 1937. A proposed mechanism of emotion. Archives of Neurology and Psychiatry 38, 725–743.

Phelps, E., 2006. Emotion and cognition: insights from studies of the human amygdala. Annual Review of Psychology 57, 27–53.

Phillips, R.G., LeDoux, J.E., 1992. Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning. Behavioral Neuroscience 106 (2), 274–285.

Powley, T.L., Laughton, W., 1981. Neural pathways involved in the hypothalamic integration of autonomic responses. Diabetologia 20 (3), 378–387.

Raisman, G., 1966. The connexions of the septum. Brain 89 (2), 317–348.

Sarter, M., Markowitsch, H.J., 1985. Involvement of the amygdala in learning and memory: a critical review, with emphasis on anatomical relations. Behavioral Neuroscience 99 (2), 342.

Schultz, W., 2006. Behavioral theories and the neurophysiology of reward. Annual Review of Psychology 57, 87–115.

Scoville, W.B., Milner, B., 1957. Loss of recent memory after bilateral hippocampal lesions. Journal of Neurology, Neurosurgery and Psychiatry 20 (11), 11–21.

Shors, T.J., 2006. Stressful experience and learning across the lifespan. Annual Review of Psychology 57, 55–85.

Squire, L.R., 1992. Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. Psychological Review 99 (2), 195.

Whitlock, J.R., Heynen, A.J., Shuler, M.G., Bear, M.F., 2006. Learning induces long-term potentiation in the hippocampus. Science Signaling 313 (5790), 1093.

Wise, R.A., 2004. Dopamine, learning and motivation. Nature Reviews Neuroscience 5 (6), 483–494.

Young, M.P., Scanneil, J.W., Burns, G.A., Blakemore, C., 1994. Analysis of connectivity: neural systems in the cerebral cortex. Reviews in the Neurosciences 5 (3), 227–250.

Zahm, D.S., 2000. An integrative neuroanatomical perspective on some subcortical substrates of adaptive responding with emphasis on the nucleus accumbens. Neuroscience & Biobehavioral Reviews 24 (1), 85–105.

Zola-Morgan, S., Squire, L.R., Amaral, D.G., Suzuki, W.A., 1989. Lesions of perirhinal and parahippocampal cortex that spare the amygdala and hippocampal formation produce severe memory impairment. The Journal of Neuroscience 9 (12), 4355–4370.

Liminality

Hazel Andrews, Liverpool John Moores University, Liverpool, UK
Les Roberts, University of Liverpool, Liverpool, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by B.A. Babcock, volume 13, pp. 8862–8864, © 2001, Elsevier Ltd.

Abstract

This article explores the concept of 'liminality,' tracing its development from the work of the French ethnologist Arnold van Gennep, to the pioneering theoretical anthropology of Victor Turner, and goes on to consider ways that ideas of liminality have informed more recent discussions and debates, particularly in relation to the anthropology of experience, tourism and mobility, and place and space. Arguing that what is understood by 'liminality' today encompasses a wide array of meanings and theoretical associations, the article examines the extent to which the concept finds resonance with a rich and diverse field of social, cultural, and geographical enquiry.

The concept of 'liminality,' much like the phenomena to which it is generally held to ascribe, is itself passing through something of a transitional phase. From its origins in the work of the French ethnologist Arnold van Gennep, through to the pioneering theoretical anthropology of Victor Turner in the 1960s and 1970s, what is understood by the term liminality today encompasses a wide array of meanings and theoretical associations that go far beyond those more traditionally linked with anthropological discussions on ritual, performance, pilgrimage and cross-societal 'rites of passage.' As Meethan (2012: p. 69) notes, the concept can be found in studies on management, health, education, cyber space, governance, sexuality, and tourism, to name but a few. One of the foremost points of critical intersection that has propelled renewed interest in liminality and liminal phenomena in recent years is the relationship between liminality and place and space. Indeed, those surveying the abundant literature on border zones, nonplaces, interstitial spaces, edgelands, contact zones, spaces of hybridity, third space, heterotopias, and so on, could be forgiven for thinking that liminality is and ever has been intrinsically spatial in terms of the discursive landscapes within which discussions on liminality are invariably situated.

While undoubtedly such trends can in part be attributed to the impact of a much broader 'spatial turn' (Warf and Arias, 2009) that has left its mark on social and cultural theory in recent decades, it can also be taken as evidence of a concerted push to reengage with the concept of liminality in ways that acknowledge, but which at the same time are by no means bound by the foundational work of figures such as van Gennep and Turner. As Thomassen (2009) notes, the concept has 'traveled' since it was first introduced in van Gennep's influential book *Rites de passage*. Accordingly, in the following text the authors set out some of the important theoretical background to liminality, acknowledging that van Gennep himself has often remained a marginal figure in the intellectual history of a concept that was to become so indelibly linked with the work of Turner. The authors then turn to some of the ways that liminality can more productively be understood in relation to wider discursive fields to which it is affiliated or into which it has encroached, particularly those that cluster under the rubric of the anthropology of experience, tourism and mobility, emotions and the self, and those linked to discussions of landscape, place and space. Given the more diffuse theoretical terrain that has come to frame discussions of liminality and liminal phenomena, and the attendant difficulties this can bring with it in terms of clearly delineating the meanings and practices that potentially fall under its banner, there is, as the authors have argued elsewhere (Andrews and Roberts, 2012), a need to rethink and 'remap' liminality. As the cartographic reference here alludes, this is a process which has most notably sought not only to explore the constitutive spatialities of liminality, but also to examine the different ways, and the extent to which, 'in-between spaces' can be said to represent one of the defining characteristics of modern topographies of self and society.

Arnold van Gennep

What the authors have come to understand as liminality in relation to social and cultural practices stems in no small part from the work of Arnold van Gennep (1873–1957), whose *Rites de passage* was published in 1909, but not translated into English until 1960. Measured against the legacy of sociological and anthropological 'founding fathers' such as Emile Durkheim, van Gennep's body of work has remained on the periphery of theory and debate (Thomassen, 2009: p. 7). With an impressive 437 titles in his bibliography, van Gennep, who mastered 18 different languages (Thomassen, 2009: p. 8), published on a wide range of topics but it is his work on liminality that has by far proved the most influential.

Derived from the Latin word for 'threshold' (*limen*) – an etymological reminder of the important spatial underpinnings to the concept – for van Gennep, 'liminal' is more directly related to the symbolic processes and ritual conventions that structure and define key moments of social transition, or 'rites of passage.' His research examined the relationship between liminal experiences and lasting effects of those experiences, arguing that structure/order is derived from liminality – from the crucial middle or 'in-between' states that facilitate ritual passage from one social stage to another. For van Gennep,

liminality does not presuppose certainty or offer an explanatory causality with regard to outcome, but rather describes a world of contingency "where … 'reality' itself, can be carried in different directions" (Thomassen, 2009: p. 5). Liminality is thus a means to conceptualize the moments between structure and agency: "in liminality, the very distinction between structure and agency ceases to make meaning; and yet, in the hyper-reality of agency in liminality, structuration takes place" (Thomassen, 2009: p. 5).

As elaborated by van Gennep, rites of passage form an explanatory codification or analytical tool by which to understand all rites and his work distinguishes between different types of ritual, ranging from change in status to those that mark the passage of time. Rites of passage become a special category with three subcategories: separation, transition (the liminal phase), and reincorporation (the postliminal stage). Basing his analysis on existing empirical and ethnographic data, van Gennep observed a pattern in rituals that he argued is universal and replicable across cultures and societies (Thomassen, 2009: p. 6). Celebrations or initiations are important features of social rites of passage that are classified into four types of social movement:

1. The passage of people from one status to another, e.g., marriage (moving from fiancée to wife or husband), initiation ceremonies in which the outsider of a group becomes an insider.
2. The passage from one place to another, e.g., moving home.
3. The passage from one situation to another, e.g., starting a new job, school, joining university.
4. The passage of time – typically the passage of a whole social group moving from one period to another, e.g., New Year, New Government or Ruling authority (King/Queen/Emperor) (see Andrews and Leopold, 2013: p. 36).

van Gennep's analysis identified certain characteristic patterns that occurred in the order of the ceremonies:

- Separation/preliminal – the physical detachment of the participant from normal life. As Gunnell (2007) observes of Busar initiation traditions in Icelandic Gymnasia, new pupils are only allowed to walk along marked pathways which separate them from the spaces of movement used by the rest of the school's students and teachers.
- Liminal/transitional – the most important period according to Victor Turner in which the participant is literally and symbolically marginalized. During this time, the participants in the ritual have a sense of *Communitas*. Communitas – a concept later developed by Turner (1969) – is a feeling of camaraderie between those who are the focus of the ritual. Gunnell's (2007) study again proves useful here as the new pupils are argued to feel a sense of bonding with each other as they share the same ordeals – wearing odd clothes and participating in rituals – before they become fully accepted into the life of the school.
- Incorporation/postliminal – the participant is reincorporated into society/returned to society. In Gunnell's Icelandic example this occurs when, having completed their challenges, the new pupils are finally accepted into the life of the school, which is celebrated with a meal or dance (see Andrews and Leopold, 2013: p. 36).

Melford Weiss's 1967 article 'Rebirth in the Airborne' applies van Gennep's tripartite model of liminality to the training and initiation rituals involved in becoming an American paratrooper. As part of their training, trainees are separated from and discouraged from association with those in the outside world. The liminal phase is typically accompanied by the use of lucky charms and totems, and, in an inversion of normal social experience, initiates are subjected to continuous periods of anxiety and stress, which helps bind them together as a group and helps inform feelings of communitas. Lastly, reincorporation, which includes the playing of the national anthem and the reciting of scriptures, culminates in the 'prop blast,' a ceremonial rebirth in which initiates reenact their jump in front of the flight sergeant, and the ritual drinking of 'blast juice,' which is quaffed within the count of '1000, 2000, 3000,' the time between an actual jump from a plane and the opening of the parachute. On successful completion, the initiate is ritually joined with his fellow paratroopers. Although paratrooper training is an essentially secular affair, "certain superstitious practices which are interwoven show that, in the broadest sense, it is also a religious rite" (Weiss, 1967: p. 24).

Building on ideas van Gennep outlined in *Rites de passage*, Turner argued that each rite of passage has these elements (separation – transition – reincorporation) within it, although some may be more developed than others depending on the purpose of the rite. In a funeral, for example, rites of separation, would be more pronounced than those typically associated with a wedding or graduation ceremony.

Victor Turner

As previously noted, van Gennep's ideas were developed most notably by the British social anthropologist Victor Turner (1920–83). In particular, Turner explored the concept of liminality that van Gennep had developed in relation to both the 'middle' phases of all rites and especially to rites of transition. Turner's first published work drawing on van Gennep's theory was an article on Ndembu circumcision ritual (1962). The Ndembu tribe of Zambia was the main focus of Turner's early work on ritual and the role of symbols in social processes, which he went on to develop more fully in *The Forest of Symbols*, published in 1967. For Turner, liminality is constituted in a condition or happenstance that has attributes of doubt and lack of inevitability between what is known and has gone before and future outcomes. The particular importance of Turner's work is its focus on the creative potential of ritual. In so doing, Turner put emphasis on the liminal phase connecting rites with notions of performance and investigating rites as times of suspension during which the way forward is characterized in terms of possibility and becoming. During the liminal phase of a ritual, initiates are in a 'betwixt and between' status, a threshold from which a new identity or social status can emerge. They, therefore, have the possibility to change from their former, preritual status into something different when they are reincorporated back into the everyday social world.

Turner became aware of van Gennep's work on ritual and liminality following the publication of the English translation of

Rites de passage in 1960. As one of the main proponents of van Gennep's theory, Turner played a pivotal role in drawing it to wider attention amongst anthropologists and others. Thomassen (2009: p. 14) argues that Turner discovered the work experientially, as the anthropologist was himself, at that time, betwixt and between academic posts at Manchester and the United States. This important transitional period would become manifest in the Chapter 'Betwixt and Between: the Liminal period in rites of passage,' in *Forest of Symbols* (1967). As Thomassen (2009) notes, Turner argued against the Durkheimian view that rituals mirror the dominant social order. Turner had already introduced the idea of 'social drama' in his work and found that van Gennep's formulations on rites of passage sat well alongside the idea. Turner (1982: p. 27) argued that during the liminal phase of rituals, play is often involved: "Liminality may involve a complex sequence of episodes in sacred space-time, and may also include subversive and ludic (or playful) events … in liminality people 'play' with the elements of the familiar and defamiliarize them." With this emphasis on playful action, liminality becomes significant in understanding individual and group responses to the 'betwixt and between' rather than emphasizing only its role as part of a sequence of events. For Turner, the liminal phase presents the opportunity of being reshaped or 'fashioned anew' (1969: p. 81) in which the persons at the center of the experience are prepared for and embody their new statuses. In short, liminal experiences shape personality, focus mind and experience, and foreground human agency (Thomassen, 2009: p. 14). Drawing on the work of philosophers such as Dilthey, the emphasis on experience and performance would become key elements in Turner's writings on liminality and liminal phenomena (see Section Liminality and Anthropology of Experience).

In *The Ritual Process*, Turner argues that the period of liminality or liminal people (that is people on the threshold of becoming different) 'are necessarily ambiguous' (1969: p. 81) defying classification in terms of the world of the everyday: "liminal entities are neither here nor there; they are betwixt and between the positions assigned and arrayed by law, custom, convention, and ceremonial" (1969: p. 81). The 'threshold' people Turner describes are typically stripped of property or clothing which would mark their rank or status in the liminal phase making them indistinguishable from others who are sharing that experience.

A particularly notable innovation in Turner's theoretical updating of van Gennep's earlier ideas is the distinction he makes between liminal and 'liminoid.' Turner (1982) suggests that in modern consumerist societies in which work and play are more clearly defined than in traditional societies the possibilities for liminal experiences to arise are reduced, being supplanted mainly by liminoid moments found in artistic performances and the practices of leisure consumption. The ritual and experiential liminal characteristics ascribed to so-called traditional societies are transmuted into liminoid experiences through play, creativity, drama, and their associated art forms, such as theater and literature. Although perhaps based on too basic a distinction between traditional and nontraditional, Turner's concept of the liminoid allows for a more qualified or conditional sense of the liminal insofar as it defines a transitional moment that is optional and which does not involve a change of status or offer a resolution (Thomassen, 2009: p. 15). The playfulness and experiential qualities of liminoid phenomena can be transformative but they are not transitional in the same way that a rite of passage represents a liminal state of 'betwixt and between.' In 'From Ritual to Theatre: The Human Seriousness of Play' (1982), Turner notes that 'oid' derives from the Greek word *eidos*, meaning a form or shape. Liminoid thus takes the form of or resembles but is not the same as the liminal. It represents "an independent domain of creative activity" (1982: p. 33). Thus Turner argued theoretical and experimental science is liminoid as it is conducted in or 'privileged areas' (laboratories or studies) separate from society in general and issues of central political concern: "Universities, institutes, colleges, etc., are 'liminoid' settings for all kinds of freewheeling, experimental cognitive behavior as well as forms of symbolic action, resembling some found in tribal society, like 'rushing' and 'pledging' ceremonies in American college fraternity and society houses" (1982: p. 33).

The distinction between pilgrimage and tourism as, respectively, forms of liminal and liminoid ritual is one that has generated discussion in anthropological studies of tourism (see below), where the distinction is at times difficult to draw. Graburn (1989), for example, has argued that tourism is a form of pilgrimage in that the experience is that of a liminal rite of passage. Pilgrimage, a focus of some of Turner's later work with his wife Edith on liminality and ritual (1978), entails an equaling out of participants: status hierarchies are temporarily suspended and the pilgrims share a common experience of communitas, building on ideas developed in Turner's (1969) earlier work. Similar arguments have been made in respect of rave, dance, and techno cultures and subcultures (see, for example, Gerard, 2004; St. John, 2008), prompting debate as to whether these represent liminal or liminoid experiences of transition and communitas.

An important but often overlooked characteristic of the liminal phase is its association with danger. The attendant risk attached to the entering of a liminal phase or liminal space is of losing a stable sense of self and identity, or of not being reincorporated into the social world in ways that had been anticipated or desired. For those seeking drug-induced and consciousness-altering experiences of communitas, for example, the psychological and physical impacts may have permanent or long-term consequences that take the experience outside of the temporary and otherwise contained environment of the liminal phase. Death itself is also strongly associated with liminality (see Das, 1977). Lastly, the suspended nature of social life normally associated with the state of liminality can become permanent in circumstances where it is institutionalized and incorporated into dominant social and political structures of power (Thomassen, 2009: p. 22).

Liminality, Place, and Space

In his account of a walk along an ancient and notoriously hazardous tidal path situated off the Essex coast, the writer Robert Macfarlane observes, "[w]e lack – we need – a term for those places where one experiences a 'transition' from a known landscape … [to] somewhere we feel and think significantly

differently. I have for some time been imagining such transitions as 'border crossings.' These borders do not correspond to national boundaries, and papers and documents are unrequired at them. Their traverse is generally unbiddable, and no reliable map exists of their routes and outlines" (2012: p. 78).

Although he does not use the term, in this quotation Macfarlane succinctly outlines some of the defining characteristics of a liminal landscape and experience. Liminality is of inferential significance here insofar as it is 'located' both in terms of its geography and its constitutive temporality. To experience the transition, Macfarlane describes entails a process of traversing space, whether real (the physical landscape as a site of wayfinding or pilgrimage) and/or imagined (the social, symbolic, and textual spaces of transition that invest the landscape with meaning). It is also an experience that has duration: it unfolds over a period of time. But crucially, like the landscape itself, the 'routes and outlines' that otherwise order and rationalize time are – experientially at least – difficult to 'map' or reliably navigate. The pilgrim or wayfarer enters 'another place,' which, like artist Anthony Gormley's installation of the same name located on Crosby beach near Liverpool (Andrews, 2012), is governed by the contingent workings of the natural elements. By the same token, the quintessential liminality of the spatiotemporal moment is marked by the threat of danger. In the case of the tidal mudflats that Macfarlane describes this, of course, has very obvious connotations. But the extent to which borders and boundaries are elusive and 'unbiddable' may also invoke fears rooted in an incipient sense of existential or ontological displacement. In the same way, the initiate entering the liminal phase of a social ritual may experience the undoing of a fixed or habituated sense of selfhood, the navigator of liminal landscapes potentially 'crosses over' to states of being or consciousness that draw their precarious affectivity from the uncertain and 'unmappable' geographies that are temporally inhabited.

That these transitional spaces might represent 'border crossings' both highlights the structural and experiential liminality of border zones and hints at ways in which liminal landscapes do not necessarily correlate with conventional territorial understandings of borders such as those that relate to national boundaries. Border zones, like all liminal phenomena, are ambivalent in that they are at once spaces of mobility and immobility, transition and stasis (Roberts, 2002), hope and violence (Harvey, 2000), place and nonplace (Augé, 1995), memory and oblivion (Augé, 2004): spaces where the 'play' of horizons and frontiers (Marin, 1984) reflects the differential power relations that govern the increasingly complex patterns of transnational mobility. As spaces of 'radical openness' (Hooks, 1990), border zones can engender what Gloria Anzaldúa has described as a 'mestiza consciousness,' a consciousness of the borderlands that is the "product of the transfer of the cultural and spiritual values of one group to another" (1987: p. 78). Equally they may take the form of 'smooth spaces' of flow, the domain of the deterritorialized 'nomad,' the figurative (and often romanticized) embodiment of postmodern homelessness who, at home everywhere and nowhere, exists in a constant state of flux and becoming. For the nomad, the border (geographical or metaphorical) represents an otherwise striated space of power that is there to be crossed or 'smoothed out' (Deleuze and Guattari, 1986; Braidotti, 1994). The idea that these and other liminal zones are antiessentialist spaces where identities are contingent, malleable, and fluid, is one that also underpins Bhabha's concept of 'third space,' conceived of as a cultural space where hybrid identities are negotiated and forged, often in resistance to hegemonic power structures. For Bhabha, "[t]he non-synchronous temporality of global and national cultures opens up a cultural space – a third space – where the negotiation of incommensurable differences creates a tension peculiar to borderline existences" (1994: p. 218).

Although concepts such as 'third space,' 'nomad,' and *la mestiza* are predicated on a radically deessentialized space of liminal encounter, arguably they are only able to sustain their deterritorializing logic through a concomitant essentialization of space. As Smith and Katz have commented, "an absolutist spatial ontology … provides the missing foundation for everything else in flux" (1993: p. 79). By reducing the complexities and dynamics of social space (Lefebvre, 1991) to a largely rhetorical space of flow, hybridity and becoming the extent to which identities are caught within – and often at the mercy of – material and economic geographies of transnational liminality can be overlooked. An instructive and tragic case in point is that of the 23 Chinese migrant workers who drowned while picking cockles in the vast expanse of sandflats and mudflats at Morecambe Bay in 2004. Located on the northwest coast of England in Lancashire, Morecambe Bay is notorious for its fast moving tides and treacherous quicksand, as well as its lucrative cockle beds. Despite the dangers posed by this stretch of coastline, the Bay has long attracted itinerant workers, many of whom, as with the Chinese cockle pickers, remain at the mercy of exploitative and illegal gangmasters. While the Morecambe Bay incident brought to light the appalling conditions faced by these migrant workers, it also forced an awareness of the extent to which such groups are an invisible but immutable presence, occupying a ghostly liminal zone on the social and geographic margins of the nation; caught in the interstices of transnational space.

Situated 'betwixt and between' land and sea, Morecambe Bay Sands represent an ambiguous, and thus potentially dangerous zone on account of its unbiddable and 'unmappable' physical terrain. But, as the 2004 incident made clear, such landscapes are also liminal to the extent that they are heterotopias, or spaces of otherness, "juxtaposing in a single real place several spaces, several sites that are in themselves incompatible" (Foucault, 1986: p. 25; Hetherington, 1997). As a space of transnational migrant labor and increasingly globalized social, cultural, and economic relations, coastal resorts such as Morecambe Bay expose the underlying tensions and social divisions between representations that play on the ludic, touristic heritage of these resorts and those which address the marginality and exclusion that characterize the other set of mobilities and meanings evoked by these spaces. In this instance, the precarious and unnavigable natural landscape of Morecambe Sands becomes a metonym for the increasingly destabilizing landscapes of global capitalism. Compare this example with the hybrid spaces depicted in the film *Bhaji on the Beach* (dir: Gurinder Chadha, 1993), in which a group of British-Asian women go on a day trip to Blackpool, also on the Lancashire coast. The sense of

shared communitas and identity that the beachscape connotes in the latter stands in stark contrast to the fear, exploitation, and uncertainty experienced by those occupying very different spaces of marginality and transnational liminality in twenty-first century Britain.

Liminality and Psychotherapy

When planning his walk across Maplin Sands on the Thames Estuary (see above), Robert Macfarlane initially and very sensibly sought the services of a guide who had the detailed local knowledge necessary to navigate the route safely. Wayfarers crossing the sandflats of Morecambe Bay are also encouraged to only undertake the walk with the accompaniment of a guide. An official guide, known as the Sandpilot of Morecambe Bay, regularly leads parties on the 6-mile trek across the sands. As the authors have elsewhere noted (Andrews and Roberts, 2012: p. 8), the role of the guide highlights another common attribute of liminal landscapes in that initiates are required to place their trust in the knowledge of a 'ritual elder' or 'master of ceremonies' so as to ensure safe navigation and transit. Analogies here can be drawn with the relationship between psychotherapist and patient, the former often held to be providing a form of 'ritual leadership' (Moore, 1991: p. 25). In *From Ritual to Theatre*, Turner casts the experimental theater director Jerzy Grotowski in the role of ritual guide-cum-psychotherapist: "Let us create a liminal space-time 'pod' or pilgrimage centre, [Grotowski] seems to be saying, where human beings may be disciplined and discipline themselves to strip off the false personae stifling the individual within" (1982: p. 120). It is important to note, however, that the ritual leader is not the master or controller of transformative space. Such space "cannot be commanded – it can only be invoked" (Moore, 1991: p. 27). As such, liminality "is always within a context of containment in which the boundaries are not tended by the ego of the individual involved" (Moore, 1991: p. 29), but nor, by the same token, are they overly determined by the intervention of the leader or guide.

Navigating the psycholiminal landscapes of the patient, the therapist, like the Sandpilot, is responsible for ensuring the careful stewardship of the boundary beyond which safe and secure transformation cannot be assured. Stressing the importance of the boundary in an article examining the psychoanalyst as ritual elder, Moore argues that:

> when sensing the need for a liminal space, [individuals] will seek out boundary and containment wherever they can. If knowledgeable ritual elders are not present to invoke liminal space and lead them through it, then they gravitate to the liminoid and try to find containment or generate it on their own. The attraction of geographical boundaries such as seashores to those in transition states is a striking example of this intuitive quest for the boundary. Finding a natural boundary is, of course, relatively simple. Locating an appropriate transformative container, however, is much more difficult. (1991: p. 28)

The concepts of liminal and liminoid, and particularly the work of Turner, have been influential in the development of Jungian-based psychoanalysis in which the therapeutic contract between therapist and client is conceived of as a ritual process. The edited collection 'Liminality and Transitional Phenomena' (Schwartz-Salant and Stein, 1991) brings together work by scholars of psychology and religion alongside that of psychotherapists and counselors who all approach the theory and practice of Jungian psychotherapy by engaging closely with ideas of liminality. The importance of the 'container' or 'vessel' as part of the psychotherapeutic ritual process is one that several contributors draw attention to. Hopcke, for example, notes that the concept of liminality is 'felicitously consonant' with Carl Jung's own approach to analysis, which was to provide a space, "a temenos, a magic circle, a vessel, in which the transformation inherent in the patient's condition would be allowed to take place" (1991: p. 117). For Hopcke, liminality is the very essence of effective analytic work, the aim of which is transformation and change.

Liminality and Tourism

The concept of liminality has found its way into more areas of academic enquiry through its use in tourism. In the first place, the idea that tourism marks some form of hiatus in the (yearly) life was expounded in the early days of the anthropological study of tourism in Nelson Graburn's text *Tourism as a Sacred Journey* (1989 [1976]). In this work, Graburn follows van Gennep's formulation of *rites de passage* to understand the touristic experience as a leaving of the normative world, to enter a liminal phase of separation in which the holiday experience allows for the individual to feel as if they have been renewed in some way before their return home upon which they are reincorporated back into the everyday, heavily routinized world they left behind. As part of the liminality of the vacation, the holidaymakers may feel some sense of communitas with their fellow travelers in which they share the trials and tribulations of being in an unfamiliar environment.

The ritualistic and episodic role of tourism as a period which permits different attitudes and behaviors to emerge from that found in the home world was used as a basis to understand the apparently ludic behavior of tourists in relation to their sexual conduct aboard charter tourism yachts in the Caribbean. In his study, James W. Lett Jr. (1983) argues that tourists to the British Virgin Islands treat the place as a playground that allows them to indulge in a form of 'playground culture' (1983: p. 55). The emphasis on play marks the experience as liminoid rather than liminal and draws attention to the distinctions highlighted by Turner (1982) – see above.

Another major contributor to the linking of the concept of liminality to tourism can be found in the work of Rob Shields (1991). In *Places on the Margin*, Shields notes that following the patronage of the Prince Regent, the southern English seaside town of Brighton developed from a fishing village into a seaside resort where dominant discourses could be subverted. For example, Brighton became connected with the 'dirty weekend,' often represented in the form of the 'dirty postcard.' What this referred to in practice was the conduct of illicit sexual relationships. Such was the reputation of

Brighton that it was often used to 'stage' adulterous encounters to further divorce proceedings. Shields argues that the beach took on a carnivalesque role, which allowed the transgression of propriety and societal norms and that this helped to mark it as liminal zone. In her study of charter tourists to Magaluf and Palmanova on the Mediterranean island of Mallorca, Andrews (2011) applies the concept of liminality not only to the beach but to the resorts as a whole, in which the practice of the holiday is marked out as different from the 'everyday' in the actions of the tourists and those mediating their experiences. This can be illustrated by the purchase of clothes especially for the occasion and the amplification of difference (and attendant dangers) promoted by the tour operators.

The degree to which holiday taking can be seen as markedly different from the everyday activities or out of the ordinary has been criticized (see, for example, Veijola and Jokinen, 1994). Andrews (2009) suggests that it is more of a disruption to habitus rather than a complete change per se, arguing that this line of thought can be developed further by thinking more carefully about the concept of experience and in particular the nuances of what it might mean and how it can be used.

Liminality and Anthropology of Experience

In their edited collection *The Anthropology of Experience*, Victor Turner and Edward Bruner (1986) discuss the nature of experience within a dramaturgical context of performance and narrative. In their volume the meaning of experience, as Geertz suggests, remains "the elusive master concept of this collection, one that none of the authors seems altogether happy with and none feels able to really do without" (1986: p. 374). In the English-speaking world, one of the problems encountered when trying to define or discuss experience is the limitations of having only one word of use in the English language, unlike in German, for instance, where there are two: *Erfahrung* and *Erlebnis*. *Erfahrung* refers to the idea of encounters and *Erlebnis* to the world of senses and emotions, intuition, and movement. The *Erlebnis* is not fixed, rather it is a process in which there are a number of possibilities that arise within or through a particular situation; it is the world of the subjective or of being affected by something (see Andrews, 2009). Drawing on the work of the German philosopher Wilhelm Dilthey (1833–1911) Turner argues that the liminal phase allows for "unique structures of experience (Dilthey's *Erlebnis*)" (1983: p. 41). Thus liminality provides the *Erlebnis* form of experience, which is a world of possibilities and becomings: "Liminality can perhaps be described as a fructile chaos, a storehouse of possibilities, not a random assemblage but a striving after new forms and structures, a gestation process, a fetation of modes appropriate to postliminal existence" (1983: p. 42). Turner compares this to 'ordinary' life, which is lived in the indicative mood and argues that liminality, in its abilities to offer possibilities, operates in the subjective. It is these ideas of the exploratory nature and potential of becoming in liminality that link it to the idea of social drama developed most notably by the philosopher John Dewey (1859–1952).

The link between Turner's ideas of liminality and *Erlebnis* had been explored in an earlier text *From Ritual to Theatre*, published in 1982. In this, Turner argues that an anthropology of performance is an integral part of the anthropology of experience and that performance is found in many different types of cultural expression: ritual, carnival, and theater to name but a few. Such performances are themselves processes and it is in this process that "what is normally sealed up, inaccessible to everyday observation and reasoning, in the depth of sociocultural life, is drawn forth" (1982: p. 13). The link to the idea of social drama and liminality being in the subjunctive is again clear: that such experiences allow for the emergence of a number of possibilities. This is echoed in the work of Michael Jackson in his work on existential anthropology, where experience is entwined with notions of 'being.' As Jackson observes "[b]eing is (…) in continual flux, waxing and waning according to a person's situation" (2005: p. x).

Conclusion

The concept of liminality is, as the authors have discussed, closely bound up with experience and the processes of change, transformation, and transition that individuals undergo as part of that experience. Tropes and practices of play, performance, flow, becoming, potential, gestation, passage, movement: these are all in some way constitutive elements that variously inform the essential 'in-between-ness' of liminal experiences and states. Socially, psychologically, and ontologically this in-between-ness entails a process of letting go of stable topographies of selfhood; of 'striving after new forms and structures' as Turner puts it. As a 'fructile chaos' liminality can also, therefore, be characterized in terms of danger (Thomassen, 2012) and of the anxieties and emotional turbulence that accompanies the passage through liminal states or spaces. Increasingly, however, the concept has taken on a far looser set of meanings that has at times reduced what might be understood as 'liminal' to a rather vague sense of being 'on the edge' or 'in-between' in ways that have become decoupled from the concept's attendant (and requisite) social underpinnings. Paying closer critical attention to the spatialities of liminal phenomena, an issue the authors addressed in their book *Liminal Landscapes: Travel, Experience and Spaces In-between* (Andrews and Roberts, 2012), is one of the ways that liminality can be regrounded in the foundational work of van Gennep, Turner and others, whilst at the same time acknowledging the extent to which the concept has 'traveled' and aligned itself with a far broader field of social, cultural, and geographical enquiry.

See also: Borders, Anthropology of; Carnival; Humor, Anthropology of; Performance: Anthropological Aspects; Play, Anthropology of; Tourism; Trickster Ethnography.

Bibliography

Andrews, H., 2009. Tourism as a 'moment of being'. Suomen Antropologi: Journal of the Finnish Anthropological Society 34 (2), 5–21.

Andrews, H., 2011. The British on Holiday Charter Tourism, Identity and Consumption. Channel View, Bristol.

Andrews, H., 2012. Another place or just another space? Liminality and Crosby Beach. In: Andrews, H., Roberts, L. (Eds.), Liminal Landscapes: Travel, Experience and Spaces In-between. Routledge, London, pp. 152–166.
Andrews, H., Leopold, T., 2013. Events and the Social Sciences. Routledge, London.
Andrews, H., Roberts, L., 2012. Introduction: re-mapping liminality. In: Andrews, H., Roberts, L. (Eds.), Liminal Landscapes: Travel, Experience and Spaces In-between. Routledge, London, pp. 1–17.
Anzaldúa, G., 1987. Borderlands/La Frontera: The New Mestiza. Aunt Lute Books, San Francisco.
Augé, M., 1995. Non-places: Introduction to an Anthropology of Supermodernity. Verso, London.
Augé, M., 2004. Oblivion. University of Minnesota Press, Minneapolis.
Bhabha, H., 1994. The Location of Culture. Routledge, London.
Braidotti, R., 1994. Nomadic Subjects: Embodiment and Sexual Difference in Contemporary Feminist Theory. Columbia University Press, New York.
Das, V., 1977. Structure and Cognition: Aspects of Hindu Caste and Ritual. Oxford University Press, Delhi.
Deleuze, G., Guattari, F., 1986. Nomadology: The War Machine. Semiotext(e), New York.
Foucault, M., 1986. Of other spaces. Diacritics 16 (1), 22–27.
Geertz, C., 1986. Making experience, authoring selves. In: Turner, V.W., Bruner, E.M. (Eds.), The Anthropology of Experience. University of Illinois Press, Chicago, pp. 373–380.
Gerard, M., 2004. Selecting ritual: DJs, dancers and liminality in underground dance music. In: St. John, G. (Ed.), Rave Culture and Religion. Routledge, London, pp. 167–184.
Graburn, N.H.H., 1989. Tourism: the sacred journey. In: Smith, V. (Ed.), Hosts and Guests: The Anthropology of Tourism, second ed. Blackwell, Oxford, pp. 17–32.
Gunnell, T., 2007. Busar: initiation traditions in Icelandic Gymnasia. In: Midholm, L., Nordstöm, A. (Eds.), The Ritual Year and Ritual Diversity, Proceedings from the International Conference of the SIEF Working Group on the Ritual Year, Gothenburg, Sweden, 7–11 June 2006, pp. 287–297.
Harvey, D., 2000. Spaces of Hope. Edinburgh University Press, Edinburgh.
Hetherington, K., 1997. The Badlands of Modernity: Heterotopia and Social Ordering. Routledge, London.
Hooks, B., 1990. Yearning: Race, Gender and Cultural Politics. South End Press, Boston, MA.
Hopcke, R.H., 1991. On the threshold of change: symbolization and transitional space. In: Schwartz-Salant, N., Stein, M. (Eds.), Liminality and Transitional Phenomena. Chiron, Wilmette, Illinois, pp. 115–132.
Jackson, M., 2005. Existential Anthropology: Events, Exigencies and Effects. Berghahn, Oxford.
Lefebvre, H., 1991. The Production of Space. Oxford, Blackwell.
Lett Jr., J.W., 1983. Ludic and liminoid aspects of charter yacht tourism in the Caribbean. Annals of Tourism Research 10, 35–56.
Macfarlane, R., 2012. The Old Ways: A Journey on Foot. Hamish Hamilton, London.
Marin, L., 1984. Utopics: The Semiological Play of Textual Spaces. Humanities Press International, Atlantic Highlands, NJ.
Meethan, K., 2012. Walking the edges: towards a visual ethnography of beachscapes. In: Andrews, H., Roberts, L. (Eds.), Liminal Landscapes: Travel, Experience and Spaces In-between. Routledge, London, pp. 69–86.
Moore, R.L., 1991. Ritual, sacred space, and healing: the psychoanalyst as ritual elder. In: Schwartz-Salant, N., Stein, M. (Eds.), Liminality and Transitional Phenomena. Chiron, Wilmette, Illinois, pp. 13–32.
Roberts, L., 2002. 'Welcome to Dreamland': from place to non-place and back again in Pawel Pawlikowski's Last Resort. New Cinemas: Journal of Contemporary Film 1 (2), 78–90.
Schwartz-Salant, N., Stein, M. (Eds.), 1991. Liminality and Transitional Phenomena. Chiron, Wilmette, Illinois.
Shields, R., 1991. Places on the Margin: Alternative Geographies of Modernity. Routledge, London.
Smith, N., Katz, C., 1993. Grounding metaphor: towards a spatialised politics. In: Keith, M., Pile, S. (Eds.), Place and the Politics of Identity. Routledge, London.
St. John, G., 2008. Trance tribes and dance vibes: Victor Turner and electronic dance music culture. In: St. John, G. (Ed.), Victor Turner and Contemporary Cultural Performance. Berghahn, New York, pp. 149–173.
Thomassen, B., 2009. The uses and meanings of liminality. International Political Anthropology 2 (1), 5–27.
Thomassen, B., 2012. Revisiting liminality: the danger of empty spaces. In: Andrews, H., Roberts, L. (Eds.), Liminal Landscapes: Travel, Experience and Spaces In-between. Routledge, London, pp. 21–35.
Turner, V.W., 1962. Three symbols of passage in Ndembu circumcision ritual: an interpretation. In: Gluckman, M. (Ed.), Essays on the Ritual of Social Relations. Manchester University Press, Manchester.
Turner, V.W., 1967. The Forest of Symbols, Aspects of Ndembu Ritual. Cornell University Press, New York.
Turner, V.W., 1969. The Ritual Process: Structure and Anti-structure. Cornell University Press, New York.
Turner, V.W., 1982. From Ritual to Theatre: The Human Seriousness of Play. PAJ Publications, New York.
Turner, V.W., Bruner, E. (Eds.), 1986. The Anthropology of Experience. University of Illinois Press, Chicago.
Turner, V., Turner, E., 1978. Image and Pilgrimage in Christian Culture: Anthropological Perspectives. Columbia University Press, New York.
van Gennep, A., 1960. The Rites of Passage. Chicago University Press, Chicago.
Veijola, S., Jokinen, E., 1994. The body in tourism. Theory. Culture and Society 11 (3), 125–151.
Warf, B., Arias, S. (Eds.), 2009. The Spatial Turn: Interdisciplinary Perspectives. Routledge, Abingdon.
Weiss, W.S., 1967. Rebirth in the airborne. Transaction 4 (6), 23–26.

Limits to Growth

Ugo Bardi, Università di Firenze, Firenze, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the findings of the 1972 report titled 'The Limits to Growth' and the debate that followed its publication. This study was one of the first attempts at quantifying how the world's economic system evolves as a function of the availability of natural resources, pollution, and population growth. The general conclusion was that the combination of resource depletion, persistent pollution, and overpopulation was likely to drive the system to economic collapse and then to a population reduction at some moment during the twenty-first century. Collapse could be avoided only by means of drastic changes in the way the world's economic system was run. The reaction to the study was initially favorable but then turned to general disbelief. At present, however, there is clear evidence of a return of interest in the study's methods and findings.

Introduction

The 1972 'Limits to Growth' (LTG) study (Meadows et al., 1972) was little less than a major scientific revolution for its times. Arriving in the general atmosphere of optimism of the years after World War II, it was the first attempt of a quantitative determination of how exactly the limits of the planet would affect the economy. The book carried a somber message: given the assumption of 'business as usual' in the way the economy was run, the collapse of the world's economic system was unavoidable as the result of a combination of resource depletion, persistent pollution, and overpopulation. The model also indicated that the human population would start declining some decades after the collapse of the industrial and agricultural systems. The study had no pretense to be able to determine when exactly would the collapse start but, using the best data available for the amounts of resources available and assuming 'business as usual,' the model indicated that the beginning of the productive decline could be expected to occur within the first two decades of the twenty-first century. The study also indicated that technological solutions that would, for instance, increase the availability of resources, could only postpone the collapse but could not prevent it. Only a radical change in the way the economy was run could stabilize the economic system throughout the twenty-first century.

The LTG study generated remarkable interest and the total number of copies sold all over the world is estimated to be more than one million (Meadows, 2010). Initially, the study generated great interest and a positive response but, in the long run it led to widespread rejection of the study's methods and results. However, there is evidence that the debate had not been exhaustive and that rejection was too hasty (see, e.g., the book 'The Limits to Growth Revisited' (Bardi, 2011). A return of interest in the LTG study is evident and it appears that its results are more and more relevant for the present situation of the world's economy.

Origins of the LTG Study

The 'LTG' was not the first study aimed at evaluating the effect of limited natural resources on the economic system. In 1866, William Stanley Jevons was perhaps the first to examine the effect of mineral depletion on the economy with his 'The Coal Question' (Jevons, 1866). Jevons surmised that the depletion of coal reserves would raise prices and gradually make coal not affordable for the British industry and that, at that point, coal production would start declining. This qualitative interpretation is perfectly compatible with the results of the LTG study, obtained more than a century later. However, in time the mainstream thought in economics moved to a more optimistic vision. A model developed by Harold Hotelling in 1931 (Hotelling, 1931) (known nowadays as 'Hotelling's rule') assumed that a 'backstop resource' would always be available to replace a depleted resource. In 1956, Robert Solow (Solow, 1956) proposed a model – later modified by Stiglitz to include the effect of mineral depletion (Stiglitz, 1974a,b) – that featured technological progress as a parameter that could always keep the economy growing. Models developed by the Austrian School of economics also emphasized technological progress as a way to overcome the depletion problem (see e.g., (Boettke, 1998) for a general description of the school, (Zimmermann, 1933) for an early assessment, (Bradley, 2004), and (Bradley, 2007) for a modern interpretation of the depletion question).

But the ancestry of LTG lies not just in economics; it can be found also in biology with the work of Alfred Lotka (Lotka, 1925) and Vito Volterra (Volterra, 1931) who laid down the foundations of quantitative studies of population dynamics. By revisiting and expanding these methods, Jay Forrester developed in the 1960s the field that became known as 'system dynamics' (Forrester, 1989). These studies were the origin of the LTG model as well as of a world modeling study developed by Forrester himself and published in 1971 with the title 'World Dynamics' (Forrester, 1971). Another study that prefigured some of the concepts later developed in the LTG study was 'The Tragedy of the Commons' by Garrett Hardin (Hardin, 1968). Hardin's study and others that followed showed that natural resources – even renewable ones – tend to be exploited in a free market economy at rates higher than their capability to reform. This tendency (termed 'overshoot' by Catton (Catton, 1982)) leads to the collapse of the production when the capital returns become insufficient to create investments in further exploitation. This concept can be seen as the basis of all the LTG results.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.91047-X

In a different field of knowledge, in 1956, the American geologist Marion King Hubbert developed an empirical model (Hubbert, 1956) that described how fast mineral resources were exploited, proposing a 'bell-shaped' production curve that, later on, became known as the 'Hubbert Curve.' This model was purely empirical, but it was later found that it could be described as a simplified version of a dynamic model (Bardi and Lavacchi, 2009). Another field that can be considered as part of the LTG ancestry is that of pollution studies. The concept had been discussed from the start of the industrial revolution and it began to be considered as a serious problem after the end of World War II, in particular as the result of the great smog crisis in London in 1952 (Davis, 2002). Shortly afterward, the book titled 'Silent Spring' written by the American biologist Rachel Carson (Carson, 1962), was the origin of the modern 'environmental movement.'

The Limits to Growth Model and Its Main Results

The LTG study examined the future by means of a model called 'World3' based on system dynamics, a method of calculation originally developed by Jay Forrester (Forrester, 1989) to describe the evolution of complex systems. The method is based on the concepts of 'stocks' and 'flows.' The first term, stocks, describes quantitative physical entities, and the second, flows, how they change in time. System dynamics models are usually implemented as a set of correlated differential equations that describe how stock sizes vary with time. Feedback effects (that is how flows depend on stock sizes) are a fundamental element of this approach (Forrester, 1994).

The model used for the LTG study contained five main stocks: (1) natural resources, (2) industrial capital, (3) agricultural capital, (4) population, and (5) pollution. The model assumed that the exploitation of mineral resources would generate more industrial capital which, in turn, would lead to more rapid exploitation, resulting in an exponential growth trend. However, the finite amount of mineral resources would gradually reduce the accumulation of capital and, eventually, depreciation would play an increasingly important role in reducing the capital stock. As a consequence, the flow of resources into the economic system would stop growing and eventually decline, generating also the decline of the industrial production. A similar mechanism was assumed to take place for renewable resources, e.g., agriculture; which could be depleted if their exploitation rate was too fast, for instance, depleting the fertile soil. Finally, the effect of growing population on the economy was taken into account assuming that population growth did enhance production but also that it would increase the problems of resource overexploitation and pollution – eventually becoming one of the major factors of the collapse of the economic system. Population growth also depended on the availability of food and services, but the dependency was not simply proportional, but rather modeled according to historical data. These parameters and their reciprocal interaction were defined and adjusted according to the available data and the model was tuned in such a way to be able to reproduce known historical trends.

The 1972 LTG study presented a set of 12 scenarios, which examined the behavior of the world's system according to various parameters in input. A robust feature of these scenarios was that the gradual depletion of mineral resources and the increasing cost of pollution tends to generate the collapse of industrial and agricultural production, followed later on by a decline in the world's population. The so-called 'base case' or 'standard run' scenario, the one that used the data judged to be the most reliable at the time, placed the start of the productive decline for some moment during the first two decades of the twenty-first century (**Figure 1**).

Other scenarios presented in the study took into account the effects of different parameters. Various hypotheses were tested; among these the availability of larger resource stocks, different sensitivity of the system to pollution and the effect of technological breakthroughs. None of these hypotheses could avoid collapse, they could only postpone it. Only a careful set of policies aimed at reducing resource consumption, fighting pollution, and stopping population growth could stabilize the system and maintain a constant level of production and population throughout the twenty-first century. Later versions of the study, (Meadows et al., 1992) (Meadows et al., 2004) updated the results of the 1972 study, but generated basically similar results.

The Scientific Discussion

After publication, in 1972, 'The LTG' was the object of a rigorous scrutiny, as it is normal for a major and innovative study. The main elements of the debate are reported in detail in later works (see, e.g., (Myrtveit, 2005) and (Bardi, 2011)). To summarize, a model that attempts to predict (or at least evaluate) the future can be judged on the basis of different criteria. A first one is based on the model's internal consistence and compatibility with what we know of the world. A second criterion consists in verifying whether the theory's predictions actually turn out to be correct. Obviously, in 1972 there was no way to judge the correctness of the LTG model by comparing it to historical data since its scenarios spanned more than one century in the future. However, a number of attempts were made to evaluate the model in terms of its internal consistency and compatibility with the general knowledge of the time.

The immediate reaction of several economists was to remark that the LTG model was not consistent with the macroeconomical models of the time. As an example, in a 2008 newspaper article, (Krugman, 2008) Paul Krugman describes this reaction in these terms (in this case, referred to the studies by the developer of system dynamics, Jay Forrester, but aimed in general also against LTG) "Forrester, an eminent professor of engineering, decided to try his hand at economics, and basically said, 'I'm going to do economics with equations! And run them on a computer! I'm sure those stupid economists have never thought of that!' And he didn't walk over to the east side of campus to ask whether, in fact, any economists ever had thought of that, and what they had learned." A number of similar reactions are reported in (Bardi, 2011).

However, critics also went more in depth in their examination. Among them, William Nordhaus published a strong criticism of Forrester's 'World Dynamics' (Forrester, 1971) that was also directed against the LTG report. In his paper, titled 'World Dynamics: Measurements without Data,' (Nordhaus,

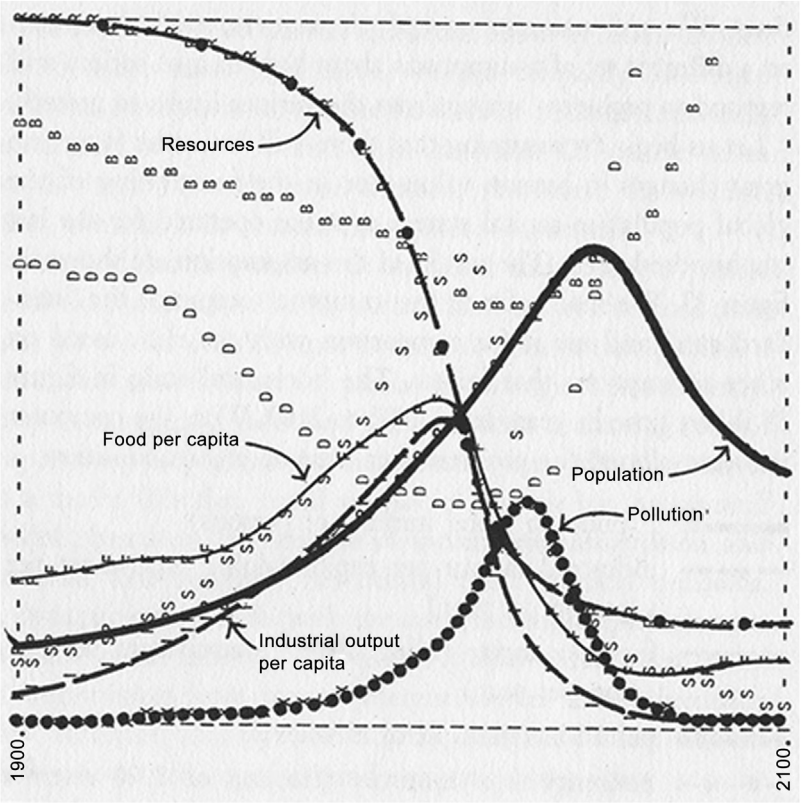

Figure 1 Scenario for the 'base case' assumptions for the 'limits to growth' study published in 1972 Courtesy of Dennis Meadows. This figure has been scanned from the original 1972 book.

1973) Nordhaus claimed that Forrester's models were not based on historical data and attempted to prove that some of the equations of Forrester were unable to describe the behavior of the real world. In his rebuttal (Forrester et al., 1974), Forrester replied that Nordhaus had incorrectly interpreted the model and showed that its equations could indeed reproduce historical data. One of the most in-depth contemporary examination of the 1972 LTG model was performed in the book titled *Models of Doom* (Cole, 1973). The debate is reviewed in detail in (Bardi, 2011) and the author arrives at the conclusion that most of the criticism proposed by the various authors of the book was based on fundamental misunderstandings. Mainly, the criticism was based on the claim that the LTG authors had been too pessimistic in their estimates of the availability of resources and that they had ignored the possibility of technological advances to increase the resource stock. But this criticism fails because the LTG study included scenarios that considered much larger resource stocks than those available according to current data and also the effect of technological breakthroughs that would increase the efficiency of use of the existing stocks.

However, the editors of the 'Models of Doom' book, H.S.D. Cole and R.C. Curnow, also examined in depth the model of the LTG study. Both in the book and in an article published in the same year (Cole and Curnow, 1973a) they emphasized several points that they described as weaknesses of the LTG model. In particular, the inability of the model to 'simulate backwards' ('backcasting') when forced to do so (Cole and Curnow, 1973b). In this case, the LTG authors noted in their rebuttal in the same publication (Cole and Curnow, 1973b) that their model had not be designed for backcasting and that this criticism had little or no relevance to the result of simulations performed the way the model was designed to work. On this controversy, Bardi notes (Bardi, 2011) that the alleged flaws of the LTG model were minor, but that the debate could have led to an improvement of this and other system dynamics models. However, as it will be discussed later on, the scientific debate on the LTG study was soon swamped by a political quarrel that prevented further meaningful discussion.

Several years after the first flaring of this debate, a new attempt of an in-depth analysis of the LTG study was performed by William Nordhaus (Nordhaus, 1992). This study was one of the few (if any) that attempted a serious comparison of models based on system dynamics with the current models used in economics, in particular, with the 'Solow–Stiglitz' model (Stiglitz, 1974 a,b). A detailed discussion of this article and of its claims can be found in (Bardi, 2011) where the author summarizes it as follows: "What Nordhaus was really doing with his 1992 paper was to present a version of the standard neoclassical model of the economy and contrast it to the LTG model. As obvious, the two models have points in common and differ in others. Both models can describe the nearly exponential growth of the economy from the beginning of the twentieth

century, all the way to the present time. Where the two models diverge is, roughly, with the first decades of the twenty-first century. The Solow-Stiglitz model states that the economy will keep growing forever; the LTG model, instead, indicates that at some point it will reach a maximum and then start declining." Obviously, no conclusion on this point could be reached in 1992. At present, however, we are close to the dates for which some of the LTG scenarios had proposed the start of the world's economic decline. During the coming years, it will be possible to tell if these scenarios have been disproved by reality.

After 1972, only a few studies attempted to use the LTG methods for new studies of world modeling. The main exception was 'The Global 2000 report to the president of the U.S.' (Barney, 1982) requested by president Carter which, however, had very little impact. System dynamics continued to be practiced and studied for various applications dealing with physical, business, and social systems (see, for instance, the site of the System Dynamics Society, http://www.systemdynamics.org/) but world modeling tended to decline and slowly disappeared from the scientific horizon. The final blow that condemned the LTG study to oblivion was an article by Ronald Bailey that appeared on Forbes in 1989. In this article, Bailey revisited criticism that had appeared in 1972 on the New York Times (Passel et al., 1972) to accuse the LTG study of having made 'wrong predictions' stating that some mineral commodities (Gold, Mercury, copper, fossil fuels, and others) should have already run out at that time. Later on, Bailey repeated his accusations in the 1993 book titled 'Ecoscam' (Bailey, 1993). As discussed in (Bardi, 2011) these accusations are groundless: there were no such predictions in the LTG book; on the contrary, no mineral resources were supposed to 'run out' before the end of the twenty-first century. Nevertheless, Bailey's accusation had an incredible success. With the Internet era arriving, it went viral and is still popular in the Web today and was stated again, recently, in a 2013 article on 'The Australian' (Burrell, 2013).

The Political Debate

The 'LTG' study was not just an exercise in solving differential equations. It was a study with considerable political relevance. And it was not just the result of the scientific curiosity of a group of researchers, but it had been commissioned by a think tank (The Club of Rome), which had as its stated objective that of influencing governments' policies (a description of the origins and the story of the Club of Rome can be found in the book by Tapio Kanninen 'Crisis of Global Sustainability' (Kanninen, 2013)). The LTG book itself did not take any political position in a traditional sense, but it could not avoid to be interpreted through a political lens. After having presented several possible future scenarios based on 'business as usual' assumptions and concluded that all of them led to economic collapse, the authors proceeded to use their model to find conditions that could avoid economic overshoot and subsequent collapse. These included limiting the exploitation of natural resources, tight controls in industrial emissions, birth control, and more. Of course, these proposals had a strong political relevance. Giorgio Nebbia (Nebbia, 1997) and Mauricio Schoijet (Schoijet, 1999) have examined the issue and they tell us how several groups united in the rejection of the findings of the LTG book, despite their widely different ideological positions. In the Western world, the left saw the study as a threat to the working class, while the right saw it as an attack against free market and liberal policies. Many industrial lobbies saw the study as a direct threat to their business, while in many poor countries, the report was seen as an attempt to perpetuate the dominance of the rich West and maintain the wide gap between rich and poor countries. Finally, religious groups strongly objected to all proposals aimed at limiting birth rates.

To these already formidable political obstacles, we should add the question of how exactly the measures recommended by the LTG study could have been implemented. This fundamental issue was never discussed in the 1972 LTG book and only in the 'Commentary' section, written by the executive committee of the Club of Rome, we read a statement (p. 194) that says "The achievement of a harmonious state of global economic, social, and ecological equilibrium must be a joint venture based on joint conviction, with benefits for all." From this and other statements by the authors of the LTG study and prominent members of the Club of Rome (see, e.g., the book 'Before it is too Late' by A. Peccei and D. Ikeda (Peccei and Ikeda, 2008)), it is clear that the needed changes were seen as generated (and not imposed) by a concerted decision that would involve governments and other world organizations. But the lack of details in these proposals made sure that they would be misunderstood and, indeed, the Club of Rome was accused (and it is still accused) of all sorts of evil plans, including that of working at a program of selective extermination of the 'darker races' (as stated in the 1983 book by Lyndon H. LaRouche 'There are no Limits to Growth' (Larouche, 1983)). This kind of accusation even spilled to the scientific debate. For instance, in 1977, an article appeared on 'Social Studies of Science' (Golub and Townsend, 1977) where the Club of Rome and the authors of the LTG report were accused to be part of a global conspiracy aimed at creating 'a kind of fanatical military dictatorship.' If this statement could be read in a peer-reviewed international journal, one can imagine what could be read in the press at that time and, in later years, in the Web.

These considerations raise the question of whether the outcry that condemned the LTG study was spontaneous or orchestrated by special interest groups. The question is legitimate because modern analysis has discovered several cases during the twentieth and twenty-first centuries where smear campaigns were created and directed by industrial lobbies to protect their economic interests. The best known case is that of the debate on the health effects of tobacco smoking. In this case, the documents that the tobacco industry released as part of the 1998 'Master Settlement Agreement' clearly show that such a campaign was organized against by the tobacco lobby against science and scientific researchers. The story is told in detail, for instance, in two books focused on of the tobacco debate: 'Doubt is their Product' (Michaels, 2008) and 'Bending Science' (McGarity and Wagner, 2008). The same methods appear to have been used unchanged in other cases even today, for instance, against climate science and climate scientists. Details on the ongoing assault on science can be found in several recent books, including 'Merchants of Doubt,' (Oreskes

and Conway, 2010) 'Climate Cover-Up,' (Hoggan, 2009) and 'Climate Wars' (Mann, 2012). For the LTG case, we have no equivalent documentation. We can, however, note that the smear campaign against the study was based on the same methods and we may at least suspect that special interest groups played a role in it.

The Rebirth

In the 1990s, the rejection of 'The Limits to Growth' could be considered complete. The study and its methods seemed to have completely lost interest by universal agreement and to have been consigned to the dustbin of wrong science. However, with the new century, the economic crisis and the political unrest led to a gradual reconsideration of the study that went in parallel with research on the depletion of a specific resource: crude oil. This subject had been studied in depth already in the 1950 by Marion King Hubbert (Hubbert, 1956) and these early studies gave rise to a new assessment of the world's oil resources by Jean Laherrere and Colin Campbell. In their article published in 1998 (Campbell and Laherrere, 1998), they concluded that the world's conventional oil production could have reached a peak and then declined during the first decade of the twenty-first century. This global peak would later become a popular concept with the term 'peak oil' coined by Colin Campbell. It may have been this line of thought that led an oil expert, Matthew Simmons, to write the first paper of twenty-first century, which favorably reviewed the LTG study (Simmons, 2000). Several other authors favorably reexamined the 1972 LTG study (Bardi, 2011), (Myrtveit, 2005), (Gardner, 2004) (Hall and Day, 2009), (NØrgaard et al., 2010), (Heinberg, 2010). In particular, in 2008, Turner (Turner, 2008) found that the 'base case' scenario of LTG has described remarkably well the behavior of parameters such as population growth, resource depletion, and pollution during the past 30 years.

These results are in stark contrast with the negative opinion of the LTG study, which is still widespread today. The debate has been ongoing and the negative opinion on the LTG study remains probably prevalent. Nevertheless, it appears that the time is past when it was considered that the study could not even be mentioned, except to say that it was wrong. The debate is not likely to be solved very soon but, looking back at the more than 40 years that passed from the publication of the first LTG book, we can at least say that it was not so 'wrong' as many had assumed it to be.

More than all, we can say that LTG was a warning, not a prediction and – as such – it could not be really 'wrong.' Clearly, predicting the future is difficult and mistakes are always possible. But, the worst possible mistake that we can make with the future is to ignore it or to assume the recent trajectories always continue unabated. The problem of depletion that the LTG study had foreseen for the first decades of the twenty-first century seems to be here in the form of high prices of all natural resources. At the same time, the climate problem seems to be growing more and more intractable. It may be that we will soon have a chance to test whether the LTG scenarios of 1972 had really foreseen the trajectory of the economy of the real world as it has been developing during the past 40 years.

See also: Ecological Economics; Economic Growth; Environment and Development; Environmental Economics; Environmental History; Environmental Sciences; Environmental Sociology; Mineral Resources, Limits to: The Case of Peak Oil; Population Dynamics: Mathematic Models of Population, Development, and Natural Resources; Resource Geography; Water Resources and Sustainable Water Management.

Bibliography

Bailey, R., 1989. Dr doom. Forbes Magazine, 45.

Bailey, R., 1993. Eco-scam: The False Prophets of Ecological Apocalypse. St. Martin Press, New York. Available at: http://www.osti.gov/energycitations/product.biblio.jsp?osti_id=6242018 (accessed 01.01.14.).

Bardi, U., 2011. The Limits to Growth Revisited. Springer, New York. Available at: http://books.google.com/books?hl=en&lr=&id=Gj0RmQWGfXcC&pgis=1 (accessed 02.12.12.).

Bardi, U., Lavacchi, A., 2009. A simple interpretation of Hubbert's model of resource exploitation. Energies 2 (3), 646–661. Available at: http://www.mdpi.com/1996-1073/2/3/646 (accessed 03.12.12.).

Barney, G., 1982. The Global 2000 Report to the President of the US. Pergamon Press, New York. Available at: http://lavoixoff.info/lavoixofffr/politique/GLOBAL2000REPORT.pdf (accessed 02.01.14.).

Boettke, P., 1998. Elgar Companion to Austrian Economics. Edward Elgar Publishing Limited, Cheltenham, UK and Northampton, MA. Available at: http://books.google.it/books?hl=en&lr=&id=5Zjzb9cDsMIC&oi=fnd&pg=PR1&dq=Boettke&ots=hOJSzFRUyW&sig=G5uncROFtmmn5c2e9YOSpxPltOo (accessed 04.01.14.).

Bradley, R.L., 2007. Resourceship: an Austrian theory of mineral resources. The Review of Austrian Economics 20 (1), 63–90. Available at: http://link.springer.com/10.1007/s11138-006-0008-7 (accessed 26.09.13.).

Bradley, R.L.J., 2004. PERC Report: Are We Running Out of Oil?. Available at: http://perc.org/articles/are-we-running-out-oil (accessed 23.08.13.).

Burrell, A., 2013. IPCC This Century's "Chicken Little". The Australian. Available at: http://www.theaustralian.com.au/national-affairs/policy/ipcc-this-centurys-chicken-little/story-e6frg6xf-1226752383735#.

Campbell, C.J., Laherrere, J.F., March 1998. The end of cheap oil. Scientific American, 80–86.

Carson, R., 1962. Silent Spring. Honghton Mifflin Co, Boston, MA. Available at: http://scholar.google.it/scholar?start=20&q=carson+silent+spring&hl=en&as_sdt=0,5#1 (accessed 10.12.13.).

Catton, W., 1982. Overshoot, the Ecological Basis of Revolutionary Change. Illinin Books Edition, Chicago, IL. Available at: http://books.google.it/books?hl=en&lr=&id=jCKXpv-E5HsC&oi=fnd&pg=PA1&dq=overshoot+catton&ots=X-5InGH1_l&sig=OcUxeV5cYAkO4Vzbj5_VAOuV_GI (accessed 10.12.13.).

Cole, H., Curnow, R., 1973a. An evaluation of the world models. Futures 5 (1), 108–134. Available at: http://www.sciencedirect.com/science/article/pii/001632877390061X (accessed 01.01.14.).

Cole, H., Curnow, R., May 11, 1973b. "Backcasting" with the world dynamics models. Nature 243, 63–65. Available at: http://www.nature.com/nature/journal/v243/n5402/abs/243063a0.html (accessed 01.01.14.).

Cole, H.S.D., 1973. Models of Doom: A Critique of the Limits to Growth. Universe Books, New York.

Davis, D., 2002. A look back at the London smog of 1952 and the half century since. Environmental Health Perspectives 110 (2), A734–A735. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1241116/ (accessed 03.01.14.).

Forrester, J., 1994. System dynamics, systems thinking, and soft OR. System Dynamics Review 10 (2–3), 245–256. Available at: http://onlinelibrary.wiley.com/doi/10.1002/sdr.4260100211/abstract (accessed 04.01.14).

Forrester, J., 1989. The beginning of system dynamics Banquet talk at the international meeting of the system dynamics society. McKinsey Quarterly. Available at: http://leml.asu.edu/jingle/Web_Pages/EcoMod_Website/Readings/SD+STELLA/Forrester-Begin'g-SD_1989.pdf (accessed 03.06.13.).

Forrester, J., 1971. World Dynamics. Wright-Allen Press, Cambridge, MA. Available at: http://documents.irevues.inist.fr/handle/2042/29441 (accessed 03.06.13.).

Forrester, J.W., Low, G.W., Mass, N.J., 1974. The debate on "world dynamics": a response to nordhaus. Policy Sciences 5 (2), 169–190.

Gardner, T., 2004. Limits to growth?–A perspective on the perpetual debate. Environmental Sciences 1 (2), 121–138. Available at: http://www.tandfonline.com/doi/abs/10.1080/15693430512331342592 (accessed 02.01.14.).

Golub, R., Townsend, J., 1977. Malthus, multinationals and the club of Rome. Social Studies of Science 7, 201–222.

Hall, C., Day, J., 2009. Revisiting the limits to growth after peak oil in the 1970s. American Scientist. Available at: http://www.americanscientist.org/issues/num2/2009/3/revisiting-the-limits-to-growth-after-peak-oil/1 (accessed 02.01.14.).

Hardin, G., 1968. The tragedy of the commons. Science 162 (13 December), 1243–1248. Available at: http://scholar.google.it/scholar?hl=en&q=garrett+hardin&btnG=&as_sdt=1,5&as_sdtp=#0 (accessed 10.12.13.).

Heinberg, R., 2010. Beyond the Limits to Growth. Post Carbon Institute, Founation series. Available at: http://files.uniteddiversity.com/Measuring_Progress_and_Eco_Footprinting/PCReader-Heinberg-Beyond-the-Limits-to-Growth.pdf (accessed 02.01.14.).

Hoggan, J., 2009. Climate Cover-up: The Crusade to Deny Global Warming. Greystone Books Ltd, Vancouver, Toronto, Berkeley. Available at: http://books.google.com/books?id=tQYjQzOkYK0C&pgis=1 (accessed 12.06.13.).

Hotelling, H., 1931. The economics of Eshaustible resources. Journal of Political Economy 39 (2), 137–175.

Hubbert, M.K., 1956. Nuclear energy and the fossil fuels. In: Spring Meeting of the Southern District Division of Production American Petroleum Institute. American Petroleum Institute, Sant Antonio, TX. Available at: energycrisis.biz/hubbert/1956/1956.pdf (accessed 03.12.12.).

Jevons, W.S., 1866. The Coal Question, second revised ed. Macmillan and Co. Available at: http://www.econlib.org/library/YPDBooks/Jevons/jvnCQ.html.

Kanninen, T., 2013. In: Weiss, T.G. (Ed.), Crisis of Global Sustainability. Routledge, London and New York. Available at: http://www.routledge.com/books/details/9780415694179/.

Krugman, P., 2008. Limits to growth and related stuff. The New York Times. Available at: http://krugman.blogs.nytimes.com/2008/04/22/limits-to-growth-and-related-stuff/?_r=0.

Larouche, L.H., 1983. There Are No Limits to Growth. The New Benjamin Franklin House, New York.

Lotka, A.J., 1925. Elements of Physical Biology. Williams and Wilkins, Baltimore.

Mann, M., 2012. The Hockey Stick and the Climate Wars. Columbia University Press, New York. Available at: http://books.google.it/books?hl=en&lr=&id=HK0CN6FVtfgC&oi=fnd&pg=PR7&dq=climate+wars&ots=ehA4dS_FKt&sig=y5-a1VG_UQ6Xi2Zbp8g2_rWknCY (accessed 31.12.13.).

McGarity, T., Wagner, W.E., 2008. Bending Science: How Special Interests Corrupt Public Health Research. Harvard University Press, Cambridge, MA and London, England. Available at: http://books.google.it/books?hl=en&lr=&id=Ah6-__otORAC&oi=fnd&pg=PR7&dq=McGArity+Wagner&ots=TnfuhvQbrt&sig=WC-X7xkCkOjBxru6sTQGrrFHkgw (accessed 03.01.14.).

Meadows, D.H., et al., 1972. The Limits to Growth. Universe Books, New York.

Meadows, D.H., Meadows, D.L., Randers, J., 1992. Beyond the Limits, Post Mills. Chelsea Green, Vermont, USA.

Meadows, D.H., Randers, J., Meadows, D.L., 2004. Limits to Growth: The 30 Year Update. Chelsea Green, White River Junction.

Meadows, D.L., 2010. Private Communication.

Michaels, D., 2008. Doubt Is Their Product: How Industry's Assault on Science Threatens Your Health. Oxford University Press, New York. Available at: http://books.google.it/books?hl=en&lr=&id=Qedeu_5_lEEC&oi=fnd&pg=PA13&dq=doubt+is+their+product&ots=d8KbENql4j&sig=WDcvaozCsQCZjTR54wZ5LqUalyQ (accessed 31.12.13).

Myrtveit, M., 2005. Working Papers in System Dynamics. Available at: folk.uib.no/sinem/WPSD/WPSD1.05WorldControversy.pdf.

Nebbia, G., 1997. Bisogno di storia e di futuro. Futuribili, New Series, Gorizia, vol. 4, pp. 149–182.

Nordhaus, W., 1992. Lethal Models. Brookings Papers on Economic Activity pp. 1–59. Available at: http://scholar.google.it/scholar?hl=en&q=nordhaus+lethal+models&btnG=&as_sdt=1,5&as_sdtp=#1 (accessed 03.06.13.).

Nordhaus, W.D., 1973. World dynamics: measurement without data. The Economic Journal 83 (332), 1156–1183.

NØrgaard, J., Peet, J., Ragnarsdottir, V., 2010. The history of the limits to growth. Solutions 1 (2), 59–63. Available at: http://www.thesolutionsjournal.com/node/569 (accessed 02.01.14.).

Oreskes, N., Conway, E.M., 2010. Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming. Bloomsbury Publishing USA, New York. Available at: http://books.google.com/books?id=fpMh3nh3JI0C&pgis=1 (accessed 12.06.13.).

Passel, P., Roberts, M., Ross, L., 1972. Review of "the limits to growth". New York Times Book Review.

Peccei, A., Ikeda, D., 2008. In: Gage, R.L. (Ed.), Before it Is Too Late. I.B. Tauris, London and New York.

Schoijet, M., October 4, 1999. Limits to growth and the rise of Catastrophism. Environmental History 4, 515–530. Available at: http://www.jstor.org/stable/10.2307/3985399 (accessed 31.12.13).

Simmons, M., 2000. Revisiting the Limits to Growth: Could the Club of Rome Have Been Correct, after All? an Energy White Paper. Available at: http://www.greatchange.org/ov-simmons,club_of_rome_revisted.html (accessed 01.01.14.).

Solow, R., 1956. Technical change and the aggregate production function. Quarterly Journal of Economics 70 (1), 65–94.

Stiglitz, J., 1974a. Growth with exhaustible natural resources: efficient and optimal growth paths. The Review of Economic Studies 41 (1), 123–137. Available at: http://www.jstor.org/stable/10.2307/2296377 (accessed 24.08.13.).

Stiglitz, J., 1974b. Growth with exhaustible natural resources: the competitive economy. The Review of Economic Studies 41, 139–152. Available at: http://www.jstor.org/stable/10.2307/2296378 (accessed 24.08.13.).

Turner, G., 2008. A comparison of the limits to growth with 30 years of reality. Global Environmental Change 18 (3), 397–411. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0959378008000435 (accessed 06.11.12.).

Volterra, V., 1931. Variations and fluctuations of the number of individuals in animal species living together. In: Chapman, R.N. (Ed.), Animal Ecology. McGraw-Hill.

Zimmermann, E.W., 1933. World Resources and Industries: A Functional Appraisal of the Availability of Agricultural and Industrial Resources. Harper & Brothers, New York. Available at: http://www.amazon.com/World-resources-industries-availability-agricultural/dp/B000855EA8 (accessed 01.01.14.).

Linear Hypothesis

Ronald Christensen, University of New Mexico, Albuquerque, NM, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 13, pp. 88747–88881, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

The term 'linear hypothesis' is often used interchangeably with the term 'linear model.' Statistical methods using linear models are widely used in the behavioral and social sciences, e.g., regression analysis, analysis of variance, analysis of covariance, multivariate analysis, time series analysis, and spatial data analysis. Linear models provide a flexible tool for data analysis and useful approximations for more complex models. This article defines linear models, introduces the commonly used matrix notation for linear models, discusses estimation, testing, and model checking, and briefly looks at models for dependent data. It also mentions how the linear hypothesis is used in generalized linear models.

Introduction

The term 'linear hypothesis' is often used interchangeably with the term 'linear model.' Statistical methods using linear models are widely used in the behavioral and social sciences, e.g., regression analysis, analysis of variance, analysis of covariance, multivariate analysis, time series analysis, and spatial data analysis. Linear models provide a flexible tool for data analysis and useful approximations for more complex models.

A common object of linear modeling is to find the most precise linear model that explains the data, to use that model to predict future observations, and to interpret that model in the context of the data collection. Traditionally, analysis of variance models have been used to analyze data from designed experiments while regression analysis has been used to analyze data from observational studies but the techniques of both analysis methods apply to both kinds of data. See also Experimental Design: Overview and Observational Studies: Overview.

Definition

The linear hypothesis is that the mean (average) of a random observation can be written as a linear combination of some observed predictor variables. For example, Coleman et al. (1996) provides observations on various schools. The dependent variable y consists of the average verbal test score for sixth-grade students. The report also presents predictor variables. A composite measure of socioeconomic status x_1 is based on fathers' and mothers' education, family size and intactness, home items, and percent of fathers who are white collar. Staff salaries per pupil is x_2. The average score on a verbal test given to the school's teachers is x_3. Denoting different schools using the subscript i and the mean of y_i by m_i, a linear hypothesis states that for some unknown numbers (parameters) β_0, β_1, β_2, and β_3,

$$m_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \beta_3 x_{i3}$$

Mosteller and Tukey (1977, pp. 326, 566) and Mosteller et al. (1983, pp. 408–20) give excerpts and analysis of the data.

In other applications, the predictors only identify whether an observation is in some group. For example, in 1978 observations y_{ij} were collected on the age at which people in Albuquerque committed suicide, see Koopmans (1987, p. 409). Here i is used to identify the person's group membership (Hispanic, Native American, non-Hispanic Caucasian), and j identifies individuals within a group. The three categories are taken to be mutually exclusive for the present discussion (although the US government now allows for individuals to identify themselves with multiple races in various surveys and the decennial census). We can define group identifier predictor variables. Let δ_{1i} take the value 1 if an individual belongs to group 1 (Hispanic) and 0 otherwise, with similar predictors to identify other groups, say δ_{2i} and δ_{3i} for Native Americans and non-Hispanic Caucasians. Note that the predictor variables do not depend on the value of j identifying individuals within a group. Denoting the mean of y_{ij} by m_{ij}, a linear hypothesis states that for some unknown parameters μ, α_1, α_2, α_3,

$$m_{ij} = \mu + \alpha_1\delta_{1i} + \alpha_2\delta_{2i} + \alpha_3\delta_{3i}$$

Since two of the δs are always zero, this model is often written more succinctly as

$$m_{ij} = \mu + \alpha_i$$

A linear hypothesis is usually combined with other assumptions about the observations y. Most commonly, the assumptions are that the observations are independent, have the same (unknown) variance σ^2, and have normal (Gaussian) distributions. For the two examples, these assumptions are written

$$y_i \text{ indep. } N(m_i, \sigma^2) \text{ and } Y_{ij} \text{ indep. } N(m_{ij}, \sigma^2)$$

where for example, $N(m_i, \sigma^2)$ indicates a normal distribution with mean m_i and variance σ^2. Incorporating these additional assumptions, the linear hypothesis becomes a full-fledged linear model, traditionally written

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \varepsilon_i \text{ indep. } N(0, \sigma^2) \quad [1]$$

and

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \varepsilon_{ij} \text{ indep. } N(0, \sigma^2) \quad [2]$$

where the εs are unobservable random errors.

Both theoretical and computational work with linear hypotheses and linear models is facilitated by the use of

matrices. If $i = 1, 2, 3, 4$, the model (1) can be written as the equality of two 4×1 matrices

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \varepsilon_1 \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{33} + \varepsilon_2 \\ \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \beta_3 x_{33} + \varepsilon_3 \\ \beta_0 + \beta_1 x_{41} + \beta_2 x_{42} + \beta_3 x_{43} + \varepsilon_4 \end{bmatrix}$$

Using properties of matrix algebra, this can be rewritten as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 x_{11} & x_{12} & x_{13} \\ 1 x_{21} & x_{22} & x_{33} \\ 1 x_{31} & x_{32} & x_{33} \\ 1 x_{41} & x_{42} & x_{43} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

Similarly, model (6) for $i = 1, 2, 3$, j=1, 2 can be written using δ as

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

In general, a linear model is traditionally written using matrices as

$$Y = X\beta + e \qquad [3]$$

where Y is a $n \times 1$ matrix containing the observable random variables y, X is a $n \times p$ matrix containing observed predictors, β is a $p \times 1$ matrix containing unobservable parameters, and e is a $n \times 1$ matrix containing unobservable random errors. To have a full-fledged model, assumptions must be made about the distribution of e. One standard assumption is that the mean value of each element of e is 0. Denote the $n\times1$ matrix of mean (expected) values as E(e). This assumption is written

$$E(e) = 0$$

The corresponding linear hypothesis involves the mean values of the entries of Y, i.e., E(Y). The linear hypothesis is that

$$E(Y) = X\beta$$

for some matrix β. Another standard assumption about e is that the covariance matrix, the $n \times n$ matrix of variances and covariances between the ε_is or equivalently between the y_is is

$$Cov(e) = Cov(Y) = \sigma^2 I$$

where I indicates the $n \times n$ identity matrix that has 1s on the diagonal and 0 elsewhere.

An alternative way to write model (3) in terms of each y_i using some matrices is

$$y_i = x_i'\beta + \varepsilon_i$$

$i = 1, \ldots, n$ where β is as before and x_i' is the ith row of the X matrix. See also Analysis of Variance and Generalized Linear Models and Linear Hypothesis: Regression (Basics).

Estimation

The unknown parameters of the linear model are the elements of β, the βs, and the variance of an observation, σ^2; they must be estimated from the data. The β_js are generally estimated using least squares estimation. For given data Y, the least squares estimates are the values of β_j that minimize

$$\sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 \text{ or equivalently } (Y - X\beta)'(Y - X\beta)$$

where the prime on $(Y–X\beta)'$ indicates the matrix transpose of $(Y–X\beta)$. (The transpose of a matrix turns the columns of the original matrix into the rows of a new matrix.)

Under the standard assumptions, least squares estimates have optimal statistical properties. If the ε_is are independent $N(0, \sigma^2)$, then the least squares estimates provide maximum likelihood estimates (MLEs) and minimum variance unbiased estimates (MVUEs). A MVUE is an estimate that has less variability than any other unbiased estimate. If we drop the assumption of normality but retain independence (actually 0 covariances) and equal variances, least squares estimates are best linear unbiased estimates (BLUEs). A BLUE is an estimate based on a linear combination of the observations that has less variability than any other unbiased estimate based on a linear combination of the observations.

In matrix terms, the least squares estimates, say, β, are obtained as solutions to the normal equations

$$X'X\beta = X'Y$$

Typically, in regression problems, the unique solution is

$$\widehat{\beta} = (X'X)^{-1Y}X'$$

In general, if $X'X$ is singular, so that $(X'X)^{-1}$ does not exist, an infinite number of solutions exist, giving an infinite number of least squares estimates. Any one of these estimates is just as valid as any other, but fortunately, for all practical purposes, the estimates are unique. This means that whenever a linear parameter $\lambda'\beta$ is identifiable in the model (estimable), the least squares estimate of the parameter is uniquely defined. See also Statistical Identification and Estimability and Analysis of Variance and Generalized Linear Models.

The other unknown parameter to be estimated is σ^2, the variance of an observation. To estimate the variance, first rearrange the model into

$$\varepsilon_i = y_i - x_i'\beta \text{ or } e = Y - X\beta$$

Then estimate β to get the residuals

$$\widehat{\varepsilon}_i = y_i - x_i'\widehat{\beta} \text{ or } \widehat{e} = Y - X\widehat{\beta}$$

the sum of the squared residuals is called the sum of squares error,

$$SSE = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \left(Y - X\widehat{\beta}\right)'\left(Y - X\widehat{\beta}\right)$$

The degrees of freedom error (dfE) is n minus the rank of the matrix X, i.e.,

$$dfE = n - r(X)$$

The rank of X is the number of free parameters that can be fitted to the model by least squares. The variance σ^2 is estimated by the mean squared error,

$$MSE = \frac{SSF}{dfE}$$

See also Estimation: Point and Interval.

Testing

The basic idea of testing is to create some model for the process of generating the data, and evaluate whether the observed data are consistent with that model. For example, such a model might be

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

One can think of this model has having four parts: (a) ε_i are independent, (b) ε_i have E(ε_i)=0, i.e.,

$$E(y_i) = m_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

(c) ε_i have common variance σ^2, (d) ε_i are normally distributed. Data that are inconsistent with the model may occur whenever any of these four parts is inappropriate.

In the context of testing the linear hypothesis, it is assumed that these four specifications are correct and a further refinement of the model is then imposed by placing an additional restriction on

$$E(y_i) = m_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

The test is usually thought of as a test of this additional restriction, generally called the null hypothesis. For example, the additonal restriction could be the null hypothesis H_0: $\beta_1 = \beta_3 = 0$. In the sixth grade verbal test example, this hypothesis indicates that the variables x_1, socioeconomic status, and x_3, teachers mean verbal test score, have no ability to help predict sixth graders' mean verbal test scores. The test of whether the data are consistent with a model that incorporates $\beta_1 = \beta_3 = 0$ is based on assuming that the null hypothesis and items (a) through (d) are all correct. If assumptions (a) through (d) are correct, then data that are inconsistent with the model suggest that either $\beta_1 \neq 0$ or βor both, i.e., that the null hypothesis is not true. However, in reality, data that are inconsistent with the model can result if any of the assumptions are violated, i.e., any of the four itemized assumptions or the null hypothesis. Because of this, if one is interested in testing only the null hypothesis, it is crucial to do everything possible to validate the assumptions embodied in items (a) through (d). See also Hypothesis Testing in Statistics and Significance, Tests of.

There are two ways to think about testing a linear hypothesis. One is to think in terms of testing models and the other is to think in terms of testing parameters.

Models

In testing models one assumes the validity of a (full) model and tests to see if some smaller (reduced) model is consistent with the data. For example, with the verbal test score data a full model might be

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

incorporating all three predictors. The reduced model may incorporate the idea that the socioeconomic and mean teacher verbal score variables x_{i1} and x_{i3} are not important, yielding

$$y_i = \beta_0 + \beta_2 x_{i2} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

In this context, the null hypothesis is that the reduced model provides an adequate explanation for the data. In any particular test, the full model need not incorporate all of the variables that have been measured. It need only be a model that is plausibly correct.

Both the full model and the reduced model have sums of squares error, degrees of freedom error, and mean squared errors associated with them. The reduced model is rejected as being inconsistent with the data if the following F statistic is too large,

$$F = \frac{SSR(\text{Red}) - SSE(Full)}{dfE(\text{Red}) - dfE(Full)} / MSE(Full)$$

If the reduced model is true, then both the full model and the reduced model are true, so both MSE(Red) and MSE(Full) are estimates of σ^2. It can then be shown that (SSE(Red)-SSE(Full))/(dfE(Red)-dfE(Full)) is also an estimate of σ^2, so both the numerator and the denominator of F are estimates of σ^2, and the ratio should be about 1. Of course, there is sampling variability associated with the F statistic. The theoretical F distribution describes the usual distribution of values one will encounter when making such calculations. It depends on two parameters: dfE(Red)-dfE(Full) degrees of freedom for the numerator and dfE(Full) degrees of freedom for the denominator.

On the other hand, if the reduced model is an inadequate explanation of the data, it can be shown that (SSE(Red)-SSE(Full))/(dfE(Red)-dfE(Full)) estimates σ^2 plus a positive number, so that if the reduced model is bad, the F statistic should be larger than 1. If the observed value of the F statistic is so much larger than 1 as to be inconsistent with the theoretical F distribution, one can reasonably conclude that the reduced model is an inadequate explanation of the data, i.e., one rejects the null hypothesis.

Parameters

The other way to perform tests in the linear hypothesis is to specify a null hypothesis in terms of the parameters. For example, if we assume the verbal test score model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

we might be interested in testing the null hypothesis H_0: $\beta_1 = \beta_3 = 0$ in which socioeconomic status and teachers mean verbal test score have no ability to predict. One way to test this hypothesis is to identify the reduced model associated with it, i.e.,

$$y_i = \beta_0 + 0x_{i1} + \beta_2 x_{i2} + 0x_{i3} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

Or

$$y_i = \beta_0 + \beta_2 x_{i2} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0, \sigma^2)$$

For a more complicated example, suppose two previous studies had suggested the hypothesis H_0: $\beta_1 + \beta_3 = 5$ and $\beta_2 = 0$.

We can rewrite the hypothesis as H_0: β_1=5-β_3 and β_2=0, and create a reduced model by substitution,

$$y_i = \beta_0 + (5-\beta_3)x_{i1} + 0x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i, \text{indep. } N(0,\sigma^2)$$

Rearranging terms gives a reduced model

$$y_i - 5x_{i1} = \beta_0 + \beta_3(x_{i3} - x_{i1}) + \varepsilon_i, \quad \varepsilon_i \text{ indep. } N(0,\sigma^2)$$

This is just a new linear model with a new dependent variable y_i−5x_{i1} and one predictor variable (x_{i3}−x_{i1}). One can compute *SSE*(Red) and *dfE*(Red) in the usual way for this new model, and simply substitute these quantities into the usual formula for the *F* statistic to get the test.

If the data are consistent with the reduced model, we have a more precise model than the one we started with. While testing provides no assurance that the reduced model is correct, the data are at least consistent with it. This more precise model can be investigated for the validity of its predictions and its usefulness in explaining the data collection process.

In addition, (estimable) parametric hypotheses can be tested directly using matrices, without identifying the reduced model. For example, the hypothesis H_0: β_1+β_3 = 5 and β_2 = 0 can be written in matrix form by creating a matrix Λ' that, when multiplied by β, isolates β_1+β_3 and β_2. The hypothesis is

$$\wedge'\beta \equiv \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} \beta_1 + \beta_3 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \end{bmatrix} \equiv d$$

In matrix terms the *F* statistic for a null hypothesis H_0 : $\Lambda'\beta; = d$ is

$$F = \frac{\left(\wedge'\widehat{\beta} - d\right)'\left[\wedge'(X'X)^{-}\wedge\right]^{-}\left(\wedge'\widehat{\beta} - d\right)}{r(\wedge)MSE} \qquad [4]$$

where β is a least squares estimate of β, $r(\Lambda)$ is the rank of the matrix Λ, and A^{-} is a generalized inverse of the matrix *A*. If A^{-} exists, it is the unique generalized inverse. The reduced model test and this direct test give the same *F* statistic and use the same degrees of freedom.

Confidence regions can be constructed using the *F* distribution. For the example, we can construct a simultaneous confidence region for $\beta_1 + \beta_3$ and β_2 as follows. Let $F(0.95, r(\Lambda), n–r(X))$ be the number so that with probability 0.95

$$F = \frac{\left(\wedge'\widehat{\beta} - \wedge'\beta\right)'\left[\wedge'(X'X)^{-}\wedge\right]^{-}\left(\wedge'\widehat{\beta} - \wedge'\beta\right)}{r(\wedge)MSE} \leq F(0.95, r(\wedge), n - r(X))$$

Here, $MSE(X'X)^{-}$ is an estimate of the covariance matrix of β and $MSE[\Lambda'(X'X)^{-}\Lambda]$ is an estimate of the covariance matrix of $\Lambda'\beta$. After substituting the observed values of $\Lambda'\beta$ and *MSE* into the inequality, the only thing that is unknown in this inequality is the value of $\Lambda'\beta$. A confidence region for $\Lambda'\beta$ consists of all the values that can be substituted for $\Lambda'\beta$ while maintaining the truth of the inequality. A precise definition of the 95 percent confidence region is as the collection of all values *d* in (4) that would not be rejected by an $\alpha = 0.05$ level *F* test. If Λ' has only one row, the confidence region simplifies to a confidence interval. See also Estimation: Point and Interval.

Bayesian Estimation

An alternative to least squares estimation is Bayesian estimation. In Bayesian estimation the parameters of the linear model are considered to be random and probability distributions reflecting the investigator's prior knowledge about the parameters are specified. Using Bayes theorem, the prior probability distributions are updated into posterior distributions that reflect the information embodied in the data. All statistical inferences are then based on the posterior distributions. Sometimes, a noninformative (improper) prior is used in which case, interval estimates arrived at using Bayesian methods are numerically equal to the interval estimates obtained from classical methods, only the interpretations of the intervals change. Alternatively, real prior information is often specified by taking the prior distribution of β given σ^2 to be a multivariate normal distribution and the prior distribution of σ^2 to be an inverse gamma distribution. This leads to modified least squares estimates that are shrunk towards the mean of the prior distribution β with corresponding changes to interval estimation. See also Bayesian Statistics; Distributions, Statistical: Special and Continuous; Elicitation of Probabilities and Probability Distributions.

Model Checking

Valid testing of a null hypothesis requires that all of the four itemized assumptions embodied in the model be true. Confidence intervals and regions are collections of parameter values that cannot be rejected by tests, so they also require the validity of the basic assumptions in the model as do Bayesian and non-Bayesian estimation. Model checking is considered an integral part of all statistical inference. A variety of methods have been developed for model checking. Most are based on plotting the residuals, see Linear Hypothesis: Regression (Graphics). These methods include checking the normality assumption by plotting the ordered (standardized (Studentized)) residuals against (some approximation to) the expected order statistics of a standard normal distribution. Such a plot should be approximately linear if the data are normal.

The (standardized (Studentized)) residuals are also plotted against their corresponding predicted variables or any other variable (that does not depend on y) for which a measurement is available for each observation. In these plots, any identifiable systematic structure is an indication of problems with the assumptions.

Heteroscedasticity refers to having observations with different variances. In some models, formal tests for equal variances are available, e.g., Bartlett's test and the model-based tests discussed in Carroll and Ruppert (1988).

Lack of fit refers to the problem of having specified an incorrect linear mean structure, i.e., an incorrect linear hypothesis. Formal tests for lack of fit are often based on

identifying clusters of observations that have similar predictor variables, i.e., similar rows of the X matrix, see Christensen (1996a, Sect. 6.6). Miller et al. (1998) have provided optimal methods for grouping observations into near replicate clusters.

The independence assumption is perhaps the hardest to evaluate. If the observations are taken at equally spaced time intervals, standard time series methods are often applied to the residuals. More generally, Christensen and Bedrick (1997) proposed creating rational subgroups (clusters) of the observations in which one suspects the observations may be more similar within clusters than between clusters. The methods of near replicate lack of fit tests can then be applied to test the independence assumption.

Weighted Models and Covariance Estimation

The standard assumption in linear models is that the observations all have the same variance and that the observations are all independent. If either or both of those assumptions is invalidated, different methods of estimation and testing are required. The usual linear model has

$$Y = X\beta + e, \quad E(e) = 0, \quad Cov(e) = \sigma^2 I$$

Writing a model

$$Y = X\beta + e, \quad E(e) = 0, \quad Cov(e) = \sigma^2 V$$

where $\sigma^2 V$ is the variance–covariance matrix of the observations (and also of the unobservable errors), V can incorporate unequal variances for observations as well as correlations between observations. Assume that V has an inverse matrix V^{-1}.

Rather than using least squares estimates of β, we use weighted least squares estimates that minimize

$$(Y - X\beta)'V^{-1}(Y - X\beta)$$

for given Y and X. The estimates can be obtained by solving the normal equations

$$X'V^{-1}X\beta = X'V^{-1}Y$$

Typically, in regression problems, the unique solution is

$$\widehat{\beta} = \left(X'V^{-1}X\right)^{-1}X'CV^{-1}Y$$

These estimates are also BLUEs. Moreover, under the assumption that the observations have a multivariate normal distribution, the weighted least squares estimates are MLEs and MVUEs. The variance σ^2 is estimated by $MSE{=}SSE/dfE$ as before, except now

$$SSE = \left(Y - X\widehat{\beta}\right)'V^{-1}\left(Y - X\widehat{\beta}\right)$$

Testing a full model and a reduced model, works as before, with the F statistic still being

$$F = \frac{SSE(\text{Red}) - SSE(Full)}{dfE(\text{Red}) - dfE(Full)} / MSE(Full)$$

Note that incorporating the more general matrix V has no effect on the process of identifying reduced models—which only depend on the mean structure and not on the variance–covariance structure of the data.

All of these methods presume that V is known. In practice, V is often unknown and must be estimated from the data. This greatly complicates the entire process of data analysis. Typically, when V is estimated, weighted least squares estimates are not BLUEs or MVUEs, nor does the F statistic have an F distribution under the null hypothesis that the reduced model is true. Nonetheless, the ideas behind weighted regression are fundamental to many applications of the linear hypothesis.

A common application of weighted regression is when the observations are independent but have different variances. If the relative sizes of the variances are known, then V is known. Often the variance of each observation is modeled as a function of some predictor variables, see Carroll and Ruppert (1988).

Split plot models involve groups of observations that all have the same correlation. For example, in an experiment that involves different ways of scheduling classes in schools and different ways of assigning homework to students, the behavior of the schools might be considered independent of each other, but a group of five students taken from the same school should have similarities with each other, while being independent of students from other schools. This correlation structure must be incorporated into V. If the same number of students are measured from each school, i.e., if the groups of observations having the same correlation are all of the same size, many linear models still allow one to use least squares estimates and their associated F tests.

Mixed models involve the use of random effects, Partition X and β into two parts, so as to write

$$Y = X_1\beta_1 + X_2\beta_2 + e$$

Instead of considering β_1 and β_2 as fixed unknown parameters, treat β_2 as random. Assume that $\mathrm{E}(\beta_2) = 0$ and $\mathrm{Cov}(\beta_2) = \sigma^2 D$ with, say, e and β_2 independent. Consolidate all of the random parts into $\xi{=}X_2\beta_2{+}e$ and write a linear model

$$Y = X_1\beta_1 + \zeta, \quad E(x) = 0, \quad Cov(\xi) = \sigma^2\left(X_2 D X_2' + 1\right)$$

where $V = X_2\, DX_2' + I$. Typically, D has to be estimated. To estimate β_1, simply use the estimate of D to get an estimate of V which is then treated as the true V.

Mixed models get much more complicated. X_2 is often partitioned into submatrices $Z_1, \ldots, Z_r$ and β_2 is correspondingly partitioned into γ_k, so as to write

$$Y = X_1\beta_1 + \sum_{k=1}^{r} Z_k\gamma_k + e$$

where $\mathrm{E}(\gamma_k){=}0$, $Cov(\gamma_k) = \sigma_k^2 I$, and the γ_k are independent. The γ_k are unknown parameters, and various methods such as residual (restricted) maximum likelihood (REML) and minimum norm quadratic unbiased estimation (MINQUE) are used to estimate them. See also Hierarchical Models: Random and Fixed Effects; Longitudinal Data.

Another application of weighted estimation is in the analysis of spatial data. With spatial data, observations that are taken close together spatially are usually more correlated than observations taken far apart. V reflects the spatial correlations of the observations and X may reflect the locations of the observations. Models for V are developed based on the distances between observations. A common application in geographic information systems is map making, in which

predictions are made for a very fine grid of locations. An individual best linear unbiased predictor for an observation on the grid, say, y_0 corresponding to location x_0' is

$$\widehat{Y}_0 = x_0'\widehat{\beta} + V_{oY}V^{-1}\left(Y - X\widehat{\beta}\right)$$

where V_{0Y} is a matrix of covariances between y_0 and the original observations in Y. In practice, both V and V_{0Y} must be estimated from the data. See also Spatial Statistical Methods.

Multivariate linear models involve creating linear models for several variables at once. For example, y_{i1} might measure verbal ability and y_{i2} mathematical ability. Obviously, for any individual i, y_{i1} and y_{i2} may be correlated. The multivariate linear model fits

$$y_{i1} = x_i'\beta_1 + \varepsilon_{i1} \quad \text{and} \quad y_{i2} = x_i'\beta_2 + \varepsilon_{i2}$$

simultaneously. See also Multivariate Analysis: Discrete Variables, Overview.

Time series data involve observations $y_1, y_2, y_3, \ldots$, taken at regular time intervals. The correlation between two observations depends on the amount of time between them. One simple linear model for explaining the behavior of such data is a first order autoregression model.

$$y_i = \beta_0 + \beta_1 y_{i-1} + \varepsilon_i$$

in which the observation at time i-1 is used to predict the observation at time i. See also Time Series: General and Time Series: Arima Methods.

Generalized linear models involve transforming the mean before making the linear hypothesis. For example, if observations y_i have means m_i one might make the hypothesis that

$$\log(m_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

This hypothesis together with assumptions about the distribution of y_i form a generalized linear model. In particular, since the linear hypothesis is based on the log transformation, this would be called a log-linear model. Parameter estimates are typically found by performing a series of weighted least squares analyses. Generalized linear models allow for the analysis of a variety of nonnormally distributed data. See also Analysis of Variance and Generalized Linear Models; Multivariate Analysis: Discrete Variables (Loglinear Models); Multivariate Analysis: Discrete Variables, Overview.

Conclusion

Models based on the linear hypothesis, and their generalizations provide tools for analyzing a wide variety of data. Linear hypotheses allow one to address numerous interesting questions about the mean structure of the data. Moreover, the models can adjust for very general forms of dependence among the data. Linear models and their generalizations continue to be the most important and widely used models for the statistical analysis of data.

See also: Analysis of Variance and Generalized Linear Models; Bayesian Statistics; Distributions, Statistical: Special and Continuous; Elicitation of Probabilities and Probability Distributions; Estimation: Point and Interval; Experimental Design: Overview; Hierarchical Models: Random and Fixed Effects; Hypothesis Testing in Statistics; Linear Hypothesis: Regression (Basics); Linear Hypothesis: Regression (Graphics); Longitudinal Data; Multivariate Analysis: Discrete Variables (Loglinear Models); Multivariate Analysis: Discrete Variables (Overview); Observational Studies: Overview; Significance, Tests of; Spatial Statistical Methods; Statistical Identification and Estimability; Time Series: ARIMA Methods; Time Series: General.

Bibliography

Carroll, R.J., Ruppert, D., 1988. Transformation and Weighting in Regression. Chapman and Hall, New York.

Christensen, R., Bedrick, E.J., 1997. Testing the independence assumption in linear models. Journal of the American Statistical Association 92, 1006–1016.

Christensen, R., 1991. Linear Models for Multivariate, Time Series, and Spatial Data. Springer, New York.

Christensen, R., 1996a. Plain Answers to Complex Questions: The Theory of Linear Models, second ed. Springer, New York.

Christensen, R., 1996b. Analysis of Variance, Design, and Regression: Applied Statistical Methods. Chapman and Hall, New York.

Christensen, R., 1997. Log-linear Models and Logistic Regression, second ed. Springer, New York.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., York, R.L., 1996. Equality of Educational Opportunity. US Dept. of Health, Education, and Welfare, Office of Education, Washington, DC.

Cook, R.D., Weisberg, S., 1994. An Introduction to Regression Graphics. Wiley, New York.

Cressie, N.A.C., 1993. Statistics for Spatial Data, revised ed. Wiley, New York.

Dobson, Annette J., 2002. An introduction to generalized linear models. CRC press, Boca Raton, FL.

Koopmans, L.H., 1987. Introduction to Contemporary Statistical Methods, second ed. Duxbury Press, Boston, MA.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman and Hall, London.

Miller, F.R., Neill, J.W., Sherfey, B.W., 1998. Maximin clusters for near-replicate regression lack of fit tests. Annals of Statistics 26, 1411–1433.

Montgomery, Douglas C., Peck, Elizabeth A., Geoffrey Vining, G., 2012. Introduction to linear regression analysis. John Wiley & Sons, Hoboken, NJ.

Mosteller, F., Tukey, J.W., 1977. Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley, Reading, MA.

Mosteller, F., Fienberg, S.E., Rourke, R.E.K., 1983. Beginning Statistics with Data Analysis. Addison-Wesley, Reading, MA.

O'Hagan, Anthony, Forster, Jonathan, Kendall, Maurice G., 2004. Bayesian inference. Arnold, London, UK.

Ravishanker, Nalini, Dey, Dipak K., 2001. A first course in linear model theory. CRC Press, Boca Raton, FL.

Searle, S.R., Casella, G., McCulloch, C.E., 1992. Variance Components. Wiley, New York.

Searle, S.R., 1971. Linear Models. Wiley, New York.

Seber, G.A.F., 1977. Linear Regression Analysis. Wiley, New York.

Seber, G.A.F., 1984. Multivariate Observations. Wiley, New York.

Senter, Herman F., 2008. Applied Linear Statistical Models. Journal of the American Statistical Association 103 (482) 880–80.

Shumway, R.H., Stoffer, D.S., 2000. Time Series Analysis and its Applications. Springer, New York.

Linear Hypothesis: Fallacies and Interpretive Problems (Simpson's Paradox)

DV Lindley, Minehead, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8881–8884, © 2001, Elsevier Ltd.

Abstract

In the study of the relationship between two quantities it can happen that the relationship is seriously changed, even reversed, when a third quantity is additionally considered. This is usually known as Simpson's paradox. The article explains how the change can arise, how the true form of the relationship may be deduced, and how good experimental design can avoid the difficulty. The article is limited to situations where the relationships are linear in the expectation.

The study of relationships between two quantities has been important since earliest times: from the connection between weather and the hunted animals, down to cigarettes causing lung cancer. In the laboratory sciences it is possible to determine the relationship by controlling the conditions, changing one quantity, and observing the effect of the change on the other. In the social and behavioral sciences this determination is not possible for three, related, reasons. First, conditions can rarely be controlled. Second, it is often not possible to set a quantity at an assigned value. Third, the relationship may not be exact but possess a random element, as with the heights of parents and their children. As a result, it is not easy to determine the nature of any possible relationship between quantities. Often an apparent connection has been found, on further study, to be spurious. We begin by looking at one of the more bizarre possibilities, usually known as Simpson's paradox (Simpson, 1951), but known to earlier writers. A more recent reference is Lindley and Novick (1981). Chapter 6 of Pearl (2000) contains a fine discussion in terms of causation. It is explained using some illustrative, medical data.

Patients were given either a treatment T, or a placebo T', to alleviate a condition. After a fixed time, it was observed whether they had recovered, R, or not, R'. Table 1 gives the data in the form of a 2×2 contingency table: thus, 16 patients recovered, even with the placebo. The table also provides the marginal totals and the two recovery rates. The treatment appears to have been effective in increasing the recovery rate over the placebo by 10%. (As an aside, the numbers here are small, 80 patients in all; but the phenomenon could persist with 8000. It is not the effect of small samples that is at issue.) There is an apparent relationship between treatment and recovery, but it has a random element and outside conditions have not been controlled.

One of these conditions is the patient's sex. Table 2, in the same form, gives the data for the men who participated in the study. Again the treatment has an effect but in the opposite direction, reducing the recovery rate by 10%. Common sense might suggest that a treatment that was overall beneficial, but harmful to men, would be highly beneficial for women. The data for them can be found by subtracting the cell entries in Table 2 for men from those for all patients in Table 1. Thus, no new information is needed. The results are given in Table 3. Common sense misleads: the recovery rate for women also drops by 10% as a result of the treatment, just as with the men. This is a treatment which is bad for men, bad for women, but good for all of us. This is the paradox. It is here in an extreme form of complete reversal when an additional factor is included. It often arises in the form where there is some change in relationship when extra information is included. How can this happen?

It is easy to see what has gone wrong in the medical example. The condition is more serious for the women, with lower recovery rates than the men. Yet, the treatment has been mainly given to the men – 30 of them, against only 10 women. Perhaps the doctor was suspicious of the treatment and felt it was too dangerous to give to the susceptible women. The consequence is that overall the treatment appears beneficial because it has been mainly given to the men, who were much more likely to recover in any case than were the women. Once

Table 1

	R	R'		Rate
T	20	20	40	50%
T′	16	24	40	40%
	36	44	80	

Table 2

Males	R	R'		Rate
T	18	12	30	60%
T′	7	3	10	70%
	25	15	40	

Table 3

Females	R	R'		Rate
T	2	8	10	20%
T′	9	21	30	30%
	11	29	40	

 http://dx.doi.org/10.1016/B978-0-08-097086-8.42139-2

this is appreciated, it is clear, from the data for the two sexes separately, that the treatment is harmful. Or is it clear? Here are two considerations that raise doubts:

1. Suppose that another quantity in addition to sex is included: for example, whether the patient came from a rural or urban environment. There would then be four contingency tables: one for rural males, one for urban females, and so on. We do not give the data, but it could happen that for each of these four tables there was a 10% improvement in the recovery rate, as with the overall data in Table 1. For this to happen, the allocation of patients would need to depend heavily on sex and environment, but it could happen. Furthermore, it could happen for some factor that was unrecognized by the scientists responsible for the experiment. The lesson is that, without some care in the consideration of quantities beyond the two that are the prime object of study, it is easy to be misled about the relationship between them.
2. Suppose the numbers in the three tables remain unaltered but they apply to an agricultural, rather than medical, example. The treatment and placebo are replaced by two varieties of maize; recovery is now high yield as against low; and being male becomes growing tall as opposed to short. Thus, the phenomenon of males having a higher recovery rate than females corresponds to tall plants giving a greater proportion of high yielders. Finally, the recovery rate is replaced by the proportion of plants with a high yield. In the medical example, it was suggested that the treatment was harmful. This translates, in the agriculture case, to T' giving the better yield. Is that really right? Surely T gave the better yield (Table 1) and grew taller (Tables 2 and 3). The suggestion is that the conclusions, from the same data, are quite different in the two contexts. The reason is that, whereas the medical treatment has no influence on sex, which maize variety is grown does influence the heights of the plants. The causal patterns in the two cases are not the same.

There is another way of appreciating the distinction between the agricultural and medical examples that avoids the difficult concept of causation. Suppose, on the evidence of the tables, and no other, you had another patient with the condition: would you give them the treatment? The answer is No, because the chance of recovery for treated males is less than for males having the placebo, and similarly for females so that sex (which is outside your control) is irrelevant, and not influenced by the treatment applied. But when it comes to which variety of maize to plant, the replacement for sex, growing height, is influenced by the variety chosen. Hence the chances referred to above are irrelevant; only the chances of high yield for each variety are relevant and the growing height does not matter, since it is determined by the choice of variety. Thus T is the better maize.

Notice that in all these examples, one of the quantities, T, could be set at a chosen value. Often this is not possible and the scientist merely passively observes whether T or T' obtains. In this case, transferring from the data to a new patient, in the manner of the last paragraph, becomes even more difficult. There can be a substantial difference between observing that an event occurs and arranging that it occurs. A person who is naturally a nonsmoker can behave quite differently from one who has been pressured into giving up smoking.

In order to fully appreciate what is going on, it is necessary, as always, to argue abstractly, to escape from the contexts of medicine or agriculture, and to examine what principles are involved. This requires some mathematics and here the specific tool of probability. We write $p(A/B)$ for the probability that event A occurs, when it is known, or supposed, that event B has occurred. AB denotes the event that occurs if, and only if, both A and B occur. Then a reasonable interpretation of Table 1 is that $p(R/T) = 0.5$, $p(R/T') = 0.4$, and that T is preferred to T'. From now on, think of R, T, and M (for male) as any three events and R', T', and M' as their negations, respectively. The medical and agricultural examples provide illustrations. A third, social example with racist overtones is provided by the case, T is White (T' is Black), R is intelligent, and M is upper, social class. All data are hypothetical.

Tables 2 and 3 refer, in this abstract form, to probabilities like $p(R/TM)$, differing from those in Table 1 by the inclusion of the conditioning event M or M'. These are related to the probabilities of Table 1 by the standard results

$$p(R/T) = p(R/TM)p(M/T) + p(R/TM')p(M'/T) \qquad [1]$$

and

$$p(R/T') = p(R/T'M)p(M/T') + p(R/T'M')p(M'/T') \qquad [2]$$

These introduce two extra probabilities, $p(M/T)$ and $p(M/T')$. Table 1 compared the probabilities on the left-hand sides of the two equations. Tables 2 and 3 compared the probabilities vertically above one another on the right. Those on the left were larger in the upper equation; those on the right were smaller there. The reason for the paradox is that, on the right-hand sides of the two equations, the weights attached to those probabilities in order to combine them to produce those on the left, namely $p(M/T)$ in the upper, and $p(M/T)$ in the lower, are different. The difference was noted earlier with the treated patients being predominately male.

The paradox can be avoided if the weights are the same; namely $p(M/T) = p(M/T')$, for then any differences on the right will be reflected on the left. The general conclusion is that in studying a relationship involving T, any extraneous factor, M, should, if possible, arise with the same probability whether T is true or not. M and T should be independent. Statisticians say M is confounded with T if the probabilities are not equal. The sound recipe is to avoid confounding. Much effort has been put into designing experiments with little or no confounding.

We referred above to the difference between the medical and agricultural examples, in that the former resulted in a preference for T', and the latter for T. The reason is that in the medical case, the scientist has control over $p(M/T)$ and, in particular, can avoid confounding by having the same proportion of males among treated patients as among those receiving the placebo. With the two varieties of maize, the proportion growing tall, M, is outside the scientist's control, so $p(R/T)$ is what matters. The reader may like to consider the politically sensitive example mentioned above where the same data refer to the relative intelligences of Black and White people.

The discussion so far has referred to events occurring or not. An event can be thought of as a quantity, which takes only two

values, 1 when it occurs, 0 when it does not. The probability that it occurs is then the expectation of the quantity. Looked at from this viewpoint, probability is a special case of expectation. The ideas above, including Simpson's paradox, extend to general quantities, X, Y, and Z, replacing T, R, and M, respectively, whenever the expectations are linear, as they necessarily must be when only two numbers, 1 and 0, are involved. Suppose then that the expectation of X, given Y and Z, is linearly dependent on their values,

$$E(X/Y,Z) = \alpha + \beta Y + \gamma Z \qquad [3]$$

If, still retaining the dependence on Y, expectations are taken over Z, the result is

$$E(X/Y) = \alpha + \beta Y + \gamma E(Z/Y)$$

If $E(Z/Y)$ is also linear, equal to $\lambda + \mu Y$, then

$$E(X/Y) = \alpha' + \beta' Y \qquad [4]$$

where

$$\alpha' = \alpha + \gamma\lambda \ \textit{and} \ \beta' = \beta + \gamma\mu$$

The contrast between eqns [4] and [3] is similar to that between **Tables 1** and **2**. For example, it could happen that β was positive but β' negative. If so, eqn [3] suggests that an increase in Y would be expected to produce an increase in X. Equation [4] similarly suggests a decrease in X.

More careful analysis modifies these statements. Changing γ in eqn [3] provides a measure of the effect on X, while Z is constant. No such restraint on Z is present in eqn [4]. This is similar to Simpson's paradox, where **Tables 2** and **3** held Z, now M or $M,'$ constant, whereas **Table 1** did not. It explains the difference between the agricultural and medical examples. Z, height, cannot be held constant; whereas Z, sex, can.

It is common for eqn [3] to be included in a statement like

$$X = \alpha + \beta Y + \gamma Z + \varepsilon$$

where ε is a random error with zero mean. Similarly eqn [4] becomes

$$X = \alpha' + \beta' Y + \varepsilon'$$

Statements in this form, unlike eqns [3] and [4], disguise the fact that the real distinction between them lies in the conditions under which X is being studied. Just as with eqns [1] and [2], the important subtleties arise because of the conditions. Thus ε' hides the fact that the variation of Z is included within it, while ε does not. It is typically advantageous, when thinking about an expectation, or probability, to recognize the conditions that are assumed to hold, as well as the quantity being expected. In particular, probability depends on two arguments, not one: the uncertain event whose probability is being evaluated, and the events that are supposed to have occurred, the conditions. Distinction is sometimes made between probability and conditional probability. In reality, all probabilities are conditional. Care is needed in working with conditional expectations.

The main conclusion is that it is difficult to establish whether or not there exists a connection between two quantities and, if there does, to determine its precise form. The difficulty principally arises from the presence of other quantities that may disturb any apparent relationship. If controlled experimentation is possible, then consideration of interactions and the avoidance of confounding can be helpful in producing sound conclusions. With observational studies, in contrast, great care is necessary in avoiding inferences that the data do not fully support. An important reason why the social and behavioral sciences are less advanced than the physical ones is that controlled experimentation is more often possible in the latter than in the former.

These considerations do not deny the possibility of establishing relationships in society; merely that more care is required. Emphasis needs to be placed on other quantities and the influence they might have on the relationship. These should be measured and the data incorporated into the analysis. On the other hand, there will be factors, like height in the agricultural example, which, although influential, can be left out of the analysis. All the time it is necessary to bring to the fore the conditions under which the basic relationship obtains. In consideration of that between X and Y, it is essential to take into account Z.

See also: Linear Hypothesis: Regression (Basics); Linear Hypothesis.

Bibliography

Lindley, D.V., Novick, M.R., 1981. The role of exchangeability in inference. Annals of Statistics 9, 45–58.

Pearl, J., 2000. Causality: Models, Reasoning and Inference. Cambridge University Press, Cambridge, UK.

Simpson, E.H., 1951. The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society, Series B 13, 238–241.

Linear Hypothesis: Regression (Basics)

Sanford Weisberg, University of Minnesota, St Paul, MN, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8884–8888, © 2001, Elsevier Ltd.

Abstract

Regression analysis is the study of how a response variable depends on one or more predictors. Usually, dependence is assumed to be through the mean, and we think of the mean or regression function as describing how the mean of the response depends on the predictors. Although some regression problems can be usefully summarized using nonparametric regression methods that make a minimum of assumptions, adding a few plausible assumptions allows using parametric models for regression that give elegant and simple results. The important class of linear regression models is described in this article. These models are frequently used in practice, and can often lead to very useful results. Topics covered include both simple and multiple regression, the basic forms of the models, interpretation of estimates and parameters, testing ideas, including comparison of groups, and diagnostic methods.

Regression analysis is the study of how a response variable depends on one or more predictors, for example how crop yield changes as inputs such as amount of irrigation or type of seed are varied, or how student performance changes as factors such as class size and expenditure per pupil are varied. In this article, we present the basics of the *linear regression model*, in which (1) the dependence of the response on the predictors is through the mean and (2) the mean of the response given the predictors is a linear function of the predictors. This is probably the most important and frequently encountered regression model.

Introduction

Regression analysis is the study of *dependence*: how does a response variable change as the values of one or more predictor variables are changed? In the general case, we have a response variable y, a set of predictor variables x, and we seek to understand the *conditional distributions* given by $F(y|x)$ as the predictors x are varied. Although the study of dependence probably began even earlier, two principal ideas lead to modern regression analysis (Stigler, 1986). At the beginning of the nineteenth century, Legendre in France and Gauss in Germany studied the mean of a response as the predictors are varied using parametric models with only a few parameters. They suggested estimating the parameters by minimizing the sum of squared errors, which is the foundation for the least squares methods commonly used today. Toward the end of the nineteenth century in England, Galton, and later Pearson and Yule, provided many of the modern notions of correlation that provide a language for understanding dependence.

An example of a regression problem with one predictor is given graphically in **Figure 1**, where the response variable $y = \text{RCC}$ is the red cell count and $x = \text{LBM}$ is the lean body mass in kilograms for 202 elite athletes that trained at the Australian Institute of Sport in the early 1990s (from Cook and Weisberg, 1999). This scatterplot serves two general purposes in a regression context. First, it can help the viewer find separated or otherwise unusual points. In **Figure 1**, one of the athletes had an unusually high value of RCC, about 7, with a relatively modest value of LBM. Such separated points are often of interest in their own right, and in any case they may have substantial effects on any computed summaries. The second use for this graph is to visualize how the conditional distribution $F(\text{RCC}|\text{LBM})$ changes as the value of LBM is varied. Conditioning on LBM means looking at narrow vertical slices in the plot. The points within a slice represent approximately a sample from the distribution of RCC given that LBM is 'fixed' at the midpoint of the interval. We can visually integrate the contents of many slices to summarize changes in conditional distributions as we move from left to right. In **Figure 1**, for example, we can see that the *mean function*, $E(\text{RCC}|\text{LBM})$ that joins the mean in each slice appears to increase as we move from left to right, and furthermore that the increase appears to be approximately the same for all values of LBM. In addition, we have enough points to say something about the conditional variability by examining the *variance function* $\text{Var}(\text{RCC}|\text{LBM})$. The variability appears to be approximately the same across the plot, or apparently $\text{Var}(\text{RCC}|\text{LBM} = 50) \approx \text{Var}(\text{RCC}|\text{LBM} = 80)$, or any other value.

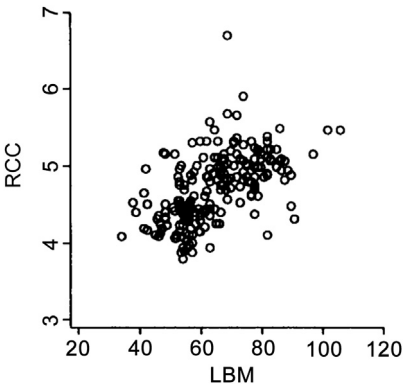

Figure 1 Scatterplot for the regression of RCC on LBM.

Simple Linear Regression

The simple linear regression model makes assumptions concerning both the mean function $E(y|x)$ and the variance function. In particular, we assume that for some transformation $u = u(x)$, we can write

$$E(y|u) = \beta_0 + \beta_1 u \qquad [1]$$

This equation requires some careful explanation. First, following Cook and Weisberg (1999), we allow the predictor x to be replaced by any monotonic transformation of it, such as the logarithmic, or inverse transformation. Doing so considerably expands the range of problems that can be handled using simple regression. Given this transformation, eqn [1] says that the expected increase in y for a unit increase in u is always equal to β_1 units, regardless of the initial value of u. Thus in a graph of y vs u, the mean function specified by eqn [1] is a straight line with intercept β_0 and slope β_1.

The mean function of eqn [1] concerns average behavior, but says nothing about the variation. Suppose we have observed data (x_i, y_i), $i = 1, \ldots, n$, observed on n units, and suppose further that we have computed $u_i = u(x_i)$ for the known transformation u. Of course the points (u_i, y_i) will not fall exactly on the line (they certainly do not do so in Figure 1). The unobservable differences

$$\varepsilon_i = y_i - E(y_i|u_i)$$

are called *errors*. By construction, if the model is correct the errors are random variables with mean zero and variance given by a positive constant σ^2. This amounts to assuming that the variance function $\mathrm{Var}(y|x)$ is constant. We assume that the ε_i are independent random variables. They are often assumed to be normally distributed, but the normality assumption is required only for inference, not for estimation.

The model given by eqn [1] and the assumptions concerning variability just described greatly simplify the regression problem by reducing the mean function to only two parameters, and the variance function to only one parameter. When this model is appropriate, it gives a very elegant summary to a regression problem.

Fitting

The most common method of estimating the unknown parameters in eqn [1] based on data uses ordinary least squares. Find the values $\widehat{\beta}_0$ and $\widehat{\beta}_1$ of (β_0, β_1) that minimize the residual sum of squares function $\mathrm{RSS}(\beta_0, \beta_1)$, given by

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \mathrm{fit}_i)^2 \qquad [2]$$

where $\mathrm{fit} = \beta_0 + \beta_1 u_i$. Finding the estimates is an elementary calculus problem, and in general the estimates are unique and are given by a formula. The estimate of σ^2 is just $\widehat{\sigma}^2 = \mathrm{RSS}/n - 2$, where RSS is the value of the residual sum of squares function evaluated at the estimates, and $n - 2$ is the degrees of freedom, which is equal to the number of observations n minus the number of parameters estimated in the mean function, which is 2 for simple linear regression in eqn [1].

Table 1 gives typical computer output for fitting the simple linear regression model to the athletes data. The column marked Estimate gives the estimated coefficients, so the intercept is estimated to be about 3.47. The estimated slope of 0.019 suggests that we expect RCC to increase on the average by about 0.019 units for each 1 kg increase in LBM. The column marked Std. error gives the estimated standard errors of these values, which are computed as $\widehat{\sigma}$ a times a constant that depends on the values of the u_i but not on the response. The t-values are the ratio of the estimate to its standard error. These are appropriate test statistics for the hypotheses that each of the coefficients in turn is equal to zero. Assuming normal errors, these values can be compared with the t_{n-2} distribution to get significance levels. In the athletes data, tests concerning the intercept are of little interest. The test that the slope is equal to zero vs the general alternative has value $t = 9.34$ and a corresponding tiny significance level, suggesting clearly that RCC increases with LBM.

Table 1 Simple linear regression of RCC on LBM

	Coefficient estimates		
Label	*Estimate*	*Std. error*	*t-value*
Constant	3.46616	0.136820	25.334
LBM	0.0193060	0.00206768	9.337
R squared		0.303574	
Sigma hat		0.383145	
Number of cases		202	
Degrees of freedom		200	

A common summary statistic for the fit of a regression is the quantity R^2, which is defined to be the *fraction of the variability in the response explained by the model*, so

$$R^2 = \frac{\mathrm{SYY} - \mathrm{RSS}}{\mathrm{SYY}} \qquad [3]$$

where $\mathrm{SYY} = E(y_i - \overline{y})^2$ is the total variability in the response ignoring the predictors. If $R^2 = 1$, then $\mathrm{RSS} = 0$ and the mean function matches the data perfectly. If $R^2 = 0$, then $\mathrm{SYY} = \mathrm{RSS}$, and knowledge of the predictors tells us nothing about the response. For the athletes data, $R^2 = 0.30$, so about 30% of the variability in RCC is explained by LBM. R^2 is a most appropriate measure when the corresponding scatterplot appears to be an elliptical swarm of points which in turn suggests bivariate normality of both the response and the predictor. This is a very strong assumption that appears to be appropriate for this example.

Diagnostics

The residuals are defined to be the differences between the observed values y_i and the fitted values $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 u_i$, so $e_i - y_i - \widehat{y}_i$. Unlike the errors ε_t, the residuals e_i are observable, and so they can be used in graphs and other procedures as a check on the assumptions of the fitted model. Imagine drawing a scatterplot of the residuals vs u. If the model is correct, then we should expect that the conditional distribution of $e|u$ should be the same as we move from left to right: the mean function should be constant and the variance function should be constant. If the model is not correct, then either of these two characteristics may not hold. The residual plot for the athletes data is shown in Figure 2. As one would have expected,

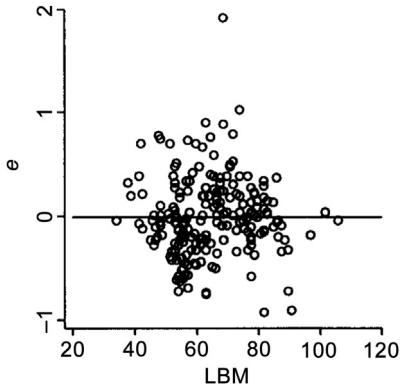

Figure 2 Residual plot from simple regression of RCC on LBM.

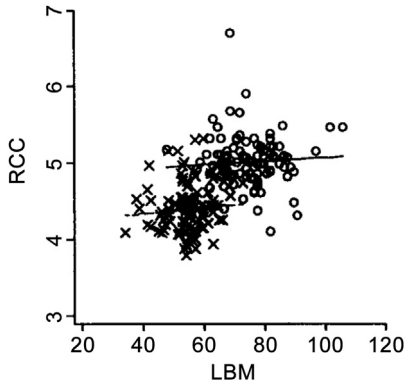

Figure 3 Regression of RCC on LBM controlling for gender. Points for females are marked with a X and have a fitted line given by the dashed line. Points for males are marked with a 0 and the fitted line is the solid line.

the model for the mean function matches the data fairly well, as both the mean and variance functions appear to be constant. The other interesting feature is the poor fit for one athlete with very high RCC. Deleting this case and refitting can tell us about the impact of this point on the fitted equation; doing so results in only minor changes in the estimated coefficients or in $\widehat{\sigma}$. Other separated points can cause very large changes in estimates, and such points are called influential.

Multiple Regression

Many regression problems require consideration of more than one predictor, and it is required to understand how the response y depends simultaneously on the predictors $x_1, x_2, \ldots, x_n$. For example, the dependence of RCC on LBM discussed so far ignores the fact that these data consist of both male and female athletes, and one might reasonably assume that gender differences are important. This will require either fitting separate models for the two populations of males and females, or more generally including gender effects in the mean function. Other problems might require additional predictors; in the athletes data, for example, including height, weight, or some blood chemistry variables might be appropriate.

We begin with p predictors $x_1, \ldots, x_p$, and from these we will build a set of k *terms* $u_0, u_1, \ldots, u_{k-1}$, to get the multiple linear regression mean function

$$\mathrm{E}(y|u_1, \ldots, u_{k-1}) = \beta_0 u_0 + \beta_1 u_1 + \ldots + \beta_{k-1} u_{k-1} \qquad [4]$$

Including terms, which are functions of the predictors, makes this model very general. For example, a model with one predictor x might include the terms $u_0 = 1$ for the intercept, $u_1 = x$, $u_2 = x^2$, and $u_3 = x^3$, so polynomial models are included in the multiple linear regression class. In addition, the u terms can include interactions (e.g., $u = x_1 x_2$), transformations (e.g., $u = \log(x_1)$), and other functions of the predictors. Quantitative variables like gender can be included in a model by defining *factors*, which are sets of indicator variables (often having the values 0 or 1) that correspond to the levels of the factor. For example, the variable *Treatment* with three levels can be included in a model with two indicator variables, the first with the value 1 for units that received treatment T1 and 0 otherwise; the second with the value 1 for units getting treatment T2, and 0 otherwise; T3 is indicated if both indicators are equal to 0. Since gender has only two levels, only one indicator variable is required.

Suppose that Gender has the value 1 for females and 0 for males. Figure 3 shows a fit of the mean function

$$\mathrm{E}(\mathrm{RCC}|\mathrm{LBM},\ \mathrm{Gender}) = \beta_0 + \beta_1\ \mathrm{Gender} + \beta_2\ \mathrm{LBM} + \beta_3\ \mathrm{LBM} \times \mathrm{Gender} \qquad [5]$$

to the athletes data. This is a multiple regression model with two predictors, LBM and Gender, and four terms, an intercept, an interaction, and terms for LBM and Gender. For males, Gender = 0 and eqn [5] becomes

$$\mathrm{E}(\mathrm{RCC}|\mathrm{LBM},\ \mathrm{Gender} = 0) = \beta_0 + \beta_2\ \mathrm{LBM}$$

while for females, Gender = 1 and

$$\mathrm{E}(\mathrm{RCC}|\mathrm{LBM},\ \mathrm{Gender} = 1) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)\mathrm{LBM}$$

We recognize that this model fits straight lines separately for each gender with intercept and slope β_0 and β_2 for males, and $\beta_0 + \beta_1$ and $\beta_2 + \beta_3$, for females. We see from the fitted least squares lines shown in Figure 3 that within gender the regression have different intercepts (males appear to have higher RCC than females), but the two groups appear to have approximately equal slopes, and those two slopes may well be zero. Thus, by conditioning on group, the relationship between RCC and LBM disappears. Table 2 summarizes the ordinary least squares fit based on model eqn [5]. The fitting method is

Table 2 Multiple linear regression of RCC on LBM and Gender

	Coefficient estimates		
Label	*Estimate*	*Std. error*	*t-value*
Constant	4.85114	0.255403	18.994
Gender	−0.666013	0.372175	−1.790
LBM	0.00234984	0.00339170	0.693
LBM × gender	0.00164645	0.00595363	0.277
R squared		0.466583	
Sigma hat		0.337009	
Number of cases		202	
Degrees of freedom		198	

a direct generalization of the method for simple regression given at eqn [2], except that fit_i is now defined using eqn [5]. We now see that the estimates for LBM and LBM × Gender are very small both in absolute value, and small relative to their standard errors. All this suggests that, given gender, RCC and LBM are independent.

This example illustrates that choice of predictors in a regression problem can be very important. The initial analysis, ignoring gender, suggested that RCC increased with increasing LBM. This spurious dependence appears to be due to gender differences: women have lower values of LBM than men, on the average, and they also appear to have lower values of RCC than do men (at least for elite athletes). The initial analysis reflected gender differences, not a meaningful relationship between the response and the predictor.

Interpreting Parameters

Weisberg (1985) reported the results of a regression to study per capita fuel consumption for the United States as a function of four predictors: *Tax* rate, in cents per gallon, *Dlic*, the fraction of adults with drivers licenses; *Inc*, the per capita income in thousands of dollars, and *Road*, the number of miles of roadway in the state. The fitted linear regression model was:

$$\text{Fuel} = 377 + 35\ \text{Tax} + 13\ \text{Dlic} - 66\ \text{Inc} - 2.4\ \text{Road}$$

Assuming that this mean function is appropriate for these data (a question that should always be considered in trying to understand a regression problem), how are the parameter estimates to be interpreted? For example, the coefficient of −66 for *Inc* suggests that for every $1000 increase in per capita income, fuel consumption will decrease by about 66 gallons per person, assuming that all other predictors could be held fixed at their current values. Similarly, the value of +35 for *Tax* suggests that for every penny increase in fuel tax, consumption goes *up*, contrary to the common sense result that increasing cost through taxes should decrease consumption, assuming all other factors are held fixed. How can this happen?

These data are observational: tax rates are determined by state legislatures in response to many other factors, possibly including the per capita fuel consumption. Thus the observed positive relationship may be due to high consumption causing an increase in taxes, rather than the other way around. There is no way to sort out an issue like this one without actually experimentally changing variables and observing what happens; this is surely not possible in setting tax rates.

As another example, Weisberg (1985) reported on the regression summary of the dependence of accident rate per million vehicle miles on speed limit *Slim*, number of access points *Acpt* and length *Len* of a segment for 39 segments of major Minnesota highways. The fitted model was

$$\text{Rate} = 9.32 - 0.77\ \text{Len} - 0.10\ \text{Slim} + 0.10\ \text{Acpt}$$

We get the sensible result that increasing the number of access points in a segment, while keeping the other two variables fixed, will increase the accident rate by 0.1. The other two coefficients are less obvious. The negative sign for *Slim* suggests that increasing speed limits will lower accident rates (these data predate the original reduction of the maximum speed limit to 55 mph), but once again we cannot tell if higher speeds cause fewer accidents, or if dangerous segments of road encourage the highway department to lower the speed limit; both would give rise to the negative coefficient.

The coefficient for *Len* is also curious; why should accident rate be related to the length of the segment, since the rate is already adjusted for the miles driven? If most accidents occur at isolated dangerous spots that are relatively rare, then increasing the length of a segment by a small amount is unlikely to include another dangerous spot. The longer segment will then have about the same number of accidents, but a lower value for accidents per million vehicle miles. Consequently, the negative association between the rate and the length, adjusted for the other variables, is an artifact of the way the data are collected. It has no intrinsic interest on its own, but it must be included in all regression models to remove the effect of this artifact.

Conclusion

Regression methodology generally seeks to understand the dependence of a response variable on a set of predictors as the predictors are varied. This very general methodology finds application in almost every branch of human inquiry. The approach to regression briefly outlined here is expanded at a relatively low mathematical level in Cook and Weisberg (1999). Fox (1997) provides another entry to the literature on regression, at about the same level. Draper and Smith (1998) requires a modestly higher level of mathematical sophistication.

Almost any of the popular statistical computing package will include programs for fitting the type of models discussed in this article. The larger programs like SAS offer several seemingly different procedures with names like 'proc reg' and 'proc glm' that in reality are just fitting the regression models described here. Newer approaches to regression require extensive use of graphics, and several programs are specifically designed for graphics, for example, the free program Arc that is described in Cook and Weisberg (1999).

See also: Linear Hypothesis: Regression (Graphics); Linear Hypothesis.

Bibliography

Cook, R.D., Weisberg, S., 1999. Applied Regression Including Computing and Graphics. Wiley, New York.
Draper, N., Smith, H., 1998. Applied Regression Analysis, third ed. Wiley, New York.
Fox, J., 1997. Applied Regression Analysis, Linear Models, and Related Methods. Sage, Thousand Oaks, CA.
Stigler, S.M., 1986. The History of Statistics. Belknap, Cambridge, MA.
Weisberg, S., 1985. Applied Linear Regression, second ed. Wiley, New York.

Linear Hypothesis: Regression (Graphics)

R Dennis Cook, University of Minnesota, St. Paul, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 13, pp. 8888–8893, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Regression analysis is the study of how a response variable depends on one or more predictors. In regression graphics we pursue low-dimensional sufficient summary plots. These plots, which do not require a model for their construction, contain all the information on the response that is available from the predictors. They can be used to visualize dependence, to discover unexpected relationships, to guide the choice of a first model, and to check plausible models. This article covers the foundations for sufficient summary plots and how they can be estimated and used in practice. Their relationship to standard model-based graphics such as residual plots is covered as well.

Regression analysis is the study of how a response variable depends on one or more predictors. For example, as a part of an ecological study of New Zealand mussels, we might wish to study the dependence of muscle mass Y (g), the edible portion of a mussel, on the size of the shell as represented by the shell height X_1, shell length X_2, and shell width X_3 (mm). We later use a data set from this study containing observation on 82 mussels.

Graphical displays can be quite useful in regression analysis because they can be used to visualize dependence, to discover unexpected relationships, to guide the choice of a model, and to check and refine plausible models. In this article, we present the basic ideas and methods of regression graphics.

Introduction

The overarching goal of a regression analysis is to understand how the conditional distribution of the response variable Y given a vector of p predictors $\mathbf{X} = (X_j, j = 1,\ldots, p)$ depends on the value assumed by $\mathbf{X}$. While attention is often restricted to the mean function $\mathrm{E}(Y|\mathbf{X})$ and perhaps the variance function $\mathrm{Var}(Y|\mathbf{X})$, in full generality it is the conditional distribution of $Y|\mathbf{X}$ that is of interest. The population of New Zealand mussels can be viewed as consisting of several subpopulations, one for each distinct shell size as represented by the value of $\mathbf{X}$. In the regression we wish to study how the distribution of muscle mass changes as we move between subpopulations.

Linear regression is a restrictive but common form of analysis in which $\mathrm{E}(Y|\mathbf{X})$ is expressed as a linear function of the predictors and the data are used to estimate the unknown coefficients of the predictors (see Linear Hypothesis: Regression (Basics)). For example, we might begin an analysis of the mussel regression by fitting the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \sigma\varepsilon \qquad [1]$$

where ε is an unobservable random error with mean 0 and variance 1, the βs are unknown regression coefficients, and σ is the unknown standard deviation. In this particular model, the distribution of $Y|\mathbf{X}$ depends on $\mathbf{X}$ only through the mean function

$$E(Y|\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \qquad [2]$$

which is linear in $\mathbf{X}$. The variance function is constant $\mathrm{Var}(Y|\mathbf{X}) = \sigma^2$ under the model.

The choice of a model can be problematic, and is often the most difficult part of a regression analysis. Beyond the general notion that linear models may serve as useful approximations, there is little prior information to support model (1) for the mussel regression and data-based modifications may well be required. Graphics can be useful in the choice of a first model. If the mussel regression had only one predictor X_1, a two-dimensional scatter plot of Y vs. X_1 could be used to visualize how the conditional distribution $Y|\mathbf{X}_1$ changes with the value X_1. Similarly, if the regression had only two predictors, say X_1 and X_2, we could use a computer-generated three-dimensional plot of Y vs. (X_1, X_2) to visualize the conditional distributions. But the mussel regression has three predictors, requiring a four-dimensional plot to view the data in full. Four-dimensional plots can not be viewed in their entirety in our three-dimensional world, so we must find indirect low-dimensional methods of viewing regression data when there are more than two predictors.

For instance, a scatterplot matrix of the mussel data is shown in Figure 1. The diagonal cells in the scatterplot matrix serve to name the axes for the individual plots and show the observed range for each variable. The top row contains plots of the response vs. each of the predictors, and the predictors are plotted in the lower left 3×3 array. This scatterplot matrix contains much information: we see that each of the mean functions for the marginal regressions in the top row is curved and that the variance functions are increasing. There are fairly strong linear relationships between the predictors, indicating that collinearity may be an issue. This type of information from low-dimensional plots can be useful but unfortunately may have little to do with the regression of Y on $\mathbf{X}$. The fact that the marginal mean functions $\mathrm{E}(Y|X_j)$, $j = 1, 2, 3$, are curved does not necessarily imply that the full mean function $\mathrm{E}(Y|\mathbf{X})$ is a nonlinear function of $\mathbf{X}$. Similarly, we cannot safely conclude that the variance function $\mathrm{Var}(Y|\mathbf{X})$ is not constant.

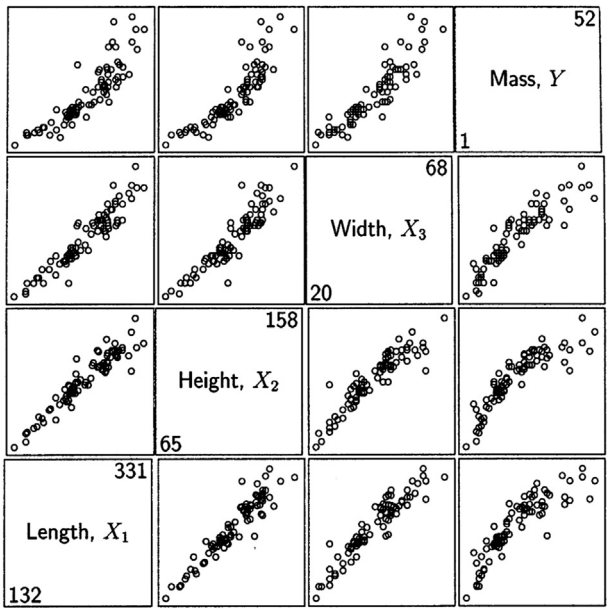

Figure 1 Scatterplot matrix for the mussel regression.

Plots of convenience such as the scatterplot matrix of Figure 1 may not portray the regression accurately and can misdirect the analysis. Low-dimensional plots must be constructed and interpreted carefully if we are to avoid such possibilities. In regression graphics we pursue low-dimensional *sufficient summary plots* which contain all the information on Y that is available from $\mathbf{X}$, and which avoid the distortion that can result from simple convenient displays of the data.

Figure 2 shows a summary plot of muscle mass Y vs. the linear combination of the predictors

$$\mathbf{b}^T\mathbf{X} = 0.05X_1 + 0.08X_2 + 0.86X_3 \qquad [3]$$

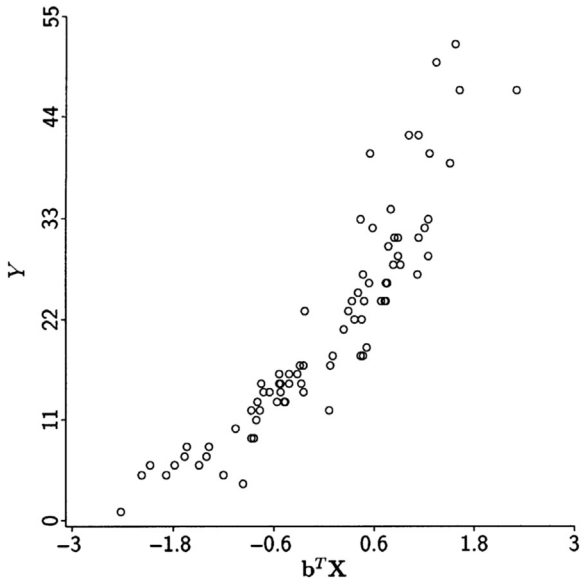

Figure 2 Summary plot for the mussel regression.

Methods of regression graphics discussed later in this article enable us to infer that this plot is *sufficient* and thus contains all the information on muscle mass Y that is available from the three predictors in $\mathbf{X}$, namely shell height, length, and width. Because the plot is inferred to be sufficient and because no model was used in its construction, we can use it to guide the remaining analysis including the choice of a first model. We conclude from the plot that the mean function is curved and that the variance function is very likely not constant. Both conclusions contradict model (1).

Sufficient Summary Plots

Dimension reduction is a leitmotif of statistics. For instance, starting with a sample $z_1,\ldots, z_n$ from a univariate normal population with mean, μ and variance 1, we know that the sample mean $\bar{z}$ is sufficient for μ (see Statistical Sufficiency). This means that we can replace the original n-dimensional sample with the one-dimensional mean $\bar{z}$ without loss of information on μ.

In the same spirit, dimension reduction without loss of information is a dominant theme of regression graphics. The goal of a regression study is to infer about the conditional distribution of the response Y given the $p \times 1$ vector of predictors $\mathbf{X}$. A central goal of regression graphics is to reduce the dimension of $\mathbf{X}$ without loss of information on the conditional distribution of $Y|\mathbf{X}$ and without requiring a model. We call this sufficient dimension reduction, borrowing terminology from classical statistics. Sufficient dimension reduction leads naturally to sufficient summary plots which contain all of the information on the regression that is available from the sample.

In regression graphics we search for the smallest number $d \leq p$ of linear combinations of the predictors, say $\mathbf{b}_1^T\mathbf{X}, \ldots, \mathbf{b}_d^T\mathbf{X}$, that contain all the information about Y that is available from $\mathbf{X}$. To formalize this idea, we re-express it in the form of a conditional independence statement

$$Y \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^T\mathbf{X} \qquad [4]$$

Here columns of the $p \times d$ matrix $\mathbf{B}$ are the vectors $\mathbf{b}_1,\ldots, \mathbf{b}_d$ so that $\mathbf{B}^T\mathbf{X}$ is the $d \times 1$ vector of linear combinations $\mathbf{b}_1^T\mathbf{X}, \ldots, \mathbf{b}_d^T\mathbf{X}$. The notation '$\perp\!\!\!\perp$' means independence. The statement is, thus, that Y is independent of $\mathbf{X}$ given any values for the d linear combinations $\mathbf{B}^T\mathbf{X} = (\mathbf{b}_1^T\mathbf{X}, \ldots, \mathbf{b}_d^T\mathbf{X})^T$. The value of d is called the structural dimension of the regression of Y on $\mathbf{X}$, and we refer to a regression as having kD structure for $k = 0, 1,\ldots, p$. We can always find a value of d and a corresponding matrix $\mathbf{B}$ so that Egn(4) is true. For example, (4) is always true when $d = p$ and $\mathbf{B} = \mathbf{I}$.

If a regression has 0D structure, then $Y \perp\!\!\!\perp \mathbf{X}$, and the predictors contain no information about the response. In such cases, a simple histogram or other graphical representation of the marginal distribution of Y is a sufficient summary plot. If $d = 1$ then all the information about Y that is available from $\mathbf{X}$ is contained in a single linear combination $\mathbf{b}_1^T\mathbf{X}$, and a plot of Y vs. $\mathbf{b}_1^T\mathbf{X}$ is a sufficient summary plot for the regression. Under 2D structure all the information about Y that is available from $\mathbf{X}$ is contained in two linear combinations $\mathbf{b}_1^T\mathbf{X}, \mathbf{b}_2^T\mathbf{X}$, and

a three-dimensional plot of Y vs $\mathbf{b}_1^T\mathbf{X}, \mathbf{b}_2^T\mathbf{X}$ is a sufficient summary plot for the regression.

Each of the following regression models is an example of a regression with 1D structure.

$$Y = b_0 + \mathbf{b}^T\mathbf{X} + \sigma\varepsilon \quad [5]$$

$$Y^{(\lambda)} = b_0 + \mathbf{b}^T\mathbf{X} + \sigma\varepsilon \quad [6]$$

$$Y = f(\mathbf{b}^T\mathbf{X}) + \sigma\varepsilon \quad [7]$$

$$Y = f(\mathbf{b}^T\mathbf{X}) + \sigma(\mathbf{b}^T\mathbf{X})\varepsilon \quad [8]$$

Model (5) is the linear regression model. In model (6) we have a linear regression model after some response transformation represented by the unknown parameter λ. The mean function in model (7) is $E(Y|\mathbf{X}) = f(\mathbf{b}^T\mathbf{X})$, where f may be unknown or known, but it depends on only one linear combination of the predictors and thus the regression still has 1D structure. Similarly, in model (8) both the mean and standard deviation functions may depend on $\mathbf{b}^T\mathbf{X}$. Generalized linear models with corresponding link functions also have 1D structure. The class of regression models with 1D structure is quite large and covers many models used in practice. A regression model with 2D structure can be considerably more complex that a regression with 1D structure. The model

$$Y = f(\mathbf{b}_1^T\mathbf{X}, \mathbf{b}_2^T\mathbf{X}) + \sigma(\mathbf{b}_1^T\mathbf{X}, \mathbf{b}_2^T\mathbf{X})\varepsilon$$

is one example of a regression with 2D structure.

The structural dimension d of a regression and the corresponding linear combinations $\mathbf{B}^T\mathbf{X}$ will generally be unknown. In the next section we discuss methods for estimating these quantities, leading to estimated sufficient summary plots.

Estimating Sufficient Summary Plots

There are several methods for estimating sufficient summary plots or properties thereof. While there are no restrictive constraints placed on the foundations of regression graphics as represented by Eqn. [4], some constraints are necessary for estimation. The most notable is that there should be no strong nonlinear relations among the predictors. This requirement constrains the marginal distribution of $\mathbf{X}$ and not the conditional distribution of $Y|\mathbf{X}$ as is usual when modeling in practice. It is part of the cost of insuring that certain low-dimensional plots contain useful information on the regression. There are several methods for insuring that this requirement holds in practice, including the possibility of using power transformations of the predictors to remove extreme nonlinearities.

Returning to the scatterplot matrix in Figure 1, we see that the mean functions in the predictor plots are all strongly linear, and thus we can proceed to infer about properties a sufficient summary plot. A first inference is available from the *inverse response plots* of X_j vs. Y in the last column of the scatterplot matrix. Since all of these plots have the same shape we infer that $d = 1$ and thus that there is a two-dimensional sufficient summary plot. The justification for this conclusion is based on the vector of inverse mean functions $E(\mathbf{X}|Y)$. If all inverse mean functions are linear related so that for some function $m(Y)$ of the response $E(X_j|Y) = E(X_j) + \alpha_j m(Y)$, $j = 1,\ldots, p$, then 1D structure is indicated. Next, if $d = 1$ as we have just inferred then there is a $p \times 1$ vector $\mathbf{b}$ such that (a) $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{b}^T\mathbf{X}$ and (b) a plot of Y vs. $\mathbf{b}^T\mathbf{X}$ is a sufficient summary of the regression.

The next issue in our discussion of the mussel data is how to estimate $\mathbf{b}$. If there are no strong nonlinear relations among the predictors then $\mathbf{b}$ can be estimated as the coefficient vector of $\mathbf{X}$ from the ordinary least squares regression of Y on $\mathbf{X}$. And the estimated sufficient summary plot can be taken as a plot of Y vs. the fitted value $\hat{Y}$ from this same regression. This is how the summary plot of Figure 2 was produced. No model is being assumed in this use of ordinary least squares. Indeed, we saw previously that the multiple linear regression model (1) is not sustained by the data. Nevertheless, the estimated coefficient vector $(\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)^T$ still provides a consistent estimate of $\mathbf{b}$. Estimation methods other than least squares can be used as well, provided the objective function is convex.

There are several more comprehensive methods for inferring about sufficient summary plots. All are based on transforming the original vector $\mathbf{X}$ of p predictors to p estimated linear combinations

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \rightarrow \mathbf{Z} = \begin{pmatrix} \widehat{\mathbf{b}}_1^T\mathbf{X} \\ \widehat{\mathbf{b}}_2^T\mathbf{X} \\ \vdots \\ \widehat{\mathbf{b}}_p^T\mathbf{X} \end{pmatrix}$$

The p estimated linear combinations $\widehat{\mathbf{b}}_j^T\mathbf{X}$ are ordered on the likelihood that they will be needed in a sufficient summary plot. If it is inferred that $d = 0$ then none of the linear combinations is needed. If it is inferred that $d = 1$ then only the first is needed and a plot of Y vs. $\widehat{\mathbf{b}}_1^T\mathbf{X}$ is the estimated summary plot. Similarly, if the structural dimension is inferred to be $d = 2$ then only the first two estimated linear combinations are needed and a three-dimensional plot of Y vs. $(\widehat{\mathbf{b}}_1^T\mathbf{X}, \widehat{\mathbf{b}}_2^T\mathbf{X})$ is the estimated summary plot.

There are three primary numerical methods for producing the vector $\mathbf{Z}$ of linearly transformed predictors, each with its own advantages and disadvantages

a. Sliced Inverse Regression (SIR)
b. Sliced Average Variance Estimation (SAVE)
c. Principal Hessian Directions (pHd)

SIR and SAVE use the column of inverse response plots in a scatterplot matrix of the predictors and the response, as discussed previously in connection the mussel data. They measure the shape of the inverse response plots by first replacing the response Y with a discrete version $\tilde{Y}$ constructed by partitioning the range of Y into K *slices*. SIR then gauges the shapes of the inverse mean functions $E(\mathbf{X}|\tilde{Y})$, while SAVE gauges the shapes of the inverse functions $E(\mathbf{X}|\tilde{Y})$ and the inverse variance functions $\mathrm{Var}(\mathbf{X}|\tilde{Y})$ simultaneously. The number of slices is a tuning parameter in both approaches. pHd is based on third moments between the response and the predictors, and does not have a tuning parameter. Each method comes with an inference procedure for the structural dimension d of the regression. In particular, beginning with $\mathbf{Z}$ predictors, graphical regression is

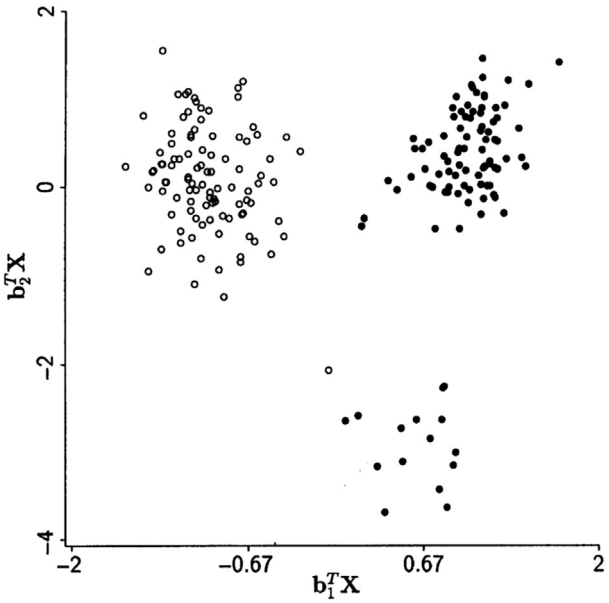

Figure 3 Estimated sufficient summary plot from SAVE for the bank note data.

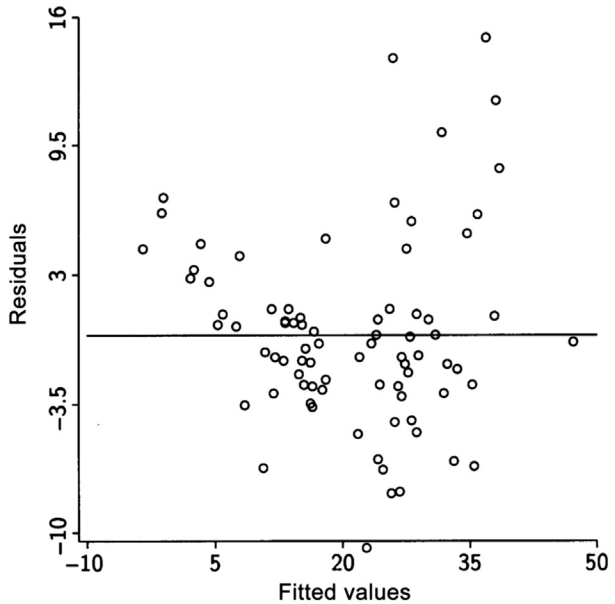

Figure 4 Scatterplot of residuals versus fitted values from the ordinary least squares fit of model.

a method for visually estimating the structural dimension of a regression and the corresponding summary plot.

SIR is good at estimating sufficient summary plots with linear trends, but it does not generally work well for finding plots with curvature and little linear trend. pHd works best for finding quadratic trends. SAVE is the most comprehensive, but may miss information when it is concentrated in linear trends where SIR works well.

We use a data set on 200 Swiss bank notes for illustration. The response is binary with $Y = 1$ indicating a counterfeit note and $Y = 0$ indicating a genuine note. The methods of regression graphics discussed previously apply equally to regressions with continuous and binary responses. There are six predictors, each giving a different aspect of the size of a note: $X_1 =$ length of bottom edge, $X_2 =$ diagonal length, $X_3 =$ left edge length, $X_4 =$ length at center, $X_5 =$ right edge length, and $X_6 =$ top edge length.

Application of SAVE to the bank note data indicated 2D structure. Shown in Figure 3 is a binary response plot of the first two linear combinations $(\widehat{\mathbf{b}}_1^T\mathbf{X}, \widehat{\mathbf{b}}_2^T\mathbf{X})$. We infer that this plot contains all the information about a note's authenticity that is available from the six predictors. The summary plot shows an outlying genuine note, good separation between the genuine and counterfeit notes, and a bimodal distribution within the counterfeit notes. The outlying genuine note could be a mislabeled counterfeit note, and the bimodal distribution within the counterfeit notes might reflect two different production processes.

Special Purpose Graphics

In addition to sufficient summary plots, there are a variety of special purpose graphics that can be used to address specific issues. These visualization methods mostly center on assessing or improving a target model.

Perhaps the most common graphical methods of model assessment are based on searching for dependence between residuals from the fitted model and the predictors. The various methods for estimating sufficient summary plots can be used to detect such dependence by simply using residuals in place of the response. In this setting the model is sustained when no information is found to contradict 0D structure.

A plot of residuals vs. fitted values from the ordinary least squares fit of model (1) is shown in Figure 4. The horizontal line intersecting zero on the vertical axis was added as a visual reference. If the model was correct, this plot should leave the impression that the residuals are independent of the fitted values. Because the residuals and fitted values in the plot are clearly dependent, we have information to contradict 0D structure and the model is called into question.

CERES plots, including the class of partial residual plots can be used to study the need for predictor transformations in a multiple regression model such as model (1). Similar graphical methods are available for studying the need for response transformations. Net-effect plots can be used to study the contribution of a selected predictor to the regression. Versions are available for both model-based and model-free regression studies. Marginal model plots can be used to assess model adequacy.

Further Reading

The approach to regression graphics based on sufficient summary plots and many special purpose graphics were discussed at a relatively low mathematical level by Cook and Weisberg (1999). A more advanced treatment that requires knowledge of linear models and linear algebra was given by

Cook (1998a). Sliced inverse regression and the method of principal Hessian directions were proposed by Li (1991, 1992). Sliced average variance estimation was proposed by Cook and Weisberg (1991). Extensions of these methods were studied by Cook (1998b), Ferré (1998), Filliben and Li (1997), Hsing and Carroll (1992), Schott (1994), Velilla (1998), and Zhu and Fang (1996). Graphical regression was proposed by Cook (1998a). The mussel data were taken from Camden (1989). The bank note data were taken from Flury and Riedwyl (1988); the analysis follows Cook and Lee (1999) who studied regressions with a binary response.

See also: Linear Hypothesis: Fallacies and Interpretive Problems (Simpson's Paradox); Linear Hypothesis: Regression (Basics); Linear Hypothesis.

Bibliography

Camden, M., 1989. The Data Bundle. New Zealand Statistical Association, Wellington, New Zealand.

Cohen, J., Cohen, P., West, S.G., Aiken, L.S., 2003. Applied multiple regression/correlation analysis for the behavioral sciences. L. Erlbaum Associates, Mahwah, NJ.

Cook, R.D., 1998a. Regression Graphics: Ideas for Studying Regressions through Graphics. Wiley, New York.

Cook, R.D., 1998b. Principal Hessian directions revisited (with discussion). Journal of the American Statistical Association 93, 84–94.

Cook, R.D., Lee, H., 1999. Dimension reduction in binary response regression. Journal of the American Statistical Association 94, 1187–1200.

Cook, R.D., Weisberg, S., 1991. Sliced inverse regression dimension reduction comment. Journal of the American Statistical Association 86, 328–332.

Cook, R.D., Weisberg, S., 1999. Applied Regression Including Computing and Graphics. Wiley, New York.

Cook, R Dennis, Li, Bing, 2002. Dimension Reduction for Conditional Mean in Regression. Annals of Statistics 30 (2), 455–474.

Ferre, L., 1998. Determining the dimension of sliced inverse regression and related methods. Journal of the American. Statistical Association 93, 132–140.

Filliben, J.J., Li, K.C., 1997. A systematic approach to the analysis of complex interaction patterns in two-level factorial designs. Technometrics 39, 286–297.

Flury, B., Riedwyl, H., 1988. Multiariate Statistics: A Practical Approach. Chapman and Hall, London.

Hsing, T., Carroll, R.J., 1992. An asymptotic theory for sliced inverse regression. The Annals of Statistics 20, 1040–1061.

Li, K.C., 1991. Sliced inverse regression for dimension reduction (with discussion). Journal of the American Statistical Association 86, 316–327.

Li, K.C., 1992. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. Journal of the American Statistical Association 87, 1025–1039.

Sarkar, Deepayan, 2008. Lattice: multivariate data visualization with R. Springer, New York.

Schott, J.R., 1994. Determining the dimensionality in sliced inverse regression. Journal of the American Statistical Association 89, 141–148.

Velilla, S., 1998. Assessing the number of linear components in a general regression problem. Journal of the American Statistical Association 93, 1088–1098.

Zhu, L.X., Fang, K.T., 1996. Asymptotics for kernel estimates of sliced inverse regression. The Annals of Statistics 24, 1053–1068.

Linear Mixed Models

Garrett M Fitzmaurice and Nan M Laird, Harvard School of Public Health, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The focus of this article is on a class of regression models that are widely used for the analysis of cluster-correlated data. Linear mixed models are a natural extension of classical linear regression models that allow for the incorporation of random effects to account for the correlation among repeated measures on the same individual or cluster. Linear mixed models are used extensively for the analysis of longitudinal data as well as cluster-correlated data arising from other types of study designs. This article describes a general linear mixed model for clustered continuous data, outlining the model development and its underlying assumptions. Many of the conceptual ideas that are fundamental to understanding mixed models are reinforced using graphical representations of the linear mixed model equations. Finally, an example, based on data from a longitudinal study of growth in body fat in girls before and after menarche, is used to illustrate the application of linear mixed models.

This article discusses a class of regression models that are widely used for the analysis of cluster-correlated data. Linear mixed models are used extensively in settings where multiple correlated measurements are made on the sampling unit of interest (e.g., repeated measurements on a subject over time or observations of subjects nested within larger units, such as employees within workplaces, patients within medical clinics). In the social and behavioral sciences, these models are routinely applied to data from longitudinal studies and to cluster-correlated data arising from other types of study designs (e.g., measurements of children who are nested within different classrooms). This article presents the essential features of a general linear mixed model for longitudinal and cluster-correlated Gaussian data. In Section Longitudinal and Clustered Data we describe a number of study designs that give rise to data with outcomes measured repeatedly on each subject or cluster, highlighting some of the potential complexities for analysis. In Section Linear Mixed Models the main features of the linear mixed model, and its underlying assumptions are discussed. Key ideas are reinforced using graphical representations of the mixed model equations. In Section Example: MIT Growth and Development Study the model and methods are illustrated through an example using data from the MIT Growth and Development Study, a longitudinal study of growth in body fat in girls before and after menarche.

Introduction

Linear mixed models can be considered an extension of classical linear regression models that allow for the incorporation of random effects to account for the correlation among repeated measures on the same subject or cluster. Mixed models are widely used in the analysis of longitudinal and cluster-correlated data. Recall that in a longitudinal study, measurements of the outcome on the same individuals are taken repeatedly over time. A distinctive feature of longitudinal data is that they are *clustered*. In a longitudinal study the cluster is composed of repeated measurements obtained from a single individual. There are many study designs that are not longitudinal but give rise to data with outcomes measured repeatedly on each *cluster*. For example, clustered data commonly arise when intact groups are randomized to interventions or when naturally occurring groups in the population are randomly sampled. An example of the former is group-randomized trials. In a group-randomized trial, also known as a cluster-randomized trial, groups of individuals, rather than each individual alone, are randomized to different treatments or interventions. Data on the outcome of interest are obtained on all individuals within a group. Alternatively, clustered data can arise from random sampling of naturally occurring groups in the population. Families, households, hospital wards, medical practices, neighborhoods, and schools are all instances of naturally occurring clusters in the population that might be the primary sampling units in a study. Also, there are certain study designs that are both clustered and longitudinal, for example, multicenter longitudinal clinical trials where subjects, nested within clinics, are measured repeatedly over time.

In all these examples where clustered data arise, we might reasonably expect that measurements on units within a cluster are more similar than measurements on units in different clusters. That is, observations between distinct clusters are independent of one another, but observations within a cluster are emphatically not. The degree of clustering can be expressed in terms of correlation among the measurements on units within the same cluster. Importantly, this correlation invalidates the crucial assumption of independence that is the cornerstone of so many standard statistical techniques. Instead, statistical models for clustered data must explicitly account for the intracluster correlation. One popular method of accounting for the correlation is via the incorporation of random subject or cluster effects into the model. Specifically, the linear mixed model is an extension of the classical linear regression model for cross-sectional data in which random effects are introduced in the model. The random effects can be thought of as underlying (or latent) individual or cluster effects that account for the correlation among repeated measures on the same individual or cluster. The resulting regression model is a 'mixed model,' in the sense that it combines the usual fixed effects or population regression parameters with random effects (i.e., models that contain both fixed and random effects are called *mixed models*).

Mixed models for continuous normal outcomes have been extensively developed since the seminal papers by Harville (1977) and Laird and Ware (1982). The incorporation of random effects offers several benefits when modeling clustered data. First, it provides great flexibility in modeling the correlation in *unbalanced* designs, for example, in longitudinal studies where measurements are made at arbitrary, irregularly spaced intervals. Second, random effects can be used to estimate or predict subject-specific or cluster-specific effects. Finally, it provides an optimal way to combine the information from two distinct sources of variability in the outcome and covariates: within-subject (or within-cluster) and between-subject (or between-cluster) variability. In this article we review the main features of linear mixed models. To fix ideas, we first introduce a simple special case of the linear mixed model with a single random effect. Later, we describe a general linear mixed model for correlated data with multiple random effects and discuss its underlying assumptions. The model and methods are illustrated through an example from a longitudinal growth study. Before discussing linear mixed models, we highlight a number of study designs where clustered data commonly arise.

Longitudinal and Clustered Data

In this section we briefly discuss various types of study designs that give rise to data with outcomes measured repeatedly on each subject and highlight some of the potential complexities for analysis. Longitudinal study designs are widely used in the social and behavioral sciences to measure intraindividual change over time. In longitudinal studies subjects are measured repeatedly throughout the duration of the study, thereby permitting the direct assessment of changes in the outcome variable over time. Longitudinal studies are popular because many factors, which can confound age effects estimated in cross-sectional studies, are automatically controlled for by measuring change within-subjects. The main goal of a longitudinal study is to characterize within-individual changes in the outcome over time and to determine whether they are related to selected covariates. Often we wish to study change as a function of an intervention (e.g., drug treatment), an event (e.g., onset of puberty), or individual characteristics (e.g., gender, exposures). Typically, although not always, longitudinal study designs call for a fixed number of repeated measurements to be made on all study participants at a set of common time points. However, it is an almost inescapable feature of longitudinal studies in the social and behavioral sciences, especially those where the repeated measurements extend over a relatively long duration, that some individuals will miss their scheduled visit or date of observation. In some studies this may necessitate that observations be made some time before or after the scheduled time. Consequently the sequence of observation times is no longer common to all individuals in the study due to mistimed and/or missing measurements. In that case we refer to the data as being 'unbalanced' over time, that is, the repeated measurements are not obtained at a common set of occasions. As noted earlier, a distinctive feature of longitudinal data is that they are clustered. In longitudinal studies the clusters are composed of the repeated measurements obtained from a single individual at different occasions. That is, subjects can be thought of as the basic sampling unit, and the outcome is measured repeatedly on subjects.

There are numerous study designs that produce clustered data. Data of this kind frequently arise in the social and behavioral sciences because individuals can be grouped in many different ways. For example, in studies of educational performance, assessments are often obtained from children who are nested within different classrooms. Such data are clustered, with classroom considered to be the sampling unit and the repeated observations are the children within the classroom. Data from studies of nuclear families are also clustered, with observations on the mother, father, and children nested within families. Other common examples of naturally occurring clusters in the population are households, neighborhoods, districts, workplaces, departments, clinics, and litters. In many (although not all) of these study designs that produce clustered data the main interest is in the effects of covariates that vary between-clusters (i.e., between-cluster effects); this is in contrast to longitudinal designs, where interest is in within-subject (within-cluster) effects. Also, we note that clustering can occur at numerous levels in a data structure. For example, observations may be obtained on patients nested within clinics, which in turn are nested within different geographical regions of the country. Another example is when observations are obtained on children nested within classrooms, nested within schools. Regardless of the exact nature of the clustering, its consequence for analysis is the same: measurements on units within a cluster tend to be more similar (i.e., positively correlated) than the measurements on units in different clusters. In general, the degree of clustering can be expressed in terms of correlation among the observations on units within a cluster. To make valid inferences, statistical models for clustered data must appropriately account for this intracluster correlation. The magnitude of the positive intracluster correlation has important implications for precision. In longitudinal studies where the focus is on within-subject effects, the positive correlation among repeated measures allows for the estimation of the effects of change over time with greater precision (smaller standard error). In contrast, the effects of covariates that vary between-clusters are estimated with less precision (larger standard error) due to the positive intracluster correlation. For ease of exposition, in the remainder of this article we focus on applications of linear mixed models to longitudinal data, but it should be apparent how the models can also be applied to cluster-correlated data more broadly defined. Before turning to model development, we introduce an example that will be used later to illustrate an application of linear mixed models.

The MIT Growth and Development Study (Bandini et al., 2002; Phillips et al., 2003; Naumova et al., 2001; Fitzmaurice et al., 2011) was an observational study designed to characterize how measures of body fat change in girls before and after the onset of menarche. While it is accepted that body fat increases about the time of menarche, of interest was whether the increase persisted up to 4 years then. Premenarche, non-obese girls were selected in the study and measured annually until 4 years after menarche, or the end of follow-up. At each annual assessment, a measure of percent body fat was obtained based on bioelectric impedance analysis, a widely used method

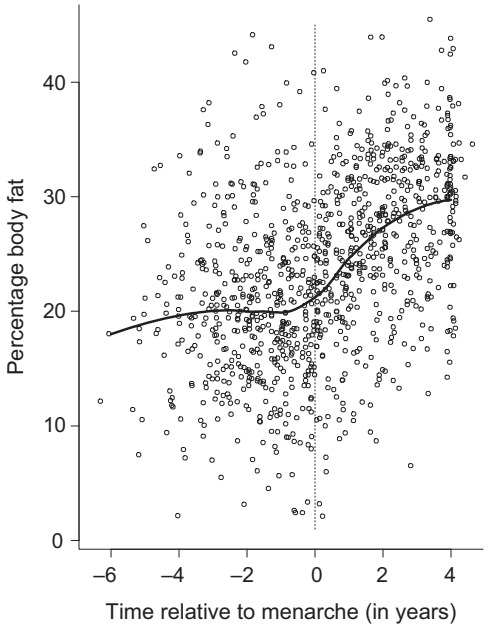

Figure 1 Scatter plot of percent body fat against time, relative to age of menarche (in years), with *lowess* smoothed curve.

for estimating body composition. The time of menarche was also recorded for each participant. While the design is almost balanced with respect to calendar time (with the exception of the final measurement, the design is balanced), what is of greater relevance to the scientific question of interest is the time relative to menarche. That is, for the purpose of analysis, the timing of measurement must be rescaled so that a value of zero corresponds to menarche, a positive value indicates time after menarche, and a negative value denotes time before menarche. Since menarche is a random event, differing for each girl, the resulting sample design is quite unbalanced. Although most girls had four measurements after menarche, the number of measures taken before was not fixed by design. Figure 1 illustrates all of the data points plotted against time relative to menarche. Note the cluster of measures at 4 years post-menarche is by design, but the other measures are scattered at random times. The trend in the mean response over time can be assessed using a *lowess* smoothed curve. Lowess is a nonparametric, robust regression method that traces the salient features of the mean response as a function of time while making only minimal assumptions about the form of the relationship. The lowess curve in Figure 1 reveals that the mean response remains relatively flat during the premenarcheal period and then rises sharply during the postmenarcheal period. We will return to this example in Section Example: MIT Growth and Development Study.

Linear Mixed Models

We first introduce some notation that will be adopted for the remainder of the article. Laird and Ware (1982) established a notation that is very commonly used in longitudinal studies, where Y_{ij} denotes the outcome measurement taken on the ith subject at the jth occasion. Notice that in balanced designs, we could choose to let j index time itself (e.g., $j = 1$ denotes time 1, $j = 2$ denotes time 2, etc.), but this is not workable in unbalanced designs where different individuals are measured at different times. Hence, we let j index measurement number rather than time or occasion; we use the terms time or occasion somewhat interchangeably to denote the specification of measurements in the design. Thus j varies from 1 to n_i, where n_i is the total number of observations obtained on the ith subject. Adopting this notation means that it will be necessary to explicitly introduce time of measurement in the models for the mean response and the covariance among repeated measurements. However this provides us with a flexible notation for handling missing data and unbalanced designs.

Before we discuss the general linear mixed model for Y_{ij}, we first consider the simplest possible case of a linear mixed model, a model with randomly varying subject effect or random intercept:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + b_{0i} + \varepsilon_{ij} \\ &= (\beta_0 + b_{0i}) + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + \varepsilon_{ij} \end{aligned} \quad [1]$$

where the βs are the fixed regression coefficients, the X_{ij}s are vectors containing the p covariates for the ith subject at the jth occasion, b_{0i} is the random subject effect (or random intercept), and ε_{ij} is a random error term. In this model b_{0i} and ε_{ij} are assumed to be mutually independent, with mean zero, and with variances, $\mathrm{Var}(b_{0i}) = \sigma_b^2$ and $\mathrm{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2$. Because the mean of the random subject effect is assumed to be zero, b_{0i} represents the deviation of the ith individual's intercept $(\beta_0 + b_{0i})$ from the population intercept, β_0. The assumption of zero mean for the random effect also implies that an ordinary linear regression model for the mean of Y_{ij} holds:

$$E(Y_{ij}|X_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} \quad [2]$$

when averaged over the distribution of the random effect. That is, the linear mixed model for Y_{ij} in eqn [1] also implies a linear regression model for the marginal mean (averaging over the random effect) of Y_{ij} in eqn [2]. Here the βs model the so-called *fixed effects*, that is, the time effects, as well as the effects of interventions or covariates, and their interactions.

For this simple case of a linear mixed model the fundamental ideas can be best understood by considering the graphical representation of the model equations for the simple case where a subject's expected response varies linearly over time. Figure 2 displays how the marginal mean response over time in the population changes linearly with time (denoted by the solid line), but also indicates how the subject-specific mean response for two individuals (denoted by the broken lines), say subjects A and B, deviate from the population trend. In Figure 2, individual A responds 'higher' than the population average and thus has a positive b_{0i}. In contrast, individual B responds 'lower' than the population average and has a negative b_{0i}. The inclusion of the errors, ε_{ij}, allows the response at any occasion to vary randomly above and below the subject-specific trajectories.

In general, the random effect, b_{0i} can be interpreted in several ways. As described above, it provides a measure of how the ith subject deviates from the marginal mean. Alternatively, we can think of b_{0i} as representing the ith subject's underlying

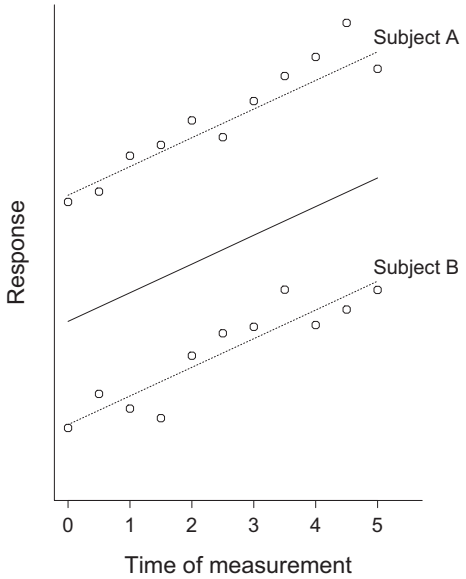

Figure 2 Graphical representation of the marginal (averaged over subjects) and subject-specific mean response over time, plus errors.

(or latent) level of response that persists over time and influences all repeated measurements on that subject, thereby providing a device for introducing correlation among the repeated measures. The latter follows from the fact that any two observations on the ith subject, say Y_{ij} and Y_{ik}, will be correlated because they share the same random effect, b_{0i}.

Next, consider the marginal covariance (variances and correlations) among the repeated measurements. For the simple model with random intercepts, the marginal variance at each occasion is simply $\mathrm{Var}(Y_{ij}) = \sigma_b^2 + \sigma_\varepsilon^2$. This expression implies that $\mathrm{Var}(Y_{ij})$ is constant over time. The correlation between any pair of repeated measurements, say Y_{ij} and Y_{ik}, is

$$\mathrm{Corr}\left(Y_{ij}, Y_{ik}\right) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\varepsilon^2}$$

This implies that all pairwise correlations are the same, regardless of the time separation between any pair of repeated measurements. One justification for this correlation structure is it assumes that conditional on a specific subject, repeated observations taken on that subject are independent. This may be reasonable in many cluster-correlated data settings. In these settings, one can consider the within-cluster subjects to be exchangeable. That is, all subjects have the same level of variability and the correlation for all pairs of subjects is the same. For example, this assumption is reasonable when the data are clustered, with classroom considered to be the sampling unit and the repeated observations are the children within the classroom. However, it is rarely useful in longitudinal studies because the assumption that variances and correlations are constant throughout the period of the study is generally unrealistic. Some degree of heterogeneity of the variability over time is not uncommon in longitudinal studies. In addition, ordinarily the pairwise correlations among repeated measures tend to decrease with increasing time separation.

More useful structures for the variances and correlations among the repeated measurements are obtained by enriching the model with multiple random effects. One popular approach is to assume each individual has her own subject-specific intercept and slope. If we assume a linear trend to model change over time, the corresponding mixed model is given by:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij} \\ &= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) t_{ij} + \varepsilon_{ij} \end{aligned} \qquad [3]$$

where t_{ij} denotes the time of the jth measurement on the ith subject, the random effect, b_{0i} represents the deviation of the ith subject's intercept from the population intercept, β_0, and the random effect, b_{1i} represents the deviation of the ith subject's slope from the population slope, β_1. A graphical representation of the model equations is displayed in Figure 3. Figure 3 illustrates how the marginal mean response in the population changes linearly with time (denoted by the solid line), but also indicates how the subject-specific mean responses (denoted by the broken lines) for two specific individuals, say subjects A and B, deviate from the population trend. In this simple illustration, individual A has a 'higher' baseline level of response (when $t_{ij} = 0$) than the population average and thus has a positive b_{0i}. In contrast, individual B has a 'lower' baseline level of response than the population average and thus has a negative b_{0i}. In addition individual A has a steeper rate of increase over time than the population average and thus has a positive b_{1i}. Individual B has a less steep rate of increase over time than the population average and thus has a negative b_{1i}.

The inclusion of both randomly varying intercepts and slopes has important ramifications for the variances and correlations among the repeated measurements. The variance of the response at the jth occasion is

$$\mathrm{Var}\left(Y_{ij}\right) = \mathrm{Var}(b_{0i}) + 2t_{ij}\mathrm{Cov}(b_{0i}, b_{1i}) + t_{ij}^2\mathrm{Var}(b_{1i}) + \mathrm{Var}\left(\varepsilon_{ij}\right)$$

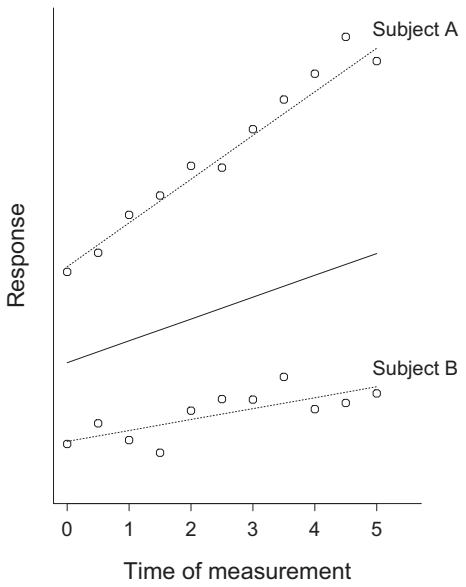

Figure 3 Graphical representation of the marginal (averaged over subjects) and subject-specific mean response over time, plus errors.

Similarly, it can be shown that the covariance between any pair of repeated measurements, say, Y_{ij} and Y_{ik} obtained at times t_{ij} and t_{ik}, respectively, is

$$\mathrm{Cov}(Y_{ij}, Y_{ik}) = \mathrm{Var}(b_{0i}) + (t_{ij} + t_{ik})\mathrm{Cov}(b_{0i}, b_{1i}) + t_{ij}t_{ik}\mathrm{Var}(b_{1i})$$

By incorporating both randomly varying intercepts and slopes, the variances and correlations are functions of time, t_{ij}. Specifically, the variance can increase or decrease over time as a quadratic function of the times of measurement. Similarly the magnitude of the covariance (and correlation) between a pair of responses, say Y_{ij} and Y_{ik}, depends on the time separation between them (t_{ij} and t_{ik}).

In the simple illustration above there are randomly varying intercepts and slopes. However, the linear mixed effects model can be generalized to incorporate additional randomly varying regression coefficients and to allow the means of the random effects to depend on covariates (e.g., interventions, gender). To handle this additional complexity, a general linear mixed model for Y_{ij} can be written as

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + b_{0i} + b_{1i} Z_{ij1} + b_{2i} Z_{ij2} + \cdots + b_{qi} Z_{ijq} + \varepsilon_{ij} \quad [4]$$

where the βs are the fixed regression coefficients, the X_{ij}s are the p covariates for the ith subject at the jth occasion, the b_is are the q random effects for the ith subject, the Z_{ij}s are the corresponding covariates (or design) for the random effects, and ε_{ij} is a random error term. Several important assumptions are commonly made for the linear mixed model. First, it is assumed that the collection of b_is has a multivariate normal distribution, with zero mean, and unstructured covariance matrix (i.e., arbitrary variances and pairwise correlations among the random effects). It is standard to assume the random effects have a normal distribution, although, in principle, other distributional assumptions could be made depending on the application (see, for example, Lee and Nelder, 1996). It is also assumed that the errors, ε_{ij} have a normal distribution with zero mean and that ε_{ij} and the b_is are mutually independent of each other. As discussed above, the assumption of zero means for the random effects implies that a linear regression model for the marginal mean (averaging over the random effects) of Y_{ij} holds:

$$E(Y_{ij}|X_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} \quad [5]$$

For the remainder, we focus on the specification of the random error terms, since modeling the mean response is not fundamentally different from choosing an ordinary linear regression model for the mean response over time. A consequence of the zero mean assumption for the random effects is that ordinarily the Z_{ij}s must be a subset of the X_{ij}s. The reason for this is not immediately obvious, but stems from the fact that $b_{0i} + b_{1i}Z_{ij1} + b_{2i}Z_{ij2} + \cdots + b_{qi}Z_{ijq}$ can then be usefully interpreted as an 'error' defined with respect to the marginal mean curve defined by eqn [5].

In principle, the random effects part of the model can be made as complex as desired through the inclusion of additional Z_{ij}s and random effects to the model. For example, Donnelly et al. (1995) discuss a model for longitudinal growth in lung function where the subject-specific trajectories depend not only on age of measurement, but also on height at the jth measurement. However, in practice, some care must be exercised in specifying random effects in the model because the variance components parameters may be poorly estimated if the model is too large. An alternative approach is to keep the random effects part of the model simple (e.g., random intercepts) but incorporate serial correlation structure in the errors, ε_{ij} (e.g., autoregressive correlated errors); see, for example, Diggle (1988) and Diggle et al. (2002). In general, it is not possible to enrich both the complexity of the random effects and the correlation structure of the errors as this can lead to identifiability problems due to an overparameterized covariance matrix for the repeated measurements. For example, in many longitudinal studies there may be insufficient information in the data at hand to support separate estimation of randomly varying intercepts and slopes in addition to a complex serial correlation structure (e.g., autoregressive) for the within-subject errors.

Finally, we note that in many applications where linear mixed models are applied, inference is focused on the βs, the fixed effects. However, in some applications, we may want to estimate the subject-specific effects. For example, in studies of growth it may be of interest to obtain subject-specific growth trajectories. In other types of longitudinal studies, it may be of interest to identify those individuals who showed the greatest increase or decrease in the outcome over time. Because the linear mixed model explicitly distinguishes between fixed and random effects, it is possible to estimate (or predict) the subject-specific effects, the b_is, or the subject-specific response trajectories $(\beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + b_{0i} + b_{1i}Z_{ij1} + b_{2i}Z_{ij2} + \cdots + b_{qi}Z_{ijq})$. In general, the statistical problem of predicting a random variable can be shown to be that of predicting its conditional mean, given the available data. Thus the best predictor of the b_is is their conditional mean, given the collection of repeated measures Y_{ij} (for $j = 1,\ldots,n_i$) on the ith subject. Under the assumptions of the linear mixed model, this is known as the 'best linear unbiased predictor' (BLUP). We note that the BLUP estimators of the b_is are often called 'shrinkage' estimators because they tend to 'shrink' the predictions of the b_is toward zero (the population mean of the random effects). The amount of shrinkage depends on the relative magnitudes of the within-subject and between-subject sources of variability. In general, when the within-subject variability is relatively larger than the between-subject variability there can be discernible shrinkage. Conversely, when the between-subject variability is relatively larger than the within-subject variability, there is far less shrinkage. In addition, the amount of shrinkage depends also on n_i, the number of observations on the ith subject. In general, there is more shrinkage when n_i is small. Intuitively, this is quite sensible since less weight should be given to the ith subject's observations when fewer are available.

Example: MIT Growth and Development Study

To highlight some of the main ideas discussed so far, we consider a longitudinal analysis of the data on percent body fat from the MIT Growth and Development Study. Recall from Figure 1 that the lowess curve reveals that the mean response remains relatively flat during the premenarcheal period and

then rises sharply during the postmenarcheal period. In the following analysis we consider the hypothesis that percent body fat increases linearly with age, but with different slopes before and after menarche. A linear mixed model approach was taken to fit a linear spline, with a knot at time zero (here denoting menarche) to the measurements of percent body fat over time. Recall that the main question of interest was whether the slope on time changes after menarche, hence using a linear spline makes it easy to address this question. Specifically, we assumed that each girl has a piecewise linear spline growth curve with a knot at the time of menarche. That is, each girl's growth curve can be described with an intercept and two slopes, one slope for the changes in percent body fat before menarche, another slope for the changes in percent body fat after menarche. We fit the following linear mixed model using (restricted) maximum likelihood estimation, assuming the distribution of the random terms is normal:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij})_+ + b_{0i} + b_{1i} t_{ij} + b_{2i} (t_{ij})_+ + \varepsilon_{ij}$$

where t_{ij} denotes the time, relative to menarche (i.e., before or after menarche), of the jth measurement on the ith subject (with $t_{ij} = 0$ at menarche), and $(t_{ij})_+ = t_{ij}$ if $t_{ij} > 0$ and $(t_{ij})_+ = 0$ if $t_{ij} \leq 0$. The three random effects are assumed to have a multivariate normal distribution, with zero mean, and unstructured covariance matrix.

Table 1 presents the fixed effects estimates of the overall mean curve, the intercept, and two slopes, before (β_1) and after menarche ($\beta_1 + \beta_2$). Notice that this approach allows us to estimate mean percent body fat at menarche (when $t_{ij} = 0$), β_0, even though few if any girls were actually measured then. From Table 1 the model-based estimate of the mean percent body fat at menarche is approximately 21%. Regarding the main question of scientific interest, based on the magnitude of the estimate of β_2, relative to its standard error, there is a significant difference between the slopes before and after menarche. The estimate of the population mean premenarcheal slope is 0.42 and indicates that the annual rate of body fat growth is less than 0.5%. In contrast, the estimate of the population mean postmenarcheal slope is 2.46(0.417 + 2.047) and indicates that the annual rate of body fat growth is approximately 2.5%, almost six times higher than the corresponding rate in the premenarcheal period.

The covariance matrix for the random effects has six parameters, three parameters for the variances of the random intercepts and slopes, and three parameters for their covariances. Estimates of the variance and covariance parameters are given in Table 2. Much of the subject-to-subject variability is contained in the intercept. The covariance between a subject's intercept and postmenarche slope is negative, suggesting that girls with higher body fat at menarche, tend to gain less after menarche. Results from the mixed model analysis also allow us to estimate values of the variances and covariances for measures of percent body fat at arbitrary times; this is illustrated in Table 3 for 4 years pre- and postmenarche. Here we see a decline in variance over time; in addition, the strength of the correlation declines with increasing time separation. The latter feature of the correlation is not uncommon in longitudinal studies.

Table 1 Estimated linear regression coefficients (fixed effects) and standard errors for the percent body fat data

Variable	*Estimate*	*SE*	Z
Intercept	21.361	0.565	37.84
Time	0.417	0.157	2.65
$(\text{Time})_+$	2.047	0.228	8.98

Table 2 Estimated variances and covariances of the random effects and standard errors for the percent body fat data

Parameter	*Estimate*	*SE*
$\text{Var}(b_{0i})$	45.941	5.739
$\text{Var}(b_{1i})$	1.631	0.433
$\text{Var}(b_{2i})$	2.750	0.964
$\text{Cov}(b_{0i}, b_{1i})$	2.526	1.219
$\text{Cov}(b_{0i}, b_{2i})$	−6.110	1.873
$\text{Cov}(b_{1i}, b_{2i})$	−1.751	0.598
$\text{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2$	9.473	0.544

Table 3 Estimated marginal correlations (on the lower diagonals) among repeated measures of percent body fat between 4 years pre- and postmenarche, with estimated variances along the main diagonal

Time in years (relative to menarche)								
−4	*−3*	*−2*	*−1*	*0*	*1*	*2*	*3*	*4*
61.3								
0.82	54.9							
0.78	0.81	51.8						
0.71	0.76	0.80	52.0					
0.61	0.70	0.76	0.81	55.4				
0.60	0.68	0.74	0.79	0.81	49.1			
0.57	0.64	0.71	0.76	0.78	0.79	44.6		
0.52	0.60	0.66	0.71	0.73	0.76	0.77	41.8	
0.47	0.54	0.60	0.64	0.66	0.70	0.74	0.76	40.8

Summary

Linear mixed models are a natural extension of classical linear regression models that allow for the incorporation of random effects to account for the correlation among repeated measures on the same individual or cluster. The widespread use of mixed models in so many areas of application in the social and behavioral sciences can be explained in part by the great flexibility they offer in modeling the correlation in *unbalanced* designs and by the increasing availability of statistical software for implementing mixed model methods. This article has described a general linear mixed model for longitudinal and cluster-correlated Gaussian data. When the outcome is discrete (e.g., binary, ordinal, or a count) rather than continuous, linear mixed models are no longer appropriate for relating the mean response to covariates. Instead, there are extensions of generalized linear models (e.g., logistic regression) for analyzing cluster-correlated discrete data known as generalized linear mixed models (e.g., *see* Generalized Linear Mixed Models). These models account for the correlation among the repeated

measures via the incorporation of random effects into generalized linear models (e.g., logistic regression models with randomly varying intercepts). Generalized linear mixed models extend in a natural way the conceptual approach represented by the linear mixed models discussed in this article.

See also: Generalized Linear Mixed Models; Hierarchical Models: Random and Fixed Effects.

Bibliography

Bandini, L.G., Must, A., Spadano, J.L., Dietz, W.H., 2002. Relation of body composition, parental overweight, pubertal stage, and race-ethnicity to energy expenditure among premenarcheal girls. American Journal of Clinical Nutrition 76, 1040–1047.

Diggle, P.J., 1988. An approach to the analysis of repeated measurements. Biometrics 44, 959–971.

Diggle, P.J., Heagerty, P.J., Liang, K.-Y., Zeger, S.L., 2002. Analysis of Longitudinal Data, second ed. Clarendon Press, Oxford.

Donnelly, C.A., Laird, N.M., Ware, J.H., 1995. Prediction and creation of smooth curves for temporally correlated longitudinal data. Journal of the American Statistical Association 90, 984–989.

Fitzmaurice, G.M., Laird, N.M., Ware, J.H., 2011. Applied Longitudinal Analysis, second ed. Wiley, New Jersey, NJ.

Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 72, 320–338.

Laird, N.M., Ware, J.H., 1982. Random effects models for longitudinal data. Biometrics 38, 963–974.

Lee, Y., Nelder, J.A., 1996. Hierarchical generalized linear models (with discussion). Journal of the Royal Statistical Society, Series B 58, 619–678.

Naumova, E.N., Must, A., Laird, N.M., 2001. Evaluating the impact of "critical periods" in longitudinal studies of growth using piecewise mixed effects models. International Journal of Epidemiology 30, 1332–1341.

Phillips, S.M., Bandini, L.G., Compton, D.V., Naumova, E.N., Must, A., 2003. A longitudinal comparison of body composition by total body water and bioelectrical impedance in adolescent girls. Journal of Nutrition 133, 1419–1425.

Linguistic Anthropology

Marcel Danesi, Victoria College, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Duranti, volume 13, pp. 8899–8906, © 2001, Elsevier Ltd.

Abstract

Linguistic anthropology is a branch of both anthropology and linguistics, aiming to document and study the use of language as a social, cognitive, communicative, and cultural tool. The viewpoint that guides most of the work in the field is the premise that language structures reflect cultural concepts which, in turn, reflect social and communicative structures. This article presents the background and basic notions of linguistic anthropology and then looks at its areas of research including ethnosemantics, conversation, texts, and dialects, among others.

Introduction

Linguistic anthropology (LA) is an approach to the study of language that focuses on the relation between language, society, and culture. It is considered by some to be a branch of general linguistics, by others a branch of anthropology, and by still others as an autonomous discipline. Its roots are in the foundation of linguistics as a science in the nineteenth century after English scholar Sir William Jones (1786) remarked that Sanskrit, Persian, Greek, and Latin sprang from the same source and thus belonged to the same 'language family.' This suggested that languages shared structural features and evolved over time from older forms to develop their own distinct forms. A full-fledged science of language emerged after Swiss philologist Ferdinand de Saussure (1916) provided a theoretical framework for studying language as a system in his book, titled *Cours de linguistique générale*, which was published posthumously from the notes taken by his students at the University of Geneva.

Saussure's framework came under the rubric of structuralism (paralleling developments in psychology and other emerging human and social sciences at the time). In Europe, it was adopted and elaborated by a number of linguists who congregated in the Czech city of Prague, known as the Prague Circle, in the 1920s and 1930s. Structuralist linguistics became the basis for studying indigenous North American languages at Columbia University in the first two decades of the twentieth century under the leadership of Franz Boas (1940), who founded the *International Journal of American Linguistics* to publish research on the languages native to North, Central, and South America in 1917. With the foundation of the Linguistic Society of America in 1924, and its journal, *Language*, a year later in 1925, linguistics developed into a flourishing discipline with a basis in anthropological methods. It received its first comprehensive treatment in America with Leonard Bloomfield's *Language* (1933).

In 1957, the American linguist Noam Chomsky argued that any serious scientific approach to language would have to focus on why all languages reveal a similar pattern for constructing their grammars, independently of culture. His perspective became a dominant one, and still has a large following. Chomsky's claim has always been that cultural and communicative phenomena are separate from strictly grammatical ones (Chomsky, 2002). Since the late 1960s, various arguments have come forward to challenge this stance, reinstating the original view of Boas and others that grammar develops in relation to communicative and cultural forces, not separately from them. This approach came to be called *anthropological linguistics* (to distinguish it from *theoretical linguistics*) and was renamed *linguistic anthropology* by Dell Hymes (1962). To this day, the two terms are used interchangeably, although the latter is now more widespread. LA is a thriving field today, sometimes overlapping with so-called *cognitive linguistics*, which focuses on the relation between language, cognition, and culture. The most prominent figure in the latter approach is the American linguist George Lakoff (1987). Lakoff's main argument is that the conceptual backbone of language is figurative and thus tied to the specific historical experiences and worldviews of the users of language. He cites, as an illustration, the emergence of grammatical gender categories in an indigenous Australian language called Dyirbal. In European languages, the gender of a noun is often unpredictable from its meaning and, thus, is typically thought to be arbitrary. For example, the word for 'table' is masculine in German (*der Tisch*), feminine in French (*la table*), and neuter in Greek (*to trapézi*). But in Dyirbal this is hardly the case. One of its four gender categories includes all nouns pertaining to *women*, all those pertaining to *fire*, and all those indicating things that are *dangerous* (snakes, stinging nettles, and the like). Gender assignment is culturally based, being determined by Dyirbal's cultural worldview.

Background

In the seventeenth and eighteenth centuries, the first surveys of languages were conducted in order to determine which features of language were universal and which were specific, independent of the cultures that spoke them. A group of French scholars, known as the Port Royal Circle, put forward the idea of a universal faculty of language that provided every different language with the same set of principles from which they could construct their individual grammars. Opposing this stance, the German scholar Wilhelm von Humboldt (1836) viewed language as springing from the historical needs of its speakers, conditioning how subsequent

generations came to view reality. An example of this is the phenomenon of color. The color names developed by a particular group of people are those that they have required for their own functional purposes. Given the fact that there are millions of visual discriminations that the eye can make, languages would have to otherwise come up with just as many words to encode them. So they name only those color discriminations that are useful and meaningful to them. It is true that there are patterns in how color categories have evolved across time and across cultures (Berlin and Kay, 1969), but this fact in no way opposes the fact that specific color vocabularies influence how people view the spectrum. Humboldt (1836: p. 43) summarized this line of reasoning as follows:

> The central fact of language is that speakers can make infinite use of the finite resources provided by their language. Though the capacity for language is universal, the individuality of each language is a property of the people who speak it. Every language has its *innere Sprachform*, or internal structure, which determines its outer form and which is a reflection of its speakers' minds. The language and the thought of a people are thus inseparable.

The disciplines of linguistics and anthropology have always overlapped considerably. Boas founded the first major department for the teaching of anthropology in 1899 at Columbia University in New York City, making linguistics the core of the discipline. The break between linguistics and anthropology came with Chomsky (1957). Boas amassed data on the Kwakiutl, a native society on the northwestern coast of North America, showing how the grammar and vocabulary of their language served their needs perfectly. Boas also introduced *ethnography* as the central method of linguistics. Ethnography is the study of some cultural phenomenon through systematic observation, consisting primarily in gathering empirical data by observing a group or society with interviews, questionnaires, and interaction.

Boas's student Edward Sapir (1921), and later Sapir's own student Benjamin Lee Whorf (1956), elaborated the Humboldtian view, suggesting that specific languages filter the reality perceived by their speakers, presenting them with differential cognitive slices of the same reality. Their approach generally falls under the rubric of *linguistic relativity*, the *Sapir–Whorf Hypothesis*, or simply the *Whorfian Hypothesis* (WH). The WH posits that languages predispose their speakers to attend to certain concepts (rather than others found in different languages) as being necessary, without blocking understanding between speakers of different languages. While this orientation is, and always has been, controversial, it has generated a lot of interesting findings, debates, and applications to the present day (Hoijer, 1954; Hill and Mannheim, 1992; Sidnell and Enfield, 2012; Danesi, 2013). Most importantly, it brought the powerful role of figurative language in cognition and culture to the forefront (Pollio et al., 1977; Ortony, 1979; Honeck and Hoffman, 1980; Lakoff and Johnson, 1980, 1999). Even a simple gustatory object like a piece of chocolate has symbolic status that is connected to love through the channel of cultural metaphors such as *love is a sweet taste*. These are also imprinted in common utterances such as *She's my sweetheart; They went on a honeymoon*; and so on.

Research Topics

The research agenda in LA is a vast and varied one. What ties all the work together is the focus on the language–culture nexus and in the many ways that it manifests itself in everyday social phenomena such as vocabulary choices and conversations. In the earliest research efforts, the focus was on how cultural classification and language interrelated because, as Sapir (1921: p. 75) clearly maintained, reality is organized cognitively by the specific categories of the language that speakers are familiar with:

> Human beings do not live in the object world alone, nor alone in the world of social activity as ordinarily understood, but are very much at the mercy of the particular language system which has become the medium of expression for their society. It is quite an illusion to imagine that one adjusts to reality essentially without the use of language and that language is merely an incidental means of solving specific problems of communication or reflection. The fact of the matter is that the "real world" is to a large extent unconsciously built up on the language habits of the group.

With different kinds of research paradigms and techniques, work in this area, known more specifically as *ethnosemantics*, continues to this day. For instance, John Lucy (1996) studied the effect of grammar on memory tasks on English and Yucatec (a Mayan language) speakers, focusing on the presence of a dichotomy between the two languages: English requires a plural marker for count nouns (*dog – dogs*); Yucatec does not, except for a small number of nouns. Lucy presented pictures of Yucatec village scenes to both speakers and asked them to perform recall tasks. He found that English speakers paid attention to the number of animate beings and objects, but ignored number for substances; Yucatec speakers paid attention to number only for animate beings. The experiment showed that grammatical categories did indeed condition how people recalled the world.

Overall, this line of inquiry has shown that, once classified, 'reality' is an ideological system passed on through language forms to subsequent generations. Of course, they can change their views of the world any time they want, by simply inventing new words for new realities that are meaningful to them. This is why languages change over time.

Identity, Socialization, and Ideology

Research in LA today has gone far beyond its ethnosemantic origins. One area concerns how language, socialization, and identity are interlinked. Don Kulick (1992), for example, conducted ethnographic work in a village in Papua New Guinea called Gapun, which spoke Taiap, a language known only to the villagers. He was able to document how that language, although spoken by a small group of people, constituted a source of pride, constituting an 'index' (a key) to identity and thus allowing the villagers to feel and claim autonomy from Tok Pisin speakers (the pidgin language spoken by others). Taiap allowed the villagers to express shared values, especially during ritualistic verbal performances that garnered a high level of emotional resonance.

Language as a socializing agent has always been at the core of LA. Prominent in this subfield is the work of Elinor Ochs and Bambi Schieffelin (Ochs and Schieffelin, 1984; Ochs, 1988; Schieffelin, 1990, 1995, 2000, 2002; Ochs and Taylor, 2001), who have looked extensively at how socialization is, essentially, a linguistic-adaptive process – allowing individuals to participate in a community and its culture in a meaningful way. Children are socialized into belief systems, historical content, and so on with and through oral narratives. In one study (Ochs and Taylor, 2001), it was shown that the narratives told (informally) at dinner time in white middle-class households in southern California typically presented role structures, gender distinctions, and other socializing and enculturation themes to family members.

Michael Silverstein's work on the role of language in the formation and expression of ideologies has made an impact within both LA and other branches (Silverstein, 1976, 1979, 1985; Rumsey, 1990; Kroskrity, 1998). Silverstein defines *ideology* as a shared body of commonsensical notions about the nature of the world. He rejects the Marxist view that ideologies create 'false consciousness' deriving from the perception in capitalist societies that material objects and institutions induce people to experience social relations in terms of the amount of capital they possess and the things they wish to purchase. Silverstein argues that, on the contrary, the form of language used by a society has more impact on ideological formations than does the socioeconomic system in place. Ideology is really 'ideation' or the formation of ideas connected to the semantic system of a language. The evolution of this system is not directly tied to political or economic forces but mirrors them, encodes them, and then filters perceptions about them. In a similar vein Alessandro Duranti (2003) has argued that language creates an 'exemplary center' or 'social space' of political and ritualistic power, not a hegemonic one. This space constrains the types of speech and relations that are relevant to a specific group. In it, refined speech, for example, is hardly part of political and economic ideology, but rather a tool in a broader system of socialization. At play in all this are semiotic forces, such as iconicity and indexicality. The former is essentially mimetic or imitative behavior that reflects a continuum or scale of prestige and power, and the latter is a relational form of behavior and understanding that reveals who a speaker is and what he or she aspires to be as well as which values he or she espouses.

Text Structure

Starting in the late 1960s, some researchers in LA started putting the focus on the role of conversations, discourses, and texts in shaping grammar. British linguist Michael Halliday (1985), for example, showed that some parts of speech are 'text-governed' structures, not sentence-based ones. Consider, for example, the following stretch of conversation, which has no pronouns in it. Even though it is completely understandable, we nevertheless perceive it as being awkward:

Speaker A: Alex is a nice young man.
Speaker B: Yes, Alex is a nice young man.
Speaker A: But Alex always likes to talk about Alex.
Speaker B: Yes, Alex does indeed always talk about Alex.

The more socially appropriate version of the conversation is one where pronouns are used instead:

Speaker A: Alex is a nice young man.
Speaker B: Yes, he is.
Speaker A: But he always likes to talk about himself.
Speaker B: Yes, he does indeed always talk about himself.

The use of pronouns is thus conversationally 'systemic'; it connects the various parts of the conversation, linking them logically like a trace device. In other words, pronoun choice is tied to text structure, not grammar, and is part of the style that is appropriate in a specific language. Some cultures encourage repetition; others, like English culture, do not. The repertoire of language forms available to a group of speakers thus does not form an arbitrary system of use; that is, the forms are not constructed artificially but in tandem with their use. As William Dwight Whitney emphasized, a precursor to this kind of approach in America (Shuy, 1997: p. 12), speakers of a language are extremely sensitive to socially coded textual cues, using them to evaluate other speakers and to present themselves in specific situations. In societies across the world, these cues are used for the whole array of social meanings that we see as critical.

Social Markers

The role of language in gender distinction is another major area of research – an area that has led to the entrenchment of the notion of *social markers*. In Japanese the word for 'stomach' is *hara* for men, but *onaka* for women, probably alluding to the fact that, biologically, the two stomachs are different. In Koasati (an indigenous language spoken in Louisiana), men say *lawawhol* to refer to 'lifting,' while women say *lakawhos*, perhaps implying that the lifting abilities of the two sexes are different. The point to note here is that the use of one or the other word according to the gender of the speaker is considered to be socially important. In some languages, such as the one spoken on the Island of Carib in the West Indies (Taylor, 1977), there exist a large number of prescribed doublets (pairs of words with the same meaning) that are marked for gender use. So, the word for 'sun' is *kachi* or *hueyu*, spoken respectively by the women and the men. Doublets are as important as differential modes of dressing, hairstyle, and other signifying resources that are used to keep the genders distinct socially.

Doublets are examples of 'social markers,' delineating those differences that are considered to be relevant or historically important to a community. Analogous markers are found across the whole spectrum of human relations, from politeness protocols to class structure. The misuse of one or the other would be perceived as a breach of social etiquette or as a sign of anomalous communication. It would be considered rude to address a superior at work with an informal mode of speech (unless the superior permits it); and it would be aberrant or strange to address a close friend with titles. Class differences, like gender ones, are mirrored in linguistic forms called *registers*. In the traditional Javanese society of Indonesia, for instance, different social classes are expected to utilize a distinct register of speech (Errington, 1988). At the top of the social hierarchy are the aristocrats; in the middle the townsfolk; and

at the bottom the farmers. The highly formal register is used by aristocrats who do not know one another very well, but can also be used by a member of the townsfolk if he or she happens to be addressing a high government official. The mid formal register is used by townsfolk who are not friends, and by farmers when addressing their social superiors. The low register is used by farmers among themselves, or by an aristocrat or town person talking to a farmer, and among friends on any social level. It is also the register used to speak to children of any class.

Change in social markers involves change in society as a whole. The use of the masculine gender in English was not only a simple grammatical device, but it reflected social inequalities. Terms like *chairman* and *spokesman* revealed how the English language predisposed its users to view certain social roles in gender terms. English grammar was organized from the perspective of those at the center of the society – the men. In matriarchal societies the reverse seems to be true. Investigating grammatical gender in the Iroquois language, Alpher (1987) found that in that language the feminine gender was the default one, with masculine items being marked by a special prefix. Alpher related this to the fact that Iroquois society is matrilineal. The women hold the land, pass it on to their heirs in the female line, are responsible for agricultural production, control the wealth, arrange marriages, and so on. Iroquois grammar is clearly organized from the viewpoint of those at the center of that particular society – the women.

Conversation and Discourse

Research interest within LA with conversations and discourse overlaps somewhat with the branch often called *pragmalinguistics* or *pragmatics*. Conversation analysis (CA) surfaced in the 1980s and 1990s, becoming a useful tool that both anthropological linguists and sociolinguists use commonly. CA aims to show that how people talk not only taps into a system of implicit social rules and patterns, but also shapes and changes the formal language system itself. Hutchby and Wooffitt (1998: p. 14) articulate the goal of CA as follows:

> CA is the study of recorded, naturally occurring talk-in-interaction. Principally, it is to discover how participants understand and respond to one another in their turns at talk, with a central focus being on how sequences of interaction are generated. To put it another way, the objective of CA is to uncover the tacit reasoning procedures and sociolinguistic competencies underlying the production and interpretation of talk in organized sequences of interaction.

The tacit reasoning procedures to which Hutchby and Wooffitt refer are manifestations of what Dell Hymes called *communicative competence* in 1971, a term that came forward to complement the notion of *linguistic competence* – a distinction that goes back to Saussure's (1916) notion of *langue* (linguistic competence) and *parole* (knowledge of how to use language). Hymes suggested that knowing how to use language during conversation is as systematic as knowing the rules of the grammar of the language being employed. Recalling the work of Charles Morris (1938), this line of inquiry shows that communication and grammar form an integrated system. In CA the target of analysis is to live among, or simply to interview (and/or record and tape), subjects in order to get not the singular object of speech (word or sentence), but the conversational text. The research has shown that conversational structure is constrained by factors such as situation, social conventions, and style (Sacks et al., 1995). Above all else it is sensitive to *sequence* structure. We anticipate how the forms in a text relate to each other and cohere sequentially into a message-making system.

Various theories have been put forward to explain the social nature of conversations. Erving Goffman (1959, 1981) referred to them as strategies in the 'framing' of identity, whereby language is used to fit a situation deliberately by speakers. Another theory is that conversation is intended to play a group function, so as to impart a sense of togetherness among interlocutors as well as a sense of security. As Robin Lakoff (1975) has discussed, speakers will even refrain from saying what they mean in some situations in the service of the higher goal of identity preservation or the conservation of group solidarity.

The notion of speech acts also has become useful in this domain of research (Austin, 1961). This concept has provided a taxonomy for pigeonholing utterances in terms of their social functions. A speech act is an utterance that aims to bring about, modify, curtail, or inhibit some real action. The utterance 'Be careful,' for instance, would have the same effect as putting a hand in front of someone to block him or her from crossing the road carelessly. The statement 'I sentence you to life imprisonment' uttered by a judge has the same effect as marching the accused directly to prison and locking him or her up. The interesting thing about speech acts is that they influence the composition and structure of sentences, showing again that linguistic and communicative competence intersect constantly in interactions.

There are various versions of speech act theory. The central idea in all of these is that grammatical and lexical structures allow people to do things with words. The main types of speech act are as follows:

- *locutionary*, the act itself of saying something (*Mary said, 'I can't come out tonight. It's my birthday'*);
- *illocutionary*, an utterance that indicates the speaker's purpose in saying something (asking and answering questions, giving information, assurance, identifying, etc.) (*Mark protested against what you said*); and
- *perlocutionary*, an utterance that produces sequential effects on the feelings, thoughts, or actions of interlocutors (*Lucy stopped me, bringing me around*).

Each of these main categories can be broken down further (Searle, 1969):

- *representatives*, utterances that commit the speaker to something: stating, concluding, representing, deducing;
- *directives*, attempts by the speaker to get an interlocutor to do something: command, offer, invite, ask, request, beg, permit, dare, challenge;
- *commissives*, utterances committing the speaker to some future course of action: promise, pledge, threaten;
- *expressives*, utterances revealing the psychological state of the speaker toward something: thank, congratulate, apologize, condole, deplore, welcome; and

- *declarations*, utterances connecting propositional content (assumptions in the utterance) to reality: appoint, resign, nominate, pronounce.

The actual structure and contents of conversations are connected to *discourse* or larger linguistic frames of reference. The term refers to the unconscious use of language to connect people or groups in terms of shared values, worldviews, beliefs, biases, and so on. Discourse is typically characterized by keywords that appear frequently in relevant conversations and by other interlinked structures such as logical argumentation styles, conversational styles, and so on (Tannen, 1989, 1993; Fairclough, 1995; Van Dijk, 1997; Scollon and Wong Scollon, 2001).

Schools, corporations, universities, the media, and other groups all develop discourse styles that determine how members within the groups speak to each other. As Russian literary critic and semiotician Mikhael Bakhtin (1981) argued, discourse styles are the products of unconscious processes of socialization and ideological reproduction. They imply an implicit knowledge system based on cultural presuppositions and a shared evaluation of the meanings of a context where speech is used. Discourse thus manifests itself in specific kinds of performative styles. Religious rites, prep rallies, political debates, and formal academic lectures, among others, are anchored in such styles, either traditionally worded or specifically composed for the occasion. Discourse creates allegiances and reinforces identity functions of a group (Hymes, 1975; Bauman, 1977; Wilce, 2006).

Other Traditional Areas

A traditional area of interest within LA, which overlaps with both sociolinguistics and historical linguistics, is the study of dialects or regional, social, or ethnic varieties of a language. By this definition, the English taught in school is only one dialect of contemporary American English, usually called Standard American English. The questions that relate dialect study to social structure include:

- Which dialect is perceived as more prestigious than others and why? This type of study comes under the rubric of *diglossia*.
- Do dialectal differences in artistic venues and expressive forms (such as hip-hop) reflect broader discourse and social tendencies, including political ones?
- Which local pronunciations show local allegiance?
- What is the role of slang and other forms in the construction of identity?

There are many societies where more than one language is used and even taught at school. The study of such phenomena and their social implications has always been an intrinsic part of LA. Spanish-speaking communities in the United States, for example, develop differential cultural realities vis-à-vis the broader English-speaking culture. This produces a whole series of issues that can be studied systematically in LA. Hispanics share many of the traditional values claimed by Americans, but they are by and large also proud of their Latin American heritage. Many Hispanic Americans feel that they should not lose contact with their cultures of origin or their language, seeking to be bicultural and bilingual. Many hope that their cultures will someday be accepted as being part of American culture.

In early school bilingual programs serving Hispanics, students were taught in Spanish in such basic subjects as mathematics and science. Meanwhile, they studied English as a second language. When they were ready, they transferred to classes taught only in English. By the 1990s, many schools had replaced that transitional approach to education with the two-way bilingual approach, combining native Spanish speakers with English-speaking students in all classes. The students progressed together through the grade levels, with some subjects taught in Spanish and others in English. The participants helped one another learn both the new language and the subjects being taught. Since bilingual education began, it was controversial. Critics claimed that it encouraged students to rely too much on Spanish. Supporters argued that it never received enough support. They claimed that lack of funding and of teacher-training programs hindered such education. In 1998, California voters approved a controversial measure requiring that all public school classes be taught only in English. But bilingual classes remained available in California to students whose parents or school boards requested them and to students with special needs. Today, most educators see bilingual education as an advantage, thanks in large part to the findings of LA.

The broader issues of bilingualism and multilingualism are also part of LA. Language change spreads through networks of people who talk with one another. Tight-knit groups that keep to themselves tend not to promote change. People can live next door to one another and not participate in the same network. In the United States, English became the popular language from coast to coast, largely replacing colonial French and Spanish and the languages of Native Americans. In the Caribbean and perhaps in British North America where slavery was practiced, Africans learned the English of their masters as best they could, creating a language for immediate and limited communication called a pidgin. When Africans forgot or were forbidden to use their African languages to communicate with one another, they developed their English pidgin into their native tongue. A language that develops from a pidgin into a native language is called a Creole. African American Vernacular English may have developed this way.

The above discussion suggests that people form social networks, which can be loose or tight depending on how members interact with each other. For instance, an office or factory may be considered a tight community because all members interact with each other. A large course of students would be a looser community because students may only interact with the instructor and maybe a few other students. A multiplex community is one in which members have multiple relationships with each other. For instance, in some neighborhoods, members may live on the same street, work for the same employer, and even intermarry.

Basil Bernstein (1971), a well-known British linguist, has distinguished between *elaborated* and *restricted* codes. These are ways in which speech varieties reflect, and are classified, as patterns in different social classes. Bernstein claimed that

members of the upper and middle classes had ways of organizing their speech into *codes* that were fundamentally different than how the lower classes did. The former were *elaborated* and the latter *restricted*. Codes allow people in classes to bond and to make differentiations that are crucial to them. In Britain he saw a basic social dichotomy – restricted codes emphasized the 'we' as a social group, whereas in elaborated codes the emphasis is on the 'I' and thus on the individual. Restricted codes thus fostered greater group adhesion. He also studied how children with a restricted code struggled at school where the elaborated code is the default form of language.

Elaborated codes are more in sync with standard forms of the language. It is generally assumed that nonstandard language is low-prestige language. However, in certain groups, such as traditional working-class neighborhoods, the standard language may be considered undesirable in many contexts. This is because the working-class code is a powerful in-group marker, and especially for nonmobile individuals, the use of nonstandard varieties (even exaggeratedly so) expresses neighborhood pride and class solidarity. There will thus be a considerable difference in the use of nonstandard varieties when going to the pub or having a neighborhood barbecue (high), and going to the bank (lower) for the same individual.

New Areas

More and more, the tools of LA are being applied to examine the evolution of linguistic forms in cyberspace. The concepts of register and style, to mention but two, are being revamped by the study of communications in online media (Crystal, 2006; Baron, 2008).

Studies such as those by Farman (2012) on Mobile Interface Theory and Papacharissi (2011) on the recreation of the Self in social media are leading the way in showing how to apply traditional linguistic and ethnographic methods to cyberspace. Facebook offers a range of new registers that allow for a new way of framing speech, including messaging, commenting, posting, as well as nonverbal modalities such as photos, sharing links, the 'like' button, and so on. Studying the 'post + commenting' section, which users treat as a quasi-conversational space, can provide new insights into older theories about sequential unfolding, turn-taking, coherent exchanges, and so on within CA. The theoretical significance of this research is the exploration of how humans maintain social order through writing and reading practices in cyberspace.

Communication that involves more than two parties is typically characterized by assigning and adopting roles, a phenomenon that Goffman (1981) called 'footing.' These roles are partly achieved in the conversational situation rather than being something that preexists ritualistically. On Facebook, the number of potential participants in a 'commenting event' tends to be much higher than in regular conversational interaction, and thus managing Facebook conversations changes the structure of footing. Anyone on a user's friends' list can leave a comment at any time, whereas in face-to-face conversation, only those who can hear another's voice will be able to participate. A different set of conversational skills is thus emerging on Facebook and other social media sites and it is starting to show that the traditional theories in LA need adaptation and elaboration.

Concluding Remarks

Since LA focuses on how language is used and on how use influences worldview and social relations, ultimately it allows us to gain valuable insight into the overall enigma of language. The goal of the puzzle-solver is to figure out how language fits into human life and what it does to that life. The anthropological linguist seeks to figure out how the bits and pieces (sounds, words, and so on) cohere into the organism of language. But having described how they do, we are still left with the dilemma of why such an organism came into being in the first place. The research in LA has shown that diversity is the norm, but that diversity only reflects differentiated attempts to solve similar problems across the world. It shows how we go about tackling the same problems of classification, understanding, and communication coming up with different solutions according to situation, time, and place. The fact that we can understand each other nonetheless shows how much we are really all one race, equal and seeking similar objectives and outcomes to the experience of existence.

See also: Code Switching: Linguistic; Ethnic Identity and Ethnicity in Archaeology; Ethnicity and Ethnic Groups: Historical Aspects; Ethnicity: Anthropological Aspects.

Bibliography

Alpher, B.J., 1987. Feminine as the unmarked grammatical gender: Buffalo girls are no fools. Australian Journal of Linguistics 7, 169–187.
Austin, J.L., 1961. How to Do Things with Words. Harvard University Press, Cambridge, MA.
Bakhtin, M., 1981. The Dialogic Imagination (C. Emerson, M. Holquist, Trans.). University of Texas Press, Austin.
Baron, N.S., 2008. Always on. Oxford University Press, Oxford.
Bauman, R., 1977. Verbal art as performance. American Anthropologist 77, 290–311.
Berlin, B., Kay, P., 1969. Basic Color Terms. University of California Press, Berkeley.
Bernstein, B., 1971. Class, Codes and Control: Theoretical Studies towards a Sociology of Language. Routledge & Kegan Paul, London.
Bloomfield, L., 1933. Language. Holt, New York.
Boas, F., 1940. Race, Language, and Culture. Free Press, New York.
Chomsky, N., 1957. Syntactic Structures. Mouton, The Hague.
Chomsky, N., 2002. On Nature and Language. Cambridge University Press, Cambridge, MA.
Crystal, D., 2006. Language and the Internet, second ed. Cambridge University Press, Cambridge.
Danesi, M., 2013. On the metaphorical connectivity of cultural sign systems. Signs and Society 1, 33–50.
Duranti, A., 1992. Language and bodies in social space: Samoan greetings. American Anthropologist vol. 94, 657–691.
Duranti, A., 2003. Language as culture in U.S. anthropology: three paradigms. Current Anthropology 44 (3), 323–348.
Errington, J.J., 1988. Structure and Style in Javanese: A Semiotic View of Linguistic Etiquette. University of Pennsylvania, Philadelphia.
Fairclough, N., 1995. Critical Discourse Analysis: The Critical Study of Language. Longman, London.
Farman, J., 2012. Mobile Interface Theory: Embodied Space and Locative Media. Routledge, London.
Goffman, E., 1959. The Presentation of Self in Everyday Life. Doubleday, Garden City.
Goffman, E., 1981. Forms of Talk. University of Pennsylvania Press, Philadelphia.
Halliday, M.A.K., 1985. Introduction to Functional Grammar. Arnold, London.
Hill, J., Mannheim, B., 1992. Language and worldview. Annual Reviews in Anthropology 21, 381–406.

Hoijer, H., 1954. The sapir-whorf hypothesis. In: Hoijer, H. (Ed.), Language in Culture. Chicago University Press, Chicago, pp. 92–105.
Honeck, R.P., Hoffman, R.R. (Eds.), 1980. Cognition and Figurative Language. Lawrence Erlbaum Associates, Hillsdale, NJ.
Humboldt, W. von, 1836. On Language: The Diversity of Human Language Structure and Its Influence on the Mental Development of Mankind (P. Heath, Trans.). Cambridge University Press, Cambridge, 1988.
Hutchby, I., Wooffitt, R., 1998. Conversation Analysis. Polity Press, Cambridge.
Hymes, D., 1962. The ethnography of speaking. In: Gladwin, T., Sturtevant, W. (Eds.), Anthropology and Human Behavior. The Anthropology Society of Washington, Washington, pp. 13–53.
Hymes, D., 1971. On Communicative Competence. University of Pennsylvania Press, Philadelphia.
Hymes, D., 1975. Breakthrough into performance. In: Hymes, D. (Ed.), In Vain I Tried to Tell You: Essays in Native American Ethnopoetics. University of Pennsylvania Press, Philadelphia, pp. 79–141.
Jones, W., 1786[1967]. In: Lehmann, W.P. (Ed.), A Reader in Nineteenth-Century Historical Linguistics. Indiana University Press, Bloomington, pp. 19–34.
Kroskrity, P.V., 1998. Arizona Tewa Kiva speech as a manifestation of linguistic ideology. In: Schieffelin, B.B., Woolard, K.A., Kroskrity, P. (Eds.), Language Ideologies: Practice and Theory. Oxford University Press, New York, pp. 103–122.
Kulick, D., 1992. Language Shift and Cultural Reproduction: Socialization, Self and Syncretism in a Papua New Guinea Village. Cambridge University Press, Cambridge.
Lakoff, G., 1987. Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. University of Chicago Press, Chicago.
Lakoff, G., Johnson, M., 1980. Metaphors We Live by. Chicago University Press, Chicago.
Lakoff, G., Johnson, M., 1999. Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. Basic Books, New York.
Lakoff, R., 1975. Language and Woman's Place. Harper and Row, New York.
Lucy, J., 1996. The scope of linguistic relativity: an analysis and review of empirical research. In: Gumperz, J., Levinson, S. (Eds.), Rethinking Linguistic Relativity. Cambridge University Press, New York, pp. 37–70.
Morris, C.W., 1938. Foundations of the Theory of Signs. Chicago University Press, Chicago.
Ochs, E., 1988. Culture and Language Development: Language Acquisition and Language Socialization in a Samoan Village. Cambridge University Press, Cambridge.
Ochs, E., Schieffelin, B.B., 1984. Language acquisition and socialization: three developmental stories and their implications. In: Shweder, R., LeVine, R.A. (Eds.), Culture Theory: Essays on Mind, Self, and Emotion. Cambridge University Press, New York, pp. 276–320.
Ochs, E., Taylor, C., 2001. The father knows best dynamic in dinnertime narratives. In: Duranti, A. (Ed.), Linguistic Anthropology: A Peader. Blackwell, Oxford, pp. 431–449.
Ortony, A. (Ed.), 1979. Metaphor and Thought. Cambridge University Press, Cambridge.
Papacharissi, Z., 2011. A Networked Self: Identity, Community and Culture on Social Network Sites. Routledge, London.
Pollio, H., Barlow, J., Fine, H., Pollio, M., 1977. The Poetics of Growth: Figurative Language in Psychology, Psychotherapy, and Education. Lawrence Erlbaum Associates, Hillsdale, NJ.
Rumsey, A., 1990. Word, meaning, and linguistic ideology. American Anthropologist 92 (2), 346–361.
Sacks, H., Jefferson, G., Schegloff, E.A., 1995. Lectures on Conversation. Blackwell, Oxford.
Sapir, E., 1921. Language. Harcourt, Brace, and World, New York.
Saussure, F. de, 1916. Cours de linguistique générale. Payot, Paris.
Schieffelin, B.B., 1990. The Give and Take of Everyday Life: Language Socialization of Kaluli Children. Cambridge University Press, Cambridge.
Schieffelin, B.B., 1995. 'Creating evidence: making sense of written words in Bosavi. Pragmatics 5 (2), 225–244.
Schieffelin, B.B., 2000. Introducing Kaluli literacy: a chronology of influences. In: Kroskrity, P. (Ed.), Regimes of Language. School of American Research Press, Santa Fe, pp. 293–327.
Schieffelin, B.B., 2002. Marking time: the dichotomizing discourse of multiple temporalities. Current Anthropology 43 (Suppl.), 5–17.
Scollon, R., Wong Scollon, S., 2001. Intercultural Communication, second ed. Blackwell, Oxford.
Searle, J., 1969. Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press, Cambridge.
Shuy, R., 1997. A brief history of American sociolinguistics: 1949–1989. In: Paulston, C.B., Tucker, G.R. (Eds.), The Early Days of Sociolinguistics: Memories and Reflections. Summer Institute of Linguistics, Dallas, pp. 11–32.
Sidnell, J., Enfield, N.J., 2012. Language diversity and social action: a third locus of linguistic relativity. Current Anthropology 53 (3), 33–50.
Silverstein, M., 1976. Shifters, linguistic categories, and cultural description. In: Basso, K., Selby, H.A. (Eds.), Meaning in Anthropology. University of New Mexico Press, Albuquerque, pp. 11–56.
Silverstein, M., 1979. Language structure and linguistic ideology. In: Cline, R., Hanks, W., Hofbauer, C. (Eds.), The Elements: A Parasession on Linguistic Units and Levels. Chicago Linguistic Society, Chicago, pp. 193–247.
Silverstein, M., 1985. Language and the culture of gender: at the intersection of structure, usage, and ideology. In: Mertz, E., Parmentier, R. (Eds.), Semiotic Mediation: Sociocultural and Psychological Perspectives. Academic Press, Orlando, pp. 219–259.
Tannen, D., 1989. Talking Voices. Cambridge University Press, Cambridge.
Tannen, D., 1993. Framing in Discourse. Oxford University Press, Oxford.
Taylor, D.M., 1977. Languages of the West Indies. Johns Hopkins, Baltimore.
Van Dijk, T. (Ed.), 1997. Discourse as Social Interaction. Sage, London.
Whorf, B. Lee, 1956. Language, Thought, and Reality. In: Carroll, J.B. (Ed.). MIT Press, Cambridge, MA.
Wilce, J.M., 2006. Magical laments and anthropological reflections: the production and circulation of anthropological text as ritual activity. Current Anthropology 47 (6), 891–914.

Relevant Websites

http://www.aaanet.org/ – American Anthropological Association.
http://www.anthrosource.net/Issues.aspx?issn=1055-1360 – Journal of Linguistic Anthropology.
http://www.oxfordbibliographies.com/view/document/obo-9780199772810/obo-9780199772810-0051.xml – Oxford Bibliographies: Linguistic Anthropology.
http://linguisticanthropology.org/about/ – Society for Linguistic Anthropology (SLA).

Linguistic Approaches to Agrammatism

Alan Beretta, Michigan State University, East Lansing, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses linguistic approaches to agrammatism. First, we consider the possible contributions that the study of agrammatism can make to an understanding of the ways a brain computes language. To understand how a brain computes language, we minimally need to know something about language and something about language processing so that we can pose coherent questions to a brain. We consider in some detail how one well-known model of agrammatism accommodates what we know about language and language processing and find that the basic challenges have not been met. We compare this to a developing range of inquiry into agrammatism that is potentially more useful in moving the larger project forward.

Introduction

Agrammatism is a somewhat general term that is used to refer to an aphasic production deficit that results in slow, effortful speech in which functional categories, such as tense markers, while they might make an occasional appearance, are often missing or wrongly substituted. This characterization is traditionally associated with Broca's aphasia, as the following exchange with a patient suggests:

Patient: Farmer, uh, Ohio, farm, soy beans and corn, uh, 400 acres, yeah, miles and miles and miles, uh 32 years ago, me, supervisor
Q: Do you still farm?
Patient: No, well, uh, lawnmower (laughs).
Q: Your children, what do they do?
Patient: Uh, uh, Harry, uh, Minnesota, uh, Minneapolis, and uh, Indians, dead, oh, many, 200, 300, 400, 500, dead.
Q: What does Harry do with Indians?
Patient: Grounds, uh, Indian grounds.
Q: A reservation?
Patient: No, uh, die, uh, goddammit.
Q: Oh, burial grounds?
Patient: Yeah! Yeah!
Q: Harry works in burial grounds. What job does he do?
Patient: Uh, uh, oh boy, arrow, uh, Indian pieces.
Q: He looks for these things?
Patient: Yeah! Yeah!
Q: So he's an archaeologist?
Patient: Yeah! Yeah!

This extract also supports another traditional observation, namely, that Broca's aphasic patients have preserved comprehension. As Broca himself noted a century and a half ago, they operate with full intelligence; they hear and understand everything that is addressed to them perfectly (1861).

Inquiry into unimpaired sentence production lags behind inquiry into unimpaired sentence comprehension; so agrammatic comprehension if it were not in fact perfectly intact, would have the greater potential to make a useful contribution to an understanding of normal parsing. As it turned out, a landmark study in the 1970s (Caramazza and Zurif, 1976) revealed that Broca's aphasics could not understand certain types of sentences once their 'full intelligence,' including world knowledge, conversational context, and so forth, was rendered irrelevant and they were forced to rely on their knowledge of sentence structure to determine who was doing what to whom. On a sentence-picture matching task (in which they point to one of two or three pictures that best matches a sentence they have just heard), they could understand a passive sentence, such as *the ball was kicked by the boy*, because they know that boys can kick balls but not the reverse, but if the sentence was reversible, such as *the girl was chased by the boy*, they could not tell who was doing the chasing. In semantically reversible sentences, they could understand actives, but not passives; and they could understand subject relative clauses but not object relative clauses.

Even if this asymmetric agrammatic comprehension had been recognized in Broca's time, he would not have been able to say why such a pattern should have occurred. The problem could not be addressed in Broca's time because there was no theory available that could make sense of the observations. In fact, Broca was acutely aware of this: "the biggest obstacle hindering progress in this part of physiology [cerebral localization] results from the insufficiency and uncertainty of the functional analysis, which must precede the search for identifying organs for each function. Science is so behind in this respect, that it has not even found its foundation yet …" (1861/2006, p. 294). Broca did not envisage that this need for theory applied to language; he considered a notion such as 'spoken language' to be a neatly circumscribed, self-evident function not requiring any explanation. It was other faculties of mind that were in need of theory. However, if we take up Broca's realization that we cannot relate brain states to some function unless we know something about that function, and if we assume – surely uncontroversially – that the only relevant knowledge is theoretical, then we might expect to benefit from attention to a theory of language in seeking to understand the nature of Broca's aphasia. We might also expect to benefit from an appreciation of how language conspires with other faculties involved in comprehension, such as memory, that is, a theory of sentence processing.

Fortunately, we have an increasingly good appreciation of the factors that affect sentence processing and we have an increasingly good understanding of the nature of the human language faculty. In light of this, how might studies agrammatic comprehension contribute to what is known about how a brain processes the sentences it is confronted with? What has been accomplished to date? What challenges lie ahead? Let us address each of these questions in turn.

What Contribution Can Inquiry into Agrammatic Comprehension Make?

To answer this question, it is useful to step back and view the issue from a broader perspective and ask what the study of brain and language, that is, neurolinguistics could contribute. If we had no theory of language, then we would be reduced to posing pretheoretical, and therefore incoherent, questions to the brain. But at least since Chomsky (1981), understanding and evidence have reached a point where we can refer to a theory of language. Similarly, the interplay of hypothesis and experiment in neurobiology, at least since the work of Hodgkin and Huxley (1952a, b, c, d, e), has attained a level of explanation that can reasonably be referred to as a theory of brain. The problem is that our best theory of language and our best theory of brain do not have a single property in common. It is what Chomsky (1995, 2000) has called a theory unification problem, a standard kind of problem throughout the history of science, but that is not to minimize the challenge. Unifying our theories of language and brain currently seems hopelessly remote.

We know how we could find out more. We could proceed invasively, as is routinely done in inquiries into brain function with nonhuman animals. Single cells in precisely identified regions of normally functioning intact brains are recorded to see if they responded to specific stimuli. To complement such findings, a whole range of techniques for inactivating those regions are deployed. For example, in avian song research, tissue from a target region might simply be ablated; or toxins introduced to block sodium channels, thus preventing cells from firing. Or, an antagonist to a particular neurotransmitter might be injected into the region, which has the effect of blockading receptors for that neurotransmitter, but leaving the receptors for other neurotransmitters unaffected. Or, brain regions might be systematically cooled, which has the effect of slowing function, allowing localization of behaviors with structures on many timescales, such as the motif, syllabic, and subsyllabic structures of zebra finch song (Long and Fee, 2008). The range, precision, and discriminability of the techniques available to research on intact and impaired brains in research on birds have yielded a great deal of knowledge in a very short period of time.

But for obvious reasons, they are not available for research on humans; there can be no direct electrical stimulation, no ablation of tissue, no severing of fibers, no single-cell recording, no pharmacological interventions to temporarily disable either pre- or postsynaptic neurons, and no cooling to analyze complex temporal dynamics. There can be no research at the cellular or molecular levels at all (except adventitiously, and in a very limited way, when patients have electrodes implanted for clinical reasons prior to surgery, for example, for intractable epilepsy).

So, the problem that confronts us is twofold: first, current theories of language and brain share not a single property; and second, language function, unlike, say, vision, in core respects is not available in other species, so we cannot avail ourselves of the invasive techniques that would allow inquiry at cellular and molecular levels.

With regard to the second problem, we have had to be more subtle. The tools at our disposal are the various noninvasive imaging technologies with which we attempt to identify regions involved in language function, or the timing of those functions, in living brains; and opportunistic use is made of naturally occurring brain damage that affects language function in some way, that is, in aphasia.

Addressing the first problem, the remoteness of language – brain theory-unification, it is a little less obvious how to proceed. One approach would be to let linguists and neurobiologists get on with the job or developing better and better theories and return in 50 years to see if some means of unification presents itself. And for a satisfying resolution on a par with Pauling's unification of chemistry and physics, we may have to wait for one theory or the other to undergo a parallel to the quantum revolution. But resolution might be arrived at differently. One possible approach is to target intermediate levels of unification, and the most obvious is for neurolinguists to deploy not only linguistic theory but also whatever understanding has been gained in psycholinguistics. Seeking to integrate what we know about the nature of language and what we know about language processing is major challenge but one that is at least realistic in the medium term; indeed many researchers have been actively investigating the relation. Precisely how to envisage the relation varies widely among individuals, and it is beyond the scope of this article to discuss the issues involved, but see, for example, Phillips (2012) and Phillips and Lewis (2013) for relevant discussion. For current purposes, it is sufficient to note that the relation between linguistics and psycholinguistics is firmly on the agenda, as it is in this respect that syntactic approaches to agrammatism may have more to offer in the coming years.

What Has Been Accomplished to Date?

In what follows, I will examine just one model of agrammatic comprehension. It is worth examining one model in some depth rather than cursorily describing several (and there are many) because much can be learned from the ways in which it has failed conceptually and empirically and where progress may be made. We will see that it runs into serious empirical problems, but more interestingly, we will see that it avoids the constraints of linguistic theory and uses theory-free strategies to accommodate anomalies. In addition, we will see that it does not set itself up in relation to normal, unimpaired processing. Seeing these limitations clearly ought to make it easier to see an alternative approach to agrammatic comprehension data.

The Trace-Deletion Hypothesis

In discussing many current models of agrammatic comprehension, we must first confront an epiphenomenon. Consider the following three facts:

1. In the world's languages, subjects typically precede objects.
2. Subjects are typically assigned thematic roles that are higher in thematic hierarchies than the thematic roles assigned to objects.

3. Agents (the do*ers* of an action) typically occur before Themes (the do*ees* of an action).

The canonical order of thematic roles is a statistical fact. It does not follow from any principle in linguistic theory. It is, that is to say, merely epiphenomenal, as far as linguistic theory is concerned. However, many models of agrammatic comprehension have been based squarely on the epiphenomenon of linearity, among them, Avrutin (2006), Caplan and Futter (1986), Linebarger (1995), and Piñango (2000). The guiding notion is that Broca's aphasics resort to a strategy in which thematic roles are assigned in a linear order. The Agent role, the highest role in all thematic hierarchies, is assigned to the first relevant noun phrase (NP) in a sentence. A role that is lower in a thematic hierarchy is assigned to a second relevant NP, and so on.

This reliance on linearity is true also of the trace-deletion hypothesis, henceforth TDH (Grodzinsky 1986, 2000).

The TDH invoked a concept from linguistic theory, the concept of syntactic movement and trace. In the 1980s, when the TDH was first formulated, very few researchers working with aphasic data were inclined to regard linguistic theory as relevant; so, viewed against that atheoretical background, the TDH, in opting to construct a hypothesis based on something theoretical, amounted to a trail-blazing proposal.

An early formulation of the TDH (Grodzinsky, 1986) proposed that in an otherwise normal representation, traces in theta positions (i.e., positions to which thematic roles such as Agent or Theme are assigned) are deleted. Since the trace is lost, thematic role information is also lost in NPs that have undergone movement. Broca's aphasics are apparently unable to infer the missing role by consulting their knowledge of the roles that the relevant verb requires. Instead, and this is where the reliance on linearity asserts itself, they resort to a nonlinguistic default strategy which assigns a role to the moved NP according to its serial position in the surface string (e.g., in a sentence with two NPs, if the verb is *bite*, and if the first NP in the surface string is the one that has moved, the strategy assigns it the role of Agent; if the second NP in the surface string is the one that has moved, the strategy assigns it the role of Theme; and so forth).

How does this play out in a passive sentence such as (1) which many Broca's aphasics tend to guess at, when the task is sentence-picture matching? What happens in the unimpaired representation is that the NP argument of the *by*-phrase (*the baker*) receives the Agent theta role, but not via movement. The subject (*the zebra*), by contrast, is derived by movement, and so is linked to a trace. In (1), the normal role assignments are given:

(1) [The zebra]$_i$ was bitten t_i by [the baker]
THEME AGENT

Because no movement is involved in the assignment of the Agent role to *the baker*, it is not linked to a trace, and the prediction is that Broca's aphasics will have no difficulty with that assignment; they get it for free via the syntax, which is by hypothesis unimpaired except with respect to movement. The NP (*the zebra*) is derived by movement, but since the trace it is linked to is deleted in a Broca's representation, thematic role assignment cannot be accomplished syntactically. The parser outputs an incomplete representation and its job is done.

Enter the default strategy. It assigns the role of Agent to the moved NP. The result is that the Broca's aphasic has a representation of the passive that contains two Agents, as in (2), where '*' indicates loss of trace:

(2) [The zebra] was bitten* by [the baker]
AGENT AGENT
(via default) (via syntax)

When asked to complete a sentence-picture matching task, the representation in (2) compels Broca's aphasics to guess which NP to interpret as Agent. Thus, their performance is random. Consequently, in the passive, we see that random performance is accomplished by means of trace deletion, the application of a linear default heuristic, and the resulting competition for Agenthood.

How is the frequently observed above-chance performance on active sentences by Broca's aphasics explained? If, as in earlier versions of syntactic theory, it is assumed that there is no movement in sentences such as (3), then the representation of a Broca's aphasic is complete and thus there is no reason for the linear default strategy to be triggered.

(3) [The zebra] bit [the baker]
AGENT THEME
(via syntax) (via syntax)

Thus far, we have seen how the TDH explains the core findings that Broca's aphasics perform above chance on actives and randomly on passives. Now, we turn to the other core findings, above-chance performance on subject relatives in sentence-picture matching tasks, but random on object relatives. Let us see how the TDH accounts for these. With respect to object relatives, earlier versions of syntactic theory assumed only one movement, as in (4):

(4) Point to [the zebra]$_i$ Op$_i$ that [the baker] bit t_i
THEME AGENT

In (4), the movement of *the zebra* is from the object position, now occupied by the trace. In an aphasic representation, lacking trace, the default strategy assigns the Agent role to the first NP (*the zebra*), and normal assignment obtains for the subject, *the baker*, as in (5):

(5) Point to [the zebra] that [the baker] bit*
AGENT AGENT
(via default) (via syntax)

As in the passive, there are two Agents. The aphasic guesses and random performance follows.

In the subject relative, the normal representation is as follows:

(6) Point to [the zebra]$_i$ Op$_i$ that t_i bit [the baker]
AGENT THEME

In a Broca's representation, in which trace is lost, the Agent role is assigned to the NP (*the zebra*) by the default strategy,

which happens to be the correct assignment, compensating for the loss of trace, and yielding a normal interpretation:

(7) Point to [the zebra] that* bit [the baker]
AGENT THEME
(via default) (via syntax)

The above examples show how the TDH succeeded in partitioning the core data. At this point, note only that in the three examples in which a trace is involved, the application of the atheoretical default strategy is necessary for the data to be partitioned in the desired way. No default strategy, no division. The strategy is already doing much of the heavy lifting, but now we will see that it works overtime, changing freely, to keep pace with selected empirical anomalies and theoretical developments.

Here is an example of an empirical anomaly which prompted a change in the default strategy. A study by Hickok and Avrutin (1995, 1996) on *wh*-questions reported that Broca's aphasics performed above chance on *which-N* questions involving subject extraction, such as (8), but at chance on *which-N* questions involving object extraction, such as (9).

(8) [Which horse]$_i$ t_i kicked [the giraffe]
AGENT THEME

(9) [Which horse]$_i$ did [the giraffe] kick t_i
THEME AGENT

In (10), application of the default assigns the Agent role to *which horse*, and thus aphasics perform in a manner similar to normal subjects, above chance. In (11), the default assigns the Agent role to *which horse*, but now both NPs in the sentence have the Agent role. Forced to guess, random behavior is assured in Broca's aphasics.

(10) [Which horse]* kicked [the giraffe]
AGENT THEME
(via default) (via syntax)

(11) [Which horse] did [the giraffe] kick*
AGENT AGENT
(via default) (via syntax)

So far, so good. However, Hickok and Avrutin also reported above-chance performance for *both* subject- *and* object-extracted *who* questions, (12) and (13).

(12) [Who]$_i$ t_i chased [the horse]
AGENT THEME

(13) [Who]$_i$ did [the horse] chase t_i
THEME AGENT

The problem is with (13). According to the TDH, the default should assign the Agent role to the moved wh-phrase, but this would mean that both NPs are Agents, and therefore subjects should behave randomly, contrary to fact:

(14) [Who] did [the horse] chase*
AGENT AGENT
(via default) (via syntax)

Hickok and Avrutin (1995, 1996) account for the differences in *who* questions and *which-N* questions by appealing to Cinque's (1990) distinction. He proposed that *which-N* phrases are referential and yield one kind of chain, a binding chain, whereas *who* phrases are nonreferential and yield a different kind of chain, a government chain. So, there is a theoretical distinction between the two kinds of questions, consistent with Hickok and Avrutin's data; but inconsistent with the TDH.

To address the problem, Grodzinsky (1995, p. 46) elected to revise the strategy (which, being atheoretical, can change without constraint) by asserting that the default strategy applies only to referential NPs. Since it could be argued that *who* is nonreferential, the TDH does not apply.

Another theoretical development was about to test the flexibility of the default strategy like never before: the VP-internal subject hypothesis, henceforth VPI.

A consensus had formed that subjects are base generated within VP and undergo movement. An immediate consequence is that traces occur in virtually every sentence, including simple actives. Apply this to the TDH and traces are deleted in virtually every sentence Broca's aphasics hear. And yet, Broca's aphasics do not have problems understanding virtually every sentence they hear, so trace-deletion per se makes no predictions at all. More and more obviously, what plays the crucial role in distinguishing sentences Broca's aphasics understand from those they do not is the theory-free default strategy. In fact, only by virtue of the strategy does the TDH have any consequences.

The VPI not only exacerbated the conceptual difficulties faced by the TDH, but it multiplied its empirical problems exponentially, as we will now see. In a single-clause sentence with a transitive verb such as *chase* and two referential NPs {Agent; Theme}, the three logical possibilities for linear role assignment are as follows:

Option A	NP1 moved Agent assigned by default	NP2 not moved Correct role assigned normally
Option B	NP1 not moved Correct role assigned syntactically	NP2 moved Theme assigned by default
Option C	NP1 moved Agent assigned by default	NP2 moved Theme assigned by default

Instances of Option A are easy to find. One such example is the passive sentence in (2) above, repeated for convenience as (15):

NP1 moved NP2 not moved
(15) [The zebra] was chased* by [the baker]
AGENT AGENT
(via default) (via syntax)

An example of Option B is not so easy to find, but for Option C, there are examples in abundance, especially from languages in which word order is freer than it is in English. Even with the English sentences above, the TDH unravels. Consider the object relative clause in (4) and (5). There was

only one movement prior to VPI, but now there are two, as shown in (16):

(16) Point to [the zebra]$_i$ Op$_i$ that [the baker] t$_j$ chased t$_i$
THEME AGENT

The core finding is that Broca's aphasics perform randomly on such sentences. When there was only one movement, the TDH could account for the data, as shown in (4) and (5). Now that there are considered to be two movements that result in two NPs that do not receive theta roles syntactically, but must acquire them both via the default strategy, what does the TDH predict? Application of the default strategy in a linear fashion to both NPs must give the following aphasic representation, as indicated in Option C above:

NP1 moved NP2 moved
(17) Point to [the zebra] that [the baker]* chased*
AGENT THEME
(via default) (via default)

In (17), since the TDH default strategy assigns roles by linear positions, the Agent role is assigned to the first NP (*the zebra*) and the Theme role to the second NP (*the baker*). Thus, subjects would be predicted to perform reliably *below chance*, as their role assignment is exactly the opposite of the unimpaired subject's role assignment (16). Since Broca's aphasics in fact perform at chance, this prediction is wrong.

The TDH now also makes the wrong prediction with the *which-N* question in (9), *Which zebra did the baker chase*. Prior to ISH, the TDH could arrive at the chance (random) prediction that is consistent with fact, but now that the subject NP (*the giraffe*) moves in addition to the wh-phrase, it would yield a below-chance prediction, contrary to fact, as shown in (18):

(18) [Which zebra] did [the baker]* chase*
AGENT THEME
(via default) (via default)

In (18), once again, since the TDH default strategy assigns roles by linear positions, the Agent role is assigned to the first NP (*which zebra*) and the Theme role to the second NP (*the baker*). Thus, subjects would be predicted to perform reliably *below chance* (their role assignment is exactly the opposite of the unimpaired subject's role assignment).

Another problem, for the end is not yet: The combination of VPI and the default strategy predicts anomalous results with any sentence type which does not assign an Agent role and which only involves a single dependency. In this respect, adjectival passives are of interest. Adjectival passives are derived in the lexicon, unlike verbal passives, which involve syntactic movement from object position (Levin and Rappaport, 1986). An example of the structure of an adjectival passive is given in (19). The negative prefix *un-*, which only attaches to adjectives, is often used as a test of an adjectival passive.

(19) The man$_i$ [was [$_{AP}$ t$_i$ unhurt] in the accident]

Because it is lexically derived, the adjectival passive has a structure involving a predicate-internal trace, as can be seen in (19). What is interesting here, however, is that adjectival passives assign a Theme role to their predicate-internal position.

With that in mind, consider the finding by Grodzinsky et al. (1991) that Broca's aphasics performed above chance on adjectival passives involving negative *un*-prefixation. In this study, the VPI was not assumed, and above-chance performance was predicted because no trace was thought to be involved. However, as we see in (19), adjectival passives with negative un-prefixation *do* involve movement and contain a trace. Application of the default strategy would assign the role of Agent, not Theme, to the moved subject of an adjectival un-passive, and therefore comprehension should be seriously impaired, which it is not.

The problems posed by the VPI with English sentences is challenging enough, but once we look at Broca's aphasia in languages with less restrictive word order than English, the problems mount up exponentially for the TDH and its linear strategy. We will consider only two examples here, but see Beretta (2001) for other examples.

Japanese scrambled actives (in which the object NP moves to the front), such as (20) yielded chance performance by Broca's aphasics in a study by Hagiwara and Caplan (1990).

(20) Hanako$_i$-o Taro$_k$-ga t$_k$ t$_i$ nagutta
Hanako-Acc Taro-Nom hit
'Taro hit Hanoko'

According the TDH, however, the prediction is for below-chance performance:

(21) [Hanoko]-o[Taro]-ga**nagutta
AGENT THEME
(via default)(via default)

In Korean actives (22), in which the object NP is dislocated to a sentence-initial position, Broca's aphasics perform at chance (Beretta et al., 2001).

(22) saja-lul$_i$ key-ka$_j$ t$_j$ t$_i$ mul-eyo
lion-Acc dog-Nom bite-COMP
'The dog bit the lion'

As with the Japanese case, the TDH wrongly predicts below-chance performance, as in (23):

(23) [saja]-lul [key]-ka** mul-eyo
AGENT THEME
(via default) (via default)

There are many other crosslinguistic data one could discuss here, but the moral of the story is by now abundantly clear. With the VPI and frequently *two* dependencies for the default strategy to deal with, data from languages with relatively free word order were always going to raise difficulties for the TDH and its commitment to a linear heuristic to do the work of partitioning data, and so it proved.

The TDH solution to the two-dependency problem posed by the VPI was to simply add another strategy. This new strategy, called *theta-bridging* (Grodzinsky, 2006), in effect absolves Broca's aphasics of any constraints imposed by the linguistic theory, in this case, the VPI. By stipulation, they do not have to attend to the dependency involved in subject movement out of VP.

But by now, we are very far indeed from any commitment to theoretical relevance.

What Challenges Lie Ahead?

If we can agree that the ultimate goal of neurolinguistic inquiry is theory unification, then muzzling linguistic theory by shackling it to a theory-free strategy is inimical to that goal. The linear models do not entirely dismiss linguistic theory, but they have set out their stall in such a way that they fail to make a single prediction without the effective agency of a linear strategy. The strategy, in all of the linear models, is designed to compete with syntactic processes in the sentences Broca's aphasics do not understand, but not to compete in those they do understand. However, nothing constraints strategies, and linearity of role assignment does not follow from anything in linguistic theory. And, absent a linear *deus ex machina*, there is simply no way that the linear accounts can divide the Broca's comprehension data in the required manner.

Linearity, we must conclude, is not a useful concept to invoke in inquiry into aphasic comprehension deficits, either empirically or conceptually. But independently of the arguments and evidence above that lead to this conclusion, there is another line of reasoning that lends support to it: relating language disorders to processing in the normal, unimpaired population.

Many commentators have observed that Broca's aphasics retain much, conceivably all, of their syntactic knowledge. After all, it has been reported that they can understand not just simple actives, subject relatives, but also clitic-doubling actives, subject clefts, raising constructions, and adjectival passives, among many other constructions (e.g., Avrutin, 2006; Beretta et al., 1999, 2001; Grodzinsky et al., 1991). If structural knowledge is more or less intact in these cases, then surely structure-building processes are also largely preserved in Broca's comprehension. After all, it would be entirely implausible that Broca's aphasics could understand all of the structures that they do if they did not possess extensive resources for structural analysis. Assuming that they do in fact have such resources, we are confronted with the question as to where these resources come from. Parsimony suggests that they were always there, but now perhaps restricted following neural insult. If we wish to press this view into service, then we have no choice but to commit ourselves to grounding our approach to Broca's comprehension in relation to a model of normal comprehension.

Linear models are not sympathetic to this view. As Friederici and Gorrell have pointed out, the TDH "lacks any connection to theories of sentence comprehension in normals" (1998, p. 260), a point also made with some force by Mauner (1995, p. 364). There seems to be little motivation to propose and defend extralinguistic models of linear heuristics in the absence of any evidence that Broca's aphasics resort to processes that are unlike those used by the normal, unimpaired population.

In this connection, Lukatela et al. (1991) have proposed that we would do better to relate aphasic deficits to the normal case. They reported that both Broca's aphasics and Wernicke's aphasics exhibit an identical pattern of comprehension errors in relative clauses in Serbo-Croatian, and furthermore, they are both different from controls only in the quantity of errors they make, but not in the kind of errors they make. In a similar vein, considering experimental reports in language development, dyslexia, aphasia, and normal adults sentence processing, Crain et al. (2001, p. 294) conclude that, "the results of experimental investigations across several subject populations reveal parallel patterns of linguistic behavior on [relative clauses]". Crain et al. believe that a processing account that encompasses both disordered and normal sentence processing would do justice to the facts. Caplan and Waters (1999), considering the role of memory in sentence processing, observe that in "many experimental paradigms, the processing of syntactically more complex sentences is not disproportionately affected by a concurrent verbal memory load, either in normal subjects or in subjects with extremely reduced working memory capacity or in aphasic patients" (p. 92).

If we are to meet the challenge of relating disordered language processing to normal language processing, it will entail a more concerted effort to carry out online rather than offline studies. In an excellent review of real-time comprehension in agrammatic aphasia, Schumacher (2009) presents a compelling case for using online techniques. There has been a steady tradition of this approach in studies of agrammatism (e.g., Blumstein et al., 1998; Burkhardt et al., 2008; Haarman and Kolk, 1994; Love et al., 2008; Prather et al., 1997), though these are more the exception than the rule.

To conclude, in order to rescue many current models from a parochial dependence on atheoretical concepts, the major challenge that lies ahead is to view agrammatic comprehension as complementary to normal processing, and to pursue online inquiry rather than offline. There are already efforts in that direction and it would be surprising if this were not to become the norm.

See also: Aphasia; Control, Syntactic; Functional Brain Imaging of Language processes; Language Processing, Functional Magnetic Resonance Imaging of; Neurolinguistic Processing of Psychological Verbs; Neurological Approaches to Agrammatism; Speech Errors, Psychology of; Speech Processing, The Cortical Organization of; Speech Production, Psychology of; Syntax (General).

Bibliography

Avrutin, S., 2006. Weak syntax. In: Grodzinsky, Y., Amunts, K. (Eds.), Broca's Region. Oxford University Press, Oxford, pp. 49–62.

Beretta, A., 2001. Linear and structural accounts of theta-role assignment in agrammatic aphasia. Aphasiology 15, 515–531.

Beretta, A., Piñango, M., Patterson, K., Harford, C., 1999. Recruiting comparative crosslinguistic evidence to address competing accounts of agrammatic aphasia. Brain and Language 67, 149–168.

Beretta, A., Schmitt, C., Halliwell, J., Munn, A., Cuetos, F., Kim, S., 2001. The effects of scrambling on Spanish and Korean agrammatic interpretation: why linear models fail and structural models survive. Brain and Language 79, 407–425.

Blumstein, S.E., Byma, G., Kurowski, K., Hourihan, J., Brown, T., Hutchinson, A., 1998. On-line processing of filler-gap constructions in aphasia. Brain and Language 61, 149–168.

Broca, P., 1861. Remarques sur le siège de la faculté du langage articulé suivie d'une observation d'aphémie (perte de la parole). Bulletin de la Société d'Anatomie (Paris), 2e serie 36, 330–357.
Burkhardt, P., Avrutin, S., Piñango, M.M., Rulgendijk, E., 2008. Slower-than-normal syntactic processing in agrammatic Broca's aphasia: evidence from Dutch. Journal of Neurolinguistics 21, 120–137.
Caplan, D., Futter, C., 1986. Assignment of thematic roles by an agrammatic aphasic patient. Brain and Language 27, 117–135.
Caplan, D., Waters, G.S., 1999. Verbal working memory and sentence comprehension. Behavioral and Brain Sciences 22, 77–126.
Caramazza, A., Zurif, E.B., 1976. Dissociation of algorithmic and heuristic processes in language comprehension: evidence from aphasia. Brain and Language 3, 572–582.
Chomsky, N., 1981. Lectures on Government and Binding. Foris, Dordrecht.
Chomsky, N., 1995. Language and nature. Mind 104, 1–61.
Chomsky, N., 2000. Linguistics and brain science. In: Marantz, A., Miyashita, Y., O'Neill, W. (Eds.), Image, Language, Brain. MIT Press, Cambridge, MA, pp. 13–28.
Cinque, G., 1990. Types of A' dependencies. MIT Press, Cambridge, MA.
Crain, S., Ni, W., Shankweiler, D., 2001. Grammatism. Brain and Language 77, 294–304.
Friederici, A.D., Gorrell, P., 1998. Structural prominence and agrammatic theta-role assignment: a reconsideration of linear strategies. Brain and Language 65, 252–275.
Grodzinsky, Y., 1986. Cognitive deficits, their proper description, and its theoretical relevance. Brain and Language 27, 178–191.
Grodzinsky, Y., 1995. A restrictive theory of agrammatic comprehension. Brain and Language 50, 27–51.
Grodzinsky, Y., 2000. The neurology of syntax: language use without Broca's area. Behavioral and Brain Sciences 23, 1–71.
Grodzinsky, Y., 2006. A blueprint for a brain map of syntax. In: Grodzinsky, Y., Amunts, K. (Eds.), Broca's Region. Oxford University Press, Oxford, pp. 83–107.
Grodzinsky, Y., Pierce, A., Marakovitz, A., 1991. Neuropsychological reasons for a transformational analysis of verbal passive. Natural Language and Linguistic Theory 9, 431–453.
Haarman, H.J., Kolk, H.H.J., 1994. On-line sensitivity to subject-verb agreement violations in Broca's aphasics: the role of syntactic complexity and time. Brain and Language 46, 493–516.
Hagiwara, H., Caplan, D., 1990. Syntactic comprehension in Japanese aphasics: effects of category and thematic role order. Brain and Language 38, 159–170.
Hickok, G., Avrutin, S., 1995. Representation, referentiality, and processing in agrammatic comprehension: two case studies. Brain and Language 50, 10–26.
Hickok, G., Avrutin, S., 1996. Comprehension of *wh*-questions in two Broca's aphasics. Brain and Language 52, 314–327.
Hodgkin, A.L., Huxley, A.F., 1952a. Measurement of current-voltage relations in the membrane of the giant axon of Loligo. Journal of Physiology 116, 424–448.
Hodgkin, A.L., Huxley, A.F., 1952b. Currents carried by sodium and potassium ions through the membrane of the giant axon of Loligo. Journal of Physiology 116, 449–472.
Hodgkin, A.L., Huxley, A.F., 1952c. The components of membrane conductance in the giant axon of Loligo. Journal of Physiology 116, 473–496.
Hodgkin, A.L., Huxley, A.F., 1952d. The dual effect of membrane potential on sodium conductance in the giant axon of Loligo. Journal of Physiology 116, 497–506.
Hodgkin, A.L., Huxley, A.F., 1952e. A quantitative description of membrane current and its application to conduction and excitation in nerve. Journal of Physiology 117, 500–544.
Levin, B., Rappaport, M., 1986. The formation of adjectival passives. Linguistic Inquiry 17, 623–671.
Linebarger, M., 1995. Agrammatism as evidence about grammar. Brain and Language 50, 52–91.
Long, M.A., Fee, M.S., 2008. Using temperature to analyze temporal dynamics in the songbird motor pathway. Nature 456, 189–194.
Love, T., Swinney, D., Walenski, M., Zurif, E., 2008. How left inferior frontal cortex participates in syntactic processing: evidence from aphasia. Brain and Language 107, 203–219.
Lukatela, K., Shankweiler, D., Crain, S., 1991. Syntactic processing in agrammatic aphasia by speakers of a Slavic language. Brain and Language 49, 50–76.
Mauner, G., 1995. Examining he empirical and linguistic bases of current theories of agrammatism. Brain and Language 50, 339–368.
Phillips, C., 2012. Parser-grammar relations: we don't understand everything twice. In: Sanz, M., Laka, I., Tanenhaus, M. (Eds.), Language Down the Garden Path: The Cognitive and Biological Basis for Linguistic Structure. Oxford University Press, New York (Papers in honor of Thomas G. Bever).
Phillips, C., Lewis, S., 2013. Derivational order in syntax: evidence and architectural consequences. Studies in Linguistics 6, 11–47.
Piñango, M., 2000. Canonicity in Broca's sentence comprehension: the case of psychological verbs. In: Grodzinsky, Y., Shapiro, L., Swinney, D. (Eds.), Language and the Brain. Academic Press, New York.
Prather, P.A., Zurif, E., Love, T., Brownell, H., 1997. Speed of lexical activation in nonfluent Broca's aphasia and fluent Wernicke's aphasia. Brain and Language 59, 391–411.
Schumacher, P., 2009. Real-time comprehension in agrammatic aphasia. In: Ibanescu, G., Pescariu, S. (Eds.), Aphasia: Symptoms, Diagnosis and Treatment, Ch.4. Nova Science Publishers, Inc.

Linguistic Fieldwork

Andrea L Berez, University of Hawai'i at Mānoa, Honolulu, HI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by U. Mosel, volume 13, pp. 8906–8910, © 2001, Elsevier Ltd.

Abstract

This article describes the undertaking of linguistic fieldwork, in which linguists study languages as they are spoken in a naturally occurring domain, rather than in a laboratory or via self-introspection. This article presents the history of linguistic fieldwork, the rise of language documentation as a major paradigm for fieldwork, fieldwork research methods, ethics and community relations, and data management issues.

What Is Linguistic Fieldwork?

The traditional understanding of *linguistic fieldwork* involves a trained linguistic scientist traveling to a remote location to live in a speech community in order to study, describe, and theorize about that community's language or the human language capacity in general. In describing the nature of prototypical fieldwork, Hyman (2001) lists three features that linguistic fieldwork exhibits. The first of these is *distance*, meaning that the research usually takes place at some distance from one's home ("That is, 'fieldwork on Kikuyu' on one's vitae becomes synonymous with 'I did fieldwork on Kikuyu in Kenya'" (Hyman, 2001, p. 16)). The second of these is *exoticism*, meaning that the language under study is usually a minority language, an understudied language, an endangered language, etc. Hyman's third characteristic of prototypical fieldwork is *duration*, meaning that the research must take place over an extended period of time, with a commitment to spending many months in the field and years investigating the language.

While these traditional characteristics of fieldwork are considered by some to be the hallmarks of true linguistic fieldwork (e.g., Crowley, 2007; Samarin, 1967), in the last decade other less-prototypical scenarios are becoming increasingly accepted and even encouraged as valid examples of linguistic fieldwork. For example, fieldwork may take place in a speech community that is living in a diaspora situation away from its homeland, perhaps in a large urban setting like New York City or London. Fieldwork may also take place in a culturally defined subcommunity of the majority language community; for example, Eckert's (1989, 2000) and Bucholtz's (2011) work on language use among American high school students. Fieldwork may even be considered to start in the college classroom in a linguistics field methods course or by working with a lone individual, although these scenarios usually lead to the more traditional model of time spent in-community by the researcher later, in order to observe the language in use in a naturalistic setting.

What all these definitions have in common is that in fieldwork, the linguist attempts to study a language by observing it in one or more of its naturally occurring domains of use. While the notion of a *domain of use* can vary from a remote village where the language has been spoken for generations, to an urban marketplace, to a religious ceremony, to a staged conversation between two fluent speakers who are far from the home community, fieldwork is distinct from other linguistic research methods in that it strives to collect data by working with people who speak the language under study (as opposed to, say, introspective or corpus-based methods).

The products or outcomes expected from linguistic fieldwork are also changing. Previously the expected products of fieldwork were limited to those that primarily served linguistic science, whether descriptive (i.e., the so-called 'Boasian trilogy' of a dictionary, a descriptive grammar, and a collection of texts) or theoretical (examples from an 'exotic' language to prove or disprove theories) in aim. Today, however, the products of fieldwork can reach a much broader audience. The focus since Hale et al. (1992) on the endangered language crisis has led to an increase in the production of language teaching materials and revitalization programs as a goal of fieldwork, as well as increased attention to capacity building activities in language communities. The emerging discipline of language documentation, described below (and see, e.g., Gippert et al., 2006; Himmelmann, 1998, 2012; Grenoble and Furbee, 2010; Woodbury, 2011), has also led to the production of annotated digital media corpora consisting of audio and video recordings of language in use, along with time-aligned transcription, translation, and analysis as a central goal of linguistic fieldwork. These documentary corpora are intended to serve as wide an audience as possible, including linguists, language communities, teachers, and others.

Furthermore, the notion of who is a fieldworker is changing. Traditionally, linguistic fieldworkers have been professional linguists, graduate students, or Christian missionaries studying a language in order to facilitate Bible translation. Now, however, it is common to find members of the speech community themselves doing fieldwork; these people may or may not be trained in linguistic science, and may be teachers, clergy, or interested laypersons with a personal stake in the study of their native or heritage language.

In fact, the term *fieldwork* has recently come under scrutiny. This is because for many people who undertake the study of language *in situ*, the traditional notion of 'going to the field' does not hold – these people are often working in their home communities, with relatives or friends serving as language consultants. However, this article continues to use the term *fieldwork* as it is used in the sciences, to distinguish the study of language in its natural environment from that in

a laboratory setting; the terms *fieldworker* and *field linguist* here are used to indicate the person who undertakes such activities, whether or not that person is affiliated with an academic institution.

History of Linguistic Fieldwork

The history of linguistic fieldwork, broadly defined, goes back at least five centuries, when Spanish and Portuguese colonists in Africa, Asia, and the Americas embarked on the study of indigenous languages in the service of exploration and Christian mission work. Philologically grounded interest in the collection of wordlists was also popular in eighteenth- and nineteenth-century Europe among scholars and laypersons alike; see Chelliah and De Reuse (2010) for a full account of the early history of linguistic fieldwork.

Contemporary fieldwork is generally considered to have started with Franz Boas in the late nineteenth and early twentieth centuries (see Rosenblum and Berez, 2010; Sakel and Everett, 2012; Woodbury, 2011). In his anthropological work in indigenous communities along the Pacific Northwest coast of North America, Boas stressed the importance of language as an inextricable component of ethnological studies as a whole. He advocated the collection of texts, especially those written by native speakers, as part of the complete description of a language. This was a sharp turn from previous linguistic fieldwork, which focused on the production of generalizing descriptive grammars and dictionaries or wordlists as the essential products of fieldwork. Boas also became the progenitor of a tradition of American fieldwork-based linguistic inquiry via his arguably most-notable student Edward Sapir, who trained, among others, Mary Haas and Morris Swadesh. Haas then went on to train at least 50 graduate students in linguistic fieldwork through her position at the University of California, Berkeley, and the Survey of California and Other Indian Languages (Turner, 1997).

Turning to other regions, the history of fieldwork on the Aboriginal languages of Australia, summarized in McGregor (2008), can roughly be divided into three periods: an early period from the 1770s until about the 1930s that was marked by the collection of wordlists and a few descriptive grammars by amateurs; a middle period from approximately the mid-1930s until the 1960s, in which professional linguists conducted fieldwork, equipped with new ideas and institutional backing; and a late period from the 1960s to the present, in which the presence of missionary linguists and Aboriginal involvement in fieldwork has expanded rapidly (see Blake, 1981; Capell, 1970; Dixon, 1980; Dixon and Blake, 1979; McGregor, 2004; O'Grady et al., 1966; Wurm, 1972 for more details).

Furthermore, fieldwork in the Pacific region – and especially on languages of the Austronesian family – was facilitated during the second half of the twentieth century in large part by linguists working at the University of Hawai'i at Mānoa. There, the Pacific Lexicography Center and, later, the Pacific and Asian Linguistics Institute, supported fieldwork in languages of Micronesia, Polynesia, and the Philippines (Grace, unpublished).

During the rise of generative linguistics in America in the second half of the twentieth century, fieldwork, while still a popular research methodology in Australia and the Pacific, fell out of favor in the United States (with the exception of research on languages of the Americas as part of the Boasian tradition described above). As Chelliah and De Reuse (2010) note:

> The Chomskyan program saw the biological unity of a language learning device and a universal grammar underlying that device. Since the grammar of all languages was presumed to be the same, the study of exotic unwritten languages was not seen as necessary and it was thought that the goals of understanding universal grammar could just as well be accomplished by studying well-known languages. Another advantage of using well-known languages was that speakers were readily available and easily trained in providing grammaticality judgments or introspective comments on a language.
>
> Chelliah and De Reuse, 2010, pp. 60–61

The Rise of Language Documentation as a Paradigm for Fieldwork

However, since the early 1990s, attention has once again turned to the value of linguistic fieldwork, spurred in particular by the crisis of language endangerment and extinction (e.g., Austin and Sallabank, 2011; Evans, 2009; Grenoble and Whaley, 1998; Hale et al., 1992; Harrison, 2007; Nettle and Romaine, 2000). Krauss's (in Hale et al., 1992) admonition that linguists not let the discipline "go down in history as the only science that presided obliviously over the disappearance of 90% of the very field to which it is dedicated" (Krauss in Hale et al., 1992, p. 10) has given rise to a new paradigm of linguistic fieldwork called alternately *language documentation* or *documentary linguistics*. (Some suggest that the term *documentary linguistics* situates this new paradigm as a subfield of linguistics (like *sociolinguistics* or *historical linguistics*), while *language documentation* remains agnostic or relegates it to an applied status. Except in direct quotes, for this article I use *language documentation* because it has become the more widely used of the two terms, although I still maintain that the practice is as worthy of theorization as any other subfield of linguistics (cf Woodbury, 2011).)

Himmelmann (1998) distinguishes the *documentation* of languages from the *description* of languages, both of which arise from fieldwork:

> Much of the work that is labeled 'descriptive' within linguistics comprises two activities, the collection of primary data and a (low-level) analysis of these data. These are indeed two separate activities as shown by the fact that the methods employed in each activity differ substantially. To date, the field concerned with the first activity – called 'documentary linguistics' here – has received very little attention from linguists. [I propose] that documentary linguistics be conceived of as a fairly independent field of linguistic inquiry and practice that is no longer linked exclusively to the descriptive framework.
>
> Himmelmann, 1998, p. 161

The rise of language documentation has had a tremendous affect on linguistic fieldwork in the last two decades. Funding agencies have arisen exclusively to fund fieldwork projects

that are documentary in nature (e.g., The Hans Rausing Endangered Language Documentation Programme, the Dokumentation Bedrohter Sprachen of the VolkswagenStiftung, the National Science Foundation/National Endowment for the Humanities Documenting Endangered Languages Program, the Foundation for Endangered Languages, and the Endangered Language Fund). Fieldwork textbooks now discuss documentation methodology (e.g., Bowern, 2008), and academic programs to grant postgraduate degrees in language documentation have been developed (e.g., in the Department of Linguistics at the University of Hawai'i at Mānoa and the Endangered Language Academic Programme at the School of Oriental and African Studies of the University of London). Major research initiatives have explored the infrastructure needed to support language documentation (e.g., The Electronic Metastructure for Endangered Languages Data project (EMELD); inNET). Several edited volumes on documentation and documentary fieldwork (e.g., Gippert et al., 2006; Grenoble and Furbee, 2010; Thieberger, 2012) as well as journals and series (*Language Documentation & Conservation*, *Language Documentation and Description*) have been published.

Summarily, the goals of language documentation fieldwork include the creation of a corpus of audio and video records of observable linguistic behavior, the annotation of those records with enough translation, analysis, and metadata that the records will be useful and accessible to a wide audience, and the responsible preservation and dissemination of those records and annotations (see, e.g., Himmelmann, 1998, 2006; Woodbury, 2011 for discussion). Documentation also goes hand in hand with the field of language revitalization, maintenance, and conservation.

Fieldwork Research Methods and Activities

A great deal of work goes into preparing for fieldwork. These include selecting the language of study, finding an appropriate fieldsite, developing a research agenda, getting visas, finding funding, purchasing equipment, and learning as much as possible about the language, culture, and field site.

Once in the field, fieldworkers typically think of their work as occurring in *sessions*, usually 1–3 h of time spent working with one or more speakers of the language (these speakers are usually called *consultants*, *informants*, *teachers*, *assistants*, etc.). A typical day of fieldwork will consist of one or two sessions, plus preparation time before each session and processing time after each session. A range of activities can take place during a session: *elicitation* of lexical, grammatical, and metalinguistic knowledge; *recording* examples of language in use; and *transcription and translation* of those recordings. In reality the division between these activities is not so clearly defined. For example, during transcription and translation work the linguist and consultant may break away to investigate an interesting grammatical phenomenon or to collect lexical data from a particular semantic field.

Elicitation is the direct questioning of the consultant(s) about a particular aspect of the language. Usually elicitation of lexical items (vocabulary) is the first kind of work to take place between a researcher and a consultant; the researcher will ask the consultant for words of simple nouns like body parts and use this information to understand the phonology of the language in question (see Bowern, 2008 for advice on initial phonetic/phonological fieldwork; Ladefoged, 2003 for advanced instrumental fieldwork on phonetics). Lexical elicitation can also be a part of more advanced fieldwork, in service of producing dictionaries (Frawley et al., 2002; Mosel, 2004; Nichols and Sprouse, 2003). Elicitation can also be used to investigate other parts of grammar, including morphology and syntax. The linguist may ask the consultant 'how do you say X in your language?' or she may provide a scenario and ask the consultant how that scenario is usually expressed in the language ('if you were thirsty, how would you ask for something to drink?'). In some cases the linguist may provide constructions in the language and ask the consultant to explain the difference, or the linguist may ask whether a certain construction is grammatical or not. Some consultants can quickly become adept at providing paradigms. Elicitation can also be aided with the use of short video clips of actions or states as stimulus, like those provided online by the Language and Cognition Group at the Max Planck Institute for Psycholinguistics in Nijmegen.

Elicitation, while a useful method for investigating some aspects of language, requires the consultant to reflect consciously on his or her knowledge of the language. This is cognitively a different kind of linguistic behavior from using language for communication, and thus it is unwise for linguists to base their research solely on elicitation. It is very important to also include the *recording of examples of language in use* in a range of genres and modalities – indeed these kinds of data make up the bulk of a language documentation corpus. Examples of language in use include nearly any kind of running speech like narration, conversation, instructional discourse, oration, etc. These kinds of language use may be more or less spontaneous (see Himmelmann, 1998) and may be either staged for the researcher or naturally occurring.

Staged or controlled speech events for recording may include asking a consultant to tell a traditional story in the language, or asking a consultant to talk about a personal experience, either of her own choosing or in response to a question like 'have you ever been in mortal danger?' or 'tell me about your first day of school.' Visual stimuli may also be used; entire research paradigms have grown out of the use of the 'pear film' (Chafe, 1980) and the 'frog story' (Strömqvist and Verhoeven, 2004) in fieldwork. (The 'pear film' is a wordless video about a pear farmer, and a boy, and a bicycle, while the 'frog story' is a wordless picture book about a boy, a dog, and a frog (Meyer, 1969). In these research paradigms, consultants are asked to narrate these in their language.) Conversational speech events may also be somewhat staged by providing consultants with a topic for discussion; Bowern (2008, p. 123) suggests games like 'what am I thinking of?' and giving directions. Naturally occurring speech events are also potentially recordable and should be included in a well-rounded corpus. These include public events like ceremonies, performances, lectures, radio programs and the like, as well as conversations among consenting participants (not necessarily in the presence of the linguist).

Recordings of language are of little use if they are not accessible to people who do not speak the language, and therefore transcription and translation are key fieldwork activities.

Transcription is the writing (or typing) of the contents of the recording, usually either in the International Phonetic Alphabet or in a locally accepted orthography. Translation involves translating the transcription into another language (usually the language of wider communication and/or a major language like English), which can happen in units of various sizes. Transcriptions can be translated at the level of phrase or sentence (usually referred to as a *free translation*), at the word level, or at the morpheme level. In all cases, transcription and translation involve analysis, of phonology during transcription and of morphology, syntax, semantics, and discourse during translation.

If the field linguist is also a fluent speaker and writer of the language under study, transcription can be a fairly rapid process. However, if the field linguist is not a speaker of the language, then she must work with a consultant, who can repeat the words on the recording slowly and clearly to the linguist, who will write them down. Translation is also a time-consuming task, and the linguist should take care to avoid the range of pitfalls that come with translating between one language and another (Crowley, 2007; Gil, 2001).

Note that no single method is enough. Consultants who are skilled at metalinguistic activities may not be fluent enough or comfortable enough to provide running speech, and fluent speakers are not necessarily adept at introspection. Importantly, a phrase that sounds ungrammatical out of context may indeed be a licit part of a naturalistic speech event, and thus grammaticality judgments can be suspect. In reality, fieldwork involves all research methods described here happening simultaneously and circularly. As the field researcher learns more about the language, she may need to go back and re-elicit, retranslate, or reanalyze previous work.

Ethics and Linguist–Community Relations

As in medical and biological research, linguists are bound to practice ethical research methods during fieldwork. Ethics has been the subject of a great deal of discussion in the last few decades, with special attention to the notion of community collaboration (discussed below) in language documentation. Because they work with human subjects, most linguists are bound by university-based ethics panels, also referred to as Human Subject Committees or Internal Review Boards. Linguistic fieldwork projects must be approved or exempted by such a committee before the researcher can commence work.

However, given that linguistic fieldwork takes place in parts of the world that may be culturally quite diverse, most researchers agree that it is imperative to consider one's ethical responsibilities beyond what the university ethics panel has approved, since practices for, e.g., consent, may differ from those in the researcher's home country. The Linguistic Society of America has provided some guidelines for ethical field-based research:

- Linguists should do everything in their power to ensure that their research poses no threat to the well-being of research participants.
- Research participants have the right to control whether their actions are recorded in such a way that they can be connected with their personal identity. They also have the right to control who will have access to the resulting data, with full knowledge of the potential consequences.
- Linguists are responsible for obtaining the informed consent of those who provide them with data (regardless of whether and how that consent is documented), for maintaining any expected confidentiality in storing data and reporting findings, and for ensuring that any individual's participation in their research is completely voluntary at every stage. Anonymous observations of public behavior, which often cannot involve consent, should include no information that could inadvertently identify individuals or, where sensitive, the community.
- Linguists should carefully consider whether compensation of some kind is appropriate, be it remuneration for time and effort, or use of their knowledge and expertise to benefit participants or their communities.
- Where feasible, linguists should facilitate participants' access to their research results (Linguistic Society of America, 2009: pp. 2–3, reprinted in Rice, 2012, p. 410).

These guidelines are designed to protect the individual research participant while still being flexible enough to take cultural differences into account. In traditional social science research, anonymity of a participant is often an inviolable component of project design, but in some language communities, speakers may take pride in having their names associated with their utterances; the second bullet point above allows the speaker to make decisions with regard to his or her anonymity. Bullet point three requires that researchers ensure that participants consent to the research activity, but does not require a signature to that effect, since in many communities paperwork is regarded with suspicion (in such cases, verbal consent would be recorded). Bullet point four allows researchers and consultants to determine locally appropriate compensation, to avoid disrupting the local economy or causing offense (McLaughlin and Seydou Sall, 2001).

In addition to the ethical considerations between linguist and individual research participants, fieldwork also requires ethical consideration between the linguist and the speaker community. Although in many cases it is difficult to determine which individuals or bodies best represent the community, it is now well accepted that fieldworkers must strive to also take the needs, interests, and expertise of the speakers themselves into consideration when designing and conducting field-based projects.

In recent years the discussion about linguist–community relations in fieldwork and language documentation has tended to embrace a collaborative model (Cameron et al., 1992; Czaykowska-Higgins, 2009; Dorian, 2002; Dwyer, 2006; Gerdts, 1998; Grinevald, 1998, 2003; Hale, 2001; Hill, 2002; Penfield et al., 2008; Rice, 2006, 2011, 2012; Stebbins, 2003; Yamada, 2007), in which the linguist works cooperatively with the community to mutually meet the language goals of the researchers and the speakers simultaneously. At the very least, the linguist integrates into her research plans to return something of potential linguistic or educational value back to the community while also working toward her own research goals. For example, a field linguist writing a grammar may also use some of the information she collects to create language

teaching materials for teachers to use in local language maintenance programs, or she may install digital 'jukeboxes' of language recordings of traditional narratives on a computer in the local library for all to listen to. Other publications on collaborative research support a much more proactive and equitable relationship between the researcher and the community. Czaykowska-Higgins (2009), for example, extends the notion of *community-based research* (Cameron et al., 1992; Strand et al., 2003) to field linguistics, advocating research that is conducted *for, with, and by* community members. In this model, members of the speech community have an equal say in the design, implementation, and ownership of a fieldwork project:

> In its fullest form, Community-Based Language Research involves training members of the language-using community to do the research themselves, and can have as one of its goals the aim of making redundant the presence in the community of academic linguists who are not from the community.
>
> Czaykowska-Higgins, 2009, p. 25

Rice (2011) points out that in some parts of the world, communities already have guidelines for ethical research and/or research protocols in place to which outside field linguists must adhere. Canadian First Nations communities have been especially active in designing research protocols, see for example the discussion of the development of the Clearwater River Denesųłiné protocol in Cheecham and Wilhelm (2013) and the Gwich'in protocol for traditional knowledge research (Gwich'in Social and Cultural Institute, 2004). In the United States, the Alaska Native Knowledge Network provides guidelines for respecting cultural knowledge and strengthening indigenous languages when working with Alaska Native language communities (Alaska Native Knowledge Network, 2000, 2001).

The discussion on ethics and linguist–community relations has also recently included viewpoints that are distinct from the collaborative research model. Most notably are Dobrin (2008) and Holton (2009), who point out that while collaboration toward linguistic goals is laudable in places like North America, Australia, and New Zealand, in other parts of the world like Melanesia and Indonesia, there are other standards for ethical behavior by outsiders. Communities may have nonlinguistic goals for their relationships with field linguists. Dobrin describes experiences with fieldwork in Papua New Guinea, where having a positive relationship with an outside researcher was perhaps more important than actual language survival, and where 'the right to language choice includes the right to choose against language' (Dobrin, 2008, p. 300). In the case of Holton (2009), differing expectations for his role as patron as well as linguist led to awkwardness in his relations with a speech community in eastern Indonesia.

Finally, there is an ethical concern for the long-term care of the data collected during fieldwork. Especially when so many of the world languages, and the knowledge systems encoded within them, are endangered, fieldworkers have a responsibility to make recordings that are of the highest possible quality, to describe languages accurately, to ensure the longevity of the data collected, and to disseminate them responsibly. *Responsibly* in this case means finding a balance between the desires of the speakers for privacy and ownership of languages and recordings, and the need to manage one's data according to accepted best practices for fieldwork data sustainability. These best practices are discussed in the next section.

Data Management, Technology, and Archiving in Fieldwork

In their 2003 article, Bird and Simons brought to the forefront of the discipline of linguistics the importance of fully exploiting technology to create documentations of language that are long-lasting and multipurposeable. As field linguists increasingly took advantage of new technologies for capturing, storing, and disseminating language data, "new digital language resources of all kinds – lexicons, interlinear texts, grammars, language maps, field notes, recordings – [were] proving difficult to reuse and less portable than the conventional printed resources they replace" (Bird and Simons, 2003, p. 557). Since then, a great deal of research on digital processes, software, and infrastructure for linguistic fieldwork has taken place (e.g., EMELD, inNet). Technology and data management are primary concerns for the field linguist, who is responsible for not only documenting language but also seeing to the longevity and appropriate dissemination of the data.

The result is that field linguists must plan for and incorporate robust data management tasks into their fieldwork plans, and they must attend to these tasks both in the field while collecting data, and out, when undertaking postfieldwork processing. First, the linguist must plan for good-quality audio and video recording, including becoming familiar with the latest digital recording devices and microphones, understanding digital file formats like WAV and MP3, and learning how to arrange a recording environment for optimal sound or video quality with no background noises, disruptive light sources, etc. (see Margetts and Margetts, 2012).

The field linguist must also have a consistent plan for file management and metadata collection. File management tasks include selecting an appropriate file naming convention that gives each file a unique identifier, and developing a strategy for backing up files in the field. This may be especially challenging in field locations where electricity or dry storage conditions are in short supply. Metadata, which is key information about each file (e.g., date, language, location, demographic information about the people on the recording, etc.), must also be diligently collected (see Thieberger and Berez, 2012 for advice on data management in fieldwork).

The field linguist must also learn how to use and evaluate different pieces of software that she might use during data processing. For example, two of the most common software tasks for field linguistics are (1) time-aligning transcriptions and translation to audio and/or video, and (2) developing a lexical database. (*Time alignment* is the linking of bits of text (here, transcriptions and translations, but also possibly glosses, comments, notation of gesture, etc.) to the time code of a particular event in a piece of media. In many cases this is done by (a) referencing the media file and (b) referencing the

start and end times of, say, particular utterances within a marked-up text document of a transcription of that media file. See Berez (2007) and Thieberger and Berez (2012).) and (2) developing a lexical database. Before selecting software, field linguists must determine whether the output of the software is in an interoperable, human- and machine-readable format like XML, lest data eventually become inaccessible in the future, if the developer ever stops supporting the software. (Fortunately, there are already prominent pieces of software available for these linguistic tasks that field linguists use quite a bit: ELAN for time alignment of text to media and Fieldworks Language Explorer (FLEx) for building lexical databases and interlinearizing texts.)

A typical workflow for processing data collected during linguistic fieldwork is as follows. After the initial planning phases (which include determining a file name convention; coming up with a plan for backing up data in the field; and finding an archive to work with (more on this last below)), the fieldworker will make recordings during a session and also immediately collect metadata for that session. Audio and video data must then be transcribed, translated, and interlinearized, and lexical items must be entered into a lexical database. These items can then be used for analysis and in printed, online, and teaching materials, etc.

At each stage of the process – that is, after making recordings, after translating them, and after analyzing them – data should be regularly deposited in an archive. Field linguists should find a dedicated repository to work with before they leave for fieldwork, because the archive staff can make recommendations for proper depositing procedures. Many field linguists now make deposits via post directly from the field, or immediately upon returning home. There are a number of accepted digital language archives in the world, like those listed with the Open Languages Archive Community. Note that not every website, hard drive, or computer counts as an acceptable archive, and care should be taken that materials are safeguarded in a repository whose staff is knowledgeable about archiving language materials. Materials that are deemed private can be embargoed for a determined period of time, or access to materials can be made open to the public or restricted to a particular group of people (like members of the speech community).

See also: Corpus Linguistics; Fieldwork in Social and Cultural Anthropology; Language Policy; Language and Society; Linguistic Anthropology; Multilingualism; Pidgin and Creole Languages; Sapir-Whorf Hypothesis.

Bibliography

Alaska Native Knowledge Network, 2000. Guidelines for Respecting Cultural Knowledge. Assembly of Alaska Native Educators, Anchorage.

Alaska Native Knowledge Network, 2001. Guidelines for Strengthening Indigenous Languages. Assembly of Alaska Native Educators, Anchorage.

Austin, P.K., Sallabank, J. (Eds.), 2011. The Cambridge Handbook of Endangered Languages. Cambridge University Press, Cambridge.

Berez, A.L., 2007. Software review of EUDICO Linguistic Annotator (ELAN). Language Documentation & Conservation 1, 283–289.

Bird, S., Simons, G., 2003. Seven dimensions of portability for language documentation and description. Language 79, 557–582.

Blake, B.J., 1981. Australian Aboriginal Languages. Angus and Robertson, Sydney.

Bowern, C., 2008. Linguistic Fieldwork: A Practical Guide. Palgrave Macmillan, New York.

Bucholtz, M., 2011. White Kids: Language, Race, and Styles of Youth Identity. Cambridge University Press, Cambridge.

Cameron, D., Frazer, E., Harvey, P., Rampton, M.B.H., Richardson, K., 1992. Researching Language: Issues of Power and Method. Routledge, London and New York.

Capell, A., 1970. History of research in Australian and Tasmanian languages. In: Sebeok, T.A. (Ed.), Current Trends in Linguistics, Linguistics in Oceania, vol. 8. Mouton, The Hague, pp. 661–720.

Chafe, W., 1980. The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production. Ablex, Norwood, New Jersey.

Cheecham, C., Wilhelm, A., 2013. Sharing Worlds of Knowledge: Research Protocols for Communities. http://hdl.handle.net/10125/26110 (accessed 29.12.13.).

Chelliah, S.L., De Reuse, W.J., 2010. Handbook of Descriptive Linguistic Fieldwork. Springer, Dordrecht.

Crowley, T., 2007. Field Linguistics: A Beginner's Guide. Oxford University Press, Oxford.

Czaykowska-Higgins, E., 2009. Research models, community engagement, and linguistic fieldwork: reflections on working within Canadian indigenous communities. Language Documentation & Conservation 3, 15–50.

Dixon, R.M.W., 1980. The Languages of Australia. Cambridge University Press, Cambridge.

Dixon, R.M.W., Blake, B.J. (Eds.), 1979. Handbook of Australian Languages, vol. 1. Australian National University Press, Canberra.

Dobrin, L.M., 2008. From linguistic elicitation to eliciting the linguist: lessons in community empowerment from Melanesia. Language 84, 300–324.

Dorian, N., 2002. Commentary: broadening the rhetorical and descriptive horizons in endangered-language linguistics. Journal of Linguistic Anthropology 12, 134–140.

Dwyer, A.M., 2006. Ethics and practicalities of cooperative fieldwork and analysis. In: Gippert, J., Himmelmann, N.P., Mosel, U. (Eds.), Essentials of Language Documentation. Mouton de Gruyter, Berlin and New York, pp. 31–66.

Eckert, P., 1989. Jocks and Burnouts: Social Categories and Identity in the High School. Teachers College Press, New York.

Eckert, P., 2000. Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten. Wiley Blackwell, Malden, MA.

Evans, N., 2009. Dying Words: Endangered Languages and What They Have to Tell Us. Wiley-Blackwell, Malden, MA.

Frawley, W., Hill, K., Munro, P., 2002. Making Dictionaries: Preserving Indigenous Languages of the Americas. University of California Press, Berkeley and Los Angeles.

Gerdts, D., 1998. Beyond expertise: the role of the linguist in language revitalization programs. In: Ostler, N. (Ed.), Endangered Languages: What Role for the Specialist? Proceedings of Second Foundation for Endangered Languages Conference. Foundation for Endangered Languages, Bath, pp. 13–22.

Gil, D., 2001. Escaping Eurocentrism: fieldwork as a processs of unlearning. In: Newman, P., Ratliff, M. (Eds.), Linguistic Fieldwork. Cambridge University Press, Cambridge, pp. 102–132.

Gippert, J., Himmelmann, N.P., Mosel, U. (Eds.), 2006. Essentials of Language Documentation. Mouton de Gruyter, Berlin and New York.

Grace, G. On the History of the Department of Linguistics at the University of Hawai'i at Mānoa, unpublished manuscript.

Grenoble, L.A., Furbee, N.L. (Eds.), 2010. Language Documentation: Practice and Values. John Benjamins, Amsterdam.

Grenoble, L.A., Whaley, L. (Eds.), 1998. Endangered Languages: Current Issues and Future Prospects. Cambridge University Press, Cambridge.

Grinevald, C., 1998. Language endangerment in South America: a programmatic approach. In: Grenoble, L.A., Whaley, L. (Eds.), Endangered Languages: Current Issues and Future Prospects. Cambridge University Press, Cambridge, pp. 124–160.

Grinevald, C., 2003. Speakers and documentation of endangered languages. In: Austin, P.K. (Ed.), Language Documentation and Description, vol. 1. School of Oriental and African Studies, London, pp. 71–101.

Gwich'in Social and Cultural Institute, 2004. Working with Gwich'in Traditional Knowledge in the Gwich'in Settlement Region. http://www.grrb.nt.ca/pdf/GTCTKPolicy.pdf (accessed 29.12.13.).

Hale, K., 2001. Ulwa (Southern Sumu): the beginnings of a language research project. In: Newman, P., Ratliff, M. (Eds.), Linguistic Fieldwork. Cambridge University Press, Cambridge, pp. 71–101.

Hale, K., Krauss, M., Watahomigie, L.J., Yamamoto, A.Y., Craig, C., Jeanne, L.M., England, N.C., 1992. Endangered languages. Language 68, 1–42.

Harrison, K.D., 2007. When Languages Die: The Extinction of the World's Languages and the Erosion of Human Knowledge. Oxford University Press, Oxford.
Hill, J., 2002. "Expert rhetorics" in advocacy for endangered languages: who is listening and what do they hear? Journal of Linguistic Anthropology 12, 119–133.
Himmelmann, N.P., 1998. Documentary and descriptive linguistics. Linguistics 36, 161–195.
Himmelmann, N.P., 2006. Language documentation: what is it good for? In: Gippert, J., Himmelmann, N.P., Mosel, U. (Eds.), Essentials of Language Documentation. Mouton de Gruyter, Berlin, pp. 1–30.
Himmelmann, N.P., 2012. Linguistic data types and the interface between language documentation and description. Language Documentation & Conservation 6, 187–207.
Holton, G., 2009. Relatively ethical: a comparison of linguistic research paradigms in Alaska and Indonesia. Language Documentation & Conservation 3, 161–175.
Hyman, L., 2001. Fieldwork as a state of mind. In: Newman, P., Ratliff, M. (Eds.), Linguistic Fieldwork. Cambridge University Press, Cambridge, pp. 15–33.
Ladefoged, P., 2003. Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques. Wiley-Blackwell, Walden, Massachusetts.
Linguistic Society of America, 2009. Ethics Statement. http://www.linguisticsociety.org/files/Ethics_Statement.pdf (accessed 29.12.13.).
Margetts, Anna, Margetts, Andrew, 2012. Audio and video recording techniques for linguistic research. In: Thieberger, N. (Ed.), The Oxford Handbook of Linguistic Fieldwork. Oxford University Press, Oxford, pp. 13–53.
McGregor, W.B., 2004. The Languages of the Kimberley, Western Australia. RoutledgeCurzon, London.
McGregor, W.B. (Ed.), 2008. Encountering Aboriginal Languages: Studies in the History of Australian Linguistics. Pacific Linguistics, Canberra.
McLaughlin, F., Seydou Sall, T., 2001. The give and take of fieldwork: noun classes and other concerns in Fatick, Senegal. In: Newman, P., Ratliff, M. (Eds.), Linguistic Fieldwork. Cambridge University Press, Cambridge, pp. 189–210.
Meyer, M., 1969. Frog, Where Are You? Dial Books for Young Readers, New York.
Mosel, U., 2004. Dictionary making in endangered speech communities. In: Austin, P.K. (Ed.), Language Documentation and Description, vol. 2. School of Oriental and African Studies, London.
Nettle, D., Romaine, S., 2000. Vanishing Voices: The Extinction of the World's Languages. Oxford University Press, Oxford.
Newman, P., Ratliff, M. (Eds.), 2001. Linguistic Fieldwork. Cambridge University Press, Cambridge.
Nichols, J., Sprouse, R.L., 2003. Documenting lexicons: Chechen and Ingush. In: Austin, P.K. (Ed.), Language Documentation and Description, vol. 1. School of Oriental and African Studies, London.
O'Grady, G.N., Voegelin, C.F., Voegelin, F.M., 1966. Languages of the world: Indo-Pacific. Anthropological Linguistics 8, 1–197.
Penfield, S.D., Serratos, A., Tucker, B.V., Flores, A., Harper, G., Hill Jr., J., Vasquez, N., 2008. Community collaborations: best practices for North American indigenous language documentation. International Journal of the Sociology of Language 191, 187–202.
Rice, K., 2006. Ethical issues in linguistic fieldwork: an overview. Journal of Academic Ethics 4, 123–155.
Rice, K., 2011. Documentary linguistics and community relations. Language Documentation & Conservation 5, 187–207.
Rice, K., 2012. Ethical issues in linguistic fieldwork. In: Thieberger, N. (Ed.), The Oxford Handbook of Linguistic Fieldwork. Oxford University Press, Oxford, pp. 407–429.
Rosenblum, D., Berez, A.L., 2010. Introduction: the Boasian tradition and contemporary practice in linguistic fieldwork in the Americas. In: Berez, A.L., Mulder, J., Rosenblum, D. (Eds.), Fieldwork and Linguistic Analysis in Indigenous Languages of the Americas (Language Documentation & Conservation Special Publication 2). University of Hawai'i Press, Honolulu, pp. 1–8.
Sakel, J., Everett, D.L., 2012. Linguistic Fieldwork. Cambridge University Press, Cambridge.
Samarin, W.J., 1967. Field Linguistics: A Guide to Linguist Field Work. Holt, Rinehart and Winston, New York.
Stebbins, T., 2003. Fighting Language Endangerment: Community Directed Research on Sm'algyax (Coast Tsimshian). Endangered Languages of the Pacific Rim Publication Series A2–026.
Strand, L., Marullo, S., Cutforth, N., Stoecker, R., Donohue, P., 2003. Community-Based Research and Higher Education: Principles and Practices. Jossey-Bass, San Francisco.
Strömqvist, S., Verhoeven, L., 2004. Relating Events in Narrative. In: Typological and Contextual Perspectives, vol. 2. Lawrence Erlbaum Associates, Mahwah, New Jersey; London.
Thieberger, N. (Ed.), 2012. The Oxford Handbook of Linguistic Fieldwork. Oxford University Press, Oxford.
Thieberger, N., Berez, A.L., 2012. Linguistic data management. In: Thieberger, N. (Ed.), The Oxford Handbook of Linguistic Fieldwork. Oxford University Press, Oxford, pp. 90–118.
Turner, K., 1997. Mary R. Haas: teacher. Anthropological Linguistics 39, 544–549.
Woodbury, A.C., 2011. Language documentation. In: Austin, P., Sallabank, J. (Eds.), The Cambridge Handbook of Endangered Languages. Cambridge University Press, Cambridge, pp. 159–186.
Wurm, S.A., 1972. Languages of Australia and Tasmania. Mouton, The Hague.
Yamada, R., 2007. Collaborative linguistic fieldwork: practical application of the empowerment model. Language Documentation and Conservation 1, 257–282.

Relevant Websites

http://dobes.mpi.nl/ – Dokumentation Bedrohter Sprachen (DoBeS).
http://www.emeld.org – EMELD.
http://tla.mpi.nl/tools/tla-tools/elan/ – ELAN Software for Time Alignment of Text to Media.
http://www.endangeredlanguagefund.org/ – Endangered Language Fund.
http://www.ogmios.org/ – Foundation for Endangered Languages.
http://fieldworks.sil.org/flex/ – Fieldwork Language Explorer Software.
http://www.hrelp.org/ – Hans Rausing Endangered Language Documentation Project (ELDP).
http://www.nflrc.hawaii.edu/ldc/default.html – Language Documentation & Conservation (Journal).
http://fieldmanuals.mpi.nl/ – L&C Field Manuals and Stimulus Materials, Max Planck Institute, Nijmegen.
http://www.innet-project.eu – inNet.
http://www.language-archives.org/ – OLAC.

Linguistic Presupposition

Rob A van der Sandt, University of Nijmegen, Nijmegen, The Netherlands

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8910–8913, © 2001, Elsevier Ltd.

Abstract

Presuppositional information is encoded in linguistic expressions by means of so-called presupposition triggers or presupposition inducers. These items are linguistic means that signal that certain information is taken for granted in the linguistic or nonlinguistic context. One task for presupposition theory is to identify these triggers, explain what type of information they induce, and why they do so. A second task is to explain why and under what circumstances the information thus induced is preserved as an intuitive inference of the inducing sentence. The latter problem is the projection problem for presuppositions. Theories of presupposition can be divided into the following main classes: semantic, pragmatic, and dynamic theories. Semantic theories explain presupposition as a semantic relation between sentences. Pragmatic theories take the information a speaker takes for granted in making an assertion as their starting point. Dynamic theories locate sentence meaning in its context-change potential: presuppositional information should be part of input context as a precondition for contextual update. The two major variants are the satisfaction account and the anaphoric account. The former construes presuppositions as definedness conditions on input contexts. According to the anaphoric account presuppositional expressions are anaphors that should be bound to a contextually given antecedent.

Presuppositional information is information that is taken for granted by the participants in a discourse or a conversation. As such it has a different status from information the participants present as new or relevant. The latter information is said to be asserted. This distinction between presupposition and assertion is reflected in the linguistic structure of sentences. For example, by his use [1] a speaker typically indicates that he takes for granted that Adam ate the apple, and asserts that Thomas regrets that.

Thomas regrets that Adam ate the apple. [1]

The information that Adam ate the apple is called a presupposition of [1]. This presupposition is said to be linguistically induced by the use of the factive verb 'regret.'

This way of introducing the notion of presupposition calls for a distinction. And the basic distinction that should be made is the distinction between presupposition as information taken for granted and presupposition as induced, invoked, or triggered by a linguistic expression. The first notion is the notion of presupposition as background information, which is already given or taken for granted in a conversation. Stalnaker and Karttunen call it the common ground. The second notion is the notion of presupposition as information that is conventionally associated with linguistic expressions or syntactic constructions. It is found in the literature under a variety of names, potential or presupposition (Gazdar, 1979), conventional implicature (Karttunen and Peters, 1979), or elementary presuppositions (van der Sandt, 1988). Though the two notions of presupposition are fundamentally different there is a straightforward connection. If a linguistic element induces a presupposition, the sentence containing the inducing element will typically be appropriate in a context, which already contains the presuppositional information. Thus [1] will normally be judged appropriate in a context in which the information that Adam ate the apple is somehow established or taken for granted. If not, one should rather convey this information in an assertoric way as in [2].

Adam ate the apple and Thomas regrets that. [2]

Linguistic Encoding of Presuppositions

Linguistic presuppositions can thus be seen as pieces of information triggered or induced by certain linguistic structures. Such triggers can be identified by using specific tests. These tests rely on the fact that presuppositional information tends to escape from environments, which block ordinary entailments. The traditional test is constancy under negation. Other tests are embedding under epistemic modals, embedding in the antecedent of a conditional construction and questioning. Of these the negation test is the least reliable since negative sentences can, in many cases, be interpreted in a metalinguistic way so as to defeat the presuppositional inference (*see* Negation, Linguistics of).

Applying these tests to the verb *stop* identifies it as a presupposition inducer.

Hubert stopped smoking (at t). [3]
Hubert used to smoke (during some period preceding t). [4a]
Hubert did not smoke (after t). [4b]
Hubert did not stop smoking. [5a]
It is possible that Hubert stopped smoking. [5b]
If Hubert stopped smoking, he must be saving a lot of money. [5c]
Did Hubert stop smoking? [5d]

The presuppositional content of [3] is given in [4a], the nonpresuppositional content in [4b]. Only the presuppositional information is preserved as an inference from the embeddings in [5a]–[5d].

The following linguistic structures are widely recognized as presupposition inducers. The presuppositional information that is triggered by the relevant items is given in the b-sentences:

1. Definite Noun Phrases (definite descriptions often including proper names and indexicals; they induce a presupposition to the effect that there is some entity satisfying the description).

The King of France is bald. [6a]
There is a King of France. [6b]

 http://dx.doi.org/10.1016/B978-0-08-097086-8.52036-4

2. Cleft constructions (including 'that-' and pseudoclefts; they trigger an existential presupposition that there is some entity satisfying the Verb Phrase).

It was Adam who ate the apple. [7a]
Someone ate the apple. [7b]

3. Focal stress (induces the same existential presupposition as clefts, though many authors want to account for this inference by an independent theory of focus).

Adam ate the apple. [8a]
Someone ate the apple. [8b]

4. 'Wh'-questions (the presuppositional material is obtained by existential quantification over the variable introduced by the 'wh'-expression).

Who ate the apple? [9a]
Someone ate the apple. [9b]

5. Quantifiers (the strong quantifiers 'all,' 'most,'…); weak quantifiers like 'some' are considered to be ambiguous between a presuppositional and a nonpresuppositional reading; the presupposition induced is that the domain of quantification is nonempty.

All of John's children are bald. [10a]
John has children. [10b]

6. Factive verbs ('realize,' 'discover,' 'regret,'…; they induce a presupposition that the presuppositional complement is true).

Thomas regrets that Adam ate the apple. [11a]
Adam ate the apple. [11b]

7. Temporal clauses ('before,' 'while,' 'since,' 'after,'…); as in [6] they induce information of a propositional nature.

After Adam ate the apple, he was expelled from Paradise. [12a]
Adam ate the apple. [12b]

8. Sortal restrictions on predicates ('bachelor,' 'close,'…; the presuppositional information consists of the requirement that the arguments of the predicate are of the appropriate sort).

John is a bachelor. [13a]
John is male, adult,…. [13b]

9. Implicative verbs ('manage,' 'succeed,'…; put a presuppositional requirement on the action denoted by the VP).

John managed to open the door. [14a]
It took some effort to open the door. [14b]

10. Verbs of transition ('start,' 'stop,' 'continue,'…; see [3] above).

The items from [1]–[4] are examples of existential presuppositions. They all involve an existential quantification over some individual or set variable. In [6] and [7] the presuppositional information is of a propositional nature. The items from [8]–[10] involve requirements on the applicability of a predicate. This is only a small sample of the linguistic structures that have been identified as presupposition inducing. Further examples include focus particles ('only' and 'even'), adverbs of repetition ('again,' 'too,' 'back'), but also, more controversially, counterfactual conditionals, nonrestrictive relative clauses, and discourse particles like 'because' and 'although.'

Projection

The presuppositions induced by simple sentences tend to survive under embedding, that is whenever a subsentence of a compound φ induces the presuppositional information χ this information will surface as an inference of φ. Karttunen (1973) was the first to note and catalog numerous exceptions to this generalization. Here are some examples. In the a-sentences the presupposition that John has children escapes from its embedded position, in the b-sentences the same presupposition does not surface as an intuitive inference.

If baldness is hereditary, John's children are bald. [15a]
If John has children, John's children are bald. [15b]
Either baldness is not hereditary, or John's children are bald. [16a]
Either John has no children or John's children are bald. [16b]
It is possible that baldness is hereditary and that John's children are bald. [17a]
It is possible that John has children and that John's children are bald. [17b]

In the terminology prevalent in the 1970s and 1980s, it is said that, in the a-sentences, the presupposition that John has children is projected to the main sentence, whereas it is said to be filtered or canceled in the corresponding b-sentences.

Semantic, Pragmatic, and Discourse Presupposition

We may distinguish between three main types of theories that deal with presupposition and projection: semantic, pragmatic, and discourse theories.

Semantic Theories

The semantic view originates in Frege and was revived by Strawson. It construes presupposition as a binary relation between sentences (or propositions). A presupposition is a requirement that has to be fulfilled for a sentence to have a truth value. Thus ψ is taken to be a presupposition of φ just in case ψ has to be true in order that φ has a truth value. This requires a semantics that allows truth-value gaps.

Assuming the standard definition of negation as an operator, which maps truth onto falsity and vice versa, semantic presupposition can then be defined as follows:

φ semantically presupposes χ if φ entails χ and not φ entails χ.

The defining characteristic is constancy under negation. The projection behavior of presuppositions in other embeddings follows from the definitions of the logical connectives and other operators. One problem for such theories is that they cannot handle the presuppositions triggered by adverbs like 'too' and 'even.' Here presupposition failure arguably does not affect the truth conditions of the sentence, but the appropriateness of the utterance. It also turns out that no definition of the logical connectives can be given which accounts for all the projection data. Defenders of the semantic account are thus forced to either postulate an ambiguity in the logical connectives or transfer the burden of explanation to pragmatics.

Pragmatic Accounts

Pragmatic accounts (*see* Pragmatics, Linguistic) view presupposition not as a semantic relation between sentences. Instead presupposition is viewed as a property of utterances. The basic notion is that of a speaker presupposing a proposition in a context of utterance. This allows for certain variability of presuppositions with the context of utterance. The first formally developed theory of this kind is found in Gazdar (1979). The basic idea here is that a presupposition of a sentence is preserved unless it conflicts with contextual information or with an implicature invoked by its utterance. In the case of [15]–[17] this amounts to the following. A typical utterance of one of the b-sentences will invoke the implicature that it is possible that John has no children. This implicature conflicts with the presuppositional assumption that he does and so the latter is canceled. Since no such implicature is invoked for the a-sentences, they will be interpreted as presupposition preserving.

Dynamic Theories

A different style of modeling is found in the dynamic approaches. The two main variants are the so-called satisfaction theory and the anaphoric or binding theory. The first derives from Stalnaker (1974) and Karttunen (1974) and has been developed in particular by Heim (1983) and Beaver. The anaphoric account is found in van der Sandt (1992) and Geurts (1999). Comparisons and different assessments of their relative merits are given in Beaver (1997) and Geurts (1999).

Both accounts start off from the idea that an utterance changes the context in which it is made. The basic requirement of the satisfaction account is that the presuppositional information should be entailed by the context of utterance. If so, the input context is incremented with the information contained in the inducing sentence. An utterance of [1] thus requires a context, which entails that Adam ate the apple. If this requirement is not met, the input context is not a suitable one. In this case a mechanism of accommodation is invoked. This means that the input context will (under appropriate conditions) be revised so as to incorporate the relevant presuppositional information thus restoring interpretability.

The anaphoric or binding account is formulated as an extension of discourse representation theory. The basic tenet is that presuppositions are anaphoric expressions (*see* Anaphora). Presuppositions only differ from pronouns and other attenuated anaphors in that they have internal structure and semantic content. The basic requirement is not that the presuppositional information is entailed by the context of utterance, but that the presuppositional anaphor should find a suitable antecedent. If no such antecedent is available, the semantic material carried by the presuppositional anaphor will be accommodated in the discourse structure so as to provide an antecedent after all.

The context-change semantics of the satisfaction theory generally yields weaker predictions than are found in the anaphoric account. This is due to the different status that presuppositional material has in both theories and to the different ways in which this material is processed. For example, for a conditional of the form $\varphi \rightarrow \chi_{<\psi>}$ where ψ is triggered by χ the satisfaction theory will predict $\varphi \rightarrow \psi$ as presupposition for the whole sentence. This yields for [15a] not the presupposition that John has children, but the presupposition that he has children *provided* that baldness is hereditary. Just like the classic semantic account, the satisfaction theory has thus to appeal to some pragmatic strengthening mechanism to derive the stronger presupposition that John has children. On the anaphoric account, presuppositional information is never transformed into a new semantic construct. Consider [15a] again. The presuppositional anaphor associated with the consequent cannot be bound to some preestablished antecedent. So some entity representing the set of John's children will be accommodated in the main context, thus yielding the stronger prediction that the utterance of [15a] presupposes that John has children.

See also: Information Processing Architectures: Fundamental Issues; Linguistics: Overview; Pragmatics, Linguistic; Semantics; Sentence Comprehension, Psychology of.

Bibliography

Beaver, D.I., 1997. Presupposition. In: van Benthem, J.A., ter Meulen, A. (Eds.), Handbook of Logic and Language. Elsevier, Amsterdam, pp. 939–1008.

Gazdar, G., 1979. Pragmatics: Implicature, Presupposition, and Logical Form. Academic Press, New York.

Geurts, B., 1999. Presuppositions and Pronouns. Elsevier, Oxford, UK.

Heim, I., 1983. On the projection problem for presuppositions. In: Barlow, M., Flickinger, D.P., Wescoat, M.T. (Eds.), Proceedings of the West Coast Conference on Formal Linguistics, vol. 2. Stanford Linguistics Association, Stanford, CA, pp. 114–125. Reprinted from: Davis, S. (Ed.), Pragmatics. Oxford University Press, Oxford, UK, pp. 397–405.

Karttunen, L., 1973. Presuppositions of compound sentences. Linguistic Inquiry 4, 167–193.

Karttunen, L., 1974. Presupposition and linguistic context. Theoretical Linguistics 1, 181–194.

Karttunen, L., Peters, S., 1979. Conventional implicature. In: Oh, C.K., Dineen, D. (Eds.), Presupposition. Academic Press, New York, pp. 363–385.

van der Sandt, R.V., 1988. Context and Presupposition. CroomHelm/Routledge, London.

van der Sandt, R.A., 1992. Presupposition projection as anaphora resolution. Journal of Semantics 9, 333–377.

Stalnaker, R., 1974. Pragmatic presuppositions. In: Munitz, M.K., Unger, P.K. (Eds.), Semantics and Philosophy. New York University Press, New York, pp. 197–214.

Linguistic Synesthesia

Sean A Day, Trident Technical College, Charleston, SC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Congenital linguistic synesthesia involves aspects of language producing atypical automatic, involuntary, and consistently invariable sensory perceptions or associations, such as colors or flavors. While most commonly found with graphemes as the inducer and colors as the concurrent, other types involve phonemes and lexemes. Synesthesia differs from hallucinations or imagery in that the sensory perceptions are real and verifiable. Some forms of these neurological traits involve higher levels of myelination along unpruned neural pathways, producing atypical rates of feed-forward and feedback.

Definition

The word 'synesthesia' (British spelling, 'synaesthesia') comes directly from Greek συν- (syn-) 'union,' and αίσθησις (aisthesis) 'sensation,' thus meaning 'a union of the senses.' Synesthesia (pl. synesthesiae) is the general name for two complexes of over 60 related cognitive traits. In the first set, 'sensorial synesthesiae,' stimuli to one sense, such as smell, are involuntarily simultaneously perceived by one or more other, additional senses, such as sight and/or hearing. One highly documented case of synesthesia involved Michael Watson (Cytowic, 1993), who synesthetically felt at or within his right hand shapes and textures corresponding to different flavors.

With the second set of synesthesiae, called 'cognitive (category) synesthesiae,' which contains 'ordinal sequence' synesthesiae, certain sets of things that cultures teach us to put together, categorize, and also usually serialize in some specific way – such as letters, numbers, or people's names – get sensory addition, such as a smell, color, or flavor. Common forms of ordinal sequence synesthesiae involve graphemes, numbers, time units, and musical notes. Synesthesiae are usually talked about in terms of the 'inducer,' the sensory stimulus, or cognitive concept which initiates the synesthetic perception, and the 'concurrent,' the synesthetic perception itself. **Figure 1** represents a common set of synesthetically colored graphemes; note that some colors may repeat, that the entire color spectrum may not be represented, that many letters are faint and in shades of gray, and that numbers tend to be brighter primary colors.

The diagnostic criteria for synesthesia are as follows: synesthesia is (1) involuntary and must be elicited; (2) discrete; (3) generic regarding its perceptions; (4) typically unidirectional, but there are rare occurrences of bidirectionality (see, e.g., von Melchner et al., 2000); and (5) noetic (see Cytowic, 1989, 2002; Cytowic and Eagleman, 2009). Synesthesia is known for the persistent invariance of the association between an inducer and a concurrent over time; thus, one of the most common tests toward determining a diagnosis of synesthesia is a 'consistency test' (Baron-Cohen et al., 1987; Rich et al., 2005; Eagleman et al., 2007). Synesthesia is also highly idiosyncratic; the set of correspondences between inducers and concurrents held by one synesthete will have few overlaps with that held by another with the same type. Synesthesia is not a disease, nor, in most cases, a deficit; it is more correctly referred to as a 'trait.'

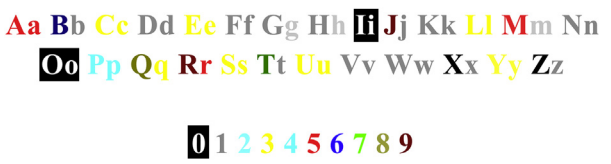

Figure 1 A typical synesthetically colored set of letter and number graphemes. Copyright by Joy A. Day, 2013.

Regarding visual synesthetic perceptions, about 80–85% of synesthetes who 'see things' do so 'inside the head,' as if the image is projected onto a screen, or just 'in the mind's eye.' The 'screen' is usually located 'on the inside of the forehead' or 'in the eyes,' although, for some, it may be at the top of the head or even in the back, at the base of the skull. For some in this group, there is instead a 'feeling of the color'; for example, the sound of middle C on a piano 'feels' red. These synesthetes have been labeled 'associators.' For those who do see things 'out there,' the images may range from less than a half meter away to fifty meters or more, although most are within the one-half to three-meter range. This latter group of synesthetes is labeled 'projectors' (see Dixon et al., 2004; Rouw and Scholte, 2010). It is possible for a multiple synesthete to be both a projector of one or more types and a nonprojector of other types, or even for a synesthete to be both with regard to one type, varying between projecting and nonprojecting with situations.

Recent studies place the ratio at about 1 in 27 people for finding any type of synesthesia in general across the population (Simner et al., 2006). The percentage of the general human population which has synesthesia varies with the type involved; estimates run from 1 in 250 for basic types of cognitive synesthesia, such as colored graphemes or colored musical pitches, to 1 in 3000 for more common forms of sensorial synesthesia, such as colored musical sounds or colored taste sensations, to 1 in 25 000 or more for people with rare or multiple forms of synesthesia, such as synesthetes who taste the things they touch. Simner et al. (2006) find a ratio of male to female at about 1:1. There are case reports of monozygotic twins, both female (Smilek et al., 2001a) and male (Smilek et al., 2005), in which one had synesthesia while the other did not.

Grapheme to Color

Roughly 1 in every 43 people worldwide has 'grapheme to color' synesthesia (Simner et al., 2006; Day, 2013). A grapheme is a written mark that designates a phoneme, syllable, morpheme, numerical concept, or other concept (e.g., a hand or body movement) via an arbitrary or iconic connection established and maintained by a social group. Phonemes are language-specific categories, each phoneme being a grouping of close, similar sounds that are cognitively generalized to be considered all variants of the same one 'speech sound.' While 'grapheme to visual' synesthesia is commonly referred to as 'colored letters (and numbers),' graphemes also include syllabic symbols, such as in the Japanese katakana system (e.g., 'カ') and ideograms (e.g., Chinese 猫). Likewise, →, ✉, and ☺ are also graphemes, but do not represent phonemic, syllabic, or morphemic elements of any currently written language.

Most research on grapheme-related synesthesia involves phonetic systems such as variants of the Roman or Cyrillic alphabets; yet synesthesia also occurs with systems such as Chinese and Egyptian ideograms, syllabary systems such as Japanese hiragana and Cherokee, and abugidas such as Bengali and Thai. Likewise Braille systems. Thus 'grapheme' transcends the visual to also include (at least) tactile modes. This is said with the awareness that, for many if not most initially sighted people who go blind or learn Braille for other reasons, vision is a major if not primary aspect of reading Braille; and also with the realization that, for those who have never been sighted, 'X to color,' or more broadly 'X to vision' synesthesia will most likely never occur. Nevertheless, for someone congenitally blind, we could encounter, for example, 'grapheme to flavor' synesthesia developing as the person learns a Braille system.

Synesthetes are not uniform in their correspondences between specific inducers and concurrents; for example, not all synesthetes see the letter 'S' as yellow. However, this has led some to the erroneous conclusion that synesthetes are wholly idiosyncratic, and that synesthetes' surrounding cultures and environments will not and cannot play a role in shaping their synesthesiae. Although it is a very rare occurrence, it is possible that the entire set of a synesthetes colored letters may have been shaped by the influence of patterned childhood objects, such as alphabet blocks or colored refrigerator magnets (Witthoft and Winawer, 2006, 2013; Hancock, 2006). However, if such influence does occur, it is more likely to only shape the first four or five items in the series; that is, for example, 'A, B, C, D, E' or '1, 2, 3, 4.'

It is possible for a synesthete to have colors connected to ideographic characters, such as are used in Chinese and Japanese (kanji). In one case study (Day, 2013), a native of Singapore reported on his synesthetic perceptions of 'ma,' which has many different tonal variations, depending upon the language in question (many dialects of Mandarin have at least five tonal forms of 'ma'). The subject provided colors for three of the five basic 'ma' variants: the character for the word 'ma' with a high tone, meaning 'mother,' 媽, is 'strawberry pink'; the character for 'ma' with a fall-rise tone, meaning 'horse,' 馬, is 'chocolate brown'; and for 'ma' with a falling tone, meaning 'to scold,' 罵, is 'charcoal gray.'

For some 'grapheme to X' synesthetes, punctuation and diacritic marks can also adopt synesthetic qualities; an exclamation point (!) might be red, quotation marks (") yellow, and a minus sign (−) in mathematics blue. For most 'grapheme to visual' synesthetes, however, punctuation marks do not take on color, or are all black. Here, we could also consider graphemes such as a 'smiley face' (☺), the dollar symbol ($), or an arrow (→), which may also take on synesthetic colors.

Ramachandran and Hubbard (2003) write, "[t]he font that evokes the most optimally saturated color is usually the simplest – for example, the clean lines of the typeface Helvetica rather than the ornate Gothic. But we have seen rare examples in which an unusual font was more effective at invoking strong color. We suggest that such fonts might serve as ultranormal stimuli that evoke even higher responses from the neurons involved with graphemes […] than more prototypical fonts."

For some 'grapheme → color' synesthetes, uppercase versus lowercase can make a difference; for example, 'B' may be green, but 'b' brown, or 'E' yellow, and 'e' red. It is common for one grapheme to influence the color of an adjacent grapheme; for example, the purple color of '4' could be so impressive and such a favorite that it will change the color of the preceding '2' in the number '24.'

For some synesthetes, not only do graphemes have an associated color, but they might also have a texture, which might only be perceived by a visual aspect (that is, interpreted via shading, color contrasts, and borders), or via synesthetic touch perception. A synesthete, for example, might not only associate the color sky blue with the letter 'E,' but 'E' will also have a 'sandy' texture; the letter 'M' not only magenta, but also a texture of 'wet dew,' or 'like grass with lots of water on it.'

Studies have been done on ambiguous graphemes which can look like either of two letters (see, for example, **Figure 2**), or as either as letters in words or numbers in mathematical figures (such as a line which could be either a '1' (one) or the letter 'I,' or a squared-off angular figure which could be either the letter 'S' or the number '5'). Dixon et al. (2006) found that the synesthetic color of the letter will indeed change for given synesthetes depending upon which grapheme – the letter or the number – the figure is perceived as. This raises the question of whether it is the form or the meaning of the grapheme which elicits synesthetic colors. Dixon et al.'s conclusion was that, for 'projector' synesthetes, it was meaning. However, Ramachandran and Hubbard argue that it is mainly the shape, and not the meaning, which determines the synesthetic response (however, see Ramachandran and Hubbard, 2001a,b). A key element here may be whether the specific test subject observed during a given test was a projector or associator

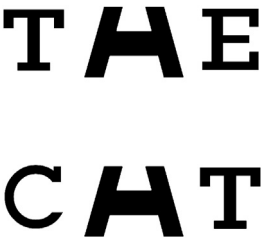

Figure 2 Ambiguity with the grapheme representing either 'H' or 'A.' Copyright by Joy A. Day, 2013.

synesthete. The question of whether it is meaning or shape which determines the color ties directly to the question of which part of the brain is cross-talking with color-processing areas.

Ordinal Linguistic Personification

Ordinal linguistic personification (OLP) (Simner and Holenstein, 2007) is a form of language-based synesthesia in which graphemes, days, and months are associated with aspects of gender and personalities; for example, the letter 'V' might be perceived to be a confident and indifferent girl, and the number '2' a shy and wimpy boy (see **Figure 3**). Although now seen to be a very common type of language-based synesthesia, we still know very little about OLP. It was first reported in 1893 (see Calkins, 1893; Flournoy, 1893), but was dismissed until recently by most researchers (see Cytowic, 2002; also Day, 2005; Simner and Hubbard, 2006; Amin et al., 2011). It is possible for a synesthete to have numbers which have personalities, but no colors.

Lexeme to Flavor

Another type of synesthesia involves speech sounds or whole words evoking synesthetic flavors (see Ferrari, 1907, 1910; Pierce, 1907; Ward and Simner, 2003; Jones et al., 2011). Here, it is not as the word 'coffee' making you *think* of the flavor of coffee; rather, someone actually tasting the flavor, in his or her mouth, in high detail. But, furthermore, the word 'confess' may likewise have the flavor of coffee, or the word 'microscope' evoke the flavor of carrots. For some of these synesthetes, certain specific speech sounds produce corresponding flavors in the mouth. For example, the hard /g/ sound in the words 'argue' and 'begin' may produce the flavor of yoghurt; the combination of the /s/ and /p/ sounds in the words 'super' or 'peace' may result in tasting tomato soup. This may happen whether the synesthete speaks, hears, reads, or just thinks about the word. The names of food items themselves will tend to taste like the food being named; for example, the word 'sausage' will taste like sausage. Then, certain phonemes of that word will be taken, and those sounds will extend into other words. For example, the 'J' – /d_3/ – sound in 'sausage' also appears in the word 'college,' making it also taste like sausage. 'Phoneme to flavor' and 'lexeme to flavor' synesthetes often also report aspects of texture and temperature as part of their synesthetic experiences; for example, the word 'rake' not only tastes like fried bacon, but is crisp, oily, and rather cold (see Ward et al., 2005). These sensations are felt on the tongue and in the mouth. Unlike with 'grapheme to color' synesthesia, in which, for example, the word 'red' might induce a synesthetic perception of green, a phenomenon known as the 'alien color effect' (see Gray et al., 2006), 'lexeme to flavor'

Figure 3 An alphabet showing an example of ordinal linguistic personification (OLP). Copyright by Joy A. Day, 2013.

synesthesia displays no such effects for the names of foods, such as the word 'bacon' tasting of spinach (Ward et al., 2005).

Ticker Tape

A very rare variant of projected language-based synesthesia is known as 'ticker tape,' in which words, whether spoken, heard, or thought about, are transformed into written text, seen in front of the synesthete. In different case reports, ticker tape synesthesia has appeared as scrolling tape issuing from the mouths of speakers, as cartoon-like balloons, and as subtitles placed slightly above or below the synesthetes line of vision, very similar to subtitled text in movies or on television. Synesthetic ticker tape text is not necessarily always colored.

Other Types of Linguistic Synesthesia

Lexeme to Visual

Lexeme to color synesthesia appears to be significantly rarer than phoneme to color synesthesia. However, case reports do exist, such as that by Weiss et al. (2001), who write about a woman who synesthetically connects colors to the names of only famous people.

Phoneme to Visual

The realm of phoneme or morpheme to visual synesthesia also includes signed symbols in various sign languages. There have been accounts of synesthetes perceiving colors and shapes for signs in American Sign Language, British Sign Language, and Japanese Sign Language (Day, 2013).

Others

Although extremely rare, 'color to grapheme' synesthesia does exist, and can exist without the synesthete having 'grapheme to color.' Other extremely rare types of language-induced synesthesia include 'grapheme to flavor,' 'grapheme to touch,' 'lexeme to odor,' 'lexeme to touch,' and 'phoneme to touch.' Regarding the last, an example given from such a synesthete: "[F]or example, tomato. If it's pronounced 'to-may-to', it's rough like gravel. If it's pronounced like 'to-mah-to', it's softer, like a warm cotton shirt that just came out of the dryer."

For some multiple synesthetes, 'colored grapheme' synesthesia can interact and interfere with other closely related types of synesthesia. Collins mentions her subject *S*, who apparently had colored graphemes, colored phonemes, and perhaps colored names and/or colored geometric shapes. For *S*, with regard to the consonant 'B,' the "[l]ook of *B* is blue (because of liquid roundness of loops making the letter). Sound of *B* is white (due to vowel *e*, which is white). Idea of *B* is blue (is associated with name Betty and a childhood experience)" (Collins, 1929, p. 15).

What Causes Synesthesia?

There are two main overarching theories of synesthesia causation. Both incorporate the idea of there being dedicated areas of the brain which specialize in one or more specific functions. For example, the visual cortex can be further subdivided so as to note those areas that deal specifically with color processing (called the fourth visual areas, or V4 area), or those areas that deal with motion processing (V5).

Many different proposals have been offered toward grapheme to visual synesthesia causation. One theory suggests that cross-activation between grapheme areas and color areas in the fusiform gyrus plays a major role (see, e.g., Ramachandran and Hubbard, 2001a,b; Hubbard et al., 2005). Another theory offers that this form of synesthesia results during processing of graphemic meaning in the anterior fusiform gyrus, when abnormal feedback travels back to posterior inferior temporal regions and the V4 color area (Smilek et al., 2001b). This latter model has been extended by others to consider disinhibited feedback "from a multisensory nexus such as the superior temporal sulcus, posterior parietal lobe, intraparietal cortex or temporo-parietal-occipital junction" (Rouw and Scholte, 2007, p. 792; see also Grossenbacher and Lovelace, 2001; Esterman et al., 2006; Weiss et al., 2005). These two models of grapheme to color synesthesia are not mutually exclusive; both processes could be involved, and could even operate simultaneously, including in a given individual.

It should be pointed out that synesthesia can also result from certain types of brain injury (see, e.g., Jacobs et al., 1981; Vike et al., 1984). This includes cases of synesthesia resulting as part of epileptic seizures (see, e.g., Ulrich, 1903). However, all reported cases of synesthesia induced through injury or degeneration are of the sensorial synesthesia form; there are no such cases resulting in, for example, colored letters, numbers, or days of the week.

Research by Asher et al. (2009) suggests that the following chromosome regions might be involved in specifically 'auditory → visual' synesthesiae: 2q24, 5q33, 6p12, and 12p12. Also, Asher et al. note that the X chromosome apparently is not involved. "The region on chromosome 6 […] has been strongly linked to dyslexia and specifically to difficulties in phonological decoding and orthographic processing, which is of particular interest given the importance of linguistic stimuli in synesthesia. Two genes in this region, KIAA0319 (MIM 609269) and DCDC2 (MIM 605755), have been proposed as candidate genes for dyslexia; both have been shown to play roles in neuronal migration, which has important implications for synesthesia given the evidence for altered neural architecture in the disorder" (Asher et al., 2009, pp. 282–283).

Meanwhile, Brang and Ramachandran (2008) have focused on gene HTR2A, on chromosome 13q. This gene is involved in serotonin reception/inhibition. Brang and Ramachandran are looking specifically at activation of S2a receptors as producing synesthesia, and (hyper)activation of S1 receptors as inhibiting S2a receptors and thus inhibiting synesthesia.

Ramachandran and Hubbard (2001b) have suggested that 'colored number' synesthesia for associator-type could result from abnormal cross-activation between the angular gyrus, which is where concepts of numbers are processed, and the superior temporal gyrus, which processes concepts of color. On the other hand, for projector-type synesthetes, colored numbers might result from hyperconnectivity between the fusiform gyrus, which processes forms, and the V4 areas of the fusiform which processes colors.

There is indication (see Dovern et al., 2012) that those with 'grapheme to color' synesthesia have a higher level of neuronal intranetwork connectivity, and higher rates (about three times that of nonsynesthetes) of both global and specific (medial and lateral visual networks to a right frontoparietal network) intrinsic network connectivity. Diffusion tensor imaging study reveals regionally increased structural connectivity in inferior temporal, parietal, and frontal cortices of grapheme to color synesthetes (Rouw and Scholte, 2007). Research indicates "that in grapheme–color synesthetes, color area V4 showed a significant FC [functional connectivity] with other synesthesia-relevant regions, namely primary and secondary visual areas, auditory cortex bilateral, and right parietal cortex, while in control subjects, V4 was also functionally connected to primary and secondary visual areas but showed no FC with the auditory cortices and the right parietal cortex" (Dovern et al., 2012, p. 7617).

Using diffusion tensor imaging (DTI), Rouw and Scholte (2007) have produced data indicating that hyperconnectivity of white matter structures is involved for at least some grapheme to visual synesthetes. Both 'projector' and 'associator' synesthetes showed greater connectivity in the left superior parietal or frontal cortex than the average for nonsynesthetes. Furthermore, projector synesthetes (but not associators) showed a significantly strong amount of connectivity in the right inferior temporal cortex; thus, projectors could be differentiated from associators based upon the DTI readings. Note that this increase in white matter ties in with theories proposing a genetic predisposition to focused decreased pruning of connections (see, e.g., Maurer, 1997). "Of particular interest is the cluster of greater connectivity found in the inferior temporal cortex near the fusiform gyrus. The fusiform gyrus is involved in the perception and categorization of visual stimuli. Furthermore, an area specialized in the perception of the (form of) graphemes is located adjacent to an area specialized in the perception of colors. This has spurred the idea that grapheme-color synesthesia is caused by cross-connections between adjacently located grapheme and color areas in the fusiform gyrus" (Rouw and Scholte, 2007, p. 795).

Repetitive transcranial magnetic stimulation to the right posterior parietal lobe has been seen to attenuate 'grapheme → color' synesthesia for a short period of time for some (Esterman et al., 2006).

However, brain scan research (see, e.g., Hupé et al., 2011; Hänggi et al., 2011) indicates that the neural bases for grapheme to visual synesthesia are apparently not localized to any specific visual processors, such as the V4 area in the occipital lobe; rather, grapheme to visual is a global neurological process, involving hyperconnectivity and interactions throughout many areas of the brain, some relatively quite distant from each other. Grapheme–color synesthetes might not just have hyperconnectivity between the fusiform gyrus and the intraparietal sulcus but throughout the entire brain, such that synesthesia is only one of many byproducts of these hyperconnections.

See also: Color Classification and Symbolism; Color Terms, Linguistics of; Functional Brain Imaging of Language processes; Phonology; Reading Nonalphabetic Scripts; Sign Language.

Bibliography

Amin, Maina, Olu-Lafe, Olufemi, Claessen, Loes E., et al., 2011. Understanding grapheme personification: a social synaesthesia? Journal of Neuropsychology 5, 255–282.

Asher, Julian E., Lamb, Janine A., Brocklebank, Denise, et al., 2009. A whole-genome scan and fine-mapping linkage study of auditory-visual synesthesia reveals evidence of linkage to chromosomes 2q24, 5q33, 6p12, and 12p12. American Journal of Human Genetics 84, 279–285.

Baron-Cohen, Simon, Wyke, Maria A., Binnie, Collin, 1987. Hearing words and seeing colours: an experimental investigation of a case of synaesthesia. Perception 16, 761–767.

Brang, David, Ramachandran, Vilayanur S., 2008. Psychopharmacology of synesthesia; the role of serotonin S2a receptor activation. Medical Hypotheses 70 (4), 903–904.

Calkins, Mary W., 1893. A statistical study of pseudo-chromesthesia and mental forms. American Journal of Psychology 5, 439–464.

Collins, Mary, 1929. A case of synaesthesia. Journal of General Psychology 2, 12–27.

Cytowic, Richard E., 1993. The Man Who Tasted Shapes. Putnam, New York.

Cytowic, Richard E., 1989. Synesthesia: A Union of the Senses. Springer-Verlag, New York.

Cytowic, Richard E., 2002. Synesthesia: A Union of the Senses, second ed. MIT Press, Cambridge, MA.

Cytowic, Richard E., Eagleman, David M., 2009. Wednesday Is Indigo blue. MIT Press, Cambridge, MA.

Day, Sean A., 2005. Some demographic and socio-cultural aspects of synesthesia. In: Robertson, Lynn, Sagiv, Noam (Eds.), Synesthesia: Perspectives from Cognitive Neuroscience. Oxford University Press, Oxford, pp. 11–33.

Day, Sean A., 2013. Synesthesia (Online). Available at: http://www.daysyn.com/ (accessed 17.03.13.).

Dixon, Mike J., Smilek, Daniel, Duffy, Patricia L., Zanna, Mark P., Merikle, Philip, 2006. The role of meaning in grapheme-colour synaesthesia. Cortex 42, 243–252.

Dixon, Mike J., Smilek, Daniel, Merikle, Philip, 2004. Not all synaesthetes are created equal: projector versus associator synaesthetes. Cognitive, Affective and Behavioral Neuroscience 4 (3), 335–343.

Dovern, Anna, Fink, Gereon R., Fromme, A. Christina B., Wohlschläger, Riedl, Valentin, 2012. Intrinsic network connectivity reflects consistency of synesthetic experiences. Journal of Neuroscience 32 (22), 7614–7621.

Eagleman, David M., Kagana, Arielle D., Nelson, Stephanie S., Sagaram, Deepak, Sarma, Anand K., 2007. A standardized test battery for the study of synesthesia. Journal of Neuroscience Methods 159, 139–145.

Esterman, Michael, Verstynen, Timothy, Ivry, Richard B., Robertson, Lynn C., 2006. Coming unbound: disrupting automatic integration of synesthetic color and graphemes by transcranial magnetic stimulation of the right parietal lobe. Journal of Cognitive Neuroscience 18 (9), 1570–1576.

Ferrari, G.C., 1907. Una varieta nuova di sinestesia. Rivista di Psicologia 3, 297–317.

Ferrari, G.C., 1910. Un nuovo caso di sinestesia uditivo-gustativa. Rivista di Psicologia Applicata 6, 101–104.

Flournoy, T., 1893. Des phenomenes de synopsie. Felix Alcan, Paris.

Gray, Jeffrey A., Parslow, David M., Brammer, Michael J., et al., 2006. Evidence against functionalism from neuroimaging of the alien colour effect in synaesthesia. Cortex 42, 309–318.

Grossenbacher, Peter G., Lovelace, Christopher T., 2001. Mechanisms of synesthesia: cognitive and physiological constraints. Trends in Cognitive Science 5 (1), 36–41.

Hancock, Peter, 2006. Monozygotic twins' colour-number association: a case study. Cortex 42, 147–150.

Hänggi, Jürgen, Wotruba, Diana, Jäncke, Lutz, 2011. Globally altered structural brain network topology in grapheme-color synesthesia. Journal of Neuroscience 31 (15), 5816–5828.

Hubbard, Edward M., Piazza, Manuela, Pinel, Philippe, Dehaene, Stanislas, 2005. Interactions between number and space in parietal cortex. Nature Reviews Neuroscience 6, 435–448.

Hupé, Jean-Michel, Bordier, Cécile, Dojat, Michel, 2011. The neural bases of grapheme-color synesthesia are not localized in real color-sensitive areas. Cerebral Cortex.

Jacobs, Lawrence, Karpik, Alice, Bozian, Diana, Gøthgen, Svend, 1981. Auditory-visual synesthesia: sound-induced photisms. Archive of Neurology 38, 211–216.

Jones, C., Gray, M.A., Minati, L., et al., 2011. The neural basis of illusory gustatory sensations: two rare cases of lexical-gustatory synaesthesia. Journal of Neuropsychology 5, 243–254.

Maurer, Daphne, 1997. Neonatal synaesthesia: implications for the processing of speech and faces. In: Baron-Cohen, Simon, Harrison, John (Eds.), Synaesthesia: Classic and Contemporary Readings. Blackwell, Oxford, pp. 224–242.

von Melchner, Laurie, Pallas, Sarah L., Sur, Mriganka, 2000. Visual behaviour mediated by retinal projections directed to the auditory pathway. Nature 404, 871–876.

Pierce, Arthur H., 1907. Gustatory audition: a hitherto undescribed variety of synaesthesia. American Journal of Psychology 18, 341–352.
Ramachandran, Vilayanur S., Hubbard, Edward M., 2001a. Psychophysical investigations into the neural basis of synaesthesia. Proceedings of the Royal Society of London, B 268, 979–983.
Ramachandran, Vilayanur S., Hubbard, Edward M., 2001b. Synaesthesia – a window into perception, thought and language. Journal of Consciousness Studies 8 (12), 3–34.
Ramachandran, Vilayanur S., Hubbard, Edward M., 2003. More Common Questions about Synesthesia (Online). Available at: http://www.scientificamerican.com/article.cfm?id=more-common-questions-abo-2003-04-14 (accessed 22.09.12.).
Rich, Anina N., Bradshaw, J.L., Mattingley, Jason B., 2005. A systematic, large-scale study of synaesthesia: implications for the role of early experience in lexical-colour associations. Cognition 98, 53–84.
Rouw, Romke, Scholte, H. Steven, 2007. Increased structural connectivity in grapheme-color synesthesia. Nature Neuroscience 10 (6), 792–797.
Rouw, Romke, Scholte, H. Steven, May 5, 2010. Neural basis of individual differences in synesthetic experiences. Journal of Neuroscience 30 (18), 6205–6213.
Simner, Julia, Holenstein, Emma, 2007. Ordinal linguistic personification as a variant of synesthesia. Journal of Cognitive Neuroscience 19 (4), 694–703.
Simner, Julia, Hubbard, Edward M., 2006. Variants of synesthesia interact in cognitive tasks: evidence for implicit associations and late connectivity in cross-talk theories. Neuroscience 132, 805–814.
Simner, Julia, Mulvenna, Catherine, Sagiv, Noam, et al., 2006. Synaesthesia: the prevalence of atypical cross-modal experiences. Perception 35, 1024–1033.
Smilek, Daniel, Dixon, Mike J., Cudahy, Cera, Merikle, Philip M., 2001b. Synaesthetic photisms influence visual perception. Journal of Cognitive Neuroscience 13 (7), 930–936.
Smilek, Daniel, Dixon, Mike J., Merikle, Philip M., 2005. Synaesthesia: discordant male monozygotic twins. Neurocase 11, 363–370.
Smilek, Daniel, Moffatt, B.A., Pasternak, J., et al., 2001a. Synaesthesia: a case study of discordant monozygotic twins. Neurocase 8, 338–342.
Ulrich, Alfred, 1903. Phénomènes de Synesthèsies chez un Épileptique. Revue Philosophique 56, 181–187.
Vike, Jean, Jabbari, Bahman, Maitland, Charles G., 1984. Auditory-visual synesthesia: report of a case with intact visual pathways. Archive of Neurology 41, 680–681.
Ward, Jamie, Simner, Julia, 2003. Lexical-gustatory synaesthesia: linguistic and conceptual factors. Cognition 89, 237–261.
Ward, Jamie, Simner, Julia, Auyeung, Vivian, 2005. A comparison of lexical-gustatory and grapheme-colour synaesthesia. Cognitive Neuropsychology 22 (1), 28–41.
Weiss, Peter H., Shah, N. Jon, Toni, Ivan, Zilles, Karl, Fink, Gereon R., 2001. Associating colours with people: a case of chromatic-lexical synaesthesia. Cortex 37, 750–753.
Weiss, Peter H., Zilles, Karl, Fink, Gereon R., 2005. When visual perception causes feeling: enhanced crossmodal processing in grapheme-color synesthesia. NeuroImage 28, 859–868.
Witthoft, Nathan, Winawer, Jonathan, 2006. Synesthetic colors determined by having colored refrigerator magnets in childhood. Cortex 42, 175–183.
Witthoft, Nathan, Winawer, Jonathan, January 10, 2013. Learning, Memory, and Synesthesia. Psychological Science. http://dx.doi.org/10.1177/0956797612452573.

Relevant Websites

http://www.synesthesia.info/index.html – The American Synesthesia Association.
http://synesthete.org/ – The Synesthesia Battery (Tests for Synesthesia).
http://www.uksynaesthesia.com/ – The UK Synaesthesia Association.

Linguistic Turn

C Fynsk, Binghamton University, Binghamton, NY, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8913–8916, © 2001, Elsevier Ltd.

Abstract

The 'linguistic turn' is presented in this article from the basis of modern philosophies of language (particularly the contributions of Ludwig Wittgenstein, Martin Heidegger, and Walter Benjamin) and in relation to contemporary adaptations of these philosophical critiques of representation in post-structuralism. The article argues that the linguistic turn occurs when language is taken not as an instrument in social practice but rather as its ground. However, it takes its distance from most 'constructivist' accounts ('social' or 'linguistic') in asserting that such accounts tend to avoid a fundamental engagement with language like the one that characterizes Michel Foucault's study, *The Archaeology of Knowledge*. The argument is made that only a fundamental analysis of this kind can approach the grounds of the social relation.

'The linguistic turn' generally designates in the social sciences a movement away from positivist or historicist accounts of social and historical phenomena toward analyses of the structures of representation that are understood to have helped shape those phenomena. The phrase is frequently (and sometimes loosely) used to characterize contemporary forms of the critique of ideology and the critical 'demystification' or 'demythologization' of historical and socio-political representations that offer themselves as transparent or natural accounts. Since such critical work often proceeds by way of attention to the rhetorical and conceptual articulations of representations that tend to efface their very character as representations, and since it frequently draws from philosophical and even literary theories of language and representation, it is understood to participate in a turn from socio-historical 'realities' to their linguistic presentation.

The phrase is used so broadly, however, that its strong meaning for theoretical methodology is frequently lost. Frequently, the critique of representational structures proceeds without a consequent attention to the question of language, and even some of the strongest versions of 'linguistic constructivism' fail to interrogate their own presuppositions. Thus, the linguistic turn will be defined here fairly strictly from the basis of its foundations in order to recover its meaning and implications for the social sciences. It will be defined as beginning when language is no longer understood as an instrument in social practice, but rather as the site or 'ground' for such practice. It will be understood to have been fully undertaken when such a fundamental understanding of language shapes the procedures of the sciences themselves.

The beginnings of the linguistic turn may easily be traced to the eighteenth century and seen to be lurking in the very rich debate concerning the origins of language. It is clearly occurring in the research of Wilhelm von Humboldt and in the writings of the German Romantics, and then actively recognized – some would say, practiced – in the later works of Friedrich Nietzsche. Thus, when the official occurrence of a linguistic turn is registered in Vienna for philosophy of language, a more 'fundamental' turning is already long underway in the tradition of German philosophy that Walter Benjamin and Martin Heidegger brought to the fore in their remarkably convergent speculative writings on language. Indeed, if Ludwig Wittgenstein himself was ready to grasp in 1929 the import of Heidegger's still largely latent meditation on language and the significance of Heidegger's own linguistic practice, it is because the philosophical ground was well prepared for Heidegger's subsequent turn (his famous *Kehre)* from an existential philosophy still marked by the metaphysics of the human subject to a thought of what is given of Being in and by language.

Historians of the social sciences will provide ample testimony of methodological tremors linked to the developments noted here in philosophy of language – in short, a displacement of the human subject or human consciousness as the grounding source of meaning in signifying practice, and the requirement of rethinking the subject/object relation and relation itself. However, it may well be argued that the strong turn realized by authors such as Benjamin and Heidegger is not assumed in the social sciences in Europe until the period of the structuralist movement (a movement that explicitly unseated human consciousness from its founding role in favor of broader structures of determination: kinship systems, linguistic structures, and the laws of sexuality or desire). Even here, however, the turn to language remained frequently at no more than the level of analogy, and systems of myth or kinship were shown to be formalized equally effectively with recourse to mathematical logic (another form of appeal to a *logos*, some would argue). Thus, while structural linguistics functioned as a kind of 'point science' for the initiatives throughout the human sciences in Europe for over a decade, only the endeavors of 'post-structuralist' authors such as Michel Foucault, Jacques Lacan, Jacques Derrida, or Gilles Deleuze actually carried the notion of the linguistic constitution of being to its full import for the social sciences. Only from the time of post-structuralism did it become possible to entertain thinking the social relation from language.

The linguistic turn has thus occurred, for example, when Michel Foucault, in his methodological treatise *The Archaeology of Knowledge* (1969), announces that the fundamental element for his study of discursive formations will be the *énoncé,* the statement. In an effort to displace the governing categories of the history of thought (totalizing categories that establish continuity and that all lead back, in Foucault's view, to the sovereignty of the human consciousness), Foucault attempts a theoretical elaboration of a discursive 'object' that is always relational and always to be thought from its material

emergence. 'Object' must be left here in quotation marks because the statement, as Foucault understands it, is the 'modality' or 'function' of existence of discursive facts, what 'gives' them or enables them to occur as language and thus as (1) a relation to a domain of objects; (2) a set of possible positions for an enunciative subject; (3) an element in a coexisting discursive field; and (4) a repeatable materiality. A statement, in Foucault's definition, cannot be understood as a sentence that lends itself to hermeneutic investigation or logical analysis, but must rather be seen as the material, always relational, condition for something like a sentence or a proposition. Historical analyses that seek the 'signified' of any group of signs (reading any document for the meaning it bears, any text for its 'allegorical' sense), like linguistic studies that seek rules of formation, invariably efface the irruptive, eventful character of the fact of language, returning all linguistic events to their organizing conditions. Thus they participate in what Foucault calls, in his famous *Discourse on Language*, the 'ancient elision of the reality of discourse'. Foucault's effort in thinking the *énoncé* as the material function of existence of any discursive act or organization is directed at honoring the 'anarchic' character of language – a potential for disruption which 'disciplinary' and institutional formations seek to channel and control. At the heart of his project is an assumption that the orders of knowledge and the social institutions in which they are embedded are organized by a fear of the undermining, or even violent dimensions of discourse for the orders in which it emerges (undermining for the mere fact that it can reveal the inherent lack of foundations for the social order, violent for the fact that it can carry the irruption of desire):

> There is undoubtedly in our society, and I would not be surprised to see it in others, though taking different forms and modes, a profound logophobia, a sort of dumb fear of these events, of this mass of spoken things, of everything that could possibly be violent, discontinuous, querulous, disordered even and perilous in it, of the incessant, disorderly buzzing of discourse.

Of course, Foucault's own study, fiercely abstract in its unfolding design, is not without its own violence. Indeed, his *Archaeology* sometimes reads as a veritable storm of theoretical elaboration; its winds tear unrelentingly at the foundations of the history of thought. But a glimpse at the virulence of reaction on the part of proponents of rational norms of 'communicative transparency' and/or pragmatic 'consensus' provide a clue to the explosive potential of Foucault's effort to honor the disruptive dimensions of language.

Numerous authors have pursued Foucault's 'archeo-logical' analyses of discursive formations; in fact, significant currents in the growing field of cultural and 'post-colonial' studies principally draw upon Foucault for their understanding of the discursive constitution of social reality. Theories of social construction are deeply indebted to his analyses of 'disciplinary' knowledge and mechanisms of control. Far fewer authors have carried forward his understanding of the subversive dimensions of the enunciative act, though one sees its impact in the fields of gender studies and sexuality or in a topic such as the study of 'the practices of the everyday' by Michel de Certeau. It is for this reason that one might regard the 'linguistic turn' – which many would undoubtedly anchor in the widespread acceptance of the theory of linguistic or social 'constructionism' – as incomplete and even faltering. Theories of social construction, and even theories that stress the necessary 'performative iteration' of social norms, almost inevitably turn short of the engagement with language that marks Foucault's work (or that of the other 'post-structuralist' authors to which I referred). They fail, in particular, to attend to the materiality of discursive usage in that dimension that involves what I might term an enunciative experience with language. This avoidance is easily attributable to Foucault's own wariness regarding the 'ontologizing' and 'transcendentalizing' sources upon which he silently draws (principally Heidegger), not to speak of his strictures against anthropomorphism and humanism. But Foucault's own turn to ethical questions in his late work point firmly to the necessity of an exploration of the subject's engagement with the fact of language at the fundamental level evoked in the *Archaeology* – with the fact and manner of the way the *énoncé* is given in its material singularity. What dimension of experience is engaged with this 'gift,' and what engagements (to evoke a motif pursued throughout Jacques Derrida's late work) does it presuppose if the relations offered by the *énoncé* must be thought from the way they are brought into existence in singular instances of usage?

Heidegger had attempted in his later meditations on language to name with precisely the term 'usage' the human relation to the fact of language. To avoid turning language into a kind of transcendental horizon (a trap into which many 'linguistic constructionists' inevitably fall), and to hold to a thought of its finitude, Heidegger attempted to think through the necessity of a human 'accession' to language – i.e., the conditions for any 'interpellation' of the human subject by language, and any human response. He did not avoid a certain anthropocentrism in this undertaking, but he also pointed beyond the purview of the metaphysics of the human subject. Unfortunately, few authors have followed through Heidegger's thinking on this crucial point, though among them we may count the French psychoanalyst Jacques Lacan (whose thought of the category of the 'real' presupposes this understanding of the 'human relation to the signifier') and the French writer Maurice Blanchot, whose enormous influence on 'post-structuralist' thought still remains to be measured. But their work, taken together with that of the distinguished Jewish philosopher Emmanuel Levinas and the more recent work of authors such as Jean-Francois Lyotard and Jean-Luc Nancy, make it possible to pursue a thought of the manner in which the social relation opens in and with language. From a notion of the human engagement with language, they elaborate an understanding of the event by which human speech communicates a form of 'finite transcendence': a transcendence of determined orders of meaning that is first of all an exposure to the other human being. They thus make it possible, in effect, to return to a thought of the social relation like that approached by Hannah Arendt in *The Human Condition* (1998) in her effort to rethink the notion of public space and the political act itself.

This latter thought of 'communication' (not a transparent communication of signified meaning, but rather, as Walter Benjamin argued so powerfully, a communication of the very possibility of relation) clearly carries with it the obligation of

defining the conditions for its engagement in critical or theoretical practice. For it is clear that the pertinence of a traditionally 'theoretical' mode of inquiry must come into question once language is engaged at levels that exceed that of its representational or instrumental function. Once a 'saying' that bespeaks the irreducible presence of an enunciating subject becomes the object of a meditation on the grounds and possibility of community (but this can be no 'object' for description), theory can no longer claim an exclusive purchase on the question of the social relation. Or it must rethink its own discursive modality and its relation to others (to the language of literature, for example). The challenge of the linguistic turn for the social sciences thus bears upon the discursive practices of the sciences themselves. It is not a challenge that requires the social sciences to abjure their commitment to clarity and rigor in analysis, but it does require them to entertain the question of their own discursive usage and the possibility of translation between discursive modes or between their own usage and that of a singular event of language. Here, it becomes clear that a fundamental reflection on linguistic usage is the very condition for a consequent engagement of the social relation – a relation that is not reducible to its linguistic determinations, but also not thinkable apart from the site of its emergence and articulation. The 'linguistic turn' is not a methodological 'given,' a path already secured; it is a required turn to the grounds of meaning that must be repeatedly undertaken if social analysis is to achieve concreteness in its effort to regain the social itself.

See also: Foucault, Michel (1926–84); Linguistic Turn and Discourse Analysis in History; Logical Positivism and Logical Empiricism.

Bibliography

Arendt, H., 1998. The Human Condition. University of Chicago Press, Chicago.
Butler, J., 1993. Bodies that Matter. Routledge, New York.
Benjamin, W., 1972. In: Tiedemann, R., Schweppenhäuser, H. (Eds.), Gessamelte Schriften. Suhrkamp, Frankfurt am Main, Germany (Demetz, P. (Ed.), 1986. Reflections. Schocken Books, New York).
De Certeau, M., 1984. The Practice of Everyday Life (S.F. Rendall, Trans.). University of California Press, Berkeley, CA.
Foucault, M., 1969. L'Archéologie du savoir. Editions Gallimard, Paris (1972. The Archaeology of Knowledge and The Discourse on Language. Pantheon Books, New York).
Fynsk, C., 1996. Language and Relation…That There is Language. Stanford University Press, Stanford, CA.
Heidegger, M., 1985. In: von Hermann, F. (Ed.), Unterwegs zur Sprache, Gesamtausgabe, vol. 12. Klostermann, Frankfurt am Main, Germany (P.D. Hertz, Trans., 1971. On the Way to Language. Harper and Row, New York).
Lacan, J., 1986. Le Séminaire, L'Éthique dans la psychanalyse. Seuil, Paris.
Lyotard, J.-F., 1983. Le Diffrend. Editions de Minuit, Paris.
Rorty, R. (Ed.), 1992. The Linguistic Turn: Essays in Philosophical Method. University of Chicago Press, Chicago.
Wittgenstein, L., 1965. Lecture on ethics. Philosophical Review 74, 8–15.

Linguistic Turn and Discourse Analysis in History

John E Toews, University of Washington, Seattle, WA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8916–8922, © 2001, Elsevier Ltd.

Abstract

Discourse analysis is the preferred method of historians who have taken the turn toward a view of language as the dense and autonomous site in which the objective reality of the past and the subjective identity of the historian is constructed. 'Discourse' refers to the framing assumptions of thought, speech, and action in particular epochs and cultures and/or to the communicative processes that characterize both development within and interaction among semiotic or cultural systems. Historians affirm the plurality and heterogeneity of discourses as a condition of historical change. They are divided about the relation between discourse and its 'others,' particularly the 'other' of material practices, and about the ways in which the production of individual and collective identities within discourse affects the status of historical agency.

The Linguistic Turn and the Post-Modern Epistemological Crisis

The term 'linguistic turn' emerged as common currency in discussions among historians during the 1980s and 1990s to designate a sweeping, but vehemently contested, shift in the framework within which both the object or content of historical investigation (the 'past') and the methods of achieving historical knowledge of this content were defined. In its weaker, least contentious, form the linguistic turn referred to a more intense and focused attentiveness to the opacity and complexity of the linguistic forms through which all evidence of the past was mediated, and the 'truth' of the past represented, reconstructed, and communicated. The term itself was appropriated from discussions among philosophers, where it had been used as a general description of the turn in twentieth-century philosophy toward reformulating the epistemological problem of the relationship between thought and reality in terms of a relationship between language and reality. In the stronger and more contentious form that surfaced in debates among historians, however, the linguistic turn came to designate a shift away from the conventional view that language was a passive, potentially transparent instrument for transmitting content, for organizing, representing or expressing the truth of a reality outside of itself, and toward a view of language as the dense, fertile and autonomous site in and through which the objective reality of the world as well as the subjective reality of language users was actively produced or constructed. From this latter perspective, the act of taking the linguistic turn was synonymous with the claim that the production and communication of historical knowledge was inextricably entangled in the epistemological crisis associated with the transition from a modern to a 'postmodern' cultural condition.

As a particular disciplinary expression of this epochal transition to postmodernism the linguistic turn in history was grounded on three major premises. First, it accepted the ultimate futility of the historical quest to reconstruct a single, coherent narrative of human history whose pattern or structure could somehow be grasped and mapped in its totality from a transcendent position outside of the particular, contingent meanings that human beings in specific times and places had imposed on the material traces of the past, a quest most recently articulated in the national liberation narratives, global modernization theories, and Marxian historical dialectics of the Cold War decades. Second, the linguistic turn appropriated from poststructuralist theory the sweeping critique of residual beliefs in the existence of noncontingent stable identities (whether natural kinds or metaphysical essences) whose fates might be traced historically through paths of accident, evolution, repression, resistance, and liberation. National, ethnic, racial, sexual, gender, and class identities were not treated as material or metaphysical realities expressed in language and culture, but contingent, constantly changing, constructions produced in language and culture. Third, this dissolution of essential identities implied a transformation in the methods of historical representation. Access to a reality not already shaped and constructed by the linguistic medium in which it was always presented to the knower, either as subjectively lived experience or objectively perceived external object or event, was considered impossible. The historian could only interpret interpretations, could only redescribe and thus reconstruct in his or her own language the linguistic constructions of others.

Various elements of the epistemological crisis associated with the linguistic turn had been present in muted and sometimes more extreme forms within the Western historical tradition since the beginnings of the modern professionalization of history in the early nineteenth century. During the post-World War II era the disconcerting relativist implications of self-consciousness about the density and relative autonomy of the linguistic medium that mediated reality as meaningful was reined in by a broadly shared disciplinary consensus concerning the reality of the past and the methods that provided access to its truth. The more radical linguistic turn of the 1980s, with its implications that the historian's past was a construction grounded on ethical and political choices, rather than on epistemological criteria, was associated with the breakdown of this consensus, as the increasingly effective voices of previously excluded perspectives, particularly those of women, ethnic minorities, and former colonial subjects, mounted a powerful critique of established professional conventions. The linguistic turn was embedded deeply in the feminist, multiculturalist, and postcolonial 'turns' among dissident members of the disciplinary community, and connected to a new series of

interdisciplinary liaisons with literary theory, cultural anthropology, and philosophy.

A major difficulty standing in the way of providing a concise description of the linguistic turn in history stems from the diverse meanings assigned to the term 'language.' The everyday conventional meaning of language as a native or vernacular language is rarely what contemporary historians have in mind. In fact, the use of the term has been extended broadly beyond the rules and practices of verbal communication in speech and writing to include nonlinguistic structures of exchange and all signifying practices through which meaning is produced, transmitted, and communicated. As the totality of meaningful action or signifying practices, 'language' becomes virtually indistinguishable from 'culture.' The favored model for organizing discussions of language by proponents of the linguistic turn during the 1980s and 1990s was the semiotic model derived from the linguistic theories of Ferdinand de Saussure and Charles Sanders Pierce, in which language is conceived as a structured system of 'signs' whose meanings are determined by their relations to each other rather than their relation to a 'transcendental' object or subject. Most influential for historians has been the development of this semiotic model in cultural anthropology and especially the French structuralist and poststructuralist theory of Louis Althusser, Roland Barthes, Jacques Derrida, Jacques Lacan, and Michel Foucault. Although 'language' and 'discourse' are still often used interchangeably in debates concerning the linguistic turn, the attempt to provide some kind of theoretical focus on the analysis of language in history has increasingly been organized around the term 'discourse.' Historians who have taken the linguistic turn tend to define the distinctiveness of their own signifying practice as 'discourse analysis.'

Discourse Analysis: Enabling Structures and Communicative Practices

Despite the widespread use of 'discourse' as an organizing concept within a variety of the humanistic and social scientific disciplines, its meanings have remained fluid, flexible, and often indefinite. However, the two general meanings of the term in common usage – (1) a formal speech or treatise that presents a theme or object in a systematic and authoritative manner, and (2) a conversation or discussion involving a communicative exchange among a number of speakers – provide a framework for the more technical, disciplinary uses within current historical work.

Discourse as the Historical a priori of Thought, Speech, and Action

The more formal, systematic, and authoritative dimension of discourse is articulated in historical studies in the attempt to outline the formal structures of rules, codes, categories, and procedures (the relatively stable aggregate of conventions, usages, idioms, tropes, vocabularies) that define a particular communicative space, delimiting the field of possible statements and actions, enabling what can be meaningfully and effectively articulated. Historical discourses in this sense are impersonal formal structures operating anonymously behind the backs of speakers and actors, constituting the world of objective reality and subjective agency by making certain types of speech or behavior possible and other types impossible, by recognizing certain speakers and actors as legitimate or effective, and excluding others as illegitimate or ineffective, by defining certain objects and events as real or objective, and others as fictional and subjective. Different discourses constitute the world of experience in different ways, and are thus incommensurable with each other. They produce homogeneity (within their field of power) and heterogeneity in relation to other discursive fields. Specific statements or actions are only meaningful as instances of the rules and regularities governing the discourse in which they occur. Discourses become markers of the stable, synchronic, self-contained, incommensurable, discontinuous dimension in historical systems of meaning or cultural formations. For the historian, mapping the structures or codes of particular discourses defines historical action and historical actors as determined by those structures, as events and agents in a foreign world that has its own criteria for determining what is true or real, who is an effective agent, whose speech is meaningful. Discourses may thus, in Foucault's terms, create specific historical 'regimes' of truth and systems of power.

Discourse analysis can be a way for the historian to establish the otherness of the past, as it has been by practitioners of the new cultural history who imagine the past as a 'foreign country.' Textual traces from the past can be provided with their own reality, differentiated from the realities of other times and places, by inserting them into the discourses that originally defined their significance. By recognizing the strangeness, the alien character of the 'voices' or 'traces' from the past deposited in the archival record, the present also comes into focus as a particular historical moment with its own sets of systematic enabling and limiting conditions. In this sense the deployment of an ahistorical, formal conception of discourse can produce a very concrete historical consciousness, an awareness of the contingent historicity of present existence. It provides few guidelines for reconstructing the developmental processes that lead from the past to the present.

Discourse as Communicative Process

As a constitutive system of categories and codes that differentiates historical worlds or cultures from each other, discourse becomes a tool for the reconstruction of past as different, for marking historical distance and avoiding anachronism. But it also freezes systems of meaning into static cross-sectional snapshots that allow analysis of the self-relating structures of individual systems, but do not illuminate the interaction among systems or the processes of transformation from one system to another. Moreover, the employment of the structural definition of discourse as a closed semiotic system seems to dissolve historical agency. The identity of agents is called forth or constituted by the discourse itself and their statements and actions remain mere instances of the discursive rules that produce the reality in which they operate. A second dimension of discourse analysis, however, breaks apart the apparent stasis and structural determinism of this stance. Conceived on a conversational or dialogical model in which signifying practices operate in continuous processes of confrontation, exchange, and transformation, discourse becomes a term for

conceptualizing historical change, communicative interaction, and the dynamics of historical agency. From this point of view, linguistic or signifying actions are not authorless events in language but actions on language by individual and collective agents who appropriate inherited possibilities contained in existing discursive structures in order to both modify those structures and to transform themselves as agents. But how could the agents, the individual, and collective identities contingently and historically constructed within the structures of discourse, become themselves the creative transformers of discourse? One possible answer to this question could be constructed from within discourse theory itself, but ultimately it does raise the issue: what is the 'other' of discourse? and how is discourse changed and transformed in its relationship to its other?

Discourse and Discourses: Pluralism and Heterogeneity

In an early manifesto of discourse analysis, Foucault (1972) distinguished between a general domain that included all meaningful or signifying actions and their formal or structural conditions, and more individualized, singular groups of signifying practices organized in a focused disciplinary fashion around the construction of a coherent regime of truth. Historians have not been completely disinterested in the more sweeping, universalizing or totalizing notions of discourse that define the historical a priori of whole epochs or cultures, like the discourse of 'man' or secular historical humanism that rules over the 'modern' epoch, but they have tended to gravitate toward a practice of discourse analysis that emphasizes both the plurality of discourses and the conflict among discourses within a general discursive domain. Three forms of such heterogeneity in discourse analysis are particularly obvious. First, discourses can be differentiated in terms of the social and cultural dimensions or spheres in which they operate or are effective. Thus, one might differentiate among public, official, popular, literary, elite, urban, or religious variants of discourses of class or madness or justice or gender at any particular time or in any particular geographic area. Second, discourses obviously are differentiated in terms of the specific domains of meaning and action – the objects, subjects, and rules of relationship – which they construct. Thus, orientalist discourse constructs 'the east' as well as the 'orientalist'; psychiatric discourse constructs the psyche and the psychiatrist. Differentiations of this type are often imagined as analogous to academic disciplines that produced their putative objects of analysis in the way that the discipline of history constructs 'the past,' or sociology constructs 'society' or 'the social.' Third, historians tend to differentiate competing or alternative discourses that operate within these various dimensions and domains. Within the dimension of public discourse, for example, there may not only exist separate discourses pertaining to the spheres of political authority, citizenship, social hierarchy, or legal structures, but also dominant and oppositional, liberal and conservative, patriarchal and feminist, bourgeois and working class discourses that delineate the boundaries of this sphere and organize and define its contents and relations in competing and conflicting ways.

This proliferation of difference in the discursive realm is one of the major conditions of creativity and historical change. At any particular moment in time an historical actor will inhabit a heterogeneous compound of discourses, dissolving the apparent formal disassociation of discourses from each other, allowing for creative performances in which the inner logics of different discourses are brought in contact with each other in transformative ways. Just as a scholar can inhabit an interdisciplinary space in which various disciplinary discourses can be brought into interaction in order to transform conventional, dominant discourses about a particular issue or problem, an actor in the public sphere might exploit the plurality of existing constructions of citizenship in various inherited discourses to imagine and act out innovative forms of membership and responsibility in a political community. The first task for the historian engaged in discourse analysis is to identify a discourse within a given body of evidence, to map out its boundaries, and reconstruct its inner logic or structure. A discourse is not simply a synoptic summary of the aggregate of statements that refer to some specific theme, like references to the work of women in France between 1815 and 1848. To identify the discourse in such thematically associated material would be to reconstruct the formal dimensions within the signifying practices informing the evidence. Discourse would be the organizing power that constructs the 'world' the evidence assumes and provides effective power to certain statements. Once such a discourse has been mapped or reconstructed, the map can be used, like an 'ideal type' or 'paradigm' to identify or discover its presence in other texts or bodies of evidence. Such reconstructive mapping of historical a prioris establishes historical difference, marks off the conditions of possible action and expression in different worlds. The second task of the historian of discourse, however, would be to cut through or modify the radical difference produced in the first operation – to describe the ways in which the specific actions – linguistic and nonlinguistic performances with communicative intent – within the formal conditions of a discourse create changes in that discourse and reveal the porous, contingent nature of discursive boundaries. This model of discourse analysis as a constant oscillation between the reconstructions of formal contexts and concrete, innovative performances in a dialogical, communicative context has been highly developed within specific realms of intellectual history – like the history of political theory, public opinion, and ideology, but its expansion into a broader analysis of the whole range of historical practices has highlighted a number of problems, problems that have become the center of the controversies dividing historians about the validity of discourse analysis per se.

Discourse and its 'Others'

Discourse and discourse analysis only attain distinctive, specifiable definition if they are placed in relation to something which they are not. What are the 'others' of discourse? This question is directly relevant to two questions that are of particular interest to historians: How can one explain the stability and/or hegemony of certain discursive formations among specific populations over extended periods of time, and how can one explain changes within specific discourses, the shifting relationships among discourses, and the emergence of new discourses in time?

Discourse and the Material World

The polemical thrust of discourse analysis when it emerged as an articulated position in debates among historians during the 1980s was directed primarily against dominant conventions of historical analysis which portrayed the realm of the production, reproduction and transmission of meaning in linguistic and other signifying actions as a product, effect or expression of the structures and transformations of material practices. Despite the complexities within and obvious differences among class oriented British cultural Marxism, the sociologically grounded modernization narratives prevalent in the US and Germany, or the structuralist 'total histories' of the French Annales school, all were viewed as ultimately reducing the production and transformation of meaning to an effect, however indirect, of changes occurring outside of the linguistic and more broadly cultural sphere. Concepts like 'ideology' or 'collective mentality' were criticized and often rejected outright because of their implication that the production of meaning in language could only be grasped with reference to a reality or truth that was situated in material social practices outside of language. In its initial stages the linguistic turn often appeared as an attempt to resist the absorption of the cultural into the social, to insist on the autonomy of processes of meaning production, to assert the autonomy (no longer 'relative') of intellectual and cultural history. More serious tensions among historians developed, however, when the realm of discourse was conceived as creating the categories and identities through which people lived their social being and defined their social interests. Class identities were not defined as public expressions of the ways in which groups articulated social interests grounded in the way they lived their relation to nature in systems of production, but identities produced in discourse that determined how interests were defined and the ways in which economic and social relations were experienced. Gender differentiation in public discourse, or medical discourse, or class discourse was not conceived as representing the contrasting realities of social experience of physically differentiated bodies but as constructed forms in which bodies were experienced as sexually different. Histories of the experiences of classes, of the sexes, of races, ethnic groups, nationalities, colonial subjects, etc., were transformed into historical reconstructions of the discursive processes through which class, sexual, racial, ethnic, national, and colonial identities were constructed.

It would be misleading however, to treat this new cultural history, grounded in discourse analysis, as some historians have tried to do, as a new form of idealism based on a belief that language makes the material world, that thought creates existence. First, in discourse analysis, discourse itself is conceived as a dense network of signifying practices, a material activity embedded in the communicative interactive relations between producers and consumers of meaning and articulated in institutional structures. The linguistic turn did not simply exalt the status of language but changed the conception of what language is and how it works in history. Second, the 'other' of discourse analysis is not the empirical analysis of concrete evidence: discourse analysts certainly are as attentive to archival evidence as any social historian. The other of discourse analysis is a belief in the ultimate accessibility of an 'unmediated' meaningful reality on the other side of discourse which determines the truth of the meanings produced within discourse. Discourse analysts restrict the historical task to reconstructing the conflicted process of the cultural mediation of meaningless 'reality,' the practical human activity of meaning construction, transmission, and transformation. Yet the actual practice of discourse analysis may raise doubts about historians' ability to finally extirpate the metaphysical impulse to represent a truth beyond the represented 'truths' of human cultural practices. In the work of Michel Foucault and many of the historians who follow his lead, for example, one finds an assumption that the actual workings of discourse are determined by a prior will-to-power, by a universal drive to impose meaningful order on meaningless 'reality,' to narrativize the disassociated fragments of experience, to stabilize the endless proliferation of contingent identities. Discourse analysis, thus, not only marks a turn from the social to the cultural, but also a turn toward the political in the ubiquitous sense that all practices in the world have a political intent, that the legitimation, expansion, and consolidation of power is not a particular sphere of historical action, but the form of all historical practice. Discourse analysis has also been characterized by a renewed and transformed interest in the history of the body and of nature. Both the human body and the world of external nature are constituted within discourse, but at the same time are primary sites of resistance to any particular discursive construction. Within the framework of discourse analysis human and nonhuman physicality cannot have a positive unmediated reality of their own, a reality capable of self-representation within its own nondiscursive, transcendental 'language.' Bodies may be constantly inscribed and re-inscribed as meaningful in various discursive networks, as laboring organisms, as reproductive mechanisms, as emblems of spiritual content, but their sheer intractable existence also marks the limits of discursive constitution and thus fuels an infinite process of re-inscription and re-constitution.

Discourse Analysis and the Problem of Historical Agency

Discourse analysis is premised on the claim that effective historical agency, and thus the existence of individual and collective subjects who are capable of productive, creative, transformative action in the world, is produced in history rather than expressed in history. It is not subjects who produce meaning, but the networks of meaning that create subjects. In some earlier forms of discourse analysis, which was heavily tilted toward an analysis of the homogenizing, totalizing regularities in discursive structures, subjects, or authors often appeared as nothing more than anonymous, replaceable bearers (*Träger*) of the structures that defined them. In many ways this conception of discourse as a systematic process of meaning production mirrored the position of social historians who defined individual and collective agents in history as the inhabitants of social functions, or as bearers of the process of material production. Historians who took the linguistic turn in the 1980s and 1990s, however, were concerned with revealing the historical contingency of determining structures, with deconstructing systems of meaning production that allowed

certain identities – of class, race, gender, ethnicity, and so on – to be experienced as natural realities or metaphysical essences. Attentiveness to the conflic tual, dialogical, communicative dimensions of discourses in which historical identities were constantly constructed, modified, dissolved, and transformed did assume the effective agency of historical subjects (both individual and collective) who were able not only to negotiate among conflicting, multivocal, sometimes contradictory discourses, but also to redefine themselves as subjective agents.

Discursive agency, however, was decentered and contingent, the agency of constantly shifting 'identities' that were the creations of discursive practices themselves. Historical subjects were 'subjected' to the logics of the various discursive networks that they inhabited. But within the limitations defined by historical situation and status they made choices that might ultimately change that situation and status, and thus also their own identities. The question of agency in discourse analysis was not so much whether or not agency was possible within its terms, but what kind of agency was imaginable. Within the complex heterogeneous compounds of any cultural or discursive formation, a particular individual might simultaneously inhabit a whole series of dispersed, fragmented subject positions. Within each of them – as parent, worker, consumer, or as sexual, racial, or national subject – the individual could act, either individually or by identifying himself with other individuals similarly positioned in various discourses, to shift the hierarchical relations among identities, and to change the identities themselves. In discourse analysis, human agency, the identity and effectiveness of the individual and collective subject, was construed as an always coming-to-be in worldly particularly, as historical through and through.

This historicization of agency was at the same time a politicization of agency. The effective construction of specific collective identities, like nation, class, or race was described as a conflictual political process in which certain subject positions were excluded or privileged, became dominant or were repressed, attained stability through forceful imposition on others, or were negotiated in a more consensual fashion. The historicization and de-centering of human agency in the many historical works devoted to the cultural or discursive constructions of class, gender, national, or racial identities written during the late 1980s and through the 1990s did not fully expunge questions concerning the extralinguistic, culturally 'transcendent' grounding or foundation of human agency among historians. The writings of intellectual mentors of the linguistic turn, like Foucault, Walter Benjamin, or Jacques Lacan (and, in the1990s, Slavoj Zizek), were replete with references to the problematic nature not only of the physicality of the body but of the transcendent autonomy of some form of un mediated subjectivity – that which sought or was given shape and identity in the various 'subject positions' constituted in discourse – which appeared to be a necessary condition for the historical differentiation and transformation of subject identities. Historians might modestly decline professional competency in addressing such questions. But that which was outside discourse analysis inevitably shaped what occurred inside discourse analysis. The metaphysical limits of discourse analysis suggested that discourse analysis was itself a historical, that is, contingent and particular identity, a 'subject position' for historians of a particular place and time.

Discourse Analysis and the Discourse of Historians

The linguistic turn in history began with, and has continued to be characterized by, an intensive self-reflexive focus. With the publication of *Metahistory*, White (1973) introduced many historians to discourse analysis as an analysis of the discourse of historians. The primary object of concern was not so much the reconstruction of the way past identities were produced in language, as the construction of the individual and collective identities of historians through their discourses about the past. Stories historians told about the past were unveiled as stories in which they fashioned themselves, and possibly their readers, in the present. In the discourse on historical theory and historiography instigated by White's work, historians' texts are analyzed as linguistic artifacts akin to the rhetorical and fictional texts of literature, in which the structures of the linguistic form determine the 'objectivity' of the content. The discourse of objective realistic representation that had conventionally legitimated entry into the disciplinary community devoted to knowledge of the past was displayed as a mystifying, self-deceiving attempt to portray the subjective (that is, based on ethical and political considerations) construction of its object as a true representation of its object. The goal of such analysis was self-recognition and emancipation.

Once the burdens of determination by the putative objective reality of the past were recognized as self-imposed limitations, historians were free to imagine and construct alternative pasts in terms of their ethical and political commitments in the present and their visions of alternative futures. Moreover, as soon as the texts inherited from the past were grasped as historically contingent discursive constructions of meaningful identity in time, the inhabitants of the present could recognize themselves in them, not just constitute themselves through them. History could become a discovery of the 'reality' of the human past which was in itself a constantly reformulated attempt to represent reality as meaningful.

Two major questions continue to fuel the debate instigated by the analysis of the historical discipline as a discursive community. The first concerns the politics of this disciplinary community. How should the boundaries of legitimate historical discourse be determined in relation to other disciplinary discourses, or what are the limits of interdisciplinary crossing if history is to maintain a distinctive identity? Within the terms of discourse analysis this question must be decided through the communicative practices of historians rather than through an appeal to the object of analysis. The question then becomes one of defining the discursive process in which historical reality and truth are represented. Who is recognized as a 'subject' with rights to participate fully in the discourse, and how are the negotiations within this discourse to be regulated? The problem of reconstituting a common past becomes a political issue of extending membership and of democratizing power within the profession. Solidarity and objectivity are inextricably bound up with each other. One of the reasons that the linguistic turn and discourse analysis have become such

contentious issues among historians is that they pertain directly to the identity of history and the identity of historians.

The second issue raised by the reflexive moment in historical discourse analysis relates to the more general issue that troubles the whole debate surrounding the linguistic turn. What are the ethical (and political) implications of construing the past as a construction of discourse? Is the creation of an inclusive, democratically negotiated consensus about the nature of the past adequate to the goal of recreating the reality of the object of historical investigation and defining the criteria of historical knowledge? As the ongoing debates about historicizing the Holocaust have suggested, discourse about the truth of the past must involve the dead as well as the living. To converse with the dead entails a recognition not only of their physical existence, their extradiscursive 'reality' as bodies that transcend all the meanings inscribed on them, but also their status as actual or potential subjects of a discourse of their own that must be taken into account in any discourse we might construct about them. It would seem that an adequate conception of historical discourse would need to include some recognition that historians' discursive production of the past can be interrupted by the voices of the dead, and that a part of the ethical obligation of the historian is to listen to those voices.

See also: Community Studies: Anthropological; Discourse, Anthropology of; Foucault, Michel (1926–84); Historical Thought and Historiography: Current Trends; History: Representation Within and Functions of; History: Theories and Methods; Linguistic Turn; Maroons in Anthropology; Postmodernism: Methodology.

Bibliography

Baker, K.M., 1990. Inventing the French Revolution: Essays on French Political Culture in the Eighteenth Century. Cambridge University Press, Cambridge, UK.

Canning, K., 1994. Feminist history after the linguistic turn: Historicizing discourse and experience. Signs: Journal of Women in Culture and Society 19 (2), 368–404.

Central European History, 1989. German Histories: Challenges in Theory, Practice, Technique. 22 (3–4) (special issue).

Chartier, R., 1988. Cultural History: Between Practices and Representations. Cornell University Press, Ithaca, NY.

Eley, G., 1996. Is all the world a text? From social history to the history of society two decades later. In: McDonald, T.J. (Ed.), The Historic Turn in the Human Sciences. University of Michigan Press, Ann Arbor, MI, pp. 193–243.

Friedlander, S. (Ed.), 1992. Probing the Limits of Representation: Nazism and the Final Solution. Harvard University Press, Cambridge, MA.

Foucault, M., 1972. The Archaeology of Knowledge. Pantheon, New York.

Goldstein, J. (Ed.), 1994. Foucault and the Writing of History. Blackwell, Oxford, UK.

Hunt, L. (Ed.), 1989. The New Cultural History: Essays. University of California Press, Berkeley, CA.

Mills, S., 1997. Discourse. Routledge, London.

Palmer, B.D., 1990. Descent into Discourse: The Reification of Language and the Writing of Social History. Temple University Press, Philadelphia, PA.

Schoettler, P., 1989. Historians and discourse analysis. History Workshop 27, 37–65.

Scott, J.W., 1988. Gender and the Politics of History. Columbia University Press, New York.

Jones, G.S., 1983. Languages of Class: Studies in English Working class History 1832–1892. Cambridge University Press, Cambridge, UK.

Torfing, J., 1999. New Theories of Discourse: Laclau, Mouffe and Zizek. Blackwell, Oxford, UK.

White, H.V., 1973. Metapolitics: The Historical Imagination in Nineteenth-Century Europe. Johns Hopkins University Press, Baltimore, MD.

White, H., 1978. Tropics of Discourse: Essays in Cultural Criticism. Johns Hopkins University Press, Baltimore, MD.

Williams, G., 1999. French Discourse Analysis: The Method of Post-structuralism. Routledge, London.

Linguistic Typology

Masayoshi Shibatani, Rice University, Houston, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Modern linguistic typology is a method in language universals research. Implicational statements in the form of 'if X, then Y' are drawn based on cross-linguistic distributional patterns of existing and nonexisting types. These statements constitute language universals delimiting the range of possible human languages; for example, a language with X property but without Y property is excluded from the possibility. A functional, as opposed to a structuralist, typological approach allows comparison of languages in term of functional equivalence/disparity rather than in terms of categorial and structural properties, which differ from one language to another.

Linguistic typology evolved from a nineteenth-century attempt to classify the languages of the world and to ascertain possible variation among them. While the current practice in typology essentially retains these goals, modern typology differs from classical typology in its scope and methodology. Instead of typological characterization of an entire language, specific phenomena or individual constructions are studied. Constraints on possible variation are stated in terms of implicational universals derived from the distributional patterns of existing and nonexisting types. Typology is nowadays considered as a particular method of linguistic investigation. In contrast to other formal approaches to grammar, such as Chomskyan generative grammar, the typological method is functionally oriented in that what is sought is the typology of the relations between function and form and in that it seeks external explanation for the observed patterns. A careful study of different types of language suggests that languages may differ more radically in the organization of grammar than believed in the latter half of the twentieth century. It is likely that cross-linguistic similarity is more profitably sought in the functional equivalence and similarity than in the ways languages define and organize grammatical elements (see Section Typological Differences in Clausal Organization below).

The term typology in linguistics, as in other fields, is generally understood to mean the classification of languages in terms of structural types. However, from the very beginning in the history of typological investigation, classification itself has never been the ultimate goal. For the grammarians of the early nineteenth century, exhaustive classification of the world's languages was a way to answer the question of what is possible variation in human language. For modern linguists, typological classification is part of a method for discovering the constraints governing possible human languages or possible grammatical constructions.

Typological classification is frequently contrasted with the genetic classification of languages and is often characterized as a historic in its orientation. This was also not the case in the beginning of the typological studies in the early nineteenth century, when genetic and comparative approaches were not clearly distinguished from the typological one. Indeed, it was in the context of developing a program of comparative historical research in his 1808 book, *Über die Sprache und Weisheit der Indier* (On the Language and Wisdom of the Indians), that Friedrich Schlegel (1772–1829) developed a typological framework most famous in the history of typology. It was mainly due to the structuralist context, in which modern typology arose, that it countenanced a synchronic outlook. In contemporary typological scenes, dynamicization of typology, as propounded most forcibly by Greenberg (1969, 1995), is fast obliterating the view that typology is a purely synchronic discipline.

Classical Typology

In the abovementioned book, Friedrich Schlegel proposed what is believed to be the first typological framework, according to which the languages of the world are classified into two types – flexional (or inflectional) type and affixing (or agglutinative) type. Though later typologists, most notably August Wilhelm Schlegel (1767–1845) and Wilhelm von Humboldt (1767–1835), elaborated on his bipartite classification, Friedrich Schlegel, alluding to the ultimate goal of typological classification, made a bold claim that the opposition between *Sprachen durch Flexion* and *Sprachen durch Affixa* represents the two main categories of language, all other cases being mere modifications and variants of these.

The typological studies of the early nineteenth century are also known as morphological typology, since the classificatory basis was the morphological structure of words. A beginning linguistics textbook typically introduces four types of language from this early endeavor: (1) inflectional (or fusional) type: relational meaning is expressed by internal modification of the lexical base such that component parts are not easily isolable (e.g., Indo-European languages, English *sing*:*sang*:*sung*); (2) agglutinative type: words consist of easily segmentable parts, each indicating a distinct grammatical meaning (e.g., Turkish *ev-ler-im-de* house-PL-1SG.POSS-LOC 'in my house'); (3) isolating type: words do not give overt expression of relational meaning or else do so by employing to the end the same kind of unit as is used for encoding lexical concepts (e.g., in Chinese the verb *gěi* 'to give' is used also as an indirect object marker); and (4) incorporating (or polysynthetic) type: nominal arguments (typically objects) and pronominal forms referring to them are incorporated in the verb form (e.g., Ainu *a-wakka-ta-re*1SG-water-draw-CAUS 'I make someone draw water').

 http://dx.doi.org/10.1016/B978-0-08-097086-8.52009-1

Classical typology embodied system internal inconsistencies and had associated value judgments that caused its decline in the second half of the nineteenth century, when the study of genetically related languages was viewed as the only possible scientific comparative study of languages. Different languages were thought to represent different stages of the development of the languages and language families with the belief that the inflectional type, which fused distinct morphological information into single morphemes, represented the highest degree of linguistic sophistication. This assumption put classical Indo-European languages at the top of the evolutionary ladder and isolating languages of Asia, e.g., Chinese, at the bottom, contradicting the long history of the Chinese language and the advanced civilization associated with classical Chinese.

The quadripartite classification has shortcomings as a true classificatory scheme in that a language may have morphological characteristics of several types simultaneously. English, for example, shows an inflectional characteristic in some irregular forms (e.g., *bring*:*brought*, *man*:*men*), an agglutinative characteristic in regular verbal and nominal inflections (e.g., *walk*:*walked*, *book*:*books*), an isolating characteristic in certain modal expressions (e.g., *cannot walk*), and an incorporating feature in some compound forms (e.g., *to hand-knit a sweater*). Edward Sapir's multidimensional typology developed in his 1921 book *Language* (Sapir, 1921) was in part response to the need to accommodate gradient characterizations of linguistic types, as he saw that languages in their entirety could not be neatly pigeonholed into a given type, the matter being a question of tendency.

While classical typology saw the culmination in Sapir's highly elaborate classificatory scheme, his followers have not generally put it into practice. Indeed, the only place where classical, morphology-based typology has been seriously pursued is Prague, where Vladimir Skalička (1909–91), who recognized five types of languages (the four previous types plus the introflexive type), was a pillar of the so-called Prague School of Typology (see Sgall, 1995).

While morphological typology as formulated by the nineteenth-century grammarians is no longer considered to be adequate in meeting the ultimate goal of language typology, classical typology saw several developments of contemporary relevance. It is perhaps even fair to say that the followers of classical typology have contemplated most contemporary issues of language typology. For example, August Schleicher (1821–68) recognized the possible connections between morphological characteristics and the manner in which grammatical relations are expressed, thereby pointing to the current view that a language type is to be defined in terms of a combination of properties that exhibit correlative patterns. The evolutionary chain of isolating agglutination inflection envisioned earlier is now recast as a cyclic development, and such a model is integrated in the current study of grammaticalization (see Bybee, 1985; Hagège, 1988). The Humboldtian claim that the structure of a particular language reflects its speakers' *Weltanschauung* has stimulated a line of research on the relationship between language types and the speakers through the nineteenth century leading up to the so-called Sapir–Whorf hypothesis in the early twentieth century. This hypothesis claims a connection between the linguistic structure and the way the speakers perceive the outside world. As typological studies take on more cognitive outlook (see DeLancey, 1981), a reappraisal of the Humboldtian tradition may be in order. An interesting development along this line is the recent revival of interest in ethnopsychology and the development of what is dubbed as 'ethnosyntax' by Wierzbicka (1992) and others.

Modern Typology

A major difference between classical typology and modern typology, developed in the middle of the twentieth century, lies in its scope and methodology. Classical typology sought a global characterization of the entire language according to a small number of typological characteristics. This holistic approach has proved to be too ambitious, and modern typologists practice partial typology, where specific phenomena or individual grammatical constructions are studied. Methodologically, modern typology heavily relies on cross-linguistic distributional patterns. Implicational universals are drawn from the distribution of existing types and nonattested types.

Roman Jakobson (1896–1982) is generally regarded as the founding father of modern typology. In his seminal paper, Jakobson (1958/1971) recognized that typological analysis leads to the discovery of the dependency relation between two elements such that the presence of one element, X, implies the existence of another element, Y, in the same system (but not vice versa). For example, the parameters of voicing and aspiration in consonants yield the following four types of consonant system (illustrated with dentals): (1) /t/ and /d/; (2) /t/, /t^h/, and /d/; (3) /t/, /t^h/, /d/, and /d^h/; and (4) /t/, /d/, and /d^h/. Of these four logically possible types, only the first three are actually observed – no language has voiced aspirates in the absence of voiceless aspirates, i.e., the fourth system. From this, it is possible to formulate an implicational statement that the presence of voiced aspirates implies the presence of voiceless aspirates. This statement constitutes a universal constraint on human language if no language indeed violates the pattern implied by this statement – that no consonantal system can contain voiced aspirates in the absence of voiceless aspirates.

Developing Jakobson's insight in a systematic and overarching manner, Joseph H. Greenberg (1915–2000) has changed the whole perspective of linguistic typology, transforming it into a new method of universals research. Greenberg's (1966) study on word-order typology drew a large number of implicational universals encompassing both morphological and syntactic levels, and some transcending them. That languages differ in word order is a commonplace experience of those who are exposed to foreign languages. English is an SVO (Subject–Verb–Object) language, while Japanese is an SOV language. The less frequently observed VSO order is instantiated by Welsh. The impact of Greenberg's work was very strong because of the far-reaching predictions it made on the basis of this kind of simple, basic feature of language. For example, it has been shown that if a language has dominant VSO order, (1) it has prepositions (G's *Universal 3*), (2) the genitive follows the governing noun (G's *Universal 2*, originally formulated in terms of the presence of prepositions), (3) question particles that are specified in position by reference to a particular word in the sentence do

not occur (G's *Universal 10*), (4) it always puts interrogative words or phrases first in interrogative word questions (G's *Universal 16*), (5) an inflectional auxiliary precedes the main verb (G's *Universal 16*), and (6) adjectives come after the noun (G's *Universal 17*).

Greenberg's universals have the following two characteristics. They are implicational such that they are restricted universals. The universals cited above obtain only with regard to VSO languages, and they have nothing to say about those languages having a different word order. Secondly, Greenberg recognizes statistical universals or high cross-linguistic tendencies, which do not obtain absolutely. For example, Greenberg's *Universal 4* is formulated as follows: "With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional," which allows a language like Persian, which has SOV order with prepositions. Beside implicational and statistical universals, linguists, in general, recognize unrestricted universals, which are believed to obtain absolutely; for example, all languages have vowels and consonants, and all languages have nouns and verbs.

Implicational statements are also obtainable from hierarchies that grammatical categories exhibit. The most well-known grammatical hierarchy is that of grammatical relations formulated by Keenan and Comrie (1977): Subject > Direct Object > Indirect Object > Genitive > Oblique Object > Object of Comparison. This hierarchy, derived from the cross-linguistic relativization patterns, predicts, for example, that if a language allows relativization on an oblique object (i.e., an object of a preposition), then it also allows the same relativization strategy to apply to all those nominals holding the grammatical relations higher in the hierarchy than the oblique object.

Another hierarchy with far-reaching ramifications is the so-called animacy hierarchy: first/second person > third person > pronoun > human noun > animate noun > inanimate noun.

In a language such as the Australian language Dyirbal, where an agentive nominal of a transitive clause may receive either ergative or nominative marking – split-ergative marking – those high on the hierarchy tend to be marked nominative, while those low on the hierarchy tend to be marked ergative such that we can draw implicational statements to the effect that if an agentive nominal X in the hierarchy is marked nominative, all those higher in the hierarchy are also marked nominative. The animacy hierarchy is also at work in voice phenomena such that an agentive nominal high in the hierarchy tends to be expressed in subject position, i.e., in the active voice, while when a patient nominal is high in the hierarchy, passivization tends to occur. A corollary of these tendencies is that those nominals high in the hierarchy are less likely to be expressed overtly as an agentive phrase (as in the English *by* agentive phrase) in the passive clause.

Recent studies of prototype categories have revealed that grammatical categories and constructions form continua rather than exhibiting rigid category boundaries. For example, studies on transitivity have shown that the traditional binary distinction between intransitive and transitive verbs (or constructions) needs to be recast as a multivalued continuum (see Hopper and Thompson, 1980). A simpler and much coarser form of this continuum yields the following hierarchy: transitive > active intransitive > inactive intransitive, where active intransitive verbs/constructions are those expressing agentive activities (e.g., 'to dance') and inactive intransitive verbs simple nonagentive processes (e.g., 'to fall'). An implicational statement can be also derived from this hierarchy; for example, if a language permits passivization of an inactive intransitive, then it also permits active intransitives and transitive verbs to undergo passivization. Also, noun incorporation (see above) basically follows this hierarchy such that the object of a transitive clause is the easiest to incorporate, followed by the subject of an inactive intransitive clause, the subject of an active intransitive clause, and finally the subject of a transitive clause. Thus, if a language incorporates the subject of an active intransitive clause, then it is likely to incorporate the subject of an inactive intransitive clause and the object of a transitive clause.

Functional Typology

The term typology is used in reference to a theoretical and methodological approach that contrasts with other, especially Chomskyan generative, approaches. The Chomskyan generative approach is said to be formal in two senses. In one sense, generative grammar is formal in dealing with the "abstract formal skeleton" (Chomsky, 1975, p. 55) of syntactic structures as an autonomous system independent of semantic or functional factors. In another sense, generative grammar is formal in that its explanations are in terms of the system internal formal categories (e.g., noun phrases, verb phrases) and structural relations. The typological approach sharply contrasts with these two characteristics of generative grammar. Its basic orientation is functional in that it assumes that the central function of language is to represent cognition and to communicate experience and thought. The task of typology is to determine the range of variation that language permits in achieving this central function. In this framework, a language is seen as a problem-solving system: how languages conceptualize a given cognitive domain and express the concepts by linguistic means (see Seiler, 1995). Studies dealing with specific constructions such as causative and possessive constructions thus start out with the formulation of the concepts or the cognitive domains to be typed in terms of the correlations between grammatical form and semantic function (see Shibatani, 1976; Seiler, 1995). This cognitive-functional stance has a long and venerable history linking the current works with Sapir's and ultimately with the Humboldtian conception of language as an '*enérgeia*' (a creative activity) rather than an '*érgon*' (a product).

The term 'functional typology' is used in recent years in order to underscore the functional orientation countenanced by the current typological research. This approach also differs from generative grammar in its recourse to external explanations. Explanations for cross-linguistic patterns of linguistic structure are sought in functional motivations such as iconicity – preference for the structure of language to reflect and the conceptual structure in a parallel manner (see Haiman, 1985). Haiman has shown that conceptual distance is reflected in grammatical distance. Thus, inalienable possession, which involves an inseparable or a closely connected whole-part relation, is expressed by a more tightly integrated form than

alienable possession; for example, *ji syim* 'my arm' versus *ji bi nggwe* (my of house) 'my house' (Abun; West Papuan). Similarly, in causative expressions, shorter forms tend to express direct causation and the longer forms indirect causation, e.g., *kill* versus *cause to die*. In recent years, however, these iconicity-based explanations have been challenged and alternative usage-based explanations have been explored. These efforts are informed of some earlier observations on economically motivated tendencies such as Zipf's: "the length of a word tends to bear an inverse relationship to its relative frequency," "the degree of distinctness of meaning seem[s] to bear an inverse relationship to F[requency] and C[rystalization] [of the configuration]" (Zipf, 1935; see also Haspelmath, 2008).

Another line of research along the functional typological premise has been developed by Hawkins and his associates in which processing mechanisms are invoked as an explanation for observed linguistic patterns (see, e.g., Hawkins, 1990; Hawkins and Cutler, 1988).

Typology and Language Change

The importance of typology to one of the major concerns of historical linguistics, namely reconstruction of protolanguages, has long been demonstrated. The basic idea here is that where a reconstruction conflicts with the known typological facts, there is a *prima facie* reason to suspect its validity. The most celebrated phonological example of this use of typology as a control on reconstruction is concerned with the Proto-Indo-European consonant system. The standard reconstruction posited a threefold system of voiced and aspirated consonants (illustrated with dentals /t/, /d/, /d^h/). However, Jakobson (1958/1971) noted that typological investigation revealed that no language would contain /d^h/ in the absence of the voiceless counterpart /t^h/, casting doubt on the validity of the Proto-Indo-European system reached by the standard reconstruction. This has led to either reinstating the neogrammarians' earlier four-point stop system of /t/, /t^h/, /d/, /d^h/, or developing alternative hypotheses. The so-called glottalic theory, which replaces the traditionally reconstructed voiced stops with ejectives, has also been informed of typological results. Phonological systems comprising a voiceless, a glottalized, and an aspirated series (/t/, /t^ʔ/, /t^h/) conform to synchronic typological findings, but the system arrived at by the conventional reconstruction does not (see Gamkrelidze and Ivanov, 1973).

The move to bring typology to the heart of diachronic linguistics, namely linguistic change itself, has been spearheaded by Greenberg (1969, 1995), who advocates the dynamicization of a synchronic typology. Greenberg offers what he calls the 'state-process model,' which views attested language types as possible stages through which languages can pass and nonattested language types as impossible stages into which a language can change. In other words, just as synchronic typology puts constraints on language types, constraints on typological change can be viewed as limitations on language change.

The connection between synchronic typology and possible change can be illustrated by the typology of number categories in the noun. It has been shown that the category of dual is only found in a system having a singular–plural distinction – no language has a singular–dual system or a plural–dual system. When a language with no number distinction develops a number system, it involves a singular–plural distinction; no language develops a system with a tripartite singular–dual–plural distinction without going through a stage in which a singular–plural distinction has been established. In this way, diachronic typology seeks to identify allowable and non-allowable transition from one state to another so as to determine possible paths of language change.

In the Greenbergian diachronic typology, typological states and transitions between states are characterized by two logically independent factors – stability and frequency. Stability represents a probability that a language, which is in a particular state, will exit from that state, while frequency is the likelihood that a language will enter a particular state. For example, a state of having nasal vowels is characterized as having high frequency but low stability, because a vowel frequently acquires the nasality feature from an adjoining nasal consonant, but once a system acquires nasal vowels, it tends to retain them. The dimensions of frequency and stability of typological states render the state-process model a probabilistic one, which countenances the gradualness of the transition from one language state to the next. In this model, synchronic exceptions are accounted for as transitional states arising from gradual diachronic process of change from one state to another.

The two main concerns of diachronic typology – gradualness of historical change and its directionality – have received focused attention in the recent developments in grammaticalization research (see Bybee, 1985; Heine et al., 1991; Hopper and Traugott, 1994; Lehmann, 1982). Grammaticalization refers to the process by which lexical items come to take on grammatical functions such as marking of tense, aspect, grammatical relations, definiteness, modality, etc. In Chinese, the verb *gěi* 'to give' is used as an indirect object marker, and the verb *iru* 'to be/to exist' in Japanese has developed an aspectual usage of indicating an ongoing activity, the progressive aspect.

Grammaticalization typically involves a synchronized change in several different grammatical dimensions. As a lexical item gets grammaticalized, it tends to undergo phonological reduction, which in turn renders the form less independent, downgrading it to a morphologically dependent status, as in the case of enclitics and affixes. Semantically, a lexical item gradually loses its original referential meaning and acquires such relational meanings as those previously noted as well as subjective and interpersonal meanings. As in the case of the Chinese verb *gěi* and the Japanese verb *iru*, grammaticalization is a gradual change such that these forms coexist as verbs along with the newly developed grammatical functions. The three main concerns of grammaticalization theory are as follows:

1. the relationship between the input and the output of grammaticalization – What kind of lexical items give rise to what kind of grammatical functions?,
2. the directionality of change – Is the change unidirectional from concrete to abstract?, and
3. the gradualness of change – Is grammaticalization a gradual change or can it happen instantaneously?

The gradual aspect of historical change brought to the fore by diachronic typology and grammaticalization research has an important implication to the matter of definition as well as the synchronic characterization of a language. Many phenomena are in transition from one state to another, and as such they are not amenable to categorical definitions. For example, whether Chinese has prepositions or not is not an easy question to answer, as forms such as the indirect object maker evolving from the verb *gěi* 'to give' can be construed either as verbs or prepositions. The existence of transitional phases of this kind calls into question the traditional distinction between synchronic linguistics and diachronic linguistics. Dynamicization of a synchronic description is the next logical step that diachronic typology suggests.

Typological Differences in Clausal Organization

As described in Section Typology and Language Change, modern typology typically examines specific domains or individual grammatical constructions in order to ascertain possible variation in human language. This leaning toward partial typology stems from the difficulty encountered in attempting to typologically characterize languages in their entirety. Despite this, there is a lingering sense among typologists that a holistic characterization of a language is a desideratum. Georg von der Gabelentz most eloquently expressed this feeling at the beginning of the twentieth century:

> But what an achievement would it be were we to be able to confront a language and say to it: "you have such and such a specific property and hence also such and such further properties and such and such an overall character"—were we able, as daring botanists have indeed tried, to construct the entire lime tree from its leaf.
>
> Gabelentz, 1901, p. 481

Whether or not the ideal entertained by Gabelentz can ever be reached remains to be seen. However, one area that merits further investigation is the differences that languages show in the organization of clause elements. In compensation for the emphasis on methodological matters in the preceding *exposé*, the following allows a glimpse of what actual typological investigation or comparison of languages in general involves.

In expressing a transitive situation involving an agent and a patient participant, languages typically distinguish these in some way so that who is acting on whom is made clear. English, for example, achieves this by using different pronominal forms, e.g., 'He hit him.' In the case of intransitive clauses, its sole nominal argument (S) can be assimilated either to the form of an agentive nominal (A) or to that of a patient nominal (P). English assimilates it to the nominal form of A – 'He fell, He ran/*Him fell, *Him ran.' A clause that treats S and A alike and distinguishes P from these is said to display the nominative-accusative coding pattern ($S = A \neq P$), and the languages possessing such clauses predominantly are called nominative-accusative-type languages.

Some languages organize clauses differently from the nominative-accusative pattern. Many Australian Aboriginal languages follow the absolutive-ergative pattern, whereby S is identified with the P nominal, a transitive clause, while distinguishing A from these, i.e., the $S = P \neq A$ pattern. Clauses may be organized in yet another way. The American Indian language Lakhota, for example, splits S nominals into an agent-like category (S_a) and a patient-like category (S_p), yielding the coding pattern, known as active-type: $S_a = A \neq P = S_p$.

All of these languages have a single type of dominant transitive clause, where the coding of A and P is fixed. Some languages have more than one type of transitive clause, making a unique comparison of intransitive clauses with transitive clauses problematic. Philippine languages are such languages, where a simple transitive clause comes in two variant forms. Cebuano, for example, renders the sentence 'Juan read the book' in two ways: *Ni-basa si Juan sa libro* (AF-read TOP Juan the book) and *Gi-basa ang libro ni Juan* (GF-read TOP book ACT Juan). (AF = actor-focus marking; GF = goal-focus marking; TOP = Topic marker; and ACT = Actor marker.) That is, either A or P can be coded as what is called Topic (marked *si* for proper names and *ang* for common nouns) of a transitive clause. The S nominal of an intransitive clause is coded as a Topic as in *Ni-dagan si Juan* (AF-run TOP Juan) 'Juan ran.' The focus marking in the verb indicates what semantic role is being coded as Topic such that the prefix *ni-* in Cebuano indexes an agentive Topic, and *gi-* a patient Topic. This applies to both transitive and intransitive clauses such that an intransitive clause with a patient S Topic would trigger *gi-* marking; *Gi-kapoy si Juan* (GF-tired TOP Juan) 'Juan is tired.' This split treatment of S Topic marking in the verb is reminiscent of the active-type coding pattern. However, Philippine languages are not active-type languages in that the nominal marking is consistent for S – it is always realized as *si*/*ang*-marked Topic – and that there are two competing forms of a transitive clause (see Shibatani (1985) on the Philippine voice system).

The four types of coding pattern examined have important implications for some basic grammatical categories. First, the notion of Subject is problematic in those clause types displaying coding patterns other than the nominative-accusative pattern. Subject, as found in European languages, refers to the category obtained from the $S = A$ grouping in the nominative-accusative pattern. It is a category deriving from the assimilation of S to A, and accordingly is agent-based.

Morphological identification of S and P obtains in an ergative-type language. A long-standing question in typology is whether this grouping of S and P should be identified as Subject. One obvious problem with this move is that whereas Subject is an agent-based category, the $S = P$ grouping is a patient-based category, since it obtains by assimilating S to P. Thus, a conservative view is that the $S = P$ grouping forms a category distinct from Subject. Since S and P typically involve zero marking, the term Absolutive suggests itself as the label for this category.

In active-type languages, there is no unique identification between the totality of S and A or P, and therefore, neither Subject nor Absolutive category obtains. In Philippine linguistics, Topic is often likened to Subject, but since it is based on neither A nor P – either can be Topic, and S can be identified with either kind of Topic – it is neither an agent-based nor a patient-based category.

While different clausal types of the world's languages may display grammatical relations resembling Subject, they are not entirely identical. Subject is agent-based, whereas Absolutive is patient-based. Philippine Topic is neutral in its alignment with the semantic roles. These different grammatical relations, however, can be identified at a functional level in that they are the essential relations that a nominal must bear in clause formation. That is, Subject, Absolutive, and Topic are an indispensable nominal category in a clause over which a verb predicates. Predication will not be complete without a nominal category about which the verb says something. The primary status of these grammatical relations is reflected in the primacy they enjoy in such grammatical processes as relativization and focusing. Subject, Absolutive, and Topic, in other words, can be identified as a primary relation at the functional level, despite the differences at the substantive level.

The differences in grammatical relations have some additional implications. Firstly, it predicts that those languages that do not have the Subject category do not show the active–passive voice opposition. Active voice (as expressed by 'John hit Bill') is defined in terms of the alignment of A and Subject and of P and Object, while passive voice entails the alignment of P and Subject and of A with Oblique (a peripheral, typically optional syntactic element), as in 'Bill was hit by John.' From this, it is obvious that languages without the Subject category cannot have the active–passive opposition in exactly the same way as in nominative-accusative-type languages. In place of the active–passive opposition, ergative languages often display ergative-antipassive voice alternation, characterized by the realignment of A from Ergative to Absolutive and of P from Absolutive to Oblique. Some ergative languages, in addition, display an analogue of the passive construction aligning P with Absolutive and A with Oblique.

In both active–passive and ergative-antipassive oppositions, there is a basic voice construction associated with the basic form of the verb. In the former, the nominative-accusative active form is the basic voice, and in the latter, the form coded in the absolutive-ergative pattern is basic. In this regard, the active and the ergative voice are a primary voice category and are functionally equivalent. Derived voice forms, the passive and the antipassive, typically contain morphologically derived verb forms, and they are used far less frequently than the unmarked, basic voice forms. Philippine-type languages, in contrast, have no unmarked voice category – the alignment of Topic with A or P is equally marked in the verb. A characteristic of the fluid-voice system of this kind is that both A-Topic form and P-Topic form are used with equal facility such that there is no lopsided distribution of voice forms as in the other two systems discussed previously. In contradistinction to these types of languages, active-type languages are known to lack voice opposition of the kind discussed here. The three voice mechanisms studied – the active–passive system, the ergative-antipassive system, and the fluid-voice system of Philippine languages – are different in their substance. Yet, they are comparable at the functional level in that they all involve realignment of a primary argument.

Languages can differ with respect to the nature of grammatical relations and voice mechanisms depending on the way clauses are organized. Besides these, cross-linguistic differences in other grammatical phenomena have been reported, some of which may be attributable to the typological differences outlined here. The question of whether these differences have profound implications for the overall organization of grammar aside, it is becoming increasingly clear that languages are likely to differ at the level of grammatical organization far more greatly than was believed in the latter half of the twentieth century, when categories and construction types were transferred from one language to another rather carelessly on the basis of partial similarity. Once a clear distinction is made between the substantive aspects of categories and constructions and their functional status within each language, languages can be profitably compared in terms of the functional equivalence or similarity even if they define and organize grammatical elements in quite different ways.

See also: Color Classification and Symbolism; Color Terms, Linguistics of; Functional Brain Imaging of Language processes; Phonology; Reading Nonalphabetic Scripts; Sign Language.

Bibliography

Bybee, J.L., 1985. Morphology: A Study of the Relation between Meaning and Form. John Benjamins, Amsterdam.

Chomsky, N., 1975. Reflections on Language. Pantheon Books, New York.

Comrie, B., 1989. Language Universals and Linguistic Typology, second ed. University of Chicago Press, Chicago.

Croft, W., 1990. Typology and Universals. Cambridge University Press, Cambridge, UK.

DeLancey, S., 1981. An interpretation of split ergativity and related patterns. Language 57, 626–657.

Gabelentz, G., 1901. Die Sprachwissenschaft. Ihre Aufgaben, Methoden und bisherigen Ergebnisse. Tauchnitz, Leipzig, Germany.

Gamkrelidze, T.V., Ivanov, V.V., 1973. Sprachtypologie und die Rekonstruktion der gemeinindogermanischen Verschlüsse. Phonetika 27, 15–56.

Givón, T., 1984/1991. Syntax: A Functional-Typological Introduction, 2 vols. John Benjamins, Amsterdam.

Greenberg, J.H., 1966. Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, J.H. (Ed.), Universals of Grammar, second ed. MIT Press, Cambridge, MA, pp. 73–113.

Greenberg, J.H., 1969. Some methods of dynamic comparison of linguistics. In: Puhvel, J. (Ed.), Substance and Structure of Language. University of California Press, Berkeley, CA, pp. 147–203.

Greenberg, J.H., 1974. Language Typology: A Historical and Analytic Overview. Mouton, The Hague, The Netherlands.

Greenberg, J.H., 1995. The diachronic typological approach. In: Shibatani, M., Bynon, Th. (Eds.), Approaches to Language Typology. Clarendon Press, Oxford, UK, pp. 143–166.

Greenberg, J.H., Ferguson, C.A., Moravcsik, E.A. (Eds.), 1979. Universals of Human Language, 4 vols. Stanford University Press, Stanford, CA.

Hagège, C., 1988. Contribution des recherches typologiques. In: Joly, A. (Ed.), Linguistique génétique. Historie et théorie. Presses Universitaire de Lille, Lille, France.

Haiman, J., 1985. Natural Syntax: Iconicity and Erosion. Cambridge; University Press, Cambridge, UK.

Haspelmath, M., 2008. Frequency vs. iconicity in explaining grammatical asymmetries. Cognitive Linguistics 19 (1), 1–33.

Haspelmath, M. (Ed.), 2001. Language Typology and Language Universals: An International Handbook. Walter de Gruyter, Berlin.

Hawkins, J., 1990. A parsing theory of word order universals. Linguistic Inquiry 21, 223–261.

Hawkins, J., Cutler, A., 1988. Psycholinguistic factors in morphological asymmetry. In: Hawkins, J. (Ed.), Explaining Language Universals. Blackwell, London, pp. 280–317.
Heine, B., Claudi, U., Hünnemeyer, F., 1991. Grammaticalization: A Conceptual Framework. The University of Chicago Press, Chicago.
Hopper, P., Thompson, S.A., 1980. Transitivity in grammar and discourse. Language 56, 251–299.
Hopper, P., Traugott, E., 1994. Grammaticalization. Cambridge University Press, Cambridge, UK.
Jakobson, R., 1958/1971. Selected Writings. Typological Studies and Their Contribution to Comparative-Historical Linguistics, vol. 1. Mouton, The Hague, The Netherlands.
Keenan, E., Comrie, B., 1977. Noun phrase accessibility and universal grammar. Linguistic Inquiry 8, 63–99.
Lehmann, C., 1982. Thoughts on Grammaticalization. A Programmatic Sketch, vol. 48. Arbeiten des Kölner Universalien-Projekts.
Ramat, P., 1987. Linguistic Typology. Mouton de Gruyter, Berlin.
Sapir, E., 1921. Language. Harcourt, Brace and World, New York.
Seiler, H., 1995. Cognitive-conceptual structure and linguistic encoding: language universals and typology in the UNITYP framework. In: Shibatani, M., Bynon, Th. (Eds.), Approaches to Language Typology. Clarendon Press, Oxford, UK, pp. 273–325.
Sgall, P., 1995. Prague school typology. In: Shibatani, M., Bynon, Th. (Eds.), Approaches to Language Typology. Clarendon Press, Oxford, UK, pp. 49–84.
Shibatani, M. (Ed.), 1976. The Grammar of Causative Constructions (Syntax and Semantics 6). Academic Press, New York.
Shibatani, M., 1985. Voice in Philippine languages. In: Shibatani, M. (Ed.), Passive and Voice. John Benjamins, Amsterdam.
Shibatani, M., Bynon, Th. (Eds.), 1995. Approaches to Language Typology. Clarendon Press, Oxford, UK.
Song, J.J. (Ed.), 2011. Oxford Handbook of Linguistic Typology. Oxford University Press, Oxford.
Wierzbicka, A., 1992. Semantics, Culture, and Cognition. Oxford University Press, Oxford.
Zipf, G.K., 1935. The Psycho-Biology of Language. Houghton Mifflin Company, Boston.

Linguistics, History of

EF Konrad Koerner, Zentrum für Allgemeine Sprachwissenschaft (ZAS), Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The development of the history of linguistics as a serious scholarly activity has its own long history that can be traced in the textbooks from mid-nineteenth to the mid-twentieth century. It is, however, only since the 1970s that the history of linguistics/language sciences has developed into a subdiscipline of linguistics in general and a subject to which the beginning level students of the discipline would be introduced as part and parcel of their formal education. In this article four main approaches to writing the history of linguistics are distinguished and discussed: (1), the summing up of the history, which presents the results of established procedures within the discipline; (2) the revolutionary history intended as propaganda for a new paradigm; (3), the non-partisan constructive history, which emphasizes continuity and progress in the discipline; and (4) the historiography of linguistics, a relatively new approach that seeks to treat the history of linguistics as a scholarly endeavor in its own right which may also inform work in other linguistic subdisciplines.

A discipline comes of age when it seriously contemplates its own past. The History of Linguistics (HoL) – now frequently referred to as 'History of the Language Sciences' – is an attempt to steer away from a narrow view of 'linguistic science.' HoL as a bona fide subject of academic research (in which doctoral dissertations can be written, for instance) began to develop only during the late 1960s, although such work had previously been done in departments of Germanic, Romance, or Slavic where such research surveys were at times undertaken to delineate the course a particular field had taken or the evolution of a specific idea or research project. It has been customary, at least since the 1880s, to add an historical introduction to textbooks in linguistics, but usually the intent has been to show the significance of recent advances in the field compared with previous endeavors. During the 1960s, following various claims made by Noam Chomsky that his own theories had little to do with the pursuits of his immediate predecessors and contemporaries, but instead followed quite different traditions such as those of the Port Royal Grammar and of Wilhelm von Humboldt, the bulk of the dissertations written in HoL were devoted to just these areas of interest, at times seriously distorting the true intent and purpose of these earlier authors. Only from the 1970s onwards, following the creation of the first journal for this particular field of interest in 1973, *Historiographia Linguistica*, and the associated monograph series united under the umbrella titled 'Amsterdam Studies in the Theory and History of Linguistic Science,' did serious work begin to emerge that challenged this *pro-domo* type of history writing. These and other organized activities (which we discuss in the subsequent sections) led to much more recent field of study, now generally referred to as Historiography of Linguistics (see Koerner, 2006 for an extended discussion) or 'Linguistic Historiography' for short, an approach to HoL which is conscious of methodological and epistemological requirements in adequate history writing in linguistics as in any science. (Most of the contributions to the two-volume *Historiography of Linguistics* published in 1975 under the main editorship of Thomas A. Sebeok were little else but surveys of previous scholarship; 'historiography' there being used in the old sense of the term; the volume edited by Parret in 1976 followed largely the Chomskyan manner of misappropriating the history of linguistics for 'political' purposes.)

It is true that we could perhaps by now speak of a 200-year tradition of linguistic history writing, perhaps beginning with François Thurot's (1768–1832) 1796 *Tableau des progrès de la science grammaticale* (cf Andresen, 1978), although several earlier works have been cited, for instance Elias Caspar Reichard's (1714–1791) *Versuch einer Historie der deutschen Sprachkunst* of 1747 (cf Koerner, 1978a: vi, for references to other eighteenth-century works). However, as the record suggests (Koerner, 1978a: pp. 1–4), it is only from the late 1860s onwards that a more thorough type of treatment of HoL emerges of which Theodor Benfey's (1809–81) monumental *Geschichte der Sprachwissenschaft* of 1869 may be regarded as the most outstanding example. It had been preceded by Heymann Steinthal's (1823–99) work of 1863, which sought to supersede Laurenz Lersch's (1811–49) three-volume *Die Sprachphilosophie der Alten* (1838–41), but which dealt only with the contributions of Greece and Rome to linguistic thought. Benfey's history of linguistics was followed by other influential works such as Raumer (1870), Delbrück (1882[1880]), Bursian (1883), which, however, were more limited in scope. The same could be said of books by scholars such as Thomsen (1902; German transl., 1927), Delbrück (1904), Trabalza (1908), Jellinek (1913), Pedersen (1916; English transl, 1983), Pedersen (1924; English transl., 1931), Dràganu (1945; Italian transl., 1970), or Robins (1951) from the first half of the twentieth century. (For detailed descriptions of these books, see Koerner, 1978a.)

New endeavors and, at times, more insightful studies in HoL appeared in the 1960s, perhaps beginning with Paul Diderichsen's (1905–64) work on his compatriot Raskof 1960 (German transl., 1976). It was followed by works such as Ivić (1963; English transl., 1965), Leroy (1963; English transl., 1967), Tagliavini (1963), Malmberg (1964; French transl., 1959), Lepschy (1965; English transl., 1970), Mounin (1967), Robins (1967), Coseriu (1969, 1972 see now Coseriu, 2003), Helbig (1970), Szemerényi (1971), and other thematically restricted books (cf Koerner 1978a for details, Jankowsky (1972), and others (see Koerner, 1978a for a full

list). Yet most of them relied uncritically on earlier accounts and rarely ventured into questions of historiographical method or touched upon matters concerning the philosophy of science, except perhaps for a fashionable nod to Kuhn's *Structure of Scientific Revolutions* of 1962.

Looking back on the past 140 or so years of history writing in linguistics, it is possible to discern three distinct types, each associated with particular motives for engaging in such activity and each occurring at specific periods in the development of the discipline. A fourth type (argued for in, e.g., Koerner, 1976) has begun developing only in more recent years. Detailed description of each of these follows.

1. First, there is the type of history written at a time when a particular generation or an individual representing the ideas, beliefs, and commitments of his or her generation to a significant extent is convinced that a desired goal has been reached and that subsequent work in the field will mainly be concerned with what Thomas S. Kuhn has referred to as 'mopping-up operations.' These accounts assumed that the theoretical framework had been sufficiently mapped out for the ordinary member of the scientific community to conduct his or her investigations and that there was no longer any need for a revision of the methodology or the approach to the subject matter under analysis; they were summing up histories that viewed the evolution of the field as growing in a more or less unilinear fashion.

This idea of or motive for writing such a history seems to be best expressed in Benfey's voluminous *Geschichte der Sprachwissenschaft und orientalischen Philologie* (1869), appearing one year after Schleicher's untimely death, and also by Raumer's *Geschichte der germanischen Philologie* (1870). It is difficult today to recreate the atmosphere of the late 1860s even if we limit ourselves to linguistic matters abstracting from external, for example, socio-political, currents: histories available today supply us with little, if any, information on this pre-neogrammarian period. Suffice it to recall for the present purpose that the works of Bopp, Rask, Grimm, and others had been significantly synthesized and methodologically developed by the generation of Georg Curtius (1820–85) and, especially, August Schleicher (1823–68), to the extent that one might speak of a 'paradigm' change having taken place at that time, of which the neogrammarian tenets of historical linguistic research associated with the names of their former students, notably Karl Brugmann (1849–1919) and August Leskien (1840–1916), were the logical, if somewhat overstated, outcome.

A similar observation about the motives of history writing, it would seem, could be made about Holger Pedersen's (1867–1953) 1924 history of the achievements of nineteenth-century Indo-Europeanists, which was preceded by a similar and somewhat shorter account of his first published work in 1916, the year of the completion of the second edition of Brugmann and Delbrück's *Grundriss*, as well as the appearance of Saussure's *Cours de linguistique générale*. The feeling of the need of such a summing-up history is expressed more clearly, perhaps, in Wilhelm Streitberg's (1865–1925) voluminous undertaking, entitled *Geschichte der indogermanischen Sprachwissenschaft seit ihrer Begründung durch Franz Bopp*. However, this is in fact more of a résumé of the work accomplished in the various branches of Indo-European philology by that time than a regular HoL (cf Koerner, 1978a: 16–17, for details). Begun in 1916, the enterprise was abandoned after 1936.

After World War II, it appears that the histories by Malmberg (1964), Ivić (1965), Leroy (1963), and others fulfilled a similar function of summing up previous attainments in linguistic science. This time the focus of attention was the post-1916 period in the history of linguistics, following the success story of Saussure's *Cours*, with its perceived emphasis on a non-historical approach to language. The neogrammarian framework of linguistic research was propounded in the histories of Pedersen, the organizational efforts of Streitberg from 1916 onwards, and other less influential books – and one may add that Pedersen, a second-generation Neogrammarian, reflects the alleged 'data-orientation' of that school much more emphatically than the original group of scholars (note that neither Delbrück's *Einleitung* nor Paul's *Prinzipien* are even mentioned in his 300-page study of 1924). Only in recent years (Jankowsky, 1972; Einhauser, 1990) have the Neogrammarians received a more adequate treatment. By the same token, the histories by Malmberg, Ivić, Leroy, and other similarly slanted studies of the 1960s put forward particular post-Saussurean trends as the most significant achievements of the discipline to date, whether Copenhagen-type, Praguean, or Bloomfieldian. Their endeavor, like that of Benfey, Raumer, Pedersen, Streitberg, and others from an earlier stage in the development of linguistics, was to a large extent the presentation of a framework of research in which they themselves had been brought up, and, concomitantly, an attempt to maintain the strength and impact of the structuralist mode of thought.

2. The second type of history-writing activity may be characterized by the intention on the part of an individual usually in his thirties (not late forties or above, as is generally the case with the first type), again representing a particular group, to launch a campaign opposing previously cherished views and still prevailing doctrines. Thus in contrast to Benfey (1869), for example, Berthold Delbrück's (1842–1922) 1880 *Einleitung* served, together with Paul's *Prinzipien* of the same year, as the mouthpiece of a new generation of scholars eager to demonstrate that their achievements significantly surpassed previous attainments in the field, and that their theories rightfully replaced those taught by the preceding generation of linguists. The claim in favor of discontinuity is what characterizes this type of activity, and Delbrück's book is a prime example of this endeavor. Typically, Schleicher was depicted by Delbrück as representing the conclusion of the phase of comparative-historical grammar inaugurated by Franz Bopp in 1816, and the *Junggrammatiker*, with whom he associated himself early in his career (soon after he had received Schleicher's chair at the University of Jena in 1873), as marking a decisive new turn in the field.

No comparable history of linguistics was written in the 1930s and 1940s with regard to structuralism, but a look into Bloomfield's *Language* (1933) or Gray's *Foundations of Language* (1939) clearly suggests that the chapters devoted to HoL were an attempt to redress the development of the

discipline and to document the superiority of the structuralist approach to any other theory or method hitherto put forward. This endeavor to prove earlier views to be utterly insufficient and inadequate has by no means been abandoned by adherents of the prevailing modes of linguistic thinking of today. On the contrary, it can easily be shown that their advocates have been eager not only to revive an interest in the history of linguistics itself but also to rewrite it to an extent that the ideas of the generation immediately preceding the present one appear the least worthy of attention. As a matter of fact, what C.F. Voegelin felicitously termed the 'eclipsing stance' that transformational-generative grammar had embarked on was best illustrated by Noam Chomsky himself, for instance in his 1962 lecture at the Ninth International Congress of Linguists held in Cambridge, MA (Chomsky, 1964). Soon thereafter, a number of his followers ardently engaged themselves in writing their particular view of history; compare the articles by Dingwall (1963), Bach (1965), or Bierwisch (1966). More recently, Frederick Newmeyer has published a book that constitutes the best example to date of this *pro domo*, Whiggish type of history writing. It selects and reinterprets past linguistic research in an attempt to prove his view that linguistics was made a science only in 1955 or in 1957, and by Chomsky, and that previous work was totally inadequate, barring a few minor incidental insights foreshadowing the 'revolution' in the field (cf Koerner, 1983; for a critical assessment of this kind of activity). Newmeyer's *Linguistics in America* (1980) appeared exactly 100 years after Delbrück's *Einleitung*, and the parallels between their authors are striking indeed: both were less than 40 when they wrote their books; both were primarily interested in syntax, not phonology, and neither had done his doctorate at the respective centers of the schools whose success story they depicted.

While Type I history writing may appear more benign as it seems to represent matter-of-fact accounts (although one should not be too sure about this), Type II HoL can best be described as propagandistic in nature; the most successful example of this type is Chomsky's 1966 *Cartesian Linguistics*. This book presents the author's views regarding the ancestry of his own theories so brilliantly that many a young students of language were carried away by this new vision of history. Today, we still find a lot of useful information in the Type I histories of the field despite their biases and shortcomings (see, e.g., Hoenigswald, 1986; Koerner, 1989); Type II histories, in contrast, although written much more recently, are already dated. It appears that, once their propagandistic purpose has been satisfied in proselytizing for the new ideology, the remainder rapidly loses its initial interest and apparent informational worth.

3. There is a third type of HoL which is intended neither to advocate a particular framework or 'paradigm' nor to attempt to provide an argument in favor of a scientific revolution within the discipline. This type may occur at any time in the development of a particular field of research since its ultimate intent is less partisan than in the other two instances and often more holistic in attitude, although the motivation behind the work may be fairly personal. To my mind, the best example is Arens' 1955 *Sprachwissenschaft*, in which an attempt was made to delineate the development of western linguistic thought from the early discussions of the Greeks about the nature of language to contemporary linguistic work, certainly with a view to indicating not only that our discipline has come a long way to gain those insights we now cherish and the methods we have developed, but also that we all have built, knowingly or not, on the findings of previous generations of linguists, and that we owe much more to these scholars than we might ever become fully aware of.

While this Type III manner of presenting the HoL might well have been the result of a very personal choice, it appears that it expressed the endeavor of a whole generation of scholars, namely, the rebuilding of a discipline after its almost total destruction through a world war. Taken in this way, Pedersen's 1931 book may well be included in the third category in that it sought to reestablish a linguistic tradition which in his belief could continue to serve as a sound basis for subsequent work in the field.

Undoubtedly, other, often nonlinguistic, motives played a role in presenting the history of the discipline in one way or another. Thus, it should be recalled that particular socioeconomic conditions, historical events, or political situations have often had a considerable influence on the motivation of writing the history of a particular discipline or the acceptance of a seemingly new theoretical framework of research or mode of thought – and in this respect histories of linguistics have failed to increase our awareness of the impact of matters or events outside the field. The works of Benfey and Raumer, for instance, were highly motivated by the rise of German nationalism and the aspiration to national unity if not superiority. Similarly, Malkiel (1968: 557) rightly observed that the success of Saussure's *Cours* "cannot be properly measured without some allowance for the feelings of that time: The acceptance of the leadership of a French-Swiss genius connoted for many Westerners then opposed to Germany a strongly desired, rationalized escape from the world of Brugmann, Leskien, Osthoff, and Paul."

4. Despite the respect scholars may have for works of the third type, in particular Arens' *Problemgeschichte*, some have felt a need for yet a fourth type of history writing (e.g., Koerner, 1976; Simone, 1975), namely the presentation of our linguistic past as an integral part of the discipline itself and, at the same time, as an activity founded on well-defined principles that can rival, in terms of soundness of method and rigor of application, those of linguistics itself. This fourth type, now usually referred to as 'linguistic historiography' or, more properly, the *historiography of linguistics* holds that the HoL should not merely be subservient to the discipline, but should assume a function comparable to that of the history of science for the natural scientist. In short, while recognizing the important distinction between chronicle and history which Benedetto Croce is credited with, recent contributors to HoL have gone a step further by distinguishing between history and historiography. This is partly an attempt to make clear a departure from previous endeavors in the field, which only too often tended to be partisan histories, if not what Henry Butterfield termed 'Whig-histories,' and partly because previous histories did

not provide a usable guide for the adequate treatment of past developments in the history of the language sciences, and thus failed to provide a better understanding of where current theories may lead us to.

Another indication of the History of Linguistics having become a mature field of scholarly endeavor is its professionalization. In 1978, the first International Conference on the History of the Language Sciences (ICHoLS) was held in Ottawa, Canada, the same year in which the Société d'Histoire et d'Épistémologie des Sciences du Langage (S.H.E.S.L.) was founded in Paris. In 1984, the Henry Sweet Society for the History of Linguistic Ideas (HSS) was established in Oxford, and several similar international and regional societies have been launched since; for instance, the North American Association for the History of the Language Sciences (NAAHoLS) which was launched late in 1987. In the meantime further ICHoLS meetings have been held in 1981 (Lille, France), 1984 (Princeton, N.J.), 1987 (Trier, Germany), 1990 (Galway, Ireland), 1993 (Washington, D.C.), 1996 (Oxford), 1999 (Paris), 2001 (São Paulo, Brazil), 2004 (Urbana Champaign, USA), 2008 (Potsdam, Gemany), and 2011 (St. Petersburg, Russia), with the 13th such International Conference scheduled to take place in 2014 (Vila Real, Portugal). In addition to *Historiographia Linguistica*, a second journal with much less focused goals, entitled 'Histoire – Épistémologie – Langage,' was established in Paris in 1979, and more recently in 1991, a third journal, *Beiträge zur Geschichte der Sprachwissenschaft*, was launched in Münster, Germany. Subsequently, scholarly societies devoted to HoL were founded in a number of other countries, notably the Sociedad Española de Historiographia Linguistica of Madrid.

Although a couple of scholars were courageous enough to attempt an account of the development of linguistics from the Ancients to modern times (Swiggers, 1997; Allan, 2009), by the mid-1990s it became obvious that the history of linguistics – and not only the Western tradition – could no longer be adequately treated by individual scholars, but that international collaboration was required. This was recognized by Lepschy (1990–98), which treated only selected periods, with the concluding volume covering nineteenth-century historical-comparative philology. Selective accounts of HoL are increasingly focused on specific linguistic subdisciplines, for example, semantics (Gordon, 1982; Nerlich, 1992), pragmatics (Nerlich and Clarke, 1996), and syntax (Forsgren, 1992), and on specific theoretical stances (e.g., Matthews, 1993, 2001). The first more successful attempt to account for the diversity of linguistic traditions and research interests was Koerner and Asher (1995) culminating in a three-tome undertaking of close to 3000 pages (Auroux et al., 2000–2006). In short, if the professionalization of the subject is any guide, the History of Linguistics has become a widely recognized and practiced field of scholarly research (cf also Allan, 2013).

See also: Chomsky, Noam (1928–); Databases and Statistical Systems: Linguistics; Historical Linguistics: Numerical Methods; Historical Thought and Historiography: Current Trends; Logic and Linguistics; Saussure, Ferdinand de (1857–1913).

Bibliography

Allan, Keith, 2009[2007]. The Western Classical Tradition in Linguistics, second ed. Continuum, London.

Allan, Keith (Ed.), 2013. The Oxford Handbook of the History of Linguistics. Oxford University Press, Oxford.

Andresen, Julie T., 1978. François thurot and the first history of grammar. Historiographia Linguistica 5, 45–57.

Arens, Hans, 1955. Sprachwissenschaft: Der Gang ihrer Entwicklung von der Antike bis zur Gegenwart. K. Alber, Freiburg and Munich (second english ed., 1969).

Auroux, Sylvain, Koerner, E.F.K., Niederehe, Hans-J., Versteegh, Kees (Eds.), 2000–2006. History of the Language Sciences: An International Handbook on the Evolution of the Study of Language from the Beginnings to the Present/Geschichte der Sprachwissenschaften: Ein internationales Handbuch zur Entwicklung der Sprachforschung von den Anfängen bis zur Gegenwart/Histoire des Sciences du Langage: Manuel international d'histoire des études linguistiques des origines à nos jours, 3 vols. Walter de Gruyter & Co, Berlin and New York.

Bach, Emmon, 1965. Structural linguistics and the philosophy of science. Diogenes 51, 111–128.

Benfey, Theodor, 1869. Geschichte der Sprachwissenschaft und orientalischen Philologie in Deutschland seit dem Anfange des 19. Jahrhunderts, mit einem Rückblick auf die früheren Zeiten. J.G. Cotta, Munich (Repr. Johnson, New York, 1965).

Bierwisch, Manfred, 1966. Strukturalismus: Geschichte, Probleme und Methoden. Kursbuch 5, 77–152 (English Translation: Modern Linguistics: Its Development, Methods and Problems. Mouton, The Hague 1971.).

Bursian, Conrad, 1883. Geschichte der classischen Philologie in Deutschland. Munchen und Leipzig.

Chomsky, Noam, 1964. Current Issues in Linguistic Theory. Mouton, The Hague.

Chomsky, Noam, 1966. Cartesian Linguistics. A Chapter in the History of Rationalist Thought. Harper & Row, New York and London.

Coseriu, Eugenio, 2003[1969, 1972]. Geschichte der Sprachphilosophie: Von den Anfängen bis Rousseau. Revised and enlarged by Jörn Albrecht. A. Francke, Tübingen and Basel.

Delbrück, Berthold, 1880. Einleitung in das Sprachstudium: Ein Beitrag zur Methodik der vergleichenden Sprachforschung. Breitkopf & Härtel, Leipzig (second ed., 1884; third ed., 1893).

Delbrück, Berthold, 1882. Introduction to the Study of Language: A Critical Survey of the History and Methods of Comparative Philology of Indo-European Languages (Transl. into English by Eva Channing). Breitkopf & Härtel, Leipzig (New ed., with a Foreword by E.F.K. Koerner, John Benjamins, Amsterdam, 1974; 2nd printing, 1989).

Delbrück, Berthold, 1904. Einleitung in das Studium der Indogermanischen Sprachen. Druck und Verlag von Breitkopf & Hartel, Leipzig.

Dingwall, William Orr, 1963. Transformational grammar: form and theory. A contribution to the history of linguistics. Lingua 12, 233–275.

Einhauser, Eveline, 1990. Die Junggrammatiker: Ein Problem für die Sprachwissenschaftsgeschichtsschreibung. WVT Wissenschaftlicher Verlag, Trier.

Forsgren, Kjell-Åke, 1992. Satz, Satzarten, Satzglieder: Zur Gestaltung der deutschen traditionellen Grammatik von Karl Ferdinand Becker bis Konrad Duden. Nodus, Münster.

Gordon, W. Terrence, 1982. A History of Semantics. John Benjamins, Amsterdam.

Helbig, Gerhard, 1970. Geschichte der neueren Sprachwissenschaft: unter dem besonderen Aspekt der Grammatik-Theorie. VEB Bibliographisches Institut, Leipzig; Max Hueber, Munich.

Hoenigswald, Henry M., 1986. Nineteenth-century linguistics on itself. In: Bynon, Theodora, Palmer, Frank R. (Eds.), Studies in the History of Western Linguistics. Cambridge University Press, Cambridge, pp. 172–188.

Ivić, Milka, 1965[1959]. Trends in Linguistics (Muriel Heppell, Trans.). Mouton, The Hague.

Jellinek, Max Hermann, 1913. Geschichte der Neuhochdeutschen Grammatik von den Anfangen bis auf Adelung, Volumes I and 2. Carl Winter, Heidelberg.

Jankowsky, Kurt R., 1972. The Neogrammarians: A Re-evaluation of Their Place in the Development of Linguistic Science. Mouton, The Hague.

Koerner, E[rnst] F[rideryk] K[onrad], 1973. Ferdinand de Saussure: Origin and Development of His Linguistic Thought in Western Studies of Language. A Contribution to the History and Theory of Linguistics. Friedrich Vieweg & Sohn, Braunschweig.

Koerner, E.F.K., 1976. Towards a Historiography of Linguistics: 19th and 20th century Paradigms. Parret 685–718.

Koerner, E.F.K., 1978a. Western Histories of Linguistic Thought: An Annotated Chronological Bibliography 1822–1976. John Benjamins, Amsterdam.

Koerner, E.F.K., 1978b. Toward a Historiography of Linguistics: Selected Essays. Foreword by R.H. Robins. John Benjamins, Amsterdam.

Koerner, E.F.K., 1983. The chomskyan 'Revolution' and its historiography: a few critical remarks. Language & Communication 3, 147–169 (Rev. and extended version in Koerner 1989. 101–146.).
Koerner, E.F.K., 1989. Practicing Linguistic Historiography: Selected Essays. John Benjamins, Amsterdam and Philadelphia.
Koerner, E.F.K., 1995. Professing Linguistic Historiography. John Benjamins, Amsterdam and Philadelphia.
Koerner, E.F.K., 1999. Linguistic Historiography: Projects & Prospects. John Benjamins, Amsterdam and Philadelphia.
Koerner, E.F.K., 2002. Toward a History of American Linguistics. Routledge, London and New York.
Koerner, E.F.K., 2004. Essays in the History of Linguistics. John Benjamins, Amsterdam and Philadelphia.
Koerner, E.F.K., 2006. Historiography of linguistics. In: Brown, Keith (Ed.), Encyclopedia of Language and Linguistics, second ed. Elsevier, Oxford, pp. 332–334.
Koerner, E.F.K., Asher, R.E. (Eds.), 1995. Concise History of the Language Sciences: From the Sumerians to the Cognitivists. Pergamon Press, Oxford and New York.
Kuhn, Thomas S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago (second corrected and extended ed., 1970).
Lepschy, Giulio C., 1965. Problems of semantics. Linguistics 3 (15), 40–65.
Lepschy, Giulio (Ed.), 1994–1998. History of Linguistics, 4 vols. Longman, London and New York.
Leroy, Maurice, 1963. Les grands courants de la linguistique moderne. Presses Universitaires de Bruxelles, Brussels (second english ed., 1971).
Malkiel, Yakov, 1968. History and histories of linguistics. Romance Philology 22, 530–566.
Malmberg, Bertil, 1964[1959]. New Trends in Linguistics: An Orientation. Naturmetodens Språkinstitut, Stockholm and Lund.
Malmberg, Bertil, 1990. Histoire de la linguistique: Sumer à Saussure. Presses Universitaires de France, Paris.
Matthews, Peter H., 1993. Grammatical Theory in the United States from Bloomfield to Chomsky. Cambridge University Press, Cambridge.
Matthews, Peter H., 2001. A Short History of Structural Linguistics. Cambridge University Press, Cambridge.
Nerlich, Brigitte, 1992. Semantic Theories in Europe, 1830–1930: From Etymology to Contextuality. John Benjamins, Amsterdam and Philadelphia.
Nerlich, Brigette, Clarke, David D., 1999. Language, Action and Context. John Benjamins Publishing, Amsterdam.
Newmeyer, Frederick J., 1980. Linguistics in America: The First Quarter-Century of Transformational-Generative Grammar. Academic Press, New York (second revision ed., 1986).
Parret, Herman (Ed.), 1976. History of Linguistic Thought and Contemporary Linguistics. Walter de Gruyter, Berlin and New York.
Paul, Hermann, 1880. Prinzipien der Sprachgeschichte. Max Niemeyer, Halle/S. (fifth ed., 1920).
Pedersen, Holger, 1931[1924]. Linguistic Science in the Nineteenth Century. Translation into English by John Webster Spargo. Harvard University Press, Cambridge, MA.
Pedersen, Holger, 1983[1916]. A Glance at the History of Linguistics, with Particular Regard to the Historical Study of Phonology. Transl. from the Danish by Caroline C. Henriksen, ed. with an introd. by Konrad Koerner. John Benjamins, Amsterdam and Philadelphia.
Raumer, Rudolf von, 1870. Geschichte der germanischen Philologie. J.G. Cotta, Munich (Repr., New York: Johnson, 1965.).
Robins, Robert Henry, 1967. A Short History of Linguistics. Longmans, London. Indiana University Press, Bloomington, 1968, (fourth ed., 1996).
Sebeok, Thomas A. (Ed.) 1975. Current Issues in Linguistics. Vol. XIII: Historiography of Linguistics, 2 vols. Mouton, The Hague.
Simone, Raffaele, 1975. Théorie et histoire de la linguistique. Historiographia Linguistica 2, 353–378.
Streitberg, Wilhelm (Ed.), 1916–36. Geschichte der indogermanischen Sprachwissenschaft seit ihrer Begründung durch Franz Bopp, 6 vols. Karl J. Trübner, Strassburg (later on Walter de Gruyter, Berlin).
Swiggers, Pierre, 1997. Histoire de la pensée linguistique: analyse du langage et réflexion linguistique dans la culture occidentale, de l'Antiquité au XIXe siècle. Presses Universitaires de France, Paris.
Szemerényi, John, Louis, 1971. Richtungen der modernen Sprachwissenschaft. Carl Winter, Heidelberg.
Tagliavini, Carlo, 1963. Panorama di storia delia linguistica. Pàtron, Bologna.
Thomsen, Vilhelm Ludvig Peter, 1902. Sprogvidenskabens historie. G.E.C. Gad, Copenhagen.
Trabalza, Ciro, 1908. Storia della Grammatica Italiana. Ulrico Hoepli, Milano.

Linguistics: Incorporation

Elke Nowak, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Incorporation is a phenomenon whereby what might be expected to appear as a separate word in a sentence in fact shows up as part of the predicate: when incorporated, a noun becomes part of the incorporating verb and specifies it. The special nature of incorporation has been known to linguists for some time. Although it is known that the phenomenon is not restricted to indigenous American languages, it was first and foremost with respect to these languages that incorporation was, and still is, being discussed. In the following, a three-way distinction between noun incorporation, argument incorporation, and grammatical incorporation as a continuum of morphological complexity is introduced.

Introduction

Incorporation is a phenomenon whereby what might be expected to appear as a separate word in a sentence in fact shows up as part of the predicate: when incorporated, a noun becomes part of the incorporating verb and specifies it. The special nature of incorporation has been known to linguists for some time. Although it is known that the phenomenon is not restricted to indigenous American languages, it was first and foremost with respect to these languages that incorporation was, and still is, being discussed. In the following, a three-way distinction between noun incorporation, argument incorporation, and grammatical incorporation as a continuum of morphological complexity is introduced.

Traditionally, noun incorporation is understood as a choice of whether the direct object argument of a verb is retained on the syntactic level or not. In addition to it, argument incorporation refers to the realization of the verb's arguments, i.e., subject and direct object, exclusively on the morphological level, i.e., verb internally. In that case, no arguments can be identified in syntax, since the basic grammatical relations exclusively materialize on the word level. Going a step further, functional items may be incorporated as well. In languages which exhibit grammatical incorporation, various other kinds of grammatical relations, e.g., passive, causative, relative and contact clauses, complementation, adverbial and prepositional information, are realized verb internally, i.e., syntax is largely replaced by morphological complexity. The term 'incorporation' will be used for all three types, regardless of whether the involved items are independent lexical units, i.e., words, or bound morphemes. Such a distinction becomes instrumental for differentiating the different types of incorporation.

Noun Incorporation

Since Kroeber (1909, p. 569) and Sapir (1989/1911, p. 254), noun incorporation has been characterized as stem combination, i.e., the combination of independent lexical items into a complex verb. The most basic manifestation of incorporation brings about the detransitivization of a transitive expression: the direct object argument of the verb is incorporated, as shown in ex. (1a) and (1b). The verb stem determines the newly built complex word categorically, while the incorporated noun loses its categorical features, as well as any grammatical markings such as case or plural, see ex. (3b) below. Nouns do not incorporate as syntactic arguments, but as stems, specifying the incorporating verb.

(1) (a) Mohawk

/wa?kyvtho?	oji:ja?/	'I planted a flower'
wa?	aorist modal prefix	
-k-	1st person singular agent	
- yvtho?	'plant	

(b) Mohawk

/wa?kji?jayvtho?/	'I flower-planted'
-ji?ja:	'flower'

(Bonvillain, 1972, p. 21)

Although the most general features of the construction might at first suggest so, noun incorporation is more than just another variety of compounding. Conspicuously, languages featuring regular and productive compounding, such as German, frequently do not allow for incorporation (Nowak, 2011, pp. 233–235). Contrariwise, incorporating languages do not necessarily allow for compounding.

In many languages, noun incorporation is not a fully productive process, but restricted to certain types of nouns, most often related to the degree of animacy, or the language provides special sets of suppletive forms for incorporation. See, e.g., Allen et al., 1984 on Southern Tiwa; Mithun, 1984, pp. 863, 876, 877; 2001, pp. 44–47.

With respect to discourse and information structure, the incorporating process often entails a loss of referentiality of the incorporated noun and its being backgrounded as aforementioned information. From the perspective of the speaker, incorporation represents an option, since it offers an alternative way of expression, a strategy to decrease the informational content of a proposition or to bring about a focus shift when proceeding in discourse (cf Mithun, 1984, pp. 869ff. on Mohawk). Thus, the incorporated variant must always be seen in contrast to its non-incorporated counterpart. Being a matter of discourse and style, noun incorporation is expected to be a nonobligatory process. As has been pointed out frequently, it represents an elaborate register and, with bilingual speakers,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.52035-2

characteristically is lost with decreasing fluency and competence (see, e.g., Mithun, 1984, p. 879ff.).

It is just this optionality, allowed for by the lexical independence of the constituents involved, which has motivated the interpretation of incorporation as a syntactic process, as a grammatical function-changing operation (Baker, 1988). Viewed from a syntactic point of view, the incorporated noun vacates its original place in the sentence as direct object. Yet it does not incorporate as argument of the verb, but as bare lexical item. Consequently, incorporation is understood as causing the loss of an argument position in syntax; the transitive expression becomes intransitive.

If at all possible, the incorporation of the external argument, i.e., the subject, typically is restricted to nonagentive noun phrases (Sapir, 1989/1911, pp. 252, 266; Mithun, 1984, p. 853; Gerdts, 1998, p. 87; Cook and Wilhelm, 1998).

(2) (a) Chukchee

nejə-k	ʔalʔal	təlgə-gʔi
hill-LOC	snow.ABS	thaw-3SG.S

(b) Chukchee

nejə-k	ʔələ = lgə-gʔi
hill-LOC	snow = thaw-3SG.S

'the snow thawed on the hill'
(Spencer, 1995, p. 451)

As has been discussed so far, in its weakest form, noun incorporation simply causes detransitivization, as in ex. (1a) and (1b). But many degrees of complexity are to be found. In its most dramatic version, phrasal constituency is dissolved and the constituents of the former noun phrase are distributed inside and outside the complex verb. As can be seen in ex. (3b), the head of the noun phrase, 'men,' is incorporated, while its modifiers remain in their syntactic, i.e., external position. This phenomenon is often referred to as 'external modification' or 'classifying incorporation.'

(3) (a) Southern Tiwa

Wisi	seuanin	bi-mū-ban
Two	man:pl	1sB-See-past

'I saw two men'

(b) Southern Tiwa

Wisi	bi-seuan- mū-ban
Two	1sB-man-see-past

'I saw two men'
(Allen et al., 1984, p. 295)

It must be kept in mind that it is exactly the distinction between syntax and morphology, and the interaction between these two levels, on which the specificity of noun incorporation rests.

Argument Incorporation

It is appropriate to distinguish noun incorporation from what was initially called 'pronominal incorporation' by Sapir. Contrary to what the term might suggest, usually no lexical pronoun is incorporated. The morphemes under consideration most often cannot even be reconstructed as historically independent forms, but are, and always were, bound morphemes, obligatory, and paradigmatically ordered (Sapir, 1989/1911, pp. 250–252). Following Evans (1999), it may also be assumed that their semantic properties are different from those of lexical pronouns, which play such an important part in syntax. The more recent term "pronominal argument" (Jelinek, 1984) draws on the relation to syntax again: the arguments of the verb are fully and exclusively expressed in form of bound morphemes, i.e., by morphological marking. 'Pronominal' refers to the fact that it is the indication of 'person,' often fused with other grammatical features such as number, which is expressed. But in fact, these markings do not replace a noun, as pronouns in European languages do. If noun incorporation offers a choice of expression to the speaker, argument incorporation does not. Since the arguments of the verb are expressed as bound morphemes joint to the verbal root or stem, no lexical arguments are necessary. Most often, a lexical manifestation/specification of the pronominal argument outside the verbal complex is highly marked and extraordinary. The way lexical content of third-person arguments is expressed differs widely from language to language, noun incorporation being one, as can be seen with ex. (4). The expression of grammatical relations is shifted from the syntactic level to the word level, as is illustrated by all examples to come. As opposed to noun incorporation, 'pronominal', i.e., morphological, arguments are not potential syntactic arguments, but represent the fundamental grammatical relations of the proposition. They must not be confused with inflection, which merely reflects the grammatical relations constituted on the syntactic level. In order to decide whether a person-marking morpheme has inflectional status or whether it actually represents the argument(s) of the verb, careful investigation is needed. To decide whether syntactic argument positions are available at all in a particular language, the application of test procedures identifying the status of syntactic constituents is helpful. Such investigations have been carried out for only comparatively few languages, such as the Australian language Warlpiri (Hale, 1983; Jelinek, 1984; Simpson, 1991; see also Austin and Bresnan, 1996), the Iroquoian language Mohawk (Baker, 1996), Salishan languages (Jelinek and Demers, 1994; Jelinek, 1995), the Athabaskan language Navajo (Jelinek, 2000), and Inuktitut (Eskimo) (Nowak, 1996a,b). Discussions of Algonquian languages seem to point in the same direction, but do not yet cover a broad range of relevant tests. Other languages suggest themselves as promising candidates, but have not been scrutinized, as is the case with most of the languages discussed by Mithun (1984, 2001).

The morphological representation of the core grammatical relations is frequently accompanied by fundamental shifts concerning the nature of lexical categories. Many argument incorporating languages feature a wealth of lexically and/or functionally bound morphemes (affixes) that have no independent lexical counterparts, and the proportion of roots and stems as compared to bound morphemes is dramatically changed in favor of the affixes (Nowak, 2005, 2004, pp. 35–61).

In Inuktitut (Eskimo), whole semantic fields denoting activities, states, events, processes, including the most elementary predications 'being' and 'having,' as well as the indication of reported speech and epistemic specifications are expressed exclusively by bound morphemes that are suffixed to the stem they modify or supplement (Nowak, 2004).

(4) Inuktitut
qimmiqaqtunga
qimmiq –qaq- -tunga
dog –have.itr. – 1s.itr 'I have a dog/dogs'

(5) Inuktitut
nirijumajara
niri- -juma- -jara
eat – want – 1s.3s.tr 'I want to eat it'

(6) Woods Cree
sipwîhcîhkâsow
away = by-foot\AI=pretend\AI=3I
'he pretends to leave'
(Starks, 1992, p. 127)

In the Salishan languages of the American Northwest Coast, bound morphemes cover the domain of referential expressions and do so exclusively; so-called 'field suffixes' such as -éné, -né, 'ear' in ex. (7) lack independent counterparts.

(7) Kalispel (Salishan)
na•céné 'he gets an earring'
(Vogt, 1940, p. 53)

(8) Shuswap (Salishan)
s –cy –cyép –ø
STAT-PLUR/REDUP-stand.upright-3.itr
'the trees'
(Kuipers, 1974, p. 100)

Complex predicates are lexicalized as referential expressions, as in ex. (8). The same has been reported of Iroquoian languages (Lounsbury, 1953; Sasse, 1988).

In cases where items with lexical content are incorporated, no matter whether they are free or bound morphemes, typically the incorporating argument and the incorporated lexical items do not exhibit agreement. Yet some cases of such internal agreement relations have been reported. For Southern Tiwa, obligatory incorporation is reported. In this language, person marking on the complex verb also reflects the person, number, and class of the incorporated noun (Allen et al., 1984, pp. 295–298). The same seems to be the case with Bininj Gun-wok, a language of North Australia.

(9) (a) Bininj Gun-wok
ba-ginje-ng gun-ganj
3/31-cook-PP IV-meat
'(S)he cooked (the) meat'
(b) Bininj Gun-wok
ba-ganj-ginje-ng
3/31-meat-Cook-PP
'(S)he cooked (the) meat'
(Evans, 1999, p. 259)

Grammatical Incorporation

The most complex incorporating processes cover major grammatical operations. Syntactic operators such as auxiliaries, lexical pronouns, adpositions, complementizers, and other syntactic function words do most often not exist at the word level at all. Consequently, syntax is greatly reduced. Most important, syntax does not provide expressional alternatives for the speaker.

(10) Warlpiri
pingka-rlipa mata-ma-ninja-kujaku ya-ni
slow -1pis tired-CAUS_INF_PURP_NEG go-NPST
'We'll go slowly so that it (walking a long distance) doesn't tire us out.'
(Simpson, 1991, p. 108)

Lexicalization processes and lexical productivity equally take place on the morphological level. Affixes do not constitute closed classes.

(11) (a) Inuktitut
niuvirvingmuaqtunga
niuvirvik –muaq- -tunga
shop -go.to- -1s.itr
'I am on the way to the shop'

The term *niuvirvik* 'shop' is a lexicalized affix combination, as is *-muaq-* 'go to'. *niuvirvik* is composed of the verb stem *niuviq-* 'acquire' and a nominalizing affix *-vik* 'place where s.th. happens'; *-muaq-* is a highly frequent combination of the (singular) prepositional case *-mut* 'to' and the verbal affix *-aq-* 'go/move.' In Inuktitut, such lexicalizations are frequent and a source of lexical productivity. While some combinations became partly or totally opaque, many lexicalizations are structurally transparent and the affixes involved remain perfectly productive, as is the case with *–mut* and *–aq-*:

(11) (b) Inuktitut
illutinuaqtunga
[illu –tinnu(t)] –aq- -tunga
[house-term.1 pl.poss] –move- -1s.itr
[to our house]
I got to our place/I went home
(Nowak, 2004, p. 84).

Interestingly enough, the noun *illu*, 'house', marked by the prepositional case, fused with possessive marking, is not stripped of its plural *–nut* or the possessive when incorporated by *–aq-*, but the fully inflected adverbial expression is incorporated.

Beyond features such as person, number, and transitivity, incorporating arguments in many instances are fused with other features such as the indication of clause linkage, often referred to as switch reference, or anaphoric coreference, as can be seen in ex. (12).

(12) (a) Inuktitut
aniqumagu aullaqpuq
ani- -qu- -magu aullaq- -puq
go.out – order/ask/tell – 4s.3s.caus.tr leave – 3s.itr
because he (4s) told him(3s) to go out he left (3s)
'he told him to go out and so he left'
(Nowak, 2008, p. 29)

(b) Inuktitut
iqalliaqłuni anuraalikaallaktualuungmat …
iqalliaq- -łuni anuraa(q) –liq- -kallak- -tualuq- -mat
go.fishing -3s.part.itr wind -begin- -suddenly- - mighty- -4s.caus.itr
he going fishing because suddenly a mighty wind begins/began..
'he had gone fishing when suddenly a storm came up…'
(Nowak, 2002, p. 50)

Grammatical incorporation must be seen as the word internal establishment of grammatical relations without syntactic spell out. It results in highly complex forms equivalent not only to full propositions, the establishment of coherent discourse rests on it. Repercussions on information distribution and information structure are inevitable. In Inuktitut, e.g., incorporated lexical items are frequently picked up, referred to, and further elaborated in the course of discourse. Most important, they constitute preferred sites for the introduction new discourse participants/new information (Nowak, 2009).

Conclusion

It must be kept in mind that languages hardly ever represent ideal types. This applies to what has been said about syntax and morphology and the degree to which they interact. The types of incorporation discussed here exemplify an increasing complexity, configurationality, on the level of morphology, illustrating a unidirectional relationship between increasing morphological complexity and the morphological realization of arguments and grammatical structuring. Languages of high morphological complexity also feature the morphological realization of arguments, but it is not necessary that all languages realizing their arguments morphologically must be morphologically complex in other respects.

The specific morphological strategies employed range from concatenative (Eskimo) through template (Athabaskan) to morphophonemic and prosodic devices (Salishan) and often exhibit mixtures of all these types. A unified account of such complex and multilayered morphological patterns is a desideratum that must be left for future research.

While noun incorporation is a matter of choice in discourse, argument incorporation and grammatical incorporation are of a fundamentally different nature. They represent the sole way of expression and cover the full range of structural capacities. The last type, grammatical incorporation, extends morphological configurations to other than argument relations, i.e., to grammatical relations in general. This is the content of the traditional term 'polysynthetic.'

See also: Discourse and Indentity; Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects; Linguistic Turn and Discourse Analysis in History; Morphology in Linguistics; Parts of Speech; Valency and Argument Structure in Syntax; Visual Word Recognition, Neurocognitive Psychology of.

Bibliography

Allen, B., Gardiner, D., Frantz, D., 1984. Noun incorporation in Southern Tiwa. International Journal of American Linguistics 50, 292–311.
Austin, P., Bresnan, J., 1996. Non-configurationality in Australian aboriginal languages. Natural Language and Linguistic Theory 14 (2), 215–268.
Baker, M., 1996. The Polysynthesis Parameter. Oxford University Press, Oxford, UK.
Baker, M., 1988. Incorporation. University of Chicago Press, Chicago.
Bonvillain, N., 1972. Noun incorporation in Mohawk. In: Forster, M. (Ed.), Papers in Linguistics from the 1972 Conference on Iroquoian Research. Mercury Series, Ethnology Division Paper No. 10. National Museum of Man, Ottawa, pp. 18–26.
Cook, E.D., Wilhelm, A., 1998. Noun incorporation. New evidence from Athabaskan. Studies in Language 22 (1), 49–81.
Evans, N., 1999. Why argument affixes in polysynthetic languages are not pronouns: evidence from Bininj Gun-wok. Sprachtypologie und Universalienforschung 52, 255–281.
Gerdts, D., 1998. Incorporation. In: Spencer, A., Zwicky, A. (Eds.), The Handbook of Morphology. Blackwell, Oxford, UK, pp. 84–100.
Hale, K., 1983. Warlpiri and the grammar of non-configurational languages. Natural Language and Linguistic Theory 1, 5–47.
Jelinek, E., 1984. Empty categories, case, and configurationality. Natural Language and Linguistic Theory 2, 39–76.
Jelinek, E., 2000. Navajo as a discourse configurational language. In: Fernald, T., Platero, P. (Eds.), The Athabaskan Languages: Perspectives on a Native American Language Family. Oxford University Press, Oxford, UK, pp. 252–287.
Jelinek, E., 1995. Quantification in Straits Salish. In: Bach, E., Jelinek, E., Kratzer, A., Partee, B.H. (Eds.), Quantification in Natural Languages. Kluwer, Dordrecht, The Netherlands, pp. 487–540.
Jelinek, E., Demers, R., 1994. Predicates and pronominal arguments in Straits Salish. Language 70, 697–736.
Kroeber, A., 1909. Noun incorporation in American languages. In: Heger, F. (Ed.), Internationaler Amerikanisten-Kongress, vol. XVI. Hartleben, Vienna, Austria, pp. 569–576.
Kuipers, A., 1974. The Shuswap Language. Mouton, The Hague.
Lounsbury, F., 1953. Oneida Verb Morphology. Yale University Press, London.
Mithun, M., 2001. The Languages of Native North America. Cambridge University Press, Cambridge.
Mithun, M., 1984. The evolution of noun incorporation. Language 60, 847–894.
Nowak, E., 2011. Die Redeteile in Inkorporation und Polysynthese. Jahrbuch Deutsch als, Fremdsprache 37, 228–243.
Nowak, E., 2009. Arguments and information management in Inuktitut - Aleut. In: Tersis, N., Mahieu, M.-A. (Eds.), Variations on Polysynthesis: The Eskimo-aleut Languages. Benjamins, Amsterdam, pp. 201–213.
Nowak, E., 2008. Inuktitut. Eine grammatische Skizze. Lincom, München.
Nowak, E., 2005. Lexical categories in polysynthetic languages. In: Cruse, A., Hundsnurcher, F., Job, M., Lutzeier, P.R. (Eds.), Lexicology. An International Handbook on the Nature and Structure of Words and Vocabularies. Mouton de Gruyter, Berlin, pp. 981–986.
Nowak, E., 2004. Inkorporation und Polysynthese. In: Nowak (Ed.), Komplexe Wortstrukturen, Komposition-Inkorporation-Polysynthese. Working Papers in Linguistics, vol. 38. TU Berlin, Berlin, pp. 1–106.
Nowak, E., 2002. Creating discourse in inuktitut. In: Nowak (Ed.), Morphology in Comparison, Working Papers in Linguistics, vol. 37. TU Berlin, Berlin, pp. 38–69.
Nowak, E., 1996a. Transforming the Images. Transitivity and Ergativity in Inuktitut (Eskimo). Mouton de Gruyter, Berlin.
Nowak, E., 1996b. Über die Organisation komplexer Aussagen. In: Harras, G., Bierwisch, M. (Eds.), Wenn die Semantik arbeitet. Festschrift für Klaus Baumgärtner. Niemeyer, Tübingen, Germany, pp. 29–56.
Sasse, H.J., 1988. Der irokesische Sprachtyp. Zeitschrift für Sprachwissenschaft 7.2, 173–213.
Sapir, E., 1989/1911. The problem of noun incorporation in American Languages. In: Bright, W. (Ed.), The Collected Works of Edward Sapir, vol. 5. Mouton de Gruyter, Berlin, pp. 27–60.
Simpson, J., 1991. Warlpiri Morpho-Syntax. A Lexicalist Approach. Kluwer, Dordrecht, The Netherlands.
Spencer, A., 1995. Incorporation in chukchee. Language 71 (3), 439–489.
Starks, D., 1992. Aspects of Woods Cree Syntax (Ph.D. thesis). University of Manitoba, Winnipeg, Canada.
Vogt, H., 1940. The Kalispel Language. Det Norske Videnskaps Akademi i Oslo, Oslo, Norway.

Linguistics: Overview

David Crystal, University of Wales, Bangor, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8948–8954, © 2001, Elsevier Ltd.

Abstract

Linguistics, the scientific study of language, has generated a wide range of branches, subdisciplines, schools of thought, and applications. This article provides a perspective within which these aspects can be interrelated, insofar as they use the word linguistics as part of their identity. It reviews the history of linguistics in the twentieth century, focusing on its claims to scientific status, and providing a classification of the subject in terms of its synchronic/diachronic, theoretical, descriptive, and comparative orientation. A second level of classification is based on the kinds of interaction which have emerged between linguistics and other disciplines, notably in the domains of culture, biology, and numerical sciences. The various branches of applied linguistics are reviewed, and several terminological extensions of the subject identified.

Linguistics, conventionally defined as the scientific study of language, has generated a wide range of branches, subdisciplines, schools of thought, and applications, many of which are given separate treatment in the encyclopedia. This article provides a perspective within which these and other aspects can be interrelated, insofar as they use the word 'linguistics' as part of their identity.

Linguistics as a Science

Linguistics achieved its academic identity during the twentieth century, shaped in the first few decades by a series of pioneering and influential personal statements (notably de Saussure, 1916; Sapir, 1921; Bloomfield, 1933), consolidated in the middle decades as a university subject at both undergraduate and graduate levels, and reinterpreted in the later decades by a new generation of scholars concerned to give the subject a sound theoretical foundation (notably Chomsky, 1957, 1965). In Europe, its origins lay chiefly in a reaction to the philological curiosity about the history and origins of language, which had preoccupied the nineteenth century; in the USA, the primary impetus came from a concern to provide a linguistic dimension to the anthropological investigations of American Indian cultures. But the subject also benefited from an increased general interest in the study of language in relation to human beliefs and behavior (e.g., in philosophy, theology, information theory, literary criticism, communication studies), which manifested itself at both popular and specialized levels. The development of the media, especially broadcasting, further helped to bring language issues to the forefront of national consciousnesses, as did the demands made by foreign-language teachers and other 'language professionals' for more sophisticated information about their subject.

The multiplicity of interests in the subject, each with its own intellectual agenda and methodologies of inquiry, brought a profound appreciation of the range and complexity of linguistic phenomena, and thus of the need to develop a research discipline with its own principles and procedures. Early on, linguistics came to be viewed as a 'science', and attempts were made to define its axioms and make its investigative procedures explicit (e.g., Bloomfield, 1926; Hjelmslev, 1943). Linguists emphasized their concern to make their studies systematic and objective, by contrast with the selective impressionism which had characterized much previous literary and pedagogical work on language. A great deal of attention was devoted to making investigative procedures clear and well defined. There was a significant growth, fostered by progress in acoustic technology, of experimental techniques, especially in phonetics. And there was an unprecedented concern for precise definition, clear and comprehensive description, and more powerful explanation. An emphasis in the first half of the century on the procedures for describing individual languages, or the features of language families, changed during the second half into an emphasis on the universal, defining properties of language, and on the nature of language viewed as part of human biology and psychology. Yet there were significant continuities too. Chomsky's focus on the distinction between linguistic competence and performance, and on the status of language as an 'organ of the mind', echoed Saussure's notions of the collective language system (*langue*) and its concrete manifestation in utterance (*parole*), and of the human language faculty (*faculté de langage*).

Because of its claims to scientific stature, the subject has often been referred to as 'linguistic science', but sometimes a designation of 'linguistic sciences' will be encountered. The choice hinges on the perceived standing of phonetics. In the 'plural' interpretation, phonetics is seen as a separate discipline, which along with linguistics makes up the linguistic sciences. The reasoning is that the subject-matter of phonetics – which includes acoustics, anatomy, physiology, neurology, and auditory perception – is so different from the 'rest' of linguistics that it cannot sensibly be subsumed under the same heading. Moreover, its methodology is very different, both in its experimental approach and in its reliance on technology. It is, some argue accordingly, a prelanguage study, and its findings need to be brought into relationship with those of linguistics. Others, however, point to the fact that the bulk of the subject deals only with those aspects of sound production and processing that are relevant to the production of spoken language.

Phonetics focuses on such entities as vowels, consonants, and pitch movements, and not on such noises as whistles, hiccups, and burps, even though these are also made by the vocal apparatus. Its aim is to explicate speech sound, and as it is speech which is the ground of being for all languages (writing being present only in a proportion of instances), this alternative argument sees phonetics as the 'indispensable foundation' (Henderson, 1971) for linguistics – one branch of a single linguistic science. (The term 'linguistics', of course, avoids the need to make a decision.) A similar issue appears when the focus is on speech alone. Speech science (or speech and hearing science) is the study of all the factors involved in the production, transmission, and reception of speech; but speech sciences will also be found, when there is a need to emphasize the constituent disciplines involved.

Classification by Orientation

These changes of emphasis and direction, which are an inevitable part of the intellectual development of a subject, have given rise to a range of designations for linguistics, each of which captures a particular orientation or 'universe of interest'. The earliest of these, introduced by Saussure, was chronological, distinguishing diachronic linguistics from synchronic linguistics. This terminology reflects his concern to move away from an exclusively historical approach for the study of language to one in which a language is seen as existing at a particular moment (or 'state') in time, 'syn-chronically', regardless of whatever may have happened to it previously or would happen to it subsequently. Most linguistic research is synchronic, in the sense that linguists are aiming to establish the nature of the system of rules which define a (part of a) language's expressive potential at any one time. For example, the word 'balcony' currently has a stress on its first syllable, and this is what would be noted as part of a synchronic description of modern English stress; the fact that, over a century ago, the stress was on the second syllable, would not be considered relevant. On the other hand, in a diachronic description, the way the stress has shifted is of central relevance – not least because the process is continuing in many other words, such as 'dispute' and 'research'. Saussure saw the two chronological orientations as intersecting – an 'axis of simultaneities' crossing an 'axis of successivities' – and this point of intersection must not be ignored. Uncertainties about usage are one example of the way in which ongoing language change complicates a synchronic analysis. But the point of the distinction is widely appreciated, and Saussure's refocusing has been so successful that it has long been standard practice to interpret the word 'linguistics' as meaning 'synchronic linguistics', with references to the other orientation requiring a more explicit phrase – usually, these days, 'historical linguistics'.

Another type of orientation relates to the generality of a linguistic inquiry. As already mentioned, a contemporary focus on language is to make statements of maximal generality: the aim is to determine what it is that defines the notion of 'human language'. All languages have a great deal in common in the way they produce sounds, organize their grammars, and construct systems of meaning in words. Identifying the universal principles which govern language, and demonstrating how these principles apply in individual circumstances to produce the languages known to us as French, English, Chinese, Swahili, etc., is the primary aim of theoretical linguistics. The term 'theory' here has its standard scientific application, reflecting the need for explanation of a maximally general kind, and the provision of fruitful models which will generate testable hypotheses about, in this case, linguistic structure and use. Theoretical linguists do not study a language because they are interested in that language per se; they study it because of what it can tell them about the nature of language in general. A particular language may be especially 'interesting' because it makes use of a type of sound or grammatical construction that other languages do not. The label 'general linguistics' is often used to capture this breadth of vision. A further consequence of this approach is that, once some linguistic data has been accumulated, a special concern is to find the best way of analyzing it. There are so many variables in pronunciation, grammar, and vocabulary that it is possible to envision an indefinitely large number of ways of showing how these domains are organized. Theoretical linguists spend much of their time searching for plausible and elegant analyses, and identifying criteria for their evaluation.

One of the controversies in the profession relates to this point. According to some linguists, as a result of the above orientation, the subject has become 'too general'. Analyses which try to achieve maximal generality inevitably make statements which are highly abstract, and seem to be far removed from the description of individual languages, and of the way these are used in everyday speech and writing. The emphasis on achieving generality, according to this view, has brought a de-emphasis on individuality – of what it is that makes one language unique, different from all the others. While it is accepted that, in the long term, all the idiosyncrasies of individual languages may be explicable with reference to general rules, there is seen to be no likelihood of this happening in the foreseeable future. In the meantime, the need to provide descriptions which are as full as possible is as important now as it ever was. Indeed, this focus on the special character of a particular language was given special urgency in the 1990s, when it was realized just how many of the world's languages are near the point of extinction. Of the 6000 or so languages in the world, it is widely thought that half are likely to die out within the next 100 years (Crystal, 2000). The need to make descriptions of the languages most at risk is therefore a priority for the subject, for once a language dies which has never been written down or recorded in any way, it is as if it had never been. This orientation for the subject, where the aim is to make as complete as possible a description of a language, in all its varieties, is usually referred to as descriptive linguistics.

It is not only the endangered languages which benefit from a descriptive approach, therefore. Descriptive accounts of English, German, Japanese, and all the other 'safe' languages are also important. 'Descriptive' here is being seen in contrast with 'traditional', referring to the tradition of language inquiry which antedated the emergence of twentieth – century linguistics, and which was characterized – especially in the field of grammar – by prescriptive and proscriptive rules about usage, purist condemnation of language change, and the distorted

description of modern languages through the use of a model of description derived from Latin. In the field of English, the need for more realistic description was the driving force behind many grammars, pronunciation guides, and dictionaries, especially from the 1950s on (e.g., Jones, 1956; Fries, 1952; Quirk et al., 1972). A particular consequence of this orientation was the provision, in some countries, of new syllabuses and materials for use in schools and teacher-training courses – a change of direction which in the later decades of the century brought not a little heart-searching, as teachers found themselves engaging in a close encounter with a linguistically based conceptual apparatus and a partly unfamiliar terminology. At a research level, since the 1980s the descriptive orientation of linguistics has been given additional impetus through the use of computational methods of data collection and analysis. Large bodies of data, in some cases consisting of several 100 million words, are now available as sources of evidence for descriptive statement, in a field that has come to be called corpus linguistics.

A considerable conceptual distance separates the description of an individual language from a general account of the properties of human language, so it is not surprising to find orientations which occupy aspects of the 'middle ground'. It is possible to extend the descriptive orientation 'upwards', bringing individual languages together in specific ways but without the inquiry necessarily bearing on a hypothesis about linguistic universals. This is the motivation behind comparative linguistics, where the purpose is to identify the chief similarities and differences between historical states of a language, between modern languages (e.g., those within a particular family, such as Romance or Germanic) or, in a more ambitious mode, between language families as wholes. This use of 'comparative' in fact antedates modern linguistics, being part of the designation of comparative philology, the subject which grew up at the end of the eighteenth century devoted to the historical comparison of Greek, Latin, Sanskrit, and their hypothetical antecedents, whence arose the concept of an Indo-European family of languages.

Applying the 'comparative method' to other languages which had written records was a natural extension, and philology (as it is today usually called) is practiced still, being distinguished from the orientation of historical linguistics partly by its aims and methods, and partly by the contrasting intellectual background (often from literature or history) of its practitioners. A related designation is typological linguistics, where the aim is to study the structural similarities between languages, regardless of their history, in order to locate them within a classification (or 'typology'). This is the only possible method in cases where there are no written historical records – a situation in which most of the world's languages find themselves.

When linguists talk about the structure of language, they are usually referring to the way a language can be analyzed into several dimensions susceptible of independent study. These dimensions are commonly referred to as 'levels' or 'components', though terminology varies widely between linguistic models. Most accounts recognize three major levels of linguistic organization: a sound system (or phonology, including its interface with phonetics), a grammatical system (with a division often made between sentence-structure (syntax) and word-structure (morphology)), and a semantic system (typically focusing on vocabulary (the lexicon), though not thereby excluding other ways in which a language's meaning is organized). The study of the written language necessitates the recognition of an additional level: an orthographic system (or graphology). And some linguists recognize other levels of analysis, such as one which extends linguistic inquiry above the sentence (a level of discourse, or text) or one which takes into account the way structure is altered when language is used interactively (pragmatics). Each of these terms has a variety of interpretations, depending on the theoretical approaches involved, but they all have one thing in common: they aim to explicate a domain of language structure, and are thus part of an orientation to the subject that might be legitimately called structural linguistics.

However, this designation has had a controversial position within the history of twentieth-century linguistic thought. In its early use, 'structural' linguistics referred to the particular approaches to phonology and grammar which were current in the 1940 and 1950s, in the work of such linguists as Bloomfield. There, the emphasis was on finding methods which would establish the contrasting units in a language's sound system (the phonemes), identify the elements of word formation (the morphemes), and analyze sentence constructions into their constituent parts. In a somewhat broader sense, developing the initiatives of Saussure, a 'structuralist' ethos prevailed, which aimed to establish systems of relations between linguistic units – an approach which, under the heading of structuralism, came to influence thinking throughout the arts and social sciences. Within linguistics, it was in due course applied to further areas of linguistic structure (especially semantics), and extended by some linguists (such as Pike, 1954–1960) to other domains of human communicative behavior. The notion of system continued to play a central role in linguistic theories which emerged during the postwar period, notably in the work of Firth (1957) and his students, and later became central in the systemic linguistics of Halliday (Halliday and Fawcett, 1987). It was the perceived limitations of the structuralist orientation, with its focus on the classification of structures and units at a 'surface' level of description within a corpus of data (linguistic 'performance'), which prompted its labeling as 'taxonomic' (with a pejorative implication) by Chomsky, and which motivated his initial formulation of generative linguistics (Chomsky, 1957), with its concern to explicate underlying linguistic relations in ways that could account for the creative linguistic abilities (or 'competence') of the individual.

Classification by Interaction

The generative orientation constituted a revolution in linguistic thought; and although other kinds of linguistic theories and models continued to be devised and developed in the latter part of the century, it was the generative approach, in its various formulations, which provided the subject's center of gravity. At the same time, partly independent of this state of affairs and partly as a consequence of it, several areas of interaction with other academic disciplines emerged. These 'hybrid' subjects were the result of a synergy between two intellectual streams.

On the one hand, linguists sensed that they needed to take their subject in fresh directions if they wished to account for everything that influenced the way people spoke and wrote (and also signed, for the analysis of deaf signing was also emerging as a new research field). On the other hand, scholars in other academic disciplines were finding that they needed to provide more sophisticated accounts of language structure and use in order to maintain their own research agendas. The potential for interaction had in some cases been recognized for many years. In the case of the interaction between linguistics and anthropology, it had been there from the earliest days of the subject (Boas, 1911). The difference was partly in the way hybrid labels came to be institutionalized as 'recognized' domains for teaching and research (anthropological linguistics, in this case), and partly in the cross-disciplinary climate of the times, which gave rise to nearly two dozen such hybrids by the turn of the century.

There is not always a total meeting of minds in these interdisciplinary encounters: the orientation given to the hybrid subject within a linguistics department is by no means identical with the orientation it is given within the department which constitutes the other half of the hybrid. For example, one of the areas of inquiry in psycholinguistics is the relationship between language and memory. When a linguist investigates how language is affected by memory and a psychologist investigates how memory is affected by language, it might be thought that the two directions of inquiry would 'meet in the middle'. In practice, differences in the researchers' intellectual histories, theoretical aims, and research procedures can produce widely divergent accounts. One need only look at the list of references at the end of journal articles to see how authors, ostensibly investigating the same subject, can come from different directions. There may be a difference of labeling, which may reflect a difference of orientation – as in sociolinguistics (a branch of linguistics) alongside the sociology of language (a branch of sociology), or philosophical linguistics (a branch of linguistics) alongside linguistic philosophy (a particular approach within philosophy). Nonetheless, the shared subject-matter, and a growing mutual familiarization with the intellectual background of the contributing disciplines, produced in the last quarter of the twentieth century a growing number of research collaborations, interdisciplinary conferences, jointly authored textbooks, and shared teaching courses in many of these hybrid subjects, resulting in a growing identity and coherence.

Hybrid subjects can be placed in clusters, on the basis of an overlap in their subject-matter. The earliest clusters to be given clear definition all relate to a concern to see language in its cultural and social context. Anthropological linguistics (or linguistic anthropology) focuses on language variation and use in relation to human cultural patterns and beliefs (in its early years, especially on the Amerindian peoples of North America). Overlapping with this is ethno – linguistics, studying language in relation to ethnic types and behavior, and widely practiced through approaches (variously called the ethnography of communication or ethnography of speaking) which analyze the entire range of variables, extralinguistic as well as linguistic, involved in social interaction. Sociolinguistics, studying all aspects of the relationship between language and society, is the most well established of this cluster of subjects, and is sometimes used as a covering label for all domains of inquiry in which there is a preoccupation with the social functions of language. However, sociological linguistics is sometimes differentiated from it, particularly in mainland Europe, reflecting a concern to see language as an integral part of sociological theory. So too is interactional sociolinguistics, referring to the study of speech in face-to-face communication.

Several other facets of the sociolinguistic cluster have maintained their individual nomenclature. One example is dialectology (sometimes called dialect geography), which focuses on the properties of regional (and more recently social) dialects. Another is areal linguistics, which studies the linguistic properties (languages as well as dialects) of large geographical regions – such as Western Europe or the British Isles. Closely related is geolinguistics (also geographical linguistics), which studies the geographical distribution of languages throughout the world with reference to their political, economic, and cultural status. Pragmalinguistics studies language use from the viewpoint of a language's structural resources; it contrasts with those pragmatic studies which examine the conditions on language use deriving from the social situation (sometimes referred to as sociopragmatics). Finally, there is the study of (linguistic) style, stylistics – a study which, though chiefly concerned with the nature of the personal linguistic identity of individuals (especially in a literary context), has often been extended to include the distinctive linguistic features of group-identifying functions of language, such as are found in advertising, science, and the media. When the focus is specifically on the distinctive language used in social institutions, such as law, medicine, and religion, some linguists use the label institutional linguistics.

Another cluster of hybrid subjects looks at language in relation to human biology. In its broadest characterization, the subject is referred to as biological linguistics (or biolinguisics), focusing on the preconditions for language development and use from the viewpoints of both the history of language in the human race and the development of language in the individual (the latter topic usually being distinguished by the name developmental linguistics or child language acquisition). Its topics include the genetic transmission of language, neurophysiological models of language production, and the anatomical comparison of humans and other species. It also subsumes the study of pathological forms of language behavior (as in aphasia, dyslexia, and language delay), though the complexity of these conditions, and the interdisciplinary nature of their diagnosis and treatment, has prompted the development of separate specialisms. Chief among these is neurolinguistics (sometimes referred to as neurological linguistics), which focuses on the neurological basis of language development and use, in particular aiming to understand the way the brain controls the processes of speaking, listening, reading, writing, and signing. Although its focus is on clinically normal states, much of its evidence comes from the investigation of abnormal language development and language breakdown. It thus overlaps with clinical linguistics, which is exclusively devoted to the diagnosis, assessment, and treatment of all forms of language disability in children and adults.

Any of these domains can of course be studied from the point of view of 'mind' as distinct from 'body', and

psycholinguistics (sometimes called psychological linguistics) has emerged as a major subject, investigating the interaction between language and such topics as memory, perception, attention, and personality, in normal and abnormal states. Depending on the domain of inquiry involved, compound names will be encountered, such as clinical psycholinguistics and developmental psycholinguistics. Psychology shares with education an interest in learning, and there has been special concern to understand how language is dealt with in schools and other educational institutions. Educational linguistics (sometimes pedagogical linguistics) focuses on the way government policies and planning, teacher training, curricula, syllabuses, and pedagogical materials handle the teaching of the mother tongue and foreign languages. The later decades of the twentieth century also saw a growing metalinguistic concern, with researchers investigating the ways in which children become increasingly aware of language and its elements, in a domain which by the 1990s had been labeled language awareness. Within the teaching profession, a corresponding fresh interest in the study of language, largely motivated in the UK by the arrival of the National Curriculum, introduced a new acronym, KAL (knowledge about language).

A further cluster of hybrids relates to the numerical sciences. Mathematical linguistics studies the formal properties of language, usually employing concepts of an algebraic or statistical kind. Its main application has been in the formalization of linguistic theory, as developed in relation to generative linguistics, but several other areas of language study have received mathematical investigation. In particular, statistical (or quantitative) linguistics studies the frequency and distribution of linguistic units in texts, with the aim of establishing general laws concerning the statistical properties of language (such as the relationship between word types and tokens), and identifying the distinctive characteristics of authors (stylostatistics). In the later part of the twentieth century, computational techniques and concepts came to be increasingly applied to the elucidation of linguistic problems, under the heading of computational linguistics. Several research areas have been developed, including natural language processing, automatic speech synthesis and recognition, machine translation, the making of concordances, and the many areas where statistical counts and analyses are required. The provision of large machine-readable texts (computer corpora) was a special development of the 1990s, and gave rise to the domain of corpus linguistics. Linguistic involvement in the other direction – in relation to such matters as computer programing, information retrieval, document identification, indexing, reference science, and knowledge management – has been limited to isolated projects, and has not yet resulted in a hybrid label, though one imagines that, given the growth in this area, the emergence of Internet linguistics can only be a matter of time.

Not all hybrid terms can be neatly clustered. Some, by their nature, stand outside of classification, because their role is to evaluate aspects of the field as a whole. The study of the history of ideas in linguistics is linguistic historiography. Philosophical linguistics studies the role of language in relation to the understanding and elucidation of philosophical concepts, as well as the philosophical status of linguistic theories, methods, and observations. Critical linguistics emerged in the 1990s with the aim of revealing hidden power relations and ideological processes at work in spoken or written texts. Critical linguists criticize mainstream linguistics for its formalist preoccupations, its lack of adequate social explanations, and its obscuring of ideological and political issues. This 'critical' perspective may be applied to individual branches of the subject. In particular, critical discourse analysis studies the relationship between discourse events and sociopolitical and cultural factors, especially the way discourse is ideologically influenced by, and can itself influence, power relations in society.

Applications and Extensions

The list of possible hybrid subjects involving linguistics is by no means exhausted by the above itemization, and new ones will doubtless emerge. A similar potential for expansion is also found in the various domains of applied linguistics. In its broadest definition, this is the application of linguistic theories, methods, and findings to the elucidation of language problems which have arisen in other areas of experience. The most established and well-developed branch is the teaching and learning of foreign languages, and sometimes the term is used as if this were the only domain involved. But several other domains of application have emerged since the middle of the twentieth century, including speech pathology and therapy, the teaching of the deaf, mother-tongue education, the writing of dictionaries (lexicography), the analysis of literary texts (literary stylistics), and translation studies. Quite restricted and specialized domains may be encountered. For example, theo-linguistics has as its focus the study of the way language is used in theological, biblical, and other branches of religious studies. Forensic linguistics is the use of linguistic techniques to investigate crimes in which language data forms part of the evidence, such as in the use of grammatical or lexical criteria to authenticate police statements. The field of forensic phonetics is often distinguished as a separate domain, dealing with such matters as speaker identification, voice line-ups, speaker profiling, tape enhancement, tape authentication, and the decoding of disputed utterances.

There is an uncertain boundary between applied linguistics and the various interdisciplinary subjects reviewed in Section Classification by Interaction, especially as several of those concerns involve practical outcomes (e.g., planning a national language policy within sociolinguistics). On the other hand, as these hybrid subjects develop their own theoretical foundations, the distinction between 'pure' and 'applied' is becoming more evident, and terminological distinctions such as that between (theoretical) psycholinguistics and applied psycholinguistics are now more regularly encountered.

The term 'linguistics' will also be encountered in a wide range of other contexts, where it is often no more than a loosely used synonym for 'school of thought' or 'approach', or even a stylistic replacement for the word language or grammar. Very general states of mind on the part of the investigator can be given a linguistics designation. For example, mentalist(ic) linguistics is contrasted with behaviorist linguistics. Macrolinguistics, representing a relatively broad frame of reference for linguistic inquiry, is opposed to

microlinguistics. In relation to schools, we find such usages as Cartesian linguistics (showing the influence of Descartes), Chomskyan linguistics (from N. Chomsky), Whorfian linguistics (from B.L. Whorf), Saussurean linguistics (from F. de Saussure), Firthian linguistics (from J.R. Firth), and neo-Firthian linguistics (scholars whose teaching derived from Firth's influence, notably M.A.K. Halliday (whose name has, in turn, been given to Hallidayan linguistics).

At this level, there is an appreciable overlap with the relevant aspect of language which is the focus of an approach – for example, the terms Chomskyan linguistics and Chomskyan grammar are often used synonymously, as are Whorfian linguistics and Whorfian semantics. The same point applies to such general approaches as functional(ist) linguistics, which treats the notion of function as central, especially in grammar, and thus allows a terminological overlap with functional(ist) grammar. Systemic and poly-systemic linguistics (which treat the notion of system as central) may also be found alongside systemic and polysystemic grammar. Virtually any variety or use of language can be focused in this way, as in creole linguistics (= the linguistic study of creole languages) and paralinguistics (= the linguistic study of *para*-language – tones of voice, facial expressions, etc.). But eventually, as increasingly restricted areas of inquiry are encountered, the use of the general term ceases to have a point, and linguists replace it by specific terms such as morphology, phonology, lexicology, semantics, onomastics, and other structural labels best explicated under the heading of 'language'.

See also: Evolution and Language: Overview; Human Language: Evolutionary Precursors; Psycholinguistics: Overview; Saussure, Ferdinand de (1857–1913); Sociolinguistics; Structuralism.

Bibliography

Bloomfield, L., 1926. A set of postulates for the science of language. Language 2, 153–164.
Bloomfield, L., 1933. Language. Holt Rinehart & Winston, New York.
Boas, F. (Ed.), 1911. Handbook of American Indian Languages. Smithsonian Institution, Washington, DC.
Bright, W. (Ed.), 1992. International Encyclopedia of Linguistics. Oxford University Press, Oxford, UK.
Chomsky, A.N., 1957. Syntactic Structures. Mouton, The Hague, The Netherlands.
Chomsky, A.N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.
Crystal, D., 1985. Linguistics, second ed. Penguin, Harmondsworth, UK.
Crystal, D., 1996. A Dictionary of Linguistics and Phonetics, fourth ed. Blackwell, Oxford, UK.
Crystal, D., 1997. The Cambridge Encyclopedia of Language, second ed. Cambridge University Press, Cambridge, UK.
Crystal, D., 2000. Language Death. Cambridge University Press, Cambridge, UK.
Firth, J.R., 1957. A synopsis of linguistic theory, 1930–55. In: Philological Society (Great Britain), Special Volume, Studies in Linguistic Analysis. Blackwell, Oxford, UK, pp. 1–32.
Fries, C.C., 1952. The Structure of English. Harcourt Brace, New York.
Halliday, M.A.K., Fawcett, R.P. (Eds.), 1987. New Developments in Systemic Linguistics. Pinter, London.
Henderson, E.J.A. (Ed.), 1971. The Indispensable Foundation: A Selection from the Writings of Henry Sweet. Oxford University Press, London.
Hjelmslev, L., 1943. Omkring Sprogteoriens Grundlœggelse. Lunos, Copenhagen (Trans, as Prolegomena to a Theory of Language, 1961). University of Wisconsin Press, Madison, WI.
Jones, D., 1956. An Outline of English Phonetics, eighth ed. Heffer, Cambridge, UK.
Lyons, J., 1981. Language and Linguistics: An Introduction. Cambridge University Press, Cambridge, UK.
Pike, K.L., 1954–1960. Language in Relation to a Unified Theory of the Structure of Human Behavior. Summer Institute of Linguistics, Glendale, CA.
Quirk, R., Greenbaum, S., Leech, G.N., Svartvik, J., 1972. A Grammar of Contemporary English. Longman, London.
Robins, R.H., 1979. General Linguistics: An Introductory Survey, third ed. Longman, London.
de Saussure, F., 1916. Cours de Linguistique Générate. Payot, Paris (Trans, as Course in General Linguistics, 1959. Philosophical Library, New York).
Sapir, E., 1921. Language. Harcourt Brace, New York.

Linked Lives

Maja Djundeva, University of Groningen, Groningen, Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article refers to the Linked Lives, following the development of the life course theory during the 1980s and the 1990s and the seminal works of Elder and colleagues. In the first part the concepts linked to Linked Lives: timing, living in time and place, human agency are described in detail with examples from sociological, psychological, and developmental literature; whereas the second part reviews recent research that has employed advanced quantitative and qualitative methods using explicitly the theoretical principle of Linked Lives in the domains of family, migration, work and labor markets, and health.

The study of family and work life has been greatly enriched by the life course perspective, where the location of families in the broader social structure influences the events family members experience with the passage of time. Changes in mortality rates over the late nineteenth and all over the twentieth century (also referred to this as the Second Demographic Transition) coupled with changes in the macro- and microsocial environment not only expanded the human life span enormously, but also made lives intertwined in ways never imagined before. Life events such as marriage, giving birth, transition to work, becoming grandparent, or retirement have been transformed in normative social institutions that in their contemporary form diverge from the normative transitions in the pre-nineteenth century life, both with regards to their occurrence and timing (*see* Partnership Formation and Dissolution in Western Societies).

To understand the potential causes and consequences of change over time in the behaviors of individuals and families, and the ways the social structure plays a part in the changing nature of kin relationships, linked lives stemmed as a concept from the life course perspective and ecological paradigms that have emerged primarily from the intellectual traditions of sociology and psychology (*see* Life Course: Sociological Aspects). Previous analyses of families and individuals have mainly focused on discerning (1) the effects related to aging, (2) the effects of secular events, or historical period events, and (3) the interplay between aging and historical period events, also known as cohort effects. A debate between sociologists and psychologists in the 1980s over the dialectic of change and the importance of human agency and self-consciousness in the change process brought the focus on the need of a more holistic perspective in order to understand or explain family development (Marshall, 1986; Marshal and Rosenthal, 1991; Osmond, 1987). Developmental research that addressed individual life transitions as they relate to family changes, the interaction of individual and family changes with historical changes, and the relationship of earlier transitions to transitions occurring later in the life cycle (Harevan, 1987) cleared the way for developing the theme of families as agents and mediators of family change (Elder, 1977, 1981, 1991). Drawing upon symbolic interaction exchange and ecological theories, Elder and Caspi (1990) suggested that intra- and intergenerational processes can link the behaviors with specific family outcomes. Their study examined the complex interplay of historical and structural influences on individuals and families as they develop over time and paved a road to the understanding of linked lives.

Linked Lives in Life Course Perspective

As one of the main principle of the life course perspective, linked lives emphasizes the interconnectedness of lives, particularly but not exclusively linked by bonds of kinship. Proposed by Elder and Caspi (1990) as one of the five principles to examine the link between social change and the life course, it encompasses that "all lives are lived interdependently, and this connectedness defines a medium through which historical change plays out its influence over time" (Elder and Caspi, 1990: p. 221). In order to understand linked lives, the principles of lives in time and place (historical time and place), human agency, the timing of lives, and intergenerational transmission of behaviors are essential paradigmatic constituents of the life course perspective (Elder, 1994). Linked lives has provided a theoretical framework for researchers studying family, social stratification, labor markets, health and nutrition, and migration. More recently, researches have implemented the concept of linked lives to study physical activity, political activism, and social capital.

The principle of linked lives refers to the interdependence of lives – that the lives of individuals can be understood only as they are intimately bound to the lives of others. Social interdependence conveys that all individuals are embedded in a social structure where personal actions have consequences for others and the actions of others reflect on the self. The social regulation, structuring and support of human lives occur through interlocking relationships with significant others, tied by bonds of kinship and friendship. Across the life cycle of socialization and behavioral exchange in generational succession, opportunities and constrains become intergenerational, thus leading to interactive social worlds and networks of relationships over the life span (*see* Intergenerational Solidarity). Whereas social networks refer to the structural characteristics of social relationships (e.g., density, reciprocity) and support refers to the resources that flow from those relationships (e.g., emotional and financial assistance), the concept of linked lives considers that events that affect one person also affect other individuals in their networks.

Linked lives emphasizes the role of social institutions to define, segment, and construct individual lives in the interconnected domains of education, family, and work over the life course. Elder's study Children of the Great Depression (1974) traced the effects of drastic income loss to the life-experiences of the longitudinal subjects on their adolescent and adult years. This study that followed individuals and their families from the prosperous 1920s across the economic depression of the 1930s showed that not all families were devastated by the collapse of the economy. Namely, working-class and middle-class families had different experience due to individual-time and family-time differences and its conclusions were vital in the inception of linked lives. According to Elder et al. (2003), linked lives captures the interdependence of lives and the social and historical influences that are expressed through the network of shared relationships.

Linked Lives and Timing

In complementary ways with timing, linked lives addresses the temporality, process, and context of human lives and development. Both concepts also address the role of significant others in regulating life trajectories through networks of informal control (see Hagan, 1991). Life course trajectories are pathways, similar to that of a career (*see* Motivation: Life Course and Sociological Perspectives).

Timing refers to the way social roles and events are organized across the life course in a concept of age-grading, to a process of life management that schedules transitions in an appropriate way and to a location, both across the life course and history.

The trajectories of individuals are comprised of the sequences of roles and experiences that make up an individual's development. This critical developmental period specifies that events or exposures that occur at a particular or more vulnerable period of early life may have long-lasting effects on how an individual develops. The order of occurring of states is closely related to the meaning of age, and in essence to the stratification of the ways in which opportunities are made available to individuals in a population. Through family and intergenerational processes, adult development and life patterns are an integral part of child development (Parke et al., 1994: p. 35). Human capital and social capital are derived from the social relationships and working arrangements that sustain community life. The variations in human and social capital present a major turning point in the lives of children that will ultimately face different adult outcomes. A prime example of the interdependence of lives is the family world of farm children (Salamon, 1992). The link between farm households and community institutions such as the church, schools, and civic organizations contributes to the social capital essential for the maximization of the options of children and minimizing their risks (Furstenberg and Hughes, 1997).

Another meaning often applied to the timing of events in lives is the impact of experience and learning in adapting to external demands for change (Lazarus, 1999). These time role changes, such as becoming a parent during the child bearing years or becoming a widow in old age are considered normative transitions. However, turning points can involve changes in the direction of an individual's life and generally reflect nonnormative or unexpected events (*see* Life Course and Event History Analysis). Therefore, an account of the correspondence or goodness of fit of the individual with the environment is essentially connected to how normative or nonnormative the life course transitions in individual's life play out. The idea of person–environment link is closely related to the timing considered as normative transition of social institutions, or the right time in the ontogeny of individual's age for events like marriage, birth of a child, or retirement. Social institutions are predictable states or periods of the life course as the majority of the population is expected to experience such states, and their on- or off-timing is considered to be a significant predictor of the developmental trajectories of individuals. There can be a 'best fit' in the timing of individual development and family life stage, and their temporal convergence with structural and historically created opportunities (Elder et al., 2003). The changes in states or roles refer specifically to times in the developmental trajectory when the individuals make normative changes.

The interactive nature of relationships represents a life-long process, linking developmental trajectories, while the temporal arrangement of social roles provides a way of thinking about time as sequence. Children, spouses, partners, or parents, exert considerable influence on individual experiences and attainments over the life course. Of particular significance are the conjoint life course dynamics that span over generations and couples. Social roles expose individuals to the stresses and strains of others, as well as to the possibility of social support, therefore linking the concept of timing with linked lives.

Linked Lives and Historical Time and Place

Historical time and place captures the large events such as depression and war that can shape not only individual psychology, family interactions, and world views, but also create the opportunities and constraints that circumscribe the available choices and behaviors (Bengtson et al., 2005). Linked lives in time emphasizes the importance of social and historical context in shaping individual lives by affecting the timing and occurrence of role transitions over childhood and adult life.

Assigning people to cohorts according to their birth or other times of entry into society has been a starting point to distinguish between the explanations why some children grew up to be less affected by the negative events in their lives than others. Research comparing the life patterns of successive cohorts is prompted by the thesis that individuals can develop in different ways according to the nature of social change (*see* Age Stratification). Older and younger children will not experience an economic recession or depression in the same way because they differ significantly in their resources and social roles. Younger children, especially boys, were the most adversely influenced by the economic stresses of the Great Depression (Elder, 1979). Similar example is Hogan's (1981) study of conventional event sequences in the transition of adulthood among cohorts of American men. The conventional sequence of the end of education, work entry, and marriage has been

changed for a large number of young men who served in the Second World War, where more frequently, marriage came before the completion of education as veterans returned back to school. Cohort effects relate to linked lives not only through the developmental trajectories of children, but also through the timing and order of the life course transitions of adults. Historical events can produce different life course changes in successive birth cohorts. For example, a study of the four birth cohorts of women who settled in Elmira, New York (Bronfenbrenner, 1958) found employment before marriage to be the primary source why women delayed childbearing in the first two cohorts, who came of age during the Great Depression and during the Second World War, respectively. In the two younger cohorts, education played more important role in the lives of women, seen also by the trends of return to school following marriage and the birth of a child. Historical influence takes the form of a cohort effect when social change differentiates the life patterns of successive cohorts.

Another important aspect of historical time and place is the geographic setting linked to specific places and their properties. Qualitative studies that have brought evidence of the changing communities and their institutions give valuable insights to the understanding of families' and individuals' linked lives. Social interdependence linked by sharing the same geographic or physical space embeds individuals in a social structure where personal actions have consequences for others, therefore relating to linked lives. The social regulation, structuring and support through kinship and other relationships embodies the meaning of linked lives. Biographical and historical timing had consequences for the demographic behavior, occupational outcomes, and psychological well-being. Regardless of the importance of historical events, which can be operationalized as cohort effects or period effects (that occur when a historical event has a similar effect on multiple birth cohorts), living in place proves to be a challenging concept in times of great special mobility.

Linked Lives and Human Agency

Human agency is based on the assumption that humans are not passive recipients of a predetermined life course but make decisions that determine their lives. Within the constraints of the individuals' own worlds, people are competent to make choices among alternatives that form and can change their life course. Choices have transitional consequences and they may entail a turning point, thus, some decisions require intense focus on the present and others are influenced by long-term goals, giving agency a distinct temporal orientation (Hitlin and Elder, 2007). Linked lives refers to the embeddedness of the life course, where individuals are actors who behave or decide in a social context in which they may not adhere to a previously written script, written for them by the particular social categories that they occupy (Granovetter, 1985: p. 487). In addition, human agency necessarily transcends individuals as the consequences from individual actions necessarily are ascertained by their broader social networks.

The continuity of behaviors across the life course is sustained in social interaction (by its cumulative consequences) and there is a tendency these styles to evoke maintaining reciprocal responses from the environment (Caspi et al., 1989). This refers directly to linked lives as the agency in the selection of particular roles or situations represents a mechanism through which life advantages and disadvantages may begin to cumulate. In that way, behavior is sustained or changed by its consequences. An example for that is how competence in adolescence is rewarded by others, which in turn predicts adult competence in the late middle years (Clausen, 1993). Both individual dispositions and family values may favor a choice of compatible environments, which in turn should reinforce and sustain the match. For example, antisocial youth tends to affiliate with similarly problematic youth, and this interaction accentuates the antisocial behavior, which over time produces cumulative disadvantages. The cycle of cumulative disadvantage can be broken by new life transitions into different environments that will assist the process of change by establishing potential turning points for a troubled life course (Laub and Sampson, 1993). For example, marriage and employment can help young men who previously exhibited antisocial behaviors to transition toward more conventional lives by providing a network of individuals to reinforce conventional behavior. Conversely, negative life events such as spouse bereavement or even divorce can put individuals on a track of cumulative disadvantage, also affecting the lives linked to those who experience the events firsthand.

Linked Lives in Changing Context: Overview of Quantitative Studies

Cross-sectional research brought positive evidence for the notion of 'linked lives' (Greenfield and Marks, 2006; Pillemer and Suitor, 1991), which investigated how circumstances in adult children's lives would have implications for the lives of their parents and, accordingly, adult children's problems would affect their parents' psychological and relational well-being. Using the 1995 National Survey of Midlife in the United States, Greenfield and Marks's study (2006) concluded that parents who reported having adult children with more types of problems demonstrated poorer well-being than parents who reported having adult children with fewer types of problems. Other cross-sectional research also showed that parents feel more ambivalent toward children when children have more problems in their relationships (Birditt et al., 2010). On the other hand, Milkie, Bierman, and Schieman's study (2008) using longitudinal data found no effect of children's divorce on changes in parents' level of depression.

The social embeddedness of the life course has also been researched with longitudinal data on parents and young children (Sigle-Rushton et al., 2005) and findings suggest that the relationship between parental divorce and subsequent childhood and adult disadvantage is not straightforward. Little evidence was found that parental divorce is linked with subsequent childhood and adult disadvantage. In research on partners in the same household, another study showed that spouses' resources do affect occupational attainment, especially in the case of the husband's resources on his wife's attainment than vice versa (Bernasco et al., 1998). In addition, the study also found negative effects of a wife's labor-market resources on

her husband's employment exit and positive effects on his occupational prestige score after a job change.

Studying the relationship between family members outside the household, Aquilino (1994) investigated how family disruption that occurred after children were grown had sizable effects on parent–adult child relations. He found that later life parental divorce lowered relationship quality and contact between adult children and parents, with particularly stronger effects for father–child than for mother–child relations, and stronger for father–daughter than for father–son relations. Kalmijn and De Graaf (2012) also studied the relationship between family members outside the household and their study showed that parents are affected by the demographic life course transitions of their children. Mothers experienced a decline in well-being when their children separated, and both fathers and mothers increased their well-being when their children married or begin living together. Fathers also experienced an increase in well-being when their children became parents.

Stretching over interdisciplinary borders, researchers have applied the concept of linked lives to the research of health. For example, living partners may provide a lifestyle environment for each other, sharing health habits and behavior. Being married is associated with more positive health habits in general and a review of the relationship between health and marriage found that married people practice more healthful behaviors (Schoenborn et al., 2004). The application of linked lives to the study of physical activity revealed that married adults were found to exercise less than unmarried adults, and having children under the age of five was negatively associated with exercising (Nomaguchi and Bianchi, 2004). With regards to health behaviors and physical activity, these relationships are important intervening units for affecting individuals – as part of couples, as members of families, and through networks of friends. For example, a physical activity and nutrition intervention for couples has been shown to be effective in improving health behaviors and potentially reducing the risks of lifestyle-related diseases (Burke et al., 2004).

In the context of family migration, Bailey et al. (2004) explored how family relations influence the migration decisions of partners in dual-earner households. They found that intergenerational links may increase the incidence of a return migration of later-life dual-earner households. In addition, scholars have posited a link between historical cultural losses and contemporary accounts of disruptive behaviors and mental health problems among indigenous people in the United States. Recent empirical evidence indicates associations between intrusive thoughts of historical losses (e.g., loss of land, loss of language) and affective outcomes including guilt, hopelessness, despair, anger, substance abuse, and depressive symptoms (Whitbeck et al., 2004). The concept of linked lives has also been applied to the research on racism, examining how experiences of discrimination encountered by one person may indirectly affect another. Discrimination may result in secondary effects such as financial strain, decreased relationship quality, and intimate partner violence (see review Gee et al., 2012). For example, past experiences with discrimination can shape a Black mother's involvement with her child's schooling (Rowley et al., 2010).

In order to fully understand stratification and labor-market research scholars started to acknowledge the degree to which employment histories and family processes are related. For that purpose, linked lives is used as an additional aspect that to large extent has been neglected in stratification and labor-market research, addressing the relationship between individual lives and social structure in a way that simultaneously seeks to uncover mechanisms of change. Blossfeld and Drobnic's work (2001, Drobnic and Blossfeld, 2004) incorporating a linked lives perspective found that interdependencies in couples exist with regards to women's labor market integration, with variations spanning across marriage cohorts and various countries. They also found evidence for the importance of class-specific differences in the division of labor between husbands and wives over the family life cycle in differing marriage cohorts and countries.

The various contexts of family, work, education, migration, and other fields of research that have employed linked lives are of great relevance and can gain the appropriate visibility through multilevel studies of not only families, but also neighborhood and community effects in the lives of others (Furstenberg and Hughes, 1997; Sampson et al., 1997). Taking into account country and cohort variations, longitudinal studies of individual change started to incorporate more a type of multilevel models, in which multiple time points are 'nested' within a person; or multiple persons are nested within families or within countries. A model with multiple levels has the ability to estimate an individual's growth within the larger social context. With advances in statistical models, there is an increasing ability to investigate lives by emphasizing the social dynamic of linked lives. Individuals' trajectories and transitions can be better understood using advanced techniques such as piecewise growth curve analyses, multilevel models, event history models, sequence analysis, and Geographic Information Systems analysis. These techniques suit the examination of changes that are associated with a particular event or critical period, but also have the ability to operationalize macrosocial indicators of influence, thus capturing the core of linked lives (*see* Life Course and Event History Analysis).

Qualitative Studies and Life Course Analysis

The potential for an analysis of interdependency through a focus on linked lives (Elder, 1974; Mason, 2004) resonates with the way well-being could be measured in developing countries. This has the potential to help counter the tendency toward methodological individualism that focuses solely on quantitative methods. In contrast to static measures of well-being, life course analysis facilitates a view that is particularly valuable given development studies' focus on how processes of change affect human experience over time. Life history research mainly employs semistructured interviews that aim to 'capture the concrete details of lives in context' (Cairns and Cairns, 2002: p. 292) without allowing data collection to be overwhelmed by standardized procedures in way that might destroy the 'intimacy' of narrative accounts (Davis and Baulch, 2011).

Therefore, the study of the connections that extend across generations and serve to integrate young and old makes use

of the life histories research. Qualitative life course analysis has great potential to examine both the perceived intentionality of individual's action (agency) and to evaluate individual's experience (subjective meaning) in the context of linked lives and changing times (Locke and Lloyd-Sherlock, 2011). Using narratives of linked lives, Locke and Lloyd-Sherlock (2011) explore the lives of 22 older people in Buenos Aires with respect to their lifetime relations with their children and their current well-being. In addition, they also conduct 60 in-depth semistructured interviews of young adults in Zambia focusing on their life trajectories from birth, through youth to early adulthood with an emphasis on education and employment. In both cases a strong emphasis on linked lives is present by exploring the role of family, friends, and other individuals, as well as on the influence of shifting social, political, and economic contexts on individual experiences.

Locke and Lloyd-Sherlock take an approach of tracking linked lives in a respondent-centric way, focusing on what respondents have to say about those who were important in their lives at various times, their relationships with them, and what it meant for them. By analyzing commonalities and differences in how informants referred to particular relationships, they expose the turning points or changes in circumstances of the respondents. These methods are more established in development studies where the value of qualitative research is well accepted. Developing the credibility of this methodology for a wider audience is still seen as a challenge as only a relatively small proportion of the data is available for academic scrutiny by others.

Concluding Remarks

The need for new concepts of human development that should apply to processes across the life span gave birth to the idea of linked lives. The influence of social and historical factors on networks of shared relationships is a crucial dynamic for the exploration of 'linked lives' (Elder, 2002). To understand how disadvantage can be linked both within and across the generations, it is necessary to consider contextual factors in individual accounts. Interpretations of these accounts must be situated within a robust analysis of 'the changing times' (Elder, 2002: p. 210), as well as the changing places they inhabit. This would not have been able without the pioneering longitudinal studies of child development (launched in the 1920s and 1930s) that have extended well beyond childhood. These studies followed their young study members up to the middle years and later life. In doing so, they generated issues that recognized that individual lives are influenced by an ever-changing historical context, noticing that the study of human lives calls for new ways of thinking about their pattern and dynamic.

In addition, next to studying the right time in the ontogeny of individual's age for events like paid work and marriage (Moen and Sweet, 2003), future avenues of research should devote greater attention to broader conceptualizations that would simultaneously incorporate school and parenthood (Macmillan and Copher, 2005). Delaying parenthood together with the increasing level of education, especially for women requires a hands-on investigation of the current trends of combining work and family. Further, contemporary societies are facing challenges such as population aging and aging workforce, therefore exploring linked lives in the context of prolonging working life and living healthy would be essential in understanding the intertwined ways in which individuals' lives meet societal change.

See also: Adulthood: Dependency and Autonomy; Age Stratification; Age Structure; Aging and Health in Old Age; Demographic Models; Discrimination: Racial; Event History Analysis in Continuous Time; Intergenerational Solidarity; International Migration; Labor, Division of; Life Course: Sociological Aspects; Lifespan Development, Theory of; Migration and Fertility; Mobility, Social; Multilevel Models in Demography; Social Changes: Models; Social Influence, Sociology of; Social Networks; Social Relationships in Adulthood; Sociological Traditions.

Bibliography

Aquilino, W.S., 1994. Later life parental divorce and widowhood: impact on young adults' assessment of parent-child relations. Journal of Marriage and Family 56 (4), 908–922.

Bailey, A.J., Blake, M.K., Cooke, T.J., 2004. Migration, care, and the linked lives of dual-earner households. Environment and Planning A 36 (9), 1617–1632.

Bengtson, V.L., Putney, N.M., Johnson, M.L., Malcolm, L., Bengtson, V., Coleman, P., Kirkwood, T., 2005. The problem of theory in gerontology today. In: The Cambridge Handbook of Age and Ageing, pp. 3–20.

Bernasco, W., de Graaf, P.M., Ultee, W.C., 1998. Effects of spouse's resources on occupational attainment in the Netherlands. European Sociological Review 14 (1), 15–31.

Birditt, K.S., Fingerman, K.L., Zarit, S.H., 2010. Adult children's problems and successes: implications for intergenerational ambivalence. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 65B (2), 145–153.

Blossfeld, H., Drobnic, S., 2001. Careers of Couples in Contemporary Society: From Male Breadwinner to Dual-earner Families. Oxford University Press, Oxford.

Bronfenbrenner, U., 1958. Socialization and social class through time and space. In: Readings in Social Psychology. Holt, New York, pp. 400–425.

Burke, V., Giangiulio, N., Gillam, H.F., Beilin, L.J., Houghton, S., 2004. Changes in cognitive measures in a randomized controlled trial of a health promotion program for couples targeting diet and physical activity. American Journal of Health Promotion 18 (4), 300–311.

Cairns, R.B., Cairns, B.D., 2002. Plotting developmental pathways: methods, measures, models, and madness. In: Looking at Lives. American Longitudinal Surveys in the Twentieth Century. Russell Sage Foundation, New York, pp. 267–296.

Caspi, A., Bem, D.J., Elder, G.H., 1989. Continuities and consequences of interactional styles across the life course. Journal of Personality 57 (2), 375–406.

Clausen, J.A., 1993. American Lives: Looking Back at the Children of the Great Depression. University of California Press, Berkeley.

Davis, P., Baulch, B., 2011. Parallel realities: exploring poverty dynamics using mixed methods in rural Bangladesh. The Journal of Development Studies 47 (1), 118–142.

Drobnic, S., Blossfeld, H.P., 2004. Career patterns over the life course: gender, class, and linked lives. Research in Social Stratification and Mobility 21, 139–164.

Elder, G.H., 1974. Children of the Great Depression: Social Change in Life Experience. University of Chicago Press, Chicago.

Elder, G.H., 1979. Historical change in life patterns and personality. Life-Span Development and Behavior 2, 117–159.

Elder, G.H., 1977. Family history and the life course. Journal of Family History 2, 279–304.

Elder, G.H., 1987. War mobilization and the life course: a cohort of World War II veterans. Sociological Forum Springer 2 (3), 449–472.

Elder, G.H., 1981. History and the family: the discovery of complexity. Journal of the Marriage and the Family 43, 489–519.
Elder, G.H., 1991. Life course. In: Borgatta, E.F., Borgata, M.L. (Eds.), The Encyclopedia of Sociology. Macmillan, New York, pp. 281–311.
Elder, G.H., 1994. Families in Troubled Times: Adapting to Change in Rural America. Transaction Publishers, Brunswick.
Elder, G.H., 2002. Historical times and lives: a journey through time and space. In: Looking at Lives: American Longitudinal Studies of the Twentieth Century, pp. 194–218.
Elder, G.H., Caspi, A., April 1990. Studying lives in a changing society: sociological and personological explorations. In: Rabin, A.I., et al. (Eds.), Studying Persons and Lives. Springer, New York, pp. 203–247.
Elder, G.H., Johnson, M.K., Crosnoe, R., 2003. The emergence and development of life course theory. In: Handbook of the Life Course, pp. 3–19.
Furstenberg Jr., F.F., Hughes, M.E., 1997. The influence of neighborhoods on children's development: a theoretical perspective and a research agenda. In: Indicators of Children's Well-Being, pp. 346–371.
Gee, G.C., Walsemann, K.M., Brondolo, E., 2012. A life course perspective on how racism may be related to health inequities. American Journal of Public Health 102 (5), 967–974.
Granovetter, M., 1985. Economic action and social structure: the problem of embeddedness. American Journal of Sociology 481–510.
Greenfield, E.A., Marks, N.F., 2006. Linked lives: adult children's problems and their parents' psychological and relational well-being. Journal of Marriage and Family 68 (2), 442–454.
Hagan, J., 1991. Destiny and drift: subcultural preferences, status attainments, and the risks and rewards of youth. American Sociological Review 567–582.
Harevan, T., 1987. Historical analysis of the family. In: Susman, M.S., Steinmetz, S.K. (Eds.), Handbook of Marriage and the Family. Plenum, New York (Chapter 2).
Hitlin, S., Elder, G.H., 2007. Time, self, and the curiously abstract concept of agency*. Sociological Theory 25 (2), 170–191.
Hogan, D.P., 1981. Transitions and Social Change: The Early Lives of American Men. Academic Press, New York.
Kalmijn, M., De Graaf, P.M., 2012. Life course changes of children and well-being of parents. Journal of Marriage and Family 74 (2), 269–280.
Laub, J.H., Sampson, R.J., 1993. Turning points in the life course: why change matters to the study of crime. Criminology 31 (3), 301–325.
Locke, C., Lloyd-Sherlock, P., 2011. Qualitative life course methodologies: critical reflections from development studies. Development and Change 42 (5), 1131–1152.
Lazarus, R.S., 1999. Stress and Emotion. Springer, New York.
MacMillan, R., Copher, R., 2005. Families in the life course: interdependency of roles, role configurations, and pathways. Journal of Marriage and Family 67 (4), 858–879.
Marshall, V.W., 1986. Dominant and emerging paradigms in the social psychology of aging. In: Marshall, V.W. (Ed.), Later Life: The Social Psychology of Aging. Sage, Beverly Hills, CA, pp. 9–33.
Marshal, V.W., Rosenthal, C.J., 1991. Aging and later life. In: Hagedorn, R. (Ed.), Sociology, fourth ed. Holt, Rineheart and Winston, Toronto, pp. 133–162.
Mason, J., 2004. Personal narratives, relational selves: residential histories in the living and telling. The Sociological Review 52 (2), 162–179.
Milkie, M.A., Bierman, A., Schieman, S., 2008. How adult children influence older parents' mental health: integrating stress-process and life-course perspectives. Social Psychology Quarterly 71 (1), 86–105.
Moen, P., Sweet, S., 2003. Time clocks: work-hour strategies. It's About Time: Couples and Careers. Cornell University Press, Ithaca, NY, pp. 17–34.
Nomaguchi, K.M., Bianchi, S.M., 2004. Exercise time: gender differences in the effects of marriage, parenthood, and employment. Journal of Marriage and Family 66 (2), 413–430.
Osmond, M., 1987. Radical-critical theories. In: Susman, M.S., Steinmetz, S.K. (Eds.), Handbook of Marriage and the Family. Plenum, New York, pp. 103–124.
Parke, R.D., Ornstein, P.A., Rieser, J.J., Zahn-Waxler, C.E., 1994. A Century of Developmental Psychology. American Psychological Association, Washington, DC.
Pillemer, K., Suitor, J.J., 1991. "Will I ever escape my child's problems?" Effects of adult children's problems on elderly parents. Journal of Marriage and Family 53 (3), 585–594.
Rowley, S.J., Helaire, L.J., Banerjee, M., 2010. Reflecting on racism: school involvement and perceived teacher discrimination in African American mothers. Journal of Applied Developmental Psychology 31 (1), 83–92.
Salamon, S., 1992. Praire Patrimony: Family, Farming, & Community in the Midwest. University of North Carolina Press, Chapel Hill.
Sampson, R.J., Raudenbush, S.W., Earls, F., 1997. Neighborhoods and violent crime: a multilevel study of collective efficacy. Science 277 (5328), 918–924.
Schoenborn, C.A., Adams, P.F., Barnes, P.M., Vickerie, J.L., Schiller, J.S., 2004. Health Behaviors of Adults: United States, 1999–2001. Vital and Health Statistics. Series 10, Data from the National Health Survey, no. 219, pp. 1–79.
Settersten, R.A., 2003. Age structuring and the rhythm of the life course. In: Handbook of the Life Course, pp. 81–98.
Sigle-Rushton, W., Hobcraft, J., Kiernan, K., 2005. Parental divorce and subsequent disadvantage: a cross-cohort comparison. Demography 42 (3), 427–446.
Whitbeck, L., Adams, G., Hoyt, D., Chen, X., 2004. Conceptualizing and measuring historical trauma among American Indian people. American Journal of Community Psychology 33 (3–4), 119–130.

Literacy and Illiteracy, History of

Istvan G Tóth, Hungarian Academy of Sciences, Budapest, Hungary

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8961–8967, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The period from the early sixteenth century up to the end of the eighteenth century saw a genuine revolution in the cultural history of the states of Europe. This article charts the development of literacy, especially in Europe and North America, in the early modern period, arguing that this was the most important time in the history of literacy, marking the period of transition from restricted to mass literacy in the West. It also considers problems of methodology in measuring literacy and reviews some of the statistical data from various European countries.

The period from the early sixteenth century up to the end of the eighteenth century saw a veritable revolution in European cultural history. In the Middle Ages, it was taken for granted that only a select few knew how to read and write and that many of them made their living out of these skills as scribes and clerks. By the late eighteenth century, however, in the most developed regions of Europe, about 50% of men were able to acquire information from reading the newspapers, and the number of people who could write was steadily rising. The skills of reading and writing and, beyond that, the use of literacy, were rapidly gaining ground, to the detriment of oral communication. The body of written documents increased drastically; literacy made its presence felt more and more strongly throughout society. Printed books grew in number while becoming cheaper and more things were thought worthy of being committed to paper. Literacy extended its power to ever-newer sectors of a developing state administration and bureaucracy in early modern Europe. Surveys and certificates were drawn up and the use of written documents gradually squeezed out verbal evidence. Inevitably, people's mentality and ways of thinking underwent a simultaneous change.

Orally transmitted knowledge and the memory of the elderly lost a good deal of their former prestige, while respect for information set down in writing, which increasing numbers of people were able to decipher, strengthened. With the spread of calendars and printed maps, the concepts of space and time altered; numerically articulated, measurable time came to replace an image of a past drawn with vague, uncertain outlines. Acquiring writing skills alone did not necessarily involve a change in mentality. Owing to the gradual spread of literacy in society, however, more and more people were able to read the dates on printed works and this affected their way of thinking. People knew what the date was and how old they were, and tied events to specific days, months, and years. The traditional concept of time had rested on oral tradition and memory, associating the passage of time with events such as wars and plagues, and the changing seasons. This slowly gave way to the new concept based on literacy, supported by written documents. This article tracks the development of literacy, especially in Europe and North America in the early modern period, the most important time in history of literacy, marking the period of transition from restricted to mass literacy in the West.

Literacy: Definitions and Mutual Relations

Functional literacy, i.e., the capacity to read a newspaper, sign a check and write a short postcard at least, was an essential and indispensable precondition of the functioning of nineteenth- and twentieth-century society. These skills were, however, lacking in the majority of the population in early modern Europe.

Literacy is much more than just the capacity to read and write. It is the constant practical use of writing and of the written record. Literacy represents a different way of seeing the world and storing information from that used by illiterate people.

During the 1990s, very important research was carried out on literacy in ancient societies, evaluating in a new light data that had been available for centuries, but not regarded as sources of literacy: descriptions of reading and writing, orthography of mural epigraphs, and so on.

Literacy remains a very real problem in large parts of the world outside the West (and in parts of the Western world where there are poor schools, and where the phone and television have completely replaced letter writing and newspaper reading, etc.). However, literacy played a central part in European cultural and social history in the centuries between 1500 and 1800. Literacy is typically a phenomenon of transition – these centuries saw in Europe the transition from a restricted medieval literacy to modern mass literacy. After the invention of the printing press in the mid-fifteenth century, the quantity of written documents, from books to newspapers and leaflets, grew very quickly, and this encouraged more and more people to learn to read and, to a lesser extent, to write.

Illiteracy is by no means a synonym for lack of culture. The oral tradition and the oral culture of traditional societies are as important and precious as the written culture and literate world of modern societies. The oral tradition did not disappear with the onset of literacy: in early modern Europe there was a coexistence of the two cultures. The oral culture was not, however, an isolated island in the dominating literate world; illiterate people using orality lived very much in the middle of a society where the most important decisions were put into a written form, and where more and more people could read and write. There was a continuous interaction between the two worlds. Illiterates used written documents that they could not decipher, e.g., passes and documents concerning ownership of their

houses and land. Often the content of lost written documents was conserved only by orality, e.g., those concerning the migration of population or the origin of settlements. Such information was often kept in memory only by the tales of the old people in the villages who may once have read such lost documents.

General illiteracy influenced the exchange of information, making it difficult for the masses to maintain contact with the world outside their town or village. In the twenty-first century, it is natural for a literate man to write a letter and read a notice or a brochure. For the unlettered, though, this causes grave difficulties and requires the assistance of others. Until the widespread use of the telephone, large sections of society practiced writing mainly through corresponding with relatives. Even illiterates tried to exchange letters, which meant sharing their innermost thoughts with those who offered to commit them to paper.

Measuring Literacy: Problems of Methodology

The best sources for measuring literacy are census records. During a census every citizen was asked if he or she could read and write, or only read, or neither read nor write. However, such censuses were carried out only from the second half of the nineteenth century. As literate people generally came from the upper, richer strata of society, they had a higher general life expectancy than illiterates. For this reason we should not draw retrospective conclusions on literacy from the age groups of censuses, as it is not absolutely certain that literate and illiterate people died in the same proportion.

We can estimate the rates of literacy before the beginning of censuses by consulting descriptive sources and statistics of crosses and signatures. Descriptive sources such as travelogs, reports of state officials, data on elementary education, book production, and book ownership, cannot furnish reliable quantitative data on the literacy rates. However, these are valuable indirect sources because we can analyze the competence of the writing. It is only exceptionally recorded that someone had authenticated a document 'by alien hand, not knowing the letters'; in most cases, only a cross by way of signature or some characters suspiciously identical with those in other signatures show that somebody was illiterate. In sources such as witness depositions, however, hundreds of nobles, peasants, and burghers declared that they could not read, or this fact about them was recorded by travelers, state administrators, etc.

Good statistics on literacy rates in early modern Europe can be drawn up, however, only on the basis of enumeration of people who signed their names or who could only make a cross instead of a signature. For this purpose we need a representative cross-section of the society where the social rank and geographical position (and if possible, age) of the persons in question can be determined. Often, however, only males and only the better-off figure in the documents, or only one professional group is represented.

Statistical data hardly reflect the real state of affairs concerning literacy. The method of counting crosses and signatures rests tacitly on the supposition that all those whose signatures have been found could indeed write, while those who left merely a cross on documents were all lacking the skill of even writing their names. Both suppositions could be erroneous, because examples abound of counterfeit signatures being entered on behalf of illiterate people and of crosses made by people well trained in penmanship. This reflects the mentality of a given society: if literacy has already acquired prestige, even illiterates try to fake writing capacity; however, if the great majority in society are illiterate, handwriting has no authentication function and thus even village notaries make only a cross, even if they can write properly. This would not come to light unless we have more than one sample of the signatures or crosses left behind by the same person, which, in turn, presupposes a recurrence of the same name, whose bearer gave a signature at one time and a cross at another. The statistical data concerning crosses and signatures, therefore, reveal the attitudes of a community toward literacy, rather than the concrete spread of literacy: they speak about the community expectations concerning literacy, the pressures on semiliterate persons to sign the documents themselves, and on the genuinely unlettered to forge a signature. Such mentality only emerged, though, if literacy was on the rise in a society. In their totality, therefore, these statistics more or less do mirror the advance of literacy. There is no doubt that if we divide society strictly into literate and illiterate people, the major part remains concealed: the semi-illiterate layer of society that did learn something of the alphabet and later, in accordance with the community's wishes, either made crosses or wrote their name on official documents. The same applied to reading, too: those who were taught the letters at school but hardly ever read, could profit from the printed prayer books whose text they knew by heart (extensive reading). However, they did not fare well when they had to read a handwritten circular whose contents were unknown to them (intensive reading). The unlettered and the literate were in fact never separated as sharply as may be suggested by the rows of crosses and signatures. No clear-cut distinction can be made between fully illiterates, who could neither read nor write, and semiliterates, who could read but not write: both groups left crosses on documents. Yet there is no greater gap in levels of education than that separating those able at least to read, to gather information from written texts, from those to whom the world of letters is literally a closed book.

The greatest number of written sources and of sources on literacy are to be found where the levels of literacy were the highest, and obviously this greatly influences research on literacy overall. In England from the seventeenth century onwards, witnesses authenticated their testimony by putting their signature or mark on the records and people signed oaths of fidelity en masse. In several Western European countries, spouses and their witnesses also confirmed a marriage by signing or marking the marriage register. This explains why the marriage register is a major source for the spread of literacy. In France, marriage registers had to be signed from 1686, in England from 1754. The city of Amsterdam, one of the most literate places of the world in this time, required marriage registers to be signed from as early as 1578. In 1630, 43% of bridegrooms and 68% of brides made a cross; 100 years later in 1730, the proportion is 24 and 49% respectively. In 1780, only 15% of men and 36% women could not sign their names at their marriage.

The conquests of Napoleon Bonaparte open up new possibilities for the historians of literacy in Germany and Italy. These two countries, so important for the cultural history of early modern Europe, lacked similar mass sources in the sixteenth to eighteenth centuries for evaluation of literacy levels. After they became part of Napoleon's empire, however, the French system was introduced – thus, around 1800, we have a statistically relevant quantity of crosses and signatures from many regions of Germany and Italy, as they were parts of France at this time. In Central and Southern Europe, however, only the minister conducting the marriage ceremony was expected to write in the register, and it was not compulsory for eyewitnesses to sign their accounts in court. Witnesses confirmed their accounts only by taking an oral oath, without signing the document. This fact in itself is indicative of the lesser role literacy and writing played in these regions compared to Western Europe.

Reading, in contrast to writing, leaves no traces behind; that is why researchers into literacy almost invariably examine the ability to write – or more specifically, to produce a signature – and give up trying to gauge reading abilities as a hopeless venture. In several European countries, however, an extraordinary source lends scope for the examination of reading skills en masse: the Lutheran pastors in those areas made an annual survey of reading skills in their flock from the late seventeenth century. In the history of European literacy, a position of eminence is certainly due to Sweden (with Finland and Latvia, which still belonged to it at the time), Zurich and vicinity in Switzerland, and certain regions in northern Germany (seventeenth and eighteenth centuries) where the Lutheran faith thrived. In the course of Lutheran church visitations in these territories, records were made of, among other things, the reading skills of all adults, even indicating the grades – so an otherwise nonmanifest skill can be examined on the basis of a large sample. The Lutheran church believed it to be of great importance for believers to read the catechism and the Bible. Those reluctant to learn to read were denied communion at Easter. The Swedish model is especially interesting, as the church was pushing for better reading skills without enforcing writing capacities. As a result there was a huge gap between reading and writing abilities – in the eighteenth century almost everybody (around 90%) was able to read, but only few (around 10%) could write. Female literacy was rising, too, and the role of the mother as a teacher at home became very important, especially in the nineteenth century.

The Lutheran church records are, however, not the only sources for examining reading capacities before the age of censuses. In Hungary, when the county administration investigated claims to noble title, it asked the nobleman making the claim whether he had read, or only seen, the lost letters patent that would prove his nobility. In three different regions of eighteenth-century Hungary, 71, 74, and 55% respectively of the lower noblemen declared that 'as they cannot read at all' they had only seen, but not read, the lost document.

Literacy in Europe

England together with the Netherlands, was the most literate country in the world during the period. In the mid-seventeenth century, during the stormy years of the Civil War, Englishmen were required to sign oaths of allegiance several times; consequently, there are ample data available concerning the state of literacy in various parts of the county, in the form of thousands of signatures and signs. On examining these, David Cressy (1980) has found that in England in the 1640s, some 70% of men and 90% of women were illiterate. As regards the sixteenth-century situation, only estimates can be made. Cressy avers that in the middle of the sixteenth century, 80% of men and 95% of women could neither read nor write. The percentages around 1500 may have been 90 and 99 respectively, though this is a mere surmise as even in England no statistical data about literacy can be culled so early. The situation in eighteenth-century Britain is much better known. According to David Cressy's estimates based on the available sources, at the beginning of the century 55% of men and 75% of women could write their own names. In 1754, it was stipulated by law in Britain that the marriage register had to be signed by both the bridegroom and the bride. From that time onwards, therefore, the development of literacy can be traced on the basis of precise data. By the middle of the eighteenth century, the majority of men knew how to write, only 40% being unlettered, while the rate among women was 60%. In the second half of the eighteenth century, however, this dynamic progress came to a sudden halt. Until the early nineteenth century, illiteracy stopped falling in Britain. The social transformation that was taking place as a consequence of the Industrial Revolution was probably to blame for this – inner migrations and early child labor were not really conducive to literacy. In the second half of the nineteenth century, though, illiteracy in Britain all but disappeared. Between 1850 and 1911, the proportion of illiterate men decreased from 30 to 1%, and that of women from 45% to less than 2%.

France is the best researched of all countries in terms of literacy in the early modern age. From the age of Louis XIV, it was compulsory for brides, bridegrooms, and witnesses to put their signature in the register at weddings, thus leaving a wide scope for researches into literacy. In France in the eighteenth century, the number of literate people almost doubled, even though at the time of the outbreak of the French Revolution they were in the minority, the proportion of literate people in France grew from 21% in 1686–90 to 37% in 1786–90. Frenchwomen also lagged behind men – literacy among them grew from 14 to 27% within 100 years, while among the men it rose from 29 to 47%. However, averages that apply to entire countries hide what is really relevant – the huge regional differences that were so characteristic of this large country. That imaginary dividing line connecting Geneva and Saint Malo, which in the seventeenth century separated the literate northern and eastern parts of France from the more backward western and southern parts, was still clearly visible on the map of literacy at least a century later, and in the case of women's literacy, as late as 1866. This indicates that regional imbalances in the spread of literacy, as an important index of development, take a long time to change. When the French Revolution broke out, there were huge regional differences in the literacy of men. In Normandy and in the countries north and east of Paris, 90% bridegrooms signed the register in Brittany and on the Atlantic coast, their proportion was less than 10%. In the secluded regions of the Massif Central, the percentage of men who put

their signature in the register was 16, of the women a mere four while in Alsace-Lorraine the percentages were 83 and 44 respectively. The differences did not diminish until the second half of the century.

By the end of the nineteenth century, illiteracy had in effect been wiped out in France. While in 1854, 13% of bridegrooms were unable to sign the marriage register, the number decreased by 1900 to a mere 5%. Within the same period the proportion of illiterate brides decreased from 46 to 6%. Compulsory schooling put an end to the difference between male and female literacy, which had been a marked feature in earlier centuries.

The newly unified Germany presented a very mixed picture from the point of view of literacy, too. In the Prussian kingdom in 1871, 10% of men above 10 years could neither read nor write. However, this average was calculated again on the basis of very different regions: the proportion was only 2% in Hessen and 3% in Bradenburg, while in the Polish-inhabited region around Posen (Wrocław) one-third of the males (32%) could not even read.

For lack of relevant researches, there are no statistical data on literacy available from the Austrian provinces of the Habsburg Empire in the sixteenth to eighteenth centuries. After the modern census had been introduced, the available data testify to huge discrepancies between the provinces, dwarfing the differences between the various French or Prussian regions. In 1900, in the most developed provinces of the Habsburg Empire, in Upper and Lower Austria, and in Bohemia and Moravia, the rate of illiteracy in the population of 10 years of age and above was less than 4%. In the same time, in Galicia 56%, in Bukovina 66%, and in Dalmatia no less than 73% of the population above 10 was illiterate.

There are ample sources in Italy on the spread of education and culture and on schools in the early modern period; however, except for the provinces conquered by Napoleon, no statistics are available about literacy before the introduction of censuses in the second half of the nineteenth century. In the vicinity of Turin in 1710, excluding the town itself, 94% of brides and 81% of bridegrooms were illiterate; by 1790, the percentages had decreased to 70 and 35 respectively. In the Neapolitan kingdom in Campania province, in the mid-eighteenth century, 76% of those submitting tax returns could put only their cross on the document, even though only the well-off, and then only men, were expected to prepare such a return at all. The marked difference between the northern and the southern provinces is evident even from this single piece of information. Censuses reveal, on one hand, that in the whole of Italy in the late nineteenth century, illiteracy, which earlier had reached astonishingly high proportions, was rapidly diminishing; on the other hand, however, the gap between northern and southern Italy was huge. It had not disappeared by the early twentieth century – in 1901 in the northern region of Piedmont 18% of the population above six was illiterate, in Lombardy 20%, in Veneto 28%, while in Sicily the figure was 71%, in Sardinia 68%, and in Calabria 79%.

In Poland, we can observe huge regional differences, too. In the Voivodeship of Cracow at the end of the sixteenth century, the landed nobility was largely literate: 96.7% of the nobles who bore the title *generosus* and usually owned several villages, could sign their name, and the percentages remained unchanged over the next two centuries. Among women, however, literacy rose rapidly in the same period: while in the late sixteenth century, 51% were unlettered, by the seventeenth century only 22% remained so. Women continued to catch up with men in the eighteenth century, when a mere 15% put a cross on documents instead of signing their names. In the lower ranks of landed gentry or *nobilis*, who owned one or at most two villages, the proportion of illiterates made up 24% in the sixteenth century and 16% in the following century. In another study, the literacy of Polish noblemen was examined on the basis of tax returns submitted in and around Cracow. In 1564–64, 43% of the noblemen were illiterate. The vast Polish kingdom, however, comprised regions with very different levels of development. In the Voivodeship of Sandomierz in the seventeenth century we find circumstances very similar to those of the Voivodeship of Cracow: 99% of the noblemen bearing the title of *generosus* and 75% of the *nobilis* could write. On the other hand, the Voivodeship of Lublin lagged behind: while those titled *generosus* were almost entirely literate, only 65% of the women could write. Of the *nobilis*, compared to the 84% in and around Cracow, only 70% signed the documents, while barely 1 out of 10 women could write their names. Even this proved to be a highly developed state of affairs when compared to Eastern Galicia, where 8% of the magnates (*magnifici*), 11% of the *generosi* and 72% of *nobili* put a cross instead of their names as late as the early eighteenth century. In 1764, two-thirds of the participants of the diet called *sejm* held here could not authenticate documents by signing them.

In Vas county, Hungary, in the eighteenth century, only 2.5% of peasants could sign their names, and among the lower strata of the gentry only 26% could sign their names with 74% putting only a cross on documents. The first census to give an overall picture about the literacy of the population in the Hungarian kingdom was carried out in 1870. At this time three-fifths (58%) of the population of above 6 years of age in the historical Hungarian kingdom (part of the Habsburg Empire, consisting of Hungary, Transylvania, and Croatia) could neither read nor write. The figure for Hungary alone was 51%, 78% in Transylvania, and over four-fifths of the population (84%) in Croatia and Slavonia. There were huge differences between the various nationalities, however, in 1890, 63% of Germans and 53% of Hungarians could read and write, but only 9.7% of Ukrainians living in Hungary, 14% of Romanians, and 30% of Serbians. If Central Europe was lagging very much behind the developed Western parts of Europe, the difference between Central Europe and Russia, and the countries of the Balkan peninsula, was as huge.

No reliable data are available on the measure of illiteracy in Russia for the sixteenth to eighteenth centuries, but Jeffrey Brooks, historian of Russian literacy, gave his 1977 work on the period between 1861 and 1917 a telling title: *When Russia Learnt to Read*. At the 1897 census, only 21–29% of men and 13% of women could read. Literacy attained the highest proportions in the Polish provinces of the empire, even though this favorable position was relative indeed: 29% of men and 23% of women had a passive knowledge of the letters. In the European part of Russia, 28% of men and 10% of women could read, while the relevant percentages in Siberia were only 16 and 3 respectively. Development in the 50 years preceding World War I was rapid; however, while in 1880, four-fifths of new

conscripts did not know the letters, in 1913, only 32% of these young men, who were just above school age, were illiterate.

The line of the River Danube and the Carpathian Mountains, which separated Hungary from the Balkan Peninsula, was also a demarcation line between two worlds as regards literacy. Missionaries and travelers in seventeenth-century Moldavia and Wallachia noticed that hardly anybody knew how to read or write, and even some of the Greek Orthodox bishops could read only with difficulty. 'Of the nobles and the tradesmen, few know the Ruthenian or the Slav script,' and even the majority of priests were illiterate according to the Jesuits, although they may not have been impartial. The population of Serbia was largely unlettered as late as the early nineteenth century. Hardly any of the leaders of the Serbian uprising could read or write; the ruling Prince Milos was probably the only illiterate head of state in Europe at the time. According to contemporaries, when Serbia became independent, neither the priests nor the monks could read and write. There was little improvement in this situation in the following decades: even the 1866 census found that 96% of the population was illiterate; at the turn of the century, the number was 83%. The Balkan Peninsula remained very much a region of illiteracy even at the turn of the twentieth century: around 1900, of the population above six, 72% in Bulgaria, 60% in Greece, 78% in Romania, and 87% in Bosnia were illiterate.

For these parts of Europe, literacy remained very much a problem of the twentieth century – the Bolshevik Party launched huge literacy campaigns after coming to power in Russia. In North America, there were very high levels of literacy for New England – male literacy rose here from 70% in 1710 to 90% at the end of the eighteenth century. However, as these data were calculated on the basis of the wills of the wealthy, doubts must be raised about this very high proportion of literates. The 1840 census showed huge differences within the country: around 4% illiterate in the state of New York, compared with 20–28% in Georgia, North Carolina, Tennessee, and Arkansas. However, the African-American population showed high levels of illiteracy even after slavery had been abolished: in 1870, some 80% of adult blacks were illiterate and the figure was 44% in 1900. Massive immigration into the USA from Europe before the World War I, imported mass illiteracy: in 1909, 99% of Danish or Scottish workers in the United States were literate, but only 45% of Portuguese and 43% of Polish workers could read and write at this time.

In other parts of the world, literacy is a problem of today and tomorrow. Around 1960, the time of dissolution of the colonial empires, 81% of the population of Africa, and 55% of Asia's population were illiterate. According to various estimates, at this time roughly 40% of the world's inhabitants still could not read or write. There remains much to do in 1992 the illiteracy figures for Burkina Faso were 86%, in Bhutan 85%, and in Afghanistan 76%. Illiteracy remains a problem for mankind even in the third millennium.

See also: Cultural History; Education, History of; Illiteracy, Sociology of; Literacy Education; Literacy and Orality; Media and History; Social History; Social Inequality in History (Stratification and Classes).

Bibliography

Arnove-Harvey, R.F. (Ed.), 1987. National Literacy Campaigns. J Graff, London/New York.
Brooks, J., 1985. When Russia Learned to Read. Literacy and Popular Literature 1861–1917. Princeton University Press, Princeton, NJ.
Blaak, J., Jackson, B., 2009. Literacy in Everyday Life: Reading and Writing in Early Modern Dutch Diaries. Brill, Leiden.
Baldwin, Y.H., 2006. Cora Wilson Stewart and Kentucky's Moonlight Schools: Fighting for Literacy in America. University Press of Kentucky, Lexington.
Cavallo, G. (Ed.), 1977. Libri e lettori nel medioevo. Rome.
Chartier, Roger (Ed.), 1985. Pratiques de la lecture. Marseille.
Chartier, R. (Ed.), 1987. Les usages de l'imprimé. Paris.
Cipolla, C.M., 1969. Literacy and Development in the West. Harmondsworth, UK.
Cressy, D., 1980. Literacy and the Social Order. Reading and Writing in Tudor and Stuart England. Cambridge University Press, Cambridge.
Darnton, O., 1971. Reading, writing and publishing in eighteenth century France. Daedalus.
Furet, F., Jacques, O., 1977. Lire et écrire: l'alphabétisation des Franais de Calvin à Jules Ferry Paris.
Goody, J. (Ed.), 1968. Literacy in Traditional Societies. Cambridge University Press, Cambridge.
Graff, H.J., 1987. The Legacies of Literacy. Bloomington, IN.
Graff, H.J., Mackinnon, A., Sandin, B., Winchester, I., 2009. Understanding Literacy in Its Historical Contexts: Socio-cultural History and the Legacy of Egil Johansson. Nordic Academic Press, Lund.
Greer, J., 2003. Girls and Literacy in America: Historical Perspectives to the Present. ABC-CLIO, Santa Barbara, CA.
Houston, R., 1985. Scottish Literacy and the Scottish Identity. Illiteracy and Society in Scotland and Northern England 1600–1800. Cambridge University Press, Cambridge.
Houston, R., 1992. Literacy in Early Modern Europe: Culture and Education (1500–1800). London.
Houston, R.A., 2002. Scottish Literacy and the Scottish Identity: Illiteracy and Society in Scotland and Northern England, 1600–1800. Cambridge University Press, Cambridge.
Johansson, E., 1981. The history of literacy in Sweden. In: Graff, H.J. (Ed.), Literacy and Social Development in the West. Cambridge University Press, Cambridge.
Krätli, G., Lydon, G., 2011. The Trans-Saharan Book Trade: Manuscript Culture, Arabic Literacy and Intellectual History in Muslim Africa. Brill, Leiden.
Lurie, D.B., 2011. Realms of Literacy: Early Japan and the History of Writing. Harvard University Asia Center, Cambridge, MA.
Marchesini, D., 1992. Il bisogno di scrivere. Usi della scrittura nell'Italia moderna. Rome, Italy.
McKitterick, R., 1992. The Uses of Literacy in Early Medieval Europe. Cambridge University Press, Cambridge.
Muchembled, R., 1978. Culture populaire-culture des élites dans la France. Paris.
Parker, G., 1980. An educational revolution? The growth of literacy and schooling in early modern Europe. Tijdschrift voor Geschiedenis 93.
Pelizzari, M.R., 2000. La penna e la zappa. Alfabetizzazione, culture e generi di vita nel Mezzogiorno moderno. Salerno, Italy.
Ruwet, J.Y.W. (Ed.), 1978. L'analphabtisme en Belgique. Louvain, Belgium.
Sanderson, M., 1972. Literacy and social mobility in the industrial revolution in England. Past and Present 56 (1), 75–103.
Schofield, R.S., 1981. Dimensions of illiteracy 1750–1850. In: Graff, H.J. (Ed.), Literacy and Social Development in the West. Cambridge University Press, Cambridge.
Spufford, M., 1979. First steps in literacy: the reading and writing experiences of the humblest seventeenth-century spiritual autobiographers. Social History 4, 407–435.
Stone, L., 1969. Literacy and education in England 1640–1900. Past and Present 42.
Tóth, I.G., 2000. Literacy and Written Culture in Early Modern Central Europe. Budapest, Hungary.
Tóth, I.G., 2000. Literacy and Written Culture in Early Modern Central Europe. Central European University Press, Budapest, Hungary.
Vincent, D., 2000. The Rise of Mass Literacy: Reading and Writing in Modern Europe. Polity Press, Cambridge, UK.
Van de Woude, A.M., 1980. Der alfabetisering. In: Algemeine Geschiedenis der nederlanden, vol. 7. Haarlem, Holland.
Wyczanski, A., 1974. Alphabetisation et structure sociale en Pologne au XVIe siecle. Annales: Économies, Sociétés, Civilisations 29 (3), 705–713.

Literacy and Orality

Wolfgang Raible, Freiburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Orality and literacy are dialectical concepts, the meaning of one term depending on the assessment of the other one. Both have a medial and a conceptual side. While mediality is trivial, the conceptual aspect is best conceived of as a continuum, thus blurring any clear-cut distinction between orality and scripturality. The scale corresponds to a series of ever more complex and demanding textual genres.

In order to unfold its immense creative and changing potential, script has to meet the appropriate cultural – e.g., institutional – conditions. The respective processes of social transition and change tend to be slow.

One of them is the enrichment of alphabetic script by an increasing number of ideographic features and a layout taking profit of the two dimensions of the written page. In mathematics, this resulted in a revolution in the seventeenth century. Informatics, one of its more recent offshoots, leaves deep imprints in actual everyday life. Recently, it made possible the Internet that relies heavily on script, integrating other media as well.

Alphabetic script was and is also a most powerful metaphor in Western thinking, leaving deep imprints, e.g., in molecular biology.

The increase in literacy characteristic of our modern societies may prompt nostalgic attitudes hailing former states of presumably authentic orality.

Definitions

Concepts that are mutually dependent are called dialectical. This label applies to 'orality' and 'literacy': 'literacy' cannot be conceived of without 'orality,' 'orality' not without 'literacy.' Hence, their respective meanings depend on how the opposite notion is assessed: if orality is given a positive value, literacy appears as negative, and vice versa.

To give some examples: Marshall McLuhan (1962) started from an oral society which he thought was originally closed, attached importance to the spoken word, and had relatively fixed social roles. He opposed this idealized form of community to Western literate societies which, having transformed oral language into a visible code, were seen as depreciating the spoken word, practicing division of labor, compelling us to cope with different social roles. In this context, a particularly negative influence is attributed to the invention of printing. Whereas Jack Goody and Ian Watt (1962) saw the transition from orality to literacy more in the sense of a profit-and-loss account, Walter J. Ong (1982) took an opposite stance underlining the positive aspects of literacy, thus suggesting to us the positive kind of innovation McLuhan was only willing to attribute to the New Age of television which, in his view, takes already the place of literacy (Raible, 1994).

The rash and far too general judgments we can find in recent and even in current scholarly discussion suggest a more cautious attitude. The first question to be asked in the present context is: what do 'orality' and 'literacy' mean? Here we have to tell a 'medial' from a 'conceptual' or 'cognitive' aspect: the medial one is trivial – something we utter is either spoken or written – whereas the conceptual side is a continuum, intertwining orality and literacy. This 'conceptual' side needs some more consideration.

Communication by language is based on social convention. Hence, the utterances we engage in or exchange with others always reflect a socially accepted activity: somebody has to be 'welcomed'; we engage in 'small talk' with her or him, are asking 'questions,' giving 'answers,' making 'jokes,' engaging in 'gossip,' etc. Some more privileged persons 'pass judgment' on somebody, make an 'injunction,' pronounce a 'verdict,' or proclaim an 'encyclical.'

The highlighted terms in the preceding paragraph describe social activities accompanied – or even brought into being – by the use of language ('speech acts'). They represent a continuum of text types that can best be arranged on a scale. Its one end can be termed 'interactive text,' the other one 'edited text' (Biber, 1988); other labels are 'conceptually oral' and 'conceptually written,' i.e., highly planned and edited text (Koch and Oesterreicher, 1994). For a similar distinction, Karl Bühler (1934) had already introduced a pair of Greek terms used by Wilhelm von Humboldt in the nineteenth century: language as 'enérgeia' (activity) versus language as 'érgon' (the result of activity: a work), i.e., language as an activity or process (involving present partners), as opposed to language as something created, produced, edited, for instance in the form of a book. What is increasing from one end of this scale to the opposite one is, among other things, the degree of planned structuring and, at the same time, the degree of intersubjectivity. Conceptually scriptural utterances should be accessible to a larger and less specified public, not just to the person I am speaking to.

The implications of the conceptual scale can best be shown within a three-level approach to 'language' (**Table 1**):

Each one of levels (a) to (c) should be conceived of as a scale, reflecting exactly the conceptual and cognitive scale holding between orality and literacy: on level (b), we are confronted with 'text types' showing an ever increasing degree of editing and planning. Stylistics and rhetoric belong to this level, rhetoric being, indeed, the art of planning effective texts in successive steps. We start with a general idea and a global strategy; then follow linear organization and memorizing; even the delivery is planned.

Table 1 The conceptual continuum and its exemplification on three levels of 'language'

	Cognitive scale		
	Wilhelm von Humboldt		
	Enérgeia	← →	*Érgon*
	Peter Koch and Wulf Oesterreicher		
	Conceptually oral	← →	*Conceptually written*
	Douglas Biber		
	Interactive text	← →	*Edited text*
(a) 'Language' as utterance	e.g., chat with a neighbor	← →	e.g., papal encyclical
(b) 'Language' as speech act types	e.g., small talk, gossip, joke	← →	e.g., novel, textbook, judgment
(c) 'Language' as the system	enables dialogue, etc.	← →	enables judgment

On level (a), we find the respective realizations, i.e., 'utterances' occupying – like small talk – one end of the scale, or, in case of a judgment or a patent specification, the other one. (c) is the level of the corresponding 'systemic prerequisites': composing texts belonging to elaborate text types demands a much higher mastery of language than buying bread or joking with our neighbor – witness legal texts: a French judgment comes as one single sentence with up to ten degrees of syntactic embedding (Krefeld, 1985). As we may observe in diachrony, language systems develop according to the communicative requirements they have to meet (Raible, 2001b).

Hence, conceptually speaking, orality and literacy are thoroughly intertwined. The often-postulated 'great divide' between orality and literacy, between 'oral' and 'literate' societies, simply does not and cannot exist (Hornberger, 1994).

Transitional Aspects

The mutual relationship between orality and literacy holds all the more as the existence of a script does not change or achieve anything by itself. Beyond doubt there is an immense creative and changing potential in literacy. But in order to unfold it, literacy has to meet the appropriate cultural and institutional conditions, above all the existence of an adequate education system (Scribner and Cole, 1999).

There are, nevertheless, some evolutionary tendencies. One of them is that literate societies are likely to develop, starting from the "communicative genres of everyday life" (Knoblauch and Luckmann, 2004 in the aftermath of Alfred Schütz), a far greater number of textual genres, along with new – e.g., legal – institutions. This enlarges the conceptual space shifting considerably the societal achievements of conceptual scripturality (and augmenting the time it takes us to acquire them). There will be neither essays, nor editorials, patent specifications, or testaments in an oral society (e.g., Biber and Vásquez, 2007; Bazerman, 2012; Bakhtin 1953/1986 is fundamental.).

The proliferation of textual genres accompanied by the extension of the conceptual scale automatically requires new possibilities on the systemic side. This is shown, among other things, e.g., by the fact that a considerable number of English speech act verbs appear for the first time in Early New English: 'to acknowledge,' 'to advocate,' 'to assert,' 'to concede,' 'to remind,' 'to apologize,' 'to question,' 'to request'; or even only in New English: 'to remark,' 'to retort,' 'to state,' 'to accept,' 'to guarantee,' 'to volunteer' (Traugott, 1987). Together with the speech act types already mentioned, these examples enhance at the same time institutional aspects of literacy.

One further general tendency is the slowness of all processes bringing about cultural change. To say it with a concept of Fernand Braudel (1902–85), they tend to be phenomena of the 'longue durée' (long-term). Some examples shall illustrate this.

Alphabetic Writing, Orthography, and the Invention of Printing

There are two kinds of writing systems: those representing formal aspects of speech ('syllabic,' 'alphabetic script'), and those rendering content ('ideographic script'). Writing proper names is a task that regularly leads to syllabic or even alphabetic elements in ideographic script (witness Egyptian or Maya hieroglyphs), whereas alphabetic script tends to become ever more ideographic. Alphabetic script with its reduction of written language to a series of 20–30 graphemes is an elegant invention. But whereas words or syllables are units immediately audible, the phonemes corresponding to graphemes are not. This leads in numerous cases to a developmental reading disorder, or dyslexia, a condition unknown in this form in ideographic or syllabic script.

By reflecting essential aspects of spoken language, early Greek script is characteristic of early alphabetic writing in general: since we do not hear pauses between words, 'scriptio continua' (continuous script) is, with some exceptions, a quite natural outcome. It prevails in Western texts up to the eighth century AD and is necessarily linked with reading out aloud.

By 1200, all the achievements of what most of us would call modern layout are present in scholastic texts: spaces between words, punctuation, capitals at the beginning of a new sentence, paragraph indention, emphasis by means of different colors and different script (all of them ideographic features), chapter headings, short summaries in the margin, footnotes, a table of contents, and alphabetic registers. As a result, the reader is not lost in an amorphous text. It is only at this time that the general practice of silent reading can begin (Parkes, 1976, 1992).

Such developments are due to a feedback loop: the higher the number of readers, the more will their needs be taken into account by the invention of features facilitating the reading process – as are ideographic elements. This explains why such phenomena can be observed at the end of the twelfth century: it was not until then that lay literacy developed to any significant extent.

By this time, we observe an incredible increase in the production of written documents. In the State Archives of Siena, there are for instance about 500 documents from the twelfth century, but nearly 17 000 for the thirteenth century. This acceleration of text production coincides with major changes in social, commercial, and agricultural organization. Thanks to doings and dealings, the city states of Northern Italy – a similar evolution can be observed in large parts of central Europe (for

England: Clanchy, 1993) – had already a population surplus. This compelled them to organize life and everyday supply very strictly by written statutes.

As a consequence, there was an immense need for knowledge: how shall I write a letter, make a speech, a treaty, how do I behave well, in short: 'how to win friends and influence people?' This was the time when, at Bologna, Roman Law was rediscovered; when the great Encyclopedias of the Middle Ages were written, when Italian merchants came to invent paper money as well as the new juridical forms of joint stock companies and insurances. In Art and Science, this was the epoch of Gothic architecture and Scholasticism (1100–1450), mediated by the contact with Arabic culture and the discovery of Aristotle. In this 'longue durée' perspective, the invention of printing presents itself much more as a necessary consequence than as an "agent of change" (Eisenstein, 1979) in the domain of literacy.

There was, nevertheless, at least one topic fostered by printing: discussion on orthography. On the one hand, printing books for a great variety of unspecified readers inspires thoughts on standardization. On the other, it goes without saying that all the achievements made in writing and layout increase the intellectual and the practical effort of the writer. This is why discussions on orthography tend to be dominated, since the sixteenth century, by elementary school teachers who would like to reduce – for those learning to write – the difficulties brought about by the ideographic tendency under discussion (Strobel-Köhl, 1994).

Material Aspects of the Written Text

The evolution of layout has an aspect that merits some more attention. Unlike one-dimensional spoken language, written text comes in two dimensions. Layout implements the potential of this two-dimensionality by making visible chapters, paragraphs, and so on.

The possibilities given with two-dimensionality are not limited to layout, though: Its potential enabled, among other things, modern mathematics. Thirteenth-century mathematicians were far from being familiar with the system of notation their successors use today. This system only took shape, between the thirteenth and the seventeenth century, in one of the transformation processes whose slowness was already mentioned. Most of today's symbols are the outcome of an intermediate stage presupposing alphabetic literacy: the stage of abbreviation. It leads not only to most familiar signs like '&' ('et'), but also to the plus and minus ('+', '−'), the symbol for the square root, etc. Eventually, it was Descartes who, at the beginning of the seventeenth century, introduced the modern conventions holding for the symbolization of the unknown element. All of this facilitated the highly momentous discoveries made during the seventeenth century: analytic geometry and calculus, resulting in a new ideographic system (Raible, 1991).

Mathematical ideograms, consequently exploiting the two dimensions of the page, are accessible to simultaneous rather than to linear perception – provided one knows the system. This enormous progress made mathematics the most important complementary discipline to natural sciences, contributing in a decisive way to their role in our modern world (Krämer, 1991). This is why, at the beginning of the seventeenth century, Galileo Galilei tells us that the *Book of Nature* is written in cipher and that those intending to read it have to master the language of mathematics.

In recent times, this role has partly passed to a new branch of mathematics, informatics. Texts written in programming languages are at least as demanding as complex written texts in natural languages (Raible, 1999). At the same time, their influence on modern life is pervasive. Retrospectively, the second part of the twentieth century is likely to appear as the age of Claude E. Shannon (1916–2001), among others the father of information theory.

The Written Text as a Metaphor

The use of written language as a metaphor has a long history, too – think of the *Book of Nature* (Blumenberg, 2007[1986]). The metaphorical use starts with the early atomists, Leucippus and Democritus (fifth century BC). Their basic idea was that the complex, variegated, and manifold world surrounding us is nothing but appearance, whereas in reality all was thought to consist of atoms and the void between them. According to what Aristotle tells us in his *Metaphysics* (985b15ff.), their visible model for the invisible structure of matter was alphabetic script. The diversity of the visible world would be due to the fact that the atoms are differently shaped – just as an A differs in shape from an N; that their order may be different – as the sequence AN is different from NA; finally, that their relative position in space may differ: a rotation of 90° makes an N out of a Z.

The central idea underlying this conception is to reduce multifold diversity onto a restricted set of elements (here the 20-odd letters of the Greek alphabet) making up for a possibly infinite number of written texts.

While the *Book of Nature* became a powerful and influential metaphor since the Middle Ages, the idea of the alphabet as the basis of a code reappears at the end of the nineteenth century in cell biology (Friedrich Miescher, 1844–95). It was again put forward in 1943, this time by the Austrian physicist Erwin Schrödinger (1887–1961), in a series of lectures given under the heading "What is life?" where he suggested a genetic alphabet similar to the Morse code. His suggestion was confirmed in 1953 by Francis Crick and James Watson, who showed that the long strands of DNA (Schrödinger's punched Morse tapes) are structured as a double helix, and by the series of momentous discoveries following this breakthrough.

Since then, the metaphor of written language in its alphabetic form is pervasive in molecular biology. The four nucleotide bases abbreviated by A, T, G, and C were called 'letters of the genetic alphabet.' RNApolymerase is 'reading' DNA sequences with their 'reading frames.' This process is called 'transcription,' and it happens thanks to 'transcription factors'; transcription is followed by an immediate process of 'proofreading.' The result is called a 'copy,' subject to further 'editing.' The resulting string of mRNA will be 'translated' into a polypeptide. This is made possible because the triplets of DNA 'encode' or are 'coding for' amino acids. The whole process is called 'gene expression.' It presupposes a 'grammar of biology.'

As the genome of humans and of other species is being 'deciphered,' the result is stored in large databases modeling the

sequences of nucleotides as sequences of the letters A, T, G, and C. In this context, molecular biologists are speaking of 'gene libraries.' Recurrent sequences of nucleotide 'letters' as well as recurrent sequences of amino acids in proteins are called 'motifs,' and the genome itself is seen as an 'encyclopedia' (Raible, 2001a).

Some Other Transition Processes

One of the disadvantages of 'oral' cultures is what was called a 'collective censorship.' Narratives the audience dislikes fall into oblivion. In literate societies all written texts can, in principle, be preserved and rediscovered by future generations. This advantage changes into a serious problem with mass media, including the Internet – this is not least a true consequence of printing.

Unless the public simply refrains from reading, there are at least three reactions. Long texts are transformed into shorter (and cheaper) versions. A lot of these text types were already created in Antiquity or in the Middle Ages, for instance 'epitome,' 'summarium,' 'argumentum,' 'catechismus,' 'breviarium,' and 'compendium.' Other genres create a new text out of parts of others (e.g., 'anthology'). Modern literary scholars cannot but use one of the vernacular versions of Valentino Bompiani's 'Dizionario delle Opere di tutti i Tempi e di tutte le Letterature' (Dictionary of the Works of all Times and Literatures).

A second reaction is an institutional censorship as embodied by all systems of education. They make use of specific 'curricula,' 'syllabi,' 'textbooks,' 'introductions,' and 'primers' in order to transform youth into fully fledged members of society. It will not come as a surprise that this kind of selection is subject to permanent discussion.

Making texts as attractive as possible, alongside with making their content easily accessible, is a third answer to the problem. The most radical changes in writing can be observed in the history of mass media, i.e., in journalistic writing and in the corresponding genres of television. Scientific writing is another good example: today, articles published in journals are structured according to an identical model and endowed with tables, schemes, pictures, whose captions and legends, together with the obligatory 'abstract' of the article itself, swiftly inform the readers (Bazerman, 1991; Gross et al., 2002).

The last point shows one of the problems, though: facilitating reception holds only for those who are already experts in the matter, with the cognitive demands of such texts increasing at the same time. Nevertheless even this strategy does not solve the basic problem. Thanks to specialization and to the accelerated growth of science, the number of scientific journals and articles published daily is so vast that we simply cannot cope with it. We should have read the relevant texts in order to know at least the State of the Art, though, before publishing, e.g., the findings of a new experiment. As a result, formerly well-established facts may fall into oblivion and are rediscovered a 100 years (or even more) later.

The presentation of information in the Internet repeats an evolution that took place in the layout of written text. Witness the plethora of new elements in subsequent releases of hypertext markup language (HTML or XML – they correspond to the third level in the above scheme) and the introduction of metatags leading, for web crawling machines, to the equivalent of subject catalogs in our libraries. The web pages of the Internet basically exploit the achievements of literacy (in traditional texts, footnotes, page references, a table of content, etc. correspond to 'links,' a major difference being only the quick availability of the objects referred to). New features are the availability of color, sound, and moving items.

At the same time, the Internet shows once again the potential inherent in a new medium, this time with a potentially unlimited communication radius. Thanks to the Net, the collaborative undertaking Wikipedia became, over a period of only 10 years, a most important source of encyclopedic information. Whoever has used an online dictionary will scarcely feel a need to revert to the printed version. Email, SMS, and Skype increasingly supersede phone calls. Chat rooms have more often than not become places where we can play with language thanks to ingenious abbreviation techniques (Strätz, 2010). Internet forums or the possibility to comment on newspaper articles show the double-edged nature of new forms of communication, though: on the one hand we can react almost immediately, on the other we tend to give way to unfiltered emotions and prejudices, do not care about thorough information, let alone orthography, our anonymity fostering even insults we would rather not make in front of a visible addressee.

A Precarious Equilibrium

There is no great divide, then, between 'oral' and 'literate societies,' orality and literacy being intertwined on the conceptual level. In addition to this, there is another factor of cultural variation: a social community can dispense with script, not with spoken language. Apart from entirely oral cultures, we are thus always confronted with more or less literate societies. Given the dialectical relationship between orality and literacy, this means a steadily changing state of equilibrium.

There are indicators showing us that the precariousness of this equilibrium is felt. As the evolution of culture in general may lead to the slogan 'back to nature,' the increase in literacy may make us long for the lost paradise of 'innocent' orality. This desire may materialize in a motto like "let us write the way we speak" (Bader, 1994). First examples were Plato's critique of literacy ('Phaidros' 275d) and the effectiveness of highly planned Sophist rhetoric giving rise to suspicion.

Anyhow, the consequence of such reactions is even more literacy. Since Antiquity, theoreticians of rhetoric know that among the genres of speech 'ex improviso speech' is the most artful one. Baldassar Castiglione (1478–1529) whose Courtier was a model for Europe since Italian Renaissance, tells us that true art means dissimulation of art. The same thing holds for the impression of orality called 'skaz' (derived from *skazat*, 'to tell') that was discovered and recommended by some of the Russian formalists and refers to oral characteristics showing up in written literary texts.

What at first sight might appear as paradoxical is quite obvious on closer inspection. Given the dialectical relationship, in a literate society orality has to be artificially created – just as seventeenth century authors conceive of nature – in literature and in the Fine Arts – as of 'the negation of the negation of nature.' Once a culture has adopted literacy, returning to 'authentic' orality is impossible.

See also: Dyslexia, Developmental; Dyslexias, Acquired and Agraphia; Illiteracy, Sociology of; Literacy Education; Literacy and Illiteracy, History of.

Bibliography

Bader, Eugen, 1994. Rede-Rhetorik, Schreib-Rhetorik, Konversationsrhetorik. Eine historisch-systematische Analyse. Tübingen: Narr.

Bakhtin, Mikhail M., 1986[1953]. The problem of speech genres. In: Emerson, C., Holquist, M. (Eds.), Speech Genres and Other Late Essays. University of Texas Press, Austin.

Bazerman, Charles (Ed.), 1991. Textual Dynamics of the Professions: Historical Analysis and Contemporary Studies of Writing in Professional Communities. University of Wisconsin Press, Madison, Wisconsin.

Bazerman, Charles, 2012. 'Genre as social action'. In: James Paul Gee and Michael Handford (Eds.). The Routledge Handbook of Discourse Analysis. Routledge, London, pp. 226–238.

Biber, Douglas, 1988. Variation across Speech and Writing. Cambridge University Press, Cambridge.

Biber, Douglas, Vásquez, Camilla, 2007. In: Bazerman, Charles (Ed.), Writing and Speaking. Handbook of Research on Writing: History, Society, School, Individual, Text. Erlbaum, New York, pp. 657–672.

Blumenberg, Hans, 2007[1986]. Die Lesbarkeit der Welt. Suhrkamp, Frankfurt.

Bühler, Karl, [1934]1990. Theory of Language: The Representational Function of Language. (Original of 1934. Donald Fraser Goodwin, Trans.) Benjamins, Amsterdam and Philadelphia.

Clanchy, Michael T., 1993. From Memory to Written Record: England 1066–1307. Blackwell, Oxford.

Eisenstein, Elizabeth L., 1979/2009. The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early modern Europe, II vols. Cambridge University Press, Cambridge.

Goody, Jack, Watt, Ian, 1962/1963. The consequences of literacy. Comparative Studies in Society and History 5, 304–345.

Gross, Alan G., Harmon, Joseph E., Reidy, Michael, 2002. Communicating Science: The Scientific Article from the 17th Century to the Present. Oxford University Press, Oxford.

Günther, Hartmut, Ludwig, Otto (Eds.), 1994. Writing and Its Use, vol. I. De Gruyter, Berlin.

Hornberger, Nancy H., 1994. Oral and Literate Cultures. In: Günther, H., Ludwig, O. (Eds.), Schrift und Schriftlichkeit. Writing and Its Uses, vol. I. Berlin, New York, pp. 424–432.

Knoblauch, Hubert, Luckmann, Thomas, 2004. Genre analysis. In: Flick, Uwe, et al. (Eds.), A Companion Tool to Quantitative Research, Ed. by London, etc. Sage, pp. 303–307.

Koch, Peter, Oesterreicher, Wulf, 1994. Schriftlichkeit und Sprache. In: Günther, Hartmut, Ludwig, Otto (Eds.), Schrift und Schriftlichkeit. Ein internationales Handbuch / Writing and Its Use. An International Handbook, vol. I. Berlin, New York, pp. 587–604.

Krämer, Sybille, 1991. Berechenbare Vernunft: Kalkül und Rationalismus im 17. Jahrhundert. de Gruyter, Berlin.

Krefeld, Thomas, 1985. Das französische Gerichtsurteil in linguistischer Sicht. Zwischen Fach- und Standessprache. Lang, Frankfurt, Bern.

McLuhan, Marshall, 1962. The Gutenberg Galaxy. The Making of Typographic Man. The University of Toronto Press, Toronto.

Ong, Walter J., S.J., 1982. Orality and Literacy: The Technologizing of the Word. London; Methuen, New York.

Parkes, Malcolm B., 1976. The influence of the concepts of ordination and compilation on the development of the book. In: Alexander, J.J.G., Gibson, M.T. (Eds.), Medieval Learning and Literature. Clarendon Press, Oxford, pp. 115–141.

Parkes, Malcolm B., 1992. Pause and Effect: An Introduction to the History of Punctuation in the West. Scolar Press, London.

Raible, Wolfgang, 1991. Die Semiotik der Textgestalt. Erscheinungsformen und Folgen eines kulturellen Evolutionsprozesses. Winter, Heidelberg.

Raible, Wolfgang, 1994. Orality and Literacy. In: Günther, H., Ludwig, O. (Eds.), Schrift und Schriftlichkeit. Writing and its Uses, vol. I. Berlin, New York, pp. 1–17.

Raible, Wolfgang, 1999. Kognitive Aspekte des Schreibens. Winter, Heidelberg.

Raible, Wolfgang, 2001a. Linguistics and genetics: systematic parallels. In: Haspelmath, Martin, et al. (Eds.), Language Universals and Language Typology. de Gruyter, Berlin and New York. Handbücher zur Sprach- und Kommunikationswissenschaft vol. 20.1, pp. 103–123.

Raible, Wolfgang, 2001b. Linking clauses. In: Haspelmath, Martin, et al. (Eds.), Language Typology and Language Universals. de Gruyter, Berlin and New York. Handbücher zur Sprach- und Kommunikationswissenschaft vol. 20.1, pp. 590–617.

Scribner, Sylvia, Cole, Michael, 1999. The Psychology of Literacy. Harvard University Press, Cambridge, MA.

Strätz, E., 2010. Sprachverwendung in der Chat-Kommunikation. Eine diachrone Untersuchung französischsprachiger Logfiles aus dem Internet Relay Chat. Tübingen: Narr.

Strobel-Köhl, M., 1994. Die Diskussion um die 'ideale' Orthographie. Das Beispiel der Kreolsprachen auf französischer Basis in der Karibik und des Französischen im 16 und 20. Jahrhundert, Tübingen: Narr.

Traugott, E.C., 1987. Literacy and language change. The special case of speech act verbs. Interchange 18, 32–47.

Literacy Education

Maureen Walsh, Australian Catholic University, and University of Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C.A. Perfetti, R. Sandak, volume 13, pp. 8975–8981, © 2001, Elsevier Ltd.

Abstract

The concept of 'literacy,' while often contentious, has constantly evolved along with social change and technological advances. Even more so in the twenty-first century the unprecedented impact of digital technologies has produced new theories and research around the nature of literacy. We are able to communicate instantly with combinations of text, photos, or videos through mobile technologies and obtain instant information from the Internet. The phenomenon of Web 2.0 has raised levels of online interaction with social changes accompanying different patterns of communication. Due to these changes key questions are raised about literacy education in classrooms. Is reading and writing different in this digital, multimodal, and multimedia environment? How can we work across traditional views of reading and writing within this new communications environment? This article examines new views of literacy that are supported by worldwide research in schools and provides examples of practical articulation between new theories of communication and classroom practice.

Scenario

Susan, a pupil in a Grade 6 class in Sydney, Australia, uploads a digital narrative onto a wiki shared by students in a Year 6 class in Glasgow, Scotland. Susan's fictional narrative is about a family in crisis from the father's gambling addiction. The text has been created after work in her class on the structure of literary narratives along with research into social issues that affect children in families. It is a first-person narrative written from the perspective of the main character, a girl of the writer's age as she witnesses the impact of parental tension and loss of income on the family. The narrative is created on a touch pad with written and visual text, photos, music, and video clips. It is interspersed with factual information that is a result of Susan's research into gambling addiction. Each screen or page of her text displays web links, text boxes, photos, and graphics. The students in Glasgow read Susan's story and give her feedback, commenting on aspects they consider to work well and making suggestions for further improvement in her writing and digital design.

Introduction

This scenario is one small example of a literacy activity in a global classroom made possible with digital communications technology. The increases of digital, mobile media and touch pad devices have been predicted as precursors of technology that may be available in future classrooms, such as "touch screen tables, robot study buddies, and 3D virtual learning environments" (SMH, 2012). Today's students are able to communicate instantly with combinations of text, photos, or videos through smart phone technology, different types of computers, and multimedia devices and obtain instant information from the internet using these devices and their applications. With Web 2.0, social changes have accompanied these technological developments. Students can participate in twittering, wikis, blogs, and various social networking sites or engage in a virtual environment through gaming or joining virtual worlds. These communication environments are changing the way people present themselves and the way relationships are developed. Anticipations are that aspects of Web 3.0 are already occurring (Exa, 2011) in the way data information is integrated and accessed. These changes raise important questions. What do they mean for the notion of 'literacy' as we have understood it for several decades, and what are the implications for literacy education in schools?

Digital communication technologies have affected society so rapidly that educational researchers, policy makers, and teachers have only begun to consider the application of these changes for curriculum design and assessment. The multimedia facilities of digital technologies, particularly mobile hand held devices and touch pads, encourage the processing of several modes simultaneously. Thus reading and writing rarely occur in isolation within such digital communication, yet, traditionally, they are seen as separate processes. Teachers are challenged to adjust their pedagogy to this new environment and balance the incorporation of new modes of communication with those skills of reading and writing that are seen as core for a literate person.

There is often polarization of 'traditional' and 'new' views of literacy but in fact conceptualizations of literacy have always changed even though the term itself was not often used until the late nineteenth century (e.g., Barton, 1994; Christie, 2010) and then mostly in relation to mass elementary schooling. Even if we consider the commonly quoted view of being literate as the 'ability to read and write,' this is a reductive concept that does not allow for 'thinking' and 'meaning' (Green, 2012: p. 4) that has always been part of human communication. Recent terms such as 'new literacies' (Lankshear and Knobel, 2003), 'multimodality' (Kress and van Leeuwen, 2001), and 'visual literacy' (Arizpe and Styles, 2003) have been used in educational contexts, yet each of these refers to ways of interacting that have been part of basic human communication and meaning making for centuries.

There have always been 'new' literacies when new technologies have been invented, whether we consider the printing press, the fountain pen, biro, typewriter, or computer. Each of these physical, technical inventions contributed to sweeping social changes in the century in which each was invented and gave people wider access to communication and education.

Multimodal combinations of language, image, sound, movement, and gesture have always been part of communication between people as well as in more formal ways through theater, dance, and music. Visual literacy has always occurred where images have represented people, activities, and events such as in prehistoric cave paintings, in the complex art of the Egyptians, and in religious art and illustrated manuscripts such as the Bible and the Koran. It is not the purpose of this article to review the history of art, technology, or communication, yet it is interesting to remember that communication has always had a basis in language, art, movement, sound, and music. There have been different times when the combination of these has been more significant when enhanced by technological developments such as the phonograph, photography, film, radios and transistors, television, computers, and multimedia. Such developments along with associated social change have contributed to the organic, ecological nature of literacy.

There have always been new ways of communicating or 'new' literacies; however, as recently stated by Donald Leu (2012), we are currently in an 'epoch' of change that is more sophisticated and complex than ever before. This article provides a brief overview of 'new' aspects of literacy that have been enabled by technological and social changes. They are considered along with what is nominally called the 'traditional' view of literacy, which is based on written language in paper-based texts. In the previous edition of this Encyclopedia, Perfetti and Sendak (2001) presented a detailed and informative review of literacy education. They showed that concepts such as 'universal literacy' and 'functional literacy' vary within different cultures and countries and demonstrated the difference between 'narrow' and 'broader' views of reading (p. 8976). They provided a detailed overview of the relationship of reading to the writing systems of a language and the importance of decoding, particularly phonemic awareness for young children learning to read. These are foundational skills that are essential for literacy education; however, students now require a repertoire of both print and digital literacy practices for their future workplace and life. With the multiple media available for communication in the twenty-first century the potential for literacy education in schools is both challenging and exciting.

Literacy in a Digital Environment

If we consider the scenario at the beginning of this article, we can examine the language and literacy that was needed to achieve the outcome of Susan's work. Susan had to understand the features of a narrative, its purpose, and structure to write a fictional story using a first-person narrator. In addition, she needed different levels of comprehension as she researched, read, and selected relevant information on social problems caused by gambling. These facts were then woven into the themes of her narrative, along with appropriate use of grammar, vocabulary, and correct spelling. These are aspects of 'traditional' literacy that would be expected of a student in the final year of primary school. The traditional skills occurred along with the technical proficiency required for using and producing text on a touch pad, e.g., browsing and selecting websites by touching and tapping hyperlinks; navigating menus and icons; scrolling through text; moving photos from one screen to another; inserting video and music; composing on screen; and designing pages or screens with color, graphics, text, text boxes, and images. These are the combination of traditional with technical skills that have been referred to for some time as 'new' literacies (e.g., Luke, 1995; Street, 1997; Lankshear and Knobel, 2003).

The scenario about Susan demonstrates the integrative potential of digital communication so that it is difficult to separate the physical from cognitive processes, or reading from writing, within new communication practices. There are, however, significant aspects that need to be considered when contemplating the differences between traditional practices of reading and writing and those that are possible with digital communication. Thus reading and writing are discussed separately in the next sections to demonstrate key features that we need to consider for educational implications. Then their interrelationship will be examined within new communication contexts and implications considered for literacy education.

New Ways of Reading

Reading with screen-based texts entails the reading of text that is not usually in a linear format. On-screen reading incorporates multisensory activities such as searching, viewing, browsing, scrolling, hyperlinking, and navigating together with the clicking and scrolling of a mouse (or the touching, tapping activities with mobile, and touch technologies). Other modes that are interpolated along with written text are image, sound, movement, and gesture. Thus the term 'multimodal' has been used in recent years to describe the nonlinear construction of texts that occur primarily on a screen. The theory of multimodality (Kress and van Leeuwen, 2001; Kress, 2003, 2010) has been the basis for the contention that the simultaneous processing of different modes of text, image, sound, and gesture in visual, media, or digital texts is a different behavior from the linear, sequential processing of print-based texts. Several theorists and researchers have examined the nonlinear reading pathways of screen-based texts (e.g., Snyder, 1997; Kress, 2003; Leu, 2000; Unsworth, 2001; Walsh, 2006, 2010) while the 2009 Program for International Student Assessment (PISA) report (OECD, 2011) identifies differences between 'static' and 'dynamic' electronic texts (pp. 27–28) to identify different ways meaning can be processed.

The navigation of screen-based texts frequently involves various aspects of online processing that includes responding to animated icons, hypertext, sound effects, and navigating pathways between and within screens. A UK study (Bearne et al., 2007) provided evidence to show that students across an age range from 3 to 16 years use a range of skills and strategies for reading screen-based texts. Researchers' observations confirmed that students were involved in 'radial browsing.' They were able to trawl language, image, and music as well as highlight key sections to retrieve information. Students of today quickly adapt to the navigation potential and the processing of different modes within digital texts. This processing itself often incorporates a merging and synchronizing of text, images, sound, and movement as these occur in recent digital products, such as

touch phones or touch pads. The touch features of recent products rely much more on gestural, spatial, and kinaesthetic movements that need further investigation in the way this processing is affecting cognitive processes. New research (e.g., Walsh and Simpson, 2013; Rowsell et al., 2013) is examining the impact of touch on cognitive processes when reading and writing with touch pads.

There are important implications for literacy education when considering reading in a digital environment. Students need to be able to decode written language, understand vocabulary, and develop the same types of comprehension that have always been required for reading literary or factual texts. At the same time, they need to decode and understand the structure of different features of digital texts, particularly the nature and purpose of visual, gestural, or audio modes. They need to be able to navigate effectively between varieties of texts and understand the difference between scanning for information and sustained reading. Teachers need to model and scaffold students' reading experiences on and between screens, and recognize the need to build students' metacognitive processing and evaluation of digital texts. At the same time, critical and evaluative skills are needed to discriminate between sites and to detect authenticity and bias in the extensive material on the internet.

New Practices of Writing

Writing on screen has existed for a long time with word-processing facilities but now 'writing' very often entails designing and assembling a product that may contain written text as well as quite sophisticated layout, graphics, photographs, and images. Similarly, a 'writer' can become a 'producer' (Sheridan and Rowsell, 2011) by designing and producing a text that combines images and graphics with written text as well as sound and movement on screen. Moreover, the whole notion of 'authoring' has changed in the way texts are 'mashed' or hybrid texts created. As shown by Susan's text, the facilities of Web 2.0 have further changed the possibilities of students' writing and text production. Her digital narrative uploaded onto a wiki was carefully designed to engage the audience who could respond with text, images, or video.

For students to produce multimodal texts they need to consider and understand features of design such as composition, use of text, and image or graphics – including aspects such as color, size, medium, angles, placement on screen – and the way these are appropriate for a specific audience. It is significant that other researchers have been investigating design as integral to literacy pedagogy (e.g., Kalantzis and Cope, 2005; Healy, 2008).

Implications for literacy education are for teachers to realize that the teaching of writing is much more diverse with the potential for even young students to produce sophisticated products. Teachers need to demonstrate the structures and styles of genres that can be produced as digital texts, for example podcasts, vodcasts, wikis, blogs, and digital narratives. These are able to be augmented with photos, graphics, video, and sound. Thus design is an important aspect of writing and creating. A challenge for teachers is to consider how the use of design can be integrated with the teaching of writing while essential skills of language, grammar, and spelling are maintained.

The Holistic Potential of New Literacy Practices

While considering the differences in both reading on screen and reading print-based texts, reading and writing frequently occur together and are interdependent. As with the scenario of Susan, students are able to research and read information and then apply their learning by producing a response in a digital format. Nor is it sensible to separate the technical, functional processes of reading or writing on screen from the social practices that accompany these processes. The tasks will often involve collaboration and discussion between teachers and students as well as between students. Students are motivated to exchange information and to show their products to peers. This participatory culture is enabled by digital technology and reflects the sharing and exchange of information that occurs now in the wider society.

Thus literacy in a multimodal, digital environment has the potential to be holistic with an interdependency and fluidity between language and literacy functions. **Figure 1** depicts the process of reading and writing in classroom contexts where students work with traditional print-based texts and digital texts. Students can be required to read and write both print and digital texts and will therefore be interacting with a variety of modes. Reading and writing occur within a particular social context and for a specific purpose. They are often interrelated with reading leading to a writing task that may involve the design and production of a multimodal text or digital product for an audience.

The diagram in **Figure 1** shows that reading and writing with print and digital texts are interchangeable resources within classrooms. Making and communicating meaning are still the core purpose of literacy; however, reading and writing have the potential to be communicated through and with a variety of modes. Along with the interaction between modes of written texts, image, sound, and movement there is often interaction between students in this environment so there is an interrelationship between language and literacy activities. Oral language, thinking, and problem solving is an important factor as students plan, investigate, and collaborate to interpret or produce digital texts.

The technologies related to Web 2.0 along with the facilities of touch pad devices ensure the interrelatedness of reading and writing. These often occur along with the viewing and posting of images, the blending of sound, and the constant interchange and connecting of messages. Over a decade ago, emails revolutionized communication, taking over from letter writing, particularly for adults in the workplace. Communication through social networking has taken over from emails for many are being used in the commercial world. In such communication we are reading, writing, and responding asynchronously, although features of new technologies such as *SKYPE* and *Face Time* allow for synchronous communication through oral language, image, and gesture.

We do not know how such processing of messages and texts is affecting the way children learn or if the processes involved in

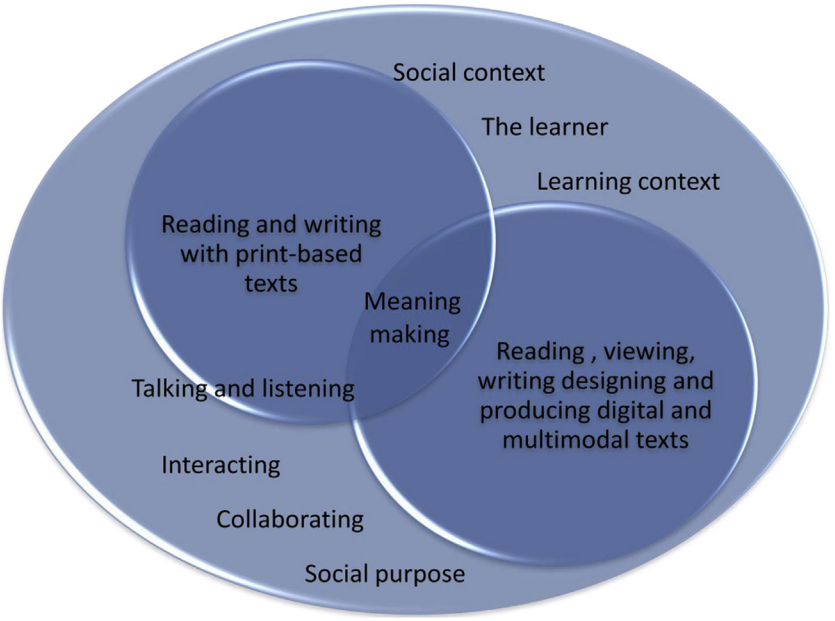

Figure 1 The holistic potential of new literacy practices. Source: Walsh, M., 2011. Multimodal Literacy: Researching Classroom Practice. PETAA, Newtown, p. 13.

activities such as texting, blogging, or communicating online are developing different cognitive abilities than those required for reading and writing traditional print-based texts. Gee's research (2003) on video gaming suggests that the procedures involved offer cognitive advantages with intricate literacy and learning opportunities. The touch features of recent products rely much more on gestural, spatial, and kinaesthetic movements that need further investigation in the way the embodied nature of this processing is affecting cognitive processes.

These unknowns are challenges for education. The study previously referred to (Bearne et al., 2007) on reading screens showed that while students were able to apply aspects of comprehension to obtain screen-based information it was "orchestrating the different modes to make meaning" (p. 20) that was seen as a different process that could not be assessed in the same way as the assessment of reading print-based texts. This is similar to Lemke's description of the 'multiplicative' (Lemke, 1998) effect that occurs when several modes are processed. A further study by Bearne and Wolstencroft (2007) demonstrated ways of teachers programming and assessing writing through students' multimodal texts. Bearne and Wolstencroft's work and classroom research in Sydney, Australia (Walsh, 2006, 2008, 2009, 2010) has emphasized how students need to understand the meaning-making potential of different modes, particularly the relationship between words and images, in reading, writing, and producing multimodal texts. Recent research by Walsh and Simpson (2013, 2014; Simpson and Walsh, 2014) shows how we need to distinguish between the technical skills of using digital technologies and the cognitive processes of interpreting and communicating meaning. To offset an overemphasis on technical skills, educators and researchers need to focus on using technologies together with a metaunderstanding of technology, particularly the way separate and multiple modes impact on meaning making.

Conclusion

The scenario at the start of this article demonstrates the way digital technology can allow communication to occur in a holistic, meaningful way. The implications of this discussion are that reading and writing with twenty-first century texts requires the processing of multiple modes and multiple reading pathways. Yet educational policy and curriculum documents are still based on theories of teaching reading and writing with print-based texts. Standardized national testing is currently based on traditional print-based literacy although there is a recognition of the importance of multimodal literacy in the move toward online testing, such as those developed in the PISA assessments (OECD, 2011). Issues of access and equity complicate attempts to develop curricula that encompass the potential of digital communication. Teachers need to adapt pedagogy so that, while they are teaching traditional skills of decoding and meaning making, they understand and examine the way different modes, especially visual, can engage students in meaning making. Similarly, writing in contemporary contexts requires further skills of design and production. New ways of programming and assessment are needed.

In this article I have raised some of the implications of changes in digital communication for literacy education. An examination of differences and similarities in reading and writing with print-based texts compared with multimodal texts has shown that, while multiple modes may need to be processed for reading and writing screen-based texts, meaning making is still the main purpose of communication. The

facilities of new technologies offer the potential for dynamic literacy learning. However, it is not the technology itself that will create a vibrant, engaging learning environment but the way teachers plan and structure literacy education. The future can be exciting for literacy in twenty-first century schools.

Bibliography

Arizpe, E., Styles, M., 2003. Children Reading Pictures. Interpreting Visual Texts. RoutledgeFalmer, London.

Barton, D., 1994. Literacy. An Introduction to the Ecology of Written Language. Blackwell Publishers Ltd, London.

Bearne, E., Clark, C., Johnson, A., Manford, P., Mottram, M., Wolstencroft, H., Anderson, R., Gamble, N., Overall, L., 2007. Reading on Screen. Research Undertaken by the United Kingdom Literacy Association with Support from the Qualifications and Curriculum Authority January–June 2006.

Bearne, E., Wolstencroft, H., 2007. Visual Approaches to Teaching Writing. Multimodal Literacy 5–11. Sage, London.

Christie, F., 2010. Literacy as a theme in educational theory and policy. In: Christie, F., Simpson, A. (Eds.), Literacy and Social Responsibility. Multiple Perspectives. Equinox, London, pp. 9–23.

Exa, October 2011. Web 3.0. Are You Ready for the Next Generation Internet? Exa News Update. http://www.exa.com.au/articles/october_11/web/ (accessed 21.01.13.).

Gee, J., 2003. What Video Games Have to Teach Us about Learning and Literacy. Palgrave Macmillan, New York.

Green, B., 2012. Subject-specific literacy and school learning: a revised account. Chapter 1. In: Green, Bill, Bevis, Catherine (Eds.), Literacy in 3D. An Integrated Perspective in Theory and Practice. ACER Press, Victoria, pp. 2–21.

Healy, A. (Ed.), 2008. Multiliteracies and Diversity in Education. Oxford University Press, Melbourne.

Kalantzis, M., Cope, B., 2005. Introduction: the learning by design approach. In: Kalantzis, M., Cope, M.B., and the Learning by Design Project Group (Eds.), Learning by Design. Victorian Schools Innovation Commission and Common Ground Publishing, Altona, VIC.

Kress, G., van Leeuwen, T., 2001. Multimodal Discourse. Routledge, London.

Kress, G., 2003. Literacy in the New Media Age. Routledge, London.

Kress, G., 2010. Multimodality. A Social Semiotic Approach to Contemporary Communication. Routledge, London.

Lankshear, C., Knobel, M., 2003. New Literacies Changing Knowledge and Classroom Learning. Open University Press, Buckingham.

Lemke, J., 1998. Metamedia literacy: transforming meanings and media. In: Reinking, D., McKenna, M., Labbo, L., Keiffer, R. (Eds.), Handbook of Literacy and Technology. Lawrence Erlbaum Associates, London, pp. 283–301.

Leu, D., 2000. Literacy and technology: deictic consequences for literacy education in an information age. In: Kamil, M.L., Mosenthal, P., Pearson, P.D., Barr, R. (Eds.), Handbook of Reading Research, vol. III. Erlbaum, Mahwah, NJ, pp. 743–770.

Leu, D., 2012. Preparing students for their literacy and learning future: the new literacy of online communication and information. In: Plenary Address at the ALEA National Conference, 6–9 July 2012, Sydney.

Luke, A., Fall 1995. When basic skills and information processing aren't enough: rethinking reading in new times. Teachers College Record 97, 1.

OECD, 2011. PISA 2009 Results: Students on Line: Digital Technologies and Performance, vol. VI. http://dx.doi.org/10.1787/9789264112995-en.

Perfetti, C.A., Sendak, R., 2001. Literacy education. In: International Encyclopedia of the Social & Behavioral Sciences. Elsevier Science Ltd, pp. 8975–8981.

Rowsell, J., Saudelli, M., Scott, R., Bishop, A., May 2013. iPads as placed resources: community literacies in global domains. Language Arts 90 (5), 351–360 (special issue edited by R. Rogers).

Sheridan, M.P., Rowsell, J., 2011. Design Literacies: Learning and Innovation in the Digital Age. Routledge, London.

Simpson, A., Walsh, M., 2014. Pedagogic conceptualisations for touch pad technologies. Australian Journal of Language and Literacy 37 (2), 128–138.

Snyder, I., 1997. Beyond the hype: reassessing hypertext. In: Snyder, I. (Ed.), Page to Screen. Taking Literacy into the Electronic Era. Allen & Unwin, Sydney, pp. 125–143.

Street, B., 1997. The implications of the new literacy studies for education. English in Education 31 (3), 45–59.

Sydney Morning Herald (SMH), 2012. The classroom, but not as we know it: technology to revolutionise schools. In: Report on the International Conference of the Learning Sciences, July 16, University of Sydney, p. 7.

Unsworth, L., 2001. Teaching Multiliteracies across the Curriculum. Changing Contexts of Text and Image in Classroom Practice. Open University Press, Buckingham.

Walsh, M., 2006. The 'textual shift': examining the reading process with visual and multimodal texts. Australian Journal of Language and Literacy 29 (1), 24–37.

Walsh, M., 2008. Worlds have collided and modes have merged: classroom evidence of changed literacy practices. Literacy 42 (2), 101–108.

Walsh, M., 2009. In pedagogic potentials of multimodal literacy. In: Tan Wee Hin, L., Subramanian, R. (Eds.), Handbook of Research on New Media Literacy at the K-12 Level: Issues and Challenges, vols. 1 and 11. IGI Global, USA, pp. 32–47 (Chapter 3).

Walsh, M., October 2010. Multimodal literacy. What does it mean for classroom practice? Australian Journal of Literacy 33 (3), 211–239.

Walsh, M., 2011. Multimodal Literacy: Researching Classroom Practice. PETAA, Newtown, p. 13.

Walsh, M., Simpson, A., 2013. Touching, tapping. thinking? Examining the dynamic materiality of touch pad devices for literacy learning. Australian Journal of Language and Literacy 36 (3), 148–157.

Walsh, M., Simpson, A., 2014. Exploring literacies through touch pad technologies: the dynamic materiality of modal interactions. Australian Journal of Language and Literacy 37 (2), 96–106.

Literary and Philosophical Utopian Thought

Adrienne Redd, Schools for Sustainability, Philadelphia, PA, USA
Tsvi Bisk, Center for Strategic Futurist Thinking, Ziv Group, Ltd, Kfar Saba, Israel

Published by Elsevier Ltd.

Abstract

This article surveys *imagined*, i.e., fictional, societies that have speculated on solutions to the problems that humans face when they live together in population centers. It examines the myth of the Golden Age in the ancient world that gave rise to speculation about how to best organize society. It examines how major cultural shifts, such as the age of exploration, the Protestant Reformation, and Industrial Revolution, led authors to imagine how to stimulate reform and facilitate well-being and social justice in the modern era. Finally, the totalitarian dangers of utopian thinking are considered.

Literary and Philosophical Utopian Thought

Utopian literature envisions fictional societies in which human conflict and deprivation have been resolved by speculative rearrangement of existing political, cultural, and economic arrangements. Some examples of utopian speculation have been written with the explicit goal of popularizing such rearrangements of power, social norms, and distribution of resources in the real world. Other ideal fictional worlds have arisen as authors pursued their own curiosity, but subsequently captured the public imagination and sparked social and political reform movements. This was the case with Edward Bellamy's *Looking Backward* (1888), which was translated into dozens of languages and was so popular that by 1900 only *Uncle Tom's Cabin* had sold more copies.

While there are Asian and Middle Eastern literary works depicting idyllic states of human existence, these are dreams, fantasies, or descriptions of consciousness beyond the earthly plane, not political or philosophical tracts. It is well accepted that utopian literature as a genre has predominantly flourished in the West (Manuel and Manuel, 1979).

A combination of the Judeo-Christian paradise myth of the Garden of Eden and the afterlife, political writings of Hellenic antiquity, and the reform impetus of the Protestant Reformation have offered fertile ground for description of an improved, even perfected society, attainable *without* supernatural intervention or altered states of consciousness.

The Greeks wrote about a better future based on human reason and effort; Hebrew prophets theorized about a better future based on right action; and utopian thought and utopian social experiments have been influenced by these two ancient cultural sources. Post-Renaissance utopian thinking builds on the idea of progress (Bury, 1920/1932/1955) or inexorable scientific advancement and employs the idea of the future as an improvement on the present because of rational human efforts and technological advancements.

The Myth of the Golden Age

The first task of identifying and analyzing utopian treatises is to discern what they are not. In many cultures, societies, religions, and cosmogonies, but most prevalently in Judeo-Christian and Hellenic cultures, there exist myths of a distant past when humankind lived in a primitive state, free of trouble, want, or conflict. Manuel and Manuel call this idea of paradise "the deepest archaeological layer of Western utopia" (Manuel and Manuel, op. cit., p. 33).

The earliest known account of an ideal society occurs in a Sumerian epic poem of the fourth millenium BC in which Enmerkar, son of the sun god Utu, rules the orderly city of Erech (variously referred to as Uruk) (Langdon, 1915).

The University of Pennsylvania Museum of Archaeology and Anthropology houses some 30 000 clay tablets dating to 2900–500 BC inscribed in Sumerian and Akkadian cuneiform excavated in the late nineteenth century in Nippur (present-day Iraq). One of these describes a blissful and peaceful era before humans knew fear and before the 'confusion of tongues' (similar to the story of the Tower of Babel in Genesis 11, pp. 1–9) (Kramer, 1968):

In those days there was no snake, there was no scorpion, there was no hyena
There was no lion, there was no wild dog, no wolf
There was no fear, no terror
Man had no rival.
In those days the land Shubur (East), the place of plenty, of righteous decrees
Harmony-tongued Sumer (South), the great land of the 'decrees of princeship,'
Uri (North), the land having all that is needful
The land Martu (West), resting in security
The whole universe, the people in unison
To Enlil in one tongue gave praise.

Subsequent myths from many cultures describe an ancient state of complete fulfillment and harmony in which humans are totally satisfied by the abundance provided by nature, are free from disease, do not quarrel, and are released from the need for grinding drudgery to survive. In such a state, there is no reason for war or oppression. The natural order flows from a divine source and is just and gentle. Such myths of an idea, happy time, and place fall into three categories: Edenic human states (before sin or corruption), the reward of a life after corporeal life, and the domain of Gods (not human societies at all). Unlike utopian literature that followed, these accounts are not social commentaries or proposals for solutions

to problems; they constitute yearning for something before this era or beyond the terrestrial world.

Roots of Utopian Thinking in the Myth of the Golden Age

These legends of the earliest stage of humanity, such as the Biblical Garden of Eden, appear in myriad cultures. The Greek poet Hesiod, around the eighth century BC, in his compilation of mythological traditions (the poem *Works and Days*), explains that, prior to his present era, there were four progressively more perfect epochs, the oldest of which was the Golden Age (Hesiod, 1914). Plutarch, the Greek historian and biographer of the first century AD, also made reference to a blissful past of humanity. That Hesiod, Plato, and other Greek thinkers see the world disintegrating and becoming tainted and increasingly dysfunctional is consistent with Classical thinking that the world starts as perfect (utopian) and degenerates over time.

What Hesiod's and the Judeo-Christian story of Eden share is that in both, knowledge disrupts perfection: the tree of knowledge in the case of Eden and Pandora's Box in the case of Hesiod. Perfection is final and complete, while the introduction of knowledge changes the social order. Closed societies are suspicious of outside influences that introduce new knowledge and generate dissatisfaction with the status quo and the impetus for reform. It is a common characteristic of literary utopias that they are frequently located on isolated islands or distant planets, or in the distant future – far removed from dissenters or innovators.

Virgil's *Georgics* (29 BC) was modeled on Hesiod's *Works and Days* and also contemplated the characteristics of an ideal society. It describes the consequences of humanity's alteration of nature. Subsequent utopias have shared this environmental theme. Ecological utopias decry a perceived widening gap between the modern way of living that destroys nature and preagricultural or early agrarian society thought to be more in harmony with nature.

Ovid's *Metamorphoses* (AD 7) describes a past era when, because nature and reason were synchronized, human beings were naturally virtuous:

> The Golden Age was first; when Man, yet new,
> No rule but uncorrupted Reason knew:
> And, with a native bent, did good pursue.
> Unforc'd by punishment, un-aw'd by fear.
> His words were simple, and his soul sincere;
> Needless was written law, where none opprest:
> The law of Man was written in his breast.

As well attempting to assert timeless principles of a proposed ideal society, many works of utopian literature reflect the sensibilities and concerns of the era in which they were written. Some researchers regard Hesiod's *Works and Days* as a reaction to economic crisis in Greece at the time (Austin and Vidal-Naquet, 1977). This crisis stimulated colonial expeditions to find new lands of abundance and ease of living – the sorts of mythical places described in western Golden Age literature. Golden Age myths were so powerful that they impressed themselves on the Western persona with profound historical consequences. The cultural belief in the existence of a better time in a better place introduced a restlessness that drove succeeding generations to search for different lands where abundance and ease of life were believed to prevail.

Utopian Treatises of Antiquity

The earliest known utopian text that is truly a political treatise is *The Republic*, written by the Greek philosopher Plato around 380 BC. Part conversation, part fictional depiction, and part educational policy proposal, it slotted citizens into a rigid class structure of 'gold,' 'silver,' 'bronze,' and 'iron' vocational and social classes. The gold citizens were to be trained in a rigorous (five-decade long) educational program to serve as benign oligarchs or 'philosopher-kings.' Their wise leadership would supposedly eliminate poverty and deprivation through fairly distributed resources. However, Plato does not include details of the economic system or explain how the transition to his proposed society will occur.

Lack of explanation of how the transition is to occur is a common omission of utopian proposals. Many fictional utopias rely on the device of placing their imagined society in the remote future or past or on an undiscovered island to avoid this problem.

Approximately two generations later, a different work entitled *The Republic* was written in deliberate opposition to Plato's. The work of the same title was by Zeno, the founder of Stoic philosophy. Although it has not survived, it was his most admired work and is best understood in contrast to Plato's *Republic*. In Plato's society the wise citizen rules the lower classes, reflecting an aristocratic, oligarchic perspective, in which the subordination of lower classes to the rulers is central to the ideal arrangement of society. Zeno's ideal society, in contrast, contains *only* wise citizens and they are subjected to very little authority, making Zeno's work an early expression of libertarian ideas. Other striking aspects of the work are that the author opposed money even for the purposes of trade, but regarded women as property that should be held in common.

Another utopian work of antiquity was written by Euhemerus approximately a century after Plato wrote *The Republic*. Only fragments of the *Sacred History* exist, but from references to it by later historians, we know that it is a travelogue to an Arabian island, Panchaea, where diverse ethnic groups coexist peacefully and where rational and consistent rules underpin the political system. It too draws on the Hellenic tradition of a primordial era of happiness, before Pandora was tricked into unleashing a box of sins and suffering upon humanity.

The last notable work written before the fall of Rome that overlaps with the utopian genre is *On the City of God* by Augustine, 354–430 CE (Augustine of Hippo, 426/1998). It expounds on the superiority of Christian beliefs over pagan worship during times of adversity. Augustine was the son of a pagan. However, while studying in Africa to become a rhetorician he plunged into a turmoil of philosophical doubt in search of truth and famously converted to Christianity. As Hesiod's *Works and Days* was written in times of crisis in Greece, so was Augustine's work a response to the coming fall of Rome. The year 410 was a pivotal moment in Western history; the Vandals, under the command of Alaric, captured the city of

Rome. It is the first work that pits the state versus religion in a political treatise, arguing that in a struggle between two overlapping cities, one earthly and one heavenly, worship of the one Christian god should take precedence and worldly and political stability and prosperity would follow.

This work is followed by a thousand years of silence on the subject of how to imagine an improved or ideal society given mundane realities. Readers might draw the conclusion that people were too distracted with mere survival to contemplate an ideal society, or that whatever significant texts were written have been lost. Sociologists would assert that the very idea of solving social problems would have been entirely alien to the medieval mind. Though there are Renaissance writings on making human society better and more just, the acknowledgment of social problems as caused by human collective action and the idea of their solvability did not arise until the nineteenth century (Rubington and Weinberg, 2003).

Non-Western 'Utopianism'

In ancient China, utopian elements appeared in legalist, neo-Confucian, and Taoist thought. Confucius wrote about *Datong* ('the Great Unity'); *Taiping* was the term he gave to the most advanced stage of society. Another Chinese utopian conception occurs in *The Peach Blossom Spring*, a poem by T'ao Yüan Ming (365–427) about a pacific agrarian society.

Far more frequently than Eastern literature, Western utopian writing has the goal of presenting an alternative to the status quo and thereby nudging society toward improvements. And it has actually produced a plethora of intentional communities that directly affected western society and history and indirectly influenced non-western societies.

Depicting an imaginary paradise of self-indulgence, *The Peach Blossom Spring* (like the medieval poem *The Land of Cockaigne or* Cokaygne) is escapist fantasy without a proposed program for bringing about social justice.

Furthermore, the Hindu concept of *moksha* and the Buddhist concept of *nirvana* cannot be thought of as utopian (in the sense of offering a model for terrestrial society) because they do not describe a potential future, but a state of mind – a belief that if humans practice meditation to escape a continuous stream of nagging thoughts and desires, they can attain an enlightenment that leads to exit from the cycle of life and death. This concept is about individual enlightenment while true utopia literature envisions shifts in culture and political organization that will move the society overall to an ideal state.

There are examples of non-Western utopian writing that contradict the convention that utopian literature is solely Western. These include Li Ju-Chen's *Flowers in the Mirror* (1828) advocating rights for women, Kang Youwei's *DaT'ung Shu* (1935) advocating world unity, as well as African utopias and dystopias such as Bessie Head's *When Rain Clouds Gather* (1969), Ali A. Mazrui's *The Trial of Christopher Okigbo* (1971), Buchi Emechta's *The Rape of Shavi* (1983), Wole Soyinka's *Seasons of Anomy* (1973), and Ben Okri's *Astonishing the Gods* (1995), These examples of non-Western utopian literature all draw on Western intellectual history, starting with the appropriation of the literary genre of the novel (a Western literary innovation) to convey their ideas.

One towering counterexample is *On the Perfect State*, written by one of Islam's most renowned philosophers, Abū Nasr Muhammad bin al-Farakh al-Fārābi (known in European medieval texts as known in medieval Latin texts as Alfarabius or Avennasar) around AD 1000. See Attempted Utopias and Intentional Communities for more discussion of the impact of the writing of al-Fārābi. The treatise by the logician and cosmologist of in Damascus bears a strong relationship to *The Republic* of Plato in that it touts the discipline of philosophy as the path for discovering the correct views about free will and how to live a morally respectable and fulfilling earthly life, and that philosophers ought to be the sovereigns of the perfect state (Al-Farabi, 1985).

Early Renaissance Utopian Literature

Beginning in the fourteenth century, the European age of exploration was partially driven by a vision of unspoiled and beautiful new lands to explore. More's *Utopia* and Bacon's *The New Atlantis* were written during this era of discovery and embody the excitement of encountering new, unspoiled worlds – tabulae rasae for fresh new attempts at human civilization.

Thomas More was an early Renaissance humanist who celebrated reason and its power to create a rational and just society in his work *Utopia*, published in Latin in 1516 and translated into English in 1551. More may have been inspired by the possibilities of creating new societies in the New World. He describes a fictional island in the Atlantic Ocean in proximity to the Americas.

The title of More's novel is pivotal to understanding the utopian literary and philosophical tradition. The etymology of More's invented title is Greek: οὐ, meaning 'not' plus τόπος, 'place.' However, εὖ also means 'good.' Therefore 'eutopia/utopia' means 'good place' as well as 'no place' and signifies a double meaning of an ideal that is unattainable. Therefore, the idea of the impossibility of a perfect society is encoded into the very name of the genre. In common parlance, people will refer to an impossible or unrealistic proposal as 'utopian.' In this sense, the term emphasizes not the perfection but the naïveté of such projects. In fact, critics disagree as to whether the tone of More's travelogue and description is satirical or an earnest proposal for a society that values religious tolerance.

Friedrich Engels (1880/1970) wrote extensively on the significance of Thomas More's *Utopia*, noting that More was a contemporary witness to the emergence of the capitalist categories of bourgeoisie and proletariat emerging from a feudal society. The assertion of that More, a scion of London artisans, sympathized with the peasant masses being driven from their land and livelihoods and that he invented a way out of their wretched state in the travelogue of *A Fruitful and Pleasant Worke of the Best State of a PubliqueWeale, and of the NeweYle Called Utopia* (1516). More's fictional solution was that extended families of craftsmen cooperated and controlled the means of production of goods. Though he did not live to see this manifested in Soviet socialism, Engels saw More's solution as a precursor to cooperative villages.

In *Utopia*, More emphasizes social order, conformity, meaningful work, economic egalitarianism, and religious tolerance. However, in his own life, More was a devout Catholic who persecuted Protestants and burned the books of Martin Luther and William Tyndale. Six Protestants were burned at the stake for heresy during More's chancellorship and he is thought to have played a role in the burning of Tyndale. Some claim he advocated the extermination of Protestants.

This kind of moral dissonance was not uncommon in utopian thought. The impulse to design a society around one, or a small number of unifying principles can be oppressive. The goal of instilling order often takes precedence over more subtle and complex values like individual rights. Views and behaviors other than those dictated by the rules of society and those of the majority may introduce dissatisfaction and upset the smooth running of society. In *The Republic*, Plato's suggestion for how to manage dissenters and poets is that they be exiled. Utopianism tends toward a conformist majoritarian view of society rather than upholding the constitutionalist value of protecting the free speech of citizens because of the value of dissent in correcting and reforming society.

Francis Bacon was called the father of the scientific method by Voltaire, and Thomas Jefferson considered him one of the three greatest men to have ever lived (along with Locke and Newton) (Jefferson, 1789). His utopian *New Atlantis* (1624) depicted a society attuned to the possibilities of rational, organized, intentional, scientific method and places the research university at the center of the imagined society. Both More and Bacon place their utopias on an island near the Americas. Like More, Bacon was clearly inspired by the possibilities of creating new societies in the New World; in fact, he was influential in creating the Virginia and Carolinas colonies. Unlike *Utopia*, *New Atlantis* is technocratic – based on science and invention. In this sense it was an anticipation of modernity while *Utopia* was premodern.

Contrasting the 'knowledge for its own sake' approach of ancient Greek inquiry, Bacon was an advocate for investigating nature to gain knowledge that would materially benefit humankind. This is a fundamental difference between the Classical and Renaissance perspective on the world. Bacon called for a "progeny of inventions, which shall … subdue our needs and miseries." Because of this he is sometimes called the father of the industrial revolution. *New Atlantis* was secularly subversive. Its aim was to substitute spiritual salvation by serving God with a well-run state dedicated to material well-being created by and for humans. It is a quintessential example of the idea of progress, as Bury examines it.

The story reveals the Judeo-Christian roots of some utopian thinking. The name of Bacon's fictional island is Bensalem, a Hebrew word that can be translated as either 'Son of Peace' or 'Son of Jerusalem.' The primary educational or research institution on the island is called 'Solomon's House' ('…denominate of the king of the Hebrews'). It is a description of what eventually came to be the modern research university and ultimately inspired the creation of 'The Royal Society.' A significant character is Joabin the Jew who claims that the Bensalemites are descendants of Abraham and their laws came from Moses. Numerous symbols in the story associate Bensalem with God's 'Chosen People' and a New Jerusalem, images that also inform the Pilgrim and Puritan experiments in New England.

The Idea of Progress

Modern utopian thought coincided with Renaissance critical thinking which produced the Reformation and Scientific Revolution which evolved into the Enlightenment or the Age of Reason. The sixteenth through twentieth centuries fostered the concept that human beings could harness effort and reason to create just societies; that is they could progress toward a better life in the corporeal world rather than wait for the better life in the imagined world to come. Prior to the Enlightenment, people thought of suffering and deprivation as part and parcel of their brief life in preparation for the world to come. The forward-looking, human-centered, and optimistic perspective of the Enlightenment constitutes an important break with the backward-looking that any ideal or uncorrupted collective state was lost or hopelessly buried in past time.

Utopianism from the Renaissance on is forward-thinking. It harnesses rationality into practical application and reveals a secular mind-set even when advanced by religious thinkers. For Protestants in particular, access to the 'world to come' depends on behavior in this world – not only as individuals but as organized societies.

Influenced by Plato's idea of a benign oligarchy, Tommaso Campanella's *City of the Sun* (1601) depicts a unified, peaceful world governed by a theocratic monarchy. He advocates both communism (within which children and women are shared, along with inanimate property) and eugenics to keep society fit and healthy. Campanella anticipated Dewey's 'Learning by Doing' with 'Learning by Observing.' Society itself was the classroom and the walls of the buildings were like a huge illustrated encyclopedia, featuring murals and alphabets. Campanella's conception of the inhabited environment as a pedagogical tool inspired Lenin's propaganda strategy. As totalitarian as Campanella's vision may seem in retrospect, it dramatizes an organized social system to maximize human potential.

Louis-Sébastien Mercier published *The Year 2440: A Dream If Ever There Was One* in 1770. Mercier's eighteenth-century hero falls asleep and wakes more than 500 years later where he sees the contrast between the decay of eighteenth-century Paris and Paris of 2440. Public space is accessible and elegant; the justice system is fair; clothing is comfortable and practical; medical care is based on science and available to all. There are no monks, priests, prostitutes, beggars, slavery, taxes, guilds, armies, or arbitrary arrest. But Mercier's utopia also gets rid of foreign trade, coffee, tea, dancing teachers, and pastry chefs, perhaps because they are purveyors of temptation and emblematic of decadence. Anticipating twentieth-century dictators and reflecting Plato's dislike of poets, Mercier's Paris also abjures 'useless' or 'immoral' literature. This illustrates a commonality of utopian writers – raising personal affinity or distaste to the level of ideology and principle, embodied in the idea that if society merely rids itself of pastry chefs all will be well.

Utopias in the Industrial Age

Equality; or, A History of Lithconia (1802) was the first American utopian novel. Its author is unknown. It depicted an island with a liberal attitude toward sex: the author believed that

the institution of marriage as the only legitimate way to enjoy sex is what caused prostitution. In Lithconia, children were the property of the state. Religion was criticized for meddling in intimate relations. Leisure was more important than work – one worked to gain sustenance, one had leisure in order to live: "…there was no money in the country…lands are in common…labor is a duty required of every citizen… and that duty being performed the remainder of his time is his own…".

Three Hundred Years Hence (1836) by American writer and scientist Mary Griffith was the first utopian fiction written by an American woman and the first to use the future to depict an improved social order. The hero falls asleep and wakes up in the Utopian states of Pennsylvania, New Jersey, and New York 'three hundred years hence.'

Étienne Cabet, the founder of the *Icarian* Movement (see 'Utopian Practice'), wrote *Travel and Adventures of Lord William Carisdall in Icaria* in 1840. It depicted an ideal society in which a democratically elected governing body controlled all economic activity and closely supervised social life. The book's popularity inspired him to establish his utopian movement.

In John Macnie's *The Diothas; or, A Far Look Ahead* (1883), the narrator wakes up in the ninety-sixth century where he finds himself in a New York-like megalopolis. The book predicts paved roads with white lines down the center to separate traffic, a global telephone network, recorded university lectures, electric cars, computers and computer graphics as well as rooftop gardens on public and private buildings. The genders are equal and women have invented television and motion pictures. Society is quasi-socialist.

Like the European age of Exploration, the great emigrations to the United States in the late nineteenth and early twentieth centuries as well as the subsequent migrations to the American West are manifestations of the search for a better life that shares something with the myth of agrarian peace and plenty. The utopian/Golden Age mind-set has historical consequences that are also described in "Attempted Utopias and Intentional Communities."

In the nineteenth century, utopian literature looked to rationality as a way to reshape society. Utopian works also gave voice to the perception that social disruption following the industrial revolution was created and caused by the rise of capitalism and profit motive as an organizing principle of society. These works are often grouped in a greater 'utopian socialist' movement, due to their shared characteristics: an egalitarian distribution of goods, frequently with the total abolition of money, and citizens only doing work which they enjoy and which is for the common good, leaving them ample time for the cultivation of the arts and sciences.

Along with myriad reform movements, the nineteenth century saw an increase in utopian works of fiction, many expressing the belief that social disruption was caused by the rise of capitalism and the profit motive as an organizing principle of society. These works are often grouped in a greater 'utopian socialist' movement, due to their shared characteristics: an egalitarian distribution of goods, the frequent abolition of money, and citizens only doing work that they enjoy and that is for the common good, leaving them with ample time for the cultivation of the arts and sciences.

As in Macnie's *The Diothas* and Mercier's view of a future utopian Paris, Bellamy's vision of a socialist England (and Europe) is told through the device of a citizen of England in the late nineteenth century who sleeps for more than a century and wakes up in a world with equally distributed wealth, in which there is no currency. Citizens merely receive the food, clothing, housing, education, etc. that they need.

Though they are offering an alternative vision of the future, utopian novels such as Bellamy's have used the device of a protagonist (contemporary with the reader) looking backward to the troubled present from a future point when most or all problems have been solved. Yet, the navigation to that future point of social and political transformation is seldom explained. In *Looking Backward*, the genesis of the socialist utopia in which all private property has been nationalized is simply described as the peaceful merging of monopolistic business concerns with the state so that 'the nation is the sole employer and capitalist' in contrast to the turmoil of Soviet and Maoist state socialism that the novel anticipated.

The World a Department Store: A Story of Life under a Cooperative System (1900) by Bradford C. Peck is another example of a protagonist going to sleep and waking up in the future. The author, Peck, was a self-made millionaire who became an advocate of a cooperative society as a palliative to the excesses of capitalism following the panic of 1893 and subsequent economic depression. He admired Bellamy's *Looking Backward* and Henry George's *Progress and Poverty*. The protagonist Percy Brantford (a play on the author's name) is a successful businessman who, after taking an overdose of a sleeping potion, lapses into a coma and wakes up 25 years later. He finds that his hometown has become Cooperative City, and has eliminated poverty, slums, and litter.

Ionia: Land of Wise Men and Fair Women (1898) by Alexander Craig describes a utopian society in the Himalayas in which electricity drawn from the earth's magnetic field supplies all energy needs. Land is owned by the state and leased to businesses and individuals. Poverty and debt have been eradicated and control of pollution and recycling of waste is paramount. The average lifespan is over 100 – twice the lifespan of Europe at the time.

Parallels to Ionia in James Hilton's novel *Lost Horizon* (1933) are too numerous to ignore; interestingly Craig wrote his book during the severe economic depression of 1893–1898 while Hilton wrote his at the outset of the Great Depression of the 1930s. Crisis, whether economic, social, or political, seems to stimulate the imagination to invoke a better way of life.

Earth Revisited (1893) by Byron Brooks also projects its nineteenth-century protagonist into the future where he encounters communal food production, solar power, electric cars, color photography, and a reclamation project that turns the entire Sahara desert into rich farmland. It also describes a juvenile justice system authorized to remove children from unfit parents.

The Human Drift (1894) by King Camp Gillette (the founder of the Gillette razor blade company) advocated a single publicly owned corporation that would control all economic activity (a variation on *Looking Backward*, in which the state subsumes monopolies). At the time the United States had 60 million citizens and the book proposed they all be concentrated in one giant metropolis built over Niagara Falls where the first large electrical generating facilities were being constructed by Tesla and Westinghouse. Gillette thought the city could absorb an additional 30 million people of future

population growth. Intensive agriculture would be in proximity to the city limits. The rest of the continent would revert back to nature. Gillette was a business tycoon who believed that the business magnate would be the natural leader of social reform, because of his rationality, his managerial ability, and his understanding of the power of capital.

Aristopia: A Romance-History of the New World (1895) by Castello Holford is a 'what if?' utopian speculation. The novel asks, "What if the founding of the United States had occurred under different conditions? Would American society have developed in a superior way?"

In His Steps: What Would Jesus Do? (1897) by Congregationalist Minister Charles Sheldon is not about personal redemption but rather about moral choices leading to or avoiding poverty and deprivation. The protagonist Rev. Henry Maxwell challenges his congregation to not do anything for an entire year without asking: "What Would Jesus Do?" Though overtly religious, it is based in the modern and secular ideas of rationality and personal responsibility.

The meaning of 'utopia' as 'ou-topia' or 'noplace' has had far-reaching resonance in the titles and themes of novels written hundreds of years after More provided the genre with its label. *The News from Nowhere* (1890) by William Morris presented a more libertarian (non-state) socialism in which there was no private property, big cities, or state authority. There was also no money, divorce, courts, prisons, or class system. Samuel Butler titled his satire and proto-dystopia *Erewhon*, which is an anagram (or nearly backward spelling) of 'nowhere.'

The Year 3000: A Dream was published in 1897 by Italian anthropologist Paolo Mantegazza. It is a work of utopian early science fiction in which a couple on their honeymoon ('mating union') in the capital of the United Planetary States visit a socialist utopia with airplanes, artificial intelligence, CAT scans, and credit cards as well as serious debates about divorce, abortion, and euthanasia. Mantegazza also predicted a unified Europe, world peace, and the elimination of disease.

Other works in this vein include *A Traveler from Altruria* by William Dean Howells (1892–93), which criticized the robber baron capitalism of the age. Altruria is an isolated island in which life is based on altruism (hence its name). *A Modern Utopia* (1905) by H.G. Wells is a modern takeoff on Plato's *Republic* in that benign Samurai on a different planet (a kind of cosmic island) have found a much better way to live.

Less frequently, utopias have also been imagined on the opposite side of the political spectrum. For example, Robert A. Heinlein's *The Moon Is a Harsh Mistress* portrays an individualistic and libertarian utopia. Capitalist utopias of this sort are generally based on free-market economies, in which the presupposition is that private enterprise and personal initiative without an institution of coercion, government, provides the greatest opportunity for achievement and progress of both the individual and society as a whole. 'Galt's Gulch' in Ayn Rand's *Atlas Shrugged* is another presentation of a libertarian and individualistic utopia.

The Soviet writer Ivan Efremov produced during the 'Thaw' period the science-fiction utopia in which a united humanity communicates with a galaxy-wide Great Circle and develops its technology and culture within a social framework characterized by vigorous competition between alternative philosophies.

One example of utopian thought that inspired a massive social movement was Theodore Herzl's utopian novel *Old New Land* in which Herzl, the founder of Political Zionism, painted a utopian vision to help sell the political idea of repatriating Israel as an attainable possibility.

Freeland: A Social Anticipation (1891) by Theodor Hertzka (the Austrian counterpart to Bellamy) was hugely successful throughout Europe and inspired a utopian movement of 1000 members in 24 chapters. A small group also tried to establish a utopian community based on Hertzka's work in Kenya but failed.

Ecological Utopias

William Henry Hudson was a naturalist who wrote *A Crystal Age* (1887), which describes a dystopian collapse of modern civilization as a prelude to a back-to-nature utopia.

In his final work *Island* (1963), Aldous Huxley depicts a gentle and doomed utopian, fictional island Pala. He uses the form of the novel to issue a warning about overpopulation, coercive politics, militarism, and mechanization, and highlights the destruction of the environment and the worship of science in contrast to Pala.

In the novelette *Rumfuddle* (1973), Jack Vance presents an inventive twist on the ecological utopia. His hero invents paratime travel and becomes effectively the ruler of earth by giving everyone their own alternate earth wilderness world as vacation retreats/suburbs without neighbors. However, he requires them to work during the week cleaning up the original earth and restoring its pristineness. A typical job is driving a bulldozer that shoves the detritus of industrial civilization through a portal into the oceans of a paratime garbage world.

Ecological planning is the central thrust of *Ecotopia: The Notebooks and Reports of William Weston* (1975) by Ernest Callenbach, in which the Pacific Northwest has seceded from the union to set up a new society. Callenbach subsequently wrote a prequel (1981) to better explain how Ecotopia had come to secede from the United States. Several fictional utopias, notably *Woman on the Edge of Time* (1976) by Marge Piercy, attend closely to environmental sustainability in the design of their society.

Feminist Utopias

Elizabeth Corbett's *New Amazonia: A Foretaste of the Future* (1889) envisions a successful suffragette movement that eventually produces a breed of 'Amazonians' who create a female utopian society. The book's female narrator wakes up in the year 2472 and finds that the Amazonians are vegetarians who use euthanasia to eliminate malformed and illegitimate children. Their economy is a version of state socialism and while men are allowed to live on the island, they are forbidden to hold public office.

Utopias have been used to explore whether gender is a societal construct or a hard-wired imperative. The fictional aliens in Mary Gentle's *Golden Witchbreed* (1985) start out as gender-neutral children and do not develop into men and women until puberty, and gender has no bearing on their social roles.

In contrast, Doris Lessing's *The Marriages Between Zones Three, Four and Five* (1980) suggests that men's and women's values are inherent to the sexes and cannot be changed, making compromise between them essential. In *My Own Utopia* (1961) by Elizabeth Mann Borghese, gender exists but is dependent upon age rather than sex; genderless children mature into women, some of whom subsequently become men.

Utopian single-gender worlds or single-sex societies have long been one of the primary ways to explore gender inequality in work. In speculative fiction, female-only worlds have been imagined to come about by the action of disease that wipes out men, along with the development of female parthenogenic reproduction. The resulting society is often shown to be utopian by feminist writers. Many influential feminist utopias of this sort were written in the 1970s: notably Joanna Russ's *The Female Man* (1975) and Suzy McKee Charnas's *Walk to the End of the World* (1974) and *Motherlines* (1981).

Utopias imagined by male authors have generally included equality between sexes, rather than separating them. Worlds without men have been portrayed almost exclusively by lesbian or feminist authors; their use of female-only worlds allows the exploration of female independence and freedom from patriarchy. The societies may not necessarily be lesbian, or sexual at all – a famous early example of a celibate, females-only utopia being *Herland* (1915) by Charlotte Perkins Gilman. This was followed by the sequel *With Her in Ourland* (1916).

Gilman had previously written a nonfeminist utopian novel *Moving the Mountain* (1911), which, similarly to Bradford Peck, moves its protagonist into the immediate (1940) rather than the distant future. This shorter term gambit demonstrates that the author wished to influence current society by showing it could achieve with imagination and political will. In the novel, by 1940, the United States had eliminated poverty and prostitution; had solved the labor problem, the color problem, and the sex problem; and had almost eliminated disease, accidents, and fires. Most people worked 4 h a day even though they only had to work 2 h a day to maintain themselves. Deforestation and soil erosion were in advanced stages of remediation. Tidal, wind, and solar power were the primary means of generating energy and the skies were filled with aircraft – a stunning vision for a book written in 1911.

Ambiguous Utopias and Dystopias

Focusing on a central problem in an attempt to eradicate it produces social distortion in the hypothetical society's treatment of other problems, such as the need for individuality and creative expression.

The Dispossessed (1974) by Ursula LeGuin is one of the few utopian novels that takes this balancing act into account. The novel is nuanced in that it touts a 'nonpropertarian' society, while simultaneously criticizing this egalitarian society for its isolation of people who are exceptional (principally the protagonist). *The Dispossessed* confronts the question of how to balance the social flattening of the fictional society of Anarres with the urge of the individual (the protagonist, Shevek) to fully explore his extraordinary talents. The main character of *The Dispossessed* is a mold-breaking physicist who is impeded in his work by his society's inability to balance the ideal of equality and the pursuit of the common good promoted by his imagination and need for solitude and the resources to conduct research.

Utopian literary thinking spawned other speculative fiction, most prominently that of dystopia, a subgenre that examines the totalitarian penumbra of utopian projects. Notable examples include *1984* by George Orwell (1949), *Brave New World* (1932) by Aldous Huxley, and *Anthem* (1938) by Ayn Rand. Very few earnest utopian novels have been written since the early 1970s; *The Dispossessed* (1974) by Ursula LeGuin is among the last.

Like fictional utopias, dystopias also depict the vision and planning necessary to bring the imagined society under one philosophical tent. However, in dystopian fiction, something has gone terribly wrong for some or all of the characters. For example, *Brave New World* is arguably utopian for nearly all of its characters. In the grand scheme of mechanized manufacturing applied to human reproduction, the citizens of London in 2540 feel included; their needs are met; they are provided with productive work, an ample supply of Soma (a mild tranquilizer), and attachment-free sex. However, the protagonist John the Savage cannot fit in or find sufficient meaning in its planned society and hangs himself at the end of the novel.

Walden Two is a novel written by behavioral psychologist B. F. Skinner. In 1948 when it was written, it was science fiction, since science-based methods for altering behavior did not yet exist. Such methods are now known as applied behavior analysis. The novel might seem utopian to some and terrifying to others because it relies on a rejection of free will. *Walden Two* embraces the proposition that the behavior of organisms, including humans, is absolutely determined by genetic and environmental variables and that systematically altering environmental variables can generate a sociocultural system that approximates utopia. The dark side of such a proposition can be seen in *A Clockwork Orange* (1962) by Anthony Burgess, in which juvenile delinquents are subjected to harsh conditioning to break them of their violent impulses or the brutal conformity of the scenes in Camazotz in *A Wrinkle in Time* by Madeleine, published in the same year.

The Truman Show (1998), a film directed by Peter Weir, depicts a utopia created for one man, the star of a reality television show, a utopia of which he is unwittingly the center. His town and his life are seemingly clean, happy, and prosperous, until he begins to suspect that his life is completely artificial. This underlines the idea that people must choose to join a community, no matter how idyllic, of their own free will.

One of the most chilling exposés of the dark side of the utopian impetus is the science fiction film *Serenity* (2005) written and directed by Joss Whedon. In the film, the central government wafts a tranquilizing gas into the environment of a newly colonized planet (to suppress their aggressive urges), only to find that some citizens exposed to it simply lose their will to live while others become hyperaggressive. Finding that it has created homicidal rapists and cannibals, it then exerts extreme control to keep the effects of the gas a secret.

An exploration of the dark side of utopia is Ursula LeGuin's short story *The Ones Who Walk Away from Omelas* (1973). A utopian city of peace and plenty is the main 'character' of the

story in which characters are simply used to dramatize the philosophical idea the story explores. In a kind of Taoist balance, the pleasant lives of the citizens require that a single child be held captive in filth and deprivation. What may be a utopia for some is dystopia for another. Citizens are informed of the trade-off when they reach the age of majority and may choose to stay or leave in the nearly ideal city.

The Totalitarian Impulse

Not surprisingly, most utopian novels fail to deal with problems faced in the contemporary, globalized, multicultural society we actually inhabit. For example, there are very few comprehensive speculative responses to racial discrimination or oppression of subordinate cultural groups in multiethnic or polylingual societies. The anti-Semitism of Fourier and Craig's utopias is the canary in the coal mine of a type of utopian thinking that prefers a homogeneous world free of dissenters, visionaries, or nonconformists.

One of the few mentions of racial strife occurs in LeGuin's *The Lathe of Heaven* (1971) in which the protagonist telekinetically dreams of solutions to problems in a world very like our own. His actualized dream to address racism makes inhabitants of the world undifferentiated by facial structure or pigmentation. Not surprisingly, the narrator and the characters perceive this as an unsatisfactory solution. Marge Piercy's *Woman on the Edge of Time* (1976) addresses racism through a combination of assigned rather than biological families and genetic engineering so that no character is 'white'; everyone has family members of various shades of brown.

Total control of society is vital to the idea of perfection or solution of *all* problems. Centralized and strong, though benign, authority is therefore a distinguishing characteristic of utopian literature. Plato's vision of the 'good society' is predicated on intolerance of disorder or dissent. For Plato, poets are corruptors of morality because they imitate reality rather than seeking ultimate truth. Artists deal with the world of imagination appearance and not absolute reality. They were to be banished from his imagined commonwealth. This begs the question 'Whose truth; whose reality?' Utopians often present their subjective intellectual predilections as objective truth and utopian intolerance for alternative visions of life have a long pedigree – with deadly consequences. Plato did not like other poets (he was a poet himself by any standard) so he wanted to banish poets. Plato most likely pictured himself as a philosopher-king so his 'utopia' had him in charge. Karl Popper went so far as to brand Plato as one of the antecedents of Fascism by quoting him:

> The greatest principle of all is that nobody, whether male or female, should be without a leader. Nor should the mind of anybody be habituated to letting him do anything at all on his own initiative; neither out of zeal, nor even playfully. But in war and in the midst of peace – to his leader he shall direct his eye and follow him faithfully. And even in the smallest matter he should stand under leadership. For example, he should get up, or move, or wash, or take his meals … only if he has been told to do so. In a word, he should teach his soul, by long habit, never to dream of acting independently, and to become utterly incapable of it.

Popper concludes that "Plato's political program, far from being morally superior to totalitarianism, is fundamentally identical with it" (Popper, 1965).

The Utopian Quandary

The island as a trope in utopian literature expresses poetically the nearly universal hope for new, undiscovered lands and social systems, but also dramatizes the degree of isolation and control such social orders may require. Both literary and attempted utopias depend on being cut off from the larger world. When pressed, the mouthpiece of Bellamy's *Looking Backward* admits to the visitor from the past that if the social utopia were to traffic with the as-yet unconverted portion of the world, the moneyless society could not survive. Only through segregation from other groups with different values and access to resources are such isolated utopias possible. As the world becomes more interconnected and interdependent, the thought experiments of fictional utopias become more valuable than ever, but such speculation must find a way to integrate order and justice with the variety of faiths, cultures, and norms that global society brings together.

See also: Attempted Utopias and Intentional Communities.

Bibliography

Augustine of Hippo, 426/1998. In: Dyson, Robert W. (Ed.), The City of God Against the Pagans. Cambridge University Press.

Austin, Michel, Vidal-Naquet, Pierre, 1977. Economic and Social History of Ancient Greece: An Introduction. University of California Press, Berkeley.

Bellamy, Edward, May 1889. How I came to write looking backward. The Nationalist (Boston) 1 (1), 1–4. Available: http://www.marxisthistory.org/history/usa/parties/nationalists/1889/0500-bellamy-howicametowrite.pdf.

Bury, John Bagnell, 1920/1932/1955. The Idea of Progress: An Inquiry into its Origin and Growth. Macmillan, London. New York: Macmillan reprint; New York: Dover.

Engels, Friedrich, 1880/1970. Socialism: utopian and scientific. In: Marx/Engels Selected Works, vol. 3. Progress Publishers, London, pp. 95–151. Available: http://www.marxists.org/archive/marx/works/1880/soc-utop/.

Goodwin, Barbara, Taylor, Keith, 1982/2009. The Politics of Utopia: A Study in Theory and Practice. Peter Lang AF, International Academic Publishers, Bern.

Jefferson, Thomas, 1789. To Richard Price Paris, January 8, 1789. Letters of Thomas Jefferson, 1743–1826. University of Virginia Library.

Kramer, Samuel Noah, Januray–March 1968. The "babel of tongues": a Sumerian version. Journal of the American Oriental Society 88 (1), 108–111. Portions of the original Sumerian texts available: http://etcsl.orinst.ox.ac.uk/cgi-bin/etcsl.cgi?text=t.1.8.2.3.

Manuel and Manuel, op. cit., p. 33.

Manuel, Frank E., Manuel, Fritzie P., 1979. Utopian Thought in the Western World. Harvard University Press, Cambridge.

Popper, Karl, 1965. The Open Society and Its Enemies. In: The Spell of Plato, vol. I. Princeton University Press, Princeton.

Rubington, Earl, Weinberg, Martin S., 2003. The Study of Social Problems: Seven Perspectives. Oxford University Press, New York.

Salerno, Anthony, 1979. The social background of seventeenth-century emigration to America. Journal of British Studies 19, 31–52.

Sargent, Lyman Tower, 2010. Utopianism: A Very Short Introduction. Oxford University Press, Oxford.

Smith, Craig, Summer 2005. "Great reformation in the manners of mankind": Utopian thought in the Scottish reformation and enlightenment. Utopian Studies 16 (2), 221–245. Available at: http://www.highbeam.com/doc/1G1-137362209.html.

Primary Texts

Al-Farabi, Abu Nasr, 1985. On the Perfect State (Richard Walzer and Seyyed Hossein Nasr, Ed. and Trans.). Clarendon Press, Oxford. Portions available: http://www.sacred-texts.com/isl/palf/palf10.htm.

Bellamy, Edward, 1888. Looking Backward 2000–1887. Ticknor and Company, Boston.

More, Thomas, 1518/1895. The Utopia of Sir Thomas more. In: Latin from the Edition of March 1518, and in English from the 1st Ed. Of Ralph Robynson's Translation in 1551. Clarendon Press, Oxford. Available as an: http://books.google.com/books?id=6REuAAAAMAAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false.

Plato, 1968. The Republic (Allan Bloom, Trans.). Basic Books, New York.

Hesiod, 1914. The Homeric Hymns and Homerica with an English: Works and Days (Hugh G. Evelyn-White, Trans.). Harvard University Press and London: William Heinemann Ltd, Cambridge. Available: http://www.sacred-texts.com/cla/hesiod/works.htm.

The Land of Cokaygne. *c.*1330. British Library/British Museum. Available: http://www.thegoldendream.com/landofcokaygne.htm.

Langdon, Stephen, 1915. Sumerian Epic of Paradise, the Flood and the Fall of Man. The University Museum, Philadelphia. Available: https://archive.org/stream/sumerianepicpar00langgoog#page/n14/mode/2up%20.

Leguin, Ursula, 1975. The ones who walk away from Omelas. In: The Wind's Twelve Quarters. Twayne Publishers, Boston. Available: http://www-rohan.sdsu.edu/faculty/dunnweb/rprnts.omelas.pdf.

Lithics and Archaeology

Jacques Pelegrin, Allée de l'Université, NANTERRE cedex, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In archaeology, 'lithics' refers to all stone artifacts fabricated by humans from the beginning of prehistory until the present time. Stone artifacts play a major role in the understanding of prehistory. They are the most abundant, and often the only, human-worked material preserved on prehistoric sites, and they retain traces informative of their fabrication and use. From an anthropological perspective, stone tool making is an essential criterion indicating human evolution. For the Paleolithic, and along with ceramics and metal for later periods, it is a fundamental source of information concerning past human societies' traditions and economy.

Historical Aspects

The Significance of Lithics in the Discovery of Prehistory

At the end of the sixteenth century, Mercati recognized the human origin of polished stone axes, which were previously believed to have been created by lightning bolts. In 1723, Jussieu compared stone axes to ethnographic objects from Canada and the Caribbean. Most notably, as early as the beginning of the nineteenth century, Thomsen classified objects from the Royal Museum of Copenhagen into three successive phases: Stone Age, Bronze Age, and Iron Age. In 1866, Lubbock divided the Stone Age into the Paleolithic, or knapped Stone Age, and Neolithic, or polished Stone Age. Meanwhile, the Stone Age, which encompassed all of humanity, remained undated until the development of absolute dating methods (*see* Chronology, Stratigraphy, and Dating Methods in Archaeology).

However, the notion that stone tools were very ancient was suggested as early as the late eighteenth century when John Frere identified stone tools in association with the bones of pachyderms at Hoxne (Suffolk, England). Soon after, the first excavations of caves in France, Belgium, and England allowed for arguments of a great antiquity of humanity. In 1859, Boucher de Perthes, with support from British geologists, finally succeeded in convincing the scientific community of the existence of an ancient, properly 'antediluvian' epoch of humanity. He demonstrated that flint hand axes could be found *in situ* with the bones of extinct animals deep within gravel quarries in northern France. The idea of very early prehistoric humans equipped with only stone tools was born.

Lithics as a Criterion for the Definition of Early Humanity?

Chimpanzees, our closest primate relatives, can use natural objects as tools (stone hammers and anvils to break nuts) and even alter natural objects to make tools (removing leaves from twigs to capture termites or ants), but they do not fabricate stone tools in the wild (Berthelet and Chavaillon, 1993). The production of stone tools is thus a tempting criterion for distinguishing the earliest species of *Homo* (*Homo habilis*, *Homo rudolfensis*, etc.) from their ancestors and from other contemporary hominids, Australopithicenes, who became extinct between 3 and 1 million years ago. Unfortunately, this criterion remains putative since it is rare to find stone tools and early hominid fossils in direct association. There is as well no a priori reason to believe that some of the Australopithecines did not also make stone tools (Hovers and Braun, 2009). Currently, the oldest known assemblage of knapped stones, discovered by H. Roche at the site of Kada Gona in Ethiopia, and dated to at least 2.6 million years ago, already shows mastery of conchoïdal fracture (the basis of controlled stone knapping), perhaps indicating that stone tool technology had already existed for a considerable time (Roche et al., 2009).

The Role of Lithics in the Construction of a Prehistoric Chronology

Once the great antiquity of humankind was finally accepted, researchers put their efforts into developing a chronological classification of the Stone Age. After the first paleontological subdivisions of E. Lartet (1858; Cave Bear Epoch, Reindeer Epoch), classifications were based on stone tools, along with bone tools for more recent periods (G. de Mortillet in 1869, E. Piette in 1889, H. Breuil in 1912, and D. Peyrony in 1933). These classifications emphasized the most characteristic tools found in strata of different sites. Following a paleontogeological perspective, these tools were called 'index fossils.'

This procedure led to two excesses that sometimes persist in current research. First, these 'index fossils' were regarded as exclusively representative of distinct time periods and 'cultures,' which were believed to represent different human groups in an ethnic sense, following an overly simplistic historic schematization (*see* Ethnic Identity and Ethnicity in Archaeology). Much work has since been devoted to determining precise dates and geographic origins for these cultures and groups, which are seen to follow a rigid succession of 'evolutionary cycles' based on the appearance and disappearance of a few tool types. Second, these successive cultures, which were originally defined in France following the excavation of numerous stratified sites, were abusively extended to other European regions, thus masking the originality of archaeological data over broad areas.

A Slowly Developing Technological Perspective

As early as the nineteenth century, some archaeologists were becoming interested in what stone tools could tell us about the

daily activities of our ancestors. They asked a first, seemingly accessible, question: how were these tools made?

This question immediately led to attempts to reproduce stone tools, yet the results remained superficial due to poor experimental procedure and lack of rigor, according to L.L. Johnson (1978). Numerous prehistorians observed modern gunflint knappers at Meusnes, France, and Brandon, Suffolk, at the beginning of the twentieth century. However, an extreme production standardization, and exclusive use of iron tools, limited the utility of the comparison. Other prehistorians turned to early accounts of Native Americans manufacturing stone tools. However, many of the descriptions they consulted were second-hand, and thus too imprecise to be useful. Nonetheless, the record in the Spanish Codex of obsidian pressure debitage by the Aztecs permitted D. Crabtree in 1968, and then J.E. Clark in 1982, to successfully reproduce this production process.

Some prehistorians, such as J. Evans, were able to synthesize numerous observations of archaeological collections, ethnographic records, and their own knapping experiments. As a result, the three major knapping operations (shaping, debitage, and retouch) and several knapping techniques (direct, hard stone percussion; indirect 'punch' percussion; and pressure retouch) were recognized very early, at about 1870. The research of W.H. Holmes during the early twentieth century concerning the lithic production activities of ethnographic and prehistoric North Americans was also remarkably modern in its approach to and interpretation of technology.

Meanwhile, the functions of stone tools remained hypothetical except for a few objects such as polished axes and certain arrowheads, the form of which suggests their potential use. Lacking other methods for determining function, the principal categories of stone tools were named after their similarity of shape with traditional tools (e.g., scrapers, borers, and knives).

Recent Methodological Developments and Research Orientations

In conjunction with the genealogical perspective described above, early research was essentially aimed at identifying the major evolutionary stages of the Paleolithic cultures. It was not until the middle of the twentieth century that broader and more analytical perspectives were developed.

The Typotechnological Approach

During the 1950s, the French prehistorian François Bordes developed a method of studying and comparing assemblages of lithic artifacts, as opposed to single objects, which was both quantitative and qualitative. The technological understanding of the objects that he gained through personal knapping experience enriched the methodology he developed. While in the past it had been considered sufficient to discuss only the 'index fossils,' this new method entailed study of all the tools recovered from the site or the layer, as well as the technological processes by which they were made. The objective of this broader characterization of lithic industries was the identification of the conceptions of the prehistoric groups that produced them. This became the basis of what is referred to as the 'cultural' approach, which has since been subject to long debates. In particular, Bordes distinguished several different, contemporaneous 'facies' in the Late Mousterian of southwest France. He interpreted these as being representative of different human groups, with different technological traditions, occupying the same region at the same time (Bordes, 1950). L. Binford and S. Binford contested this interpretation, proposing instead that these different facies represented different functional, perhaps seasonal, activities. Other researchers have proposed that these facies could be chronological phases or that they reflect the differential availability of raw materials. None of these interpretations has received unanimous acceptance. In addition, F. Bordes' method of distinguishing tool types was rather intuitive; some authors, like H.L. Dibble, have suggested that several of these tool types actually represent different stages of use and resharpening of a single type of tool. Presently, economic and behavioral factors are proposed as accounting for at least some of the diversity of Mousterian assemblages (see Delagnes and Rendu, 2011; Kuhn, 1995).

Meanwhile, advances in computer technology that have allowed for more elaborate statistical analyses permitted H.L. Movius to confirm the reality of several tool types distinguished by F. Bordes and D. de Sonneville-Bordes, particularly for the Upper Paleolithic. The 'reality' and meaning of types, and the possible use of typological analysis continue to be debated.

The Paleoethnological Approach

André Leroi-Gourhan, a French contemporary of Bordes, implemented a paleoethnological approach to prehistoric human groups. This perspective stemmed from his more general studies of techniques in ethnographic contexts, following in the mid of the twentieth century a broad anthropological perspective (Leroi-Gourhan, 1993). Along with other members of his research team, his objective was to reconstruct the spatial dimensions of human activities, which are principally represented by the remains of their technological activities and the organization of the artifacts in relation to various structures (fireplaces, traces of huts or tents, etc.). Such spatial analysis could be performed only at those rare sites, like the famous 'Pincevent' in the Paris basin where occupation floors are well preserved. In addition, excavation techniques had to be modified from the primarily vertical, stratigraphic section approach that was well adapted to the chronological perspective described above, to a more horizontal approach to allow for exposure of contemporaneous objects on an occupation floor (*see* Survey and Excavation (Field Methods) in Archaeology). Under these conditions, this research objective led to two important methodological developments.

The first development was the systematic search for refits between broken objects, retouch flakes and the tool from which they originate, and between lithic items successively detached from their core. As sporadically shown since the nineteenth century, refitting makes it possible to detect disturbances within and between archaeological levels, and also to interpret certain archaeological structures and activity areas. In addition to better grasping the spatial and chronological relations between isolated objects, the method of refitting is increasingly

used to address behavioral and technological questions (Schurmans and De Bie, 2007).

The second development has been the emergence, in France and increasingly in the Anglo-Saxon world, of an approach that considers all lithic remains, including waste products of production, finished tools, as well as used and abandoned or broken tools, in terms of their technical process of production and use. This analysis, originally developed in ethnological research and the anthropology of techniques, following the inspiration of Marcel Mauss and André Leroi-Gourhan, is called the *chaîne opératoire* (operational sequence) approach (Schlanger, 2005; Soressi and Geneste, 2011). This sequence is composed of the succession of technical actions beginning with raw material acquisition and continuing to the abandonment of the manufactured tools. The detailed, empirical, and experimentally based analysis of this sequence, including its invariant and flexible moments, its constraints, and its choices, makes it possible to reconstruct, at least in part, the social, cognitive, economic, and environmental factors implicated in the transformation of naturally occurring materials into finished cultural products (Dobres and Hoffman, 1999; Inizan et al., 1999–2009).

The Processual Approach

Beginning in the 1960s, L. Binford and other American archaeologists developed the theoretical approach called 'processual archaeology.' According to this approach, archaeological artifacts must be considered in terms of the dynamic cultural and natural processes by which they reached the state in which they are found by archaeologists – rather than as static end products, as they are when a typological approach is used. As with the paleoethnological approach, questions can thus be asked concerning the technological behavior of prehistoric peoples, and one can attempt to situate this behavior within its environmental context (*see* Hunter-Gatherer Societies, Archaeology of). The concept of 'middle-range theory' was borrowed from sociology to provide a link between static archaeological remains and dynamic processes responsible for them. This involves the development of explanatory 'models' based primarily on ethnographic references and experimentation. In addition, processual archeologists placed a new emphasis on the natural formation processes (geological, biological, chemical, etc.) of archaeological sites and objects, as well as on the rigorous application of the hypotheticodeductive method to archaeological interpretation. This led to a more critical view of the ability to interpret the past than that espoused by most European archaeologists, and thus to a divergence of research orientations. In terms of lithic studies, American archaeologists became increasingly concerned with the broader contextual aspects of the objects, including the role of lithic technology within the entire technological system of a cultural group. This approach, known as 'technological organization,' addresses issues such as group mobility, tool design needs and constraints, time allocation, risk factors, and degrees of technological investment (Bamforth and Bleed, 1997; Odell, 1996). It is still the dominant approach followed by American archaeologists today, besides other theoretical or paradigmatic approaches (Bleed, 2001; Hayden et al., 1996; Odell, 2000) (*see* Environmental Archaeology, Ethnoarchaeology, Cognitive Archaeology, and Archaeology, Theory In).

Developments in Technological Analysis and Classification

Simultaneously, and until the present time, continental European archaeologists, the French in particular, continued to develop their inductive approach to lithic artifacts. This perspective is based on profound practical knowledge of Old World lithic industries, whose high variability and informational value (due to an abundance of complex and sophisticated technologies) is particularly well adapted to this type of approach.

As early as 1963, J. Tixier had made the pertinent distinction between method and technique in stone knapping. Technique designates the modes of execution in tool fabrication, including how force is applied (direct percussion, indirect percussion, or pressure), the tool employed (e.g., stone cobble, wood, antler or bone hammer, and copper point), and the gestures that are used (e.g., direction of blow and manner of holding the stone and tools). Method, which is performed according to a systematic and more or less reasoned process, concerns the organization (order and placement) of series of removals from the object being worked, and thus forms the structure of the operational sequence of production. Systematic experiments, which developed largely since then, are essential to the recognition of techniques that can be identified based on particular stigmata seen on the objects (Pelegrin, 2012). In contrast, the knapping method can be identified only through a technological 'reading' of all the objects that entails determination of the direction and order of removals from each piece.

An experienced lithic analyst can mentally perform this technological reading, even in the absence of refit pieces. It is thus possible to classify an assemblage of lithic artifacts according to their original order in the operational sequence of tool fabrication, and distinguish between the products desired by the stone knappers and the waste resulting from their production (Andrefsky, 2009; Inizan et al., 1999–2009).

Economic Perspectives

The analytical principles described above, along with the technological classification of lithic artifacts, have revealed that in certain assemblages particular types of 'blanks' were selected for the fabrication of particular types of tools. This realization led to the concept of debitage economy (economy of the products of knapping). This concept was extended to the end of the operational sequence to include tool economy, via which certain blanks were successively used and/or resharpened in different ways. At the beginning of the operational sequence, in a single lithic assemblage, different stone raw materials were often differentially acquired and/or used. This led to the concept of raw material economy. These perspectives are especially useful for identifying possible choices and decisions made by prehistoric peoples during technological activities (Inizan et al., 1999–2009; Perlès, 1992).

Archaeologists also recognized the need for precise determinations of the geographic origins, and modes of acquisition, of the stone materials found in archaeological sites. Following Renfrew's research on the circulation of obsidian in the Near East, surveys and petrographic characterizations (*see* Archaeometry) were conducted in many regions of the world where

diverse workable stone materials are found. The identification of these materials in prehistoric lithic assemblages allows us to determine the extent of the territories frequented by foragers, from several kilometers for the Early Paleolithic human groups, to several hundreds of kilometers during the Upper Paleolithic (Féblot-Augustins, 2009). Later, during the European Neolithic, stone axes and long blades produced in specialized workshops were exchanged over distances up to 1000 km (e.g., Bradley and Edmonds, 1993). Once neglected, compared to ceramics, the specialized lithic productions from the Neolithic and even 'classical' periods are presently receiving more interest and provide precious information both on social structure (Hayden, 1998) and exchange networks (*see* Trade and Exchange, Archaeology of).

The Development of Methods for Determining the Function of Stone Tools

In 1964, with the English publication of his long-standing research, the Russian archaeologist S.A. Semonov demonstrated that the use of stone tools left different macroscopic and microscopic traces on their surfaces. V.E. Schelinskiï and later L.H. Keeley further developed Semenov's results using a high magnification method that allows for identification of the materials worked (e.g., wet or dry skin, bone, antler, meat, and plants) on the bases of comparison with experimental samples. In addition, lower magnification methods can indicate the manner in which the tool was used (e.g., scraping, drilling, sawing, and striking), and how it was possibly hafted, according to the direction of striations and location of polish and microremovals (Keeley, 1980).

Unfortunately, these time-consuming methods are not completely reliable. Traces of use are quite often damaged, especially on older tools. Another problem is the complexity of the physical processes involved in use wear. Nonetheless, certain important results have been obtained, for example, the illumination of harvesting techniques in the Near East (Anderson, 1999) (*see* Food Production, Origins of). Other studies, also based on experimentation, have focused on confirming the function of stone and bone points as hunting weapon heads. Such analyses attempt to determine how weapon heads were hafted and used (e.g., thrust or thrown by hand, or with a spear thrower or bow, from long or short distance), and test their durability and performance capabilities (Knecht, 1997).

The Evaluation of Skill in Lithic Technology

Stone tool fabrication also can be analyzed in terms of cognitive processes (*see* Cognitive Archaeology). Being an acquired ideomotor activity, stone knapping requires a period of apprenticeship, and involves the operation of certain psychomotor capacities. Research questions such as the following can thus be addressed:

1. The development of psychomotor capacities during the course of human evolution. For example, it has been proposed that the production of bifacial hand axes with regular edges and a symmetrical form, by evolved *Homo erectus*, required a propositional level of reasoning (Pelegrin), and/or Piaget's rational degree of intelligence (Wynn) (Beaune et al., 2009; de la Torre et al., 2003; Gibson and Ingold, 1993; Nowell and Davidson, 2010).
2. The recognition of different skill levels within a Paleolithic human group, indicating the activities of children and adolescents at different stages of apprenticeship, or a sexual division of activities (Pigeot, 1990; Stapert, 2007).

Conclusion

In the past 30–40 years, approaches to lithic artifacts have undergone a genuine revolution. Starting from the former perspective, wherein the most 'characteristic' tools were selected a priori for the purpose of constructing 'chronocultural' sequences, the approaches to lithic technology have become both more refined and broader. They have been refined by considerable progress in methods of technological analysis of individual objects and of assemblages of objects. The development of practical knowledge through experimental analyses has played a major role in this progress. Approaches have broadened due to the ambition of archaeologists to enhance their field as a human science. From a more ecological and social perspective, researchers are interested in tracing the extraordinary diversity of human behavior since the dawn of humanity, and in understanding its underlying evolutionary mechanisms.

See also: Archaeology, Theory In; Archaeometry; Ceramics in Archaeology; Chronology, Stratigraphy, and Dating Methods in Archaeology; Cognitive Archaeology; Environmental Archaeology; Ethnic Identity and Ethnicity in Archaeology; Ethnoarchaeology; Food Production, Origins of; Hunter-Gatherer Societies, Archaeology of; Survey and Excavation (Field Methods) in Archaeology; Trade and Exchange, Archaeology of.

Bibliography

Anderson, P.C. (Ed.), 1999. Prehistory of Agriculture. New Experimental and Ethnographic Approaches. Monograph 40. Institute of Archaeology, University of California, Los Angeles.

Andrefsky, W., 2009. The analysis of stone tool procurement, production and maintenance. Journal of Archaeological Research 17, 65–103.

Bamforth, D.B., Bleed, P., 1997. Technology, flaked stone technology, and risk. Number 7. In: Barton, C.M., Clark, G.A. (Eds.), Rediscovering Darwin: Evolutionary Theory in Archeological Explanation. Archeological Papers of the American Anthropological Association, pp. 109–139.

de Beaune, S.A., Coolidge, F.L., Wynn, T. (Eds.), 2009. Cognitive Archaeology and Human Evolution. Cambridge University Press, Cambridge, UK.

Berthelet, A., Chavaillon, J. (Eds.), 1993. The Use of Tools by Human and Non-Human Primates. Clarendon Press, Oxford (A Fyssen Foundation Symposium, Oxford Science Publications).

Bleed, P., 2001. Trees or chains, links or branches: conceptual alternatives for consideration of stone tool production and other sequential activities. Journal of Archaeological Method and Theory 8, 101–127.

Bordes, F., 1950. Principes d'une méthode d'étude des techniques de débitage et de la typologie du Paléolithique ancien et moyen. L'Anthropologie 54, 19–34.

Bradley, R., Edmonds, M., 1993. Interpreting the Axe Trade: Production and Exchange in Neolithic Britain (New Studies in Archaeology series). Cambridge University Press, Cambridge, UK.

de la Torre, I., Mora, R., Dominguez-Rodrigo, M., de Luque, L., Alcala, L., 2003. The Oldowan industry of Peninj and its bearing on the reconstruction of the technological skills of Lower Pleistocene hominids. Journal of Human Evolution 44, 203–224.

Delagnes, A., Rendu, W., 2011. Shifts in Neandertal mobility, technology and subsistence strategies in western France. Journal of Archaeological Science 38 (8), 1771–1783.

Dobres, M.A., Hoffman, C.R. (Eds.), 1999. The Social Dynamics of Technology: Practice, Politics, and World Views. Smithsonian Institution Press, Washington D.C.

Féblot-Augustins, J., 2009. Revisiting European Upper Paleolithic raw material transfers: the demise of the cultural ecological paradigm? In: Adams, B., Blades, B.S. (Eds.), Lithic Materials and Paleolithic Societies. Wiley-Blackwell Publishing, Hoboken, NJ, pp. 25–46.

Gibson, K.R., Ingold, T. (Eds.), 1993. Tools, Language and Cognition in Human Evolution. Cambridge University Press, Cambridge, UK.

Hayden, B., 1998. Practical and prestige technologies: the evolution of material systems. Journal of Archaeological Method and Theory 5 (1), 1–55.

Hayden, B., Nora, F., Spafford, J., 1996. Evaluating lithic strategies and design criteria. In: Odell, G.H. (Ed.), Theory and Behavior from Stone Tools. Plenum Publishing, New York, pp. 9–49.

(Series: Vertebrate Paleobiology and Paleoanthropology). In: Hovers, E., Braun, D.R. (Eds.), 2009. Interdisciplinary Approaches to the Oldowan. Springer, New York.

Inizan, M.-L., Reduron-Ballinger, M., Roche, H., Tixier, J., 1999. Technology and Terminology of Knapped Stone (Préhistoire de la pierre taillée, 5). CREP, Nanterre (published online 2009 on the ARKEOTEK website).

Johnson, L.L., 1978. A history of flintknapping experimentation: 1838–1976. Current Anthropology 19 (2), 337–372.

Keeley, L.H., 1980. Experimental Determination of Stone Tool Uses. The University of Chicago Press, Chicago and London.

Knecht, H. (Ed.), 1997. Projectile Technology. Plenum Press, New York.

Kuhn, S., 1995. Mousterian Lithic Technology: an Ecological Approach. Princeton University Press, Princeton, NJ.

Leroi-Gourhan, A., 1993. Gesture and Speech (*Le geste et la parole*, 1964, Albin Michel, Paris, Trans.). MIT Press, Cambridge, MA.

Nowell, A., Davidson, I. (Eds.), 2010. Stone Tools and the Evolution of Human Cognition. University Press of Colorado, Boulder, CO.

(Interdisciplinary Contributions to Archaeology Series). In: Odell, G.H. (Ed.), 1996. Stone Tools: Theoretical Insights into Human Prehistory. Plenum Press, New York.

Odell, G.H., 2000. Stone tool research at the end of the millenium: procurement and technology. Journal of Archaeological Research 8 (4), 269–331.

Pelegrin, J., 2012. New experimental observations for the characterization of pressure blade production techniques. In: Desrosiers, P. (Ed.), The Emergence of Pressure Blade Making: from Origin to Modern Experimentation. Springer Editions, New York, pp. 465–501.

Perlès, C., 1992. In search of lithic strategies: a cognitive approach to prehistoric chipped stone assemblages. In: Gardin, J.-C., Peebles, C.S. (Eds.), Representations in Archaeology. Indiana University Press, Bloomington, IN, pp. 223–247.

Pigeot, N., 1990. Technical and social actors: flintknapping specialists at Magdalenian Etiolles. Archaeological Review of Cambridge 9 (1), 126–141.

Roche, H., Blumenshine, R.J., Shea, J.J., 2009. Origins and adaptations of early Homo: what archaeology tells us. In: Grine, F.E., Fleagle, J.G., Leakey, R.E. (Eds.), The First Humans – Origin and Early Evolution of the Genus *Homo*. Springer, New York, pp. 135–147.

Schlanger, N., 2005. The chaîne opératoir. In: Renfrew, C., Bahn, P. (Eds.), Archaeology: the Key Concepts. Routledge, London, pp. 25–31.

(British Archaeological Reports International Series 1596). In: Schurmans, U.A., De Bie, M. (Eds.), 2007. Fitting Rocks: Lithic Refitting Examined. Archaeopress, Oxford, UK.

Soressi, M., Geneste, J.-M., 2011. The history and efficacy of the chaîne opératoire approach to lithic analysis: studying techniques to reveal past societies in an evolutionary perspective. PaleoAnthropology. http://dx.doi.org/10.4207/PA.2011.ART63 (Special issue: Tostevin, G.B. (Ed.), Reduction Sequence, Chaîne Opératoire, and Other Methods: the Epistemologies of Different Approaches to Lithic Analysis), pp. 334—350. (Online by the Paleoanthropology Society.)

Stapert, D., 2007. Neanderthal children and their flints. PalArch's Journal of Northwest Europe 1 (2), 16–39.

Litigation

Herbert M Kritzer, University of Minnesota Law School, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Litigation is a highly structured process of dispute resolution that invokes the power of the state, or a contractually agreed-upon private decision-maker, to provide a means to authoritatively adjudicate a dispute between two or more parties. Key theoretical issues for understanding litigation include rights, procedures, incentives, and uncertainty. Rights focus on who has access to litigation and under what circumstances; procedures deal with the processes employed in the litigation process (with a major distinction between adversarial and investigatory procedures); incentives turn on both what can be achieved through litigation and how the costs of employing litigation are determined and distributed; uncertainty reflects the strategic nature of litigation. The frequency of litigation varies among countries, but precise estimates of the quantity or rate of litigation are difficult to determine, due to differences in what is included and what is counted. Litigation is typically conducted through agents (lawyers) who have their own set of incentives. Litigants can be divided into one-shot and repeat players, with the latter enjoying a range of advantages reflecting resources, experience, and time horizon. Litigation procedures are frequently criticized as costly and slow; for these reasons, reform of those procedures occurs with some regularity.

Definition

Litigation is a highly structured process of dispute resolution that invokes the power of the state, or a contractually agreed-upon private decision-maker, to provide a means to authoritatively adjudicate a dispute between two or more parties. In many countries in a large proportion of disputes or cases involving litigation, the parties reach resolutions without completing the full litigation process, either by voluntary settlement, abandonment by one party, or some intermediate authoritative decision short of the full process. Unless there is some legal requirement that the adjudication process be invoked, there is typically an even larger set of disputes that never enter into the formal litigation process but are resolved with the knowledge that the process could be invoked and that anticipate the result the litigation process is likely to produce. Most cases of litigation involve disputes among private parties, either individuals or organizations (businesses, corporations, and so forth). Some cases involve the government as a party to the dispute, and especially in federal systems, the parties on both sides of a case can be governmental entities.

Adjudication involves three key elements: (1) authoritative resolution of a dispute (2) by a neutral third party (3) through the application of preexisting norms or rules. Although the courts customarily are associated with adjudication, there is nothing that limits adjudicatory dispute resolution to courts or other governmental bodies; private adjudication systems are common in many countries and transnationally. Also, while the preexisting norms or rules most often associated with adjudication comprise what is normally called 'law,' there is nothing that precludes adjudication from relying upon any agreed set of norms or rules, including those derived from custom or religious texts.

Litigation is a particular form of adjudication that is based on a set of formalized procedures. Formalization is a matter of degree. There are forms of adjudicative dispute resolution that dispense with most aspects of formalized process; examples include small claims courts, qadi courts, or other kinds of community-based courts. On the formality continuum, there is no clear line separating what is appropriately labeled litigation and what is not.

When a governmental entity serves as the adjudicator and oversees the litigation process, it usually carries the label 'court,' although there are examples where the labels 'commission,' 'tribunal,' 'board,' or 'ombudsman' appear. In some governmental systems, some or all of the courts are formally independent of the other branches of government, while in other systems the courts are more integrated into the executive. Even in systems where the courts are independent of the executive, there are often specialized adjudicatory forums within the executive, possessing varying degrees of independence. What is labeled a 'court' in one system may be labeled a 'tribunal' or a 'commission' in another. For example, in Germany there is a system of 'labor courts' to handle litigation involving many types of disputes arising from the workplace; in Britain some of the cases that would go to labor courts in Germany go to 'Industrial Tribunals,' and in the United States the forum is often called a Labor Relations or Employment Relations 'Commission' or Workers' Compensation Board. Even within a single country one finds entities labeled as 'courts' (the U.S. Tax Court) or as something else ('Tax Appeals Commission'), although both perform similar adjudicatory activities.

For a dispute-resolution process to be a form of litigation requires only that the method of resolution be adjudication, as defined above, and that there is a formalized, structured process governing the forum's activities. Litigation involves a wide range of issues, and the cases range from the most routine, mundane disputes to mega-disputes involving large numbers of people, huge sums of money, or both. Because of this range, generalizations are difficult, and it is best to think in terms of multiple worlds of litigation reflecting differing scales of activities and issues (Hensler, 1993).

Theoretical Issues in Understanding Litigation

The highly structured nature of litigation derives from a combination of rights, procedures, and incentives. Each of

these elements influences who uses litigation, when they choose to use it, and the specific nature of the litigation process.

Rights

Rights refer first to who is granted access to the mechanism of litigation. In the United States and Canada, there are complex rules governing who has 'standing' to bring an action in court, and under what circumstances they may bring that action. For example, under the US Constitution, the jurisdiction of federal courts is limited to 'cases and controversies,' and only those who are in fact affected by the case or controversy may bring an action. In other countries (e.g., Israel or Germany), many kinds of actions can be brought by persons who have only a tangential connection to the harm or the issue in dispute. In still other countries, litigation is used regularly as a means of obtaining advisory opinions from judicial bodies, and there need not be any immediate case or controversy in the American sense.

The legal rules and norms within a country or jurisdiction control the kinds of substantive issues subject to litigation. For example, in England tenants often have what amount to a property interest in their tenancy; this can lead to costly disputes and litigation over whether a property owner can obtain the use of property held under such a tenancy. In the United States, there are very detailed protections afforded certain categories of employees, and there is significant litigation over those protections; in England, there are many fewer specific protections, but a more general limitation on an employer's ability to discharge employees without just cause.

In addition to whether the issues are properly subject to litigation, the question of whether there are avenues of redress other than litigation is important. In England, consumers may bring claims for defective goods, but they may also turn to authorities to enforce criminal laws vis-à-vis the duties of sellers of consumer goods (i.e., consumer protection can be dealt with under criminal law as well as through civil litigation). Despite substantial social and cultural similarities, litigation over torts arising from auto accidents is lower in the Netherlands than in neighboring areas of Germany in part because insurance companies in the Netherlands have institutionalized mechanisms for resolving claims without litigation (Blankenburg, 1998).

The remedies available also have a great deal to do with whether litigation is invoked. In the United States, mass disasters such as fires, airplane crashes, and the like typically result in litigation because American law provides substantial remedies to the victims of such events. In contrast, where remedies are severely limited, litigation is less likely. For example, in 1987, an English cross-channel ferry capsized due to crew negligence, resulting in almost 200 deaths; English law strictly limited damages, and the ferry company and its insurers quickly resolved the victims' claims. A second example of the impact of available remedies involves persons in the United States claiming job discrimination on some basis other than race. Until the early 1990s, civil rights laws limited remedies to lost wages and/or reinstatement; this typically made litigation an unattractive option. In 1991, legislation was passed permitting punitive damages in intentional discrimination cases, making redress through litigation more feasible (Farhang, 2010).

Procedures

A set of rules of procedure govern litigation, whether in a court, an administrative tribunal, or a private setting. These rules determine the form by which litigants initiate claims, provide for mechanisms to inform and establish jurisdiction over defendants or respondents, set procedures for gathering the information necessary for the adjudicator to reach a decision, and establish decision-making rules (standards of proof, who makes the decision, and so forth). Procedural rules also determine the availability of aggregation procedures (e.g., class actions), who bears the burdens of the cost of litigation (i.e., fee shifting), and who actually serves as the adjudicator of issues of fact and issues of law (professional judges, lay juries, or some mixed form). Specific rules may be set by the forum, by the parties, or by some combination of the two, depending upon the structure of the forum. Although there are many variations in specific rules, there are two very broad procedural approaches to litigation that are associated with the two dominant Western legal traditions: common law systems and civil law systems.

Procedure in common law systems is usually based on a model of party adversity, with the adjudicator playing the role of umpire. The parties and their agents have primary control over the flow and process of litigation, with the adjudicator receiving what is offered and forming a decision based on what the parties present. Procedure in civil law systems relies on a model variously described as 'inquisitorial' or 'investigatory,' where the primary responsibility for determining and requesting information lies with the adjudicator. Under the common law system, the formal adjudication tends to be concentrated around an event called a 'trial,' when the two sides present their evidence and witnesses to the adjudicator. Under the civil law system, the formal proceedings are more episodic, with the adjudicator determining what additional evidence or 'proofs' he or she would like to see based on what transpires at a particular session. A second major difference between the two traditions is the relative balance placed on oral versus written processes, with common law procedures traditionally involving a heavier oral component and civil law countries placing a greater emphasis on written materials. Developments in recent years have moved the two systems closer together, with civil systems reducing the episodic nature of adjudication procedures and reducing the reliance on written materials, while common law systems have moved toward increased reliance on written materials.

The specific rules create a basis for analyses of litigation by social scientists. These analyses may be process based, building on the flow of litigation. One approach models litigation as a series of stages and transitions, with litigation progressing from the initial harm through the initiation of a dispute into the formal litigation process, culminating in a trial and possibly an appeal (Felstiner et al., 1980–81; Kritzer, 1991b: pp. 401–402). Other analyses rely on tools such as game theory and economic modeling (e.g., Bebchuk, 1984; Hause, 1989). These latter analyses provide predictions of relative advantages and

disadvantages created by the rules, as well as models of how the parties and their agents should behave under assumptions of economic rationality and uncertainty.

Incentives

Litigation frequently has significant financial implications, in terms of the costs of litigation and/or the resulting remedies. The result is a set of incentives that serve to structure the decision making of the parties and their agents. The first incentive is the likelihood of success, both in terms of 'winning' and in terms of obtaining the desired redress; those who must bear the cost of litigation are sensitive to whether the litigation will succeed. The second incentive is the size of potential remedies, which is a function of the severity of the alleged harm, the law, or rules governing the awarding of remedies (Atiyah, 1987), and possibly the nature of the adjudicator of the remedy (professional or lay, single or group, expert or nonexpert). The third incentive is the costs associated with litigation: institutional fees, costs of representation, and expenses for investigation, experts, and so forth. The fourth incentive is the fee regime governing the payment system for legal representation and associated costs (Kritzer, 2009; Prichard, 1988). A fee regime is defined by three elements: who pays fees and costs, how fees and costs are computed, and how fees and costs are regulated or reviewed. In a number of countries, recent years have seen a range of changes to factors relating to the litigation incentive structure. These changes include limitations on damages that can be obtained ('damage caps'), the adoption of various forms of no win, no pay fees (e.g., 'conditional fees' in England, percentage-based fees in Ontario), the reduction of government-funded legal aid to pay these in some countries, movement away from traditional 'loser pays' rules in England, the development of new forms of insurance such as 'after-the-event' insurance to cover the potential costs payable to the winning party in a 'loser pays' system (Lewis, 2011), and the emergence of third-party litigation finance companies, which essentially fund some or all of a party's costs in return for a share of any recovery (Tuil and Visscher, 2010; Hodges et al., 2012; Shepherd, 2012).

As with procedure generally, incentive structures lend themselves to economic analyses. Extensive formal analyses have been made of rules governing cost allocation and their expected impacts on decisions to litigate, decisions to go to trial or to settle (Reinganum and Wilde, 1986), the likelihood of frivolous litigation (Polinsky and Rubinfeld, 1993), and the role of attorney fee arrangements (Johnson, 1980–81; Shavell, 1982; Hylton, 2012). In addition to theoretical analyses, procedural rules and the resulting incentives have been the subject of extensive empirical analyses, most heavily in the United States but also in England and at least one country on the continent of Europe (Kritzer, 2004; Helland and Tabarrok, 2003; Fenn et al., 2002; Yarrow, 1998; Moorhead, 2010; Faure et al., 2006; Lewis et al., 2006; Morris, 2007).

The Role of Uncertainty

One of the central features of litigation is uncertainty, and the decisions of actors in the litigation system must be made in light of this uncertainty (Bebchuk, 1984; Daughety and Reinganum, 2005). Both the degree and source of uncertainty vary depending on rules, the substance of the case, and the general structure of the system. Key elements of uncertainty include which side the adjudicator will favor, the size or nature of the remedy that the adjudicator will award, the length of time the litigation process will take, and the costs that will be incurred. Part of the uncertainty reflects the strategic nature of all conflict, and part of it reflects the introduction of a third-party, neutral decision-maker, combined with formalized procedures for influencing that decision-maker. Uncertainty creates risk, and the parties involved in litigation differ significantly in their ability and willingness to bear the burdens created by risk (Galanter, 1974).

Empirical Results Concerning Litigation

The most extensive empirical research on litigation has been carried out in the United States, although there is also a growing body of empirical literature focusing on litigation in other common law countries, many European civil law countries, and countries from other regions of the world (Japan, India, the Philippines, and in recent years, China).

Although there is a tendency to speak of litigation as a singular phenomena, there are clearly multiple worlds of litigation (Hensler, 1993). Those worlds can be divided along at least three dimensions: the amount at issue, the types and numbers of parties involved, and the substance of the dispute leading to litigation. Some studies focus on a particular type of litigation (e.g., medical negligence or divorce) while other studies speak in broad terms. Some studies look at prominent cases or sets of cases, such as the litigation surrounding Agent Orange (Schuck, 1986) or the litigation concerning Bendectin (Sanders, 1992), while others focus on routine litigation (Kritzer, 1990) or small claims (Yngvesson and Hennessey, 1975; Vidmar, 1984; Baldwin, 1997). Regardless, one must be careful to take note of the types of cases being examined and avoid drawing general conclusions about litigation based on a study that looks at only a slice of the broad litigation phenomenon.

The Likelihood that Disputants Will Turn to Litigation

Research from a variety of countries consistently has shown that relatively small proportions of most disputes eventually involve formal litigation procedures (Genn, 1999; Pleasence et al., 2010; Murayama, 2007; Michelson, 2007). The exceptions involve certain types of cases that by law require formal court action (e.g., dissolution of marriage in many countries), but those types of cases frequently involve only *pro forma* court action. Particularly in common law countries, the image of the filtering process has often been cast as a 'Disputing Pyramid,' with injuries occupying the base of the pyramid and final adjudicatory resolution at the top (Engel and Steele, 1979; Miller and Sarat, 1980–81). The base of the pyramid is broad and the upper areas are narrow; most of the variation, whether by type of dispute or by country, tends to come in the middle regions of the pyramid, reflecting a combination of incentive structures and alternative avenues of redress.

The initiation of litigation presupposes that an injury is perceived, that responsibility for the injury is externalized, and

that a grievant (or an agent of the grievant) seeks redress (Felstiner et al., 1980–81). Most research shows substantial attrition from potential claims to actual claims (Kritzer, 2010), often reflecting that injured parties are never aware that they have suffered an injury, do not attribute to someone else the responsibility for the injury, or do not initiate a claim because of a belief that redress is not practical or it is not needed given alternative sources of compensation. As suggested above, even if litigation is initiated, most cases end before final adjudication. Settlements occur in what has been dubbed the 'shadow of the law' (Mnookin and Kornhauser, 1979), and they reflect anticipations of what judges or juries will decide.

In common law countries, the emphasis in the civil justice process is on proof provided by the parties; in civil law systems, the emphasis is on logical reasoning by the judge. Thus, in civil law countries, the adjudicator plays a more central role, and multilevel review is a more integral and usual part of the process. As a result, there are many more adjudicators in civil law countries. Moreover, the perceived quality of judges making the initial decision is lower in civil law countries compared to common law countries, where the numbers of judges are relatively small and are chosen selectively after years of legal practice. In civil law countries, it is relatively easy to get an initial court decision because the judge controls the process; in common law countries, obtaining an initial court decision is more difficult and costly because the parties can fight each other every step of the way. However, in civil law countries, appeals are more common, making the economic advantage of one system over another unclear.

Cross-National Patterns in Litigiousness

The United States has a reputation as particularly litigious. This in part reflects the important role played by law and courts in the United States from its inception. However, one study that attempted to generate litigation rates *c.*1990 across a wide

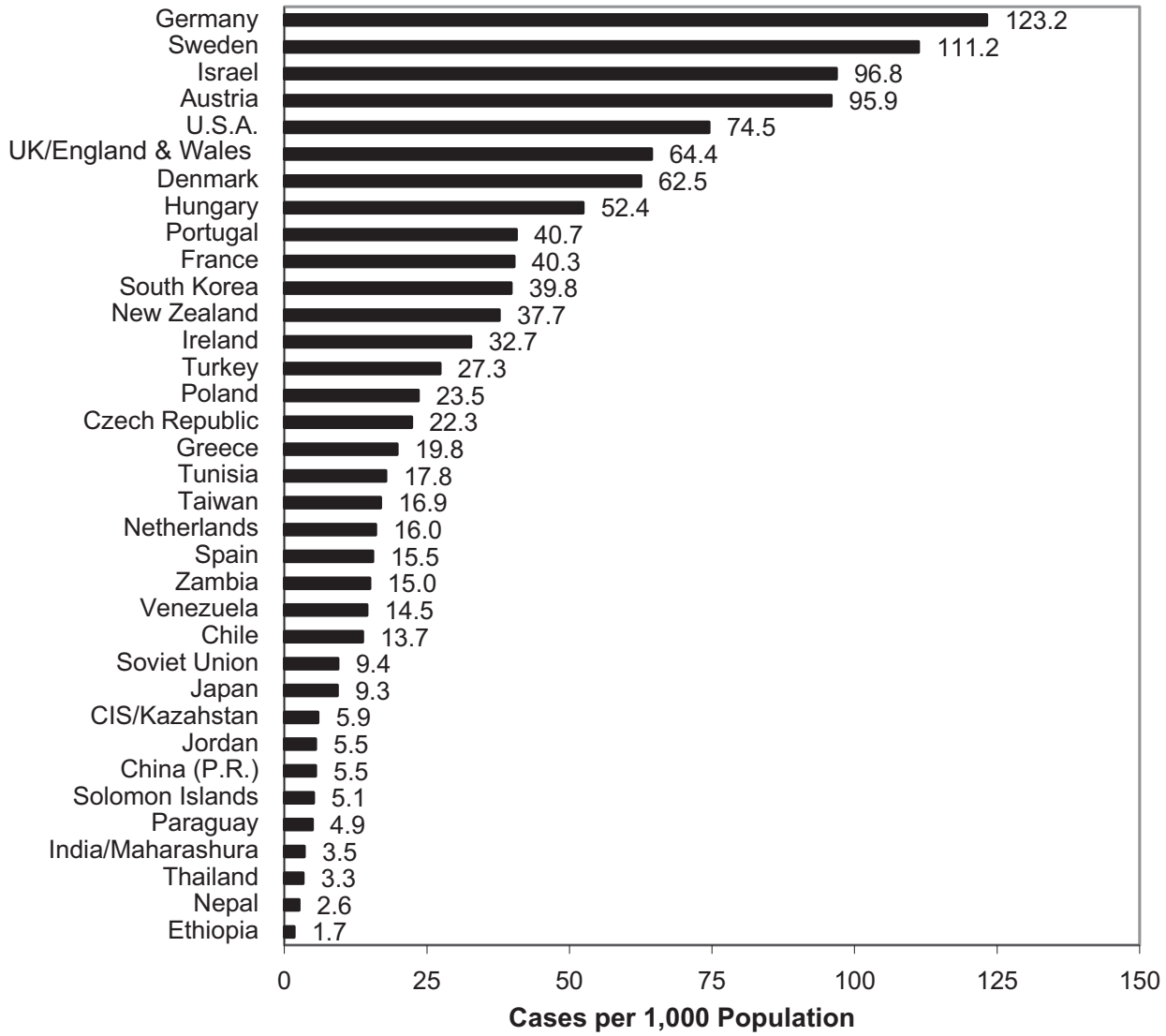

Figure 1 Litigation rates, *c.*1990. Wollschläger, C., 1998. Exploring global landscapes of litigation rates. In: Jurgen Brand and Dieter Strempel (Eds.), Soziologie des rechts: Festschrift für Erhard Blankenburg zum 60 geburtstag. Nomos Verlagsgesellschaft, Baden Baden, pp. 577–588.

range of countries cast doubt on that reputation (Wollschläger, 1998). Figure 1 graphically represents the findings of that study and shows substantial variation in the use of litigation. While the United States is at the higher end of the distribution, it does not stand out as exceptional. However, caution must be exercised in interpreting the information in Figure 1 (and in Figure 2, which will be discussed later) because of the difficult issues that arise in estimating litigation rates across countries.

A first general problem with comparing litigation rates is obtaining comparable information across countries. Countries, and even subunits of countries, differ in institutional arrangements for dealing with various types of problems. For example, in Germany there is a separate group of labor courts to deal with a range of workplace issues; in England, these types of matters are routed to industrial tribunals rather than to courts. In the most states of the United States, claims arising from workplace injuries go to administrative tribunals, but there are a small number of states that have established workers' compensation courts. Tax disputes may go to an administrative body, to a specialized tax court, or in some situations to a general jurisdiction court. A further complication is that in some countries many civil matters that come to court are not litigious; separating out the litigious and nonlitigious matters presents a challenge.

There has been an effort among a range of European countries to develop a system of reporting on those countries' judicial

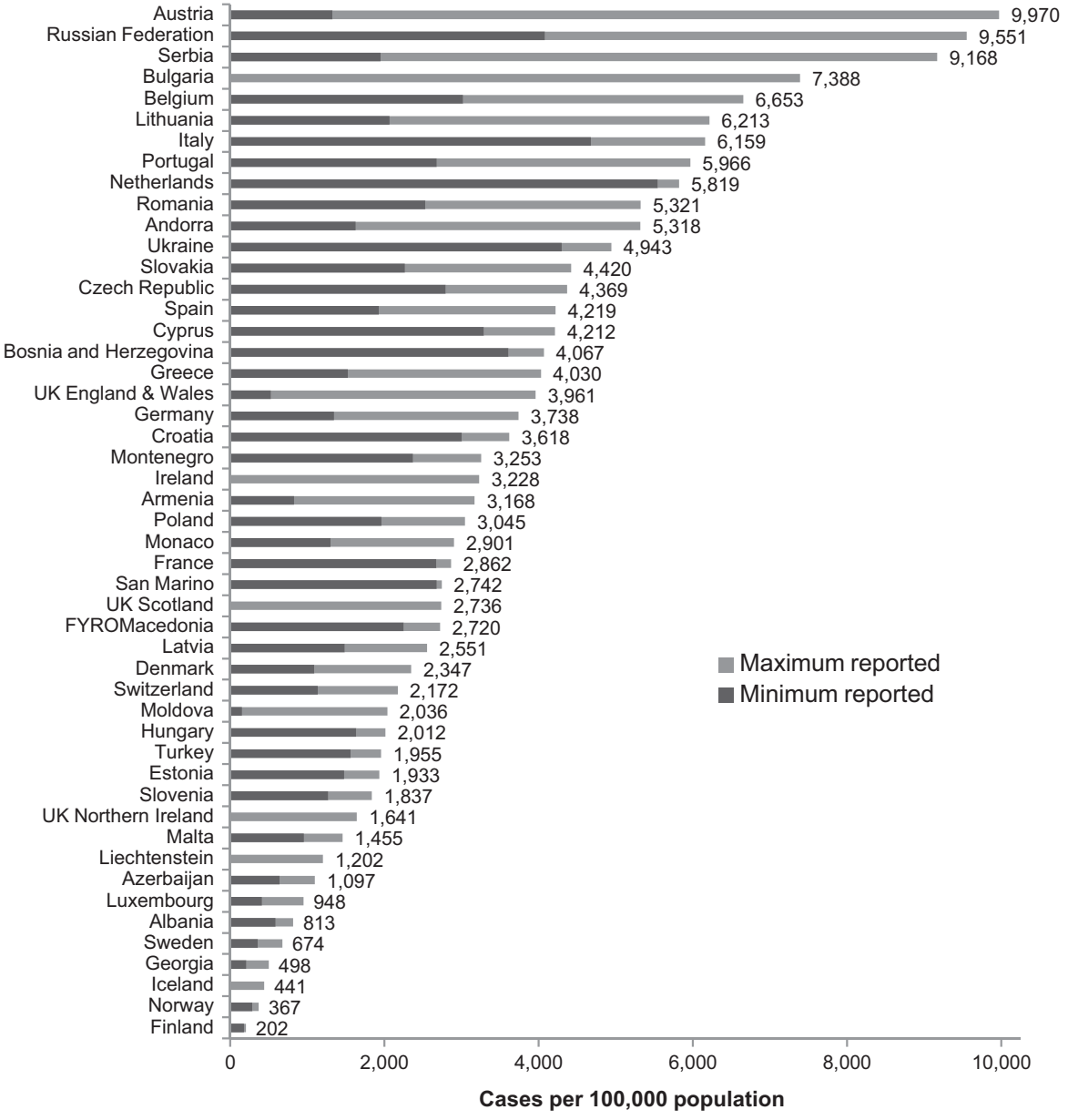

Figure 2 European litigation rates, 2004–10. Reports published by the European Commission for the Efficiency of Justice (CEPEJ).

systems to allow for a measure of comparison (Contini, 2000). The European Commission for the Efficiency of Justice has issued four reports providing estimated litigation rates for 2004, 2006, 2008, and 2010, covering as many as 49 countries in Europe. Figure 2 shows the range of litigation rates reported for each of the 49 countries, with the lighter bar showing the maximum rate reported (or the single rate where only one report was available) and the darker bar showing the minimum rate. For a few countries, it is clear that some change occurred in what was being reported (i.e., the litigation rate for England and Wales was on the order of 3500 per 100 000 population for the first two reports but dropped to less than 550 per 100 000 for the last two reports). Figure 2 is striking for the variation among the European countries, and the inconsistency with patterns in Figure 1. According to Figure 1, Germany was the most litigious of the European countries *c*.1990, followed closely by Sweden and Austria; Spain and the Netherlands stood at the bottom of that figure among European countries. In contrast, Figure 2 puts the Netherlands as third-most litigious and Spain in the top half, with Sweden close to the bottom and Germany in the middle.

A second general problem with comparing litigiousness is determining the baseline that should be used. Almost all studies rely on per capita litigation rates. However, litigation arises out of very specific types of events, the number of which is not simply a function of population. For example, product liability litigation arises from events involving products that are purchased; a country with a higher rate of material consumption may have a higher per capita rate of product liability litigation simply because the transactions that might give rise to such litigation are more common. The problem of using baselines other than simple population is almost never addressed because of the difficulties of developing workable alternatives.

Many explanations have been suggested for whatever variation actually exists. Japan is frequently cited as a country with a very low litigation rate. One explanation for this is that within Japanese culture there is an imperative to avoid conflict (Kawashima, 1973). An alternative explanation is that events such as automobile accidents are dealt with in a way that avoids the need for litigation (Tanase, 1990). A third explanation is that the Japanese elite has intentionally restricted access to litigation as a form of redress; this may be starting to change with the decision in Japan to increase the number of lawyers (*bengoshi*) who are authorized to bring court cases.

The Role that Litigants and Their Agents (Lawyers) Play

Decisions to initiate litigation typically are made by litigants in consultation with their legal advisors. Two elements dominate the decision to litigate. The first is the need to finance the litigation. The second is the likelihood of success, both in terms of whether the desired decision can be obtained and whether the decision will be enforceable. Research has shown consistently that a key element in obtaining redress in tort cases is the availability of a source of payment, either a wealthy defendant or insurance coverage.

Where the potential litigant must directly finance the litigation, the decision to initiate litigation will lie almost entirely with the potential litigant. Lawyers may seek to dissuade potential litigants from pursuing the claim, citing the financial, time, and emotional costs as well as the uncertainty inherent in the process. Where the cost, or at least the risk, of the litigation is to be borne by some other actor, these actors will typically be more involved in the decision to proceed. For example, in the United States, where much litigation is funded by plaintiffs' lawyers through the mechanism of a contingency fee, the lawyer's decision whether to accept a case is crucial; lawyers turn down at least as many cases as they accept (Kritzer, 2004; Daniels and Martin, 2002). When some third party, such as an insurance company (through litigation expense insurance), legal aid, or a trade union, will cover the litigation expense, the third party will make the decision on whether to proceed or require that it be allowed to review and veto any decision to proceed. These decisions are determined by a combination of likelihood of success and the amount at stake.

Nature of Litigants

The nature of litigants typically varies depending on the type of case. Tort (injury) cases are dominated by individuals as plaintiffs, while contract cases more often involve businesses as plaintiffs. Defendants in tort cases are often nominally individuals, but most individual defendants are insured and the insurance company becomes the effective defendant. In contract cases, defendants are distributed among individuals and organizations, with variations depending on the exact nature of the issue at stake.

Litigants are not created equal. Litigation typically requires resources and expertise, and analysts have argued that this conveys an advantage to the wealthy. Galanter (1974) distinguished between "one-shot" and "repeat players" in litigation, arguing that the repeat players, who tend to be 'haves,' should be expected to win more often than the 'one-shotters,' who are likely to be 'have-nots.' While some of the advantages of repeat players reflect resources, other advantages reflect the ability of the repeat player to take a long-run view of litigation. The result is that repeat players are more able to 'play the odds' (i.e., they can act as risk-neutral participants), and more able to 'play for rules' (i.e., try to establish precedents that will work to their advantage over the long run). Galanter's distinction has been examined in a significant body of research conducted in a range of countries (Glenn, 2003).

The importance of the one-shot versus repeat player distinction for understanding how litigants behave is further enhanced by dividing the repeat player into the 'government player,' which may have even fewer economic constraints than other repeat players and which may play by a set of special rules, and the 'routine player,' for whom litigation is genuinely a part of doing business (e.g., the liability insurance company) and which cares little about the outcome of individual cases as long as it is able to comfortably predict the likelihood of winning and losing. Yet another type of repeat player is the political interest group, which uses litigation for explicitly political purposes. Although this is most often discussed in the American context (Kagan, 2001; Silverstein, 2009), it is by no means unique to the United States; particularly as politics becomes more judicialized (Harlow and Rawlings, 1992; Tate and Vallinder, 1995; Kelemen, 2006), these types of litigants are becoming both more common and more important. Developments such as the increasing role of the European

Court of Justice reflect and encourage these changes. Notably, although there has been an increase in interest group litigation, the evidence is inconsistent on whether interest group involvement increases the likelihood of success (Epstein and Rowland, 1991) or whether interest groups actually achieve their ultimate goals even when successful in court (Rosenberg, 1991; Hall, 2011).

Litigation Process

Research on the process of litigation is generally nation specific, because the process is so closely tied to the rules governing that process. Some of the research focuses on the handling of specific types of cases, such as injury cases or divorce cases. Much of this research focuses on the role of the lawyer in the litigation process (Kritzer, 1990) and how lawyers relate to their clients (Rosenthal, 1974; Sarat and Felstiner, 1995); there is disagreement over the degree to which lawyers dominate in the decision-making process (Kritzer, 1998). Other studies focus on specific aspects of litigation, including the impact of specific rules and procedures such as discovery (Brazil, 1980b, 1980a; McKenna and Wiggins, 1998; Mullenix, 1994; Willging et al., 1998), summary judgment (Cecil et al., 2007; Eisenberg and Lanvers, 2009; Kritzer, 1986), the declining incidence of trials (Galanter, 2004; Dingwall and Cloatre, 2006), the rules regarding expert testimony (Cecil, 2005; Edmond, 2008; Cheng and Yoon, 2005; Gatowski et al., 2001), the role the judge in settlement processes (Provine, 1987; Robinson, 2009; Galanter and Cahill, 1994), and procedures and problems in aggregating claims (Hensler et al., 2009). Recent changes in the United States regarding pleading standards in the federal courts are beginning to generate research on how those changes may be impacting the litigation process (Cecil et al., 2011; Moore, 2012). There is no theoretical framework or argument tying together this disparate research, which is heavily US-based.

One area that is ripe for cross-national research is the impact of fee regimes, given that there is more variation in those regimes across countries than within countries. However, any cross-national research on this topic confronts a range of confounding factors that make isolating the impact of the elements of fee regimes extremely difficult. The most successful research has relied upon internal variation in fee regimes (Fenn et al., 2002) or natural experiments arising from changes in fee regimes (Snyder and Hughes, 1990; Hughes and Snyder, 1995). England and Wales are particularly fertile grounds for research in this area, given the changes that have occurred since 1995, when conditional fees were first permitted (Fenn et al., 2002; Yarrow, 1998; Moorhead, 2010; Lewis et al., 2006; Morris, 2007). It is too early to know whether research on the impact of these changes will heighten our empirical understanding of the impact of fee regimes on the litigation process.

Winning and Losing in Litigation

A natural issue in the study of litigation concerns the outcome: who wins, who loses, and why? The definition of winning and losing is problematic: should outcomes be measured in absolute terms or relative terms? If a plaintiff (the initiator of litigation) obtains some redress from the court, has the plaintiff 'won'? What if that very same result could have been achieved without litigation, and what if what the plaintiff sought was something more that it failed to achieve? What if the only reason it could have achieved even that lesser result was because the option of litigation existed? With relatively few exceptions, studies of litigation outcome ignore these complications. A second problem is that the final resolution of litigation is most often a private agreement among the parties. Those cases resolved through formal, final adjudication tend to be a small, nonrepresentative subset. Some scholars argue that the cases resolved through trial are those where uncertainty is high or where there are strongly conflicting views of the likely outcome (Priest and Klein, 1984). Others argue that some types of parties avoid adjudicatory outcomes intentionally when the result will produce unfavorable results, and concentrate their adjudicatory efforts on cases that will produce favorable precedential outcomes (Galanter, 1974).

Many studies have looked at the relative success of different types of litigants (Kritzer and Silbey, 2003), and while the majority of studies have found that repeat players are more successful than one-shot players, the patterns reveal major complications. First, the largest gaps tend to come not between repeat players and one-shot players but between government litigants and nongovernmental litigants (Kritzer, 2003). Second, few researchers ask the question of whether one-shotters are better off with or without litigation as a means of dispute resolution; that is, litigation must be assessed not just internally but externally vis-à-vis its alternatives. For example, are persons injured through medical negligence better off using litigation as the means of redress, relying upon the good will of medical providers, or using a system of insurance that provides for medical expenses, loss of wages, and so forth?

In the United States, there has been substantial research on the outcomes of civil trials, particularly those trials decided by juries (Vidmar and Hans, 2007: pp. 267–338; Daniels and Martin, 1995; Seabury et al., 2004; Litan, 1993). Some of these studies focus on civil cases generally, while others focus on single areas such as medical malpractice (Vidmar, 1995) or product liability (Viscusi, 1986). Typically, these studies show a large proportion of verdicts or decisions involving modest amounts of money, but with a distribution showing an extreme tail consisting of a small number of very large verdicts (Cohen, 2004, 2005a, 2009; Langton and Cohen, 2008). Although over the broad range of cases plaintiffs win about half of the time and defendants win about half of the time (Priest and Klein, 1984), the studies also show a range of plaintiff success depending on the nature of the case, reflecting in part whether the uncertainty is over liability or the amount of damages (Wittman, 1985; Gross and Syverud, 1991).

One specialized theme has been the frequency and nature of punitive (exemplary) damage awards. Research consistently shows that such awards are infrequent, that they tend to be relatively modest, and that there is a strong correlation between compensatory damages and punitive damages when the latter are awarded (Eisenberg et al., 2002; Vidmar and Holman, 2010). Variability in punitive damage awards is greatest for cases with small compensatory awards (Eisenberg et al., 2010). Moreover, the likelihood of a punitive damage award varies depending on the type of case, with such awards most

likely in intentional torts and financial injury cases (Cohen, 2005b; Peterson et al., 1987: pp. 10–12).

Research looking at decisions of American trial judges vis-à-vis questions such as liability and damages is much less common. Research there shows that judges are more likely to find for the plaintiff, although the typical judge's award is about 20% less than the typical jury's (Cohen, 2009; Langton and Cohen, 2008; Clermont and Eisenberg, 1992; Kalven and Zeisel, 1971: pp. 63–65). A danger in all of these comparisons of decisions by judges and decisions by juries is that the cases decided by judges and juries are probably not comparable, because litigants themselves choose between the two types of trials.

Another line of research in the United States focuses on the substantive decisions of judges in civil cases (Carp and Rowland, 1983; Rowland and Carp, 1996; Epstein and Rowland, 1991). This research typically relies upon examination of published decisions by federal trial judges, and distinguishes between decisions labeled as 'liberal' and those labeled 'conservative' (using contemporary American notions of this continuum). The core results of the research demonstrate that there are measurable differences in outcomes depending on whether judges are identified with the Republican or Democratic parties based on either their own affiliation prior to appointment or the affiliation of the President who appointed them. One recognized problem with this research is that published decisions differ from the full population of decisions made by federal district judges in civil cases (Siegelman and Donohue, 1990; Olson, 1989), and results differ when unpublished decisions are included in the analysis, reducing or eliminating the relationship between decisions and partisanship (Ashenfelter et al., 1995).

While substantial attention has focused on formal adjudication by judges and juries, less attention has been directed at the outcomes of cases resolved short of adjudication. Part of the reason for this is that determining outcomes is more difficult because they are not part of the public records. A number of empirical studies (Genn, 1988; Kritzer, 1991a; Hyman et al., 1995; Fenn and Rickman, 1997; Swanson and Mason, 1998) have focused on the settlement process, both in the United States and elsewhere, generally showing that economic considerations (likelihood of success, who is bearing costs, the magnitude of costs both absolutely and relative to stakes) combined with risk preference are central to understanding settlement. Simply put, cases will tend to settle when the plaintiffs net expected gain (recovery discounted by probability of success *minus* expected costs) is less than or equal to the defendants' expected loss (payout discounted by probability of plaintiff success *plus* costs). When stakes are relatively small, this is almost always true (Kritzer, 1991a), and it is frequently true for larger cases even when there is substantial disagreement about the likely outcome. It is less clear to what degree the intervention of court officials can influence the settlement process, in no small part because such a large proportion of cases settle regardless.

Litigation Reform

How to improve the litigation process is a perennial issue (Zuckerman, 1999; Reda, 2012), and a final body of research deals with litigation reform. Although the term 'reform' suggests improvements, whether the changes proposed or implemented constitute improvements is a matter of debate. 'Reform' is used here to refer to change brought about through legislative or administrative action rather than changes resulting from decisions by courts in particular cases. One thread examines the politics of litigation reform including changes that create new opportunities for litigation (Daniels and Martin, 1995; Haltom and McCann, 2004; Burke, 2002; Farhang, 2010).

A second thread of research provides empirically based assessments of some aspect of the litigation system in order to determine what changes might be called for. In England, there have been a series of procedural reforms in recent years. In the mid-1980s, the then Lord Chancellor's Department commissioned a series of studies for what was called the Civil Justice Review. In the mid-1990s, Master of the Rolls, Lord Harry Woolf, undertook a review of the existing civil procedure rules and commissioned a study of litigation costs (Genn, 1996). Most recently, Lord Rupert Jackson of the English Court of Appeal conducted a review of cost-related rules, and again commissioned some empirical research (see the References for the link to the website with materials from Lord Jackson's inquiry).

A third thread of research has examined the impact of various reforms. In the UK, several studies have sought to assess the impact of the Woolf reforms (Fenn et al., 2009; Goriely et al., 2002; Zander, 2009); a second line of UK research has examined the impact of changes to cost and funding rules (Fenn and Rickman, 2011), and whether those changes have encouraged the development of what critics have termed a 'compensation culture' (Moorhead, 2010; Hand, 2010; Morris, 2007; Lewis et al., 2006). In the United States, there has been extensive research seeking to assess the impact of various changes regarding the handling of medical malpractice claims (Hyman et al., 2009; Daniels and Martin, 2010; Studdert et al., 2004; Sharkey, 2005; Pace et al., 2004; Avraham, 2007) as well as research on the impact of tort reform more generally (Eisenberg, forthcoming). One major US study sought to assess the impact of the federal Civil Justice Reform Act, which included the introduction of a variety of procedural changes with an emphasis on increased use of intensive case management (Kakalik et al., 1996).

A final thread of research linked to litigation reform considers the rise of alternative dispute resolution either as a part of the litigation process (e.g., the use of mediation in aid of settlement) or as a replacement for court-based litigation (Genn, 2010: pp. 78–125). Arbitration has a long history but has come to particular prominence in recent years as a means of dealing with everything from large complex international commercial disputes (Dezalay and Garth, 1996) to routine consumer problems (Drahozal and Zyontz, 2010, 2011; Eisenberg et al., 2008) and a range of issues in between (Gross and Black, 2009; Choi and Eisenberg, 2010; Yoon, 2004). Mediation has become extremely popular in recent years, both in the United States and elsewhere, and a substantial body of research concerning mediation now exists (Wissler, 2002; Relis, 2009; Genn et al., 2007; Bogoch and Kaddari, 2007).

See also: Class Actions, Legal; Corporate Law: Publicly Traded Corporations; Courts and Adjudication; Disputes, Social Construction and Transformation of; Empirical Legal Studies; Insurance and the Law; Justice, Access To; Law, Mobilization of; Law, Sociology of; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Negotiation and Bargaining: Role of Lawyers; Torts in Law.

Bibliography

Ashenfelter, O., Eisenberg, T., Schwab, S.J., 1995. Politics and the judiciary: the influence of judicial background on case outcome. Journal of Legal Studies 24, 257–281.

Atiyah, P.S., 1987. Tort law and the alternatives: some Anglo-American comparisons. Duke Law Journal 1987, 1002–1044.

Avraham, R., 2007. An empirical study of the impact of tort reforms on medical malpractice settlement payments. Journal of Legal Studies 36, S183–S229.

Baldwin, J., 1997. Small Claims in the County Courts in England and Wales: The Bargain Basement of Civil Justice? Clarendon Press, Oxford.

Bebchuk, L.A., 1984. Litigation and settlement under imperfect information. Rand Journal of Economics 15 (3), 404–415.

Blankenburg, E., 1998. Patterns of legal culture: the Netherlands compared to neighboring Germany. American Journal of Comparative Law 46, 1–41.

Bogoch, B., Kaddari, R.H., 2007. Co-optation, competition and resistance: mediation and divorce professionals in Israel. International Journal of the Legal Profession 14, 115–145.

Brazil, W.D., 1980a. Civil discovery: lawyers' views of its effectiveness, its principal problems and abuses. American Bar Foundation Research Journal 80, 787–902.

Brazil, W.D., 1980b. Views from the front lines: observations by Chicago lawyers about the system of civil discovery. American Bar Foundation Reserach Journal, 217–251.

Burke, T.F., 2002. Lawyers, Lawsuits, and Legal Rights: The Battle over Litigation in American Society. University of California Press, Berkeley.

Carp, R.A., Rowland, C.K., 1983. Policymaking and Politics in the Federal District Courts. University of Tennessee Press, Knoxville, Tenn.

Cecil, J.S., 2005. Ten years of judicial gatekeeping under Daubert. American Journal of Public Health 75, S74–S91.

Cecil, J.S., Cort, G.W., Williams, M.S., Bataillon, J.J., 2011. Motions to Dismiss for Failure to State a Claim after Iqbal: Report to the Judicial Conference Advisory Committee on Civil Rules. Federal Judicial Center, Washington, DC.

Cecil, J.S., Eyre, R.N., Miletich, D., Rindskopf, D., 2007. A quarter-century of summary judgment practice in six federal district courts. Journal of Empirical Legal Studies 4, 861–907.

Cheng, E.K., Yoon, A.H., 2005. Does *Frye* or *Daubert* matter? A study of scientific admissibility standards. Virginia Law Review 91, 471–513.

Choi, S.J., Eisenberg, T., 2010. Punitive damages in securities arbitration: an empirical study. The Journal of Legal Studies 39, 497–546.

Clermont, K.M., Eisenberg, T., 1992. Trial by jury or judge: transcending empiricism. Cornell Law Review 77, 1124–1177.

Cohen, T.H., 2004. Tort Trials and Verdicts in Large Cities, 2001. Bureau of Justice Statistics, Washington, DC.

Cohen, T.H., 2005a. Federal Tort Trials and Verdicts, 2002–03. Bureau of Justice Statistics, Washington, DC.

Cohen, T.H., 2005b. Punitive Damages Awards in Large Counties, 2001. Bureau of Justice Statistics, U.S. Department of Justice, Washington, DC.

Cohen, T.H., 2009. Tort Bench and Jury Trials in State Courts, 2005. Bureau of Justice Statistics, Washington, DC.

Contini, F., 2000. European Data Base on Judicial Systems. Edizioni Scientifiche Lo Scarabeo, Bologna, Italy.

Daniels, S., Martin, J., 1995. Civil Juries and the Politics of Reform. Northwestern University Press, Evanston.

Daniels, S., Martin, J., 2002. It was the best of times, it was the worst of times: the precarious nature of plaintiff's practice in Texas. Texas Law Review 80, 1781–1828.

Daniels, S., Martin, J., 2010. "It is no longer viable from a practical and business standpoint": damage caps, "hidden victims," and the declining interest in medical malpractice cases. International Journal of the Legal Profession 17, 59–82.

Daughety, A.F., Reinganum, J.F., 2005. Economic theories of settlement bargaining. Annual Review of Law and Social Science 1, 35–59.

Dezalay, Y., Garth, B., 1996. Dealing in Virtue: International Commercial Arbitration and the Construction of a Transnational Legal Order. University of Chicago Press, Chicago, IL.

Dingwall, R., Cloatre, E., 2006. Vanishing trials: an English perspective. Journal of Dispute Resolution 2006, 51–70.

Drahozal, C.R., Zyontz, S., 2010. An empirical study of AAA consumer arbitrations. Ohio State Journal on Dispute Resolution 25.

Drahozal, C.R., Zyontz, S., 2011. Creditor claims in arbitration and in court. Hastings Business Law Journal 7, 77–116.

Edmond, G., 2008. Secrets of the 'hot tub': expert witnesses, concurrent evidence and judge-led law reform in Australia. Civil Justice Quarterly 27, 51–82.

Eisenberg, T., forthcoming. The empirical effects of tort reform. In: Arlen, J. (Ed.), Research Handbook on the Economics of Torts. Edward Elgar, Cheltenham, UK.

Eisenberg, T., Heise, M., Wells, M.T., 2010. Variability in punitive damages: an empirical assessment of the U.S. Supreme Court's decision in Exxon Shipping Co. v. Baker. Journal of Institutional and Theoretical Economics 166, 5–26.

Eisenberg, T., Lanvers, C., 2009. What is the settlement rate and why should we care? Journal of Empirical Legal Studies 6, 111–146.

Eisenberg, T., Miller, G.P., Sherwin, E., 2008. Mandatory arbitration for customers but not for peers: a study of arbitration clauses in consumer and nonconsumer contracts. Judicature 92, 118–123.

Eisenberg, T., Rachlinski, J.J., Wells, M.T., 2002. Reconciling experimental incoherence with real world coherence in punitive damages. Stanford Law Review 54, 1239–1271.

Engel, D.M., Steele, E.H., 1979. Civil cases in society: legal process, social order, and the civil justice system. American Bar Foundation Research Journal 2, 295–346.

Epstein, L., Rowland, C.K., 1991. Debunking the myth of interest group invincibility in the courts. American Political Science Review 85, 205–217.

Farhang, S., 2010. The Litigation State: Public Regulation and Private Lawsuits in the U.S. Princeton University Press, Princeton, NJ.

Faure, M.G., Hartlief, T., Philipsen, N.J., 2006. Funding of personal injury litigation and claims culture: evidence from the Netherlands. Utrecht Law Review 2, 1–21.

Felstiner, W.L.F., Abel, R.L., Sarat, A., 1980–81. The emergence and transformation of disputes: naming, blaming, claiming. Law & Society Review 15, 631–654.

Fenn, P., Gray, A., Rickman, N., Carrier, H., 2002. The impact of conditional fees on the outcome of personal injury cases. Insurance Research & Practice 17, 43–48.

Fenn, P., Rickman, N., 1997. Four offers and a trial: the economics of out-of-court settlement. In: Hawkins, K. (Ed.), The Human Face of Law: Essays in Honour of Donald Harris. Clarendon Press, Oxford, pp. 103–119.

Fenn, P., Rickman, N., 2011. Fixing lawyers' fees ex ante: a case study in policy and empirical legal studies. Journal of Empirical Legal Studies 8, 533–555.

Fenn, P., Rickman, N., Vencappa, D., 2009. The Impact of the Woolf Reforms on Costs and Delay. Nottingham University Business School, Centre for Risk & Insurance Studies.

Galanter, M., 1974. Why the 'haves' come out ahead: speculations on the limits of legal change. Law & Society Review 9, 95–160.

Galanter, M., 2004. The vanishing trial: an examination of trials and related matters in federal and state courts. Journal of Empirical Legal Studies 1, 459–570.

Galanter, M., Cahill, M., 1994. 'Most cases settle': judicial promotion and regulation of settlements. Stanford Law Review 46, 1339–1391.

Gatowski, S.I., Dobbin, S.A., Richardson, J.T., et al., 2001. Asking the gatekeepers: a national survey of judges on judging scientific evidence in a post-Daubert world. Law and Human Behavior 25, 435–458.

Genn, H., 1988. Hard Bargaining: Out of Court Settlement in Personal Injury Actions. Oxford University Press, Oxford.

Genn, H., 1996. Survey of Litigation Costs. Lord Chancellor's Department, London.

Genn, H., 1999. Paths to Justice: What People Do and Think about Going to Law. Hart Publishing, Oxford.

Genn, H., 2010. Judging Civil Justice. Cambridge University Press, Cambridge.

Genn, H., Fenn, P., Mason, M., et al., 2007. Twisting Arms: Court Referred and Court Linked Mediation under Judicial Pressure. Ministry of Justice, London.

Glenn, B.J., 2003. The varied and abundant progeny. In: Kritzer, H.M., Silbey, S.S. (Eds.), In Litigation: Do the "haves" Still Come Out Ahead? Stanford University Press, Stanford, CA, pp. 371–409.

Goriely, T., Moorhead, R., Abrams, P., 2002. More Civil Justice? The Impact of the Woolf Reforms on Pre-action Behaviour. The Law Society, London.

Gross, J., Black, B., 2009. When perception changes reality: an empirical study of investors' views of the fairness of securities arbitration. Journal of Dispute Resolution 2008, 349–410.

Gross, S.R., Syverud, K.D., 1991. Getting to no: a study of settlement negotiations and the selection of cases for trial. Michigan Law Review 90, 319–393.

Hall, M.E.K., 2011. The Nature of Supreme Court Power. Cambridge University Press, New York.

Haltom, W., McCann, M., 2004. Distorting the Law: Reform Politics, Mass Media, and the Litigation Crisis. University of Chicago Press, Chicago, IL.

Hand, J., 2010. The compensation culture: cliché or cause for concern? Journal of Law and Society 37, 569–591.

Harlow, C., Rawlings, R., 1992. Pressure through Law. Routledge, London.

Hause, J.C., 1989. Indemnity, settlement, and litigation, or I'll be suing you. Journal of Legal Studies 18, 157–179.

Helland, E., Tabarrok, A., 2003. Contingency fees, settlement delay, and low-quality litigation: empirical evidence from two datasets. Journal of Law, Economics & Organization 19, 517–542.

Hensler, D.R., 1993. Reading the tort litigation tea leaves: what's going on in the civil liability system? Justice System Journal 16, 139–154.

Hensler, D.R., Hodges, C., Tulibacka, M., 2009. The globalization of class actions. The Annals 622 (entire issue).

Hodges, C., Peysner, J., Nurse, A., 2012. Litigation Funding: Status and Issues. Oxford University.

Hughes, J.W., Snyder, E.A., 1995. Litigation and settlement under the English and American rules: theory and evidence. Journal of Law and Economics 38, 225–250.

Hylton, K.N., 2012. The economics of third-party financed litigation. Journal of Law, Economics & Policy 8, 701–741.

Hyman, D.A., Black, B., Silver, C., Sage, W.M., 2009. Estimating the effect of damages caps in medical malpractice cases: evidence from Texas. Journal of Legal Analysis 1, 355–409.

Hyman, J.M., Heumann, M., Dautrich, K.J., Rubenstein, H.L., 1995. Civil Settlement: Styles of Negotiation in Dispute Resolution. New Jersey Administrative Office of the Courts, New Brunswick, NJ.

Johnson Jr, E., 1980–81. Lawyer's choice: a theoretical appraisal of litigation investment and decisions. Law & Society Review 15, 567–610.

Kagan, R.A., 2001. Adversarial Legalism: The American Way of Law. Harvard University Press, Cambridge.

Kakalik, J.S., Dunworth, T., Hill, L.A., et al., 1996. An Evaluation of Judicial Case Management under the Civil Justice Reform Act. RAND Corporation, Santa Monica.

Kalven, H., Zeisel, H., 1971. The American Jury. University of Chicago Press, Chicago, IL.

Kawashima, T., 1973. Dispute resolution in contemporary Japan. In: Mehren, A. v (Ed.), Law in Japan. Harvard University Press, Cambridge, MA, pp. 48–67.

Kelemen, R.D., 2006. Suing for Europe: adversarial legalism and European governance. Comparative Political Studies 39, 101–127.

Kritzer, H.M., 1986. Adjudication to settlement: shading in the gray. Judicature 70, 161–165.

Kritzer, H.M., 1990. The Justice Broker: Lawyers and Ordinary Litigation. Oxford University Press, New York.

Kritzer, H.M., 1991a. Let's Make a Deal: Negotiations and Settlement in Ordinary Litigation. University of Wisconsin, Madison, WI.

Kritzer, H.M., 1991b. Propensity to sue in England and the United States: blaming and claiming in tort cases. Journal of Law and Society 18, 400–427.

Kritzer, H.M., 1998. Contingent-fee lawyers and their clients: settlement expectations, settlement realities, and issues of control in the lawyer-client relationship. Law & Social Inquiry 23, 795–822.

Kritzer, H.M., 2003. The government gorilla: why does government come out ahead in appellate courts? In: Kritzer, H.M., Silbey, S.S. (Eds.), In Litigation: Do the 'haves' Still Come Out Ahead? Stanford University Press, Stanford, CA, pp. 342–370.

Kritzer, H.M., 2004. Risks, Reputations, and Rewards: Contingency Fee Legal Practice in the United States. Stanford University Press, Stanford, CA.

Kritzer, H.M., 2009. Fee regimes and the cost of civil justice. Civil Justice Quarterly 28, 344–366.

Kritzer, H.M., 2010. Claiming behavior as legal mobilization. In: Cane, P., Kritzer, H.M. (Eds.), Oxford Handbook of Empirical Legal Research. Oxford University Press, Oxford, pp. 260–285.

Kritzer, H.M., Silbey, S.S. (Eds.), 2003. In Litigation: Do the 'haves' Still Come Out Ahead? Stanford University Press, Stanford, CA.

Langton, L., Cohen, T.H., 2008. Civil Bench and Jury Trials in State Courts, 2005. U.S. Department of Justice, Bureau of Justice Statistics, Washington, DC.

Lewis, R., 2011. Litigation costs and before-the-event insurance: the key to access to justice? Modern Law Review 74, 272–286.

Lewis, R., Morris, A., Oliphant, K., 2006. Tort personal injury claims statistics: is there a compensation culture in the United Kingdom? Tort Law Journal 14, 158–175.

Litan, R.E. (Ed.), 1993. Verdict: Assessing the Civil Justice System. Brookings, Washington, DC.

McKenna, J.A., Wiggins, E.C., 1998. Empirical research on civil discovery. Boston College Law Review 39, 785–807.

Michelson, E., 2007. Climbing the dispute pagoda: grievances and appeals to the official justice system in rural China. American Sociological Review 72, 459–485.

Miller, R.E., Sarat, A., 1980–81. Grievances, claims, and disputes: assessing the adversary culture. Law & Society Review 15, 525–565.

Mnookin, R.H., Kornhauser, L., 1979. Bargaining in the shadow of the law: the case of divorce. Yale Law Journal 88, 950–997.

Moore, P.H., 2012. An updated quantitative study of Iqbal's impact on 12(b)(6) motions. University of Richmond Law Review 45, 603–655.

Moorhead, R., 2010. An American future? Contingency fees, claims explosions and evidence from employment tribunals. Modern Law Review 73, 752–784.

Morris, A., 2007. Spiralling or stabilising? the compensation culture and our propensity to claim damages for personal injury. Modern Law Review 70, 349–378.

Mullenix, L., 1994. Discovery in disarray: the pervasive myth of pervasive discovery abuse and the consequences of unfounded rulemaking. Stanford Law Review 46, 1393–1445.

Murayama, M., 2007. Experiences of problems and disputing behaviour in Japan. Meiji Law Journal 14, 1–59.

Olson, S.M., 1989. Studying federal district courts through published decisions.

Pace, N.M., Golinelli, D., Zakaras, L., 2004. Capping Non-economic Awards in Medical Malpractice Trials. RAND Corporation, Santa Monica, CA.

Peterson, M.A., Sarma, S., Shanley, M., 1987. Punitive Damages: Empirical Findings. Rand Corporation, Santa Monica, CA.

Pleasence, P., Balmer, N., Patel, A., Denvir, C., 2010. Civil Justice in England and Wales 2009. Legal Services Commission, London.

Polinsky, A.M., Rubinfeld, D.L., 1993. Sanctioning frivolous suits: an economic analysis. Georgetown Law Journal 82, 397–435.

Prichard, J.R.S., 1988. A systematic approach to comparative law: the effect of cost, fee, and financing rules on the development of substantive law. Journal of Legal Studies 17, 451–475.

Priest, G.L., Klein, B., 1984. The selection of disputes for litigation. Journal of Legal Studies 13, 1–55.

Provine, D.M., 1987. Managing negotiated justice: settlement procedures in the courts. Justice System Journal 12, 91–112.

Radbruch, G., 1946, Gesetzliches Unrecht und übergesetzliches Recht, Süddeutsche Juristen-Zeitung, pp. 105–108.

Reda, D.S., 2012. The cost-and-delay narrative in civil justice reform: its fallacies and functions. Oregon Law Review 90, 1085–1133.

Reinganum, J.F., Wilde, L.L., 1986. Settlement, litigation, and the allocation of litigation costs. Rand Journal of Economics 7, 557–566.

Relis, T., 2009. Perceptions in Litigation and Mediation: Lawyers, Defendants, Plaintiffs and Gendered Parties. Cambridge University Press, New York.

Robinson, P.R., 2009. Settlement conference judge - legal lion or problem solving lamb: an empirical documentation of judicial settlement conference practices and techniques. American Journal of Trial Advocacy 33, 113–165.

Rosenberg, G.N., 1991. The Hollow Hope: Can Courts Bring about Social Change? University of Chicago Press, Chicago, IL.

Rosenthal, D.E., 1974. Lawyer and Client: Who's in Charge. Russell Sage Foundation, New York.

Rowland, C.K., Carp, R.A., 1996. Politics and Judgment in Federal District Courts. University Press of Kansas, Lawrence, KS.

Sanders, J., 1992. The bendectin litigation: a case study in the life-cycle of mass torts. Hastings Law Journal 43, 301–418.

Sarat, A., Felstiner, W.L.F., 1995. Divorce Lawyers and Their Clients: Power and Meaning in the Legal Process. Oxford University Press, New York.

Schuck, P.H., 1986. Agent Orange on Trial: Mass Toxic Disasters in the Courts. Belknap Press of Harvard University Press, Cambridge.

Seabury, S., Pace, N.M., Reville, R.T., 2004. Forty years of civil jury verdicts. Journal of Empirical Legal Studies 1, 1–25.

Sharkey, C.M., 2005. Unintended consequences of medical malpractice damages caps. NYU Law Review 80, 391–512.

Shavell, S., 1982. Suit, settlement, and trial: a theoretical analysis under alternative methods for the allocation of legal costs. Journal of Legal Studies 11, 55–81.

Shepherd, J.M., 2012. Ideal versus reality in thrid-party litigation funding. Journal of Law, Economics & Policy 8, 593–611.

Siegelman, P., Donohue III, J.J., 1990. Studying the iceberg from its tip: a comparison of published and unpublished employment discrimination cases. Law & Society Review 24, 1133–1170.

Silverstein, G., 2009. Law's Allure: How Law Shapes, Constrains, Saves, and Kills Politics. Cambridge University Press, New York.

Snyder, E.A., Hughes, J.W., 1990. The English rule for allocating legal costs: evidence confronts theory. Journal of Law, Economics, & Organization 6, 345–380.

Studdert, D.M., Yang, Y.T., Mello, M.M., 2004. Are damages caps regressive? A study of malpractice jury verdicts in California. Health Affairs 23, 54–67.

Swanson, T., Mason, R., 1998. Nonbargaining in the shadow of the law. International Review of Law and Economics 18, 121–140.

Tanase, T., 1990. The management of automobile accident compensation in Japan. Law & Society Review 24, 651–692.

Tate, C.N., Vallinder, T. (Eds.), 1995. The Global Expansion of Judicial Power. New York University Press, New York.

Tuil, M., Visscher, L. (Eds.), 2010. New Trends in Financing Civil Litigation in Europe: A Legal, Empirical, and Economic Analysis. Edward Elgar, Cheltenham, UK.

Vidmar, N., 1984. The small claims court: a reconceptualization of disputes and an empirical investigation. Law & Society Review 18, 515–550.
Vidmar, N., 1995. Medical Malpractice and the American Jury: Confronting the Myths about Jury Incompetence, Deep Pockets, and Outrageous Damage Awards. University of Michigan Press, Ann Arbor, MI.
Vidmar, N., Hans, V.P., 2007. American Juries: The Verdict. Prometheus Books, Amherst, NY.
Vidmar, N., Holman, M., 2010. The frequency, predictability, and proportionality of jury awards of punitive damages in state courts in 2005: a new audit. Suffolk University Law Review 43, 855–895.
Viscusi, W.K., 1986. The determinants of the disposition of product liability claims and compensation for bodily injury. Journal of Legal Studies 15 (2), 321–346.
Willging, T.E., Stienstra, D., Shapard, J., Miletich, D., 1998. An empirical study of discovery and disclosure practice under the 1993 federal rules amendments. Boston College Law Review 39, 525–596.
Wissler, R.L., 2002. Court-connected mediation in general civil cases: what we know from empirical research. Ohio State Journal on Dispute Resolution, 641–703.
Wittman, D., 1985. Is the selection of cases for trial biased? Journal of Legal Studies 14, 185–214.
Wollschläger, C., 1998. Exploring global landscapes of litigation rates. In: Brand, J., Strempel, D. (Eds.), Soziologie des rechts: Festschrift für Erhard Blankenburg zum 60 geburtstag. Nomos, Baden Baden, pp. 577–588.
Yarrow, S., 1998. The Price of Success: Lawyers, Clients and Conditional Fees. Policy Studies Institute, London.
Yngvesson, B., Hennessey, P., 1975. Small claims complex disputes: a review of the small claims literature. Law & Society Review 9, 219–274.
Yoon, A., 2004. Mandatory arbitration and civil litigation: an empirical study of medical malpractice litigation in the west. American Law and Economics Review 6, 95–134.
Zander, M., 2009. The Woolf reforms: what's the verdict. In: Dwyer, D. (Ed.), Civil Procedure Rules Ten Years on. Oxford University Press, Oxford, pp. 417–434.
Zuckerman, A.A.S. (Ed.), 1999. Civil Justice in Crisis: Comparative Dimensions of Civil Procedure. Oxford University Press, Oxford.

Relevant Websites

http://www.uscourts.gov/FederalCourts/UnderstandingtheFederalCourts/AdministrativeOffice.aspx – Administrative Office of the United States Courts.
http://www.americanbarfoundation.org/index.html – American Bar Foundation.
http://bjs.ojp.usdoj.gov/index.cfm?ty=dcdetail&iid=242 – Bureau of Justice Statistics, Civil Justice Survey of State Courts.
http://www.justice.gc.ca/eng/pi/rs/date.cfm – Department of Justice (Canada).
http://www.coe.int/t/dghl/cooperation/cepej/evaluation/default_en.asp – European Commission for the Efficiency of Justice.
http://www.fjc.gov/ – Federal Judicial Center.
http://iaals.du.edu/ – Institute for the Advancement of the American Legal System (University of Denver).
https://www.gov.uk/government/organisations/ministry-of-justice/about/statistics – Ministry of Justice (England and Wales).
http://www.ncsc.org – National Center for State Courts.
http://www.rand.org/jie/research/civil-justice.html – RAND Institute for Civil Justice.
http://www.judiciary.gov.uk/publications-and-reports/review-of-civil-litigation-costs/index – Review of Civil Litigation Costs (England and Wales).
http://www.esl.eur.nl/arw/sectie_rechtseconomie/rotterdam_institute_of_law_and_economics_rile/ – Rotterdam Institute of Law and Economics (RILE).
http://www.judiciary.gov.uk/publications-and-reports/statistics – United Kingdom Judiciary.

Llewellyn, Karl Nickerson (1893–1962)

Brian Leiter, University of Texas, Austin, TX, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 8999–9001, © 2001, Elsevier Ltd.

Abstract

Karl Llewellyn was the most important figure in American Legal Realism and one of the most influential figures in Anglo-American law of the twentieth century. After some attention to basic biographical facts (education at Yale, time in Germany influence of 'free law' movement, professorships at Columbia and Chicago, role in drafting the Uniform Commercial Code), the article will focus on his major jurisprudential claims concerning the indeterminacy of legal reasoning, and the role of 'situation-types' and un-codified norms in judicial decision-making. The influence of these jurisprudential claims on his seminal work in commercial law will then be explored. Finally, some attention will be given to the influence and reception of his work in analytic jurisprudence, law and economics, and Critical Legal Studies

Biography

Karl Nickerson Llewellyn was born in Seattle, Washington, on 22 May 1893, though the family soon moved to Brooklyn, New York. At age 16, he went to Germany to study for 2 years, before entering Yale College in 1911. When World War I broke out, he was studying in Paris. Staunchly Germanophilic, he joined up with the 78th Prussian Infantry, was wounded later in 1914, and subsequently received the Iron Cross! Returning to Yale, he completed his undergraduate studies and enrolled in Yale Law School, from which he graduated in 1918. Wartime absences of faculty and students led Llewellyn to stay on at Yale for 1918–19 as editor-in-chief of the *Yale Law Journal* and as a part-time instructor teaching, due to curricular needs, commercial law – the field in which he would later become the most influential American figure of the twentieth century. He began working as a lawyer for a New York City bank in 1920, soon joining the prestigious corporate law firm of Shearman and Sterling. He joined the Yale faculty in 1922, moving to Columbia Law School in 1925; Columbia was then a fertile ground of the new 'Legal Realism,' a movement in which Llewellyn soon emerged as the major figure. He taught at Columbia until 1951, though he spent the academic year 1928–29 once again in Germany, as a visiting professor at Leipzig. During this time, he became better acquainted with the German 'free law' movement, which influenced Llewellyn's conception of Legal Realism. In the 1930s, he worked with an anthropologist doing field-work on American Indian law, as well as producing voluminous scholarship in commercial law. From the 1940s until his death, he made his seminal contributions to the Uniform Commercial Code. He moved to the University of Chicago Law School in 1951, which also hired his second wife, Soia Mentschikoff, herself an accomplished legal scholar. He spent the remainder of his career at Chicago, where he died of a heart attack on 13 February 1962.

Llewellyn and American Legal Realism

Llewellyn was the most important and famous figure associated with the American Legal Realism of the 1920s and 1930s, a movement that transformed the way both academics and lawyers thought about law. The Realists argued against the dominant 'mechanical jurisprudence' or 'formalism' of the day, which held that judges decide cases on the basis of distinctively *legal* rules and reasons, which justify a unique result in every case. The Realists argued, instead, that a careful look at how judges *really* decide cases reveals that they decide not primarily because of law, but based on their sense of what would be 'fair' on the facts of the case. Legal rules and reasons figure simply as post-hoc rationalizations for decisions reached on the basis of nonlegal considerations.

Legal Indeterminacy

The Realists argued that the law was 'indeterminate.' By this, they meant that the available class of legitimate legal reasons did not *justify* a unique decision, especially in those cases that reached the stage of appellate review. Llewellyn argued for indeterminacy based on the existence of conflicting, but equally legitimate, canons of interpretation for precedents, and statutes. So, for example, courts had endorsed *both* the principle of statutory construction that, 'A statute cannot go beyond its text,' but also the principle that 'To effect its purpose a statute must be implemented beyond its text' (Llewellyn, 1950: p. 401). But if a court could properly appeal to either canon when faced with a question of statutory interpretation, then the 'methods' of legal reasoning (including principles of statutory construction) would justify at least two different interpretations of the meaning of the statute. In that case, the question for the Realists was: why did the judge reach that result, given that law and legal reasons did not require him to do so? Llewellyn (1930) made a similar argument about the conflicting, but equally legitimate, ways of interpreting precedent, which he called the 'strict' and 'loose' views of precedent. According to Llewellyn, a judge almost always had the latitude to characterize the decision of a prior case in a highly fact-specific way, so as to distinguish it from the present case, or to characterize its holding in a way that abstracted from the specific facts of the earlier case, so as to make it binding in the present case. Thus, a precedent never really constrained a judge in a later case, because he could construe it 'strictly' to distinguish it, or 'loosely' to treat it as binding.

Two Branches of Realism

All the Realists agreed that the law and legal reasons are often indeterminate, so that the best explanation for why judges decide as they do must look beyond the law itself. In particular, all the Realists endorsed the following 'core claim': in deciding cases, judges respond primarily to the stimulus of the facts of the case, rather than to legal rules and reasons.

There was, however, a division among Realists over the question of how to explain why judges respond to the underlying facts of the case as they do. The 'Sociological' Wing of Realism – represented by Llewellyn (1930, 1960, 1962) – thought that judicial decisions fell into *predictable* patterns (though *not*, of course, the patterns one would predict just by looking at the existing rules of law). From this fact, these Realists inferred that various 'social' forces must operate upon judges to force them to respond to facts in similar, and predictable, ways. These social forces might include the economic background of the judges and their professional socialization experiences. It was these factors, rather than rules of law, that account for decisions and determine the responses judges have to the underlying facts of the case.

Llewellyn's focus was on the professional socialization and background of judges that would lead them to be sensitive to the norms of ordinary commercial practice. Focussing primarily on commercial law disputes, Llewellyn argued that what judges really do in these cases is try to enforce the uncodified but prevailing norms from the commercial culture in which the dispute arose. Llewellyn (1960: pp. 122–124) gave this famous example: a series of New York cases applied the rule that a buyer who rejects the seller's shipment by formally stating his objections thereby waives all other objections. Llewellyn notes that the rule seems to have been applied rather harshly in the NY cases, since either the buyers may not have known at the time of rejection of other defects or the seller could not have cured the defects anyway. A careful study of the underlying facts, however, revealed that in each case where the rule seems to have been harshly applied what had *really* happened was that the market had gone sour and the buyer was looking to escape the contract. The judge, being 'sensitive to commerce or to decency' (1960: p. 124) applies the unrelated rule about rejection to frustrate the buyer's attempt to escape the contract. Thus, the commercial norm – buyers ought to honor their commitments even under changed market conditions – is enforced by the courts through a seemingly harsh application of an unrelated rule concerning rejection. It is these 'background facts, those of mercantile practice, those of the situation-type' (1960: p. 126) which determine the course of decision.

Llewellyn's seminal work on the Uniform Commercial Code was informed by his Realist perspective (cf White, 1994). For example, the Code imposes an obligation of 'good faith' in all contractual dealings, which is defined as requiring not only honesty but also 'the observance of reasonable commercial standards of fair dealing in the trade.' But for a judge, then, to enforce the rule requiring good faith is just for the judge to enforce the norms of commercial culture! The reliance of the Code throughout on norms of good faith and 'reasonableness' is a constant invitation to the judge to do what he would, on the Realist theory, be likely to do anyway: enforce the norms of the prevailing commercial practice.

The 'Idiosyncracy Wing' of Realism, by contrast – exemplified most prominently by Frank (1930) – claimed that what determines the judge's response to the facts of a particular case are idiosyncratic facts about the psychology of that individual. Thus, Frank notoriously asserted that 'the personality of the judge is the pivotal factor in law administration' (1930: p. 111). This view is hard to square though with lawyers' ability to predict rather well what judges will do (if they could not do that, litigators, at least, would be out of business). Despite its extreme nature and limited appeal to most Realists, Frank's view has come to dominate the current picture of Realism, to the extent that Llewellyn's far more typical 'sociological' approach is often neglected.

Impact and Current Importance

By calling attention to the role of nonlegal factors in judicial decision-making, Llewellyn and the Realists initiated the interdisciplinary turn in American legal education, and made clear the need for lawyers to draw on the social sciences in understanding the development of law and what judges do. Much contemporary political science literature on law and courts takes its inspiration from Realism by seeking to explain decisions not by reference to legal reasons (which are assumed to be indeterminate) but by reference to facts about the politics, background, and ideology of the judges (Cross, 1997). Recently, it has even been argued that Llewellyn's own approach to commercial law anticipated the now dominant economic analysis of the business law fields (Schwartz, 2000). This should hardly be surprising, though, since the economists seek to structure legal rules to maximize efficient outcomes, but efficiency is surely an important norm of all mercantile practice: thus, if judges enforce the norms of mercantile practice (as Llewellyn claims they do and should do), it will turn out that judges will try, among other things, to produce efficient outcomes.

Llewellyn's work on Native American law and dispute resolution practices (Llewellyn and Hoebel, 1941) remains an anthropological classic, while his introductory lectures to Columbia law students (Llewellyn, 1930) may lay claim to being the most widely read book on law by generations of American law students. While the Uniform Commercial Code (UCC), the single most influential piece of commercial law in the US, bears Llewellyn's imprimatur, it is often lamented in recent years that revisions to the UCC have moved it away from Llewellyn's original vision of a Code sensitive to actual commercial practice. Even so, the lawyers and scholars effecting these changes continually feel the need to grapple with Llewellyn's conception.

Within jurisprudence, Llewellyn's ideas have been subjected to extensive criticism by analytical philosophers (notably H.L.A. Hart), though more recently a philosophical defense has been mounted (Leiter, 1997). Among legal scholars, however, Llewellyn's ideas remain influential, both in scholarly movements (like 'Critical Legal Studies' and 'Law and Society') that emphasize law's indeterminacy and the need to look at extra-legal factors to explain law, but also in the work of judges, who now attend to the specific facts of cases (the 'situation-types' as Llewellyn called them) and general nonlegal policy considerations much more openly than before the triumph of Realism in American law.

See also: Courts and Adjudication; Law and Society: Development of the Field; Law, Sociology of; Law: History of Its Relation to the Social Sciences; Legal Realism; Legal Scholarship.

Bibliography

Cross, F., 1997. Political science and the new legal realism. Northwestern University Law Review 92, 251–326.

Frank, J., 1930. Law and the Modern Mind. Brentano's, New York.

Leiter, B., 1997. Rethinking legal realism: Toward a naturalized jurisprudence. Texas Law Review 76, 267–315.

Llewellyn, K., 1930. The Bramble Bush. Oceana, New York.

Llewellyn, K., 1950. Remarks on the theory of appellate decision and the rules and canons about how statutes are to be construed. Vanderbilt Law Review 3, 396–406.

Llewellyn, K., 1960. The Common Law Tradition: Deciding Appeals. Little, Brown & Co, Boston.

Llewellyn, K., Hoebel, E.A., 1941. The Cheyenne Way: Conflict and Case Law in Primitive Jurisprudence. University of Oklahoma Press, Norman, OK.

Schwartz, A., 2000. Karl Llewellyn and the origins of contract theory. In: Kraus, J., Walt, S. (Eds.), The Jurisprudential Foundations of Corporate and Commercial Law. Cambridge University Press, New York.

Twining, W., 1973. Karl Llewellyn and the Realist Movement. University of Oklahoma Press, Norman, OK.

White, J.J., 1994. The influence of American legal realism on Article 2 of the uniform commercial code. In: Krawietz, W., et al. (Eds.), Prescriptive Formality and Normative Rationality in Modern Legal Systems. Duncker and Humblot, Berlin.

Local Economic Development

Christian M Rogerson, University of Johannesburg, Johannesburg, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Local Economic Development (LED) has emerged globally as a vital planning approach for local economies. The emergence and growing significance of LED planning approaches are inseparable from the changing world economy and especially the advance of globalization. The question of catalyzing LED planning initially emerged in Europe and spread to other advanced economies. The geographical diffusion of LED has occurred to late-developing economies in sub-Saharan Africa and Latin America. The relevance of this place-based approach to contemporary development planning in the global South is widely confirmed. New evidence demonstrates that the implementation of LED approaches can raise human development prospects in the global South.

Globalization is a vital driver for reconfiguring planning approaches toward local and regional development (Christensen and van der Ree, 2008; Rodriguez-Pose, 2008). Arguably, a major facet of globalization is the increasing pervasiveness of markets around the world (Christensen and van der Ree, 2008: 2). Even the most remote places must take steps to adjust to the increased competition that globalization brings (Pike et al., 2006). Amidst circumstances of global economic turmoil, localities are compelled to find new solutions to support local competitiveness as well as to create inclusive development (ILO, 2008). Indeed, local economies are affected more than ever before "by policies and processes formed at the supranational level, such as market liberalisation, expanding global production systems and the changing terms of trade" (ILO, 2006: 2). Overall, therefore, the advance of globalization accords Local Economic Development (LED) strategies "a bigger role to play in international development" (Rodriguez-Pose, 2008: 24). More particularly, LED "offers a means to counteract or take advantages of the forces of globalisation by maximising local potentials" (ILO, 2006: 2). In the reshaped landscape of development planning, LED emerges as a vital new planning focus particularly in the context of pervasive trends toward decentralization – the deliberate and planned transfer of resources away from central state institutions – and of shifting structures of government and governance (Rodriguez-Pose and Tijmstra, 2007; Christensen and van der Ree, 2008; Rodriguez-Pose and Tijmstra, 2009).

The Emergence of LED

The developmental challenges posed by globalization precipitated a serious rethink about the validity of former approaches to development planning. LED approaches are generally distinguished from the so-termed traditional development policies.

The differences between traditional development policies and LED are captured for five domains by Rodriguez-Pose (2001):

1. Whereas in traditional top-down approaches the decision on where to implement development strategies is taken by central government planners and developers, with little or no involvement of local actors, LED practices favor the promotion of development in all territories by using the economic potential and the competitive advantage of every space. The initiative to launch the development strategy is taken locally or with strong local support.
2. As a result of where and how the decisions are taken, traditional policies have been generally designed, managed, and implemented by ministries or central government agencies. The involvement of local actors in LED strategies implies, in contrast, a much greater degree of vertical and horizontal coordination of all the actors involved. Vertical coordination entails the synchronization of local, regional, national, and supranational or international institutions. Horizontal coordination comprises local public and private actors concerned with development issues.
3. The LED approach is territorial rather than sectoral. Traditional policies have tended to adopt a sectoral approach. The promotion of specific industrial sectors that contribute to generate economic dynamism has been one of the main objectives of traditional policies. By contrast, LED uses a territorial approach as a means of achieving economic development. The diagnosis of the economic, social, and institutional conditions of every territory and the identification of the local economic potential are the foundations upon which a local development strategy is constructed.
4. Closely allied to the sectoral approach of most traditional development policies is the development of large industrial projects that were expected to promote additional economic activity and generate the networks and value chains needed to achieve sustainable development. The problems of this type of practice encouraged LED practitioners to identify and use the economic development potential of each area and to stimulate the progressive adjustment of the local economic system to changing economic conditions (Rodriguez-Pose, 2001).
5. The two approaches are set apart by their ways of attracting economic activity. While traditional approaches often rely on a national government for financial support, incentive packages, and subsidies in order to attract and maintain economic activity, LED tends to shun such activities and concentrate on the improvement of basic conditions in the locality for the development and attraction of further economic activity (Rodriguez-Pose, 2001).

As pointed out by several observers, LED is a contested concept and several definitions exist for it. For Bartik (2003: 1), LED represents increases in a "local economy's capacity to create wealth for local residents". The World Bank (2003) considers the activity of LED is about local people working together to achieve sustainable economic growth that brings economic benefits and quality of life improvements for all in the community. From a perspective rooted in the global South, Helmsing and Egziabher (2005: 1) consider LED to be a process wherein development partnerships are initiated among local governments, NGOs, community-based groups, and the private sectors in order to manage existing resources, create employment, and stimulate the economy of a well-defined territory. Other analysts forward that in the absence of a clearly defined theoretical model LED strategies often resort to the essential features of the approach to specify their content (Rodriguez-Pose, 2001: 8). The ILO distinguishes four core features to characterize LED strategies, namely that they require participation and social dialogue; are anchored on a specific territory; entail mobilization of local resources and competitive advantages; and are locally owned and managed. These characteristics coalesce in a definition that LED is a "participatory development process that encourages partnership arrangements between the main private and public stakeholders of a defined territory, enabling the joint design and implementation of a common development strategy, by making use of the local resources and competitive advantage in a global context, with the final objective of creating decent jobs and stimulating economic activity" (ILO, 2006: 2).

Finally, the German International Development Agency (GIZ), which has been actively promoting LED approaches in the global South, holds a similar view while specifying that LED should "tackle market failures, remove bureaucratic obstacles for local businesses and strengthen the competitiveness of local firms" (Ruecker and Trah, 2007: 15). The GIZ also sees LED as an ongoing process and recognizes that "stimulating local economic growth will also involve systemic change at larger scales" (Meyer-Stamer, 2008: 1). Of critical importance is the notion of the 'systemic competitiveness of a territory', which is viewed as the capability of a locality to generate rising income levels and enhance local livelihoods (Meyer-Stamer, 2008: 7).

A scan of international experience reveals different approaches to LED as well as different entry points for starting an LED process. Arguably, LED offers an integrated, context-sensitive approach to development rather than a 'one-size-fits-all' solution. Specific LED strategies for mobilizing local economic potential include innovative interventions relating to inter alia, infrastructure, strengthening local small and medium enterprises (SMEs) and their skills; attraction of foreign direct investment; fostering territorial competitiveness; capacitating local institutions; improved management of the development process; as well as internalizing local resources (Rodriguez-Pose, 2008: 23). Likewise, the ILO (2008: 2) forwards that bringing together local governments, the private sector, and civil society in a search for the right LED formula facilitates "the community to build from the 'inside-out', capitalizing on local assets rather than from the 'outside-in' relying on external interventions" (ILO, 2008: 2). "What links different approaches together is the quest for making local economies robust and creating productive jobs and also the recognition that local or regional competitive advantage rests on local interactions, knowledge spillovers and institutional synergies" (Salazar-Xirinachs, 2008: v). Although LED embodies a clear economic focus, it is not simply about economic growth; rather it is geared ultimately toward attaining a sustainable development trajectory that reconciles often conflicting economic, social, and ecological issues and objectives (Ruecker and Trah, 2007: 15). For some observers, the key contemporary challenge facing LED is "how to make the most of local resources in a way that improves returns from global markets" (Christensen and van der Ree, 2008: 2).

Rodriguez-Pose (2001, 2008) identifies numerous economic and social advantages of LED strategies over those of traditional development programs. The social advantages are that LED strategies empower local societies and generate local dialogues. Until recently people living in many areas of the developing world had little say or control over economic activity occurring in their territory; now they are beginning to adopt a more proactive stance with respect to their own future. Another social benefit is that the preparation and implementation of LED strategies assist in making local institutions more transparent and accountable thereby contributing to developing civil society. The economic advantages of the approach of LED are, however, its most significant strengths (Rodriguez-Pose, 2001, 2008). As LED strategies seek to embed economic activity in a territory and make economic activity dependent on the specific economic conditions and comparative advantages of that place, they generate sustainable employment in enterprises more capable of withstanding changes in the global economic environment.

LED as a Place-Based Policy Intervention

LED policies are part of a broader suite of debates on the validity of so-termed place-based as opposed to 'spatially blind' policies. To a large extent, these debates have been precipitated by the appearance of the World Bank's (2009) *World Development Report*. The Report is highly critical of the heavy focus by certain policy makers in searching for 'spatially balanced development' (World Bank, 2009: 73). The World Bank Report offers the core message that economic development inevitably will be geographically uneven and that to introduce or continue with policies designed to spread out economic activities – spatially or across a national urban system – carries with it the inherent danger that economic growth may be discouraged or arrested (World Bank, 2009). This said, the Report asserts that "development can still be inclusive, in that even people who start their lives far away from economic opportunity can benefit from the growing concentration of wealth in a few places" (World Bank, 2009: xxi). The essential parameters for achieving inclusive development are spelled out as follows: "With good policies, the concentration of economic activity and the convergence of living standards can happen together. The challenge for governments is to allow – even encourage – 'unbalanced' economic growth, and yet ensure inclusive development. They can do this through economic integration – by bringing laggard and leading places closer in economic terms" (World Bank, 2009: 20). It is explained this integration can most

appropriately be undertaken "by unleashing the market forces of agglomeration, migration and specialization, not by fighting and opposing them" (World Bank, 2009: 21).

A vital question posed in the Report is whether governments should invest in people or places. The World Bank favors investments in "activities that produce the highest economic and social returns nationally" (World Bank, 2009: 231). World Bank advice is "in economically leading regions, to invest in durable, export-oriented activities aimed at increasing national growth whereas in lagging areas investments should be in people to poverty reduction" (World Bank, 2009: 231). The key message of the Report is the need to reframe policy debates away from a sole focus on spatial targeting and instead to promote a menu of policy instruments, which might include but will necessarily foreground spatial targeting, in the overarching search for economic integration.

The Report advocates *'spatially blind'* rather than spatially targeted policy interventions. The term 'institutions' is used as shorthand to embrace "all the policy instruments that are spatially blind" (World Bank, 2009: 38); it includes the "amenities that governments should provide to everyone, regardless of place". This is explained as embracing mechanisms for the universal delivery of "basic amenities such as the administration of justice, public security, the regulation of land, labor and capital markets, education and health and electricity, water and sanitation" (World Bank, 2009: 38). Spatially blind policies include "such national policies as the income tax system, intergovernmental fiscal relations, and the governance of land and housing markets as well as education, health care, basic water and sanitation" (World Bank, 2009: 240). By contrast to spatially targeted interventions, it is considered that spatially blind interventions involving universal and progressive income taxation, social security, and unemployment benefits 'work less noisily' (World Bank, 2009: 238). Moreover, "as the base of economic integration, such 'institutions' capture the benefits of spatial concentration of production and deliver convergence in living standards" (World Bank, 2009: 238).

The World Bank's advocacy of spatially blind policies as opposed to *place-based interventions* has created controversy. Among others, Turok (2013: 1) sees 'serious omissions and shortcomings' in the Report as a consequence of reducing the complexity of spatial development to a few simple distinctions. In addition, the Report evidences undue faith in the ability of market forces to reduce spatial inequalities and a "notable lack of faith in the capabilities of the State to stimulate and sustain economic development in lagging areas" (Turok, 2013: 6). For Garcilazo (2011: 466) also the case for place-based policy responses is clear. Although the largest cities contribute disproportionately to national and aggregate economies, the contributions of smaller cities and rural areas should not be overlooked. Accordingly, there is a need for a differentiated approach capable of unleashing the growth potential of these other regions and localities, which would be through place-based policy interventions (including LED). It is maintained therefore that place-based policies are 'best equipped' for the task of unleashing and maximizing the growth potential of all regions and localities (Garcilazo, 2011: 466).

Place-based interventions highlight the importance of place-to-place differences by considering the array of different contextual factors that can influence the potential returns of intervention (Barca et al., 2012). The underlying principle is that tackling underutilized potential and reducing social exclusion may be best achieved by using place-based policy to maximize the growth potential of each locality or region (Tomaney et al., 2011: 626). Barca et al. (2012) note that a place-based development policy builds on local knowledge, incorporates local values, and mobilizes partnerships across different levels of governance, which strengthens institutions (Barca et al., 2012: 147). The success of place-based policies demands that local (and regional) actors be galvanized to assume a constructive and leading role in policy but also that these actors work with others from more distant locales to design and deliver the policies.

Overall, therefore, in a fluid global context of financial crisis, food crises, and the rising political and economic significance of (parts) the global South, place-based solutions are seen as better suited to current and future circumstances than the spatially blind approach advocated by the World Bank (Tomaney et al., 2010, 2011). In particular, support for place-based approaches promotes not just economic growth but also reduced inequality and lower territorial polarization (Tomaney et al., 2011: 626–627).

The Geographical Spread of LED

The origins of modern LED practices trace back to the 1960s. Pressures to stimulate LED came from at least two main directions: the persistence of uneven and inequitable economic outcomes and concerns about the dearth of local investment in certain regions (Le Heron, 2009: 93). Although interest in LED first emerged in Western Europe and North America, during the 1990s in particular LED ideas and practices spread from the global North to the global South (Nel and Rogerson, 2005; Rodriguez-Pose and Tijmstra, 2009).

Several recent writings affirm the continued global relevance, if not increased importance, of local (and regional) development policies in the context of the rapid unfolding of events that are reconfiguring the global economy (Rodriguez-Pose and Tijmstra, 2009; Pike et al., 2011). A combination of seismic international events has altered the prospects for local and regional developments (Tomaney et al., 2011: 618). The global financial crisis that unfolded after 2007 has profoundly influenced the prospects for LED internationally (Tomaney et al., 2011). Other critical considerations are the growing significance and expansion of South–South trade with the formation of new economic blocs, such as Brazil, Russia, India, China, and South Africa (BRICS) and the mounting acknowledgment of the need to implement adaption and mitigation strategies in relation to the consequences of global climate change. Furthermore, changes in international financial regulations, investment flows, and trade are also anticipated to affect local development futures (Tomaney et al., 2010: 773). Other issues relevant to LED futures include the management of natural resources, climate change, food security considerations, and access to water and energy.

In the contemporary, global economy localities are increasingly viewed as pivotal sites of competitiveness. Most importantly, the works of Pike et al. (2006), Rodriguez-Pose (2008), and Crescenzi and Rodriguez-Pose (2011) show that

in a world of rapidly reorganizing economic activity, the ability of the 'top-down' policy paradigm to deal with these challenges is called into question. Instead, as argued by Crescenzi and Rodriguez-Pose (2011: 2), the category of 'bottom-up' local development policies have been mobilized to address this operational vacuum. Bottom-up policies can have an impact on the natural economic potential, which is viewed as growing from within localities (Pike et al., 2006: 155). Through empowering local actors and making them directly responsible for the design of their own collective response to the challenges of the changing global economy, bottom-up policies often are considered effective alternative pathways to those offered by traditional approaches to development (Crescenzi and Rodriguez-Pose, 2011: 2).

Overall, LED policies are considered to represent a critical response to the territorial challenges produced by globalization, a response that is relevant to both advanced and developing countries. This said, the greatest challenges for LED are in the global South, in particular Latin America and sub-Saharan Africa. In poor and late-developing countries, including those in Latin America and sub-Saharan Africa, the recent emergence of local development policies represents a similar response to that which occurred in the global North, namely "an answer on behalf of localities and territories to the challenges of poverty, productive restructuring and increased competition" (Vazquez-Barquero, 2011: 509). At the local scale in Latin America, experiments in several countries have been undertaken with the innovation of new radical local development interventions to directly address high local levels of poverty, informality, inequality, and development (Bateman et al., 2011). Colombia is seen as in the vanguard of this new movement starting with its capital city of Bogota. Nevertheless, Bateman et al. (2011: 1–2) identify the city of Medellin as the "pioneer in developing a new Post-Washington consensus local economic policy". At its core is the notion that follows that of the successful industrial districts in Western Europe that subnational levels of government can, and should, be proactive in building the institutional and organizational infrastructures required for the successful emergence, expansion, and upgrading of growth-oriented micro-, small, and medium enterprises (Bateman et al., 2011: 2). Of special importance is that in Medellin the local developmental state model was anchored upon what is described as "creative forms of local state ownership and control of key enterprises" (Bateman et al., 2011: 2). In Medellin, financing was secured through local state ownership of key enterprises and assets that produced a revenue flow and specifically involved city ownership and effective management of the main energy company, which was mandated to channel 30% of its net annual profit to the city budget. This financial bounty allowed the city the fiscal space to engage in economic and social development areas beyond those that otherwise have been possible. Two examples are the establishment of a propoor transport system that allows the poor to work in more distant areas as well as for informal enterprises to sell their goods through improved access to the city center and the initiation of a network of publicly funded business support centers located in the city's poorest areas and targeted to support business development by the poor with free business support services and technical advice to aspirant entrepreneurs as well as a microcredit facility.

Bateman et al. (2011) have argued that this propoor model of LED from Medellin offers significant lessons for Latin America and elsewhere in relation to the potential and pitfalls of proactive LED policy operating in extremely marginal communities. The intervention support for microenterprises resulted in increasing the number and quality of microenterprises operating in the poorest communities creating an equality of opportunity for establishing a functioning business. One shortcoming has been that policy support so far has mainly been upon microenterprise as opposed to linking to issues of higher productivity SMEs, which is advocated in the concept of productive development policies. Nevertheless, the Medellin case is viewed as a promising demonstration of the local developmental state and a positive response to what are considered the failures of neoliberalism. Improved intervention is necessary, however, in particular with regard to improving transfer of technology, innovation, and knowledge to enhance the productivity of small- and medium-sized businesses (as opposed to microenterprises) (Bateman et al., 2011).

Across sub-Saharan Africa the rising importance of LED is signaled by the activities of the United Cities and Local Governments of Africa (UCLGA) (Swinburn and Yatta, 2006). The UCLGA is the united voice and representative of local government in Africa, an association of local governments, individuals, and entities interested in African local government. As part of UCLGA's mission of "Building African Unity from within and Driving Development through the Grass Roots", LED is identified as one of six key priority areas for action for the period 2006–15. By identifying LED as a priority for urban Africa, Swinburn and Yatta (2006) concede national-level macroeconomic policies no longer are sufficient for attaining sustainable broad-based economic growth. Moreover, as globalization diminishes the powers of the nation state in Africa and correspondingly underlines the role of localities and wider city regions as drivers of national growth, traditional sectoral supply side development strategies are not achieving hoped-for outcomes (Swinburn and Yatta, 2006). Within Africa, the South African case is usually highlighted as exceptional for the extensive devolution and decentralization of powers that have taken place (Nel and Rogerson, 2005; Rodriguez-Pose and Tijmstra, 2007; Rogerson and Rogerson, 2010). Certainly, within Africa, planning for LED in South Africa is the most advanced and longest established and since the 1994 democratic transition, LED has been elevated from isolated local development intervention, mainly in its leading cities, to an obligatory mandate for all local authorities (Nel and Rogerson, 2005). The well-documented South African experience offers a wealth of potential lessons for late-developing countries variously concerning the merits of progrowth vs propoor approaches, institutional arrangements, and potential roles of local government for energizing LED (Rogerson, 2008, 2010; Rogerson and Rogerson, 2010).

Reaffirming the Case for LED

Although a range of international development organizations has recommended LED strategies for the past two decades, monitoring and evaluating the success/failure of such interventions remains problematic. Twenty years ago Clarke and

Gaile (1992: 193) pointed out efforts so far undertaken to evaluate LED policy had been "a quagmire of good intentions and bad measures". The popular trend by LED analysts and practitioners has been to search out 'good practice' case studies. Rightly, Valler (2011: 569) points to the paucity of evidence regarding what constitutes success or failure of LED in particular localities. A major challenge in evaluating LED outcomes concerns assessment of what changes, if any, might have happened in the absence of LED interventions (Valler, 2011). On the whole, systematic monitoring of different LED strategies has received little attention (Rodriguez-Pose and Palavicini-Corona, 2013: 303). As a result, whether LED strategies truly change development outcomes in either the developed or the developing world remains in question.

The most powerful evidence from an evaluation of the impact of LED strategies is the recent work in Mexico by Rodriguez-Pose and Palavicini-Corona (2013). They conduct a quantitative analysis of 898 municipalities in Mexico to assess the impact – success or failure – of LED activities. They conclude that "pursuing or even thinking about LED strategies has paid off for local authorities in Mexico in the last two decades" (Rodriguez-Pose and Palavicini-Corona, 2013: 313). A major conclusion is perhaps that the "main difference remains between thinking and/or implementing one basic element of a LED strategy or doing nothing" with those municipalities doing nothing seen as "as a result, worse off in development terms" (Rodriguez-Pose and Palavicini-Corona, 2013: 313). This Mexican evidence suggests that the greatest impact on human development emerges through the design of a development plan, the implementation of capacity building measures, and the forging of new development links both within and beyond the municipality. Overall, this seminal study offers the first definitive confirmation that greater human development in the global South is possible through local development strategies.

See also: Culture and Economic Development; Development Geographies; Development Theory in Geography; Development and Urbanization; Development and the State; Development: Social-Anthropological Aspects; Ethical Turn in International Development; Global Development: Markets and Wellbeing; Migration and Development; Peasantry in the Twenty-First Century; Planning and Development.

Bibliography

Barca, F., McCann, P., Rodriguez-Pose, A., 2012. The case for regional development intervention: place-based versus place-neutral approaches. Journal of Regional Science 52, 134–152.

Bartik, T.J., 2003. Local Economic Development Policies. Upjohn Institute, Kalamazoo. Staff Working Paper No. 03–91.

Bateman, M., Ortiz, J.P.D., Maclean, K., 2011. A Post-Washington Consensus Approach to Local Economic Development in Latin America? An Example from Medellin, Colombia. Overseas Development Institute, London (Background Note).

Christensen, J.D., van der Ree, K., 2008. Building inclusive local economies through promoting decent work. @local.glob 5, 2–5.

Clarke, S., Gaile, G., 1992. The new wave: postfederal local economic development strategies. Economic Development Quarterly 6, 187–198.

Crescenzi, R., Rodriguez-Pose, A., 2011. Reconciling Top-down and Bottom-up Development Policies. Working Papers Series in Economics and Social Sciences 2011/03. IMDEA, Madrid.

Garcilazo, E., 2011. The evolution of place-based policies and the resurgence of geography in the process of economic development. Local Economy 26, 459–466.

Helmsing, A.H.J., Egziabher, T.G., 2005. Local economic development in Africa: introducing the issues. In: Egziabher, T.G., Helmsing, A.H.J. (Eds.), Local Economic Development in Africa: Enterprises, Communities and Local Development. Shaker Publishing BV, Maastricht.

ILO, 2006. A Local Economic Development Manual for China. ILO, Geneva.

ILO, 2008. Local Economic Development Outlook. ILO, Geneva.

Le Heron, R., 2009. 'Globalisation' and 'local economic development' in a globalizing world: critical reflections on the theory-practice relation. In: Rowe, J.E. (Ed.), Theories of Local Economic Development: Linking Theory to Practice. Ashgate, Farnham.

Meyer-Stamer, J., 2008. Systemic Competitiveness and Local Economic Development. Mesopartner, Duisburg.

Nel, E., Rogerson, C.M. (Eds.), 2005. Local Economic Development in the Developing World: The Experience of Southern Africa. Transaction, New Brunswick, NJ and London.

Pike, A., Rodriguez-Pose, A., Tomaney, J., 2006. Local and Regional Development. Routledge, London.

Pike, A., Rodriguez-Pose, A., Tomaney, J., 2011. Introduction: a handbook of local and regional development. In: Pike, A., Rodriguez-Pose, A., Tomaney, J. (Eds.), Handbook of Local and Regional Development. Routledge, London, pp. 1–14.

Rodriguez-Pose, A., 2001. The Role of the ILO in Implementing Local Economic Development Strategies in a Globalized World. London School of Economics, London.

Rodriguez-Pose, A., 2008. Milestones and challenges of LED practice and academic research. @local.glob 5, 22–24.

Rodriguez-Pose, A., Palavicini-Corona, E.I., 2013. Does local economic development really work? assessing LED across Mexican municipalities. Geoforum 44, 303–315.

Rodriguez-Pose, A., Tijmstra, S., 2007. Local economic development in sub-Saharan Africa. Environment and Planning C: Government and Policy 25, 516–536.

Rodriguez-Pose, A., Tijmstra, S., 2009. On the Emergence and Significance of Local Economic Development Strategies. CAF Working Paper, No 2009/07. Financiando El Desarrollo, Caracas.

Rogerson, C.M., 2008. Consolidating local economic development in South Africa. Urban Forum 19, 307–328.

Rogerson, C.M., 2010. Local economic development in South Africa: key strategic challenges. Development Southern Africa 27, 481–495.

Rogerson, C.M., Rogerson, J.M., 2010. Local economic development in Africa: global context and research directions. Development Southern Africa 27, 465–480.

Ruecker, A., Trah, G., 2007. Local and Regional Development. German Agency for Technical Cooperation (GIZ), Eschborn.

Salazar-Xirinachs, J., 2008. Preface. @local.glob 5 (v).

Swinburn, G., Yatta, F., 2006. Furthering the Local Economic Development Agenda in Africa. Draft Discussion Paper Presented by the Municipal Development Partnership at Africities Congress, September (Nairobi).

Tomaney, A., Pike, A., Rodriguez-Pose, A., 2010. Local and regional development in times of crisis. Environment and Planning A 42, 771–779.

Tomaney, A., Pike, A., Rodriguez-Pose, A., 2011. Local and regional development: reflections and futures. In: Pike, A., Rodriguez-Pose, A., Tomaney, J. (Eds.), Handbook of Local and Regional Development. Routledge, London, pp. 618–630.

Turok, I., 2013. People-based versus place-based policies. The 2009 World Development Report. Local Economy 28, 3–8.

Valler, D., 2011. The evaluation of local and regional development policy. In: Pike, A., Rodriguez-Pose, A., Tomaney, J. (Eds.), Handbook of Local and Regional Development. Routledge, London, pp. 569–580.

Vazquez-Barquero, A., 2011. Local development – a response to the economic crisis: lessons from Latin America. In: Pike, A., Rodriguez-Pose, A., Tomaney, J. (Eds.), Handbook of Local and Regional Development. Routledge, London, pp. 506–514.

World Bank, 2003. Local Economic Development: Quick Reference. The World Bank, Washington, DC.

World Bank, 2009. Reshaping Economic Geography: World Development Report 2009. The World Bank, Washington, DC.

Location, Absolute and Relative

Harvey J Miller, The Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.D. Nystuen, volume 13, pp. 9011–9013, © 2001, Elsevier Ltd.

Abstract

Absolute and relative location are fundamental geographic concepts. Absolute location facilitates the organization of geographic data and information for analysis and mapping. Relative location relationships such as inclusion, connectivity, direction, and distance are important to many real-world physical and human phenomena. This article reviews the geographic concepts of absolute and relative locations. This includes fundamental concepts of geographic space, describing absolute location through geographic referencing, and describing relative location through spatial relations.

Introduction

Location refers to a place within a system of places or *space* where an entity can exist. *Absolute location* defines that place with respect to all possible places within the space, typically using a referencing coordinate system. Examples include latitude and longitude or map coordinates. *Relative location* defines that place with respect to its connection, direction, or distance to other places, typically those occupied by other entities of interest. Examples include statements such as 'The road intersects the park,' 'Berlin is north of Vienna,' and 'My house is 3 km from the university.'

Absolute and relative location are central concepts in geography and geographic information science. Absolute location is a fundamental way of ordering information about physical and human properties of the Earth surface, both in the traditional format of paper maps as well as digital formats such as those found in geographic information system (GIS) software. Relative location is critical for everyday use of geographic concepts and information such as navigation, wayfinding, and other human behavior (Egenhofer and Mark, 1995; Freundshuh and Egenhofer, 1997). In addition, relative location conditions the physical and human phenomena at particular geographic locations; for example, neighborhoods that are proximal tend to have similar crime rates, and wind direction affects the risk of exposure to airborne toxins from a source. Relative location is central to *spatial analysis* or the quantitative measurement and analysis of phenomena based on relative location (Miller, 2004; Miller and Wentz, 2003).

Geographic Space

The concepts of space and location are highly general and not limited to geography and geographic information science. We can define a space by designating a set of objects and relations between these objects (Gatrell, 1983). Since this definition can encompass any structured collection of objects, it is too general to describe the spaces of interest in geography (Worboys and Duckham, 2004).

Geographic space or *geospace* is a type of space that represents some portion of the Earth's surface. A geospace consists of a set of *places* where attributes or objects can exist and a *spatial relation* measuring the separation between pairs of places. The set of locations can be *strictly bounded* or *countably bounded*. A strictly bounded set of locations is isomorphic to the set of integer numbers; this treats space as *discrete* or consisting of a distinct lattice of places. A countably bounded set of locations is isomorphic to the set of real numbers; this conceptualizes space as *continuous* or consisting of dense fabric of places. In practice, however, all spaces represented in digital computers are strictly bounded since these computers have finite resolution for representing numbers, including those describing places (Beguin and Thisse, 1979; Miller and Wentz, 2003).

A configuration of places supports the concept of *absolute location*. In geography, the configuration of places typically corresponds to a *plane* in low-dimensional space; examples include networks (one-dimensional space), map space (two dimensions), or natural space (three dimensions). Places can be referenced spatially using *Cartesian coordinates* or positions along defined axes representing the spatial dimensions. Equivalently, we can reference a place with respect to the axes' origin using *polar coordinates*. When representing large portions of the Earth's surface (such as continents), it is more natural to configure these places in *spherical* rather than planar space. In this case, we reference each place with respect to the sphere's center, poles, and a reference plane at the midpoint between the poles. If we give each place in a planar or spherical configuration of places a unique reference based on its position, we can refer to its position exactly or absolutely.

The concept of *relative location* requires defining a spatial relation that describes the separation among places in the system. A common spatial relation is *distance*: this describes the length of the shortest path between any pair of locations. The distance relation typically satisfies the *metric space* properties of *nonnegativity* (the distance between any two locations is never less than zero), *identity* (the distance of a location to itself is zero), symmetry (the distance between two locations is the same in either direction), and *triangular inequality* (the indirect distance between two locations is never smaller than the direct distance). A general distance function for planar space that satisfies these requirements is

$$d_p = \left[\sum_{k \in X} \left| x_i^k - x_j^k \right|^p \right]^{\frac{1}{p}}$$

where x_i^k is a k-dimensional vector describing the coordinates of location i and p is a parameter indicating the *distance norm*. Important special cases include $p = 2$: This corresponds to the Euclidean distance norm, with the shortest path corresponding to the straight-line segment between the two locations. Another special case occurs when $p = 1$: this corresponds to the so-called Manhattan metric where all segments of any path between two locations are parallel to the axes, similar to navigation of a gridiron road network. The distance norm parameter p can be any real number equal to or greater than one; values for p less than one means that the distance relations do not satisfy the triangular inequality and the space is no longer metric. The value of p can also be estimated from observed travel distances, including distances within transportation networks (Love et al., 1988).

A Euclidean distance norm is typical in geography and geographic information science: almost every paper or digital map assumes Euclidean distance. Euclidean distance as well as any distance norm also encompasses other relative location concepts such as *set-based* (inclusion vs exclusion in a set of locations), *topology* (connectivity), and *direction*. However, we can refer to these more primitive spatial Euclidean distance norm relations without requiring that distances satisfy the Euclidean or any other spatial metric; in fact, we do not need the concept of distance at all to describe set-based, topological, and directional relations (Worboys and Duckham, 2004). We will describe these relative location concepts more fully below.

In spherical space, the shortest paths between any two locations is the smaller arc of the *great circle* passing through the locations, where a great circle is a circle on the surface of a sphere with its center coinciding with the center of the sphere (Love et al., 1988). Spherical distance functions can obey the metric space properties identified above (nonnegativity, identity, symmetry, and triangle inequality).

Describing Absolute Location: Geographic Referencing

Describing absolute location on the Earth's surface involves five issues: they are geodetic datum, map projection, coordinate system, map scale, and precision/accuracy (Campbell, 2000). The *geodetic datum* is a model of the Earth's shape. Although we often conceptualize the Earth as a sphere, an ellipsoid is a better representation as rotation causes the Earth to be slightly flattened. An even better representation is a *geoid*: this is an ellipsoid-like model of the Earth that considers its irregular gravity surface. A range of datums exists, both global and local; these differ with respect to their description of the Earth. One common datum is the World Geodetic System 1984 (WGS 84); this datum is geocentered and supports the global positioning system.

It is possible to use geodetic datum coordinates directly; an example is the latitude and longitude systems. However, we often transform these measurements to a plane for ease of use in both paper and digital form. A *map projection* transforms locations on a curved surface to a flat plane. A map projection cannot simultaneously preserve all geometric properties such as area, angles, distances, and directions; therefore, one must choose which properties to preserve, thereby distorting the others. *Conformal* projections preserve angles; a well-known example is the Mercator projection. *Equal-area* projections preserve areas; an example is the Lambert projection. *Equidistant* projections maintain distance relation relative to a designated location; an example is conic projections.

Coordinate systems reference locations in geodetic (unprojected) or planar (projected) space. Most coordinate systems use the Cartesian coordinate system, although some land surveying maps use polar coordinates. The previously mentioned latitude and longitude system is a Cartesian coordinate system referring directly to the geodetic datum. Latitude expresses location relative to the equator while longitude expresses location relative to the Prime Meridian at Greenwich, U.K. A common planar coordinate system is the Universal Transverse Mercator system based on a type of Mercator projection.

Describing absolute location with respect to the Earth also involves a choice of *map scale*. Map scale is the ratio of map distance to ground distance. For example a 1:24 000 scale means 1 inch on the map is equivalent to 24 000 inches on the ground while a 1:100 000 scale means 1 inch on the map is equivalent to 100 000 inches on the ground. The smaller the ratio, the larger the map scale and the more detail can be represented on the map. However, larger map scales involve bigger data volumes.

Positional *precision and accuracy* are two distinct but related concepts. Precision refers to the resolution of the measurement, while accuracy refers to the degree to which the information on the map matches reality. Both precision and accuracy are functions of map scale, with larger map scales having higher precision and therefore higher potential (but not necessarily actual) accuracy. Map accuracy standards for horizontal and vertical errors are consequently related to map scale. For example, the US Federal government requires for maps larger than 1:20 000 to have not more than 10% of the points tested to have horizontal error of more than 1/30 inch measured using the map publication scale (USGS, 2006).

Describing Relative Location: Spatial Relations

Relative location describes places with respect to their inclusion, connection, direction, or distance to other places, typically those occupied by other entities of interest. There are four major classes of *spatial relations*: they are set-oriented, topological, directional, and distance relations (Worboys and Duckham, 2004). These are not mutually exclusive, but the ordering above indicates increasing sophistication and therefore inclusion. For example, set orientation by itself is the most primitive conceptualization and does not include richer spatial relations related to topology, direction, and distance. However, a space with metric distances also has implicit set-oriented, topological, and directional relations.

Set-oriented relations use set theory to conceptualize relative location. Each place is an element, and a set of places represents the geometric footprint of a spatial object such as a polygon. Based on this conceptualization, set theory operators such as intersection, union, difference, and complement can be used to characterize the relationship between different spatial sets.

Topological relations describe properties that do not vary under topological or 'rubber sheet' transformation of the space

(that is, continuous transformations that do not involve tearing or punching holes in the space). Topological relations focus on connectivity, interior, exterior, and boundary relationships between places. Potential topological relationships among spatial objects such as points, lines, and polygons are surprisingly rich and can differentiate among a wide range of configurations (Egenhofer and Franzosa, 1991).

We can classify *directional* relations between spatial objects based on the frame of reference. *Cardinal* directional relationships use an external reference frame, typically the directions north, south, east, and west. An example is the statement 'the university is north of the city center.' *Object-centered* directional relations use a designated object as a reference frame; an example is 'the park is behind the school.' *Ego-centric* directions use a viewer's perspective as a reference frame; an example is 'the restaurant is on the left.'

As discussed above, *distance* relations are fundamental to most conceptualizations of geographic space although it is a sophisticated spatial relation that encompasses other, more primitive relations. Distance is a straightforward concept when applied to individual places but more complicated when applied to spatially extensive collections of places such as lines, polylines, and polygons. Except for points, a distribution of distances exists between two spatial objects; summaries of this distribution include the minimum, maximum, average, and Hausdorff (maximum of the minimum) distances. Commercial GIS software typically uses the centroid-to-centroid distance between two spatial objects to approximate the average distance, but this approximation is less accurate as the objects become larger and closer together (Okabe and Miller, 1996).

See also: Geographic Information Systems; Scale in Geography; Spatial Analysis in Geography; Spatial Association, Measures of; Spatial Data.

Bibliography

Beguin, H., Thisse, J.F., 1979. An axiomatic approach to geographical space. Geographical Analysis 11, 325–341.

Campbell, J., 2000. Map Use and Analysis, fourth ed. McGraw-Hill, Boston, MA.

Egenhofer, M.J., Franzosa, R.D., 1991. Point-set topological spatial relations. International Journal of Geographical Information Systems 5, 161–174.

Egenhofer, M.J., Mark, D., 1995. Naive geography. In: Frank, A., Kuhn, W. (Eds.), Spatial Information Theory a Theoretical Basis for GIS, Lecture Notes in Computer Science, vol. 988. Springer, Berlin, pp. 1–15.

Freundshuh, S.M., Egenhofer, M.J., 1997. Human conceptions of space: implications for GIS. Transactions in GIS 2, 361–375.

Gatrell, A.C., 1983. Distance and Space: A Geographical Perspective. Clarendon Press, Oxford.

Love, R.E., Morris, J.G., Wesolowsky, G.O., 1988. Facility Location: Models and Methods. North-Holland, New York.

Miller, H.J., 2004. Tobler's first law and spatial analysis. Annals of the Association of American Geographers 94, 284–289.

Miller, H.J., Wentz, E.A., 2003. Representation and spatial analysis in geographic information systems. Annals of the Association of American Geographers 93, 574–594.

Okabe, A., Miller, H.J., 1996. Exact computational methods for calculating distances between objects in a cartographic database. Cartography and Geographic Information Systems 23, 180–195.

United States Geological Survey (USGS), 2006. Map Accuracy Standards. Fact Sheet FS-171-99. http://egsc.usgs.gov/isb/pubs/factsheets/fs17199.html.

Worboys, M., Duckham, M., 2004. GIS: A Computing Perspective. CRC Press, Boca Raton, FL.

Location Theory

Cees Gorter[†] **and Peter Nijkamp,** VU University, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This survey of location theory offers a brief overview of both traditional locational analysis and a modern orientation toward trade and growth theory. Decisions of households are also reviewed briefly, followed by a broader exposition of modern search theory and of the potential of the modern information and communication technology. Finally, it offers a survey of classical facilities location theory and an exposition of recent advances.

The 'Where' Question: Space Is Crucial

Geographical space is the action space of individuals, households and firms. Location and allocation are two sides of the same coin. This seemingly simple observation leads to a wealth of spatial – economic phenomena that are well covered in the recent literature (see Fischer and Nijkamp, 2013). Examples are housing and labor markets, regional growth disparities, spatial innovation and entrepreneurship conditions, urban agglomerations, industrial clusters, or energy and environmental resource scarcity. All such phenomena are critically determined by the location of human and industrial activity. There is an abundance of both conceptual and applied studies that address the above-mentioned spatial mapping of activities. The present contribution aims to highlight the essential elements of standard location theory.

Location theory is concerned with the intriguing question of where economic activity can be found. It addresses the geographic dimension of economic decisions and analyzes the behavior of firms and households. Thus the locus, the geographic place, is of prime interest in location analysis.

More specifically, the main question in location theory is what are the motives to choose a particular location and what are the geographic implications? Clearly, there is not a single, conclusive answer. A student who has just left the university will usually have different aims in mind when seeking a residence than will his parents who have just retired. Despite the variety of motives and goals, many factors are common to anyone seeking a residential location (such as affordability, accessibility, quality of life, etc.).

The same applies to business firms. An ice-cream vendor will not open a shop in a dead-end street, and a diamond seller will not opt for a location in an underground station. Both may have profit motives, but their decisions regarding where to locate may be totally different. Nevertheless, the actual location patterns of business firms reveal some similarities. The location patterns of households and firms are not random worldwide, but display many regularities. For instance, all over the world we find a population concentration and business concentration in river basin areas and delta areas, a phenomenon observed by Adam Smith more than 200 years ago. Many theoretical and methodological contributions on location analysis can be found in the recent *Handbook* by Fischer and Nijkamp (2013).

Location theory does not investigate the spatial point decision of only an individual actor (household, business firm), but looks also into the regularities (patterns, geographic structures) that emerge out of the decisions of many individual actors. Locational decisions are usually influenced by so-called externalities, which mean that the decisions made previously by a group of actors affect the current decisions of new market entrants. This explains why cities emerge as a spatial conglomerate of many individual actors seeking urbanization economies, or why industrial complexes arise as a consequence of agglomeration advantages (such as a joint use of infrastructure). Thus, location theory has both micro- and macro (collective or social) aspects; locational decisions are in the end usually interactive decisions. The importance of scale advantages in locational choices and the influence of distance friction costs have stimulated much research on spatial flows (often mentioned as spatial interaction models). In this context, the Newtonian gravity principle plays often a central role. Gravity models have been a key instrument in location-allocation analysis (see for a review Nijkamp, 2013).

Clearly, once a spatial pattern of economic activity has been realized, it will influence other patterns as well. Consumers tend to be attracted by the proximity of shopping facilities, business centers tend to orient themselves toward the presence of a critical mass of clients with a high purchasing power, etc. Thus, there is interactivity in location at the micro- and the meso-/macroscales. This spatial interactivity does not lead to a random spatial configuration, but to a regular – often hierarchical – organization of spatial systems. This is clearly reflected in central place theory, and in related statistical regularities described by the Zipf's Law (or its companion, the Gibrat's Law) (see also Reggiani and Nijkamp, 2014).

The next section offers a concise presentation of some highlights from traditional location theory, followed by an exposition of some modern developments. After a discussion of firms' location decisions, households' decisions will be looked at. The impact of the modern network economy and of information and communication technology (ICT) will be dealt with as well. This article concludes with some prospective remarks.

Conventional Wisdom on Location Decisions of Firms

Location theory forms a central element in regional economics and economic geography. Location and trade were regarded by the grandfather of economics, Adam Smith, as

[†] Deceased.

two interwoven phenomena: location influences trade flows and trade affects location decisions. In the history of economics a great variety of contributions on the nature, motives, and implications of the location of production can be found. We will offer only a few examples here from this rich history.

Location and land use in agricultural production was studied extensively by Von Thünen (1842), by using a simple profit maximization scheme for spatial product choices, as a function of revenues, product types, and transport costs. He was able to demonstrate the existence of a set of concentric rings of agricultural products, based on Ricardian comparative cost theory leading to product specialization. Von Thünen's rent theory has formed the basis for urban land rent theory, which has become a central focus in urban economics.

The location of industrial firms was investigated by Weber (1909) by looking for a cost minimizing location solution of a new firm on the basis of transport costs for resource inputs (and other material inputs) and for product outputs (to be shipped to the market). This intriguing and complex question was solved by Weber (1909) by using a triangular force field of input and output points, within which the optimal location has to be found.

The spatial concentration of economic activity in a city has also been studied intensively, starting from the work of Marshall (1925), who argued that agglomeration benefits are central for city formation. He referred to externalities which caused such benefits, in particular better information and skills, trade growth, specialized equipment, and availability of skilled labor.

The shift toward multiproduct economies and the spatial interrelationships between multiple products received due attention some 50 years ago in the context of the so-called central place theory, developed by Christaller (1933) and Lösch (1954). The basic idea is that the location of economic activity is subject to agglomeration advantages. As a consequence, economic activities are not spread randomly, but manifest themselves in clusters. Moreover, in the light of product diversity, price elasticity, and the existence of daily and nondaily goods, a spatial hierarchy is a logical outcome of a series of profit-maximizing location decisions of firms. This hierarchy means that a place with a certain ranking order also includes firms producing goods with a lower order. This theory has become a landmark in location theory and has had far-reaching implications for spatial planning.

In the postwar period, a wide range of new and refreshing contributions has also been offered, such as the growth pole theory, the cumulative causation model, and the forward–backward linkage view on regional development (see also Thisse et al., 1996).

The complexity of location decisions in a continuous space or in a discrete network configuration has also prompted a rich research field in the operations research discipline, where many attempts have been made to derive mathematically the optimal site of public facilities or private goods, especially in the context of a multiproduct location decision. It should be noted that in a discrete space with a finite number of location options the number of combinatorial possibilities may be formidable, so that rather advanced mathematical software has to be developed.

In the 1990s much attention was given to a rigorous analysis of the link between trade and locations, especially in the spirit of the monopolistic competition approach advocated by Krugman (1991). In this analysis, spatial accessibility does not only favor transport and trade, but geographical clustering is also positively influenced by a strong critical mass of local demand, by increasing returns to scale in the production sector and by a good transportation system in the area. In a later publication, Krugman (1995) has emphasized that locational factors and spatial interaction effects may be understood from the perspective of two main drivers, viz., agglomeration forces and distance friction costs. Such forces are the main determinants of the complex space-economy. This issue will be discussed in more detail in Section Location and Growth.

Location and Growth

Locational decisions are usually not stand-alone choices of a firm, but are part of a broader set of firm decisions, e.g., on market areas, marketing channels, technology to be used, image, etc. Against this background transport – and thus trade – plays an important role. This fact was recognized in the well-known Heckscher–Ohlin theory on international (or interregional) trade flows, where trade intensity between two countries is dependent on the comparative advantage of the use of factor inputs. Clearly, also the degree of homogeneity of products to be traded plays a role here. Based on the assumptions of monopolistic competition, Dixit and Stiglitz (1977) have tried to make the Heckscher–Ohlin theory more realistic, inter alia by allowing for product differentiation. In this way, both inter- and intraindustry trade flows and locational patterns can be analyzed. This integration of trade theory and location theory forms one of the main characteristics of modern growth and trade theory.

In this context there is also a need for a more dynamic perspective to spatial equilibrium phenomena. Such a perspective presupposes a more integrated view on locational decisions and emerging market patterns, thus necessitating an economic growth perspective. Consequently, trade and location are endogenously determined in modern economic growth theory, as developed inter alia by Krugman (1991). This also means that factor mobility, transportation costs, and transaction costs become an integral part of modern theories on location and urban or regional growth.

In a recent study by Fujita et al. (1999), the authors offer an attempt to retrack economic geography (and regional science), while they also aim to build a 'new economic geography' based on a few rigorous economic principles. The authors signal a regrettable dividing line between mainstream economics and the economics of location. They aim to build a bridge on the basis of a few simple (certainly not universally valid) concepts, in particular the imperfect competition model marked by increasing returns to scale (originating from Dixit and Stiglitz, 1977). In an open (multiregion or multicountry) system various types of spatial agglomeration patterns may emerge, depending inter alia on transport costs, forward and backward linkages, and immobility of resources. The authors offer a new perspective in their effort to bring general economics closer to regional economics. They illustrate their arguments by referring

to core-periphery phenomena, agriculture, urban systems, city size, transportation, international trade, and industrial clustering. The book forms a clear manifestation of the rigor of solid economic analysis for the explanation of the spatial patterns and evolution of economic activity.

The two pillars of regional economics are certainly formed by agglomeration economies and generalized transportation costs. Much emphasis is made in the book on the economic analysis of urbanization and scale advantages, but less attention is given to transportation costs. In our era of ICT development where many economists advocate the 'death of distance,' it would have been preferable to pay more attention to both psychological and virtual distance costs, and their implications for the spatial organization of our world (see Section Modern Residential Location Behavior in a Network Society: The Death of Distance).

The spatial concentration of economic activity has also prompted an accelerated interest in innovation and entrepreneurship in large urban or industrial agglomerations. The recent literature in geography and regional science has focused much attention on the innovative behavior of entrepreneurs, as well as on the urban seedbed conditions for entrepreneurship and innovativeness (see Kourtit et al., 2011).

It is clear that a further integration of location theory, transportation science, trade theory, innovation theory, and industrial organization seems to be a necessary condition for a rigorous way forward.

Where Do People Wish to Live?

The locational choices of households can be looked at from many perspectives exemplified by a great variety of historical, cultural, social, economic, and geographical views. The most prominent and visible pattern in the spatial distribution of households is the clustering of people in the vicinity of firms that demand workers for their economic activities. For example, the Industrial Revolution, based on new, steam-driven technology, created mass-produced goods at particular geographical locations where the inputs for the industrial production process were found. The workers needed for this new mass production of goods were forced to settle near the factories, given the limited transport facilities at the time and their low financial means.

The ultimate example of spatial clustering of economic agents (firms and households) is observed through city formation and city growth, both in developed (industrialized) countries and in developing countries. Nevertheless, some people do not live in urbanized areas with major employment centers, but commute from their home regions to their workplace in the city. This travel distance of households from home to work forms the principal focal point in the classical urban location theories that were developed to explain the spatial structure of agglomerations (see Alonso, 1964; Muth, 1969; Mills, 1972).

The old economic wisdom on urban location decisions of households builds on the functioning of the market for land from which urban location patterns were derived through the application of standard economic laws of supply and demand. The classical, elementary model for the urban location of households considers the decision-making process of households as a place-utility maximizing process in which the travel distance from home to work (commuting distance) plays a crucial role. For reasons of expositional convenience, company businesses are supposed to be located in the center of the agglomeration (but this was relaxed in later versions of the classical models). In this case, all households would prefer to live in the center, if differences in land prices across the urban space were not yet taken into account (thereby minimizing their commuting costs). However, the excess demand of households for locations in the vicinity of the business center forces the prices for this scarcely available land up, because the households are willing to pay more for central locations. Now, a fundamental locational choice for households emerges in which they have to trade-off the conditions of living near the center (low commuting costs, but high costs for housing) with the countervailing features of living at some further distance of the center (high commuting costs, low housing costs). The difference in the costs for land (housing) will be determined – in the spirit of classical urban models – by an equilibrium state of the land market in which the household's well-being (or utility) is equalized over all possible locations within the urban space. The main economic principle of bidding for those (urban) locations that households consider to be attractive in some sense remains valid, when differences across households (e.g., size, income, preferences), the existence of urban employment subcenters, governmental regulations of land and housing markets, and other locational variations or neighborhood characteristics (public and private facilities, environmental amenities, green areas, socioeconomic composition) are taken into account (see Straszheim, 1986). This means that besides valuing shorter distances to the city (employment) center, households bid up housing prices in higher-valued neighborhoods.

Residential Relocation of Households: Search and Thou Shall Find

People do frequently move to other locations to obtain better housing during the course of their expected lifetime. There are several reasons for this change of residence and most often the motives are job-related (labor migrants) or housing-related (size and facilities of the dwelling unit). The residential move could, of course, also be due to health, family moves, change in household composition (marriage or living together, childbirth), or improved transport facilities. Short-distance moves are usually associated with housing-related reasons (such as a change in household income), whereas long-distance residential moves are likely to be triggered by job changes (see Clark and Van Lierop, 1986). In general, people consider moving when some external shock makes the current location suboptimal, that is, the household becomes dissatisfied with the dwelling itself, the housing environment (social bonds with people living in the neighborhood, availability of parks, recreational facilities, playgrounds, and schools), or the employment opportunities that are accessible from the current location through commuting. This shock may originate from the internal household circumstances (e.g., childbirth or a divorce), be related to market forces (rent, interest rate,

commuting costs) or be due to externalities such as noise caused by new neighbors or, e.g., pollution of the local environment (air, water, soil).

The propensity to move creates a search process in which households try to improve their current situation. At the heart of this search process, households face a lack of information on potential dwellings at alternative locations and are confronted with time and money costs to acquire such information. Moreover, in the search process the household has to decide on the spatial area of search. A residential move will then be made when the expected gain is higher than the moving costs (this may also include nonmonetary costs related to the psychological value attached to the local social network of contacts).

Accordingly, it is well-known that residential moves are affected predominantly by age which represents the stage in the life cycle. Furthermore, transaction (moving) costs often prevent households from moving in response to small changes in the environment. From a methodological point of view, the search behavior of households on the housing market can be investigated appropriately by employing duration modeling techniques in which variations in residential duration are used to analyze housing market dynamics (preferably in a spatial context). Recent theoretical developments along the search-theoretical angle have also stressed the simultaneity of the decision to move and to change jobs, thereby taking into account commuting costs (see Van Ommeren et al., 1996). One of the central recent issues in this respect is also whether people follow jobs (as outlined in the classical urban models) or jobs follow people (as argued in previous sections on industrial location). Empirical research has established evidence that firms tend to follow – but not to lead – workers to suburban areas (see White, 1999).

An important recent megatrend is the rise in foreign migration. We live in the 'migration age.' Modern urban agglomerations attract an unprecedented mass of migrants in search of new opportunities. The consequences of this urban migration influx are sometimes uncertain, but lead to significant changes in housing markets, labor markets, consumption patterns, and social cohesion. A systematic methodology is needed to investigate this multitude of impacts, called Migration Impact Assessment (see Nijkamp et al., 2013).

Finally, it should be emphasized that in many Western societies the possibilities to realize the demand for particular locations are limited by regulatory measures on the housing markets and restrictive spatial planning policies. In many Western countries, both housing demand and supply are regulated extensively (by means of price regulation, allocation devices by local communities, etc.).

Modern Residential Location Behavior in a Network Society: The Death of Distance

Many households now live and work in an information society in which economies are characterized by networks and increasingly build on the input of knowledge (and less on the classical production factors capital, labor and land). This new paradigm for the world economy – in which ICT plays a decisive role (see Castells, 1996) – has significant consequences for the spatial distribution of activities (at the macrolevel) and for the location decisions of agents (at the individual level). The globalizing world creates footloose agents that are not tied to specific geographical places (either market- or input-related), but at the same time localization trends can be noticed with an increasing importance of local locational factors (such as favorable living conditions, the environmental quality of life, and housing facilities).

So what are the main effects of ICT on household location and relocation decision-making? First, ICT greatly influences the search process for new dwellings (and alternative locations), since, search costs are greatly reduced by easy and full access to information (e.g., via the use of Internet). It may, however, be difficult to filter adequately information through these new media and also to obtain reliable information. Moreover, the actual moves still remain to be costly because of monetary and nonmonetary (psychological) moving costs. Second, the ICT-revolution, based on rapidly improving information transport, has initiated a process of internationalization and globalization, and the emergence of the service (network) economy. More specifically, technological progress leads to larger communication flows between firms and households in different regions and cities. In addition, the immaterial character of information removes physical barriers for households in their locational preferences. As a consequence, geographical distance becomes less and less important (because of the Internet, e-mail, and fax); this seems to be the death of distance because of the enormous decrease in communication costs. At first sight, people may choose to work at any place they prefer, thereby enabling them to live where they want without taking into consideration the transportation (commuting) cost, as was highlighted in the classical view on household location behavior. However, it also appears that local networks of firms are vital to the exchange of so-called tacit knowledge (embodied in people) by means of face-to-face contacts. For example, headquarters of major companies often prefer to be located in large agglomerations with multimodal transport facilities (especially by air). This implies, then, that geographical proximity remains an important factor in the (re)location decision of particular groups of households. Consequently, the basic trade-off for households between commuting costs and housing costs does not disappear in their locational choice process.

Location of Public Facilities

As has been argued in the previous sections on household locations, the proximity of private and public facilities will affect the locational decisions of households. From the household's perspective, the location of facilities is taken for granted. The suppliers of facilities do also have to consider the geographical location of their services (given the location of potential customers). In the real world, one observes that a wide range of services to local inhabitants is controlled by public authorities either at the local or at a regional/national level. The various types of public facilities can be classified as follows (see Johansson and Leonardi, 1986).

Travel-to facilities: consumers visit services offered such as schools, libraries, concert buildings, parks, hospitals, etc.
Insurance service facilities: consumers are protected against emergencies due to the existence of certain facilities as fire brigades, dikes, police force, army, etc.
Delivered-from facilities: consumers obtain benefits (or costs) at the place of residence as a result of certain facilities such as broadcasting, pollution from airports, etc.

In case of the first two classes of facilities, consumers travel to the service concerned and the received benefits will be distance-dependent (due to transportation costs). For the third class, the consumer obtains the services at home.

In general, facilities are considered to be public if the services supplied by the public authorities also have a 'public good' character. In short, consumption of a public facility does not hinder the consumption by other consumers of that same good (in technical terms, this means that public goods are nonexcludable, nonrejectable, and allow for joint consumption). Furthermore, the public authorities decide on the particular location of the public facilities. Therefore a major concern in this location decision is given to the distance of consumers to the service at hand. The distance dependence does, however, also imply that the public goods character is not complete in the sense that some will have easier (less expensive) access to the service than others. To sum up, the authorities first have to select their set of public facilities (that is, they have to estimate the demand for public goods) and next, they have to choose how to locate these facilities, given the spatial distribution of potential consumers.

A well-known hypothesis on the allocation of public facilities across space originates from the early work of Tiebout (1956) who postulated that people reveal their preferences for local (municipal) services by 'voting with their feet.' By this 'feet-voting' behavior, Tiebout meant that households migrate to municipalities that offer the best combinations of preferred services and associated taxes. It also gives municipalities the scope to affect the supply of facilities. As a consequence, local communities can influence the size and composition of their inhabitants (through the in- and outmigration of households) as well as their local tax base to finance these facilities. Of course, households will differ in their spatial ('voting by feet') reactions to the provision of various types of facilities. For example, young households with children will probably be attracted by places of residence offering educational services, playgrounds, and recreational amenities, whereas older households are more likely to move to places with support and care services and social-cultural amenities. One of the most important implications of the concept of 'voting by feet' is that the allocation of goods takes place more efficiently than with majority voting. This conclusion does, however, presuppose that housing search and moving are costless, ideal places can be found for each household, and that local goods do not yield benefits to people living outside the community. This set of conditions is hard to meet in practice.

To conclude, the ultimate challenge to analyze jointly household location and public facility location would require an interactive modeling framework of spatial equilibrium in which household locational choices affect governmental decisions on public facility location and vice versa.

Concluding Remarks

Location theory is essentially the heart of economic geography and regional economics, but it is also linked increasingly to industrial organization and trade theory. In particular against the background of the current globalization process, location theory is positioned in a global force field. This force field leads to many rapid responses and behavioral adjustments of business firms, so that stable and robust locations are increasingly replaced by nomadic types of business behavior. The rapidly emerging technological changes (including ICT) encourage location theory to become a battlefield of industrial forces. Clearly, this dynamic evolution has both advantages and disadvantages, and much empirical work would be needed to find out the pros and cons of this dynamic change.

In recent years, we have witnessed the emergence of 'new economic geography' as a conceptual framework for a better understanding of the complex space-economy (see Fujita et al., 1999). In this context, an attempt is made to relate spatial economic behavior of agents to scale and agglomeration economies as well as to open spatial interaction movements (trade, transport, etc.).

It also ought to be noted that from an economic perspective much attention is given to rational behavior of individual firms or households. In reality, we observe many behavioral responses which are instigated by derived factors, such as neighborhood effects, uncertainty analysis, intrafirm transaction strategies, outsourcing mechanisms, etc. This suggests that also the behavioral basis of location theory certainly needs to be reinforced.

Finally, the space-economy is by no means static. It offers a dynamic picture, with large changes in the spatial pattern of human activity. An important modern trend is the spatial movement of people in the current urbanization trend. More than one half of the world population is living in cities, and there is for the time being no reason to expect a declining trend in urbanization. The creativeness climate in modern cities (see Florida, 2001), the knowledge base offered by large agglomerations (see van Geenhuizen and Nijkamp, 2012), and the concentrated variety in choice options in urban areas (see Glaeser et al., 2001) shape a geographical arena where people wish to live. Consequently, locational behavior and urbanization will increasingly have to be seen as two sides of the same coin.

See also: Economic Geography; Housing Economics; Industrial Ecology; Industrial Geography; Labor Market Institutions and Labor Market Outcomes in Transition Economies; Residential Segregation: Geographic Aspects; Spatial Optimization Models; Spatial Pattern, Analysis of.

Bibliography

Alonso, W., 1964. Location and Land Use. Harvard University Press, Cambridge, MA.
Castells, M., 1996. The Rise of the Network Society. Basil Blackwell, Malden, UK.
Christaller, W., 1933. Die Zentralen Orte in Süddeutschland (Trans. Baskin, C.W., 1966. Central Places in Southern Germany. Prentice-Hall, Englewood Cliffs, NJ.). Fischer, Jena, Germany.

Clark, W.A.V., Van Lierop, W., 1986. Residential mobility and household location modelling. In: Nijkamp, P. (Ed.), In Handbook of Regional and Urban Economics, Regional Economics, vol. I. North-Holland, Amsterdam, pp. 97–132.
Dixit, A.K., Stiglitz, J.E., 1977. Monopolistic competition and optimum product diversity. American Economic Review 67, 297–308.
Fischer, M.M., Nijkamp, P. (Eds.), 2013. Handbook of Regional Science. Springer-Verlag, Berlin.
Florida, R., 2001. The Rise of the Creative Class. Basic Books, New York.
Fujita, M., Krugman, P., Venables, A., 1999. The Spatial Economy: Cities, Regions, and International Trade. MIT Press, Cambridge, MA.
Glaeser, E., Kolko, J., Saiz, A., 2001. Consumer city. Journal of Economic Geography 1 (1), 7–50.
van Geenhuizen, M.S., Nijkamp, P. (Eds.), 2012. Creative Knowledge Cities. Edward Elgar, Cheltenham.
Johansson, B., Leonardi, G., 1986. Public facility location: a multiregional and multi-authority decision context. In: Nijkamp, P. (Ed.), In Handbook of Regional and Urban Economics, Regional Economics, vol. I. North-Holland, Amsterdam, pp. 133–170.
Kourtit, K., Nijkamp, P., Stough, R. (Eds.), 2011. Drivers of Innovation, Entrepreneurship and Regional Dynamics. Springer Verlag, Berlin.
Krugman, P., 1991. Increasing returns and economic geography. Journal of Political Economy 99, 483–499.
Krugman, P., 1995. Development, Geography, and Economic Theory. MIT Press, Cambridge.
Lösch, A., 1954. The Economics of Location. Yale University Press, New Haven.
Marshall, A., 1925. Principles of Economics. MacMillan, London.
Mills, E., 1972. Studies in the Structure of the Urban Economy. Johns Hopkins University Press, Baltimore, MD.
Muth, R., 1969. Cities and Housing. University of Chicago Press, Chicago.
Nijkamp, P., 2013. The universal law of gravitation and the death of distance. Romanian Journal of Regional Science 7 (2), 1–10.
Nijkamp, P., Pool, J., Sahin, M. (Eds.), 2012. Migration Impact Assessment: New Horizons. Edward Elgar Publishing, Cheltenham.
Paelinck, J.H.P., Nijkamp, P., 1985. Operational Theory and Method in Regional Economics. Avebury, Aldershot, UK.
Reggiani, A., Nijkamp, P., 2014. Did Zipf anticipate socio-economic spatial networks? Environment & Planning B.
Smith, A., 1776. An Inquiry into the Nature and Causes of the Wealth of Nations. Liberty Classics, Indianapolis, IN.
Straszheim, M., 1986. Urban residential location. In: Mills, E.S. (Ed.), In Handbook of Regional and Urban Economics, Urban Economics, vol. II. North-Holland, Amsterdam, pp. 717–757.
Thisse, J.P., Button, K., Nijkamp, P. (Eds.), 1996. Location Theory, vols. I and II. Edward Elgar, Cheltenham.
Tiebout, C., 1956. A pure theory of local expenditures. Journal of Political Economy 64, 416–424.
Van Ommeren, J., Rietveld, P., Nijkamp, P., 1996. Residence and workplace relocation: a bivariate duration model approach. Geographical Analysis 28, 315–329.
Von Thünen, J.H., 1842. Der isolierte Staat in Beziehung anf Landwirtschaft and Nationalökonomie (Trans. Wartenberg, C.M., 1966. Von Thünen's Isolated State. Pergamon, Oxford, UK.). Léopold, Rostock, Germany.
Weber, A., 1909. Über den Standort der Industrien. Tübingen, Germany (Trans. Friedrich, C.F., 1929. Alfred Weber's Theory of Location of Industries. University of Chicago Press, Chicago.).
White, M.J., 1999. Urban areas with decentralized employment: theory and empirical work. In: Cheshire, P., Mills, E.S. (Eds.), In Handbook of Regional and Urban Economics, Applied Urban Economics, vol. III, pp. 1375–1412.

Logic and Linguistics

Thomas Ede Zimmermann, Goethe University Frankfurt, Frankfurt, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Logical concepts and techniques are widely used (1) as tools for linguistic analysis and (2) in the study of the structure of grammatical theories. As to (1), the classical field of applying logic to linguistics is the analysis of meaning. The key concepts of linguistic semantics, including the very notion of meaning, are frequently defined in terms of, or in analogy with, concepts originating in formal logic. Moreover, algorithms translating natural language into logical symbolism are used to give precise accounts of, and allow comparison with, the semantic structure of natural language. Concerning (2), a central idea of generative grammar has been to specify logical characteristics of linguistic theory and study their repercussions on the complexity of the object under investigation; i.e., language.

Although the founding fathers of modern logic showed some interest in the analysis of meaning in natural language, solid links between the two disciplines of logic and linguistics had not been established before the middle of the twentieth century. In the late 1950s, logical techniques were first applied to the analysis of grammars as theories of linguistic competence. These applications rapidly developed into the field of mathematical linguistics and subsequently became a central part of theoretical computer science. They will be briefly addressed in the second part of this article. The tightest connections between logic and linguistics, however, grew out of the (re)discovery of logical analysis as a tool of natural language semantics, starting with the advent of Montague Grammar in the early 1970s. The role of logic in the analysis of meaning forms the subject of the first part of this article.

Logic and Natural Language Semantics

The influence of logic on natural language semantics starts with the key notion of the field: meaning. Following the tradition of analytic philosophy (cf. Ammerman, 1965), many linguists take a 'realistic' approach to meaning that makes abundant use of logical apparatus. The basic ideas of the logical analysis of meaning will be explained in The Logical Approach to Meaning section. The Logic and Language section is concerned with some more specific relations between languages of formal logic and natural languages such as English (see Barwise and Etchemendy, 1990; Gamut, 1991 for introductions to modern logic from a semantic perspective).

The Logical Approach to Meaning

A primary goal of semantic theory is to describe the *sense relations* that competent speakers perceive between expressions of a language (cf. Lyons, 1968: Chapter 10; Larson and Segal, 1995: Chapter 1). Thus, for example, knowledge of the following facts about English phrases is part of native speakers' semantic competence: (1) *brown paper bag* is more specific than, or *hyponymous* to, *paper bag*; (2) *blonde eye-doctor* and *fair-haired ophthalmologist* are *synonymous* in that they necessarily apply to the same persons; and (3) *Mary owns a dog* and *Nobody owns a dog* are *incompatible* with each other. Many sense relations can be expressed in terms of *entailments* between sentences: (1) holds because *This is a brown paper bag* entails *This is a paper bag*; (2) holds because *This is a blonde eye-doctor* and *This is a fair-haired ophthalmologist* entail each other; and (3) holds because *Nobody owns a dog* entails *It is not the case that Mary owns a dog*. A theory that correctly predicts the entailment relations holding between the sentences of a particular language may therefore be expected to go a long way toward accurately describing its sense relations in general.

In modern logic, the notion of entailment has been successfully formalized by means of *truth conditions* (cf. Zimmermann, 2011): statement *A* entails statement *B* if, and only if, any way of making *A* true is also a way of making *B* true. Ways of making statements true are called *models*, which is why the enterprise of describing meaning in terms of truth conditions is known as *model-theoretic semantics*. More specifically, a model consists of a set of objects, the model's *Universe*, plus an *Interpretation* assigning to every nonlogical word a suitable denotation. For instance, the Universe of a specific model might be the set of all numbers, past US presidents, plants, or whatever – any set would do, as long as it is not empty. Similarly, the Interpretation could assign any individual to the name *Mary* – the number 20, Jimmy Carter, my little green cactus – as long as that individual is in the model's Universe; and it could assign any set (even the empty set) to the noun *ophthalmologist*, as long as the members of that set are all in the model's Universe, and as long as it is the same set as that which the model assigns to *eye-doctor*. The idea is that any model in which the individual assigned to *Mary* happens to be a member of the set assigned to *ophthalmologist* makes the sentence *Mary is an ophthalmologist* true; and any model in which the set assigned to *ophthalmologist* happens to be empty, is one that makes the sentence *Nobody is an ophthalmologist* true.

In general, then, a model assigns to each expression of the language a denotation, thereby deciding the truth and falsity of each (declarative) sentence. In fact, the denotation of a sentence is identified with its *truth value*: the number 1 if the sentence is true, and the number 0 otherwise. (Other choices of truth values are possible but less convenient.) Using this

numerical convention and writing $[\![X]\!]^M$ for the denotation of an expression X according to a model M, a statement A thus turns out to entail B just in case, for any model M, $[\![A]\!]^M \leq [\![B]\!]^M$: whenever A's truth value $[\![A]\!]^M$ according to any model M is 1, then so is B's truth value $[\![B]\!]^M$ according to the same model M.

Traditionally, the statements made true by models are logical formulas. In linguistic semantics, there are two ways of adapting models to the interpretation of natural language: either natural language expressions are translated into logical notation (*indirect interpretation*) or models directly assign denotations to them (*direct interpretation*). On either approach, complex expressions, i.e., anything longer than one word, are usually interpreted following the *Principle of Compositionality* (Janssen, 1997): the translation or the denotation of a complex expression consisting of more than one word [like a sentence] is obtained by combining the translations or, respectively, the denotations of its immediate parts [like subject and predicate]. As a consequence, the denotations models assign to (the translations of) natural language expressions tend to be rather abstract. If, for instance, the denotation of the noun phrase *my car*, i.e., a concrete object, is to be obtained by combining the denotations of the noun *car*, which is a set of concrete objects, and that of the possessive *my*, the latter would have to be some 'selection mechanism' that picks out the unique object belonging to the speaker when combined with a set of objects. More abstract denotations turn out to be needed in a compositional treatment of determiners (like *every* and *most*), adverbs (like *quickly*) and prepositions (like *between*). This is one of the reasons why the most common systems of logic, propositional calculus and first-order logic, do not suffice as the basis of indirect interpretation.

The Principle of Compositionality has an important consequence, known as the *Substitution Principle*: if the denotation of a complex expression is completely determined by the denotations of its parts, then one may replace any part by something with the same denotation without thereby changing the denotation of the whole. In fact, the two principles turn out to be largely equivalent (Hodges, 2001). A case in point is the sentence (*a*) *Mary is an ophthalmologist*. If Mary happens to be the speaker's neighbor, then the name *Mary* and the noun phrase *my neighbor* have the same denotation. Hence, if (*a*) is true and thus denotes the truth value 1, then, by the Substitution Principle, (*b*) *My neighbor is an ophthalmologist* has the same denotation as (*a*), i.e., it is also true. While the Substitution Principle appears to make the right prediction in this case, this is not always so. Together with the assumption that sentences denote truth values, the Substitution Principle immediately leads to what, in the philosophy of language, is known as *Frege's Problem* (Frege, 1892): replacing any sentence by another one with the same truth value ought to preserve denotation, which it does not always do. For instance, the denotation, i.e., the truth value, of (*b*) *Nobody doubts that Mary is an ophthalmologist* ought to be preserved when (*a*) is replaced by a sentence with the same truth value. If Mary is an ophthalmologist, the truth value of (*a*) is 1 and hence the same as that of (*c*) *Paris is the capital of France*. However, (*b*) might well be false while (*b′*) *Nobody doubts that Paris is the capital of France is* true. Clausal embedding under so-called *attitude verbs* like *doubt*, *know*, or *tell*, thus turns out to be an *intensional construction*; i.e., one that defies the Substitution Principle.

Various solutions to Frege's Problem have been proposed. Practically all of them employ the idea that, at least in certain constructions, the contribution that a sentence makes to the denotations of larger expressions is not its truth value but something else. According to a widely accepted view (originating with Frege), it is its *informational content*: attitude verbs like *doubt* denote a relation between the individual denoted by the subject and the information conveyed by the embedded clause. A common way of modeling informational content is by means of *possible worlds* – whence the term *possible worlds semantics* (see Barwise, 1989; Devlin, 1991 for alternative approaches). The basic idea is akin to that of measuring the expectance value of an event by the proportion between favorable and unfavorable cases. The cases become possible worlds, the event is the information (as expressed by a sentence) being correct and instead of the quantitative relation between the sets of cases expressed by a ratio, it is the qualitative difference that is emphasized; i.e., which worlds are favorable and which ones are not. Following this line of thought, the informational content of a sentence can be represented by a function (in the mathematical sense) mapping worlds to truth values. For example, the informational content of (*a*) above becomes a function assigning to any possible world the truth value 1 if that world is such that, in it, Mary is an ophthalmologist; all other worlds will be assigned the truth value 0. Each possible world is thought of as representing a specific way the world could be, or might have been – down to the last detail. So the informational content of a sentence will typically assign 1 to an infinity of possible worlds differing in all sorts of aspects that are irrelevant to the truth of that sentence; and it will typically assign 0 to innumerably many worlds differing in matters that do not affect the falsity of the sentence.

In order to give a full solution to Frege's Puzzle within possible worlds semantics, the informational content of each sentence must be systematically determined, which is again done compositionally, replacing denotations by *intensions*. The intensions of sentences are their informational contents, but the notion of intension is more general in that it applies to all kinds of expressions (Carnap, 1947; Montague, 1970). The intension of any expression is its denotation (or *extension*) as it varies from world to world; i.e., a function (in the mathematical sense) assigning to each possible world the extension of the expression in that world. As a case in point, one may compare the descriptions (*d*) *the present capital of France* and (*e*) *the largest city in France*: in some worlds, including our actual reality, their extensions coincide, but given a different course of history, as represented by a possible world w, Reims, though smaller than Paris, would be the capital of France. So the intensions of (*d*) and (*e*) cannot be the same, because they disagree on w (and on many other worlds). In general, then, expressions with the same denotations may still differ in their intensions and hence, given compositionality, in the contribution they make to the informational contents of the sentences in which they occur.

The method of extension and intension complements, rather than replaces, the model-theoretic approach to meaning. Instead of directly assigning arbitrary denotations to expressions, the models of possible worlds semantics couple

each expression with an arbitrary intension, which then, being a function, uniquely determines an extension relative to a given world. Moreover, just like models come with their own Universes, in possible worlds semantics each model is also equipped with its own Logical Space of worlds. This construction originates from *modal logic* (see below) and offers an alternative to the model-theoretic reconstruction of entailment: a sentence *A strictly implies* a sentence *B* if *B* is true in all worlds in which *A* is true; i.e., if the intension of *B* yields the truth value 1 for any possible world for which the intension of *A* yields 1.

Logic and Language

The basic tools of logic-based linguistic semantics – models and possible worlds – had originally been developed for artificial languages designed for specific purposes, like the analysis of mathematical proofs. Logical formulas can usually be paraphrased in colloquial language; the relation between formula and paraphrase is taught in introductory logic classes. Conversely, many constructions in natural language (passive, relative clauses, …) have long been known to be systematically expressible in languages of logic (cf. Quine, 1960); *formalization*, i.e., the art of translating ordinary sentences into logical notation, is also something to be picked up in an introduction to logic. As a side effect of the logical approach to natural language semantics, these relations between natural language grammar and logical notation can be explored in a more rigorous way. We will briefly go through some pertinent examples; more detailed case studies can be looked up in any textbook or handbook on formal semantics (Heim and Kratzer, 1998; Zimmermann and Sternefeld, 2013; von Stechow and Wunderlich, 1991; van Benthem and ter Meulen, 1997; Maienborn et al., 2011/2012).

The most obvious connections between logic and language concern the meaning of *logical words*; i.e., those lexical items that translate the basic inventory of logical formalism: the ('Boolean') connectives '¬'; '∧'; '∨'; … translated as *[it is] not [the case that]*; *and*; *or*; … and the *quantifiers* '∀' and '∃' corresponding to *everything* and *something*. The logicians' versions of the former are combinations of truth values. For example, just like a sentence of the form '*A and B*' is true just in case both *A* and *B* are, '∧' combines 1 and 1 into 1 and the other pairs of truth values into 0. Similarly, '¬' (glossed as *not*) turns truth (1) into falsity and vice versa; '∨' (∼*or*) combines 0 and 0 into 0 and everything else into 1, etc. These combinations of truth values can thus be given by means of the *truth tables* in Table 1.

Even though truth tables work fine as first approximations to a semantic analysis of the words *not*, *and*, and *or* as they are used in mathematical jargon, there is more to the meaning of these words. To begin with, *and* and *or* do not always stand between sentences but may be used to connect almost all kinds of expressions: *every book and a pen*; *see and/or hear*; *true or false*; etc. However, in many cases, these nonsentential uses abbreviate coordinated sentences: *John owns every book and a pen* is short for *John owns every book and John owns a pen*, etc. In fact, nonsentential *and* and *or* can frequently be systematically reduced to the relevant truth tables by a *type shift*, a semantic rule that systematically generalizes combinations of truth values to combinations of other denotations (see Partee and Rooth, 1983; Keenan and Faltz, 1985; Hendriks, 1993). However, even this generalization does not cover all nonsentential uses of these words. In particular, there is no obvious reduction of the group reading of *and*, as in *John and Mary are a married couple*, to the truth table of Boolean conjunction; for a less obvious reduction see, however, Winter (2001). Neither does there seem to be a straightforward reduction of alternative *or*, as in questions like *Shall we leave or do you want us to stay?*, to Boolean disjunction. Certain other semantic aspects that the truth table analysis seems to miss may be captured by pragmatic principles governing efficient use, as opposed to literal meaning (Grice, 1989; Levinson, 1983). As a case in point, the difference between *John lit a cigarette and Mary left the room* and *Mary left the room and John lit a cigarette* may be explained by a general principle of discourse organization, according to which the order of sentences should match the order of the events they describe. More involved pragmatic reasonings have been given to account for the exclusive sense of *or*, according to which *John missed the train, or he decided to sleep late* implies that John did not both miss the train and decide to sleep late.

In addition to the Boolean connectives, standard formal languages of logic contain both a *universal* and an *existential quantifier*, usually translated as *everything* and *something* and expressing properties of predicates: being universal, i.e., true of everything, and being nonempty, i.e., applying to something. Natural language statements are only expressible in standard (first order) logic if they can be rephrased by combinations of Boolean connectives and the two quantifiers. A case in point is the sentence (*N*) *Every boy loves a girl* that can be paraphrased by (*L*) *For everything it holds that either it is not a boy, or for something it holds that it is a girl and the former loves the latter* – which translates into predicate logic as (*F*) $(\forall x)\ [\neg\ B(x) \vee (\exists y)\ [G(y) \wedge L(x,y)]]$. As the logician's paraphrase (*L*) indicates, the internal structure of quantified statements in natural language is quite different from of the corresponding logical formulas. In particular, it usually involves *determiners* like *every* and *a* instead of the natural language quantifiers *everything* and *something*. The difference is that a determiner relates *two* sets, whereas a quantifier makes a statement about *one* set. Thus, in (*N*), *every* relates the set of boys with the set of girl lovers, whereas the quantifier *everything* in (*L*) as well as the universal quantifier '∀' in (*F*) attribute universality to the set of individuals that love girls if they are boys; i.e., the set of those that are either not boys or else love a girl.

Another means of expressing quantification in natural language is by way of certain *adverbs*. In particular, in so-called *donkey sentences* like *If a farmer owns a donkey, he never beats it*, the adverb *never* (on one reading) expresses that the set of donkey-owning farmers that beat their donkeys is empty.

Table 1 Truth tables for Boolean connectives

A	B	¬A *[negation; gloss: not* A*]*	A∧B *[conjunction;* A *and* B*]*	A∨B *[disjunction;* A *or* B*]*
1	1	0	1	1
1	0	0	0	1
0	1	1	0	1
0	0	1	0	0

The very form of this *adverbial quantification* is rather remote from the usual logical formalization; it is not even clear whether in this case, as is usually assumed (see above), the indefinites *a farmer* and *a donkey* correspond to existential quantifiers. In natural language semantics special logical systems have been developed to give a more adequate logical account of indefinites in donkey sentences and related constructions (see Dekker, 2012).

Apart from these structural differences in the formulation of quantified statements, there are also differences in expressive power. In particular, not all natural language determiners can be paraphrased by combining Boolean connectives and logical quantifiers. For instance, it can be shown that a statement like *Most boys are asleep* cannot be expressed in first-order logic (Barwise and Cooper, 1981). Adequate formalizations of such sentences, and hence of natural language in general, again call for more powerful and complex languages of logic (*see* Quantifiers for more about quantification in natural language and its relation to logic).

Logicians have devised various extensions of classical propositional and predicate logic as tools of formalizing reasoning involving specific locutions (cf. Haack, 1974; Gabbay and Guenthner, 2001, vol. 2). For instance, in addition to Boolean connectives, languages of (propositional) *modal logic* (Chellas, 1993; Blackburn et al. 2001) contain a *necessity* operator, usually written as '$\Box$'; a formula '$\Box\varphi$' [read: it is necessarily so that φ] is true *in a world w* of a given model if the formula φ itself is true in all worlds that are possible for (or, technically: *accessible* from) *w*. Combining negation and necessity, it is also possible to express a corresponding notion of possibility as '$\neg\Box\neg$.' Models of modal logic thus contain domains of possible worlds organized by an accessibility relation. Depending on the precise structure of the latter, modal logic can be used to formalize various uses of modal verbs as in *Someone must have been here* (*epistemic* must) or *You may now leave the room* (*deontic* may). However, detailed studies of modality have revealed that these formalizations are at best approximations of the meanings of modal verbs in natural language (cf. Kratzer, 1991). A similar relation is that between *temporal logic* (Venema, 2001) and the tenses and other means of reference to time in natural language (*see* Tense, Aspect, Mood, and Evidentiality, Linguistics of).

Various phenomena have led semanticists to explore alternatives to classical logic with its model-theoretic interpretation assigning one of two truth values to each formula or sentence (see Beaver, 1997). A sentence like (*1*) *John knows that there will be a party tonight* seems to imply (*P*) *There will be a party tonight*; on the other hand, so does its negation (*2*) *John does not know that there will be a party tonight*. But if (*1*) and (*2*) both implied (*P*) in the sense of classical model theory, then (*P*) would be true throughout the models in which (*1*) is true – as well as in those models in which (*2*) is true; i.e., the ones in which (*1*) is false. But then (*3*) would have to be a *tautology*, i.e., a statement that is true in all models; for any model would have to make (*1*) true or false. Giving up this last assumption, one enters the realm of (certain) *nonclassical logics*, the simplest among which are *three-valued logics* with an additional truth value *U* (read: *undefined*) which models assign to sentences like (*1*) if their *presupposition (P)* is not true in them (*see* Linguistic Presupposition for more on this).

The Logic of Grammar

As theories of language structure, grammars themselves have been the objects of logical analysis. The earliest such investigations mark the beginning of generative grammar and were directed to shedding a light on the complexity of human language (see Syntactic Complexity section). subsequent approaches focused on the role of grammatical rules in the activity of parsing, which bears some resemblance to the process of deduction in logic; and more recently, grammatical reasoning has been seen as a nonmonotonic process, thus reaching beyond the limits of classical logic (see Further Issues section).

Syntactic Complexity

A syntactic theory of a particular language may be seen as a body of theoretical statements describing the structure of grammatically correct phrases. Formal reconstruction reveals that these theories come in different formats and complexities (cf. Pullum and Scholz, 2001 for more on this). For instance, a *context-free* description of a language only contains grammatical rules of the form: 'If a phrase of category *B* is immediately followed by a phrase of category *C*, the result will be a phrase of category *A*.' The standard notation for a rule of this form is: '$A \rightarrow B + C$.' Thus the following rule of English grammar is context-free: *if a preposition* (like *about*) *is followed by a noun phrase* (like *corrupt politicians*), *they form a prepositional phrase* (*about corrupt politicians*). In standard notation, this rule reads: '$PP \rightarrow Prep + NP$.' On the other hand, the following would-be rule is *not* context-free: *after a ditransitive verb* (like *give*), *two noun phrases* (like *every kid* and *a penny*) *form an object-phrase* (*a kid a penny*). Rather, the rule comes under the more complex scheme: 'If a phrase of category *B* is immediately followed by a phrase of category *C*, the result will be a phrase of category *A*, provided that it follows a phrase of category *D*'; one just has to put $A =$ *object-phrase*, $B = C =$ *noun phrase*, and $D =$ *ditransitive verb*. Rules of this form are called *context-sensitive*. Using standard notation, they are written as: '$A \rightarrow B + C/D-$.'

The distinction between context-free and context-sensitive rules – and similar distinctions between other kinds of grammatical rules – can be used to define an abstract complexity ranking between languages, the *Chomsky Hierarchy* (after Chomsky, 1957, 1959). A language is context-free (or, as the case may be, context-sensitive) if it can in principle be described by a body of context-free (or, respectively, context-sensitive) rules. Languages in the sense of this definition are arbitrary sets of sentences, which themselves are arbitrary strings of words. It turns out that, in this abstract mathematical setting (known as *mathematical linguistics* or *formal language theory*), though all context-free languages are trivially context-sensitive, the converse is not the case: one can construct languages that are context-sensitive without being describable by any body of context-free rules whatsoever. A famous example is the 'language' $a^n b^n c^n$ whose 'sentences' consist of arbitrarily many occurrences of the 'word' *a* followed as many *b*s, which are again followed by the same number of *c*s. (A proof can be found in Kracht, 2003: Section 1.6.)

In applying these abstract concepts, a language like English is identified with the set of its grammatical sentences, thereby abstracting from meaning or grammatical structure (see Kracht, 2011 for a wider perspective). Given this identification, it is natural to ask for the exact location of human languages in the Chomsky Hierarchy. It may seem that an answer to this question requires complete knowledge of the grammatical sentences of any given language – something that current linguistic research has not attained. However, at least some of these complexity issues can be settled on the basis of partial information about the languages under investigation. In particular, it has been shown that, given the intricate structure of its verb complexes, the Swiss German dialect spoken in Zurich cannot be captured by context-free means (Shieber, 1985); a similar result concerning word formation has been obtained for the African language Bambara spoken in Mali (Culy, 1985), thus proving that natural languages in general are not context-free. See also the discussion in Pullum and Rawlins (2007) on the English *X or no X* construction (as in *Rain or no rain, we are going to have our carnival*) as a possible source of noncontext-freeness.

Further Issues

The fact that ordinary syntactic notation looks rather different from logical symbolism should not distract from the close connections between the two. For instance, as indicated above, a context-free rule only abbreviates a lengthy statement about the internal structure of phrases; and this statement can be easily (though somewhat clumsily) expressed in standard logical symbolism (i.e., first-order predicate logic). This fact becomes particularly important when it comes to the task of *parsing* phrases of a given language according to a given grammar. As a case in point, parsing the prepositional phrase *about Mary* on the basis of the above context-free rule boils down to making a connection between the following three pieces of information: (*I*) *about* is a preposition; (*II*) *Mary* is a noun phrase; and (*III*) if x is a preposition and y is a noun phrase, then $x\ y$ (i.e., x followed by y) is a prepositional phrase. (*I*) and (*II*) are basic facts about English words to be recorded in a *lexicon* (which may be thought of as part of a syntactic description); (*III*) is just a paraphrase of the above context-free rule: '$PP \rightarrow Prep + NP$.' (*I*)–(*III*) obviously entail that *about Mary* is a prepositional phrase, and this is precisely the kind of entailment that can be formalized by means of elementary logic. Hence it should not come as a surprise that there are close relations between parsing strategies and procedures for verifying logical entailments. This connection has been studied in computational linguistics and in categorial grammar under the heading *Parsing as Deduction* (see Johnson, 1991).

The role of logical deduction in grammatical theory is more complex than can be indicated in an article of this length. In particular, certain phenomena like the markedness of grammatical features or the interaction of phonological constraints, seem to call for nonclassical logics that violate the *Law of Monotonicity*, according to which a conclusion once drawn remains correct when more premises are added (cf. Besnard et al. 2003; Prince and Smolensky, 2004).

See also: Generative Grammar; Lexical Semantics; Logics for Knowledge Representation; Meaning and Rule Following; Montague Grammar; Semantics.

Bibliography

Ammerman, R.R. (Ed.), 1965. Classics of Analytic Philosophy. McGraw-Hill, New York.
Barwise, J., 1989. The Situation in Logic. CSLI, Stanford, CA.
Barwise, J., Cooper, R., 1981. Generalized quantifiers and natural language. Linguistics and Philosophy 4, 159–219.
Barwise, J., Etchemendy, J., 1990. The Language of First-Order Logic. CSLI, Stanford, CA.
Beaver, D., 1997. Presupposition. In: van Benthem, J., ter Meulen, A. (Eds.), The Handbook of Logic and Language. Elsevier, Amsterdam, pp. 939–1008.
Besnard, Ph., Fanselow, G., Schaub, T., 2003. Optimality Theory as a family of cumulative logics. Journal of Logic, Language and Information 12, 153–182.
Blackburn, P., de Rijke, M., Venema, Y., 2001. Modal Logic. Cambridge University Press, Cambridge, UK.
Carnap, R., 1947. Meaning and Necessity. University of Chicago Press, Chicago.
Chellas, B., 1993. Modal Logic. Cambridge University Press, Cambridge, UK.
Chomsky, N., 1957. Syntactic Structures. Mouton, The Hague, The Netherlands.
Chomsky, N., 1959. On certain formal properties of grammars. Information and Control 2, 137–167.
Culy, C., 1985. The complexity of the vocabulary of Bambara. Linguistics and Philosophy 8, 345–351.
Dekker, P., 2012. Dynamic Semantics. Springer, Dordrecht.
Devlin, K., 1991. Logic and Information. Cambridge University Press, Cambridge, UK.
Frege, G., 1892. Über Sinn und Bedeutung. Zeitschrift für Philosophie und Philosophische Kritik 100, 25–50.
Gabbay, D., Guenthner, F. (Eds.), Handbook of Philosophical Logic, second ed., vol. 2. Springer, Dordrecht, The Netherlands.
Gamut, L.T.F., 1991. Logic, Language, and Meaning, 2 vols. University of Chicago Press, Chicago.
Grice, P., 1989. Studies in the Way of Words. Harvard University Press, Cambridge, MA.
Haack, S., 1974. Deviant Logic. Cambridge University Press, Cambridge, UK.
Heim, I., Kratzer, A., 1998. Semantics in Generative Grammar. Blackwell, Oxford, UK.
Hendriks, H., 1993. Studied Flexibility: Categories and Types in Syntax and Semantics. PhD thesis, University of Amsterdam.
Hodges, W., 2001. Formal features of compositionality. Journal of Logic, Language and Information 10, 7–28.
Janssen, T.M.V., 1997. Compositionality (with an appendix by B. H. Partee). In: van Benthem, J., ter Meulen, A. (Eds.), Handbook of Logic and Language. Elsevier, Amsterdam, NY.
Johnson, M., 1991. Deductive parsing: the use of knowledge of language. In: Berwick, R.C., et al. (Eds.), Principle-Based Parsing: Computation and Psycholinguistics. Kluwer, Dordrecht, The Netherlands, pp. 39–64.
Keenan, E., Faltz, L., 1985. Boolean Semantics for Natural Language. Reidel, Dordrecht, The Netherlands.
Kracht, M., 2003. The Mathematics of Language. de Gruyter, Berlin.
Kracht, M., 2011. Interpreted Languages and Compositionality. Springer, Dordrecht.
Kratzer, A., 1991. Modality. In: von Stechow, A., Wunderlich, D. (Eds.), pp. 639–650.
Larson, R., Segal, G., 1995. Knowledge of Meaning. MIT Press, Cambridge, MA.
Levinson, S.C., 1983. Pragmatics. Cambridge University Press, Cambridge, UK.
Lyons, J., 1968. Introduction to Theoretical Linguistics. Cambridge University Press, Cambridge, UK.
Maienborn, C., et al. (Eds.), 2011/2012. Semantics. An International Handbook of Natural Language Meaning. de Gruyter, Berlin.
Montague, R., 1970. Universal grammar. Theoria 36, 373–398.
Partee, B., Rooth, M., 1983. Generalized conjunction and type ambiguity. In: Bäuerle, R., et al (Eds.), Meaning, Use, and Interpretation of Language. Walter de Gruyter, Berlin, pp. 361–383.
Prince, A., Smolensky, P., 2004. Optimality Theory. Blackwell, Oxford.
Pullum, G.K., Rawlins, K., 2007. Argument or no argument? Linguistics and Philosophy 30, 277–287.
Pullum, G.K., Scholz, B.C., 2001. On the distinction between model-theoretic and generative–enumerative syntactic frameworks. In: de Groote, P., Morrill, G., Retoré, C. (Eds.), Logical Aspects of Computational Linguistics: 4th International Conference (Lecture Notes in Artificial Intelligence, 2099). Springer, Berlin, pp. 17–43.
Quine, W.V., 1960. Word and Object. MIT Press, Cambridge, MA.
Shieber, S.M., 1985. Evidence against the context-freeness of natural language. Linguistics and Philosophy 8, 333–343.

Stechow, A.von, Wunderlich, D. (Eds.), 1991. Semantics. An International Handbook of Contemporary Research. de Gruyter, Berlin.

van Benthem, J., ter Meulen, A. (Eds.), 1997. Handbook of Logic and Language. Elsevier, Amsterdam, NY.

Venema, Y., 2001. Temporal logic. In: Goble, L. (Ed.), The Blackwell Guide to Philosophical Logic. Blackwell Publishers, Malden, USA, pp. 203–223.

Winter, Y., 2001. Flexibility Principles in Boolean Semantics: Coordination, Plurality and Scope in Natural Language. MIT Press, Cambridge, MA.

Zimmermann, T.E., 2011. Model-theoretic semantics. In: Maienborn, C., et al. (Eds.), Semantics: An International Handbook of Natural Language Meaning. de Gruyter, Berlin, pp. 762–801.

Zimmermann, T.E., Sternefeld, W., 2013. Introduction to Semantics. de Gruyter, Berlin.

Logic Models

Joy A Frechtling, Westat, Rockville, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The purpose of this article is to discuss the use of logic models as a tool and consider how logic models can best be used to strengthen research and evaluation studies. We define logic models as a range of visual depictions of a project or program's theory of change or theory of action. The article provides examples of logic models, suggests some historical roots, discusses the development of logic models and how they can be used, and offers a list of their strengths and weaknesses.

Background

The purpose of this article is to discuss the use of logic models as a tool and consider how logic models can best be used to strengthen research and evaluation studies in the social and behavioral sciences. We define logic models as a range of visual depictions of a project or program's theory of change or theory of action. Knowlton and Phillips (2013) offer the following descriptions

> Logic models are a graphic way to organize information and display thinking. They are a visual approach to the implicit maps we all carry in our minds about how the world does or *should* work … Logic models describe the planned action and its expected results.

Critical in their definition and the definition assumed in this article is the belief that a logic model is an approach more than a single template. It is a way of thinking about or deconstructing how one believes "you can get there from here." A logic model makes assumptions explicit and can help to identify gaps or differences in perspectives that otherwise might go unrecognized. Although there are some formats that are more frequently used than others, the approach offers flexibility and is modifiable for different purposes and contexts.

The discussion of logic models in the literature has increased dramatically over the last two decades. An examination of citations in Google Scholar indicates that the term was cited 566 times in 2000 and 3060 times in 2012. A search of ERIC, focusing on the education field, showed that the term was used in the text of studies 16 times in 2000 and 56 times in 2010. While these numbers obviously differ considerably in absolute terms, the message is the same. Logic models are far more in the mainstream today than they were 10 years ago. The idea of logic models has, however, a longer history. According to Funnell and Rogers' (2011) discussions of program theory, the underpinnings of logic models were taking place in the 1950s. In their book, *Purposeful Program Theory*, they provide an illuminating review of the history of program theory leading to the development of visual depictions by the United Way and the W.K. Kellogg Foundation in the late 1990s and early 2000s. Contributors in the last half of twentieth century to the development of this theory include many well-recognized researchers and evaluators such as Daniel Stufflebeam, Carol Weiss, Joseph Wholey, Huey-Tsyh Chen, Peter Rossi, and Len Bickman.

While the nonprofit sector (specifically the United Way and the W.K. Kellogg Foundation) can be credited with leading the charge to use logic models, their acceptance today extends beyond the nonprofit world. Increasingly federal agencies such as Substance Abuse and Mental Health Services Administration, Centers for Disease Control, Department of Education, and the National Science Foundation see the inclusion of a logic model to be an important part of a program description and many different technical assistance efforts have been devoted to helping grantees develop and use logic models as part of their program development. A Google search using the key words 'logic models and the federal government' resulted in a listing of over 7 000 000 potential links (count taken on 13 October 2013).

Logic models are very versatile tools. They can be applied to a range of questions or disciplines. While the examples in this article are taken largely from education programs, logic models are used in such differing contexts as community-based research, studies of judicial procedures, research on health and health interventions, studies of adult learners, and evaluations of preschool service. In a sense while logic models are used to clarify and explicate the content of a program, they themselves are content neutral, being applicable to almost any area when activities or strategies are believed to relate to a specific set of intended or desired outcomes. Additionally, while the examples in this article focus on evaluation studies, logic models can be used for a variety of purposes including needs assessment, structuring research syntheses, supporting program management and monitoring, describing organizational change efforts, and designing dissemination and communication. Logic models have even been applied to describing different evaluation frameworks or approaches (Hansen et al., in press).

What Is a Logic Model?

While there are variations in how logic models are presented, at a minimum they include a set of components – inputs, activities, outputs, outcomes, impacts, and context – and connections between the components (Kellogg Foundation, 2004; Frechtling, 2007) **Figure 1** presents a simple, generic logic model.

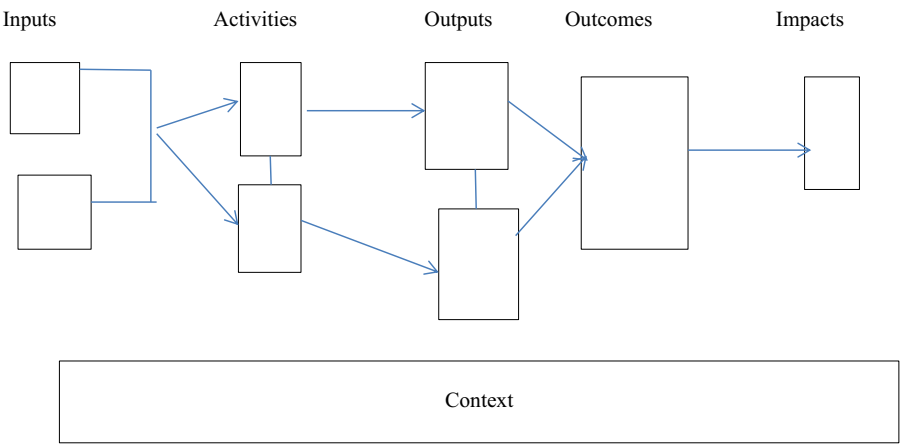

Figure 1 Generic logic model.

The components are defined as follows:

Inputs are the resources brought to a project, typically funding sources or the experiences and knowledge of the individuals and institutions involved.

Activities are the components of the treatments that a project is providing. They are the actions that will be taken to move toward the project's goals and outcomes.

Outputs are the products of activities. Outputs document the implementation of an activity in simple, itemized ways. They document the size and scope of the activity.

Outcomes are results or changes in behaviors or learning. Outcomes identify the goals and objectives of the project. Outcomes are typically specified in the short and longer term. How to differentiate short- and long-term outcomes is project specific and determined in large part by the complexity of the changes that are sought.

Impacts are broader changes in systems that can be linked to the interventions or activities depicted in the model. Typically such impacts include, but also extend beyond, the specific units that are the focus of study.

Contextual factors are the special characteristics of a grantee setting that need to be considered in understanding how a set of interventions may play out. Context can include both current context and historical context, as previous experiences may color how a proposed change is perceived. Contextual factors help clarify the nature of the setting in which the project is being implemented and can play a role in identifying challenges that need to be addressed. Understanding of context also helps inform where approaches and findings might and might not be generalizable.

Connecting the components and showing relationships between and among them are a series of arrows and lines which show expected interdependences or consequences. The arrows are a critical part of the logic model, showing what is hypothesized to lead to what. The lines are also important and portray connections between potentially isolated pieces.

Figure 2 presents a simple logic model showing the expected sequencing of events for a program that provides professional development for teachers with the goal of improving skills and knowledge.

By looking at the outcome we see that the objective of the project is to build human capacity. In this case, it is to produce an increase in skills or knowledge. The strategy for achieving this outcome is providing professional supports. The

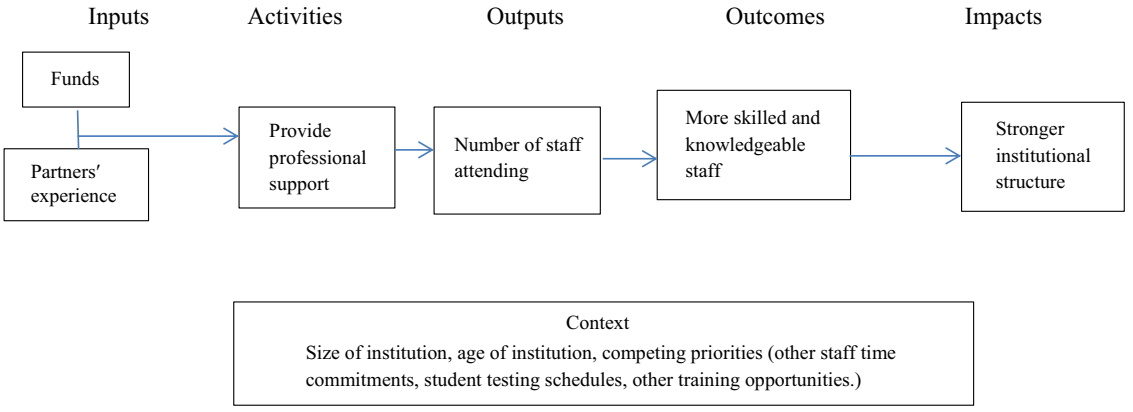

Figure 2 Simple logic model for a professional development program.

output – counts of attendees – is proposed as a way to document that the professional supports have been offered and have attracted participants. There are two inputs or resources named as critical to supporting the work, funds and the partners' prior experience. Finally, the contextual variables that are hypothesized to be important to consider are the size and age of the institution and competing priorities.

It is important to note that an output is different from an outcome. While the latter captures the result or change that is expected from the activity, the former is just one way of documenting that the activity took place. A common error in the development of a theory of change is confusing outputs with outcomes. In this case, it would be suggesting that number of attendees is the 'outcome.' Attending a professional development session and changing in some way because of that attendance are quite different things.

The logic model portrayed in Figure 2, while useful, is quite simple. It clearly shows what might be considered a sequence of key activities, but fails to express in any detail the theory of action that underlies it. Figure 3 presents another kind of logic model, one that is intended to capture not only the sequence of activities but more of the underlying theory that scaffolds it.

What are the key differences in these two depictions? First, Figure 3 is more explicit about how the professional support is developed and who is engaged in it. The purpose is not to merely provide development but to utilize disciplinary faculty in doing so, reflecting a research-based belief that disciplinary knowledge is essential to high-quality professional development. Second, the graphic also makes a distinction between two complementary sets of activities: the development and implementation of the professional development and the purchase of materials. Both are called out as being part of the theory of action – one addressing the need for enhanced human resources, the other for new and better material resources. Third, the logic model in Figure 3 is more specific about the kinds of skills and knowledge that are expected to occur. Teachers are not expected to increase their skills and knowledge generally, but rather with regard to algebraic concepts. Professional development is not expected to have a generic or broad-based impact, but rather change is expected to be specific to a given set of knowledge and skills. Fourth, this

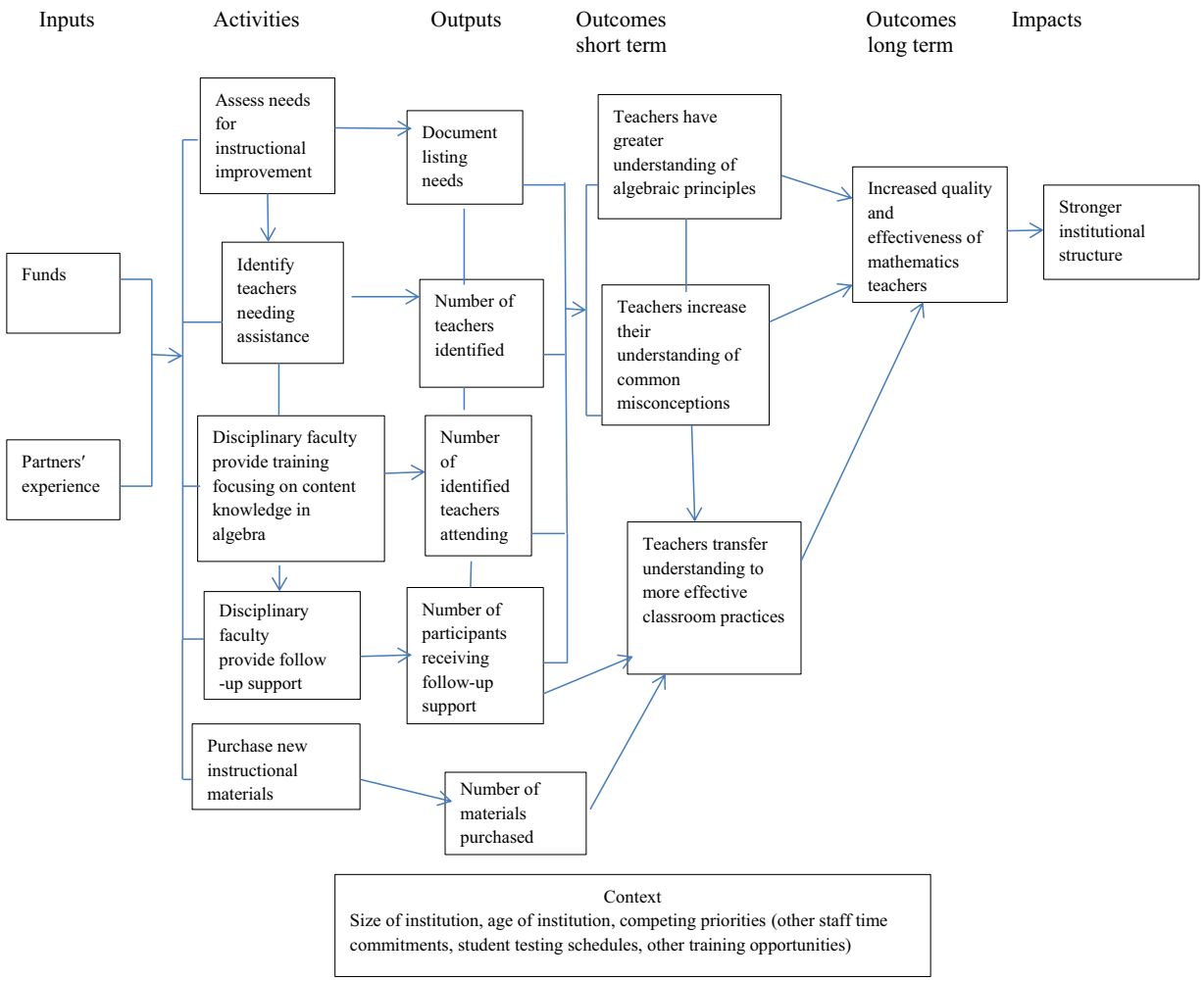

Figure 3 More complex depiction of a theory of action.

increase in algebraic understanding occurs as a result of both the training and follow-up activities. One without the other is not expected to lead to significant changes. Finally, the increase in understanding of algebraic concepts and misconceptions themselves are an antecedent of additional changes – classroom instruction. However, they alone do not lead to this change. Rather both follow-up activities and new instructional materials are needed to bring instructional change about. Figure 3 is considerably more explicit about what is happening and how these activities, alone and in combination, are expected to lead to desired outcomes.

As illustrated in both the simple and more complex examples given above, logic models are typically contained on one page and constructs are portrayed at a relatively high level. Sometimes, however, it is useful to create a family of logic models, which provide a 'drill down' from this one-page view. For example, if the target of study is a program that encompasses several strands of activities that may be carried out independently, it can be useful to use this family approach to clarify both the theory of action underlying the program and its subparts. Suppose a program is intended to apply cutting edge knowledge about how students learn and the contextual factors that influence this learning to instructional opportunities within and outside of the regular academic environment. Such a program could have different strands, including setting – formal or informal educational environments; focus – curriculum, professional development, technology; and targets – students, parents/community, instructional leaders, each with their own theory of change. While the overall outcome of the program is educational improvement, to understand the program and evaluate it, it is useful to look at the different funding strands and their theories of change.

There are many ways to actually create the visual that captures a logic model. Many templates for creating a logic model have been developed and can be easily accessed using a search engine. The University of Wisconsin is just one site that offers this resource (http://www.uwex.edu/ces/pdande/evaluation/evallogicmodelworksheets.html). Software such as Word, Visio, Prezi, PowerPoint, or Excel can all be used to manually develop your model on the computer. There are also several different software packages that have specifically been developed to support logic model development. These include DoView, Innovation Network Logic Model Builder, and ImpactMap.

As stated above, the logic model graphic described above is only one of the ways of visualizing a theory of change. There are a number of other formats that can be used. For example, some like to use a vertical approach, with inputs on the bottom and impacts on top (see Figure 4). Having the impacts and outcomes at the top of the visual helps promote a focus on what the project is trying to accomplish.

Others prefer a nonlinear approach, feeling that the nonlinear approach conveys a more interactive and dynamic image (Figure 5). Of course, capturing the interactions and potential feedback opportunities that might be expected to occur can lead to the development of a visual which, while accurate, loses the ability to communicate in any easy fashion (Figure 6). Logic model developers must always balance the desire to present all critical details with the goal of creating a visual image that clarifies, rather than confuses, communication.

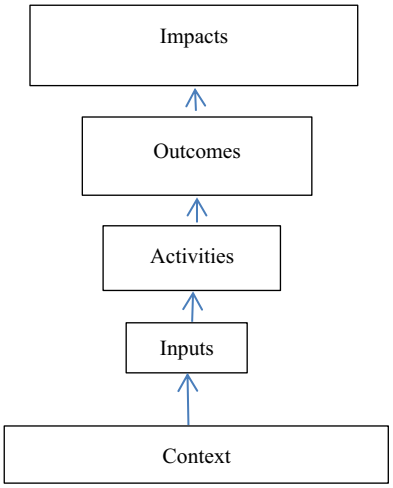

Figure 4 A vertical logic model.

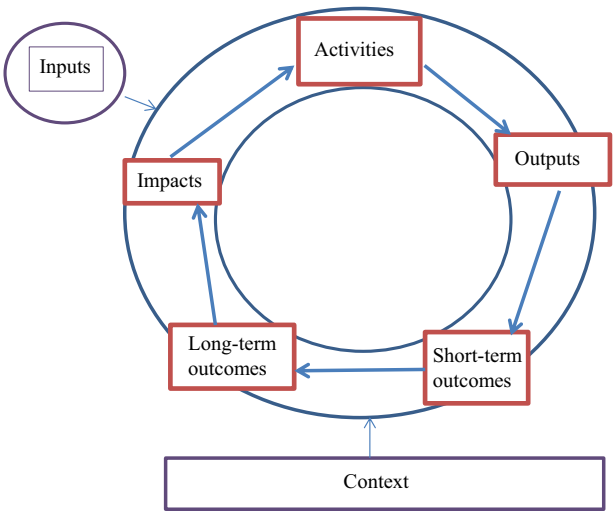

Figure 5 Simple nonlinear logic model.

Logic Models and Methodologies for Research and Evaluation

The logic model visual has applicability within wide range of research and evaluation approaches. Earlier on we noted the philosophical similarity between program theory and logic models. Both are also similar in that they are applicable to many different evaluation contexts and methodologies. They provide a way of clarifying how a series of activities is expected to lead to a set of desired outcomes, but do not prescribe a specific strategy for assessing whether the theory is correct. Logic models are compatible with a range of approaches to studying a problem, as long as goals for a program can be specified. They can be used to scaffold different evaluation approaches, including participatory evaluation, developmental evaluation, randomized treatment control studies, cost–benefit analyses, needs assessment, and evaluability assessment to name just a few approaches. That said, evaluation type might influence how a logic model was developed with, for example,

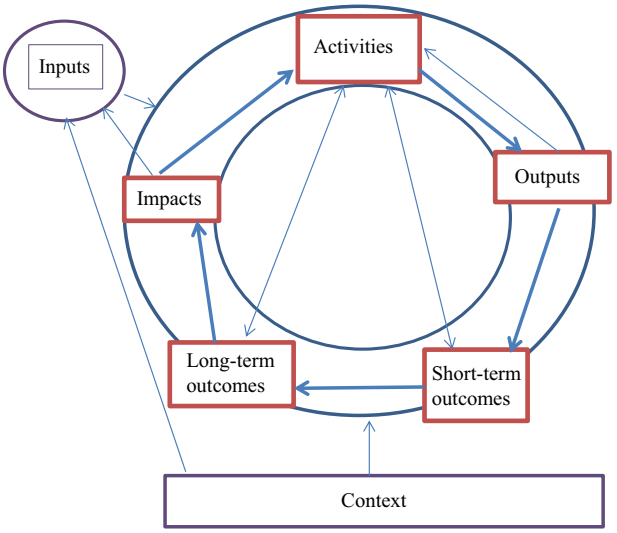

Figure 6 Nonlinear logic model with feedback loops.

participatory evaluation including a broader range of stakeholders than one that does not take the participatory approach (see discussion in next section).

Logic models are also useful tools to support investigation at all stages of program development from design to implementation to completion and even scale up or sustainability. The Community Toolbox (http://ctb.ku.edu/, retrieved 14.10.13.) suggests using logic models during planning, implementation, staff and stakeholder orientation, evaluation, and advocacy.

When a logic model is used as part of an evaluation study, it can provide the foundation for looking at implementation, as well as outcomes. Typically formative evaluation which examines implementation focuses on the left-hand part of the logic model, specifically activities and outputs, and summative assessment examines the right-hand part of the logic model, outcomes and impacts. In the example provided earlier, we can use the logic model as a guide for identifying what should be assessed when. A formative evaluation of the program depicted in this theory of action would examine the development and implementation of the needs assessment, identification of program participants, the initial professional development and follow-up support provided by disciplinary faculty, and the purchase of new instructional materials. The summative evaluation would focus on the results of these activities, first examining whether there are changes in understanding of algebraic principles, common misconceptions, and effective classroom practices; and then consider the longer term issues of quality and effectiveness of mathematics teachers, and ultimately the strength of the educational institutions in which they work.

Logic models are, however, silent with regard to how something can be assessed. Whether the theory of change is examined quantitatively or qualitatively is not prescribed by the logic model format. While the logic model helps to identify what should be measured, it does not specify the particular data collection tools that should be used. Again turning to our illustrative model, we note that while the model suggests that the assessment of the 'follow-up support' provided by faculty is important, there is nothing in the model per se that prescribes how to examine this feature. Indeed, this feature could be addressed in several ways. In addition to information on the number of participants receiving follow-up support (which documents implementation), an evaluation could include assessment of the perceptions of the participants regarding this follow-up or even the nature of the follow-up discussions, using surveys, interviews, or focus group. It is important to remember that decisions regarding how to structure the theory of change are different and separate from decisions regarding how enactment of the theory is to be measured.

Developing and Using Logic Models

The development of a logic model is a 'social' activity. It is a process that involves multiple actors, fosters new relationships and understandings, can be contentious as well as smooth, and can look different at different times and in different contexts. While a single person or group may lead the development of a logic model or create an initial draft, the development of a logic model benefits from the inputs and interactions of multiple, key stakeholders who play different roles and may have different perceptions. Of course when collaborative activity is undertaken in developing the model for a specific activity, it is important to first develop a joint understanding of what a logic model is and is not. Partners need to understand what is meant by the key components – inputs, activities, outputs, outcomes, impacts, and contexts – as well as the importance of tracing specific linkages within and among components.

While developing a logic model can be done at any point in a project's evolution, it can be maximally useful if logic model development is initiated at the planning stage and then revisited at strategic points during the project's life cycle. What is strategic depends on the nature of the project and when key reporting events take place. Some projects only look at their theory of change at the beginning and end of the project; others review the logic model each time new data come in or when a reporting event is looming. The important point is recognizing that a logic model should be expected to be a 'living' and developmental depiction. Information gathered from outputs may well lead to suggestions for changes in activities; examination of actual short-term outcomes against the original short-term outcomes that were expected to occur can also lead to changes in project design, timeline, or level of anticipated outcomes. Modifications are to be expected and even celebrated as new information is acquired.

Who is involved in the development of a logic model will differ depending on the project and who is considered a key stakeholder. In some cases, it is sufficient to have program leadership and the researcher/evaluator at the table. In other cases, it may be necessary to include participants and other key stakeholders, such as representative from the broader community in the process. If the logic model is to be part of a participatory evaluation, then the evaluation participants should also be represented at the logic modeling phase. Including representation from a broad range of groups can have a significant payoff in long run, if ownership and

engagement is established. Decisions about the value or worth of an activity can be facilitated when groups with varied roles and initial understandings have been given the opportunity to create a shared vision. However, the more diverse the groups engaged in logic model development, the more challenging the process and the greater the demands on time and resources.

There are many different ways that the discussion can be structured. Sometimes it is useful for the researcher/evaluator to draft an initial version of the logic model and use this model to kick off broader discussion. If the participants have a general understanding of what a logic model is, they can be asked to create their own logic models and bring them for sharing and discussion. In other cases, a meeting might start with a clean slate, allowing the development to occur as a group activity. Each strategy has strengths and weaknesses and may work well or badly depending on the particular nature of the individuals involved. If time is limited, the first approach – starting with a logic model generated by the researcher/evaluator – is likely to be the most cost-effective.

Although the traditional logic model has a linear appearance, starting with inputs and ending in impacts, the development of a logic model rarely starts with inputs and moves successively through the various components, ending with impacts. Especially, when the project is in its planning or early implementation stages, it is more productive to start with what a project is trying to achieve – desired outcomes or impacts – and work backward, identifying more proximate outcomes and assessing their linkage to the activities proposed to accomplish them (taking into account the inputs or resources available). It can be helpful to pose a series of questions such as what are you trying to ultimately achieve? Who or what are you targeting to bring about these changes? Do the activities that have been proposed appear likely to lead to these outcomes? If not, what else do you need to do to bring these changes about? How long will it take for these changes to occur? What factors in your specific environment might be expected to influence the course and speed of change? Can you do what you need to do with the resources you have available? If not, what are the alternatives? In essence, the discussion needs to bounce back and forth or toggle among the components to assess the extent to which they hang together and work as an integrated whole.

Part of the development of a logic model is also clarifying what is meant by achieving a goal and the value of different types of evidence for documenting goal attainment. Especially when a goal is stated at a fairly high level of abstraction, stakeholders may hold different perceptions of how 'success' should be defined. If, for example, a project goal is to improve student learning, participants may define learning in vastly different ways, including solving complex problems, generating new and innovative ideas, performing well on tests, or acquiring new habit of mind. Each of these has value, but would in all likelihood require different strategies or activities to be undertaken. Activities that might be judged appropriate for increasing test score performance might be quite different than those for generating new and innovative ideas.

Development of a logic model and clarifying the meaning of terms can help stakeholders develop a shared vision of what the project is trying to accomplish and how desired goals are to be reached. At the same time, the discussions can unearth differences in perceptions and understandings that in the short run, at least, may cause anxiety and tensions among groups to emerge. Sometimes accord has reached among different stakeholders only by deliberately allowing some terms to remain fuzzy – a potentially dangerous, but sometime necessary tactic. Logic model developers need to be aware of this potential for creating disequilibrium and be prepared to take the time to hear out the different voices and work with participants to reach a solution. Even if there is agreement, however, it is possible that the discussion that goes hand in hand with logic model development will unearth gaps in the projects' theory of change that no one has recognized. For example, it is not atypical for the development of a set of activities to take on a life of their own, with the connection to project goals getting lost in the process. There may be both activities that do not seem to have any necessary connections to desired outcomes and outcomes that do not link back reasonably to activities identified once the theory of change is systematically examined. In such cases, it may be useful to rethink why such activities were included and whether or not some project redesign is needed. Several sessions of discussion separated by time for reflection and regrouping may be needed before consensus can be achieved.

Strengths and Weaknesses of Logic Models

As this article documents, the logic model is a versatile and useful tool with broad applicability to project design, implementation, evaluation, documentation, and dissemination. However, before crowning the logic model as a 'magic bullet,' it is important to recognize that developing and using logic models has both benefits and limitations. On the benefits side, logic models

- make explicit the theory underlying a project and why certain actions should be expected to lead to specific outcomes,
- force clarification and definition of terms that might otherwise remain vague and subject to multiple interpretations,
- identify gaps in a project's structure or logic,
- assist stakeholders with different perspectives to develop a shared vision and understanding,
- clarify roles and responsibilities,
- provide a tool for management and monitoring,
- identify questions that should be addressed and critical junctures for which data should be collected, and
- are flexible and can be modified as new information is received.

On the limitations side, logic models

- can take considerable time and resources to develop, especially if multiple stakeholder groups with different viewpoints are part of the discussion,
- may not serve to clarify a theory of change if overly complex and detailed,
- can become irrelevant if developed and not revisited,
- can become an obsession, if revisited too frequently and changes are made reflecting small modifications to the program,

- are only as good as the thinking that goes into them and may be flawed if participants thinking is flawed, and
- can be misinterpreted, if users confuse terms such as outputs and outcomes or fail to take into account contextual factors.

See also: Applied Sociology; Cost-Benefit and Cost-Effectiveness Analyses in Evaluation Research; Evidence-Based Practice; Implementation Science; Mixed Methods; Prevention Research.

Bibliography

Bickman, L., 1990. Advances in Program Theory. New Directions for Evaluation, No. 47. Jossey-Bass, San Francisco.

Chen, H., 1990. Theory Driven Evaluation. Sage, Thousand Oaks, CA.

Frechtling, J., 2007. Logic Modeling Methods in Program Evaluation. John Wiley and Sons, San Franciso.

Funnell, S.C., Rogers, P.J., 2011. Purposeful Program Theory: Effective Use of Theories of Change and Logic Models. John Wiley & Sons, San Francisco, CA.

Hansen, M., Alkin, M., Wallace, T., 2013. Depicting the logic of three evaluation theories. Evaluation and Program Planning 38, 34–43.

Knowlton, L.W., Phillips, C.C., 2013. The Logic Model Guidebook. Sage, Los Angeles.

Stufflebeam, D.L., 1983. The CIPP model for program evaluation. In: Madaus, G.F., Scriven, M., Stufflebeam, D.L. (Eds.), Evaluation Models: Viewpoints on Educational and Human Services Evaluation. Kluwer, Boston.

United Way of America, 1996. Measuring Program Outcomes: A Practical Approach. United Way of America, Alexandria, VA.

Weiss, C.H., 1997. Theory-based evaluation: past, present, and future. In: Rog, D., Fournier, D. (Eds.), Progress and Future Directions in Evaluation: Perspectives on Theory, Practice, and Methods. New Directions for Program Evaluation, No.76. Jossey-Bass, San Francisco.

W.K. Kellogg Foundation, 2004. Logic Model Development Guide. Battle Creek, MI.

Wholey, J.S., 1987. Evaluability assessment: developing program theory. In: Bickman, L. (Ed.), Using Program Theory in Evaluation. New Directions for Program Evaluation. No.33. Jossey-Bass, San Francisco.

Logical and Hypothetical Reasoning in Adolescence, Development of

Eric Amsel, Weber State University, Ogden, UT, USA
David Moshman, University of Nebraska–Lincoln, Lincoln, NE, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the development of logical and hypothetical reasoning from early childhood through adolescence and young adulthood. These two forms of reasoning enable adolescents to entertain and coordinate deductively valid and ontologically possible propositions. The development of logical and hypothetical reasoning in adolescence involves both the construction of metalogical knowledge about logical necessity and truth and executive regulation of inferences for systematic thinking. A variety of biological, psychological, social, and cultural factors are known to influence developmental changes in such reasoning.

Adolescence is a time during which new and powerful forms of reasoning emerge, resulting in fundamental transformations in how people think about themselves, others, and the world. Central to adolescent progress is the ongoing development of logical and hypothetical reasoning (Amsel, 2011; Barrouillet and Gauffroy, 2013; Markovits, 2013). Logical reasoning, in its core meaning, is the deduction of conclusions that follow necessarily from given premises. Hypothetical reasoning is the process of creating, and making inferences about, possible worlds, such that (what is believed to be) reality can be compared to alternatives.

The significance of logical and hypothetical reasoning is that they allow the mind to go beyond perceptions of immediately available information to formally and systematically entertain deductively valid and ontologically possible propositions. These two inference skills are implicated in a range of uniquely human activities, from the make-believe play of children to the research practices of scientists. Both the pretending child and researching scientist have been characterized as generating hypothetical alternatives to the way the world is known or believed to be and reasoning logically about their implications and consequences.

The focus on adolescence as a period of major progress in logical and hypothetical reasoning needs justification in light of claims that these inference skills are applied spontaneously by toddlers or, quite the contrary, that they are acquired only with graduate training in science. In contrast to both these positions, Jean Piaget (1972) and Inhelder and Piaget (1958) proposed a qualitative transition in early adolescence (beginning about age 11 or 12 years) to a new stage of 'formal operations' involving advanced levels of logical and hypothetical reasoning. In this article, the authors review research generated over half a century that has supported and challenged Piaget's claims and tested alternative theoretical accounts. The review is organized chronologically, first addressing the roots of these reasoning abilities in young children, their increasingly sophisticated use by older children and young adolescents, the potential for increased control and coordination beyond that, and their promotion with formal education. The review also considers biological, psychological, social, and cultural factors associated with developmental changes in reasoning.

Logical and Hypothetical Reasoning in Young Childhood

In their classic book, *The Growth of Logical Thinking from Childhood to Adolescence*, Inhelder and Piaget (1958) describe major advances in adolescents' ability to reason logically and hypothetically and explain these advances as outcomes of constructing the mental structures of the formal operational stage (Moshman, 2009). Formal operations is the fourth and final stage of cognitive development, when thinking transcends earlier constraints that limited understanding and inference to actions on the world (sensorimotor stage), symbolic representations of the world (preoperational stage), or formal relations (categorical, mathematical) applied to actual situations in the world (concrete operational stage). Inhelder and Piaget's conclusions were based on individual interviews of over 1500 children and adolescents, ages 5 through 16 years, as they solved a variety of physics problems. Their performance reveals an increasing depth of appreciation with age of the logical and mathematical features of the problems and their growing skills to solve them by systematically constructing hypothetical possibilities and inferring their logical consequences. This strict coordination of logical and hypothetical reasoning, which Piaget called hypotheticodeductive reasoning, was seen as central to formal operations.

The Piagetian focus on adolescence has been challenged by information processing theorists, for whom cognitive development in general and the acquisition of inference skills in particular can be explained by the regular and continuous improvement of cognitive processes of memory, attention, and executive control, among others. Information processing researchers have examined young children's logical and hypothetical inference skills to identify the nature of their underlying inferential competence and sources of development. For example, there is evidence that young children could successfully solve simple combinatorial reasoning tasks in which they were asked to generate all the six possible combinations of clothing for a teddy bear who had two tops and three pairs of pants (English, 1993). But there was a substantial decrement in children's performance when the cognitive demands of the task increased by adding more combinations to the two-variable problem (3×3) or by adding a third variable ($2 \times 2 \times 2$ or

2 × 2 × 3). Only adolescents demonstrated inferential competence on the more complex combinatory tasks.

Similarly, there is evidence that young children can solve simple logical reasoning tasks in which they deduce a proposition q from a conditional (If p then q) or universally qualified (all p are q) major premise and the minor premise p (Hawkins et al., 1984; Kuhn, 1977). Their performance on these inference tasks is almost flawless when the content of the premises are congruent with that they know or believe (If the wind is blowing, the flag on the pole is waving; The wind is blowing; Is the flag waving?). However, their performance is below chance when the premise content is incongruent (All birds have wheels; Robin is a bird; Does Robin have wheels?) as such content increases the cognitive demands by requiring that children accept and reason from a disbelieved premise (hypotheticodeductive reasoning). Moreover, even using congruent content, children have more difficulty making the equally valid *modus tollens* inference (if p then q and not q → not p) compared to their performance on the *modus ponens* inference (if p then q and p → q), perhaps because of the additional linguistic demand of processing negations (Note: The expressions on the left of the → are premises and the expression on the right is the conclusion, which may or may not be judged to follow logically from the premises) (Roberge, 1971).

Children also performed particularly poorly when required to inhibit inferences due to an invalid argument, as in the cases of standard fallacious arguments that assert the consequent (if p then q and q → p) or deny the antecedent (if p then q and not p → not q). Invalid inferences are often made because of a failure to distinguish a conditional premise (if p then q), where p is a sufficient but not necessary condition for q, from the biconditional premise (if and only if p then q) where p is both necessary and sufficient for q and thus q is also necessary and sufficient for p (Rumain et al., 1983). Children were shown to inhibit invalid inferences when presented with multiple premises (if p then q, if a then q, if b then q), which highlight that a given antecedent condition (p) is not the only one sufficient for consequent (q), countermanding the invited inference that p is also necessary for q (Rumain et al., 1983). Other research demonstrates that children's performance on conditional reasoning tasks with the premise if p then q improves if they are asked to generate examples of other sufficient conditions for q in the form of instances of not p and q (Markovits, 2013).

Other cases abound of researchers seeking to demonstrate young children's inferential competence. In one line of research, young children have demonstrated the ability to make inferences from premises with incongruent content when they are presented in a pretend context (Amsel et al., 2005b; Dias and Harris, 1988). For example, when the premises *all cats bark* and *Rex is a cat* was presented as part of pretend play activity, 6-year-olds affirmed that *Rex barks* at a rate above chance (Dias and Harris, 1988). Framing empirically false premises as 'pretend' invites children to use their imagination and create a fanciful make-believe world from which they suppose the truth of the premise and infer causes and consequences of the pretend states of affair (Harris, 2000). Make-believe worlds are fanciful in the sense that they are not constrained by and should have no influence on understanding the real world (Leslie, 1987). Young children generally respect this by carefully distinguishing between fantasy and reality, limiting cases where the former impacts the latter (Weisberg, 2013).

Despite limitations in their information processing capacities, young children demonstrate logical and hypothetical inference skills, at least in certain contexts. But most developmentalists, following Piaget, believe logical and hypothetical reasoning continue to develop at least through early adolescence (Amsel, 2011; Barrouillet and Gauffroy, 2013; Klaczynski, 2009; Kuhn, 2009; Markovits, 2013; Moshman, 2011, 2013a), although differing with Piaget and with each other about the details of this developmental transition.

Logical and Hypothetical Reasoning in Older Children and Young Adolescents

Although even young children have logical and hypothetical competencies, older children and young adolescents demonstrate increasing insight about and flexibility in use of these inference skills across different contexts. For example, Andrew Shtulman and Susan Carey (2007) examined the development of judgments about ontological possibilities with 4-, 6-, and 8-year-olds, and college students. Participants were asked to judge whether ordinary (eating chocolate ice cream), improbable (eating pickle ice cream), and impossible (eating lightning) events could occur in real life. Only the college students judged ordinary and improbable events as equally possible, distinguishing them from impossible events. The ability to distinguish improbable from impossible events improved over age as children relied less on whether these events were experienced and more on identifying real-world facts that would preclude impossible events from occurring. The growing capability of older children to reflect on and imagine the real-world status of nonexperienced events was found to generalize across different types of impossible and improbable biological, psychological, and social events (Shtulman, 2009).

The ability to reflect on the ontological status of events is implicated in counterfactual reasoning about states of affairs that could have occurred but did not. Counterfactual reasoning requires the formation of serious possible worlds, which – unlike fanciful make-believe ones – are constrained by reality and can be used to compare it to possible alternatives. Although creating mental worlds of any sort requires the work of the imagination (Harris, 2000), possible worlds can be much more challenging to create than are make-believe ones because of the need in the former to reflect on and evaluate event sequences as ontologically plausible (Amsel, 2011).

Counterfactual reasoning can have emotional consequences of regret upon the realization that one could have acted differently in a situation and thereby avoided an undesirable outcome that has occurred. Although there has been some controversy about when regret is first experienced and judged in others (Rafetseder and Perner, 2012), anticipating potential future regret and taking actions to avoid it appears to develop only in adolescence (Amsel et al., 2005a; Guttentag and Ferrell, 2008).

The greater insight into and flexibility in the use of inference skills by young adolescents is further illustrated in logical reasoning. Older but not younger children distinguish deductive from inductive inferences on the basis of the certainty

associated with the former, an insight that appears to develop further in adolescence (Pillow et al., 2010). Furthermore, young adolescents but not children can explicitly distinguish arguments by their logical validity (e.g., spontaneously categorizing argument 1 as logically valid but argument 2 as logically invalid) irrespective of the truth of the conclusions (Moshman and Franks, 1986).

1. If dogs are bigger than elephants, and elephants are bigger than mice, then dogs are bigger than mice.
2. If dogs are bigger than mice, and elephants are bigger than mice, then dogs are bigger than elephants.

The initial emergence of explicit distinctions between deductive and inductive inferences and between valid and invalid arguments in early adolescence has been characterized as part of a broader acquisition of metalogical knowledge (Moshman, 2011). Metalogical knowledge is explicit conceptual or epistemological understanding about the justifiability of inferential processes and the necessary (as distinct from empirical) nature of logical truth. Correct use of inferential skills is no guarantee that the meaning and significance of the inferences are appreciated or that their justification is understood. The dramatic advance of metalogical understanding in early adolescence is likely associated with the increase in other inference skills demonstrated on logical reasoning tasks. For example, young adolescents no longer need the support of a pretend play context in order to mentally disregard their prior beliefs and knowledge and correctly infer conclusions in conditional reasoning with incongruent premises (Amsel et al., 2005b).

The emergence of increasingly explicit knowledge about logic and reasoning can be attributed most directly to active reflection on inferential processes, including abstraction of their logical form and construction of advanced conceptions of necessity and possibility. Such reflection generally occurs in social contexts of argumentation and justification and is likely influenced by a variety of biological and cognitive factors. For example, brain changes in adolescence are associated with increases in working memory and cognitive control (Paus, 2005) and may enable the construction of higher order metalogical knowledge.

In addition to biological and cognitive considerations, there are also changes in the schooling of older children and young adolescents, at least in cultures with compulsory education. Middle or Junior High school students (grades 6–8 or 7–9) have a much more rigorous curriculum than younger students. Notably, for many, postelementary education is their first exposure to algebra and geometry. Mathematical instruction in these topics has been shown to be related to improvements in students' formal concepts of justification, proof, and equivalence (Alibali et al., 2007; Bieda, 2010), which may also have consequences for their thinking about their own reasoning and, more generally, about the logical nature of mathematics as distinct from empirical matters of science.

Finally, culture itself, and the value it places on higher ordered thinking, may play a role in older children's and young adolescents' tendency to think about their reasoning. Cross-cultural researchers argue that the acquisition and use of higher ordered inference skills depends on whether or not a given culture views them as useful (Cole, 2006). For example, traditional villagers in third-world cultures resist attempts to get them to engage in the simplest forms of logical reasoning when the premises are not known to be true (Luria, 1976). Consider the dialogue between Alexander Luria (a Russian psychologist) and a rural Uzbekistan farmer (Luria, 1976: pp. 108–109). Luria presented the following logical argument: In the Far North, where there is snow, all bears are white. Novaya Zemlya is in the Far North and there is always snow there. What color are the bears there? The Uzbekistan adult replied, "There are different sorts of bears." Luria then repeated the syllogism, and the farmer responded: "I don't know. I have never seen any others…Each locality has its own animals: if it's white, they will be white; if it's yellow, they will be yellow." Luria continued: "But what kind of bears are there in Novaya Zemlya?" The farmer replied: "We always speak only of what we see; we don't talk about what we haven't seen."

In summary, early adolescence is marked by advances in metalogical knowledge and in its use to guide logical and hypothetical reasoning. Inferences come to be intentionally deployed in the service of critically evaluating the logical consistency and ontological plausibility of claims made about the real world. Reasoning, which can be defined as epistemologically self-regulated thinking (Moshman, 2013b), becomes increasingly reflective in its orientation toward justifiable conclusions. This is a notable development beyond the child's use of logical and hypothetical inference skills to create and reason within alternative make-believe worlds.

The progression from fantasy to formality in the scope of reasoning is consistent with the Piagetian account of development. It remains unclear, however, whether all individuals in all cultures achieve something like formal operations. The rural Uzbekistan farmer made famous by Luria gives perspective on the importance of the biocultural context on the development of inference skills. Rather than being intrinsically driven to construct metalogical knowledge or to refine their inference skills in light of such knowledge, adolescents and adults may come to do so because of the biocultural context in which they grow up. The complex interactions between the society, schooling, and social context of young adolescents as they themselves go through the hormonal changes associated with puberty affect their acquisition of inferential skills. In a biocultural context that is supportive of the acquisition of metalogical knowledge and the refinement of logical and hypothetical reasoning skills, inferential abilities reach levels not seen in childhood and may continue to develop well beyond early adolescence.

Further Development of Logical and Hypothetical Reasoning in Adolescence

Many individuals demonstrate improvements in logical and hypothetical reasoning beyond early adolescence. Hypothetical thinking improves as adolescents become capable of making more complex inferences about alternative possible worlds on counterfactual reasoning tasks. For example, Ana Rafetseder and her colleagues presented counterfactual reasoning tasks requiring simpler and more complex inferences (exemplified below) to 6-, 10-, and 14-year-olds and adults (Rafetseder et al.,

2013). Researchers acted out stories with dolls and props and carefully checked participants' understanding of the details of each story before asking the counterfactual question, which in the versions below was "What would have happened if Suzi had taken her shoes off? Would the floor be clean or dirty?" Overall correct performance was much lower on the more complex version 2 (the floor would be dirty, 64%) than on the simpler version 1 (the floor would be clean, 98%) story. Version 2 requires mentally inferring a complex counterfactual outcome in which Suzi keeps the floor clean but Max dirties it when he subsequently walks through it. Only the adults performed equally well on both versions; 14-year-olds showed a very high level of correct performance but still performed better on the simpler version 1 than on the more complex version 2 story. The 10- and 6-year-old children performed significantly worse than the 14-year-olds and adults, with approximately 50% not making a single correct response on version 2 stories. The authors concluded that there remain challenges in creating possible worlds that extend into adolescence.

Version 1. One day the floor is nice and clean, but then something happens. Suzi comes home and does not take her shoes off and makes the floor all dirty.

Version 2. One day the floor is nice and clean, but then something happens. Suzi and Max come home and they do not take their shoes off. Suzi walks in (followed immediately by Max as depicted by dolls) and makes the floor all dirty.

Logical thinking also improves as adolescents become better able to make deductive inferences on abstract conditional reasoning tasks. These tasks provide no knowledge-based counterexamples to support a sufficiency relation between antecedent and consequent. For example, Henri Markovits and Hugues Lortie-Forgues (2011) had participants aged 12 and 15 years make inferences about abstract conditionals (e.g., if a person daigues then they will become a gadoro), which were presented as conditions on the planet Kronus, where there were different names for things than on Earth. The older group outperformed the younger, on average, particularly when resisting making invalid inferences on denying the antecedent and asserting the consequent items. Such errors would be expected if participants assume a necessary and sufficient relation between the abstract antecedent and consequent. These and other results suggest that abstract understanding of conditional premises as sufficient relations between antecedents and consequents may continue to improve at least through age 15 years (Markovits, 2013).

These changes in reasoning documented above occur as part of a broader set of cognitive changes affecting adolescents (Kuhn, 2009). For example, problem solving can be performed more quickly and efficiently by adolescents than by children because a larger range of relevant experiences, knowledge, and heuristics (simplified inference strategies or 'rules of thumb' that lead to quick but sometimes incorrect conclusions) can be automatically activated in memory (Reyna and Farley, 2006). However, the unique challenge of ontologically complex and logically abstract inferences is to inhibit automatically activated knowledge, experiences, and heuristics in favor of making carefully controlled inferences that are intentionally guided by metalogical knowledge. That is, although metalogical knowledge aids in adolescents' understanding what inference to make, adolescents need to exercise executive control over their information processing system in order to make such inferences free from bias resulting from automatically activated experiences, knowledge and heuristics.

It is notable then that middle adolescence is for many a period of improvements in executive function, which is a "deliberate top-down neurocognitive process involved in the conscious, goal-directed control of thought, action, and emotion" (Zelazo and Carlson, 2012: p. 354). The improvement of executive function is linked to neurological changes in the speed and efficacy of neurons in a specific area of the brain's prefrontal cortex and their connections to other brain areas (Blakemore, 2012). Changes in executive function are thought to better enable middle adolescents to inhibit automatic processes and to rely instead on more controlled and deliberate processes in order to make intentional inferences that are guided by metalogical knowledge. Direct and indirect measures of prefrontal cortex functioning have been shown to be related to correctly resolving conflicts between activated beliefs about the truth versus the logical validity of conclusions on logical reasoning tasks (Steegen and De Neys, 2012; Stollstorff et al., 2012).

In addition to executive function and other information processing abilities necessary for improved inferential skills, many adolescents continue to construct metalogical knowledge and increasingly use it to guide their inferences. More broadly, adolescents' reasoning reflects their developing epistemologies, which address the nature of and foundations for their own and others' knowledge claims about the world, especially with respect to normative issues of truth and justification. Many adolescents become increasingly focused on whether knowledge claims are best treated as objective statements that are either factually true or validly deduced, subjective tastes that are merely more or less preferred, or reasonable interpretations that can be rationally evaluated as convincingly justified (Chandler et al., 1990; Kuhn, 1991, 2009; Kuhn et al., 1988; Moshman, 2011, 2013b).

Besides changes at the personal level, there are also changes at the interpersonal level that contribute to adolescent reasoning, including disputes and disagreements with peers and parents. Effective argumentation requires coordinating hypothetical and logical reasoning by highlighting the quality (strengths and weaknesses) of logically possible justifications of knowledge claims (Clark and Sampson, 2008; Kuhn, 1991). Furthermore, argumentation implicates both (1) the availability of metalogical knowledge about the quality of arguments and (2) inferential skills to strategically apply that knowledge in ongoing arguments to defend one's own and undermine one's opponent's positions. Argumentation skills have been shown to improve from childhood to middle adolescence in both ways. For example, 11th grade adolescents, who on average had more sophisticated epistemological understanding than their 7th and 9th grade counterparts, were also generally better able to detect informal reasoning fallacies than younger students (Weinstock et al., 2006). Similarly, adolescents improve in understanding the need to strategically undermine their opponents' positions and not just to defend their own positions (Kuhn and Udell, 2007), although they

may not always implement such a strategy (Stanovich et al., 2012). Finally, argumentation has been shown to improve with practice, particularly with interventions that help arguers reflectively assess the quality of their performance (Chinn and Anderson, 1998; Iordanou, 2010; Udell, 2007). However, cultural variation exits in the perceived value of argumentation as a mean of promoting cognitive development or educational outcomes (Kuhn et al., 2011).

At the societal level, high school education may contribute to improving metalogical understanding of and skills for inferential reasoning. The increased premium placed on STEM (Science, Technology, Engineering, and Mathematics) preparation has resulted in instruction promoting students' abilities to engage in scientific practices (NRC, 2012). Like argumentation, the forms of reasoning practiced by scientists necessitate the coordination of logical and hypothetical reasoning by focusing on generating hypotheses, designing experiments, and evaluating evidence (Inhelder and Piaget, 1958; Kuhn et al., 1988). Proficiency in scientific reasoning involves both the metalogical understanding of formal theory–evidence relationships and skills to apply such understanding to the tasks of formulating and testing hypotheses and evaluating evidence.

Although children have been shown to have a variety of scientific reasoning skills (Morris et al., 2013), evidence points to further improvements beyond childhood and often well into adolescence. Among these improvements are increases in metalogical and epistemological understanding of scientific reasoning and in associated inferential skills. Metalogical developments include more carefully delineated distinctions between theory and evidence, particularly in the forms of evidence that would uniquely test theoretically based hypotheses (Bullock et al., 2009; Kuhn et al., 1988). Epistemological developments include insights into the nature of scientific knowledge, such as recognizing the tentativeness of empirical (as distinct from logical) claims (Sandoval, 2005). There are also corresponding improvements from childhood to adolescence in skills to minimize inferential biases in scientific reasoning. These 'debiasing' skills include reducing the effect of prior beliefs in the evaluation of evidence, generating and evaluating a broader set of possible hypotheses, generating disconfirming and unconfounded tests of hypotheses, and keeping accurate and systematic records (Amsel and Brock, 1996; Klahr et al., 1993; Kuhn et al., 1995; Schauble, 1990, 1996). Although scientific reasoning has been shown to improve with practice and instruction (Kuhn et al., 1995; Schauble, 1996), metalogical understanding of and skills for making unbiased inferences remain less than optimal.

In summary, both logical and hypothetical reasoning tend to become more sophisticated over the course of adolescence, not only when inferences are made individually but also when the two are coordinated, as in argumentation and scientific reasoning. This reflects a developmental trend toward consolidation in the sense that reasoning is better guided by increased metalogical understanding of the logical validity and ontological consistency of inferences and more effectively controlled by an increasingly powerful information processing system. The more powerful information processing system enables the activation and use of justified inference strategies and the corresponding inhibition of automatically activated but potentially biasing experiences, knowledge, and heuristics. But the powerful cognitive tendency to respond automatically remains, resulting in biased reasoning, even when one has metalogical knowledge of valid alternatives. Dual process theoretical accounts of reasoning explain how inferences may be biased by automatically activated information and also identify conditions under which one more readily deploys metalogical knowledge to foster valid reasoning (Amsel et al., 2008; Barrouillet and Gauffroy, 2013; Klaczynski, 2009; Stanovich et al., 2012).

These developments in reasoning beyond early adolescence are neither universal nor inevitable. They are a product of complex interactions among many factors, including cognitive reflection and self-regulation, preferences for deliberative reasoning, neurological functioning, interpersonal interactions, schooling, and cultural context.

Educational and Professional Contributions to Logical and Hypothetical Reasoning

As noted previously, the trajectory of logical and hypothetical reasoning is affected by educational experiences. This effect becomes more pronounced over age. For example, college-educated adults are better able than non-college-educated adults to ignore their prior beliefs about a causal relation and validly evaluate data that contradict those beliefs (Amsel and Brock, 1996; Kuhn et al., 1988). A similar effect of education has been found for argumentation, with a college education related to improved performance on a variety of measures of argumentative skills (Kuhn, 1991), including a reduction of the so-called myside bias, a preference for supporting one's own position over opposing one's opponent (Toplak and Stanovich, 2003).

An intriguing suggestion of the effect of education on reasoning was provided by Demetriou and Bakracevic (2009), who presented conditional reasoning tasks to adolescents (13–15 years of age) and groups of young (23–25 years of age), middle (33–35 years of age), and mature (43–45 years of age) adults who had or had not completed college. The college educated adult groups made more correct responses than the non-college-educated adult groups, with adolescents scoring lower than the educated adults and higher than the noneducated adults. College-educated adults also were more certain of their answers than non-college-educated adults or adolescents, perhaps reflecting their greater understanding of the logical necessity of the inferences they were making.

The authors concluded that education had a beneficial effect on inference skills and understanding. Their data, however, do not rule out the alternative interpretation that students with greater reasoning abilities are simply more likely to successfully complete college. Darrin Lehman and Richard Nisbett (1990) provided stronger evidence for the causal effect of schooling on reasoning as part of a research program exploring the trainability of reasoning. College freshman in the Natural Sciences, Humanities, Social Science, and Psychology were given tests of probabilistic reasoning (requiring hypotheticodeductive inferences) and conditional reasoning (requiring deductive inferences). Although no different in their reasoning skills as freshmen, these students were retested as seniors to assess the changes in their reasoning over time. Students in the natural

sciences and humanities improved substantially in conditional reasoning compared to those in social science and psychology. At least for the natural science students, this improvement was related to the number of math and computer science courses taken, which require understanding abstract formal relations. Students in the social sciences and psychology improved substantially in probabilistic reasoning compared to those in the natural sciences and humanities, which was related to the number of statistics courses taken.

The impact of schooling also applies to professional training in science and the law leading to expertise in forms of logical and hypothetical reasoning. Expertise in science involves acquiring both general hypotheticodeductive inference abilities (e.g., designing simple and direct tests of hypotheses and drawing valid conclusions about theories in light of the evidence) and specific knowledge about the phenomenon of interest (e.g., knowing which variables may be possible causes and how best to test them). One study compared practicing scientists (experts) and college students (novices in how they designed tests in a computer simulation of the causes of the memory spacing effect – that better memory results when studying is spaced over time than massed together (Schunn and Anderson, 1999). Half the experts were psychology researchers in cognitive psychology and so were domain experts and the other half were psychology researchers in other domains and so were task experts. Half the novices had high mathematics SAT scores (high ability novices) and half had average scores (average ability novices). The results demonstrated that there were specific knowledge differences between the domain experts and all other groups. However, the two expert groups made valid hypotheticodeductive inferences more often than the two novice groups. The results suggest that scientists are experts in general hypotheticodeductive reasoning and can apply it outside their immediate domain of expertise.

Another study of professional training in law and psychology demonstrated a similar generality in inferential expertise. Amsel et al. (1991) compared novices and experts in law and psychology in how they evaluate everyday causal statements (e.g., 'kicking the television (TV) set fixed the picture'). As experts in science, psychologists and psychology graduate students tended to make causal inferences more often than experts in law by appealing to statistical evidence of the association between the cause and the effect (e.g., every time the TV is kicked the picture is fixed). As experts in law, lawyers and law students tended to make causal statements more often than experts in science by appealing to counterfactual states of affair (e.g., if the TV had not been kicked it would not have been fixed). The novice groups (police officers and psychology undergraduates) tended not to find the statistical or counterfactual justifications particularly convincing and preferred to make causal statements by appealing to physical mechanisms (e.g., by kicking the TV, the connection between the wires got better).

Conclusion

Children's logical and hypothetical reasoning become increasingly epistemologically self-regulated due to the development of metalogical knowledge and information processing skills, a process that often continues well into adolescence and sometimes beyond. This developmental trajectory is the product of a broad set of factors ranging from biology to culture and includes a central role played by the individual's own motivated search to understand the epistemological basis for valid inference. Logical and hypothetical inferences increasingly come under metacognitive control, and this process may continue for older adolescents and adults who select experiences like attending university and professional school where they acquire expertise in the use of advanced reasoning skills. Logical and hypothetical reasoning is a life span phenomenon: It is present in young children in a variety of forms and continues to be refined well into adulthood in other forms. However, adolescence is a time of especially great developmental importance and potential in the development of such reasoning.

See also: Childhood and Adolescence: Developmental Assets; Counterfactual Reasoning, Qualitative: Philosophical Aspects; Deductive Reasoning Systems; Epistemic Doubt During Adolescence; Piaget's Theory of Cognitive Development; Piaget's Theory of Human Development and Education; Piaget, Jean (1896–1980); Practical Reasoning: Philosophical Aspects; Problem Solving and Reasoning: Case-based; Stress in Adolescence: Effects on Development.

Bibliography

Alibali, M.W., Knuth, E.J., Hattikudur, S., McNeil, N.M., Stephens, A.C., 2007. A longitudinal examination of middle school students' understanding of the equal sign and equivalent equations. Mathematical Thinking and Learning 9 (3), 221–246.

Amsel, E., 2011. Hypothetical thinking in adolescence: its nature, development, and applications. In: Amsel, E., Smetana, J.G. (Eds.), Adolescent Vulnerabilities and Opportunities: Developmental and Constructivist Perspectives. Cambridge University Press, New York, NY, US, pp. 86–113.

Amsel, E., Brock, S., 1996. The development of evidence evaluation skills. Cognitive Development 11 (4), 523–550.

Amsel, E., Smalley, J., 2000. Beyond really and truly: children's counterfactual thinking about pretend and possible worlds. In: Riggs, K., Mitchell, P. (Eds.), Children's Reasoning and the Mind. Psychology Press, Brighton, UK, pp. 99–134.

Amsel, E., Langer, R., Loutzenhiser, L., 1991. Do lawyers reason differently from psychologists? A comparative design for studying expertise. In: Sternberg, R.J., Frensch, P.A. (Eds.), Complex Problem Solving: Principles and Mechanisms. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, England, pp. 223–250.

Amsel, E., Cottrell, J., Sullivan, J., Bowden, T., 2005a. Anticipating and avoiding regret as a model of adolescent decision-making. In: Jacobs, J., Klaczynski, P. (Eds.), The Development of Judgment and Decision-Making in Children and Adolescence. Erlbaum, Mahwah, NJ, pp. 119–154.

Amsel, E., Trionfi, G., Campbell, R., 2005b. Reasoning about make-believe and hypothetical suppositions: towards a theory of belief-contravening reasoning. Cognitive Development 20, 545–575.

Amsel, E., Klaczynski, P.A., Johnston, A., Bench, S., Close, J., Sadler, E., Walker, R., 2008. A dual-process account of the development of scientific reasoning: the nature and development of metacognitive intercession skills. Cognitive Development 23, 452–471.

Barrouillet, P., Gauffroy, C., 2013. Dual processes and mental models in the development of conditional reasoning. In: Barrouillet, P., Gauffroy, C. (Eds.), The Development of Thinking and Reasoning. Psychology Press, New York, pp. 95–121.

Bieda, K.N., 2010. Enacting proof-related tasks in middle school mathematics: challenges and opportunities. Journal for Research in Mathematics Education 41 (4), 351–382.

Blakemore, S., 2012. Imaging brain development: the adolescent brain. Neuroimage 61 (2), 397–406. http://dx.doi.org/10.1016/j.neuroimage.2011.11.080.

Bullock, M., Sodian, B., Koerber, S., 2009. Doing experiments and understanding science: development of scientific reasoning from childhood to adulthood. In: Schneider, W., Bullock, M. (Eds.), Human Development from Early Childhood to Early Adulthood: Findings from a 20 Year Longitudinal Study. Psychology Press, New York, NY, US, pp. 173–197.

Chandler, M.J., Boyes, M., Ball, L., 1990. Relativism and stations of epistemic doubt. Journal of Experimental Child Psychology 50 (3), 370–395.

Chinn, C., Anderson, R., 1998. The structure of discussions intended to promote reasoning. The Teachers College Record 100, 315–368.

Clark, D.B., Sampson, V., 2008. Assessing dialogic argumentation in online environments to relate structure, grounds, and conceptual quality. Journal of Research in Science Teaching 45 (3), 293–321.

Cole, M., 2006. Culture and cognitive development in phylogenetic, historical, and ontogenetic perspective. In: Kuhn, D., Siegler, R.S., Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology, Cognition, Perception, and Language, sixth ed., vol. 2. John Wiley & Sons Inc., Hoboken, NJ, US, pp. 636–683.

Demetriou, A., Bakracevic, K., 2009. Reasoning and self-awareness from adolescence to middle age: organization and development as a function of education. Learning and Individual Differences 19, 181–194.

Dias, M.G., Harris, P.J., 1988. The effect of make-believe play on deductive reasoning. British Journal of Developmental Psychology 6, 207–221.

English, L.D., 1993. Children's strategies in solving two- and three-dimensional combinatorial problems. Journal for Research in Mathematics Education 24, 255–273.

Guttentag, R., Ferrell, J., 2008. Children's understanding of anticipatory regret and disappointment. Cognition and Emotion 22, 815–832.

Harris, P.L., 2000. The Work of the Imagination. Blackwell, Oxford.

Hawkins, T., Pea, R.D., Glick, J., Scribner, S., 1984. "Merds that laugh don't like mushrooms": evidence for deductive reasoning by preschoolers. Developmental Psychology 20, 584–594.

Inhelder, I., Piaget, J., 1958. The Growth of Logical Thinking. Basic Books, New York, NY.

Iordanou, K., 2010. Developing argument skills across scientific and social domains. Journal of Cognition and Development 11 (3), 293–327.

Klaczynski, P.A., 2009. Cognitive and social cognitive development: dual-process research and theory. In: Evans, J.B. St. T., Frankish, K. (Eds.), In Two Minds: Psychological and Philosophical Theories of Dual Processing. Oxford University Press, Oxford, UK, pp. 265–292.

Klahr, D., Fay, A., Dunbar, K., 1993. Heuristics for scientific experimentation: a developmental study. Cognitive Psychology 25, 111–146.

Kuhn, D., 1977. Conditional reasoning in children. Developmental Psychology 13, 342–353.

Kuhn, D., 1991. The Skills of Argument. Cambridge University Press, Cambridge, England.

Kuhn, D., 2009. Adolescent thinking. In: . In: Lerner, R.M., Steinberg, L. (Eds.), Handbook of Adolescent Psychology, third ed., vol. 1. Wiley, Hoboken, NJ, pp. 152–186.

Kuhn, D., Udell, W., 2007. Coordinating own and other perspectives in argument. Thinking & Reasoning 13 (2), 90–104.

Kuhn, D., Amsel, E., O'Loughlin, M., 1988. The Development of Scientific Thinking Skills. Academic Press, Orlando, FL.

Kuhn, D., Wang, Y., Li, H., 2011. Why argue? Developing understanding of the purposes and value of argumentive discourse. Discourse Processes 48, 26–49.

Kuhn, D., Garcia-Mila, M., Zohar, A., Andersen, C., 1995. Strategies of knowledge acquisition. Monographs of the Society for Research in Child Development 60, 1–128.

Lehman, D., Nisbett, R.E., 1990. A longitudinal study of the effects of undergraduate education on reasoning. Developmental Psychology 26, 952–960.

Leslie, A.M., 1987. Pretense and representation: the origins of "theory of mind". Psychological Review 94, 412–426.

Luria, A.R., 1976 (M. Lopez-Morillas, L. Solotaroff, Trans.). In: Cole, M. (Ed.), Cognitive Development: Its Cultural and Social Foundations. Harvard University Press, Cambridge, MA.

Markovits, H., 2013. The development of abstract conditional reasoning. In: Barrouillet, P., Gauffroy, C. (Eds.), The Development of Thinking and Reasoning. Psychology Press, New York, pp. 71–91.

Markovits, H., Lortie-Forgues, H., 2011. Conditional reasoning with false premises facilitates the transition between familiar and abstract reasoning. Child Development 82 (2), 646–660.

Morris, B.J., Croker, S., Masnick, A.M., Zimmerman, C., 2013. The emergence of scientific reasoning. In: Kloos, H., Morris, B.J., Amaral, J.L. (Eds.), Current Topics in Children's Learning and Cognition. InTech, Rijeka, Croatia, pp. 61–82.

Moshman, D., 2009. Adolescence. In: Müller, U., Carpendale, J.I.M., Smith, L. (Eds.), The Cambridge Companion to Piaget. Cambridge University Press, Cambridge, UK, pp. 255–269.

Moshman, D., 2011. Adolescent Rationality and Development: Cognition, Morality, and Identity, third ed. Psychology Press, New York.

Moshman, D., 2013a. Adolescent Rationality. Advances in Child Development and Behavior, vol. 45. Elsevier, Amsterdam, 155–183.

Moshman, D., 2013b. Epistemic cognition and development. In: Barrouillet, P., Gauffroy, C. (Eds.), The Development of Thinking and Reasoning. Psychology Press, New York, pp. 13–33.

Moshman, D., Franks, B.A., 1986. Development of the concept of inferential validity. Child Development 57 (1), 153–165. http://dx.doi.org/10.2307/1130647.

National Research Council (NRC), 2012. A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. The National Academies Press, Washington, DC.

Paus, T., 2005. Mapping brain maturation and cognitive development during adolescence. Trends in Cognitive Sciences 9 (2), 60–68.

Piaget, J., 1972. Intellectual evolution from adolescence to adulthood. Human Development 15, 1–12.

Pillow, B.H., Pearson, R.M., Hecht, M., Bremer, A., 2010. Children's and adults' judgments of the certainty of deductive inferences, inductive inferences, and guesses. The Journal of Genetic Psychology: Research and Theory on Human Development 171 (3), 203–217. http://dx.doi.org/10.1080/00221320903300403.

Rafetseder, E., Perner, J., 2012. When the alternative would have been better: counterfactual reasoning and the emergence of regret. Cognition and Emotion 26 (5), 800–819. http://dx.doi.org/10.1080/02699931.2011.619744.

Rafetseder, E., Cristi-Vargas, R., Perner, J., 2010. Counterfactual reasoning: developing a sense of nearest possible world. Child Development 81 (1), 376–389. http://dx.doi.org/10.1111/j.1467-8624.2009.01401.x.

Rafetseder, E., Schwitalla, M., Perner, J., 2013. Counterfactual reasoning: from childhood to adulthood. Journal of Experimental Child Psychology 114 (3), 389–404. http://dx.doi.org/10.1016/j.jecp.2012.10.010.

Reyna, V.F., Farley, F., 2006. Risk and rationality in adolescent decision making: implications for theory, practice, and public policy. Psychological Science in the Public Interest 7 (1), 1–44.

Roberge, J.J., 1971. Some effects of negation on adults' conditional reasoning abilities. Psychological Reports 29 (3), 839–844.

Rumain, B., Connell, J., Braine, M.D.S., 1983. Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults. Developmental Psychology 19, 471–481.

Sandoval, W.A., 2005. Understanding students' practical epistemologies and their influence on learning through inquiry. Science Education 89 (4), 634–656.

Schauble, L., 1996. The development of scientific reasoning in knowledge-rich contexts. Developmental Psychology 32, 102–119.

Schunn, C.D., Anderson, J.R., 1999. The generality/specificity of expertise in scientific reasoning. Cognitive Science 23 (3), 337–370.

Shtulman, A., 2009. The development of possibility judgment within and across domains. Cognitive Development 24, 293–309.

Shtulman, A., Carey, S., 2007. Improbable or impossible? How children reason about the possibility of extraordinary events. Child Development 78, 1015–1032.

Simoneau, M., Markovits, H., 2003. Reasoning with premises that are not empirically true: evidence for the role of inhibition and retrieval. Developmental Psychology 39 (6), 964–975.

Stanovich, K.E., West, R.F., Toplak, M.E., 2012. Judgment and decision making in adolescence: separating intelligence from rationality. In: Reyna, Valerie F., Chapman, Sandra B., Dougherty, Michael R., Confrey, Jere (Eds.), The Adolescent Brain: Learning, Reasoning, and Decision Making. American Psychological Association, Washington, DC, US, pp. 337–378.

Steegen, S., De Neys, W., 2012. Belief inhibition in children's reasoning: memory-based evidence. Journal of Experimental Child Psychology 112 (2), 231–242.

Stollstorff, M., Vartanian, O., Goel, V., 2012. Levels of conflict in reasoning modulate right lateral prefrontal cortex. Brain Research 1428, 24–32.

Toplak, M.E., Stanovich, K.E., 2003. Associations between myside bias on an informal reasoning task and amount of post-secondary education. Applied Cognitive Psychology 17 (7), 851–860.

Udell, W., 2007. Enhancing adolescent girls' argument skills in reasoning about personal and non-personal decisions. Cognitive Development 22, 341–352.

Weinstock, M.P., Neuman, Y., Glassner, A., 2006. Identification of informal reasoning fallacies as a function of epistemological level, grade level, and cognitive ability. Journal of Educational Psychology 98 (2), 327–341.

Weisberg, D.P., Beck, S.R., 2010. Children's thinking about their own and others' regret and relief. Journal of Experimental Child Psychology 106 (2), 184–191.

Weisberg, D.S., 2013. Distinguishing imagination from reality. In: Taylor, M. (Ed.), The Oxford Handbook of the Development of Imagination. Oxford University Press, New York, NY, pp. 75–93.

Zelazo, P., Carlson, S.M., 2012. Hot and cool executive function in childhood and adolescence: development and plasticity. Child Development Perspectives 6 (4), 354–360.

Logical Positivism and Logical Empiricism

Herbert Keuth, Eberhard Karls Universität, Tübingen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Logical positivism and logical empiricism developed in the early twentieth century. The Vienna Circle, the Berlin Society for Empirical Philosophy, and the Lvov-Warsaw School of Logic made important contributions to logic, mathematics, philosophy of science, and language analysis. The Vienna Circle proposed a scientific world-conception. Its goal is unified science; its method is logical analysis, which will unmask traditional philosophical problems as pseudoproblems or transform them into empirical problems. Factual knowledge results only from experience (empiricist), which rests on what is immediately given (positivist). In the 1930s, phenomenalist neopositivism was replaced by realist neoempiricism.

The terms 'logical positivism' and 'logical empiricism' – alternatively 'neopositivism' and 'neoempiricism' – are mainly used synonymously. They refer to a modern version of empiricism, the doctrine that all knowledge about matters of fact stems from and relies on sensory experience and remains hypothetical. Like classical empiricism neoempiricism is 'positive' in a broad sense: it postulates scientific enquiry and rejects metaphysics. In the late nineteenth and the early twentieth centuries, most of modern empiricism was 'positive' also in a narrower sense: all factual knowledge was considered a construction from what is positively given, i.e., from immediate impressions or elements. In the 1930s, this phenomenalist neopositivism has been replaced by a physicalist neoempiricism. Both are called 'logical' because they apply logic for analyzing scientific concepts and clarifying philosophical problems.

Historical Background

Central issues of neoempiricism have their roots in classical empiricism. Empiricism rules out synthetic statements a priori, i.e., necessarily true factual statements. David Hume (1711–76) divided the objects of human enquiry into relations of ideas and matters of fact. Geometry, algebra, and arithmetic deal with relations of ideas. Their statements are intuitively or demonstratively certain, hence a priori. They do not depend on or tell about what exists anywhere in the universe. On the other hand, the contrary of every matter of fact is still possible, because it can never imply a contradiction. As the testimony of our senses is not absolutely reliable, statements about matters of fact are fallible (*An Enquiry Concerning Human Understanding*, Section IV). Hume also proposed a meaning criterion. When we suspect that a philosophical term is employed without any meaning or idea, we need only enquire, from what impression that supposed idea is derived, and if we cannot find any such impression, we will conclude that this term is meaningless (Section II). Hume's discussion of causality showed how to avoid metaphysical hypostatizations of concepts. Causes and effects can be discovered by experience, not by reason. However, no examination of a single object can show that it is a cause or an effect. Therefore, Hume proposed a regularity theory of causation, defining a cause to be "an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second" (Section VII).

Auguste Comte (1798–1857) proposed the terms 'positivism' and 'sociology.' In *Cours de Philosophie Positive*, he maintained that human intellectual development had moved from a theological stage through a metaphysical stage to the positive stage, which relies on scientific method and discovers laws on the basis of relations between phenomena. This holds for natural and social sciences. John Stuart Mill (1806–73) proposed in *A System of Logic, Ratiocinative and Inductive* an empiricist methodology regarding logic and mathematics as the most general empirical sciences, an idea found later in Quine's pragmatism.

Hermann Helmholtz sketched in *Treatise on Physiological Optics* (Helmholtz [1867]1925, [§26, §33]) a realist empiricist epistemology. According to his inference theory of perception we explain our sensations by assuming that they are common effects of external objects and of our nature, but we do not immediately perceive the objects, or our nature, or the circumstance that the objects have an effect on our nature. This realist theory involves the metaphysical assumption of an external world acting on us, which Mach's phenomenalism tries to avoid. All properties that we assign to objects of the external world "may be said to be simply *effects* exerted by them either on our senses or on other natural objects" (1925, 20). As "we see invariably the peculiar sort of interaction occurring, we attribute to the objects a permanent capacity for such effects …. This permanent capacity is … called … *property*" (21). Accordingly, properties are dispositions to lawlike behavior. As we do not perceive the dispositions but only what we take to be their effects, statements that ascribe properties to objects transcend experience. Hence they are fallible.

The physicist and philosopher Ernst Mach (1838–1916) and the philosopher Richard Avenarius (1843–96) proposed a phenomenalist positivism. Avenarius called it empiriocriticism. Lenin criticized it in *Materialism and Empirio-criticism*. Mach rejected Kant's transcendental idealism, in particular the ideas of things-in-themselves and of synthetic a priori knowledge. He considered the atom, which had never been observed, merely a tool for representing phenomena, and he criticized Newton's idea of absolute space and time. Mach's philosophy of science tries to avoid all metaphysical concepts and assumptions. Factual knowledge is considered a conceptual organization of the data of

sensory experience. Hence statements on facts must be empirically verifiable. Mach's phenomenalist construction of the world in *The Analysis of Sensations* ([1886]1897) starts from his contemplation of the spatial boundary of his body. His senses provide findings from outside and inside this boundary. The latter are psychical; the former are physical and can be dissolved into elements like colors and tones. Elements that depend on findings both from outside and from inside are sensations. The complex of what is psychical to Mach is his ego. As the elements of his neighbors are not given to Mach and vice versa, there is no intersubjectively common boundary between the physical and the psychical, rather there are common elements of both districts. Taken by themselves the elements are neither physical nor psychical. This neutral monism avoids the body–mind problem. Obviously Mach is not an ontological but a 'theoretical' solipsist. The objects of the real world are complexes of sensations; hence, there is no need to postulate metaphysical things-in-themselves. Science describes functional interrelationships between complexes of sensations. By learning science we profit from the experience that others have made. Science is a process of adaptation. The most exact and most economical descriptions are most helpful in everyday life.

Gottlob Frege (1848–1925) laid in *Conceptual Notation* ([1879]1972) the foundations of mathematical logic. He introduced variables and functions into logic in an attempt to show that arithmetic is a branch of logic. Concepts are a special class of functions. His program of logicism is worked out in *Basic Laws of Arithmetic* ([1893]1964) where he tries to derive the laws of arithmetic from logical axioms. In 1902, Russell informed Frege that his system implies the antinomy of the set of all sets that do not contain themselves. Frege's work is the first complete example of the logical analytic method in philosophy.

In *The Principles of Mathematics* (1903), Bertrand Russell (1872–1970) stated the philosophical case for logicism. To avoid the antinomy, he proposed the theory of types. In *Principia Mathematica* (1910–13), Russell and Whitehead tried to demonstrate that mathematics is a branch of logic. However, Gödel's first incompleteness theorem (1931) shows that this demonstration is impossible. In *Our Knowledge of the External World* (1914), Russell writes, "the study of logic becomes the central study in philosophy: it gives the method of research in philosophy, just as mathematics gives the method in physics … all this supposed knowledge in the traditional systems must be swept away, and a new beginning must be made" (Section 6). Carnap felt as if this appeal had been directed to him personally.

After reading Frege and Russell, Ludwig Wittgenstein (1889–1951) turned from engineering to logic and mathematics. In *Tractatus Logico-philosophicus* ([1921]1922), he proposed a picture theory of language, according to which a proposition can picture a fact because fact and proposition share a common structure. For each atomic fact there is a corresponding atomic proposition. More complex molecular propositions are constructed from atomic propositions by means of logical connectives. What cannot be pictured cannot be said. The sentences of logic and mathematics do not make statements on the world or on numbers. They do not have linguistic meaning (6.21). They all say the same, i.e., nothing (5.43). According to Wittgenstein philosophy is not a set of theories but an activity. Its task is to clear up confusions arising from misunderstandings of language (4.112). His logical atomism strongly influenced the Vienna Circle. Wittgenstein later rejected logical atomism and argued that the meaning of words depends on their role in language games.

Neoempiricist Circles

From 1907 to 1912, a group that is sometimes called the First Vienna Circle met informally to discuss epistemological and methodological problems of modern physics. It had gathered around the physicist Philipp Frank (1884–1966), the mathematician Hans Hahn (1879–1934), and the economist and sociologist Otto Neurath (1882–1945). The mathematician Richard von Mises (1883–1953) was a corresponding member. They tried to retain the most essential points of Mach's philosophy of science and supplement it by Henri Poincaré's conventionalism and Pierre Duhem's holism. Other fields of their interest were the foundations of mathematics and logistic. Poincaré (1854–1912) had argued that, as we cannot tell whether physical space is Euclidian or non-Euclidian, our choice between geometries is a matter of convention. There is no true scientific hypothesis, rather we choose the hypothesis that is most advantageous (*Science and Hypothesis* [1902]1905). Duhem (1861–1916) had argued that it is impossible to test any physical theory separately. Rather in each case the whole of physics is tested (*The Aim and Structure of Physical Theory* [1906] 1954). In 1912, Frank left Vienna to succeed Einstein in Prague.

The Vienna Circle

In 1922, the physicist and philosopher Moritz Schlick (1882–1936) obtained the chair for the philosophy of the inductive sciences at the University of Vienna previously held by Mach and Boltzmann. A discussion group gathered around his seminar and met regularly since 1923. It was called the 'Schlick Circle' from within and soon the 'Vienna Circle' from outside. Prominent members were Hans Hahn; Otto Neurath; the philosophers Herbert Feigl (1902–88) (who in 1931 proposed the label 'logical positivism'), Friedrich Waismann (1896–1959), Victor Kraft (1880–1975), Felix Kaufmann (1895–1949), and Béla Juhos (1901–71); and the mathematicians Karl Menger (1902–85) and Kurt Gödel (1906–78). Philipp Frank cooperated while in Prague. In 1926, Rudolf Carnap (1891–1970), who had studied mathematics with Frege, joined the circle.

Among the subjects discussed were Einstein's theories of relativity; Wittgenstein's *Tractatus*, which suggests a verifiability criterion of meaning (4.063) and the idea that the sentences of logic and mathematics do not express thoughts; Mach's phenomenalism, which inspired Carnap's *Aufbau*; the character of observation or protocol sentences; and the construction of a metalanguage for the syntactic analysis of scientific languages.

In 1928, the Ernst Mach Society (Verein Ernst Mach) was founded. The choice of its name shows its basic orientation: science free of metaphysics. Originally initiated by the Austrian Freethinker Alliance (Freidenkerbund Österreichs) it became

the platform from which the Vienna Circle addressed a wider public in order to disseminate the findings of the exact sciences and to further modern empiricism and a new organization of economic and social relations. Several members, in particular Neurath, Carnap, and Edgar Zilsel (1891–1944), supported anticlerical and socialist positions. In 1929, the Ernst Mach Society published the booklet *The Scientific Conception of the World: The Vienna Circle* ([1929]1973), which contains a summary of positions common to the members of the Vienna Circle and was dedicated to Schlick. It is the first volume of a series of publications of the Ernst Mach Society. In 1931, Carnap went to Prague, where he and Frank formed a branch of the Vienna Circle, and Feigl emigrated to the United States. In 1934, the Ernst Mach Society was dissolved by the Dollfuss government, Neurath emigrated, and Hahn died. In 1936, Schlick was assassinated and Carnap left Prague for Chicago. The circle continued in a small private discussion group led by Waismann until he emigrated to England in 1938.

The Berlin Society for Empirical Philosophy

In 1928, the physicist Hans Reichenbach (1891–1953), the mathematician Kurt Grelling (1886–1942), and the logician Walter Dubislav (1895–1937) founded in Berlin the Society for Empirical Philosophy (Gesellschaft für empirische Philosophie). Other members were the philosophers Carl Gustav Hempel (1905–97) and Joseph Petzoldt (1862–1929) and the mathematicians David Hilbert (1862–1943) and Richard von Mises. In 1930, Carnap and Reichenbach took over the journal *Annalen der Philosophie* and renamed it *Erkenntnis*. Its new program was to pursue philosophy not as an isolated discipline but in closest connection with the sciences. Reichenbach called this project 'logical empiricism.' On the proposal of Hilbert the Berlin Society was in 1931 renamed 'Society for Scientific Philosophy.'

Together the Vienna Circle and the Berlin Circle organized congresses on epistemology and philosophy of science, some of them preparatory: in Prague 1929; in Königsberg 1930, where Gödel announced his proofs of the completeness of first order logic (his doctoral dissertation of 1929) and of the incompleteness of arithmetic (published in 1931); in Prague 1934; then the international congress for scientific philosophy in Paris 1935; and the international congresses for the unity of science in Copenhagen 1936, in Paris 1937, in Cambridge, UK, 1938, and in Cambridge, USA, in 1939.

The Lvov-Warsaw School

Kazimierz Twardowski (1866–1938) established the Lvov-Warsaw School of logic, which studied the foundations of mathematical logic and analyzed the relationship of language and knowledge. Among its members were his doctoral students Jan Łukasiewicz (1878–1956) who developed the Polonian notation and formalized a system of three valued logic, Stanisław Leśniewski (1886–1939), Tadeusz Kotarbiński (1886–1981), and Kazimierz Ajdukiewicz (1890–1963) who in 1920 developed a deductive theory for the study of logic based on syntax. Alfred Tarski (1901–83), doctoral student of Leśniewski, established scientific semantics. In 1933, he defined the concept of truth in formalized languages. To avoid antinomies like that of the liar he separated object- and metalanguage. In Vienna, Tarski met Gödel and worked with Menger. Tarski's metalogic influenced Carnap. Ajdukiewicz and Kotarbiński also cooperated with the Vienna Circle.

The Scientific Conception of the World

The scientific conception of the world is not a collection of theses; rather it consists of a basic attitude, common points of view, and a direction of research. The task of philosophical work is the clarification of problems, concepts, and assertions, not the production of specifically philosophical theses or even dogmas. The scientific world-conception can best be characterized by two main features. First, it is empiricist and (in 1929) positivist; it expects factual knowledge only from experience (empiricist), which rests on what is immediately given (positivist). Therefore, it rejects all metaphysical philosophy. Second, its method is logical analysis of concepts and statements both of natural and formal languages. This method essentially distinguishes modern empiricism from classical empiricism. Applied to traditional philosophical problems it will either unmask them as pseudoproblems or transform them into empirical problems.

Unified Science

The goal of the scientific world-conception is unified science. In 1938, a series of publications started, the *International Encyclopedia of Unified Science*, which was never completed. Nineteen monographs appeared from 1938 to 1969 and were finally collected in two volumes. Unified science may be characterized by two requirements. The weaker one is all sciences apply empirical methods and use the physicalist language or thing language (which is not identical with the language of physics). The stronger requirement is the laws of all sciences are derivable from physical laws. This program is called reductionism. Since 1930, parts of medicine have been reduced to biology, parts of biology to chemistry, and parts of chemistry to physics. As long as sciences produce new theories, the reduction of all sciences to physics will never be complete, and neoempiricists cannot want that science should ever come to a standstill just to complete unified science. Therefore, they will have to be content with sporadic partial reductions. However, most criticism of the idea of unified science is directed against the weaker requirement, because it rules out scientific statements on specifically mental entities (as e.g., minds that are independent of brains) and specific nonempirical methods postulated in the humanities (as e.g., Verstehen [understanding], introspection, or intuition).

The Kernel of Neopositivism

Carnap's *Der logische Aufbau der Welt* ([1928]2003, *The Logical Structure of the World*) is a crystallization of logical positivism, its most influential book. It is empiricist, phenomenalist (positivist), and antimetaphysical. Frege and Russell had attempted to reduce all logical and mathematical concepts to a few fundamental concepts. In the *Aufbau* Carnap tries to sketch how the same can be done for empirical concepts.

A system of concepts could be built on a physicalist basis, but he chooses a phenomenalist basis, because it represents not only logical but also epistemological relations among the concepts. The choice of a language is a matter of convenience. This 'principle of tolerance' is stated in Carnap's second important book *The Logical Syntax of Language* ([1934] 1937). By the logical syntax of a language he means the systematic statement of the formal rules that govern it together with the development of the consequences that follow from the rules.

In the *Aufbau* Carnap defends, like Mach and Russell, the positivist thesis that – in principle – it is possible to reduce all scientific concepts stepwise to (concepts referring to) the immediately given. While Mach lacked the logical theory of relations developed in *Principia Mathematica* and Russell left it in *Our Knowledge of the External World* (1914) at a very readable informal sketch, Carnap uses logic to design a 'constitutive theory' and work it out in some detail. Its main thesis says that it is possible to deduce all scientific concepts by means of 'constitutional definitions' from a few fundamental concepts. When this logical analysis is completed for all concepts, they can be ordered into a 'constitutive system.' At its basis, concepts refer directly to the given; in the next layer they refer to physical objects, then follow other minds and finally the objects of social science. Instead of saying that concepts are constituted he could as well say that objects are constituted. This is merely another mode of speaking.

The Gestalt psychology of Wertheimer, Köhler, and Koffka had shown that simple sense data result from a process of abstraction, while instantaneous visual fields may be given as units. Therefore, Carnap chooses total instantaneous experiences (Elementarerlebnisse) as fundamental elements. They are momentary cross sections of the total stream of experience. At this level the ego has not yet been constituted. Still subjectless fundamental elements are individual (eigenpsychisch). Hence the basis of the system is solipsistic, but this is a methodic, not an ontological solipsism. Carnap's sole fundamental relation is a similarity relation between fundamental elements. In his exposition, a process of definitions starting from recollections of similarities (Ähnlichkeitserinnerungen) between pairs of visual fundamental elements leads first to the visual domain, then to positions in the visual field, to colors, the temporal order, etc. Carnap calls this process 'quasi-analysis.' Next, things in the perceptual space are constructed, among them the observer's body and other persons' bodies. The class of all individual experiences of the observer is his 'ego.' Then follow other minds and finally cultural products. Thus any scientific statement could – in principle – be transformed into a statement on similarity relations between total instantaneous experiences. This construction avoids metaphysical hypostatizations of physical or mental entities. Although statements on bodies and minds remain possible, the body–mind problem has been dissolved.

Later Carnap admitted that a definitional reduction of concepts is not possible and that the extensional language that he proposed is not sufficient. In his article 'Die physikalische Sprache als Universalsprache der Wissenschaft' ([1932a], 1934a, literally: The physicalist language as a universal language of science) Carnap prefers the physicalist language. Now phenomenalist neopositivism has come to an end, but physicalist neoempiricism lives on.

The Basic Thesis of Empiricism

In *General Theory of Knowledge* ([1918]1985), Schlick rejects both Kant's thing-in-itself and Mach's phenomenalism. Instead he proposes a critical realism. He also gives a thorough critique of Kant's synthetic a priori. While Kant believes "When … strict universality belongs to a judgment essentially; this points to a special source of cognition for it, namely a faculty of a priori cognition" (*Critique of Pure Reason*, B4), empiricists take the notion that thinking can lead to factual knowledge out of its own resources without using any empirical material to be a basic error of metaphysics. Only the formal (analytic) truths of logic and mathematics can be known a priori, all factual (synthetic) knowledge is a posteriori. This is the basic thesis of empiricism. In Einstein's words: "As far as the laws of geometry refer to reality, they are not certain; and as far as they are certain, they do not refer to reality."

The Verifiability Criterion of Meaning

The best known proposal of logical positivism is the verifiability criterion of meaning (or meaningfulness). Only if we know what a statement means, can we assess whether it is true. However, the verifiability criterion implies that we can only know what it means, if we can assess whether it is true. Its best known wording by Waismann (1930, 229, my translation) reads as follows: "If it can in no way be indicated when a statement is true, then this statement does not have any meaning at all; for the meaning of a statement is the method of its verification. … A statement which cannot be definitely verified is not verifiable at all; it just lacks any meaning." Waismann's wording presupposes that statements about the empirically given can be definitely verified or are for some reason certain and therefore do not need verification. This presupposition is subject of the debate on protocol sentences.

Serious problems of the verifiability criterion of meaning were evident. First, we could not answer questions, if they had no linguistic meaning. As they are neither true nor false, they cannot be verified. If, however, questions can be meaningful without being verifiable, why can statements not be meaningful without being verifiable? Second, universal statements like 'All ravens are black' cannot be definitely verified because of the problem of induction. Accordingly they are considered meaningless. However, we use them in everyday life, and classical physics is deterministic.

To solve the problem of universal statements, various proposals have been made. They all have their specific defects, and they all have a common defect: not even singular observation statements can be definitely verified, for the properties that they attribute to objects are dispositions to lawlike behavior, but we observe only instances of this behavior. Hence according to the verifiability criterion of meaning there are no meaningful statements at all. Karl Popper (1902–94) thoroughly criticized these proposals in a manuscript that was in 1933 accepted for publication by Schlick and Frank but was rejected by the publisher Springer because of its length. An incomplete edition appeared only in 1979. What Helmholtz had said about dispositions Popper said about disposition predicates: the statement "Here is a glass of water" cannot be

verified, because the universals "… is glass" and "… is water" express lawlike behavior ([1935]1959a, 95), dispositions, and each of these laws transcends experience (424). The verifiability criterion of meaning was soon given up when Popper proposed in his *Logic of Scientific Discovery* a falsifiability (later testability) criterion of demarcation between empirical science and metaphysics, according to which those, and only those, synthetic statements are considered to be empirically scientific that are empirically falsifiable ([1933]1959b, sect. 6). Metaphysical statements may be meaningful but are not scientific.

The Debate on Protocol Sentences

The debate on protocol sentences brought another radical change. Its most important arguments were exchanged in three articles. According to Carnap's 'Die physikalische Sprache als Universalsprache der Wissenschaft' ([1932]1934), protocol sentences like 'here, now, blue; there, red' refer to the given, need no justification, and serve as foundation for all other statements of science. Neurath objects in 'Protocol Sentences' ([1932]1959) that there are no conclusively established protocol sentences. His protocol sentences refer to physical objects and contain the name of a person drafting the protocol, an indication of place and time, and a perception term. They are subject to verification and may be changed or even discarded. In his rejoinder 'On Protocol Sentences' ([1932] 1987) Carnap argues that both phenomenalist and physicalist protocol sentences are possible, but he now prefers an intersubjective, physicalist language. He rejects Neurath's proposal to admit only protocol sentences of a certain form in favor of Popper's proposal that every concrete sentence of the physicalist language of science can – under certain circumstances – serve as a protocol sentence. All sentences are fallible and must be empirically tested. As there are no absolute initial sentences of science left, Popper's testing procedure (in [1933] 1959b) involves, according to Carnap, the most radical elimination of the absolutism of the given. This ends the debate on protocol sentences, while the rise of Popper's critical rationalism, an advanced version of neoempiricism, begins.

The quest for certainty is not yet over. Schlick never was a phenomenalist, but he still clings to the absolutism of the given. According to him, understanding and verifying coincide in the case of 'confirmations' (Konstatierungen) like 'Here now so and so' ([1933]1959, 221, 224). 'Confirmations' are absolutely true and certain. On the other hand, they are only subjective, and their function lies only in the immediate present. Accordingly, they play only a pragmatic part in the evaluation of scientific hypotheses.

The Duhem-Quine Thesis

The Duhem-Quine thesis is the most serious attack on neoempiricism. According to Willard Van Orman Quine (1908–2000), modern empiricism has been conditioned in large part by two dogmas (1951). One is the analytic/synthetic distinction; the other is reductionism, the belief that each meaningful statement is translatable into a statement about immediate experience. Both are ill-founded. If they are given up, the border between science and metaphysics falls and the philosophy of science shifts toward pragmatism.

Quine calls a statement 'analytic' if it is true by virtue of meaning alone. Now there is no precise meaning of 'meaning.' Moreover, we cannot in all languages give the word 'analytic' a precise meaning. Therefore, he concludes, the analytic/synthetic distinction is not tenable. However, Quine's argument is not conclusive. Tautologies have no meaning; they are analytic because of their form alone, explicit definitions create new synonymies, any author can state which of his sentences are meant to be linguistic conventions, and it would be sufficient to give the word 'analytic' a precise meaning only for that language whose statements we are actually examining.

According to Quine the reductionism of Carnap's *Aufbau* survives in the idea that statements can be tested separately. While Duhem had argued that we never test a single physical theory but always test the whole of physics, Quine argues that we always test the whole of our knowledge including pure mathematics and logic. We can hold true any statement come what may, if we make drastic enough adjustments elsewhere in the system of our knowledge, even among logical laws. This argument is even less conclusive. Laws of logic and mathematics are true or false according to their form, while laws of physics fit or do not fit reality. Only if we had reasons to accept Quine's (and Mill's) pragmatist or instrumentalist interpretation of both kinds of laws, could it seem reasonable to reject the analytic/synthetic distinction. Even though not conclusive the holism of the Duhem-Quine thesis is the best known trigger of the pragmatist turn in analytic philosophy.

Prospects

Phenomenalist neopositivism ended in the 1930s. Neoempiricists have not only given up the idea that metaphysical statements are literally meaningless but also admit that there is no sharp boundary between science and metaphysics. They recognize that most scientific theories are not from the beginning empirically testable but start as metaphysical hypotheses in the sense of Popper's criterion of demarcation and become scientific, i.e., empirically testable, only after some development of the theories themselves or of the means of testing them. Before the thesis of the Higgs-boson's existence – possibly – became testable, scientists had to specify which traces it would leave and the Large Hadron Collider had to be built.

There is still a difference between scientific speculation, which aims at testable hypotheses, and philosophical speculation, whose results never become intersubjectively testable, and there are differences between philosophical speculations. In his 1929 lecture '*Was ist Metaphysik?*' Martin Heidegger pronounced that the nothing (or nothingness) nothings and that this activity is the origin of negation. Carnap diagnosed that Heidegger had turned the negation 'nicht' (not) both into a substantive 'das Nichts' (the nothing or nothingness) and into a verb 'nichten' (to nothing) and that his pronouncement was therefore nonsensical. Immanuel Kant asserted in *Critique of Pure Reason* that (Euclidian) space is an a priori intuition, a necessary presupposition of empirical knowledge. However, according to Einstein's general theory of relativity physical space is elliptic and visual space is taken to be by and large hyperbolic. Does this mean that what Kant considered a priori necessarily true has turned out a posteriori to be factually false?

Did Kant mean a third kind of space, and what can be said about it other than that it is Euclidian? Kant's theses on the contribution of categories and of the a priori intuitions of space and time to our knowledge filled a gap that phenomenalist empiricism had left. Insofar they furthered the evolution of epistemology, even though they turned out to be untenable.

Neoempiricists will rather trust in Helmholtz's critical realism than in Kant's transcendental idealism and learn from psychology of perception, neurophysiology, and – possibly – electronic modeling of brain processes. As these disciplines develop, so will empiricism. Again and again new aspects of modern empiricism will be debated. On the other hand, two issues have long been disputed, sometimes bitterly: the methodological unity of science and the possibility of cognitive ethics.

There is empirical research in the humanities, but there are still authors claiming to know methods peculiar to the humanities. Neoempiricists will accept any way to discovery, whether it is methodical or not, but they will insist that all results remain hypothetical and must in the end be intersubjectively tested, i.e., by means of our physiological senses. Insofar they uphold the idea of the methodological unity of science. Their opponents will insist that they know methods of evaluating results of the humanities that do not involve sensory experience but are nevertheless intersubjective among the knowing, and part of them will even call these methods empirical because they involve a different kind of experience.

Since antiquity, countless attempts have been made to show that there is ethical knowledge, that we can know what is objectively virtuous, good, or right. Up to now, none of them has withstood serious criticism. Nevertheless, the claim to possess ethical knowledge, religious or temporal, has proved to be an efficient means of influencing people. Open criticism tends to reduce its efficiency and is therefore much resented. In the dispute with the skeptic a derogatory label can be an efficient weapon. So the skeptical, neoempiricist is often called a 'positivist,' suggesting that his judgment is unreliable because he still maintains a position that has long been refuted. Others may call him an 'infidel.' Probably, the dispute about ethical knowledge will never end.

See also: Critical Rationalism; Knowledge (Explicit, Implicit and Tacit): Philosophical Aspects; Methodological Individualism: Philosophical Aspects; Objectivity: Philosophical Aspects; Physicalism and Alternatives; Pragmatism: Philosophical Aspects; Relativism: Philosophical Aspects; Scientific Explanation; Social Science, The Idea of.

Bibliography

Carnap, R., 1928. 2003. Der logische Aufbau der Welt. Weltkreis-Verlag: Berlin-Schlachtensee. Trans. The Logical Structure of the World and Pseudoproblems in Philosophy. Open Court, Chicago.

Carnap, R., 1932a. 1934a. Die physikalische Sprache als Universalsprache der Wissenschaft. Erkenntnis 2, 432–465. Trans. The Unity of Science. Kegan Paul: London.

Carnap, R., 1932b. 1987. Über Protokollsätze. Erkenntnis 3, 215–228. Trans. On protocol sentences. Noûs XXI, 457–470.

Carnap, R., 1934b. 1937. Logische Syntax der Sprache. Julius Springer, Wien. Trans. The Logical Syntax of Language. Kegan Paul, London.

Duhem, P., 1906. 1954. La théorie physique, son objet et sa structure. Chevalier & Rivière, Paris. Transl. by P. Wiener from the Second French Edition 1914. The Aim and Structure of Physical Theory. Princeton University Press, Princeton.

Frege, G., 1879. 1972. Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens. Halle: Louis Nebert; Bynum, T. W. (trans. and ed.), Conceptual notation and related articles. Oxford UP, Oxford.

Frege, G., 1893. 1964. Grundgesetze der Arithmetik, begriffsschriftlich abgeleitet. Jena, vol. 1 1893, vol.2 1903. In: Hermann, Pohle, partial translation by Furth, M. (Eds.), The basic laws of arithmetic. University of California Press, Berkeley.

Helmholtz, H., 1867. 1925. Handbuch der physiologischen Optik. Leopold Voss, Leipzig. Translated from the Third German Edition [1910]. In: Southall, J.P.C. (Ed.), Helmholtz's Treatise on Physiological Optics. George Banta Publishing Company, Menasha, Wisconsin published by the Optical Society of America, vol. III.

Mach, E., 1886. 1897. Die Analyse der Empfindungen und das Verhältnis des Physischen zum Psychischen. G. Fischer, Leipzig. Trans. In: Williams, C.M. (Ed.), Contributions to the Analysis of the Sensations. Open Court, Chicago.

Neurath, O., 1932. 1959. Protokollsätze. Erkenntnis 3, 204–214. Trans. In: Ayer, A.J. (Ed.), Protocol Sentences. The Free Press, New York, pp. 199–208.

Poincaré, H., 1902. 1905. La science et l'hypothèse. Flammarion, Paris. Trans. Science and Hypothesis. Walter Scott, London.

Popper, K.R., 1933. 1959b. Ein Kriterium des empirischen Charakters theoretischer Systeme Erkenntnis 3, 426-427. Trans. In: Popper, K.R. (Ed.), A Criterion of the Empirical Character of Theoretical Systems (appendix *i).

Popper, K.R., 1935. 1959a. Logik der Forschung, Wien: Springer. Trans. The logic of scientific discovery. Basic Books, New York.

Popper, K.R., 1979. Die beiden Grundprobleme der Erkenntnistheorie. J. C. B. Mohr (Paul Siebeck), Tübingen.

Quine, W.V.O., 1951. Two dogmas of empiricism. The Philosophical Review 60, 20–43.

Russell, B., 1903. The Principles of Mathematics. Cambridge UP, Cambridge UK.

Russell, B., 1914. Our Knowledge of the External World as a Field for Scientific Method in Philosophy. Open Court, Chicago and London.

Russell, B., Whitehead, A.N., 1910–1913. Principia Mathematica, 3 vols. Cambridge UP, Cambridge UK.

Schlick, M., 1933. 1959. Über das Fundament der Erkenntnis. Erkenntnis 4, 77 -99. Trans. In: Ayer, A.J. (Ed.), The foundation of knowledge. The Free Press, New York, pp. 209–227.

Ernst Mach, Verein, 1929. 1973. Wissenschaftliche Weltauffassung: Der Wiener Kreis. Artur Wolf, Wien. Trans. The Scientific Conception of the World: The Vienna Circle. D. Reidel, Dordrecht.

Waismann, F., 1930. Logische Analyse des Wahrscheinlichkeitsbegriffs. Erkenntnis 1, 228–248.

Wittgenstein, L., 1921. 1922. Logisch-philosophische Abhandlung. Ostwalds Annalen der Naturphilosophie 14. Trans. Tractatus Logico-philosophicus. Kegan Paul, London.

Relevant Websites

http://www.univie.ac.at/ivc/ and http://wienerkreis.univie.ac.at/ – Institute Vienna Circle; Institut Wiener Kreis. Verein zur Förderung wissenschaftlicher Weltauffassung, Universität Wien.

http://viennacirclefoundation.nl/ – Vienna Circle Foundation, Amsterdam.

http://www.austrian-philosophy.at – Forschungsstelle und Dokumentationszentrum für österreichische Philosophie (FDÖP), Graz.

http://www.library.pitt.edu/libraries/special/asp/archive.html – University of Pittsburgh, Archives of Scientific Philosophy.

http://ww.uni-klu.ac.at – Karl Popper Foundation, Klagenfurt.

http://kgs.logic.at/ – The Kurt Gödel Society, Vienna.

http://www.uni-konstanz.de – Philosophisches Archiv der Universität Konstanz.

http://www.moritz-schlick.de – Moritz-Schlick-Forschungsstelle, Rostock.

Logics for Knowledge Representation

Bernhard Nebel, Institut für Informatik, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Knowledge representation (*see* Knowledge Representation) and reasoning plays a central role in Artificial Intelligence. Research in Artificial Intelligence (henceforth AI) started off by trying to identify the *general mechanisms* responsible for intelligent behavior. However, it quickly became obvious that general and powerful methods are not enough to get the desired result, namely, intelligent behavior. Almost all tasks a human can perform which are considered to require intelligence are also based on a huge amount of knowledge. For instance, understanding and producing natural language heavily relies on knowledge about the language, about the structure of the world, about social relationships, and so on (*see* Logic and Linguistics).

One way to address the problem of representing knowledge and reasoning about it is to use some form of logic. While this seems to be a natural choice, it took a while before this 'logical point of view' became the prevalent approach in the area of knowledge representation. Below, we will give a brief sketch of how the field of knowledge representation evolved and what kind of logical methods have been used. In particular, we will argue that the important point about using formal logic is the *logical method*.

Logic-Based Knowledge Representation: A Historical Account

McCarthy (1968) stated very early on that mathematical, formal logic appears to be a promising tool for achieving human-level intelligence on computers. However, in the early days of AI, there were also a number of researchers with a completely different opinion. Minsky (1975), for example, argued that knowledge representation formalisms should be flexible and informal. Moreover, he argued that the logical notions of *correctness* and *completeness* are inappropriate in a knowledge representation context.

While in those days heated arguments of the suitability of logic were exchanged, by the end of the 1980s, the logical perspective seem to have gained the upper hand (Brachman, 1990). Since the 1990s almost all research in the area of knowledge representation and reasoning was based on formal, logical methods as demonstrated by the papers published in the biannual international conference on *Principles of Knowledge Representation and Reasoning*, which started in 1989.

It should be noted, however, that two perspectives on logic are possible. The first perspective, taken by McCarthy (1968), is that logic should be used to *represent* knowledge. That is, we use logic as the representational and reasoning tool inside the computer. Newell (1982) on the other hand proposed in his seminal paper on the *knowledge level* to use logic as a formal tool to *analyze* knowledge. Of course, these two views are not incompatible. Furthermore, once we accept that formal logic should be used as a tool for analyzing knowledge, this could also lead to using logic for representing knowledge and for reasoning about it as well.

Knowledge Representation Formalisms and Their Semantics

Saying that logic is used as the main formal tool does not say which kind of logic is used. In fact, a large variety of logics (Gabbay et al., 1995) have been employed or developed to solve knowledge representation and reasoning problems. Often, one started with a vaguely specified problem, developed some kind of knowledge representation formalism without a formal semantics, and only later started to provide a formal semantics. Using this semantics, one could then analyze the complexity of the reasoning problems and develop sound and complete reasoning algorithms (*see* Algorithm). I will call this the *logical method*, which proved to be very fruitful in the past and has a lot of potential for the future.

One good example for the evolution of knowledge representation formalisms is the development of *description logics*, which have their roots in so-called *structured inheritance networks* formalisms such as KL-ONE (Brachman, 1979). These networks were originally developed in order to represent word meanings. A *concept node* connects to other concept nodes using *roles*. Moreover, the roles could be structured as well. These networks permit, for example, the definition of the concept of a bachelor.

Later on, these structured inheritance networks were formalized as so-called *concept languages*, *terminological logics*, or *description logics*. Concepts were interpreted as unary predicates, roles as binary relations, and the connections between nodes as so-called *value restrictions*. This leads for most such description logics to a particular fragment of *first-order predicate logic*, namely, the fragment L_2. In this fragment only two different variable symbols are used. As it turns out, this is a *decidable fragment* of first-order logic.

However, some of the more involved description logics go beyond L_2. They contain, for example, relational composition or transitive closure. As it turns out, such description logics can be understood as variants of multimodal logics (Schild, 1991), and decidability and complexity results from these multimodal logics carry over to the description logics.

All these insights, that is, determination of decidability and complexity as well as the design of decision algorithms, are based on the rigorous formalization of the initial ideas (Baader et al., 2003). In particular, it is not just one logic that is used to derive these results, but it is the *logical method* that

led to the success. One starts with a specification of how expressions of the language or formalism have to be interpreted in formal terms. Based on that one can specify when a set of formulae is *satisfiable*, that is, describes structures. Based on that, *decidability* and *complexity* (*see* Algorithm) – perhaps of fragments of the languages – can be determined. In parallel, one can design inference methods or decision procedures that enable us to determine *satisfiability* of a formula on purely syntactic grounds. Finally, one can then prove these methods to be *correct* and *complete*.

Based on these formal foundations, description logics have made it into the 'real life.' Meanwhile they have become the formal foundation for the *semantic web* and the *web ontology language* (Antoniou and van Harmelen, 2004).

Nonmonotonic Logics

Another interesting area where the logical method has been applied is the development of so-called *nonmonotonic logics* (*see* Nonstandard Reasoning). These are based on the intuition that sometimes a logical consequence should be retracted if new evidence becomes known. For example, we may assume that our car will not be moved by somebody else after we have parked it. However, if new information becomes known, such as the fact that the car is not at the place where we have parked it, we are ready to drop the assumption that our car has not been moved.

This general reasoning pattern was used quite regularly in early AI systems, but it took a while before it was analyzed from a logical point of view. In 1980, a special issue of the *Artificial Intelligence* journal appeared, presenting different approaches to nonmonotonic reasoning, in particular Reiter's (1980) *default logic* and McCarthy's (1980) *circumscription* approach.

A disappointing fact about nonmonotonic logics appeared to be that it is very difficult to formalize a domain such that one gets the intended conclusions. In particular, in the area of *reasoning about actions*, McDermott (1987) has demonstrated that the straightforward formalization of an easy temporal projection problem (the 'Yale shooting problem') does not lead to the desired consequences. However, it is possible to get around this problem. Meanwhile, the formalization of dynamic domains is based either on an approach using the *situation calculus* using so-called *successor-state-axioms* (Reiter, 2001) or they are based on *action languages* using the semantics of state transition systems (Gelfond and Lifschitz, 1993).

It took more than a decade before people started to analyze the computational complexity (of the propositional versions) of these logics. As it turned out, these logics are usually somewhat more difficult than ordinary propositional logic (Gottlob, 1992). This, however, seems tolerable since we get much more conclusions than in standard propositional logic. At the same time, the tight connection between nonmonotonic logic and *belief revision* (Gardenfors, 1988) was noticed. Belief revision – modeling the evolution of beliefs over time – is just one way to describe how the set of nonmonotonic consequences evolve over time, which leads to a very tight connection on the formal level for these two forms of nonmonotonicity (Nebel, 1991). Again, all these results and insights are mainly based on the logical method to knowledge representation.

One of the important results on the research of nonmonotonic logics is *answer set programming* (ASP) (Martin Gebser et al., 2012), which is based on the stable model semantics of *logic programming*. This kind of logic programming allows for straightforward formulations of many *search and optimization problems* and is very efficient in computing the solutions to such problems in practice.

Qualitative Spatial and Temporal Representation and Reasoning

The formalization of knowledge about qualitative aspects of space and time is an area that reaches from philosophical logic over AI to psychology. Again, the logical method is prevalent here, as, for example, demonstrated by the handbook on spatial logics (Aiello et al., 2007).

The development of this area started probably with Allen's (1983) paper about the formalization of qualitative interval relations. With this formalization one could express the relative position of one-dimensional objects, which were interpreted as time intervals. This so-called *Allen calculus* was intensively analyzed, in particular from a computational complexity point of view (Nebel and Bürckert, 1995), and led to the development of a number of similar calculi for different aspects of spatial relations, for example, the RCC-8 calculus (Randell et al., 1992). For most of these calculi symbolic constraint-propagation methods (*see* Constraint-Satisfaction Models) are employed to address the above-mentioned satisfiability problem (Renz and Nebel, 2007). However, as it turns out, these inference methods are often not complete.

Using the Logical Method in Cognitive Science

While in general Artificial Intelligence and Cognitive Science have a lot in common, the general research goals are usually quite distinct. Artificial Intelligence focuses on synthesizing intelligent systems, Cognitive Science concentrates on analyzing human cognition. Nevertheless, there are often points where these two disciplines can help each other.

In particular in the area of *spatial cognition*, the logical analysis of temporal and spatial concepts was quite successful. In particular, Allen's temporal relations interpreted as one-dimensional spatial relations were in the focus of attention. Based on an analysis of how many possible configurations arise when composing two interval relations, experiments showed that humans systematically preferred only one particular subset of the possible relations. Here the logical formalization was used to provide the normative result against which the observed behavior could be compared. Furthermore, systematic deviations could be correlated to the number of possible configurations and to the subset of what has been termed preferred relations (Knauff et al., 1998). Similar results were achieved with respect to RCC-8 (Renz et al., 2000).

Outlook

The above description of the use of logics for knowledge representation is necessarily incomplete. Nevertheless, one should have got an idea of how logics are used in the area of

knowledge representation. As mentioned, it is the idea of providing knowledge representation formalisms with formal (logical) semantics that enables us to communicate their meaning, to analyze their formal properties, to determine their computational complexity, and to devise reasoning algorithms.

While the research area of knowledge representation is dominated by the logical approach, this does not mean that all approaches to knowledge representation must be based on logic. Probabilistic, decision theoretic, and game theoretic approaches have become quite popular. However, even here, the *logical method* prevails. For example, Lutz and Schröder (2010) combine in their approach description logics with probability theory to model subjective uncertainty.

See also: Algorithm; Algorithmic Complexity; Artificial Intelligence: Connectionist and Symbolic Approaches; Artificial Intelligence: Genetic Programming; Cognitive Neuroscience; Constraint-Satisfaction Models; Knowledge Representation; Logic and Linguistics; Nonstandard Reasoning; Psycholinguistics: Overview.

Bibliography

Aiello, M., Pratt-Hartmann, I., van Benthem, J. (Eds.), 2007. Handbook of Spatial Logics. Springer-Verlag, Berlin, Heidelberg, New York.

Allen, J.F., 1983. Maintaining knowledge about temporal intervals. Communications of the ACM 26 (11), 832–843.

Antoniou, G., van Harmelen, F., 2004. Web ontology language: OWL. In: Staab, S., Studer, R. (Eds.), Handbook on Ontologies, International Handbooks on Information Systems. Springer, pp. 67–92.

Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (Eds.), 2003. The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press.

Brachman, R.J., 1979. On the epistemological status of semantic networks. In: Findler, N.V. (Ed.), Associative Networks: Representation and Use of Knowledge by Computers. Academic Press, New York, NY, pp. 3–50.

Brachman, R.J., 1990. The future of knowledge representation. In: Proceedings of the 8th National Conference of the American Association for Artificial Intelligence (AAAI-90). MIT Press, Boston, MA, pp. 1082–1092.

Gabbay, D.M., Hogger, C.J., Robinson, J.A. (Eds.), 1995. Handbook of Logic in Artificial Intelligence and Logic Programming, vols. 1–5. Oxford University Press, Oxford, UK.

Gardenfors, P., 1988. Knowledge in Flux – Modeling the Dynamics of Epistemic States. MIT Press, Cambridge, MA.

Gelfond, M., Lifschitz, V., 1993. Representing action and change by logic programs. Journal of Logic Programming 17 (2–4), 301–321.

Gottlob, G., 1992. Complexity results for nonmonotonic logics. Journal for Logic and Computation 2 (3), 397–425.

Gebser, Martin, Kaminski, Roland, Kaufmann, Benjamin, Schaub, T., 2012. Answer Set Solving in Practice, Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool, San Rafael, CA.

Knauff, M., Rauh, R., Schlieder, C., Strube, G., 1998. Mental Models in Spatial Reasoning, Spatial Cognition – An Interdisciplinary Approach to Representation and Processing of Spatial Knowledge. In: Lecture Notes in Artificial Intelligence, vol. 1404. Springer-Verlag, Berlin; Heidelberg; New York, pp. 267–291.

Lutz, C., Schröder, L., 2010. Probabilistic description logics for subjective uncertainty. In: Lin, F., Sattler, U., Truszczynski, M. (Eds.), Principles of Knowledge Representation and Reasoning: Proceedings of the 12th International Conference (KR-10). AAAI Press, Toronto, ON.

McCarthy, J., 1968. Programs with common sense. In: Minsky, M. (Ed.), Semantic Information Processing. MIT Press, Cambridge, MA, pp. 403–418.

McCarthy, J., 1980. Circumscription – a form of non-monotonic reasoning. Artificial Intelligence 13 (1–2), 27–39.

McDermott, D.V., 1987. A critique of pure reason. Computational Intelligence 3 (3), 151–160.

Minsky, M., 1975. A framework for representing knowledge. In: Winston, P. (Ed.), The Psychology of Computer Vision. McGraw-Hill, New York, NY, pp. 211–277.

Nebel, B., Bürckert, H.-J., 1995. Reasoning about temporal relations: a maximal tractable subclass of Allen's interval algebra. Journal of the Association for Computing Machinery 42 (1), 43–66. URL: ftp://ftp.informatik.uni-freiburg.de/papers/ki/nebel-burckert-jacm95.ps.gz.

Nebel, B., 1991. Belief revision and default reasoning: syntax-based approaches. In: Allen, J.A., Fikes, R., Sandewall, E. (Eds.), Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference (KR-91). Morgan Kaufmann, Cambridge, MA, pp. 417–428.

Newell, A., 1982. The knowledge level. Artificial Intelligence 18 (1), 87–127.

Randell, D.A., Cui, Z., Cohn, A.G., 1992. A spatial logic based on regions and connection. In: Nebel, B., Swartout, W., Rich, C. (Eds.), Principles of Knowledge Representation and Reasoning: Proceedings of the 3rd International Conference (KR-92). Morgan Kaufmann, Cambridge, MA, pp. 165–176.

Reiter, R., 1980. A logic for default reasoning. Artificial Intelligence 13 (1), 81–132.

Reiter, R., 2001. Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems. MIT Press, Cambridge, MA.

Renz, J., Nebel, B., 2007. Qualitative spatial reasoning using constraint calculi. In: Aiello, M., et al. (Eds.), pp. 161–215.

Renz, J., Rauh, R., Knauff, M., 2000. Towards cognitive adequacy of topological spatial relations. In: Freksa, C., Brauer, W., Habel, C., Wender, K.F. (Eds.), Spatial Cognition II – Integrating Abstract Theories, Empirical Studies, Formal Models, and Practical Applications, Lecture Notes in Computer Science, vol. 1849. Springer-Verlag, Berlin; Heidelberg; New York.

Schild, K., 1991. A correspondence theory for terminological logics: preliminary report. In: Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-91). Morgan Kaufmann, Sydney, Australia, pp. 466–471.

Log-Linear and Marginal Models

Tamás Rudas, Faculty of Social Sciences, Eötvös Loránd University, Budapest, Hungary

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Log-linear models for the joint distribution of several categorical variables assume that the cell frequencies or probabilities are influenced by multiplicative effects associated with groups of variables. These effects are conditional on the remaining variables in the analysis. Marginal log-linear models provide more flexibility, by allowing some of the effects to depend only on subsets of variables. These models have important applications in social mobility, longitudinal data or causal analysis, and are the basis of doing regression analysis and graphical modeling for categorical variables. This article reviews several of the reasons why log-linear models are considered, discusses the interpretation of such models and of the related parameters, introduces marginal log-linear models, and describes several of the areas of application.

Log-Linear Models

Any model class used in statistical analysis aims to provide simple structure to ease interpretation and flexibility to provide good fit to data. While model fit to data is usually considered a technical issue (although there are some ad-hoc suggestions), the perception of simplicity is often influenced by conventions, education, and trends. The class of log-linear models may be looked at in a number of different ways, each providing a different aspect of simplicity, and a possibility for interpretation.

Log-linear models apply to the joint analysis of several categorical (discrete) variables, where the data take the form of observed counts, usually presented as entries in the cells of a contingency table. The goal of the analysis, broadly speaking, is to find structure behind the population probabilities or population frequencies.

For many researchers, linear (more precisely: additive) effects appear simple, but these are difficult to confine to (0, 1), where probabilities are, so one possible idea is to assume that the logarithms of the cell probabilities are sums of certain effects. For example, if p_{ij} is the probability in the cell defined as the combination of the ith category of the first variable and the jth category of the second variable, if there are two variables only, one may assume that

$$log\, p_{ij} = a + b_i + c_j \qquad [1]$$

where a is the overall effect, present in every cell, b_i is the effect of the ith category of the first variable (often called the effect of the ith row of the contingency table) and c_j is the effect of the jth category of the second variable (often thought of as the effect of the jth column of the contingency table). The formula in [1] is analogous to the model of no interaction in a two-way analysis of variance. The formula in [1] can be easily rewritten as

$$log\, p_{ij} = d_i + e_j \qquad [2]$$

and the model remains the same. One might think that a more restrictive assumption is that the row and column effects are linked to the row and column marginal probabilities as

$$d_i = log\, p_{i+},\ e_j = log\, p_{+j}$$

giving

$$log\, p_{ij} = log\, p_{i+} + log\, p_{+j} \qquad [3]$$

which, after exponentiating gives

$$p_{ij} = p_{i+}p_{+j} \qquad [4]$$

that is, the well known independence model. In fact, [3] or [4] which are equivalent, if the probabilities are assumed to be positive, are not more restrictive than [1] or [2], rather they are equivalent.

Formula [4] may also be seen as linking the joint distribution p_{ij} to its own one-dimensional marginal distributions. One may also ask the following question. Out of all the two-dimensional distributions having the one-way marginal distributions p_{i+} and p_{+j}, which one is the closest to uniform, that is, which one contains the least amount of information, in addition to the one-way marginals. The answer is the distribution of the form [4].

It is immediate from [4], that the odds ratio (for simplicity of presentation, thinking only of the case when both variables are binary) is also equal to 1:

$$OR = \frac{p_{11}p_{22}}{p_{12}p_{21}} = 1 \qquad [5]$$

and it can be seen that [5] also implies [4].

These are four different aspects (see Goodman, 1970; Bishop et al., 1975; Haberman, 1978; Gokhale and Kullback, 1978; Rudas, 1998), which all point to the same model, and the interpretations of the log-linear model implied by these properties, extend to the general log-linear model.

In general, the log-linear model is defined by choosing a class of subsets of variables, say $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_k\}$ and assuming that in every cell ω of the contingency table, the cell probability $p(\omega)$ may be written as.

$$log\, p(\omega) = \sum_{\gamma_h \in \Gamma} f_{\gamma_h}(\omega_{\gamma_h})$$

where the functions f_{γ_h} depend on the cell ω through the indices of the variables, which are in γ_h. For instance, in the case of [1], the variables are, say, A and B, $k = 3$, $\gamma_1 = \varnothing$, $\gamma_2 = A$, $\gamma_3 = B$, and

$$f_{\gamma_1} = a,\ f_{\gamma_2}(i) = b_i,\ f_{\gamma_3}(j) = c_j$$

The values of the functions f_{γ_h} are called the log-linear parameters, and they are not uniquely defined. One way to make them unique, is to assume that the values of the parameter $f_{\gamma_h}(\omega_{\gamma_h})$ sum to zero, when summed in any of the indices of the variables in γ_h. In this case, the parameters are usually denoted as

$$\lambda_{\gamma_h}(\omega_{\gamma_h})$$

Another important way to make the parameters unique, is to choose a reference category from among all the categories of every variable, and to assume the parameter is zero, whenever ω_{γ_h} is such that any of the variables in γ_h is equal to its reference category. In this case, the parameters are usually denoted as.

$$\mu(\omega_{\gamma_h})$$

For several variables, each of the above properties may be used as a basis of generalization. For example, in the case of three variables, A, B and C, a very important model is conditional independence of A and B, given C. This may be written using $\Gamma = \{\varnothing, A, B, C, AC, BC\}$, as

$$log\ p_{ijk} = f_{\varnothing} + f_A(i) + f_B(j) + f_C(k) + f_{AC}(i,k) + f_{BC}(j,k)$$

With reparameterization, the first four terms may be omitted, and also the model may be written as

$$p_{ijk} = \frac{p_{i+k}p_{+jk}}{p_{++k}}$$

as if one had $\Gamma = \{C, AC, BC\}$, or with a conditional odds ratio as

$$COR(A,B|C=k) = \frac{p_{11k}p_{22k}}{p_{12k}p_{21k}} = 1 \qquad [6]$$

A variant of the most general definition of log-linear models (cf Rudas, 1998) is that given a class of subsets of the variables (the so-called maximal interactions) Γ, if a subset of variables is contained in, for some γ_h, then these variables may be associated, but if no γ_h contains this subset, then these variables, conditioned on all other variables, are not associated. Not associated defined here using generalizations of the conditional odds ratio in [6]. The odds ratio is defined for several variables in a recursive way. For three variables,

$$OR(A,B,C) = \frac{COR(A,B|C=1)}{COR(A,B|C=2)} = \frac{p_{111}p_{221}p_{122}p_{212}}{p_{121}p_{211}p_{112}p_{222}}$$

For example, in the case of three variables, if the maximal interactions are AB, AC, and BC, then there is no association among all three variables, but there may be association among any pair of the variables. In particular setups, more straightforward interpretations are available. When the variables are not binary, a bit more elaborated definition is applied.

Applications and Interpretations of Log-Linear Models

Every log-linear model (or a distribution in a log-linear model) may be interpreted using vanishing higher order interactions. There are two very important applications of log-linear models, where a more specific interpretation is also possible. These are presented next.

Regression for Categorical Variables

In a regression type statistical problem, one of the variables is designated to be a response to the other, explanatory, variables. In the traditional setup when all variables are numerical, very often a linear effect on the expectation of the response is assumed, and when the variables have a joint normal distribution, the ordinary least squares estimation technique has many optimality properties. Such an approach is clearly not appropriate when the variables are categorical. In these cases, the best approach is to fit a log-linear model and use this to obtain a simple formula for the dependence of the conditional odds for the response variable, given the explanatory variables.

For example, out of three variables, let A be response, B and C explanatory. Also assume for simplicity, that A is binary. If the model AB, AC, BC is used, then the multiplicative form of this model, using λ parameters and $\beta = \exp\lambda$, is that

$$p_{ijk} = \beta_{\varnothing}\beta_A(i)\beta_B(j)\beta_C(k)\beta_{AB}(i,j)\beta_{AC}(i,k)\beta_{BC}(j,k)$$

Then, the conditional odds of $A=1$ versus $A=2$ may be written as

$$\frac{p_{1jk}}{p_{2jk}} = \frac{\beta_{\varnothing}\beta_A(1)\beta_B(j)\beta_C(k)\beta_{AB}(1,j)\beta_{AC}(1,k)\beta_{BC}(j,k)}{\beta_{\varnothing}\beta_A(2)\beta_B(j)\beta_C(k)\beta_{AB}(2,j)\beta_{AC}(2,k)\beta_{BC}(j,k)}$$

which simplifies to

$$\frac{p_{1jk}}{p_{2jk}} = \frac{\beta_A(1)\beta_{AB}(1,j)\beta_{AC}(1,k)}{\beta_A(2)\beta_{AB}(2,j)\beta_{AC}(2,k)} = \frac{\beta_A(1)}{\beta_A(2)}\ \frac{\beta_{AB}(1,j)}{\beta_{AB}(2,j)}\ \frac{\beta_{AC}(1,k)}{\beta_{AC}(2,k)}$$

The three terms in the last formula may be identified with the overall odds of $A=1$ versus $A=2$ and how $B=J$ and $C=K$ modify this odds. Under the model AB, AC, BC, the effects of B and C on A can be separated. The farther is the value of $\beta_{AB}(1,j)/\beta_{AB}(2,j)$ from 1, the stronger is the effect of $B=j$ on the conditional odds, and similarly for the other effects.

If, however, the AB, AC, BC model cannot be assumed to describe the population, then one has to work with the model specified by ABC, which means no restriction, and then one would have

$$\frac{p_{1jk}}{p_{2jk}} = \frac{\beta_A(1)}{\beta_A(2)}\ \frac{\beta_{AB}(1,j)}{\beta_{AB}(2,j)}\ \frac{\beta_{AC}(1,k)}{\beta_{AC}(2,k)}\ \frac{\beta_{ABC}(1,j,k)}{\beta_{ABC}(2,j,k)}$$

where the last term is the joint effect of B and C on (the odds of) A, when $B=j$ and $C=k$, which is there in addition to their individual (separable) effects.

For four variables, the model AB, ACD, BCD means that out of the three explanatory variables, B has a separable effect, and C and D also have a joint effect on A. Or AB, AC, BCD means that D does not have a direct effect on A.

This approach gives clear meaning of and answer to the questions of which out of the potential explanatory variables have effects, how strong are the existing effects and whether there are joint effects, in addition to the individual effects.

Markov Models

Many, but not all, of the log-linear models may be interpreted using conditional independences. One criterion that may be

used to decide whether or not the log-linear model specified by Γ has this property is the following one. The log-linear model may be interpreted using conditional independences, if and only if Γ does not have a forbidden configuration. A forbidden configuration is a subset of variables γ, such that Γ does not contain γ, but it does contain all proper subsets of it. To check the latter property, Γ needs to be completed with all subsets of the interactions in it, which does not affect the model, only its parameterization. For example, $\Gamma = \{AB, AC, BC\}$ has a forbidden configuration, namely ABC, $\Gamma = \{AB, AC, ABC\}$ does not have one.

The conditional independence properties, also called Markov properties, have many variants. The simplest one, called the pairwise Markov property, is that if variables A and B are not contained in any interaction of the model, then A and B are conditionally independent given all other variables.

Log-linear models that do not have a forbidden configuration are called graphical models, because they can be well-represented by graphs. The nodes of such a graph are the variables, and two nodes are connected if and only if there is an interaction in the model, which contains these two variables. For example, in the case of the model defined by $\Gamma = \{AC, BC\}$, the nodes A and C and the nodes B and C are connected, but A and B are not connected. Such a graph represents (or many of us have learned to see that way) the essence of conditional independence between A and B, given C: there may be direct relationships between A and C and between B and C, but A and B are not directly related, rather only through C. In general, according to the Markov property, two variables are conditionally independent, given all other variables, if and only if they are not connected by an edge. It has to be emphasized, that the graph is primarily a tool to communicate assumptions defining a statistical model: it has no inherent conditional independence meaning, rather one is associated with it in the graphical modeling framework. In the linear structural equations tradition, a fundamentally different interpretation is associated with a graph.

Other Markov properties, which are equivalent to the local one under the assumption that all cell probabilities are positive, are the local and the global Markov properties, see Lauritzen (1996) or Edwards (2000). The former property says that a variable and those to which it is not connected, are conditionally independent, given those, to which it is connected. The latter one says, that if three disjoint groups of variables are such, that all paths connecting a variable in the first one with a variable in the second one, goes through a variable in the third one, then the first and second groups of variables are conditionally independent given the third group of variables.

An important property of graphical Markov models is that they apply, not only to categorical variables, but to Gaussian variables, as well, and even to situations, when some of the variables are categorical, and the conditional distribution of the other variables, given the categorical ones, is Gaussian. Edwards (2000) and Whittaker (2009) contain several interesting examples of using graphical Markov models.

Estimation and Testing

In the context of log-linear models, testing of model fit is almost always based on comparing observed and estimated frequencies. The estimated frequencies are usually determined based on the principle of maximum likelihood. When all observed cell frequencies are positive, there is a unique maximum likelihood estimate of the cell probabilities, whether sampling was multinomial or Poisson. These estimates may be obtained by the Iterative Proportional Fitting algorithm.

The maximum likelihood estimates under a log-linear model determined by the interactions $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_k\}$ are uniquely characterized by the following two properties:

The conditional odds ratios of the variables γ, given any categories of all other variables, are equal to 1, if γ is not a subset of any γ_h and the marginal distributions of the variables γ are equal to their observed distributions, if γ is a subset of any γ_h.

In other words, the observed distributions of the interactions are preserved by the maximum likelihood estimate (MLE), and are combined with the no higher order interaction property defining the log-linear model. That is, the MLE is the most uniform distribution out of those that have the observed marginal distributions of the subsets of variables, which are allowed interactions.

The comparison of the observed and estimated frequencies is usually done by test statistics with an asymptotic chi-squared distribution (Read and Cressie, 2011). The application of those test statistics is known to be limited to sample sizes relatively large compared to the size of the contingency table, see, e.g., Rudas (1986) and in case of doubt, resampling methods should be used.

The log-linear parameters computed from the MLE are the maximum likelihood estimates of the true parameters and may be used in analyses like the one described in Section Regression for Categorical Variables.

Marginal Models

Marginal models generalize log-linear models, providing larger flexibility in modeling. Log-linear models restrict the association among certain variables, conditioned on the categories of all other variables. Conditioning is one way to reduce the number of variables involved in the analysis. Conditioning means selection: the analysis is confined to the cases which are in the specified categories of the conditioning variables. Another way to reduce the number of variables involved in the analysis is marginalization. Some of the variables are omitted from consideration. Several statistical applications involve marginalizing, before conditional associations could be considered. For such examples, see Bergsma et al. (2009, 2013) and Rudas and Bergsma (2004).

One approach, initiated in Bergsma and Rudas (2002), is based on defining a parameterization of all distributions using marginal log-linear parameters, and then imposing restrictions by setting some of these parameters equal to zero. A marginal log-linear parameter is a log-linear parameter, which is computed from a marginal distribution of the table. For

instance, assume that A and B are independent, and also B and C are independent, but these independences are meant to be marginal, i.e., to hold when the third variable is disregarded. Then, one may consider the AB, BC, ABC marginals of the table (the last marginal always has to be the entire table). Then the following log-linear parameters are considered:

$$\text{In the marginal } AB\text{: } \lambda_{\varnothing}^{AB}, \lambda_{A}^{AB}(i), \lambda_{B}^{AB}(j), \lambda_{AB}^{AB}(i,j)$$

$$\text{In the marginal } BC\text{: } \lambda_{C}^{BC}(k), \lambda_{BC}^{BC}(j,k)$$

$$\text{In the marginal } ABC\text{: } \lambda_{AC}^{ABC}(i,k), \lambda_{ABC}^{ABC}(i,j,k)$$

that is, in every marginal, the log-linear parameters of those interaction are computed, which have not been computed in previous marginals. The advantage of this parameterization is that the required marginal independences may now be imposed by assuming that

$$\lambda_{AB}^{AB}(i,j) = 0 \text{ and } \lambda_{BC}^{BC}(j,k) = 0$$

The above approach may always be applied, when the marginals of interest (AB and BC in the example), may be ordered in such a way, that no smaller marginal comes after a larger one. When there is only one marginal of interest, the entire table, i.e., no marginalization is involved, one obtains from this approach the standard log-linear models.

Estimation and testing of marginal log-linear models are quite well understood, and for the user, they are not more complex than the same procedures are for standard log-linear models. In general, the marginal log-linear parameters, which are not set to zero are estimated and interpreted.

In the above example, the value of $\lambda_{AC}^{ABC}(i,k)$ is a measure of how strong is the association between variables A and C, given the categories of B (in fact, the parameter gives the average value for the different categories of B), and $\lambda_{ABC}^{ABC}(i,j,k)$ is a measure of the second-order association among the three variables. If, however, interest lied in estimating the strength of marginal (as opposed to conditional) interaction between A and C, under the same model, one could use the marginals AB, AC, BC, ABC, and could parameterize as follows:

$$\text{In the marginal } AB\text{: } \lambda_{\varnothing}^{AB}, \lambda_{A}^{AB}(i), \lambda_{B}^{AB}(j), \lambda_{AB}^{AB}(i,j)$$

$$\text{In the marginal } AC\text{: } \lambda_{C}^{AC}(k), \lambda_{AC}^{AC}(i,k)$$

$$\text{In the marginal } BC\text{: } \lambda_{BC}^{BC}(j,k)$$

$$\text{In the marginal } ABC\text{: } \lambda_{ABC}^{ABC}(i,j,k)$$

The model would be defined by the same assumptions, but one would get an estimate for the strength of the marginal association between A and C from $\lambda_{AC}^{AC}(i,k)$.

Applications of Marginal Models

As the foregoing example illustrates, the marginal log-linear modeling approach is quite flexible both in model definition and in the parameters used. Next, some applications relevant in the social sciences are discussed. For several other examples, see Bergsma et al. (2009), Rudas and Bergsma (2004), and Németh and Rudas (2013a).

Panel Studies

In panel studies, the same individuals are observed repeatedly, for example, they are interviewed about their political party preferences, or their television viewing behavior is recorded. The main advantage of such designs is that transitions may be observed and analyzed. For example, how supporters of a given political party change their preferences over time or where the current supporters of a political party came from. The model of interest is often that of a Markov chain. For instance, a one-step Markov chain assumes that past positions influence the current position only through the immediate past position. More precisely, the assumption is that the conditional probabilities of the different categories of the current position, when conditioned on the previous position, do not change upon further conditioning on earlier positions. If the variable observed in time point t is denoted as A_t, the assumption is that

$$A_t \text{ is conditionally independent from } A_{t-2}, A_{t-3}, \ldots, A_1, \text{ given } A_{t-1}$$

for all time points t. If there are, say six time points, the assumptions are

$$A_3 \text{ is conditionally independent from } A_1, \text{ given } A_2$$

$$A_4 \text{ is conditionally independent from } A_2, A_1, \text{ given } A_3$$

$$A_5 \text{ is conditionally independent from } A_3, A_2, A_1, \text{ given } A_4$$

$$A_6 \text{ is conditionally independent from } A_4, A_3, A_2, A_1, \text{ given } A_5$$

and the relevant marginal log-linear parameterization is obtained using the marginals

$$A_1A_2A_3, \quad A_1A_2A_3A_4, \quad A_1A_2A_3A_4A_5, \quad A_1A_2A_3A_4A_5A_6$$

The rule to impose the required conditional independences is to set every marginal log-linear parameter to zero, which contains variables that are assumed to be conditionally independent from each other. For example, in the $A_2A_2A_3A_4$ marginal, the following log-linear parameters are used in the parameterization (the indices are omitted):

$$\lambda_{A_4}^{A_1A_2A_3A_4}, \lambda_{A_1A_4}^{A_1A_2A_3A_4}, \lambda_{A_2A_4}^{A_1A_2A_3A_4}, \lambda_{A_3A_4}^{A_1A_2A_3A_4}, \lambda_{A_1A_2A_4}^{A_1A_2A_3A_4}, \lambda_{A_1A_3A_4}^{A_1A_2A_3A_4},$$
$$\lambda_{A_2A_3A_4}^{A_1A_2A_3A_4}, \lambda_{A_1A_2A_3A_4}^{A_1A_2A_3A_4}$$

Note that interactions involving only the first three variables are parameterized in the previous marginal. The conditional independence involving the first four variables is imposed by assuming that

$$\lambda_{A_1A_4}^{A_1A_2A_3A_4} = 0, \lambda_{A_2A_4}^{A_1A_2A_3A_4} = 0, \lambda_{A_1A_2A_4}^{A_1A_2A_3A_4} = 0, \lambda_{A_1A_3A_4}^{A_1A_2A_3A_4} = 0,$$
$$\lambda_{A_2A_3A_4}^{A_1A_2A_3A_4} = 0, \lambda_{A_1A_2A_3A_4}^{A_1A_2A_3A_4} = 0$$

The parameters not set to zero are estimated from the data and model fit is tested. If the model is not rejected, the estimated values of $\lambda_{A_3A_4}^{A_1A_2A_3A_4}$ provide information for

the effect of the position at time point 3, on the position at time point 4.

In fact, panel studies may be seen as repeated measurements, where variables are measured before and after a treatment, with treatment being the events that occurred in between two measurements. In another repeated measurement setup, one may be interested in whether or not brand preferences are altered as a result of information regarding the physiological effects of the consumption of certain products. In an experiment, subjects are interviewed about their brand preferences, then they study information related to the effects of consumption, then they are interviewed again. Let C be a covariate (like income categories), and B_b and B_a. The declared brand preferences before and after the treatment was applied. Then, the assumption of the brand preference being independent from the covariate, both before and after the treatment, is exactly the first model discussed in Section Marginal Models: B_b and C are independent and also B_a and C are independent.

Directed Markov Models

Currently, much research interest is concentrated around models describing effects among variables, be these casual or evidential. Such models are discussed under various names in the literature, including directed Markov models, directed graphical models, DAG models, Bayesian networks. The assumptions in these models may also be formulated as conditional independences, i.e., Markov properties, but these take into account the direction in which the assumed effects go.

The graphs that are often used to illustrate or communicate such models are directed, which means that two nodes, representing two variables, if connected, are connected by an arrow, the direction of which shows the direction of the effect. The graphs are, further, acyclic, which means that it is not possible to start in a node, follow the direction of the arrows, and return into this node (directed acyclic graph = DAG). In a DAG, the parents of a node are those nodes, from which an arrow points to this node. The nondescendants of a node are those nodes, into which one may not arrive from this node, following the direction of the arrows. Parents of a node are all nondescendants of it, otherwise the graph would not be acyclic.

The conditional independence property associated with a DAG, the directed Markov property is that every node is conditionally independent from its nondescendants, which are not parents of it, given its parents.

For example, the Markov chain discussed in Section Panel Studies may also be described by a DAG, if one draws an arrow from A_t to A_{t+1}, for all time points t, and the directed Markov property coincides with the Markov property discussed in Section Panel Studies, and the parameterization and interpretation using marginal log-linear models also applies.

In fact, every directed Markov model is a marginal log-linear model (see Rudas and Bergsma, 2004 or Németh and Rudas, 2013a) and the estimated values of the marginal log-linear parameters may be used to assess the magnitude of effects suggested by the data, after appropriate model testing. There are several areas in social and behavioral research, where DAG models describe relevant substantive hypotheses. Németh and Rudas (2013a,b) apply marginal models for the analysis of the social status attainment process in various setups, and give detailed description of interpretations and computations, using DAG and some other related models.

Other Generalizations and Applications

From a more general perspective, marginal log-linear models may be seen useful, when one deals with variables, which are related in some way, and the relationship may be caused by the sampling procedure or time dependence or by other reasons, and the immediate goal is to do an analysis with reference to these relationships. Several such applications, including information about relevant software, are presented in Bergsma et al. (2009). The book also discusses the relationship between marginal log-linear modeling and structural equations modeling (SEM).

Another interesting application of marginal log-linear models is the specification and analysis of path models for categorical data. These models assume that only effects associated with the arrows present in a DAG exist among variables (Goodman, 1973). In a SEM approach, this is usually interpreted as the existence of some kind of regression equations, linking variables to their parents. Such regressions are hard to interpret in the case of categorical variables, and the path model assumption is more naturally taken as meaning that no higher order associations exists, than those including variables and one of their parents (see Rudas et al., 2006). Under such a model, the existing effects are fully quantified by the marginal log-linear parameters involving a variable and one of its parents.

Marginal log-linear models may be seen as an approach, where effects present in the joint distribution of several categorical variables, often thought of as a contingency table, include such that they involve only a group of variables (a marginal), not all. It is also possible to consider models, where effects are associated not with groups of variables, rather with arbitrary subsets of the cells of the contingency table, which have some common property. For example, Klimova and Rudas (2012) present an analysis of social mobility tables, where the effect of the distance of the son's position from the father's position is also taken into account and is quantified.

Acknowledgment

The writing of this article was supported in part by Grant K-106154 from the Hungarian National Scientific Research Fund (OTKA).

See also: Bayesian Graphical Models and Networks; Binary Response Models and Logistic Regression; Graphical Markov Models: Overview; Markov Models and Social Analysis; Multivariate Analysis: Discrete Variables (Loglinear Models); Multivariate Analysis: Discrete Variables (Overview).

Bibliography

Bergsma, W.P.V., Croon, M., Hagenaars, J.A., 2009. Marginal Model: For Dependent, Clustered, and Longitudinal Categorical Data. Springer, New York.

Bergsma, W.P.V., Croon, M., Hagenaars, J.A., 2013. Advancements in marginal modeling for categorical data (with discussion). Sociological Methodology 43, 1–41.
Bergsma, W.P., Rudas, T., 2002. Marginal models for categorical data. Annals of Statistics 30, 140–159.
Bishop, Y.V.V., Fienberg, S.E., Holland, P.W., 1975. Discrete Multivariate Analysis. MIT Press, Cambridge.
Edwards, D., 2000. Introduction to Graphical Modeling. Springer, New York.
Goodman, L.A., 1970. The multivariate analysis of qualitative data: interactions among multiple classifications. Journal of the American Statistical Association 50, 226–256.
Goodman, L.A., 1973. The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. Biometrika 60, 179–192.
Gokhale, D.V., Kullback, S., 1978. The Information in Contingency Tables. Marcel Dekker, New York.
Haberman, S.J., 1978. Analysis of Qualitative Data, vol. 1 & 2. Academic Press, New York.
Klimova, A., Rudas, T., 2012. Coordinate free analysis of trends in British social mobility. Journal of Applied Statistics 39, 1681–1691.
Lauritzen, S.L., 1996. Graphical Models. Oxford University Press, New York.
Németh, R., Rudas, T., 2013a. On the application of discrete marginal graphical models (with discussion). Sociological Methodology 43, 70–100.
Németh, R., Rudas, T., 2013b. Discrete graphical models in social mobility research – a comparative analysis of American, Czechoslovakian and Hungarian mobility before the collapse of state socialism. Bulletin of Sociological Methodology 118, 5–21.
Read, T.R.C., Cressie, N.A.C., 2011. Goodness-of-Fit Statistics for Discrete Multivariate Data. Springer, New York.
Rudas, T., 1986. A Monte Carlo comparison of the small sample behaviour of the Pearson, the likelihood ratio and the Cressie-Read statistics. Journal of Statistical Computation and Simulation 24, 107–120.
Rudas, T., 1998. Odds Ratios in the Analysis of Contingency Tables. Sage, Newbury Park.
Rudas, T., Bergsma, W.P., 2004. On application of marginal models for categorical data. Metron 17, 1–25.
Rudas, T., Bergsma, W.P.V., Némenth, R., 2006. Parameterization and estimation of path models for categorical data. In: Rizzi, A., Vichi, M. (Eds.), COMPSTAT 2006. Physica, Heidelberg, pp. 383–394.
Whittaker, J., 2009. Graphical Models in Applied Multivariate Statistics. Wiley, New York.

Longitudinal Analyses of Sexual Development through Early Adulthood

Joseph Gonzales and Emilio Ferrer, University of California, Davis, Davis, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Longitudinal data analysis is a category of analytic techniques used to assess data in which units (e.g., persons) are measured more than once. These techniques can be used to inform specific questions such as (1) what are the trajectories and rates of change in a behavior (e.g., latent growth modeling); (2) what are the relations of changes in behavior (e.g., bivariate latent growth models); (3) can we predict outcomes using measures of change (e.g., latent growth models with exogenous outcomes); and (4) when are events most likely to occur for individuals along some temporal metric (e.g., discrete-time survival analysis). While these questions are unique, they are all concerned with understanding patterns and relations of intraindividual change utilizing longitudinal data. In this article, we apply these methods in order to demonstrate different ways of interpreting longitudinal change in sexual behavior from late childhood through young adulthood.

Longitudinal data analysis affords researchers access to the dimension of time, that space within which developmental changes occur. While this dimension is not required to infer meaningful relations and experimental effects, it is necessary if a researcher wants to understand the dynamics of the processes in question, as it unfolds over time.

Consider, for example, the established and growing body of literature on sexual behaviors. In particular, risky sexual behaviors (e.g., Coker et al., 1994; Grello et al., 2006; Gute and Eshbaugh, 2008; Paul et al., 2000) have been linked to substance use (e.g., alcohol; Grello et al., 2006; Paul et al., 2000; Traeen and Lewin, 1992). Much of this research is based on methods utilizing single observations. Although informative in some ways, such methods limit our understanding of the intraindividual aspects of development and change. In this article, we demonstrate longitudinal modeling methods based on within-person changes (e.g., McArdle, 2009; Singer and Willet, 1993) and evaluate questions that pertain to developmental processes and their interrelations over time (Baltes and Nesselroade, 1979). To demonstrate these methods, we apply them to empirical data on sexual behavior over time to examine both behavioral change and its correlates from late childhood through early adulthood.

Behavioral Change Over Time

During the teen years, many adolescents engage in sexual behaviors for the first time (Chandra et al., 2005; Martinez et al., 2006). While these experiences seem normative and do not necessarily impair adolescents' ability to maintain various adaptive goals (e.g., educational goals and personal health), subsets of adolescents seem prone to encountering negative consequence (e.g., contraction of STDs and unplanned pregnancies) as a result of high-risk sexual behavior. One factor that has been identified as contributing to participation in high-risk sexual behaviors is alcohol use (e.g., Grello et al., 2006; Gute and Eshbaugh, 2008; Miller, 2012; Paul et al., 2000). However, the unique and interrelated development of these behaviors is generally not explicitly explored in the extant literature.

To characterize changes in sexual behavior and alcohol use, as well as their interrelations, methods based on within-person changes are needed. One such method is latent growth modeling (LGM; e.g., McArdle, 2009; McArdle and Hamagami, 2001). LGM can be used to identify changes occurring over various time metrics (e.g., age) to describe the trajectory of change (e.g., changes in the number of novel sexual partners and frequency of alcohol use; Dogan et al., 2010) and their interrelations over time. An LGM can be defined (e.g., Ferrer et al., 2004) as follows:

$$Y(t)_i = \gamma_{0i} + A(t) * \gamma_{si} + e(t)_i \qquad [1]$$

$$\gamma_{0i} = \mu_0 + e_{0i} \quad \text{and} \quad \gamma_{si} = \mu_s + e_{si} \qquad [2]$$

Where Y is an outcome measured over time ($t = 1, \ldots, T$) across the same individual ($i = 1, \ldots, I$). The coefficients γ_0 and γ_s represent the initial level and the slope of change in the outcome variable over time, respectively. The basis coefficients, $A(t)$, represent the timing or shape of growth from the slope over time, and $e(t)$ represents the unobserved error of measurement. Both the level and slope include individual differences such that they have fixed group means, μ_0 and μ_s, and residuals, e_{0i} and e_{si}.

From this initial model, several modifications can be made to explore other research questions. For example, a bivariate growth model can be built with two LGMs describing how two different growth trajectories are related to each other with respect to their starting values and their slopes (McArdle, 1988; 2009). Consider the bivariate case. This model can be written for variables $Y(t)$ and $X(t)$ over time as a model of two growth curves (e.g., Ferrer et al., 2005).

$$\begin{aligned} Y(t)_i &= \gamma_{0i} + \text{Age}_y(t)_i \cdot \gamma_{si} + e_y(t)_i \quad \text{and} \\ X(t)_i &= x_{0i} + \text{Age}_x(t)_i \cdot x_{si} + e_x(t)_i \end{aligned} \qquad [3]$$

with all covariances allowed among the latent variables ($\sigma_{y(i),\, x(i)}$). Among these covariances, $\sigma_{ys,\, xs}$ represents the covariance between the age slopes of the two variables; the degree that total change in one variable is related to total change in the other.

This bivariate model can also be written using a random coefficients, or multilevel, notation as a model with a single outcome variable Y_{itk} for individual i at occasion t on variable k, where the variables k do not need to be measured at the same occasions for all individuals as,

$$Y_{itk} = \sum_{k=1}^{m} d_k(\beta_{0ik} + \beta_{1ik} \cdot \text{Age}_{itk} + \beta_{2ik} \cdot p_{itk} + e_{itk}) \quad [4]$$

where $k = 1, \ldots, m$ represents the number of variables (two in this study), d_k is a dummy variable d_1 to d_m (1 or 0) to indicate the dependent variables, β_{0ik} is the intercept for person i on variable k, β_{1ik} is the slope associated with age for person i on variable k, and e_{itk} is the residual for person i at occasion t for variable k. Similar to the univariate case, the intercepts and slopes for the k variables can be decomposed as,

$$\begin{aligned} \beta_{0ik} &= \beta_{0k} + u_{0ik}, \quad \text{and} \\ \beta_{1ik} &= \beta_{1k} + u_{1ik}, \end{aligned} \quad [5]$$

which indicates that, for each variable k, the intercept and slope have fixed group means (β_{0k}, β_{1k}) and random residuals (u_{0k}, u_{1k}), and these residuals have variance components ($\sigma^2_{0k}, \sigma^2_{1k}$) and could covary among themselves. In addition to these within-variable k random effects, covariances between intercepts and slopes are also allowed between variables. To test specific hypotheses – or merely due to data constraints – restrictions can be imposed as needed in the covariance matrix of individual residual variances $\mathbf{T} = \text{cov}(\mathbf{E}_{nt})$ and in the between-individual covariance matrix $\Sigma = \text{cov}(U_t)$. This model can now be estimated using any standard multilevel program.

Event Occurrence or Onset

Another substantive example to consider is when are people most likely to begin engaging in sexual behavior, and what factors will delay or expedite their sexual debut. This information may be particularly helpful in timing interventions such as sex education (e.g., Somers and Surmann, 2005) by identifying periods of greater behavioral risk for various subpopulations. A longitudinal analytic approach that can be used to address this question is discrete-time survival analysis. This technique can be applied to discrete intervals of time (e.g., years) to examine the likelihood of a one-time behavioral occurrence (Singer and Willet, 1991, 1993).

Singer and Willet (1993) present a survival analysis model where survival for a given period is defined as follows,

$$s_t = s_{t-1} * (1 - h_t) \quad [6]$$

such that t represents the time period, hazard, h_t, values represent the risk of an event occurring in the (remaining) period population during a given period, and survival, s_t, represents the proportion of the surviving population during a given period.

The discrete-time hazard, h_t, is defined as the probability that a randomly selected individual, i, will have the target event occur during time period, t, given that he or she has not already experienced the event.

$$h_{it} = \frac{1}{1 + e^{-\left[\left(\alpha_1 D_{1it} + \alpha_2 D_{2it} + \cdots + \alpha_t D_{T_{it}}\right) + \left(\beta_1 Z_{1it} + \beta_2 Z_{2it} + \cdots + \beta_p Z_{Pit}\right)\right]}} \quad [7]$$

With D representing dummy coding for periods, and Z representing covariates that can be either time invariant or time variant for individuals between periods. Both invariant and variant covariates are assumed to be invariant during each time period. The intercept parameters ($\alpha_1, \ldots, \alpha_T$) describe the baseline hazard rate for each time period and the slope parameters ($\beta_1, \ldots, \beta_P$) describe the effects of the covariates on the baseline hazard rate.

Empirical Examples

In the following examples, we demonstrate how different growth models can accomplish some of the objectives of longitudinal research (Baltes and Nesselroade, 1979): (1) the identification of intraindividual change (i.e., how does the number of total sexual partners change across observation periods); (2) the analysis of interrelations in behavioral change (i.e., how do the initial levels and changes of total sexual partners and alcohol use relate to each other); and (3) how do the initial levels and changes for both total number of sexual partners and drinking behavior predict an exogenous outcome of interest (i.e., number of pregnancies). Additionally, we illustrate the application of discrete-time survival analysis with respect to a critical event in sexual development, the sexual debut. As part of this illustration, we incorporate time-invariant (i.e., biological sex) and time-variant (i.e., age of menarche) covariates to evaluate differences in risk during any given period of observation as a function of heterogeneity between individuals.

Method

Participants

We use data from the National Longitudinal Study of Adolescent Health (Add Health) (Harris and Udry, 2013). Data were collected as part of a longitudinal study of a nationally representative sample of the United States adolescents in grades 7–12 during the 1994–95 school years. The data set consists of four waves of data separated by approximately 1, 4, and 7 years (waves 1–2, 2–3, and 3–4, respectively).

A subset of participants were selected for inclusion in the present study, if their previous interview responses regarding their number of total sexual partners were less than or equal to their later responses (i.e., wave 1 $\leq$ wave 2 $\leq$ wave 3 $\leq$ wave 4). A total of 5200 participants were included in this subset, of which 47% were male and 53% were female. Ages ranged 12–21 years ($M = 16$, $\text{SD} = 1.80$) at wave 1, 13–22 years ($M = 16.5$, $\text{SD} = 1.59$) at wave 2, 18–28 years ($M = 22.22$, $\text{SD} = 1.87$) at wave 3, and 24–34 years ($M = 28.87$, $\text{SD} = 1.79$) at wave 4.

Materials and Procedure

At each wave of the study, participants responded to a battery of items both with an interviewer and independently via a computer with prompts.

Total Number of Sexual Partners

Wave 1–4 interviews utilized different methods for assessing total number of sexual partners. Wave 1 asked participants to indicate how many total sexual partners participants had; wave 2 asked for participants to indicate how many new sexual partners they had since their last interview; wave 3

asked how many total partners participants had; and wave 4 asked two questions: (1) how many total female sexual partners participants had, and (2) how many total male sexual partners participants had. Wave 2 total sexual partner counts were computed by adding wave 1 total partner counts and wave 2 new partner counts. Wave 4 total sexual partner counts were computed by summing the total male partner and total female partner values.

Participant Drinking Behavior

Wave 1–4 interviews included a question about participant drinking behavior, "During the past 12 months, on how many days have you gotten drunk?" For waves 1 and 2, participants were asked to respond using a 7-point likert scale ranging from 1 ('every day') to 7 ('never'). In waves 3 and 4, participants responded to the same question using a 7-point likert scale ranging from 0 ('never') to 6 ('every day'). Waves 1 and 2 responses were reverse coded to conform to the waves 3 and 4 reporting format.

Participant Age at Time of Interview

Participants' birth month and year were recorded as part of their participation in the study. Additionally, the month and year of each interview was also recorded. Birth month and interview month were scored 1–12 (January–December, respectively). Age at time of interview was computed by dividing birth month and interview months by 12 and adding them to birth year and interview years, respectively. This new birth year score was then subtracted from the new interview year score for waves 1–4, yielding an age at interview score for each interview wave. Two age score formats were used: (1) the computed age scores and (2) the computed age scores rounded to the nearest year.

Age of Sexual Debut

During wave 4, participants were asked at what age they first had vaginal intercourse. Reported ages ranged from 1 to 30 years old ($M = 16.82$, SD = 2.99).

Age of Menarche

Female participants were asked at wave 1–3 interviews what age, in years, they had their first menstrual period. Disagreement in age of first menstrual period was resolved by taking the average of ages reported by participants and rounding to the nearest year. Ages ranged from 7 to 17 ($M = 12.26$, SD = 1.36).

Additionally, unrounded age of first menstrual period scores for each participant were used to center participant unrounded ages to yield a participant-centered menstrual age ranging from 2.42 to 22.67 ($M = 8.36$, SD = 5.67). Negative scores indicate years prior to age of menarche, zero indicates age at menarche, and positive scores indicate years after menarche.

Number of Participant Pregnancies

During the wave 4 interview, participants were asked how many times they either became pregnant or impregnated someone else; regardless of the outcome of the pregnancy. Number of total pregnancies reported ranged from 0 to 12 ($M = 1.35$, SD = 1.55).

Results

Growth Modeling: Wave-Based Time Metric

Total Number of Sexual Partners

In the first set of analyses, we examined the trajectories of total number of sexual partners over interview waves. Figure 1 depicts the number of sexual partners across waves 1–4 and suggests that a linear growth may be an appropriate starting point. We fit no growth, linear growth, and quadratic growth models using wave as the measurement occasion. All analyses were conducted using maximum likelihood estimation. Fit for the no growth, linear growth, and quadratic growth models were poor and they were rejected as untenable models. A nonlinear LGM was estimated by constraining the wave 1 and 4 loadings from the slope to zero and one, respectively (see Figure 2), and estimating the slope's basis coefficients for waves 2 and 3.

Fit to the data with the estimated slope loadings was good (see Table 1), suggesting that this fourth model was a tenable representation of the change in number of sexual partners. Moreover, estimates of the slope loadings for waves 2 and 3 ($\alpha = 0.07$ and 0.36, respectively) conform to expectations from the graphs of the data: there is little growth from wave 1 to 2, and transitions from wave 2 to 3, and wave 3 to 4 represent the majority of change in total number of sexual partners.

While this rate of change may not seem linear, it should be noted that time from wave 1 to 2 is about 1 year. In contrast, from wave 2 to 3 is about 4 years, and from wave 3 to 4 is about 7 years. Thus, the average yearly rate of change from wave 2 to 3 was about 0.09, and the average yearly rate of change from wave 3 to 4 was also about 0.09. While we cannot use waves to evaluate if this rate of change was constant between unobserved years, it does suggest that growth may have been relatively constant across the overall period of observation.

Ratings of Days Drunk

The same no growth, linear growth, and quadratic growth models were next fit to participants' drinking behavior ratings. Again, fit for each model was poor and they were rejected. As with total number of sex partners, a nonlinear LGM was fit to the data constraining the wave 1 and 4 loadings from the slope to zero and one, respectively.

Fit to the data was good (see Table 1), suggesting that this fourth model was a tenable representation of the change in ratings of drinking behavior. Again, the estimates of slope loadings for waves 2 and 3 ($\alpha = 0.19$ and 1.08, respectively) conform well to expectations based on plots of the data (see Figure 1), such that there was a peak in drinking behavior at wave 3, with a slight decline at wave 4.

As with total number of sexual partners, it is worth noting that these rates of change are not for equivalent periods of time if considered with respect to the number of years that pass between waves of data collection. Here, the average yearly rate of change from wave 2 to 3 is about 0.22, and the yearly rate of change from wave 3 to 4 is about 0.01. This suggests that the average yearly rate of change from wave 1 to 3 may be very similar, and that there was very little yearly change from wave 3 to 4.

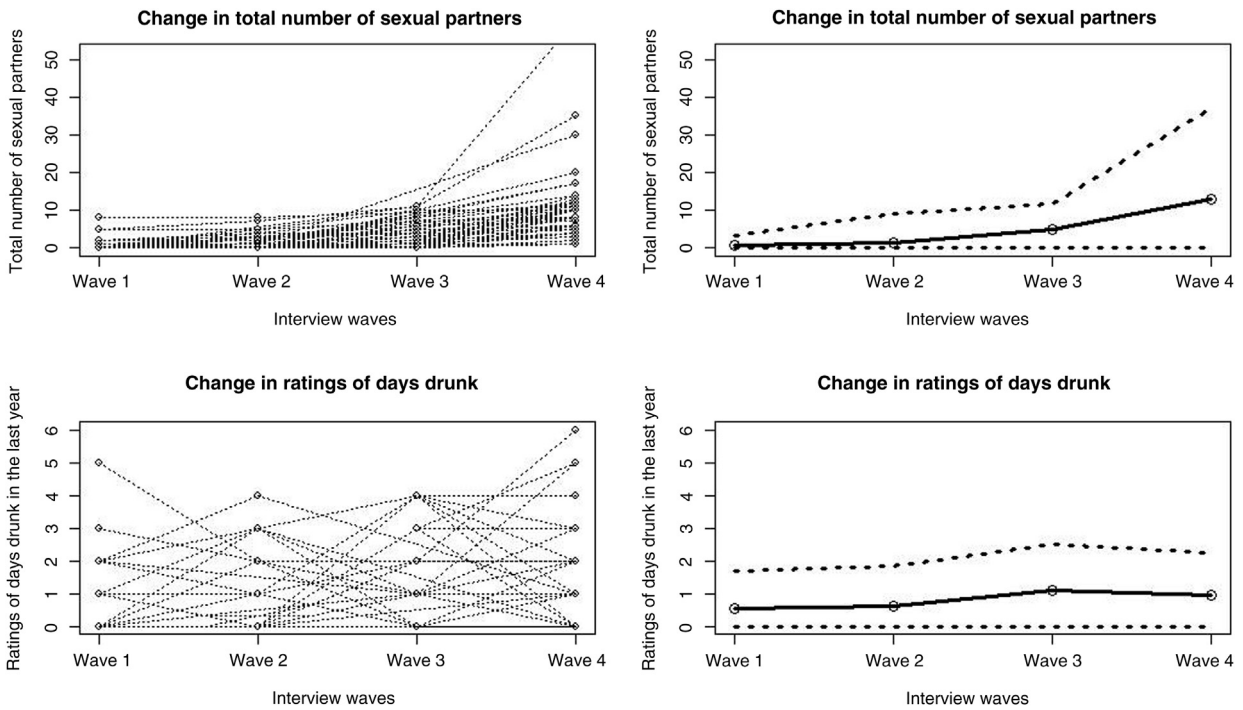

Figure 1 Changes in the total number of sexual partners and drinking behavior over waves of participant interviews. The graph on the top-left depicts the individual increase in total number of sexual partners for 50 randomly selected participants. The graph on the top-right depicts the average change in total number of sexual partners using the solid black line, with 1 SD errors around the mean value indicated by dashed lines. The graph on the bottom-left depicts individual drinking behavior trajectories of 50 randomly selected participants. The graph on the bottom-right depicts the average change in drinking behavior using a solid black line, with 1 SD errors around the mean value indicated by dashed lines. Note that the error bars have no values below zero as a score less than zero is not possible. Also note that variability in total number of sexual partners is very large in wave 4 relative to waves 1–3, and that variability for this measure is lowest in wave 1 relative to waves 2–4.

Bivariate Model for Number of Sex Partners and Drinking

Having developed tenable growth models for total number of sex partners and drinking behavior across waves, we next fit a bivariate growth model to examine the relation between these two processes with respect to their initial levels and changes over time. For this, we specified a model in which the intercepts and slopes of both variables were allowed to covary (see Figure 3). These covariances describe relations between latent growth factors for total number of sexual partners and drinking behavior.

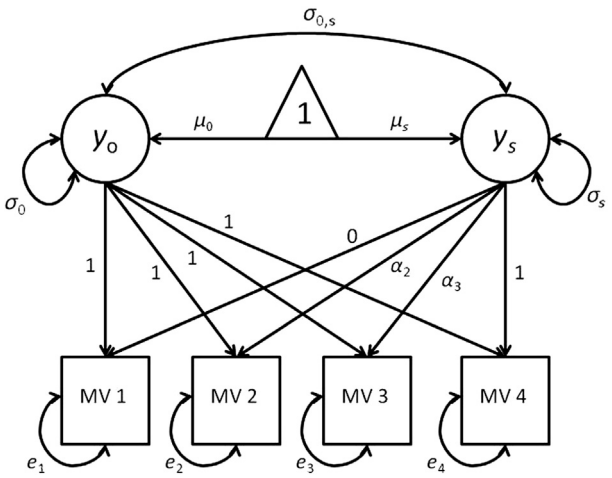

Figure 2 Path diagram of a growth model with four points of observation for a single manifest variable. Lines with double-arrows represent covariances and variances. Lines with single arrows represent regressions or factor loadings. Y_0 represents the level and Y_S represents the slope.

Model fit to the data was excellent (see Table 2). To ease interpretation, covariances were standardized into correlations. Parameter estimates of bivariate relations between both variables indicate a positive correlation between the initial ratings of drinking behavior and number of sexual partners ($r = 0.38$, $p < 0.01$), but no relation between total change in number of sexual partners and change in drinking behavior ($r = 0.04$, $p = 0.12$). These results indicate that drinking behavior and total number of sexual partners are related initially, but that their changes are not.

Bivariate Growth with an Exogenous Outcome

Next, we added an exogenous outcome that is regressed upon the levels and slopes for total number of sexual partners and drinking behavior. Given interest in the possible relation between number of pregnancies and number of sexual partners, a substantively interesting outcome is the total number of pregnancies reported by the participants. Again, model fit was excellent, and estimates varied very little with the previous bivariate model (see Table 2).

Estimates indicate that both the level and slope of drinking behavior are negatively related to number of pregnancies,

Table 1 Estimates for number of sexual partners and drinking behavior growth models

Total number of sexual partners					*Ratings of days drunk during last 12 months*				
		Estimate	*SE*	p-*value*			*Estimate*	*SE*	p-*value*
Y_0	Tsex1	1	–	–	Y_0	Drink1	1	–	–
	Tsex2	1	–	–		Drink2	1	–	–
	Tsex3	1	–	–		Drink3	1	–	–
	Tsex4	1	–	–		Drink4	1	–	–
Y_s	Tsex1	0	–	–	Y_s	Drink1	0	–	–
	Tsex2	0.07	0.01	<0.01		Drink2	0.19	0.03	<0.01
	Tsex3	0.36	0.01	<0.01		Drink3	1.08	0.04	<0.01
	Tsex4	1	–	–		Drink4	1	–	–
Cor	(Y_0, Y_s)	0.35	0.04	<0.01	Cor	(Y_0, Y_s)	−0.57	0.02	<0.01
Means	Y_0	0.85	0.03	<0.01	Means	Y_0	0.54	0.02	<0.01
	Y_s	12.57	0.36	<0.01		Y_s	0.48	0.02	<0.01
Variance	Y_0	5.78	0.35	<0.01	Variance	Y_0	0.81	0.04	<0.01
	Y_s	238.36	10.82	<0.01		Y_s	1.12	0.08	<0.01
Residual	Tsex1	0.31	0.34	0.36	Residual	Drink1	0.5	0.04	<0.01
	Tsex2	53.08	1.37	<0.01		Drink2	0.9	0.03	<0.01
	Tsex3	7.63	1.15	<0.01		Drink3	1.03	0.05	<0.01
	Tsex4	316.32	11.42	<0.01		Drink4	0.85	0.04	<0.01
CFI	0.962				CFI	0.972			
RMSEA	0.079 (0.066, 0.092)				RMSEA	0.059 (0.047, 0.073)			

Note: Y_0 denotes the latent level for each growth model and *Ys* denotes the latent slope for each growth model. Tsex1–4 indicates manifest scores of total number of sexual partners at waves 1–4, respectively. Drink 1–4 indicates manifest scores of self-reported ratings of days the participant was drunk over the last 12 months at waves 1–4, respectively. Cor indicates the correlation between the level and slope for each growth model. Level loadings for waves 1–4 were constrained to equal one in both models, and wave 1 and 4 slope loadings were constrained to equal 0 and 1, respectively, in both models. CFI and RMSEA are measures of model fit. RMSEA consists of its score followed by its 90% CI in parentheses.

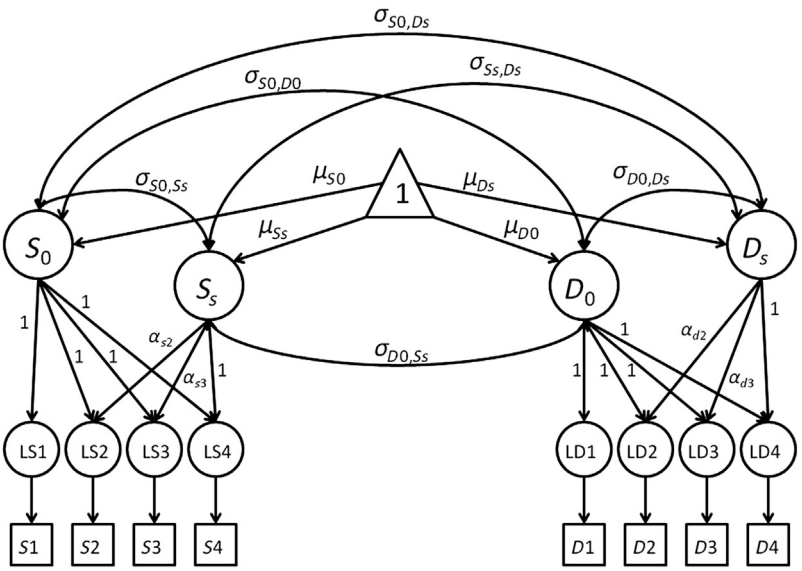

Figure 3 Path diagram of bivariate latent growth model. Lines with double-arrows represent covariances and variances. Lines with single arrows represent regressions or factor loadings. Note the covariances between the levels and slopes for total number of sex partners (S_0 and S_S) and drinking behavior (D_0 and D_S). For convenience, residuals were not depicted in the model.

whereas level and slope of number of sexual partners are positively related to number of pregnancies. Differences in the reported (unstandardized) estimates may lead one to suspect that latent growth factors for ratings of drinking behavior are stronger predictors of number of pregnancies. However, such differences are partially a result of variable metrics. The standardized coefficients for level and slope of number of sexual partners ($\beta = 0.08$ and 0.19, respectively) and drinking behavior

Table 2 Bivariate growth model for number of sexual partners and drinking behavior

Without pregnancy outcome					*With pregnancy outcome*				
		Estimate	*SE*	p-*value*			*Estimate*	*SE*	p-*value*
Alc Y_1 BY	LDrinkl	0	–	–	Alc Y_1 BY	LDrink1	0	–	–
	LDrink2	0.22	0.02	<0.01		LDrink2	0.2	0.02	<0.01
	LDrink3	1.1	0.04	<0.01		LDrink3	1.1	0.04	<0.01
	LDrink4	1	–	–		LDrink4	1	–	–
Tsex Y_1 BY	LTSex1	0	–	–	Tsex Y_1 BY	LTSex1	0	–	–
	LTSex2	0.07	0.01	<0.01		LTSex2	0.07	0.01	<0.01
	LTSex3	0.35	0.01	<0.01		LTSex3	0.35	0.01	<0.01
	LTSex4	1	–	–		LTSex4	1	–	–
		Correlations					*Correlations*		
Tsex Y_0	Alc Y_0	0.38	0.02	<0.01	Tsex Y_0	Alc Y_0	0.39	0.02	<0.01
	Alc Y_1	−0.28	0.02	<0.01		Alc Y_1	−0.28	0.03	<0.01
	Tsex Y_1	0.39	0.04	<0.01		Tsex Y_1	0.39	0.04	<0.01
Alc Y_0	Alc Y_1	−0.59	0.02	<0.01	Alc Y_0	Alc Y_1	−0.57	0.02	<0.01
	Tsex Y_1	0.28	0.02	<0.01		Tsex Y_1	0.29	0.02	<0.01
Tsex Y_1	Alc Y_1	0.04	0.03	0.12	Tsex Y_1	Alc Y_1	0.04	0.03	0.1
		Regressions					*Regressions*		
Npreg	Tsex Y_0	–	–	–	Npreg	Tsex Y_0	0.06	0.02	<0.01
	Alc Y_0	–	–	–		Alc Y_0	0.02	0.003	<0.01
	Tsex Y_1	–	–	–		Tsex Y_1	−0.33	0.05	<0.01
	Alc Y_1	–	–	–		Alc Y_1	−0.56	0.05	<0.01
		Model Fit					*Model Fit*		
	CFI	0.963				CFI	0.963		
	RMSEA	0.046 (0.040, 0.051)				RMSEA	0.043 (0.038, 0.048)		

Note: In both models the levels and slopes of total number of sexual partners and days drunk ratings were allowed to covary. Y_0 denotes the intercept and Y_1 denotes the slope. Tsex is total number of sex partners, Alc is ratings of days drunk, and Npreg is the total number of pregnancies participants reported having or caused during the wave 4 interview. Number of pregnancies was regressed on the total number of sexual partners and days drunk rating levels and slopes in the 'With Pregnancy Outcome' model. CFI and RMSEA are measures of model fit. RMSEA consists of its score followed by its 90% CI in parentheses.

($\beta = -0.19$ and -0.38, respectively) show that these differences in magnitude are not as pronounced as they initially appeared.

Growth Modeling: Age-Based Time Metric

One limitation of the previous models is that while they evaluate the temporal interrelations of two variables of interest, they do so using a time metric of little substantive value. Saying that the number of total sex partners changed from wave 1 to wave 2 means little to a substantive researcher who may be interested in how these changes occur over a given period of time (e.g., adolescence). Yet, many longitudinal studies produce panel data much like this, consisting of a cross-sectional sampling of participants at different ages over a period of time.

A more meaningful approach, especially from a developmental perspective, would be to model change in total number of sex partners and drinking behavior using age as the underlying metric (see Figure 4). To illustrate this possibility, we fit the individual and bivariate models using age as the time metric. Additionally, rather than using discrete-time intervals, we now treat the time interval as continuous with all age scores centered using the lowest observed age in the sample, 1 158 333 years.

Unlike the wave-based models using discrete time, there are no longer individual slope loadings estimated for time intervals (see Table 3). Similar to the wave-based models, we still estimate an initial level for both total number of sexual partners and drinking behavior ($\gamma_o = -3.11$ and 0.45, respectively), which now represents the expected values of the outcome variable when age is 1 158 333 years. Additionally, we estimate a slope for number of sexual partners and drinking behavior ($\gamma_s = 0.9$ and 0.04, respectively), which now represents the expected amount of change in the outcome for each year of age.

Also like the wave-based models, we can model both outcomes at the same time and investigate the relation between the total number of sexual partners and drinking behavior with respect to their levels and slopes. These relations differ from the previous model in several interesting ways. First, the levels of number of sexual partners and drinking behavior are no longer related ($r = -0.1$, $p = 0.06$). This could be because at the initial age in this sample there is not much occurrence of either behavior. In contrast, the wave-based format includes participants in wave 1 who were an average of 4 years older than the continuous age intercept, with higher corresponding scores and variability of both total sex partners and drinking behavior. Second, the association between the changes in total number of sexual partners and those of drinking behavior is now positively related ($r = 0.22$, $p < 0.01$), suggesting that changes in both behaviors are related.

It is evident that by using a different time metric, we find differences that suggest distinct relations between total number of sex partners and drinking behavior. Specifically, when examined across waves the initial levels of both variables are related but their changes are not. When evaluated over age, this is reversed and the changes are correlated but the initial levels are not. In spite of these otherwise unsurprising discrepancies, when we regressed the exogenous variable, total

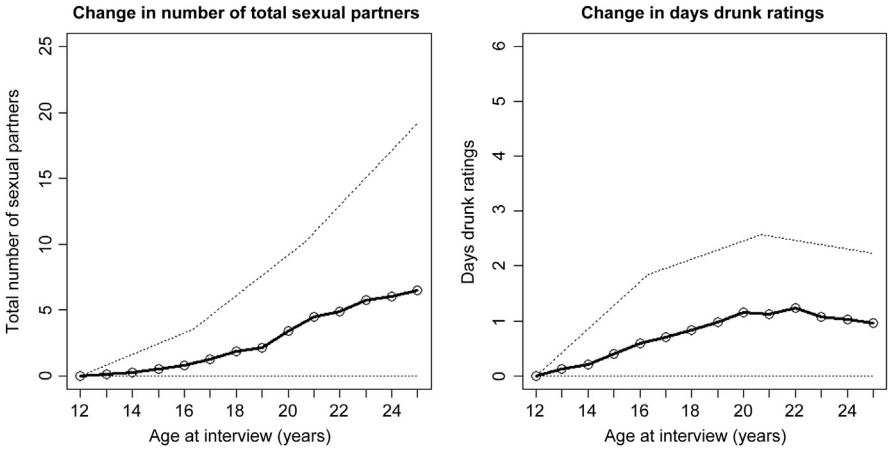

Figure 4 Change in total number of sex partners and ratings of days spent drunk across participant age. This demonstrates an alternative and developmentally meaningful time metric across which to plot changes in total number of sexual partners and ratings of days spent drunk.

number of pregnancies, on our levels and slopes for total number of sexual partners and drinking behavior we find that results are fairly similar.

Growth Modeling: Menstrual Age-Based Time Metric

To further emphasize the possibility of using different time metrics when modeling longitudinal data, we now consider a model of female participants based upon menstrual age (see Figure 5). We centered female participants' age by their age of menarche. This yields a menstrual age score for each participant where negative age scores indicate premenarche age, zero represents age of menarche, and positive scores indicate postmenarche age. While each female participant may have experienced menarche at a different age, in this format they are centered to be on the same menstrual-age metric.

Again, we modeled the menstrual age continuously, and therefore do not have individual slope loadings for individual ages (see Table 4). However, as with the previous age model,

Table 3 Growth models using continuous age in years as time metric

Total number of sexual partners					*Drinking behavior*				
		Estimate	*SE*	p-*value*			*Estimate*	*SE*	p-*value*
Cor	(Y_0, Y_1)	−0.999	0.02	<0.01	Cor	(Y_0, Y_1)	−0.37	0.01	<0.01
Means	Y_0	−3.11	0.13	<0.01	Means	Y_0	0.45	0.02	<0.01
	Y_1	0.9	0.36	<0.01		Y_1	0.04	0.002	<0.01
Variance	Y_0	43.04	3.26	<0.01	Variance	Y_0	0.46	0.04	<0.01
	Y_1	2.99	0.08	<0.01		Y_1	0.003	<0.000	<0.01
Residual	Tsex	33.07	0.58	<0.01	Residual	Ddrunk	1.03	0.02	<0.01
Bivariate model without covariate					*Bivariate model with covariate*				
Correlations		*Estimate*	*SE*	p-*value*	*Correlations*		*Estimate*	*SE*	p-*value*
Tsex Y_0	Alc Y_0	−0.1	0.01	0.06	Tsex Y_0	Alc Y_0	−0.09	0.01	0.06
	Alc Y_1	−0.21	0.01	<0.01		Alc Y_1	−0.19	0.01	<0.01
	Tsex Y_1	−0.998	0.001	<0.01		Tsex Y_1	−0.997	0.001	<0.01
Alc Y_0	Alc Y_1	−0.35	0.01	<0.01	Alc Y_0	Alc Y_1	−0.35	0.01	<0.01
	Tsex Y_1	0.15	0.01	<0.01		Tsex Y_1	0.15	0.01	<0.01
Tsex Y_1	Alc Y_1	0.22	0.01	<0.01	Tsex Y_1	Alc Y_1	0.19	0.01	<0.01
Regressions		*Estimate*	*SE*	p-*value*	*Regressions*		*Estimate*	*SE*	p-*value*
Npreg	Tsex Y_0	–	–	–	Npreg	Tsex Y_0	−0.02	0.01	<0.01
	Alc Y_0	–	–	–		Alc Y_0	0.17	0.08	0.03
	Tsex Y_1	–	–	–		Tsex Y_1	0.1	0.02	<0.01
	Alc Y_1	–	–	–		Alc Y_1	−10.07	0.94	<0.01

Note: In the bivariate models, the levels and slopes of total number of sexual partners and days drunk ratings were allowed to covary. Y_0 denotes the intercept and Y_1 denotes the slope. Tsex is total number of sex partners, Alc is ratings of days drunk, and Npreg is the total number of pregnancies participants reported having or caused during the wave 4 interview. Number of pregnancies was regressed on the total number of sexual partners and days drunk rating levels and slopes in the 'Bivariate Model with Covariate' subtable.

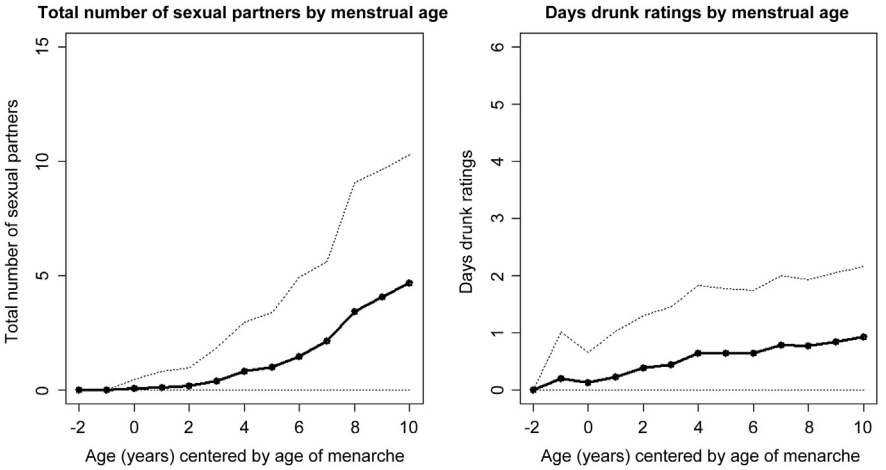

Figure 5 Change in female total sex partners and ratings of days drunk across participant menstrual age. This demonstrates an alternative and developmentally meaningful time metric across which to plot changes in total number of sexual partners and ratings of days spent drunk.

we still estimate an initial level and a slope for both total number of sex partners and drinking behavior, and these estimates can be evaluated for interrelations and predictive ability of exogenous variables.

While estimates may differ from previous models due to the use of an alternative time metric, variations could also be a result of heterogeneity between male and female participants (i.e., average female and male trajectories may differ). In this case, relations between the slopes and levels of total sex partners and drinking behavior were quite similar to the previous age-based model. Differences included a positive relation between the levels for total sex partners and drinking behavior ($r = 0.13$, $p < 0.01$), and a non-significant relation between the level of drinking behavior and the slope of total sex partners ($r = 0.06$, $p = 0.11$). This suggests that females have a positive relation between drinking behavior and number of sexual partners at their age of menarche, but not between their age of menarche drinking behavior and rate of change in total sex partners. Finally, evaluation of our bivariate model with our exogenous

Table 4 Growth model using menstrual age in years as time metric

Total number of sexual partners					*Drinking behavior*				
		Estimate	*SE*	p-*value*			*Estimate*	*SE*	p-*value*
Cor	(Y_0, Y_1)	−0.996	0.002	<0.01	Cor	(Y_0, Y_1)	−0.56	0.02	<0.01
Means	Y_0	−1.76	0.1	<0.01	Means	Y_0	0.48	0.02	<0.01
	Y_1	0.69	0.02	<0.01		Y_1	0.02	0.002	<0.01
Variance	Y_0	9.7	0.77	<0.01	Variance	Y_0	0.52	0.05	<0.01
	Y_1	1.26	0.04	<0.01		Y_1	0.003	<0.001	<0.01
Residual	Tsex	13.16	0.32	<0.01	Residual	Ddrunk	0.77	0.02	<0.01
Bivariate model without covariate					*Bivariate model with covariate*				
Correlations		*Estimate*	*SE*	p-*value*	*Correlations*		*Estimate*	*SE*	p-*value*
Tsex Y_0	Alc Y_0	0.13	0.02	0.03	Tsex Y_0	Alc Y_0	0.09	0.02	0.09
	Alc Y_1	−0.37	0.02	<0.01		Alc Y_1	−0.32	0.02	<0.01
	Tsex Y_1	−0.97	0.005	<0.01		Tsex Y_1	−0.97	0.005	<0.01
Alc Y_0	Alc Y_1	−0.55	0.02	<0.01	Alc Y_0	Alc Y_1	−0.53	0.02	<0.01
	Tsex Y_1	0.06	0.03	0.11		Tsex Y_1	0.07	0.02	0.05
Tsex Y_1	Alc Y_1	0.24	0.02	<0.01	Tsex Y_1	Alc Y_1	0.24	0.02	<0.01
Regressions		*Estimate*	*SE*	p-*value*	*Regressions*		*Estimate*	*SE*	p-*value*
Npreg	Tsex Y_0	–	–	–	Npreg	Tsex Y_0	−0.03	0.01	<0.01
	Alc Y_0	–	–	–		Alc Y_0	0.18	0.07	<0.01
	Tsex Y_1	–	–	–		Tsex Y_1	0.11	0.03	<0.01
	Alc Y_1	–	–	–		Alc Y_1	−10.1	1.56	<0.01

Note: In the bivariate models, the levels and slopes of total number of sexual partners and days drunk ratings were allowed to covary. Y_0 denotes the intercept and Y_1 denotes the slope. Tsex is total number of sex partners, Alc is ratings of days drunk, and Npreg is the total number of pregnancies participants reported having or caused during the wave 4 interview. Number of pregnancies was regressed on the total number of sexual partners and days drunk rating levels and slopes in the 'Bivariate Model with Covariate' subtable.

outcome, number of pregnancies, regressed on slopes and levels yielded near identical estimates as with our previous model using continuous age. This suggests very little change in prediction of the exogenous outcome as a result of the change in time metric or focus on a subpopulation of female participants.

Discrete-Time Survival Analysis

Data Preparation

Survival analysis required the creation of a person-period data set (Singer and Willet, 1991, 1993). This data set contains a series of individual observations indicating sexual debut status, biological sex, and menarche status. The target event is coded as zero for time periods during which sexual debut did not occur, and one for the time period in which it did occur for each person. The records are censored such that after sexual debut occurs for a person, no further observations for that person are included in the data set. Finally, due to the repeated observations across periods, covariates are capable of being modeled as invariant or variant. For all participants, biological sex was coded as invariant for each observation. In contrast, female participants' menarche status was variant, and indicated if female participants' observations were pre- or postmenarche onset.

Male and Female Sexual Debut

First, a baseline survival model of male and female participants' sexual debut was created using year of observation. Figure 6 shows the survival and hazard rates for sexual debut from 12 to 25 years of age. The survival portion of the figure shows the proportional decline of the remaining population who have not had their sexual debut for each year relative to the starting population of participants. The hazard portion of the figure shows the proportion of surviving participants having their sexual debut during each year.

Results indicated that risk of sexual debut began increasing with the onset of adolescence. By 15 years of age, hazard rates increased to 18% with 23% of participants having reported having their sexual debut. By 16 years of age 51% of participants reported having their sexual debut, and by 18 years of age 78% of participants had their sexual debut. Note that estimates for 25 years of age were not significant (see Table 5). This may be due to data sparseness at the more extreme age of target behavior occurrence. As is expected, this suggests that estimates may be less robust where there are fewer data points.

A second model of sexual debut was developed as an extension of the previous model with the addition of a time-invariant covariate – biological sex. The variable was dummy coded with male as the reference (coded as zero), and female as the change in value (coded as one). The parameter estimate for the 'Female' covariate was not significant (see Table 5). Figure 7 depicts this difference in survival and hazard rates where the solid line with circles indicates the projected rates for participants who are male and the dashed line with pluses indicates the projected rates for participants who are female. Note the almost perfect overlap between lines indicating the lack of a biological sex effect. These models were compared for improvement in fit using change in -2 log likelihood evaluated with a chi square distribution. Given the lack of effect of biological sex, the significance of improved fit between models was unexpected ($\Delta\chi^2 = 7.8$, $\Delta df = 1$, $p < 0.01$).

Female Sexual Debut and Menarche

A third survival analysis model was fit exclusively to data from females. As with the male and female survival analyses, a baseline survival model was fit to the data (see Table 6). As expected given the previous comparison between males and females, the female-only baseline model was essentially identical to the previous models. This time we expanded on the baseline model by including a time-varying covariate – participants' menarche status. Menarche status is time-varying,

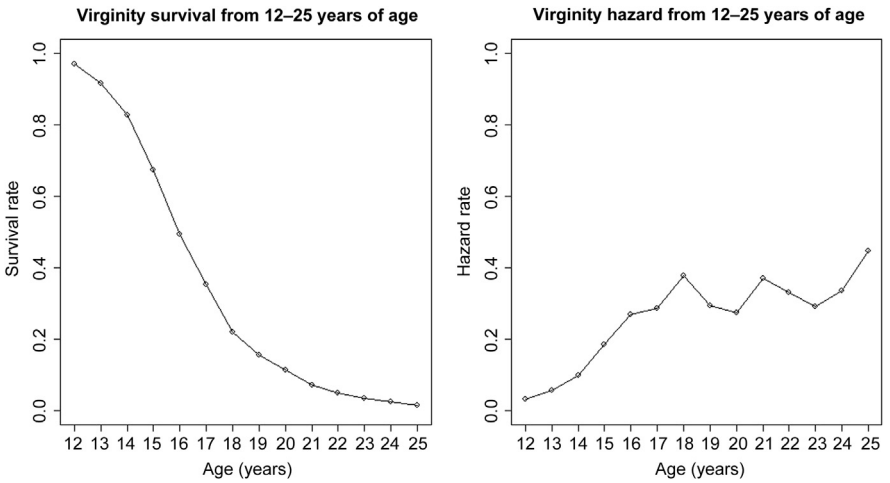

Figure 6 Virginity survival and hazard rates for participants from 12 to 25 years of age. Rate of decline in survival accelerated as participants entered the teen years with 23% of participants having engaged in first-time vaginal intercourse by the age of 15. It is also at age 15 that the hazard rate rises above 20%, which it will not drop below for the remainder of observations. The peak surges in hazard are at ages 18, 21, and 25.

Table 5 Male and female sexual debut hazard estimates

	No covariate			*Time-invariant covariate*		
Parameter	*Estimate*	*SE*	p-*value*	*Estimate*	*SE*	p-*value*
y_{12}	−3.48	0.10	<0.01	−3.50	0.10	<0.01
y_{13}	−2.82	0.07	<0.01	−2.85	0.08	<0.01
y_{14}	−2.22	0.06	<0.01	−2.25	0.06	<0.01
y_{15}	−1.49	0.05	<0.01	−1.52	0.05	<0.01
y_{16}	−1.00	0.05	<0.01	−1.03	0.05	<0.01
y_{17}	−0.91	0.05	<0.01	−0.94	0.06	<0.01
y_{18}	−0.50	0.06	<0.01	−0.53	0.06	<0.01
y_{19}	−0.88	0.08	<0.01	−0.90	0.08	<0.01
y_{20}	−0.98	0.09	<0.01	−1.02	0.10	<0.01
y_{21}	−0.54	0.10	<0.01	−0.56	0.10	<0.01
y_{22}	−0.70	0.13	<0.01	−0.73	0.13	<0.01
y_{23}	−0.90	0.17	<0.01	−0.92	0.17	<0.01
y_{24}	−0.68	0.19	<0.01	−0.71	0.19	<0.01
y_{25}	−0.22	0.22	0.32	−0.25	0.22	0.26
Female	–	–	–	0.05	0.04	0.22

Note: −2LL (2 Log Likelihood) for the 'No Covariate' model was 17 458.59, and −2LL for the 'Time-Invariant Covariate' model was 17 450.74. Their difference is 7.84 with 1 d*f*, which is a significant change in fit ($p < 0.01$). Parameters y_{12}–y_{25} are the dummy coded period variables. The nonsignificant estimate for 25 years of age may be a consequence of sparse data, suggesting that greatest confidence be placed in describing hazard and survival rates in late childhood through early adulthood.

as participants may begin observations having not experienced menarche (coded as zero) and may change to having experienced menarche (coded as one) prior to the occurrence of the target event and data censorship.

Results indicate that having experienced menarche significantly predicts occurrence of sexual debut. Female participants are 1.57 times more likely to have their sexual debut during a time period once they have experienced menarche compared to time periods prior to menarche. Figure 8 depicts this difference in survival and hazard rates. The solid line with circles indicates the projected rates for participants who have not experienced menarche and the dashed line with pluses indicates the projected rates for participants who have experienced menarche.

Comparisons of model fit show that the second model including the menarche status covariate fit the data significantly better than the baseline model ($\Delta\chi^2 = 23.82$, $\Delta df = 1$, $p < 0.001$).

Discussion

Summary of Findings

In this article, we used LGM to describe changes in the number of sexual partners, drinking behavior, their interrelations, and their power to predict number of pregnancies. Using measurement occasion as the time metric in our analyses, we concluded that neither number of sexual partners nor drinking behavior could best be described as changing linearly or quadratically over time. Rather, we found that increases in total number of sexual partners was greatest from waves 2 to 3 and 3 to 4, and that drinking behavior peaked in wave 3 and then declined slightly in wave 4. We also found that change in drinking behavior and number of sexual partners was not significantly related, but that their levels and slopes were. This suggests that there is an initial relation and effect on rate of change, but that overall change is not related. Finally, we determined that drinking behavior seems to be prohibitive to individuals' total number of pregnancies, but that number of sexual partners is a positive predictor of total number of pregnancies. It is not clear if the prohibitive effect of drinking has to do with decreasing the number of planned pregnancies with committed partners, nor is it clear if the positive relation between the number of sexual partners and pregnancies is indicative of pregnancies during uncommitted or committed relationships that were planned or unplanned.

In a second set of analyses we used more substantive time metrics (i.e., participants' age and female participants' menstrual age). We found that with the recoding of the time metric interrelations between the level and slope of total

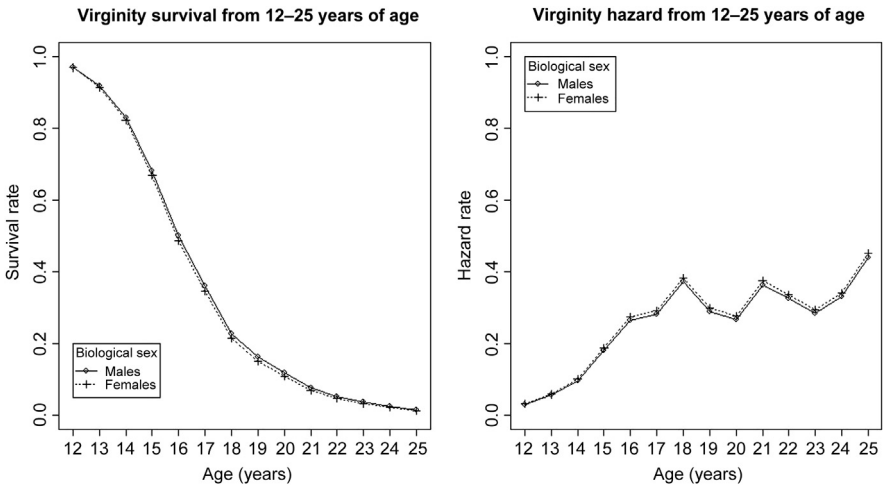

Figure 7 Comparison of male and female virginity survival and hazard rates for participants from 12 to 25 years of age. There was no significant difference in model fit with the inclusion of the 'Female' covariate. As a result, there is virtually perfect overlap between male and female survival and hazard rates.

Table 6 Female sexual debut hazard estimates

	No covariate			*Time-varying covariate*		
Parameter	*Estimate*	*SE*	p-*value*	*Estimate*	*SE*	p-*value*
y_{12}	−3.77	0.15	<0.01	−4.06	0.16	<0.01
y_{13}	−2.84	0.10	<0.01	−3.21	0.13	<0.01
y_{14}	−2.18	0.08	<0.01	−2.59	0.12	<0.01
y_{15}	−1.37	0.06	<0.01	−1.79	0.11	<0.01
y_{16}	−0.94	0.06	<0.01	−1.36	0.11	<0.01
y_{17}	−0.89	0.07	<0.01	−1.31	0.11	<0.01
y_{18}	−0.50	0.08	<0.01	−0.92	0.12	<0.01
y_{19}	−0.85	0.10	<0.01	−1.28	0.14	<0.01
y_{20}	−1.05	0.13	<0.01	−1.46	0.16	<0.01
y_{21}	−0.52	0.14	<0.01	−0.94	0.16	<0.01
y_{22}	−0.88	0.18	<0.01	−1.30	0.20	<0.01
y_{23}	−1.22	0.24	<0.01	−1.63	0.25	<0.01
y_{24}	−0.75	0.24	<0.01	−1.16	0.26	<0.01
y_{25}	−0.11	0.28	0.68	−0.51	0.29	0.08
Menarche	–	–	–	0.45	0.10	<0.01

Note: −2LL (2 Log Likelihood) for the 'No Covariate' model was 9763.84, and −2LL for the 'Time-Varying Covariate' model was 9740.02. Their difference is 23.82 with 1 d*f*, which is a significant change in fit ($p < 0.01$). Parameters y_{12}–y_{25} are the dummy coded period variables. The nonsignificant estimate for 25 years of age may be a consequence of sparse data, suggesting that greatest confidence be placed in describing hazard and survival rates in late childhood through early adulthood.

number of sex partners and drinking behavior differed. Using the wave-based metric the correlation between the two slopes was not significant, but this relation was significant when using both the age-based and menstrual age-based metrics. Additionally, there was some disagreement between the age-based and menstrual age-based models with respect to the relation between the two levels ($r = -0.10$ and 0.13, respectively). But it is not clear if this is a result of change in time metric, heterogeneity between males and females, or both.

Finally, we used discrete-time series analysis to examine the likelihood of sexual debut during adolescence and early adulthood. Our findings indicated that the risk of sexual debut begins to increase in early adolescence for both males and females. By the age of 15, 23% of the participants in our sample had their sexual debut. By 16, that proportion more than doubled to 51%. We also found that there was no significant difference in hazard and survival rates between males and females in our sample, but model fit improved significantly when the covariate was included.

However, when applying the same model to females only, we found that having experienced menarche increased the odds of having their sexual debut by 1.57 times. While it is not clear what may be driving this effect, it is likely that the onset of ovulation has a direct effect on females' own interest in, and attraction to, romantic partners and sexual activity. Additionally, it is possible that postmenarche females are physically more developed than their premenarche peers, making them more desirable partners to males.

Substantive Implications

An overarching implication of our results is that the conclusions one can reach about the relation between drinking behavior and number of sexual partners will vary depending on how the question is asked, and consequently analyzed. While there are scenarios in which the wave metric can be of interest (e.g., evaluating practice effects), in our illustration such effects are unlikely and evaluations of changes over age and menstrual age are more meaningful to understanding developmental change. Estimates from the continuous age model suggest that there is a positive relation between changes in drinking behavior and number of sexual partners over this period.

These models do not allow inferences about relations between these behaviors beyond the observed range of ages. However, over the developmental period studied, we observed that participants who had greater changes in drinking behavior had corresponding changes in their number of sexual partners. Similarly, while this relation is often expected given single-observation research, this model suggests that this is a relation that may exist within individuals over time, such

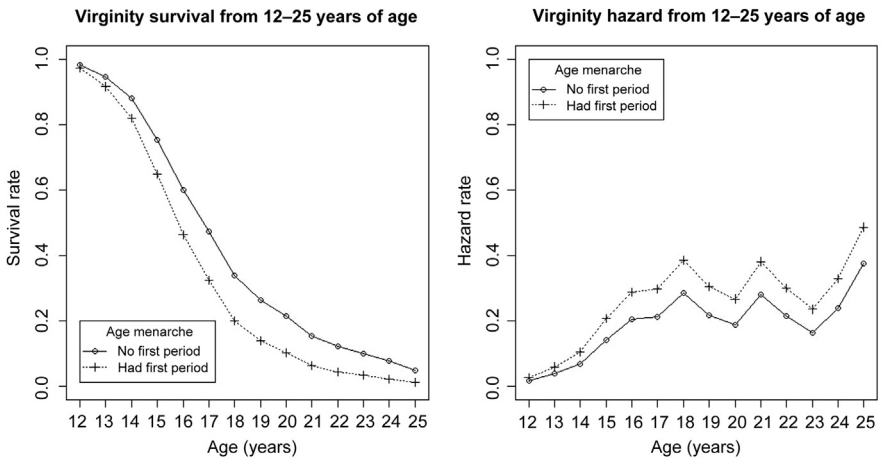

Figure 8 Comparing female virginity survival and hazard rates based on age of menarche. There was a significant difference in model fit with the inclusion of the 'age of menarche' covariate. Females were 1.57 times as likely to engage in first-time vaginal intercourse if they had experienced menarche compared to their peers who had not. This difference emerges in late childhood and early adolescence, becomes pronounced around ages 15–16, and is maintained through the remaining observation periods.

that changes in drinking behavior over time are related to the total number of sexual partners over time as well.

Our survival analysis indicates that there was relatively little difference between male and female participants with respect to risk of sexual debut. This suggests that the optimum timing for sexual education may not differ between male and female adolescents. Additionally, depending on the education strategy to be employed, these data may also indicate periods for optimum repetition of education with respect to increasing risk of sexual debut.

Methodological Limitations

The sexual debut data do not represent yearly observations. Rather, it was derived from a single observation for each participant. This may be problematic given potential inaccuracy due to retroactive reporting. For example, participants may retroactively alter the age of their sexual debut, or recall the timing differently than they would have around the time of the actual event.

Generally, this approach has two limitations with respect to the length of the time periods and estimation of covariate effects. Regarding the former, we were able to use time in years because that was the precision of time intervals inherent in the event record. However, this may not be the best time metric to describe occurrence of sexual debut. As for the latter, the effect of covariates is implicitly modeled as constant across time periods. However, this assumption may not be reasonable. For example, postmenarche status likely impacts 16-year-old females differently than 21-year-old females.

Conclusions

Overall, this article is meant to serve as an example of how longitudinal data can be analyzed using discrete and continuous time analyses aimed at answering questions about intra-individual change over time. Specifically, our objective was to demonstrate how developmental processes can be differently described and understood via a series of similar analytic techniques, such that small variations in modeling can address a breadth of substantive research questions. While these examples are by no means exhaustive, they demonstrate well how the same modeling framework can be used to approach questions about longitudinal data in several substantively different ways.

See also: Alcohol Use among Young People; Cross-Sectional Methodologies in Developmental Psychology; Disability: Social and Psychological Aspects; Hierarchical Models: Random and Fixed Effects; Hooking up and Casual Sex; Longitudinal Data: Event–History Analysis in Discrete Time; Methodological Issues in the Study of Sexuality; Sexual Debut; Sexual Risk Behaviors; Sexuality over the Life Course; Survival Analysis: Introduction.

Bibliography

Baltes, P.B., Nesselroade, J.R., 1979. History and rationale of longitudinal research. In: Nesselroade, J.R., Baltes, P.B. (Eds.), Longitudinal Research in the Study of Behavior and Development. Academic Press, New York, pp. 1–39.

Chandra, A., Martinez, G.M., Mosher, W.D., Abma, J.C., Jones, J., 2005. Fertility, family planning, and reproductive health of U.S. women: data from the 2002 national survey of family growth. Vital and Health Statistics, vol. 23, no. 25. National Center for Health Statistics, Hyatsville, MD.

Coker, A.L., Richter, D.L., Valois, R.F., McKeown, R.E., Garrison, C.Z., Vincent, M.L., 1994. Correlates and consequences of early initiation of sexual intercourse. Journal of School Health 64, 372–377. http://dx.doi.org/10.1111/j.1746-1561.1994.tb06208.x.

Dent, C.W., Grube, J.W., Biglan, A., 2005. Community level alcohol availability and enforcement of possession laws as predictors of youth drinking. Preventive Medicine 40 (3), 355–362. http://dx.doi.org/10.1016/j/ypmed.2004.06.014.

Dogan, S.J., Stockdale, G.D., Widaman, K.F., Conger, R.D., 2010. Developmental relations and patterns of change between alcohol use and number of sexual partners from adolescence through adulthood. Developmental Psychology 46 (6), 1747–1759. http://dx.doi.org/10.1037/a0019655.

Ferrer, E., Hamagami, F., McArdle, J.J., 2004. Modeling latent growth curves with incomplete data using different types of structural equation modeling multilevel software. Structural Equation Modeling 11 (3), 452–483.

Ferrer, E., Salthouse, T.A., McArdle, J.J., Stewart, W.F., Schwartz, B.S., 2005. Multivariate modeling of age and practice in longitudinal studies of cognitive abilities. Psychology and Aging 20, 412–422.

Grello, M., Welsh, D., Harper, M., 2006. No strings attached: the nature of casual sex in college students. The Journal of Sex Research 43 (3), 255–267.

Gute, G., Eshbaugh, E.M., 2008. Personality as a predictor of hooking up among college students. Journal of Community Health Nursing 25, 26–43. http://dx.doi.org/10.1080/07370010701836385.

Harris, K.M., Udry, J.R., 2013. National Longitudinal Study of Adolescent Health (Add Health), 1994–2008. ICPSR21600-v11. Inter-university Consortium for Political and Social Research (Distributor), Ann Arbor, MI. http://dx.doi.org/10.3886/ICPSR21600.v11.

Martinez, G.M., Chandra, A., Abma, J.C., Jones, J., Mosher, W.D., 2006. Fertility, contraception, and fatherhood: data on men and women from cycle 6 (2002) of the national survey of family growth. Vital and Health Statistics, vol. 23, no. 26. National Center for Health Statistics, Hyatsville, MD.

McArdle, J.J., 1988. Dynamic but structural equation modeling of repeated measures data. In: Nesselroade, J.R., Cattell, R.B. (Eds.), The Handbook of Multivariate Experimental Psychology, second ed. Plenum Press, New York, NY, pp. 561–614.

McArdle, J.J., 2009. Latent variable modeling of differences and changes with longitudinal data. Annual Review of Psychology 60, 577–605. http://dx.doi.org/10.1146/annurev.psych.60.110707.163612.

McArdle, J.J., Hamagami, F., 2001. Latent difference score structural models for linear dynamic analysis with incomplete longitudinal data. In: Collins, L.M., Sayer, A.G. (Eds.), New Methods for the Analysis of Change. American Psychological Association, Washington, DC, pp. 139–175.

Miller, K.E., 2012. Alcohol mixed with energy drink use and sexual risk-taking: casual, intoxicated, and unprotected sex. Journal of Caffeine Research 2 (2), 62–69. http://dx.doi.org/10.1089/caf.2012.0015.

Paul, E.L., McManus, B., Hayes, A., 2000. "Hookups": characteristics and correlates of college students' spontaneous and anonymous sexual experiences. The Journal of Sex Research 37 (1), 76–88.

Singer, J.D., Willet, J.B., 1991. Modeling the days of our lives: using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. Psychological Bulletin 110 (2), 268–290.

Singer, J.D., Willet, J.B., 1993. It's about time: using discrete-time survival analysis to study duration and the timing of events. Journal of Educational Statistics 18 (2), 155–195.

Somers, C.L., Surmann, A.T., 2005. Sources and timing of sex education: relations with American adolescent sexual attitudes and behavior. Educational Review 57 (1), 37–54. http://dx.doi.org/10.1080/0013191042000274178.

Traeen, B., Lewin, B., 1992. Casual sex among Norwegian adolescents. Archives of Sexual Behavior 21 (3), 253–269.

Longitudinal Causal Inference

Miguel A Hernán and James M Robins, Harvard School of Public Health, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article introduces fundamental concepts for making inferences about treatments with values that may change over time. We define time-varying treatments and confounders, and provide a classification of treatment strategies. We also describe key identifiability conditions, and the bias of traditional adjustment methods when the time-varying confounders are affected by previous treatment.

Time-Varying Treatments

Many causal questions involve treatments that vary over time (*see* Longitudinal Data). For example, we may be interested in estimating the causal effects of medical treatments, lifestyle habits, employment status, marital status, occupational exposures, etc. Because these treatments may take different values for a single individual over time, we refer to them as time-varying treatments. Robins (1986, 1987, 1997) proposed a general counterfactual model that permits the estimation of total and direct effects of fixed *and* time-varying treatments in longitudinal studies, whether randomized or observational (*see* Observational Studies: Overview).

Unlike causal contrasts for fixed treatments, causal contrasts that involve time-varying treatments need to incorporate time explicitly. To see this, first consider a fixed (non-time-varying) dichotomous treatment A that takes value 1 if an individual i receives treatment and 0 otherwise. (We assume subjects are independent and identically distributed and thus suppress the i subscript in A_i) The average causal effect of A on the outcome Y is defined as the contrast between the mean counterfactual outcome $Y^{a=1}$ under treatment and the mean counterfactual outcome $Y^{a=0}$ under no treatment, e.g., $\mathrm{E}(Y^{a=1}) - \mathrm{E}(Y^{a=0})$ (*see* Causal Inference: Overview). That is, questions about the average causal effect are questions about whether the mean under treatment differs from the mean under no treatment.

Now consider a time-varying dichotomous treatment A_k that takes value 1 if an individual receives treatment at time k and 0 otherwise. We use an overbar to denote treatment history, that is, $\overline{A}_k = (A_0, A_1, \ldots A_k)$ is the history of treatment through time k, where k can take values 0, 1, 2…K. When we refer to the entire treatment history through K, we often represent $\overline{A}_K$ as $\overline{A}$ (without a time subscript). The average causal effect of the time-varying treatment $\overline{A}$ on the outcome is not uniquely defined because there are many possible contrasts of interest that depend on the values of each of the components of $\overline{A}$. In particular, the average causal effect of $\overline{A}$ cannot be defined as a contrast between the mean counterfactual outcome under treatment and the mean counterfactual outcome under no treatment, because such contrast does not specify the time of treatment. Do we mean treatment at time 0? At time 3? At all times between 0 and K?

As an example, consider an observational 5-year follow-up study of individuals infected with the human immunodeficiency virus (HIV) who have never received antiretroviral therapy A_k before the start of the study at time 0, i.e., $A_{-1} = 0$ for all individuals. The time subscript k for treatment A_k can take 60 values: 0, 1, 2…59 months. An individual who receives treatment continuously throughout the follow-up has treatment history $\overline{A} = (A_0 = 1, A_1 = 1, \ldots A_{59} = 1) = (1, 1, \ldots 1)$, or $\overline{A} = \overline{1}$. Analogously, an individual who never receives treatment during the follow-up has treatment history $\overline{A} = (0, 0, \ldots 0) = \overline{0}$. Most individuals are treated only during part of the follow-up, and therefore have intermediate treatment histories with some 1s and some 0s – which we cannot represent as compactly as $\overline{1}$ and $\overline{0}$. For simplicity, let us assume that no individuals are lost to follow-up or die during this period, and that all variables are perfectly measured.

Now let $Y_{K+1} = Y_{60}$ be the concentration of HIV in the blood – the viral load – at the end of follow-up. For simplicity of notation we will refer to Y_{K+1} as Y. We would like to know whether treatment $\overline{A}$ affects the mean of Y. But we can no longer define the average causal effect of a time-varying treatment as a contrast of the sort $\mathrm{E}(Y^{a_k=1}) - \mathrm{E}(Y^{a_k=0})$, because such contrast only involves treatment at a single time k. Rather, we will have to define the average causal effect as a contrast between the counterfactual mean outcomes under two treatment strategies that involve treatment at all times between the start ($k = 0$) and the end ($K = 60$) of the follow-up.

Treatment Strategies

A treatment strategy – also referred to as a treatment plan or a treatment regime – is a rule that assigns treatment to each individual at each time k. In our HIV example, two treatment strategies are 'always treat' and 'never treat' during the follow-up. The strategy 'always treat' is represented by $\overline{a} = (1, 1, \ldots 1) = \overline{1}$, and the strategy 'never treat' is represented by $\overline{a} = (0, 0, \ldots 0) = \overline{0}$. Following conventional statistical terminology, we use lowercase letters to denote possible realizations of a random variable, e.g., a_k is a possible realization of treatment A_k.

We can now define an average causal effect of $\overline{A}$ on the outcome $Y_{K+1} = Y$ as the contrast between the mean counterfactual outcome $Y^{\overline{a}=\overline{1}}$ under the strategy 'always treat' and the mean counterfactual outcome $Y^{\overline{a}=\overline{0}}$ under the strategy 'never treat,' e.g., $\mathrm{E}(Y^{\overline{a}=\overline{1}}) - \mathrm{E}(Y^{\overline{a}=\overline{0}})$. There are many other possible causal effects for the time-varying treatment $\overline{A}$, each of them defined by a contrast of outcomes under two particular treatment strategies. For example, we might be

interested in the average causal effect defined by the contrast $E(Y^{\bar{a}}) - E(Y^{\bar{a}'})$ that compares the strategy 'treat at every other month' = (1, 0, 1, 0…) with the strategy 'treat at all months except the first one' $\bar{a}' = (0, 1, 1, 1\ldots)$. There are 2^K possible combinations of values rule $(a_0, a_1, \ldots a_K)$ for dichotomous a_k, and therefore 2^K such treatment strategies. However as we next explain these 2^K such strategies do not exhaust all possible treatment strategies.

In our HIV example, the time-varying covariate CD4 cell count, L_k, is measured (in cells per μl) each month k in all individuals. A low CD4 cell count is an indicator of bad prognosis among HIV-infected individuals. We could then consider the strategy "do not treat before CD4 cell count drops below 500, always treat after CD4 drops below 500, regardless of future CD4 cell counts." This treatment strategy is different from the ones considered in the previous paragraph because we cannot represent it by a rule $\bar{a} = (a_0, a_1, \ldots a_K)$ under which all individuals get treatment a_0 at time $k = 0$, a_1 at time $k = 1$, etc. Now, at each time, some individuals will be treated and others will be untreated. This is an example of a *dynamic treatment strategy*, a rule in which the treatment a_k at time k depends on the evolution of an individual's history of time-dependent covariate(s) $\bar{L}_k$ and possibly on treatment $\bar{A}_{k-1}$. Strategies $\bar{a}$ for which treatment does not depend on the evolution of time-dependent covariates we refer to as nondynamic or *static treatment strategies*.

Formally, a dynamic strategy is a rule $g = [g_0(\bar{a}_{-1}, \bar{l}_0), g_1(\bar{a}_0, \bar{l}_1), \ldots g_K(\bar{a}_{K-1}, \bar{l}_K)]$, where $g_k(\bar{a}_{k-1}, \bar{l}_k)$ is a function that specifies the treatment assigned at k to an individual with past history $(\bar{a}_{k-1}, \bar{l}_k)$. In our HIV example, $g_k(\bar{a}_{k-1}, \bar{l}_k)$ is 1 if an individual's $\bar{l}_k$ implies that his CD4 cell count dropped below 500 at or before k; otherwise $g_k(\bar{a}_{k-1}, \bar{l}_k)$ is 0. To make the notation uniform, we will also refer to static strategies as g, with the understanding that the functions $g_k(\bar{a}_{k-1}, \bar{l}_k)$ for static strategies do not depend on the covariate history $\bar{l}_k$.

All of the strategies described above are *deterministic treatment strategies* because they assign a particular value of treatment (0 or 1) to each individual at each time. More generally, we could consider *random treatment strategies* that do not assign a particular value of treatment, but rather a probability of receiving a treatment value. Random treatment strategies can be static (e.g., "independently at each month, treat a subject with probability 0.3 and do treat with probability 0.7") or dynamic (e.g., "independently at each month treat subjects whose CD4 is currently below 500 with probability 0.3, but do not treat subjects with current CD4 count exceeding 500"). See Young et al., (2014) for a taxonomy of treatment strategies.

In summary, causal inference with time-varying treatments involves the contrast of counterfactual outcomes under two or more treatment strategies. The average causal effect of a time-varying treatment is not defined unless we define the strategies that are being compared. In our HIV example, the difference $E(Y^g) - E(Y^{g'})$ measures the average causal effect of strategy g (say, 'always treat') versus strategy g' (say, 'never treat'), whereas the difference $E(Y^g) - E(Y^{g''})$ measures the average causal effect of strategy g ('always treat') versus strategy g'' (say, "treat after CD4 drops below 500"). There is not a single definition of average causal effect for time-varying treatments.

Time-Varying Confounders

Consider again our follow-up study of the effect of antiretroviral therapy $\bar{A}$ on viral load Y at the end of follow-up in HIV-infected subjects. The time-varying covariate CD4 cell count, L_k, is a proxy for the true, but unmeasured and unknown, immunosuppression level U. The greater a subject's *unmeasured* immunosuppression level U, the greater her viral load Y and the lower the CD4 cell count L_k. Every month k, a treatment decision A_k is made based on the value of CD4 cell count L_k. The lower the CD4 cell count, the more likely a patient is to be treated. Note that both the variable U and the outcome Y could be time varying too, but that is irrelevant to our current discussion. (Time-varying indicators of failure would be needed to represent a survival analysis setting (*see* Survival Analysis: Introduction) in which the outcome is, say, death.)

The causal diagram (Spirtes et al., 1993; Pearl, 1995) in Figure 1 depicts the first two time points of our study. To avoid clutter, let us assume that, in truth but unknown to the investigators, treatment has no causal effect on the outcome of any individual, i.e., the sharp null hypothesis holds and hence there are no causal paths from any of the treatment nodes A_k to the outcome Y. Also to avoid clutter, we have omitted the node L_0 for baseline CD4 cell count. (Baseline treatment A_0, and perhaps subsequent treatment, would generally be decided after observing L_0, but the inclusion of L_0 in the causal diagram would not add any conceptual insights to the discussion in this article.)

Under the structure depicted in Figure 1, we say that there is confounding for the effect of the treatment A_1 on the outcome Y because A_1 and Y share the cause U, i.e., because there is a backdoor path between A_1 and Y through U. This backdoor path $A_1 \leftarrow L_1 \leftarrow U \rightarrow Y$ cannot be blocked by conditioning on the common cause U because U is unmeasured and therefore unavailable to the investigators. However, this backdoor path can be blocked by conditioning on L_1, which is measured. Note that a second backdoor path gets open after conditioning on L_1. This second backdoor path $A_1 \leftarrow A_0 \rightarrow L_1 \leftarrow U \rightarrow Y$ can be blocked by conditioning on prior treatment A_0, which should be available to investigators. Thus, if the investigators collected data on L_1 for all individuals, there would be no unmeasured confounding for the effect of A_1 given L_1 and A_0. We say that L_1 is a confounder for the effect of A_1 because it is needed to eliminate confounding, even though the actual common cause of A_1 *and* Y was the unmeasured U.

Now imagine the very long causal diagram that contains all time points k. At each time k, the covariate history $\bar{L}_k$ will be needed, together with the treatment history $\bar{A}_{k-1}$, to block the backdoor paths between the treatment A_k (and perhaps future treatment A_t, for $t > k$) and the outcome Y. We say that

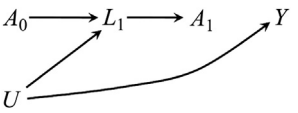

Figure 1 Causal diagram of a study with time-varying treatment A_k and confounders L_k, and with treatment–confounder feedback. For simplicity, only the first two time points are shown.

the time-varying covariates in $\overline{L}_k$ are *time-varying confounders* for the effect of the treatment strategy g on Y because $\overline{L}_k$ includes postbaseline covariates that are needed to eliminate confounding at several, or all, times k in the study. To validly estimate the effect of a treatment strategy g without bias, we need to be able to validly estimate the joint effect of all A_k simultaneously and without bias (Robins, 1986). It can be shown that, for static regimes, blocking the backdoor path for every component k of the time-varying treatment A_k is a sufficient (but not necessary) condition for valid estimation of the causal effect of treatment strategies g that involve the time-varying treatment A_k (Pearl and Robins, 1995).

Suppose treatment A increases future CD4 cell count, and the presence of low CD4 count increases the probability of receiving treatment. That is, in our example there is a feedback loop between treatment and confounders. Interestingly, a causal diagram like the one in **Figure 1** – an acyclic graph – can be used to represent the treatment–confounder feedback loop – a cycle – by elaborating the feedback loop in time. This approach to visually representing feedback cycles with acyclic graphs requires that time be represented as a discrete variable. That is, the causal diagram implies that treatment and covariates only change at each prespecified time k from 0 to K. This forced discretization of time is not a limitation in practice: the length of the intervals k can be chosen to be as short as the granularity of the data requires. For example, in a study where individuals see their doctors once per month or less frequently (as in our HIV example), time may be safely discretized into month intervals. In other cases, year intervals or day intervals may be more appropriate.

The treatment–confounder feedback, however, is not a necessary feature of time-dependent confounders. That is, the time-varying covariates $\overline{L}_k$ would be time-varying confounders even if they were not affected by prior treatment. We chose to present a causal diagram with treatment–confounder feedback because that setting creates an interesting problem for causal inference: when the time-varying confounders $\overline{L}_k$ are affected by prior treatment, confounding adjustment cannot be handled by conventional methods to adjust for confounding (Robins, 1986). Rather, g-methods are needed to appropriately adjust for time-varying confounders affected by prior treatment. We explore this problem below. First, we describe the identifiability conditions for the effects of time-varying treatments.

Sequential Exchangeability and Identifiability Conditions

Making causal inferences about fixed treatments requires measuring and adjusting for a set of covariates L – informally, the confounders – required to achieve conditional exchangeability $Y^a \perp\!\!\!\perp A|L$. This condition holds in conditionally randomized experiments in which individuals are assigned treatment with a probability that depends only on the values of their covariates L. The condition also holds in observational studies in which the probability of receiving treatment depends on the measured covariates L and, conditional on L, does not depend on any unmeasured, independent predictors U of the outcome. When exchangeability $Y^a \perp\!\!\!\perp A|L$ holds, we can obtain consistent estimates of the average causal effect of treatment via any methods that appropriately adjust for the variables in L, such as standardization and inverse probability (IP) weighting.

Similarly, making causal inferences about treatment strategies requires measuring and adjusting for the time-varying covariates $\overline{L}_k$ – the time-varying confounders – required to achieve a sequential version of conditional exchangeability that encompasses all time points in the study. This condition holds in (ideal) sequentially randomized experiments in which individuals are assigned treatment at each time k with a probability that depends only on the values of their prior covariate history $\overline{L}_k$ and treatment history $\overline{A}_{k-1}$. The condition also holds in observational studies in which the probability of receiving treatment at each time depends on their treatment and covariate history $(\overline{A}_{k-1}, \overline{L}_k)$ and, conditional on this history, does not depend on any unmeasured predictors U of the outcome.

In the HIV example summarized in **Figure 1**, this sequential exchangeability holds because there are no arrows into treatment A_1 from any unmeasured cause U of the outcome: all arrows into treatment A_1 come from prior treatment A_0 and covariates L_1. That is, **Figure 1** represents a sequentially randomized trial in which treatment A_0 is marginally randomized and treatment A_1 is conditionally randomized given A_0 and L_1. The treated and the untreated at time 1 are exchangeable conditional on A_0 and L_1. Sequential exchangeability – the combination of $Y^g \perp\!\!\!\perp A_0$ and $Y^g \perp\!\!\!\perp A_1|L_1, A_0$ – holds for all treatment strategies g that involve the time-varying treatment (A_0, A_1).

Now imagine the very long causal diagram with many time points $k = 0, 1 \ldots K$ that represents a sequentially randomized trial in which treatment A_k at each time k is randomly assigned conditional on prior treatment and covariate history $(\overline{A}_{k-1}, \overline{L}_k)$. The causal diagram would include no direct arrows from any unmeasured risk factors U into treatment A_k at any time k. In this sequential randomized trial, the treated and the untreated at each time k are exchangeable conditional on prior covariate history $\overline{L}_k$ and treatment history $\overline{A}_{k-1}$. That is, sequential exchangeability $Y^g \perp\!\!\!\perp A_k|\overline{L}_k, \overline{A}_{k-1}$ holds for all times $k = 0, 1 \ldots K$, and we say that there are no unmeasured time-varying confounders. Under sequential exchangeability, we can obtain consistent estimates of the average causal effect of a treatment strategy via any methods that appropriately adjust for history $(\overline{A}_{k-1}, \overline{L}_k)$, such as standardization and IP weighting. The condition of sequential exchangeability $Y^g \perp\!\!\!\perp A_k|\overline{L}_k, \overline{A}_{k-1}$ can be slightly weakened. We can use single world intervention graphs (Richardson and Robins, 2013) to graphically derive these weaker sequential exchangeability conditions for static and dynamic treatment regimes.

Causal inference for time-varying treatments also requires a generalization of the positivity condition from the fixed version 'if $f_L(l) \neq 0$, $f_{A|L}(a|l) > 0$ for all a' to the sequential version 'if $f_{\overline{A}_{k-1}, \overline{L}_k}[\overline{a}_{k-1}, \overline{l}_k] \neq 0$, then $f_{A_k|\overline{A}_{k-1}, \overline{L}_k}[a_k|\overline{a}_{k-1}, \overline{l}_k] > 0$ for all a_k.' The definition of consistency of counterfactuals can also be generalized in a sequential fashion from the fixed version 'if $\overline{A} = \overline{a}$ for a given subject, then $Y^{\overline{a}} = Y$ for that subject' to the sequential version '$Y^{\overline{a}} = Y^{\overline{a}*}$ if $\overline{a}^* = \overline{a}, Y^{\overline{a}} = Y$ if $\overline{A} = \overline{a}, \overline{L}_k^{\overline{a}} = \overline{L}_k^{\overline{a}*}$ if $\overline{a}_{k-1}^* = \overline{a}_{k-1}$, $\overline{L}_k^{\overline{a}} = \overline{L}_k$

if $\overline{A}_{k-1} = \overline{a}_{k-1}$,' where $\overline{L}_k^{\overline{a}}$ is the counterfactual L-history through time k under regime $\overline{a}$. As always, causal inference requires well-defined interventions. Otherwise, a meaningful comparison of the treatment strategies is not possible.

Unfortunately, sequential exchangeability and positivity are not guaranteed to hold in observational studies. Achieving approximate exchangeability requires expert knowledge, which will guide investigators in the design of their studies to measure as many variables in $\overline{L}_k$ as possible. For example, in an HIV study, experts would agree that time-varying variables like CD4 cell count, viral load, and other joint predictors of treatment and outcome need to be measured and appropriately adjusted for. However, in observational studies the condition of sequential exchangeability cannot be empirically tested. Like for fixed treatments, causal inference for time-varying treatments requires the untestable assumption of conditional exchangeability – only now sequentially during the follow-up rather than at baseline only.

The Bias of Traditional Methods for Time-Varying Treatments

Let us assume for now that sequential exchangeability and positivity hold in our observational HIV study. That is, the measured time-varying confounders are sufficient to validly estimate the average causal effect of any treatment strategy. Then the question is what confounding adjustment method to use. The answer to this question highlights a key problem in causal inference about time-varying treatments: treatment–confounder feedback. When time-varying confounders are affected by prior treatment, traditional adjustment methods – stratification, regression, matching – may introduce bias. That is, even if we had all the information required to validly estimate the average causal effect of a treatment strategy, we would be generally unable to do so unless we used g-methods – inverse probability weighting, g-estimation, g-formula. The bias of traditional, stratification-based methods for time-varying treatments is conceptually identical to the bias of traditional methods for direct effects.

To illustrate this problem, consider again the HIV study depicted in Figure 1. Under this causal diagram, CD4 cell count L_1 is sufficient to block all backdoor paths between treatment and outcome (there is only one such path), so sequential exchangeability holds. Say that we want to estimate the average treatment effect of treatment strategy 'always treat,' $g = (a_0 = 1, a_1 = 1)$, compared with treatment strategy 'never treat,' $g' = (a_0 = 0, a_1 = 0)$ on the outcome Y. According to Figure 1, the true, but unknown to the investigator, average causal effect $E(Y^g) - E(Y^{g'})$ is 0.

The data analysis might start by computing the cumulative treatment variable $A = A_0 + A_1$, which can take 3 values: 0 (if the subject remains untreated at both times), 1 (if the subject is treated at time 1 only or at time 2 only), and 2 (if the subject is treated at both times). The treatment strategies of interest can then be expressed as 'always treat' $A = 2$, and 'never treat' $A = 0$. Because the average causal effect of interest is null, any valid method should estimate that the value of the difference $E(Y^{a=2}) - E(Y^{a=0})$ is 0.

A commonly used method to adjust for confounding is estimating the association between treatment and outcome in each subset – stratum – of the study population defined by L_1. Say that the variable L_1 can take only two values: 1 if the CD4 cell count is low, and 0 otherwise. To estimate the effect in those with low CD4 cell count, we would have to select (i.e., condition on) the subset of the population with value $L_1 = 1$. That is, we would compute the difference $E(Y|A = 2, L_1 = 1) - E(Y|A = 0, L_1 = 1)$. And similarly for the stratum $L_1 = 0$. But stratification can have unintended effects when the association measure is computed within levels of a variable L_1 that is caused by or shares causes with a component A_0 of the time-varying treatment.

Figure 2 shows that conditioning on L_1 – a collider – opens the path $A_0 \rightarrow L_i \leftarrow U \rightarrow Y$. That is, stratification induces a noncausal association between the treatment A_0 at time 0 and the outcome Y within levels of L_1. Among those with low CD4 count ($L_1 = 1$), being on treatment ($A_0 = 1$) makes it more likely that the person is severely immunodepressed; among those with a high level of CD4 ($L_1 = 0$), being off treatment ($A_0 = 0$) makes it more likely that the person is not severely immunodepressed. Thus, the side effect of stratification is to induce an association between prior exposure A_0 and U, and therefore between A_0 and the outcome Y.

Stratification eliminates confounding for A_1 at the cost of introducing selection bias for A_0. The associational differences $E(Y|A = 2, L_1 = l) - E(Y|A = 0, L_1 = l)$ could be different from 0 even if treatment has no effect on the outcome of any individuals. This bias arises from choosing a subset of the study population by selecting on a variable L_1 affected by (a component A_0 of) the time-varying treatment. The net bias for any particular summary of the time-varying treatment that is used in the analysis (cumulative treatment in our example) depends on the relative magnitude of the confounding that is eliminated and the selection bias that is created.

The same selection bias arises when fitting an outcome regression model for the mean outcome within levels of treatment history $\overline{A}$, summarized by the cumulative treatment A, and covariate history $\overline{L}$, which in our oversimplified example includes only L_1. One possible outcome regression model is

$$E(Y|A, L_1) = \beta_0 + \beta_1 I(A = 2) + \beta_2 I(A = 1) + \beta_3 L_1$$

where $I(A = a)$ takes value 1 if $A = a$ and 0 otherwise. That is, $I(A = 2)$ and $I(A = 1)$ are indicators – dummy variables – for values 2 and 1, respectively, of the trichotomous variable $A = A_0 + A_1$. When both β_1 and β_2 equal 0, the model estimates the conditional mean in those with $A = 0$, the reference value. The parameter β_1 is equal to the difference of conditional means $E(Y|A = 2, L_1) - E(Y|A = 0, L_1)$. This model assumes that the value of the difference is the same in the strata $L = 1$

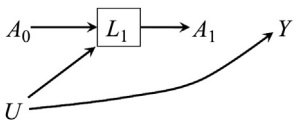

Figure 2 Causal diagram of a study with time-varying treatment A_k and confounders L_k, and with conditioning on confounders affected by treatment. For simplicity, only the first two time points are shown.

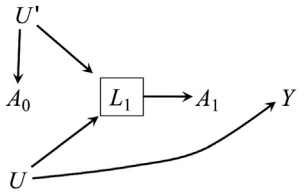

Figure 3 Causal diagram of a study with time-varying treatment A_k and confounders L_k, and with conditioning on confounders that share causes with treatment. For simplicity, only the first two time points are shown.

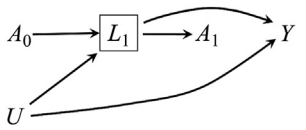

Figure 4 Causal diagram of a study with time-varying treatment A_k and confounders L_k, and with conditioning on confounders that are on a causal pathway from treatment to outcome. For simplicity, only the first two time points are shown.

and $L = 0$. Some might want to interpret the associational parameter β_1 as the effect of 'always treat' $A = 2$ versus 'never treat' $A = 0$ within levels of the covariate L_1. Unfortunately, such causal interpretation is unwarranted because, as Figure 2 shows, conditioning on L_1 induces an association between a component of treatment and the outcome.

We said before that, in order to estimate the effect $E(Y^{a=2}) - E(Y^{a=0})$ without bias, we need to be able to estimate the joint effect of the treatment components A_0 and A_1 simultaneously and without bias. As we have just seen, this is something impossible to achieve using traditional methods such as stratification and outcome regression (and matching), even when data on L_1 are available. In the presence of treatment–confounder feedback, if we do not adjust for L_1, there will be confounding for the effect of A_1 on Y, and if we do adjust for L_1 by traditional methods, there will be selection bias for the causal effect of A_0 on Y.

Technically speaking, the bias of traditional methods will occur not only when time-varying confounders are affected by prior treatment, but also when time-varying confounders share an unmeasured cause U' with prior treatment, as shown in Figure 3 (Hernán et al., 2004). In this setting, conditioning on the collider L_1 opens the path $A_0 \leftarrow U' \rightarrow L_1 \leftarrow U \rightarrow Y$. For brevity, we refer to both settings in Figures 2 and 3 as treatment–confounder feedback.

Conditioning on confounders L_1 which are affected by previous treatment can create selection bias even if the confounder is not on a causal pathway between treatment and outcome. In fact, no such causal pathway exists in Figures 2 and 3. On the other hand, in Figure 4 the confounder L_1 for subsequent treatment A_1 lies on a causal pathway from earlier treatment A_0 to outcome Y, i.e., the path $A_0 \rightarrow L_1 \rightarrow Y$. Were the potential for selection bias not present in Figure 4 (e.g., were U not a common cause of L_1 and Y), the association of cumulative treatment A with the outcome Y within strata of L_1 could be an unbiased estimate of the direct effect of A on Y not through L_1, but still would not be an unbiased estimate of the overall effect of A on Y, because the effect of A_0 mediated through L_1 is not included.

In summary, adjustment for time-varying confounders is generally necessary when comparing treatment strategies that involve time-varying treatments either in observational studies or in randomized trials. Because traditional adjustment methods cannot appropriately adjust for time-varying confounders that are affected by treatment, g-methods (Robins and Hernán, 2008) are required to handle treatment–confounder feedback.

See also: Causal Inference: Overview; Longitudinal Data; Observational Studies: Overview; Survival Analysis: Introduction.

Bibliography

Hernán, M.A., Hernandez-Diaz, S., Robins, J.M., 2004. A structural approach to selection bias. Epidemiology 15 (5), 615–625.

Pearl, J., 1995. Causal diagrams for empirical research. Biometrika 82, 669–710.

Pearl, J., Robins, J.M., 1995. Probabilistic evaluation of sequential plans from causal models with hidden variables. In: Uncertainty in Artificial Intelligence, Proceedings of the 11th Conference, pp. 444–453.

Richardson, T.S., Robins, J.M., 2013. Single World Intervention Graphs (SWIGs): A Unification of Counterfactual and Graphical Approaches to Causality. Working Paper Number 128. Available at: http://www.csss.washington.edu/Papers/.

Robins, J.M., 1986. A new approach to causal inference in mortality studies with a sustained exposure period – application to the healthy worker survivor effect. Mathematical Modelling 7, 1393–1512 (published errata appear in Mathematical Modelling 14, 917–921(1987)).

Robins, J.M., 1987. Addendum to "a new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect". Computers and Mathematics with Applications 14 (9–12), 923–945 (errata appeared in Computers and Mathematics with Applications 18, 477(1987)).

Robins, J.M., 1997. In: Berkane, M. (Ed.), Causal Inference from Complex Longitudinal Data. Latent Variable Modeling and Applications to Causality. Springer Verlag, New York, pp. 69–117.

Robins, J.M., Hernán, M.A., 2008. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice, G., Davidian, M., Verbeke, G., Molenberghs, G. (Eds.), Longitudinal Data Analysis. Chapman and Hall/CRC Press, New York, pp. 553–599.

Spirtes, P., Glymour, C., Scheines, R., 1993. Causation, Prediction, and Search. Lecture Notes in Statistics 81. Springer-Verlag, New York.

Young, J.G., Hernán, M.A., Robins, J.M., 2014. Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data. Epidemiologic Methods 52 (Suppl. 3), S132–S139.

Longitudinal Data

Burton H Singer, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 9058–9063, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

This article discusses cross-time, multiple life-domain representations of life histories and their relationship to a defined set of outcomes. A primary objective is to demonstrate how to integrate numerical and narrative information derived from large data sets, where the individual history -or major segments of a whole life – is the basic unit of analysis. We distinguish bottom-up, or sequential aggregation strategies, from top-down, or partitioning strategies. The ideas and methods are illustrated in the context of health outcomes.

Longitudinal data analysis is a subspecialty of statistics in which individual histories use spaced en rules throughout for parenthetical dashes interpreted as sample paths, or realizations, of a stochastic process – are the primary focus of interest. A wide variety of scientific questions can only be addressed by utilizing longitudinal data together with statistical methods that facilitate the detection and characterization of regularities across multiple individual histories. Despite the stated focus on individual histories, most of the analytical strategies either invoke strong assumptions about stochastic dependencies across time and/or use variable-centered methods to study the influence of multiple covariates on process dynamics and its relationship to diverse outcomes. An exception to this pattern is the literature on growth curve analysis, where the full individual histories are modeled and frequently aggregated into categories of similar histories. This set of techniques, as currently constituted, has severe limitations when a study objective is to understand the interrelationships between multiple dimensions of a multivariate time series. Problems arise due to the utilization of interaction terms, as in regression analyses, which represent the imposition of a variable-centered idea in an inherently person-centered question.

The purpose of this article is to describe some person-centered methods that respond to the challenge of characterizing life histories, ascertained with longitudinal data, placing emphasis on the effective utilization of both cross-time and multiple-life-domain information. We motivate the methods and illustrate their use in the context of explaining how a variety of health outcomes occur. However, the methodological issues are generic and transfer to a wide range of problems in the behavioral and social sciences.

Motivation

Extensive empirical evidence documents associations between early and mid-life risk factors and later-life chronic disease and disability (Barker and Osmond 1986, Wadsworth and Kuh 1997). These findings generate considerable interest in understanding the subtle nuances of life histories that might explain how such outcomes occur. Broad-gauged interest in successful aging (Rowe and Kahn 1998), with particular emphasis on how some individuals manage to adjudicate effectively multiple difficult challenges, also suggests in-depth comparative studies of life histories. Detailed accounts of whole lives have previously been studied primarily as individual cases, with evidence presented in forms ranging from rich ethnographies to focused interviews and, in a medical context, from open-ended responses to structured inquiries by clinicians.

Complementing such narrative accounts of life histories are structured longitudinal surveys that contain – for each individual – hundreds, or even thousands, of responses to questions about experiences over time and across multiple life domains (Hauser et al. 1993, Power et al. 1991, Wadsworth and Kuh 1997). In principle, those interested in the impact of multiple facets of life histories on a given outcome(s) should be able to construct representations of whole lives from the survey responses. Despite considerable interest in this task across multiple disciplines (Elder 1974, Vaillant 1983), there has been a dearth of analytic strategies for producing representations of whole lives and for aggregating them into meaningful taxonomies that facilitate understanding of how given outcomes occur.

With these observations at hand, our primary objective is to describe top-down and bottom-up methods for constructing taxonomies of life history representations and use them as the basis for explaining how given outcomes occur. We will also emphasize the construction of narratives from numerical survey responses and their utilization as part of an analytical strategy to understand 'whole lives.' We illustrate the ideas and methods in the context of health outcomes. However, the methodological issues are quite general and by no means restricted to this class of outcomes.

Classes of Methods

The techniques discussed in this article are called person-centered in the sense that the individual life history is the unit of analysis. Within this broad class of methods are two categories of techniques called 'bottom-up' and 'top-down.' The bottom-up strategies begin with highly idiosyncratic individual histories, and the analytical steps then identify important commonalities and differences across lives, leading to

aggregation of histories into relatively homogeneous groups. Optimal matching (Abbott and Hrycak 1990), p-technique integrated with dynamic factor analysis (Shifren et al. 1997), and the integrated qualitative–quantitative aggregation strategy of Singer et al. (1998) are examples of bottom-up strategies.

Top-down techniques sequentially partition a heterogeneous population into progressively more homogeneous subgroups. Recursive partitioning (Zhang and Singer 1999), Grade-of-Membership models (Berkman et al. 1989), and the clustering methods of Bergman (1998) are generic examples of these methods.

A Bottom-Up Example: Understanding Psychological Resilience

The Problem

Among those who experience major episodes of depression in adult life, or earlier, some manage to subsequently regain high levels of well-being – that is, they demonstrate resilience. We ask whether those who show recovery have particular features that distinguish them from those who continue to suffer depression. An affirmative answer could suggest prevention strategies which, if applied early in life, might serve to bypass episodes of depression, or minimize their severity and facilitate recovery from them.

Longitudinal Data

The Wisconsin Longitudinal Study (WLS) began with a random sample of 10,317 men and women who graduated from Wisconsin high schools in 1957. Survey data were collected from the original respondents in 1957, 1975, and 1992/93. Telephone interviews were conducted in 1992/93 with 8 020 of the 1975 respondents, 6 535 of whom also responded to a mail survey. Data have been collected on respondents' family background, starting resources, academic abilities, youthful aspirations, social support, social comparisons, and the timing and sequencing of adult educational and occupational achievements, work events and conditions, family events, and physical and mental health. We focus on a subset of the 5 009 primary respondents who participated in all three waves of the WLS and who responded to questions assessing affective disorder and well-being in the most recent survey. The affective questions were administered to a randomly selected 80 percent of respondents. For each individual, there were responses on over one thousand variables covering all three waves of the survey and information across the multiple life domains indicated above.

The resilience criteria (see Singer et al. 1998) were satisfied by 168 women. There were 59 women who satisfied the depressed/unwell criteria. The analytical task is the development of a taxonomy of life histories for each group such that the characterization of any history is sufficiently nuanced to capture the richness of individual experience, but not so detailed as to represent only very few women. For a more broad-based discussion of resilience, using both narrative and numerical data on individual histories to the study of racial and ethnic inequalities in health, see Singer and Ryff (1997).

A Person-Centered Strategy

Construction of an empirical taxonomy of life histories requires a set of organizing principles – based on prior literature – that constrain the distillation of over one thousand responses for each individual into composite variables. Such composites are the basis for the ultimate representations of categories of histories. The following five principles were used to specify combinations of individual characteristics essential to explain how the outcomes come about.

a. Adversity and its cumulation over time has negative mental health consequences.
b. Advantage and its cumulation over time has positive mental health consequences.
c. Reactions to adversity or advantage influence the impact of life experiences.
d. Position in social hierarchies has consequences for mental health.
e. Social relationships influence the impact of life experiences.

These propositions are not new to the social sciences, mental health, or aging fields. What is new is the integration of them into an analytical strategy to guide life-history analysis. The methodology is usefully decomposed into five principle steps:

1. guided by the organizing principles, life history narratives are constructed for a sub-sample of cases from each outcome group (i.e., resilient and depressed/unwell);
2. abstracted chronological charts, each one representing the synthesis of several narratives, were constructed;
3. reduced response vectors, **b**, consisting of composite variables suggested by the chronological charts and individual variables whose rationale derives from the organizing principles are specified;
4. aggregates (categories) of histories, in terms of response vectors judged to be close to one another, are specified; and
5. tests of distinguishability are conducted.

Step (5) reduces to ascertaining whether or not the categories of histories characterizing the resilient women occur with low frequency in the depressed/unwell group and, conversely, whether histories characterizing the depressed/unwell women occur with low frequency among the resilient women. If the taxonomies of histories are to be a basis for explanation of the mental health outcomes, there should be distinguishability of categories of histories between the two groups.

Narratives

A small random sample (~10 cases) is selected from each outcome group, and a narrative biographical story is written about each individual. We adopt Stone's (1979) conception of narrative, defined as 'the organization of material in a chronological sequential order and the focusing of the content into a single coherent story, albeit with subplots.' Our rationale for writing narratives based on survey data (with over one thousand responses per individual) is that the human mind has difficulty processing enormous amounts of information in parallel (i.e., long lists of responses representing a single life), although it is well suited to process a coherent story (i.e., a

written narrative about a life) (Turner 1996). The use of narratives here is qualitatively different from their role when they are part of the raw data. Specifically, the construction of narratives – by the analyst and starting with numerical response data – is part of an analytical strategy. The narratives do not represent empirical data derived from focused interviews, recorded conversations, or text written by the respondents, but stories written by investigators to describe 'whole lives through time.' In generating the narratives, responses selected for inclusion were those expected to impact mental health, based on prior theoretical and empirical literature. The goal is to tell a life story in such a way that omitted information would make little or no difference for understanding the relevant life experiences preceding and predicting midlife mental health outcomes.

Abstract Chronological Charts

Two life histories can be aggregated and called similar, or even approximately indistinguishable, when the response vectors representing them are similar. At the level of individual cases, represented by raw response data on more than one thousand variables or by narratives, there is so much idiosyncratic detail that meaningful aggregation seems almost hopeless. However, via the organizing principles, it is possible to start with three or four narratives and begin specifying composite variables that ultimately lead to the specification of response vectors, some of which are sufficiently similar that they can be aggregated into a common category.

The composite variables are the initial basis for revised response vectors for all individuals that need to satisfy two criteria. First, the revised response vector for a given individual should provide for a nuanced summary of the person's life history as it relates to mental health outcomes. Second, the response vector should not be too idiosyncratic. The latter condition is essential if we are to aggregate life histories – represented as reduced response vectors – so that individuals with 'similar' response vectors can be grouped into common categories of histories.

Reduced Response Vectors

Guided by the abstract chronological charts, we specify a provisional reduced vector of variables, whose values for a single individual define the person's reduced response vector. The provisional reduced vector of variables will, clearly, represent all persons for whom we have constructed narratives. However, this is only a small sub-sample of a much larger population. We augment this group with 10–15 new cases for whom we do not have narratives. Their provisional reduced response vectors augmented by responses on additional variables that are clearly linked to the organizing principles are displayed. Visual examination of these augmented response vectors for the new cases AND those for which there are narratives serves to suggest new composite variables and revisions of those specified for the provisional reduced vector of variables. The result of this visual inspection and judgment is a new second-stage reduced vector of variables. We continue bringing in blocks of 10–15 new cases at a time, repeating the above step, and thereby iteratively refining the reduced vector of response variables until all cases for a given mental health outcome group have been examined. The final reduced vector of response variables, **b**, is the basis for initiating the formal specification of a taxonomy of histories.

Aggregation of Reduced Response Vectors: Pathway Specification

All variables are identified by binary codes in **b**. Next, we identify the longest AND statement common to at least 20 individuals. This identifies a set of response vectors – equivalently, a set of life histories – that exactly match one another on the components of the AND statement. If the components of the AND statement are distributed across the whole lives and they are interpretable as meaningful pathways to a given mental health outcome, then we have succeeded in aggregating a set of life histories into a well-defined group. For the resilient women in WLS, the binary coded version of **b** had 17 components, and the longest AND statement common to at least 20 individuals had seven components.

With an initial aggregate of histories at hand, we eliminate this group from further consideration and identify the longest possible AND statement characterizing at least 20 – and hopefully more – of the remaining men. In the WLS resilient women, this did not lead to a long enough AND statement that was sufficiently nuanced to characterize possible routes into episodes of depression, as well as the cumulation of advantage that would promote recovery. This led to specification of more complex Boolean statements that were both consistent with our organizing principles and represented the life histories of a substantial number of women. At least two strategies can be used for this specification. First, visual inspection of the 17-component response vectors revealed a group of 48 women, all of whom grew up in a household with at least one alcoholic parent AND who had a diversity of different forms of compensating advantage in adulthood (represented by logical OR statements). For a narrative description of the resulting Boolean statements, see Singer et al. (1998).

A more formally mechanized approach to this step – using, for example, Ragin's QCA software (1987) – would be to minimize the Boolean expression representing all women who are not in the first group identified via the 'longest possible AND statement criterion.' The resulting minimal Boolean expression would, hopefully, have clearly interpretable components that would represent distinct pathways – hence, additional members of the desired taxonomy of life histories – to given mental health outcomes. Although the analysis in Singer et al. (1998) proceeded beyond the first specification of an initial group by inspection and judgment, linked to organizing principals, we recommend application of both strategies en route to a final taxonomy. Indeed, as illustrated in Ragin (1987), minimal Boolean expressions that are not readily interpretable on a first pass through reduced response vectors often suggest recoding and construction of alternative composite variables that ultimately lead to a useful taxonomy.

Tests of Distinguishability

Having constructed a taxonomy of pathways for each outcome category – resilient and depressed/unwell in the present

example – it is necessary to demonstrate that the routes to any one outcome differ in important ways from the set of routes to a different outcome. To this end, we first ask whether the aggregate pathways for the resilient women differ from the pathways for the depressed/unwell population. In particular, we tabulate the frequency with which the resilient pathways occur in the depressed/unwell population. Conversely, we take the aggregate pathways representing the depressed/unwell women and tabulate the frequency with which those pathways actually occur among the resilient women. We expect to find that the resilient pathways are rare in the depressed/unwell population and also that the depressed/unwell pathways are rare in the resilient population. For example, 16 percent of the resilient women are on the pathway indicated above by the seven-component logical AND statement. Only 8 percent of the depressed/unwell have the same pathway ($p < .10$). Twenty percent of the resilient women are on the pathway where at least one parent is a chronic alcoholic AND there is a diversity of compensating positive experiences in adulthood. This pathway only occurs for 5 percent of the depressed/unwell population ($p < .05$). There are persistently negative-experience pathways in the depressed/unwell population that have no counterpart in the resilient group. Sharper discrimination occurs depending upon the particular application (see e.g., Zhao et al. 2000). The central point, however, is that assessment of distinguishability is important for making the case that there are, as hypothesized, qualitatively different routes to distinct mental health outcomes.

An Alternative Bottom-up Strategy

An analytical strategy that might serve as an alternative to what is put forth in the above example is optimal matching (Abbott and Hrycak 1990, Abbott and Barman 1997). Optimal matching derives from the literature on comparison of DNA sequences. It takes univariate time series for individuals and groups them together into sets of what are regarded as equivalent histories. The grouping requires the availability of a metric on the space of possible sequences (histories); and two sequences are defined to be members of the same group if their distance apart – in terms of the metric – is sufficiently small. To apply the automated optimal matching technique requires that input data be in the form of what we have called 'reduced response vectors.' The analogue of steps leading from more complex longitudinal data to this stage, as we have described above, must still be carried out for univariate time series prior to carrying out the formal aggregation step (Abbott and Hrycak 1990). Although optimal matching is a natural way to proceed in the construction of a taxonomy of histories, the problem with its implementation on cross-time, multi-life-domain data is that there is no obvious metric yet developed that is suited to the complex multidimensional time series corresponding to individual histories, as discussed in the above example. Furthermore, a single measure of distance between two multidimensional histories seems too coarse for decisions about what constitute equivalent histories. The strategy for aggregation of reduced response vectors – in the above example – uses information about multiple features of life histories. This suggests that any automated aggregation technique, analogous to optimal matching, should be based on a multidimensional measure on the space of histories.

Top-Down Methods

There is a considerable diversity of top-down, person-centered analytical strategies. A valuable review of both concepts and methods is contained in Cairns et al. (1998). One common strategy focuses on clustering techniques, where a critical input to the algorithms is a metric on the space of histories. Not surprisingly, applications of these techniques, which partition a heterogeneous population into relatively homogeneous clusters, have been restricted to low-dimensional data and rather simple summaries of histories. To reach the level of nuance and multi-domain representation of the kind we have discussed above requires a quite different set of measures of similarity of histories to be constructed than heretofore. This is an important research topic for the future.

Recursive partitioning (RP) classification (Breiman et al. 1984, Zhang and Singer 1999) is an automated procedure that fits trees to multi-dimensional longitudinal data such that pathways down the tree to a given outcome (terminal node) identify homogeneous sub-populations with common histories. Trees are constructed with the objective of maximizing ability to predict membership in the outcome categories. Each pathway down the tree ending in a given terminal node (identified, for example, with a health outcome) corresponds to a logical AND statement incorporating information across multiple life domains. Two or more pathways into the same terminal node are identified with a logical OR statement describing how that particular outcome occurred. The judgment of the investigator is required to determine whether all pathways to the nodes identified with a given outcome should be summarized by a single complex Boolean expression, or partitioned into two or more sets of pathways defining qualitatively different kinds of histories. Recursive partitioning (RP) algorithms could accept either the original raw response vectors or reduced response vectors, as described in the above example.

From our perspective, a limitation of RP for generating a taxonomy of life histories is the fact that the performance criteria and the splitting rules for tree construction are not constrained by substantive knowledge from any field of scientific inquiry (e.g., a set of organizing principles). Thus, pathways down the best predictive tree are not always interpretable within the knowledge of a given field. This disjunction between statistical performance criteria, based solely on numerical goodness-of-fit, and constraints on tree structure and splitting rules derived from extant scientific knowledge have proven problematic in past applications of RP (Levy et al. 1981, 1985), for example, applied RP to neurological assessments of comatose patients to predict which patients were likely to recover with moderate disabilities as opposed to remaining vegetative or dying within one year of hospital admission. The numerically best predictive trees were virtually never neurologically interpretable. Rather, inspection of 20+ nearly optimal trees led to selection, based on scientific judgment, of a prognostic tree that had interpretable pathways. For a start at more interactive – i.e., allowing the investigator to modify trees during the construction process on the basis of organizing

principles – and flexible strategies based on RP technology, the reader should consult Zhang and Singer (1999).

Conclusions

We have described a variety of person-centered methods for the analysis of longitudinal data. Particular emphasis has been given to a relatively new strategy – the bottom-up example – that takes a full complex individual history as the unit of analysis. The principal and novel features of this strategy are:

a. It includes the writing of narratives at the beginning and terminal stages of the analysis of survey data.
b. It is person- not variable-centered.
c. It produces pathways that represent aggregates of individual life histories.
d. It works simultaneously across multiple life domains.
e. It is capable of identifying multiple pathways to a given outcome.
f. It systematizes the decision making en route to delineation of the pathways.
g. It tests hypotheses about whole lives that are the integration of what is typically studied as separate pieces.

Point (a) underscores the importance of examining whole lives in story form as a means of comprehending the volume of detail that necessarily comprises human lives. Although point (b) is a clear objective in much of the life-history literature (e.g., Elder 1974, Magnusson and Bergman 1990), the lack of person-centered methodological tools has prevented investigators from proceeding to point (c). Attempts to get at aggregates have typically involved the use of variable-centered techniques, since these were available. However, the investigators mentioned above and many others were well aware that the lack of preservation of individuals as units of analysis severely limited the inferences they could draw about the dynamics of individual (or aggregates of individuals) change. Point (d) is conventionally dealt with, if at all, by variable-centered techniques. The bottom-up example strategy is, to the best of our knowledge, the first incorporation of multiple life domain information in a process of aggregation of whole life representations. Many investigators (e.g., Harris et al. 1990, Magnusson and Bergman 1990) have struggled with point (e); however, the absence of tools for aggregating whole life representations has prevented their delineating full pathways. Harris et al. (1990) set forth a template of several pathways to depression, and Magnusson and Bergman (1990) looked for 'patterns' connecting childhood problems with adult maladjustment. Nevertheless, their variable-centered empirical analyses (using logistic regression with interaction terms in the former, and cluster analysis in the latter) did not achieve the desired objective. What was needed was a person-centered methodology. Point (f) emphasizes the critical role of scientific judgment embedded in the progression of methodological steps, something that is an essential part of any data analysis, but is infrequently reported. Finally, point (g) carries hypothesis testing into the uncharted territory of whole lives.

Recommended reading for more detail on the person-centered methods emphasized in this article and on the integration of numbers and narratives to study whole lives, including narratives as data, see Singer et al. (1998), Singer and Ryff, (2001). An in-depth exposition of recursive partitioning, including online software, is contained in Zhang and Singer (1999). An instructive example of formal modeling of multiple-episode event history data is given in Heckman and Walker (1998) and a useful illustration of modeling of multivariate event history data, with both observed and unobserved covariates, is contained in Manton et al. (1987).

See also: Factor Analysis and Latent Variable Models in Personality Psychology; Growth Curve Analysis; Life Management, Developmental Psychology of; Lifespan Development, Theory of; Linear Hypothesis: Regression (Basics); Longitudinal Data: Event–History Analysis in Discrete Time; Markov Models and Social Analysis.

Bibliography

Abbott, A., Barman, E., 1997. Sequence comparison via alignment and Gibbs sampling: A formal analysis of the emergence of the modern sociological article. In: Raftery, A. (Ed.), Sociological Methodology. American Sociological Association, Washington, DC, pp. 47–87.

Abbott, A., Hrycak, A., 1990. Measuring resemblance in sequence data: An optimal matching analysis of musician's careers. American Journal of Sociology 96, 144–185.

Barker, D., Osmond, C., 1986. Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales. Lancet I, 1077–1081.

Bergman, L., 1998. A pattern-oriented approach to studying individual development: Snapshots and processes. In: Cairns, R., Bergman, L., Kagan, J. (Eds.), Methods and Models for Studying the Individual. Sage Publications, Thousand Oaks, CA, pp. 83–122.

Berkman, L., Singer, B., Manton, K., 1989. Black/white differences in health status and mortality among the elderly. Demography 26 (4), 661–678.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

Cairns, R., Bergman, L., Kagan, J. (Eds.), 1998. Methods and Models for Studying the Individual. Sage Publications, Thousand Oaks, CA.

Diggle, Peter, Heagerty, Patrick, Liang, Kung-Yee, Zeger, Scott, 2002. Analysis of longitudinal data. Oxford University Press, Oxford, UK.

Elder, G., 1974. Children of the Great Depression: Social Change in Life Experience. University of Chicago Press, Chicago.

Hancox, Robert J., Milne, Barry J., Poulton, Richie, 2004. Association between Child and Adolescent Television Viewing and Adult Health: A Longitudinal Birth Cohort Study. The Lancet 364 (9430), 257–262.

Harris, T., Brown, G., Bifulco, A., 1990. Loss of parent in childhood and adult psychiatric disorder: A tentative overall model. Development and Psychopathology 2, 311–328.

Hauser, R., Carr, D., Hauser, T., Hayes, J., Krecker, M., Kuo, H., Magee, W., Presti, J., Sweeney, M., Thompson-Colon, T., & Uhrig, S., 1993. The Class of 1957 After 35 years: Overview and Preliminary Findings. Madison, WI: University of Wisconsin-Madison Working paper 93–17, Center for Demography and Ecology.

Heckman, J., Walker, J., 1998. Using goodness of fit and other criteria to choose among competing duration models: A case study of Hutterite data, 1987. In: Cloog, C. (Ed.), Sociological Methodology. American Sociological Association, Washington, DC, pp. 247–307.

Hedeker, Donald, Gibbons, Robert D., 2006. Longitudinal data analysis. John Wiley & Sons, Hoboken, NJ.

Levy, D., Bates, D., Caronna, J., Cartlidge, N., Knill-Jones, R., Lapinski, R., Singer, B., Shaw, D., Plum, F., 1981. Prognosis in non-traumatic coma. Annals of Internal Medicine 94, 293–301.

Levy, D., Caronna, J., Singer, B., Lapinski, R., Frydman, H., Plum, F., 1985. Predicting outcome from hypoxic-ischemic coma. Journal of American Medical Association 253, 1420–1426.

Magnusson, D., Bergman, L., 1990. A pattern approach to the study of pathways from childhood to adulthood. In: Robins, L., Rutter, M. (Eds.), Straight and Devious Pathways from Childhood to Adulthood. Cambridge University Press, Cambridge, UK, pp. 101–115.

Manton, K., Stallard, E., Woodbury, M., Tolley, D., Yashin, A., 1987. Grade-of-membership techniques for studying complex event history processes with unobserved covariates. In: Clogg, C. (Ed.), Sociological Methodology. American Sociological Association, Washington, DC, pp. 309–346.

Power, C., Manor, O., Fox, A., 1991. Health and Class: The Early Years. Chapman and Hall, London.

Ragin, C., 1987. The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies. University of California Press, Berkeley, CA.

Rowe, J., Kahn, R., 1998. Successful Aging. Pantheon Books, New York.

Shifren, K., Hooker, K., Wood, P., Nesselroade, J., 1997. Structure and variation of mood in individuals with Parkinson's disease: A dynamic factor analysis. Psychology and Aging 12 (2), 328–339.

Singer, B., Ryff, C.D., 1997. Racial and ethnic inequalities in health: Environmental, psychosocial, and physiological pathways. In: Devlin, B., Fienberg, S., Resnick, D., Roeder, K. (Eds.), Intelligence, Genes, and Success: Scientists Respond to the Bell Curve. Springer-Verlag (Copernicus), New York, pp. 89–122.

Singer, B., Ryff, C.D., 2001. Person-centered methods for understanding aging: The integration of numbers and narratives. In: Binstock, R., George, L.K. (Eds.), Handbook of Aging and the Social Sciences. Academic Press, San Diego, CA, pp. 44–65.

Singer, B., Ryff, C., Carr, D., Magee, W., 1998. Linking life histories and mental health: A person-centered strategy. In: Raftery, A. (Ed.), Sociological Methodology. American Sociological Association, Washington, DC, pp. 1–51.

Stone, L., 1979. The revival of narrative: Reflections on a new old story. Past and Present 85, 3–24.

Strawbridge, William J., Deleger, Stéphane, Roberts, Robert E., Kaplan, George A., 2002. Physical Activity Reduces the Risk of Subsequent Depression for Older Adults. American Journal of Epidemiology 156 (4), 328–334.

Turner, M., 1996. The Literary Mind. Oxford University Press, New York.

Vaillant, G., 1983. The Natural History of Alcoholism. Harvard University Press, Cambridge, MA.

Wadsworth, M., Kuh, D., 1997. Childhood influences on adult health: A review of recent work from the British 1946 national birth cohort study, the MRC national survey of health and development. Pediatric Perinatal Epidemiology 11, 2–20.

Warren, John Robert, Sheridan, Jennifer T., Hauser, Robert M., 2002. Occupational Stratification across the Life Course: Evidence from the Wisconsin Longitudinal Study. American Sociological Review 67 (3), 432–455.

Zhang, H., Singer, B., 1999. Recursive Partitioning in the Health Sciences. Springer-Verlag, New York.

Zhao, H., McLanahan, S., Brooks-Gunn, J., Singer, B., 2000. Studying the real child rather than the ideal child: Bringing the person into developmental studies. In: Bergman, L., Cairns, R., Nilsson, L.-G., Nystedt, L. (Eds.), Developmental Science and the Holistic Approach. Lawrence Erlbaum, Mahwah, NJ, pp. 393–419.

Longitudinal Data: Event–History Analysis in Discrete Time

Elja Arjas, University of Helsinki, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article, volume 13, pp. 9063–9067, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article first gives a practical motivation for considering event-history models in discrete-time. The main point in the subsequent development is in explaining the structure of the corresponding likelihood function, including situations in which the considered individual status variables are related to covariates via a regression model. Finally, a brief description is given about how discrete-event history data can give rise to truly continuous time models, by applying Bayesian modeling for data augmentation.

The practical motivation for considering discrete-time models of event histories (Allison, 1982; Thompson, 1977) comes from the fact that such longitudinal data are often recorded and presented in a discrete-time form (*see* Longitudinal Data). Thus, for example, it is common to collect information on the values of considered individual status variables at regularly spaced time points and to report such measurements directly as the data. A time unit is then fixed, typically to be a day, a week, a month, or a year, thus creating a regular time lattice. A slightly different but related reason why event-history data often appear in a discrete-time form is that the event times themselves are reported in a rounded form. Then the events which happened inside a time interval of unit length are somewhat arbitrarily assigned to either of its end points or, in so-called *actuarial methods*, to the middle. As a consequence, the reported event times are often *tied*, with several individuals sharing the same value. As an example, one can consider the changes of a person's employment status during the work history, typically consisting of spells of employment and unemployment, and finally ending with a permanent transition to outside of the work force. In order to be completely accurate, the reporting might have to done day-by-day. More realistically in practice, it can only be done monthly, and then a number of different alternatives exist.

For example, one could directly report the employment status on the last day of the month. A drawback of such reporting would be that, if the status can undergo frequent changes, the status at the end of the month might be purely accidental and not represent the entire month in an adequate way. An alternative would be to set a threshold x and ask whether the person has been employed for at least x days during the month. Such a solution may not be completely satisfactory either, as a low value of x does not distinguish between people who were employed only briefly from those who were employed during the entire month, with a similar but opposite problem arising for a high value of x. Such problems are of less concern if changes in the status variable are less frequent (such as marital status), or even irreversible.

Discrete-Time Hazards

The main advantages from linking a statistical model directly with such data are that model construction becomes relatively straightforward, and that no special methods are required for dealing with the ties. For an illustration, suppose for simplicity that there are only two states for each individual, say *alive* and *dead*, and that the chosen time unit is one month. Then the lifetime T of an individual, in a discretized time scale, can be thought of as being determined by a sequence of Bernoulli trials, resulting in a sequence of 'zeros' as long as the individual is alive and ending with a 'one' for the month during which the individual dies. The basic element for setting up a corresponding statistical model is the conditional probability of recording 'one' in a unit interval given that all preceding intervals resulted in 'zeros', that is, the probability of death during the tth month for a person still alive after t-1 months (*see* Survival Analysis: Overview). Because of the discrete-time lattice used, such *discrete-time hazards* can be simply expressed as conditional probabilities:

$$h(t) = Pr(T = t | T \geq t) \qquad [1]$$

By applying the chain multiplication rule of conditional probabilities progressively for increasing values of t, one finds readily that the unconditional probability for the survival is given by the product

$$Pr(T > t) = \prod_{s=0}^{t} [1 - h(s)] \qquad [2]$$

Note that moreover

$$Pr(T = t) = Pr(T > t - 1)h(t) \qquad [3]$$

These are discrete-time analogues of the well known *exponential formulas* for the survival probability and density, respectively, which are valid in continuous time and express a lifetime distribution in terms of the corresponding *hazard rate* (see, e.g., Andersen et al., 1993; Kalbfleisch and Prentice, 1980). In fact, the exponential formulas can be obtained as limits when the time lattice used in the discretization becomes more and more dense, thereby making the discrete-time hazards $h(t)$ numerically small.

Regression Models and Likelihood Inference

When considering data consisting of lifetimes of several individuals, the number of such Bernoulli trials equals the total number of person-months at risk, and under natural independence conditions across individuals, a statistical model for the entire data can be formulated in terms of equally many

Bernoulli trials, with the corresponding discrete-time hazards depending on age. This gives rise to a simple likelihood expression of a product form, without that tied lifetimes in the data would be a problem. The structure of the likelihood remains essentially unchanged if the discrete-time hazards of the considered individuals depend on covariates, as long as the covariates are fixed in time. One is then dealing with a regression problem, in which the observed development of the response variables (lifetimes) is explained in terms of the corresponding realized covariate values in the data (*see* Survival Analysis: Overview). A convenient framework for considering such regression models, including the possibility of using standard software, is offered by *generalized linear models* (for a general account and references, see Fahrmeir and Tutz (1994), and for asymptotic properties of the estimators, see, e.g., Arjas and Haara (1987)). In the important special case of a *logistic link* function, if the considered time unit is short enough to make the discrete-time hazards $h(t)$ small and the intercept of the linear expression is allowed to depend on the interval, the model corresponds closely to the proportional hazards model of Cox (1972). If the covariates can themselves change in time, the same logic and structure of the likelihood can be used as long as the covariates are *external* (Kalbfleisch and Prentice, 1980) in the sense that their development can be thought of as being exogenous from the point of view of the studied event-history data of individuals.

This is because there is then no need for an explicit modeling of the process governing the covariate process, as the resulting contributions to the likelihood expression do not depend on the actual parameters of interest and can therefore be coalesced into a proportionality constant. The same logic applies if the individual event histories are right censored and the censoring mechanism is assumed to be noninformative. If the model also involves time-dependent *internal* covariates, special care is needed in modeling and inference to make sure that their development in time does not *confound* the regression problem and the corresponding conclusions, which are often given a causal form. The issue of confounding, in the case of sequential designed experiments, has been considered in a series of papers by Robins (e.g., Robins, 1997; Robins and Hernán, 2009). Sometimes, in order to give useful causal interpretations, it becomes necessary to build the regression model in two or more levels representing, for example, group level and individual level effects. For a recent contribution in this direction, see Barber et al. (2000). The same general ideas apply also in situations in which the considered status or response variables are no longer just binary. For example, the status of an individual could assume more than two possible values, and the same status could repeat itself many times during the follow-up. For example, as noted above, a person's employment status could first alternate several times between the states *employed* and *unemployed*, until it finally reaches the absorbing state *outside of work force* because of retirement. What is needed, however, is a convenient form of conditioning, which carries information about an individual's earlier history and present status and thereby influences the assessment of the discrete-time hazards associated with the future development, always by one time unit ahead. For an interesting case study, applying a variety of inferential techniques, see Fahrmeir and Wagenpfeil (1996).

Structure of the Likelihood

Let us now make this idea more explicit. Indexing the individuals in the data by $i \in I$ and denoting the value of the corresponding status or response variable at time t by $X_t(i)$, we are led to consider conditional probabilities for a transition from a state $X_t(i)$ at time t to a state $X_{t+1}(i)$ at a later time $t+1$. Denoting by $H_t(i)$ the pre-t-history of individual i, which contains all observations made on i up to and including time t, including the past values of $X_s(i)$ $(s \leq t)$ and of possible covariates considered in the analysis, the first task is to specify the conditional probabilities $\Pr[X_{t+1}(i)|H_t(i)]$.

The following natural conditional independence assumptions will then often be valid:

1. The variables $X_{t+1}(i)$, $i \in I$, are conditionally mutually independent given the collective pre-t-past $\{H_t(i),\ i \in I\}$ in the data.
2. Knowledge of the individual's own pre-t-history $H_t(i)$ is sufficient for predicting $X_{t+1}(i)$ in the sense that

$$Pr(X_{t+1}(i)|H_t(j), j \in I) = Pr[X_{t+1}(i)|H_t(i)] \qquad [4]$$

Under conditions (a) and (b), the likelihood expression arising from discrete-time event history data becomes a product of terms of the form $\Pr[X_{t+1}(i)|H_t(i)]$, where the product is taken over all individuals i and all times t. (A tacit assumption here has been that if the covariates registered in $H_t(i)$ are themselves time dependent, they are either exogenous or their evolution is modeled in a way which does not involve the actual parameter of interest.) This makes the statistical inference of the model parameters in principle a straightforward problem, for example, by applying maximum likelihood. If, moreover, the condition

3. Knowledge of the current state $X_t(i)$ is sufficient in the sense that

$$Pr[X_{t+1}(i)|H_t(i)] = Pr[X_{t+1}(i)|X_t(i)] \qquad [5]$$

holds, the model can be formulated in terms of a discrete-time *Markov chain* (see, e.g., Cinlar, 2013 and Markov Models and Social Analysis). Information contained in the observed past history $H_t(i)$ can then be stored into a covariate vector of a fixed length, and the likelihood becomes a product of transition probabilities $\Pr[X_{t+1}(i)|X_t(i)]$, again taken over both i and t.

In spite of the seemingly general applicability of such a discrete-time approach to likelihood inference, the following two conventions, both usually involving an approximation, seem necessary in interpreting such models. First, censoring and possible other forms of controlling the number of individuals at risk are always thought to happen at times which are integer multiples of the chosen time unit. Second, if the model involves time-dependent covariates which are used as regressors of the observed status or response variable, their measured values need to be interpreted as 'prevailing (constant) conditions for exposure' during each considered time interval of unit length (see Sandefur and Tuma, 1987). In particular, if the response variable can feed back and influence the values of the covariates, in order to save a causal interpretation of the regression model, the covariate values must be determined already at the beginning of each time interval, thus avoiding

the potential problem of conditioning on variables which are intermediate between a contemplated cause and the effect.

Embedding into Continuous-Time Models

Another, and perhaps a more serious, drawback from a modeling perspective concerns the interpretation of the discrete-time parameter. In almost any conceivable applied context, and corresponding the common human perception, time is viewed as being continuous. Based on this perception, and even in situations in which the data are in a discrete-time form, a more natural alternative for statistical modeling of event histories starts from the idea that there is a true underlying status process in continuous time, and that the recorded event-history (panel) data consist then of a series of measurements from this underlying process taken at a number of typically prescheduled time points. The term *current status data* is sometimes used to emphasize this aspect. The development of a person's true employment status, marital status, support of political parties, or presence or absence of a chronic disease, as functions of time, would be natural examples. Indeed, such individual status histories would be the focus of a statistical analysis if they only could be observed, while discrete-time event history data consist merely of incomplete information, typically drawn from a sequence of cross-sections of such histories. As a consequence of such imprecision, individual event histories in continuous time are generally unidentifiable from discrete-time data. In particular, the exact times of occurrence are then not recorded, and if the status of a person has changed back and forth within a unit interval, this possibility may not even become explicit in the data (see, e.g., Stasny, 1986). Nevertheless, it may be possible to embed the data successfully into a larger statistical model in continuous time. The fact that the individual underlying continuous-time developments cannot generally be determined uniquely from the data can often be compensated with some form of *data augmentation*. In other words, we can then consider an entire collection, or distribution, of sample paths which are consistent with the data. The practical solution may involve some form of simulation (multiple imputations), while a theoretical motivation to such techniques comes from Bayesian statistical inference. Exact numerical computations become quickly infeasible in practice, however. This leads to an application of sampling-based *Markov chain Monte Carlo* methods, in which the possible realizations of the individual processes are drawn by simulation (*see* Monte Carlo Methods and Bayesian Computation: Overview). Such data augmentation methods have been applied to data describing phenomena such as recurrent bacterial infections (Auranen et al., 2000) and influenza transmission (Cauchemez et al., 2004), but the same approach applies more generally to problems in which it is natural to relate discrete-event history data to underlying individual developments in continuous time.

Conclusion

Depending on the context, modeling of event-history data may emphasize more the *event times* at which some particular events of interest occurred, or the *states* in which some studied process happened to be at the times at which it was measured. To stress the former aspect, one often speaks of survival or duration data. The second aspect is similarly stressed by talking about repeated measurements or longitudinal data. When the time variable in the data and in the model is discrete, it may sometimes be difficult to make a clear distinction between these two basic forms. On the other hand, the likelihood expressions will be of the same canonical product form in both cases, so that likelihood inference will remain the same. Recommended reading on these and closely related issues include Allison (1982), Barber et al. (2000), Diggle et al. (1994), Kalbfleisch and Prentice (1980), and Sandefur and Tuma (1987).

See also: Causal Inference: Overview; Longitudinal Causal Inference; Longitudinal Data: Event–History Analysis in Discrete Time; Longitudinal Data; Longitudinal Research: Panel Retention; Markov Chain Monte Carlo Methods; Markov Models and Social Analysis; Monte Carlo Methods and Bayesian Computation: Overview; Panel Surveys: Uses and Applications; Survival Analysis: Introduction; Survival Analysis: Overview.

Bibliography

Allison, P.D., 1982. Discrete-time methods for the analysis of event histories. In: Leinhardt, S. (Ed.), Sociological Methodology. Jossey-Bass, San Francisco, CA, pp. 61–98.

Andersen, P.K., Borgan, O., Gill, R.D., Keiding, N., 1993. Statistical Models Based on Counting Processes. Springer, New York.

Arjas, E., Haara, P., 1987. A logistic regression model for hazard: Asymptotic results. Scandinavian Journal of Statistics 14, 1–18.

Auranen, K., Arjas, E., Leino, T., Takala, A.K., 2000. Transmission of pneumococcal carriage in families: A latent Markov process model for binary longitudinal data. Journal of the American Statistical Association 95, 1044–1053.

Barber, J.S., Murphy, S., Axinn, W.G., Maples, J., 2000. Discrete-time multilevel hazard analysis. Sociological Methodology 30, 201–235.

Cauchemez, S., Carrat, F., Viboud, C., Valleron, A.J., Boelle, P.Y., 2004. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. Statistics in medicine 23, 3469–3487.

Cinlar, E., 2013. Introduction to stochastic processes. Courier Dover Publications.

Cox, D.R., 1972. Regression models and life tables (with discussion). Journal of the Royal Statistical Society, Series B 34, 187–220.

Diggle, P.J., Liang, K.-L., Zeger, S.L., 1994. Analysis of Longitudinal Data. Oxford University Press, Oxford, UK.

Fahrmeir, L., Tutz, G., 1994. Multivariate Statistical Modeling Based on Generalized Linear Models. Springer, New York.

Fahrmeir, L., Wagenpfeil, S., 1996. Smoothing hazards and time-varying effects in discrete duration and competing risk models. Journal of the American Statistical Association 91, 1584–1594.

Kalbfleisch, J.D., Prentice, R., 1980. The Statistical Analysis of Failure Time Data. Wiley, New York.

Robins, J., 1997. Causal inference from complex longitudinal data. In: Berkane, M. (Ed.), Lecture Notes in Statistics, Latent Variable. Modeling and Applications to Causality, vol. 120. Springer, New York, pp. 69–117.

Robins, J.M., Hernán, M.A., 2009. Estimation of the causal effects of time-varying exposures. Longitudinal Data Analysis 553–599.

Sandefur, G.D., Tuma, N.B., 1987. How data type affects conclusions about individual mobility. Social Science Research 16, 301–328.

Stasny, E.A., 1986. Estimating gross flows using panel data with nonresponse: An example from the Canadian labor force survey. Journal of the American Statistical Association 81, 42–47.

Thompson Jr., W.A., 1977. Treatment of grouped observations in life studies. Biometrics 33, 463–470.

Longitudinal Research: Panel Retention

Jennifer A Parsons, University of Illinois at Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.B. Warnecke, J.A. Parsons, volume 13, pp. 9067–9071, © 2001, Elsevier Ltd.

Abstract

Maintaining a panel in longitudinal surveys requiring reinterviews of participants over time is the focus of this article. The consequences and correlates of panel attrition are reviewed, and strategies for retaining sample members throughout the interview period are offered. Specifically, a set of recommendations for the initial design of panel surveys and tracking respondents between interview waves in order to minimize bias due to attrition is offered.

Introduction

Panel studies involve the collection of data over time from the same sample of respondents. Unlike other forms of longitudinal studies, panels allow for the study of individual behavior change over time. However, because the same individuals are followed, there is eventual attrition, or nonresponse, after the baseline data collection wave. Attrition is either due to the researcher's inability to locate the respondent for additional waves of data collection, or when located, the respondent declines participation. Since the value of panel surveys is dependent upon the ability to study the same respondents at different points in time, reducing attrition is of major concern in social and behavioral research. Loss of respondents over time raises the possibility of bias if those who are lost to follow-up differ from those who remain in the panel on key dependent variables. Therefore, panel attrition can affect both the internal and external validity of the study (Cook and Campbell, 1979). In this article, the author reviews some recommended strategies for minimizing panel attrition with general populations. Some of the strategies depend on institutions like the post office whose functions vary from country to country. The information provided here is about the United States, but it is expected that much of it is adaptable to other countries. The recommendations discussed here for minimizing panel attrition apply mainly to population-based samples. Panel surveys with more hard-to-locate or high-risk populations (such as the homeless, drug users, or runaway youth) will require more specific, tailored strategies (Wright et al., 1995; Scott, 2004).

Factors Associated with Panel Retention

Panel survey success depends on location, contact, and cooperation (Lepkowski and Couper, 2002). Failure to locate a respondent after the baseline wave can be attributed to residential mobility. According to the 2009 American Community Survey, 15.4% of the US population aged 1 year and over lived in a different residence than 1 year ago (Ihrke et al., 2011). The majority (67%) moved within the same county, 17% moved to a different county in the same state, and 12.6% moved to a different state. Age is correlated with moving; the highest rates of residential mobility are found among those aged 18–29 years, and they decline as age increases. Rates of residential mobility also vary with race and ethnicity; they are the lowest for non-Hispanic whites (13.8%) and the highest among Hispanics (17.9%). People who rent their homes are more likely to relocate than those who own their residences. Nearly one-third of people living in rental units reported having moved in the previous year compared with 6.7% living in owner-occupied housing units. Understanding these correlates of residential relocation will help investigators plan for the level of tracking they will need to do during follow-up interviews, and also guide the amount of locating and secondary source tracking information that is collected at baseline. Including these variables as indicators at the baseline wave of data collection is important for postsurvey attrition assessment and weighting to adjust for it.

When respondent location is successful, contact is the next step. Difficulty in making contact with successfully located panel subjects is a greater factor in the first follow-up waves than in later waves (Lepkowski and Couper, 2002). Detailed information on the successful subject contact from the baseline wave (number of contact attempts, time and day of interview, etc.) can be used to increase successful contact at subsequent waves. Once contact has been established, the subject's cooperation is largely going to be based on past experience with the study: salience, perception of burden, and degree of 'panel fatigue,' which is the term used to describe subjects who drop out of longitudinal studies after several waves of data collection due to boredom or loss of interest in the study.

Study Design Features to Maximize Panel Retention

There are three main factors that will affect the degree of attrition in any panel study: (1) recruiting the respondent into the study; (2) successfully locating the respondent for subsequent interviews; and (3) maintaining the respondent's commitment to the panel.

Recruiting Respondents to the Panel

Plans for recruitment and retention should be developed at the time the study is designed and budgeted. They are as important as any other feature of the study design, but often do not get the

serious attention – and importantly, funding allocation – they deserve. Preplanning will ensure not only that there are sufficient resources to successfully recruit and retain the respondents, but also that the information needed to track respondents is collected during the baseline interview. It will also guide the selection of the sample size since this must allow for attrition if locating respondents is anticipated to be a serious problem. It may also affect decisions about the mode of interview, because certain segments of the population may not be available by one mode or another. For example, a longitudinal study of seniors may net few successful follow-up contacts via e-mail, whereas e-mail, Facebook, and other social technologies would be critical for college-aged populations.

Successful recruitment and retention can be enhanced with the creation of a recognizable identity for the study. A study logo and identifiable study title or acronym that is used in all correspondence (written and oral, including the study's Web site) and that appears on self-administered questionnaires, where applicable, have also been suggested as a way to create an identifiable image of the study for the respondent (Given et al., 1990). With the right population, a study Facebook page can also be used to provide general information and next steps for study participants.

Depending on the sample frame, an advance letter and/or advance e-mail should be sent to respondents at baseline, prior to the first interviewer contact, which explains the purpose of the study, and where they can go for further information (e.g., a phone number or study Web site, depending on the population). Clearly, advance letters or e-mails require that the investigator have these addresses for the population; if unavailable, they should be obtained at the baseline interview for follow-up purposes.

Locating and Tracking Respondents for Follow-Up Waves

Once respondents are recruited into the panel, several methods can be used to increase the likelihood of contact at designated follow-up waves. These generally can be classified into proactive, field-based, and centralized tracking efforts (Couper and Ofstedal, 2009). Attempts at locating the respondent should persist until a final disposition of the case can be ascertained for the purposes of calculating response rates.

The proactive approaches refer to the essential information that needs to be collected at each interview wave in order to increase the likelihood of locating them later. During the baseline interview, information should be obtained that will help locate respondents later. One invaluable resource is the secondary contact source. Secondary sources are usually one or two persons, not residing with the respondent, who will know how to contact the respondent at any point in time. At a minimum, the relationship of the secondary source to the respondent and that person's telephone number and e-mail address should be collected. Allow for the contingency that the subject may want to get their source's permission to share their contact information first; this will involve follow-up calls, and perhaps mailings, to ensure that the subject secured this information. A form in a postage-paid envelope could be left with the respondent to complete and mail back once they have received permission from their contacts.

In addition, during the baseline interview all respondents should be asked to provide their dates of birth, social security numbers (for use with credit-based tracking programs), and driver's license numbers. Depending on the population, respondents' e-mail addresses should also be collected as they can be used as an inexpensive means of tracking respondents. Note that the subject will need to be assured that this information will only be used for tracking and locating purposes. In addition, depending on the population, you can ask for the subject's permission to contact them via social media (Facebook and Twitter).

Even the most conscientiously designed study will not prevent respondents from being lost to follow-up. If all secondary source information does not lead to the whereabouts of the respondent, there are some field-based strategies that can be employed. In face-to-face studies, an excellent source of information about respondents comes from people who are connected in some way with the respondent. Roommates, neighbors, property managers, mail carriers, family, and friends might be available to assist in searching for, and contacting, a hard-to-find study participant. When finding the respondent through direct contact does not lead to complete or solid information, personal informants can sometimes provide essential information about how to reach the respondent, what kind of schedules they keep, and whether they have moved.

In addition to 'address correction requested' stamped on advance letter envelopes, a number of other useful services are available through the United States Postal Service (USPS) to improve mailing addresses and locate respondents who have moved. Through the Address List Correction Service, business and resident mailing lists can be submitted for correction. Services provided include cross-listing of names to which mail cannot be delivered or forwarded; correcting ZIP codes, house numbers, rural, or post office box numbers, street names, and suffixes; furnishing new addresses for customers who have moved; and adding building, apartment, suite, or room numbers when known. There is a minimum charge for correcting each submitted list, which can be arranged through local post office branches. In addition, the USPS National Change of Address (NCOA) program processes change of address data nationwide. The NCOA service is provided by private sector companies, certified and licensed by the USPS, who charge a fee for the service. These companies receive updated, computerized COA information on a regular basis. To search for lost or unlocatable study respondents, a file of names and addresses is submitted to one of the NCOA vendors, who search for matches in the NCOA. New addresses are provided when an exact match is found. Most other countries will have systems with many of these attributes.

Moreover, centralized tracking may also be useful in locating respondents. Property records, credit bureaus, and vital statistics databases may be accessed, usually for a fee. It is important to investigate these databases during the planning stages, not only to properly allocate fees in the budget for searchers, but also to ensure that the correct information is collected. Accessing public records is necessary for obtaining missing information about respondents. Public records contain data on births, deaths, marriages, criminal activity, military service, motor vehicle registration, and employment. Of particular use in panel studies that span a long period of time is the National Death Index,

maintained by the National Center for Health Statistics. For a fee, the index will match respondents with their file of deceased persons using the name and date of birth. The likelihood of a match increases if the social security number or middle initial is available. In some states, it is possible to search a Web site from the department of corrections to ascertain if any study participants have been incarcerated.

Retaining Respondents between Interview Waves

Being a member of a panel is burdensome for the respondent. Thus, it is critical to the successful maintenance of the sample that the relationship be reinforced between contacts.

Regardless of the duration of time between data collection waves, it is advisable to maintain contact with respondents every 3–6 months as resources allow, relative to the total length of data collection. Regular contact keeps the respondents engaged and reminds them of the value placed by the investigator on their participation. It also helps the investigator keep track of the respondent. The extent of the contact need not be more than the advance letter or a thank you note. Besides a friendly greeting and an expression of thanks for the respondents' participation, the content should include a reminder of the time of the next follow-up survey and a request for new contact information if the respondent has moved or will be moving. Investigators can also personalize the contact between interviewing waves by using birthday cards, holiday greetings, or other forms of written correspondence (Walton et al., 1998). These formats reinforce that the study staff is interested in and cares about the participants.

The Use of Incentives to Recruit and Retain Subjects in a Panel Study

Overall, the literature favors the use of incentives in cross-sectional surveys (Church, 1993; Singer et al., 1999), particularly prepaid, monetary incentives, to increase response rates; these results are generally consistent across all modes of data collections and different populations (Toepoel, 2012). Incentives are commonly used in longitudinal studies as well as a strategy for encouraging participation, given the respondent burden associated with repeated involvement and the value of the subject's ongoing participation to the researcher (Laurie and Lynn, 2009). Not only do they represent a concrete expression of appreciation to the respondent, but they also have the effect of keeping the respondent interested and involved over time.

The work reported by Singer et al. (1999) is also applicable to panel respondents. It suggests that informing the respondent at the start of the study that they will be offered an incentive for participation increases cooperation at later waves. Thus, at the conclusion of each interview, and before collecting the data that will help track the respondent at subsequent waves, it is useful to inform respondents that they will be paid an incentive for their future participation. However, it is important to emphasize that the respondent is not being paid for the data just provided. Rather, the incentive is a token of appreciation for having provided something that is invaluable and therefore cannot be purchased.

It is not uncommon for longitudinal surveys to increase the incentive amounts with each wave of participation, or to pay a bonus if all waves are completed. For longitudinal surveys that span several years, an increase in the incentive payment at each wave acknowledges cost of living increases and is a symbolic gesture to the respondent that their ongoing cooperation is valued and important (Laurie and Lynn, 2009). Anecdotal evidence also suggests that respondents are more likely to notify investigators of new addresses when they have moved if they expect an incentive, especially when the subject receives between-wave reminders from the investigator.

Summary

Panel retention is an important part of longitudinal studies that involve reinterviewing a cohort of respondents over time. While participation in a panel is a burden for the respondent, maintaining an unbiased panel is a challenge for the investigator. Thus, before such a project is undertaken, planning is critical and the level of effort and associated costs need to be considered from the outset.

A key aspect of a successful panel is the study design. The sample, questionnaire, and associated budget must all accommodate the need to recruit an unbiased panel and maintain their participation over a lengthy period of time. Recruitment involves attracting individual respondents to agree to participate in the baseline and to be followed for some period of time afterward. Locating respondents is critical to ensuring the continued quality of the panel data. Finally, recruiting and maintaining the panel will be facilitated if respondents develop a commitment to it. Ensuring that the panel is an unbiased sample of the population of interest requires effort to not only convince individuals to participate but provide sufficient information to enable the panel to continue.

Achieving a high level of cooperation to any survey has become increasingly challenging over the last decade; the Pew Research Center reports that their telephone surveys that would have had a response rate near 40% in the late 1990s was lower than 10% by 2012 (Kohut et al., 2012). A successful longitudinal study will require careful planning and budget allocation from the outset. Plans for follow-up must be established at the onset of the study so that not only all the relevant information is collected from the respondents as necessary resources are allocated, but also detailed nonresponse analyses can be performed. Without sufficient data on respondents, when they change residence or telephone, they are likely to be lost to follow-up, introducing bias into the data. Convincing respondents that they are important to the success of the panel and that the investigator values their participation ultimately helps reduce the cost of locating and following the respondents. Overall, participation in research is a process of exchange. Thus, it is important to show respect and consideration for the sources of the data.

See also: External Validity; Longitudinal Causal Inference; Missing Data; Panel Surveys: Uses and Applications; Reliability and Validity Assessment: New Approaches; Selection Bias, Statistics of.

Bibliography

Church, A., 1993. Estimating the effect of incentives on mail survey response rates: a meta-analysis. Public Opinion Quarterly 57, 62–79.

Cook, T.D., Campbell, D., 1979. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Houghton-Mifflin Company, Geneva, IL.

Couper, M., Ofstedal, M., 2009. Keeping in contact with mobile sample members. In: Lynn, P. (Ed.), Methodology of Longitudinal Surveys. Wiley, New York, pp. 183–203.

Given, B., Keilman, L., Collins, C., Given, C., 1990. Strategies to minimize attrition in longitudinal studies. Nursing Research 39 (3), 184–186.

Ihrke, D., Faber, C., Koerber, W., 2011. Geographical Mobility: 2008 to 2009. Current Population Reports, P20-565. U.S. Census Bureau, Washington, DC.

Kohut, A., Keeter, S., Doherty, C., Dimock, M., Christian, L., 2012. Assessing the Representativeness of Public Opinion Surveys. The Pew Research Center for the People and the Press, Washington, DC.

Laurie, H., Lynn, P., 2009. The use of respondent incentives on longitudinal surveys. In: Lynn, P. (Ed.), Methodology of Longitudinal Surveys. Wiley, New York, pp. 205–233.

Lepkowski, J., Couper, M., 2002. Nonresponse in the second wave of longitudinal household surveys. In: Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (Eds.), Survey Nonresponse. John Wiley and Sons, New York.

Scott, C.K., 2004. A replicable model for achieving over 90% follow-up rates in longitudinal studies of substance abusers. Drug and Alcohol Dependence 74, 21–36.

Singer, E., Van Howeyk, J., Gebler, N., Raghunathan, T., McGonagle, K., 1999. The effect of incentives on response rates in interviewer-mediated surveys. Journal of Official Statistics 15 (2), 217–230.

Toepoel, V., 2012. Effects of incentives in surveys. In: Gideon, L. (Ed.), Handbook of Survey Methodology for the Social Sciences. Springer, New York, pp. 209–223.

Walton, M.A., Ramanathan, C.S., Reischel, T.M., 1998. Tracking substance abusers in longitudinal research: understanding follow-up Contact difficulty. American Journal of Community Psychology 26 (2), 233–253.

Wright, J.D., Allen, T.L., Devine, J., 1995. Tracking non-traditional populations in longitudinal studies. Evaluation and Program Planning 18 (3), 267–277.

Long-Term Potentiation and Long-Term Depression

Tammy L Ivanco, University of Manitoba, Winnipeg, MB, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Ito, volume 13, pp. 9071–9074, © 2001, Elsevier Ltd.

Abstract

Long-term potentiation (LTP) and long-term depression (LTD) have been studied extensively as experimentally induced phenomena in the brain, and may share mechanisms with learning and memory, which most of us are more familiar with. Stimulation to induce LTP and LTD, an increase or decrease in synaptic strength, respectively, has indicated a sequence of overlapping molecular events from induction to long-term maintenance of the change in synaptic strength. Many studies have focused on the roles of *N*-methyl-D-Aspartate and mGluRs in LTP and LTD, and the subsequent intracellular cascades the receptors help control. LTP and LTD are associated with mechanisms that indicate a complex and flexible memory system, which is also dependent on the transition from simple to more complex connections at synapses. As the mechanisms in LTP and LTD parallel to those found with learning and memory, there is potential that LTP and LTD studies may reveal how we learn and remember, and how learning and memory may fail. Pharmacological manipulations that are used to study LTP and LTD may suggest potential drug treatment for brain disease and aging, and LTP/LTD stimulation protocols may become useful in the treatment of brain disorders, such as Epilepsy and Fragile X Syndrome.

Introduction

Long-term potentiation (LTP) and long-term depression (LTD) have been broadly studied in many brain areas and circuits. LTP and LTD are characterized by the persistent increase or decrease in synaptic strength, respectively. LTP, discovered in the early 1970s, met with much acclaim because it satisfies criteria proposed for a synaptic mechanism involved in learning and memory proposed in Donald Hebb's book "The Organization of Behavior." Hebb had suggested that learning and memory were the outcomes of the selective strengthening of synapses in response to activity. Later, theories on the weakening of synapses were also proposed, which could provide the basis for forgetting and general memory loss. As LTP and LTD have characteristics that mirror learning and memory, and are easily studied in the laboratory, they rapidly became popular models of synaptic plasticity. Evidence suggests the modification of connections throughout brain circuitry occurs in response to repeated experiences using LTP- and LTD-like processes. Whether LTP and LTD are actual synaptic processes underlying learning and memory remains to be determined, but research suggests many forms of memory rely on similar molecular mechanisms also used by LTP and LTD.

Generally, there are two approaches to inducing lasting change in synaptic connections experimentally to simulate the brain's response to repeated experiences. The first approach includes using high- or low-frequency electrical stimulation to induce LTP or LTD, respectively. The second includes causing activation though pharmacological or alternate methods to mimic the stimulation induction protocols used to induce LTP or LTD. More recently, LTP and LTD have been investigated as tools to assess and treat disorders and diseases of the central nervous system, particularly in transgenic mice. A number of neurological disorders and diseases appear to result from a loss or increase in synaptic excitability, resulting in changes in connections to and between brain regions. The ability to use both old and new technologies to influence synaptic strength is promising for altering synaptic plasticity and promoting better long-term outcome in disease and even with aging.

History

Virtually, all neuroscience textbooks highlight the basic rule proposed by Donald Hebb (1949), citing something like "When cell A repeatedly takes part in firing cell B, the connection between the two becomes strengthened." It took over 20 years for an instantiation of Hebb's Rule to be demonstrated experimentally, however. Per Andersen had laid some of the early ground work for the discovery with his studies on frequency potentiation (Andersen et al., 1966), in which repeated stimulation of the perforant path input to the dentate gyrus resulted in increased excitability of the granule cells lasting several seconds. Andersen and his graduate student (Andersen and Lomo, 1967) had found that basic frequency potentiation was insufficient to account for some of the plastic changes necessary for learning and memory because the enhancements were short lived, but the idea for trains inducing a lasting effect was evident (reviewed in Lomo, 2003). After Bliss joined the lab, the experimental support for Hebb's Rule came when LTP was observed (Bliss and Lomo, 1973). The two studied the effect of high-frequency stimulation in the *in vivo* hippocampus of anesthetized rabbits at synapses between the medial perforant pathway and the granule cells of the dentate gyrus. In this study, a stimulating electrode had been implanted into the perforant path and a brief, high-frequency train of electrical pulses was used to induce action potentials in the afferent pathway of the dentate gyrus. The result ensured coincident pre- and postsynaptic depolarization at the perforant path-dentate gyrus synapses – cell A repeatedly firing cell B – as described by Hebb for synaptic strengthening. Recording electrodes implanted into the cell body layer of the dentate gyrus recorded the synaptic response in the form of an excitatory postsynaptic potential (EPSP), induced by the activated population of

granule cells. These recording electrodes also revealed the lasting synaptic enhancement resulting from the tetanus (high-frequency stimulation pulses). In this preparation, LTP lasted between 30 min and 10 h. Bliss and Gardner-Medwin (1973) retested LTP in the awake and freely moving rabbit hippocampus, using similar parameters. In the chronically implanted preparation, LTP lasted between 1 h and 3 days, sufficient to allow for comparisons with short-term memory.

The phenomena of LTP represented an entirely new way to investigate plasticity in the mammalian brain. Ensuing experiments examining the intact hippocampus focused primarily on rats and mice. The development of the transverse hippocampus slice preparation, with the easy placement of electrodes in defined areas and patch clamp opportunities, resulted in the ability to experimentally focus on determining the rules for LTP induction and the molecular mechanisms of LTP, and subsequently LTD. An *in vitro* preparation, or slice preparation, requires the dissection of a very thin slice of brain tissue (most often hippocampal brain tissue) that is placed into chamber with oxygen and a cocktail of chemicals that keeps the tissue viable for several hours. Research has now demonstrated that LTP can be induced across the brain, in subcortical (e.g., Racine et al., 1983) and cortical locations (e.g., Ivanco et al., 2000), and even between these locations (e.g., Ivanco and Racine, 2000). Investigators considered the rules determined in LTP studies and sought out whether LTD followed similar rules. The answer is both yes and no.

Alger and Teyler (1976) found low-frequency monosynaptic stimulation between the perforant path and dentate resulted in a long lasting depression of the evoked response in hippocampal slice preparations. LTD has now been reported in other pathways of the hippocampus and cortex, *in vivo* and *in vitro* preparations (e.g., Froc and Racine, 2004). Since the first early reports of LTD, it has become clear that LTD can be divided into two types. The first type is LTD that is observed *de novo* in response to low-frequency stimulation. The second type is LTD that is observed only after the induction of LTP. Often the second category is called depotentiation, due to the enhanced response of the synaptic connections that must first be generated. Further, with depotentiation, the response does not depress beyond the baseline evoked response. There is some debate whether LTD is really a mixture of LTD and depotentiation.

Activity-dependent modification of synaptic connections is fundamental to the storage of information within the brain, but there may be additional considerations when studying plasticity consistent with how we seem to store information. Bidirectional modifications in synaptic plasticity are important in models of learning and memory, as they highlight both our ability to learn new information, but also forget or lose access to information that is no longer used. The ability of new and old information to interfere or help with learning indicates the underlying mechanisms must be very dynamic. The theory of metaplasticity (reviewed by Abraham and Bear, 1996), which can be thought of as a higher form of plasticity, suggests plasticity can be modulated or modified by prior synaptic activity. Evidence for an electrophysiological correlate comes from studies investigating the induction of LTP by a strong tetanus after first providing a weaker tetanus. Huang et al. (1992) reported that a weak stimulus delivered to the Schaffer Collaterals prior to a stronger tetanus inhibited LTP in the CA1 region of the hippocampus. They suggest that a low-level activation of the *N*-methyl-D-Aspartate (NMDA) receptors resulted in this inhibition of subsequent LTP. The low-frequency priming stimulus can also facilitate LTP induction by high-frequency trains if the priming tetanus is delivered at an intensity that is near the threshold for inducing LTP (Christie et al., 1995). The data suggest a complex interaction between low- and high-frequency stimulation. Further adding to this complexity, a priming stimulus has been shown to result in LTD. Christie and Abraham (1992) discovered an associative form of LTD occurred in the lateral perforant path to dentate gyrus pathway. Associative LTD only appeared when the priming, low-frequency stimulation was provided before and out of phase with the high-frequency stimulation, which alone would induce LTP. One of the general predictions from this metaplasticity argument is that the prior induction of LTP would also facilitate LTD. Indeed, synaptic depression in the form of LTD is more reliably produced in the CA1, if LTP is first induced (e.g., Staubli and Lynch, 1990). Exploring metaplasticity may actually prove valuable for noninvasive stimulation techniques for use in diseased or damaged brain. LTP and LTD allow for a set of dynamically altered connections, providing evidence for a very flexible memory system.

Induction

LTP and LTD induction is the sequence of early events that lead to the initial increase or decrease in synaptic efficiency, respectively. Racine et al. (1983) provided some of the first evidence that LTP in the hippocampus and other brain regions was not a unitary process, but involves a sequence of overlapping phases that rely on different mechanisms. Racine et al. (1983) defined an early LTP phase (LTP1, also commonly called E-LTP) lasting 1–2 h in the chronically implanted animal and a late phase of LTP (LTP2, also commonly called L-L) lasting about 5 days. The mechanisms involved in LTP induction are distinct from those involved in the phase that leads to persistent LTP and a sequence of synaptic changes parallel to the electrophysiological phases. Most of the research on the processes of LTP induction has resulted from the study of the CA3-CA1 synapses and the synapses between the medial perforant path and dentate gyrus, but generally the principles found in these synapses are consistent with other areas. Focus on these same synapses indicated, like LTP, LTD occurs as a sequence of overlapping phases. In both LTP and LTD, the postsynaptic NMDA receptors were found to be highly involved in the plasticity. Before discussing how receptors come into play, some of the characteristics of LTP and LTD may be helpful for understanding their actions and why particular receptors have been most well studied.

Input specificity and associativity fulfill criteria for an efficient memory system proposed by Hebb (1949) and also provide evidence for mechanisms underlying LTP and LTD. Independently stimulated afferent pathways demonstrate synapses that are located close together may still be independently potentiated or depressed, indicating input specificity. Using the same approach of independent stimulation, it

was found weak stimulation to a pathway that would not induce LTP on its own was sufficient when it was paired with stronger stimulation to a second pathway with synapses on the same target cells, indicating associativity similar to the associate nature of human memory first implied by Hebb. A third characteristic of LTP is cooperativity, which refers to the idea that when a train of stimulation (e.g., 200 Hz) is delivered to a pathway it induces depolarization in the postsynaptic side at the same time a new pulse reaches the presynaptic terminal. Cooperativity suggests that the synapse has a coincidence detector that both detects the postsynaptic depolarization and the next pulse from the stimulation train. The basis for cooperativity is the action and mechanism of the NMDA receptor. Like LTP, some forms of LTD also rely on the action of the NMDA receptor at many synapses.

The NMDA receptor is a glutamate receptor located on the postsynaptic terminal that is a calcium permeable channel activated by glutamate, but only when the postsynaptic membrane potential is sufficiently depolarized (reviewed in Bliss and Cooke, 2011). Depolarization allows for a magnesium ion blocking the channel to be expelled and glutamate released from the presynaptic terminal is bound, resulting in calcium and other cations into the synaptic terminal. The calcium influx into the cell through the NMDA receptor is essential for the induction of LTP and LTD. Indeed, the application of calcium chelators intracellularly prevents LTP (e.g., Lynch et al., 1983) and LTD (e.g., Mulkey and Malenka, 1992). The time frame for the NMDA-receptor mediated calcium influx is brief, but the LTP inducing tetanus (another term used for the stimulation train inducing plastic change) may activate receptors repeatedly, if the train is sufficiently long. The rise in intracellular calcium activates calcium-dependent enzymes. One of these is the calcium calmodulin-dependent protein kinase II (CaMKII), which has been noteworthy in the learning and memory literature, as well as the LTP literature. The NMDA receptor is actually a complex made of subunits and different LTD stimulation protocols seem to recruit the action of subunits differently. NMDA receptors are triggers for synaptic plasticity, but do not themselves determine the change in synaptic efficacy (reviewed in Collingridge et al., 2010). Thus, there must be other mechanisms that are involved in determining whether the resulting change following stimulation is an increase or a decrease in synaptic efficacy. One of the more recent discoveries is the incorporation of α-amino-3-hydroxy-5-methylisoxazole-4-propionic acid (AMPA)-type glutamate receptors into synapses that previously lacked them following coincident pre- and postsynaptic activity inducing LTP. Granger et al. (2013) investigated the AMPA receptor in LTP induction and found LTP was reduced when there was a decrease in AMPA receptors. LTP induces AMPA receptors located in a reserve pool until the LTP expression moves from the receptor level to changes in the synaptic structure. Some forms of LTD may reflect the removal of AMPA receptors from the synapse or internalization of NMDA receptors with low-frequency stimulation and chemically induced LTD, respectively (reviewed in Collingridge et al., 2010). Blockade of the NMDA receptor with antagonists prevent the induction of LTP, but have no effect once LTP is established, indicating an effect limited to the induction phase. NMDA-dependent LTP and LTD are the most commonly discussed forms of synaptic plasticity, but there are other forms of LTP and LTD that do not rely on the NMDA receptor. Froc and Racine (2004) found cortical LTP in the chronically implanted rat was dependent on the NMDA receptor, but depotentiation and LTD were not. Pairing of pre- and postsynaptic depolarization results in coincidence detection in the hippocampus and elicits LTP, but the same form of stimulation in the cerebellum results in LTD at synapses between parallel fibers and Purkinje cells, indicating alternate mechanisms in some synapses. Although a number of receptors have been tested, the mechanisms underlying induction of LTP and LTD in a range of brain areas continues to be investigated.

One characteristic that may lead to confusion in the LTP/LTD literature is related to how long the induction phase lasts. Although the induction phase is described as occurring prior to the expression and maintenance phases of LTP and LTD, the preparation being used has significant impact on how the timelines are described. In the transverse slice preparation, the induction period may be described as lasting minutes to about an hour. In the intact, but anesthetized preparation, the timeline for induction is similar to that of the slice (see **Figure 1**). The real difference occurs in the intact freely moving animal, with electrodes implanted and cemented in place so the animal can be anesthetic-free during induction, which can be done over a much longer period. In the case of the intact, freely moving animal, the induction phase for LTP may be minutes to few hours in the hippocampus (e.g., Racine et al., 1983), or even days for cortical LTP induction (e.g., Ivanco et al., 2000). The different preparations have provided support to the theory that hippocampal LTP is similar in nature to short-term memory, which is rapid, but information for long-term storage transferred to cortical areas through consolidation, and cortical LTP is similar in nature to long-term memory, which is more permanent, but requires more activation over time. Ivanco and Racine (2000) also found the structures between the entorhinal cortex, the origins of the perforant path, and the cortex could support LTP indicating an LTP-like mechanism could be used for consolidation.

A second challenge in reviewing the literature is defining the transition between induction, which seems mostly receptor dependent, and a more persistent form of synaptic change that is more dependent on protein synthesis and more obvious structural changes. Weeks et al. (2000) investigated morphological correlates of the induction phase by examining the synaptic morphology 1 h following LTP inducing stimulation in the chronically implanted rat. With this delay, the authors report no change in total synaptic number, but a change in the synapse profile, where more perforated (complex synapses) and irregularly shaped synapses were present in the dentate gyrus after perforant path stimulation to induce LTP. Concave synapses were found to be larger, as well. Although the preparation is different, Oh et al. (2013) found shrinkage in the size of individual spines within 60 min in response to low-frequency stimulation leading to LTD in hippocampal slice cultures from young rats. Structural changes in the window of the induction phase suggest an enlargement of synapses occurs with LTP and shrinkage of synapses with LTD. Thus, morphological changes can occur within the window defined as the induction phase.

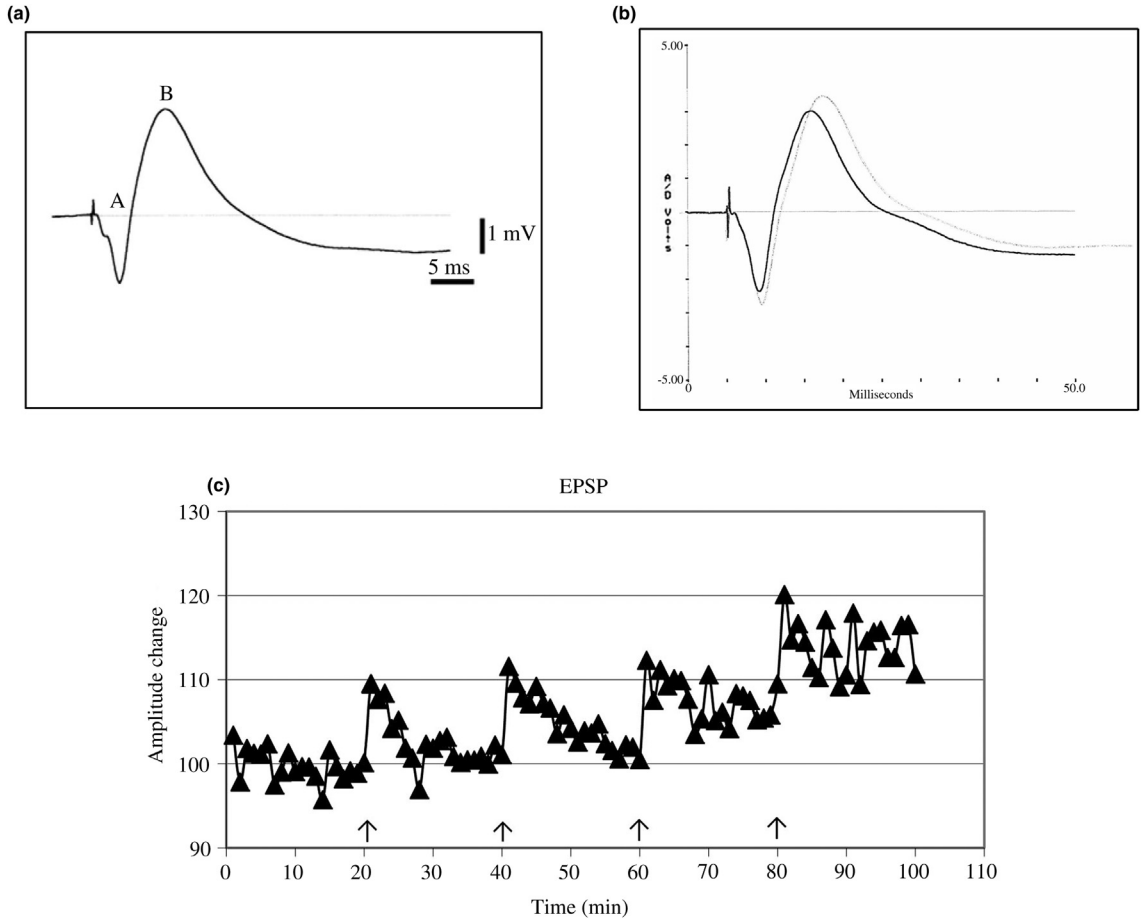

Figure 1 Induction of cortical long-term potentiation (LTP) in the intact anesthetized preparation. An evoked potential from the rat cortex is shown in (a). Within (a), the population spike recorded in this potential is represented by A and the excitatory postsynaptic potential (EPSP) is represented by B. An evoked potential before (solid line) and after LTP induction (dashed line) is shown in (b). The time course for induction is shown in (c) with the arrows representing stimulation train delivery. LTP in this preparation is not immediate, but requires several trains to be observed.

Expression

Calcium detection mechanisms are involved in the transition from LTP and LTD induction to expression. The activation of calcium CaMKII and cyclic AMP (cAMP)-dependent kinase (PKA) is an essential step in the process of NMDA-dependent LTP in the hippocampus, and likely other areas where NMDA-dependent LTP is induced. The kinases are able to detect elevations in calcium and in turn phosphorylate proteins involved in expression of LTP, such as AMPA receptors. The glutamate receptor 1 (GluR1) subunit of the AMPA receptor is regulated by CaMKII and PKA. Lee et al. (2000) investigated the role of the GluR1 subunit on bidirectional communication using LTP and LTD induction protocols, but examined both naïve synapses and synapses that had a stimulation history. LTP induction in naïve synapses and depressed synapses increased phosphorylation to the CaMKII site and to the PKA site, respectively. LTD induction in naïve synapses and previously potentiated synapses led to dephosphorylation at the PKA site and CaMKII site, respectively. The data suggest LTP and LTD are not simply the functional inverse of each other, but that they have unique roles in modifying the AMPA receptor GluR1 subunits.

With the persistent activation of the detection mechanisms and kinase activation, a cascade of signaling events result in the production of new proteins and changes in gene expression (e.g., Dash et al., 1991) and allow LTP and LTD to have sustained expression beyond the first few hours. With the transition of early plasticity (E-LTP and E-LTD) to late plasticity (L-LTP and L-LTD), the signaling involved moves from the synapse itself to the cell nucleus, which is able to initiate the novel gene transduction protocols necessary for a longer term plasticity. This transition is accomplished by a complex cAMP signaling cascade that recruits PKA, mitogen activated protein kinases (MAPK), and cAMP-response element-binding protein (CREB), as well as a host of modulatory neurotransmitters including dopamine, serotonin, noradrenaline, and acetylcholine acting on their receptors.

Although most of the literature on the transition of LTP/LTD from an expression to a maintenance phase focuses on the processes occurring in the cell body, there have been recent findings that suggest that cells also have dendritic machinery that allows for a more selective set of alterations that would occur at the activated synapses independently. Weiler et al. (2004) found that the Fragile X Mental Retardation Protein

(FMRP) was involved in neurotransmitter-activated protein translations at the synapse. The translation of existing or new mRNA at polyribosomes located in dendrites or in the synapse would drastically reduce the time and need for additional proteins for plastic change at the synapse after activation. One of the more recent findings is the upregulation of micro-RNAs (miRNAs) in response to stimulation (e.g., Lee et al., 2012). The actions of the miRNAs are not fully known, but they may act to upregulate or downregulate targets to influence spine structure and function, and change plasticity locally at synapses.

Less research has focused on the structural alterations in this intermediate phase, than immediately after LTP induction or in the longer term. Weeks et al. (1999) investigated synaptic profiles 24 h after LTP induction in chronically implanted animals. LTP was not associated with an increase in synaptic number, but there was an increase in the proportion of presynaptically concave-shaped synapses 24 h following LTP induction. The concave synapses were smaller in the stimulated tissue than they were in control tissue and the synapses were more likely to be perforated (suggesting a more complex form). In second paper Weeks et al. (1998) used the same tissue, but correlated synaptic number with the degree of LTP induced. Although synapses per neuron had not shown up as significant in the initial study, synapses per neuron were significantly related to the degree of potentiation. Since potentiation varies in each animal dependent on electrode placement and the targeted synapses, it is noteworthy that as more LTP is induced, more synapses are present. Ivanco et al. (2000) examined spine density and dendritic arborization 24 h after the last tetanus of LTP inducing stimulation in the somatomotor cortex in chronically implanted rat. LTP expression in this area was correlated with greater spine density and alterations in the dendritic tree, which is consistent with anatomical changes induced by animal exposures to enriched environments. The data from the anatomical studies are consistent with a pattern of structural changes (reviewed in Weeks et al., 2001), whereby LTP induction, expression, and long maintenance are paralleled by an increasing level of overall complexity in the synaptic connections. There have been fewer studies of synaptic ultrastructure following LTD inducing stimulation.

Maintenance

CaMKII has been emphasized as one of the mechanisms involved in the transition to long-term maintenance for LTP. CaMKII can continue to phosphorylate itself, so it has the ability to remain active for a long time without other events occurring, even after calcium levels return to normal. In many ways, it is the memory within the system. Less, however, is known about the mechanisms leading to the maintenance of LTD.

Brain-derived neurotrophic factor (BDNF) is made in response to the MAPK and CREB signaling following presynaptic glutamate release. BDNF is a protein that regulates different cellular events, ranging from selection of neural progenitors to the terminal differentiation of dendrites and connectivity of neurons (Yoshii and Constantine-Paton, 2010). BDNF plays a role in both LTP and LTD, but not by playing an inverse role (reviewed in Bliss and Cooke, 2011). Early after it is made, BDNF (pro-BDNF) has the ability to bind to the tyrosine kinase B (TrkB) receptor and the P75 receptor. Activation of the TrkB receptor by BDNF binding encourages LTP induction. Binding of BDNF to the P75 receptor results in an alteration of the NMDA receptor, subsequently promoting LTD. The role of BDNF is complex, but it is a substance that has garnered a lot of attention as a key player, not just in LTP and LTD, but in learning and memory. Thus, BDNF has provided significant support for the theory that understanding mechanisms involved in LTP and LTD may translate into understanding mechanisms of learning and memory.

An atypical isoform of calcium-dependent kinase PKC, PKMzeta (PKMζ), may act locally to maintain LTP. PKMζ is expressed after LTP induction, but, like CAMKII, it continues to be active, even in the absence of calcium. Migues et al. (2010) found that PKMζ facilitates the maintenance of LTP by blocking a GluR2-dependent pathway and results in the increase of AMPA receptors in the synapse. The additional AMPA receptors at the postsynaptic site encourage additional synaptic transmission. Migues et al. also found evidence for a similar role in fear conditioning, indicating a common mechanism in LTP and learning. Further, they investigated the amygdala and hippocampus, so extended the role of this mechanism across the brain. Inhibition of PKMζ also resulted in a decrease in GluR2, but not GluR1 subunits. Downregulation of PKMζ has been implicated in LTD in the hippocampus 2 h after LTD induction (Hrabetova and Sacktor, 2001). The translocation of other PKC isoforms occurs transiently, but the downregulation of PKMζ is long lasting. PKMζ appears to be bidirectionally modified based on the type of stimulation and afferent activity. Hrabetova and Sacktor suggest that PKMζ may reflect one of the molecular correlates of a sliding threshold system triggered by stimulation where long-term dephosphorylation becomes associated with LTD and increased phosphorylation with LTP. The suggestion unifies a maintenance mechanism for both forms of synaptic plasticity, which would be biologically efficient.

The long-term maintenance of LTP and LTD may depend on long lasting structural alterations at synapses. In tissue collected 5 days after LTP induction in chronically implanted rats, the synaptic profile from stimulated animals indicated longer concave-shaped synapses and more perforations in synapses (e.g., Weeks et al., 2001). The longer concave surface area would support a greater number of NMDA receptors favoring a lasting and efficient synapse, following Hebb's rule.

Studying LTP/LTD in Human Brains

Many of the studies described stress activity-dependent plasticity found in laboratory animal brains. There has, however, been verification of similar forms of plasticity in human brains. One of the challenges for arguing parallel mechanisms in the human brain to the animal brain is that the bulk of the data from humans have been collected from tissue that may have been exhibiting neuropathology and was extracted following temporal lobe epilepsy. Key molecular signaling mechanisms and cells themselves are directly impacted by epilepsy. Despite this limitation, the human brain does exhibit activity-dependent plasticity we have seen in animals.

Beck et al. (2000) collected tissue that had been removed from patients that had drug-resistant temporal lobe epilepsy. A subgroup of the patients' hippocampi did not show severe damage. In this subgroup, LTP could readily be induced in the perforant path to dentate gyrus synapses by theta burst stimulation. A second subgroup with epileptogenic lesions outside of the hippocampus also showed LTP induced by stimulation. This LTP was confirmed to be NMDA dependent because a competitive NMDA receptor antagonist introduced into the slice bath blocked LTP. In patients with hippocampal cell loss or mossy fiber sprouting (a pathological correlate to seizures), but an extrahippocampal epilepsy focus, theta burst stimulation lower levels of NMDA-dependent LTP. Stimulation to induce a depotentiation effect 1 h after LTP, however, resulted in only a partial reversal of the LTP. Pharmacologically, increasing the cAMP concentration also resulted in a lasting LTP in the human hippocampal tissues, but LTP was lower than with stimulation. When the hippocampus was the primary source of the epileptic activity, however, LTP and depotentiation could not be induced, suggesting a severe impairment of the bidirectional communication. Loss of synaptic plasticity is a candidate mechanism contributing to the loss of declarative memory in humans and animals.

One of the treatments known to be effective in drug-resistant depression is electroconvulsive therapy (ECT), which has some parallels with LTP/LTD protocols. In this treatment, stimulation is delivered across the brain with the use of external electrodes, under anesthesia. There are theories as to how ECT might act on the brain to eliminate depression, but one of the most common theories is related to how ECT might reduce cortical excitability in a brain that has increased excitability. Kato (2009) has outlined the molecular cascades resulting from ECT and draws parallels with biochemical markers of LTD seen in animal studies. Thus, ECT may cause the induction of an LTD-like or depotentiation-like process in the active synapses in the cortex. In the clinic, a newer technology is potentially mimicking the effect of ECT for patients with depression. Low-frequency stimulation delivered through repetitive transcranial magnetic stimulation (rTMS) may provide a less invasive method, as it can be delivered without anesthesia and can be used over multiple sessions, and even at home with smaller direct current devices.

Epilepsy results from a hyperexcitability of synapses and may reflect an excessive level of LTP-like activation in glutamatergic synapses. For the same reason the rTMS works in depression, it may have a role to play in epilepsy treatment. Indeed, a method to induce LTD following endogenous LTP induction would have therapeutic benefits. One of the challenges for this type of treatment is related to the location of the primary epileptogenic focus. The coil device used to deliver rTMS is such that penetration into the brain is limited to cortical regions. For foci located in the cortex, rTMS may have more direct outcome, but for deeper foci, such as in the hippocampus, rTMS would be more limited. Data for the use of rTMS in epilepsy is not more fully discussed here, as there are mixed results in the literature, likely depending on the type of epilepsy and focus region. It would be expected the future research using rTMS and magnetic resonance imaging (MRI) to collect structural information and functional MRI (fMRI) to evaluate brain regions impacted by the epilepsy will reveal less mixed data on outcome.

In a review of the literature, Bliss and Cooke (2011) suggest that in intact brains, the readout of plasticity after activation is interpreted as synaptic plasticity, but not directly measured or without potential caveats. There have been debates about whether the synaptic plasticity seen in animal brains as a result of stimulation is paralleled in the human brain in the form of LTP and LTD. One of the future findings they indicate will be important for stimulation protocols to be more widely used in humans to induce LTP or LTD, is noninvasively induced plasticity that follows Hebbian rules. Longevity, input specificity, and associativity are key to a system that follows the rules Hebb laid out and would be necessary to move from LTP/LTD-like to discussing a noninvasive LTP/LTD in the human brain. Further to this is the identification of key synapses that are changed and how they are changed. The knowledge of the synaptic changes and mechanisms would also be important for using a noninvasive treatment for the plasticity-related challenges, the brain encounters both in disease and aging, but may also lead to more targeted pharmacological treatments when noninvasive techniques would be insufficient to recover lost or impaired function.

Putative Roles in Disease and Memory Impairment

Parallel investigations of LTP and memory are a major tool for examining the relationships between synaptic plasticity and memory. There is less investigation into the role of LTD in learning and memory, but one of the reasons for this is that many of the mechanisms involved in LTD and inhibitors of LTD also impact LTP, contributing to a less clear specific role of LTD. Understanding the molecular underpinnings of LTP and LTD in models may translate into a greater understanding of pathogenesis of disease.

Alzheimer's disease (AD) is a progressive neurodegenerative disease characterized by cognitive decline and loss of memory function over time. Additional characteristics include cortical atrophy, neuronal death, neuroinflammation, synaptic loss, and deposits within cells (neurofibrillary tangles) and the amino acid peptide amyloid-beta (Aβ). Although there has not been a disease-modifying drug developed, there has been significant discovery related to AD pathogenesis and how neuronal synapses are impacted by disease progression. Koffie et al. (2011) review the lines of evidence supporting the theory of alteration in amyloid processing leading to AD including the relationships between the Aβ precursor protein (APP) and trisotomy 21 (Down syndrome), and between Aβ and familial early onset AD. Further, Aβ in its oligomeric (multiple polymer) state may be toxic to synapses by inducing a partial blockade in NMDA receptor currents, changing calcium entry, promoting LTD over LTP, and inducing a cascade resulting in the internalization of NMDA receptors to inhibit future LTP. The long-term effects of Aβ would include a morphological change whereby spines would be inhibited from the alterations associated with LTP. Transgenic mice with an APP mutation have increased Aβ and generally have dramatic cognitive impairment. Although at physiologic levels Aβ would not lead to synaptic dysfunction, higher levels would lead to the clinical manifestation of AD. At low levels, Aβ may promote LTP and function, but higher levels lead to synaptic inefficiencies and

changes in the morphology, potentially leading to elimination of synapses, and later neuronal loss.

Neuroinflammation is a characteristic of AD, but neuroinflammation also increases in the healthy brain when an individual ages. Cytokines are immune system molecules that communicate between the immune system and brain. Unfortunately, inflammatory cytokines may also be one of the key contributors to the deficits in synaptic plasticity identified in aging rodents (reviewed in Lynch, 2010). One of the findings in aging rats is that as a rat gets older, there is less LTP. Inflammatory cytokines may impair the ability of aging animals to sustain LTP. Activated microglia are responsible for the release of cytokines, chemokines, and neurotrophins (e.g., BDNF). An important consequence of the release of inflammatory cytokines is a change in the brain that can profoundly effect behavior including a modulation of hippocampal and cortical learning and memory in the rodent (e.g., Larson et al., 2007). Lynch (2010) indicated that inflammatory cytokines introduced directly and indirectly into the brain influence LTP in the perforant path to dentate gyrus synapses, decreasing LTP. Further, impairment in LTP was directly in proportion to the level of inflammatory cytokines and inversely related with noninflammatory cytokines. Lynch was able to make a strong case for the causal link between microglial activation and impaired LTP, but also for the ability of inhibition of microglial activation to partially rescue LTP. Understanding how microglial activation inhibits plastic change may facilitate the discovery one of the most beneficial treatments, a treatment for the cognitive decline and loss of flexible learning that occurs with aging.

Cognitive impairment is a feature in many genetic brain disorders. Fragile X Mental Retardation Syndrome (FXS) is the leading inherited cause of mental retardation. In this disorder, an extension occurs to a CGG island of the Fragile X Mental Retardation 1 (FMR1) gene. The expansion to over 200 repeats leads to hypermethylation of the gene and a reduction or elimination in FMRP through transcriptional silencing. FMR1 knockout mice show learning deficits and hyperactivity and appear to represent a good model of the human disorder (Dutch/Belgium Consortium, 1994). These mice also exhibit protein synthesis independent LTD mediated by a group of mGluRs, which may also account for the abnormally long and thin dendritic spines seen in both the mice and the human population with the disorder (reviewed in Irwin et al., 2000). LTD-like processes mediated by mGluRs may contribute to the learning and memory problems in FXS.

The Fragile X premutation is a trinucleotide expansion in the CGG island of 55–200 repeats on the FMR1 gene. It was widely thought that the premutation carriers were not cognitively impaired, but this may not be the case. Hunsaker et al. (2012) investigated an animal model of the premutation to the FMR1 gene that contained 70–198 repeats. Increasing CGG repeats modulated the induction of LTP, LTD, mGluR1/5 LTD, and behavioral tasks that required spatiotemporal processing. Although an increasing number of CGG repeats was associated with impaired plasticity, the plasticity once induced did not show any difference in expression. Additional research is necessary to discover the mechanism underlying the relationship between CGG repeat number and impairment, to more fully explain the expected outcomes on plasticity and cognitive function.

Conclusions

Since the discovery of LTP in the hippocampus, a substantial amount of progress has been made to understand bidirectional plasticity in the brain. Most of the data has focused on the hippocampus, but many studies extend the findings on plasticity to other areas of the brain suggesting that some of the mechanisms that have been found are pervasive throughout the brain. LTP and LTD are a result of the activation of postsynaptic receptors that mediate synaptic transmission. The signaling cascades often suggest a unitary system that has flexibility to result in both LTP and LTP depending on the induction stimulus, but not in always in a reverse manner. Novel pharmacological and genetics approaches have allowed for a greater understanding, but also do much to highlight the role of an LTP- and LTD-like process in both health and disease states, as well as the role stimulation may play in interventions.

See also: Electroencephalography: Basic Principles and Applications; Habituation; Kindling; Neural Plasticity; Neural Synchrony as a Binding Mechanism.

Bibliography

Abraham, W.C., Bear, M.F., 1996. Metaplasticity: the plasticity of synaptic plasticity. Trends in Neurosciences 19 (4), 126–130.

Alger, B.E., Teyler, T.J., 1976. Long-term and short-term plasticity in the CA1, CA3, and dentate regions of the rat hippocampal slice. Brain Research 110, 463–480.

Andersen, P., Homquvist, B., Voorhoeve, P.E., 1966. Entorhinal activation of dentate granule cells. Acta Physiologica Scandinavica 66, 448–460.

Andersen, P., Lomo, T., 1967. Control of hippocampal output by afferent volley frequency. Progress in Brain Research 27, 400–412.

Beck, H., Goussakov, I.V., Lie, A., Helmstaedter, C., Elger, C.E., 2000. Synaptic plasticity in the human dentate gyrus. The Journal of Neuroscience 20, 7080–7086.

Bliss, T.V.P., Cooke, S.F., 2011. Long-term potentiation and long-term depression: a clinical perspective. Clinics 66, 3–17.

Bliss, T.V., Gardner-Medwin, A.R., 1973. Long-lasting potentiation of synaptic transmission in the dentate area of the unanaestetized rabbit following stimulation of the perforant path. Journal of Physiology 232 (2), 357–374.

Bliss, T., Lomo, T., 1973. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. Journal of Physiology (London) 232, 331–356.

Christie, B.R., Abraham, W.C., 1992. Priming of associative long-term depression in the dentate gyrus by theta frequency synaptic activity. Neuron 9 (1), 79–84.

Christie, B.R., Stellwagen, D., Abraham, W.C., 1995. Reduction of the threshold for long-term potentiation by prior theta-frequency synaptic activity. Hippocampus 5 (1), 52–59.

Collingridge, G.L., Peineau, S., Howland, J.G., Wang, Y.T., 2010. Long-term depression in the CNS. Nature Reviews 11, 459–473.

The Dutch-Belgian Fragile X Consortium, 1994. Fmr1 knockout mice: a model to study fragile X mental retardation. Cell 78 (1), 23–33.

Dash, P.K., Karl, K.A., Colicos, M.A., Prywes, R., Kandel, E.R., 1991. cAMP response element-binding protein is activated by Ca2+/calmodulin- as well as cAMP-dependent protein kinase. Proceedings of the National Academy of Sciences of the United States of America 88 (11), 5061–5065.

Froc, D.J., Racine, R.J., 2004. N-methyl-D-aspartate receptor-independent long-term depression and depotentiation in the sensorimotor cortex of the freely moving rat. Neuroscience 129 (2), 273–281.

Granger, A.J., Shi, Y., Lu, W., Cerpas, M., Nicoll, R.A., 2013. LTP requires a reserve pool of glutamate receptors independent of subunit type. Nature 493 (7433), 495–500.

Hebb, D.O., 1949. The Organization of Behavior. Wiley, New York.

Hrabetova, S., Sacktor, T.C., 2001. Transient translocation of conventional protein kinase C isoforms and persistent downregulation of atypical protein kinase Mzeta in long-term depression. Brain Research. Molecular Brain Research 95 (1–2), 146–152.

Huang, Y.Y., Colino, A., Selig, D.K., Malenka, R.C., 1992. The influence of prior synaptic activity on the induction of long-term potentiation. Science 255 (5045), 730–733.
Hunsaker, M.R., Kim, K., Willemsen, R., Berman, R.F., 2012. CGG Trinucleotide repeat length modulates neuronal plasticity and spatiotemporal processing in a mouse model of the fragile X premutation. Hippocampus 22, 2260–2275.
Ivanco, T.L., Racine, R.J., 2000. Long-term potentiation in the reciprocal cortico-hippocampal and corticocortical pathways in the chronically implanted, freely moving rat. Hippocampus 10 (2), 143–152.
Ivanco, T.L., Racine, R.J., Kolb, B., 2000. Morphology of layer III pyramidal neurons is altered following induction of LTP in sensorimotor cortex of the freely moving rat. Synapse 37 (1), 16–22.
Irwin, S.A., Galvez, R., Greenough, W.T., 2000. Dendritic spine structural anomalies in fragile-X mental retardation syndrome. Cerebral Cortex 10 (10), 1038–1044.
Kato, N., 2009. Neurophysiological mechanisms of electroconvulsive therapy for depression. Neuroscience Research 64, 3–11.
Koffie, R.M., Hyman, B.T., Spires-Jones, T.L., 2011. Alzheimer's disease: synapses gone cold. Molecular Degeneration 6, 63–72.
Lee, H.-K., Bararosie, M., Kameyama, K., Bear, M.F., Huganir, R.L., 2000. Regulation of distinct AMPA receptor phosphorylation sites during bidirectional plasticity. Nature 405, 955–959.
Larson, S.J., Hartle, K., Ivanco, T.L., 2007. Acute administration of interleukin-1β disrupts motor learning. Behavioral Neuroscience 121 (6), 1415–1420.
Lee, K., Kim, J.-H., Kwon, O.-B., An, K., Ryu, J., Cho, K., Suh, Y.-H., Kim, H.-S., 2012. An activity-regulated microRNA, miR-188, controls dendritic plasticity and synaptic transmission by downregulating neuropilin-2. The Journal of Neuroscience 32, 5678–5687.
Lomo, T., 2003. The discovery of long-term potentiation. Philosophical Transactions of the Royal Society of London 358, 617–620.
Lynch, G., Larson, J., Kelso, S., Barrionuevo, G., Schottler, F., 1983. Intracellular injections of EGTA block induction of hippocampal long-term potentiation. Nature 305 (5936), 719–721.
Lynch, M., 2010. Age-related neuroinflammatory changes negatively impact on neuronal function. Frontiers in Aging Neuroscience 1 (6), 1–8.
Migues, P.V., Hardt, O., Wu, D.C., Gamache, K., Sacktor, T.C., Wang, Y.T., Nader, K., 2010. PKMzeta maintains memories by regulating GluR2-dependent AMPA receptor trafficking. Nature Neuroscience 13 (5), 630–634.
Mulkey, R.M., Malenka, R.C., 1992. Mechanisms underlying induction of homosynaptic long-term depression in area CA1 of the hippocampus. Neuron 9 (5), 967–975.
Oh, W.C., Hill, T.C., Zito, K., 2013. Synapse-specific and size-dependent mechanisms of spine structural plasticity accompanying synaptic weakening. Proceedings of the National Academy of Sciences of the United States of America 110 (4), E305–E312.
Racine, R.J., Milgram, N.W., Hafner, S., 1983. Long-term potentiation phenomena in the rat limbic forebrain. Brain Research 260 (2), 217–231.
Staubli, U., Lynch, G., April 9, 1990. Stable depression of potentiated synaptic responses in the hippocampus with 1–5 Hz stimulation. Brain Research 513 (1), 113–118.
Weeks, A.C.W., Ivanco, T.L., LeBoutillier, J.C., Racine, R.J., Petit, T.L., 1998. The degree of potentiation is associated with synaptic number during the maintenance of long-term potentiation in the dentate gyrus. Brain Research 798, 211–216.
Weeks, A.C.W., Ivanco, T.L., LeBoutillier, J.C., Racine, R.J., Petit, T.L., 1999. Sequential changes in the synaptic structural profile following long-term potentiation in the rat dentate gyrus: I. The intermediate maintenance phase. Synapse 31, 97–107.
Weeks, A.C.W., Ivanco, T.L., LeBoutillier, J.C., Racine, R.J., Petit, T.L., 2000. Sequential changes in the synaptic structural profile following long-term potentiation in the rat dentate gyrus: II: Induction/early maintenance phase. Synapse 36, 286–296.
Weeks, A.C.W., Ivanco, T.L., LeBoutillier, J.C., Racine, R.J., Petit, T.L., 2001. Sequential changes in the synaptic structural profile following long-term potentiation in the rat dentate gyrus: III: long-term maintenance phase. Synapse 40, 74–84.
Weiler, I.J., Spangler, C.C., Klintsova, A.Y., Grossman, A.W., Kim, S.H., Bertaina-Anglade, V., Khaliq, H., de Vries, F.E., Lambers, F.A., Hatia, F., Base, C.K., Greenough, W.T., 2004. Fragile X mental retardation protein is necessary for neurotransmitter-activated protein translation at synapses. Proceedings of the National Academy of Sciences of the United States of America 101 (50), 17504–17509.
Yoshii, A., Constantine-Paton, M., 2010. Postsynaptic BDNF-TrkB signaling in synapse maturation, plasticity, and disease. Developmental Biology 70, 304–322.

Relevant Websites

http://dnalc.org/view/1099-Long-term-Depression.html – DNA Learning Center – Long-Term Depression.
http://dnalc.org/view/549-Long-term-Potentiation.html – DNA Learning Center – Long-term Potentiation.
http://dnalc.org/view/1210-Long-term-Potentiation-and-Learning.html – DNA Learning Center – Long-term Potentiation and Learning.
http://www.nature.com/nrn/journal/v6/n9/fig_tab/nrn1749_F1.html – Nature Reviews Neuroscience – Fig. 1 What is late-phase long-term potentiation?.
http://www.ncbi.nlm.nih.gov/books/NBK10899/ – NCBI Resources Bookshelf – Long-Term Synaptic Depression.
http://www.ncbi.nlm.nih.gov/books/NBK10878/ – NCIB Resources Bookshelf – Neuroscience 2nd Ed Long-Term Synaptic Potentiation.
http://www.sagepub.com/garrettbb2study/animations/12.10.htm – Sage Publications Brain and Behavior Figure 12.10 Associative Long-Term Potentiation.

Lorenz, Konrad (1903–89)

Wolfgang M Schleidt, University of Vienna, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Konrad Lorenz, cofounder of ethology (the study of the biology of organism behavior), ranks among the most important stewards of Darwin's heritage. A stern critic of the vitalistic concept of instinct and of the behaviorists' doctrines of the 'empty organism' that is predominantly environmentally controlled, Lorenz advocated a bottom-up analysis of behavior, stressing the importance of observation and inductive reasoning. This led to the discovery of a basically modular structure of behavior including 'fixed action patterns' controlled by internal variables and external 'key stimuli' coordinated by specific 'innate releasing mechanisms.' Learning occurs on these modular levels, e.g., as in 'imprinting.' Lorenz became one of the founding directors of the Max Planck Institute for Behavioral Physiology (Seewiesen), the leading institution for behavioral research in Germany. Lorenz, Niko Tinbergen, and Karl von Frisch shared the 1973 Nobel Prize for Medicine 'for their discoveries concerning the organization and elicitation of individual and social behavior patterns.' After his retirement, Lorenz returned to his native Austria, continued his work on graylag goose's social behavior and in evolutionary epistemology, and became an influential opponent of nuclear power and environmental degradation.

Biography

Family

Konrad Zacharias Lorenz (born 7 November 1903 in Vienna, Austria) grew up within the upper crust of pre-World War I Vienna: his mother Emma (née Lecher) from a family of writers and scientists; his father, Adolf Lorenz, world-renowned orthopedic surgeon, Nobel Prize nominee (Lorenz, 1936). Konrad spent his formative years in Vienna and at his family's country estate in Altenberg on the banks of the Danube. At that time, the river was still lined by large tracts of virgin riverine forest, and his inquisitive mind became fascinated by nature. He became a passionate observer and skilled caretaker of any swimming, crawling, or flying creature. These interests did not interfere, however, with his early socialization into the network of learned relatives and their friends (e.g., Karl Popper).

Education

After visiting a 'progressive' elementary school, he was sent to the esteemed *Schottengymnasium*, where his biology teacher, an enthusiastic Darwinist, introduced him to evolutionary reasoning. At his father's suggestion, he began to study medicine at Columbia University, New York, in 1922. Homesick, especially for his wife-to-be, he continued his studies in Vienna, receiving his Dr med. degree in 1928. During his first appointment as *Demonstrator* (eqivalent to assistant professor) at the Department of Anatomy, University of Vienna, he studied zoology (J. Versluys), paleontology (O. Abel), and psychology (K. Bühler, his assistant E. Brunswik); he received his Dr phil. degree in 1933.

Scientific Career

Lorenz's meager salary as *Demonstrator* and his wife's (Margarethe née Gebhardt) income as physician at a local hospital provided just enough to raise their family. Lorenz was always short of funds for the private research facility he had established in Altenberg. His negotiations with the (German) Kaiser-Wilhelm-Gesellschaft for funds to create a special institute for behavioral studies date back to 1937, but at the outbreak of World War II this project was shelved. In 1940, E. Baumgarten (a nephew of Max Weber and a pupil of J. Dewey and, hence, a representative of the pragmatist school of philosophy) accepted the famed chair of philosophy at Königsberg University, once occupied by Immanuel Kant, and looked for a biologically oriented psychologist who, at the same time, was interested in theory of knowledge to share with him the prestigious appointment. O. Koehler and E. von Holst became the matchmakers, and Lorenz was chosen (Lorenz, 1974; Leyhausen, 2000; Burkhardt, 2005).

One year later, however, Lorenz was drafted into the German Army as a motorcycle rifleman (in his papers he had not mentioned his medical training because of his limited experience with patients, and stressed only his proficiency in motorcycle racing and repair). When the authorities discovered that he held a medical degree and was a professor of psychology, he was assigned to a unit of the *Heerespsychologie* (Army Psychology) in Posen. Two months later, however, this unit was dissolved, and he was transferred to the Posen Army field hospital for neurology and psychiatry, where he worked as *Unterarzt* (medical noncommissioned officer (NCO)) in a unit of 285 beds with patients suffering from war injuries, psychoses, and psychosomatic disorders. During a field assignment on the Russian front in 1944, he was captured by the Red Army. After that, he served in various POW camps as field surgeon until his release in 1948.

In post-World War II Austria, career prospects for a biologist were worse than ever, while in Britain Lorenz's colleagues became especially interested in the Darwinian approach to behavioral studies. Niko Tinbergen became reader at Oxford University and, in 1950, Lorenz received a similar offer from the University of Bristol. German colleagues, notably E. von Holst and G. Kramer, both meanwhile directors at the newly founded Max Planck Institute for Marine Biology (Wilhelmshaven), were able to revive the old project of an institute for behavioral studies.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61078-4

A matching offer was made, and Lorenz accepted the appointment as head of a small behavior research unit in Buldern, Westfalia, founded in spring of 1951 as precursor to a special Max Planck Institute for Behavioral Physiology. Construction started in Seewiesen, Bavaria in 1955, and the institute opened in 1958; Lorenz was Director of his own division there until he retired in 1973.

Retirement

Upon retirement, Lorenz returned to his home in Altenberg, Austria. Shortly thereafter, he was awarded, together with Niko Tinbergen and Karl von Frisch, the 1973 *Nobel Prize for Medicine*. During retirement, he tidied up his work on evolutionary epistemology, on the social behavior of graylag goose and reef fishes, and became an eloquent spokesman for environmental conservation (Schleidt, 1988). Lorenz died in Vienna on 27 February 1989 at the age of 85 and was buried near Altenberg at St. Andrä-Wördern cemetery (Wuketits, 1990).

Contribution to Knowledge

Lorenz, cofounding with Niko Tinberger the science of ethology as 'the study of the biology of organism behavior,' ranks among the most important stewards of Darwin's heritage of the twentieth century. In the 1930s and 1940s, most psychologists, sociologists, and biologists tried to understand behavior within the framework of Aristotelian philosophy, theology, Marxism, or Behaviorism. Lorenz, on the other hand, viewed behavior as playing a central role in the process of evolution, and shaped primarily by 'natural selection,' rather than by learning. By applying the methods of comparative morphology and paleontology to the study of 'species specific' behavior, he created a scientific approach that rapidly provided a wealth of data and new insights (Lorenz, 1931, 1932, 1950b, 1965b, 1978). Contrary to the claim of his critics, Lorenz ignored neither the conventional schools of thought nor their observations, but was cautious in accepting their theories and conclusions. For example, he never denied the importance of environmental factors in learning, but he challenged the behaviorists' claim that innate behavior was restricted to only a few basic vital reflexes. In fact, he emphasized especially the interlocking of modular innate and learned behavioral (e.g., as in 'imprinting,' Lorenz, 1935; of the two alternative channels through which information can enter a living system, Lorenz, 1961, 1965b).

Observation as Research Tool

Fundamental to Lorenz's approach to behavior is his rigorous adherence to the doctrine of *induktive Naturwissenschaft* (inductive natural science; Lorenz, 1942, 1959, 1978): first establish a solid base with meticulous observations, supported by objective records (photos, film), tainted as little as possible by theories, and – only after a solid data base (*Induktions-Basis*) has been established – let patterns emerge from the data. He became a pioneer of scientific cinematography (Lorenz, 1950a, 1952; Schleidt and Oeser, 2011) and was founding member of Encyclopedia Cinematographica (Wolf, 1967).

Behavioral Modules

Lorenz considered behavior to be composed of basic modules of classical reflexes, species specific *Instinktbewegung* 'fixed action pattern' (FAP) (in later years *Erbkoordination*, *erbkoordinierte Bewegung* (1978) (in English: 'fixed motor pattern')), specific circuits for the environmental control of reflexes and FAPs, which were 'innate releasing mechanisms' or IRM (*angeborener Auslösemechanismus*), 'appetites and aversions' (*Stimmungen*), etc. Such modules provide each organism with basic behavioral skills to navigate within its environment, detect necessary resources, avoid danger, and interact with conspecifics. Learning can enter in various ways at the level of each of these modules, e.g., in the case of a reflex as 'conditioned reflex.'

This FAP/IRM dyad is basically an extension of the classical stimulus–response concept, enhanced by several new features because the classical reflex, activated only by a specific external stimulus, cannot account for the observation that "a healthy animal is up and doing" (as W. McDougall so aptly remarked). Consequently, the spontaneous nature of behavior, in general, and especially in many typical FAPs (locomotion, courtship behavior, certain types of bird song, etc.) became a topic of special interest (FActionP). FAPs constitute much more complex patterns than any of the classical reflexes (FAPatterns), and particular spatiotemporal parameters of FAPs are highly stereotyped within any one species (FixedAP), but different between species. Lorenz exemplified the complex interaction between internal and external variables in his 'psychohydraulic' model (1950b, 1978), and applied this reasoning to explain the control of aggression in animals and humans (1963).

It was basically for this work on behavior modules, which are central to ethology, that Lorenz, Niko Tinbergen, and Karl von Frisch shared the distinction of winning the 1973 Nobel Prize for Medicine "for their discoveries concerning the organization and elicitation of individual and social behavior patterns."

Ethograms

Given the existence of species specific behavior, a behavioral study of individuals of any given species starts with an assessment of the modules of its behavioral repertoire, working toward the open-ended task of a complete 'ethogram' of that species. Lorenz had already touched on this in his first publication (1931), and a decade later published an exemplary comparative study of the courtship modules of ducks (1941b). He delayed publishing of his ethogram of the graylag goose until 1988 … as a sort of swan song.

The Innate Forms of Possible Experience

This heading is a translation of *Die angeborenen Formen möglicher Erfahrung*, the title of (Lorenz's, 1943) seminal paper dealing with the evolution of cognition, as a new paradigm. The results of many years of research into 'learning' – from the classical curves of learning and forgetting by I. Pavlov and E.L. Thorndike – indicate that it is a slow process. If an animal's survival depended solely on learning, it would be dead long before it had completed its first set of learning trials. Thus, it

should not be surprising that an animal, entering into its world, is equipped with: (1) certain 'innate' skills (behavior patterns to satisfy vital needs and escape danger), and (2) 'a priori knowledge' to detect essential features of resources and danger – innate in the sense that particular patterns of behavior or relevance of particular environmental features are available prior to individual experience. In other words a reflex, FAP, or IRM not only enables survival enhancing action, but it also constitutes organismic 'knowledge' about the environment and its contingencies to which that individual's ancestors had been exposed in the evolutionary process (Lorenz, 1941a, 1943, 1953, 1965b, 1973, 1978).

Imprinting

This peculiar process of forming an irreversible association between a behavior pattern and a 'stimulus' (object), without any reward other than the presence of that object, has been used for ages in animal breeding, and noted in the literature, but Lorenz was the first to recognize imprinting's far-reaching theoretical significance (Lorenz, 1935, 1965b, 1978).

Nature–Nurture

Much ink has been spilled on this duality, not much because of any intellectual challenge posed by these two concepts, but because of deeply rooted beliefs, religious or political, touching upon free will, accountability, and fate. Lorenz was attacked early on for his use of 'innate' as though he had excluded the possibility of learning, while all he suggested was that innate behavior was essential for the survival of individual organisms. Early on, he pointed to the intertwining of innate and learned behaviors (e.g., *Trieb-Dressur-Verschränkung*, drive-training-interlacing 1935) and, with imprinting, provided a splendid example of such a process; an example that was ignored by the ruling psychologists of his time. He was attacked relentlessly for defending the concept of innateness, despite numerous attempts to respond to his critics (1961, 1965b, 1978, etc.).

Epistemology

"Philosophy is the academic abuse of those human potentials which have evolved by natural selection for the individual's survival in his struggle with the daily contingencies of his world." This definition, attributed to Lorenz's mentor O. Heinroth, encapsulates the insight that human ability to acquire knowledge is not merely a matter of the reasoning individual *Homo sapiens* within a literate society, but is based on a long, evolutionary process, preceding the emergence of humanity by millions of years. Lorenz, having been elevated to Immanuel Kant's former chair in Königsberg in 1940, and facing the challenges by contemporary philosophy, and specifically by resident neo-Kantians, took a radically new approach to Kant's concept of a priori (Lorenz, 1941a, 1943), pioneering what is now called Evolutionary Epistemology (Popper, 1972; Vollmer, 1980). The most detailed discussion of this is contained in his posthumously published *Russian Manuscript* 1944–48 (1992), written with the fresh memories of his encounter with Kant's Königsberg in mind, while a more concise discussion can be found in the English edition of his textbook (1978/81).

Moral-Analogous Behavior, Human Morality, and Environmental Awareness

Lorenz, throughout his life, struggled with his belief that morality could well have an evolutionary basis (as he had shown convincingly for the evolutionary basis of cognition in his evolutionary epistemology). He coined the term 'moral-analogous behavior' (moral-analoges Verhalten): special behavior that channels the individual's selfish struggle for its individual existence and reproductive success into an altruistic contribution for the good of its community. Based on his observations of a variety of highly social animals, Lorenz noticed that domesticated forms had lost certain elements of their moral-analogous social behaviors, which their wild ancestors had evolved as essential part of their sociality. Lorenz concluded, by analogy, that human evolution in general, and especially our own modern civilization, has led to comparable domestication effects ('self-domestication'), which have become mixed blessings. Modern humans have freed themselves from the firm guide of instincts and consequently have more freedom to learn and to make their own choices, but they also have lost some of their instinctive inhibitions to inflict damage to, and even kill conspecifics to an extent not known in any other species (1940, 1943, 1963, 1983).

In his early papers, Lorenz proposed to counter these dangers to humanity with eugenic measures (before the atrocities of Hitler's Germany, eugenics was a widely respected field, especially in England and the United States), especially by encouraging ethically guided (not ethnically, as often misread) mate choice (1940, 1943). Starting in the mid-1950s, Lorenz's proposal of an evolutionary basis of human morals were attacked for various reasons or dismissed as unfounded. As competing hypotheses – especially those claiming sole environmental control of ethical behavior – have failed so far to resist falsification, these issues are, on the scientific level, far from resolved.

Lorenz's concern about human evolution running awry and leading us to extinction by our destruction of the environment was not merely a consequence of the atrocities of World War I, World War II, and of the nuclear demise scenarios of the Cold War, but had been deeply rooted in both his trust in science and in his allegiance to the Hippocratic oath. Thus, when by the vagaries of political and economic interests a nuclear power plant was built near his home in Altenberg, at a location that had been flawed by previously undisclosed risks, he opposed completion of its construction violently. Soon he found himself the revered leader in this fight that ultimately prevented the key-ready nuclear power plant from going into service. A few years later, Lorenz was equally successful in a fight to save the last stretch of wetlands along the Danube.

Relevance

Lorenz's impact on behavioral and social science is linked closely to his collaboration and friendship with Niko Tinbergen, and it is difficult to measure each one's individual contribution. Both were curious naturalists endowed with outstanding skills

in observing, in detecting regularities – patterns of nature – and in formulating testable hypotheses; and, both were outstanding teachers (e.g., Lorenz, 1949, 1950c). While Tinbergen was more inclined toward an experimental approach, as exemplified in their joint paper (Lorenz and Tinbergen, 1938), Lorenz's imprint on ethology is obvious in matters of pure observation, inductive reasoning, and formulating a comprehensive theory of the modular concept of action patterns and releasing mechanisms, of these modules being shaped by evolution as well as by learning, and of imprinting as a special type of learning.

A first attempt to estimate Lorenz's impact on his field from citation frequencies (Leydesdorff, 2001) indicates an increase of citation frequencies well above the current average.

Selected Topics

Lorenz achieved his scientific breakthrough in 1936. At an invited lecture before the German scientific elite of the Kaiser-Wilhelm-Gesellschaft in Berlin, he proposed to stop arguing about how many drives and instincts one can define, and instead start describing and analyzing behavior patterns as though they were anatomical features. In a discussion following the lecture, Max Hartmann, doyen of German biology, remarked that "Konrad Lorenz's approach opens up to causal analysis a field that was previously just a playground for fruitless philosophical speculation" (Lorenz, 1992).

World War II severely restricted research and exchange. However, soon after the war had ended, Tinbergen traveled in the UK and the United States presenting the new, ethological approach to behavioral studies, his effort culminating in his book *The Study on Instinct* (1951). Lorenz, who, up to then, had established an international reputation only among ornithologists, argued the case for a biological approach to the study of behavior and soon became the leading theoretician in the new science of ethology.

The high esteem for ethological research resulted in generous support by the Max-Planck-Gesellschaft, and subsequently by various national funding agencies, notably National Institutes of Health in the United States. Small informal meetings soon turned into international affairs. The biennial *International Ethological Conferences*, dominated by the Lorenz–Tinbergen School and 'converts' from a variety of related fields, became the central event for discussing new results and theories. The German journal *Zeitschrift für Tierpsychologie*, founded in 1937 with Lorenz as one of its editors, the first journal to favor ethological publications, soon got an English language companion in the Dutch journal *Behaviour*. As publications in the new science steadily increased *Animal Behaviour* was introduced a few years later as the official organ of Animal Behavio(u)r Societies that were founded in the United States and UK. While *animal psychology* and *comparative psychology* had been little-known specialties, dealing mainly with the learning abilities of rats, the Lorenz–Tinbergen-inspired new *animal behavior* became a hot topic for lectures in academic departments both in biology and psychology, resulting in a demand for teachers and textbooks (e.g., Eibl-Eibesfeldt, 1997, 1999), and spawning new opportunities and funding for behavioral research.

Since Lorenz, Tinbergen, and their colleagues were not only qualified biologists, but are also equally well established in neighboring fields, the popularity of ethology soon led to fruitful cooperation, notably with sensory physiology, brain sciences, and psychiatry (e.g., Lorenz, 1953). This process had considerable impact on the emergence of neurobiology, psychobiology, and evolutionary psychiatry, emphasizing a naturalistic, Darwinian approach at a time when the general trend in biology was toward molecular biology and biochemistry.

Subjects of ethological research at first were mainly vertebrates – birds, mammals, and fishes – and insects. The new approach was tested successfully and applied to all kinds of species, including our own. Even though the magnitude and difficulties of a task like establishing a human ethogram may appear mind-boggling, 'human ethology' becomes workable as a special case of primate ethology (e.g., Eibl-Eibesfeldt, 1997). A healthy human newborn faces similar tasks as any other primate baby and behaves very similarly.

By applying ethological methods within the framework of natural science and evolutionary reasoning, similarities and differences between human and nonhuman animals are no longer a matter of poetic license, but causal analysis – not, however, without drawing heavy critique from the old religious and political establishment. *On Aggression* and *The Waning of Humaneness* (Lorenz, 1963, 1983) are examples of an ethological approach to the moral strongholds of the humanities.

Human ethology has become a domain of its own, of broad overlap not only with primatology, but also anthropology (physical and cultural), psychology, and sociology, leading toward a new synthesis of knowledge, which proves that the old distinction between hard sciences and humanities has become obsolete.

Politics

Fame has its price. Especially since having been awarded the Nobel Prize, Lorenz has been branded a racist and an anti-Semite and alleged to have provided a pseudoscientific justification for the Holocaust (e.g., Klopfer, 1999).

In regard to these allegations, we must consider first that Lorenz grew up within a part of the Viennese society that was enmeshed intimately by friendship, marriage, and professional ties with the wealthy Jewish elite. He maintained these connections throughout his life by hosting many guests and being a guest on many in his frequent travels of old and new friends especially in England and the United States, and was certainly more pro-Semitic than average, and far less racist, especially than some of those who are now raising their voice. Thus, the charge that he advocated racist and anti-Semitic genocide is unfounded.

With respect to the allegation that Lorenz provided a pseudoscientific justification for the Holocaust, it must be considered that there is a long chain of inferences necessary to connect his controversial writings on the potentially dangerous effects of human self-domestication on the biological basis of ethical (not, and never ethnical!) behavior, or Lorenz's proposed use of eugenics to combat the rise of unethical behavior, to the atrocities of the Third Reich. And, there is no indication – in spite of many inquiries and searches (summarized in Klopfer, 1999) that prior to the end of World War II anybody in power was aware of such a link. Lorenz being called to the army early in World War

II, his first assignment as a rifleman and later as medical NCO (never promoted to a higher rank) gives ample evidence that he lacked political pull (Schleidt, 2000, 2003).

Lorenz became active in politics only after his retirement. As a Nobel Laureate, he was 'welcomed home' by the Austrian socialist elite. Given all the academic and public honors he had received, people listened when he spoke – as he proudly proclaimed – and he spoke loudly and clearly for environmental preservation and moral responsibility. He became the figure head of the growing 'Green' movement, both in a referendum against the use of nuclear power (5 November 1978) and in the "Konrad-Lorenz-Volksbegehren" (March 4–11, 1985) against the hydroelectric power plant in Hainburg that threatening the destruction of the surrounding wetland (Schleidt, 2000, 2003).

Concluding Remarks

Konrad Lorenz had been interested in a wide variety of animals, ranging from ciliates to primates, and among the fishes his special interest had been in Cichlids and the native cyprinids of Austria, however, most of his research had been focused on birds, notably corvids and waterfowl.

He established a breeding colony of Jackdaw in the attic of his home in Altenberg, and built there a pond for his ducks and geese. When in 1955, the dream of his life came true and he was allowed to choose – together with Erich von Holst – the site for their 'Max Planck Institute for Behavioral Physiology,' the banks of a little Bavarian lake, the 'Esssee' were picked for the location, as the most suitable place to study waterfowl in their natural environment.

When in the mid-1970s sociobiology made the headlines and behavioral ecology started riding the new wave of environmental awareness, the glamour of ethology began to fade. In 1999, the crown jewel of ethology, the Max Planck Institute at Seewiesen, was threatened to be closed, but offered a new lease on life as an ornithological research station. In 2004, this research unit was renamed and rebuilt as Max Planck Institute for Ornithology.

The ideas conceived and publicized by Lorenz, Tinbergen, and their contemporaries as the foundations of ethology have become public property. This may be the best indicator of the importance of the work of Konrad Lorenz, and of ethology.

See also: Bowlby, John (1907–90); Campbell, Donald Thomas (1916–96); Darwin, Charles Robert (1809–82); Erikson, Erik Homburger (1902–94); Ethics and Values; Evolution, Natural and Social: Philosophical Aspects; Evolutionary Approaches to Human Behavior: Philosophical Aspects; Freud, Sigmund (1856–1939); Kant, Immanuel (1724–1804); Kantian Ethics and Politics; Knowledge (Explicit, Implicit and Tacit): Philosophical Aspects; Pavlov, Ivan Petrovich (1849–1936); Popper, Karl Raimund (1902–94); Psychohistory; Science, History of; Skinner, Burrhus Frederick (1904–90); Watson, John Broadus (1878–1958).

Bibliography

Burkhardt Jr., R.W., 2005. Patterns of Behavior: Konrad Lorenz, Niko Tinbergen, and the Founding of Ethology. The University of Chicago Press, Chicago.

Eibl-Eibesfeldt, I., 1999. Grundriss der vergleichenden Verhaltensforschung—Ethologie, eighth ed. Piper, Munich/Zürich (1970 Ethology—The Biology of Behavior. Holt, Rinehart & Winston, New York).

Eibl-Eibesfeldt, I., 1997. Die Biologie des menschlichen Verhaltens, fourth ed. Piper, Munich/Zürich (1989 Human Ethology. Aldine de Gruyter, New York).

Klopfer, P.H., 1999. Konrad Lorenz and the national socialists: on the politics of ethology. International Journal of Comparative Psychology 7, 202–208.

Leydesdorff, L., 2001. Scientometrics Smelser, Neil J., Baltes, Paul B. (Eds.), 2001. International Encyclopedia of the Social and Behavioral Sciences, vol. 20. Elsevier, Oxford, UK, pp. 13752–13755.

Leyhausen, P., 2000. Das Königsberger Jahr [The year in Königsberg]. In: Kotrschal, K., Müller, G., Winkler, H. (Eds.), Konzepte der Verhaltensforschung—Konrad Lorenz und seine Folgen. Filander Verlag, Fürth.

Lorenz, K., 1931. Beiträge zur Ethologie sozialer Corviden. Journal für. Ornithologie 79 (1), 67–127.

Lorenz, K., 1932. Betrachtungen über das Erkennen der arteigenen Triebhandlungen der Vögel. Journal für Ornithologie 80 (1), 50–98.

Lorenz, A., 1936. My Life and Work. Scribner, New York.

Lorenz, K., 1935. Der Kumpan in der Umwelt des Vogels. Journal für Ornithologie 83 (2–3) 137–215 und 289–413.

Lorenz, K., 1940. Durch Domestikation verursachte Störungen arteigenen Verhaltens. Zeitschrift für angewandte Psychologie und Charakterkunde 59, 2–81.

Lorenz, K., 1941a. Kants Lehre vom apriorischen im Lichte gegenwärtiger Biologie. Blätter für Deutsche Philosophie 15, 94–125 (1962 Kant's Doctrine of the A Priori in the Light of Contemporary Biology. General Systems 7, 23–35).

Lorenz, K., 1941b. Vergleichende Bewegungsstudien an Anatiden. Journal für Ornithologie 89 (3), 194–293. Festschrift O. Heinroth.

Lorenz, K., 1942. Induktive und teleologische Psychologie. Die Naturwissenschaften 30 (9-10), 133–143.

Lorenz, K., 1943. Die angeborenen Formen möglicher Erfahrung. Zeitschrift für Tierpsychologie 5, 235–409.

Lorenz, K., 1949. Er redete mit dem Vieh, den Vögeln und den Fischen. Borotha Schoeler, Vienna (1952 King Solomon's Ring. Methuen, London).

Lorenz, K., 1950a. Ethologie der Graugans. Beiheft zum gleichnamigen Film, Hochschulfilm C 560. Institut für Film und Bild in Wissenschaft und Unterricht, Göttingen.

Lorenz, K., 1950b. The comparative method in studying innate behavior patterns. Symposia of the Society for Experimental Biology 4, 221–268.

Lorenz, K., 1950c. So kam der Mensch auf den Hund. G. Borotha-Schoeler, Vienna (1954 Man Meets Dog. Methuen, London).

Lorenz, K., 1952. Balz und Paarbildung bei der Stockente (Anas platyrhynchos L.). Begleitheft zum gleichnamigen Hochschulfilm C 626. Institut für Film und Bild in Wissenschaft und Unterricht, Göttingen.

Lorenz, K., 1953. Über angeborene Instinktformeln beim Menschen. Deutsche medizinische Wochenschrift 78 (45–46), 1566–1569, 1600–1604.

Lorenz, K., 1959. Gestaltwahrnehmung als Quelle wissenschaftlicher Erkenntnis. Zeitschrift für experimentelle und angewandte Psychologie 6 (1), 118–165.

Lorenz, K., 1961. Phylogenetische Anpassung und adaptive Modifikation des Verhaltens. Zeitschrift für Tierpsychologie 18 (2), 139–187.

Lorenz, K., 1963. Das sogenannte Böse. Zur Naturgeschichte der Aggression. G. Borotha-Schoeler, Vienna (1966 On Aggression. Harcourt, Brace & World, New York).

Lorenz, K., 1965a. Über tierisches und menschliches Verhalten. Aus dem Werdegang der Verhaltenslehre. Gesammelte Abhandlungen. Bd. I (1931, 1932, 1935, 1937, 1938, 1942), Bd. II (1941b, 1950c, 1954, 1959, 1961, 1963b). Piper, Munich (R. Martin, Transl.). (1970–71 Studies in Human and Animal Behavior. In Vol. II, the 1961 paper has been omitted, since an expanded translation by the author was published the same year as Lorenz 1965b. Harvard University Press, Cambridge, MA).

Lorenz, K., 1965b. Evolution and Modification of Behavior. Chicago University Press, Chicago.

Lorenz, K., 1973. Die Rückseite des Spiegels. R. Piper, Munich (1977 Behind the Mirror. Harcourt, Brace & World, New York).

Lorenz, K., 1974. Konrad Lorenz (Autobiographical Sketch). In: Frängsmyr, T. (Ed.), Les Prix Nobel en 1973. The Nobel Foundation, Stockholm, Sweden.

Lorenz, K., 1978. Vergleichende Verhaltensforschung: Grundlagen der Ethologie. Springer Verlag, Vienna/New York (1981 The Foundations of Ethology. Springer Verlag, New York).

Lorenz, K., 1983. Der Abbau des Menschlichen. R. Piper, Munich (1987 The Waning of Humaneness. Little, Brown, Boston).

Lorenz, K., 1988. Hier bin ich-wo bist Du ? Ethologie der Graugans. Piper, Munich/Zürich (1991 Here Am I—Where Are You? The Behavior of the Greylag Goose. Harcourt Brace Jovanovich, New York).

Lorenz, K., 1992. Die Naturwissenschaft vom Menschen. Das 'Russische Manuskript.' Aus dem Nachlaß herausgegeben von Agnes von Cranach. Piper, Munich/Zürich (1996 The Natural Science of the Human Species. MIT Press, Cambridge, MA/London).

Lorenz, K., 2003. Eigentlich Wollte Ich Wildgans Werden. Aus Meinem Leben. Piper, München, Zürich.

Lorenz, K., Tinbergen, N., 1938. Taxis und Instinkthandlung in der Eirollbewegung der Graugans I. Zeitschrift für Tierpsychologie 2, 1–29.

Popper, K.R., 1972. Objective Knowledge: An Evolutionary Approach. Clarendon Press, Oxford, UK.

Schleidt, W.M. (Ed.), 1988. Der Kreis um Konrad Lorenz—Ideen, Hypothesen, Ansichten. P Parey, Berlin.

Schleidt, W.M., 2000. Politik gegen und mit Konrad Lorenz. In: Kotrschal, K., Müller, G., Winkler, H. (Eds.), Konzepte der Verhaltensforschung—Konrad Lorenz und seine Folgen. Filander Verlag, Fürth.

Schleidt, W., 2003. Epilog: Wer war der "Vater der Graugänse" wirklich? In: Lorenz, K. (Ed.), Eigentlich wollte ich Wildgans werden. Aus meinem Leben. Piper, München, Zürich, pp. 101–122.

Schleidt, W., Shalter, M.D., Moura-Neto, H., 2011. The hawk/goose story: the classical ethological experiments of Lorenz and Tinbergen, revisited. Journal of Comparative Psychology 125 (2), 121–133.

Schleidt, W.M., Oeser, E., 2011. Konrad Lorenz's use of cinematic film for studying dabbling ducks' courtship behaviour and the availability of historic film materials. Wildfowl 61, 45–51.

Tinbergen, N., 1951. The Study of Instinct. Oxford University Press, Oxford, UK.

Vollmer, G., 1980. Evolutionäre Erkenntnistheorie. Hirzel, Stuttgart.

Wolf, G., 1967. Der wissenschaftliche Dokumentationsfilm und die Encyclopaedia Cinematographica. Barth, München.

Wuketits, F.M., 1990. Konrad Lorenz—Leben und Werk eines großen Naturforschers. Piper, Munich.

Lotka, Alfred James (1880–1949)

Nathan Keyfitz[†], Harvard University, Cambridge, MA, USA

© 2001 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 13, pp. 9089–9091, © 2001, Elsevier Ltd.

Abstract

When A.J. Lotka died in 1949 his obituary in the *Journal of the American Statistical Association* began 'A brilliant chapter in modern demography came to a close with the death of Alfred James Lotka.' David Smith calls him 'the father of mathematical demography.' And these were understatements of his contribution, for demography was only a part of Lotka's scientific interests – he was also described as chemist, physicist, biologist, and actuary. What is common in all that he wrote was a search for better understanding through the use of mathematics. His name often appears in the literature as the author of the renewal equation – a mathematical expression of the mechanism by which the generations succeed one another – and as the codiscoverer of the Lotka–Volterra equations for predation and other interactions of multiple species in population biology. Of Lotka's many publications the two best known are the *Théorie analytique des associations biologiques* (1939) (translated by David Smith as *Analytical Theory of Biological Populations*) that provides the basis of modern mathematical demography, and *Elements of Physical Biology* (1925) that looks at the whole of the living world from a physico-chemical perspective.

When A.J. Lotka died on 5 December 1949 his obituary in the *Journal of the American Statistical Association* began 'A brilliant chapter in modern demography came to a close with the death of Alfred James Lotka.' David Smith, translator of his major work in the field (Lotka, 1939), calls him 'the father of mathematical demography.' And these were understatements of his contribution, for demography was only a part of Lotka's scientific interests – he was also described as chemist, physicist, biologist, and actuary, everywhere seeking better understanding through the use of mathematics.

Lotka was born on 2 March 1880 in Lemberg (Lvov), a Polish town of the Austro-Hungarian Empire, of American parents who were traveling Moravian missionaries; he died in Red Bank, New Jersey. His training was international: a boyhood mostly in France; a degree at the University of Birmingham (BSc 1901, DSc some 10 years later); graduate work in physics at Leipzig. Coming to the USA in 1902, he worked for the General Chemical Company, briefly for the US Patent Office as a patent examiner, and for the National Bureau of Standards as a physicist. He was editor of the *Scientific American Supplement* (1911–14), finally settling in the Statistical Bureau of the Metropolitan Life Insurance Company (1924–47). He was for long a bachelor, somewhat shy, described as quiet, learned, modest, and gently humorous. In 1935 he married Romola Beattie; they had no children. He was elected President of the Population Association of America (1938–39) and of the American Statistical Association (1942).

In his terminal year of graduate school, at Cornell in 1908–9, he began work with Professor F.R. Sharpe that led to the famous renewal equation – a mathematical description of the replacement of the generations:

$$B(t) = \int_{\alpha}^{\beta} B(t-a)p(a)m(a)da \qquad [1]$$

In the simplest case, rates of birth and death are constant over time, there is no migration to or from the defined area and the female sex only is considered. These conditions make possible a strict accounting of the relation between two generations. For on the right side is $B(t–a)$ the number of births a years earlier ($\alpha<a<\beta$ where α is the youngest age of childbearing and β the oldest); $p(a)$, the probability that a child born a years ago survives to time t; $m(a)$ is the birth rate for women aged a; integrating gives current births and so should be comparable to the births $B(t)$ on the left side. The unknown quantity that should make the two sides equal is the trajectory of births, to be provided in terms of the known functions $p(a)$ and $m(a)$. If the trajectory of $B(t)$ in Eqn [1] is simplified down to a curve of geometric increase, as Lotka initially assumed, then the question posed by the renewal equation is what rate of increase of births would equate the two sides of Eqn [1]. If that rate is r, taken either arbitrarily or from past data, then the age distribution of the population would stabilize over time to numbers proportional to $\exp(-ra)p(a)$ at the several ages a.

In finite intervals of time and age, Eqn [1] can be expressed naturally as a recurrent series and this was done by Lotka himself late in his career. P.H. Leslie formulated it in matrix terms. To these can be added a fourth version in partial differential equations (Keyfitz and Keyfitz, 1997), that also shows and compares the three other formulations.

Like most other discoveries, this one had predecessors who in part anticipated, as well as followers who marginally improved on, Lotka's work. Leonhard Euler (1760) came close on the theory; Edwin Cannan (1895) in the nineteeth century published an arithmetic calculation projecting the population of England and Wales that forecast the (then surprising) fall in births. R. Kuczynski (1931) and Richard Böckh of the Berlin Statistical Office were important predecessors working with the ratio of change per generation, the Net Reproduction Rate (NRR), $R_0 = \int p(a)m(a)da$. Willy Feller (1941), Princeton mathematician, noted that the solution to the renewal equation fell out simply and rigorously if each of the functions involved is translated into its unique Laplace Transform. Golubitsky et al. (1975) published an even more compact and more general solution.

Lotka went on to other demographic questions. He was curious to know why old populations have so few surnames and

[†]Deceased.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61079-6

found the answer in a pure probability accounting for the extinction of family lines (1931a). He ascertained the frequency of orphanhood as a function of birth and death rates (1931b) and calculated the money loss to dependents and to society when a person dies prematurely (Lotka and Dublin, 1946). In an early paper he went beyond human populations on the mode of growth of self-renewing aggregates; then on industrial replacement – for instance in a fleet of trucks of varied ages – he showed how the mathematics fits trucks as well as people.

Lotka's *Elements of Physical Biology* (1925), composed between 1922 and 1924 under a fellowship arranged by Raymond Pearl at Johns Hopkins University, is sharply distinguished from his earlier work – it looks at the whole of the living world from a physicochemical perspective. A first chapter on definition rejects a concept of life that sets it up behind an impassable boundary, such that living matter can never emerge through physical and chemical mechanisms. But it must have so emerged at least once in the history of the planet, and that emergence was of interest to Lotka.

Lotka anticipated modern ecology in his continual search for cyclical processes. One of many examples is the circulation of sulfur, that may be summarized from the *Elements of Physical Biology* as follows.

Sulfur occurs in abundance in inorganic nature in the form of the sulphates of the alkalies and the alkaline earths. Plants assimilate these compounds directly and then form proteins containing from 0.3 to 2% of sulfur. In this form it is assimilated by animals, who excrete the element chiefly in the form of sulfates. These, returned to the soil, complete the cycle.

And in the same way, Lotka portrays the cycles of water, iron, chlorine, and other components of the earth's crust. He took the science of these cycles from Swedish chemist Arrhenius and others, but he gets credit for being among the very first to see how crucial these cycles are to life and how they fit into and sometimes exacerbate one another. The number of people in the world more than doubled since his time, and production per person far more than doubled, so we are now forced to attend to such cycles as those of fresh water and carbon, and they enjoy considerable attention in legislative bodies and space in late twentieth-century newspapers and magazines. But all that is a development of the late twentieth century, so we have to remind ourselves that these issues were a preoccupation of Lotka from early in the century when there were very few far-sighted enough to think they might have practical consequences.

Lotka refers to the Swedish chemist Arrhenius who first saw from theoretical considerations the possibility of man-made carbon dioxide leading to global warming, and he cites Arrhenius' finding that the air is becoming richer in CO_2. Moreover plants, especially trees, withdraw CO_2, and the cutting of trees releases CO_2; the oceans could offset the tree-cutting by increasing their absorption of the gas. These are only three of the determinants of free CO_2 in the air and no one has yet shown a way to isolate the man-made CO_2 released by the burning of fossil fuels and cutting of forests from changes in the capacity of the oceans to hold carbon. Seventy-five years ago Lotka had a realistic view of the complexities of global equilibrium and the difficulties of forecasting what forces would dominate within it. The earth is a vast and complex physico-chemical-biological engine in constant motion, and if this is now a commonplace, Lotka was in the forefront of those who made it so.

The Lotka–Volterra equations are much referenced. Vito Volterra, an Italian mathematician, was asked by his biologist son-in-law Humberto D'Ancona whether he could explain the cycles of fish population in the Adriatic Sea. Lotka, also in the first half of the 1920s, wanted to see whether the relation of predator and prey could be assimilated to the autocatalytic processes of chemistry he had studied earlier. A pair that comes to mind as a model for both Lotka and Volterra are foxes ($F(t)$) and rabbits ($R(t)$), in which a carnivorous species is dependent on a grass-eating species. The flow of energy is *sun* → *grass* → *rabbits* → *foxes*. Lotka (1925) and Volterra (1927) seem independently to have noted and explained the same features of predation, especially the cycles it generates. The simplest form of the Lotka–Volterra equations is

$$F'(t) = -\alpha F(t) + \gamma R(t) \qquad [2]$$

$$R'(t) = \beta R(t) - \delta F(t) \qquad [3]$$

Primed symbols are derivatives, and the four positive numbers α, β, γ, δ are fixed parameters. The $-\alpha$ in the first equation tells us that in the absence of the prey (i.e., if $R(t)$ is zero) then the predator declines in numbers, to ultimate extinction. The second equation shows the prey increasing indefinitely in the absence of the predator. That γ has a plus sign in the first equation allows for the fact that with more abundant prey the predator grows faster. With other configurations of signs the same equations can represent competition, symbiosis, etc. They can incorporate limits on the amount of grass and can be extended to three or more species.

Lotka's indifference to what he considered mathematical niceties was matched by a quality found in many other original minds, great sensitivity on priorities. He engaged in a long and indecisive controversy with R.R. Kuczynski. The history of science is pockmarked with such disputes. What puts this one in a special class is that Paul Samuelson undertook a uniquely thorough investigation (1977) to sort out the issues, and credits Lotka, Böckh, and Kuczynski, among others. With access to the Princeton archives where Lotka's private papers, as well as his publications, are deposited, he demonstrates with the help of pages of solid algebra that Lotka's predecessors turned up a good part of the renewal story, but their work was incomplete. Lotka had both the theoretical interest and the command of data, plus the drive to follow through.

So, Samuelson concludes on this issue, the brighter light has drowned out the earlier light. But now that the dust has died down, it looks as though Lotka's place in history, which is a secure one, is based less on the priorities he claimed so vigorously in controversy than on his anticipation and objective treatment of a complex of problems revolving around population and environment, problems barely suspected when he lived, that we now realize will be with us for a long time. A half-century after his death he is very much alive through his books and papers, and in the demographic profession we can still say, with Frank Notestein (*see* Notestein, Frank Wallace (1902–83)), that 'his is the greatest name.'

See also: Demography, History of; Demography: History since 1900; Statistics, History of.

Bibliography

Cannan, E., 1895. The probability of a cessation of the growth of population in England and Wales during the next century. Economic Journal 5, 505.

Euler, L., 1760. A general investigation into the multiplication and mortality of the human species [1970. Theoretical Population Biology 1, 307 (trans. N Keyfitz and B Keyfitz)].

Feller, W., 1941. On the integral equation of renewal theory. Annals of Mathematical Statistics 12, 243.

Golubitsky, M., Keeler, E., Rothschild, M., 1975. Convergence of the age structure: Applications of the projective metric. Theoretical Population Biology 7, 84–93.

Keyfitz, B.L., Keyfitz, N., 1997. The McKendrick partial differential equation and its uses in epidemiology and population study. Mathematical and Computer Modelling 26, 1–9.

Kuczynski, R., 1931. Fertility and Reproduction: Methods of Measuring the Balance of Births and Deaths. Macmillan, New York.

Lotka, A.J., 1911. A problem in age distribution. Philosophical Magazine 21, 435.

Lotka, A.J., 1925. Elements of Physical Biology. Williams & Wilkins, Baltimore, MD. Republished, 1956, as Elements of Mathematical Biology. Dover Publications, New York.

Lotka, A.J., 1931a. The extinction of families. Journal of the Washington Academy of Sciences 21, 377–453.

Lotka, A.J., 1931b. Orphanhood in relation to demographic factors. Metron 19, 37–109.

Lotka, A.J., 1939. Théorie analytique des associations biologiques. Part I: Principes; Part II: Analyse démographique avec application particuliére á 1'espéce humaine. Actualites Scientifiques et Industrielles 187, 780. Hermann & Cie, Paris [1998 Analytical Theory of Biological Populations (D.P. Smith and H. Bossert, Trans.). Plenum, New York and London].

Lotka, A.J., Dublin, I., 1946. The Money Value of a Man, rev. edn. The Ronald Press, New York.

Lotka, A.J., Sharp, F., 1911. A problem in age distribution. Philosophical Magazine Ser 6 (21), 435–438.

Samuelson, P., 1977. Resolving a historical confusion in population analysis. In: Smith, D., Keyfitz, N. (Eds.), Mathematical Demography: Selected Papers. Springer-Verlag, New York.

Volterra, V., 1927. Variazioni e fluttuazioni del numero d'individui in.specie animali conviventi. C Ferrari, Venice, Italy.

Love and Intimacy, Psychology of

Panos Kordoutis, Panteion University of Social and Political Sciences, Athens, Greece

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by H.T. Reis, volume 13, pp. 9091–9094, © 2001, Elsevier Ltd.

Abstract

Intimacy occurs when two people are responsive to each other, self-disclose, and demonstrate involvement behaviors. This interactive process may culminate in love and the establishment of a close relationship. Love is the evolutionary result of innate bonding and mating systems such as attachment, caregiving, and sex. It is a variable experience taking the form of attachment, compassionate, companionate, or romantic love. Love styles are relational attitudes about love. Love types result from combining intimacy, passion, and commitment. Love stories explain the social origin of love's meaning and change. Companionate love features are prominent in laypersons' love prototypes.

If close relationships form the foundation and theme of the human condition, love and intimacy represent the invisible master plan that humans apply in building their close relationships throughout their lifetimes. The two concepts are often closely associated in popular and scientific discourse as they capture core phenomenological experiences and processes involved in the initiation, development, and maintenance of relationships. They both reflect the universal and biologically ingrained human need to create and maintain dyadic attachments, affiliations, and networks of interdependence which secure that a cohesive and self-perpetuating community supports and maximizes individual well-being. At the individual level, love and intimacy directly or indirectly affect personal well-being; self-worth; and the motivation for self-expansion, social-adjustment, and community connectedness. At the close relationship level, they affect both relationship stability and longevity, perceptions or relationship quality, satisfaction, expectations about the relationship, and the motivation of maintaining it (Berscheid and Regan, 2005).

Intimacy has for the most part been viewed as a dynamic interactive process between two people that enhances their motivation to keep engaging in behaviors that may gradually increase knowledge of each other, mutual understanding, and bonding. In contrast, love has been studied as a phenomenological cognitive–affective state, resulting from a broad range of processes including intimacy. In our discussion of intimacy, we will first analyze the specific qualitative characteristics of the behavioral patterns that people elicit in interpersonal contacts. Subsequently, we will describe a model that attempts to explain by means of what process, the interacting behavioral patterns elicited by the dyad involved in the interpersonal contact, generate, maintain, and augment intimacy.

Intimacy as a Relationship Process

Intimacy may well refer to a transient experience of sharing private personal information with another fellow human being, such as an unknown fellow traveler, sitting next to us, on a long transatlantic flight. However, it usually refers to entire patterns of verbal and nonverbal behaviors occurring at the onset or within a relationship, which bear characteristics consistently reflecting motivation for (1) self-disclosure, (2) self-involvement, and (3) responsiveness, that is for seeking signs of interest and involvement on behalf of the other person (Laurenceau et al., 2004a,b; Prager, 2000; Prager and Roberts, 2004; Reis et al., 2004).

Self-Disclosure

Self-disclosure may include revealing objective private information about the self as well as expressing motives, needs, wants, goals, fears, and feelings in general (Laurenceau et al., 2004a,b). For instance, individuals may talk about their personal history, relationships with family members, important career milestones, people who have influenced their lives, past positive and negative/traumatic experiences, past failed relationships, sadness, risks, hurt, pain, emotional nakedness, feelings of being a special person, superior or inferior to others, privileged or flawed, and so on. At the initiation of a close relationship, the objective of self-disclosure behaviors is to establish (1) the subjective perception of a virtual safety zone between the two interacting individuals, so that the more vulnerable aspects of the self can be communicated and exposed without the apparent risk of getting hurt, (2) to make self fully known to other for what it is, for its brighter and darker sides, and (3) to seek understanding and acceptance and ultimately claim unconditional positive regard for what the self is as a whole. To the extent that these objectives have been attained, intimacy's role shifts to transforming each relationship member's perception of his/her individual outcome interdependence in the relationship. The interdependent partners acquire a 'we' perspective and by perceiving relationship outcomes as own, they tend to pursue maximization of joint gain. Intimacy transforms the relationship into a union of trust (Kelley, 1983/2002; Rusbult et al., 2004).

Involvement Behaviors

Involvement behaviors are, for the most part, nonverbal and convey attentiveness, interest, and active participation in the interaction with other (Prager, 2000). People may thus actively seek eye contact and gaze at each other for an extended time, may sit closer to one another, assume open posture, lean

forward, gesture, smile or in other ways show facial animation, and nod their heads while listening and talking. As interactions among the intimate couple unfold and multiply, touch is among the behaviors that may further intensify the intimacy experience. Invited touch of parts of the body that have been identified as vulnerable, such as the face or the torso is very likely to yield strong feelings of intimacy. These feelings are intensified when individuals sustain the touch, as in the case of embracing other or when placing arms around the shoulder or the waist. When the interaction between two individuals has ensued from initial sexual attraction, or intimacy has brought about sexual attraction, sexual contacts and sexual intercourse among the couple can escalate behaviors of intimacy and the breadth and depth of the experienced intimacy. However, sexual touch or sexual expressions do not necessarily bring about intimacy. When interacting parties have the subjective feeling that the 'safety zone,' mentioned above, protecting the more vulnerable aspects of self from exposure, is not in place, sexual advances or any involvement behavior may irreversibly hamper intimacy or destroy the relationship per se.

Involvement may also be expressed by means of verbal behavior that conveys either the message of desired proximity or simply positive affect for being close. The functional objective of involvement behaviors in establishing intimacy is to demonstrate behavioral intention for approaching and getting involved with other and to test other's responsiveness and behavioral intention for approaching self. In developing and maintaining a relationship, involvement behaviors mainly function to maintain immediacy of emotional reward among the couple and support mutual interdependence with positive affect and management of negative affect.

When people are motivated to initiate and maintain intimate relationships, they increase the frequency of involvement behaviors, actively employ a broad range of modes of communicating (phone, texting, e-mailing, social media, etc.), and subjectively 'make time' to engage in interactions.

Responsiveness

Responsiveness refers to the active and focused attention, the discerning perception, and the timely and sensitive response to what is being said and done by other during an intimate interaction (Prager, 2000; Reis et al., 2004). Responsiveness in intimate interactions requires that the nature of disclosure and involvement behaviors emitted by other, first be identified for what they are, second be discerned for the specific intention or content expressed, and third be addressed in one way or another or ignored.

Responsiveness behaviors may be both verbal and nonverbal and their function is to denote that the disclosing person is followed and understood with attention and empathy. In particular, nonverbal behaviors play a substantial role in people's experience of intimacy perhaps due to their immediate association with affective reactions and the fact that they are, at least, partially involuntary. Such behaviors may include attentive listening, head nodding, eye contact and gazing, variation of voice tones, facial expressions, smiles, gestures and body postures such as leaning forward or assuming an open posture. Occasional bodily contact, such as hugging and touching other in a heartening and reassuring manner during critical disclosing moments, may also occur. Although such behaviors superficially resemble behaviors of involvement, they are essentially different because their function is to facilitate disclosure (e.g., by establishing appropriate physical distance) and other's voluntary choice for testing engagement in involvement behaviors. Short and extensive or elaborate verbal responses in intimate interactions also function to communicate that the listener is interested in what is being said and willing to listen further and invest time in the interaction. Moreover, they can overtly convey understanding, acceptance, respect, and genuine care.

Partners of a close relationship recognize each other's responsiveness by using four procedural criteria: (1) How often other responds to one's communicative behaviors and how predictable are their responses. High frequency indicates high responsiveness, whereas predictability creates a sense of familiarity and warmth, as other seems to 'always' understand own behavior; (2) How many of own behaviors are responded to by other appropriately, that is taking into consideration their exact content (obviously, the statement 'I am desperate' cannot be responded with 'shall we have something to eat?'); (3) How many of the responses are suitably elaborate; in some interactions initially responding 'I understand' or 'I know' to 'I am sad' may be fitting, whereas in others a more extended answer would have been more welcome; and (4) How timely are the responses; timely response suggests reassuring predictability of other's behavior toward self, which in turn implies controllability. If these criteria are met, relationship partners feel that other values the relationship and their interaction, otherwise partners feel frustrated and the relationship may lead to dissolution.

Disclosure, involvement, and responsiveness behavioral patterns do not function independently, but are sensitive and proactive/reactive to the onset of each other, the timing and the source of expression. Actually, intimacy seems to result from a highly interactive, recurrent cybernetic loop among the behavioral patterns elicited by the two communicating individuals. Thus, the dyadic generation of intimacy needs to be effectively depicted by a process model. Below, we roughly outline such a model based on the influential interpersonal model of intimacy that has been proposed by Reis and Shaver (1988), Laurenceau et al. (2004b).

A Dyadic Intimacy Process Model: Disclosure, Involvement, and Responsiveness at Work

The model describes intimacy as an interpersonal interaction between A and B that goes through a preconditional proactive–reactive stage (I), a mediating perceptive stage (II), and an outcome stage (III). The latter stage may feed back into stage I, regulating the continuation of the intimacy interaction cycle, the shifting of the proactive or reactive roles (e.g., disclosing vs listening), its climaxing or anticlimaxing, and/or eventual stop:

I. This stage starts off with proactive self-disclosure behaviors on the part of A. It may be a piece of private information or a feeling or both e.g., 'I am so sad…' or 'I am so sad, I just broke up with my relationship.' It continues with reactive or responsive behaviors on the part of B. They could

be a verbal response timely expressed, e.g., 'Why? What happened?' accompanied by an eye contact showing sympathy. B may also sit closer, showing readiness to listen and encourage A to talk about his or her feelings. While listening, B may follow A by maintaining eye contact, head nodding, and using facial expressions to demonstrate understanding, sympathy, and support. On the other hand, B's responsive behaviors may denigrate or 'invalidate' the feeling expressed or show disinterest and change the subject, e.g., by saying 'Oh, breaking up is always sad for everyone! Are you going to watch the Oscars tonight?'

II. If A perceives B's responsive behaviors as demonstrating that (1) she is being followed and understood for what she says, (2) she is validated, that is the significance of the contents of self-disclosure is recognized for what it is and the discloser is respected for her experiences and feelings, (3) she is cared for, as a person, that is B seems to be to a greater or lesser extent emotionally involved with her, then stage III, the outcome stage is activated.

III. A perceives B's reactions as a positive outcome of the interaction and a sign that more proactive intimacy behaviors may be rewardingly tried out by him or her (that is, by A). At the same time, such reactions work as a predictive sign that similar proactive behaviors are to be expected from the other party (B). In this case, the process is repeated across time and even across different settings with A and B exchanging roles (proactive vs reactive). The feedback loop (I–III) may climax with even more intense and diverse self-disclosure, responsiveness, and involvement behaviors.

When the recurrent functioning of the intimacy model's feedback loop, and the spiraling exchange of disclosure, responsiveness, and involvement behaviors is mutually rewarding, the intimacy process may be sustained for a long period of time, across multiple and diverse interactions. Eventually, the interacting parties start exerting strong, frequent, and diverse effects on one another (Kelley, 1983/2002), establishing a close or 'intimate' relationship.

Love

Love can better be understood as the subjective perception of an affective–cognitive state that may have resulted from dyadic intimacy as well as any of several relationship processes occurring in multiple dyadic, community, or other contexts, ranging from the romantic relationships, family relationships, one's fellow human beings, or even one's relationship with God (Berscheid and Regan, 2005). Indeed, love is a broad, multifaceted concept and although it is relatively easy to self-report when experienced, it is nearly impossible to define in abstract terms, without reference to its specific characteristics and the relationship settings in which it occurs (Kelley, 1983/2002).

In our discussion of love, we will first present the biopsychological perspective, which views love as the by-product of two biological processes that have assisted humans in successfully responding to the evolutionary challenges of social living and reproduction; one refers to the human bonding, the other to mate selection. Then, we will present the taxonomies of love, that is theorizing about the contextual variations in the meaning of love from both a scientific and a lay perspective.

The Biopsychological Perspective

Shaver et al. (1988) in their highly influential theory on romantic love, based in part on Bowlby's (1979) ethological theory of attachment, suggest that there are three innate motivational systems: attachment, caregiving, and sex. All three have evolved over the course of human development to increase the likelihood that the vulnerable human infant, which requires nurturing support for an extended period of time, will survive and reach a reproductive age. Each system has different aims and dictates different strategies in achieving these aims. The aim of the attachment system is to help the person establish a psychological bond with an attachment figure, so that he or she can seek shelter when coming across a threat and feel that relationships, in general, can be more or less predictable and trustworthy. The caregiving system yields protection and support for significant others in need. The sexual system secures that sexual partners approach and focus on each other at the expense of other activities or others with the aim of reproduction. Each of the three motivational systems may receive different feedback from the infant's or child's caregiver (usually the mother) and his or her significant others. In turn, the infant or child, based on this feedback, establishes behavioral strategies or patterns of probabilistic responses to relational stimuli that guide his or her relating behavior later, in adult life, including mating and romantic relationship behavior.

Research (Shaver and Mikulincer, 2006, 2011) has examined individual differences in attachment strategies (secure, avoidant, and anxious) and their association to the other two systems. More specifically, securely attached people are more likely than nonsecurely attached ones to provide effective care to a needy partner as they have a stringer sense of efficacy and controllability in coping with distress. They also tend to be more responsive and caring. In contrast, avoidant people are less willing to be compassionate to others and provide care. Anxiously attached people are preoccupied with their own needs and are thus unable to react to others' suffering with empathy. In contrast, they react with own distress which hampers their ability to provide effective, sensitive, and unintrusive care. In terms of sexual expression, anxiously attached people seek to maintain their own protection and security and would have trouble attending to their partner's sexual needs and preferences. They tend to use sex to reassure themselves and cast off their fear of abandonment. Avoidants may feel uneasy during sexual intercourse that requires even the minimum of intimacy, as they dislike exposure to vulnerability and openness; they tend to dismiss care and avoid intimacy often involved in sexual intercourse. Avoidant people may also be using sex instrumentally, as a way of manipulating others in pursuing ulterior goals.

Fisher's (2006) neurological approach also proposes three different brain-based systems: the sex drive, attachment, and attraction. Sexual desire springs out of the sex drive system, and aims at successful reproduction. Attachment refers to nest building, the establishment of proximity with other, and mutual

caregiving. Attraction describes investing energy and time at focusing on one particular mate and on mate-guarding. Fisher's sex drive and attachment systems resemble Shaver's sexuality and attachment systems, respectively. However, Fisher's definition of 'attraction' is more clearly associated with passionate and romantic love. By contrast, Shaver's conceptualization of the caregiving system is more related to compassionate or companionate love. All three kinds of love will be discussed in the taxonomies of love section.

The functioning of the systems proposed above seems to be supported by biobehavioral mechanisms, common in mammals, including humans (Leckman et al., 2006). For instance, the contention of attachment system (Shaver's) and attraction systems that falling in love is stressful and may thus require anxiety regulation is in consonance with the finding of increased cortisol levels among passionate lovers. On the other hand, peptides, such as oxytocin, have been found to reduce and relax the anxiety involved in both passionate love attachment and other forms of bonding. Low serotonin levels seem to be responsible for anxious obsessive behaviors often observed in people in love.

The evolutionary approach to love (Kenrick, 2006) states that different kinds of love are relevant to different problems humans have encountered during their evolution. For instance, humans need to attract potential mates; once they do so, they have to maintain a relationship with them for a significant time period so that their offspring is taken care off in the best possible way. In responding to this problem, males, generally, have acquired the bias to ascribe more importance to attractiveness in females, because attractiveness serves as a cue to fertility. In contrast, females, in selecting and, more particularly, in maintaining a mate pay attention to his status, that is to his ability to provide for the offspring. Thus evolutionary problems such as mate-seeking and parental care, that require different solutions, may have dictated the development of different kinds of love. The former may be subjectively perceived as being primarily focused on sexuality and passion, whereas the latter on caregiving and attachment.

Love Taxonomies

Taxonomies of love are the result of systematic efforts to organize the multiplicity of phenomena described under the use of the word 'love' in the broad array of social contexts it usually occurs. Love taxonomies are useful not only because they facilitate us in describing similar observable phenomena associated with each kind of love, but also because they help us understand the causes of such phenomena and predict their future course (Kelley, 1983/2002).

The Quadrumvirate of Love

Ellen Berscheid (2006) has suggested that the already discussed attachment love and three more kinds of love can cover all love-associated phenomena. Thus, below we turn to the discussion of compassionate, companionate, and romantic love.

Compassionate love motivates altruistic interest for another's welfare as well as actions to promote it (see also Fehr and Sprecher, 2009; Sprecher and Fehr, 2005). It has also been labeled 'altruistic,' 'charitable,' 'brotherly,' 'communal' love, 'agape,' and 'B-love'; behaviors associated with it have also been studied under different names, such as 'prosocial behavior,' 'caregiving,' 'support behaviors,' and 'communal responsiveness' (Clark and Monin, 2006). Individuals engaging in compassionate love do not expect that their actions will be reciprocated. Compassionate love is based on the caregiving system described in the context of Attachment Theory. Humans are born weak and vulnerable, and in order to survive and reach reproductive age, they need the 'asymmetrical' support of an older, stronger, and wiser member of their species. The survival utility of this asymmetry has been inherited to our species members over evolutionary time. Compassionate love involves responsiveness to clear or discreet signs of need, sensitivity in understanding the nature of the need, and unobtrusive behaviors to alleviate it.

In contrast to compassionate love, companionate love (see also Hatfield, 1991) is contingent upon rewards and punishments. The evolutionary and survival value of the pain–pleasure principle, basic to learning and behavioral theories, underlies this kind of love. People love those who love them back, or more precisely those who reward them. Companionate love has been discussed in the past under terms such as 'liking,' 'friendship love,' 'philias,' 'affiliation,' 'pragmatic love,' and 'affection.' Companionate love relationships are often based on characteristics that are intrinsically rewarding to both relationship partners such as physical attractiveness; familiarity; similarities in background, goals, and values; and common interests. Companionate love relationships are characterized by physical or virtual proximity, stability, positive feelings, mutually predictable and responsive behaviors, mutual recognition and acceptance, and the exchange of actual rewards. Intimate couples who experience companionate love feel that their lives are deeply interdependent; they see, understand, and accept each other for what they really are and tend to be best friends. Finally, they show mutual affection and concern for each other's welfare, have positive feelings for their partner, and tend to prioritize his or her needs against those of third parties.

Romantic love has been studied under names such as 'passionate,' 'erotic,' and 'addictive' love (see also Hatfield, 1991; Hatfield et al., 2007). It rises out of sexual desire and is based, to a large extent, on the neurochemical mechanisms we have previously discussed. Of course, romantic love is hardly limited to sexual desire. In attempting to capture its rich phenomenology, one has to unveil its cognitive, affective, and behavioral components. Cognitions reflecting romantic love include idealization of partner and the relationship, intrusive and often obsessive thinking about them, and a strong desire to know them and be known (self-disclosure). Emotions of romantic love comprise attraction, sexual attraction in particular, intense and conflicting feelings, elation and bliss, anguish and disappointment, as well as longing for reciprocity. Behaviors indicating romantic love include 'studying' the other person, trying to 'read' and determine their feelings, providing service to them, and maintaining physical closeness. Further, individuals in love tend to behave toward their loved ones according to their imaginary expectations about them, rather than according to realistic appraisals (Snyder et al., 1977; Berscheid and Regan, 2005). At the same time, as the basis of relating is fleeting and volatile emotionality, the ensuing relationship tends to be unstable, unpredictable, and sometimes unrealistic. Lovers do not see each

other as actual persons with real needs, but as projections of how they would like them to be (Hendrick and Hendrick, 2000; Berscheid and Regan, 2005).

As passion also tends to quickly deplete the organism's resources, romantic love withers away and cannot support the relationship for long. For this reason, relationships solely based on romantic love do not fare well in the long run. In the best case scenario, as romantic love grows weaker, companionate love, enhanced by intimacy built up, may gradually replace it, stabilizing the relationship and placing it on a more realistic basis. Romantic and companionate love, are not mutually exclusive; both kinds seem to be present at the initiation of a relationship and may function in a complementary rather than competitive fashion. In fact, most couples report that their lover is also their best friend (Hendrick and Hendrick, 2000; Berscheid and Regan, 2005).

Love styles

Love styles are ideologies of love or belief–attitude systems about how one loves other; they may not be necessarily subjectively perceived or consciously stated as such, but they are certainly expressed in interactions with others in intimate settings (Hendrick and Hendrick, 2006). In that sense, they can be construed as interpersonal cognitions about loving others that loosely dictate romantic social behavior. Thus, love styles are constructs comprising beliefs and attitudes about loving, relational personality traits, and a variable emotional core, reflecting the general need for affiliation. There are six dominant love styles: Eros (passionate, erotic, or romantic love), Ludus (game-playing uncommitted love), Storge (friendship or companionate love), Pragma (practical, calculating, and instrumental love), mania (possessive, obsessional, or dependent love), and Agape (altruistic or compassionate love). Plenty of studies have addressed research questions on the association between love styles and demographic, personality factors, and relationship outcome variables (e.g., satisfaction, sexuality, adjustment, quality, stability, longevity) (Hendrick and Hendrick, 2006; Neto, 2012).

The Duplex Theory of Love: The Triangular Theory of Love and the Theory of Love as a Story

The *Triangular Theory of Love* defines the concept of love structurally, in terms of its components. By understanding the functions and interrelationships of love components, Sternberg (2006) attempts to grasp the variations in the meaning of love, as well as the variability in behaviors associated with it.

The Triangular Theory of Love

According to the Triangular Theory of Love (2004), there are three components of love that can be metaphorically viewed as the three vertices of triangle. These are intimacy, passion, and decision/commitment and can be identified in various forms in older theories; moreover, they are intuitively meaningful to laypersons (Aron and Westbay, 1996). The component of intimacy comprises the phenomenological outcomes of the intimacy process that we have previously discussed. These are feelings of closeness, warmth and bonding, communication, understanding, caring, compassion, and empathy. Similarly, the passion component includes feelings familiar from our description of romantic love, such as sexual arousal and intense desire, longing for other's presence, obsession with the loved one, excitement, energy, enthusiasm, a sense that attraction is inevitable, and that attachment to other was destined. The decision/commitment component, in the short term, refers to the realization and decision that one loves other and is interested to enter and tentatively remain in a relationship. In the long term, it is the decision to be committed in maintaining it. Sternberg further proposes that the three components can be combined in eight possible limiting cases where the amount or proportion of each component is, for reasons of illustration, estimated to the maximum (or is fully present) and to the minimum (or is entirely absent). Thus, (1) absence of all the three components is nonlove, (2) intimacy alone is what we experience as liking, (3) passion alone is nothing but infatuation, (4) commitment by itself is empty love, (5) intimacy combined with passion without commitment is romantic love, (6) intimacy combined with commitment without passion is companionate love, (7) passion plus commitment without intimacy is fatuous love, and (8) intimacy combined with both passion and commitment is consummate or complete love. On an individual basis, all three components are strongly associated with happiness and satisfaction in relationships. Couples, however, tend to be happier to the extent that the balance among intimacy, passion, and commitment each of them desires approximates that of the other. For example, a couple may experience distress, if one of them wants a lot of passion but not much commitment and the other wants a lot of commitment while has no motivation to engage in full passion or wishes to withhold passion.

Structural analysis cannot inform us on the social origin and importance of love components, nor does it account about the dynamic change of love component combinations within a relationship. The theory of *Love as a Story* (Sternberg, 1995, 2006) explains where the meaning of love components comes from and how component contents and combinations may change over time in one's own relationship or across relationships.

The Theory of Love as a Story

Love Stories are guiding metaphors similar to cognitive schemata or scenarios that help people predict and possibly control the behavior of other and self within an intimate relationship. Stories come from our experiences with narratives, tales, books, TV series and movies, and observation of the relationships of our parents, our friends' parents, and our own. Stories have protagonists and require the casting of the 'supporting role,' that is of a partner who can fit the scenario and will accept to play the role. Some are willing to do so, some will not, and ask for the scenario to be edited or even changed. Compatibility or complementarity between the two basic roles and compatibility between role and scenario are of critical importance for relationship satisfaction and other relationship outcomes. Individuals differ in their flexibility to negotiate and edit or even change their preferred relationship love story. Our stories evolve and may change over time. However, we carry a hierarchy of preferred love stories and we are happier when we succeed in inviting someone to 'play' in our most preferred one. In contrast, if we eventually concede to participate in a story that is not high in our hierarchy, that relationship will probably yield less satisfaction for us and we are more likely to abandon it for a story higher in our hierarchy.

Research has identified a variety of about 26 common stories (Sternberg, 1995, 2006). Examples are (1) the fairy tale story, in which the protagonist is a prince or a princess, they expect that special other and live happily ever after, (2) the business story, according to which a relationship is run like a business with gains and losses, partners can make practical agreements and plans, (3) the travel story, in which partners travel together, along the same path, to a common destination, (4) the police story, in which one partner (the policeman) sets the rules and is always vigilant that other observes them, (5) the horror story, in which one partner abuses the other, (6) the collector story, according to which one partner collects lovers and admires them as showcase of precious objects. Research has shown that some stories are more adaptive than others (e.g., the travel and the fairy tale story vs the horror or the police story) as they help the individual adjust better to the relationship, while the relationship per se yields better outcomes. Couples have also been found to be more satisfied when their story hierarchies match.

The Prototype Approach

Seeking a breakthrough to the difficulties associated with defining love, Fehr (1994, 2006) resorted to people themselves and their own understanding of the love concept. The love prototype is based on spontaneously elicited features of love as well as on measuring how important these features are to laypersons in cognitively representing the concept. Prototype research found that companionate love features were clearly predominant in the lay perception of love. For example, features consistently found include trust, caring, honesty, friendship, and respect. Even when features were listed explicitly for romantic love, trust, honesty, and happiness ranked well above passion and sexuality. On the other hand, the same line of research has revealed that the romantic–companionate love distinction is meaningful and valid to laypersons.

Conclusion

Intimacy is the subtle process that brings to bear love, although potent feelings of romantic or passionate love may be initiated before even the timid steps of intimacy development have been taken. Indeed, romantic love only needs sexual desire, realism enough to focus on the desirable aspects of other, and a lot of imagination to augment them while disregarding undesirable aspects. Its evolutionary role is to lure the couple into the deceptive magic of love, away from daily concerns and other interests. Nevertheless, it is intimacy that will enhance the climaxing of romantic love and the one that will secure basic relationship functioning when passion inevitably subsides; intimacy will prepare the rise of companionate love that may eventually support the continuation and longevity of the close relationship.

Love, of course, is not limited to close relationships and thus intimacy is not important only for the romantic couple. Being sensitive and responsive to others, as intimacy suggests, is the basis upon which compassionate love thrives. Being 'asymmetrically' responsive to the need of others, without expecting reciprocation, has obviously been an evolutionary advantage for the survival of our species, because all people, irrespective of gender and age, will find themselves at one time or another in times of need. Attachment theory is keen on the notions of responsiveness, unconditional love, and caregiving. The mother, in the mother–infant archetypical asymmetric relationship can set with her responsive behavior, the basic precondition for the child to develop as an adult, a personal sense of autonomy, and the perception that his or her social world is predictable and trustworthy; he or she can love others and be loved without fearing separation, abandonment, and betrayal.

Love can also take on many forms and expressions. It is important to know what is intended behind each pattern of expression, love style, type of love, or love story. Although, romantic love during the peak of passion can take on a nearly irrational and obsessive character, loving settings allow partners to express their own perception of their relationship, listen to other's and negotiate upon it. Love styles, types, and stories can significantly contribute to our understanding of how we perceive the intimate relationship we are in, and what we expect from it and our partner. Individuals can alter their love stories if they feel that they have been trapped in the repetition of a destructive scenario and couples can mutually edit their love story until it becomes compatible and productive to both.

See also: Cultural Influences on Interpersonal Relationships; Integrated Attachment Theory; Interpersonal Attraction, Psychology of; Intimate Partner Abuse, Applied Research on; Prosocial Behavior and Empathy; Satisfaction and Commitment in Long-Term Heterosexual Relationships; Social Relationships in Adulthood.

Bibliography

Aron, A., Westbay, L., 1996. Dimensions of the prototype love. Journal of Personality and Social Psychology 70 (3), 535–551.

Bowlby, J., 1979. The Making and Breaking of Affectional Bonds. Tavistock, London.

Berscheid, E., 2006. Searching for the meaning of love. In: Sternberg, R., Weis, K. (Eds.), The New Psychology of Love. Yale University Press, New Haven & London, pp. 171–184.

Berscheid, E., Regan, P., 2005. The Psychology of Interpersonal Relationships. Pearson Education, Upper Saddle River.

Clark, M., Monin, J.K., 2006. Giving and receiving communal responsiveness as love. In: Sternberg, R., Weis, K. (Eds.), The New Psychology of Love. Yale University Press, New Haven & London, pp. 200–224.

Fehr, B., 1994. Prototype-based assessment of laypeople's views of love. Personal Relationships 1, 309–331.

Fehr, B., 2006. A prototype approach to studying love. In: Sternberg, R., Weis, K. (Eds.), The New Psychology of Love. Yale University Press, New Haven & London, pp. 225–248.

Fehr, B., Sprecher, S., 2009. Prototype analysis of the concept of compassionate love. Personal Relationships 16, 344–364.

Fisher, H., 2006. The drive to love: the neural mechanism for mate selection. In: Sternberg, R., Weis, K. (Eds.), The New Psychology of Love. Yale University Press, New Haven & London, pp. 87–115.

Hatfield, E., Rapson, R.L., Martel, L.D., 2007. Passionate love. In: Kitayama, S., Cohen, D. (Eds.), Handbook of Cultural Psychology. Guilford Press, New York, pp. 760–779.

Hatfield, E., 1991. Passionate and companionate love. In: Sterberg, R.J., Barnes, M.L. (Eds.), The Psychology of Love. Yale University Press, New Haven, pp. 191–217.

Hendrick, C., Hendrick, S., 2006. Styles of romantic love. In: Sternberg, R., Weis, K. (Eds.), The New Psychology of Love. Yale University Press, New Haven & London, pp. 149–170.

Hendrick, S., Hendrick, C., 2000. Romantic love. In: Hendrick, C., Hendrick, S. (Eds.), Close Relationships, a Sourcebook. Sage Publications, Thousand Oaks, pp. 203–216.

Kelley, H.H., 1983/2002. Love and commitment. In: Kelley, H.H., Bersceid, E., Cristensen, A., Harvey, J.H., Huston, T.L., Levinger, G., McClintock, E., Peplau, L.A., Peterson, D.R. (Eds.), Close Relationships, pp. 265–314. Clinton Corners, NY.

Kenrick, D., 2006. A dynamic evolutionary view of love. In: Sternberg, R., Weis, K. (Eds.), The New Psychology of Love. Yale University Press, New Haven & London, pp. 15–34.

Laurenceau, J., Feldman, L., Pietromonaco, P., 2004a. Intimacy as an interpersonal process: the importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchange. In: Reis, H., Rusbult, C.E. (Eds.), Close Relationships. Psychology Press, New York, pp. 199–212.

Laurenceau, J., Rivera, L.M., Schaffer, A.R., Pietromonaco, P.R., 2004b. Intimacy as an interpersonal process: current status and future directions. In: Masek, D.J., Aron, A. (Eds.), Handbook of Closeness and Intimacy. Lawrence Erlbaum Associates, Mahwah, New Jersey, pp. 61–80.

Leckman, J., Hrdy, S.B., Keverne, E.B., Carter, C.S., 2006. A biobehavioral model of attachment and bonding. In: Sternberg, R., Weis, K. (Eds.), The New Psychology of Love. Yale University Press, New Haven & London, pp. 116–148.

Neto, F., 2012. Perception of love and sex across the adult life span. Personal Relationships 29 (6), 760–775.

Prager, K., 2000. Intimacy in personal relationships. In: Hendrick, C., Hendrick, S. (Eds.), Close Relationships, a Sourcebook. Sage Publications, Thousand Oaks, pp. 229–244.

Prager, K., Roberts, L., 2004. Deep intimate connection: self and intimacy in couple relationships. In: Masek, D.J., Aron, A. (Eds.), Handbook of Closeness and Intimacy. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 43–60.

Reis, H.T., Clark, M.S., Holmes, J.G., 2004. Perceived partner responsiveness as an organizing construct in the study of intimacy and closeness. In: Masek, D.J., Aron, A. (Eds.), Handbook of Closeness and Intimacy. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 201–228.

Reis, H.T., Shaver, P., 1988. Intimacy as an interpersonal process. In: Duck, S. (Ed.), Handbook of Personal Relationships. Wiley, Chichester, pp. 367–389.

Rusbult, C.E., Kumashiro, M., Coolsen, M.K., Kirchner, J.L., 2004. Interdependence, closeness and relationships. In: Masek, D.J., Aron, A. (Eds.), Handbook of Closeness and Intimacy. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 137–162.

Shaver, P., Hazan, C., Bradshaw, D., 1988. Love as attachment: the integration of three behavioral systems. In: Sterberg, R.J., Barnes, M.L. (Eds.), The Psychology of Love. Yale University Press, New Haven, pp. 68–99.

Shaver, P., Mikulincer, M., 2011. An attachment-theory framework for conceptualizing interpersonal behavior. In: Horowitz, L.M., Strack, S. (Eds.), Handbook of Interpersonal Psychology. John Wiley & Sons, Hoboken, pp. 17–36.

Shaver, Ph, Mikulincer, M., 2006. A behavioral systems approach to romantic love relationships: attachment, caregiving and sex. In: Sternberg, R., Weis, K. (Eds.), The New Psychology of Love. Yale University Press, New Haven & London, pp. 35–64.

Snyder, M., Tanke, E.D., Berscheid, E., 1977. Social perception and interpersonal behavior: on the self-fulfilling nature of social stereotypes. Journal of Personality and Social Psychology 35 (9), 656–666.

Sprecher, S., Fehr, B., 2005. Compassionate love for close others and humanity. Journal of Social and Personal Relationships 22 (5), 629–651.

Sternberg, R., 1995. Love as a story. Journal of Social and Personal Relationships 12 (4), 541–546.

Sternberg, R., 2004. A triangular theory of love. In: Reis, H., Rusbult, C.E. (Eds.), Close Relationships. Psychology Press, New York, pp. 213–228.

Sternberg, R., 2006. A duplex theory of love. In: Sternberg, R., Weis, K. (Eds.), The New Psychology of Love. Yale University Press, New Haven & London, pp. 184–199.

Relevant Websites

http://www.iarr.org/ – The International Association for Relationship Research.

http://www.scienceofrelationships.com/ – The Science of Relationships Site.

http://www.sexscience.org/ – The Society for the Scientific Study of Sexuality.

http://kinseyconfidential.org/ – The blog of the Kinsey Institute on Sex, Love and Relationships.

http://www.kinseyinstitute.org/ – The Kinsey Institute for Research in Gender, Sex and Reproduction.

Luhmann, Niklas (1927–98)

Rudolf Stichweh, Forum Internationale Wissenschaft, University of Bonn, Bonn, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Niklas Luhmann (1927–98) was a German sociologist who continued and renewed the tradition of sociological systems theory started by Talcott Parsons. Niklas Luhmann was impressive in adding new insights and concepts from phenomenology, medieval scholasticism, biology, cybernetics, social psychology, legal dogmatics, and other disciplines. Luhmann's theory is best characterized in comparing it to the major achievements of Talcott Parsons: An open, inductive list of global function systems supplants the deductive, binary logic of AGIL; 'World Society' supersedes the solidarity-based national communities postulated by Parsons; neo-Darwinist evolutionary theory pushes away the Parsonian thinking in developmental trends; communication takes the place of action, and generalized media of exchange become media of communication, which do not mediate exchanges between systems but bring about the core of autopoietically closed communication systems; a loose, hierarchical arrangement of interdisciplinary borrowings substitutes for the hierarchical conceptual structure of the Parsonian paradigm.

Biography

Niklas Luhmann was born on 8 December 1927, in Lüneburg (Hanover), as the son of a brewer; his mother came from the Bernese Oberland. Luhmann's family seems to have been permissive toward the children and distant from National Socialism. Luhmann attended a humanist Gymnasium at Lüneburg where he was an assiduous student. In spring 1943, Luhmann's whole class became helpers of the German flak at airports nearby. School hours continued irregularly at the location, and in autumn 1944, he had to leave school and become a regular soldier in South Germany. In spring 1945, the American army took him as a prisoner of war and transported him first to Ludwigshafen and then to a labor camp near Marseille.

As Luhmann still was not yet 18, he was released from the camp in autumn 1945. Since his secondary school degree was not accepted for academic study, he went back to Lüneburg and took a special class, which led to the 'Abitur.' He decided to study law, which was motivated by his supposition that law is the kind of knowledge system that can help with the breakdowns of order he had experienced. From 1946 to 1949, he was a law student at Freiburg with a strong interest in Roman law and in historical and comparative aspects of law. Luhmann went back to Lüneburg, became a trainee lawyer with a legal practitioner and prepared a legal dissertation, which was never finalized. He took his second state examination in 1953 and found his first job in 1954.

Having first worked with a private legal practitioner, Luhmann acquired a certain dislike toward dependence on clients. Therefore, he preferred public service and had his first job as a lawyer at a higher administrative court in Lüneburg (1954–55), then switching to the ministry of culture in the federal state Lower-Saxony in Hanover (1955–62). In these years, he slowly, through private study, became a sociologist. His first two papers were published in 1958 and 1960 in a journal for the sciences of administration. Then Luhmann applied for and received a stipend at the 'Littauer Center for Public Administration' (Harvard) where he studied primarily at the 'Department of Social Relations' with Talcott Parsons (1960–61).

Returning to Germany, Luhmann took a job at the research institute of the University for Public Administration at Speyer in 1962. In the same year, he published his classical essay on 'Function and Causality' in the 'Kölner Zeitschrift für Soziologie und Sozialpsychologie.' In 1964, Helmuth Schelsky, at this time probably the most influential German sociologist, asked him if he wanted to become a professor of sociology at the planned University of Bielefeld. Luhmann accepted and finished his 'Doctor of Philosophy' and his 'Habilitation' in 1966, taught in Münster and, in 1968, in Frankfurt (as a substitute for Theodor W. Adorno). Finally, in 1968, he was appointed as the first professor at the University of Bielefeld, 2 years before the first student inscriptions there. He stayed in Bielefeld until his retirement in 1993. Luhmann liked traveling, but otherwise he preferred a rather uneventful life of which he made use in producing the most impressive publication record of twentieth-century sociology. He lived near Bielefeld with his wife who died early (1977) and their three children. Only a few years after his retirement, Luhmann contracted a severe illness from which he died on 6 November 1998.

The Intellectual Development

We do not know much about early intellectual influences but a relevant circumstance is that Luhmann was a private scholar between the end of his law studies (1949) and his stipend at Harvard (1960–61). The intellectual education of Luhmann probably was much broader than usually is the case for young scientists who try to start a disciplinary career. From brief forays, we know about some of Luhmann's early readings: Camus, Dostoevsky, Hölderlin, Jean Paul, Thomas Mann; in these years, he had no reason to become a disciplinary specialist.

In reconstructing this intellectual education, there are two who had lasting significance for the genesis and structure of Niklas Luhmann's theories: Edmund Husserl (1859–1938) and Talcott Parsons (1902–79). Both of them, as soon as Luhmann got to know their writings, induced a cognitive shift via key

concepts deeply embedded in the structures of the theories Luhmann built. Luhmann, who did not give much weight in intellectual reconstructions to persons, made exceptions in his lectures for Husserl and Parsons. From this dual influence arose the synthesis of phenomenology and systems theory, which is historically unique: from Husserl, Luhmann took the strict separation of psychic systems (consciousness) and social systems (communication), which he radicalized in a way nobody had done before. He further adapted from Husserl the core mechanism used on both sides of the gulf between psychic and social systems: Meaning as a mode of selectivity, which builds complexity by remembering even those possibilities that were not chosen. Such a system built on meaning as its way of dealing with selections will incessantly oscillate between references toward its own states and references toward things external to it, a distinction that is akin to the Husserlian concept of intentionality.

From these examples, we can infer the originality of the strategy. On the one hand, the distinction psychic/social is radicalized, which implies a negation of intersubjectivity, as subjectivity is a concept suitable only for psychic systems. On the other hand – as meaning is used in social systems, too – the rich vocabulary of the European philosophy of consciousness becomes instructive for the understanding of mechanisms and structures of social systems.

The concept of social system is taken from Talcott Parsons and this process of taking stock of Parsons and reintegrating his conceptual structures into a completely reformulated systems theory is a still more important cognitive undertaking than the interdisciplinary discourse with Husserl. The logic in Luhmann's way of dealing with Parsons consists in making use of nearly everything Parsons invented, and in doing this to recontextualize every concept, a strategy that maximizes as well integrative continuity as it favors building completely new conceptual structures.

There are many more authors and thinkers who have to be included as important parts of the intellectual context of Luhmann's writings: Herbert Simon's influence on Luhmann's organization theory and Luhmann's intuitions on bounded rationality. There is an early and extensive reception of many authors of General Systems Theory, among others Ludwig von Bertalanffy and Ross Ashby. Others who contributed significant insights included: Kenneth Burke, Gaston Bachelard, Robert K. Merton, Donald T. Campbell, Reinhart Koselleck, Erving Goffman, Humberto Maturana, Henri Atlan, Gregory Bateson, George Spencer Brown, and Franz Heider.

There are patterns in his ways of absorbing influences. Luhmann had encyclopedic reading interests. His main dislike applied to authors who were stronger in normative than in cognitive arguments. He did not prefer establishment figures. Even in disciplines not his own, he was willing to be inspired by outsiders. What in legal discourse is called the 'prevailing opinion' did not impress him; he took nearly forgotten authors from obscurity and gave them a central role in theory building (cf the paper on Vauvenargues, Ch. 7 in SA3). Furthermore, Luhmann had an uncanny talent for finding just the right interpretation for strange ideas from other disciplines. Though there were numerous social scientists who experimented with Maturana's concept of 'autopoiesis,' only Luhmann's variant survives today.

Theory of Social Systems

Luhmann called his contribution to sociological theorizing 'Systems Theory' (early papers in SA3), formulating the continuity to General Systems Theory and to Talcott Parsons. Systems theory is still one of the most influential paradigms of sociological thinking and research, and it is held together by the fact that the concept of social system functions as *the core concept*, which is in itself the object of incessant reformulation and interdisciplinary renewal.

Function and Causality

There was one alternative self-designation Luhmann used in the sixties. His 'functional–structural theory' was meant as an inversion and alternative to the Parsonian 'structural–functional theory.' By this, Luhmann indicated that his theorizing does not start with given social structures that are subsequently analyzed in their functionality (a *modus operandi* he attributed to Parsons). Instead, the sociologist is supposed to begin with social problems, which are understood as functional references – e.g., How to ensure future need fulfillment by present action? How to ensure interpersonal consistency in experiencing the world? Real social structures are then analyzed in their capability to contribute to the resolution of these problems. What becomes visible here is Luhmann's preference for a historical and comparative functionalism, which always compares alternative structural patterns in their ability to contribute to the solution of relevant social problems. This is near to Darwinian evolutionary biology or to certain types of evolutionary economic institutionalism (the Veblenian tradition) and articulates a preference for comparative studies against a conventional preference for the causal reduction of observed events. The methodology of this kind of equivalence functionalism was the subject of Luhmann's first extended sociological essay (SA1: Ch. 1), and he always remained true to this comparative and evolutionary interpretation of the methodological tendencies of sociological systems theory, although he did not write much about functionalism in later years.

System and Environment

As soon as the differentiation from Parsons lost its symbolic relevance, Luhmann rarely used 'functional–structural theory' as a self-description of his theory, instead using 'systems theory.' In the early years, 'cybernetic' was sometimes added by Luhmann to 'systems theory' (SA1, p. 132, n. 16) to mean selectivity of the relations of the system to its environments. There are three interesting implications in this definition. First, selectivity becomes a universal attribute of any event, which ever happens in a system. Second, the environment becomes relevant as a circumstance, which impacts on any selection event in the system. And, third, in choosing *its* selection events and observing *its* environments the system acquires a self-referential character. Therefore, Luhmann's systems theory is emphatically a system–environment theory specified by its cybernetic (i.e., self-referential) character. From the relevance of the concept of 'environment' follows one more methodological postulate. Systems theory not only needs to be comparative in cognitive operations; it must also, in observing alternative

strategies and trajectories of social systems, explain these operations on the basis of the observation of its environments by the system.

Another central term for the analysis of a system and its environments is complexity. A system consists of certain elements and realizes in a selective way relations among them. Luhmann called this property of a system 'complexity.' The complexity of a system seems to be related to the demands its environments place on it.

Meaning and Social Systems

Up to this point, we have not specified which kind of system we are speaking about. In the case of Luhmann, the primacy of social systems is obvious. Though Luhmann had broad interests in cybernetics, biological theory and numerous other disciplines, he left no doubt that he only intended to contribute to the theory of social systems.

What is the basis of the specificity of social systems? This is 'meaning' (Ch. 2 in Habermas and Luhmann, 1971), a special case of a theorem in General Systems Theory, which says that in any system one observes a production of surplus possibilities and of mechanisms reducing these surpluses. Meaning is that way of dealing with surpluses in which the possibilities not chosen are not eliminated, but are remembered and virtualized and thereby stored for future use. Meaning systems that consist of a mix of realities (realized possibilities) and virtualities (as yet unrealized possibilities) need more sophisticated mechanisms for dealing with the kind of complexity they produce.

Social Systems and Psychic Systems

Meaning allows us to distinguish social systems from biological systems, physiochemical systems, and machines, all of which are not able to produce and to process meaning. But there is one further type of systems – psychic systems – for which meaning is constitutive of its operations.

As was the case in Parsons, social and psychic systems are conceived by Luhmann as two different types of systems, separate from one another but coupled via media such as meaning and language. In Parsons, the distinction of social and psychic systems was introduced via the 'action frame of reference' and they represent two of the four types of action systems, which contribute to the emergence of action as a phenomenon constitutive of the human condition. In Luhmann, we have a different constellation. All the hierarchical levels we have in Parsons disappear in the strictly nonhierarchical theory of Luhmann. Instead, we have a clear disjunction of social and psychic systems, both of them being environments for the other type of system. This way it is more explicit that persons and their psychic systems are *only* environments of social systems (and social systems are *only* environments of psychic systems). This diagnosis was perceived as antihumanist and therefore aroused numerous controversies. Luhmann liked to turn these arguments around and to insist that for a person the autonomy from the structures of social systems is an important freedom.

But how is this strict separation of two system types, both of which operate on the basis of 'meaning' and which are connected by language, to be explained? Luhmann developed a theory that the emergence of a system happens by the self-specification of the elements, which are constitutive of the system. The differentiation of social and psychic systems is then explained by the differentiation of elements characteristic of these two systems.

In looking at psychic systems, Luhmann developed a conception akin to Husserl's. Psychic systems consist of thoughts as their elementary basis. Thoughts are connected with one another, referring to earlier thoughts and preparing ongoing considerations (Luhmann SA6: Ch. 1–4). Conceived this way, Luhmann calls psychic systems 'systems of consciousness.' From this, the implication is that there is no systematic place for a concept of the unconscious, except in postulating an observer who ascribes latencies to a psychic system, which are unobservable for the system.

Action and Experience

What are the elementary constituents of social systems? In a classical sociological understanding, one probably would have opted for 'actions,' e.g., the 'unit acts' of Talcott Parsons, and for some years Luhmann described the basic social elements in this way, making use of terms such as 'communicative action.' At the same time, he complicated the understanding by introducing a distinction between 'action' and 'experience' (Luhmann SA3: Ch. 5). Social systems are postulated to process selections for which there exist two alternatives: They are either causally attributed to one of the social systems, or alternatively are seen as representing objective circumstances in the world, which one is unable to influence in the present situation. This distinction of action and experience is no ontological distinction, which identifies ontological properties of the selection events. It is only based on attributions produced by the social systems that can be disputed and reversed. This argument demonstrates how much social systems exist in a social world entirely produced by themselves and how much objectivity (information about states of the world) and subjectivity (selections for which actors can be held responsible) can reverse their roles. But what does this say about the constitutive element of social systems? We cannot point to 'selection' since selection is a general phenomenon at the basis of natural as well as social systems. And there is a clear argument against 'action,' since 'action' is one part of the distinction of action and experience, both sides of which are in the realm of the social. Therefore, we have to look for another concept, which allows us to identify the boundaries of the social domain.

Communication and Action

Until the late 1970s, Luhmann sometimes said that he hesitated whether he should designate 'actions' or 'communications' as the elementary constituents of social systems. One could perceive a rhetorical aspect in these remarks as Luhmann had already established the thesis that in social systems only some selections are attributed to a social system as its actions and therefore the concept of social action could not claim the universal status needed for *the* constitutive element of a social system.

The other candidate for elemental status was obviously communication, a concept until that point sociologists mainly used casually. Parsons never formulated a theory of communication, though since the information theory of the 1940s communication was a probable candidate for a general sociological theory. To understand the fundamental character of information transfer in social processes, social action would have been a counterintuitive choice. Already in 1951, there was a book by Jürgen Ruesch and Gregory Bateson *Communication: The Social Matrix of Psychiatry*, which gave a good idea of how to base a social science discipline on the concept of communication.

For these reasons Luhmann, in the theoretical treatise he finally published in 1984, resolutely established his theory of social systems as a theory of communication systems (Luhmann, 1984: Ch. 4). Communication is the foundational element of social systems and as such it even constitutes the boundaries of social systems. Social systems only consist of communications and there is no communication outside of social systems, which means there is a sharp boundary separating social systems from those environments that do not consist of communications.

Luhmann regards communication as based on three selections, which are indispensable components and are necessarily intertwined: 'information,' 'utterance,' and 'understanding.' Information can be interpreted in the way Bateson proposed: an information is "a difference which makes a difference," a conception requiring a kind of minimum sociality. There must be an entity at which the first difference occurs and a second entity, which registers the first difference and attributes informative relevance to it. The first and the second entities can be identical (I observe the change of my bodily states and ascribe informative relevance to them), but even here exists an internal division, which establishes some 'internal sociality.' Information in this interpretation is not yet communication, but there is subjectivity involved that infers the informative relevance of the differences observed. For communication to arise there must be further components in the communication process. There always has to be a system that explicitly and intentionally decides to *utter* the informative difference. Besides these intentional utterances, we can concede the possibility of nonintentional utterances. For example, I may change my appearance in a way, which is perceived as a kind of utterance that informs others about changes of my mind. This utterance component in a communication process may be called the action component, since there always will be an attribution that classifies the utterance as the action of a specific system. Information and utterance still do not suffice to produce an elementary communication. To realize a communication one needs *a second system, which understands the information uttered by the first system*. Luhmann's concept of 'understanding' is a rather formal one, which does not demand that 'understanding' is a correct or good understanding. It includes the possibility of 'misunderstanding' as a case of understanding. Even if I grin in listening to sad news, this counts as an indicator for understanding to other participants and probably points to the acceptance or rejection of the news (e.g., I do not believe the news or it is good news for me). My acceptance/rejection has to be seen as a fourth component of communication, which already is part of the next communication.

Communication Theory and Double Contingency

Luhmann's three-component theory of communication is related to theories proposed by Karl Bühler and John Searle. It is important to bring to mind the most important sociological understandings assumed by Luhmann's theory: (1) Communication is not dependent on the intention to communicate, utterances can be unintentional. There are always two systems (processors) involved; for communication the second system observes a difference of information and utterance. (2) Communication's medium can be language or nonverbal communication, thus what can be done with it obviously differs. (3) For the competition between 'action' and 'communication,' Luhmann found an elegant solution. The primacy of communication is obvious. But among the components of elementary communications, there is one – the utterance – which functions as the action component. This allows a more general understanding regarding the concept of action as constituted by attributions. If for a selection one looks for someone whose responsibility one wants to claim, one will attribute this selection to this system as its action. (4) As we already saw, understanding as the third component can immediately pass into the fourth component acceptance/rejection, which is already part of the next communication. This argument makes visible communication as a flow and the recursivity built into this flow, that is, communications coming back to or referring to earlier communications. (5) A precondition of any communication is that at least two systems or processors (or 'alter' and 'ego' in the terminology from phenomenology) participate in it. Parsons as well as Luhmann thematize this condition in terms of a theory of 'double contingency,' which Parsons invented in *Toward a General Theory of Action* (1951). Double Contingency means the paradoxical reciprocity of both systems being oriented in their actions and expectations toward what the other system is probably going to do, and the uncertainty arising from this situation (Luhmann, 1984: Ch. 3). Theories of double contingency have to demonstrate – and both Parsons and Luhmann tried to solve this problem – how communication, order, and system formation happen to arise in a situation in which, at the beginning, a reciprocal blockade seems the most probable outcome.

Communication Elements as Events and Operations

Are there plausible arguments against the status of communications as elementary constituents of social systems? One should not adduce the complex character of communications as an objection since unit acts or atoms as elements also have a complex internal structure. But as elements, communications do not seem to possess the internal (temporal) stability, which one might demand as a condition for element status. Some years before he adopted communication theory, Luhmann was already beginning work on the specific temporality of the elements of social systems. In one of his most fascinating essays, *Time and Action – A Forgotten Theory* (1979), Luhmann proposed a solution he attributed

to the French moralist and enlightenment philosopher Vauvenargues (SA3: Ch. 7).

The elements of social systems are 'events.' An event is something that has only vanishing duration. As soon as it appears it is already gone, thus in any social system there is the imperative to incessantly produce new events, which connect to earlier events that are just vanishing, otherwise the system might come to an end. This is a remarkable interpretation considering the equilibrium and stability postulates of social theory in the 1950s.

Luhmann adds one more concept. This is the 'reproduction of event-based elements.' Reproduction is not identical reproduction but new event-based elements have to relate to earlier elements in the system and their constructive freedom is restricted by the history of the system. For this reproduction of event-based elements, Luhmann proposes the name 'operation.'

This operative reformulation of sociological theory necessitates significant changes in systems theory. For example, one can no longer speak about complexity in the way Luhmann introduced this concept in the late-1960s: Complexity as selectivity in the relations between the elements of a system. Now the same structure arises in each individual operation of a system, which means that already on the level of the elements the demands are fulfilled, which in earlier systems theory were seen as macroproperties of whole systems. This is clearly said in *The Society of Society*:

> In principle the classical concept of complexity is sabotaged by the concept of operation as it transforms the distinction of element and relation into one unitary concept (Operation = selective relationing of elements).
>
> 1997: 139, n. 181

Self-Reference, Autopoiesis, Operational Closure

This changeover to an operative understanding of social systems requires a number of far-reaching rearrangements. The first, going back two decades to his idea of cybernetic systems theory, is the increasing importance of the concept of 'self-reference.' From environmental perturbations the system selects those which it needs to stabilize or optimize its *Eigenstates*; self-referential control is also valid for the reproduction of the event-based elements of the system. This immediately connects to the concept of 'autopoiesis,' which Luhmann borrowed from Humberto Maturana and Francisco Varela. Autopoiesis means that everything that functions as a unity in a system – element, operation, structure, boundary – is due to the production processes of the system itself. It follows that on this level of the production of the unities constitutive of a system, no external elements can be imported and the system has to be conceived as an operationally closed system.

Operation and Observation

A last important addition regards the concept of operation. We take from the theory of meaning and from information theory the insight that operations are always based on differences. Operations choose a certain option and discard others always based on a distinction. In social systems or psychic systems, it is often such that both sides of the distinction are represented (options not chosen at least being remembered). An operation for which this is true, Luhmann calls an observation. Thus one can say that each observation is an operation as it depends on its operative performance. On the other hand, in social systems, there are many operations that one could not call an observation as they simply happen without the differences on which they are based being represented.

Like theorists from Harold Garfinkel to Anthony Giddens, Luhmann postulates that modern social systems are based in acts of observation and self-observation. Luhmann gives a more precise version using a term introduced by Heinz von Foerster: Observation in modern social systems, especially in function systems, is always 'second-order observation.' Observers do not have direct access to reality; they observe other observers and using these observations decide with which observations to connect, and from which to dissociate. Modern science that always observes on the basis of observations published by other scientists is a good example of this indirectness in the function systems of modern society.

Theory of Society

The theory of social systems presented above is, in Luhmann's view, the most general social theory sociology can produce. It has a formal similarity to the 'action frame of reference' in the Parsonian *oeuvre*. Luhmann planned this theory as one of two core projects, the other was to conceive a theory of society, society being understood as the most extensive social system including all communications and actions.

Interaction, Organization, Society

To understand the theory of society one has to know that among the multitude of social systems, there are three types on three *levels* of system formation to which he ascribes special prominence. Luhmann calls them 'interaction,' 'organization,' and 'society.'

Luhmann's theory of interaction is very close to Erving Goffman's theory of the interaction order (SA2: Ch. 2). Interaction systems only demand physical presence of the participants and reciprocity of perception. Interaction systems are limited, they can only process one subject matter at a time and when the copresence of participants ends, the interaction systems also end. But they are common, they are repetitive, and they crystallize social forms. Therefore, they account for a significant part of our everyday experience.

Organizations or *formal organizations* define an intermediary level of system formation (Luhmann, 1964, 2000). They are membership organizations based on formalized conditions for entry and exit. As a member, one is bound to the rules of an organization. As autopoietic systems, organizations are operationally closed on the level of decisions. At the end, everything that happens in an organization comes from a decision (e.g., decisions on membership) or leads to decisions and these decisions have to be attributed to the organization itself.

Society is the most extensive social system, including all interactions and organizations in its purview. Luhmann defines society via communicative attainability and derives the

conclusion that today there is only one societal system on earth, which he calls 'World Society.' That was a provocative thesis, first published in 1971 (SA2: Ch. 4), which for some time – until the onset of globalization theory – found neither much notice nor much disagreement.

What does the theory of society Luhmann finally published in the year before his death (1997) look like? Over decades Luhmann developed the theory in four big threads. In the youngest of these threads, society is understood as a system incessantly producing self-observations and self-descriptions (Luhmann GS 1–4; 1982; 1997: Ch. 5). The other three threads are closely related to Luhmann's theory of meaning (Habermas and Luhmann, 1971: Ch. 2) and distinguish a social, a temporal, and a material dimension of meaning. These three dimensions are connected to the theory of symbolically generalized media of communication (social dimension), the theory of sociocultural evolution (temporal dimension), and differentiation theory (material dimension).

Theory of Symbolically Generalized Media of Communication

Regarding the social dimension of meaning, the most important problem seems to be how it is possible to bridge differences of opinion and of interest among participants and achieve the acceptance of unexpected selections by participants. Language, especially language that is made use of in an effective way, is a medium of communication to achieve this. Since Greek and Roman antiquity language was supported by rhetoric as an art, that is, as a technique for improving the effectiveness of those intentions which one pursued, the same social function (effectiveness in communication) was reintroduced into modern society by symbolically generalized media of communication (Luhmann, 1997: Ch. 2).

In contrast to Parsons who ascribed to exchange media, a special function for mediating input–output processes between systems, Luhmann analyzes symbolically generalized communication media as mechanisms internal to systems. Such mechanisms arise primarily in situations where you need additional means of motivation for especially improbable demands. Why, for example, should I accept that another person buys goods, which I would like to have myself? Luhmann proposed the symbolically generalized medium of money. The buyer of much sought-after goods has to hand over a significant amount of money, which symbolically signals to all other owners of money that their money gives them, too, the freedom to buy because as a general medium money can be exchanged against any goods. In a similar way, but with changing motivational constellations, Luhmann analyzes the other media: *Love*: There is someone who completely realigns his actions with the way I experience the world and this motivates me to reciprocate his/her love. *Power*: I am motivated by the negative sanctions a power-holder controls and which I want to evade. *Truth:* I am motivated by reliable ways of experiencing the world promised by science and of which I know that all rational others have to accept them. This is the basic argumentative figure, which is supplemented in Luhmann by a long catalog of media properties: *binary code structures in media* (beautiful/ugly, true/wrong, lawful/unlawful); *probabilities of inflation and deflation of media symbols; mechanisms which connect the media with body perceptions of participants* (sexuality, physical force, physical needs) (Luhmann, 1975). It is easily perceived that this theory, handled as a comparative theory of different media constellations, has its analytical value in these possibilities of systematically comparing types of effective communications.

Theory of Sociocultural Evolution

The second part of the theory of society looks at *changes of social structures in long-term perspective*. Instead of 'modernization,' 'social change,' or 'development' as relevant theoretical paradigms (1997: Ch.3), he pleaded for a neo-Darwinist evolutionary theory. This was a surprising choice as 'evolution' was then a discredited position in social theory. Donald T. Campbell, an American psychologist and methodologist, was one of the few persistently trying to revivify evolutionary interpretations of social science problems. For Luhmann's approach two factors were important. First, he was fascinated that chance events can become productive for the formation of social structures. He perceived this as the conceptual core of evolutionary theories and it is well adapted to other aspects of his theorizing. Second, Luhmann saw the main task to be in a theory of evolutionary mechanisms. The catalog of evolutionary mechanisms – variation, selection, stabilization/retention – and the conviction of their conceptual centrality Luhmann took from Donald Campbell. His main questions aimed at the identification of these three mechanisms in different social systems.

On the macrolevel, Luhmann locates the *variation mechanism* in the possibility to say *no* which is given in any communicative situation and which if used always implies the possibility of changed impulses in a system. The *selection mechanism* consists in the binary codes of the media of communication to which he ascribes the function to sort new meaning elements arising on the basis of negations. The *stabilization of selected meaning complexes* happens via system formation, that is, by the differentiation of a (new) system, which specializes on the material dimensions of meaning becoming prominent in the selection processes. The stabilization of a system has the effect that it becomes easier to see where further possibilities of negation and change are located in the reorganized system. This retroactive effect from stabilization to variation demonstrates a circular connectedness of the three evolutionary mechanisms. In looking at this theoretical construction, one perceives an internal representation of the three threads of the theory of society. There is one mechanism that stands for the probability of chance, a second mechanism focused on selective acceptance of improbable meaning demands, and a third mechanism that reorganizes the material complexity of the world by differentiation processes.

Differentiation Theory

Looking at differentiation theory, the third prominent part of the theory of society, we are near the core of the sociological tradition (1997: Ch.4). The main innovation Luhmann added was interpreting differentiation as a theory of *system formation*. Differentiation processes always produce new system–environment differences. This differs from the binary- and

AGIL-logic of the Parsonian paradigm with Luhmann substituting the idea that for the formation of systems there exist only a limited number of forms. The problem of *changing forms of systems differentiation* becomes the conceptual core of his differentiation theory.

There is first of all *segmentary differentiation*. Segments can be identified by the sameness of their internal structures and by the equality of their societal importance and rank (kinship groups in a tribal society are an example). As Durkheim postulated, in a segmentary society one can add or take away segments without effecting significant changes in the structure of society.

A second form that Luhmann added on the basis of historical evidence is the *differentiation of center and periphery*. A classical illustration is the differentiation of city and country, which postulates systematic and institutionalized differences in the control of resources and information. However, the place of this form of differentiation in the systematics of differentiation theory is not entirely clear since it would be difficult to argue that center/periphery brings about new types of social systems. It seems more convincing that center/periphery is either a distinction internal to a system or a form of interconnectedness among systems that remain separate otherwise. Therefore center/periphery is probably not a differentiation form of its own and *mainly organizes historical transfers between forms of differentiation*.

The third form of differentiation is called *hierarchical differentiation or stratification*. It divides society into social systems (estates, castes, strata), which are distinguished by inequalities of rank. Each of these systems defines for its members a total life form, which does not need other contexts of participation even though there are occasional contacts with other strata. All the traditional high cultures of human history were stratified societies, which demonstrate the historical centrality of this differentiation form.

Functional differentiation is, for Luhmann, the differentiation form of a modern world society, which consists of macrosystems for economic relations, politics, law, religion, science, and numerous other functional contexts. All of these are global macrosystems of worldwide communicative extension. Among one another, they are characterized by extreme social and cultural diversity, which excludes any possibility of a rank order among them. What distinguishes them from segmentary structures is that all these function systems are indispensable. Luhmann called them *self-substitutive orders* where a deficit of any functional resource (e.g., money, power) cannot be substituted for by resources from another function system. As the hypothesis of the 'functional differentiation of world society' is the central empirical hypothesis of Luhmann's theory of society, the future impact of his theory will be dependent on the influence of this hypothesis on global sociological theorizing and research.

Profile of Theory Building and Relevance for Future Social Science

One of the most conspicuous features of the writings of Niklas Luhmann is a strong belief in the validity and cognitive strength of theoretical thinking. Theory for him primarily meant conceptual work, and in this he had close affinities to the cognitive style of Roman law and of philosophy. Luhmann never would have considered himself a philosopher, but for him the philosophical tradition was a treasure trove of conceptual resources, which sociology could neglect only at considerable cost. One example of the value of his conceptual focus is his transformation of the blunt Parsonian concept of contingency from simply dependence into the scholastic concept defined by the double negation of chance and necessity. This way he created a more flexible conceptual instrument.

Interdisciplinarity was, in Luhmann's case, a logical corollary of his stance toward the philosophical tradition. In the development of his theories, he continuously imported concepts from social psychology, cybernetics, linguistics, biology, and logic. And the conceptual structures and theories he imported were then completely sociologized. A remarkable case is *autopoiesis*, a biological concept of circular production processes of biological macromolecules, which Luhmann transferred to the communication-based function systems of modern society. This is a controversial operation but on one point, there cannot be any dispute: there are not even traces of materialism or biologism in the sociological theory resulting from the transfer.

Marxism was an important context in the development of Luhmann's theories. When Luhmann's scientific work became visible in the late-1960s, the intellectual scenery in Germany was dominated by the student movement and its intellectual accompaniments. Luhmann became famous in 1971 when he published a book with Habermas, the main representative of the Frankfurt school (Habermas and Luhmann, 1971). In the following years, the career of systems theory in Germany was coupled with the demise of Marxism. After the mundane and middle-range sociology of the 1950s and 1960s, the intellectual opposition of the late 1960s had reestablished the expectation that sociology should deliver complex theories of society. When Marxism and political economy failed, systems theory was the only alternative remaining. Luhmann had no sympathy for the intellectual and political claims of Marxism, but systems theory was perfectly adapted to an expectation structure looking for macrotheories of society and their universalizing claims.

The structural functionalism of Talcott Parsons was the theory that Niklas Luhmann started from. This was an unusual intellectual relation: On the one hand, Luhmann reconstructed point to point, often taking up minor theoretical distinctions hidden in appendices of the essays on generalized media of exchange. On the other hand, the theory Luhmann built via this reconstructive effort is completely different from the Parsonian undertaking: An open, inductive list of function systems supplanting the deductive logic of AGIL; 'World Society' superseding the solidarity-based national communities Parsons postulated; neo-Darwinist evolutionary theory pushing away the Parsonian thinking in developmental trends; a loose, heterarchical arrangement of interdisciplinary borrowings substituting for the hierarchical conceptual structure of the Parsonian paradigm. When today's systems theory is one of the few elaborated and universalistic paradigms of sociological thinking, which probably will have an influence far into the twenty-first century, this is due to

the ever changing theory-building and theory-reconstructing venture advanced by Niklas Luhmann.

Systems theory such as Luhmann took over from Parsons and the cybernetic tradition, by giving it a new twist as a nonhierarchical theoretical landscape in a nonhierarchical society, stands out among sociological paradigms because of its sociological universalism, not being tied to specific problem situations but instead being applicable to the whole range of sociological problems from minuscule microcontexts to world society, and from long-term historical interpretations of the evolution of human society to the most recent developments in communication practices and technologies. This universalism is not achieved via a rigorous simplification of the conceptual apparatus (as may be the case in rational choice sociology). Instead, it is achieved by an inclusive, interdisciplinary, and historically informed network of sociological theories, which are only loosely coupled under the umbrella of systems theory. This is probably the main attractor of present-day systems theory.

See also: Culture, Sociology of; Cybernetics; Functionalism, History of; Luhmann's Social Systems: Meaning, Autopoiesis, and Interpenetration; Parsons, Talcott (1902–79); Phenomenology in Sociology; Social System; Sociocybernetics; Sociological Theory; Sociology: Overview.

Bibliography

Habermas, Jürgen, Luhmann, Niklas, 1971. Theorie der Gesellschaft oder Sozialtechnologie. Suhrkamp, Frankfurt a.M.

Luhmann, Niklas, 1964. Funktionen und Folgen formaler Organisationen. Duncker & Humblot, Berlin.

Luhmann, Niklas, 1970/1975/1981/1987/1990/1995. Soziologische Aufklärung, vol. 1–6. Westdeutscher Verlag, Opladen (quoted as SA1 – SA6).

Luhmann, Niklas, 1975. Macht. Enke, Stuttgart.

Luhmann, Niklas, 1980/1981/1989/1995. Gesellschaftsstruktur und Semantik. Studien zur Wissenssoziologie der modernen Gesellschaft, vol. 1–4. Suhrkamp, Frankfurt a.M (quoted as GS1 – GS4).

Luhmann, Niklas, 1982. Liebe als Passion. Zur Codierung von Intimität. Suhrkamp, Frankfurt a.M.

Luhmann, Niklas, 1984. Soziale Systeme. Suhrkamp, Frankfurt a.M.

Luhmann, Niklas, 1993. Das Recht der Gesellschaft. Suhrkamp, Frankfurt a.M.

Luhmann, Niklas, 1997. Die Gesellschaft der Gesellschaft, vol. 1–2. Suhrkamp, Frankfurt a.M.

Luhmann, Niklas, 2000. Organisation und Entscheidung. Westdeutscher Verlag, Opladen.

Luhmann's *Social Systems*: Meaning, Autopoiesis, and Interpenetration

Kenneth C Bausch, Institute for 21st Century Agoras, Cincinnati, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The problem of double contingency and its accompanying parable of the black boxes informs Luhmann's conception of meaning and frequently provides illustration at critical junctures. In life, (we) 'psychic systems' and (societies) 'social systems' are constantly faced with situations that require choices. Our *meanings* develop from those choices. Each choice that we make is an element of our meaning. *Autopoiesis* forms the background of Luhmann's theory. Psychic and social autopoietic systems live by constantly maintaining their reproduction as a closed system. At the same time, they constantly interact with their environment by incorporating elements from it and releasing unneeded elements back into it. In this process, what remains the same is the reproductive process, which incorporates those elements that foster its life and evolution. Interpenetration describes how closed autopoietic systems come to share meaning and come to cooperatre and understand each other.

The Problem of Double Consistency

To a large extent, Luhmann's theory of social systems can be understood as an extended solution to the problem of double contingency.

What happens when two psychic or social systems interact? Luhmann examines this situation by reconsidering Parson's problem of double contingency. He states this paradox as follows: "Action cannot take place if alter makes his action dependent on how ego acts and ego wants to connect his action to alter's" (1995: p. 103). (All references to Luhmann, unless otherwise indicated, are to Luhmann, *Social Systems* (1995)). Double contingency involves two closed self-referential circles with no asymmetry and therefore no communication or common action. This situation "concerns a basic condition of possibility for social action as such. No action can occur without first solving this problem of double contingency" (p. 103).

Luhmann reframes the problem of double contingency in terms of 'black boxes' in order to make it independent of existing personalities and applicable to social systems as well as psychic systems.

> Two black boxes, by whatever accident, come to have dealings with one another. Each determines its own behavior by complex self-referential operations within its own boundaries. What can be seen of each is therefore necessarily a reduction. Each assumes the same about the other. Therefore, however many efforts they exert and however much time they spend (they themselves are always faster!), the black boxes remain opaque to one another.
>
> p. 109

How can these black boxes become white for one another?

When black boxes encounter one another, they are forced to integrate each other into their meaning systems. Because they lack insight into each other, they devise stratagems. Ego attempts to determine alter's behavior by assuming what it will be. Then ego sensitively observes alter's behavior to see what happens. Ego soon discovers that alter's behavior cannot be calculated by ego's expectations. In the mutuality of this discovery, ego and alter grant freedom to one another and to themselves. In addition, ego comes to realize that it is assuming how alter will act, and that alter knows that ego is assuming.

A situation of mutual recognition is established. Ego and alter concentrate on each other's input–output behavior, and learn to see themselves as functioning as the other does. They try to influence each other and learn from what happens. In this process, they create a social system. Luhmann says:

> In this way, an emergent order can arise that *is conditioned* by the complexity of the systems that make it possible *but that does not depend on this complexity's being calculated or controlled*. We call this emergent order a social system.
>
> p. 110

A social system, then, does *not* rely on ego and alter seeing through each other and definitely prognosticating each other. A social system is *not* built upon such basal certainties; it *is* built upon social uncertainties and the mechanisms that ego and alter put in place to regulate their dealings with those uncertainties. In Luhmann's words:

> The social system is a system because there is no basal certainty about states and no prediction of behavior to be built thereon. Only the uncertainties *that result from this* are controlled, and they are controlled only with reference to the participants' *own* behavior. System formation constrains (=structures) the possibilities of *safeguarding one's own behavior* in any such situation … The absorption of uncertainty runs its course by stabilizing expectations not by stabilizing behavior.
>
> p. 110

In the ongoing dynamic set up by the condition of double contingency, ego and alter build up structural values by projecting and testing their expectations. They structure their experiences according to these structural values and thereby create an emergent social reality.

Luhmann uses double contingency as a leitmotif throughout his theory of social systems. Black boxes (autopoietic systems), interacting through crafted expectations and careful scrutiny, make moment-to-moment selections that generate social systems and personal meaning. With this parable, Luhmann clearly illustrates how double contingency provides the requisite uncertainty for communication and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32201-2

evolutionary progress. Luhmann uses his solution of double contingency to build the structures of the social world, including meaning, role, norm, program, and value. The black box parable is part and parcel of Luhmann's theory of autopoiesis.

Meaning

Luhmann begins his theory by establishing a new conception of the idea of 'meaning.' He does *not* base his concept of meaning on the content of a statement, some particular fact in the world, or some 'meaning of life.' Instead, he focuses on how human experience is ordered. "A direct and presupposition-less approach to the problem of meaning is then best sought in a phenomenological description of what actually presents itself in meaningful experience" (1990: p. 25).

He begins with the fundamental insight of Husserl that every experience is an experience of something; that knowledge is *intentional*; that is, every experience intends (points to) something outside of the experiencer.

The basic experience of intentionality contains the following interrelated elements:

- transcendence,
- recognition of actuality and potentiality,
- overabundance of possibilities (complexity),
- a need to reduce this complexity that requires choices about what to notice and what to do (selection),
- the contingency of this selection, and
- the negation of possibilities that are not selected.

Underlying this experience is the assumption that a self-referential subject experiences the difference between its actuality and potentiality; makes contingent selections; and retains its negated choices in memory.

Transcendence is built into our experience because we are always attending to things that are beyond our boundaries. Not only that, we attend to outside things with an awareness that they are outside us. This outside–inside awareness is accompanied by *actual–potential awareness*. We experience what we are at the moment (our actuality) and anticipate what we might become (our potentialities) should we choose one or another of our present options.

To continue living, we have to process the *overabundance of possibilities* that confront us by attending to one and negating the others. By that *selection*, we integrate our actuality with the transcendence of one of our possibilities. We have no guarantees in our selections; we never know if we are actualizing the *right* possibility. "In practice, then, complexity means the necessity of choosing contingently, the necessity of accepting risks" (1990: p. 26).

The *negation* involved in selection has 'functional primacy' (1990: p. 27). When we negate the possibilities that we do not select, we do not eliminate them entirely; we merely bracket them; we keep them as the world background of our experience. Negation makes this bracketing possible because it has a reflexive quality; it can be applied to itself. Therefore, with every selection–negation, we maintain the assurance that our negations will not obliterate everything, and that the world will still be there when we need it.

Provisionally, then, at this point in our discussion, meaning is a selection that simplifies experience and maintains the complexity of the world. It simplifies experience by selecting the objects of experience and action. It maintains the complexity of the world through the reflexivity of negation that brackets the nonattended to world out of our immediate experience, but maintains it as background reality. Because of this reflexivity of negation, meaning emboldens us to face the risks of contingency that would otherwise immobilize us in irreversible life and death decisions. This understanding of meaning fits seamlessly into Luhmann's theory of autopoiesis.

Autopoietic systems straddle the system–environment interface. They define themselves as being in opposition to their environments. They constantly recreate themselves in their elements and processes, while they maintain their separateness from their environment and interact with it constantly. They exist only by creating themselves. If they stop creating themselves, they die.

At the same time, they gather energy, nutrients, and the building blocks of their elements from their environments. They "use the difference between system and environment within themselves for orientation" (1995: p. 9). They are closed systems that maintain their closure at all costs, but they establish that closure by the rules of selection they use in utilizing aspects of their environment.

Psychic and social autopoietic systems reproduce processes of meaning. They make selections from the information that faces them. The reflexivity of negation preserves the nonselected elements of complexity as the enduring horizon of their experience. Their self-referentiality provides them with a way to self-determine their selections, so that they are active creators of their meaning and are not merely altered by outside influences. In their self-referentiality, they process new meaning on the basis of old meanings. They are constantly circulating meanings.

The elemental act of meaning is a selection made by a psychic or societal system as it momentarily reproduces itself. This selection determines the elements and processes it will have till its next reproduction. Over time, psychic and social systems evolve together; they accumulate common successful selections. "This co-evolution has led to a common achievement… We call this evolutionary achievement 'meaning'" (1995: p. 59).

Meaning reduces the risk involved in our experiences and actions by providing redundancy, that is, the assurance that lots of people and lots of our own ideas redundantly indicate the same thing. Through redundancy, our meanings gather psychological and social conviction, and they simplify the process of making new selections. In this manner, meaning does not supply any particular answer or course of action; it supplies only the assurance of redundancy for something or another. Meaning, then, has a reach that is universal but without content.

An autopoietic system is constantly referring back and forth across the system–environment interface as it generates meaning from information. Its closure enables it to selectively appropriate (meaning) elements from the total expanse of complexity that faces it. Without closure, it would have no continuity. With closure, it can generate meanings secured by

redundancy and negatable negation from anything in its complex internal and external environment.

As it generates meanings, an autopoietic system recognizes its own relatively simple and redundant meaning as something it can change. On the basis of its autopoiesis, it separates complexity into inner ('self') and outer (environment). It neglects the overwhelming and variable complexity of its environment except when that complexity intrudes on its meaning as information, that is, as "a difference that makes a difference" (Bateson).

Because of our autopoietic status, we have to choose meanings from moment to moment. As a result, 'meaninglessness' is impossible in experience because a decision to call something 'meaningless' is itself a selection that creates meaning. "Meaning is an unnegatable category, a category devoid of difference [whose negation as] meaninglessness is … possible only in the domain of signs and resides in a confusion of signs" (p. 62).

"Meaning is basally unstable, restless, and with a built-in compulsion to self-alteration" (1995: p. 65). Moment to moment, our meaning orientation has to change to deal with new information. "The instability of meaning resides in [this] untenability of [our] core of actuality" (1995: p. 65). Still we keep our balance because our actions open new possibilities within new horizons.

[Our] ability to restabilize is provided by the fact that everything actual has meaning only within a horizon of possibilities indicated along with [it]. And to have meaning means that one of the possibilities that could be connected up can and must be selected as the next actuality, as soon as what is actual at the moment has faded away, transpired, and given up its actuality out of its own instability … Meaning is the unity of actualization and virtualization, of reactualization and revirtualization, as a self-propelling process (1995: p. 65).

Our choice negates the rest of our options by bracketing them and thereby puts us at risk and in a state of imbalance. We deal with this risk by being extremely sensitive to the new situation that confronts us with some of the old bracketed world and with new possibilities; that is, we adjust to a new horizon. Our actuality is unstable and faced with necessary selection; our potentiality is contingent and maintains stability through its gradual change. We pick our way through rough country by making steps that may advance us on our way and may not. We alertly check our relation to our new horizon with each step as we select our next movement.

Building Social Structures through Communication

In the ongoing dynamic set up by the condition of double contingency, ego and alter build up structural values by projecting and testing their expectations. They structure their experiences according to these structural values and thereby create an emergent social reality. These realities develop as meaning formulations that have a degree of transparency for reciprocal observation and communication.

Some of the basic concepts, so formed, are those of person, intelligence, memory, and learning. These realities cannot be observed within black boxes. They are projected substrates for understanding mutually observed behaviors, substrates reinforced by redundant projections. 'Person' indicates that ego cannot observe how acquaintance with alter provides security about what alter will probably do. In other words, alter acts autonomously but with a logic similar to ego's. 'Intelligence' indicates that ego cannot observe how alter chooses one and not another solution to a problem. 'Memory' indicates that "one cannot observe how one complex, actual state of a system passes over into the next, so that one must fall back instead on selected past inputs" (p. 111). 'Learning' indicates that ego cannot observe how information triggers far-reaching consequences in alter. These examples show that:

> It would be futile to seek a psychic or even organic substrate for such things as person, intelligence, memory, or learning. All this concerns observer stratagems for interpreting what cannot be observed and transferring it to the emergent level of contact between systems.
>
> p. 111

Because ego and alter cannot calculate each other's behavior, they concede freedom to one another and limit themselves to knowledge that helps them handle contingency. With this stratagem, they reduce the complexity of the selections that they face. Because they do not see into their companion's black box, they are "bound to the experience of action," what their companion does, and they are "steered by the concession of freedom" (p. 112).

The experience of contingency relates to environments outside the black boxes and therefore beyond the bounds of intrasystem relations. Ego and alter both experience themselves as having environments. Therefore, they relate their experience to those environments as well as to actions. "The generalized result of constant operation under the condition of double contingency is finally the social dimension of all meaning, namely, that one can ask for any meaning how it is experienced and processed by others" (p. 113). In this way, the improbability of the social order (how could a society be exactly such and so) is explained as being normal.

The problem of double contingency solves itself when ego tells alter, "I will do what you want if you do what I want" (p. 117). This proffered agreement is built on the freedom of both parties: "I do not allow myself to be determined by you, if you do not allow yourself to be determined by me" (p. 117). This agreement is the extremely unstable core structure around which social systems crystallize. The core structure and social system built up on it will collapse, if they are not constantly renewed.

In the experience of double contingency, ego sees alter as an alter ego, that is, as an ego with a different perspective. Ego also realizes that alter sees him as an alter ego. Both ego and alter experience this situation as unstable and unacceptable. "In *this* experience the perspectives converge, and that makes it possible to suppose an interest in negating this negativity, an interest in determination" (p. 122).

This interest in determination leads ego and alter to tackle the problem of communication. It is the factually effective catalytic agent that brings it about.

In their mutual recognition of alter egos, ego and alter reach the understanding that is basic to communication. To overcome their impasse of double contingency, they externalize their

expectations for each other with the action of utterance. The intended content of that utterance and alter's reaction to it constitute information for both ego and alter. Understanding, utterance, and information constitute the essence of communication for Luhmann. Each of these elements is a meaning selection event in the ongoing communications that constitute society. The underlying understanding of mutual recognition, for example, is a contingent selection to treat alter as alter ego. An utterance, such as a smile, is an expectation selectively put forward to express tentative friendliness and to test alter's friendliness. Alter's return smile constitutes information that is either selected or rejected by ego as a return offer of tentative friendliness.

Each of these selections is made on the basis of difference *and* unity. The overall unity is the system–environment in which systems define themselves self-referentially as different from their environment. Information is the selection from difference that enhances a system's unity. Utterance is the selection from a system's store of information and coding. Understanding is a selective action that distinguishes between information and utterance.

The difference *between information and utterance* is crucial for the emergence of communication. In recognizing this difference, ego and alter realize that their utterances are contingent, which they must be conditioned upon commonly accepted meanings. On the basis of this realization, ego and alter can build stores of shared meanings. Luhmann says,

> Communication emerges only if this last difference [between information and utterance] is observed, expected, understood, and used as the basis for connecting with further behaviors. Thus understanding normally includes more or less extensive misunderstandings; but these are always … misunderstandings that can be controlled and corrected.
>
> p. 141

Systems and Environments

An autopoietic system creates the differences between itself and its environment by the selections it makes. When a system makes selections, it differentiates its internal complexity from the greater complexity of its environment. It places greater relevance on internal events/processes and endeavors to arrange them. It considers external events/processes to be rather irrelevant, and largely ignores them. In this way, "the system acquires freedom and the autonomy of self-regulation by indifference to its environment" (p. 183). Thus, the difference between an autopoietic system and its environment is not a metaphysical either/or. It is the result of selections made by a system as it reduces its complexity. As they create boundaries and selectively process their inputs and outputs, systems create themselves as centers of their worlds.

A psychic or social system uses differences to generalize and respecify events; it builds the internal *structure* that steers its internal processes without "point for point correspondences with this relevant environment" (p. 184). It creates its identity in *reflection* by identifying itself as what it can control and what continues to recur, and identifying its environment as everything else. On all three levels, the difference of relative complexity provides the standard of autopoietic identity. "Identity in contrast to everything else is nothing more than the determination and localization of difference in relative degree of complexity" (p. 184).

In their ongoing selections of complexity, meaning systems create distinctions in their environments and cocreate the objects they perceive. They gradually build a world of objects as these objects undergo a continual flux of contradiction, reorientation, and redundancy in a constantly reorganizing environment. This understanding of the origin of differences radically deontologizes objects. Under this interpretation,

> [There is] no unambiguous localization of any sort of "items" within the world nor any unambiguous classifying relation between them. Everything that happens belongs to a *system* (or to many systems) and *always at the same time* to the environment of other systems.
>
> p. 177

Autopoietic reproduction also guides the internal differentiation of a system. These systems reproduce not only their internal elements, but also their ability to reproduce. Luhmann puts it this way:

> The connection between reproduction and differentiation becomes comprehensible if one views reproduction, not as the identical or almost-identical replication of the same (e.g., as replacing supplies), but as a constantly new constitution of events that can be connected. Reproduction always implies reproducing the possibility of reproduction.
>
> p. 189

For social systems, this autopoietic reproduction continually reenacts the condition of double contingency. Every momentary event of a social system presents it with a challenge to adapt its structure to fit new complexity, *and* the possibility to generate within itself a new subsystem. In this process, a system evolves complexity and includes system–environment differences within itself.

Luhmann illustrates this situation by describing the situation in which a man might offer a light to a woman smoker.

> At a party one sees a woman reach for a cigarette, and (if she dawdles suggestively), one may offer her a light from one's own cigarette lighter. [In this situation,] settled differentiations stabilize the possibilities for reproduction by constraining conditions on the comprehensibility of communication and the suitability of behavioral modes. But the meaning surpluses that must be produced alongside provide ever further chance for innovative system formation; in other words, they provide the chance to include new differences and new constraints and thus to increase the ability to constrain the initial situation via differentiation. Only thus can system complexity increase.
>
> p. 189

Subsystems like this social one make improbable states become probable. They make possible the increase and normalization of organized complexity for the system. In our example, the interaction that provides the light may lead to a date and may lead to many otherwise improbable eventualities (like the man and woman going to a parent–teacher conference with their child).

Such improbable internal self-differentiations build subsystems upon an already pacified environment. They presuppose a social situation that is already rather regulated. The woman's dawdling with the cigarette is understood as an invitation to conversation. The conversation is understood as the basis for further mutual self-revelation, and so on.

Because of differentiation, each subsystem orients itself only to its own system–environment difference and its part of the system's overall complexity. It assumes that other reproductive requirements of the system are fulfilled by other subsystems. At the same time, each subsystem reconstructs the whole system. Because of the circular dependence of parts and wholes, each shift within a system rearticulates the self-reference of the whole system.

Action

The story related above illustrates an ever-present sequence of how social structures come to be. The woman sends an ambiguous message; the man responds to the message. These two actions are both momentary (they have a very short duration), but they form a relationship that is the basis for further events and deeper relationships, which may develop as in the story or lead to other consequence because of different choices.

The idea of action is based upon the fundamental and radical notion of an *event*. An event is a temporal atom, an indivisible, all-or-nothing happening. An action is such a momentary all-or-nothing happening. As such, it passes away as do all events in time, yet it "brings about a total change in past, present, and future … It gives up the quality of being present to the next event and becomes a past for it (i.e., for its future)" (p. 287). By passing away, an action provides maximum freedom vis-à-vis time to its successors and to the system that they constitute.

Boundaries

Social systems constitute themselves by communicating meanings. Through meanings, they establish their boundaries. In Luhmann's words, "The difference between system and environment is mediated exclusively by *meaning-constituted boundaries*" (p. 194). A system expands its boundaries every time it reproduces itself by incorporating new meaning. At the same time, it also redefines its boundaries by excluding elements of its environment. Each such event "makes a relation and with it a boundary decision" (p. 195).

Every communication is based upon an expectation that it will be heard. It "stakes a claim. At the very least, it demands time and attention" (p. 195). If it is accepted and if it contains new themes, it extends the system's boundaries, and enmeshes its speaker in the ongoing communicative action.

Meaning-constituted boundaries are self-generated boundaries that are regulated within the system itself. Systems are guided in their processing of new themes "by what has already occurred, by what is possible in a situation, and by general structures of expectation" (p. 197). Such structures of expectation help one to "foresee in detail how and about what one should communicate in the supermarket, on the football field, at lunch, at home …" (p. 197).

The World with Many Centers

Every autopoietic system creates centrality for itself by differentiating itself from its environment. If it does not do that, it exists in the state of undifferentiated innocence, 'the state of grace,' in which centrality is meaningless. With meaning-constituted boundaries, systems begin to have a *world*. They experience themselves, their environment, and everything that functions in it as an element, as a selection within a horizon that includes all possibilities and indicates further ones, that indicates an end and a beyond, that is both necessarily and arbitrarily bounded from anywhere within it (p. 207).

Systems correlate meaning and world. They position every meaning that they create within a world of previous meanings. They assume the existence of that world as background for the specification of every meaning.

The *world* becomes an ultimate concept that describes the *unity of the difference between system and environment* … an ultimate concept, one free from further differences … [It is a] unit of closure subsequent to a difference. It is the world after the fall from grace (p. 208).

In this world, the pride of centrality is removed from the subject. The center of the world is created by each system–environment difference that constitutes the world.

Every difference becomes the center of the world, and precisely that makes the world necessary: for every system–environment difference, the world integrates all the system–environment differences that a system finds in itself and its environment. In this sense, the world has multiple centers – but only so that every difference can fit the others into its own system and environment (p. 208).

Interpenetration

Systems that are environments for each other can couple their structures in such a way that they increase their independence from environmental determination by increasing their dependence on each other. By this process of 'structural coupling' (Maturana and Varela, 1987), two or more systems form a more complex integrated whole, while retaining (even increasing) their individual capabilities. In meaning (psychic and social) systems, structural coupling is described as interpenetration.

A system can make its own complexity, which involves indeterminacy, contingency, and the pressure to select, available to another system that can use that complexity for its own system-building. In doing this, it is said to have 'penetrated' the second system. From that point on, the behavior of the penetrating system is codetermined by the receiving system. When the Internet presents itself to me, for example, what happens on the net is codetermined by the use I make of it.

Interpenetration exists when penetration occurs reciprocally, "when both systems enable each other by introducing their own already-constituted complexity into each other … This means that greater degrees of freedom are possible in spite (better: because!) of increased dependencies" (p. 213).

As individuals try to communicate, they create the noise that generates social systems. "Social systems come into being on the basis of the noise that psychic systems create in their attempts to communicate" (p. 214). This 'noise,' or complexity, creates

pressure to select a meaning without specifying what meaning. As a result, it presents risk, freedom, and negatability to a system.

In this noisy situation, a system tentatively advances an expectation and then sensitively attunes itself to psychic and social feedback. If its expectation is denied, it learns to negate that expectation. If its expectation is affirmed, it is reassured and, in the course of time, it may become redundantly reassured. With redundancy, two or more systems set the stage for further ventures into noise.

When systems interpenetrate, they do share elements, but they share those elements from individual perspectives. As Luhmann says,

> Interpenetrating systems converge in individual elements-that is, they use the same ones-*but they give each of them a different selectivity and connectivity*, different pasts and futures. Because temporalized elements (events) are involved, the convergence is possible only in the present. The elements signify different things in the participating systems, although they are identical as events: they select among different possibilities and lead to different consequences.
>
> p. 215

In this way, autopoietic systems share meaning elements while maintaining their own autonomy.

Interpenetration connects the autopoieses of organic life, consciousness, and communication. These three forms of autopoiesis condition each other. "Autopoiesis qua life and qua consciousness is a presupposition for forming social systems" (p. 218). Conversely, "the social system, based on life and consciousness, makes the autopoiesis of these conditions possible in that it enables them to renew themselves constantly in a closed nexus of reproduction" (p. 219). In all of their interpenetrations, autopoietic systems use the complexity of their contributing systems, but they do so on the basis of their being autonomous. On both sides of structural coupling, there is "the *difference* between and the *interlocking* of *autopoiesis* and *structure* (the one continuously reproducing, the other discontinuously changing) … between *organic/psychic* and *social* systems" (p. 220).

This discussion of meaning, autopoiesis, and interpenetration touches merely upon the structural aspects of Luhmann's social theory. How we maneuver within his autopoietic world would involve us in discussions of:

- morality
- socialization
- expectations
  - persons
  - roles
  - programs
  - values
  - normative
  - cognitive
- structural change
- contradiction
- self-reference
- epistemology

See also: Law, Autopoiesis in; Luhmann, Niklas (1927–98); Networks and Meaning.

Bibliography

Bateson, G., 1972. Steps to an Ecology of Mind. Bantam, New York.

Bausch, K., 1997. The Habermas/Luhmann debate and subsequent Habermasian perspectives on systems theory. Systems Research and Behavioral Science 14, 315–330.

Bausch, K., 1999. The applicability of systems theory to social processes. Dissertation Abstracts International.

Bausch, K., 2001. The Emerging Consensus in Social Systems Thinking. Kluwer Academic/Plenum Publishers, New York.

Habermas, J., Luhmann, N., 1971. Theorie der Gesellschaft oder Sozialtechnologie: Was leistet die Systemforschung? Suhrkamp, Frankfurt [Theory of Society or Social Technology: What Does Systems Research Accomplish?].

Luhmann, N., 1990. Essays on Self-reference. Columbia University Press, New York.

Luhmann, N., 1995. Social Systems (J. Bednarz, D. Baecker, Trans.). Stanford University Press, Stanford (Original work published 1984).

Maturana, H.R., Varela, F.J., 1987. The Tree of Knowledge: The Biological Routs of Human Understanding. New Science Library, Boston.

Luria, Aleksandr Romanovich (1902–77)

Michael Cole, University of California, La Jolla, CA, USA
Xavier E Cagigas, University of California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Aleksander Romanovich Luria, known best as a pioneering neuropsychologist, conducted basic research on human cognitive and emotional functions between the early 1920s and the mid-1970s. His scientific career is reviewed; special emphasis is placed on the ways in which Luria's approach to neuropsychology is best understood in light of his lifelong quest to create a unified approach to understanding human development as a biocultural, social process.

Aleksander Luria (1902–77) is best known as the father of neuropsychology. However, analysis of his full career indicates that neuropsychology occupied only part of Luria's career and life project, which was devoted to the development of a unified psychological science. This biographical sketch follows Luria from his early research on objective methods of psychodiagnosis, through a series of studies on the cultural history and ontogeny of complex psychological functions, ending with his work on neuropsychology and his efforts to combine the study of general laws with the study of individual human functioning.

Aleksander Luria was born in Kazan, an old Russian University town east of Moscow on 16 July 1902. He entered Kazan University in 1918 at an unusually early age of 16 years owing to the social disorganization (and associated opportunities) associated with the Russian Revolution. He obtained his bachelor's degree in 1921 at the age of 19. In 1923, he moved to Moscow where he remained for the ensuing 54 years (with the exception of a period between 1942 and 1945 when he directed a clinic for rehabilitation of soldiers who had suffered brain injuries). During this long career, he taught and conducted research at a variety of important institutions including the Institute of Psychology, the Department of Psychology at Moscow State University, the Institute of Defectology, and the Bourdenko Institute of Neurosurgery. He died in Moscow on 14 August 1977.

While an undergraduate student, he established the Kazan Psychoanalytic Association and planned on a career in psychology. His earliest research sought to establish objective methods for assessing Freudian ideas about abnormalities of thought and emotion as well as the effects of fatigue on mental processes. His dream was to create a unified psychology that would encompass both the intimate personal thoughts and emotions of individuals, as well as the biological and social conditions that supported such processes (Luria, 1979). As he summarized his notions about psychoanalysis:

> Here, I thought, was a scientific approach that combined a strongly deterministic explanation of concrete, individual behavior with an explanation of the origins of complex human needs in terms of natural science. Perhaps psychoanalysis could serve as the basis for a scientific reale Psichologie, one that could overcome the nomothetic-idiographic distinction.
>
> Luria, 1979

In emulation of the psychoanalytical writers, he conducted clinical research on free associations, but he mistrusted the results of such efforts, feeling that any conclusions he tried to reach about the flow of his subjects' thoughts were insufficiently grounded. As he wrote in his autobiography, "While I was able to fill notebooks with (a patient's) free associations, I was in no position to carry out my plan to use such data to capture the concrete reality of the flow of ideas" (Luria, 1979).

In response to this dissatisfaction he created a methodology designed to embody a psychodynamic theory of mind in an objective set of laboratory procedures. The centerpiece of this methodology was an experimental technique that he called the combined motor method, which, he hoped, would provide a way of rendering Freud's clinical methods accessible to experimental treatment.

The fullest existing description of this work is contained in a monograph published in English under the title *The Nature of Human Conflicts: Or Emotion, Conflict and Will* (Luria, 1932). Luria explicitly rejected mechanical determinism in psychological analysis, declaring that "The structure of the organism presupposes not an accidental mosaic, but a complex organization of separate systems … (that) unite as very definite parts into an integrated functional structure" (Luria, 1932).

This approach required Luria to solve the following problem. Since this structure is the consequence of a long complicated development, both ontogenetically and cultural-historically, and because the parts are integrated into a whole functional system, how can it be possible to isolate elements in this system for purposes of psychological analysis? Phrased differently, since no two people are constructed alike, how could one possibly obtain valid evidence about the thought processes of another person?

The answer that Luria provided was that other people's thoughts can only be revealed indirectly in so far as they can be reflected in a publicly displayable, voluntary behavior. He phrased his strategy as follows:

> We should on the one hand, produce the central process of the disorganization of behavior; on the other hand, we should try to reflect this process in some [other] system accessible and suitable for examination. The motor function is such a systematic, objectively reflected structure of the neuro-dynamic processes concealed from immediate examination. And there lies before us the use of the motor function as a system of reflected structure of hidden psychological processes. Thus we proceed along the path we call the combined motor method.
>
> Luria, 1932

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61081-4

The first phase in his technique was to induce a well-coordinated, publicly available behavior as the medium of coordination necessary for the psychological analysis to be accurate. He used various devices for this purpose. Often the subject was requested to hold the left hand steady in a device that could record its movements, while simultaneously being asked to press a button or squeeze a bulb in response to verbal stimuli presented by the experimenter. Once this behavior became stable, the analyst sought to disrupt it selectively in line with his hypothesis about particular internal psychological states.

The combined motor method was applied to psychodiagnosis in a wide variety of real-life circumstances consistent with Luria's goal of demonstrating the possibility of a methodology powerful enough to reach beyond the laboratory to engage the kinds of long-term emotions that typically organize human behavior. For example, it became the basis for early lie detector tests and later for studies of the role of speech in the development of normal and mentally retarded children as well as adults suffering various neuroses and brain damage. It also served as the foundation for a great deal of his subsequent research on the relation of culture and human thought. As Luria (1932) expressed this relationship at the end of *The Nature of Human Conflicts*, "The analysis of complex cultural mechanisms is the key to the understanding of the simple neurodynamical processes."

In 1924, Luria had met Lev Vygotsky, a developmental psychologist whose influence was decisive in shaping Luria's future career. Together with Vygotsky and Alexei Leontiev, Luria sought to establish an approach to psychology that would enable psychologists to "discover the way natural processes such as physical maturation and sensory mechanisms become intertwined with culturally determined processes to produce the psychological functions of adults" (Luria, 1979). Vygotsky and his colleagues referred to this new approach variably as 'cultural,' 'historical,' and 'instrumental' psychology. These three labels all index the centrality of cultural mediation in the constitution of specifically human psychological processes, the role of the social environment in structuring the processes by which children appropriate the cultural tools of their society in the process of ontogeny, and the need to study processes of change over time. An especially heavy emphasis was placed on the role of language, the 'tool of tools' in this process: the acquisition of language in early childhood was seen as the pivotal moment in ontogeny when phylogeny and cultural history are merged to form specifically human forms of thought, feeling, and action. From the late 1920s until his death, Luria sought to elaborate this synthetic, cultural-historical psychology in different content areas of the discipline.

During the 1930s, he conducted two lines of research aimed at elucidating the relationship between phylogeny and cultural history in children's development (this topic is often referred to as the 'nature–nurture question'). The first project falls within the scope of what is now referred to as 'cross-cultural psychology.' It seeks to hold phylogeny constant while studying the effect of rapid cultural change.

In one implementation of this strategy, Luria led two expeditions to Central Asia where he investigated changes in psychological functioning brought about by changes in economic activity, social organization, and literacy that were then taking place (this work was only published after a long delay resulting from its ideological implications). In this research, he presented psychological tasks believed to tap such areas as perception, reasoning, problem solving, and self-concept to groups of adults who had, or had not, been incorporated into the Soviet system of collectivized farming or formal system of schooling. He found a marked difference in tested performance as a function of these variations in life experience. The thinking of those people who remained isolated from modern forms of agricultural work or schooling appeared to be dominated by immediate practical concerns bound to particular situations while those who had undergone the modernizing changes sponsored by the state appeared to think in more general and abstract ways that could be applied across a wide variety of situations.

A second line of work on the nature–nurture issue focused on comparisons between identical and fraternal twins exposed to different regimes of deliberate instruction. This work was organized at the Institute of Medical Genetics where hundreds of twins attended a specially organized boarding school. One set of studies involved the performance of identical and fraternal sets of twins of different ages on tasks designed to vary in the extent to which they depended upon cultural knowledge. Luria's basic finding in this work was that the correlation between the performances of identical twins were higher than those of fraternal twins when the tasks they were presented required relatively little specialized cultural knowledge. But the more the tasks required such knowledge, and the older the children were, the closer the correlations between fraternal twins approached the correlations among identical twins. He concluded that as children grow older and acquire more cultural knowledge, the influence of phylogeny over development gives way to the influence of cultural history.

He also conducted a special case study of a pair of identical twins who had developed what appeared to be their own private language before entering the boarding school as young children. For several months one of the twins was provided specialized training in the use of language to guide thinking about a variety of intellectual tasks while the other was not. The twin who received specialized training showed markedly more rapid intellectual development than the untrained twin, although the untrained twin's development rebounded once he, too, was provided specialized training. This result provided converging evidence on the role of culturally organized experience, especially language, on cognitive development (Luria and Yudovich, 1959).

In the late 1930s, largely to remove himself from public view owing to the period of political purges initiated by Joseph Stalin, Luria entered medical school where he specialized in the study of aphasia, retaining his focus on the relation between language, thought, and self-regulation, but now in the context of the politically neutral arena of medical practice. The onset of World War II made his specialized knowledge of crucial importance to the Soviet war effort, and the tragic widespread availability of people with various forms of traumatic brain injury provided him with voluminous materials for developing his theories of brain function and methods for the remediation of focal brain lesions. It was during this period that he developed the systematic approach to brain and cognition that has come to be known as the discipline of neuropsychology.

Several ideas were central to Luria's approach to brain-behavior relationships. First, he conceived of the brain as consisting of three global 'blocks' which underpin and control three associated kinds of brain processes. The first block, sometimes referred to as the 'primary sensory areas,' controls perceptual and motor integration. The second block, which includes the prefrontal and frontal areas, undergirds programing and executive control of thought and action. The third block, which includes the brain stem and paleocortex, controls the level of arousal and activation of the cortex as a whole. In addition to considering the workings of each of these blocks, he was particularly interested in the ways in which these three functional blocks interact dynamically to reconfigure one another in the execution of various neurocognitive processes.

Within this general framework, Luria sought to overcome a classical dichotomy between those who believed in the strict localization of brain functions and those who emphasized the 'equipotentiality' of different parts of the brain in subserving various psychological functions. While he accepted the idea that elementary functions such as the registration of sensations from the retina could be localized in particular groups of cells, he insisted that the systemic structure of complex forms of mental activity characteristic of adult humans represents complex functional systems involving the coordinated action of many anatomical locales. These functional brain systems are themselves mediated by the culturally organized environment, especially its systems of signs. As a consequence, he argued, "to understand the brain foundations for psychological activity, one must be prepared to study both the brain and the system of activity" (Luria, 1979). This insistence on linking brain structure and function to the proximal, culturally organized, environment provides the thread of continuity between the early and later parts of Luria's career. Third, he combined clinical practice with basic research on brain function, taking his ability to create systems of remediation as crucial evidence concerning the adequacy of his theories. Specifically, he believed that by reorganizing the extracortical components of various cognitive processes he could affect a change in the functional neural systems which remained intact in brain-damaged populations, thus leading to remediation of neurocognitive functions thought to be lost as the result of damage to specific brain areas.

Following the war, Luria sought to continue his work in neuropsychology. However, his plans were interrupted for several years in 1948 when he was removed from the Institute of Neurosurgery during a period of particularly virulent anti-semitic repression. For the next several years, he pursued his scientific interests through a series of studies of the development of language and thought among mentally retarded children at the Institute of Defectology. This work used a variation on the combined motor method to explore how the development of self-control is linked to the development of language, and provided a theoretical basis for the development of special curricula for the mentally handicapped.

In the late 1950s, Luria was permitted to return to the study of neuropsychology, which he pursued until his death of heart failure on 14 August 1977. During this period, he continued developing the idea of the brain as composed of functional systems closely linked to the conditions of the proximal environment, producing a number of important monographs on the brain bases of language, memory, and cognition. As this work became known outside of Russia, it gave rise to the development of standardized diagnostic tests (often referred to as the Luria Neuropsychological Battery) which are still widely used in clinical work. Even with the advent of highly sophisticated neuroimaging methods which can more directly localize structural brain lesions, the neuropsychological methods Luria pioneered continue to be the gold standard for understanding the functional impact of disruptions in neurocognition secondary to cerebral insults. Despite the recent explosion of technological innovation (i.e., MRI, PET, spectroscopy, etc.) which has threatened to relegate Luria's methods to footnote status in the neurosciences, what has remained internationally influential are the basic theoretical principles concerning the organization of higher psychological functions into complex functional systems mediated by sign systems and the use of these principles to guide the development of remedial procedures for brain-damaged patients (Christensen et al., 2009). Although not formally acknowledged, Luria's footprint is also quite evident in the recent manifestation of cultural neuroscience which attempts to study *how cultural values, practices and beliefs shape and are shaped by the mind, brain and genes across multiple timescales* (Chiao and Ambady, 2007).

In the years just prior to his death, Luria returned to his early dream of constructing a unified psychology that overcame the centuries-old split between natural and cultural science approaches to psychology. He proposed to do this by showing how theoretical principles arrived at through the experimental study of groups of people ('nomothetic' knowledge) could be made relevant to understanding and coping with the concrete life circumstances of an individual human being (idiographic knowledge). He called the approach combining the two psychologies 'romantic science.' He contrasted romantic science with what he called classical science:

> Classical scholars are those who look upon events in terms of their constituent parts. Step by step they single out important units and elements until they can formulate abstract, general laws …. Romantic scholars' traits, attitudes, and strategies are just the opposite. They do not follow the path of reductionism, which is the leading philosophy of the classical group. Romantics in science want neither to split living reality into its elementary components nor to represent the wealth of life's concrete events as abstract models that lose the properties of the phenomena themselves.

He illustrated his concept of romantic science in two longitudinal case studies involving people for whom ordinary ways of mediating action in the world was impossible (Luria, 1968, 1972). One was a man with a superb, but unusually organized memory. The other was a man who had suffered unusual disorganization of memory owing to massive destruction of the right posterior part of his brain. In each case, Luria combined information from experimental studies of large groups of subjects with the peculiarities of the individual case. His ideas about how the two kinds of processes combine were evaluated by the therapeutic regimes he prescribed.

These two case studies illustrate in dramatic detail his blend of classical, experimental approaches with clinical and remediational approach, a synthesis that stands as a model for psychological research and theory to the present day.

See also: Aphasia; Cognitive Neuroscience; Cross-Cultural Psychology; Emotion: History of the Concept; Experimental Laboratories: Psychological; Neuropsychological Testing; Pavlov, Ivan Petrovich (1849–1936); Psychoanalysis, History of; Psychology, History of (Twentieth Century); Psychotherapy Process Research; Psychotherapy, History of: Psychiatric Aspects; Vygotsky, Lev Semenovic (1896–1934).

Bibliography

Chiao, J.Y., Ambady, N., 2007. Cultural neuroscience: parsing universality and diversity across levels of analysis. In: Kitayama, S., Cohen, D. (Eds.), Handbook of Cultural Psychology. Guilford Press, NY, pp. 237–254.

Cole, M., Maltzman, I. (Eds.), 1969. A Handbook of Contemporary Soviet Psychology. Basic Books, New York.

Christensen, A., Goldberg, E., Bougakov, D. (Eds.), 2009. Luria's Legacy in the 21st Century. Oxford University Press, New York.

Goldberg, E. (Ed.), 1990. Contemporary Neuropsychology and the Legacy of Luria. Erlbaum, Hillsdale, NJ.

Luria, A.R., 1932. The Nature of Human Conflicts. Liveright, New York.

Luria, A.R., 1966. Higher Cortical Functions in Man. Basic Books, New York.

Luria, A.R., 1968. The Mind of Mnemonist. Basic Books, New York.

Luria, A.R., 1970. Traumatic Aphasia: Its Syndromes, Psychology, and Treatment. Mouton, The Hague.

Luria, A.R., 1972. The Man with a Shattered World. Basic Books, New York.

Luria, A.R., 1973. The Working Brain. Basic Books, New York.

Luria, A.R., 1979. The Making of Mind. Harvard University Press, Cambridge, MA.

Luria, A.R., Yudovich, F.A., 1959. Speech and the Development of Mental Processes. Staples Press, London.

M

Macroeconomics, History of From 1933 to Present

Kevin D Hoover, Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Schabas, volume 6, pp. 4152–4158, © 2001, Elsevier Ltd.

Abstract

The history of modern macroeconomics begins when much older economic questions were reclassified by Ragnar Frisch under the headings 'microeconomics' and 'macroeconomics.' The history of macroeconomics related here is importantly a history of the relationships of macroeconomics to microeconomics, and econometrics. The emphasis is on the development of macroeconomics as an interplay among economic theory, empirical investigation, and public policy.

The Origins of Modern Macroeconomics

The problems addressed by the field now called 'macroeconomics' are ancient (*see* Macroeconomics, History of up to 1933); yet the modern era begins in 1933 only with the coinage of the terms *macroeconomics* and *microeconomics* (almost certainly) by Ragnar Frisch, who also coined the term *econometrics* at about the same time (Velupillai, 2009). All three terms circulated informally – first, through the newly formed econometric society – before slowly becoming accepted as defining the broadest conceptual divisions of economics. For Frisch, microeconomics concerned the behavior of individual firms or consumers, whereas macroeconomics concerned the economy in its entirety. Practically, macroeconomics must give up on fine detail and deal with aggregated data. The history of modern macroeconomics is importantly a history of its relationships to microeconomics and econometrics.

Econometrics for Frisch was scientific economics and involved the cooperation of economic theory, mathematics, and statistics (Bjerkholt, 1998; Louçã, 2007). Like many others at the time he was concerned especially with *dynamics*, accounts of the movements of economic quantities over time. Jan Tinbergen gave life to the econometric project with the first macroeconometric model, a model of the Dutch economy in 1936. The worldwide Great Depression brought the problem of business cycles into the forefront of economics and debates over a correct understanding of the crisis raged throughout the 1930s. Tinbergen's approach had been pragmatic and not deeply committed to a particular theory.

While John Maynard Keynes never used the term 'macroeconomics,' it was only with the publication of his *General Theory of Employment, Interest, and Money* (1936) that its common theoretical core began to take shape. Keynes argued that earlier 'classical' economists failed to offer a theory of output as a whole, believing that market coordination was sufficiently effective that full employment of labor and capital failed only when market 'imperfections,' such as government regulation or unions intervened. In contrast, Keynes argued that, in complex, monetary economy, perfectly functioning markets may fail to be self-adjusting.

As a matter of accounting logic, aggregate savings in an economy (mainly lodged in financial assets) had to equal aggregate investment (mainly in the form of machinery and other physical means of production). But savers and investors were typically distinct actors: individuals, motivated by thrift, caution, and habitual psychology, made savings decisions; while firms in the face of a radically uncertain future, motivated by subjective expectations of profits, made investment decisions. Interest rates, representing the opportunity cost of real capital relative to financial assets were an important determinant of investment; while income was the most important determinant of consumption and savings. Formal financial markets operating through interest rates were inadequate by themselves to bring the plans of investors and savers into line. When savings plans exceeded investment plans, firms would find themselves with unsold goods and would be forced to cut back production and lay off workers – reducing incomes and bringing aggregate demand (i.e., planned total expenditure) back into line with aggregate supply (i.e., the value of total production), but at the cost of increased unemployment. Highly heterogeneous labor market inhibited the ability of workers effectively to offer to work for lower wages at times when demand for the products of firms fell short of their expectations, resulting in unemployment. The economy could be stuck for long periods in recession. The government could start a virtuous circle, directly stimulating demand through direct purchases, generating incomes that, in turn, generated further demand – a *multiplier* process. More important, the improved estimation of future profits would increase private investment and push the economy toward full employment. Keynes' position was, contrary to Marxist criticism, not that capitalism was doomed to failure. Rather it often – and

sometimes severely – underperformed. Yet, helpful policy interventions were possible.

Immediately on publication, the economics profession began to try to understand and interpret Keynes' *General Theory*. What ultimately came to be called the IS-LM model, due to John Hicks (1937), became the most influential interpretation. Hicks stripped the *General Theory* down to its architectural bare bones, stressing its commonalities with its 'classical' precursors and downplaying its more radical elements, such fundamental uncertainty and coordination failures. Hicks' simple model proved to be easily taught and a useful framework for econometric modeling and mathematical elaboration. While it clearly captured some of the essentials, Keynes, himself trained as a mathematician, was skeptical of the value of detailed formal modeling.

In 1938 and 1939, Tinbergen published a two-volume study based on the first econometric model of the United States economy. Utterly convinced of the wrongheadedness of the econometric project, Keynes published a scathing review (Morgan, 1990).

The World War II and Keynesianism

The General Theory laid the intellectual groundwork for government intervention in the economy; the World War II made it a reality. The war greatly expanded the size of government and the scope of its direct management – over prices, production, labor markets, financial markets, and foreign trade. Such direct management required better data. The theory and practice of national income accounting had been begun in the 1930s by Colin Clark and Simon Kuznets, among others (Kendrick, 1970). The war gave it an enormous practical boost, so that the United States and Great Britain began to publish official accounts shortly after the war with countries around the world quickly falling into line. Electronic computers, originally developed for military purposes, proved a godsend for the development of both data collection and the practical implementation of econometric modeling.

The fear of a postwar depression and the practical success of economic intervention during the war convinced governments that it was both feasible and necessary to take Keynes' advice to actively manage the macroeconomy. The Beveridge Report in Great Britain and the Employment Act of 1946 in the United States committed their governments to a policy of aiming at high rates of growth and low levels of unemployment. Abba Lerner (1943) coined the term 'functional finance' to argue that government policy should not be judged by whether the budget is balanced but by the state of the economic variables that directly affect welfare: income, employment, inflation. Government debt in the US, the UK and many other countries, which had historically been paid off after being run-up in wartime, was in fact maintained at higher levels after World War II than after earlier wars.

Economic management was extended to international trade and finance through the Bretton Woods Conference of 1944, in which Keynes was a key player, that led to the creation of the World Bank and the International Monetary Fund. These institutions in turn gave a boost to the development of international macroeconomics as a distinct discipline.

Microfoundational Programs and Macroeconometric Models

Frisch viewed the macroeconomic context as indispensable for microeconomic analysis. Most economists, however, viewed microeconomics as prior to macroeconomics: the economy as a whole being built out of individuals – explanation must ultimately be grounded in individual behavior. In the second half of the nineteenth century, Leon Walras had provided a vision of an agent-by-agent model of economic interdependence, based on the assumptions of perfect competition. From the 1930s to 1970s, substantial progress was made in formalizing Walras' vision of *general equilibrium*. One program of 'microfoundations' for macroeconomics aimed at discovering what assumptions would be necessary to formulate general equilibrium models that displayed 'Keynesian' features: unemployed resources, scope for beneficial policy intervention, and so forth (Weintraub, 1979; Hoover, 2012). This theoretical program made closest contact to practical macroeconomics in the 'general disequilibrium' models associated with Don Patinkin, Robert Clower, Robert Barro, Herschel Grossman, and Edmund Malinvaud (Backhouse and Boianovsky, 2012). These models emphasized the consequences of *rationing*: firms see potential profit in greater production provided that there is demand; unemployed workers would be willing to buy goods provided that they are hired; but neither is a position to act unless the other acts first; so the economy is stuck.

A second microfoundational program, whose central player was Lawrence Klein, also sought a marriage of Walrasian general equilibrium with Keynesian insights, but was empirically oriented (Hoover, 2012). Klein adopted an aggregate general equilibrium macroeconomic architecture similar to Hicks' IS-LM model, which lent itself especially to quantification and estimation using the new econometric tools developed mainly by the Cowles Commission (Morgan, 1990). Klein's ideal, however, was disaggregation as far as the data would allow. By 1950, Klein had estimated a series of small macroeconometric models of the United States (Bodkin et al., 1991). The twenty-five equation 'Klein–Goldberger' model became the paradigm for later macroeconometric models. Macroeconometric modeling became a 'big science' enterprise, involving large research teams with substantial division of labor over long periods of time. Klein was involved in a series of modeling projects, the Brookings model (about 200 equations in its 1959 version and 400 equations in its 1972 version), the Wharton model, the MPS model (a joint effort of the Massachusetts Institute of Technology, the University of Pennsylvania, and the Social Science Research Council), which later evolved into the FMP model (with the Federal Reserve replacing the Social Science Research Council). Many similar models were developed by other teams – academic, governmental, and commercial – for many countries. Project LINK, which Klein directed under the auspices of the United Nations, attempted to integrate national models into a worldwide model.

The increasing size of the models reflected increasing disaggregation. A major microfoundational effort – both explicitly within macroeconometric modeling teams and as stand-alone research – aimed to provide individual optimizing accounts for each of the main elements of the Keynesian

model. Reacting to empirical findings of Simon Kuznets, for example, alternative microeconomic accounts of the consumption function were developed and evaluated on microeconomic and aggregate data: Duesenberry's (1949) relative-income hypothesis, Friedman's (1957) permanent-income hypothesis, and Modigliani and Brumberg's (1954) life-cycle hypothesis. Some typical instances of similar efforts made for other Keynesian functions, include the demand for money (Baumol, 1952; Tobin, 1958), investment (Jorgenson, 1963), the production function (Fisher, 1969), and labor supply (Lucas and Rapping, 1969).

The ultimate function of macroeconometric modeling in the eyes of Keynesian was to provide the basis for policy guidance. Tinbergen (1956) laid out the 'targets-and-instruments' framework for policy analysis: policymakers set goals and economists, using macroeconometric models, guide their setting of monetary and fiscal policy tools to reach those goals. The economy is viewed as a machine to be manipulated. Keynesian economists widely concurred with the view that the ultimate resolution of the relationship between microeconomics and macroeconomics was less a matter of establishing a theoretical connection than of achieving in Paul Samuelson's (1955: p. vi) phrase "a grand neoclassical synthesis," in which macroeconomic policy would maintain full employment of resources, at which point all the old verities of microeconomics with respect to efficient resource allocation would apply.

Economic Growth

As postwar macroeconomic management appeared to be successful and fear of another 1930s-scale depression faded, macroeconomists turned their attention from business cycles to economic growth (Boianovsky and Hoover, 2009). The turn was, in part, a reaction to the Cold War, which could be framed as a competition between the Soviet Bloc and western democracies – both over economic dominance and over providing a persuasive and workable model for developing economies. Harrod's (1939) dynamic Keynesian model was interpreted as sharing a formal structure with Domar's (1946) growth model, which became a workhorse of development economics. The 'Harrod–Domar' model pointed to high rates of investment as the key to growth in economies with unemployed resources.

In 1956, Robert Solow and Trevor Swan independently criticized the 'Harrod–Domar' model for its assumption of fixed technology and offered a flexible-technology model in which the mixture of inputs can respond to the relative price of labor to capital. A surprising consequence of the Solow–Swan model is that the long-term, sustainable rate of growth does not depend on the rate of investment but only on the underlying rates of growth of the labor force and technology and on the rate of depreciation of capital. The *level* of per capita income nonetheless does depend on the rate of investment. In an empirical study framed by this model, Solow (1957) showed the surprising result that technological progress, rather than capital investment, explained by far the largest part of the actual growth of the United States economy.

Macroeconomists throughout the 1960s elaborated the Solow–Swan model, trying to give compelling theoretical accounts of technological progress and of 'optimal growth,' conceived of as either policies that maximized consumption per head ('golden rule policies') or ones that maximized intertemporally discounted social utility. With the rise of the environmental movement, which challenged growth as a social goal, and with the severe recessions of the early 1970s, growth economics moved to the back burner. By the end of the 1980s, as the business cycle moderated in developed countries, growth economics was revived. The attention was again placed on technological progress, but also on the apparent failure of developing economies to converge toward the performance of developed economies – a prediction of the Solow–Swan model. The 'new growth models' stressed production involving increasing returns to scale in which growth rather than exhausting productive opportunities expanded them, so that the more an economy grows the faster it can grow.

The Monetarist Challenge

Keynes and the postwar Keynesians saw fluctuations in aggregate demand, largely owing to the investment behavior of firms, as the main source of the business cycle, and they saw government policy (mainly fiscal policy) as necessary remediation. *Monetarism* (a term coined in the late 1960s) offered an anti-Keynesian alternative analysis (Mayer, 1978). Unlike Keynesians, monetarists, led by Milton Friedman, presumed that the private sector left to its own devices was fundamentally stable and that ill-conceived government policy was the main source of business cycles.

Friedman revived the quantity theory of money, which is the ancient doctrine of the *neutrality of money* – i.e., the price level in the economy in the long-run is directly proportional to the stock of money. One implication is that a policy that works through increasing the supply of money will ultimately affect only prices and not the quantities of real goods and services or employment, which are the sources of genuine economic welfare. But Friedman offered an important twist: in the short-run transitions between one long-run price level and another, changes in the stock of money could have powerful real effects. Friedman and Anna J. Schwartz's *Monetary History of the United States* (1963) – exhibit number one in the monetarist case – convinced many that the Federal Reserve's having allowed the money stock to fall precipitously was the principal cause of the Great Depression and that smaller recessions generally could be laid at the feet of central banks.

Under the metallic monetary standards that prevailed before World War II, high rates of inflation were largely wartime phenomena. After the war, the world experienced unprecedented secular increases in the price level, and macroeconomists sought the cause. They distinguished, for example, between price increases owing to rising costs of inputs (cost-push inflation) and rising expenditure on output (demand-pull inflation). The Phillips curve documented empirically an inverse relationship between the level of unemployment and the rate of inflation (Wulwick, 1987). Some economists read the Phillips curve as a trade-off: unemployment rates could be lowered provided that we are willing to tolerate higher inflation rates.

Friedman (1968a) argued, first, that there was ultimately only one kind of inflation. Cost-push inflation and inflation owing to changes in the demand for real goods (e.g., increases in government expenditure) were essentially *relative* price changes that provided the information about supply and demand of productive inputs and were essential to the smooth functioning of the economy. They were self-limiting. Only demand owing to increases in the money supply had no natural upper bound and could account for the postwar inflation: "inflation is always and everywhere a monetary phenomenon"(Friedman inter alia 1968b: pp. 18, 98, 105). Friedman argued that the Phillips curve resulted from workers mistaking the rise in all prices (including their wages) for an increase in their real wage, so that they were willing to work at what firms understood actually to be a lower real wage, thus lowering the unemployment. Friedman argued that workers' expectations of inflation would catch up with actual inflation rates and the real effects would vanish. The expectations-augmented Phillips curve (developed independently by Edmund Phelps) displays only a short-term trade-off, and policymakers cannot suppress unemployment below its *natural rate*. Even nonmonetarists rapidly adopted this model of inflation, although sometimes substituting the *nonaccelerating inflation rate of unemployment* for the 'natural rate' to avoid the implication that it was impervious to every kind of economic policy and not just to the management of monetary aggregate demand.

Monetarists argued that the short-term real effects of monetary policy, as they resulted from confusions about price signals, mainly misallocated resources. The economy was too complex, the channels from policy to real outcomes were too inscrutable, and the processes of individuals forming expectations were too unstable for policymakers to calibrate their interventions in reliably beneficial ways. Monetarists therefore advocated that policy should follow simple rules. For example, Friedman proposed that the supply of money should grow at a steady rate of X percent a year, where X was related to the underlying long-term growth rate of gross domestic product.

Monetarism triggered active, and sometimes acrimonious, empirical debates. Gradually, monetarists' views came to dominate at the Federal Reserve and, subsequently, at central banks around the world. Although not a pure implementation of monetarist doctrine, Chairman Paul Volcker's reform of Federal Reserve monetary policy, starting in 1979, and the subsequent collapse of US inflation rates was largely counted as a monetarist triumph. Nonetheless, monetarism remains controversial on theoretical, empirical, and policy grounds – not least, because financial innovation undermined a conceptually stable definition of the money stock, rendering empirical evidence equivocal and money-supply policy difficult to implement.

The New Classical Macroeconomics and the Challenge to Macroeconometric Modeling

One way or another, expectations play a part in each of the principal elements of the Keynesian model, and it was understood early on that empirical measures of expectations would be essential for successful macroeconometric modeling. Early efforts were based on mechanical extrapolations of past data, but these proved to be inconsistent in the sense that the best predictions of the model – i.e., the expectation that the model itself generated – would typically differ *systematically* from the expectation assumed to hold in the model. A systematic error could be easily reduced and might even be profitable to someone who was able to bet in the market against a policymaker informed by the model. John Muth's (1961) *rational-expectations hypothesis* is the simple idea that the expectation assumed to motivate agents in a model must coincide with the best prediction of that same model. Lucas and Rapping in the context of models labor markets and Thomas Sargent in the context of financial markets introduced the rational-expectations hypothesis into macroeconomics around 1970. Model consistency implied that all relevant aspects of the model must be considered simultaneously, which is a daunting technical challenge in models as complex as the macroeconometric models in the tradition of Klein. Early rational-expectations macromodels were small and assumed that all markets clear at all times. Clearing markets and rational expectations typically implies that money is neutral even in the short run: the only real effects of monetary policy are unsystematic, so that policy can add noise to the system but not achieve systematic improvements. Policies that adopt simple, easily understood rules minimize the noise. The *policy-ineffectiveness proposition* and the commitment to policy rules suggested to some that the new approach was a radical 'monetarism mark II.' Initially, some referred to it as the 'rational-expectations school,' but it was soon seen that the characteristic results depended on clearing markets and that the rational-expectations hypothesis could be used even in models with typically Keynesian outcomes, so that *new classical macroeconomics* is a better term for the new approach (Hoover, 1988).

Monetarism had accepted the aggregative structure of Keynesian models, differing over more specific details (Gordon, 1974). Robert Lucas (1976) offered a more radical objection – the *Lucas critique*. Tinbergen saw macroeconometric models as a tool for policymakers guiding their use of policy instruments. Lucas argued that policymakers must be thought of not as standing outside the economy, as Tinbergen's framework suggested, but on a par with other actors within the economy. Those actors observe and anticipate the policies and adjust their behavior accordingly. The aggregate relationships of the macroeconometric model are not, therefore, likely to remain stable in the face of alternative policies. Lucas proposed that the only solution to this problem of *policy noninvariance* is to model explicitly the decisions of individual agents reacting to policy.

The Lucas critique shifted the landscape of macroeconomics. Despite a long history of trying to understand the microeconomics behind macroeconomics, mainstream macroeconomists for the first time insisted explicitly that only models with appropriate 'microfoundations' would be acceptable. The piecemeal approach of the Klein program was rejected in favor of a systemic approach based on a general equilibrium with rational expectations. The technical challenges were, however, insurmountable, and, without much reflection or critical discussion, the profession adopted the

representative-agent model, an approach in which a single (or sometimes a few types of) agent, who maximizes a utility or profit function in the manner of a microeocnomic agent is taken to stand for all agents in the economy (Hartley, 1997). The representative-agent model simply ignores the conceptual and technical problems of aggregation and, in fact, provides the simulacrum of microfoundations and not the genuine article.

Keynesians quickly adopted key planks of new classicism: rational expectations, the Lucas critique, representative-agent microfoundations. They challenged the assumption that markets clear rapidly and suggested various 'frictions' – institutional or economic features that inhibit rapid adjustment of prices or wages to external 'shocks' or to policy actions. Frictions were supposed to provide the microeconomic underpinning of short-run nonneutrality. *New Keynesian* approaches emphasize a range of frictions, including contracts, menu costs, efficiency wages, imperfect competition, informational asymmetry, and coordination failures (Mankiw and Romer, 1991). Some approaches require going beyond strictly representative-agent models, but the departures are rarely very large. The payoff is to better identify sources of aggregate fluctuations and to preserve a scope for beneficial policy intervention.

The weakest point of early new classical models was the inability to explain business cycles, which was the main opening for the new Keynesian challenge. Lucas constructed a new classical monetary business-cycle model in 1975. Random shocks make some periods more productive than others, and forward-looking workers and producers are more active in these periods – a boom. Long-lived capital goods add dynamics, transmit transient booms to the future, and smooth the transitions, so that booms and busts are stretched over longer periods. Lucas' model was not empirically successful, but it provided the template for further developments. Most notably, Kydland and Prescott (1982) developed a model in which random shocks to the productivity of technology was the main driving force of the cycle. For some time, such *real-business-cycle models* dominated the new classical research agenda, with successful modifications trying to improve the empirical match (Hartley et al., 1998).

The New Neoclassical Synthesis

By the mid-1990s the debate between new classicals and the new Keynesians had cooled. The effort to improve the empirical fit of real-business-cycle models had forced new classicals to consider additional sources of fluctuations than technology shocks and to integrate price adjustment or other frictions into their models. But the larger theoretical template of the real-business-cycle model – rechristened the *dynamic, stochastic general equilibrium* (or *DSGE*) model – became the workhorse for business-cycle analysis (Duarte, 2012). Soon a *new neoclassical synthesis* was declared.

For practical monetary-policy analysis, the new neoclassical synthesis focused on a simple, aggregate model that was regarded as a kind of summary for a microfoundationally adequate business-cycle model. The paradigm model included a forward-looking investment–savings relationship (IS curve), an expectations-augmented Phillips curve, and a monetary-policy rule governing the setting of short-term interest rates. The main use of the models was to evaluate counterfactually the performance of alternative monetary-policy rules. These models, along with the DSGE models, found widespread acceptance in central banks around the world, sometimes coexisting uneasily with older-style macroeconometric models, which had fallen out of favor with academic economists in the 1970s.

The development of macroeconomics has frequently been shaped by events in the world: the Great Depression, World War II, the oil price shocks of the 1970s and 1980s. The worldwide recession and financial crisis beginning in December 2007 appears to be another such shock. It exposed weaknesses in currently popular macroeconomic models. Models frequently lacked a financial sector, limiting their relevance to the main element of the crisis. The assumption of rational expectations and the absence of heterogeneity among agents made it difficult to capture or to characterize the mechanics of the asset price boom that preceded the crisis or its sudden collapse. The radical reaction among economists is that the financial crisis challenges the fundamental basis on which modeling – at least since the 1970s – has preceded: fundamentally new macroeconomic thinking is in order. The conservative reaction acknowledges the incapacities of current models, but sees their foundations as essentially sound: extension and development of the existing lines of macroeconomic thinking will suffice. Research proceeds along both lines. It is too early to offer a historical assessment.

See also: Data Bases and Statistical Systems: Economics (Macroeconomics); Economic Growth; Economics, History of; Expectations, Economics of; Macroeconomics, History of From 1933 to Present; Monetary Policy Since the 2007–2009 Financial Crisis; Unemployment: Structural.

Bibliography

Backhouse, Roger, Boianovsky, Mauro, 2012. Transforming Modern Macroeconomics Exploring Disequilibrium Microfoundations, 1956–2003. Cambridge University Press, Cambridge.

Baumol, William J., 1952. The transactions demand for cash: an inventory theoretic approach. Quarterly Journal of Economics 66 (4), 545–556.

Bjerkholt, Olav, 1998. Ragnar Frisch and the foundation of the econometric society. In: Strøm, Steinar (Ed.), Econometrics in the 20th Century: The Ragnar Frisch Centennial Symposium. Cambridge University Press, Cambridge, pp. 26–57.

Bodkin, Ronald, Klein, Lawrence, Marwah, Kanta, 1991. A History of Econometric Model Building. Edward Elgar, Aldershot.

Boianovsky, Mauro, Hoover, Kevin D., 2009. The neoclassical growth model and twentieth-century economics. In: Boianovsky, Mauro, Hoover, Kevin D. (Eds.), Robert Solow and the Development of Growth Economics. History of Political Economy 41, 1–23.

Boianovsky, Mauro, Hoover, Kevin D., 2014. In the kingdom of Solovia: the rise of growth economics at MIT, 1956–1970. In: Roy Weintraub, E. (Ed.), MIT and the Transformation of American Economics. In: Annual Supplement to History of Political Economy, vol. 46. Duke University Press, Durham, NC, 198–228.

Domar, Evsey D., 1946. Capital expansion, rate of growth, and employment. Econometrica 14 (2), 137–147.

Duarte, Pedro G., 2012. Not going away: representative-agent models and microfoundations in the making of a new consensus in macroeconomics. In: Duarte, Pedro G., Lima, Gilberto T. (Eds.), Microfoundations Reconsidered: The Relationship of Micro and Macroeconomics in Historical Perspective. Edward Elgar, Cheltenham, pp. 190–238.

Duesenberry, James S., 1949. Income, Saving and the Theory of Consumer Behavior. Harvard University Press, Cambridge, MA.
Fisher, Franklin M., 1969. The existence of aggregate production functions. Econometrica 37 (4), 553–577.
Friedman, Milton, 1957. A Theory of the Consumption Function. Princeton University Press, Princeton.
Friedman, Milton, 1968a. The role of monetary policy. American Economic Review 58 (1), 1–17.
Friedman, Milton, 1968b. Dollars and Deficits: Inflation, Monetary Policy and the Balance of Payments. Prentice-Hall, Englewood Cliffs, NJ.
Friedman, Milton, Schwartz, Anna J., 1963. A Monetary History of the United States, 1867–1960. Princeton University Press, Princeton.
Gordon, Robert J. (Ed.), 1974. Milton Friedman's Monetary Framework: A Debate with His Critics. University of Chicago Press, Chicago.
Harrod, Roy F., 1939. An essay in dynamic theory. Economic Journal 49 (193), 14–33.
Hartley, James E., 1997. The Representative Agent in Macroeconomics. Routledge, London.
Hartley, James E., Hoover, Kevin D., Salyer, Kevin D., 1998. The limits of business cycle research. In: Hartley, James E., Hoover, Kevin D., Salyer, Kevin D. (Eds.), Real Business Cycles: A Reader. Routledge, London.
Hicks, John R., 1937. Mr. Keynes and the 'Classics': a suggested interpretation. Econometrica 5 (2), 147–159.
Hoover, Kevin D., 1988. The New Classical Macroeconomics: A Sceptical Inquiry. Blackwell, Oxford.
Hoover, Kevin D., 2012. Microfoundational programs. In: Garcia Duarte, Pedro, Lima, Gilberto Tadeu (Eds.), Microfoundations Reconsidered: The Relationship of Micro and Macroeconomics in Historical Perspective. Edward Elgar, Cheltenham, pp. 19–61.
Jorgenson, Dale W., 1963. Capital theory and investment behavior. American Economic Review 53 (2), 247–259.
Kendrick, John W., 1970. The historical development of national-income accounts. History of Political Economy 2 (2), 284–315.
Keynes, John M., 1936. The General Theory of Employment, Interest and Money. Macmillan, London.
Kydland, Finn, E., Prescott, Edward, C., 1982. Time to build and aggregate fluctuations. Econometrica 50 (6), 1345–1370.
Lerner, Abba P., 1943. Functional finance and the federal debt. Social Research 10 (1), 38–51.
Louçã, Franciso, 2007. The Years of High Econometrics: A Short History of the Generation that Reinvented Economics. Routledge, London.
Lucas Jr., Robert E., 1976. Econometric policy evaluation: a critique. In: Brunner, Karl, Meltzer, Alan H. (Eds.), The Phillips Curve and Labor Markets, Carnegie-Rochester Conference Series on Public Policy, vol. 1. Elsevier, New York, pp. 19–46.
Lucas Jr., Robert E., 1975. An equilibrium model of the business cycle. Journal of Political Economy 83 (6), 1113–1144.
Lucas Jr., Robert E., Rapping, Leonard A., June 1969. Price expectations and the Phillips curve. American Economic Review 59 (3), 342–350.
Mankiw, Gregory, Romer, David, 1991. The New Keynesian Economics, vols. 1 and 2. MIT Press, Cambridge, MA.
Mayer, Thomas (Ed.), 1978. The Structure of Monetarism. Norton, New York.
Modigliani, Franco, Brumberg, Richard, 1954. Utility analysis and the consumption function: an interpretation of cross-section data. In: Kurihara, Kenneth K. (Ed.), Post-Keynesian Economics. Rutgers University Press, New Brunswick, NJ.
Morgan, Mary S., 1990. The History of Econometric Ideas. Cambridge University Press, Cambridge.
Muth, John F., 1961. Rational expectations and the theory of price movements. Econometrica 29 (3), 315–335.
Samuelson, Paul A., 1955. Economics: An Introductory Analysis, third ed. McGraw-Hill, New York.
Solow, Robert, M., 1957. Technical change and the aggregate production function. Review of Economics and Statistics 39 (3), 312–320.
Tinbergen, Jan, 1956. Economic Policy: Principles and Design. North-Holland, Amsterdam.
Tobin, James, 1958. Liquidity preference as behavior toward risk. Review of Economic Studies 25 (2), 65–86.
Velupillai, Kumaraswamy Vela, 2009. Macroeconomics – a clarifying note. Economia Politica 26 (1), 135–137.
Weintraub, Roy, E., 1979. Microfoundations: The Compatibility of Microeconomics and Macroeconomics. Cambridge University Press, Cambridge.
Wulwick, Nancy J., 1987. The Phillips curve: which? whose? to do what? how? Southern Economic Journal 53 (4), 834–857.

Macroeconomics, History of up to 1933

Robert W Dimand, Brock University, St. Catharines, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The field that became known as macroeconomics emerged from long and rich traditions of monetary theory, business cycle analysis, and discussion of economic growth. Macroeconomics is the analysis of aggregate levels and rates of change of output and employment, the price level (and its inverse, the purchasing power of money), the interest rate, and the balance of payments. This field had a long history before it received its name with Ragnar Frisch's (1933) distinction between macrodynamics and microdynamics and before John Maynard Keynes' General Theory of Employment, Interest and Money (1936). This article analyzes the 'prehistory' of the concept – the various strands of economic thinking that coalesced in about 1933 to produce modern macroeconomics, i.e., roughly the history of macroeconomics up to Keynes. The subsequent history is covered in a companion article, "Macroeconomics, history of, from 1933 to the present."

Macroeconomics, the analysis of aggregate levels and rates of change of output and employment, the price level (and its inverse, the purchasing power of money), the interest rate, and the balance of payments had a long history before the field received its name with Ragnar Frisch's (1933) distinction between macrodynamics and microdynamics and before John Maynard Keynes' *General Theory of Employment, Interest and Money* (1936).

Prehistory

Three currents of analysis and discussion, overlapping but distinguishable, emerged on subjects later considered macroeconomics, coalescing in the 1930s. The theory of money and prices began with the quantity theory of money as an explanation of the 'Price Revolution' of the sixteenth century and of balance of payments adjustment under the gold standard (Grice-Hutchison, 1952; Locke, 1991; Gervaise, 1720; Hume, 1752; Christiernin, 1761; Monroe, 1923; Hegeland, 1951; Vickers, 1959; Guggenheim, 1989), was extended to inconvertible paper currency during the French Revolution and Napoleonic Wars (Thornton, 1802; Murphy, 2009; Arnon, 2011), debated Jean-Baptiste Say's law of markets and the possibility of a general glut of commodities at the end of the Napoleonic Wars (Sowell, 1972; Link, 1959; Corry, 1962; Green, 1992; Baumol, 1999; Davis, 2005), and achieved analytical sophistication during the Banking School–Currency School–Free Banking School debates (Fetter, 1965; O'Brien, 2007), culminating in Walter Bagehot's treatment of the lender of last resort in *Lombard Street* (1873) and then in the contributions in the 1890s by Alfred Marshall, Irving Fisher, and Knut Wicksell, the three theorists appropriately pictured on the cover of David Laidler's *The Golden Age of the Quantity Theory* (1991a); the interwar flourishing of monetary analysis of saving–investment adjustment (Eshag, 1963; Bridel, 1987; Jonung, 1991; Laidler, 1999; Dimand, 2003; and sources reprinted in Wood and Capie, 2000). A second approach, brought to prominence by Clément Juglar (1862) and W. Stanley Jevons (1884), viewed that fluctuations in production and prices are cycles, oscillations with more or less regularity in amplitude and periodicity, producing a vast literature sampled by D.P. O'Brien (1999), Harald Hagemann (2001), and Mauro Boianovsky (2005), and leading to the creation in the interwar period of business cycle research institutes such as the National Bureau of Economic Research (NBER) (Mitchell, 1927; Kondratiev, 1997; Glasner, 1997). Histories of crisis theories, and particularly of theories of periodic crises, have been published as early as E. von Bergmann (1895) and Edward D. Jones (1900), with histories of the phenomenon of crises going back to M. Wirth (1858). Thirdly, François Quesnay's Physiocratic *Tableau Économique*, tracing the circular flow of spending and income in the economy as whole, was followed by the 'magnificent dynamics' of classical economists such as David Ricardo, Thomas Robert Malthus, and John Stuart Mill, exploring the paths of population, production, and wage and profit rates as a growing economy approached a stationary state (Eltis, 1984; Davis, 2005; Brewer, 2010), with the beginnings of formal algebraic models of economic growth sketched by J.C.L. Simonde de Sismondi and developed in Karl Marx's "schemes of simple and expanded reproduction" (Arthur and Reuten, 1998) and a false dawn of modern growth theory in with Allyn Young (1928) on increasing returns to scale and economic growth, Frank Ramsey (1928) on optimal capital accumulation, and Grigorii Fel'dman (1928) on the growth rate in a planned economy.

These extensive pre-1933 contributions to what would later be called macroeconomics raised issues that resonate in later macroeconomics (as the terms New Classical and New Keynesian indicate, as do Divisia monetary aggregates (François Divisia, 1925), Ricardian equivalence, Schumpeterian creative destruction). The work of Milton Friedman and his students revived the quantity theory tradition (see Friedman, 1956). Michael Woodford's *Interest and Prices* (2003) harks back to the analysis of a pure credit economy in Knut Wicksell's *Interest and Prices* (1898). Christopher Sims' vector autoregression (VAR) approach to modeling economic fluctuations without assuming too much a priori knowledge of the underlying model is, with more modern statistical techniques, in the spirit of Wesley Mitchell's NBER statistical

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03058-0

approach to business cycles, starting with time series analysis of fluctuations rather than with theoretical models (an approach that produces the concept of leading indicators). Irving Fisher, although spectacularly mistaken about whether stock prices had reached a permanently high plateau in October 1929, shaped macroeconomic analysis and practice through the Fisher equation relating real and nominal interest, the Fisher diagram for intertemporal consumption smoothing, indexed bonds, distributed lags, the Fisher ideal index number (the geometric mean of the Paasche and Laspeyres indexes), and, especially since the credit crunches of 2007 and 2008, his debt deflation theory of depressions (see e.g., Korinek and Mendoza, 2013), with Ralph G. Hawtrey's concept of a credit deadlock and the Mises–Hayek Austrian trade cycle theory also attracting renewed attention. Central bank responses to the credit crunch were informed by the fact that Ben Bernanke of the Federal Reserve, in his doctoral dissertation, and Mervyn King of the Bank of England, in his European Economic Association presidential address, had studied the US debt deflation of the early 1930s, with close attention to Fisher (1933), just as monetary policy decisions at the Bundesbank and then at the European Central Bank were informed by acquaintance with the literature surrounding the German hyperinflation of the early 1920s (including Keynes, 1923). Macroeconomics did not begin in 1933 or 1936, although certainly Keynes (1936) provided a transformative focus on the equilibrium level of aggregate output and employment rather than on prices and cycles.

The Theory of Money and Prices

European prices roughly quadrupled from 1530 to 1600, as was documented for English prices by Bishop William Fleetwood (1707), pioneer of price history and of quantification of changes in the purchasing power of money (but because the commodity prices examined by Fleetwood all changed in roughly the same proportion, he did not have to consider appropriate weights in a price index). This 'Price Revolution' occasioned the emergence of the quantity theory of money, described by Mark Blaug (in Blaug et al., 1995) as "the oldest surviving theory in economics" and by David Laidler (1991b) as "always and everywhere controversial." Martin Navarro de Azpilcueta of the University of Salamanca, writing in Latin in 1556, and the French political theorist Jean Bodin in 1568 attributed the rise in European prices to the rise in the same proportion of the quantity of money circulating in Europe, due to the inflow of gold and especially silver to Spain from Mexico and Upper Peru (now Bolivia) and to increased output of European silver mines (see Marjorie Grice-Hutchison, 1952 on the Salamanca School, and Arthur E. Monroe, 1923 and Hugo Hegeland, 1951 on Bodin). Azpilcueta and the other Salamanca writers were the last phase of a medieval scholastic tradition that included Nicole Oresme's analysis around 1360 of how the prince's seigniorage from debasement of the coinage took purchasing power away from the public (Oresme et al., 1989), a discussion that drew in turn on Aristotle's remarks on the functions of money. Isaac Gervaise (1720), Richard Cantillon, P.N. Christiernin (1761), and, above all, David Hume (1752) used the quantity theory of money to analyze the automatic restoration of international payment equilibrium and critique the mercantilist obsession with a balance of trade surplus, holding that in the long run a change in the quantity of money will cause prices and money wages to change in the same proportion, and the money supply will be distributed across countries in proportion to the demand for real money balances in each country. In Hume's price–specie flow mechanism, mercantilist interventions, using tariffs and quotas on imports, regulations, bounties on exports, and monopolies on colonial trade to increase the quantity of money in a country through a trade surplus would, in the long run, just raise prices in the home country and lower prices in its trading partners, reducing net exports until equilibrium was restored. In the short run, an increase in the quantity of money would give a temporary stimulus to economic activity until prices and wages fully adjusted (see Filippo Cesarano, 2007; Arie Arnon, 2011; Dimand, 2013). The specie (gold and silver) that flowed into Spain from its empire in the New World flowed out again in trade. In Adam Smith's critique of mercantilism, closer than Hume's of the twentieth century monetary approach to the balance of payments, the terms of trade would not change for a country that, like Scotland, was small enough to be a price taker in world markets (Humphrey, 1993; Laidler, 2003: Chapter 1; Arnon, 2011). If mercantilist interventions artificially increased the quantity of money in a country, the resulting excess supply of money would be spent on imports and on exportable goods (and on foreign lending), causing the gold and silver to flow out of the country again, until the world money supply was distributed in proportion to demand for real money balances, with money supply equal to money demand in each country.

Although John Law had experimented with inconvertible paper currency in France until the collapse of Law's system in 1720, David Hume, Adam Smith, and their contemporaries analyzed price and payments adjustment under a metallic standard, with gold and silver coin and bullion circulating (see Murphy, 2009, on Law). Following a bank run sparked by reports of a French invasion, conversion on demand of Bank of England notes into specie was suspended from 1797 to 1821. Henry Thornton, the Bullion Committee of the House of Commons, and David Ricardo extended the quantity theory of money to inconvertible paper, conclusively establishing that the rise in British prices during the Napoleonic Wars and the high price of bullion on the Hamburg market (when purchased with Bank of England notes) resulted from the increase in the Bank of England's note issue, an analysis that failed to convince either the Bank's directors or the majority in the House of Commons (Thornton, 1802; Hayek, [1931]1991; Viner, 1937; de Boyer des Roches, 2003; Davis, 2005; Arnon, 2011). The Bank directors insisted that they could not overissue notes as long as they only discounted real bills of exchange to accommodate the needs of trade, forgetting that the nominal volume of trade would grow as prices increased. The directors invoked the authority of Adam Smith for this real bills doctrine, even though Smith's analysis had dealt with bank notes that were convertible into coin on demand. Similarly, John Law had argued in 1705 that a land bank could not overissue notes as long as the notes were created by lending with land as security, overlooking the endogeneity of the price of land.

At the end of the Napoleonic Wars, the sharp reduction in military spending, availability of cheaper grain imports from Continental Europe and the United States, and the contraction of the note issue as the Bank of England moved toward restoration of convertibility at the prewar parity combined to produce industrial and agricultural depression in Britain. The classical theory of rent and of comparative advantage emerged from the classical critique of the Corn Laws, the protectionist reaction to the agricultural depression. Thomas Robert Malthus and J.C.L. Simonde de Sismondi attributed the industrial depression to a general glut of commodities, an insufficiency of effective demand. David Ricardo, James Mill, and Jean-Baptiste Say held that such a general glut was impossible, because people only supply goods, services, and assets because they wish to acquire other goods, services, and assets, so that the industrial depression was a transitional problem, as wage and price signals shifted factors of production from producing commodities in demand during wartime to producing commodities in demand during peacetime (Link, 1959; Corry, 1962; Patinkin, 1965; Sowell, 1972; Green, 1992; Hollander, 1997; Davis, 2005). The problem of transition was not considered by the classical economists as something to be ignored – Say endorsed public works to provide employment during the transition, whereas Ricardo wished to restore convertibility at the depreciated value of notes rather than the prewar parity – but it was a temporary, short-run issue, resulting from a change in the mix of commodities demanded, not from a shortfall of demand in general. Following a tradition going back to Aristotle's invocation of the myth of King Midas, Ricardo, Say, and James and John Stuart Mill held that people sell commodities for money intending to use the money as a medium of exchange, not with the hoarding of money as the ultimate goal. In the debate over what Keynes was to call Say's law of markets (see Baumol, 1999), (aggregate) supply creates its own (aggregate) demand, the classical economists (in Keynes' sense of classical as accepting some form of Say's Law) tried to say something stronger than what Oskar Lange (1942) later termed Say's Equality (the value of excess demand sums to zero in equilibrium over all markets other than money – that is, the truism that the value of aggregate excess demand is zero if excess demand is zero in each market), but weaker than what Lange termed Say's Identity (the value of excess demand sums to zero over all markets except money, at any set of prices, market clearing, or otherwise – which, together with Walras' Law, implies that the money market always clears for any set of prices). Keynes (1933a, 1936) drew inspiration from Malthus' emphasis on the concept of effective demand (demand allowing for quantity constraints: if the labor market is in excess supply, unemployed workers cannot spend on goods the wages they would have earned if employed), even though Malthus' failure to distinguish a decision to save from a decision to invest undermined his response to Ricardo's analysis.

The Bank Act of 1844 and the British banking crises of 1825, 1832, 1836–37, 1847, and 1857 occasioned the debate between the Currency School and Banking School over the appropriate structure and policy of the Bank of England (Viner, 1937; Fetter, 1965; Skaggs, 1999; Laidler, 2003; O'Brien, 2007; Arnon, 2011), with the Free Banking School denying the need for any legally privileged bank of issue acting as a central bank (Smith, 1936). Drawing on Ricardo's presentation of the quantity theory of money and the price–specie flow mechanism, Currency School writers such as J.R. McCulloch, G.W. Norman, and Lord Overstone urged that the Bank of England should act so that a mixed paper and metallic currency would behave like a metallic currency (for instance, a 100% marginal reserve requirement on note issue), and should manage the volume of currency to prevent gold outflows. The Currency School, whose ideas shaped Sir Robert Peel's Bank of 1844, attributed banking crises to mismanagement by the Bank of England. Banking School writers such as Thomas Tooke and John Fullarton emphasized the endogeneity of credit, with the volume of credit more likely to react to changes in prices than to cause them, and with the quantity of banknotes demand-determined. Rather than expecting the Bank of England to be able to prevent gold drains by varying the discount rate, the Banking School wished the Bank to hold a large enough bullion reserve to accommodate temporary balance of payments fluctuations without unsettling credit markets. These debates in monetary theory culminated in Walter Bagehot's *Lombard Street* (1873), the classic account of how the central bank should act as lender of last resort to preserve the liquidity of the banking system, rather than acting as an ordinary bank to maximize the profits of its shareholders.

1886 and All That

The Royal Commission on the Depression of Trade and Industry (Northcote, 1886: p. x) reported "that the trade and industry of the country are in a condition that may be fairly described as depressed; that by this depression is meant a diminution, and in some cases, an absence of profit, with a corresponding diminution of employment for the laboring classes; that neither the volume of trade nor the amount of capital invested therein has materially fallen off, though the latter has in many cases depreciated in value; that the depression above referred to dates from about the year 1875, and that, with the exception of a short period of prosperity enjoyed by certain branches of trade in the years 1880 to 1883, it has proceeded with tolerable uniformity." The same year, the first annual report of the US Commissioner of Labor was on *Industrial Depressions* (Wright, 1886; Barnett, 1941). Lecturing in Scotland on *Irregularity of Employment and Fluctuations of Prices* (Foxwell, 1886: pp. 48, 96), Herbert S. Foxwell, professor of political economy at University College, London (where Knut Wicksell attended some of his lectures that year) and a fellow and lecturer of St. John's College, Cambridge, lamented "the most serious effect of falling prices. They depress industry and trade" leading to what is "of all the many claims of labor, the most grave, the most pressing, and the most just, … the claim I have brought before you to-night, the claim for more regular employment." The literature on depression and unemployment grew until a bibliography on unemployment published by the London School of Economics in 1909 extended to 71 pages (not counting Sidney Webb's Introduction). Many shared Foxwell's conviction that fluctuations in the purchasing power of money caused fluctuations in employment (although formal economic theory, like Foxwell, was vague on the mechanism connecting

the two phenomena). As Foxwell (1886: p. 52 n) noted, "So long ago as 1856, Mr. William Cross ably advocated a monetary standard corrected by means of an index number. See his *Standard Pound v. the Pound Sterling* (Aberdeen, 1856). Another method for securing the same object has recently been proposed by M. Walras" (see Léon Walras' *Théorie de la Monnaie*, 1886, and Bridel, 1997). Alfred Marshall (1887), also of St. John's College, Cambridge, and one of the witnesses before the Royal Commission, tentatively suggested, as a first step toward stabilizing the purchasing power of an index number, symmetallism, a monetary unit of a certain amount of gold plus a certain amount of silver, rather than just one metal, but backed away after Sir Robert Giffen blasted 'fancy monetary standards' that would tamper with the trusted, impersonal, and automatic gold standard (Laidler, 1991a; see also Marshall, 1923, 1926). The possibility of using monetary policy to target a price index (rather than an exchange rate or the price of gold), in the interest of stabilizing output and employment, was a thought that occurred to economists before the twentieth century – and although William Cross is a forgotten figure who was never part of the mainstream of economics, the same cannot be said of Walras or Marshall.

Monetary Theorists and Economic Stabilization

Bimetallists, such as William Jennings Bryan in his "Cross of Gold" speech accepting the Democratic nomination for President of the United States in 1896, saw gradually declining prices as a burden on debtors (especially farmers with mortgages), and proposed to increase the quantity of money and thus raise prices by coining both gold and silver at a fixed price ratio (so that the monetary authority would have the daunting task of fixing the relative price between two metals, a problem that Marshall's symmetallism avoided). J. Laurence Laughlin, a hard-money opponent of bimetallism and an important influence on the design of the Federal Reserve System, responded by rejecting the quantity theory of money, emphasizing the cost of production of gold rather than the quantity of money (Skaggs, 1995). Irving Fisher (1896, 1911) offered a theoretical and statistical defense of the quantity theory (see also Kemmerer, 1907), with changes in the quantity of money determining the price level and short-run real fluctuations, but, contrary to the bimetallists, with money neutral in the long run (Dimand, 2013). Fisher (1896), equating the difference between interest rates in any two standards (gold and silver, money and commodities, two national currencies) to the exchange rate of appreciation of one standard against the other, showed that the price deflation (appreciation of the purchasing power of money) only burdened debtors if the deflation was unexpected. If the deflation was expected, it would be reflected in lower interest rates. With acknowledgment to Marshall's evidence before official inquiries (later collected by Keynes as Marshall, 1926), Fisher (1896) introduced the Fisher equation (nominal interest equal to real interest plus expected inflation), the expectations theory of the term structure of interest rates, and uncovered interest parity between currencies. Covered interest parity, equating the spread between spot and forward exchange rates to the difference between nominal interest rates in two currencies, was added by Keynes (1923) (see Flanders, 1989 on developments in open-economy monetary economics). Modeling expected inflation with a distributed lag of past price level changes, Fisher (1911: Chapter IV, 1926) attributed fluctuations in real economic activity and employment to slow and incomplete adjustment of inflation expectations to monetary shocks, a phenomenon that in a 1928 book title Fisher termed 'the money illusion.' Influenced by Fisher (the only economist he had read at the time of writing Hawtrey, 1913), Ralph Hawtrey of the British Treasury also advanced a monetary theory of the trade cycle (Hawtrey, 1913, 1919, 1932; Deutscher, 1990). From the concluding chapter of Fisher (1911) onward, Fisher advocated a 'compensated dollar,' seeking to stabilize the real economy by stabilizing the purchasing power of money as measured by a price index. To try to make this proposal operational, Fisher identified the geometric mean of the Paasche and Laspeyres indexes as the ideal index number, published a weekly price index, and pioneered the issue of indexed bonds.

While Fisher and Hawtrey stressed monetary shocks and advocated price level stabilization, Joseph Schumpeter ([1912]1934) and Dennis Robertson (1915) saw clusters of entrepreneurial innovations, introducing new products or techniques, as the real shocks driving economic fluctuations (see Presley, 1979 and Fletcher, 2000 on Robertson, and Swedberg's introduction to Schumpeter, 1991). Pigou (1927) argued that waves of optimism and pessimism caused fluctuations in investment. The Stockholm School (Karin Kock, 1929; Erik Lindahl, 1939; Erik Lundberg, 1937; Gunnar Myrdal, 1939) followed Knut Wicksell (1898, 1907, 1915) in emphasizing the distinction between the market rate of interest, set by the banking system, and the natural rate of interest, variously defined as the interest rate at which the price level would be constant, or at which investment equals saving, or which reflects the marginal productivity of capital. Monetary shocks shift the market rate of interest, while innovations and other real productivity shocks change the natural rate. As long as the market rate was less than the natural rate, entrepreneurs would borrow for new investment, causing prices to rise, and the converse occurred when the market rate was above the natural rate. Monetary policy could stabilize the economy by varying the market rate to match changes in the natural rate (which was not directly observable). The process analysis of the Swedish economists, of Robertson (1926), and of Keynes' *Treatise on Money* (1930), while motivated by concern with fluctuations in output and employment, formally modeled price level movements and interest rates, with no explicit theory of unemployment or aggregate income (Patinkin, 1982; Jonung, 1991). Beginning with the first German edition of Mises (1935) in 1912, the Austrian monetary malinvestment theory of the trade cycle drew on the Currency School tradition and on Wicksell to interpret depressions as the necessary corrections of unsound speculative investment resulting from expansionary monetary disturbances, and to reject price level stabilization as interfering with and prolonging necessary deflationary adjustments (Mises, 1935; Hayek, 1931, 1933; Ellis, 1934).

Fluctuations as Cycles

Although they were not the first to detect some more or less regular periodicity in the pattern of crises and booms (Jevons, 1884: p. 129, cited publications by Hyde Clarke in 1847), the idea that trade and prices follow a cycle gained force in the 1860s with the empirical work of Clément Juglar (1862) and W. Stanley Jevons (1884) (a collection of his articles from 1862 to 1882 edited after Jevons' death by Herbert S. Foxwell, his successor at University College), with later writers from London School of Economics (LSE) director Sir William Beveridge to NBER founder Wesley Mitchell (1927) and Joseph Schumpeter (1939) going beyond a single cycle to interpret time series as several superimposed cycles of differing amplitude and periodicity. The literature was immense (O'Brien (1999), Hagemann (2001), Boianovsky (2005) offer 11 volumes of selections from the primary sources) and, following the establishment of Wesley Mitchell's NBER in the United States and the London and Cambridge Economic Service and the National Institute of Economic and Social Research in Britain, conjuncture or business cycle research institutes flourished in Vienna, Berlin, Kiel, Sofia, Warsaw, Moscow, and other European centers in the 1920s (the Austrian Institute for Business Cycle Research became the Austrian Institute for Economic Research after World War II).

Jevons' attribution of the 11-year trade cycle to the 11-year sunspot cycle, the periodic oscillation of solar activity, was responsible for economics being allowed to remain in the British Association for the Advancement of Science, by offering an apparently objective, scientific physical explanation for economic fluctuations, but in modern macroeconomics, the term 'sunspots' refers to any intrinsically irrelevant variable. Jevons' hypothesis, while mistaken, was not unreasonable: meteorologists of the time believed that sunspots affected the weather: cycles in rainfall and temperature would cause cycles in agricultural harvests, and agriculture accounted for a larger share of the world economy in the 1860s than it does now. When sunspots were believed to follow a cycle of 11.1 years, Jevons founds an 11-year cycle in prices, "and when the sun-spot cycle was revised to 10.45 years he was able to make the average interval between English crises 10.466 years" (Mitchell, 1927: p. 384). Jevons assumed that a cyclical effect must have a cyclical cause, but Frisch (1933) drew attention to Wicksell's remark that when a rocking horse is struck with a stick, the motion of the rocking horse does not resemble the motion of the stock. Seasonality, Jevons' other discovery of an economic cycle with a cyclical physical cause (also in Jevons, 1884), has fared better, and is the progenitor of seasonally adjusted time series.

With enough cycles, any time series could be 'explained' as the summation of cycles: using the periodogram (an early version of spectral analysis), Beveridge found 19 cycles in wheat prices with periods varying from 2.735 to 68 years (Dimand, 1999). In a paper first published in Russian in 1927 with an English summary, Eugen (or Evgenii) Slutsky (1937) used a moving average of the last three digits of the winning Moscow lottery numbers to demonstrate that summation and averaging of random series would produce apparent cycles where none existed. Slutsky (1937) and Ragnar Frisch (1933), and Fisher's critique of "the so-called 'business cycle'" as a "dance of the dollar," turned the attention of economists to oscillatory responses to external shocks (real or monetary), rather than underlying cycles (see Andvig, 1981 on Frisch and V. Barnett, 2011 on Slutsky). The Moscow Conjuncture Institute led by Nikolai Kondratiev, whose long wave theory implied that capitalism would experience another upswing after the Depression, was destroyed in Stalin's purges, with Slutsky retreating from economic topics to meteorology and statistics – ironically including research on sunspots (Kondratiev, 1997; V. Barnett, 1998, 2011). Other business cycle institutes survived but, as in the case of Mitchell's NBER, evolved into general economic research centers, with seasonality and leading indicators as the most specific lasting contributions of business cycle analysis. More generally, Christopher Sims' VAR approach to macroeconomics, starting with time series analysis without assuming much knowledge of the true underlying theoretical model, is in the spirit of the statistical approach to business cycle analysis undertaken by Juglar, Mitchell, Beveridge, and others.

Dynamics and Growth

David Ricardo, Thomas Robert Malthus, and John Stuart Mill analyzed the movement of profit rate, real wages, rent, and output as economies moved toward a stationary state in which the wage would equal the 'natural price of labor' and both population and the capital stock would be constant until the next innovation changed agricultural productivity (Hoselitz, 1960; Eltis, 1984; Hollander, 1997; Brewer, 2010). More formal algebraic modeling of economic growth began tentatively with Sismondi and then, drawing inspiration from Quesnay's *Tableau Économique*, more systematically with Karl Marx's schemes of simple and expanded reproduction in Volume II of *Capital* (Arthur and Reuten, 1998), which were developed further in debates between Rosa Luxemburg ([1913]1951) and Otto Bauer about whether the capitalist mode of production would inevitably break down after a certain number of periods because of insufficient purchasing power and sectoral imbalances (see Kalecki, 1967 and Joan Robinson's introduction to the 1951 translation of Luxemburg).

In his presidential address to Section F of the British Association, Allyn Young (1928), recently transplanted from Harvard to a chair at LSE, directed attention to Adam Smith's linkage of economic growth to increasing returns to scale ("the division of labour is limited by the extent of the market"), ironically just 2 years after Piero Sraffa had, in the same journal, shown that increasing returns are problematic for modeling equilibrium of the Marshallian firm. In the same journal issue as Young (1928), the Cambridge mathematician and philosopher Frank Ramsey (1928) analyzed optimal capital accumulation (with the first use of calculus of variations in an economics journal) in the first step toward what is now taught as the Ramsey–Cass–Koopmans model, and the same year G.A. Fel'dman (1928) attempted a formal model of capital intensity and output growth rates for Soviet central planning. But by 1930, Young and Ramsey were

dead (Ramsey at the age of 26), and Fel'dman had disappeared in Stalin's purges, and the lines of research they indicated were not pursued until much later.

From Cycles and Price Dynamics to Analysis of Aggregate Output and Employment

Although both monetary theorists and business cycle analysts were motivated by concern over periods of high unemployment (as well as by the social costs of rapid inflation, with Keynes' *Tract on Monetary Reform* and Fisher stressing the collapse of monetary exchange in the German hyperinflation of the early 1920s), they lacked a macroeconomic analysis of unemployment. Pigou (1913) talked about 'effective demand' in an early short book for the general public about unemployment, but his more formal and systematic *Theory of Unemployment* (1933) offered only a one-market analysis in which unemployment just meant that the wage rate was above the market clearing level. The implication of Pigou's analysis, which Pigou steadfastly refused to draw in his advice and writings on policy, was that wage cuts would eliminate unemployment even in the Great Depression. But Edwin Cannan did draw that conclusion in his 1932 presidential address to the Royal Economic Society (reprinted in Dimand, 2002: vol. III). Beveridge's *Unemployment, a Problem of Industry (1909 and 1930)* (1930), an expansion of his 1909 book, saw unemployment in search-theoretic terms, with labor exchanges as the means to more efficient matching of workers and jobs. Fisher (1933) presented a debt deflation theory of depressions, stressing the scramble for liquidity because of fears of default on inside debts with fixed money values, which later influenced Hyman Minsky (1975), James Tobin (1980), Ben Bernanke (2000), and Mervyn King, but at that time he had lost his audience by being so memorably wrong about the stock market. The mass unemployment of the early 1930s turned public attention from such academic authorities to underconsumptionists outside the mainstream (see Bleaney, 1976) or to Marxist claims that depressions and mass unemployment were inherent in capitalism. Schumpeter held that the Great Depression was just the regrettable coincidence of the troughs of three cycles of different lengths.

John Maynard Keynes, in his pamphlet *The Means to Prosperity* (1933b) and his 1933 Cambridge lectures (reconstructed in Rymes, 1989), and then more fully and formally in *The General Theory of Employment, Interest and Money* (1936), offered a solution to this impasse. Contrary to the Marxists, depressions and unemployment could be tamed without revolution or abolition of capitalism, by appropriate management of aggregate demand to offset fluctuations in expectations of the long-period profitability of new investment, fluctuations resulting from fundamental uncertainty about the future. Unlike John Hobson or other underconsumptionists, or Michal Kalecki writing in Polish within a Marxian tradition (Kalecki, 1967; translations of essays written in the 1930s), Keynes was a well-established figure within the profession, steeped in the Cambridge heritage of Marshallian monetary theory and saving-investment analysis (Eshag, 1963; Bridel, 1987; Laidler, 1999). Drawing on the multiplier analysis of his student Richard Kahn (1931), who together with Giblin (1930), Hawtrey (1932), and others had related changes in equilibrium income to changes in autonomous investment or exports (see Davis, 1980; Patinkin, 1982; Clarke, 1988; Dimand, 1988), Keynes (1933b, 1936) offered a theory of the equilibrium level of national income as the level of income at which desired investment equals desired saving. Noting that nominal wages had fallen by 30% in the United States in the early 1930s without reducing unemployment or real wages, Keynes (1936: Chapter 19) invoked dynamic arguments to deny that wage and price flexibility would guarantee restoration of full employment after a demand shock – sticky wages were his recommendation, not his assumption (see also Minsky, 1975; Tobin, 1980). Not only did Keynes have a message welcomed by policy makers and the public, but, combined with the creation of national income accounts and advances in econometric technique, his theory of the determination of aggregate income and employment provided an agenda for macroeconomists, drawing together the traditions of monetary theory and empirical studies of economic fluctuations. Regardless of later serious doubts by macroeconomists about the shift of emphasis from inflation to unemployment, or the eclipse of search-theoretic analysis of unemployment, or Keynes' sanguine view of the motivation of policy makers, the theory of equilibrium aggregate income and employment sketched in Keynes (1933b) and in his 1933 lectures (in Rymes 1989) and pamphlet and presented in the *General Theory* marked a transformation of a field that had a long and rich history by the time it acquired the name macroeconomics.

See also: Economics, History of; Financial Crises, History of; Fisher, Irving (1867–1947); Hume, David (1711–76); Keynes, John Maynard (1883–1946); Keynesianism and the Question of State Interventionism; Macroeconomics, History of From 1933 to Present; Malthus, Thomas Robert (1766–1834); Marshall, Alfred (1842–1924); Marx, Karl (1818–83); Mill, John Stuart (1806–73); Myrdal, Gunnar (1898–1987); Ricardo, David (1772–1823); Robinson, Joan (1903–83); Schumpeter, Joseph A. (1883–1950); von Hayek, Friedrich A. (1899–1992).

Bibliography

Andvig, J.C., 1981. Ragnar Frisch and business cycle research during the Interwar years. History of Political Economy 13, 695–725.

Arnon, A., 2011. Monetary Theory and Policy from Hume and Smith to Wicksell: Money, Credit and the Economy. Cambridge University Press, Cambridge, UK.

Arthur, C.J., Reuten, G. (Eds.), 1998. The Circulation of Capital: Essays on Volume II of Marx's "Capital". Macmillan, London.

Bagehot, W., 1873. Lombard Street, a Description of the Money Market. John Murray, London. Reprinted in: Stevas, N.St.J. (Ed.), The Collected Works of Walter Bagehot, The Economist 1974–1986, London.

Barnett, P., 1941. Business-Cycle Theory in the United States, 1860–1900. University of Chicago Press, Chicago.

Barnett, V., 1998. Kondratiev and the Dynamics of Economic Development: Long Cycles and Industrial Growth in Historical Context. Macmillan, Basingstoke.

Barnett, V., 2011. E.E. Slutsky as Economist and Mathematician: Crossing the Limits of Knowledge. Routledge, London and New York.

Baumol, W.J., 1999. Retrospectives: say's law. Journal of Economic Perspectives 13, 195–204.

von Bergmann, E., 1895. Die Wirtschaftskrisen: Geschichte der nationalökonomischen Krisentheorien. W. Kohlhammer, Stuttgart.

Bernanke, B.S., 2000. Essays on the Great Depression. Princeton University Press, Princeton, NJ.

Beveridge, W.H., 1930. Unemployment, a Problem of Industry (1909 and 1930). Longmans, Green, London.

Blaug, M., Eltis, W., O'Brien, D.P., Patinkin, D., Skidelsky, R., Wood, G.E., 1995. The Quantity Theory of Money from Locke to Keynes and Friedman. Edward Elgar, Aldershot, UK; Brookfield, VT.

Bleaney, M., 1976. Underconsumption Theories: A History and Critical Analysis. International Publishers, New York.

Boianovsky, M. (Ed.), 2005. Business Cycle Theories: Selected Texts 1860–1939, vols. 5–8. Pickering & Chatto, London.

de Boyer des Roches, J., 2003. La pensée monétaire: Histoire et analyse. Les Solo, Paris.

Brewer, A., 2010. The Making of the Classical Theory of Economic Growth. Routledge, London and New York.

Bridel, P., 1987. Cambridge Monetary Thought: The Development of Saving-Investment Analysis from Marshall to Keynes. Macmillan, Basingstoke.

Bridel, P., 1997. Money and General Equilibrium Theory: From Walras to Pareto (1870–1923). Edward Elgar, Cheltenham, UK.

Cesarano, F., 2007. Monetary Theory in Retrospect: The Selected Essays of Filippo Cesarano. Routledge, London and New York.

Christiernin, P.N., 1761. Lectures on the high price of foreign exchange in Sweden (1967 Trans.). In: Eagly, R.V. (Ed.), The Swedish Bullionist Controversy. American Philosophical Society, Philadelphia.

Clarke, P., 1988. The Keynesian Revolution in the Making, 1924–1936. Cambridge University Press, Cambridge, UK.

Corry, B.A., 1962. Money, Saving and Investment in English Economics 1800–1850. Macmillan, London.

Davis, E.G., 1980. The correspondence between R.G. Hawtrey and J.M. Keynes on the *Treatise*: the genesis of output adjustment models. Canadian Journal of Economics 13, 716–724.

Davis, T., 2005. Ricardo's Macroeconomics: Money, Trade Cycles and Growth. Cambridge University Press, Cambridge, UK.

Deutscher, P., 1990. R.G. Hawtrey and the Rise of Macroeconomics. Macmillan, Basingstoke.

Dimand, R.W., 1988. The Origins of the Keynesian Revolution. Edward Elgar, Upleadon, UK. Stanford University Press, Stanford, CA.

Dimand, R.W., 1999. Beveridge on unemployment and cycles before. The General Theory. History of Economic Ideas 7 (3), 33–51.

Dimand, R.W. (Ed.), 2002. The Origins of Macroeconomics, 10 vols. Routledge, London and New York.

Dimand, R.W., 2003. Interwar monetary and business cycle theory. Research in History of Economic Thought and Methodology 21-A, 121–148.

Dimand, R.W., 2013. David Hume and Irving Fisher on the quantity theory of money in the short run and the long run. European Journal of the History of Economic Thought 20, 284–304.

Divisia, F., 1925. L'indice monétaire et la théorie de la monnaie. Revue d'Économie Politique 39, 980–1009.

Ellis, H.S., 1934. German Monetary Theory 1905–1933. Harvard University Press, Cambridge, MA.

Eltis, W.A., 1984. The Classical Theory of Economic Growth. Macmillan, London.

Eshag, E., 1963. From Marshall to Keynes: An Essay on the Monetary Theory of the Cambridge School. Basil Blackwell, Oxford.

Fel'dman, G.A., 1928. On the theory of the growth rates of national income I and II (1964 Trans.). In: Spulber, N. (Ed.), Foundations of Soviet Strategy for Economic Growth. Indiana University Press, Bloomington, pp. 174–199 and 304–331. Reprinted in Dimand (2002), vol. III.

Fetter, F.W., 1965. The Development of British Monetary Orthodoxy 1797–1875. Harvard University Press, Cambridge, MA.

Fisher, I., 1896. Appreciation and Interest. Macmillan, New York. In: Fisher (1997), vol. 1.

Fisher, I., 1911. The Purchasing Power of Money, with H.G. Brown. Macmillan, New York. In: Fisher (1997), vol. 4.

Fisher, I., 1926. A statistical relationship between unemployment and price level changes. International Labour Review 13, 785–792. Reprinted 1973 as, Lost and found: I discovered the Phillips curve – Irving Fisher. Journal of Political Economy 81, 496–502; Also in: Fisher (1997), vol. 8.

Fisher, I., 1933. The debt-deflation theory of great depressions. Econometrica 1, 337–357. In: Fisher (1997), vol. 10.

Fisher, I., 1997. In: Barber, W.J. (Ed.), assisted by Dimand, R.W., Foster, K.; Tobin, J. (Consulting Ed.). The Works of Irving Fisher, 14 vols. Pickering & Chatto, London.

Flanders, M.J., 1989. International Monetary Economics 1870–1960. Cambridge University Press, Cambridge, UK.

Fleetwood, W., 1707. Chronicon Preciosum. London, second ed. 1745, reprinted. A.M. Kelley, New York.

Fletcher, G., 2000. Understanding Dennis Robertson. Edward Elgar, Cheltenham, UK; Northampton, MA.

Foxwell, H.S., 1886. Irregularity of Employment and Fluctuations of Prices. Co-Operative Printing Company, Edinburgh. Reprinted in: Dimand (2002), vol. I.

Friedman, M. (Ed.), 1956. Studies in the Quantity Theory of Money. University of Chicago Press, Chicago.

Frisch, R., 1933. Propagation problems and impulse problems in dynamic economics. In: Economic Essays in Honour of Gustav Cassel. George Allen & Unwin, London.

Gervaise, I., 1720. The system or theory of the trade of the world. H. Woodfall, London; reprinted. with foreword by J. Viner and introduction by J.M. Letiche. Johns Hopkins Press, Baltimore. 1954. In: Dimand, R.W. (Ed.), The Origins of International Economics, vol. I. Routledge, London and New York, 2004.

Giblin, L.F., 1930. Australia, 1930. Melbourne University Press, Melbourne. Reprinted in: Dimand (2002), vol. IV.

Glasner, D. (Ed.), 1997. Business Cycles and Depressions: An Encyclopedia. Garland, New York.

Green, R., 1992. Classical Theories of Money, Output and Inflation. Macmillan, London and Basingstoke.

Grice-Hutchison, M., 1952. The School of Salamanca: Readings in Spanish Monetary Theory 1544–1605. Clarendon Press, Oxford.

Guggenheim, T., 1989. Preclassical Monetary Theories. Pinter, London and New York.

Hagemann, H. (Ed.), 2001. Business Cycle Theories: Selected Texts 1860–1939, 4 vols. Pickering & Chatto, London.

Hawtrey, R.G., 1913. Good and Bad Trade. Constable, London. Reprinted A. M. Kelley, New York, 1970, with 1962 preface by author.

Hawtrey, R.G., 1919. Currency and Credit. Longmans, Green, London.

Hawtrey, R.G., 1932. The Art of Central Banking. Longmans, Green, London.

Hayek, F.A., [1931]1991. English monetary policy and the bullion debate. In: Hayek, F.A., The Trend of Economic Thinking. University of Chicago Press, Chicago, (Chapters 9–12).

Hayek, F.A., 1931. Prices and Production, second ed. 1935. Routledge, London.

Hayek, F.A., 1933. Monetary theory and the trade cycle (N. Kaldor and H. M. Croome, Trans.). Routledge, London.

Hegeland, H., 1951. The Quantity Theory of Money. Goteborg, reprinted 1969. A.M. Kelley, New York.

Hollander, S., 1997. The Economics of Thomas Robert Malthus. University of Toronto Press, Toronto.

Hoselitz, B.F. (Ed.), 1960. Theories of Economic Growth. Free Press, New York.

Humphrey, T.M., 1993. Money, Banking, and Inflation: Essays in the History of Economic Thought. Edward Elgar, Aldershot, UK; Brookfield, VT.

Jevons, W.S., 1884. In: Foxwell, H.S. (Ed.), Investigations in Currency and Finance. Macmillan, London.

Jones, E.D., 1900. Economic Crises. Macmillan, New York.

Jonung, L. (Ed.), 1991. The Stockholm School of Economics Revisited. Cambridge University Press, Cambridge, UK.

Juglar, C., 1862. Des crises commerciales et de leur retour périodique en France, en Angleterre et aux États-Unis. Guillaumin, Paris.

Kahn, R.F., 1931. The relation of home investment to unemployment. Economic Journal 41, 173–198.

Kalecki, M., 1967. Studies in the Theory of Business Cycles, 1933–1939. Basil Blackwell, Oxford.

Kemmerer, E.W., 1907. Money and Credit Instruments in Their Relation to General Prices. Henry Holt, New York.

Keynes, J.M., 1923. A Tract on Monetary Reform. Macmillan, London. In: Keynes (1971–1989), vol. IV.

Keynes, J.M., 1930. A Treatise on Money, 2 vols. Macmillan, London. In: Keynes (1971–1989), vols. V and VI.

Keynes, J.M., 1933a. Robert Malthus, the first of the Cambridge economists. In: Keynes, J.M. (Ed.), Essays in Biography. Macmillan, London. Reprinted in: Keynes (1971–1989), vol. 10.

Keynes, J.M., 1933b. The Means to Prosperity, vol. 9. Harcourt Brace, New York. Reprinted in: Keynes (1971–1989).

Keynes, J.M., 1936. The General Theory of Employment, Interest and Money. Macmillan, London. In: Keynes (1971–1989), vol. VII.

Keynes, J.M., 1971–1989. In: Moggridge, D.E. (Ed.), The Collected Writings of John Maynard Keynes, 30 vols. Macmillan, London. Cambridge University Press, for the Royal Economic Society, New York.
Kock, K., 1929. A Study of Interest Rates, Stockholm Economic Studies No. 1. P.S. King, London.
Kondratiev, N.D., 1997. In: Samuels, W.J., Makasheva, N. (Eds.), Barnett, V., Reijnders, J., Solomou, S., Tylecote, A. (Consulting eds.), The Works of Nikolai D. Kondratiev, 4 vols. Pickering & Chatto, London.
Korinek, A., Mendoza, E., 2013. From Sudden Stops to Fisherian Deflation: Quantitative Theory and Policy Implications. NBER Working Paper 19362, August.
Laidler, D.E.W., 1991a. The Golden Age of the Quantity Theory. Phillip Allan, London; Princeton University Press, Princeton, NJ.
Laidler, D.E.W., 1991b. The quantity theory is always and everywhere controversial – why? Economic Record 67, 289–306.
Laidler, D.E.W., 1999. Fabricating the Keynesian Revolution: Studies of the Inter-war Literature on Money, the Cycle, and Unemployment. Cambridge University Press, Cambridge, UK.
Laidler, D.E.W., 2003. Macroeconomics in Retrospect: Selected Essays. Edward Elgar, Cheltenham, UK; Northampton, MA.
Lange, O., 1942. Say's law: a restatement and critique. In: Lange, O., McIntyre, F., Yntema, T.O. (Eds.), Studies in Mathematical Economics and Econometrics in Memory of Henry Schultz. University of Chicago Press, Chicago, pp. 49–68. Reprinted in: Dimand (2002), vol. III.
Lindahl, E., 1939. Studies in the Theory of Money and Capital. George Allen & Unwin, London.
Link, R., 1959. English Theories of Economic Fluctuations 1815–1848. Columbia University Press, New York.
Locke, J., 1991. In: Kelley, P.H. (Ed.), Locke on Money, 2 vols. Clarendon Press, Oxford.
Lundberg, E., 1937. Studies in the Theory of Economic Expansion. P.S. King, London.
Luxemburg, R., [1913]1951. The Accumulation of Capital (A. Schwarzchild, Trans. with introduction by J. Robinson). Routledge & Kegan Paul, London.
Marshall, A., 1887. Remedies for Fluctuations in General Prices. Contemporary Review. Reprinted in: Pigou, A.C. (Ed.), Memorials of Alfred Marshall. Macmillan, London, 1925.
Marshall, A., 1923. Money, Credit and Commerce. Macmillan, London.
Marshall, A., 1926. In: Keynes, J.M. (Ed.), Official Papers. Macmillan, London.
Minsky, H.P., 1975. John Maynard Keynes. Columbia University Press, New York.
Mises, L. von, 1935. The Theory of Money and Credit (H. Batson, Trans.). Cape, London.
Mitchell, W.C., 1913. Business Cycles. University of California Press, Berkeley.
Mitchell, W.C., 1927. Business Cycles: The Problem and Its Setting. National Bureau of Economic Research, New York.
Monroe, A.E., 1923. Monetary Theory before Adam Smith. Harvard University Press, Cambridge, MA.
Myrdal, G., 1939. Monetary Equilibrium (R. Bryce and N. Stolper, Trans.). W. Hodge, London.
Murphy, A.E., 2009. The Genesis of Macroeconomics: New Ideas from Sir William Petty to Henry Thornton. Oxford University Press, Oxford.
Northcote, S., (Lord Iddlesleigh), Chair, 1886. Depression of Trade and Industry. Final Report of the Royal Commission Appointed to Inquire into the Depression of Trade and Industry; with Minutes of Evidence and Appendices. Parliamentary Papers, Command Paper C. 4893, London.
O'Brien, D.P. (Ed.), 1999. Foundations of Business Cycle Theory, 3 vols. Edward Elgar, Cheltenham, UK; Brookfield, VT.
O'Brien, D.P., 2007. The Development of Monetary Economics. Edward Elgar, Cheltenham, UK; Northampton, MA.
Oresme, N., de Sassoferrato, B., Buridan, J., 1989. In: Dupuy, C. (Ed.), Traité des monnaies et autres écrits monétaires du XIVe siècle (F. Chartrain, Trans.). La Manufacture, Lyon.
Patinkin, D., 1965. Money, Interest and Prices, second ed. Harper & Row, New York.
Patinkin, D., 1981. Essays on and in the Chicago Tradition. Duke University Press, Durham, NC.
Patinkin, D., 1982. Anticipations of the General Theory? And Other Essays on Keynes. University of Chicago Press, Chicago.
Pigou, A.C., 1913. Unemployment. Williams & Norgate, Home University Library, London.
Pigou, A.C., 1927. Industrial Fluctuations. Macmillan, London.
Pigou, A.C., 1933. Theory of Unemployment. Macmillan, London.
Presley, J., 1979. Robertsonian Economics. Holmes & Meier, New York.
Ramsey, F.P., 1928. Mathematical theory of saving. Economic Journal 38, 543–559.
Robertson, D.H., 1915. A Study of Industrial Fluctuation. P.S. King, London.
Robertson, D.H., 1926. Banking Policy and the Price Level. P.S. King, London.
Rotwein, E. (Ed.), 1955[1752]. David Humes: Writings on Economics. University of Wisconsin Press, Madison, WI.
Rymes, T.K. (Ed.), 1989. Keynes's Lectures, 1932–35, Notes of a Representative Student. Macmillan, Basingstoke. University of Michigan Press, Ann Arbor.
Schumpeter, J.A., [1912]1934. The Theory of Economic Development, (R. Opie, Trans.). Harvard University Press, Cambridge, MA.
Schumpeter, J.A., 1939. Business Cycles, 2 vols. McGraw Hill, New York.
Schumpeter, J.A., 1991. In: Swedberg, R. (Ed.), The Economics and Sociology of Capitalism. Princeton University Press, Princeton, NJ.
Skaggs, N.T., 1995. The methodological roots of J. Laurence Laughlin's anti-quantity theory of money and prices. Journal of the History of Economic Thought 17, 1–20.
Skaggs, N.T., 1999. Changing views: twentieth century opinion on the banking school-currency school controversy. History of Political Economy 31, 361–391.
Slutsky, E.E., 1937. The summation of random causes as the source of cyclic processes. Econometrica 5, 105–146.
Smith, V.C., 1936. The Rationale of Central Banking. P.S. King, London.
Sowell, T., 1972. Say's Law: A Historical Analysis. Princeton University Press, Princeton, NJ.
Thornton, H., [1802]1939. An Enquiry into the Nature and Effects of the Papers Credit of Great Britain, with two 1811 speeches by Thornton, and introduction by F.A. Hayek. George Allen & Unwin, London.
Tobin, J., 1980. Asset Accumulation and Economic Activity. Basil Blackwell, Oxford. University of Chicago Press, Chicago.
Vickers, D., 1959. Studies in the Theory of Money 1690–1776. Chilton, New York.
Viner, J., 1937. Studies in the Theory of International Trade. Harper & Brothers, New York.
Walras, L., 1886. Théorie de la monnaie. Larose et Forcel, Paris. L. Corbaz, Lausanne.
Wicksell, K., 1898. Interest and Prices (R. Kahn, 1936, Trans.). Macmillan, London.
Wicksell, K., 1907. The influence of the rate of interest on prices. Economic Journal 17, 213–220.
Wicksell, K., 1915 (E. Claasen, 1935, Trans.). Lectures on Political Economy, vol. 2. Routledge, London.
Wirth, M., 1858. Geschichte der Handelskrisen. Frankfurt am Main, fourth ed. 1890.
Wood, G., Capie, F. (Eds.), 2000. The Development of Monetary Theory in the 1920's and 1930's, 4 vols. Routledge, London and New York.
Woodford, M., 2003. Interest and Prices. Princeton University Press, Princeton, NJ.
Wright, C.D., 1886. Industrial Depressions: First Annual Report of the United States Commissioner of Labor. Government Printing Office, Washington, DC. Reprinted A.M. Kelley, 1968, New York.
Young, A.A., 1928. Increasing returns and economic progress. Economic Journal 38, 527–542.

Macrosociology-Microsociology

Gianluca Manzo, GEMASS–CNRS and University Paris-Sorbonne, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Cherkaoui, volume 13, pp. 9117–9122, © 2001, Elsevier Ltd.

Abstract

The history of sociology shows that the distinction between macro- and microsociology is at the same time fundamental and fundamentally ambiguous. This article presents a selective overview of the solutions to the micro/macro issue that have appeared in sociology from the classics to today's developments in philosophy of social sciences and sociology. The article aims to highlight the main sources of disagreement and division among sociologists interested in the micro/macro issue, and in the end stresses the importance of mechanism-based theorizing and formal methods to advance it.

Introduction

When the history of sociological theorizing has to be systematized, the distinction between micro- and macrosociology is fundamental (see, for instance, Collins, 1988; Cherkaoui, 2005: Chapter 5). But what are micro- and macrosociology? According to Randall Collins (1981: 984), "microsociology is the detailed analysis of what people do, say, and think in the actual flow of momentary experience. Macrosociology is the analysis of large-scale and long-term social processes, often treated as self-subsistent entities such as 'state,' 'organization,' 'class,' 'economy,' 'culture,' and 'society.'" While this broad characterization of micro- and macrosociology is largely undisputed (see for instance, Calhoun et al., 2012: 27), it seems problematic to provide a more precise definition. In fact, as noted by Münch and Smelser (1987: 356), "(…) the terms 'micro' and 'macro' have been assigned a number of diverse meanings in the sociological literature and these meanings are not always consistent with one another." Along similar lines, Wippler and Lindenberg (1987: 153) admitted that "there is no agreement on the micro/macro distinction, except that 'micro' always refers to smaller units than those implied by macro. (…) the various meanings attached to this distinction have generated micro/macro problems that stand in the way of an adequate solution of the master problem." Thus, the distinction between macro- and microsociology is at once fundamental and fundamentally ambiguous (as an additional proof, see Table 1 in Krause, 2013: 142).

In this short article, I first review solutions to the micro/macro issue that appeared in sociology from the classics to the 1980s. Without any claim to exhaustiveness, I select typical, in the sense of recurrent, solutions to suggest that (1) sociologists have often opted for a substantive understanding of the notions micro/macro, thus equating a given level of analysis with a given set of entities or phenomena; (2) sociologists have recurrently attempted to give one level explanatory priority over another; (3) in order to justify their 'priority' claims, sociologists have often mixed ontological and explanatory arguments (for an early statement along similar lines, see Wiley, 1988). In my opinion, these are the main sources from which the disagreements on the micro/macro issue arise. Then, by discussing in-progress epistemological, theoretical, and methodological developments, I suggest that such disagreements can be overcome provided sociology renounces the vertical, multilayer metaphor contained in the notions of micro and macro, and puts all its energy into the detailed description of the mechanisms that constantly transform small-scale heterogeneity into large-scale, time-resistant mixtures of homogenization and differentiation, whatever level the entities under study may be assigned on the low(micro) and high(macro) continuum.

Micro/Macro Balance

Although "the effort of classic sociology in no sense could be seen as having resolved the micro-macro polarization" (Alexander and Giesen, 1987: 19), the closest readers have recognized that classical sociologists were balanced micro/macro thinkers (Ritzer, 2008: 500). To see this quality, Boudon (1995) argued, one needs to inspect the complex interplay of programatic statements, methodological reflection, and empirical-oriented analysis in classical work.

Durkheim is a paradigmatic example. In the first chapter of *The Rules of Sociological Methods*, he defended the view that social structures have a reality in themselves, are external to individuals, and constrain their actions. By analogically building on biological and chemical phenomena like the 'hardness of bronze' or the 'liquidity of water,' Durkheim (1895 [1982]: 39) argued that the collective level arises from the individual one, yet the former is *sui generis*, meaning qualitatively different, from the later: "(…) the whole does not equal the sum of its parts; it is something different, whose properties differ from those displayed by the parts from which it is formed" (Durkheim, 1895[1982]: 128). That is why, according to some commentators (see Sawyer, 2005: Chapter 6), Durkheim should be regarded as the first 'emergentist' in sociology (on modern debates on the concept of emergence, see Bedau, 2003). However, as stressed, among others by Bearman (1991), Collins (2004: 32–40), and Cherkaoui (2005: Chapter 2, 2008: Chapters 1 and 3), this view did not restrain Durkheim from formulating detailed explanations in terms of networks, interactions, and actions of complex high-level outcomes like patterns of suicides or long-term dynamics of system differentiation.

Weber provides a specular case. In the foundational essay *Some Categories of Interpretive Sociology* (1913[1981]: 158), he

defined individuals as the 'atoms' of sociology and advised that sociology should start with individuals' actions and motivations and show how high-level structures and trends ultimately derive from them. However, similarly to Durkheim, Weber's analyses of specific social phenomena yield a more complex picture. For instance, Weber's study of the origin of capitalism as it is articulated in the *General Economic History* (1927[1981]) shows a complex explanatory model, which relates several levels of analysis and focuses on formulating specific hypotheses on mechanisms bridging these levels. Contrary to Coleman's (1987: 154–157, 1990: Chapter 1) criticism of Weber's analysis, which rests on commenting *The Protestant Ethics and The Spirit of Capitalism*, even the micro-to-macro step of Weber's explanation contained detailed analyses of relational and institutional factors that transformed individuals' beliefs into stable macrostructures – a point nicely illustrated by Cherkaoui (2005: Chapter 3).

Other classics pursued in practice this balance between micro- and macrosociology. Elster (1985, 2009), Cherkaoui (2005: Chapter 1), and Edling and Hedström (2009) have shown that, similar to Durkheim and Weber, Marx and Tocqueville were also attentive to the interplay between actions, interactions, organizations, and structures, and formulated specific hypotheses on detailed mechanisms relating several levels of analysis.

Micro/Macro Extremisms

The classical micro/macro balance was lost for a large part of twentieth-century sociology. As stressed by Ritzer (2008: 500), much modern sociological theory has been characterized by "micro and macro extremists – that is, the preeminence of theorists and theories that accord overwhelming power and significance to either the micro or the macro level." To capture this phenomenon, Turner (2001: 4) coined the polemic terms of 'micro-' and 'macrochauvinism.'

Talcott Parsons's structural functionalism played a crucial role in this radicalizing trend. Although in *The Structure of Social Action* (1937), Parsons's focus is on actors, the analytical importance of normative elements is already stressed without ambiguity. These elements will progressively become predominant with Parsons's *The Social System* (1951) and later work in which he studied how supraindividual structures are internalized by individuals and how these structures satisfy specific systemic needs. Parsons's (1954: Chapters 4, 15, and 19) writings on social stratification are exemplary of this approach. According to his interpretation, occupations crystallize different values, and the system provides larger symbolic and material gratifications to those occupations expressing the most important values at a given point in time. The specific (multilevel) mechanisms by which the consensus on these values emerges are left unspecified, however (for a detailed analysis of Parsons's work and more recent developments in 'system' theory, see, for instance, Joas and Knöbl, 2009: Chapters 2–4, 11, and 13).

As noted by Calhoun et al. (2012: 28), the most important 'microchauvinisms' grew in American sociology out of a reaction to Parsons's 'macroextremism.' On the utilitarian side, this is the case of Homans's (1961) defence of psychological reductionism and his project to deduce every supraindividual structure from more elementary forms of exchanges among self-oriented actors within small groups. On the interpretive side, although different when carefully inspected, ethnomethodology (Garfinkel, 1967), symbolic interactionism (Blumer, 1969), and Goffman's (1982) frame analysis share the assumption that supraindividual structures are transient and contingent, depending on the way actors interpret and negotiate the meanings of actions and situations when they interact -Schutz's (1937[1967]) social phenomenology being a power source of inspiration for these approaches.

Let us finally consider the interesting case of shifting from micro- to macroextremism within the work of the same author. Peter Blau is perhaps the most refined example of this trajectory in that he moved from a refined variant of social exchange theory (see Blau, 1964) to an equally sophisticated version of macroextremism focused on how population heterogeneity suffices to predict the patterns of social interactions (see Blau, 1977, and, for a summary, Blau, 1987, 2001).

Micro-Macro Linkage

Compared to this overemphasis on the micro- or the macro-level between the 1930s and the 1970s, sociological theorizing began to develop a more balanced perspective – closer to the spirit of classical empirically oriented analyses (see first section) – in the 1980s. In the introductory essay to *The Micro-Macro Link*, Alexander and Giesen (1987: 31) suggested that "(…) in the present decade a quite different phase of theoretical debate has emerged, one marked by the serious ongoing effort within every theoretical tradition and from both sides of the great divide to link micro- and macro-perspectives (…) we are convinced that the scope and intensity of this search for linkage are without precedent in the history of sociology." Münch and Smelser's (1987: 385) conclusion to the same book is equally explicit: "(…) those who have argued polemically that one level is more fundamental than the other (in some kind of zero-sum way), or who have argued for the complete independence of the two levels, must be regarded as in error."

The effort to reach a multilevel theorizing has been productive in different ways, however. In US sociology, some authors started with the idea that the distinction between micro- and macrolevels is too reductive to capture the complexity of social life. Thus, Alexander (1987) and Ritzer (2008: Figures 1–2 and A.1–3), for instance, suggested that we should instead think in terms of *continua* – in particular, the former proposes an individual/collective and an instrumental/normative *continuum* whereas the latter speaks of a micro/macroscopic and an objective/subjective *continuum*. By combining them, these authors argued, multidimensional analytical schemas can be built and used conceptually to map the variety of phenomena characterizing the social world at different levels of analysis. By contrast, other American sociologists still bet on the foundational value of a specific level of analysis to close the micro/macro gap. Randall Collins (1981, 2004), for instance, building on Durkheim and Goffman, suggested that a particular type of social interactions, i.e., ritual

interactions, that actors experience in a variety of social settings triggers emotions that in turn create a feeling of social belonging, which transforms microlevel disorder into macrolevel order. American students of social networks started with a similar assumption that the intermediate level of social relations is crucial to understand the micro-macro transition (see, for instance, Granovetter, 1973, 1983, 1985). Differently from Collins's qualitative perspective, however, network analysts did not limit themselves to a specific type of interactions. They started instead by empirically describing network-related features impacting on the kind of macrooutcome observed (for an example on the field of diffusion processes, see, for instance, Valente, 1995). Over the decades, social network analysts developed a well-specified set of mechanisms and theories on how microlevel entities and attributes translate into macropatterns (for an overview, see Borgatti and Lopez-Kidwell, 2011).

During 1980s, European sociology also developed original integrative approaches to the micro/macro issue. These approaches insisted on the circularity of 'structure' and 'action,' but they conceptualize this circularity in very different ways. On the one hand, Giddens's (1984) 'structuration theory' and Bourdieu's (1977, 1980[1990]) 'genetic structuralism,' although different in many respects, share the common assumption that there is a relation of codetermination between 'structure' and 'action.' In Giddens, this 'duality' is resolved by postulating that actors' practices dynamically create their own structural conditions, whereas, in Bourdieu, structures are assumed to become incorporated in actors through invisible socialization processes, so that structures silently reappear in the way actors behave and think. On the other hand, the 'structural-individualistic' approach developed by several Dutch sociologists (see Wippler, 1978; Wippler and Lindenberg, 1987) – or the very similar 'recursive' form of methodological individualism proposed by Boudon (1987) – starts with a specific image of actors, which assumes that actors have goals and attach subjective probabilities to different courses of action, and, instead of developing a general account of the relationship between structure and action, advises formulating specific 'bridge' hypotheses (i.e., hypotheses on the way structures impact on actors' goals and opportunities) and 'transformation rules' (i.e., hypotheses on specific institutions and processes that combine actors' action into collective outcomes).

Layered and 'Flat' Social Ontologies

As anticipated by the promoters of the micro/macro-linkage debate themselves (Alexander and Giesen, 1987: 31, 37), the set of solutions that took shape during the eighties did not lead to any widely accepted 'Newtonian synthesis.' In fact two opposite reactions can be detected. Both directly refer to ontology but develop opposite views.

Turner's (2005) reaction is a good example of the first perspective. His objection to both the 'multidimensional' and 'circular' approaches deserves to be quoted extensively: "the micro-macro distinction will not be theoretically resolved by talk about multidimensionality or recourse to tired old dichotomies like agency/structure, rational/nonrational, subjective-objective, action-order and so on. Instead, we need to be much more specific on the properties of the social universe that are denoted by the labels of micro and macro" (Turner, 2005: 407). Thus, according to Turner, to develop a truly multilevel sociological theory requires us to reject the idea that the distinction between micro and macro is purely analytical. One should accept instead that "these levels *are* reality" (Turner, 2005: 409). On this basis, Turner posits the existence of three levels – i.e., 'macro,' 'meso,' and 'micro' – each of which is characterized by specific entities, driven by specific forces, and described by specific lawlike regularities. The task of a truly multilevel sociological theory, Turner (2005: 417) argues, would be "to develop theoretical models and principles explaining the nature of the linkages across levels."

Thus, Turner's reaction to the US micro/macro debate and to the European agency/structure literature is in line with the so-called 'critical realist' perspective, which, as Gorski (2013: 659) puts it, "instead of a purely conventional distinction between 'micro' and 'macro' [it] appeals to the real ontological distinctions between the various layers or 'strata' in the natural and social worlds." Within this framework, the solution to the problem of the micro/macro linkage would come from postulating the existence of a series of levels of reality with 'emergent,' meaning irreducible, properties and seeking to discover the origin of such 'originality' at each level as well as the effects of higher levels on lower levels (what is called 'downward causation' in this perspective). Archer's (1995) 'morphogenetic' approach as well as, more recently, Sawyer's (2005: Chapter 10, in particular) 'emergentist' sociology represent two sophisticated examples of this line of thought.

An equally 'radical' but opposite solution to the problem of linking micro and macro claims that the problem can be advanced by giving up the notion of level itself. This is the option chosen by Latour (1999), for instance. In an attempt to clarify the meaning of his 'actor-network theory,' he states that (1) "(…) the social possesses the bizarre property of not being made of agency and structure at all, but rather of being a circulating entity"; (2) "(…) in the social domain there is no change of scale. It is so to speak flat and folded (…)"; (3) "(…) there is no zoom going from macro structure to micro interactions (…)" (Latour, 1999: 17, 18, 19, respectively). For this reason, 'action-network theory' suggests not thinking in terms of levels but in terms of 'circulations,' 'trajectory,' or 'movements.' A similar, although more sophisticated, position is endorsed by Abbott (2007). Within the context of a discussion on mechanism-based theorizing in sociology (see last section below), he claimed "(…) the Coleman Boat presupposes a whole model of levels that it is the task of my own version of social theory to destroy. There are no levels. The levels concept is incoherent, an uncriticized inheritance from liberal political theory" (Abbott, 2007: 2). As an alternative, he suggests a relational and processual perspective that pays attention to the 'fractal' nature of the social world (see Abbott, 2001: Chapter 6) – interestingly, this 'fractal' nature is something denied by Latour (1999: 18). At the same time, Abbott (2007: 2) honestly admits that we do not really have at the moment a 'serious' theory that allows us to think of the social world with level-free eyes.

Toward a Scale-Based Thinking

It is a fact that the integrative stage of the history of the micro/macro debate was not capable of generating shared solutions. The authors who explicitly characterized the micro/macro distinction in analytical terms and at the same time attempted to integrate several levels of analysis were attacked on several grounds. First, it was observed that some of them tend conceptually to map social reality more than theoretically articulate its dimensions (see Turner's (2005: 407) objection to Alexander's or Ritzer's proposals). Second, concepts like 'duality of structure' or 'habitus' were criticized for restating the problem of structure/agency codetermination instead of providing the details of the interplay between the two (see Ritzer, 2008: 526; Turner, 2005: 406–407; Van den Berg, 1998). Third, it was remarked that, despite the programatic claim, integrative approaches ultimately end up giving one level priority over another, so that, for instance, Collins and Bourdieu finally remain a 'microreductionist' and a 'macrostructuralist,' respectively (see Ritzer, 2008: 507–510, 532). On the other hand, social theorists who overtly proposed a multilayer ontology, thus opting for the notion of 'levels of reality' rather than that of 'level of analysis,' raise the difficult question of the determination of the number, the content, and the frontiers among these levels. A quick comparison of Turner's (2005: 410–411) Table 23.1 with Sawyer's (2005: 211) Figure 10.3, for instance, would suffice to show that these choices are arbitrary, a priori, and ad hoc. Without empirical research, lists of 'levels of reality' are nothing but lists of verbal labels. Unfortunately, and somehow surprisingly, a similar objection also applies to theorists who opted for an analytical understanding of the notion of level in that micro- and macrolevels tend to be in practice equated to specific entities and activities. This is undesirable for two reasons. First, different micro- and macroperspectives usually opt for different micro- and macrocontents, thus generating conceptual ambiguity and confusion. Second, as soon as a specific content is given a specific level, one risks shifting from an analytical to a substantive, ontological characterization of this level, which in turn increases the probability that the discussion will turn to what exists or not in the social world (on this danger, see, among philosophers of social science, Little, 2006).

Recent debates in philosophy of social science seem to open a fruitful avenue to a resolution of these problems. In particular, Ylikoski (2012, 2014) proposes articulating three elements, namely the concept of 'scale,' the distinction between 'constitutive' and 'causal' statements, and the concept of 'mechanism.'

As to the first element, Ylikoski suggests that the distinction between micro and macro could be better described in terms of small- and large-scale phenomena (Ylikoski, 2014: Section 3). To think in terms of scales, he argues, would help avoid the implicit level-based vertical metaphor contained in the notions of micro and macro. The notion of level raises indeed the difficult question of knowing in what sense a social norm, for instance, is a higher level entity than the individuals believing in it. After all, the social norm, to be effective, should reside in the minds of some individuals (at least those who enforce the norm or those who are in charge of having the norm enforced). Thus, in this sense, the norm would be on the same level as the individuals. Within the scale-based framework, this ambiguity disappears; the norm is at a different scale than the actor in that, to be a social norm, it must hold for more than one actor (otherwise, it would be a private normative belief). The difference in terms of level is thus replaced by simpler differences in terms of size. The notion of 'scale' also makes it clear that the distinction between micro and macro, thus between 'levels,' can only be relative and content-free in that it depends on the target one is seeking to describe/explain – if one focuses on individuals interacting within a triad, then transitivity is a large-scale property, but transitivity becomes a small-scale phenomenon when one starts studying the dynamics of interactions within a given firm.

In Ylikoski's argument, scale-based thinking should be combined with an appreciation of the difference between 'constitutive' and 'causal' statements (see Ylikoski, 2012: 2.4; 2014: Section 4). A constitutive statement, he remarks, establishes a relation of dependence between small- and large-scale properties whereas a 'causal' statement describes a relation of dependence between the events that are responsible for the transition between phenomena at different scales. The distinction, Ylikoski argues, is important because it helps avoid the conflation between ontological and explanatory statements. The fact that a given large-scale phenomenon is *made of* smaller scale entities (constitutive viewpoint) does not imply that, to explain the former, one only needs small-scale events. From the point of view of 'causal' statements, the 'difference-makers' are not limited to a single, unique scale.

Within a scale-based framework aimed at answering 'causal' questions, Ylikoski (2012: Sections 2.2 and 2.4.1) finally suggests, the fundamental problem of scientific knowledge becomes the discovery of the 'mechanisms' that transform small-scale properties into larger-scale patterns as well as the mechanisms that clarify the way large-scale properties impact on the activities of smaller scale units. Thus, the problem is no longer what the micro or the macro is, nor what the most fundamental level of analysis is. The problem of micro- and macrosociology is reshaped in more general terms, i.e., as an enterprise of figuring out the details that make the connection between phenomena located at different scales, whatever their specific content.

Recent developments in sociological theory suggest that these ideas are progressively traveling from philosophy of sciences to sociology, and transforming the way sociologists conceive the distinction between micro- and macrosociology. I close this article by reviewing these developments.

Mechanism-Based Thinking

The explicit use of the concept of mechanism, entered sociology in the 1990s (see Hedström and Swedberg, 1998; for an analysis of precursors, see Manzo, 2010: 132–138). Over the last two decades, a distinctive sociological perspective called analytical sociology was elaborated around this concept (see Hedström, 2005). Its foundational axiom is that to explain a connection between events amounts to detailing the set of entities, properties, activities, and

relations among them that brings about the observed connection. Analytical sociology is now a complex intellectual movement that orientates empirical research in several areas (see Hedström and Bearman, 2009; Manzo, 2014a) and stimulates interesting controversies (see Abbott, 2007; Little, 2012; Opp, 2013). Analytical sociology is deeply concerned with the micro/macro issue. In this respect, three elements of this research program must be briefly discussed: (1) the form of methodological individualism that animates analytical sociology; (2) its views on the theory of action; and (3) some of its methodological proposals (for a more general and detailed discussion, see Manzo, 2014b).

As to the first point, analytical sociology's view is deeply in line with the distinction between 'constitutive' and 'causal' statements discussed in the previous section. From a 'constitutive' point of view, large-scale properties are assumed to be *made of* smaller scale entities and properties. A network-related property like transitivity, for instance, must be constitutively explained by referring to some properties of smaller scale entities such as their tendency to associate with similar others. In this sense, analytical sociology is reductionist. However, from a causal explanatory point of view, analytical sociology acknowledges the empirical existence of large-scale properties and structures, which are defined in the following way: (1) once they come into existence, large-scale properties span a temporal window wider than that characterizing the small-scale entities that have dynamically contributed to generating them; (2) once they have come into existence, large-scale properties cannot be meaningfully defined, or quantified, for a single small-scale entity (on this second feature, see Hedström and Ylikoski, 2014: 62); and (3) the specific value and form assumed by large-scale properties crucially depends on the system of interdependencies that exist among the set of small-scale entities that have dynamically contributed to generating them. Analytical sociology not only acknowledges the empirical existence of large-scale properties and structures but also that they are causally relevant; the only requirement is that it should be possible to indicate at least one small-scale element through which the large-scale effect operates – when the small-scale entity is an actor, no matter whether consciously or unconsciously.

Thus, this understanding of the micro/macro relationship is compatible with the form of structural individualism originally formulated by Dutch sociologists and Boudon (see previous section) as well as with its dynamic version proposed by Coleman (1993: 63). However, to recast this doctrine in terms of scales and pay attention to the distinction between 'constitutive' and 'causal' statements avoids the undesirable misunderstanding that the micro must necessarily be equated with actors and their activities.

With respect to the second point, i.e., the theory of action followed by analytical sociology, two novelties must be highlighted. First, contrary to what some have argued (see, for instance, Gross, 2009), there is nothing in the spirit of analytical sociology that commits this perspective to a specific action theory. On the one hand, when intentionality and rationality are mobilized, it is an open understanding of rationality in terms of reason-based behavior that is assumed (see Manzo, 2010: 150–157, 2012). On the other hand, both the 'wide' and a fortiori, the narrow version(s) of rational choice are explicitly criticized for their lack of realism (see Manzo, 2013a; Hedström and Ylikoski, 2014). The second novelty concerns a trend within analytical sociology toward the recognition that individuals and their (intentional) actions should not necessarily be regarded as the basis of final explanation (see Hedström and Ylikoski, 2014). As stressed by Ylikoski (2012: Section 2.5), 'intentional fundamentalism' is unmotivated from the point of view of the mechanism-based account of explanation, which is neutral with respect to the rock bottom of causal chains. Thus, although analytical sociology echoes previous solutions to the micro/macro issue, namely within the rational-choice tradition (see, for instance, Coleman, 1987), analytical sociology differs by not committing to a specific theory of action, or to a specific level of analysis. This choice is likely to increase the domain of application of the perspective.

One aspect of analytical sociology's methodology, which is my third point, tends in the same direction. With the partial exception of the rational-choice tradition, which built its micro/macro theorizing in part on mathematical models (for a recent statement, see Raub et al., 2011), the perspectives reviewed in the previous sections relied either on natural language or on statistical methods whose internal structure makes it impossible to represent specific mechanisms bridging different levels of analysis (for an old statement, see Coleman, 1987: 168–171). This is especially true when 'transformation' mechanisms involving dynamic interdependencies are at stake (see Coleman, 1986: 1316). By contrast, analytical sociology proposes a particular simulation tool, i.e., agent-based modeling, in order to formalize multilevel mechanism-based hypotheses and study their consequences. Note that, among theorists involved in the micro/macro debate, the potential of simulation models did not go unnoticed (see, for instance, Collins, 1988: Appendix; Turner, 2005: 420). However, compared to other simulation techniques, the flexibility and power of agent-based modeling opens new and unhoped-for opportunities for micro/macro sociological theorizing. Let me briefly explain why (for a more detailed discussion, see Manzo, 2014b: 29–35).

Because of the algorithmic nature of the technique and the specific programing language used to write such algorithms, any substantive mechanism for which one has a theoretical representation can be implemented and studied within the framework of agent-based modeling. For sociologists interested in theorizing about transitions among levels of analysis, this flexibility is especially appreciable for the following reasons: (1) Despite their name, agent-based models are not at all limited to representing 'individuals.' According to the bundle of properties and rules that one programs for a given class of objects, the latter can be used to design the behavior and interactions of a set of particles, molecules, cells, beliefs, actors, groups (of particles, molecules, etc.), organizations, or whatever else on which one may have a precise theoretical understanding and representation. Thus, the specific way in which an 'object' is programed and the specific type of relations that the programmer establishes among different classes of 'objects'

allow representation of any form of connections and exchanges among entities located at different levels of analysis. (2) Agent-based modeling is entirely agnostic about the logic of action by which agents are driven. Thus, if the small-scale entities to be represented are individuals, then self-interested maximizing agents can be designed as well as more or less sophisticated 'cognitive' agents, or even agents that are driven simply by heuristics. (3) Agent-based modeling makes it possible to handle several forms of heterogeneity. Agents can be heterogeneous in terms of attributes and attributes' values. More radically, agents can be heterogeneous in terms of activities, tasks, or the behavior rules by which they are driven. The temporal order in which these rules are performed can also vary across agents. Thus, agent-based modeling constitutes a robust formal tool to go beyond the metaphor of the 'representative agent,' an unrealistic metaphor widely used to avoid, rather than to solve, the problem of the transition from the small- to the large-scale phenomenon (see Gallegati and Kirman, 1999). (4) Agent-based modeling makes it possible to design both space- and network-based local interactions. Agents can be embedded within these interactions so that their beliefs, desires, and opportunities can be modeled as locally constrained and influenced by the choices of other agents and by the network's topology. Moreover, compared to social network analysis, agent-based modeling makes it possible to turn network-level patterns such as homophily, triadic closures, cycles, balanced structures, structural holes, or brokerages into dynamic mechanisms, and their large-scale consequences can be studied (among students of social networks the interest of agent-based simulations is increasing indeed, see Monge and Contractor, 2003).

In a nutshell, agent-based modeling represents a flexible virtual laboratory in which the consequences of specific mechanisms connecting entities at different scales can be studied in a rigorous way. To build such artificial societies requires that every detail is specified. For micro/macro theorizing, this obligation has two advantages. First, conceptual vagueness cannot exist; second, when one takes shortcuts because detailed knowledge is not available, this limitation is immediately visible, thus helping to see where our theoretical knowledge must improve.

With respect to the micro/macro issue, what kind of theorizing can analytical sociology produce on these bases? Analytical sociology does not seek a 'grand' theory explicating how the connection between small- and large-scale phenomena works in general. Analytical sociology's ambition is to provide detailed descriptions of how this connection unfolds for specific and limited portions of the social world. In multiplying these local inquiries, the goal is to detect common features across contexts and phenomena so that classes of semigeneral, in the sense of transversal, mechanisms can be found (see Hedström and Bearman, 2009: 7). Examples of this approach can be seen in the effort to establish if network-based exposure mechanisms are equally involved in the emergence of large-scale suicide patterns (Bearman and Moody, 2004; Hedström et al., 2008), inequality structures (DiMaggio and Garip, 2011; Manzo, 2013b), fertility trends (Gonzalez-Bailon and Murphy, 2013), friendship dynamics (Bearman et al., 2004), or norm enforcement (Centola et al., 2005); if matching mechanisms work similarly on the job market (Stovel and Fountain, 2009) and for mate choices (Todd et al., 2005); or if imitation-based cumulative-advantage mechanisms apply to both the origin of status hierarchies (Gould, 2002) and consumption fads (Salganik and Watts, 2009).

This may seem not ambitious enough, but I believe it should help sociology to accumulate knowledge on how small-scale differentiation is turned into large-scale mixtures of differentiation and homogenization that is more precise than general transition modes – like 'aggregation,' 'combination,' 'externalization,' 'internalization,' or 'conformity' – which, according to some observers (see Münch and Smelser, 1987: 376–385), constitute the main outputs of several decades of micromacro theorizing.

Concluding Remarks

According to Archer's (1995) reconstruction, the solutions sociologist have proposed to the micro/macro, agency/structure issue historically fall within three types, i.e., 'downward conflation' (low-level entities are epiphenomenal compared to high-level entities), 'upward conflation' (high-level entities are epiphenomenal compared to low-level entities), and 'elision' or 'central conflation' (high- and low-level entities are inseparable and mutually constitutive). The sample of authors and perspectives overviewed in this article suggests that Archer's typology captures the fundamental trends of much of the debate. It is unsatisfactory, however, in three regards.

First, Archer's account does not consider heterodox readers who, contrasting usual interpretations, have shown that the classics are balanced micro/macro thinkers whose empirically oriented analyses escape Archer's three forms of reductionism. Second, Archer does not consider specific variants of the rational-choice tradition, modern social network analysis, and, for chronological reasons, the analytical sociology research program, all of which are approaches that reject both 'downward' and 'upward' conflation and seek to detail the mechanisms responsible for the interplay between structures and actors, thus avoiding 'central conflation' as well. Finally, again for chronological reasons, Archer's constructive proposal inspired by the critical-realism school of philosophy could not take into account recent epistemological developments that raise doubts about the utility of positing a priori an n-layer ontology, and suggest that, be it analytical or realist, the cake-layer model of society implicit in the level metaphor can be fruitfully replaced with a scale-based, flat perspective.

Thus, ultimately, this article suggests that the debates on the micro/macro issue are entering a new phase in which "there is no longer any demand for rigorous strategies of reduction" (Giesen, 1987: 337), and robust methods now exist for designing bridging mechanisms instead of only talking about them.

See also: AGIL, History of; Action, Theories of Social; Agent Based Modeling, Statistics of; Analytical Sociology; Behavioral Economics; Dramaturgical Analysis; Emergent Properties; Exchange: Social; Explanation: Conceptions in the Social Sciences; Functionalism, History of; Habitus: History of a Concept; Heuristics, History of; Individualism versus Collectivism: Philosophical Aspects; Intentionality and Rationality; Interactionism, Symbolic; Methodological Individualism in Sociology; Network Analysis, History of; Rational Choice Explanation: Philosophical Aspects; Rational Choice Theory in Sociology; Social Mechanism; Social Networks; Social Structure: History of the Concept; Sociology, History of; Utilitarian Social Thought, History of; Weberian Social Thought, History of.

Bibliography

Abbott, Andrew, 2001. Chaos of Disciplines. University of Chicago Press, Chicago.

Abbott, Andrew, February 2007. "Mechanisms and relations": a response to the comments. Sociologica. http://dx.doi.org/10.2383/24754.

Alexander, Jeffrey, 1987. Action and its environments. In: Alexander, Jeffrey, Giesen, Bernhard, Münch, Richard, Smelser, Neil (Eds.), The Micro-Macro Link. University of California Press, Berkeley and Los Angeles.

Alexander, Jeffrey, Giesen, Bernhard, 1987. From reduction to linkage: the long view of the micro-macro link. In: Alexander, Jeffrey, Giesen, Bernhard, Münch, Richard, Smelser, Neil (Eds.), The Micro-Macro Link. University of California Press, Berkeley and Los Angeles, pp. 1–45.

Archer, M.S., 1995. Realist Social Theory: The Morphogenetic Approach. Cambridge University Press, Cambridge.

Bearman, Peter S., 1991. The social structure of suicide. Sociological Forum 6 (3), 501–524.

Bearman, Peter S., Moody, James, 2004. Suicide and friendships among american adolescents. American Journal of Public Health 94 (1), 89–95.

Bearman, Peter S., Moody, James, Stovel, Katherine, 2004. Chains of affection: the structure of adolescent romantic and sexual networks. American Journal of Sociology 110 (1), 44–91.

Bedau, Mark, 2003. Downward causation and autonomy in weak emergence. Principia 6, 5–50.

Blau, Peter, 1964. Exchange and Power in Social Life. Wiley, New York.

Blau, Peter, 1977. Inequality and Heterogeneity: A Primitive Theory of Social Structure. Free Press, New York.

Blau, Peter, 1987. Contrasting theoretical perspectives. In: Alexander, Jeffrey, Giesen, Bernhard, Münch, Richard, Smelser, Neil (Eds.), The Micro-Macro Link. University of California Press, Berkeley and Los Angeles, pp. 71–86.

Blau, Peter, 2001. Macrostructural theory. In: Turner, Jonathan (Ed.), Handbook of Sociological Theory. Springer, New York, pp. 343–353.

Blumer, Herbert, 1969. Symbolic Interaction: Perspective and Method. Prentice-Hall, Englewood Cliffs, NJ.

Borgatti, Stephen, Lopez-Kidwell, Virginie, 2011. Network theory. In: Scott, John, Carrington, Peter (Eds.), The SAGE Handbook of Social Network Analysis. Sage, London.

Boudon, Raymond, 1987. The individualistic tradition in sociology. In: Alexander, Jeffrey, Giesen, Bernhard, Münch, Richard, Smelser, Neil (Eds.), The Micro-Macro Link. University of California Press, Berkeley and Los Angeles, pp. 45–71.

Boudon, Raymond, 1995. Should one still read Durkheim's rules after one hundred years? Revue Suisse de Sociologie 21 (3), 559–573.

Bourdieu, Pierre, 1977. Outline of a Theory of Practice. Cambridge University Press, London.

Bourdieu, Pierre, 1980[1990]. The Logic of Practice. Stanford University Press, Stanford, CA.

Calhoun, Craig, Gerteis, Joseph, Moody, James, Pfaff, Steven, Virk, Indermohan, 2012. Contemporary Sociological Theory. Wiley-Blackwell.

Centola, D., Macy, M., Willer, R.A., 2005. The Emperor's dilemma: a computational model of self-enforcing norms. American Journal of Sociology 110 (4), 1009–1040.

Cherkaoui, Mohamed, 2005. Invisible Codes: Essays on Generative Mechanisms. The Bardwell Press, Oxford.

Cherkaoui, Mohamed, 2008. Durkheim and the Puzzle of Social Complexity. The Bardwell Press, Oxford.

Coleman, James, 1986. Social theory, social research and a theory of action. American Journal of Sociology 91, 1309–1335.

Coleman, James, 1987. Microfoundations and macrosocial behaviour. In: Alexander, Jeffrey, Giesen, Bernhard, Münch, Richard, Smelser, Neil (Eds.), The Micro-Macro Link. University of California Press, Berkeley and Los Angeles, pp. 153–177.

Coleman, James, 1990. Foundations of Social Theory. Harvard University Press, Cambridge.

Coleman, James, 1993. Reply to Blau, Tuomela, Diekman and Baurmann. Analyse und Kritik 15 (1), 62–69.

Collins, Randall, 1981. On the microfoundations of macrosociology. American Journal of Sociology 86 (5), 984–1014.

Collins, Randall, 1988. Theoretical Sociology. Harcourt Brace Jovanovith, Orlando.

Collins, Randall, 2004. Interaction Ritual Chains. Princeton University Press, Princeton and Oxford.

DiMaggio, P., Garip, F., 2011. How network externalities can exacerbate intergroup inequality. American Journal of Sociology 116 (6), 1887–1933.

Durkheim, Emile, 1895[1982]. The Rules of Sociological Method. Free Press, New York.

Edling, Christofer, Hedström, Peter, 2009. Tocqueville and analytical sociology. In: Mohamed, Cherkaoui, Hamilton, Peter (Eds.), Raymond Boudon: A life in Sociology. Essays in Honour of Raymond Boudon, part 1, vol. I. Bardwell Press, Oxford, pp. 153–172.

Elster, Jon, 1985. Making Sense of Marx. Studies in Marxism and Social Theory. Cambridge University Press, Cambridge.

Elster, Jon, 2009. Alexis de Tocqueville. The First Social Scientist. Cambridge University Press, New York.

Gallegati, M., Kirman, A.P., 1999. Beyond the Representative Agent. Edward Elgar, Aldershot.

Garfinkel, Harold, 1967. Studies in Ethnomethodology. Prentice-Hall, Englewood Cliffs, NJ.

Giddens, A., 1984. The Constitution of Society: Outline of the Theory of Structuration. Polity Press, Cambridge.

Giesen, Bernhard, 1987. Beyond reductionism: four models relating micro and macro levels. In: Alexander, Jeffrey, Giesen, Bernhard, Münch, Richard, Smelser, Neil (Eds.), The Micro-Macro Link. University of California Press, Berkeley and Los Angeles, pp. 337–356.

Goffman, Erving, 1982. Frame Analysis: An Essay on the Organization of Experience. Harper Colophon, New York.

Gonzalez-Bailon, S., Murphy, T.E., 2013. Social interactions and long-term fertility dynamics. A simulation experiment in the context of the French fertility decline. Population Studies 67 (2), 135–155.

Gorski, Philip, 2013. What is critical realism? And why should you care? Contemporary Sociology 42 (3), 658–670.

Gould, Roger, 2002. The origins of status hierarchies: a formal theory and empirical test. American Journal of Sociology 107, 1143–1178.

Granovetter, Mark, 1973. The strength of weak ties. American Journal of Sociology 78 (6), 1360–1380.

Granovetter, Mark, 1983. The strength of weak ties: a network theory revisited. Sociological Theory 1, 201–233.

Granovetter, Mark, 1985. Economic action and social structure: the problem of embeddedness. American Journal of Sociology 91, 481–510.

Gross, N., 2009. A pragmatist theory of social mechanisms. American Sociological Review 74, 358–379.

Hedström, P., 2005. Dissecting the Social: On the Principles of Analytical Sociology. Cambridge University Press, Cambridge.

Hedström, P., Bearman, P. (Eds.), 2009. The Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford.

Hedström, Peter, Swedberg, Richard (Eds.), 1998. Social Mechanisms: An Analytical Approach to Social Theory. Cambridge University Press, Cambridge.

Hedström, Peter, Ylikoski, Petri, 2010. Causal mechanisms in the social sciences. Annual Review of Sociology 36, 49–67.

Hedström, Peter, Ylikoski, Petri, 2014. Analytical sociology and rational-choice theory. In: Manzo, Gianluca (Ed.), Analytical Sociology: Actions and Networks. Wiley, Chichester, UK, pp. 57–70.

Hedström, P., Liu, K.-Y., Nordvik, M., 2008. Interaction domains and suicides: a population-based panel study of suicides in the Stockholm metropolitan area, 1991–1999. Social Forces 87 (2), 713–740.

Homans, George, 1961. Social Behavior: Its Elementary Forms. Harcourt, Brace and World, New York.

Joas, Hand, Knöbl, Wolfgang, 2009. Social Theory: Twenty Introductory Lectures. Cambridge University Press, Cambridge.

Krause, Monika, 2013. Recombining micro/macro: the grammar of theoretical innovation. European Journal of Social Theory 16 (2), 139–152.
Latour, Bruno, 1999. On recalling ANT. In: Law, John, Hassard, John (Eds.), Actor Network Theory and after. Blackwell.
Little, Daniel, 2006. Levels of the social. In: Turner, S., Risjord, M.M. (Eds.), Handbook for Philosophy of Anthropology and Sociology. Elsevier, Amsterdam, New York, pp. 343–372.
Little, Daniel, January 2012. Analytical sociology and the rest of sociology. Sociologica 1. http://dx.doi.org/10.2383/36894.
Manzo, Gianluca, 2010. Analytical sociology and its critics. European Journal of Sociology 51 (1), 129–170.
Manzo, Gianluca, 2012. Reason-based explanations and analytical sociology: a rejoinder to Boudon. European Journal of Social Sciences 50 (2), 35–65.
Manzo, Gianluca, 2013a. Is rational choice theory still a rational choice of theory? Social Science Information 52 (3), 361–382.
Manzo, Gianluca, 2013b. Educational choices and social interactions: a formal model and a computational test. Comparative Social Research 30, 47–100.
Manzo, Gianluca (Ed.), 2014a. Analytical Sociology: Actions and Networks. Wiley, Chichester, UK.
Manzo, Gianluca, 2014b. Data, generative models, and mechanisms: more on the principles of analytical sociology. In: Manzo, Gianluca (Ed.), Analytical Sociology: Actions and Networks. Wiley, Chichester, UK, pp. 4–53.
Monge, P.R., Contractor, N., 2003. Theories of Communication Networks. Oxford University Press, New York.
Münch, Richard, Smelser, Neil, 1987. Relating the micro and macro. In: Alexander, Jeffrey, Giesen, Bernhard, Münch, Richard, Smelser, Neil (Eds.), Beyond Reductionism: Four Models Relating Micro and Macro Levels, The Micro-Macro Link. University of California Press, Berkeley and Los Angeles, pp. 356–389.
Opp, K.-D., 2013. What is analytical sociology? Strengths and weaknesses of a new sociological research program. Social Science Information 52 (3), 329–360.
Parsons, Talcott, 1951. The Social System. Free Press, Glencoe, IL.
Parsons, Talcott, 1954. Essays in Sociological Theory. The Free Press, Glencoe, Illinois.
Raub, Werner, Buskens, Vincent, van Assen, Marcel A.L. M., 2011. Micro-macro links and Microfoundations in sociology. Journal of Mathematical Sociology 35, 1–25.
Ritzer, George, 2008. Sociological Theory. McGraw Hill, New York.
Salganik, M.J., Watts, Duncan J., 2009. Web-based experiments for the study of collective social dynamics in cultural markets. Topics in Cognitive Science 1, 439–468.
Sawyer, R.K., 2005. Social Emergence: Societies as Complex Systems. Cambridge University Press, Cambridge.
Schutz, Alfred, 1937[1967]. The Phenomenology of the Social World. Northwestern University Press, Evanston, IL.
Stovel, Katherine, Fountain, Christine, 2009. Matching. In: Hedström, P., Bearman, P. (Eds.), The Oxford Handbook of Analytical Sociology. Oxford, University Press, Oxford, pp. 365–391.
Talcott, Parsons, 1937. The Structure of Social Action. A Study in Social Theory with Special Reference to a Group of Recent European Writers. Free Press, New York.
Todd, Peter M., Billari, Francesco C., Simao, Jorge, 2005. Aggregate age-at-marriage patterns from individual mate-search heuristics. Demography 42 (3), 559–574.
Turner, Jonathan, 2001. Sociological theory today. In: Turner, Jonathan (Ed.), Handbook of Sociological Theory. Springer, New York, pp. 1–21.
Turner, Jonathan, 2005. A new approach for theoretically integrating micro and macro analysis. In: Calhoun, Craig, Rojek, Chris, Turner, Bryan (Eds.), The SAGE Handbook of Sociology. SAGE Publications Ltd., London, pp. 405–423.
Valente, Thomas, 1995. Network Models of the Diffusion of Innovations. Hampton Press, Cresskill, NJ.
Van den Berg, Axel, 1998. Is sociological theory too grand for social mechanisms? In: Hedström, Peter, Swedberg, Richard (Eds.), Social Mechanisms: An Analytical Approach to Social Theory. Cambridge University Press, Cambridge, pp. 204–238.
Weber, Max, [1913]1981. Some categories of interpretive sociology (E. Graber, Trans.). Sociological Quarterly 22, 151–180.
Weber, Max, [1927]1981. General Economic History. Transaction Publishers.
Wiley, Norbert, 1988. The micro-macro problem in social theory. Sociological Theory 6 (2), 254–261.
Wippler, Reinhard, 1978. The structural-individualistic approach in Dutch sociology. The Netherlands Journal of Sociology 14, 135–155.
Wippler, Reinhard, Lindenberg, Siegwart, 1987. Collective phenomena and rational choice. In: Alexander, Jeffrey, Giesen, Bernhard, Münch, Richard, Smelser, Neil (Eds.), The Micro-Macro Link. University of California Press, Berkeley and Los Angeles, pp. 135–153.
Ylikoski, Petri, 2012. Micro, macro, and mechanisms. In: Kincaid, H. (Ed.), The Oxford Handbook of Philosophy of the Social Sciences. Oxford University Press, Oxford, pp. 21–45.
Ylikoski, Petri, 2014. Rethinking micro-macro relations. In: Collin, Finn, Zahle, Julie (Eds.), Rethinking the Individualism–Holism Debate. Essays in Philosophy of Social Science. Springer.

Mafia, Anthropology of

Anton Blok, University of Amsterdam, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mafias develop in failing states. Such states often continue failing because of the very presence of mafias and their collusion with politicians and civil servants. This happened in western Sicily in the nineteenth and twentieth centuries with the development of a modern market economy that was left without protection and created a demand for alternative protection. Men skilled in the use of violence and loosely organized in autonomous, informal, local groups, called mafias, enforced alternative, nonstate ('criminal') protection. Following the abolition of feudalism and the confiscation of church domains, vast landed estates (latifundia) came up for sale. The reforms increased the number of absentee landlords and widened the gaps between them and the peasants left without compensation for the lost user rights on the estates. Mafiosi operated as powerbrokers between absentee landlords and a growing peasant proletariat. The expansion of the suffrage in the twentieth century reinforced existing power relations based on patronage. Once more, a democratic innovation favored mafia structures, organized as political machines to gather votes and keeping at bay a barely controlled class war. Mafia presence was particularly strong in western Sicily: in and around Palermo and in its vast and desolate hinterland of large cereal-pasture holdings and isolated agro-towns on hilltops that separated the peasants from the land. The near absence of mafia in the eastern part of Sicily coincides with a less colonial and a less 'feudal' past, resulting in a more equal distribution of land, more intensive and diversified cultivation patterns, an advanced infrastructure, and more democratic institutions, similar to areas in the northern parts of Italy.

Mafia

The term 'mafia' originates from western Sicily, in particular from Palermo – for long the capital of wealth and power, and also a city of great poverty among its vast proletariat – a main recruiting field of *malfattori* (evildoers). Mafia refers to enforced protection of property provided by local, informal, autonomous, loosely organized groups, called 'families,' *cosche* (plural of *cosca*), or simply 'mafias.' Each group choses its own leader (*capo*) and forms a confederation with other, similar groups without losing its autonomy. These brotherhoods are named after the territory they control and protect: a village, a town, an urban neighborhood. Mafias are essentially local: first and foremost they depend on local knowledge – not unlike anthropologists. The Sicilian mafia is a grassroot movement raised from the ranks of the desperate poor. Segmentation and democracy rather than hierarchy rule the complex world of mafia.

Because they are autonomous and do not recognize any superior power, mafias are sovereign and resemble states, as sociologist and historian Charles Tilly explains in an influential essay, "War making and state making as organized crime" ([1985]1997). As a local group, a mafia polices a territory, provides protection, settles disputes, extracts a tribute, and wages wars with other *cosche*. All these functions are interrelated and influence each other. Mafias interact in conflict and accommodation, form alliances and coalitions, and engage in bloody wars, but leave their own members free to cultivate and maintain relations with outsiders and develop their own network of 'friends' across class and place, most notably politicians. Although lacking an effective centralizing ruling body, the Sicilian mafia, like mafias elsewhere, is a unitary social, cultural, and political confederation, called '*Cosa Nostra*' ('Our Thing'). Major local chiefs are represented in a commission that has mainly coordinating tasks (Gambetta, 1993; Paoli, 2003.) The so-called '*apo di tutti capo*' is a local chief whose influence reaches beyond his own *cosca* but usually respects the autonomy of other local mafias. This also holds for the most famous ones, including Vito Cascio Ferro (1862–1943), Calogero Vizzini (1877–1954), Giuseppe Genco Russo (1893–1976), Salvatore Riina (1930), and Bernardo Provenzano (1933). Tellingly, they all came from humble families in small towns located in the latifondo area of Sicily's western interior, Bisacquino, Villalba, Mussomeli, and Corleone, respectively.

Whatever its feudal aspects, mafia is a modern phenomenon and has developed with the advance of an expanding market economy in a context where the state failed to protect property rights and settle disputes (Gambetta, 1993; Varese, 2006, 2012). Following the abolition of feudalism in 1812 and the subsequent confiscation of church domains in the early 1860s, shortly after the unification of Italy in 1861, vast landed estates came up for sale. These reforms left large sections of a booming market unprotected which created a demand for alternative ('criminal') protection, to which men skilled in the use of violence responded. They included former estate guards, bandits, outlaws, ex-soldiers, and ambitious, ruthless peasants and shepherds recruited from the poorest sections of the rural population. To make things worse for the working classes, the new state imposed taxation, conscription, and obligatory education, which further alienated them from the new government. Schools withdrew their children from the family's labor force and were, therefore, considered an imposition. As a consequence illiteracy rates remained high far into the twentieth century. Prospective conscripts fled to the mountains and formed or joined armed bands, making banditry endemic in large parts of Sicily's interior parts for decades after unification. The army fought pitched battles with assorted robber bands at the cost of human life on both sides and the loss of

natural resources as forests were cut down or set on fire to combat outlawry (Mack Smith, 1968: 445–479; Fiume, 1984; D'Alessandro, 1959).

The transition to a modern market economy deprived most peasants from any compensation for their lost user rights (*usi civici*) on former feudal and ecclesiastical domains that were essential for their survival. (Blok, 1974: 37–42). The market economy widened the already considerable gap between landowners and peasants. Throughout the nineteenth century, proletarization of the workforce and rural poverty encouraged theft and organized banditry, an abortive peasant movement in the early 1890s, and subsequent massive emigration to America (Fiume, 1984; D'Alessandro, 1959; Romano, 1959; Blok, 1974). In the absence of reliable central control over the means of violence, mafia took shape as an *instrumentum regni* of Italian politicians, who chose to rule Sicily through its dominant agrarian classes – the absentee urban aristocracy and an upwardly mobile rural bourgeoisie. The expansion of the suffrage to all males – another democratic innovation – reinforced the existing power relations based on networks of patron–client relationships, in Sicily called 'friends-of-friends.' At the local polls, mafiosi canvased more votes for their representatives in parliament. "Mafia *cosche*," noted the Sicilian political scientists Gaetano Mosca at the beginning of the twentieth century, "immediately understood the great advantage that they could draw from participation in the political and administrative elections. This participation became more effective and active after the law widened the suffrage and gave the right to vote to the members of the *cosche* and to the classes in which they had greater influence and enjoyed prestige" ([1900]1949: 243, cited in Paoli, 2003: 186–187). Explains Paoli: "The clienteles organized by the *capimafia* – their nuclei formed by mafia groups – mingled with a wider network of clientelistic relationships through which the central government integrated southern upper classes into the national political system and secured their support for liberal government majorities" (2003: 187). With the further extension of the suffrage and the rise of the Popular and Socialist parties, prime minister Giovanni Giolitti (between 1903 and 1915) favored the strengthening of mafia structures, organized as political machines to gather votes for liberal clienteles and notables – and promoted their legitimation. "The action of mafia clienteles was, in fact, openly supported by state authorities at all levels, including even the prefects – the direct representatives of the central government in the single provinces" (Paoli, 2003: 187).

This aspect of the Sicilian mafia – its 'contiguity,' its 'cohabitation,' its collusion, its interweaving with state authorities, which helps accounts for its reproduction – gets somewhat lost in the revised perspective that emphasizes "mafia as a perverse response to modernization and the emergence of a market economy in a context where the state failed to protect property rights and a supply of individuals trained in the use of violence was available to offer alternative, nonstate enforcement" (Varese, 2006: 411–413).

Mafia presence was particularly strong in and around Palermo, in the adjacent citrus fruit areas of the *Conca d'Oro* (Golden Shell) and in its vast and forbidden hinterland of the island's western interior (Lupo, 1996: 103–147; Schneider and Schneider, 1976: 19–40). This was the area of large cereal-pasture holdings (the so-called *latifondi* or *ex-feudi*, whose history goes back to Antiquity) with its agro-towns situated on hilltops and mountain slopes for reasons of safety, but paradoxically alienating the peasants from the land and creating insecurity in the countryside. On a desolate, rolling, and mountainous landscape with huge stretches of wheat fields and pastures, large herds of sheep and cattle were driven from pasture to pasture according to the season. The arrangements facilitated one of Sicily's major 'industries' called *l'abigeato*, large-scale cattle rustling. Transporting the animals over long distances required the cooperation of various accomplices, including members of local *cosche* (Fentress, 2000: 132–133; Schneider and Schneider, 2003: 24–25). Cattle rustling in Sicily's western interior throws light on mafia's origin and class – and helps understand its continual reproduction, a point *pentito* (justice collaborator) Vincenzo Marsala made when talking to judge and prosecutor Giovanni Falcone: "… if you don't start with the rural towns, you will never uproot this noxious weed. The hinterland [of Palermo] is the great reservoir of manpower for the mafia, allowing it to continually find fresh blood for its membership" (cited in Stille, 1996: 191).

In this world of large-landed estates in Sicily's western interior, absentee landlords entrusted the management of their properties to ambitious, unscrupulous, upwardly mobile leaseholders (*gabellotti*) and overseers who had a reputation for violence. Until the advent of land reform in the 1950s, land was subleased to peasants on short-term contracts. These local bosses and their retainers formed the backbone of the rural mafia. The typical mafioso was a mediator or power-broker, who thrived on the gaps in communication between landlord and peasant, seller and buyer of commodities, political candidate and electorate, and exchanging services and 'favors' with his clients. Ultimately, he held sway over the links between state and local communities.

The earliest reference to what only much later would be called 'mafia,' comes from a report of a Neapolitan magistrate, the general prosecutor in Trapani, a major town in western Sicily, who in 1838 wrote to the Minister of Justice in Naples, well before the term 'mafia' appeared in print in the mid-1860s:

"In many towns and villages [in Sicily] there are fraternities, sort of sects that are called '*partiti*' without meetings and without other links than those of dependency on a chief who is here a landowner and there an archpriest. Common funds serve to meet certain ends as, for example, to exonerate a civil servant or to accuse an innocent person …. The population colludes with convicts. When thefts take place, mediators emerge and offer their services to make up for the stolen goods …. Many high-placed magistrates cover up these fraternities with an impenetrable shield …. It is impossible to induce the town guards to scout the area, nor is it possible to find witnesses of crimes committed in full daylight. At the top of this state of dissolution is the capital [Palermo] a city in which live 40,000 proletarians whose subsistence depends on the pomp and caprices of the rich and the powerful" (cited in Marino, 1964: 56). The reader may note that Biundi's *Vocabolaria Siciliano-Italiano* (Palermo, 1851) does not have entries on 'mafia' and 'mafioso.' The etymology of 'mafia' is controversial. See Novacco (1959) and Gambetta (1993: 135–140, 259–261).

This description contains all the elements of what mafia, after more than a century and a half, still stands for: informal, loosely structured, secret local sodalities ('fraternities') that police and protect a distinct territory, mutual aid, refusal to testify, and collusion with formal authorities, later called *intreccio* (Fentress, 2000: 147–181; Schneider and Schneider, 2003: 18, 33–34).

With the state's crackdown on the Sicilian mafia in the 1960s through the 1990s much more has become known about the subject through the reports from the Parliamentary Antimafia Commission and the detailed testimonies of several *pentiti* (ex-mafiosi who turned state evidence) in the maxi-trials in Palermo in the 1980s (Arlacchi, 1993, 1994; Gambetta, 1993; Stille, 1995; Jamieson, 2000; Schneider and Schneider, 2003; Paoli, 2003, 2008).

Local mafias impose a status contract on their members which involves a ritual initiation into the group and the adoption of a new identity of 'man of honor.' Having sworn an oath, members are obliged to keep secret their new identity as well as the composition, action, and strategies of their group. Their duty of silence is absolute. As in other secret societies, secrecy is a defense strategy (Paoli, 2003: 16–18; cf Simmel, [1908]1950: 346). Criminalized by the state as *malfattori*, *delinquenti* since the 1860s, mafia groups had to protect themselves from arrest and criminal prosecution for their continuous recourse to violence and had to resort to secrecy (Paoli, 2003: 17–18). Moreover, as Paoli explains, the initiation ceremonies of affiliation additionally created ritual ties of brotherhood among the members of a mafia family: the status contract is also an act of fraternization. The new recruits become 'brothers' to all members and share what anthropologists call a "regime of [generalized] reciprocity": this presupposes altruistic behavior without expecting any short-term reward. Paoli refers to an earlier account of a chief prosecutor, Lestingi (1884), who points out that mafia groups constitute brotherhoods whose 'essential character' lies in 'mutual aid without limits and without measure, even in crime' (cited in Paoli, 2003: 17). Paoli also recognizes that as secret brotherhoods using violence, southern Italian mafia federations (including the 'Ndrangheta' in Calabria) show striking similarities to organizations like the Yakusa in Japan and the Triads in China, Thailand, and Taiwan. These federations are also multifunctional (Paoli, 2008: 17–18; Varese, 2006: 411, 2012: 5, 17, 33, 147, 150–153). In his famous fieldwork study of Sicily, undertaken fifteen years after Italian unification, Franchetti already pointed at the 'extraordinary elasticity' of the Sicilian associations of *malfattori* (evildoers): "the goals multiply, the field action widens, without the need to multiply the statutes; the association divides for certain goals, remains united for others" (Franchetti ([1876]1993: 100; cited in Paoli, 2008: 18).

Yet this very elasticity may also question the brotherhood ideology and practice in mafias. The necessity of swearing secret oaths, common in secret societies as shown in La Fontaine comparative study (1985), may also betray a deep mutual suspicion among the affiliated. In his early essay on secret societies, Simmel pointed at their weaknesses. Secrets cannot be continuously protected. The protection it offers, notes Simmel, is indeed absolute, but only temporary: "a secret known by two persons is no longer a secret" (Simmel [1908] 1950: 346).

The vast volume of intra-mafia violence attests to this principle. Unintentionally, infighting exposes the very existence of mafias – about the last thing their members are waiting for. Moreover, violence is not restricted to wars between local mafias; bloody fights also occur between factions in the local *cosca* itself. Overweening leaders may invite resentment and several of them have been brutally killed and replaced by their former protégés. This happened, for example, in Corleone in the late 1950s when the much younger and ambitious Luciano Leggio killed his influential former boss Navarra, who had first tried to have him eliminated but failed. The altercation pertained to conflicting major economic interests (Chilanti and Farinella, 1964: 41–58; Catanzaro, 1988: 198–208). About 20 years later, a coalition of mafiosi from Corleone and Catania had the chief of the latter city's *cosca* murdered in an underhanded fashion. The victim harbored different views about the composition and policy of the Commission of *Cosa Nostra*. He was killed and replaced by his former protégé, Benedetto Santapaolo. Referring to the event, Antonino Calderone, the younger brother of the victim, later acknowledged – as a *pentito*, in a long and fascinating exposé – that everything in mafia comes down to sheer power. As Stille, referring to this case, phrased it, "[In] the simple code of the brutal, ancient world of the Sicilian countryside where force is the only law, there is no contradiction between personal kindness and extreme ferocity" (Stille, 1995: 231[228–234]; Arlacchi, 1993).

In their sober comments on current representations of the role of *cosche* and the 'coordinating' Commission, the Schneiders phrase the complexities as follows: "The prototypical nucleus of Sicilian organized crime consists of loose, at times shifting coalitions between leading mafiosi (select *capi*) and corrupt elites (regional and national politicians, administrators, professionals, and business men) engaged in the accumulation of power and wealth – not the overly conjured *cosca* or Commission" (Schneider and Schneider, 2003: 48).

Rather than a 'state-within-a-state,' as the Sicilian mafia has often been represented, mafiosi successfully infiltrated public institutions, including political parties, the judiciary, banks, and legal firms. They did so through their own personal network of friends rather than as members of sodalities or through the Commission. With the extensions of the suffrage in the course of the twentieth century, mafiosi enlarged their grip on the electorate and controlled more votes. Postwar government programs to 'develop' the country's southern periphery, enabled mafiosi to swallow up even more funds particularly in the (urban) construction industry. Writes Paoli, "From the 1950s onwards, Cosa Nostra groups, and later 'Ndrangheta families as well, set up their own building companies and pressed claims to become directly involved in both small and large building sites. Since 1980s, moreover, mafia enterprises of both regions [Sicily and Calabria] have become eager to obtain a share of any public work tendered by national or local public administrations. …. Unlike other western forms of organized crime, the meaning (and danger) of the two main Sicilian and Calabrian mafia organizations cannot be limited to their involvement in illegal markets. Their peculiarity lies in their will to exercise political power and their interest in exercising sovereign control over the people in their communities" (Paoli, 2008: 20–21). This last point is confirmed by the continuous

enforcement of a tribute, the so-called *pizzo*, all shops and business had to pay for protection to the local *cosca*. For a while, starting in 2004, and as part of the broader anti-mafia movement, another popular anti-mafia campaign in Palermo called ('*Addio Pizzo*') tried to evade this imposition by encouraging co-citizens to buy only in shops that refused to pay protection money to the local mafia. They did so under the slogan, "Pago a chi non paga" (I pay those who do not pay).

The Sicilian mafia is a form of organized crime, but not all forms of organized crime are mafias. As Varese convincingly argues, "the study of the mafia cannot be separated from the study of state failures" (2006: 441). If states fail, mafias can provide alternative protection of property and effectively settle business disputes. Although enforced and 'criminal,' mafia protection is not the same as extortion. It can even protect its clients against extortion. Also states enforce their protection on their subjects (2006: 412; Tilly, 1997). Yet Varese considers mafia protection a 'social evil.' He mentions several reasons: "In the economic sphere, it promotes inefficiency and reduces competition. By protecting thieves and other criminals, it also promotes further crime. The mafia operates without consideration for justice, fairness, or the wellbeing of society at large. In the world run by the mafia, there is no such thing as a 'right' to the protection for which one had paid. Mafiosi can ask for more favors or more money and then turn against their dutifully paying clients, and there is no higher authority to which a mafia victim can appeal" (Varese, 2006: 412–413).

Moreover, blaming the government of its neglect and theft (as in the common phrase '*Governo ladro*') obscures the reciprocal relation between condition and effect. Government investments (construction of roads and providing other public services) remain abortive as a result of the very practices of patronage, friendship, and self-help. Sicilians remain caught in a vicious circle of clientelistic politics and economic stagnation. As two anthropologists point out, "What clients ask from their patrons is not economic development but personal material advantages …. Clientelism entails not just the exchange of material favors or votes, but also the exchange of those products of the public administration (permits, licences, authorizations), which do not cost the patron money, but have a significant economic impact on the clients …. Highways may be built not because of their potential effect on economic growth, but because their construction creates numerous occasions for clientelistic exchanges" (Chubb, 1882: 3–6; Piattoni, 1998: 233–234). As mentioned above, the interweaving of political, economic, and mafia interests is called '*intreccio*,' or collusion (Schneider and Schneider, 2003: 18, 32–35). Next to class and culture, collusion is one of the factors that help account for mafia's longevity and reproduction.

Reciprocity between mafiosi and their clients is a form of lopsided friendship – a hallmark of patron–client relationships (Pitt-Rivers, [1954]1971: 140). In this generalized form of reciprocity, however, it happens that the mafioso will tell his client only later how, what, and where to reciprocate for a rendered service or 'favor.' In such cases, the client is tied up ('*vincolato*') and has to meet his obligation ('*levare l'obbligo*') whatever and whenever told. The mafioso can hold his client at a loose end, as a hostage of sorts, and as long as it suits him. This comes down to what more than anything else counts for mafiosi: showing power over other people. Mafia power is often sheer *prepotenza*; that is, showing that you are more powerful than others – and, for that reason, also more *omu* (man). It is an ultimate form of *abilità*, or self-help, in a failing state. Early German anthropologists neatly captured the conditions of *Selbsthilfe*, saying it looms large whenever the individual threatens 'to be either suffocated or be abandoned by the state' (Steinmetz, 1931: 522; cf Hess, 1970).

One may agree with Varese's view on mafia as a social evil – as simply organized crime. Yet other aspects deserve attention. The mafia developed, as noted above, as a reaction to extreme social injustice, humiliation, disrespect, and exploitation of the working classes. Its organization in brotherhoods in nineteenth and twentieth century western Sicily offered a respectable way out in the form of becoming 'men of honor.' As *pentito* Francesco Marino Mannoia explained to Palermitan Prosecutor Roberto Scarpinato, "Many believe that you enter *Cosa Nostra* for money. This is only part of the truth. Do you know why I entered *Cosa Nostra*? Because before in Palermo I was Mr. Nobody. Afterwards, wherever I went, heads lowered. And this for me was worth any price" (cited in Paoli, 2003: 152). Moreover, as results from prolonged local observation and participation in two rural communities in western Sicily (Blok, 1974; Schneider and Schneider, 1976), one should acknowledge the importance of long-standing, multi-stranded relations between local mafiosi and their clients, including ties of kinship, friendship, marriage, ritual kinship, and patronage which involves reciprocity of services and favors over time (cf Wolf, 1966: 81–91, 2001: 166–183). This brings us to the issue of mafia and Sicilian cultural codes – to the norms and values mafiosi share with other Sicilians. Here we touch on the breeding ground of mafia and its reproduction.

Although mafiosi exacerbated class tensions between landlords and peasants by their rent-capitalist management of landed estates and appropriating land and became landowners in their own right, they also controlled these tensions by carving out channels of upward social mobility for ambitious and ruthless peasants and shepherds. When indicted for violent crimes, mafiosi were usually acquitted for lack of evidence because of high-level protection while local witnesses would hesitate to testify against them (a favor the perpetrators would certainly not forget), which greatly helped to enhance their power and reputation of 'men of respect.' Inspired by both fear and admiration, the local population drew up a 'wall of silence' (*omertà*), which ultimately blocked effective prosecution of mafiosi and enhanced their reputation.

Sicilians may have mixed feelings about mafiosi (who remained somewhat ambiguous figures), but they can nonetheless strongly identify with them (Hess, 1970: 72–81). As 'men of honor' they have made themselves socially respected. More importantly, they show in their entire habitus the main cultural values and norms of Sicilian popular culture, including *omertà* (from *omu*, man), *abilità*, *furbería* (cunning, recalling the *metis* of Odysseus), care of family, conjugal loyalty, virginity of unmarried daughters and sisters, friendship, avoiding persons who represent the law, and ruthless ways when dealing with offences (*vendetta*). We find these cultural codes back and expressed in everyday practices

and situations, and further clarified in the use of numerous old yet current expressions, sayings, and proverbs (Blok, 2010). Following helpful suggestions of Sicilian historian Giovanna Fiume (1989: 203–205), a few examples suggest (without the intention to criminalize an entire population) to what extent mafiosi share these cultural codes that are widely accepted in western Sicilian society.

When confronted with (a picture of) the body of a murdered person, people may take a close look and identify categorically with the perpetrator, not with the victim, and comment with the standard phrase, "If he is killed in this (brutal) way, he must have done something terrible." In a world of failing states, *vendetta* is not only accepted but also encouraged as a social necessity – if only to ward off further violence. In similar cases people may also say, "*Bisogna pensare al vivo*." In translation, "One has to take care of the living one" – the perpetrator. An anecdote may further illustrate this point. When Marco Minghetti was the prime minister of Italy (1873–76), he made a trip to Sicily and visited the small town of Camporeale, situated west of the now well-known town of Corleone. The local priest introduced himself to him as follows: "I recommend to you a poor young man who needs your protection." Minghetti asked, "And why? What does he want?" The priest answered, "Nothing, he only met with an accident, he has killed a man" (cited in Alongi, 1891: 50). The exchange between the highest authority of the state and a local authority, familiar with indigenous customs and using the word '*poveretto*' (not only 'poor,' but also 'unfortunate') for a man who from the point of view of formal law was a criminal, involves a switching of codes still current today.

The values attached to *omertà* show how mafia practices and cultural codes dovetail with Sicilian popular culture at large. As the Schneiders explain: "Mafiosi adapted to expanding state power not only through the organizational arrangements associated with the *cosca* but also by actively and self-consciously manipulating the cultural codes of western Sicily, exploiting them as ideology. In this way, they were able to promote solidarity among themselves and at the same time to influence the attitudes and behavior of others (especially those who might testify against them in court). It made sense for mafiosi to become specialists in honor, friendship, and *furbería* (shrewd and cunning). Those codes were part of the local lexicon of many people; they symbolized the island's distinctive qualities vis-à-vis northern Italy and the rest of Europe; and they were instrumental in competition for power" (Schneider and Schneider, 1976: 192).

Above all, *omertà* comes down to be silent, pretending not to know, avoiding asking or answering questions. As the proverb indicates: "*Cu è surdu, orbu e tace, campa cent'anni 'mpace*". ("Anyone who is deaf, blind, and does not talk, will live a hundred years in peace" (Pitrè, 1978: III, 215–234)). This widely used proverb neatly captures the purport of the cultural code of *omertà* and refers to a habitus of reserve, secrecy, and reticence. Ex-mafioso and *pentito*, Tommaso Buscetta epitomized this world as "una cultura di tipo omertoso" (cited in Arlacchi, 1994: 84).

In Sicily people make much of friendship, but are no less aware of its dark sides. This awareness takes form in another cultural code and finds expression in several sayings. "*Amicizia con tutti, confidenza a nessuno*" ("Friends with everyone, trust nobody"). First, friends know much more about you than anybody else and do not raise suspicion immediately. For the same reason they may be the ones to (be asked to) turn against you. This also happened among "friends-of-friends." Intra-mafia killings – especially during the so-called mafia wars in the early 1960s and the period between the late 1970s and early 1990s – surpass homicides involving outsiders. Many victims were executed under the veil of friendship and hospitality, including betrayals within betrayals (Stille, 1995). The standard phrase "*Dagli amici mi guarda Iddio*" ("God protect me from my friends") implies that one can take care of his enemies himself. A friend of a prospective victim can be mobilized to carry out a killing, assist in it, or unknowably become an accomplice – if only not to raise the alarm of the victim. When the bandit Salvatore Giuliano (1922–50) had outlived his usefulness for politicians, he was, through mediation of local mafiosi, betrayed and killed by his closest ally, Gaspare Pisciotta, who was also his cousin (Maxwell, 1957). Tommaso Buscetta, a major figure in the mafia of Palermo who later moved to South America and had seen a lot of intra-mafia violence both in prison and on the streets of Palermo, referred to the ambiguities of friendship among mafiosi in an interview when detained as a *pentito* as follows. "The man who stands besides you might take you to your grave as easily as to a party. The dearest friend may become your murderer" (Arlacchi, 1994: 155).

From the late 1970s onward, mafia in Sicily gradually lost its role as a pragmatic extension of the State and assumed the character of a hidden power. Mafiosi disappeared as public figures with public identities. Mafia's peaceful coexistence with the State (the *pax mafiosa*) came to an end with the killing of politicians, policemen, magistrates, and numerous other civil servants, not to mention later terroristic attacks in three north Italian cities in the early 1990s. Terror replaced accommodation. Like outlaws elsewhere, mafiosi had to hide from the law (if they did not cooperate as *pentiti* with the judicial authorities), living clandestine lives as fugitives and outlaws in modest accommodations, mostly in their own territories. Several factors are responsible for this major transition.

First, beginning in the late 1960s, Sicilian mafiosi moved into the international narcotic trade, which produced huge and fast profits but entailed high risks. These ventures, like the building boom earlier, resulted in internecine struggles between rival factions. In this case between upstarts from the Palermo hinterland, like the Corleonesi, first led by Luciano Leggio and later by Salvatore Riina, and Bernardo Provenzano – the "viddani" (i.e., peasants: a contemptuous reference to their humble background), on the one hand, and the urban, rich, and chic establishments in and around Palermo of which Stefano Bontade was the most striking figure and close ('*intimo*') to Salvo Lima, the main Sicilian political patron, or *favoreggiatore*, who mediated between the majority government party (Christen Democrats, in power since 1948) and the mafia (Schneider and Schneider, 2003: 33, 55–57). The sustained, effective anti-mafia policies provoked campaigns of terror against politicians and civil servants as well as against family members of *pentiti*. This new development brought mafiosi into open conflict with the Italian State, which could not stand back and close its eyes to

the massive infiltration of drug money into its economy. Nor could the government ignore the pressure from neighboring States and other countries to take a strong position against the drug-trafficking and money laundering. The narcotics trade changed the Sicilian and American mafia. Although it became the mafia's most lucrative enterprise (at least for a while), it has also generated more murder, violence, and betrayal than any other criminal activity. As the profits spiraled, so did the need to protect them (Shawcross and Young, 1987: 307).

The extent to which the tables had been turned on politicians who had either directly or indirectly protected the interests of mafiosi for decades in exchange for electoral support can best be illustrated by the following incident. After the first Sicilian politicians and magistrates who tried to withdraw from the mafia or to combat it with determination were assassinated in public places in 1979 and 1980 to be remembered as the first wave of 'excellent cadavers,' Stefano Bontade, the flamboyant mafia leader in Palermo, allegedly warned the prime minister Giulio Andreotti on his visit to the island in February 1980 with the following words: "We're in charge here in Sicily and if you don't want to destroy the DC [Christian Democrats] you will do what we say. Otherwise we will take away our votes not just here but … in the whole of southern Italy. Then you will be left with only the vote in the North, where everyone votes Communist" (deposition of Marino Mannoia in the American witness protection program, as Stille clarifies, "Although Marino Mannoia did not participate in the private discussions, his boss, Stefano Bontade, later described the contents of the meeting to him …."; "Marino Mannoia provided the first eyewitness testimony tying Andreotti to the bosses of the mafia" (Stille, 1995: 391)).

The new generation of magistrates was helped by special legislation, which included a law that made it illegal to belong to the mafia (now defined as a criminal organization: '*associazione mafiosa*') and also gave prosecutors the power to seize mafia assets accumulated through criminal activity. From the early 1980s onward, the Italian State renewed its fight against the mafia, which culminated in effective proceedings (the so-called maxi-trials) against hundreds of mafiosi. The prosecution was substantially helped by mafiosi from losing factions who, partly for reason of revenge and partly in order to survive, repented, turned state's evidence, and helped build a case against their former peers (Stille, 1995; Jamieson, 2000; Schneider and Schneider, 2003).

Behind all these changes lurked an even larger transition. This was the demise of communism in Europe, which unintentionally undermined the mafia's *raison d'etre*. Since the immediate postwar years, mafiosi had always supported political parties that opposed communist and socialist parties. Faced with the strongest communist party in Western Europe, the Italian Christian Democrats could not forego the support of an informal power that provided huge blocks of votes and, thus, effectively staved off what was perceived as 'the danger of communism.' With the disintegration of the Soviet Union and the end of the Cold War, the situation radically changed and the shield of protection that for decades had kept mafiosi out of prison began to fall apart. The situation resembled the crackdown on the mafia under Fascism. Prefect Mori's large-scale operation in the late 1920s took place after the Sicilian landed classes had obtained more effective and less costly protection of their property from the new government.

This turn exposed mafiosi and left them without protection. Those who did not manage to emigrate were rounded up, tried, and convicted (Mori, 1933; Duggan, 1989). The fall of Fascism, the Allied occupation of Sicily, and the reestablishment of democracy in the 1940s created the conditions for a return and rebirth of the mafia as an *instrumentum regni* for politicians set to prevent the advent of a communist government in Italy, a country situated on the borders of Soviet influence.

Today, in the Italian pubic debate, the Sicilian mafia is still portrayed as powerful, almost invincible. Few observers express less pessimistic views and are, therefore, criticized by anti-mafia activists, who even accuse them of indirectly serving mafia interests. Still others believe that the anti-mafia fight has slowed down over the past few years (Paoli, 2008: 15). Paoli, who documents the current crisis of Italian mafias (both in Sicily and Calabria), argues, however, that their decline is primarily the result of law enforcement repression since the early 1990s. Yet, public prosecutors in the court of justice in Palermo still have a hard time doing their work. Facing both politicians and mafiosi in the dock, continually surrounded by bodyguards, behind the door of their office and outside when going and leaving the courthouse, they suffer death threats of mafia bosses sentenced to life in prison and receive little cooperation from the government. "I am not sure of my life," reported public prosecutor, Nino di Matteo, in a recent interview. Di Matteo grew up in Palermo with the stereotype that all Sicilians are mafiosi but shares the dream that honest Sicilians can liberate themselves from the mafia. "In spite of the problems I feel that I can contribute something" (Sandra Venema, *De Volkskran*t 8 May 2014, 10–11).

As indicated above, mafiosi draw substantially on Sicilian popular culture, on cultural codes "that were part of the local lexicon of many people …." Moreover, poverty and unemployment among the working classes, both a consequence of mafia, still provides recruiting fields for prospective aspirant mafiosi in both urban and rural settings. Together with culture and class, collusion seems to continually encourage mafia reproduction in western Sicily.

Coda

Today, the term 'mafia' is widely used as a synonym for organized crime. A still larger conflation of the term includes metaphorical and ironic references to any coterie wielding illicit power in formal organizations or institutions.

Bibliography

Alongi, G., 1891. L'Abigeato in Sicilia. Studii di Patologia Sociale. Tipi. Di Luigi Giliberto, Marsala.

Arlacchi, P., 1993. Men of Dishonor. Inside the Sicilian Mafia. An Account of Antonino Calderone. W. Morrow, New York.

Arlacchi, P., 1994. Addio Cosa Nostra. La vita di Tommaso Buscetta. Rizzoli, Milano.

Biundi, G., 1851. Vocabulario Siciliano-Iitaliano. Stamperia Carini, Palermo.

Blok, A., 1974. The Mafia of a Sicilian Village, 1860–1960. A Study of Violent Peasant Entrepreneurs. Basil Blackwell, Oxford.

Blok, A., 2010. Behavior codes in Siciliy. Bypassing the law. Behemoth. A Journal on Civilisation 2, 56–70.
Catanzaro, R., 1988. Il Delitto come Impresa. Storia Sociale della Mafia. Liviana Editrice, Padova [1992. Men of Respect. A Social History of the Sicilian Mafia. The Free Press, New York.].
Chilanti, F., Farinella, M., 1964. Rapporto sulla Mafia. F. Flaccovio.
Chubb, J., 1882. Patronage, Power, and Poverty in Southern Italy. A Tale of Two Cities. Cambridge University Press, Cambridge.
D'Alessandro, E., 1959. Brigantaggio e Mafia in Sicilia. Casa Editrice G. D'Anna, Messina/Firenze, Florence.
Duggan, C., 1989. Fascism and the Mafia. Yale University Press, New Haven, CT.
Fentress, J., 2000. Rebels and Mafiosi. Death in a Sicilian Landscape. Cornell, Ithaca, NY.
Fiume, G., 1984. Le Bande Armate in Sicilia (1819–1849). Annali della Facoltà di Lettere Filosofia dell'Università di Palermo, Palermo.
Fiume, G., 1989. Interview with Giovanni Falcone: La Mafia tra Criminalità e Cultura. Meridiana 5, 199–2009.
Franchetti, L., 1993[1876]. La Sicilia nel 1876. Condizioni politiche ed amministrative della Sicilia. Donzelli, Rome.
La Fontaine, J.S., 1985. Initiation. Ritual Drama and Secret Knowledge across the World. Penguin Books, Harmondsworth.
Gambetta, D., 1993. The Sicilan Mafia. The Business of Private Protection. Harvard University Press, Cambridge, MA.
Hess, H., 1970. Mafia. Zentrale Herrschaft und Lokale Gegenmacht. Mohr, Tübingen.
Jamieson, A., 2000. The Antimafia. Italy's Fight against Organized Crime. Macmillan, London.
Lestingi, F., 1884. L'Associzione della Fratellanza nella provincia di Girgenti. Archivio di Psichiatria, Antropologia Criminale e Scienze Penali 5, 452–463.
Lupo, S., 1996. Storia della Mafia. Dalle origini ai nostri giorni, Revised and expanded ed. Donzelli, Rome.
Mack Smith, D., 1968. A History of Sicily, 2 vols. Chatto & Windus, London.
Marino, G.C., 1964. L'Opposizione Mafiosa (1870–1882). S.F. Flaccovio, Palermo.
Maxwell, G., 1957. God Protect me from my Friends. Longmans, London.
Mori, C., 1933. The Last Struggle with the Mafia (Orlo Williams, Trans.). Putnam, London/New York.
Mosca, G., 1949[1900]. Che Cosa è la Mafia?. In: Partiti e Sindacati nella Crisi del Regime Parlamentare. Laterza, Bari.
Novacco, D., 1959. Considerazioni sulla Fortuna del Termine "Mafia." Belfagor 14, 2–11.
Paoli, L., 2003. Mafia Brotherhoods. Organized Crime Italian Style. Oxford University Press, Oxford.
Paoli, L., 2008. The decline of the Italian mafia. In: Siegel, D., Nelen, H. (Eds.), Organized Crime, Culture, Markets and Politics. Springer, New York.
Piattoni, S., 1998. "Virtuous clientelism": the southern question resolved? In: Schneider, J. (Ed.), Italy's "Southern Question." Orientalism in One Country. Berg, Oxford.
Pitrè, G., 1978[1870–1913]. Proverbi Siciliani. A cura di Aurelio Rigoli, 5 vols. Il Vespro, Palermo.
Pitt-Rivers, J., 1971[1954]. People of the Sierra. University of Chicago Press, Chicago.
Romano, S.F., 1959. Storia dei Fasci Siciliani. Laterza, Bari.
Schneider, J., Schneider, P., 1976. Culture and Political Economy in Western Sicily. Academic Press, New York.
Schneider, J., Schneider, P., 2003. Reversible Destiny. Mafia, Antimafia, and the Struggle for Palermo. University of California Press, Berkeley.
Shawcross, T., Young, M., 1987. Men of Honor. The Confessions of Tommaso Buscetta. Collins, London.
Simmel, G., 1950. Secrecy & the secret society (Kurt H. Wolf, Trans., Ed., and with an Introduction). In: Simmel, G. (Ed.), The Sociology of Georg Simmel. The Free Press, New York.
Steinmetz, S., 1931. Selbsthilfe. In: Vierkandt, A. (Ed.), Handwörterbuch der Soziologie. Ferdinand Enke Verlag, Stuttgart.
Stille, A., 1995. Excellent Cadavers. The Mafia and the Death of the First Italian Republic. Jonathan Cape, London.
Tilly, C., 1997[1985]. War making and state making as organized crime. In: Tilly, Charles (Ed.), Roads from Past to Future. Rowman & Littlefield, Lanham, MD.
Varese, F., 2006. How mafia's migrate: the case of the Ndrangheta in Northern Italy. Law & Society Review 40 (2), 411–444.
Varese, F., 2012. Mafias on the Move. How Organized Crime Conquers New Territories. Princeton University Press, Princeton & Oxford.
Wolf, E.R., 1966. Peasants. Prentice-Hall, Englewood Cliff, NJ.
Wolf, E.R., 2001[1966]. Kinship, Friendship, and patron-client relations in complex societies. Pathways of Power. Building an Anthropology of the Modern World. University of California Press, Berkeley.

Mafias, Criminology of

Gavin Slade, Centre for Criminology and Socio-Legal Studies, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article gives an overview of academic writing on mafias. It provides a definition for this often sensationalized phenomenon which focuses on the governance functions mafias perform and their similarities to states. It contrasts mafias with other forms of organized crime. The article goes on to discuss how and why mafias emerge and exist in certain places in the world at certain times, what mafias do, how they organize, how they expand, and how finally they decline.

The word 'mafia' invokes vivid imagery: secret societies, bizarre initiation rituals, violent deaths, Tony Soprano, a severed horse's head at the end of a bed. Much of this imagery comes from the slew of films and television shows about mafias, which seem to have unending popularity. Of course, the cultural representation of mafias is often a long way from reality. The academic study of organized crime rather attempts to give sober analyses of the activities, structure and social organization of groups or networks of criminals, the functions they perform, and the harm they cause. This has become a burgeoning sphere in criminology though it is a notoriously difficult subject to conduct research on. While ethnographic accounts of organized crime are rare due to the nature of the subject, research can often employ qualitative methods combined with the study of police data and court cases. Wiretap data, for example, can be used to conduct quantitative and network analyses. Some academics also focus on the way 'organized crime' as a concept is constructed socially, politically, and culturally, and how it may be used and abused. In all this, it is necessary to clearly distinguish mafias as a distinct analytical object of study. In order to do this, we will first turn to definitions of what a mafia is. Then, giving examples from around the world, we will look at why mafias emerge in certain places at certain times, their activities, their organization and expansion, and, finally, their decline.

What Are Mafias?

An organized criminal group, according to the United Nations Convention against transnational organized crime (2004: p. 5), is "a structured group of three or more persons, existing for a period of time and acting in concert with the aim of committing one or more serious crimes or offences…in order to obtain directly or indirectly, a financial or material benefit." Other organizations and governments have a very similar understanding of organized criminal groups to this. In most definitions in the last decade, we can see the same familiar elements: a group must be enduring, commit certain serious crimes, and seek enrichment. Yet, over the decades there was not always such a consensus. Definitions by both academics and governments have changed. In 1963, after mafia turncoat Joseph Valachi had vividly testified to the McClellan Commission as to the existence of a structured, pervasive Italian–American network of criminals, 'organized crime' became virtually synonymous with 'Cosa Nostra' (meaning 'Our Thing,' the Sicilian, and Italian–American mob) in definitions. As research developed away from the fixation on Italian–Americans, other concepts such as 'hierarchy,' 'monopoly,' 'harm,' and 'network' all came in and out of fashion in defining organized crime (Varese, 2010).

Clearly, when we think of mafias, the United Nations definition would appear to fit very well. Yet, perhaps this is because that definition is in many regards a catch-all one and, therefore, unhelpful for analytically distinguishing between certain types of groups. Is there a difference, for example, between mafias such as the Triads (originally from Hong Kong), the Yakuza (Japan), Camorra (Naples, Italy), *vory-v-zakone* ('thieves-in-law,' or the Russian Mafia), or Cosa Nostra (Sicily and the US) and street gangs such as the Cribs and the Bloods or motorcycle gangs such as the Hell's Angels for that matter? How should we distinguish between the 'Ndrangheta, a mafia based in Calabria, Italy which controls and extracts tribute from businesses in the region through violence, and the rather less ruthless cheese smuggling network that operated across the US–Canada border selling cheap US cheese to hard-up Canadian restaurateurs on the black market?

On the UN definition both are organized crime, yet this seems to miss important differences. Moreover, the use and abuse of the word mafia does not help here either; the press dubbed the cheese smugglers the 'Poutine Mafia' after the famously cheesy and greasy Canadian dish. Just as many different phenomena can fit into the UN definition of organized crime, so 'mafias' are now everywhere and therefore nowhere. Arkady Vaksberg argued that there was a 'Soviet Mafia,' by which he meant the entire political system of the Soviet Union (Vaksberg, 1991). If mafias can be anything from cheese smugglers to governments of continent-sized countries what meaning can the term possess? Some therefore argue that the term 'mafia' should simply be discarded by criminologists, as it has at best no analytical purchase and at worst it is a dangerous slur used to tarnish undesirable groups, particularly immigrants. When attached to ethnic, national, or regional markers such as 'Italian,' 'Albanian,' 'Chechen,' 'Corsican,' or 'Turkish' to give some of the most obvious examples, the term 'mafia' stigmatizes.

Is there a sense in which the term 'mafia' can be rescued from this muddle? There are a number of scholars who argue that while mafias are certainly organized crime groups, they are a specific subspecies possessing clear features that distinguish them. Some emphasize the use of a hierarchical structure with clear bosses, others focus on the systematic use of corruption

and clientelist links with state agents as a distinguishing feature. However, perhaps the most powerful approach to understanding mafias is to focus on what they do. Diego Gambetta, in a groundbreaking work on the Sicilian Mafia, argued for seeing mafias as alternative governance providers, attempting, as states do, to monopolize the sale of protection over a given territory (Gambetta, 1993). Gambetta stressed that mafias are, furthermore, autonomous suppliers of this protection; they are not internalized within another entity, a business say, but sell protection and extract tax in return from individuals, organizations, and enterprises.

Borrowing from Gambetta and the Nobel Prize winning economist Thomas Schelling (1984), we can make the following distinction: organized crime groups attempt to monopolize the provision of some illicit service or product utilizing violence, whereas mafias are autonomous entities that attempt to monopolize the provision of protection over a given territory utilizing violence. In reality this may be quite blurred. Mafias may engage in other organized criminal activities, drug trafficking say, but as long they maintain this protection function at their core, they may be called a mafia. Clearly, the protection the mafia sells might simply be from itself, and we call this straightforward extortion. However, the evidence suggests that mafias in most cases generalize the sale of protection and are ready to actually provide the protection service should the need arise. A key insight from this distinction of mafias from other types of organized crime groups is that mafias provide governance in an unstable underworld that is naturally lacking in trust. Organized crime groups cannot turn to the police or the courts to resolve disputes and enforce agreements, and so often turn to mafias which induce the trust that criminal groups need to coordinate and operate.

This element of governance as central to understanding mafias has a clear debt to Max Weber and his definition of the state which he defined as the monopoly of legitimate force over a given territory (Weber, 1992). Indeed, this suggests that in essence mafias do just what states do; the distinction from the state is simply one of scale and legitimacy (Tilly, 1985). A number of criminological studies, inspired by Gambetta's work, have investigated groups in China, Russia, and Japan (Chu, 1999; Varese, 2001; Hill, 2006) and found that there were indeed mafias, in the strict sense that Gambetta uses the term, operating in these countries. The Yakuza, Russian Mafiya, Triads, Cosa Nostra, 'Ndrangheta, and so on, all have the protection function in common. This definition allows us to discuss and analyze certain other types of phenomena not usually associated with mafia activity. In Northern Ireland, for example, paramilitary groups which had primarily political goals also engaged in racketeering and extracted tribute over a given territory. In this sense, these groups are akin to mafias, distinguished only by their overtly political aims – something most mafias do not share. The focus then on monopoly and protection, in essence governance functions, allows us to analyze and define what groups we can legitimately call mafias.

Why Do Mafias Emerge?

The term 'mafia' first emerged into popular usage in nineteenth-century Sicily, a variation on an Arabic word that described masculine bravado and the refusal of authority. A theatrical play called 'I Mafiusi della Vicaria' about a Sicilian prison group that had its own rituals and rules toured Italy in the 1860s and popularized the words 'mafia' and 'mafioso.' But how and why did the mafia itself emerge in Sicily at this time? And why did some areas of the island prove more fertile breeding grounds for this phenomenon than others? Are the conditions that gave rise to the Sicilian Mafia the same in other countries that have seen mafias emerge?

Sicily is an island that has experienced centuries of colonization by foreign powers. By the middle of the nineteenth century, there was fierce resistance to the attempts of the Italian state to integrate the south into a newly unified nation. Trust in society was low. The end of feudalism had created a class of free peasants who could cultivate the land and profit from it, paying rents to landholders. A nascent market economy was emerging. Absent landlords needed protection of their estates and extraction of the highest rents possible, agricultural produce from the estates needed safe passage for goods from the inland to the coast. The central Italian state was weak in the region and could not provide this protection efficiently. It was in the context of low trust, weak state protection, increasing economic transactions, and class tensions that individual mafiosi, or enforcers, emerged. They autonomously resolved disputes, enabled the extortion of peasants, and provided for the safe passage of goods from the inland to the coast.

Mafiosi at this time protected the interests of anyone who would pay. Former guards of the estates who had lost their jobs during this period of economic change created a supply of violent individuals who sought out a role for themselves in the new economy and sold their services as entrepreneurs of violence. It is not clear that at this time secretive organizations – the mafia – had yet emerged but individual mafiosi had carved out a niche for themselves with a distinct way of behaving (Blok, 1974). In economic terms, these mafiosi met a demand for trust as, for a fee, they helped to reduce transaction costs by inducing trust in informal agreements, preventing defection in lieu of legal guarantees and contracts. Indeed, where the demand for such services is higher, in principle the greater should be the mafia presence. In Sicily, the west of the island produced many more mafiosi than the east, and this can be explained by the greater instances of absentee landlordism and the breakup of the land into smaller parcels at the end of feudalism creating a higher volume of economic transactions than in the east and greater unprotected movement of goods from the inland to the coast. The specificities of land reform then meant that demand for protection varied within the island. This simple fact undermines the common assumption that there is something specific to Sicilian mentality or culture that was crucial to mafia emergence.

The case of Sicily suggests that there are a number of conditions that appear necessary for the emergence of a mafia in a specific place and time. First, low trust among individuals and in the protection of the state to enforce contracts, resolve disputes, and protect individuals and entities is a usual condition of mafia emergence. Second, where a monopoly over property is broken up and there is a resulting increase in the number of transactions between newly emerging economic actors, the demand for trust increases. If the state is not able to meet this demand then alternative protection providers may

emerge. Yet this emergence is not a given. There must also be a supply of individuals who are able to meet the demand for protection, i.e., those trained in violence, as this is the key asset a mafia possesses. These conditions were present in the mid-nineteenth century in Sicily. They produced mafiosi and, finally, mafias there, which have continued in various forms up to the present day (Gambetta, 1993).

Postcommunist countries allowed a natural experiment to test this theory of mafia emergence. By the 1980s, many socialist countries were implementing moves away from the sluggish command economy, including the Soviet Union itself under Mikhail Gorbachev. With the end of the state's monopoly over production, which was complete by 1991, vivid stories began to emerge of violence-wielding groups fighting out mafia wars in newly independent Russia, reaching its bloody zenith in 1992 and 1993. After 70 years of totalitarianism, trust in the state and within society was low, Russia was virtually bankrupt, and justice institutions corrupted. Despite property rights being largely unprotected by the state, there was huge competition to take up the plethora of business opportunities in the new economy and to grab a slice of the privatization of state assets. In this context of low trust, increasing economic transactions, and virtual state collapse, mafia-like groups emerged in Russia and in most postcommunist countries.

The supply of entrepreneurs of violence came in many forms in different places depending on the reputational assets of the groups in question within their societal and cultural context. In Bulgaria, wrestlers, who were respected in society and had created wealth and connections due to travel to international competitions during the communist period, founded the strongest protection-providing groups. In Kyrgyzstan, former police officers predominated. In Russia, a range of actors, from veterans of the Soviet War in Afghanistan, to martial arts experts, and prison gang bosses known as 'thieves-in-law' or *vory-v-zakone* who had held pride of place in Soviet counterculture, competed and colluded to monopolize the provision of private protection. They were joined by the thousands of ex-secret service officers who had been released from service in order to shrink the police state in the 1990s. Due to their former links with the state, these officers were able to obtain licenses and form legitimate private protection agencies slowly forcing out erratic and violent competitors such as Chechen and Slavic gangs (Volkov, 2002).

Japan provides another example. While the Yakuza claim an ancient lineage going back to the seventeenth century, their primary activities were speculation and selling illicit goods on the black market as well as engaging in organized gambling. The Yakuza, then, for most of its history was engaged in organized crime but not mafia activities, as we have defined these. The end of feudalism in the 1860s did not change these essential functions of the Yakuza as the state maintained control over the process of economic change and absorbed potential violent entrepreneurs such as the samurai warrior class (abolished in 1873) into new military structures. Instead, the Yakuza only legitimately became a mafia in the post-1945 period when they began to provide protection services. Once again, a key feature of this time period was the decline of state power and huge upheaval in society with Japan's defeat in World War II. Under US control, Japan's police and security structures were virtually dismantled, Japanese police had no jurisdiction over Chinese and Korean former prisoners of war, and there was an influx of battle-hardened Japanese exsoldiers, as well as a growing black market demand for drugs (Hill, 2006). This is the point at which conditions were primed for the Yakuza, with its culturally specific history and imagery, to gain new resonance in society and take on the form of a mafia – to attempt to monopolize protection over given territories.

What Do Mafias Do?

As we have seen, mafias on the definition provided here, in essence, extort and sell protection to businesses and resolve disputes among both criminal and legitimate actors. In the underworld of crime where contracts mean nothing and the state cannot be used to enforce agreements, mafias find a natural role. Where they find it beneficial and not destabilizing, mafia members might also engage in illegal activities such as illegal arms or drug trafficking. They will also naturally look for ways to launder money, setting up companies and bank accounts fronted by straw men often in other jurisdictions. Certain criminal markets, heroin, for example, are seen as potentially highly violent and volatile, and mafias may be loath to engage in them. Other markets, such as human smuggling, may be difficult to enter as the nature of the activity does not benefit rather static, hierarchically structured, and relatively visible mafias and instead favors flat, highly fluid, and transitory, opportunistic networks (Zhang and Chin, 2003). Moreover, due to how sensitive mafias are to reputational damage, highly morally dubious activities such as human trafficking, running prostitution rings, or kidnapping are often discouraged or at least mafia members might engage in them individually without using the group name.

Thus, the volatile underworld and complex illegal markets might not always be the best source of revenue for the mafia. As well as governing the underworld of organized crime then, mafias find various roles in the legitimate economy. Legitimate companies, often stable and working from fixed locations, provide clear opportunities for mafias to extract tribute as well as potentially supply front and no-show jobs for individual mafiosi that bring in regular sources of income to show the taxman. The mafia can force itself into industries and maintain a presence there through the corruption of local politicians and police departments. Mafias can also be invited to provide its services in a specific market or industry where this is beneficial. Examples can be small scale, such as the Triad-enforced cartelization of bus routes in Hong Kong or of wholesalers in fruit and vegetable marketplaces in southern Italy and post-Soviet Georgia. In these cases, the mafia service is highly beneficial for market players, as the markets in question have very low start-up costs, demand remains high regardless of price, there is little product differentiation, and a large number of small firms compete (Gambetta and Reuter, 1996). The mafia prevent entry of competitors, facilitate cartelization, and organize the market. The cost of the mafia service should ideally be outweighed by the benefits of the reduction in competition.

Mafia involvement can extend to large legitimate industries and have pervasive effects. There are instances of mafias engaging in rigging tenders by enforcing a cartel agreement

among conspirators who agree not to undercut each other and to take turns in winning given tenders. Some large-scale industries become suffused with mafia activity. In the waste hauling industry in New York, for example, the mafia invested themselves in haulage companies and infiltrated professional associations. They imposed a cartel through controlling and enforcing the division of customers among haulage companies. In this way, Cosa Nostra were able to effectively divide up routes and customers, enforcing a cartel in places such as Long Island and New Jersey. Business owners were charged top rates for the unavoidable waste haulage service and, unable to find a different company willing to make a better offer, paid up, passing on the extra costs to customers.

Construction is another industry highly susceptible to mafia influence (Jacobs et al., 2001). The mechanism for control of construction in cities such as New York City and Montreal involves control over labor unions. The mafia was able to effectively threaten construction companies with delays and disruption to schedules by being able to instigate strikes. As many different types of workers are required to complete a building project – concrete pourers, scaffolders, teamsters, electricians, and so on – any disruption to any of these groups could cause delays and incur huge costs. Moreover, construction sites made for unmonitored spaces to conduct illicit business and for mafiosi to take up no-show jobs.

How Do Mafias Organize?

Mafia 'families' are usually not families at all, but hierarchies consisting of a limited number of mafiosi drawn from networks of socially close people. Classically, the hierarchy involves a boss, a small number of authoritative members close to him who control the equivalent of a team leader who, in turn, runs small groups of heavies doing the dirty work on the ground. Ritual initiation draws a boundary that establishes the rights and obligations of these different classes of members, with uninitiated members having fewer privileges and often being required to make heavy material and temporal investments. Such distribution of rights and obligations, and the expectations regarding the behavior of members are often written down or passed on verbally as a 'code of honor.' Such hierarchical structuring, the use of ritualized initiation, and codes of behavior are a rather uniform aspect of mafias from Italy to Russia, Japan, China, and in mafia-like groups such as the Hell's Angels, Mexican drug cartels, or US prison gangs.

It is no accident that hierarchy is a key component of mafias. Deferring to a boss allows conflict among members to be resolved efficiently, judgments brought and punishment meted out through established mechanisms. This creates stability, a benefit that outweighs the potential costs of a boss abusing his or her power. In case a boss misbehaves, or indeed any member, codes of honor serve a function analogous to a constitution or criminal code, defining the limits of the power of individual members, and stipulating behavioral norms and accompanying punishments for infringements. Such codes are highly detailed in the case of prison gangs such as the Mexican Mafia and La Nuestra Familia in California (Skarbek, 2011). These groups act as mafias in providing protection in prison as well as extorting and controlling drug-dealing gangs on the street. Their codes are presumably deliberately elaborate due to the highly uncertain environment of the prison and the potential for the abuse of power by members and bosses.

Ritual initiation is also a crucial element of mafias and results from the high barriers to entry mafias place on potential recruits. Recruitment is vital to the strength and resilience of a mafia, in that members need very specific skill sets and personal qualities, there is always a threat of undercover police infiltration, and recruiting poor members can affect the collective reputation of the group. As will be discussed below, reputation is utterly vital for survival. Most mafias require a significant self-imposed cost to be borne by a candidate upon entry, establishing credible commitment and binding them into the fortunes of the group. Committing murder, for example, is a cost that an undercover officer or undedicated candidate could not bear. The Numbers Gang, a South African prison gang, requires candidates to attack and wound a prison warden as part of initiation. Russian *vory* or thieves-in-law require that candidates have spent significant amounts of time in prison and have never worked or had any dealings with the state. Initiation can itself be physically excruciating. Yakuza candidates must mutilate themselves by removing their little fingers upon initiation. The Triads, on the other hand, while having relatively low barriers to entry, impose a high exit cost: death. Once joined, leaving is not an option. Again, in the Triads this potential future cost is constituted by taking oaths during elaborate ritualized ceremonies. The important functions played by these often bizarre practices perhaps explain why they are so prevalent among such a variety of groups from different parts of the world.

How Do Mafias Expand?

In the 1960s, the American criminologist Donald Cressey argued that Cosa Nostra was an expansionist rational bureaucracy slowly creeping through the American nation (Cressey, 1969). There has since been much questioning of Cressey's position. Economic arguments on the expansion of organized crime reasoned that organized criminal enterprises were much more likely to be small and localized given the conditions of risk they operated in (Reuter, 1985). These hierarchical mafia groups struggle to expand outside their locale, as their position is very much maintained by their reputation, and, being unable to advertise their services widely as a legitimate firm might, this will not travel too far. Furthermore, moving to new territory increases costs on monitoring members of the group operating at a distance and can create the grounds for deadly conflict with already established groups in new territories. Violence and conflict are undesirable, as they draw unwanted attention from law enforcement.

Despite this, there is evidence that certain groups have achieved high levels of coordination across large distances and territories as well as successful transplantation out of original locales. Wiretaps of Sicilian mafiosi in a bar in Montreal in the 1970s as well as other sources point to the existence of a mafia commission in Sicily, an organizing body over and above individual mafia families that enables coordination and regulation of activities and resolution of disputes over a large territory. Similarly, the Russian and post-Soviet mafias are

marked by high levels of mobility and appear to coordinate and communicate through the use of practices learned in the Gulag, the widely dispersed archipelago of labor camps in the Soviet Union. The ability to coordinate in that repressive environment has been transferred to post-Soviet organized crime enabling collection of regional collective funds (known as *obshchaki*) and collective decision making at so-called thieves' courts (known as *skhodki*).

However, the economic approach to mafia organization suggests that such high levels of coordination are rare. This approach places a great emphasis on the initial start-up costs of running a mafia firm in a given locale. There are initially large costs and high risks in producing violence in order to establish a level of credibility in the threats produced by potential mafia groups. Once these start-up costs have been taken on, however, mafias can reap economies of scale. Once credible commitment to use violence has been signaled, clients will pay for protection, dispute resolution, and other services without testing the enforcement mechanisms present behind these services (Reuter, 1985). In essence, at this stage just the word of the mafioso can be converted into economic value. Reputation, therefore, substitutes perfectly for the mafia's main production unit, violence, at no extra production cost. This is most likely why mafia usually have the most obvious and well-known 'brand names' ('Cosa Nostra,' 'Yakuza,' and so on), protected by violence, of any type of organized crime group. The brand does all the work in the mafia business. Expansion on the back of this reputation, therefore, makes economic sense and clients will seek it out: once one business has protection, the probability of victimization is heightened for unprotected businesses as crime is displaced. Thus, monopoly is the most likely outcome, as clients will rationally transfer their business to whoever possesses the strongest reputation based on the most credible threat of violence. Eventually, however, a geographical limit will be reached, past which the reputation does not travel, or where another mafia group has a stronger reputation for credible threats. The costs of using violence must then be taken on once again and mafiosi have to decide whether the costs of conflict and potential attention from law enforcement will be outweighed by the benefits of expansion.

Expansion does occur though, and sometimes between countries and continents. Mafias can expand into new territories through push factors, where repressive government policies in one country or region displace mafia activity into another, or through pull factors, where mafias make strategic decisions to move based on the perceived benefits of a receiving country. These factors include lax law enforcement and material opportunities. Federico Varese shows that push factors are more likely to be the reason for mafia transplantation but that mafias do not move that easily or that willingly (Varese, 2011). For example, in the mid-twentieth century, the Italian state believed it could help develop the south and reduce mafia influence by resettling mafiosi in northern regions, this had the unintended consequence of simply displacing mafia activity to northern Italy. The post-Soviet republic of Georgia enacted an extremely harsh antimafia campaign in 2005. This has partially caused a mafia war between Georgian expat groups in Moscow and expansion into Europe as witnessed by continent-wide raids in 2010 and 2013, which netted over 100 organized criminals from the post-Soviet space, the majority Georgians, as they flooded out of the country in the wake of government policy (Slade, 2012).

Once pushed out of a territory, other factors such as the structure of the local economy in the receiving country, the presence of a general migratory outflow from one country to another, and the levels of trust and demand for protection in the receiving locale, all effect the success of mafia transplantation. For example, the large exodus of Sicilians to the US at the end of the nineteenth century included some individuals who were members of mafias back in Sicily. Once in the US and in particular in New York City certain factors were conducive to the establishment of an overseas filial of the Sicilian mafia. The rise of prohibition and well-intentioned but misfiring police reform in New York in the 1910s resulted in a large unprotected market opening up. This facilitated the rise of Cosa Nostra and the establishment of mafia dynasties that are still with us today in New York, otherwise known as the Five Families (Varese, 2011).

Why and When Do Mafias Go into Decline?

Few criminologists are prepared to declare any mafia dead. It would be difficult to know for one thing whether a mafia, as a secretive organization, truly has died or it is simply dissimulating disappearance. Moreover, there are plenty of people for whom the mafia is a useful construct (Gambetta, 2009). Police departments can get extra funding and government support by playing up to the threat of a conspiratorial, hydra-headed mafia monster (Smith, 1976). Governments can blame shadowy mafia figures or immigrant alien conspiracies for problems of crime and violence and other failings, while 'transnational organized crime' can be invoked to demonize and justify meddling in other countries, particularly in the age of the 'war on terror' (Woodiwiss and Hobbs, 2009). There is no doubt that 'organized crime' and 'mafia' are terms that are used and abused for political purposes; figures concerning the economic power of organized crime groups and mafias can be but mere shots in the dark. Nevertheless, whatever the motive, states do occasionally attack the mafia. There are numerous examples of antimafia policies from different countries, all with varying levels of success. What factors might weaken the strength of mafias?

Many studies focus on state-centric factors to explain decline. Governments must make strategic decisions about when to attack the mafia, as it may prove to be an expensive activity which reaps little gains in crime control or at the ballot box. In some countries, governments face violent backlashes if the mafia is challenged. In Brazil, for example, the First Command of the Capital (the PCC, a prison gang) is able to virtually shut down the city of Sao Paolo if it chooses to oppose the state. Thus, mafias and states often simply exist in wary equilibrium of each other, often maintaining a negotiated connection through corrupt links on local levels. Still, at certain moments states can launch full-out attacks on mafias, though success is rarely a given.

In the US, the Racketeer Influenced and Corrupt Organizations Act of 1970 was a relatively successful attempt to curb organized crime. It allowed greater punitive measures against those who order commission of crime and engage in patterned acts of racketeering. On top of this, demographic change, reform

of the policing of cities granting increased federal oversight, and improved legislation have reduced corruption and the infiltration of traditional mafias on a local level in the US (Reuter, 1995). Administrative measures that aim at increasing transparency, such as the creation of trusteeships in organized crime-infiltrated industries, are credited with decreasing the influence of the mob in cities such as New York and Amsterdam.

In Italy, direct antimafia strategies, including legislation and law enforcement reform as well as civil society mobilization, implemented after the mob killings of the judges Falcone and Borsellino in 1992 have been relatively effective, at least in Sicily (Paoli, 2007). These measures include new law enforcement techniques and heightened police coordination including longer investigation times, the greater use of wiretapping, entrapment, state's witnesses, and the criminalization of activities that aid the mafia.

However, a simple emphasis on the policies of the state and the changes in the socioeconomic environment does not account for the vulnerabilities or resilient qualities that decide whether a mafia survives or succumbs to state attack. Strength in recruitment, for example, is a key variable in survival. Similarly, internal and external structuring of relations is vital for robust organization, effective adaptation, flows of information, and efficient monitoring and sanctioning. Yet, the organization of criminal activities is often risky, and constantly changing in conditions of permanent insecurity and uncertainty. Some scholars capture this by talking instead of *organizing* crime rather than the more presumptive *organized* crime (see, e.g., McIllwain, 2001; Morselli, 2008). This emphasizes the contingent-networked nature of organized crime and how these networks can possess vulnerabilities and weak points, or be poorly adapted to changing environments and disorganized in front of state power.

In explaining the decline of mafias, the primary emphasis then is on law enforcement, governance, political processes, and in some cases the structure and organizational change within mafias themselves. By combining an analysis of state policies, societal and economic change, and the adaptations of mafias in terms of recruitment, internal and external relations, depth and sources of resources available, and the shifting motivations of mafiosi, criminologists can create convincing accounts of why mafia groups can in some cases survive state attack, while in other cases be greatly weakened by it. Focusing on organizational vulnerabilities of organized criminals might help understanding of how best to tackle established and seemingly indestructible mafia groups, helping to explode the myth that the mafia is forever.

See also: Crime Trends and Debates; Crime: Organized; International and Comparative Criminology; Media, Crime and Crime Waves; Quantitative Methods and Crime; Transnational Organized Crime.

Bibliography

Blok, A., 1974. The Mafia of a Sicilian Village 1860–1960: A Study of Violent Peasant Entrepreneurs. Blackwell, New York and Oxford.
Chu, Y., 1999. The Triads in Business: A Brief Introduction, first ed. Routledge, London.
Cressey, D., 1969. Theft of the Nation: The Structure and Operations of Organized Crime in America, vol. 174. Transaction Publishers.
Gambetta, D., 2009. Codes of the Underworld. How Criminals Communicate. Oxford University Press, Oxford.
Gambetta, D., 1993. The Sicilian Mafia. The Business of Private Protection. Harvard University Press, Harvard.
Gambetta, D., Reuter, P., 1996. Conspiracy among the many: the mafia in legitimate industries. In: Fiorentini, G., Peltzman, S. (Eds.), The Economics of Organised Crime. Cambridge University Press, Cambridge.
Hill, P.B.E., 2006. The Japanese Mafia: Yakuza, Law, and the State. Oxford University Press, Oxford.
Jacobs, J.B., Friel, C., Raddick, R., 2001. Gotham Unbound: How New York City Was Liberated from the Grip of Organized Crime. NYU Press, New York.
McIllwain, Jeffery Scott, 2001. Organizing Crime in Chinatown – Race Racketeering in New York City, 1890–1910. McFarland and Company, Jefferson, NC.
Morselli, C., 2008. Inside Criminal Networks. Springer.
Paoli, L., 2007. Mafia and organised crime in Italy: the unacknowledged successes of law enforcement. West European Politics 30 (4), 854–880.
Reuter, P., 1985. Disorganised Crime: Illegal Markets and the Mafia. MIT Press.
Reuter, P., 1995. The decline of the American mafia. The Public Interest 120, 89–99.
Schelling, T.C., 1984. Choice and Consequence. Harvard University Press.
Skarbek, D., 2011. Governance and prison gangs. American Political Science Review 105 (4), 702–716.
Slade, G., 2012. No country for made men: the decline of the mafia in post-soviet Georgia. Law and Society Review 46 (3), 623–649.
Smith Jr., Dwight C., January 1976. Mafia, the prototypical alien conspiracy. Annals of the American Academy of Political and Social Science 423, 75–88.
Tilly, C., 1985. War making and state making as organised crime. In: Skocpol, T., Evans, P., Rueschemeyer, D. (Eds.), Bringing the State Back. Cambridge University Press.
United Nations Office on Drugs and Crime, 2004. Convention against Transnational Organized Crime and the Protocols Thereto. United Nations, New York.
Vaksberg, A., 1991. The Soviet Mafia. St. Martin's Press, New York.
Varese, F., 2001. The Russian Mafia: Private Protection in a New Market Economy. Oxford University Press, Oxford.
Varese, F., 2010. What is organized crime?. In: Varese, F. (Ed.), Organized Crime: Critical Concepts in Criminology, first ed., vol. 1. Routledge, London, pp. 1–33.
Varese, F., 2011. Mafias on the Move: How Organized Crime Conquers New Territories. Princeton University Press.
Volkov, V., 2002. Violent Entrepreneurs: The Use of Force in the Making of Russian Capitalism. Cornell University Press, Ithaca.
Weber, 1992. Economy and Society, new ed. University of California Press.
Woodiwiss, M., Hobbs, D., 2009. Organized evil and the Atlantic alliance. British Journal of Criminology 49 (1), 106–128.
Zhang, S., Chin, K.-L., 2003. The declining significance of triad societies in transnational perspective. British Journal of Criminology 43, 469–488.

Relevant Websites

www.primecrime.ru – Biographies and up to date news on the Russian and Post-Soviet Mafias (in Russian).
http://www.organized-crime.de/OCDEF1.htm – Definitions of Organized Crime.
http://www.exlegi.ox.ac.uk/ – Extra-Legal Governance Institute at Oxford University.
https://reportingproject.net/occrp/ – Organized Crime and Corruption Reporting Project.
http://www.blogdelnarco.com/ – Undercover reporting on Mexican cartel wars (in Spanish).
http://www.unodc.org/unodc/en/index.html – United Nations Office on Drugs and Crime.

Magic, Anthropology of

Vassos Argyrou, University of Hull, Hull, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article makes an anthropological study of anthropological interpretations of magical beliefs and practices in non-Western societies. It contends that these interpretations must be situated and understood within the discipline's own culture whose fundamental guiding principle is the notion of 'human unity.' It further contends that, although the target of these interpretations is ethnocentrism, paradoxically, they turn out to be ethnocentric themselves.

The magical, broadly and somewhat loosely understood as belief in the existence and efficacy of supernatural forces, has been an area of intense anthropological interest from the inception of anthropology as an academic discipline in the late nineteenth century to the present day. This interest has little if anything to do with a penchant for the study of the occult and the exotic and much to do with the discipline's most fundamental underlying assumption and guiding principle – the idea of human unity. On the one hand, this idea refers to what E.B. Tylor posited in the nineteenth century as 'the psychic unity of mankind,' the ability to think rationally, which many of Tylor's contemporaries doubted in the case of what they perceived as 'primitive people.' On the other, and over and above Tylor's postulate, it refers to cultural equality, implicit in the twentieth-century notion of cultural relativism and assumed in all anthropological critiques of ethnocentrism, effectively Eurocentrism – the belief in the cultural superiority of Europe and the West.

To put it schematically, the question that anthropologists were called upon to answer or themselves felt the need to answer was the following: if native populations are rational people 'like us,' as the principle of human unity suggests, how is one to explain their involvement in apparently irrational beliefs and practices like magic? Anthropology's answer to this question is entangled in a paradox, which is only now beginning to come to light and be discussed. In explaining native involvement in magical beliefs and practices as something reasonable, anthropologists also explain native culture away. Unwittingly and unwillingly, they end up being ethnocentric themselves.

This article examines anthropological explanations of magic in the context of the various efforts to uphold the principle of human unity. It begins with social evolutionism and explores the views of E.B. Tylor, the paradigmatic figure of Victorian anthropology. It then moves on to the early twentieth century and the work of two key figures, Malinowski and Evans-Pritchard, both of whom moved away from the Victorian understanding of magic as a pseudoscience and began to treat it as a symbolic phenomenon. By the 1960s, symbolic interpretation becomes dominant and the article explores the work of two anthropologists who played a pivotal role in the consolidation of this paradigm – Mary Douglas and Clifford Geertz. It then moves on to two studies within the Marxist problematic, which link magic to local experiences of capitalism: Michael Taussig's explanation of devil worship in South America and Jean and John Comaroffs' discussion of 'occult economies' in postapartheid South Africa. The article concludes with a brief reference to the recent recognition that symbolic interpretation explains native culture away and the suggestion from some quarters that the notion of culture ought to be replaced by the notion of ontology.

Social Evolutionism: Magic as a 'Pseudoscience'

The nineteenth century was the time when anthropology became a recognized academic discipline. It was also the time when the achievements of the natural sciences had captured public imagination and for many provided tangible proof of the superiority of Europe vis-à-vis the rest of the world. Anthropologists were not immune to this vision of the world. Nonetheless, some, most notably Tylor, were not prepared to accept the more extreme form of this vision, which treated the magical beliefs and practices of native populations as an indication of a radically different mentality – so different, in fact, as to cast doubt on their humanity. Tylor's problem, therefore, was to explain magic in a way that was consistent with his postulate of the psychic unity of mankind, in effect to demonstrate that, although 'primitive people' were apparently different from Europeans, in a fundamental sense they were also the same.

For Tylor (1874), there was no doubt that magic was "one of the most pernicious delusions that ever vexed mankind" and characteristic of "the lowest known stages of civilization." It was a 'pseudoscience,' a system of thought and practice that attempted to manipulate the empirical world but was unable to do so. Yet this did not mean that it was the product of inherently different minds. The human mind, he argued, operates everywhere and at all times with the same principles of thought – by associating ideas on the basis of analogy and cause and effect. 'Primitive' people, being at the lowest stage of intellectual development, a stage characterized by the kind of innocence one finds in children, were particularly prone to making errors and magic was a paradigmatic example. To use Tylor's own terminology, 'primitive' people mistake 'ideal for real connections,' that is, they treat events that are only accidentally related as if they cause one another. There is nothing in reality that connects the crowing rooster with the rising sun, but because the two events occur regularly and sequentially, 'primitive' people assume that the former causes the latter. Magic is

the practical application of such misconceptions. The rooster is made to crow so that the sun will rise. The moral of Tylor's story was that 'primitive' people, being like children, will eventually grow up and begin to think and act rationally.

Twentieth-Century Paradigms

For reasons that are yet to be fully explained, by the early twentieth century, the social evolutionism of Tylor and his contemporaries and the ethnocentrism that guided it had become less palatable. And so was their explanation of magic. The issue was no longer to demonstrate that 'primitive' people are *potentially* rational but that magic itself was, if not strictly speaking rational, certainly a reasonable way of dealing predicaments that are often universal. The first anthropologist to treat magic along these lines was Bronislaw Malinowski.

Magic as a Source of Hope and Optimism

Malinowski spent the years of the First World War among the natives of the Trobriand Islands of the southwest Pacific and developed a different understanding of the place of magic in native life. In his extended essay *Magic, Science and Religion* (1954) he sets out to show that much of what had been written about magic was speculative and had little basis in reality. To this effect, he engages both Tylor and the French philosopher-turned-anthropologist Lévy-Bruhl. The former, Malinowski argues, depicts natives as if they were intellectuals seeking to find answers to theoretical questions, and the latter as being "hopelessly and completely immersed in a mystical frame of mind" (1954). In contrast to both, Malinowski found the natives of the Trobriand Islands to be practical people concerned with practical things: fishing, gardening, building canoes, and tribal festivities. This is not to say that they did not also practice magic. Yet they did so only under specific circumstances.

In a celebrated example, Malinowski shows that when fishing in the lagoon, the Trobrianders did not resort to magical rites; they relied instead on their knowledge and skill. In open-sea fishing, however, which is "full of danger and uncertainty, there is an extensive magical ritual to secure safety and good results" (1954). This and other similar examples, Malinowski argues, demonstrate that natives turn to magic only in situations of emotional stress, when they reach the limits of their knowledge and cannot control the outcome of their actions. Malinowski concluded, therefore, that far from being a cognitive mistake or a sign of mystical mentality, magic fulfills an important psychological function: it ritualizes optimism and enhances human confidence and hope.

Evans-Pritchard: Transition to Symbolism

Although highly suspicious of evolutionism, Malinowski was not able to go beyond the other tenet of Victorian anthropology – that the primary aim of magic was the attainment of practical results. Nor, by extension, was he able to go beyond the notion that magic was ultimately a pseudoscience. The shift from this utilitarian perception to one that views magic as a meaningful and meaning-generating phenomenon came with Evans-Pritchard (1976) in his classic study of Zande witchcraft. Evans-Pritchard was working within the broad sociological problematic developed by Emile Durkheim, which emphasized social stability and treated religion as a social institution that contributed to it. In his study of the Azande of Central Africa, Evans-Pritchard extended the argument to witchcraft. At the same time, he continued the debate with Lévy-Bruhl that Malinowski began. It is in this confrontation with the French philosopher that he lays the ground for the paradigm that was to dominate subsequent anthropological studies of magicoreligious systems.

Evans-Pritchard analyzes Zande witchcraft along two axes. The first explores how witchcraft contributes to the cohesion of Zande society – its social function. The other examines the condition under which the Zande have recourse to it – its cultural function. Witchcraft embraces a system of values that regulate Zande behavior. No one knows for certain who might be a witch, but spiteful, moody, ill-tempered, and bad-mannered individuals are prime suspects. Belief in witchcraft acts in a way that encourages more positive dispositions and helps to curb antisocial behavior. If the Zande suspect that their neighbor is a witch, they are careful not to offend him/her from fear of being bewitched. Similarly, spiteful and ill-tempered individuals do not wish to be accused of being witches from fear of revenge and are therefore likely to change their behavior. In this way, says Evans-Pritchard, friction and conflict are kept in check and the cohesion of Zande society is maintained.

Yet social stability is hardly what the Zande have in mind when they make witchcraft accusations. Their aim rather is to explain misfortune in a way that makes it socially relevant and meaningful. Their explanation no doubt refers to a supernatural cause – witchcraft – but this does not mean that the Zande are ignorant of the natural causes of events, much less that they have a mystical frame of mind as Levy-Bruhl had argued about 'how natives think.' To use a well-known example, sometimes old granaries collapse and injure people and this, like other accidents, is attributed to witchcraft. The Zande are well aware that wood decays and that termites eat away the foundations of granaries; they know that it is because of these natural causes that old granaries collapse. What they try to explain through witchcraft accusations is the timing of the event. Why should a particular granary collapse "at the particular moment when these particular people were sitting beneath it? Through the years it might have collapsed, so why should it fall when certain people sought its kindly shelter?" (1976) Witchcraft offers an explanation over and above natural causation, one that links the collapse of the granary with the fact that certain people were at that point in time sitting underneath it. We, Evans-Pritchard points out, do not have an explanation for the intersection of the two events; we say that they have independent causes and that their intersection is an accident. The Zande provide the missing link; they say it is witchcraft. By making the event socially relevant they are not so hopeless. They can act on it.

Magic as a Symbolic Phenomenon

After Evans-Pritchard's ground-breaking book, the notion that magic is a symbolic phenomenon, something that reflects underlying social and cultural realities rather than an irrational

practice becomes the central argument in most anthropological studies. In what follows, the author discusses several well-known works, beginning with Douglas's (1966) famous analysis of pollution and taboo (*see* Taboo).

In this book, Douglas deals with the accusation that 'primitive' people do not make a distinction between the holy and the unclean – an accusation whose implication is that 'primitive' societies have a magical orientation toward the physical world; that people in these societies believe nature to be inhabited by unpredictable and often malevolent supernatural forces; and that contact with unclean things, such as menstrual blood or corpses, brings one perilously close to these forces.

Douglas's response to this accusation is that notions of pollution and taboo are symbolic expressions of conceptual and social disorder, the idiom through which people come to terms with unclassifiable things and intractable social contradictions. Dirt, Douglas points out, is matter out of place, the inevitable outcome of every attempt to classify things and establish a conceptual grip on the world. It is the result of a universal human predicament, the need to transform chaos into a cosmos, that is, a meaningful world. Those things that do not fit our conceptual schemes and hence contradict and undermine them are everywhere designated as dirt; they become dangerous and are either suppressed or avoided. The threat is real enough, since unclassifiable things undermine our ability to construct and lead meaningful lives. At the level of culture, this danger is expressed through beliefs in pollution and defilement and concomitant rules of avoidance.

Contradictions in conceptual schemes are often reproduced in contradictions at the social level. Take, for example, the notion of sexual pollution and the related idea that it is dangerous for men to come into contact with menstrual blood. This notion, Douglas points out, is widespread among New Guinea tribes, but it is particularly pronounced among the Mae Enga. In this tribe, men believe that sexual intercourse weakens male strength so that even within marriage it is reduced to the minimum necessary for procreation. This belief may appear as an irrational fear, but as Douglas explains, the taboo reflects conflicts embedded in Enga society. The latter consists of exogamous clans that compete fiercely for prestige and power and are hostile to another. The exogamous clan system, however, forces men to marry outside their own clan, that is, women who come from the enemy's camp. In their pollution beliefs and practices of sexual avoidance, the Enga are trying to overcome symbolically a fundamental social contradiction – building marriage and family on enmity.

Less than a decade after Douglas's book, Geertz (1973) wrote a seminal essay that was to consolidate the symbolic approach in the study of magicoreligious systems. Geertz's essay is entitled Religion as a Cultural System but the reference to 'religion' should be understood to mean any metaphysical system, including magic. Geertz's essay is inspired by, and draws on, Evans-Pritchard's discussion of Zande witchcraft. It also draws on Max Weber's religious sociology, but unlike the latter who was working within an evolutionary paradigm and treated magic as historically prior to religion, Geertz does not make a distinction between the two. Geertz's argument is that 'religion' is a cultural system that helps people to maintain faith in the ultimate meaningfulness of life. It does so by accounting for all those anomalies in human experience – paradoxes, puzzles, ambiguities – that threaten to undermine the general order of existence. There are three points, according to Geertz, where chaos, "a *tumult* of events which lack not just interpretations but *interpretability*" threatens to enter into the world: at the limits of people's ability to explain, endure suffering, and make sound moral judgments. Coming up against these limits time and again, Geertz (1973) points out, "sets ordinary human experience in a permanent context of metaphysical concern and raises the dim, back-of-the-mind suspicions that one may be adrift in an absurd world." It is at such difficult moments that religion intervenes and affirms the ultimate meaningfulness of the world. It does so not by denying that there is ignorance, suffering, and injustice in the world, but rather by denying that such things are intrinsic to reality, an inescapable fact of the world.

The last two works to be examined are Taussig's (1980) book *The Devil and Commodity Fetishism in South America* and a recent article by Comaroff and Comaroff (1999) entitled Occult Economies and the Violence of Abstraction: Notes from the South African Postcolony. Both are influential works. Although situated within the Marxist problematic, which has been historically unreceptive to the significance of magicoreligious systems, they decisively adopt a symbolic approach. They interpret native beliefs in the magical and the occult as a means of making sense of capitalism and dealing with its impact on everyday life.

In rural Colombia, native people working on sugarcane plantations often enter into a secret contract with the devil. In return for their soul, the devil helps them work faster and increase their production and hence their wages. Such contracts, however, are said to have several negative consequences. To begin with, it is said that there is no point in investing the extra money in capital goods, such as livestock or land. The money earned with the help of the devil is believed to be inherently barren: the animals will die and the land will become sterile. Moreover, those who enter into such contracts are said to die young and in pain. In short, even though people improve their conditions in the short run, the long-term effects of dealing with the devil are destructive. How are these beliefs and practices to be interpreted?

The devil, Taussig argues, signifies the way in which local people are trying to come to terms with the capitalist relations of production. It also expresses their evaluation of, and resistance to, the new way of life. Having been uprooted from their ancestral homes, the local peasants are now forced into wage labor. Their traditional way of life based on reciprocity and cooperation has been displaced by the capitalist mode of production, exploitation by the landowners, and competition among themselves. It may be the case that they are now financially better off than before, but as the devil stories suggest, the quality of their lives has drastically deteriorated. The devil is an apt symbol for expressing the new order of things.

The Comaroffs, (1999) article is concerned with the deployment of magical means for the attainment of material ends – the 'occult economies' of the title – in postapartheid South Africa. These include a wide range of phenomena: from ritual murder of people suspected of witchcraft to the sale of body parts to be used in potions for success in one's pursuits

and the production of zombies who work for free and enrich in this way their 'employers.' Although such practices are not new, there has been a dramatic increase in the postrevolutionary period and the aim of the article is to explain 'why *now*?' The authors point to what they call 'millennial capitalism,' the promise of prosperity for all and the realization that even in the postapartheid period it is still only the very few that actually prosper. Magic is a culturally familiar idiom through which people attempt to make sense of this reality and the 'mysterious mechanisms of the market.' If the majority is left out of the promise of prosperity, many people reason, it must be because 'arcane forces are intervening in the production of value, diverting its flow for selfish purposes.' This, the authors argue, explains both the pursuit of new magical means for attaining wealth and the effort to eradicate those who are thought to monopolize it through such means.

Culture or Ontology?

The author has already suggested that contemporary interpretations of magic are entangled in a paradox: in trying to combat the ethnocentrism that portrays magic as an irrational practice, they end up, unwittingly and unwillingly no doubt, being ethnocentric themselves. It may be the case that magic ritualizes hope and optimism, that witchcraft makes accidents meaningful, that sex avoidance expresses social conflicts, that the devil symbolizes the nature of capitalism, and so on but this is not how the people involved in these beliefs and practices understand them. As far as they are concerned, magic before a fishing expedition in the open sea ensures safety and a good catch, witchcraft is a force in the body of the witch that flies at night and attacks its unsuspecting victims, contact with menstrual blood causes serious harm, and the devil is a malevolent supernatural force. If anthropological interpretations are valid, the inescapable conclusion is that natives suffer from a 'cultural unconscious' (Argyrou, 2002). They may not be irrational but they are certainly ignorant of the true meaning of their magical beliefs and practices.

It is problems such as these that seem to be behind a recent call to replace the notion of culture with that of ontology (e.g., Viveiros de Castro, 2003; Holbraad, 2010). The problem with cultural analysis, according to the argument, is that it posits one reality that is open to different representations – one nature, many cultures. This is a problem because in this schema only one representation can be valid – usually the Western representation. An ontological perspective, by contrast, posits multiple realities, which means that other ways of life can no longer be treated as misrepresentations of reality. They are realities in their own right and if anthropologists are to understand them, they need to expand their conceptual horizons.

Can this theorization of other ways of life save the day? Without further elaboration of the thesis, it is too early to tell. One question, however, may be raised: does this thesis not posit a reality (of multiple ontologies) that the natives are once again not aware of?

See also: Collective Beliefs, Sociological Explanation of; Collective Memory, Anthropology of; Folklore; Mysticism; Myth in Religion; Myths and Symbols: Organizational; Tradition, Anthropology of.

Bibliography

Argyrou, V., 2002. Anthropology and the Will to Meaning: A Postcolonial Critique. Pluto Press, London.
Comaroff, J., Comaroff, J., 1999. Occult economies and the violence of abstraction: notes from the South African postcolony. American Ethnologist 26, 279–303.
Douglas, M., 1966. Purity and Danger: An Analysis of the Concepts of Pollution and Taboo. Praeger, New York.
Evans-Pritchard, E.E., 1976. Witchcraft, Oracles, and Magic among the Azande. Clarendon, Oxford.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Holbraad, M., 2010. Ontology is just another word for culture: against the motion. Critique of Anthropology 30, 179–185.
Malinowski, B., 1954. Magic, Science and Religion and Other Essays. Doubleday Anchor Books, Garden City, NY.
Taussig, M., 1980. The Devil and Commodity Fetishism in South America. The University of North Carolina Press, Chapel Hill, NC.
Tylor, E.B., 1874. Primitive Culture, vol. I. Holt, New York.
Viveiros de Castro, E., 2003. (Anthropology) and (Science). Manchester: Manchester Papers in Social Anthropology, vol. 7.

Magnetic Resonance and Molecular Imaging in Psychiatry

Oliver D Howes, MRC Clinical Sciences Centre, Imperial College-Hammersmith Hospital Campus, London, UK; and Institute of Psychiatry, King's College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Schlösser, volume 15, pp. 10148–10155, © 2001, Elsevier Ltd.

Abstract

Neuroimaging techniques provide *in vivo* measures of brain structure and biology. They have revolutionized the study of psychiatric disorders by enabling neurobiology to be linked to the clinical expression and development of disorders, and the effects of treatment or behavioral challenges. This article provides an overview of the main techniques – Magnetic resonance imaging, including functional magnetic resonance and magnetic resonance spectroscopy; positron emission tomography; and single photon emission computed tomography – and illustrative key findings in psychiatry to highlight their application. It concludes by discussing future directions.

Introduction

Psychiatric disorders are challenging to study, not least because it has proven hard to produce preclinical models that reproduce key aspects of their phenotypes. A number of neuroimaging techniques, such as computerized tomography, have been available for decades and provide a means of studying the neurobiology underlying these disorders. However, the development of magnetic resonance (MR) and molecular imaging techniques and their application to study the brain has a number of advantages over earlier techniques, and has contributed to an impressive expansion in understanding of the neurobiological basis of psychiatric disorders. Magnetic resonance imaging (MRI) provides data on brain structure that can be used to identify regional or global alterations in a disorder. Functional magnetic resonance imaging (fMRI) enables investigation of neuronal activation, generally during a cognitive or behavioral task or a drug challenge. Neurochemistry can be investigated using magnetic resonance spectroscopy (MRS). Positron emission tomography (PET) and single photon emission computed tomography (SPECT) are molecular imaging techniques that also provide data on brain chemistry, as well as enabling the interrogation of other aspects of brain function, including receptor levels, blood flow, and metabolism.

Neuroimaging has revolutionized the study of psychiatric disorders in a number of ways. First, the *in vivo* measures can be related to clinical features to provide insights into the nature of the brain alterations underlying specific aspects of the disorder. Second, they can be applied at different stages of the disorder to elucidate the biology underlying stages of the disorder and response to treatment. Third, the measures can be repeated to illuminate the nature of changes over time with the development and progression of the disorder. Fourth, neuroimaging can be used to aid drug development, for example, by identifying targets and determining whether a drug enters the brain and acts in the right region. This article discusses the major uses of neuroimaging in psychiatry and illustrates them with key examples from particular disorders. It is not intended to be all encompassing and suggestions for further reading are provided at the end.

Structural MRI

Principles

MRI takes advantage of the physical properties of atomic nuclei, such as hydrogen, that are electrically charged. These behave as charged spinning tops that produce a magnetic field. In the MRI scanner an external magnetic field is imposed, which causes these nuclei to align either parallel or antiparallel to the field. External electromagnetic radio frequency (RF) pulses are then used to move the spinning nuclei out of their initial alignment, which is accompanied by a change in their energy state. When the RF pulse is removed, the nuclei relax back to their equilibrium state and this change in their energy state is accompanied by the production of an RF signal, which is detected by coils within the scanner. This signal is reconstructed to produce an MR image. Protons in different tissues return to their equilibrium states at different relaxation rates, depending on the local environment of the protons, which provides information on the nature of the tissue. The application of additional magnetic fields results in predictable changes in the RF signal that can be used to provide spatial information. The basic unit of the resulting three-dimensional image is a voxel (volumetric picture element). Voxel size depends on the resolution of the scanner. In general, MR scanners with higher magnetic field strength (measured in Tesla) have higher resolution. The intensity of the MR signal in a given voxel depends on the proton density of the tissue in the voxel. As hydrogen atoms have protons, tissue with high water (containing two hydrogen atoms) content provides a high intensity signal.

Functional Magnetic Resonance Imaging

General Aspects

MRI techniques have been developed to provide information on the function of the brain as well as its structure. This enables functional changes to be investigated in central nervous system disorders or during behavioral, cognitive, or emotional tasks. The most widely used techniques in psychiatry are described below.

Blood Oxygen-Level-Dependent (BOLD) fMRI

Probably the most widely used fMRI approach is based on the BOLD-contrast imaging technique to provide an indirect index of change in neuronal activity. This takes advantage of the fact that there is a difference in the magnetic properties of blood depending on whether it is oxygenated or not, as deoxyhemoglobin is paramagnetic and oxyhemoglobin is diamagnetic. Neuronal activation results in a local change in blood flow that is beyond the metabolic demand, leading to a local increase in oxyhemoglobin and a decrease in deoxyhemoglobin. This change is associated with an increase in the T2* signal, which can be detected in the scanner. BOLD fMRI studies generally involve the subject performing a cognitive or other task in the scanner. The change in signal during this task relative to a baseline is then taken as an index of brain activation during the task. It is important to note two things about BOLD fMRI studies. First, they are based on signal change – this means that the baseline (control) condition can be as important as the task. Second, the signal change is generally quite small – in the order of one to five percent.

Studies usually have either a block- or event-related design. In a block design, one condition is presented for a period of time followed by the other condition, and this sequence is generally repeated. A limitation of this approach is the fixed order. Whereas, event-related designs model the hemodynamic response to each condition. This avoids the need to present conditions in a fixed order and enables timings to be varied.

Perfusion fMRI

Brain blood perfusion can be studied using MR either by using a paramagnetic contrast agent (such as gadolinium) that is injected into a vein, or changing the magnetic properties of the arterial blood entering the brain to 'label' it. The latter technique, arterial spin labeling, has the advantage of being noninvasive. Both techniques provide physiologically quantifiable measure, in contrast to BOLD fMRI.

Diffusion-Tensor fMRI

Water molecules are in constant random motion in the brain. Diffusion-weighted fMRI takes advantage of the magnetic properties of water, which can be detected in the scanner to index the motion of water molecules. In some ordered tissues, the motion of the water is constrained in certain directions. For example, in a bundle of parallel axons water can more readily diffuse along the main axis of the bundle than across the bundle – diffusion is anisotropic. Diffusion tensor imaging (DTI) is frequently applied to measure the degree of anisotropy in white matter tracts to detect differences in the integrity of tracts between patients and controls, for example.

Magnetic Resonance Spectroscopy

Magnetic resonance spectroscopy (MRS) is used to investigate neurochemistry. It is based on the fact that nuclei of one type resonate at different frequencies depending on their local chemical environment that depends on the molecular structure of a given neurochemical. This phenomenon is termed the chemical shift. The chemical shift thus provides a signature that can be used to detect particular neurochemicals in the brain. The most widely used MRS techniques in the brain are ^{1}H-MRS and ^{31}P-MRS. ^{1}H-MRS is well established for the detection of N-acetyl aspartate (NAA), creatine–phosphocreatine (Cr–PCr), glutamine, glutamate, aspartate, and lactate, and increasingly for gamma-aminobutyric acid. NAA is an amino acid generally considered to be a neuronal marker.^{31}P-MRS is used to measure brain membrane phospholipids and high-energy phosphate metabolism.

Molecular Imaging with Radiolabeled Ligands

General Aspects

PET and SPECT are molecular imaging techniques that provide *in vivo* information on biological molecules. They use very small quantities of radiolabeled molecules, known as radiotracers. These emit photons that can be measured in the scanner. Thus the localization and behavior of the radiolabeled molecule can be determined *in vivo*. Any substance that can have a radionuclide attached has the potential to be developed into a PET or SPECT ligand. The usefulness of particular ligand depends on its properties and target. Commonly available ligands target specific receptors, enzyme function, inflammatory markers, brain blood flow, and metabolism. It is also possible to index change in transmitter levels following a drug or cognitive challenge.

Principles

To obtain PET and SPECT images, the volunteer receives a radiolabeled ligand intravenously. Thus both techniques involve exposure to a small amount of radiation. PET uses positron emitting radionuclides such as carbon-11 (^{11}C), oxygen-15 (^{15}O), and fluorine-18 (^{18}F). These have short half-lives – meaning that they decay rapidly, generally within minutes and up to about 110 min in the case of ^{18}F. As a consequence, they need to be used quickly, which usually means they generally have to be made on-site and necessitates expensive equipment and skilled radiochemistry to be available. On the other hand, the common SPECT radionuclides such as technetium-99 (^{99m}Tc) or iodine-123 (^{123}I) generally have much longer half-lives – from hours to days. This means that they can be made off-site and potentially used for a number of syntheses. Another key aspect of the difference in the radionuclides used in the two techniques is that those used in PET tend to be relatively easily incorporated in many biological molecules. In contrast, the radionuclides used in SPECT tend to be rarely found naturally in biological molecules, and therefore incorporating them to make a successful radiotracer can be more difficult. Generally, it is important to know whether the addition of the radionuclide to make the radiolabeled ligand alters the properties of the molecule being labeled, or its behavior *in vivo*. This is often more of an issue in developing SPECT ligands because the radionuclides are rarely found naturally in biological molecules.

Findings

This section reviews some of the main findings from neuroimaging studies in psychiatry. It focuses on schizophrenia, affective disorders, and Alzheimer's disease, as these have been among the most extensively studied psychiatric disorders and provides a number of examples of the use and potential of neuroimaging techniques in psychiatry. It is intended to illustrate how these techniques have been used to advance understanding of psychiatric disorders and their potential for clinical application.

Schizophrenia

Structural Studies

There have been over 100 structural MRI studies in schizophrenia to date (Ellison-Wright et al., 2008). These have unequivocally demonstrated that there are structural alterations seen on average in people with schizophrenia, although it is important to note that structural alterations may not be seen at the individual level (Ellison-Wright et al., 2008). Increased ventricular volume and reduced total gray matter are among the most consistent findings (Ellison-Wright et al., 2008). The ventricular enlargement is likely to reflect loss of volume in other structures rather than an abnormality in the ventricular system per se. Gray matter volume in frontal and temporal lobe regions is consistently reduced, and generally the reductions in these regions are larger than in other regions, suggesting some regional specificity (Ellison-Wright et al., 2008). These changes are also seen in first episode patients at onset of schizophrenia and before treatment, indicating that they are not solely attributable to the effects of antipsychotics or chronicity. Furthermore similar, although generally less marked, structural alterations are also seen in the prodrome to schizophrenia, and seem to progress with the onset of the disorder (Pantelis et al., 2003). Longitudinal studies from the onset of the disorder have also shown further gray matter loss and ventricular enlargement during the course of schizophrenia (Ho et al., 2011). There is considerable controversy over whether these changes reflect a pathophysiological process or are a consequence of other factors such as antipsychotic treatment (Ho et al., 2011). DTI measures indicate that there is reduced integrity of white matter tracts in schizophrenia as well, indicating that alterations are not limited to gray matter alone, and highlighting that schizophrenia involves loss of connectivity between regions.

Functional studies

There have been a large number of functional imaging studies in schizophrenia, beginning with PET and SPECT imaging studies in the 1980s. These have shown that schizophrenia is associated with reduced cerebral blood flow and glucose metabolism, particularly in frontal regions, and to link them to symptoms (Davidson and Heinrichs, 2003). Subsequent fMRI imaging has extended these findings to show altered BOLD responses during a large number of cognitive and emotional tasks. Executive function is one domain of cognitive function that is consistently impaired in patients with schizophrenia. A meta-analysis of over 40 fMRI studies of executive function tasks in schizophrenia has found a remarkably consistent pattern of altered BOLD responses, with reduced activation in frontal cortical regions (including the dorsolateral prefrontal cortex) and thalamus, and reduced activation in the anterior cingulate cortex (Minzenberg et al., 2009). It is thus clear that there is altered brain function in schizophrenia.

Neurochemical imaging studies have provided evidence on the nature of the molecular abnormalities that may underlie the functional brain alterations in schizophrenia, and symptoms. A meta-analysis of over 50 molecular imaging studies of the dopamine system has pin-pointed the nature of the dopaminergic changes in the disorder (Howes et al., 2012). It identified that while there is no marked alteration in D2/3 receptor availability, there is a marked and consistent elevation in the capacity for striatal dopamine synthesis and dopamine release in patients with schizophrenia. The same alterations have also been recently reported in the prodromal phase prior to the onset of psychosis (Howes et al., 2011b), and found to increase with the onset of psychosis (see **Figure 1** and Howes et al., 2011a). MRS imaging has identified alterations in glutamatergic indices in schizophrenia, although, in contrast to the consistent elevation in the presynaptic dopaminergic findings, there is inconsistency (Marsman et al., 2013). The lack of consistency may reflect the stage of illness or effect of antipsychotic treatment – with some evidence suggesting that there are elevations early in the illness that are reduced with treatment.

Affective Disorders

Structural and fMRI studies

Unipolar depression is associated with reduced brain volume, although alterations are in somewhat different regions to those seen in schizophrenia and are less marked. Meta-analysis indicates that reductions are seen in regions involved in emotional processing such as the orbitofrontal cortex, cingulate cortex, hippocampus, and striatum, while there is an increase in pituitary volume (Arnone et al., 2012a). There is also consistent evidence linking the presence of increased white matter hyperintensities – an indicator of possible focal infarcts or other damage – to depression. In addition to structural alterations, there is evidence for functional brain changes in depression. Converging data from fMRI studies increased activity in a number of brain regions involved in the processing of emotion, particularly to negative stimuli in the disorder (Kupfer et al., 2012). The amygdala, ventral striatum, and medial prefrontal cortical regions show increased activation to negative emotional stimuli (Arnone et al., 2012b). Interestingly these alterations have been found to be reversed by treatment with serotonergic antidepressants (Harmer et al., 2009; Kupfer et al., 2012). Furthermore response to serotonergic antidepressants is predicted by the activity in these regions.

Molecular imaging studies

Given the importance of serotonergic drugs in the treatment of depression, it is not surprising that molecular imaging studies have particularly focused on the serotonin system. Serotonin$_{1A}$ receptors are thought to play a key role in maintaining stable serotonin transmission and to be involved in the mechanism of antidepressant treatment. Most of molecular imaging studies of these receptors have found that patients with depression have reduced 5-HT$_{1A}$ receptor density, particularly in the raphe

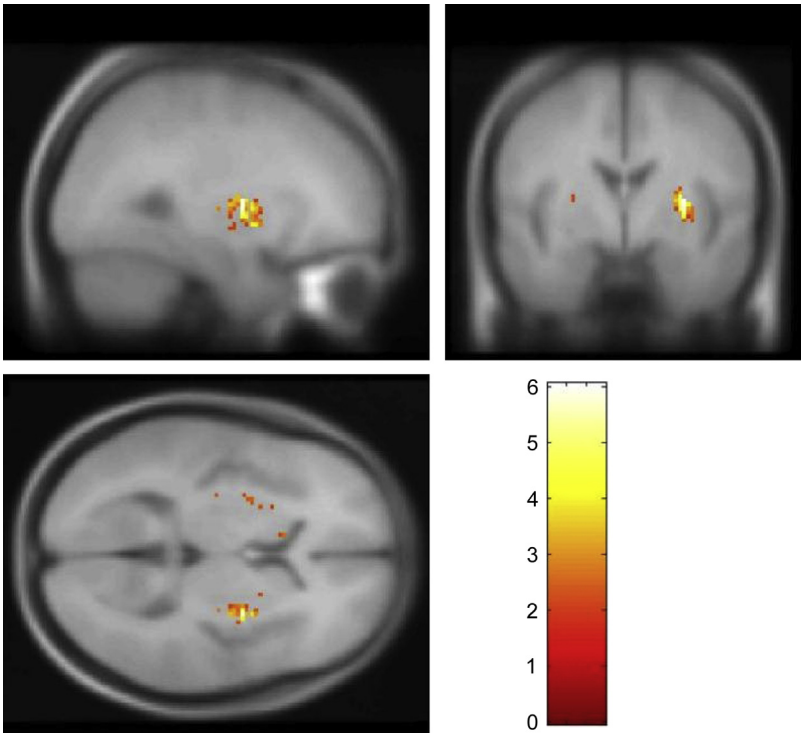

Figure 1 Showing the change in striatal dopamine synthesis capacity with the onset of psychosis – illustrating the value of neuroimaging to investigate pathophysiological changes during the course of psychiatric disorders. Taken with permission from Howes, O., Bose, S., Turkheimer, F., Valli, I., Egerton, A., Stahl, D., Valmaggia, L., Allen, P., Murray, R., Mcguire, P., 2011a. Progressive increase in striatal dopamine synthesis capacity as patients develop psychosis: a PET study. Molecular Psychiatry 16, 885–886.

nucleus. However, increased 5-HT_{1A} receptor density has also been reported. This apparent discrepancy may be due to methodological differences, particularly whether the 5-HT_{1A} binding potential is determined using the 'gold standard' arterial input function, or using a reference region (Shrestha et al., 2012). Recently serotonin$_{1A}$ receptor levels have been found to closely correlate with regional gray matter volumes, particularly those involved in emotional processing, in patients with depression, suggesting that these abnormalities are linked (Kraus et al., 2012).

The 5-HT_{1B} receptor is another presynaptic autoreceptor, which inhibits serotonin and other monoamine neurotransmitter release. A 20% decrease in 5-HT_{1B} binding potential in the ventral striatum and pallidum of patients with depression has been reported, suggesting there may be abnormalities in other autoreceptors in depression as well (Murrough et al., 2011). The serotonin transporter has been another focus of investigation in depression, particularly because of the importance of serotonin transporter inhibitors in the treatment of depression. Although there have been some conflicting results, the majority of studies have found decreased serotonin transporter binding in patients with depression (see recent examples: Selvaraj et al., 2011; Reimold et al., 2011). PET studies have also found evidence for alterations affecting one of the enzymes that break down monoamines. Specifically the density of monoamine oxidase-A (MAO-A) has been found to be elevated in depression (Meyer et al., 2006), suggesting that the metabolism of monoamines is increased in depression. Based on this, Meyer et al. (2006) hypothesized that the increased density of MAO-A in depression is the primary driver of reduced monoaminergic signal transduction in depression. This suggests that dopaminergic and norepinephrinergic systems will be affected as well. There is some evidence of alterations in the dopaminergic system in depression, although changes have not been marked with slight decreases or no change in D_1 and D_2 receptor density, and a decrease in dopamine transporter binding being reported. However, the paucity of radiotracers for the norepinephrine system has limited investigation of this system in depression, though some tracers for imaging norepinephrine transporter and adrenergic receptors are now under development.

Bipolar affective disorder is also associated with structural brain alterations compared to healthy controls but involving different brain regions to unipolar depression (Selvaraj et al., 2012). Gray matter reductions are seen in temporal cortical regions, prefrontal cortex (in particular the inferior frontal gyrus), and the claustrum, although there is heterogeneity between studies, potentially due to differences in the illness stage of the patients included in the various studies. Increased amygdala activation to emotional stimuli is seen in bipolar affective disorder, both in manic and euthymic patients (Strakowski et al., 2012). In contrast, decreased activation is seen in prefrontal cortical regions including the orbitofrontal and ventrolateral prefrontal cortex. Euthymic patients do not show the same degree of reduced prefrontal activation, suggesting that there may be recovery of prefrontal function. Additionally there is emerging evidence that there is reduced functional connectivity between prefrontal regions and the

amygdala in bipolar affective disorder. Overall, this suggests that a failure of frontal regulation of limbic areas such as the amygdala may underlie the emotional dysregulation that characterizes bipolar affective disorder.

Alzheimer's Disease

Structural Studies

Structural MRI studies have consistently shown structural changes in Alzheimer's Dementia, with cerebral atrophy typically starting in temporal regions, particularly medial temporal areas, and spreading to frontal and parietal regions. At diagnosis, whole brain volume is reduced by ~5%, and reductions in the medial temporal cortex are in the range of 10% (Jack et al., 2008; Johnson et al., 2012). Atrophy begins insidiously in temporal cortex, characteristically beginning in the entorhinal cortex, hippocampus, amygdala, and parahippocampal cortex and gradually progressing in these regions, and extending to other regions. Loss of volume in hippocampal regions is typically in the region of 3–5% per year (Johnson et al., 2012). Studies of people with mild cognitive impairment show that hippocampal atrophy is evident even before the diagnosis of Alzheimer's disease (Jack et al., 2008). MRI assessments of medial temporal region are able to distinguish early Alzheimer's disease from controls with good sensitivity and specificity (>80% in both cases) (Frisoni et al., 2010). Medial temporal atrophy on MRI also shows promise as a marker to identify people who are going to develop Alzheimer's disease in the next year or so, from those with similar mild cognitive impairments who do not go on to develop the disorder in this time frame (Frisoni et al., 2010). It is important to note, however, that while the typical pattern of progressive changes begins in the temporal cortex, atrophy may be seen in other regions and linked to atypical presentations. For example, patients presenting with visual disturbance have been found to show posterior cortical atrophy. Consequently, exclusively focusing on temporal regions may miss some cases. Regional MRI volume has been found to be related to postmortem neuronal counts, suggesting that the volumetric change at least, in part, reflects loss of neurons (Gosche et al., 2002). Furthermore, brain atrophy is directly related to the degree of cognitive impairment seen in the disease (Jack et al., 2002). At the same time that there is gray matter loss, there is widening of cortical sulci and ventricular enlargement (Johnson et al., 2012). Local white matter hyperintensities on MRI have also been reported, although these seem more indicative of vascular dementia and are also seen as part of normal aging.

Functional Cerebral Glucose and Activation Studies

Molecular and functional MR imaging studies have shown brain functional changes in Alzheimer's disease, indicating that the structural changes are accompanied by functional alterations. Cerebral glucose metabolism, as measured by [^{18}F]fluorodeoxyglucose (FDG), is reduced in temporoparietal regions, precuneus, posterior cingulate cortex, and frontal cortex in Alzheimer's disease (Minoshima et al., 1997; Foster et al., 2007). Subsequent fMRI studies have shown alterations in similar regions. In particular, fMRI studies consistently show decreased medial temporal cortical activation during memory tasks in Alzheimer's disease. Findings of fMRI are less consistent in people with mild cognitive impairment, with some studies reporting decreased medial temporal activation, and others increased activation in the same area. This may reflect the stage of illness – with very early phases – but could also be accounted for by differences in the tasks used, and/or performance of the subjects.

Molecular Imaging Studies

As amyloid plaques are a key pathophysiological feature of Alzheimer's disease, considerable effort has gone into developing molecular imaging probes for amyloid. The most extensively used is N-methyl-[^{11}C]2-(4-methylaminophenyl)-6-hydroxybenzothiazole, also known as Pittsburgh compound B (PIB). Elevated [^{11}C]PIB binding is seen in the prefrontal cortex, precuneus, and posterior cingulate of Alzheimer's patients in comparison with controls. Studies have also indicated that amyloid deposition is most active during the early phase of the disease (Engler et al., 2006). However, cognitive impairment does not seem to correlate with extent of [^{11}C]PIB binding, suggesting that while this may be an early marker of the disease process, it is the later consequences of this process that underlie the cognitive dysfunction. [^{11}C]PIB PET imaging appears to be better at detecting mild cognitive impairment and early Alzheimer's disease than changes in cerebral glucose measured with [^{18}F]FDG PET (Klunk, 2011). However, metabolic abnormalities are more regionally specific and may be more closely linked to cognitive deficits (Kadir et al., 2012).

The short half-life of carbon-11 limits the clinical application of [^{11}C]PIB and other radiotracers labeled with carbon-11 to centers with an on-site cyclotron. Fluroine-18 has a much longer half-life, approaching 2 h, which means radiotracers labeled with fluorine-18 can be made at a central site and delivered to a number of regional centers. Consequently, fluorinated tracers for amyloid plaques, such as [^{18}F]flutemetamol, [^{18}F]florbetapir, [^{18}F]florbetaben, and other fluorinated equivalents of [^{11}C]PIB have been developed. One of these, [^{18}F]florbetapir (AMYViD™, Eli Lilly), has recently received FDA approval for the imaging of beta amyloid plaques. [^{18}F]florbetapir scans show very good sensitivity (>90%) and specificity (95%) for amyloid (Yang et al., 2012).

It has been hypothesized that neuroinflammation plays a role in the neuronal loss seen in Alzheimer's disease. Microglia in the brain are typically activated as part of the inflammatory response, and microglial activation is seen in postmortem studies in Alzheimer's disease. The PET tracer [^{11}C]PK11195 indexes the activation of microglia in the brain and has been used to study the neuroinflammatory response in Alzheimer's disease. [^{11}C]PK11195 binding has been found to be elevated in the temporoparietal, cingulate, and entorhinal cortex in Alzheimer's disease, and to correlate with impairments in cognitive performance (Edison et al., 2008). [^{11}C]DED is a PET tracer that measures astrocytic activation that has also been shown to be increased in Alzheimer's disease and in mild cognitive impairment. In fact, [^{11}C]DED binding may be higher in mild cognitive impairment than in Alzheimer's disease, suggesting that astrocytic activation could be an early pathophysiological change in Alzheimer's (Carter et al., 2012).

Molecular imaging may be a useful surrogate marker to monitor treatments for Alzheimer's disease. [^{11}C]PIB PET studies have also been used to evaluate the effect of anti-amyloid treatments, finding that treatment was associated with a reduction in binding (e.g.,: Ostrowitzki et al., 2012). Similarly imaging microglia may help assess the effectiveness of drugs that are targeted toward reducing inflammation in the brain, such as nonsteroidal anti-inflammatory drugs.

Recently it has been recognized that developing and applying imaging to Alzheimer's disease requires developing standardized approaches for the collection of MRI, PET, and other measures, and this has resulted in a number of initiatives to do this and to validate the use of imaging and other biomarkers for the early detection and monitoring of Alzheimer's disease. The Alzheimer's disease neuroimaging initiative (ADNI) is an example of one such initiative that spans research centers around the world. The ADNI has found that combining MRI and PET, with cerebrospinal fluid measures and clinical tests significantly increases the ability to correctly predict conversion to Alzheimer's disease compared with clinical testing alone (Shaffer et al., 2013). This suggests that different imaging modalities provide complementary information, and that a multimodal approach may be most desirable both for research and clinical applications.

Summary

MR and molecular imaging provide *in vivo* measures of a number of aspects of brain structure and function, and enabled them to be linked to clinical measures and outcomes, including changes over time. This is a clear advantage over other approaches such as *postmortem* studies. As such, MR and molecular imaging has already made a major contribution to understanding the biological basis of psychiatric disorders, and it is likely that they will continue to do so. They have contributed to establishing these disorders as brain disorders with distinct neurobiological alterations. In contrast, the direct contribution of these techniques to routine clinical practice is less developed. This is perhaps not surprising given that the technologies are relatively young, the disorders are complex, and there are regulatory and infrastructure barriers that inhibit translation to clinical practice. Initiatives such as ADNI are addressing some of these, and the licensing of [^{18}F] florbetapir indicates that regulatory barriers can be overcome. MR and molecular imaging techniques are becoming increasingly sophisticated and sensitive, and further technological developments, such as combined MR and PET scanners, are likely to contribute to a further refined understanding of the pathophysiology of psychiatric disorders.

See also: Cognitive Neuropsychology, Methods of; Cortical Areas Engaged in Movement: Neuroimaging Methods; Functional Brain Imaging of Language processes; Functional Magnetic Resonance Imagery, Analysis of; Imaging Methods in Clinical Psychology; Imaging Methods in Psychotherapy; Language Processing, Functional Magnetic Resonance Imaging of; Neuropsychological Testing; Schizophrenia: Metabolic Consequences.

Bibliography

Arnone, D., McIntosh, A.M., Ebmeier, K.P., Munafo, M.R., Anderson, I.M., 2012a. Magnetic resonance imaging studies in unipolar depression: systematic review and meta-regression analyses. European Neuropsychopharmacology 22, 1–16.

Arnone, D., McKie, S., Elliott, R., Thomas, E.J., Downey, D., Juhasz, G., Williams, S.R., Deakin, J.F., Anderson, I.M., 2012b. Increased amygdala responses to sad but not fearful faces in major depression: relation to mood state and pharmacological treatment. American Journal of Psychiatry 169, 841–850.

Buxton, R., 2002. Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques. Cambridge University Press, Cambridge, UK.

Carter, S.F., Scholl, M., Almkvist, O., Wall, A., Engler, H., Langstrom, B., Nordberg, A., 2012. Evidence for astrocytosis in prodromal Alzheimer disease provided by 11C-deuterium-L-deprenyl: a multitracer PET paradigm combining 11C-Pittsburgh compound B and 18F-FDG. Journal of Nuclear Medicine 53, 37–46.

Davidson, L.L., Heinrichs, R.W., 2003. Quantification of frontal and temporal lobe brain-imaging findings in schizophrenia: a meta-analysis. Psychiatry Research 122, 69–87.

Edison, P., Archer, H.A., Gerhard, A., Hinz, R., Pavese, N., Turkheimer, F.E., Hammers, A., Tai, Y.F., Fox, N., Kennedy, A., Rossor, M., Brooks, D.J., 2008. Microglia, amyloid, and cognition in Alzheimer's disease: an [11C](R)PK11195-PET and [11C]PIB-PET study. Neurobiology of Disease 32, 412–419.

Ellison-Wright, I., Glahn, D.C., Laird, A.R., Thelen, S.M., Bullmore, E., 2008. The anatomy of first-episode and chronic schizophrenia: an anatomical likelihood estimation meta-analysis. American Journal of Psychiatry 165, 1015–1023.

Engler, H., Forsberg, A., Almkvist, O., Blomquist, G., Larsson, E., Savitcheva, I., Wall, A., Ringheim, A., Langstrom, B., Nordberg, A., 2006. Two-year follow-up of amyloid deposition in patients with Alzheimer's disease. Brain 129, 2856–2866.

Foster, N.L., Heidebrink, J.L., Clark, C.M., Jagust, W.J., Arnold, S.E., Barbas, N.R., Decarli, C.S., Turner, R.S., Koeppe, R.A., Higdon, R., Minoshima, S., 2007. FDG-PET improves accuracy in distinguishing frontotemporal dementia and Alzheimer's disease. Brain 130, 2616–2635.

Frisoni, G.B., Fox, N.C., Jack Jr., C.R., Scheltens, P., Thompson, P.M., 2010. The clinical use of structural MRI in Alzheimer disease. Nature Reviews Neurology 6, 67–77.

Gosche, K.M., Mortimer, J.A., Smith, C.D., Markesbery, W.R., Snowdon, D.A., 2002. Hippocampal volume as an index of Alzheimer neuropathology: findings from the Nun Study. Neurology 58, 1476–1482.

Grunder, G., 2012. Molecular Imaging in the Clinical Neurosciences. Humana Press-Springer, New York.

Harmer, C.J., Goodwin, G.M., Cowen, P.J., 2009. Why do antidepressants take so long to work? A cognitive neuropsychological model of antidepressant drug action. British Journal of Psychiatry 195, 102–108.

Ho, B.C., Andreasen, N.C., Ziebell, S., Pierson, R., Magnotta, V., 2011. Long-term antipsychotic treatment and brain volumes: a longitudinal study of first-episode schizophrenia. Archives of General Psychiatry 68, 128–137.

Howes, O., Bose, S., Turkheimer, F., Valli, I., Egerton, A., Stahl, D., Valmaggia, L., Allen, P., Murray, R., Mcguire, P., 2011a. Progressive increase in striatal dopamine synthesis capacity as patients develop psychosis: a PET study. Molecular Psychiatry 16, 885–886.

Howes, O.D., Bose, S., Valli, I., Turkheimer, F., Egerton, A., Valmaggia, L., Murray, R.M., Mcguire, P., 2011b. Dopamine synthesis capacity before onset of psychosis: a prospective [18]-DOPA PET imaging study. American Journal of Psychiatry 168, 1311–1317.

Howes, O.D., Kambeitz, J., Kim, E., Stahl, D., Slifstein, M., Abi-Dargham, A., Kapur, S., 2012. The nature of dopamine dysfunction in schizophrenia and what this means for treatment. Archives of General Psychiatry 69, 776–786.

Jack Jr., C.R., Dickson, D.W., Parisi, J.E., Xu, Y.C., Cha, R.H., O'brien, P.C., Edland, S.D., Smith, G.E., Boeve, B.F., Tangalos, E.G., Kokmen, E., Petersen, R.C., 2002. Antemortem MRI findings correlate with hippocampal neuropathology in typical aging and dementia. Neurology 58, 750–757.

Jack Jr., C.R., Weigand, S.D., Shiung, M.M., Przybelski, S.A., O'brien, P.C., Gunter, J.L., Knopman, D.S., Boeve, B.F., Smith, G.E., Petersen, R.C., 2008. Atrophy rates accelerate in amnestic mild cognitive impairment. Neurology 70, 1740–1752.

Johnson, K.A., Fox, N.C., Sperling, R.A., Klunk, W.E., 2012. Brain imaging in Alzheimer disease. Cold Spring Harbor Perspectives in Medicine 2, a006213.

Kadir, A., Almkvist, O., Forsberg, A., Wall, A., Engler, H., Langstrom, B., Nordberg, A., 2012. Dynamic changes in PET amyloid and FDG imaging at different stages of Alzheimer's disease. Neurobiology of Aging 33 (198), e1–e14.

Klunk, W.E., 2011. Amyloid imaging as a biomarker for cerebral beta-amyloidosis and risk prediction for Alzheimer dementia. Neurobiology of Aging 32 (Suppl. 1), S20–S36.

Kraus, C., Hahn, A., Savli, M., Kranz, G.S., Baldinger, P., Höflich, A., Spindelegger, C., Ungersboeck, J., Haeusler, D., Mittehauser, M., Windischberger, C., Wadsak, W., Kasper, S., Lanzenberger, R., 2012. Serotonin-1A receptor binding is positively associated with gray matter volume – a multimodal neuroimaging study combining PET and structural MRI. Neuroimage 63 (3), 1091–1098.

Kupfer, D.J., Frank, E., Phillips, M.L., 2012. Major depressive disorder: new clinical, neurobiological, and treatment perspectives. Lancet 379, 1045–1055.

Marsman, A., Van Den Heuvel, M.P., Klomp, D.W., Kahn, R.S., Luijten, P.R., Hulshoff Pol, H.E., 2013. Glutamate in schizophrenia: a focused review and meta-analysis of (1)H-MRS studies. Schizophrenia Bulletin 39, 120–129.

Meyer, J.H., Ginovart, N., Boovariwala, A., Sagrati, S., Hussey, D., Garcia, A., Young, T., Praschak-Rieder, N., Wilson, A.A., Houle, S., 2006. Elevated monoamine oxidase a levels in the brain: an explanation for the monoamine imbalance of major depression. Archives of General Psychiatry 63, 1209–1216.

Minoshima, S., Giordani, B., Berent, S., Frey, K.A., Foster, N.L., Kuhl, D.E., 1997. Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease. Annals of Neurology 42, 85–94.

Minzenberg, M.J., Laird, A.R., Thelen, S., Carter, C.S., Glahn, D.C., 2009. Meta-analysis of 41 functional neuroimaging studies of executive function in schizophrenia. Archives of General Psychiatry 66, 811–822.

Murrough, J.W., Henry, S., Hu, J., Gallezot, J.D., Planeta-Wilson, B., Neumaier, J.F., Neumeister, A., 2011. Reduced ventral striatal/ventral pallidal serotonin1B receptor binding potential in major depressive disorder. Psychopharmacology (Berl) 213, 547–553.

Ostrowitzki, S., Deptula, D., Thurfjell, L., Barkhof, F., Bohrmann, B., Brooks, D.J., Klunk, W.E., Ashford, E., Yoo, K., Xu, Z.X., Loetscher, H., Santarelli, L., 2012. Mechanism of amyloid removal in patients with Alzheimer disease treated with gantenerumab. Archives of Neurology 69, 198–207.

Pantelis, C., Velakoulis, D., Mcgorry, P.D., Wood, S.J., Suckling, J., Phillips, L.J., Yung, A.R., Bullmore, E.T., Brewer, W., Soulsby, B., Desmond, P., Mcguire, P.K., 2003. Neuroanatomical abnormalities before and after onset of psychosis: a cross-sectional and longitudinal MRI comparison. Lancet 361, 281–288.

Reimold, M., Knobel, A., Rapp, M.A., Batra, A., Wiedemann, K., Strohle, A., Zimmer, A., Schonknecht, P., Smolka, M.N., Weinberger, D.R., Goldman, D., Machulla, H.J., Bares, R., Heinz, A., 2011. Central serotonin transporter levels are associated with stress hormone response and anxiety. Psychopharmacology (Berl) 213, 563–572.

Selvaraj, S., Murthy, N.V., Bhagwagar, Z., Bose, S.K., Hinz, R., Grasby, P.M., Cowen, P.J., 2011. Diminished brain 5-HT transporter binding in major depression: a positron emission tomography study with [11C]DASB. Psychopharmacology (Berl) 213, 555–562.

Selvaraj, S., Arnone, D., Job, D., Stanfield, A., Farrow, T.F., Nugent, A.C., Scherk, H., Gruber, O., Chen, X., Sachdev, P.S., Dickstein, D.P., Malhi, G.S., Ha, T.H., Ha, K., Phillips, M.L., Mcintosh, A.M., 2012. Grey matter differences in bipolar disorder: a meta-analysis of voxel-based morphometry studies. Bipolar Disorders 14, 135–145.

Shaffer, J.L., Petrella, J.R., Sheldon, F.C., Choudhury, K.R., Calhoun, V.D., Coleman, R.E., Doraiswamy, P.M., Alzheimer's Disease Neuroimaging Initiative, 2013. Predicting cognitive decline in subjects at risk for Alzheimer disease by using combined cerebrospinal fluid, MR imaging, and PET biomarkers. Radiology 266, 583–591.

Shrestha, S., Hirvonen, J., Hines, C.S., Henter, I.D., Svenningsson, P., Pike, V.W., Innis, R.B., 2012. Serotonin-1A receptors in major depression quantified using PET: controversies, confounds, and recommendations. Neuroimage 59, 3243–3251.

Strakowski, S.M., Adler, C.M., Almeida, J., Altshuler, L.L., Blumberg, H.P., Chang, K.D., Delbello, M.P., Frangou, S., Mcintosh, A., Phillips, M.L., Sussman, J.E., Townsend, J.D., 2012. The functional neuroanatomy of bipolar disorder: a consensus model. Bipolar Disorders 14, 313–325.

Yang, L., Rieves, D., Ganley, C., 2012. Brain amyloid imaging–FDA approval of florbetapir F18 injection. New England Journal of Medicine 367, 885–887.

Majoritarianism and Majority Rule

Douglas W Rae, Yale University, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Majoritarianism and majority rule prove to be an aspiration for regimes in upcoming decades. In comparison with minority rule, majority rule proves to favor political equality and is efficient at coordinating voter preferences to outcomes. In addition, majority rule can trump formal rationality in terms of voting preferences and become skewed when incorporated with proportional representation and other complex institutions. In context to the shifting economic power occurring in the world, many states throughout the world have been slow to incorporate majority rule in their respective states. Through case studies, one sees the struggles and limits of introducing majoritarianism in political regimes.

This article will use 'majority rule' to denote the voting rule requiring $[n + 1]/2$ to carry a decision, and 'majoritarianism' to denote the belief that this rule, and others closely resembling it, are desirable. These ideas may be analyzed by posing four questions:

1. Is majority rule efficiently fair?
2. Does majority rule efficiently match preferences to outcomes?
3. Does majority rule comply with basic requirements of formal rationality?
4. How well do the more complex procedures and electoral systems used in governments match up with majoritarian requirements?

Fairness

In the long history of human governance, majoritarianism has for the most part been trumped but the rule of minorities. Often, these minorities have been elites based on military prowess, land ownership, heredity, or priestly knowledge – whether the priests be religious or, for example, socialist. In capitalist and market societies, elites with major ownership stakes and elites constituted by managerial expertise have in many respects held authority denied to popular majorities (Lindblom, 1977). In other cases, such as the American Constitution of 1789, *generic* minorities threatened by majorities seeking new uses of governmental power are granted the authority to block action. The US system builds in no bias in favor of a specific minority, and it is indeed prejudicial toward all minorities seeking new uses of government. The favorable bias is specific to defensive minorities seeking to preserve existing policy. The overall historical pattern, we may safely conclude, is that most societies have not been governed in majoritarian fashion.

Majoritarianism arises against this background, in major part, as a demand for political equality. Historical demands made for expanded voting rights, across barriers erected on class, religion, gender, and race have played a central role in the democratization of Western nation-states from the beginning. Majority rule is implicit in most of this history. Foremost among the meanings of political equality is the doctrine 'One man (one person), one vote.' This powerful idea links majority rule to the central appeal of popular democracy without specifying exactly the nature of the equality in question. Kenneth May has conveniently formalized this idea with the criterion of *anonymity*: if a decision process yields the result that X is preferred to Y in a given case, it should again do so for every other case in which the preferences of any two blocks of voters of equal size are transposed (May, 1952). Put otherwise, you should always be able to count the votes (and announce a decision) without seeing whose votes they are. Another feature of equality or equity entailed by majority rule concerns the content of the alternates under consideration. Majority rule is formally neutral in its treatment of alternatives. Whether X or Y is the status quo, for example, makes no difference to the counting of votes. Whether X or Y is prolabor, respectful of racial minorities, consistent with the tenets of a prevailing ideology, or consistent with the interests of economic elites – none of these or similar considerations about the alternatives in question will matter under majority rule. This additional requirement of 'neutrality' combines with anonymity in 'May's Theorem' to define majority rule as uniquely best among decision rules, which respond positively and nonrandomly to voter preferences (May, 1952).

The foregoing considerations constitute a strong case for the egalitarian fairness of majoritarianism on an abstractly a *priori* basis. They do not, however, say anything about the fairness of majority rule in specific historical cases where known groups have predictable or announced preferences. Indeed, the neutrality of majority rule toward alternatives implies an embedded incapacity to discriminate in favor of fair, egalitarian, or just policies as against unfair, antiegalitarian, or unjust ones. Majority rule provides no defense against majorities embracing evil. In racially and religiously divided societies, where strong conflicts predictably divide large groups from small ones, this is a crippling disability, which occasions sharp resistance to majority rule. One only needs to think of recent religious and ideological conflict in Iraq, Iran, Egypt, and Turkey to see this point: Where majority and minority are frozen, majority rule becomes a weapon, not a compromise. The time-honored remedies in cases of this sort are *partition* (to produce more homogeneous, less conflictual, and smaller polities), the protection of *minority rights* against specified forms of majority action within each polity, and, of course,

emigration opportunities for members of an afflicted minority. Where some populations enjoy low-cost exit and others do not, even these solutions may seem unfair. The development of civic cultures, which surmount and eventually attenuate conflict is probably superior to any of these more mechanical solutions, but requires historical time and good fortune. It is nevertheless worth noting that even some very hierarchical and undemocratic cultures – India, Japan, quite possibly South Africa – have made great strides in this direction (see, in particular, Shapiro, 2001).

Preference Efficiency

Quite apart from considerations of fairness, a case can be made for majority rule based on its efficiency in matching preferences to policy outcomes. Indeed, it can be shown analytically that all other procedures for making collective decisions are inferior to majority rule in their ability to maximize direct fit between preferences and outcomes (see Rae, 1969; for a critical evaluation, see Mueller, 1989). This applies to other voting rules, rules giving priority to specific types of policy, and all manner of random or arbitrary procedures. Preference-matching efficiency appears most plausible for issues, where the determination of basic policy issues and entitlements is concerned. If decisions entail starkly different stakes for different voters ('Shall we destroy the reader's home?'), nominal matching of preferences to outcomes is obviously inadequate as a criterion. Where decisions offer gains from trade among the rights, so established market and market-like arrangements appear superior on efficiency criteria. Indeed, as Ronald Coase has shown trade can turn any well-defined set of initial rights into a Paretoefficient set of rights on a voluntary basis (Coase, 1960). Insofar as such a process of exchange can be embedded in a voting rule, which would require unanimity for each change of policy (Wicksell, 1896). These market-oriented schemes of course give great leverage to groups and individuals who are privileged by the initial distribution of rights and assets, and are explicitly antimajoritarian for that reason. It must of course be accepted that neither these market-oriented schemes nor majoritarianism can purport to efficient treatment of preferences, which are manipulated, distorted, or subject to intimidation.

Majority Rule and Formal Rationality

It has been well known since the late-eighteenth century that majority rule posses some quirky features, especially when multiple alternatives are considered in one process. Most famous among these is the 'cycle of voting' identified by the Marquis de Condorcet (1785). In the simplest case, we imagine three voters and three alternatives (**Table 1**). In sequential application of majority rule, A defeats B on the strength of voters 1 and 2, B beats C via 1 and 3, yet C tops A via 2 and 3. This violates the axiom of formal rationality known as transitivity, and opens the possibility that control over agenda sequences is tantamount to having control over outcomes. This transitivity requirement, together with others, which tap additional aspects of formal rationality, defines the justly famed Arrow Theorem sowing that all decision procedures, or social welfare functions, may violate the dictates of formal rationality in one way or another (Arrow, 1951).

Table 1 Hypothetical responses of three voters to three alternatives A, B, and C

Voters	*1*	*2*	*3*
Best	A	C	B
	B	A	C
Worst	C	B	A

Majoritarianism in Complex Institutions

Far more important than these formal properties for most purposes is the task of realizing majoritarian aims in complex institutions. The American 'electoral college,' for example, contains the very real potential for reversing national majorities having done so as recently as the fall of 2000. Most single-member electoral systems place an immense premium on spatial distribution of the vote, often giving majority status to parties with well less than majority status in the electorate. Worse yet, where regional parties dominate, as in India, they may fragment outcomes so that nothing close to national majorities are formed. Systems of proportional representation, particularly those based on large districts and exact proportionality, pass the problem of majority formation on to parliaments, and create the opportunity for small parties to exert influence out of all proportions to their size, as with the case of Israel in the 1990s. Most, if not all, legislative procedures, entailing many critical veto points conspicuously violate majority rule. It follows that majoritarianism is less a blueprint than an aspiration.

Conclusion

We live in an era when democratic aspirations have mostly swept the field of direct rivals, particularly among nations that play leading roles in the world economy. Yet in practice, movement toward democratization has been slow, halting, and uncertain in many places. Most of Africa has made little progress, and many key systems from the former Soviet bloc have made minimal progress. China currently displays very less movement toward popular majority politics. Majoritarianism may usefully be seen as one of the key aspirations that should help to shape change and reform in coming decades.

See also: Democratic Theory; Democratic Transitions; Elections: Apportionment; Electoral Systems; Political Representation; Politics and Participation; Voting, Explanations of: Social Class.

Bibliography

Arrow, K.J., 1951. Social Choice and Individual Values. Wiley, New York.
Coase, R., 1960. The problem of social cost. Journal of Law and Economics 3, 1–44.
Condorcet, M de, 1785. Essai sur l'Application de l'Analyse a la Probabilite des jDecisions Rendues a la Pluraliste des Voix. De l'Improm ene royale, Paris.

May, K., 1952. A set of independent, necessary and sufficient conditions for simple majority decision. Econometrica 20, 680–684.

Mueller, D.C. Public Choice II. Cambridge University Press, Cambridge, UK.

Rae, D.W., 1969. Decision-rules and individual values in constitutional choice. American Political Science Review 63, 40–56.

Rae, D.W., 1967. The Political Consequences of Electoral Laws. Yale University Press, New Haven.

Shapiro, I., 2001. The state of democratic theory. In: Katznelson, I., Milner, H.H. (Eds.), Political Science: The State of a Discipline. American Political Science Association, Washington DC.

Majorization and Stochastic Orders

Harry Joe, University of British Columbia, Vancouver, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The simplest stochastic order is an ordering of 'largeness' on random variables or random vectors; more generally, stochastic orders are orderings that apply to compare random variables or random vectors, or probability distributions or measures. Majorization orderings are orderings of divergence from uniformity and provide a means of proving many inequalities.

Several equivalent forms are given for the stochastic largeness order for random variables, and for measures in a partially ordered space, from which generalizations come from extending one or more of the equivalent forms. Similarly, several equivalent forms are given for vector majorization and generalized majorization for functions on a measure space. Dependence orderings that result from stochastic and majorization orderings are obtained and contrasted. Examples where the orderings overlap, including the Lorenz ordering, are mentioned.

Applications to illustrate the orderings include measures of inequality and diversity, measures of dependence, preference orderings in random utility models for choice, and comparison of risks for portfolios.

Introduction

In a specific context, stochastic orders are orderings of 'largeness' on random variables or vectors: that is, a random variable X is stochastically smaller than another random variable Y or a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ is stochastically smaller than another random vector $\mathbf{Y} = (Y_1, \ldots, Y_d)$ if X tends to take smaller values than Y or the component X_i $(i = 1, \ldots, d)$ of $\mathbf{X}$ tends to take smaller values than the corresponding component Y_i of $\mathbf{Y}$. In a broad sense, stochastic orders are orderings that apply to the comparison of random variables or vectors, or probability distributions or measures, and there have been many such stochastic orders. Here, a few stochastic orders that may be of greater interest in the social and behavioral sciences are summarized. These include orderings of dispersion and dependence.

Majorization orderings are orderings of divergence of a vector or function from uniformity. They are useful as a means of proving many inequalities. The vector majorization ordering applies to vectors of the same length that have the same sum of components. It has been generalized to orderings of functions on measure spaces, and in this context, there are overlaps with stochastic orders.

Stochastic and majorization orderings are defined in Sections Stochastic orders and Majorization Orderings, starting with the simplest forms. Comparisons and contrasts of stochastic and majorization orderings are given in Section Contrast of stochastic and majorization orderings, and illustrative examples of these orderings are given in Section Applications and illustrative examples. The few references that are given provide a guide to research in majorization and stochastic orderings; they include historical perspectives on the orderings – some often cited early references are Hardy-Littlewood-Pólya, Schur, Lorenz, and Dalton for majorization and inequalities, and Lehmann for the simplest stochastic order.

For terminology and notation, increasing (decreasing) is used in place of nondecreasing (nonincreasing), $(z)_+$ denotes $\max\{0,z\}$ over reals, $\overset{d}{=}$ denotes random quantities that are equal in distribution, $F^{-1}(v) = \inf\{x : F(x) \geq v\}$ is the functional inverse of a univariate cumulative distribution function F, and $X \sim F$ denotes a random variable X having the distribution F.

Stochastic Orders

A stochastic order in general means a partial ordering or preordering (the latter need not be reflexive) that compares random variables or vectors, or probability distributions or measures. A stochastic order in a specific sense means the simplest ordering of largeness of random variables; for clarity of terminology, this is referred to as the stochastic largeness order (it is called the usual stochastic order in Shaked and Shanthikumar (2007)). Many stochastic orders generalize this stochastic largeness order in some way. Some main references for this section are Stoyan (1983), Mosler and Scarsini (1993a, 1993b), and Shaked and Shanthikumar (2007).

The definitions of the stochastic largeness orders are given next.

Definition 1 (Stochastic largeness order for random variables.) Let X, Y be random variables with respective distribution functions F, G. X is stochastically smaller than Y, written $X \prec^{st} Y$ or $F \prec^{st} G$, if one of the following equivalent conditions hold.

(1) $\Pr(X > u) \leq \Pr(Y > u)$ for all $-\infty < u < \infty$, or $\Pr(X \leq u) \geq \Pr(Y \leq u)$ for all $-\infty < u < \infty$
(2) $\mathrm{E}[\psi(X)] \leq \mathrm{E}[\psi(Y)]$ for all increasing functions ψ for which the expectations exist
(3) There exist random variables $\hat{X}, \hat{Y}$ on the same probability space such that $\hat{X} \overset{d}{=} X$, $\hat{Y} \overset{d}{=} Y$, and $\Pr(\hat{X} \leq \hat{Y}) = 1$
(4) There exists a monotone kernel $K : \Re \times \Re \to [0, 1]$ (that is, $K(t,\cdot)$ is a distribution function for all $t \in \Re$ and $K(t,u) = 0$ for $u \leq t$) such that $G(u) = \int_{-\infty}^{\infty} K(t,u)dF(t)$
(5) $F^{-1}(v) \leq G^{-1}(v)$ where F^{-1}, G^{-1} are the inverse cumulative distribution functions

Definition 2 (Stochastic largeness order in a partially ordered space.) Let P_1, P_2 be probability measures on a partially

ordered Polish space $(\mathscr{X}, \mathscr{A})$ (usually Euclidean space) with measure ν (usually Lebesgue or counting measure). Let $\prec$ be the partial order on χ; for d-dimensional Euclidean space, the usual partial order is given by $x \prec y$ if $x_i \leq y_i$, $i = 1, \ldots, d$. P_1 is stochastically smaller than P_2, written $P_1 \prec^{st} P_2$, if one of the following equivalent conditions hold.

(1) $P_1(A) \leq P_2(A)$ for all upper sets A (i.e., sets for which $x \in A$ and $x \prec y$ imply $y \in A$)
(2) $\int \psi dP_1 \leq \int \psi dP_2$ for all increasing functions ψ for which the expectations exist
(3) There exist random quantities X_1, X_2 with values in $(\mathscr{X}, \mathscr{A})$ such that $X_1 \sim P_1$, $X_2 \sim P_2$, and $X_1 \prec X_2$ almost surely
(4) There exists a Markov kernel $M: \mathscr{X} \times \mathscr{A} \to [0, 1]$ such that $M(t, \cdot)$ is a measure with support on $\{\mathbf{z} : \mathbf{t} \prec \mathbf{z}\}$ and $P_2(A) = \int M(\mathbf{t}, A) P_1(d\mathbf{t})$ for all $A \in \mathscr{A}$

Many general stochastic orders take the form of (2) with other classes of functions ψ; for example, convex functions, increasing convex functions, and increasing concave functions. Generalizations of the form of (2) are also known as dilation orderings. The class of increasing concave functions leads to second-order dominance. The class of convex functions leads to a dispersion ordering of random variables with the same mean (see Section Contrast of stochastic and majorization orderings), one of many dispersion orderings that have been proposed.

Generalizations for random variables include the hazard rate order, reverse hazard rate order, and likelihood ratio order. These are useful for proving results that are intuitively plausible with a stochastic largeness order; sometimes, the result will not hold or may be much more difficult to prove with $\prec^{st}$.

Definition 3 (Some additional univariate stochastic orders.) Let X, Y be two random variables with respective cumulative distribution functions F, G and densities f, g.

(1) G is larger than F in the likelihood ratio order, denoted by $F \prec^{LR} G$ or $X \prec^{LR} Y$, if $f(x)g(y) \geq f(y)g(x)$ for all $-\infty < x < y < \infty$
(2) G is larger than F in the hazard rate order, denoted by $F \prec^{HR} G$ or $X \prec^{HR} Y$, if $f(x)/[1 - F(x)] \geq g(x)/[1 - G(x)]$ for all $-\infty < x < \infty$
(3) G is larger than F in the reverse hazard rate order, denoted by $F \prec^{RHR} G$ or $X \prec^{RHR} Y$, if $f(x)/F(x) \leq g(x)/G(x)$ for all $-\infty < x < \infty$

It is not difficult to check that $F \prec^{LR} G$ implies both $F \prec^{HR} G$, $F \prec^{RHR} G$, and either $F \prec^{HR} G$ or $F \prec^{RHR} G$ implies $F \prec^{st} G$. Intuitively, these concepts provide different ways of comparing the event of X smaller and Y larger versus the event of Y smaller and X larger. There are multivariate versions of these orderings for random vectors. However, there are also other similar orderings that compare the amount of dependence in multivariate distributions with the set of univariate margins held fixed; these are known as dependence orderings. A basic dependence ordering is given next.

Definition 4 (Concordance ordering.) Let $\mathscr{F} = \mathscr{F}(F_1, \ldots, F_d)$ be the class of multivariate distributions with given univariate margins $F_1, \ldots, F_d$, where $d \geq 2$. Let $F, G \in \mathscr{F}$ with respective survival functions $\overline{F}, \overline{G}$ (defined below). Suppose $\mathbf{X} \sim F$ and $\mathbf{Y} \sim G$. Then G is more concordant than F, written $F \prec^c G$ or $\mathbf{X} \prec^c \mathbf{Y}$, if $\forall \ \mathbf{t} \in \Re^d$,

$$\begin{aligned} F(\mathbf{t}) &= \Pr(X_1 \leq t_1, \ldots, X_d \leq t_d) \\ &\leq \Pr(Y_1 \leq t_1, \ldots, Y_d \leq t_d) = G(\mathbf{t}) \\ \overline{F}(\mathbf{t}) &= \Pr(X_1 \geq t_1, \ldots, X_d \geq t_d) \\ &\leq \Pr(Y_1 \geq t_1, \ldots, Y_d \geq t_d) = \overline{G}(\mathbf{t}) \end{aligned}$$

Note that this concordance ordering compares specific upper sets and lower sets rather than all upper sets as in Definition 2. Hence a distribution is larger in the concordance ordering if the random variables are more likely to simultaneously take large or small values. Other dependence orderings are given in Chapter 9 of Shaked and Shanthikumar (2007) and Joe (1997).

Majorization Orderings

Vector majorization is defined, followed by a general majorization for functions on a measure space. Some main references for this section are Marshall et al. (2011), Arnold (1987), and Joe (1993).

Definition 5 (Vector majorization.) Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ be vectors in $\Re^n$ such that $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$. $\mathbf{x}$ is majorized by $\mathbf{y}$ or $\mathbf{y}$ majorizes $\mathbf{x}$, written $\mathbf{x} \prec^m \mathbf{y}$, if one of the following equivalent forms holds.

(1) $\sum_{i=1}^k x_{n-i+1:n} \leq \sum_{i=1}^k y_{n-i+1:n}$, $k = 1, \ldots, n-1$, or $\sum_{i=1}^k x_{i:n} \geq \sum_{i=1}^k y_{i:n}$, $k = 1, \ldots, n-1$, where $(x_{i:n})$, $(y_{i:n})$ are the elements of $\mathbf{x}$, $\mathbf{y}$, respectively arranged in increasing order
(2) $\sum_1^n \psi(x_i) \leq \sum_1^n \psi(y_i)$ for all convex, continuous real-valued functions ψ
(3) $\sum_1^n (x_i - t)_+ \leq \sum_1^n (y_i - t)_+$ for all reals t
(4) There is a doubly stochastic $n \times n$ matrix P (nonnegative matrix such that row and column sums are 1) such that $\mathbf{y} = \mathbf{x}P$
(5) $\mathbf{x}$ is in the convex hull of $\{(y_{i1}, \ldots, y_{in}) : (i_1, \ldots, i_n)$ is a permutation of $(1, \ldots, n)\}$

Majorization is also called the Schur order. There are weak majorization orders that do not require the vector sums to be the same. For example, in (2) of Definition 5, for weak submajorization and weak supermajorization, the class of functions ψ are increasing convex functions or decreasing convex functions, respectively. In economics (e.g., Foster and Shorrocks, 1988), in the context of poverty orderings, majorization is called Lorenz dominance and the second of the weak majorization orderings is called general Lorenz dominance.

Generalizations of Definition 5 have come from generalizations of any one of the equivalent properties. The next definition covers many but not all generalizations of vector majorization. Marshall et al. (2011) unified inequalities through majorization, and although generalized majorization leads to inequalities, inequalities have not always been the motivation for extensions.

Let $(\mathscr{X}, \mathscr{A}, \nu)$ be a measure space. For most applications, $\mathscr{X}$ will be a subset of a Euclidean space, and ν will be Lebesgue measure or counting measure. For a nonnegative integrable function h on $(\mathscr{X}, \mathscr{A}, \nu)$, let $m_h(t) = \nu(\{x : h(x) > t\})$, $t \geq 0$, and let $h^*(u) = m_h^{-1}(u) = \sup\{t : m_h(t) > u\}$, $0 \leq u \leq \nu(\mathscr{X})$; h^* is the (left–continuous) decreasing rearrangement of h.

Definition 6 Let a and b be nonnegative integrable functions on $(\mathscr{X}, \mathscr{A}, \nu)$ such that $\int a d\nu = \int b d\nu$. Then a is majorized by b, written $a \prec^m b$, if one of the following four equivalent conditions holds.

(1) $\int_0^t a^*(u)du \le \int_0^t b^*(u)du$ for all $0 \le t < \nu(\mathscr{X})$

(2) $\int \psi(a)\, d\nu \le \int \psi(b)\, d\nu$ for all convex, continuous real-valued functions ψ such that $\psi(0) = 0$ and the integrals exist

(3) $\int [a-t]_+ \, d\nu \le \int [b-t]_+ \, d\nu$ for all $t \ge 0$

(4) $\int_t^\infty m_a(s)ds \le \int_t^\infty m_b(s)ds$ for all $t \ge 0$

Note that Definition 5 follows from Definition 6 with $\mathscr{X} = \{1, \ldots, n\}$, $a = (a_1, \ldots, a_n)$, $b = (b_1, \ldots, b_n)$, and ν being counting measure. Next a dependence ordering is introduced from majorization to contrast with Definition 4.

Definition 7 (Dependence ordering.) Let $\mathscr{F} = \mathscr{F}(F_1, \ldots, F_d)$ be the class of multivariate densities (with respect to a measure ν) with given univariate margins $F_1, \ldots, F_d$. Suppose F_j has density f_j, $j = 1, \ldots, d$. The (constrained) majorization ordering $f \prec^m g$ of densities $f, g \in \mathscr{F}$ is interpretable as an ordering of dependence. The density $f_I = \Pi_j f_j$ is among those that are minimal in $\mathscr{F}$ with respect to $\prec^m$.

The interpretation of this ordering is different from the concordance ordering in Section Stochastic orders. Densities in $\mathscr{F}$ are larger in this majorization ordering if they diverge more from $f_I = \Pi_j f_j$. Densities for a random vector that almost satisfy a functional relationship are large in the majorization dependence ordering.

Contrast of Stochastic and Majorization Orderings

Some relations between stochastic and majorization orderings are summarized in this section. First consider the case of probability vectors over a finite ordered set of size n, indexed as $\{1, \ldots, n\}$. Let $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_n)$ be two probability vectors. If $\mathbf{p}$,$\mathbf{q}$ are increasing, then $\mathbf{p} \prec^{st} \mathbf{q}$ and $\mathbf{p} \prec^m \mathbf{q}$ are the same, and if $\mathbf{p}$,$\mathbf{q}$ are decreasing, then $\mathbf{p} \prec^{st} \mathbf{q}$ and $\mathbf{q} \prec^m \mathbf{p}$ are the same. Next, consider probability measures P_1,P_2 that have masses n^{-1} at $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, respectively. Then (2) of Definition 5 becomes $\int \psi dP_1 \le \int \psi dP_2$, which is of the form of (2) of Definition 2.

An example of an overlap of a stochastic ordering for dispersion and the general majorization ordering is called the Lorenz ordering in Arnold (1987), and other names in mathematics and economics. Since majorization orderings are orderings of divergence from uniformity, this connection is not surprising. In Definition 6, let $\mathscr{X} = [0, 1]$, let ν be Lebesgue measure, and let $a = F^{-1}$, $b = G^{-1}$, where F,G are the respective distribution functions of nonnegative random variables X,Y (with a common finite mean). Since a,b are monotone increasing, the decreasing rearrangements are respectively, $a^*(t) = F^{-1}(1-t)$ and $b^*(t) = G^{-1}(1-t)$. Condition (2) of Definition 6 becomes $\int_0^1 \psi(F^{-1}(t))dt \le \int_0^1 \psi(G^{-1}(t))dt$ or

$$\mathrm{E}[\psi(X)] = \int_0^\infty \psi(s)\mathrm{d}F(s) \le \int_0^\infty \psi(s)\mathrm{d}G(s) = \mathrm{E}[\psi(Y)] \quad [1]$$

for all convex continuous functions ψ. Note that this then implies that $\mathrm{E}(X) = \mathrm{E}(Y)$, by taking $\psi(s) = s$ and $\psi(s) = -s$.

An interpretation is that with a given mean wealth μ, a distribution of wealth F with a constant F^{-1} is most equitable (this corresponds to a mass of 1 at μ) and a distribution F which is larger in the (Lorenz) ordering is less equitable. In this context, the Lorenz curve of a distribution F is given by

$$L_F(s) = \left[\int_s^1 F^{-1}(1-t)dt\right] \Big/ \left[\int_0^1 F^{-1}(1-t)dt\right]$$

$$0 < s < 1$$

This is roughly the complement of the term given in (1) of Definition 6, so a Lorenz curve that is farther below the 45° diagonal is less equitable.

Applications and Illustrative Examples

Majorization and stochastic orderings are interesting to study for their own sake because of interesting mathematical properties, but they have also been widely applied, for example, in operations research, queuing theory, reliability, economics, and many areas of mathematics and statistics. For any ordering, the class of order preserving or isotonic functions is of interest. Member functions of an isotonic class can be used as measures of various types; examples are measures of inequality and measures of dependence.

Example 1. (Measures of inequality and dependence.) Measures of inequality include Gini indices and extensions, poverty indices in economics, and measures of diversity in biology (Marshall et al., 2011; Foster and Shorrocks, 1988). For summarizing data consisting of a mix of categorical and continuous variables, measures of association or dependence are useful. Concordance orderings lead to measures of monotone association like Kendall tau and Spearman rho (Joe, 1990). Constrained majorization orderings lead to measures of divergence from independence like relative entropy measures and uncertainty measures (Joe, 1989).

Example 2. (Random utility models for choice in mathematical psychology.) The setting in random utility choice models is that there is a superset of d items that can be compared in various ways. Suppose the items are assumed to be indexed as $1, \ldots, d$ and $S = \{1, \ldots, d\}$ denotes the index set. Associated with item i is a utility random variable U_i, $i = 1, \ldots, d$; $U_1, \ldots, U_d$ can be independent or dependent. If the U_i are independent, then stochastic ordering results on the U_i lead to preference orderings of various types. The following results are straightforward to prove when the U_i are independent, $U_i \sim F_i$, and there is a zero probability of ties among the U_i,

(1) (i) If $F_k \prec^{st} F_j$, then $p_{jk} = \Pr(U_j > U_k) \ge \frac{1}{2}$ or item j is pairwise preferred to item k. (ii) If $F_k \prec^{st} F_j \prec^{st} F_i$, then $p_{ij}, p_{jk} \ge \frac{1}{2}$ and $p_{ik} \ge \max\{p_{ij}, p_{jk}\}$, or items i, j, k are ordered by strong stochastic transitivity.

(2) If $F_k \prec^{st} F_j$, then $p_{j,S} \ge p_{k,S}$ for S such that $j, k \in S$, where $p_{i,S} = \Pr(U_i > U_\ell, \ell \in S \backslash \{i\})$ is a choice probability of the set S, a subset of $\{1, \ldots, d\}$ with cardinality at least 2. This means the stochastic order implies a preference ordering according to choice probabilities.

(3) Let $r(a_1, \ldots, a_d) = \Pr(U_{a_1} > \cdots > U_{a_d})$, where $(a_1, \ldots, a_d)$ is a permutation of $(1, \ldots, d)$. Let τ_{jk} be a transposition operator so that $\tau_{jk}(a_1, \ldots, a_d)$ is the vector with a_j, a_k transposed. If $F_d \prec^{LR} \cdots \prec^{LR} F_1$, then the items are preference ordered in the strong sense of $r(1, \ldots, d) \geq r(\tau_{jk}(1, \ldots, d))$ for all $1 \leq j < k \leq d$. If the weaker condition $U_k \prec^{HR} U_1$ $(2 \leq k \leq d)$ is assumed, the strongest conclusion is $r(1, \ldots, d) \geq r(\tau_{1k}(1, \ldots, d))$, and if the weaker condition $U_d \prec^{RHR} U_j$ $(1 \leq j < d)$ is assumed, the strongest conclusion is $r(1, \ldots, d) \geq r(\tau_{jd}(1, \ldots, d))$.

These results contrast the strength of conclusions that are possible with various stochastic orders of largeness. Generally, no preference ordering results follow from stochastic ordering assumptions if the U_i are dependent.

Example 3. Comparison of risks for portfolios. Let X, Y be two continuous random variables representing the returns of two portfolios A,B. B is a better portfolio than A if $\mathrm{E}(X) \leq \mathrm{E}(Y)$ and Y is less dispersed than X in some sense. If the ordering of dispersion is taken from the stochastic/majorization ordering in eqn (1), as

$$\mathrm{E}(\psi(\mathrm{Y} - \mathrm{E}[Y])) \leq \mathrm{E}(\psi(\mathrm{X} - \mathrm{E}[X]))$$

for all convex continuous functions ψ, then $\mathrm{E}[u(X)] \leq \mathrm{E}[u(Y)]$ for all increasing concave (utility) functions, and hence this comparison of risk agrees with utility theory with risk averse utilities. See Mosler (1983) for related results.

The examples presented here give an idea of how the orderings have been applied. More recent examples are given in Arnold (2007). The reader can consult the references for many other interesting applications.

See also: Order Statistics; Ordered Relational Structures; Partial Orders.

Bibliography

Arnold, B.C., 1987. Majorization and the Lorenz Order. Springer, New York.
Arnold, B.C., 2007. Majorization: here, there and everywhere. Statistical Science 22, 407–413.
Foster, J.E., Shorrocks, A.F., 1988. Poverty orderings and welfare dominance. Social Choice Welfare 5, 179–198.
Joe, H., 1989. Relative entropy measures of multivariate dependence. Journal of the American Statistical Association 84, 157–164.
Joe, H., 1990. Multivariate concordance. Journal of Multivariate Analysis 35, 12–30.
Joe, H., 1993. Generalized majorization orderings and applications. In: Shaked, M., Tong, Y. (Eds.), Stochastic Inequalities. Lecture Notes-Monograph Series, vol. 22. IMS, Hayward, CA, pp. 145–158.
Joe, H., 1997. Multivariate Models and Dependence Concepts. Chapman & Hall, London.
Marshall, A.W., Olkin, I., Arnold, B.C., 2011. Inequalities: Theory of Majorization and Its Applications, second ed. Springer, New York.
Mosler, K., 1983. Increasing multivariate risk: some definitions. In: Beckmann, M., Krelle, W. (Eds.), Risk and Capital. Lecture Notes in Economics and Mathematical Systems, vol. 227. Springer, Berlin, pp. 88–102.
Mosler, K., Scarsini, M. (Eds.), 1993a. Stochastic Orders and Decision Under Risk. Lecture Notes Monograph Series, vol. 19. Institute of Mathematical Statistics, Hayward, CA.
Mosler, K., Scarsini, M., 1993b. Stochastic Orders and Applications. A classified bibliography. Lecture Notes in Economics and Mathematical Systems, vol. 401. Springer, Berlin.
Shaked, M., Shanthikumar, J.G., 2007. Stochastic Orders. Springer, New York.
Stoyan, D., 1983. Comparison Methods for Queues and Other Stochastic Models. Wiley, New York.

Malinowski, Bronislaw (1884–1942)

Michael W Young, Australian National University, Canberra, ACT, Australia

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 9147–9151, © 2001, Elsevier Ltd.

Abstract

Born in Cracow, Austrian Poland, Malinowski studied mathematics, physics, and philosophy at the Jagiellonian University before turning to 'primitive sociology' and ethnology in London. Between 1914 and 1918 he conducted three expeditions to Papua, during which he 'discovered' the ethnographic benefits of 'living right among the natives' and mastering their language. Intensive fieldwork was to become the hallmark of British anthropology, and it was largely through Malinowski's mythicized 'heroic' example that it became a rite of passage for entry into the profession. Malinowski spent the interwar years teaching at the London School of Economics, where he was appointed Professor of Social Anthropology in 1927. He promoted the most significant paradigm shift in British anthropology since its creation in the mid-nineteenth century: from the speculatively historical study of cultural traits as 'survivals' to the synchronic study of social 'institutions' within bounded, functioning cultures. Malinowski's 'functionalism' was essentially an appeal for contextualized empirical data purified of historical speculations. It was influential during the 1920s and 1930s and helped him secure Rockefeller funding for his students' research in Africa. His enduring conceptual contributions to anthropology lay in areas such as kinship, marriage, and the family; in magic, myth, and religion; in economic anthropology; and in pragmatic linguistic theory.

A Polish-born anthropologist, Malinowski had a profound influence on the development of British social anthropology between the wars. Under the banner of Functionalism he revolutionized fieldwork methods, contributed to a wide range of academic and public issues, and trained a cosmopolitan cohort of pupils which spread his teachings throughout the world. A passionate, volatile, and sometimes abrasive personality, Malinowski provoked controversy throughout his career; but his intellectual integrity was absolute. He promoted the humanitarian uses of anthropology with entrepreneurial flair, attracting Rockefeller funding to his cause. In addition to his ancestral status as a founder of a cohesive academic discipline, Malinowski's historical legacy in Britain and the Commonwealth was an expanded institutional base for its teaching and practice. In Ernest Gellner's ennobling epithet, Malinowski was 'Anthropologist Laureate of the British Empire.'

A Polish Anglophile

Bronislaw Malinowski was born on 7 April 1884, of noble (*szlachta*) stock in Cracow, capital of the semi-autonomous province of Galicia in the Austro-Hungarian Empire. As a child he lived for several years among local peasantry and, with his mother, took lengthy trips abroad for the sake of his delicate health. To these early experiences of cultural difference Malinowski attributed his vocation to anthropology. He was also inspired by reading James Frazer's *The Golden Bough.*

During Malinowski's youth, Cracow was a vibrant intellectual and artistic center, home of the Modernist movement called Young Poland. His most important early friendship was with the surrealist artist, novelist, and dramatist, Stanislaw Ignacy Witkiewicz. Unable to match his friend's artistic genius, Malinowski resolved to excel in the sciences. At the Jagiellonian University he studied mathematics, physics, and philosophy. His brilliant doctoral thesis, 'On the Principle of Economy of Thought' (1906), was a critique of the positivist epistemology of Ernst Mach; it contained the seeds of Malinowski's later thinking about functionalism (Thornton and Skalnik, 1993).

In 1908–09, during three semesters at the University of Leipzig, Malinowski turned to the study of *Vökerpsychologie* under Wilhelm Wundt and economic history under Karl Buecher. Possessed by 'a highly developed Anglomania,' in 1910 Malinowski crossed the English Channel to enrol as a graduate student in ethnology at the London School of Economics (LSE) in the University of London. The crossing was his Rubicon, and although not granted British nationality until 1931, he was to spend his most productive years in England.

Under the supervision of Edward Westermarck, Malinowski completed his first English book, *The Family Among the Australian Aborigines* (1913), an evaluation of Aboriginal ethnography and a rigorous critique of evolutionist theories. He presented a signed copy to his Polish compatriot, Joseph Conrad, whom he deeply admired.

The Cult of Fieldwork

The dominant paradigm of late-nineteenth century anthropology (or ethnology) was evolutionism, based on the premise that mankind had passed through universal stages of development on the journey from savagery to civilization. A division of labor existed in imperial Britain between armchair theorists (like Frazer) and amateur ethnographers (explorers, missionaries, and colonial officers) who reported from outposts of the Empire.

The professionalization of anthropology began following the Cambridge University Expedition to the Torres Strait of 1898–99 led by the zoologist-turned-anthropologist A.C. Haddon. By 1910 anthropology was being taught at Cambridge, Oxford, and London. It was due to Haddon and his disciples W.H.R. Rivers and C.G. Seligman that Malinowski found an almost messianic cult of fieldwork in Britain. In Seligman's words, 'field research in anthropology is what the blood of martyrs is to the Church.' By the time Malinowski

embarked on his own fieldwork, at least a dozen students of Haddon and his colleagues had preceded him.

World War One broke out just as Malinowski arrived in Australia. Although cut off from his English funds and under suspicion as an 'enemy alien,' he was permitted entry to Papua (ex-British New Guinea) to conduct his research. His apprentice fieldwork was among the Mailu of the south coast, whence Seligman had directed him to fill a gap left by his own ethnographic survey of Papua a decade earlier. After five busy months, Malinowski returned to Australia where he dashed off *The Natives of Mailu*, a conventional ethnographic report (Young, 1988). In June 1915 he returned to eastern Papua with Australian funding for a second spell of fieldwork. Supposedly en route to another field site, Malinowski visited Kiriwina, the largest of the Trobriand Islands, about which Seligman had already reported comprehensively in *The Melanesians of British New Guinea* (1910). For various reasons Malinowski's visit to Kiriwina evolved into a sojourn of 10 consecutive months, followed by a second period of 10 months in 1917–18. Although his movements were restricted by the colonial authorities, Malinowski was not, as his popular legend had it, 'interned' in the Trobriands.

He had learned in Mailu that the richness of his data was a function of the time he had spent living in the village, shunning the company of fellow Europeans and observing everyday life at first hand. It was also imperative to master the local language well enough to have unmediated access to one's subjects. Under the rubric 'the intensive study of limited areas,' such precepts of method had been advocated by Haddon and Rivers during the previous decade, but they had not been put into proper practice until Malinowski's work in Kiriwina.

Thus, the innovative breakthrough his fieldwork effected was based on the simple expedient of living in the same location for many months and probing for information in the vernacular. This method of 'total immersion' encouraged a more intimately interactive style of ethnography, permitting an empathetic understanding which went beyond mere recording. The fieldworker could begin to know another culture from within: 'the final goal' of the ethnographer is 'to grasp the native's point of view ... to realize *his* vision of *his* world' (Malinowski, 1922: p. 25). 'Participant observation' was to become the catch-cry of his pupils' generation. It still remains the hallmark of anthropological fieldwork. Professionally, it was prolonged and intensive fieldwork, with its loneliness and hardship, that stamped the new breed of anthropologist (Stocking, 1995: Chap. 6).

Trobriand Man as Exemplar

Malinowski was criticized for making his Trobri-anders the measure of mankind, but he did so persuasively as Trobriand Man was introduced into the controversies of the day. On close inspection, this exemplar of the Savage was not so very different from Civilized Man. He was demonstrably rational and technologically competent, morally and aesthetically sensitive: an equal, therefore, and not (as imperialists and evolutionists would have it) a benighted inferior. He believed in magic, certainly, but he used it only in situations where technology could not guarantee an outcome. 'Magic helps those who help themselves.' Magic was a tool to which men resorted when the desired end of any enterprise was in doubt. This instrumental, pragmatic view of magic was typical of Malinowski's functionalism.

In 1922 Malinowski published his first Trobriand classic, *Argonauts of the Western Pacific*, which set British anthropology firmly on a new course (Malinowski, 1922). It was a richly textured study of *kula*, the circuitous ceremonial exchange of shell valuables which linked the Trobriands to neighboring archipelagoes. Malinowski's combination of ethnographic observation, methodological prescription, and literary romanticism was vivid and compelling. An enthusiastic preface by Sir James Frazer ensured *Argonauts'* successful launch.

With *Crime and Custom in Savage Society* (1926), a little book on 'the seamy side' of Trobriand law, Malinowski pioneered the study of political and legal anthropology. The following year, *Sex and Repression in Savage Society* set a cat among the Freudian pigeons by contesting the claim that the Oedipal Complex was universal (Malinowski, 1927). Malinowski argued that in the matrilineal Trobriands it was the mother's brother, not the father, who was the hated authority figure; likewise, it was the sister, not the mother, who was the main object of a boy's incestuous desire. The debate continues.

The Sexual Life of Savages (Malinowski, 1929) was a detailed study of Trobriand courtship, marriage, and domestic life; it was also a covert attack on sexual prudery and hypocrisy, and in the changing mores of the period it had a liberating appeal.

Malinowski's last Trobriand ethnography, *Coral Gardens and their Magic* (1935), dealt exhaustively with horticultural practices and their ritual accompaniments, with land tenure and its politics, and with the language of magic. The two-volume work was a triumph of thick description and a baroque monument to its author's functional method.

Malinowski fancied that his monographs did for anthropology what Joseph Conrad's novels had done for English literature. It is true that through these and many other writings, Malinowski's Trobrianders came to be known as one of the best-documented 'primitive' peoples in the world. Subsequent generations of anthropologists have reinterpreted his material in the light of different theories – the most flattering compliment posterity can pay to any ethnographer.

Functionalism

The paradigmatic theory that informed Malinowski's Trobriand monographs was functionalism. As an heuristic method which guided fieldwork, functionalism was an appeal for contextualized empirical data purified of historical speculations. Malinowski urged ethnographers to look for the use of a custom (or institution) in the present, so as to determine the part it played in the integral working of a society or culture. This would give the clue to its meaning and the explanation for its existence, one that (in contrast to the speculative explanations of 'antiquarian' anthropologists) was concretely based on empirical observation. Although functionalism was demonstrably deficient when applied to societies undergoing rapid social change, as a 'clinical' method it was immensely influential during the interwar years. The promise it offered

'practical' or applied anthropology in African colonies enabled Malinowski to secure, through the sponsorship of the International African Institute, Rockefeller funding for his students' fieldwork.

His theory of language was based on the premise that function determined form. 'Primitive' language was 'a mode of action' rather than a 'countersign of thought.' Much speech is 'phatic' (agreeable social noises such as 'How are you?') or instrumentally pragmatic, meaningful only in its 'context of situation.' Malinowski foreshadowed Ludwig Wittgenstein's axiom 'do not ask the meaning but the use.'

As grand theory, Malinowski's functional view of culture as an apparatus for the satisfaction of human needs was born out of Machian empiricism, confirmed by Jamesian pragmatism, and blessed by the Durkheimian tenet that culture was *sui generis*. Malinowski's 'scientific theory of culture' (1944a) was built on the premise that man's biological heritage provided the ground plan for all cultures; basic (or primary) needs such as reproduction and derived (or secondary) needs such as education yielded similar cultural solutions, the universal form of which was the family. In a nutshell, he argued (a) that function determines the form of cultural entities; (b) that the essential element of the science of culture is the institution; and (c) that institutions must ultimately be defined in relation to the biological needs they satisfy. As formal theory, then, functionalism explained culture in an instrumental and reductionist fashion. The schematic, utilitarian form in which he presented this theory gained few adherents and it was soon forgotten.

Teacher, Pundit, and Propagandist

Malinowski's rise to eminence at the LSE had been rapid: from Lecturer in 1923, Reader in 1924, to foundation Professor of Social Anthropology in 1927. The mode of teaching he favored was the seminar. In Socratic fashion, he deployed wit, erudition, and 'intellectual shock tactics' to provoke thoughtful argument. His electrifying seminar became a legend during the 1930s, engaging his brightest pupils and attracting from every continent students of cognate disciplines. The multiethnic roll-call of his pupils would fill a page, but among the most eminent were Raymond Firth, Edward Evans-Pritchard, Ashley Montagu, Isaac Schapera, Audrey Richards, Camilla Wedgwood, Meyer Fortes, Paul Robeson, Siegfried Nadel, Lucy Mair, Reo Fortune, Ralph Bunche, Ian Hogbin, Jomo Kenyatta, Talcott Parsons, Gregory Bateson, Edmund Leach, and Phyllis Kaberry.

Despite his fragile health, Malinowski was a ferociously hard worker who mustered others to his aid. He employed servants at home, dictated most of his writings to stenographers, and employed pupils as long-suffering research assistants. Until she was tragically incapacitated by multiple sclerosis, Elsie, his Australian-born wife, served him as a loving critic and secretary. Together with their three daughters they spent summers at the family villa in the South Tyrol, where pupils often visited to walk and talk (Wayne, 1995).

In tandem with his academic activities, Malinowski performed the role ofpublic intellectual, disseminating progressive views on marriage, the family, and sex education. He lectured missionaries and colonial officers; he gave BBC talks on religion; he actively supported the British Social Hygiene Council and Mass Observation. More importantly, he served on the board of the International African Institute for which he devised a field training program. In 1934 he visited several of his pupils in South and East Africa before conducting some fieldwork of his own in Tanganyika.

Malinowski had lectured widely in the United States in 1926, 1933, and 1936, and he returned for a sabbatical year in October 1938. When war broke out in Europe he was advised to remain in America for the duration. He taught at various universities, anxious to secure a permanent position. As 'war work' he embarked on a lecturing and writing campaign of propaganda (his own term) to convince Americans of the need to fight Hitler. Of all the honors he received he valued most the distinction the Nazis conferred upon him by burning his books. His opposition to totalitarianism was categorical and he wrote passionately against it in *Freedom and Civilization* (Malinowski, 1944b).

During the summers of 1940 and 1941 Malinowski conducted fieldwork in the Oaxaca Valley of Mexico with his new wife and a Mexican assistant. This fieldwork did not match the exacting standards he had set in the Trobriands. Just before his sudden death of a heart attack on 16 May 1942 in New Haven, CT, Malinowski was appointed to a permanent professorship at Yale University and elected President of the Polish Institute of Arts and Sciences in New York.

Malinowski's Legacy

Malinowski bequeathed a controversial legacy. He did more than anyone else of his time to establish a cohesive (even 'clannish') school of anthropology with a unique academic profile. But the year he left for America was the year A.R. Radcliffe-Brown returned to England to create a rival school at Oxford. The latter's more Durkheimian and sociologically austere version of functionalism (known as structural functionalism) attracted many of Malinowski's erstwhile disciples. Radcliffe-Brown dominated British social anthropology during the following decade.

In 1957, a dozen of Malinowski's pupils published a Festschrift which critically evaluated his contribution to different subfields. While conceding that he was an incomparable ethnographer, the majority verdict was that Malinowski was a flawed and inconsistent thinker who had failed to grasp the analytical priority of 'systems' and 'structures.' It thus became a cliche of disciplinary folk history that Malinowski was a brilliant ethnographer but a mediocre theoretician.

This view justified neglect of his work until the late 1960s, when, abetted by Raymond Firth, Malinowski's successor at LSE, the pendulum swung from a theoretical preoccupation with structure to the analysis of process and event. Malinowskian 'methodo logical individualism' became respectable again through decision-making analysis and transactional anthropology. Meanwhile there was a rapprochement between British social and American cultural anthropology as their research agendas converged. American anthropologists had always been receptive to Malinowski's field methods, and eventually the kind of fieldwork they conducted became

indistinguishable from that of their British and Commonwealth colleagues.

In 1967, Malinowski's widow published his personal field diaries. They were an embarrassment to many of his pupils and Raymond Firth's preface to the book is an exercise in damage control. Malinowski's agonized musings on the alienation he experienced in the field put the lie to the effortless rapport he was believed to have enjoyed with Trobrianders. His diaries exacerbated, if not precipitated, a moral and epistemological crisis of anthropology during the 1970s, when global decolonization was turning attention to the exploitative aspects of the discipline and its dubious historical role as a handmaid of Imperialism.

The occasion of Malinowski's centenary celebrations in 1984 produced another round of evaluations, the most notable being a collection of essays by Polish scholars. They not only rehabilitated Malinowski's reputation in his homeland, but painstakingly examined his theories in the light of his Polish roots. He was declared a scientific humanist and rationalist, a romantic positivist and political liberal who belonged in the mainstream of modernist European thought (Ellen et al., 1988).

In America the diaries stimulated fresh debates on the assumptions underlying Malinowskian fieldwork and ethnographic representation. Postmodernists appropriated Malinowski for their own rhetorical ends in deconstructing ethnographic authority. To an ironic degree, however, he had pre-empted some of their concerns by the humanistic strain in his work and, except when promoting his 'science' for institutional support on behalf of 'practical anthropology,' he was by no means a naive positivist. Nor was he a 'romantic primitivist' as some critics have charged, despite the Frazerian echoes in his earliest monographs. Fully 50 years before postmodernists deconstructed exoticism, Malinowski had come to the view that anthropology romantically 'over-sensationalized' differences between cultures, and he questioned its fascination with 'freakish' customs such as polyandry, widow burning, and headhunting.

Despite the fragmentation of socio-cultural anthropology at the millennium, Malinowski's fruitful concepts can be detected in almost every subfield, and his legacy persists wherever ethnography is practiced. His name is indelibly linked with Melanesia, but it appears too in anthropological writings from countries as far-flung as Mexico and China, Italy and New Zealand, and India and Ireland. A protean, cosmopolitan-ancestral figure, then, Malinowski continues to fascinate his intellectual descendants.

See also: Economic Anthropology; Ethnography; Fieldwork in Social and Cultural Anthropology; Functionalism, History of; Kula Ring, Anthropology of; Participant Observation.

Bibliography

Ellen, R., Gellner, E., Kubica, G., Mucha, J. (Eds.), 1988. Malinowski Between Two Worlds: The Polish Roots of an Anthropological Tradition. Cambridge University Press, Cambridge, UK.

Firth, R.W. (Ed.), 1957. Man and Culture: An Evaluation of the Work of Bronislaw Malinowski. Routledge and Kegan Paul, London.

Malinowski, B., 1922. Argonauts of the Western Pacific. Routledge and Kegan Paul, London.

Malinowski, B., 1926. Crime and Custom in Savage Society. Routledge and Kegan Paul, London.

Malinowski, B., 1927. Sex and Repression in Savage Society. Routledge and Kegan Paul, London.

Malinowski, B., 1929. The Sexual Life of Savages. Routledge and Kegan Paul, London.

Malinowski, B., 1935. Coral Gardens and their Magic. vols 2. Allen & Unwin, London.

Malinowski, B., 1944a. A Scientific Theory of Culture. Chapel Hill, University of North Carolina Press, NC.

Malinowski, B., 1944b. Freedom and Civilization. Roy, New York.

Malinowski, B., 1967. A Diary in the Strict Sense of the Term. Routledge, London.

Stocking, G.W., 1995. After Tylor. University of Wisconsin Press, Madison, WI.

Thornton, R.J., Skalnik, P. (Eds.), 1993. The Early Writings of Bronislaw Malinowski. Cambridge University Press, Cambridge, UK.

Wayne, H. (Ed.), 1995. The Story of a Marriage: The Letters of Bronislaw Malinowski and Elsie Masson. vols 2. Routledge, London.

Young, M.W. (Ed.), 1988. Malinowski Among the Magi: 'The Natives of Mailu'. Routledge, London.

Malthus, Thomas Robert (1766–1834)

Jacques Dupâquier, EHESS, Paris, France

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, pp. 9151–9156, © 2001, Elsevier Ltd.

Abstract

Thomas Robert Malthus (1766–1834) was brought up in the ideology of progress, but made a complete break with progressive thought with his *Essay on the Principle of Population* (1798), in which he argued that it is useless, even dangerous, to provide relief for the poor, as this encourages them to reproduce. The success of this work provoked a heated discussion, obliging Malthus to expand and elucidate his ideas on population. This resulted in a second *Essay* (1803), markedly different from the first. This work established Malthus's lasting reputation; he was appointed to a chair in political economy – the first of its kind in Europe – and elected a member of the Royal Society. In 1820, he published another major work: *Principles of Political Economy*. Malthus's theory soon became known on the continent, particularly in France and Germany, but not without misconceptions or misinterpretations. Most of his opponents have only taken up a caricatural model according to which population always increases in a geometrical progression, while the means of subsistence only increase in an arithmetical progression. In fact, Malthus referred only to tendencies: demographic growth is always checked, be it preventively or positively. In the first *Essay*, these checks were identified as misery (in the wider sense) and vice; in the second *Essay*, Malthus introduced the notion of moral restraint (meaning late marriage). All this was embedded in Malthus's moral and theological reflection on progress and Providence; but the *Principle of Population* also implied pessimistic views on poverty and state assistance, meaning that Malthus's reputation – among socialists and progressivists at least – was permanently tarnished. This perception was quite unjustified, since Malthus was a liberal member of the Whig Party and in favor of reform. Malthus's intellectual legacy was extremely varied. On the economic level, the theory of effective demand, which he introduced in the *Principles*, was a lasting success; on the political level, the very term 'Malthusian' was coined as early as 1822 by Francis Place, who proposed a much easier and more effective method of checking population growth than moral restraint: the limitation of births within marriage. It is in this form, representing a complete break with Malthus' theory, that neo-Malthusianism became established in the nineteenth century, gaining wide acceptance after the Second World War, even at the United Nations.

The passions aroused even today by the name of Malthus bear witness to the eminent position – almost equal to that of Marx – which he still occupies in the history of ideas.

An Unspectacular Life

Thomas Robert Malthus was born on 13 February 1766, in 'The Rookery' in Dorking, near Wooton, Surrey, England. His father, Daniel Malthus, was a prosperous, brilliant, and educated man, a convinced progressive who corresponded with Voltaire and particularly with Rousseau, who visited him on March 9 of the same year.

Very much taken with new ideas, Daniel Malthus brought up his son in accordance with the precepts of *Émile*, and put him in the charge of two dissentient intellectuals: Richard Graves (1715–1804), who had to leave Cambridge as a result of a misalliance, and Gilbert Wakefield (1756–1801), a former pastor who taught classical literature at Warrington Academy and was imprisoned in 1799 for writing that the poor would have nothing to lose from a French invasion. To crown this educational career, Malthus, who had been admitted to Jesus College, Cambridge, in 1784, became a pupil of William Frend (1757–1841), who was dismissed from the university in 1793 for publishing three pamphlets against the Church. It was perhaps Frend who aroused Malthus' interest in economics and demography (Petersen, 1979).

In 1788 Malthus took holy orders, in 1793 he was elected a fellow of Jesus College and became rector of Oakwood Chapel (Surrey), 12 or 13 km from the village of Albury, where his father Daniel had settled in 1787. Some years later, he wrote a first essay entitled *The Crisis, a View of the Recent Interesting State of Great Britain, by a Friend of the Constitution*. No publisher could be found for this work, however, and only extracts – published after Malthus' death by his friends William Empson and William Otter – survive (Otter, 1836). These reveal a very politically correct Malthus, an advocate of the provision of state assistance at home for those not capable of securing their own livelihood.

Two years later Malthus' polemic pamphlet was published, marking a complete break with the progressive ideas of his entourage: *An Essay on the Principle of Population as it Affects the Future Improvement of Society; With Remarks on the Speculations of Mr. Godwin, M. Condorcet and Other Writers*. The essay was published anonymously, but Malthus was recognized immediately as its author. The pamphlet established Malthus' name, but its central thesis, which was expressed with an apparent scientific detachment, that it is useless, even dangerous, to provide relief for the poor as this encourages them to reproduce, created a scandal.

Malthus, described by all those who met him as a gentle, courteous, sincere, and sensitive man, was thus drawn into a heated and interminable discussion, obliging him to expand and elucidate his ideas on population.

After traveling with his Cambridge friends to Scandinavia (1799), France, and Switzerland (1802), and producing a new essay on price increases (*An Investigation of the Cause of the Present High Price of Provisions* 1800), thus demonstrating his talents as an economist, Malthus published a new book in 1803. This is regarded as a second edition of the 1798

Essay; and indeed entire chapters were taken from the first edition, but the 1803 edition differs fundamentally with respect to its complete title (*An Essay on the Principles of Population, or a View of its Past and Present Effects on Human Happiness with an Inquiry into our Prospects Respecting the Future Removal or Mitigation of the Evils which it Occasions*), its (fourfold) length and, above all, its content, now constituting moral and sociological reflections on population growth. Even during the author's lifetime, the new *Essay* was revised four times with additions and corrections.

In April 1804 Malthus, now 38, married his cousin Harriet Eckersall at Claverton, near Bath. Within four years, she had given birth to one son and two daughters (Petersen, 1979).

In 1806 he was appointed professor of history and political economy at the East India Company's college, which opened at Hertford before moving to Haileybury (Hertfordshire) several years later. He held this chair in political economy – the first of its kind – until his death, and publicly defended the college when its existence was jeopardized (Petersen, 1979).

For 14 years, Malthus produced little more than a variety of pamphlets – particularly on the importance of grain – but with the five successive editions of his second *Essay* he established such a reputation that he was elected a member of the Royal Society (1819), a corresponding member of the Royal Society of Literature (1824), and later an associate member of the *Académie des Sciences Morales et Politiques* and the Royal Academy of Berlin (1833). In 1821 he was one of the founder members of the Political Economy Club and, in the year of his death, of the Royal Statistical Society.

In 1820 Malthus's second *magnum opus* was published: *Principles of Political Economy Considered with a View to Their Practical Application*. Being of a concrete nature, it was opposed to the more systematic and dogmatic views of David Ricardo (Keynes, 1939).

After this, there is little to report apart from the publication of an article entitled 'Population' in the *Encyclopaedia Britannica* (1824) and further trips to the continent and to Scotland. Having gone to Bath to spend Christmas with his family, Malthus died of a heart attack on 29 December 1834, at the age of 68. His grave is to be found in Bath Abbey.

Malthus's Contribution to the Sociology and Economy of Populations

What the general public, nonspecialists and, regrettably, many historians and economists have taken up of Malthus' theory is a caricatural model according to which population always increases in a geometrical progression, while the means of subsistence only increase in an arithmetical progression, thus making it necessary to limit births by all possible means.

In fact, Malthus' theory went through a complex evolutionary process, developing from one work to the next. His central ideas, however, were already present in the 1798 *Essay*.

Originally, this first *Essay* was merely a philosophical pamphlet intended as a criticism of the utopian optimism of Godwin and Condorcet. The author's main aim was to undermine the theory that it was possible for man and society to progress. In the conclusion to chapter VII, he resumes his arguments as follows:

> Must it not then be acknowledged by an attentive examiner of the histories of mankind, that in every age and in every state in which man has existed, or does now exist,
> –That the increase of population is necessarily limited by the means of subsistence.
> –That the population does invariably increase when the means of subsistence increase. And,
> –That the superior power of population is repressed, and the actual population kept equal to the means of subsistence by misery and vice (Malthus, 1798).

In the preface, Malthus states that "(t)he following Essay owes its origin to a conversation with a friend (very probably Daniel Malthus, the author's father, who was to die two years later, taking his generous ideas and fabulous illusions with him to the grave) on the subject of Mr. Godwin's Essay on avarice and profusion in his Enquirer."

The discussion started the general question of the future improvement of society, and the author at first sat down with an intention of merely stating his thoughts to his friend upon paper in a clearer manner than he thought he could do in conversation. But as the subject opened upon him, some ideas occurred, which he did not recollect to have met with before; and as he conceived that every the least light on a topic so generally interesting might be received with candor, he was determined to put his thoughts in a form for publication (Malthus, 1798).

A generation gap? Certainly, but one which only acquired its full significance from the fact that it occurred at a turning point in history. Indeed, in the first chapter, Malthus refers to "the new and extraordinary lights that have been thrown on political subjects, which dazzle, and astonish the understanding; and particularly that tremendous phenomenon in the political horizon the French Revolution, which, like a blazing comet, seems destined either to inspire with fresh life and vigour, or to scorch up and destroy the shrinking inhabitants of the earth, have all concurred to lead many able men into the opinion that we were touching on a period big with the most important changes, changes that would in some measure be decisive of the future fate of mankind."

Malthus was well aware of the disillusionment and skepticism engendered by this first historical assault on the prevailing ideology: "The view which he has given of human life has a melancholy hue; but he feels conscious that he has drawn these dark tints, from a conviction that they are really in the picture; and not from a jaundiced eye, or an inherent spleen of disposition" (Malthus 1798).

In this first essay, Malthus's argument is backed up by few concrete examples but for a brief examination of the different stages of civilization through which humanity has gone (Chaps. III and IV), case studies of England (Chap. V), the new colonies (Chap. VI), and a note on the speed with which even the old states recovered from the ravages of war, pestilence, famine, and natural catastrophes. The main part of the work (Chaps. VII–XV) is devoted to a rebuttal of the ideas of Wallace, Condorcet, and particularly Godwin, who had just published the third edition of his philosophical treatise *An Enquiry Concerning Political Justice and Its Influence on General Virtue and*

Happiness (Godwin 1798). However, two chapters (Chaps. XVI and XVII) initiated the criticism of Adam Smith's economic theory, and two others (Chaps. XVIII and XIX, both omitted from the second *Essay)* seem to provide the key to Malthus' system of values.

In the tradition of the Scottish moralists – particularly Abraham Tucker – whose influence he had been exposed to at Cambridge, Malthus came to believe that the constant pressure that misery exerts on man leads to a conception of life in which hope is directed towards the afterlife (Tucker, 1776, in Dupaquier and Fauve-Chamoux, 1983). In contrast to these moralists, however, Malthus refused to see life on earth as a trial, but as 'a process necessary, to awaken inert, chaotic matter into spirit; to sublimate the dust of the earth into soul; to elicit an aethereal spark from the clod of clay (Malthus, 1799).

Malthus went on to elaborate a theory of stimulation which anticipated that of Toynbee: "To furnish the most unremitted excitements of this kind, and to urge man to further the gracious designs of Providence by the full cultivation of the earth, it has been ordained that population should increase much faster than food ... consider man, as he really is, inert, sluggish and averse from labour, unless compelled by necessity ... we may pronounce with certainty that the world would not have been peopled, but for the superiority of the power of population to the means of subsistence … . The principle according to which population increases, prevents the vices of mankind, or the accidents of nature, the partial evils arising from general laws from obstructing the high purposes of the creation. It keeps the inhabitants of the earth always fully up to the means of subsistence; and is constantly acting upon man as a powerful stimulus, urging him to the further cultivation of the earth, and to enable it, consequently, to support a more extended population" (Malthus, 1798).

The Malthus of 1798 was a moralist, sociologist, and demographer, rather than an economist. He had begun to reflect on aspects of this domain, however. In his first *Essay*, he criticized Adam Smith's idea that any increase in revenue and capital leads to an increase in the size of the wage fund designated for the upkeep of the labor force, and referred to cases in which increasing wealth had in no way improved the living conditions of poor workers. On the same lines, addressing the problem of how correctly to define the wealth of a nation, he criticized the Physiocrats, and contended that the work of craftsmen and workers was productive for the individuals themselves, although not for the nation as a whole.

Malthus did not return to these thoughts in the first three editions of the second *Essay*, but the gradual development of his theory can be seen in *An Investigation of the Cause of the Present High Price of Provisions* (1800), his teachings at the East India Company's college, his article *Depreciation of Paper Money* in the *Edinburgh Review*, and his publications *Pamphlets on the Bullion Question* in the same journal, *Observations on the Effects of the Corn Laws* (1814), *Principles of Political Economy* (1820), and *Definitions in Political Economy* (1827).

How can Malthus' approach to economics be summed up in a few words? It is not a coherent, systematic and immutable whole: 'Yet we should fall into a serious error,' he wrote in the *Principles*, "if we were to suppose that any propositions, the practical results of which depend upon the agency of so variable a being as man, and the qualities of so variable a compound as the soil, can ever admit of the same kinds of proof, or lead to the same certain conclusions, as those which relate to figure and number" (Malthus, 1820).

Overall, Malthus adhered to the principles of the classical school, but was critical of the idea that human needs and desires are indefinitely and immediately expansible. According to Joseph-J. Spengler, Malthus argued that population growth is conditioned by the 'effective demand' for labor, and not primarily by the productive capacity. "Such a demand in fact only tends to occur when adequate moral and political conditions coincide, when the social structure is flexible, the landed property correctly divided and commerce active, when a sufficient number of individuals are willing and able to consume more material wealth than they have produced, and when human beings are strong enough to compensate for the inelasticity of the demand for goods and services in terms of effort" (Malthus, 1798).

It is primarily where demography, sociology, and morality are concerned that the second *Essay* marks a development in Malthus' thought.

In his first *Essay*, Malthus barely touched upon demographic science or political arithmetics as it was known at the time. He knew his Hume, Wallace, and Price, cited King, Short, and Süssmilch, but, with the exception of the latter, overlooked the work of foreign scholars. He contented himself with rough calculations on 'the unhealthy years' and the relationship between the number of births and the number of burials. In his second *Essay* on the other hand, he devoted one chapter to the fertility of marriages, another to the effect of epidemics on population movements, and a third to emigration though this cannot be defined as demographic analysis.

Where sociology is concerned, the core of the second *Essay* consists in an analysis of the checks to population in the lowest stage of human society (Book I, Chap. III), among the American-Indians (Chap. IV), in the islands of the South Sea (Chap. V), among the ancient inhabitants of the North of Europe (Chap. VI), among modern pastoral nations (Chap. VII), in different parts of Africa (Chap. VIII), in Siberia (Chap. IX), in the Turkish dominions and Persia (Chap. X), in Indostan and Tibet (Chap. XI), in China and Japan (Chap. XII), among the Greeks (Chap. XIII), among the Romans (Chap. XIV), and in the various states of modern Europe: Norway, Sweden, Russia, middle parts of Europe, Switzerland, France, England, Scotland, and Ireland (Book II, Chaps. I–VIII).

For the most part, the sources of these reflections have been located (cf. the contributions of M. Godelier, N. Broc, and J. Stagl in Fauve-Chamoux *Malthus hier et aujourd'hui* (1984), and Malthus's interpretations – especially those concerning primitive societies – have been discussed, but the central element here is the new way of approaching the issue. Malthus, thus, emerges as one of the pioneers of the sociology of populations. It is on the moral and political level, in particular, that Malthus's theory is completed, confirmed, and substantiated in the second *Essay* and the successive editions.

In the first *Essay*, he asserted that the reproductive power of the human population could only be checked by misery and vice, the word 'misery' being used in the broad sense, and 'vice' in the sense of sexuality being deviated outside of the institution of marriage. While he did not yet use the term 'moral

restraint,' he further indicated that the pressure of providing for a family operates to varying degrees in all classes of society, and termed this a 'preventive check' in opposition to the 'repressive check' of misery.

In the second *Essay*, he introduced moral restraint explicitly as another of the checks preventing the population from increasing beyond the limits of the means of subsistence. Two chapters of book IV are devoted to moral restraint, the obligation to practice this virtue, and the effect this would have on society.

All this was embedded in Malthus's reflections on the lot of the poor and the principle of aid. While in 1796 (*The Crisis*) Malthus had defended the principle of state assistance at home, 2 years later in the *Essay* he asserted not only that the immense sums of money collected in conjunction with the poor laws would not improve the living conditions of paupers, but that 'they have spread the general evil over a much larger surface,' having encouraged the poor to reproduce.

In the *Essay* of 1803, Malthus went even further. In order to illustrate the idea that paupers had no right to aid, he conceived the famous analogy of the banquet, in which the theme was carried to a ridiculous extreme. This offending passage was cut out of following editions, but was unearthed in 1820 by Godwin, who had undertaken, rather late in the day, to refute the Malthusian doctrine (Godwin, 1820). In this, he was followed by most of the later writers, either in good faith or otherwise.

Petersen, in his famous book *Malthus Reconsidered* (1979), shows why Malthus is termed reactionary by ideologists, and points out the injustice of this accusation: "Malthus was an active member of the Whig Party, and the social reforms he advocated – in addition to the crucial one of universal schooling – included an extension of the suffrage, free medical care for the poor, state assistance to emigrants, and even direct relief to casual labourers or families with more than six children; similarly, he opposed child labour in factories and free trade when it benefited the traders but not the public. Apart from such details he was an honest and beneficent reformer, committed throughout his life to the goal that he shared with every liberal of his day – the betterment of society and of all the people in it."

The Impact of Malthus and Malthusianism

Malthus's theory soon became known on the continent, particularly in France and Germany. In France, while the *Essay* of 1798 was not translated until 1980 (by Vilquin), parts of the *Essay* of 1803 were presented in the *Bibliothèque Britannique* as early as 1805 by the Genevian Pierre Prévost, who was also responsible for the translation of the third edition in 1809. The *Principles of Political Economy* appeared in French in the year of their English publication. In Germany, Malthus was translated several times from 1806 onwards. In the rest of Europe (Italy, Spain, Russia), Malthus's works were not translated until the second half of the nineteenth century, but as philosophers and economists read them in English, French, and German, they were much discussed (Fauve-Chamoux, 1984).

In view of the frequency of misquotations, mis-readings, second-hand references, silly remarks, and invectives accumulated by these commentators, however, one cannot help but wonder whether the majority of them, even Karl Marx, had really read Malthus's publications.

The most common misconception – particularly among economists – is that Malthus contended that the population really does increase in geometrical progression while the means of subsistence increase in arithmetic progression. In fact, Malthus referred only to tendencies, and refused to allow his theory to be reduced to the simplistic conclusion that population is regulated by subsistence.

An associated misconception casts Malthus as an enemy of demographic growth. In fact, as early as 1805 he made it quite explicit that "It is an utter misconception of my argument to infer that I am an enemy to population. I am only an enemy to vice and misery, and consequently to that unfavourable proportion between population and food which produces these evils." This misconception was soon so widespread that the adjective 'Malthusian' was coined to describe not only the practice of restricting births but, in the last half of the nineteenth century, the practice of limiting economic production.

As a rule, socialists have poured scorn on Malthus for his liberal ideas and his skepticism with respect to social intervention policies. In attacking Godwin, he had undermined the very basis of utopian socialism. This resulted in a vigorous, if somewhat delayed, reaction from Godwin himself (Godwin, 1820), to which Malthus responded the following year in the *Edinburgh Review*.

This debate came to a head in 1839, when Marcus (1939) accused Malthus of advocating the asphyxiation of 'surplus' newborns, a myth which was taken up and popularized in France by Leroux (1839) with expressions such as 'the somber Protestant of sad England' and 'the selfish defender of the propertied classes.' Karl Marx, champion of invective, is every bit as dismissive, denouncing Malthus as 'superficial,' 'a professional plagiarist,' 'the author of nonsense,' 'the agent of the landed aristocracy,' 'a miserable sinner against science,' 'a paid advocate,' 'the principal enemy of the people,' etc. (cf. Michelle Perrot, *Malthusianism and Socialism* in Dupaquier and FauveChamoux 1983).

However, most socialists apart from Marx agree that there is a connection between overpopulation and misery, but distance themselves from Malthus with respect to the proposed causes and remedies. The most serious discussion of Malthus's theories is to be found in the work of Karl Kautsky (Kautsky, 1880). When neo-Malthusianism came to the fore at the end of the nineteenth century, however, socialist criticism once more became more radical. Where his doctrines are concerned, Malthus has had a large number of successors, many of them illegitimate – at least in the author's system of values.

The 'legitimate' successors to Malthus's works include the analysis of the causes, processes, and consequences of the growth of populations and, in political economy, the theory of effective demand, which Keynes revived (1939), affording it great topicality (*The General Theory of Employment, Interest and Money*), as well as the references made to the *Essay on the Principle of Population* by Charles Darwin (1838) and Alfred Russel Wallace (1858).

The illegitimate successors – the neo-Malthusianists – are much more numerous and more visible, having remained in the forefront.

As early as 1822, the Englishman Francis Place, while adopting Malthus's conceptual framework, proposed a much easier and more seductive way of checking population growth than moral restraint: the voluntary limitation of births within marriage (Place, 1822). Knowlton also advocated this approach in the United States (Knowlton, 1833), and the idea met with the approval of scholars such as Carlisle and Stuart Mill.

In 1860 the journalist Charles Bradlaugh set up the Malthusian League in London. The organization went from strength to strength from 1877 onwards, when Bradlaugh and Annie Besant, one of the pioneers of feminism, were prosecuted for obscenity.

Neo-Malthusianism was taken to France by the anarchist Paul Robin, who founded the League for Human Regeneration in 1896, and to Germany by the social democrat Alfred Bernstein, who in 1913 organized two meetings in Berlin, causing quite a stir. However, the neo-Malthusians came up against not only the hostility of the authorities but the distrust of socialist theorists.

At the time, militant feminism was evolving all over Europe, and birth control was just one element of its main objective: the sexual liberation of women. The decisive turn, however, was taken in the US thanks to Margaret Sanger, who in 1916 had founded a birth control clinic in Brooklyn. In 1925 she organized an international neo-Malthusian probirth control conference in New York, and in 1927 the first international congress on population was held. This was the basis of the International Union for the Scientific Study of Populations, which however almost immediately dissociated itself from the neo-Malthusian network.

See also: Civilization, Concept and History of; Historical Demography; Political Economy, History of; Population Dynamics: Classical Applications of Stable Population Theory; Population Dynamics: Momentum of Population Growth; Population Dynamics: Theory of Nonstable Populations; Population Dynamics: Theory of Stable Populations; Population Forecasts; Ricardo, David (1772–1823); Smith, Adam (1723–90).

Bibliography

Bonar, J., 1885/1924. Malthus and His Work. Macmillan, New York.
Condorcet (Caritat M-J-A), 1795/1970. Esquise d'un tableau historique des progrès de l'esprit humain. Vrin, Paris.
Celass, D.V. (Ed.), 1953. Introduction to Malthus. Wiley, New York.
Charbit, Y., 1981. Du Malthusianisme au populationisme, les économistes français et la population (1840–1870) INED. Travaux et documents, Paris.
Dupaquier, J., Fauve-Chamoux, A. (Eds.), 1983. Malthus Past and Present. Academic Press, London.
Eversley, DEC 1959. Social Theories of Fertility and the Malthusian Debate. Clarendon, Oxford, UK.
Fauve-Chamoux, A. (Ed.), 1984. Malthus hier et aujourd'hui. Sociètè de demographie historique. Paris.
Godwin, W., 1798. An Enquiry concerning Political Justice and its Influence on Moral and Happiness. University of Toronto Press, Toronto.
Godwin, W., 1820. Of Population: An Enquiry Concerning the Power of Increase in the Numbers of Mankind, being an Answer to Mr Malthus's Essay on that Subject. Longman, London.
Gordon, L., 1976. A Social History of Birth Control in America. In: Viking, Penguin, New York James, P. (Eds.), 1966 The Travel Diaries of Thomas Robert Malthus. Cambridge University Press, Cambridge.
James, P. (Ed.), 1979. Population Malthus, His Life and Times. Routledge, London.
Kautsky, K., 1880. Der Einfluβ der Volksvermehrung auf den Fortschritt der Gesellschaft. Block und Hasbach, Vienna.
Keynes, J.M., 1939. Robert Malthus: The First of the Cambridge Economists. Essays in Biography, New York.
Knowlton, C., 1833. The Fruits of Philosophy. Harcourt Brace, New York.
Marcus, 1939. The Book of Murder. Vademecum for the Comissioners and Guardians of the New Poor Law. With a Refutation of the Malthusian Doctrine. London.
Malthus, T.R., 1798. An Essay on the Principle of Population as it Affects the Future Improvement of Society, with Remarks on the Speculations of Mr. Godwin, M. Condocet and other Writers, London.
Malthus, T.R., 1803. An Essay on the Principle of Population, or a View of his Past and Present Effects on Human Happiness, with an Inquiry into our Prospects respecting the Future Removal or Mitigation of the Evils which it occasions. J. Johnson, London.
Malthus, T.R., 1820. Principles of Political Economy, Considered with View to their Practical Application. London.
Otter, W., 1836. Memoir of Robert Malthus. In: Malthus Principle of Political Economy. John Murray, London.
Petersen, W., 1979. Malthus Reconsidered. Harvard University Press, Cambridge, MA.
Place, F., 1822. Illustrations and Proofs of the Principle of Population. Himes, London.
Smith, K., 1951. The Malthusian Controversy. Routledge & Paul Kegan, London.
Stassart, J., 1957. Malthus et la population. Faculté de droit, Liège.

Managed Care and Psychotherapy in the United States

Linda D Weaver and Clarissa C Marques, Magellan Health Services, Columbia, MD, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 13, pp. 9162–9165, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The relationship between managed care and psychotherapists is a stormy one, characterized by conflicts about professional autonomy, accountability, and finances. Both managed care professionals and psychotherapists share a commitment to the provision of effective mental health treatment to those in need. But, there is disagreement regarding the methods best to keep high quality behavioral health services affordable and available. This article traces the evolution of managed behavioral care and psychotherapy from the 1970s to the present. Provider criticisms of managed care are addressed. A patient centric approach to managed behavioral healthcare is proposed in which providers and managed care work together to enhance and sustain positive treatment outcomes while eliminating unnecessary costs.

The impact of managed care on the delivery of mental health and substance abuse services has been significant. This article traces the evolution of managed behavioral care and psychotherapy from the 1970s to the present. Provider criticisms of managed care are addressed. A patient-centric approach to managed behavioral health care is proposed in which providers and managed care work as partners to enhance treatment outcomes and keep health care affordable.

The Evolution of Managed Behavioral Health Care

Managed behavioral care is the integration of the financing and delivery of mental health and substance abuse treatment within a system that seeks to manage the accessibility, cost and quality of that care (Marques, 1998: p. 411). The managed health care industry is the result of conflicting economic and political pressures to ensure Americans unlimited access to health care services, regardless of their ability to pay, in a market-based system of health care delivery.

The USA is unique among industrialized nations in its reliance on private employers to voluntarily provide health insurance for their employees (Iglehart, 1999: p. 6). Approximately 65% of Americans receive health insurance through the workplace, 66% of them through a managed care plan (*Facts About Managed Care* 1999). In all other industrialized nations, the government provides health care or requires private employers to provide coverage (Studdert et al., 1999). Fears that double-digit health inflation will cause employers to abandon the provision of health care coverage create a government interest in keeping health care premiums low.

Managing Care in the 1970s

Managed care came into prominence in the 1970s as a result of spiraling health care costs (Mullen, 1995: pp. 2–6). Rapid technological advancement in combination with an aging population created exponential increases in health care spending. Health insurance allowed patients to receive treatment without awareness of the costs, permitting practitioners to 'set their price' without consumer flight. Employers turned to managed care as a solution to escalating insurance premiums and fears of the inefficiency of a government-run health care system.

The 1970s approach to managed care focused on utilization review of hospital care. Utilization review is a process by which a managed care organization (MCO) evaluates whether recommended treatment is medically necessary and, as a result, reimbursable under a health insurance policy. This determination is typically conducted through a comparison of proposed treatment plans with medical necessity guidelines. At first, these guidelines were considered proprietary, creating a guessing game as to the definition of medical necessity. It was not until the early 1990s that Green Spring Health Services became the first managed behavioral health care company to make medical necessity criteria public, training providers on their use.

Managing Behavioral Health Care in the 1980s

As the rapid increase in physical health costs began to slow in the 1980s, mental health costs continued to rise, representing an increasing proportion of the health care dollar (Mihalik and Scherer, 1998: p. 1). Mental health benefits typically covered 28–45 days in a psychiatric hospital. This coverage encouraged overutilization of expensive inpatient services and an overbuilding of psychiatric hospitals. Psychiatric facilities competing to fill empty hospital beds advertised all-expense-paid 'vacations' to lush locations, respite from acting-out adolescents, and retreats for alcoholics with health insurance. One study concluded that 40% of psychiatric hospitalizations were medically unnecessary (Mechanic et al., 1995: pp. 19, 20).

Also in this decade, the medical cost offset of psychotherapy became widely known. Nick Cummings, through his work with Kaiser Permanente, demonstrated that brief, effective therapy could reduce the number of doctor visits of 'high-utilizers' by as much as 65% (Cummings, 1985). These 'worried well' sought treatment despite the absence of an organic cause for

symptoms. Resolution of physical complaints was successfully achieved through psychological treatment of emotional distress. Meta-analysis of other cost-offset studies found 10–33% reductions in hospitalization costs for physical illness following mental health treatment (Mumford et al., 1984: p. 1145). These cost reductions were attributed to the alleviation of the psychological distress associated with a traumatic or chronic illness.

The desire to prevent medically unnecessary psychiatric hospitalization and the potential for medical cost savings resulted in the creation of mental health 'health maintenance organizations' (HMOs). American Biodyne Centers, California Wellness Plan, and Psychology Management Systems were pioneers in the provision of unlimited outpatient mental health treatment designed to reduce inpatient psychiatric costs and overall medical expenses. As with most staff model HMOs, however, rapid access to routine mental health treatment was problematic. To increase appointment availability managed behavioral care incorporated a network model of service delivery similar to those used in medical managed care systems.

Managing Behavioral Health Care in the 1990s

Mental health 'carve-out' companies emerged in the 1990s. These carve-outs, called managed behavioral health organizations (MBHOs), specialize in managing mental health and substance abuse treatment. Health plans pay the MBHO a percentage of the health care premium for all members to manage the mental health utilization of the approximate 5–10% of members who use the benefit. The MBHO performs medical necessity review of the mental health treatment recommended by a network of independently contracted facilities, programs, and mental health practitioners. In 1999, 177 million Americans were contracted to receive mental health services managed by an MBHO (1999).

In the early 1990s the MBHO managed care through a combination of utilization review, discounted fee arrangements, and benefit management. Of mental health treatment dollars 65–80% were paid to inpatient and residential treatment facilities (Hennessy and Green-Hennessy, 1997: pp. 340, 341). Because 80% of mental health expenditures were associated with treatment for 20% of patients, early discharge planning, strategies to improve treatment compliance, and case coordination were managed care interventions for 'high-utilizers' of psychiatric services (Patterson, 1992: pp. 51, 54). The effectiveness of these techniques resulted in savings of 25–40%, at the same time increasing access to treatment for all members.

The late 1990s saw the brief emergence of 'integrated delivery systems.' This approach was based on the premise that practitioners paid to provide all of a member's care would be clinically and financially motivated to treat a patient in the least expensive, but most effective manner possible. Ideally, as the need for treatment intensity diminished, a patient could be 'stepped down' to less restrictive settings within the same facility, maximizing treatment coordination. Instead, costs increased and providers bankrupted. Patients previously treated at one level of care received treatment at all levels of care. Outpatient providers were unable to simultaneously provide and manage care, and still make a profit.

Given the price pressures of the industry, when delivery systems could not integrate, the MBHO industry did. Magellan Health Services acquired four MBHOs covering over 71 million members. FHC purchased Value Health Services and created ValueOptions, an MBHO covering 22 million members. United Healthcare purchased United Behavioral Health covering 20 million lives. This consolidation served as a catalyst for provider political organization and activism.

Criticisms of Managed Care

Most mental health providers oppose managed care. Providers contend that managed care organizations face an inherent conflict between quality and cost control. Well-publicized stories of personal tragedy resulting from bad utilization management decisions have created an environment in which managed care has gone from being the centerpiece of health care reform to its target. Provider criticisms of managed care fall into four categories: (a) increased nonreimbursable administrative time; (b) sacrificing of quality for cost savings; (c) intrusion on professional autonomy; and (d) a lack of MBHO accountability.

Increased Nonreimbursable Administrative Time

Provider dissatisfaction with managed care companies is often reflected in complaints of increased administrative time and decreased income. Many administrative requirements originate in private accreditation bodies' concerns for the quality of care under managed care, but require providers to complete additional paperwork. For example, the National Committee for Quality Assurance (NCQA) requires 'credentialing.' Credentialing begins with the practitioner's completion of a lengthy application and ends following a one- to three-month verification process. The practitioner's education and license are validated. Malpractice history, criminal record, and complaint data banks are reviewed for patterns of misconduct. Credentialing reduces the risk that members will receive treatment from clinicians unqualified to help them or potentially harm them.

Other administrative tasks are associated with ensuring that only medically necessary treatment is reimbursed. Precertification ensures that a member's benefit covers the services to be provided, and the medical necessity of the level and intensity of care proposed. Treatment plan review evaluates patient progress and ongoing necessity for treatment. These reviews also protect the patient from incurring surprise financial liability, and the practitioner from accruing 'bad debt.' Prior to managed care, 15% of provider revenue was written off due to an inability to collect (Luft, 1999: pp. 957, 964).

Sacrificing Quality for Cost Savings

Practitioners presume that a cost containment focus precludes concerns about treatment quality. Managed care payment mechanisms such as subcapitation, case rates, or withholds are criticized for encouraging undertreatment. Yet, most studies conclude that more people access mental health treatment

under HMO capitation arrangements than under fee-for-service arrangements. Although the treatment intensity is less, outcomes appear to be the same (Mechanic et al., 1995: p. 19). There is some suggestion that the seriously mentally ill do not fare as well in the HMO setting, but research design limitations prevent generalization.

Psychologists point to increased utilization of social workers as evidence of managed care's interest in savings over quality. Yet there are no studies to suggest psychologists' results are superior. In a survey conducted by Consumer Reports (Seligman, 1995: p. 965), patients reported equal satisfaction with outcomes whether psychologists, social workers, or psychiatrists provided treatment. Because there are only 70 000 psychologists in the USA, but over 189 000 social workers (Mills, 1997) who provide services at an average of $15 less per session, social work referrals improve treatment access.

The debate over managed care's impact on treatment quality is unlikely to be 'won' because treatment outcomes in mental health are difficult to define and measure (Marques et al., 1994: pp. 22–9). Variations in treatment practices, inconsistent use of diagnostic coding, the cyclical nature of mental health symptom exacerbation, and a lack of professional consensus regarding the definition of a positive mental health outcome make it difficult to determine what works and what doesn't. Patient satisfaction with treatment, although measurable, is not correlated with outcome. Patient self-reports of improvement are inconsistent with providers' perceptions of treatment completion. Despite these limitations, research has consistently determined that mental health treatment works and that managed care has not caused deterioration in these results. While patient and provider satisfaction has diminished, the quality of care has not.

Detracting from Practitioners' Professional Autonomy to Advocate for Patients

Providers resent the imposition on professional autonomy associated with medical necessity review. It is popular to complain of clerical oversight of professional judgment. Although it is possible for unlicensed individuals to authorize or approve recommended treatment according to a professionally developed treatment guideline, the industry standard is that a clinical peer or board-certified psychiatrist renders all nonauthorization decisions. Many states legislate peer review and both NCQA and the Utilization Review Accreditation Commission (URAC) require it.

Fears of 'economic credentialing' pressure clinicians to consider financial factors when recommending treatment, conflicting with professional ethics of patient advocacy. Providers are concerned that only those who treat individuals 'cheaply' will be retained in a managed care network, while providers who favor longer-term treatment, and potentially better outcomes, will be excluded. Yet few providers are removed from managed care networks for patterns of excessive overutilization. The preauthorization requirement often detects and prevents overutilization before it occurs. More common causes for provider termination are noncompliance with credentialing requirements, or provider requests for network removal. Access requirements make narrowing a managed care network solely on utilization unlikely.

Perceived Lack of Accountability

Managed care reform began to appear on state legislative agendas in 1995 and on the congressional agenda in 1996. Providers organized to fight managed care at the state level. Symbolically targeting 'gag clauses,' physicians began talking to the media, raising consumer anxieties about the quality of health care provided by profit-seeking managed care companies. Despite research to the contrary, the public became convinced through anecdote that managed care was destroying US health care.

Public concern about a perceived lack of managed care accountability has resulted in a proliferation of regulation in the last five years. All 50 states have passed some form of managed care legislation. Federal patient rights legislation has been introduced in Congress every year since 1997. NCQA introduced MBHO quality standards in 1996.

Mental health providers, frustrated with managed care's impact on their practices, are diversifying by providing consultation, education, and forensic services. 'Cash only' practices are emerging, with promises of enhanced privacy due to the absence of managed care oversight. For managed behavioral care to continue to evolve in the new millennium, it must repair its relationship with providers. A patient-centric model of managed behavioral care provides a framework in which practitioners and managed care can partner to achieve the shared goal of improving mental health in the USA.

The New Millennium: Patient-Centric Care Management

In the patient-centric model of managing behavioral care, an MBHO hires clinician-care managers who:

improve access to needed services;
match patients to providers who have the skills and experience most likely to result in treatment success;
maximize benefit allocation through medical necessity review of recommended services;
enhance treatment outcomes through intervention with ineffective treatment plans;
sustain treatment gains through aftercare monitoring; and
identify or prevent undetected mental illness.

The patient-centric model of managing care is based on the premise that helping patients overcome an emotional disturbance quickly and keeping them well is less expensive than allowing patients to prematurely abandon treatment and subsequently return to intensive levels of care. Promoting early access to effective providers is consistent with this goal.

In a patient-centric model, individual benefits are managed in the context of the well-being of an insured population. Ensuring that care is only reimbursed when medically necessary, that a patient is treated in an effective and less costly setting, and that patients are referred to practitioners with discounted rates maximizes the value of patients' benefits. Anticipation of benefit exhaustion and transition planning prevents individuals who cannot afford needed care from going without it.

Removal of barriers to mental health treatment is the ultimate value managed care can bring to the therapeutic process. Transportation problems, poorly coordinated care, or identification of contraindicated drug combinations are the clinician care manager's province, allowing the practitioner to focus on treatment interventions. By comparing recommended treatment to practice guidelines based on outcome research, managed care can promote the updating of providers' treatment repertoires. Following treatment, periodic monitoring of a patient's condition or follow-through with an aftercare plan can be effective in sustaining treatment gains. Joint pursuit of these patient-centric objectives permits the consumer to take advantage of treatment and care management resources, while reducing the cost of ineffective or unnecessary care.

Conclusion

In order to continue to reduce the cost of mental health care while maintaining or improving its quality, managed behavioral health care systems must continue to evolve and adapt. A managed care–provider partnership is unlikely to be easily achieved. Although mental health outcome research improved dramatically over the last 20 years of the twentieth century, there is still much professional disagreement about 'what works.' These debates are often the substance of discussion between providers and care managers. At the heart of the animosity, however, is the providers' perceived loss of professional dignity and autonomy. The next evolution in managed care will not occur unless the providers' sense of professionalism is restored. A patient-centric model of managed behavioral care aligns the goals of managed care employees, mental health providers, and purchaser-employers with the best interests of the patient population.

Editor's Note

Two books from the 1990s on managed psychotherapeutic care have recently been issued in new editions: (1) Alperin and Phillips (2013[1996]), and (2) Sauber (2013[1997]). The impact of managed care on the costs of mental health treatment, an issue discussed at length by Weaver and Marques, continue to be researched. For a recent statement, see Crane and Payne (2011). Disparities in access to mental health services under managed care are another current research theme; see Akincigil et al. (2012). Also useful: Olfson and Marcus (2010).

See also: Health Care: Legal Issues (United States).

Bibliography

Akincigil, A., Olfson, M., Siegel, M., Zurlo, K.A., Walkup, J.T., Crystal, S., 2012. Racial and ethnic disparities in depression care in community-dwelling elderly in the United States. American Journal of Public Health 102 (2), 319–328.

Alperin, R.M., Phillips, D.G. (Eds.), 2013(1996). The Impact of Managed Care on the Practice of Psychotherapy: Innovation, Implementation and Controversy. Routledge, London.

Cummings, N.A., Dorken, H., Pallak, M.S., 1993. The impact of psychological intervention on health care costs and utilization. In: Cummings, N.A., Pallak, M.S. (Eds.), Medicaid, Managed Behavioral Health and Implications for Public Policy, 2. Foundation for Behavioral Health, San Francisco, pp. 1–23.

Hennessy, K.D., Green-Hennessy, S., 1997. An economic and clinical rationale for changing utilization review practices for outpatient psychotherapy. Journal of Mental Health Administration 24 (3), 340–349.

Iglehart, J.K., 1999. Tumult and transformation. Health Affairs 18 (2), 6.

Luft, H.S., 1999. Why are physicians so upset about managed care? Journal of Health Politics Policy and Law 24 (5), 957–966.

Marques, C.C., 1998. The evolution of managed behavioral health care: a psychologist's perspective on the role of psychology. In: Bellack, A.S., Hersen, M. (Eds.), Comprehensive Clinical Psychology, 411, pp. 410–418.

Marques, C., Geraty, R., Harbin, H., Hoover, K., Theis, J., 1994. Quality and access in the managed behavioral healthcare industry. Behavioral Healthcare Tomorrow 3 (5), 22–29.

Mechanic, D., Schlesinger, M., McAlpine, D.D., 1995. Management of mental health and substance abuse services: State of the art and early results. Milbank Quarterly 73 (1).

Mihalik, G., Scherer, M., 1998. Fundamental mechanisms of managed behavioral health care. Journal of Health Care Finance 24 (3), 1–15.

Mills, M., September 1997. US has 114 behavioral health professionals per 100,000 population. Open Minds Newsletter, 6.

Mullen, J.K., 1995. Intro to Managed Care.

Mumford, E., Schlesinger, J.H., Glass, V.G., Patrick, C., Cuerdon, T., 1984. A new look at evidence about reduced cost of medical utilization following mental health treatment. American Journal of Psychiatry 141 (10), 1145–1158.

Olfson, M., Marcus, S.C., 2010. National trends in outpatient psychotherapy. American Journal of Psychiatry 167 (12), 1456–1463.

Patterson, D.Y., Sharfstein, S.S., 1992. The future of mental health care. In: Feldman, J., Fitzpatrick, R.J. (Eds.), Managed Mental Health Care: Administrative and Clinical Issues. American Psychiatric Press, Washington, DC, pp. 335–346.

Richard, S.S. (Ed.), 2013(1997). Managed mental health care: Major diagnostic and treatment approaches. Brunner/Mazel, Bristol, PA.

Russell, C.D., Payne, S.H., 2011. Individual versus family psychotherapy in managed care: Comparing the costs of treatment by the mental health professions. Journal of Marital and Family Therapy 37 (3), 273–289.

Seligman, M.E., 1995. The effectiveness of psychotherapy: The consumer reports study. American Psychologist 50 (12), 965–974.

Studdert, D.M., Sage, W.M., Gresenz, C.R., Carole, R., Hensler, D.R., 1999. Expanded managed care liability: What impact on employer coverage? Health Affairs 8, 7–27.

Satcher, D., vs. Surgeon General, 1999. Financing and managing mental health care. In: Mental Health: A Report of the Surgeon General. Department of Health and Human Services, Washington, DC.

Management Accounting

Alnoor Bhimani, London School of Economics, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The modern-day management-accounting literature originates from late nineteenth-century factory cost-accounting writings. Over the past 20 years, the field has undergone an important transformation as a result of localized as well as global economic and organizational management forces. These forces of change are considered, and important areas of modern management accounting are reviewed. Two management-accounting approaches (target costing and activity accounting) are discussed at greater length. Concluding thoughts are presented on what some scholars and thinkers believe will be the fate of management accounting in the twenty-first century.

Introduction

The Chartered Institute of Management Accountants in the United Kingdom defines management accounting as an integral part of management concerned with identifying, generating, presenting, and interpreting information used for

1. formulating strategy,
2. planning and controlling activities,
3. decision taking,
4. efficient resource usage,
5. performance improvement and value enhancement,
6. corporate governance and internal control,
7. safeguarding tangible and intangible assets.

While the discipline finds its roots in industrial accounting concerns and factory-based cost accounting, it has undergone an important transformation over the past 20 years. The field presently continues to undergo significant changes and remains in a dynamic state of flux. During the 1980s, many organizations in Western countries perceived a need to alter their management practices. A variety of factors were responsible for this. In part, at the international level, competitive manufacturing methods used by major Japanese corporations and by those of newly industrialized countries were seen as superior to those used in the West (Bhimani and Bromwich, 1996). To an extent, this was ascribed to sociocultural characteristics of Japanese workers from factory-floor operators to senior executives. But what was viewed as more significant was the application of flexible manufacturing technologies by the Japanese to produce a greater diversity of high-quality, low-priced products more quickly. Specific modes of work organization, production systems, and motivational incentives were seen to differ radically from those used in the West. These forces of change on management approaches have also been viewed as having had important implications for management accounting practices. The adoption of flexible technologies in manufacturing and service-oriented enterprises, including manufacturing and enterprise resources planning systems, and features of computer-integrated manufacturing, led to conventional accounting methods for cost allocation being questioned. The manufacturing environment had become more automated, more specialized, more flexible, and requiring larger capital investments, whereas methods of costing products continued to operate under the conventional notion that company operations were geared toward repetitive production of homogeneous products relying on labor-based activities. Management-accounting techniques were thus seen to be antagonistic to the realities of the new manufacturing environment. Similar critiques were being made of cost-accounting methods in service industries.

Other calls for change in management accounting were also being voiced. North American academics thought that from the post–First World War period, progress in management accounting had virtually ceased, principally as a result of the growing importance of the external reporting priorities, especially in the area of stock valuation over those of internal organizational accounting. Some critics of traditional management accounting asserted that management accounting had, for over 70 years, been subordinated to the requirements of financial-statement preparation for external parties (Johnson and Kaplan, 1987). Advocates of this school have supported calls for management-accounting reforms based on the argument that the field must once again be allowed to develop in its own right, rather than remain subordinate to financial accounting.

Many commentators made more substantive arguments, focusing on the propriety of accounting practices, given the roles that management accounting is intended to fulfill. The usefulness of complex methods of organizational costing and pricing, in the face of the belief that the price of a product must ultimately meet the expectations of the marketplace rather than those of its producers, has been questioned. Others have given weight to the argument that management accounting has for too long remained isolated and divorced from other enterprise functions. Conventional accounting practices have put into place channels for information flow and routes for data exchange that have rendered some organizations static, inflexible, and excessively structured, especially in the case of firms operating in dynamic and fast-changing markets. Calls have thereby been made for cost-management tools appealing to wider rationales of enterprise management.

Onward with Change

During the 1990s, many organizations pursued a vigorous management change agenda. Until 1994, many organizations shied away from investing in novel enterprise management

 http://dx.doi.org/10.1016/B978-0-08-097086-8.73013-3

philosophies because of the malaise experienced in the worldwide economy. Nevertheless, cost-containment programs were widely implemented. Techniques such as activity-based costing (ABC), throughput accounting, target cost management, life-cycle costing, and the greater use of nonfinancial performance measures were seen as offering the potential of enhancing value creation and aiding cost containment (Bromwich and Bhimani, 1994; Yoshikawa et al., 1993). From 1995, as the economic engine of major Western economies began to generate more steam, organizational searches for wider reaching management technologies grew apace once more. Such pursuits were, however, tempered with a more cautious approach to large-expenditure projects on altering ways of managing. Moreover, accounting techniques for aiding managerial tasks appeared more cognizant of a need to interlink enterprises' various value-creation activities. Thus, management-accounting technologies stressed processes, as opposed to functions, and emphasized linkages between traditional activities, as well as more flexible organizational structures.

Notions of the impact of brand management on cost accountings were highlighted. Attempts to link organizational learning to customer responsivity and financial tracking were made via establishing novel performance-measure control approaches. Traditional conceptions of cost-accounting techniques for scorekeeping purposes were shunned, whereas achieving more balanced scores came to be deemed important (Kaplan and Norton, 1996). Likewise, 'value added' was a significant, although short-lived endeavor among proponents of more comprehensive financial reporting in the 1970s, whereas the pursuit of a trademarked notion of economic value added surfaced (Stern et al., 1995). Certainly, the need to be strategically perceptive in all things management accounting was widely perceived during the mid-1990s. The US-based Institute of Management Accountants changed the name of its flagship magazine *Management Accounting* to *Strategic Finance*. Likewise, the Canadian Society of Management Accountants revamped its intellectual capital whereby its published products came to be grouped as a 'strategic management series' package. No doubt many organizations, in both Europe and North America, consider themselves as engaging and supporting strategic management-accounting activities while not appealing to any single definition of the term (Bhimani and Keshtvarz, 1999; Gulding et al., 2000).

Management Accounting and Digital Technologies

Although computers have shaped business activities for over four decades, it was only during the mid-1990s that a profound second-wave digital revolution took place. Three forces became intertwined at this time:

- the spread of user friendly operating systems and interfaces,
- the rapid diffusion of the Internet and the World Wide Web, and
- the convergence of four formerly distinct industries – computers, software, communications, and media and entertainment.

These three factors led to the creation of a huge worldwide value network with attendant new business models and novel ways for enterprise architectures to form and enable the generation of wealth creation. Concurrently, business intelligence started to crowd out different sources of information.

What consumers and decision makers do with information can insightfully shed light on what drives their purchases or actions. So, traditionally, enterprises have designed information systems to produce formal information that system users deploy based on customer purchases and related economic activities. However, a shift in information design structures is currently taking place. There is now a realization that what information users discard can be equally if not more useful as a source of business intelligence for companies. A customer buying online will often leave a trail of information disclosed prior to making the purchase. This 'data exhaust,' if effectively analyzed, can help organizations determine how and what individuals rationalize and the path they take before they arrive at making a purchase or a decision. Google.com learns from every search process carried out. Amazon.com gathers information from online customer behavior irrespective of whether a purchase is made. eBay monitors buyer and seller activities even where no economic transactions take place. These companies analyze data exhaust in the provision of important financial intelligence, which can shape cost management, pricing decisions, and operational controls.

Drawing business intelligence from information not just on transactions but also on searching behavior, Website surfing, and browsing experience has, for Web-enabled companies, become a basis for competitive engagement. In the future, this will increase in significance. As 'transaction agnostic' information grows, data volume and information management will expand and place important challenges on financial systems.

Globally produced data grows 10-fold every 5 years. The pace is partly reflective of the increase in media, entertainment, and social networking possibilities online. And while much of the information is unstructured, a significant proportion is amenable to structuring in an economically purposeful sense. This growth of information alongside the possibilities presented by data exhaust enables useful analysis and assessment by businesses.

For instance, purchases made via Amazon are often tied to purchases made previously. Probabilities can be established about the likelihood of particular subsequent purchases being made based on data collected about macrolevel buying behavior, nonpurchases, and prior online interactive searches. eBay likewise continuously alters its listings based on prior listing activity, bidding behavior, pricing trends, search terms used, and purchase frequency. Google searches become continuously more relevant because search results are based on what users with prior similar searches eventually stayed with. In this sense, data trails left by surfers enable data mining, which can be deployed to directly impact pricing policy, cost-containment prioritization, and cash and working capital management activities.

While data mining increasingly shapes financial control and management possibilities, traditionally management

accounting systems have been influenced by information collection about products that trigger economic changes.

There is, however, a growing trend toward the collection of information from within products and processes themselves. For instance, real-time data collection is possible without the need for separate information systems, through use of

- pill-shaped microcameras,
- precision agricultural and industrial sensors, and
- radio frequencies identification tags.

Thus, it is possible to achieve continuous information collection on possible product purchase pathways as well as cost changes, dynamic pricing options, and working capital management choices, when products and processes themselves collect and transform information about their own behavior (see Bhimani and Bromwich, 2009; Bhimani, 2012).

This presents novel possibilities for management accounting where it can become a receptor and assessor of information for real-time analysis and decisions.

Strides in Management Accounting

Perhaps it is well to turn to some significant technologies of management accounting that emerged over the past two decades and are now in widespread use within many enterprises globally. Two of these provide the focus presently.

Target Costing

Target costing indicates for each product (and each cost element of that product) the cost reductions (over time) that must be achieved by the firm and its suppliers in order to ensure that the company's strategy is realized throughout the product's life cycle, while maintaining a competitive market price for the product offered. A preliminary step in this process is for the enterprise to come to a view over the whole product life cycle as to the product strategy in terms of the product attributes to be offered relative to competitors, including further options, quality and after sales services, the technology to be used, the desired profit, and market share. A comparison with expected prices in the market over the life cycle of the product determines the price at which the product with the planned characteristics can be offered in the market and be expected to obtain their target profit and market growth.

Target costing in some Japanese companies is seen as a cost-reduction exercise that seeks to achieve cost reduction in the preproduction stages. It seeks to force designers to reduce costs, although how far target prices should seek to include strategy and how far it should be seen predominantly as a continuing cost-reduction device is a firm-specific issue. Such an approach may be supplemented by levying extra costs on nonstandard parts and excessive tooling, as well as penalizing the use of nonstandard parts or new machines. The procedure contains a number of other lessons. One is that prices are determined in the market and reflect competitors' strategies and customers' demands. Target costing also requires that companies collect and use external information that may not be available or, if available, may not ordinarily be assessed by the Finance department. Finally, it illustrates the possibility of incorporating strategy into accounting and performance measurement more generally.

There are a number of ways of setting target costs; these include

1. determining the difference between allowable cost (selling prices – desired profit) and forecast costs determined using current costs,
2. determining the difference between allowable cost (using planning selling price net of the required return on sales) and current costs,
3. applying the desired cost reduction to current costs.

The above description of target costing should be seen as broadly representative of what ought to be done rather than the organizational reality. There is evidence that Japanese firms often fail to make the target cost, at least prior to manufacture, and either adjust the target costs or decide not to continue with the product in its existing form. Target costing practices in Western enterprise likewise evidence a high degree of organizational specificity.

Tools for Target Costing

In Japan, value engineering (VE) is an activity that helps to design products that meet customer needs at the lowest cost while assuring the required standard of quality and reliability. A number of tools are generally used in this activity.

Functional Analysis

The aim is to determine the functions of the product and indicate how far the design of these functions will satisfy the consumer (as shown by market research) and at what cost relative to that part of the target cost pricing by elements allocated to that function. This allows the variance between the target cost-per-function and its current cost to be calculated. It also allows the contribution that a function makes to consumer needs to be ascertained relative to its cost, thereby leading to redesign where two items are not well matched (see Yoshikawa et al., 1993).

Cost Tables

Cost tables are detailed simulated databases of all the costs of producing a product, either for product functions or product components, based on a large variety of assumptions. They allow 'what if' questions to be posed relating to decisions to change any aspect of the product. A cost table will show the variety of cost performances when using different types of materials for a given component. The data in cost tables will be accepted by all members of the organization, and therefore ensure common thinking throughout. There will be a cost table for each part of the product's life cycle.

Value Engineering Collections

These are libraries of case studies of the detailed experience from previous VE activity, the use of which is to improve future target costing and VE. Of course, all the other well-known techniques used in Japan, such as just-in-time methods and quality circles, are utilized, where appropriate, in target-costing exercises.

Enterprise cost structure in Japan is such that, on average, material and components account for some 60% of cost, direct

labor some 14%, and overheads some 24% (Bhimani and Bromwich, 1996), although these percentages can vary substantially across firms and industries. Comparative US figures are material and component costs (53%), direct labor (15%), and overheads (32%). Possibly, high overhead cost incursions could be a signal that for some firms much may be gained from the market-derived discipline that is evident in the target-cost-management process.

Activity Costing

Concerns about overhead allocation have been voiced by many managers and, perhaps more importantly, the need to account in a way that encompasses the activities that cause costs rather than using traditional methods of overhead allocation is recognized by professional accounting bodies.

ABC emerged as a cost-management approach during the late 1980s. It can be defined as an approach to costing that focuses on activities as the fundamental cost objects. It uses the cost of these activities as the basis for assigning costs to other cost objects such as products, services, or customers (Bhimani et al., 2012). During the 1990s, it has increasingly been linked to managerial decisions via time, market, strategy, and quality issues. Examples of ABC calculation can be found in Bhimani (2012).

The published evidence on the payoffs and costs of installing activity-based cost and management information systems is not unequivocal (Bromwich and Bhimani, 1994). Partly, this is because seeking an association between major changes in organizational and management structures and management accounting is a difficult and time-consuming exercise. The practicing management accountant must be satisfied that the benefits likely to accrue from any change will outweigh its costs, although cost-benefit quantification is not a straight forward proposition. Moreover, abstract, theoretical concepts do not tend to succeed in achieving implementation in organizations. To date, there is little evidence that any new accounting method that attempts to facilitate the management decision-making process actually increases the ability to generate greater profits, although providing such evidence is very difficult. Ascribing increased profitability to the implementation of a costing technique, such as ABC, is a difficult exercise. Problems arise in tracing changes in the 'bottom line' to managerial action associated with altered costing information. The inability to isolate the effects of a single change such as the adoption of ABC *ceteris paribus*, and to use another company as a 'control' for comparative purposes, makes any assertion of a profitability link dubious. This is especially so because changes in organizations often come in bundles, after, for example, the appointment of a new managerial team. Perhaps the ultimate test of establishing whether a switch from a traditional cost system to ABC yields increased benefits is a market test that seeks to ascertain whether such a switch becomes a permanent one for a sufficiently large number of companies adopting ABC. Given the relative novelty of this costing approach, however, this test is difficult to apply. Nevertheless, in various surveys of ABC/ABM implementations in UK firms, Innes et al. (2000) report that, in 1995, 13.2% of companies rejected the technique, whereas 15.3% did so in 1999. They conclude:

> All in all, the survey results are not indicative of any further growth in ABC/M but rather suggest at best a leveling off in the extent of its use in our largest companies.

There are many managers who believe that activity-based accounting methods have changed their approach to decision making, indirectly yielding increased profits and, more importantly, a much better understanding of their business (see Cooper and Kaplan, 1998; for case studies). Some problems still exist with overhead allocation where, for instance, the source of the overhead is difficult to trace or is sunk or is common to diverse activities. Arbitrariness and judgment will, therefore, continue to significantly affect overhead cost accounting. There is strong support, however, for the argument that dynamic and changing manufacturing (and nonmanufacturing) environments render redundant, conventional cost-accounting systems that emphasize information priorities other than those emerging as virtually indispensable in today's competitive marketplace. For example, strategic cost analysis, cost-benefit appraisal for long-term projects, life-cycle consideration, quality, and customer satisfaction data are likely to become increasingly meaningful in service and manufacturing environments (Bromwich and Bhimani, 1994).

The Future of Management Accounting

While many other management-accounting approaches have been, and continue to be, developed, aside from activity accounting and target costing, some concerns have been voiced concerning the future fate of management-accounting practices. King (1997), for instance, believes that many management accountants complain that – while their skills may be valued – they are often excluded from management decision making. Accountants are often not seen as starting players on the management team, in part because of their training, which emphasizes precision and accuracy at the expense of relevance. King (1997) suggests that recognizing that there is more judgment – and less precision – in accounting is a first step in what may have to be an extensive program to ensure that accountants will survive into the twenty-first century.

Foster (1996), likewise, believes that change is required in the way in which management accountants operate. Management accounting is in a continual state of adaptation as the knowledge base in the field increases and the business environment changes. According to Foster (1996), management accounting has often been portrayed as focusing too much on internal reporting and on providing information for statutory financial reporting. This portrayal is seen as inadequate. In the future, decisions about which activities to outsource, and how to structure joint ventures with other organizations will be critical management responsibilities. Management will seek financial information

to help plan and monitor ongoing relationships with external partners. Foster (1996) notes that, in the future, one will seek greater recognition of root-cause cost drivers as well as cost drivers arising within and across functions.

According to Flegm (1996), skill with numbers is not enough. Accountants of the future must be experts in the basic disciplines of business, yet also generalists who can manage different disciplines, communicate with clients, and motivate employees. The call for more strategic management accountants, likewise, still rings loud. Kaplan and Norton (1996, 2000) consider that advances since the 1990s have made it possible for management accountants to become part of their organization's value-added team. Management accountants must participate in the formulation and implementation of strategy (Kaplan, 1995), and help translate strategic intent and capabilities into operational and managerial measures.

Management accountants have to move away from being scorekeepers of the past and to become the designers of the organization's critical management information systems. Existing performance measurement systems, including those based on activity accounting, focus on improving existing processes. The balanced scorecard, according to Kaplan (1995), by contrast, focuses on what new processes are needed to achieve breakthrough performance objectives for customers and shareholders.

This call has been voiced in parallel to that for altering management accountants' patterns of behavior (Johnson, 1995). In the past, management accountants focused on how to disclose and report financial accounting numbers, especially cost numbers, so that managers could draw from them more useful information. According to Johnson (1995), instead of focusing on achieving targets, companies' intent on achieving positive long-term results should focus on mastering disciplined and standardized patterns of behavior. Management accountants should work to create channels through which people inquire openly about purpose and method. In this light, Borthick and Roth (1997) note that the management-accounting systems of the future will be a large, real-time database whose data managers anywhere will be able to retrieve, manipulate, and analyze for new problems as they arise. The choice that companies face is whether to be startled by what the future brings or to embrace the future by implementing organization-wide database systems to reap the benefits of better data access.

What seems evident is that each decade brings with it pressures for changing management accounting, as well as different perceptions of the implications of such pressures. The late 1990s heralded in the New Economy where past conceptions of proper organizational management are viewed as being in need of change. A renaissance of liberalization from the past is deemed essential by many old industrial firms facing competition by new giants that did not exist only 5 years before. The Internet economy has made 'modern' management techniques ancestral. Bricks and mortar cardinal principles are being eschewed and replaced by evolving bricks and mortar truths. Life-cycle coasters are now to move at an Internet time pace. Balanced scorecard adopters must tackle new forms of imbalance. Activity accounting systems must begin to appeal to cyber-based cost drivers. Target cost management must now capture the nuances of ephemeral Web markets. While little can be said of where – if anywhere – the future will take management accounting, pundits will continue to dwell on what is essential for its well-being. The finance function must today address novel ways of appealing to information. It must address altered platforms, sources, and architectures relating to information output, channels, structures, and modes of assessment. The sources of change are as multifaceted as are the effects. The global economy is affected by macro- and microlevel changes. Governmental actions are influencing the manner in which enterprises access finance, make use of it, report it, and monitor the risks of their economic activities.

See also: Information and Knowledge, Organizational; Innovation; Management: General; Organizational Decision Making; Supply Chain Management; Technology and Organization.

Bibliography

Bhimani, A., 2012. Strategic Finance. Strategy Press, London, UK.
Bhimani, A., Bromwich, M., 2009. Management Accounting: Retrospect and Prospect. Elsevier/CIMA, London.
Bhimani, A., Horngren, C., Datar, S., Rajan, M., 2012. Management and Cost Accounting, fifth ed. FT/Pearson, Hemel Hempstead, UK.
Bhimani, A., Keshtvarz, M., 1999. British management accountants: strategically inclined? Journal of Cost Management 13, 25–31.
Borthick, A.F., Roth, H.P., 1997. Faster access to more information for better decisions. Journal of Cost Management 11, 25–30.
Bromwich, M., Bhimani, A., 1994. Management Accounting: Pathways to Progress. CIMA, London.
Cooper, R., Kaplan, R.S., 1998. The Design of Cost Management Systems. Prentice-Hall, Englewood Cliffs, NJ.
Flegm, E.H., 1996. The future of management and financial accounting. Journal of Cost Management 10, 44–49.
Foster, G., 1996. Management accounting in 2000. Journal of Cost Management 10, 36–39.
Gulding, C., Cravens, K.S., Tayles, M., 2000. An international comparison of strategic management accounting practices. Management Accounting Research 11, 113–167.
Innes, J., Mitchell, F., Sinclair, D., 2000. A tale of two surveys. Research Update: The Newsletter of CIMA Research Spring, 4.
Johnson, H.T., 1995. Management accounting in the 21st century. Journal of Cost Management 9, 15–20.
Johnson, H.T., Kaplan, R.S., 1987. Relevance Lost: The Rise and Fall of Management Accounting. Harvard Business School Press, Boston.
Kaplan, K.S., 1995. New roles for management accountants. Journal of Cost Management 9, 6–14.
Kaplan, R.S., Norton, D.P., 1996. The Balanced Scorecard. Harvard Business School Press, Boston.
Kaplan, R.S., Norton, D.P., 2000. The Strategy-focused Organisation: How Balanced Scorecards Thrive in the New Business Environment. Harvard Business School Press, Boston.
Kato, Y., 1993. Target costing support systems: lessons from leading Japanese companies. Management Accounting Research 4, 33–48.
King, A.M., 1997. Three significant digits. Journal of Cost Management 11, 31–37.
Stern, J., Stewart, G., Chew, D., 1995. The EVA® financial management system. Journal of Applied Corporate Finance Summer, 32–46.
Yoshikawa, T., Innes, J., Mitchell, F., Tanaka, M., 1993. Contemporary Cost Management. Chapman and Hall, London.

Management: General

Alberto Martinelli, University of Milan, Milan, Italy

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 13, pp. 9170 – 9178, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Management describes the complex behavior of those responsible for the decisions that determine the allocation of resources, both human and physical, within an organization. The role of general managers is a multifunctional role: decision-maker, strategic leader, human relations expert, custodian of institutionalized values, mediator between business and society, negotiator and power broker. Managers act in a set of institutional contexts and display specific cultural attitudes and styles of action. According to the simplified model of firm behavior under perfect competition, the managerial function is limited to a set of highly rigid decision rules for determining price and output. In real market situations which deviate from the perfectly competitive, market forces are only one of the variables in the complex context where business management takes place. Other key variables concern the corporate organization, the larger group of organizations in which the firm is embedded, state policies and legal systems, and the stategies of both internal and external stakeholders of the firm. The general management function is therefore a very complex one, that includes a set of managerial roles and implies various forms of managerial behavior and style. The focus of this article is on general management, on managerial behavior and managerial roles in organizations. Four main questions are discussed: (a) the rise of managerial capitalism and the separation of ownership and control; (b) what managers really do, i.e., the main features of managerial behavior; (c) the role of managers in business organizations; (d) managerial ideologies and managerial styles.

Management describes the complex behavior of those responsible for the decisions that determine the allocation of resources, both human and physical, within an organization. The role of general managers is a multifunctional role: decision maker, strategic leader, and human relations expert, custodian of institutionalized values, mediator between business and society, negotiator and power broker. It is played in a set of institutional contexts and implies specific cultural attitudes and styles of action. According to the simplified model of firm behavior under perfect competition, the managerial function is limited to a set of highly rigid decision rules for determining price and output. In real market situations which deviate from the perfectly competitive, market forces are only one of the variables in the complex context where business management takes place. Other key variables concern the corporate organization, the larger group of organizations in which the firm is embedded, state policies and legal systems, and the stategies of both internal and external stakeholders of the firm. The general management function is therefore a very complex one, that includes a set of managerial roles and implies various forms of managerial behavior and style.

For specific areas of management see articles on Administration in Organizations; Corporate Finance: Financial Control. The focus of this article is on general management, on managerial behavior and managerial roles in organizations. More specifically, four main questions will be discussed: (a) the rise of managerial capitalism and the separation of ownership and control; (b) what managers really do: the research on managerial behavior; (c) the role of managers in business organizations; (d) managerial ideologies and managerial styles.

The Rise of Managerial Capitalism and the Separation of Ownership and Control

The growing importance of management in the twentieth century is related to the transformation of capitalism and to the emergence of large business organizations, in the form of public companies, that implies the separation of ownership and control. This transformation has been due to a variety of factors, such as technological innovation, generalization of market relations to preindustrial settings, international competition, industrial concentration, and vertical integration. The change has resulted in increasing organizational complexity, multifunctional structures, greater role differentiation, and more decentralized authority.

The transition from family capitalism to managerial capitalism was facilitated by the legal institution of the limited company; it was related to the timing and sequence of the industrialization process in different countries and took place earlier in the UK and the USA than in Continental Europe and Japan. Shareholding has often become very dispersed. Boards of directors have been predominantly composed of fulltime senior executives who nominate new members and take decisions concerning high management appointments and remuneration. Top management has become effectively responsible not only for operating organizations, but also for strategic decisions concerning investment, finance, internal growth and diversification, profit retentions, and acquisitions. Business firms have displayed a capacity for persistent, although irregular, long-term growth in size, as shown by the growing share of the 100 largest corporations in total industrial value-added, which from less than 10 percent in 1900 grew to 30 – 50 percent in various industries in the year 2000.

The growth of the large corporation and the increase in capital concentration, however, are not inconsistent with coexistence of other components of the modern economy, where traditional capitalist modes (markets containing large numbers of small owner-managed businesses), and new forms of productive systems (industrial districts with specific competitive-cooperative patterns among producers in the same industry) not only survive, but effectively compete and grow. In addition, in the last two decades of the twentieth

century, technological developments, such as personal computers, which increase the comparative efficiency of small organizational units, and new institutional developments, such as corporate divestitures and management buyouts, contribute to checking and even reversing concentrating tendencies.

The large corporation and managerial capitalism are neither a general pattern of business organization, nor necessary steps in economic modernization, but they have marked first the economic history of the UK and the USA and later, that of continental Europe and Japan. The 'complete businessperson' – who was at the same time owner, entrepreneur, and manager – slowly gave place to the separation of the property and control functions and to a plurality of specialized roles within the corporation.

Both Marx and Schumpeter saw the separation of the managerial role from the ownership of capital as a sign of the crisis of capitalism. The limitations of both Marx's and Schumpeter's views in this regard is due largely to their belief that the competitive economy of the individual entrepreneur is the only brand of sustainable capitalism. In reality, capitalism has proved compatible with the existence of large, complex firms, and with state intervention and regulation of the economy. What Marx and Schumpeter saw as a crisis factor, other scholars viewed as a stage in capitalist development and a feature of the fundamental process of the growth of the business firm. Influenced by Weber's analysis of the rise of the modern state, Rathenau (1918) argued that with the depersonalization of property and the rise of *'organisierte Kapitalismus*,' the firm followed the same path as, and became similar to, the modern state.

Berle and Means (1932), in their pathbreaking empirical research on two hundred US firms, argued that the separation of ownership and control was an organizational requirement of the modern corporation and an irreversible trend of economic life. Berle and Means conceived the modern corporation as an organized social group based on the interdependence of different economic interests – those of the owners, the employees, the consumers, and the controlling group. This last group has the 'active property' of the firm – as distinct from the 'passive property' of the shareholders – and performs the entrepreneurial function. Managerial power is highly concentrated in the controlling group, as religious power was in the medieval church and political power in the nation state. Conflicts of interest can arise between managers- entrepreneurs and shareholders; the former are mostly interested in reinvesting most of the profit, in order to enhance the firm's competitive power and long-term development, while the latter want short-term economic gains. A more radical view of this conflict was Burnham's (1941), who foresaw a 'managerial revolution.'

Ideas similar to those of Berle and Means's were developed by Coase (1937), Williamson (1975), and Chandler (1962, 1977), who focused on the firm as an alternative allocative mechanism to the price mechanism, and of the 'visible hand' of management as an alternative to the 'invisible hand' of the market. Chandler, in particular, argued that a fundamental reason for the success of the modern corporation was its ability to obtain higher productivity rates by internalizing activities previously performed by separate autonomous business units, by reducing information and transactions costs, and by better coordinating the flow of goods and factors of production. But the advantages of corporate organization could not be fully exploited until a managerial hierarchy was created, made of managers who are rewarded, selected, and promoted on the basis of their professional skills and performances rather than on the basis of their family ties and the money they have invested in the firm. Once formed, this managerial hierarchy becomes the source of continuity, power, and growth in the large firm, and tends to assume a life of its own, independent from the individual turnover at the top.

According to Chandler, the dividing line between entrepreneur and manager is drawn on the basis of the type of decisions they make: the minority of the general executives and owner-entrepreneurs – who are responsible for strategic decisions – are neatly separated from the majority of sectoral managers – who are in charge of specific processes and perform routine activities. Managerial roles in the modern corporation are located along a hierarchy of authority and responsibility. The control of finance, personnel, organization, research, and information is a necessary condition, but is not enough to define a full entrepreneurial role; the role implies also the ability to work out long-term general strategies and the power to allocate the resources necessary to achieve the desired goals (Mathews, 2010).

The thesis of the irreversible diffusion of managerial capitalism has been criticized on various grounds. First, it has been pointed out that managerial control applies only to large firms, whereas, as I remarked earlier, the 'other half' of the contemporary economy is made of small firms, most of which are managed by owner-entrepreneurs. Second, even for large firms, family capitalism is much more resilient than the scholars of managerial control would admit. This seems to be the case not only in Continental Europe and Japan but also in the USA, as Burch (1972) tried to show through the survey he made 40 years later of the same firms that Berle and Means studied.

Third, it has been argued that conflictual views on short- and long-term strategies actually vary according to the conception of corporate control developed by high-level executives. The finance conception of control that was pioneered by US conglomerates in the 1980s provides an instance of short-run gain strategy upheld by top managers (Fligstein, 1990). There are many instances of top managers and owners basically sharing the same interests and values, in spite of the different roles they play, due to the fact that corporate executives, while not often among the largest shareholders, receive incomes that are highly correlated with stock performance, and that a relevant part of their remuneration is the form of stock, dividend income, and capital gains (Lewellen, 1971).

Fourth, scholars like Galbraith (1967) argue that leadership in the large business firm is no longer confined to entrepreneurs, or to top managers; but it is exerted by a technostructure, that is, by all those participants who have specialized knowledge and participate in collective decision making. Technologically advanced and organizationally diversified corporations cannot be governed by single entrepreneurs, or a few top executives; they must be managed by the larger technostructure, or by Mey- naud's (1964) technocrats.

Fifth, the idea of the firm as an alternative allocative mechanism for the price mechanism has been criticized for

neglecting other important mechanisms in the regulation of economic activity, such as clans, cartels, and trusts, employers, and trade associations (Streeck and Schmitter, 1985).

Finally, the very notion of corporate control has been challenged by the studies of interlocking directorates. Mizruchi (1982) and Stockman et al. (1983), among others, argue that public companies, with ownership dispersed among many stockholders, are actually controlled by financial institutions—such as national commercial banks and insurance companies—and through corporate interaction patterns rather than by their top managers. Mintz and Schwartz (1985) found a complex division of labor among financial and nonfinancial institutions. Distinguishing among New York money market commercial banks, regional banks, and national insurance companies, they portray a system where financial institutions simultaneously unite and divide. The densely interlocked groups at the local level – centered around the largest regional banks – are connected to the larger money market banks by a series of bridging links created by the country's largest insurance companies. Control of capital flows enables banks to constrain the actions of nonfinancial corporations and to pursue to some extent centralized decision making (Davis, 2009).

A related view is that of Useem (1984), who views the formation of an intercorporate management network as the driving force in the rise of institutional capitalism, that is, the classwide organization of business. The dispersion of ownership and the superior capacity of managerial hierarchies in large firms were the engine behind the transition from family to managerial capitalism. The creation of transcorporate networks of ownership and directorships – as an extension of individual corporate strategies – generated the change for the emergence of classwide organization. Prerequisites for promotion of top executives are not only decision-making and other abilities within the firm, but also demonstrated leadership outside it; prospects of promotion to top positions are generally enhanced if the aspirant serves effectively on other company boards, plays a relevant role in business associations, and achieves a reputation for having access to higher government posts and for effectively managing the relations with relevant stakeholders of the firm.

Some studies argue that in today's global market, with chief executives operating across national boundaries in transnational corporations and financial networks, managerial capitalism is changing into international capitalism (Vernon, 1971; Makler et al., 1982); an international business class, made up of manager-entrepreneurs with interconnected careers and a common cosmopolitan culture, plays a key role in the world economy (Bottomore and Brym, 1989); and corporate power stems from the ability to control global production networks and financial flows (Held et al., 1999).

What Managers Really Do: The Research on Managerial Behavior

After Fayol's pioneering study (1916), research on managerial behavior grew in the 1950s. Early studies, described as the Work Activity School by Mintzberg (1973), recorded and classified the work activities of managers using simple categories. Carlson (1951) organized a diary analysis of the activities actually performed in four weeks by seven Swedish and two French top managers, supplemented by extensive interviews and by background knowledge. It was a participating research, since it grew out of a longstanding discussion group of Swedish managing directors with Carlson, aimed at improving selection and training procedures for top management posts. The study showed that, with a few exceptions, managers spent most of their time talking and listening (up to more than 90 percent of their time). Most of their activities, apart from formal meetings, were brief, fragmented, and interrupted; they rarely had as much as half an hour a day alone. Managers were exposed to 'administrative pathologies,' such as being overloaded with details or neglecting personal inspection tours.

A second type of empirical research is well exemplified by Mintzberg's (1973) study of the weekly activities of five chief executives in different kinds of organizations. He classifies 10 managerial roles into three main headings: informational, interpersonal, and decisional. Interpersonal roles include figurehead, leader, and liaison, and also negotiator and spokesperson. But the key function is the monitoring and dissemination of information, which makes executive managers the nerve center of the organization and, we can add, provides a major basis for their power. Top managers' foremost professional hazard is the danger of dealing with all issues at a superficial level, because of the variety and brevity of their activities.

A third kind of studies of managerial behavior focuses on differences between managerial jobs. Stewart (1982, 1991) is a good example. She stresses different managerial styles and the constraints, opportunities, and choice models in managerial activity; and she identifies as main criteria of differences the nature and the difficulty of the contacts required in the job, whether the job was a boss-dependent one, and whether the job was an exposed one where the jobholder's mistakes were personally identifiable.

Kotter's (1982) study, based on lengthy interviews and documents of 15 general managers in various companies and their key contacts, introduced agendas and networks as key concepts. He found out that general managers tend to build their agendas—made of loosely connected goals and plans—during the first six months to one year in the job and through the development of a network of contacts. Experienced managers may have the problem of 'tunnel vision,' and therefore require occupational mobility in order to mix with people from various businesses and with different perspectives.

All these studies, and many similar others, contributed to show the complexity of managerial behavior and, at the same time, the difficulty of identifying what is specific about it. Whitley (1989) gives a good summary of what can be considered managerial about managerial activities. He suggests five main features: They are highly interdependent, contextual, and systemic; they are relatively unstandardized; they are changeable and developing; they combine both the maintenance of administrative structures and their changes; they rarely generate visible and separate outputs which can be directly connected to individual inputs.

The limitation of many of these studies is that they underplay the specific organizational contexts where managerial behavior takes place. An exception in this respect is

Hannaway's study (1989), based on almost 30 000 random samples of the work activities of 52 managers, divided into three hierarchical levels, over six weeks. She was interested in how managers behave in a hierarchical social system that is characterized by uncertain tasks, ambiguous outcomes, and biased feedback. Among her most interesting findings are the attempts of managers to protect themselves from blame when things go badly by involving others in any potentially risky decision, and to protect themselves from the ambiguity of managerial work signaling their worth to others. Hannaway's study introduces the research topic of the role of managers in the organization.

The Role of Managers in the Organization

Whereas most psychological studies of management focus on behavior, most sociological studies focus on managerial roles in the organization. A major influence in the study of management in organization was Weber's theory of bureaucratic administration, which fostered a view of managerial authority based on incumbency in a legally defined office and technical competence. Still more influential for the study of management was Barnard's (1938) classical study, which identified in administration, coordination, and control the essential managerial functions. For Barnard, the core of business organizations is the conscious and deliberate cooperation of different individuals for a common goal. Organizations are superior to individuals because of their intrinsic rationality. Accordingly, the function of managers is three-fold: they must take strategic decisions, through the rational evaluation of alternatives; they must assure the consensus and the cooperation of all members of the business organization through shared goals and value premises; and they must achieve a satisfactory mediation between organizational needs and individual expectations. Barnard's view laid the ground for major conceptions of the managerial role, which I will frame in the dominant organizational paradigms.

First, there is the view of the manager as a decision maker and an organization person, as it is derived by the organizational decision-making approach (Simon, 1957; March and Simon, 1958). This paradigm has been widely accepted in management schools; and since management schools are important in shaping managerial roles, it is the most widespread among managers and the most influential in shaping their actual behavior. For Simon, organizations help decision makers both by simplifying decisions and by supporting participants in the decisions they need to make. Organizations encourage decision makers to 'satisfice' rather than maximize – that is, to settle for acceptable as opposed to optimal solutions – to attend to problems sequentially rather than simultaneously, and to utilize existing repertoires of action programs rather than to develop novel responses for each situation. Organizations are necessary because of the cognitive limits of individual decision makers; they support rational decision making by setting integrated subgoals and stable expectations and by subdividing responsibilities among managers and other firm members, providing them with the necessary information, resources, and facilities. According to this view, the manager is mainly the organization person, and managerial behavior is an instance of bounded rationality and is largely influenced by organizational structures. Another version of the manager as organizer is Marris's (1979), who sees the ability to organize the complex functioning of modern large-scale firms as the distinctive feature of management.

The second paradigm is the conception of the manager as the informal, human relations-oriented leader, as developed in the human relations model (Roethlisberger and Dickson, 1939; Mayo, 1945; Likert, 1961), and related ones. It stresses the role of managers in exerting informal leadership, and promoting worker morale, increased effort, and cooperation in the service of organizational goals. From this perspective, good managerial leadership aims primarily at influencing the behavior of individual participants, and is conceived as democratic rather than authoritarian, employee-centered rather than production-centered, concerned with interpersonal relations and informal status hierarchies rather than with bureaucratic rules and formal authority. The human relations perspective stimulated relevant research and business policies aimed at increasing productivity through changes in work organization (job enlargement) and/or workers' participation in decision making. But, it was criticized both on ideological grounds – for de-emphasizing the actual conflict of interests between workers and management – and on empirical grounds, since several decades of research have demonstrated no clear relation between worker satisfacton and productivity, between leadership style and worker productivity, and between participation in decision making and worker satisfaction (Strauss, 1963; Hollander and Julian, 1969).

The third conception, best represented by Selznick (1957), is that of the manager as custodian of institutionalized values and mediator between organization and society. It is couched in the institutional paradigm and seeks to explain changes in the operative goals of the organization in the face of environmental constraints. Intellectually indebted to Michels and Merton, Selznick proposed to study the natural history of organizations by focusing on critical decisions that foster structural change. He discussed how they develop distinctive structures, capacities, and liabilities, through which they become institutionalized, that is, 'infused with value beyond the technical requirements of the task at hand' (Selznick, 1957, p. 17). For him, informal structures can never succeed in conquering the nonrational dimensions of organizational behavior, because organizations are concrete structures that must adapt to pressures of their institutional environment, and because they are made of individuals who participate as 'wholes' in the organization and do not act merely in terms of their formal roles within the system. Accordingly, managers issue their directives, although they may be neither understood nor followed by their subordinates. The managerial role is, therefore, defined not in terms of managers' authority, but of their relation to institutionalized values. Since organizations reflect and protect societal values, managers act as catalysts in this process (Scott and Davis, 2007).

The fourth view is that of the manager as a negotiator in a conflict-ridden context and as a power holder, or as I prefer to define it, a power broker. It can be drawn from both Cyert's and March's (1963) behavioral approach, from the students of organizations as negotiated orders, and as political arenas, like Crozier and Friedberg (1977) and Daudi (1989). In this

view, the plurality of goals, and the diversity of interests and outlooks among participants, make conflict not a symptom of organizational malfunctioning, but a normal feature of organizational life. According to Cyert and March, managers can adopt various mechanisms for conflict resolution, such as the budget – which stabilizes previous negotiations and expectations – standard procedures and operational rules, informal relations and traditional practices. Real strategies within the firm do not rationally maximize corporate ends but result from compromises within the dominant coalition and between the coalition and other collective actors.

The complex relation between corporate management and control and the insitutional context is also at the core of Fligstein's (1990) study on *The Transformation of Corporate Control*. Rooted in the paradigm which sees organizations as continuously transforming open systems, Fliegstein argues that changes in corporate control take place in three relevant institutional contexts. First of all, the firm itself is the institutional context where the internal strategy and structure of existing firms reflect organized power and interests, and managers' ideas about appropriate corporate behavior in the light of existing conventional wisdom. Second, organizations are embedded in larger groups of organizations which are called organizational fields, that may be defined in terms of product line, industry, or firm size (suppliers, distributors, financial agencies and, most often, competitors). The organizations that surround the firm provide constant information as to what is occurring in the firm's product markets – which are filtered, interpreted, and greatly affect what actions are possible. Third, the state sets the rules that define what organizations can do and what the limits of legal behavior are. Legislatures and courts define the rights and privileges and set the boundaries of appropriate corporate behavior. Antitrust laws, for instance, define the limits of competitive behavior. Managers' courses of action are not so much constructed in terms of self- interest (whether it be that of the shareholder or the managers themselves), but rather in terms of the legal framework and the self-conscious version of the world that make old and new courses of action possible and desirable. Fligstein reminds us that the worlds of top managers have always been highly structured and their actions shaped by social and political contexts.

With a more radical departure from previous approaches, the perspective of the organization as a negotiated order brings the actor back at the center of organizational analysis, and focuses on concepts like power, strategic game, dominant coalition, and negotiated order. Crozier and Friedberg (1977) do not see organizations as the product of individuals and groups to established roles and procedures, but rather as the result of games in which the freedom of the various actors playing is limited by the formal and informal rules of the organization. Their vision, however, does not amount to diluting organizational action into a sum of individual actions, because the outcomes of interactive strategic games are real organizational properties, and because all participants need a certain amount of cooperation from others and share with them a common interest in organizational survival. By virtue of each actor's constrained freedom, they have power over the other players, since each actor is a source of uncertainty for them and for the organization as a whole. According to this conception, managers can be seen as special players who have greater control over uncertainty areas and are consequently more able to mobilize appropriate resources, like specialized competence, control of information, manipulation of norms, and control of the relations between the organization and the relevant environments.

More explicitly focused on power is Bacharach and Lawler's (1980) model, which introduces the conception of managers as power holders. Indebted to Michel's classical work on political parties as organizational oligarchies, this approach sees managers as power holders, both in the informal structure of influence and in the formal structure of authority. It analyzes the power games, coalitions, alliances, and conflicts organization leaders play both within the dominant coalition and with other antagonistic groups, and the relations between these leaders and their followers.

Related to the view of the manager as a power holder is the study of politics in management. The classical study, Dalton's (1959) *Men who Manage*, analyzes how a group of middle managers distorted the information that they gave to head office to forward their own political goals. Managers have more opportunities than other employees to gain resources and to form alliances in order to forward their career. In the study of politics in management, the image of the manager is often reversed. Managers are not described as acting primarily in the interest of the organization, but in their own interest. From being the heroes of the successful corporation through efficient administration, sound financial control, and profit-maximizing strategies, they become the rulers of bloated bureaucracies that do not produce sufficiently high returns on investment to stockholders, and concentrate on their personal aggrandizement.

Managerial Ideologies and Managerial Styles

Two major pioneering studies of managerial ideologies were published in the same year: that by Bendix (1956) and that by Sutton et al. (1956). The former is a comparative-historical study of managerial ideologies, i.e., of all those ideas which aim at justifying and explaining managerial power and authority within the firm and, indirectly, the social status of the business class. Managerial ideologies are internalized by entrepreneurs and top managers and are exposed on their behalf. The changes in ideologies of management during the nineteenth and twentieth centuries in Anglo-American and Russian civilizations are both similar and divergent; they are similar insofar as they share the increased managerial concern with the attitudes of workers, which presumably account for their differential productivity; they diverge because in the Anglo-American historical environments the authority relations between employees and workers remained to a large extent autonomous realms of group relations, whereas in pre- and postrevolutionary Russia, the conduct of both employers and workers was regulated by the superordinate authority of the state.

Whereas Bendix's study is a comparative historical analysis of social theories and ideas worked out by thinkers who interpret the ideological needs of business elites to legitimize managerial power, Sutton et al. study the explicit public statements of US managers and entrepreneurs. They conceive ideology not as a specific device for legitimizing power in the

business organization, but as a multifunctional mechanism capable of alleviating the strains to which business- people are inevitably subject. Three major types of strains are identified: (a) strains stemming from criticism of the dominant position of business in society, that can undermine the businessperson's belief that corporate values are consistent with US culture; (b) strains originating from the contradictory demands put to the businessperson by the different interlocutors (stakeholders in today's language), i.e., stockholders, employees, competitors, consumers, colleagues; and (c) the strains deriving from the conflicting demands of other social roles that entrepreneurs and managers play outside the firm – in family, community, and various informal groups.

Within the limits of their own cultural traditions, businesspeople adhere to their particular kind of creed to resolve emotional conflicts, overcome the doubts, and alleviate the anxieties engendered by the actions that their roles as entrepreneurs and managers compel them to take. For individuals in business, the function of the ideology is to help them maintain the psychological ability to meet the demands of their job. For instance, the conflict of values between profit and ethics can be eased in two opposite ways: either by affirming that devotion to profit is not inconsistent with social welfare but is, rather, a prerequisite of economic efficiency; or by denying that private profit is, or ought to be, the principal orientation of the business enterprise. Given the importance of business, it follows that ideology has functional importance also for those on whom the actions of business people impinge and for the whole society.

Most research on managerial behavior and roles has been limited to one country. Studies on cultural differences in managerial styles have been more comparative, like Bendix's. Relevant cultural studies of management focus on managerial styles and distinctive national traditions. They usually adopt a comparative perspective and compare Western Europe and North America with Japan, and, until the fall of the USSR, Western and Soviet managerial styles and cultures (Boddewyn, 1971). Another group of research does not compare countries and socioeconomic systems, but rather different types of enterprise, along the private – public and the national – multinational dimensions.

The comparison with Japan is framed in the notion of the diversity of capitalisms that compares the US model of 'market-driven capitalism' with the European model of 'market social economy' and with the Japanese model of 'patronage capitalism.' The Anglo-Saxon model seems more competitive and more effective in job creation, whereas the Continental European one appears more effective in combining competitivenes and social cohesion. The Japanese experience is of interest for two main reasons: the successful competitive performance of Japanese firms in the world market, and recent technological changes that have made the Taylorist – Fordist types of work organization obsolete.

Drawing on the Japanese corporations' experience, authors like Dore (1973) and Ouchi (1981) have argued that managerial styles based on hierarchical relations have clear limits to their effectiveness. Overtly bureaucratized organizations, both of the Weberian type and on the Taylorist – Fordist model, lead to internal failures of functioning because of their rigid and participation-stifling nature. Clan forms of authority and the related styles of leadership, like those existing in the work groups and quality circles of Japanese firms, are more efficient, since they rely on the close ties and personal connections among the members of the group. Besides, this style of management develops noneconomic incentives, such as those based on collective identity and a sense of collective goals, which contribute to raise workers' morale and motivation (Markus et al., 2006).

Managerial career patterns are consistent with this culture, in the sense that institutional loyalty is a core value for Japanese managers, whereas for US and European ones a professional self-definition seems to prevail, in the sense that moving from one firm to another in order to improve one's career is considered fully legitimate. How far Japanese managerial styles depend on the cultural environment of Japan – which favors an especially compliant, deferential, and hardworking labor force – can be tested by studying the experience of Japanese-run firms recently set up in Western countries. Although the number of studies carried out so far is small, it does seem that Japanese management practices can operate with some success in a more individualistic labor force. Studies of Japanese-run plants in the USA, UK, and other Western European countries (White and Trevor, 1983) indicate that bottom-up decision making does work outside Japanese society. Workers seem to respond positively to the greater level of involvement these plants provide, as compared to that of the Western style of management to which they were previously exposed. In recent years, however, the complex impact of globalization and the slower growth of the Japanese economy have put strains on the distinctive Japanese model and redirected it, to some extent, toward the market-led model of capitalism and US styles of management.

Another important instance of of cross-cultural influence in managerial styles has been the imitation, and to some extent the imposition, of US managerial values and practices in France and other Western European countries as a consequence of the USA's victory in World War II (Boltanski, 1990).

The comparison of managerial styles and cultures in private and state-controlled enterprises has been another relevant, although declining, area of study. State-controlled firms differ to some extent from private ones because of their patterns of recruitment and socializing agencies and because of the distinctive features of their business environment. Public firms, managers tend to come more often than their private firms colleagues from either government bureaucracy, as in the French case, or politics and the liberal professions, as in the Italian case. They often have a distinct political-ideological imprinting, as did the managers of postwar British nationalized industries (who had Labour Party affiliations) or the managers of postwar Italy's state-controlled firms (who had been educated at the Catholic University of Milan and had political experience in the Christian Democratic Party). They often showed different attitudes to labor relations and in corporate strategies because their dependence on state policies may have caused diffusion of targets, subordination of corporate goals to broader 'social' ends, and interferences by specific party factions and interest groups.

While some authors (Rees, 1984) point out that they can act according to more flexible cost-opportunity criteria of management efficiency, others (Martinelli, 1981) argue that they risk experiencing a threefold subordination: to private

capital, which uses state controlled firms to get rid of unprofitable businesses and to socialize losses; to political parties, who use them for quelling social tensions and granting benefits to specific interest groups; and to unions, who find in them a weaker counterpart than private companies, and a place where they can win 'easy' victories at the state budget's expense. However, these differences should not be exaggerated, especially for managers of public firms that compete with private ones in the same market environment.

A growing field is the study on executives' managerial cultures and styles in the extensive literature on multinational corporations (Weinshall, 1977) and on globalization (Held et al., 1999). The prevailing view is that managers of multinational corporations do not differ significantly from national firms' managers, except for a greater emphasis on cosmopolitan values and styles of life, a greater sensibility for the scope of their decisions, and a greater alertness to the different societal environments in which they work. An interesting area in the analysis of multinational corporations' leadership styles is industrial relations, where a unified employers' strategy confronts a number of fragmented and diversified trade-union strategies (Makler et al., 1982; Hollingsworth and Boyer, 1997).

As for the impact of modern globalization on managerial cultures and styles, contradictory tendencies are at work. On the one hand, distinct national and corporate-specific managerial styles persist, as do distinct institutional arrangements (different functions played by states although declining in their sovereign power, distinct trade-union strategies and organizations, specific organizational devices for increasing efficiency and workers' morale, etc.). On the other, the increasingly interconnected market economy fosters the diffusion of a common managerial culture and of cosmopolitan attitudes and styles among managers. It is still an open question to what extent this cosmopolitan culture is the diffusion of a single US-inspired model or the result of a hybridization of management styles and cultures. This seems the most promising line of future research in managerial roles, cultures, and behavior (Guillen and García-Canal, 2009).

See also: Administration in Organizations; Corporate Culture; Corporate Governance; Entrepreneurship and Economic Growth; Leadership in Organizations, Sociology of; Multinational Corporations; Organizational Decision Making; People in Organizations; Stockholders' Ownership and Control; Strategic Management; Strategizing.

Bibliography

Bacharach, S.B., Lawler, E.J., 1980. Power and Politics in Organizations. Jossey-Bass, San Francisco.
Barnard, C., 1938. The Functions of the Executive. Harvard University Press, Boston, MA.
Bendix, R., 1956. Work and Authority in Industry. Wiley, New York.
Berle Jr, A., Means, G.C., 1932. The Modern Corporation and Private Property. Harcourt, New York.
Boddewyn, J.J., 1971. European Industrial Managers, West and East. International Arts and Sciences Press, White Plains, NY.
Boltanski, L., 1990. Visions of American management in postwar France. In: Zukin, S., DiMaggio, P. (Eds.), Structures of Capital: the Social Organization of Economy. Cambridge University Press, New York, pp. 343–372.
Bottomore, T., Brym, R., 1989. The Capitalist Class. An International Study. Harvester Wheatsheaf, New York.
Burch, P.H., 1972. The Managerial Revolution Reassessed: Family's Control in America's Large Corporations. Lexington Books, Lexington, MA.
Burnham, J., 1941. The Managerial Revolution. John Day, New York.
Carlson, S., 1951. Executive Behavior: A Study of the Workload and Working Methods of Managing Directors. Stromberg, Stockholm.
Chandler Jr, A.D., 1962. Strategy and Structure: Chapters in the History of Industrial Enterprise. MIT Press, Cambridge, MA.
Chandler Jr, A.D., 1977. The Visible Hand: The Managerial Revolution in American Business. Harvard University Press, Cambridge, MA.
Coase, R.H., 1937. The nature of the firm: Origin, meaning, influence. Economica 4, 386–405.
Crozier, M., Friedberg, E., 1977. L'acteur et le systeme. Seuil, Paris.
Cyert, R.M., March, J.G., 1963. A Behavioral Theory of the Firm. Prentice-Hall, Englewood Cliffs, NJ.
Dalton, M., 1959. Men Who Manage. Wiley, New York.
Daudi, P., 1989. The Discourse of Power in Managerial Praxis. Blackwell, Oxford, UK.
Davis, G.F., 2009. The rise and fall of finance and the end of the society of organizations. Academy of Management Perspectives 23 (3), 27–44.
Dore, R., 1973. British Factory, Japanese Factory: The Origins of National Diversity in Industrial Relations. Allen & Unwin, London.
Fayol, H., 1916. Administration industrielle et générale. English translation 1949. Pitman, London.
Fligstein, N., 1990. The Transformation of Corporate Control. Harvard University Press, Cambridge, MA.
Galbraith, J.K., 1967. The New Industrial State. Hamish Hamilton, London.
Guillén, M.F., García-Canal, E., 2009. The american model of the multinational firm and the "New" Multinationals from emerging economies. Academy of Management Perspectives 23 (2), 23–35.
Hannaway, J., 1989. Managers Managing. Oxford University Press, New York.
Harbison, F., Myers, C.A., 1959. Management in the Industrial World. McGraw-Hill, New York.
Held, D., McGrew, A., Goldblatt, D., Perraton, J., 1999. Global Transformations. Polity Press, Cambridge, UK.
Hollander, E.P., Julian, J.W., 1969. Contemporary trends in the analysis of leadership process. Psychological Bulletin 71, 387–397.
Hollingsworth, J.R., Boyer, R. (Eds.), 1997. Contemporary Capitalism. The Embeddedness of Institutions. Cambridge University Press, New York.
Kotter, J., 1982. The General Managers. The Free Press, New York.
Lewellen, W.G., 1971. The Ownership Income of Management. Princeton University Press, Princeton, NJ.
Likert, R., 1961. New Patterns of Management. McGraw-Hill, New York.
Makler, H., Martinelli, A., Smelser, N.J., 1982. The New International Economy. Sage, London.
March, J.G., Simon, H.A., 1958. Organizations. Wiley, New York.
Markus, H.R., Uchida, Y., Omoregie, H., Townsend, S.S., Kitayama, S., 2006. Going for the gold. Models of agency in japanese and american contexts. Psychological Science 17 (2), 103–112.
Marris, R., 1979. Theory and Future of the Corporate Economy and Society. North Holland, Amsterdam.
Martmelli, A., 1981. The Italian experience. In: Vernon, R., Aharoni, Y. (Eds.), State-owned Enterprise in the Western Economies. Croom Helm, London, pp. 85–99.
Mathews, J.A., 2010. Lachmannian insights into strategic entrepreneurship: resources, activities and routines in a disequilibrium world. Organization Studies 31 (2), 219–244.
Mayo, E., 1945. The Social Problems of an Industrial Civilization. Harvard University Press, Cambridge, MA.
Meynaud, J., 1964. La technocratie: mythe au realité? PUF, Paris.
Mintz, B., Schwartz, M., 1985. The Power Structure of American Business. Chicago University Press, Chicago.
Mintzberg, H., 1973. The Nature of Managerial Work. Harper & Row, New York.
Mizruchi, M.S., 1982. The American Corporate Network: 1904–1974. Sage, Beverly Hills, CA.
Ouchi, W.A., 1981. Theory Z: How American Firms Can Meet the Japanese Challenge. Addison Wesley, Reading, MA.
Rathenau, W., 1918. Die neue Wirtschaft. S Fisher, Berlin.
Rees, R., 1984. Public Enterprise Economics. Weidenfeld and Nicholson, London.
Roethlisberger, F.J., Dickson, W.J., 1939. Management and the Workers. Harvard University Press, Cambridge, MA.
Scott, W.R., Davis, G.F., 2007. Organizations and organizing: Rational, natural and open system perspectives. Pearson Education, Upper Saddle River.
Selznick, P.H., 1957. Leadership in Administration. Harper & Row, New York.
Simon, H., 1957. Administrative Behavior. Macmillan, New York.
Stewart, R., 1982. Choices for the Manager. McGraw-Hill, Maidenhead, UK.

Stewart, R., 1991. Chairmen and chief executives: an exploration of their relationship. Journal of Management Studies 28 (5), 511–528.
Stockman, F.N., Ziegler, R., Scott, J. (Eds.), 1983. Intercorporate Structure: Comparative Analysis of Ten Countries. London.
Strauss, G., 1963. Some notes on power equalization. In: Leavitt, H.J. (Ed.), The Social Science of Organization. Prentice-Hall, Englewood Cliffs, NJ.
Streeck, W., Schmitter, P., 1985. Community, market, state and associations. In: Private Interest Government: Beyond Markets and State. Sage, London.
Sutton, F.X., Harris, S.E., Kaysen, C., Tobin, J., 1956. The American Business Creed. Harvard University Press, Cambridge, MA.
Useem, M., 1984. The Inner Circle: Large Corporations and the Rise of Business Political Activity in the U.S. and U.K. Oxford University Press, Oxford, UK.
Vernon, R., 1971. Sovereignty at Bay. Basic Books, New York.
Weinshall, T.D. (Ed.), 1977. Culture and Management. Penguin Books, London.
White, M., Trevor, M., 1983. Under Japanese Management: The Experience of British Workers. Heinemann, London.
Whitley, R., 1989. On the nature of managerial tasks and skills: their distinguishing characteristics and organization. Journal of Management Studies 26 (3), 209–224.
Williamson, O., 1975. Markets and Hierarchies: Analysis and Antitrust Implications. Free Press, New York.

Managerial and Organizational Cognition

Gerard P Hodgkinson, Warwick Business School, University of Warwick, Coventry, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Incorporating the insights of the full range of the behavioral and social sciences, managerial and organizational cognition (MOC) is an interdisciplinary endeavor that seeks to advance fundamental and applied knowledge of how individuals and social units in human organizations acquire, store, recall, interpret, and act upon information. Originating during World War II, the study of cognition in the workplace has progressed from being a largely monodisciplinary application of experimental cognitive psychology, with a view to enhance practices targeted at helping individuals to overcome fundamental information processing limitations, to a much richer conception of people and their jobs. Situated at the nexus of the cognitive and organizational sciences, MOC is continuously evolving new theories, research, and tools to better explain and predict behavior in and around the workplace, thus improving the effectiveness and well-being of individuals and intra- and interorganizational collectives.

Introduction

Managerial and organizational cognition (MOC) is an interdisciplinary endeavor that seeks to advance fundamental and applied knowledge of how individuals and social units in human organizations acquire, store, recall, interpret, and act upon information. Its scope is as broad as the entire field of management and organization studies, offering cognitive perspectives on topics and subfields as varied as entrepreneurship (Baron, 2004; Grégoire et al., 2011; Shepherd and Krueger, 2002), industrial–organizational psychology (Avolio et al., 2009; Hennessey and Amabile, 2010; Hodgkinson and Healey, 2008; Lord et al., 2010), organization theory and design (e.g., Huber, 1991, 2004), and strategic management (Powell et al., 2011; Hodgkinson and Sparrow, 2002; Huff, 1990; Porac et al., 1989). Although its main foundations were initially psychological in nature, MOC now also incorporates the insights of sociology, philosophy, behavioral economics, and social neuroscience, among other disciplines.

Historically, research on cognition in the workplace has fallen under two major traditions, namely, the human factors tradition and the organizational tradition (Hodgkinson and Healey, 2008). The human factors tradition began during World War II, with the prime intention of enhancing military efficiency and effectiveness. Advancing the human information processing approach to the analysis of skilled behavior, psychologists pursued a program of basic research, which inter alia informed the design of control and display systems pertaining to aircraft, warships, and related military equipment, the scheduling of work, and the design of wider task environments. During this foundational period, skilled performance was modeled as a series of computer-like input–throughput–output processes (e.g., sensory mechanisms, memory functions, and decision outputs), which, when operating in concert, generally culminated in faster performance with less error (see, e.g., Broadbent, 1958; Welford, 1976; Wickens and Carswell, 2006).

Post–World War II, as industrial processes became increasingly more complex, the experimental approach of cognitive psychology that had underpinned the human factors tradition of the war effort continued, as researchers turned to investigate directly the nature and impact of human operators' mental representations of their tasks and the wider systems in which they were embedded (see, e.g., Edwards and Lees, 1974). More recently, following the rise of computer-based work systems, research and theory development within this tradition has centered primarily, although by no means exclusively, on human–computer interaction (see, e.g., Card et al.'s, 1983, seminal volume). From basic desktop machines to modern-day flight decks, the impact of contemporary work in the human factors tradition has been considerable.

Soon after World War II, the late Nobel Laureate Herbert Simon's seminal volumes, *Administrative Behavior* (Simon, 1947) and *Organizations* (March and Simon, 1958), were published together with Cyert and March's (1963) classic work, *A Behavioral Theory of the Firm*. Collectively, these books laid the foundations of the second major tradition in the analysis of cognition in the workplace, namely, the organizational tradition. Work in this tradition conceived organizations as constellations of boundedly rational actors, each beset with fundamental information processing limitations of attention and perception, and which all too frequently embarked upon suboptimal courses of action. This body of work thus established the 'computational perspective' within the organizational tradition (Lant and Shapira, 2001), exemplified by the heuristics and biases program of work pioneered by the late Amos Tversky and Nobel Laureate Daniel Kahneman (see, e.g., Tversky and Kahneman, 1974).

An impressive volume of evidence has been amassed consistent with the idea that in order to render the world manageable, organizational decision makers employ a variety of 'rules of thumb,' known as 'heuristics.' Heuristic processing strategies enable the decision maker to cut through the welter of information bombarding them, by imposing a number of simplifying assumptions on the data. Simon (1957) identified one such heuristic, known as 'satisficing.' According to Simon, faced with a number of choice alternatives, decision makers select the first alternative that meets their minimum requirements, rather than choosing the best from all the alternatives on offer. Satisficing is much simpler in terms of its cognitive operations, thus placing fewer demands on scarce mental

resources. Over several decades, researchers have identified other heuristics that, in common with satisficing, involve less mental effort and there is strong evidence suggesting that heuristic modes of reasoning underpin human judgment processes in the workplace (for details, see Hodgkinson and Sparrow, 2002). Although the use of heuristics reduces the information processing requirements on the decision maker, there are also risks involved, in that their deployment may lead to suboptimal or 'biased' outcomes. In the case of satisficing, for example, once an 'acceptable' option is found, the search for, and evaluation of, additional potentially viable alternatives ceases. Consequently, 'better' options that have yet to be considered will be ignored. The practical implications arising from this work are that, decision makers should be encouraged to expend greater cognitive effort and reason more systematically with a view of debiasing their judgments. However, precisely how this might be accomplished remains the subject of considerable controversy, with only limited high-quality scientific evidence supporting some of the more popular practices in vogue.

Computation versus Interpretation

The work of Karl Weick and colleagues (e.g., Weick, 1979, 1995; Sutcliffe and Weick, 2008) has challenged a number of the core ontological assumptions on which the computational perspective within the organizational tradition is predicated, in particular the assumption that task and wider institutional environments are objective entities that can only be partially comprehended, due to limited processing capacity. Advancing an alternative account, predicated on an interpretivist philosophy, Weick and colleagues maintain that decision makers literally create their own constraints through active, constructive processes of 'enactment' and 'sensemaking,' variously rearranging, isolating, and demolishing seemingly objective features of their surroundings, in turn giving rise to intersubjective differences in perceptions of reality.

In the final analysis, as argued by Lant and Shapira (2001), the processes of bounded rationality and sensemaking and enactment, highlighted variously by scholars whose work has been informed respectively by the computational and interpretive perspectives, ultimately coexist in a dynamic interplay. The former, exemplified by the work on mental representations and behavioral decision research, draws attention to the fundamental information processing limitations of organizational decision makers and the strategies they employ in the *downstream* choice or calculation processes at the heart of decision making and problem solving, in an effort to overcome those limitations, whereas the latter, exemplified by the work of Weick and colleagues, emphasizes the *upstream* processes adopted by individuals and groups to extract patterns of meaning from ambiguous environmental cues in the social construction of organizational realities.

Upper Echelons Theory

Another classic development worthy of brief mention at this juncture is the upper echelons theory of strategic decision making (Hambrick and Mason, 1984). Drawing on many of the foregoing concepts, this theory maintains that senior executives forming the organization's dominant coalition make decisions under bounded rationality, based on their interpretations of the organizational circumstances that they are involved in, and that their interpretations of those circumstances are necessarily partial, informed by the 'administrative givens' they bring to bear on the situation at hand. These givens originate from executives' education, prior work experience, personal values, and a host of related individual differences that direct their attention selectively and shape their perceptions. Unfortunately, however, the highly sensitive nature of the issues and problems confronting executive teams, to say nothing of the time pressure under which they function, renders access for research purposes highly problematic. Hambrick and Mason (1984) thus advocated methods that obviate the need to employ direct methods of cognitive assessment, using executives' background characteristics (e.g., education and functional specialization) as proxies for cognitive variables (i.e., values and beliefs) in the prediction of organizational outcomes (e.g., innovation and firm performance). A detailed consideration of all the available evidence for and against the upper echelons perspective lies beyond the scope of an article of this nature. Here, it is sufficient to note that from the early 1990s onward, the upper echelon approach has been the subject of growing theoretical and empirical scrutiny, amid numerous contradictory and inconsistent findings. Responding to these challenges, more recent work has incorporated direct methods of cognitive assessment in an attempt to isolate the determinants and consequences of executive perceptions and beliefs (see, e.g., Chattopadhyay et al., 1999; Markoczy, 2001).

Current and Emerging Directions in Theory and Research

Foundational Concepts under Scrutiny

Each of the foregoing developments, located variously along the computation – interpretation continuum, pervades the main substantive domains of inquiry that constitute MOC. However, a number of foundational concepts at the heart of the human factors and organizational traditions are giving way to alternative conceptions. For instance, the 'taken-for-granted' notion that cognition resides exclusively in people's heads (as implied by stage-based human information processing models of decision making and skilled performance) is increasingly less tenable. Thus, for example, researchers pioneering the 'cognitive engineering' approach in the human factors tradition, emphasizing the importance of understanding the myriad of interactions that occur between task and environmental constraints on one hand and operators' knowledge structures on the other hand (e.g., Zhang and Norman, 1994), are beginning to examine the role of complex workplace technology as both a shaper and repository of knowledge (see also Hutchins, 1995), a trend paralleled in the organizational tradition (cf Walsh and Ungson, 1991).

Although the basic mental models concept, coined initially by Craik (1943), continues to play a central role in the more recent work of both traditions, it, too, has been problematized, with usage ranging from a manner akin to Johnson-Laird's

(1983) initial conception of mental models as temporary dynamic models in working memory, to the notion of enduring knowledge structures in long-term memory, akin to Bartlett's (1932) notion of schema. Not surprisingly, this basic lack of agreement pertaining to the definition and usage of the notion of mental models at the individual level has similarly hampered progress at the team, organizational, and interorganizational levels of analysis, resulting in a proliferation of constructs and techniques for operationalizing the various notions in play (for critical overviews of these developments, see Hodgkinson and Healey, 2008; Hodgkinson and Sparrow, 2002; Walsh, 1995).

Rethinking Bounded Rationality and the Role of Intuition

Work extending the basic ideas of Kahneman and Tversky – the epitome of the computational perspective – has also come under increasingly critical scrutiny on philosophical, theoretical, and methodological grounds. Predicated upon a fundamentally different conception of Simon's bounded rationality notion from that construed by traditional heuristics and biases researchers, known as ecological rationality, Gigerenzer and colleagues (e.g., Gigerenzer, 1991; Gigerenzer and Goldstein, 1996) have identified a new class of heuristics, known as 'fast and frugal heuristics,' which they maintain are adaptively better matched to the informational structure and demands of decision makers' everyday environments than those studied conventionally by decision researchers, based on the work of Kahneman and Tversky. Arguing that many of the experimental studies demonstrating the irrationality of human judgment are based on tasks involving probabilistic reasoning and other forms of abstract judgment that are far removed from the real-world environments to which humankind has readily adapted over many years, Gigerenzer and colleagues' fast and frugal heuristics make minimal computational demands on the decision maker, do not involve probabilities and utilities, and are deployed in situations that require individuals to draw inferences with limited time and knowledge. According to Gigerenzer and his colleagues, fast and frugal heuristics complement *satisficing heuristics*, i.e., heuristics for searching through available alternatives. The latter are deployed in choice situations where the decision maker does not know much about the possibilities ahead of time. However, as observed by Hodgkinson and Healey (2008), very few studies have investigated the deployment of fast and frugal heuristics in naturalistic settings and the limited work to date has yielded mixed findings regarding the extent to which decision makers actually rely on heuristics falling within this class and with what effect.

A second group of researchers, falling under the umbrella of naturalistic decision making (NDM), similarly rejects the notion of equivalency between the sparse confines of the laboratory and the infinitely richer settings in which organizational decision makers conduct their everyday affairs. Originating in studies of domain experts making complex, high stakes, and ill-structured decisions under time pressure, often in dangerous situations, NDM has gathered momentum over the past decade, both in organizational (Lipshitz et al., 2006) and human factors (e.g., Lipshitz and Cohen, 2005) applications. In theoretical terms, Klein's (1993) recognition-primed decision making (RPD) model, with its emphasis on the crucial role of pattern recognition in obviating the need for extensive deliberation of multiple alternatives, epitomizes the NDM approach. In marked opposition to the prescriptions of Kahneman and Tversky's work, which views intuition as a potential source of cognitive bias, the RPD model highlights intuition as an important foundation of skilled performance in many everyday situations. Interestingly, Kahneman and Klein (2009, 2010) have recently collaborated on a series of articles in which they have outlined a series of conditions that they agree, when present, warrant the use of intuition to enable effective decision making.

Emotionalizing MOC

In the wake of advances in social cognitive neuroscience (e.g., Lieberman, 2007) and neuroeconomics (e.g., Loewenstein et al., 2008), MOC researchers are turning to study the role of affect (i.e., feelings and emotions) in organizational decision processes and strategic adaptation. To the extent that behavioral plausibility is a desirable criterion for management and organization theory (Gavetti et al., 2007), social neuroscience has an important role to play in defining what is behaviorally plausible (Healey and Hodgkinson, 2014). Computational models of human thinking and action based purely on 'cold' cognition still predominate in management and organization theory (Hodgkinson and Healey, 2008). However, continuing to build MOC theory on computational models that are themselves a product of the experimental psychology of the mid-twentieth century is problematic because, contemporary psychological science suggests that many aspects of behavior once thought as the preserve of cold cognition rely heavily on affective mechanisms. The turn in MOC to embrace hot cognition – informed by neuroscientific studies of the mechanisms recruited by behavioral phenomena such as attention, decision, and sensemaking – thus demands a rethink of the field's fundamental assumptions concerning human cognition (Hodgkinson and Healey, 2011).

Refining Methodological Assumptions and Techniques

Going forward, techniques that have been developed over many years as a basis for studying individual and collective 'revealed representations' of explicit knowledge in and around the workplace (see, e.g., Huff, 1990; Fiol and Huff, 1992; Shadbolt and Milton, 1999; Sparrow, 1998) must now be adapted to meet these new challenges. In addition, new methods are required to study implicit representations of knowledge. To this end, innovative procedures such as the implicit association test for the assessment of implicit attitudes (Greenwald et al., 1998) and the use of subconscious priming techniques (see Shantz and Latham, 2009) to investigate subconscious goals (Latham et al., 2010) are on the ascendency; although, several nontrivial theoretical, methodological, and ethical challenges have yet to be resolved (cf Jost et al., 2009; Tetlock and Mitchell, 2009). Arguably, one of the biggest methodological challenges that has yet to be resolved is the question of how to achieve

unified accounts of cognitive–affective–behavioral functioning that interconnect individual microprocesses (e.g., basic judgments and choices) and collective behavior (e.g., group and team functioning, organizational innovation, and interorganizational dynamics), while avoiding the problem of reductionism and related philosophical pitfalls (cf Healey and Hodgkinson, 2014; Powell et al., 2011).

Concluding Remarks

As noted at the outset, the scope of MOC is as broad as the entire field of management and organizational studies. Hence, in this short article, it has only been possible to offer a thumbnail sketch of selected concepts and theoretical approaches. Important topics that are not surveyed in this brief article include identity (e.g., Ashforth and Mael, 1989; Haslam, 2004), attribution (Martinko et al., 2006), and learning (e.g., Huber, 1991; Wood and Bandura, 1989). Each of these major bodies of work, together with those discussed in greater detail in the previous sections, contributes to the vibrancy of this diverse and exciting eclectic field.

See also: Behavior Analysis, Applied; Decision Making: Nonrational Theories; Emotions and Work; Human Factors and Ergonomics; Implicit Association Test; Industrial–Organizational Psychology: Science and Practice; Personality and Values at Work.

Bibliography

Ashforth, B.E., Mael, F., 1989. Social identity theory and the organization. Academy of Management Review 14, 20–39.

Avolio, B.J., Walumbwa, F.O., Weber, T.J., 2009. Leadership: current theories, research and future directions. Annual Review of Psychology 60, 421–449.

Baron, R.A., 2004. The cognitive perspective: a valuable tool for answering entrepreneurship's basic "why" questions. Journal of Business Venturing 19, 221–239.

Bartlett, F.C., 1932. Remembering: A Study in Experimental and Social Psychology. Cambridge University Press, Cambridge.

Broadbent, D.E., 1958. Perception and Communication. Pergamon Press, London.

Card, S., Moran, T.P., Newel, A., 1983. The Psychology of Human-Computer Interaction. Erlbaum, Hillsdale, NJ.

Chattopadhyay, P., Glick, W.H., Miller, C.C., Huber, G.P., 1999. Determinants of executive beliefs: comparing functional conditioning and social influence. Strategic Management Journal 20, 763–789.

Craik, K., 1943. The Nature of Explanation. Cambridge University Press, Cambridge.

Cyert, R.M., March, J.G., 1963. A Behavioral Theory of the Firm. Prentice-Hall, Englewood Cliffs, NJ.

Edwards, E., Lees, F.P. (Eds.), 1974. The Human Operator in Process Control. Taylor & Francis, London.

Fiol, C.M., Huff, A.S., 1992. Maps for managers. Where are we? Where do we go from here? Journal of Management Studies 29, 267–285.

Gavetti, G., Levinthal, D., Ocasio, W., 2007. Neo-carnegie: the carnegie school's past, present, and reconstructing for the future. Organization Science 18, 523–536.

Gigerenzer, G., 1991. How to make cognitive illusions disappear. Beyond heuristics and biases. In: Stroebe, W., Hewstone, M. (Eds.), European Review of Social Psychology, 2. Wiley, Chichester, UK, pp. 83–115.

Gigerenzer, G., Goldstein, D.G., 1996. Reasoning the fast and frugal way: models of bounded rationality. Psychological Review 103, 650–669.

Greenwald, A.G., McGhee, D.E., Schwartz, J.L.K., 1998. Measuring individual differences in implicit cognition: the implicit association test. Journal of Personality and Social Psychology 74, 1464–1480.

Grégoire, D.A., Corbett, A.C., McMullen, J.S., 2011. The cognitive perspective in entrepreneurship: an agenda for future research. Journal of Management Studies 48, 1443–1477.

Hambrick, D.C., Mason, P.A., 1984. Upper echelons: the organization as a reflection of its top managers. Academy of Management Review 9, 193–206.

Haslam, S.A., 2004. Psychology in Organizations: The Social Identity Approach, second ed. Sage, London.

Healey, M.P., Hodgkinson, G.P., 2014. Rethinking the philosophical and theoretical foundations of organizational neuroscience: a critical realist alternative. Human Relations 67 (7), 765–792.

Hennessey, B.A., Amabile, T.M., 2010. Creativity. Annual Review of Psychology 61, 569–598.

Hodgkinson, G.P., Healey, M.P., 2008. Cognition in organizations. Annual Review of Psychology 59, 387–417.

Hodgkinson, G.P., Healey, M.P., 2011. Psychological foundations of dynamic capabilities: reflexion and reflection in strategic management. Strategic Management Journal 32, 1500–1516.

Hodgkinson, G.P., Sparrow, P.R., 2002. The Competent Organization: A Psychological Analysis of the Strategic Management Process. Open University Press, Buckingham, UK.

Huber, G.P., 1991. Organizational learning: the contributing processes and the literatures. Organization Science 2, 88–115.

Huber, G.P., 2004. The Necessary Nature of Future Firms: Attributes of Survivors in a Changing World. Sage, Thousand Oaks, CA.

Huff, A.S. (Ed.), 1990. Mapping Strategic Thought. Wiley, Chichester, UK.

Hutchins, E., 1995. Cognition in the Wild. MIT Press, Cambridge, MA.

Johnson-Laird, P.N., 1983. Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness. Cambridge University Press, Cambridge, MA.

Jost, J.T., Rudman, L.A., Blair, I.V., Carney, D.R., Dasgupta, N., Glaser, J., Hardin, C.D., 2009. The existence of implicit bias is beyond reasonable doubt: a refutation of ideological objections and executive summary of ten studies that no manager should ignore. Research in Organizational Behavior 29, 39–69.

Kahneman, D., Klein, G., 2009. Conditions for intuitive expertise a failure to disagree. American Psychologist 64, 515–526.

Kahneman, D., Klein, G., 2010. When can you trust your gut? McKinsey Quarterly March, 58–67.

Kahneman, D., Slovic, P., Tversky, A. (Eds.), 1982. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge.

Klein, G.A., 1993. A recognition-primed decision (RPD) model of rapid decision making. In: Klein, G., Orasanu, J., Calderwood, R., Zsambok, C. (Eds.), Decision Making in Action: Models and Methods. Ablex, Norwood, CT, pp. 138–147.

Lant, T.K., Shapira, Z., 2001. New research directions on organizational cognition. In: Lant, T.K., Shapira, Z. (Eds.), Managerial and Organizational Cognition: Computation and Interpretation. Lawrence Erlbaum, Mahwah, NJ, pp. 367–376.

Latham, G.P., Stajkovic, A.D., Locke, E.A., 2010. The relevance and viability of subconscious goals in the workplace. Journal of Management 36, 234–255.

Lieberman, M.D., 2007. Social cognitive neuroscience: a review of core processes. Annual Review of Psychology 58, 259–289.

Lipshitz, R., Cohen, M.S., 2005. Warrants for prescription: analytically and empirically based approaches to improving decision making. Human Factors 47, 102–120.

Lipshitz, R., Klein, G., Carroll, J.S. (Eds.), 2006. Special Issue on 'Naturalistic Decision Making and Organizational Decision Making: Exploring the Intersections' Organization Studies 27, 917–1057.

Loewenstein, G., Rick, S., Cohen, J.D., 2008. Neuroeconomics. Annual Review of Psychology 59, 647–672.

Lord, R.G., Diefendorff, J.M., Schmidt, A.M., Hall, R.J., 2010. Self-regulation at work. Annual Review of Psychology 61, 543–568.

March, J.G., Simon, H.A., 1958. Organizations. Wiley, New York.

Markoczy, L., 2001. Consensus formation during strategic change. Strategic Management Journal 22, 1013–1031.

Martinko, M., Douglas, S.C., Harvey, P., 2006. Attribution theory in industrial and organizational psychology. In: Hodgkinson, G.P., Ford, J.K. (Eds.), International Review of Industrial and Organizational Psychology, vol. 21. Wiley, Chichester, UK, pp. 127–187.

Porac, J.F., Thomas, H., Baden-Fuller, C., 1989. Competitive groups as cognitive communities: the case of Scottish knitwear manufacturers. Journal of Management Studies 26, 397–416.

Powell, T.C., Lovallo, D., Fox, C.R., 2011. Behavioral strategy. Strategic Management Journal 32, 1369–1386.

Shadbolt, N., Milton, N., 1999. From knowledge engineering to knowledge management. British Journal of Management 10, 309–322.

Shantz, A., Latham, G.P., 2009. An exploratory field experiment of the effect of subconscious and conscious goals on employee performance. Organizational Behavior and Human Decision Processes 109, 9–17.
Shepherd, D.A., Krueger, N.F.J., 2002. An intentions-based model of entrepreneurial teams' social cognition. Entrepreneurship Theory and Practice 27, 167–186.
Simon, H.A., 1947. Administrative Behavior. Macmillan, New York.
Simon, H.A., 1957. Administrative Behavior, second ed. Macmillan, New York.
Sparrow, J.A., 1998. Knowledge in Organizations: Access to Thinking at Work. Sage, London.
Sutcliffe, K.M., Weick, K.E., 2008. Information overload revisited. In: Hodgkinson, G.P., Starbuck, W.H. (Eds.), The Oxford Handbook of Organizational Decision Making. Oxford University Press, Oxford, pp. 56–75.
Tetlock, P.E., Mitchell, G., 2009. Implicit bias and accountability systems: what must organizations do to prevent discrimination? Research in Organizational Behavior 29, 3–38.
Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. Science 185, 1124–1131.
Walsh, J.P., 1995. Managerial and organizational cognition: notes from a trip down memory lane. Organization Science 6, 280–321.
Walsh, J.P., Ungson, G.R., 1991. Organizational memory. Academy of Management Review 16, 57–91.
Weick, K.E., 1979. The Social Psychology of Organizing, second ed. Addison-Wesley, Reading, MA.
Weick, K.E., 1995. Sensemaking in Organizations. Sage, Thousand Oaks, CA.
Welford, A.T., 1976. Skilled Performance. Scott Foresman, Glenview, IL.
Wickens, C.D., Carswell, C.M., 2006. Information processing. In: Salvendy, G. (Ed.), Handbook of Human Factors and Ergonomics. Wiley, Hoboken, NJ, pp. 111–149.
Wood, R., Bandura, A., 1989. Social cognitive theory of organizational management. Academy of Management Review 14, 361–384.
Zhang, J.J., Norman, D.A., 1994. Representations in distributed cognitive tasks. Cognitive Science 18, 87–122.

Relevant Website

http://division.aomonline.org/moc/ – The Managerial and Organizational Cognition (MOC) Division of the Academy of Management comprises the world's most significant professional body devoted to the development of the MOC field. Its Web site contains many useful resources for enhancing research, pedagogy and practice.

Mania

Neel Halder, Alpha Hospitals, Bury, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Bipolar disorder, previously known as manic depression, is a chronic episodic mental illness associated with behavioral disturbances. It is characterized by episodes of mania (or hypomania) and depression, and either one can occur first. A diagnosis of bipolar requires at least one depressive episode and at least one manic or hypomanic episode. This article explores the history, diagnosis, epidemiology, course and prognosis, etiology, and management of bipolar disorder. The management of the depressive phase is similar to unipolar depression, and is covered in the article Depression.

Introduction

> When I came out of the hospital … I fancied that I hadn't had anything, only afterwards did I have the feeling that I'd been ill. What can you say, I have moments when I'm twisted by enthusiasm or madness or prophecy like a Greek oracle on her tripod.
>
> Vincent van Gogh, 3 February 1889, in a letter addressed to his brother Theo

Mania is a serious mental disorder characterized typically by elevated mood, a quickening of speech and thoughts, a feeling of having more energy, and need for less sleep, grandiosity, occasionally heightened sexuality and irritability. *The Oxford Dictionary* (2010) describes mania as a "mental illness marked by periods of great excitement or euphoria, delusions, and overactivity." The word mania comes from the Greek word literally meaning 'madness.' Depression has been covered in the article 'Depression'.

History of Bipolar

In the first century BC, the Greek physicians suggested a connection between melancholia and mania. Just as they thought melancholia resulted from an excess of 'black bile,' mania was thought to result as an excess of 'yellow bile.' In the nineteenth century French doctors started thinking of mania and depression as being on the same continuum: 'la folie double forme' (Baillarger, 1854), and 'folie circularie,' a circular form of mania and depression (Falret, 1854). Although Falret (1854) had described mild forms of mania, Mendel (1881) was the first person to define 'hypomania' as a "form of mania which typically shows itself only in the mild stages, abortively, so to speak." Kahlbaum (1882) described cyclothymia as milder changes in mood shifts but not enough to become depression or mania. It was around this time when van Gogh produced his paintings, and future historians would comment on how they could tell when he was in his more elated and more depressed stages of his illness from his paintings alone (see **Figure 1**). This is made easier due to the large number of letters written by van Gogh, all dated and many detailing his state of mind and health, and therefore making it possible to match the paintings to the letters. The painting shown was completed in what was described as the most productive period of his painting career, and the use of the vivid colors and the content of blossoming trees are in sharp contrast with the darker colors shown in the article 'Depression'. Kraepelin (1921) then made the leap and separated mood disorders from others such as schizophrenia, introducing the term 'manic depressive.' Leonhard in 1957 used the word 'bipolar' to distinguish between people who experienced mania and depression, and 'unipolar' for those with depression only (Leonhard, 1957). Bipolar has now been adopted into modern psychiatric practice, and replaced the term 'manic depression.'

Diagnosis

There are two widely recognized contemporary classification systems in use worldwide for depression: the ICD-10 (WHO, 1992) and DSM-IV (APA, 1994), both of which have new versions due out in the near future. Operational definitions are provided for clinical and epidemiological research.

Table 1 shows the similarities and differences between the two main classification systems. Both systems have categories for single and recurrent episodes, and both have a category

Figure 1 Vincent van Gogh. Orchard with blossoming apricot trees. Painting, Oil on Canvas. Arles: March 1888. In Van Gogh Museum, Amsterdam.

Table 1 Classification of mania/bipolar disorder

ICD-10	*DSM-IV*
	Hypomanic episode
Manic episode	*Manic episode*
● Hypomania	● Mild
● Mania, without psychosis	● Moderate
● Mania, with psychosis	● Severe
	● Severe with psychosis
Bipolar affective psychosis	*Bipolar I and II disorders*
● Currently hypomanic	● Current (or most recent episode)
● Currently manic (with/without psychosis)	● Hypomanic
● Currently depressed (mild, moderate, or severe)	● Manic[a]
● Currently mixed	● Depressed
● In remission	● Mixed[a]
Cyclothymia	*Cyclothymic disorder*

[a]Excludes bipolar II disorder.

for cyclothymia, as a persistent instability of mood involving periods of mild depression and mild elation, but neither being severe or prolonged enough to meet the criteria for bipolar. Also both systems divide mania into whether or not it features with psychotic symptoms. In the ICD-10 (WHO, 1992), the authors write that people with cyclothymia may not come to the attention of medics, as the elevated periods may well be enjoyable to the individual.

Hypomania is more persistent (of at least 4 days) and more marked than the elation experienced in cyclothymia, but there are no hallucinations or delusions, and the symptoms do not tend to lead to a severe disruption of work or social rejection.

This is in contrast to mania, where normal social inhibitions are lost, and there is marked distractibility. This is in addition to the usual core symptoms of elation, increased energy (resulting in overactivity), pressure of speech (characterized by uninterruptability), increased self-esteem and optimism, and decreased need for sleep. There is a tendency to embark on extravagant and dangerous schemes, spend money recklessly, and make some inappropriate amorous advances. The episode should last at least 1 week before a diagnosis is made. If psychosis is present, it usually takes the form of delusions of grandiosity (e.g., having superpowers, being extremely wealthy, or related to royalty). Persecutory or religious delusions can also be present. 'Flight of ideas' can be present, whereby the individual flits from topic to topic with a tenuous connection, often ending up on an answer far removed from the question asked. Sims (1995) describes an example of flight of ideas in his book, of a female patient saying: "They thought I was in the pantry at home … Peekaboo … there's a magic box. Poor darling Catherine, you know, Catherine the Great, the fire grate, I'm always up the chimney. I want to scream with joy …. Hallelujah!" In severe cases, there may be a neglect of self-care, and of eating and drinking, and they may be placing themselves at risk from and to others.

A diagnosis of bipolar requires at least one depressive episode and at least one manic or hypomanic episode. DSM-IV further separates bipolar into:

- Bipolar I – where mania has occurred on one occasion
- Bipolar II – where hypomania has occurred (but not mania)

A 'mixed' episode of bipolar disorder can occur where a depressed mood may be accompanied by days or weeks of overactivity and pressure of speech; or conversely, a manic mood to be accompanied by agitation, loss of energy, and libido. Sometimes depressive symptoms and hypomanic or manic symptoms may alternate day to day or even hour to hour. According to the ICD-10, a diagnosis of mixed bipolar disorder should be made if the two sets of symptoms are both prominent for the greater part of the current episode, which should last at least 2 weeks (DSM-IV requires the episode to be at least 1 week).

Rapid cycling disorder can occur with both bipolar I and II disorders, and the diagnosis is made where there are at least four affective episodes (i.e., major depression, mania, hypomania, or mixed mood disorder) in the previous 12 months. Rapid cycling is more common in women, and unfortunately has a poorer response to treatment such as with lithium.

Any assessment of bipolar disorder needs to include the following:

- Detailed history of the episode (symptoms, presence of hallucinations or delusions, collateral history if possible)
- Any previous episodes of mania or depression
- Any suicidal or homicidal thoughts
- Any self-neglect
- Family history
- Substance misuse, smoking, and alcohol intake
- General physical health

In addition one needs to rule out the following that can present in similar ways to bipolar, but should not be misdiagnosed as such (Smith et al., 2011):

- Hyperthyroidism or hypothyroidism
- Anorexia nervosa
- Cerebrovascular event
- Dementia
- Other psychiatric disorders, for example, schizophrenia
- Acute drug withdrawal or illicit drug ingestion

Sims (1995) describes in his book about how the manic patient "looks at his world with complete, unshakeable, and wholly unjustified optimism." The person with mania frequently does not have insight into the illness, meaning that when others try and help slow things down, it often manifests in what Sims calls 'manic irritability', when "he feels all his brilliant ideas are frustrated by the lack of vision of everyone else" (Sims, 1995).

Epidemiology

Males and females are equally affected in bipolar disorder (but not in rapid cycling), in contrast to depression, which is more common in women. The lifetime prevalence of bipolar disorder varies from 0.3% to 1.5%. The average age of onset in Western countries is around the mid-20s, although the most common age of onset is the late teenage years, 15–19 years (Joyce, 1984). The second most frequent age range of onset is 20–24 years. Late-onset bipolar disorder is rare but it does occur.

Interesting Yutzy et al. (2012) report an increase in the prevalence bipolar affective disorders I and II, where from

mid-1970s to 2000 it ranged from 0.4% to 1.6%, and by the late 1990s to the 2000s, it had climbed from approximately 5% to 7%.

Course and Prognosis

Manic episodes last between 2 weeks and 4–5 months in bipolar disorder, although over the course of the disorder, there are many more episodes of depression than mania. Ninety percent of patients would experience more than one episode of mania. As patients become older, the time between episodes becomes shorter in duration. Unfortunately, only 20% achieve 5-year stability with good social and personal outcome (NICE, 2006).

According to Abreu et al. (2009) 25–56% present at least one suicide attempt during their lifetime and 15–19% die from the attempt. Lithium has been shown to reduce the risk of suicide and number of suicide attempts in bipolar disorder, but is less effective in rapid cycling bipolar disorder, which has a poorer prognosis (Tondo and Baldessarini, 2009).

There is an association between bipolar disorder and premature death. In a national cohort study of 6618 adults with bipolar disorder, Crump et al. (2013) found that those with bipolar disorder died about up to 9 years earlier on average than the rest of the population. They died prematurely from multiple causes, including cardiovascular disease, diabetes, chronic obstructive pulmonary disease, influenza or pneumonia, unintentional injuries, and suicide.

Several factors have been reported to indicate a better or worse prognosis for bipolar (Soreff and Ahmed, 2014).

Factors suggesting a worse prognosis include the following:

- Poor job history
- Alcohol abuse
- Psychotic features
- Depressive features between periods of mania and depression
- Evidence of depression
- Male sex

Factors suggesting a better prognosis include the following:

- Shorter manic phases
- Late age of onset
- Few thoughts of suicide
- Few psychotic symptoms
- Few medical problems

Etiology

Genetic Factors

Bipolar disorder is one of the most heritable of all psychiatric disorders. There is a genetic component to bipolar disorder, and the literature suggests the involvement of the following genes: ANK3, CACNA1C, and CLOCK genes (Sklar et al., 2008; Ferreira et al., 2008). These genes are thought to make someone susceptible to bipolar disorder, but like many other psychiatric disorders, there are other environmental factors that come into play. Further evidence of the gene–environment hypothesis comes from twin studies, where there is a concordance of 33–90% for bipolar I in identical twins (Soreff and Ahmed, 2014). If genes were the only etiological factor, one would expect 100% concordance as they share 100% of their genes. Conversely, adoption studies show that environment alone is also not the only factor. Children whose biologic parents have either bipolar I or a major depressive disorder remain at increased risk of developing an affective disorder, even if they are reared in a home with adopted parents who are not affected (Soreff and Ahmed, 2014).

A child of a parent with bipolar disorder has a 50% chance of having a major psychiatric disorder. First-degree relatives of people with bipolar I are approximately 7 times more likely to develop bipolar I than the general population (Soreff and Ahmed, 2014).

Biochemical Factors

Certain drugs such as cocaine that increase the levels of monamines (dopamine, serotonin, norepinephrine) in the brain can precipitate mania. Some authors have implicated the neurotransmitter glutamate in the etiology of affective disorders, and postmortems looking at the frontal lobes of those with affective disorders have shown elevated levels of glutamate (Hashimoto et al., 2007).

The hypothalamic–pituitary–adrenal (HPA) axis is a regulatory system that integrates neuronal and endocrine function, and manic and depressed bipolar individuals appear to have a dysregulated HPA axis. The HPA axis is linked to the serotonin system, sleep disturbance, and cortisol secretion. Exploring this further may give rise to other novel treatment strategies.

Neurophysiological Factors

A decrease in gray matter, a decreased activation in certain brain regions, and an increased activation in other regions (such as the ventral limbic area that mediate and generate emotional responses) have been discovered (Houenou et al., 2011), although more research in this area is needed.

Psychodynamic Factors

Melanie Klein proposed that mania was a defense against depression, and certain personality characteristics may be present in those with bipolar disorder (disturbances in neuroticism, extraversion, and openness) (Barnett et al., 2011). A dysregulation of self-esteem, childhood trauma, and abuse have been proposed as etiological factors for bipolar disorder (Leverich et al., 2002).

Environmental Factors

The gene–environment hypothesis dictates that the genes make an individual susceptible to the disorder, and the environment acts as an additive to cause the disease; different researchers have proposed differing amounts of interaction between the genes and the environment. A stressful circumstance can often act as a trigger for bipolar, and these could include relationship breakdown, abuse, death of a loved one, physical illness, prolonged sleep disturbance, or accumulative stress related to money, work, or relationships.

Cultural Factors

In the UK, Afro-Caribbean immigrants have been found to be at a greater risk of presenting as manic than their Caucasian counterparts (van Os et al., 1996), although this could be due to the stress of immigration or racial prejudice. In Figi, a condition called 'matiruku' has been reported, described as a 'periodic insanity' (Price and Karim, 1978). It is characterized by increased talkativeness, increased psychomotor activity, elated mood, a decreased need for sleep, and occasional violence. These episodes would occur at the same time every month, or every year, and would last about 2–7 days. Diagnoses could be influenced by the self-report of individuals belonging to certain communities, giving skewed data. For example, in Amish communities, overly inflated self-esteem is considered to be a grave sin. In both prospective and retrospective studies carried out in China and Israel in individuals with bipolar disorder, there has been a preponderance of recurrent manic episodes over depressive episodes. This is the opposite to what has been found in other areas across Europe and other Western countries. Some authors suggest that climactic factors may play a part in the course for bipolar disorder (Sanches and Jorge, 2004).

Treatment and Management

The following is largely based on the current National Institute for Health and Clinical Excellence (NICE) guidelines in the UK for adults, children, or adolescents with bipolar disorder (NICE, 2006), and the Evidence-based guidelines for treating bipolar disorder: revised second edition from British Association for Psychopharmacology (Goodwin, 2009). NICE provides national guidance and advice to improve health and social care, and the recommendations are based on the best available evidence, also taking into account of cost–benefit ratio.

Non-pharmacological Treatment

Of paramount importance is the need to establish a trusting therapeutic relationship between the individual with bipolar, and the carers and health-care professionals. The following are examples of non-pharmacological treatments that can help in bipolar (Patient.co.uk, 2012).

- Education regarding diagnosis, treatment, and side-effects
- Good communication
- Self-help groups
- Support groups
- Self-monitoring of symptoms, side-effects, and triggers
- Coping strategies
- Psychological therapy
- Encouragement of engagement in calming activities
- Telephone support
- Diet – adjunct use of omega-3 fatty acids could help bipolar depressive but not bipolar manic symptoms (Sarris et al., 2012)
- Regular exercise
- Psychological interventions such as cognitive behavioral therapy and supportive psychotherapy (Goodwin, 2009)

Pharmacological Interventions

Treatment of Acute Manic or Hypomanic Symptoms

The drug treatment of an acute manic or hypomanic episode depends on the severity of symptoms and clinicians should be guided by current medication doses and previous response (NICE, 2006). When the NICE (2006) guidelines were written only lithium, olanzapine, quetiapine, risperidone, and valproate semisodium were licensed for the treatment of acute mania in the UK. Lithium, olanzapine, or valproate should be considered for long-term treatment (2–5 years depending on risk factors) of bipolar disorder. Lithium requires monitoring of levels and monitoring of renal and thyroid function.

Pharmacological agents approved by the FDA, however, for use in treating bipolar disorder are as follows (taken from Sorref and Ahmed, Medscape, updated 2014):

- Valproate (manic)
- Carbamazepine, extended-release (manic, mixed)
- Lamotrigine (maintenance)
- Lithium (manic, maintenance)
- Aripiprazole (manic, mixed, maintenance)
- Ziprasidone (manic, mixed)
- Risperidone (manic, mixed)
- Asenapine (manic, mixed)
- Quetiapine (manic, depression)
- Chlorpromazine (manic)
- Olanzapine (manic, mixed, maintenance)
- Olanzapine–fluoxetine (depression)

Figure 2 shows the initial treatment plan for mania or mixed bipolar episode as devised by the British Association of Psychopharmacology (Goodwin, 2009).

Management of Depressive Symptoms

This is the same as in the management in unipolar depression, and is included in the articles on depression and antidepressants, and hence has not been included here. Of note, however, there is a risk of switching to a manic episode with antidepressants, and thus caution is advised and careful monitoring is necessary. In bipolar, antidepressant may not be as effective, and a combination therapy (antidepressant with an anti-manic agent) may be required.

Management of Acute Mixed Episodes

NICE (2006) advises that antidepressants should be avoided, and prescribers should consider treating patients with an acute mixed episode as if they had an acute manic episode (see above). Patients should be monitored closely for risk of suicide.

Management of Rapid Cycling

Previous treatments need to be reviewed, and the prescriber needs to check for compliance with medication. If they are on antidepressants, then this needs to be stopped. Thyroid function should be checked, and anti-manic therapy will need to be optimized. Lithium and valproate is used first line, but if this does not work, lithium can be trialed on its own (NICE, 2006). Lithium levels would need to be

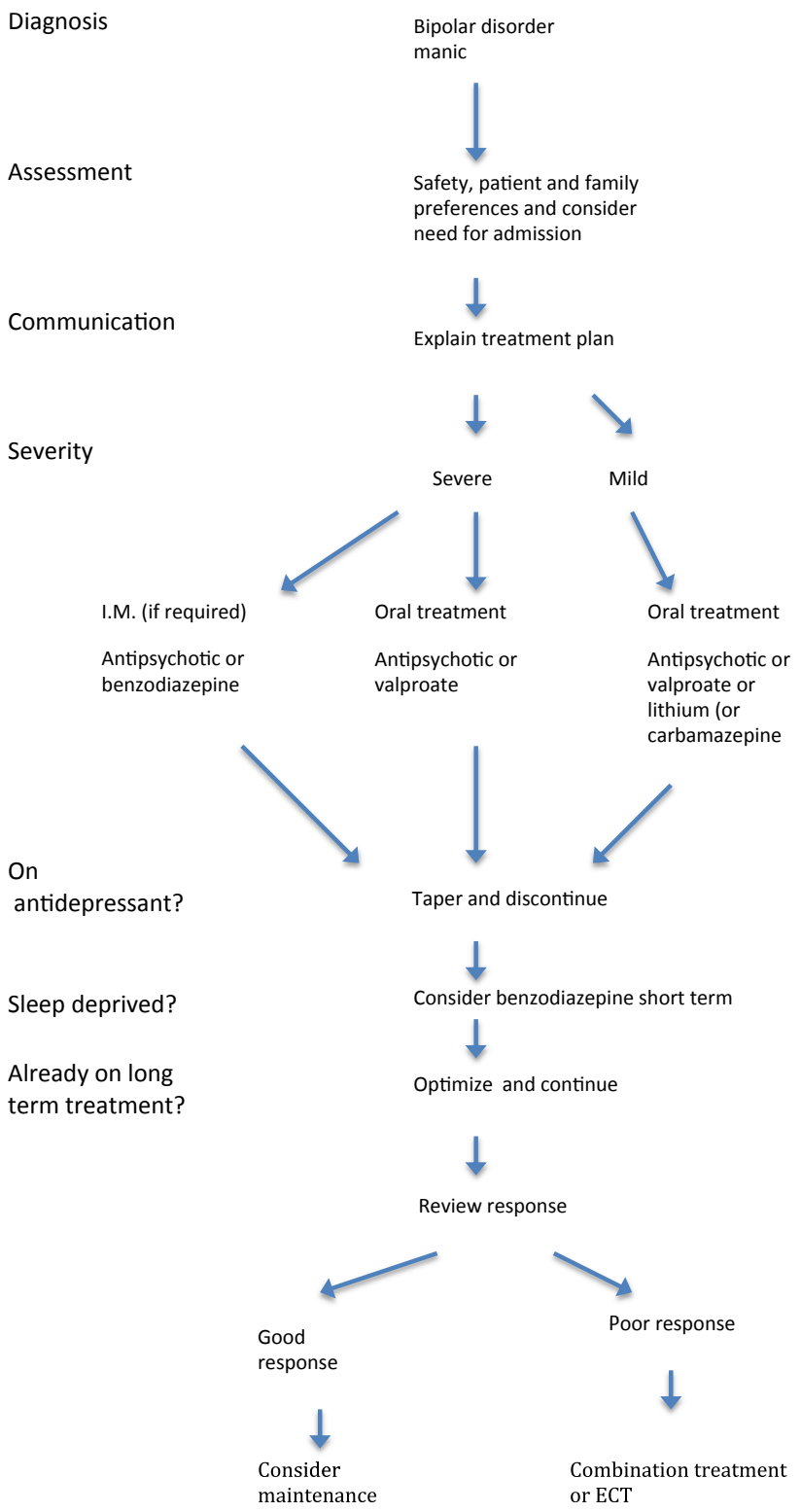

Figure 2 Evidence-based guidelines for treating manic or mixed bipolar episode. Goodwin, 2009.

checked, as both lithium withdrawal and lithium toxicity can cause rapid cycling. Psycho-education (in addition to keeping a 'mood diary' detailing aspects of their symptoms, when they occur and anything that helps) should be encouraged. The focus should be on optimizing the long-term treatment instead of focusing on each individual episode. Medication trial should last for more than 6 months (NICE, 2006).

Electroconvulsive Treatment in Bipolar

Electroconvulsive treatment (ECT) should only be considered in severe cases, with the aim of obtaining short-term and quick resolution or improvement of the symptoms. ECT should be used only when pharmacological and psychological treatment has been tried and failed. ECT can also be used if the situation is thought to be life threatening in those with severe depression, catatonia, or a prolonged or severe manic episode (NICE, 2006). The effects of ECT are thought to be short-lived.

Summary

In recent times, more people in the media and public eye are declaring their diagnoses of bipolar. This has undoubtedly increased the awareness of the condition, with a hoped decrease in stigma. Some have emphasized the positive aspects of having bipolar; for example there is evidence that it is strongly correlated to creativity and literacy (Santosa et al., 2007). Others who have experienced bipolar have commented on the enjoyable experience of having the increased energy and mood during the elated phase. This has even had the effect of a perceived increase in individuals coming to their health care specialists with a self-diagnosis of bipolar, which could be due to the increased awareness, and decreased stigma, but may also be due to the "lay person's aspirations for higher social status, as illustrated by the implicit association of bipolar disorder with celebrity status and creativity" (Chan and Sireling, 2010). The question arises about whether some of the successful people we see with bipolar would have been as talented if they did not have their illness. In one of his last letters van Gogh wrote: "If I could have worked without this accursed disease, what things I might have done." The treatment available today was not present then, and we simply do not know if his brilliance as an artist was because of his mental illness or in spite of it. What we do know is that the world is undoubtedly better off for having his paintings.

See also: Antidepressant Drugs; Depression; Mental Disorders, Epidemiology of; Mentally Ill, Public Attitudes towards; Psychopharmacotherapy: Side Effects; Psychotherapy and Pharmacotherapy; Schizophrenia; Suicide.

Bibliography

Abreu, L.N., Lafer, B., Baca-Garcia, E., Oqeundo, M.A., 2009. Suicidal ideation and suicide attempts in bipolar disorder type I: an update for the clinician. Revista Brasileira de Psiquiatria 31 (3), 271–280.
American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders, fourth ed. (DSM-IV). APA, Washington, DC.
Baillarger, J., 1854. Note sur un genre de folie. Bulletin de la Academie Imperiale de Medicine (Paris) 19, 340–352.
Barnett, J.H., Huang, J., Perlis, R.H., Young, M.M., Rosenbaum, J.F., Nierenburg, A.A., Sachs, G., Nimgaonkar, V.L., Miklowitz, D.J., Smoller, J.W., 2011. Personality and bipolar disorder: dissecting state and trait associations between mood and personality. Psychological Medicine 41 (8), 1593–1604.
Chan, D., Sireling, L., 2010. 'I want to be bipolar'… a new phenomenon. The Psychiatrist Online 34, 103–105.
Crump, C., Sundquist, K., Winkleby, M.A., Sundquist, J., 2013. Comorbidities and mortality in bipolar disorder: a Swedish national cohort study. JAMA Psychiatry 70 (9), 931–939.
Falret, J., 1854. Mémoire sur la folie circulaire. Bulletin de l'Académie Imperiale de Médecin (Paris) 19, 382–400.
Ferreira, M.A., O'Donovan, M.C., et al., 2008. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nature Genetics 40 (9), 1056–1058.
Goodwin, G.M., 2009. Evidence-based guidelines for treating bipolar disorder: revised second edition- recommendations from the British Association for Psychopharmacology. Journal of Psychopharmacology 23 (4), 346–388.
Hashimoto, K., Sawa, A., Iyo, M., 2007. Increased levels of glutamate in brains from patients with mood disorders. Biological Psychiatry 62 (11), 1310–1316.
Houenou, J., Frommberger, J., Carde, S., Glasbrenner, M., Diener, C., Leboyer, M., Wessa, M., 2011. Neuroimaging-based markers of bipolar disorder: evidence from two meta-analyses. Journal of Affective Disorders 132 (3), 344–355.
Joyce, P.R., 1984. Age of onset in bipolar affective disorder and misdiagnosis in schizophrenia. Psycholigical Medicine 14, 145–149.
Kahlbaum, K.L., 1882. Uber cyclisches Irresein. Der Irrenfreund 10, 145–157.
Kraepelin, E., 1921. Manic Depressive Insanity and Paranoia (R.E. Barclay, Trans., G.M. Robertson, Ed.). Livingstone, Edinburgh, UK.
Leonhard, K., 1957. Aufteilung der endogenen Psychosen. Akademie.
Leverich, G., McElroy, S.L., Suppes, T., Keck Jr., P.E., Denicoff, K.D., Nolen, W.A., Altshuler, L.L., Rush, A.J., Kupka, R., Frye, M.A., Autio, K.A., Post, R.M., 2002. Early physical and sexual abuse associated with an adverse course of bipolar illness. Biological Psychiatry 51, 288–297.
Mendel, E., 1881. Die Manie. Urban and Schwazenburg, Vienna.
NICE, 2006. The Management of Bipolar Disorder in Adults, Children and Adolescents, in Primary and Secondary Care. Available at: http://www.nice.org.uk/guidance/cg38/chapter/guidance (last accessed 06.08.14.).
van Os, J., Takei, N., Castle, D.J., Wessely, S., Der, G., MacDonald, A.M., Murray, R.M., 1996. The incidence of mania: time trends in relation to gender and ethnicity. Social Psychiatry and Psychiatric Epidemiology 31 (3–4), 129–136.
Patient.co.uk, 2012. Bipolar Disorder. Available at: http://www.patient.co.uk/doctor/bipolar-disorder#ref-2 (last accessed 08.05.14.).
Price, J., Karim, I., 1978. Matiruku, a Fijian madness: an initial assessment. British Journal of Psychiatry 133, 228–230.
Puri, B.K., Hall, A.D., 2004. Revision Notes in Psychiatry, second ed. Arnold, London.
Sanches, M., Jorge, M.R., 2004. Transcultural aspects of bipolar disorder. Revista Brasileira de Psiquiatria 26 (Suppl. III), 54–56.
Stevenson, Angus (Ed.), 2010. Oxford Dictionary of English, third ed. Oxford University Press.
Santosa, C.M., Strong, C.M., Nowakowska, C., Wang, P.W., Rennicke, C.M., Ketter, T.A., 2007. Enhanced creativity in bipolar disorder patients: a controlled study. Journal of Affective Disorders 100, 31–39.
Sarris, J., Mischoulon, D., Schweiter, I., 2012. Omega-3 for bipolar disorder: meta-analyses of use in mania and bipolar depression. Journal of Clinical Psychiatry 73 (1), 81–86.
Sims, A., 1995. Symptoms in the Mind. An Introduction to Descriptive Psychopathology, second ed. W.B. Sauders Company Ltd, London.
Sklar, P., Smoller, J.W., Fan, J., Ferreira, M.A., Perlis, R.H., Chambert, K., et al., 2008. Whole-genome association study of bipolar disorder. Molecular Psychiatry 13 (6), 558–569.
Soreff, S., Ahmed, I., Updated 2014. Bipolar Affective Disorder. Medscape. Available on: http://emedicine.medscape.com/article/286342-overview (last accessed 06.08.14.).
Smith, D.J., Griffiths, D.J., Kelly, M., Hood, K., Craddock, N., Simpson, S.A., 2011. Unrecognised bipolar disorder in primary care patients with depression. British Journal of Psychiatry 199, 49–56.
Tondo, L., Baldessarini, R.J., 2009. Long-term lithium treatment in the prevention of suicidal behavior in bipolar. Epidemiologia e Psichiatria Sociale 18 (3), 179–183.
World Health Organisation, 1992. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines. WHO, Geneva.
Yutzy, S.H., Woofter, C.R., Abbott, C.C., Melhem, I.M., Parish, B.S., 2012. The increasing frequency of mania and bipolar disorder: causes and potential negative impacts. Journal of Nervous and Mental Disease 200 (5), 380–387.

Relevant Websites

www.dballiance.org – Depression and Bipolar Support Alliance.
www.ibpf.org – International Bipolar Foundation.
http://emedicine.medscape.com/article/286342-overview#showall – Medscape.
www.NICE.org.uk – National Institute for Health and Clinical Excellence.
http://www.patient.co.uk/doctor/bipolar-disorder#ref-2 – Patient.co.uk. Bipolar Disorder.
http://www.rcpsych.ac.uk/healthadvice/problemsdisorders/bipolardisorder.aspx – Royal College of Psychiatrists information on bipolar.
http://bipolarhelpvideos.com/tedxterrytalks-laura-bain-living-with-bipolar-type-ii/ – TED Talks. Laura Bain – Living with Bipolar II.

Mann, Michael (1942–)

Ralph Schroeder, Oxford Internet Institute, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Michael Mann is a comparative-historical sociologist whose main work is the four volume 'Sources of Social Power.' This work provides a sociological history from the beginning to the present day, giving an account of long-term patterns of ideological, economic, and military power conceived of as overlapping and intersecting socio-spatial networks. He is particularly well-known as a political sociologist.

Michael Mann is a comparative-historical sociologist best known for the four volumes of his 'Sources of Social Power' (1986, 1993, 2012, 2013a, henceforth 'Sources'). This four volume work spans history from ancient times to the present day, distinguishing between four main types of power: ideological, economic, military, and political. Mann charts the course of these four types of power, conceived as networks, at the leading edges of social development across time and space. His approach to understanding social change, analyzing major macro-patterns in society over time, updates the tradition of the classical sociological theory of Marx and Weber. Like the works of these classical theorists, 'Sources' deals, as it moves to the present, with the origins and trajectories of modern capitalism; and this question is also still central to sociologists interested in tackling social change from a macro-historical perspective.

Mann was born in 1942 and grew up outside of Manchester in lower middle-class circumstances (Hall, 2006). He did an undergraduate degree in history at Oxford University and continued there to do his doctorate with an empirical study about industrial workers. At this time, the late 1960s and early 1970s, there was much debate about the role of ideology and class consciousness among the working class. Mann argued, in his first major journal article (1970), that ideology did not provide the 'glue' of capitalist society, nor did the alienation of workers indicate the potential for revolutionary change. This is one background to his sociology. The second was the idea, prevalent in Marxist sociology at that time, that the role of the state, which lacked autonomy, was to legitimate the economic conditions of capitalism. This idea, again, was rejected by Mann (1984) when he argued that the state came to constitute an autonomous source of power, operating differently from other sources by exercising centralized control over a particular territory, with its own dynamic.

While Mann's early work developed against the background of Marxist sociology and he departed from some of its main tenets, his politics have consistently been on the left or radical end of the spectrum. The Marxist debates, which continued into the 1980s, are now largely forgotten, displaced in the social sciences by the 'cultural' turn. Yet Mann started writing 'Sources' at a time when the major figures in social thought were the Neo-Marxists and Frankfurt School thinkers – or their positivist opponents. He taught sociology at the University of Essex in the 1970s before spending most of the subsequent decade at the London School of Economics, where at the time Ernest Gellner (1988) and John A. Hall (1986) were also working on sociological histories of power. They jointly and for many years convened a seminar called 'Patterns of History,' to which they invited a wide range of speakers from across the social science disciplines. This seminar was used to inform their comparative-historical analyses. The synthesis of many fields, including archaeology, ancient history, and anthropology, was evident in the first volume of 'Sources,' published in 1986, and has been in evidence throughout Mann's work. Mann then moved to the department of sociology at the University of California, Los Angeles in 1987, where he has remained since.

Apart from 'Sources,' Mann has written about geopolitics and the threat of nuclear war during the height of the Cold War in the 1980s (1988a), a comparative analysis of 'Fascists' (2004), and 'The Dark Side of Democracy' (2005), about the major instances of ethnic cleansing – including the aftermath of the breakdown of the Soviet Union and subsequent Balkan wars and the war in Rwanda in the 1990s. He also wrote 'Incoherent Empire' (2003; see also his conversations with John Hall, Mann, 2011), criticizing American foreign policy against the backdrop of the wars in Iraq and Afghanistan in the 1990s and 2000s. But apart from these forays into other topics, his main work has been 'Sources.' Now that this is completed with the final volume, subtitled 'Globalizations' (2013a), he has brought his account of the development of power up to the present day. And although he is planning further work refining and drawing out the implications of 'Sources,' these four volumes provide an indepth sociological account of the shape of history.

It is impossible to present anything more than a few arguments from the four volumes, which together run to more than 2000 pages. Before doing so, it is worth briefly elaborating the conceptual apparatus of 'Sources.' First, Mann thinks that societies are not 'systems' or unified wholes; rather 'societies are constituted of multiple overlapping and intersecting socio-spatial networks of power' (1986: 1). This implies that networks of power have different spatial extents or reaches: for example, the reach of contemporary American military power is global, even though it also confronts other military powers; and this network is different from the reach of capitalist markets, which cover even greater geographical reaches of the globe; and different again from political economies of welfare, which are confined to nation-states. Secondly, power, apart from being of four types, is not just 'power over,' or the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61240-0

'distributive' power of A over B (as Weber argued), but also 'power through,' joint power over third parties or over nature, which (following Parsons) Mann labels 'collective power' (1986: 6).

Thirdly, power is also divided into extensive versus intensive, and can be authoritative or diffused. These distinctions concern the organizational reach of power (1986: 9), and Mann gives the following examples for the combinations of each of these four: an army command structure is authoritative/intensive (tight control over troops), whereas a militaristic empire is authoritative/extensive (control over far-flung territories). A general strike is intensive/diffused (generating strong but not concentrated solidarity) and market exchange is extensive/diffused (consisting of lots of indirectly connected interactions over large spaces). These distinctions allow him to characterize the socio-spatial networks of all four sources, including the most geographically dispersed macro-relations of civilizations and empires; however, Mann also shows that the most spatially circumscribed power containers can provide solidarity for workers in factories or impose control over troops ranged on battlefields.

Against this background, it is possible to present some of the main arguments in the four volumes, with a particular focus on where they place Mann in contemporary sociological debates. Volume One, subtitled 'A history of power from the beginning to AD 1760,' culminates in his explanation of "the rise of the West" but it begins with the emergence of the settled societies in ancient Near East, where the need to protect an agricultural surplus required centralized power – or the 'caging' of people within a territory subject to centralized political authority. This prefigures the importance of military-cum-political power in his work, as opposed to economic class relations. Mann does not ignore the dynamics of class, especially in his account of the Greco-Roman world. But his explanation is also more state-centric, and remains so throughout the four volumes. Indeed, he is most closely associated with the 'state-centered' (Collins, 1999) school of Jack Goldstone, Theda Skocpol, and Charles Tilly, who argue that war and the preparation for war and the resulting increase in state capacity is central to the understanding of the rise of the West. This school has concentrated on modern Europe, but Mann's analysis of the military source of power also extends to other parts of the world and to other periods.

Like other 'state-centered' theorists, Mann argues that military technologies are often at the leading edge of increases in social power. Yet he does not focus on technology per se, but rather on how it organizes power in socio-spatial networks. One example is that technology controls the logistical reach of military power, such as the need to supply transport animals with food (1986: 134–142). This allows him to argue that this reach was more limited than is commonly thought. The role of the socio-spatial organization also sets his explanation of the rise of the West apart from other power networks. A key factor here, on which Volume One ends is the role of ideology, in the shape of the 'normative pacification' provided by medieval Christendom (1986: 506). Here Mann departs from Weber's explanation which focused on Protestantism and on the individual's ethic of asceticism. Instead, Mann's explanation centers on the integrating organizational force of Catholicism in late medieval Europe.

Another departure here from Weber is that although he devotes a chapter to an excursus into comparisons with the other world-religions, Mann argues that the evolution of power is more important than a comparison to ascertain which of the major civilizations achieved a breakthrough. In other words, it is more important to explain how the leading edge of power has moved on a path of increasing over time – rather than explaining why it increased in one place (the West) and not in others (the more stable 'rest'). This, too, sets him apart from other social thinkers. If we fast-forward to the present day for a moment, we can see that there is an advantage to Mann's approach in analyzing different spans of power over different periods of time: his framework whereby the spatial organization of military power is separate from the reach of economic power means that he is able to argue that the United States can remain the globally dominant military power for several decades to come – even when the leading edge of economic power may have shifted elsewhere, and particularly toward East Asia. In other words, one type of power can still be 'leading' even while other powers undergo comparative shifts.

Volume Two of 'Sources,' subtitled 'The Rise of Classes and Nation-States, 1760–1914,' concentrated on the role of the state during the nineteenth century, which covers the period from the French and American Revolutions to the First World War. The two revolutions are also the second major instance in history, apart from the unifying role of the church in the Middle Ages, of the leading role of ideological power. The third major ideological influences were in the twentieth century – the Fascist and Communist revolutionary ideologies. For the nineteenth century, a lasting contribution of Volume Two is the detailed account of comparative state capacity. Mann argues that especially in the late nineteenth century, some states expanded their infrastructural power more rapidly than their despotic power, where infrastructural power is the collective power of the state "through" society "coordinating social life," whereas despotic power entails the distributive power of state elites over civil society (1993: 59). The analysis is based on detailed data that he compiled about state spending, which shows that in some states more than others, civilian spending increased faster than military spending, which had previously been the bulk of state spending.

His account of the nineteenth century is therefore unique in giving pride of place to the state: the dominant accounts of this period focus on the industrial revolution and on technological innovations. Mann, in contrast, argues that in this period, the state caged citizens into nation-states, and these nation-states then went on to become collective actors internationally, with rival ideologies and political economies, in the twentieth century. Still, at the core of Volume Two is therefore a puzzle (recalling that Mann's explanation is state-centered and focuses on military competition): this was a relatively peaceful period in world history, yet it was also a period in which state capacity increased at an unprecedented pace. This puzzle can be resolved when it is recognized, first that the increase was mainly in infrastructural rather than despotic power. Mann thus departs from Tilly (1990) in highlighting the reduced contribution of the state to the military and its increasing role in creating

civilian infrastructures. Secondly, states took on quite different shapes during this period, some (Germany, Russia) becoming more state-centric, while in others (Britain, the United States) markets played a greater role. It can be added that although the nineteenth century was relatively peaceful in Europe, this can be contrasted with the often violent European conquests of colonies and by the settler states in America and Australia. In any event, the division of major powers into at least two kinds of states paves the way for explaining authoritarian paths in the twentieth century.

Volume Three, subtitled 'Global Empires and Revolution, 1890–1945,' goes back into the nineteenth century to revise Mann's account in the earlier volumes somewhat. Mann now argues that empires, which he did not discuss in Volume Two, become central to the twentieth century. Thus apart from the two authoritarian paths prefigured in the nineteenth century, Volume Three charts the rise and fall of the British, Japanese, American, and other empires. Imperialism also enabled Mann to add racism as a major modern political ideology. In Mann's account of the struggle between rival empires, there are echoes of the 'state-centered' school. However, Mann also thinks that the World Wars represent a culmination of intra-European warfare, with the conflict between Japan and China added in the Second World War. However, the result of the wars was the end of European imperial dominance. This was partly due to the contradiction between a racist ideology, which denied rights to colonial subjects, coupled with the need to use soldiers from the colonies during the World Wars, in turn fueling their nationalism. These factors contributed to the downfall of the colonial empires after the Second World War – except for the American empire, where the racism of denying rights to African-Americans had been a feature of domestic politics rather than of America's overseas engagements.

Another place in Volume Three where Mann departs from conventional ideas is in his explanation of the Great Depression, where he focuses not just on finance (as many others do), but rather on a series of underlying tensions: first, in geopolitics, there was a lack of international cooperation resulting from the First World War, partly due to the vacuum left because Britain no longer played a hegemonic role and the United States was not yet willing to take on this role. Secondly, in agriculture, there was overproduction and consequently, not enough demand for the new consumer goods that could now be mass produced. This approach to understanding an economic crisis by reference to deeper structural causes, also marks Mann's explanation of the Great Recession of 2008 in Volume Four, which was not just caused by financial markets but also by the lack of consumer demand in societies, especially the Anglophone ones (the United States, Britain, and other English-speaking countries where economic power is comparatively stronger than the other state), partly due to increasing inequality and with it the lack of broad-based consumer demand (the few rich are not intensive consumers). This underlying factor contributed to financial instability, ultimately affecting whole economies. Still, the lack of widespread demand and its resulting inequalities meant that the 2008 Great Recession affected the (white) Anglophone countries the worst, just as the Great Depression could be labeled the 'Great White Depression' (2013b: 78) since it mainly affected Europe and North America; other parts of the world were affected only as these crises rippled outwards.

Volume Three also continues the explanation of the divergence between different kinds of capitalist welfare states (this was already prefigured in Volume Two). A major argument here is that the United States was in fact on course to a welfare state during the 1930s, just as other countries were. Put differently, it was not yet on an exceptional course. Instead, there was under Roosevelt a compromise between capitalists and workers which created employment rights and some social insurance as a means to counteract the effect of the Depression. In Sweden, in contrast, which is often seen as the quintessential model of welfare state development, the development of the welfare state was still embryonic at this time. Here again war played a decisive role for the United States: after the Second World War, since there was a boost to economic growth and little unemployment; only a weak welfare state emerged in the United States, which confined benefits to soldiers (as opposed to other European powers, where whole societies had waged war, and so more universal welfare states were created). Thus, at the end of Second World War, with highly progressive taxation and a large state that was mobilized for the war effort no longer needed in America – unlike in other countries where a stronger state was required for economic recovery and where citizens also had to be rewarded for their war efforts – the American welfare state got off to a comparatively weak start. It can be seen that war and military power, for the most part overlooked by sociologists, is a distinctive feature of Mann's sociology, and it can play a positive as well as a destructive role.

To explain the different paths of welfare states, Mann draws on T.H. Marshall's concept of citizenship. Mann uses the three types of citizenship – civic, political, and social – to analyze the rights that people obtained from the ruling classes within the nation-state (1988b). Thus he goes beyond explaining conflict in capitalist societies solely by reference to economic classes. The paths whereby these rights were achieved now define the varieties of capitalism. But while civil and political rights have been gained across democratic nation-states, the inclusiveness and depth of social citizenship differs substantially across Western democracies and beyond. As we have seen, war has played a major role in the trajectory of citizenship rights. Another example here could be how women's suffrage (political rights) was achieved early in White Settler societies, where women had stronger roles in the conquest of the frontier; and after the First World War political rights were expanded to include women in countries where they had substantially supported the war effort. Again, an interesting implication is that unless there is either a major upheaval, such as war or a crisis that impels classes or other social groups to seek change in the social contract, these rights will not be extended further. This partly accounts for Mann's rather pessimistic view that social citizenship rights have stalled in recent decades, leaving the varieties of citizenship regimes intact.

After the end of Second World War, only two empires remained and since the end of the Cold War, only one has been left. In Volume Four (2013a), four of thirteen chapters are devoted to American empire and American exceptionalism; and America plays a large part in several other chapters as well.

There are many others who have written about American empire or America's role in the world; and here too Mann departs from other views inasmuch as he focuses on the military power of the United States, rather than on how empire is driven by the needs of capitalism to protect or expand markets (as posited by Marxists and world-systems theorists). Further, he carefully weighs the benefits and costs of the American 'informal' empire (as he does with the other empires in Volumes Three and Four); and while there have been some benefits to the United States, especially since the dollar has been the world's reserve currency ('seigniorage'), on balance, maintaining a military empire has been a drain on American power. The United States has nevertheless also had some gains from informal empire, including the facility of being able to run large trade deficits in return for the security provided for its military allies. The balance sheet of those at the receiving end of American domination abroad, on the other hand, has been much more negative. Several countries that were fought over *during* the Cold War were decimated by American military power; and Iraq and Afghanistan have had the same fate *after* the Cold War, when the United States has been the only developed democracy initiating interstate wars.

Another chapter in Volume Four that deserves to be highlighted is on climate change, where a key part of Mann's explanation is the rise of consumer citizenship. At a time when social citizenship (or the welfare state) is only being defended rather than deepened by the major powers, it is also being displaced by the idea that the benefits of market growth is among the basic rights of citizens – as consumers. This type of citizenship is taking hold not just in the West, but also in China, which, together with America, must be part of any solution for the mitigation of the potentially disastrous consequences of climate change. These two powers have the largest economies (and the United States also has military power), while the parts of the Global South, which are most adversely affected by climate change, have little power in either sense. Mann argues that the solution to this crisis will be political, with agreement between the world's nations not to focus so much on growth and economic benefits, curbing capitalist profit; as well as from the side of citizens as consumers. In other words, the solution will involve the most powerful institutions in the twentieth and twenty-first centuries – nation-states, capitalism, and citizenship. Yet we have already seen that the weaknesses of economies based on consumer demand contributed to the Great Recession of 2008, when there was not a large enough base of prosperity to sustain growth, due to increasing inequality. Here of course there is a contradiction, since growth is needed to avoid severe economic downturns – and yet is also environmentally disastrous.

The focus on empires, as well as on nation-states and capitalism, are also the key to Mann's thinking about 'globalizations.' He pluralizes the term in the subtitle for his fourth volume because all three of these (empires, nation-states, and capitalism) have become globalized in the course of the twentieth century. Capitalism won out over rival ways of organizing economies, particularly state-directed ones, and even though there are still varieties of capitalism, market-capitalism has extended to most parts of the globe. The nation-state has also become the norm across the globe: here, of course, there is great variety, but nowadays, states are only legitimate when they represent their people. The definition of people (or citizens) in terms of a hierarchy or exclusiveness of races (as with the Fascist powers) was defeated in the Second World War, and only China among the developed societies still has a single-party state claiming to represent the interests of its people. Otherwise, the model of democratic nation-states has spread throughout the North and beyond. Finally, although empires competed with each other and thus fractured the globe before the Second World War, they were consolidated into two factions during the Cold War; and now all that remains is the American informal economic empire shielded by the global reach of its military power. Although another global force and crisis, climate change, and possibly yet another, nuclear war (as Mann notes, in keeping with the centrality of military power in his work), threaten the world, political and economic globalizations are still ongoing.

In bringing 'Sources' up to the present, Mann offers a sophisticated theory of globalization. His final two volumes set him apart from other sociologists in providing a multidimensional account of contemporary capitalist democracies. Most other accounts of globalization and of contemporary capitalism are reductionist in focusing primarily on the economic dimension. Mann is also among the few sociologists to incorporate military power as part of his overall analysis rather than within the subdiscipline of military sociology or within international relations. The same goes for his analysis of climate change, which goes beyond the sociology of the environment or of risk. At the same time, Volumes Three and Four also engage with mainstream debates in sociology, especially in analyzing the varieties of capitalism (or 'political economy'), and also in social movement theory and theories of revolution. As we have seen, Mann's addition of military power plays a crucial role, for example, in his account of the varieties of capitalism and his explanation of why the American welfare state did not grow like the others.

Mann's work is also unique in focusing attention on the nation-state, arguing against the idea of its demise, which is popular among globalization theorists. Instead, his multiple 'globalizations' have different spatial reaches and the four sources of social power also operate differently: for example, while the political power and social citizenship rights of workers are 'caged' by the authoritative power of the nation-state, the diffuse economic power of capitalists, on the other hand, and especially of finance capital, is at least partly transnational. Hence Mann argues that the four sources have become 'orthogonal' to one another, which has benefits in avoiding dangerous concentrations of power, but also drawbacks insofar as it can result in stalled social development. In any event, Mann's sociology is distinctive in theorizing the transnational or global level, not just with military power, but also in charting shifts, for example, in the economic power of financial capital.

While Mann rejects deterministic accounts of social change which rely on a single source of power, he nevertheless wants to provide an encompassing theory. Yet a major current in recent social thought has been to argue against such encompassing theories or 'grand narratives.' And it is noteworthy that there have recently been no sociological theories or analyses of the macro-social processes of a scope similar to Mann's 'Sources.'

As we have seen, his grand narrative (if it can be called that) departs in important ways from other analyses of states and economies. At the same time, his four sources, operating differently and across different spatial extents over various periods, do not make for a single reductive story about the evolution of society. Thus one criticism of his work is that he is too historical and not analytical enough. And it is true that there is no single central thrust to his anatomy of contemporary social change. Perhaps the main reply to this criticism is that with globalizations, the processes of social change have become very complex, and only a multidimensional and deeply historical analysis can give this complexity its due.

Another way to highlight the distinctiveness of his sociology is to contrast him with other contemporary social theorists: Mann rejects Neo-Marxist or world-systems theories which see American empire as a way to secure capitalist interests across the globe. This puts him at odds with thinkers such as Wallerstein (2013), who foresee the demise of American power with a shift to economic power to Asia, and China in particular. For Mann, capitalism is not the only force determining social change, and thus American military power can remain unrivalled for decades to come, even while its economic power wanes. Yet Mann also rejects liberal thinkers like Gellner (1996) and Hall (2013), who think that the diffusion of capitalism and civil society (or the growth of middle classes) will, in due course, yield less authoritarian regimes. Mann, in contrast, thinks that there is a variety of capitalisms, and some with stronger states – including single-party China – have coped better with the recent economic crisis while at the same time being on a path toward greater economic inequality.

Another contrast that can be made is with international relations thinkers, who are split into realists and liberals. Realists would focus on the geopolitical or military interests of the United States, ranged against the rise of the new military power, China, in the East. Liberals would prioritize how the US and its allies continue to uphold a binding normative order that promotes democratic regimes and opens markets bringing economic growth. Yet Mann wants to see military power not as a separate order of interstate relations but analyzes these relations in terms of the positive and negative impact they have on domestic social development. At the same time, it is not just norms that uphold the international order, nor does economic growth diffuse evenly with open markets: there are also power imbalances in how financial regimes are imposed on the global South and economic inequalities are growing in some regions even while declining in others (Mann and Riley, 2007). Here, too, Mann departs from the dominant orthodoxies, in this case from the main positions in the field of international relations.

A criticism of Mann's work is that he downplays ideological power (see the chapters by Gorski and Bryant in Hall and Schroeder, 2006). There are three defences against this criticism: first, that ideological power has no force without being mobilized in organizational form, either extensive or intensive. In this respect, his theory can be labeled 'organizational materialism.' Secondly, he has added racist ideology to his explanation of empire and ideology has played a decisive role on a number of historical occasions, including modern revolutions (French, American, Russian, Chinese, and Iranian) and Fascism. Third, in terms of political ideologies, Mann points out that environmentalism is the only major new ideology in the twentieth and twenty-first centuries. Still, it remains that Mann has bucked the recent 'cultural turn' in the social sciences. It is also noteworthy that Mann prefers speak of 'ideological power' rather than 'patterns of culture,' and perhaps it is an open question whether, in addition to the ideologies and occasions when they mattered in Mann's account, it is possible to identify other instances of the force of ideologies in contemporary social change.

A somewhat different question relates to Mann's omission, among of his four sources of social power, of science and technology (or technoscience). Not only his account of the divergence between industrial powers in the West and 'the Rest' during the nineteenth century, but also his account of climate change, raise the issue of where techno-science fits in as a source of power, both in contributing to the rise of the West and also in contributing, separately, to environmental problems (and possible solutions to these). Mann mentions on occasion that science, with its cold explanations and lack of emotional succor (2012: 8), differs from ideological power. But he also argues that technoscience, rather than being independent, has been subservient to the other sources of social power. Nevertheless, it has been argued that without being treated as autonomous, the role of techno-science in his theory is incomplete (see the chapter by Goldstone in Hall and Schroeder, 2006; Schroeder, 2013).

These are just some of the distinctive contributions and criticisms of Mann's sociology. His work, drawing on several disciplines apart from sociology, has been influential beyond sociology and has reached into various social science disciplines. His impact has been particularly pronounced in political sociology, in theories of the state and nationalism, ethnic conflict and empire, and in international relations. His concept of infrastructural power has been used in examining developing societies, where states are often strong in despotic (and military) power, but weak in providing a basis for social development. Finally, in can be mentioned that social historians in particular have drawn on his work since social history has focused on social change 'from below,' a trend which can be also be seen in the current turn to global history, where also, Mann's work has been well received (Osterhammel, 2011).

What we see in his analysis of contemporary social change is not a single model on which societies are converging: rather, there is one dominant military empire, several forms of capitalism, and different kinds of nation-states, including the Chinese party-state and several others. Economic power is now globally capitalist, with capitalist elites having more extensive power while workers are nationally caged; nation-states are stably institutionalized but consist of different types with greater or less concentrations of power and different interests that they represent; and there are now multiple globalizations after the upheavals that fractured the globe in the twentieth century. Mann's main impact is perhaps yet to come: the last two volumes have only been published recently. Comparative-historical sociology has been an area of sociology marked by high-powered debates in the past, a tradition that goes back to Marx and Weber. More recently it has been reinvigorated by figures like Bendix, Tilly, Skocpol, Goldstone, Gellner, and others. This subdiscipline has not perhaps been as influential as it could be in

mainstream sociology or in connecting with other areas of sociology; the main connection with other areas of sociology has been the state. In the twenty-first century, this focus on the state, though now taken beyond the context of the West, is bound to be a key focus in sociology, now ineluctably entwined with global forces. It is inevitable that, barring climate and nuclear disaster, the world will become more globalized in the twenty-first century, with alternative economic and political models competing with each other and facing further crises. In this complex of networks of social powers, Mann's 'Sources' is bound to be an indispensable guide.

See also: Capitalism: Global; Citizenship, Sociological Aspects of; Globalization and World Culture; Militarism; Nationalism, Sociology of; Power in Society; Power; Social Networks; Sociological Theory; State, Sociology of the; War and Nationalism.

Bibliography

Collins, R., 1999. Maturation of the state-centred theory of revolution and ideology. In: Macro-sociology: Essays in Sociology of the Long Run. Stanford University Press, Stanford, pp. 19–36.

Gellner, E., 1988. Plough, Sword and Book – the Structure of Human History. Collins Harvill, London.

Gellner, E., 1996. Conditions of Liberty: Civil Society and its Rivals. Hamish Hamilton, London.

Hall, J.A., 1986. Powers and Liberties. Penguin, Harmondsworth.

Hall, J.A., 2006. Political questions. In: Hall, J.A., Schroeder, R. (Eds.), An Anatomy of Power: The Social Theory of Michael Mann. Cambridge University Press, Cambridge, pp. 33–55.

Hall, J.A., 2013. The Importance of Being Civil: The Struggle for Political Decency. Princeton University Press, Princeton.

Hall, J.A., Schroeder, R. (Eds.), 2006. An Anatomy of Power: The Social Theory of Michael Mann. Cambridge University Press, Cambridge.

Mann, M., 1970. The social cohesion of liberal democracy. American Sociological Review 35, 423–439.

Mann, M., 1984. The autonomous power of the state: its origins, Mechanisms, and results. Archives Europeans de Sociology 25, 185–213.

Mann, M., 1986. The Sources of Social Power, Volume I: A History of Power from the Beginning to 1760 AD. Cambridge University Press, Cambridge.

Mann, M., 1988a. The roots and contradictions of modern militarism. In: States, War and Capitalism. Basil Blackwell, Oxford, pp. 166–187.

Mann, M., 1988b. Ruling class strategies and citizenship. In: States, War and Capitalism. Basil Blackwell, Oxford, pp. 188–209.

Mann, M., 1993. The Sources of Social Power, Volume II: The Rise of Classes and Nation-States, 1760–1914. Cambridge University Press, Cambridge.

Mann, M., 2004. Fascists. Cambridge University Press, Cambridge.

Mann, M., 2005. The Dark Side of Democracy: Explaining Ethnic Cleansing. Cambridge University Press, Cambridge.

Mann, M., 2006. The sources of social power revisited: a response to criticism. In: Hall, J.A., Schroeder, R. (Eds.), An Anatomy of Power: The Social Theory of Michael Mann. Cambridge University Press, Cambridge, pp. 343–396.

Mann, M., 2003. Incoherent Empire. Verso, London.

Mann, M., 2011. Power in the 21st Century: Conversations with John A. Hall. Polity Press, Cambridge.

Mann, M., 2012. The Sources of Social Power, vol. 3: Global Empires and Revolution, 1890–1945. Cambridge University Press, Cambridge.

Mann, M., 2013a. The Sources of Social Power, vol. 4: Globalizations, 1945–2011. Cambridge University Press, Cambridge.

Mann, M., 2013b. The end may be nigh, but for whom? In: Wallerstein, I., Collins, R., Mann, M., Derluguian, G., Calhoun, C. (Eds.), Does Capitalism Have a Future? Oxford University Press, Oxford, pp. 71–99.

Mann, M., Riley, D., 2007. Explaining macro-regional trends in global income inequalities. Socio-economic Review 5, 81–115.

Osterhammel, J., 2011. Globalization. In: Bentley, J.H. (Ed.), Oxford Handbook of World History. Oxford University Press, Oxford, pp. 89–104.

Schroeder, R., 2013. An Age of Limits: Social Theory for the 21st Century. Palgrave Macmillan, Basingstoke.

Tilly, C., 1990. Coercion, Capital, and European States. Blackwell, Oxford.

Wallerstein, I., 2013. Structural crisis, or why capitalists may no longer find capitalism rewarding. In: Wallerstein, I., Collins, R., Mann, M., Derluguian, G., Calhoun, C. (Eds.), Does Capitalism Have a Future? Oxford University Press, Oxford, pp. 9–35.

Mannheim, Karl (1893–1947)

Volker Meja, Memorial University of Newfoundland, St. John's, NL, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Karl Mannheim, key classical sociologist, social thinker and founder of the sociology of knowledge, was born 1809 in Budapest, studied and worked there as well as in Heidelberg and Frankfurt, Germany, and died 1947 in emigration in London, England. His *Ideologie und Utopie* (1929) was the most widely debated book by a living sociologist in Weimar Germany and *Man and Society in an Age of Reconstruction* (1940) was no less influential during his life in Britain. His studies of generations, competition as a cultural phenomenon, economic ambition, structures of thinking, sociology of culture, conservatism, and his later writings in political sociology and social planning, mark him out as an ingenious and innovative thinker.

As a classic of sociology, Karl Mannheim is the author of one book, with three lives. *Ideologie und Utopie* (Mannheim, 1929) was the most widely debated book by a living sociologist in Germany during the climactic years of the Weimar Republic; *Ideology and Utopia* (Mannheim, 1936) has been a standard in American-style international academic sociology; and the quite different German and English versions of the book figure in reappraisals of Mannheim initiated by new textual discoveries (Mannheim, 1980, 1984) and republications (Wolff, 1993; Mannheim, 1952, 1953, 1956), fostered by a reopening of questions the discipline of sociology had once deemed closed. Mannheim's sociological theorizing has been the subject of numerous book-length studies, evidence of an international interest in his principal themes (e.g., Barboza, 2009; Corsten, 2010; Endreß and Srubar, 2000; Enebakk, 2000; Gabel, 1987; Kettler and Meja, 1995; Kettler et al., 2008; Kupiec, 2006; Patanè, 2000; Santambrogio, 1990). As a living social thinker, Karl Mannheim was not in fact the author of any work he himself considered a finished book, but rather of some 50 major essays and numerous smaller treatises, many later published in book form. His studies of generations, economic ambition, competition, structures of thinking and similar topics in the sociology of culture, as well as his later work in political sociology, notably on the theme of planning, mark him out as an ingenious and innovative social thinker.

Born on 27 March 1893 in Budapest, the son of prosperous Jewish parents, Mannheim studied philosophy and sociology in Budapest, Berlin, Paris, and Heidelberg, held academic posts at Heidelberg, Frankfurt, the London School of Economics, and the University of London, and died on 9 January 1947 in London. Just prior to his death he was appointed chair of the European division of UNESCO. His biography, one of intellectual and geographical migration, falls into three main phases: Hungarian (to 1919), German (1919–33), and British (1933–47). Among his most important early intellectual influences are Georg Lukács, Oscar Jászi, Georg Simmel, Edmund Husserl, Heinrich Rickert, and Emil Lask. Mannheim was also strongly influenced by the writings of Karl Marx, Max Weber, Alfred Weber, Max Scheler, and Georg Dilthey. Through these and other authors, German historicism, Marxism, phenomenology, sociology late also Anglo-Saxon pragmatism became decisive influences upon his work.

Mannheim spent his first 26 years in Budapest, where he graduated in philosophy and precociously participated in the intellectual life of the 'second reform generation' born a decade earlier. The advanced thinkers of the time were divided between proponents of modernization oriented to French and English social thought and prophets of radical cultural rebirth inspired by Russian and German models. Mannheim did not think that his dedication to the latter group, led by the philosopher Lukács, entailed a blanket rejection of the former, under the sociologist Jászi. Lukács' wartime *Sunday Circle* in Budapest may have devoted its meetings to Dostoevski and Meister Eckhardt, but Mannheim was also a member of the *Max Weber Circle* when he was in Heidelberg, and during a visiting semester in Berlin in 1914 he attended the lectures of Georg Simmel, the subtle mediator between cultural philosophy and sociology.

Mannheim lived in Germany from 1919, when he fled the counter-revolutionary regime in Hungary, until 1933, when National Socialist decrees forced him out of the University of Frankfurt and he left Germany for England. Within a few years of his move from the Budapest intellectual scene to German university life, Mannheim began work on two treatises on the sociology of culture that were only posthumously published as *Structures of Thinking* (1980, 1982) and he completed a habilitation thesis under Alfred Weber, a sociological interpretation of the rise and self-differentiation of conservatism with the programmatic subtitle *A Contribution to the Sociology of Knowledge* (1925). Its submission coincided with Mannheim's publication of an article devoted to his critical encounter with Max Scheler, who recently had brought the concept of the sociology of knowledge into discussion. Mannheim's inaugural address as university teacher (*Privatdozent*) at Heidelberg set out the parameters of 'the contemporary state of sociology in German,' dealing with Weber, Troeltsch, and Scheler. Sociology, he believed, provided the frame of reference for twentieth century thinking as a whole.

Mannheim's aloofness, however, from the specialized 'state of sociology' question, as well as his equation of the main currents of contemporary thought with the leading sociological theories, indicate that his move from philosophy to sociology cannot be understood as a simple change of academic

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61084-X

specializations. Sociology, in his view, was a more comprehensive undertaking than the academic discipline still taking form. Goaded in 1929 by a charge of 'sociologism' against *Ideologie und Utopie* by the noted literary scholar, Ernst Robert Curtius, Mannheim (Meja and Stehr, 1982, 1990) again invoked the heritage of Weber, Troeltsch, and Scheler against the literary scholar's accusation of treason to humanism, presenting these sociologists as modern German classics, as 'a great heritage, a tradition that must be built upon.'

The novelty in Mannheim's approach to the sociology of knowledge is epitomized in three distinct claims. First, and most controversial, is the contention that boundaries between manifestly ideological and ostensibly scientific modes of explaining the cultural as well as the social world are porous, with sociology of knowledge emerging in the border region, as a reflexive therapy for both domains. Second is the concomitant conception of ideologies as cognitive structures that are variously flawed, limited, perspectivistically one-sided, subject to correction from other perspectives, and nevertheless productive of knowledge. The third claim, then, is that the sociology of knowledge bears on the answers to substantive questions addressed by ideologies, consequently contributing directly to political orientation. It does so not because knowledge of social genesis can determine judgments of validity, but because systematic pursuit of such knowledge fosters a comprehensive synthesis and renders particularistic ideologies obsolete.

Mannheim's strategy involves two steps. First, the variety of ideas in the modern world is classified according to a scheme of historical ideological types, in keeping with his thesis that the ideological field has moved from atomistic diversity and competition to concentration. Liberalism, conservatism, and socialism are the principal ideological types. Second, each ideology is interpreted as a function of some specific way of being in the social world, as defined by location within the historically changing patterns of class and generational stratification. Liberalism is thus referred to the capitalist bourgeoisie in general, and various stages in its development are linked to generational changes. Similar analyses connect conservatism to social classes harmed by the rise to power of the bourgeoisie, and socialism to the new industrial working class. Each of the ideologies is said to manifest a characteristic 'style' of thinking, a distinctive response to the issues that systematic philosophy has identified as constitutive of human consciousness, such as conceptions of time and space, the structure of reality, human agency, and knowledge itself. The political judgments and recommendations of purely ideological texts must be taken in that larger structural context. The style of thinking is most apparent in the way concepts are formed and in the logic by which they are interlinked. Each style expresses some distinctive design upon the world vitally bound up with the situation of one of the social strata present in the historical setting.

Sociology of knowledge seeks to give an account of the whole ideological field, in its historical interaction and change. To have a method for seeing this means to see in a unified way what ideologically oriented viewers can only see in part. Mannheim draws on Marxism for a conception of politics as a process of dialectical interplay among factors more 'real' than the competing opinions of liberal theory. But neither the proletariat nor any other socio-political force is bearer of a transcendent rationality, historically destined to reintegrate irrationalities in a higher, pacified order. The contesting social forces and their projects in the world are in need of a synthesis that incorporates elements of their diverse social wills and visions. Syntheses in political vision and in the social sciences are interdependent. Sociology of knowledge foreshadows and fosters both.

Mannheim's call to a professorship in Frankfurt followed the remarkable recognition earned by his further work in sociology of knowledge. In his presentation on *Competition as a Cultural Phenomenon* at the Sixth Conference of German Sociologists in 1928 (in Wolff, 1993), Mannheim audaciously used the value–judgment controversy in recent sociology to illustrate his theses about the connectedness of social thought to existence (*Seinsverbundenheit*) and the operations of socially grounded competition to generate syntheses that transcend intellectual conflict. *Ideologie und Utopie*, consisting of three essays on 'politics as a science,' 'utopian consciousness,' and an explication of the concepts of '*ideology*' and '*utopia*' generated great excitement that launched Mannheim in his new setting, recognized in the wider intellectual community as a significant and controversial personality. Questions about varieties of knowing and about the ways in which new knowledge depends on authentic grounding in the contexts of existing knowledge are the major themes of much of Mannheim's work during this period. His seminal paper on *The Problem of Generations* (1927), which continues to have a lasting impact in sociological research and analysis (especially by way of cohort analysis and the sociology of the life course), also belongs here.

The sociology of knowledge dispute about *Ideologie und Utopie* (Meja and Stehr, 1982, 1990) was primarily a philosophical and political dispute, with the focus, first, on Mannheim's hope of overcoming both ideology and political distrust by way of a sociology of knowledge; second, on his conception of the intelligentsia as the social stratum uniquely equipped and even destined for this task; and third, on his activist conception of sociological knowledge, as a mode of mediating between theory and practice. In *Science as a Vocation* (1922), Max Weber had distinguished between words in politics and in science, likening the former to weapons for overpowering opponents and the latter to ploughshares for cultivating knowledge. Mannheim offers the sociology of knowledge as an 'organon for politics as a science,' as a way of bringing about the biblical transformation of swords into pruning hooks prophesied by Isaiah. His proposals were widely canvassed in the leading periodical reviews and subjected to intense criticism, but his reading of the intellectual situation was almost universally applauded. In the cultivated Weimar public and among the participants in the 'Weimar conversation' about social thought after Nietzsche and Marx, *Ideologie und Utopie* figured as the representative book of its time, whether as symptom of cultural crisis or as promise of a way out. If *Ideology and Utopia* and *Structures of Thinking* represent Mannheim's most important theoretical exercises, his work on conservatism has commonly been taken as paradigmatic for an empirical sociological of knowledge.

During his five semesters as professor in Frankfurt (where Norbert Elias was his research assistant), however, Mannheim

declined the role of public intellectual. He deliberately separated the professional aspects of his activities from his public reputation. While drawing close to Paul Tillich and his circle of religious socialists, celebrated and embattled as an 'intellectual,' he nevertheless increasingly defined himself as a professional sociologist. Never having exposed himself politically, he was unprepared for the Nazi measures that deprived him of his professorship on grounds of his foreign birth and his Jewish ethnicity. *Ideologie und Utopie* had been favorably reviewed at length in the Socialist periodical, *Die Gesellschaft*, but the four articles published there (by Hannah Arendt, Herbert Marcuse, Hans Speier, and Paul Tillich), were all more critical than the reviews in *Die Tat*, a periodical of the activist Right. The hard Left treated him as a betrayer of Marxism. Soon Mannheim was a refugee in Amsterdam. Neither his sociology nor his politics had anything to do with his exile from Germany.

In the summer of 1933, Mannheim was appointed to a special lectureship at the London School of Economics. Neither the times nor his situation were conducive to pursuing sociology of knowledge studies, however. Mannheim saw it as his mission to diagnose the general crisis he held responsible for the German disaster and to promote prophylactic and therapeutic measures in Britain. His sense of urgency and his grand theoretical ambitions impressed students, but they estranged the professional sociologists led by Morris Ginsberg, who were engaged in a difficult fight for academic respectability of a discipline still widely dismissed by the English university establishment. Although marginalized at the London School of Economics, the only British institution with a chair in Sociology, Mannheim made a place for himself as a public intellectual, especially after his acceptance by a circle of Christian thinkers that included T.S. Eliot and whose periodic discussions and publications centered on a theme of cultural crisis hospitable to Mannheim's sociological interpretations.

In Britain, Mannheim continued to focus on the relationship between knowledge and society, his lifelong topic. Writing in *Man in Society in an Age of Reconstruction*, a work no less influential during the war and in the postwar years than *Ideology and Utopia*, he now claims that the sociology of knowledge has lost its strategic centrality with the demise of ideological competition (Mannheim, 1940) and calls for "a new experimental attitude in social affairs," in view of the "practical deterioration of the ideals of Liberalism, Communism, and Fascism." The National Socialist dictatorship, Mannheim argues, exploits a socially unconscious mass response to a worldwide crisis in the institutions of liberal civilization, involving the obsolescence of its regulative social technologies – from markets to parliaments to elitist humanistic education. He pleads for a planned social order that strategically utilizes the new social technologies that undermine the spontaneous self-ordering of the previous epoch. A consensual reconstruction could save human qualities and diversities earlier privileged by liberalism, unlike the violent homogenization imposed by communist or national socialist control through command. Timely action guided by awareness of the impending crisis by leading strata, notably the English elite of gentlemanly professionals, can tame the processes that would otherwise destroy liberal civilization and condition mass populations for dictatorial domination. 'Planning for freedom' presupposes a reorientation among traditional elites, their acceptance of a sociological diagnosis of the times, and their willingness to learn prophylactic and therapeutic techniques (Mannheim, 1943). Mannheim now claims for sociology the ability to ground and coordinate interdisciplinary approaches in planning. His lectures and writings on 'planning' won him an interested audience, especially during the war and immediate postwar years, and his conception of a postideological age was never altogether submerged by the Cold War (Mannheim, 1950). As his wartime slogan of 'militant democracy' justified German measures against leftists during the middle decades of the century, his slogan of the Third Way was heard at the beginning of the new millennium in support of political designs he would have found familiar.

Among sociologists, however, Mannheim's standing was defined by the reception of *Ideology and Utopia*. In his preface, Louis Wirth casts the work primarily as a contribution to objectivity in social science. The professional consensus is formalized in R.K. Merton's authoritative essay in 1945 on 'The Sociology of Knowledge' (in Merton, 1957). Merton includes Mannheim in a group of social theorists from Marx to Sorokin whose diverse approaches to 'the relations between knowledge and other existential factors in the society and culture' he relates to questions and alternatives that provide an agenda for the theoretical clarification and empirical research required to build sociology. Merton's 'paradigm' for the sociology of knowledge sets out five key issues: the existential basis of mental productions, the varieties and aspects of mental productions subject to sociological analysis, the specific relationship(s) between mental productions and existential basis, the functions of existentially conditioned mental productions, and the conditions under which the imputed relations obtain. Crediting him with having sketched the broad contours of the sociology of knowledge with skill and insight, Merton nevertheless finds Mannheim's theory burdened with dubious philosophical claims, and insists that the relations between knowledge and social structure can be clarified only after they are "shorn of their epistemological impedimenta, with their concepts modified by the lessons of further empirical inquiry" (Merton, 1957: p. 508).

The condition for Mannheim's acceptance as deserving sociological pioneer was the discarding of the concept of total ideology and his way of questioning both social science and social knowledge. The editors and organizers of three posthumous collections of Mannheim's essays, his close friends and noted social scientists, Paul Kecskemeti, Ernest Manheim, and Adolph Loewe, reinforced the consensus view. As Mannheim became acquainted with Anglo-American social science, they argued, empirical social psychology displaced continental philosophies as the theoretical framework for his thinking, and his early writings merit consideration primarily as brilliant anticipations of later developments. In Merton's sociological theory classes at Columbia University during the 1950s, *Ideology and Utopia* often followed Machiavelli's *Prince* in the syllabus, as source material for an exercise in transmuting suggestive ideas into testable propositions. Mannheim's work had joined the list of historical authorities

celebrated as legitimating precursors of a settled way of defining and doing sociology.

We can trace the renewal of attention to the historical social thinker, Karl Mannheim, to renewed conflicts about the subject, method, and attitude of sociology. The connection is not simple, since the attack on the disciplinary consensus, where historical models were involved, was more likely to call on Marxist writers or on figures like Theodor W. Adorno, Max Horkheimer, and Herbert Marcuse, who had been among Mannheim's harshest contemporary critics. The effect, nevertheless, was to relegitimate questions about the historicity of social knowledge, the problem of relativism, and the paths of reflexivity open to social thinkers – the issues filtered out of Mannheim's thought in his American reception – and to provide a new point of connection for those sociologists (e.g., Wolff, 1993) who had continued puzzling over the issues debated when *Ideologie und Utopie* first appeared. Berger and Luckmann's influential phenomenological reformulation of the sociology of knowledge in *The Social Construction of Reality* (1966) represents a departure from Mannheim's sociology of knowledge. All knowledge, both objective and subjective (and including ideology, propaganda, science, and art), is now seen as socially constructed. Berger and Luckmann examine how forms of knowledge are maintained and modified by the institutions and individuals embodying and embracing them.

In the newly fluid state of questions appropriate for sociological theory, Mannheim's famous *Ideology and Utopia* no longer stands alone among his writings, and his insistence on essayistic experimentalism is no longer ignored. There are no propositions to be distilled out of Mannheim's work; there is just the thoughtful encounter with it. Mannheim's prediction about the consequences of uncovering the 'massive fact' of knowledge in society and society in knowledge has been borne out. Today 'knowledge society' is a widely recognized and accepted concept. Yet we are not beneficiaries of any 'progress' in thought, as post-modernists paradoxically often suppose: we may well have to think deeply about old texts, however we class them.

See also: Age, Sociology of; Knowledge, Sociology of; Phenomenology: Philosophical Aspects; Science, Sociology of; Sociology, History of.

Bibliography

Berger, P., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.

Barboza, A., 2009. Karl Mannheim. UVK, Konstanz.

Corsten, M., 2010. Karl Mannheims Kultursoziologie. Eine Einführung. Campus, Frankfurt and New York.

Endreß, M., Srubar, I., 2000. Karl Mannheims Analyse der Moderne. Jahrbuch für Soziologie. Leske und Budrich, Opladen.

Enebakk, V., 2000. Fra Wissenssoziologie til Sociology of Knowledge. Resepsjonen av Karl Mannheims kunnskapssosiologi I forbindelse med publiseringen av Ideology and Utopia (1936). Senter for teknologi, innovasion og kultur, Universitetet i Oslo, Oslo.

Gabel, J., 1987. Mannheim et le Marxisme Hongrois. Méridiens Klincksieck, Paris.

Kettler, D., Meja, V., 1995. Karl Mannheim and the Crisis of Liberalism: The Secret of These New Times. Transaction Publishers, New Brunswick, NJ.

Kettler, D., Loader, C., Meja, V., 2008. Karl Mannheim and the Legacy of Max Weber. Retrieving a Research Programme. Ashgate, Aldershot, England.

Kupiec, A., 2006. Karl Mannheim. Idéologie, utopie et connaissance. Felin, Paris.

Mannheim, K., 1929/1953. Ideologie und Utopie, third enl. ed. Schulte-Bulmke, Frankfurt. (1995. eighth ed. Klostermann, Frankfurt).

Mannheim, K., 1936. Ideology and Utopia (L. Wirth, E. Shils, Trans.). Kegan Paul, Trench, Trubner & Co., London.

Mannheim, K., 1940. Man and Society in an Age of Reconstruction (E. Shils, Trans.). Kegan Paul, Trench, Trubner & Co., London.

Mannheim, K., 1943. Diagnosis of Our Time. Wartime Essays of a Sociologist. Kegan Paul, Trench, Trubner & Co., London.

Mannheim, K., 1950. In: Bramsted, E.K., Gerth, H. (Eds.), Freedom, Power and Democratic Planning. Oxford University Press, New York.

Mannheim, K., 1952. In: Kecskemeti, P. (Ed.), Essays in the Sociology of Knowledge. Routledge and Kegan Paul, London.

Mannheim, K., 1953. In: Kecskemeti, P. (Ed.), Essays on Sociology and Social Psychology. Routledge and Kegan Paul, London.

Mannheim, K., 1956. In: Manheim, E., Kecskemeti, P. (Eds.), Essays on the Sociology of Culture. Routledge and Kegan Paul, London.

Mannheim, K., 1922–24/1980. Strukturen des Denkens. Suhrkamp, Frankfurt [Trans. 1982. In: Kettler, D., Meja, V., Stehr, N. (Eds.), Structures of Thinking. Routledge & Kegan Paul, London].

Mannheim, K., 1925/1984. Konservatismus. Ein Beitrag zur Soziologie des Wissens. Suhrkamp, Frankfurt [Trans. 1986. Kettler, D., Meja V, Stehr, N. (Eds.), Conservatism. A Contribution to the Sociology of Knowledge. Routledge & Kegan Paul, London].

Meja, V., Stehr, N. (Eds.), 1982. Der Streit um die Wissenssoziologie, 2 vols. Suhrkamp, Frankfurt.

Meja, V., Stehr, N. (Eds.), 1990. Knowledge and Politics. The Sociology of Knowledge Dispute. Routledge Library Editions: Social Theory, London and New York.

Merton, R.K., 1957. Social Theory and Social Structure. The Free Press, Glencoe, IL.

Patanè, S., 2000. Oltre la modernità. La sociologia della conoscenza di Karl Mannheim nella società postmoderna. Euroma, Roma.

Santambrogio, A., 1990. Totalità e Critica del totalitarismo in Karl Mannheim. Angeli, Milano.

Weber, M., 1919/1922. Wissenschaftslehre. J.C.B. Mohr (Paul Siebeck), Tübingen.

Wolff, K.H. (Ed.), 1993. From Karl Mannheim. Transaction Publishers, New Brunswick, NJ [second ext. ed. In: Meja, V., Kettler, D. (Eds.)].

Manual Based Treatments

Johanna Thompson-Hollands and David H Barlow, Boston University, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Treatment manuals were first developed in the 1950s and quickly became an important tool, allowing for the objective and replicable definition of a given psychological approach. Manuals greatly enhanced the rigor of treatment research studies, and they also provided an effective way to communicate treatment strategies to clients and clinicians at large. There are now dozens of manuals available, representing a range of problem areas and theoretical approaches. Although manuals have many advantages when compared to unstructured treatments, they have faced criticism from some practitioners. Nevertheless, manuals will continue to play an important role in psychological treatment in the future.

Introduction

The practice of psychotherapy has seen dramatic and sweeping changes since the 1950s. One of the most dramatic changes impacting the delivery and dissemination of specific psychotherapeutic service has been the development of detailed explanatory manuals for complex psychological treatments. In addition to manuals intended for use by therapists, client 'workbooks' are also available for use in therapist-directed treatment or as self-help resources. The main function of these treatment manuals is to provide an operational definition of a given psychotherapeutic approach, including acceptable procedures, techniques, and strategies (Chambless and Ollendick, 2001) and to make the process of psychotherapy more accessible to clients and clinicians alike. Typically, these manuals describe session-by-session strategies for specific problems such as depression or phobias for therapists to follow.

Many changes in the field led to the impetus for specific treatment manuals to be developed including the demand for 'good and better research,' in which the therapeutic intervention is specified for all to examine, and the burgeoning of the hegemonic managed behavioral health care system. In response to these many demands, the number of manuals has proliferated. Although the psychotherapeutic community has positively received the majority of manuals, some frontline clinicians raise objections to this approach, stating, among other objections, that manuals damage the therapeutic relationship, and that clinical innovation is restricted (Addis and Krasnow, 2000). Despite these obstacles, treatment manuals are now a well-established part of the dissemination and treatment landscape in clinical psychology.

Developmental History of Treatment Manuals

The initial impetus for the development of treatment manuals came from psychotherapy researchers who in the early 1960s began to test broadly the effectiveness of specific treatments in controlled outcome studies. At that time, psychotherapy outcome studies were methodologically crude, consisting largely of collections of case studies in which successful outcomes were attributed to the relevant 'school' of psychotherapy rather than to specific elements of treatment that could be examined individually (Barlow et al., 2013). The treatments administered in these studies were quite loosely defined, preventing the analysis of particular treatment components. Naturalistic studies were also conducted, incorporating many theory-driven measures of clinician and patient behaviors; such studies tended to utilize broad, heterogeneous groups of patients, and treatment decisions (use of medications, assignment of patients to particular therapists) were based on clinical judgment. Such designs produced results that were generally conflicting and of little scholarly value (May, 1973). Despite the lamentable state of outcome research during this time, a few prescient researchers recognized the need to pursue methods that could more closely tie process to outcome (Paul, 1967; Strupp, 1964).

Looking to demonstrate successfully that psychological interventions could withstand rigorous scientific investigation, similar to that of existing pharmacological treatments (Luborsky and DeRubeis, 1984), scientist-practitioners realized that they needed treatment tools that would allow for systematic replication and comparison. Psychotherapeutic procedures needed to be clearly specified in order to be tested accurately (Luborsky et al., 1975). As Bergin and Strupp (1972) noted, "In short, the therapist and the researcher must become increasingly *explicit about the operations of psychotherapy and the nature of the therapist's influence*. Existing knowledge, while undoubtedly embodying important psychological principles, is altogether too general, broadly-gauged, and imprecise …. [F]uture research in this area must rest firmly on empirical data; it must tend toward increasing explicitness and specificity; and it must seek to isolate psychological principles embedded in, and often obscured by, divergent theoretical formulations" (p. 436).

The creation of treatment manuals paved the way for a true science of treatment to emerge, for researchers were now able to outline a precise template of interventions that could be replicated with fidelity (Perepletchikova et al., 2007). Wilson (1996) more specifically pointed out that treatment manuals sought to eliminate any 'error' associated with 'clinical judgment' or intuition that might cause one therapist to behave in a substantially different manner from another. Many researchers hoped that by utilizing treatment manuals, presented in this fashion, psychological interventions would

 http://dx.doi.org/10.1016/B978-0-08-097086-8.21045-3

be able to withstand the methodological constraints of research protocols. This provided a striking contrast to the conduct of treatment outcome research prior to manualization, in which specific therapeutic techniques were often not explained and thus could not be compared to other treatments or be replicated by other investigative groups. As a consequence, federal agencies in countries such as the United States and the United Kingdom began to request that investigators specify the use of a treatment manual in their studies when applying for funding.

System-Wide Adoption of Manualized Treatment

While early manuals were usually developed in order to facilitate and strengthen research, manuals are now the primary mechanism for dissemination of empirically supported psychological treatments. In recent years, national-level programs to support the use of manualized, evidence-based interventions to treat mental illness have been established, with governments committing billions of dollars to the cause (McHugh and Barlow, 2010). In the United States, the Veterans Health Administration (VHA) produced a new Mental Health Strategic plan in 2004, including goals for disseminating evidence-based psychological treatments into practice throughout the VHA system. Several manual-based treatments for specific disorders relevant to a veteran population, including cognitive-processing therapy (CPT; Resick et al., 2007) and prolonged exposure (PE; Foa et al., 2007) for posttraumatic stress disorder, were 'rolled out' system-wide (Karlin et al., 2010). Phase I of this roll-out process for CPT and PE involved the development of treatment manuals tailored to a veteran population. All VHA mental health clinics are required to have staff members who are adequately trained in the interventions, and the explicit goal of the program is to have such treatments become the standard of care (McHugh and Barlow, 2010). In the United Kingdom, a similar program, Improving Access to Psychological Therapies (IAPT), was announced in 2007 (Clark et al., 2009); the initiative is extensive, with rigorous standards for training in and implementation of the supported treatments. The IAPT represents the single largest commitment of a national agency to dissemination of evidence-based psychological treatments thus far, and initial results have indicated that the program's goals of increasing referrals and access to empirically supported treatments are being met (Clark et al., 2009).

Across these large-scale initiatives, clinical practice guidelines and quality standards explicitly articulate the optimal strategies for assessing and treating a variety of psychological problems. Treatments recommended in Psychotherapy Clinical Practice Guidelines are typically based on two specific factors: (1) efficacy, or internal validity of the specific treatment, the determination of which is based on the results of a systematic evaluation of the intervention in a controlled setting, and (2) effectiveness, or clinical utility of the treatment, which is based on the feasibility, generalizability, and cost-effectiveness of the intervention actually being delivered in a local setting. Based on these equally important and rigorous principles, the use of treatment manuals that can produce the necessary evidence is encouraged (Chambless and Ollendick, 2001). Strosahl (1998) points out that manuals are especially appealing to managed-care settings because they are essentially an easily discernible road map for the most appropriate way to implement clinical practice guidelines. Within managed care organizations, where psychotherapy (1) must demonstrate its overall effectiveness, (2) may be limited to a certain number of sessions, and (3) increasingly is delivered by practitioners with less than doctoral degrees, treatment manuals are embraced for their ability to facilitate the delivery of empirically supported treatments at lower costs. As a result, manual-based treatments have been incorporated as one of the major components of evidence-based service delivery (Roth and Pilling, 2007).

Specific Treatment Manuals

The majority of current treatment manuals are based on behavioral techniques rather than more traditional psychotherapy models, which tend to emphasize a more individualized theory-based understanding of underlying patient problems (Wilson, 1996). Currently, specific manuals exist for many diagnostic categories specified by Diagnostic and Statistical Manual of Mental Disorders, 4th edition, including a wide variety of anxiety disorders such as panic disorder, generalized anxiety disorder, obsessive-compulsive disorder, posttraumatic stress disorder, and social anxiety disorder. Most of these manuals utilize cognitive and behavioral therapy (CBT) techniques with an emphasis on identifying and challenging maladaptive cognitions and eliminating behaviors that serve to increase and maintain anxiety (e.g., Barlow et al., 2011; Craske and Barlow, 2006; Foa et al., 2012). Other manuals with a cognitive behavioral emphasis include treatments for bulimia nervosa, weight reduction, stress reduction, depression, addictive behaviors, and sexual dysfunction (e.g., Antoni et al., 2007; Apple and Agras, 2007; Daley and Marlatt, 2006; Gilson et al., 2009). Newer approaches to cognitive-behavioral techniques that incorporate mindfulness and non-judgment in addition to some of the more traditional CBT elements have also been written into manuals for a variety of problem areas (e.g., Forsyth and Eifert, 2008; Woods and Twohig, 2008).

Although cognitive behavioral manuals tend to predominate, other forms of therapy have also followed this route as the benefits of manualization (ease of dissemination and training and replicability across research studies) have become more recognized. Family-systems approaches for disorders ranging from bipolar disorder and schizophrenia to eating disorders have been published (e.g., Lock et al., 2012; Miklowitz, 2010). Manuals also exist for social skills training, problem-solving therapy, and supportive/expressive therapies, among others (e.g., Bellack et al., 2004; Nezu et al., 2012; O'Connor and Ammen, 2012).

Pros and Cons of Treatment Manuals

One distinct advantage of treatment manuals is that they have been shown to be effective in controlled treatment outcome studies (Barlow, 2004; Nathan and Gorman, 2007). Thus,

when practitioners implement techniques, they can do so with the confidence that they are delivering services that have a high probability of success. Furthermore, manuals guide clinicians to use strategies that are spelled out and are clearly discernable. This allows for the examination of treatment integrity, including therapists' treatment adherence and competence, as well as differentiation from other treatment approaches (Perepletchikova et al., 2007). As mentioned above, poor treatment integrity compromises the validity of research findings and hinders replication attempts. By providing a specific outline of techniques, manuals impose structure upon the therapist and communicate discrete parameters regarding the treatment itself.

Wilson (1996) points out that treatment manuals can reduce the errors that might be associated with unrestrained clinical judgment. "What is typically overlooked, however, and what research on human judgment so clearly documents, is that individualized clinical judgment can just as easily produce worse results than standardized treatments, by introducing errors and inappropriate strategies that are not part of manual-based treatments" (Wilson, 1996, p. 302). Manuals, however, outline specific techniques to be used in each session, which have been created by systematic investigation, rather than relying upon clinical judgment to choose a session topic.

Due to their structure, treatment manuals may also facilitate a more highly focused and efficient therapy. With a limited number of sessions and specific goals and strategies outlined for each of these sessions, this more focused approach may actually lead to a more active engagement in therapy for patients and encourage efficiency on the part of clinicians.

Treatment manuals also make it easier to train and supervise therapists in specific clinical techniques and strategies (Calhoun et al., 1998). By providing nascent clinicians and supervisors with specific guidelines and techniques to monitor, treatment manuals streamline the learning process and delineate a specific set of therapeutic skills to be learned. This in turn may lead to a greater aptitude and ability for learning therapeutic techniques in general, and a larger armamentarium of clinical skills.

Treatment manuals, as structured interventions, can also help clinicians deliver therapy in a brief and in many cases nontraditional format (Craske et al., 1995). For example, a manualized version of brief problem-solving therapy for depression was administered by both primary care physicians and by nurses; patients' improvement was not significantly different whether a physician or a nurse treated them. Additionally, the therapy-alone conditions were not significantly different from either a medication alone condition or a combined treatment condition (Mynors-Wallis et al., 2000). Considering the high prevalence rates of clinical and subclinical mental disorders in primary care settings (Kroenke et al., 2007), such interventions are clearly cost-effective.

Despite these many advantages, a sizable segment of practicing clinicians object to the notion of treatment manuals, pointing out the many disadvantages in the delivery of what has been called derisively, robotlike therapy (Addis and Krasnow, 2000). For example, dissenters have argued that randomized controlled trials, utilized to validate most manualized treatments, exclude patients with comorbid diagnoses, among other constraints, and that the results of these types of research studies, including the treatments they produce, inherently have limited applicability to 'real-world' settings (Bologna et al., 1998; Carroll and Nuro, 2002). These same individuals note that highly structured, manualized psychological interventions used in clinical trials cannot generalize to practice settings where psychotherapy is often applied more flexibly and adapted to meet the needs of the patients, who may present with multiple problems (Barlow, 1996; Jensen-Doss et al., 2009; Nelson et al., 2006). This set of objections is based on the fundamental idea that the scientific method, as it is known, is incapable of proving the effectiveness or ineffectiveness of psychotherapy, and that some alternative methodological approach to psychotherapy research is preferable. These are important criticisms that cannot be ignored. However, manualized treatment need not be delivered rigidly; even though the therapist utilizing a treatment manual is delivering a particular set of outlined techniques, within this delivery they are also encouraged to rely on their own personal skills to convey the therapy to the patient (Kendall et al., 1998). For example, clinicians in the Treatment of Depression Collaborative Research Program were adherent to the study's treatment manual while also flexibly adapting their presentation and interaction style as appropriate to a particular patient (Gibbons et al., 2002). Importantly, research has also shown that the overall amount of therapist flexibility does not predict patient outcomes (Kendall and Chu, 2000), suggesting that the importance of tailoring interventions to a given individual may be overstated. Mounting evidence suggests a high degree of generalizability of manualized treatments to front-line clinical settings despite substantial comorbidity and a diverse population base (Barlow et al., 1999).

Other clinicians bemoan that structured treatment manuals will compromise the integrity of the therapeutic relationship and thus interfere with the therapy process (Addis et al., 1999). For example, many of these clinicians fear that manualized therapy may reduce the therapy process to mechanized and robotic delivery of techniques, thereby eliminating the importance of the therapist as a clinical tool. However, Wilson (1996) notes, "if clinical artistry is taken to connote such necessary therapeutic elements as developing a therapeutic relationship and engaging patients in the change process, then treatment manuals do not obfuscate it – rather they demand it" (p. 305). The data indicate that therapists engaged in manualized treatments in clinical practice can form strong alliances (Addis, 1997; Creed and Kendall, 2005). Simply because the therapist has to spend a segment of time conveying specific therapeutic techniques does not mean that they are not building alliances with their patients. Furthermore, a great deal of time is allotted in manual-based therapies to emphasize relationship building and exploring the patient's unique perspective. Examples include such strategies as identifying client and therapist expectations for treatment, eliciting client feedback, and exploring motivation for change (Barlow et al., 2011; Hope et al., 2010).

Criticisms of treatment manuals also encompass the belief that manuals may undermine successful case formulation. That is, many clinicians argue that manual-based treatments will be less effective because these approaches assume that all

individuals with a given disorder are uniform and have the same symptoms for the same reasons (Wilson, 1996). As such, these techniques will 'miss the boat' for a certain subset of patients and will attempt to treat them with techniques that are not applicable. The fact that manuals do not allow for clinical exploration that might lead to the discovery of what works for patients with dissimilar etiology is a commonly expressed concern that may lead to diminished treatment innovation. In fact, treatment manuals do emphasize the current symptom picture over past development processes and idiosyncratic etiology. But research has yet to suggest that therapeutic attention to past developmental processes contributes in a substantial way to the amelioration of the disorder.

Finally, utilization of treatment manuals can promote innovation in the delivery of clinical services. That is, when highly specified treatments are delivered, and assuming appropriate outcome measures are collected, clinicians can determine when a specific treatment is working and for whom. In the case of failures, either individually or systematically, it will then be clear that innovations to the treatment program are needed and should be incorporated and evaluated. If data are systematically collected regarding who responds to the treatment and who does not, along with responses to second-stage interventions, innovations in treatment will occur much more rapidly than with unstructured approaches.

Many of the criticisms of manualized treatment have some validity, and researchers continue to explore issues relating to the acceptability and external validity of manual-based treatments. However, a large subset of practicing clinicians continues to report no or low levels of familiarity with these protocols (McHugh and Barlow, 2012); moving forward, it is critical that evaluations of manualized treatments be based upon their true performance, rather than on worst fears of their potential shortfalls.

The Future of Structured Treatment Manuals

The move toward treatment manuals has radically shifted the science and practice of psychotherapy over the last half century. Despite some objections and lingering doubts, treatment manuals have many merits, with demonstrated efficacy the principal attribute, and will clearly continue to shape the treatment of mental illnesses for years to come. As one looks ahead to the future of these evidence-based psychological interventions, challenges have become apparent. The first challenge is making these treatments more available to the people who would benefit from them. Given the difficulty of successfully disseminating and implementing even one evidence-based treatment into routine practice (McHugh and Barlow, 2010, 2012), it has become apparent that large-scale dissemination of treatments focused on single psychological disorders is not feasible. One more parsimonious alternative, transdiagnostic treatment protocols, can provide a comprehensive, theory-based approach to addressing a range of diagnoses and comorbidities. This reduces training burden for clinicians while maintaining the advantages of manual-based treatment. A second challenge relates to moving treatment 'beyond the book' using stepped care models and a variety of delivery methods. The need for mental health services vastly outpaces the available number of clinicians, creating a need for creativity in the administration of treatment (Kazdin and Blase, 2011; Kazdin and Rabbitt, 2013). Stepped care approaches provide several levels of treatment depending on the needs of the individual patient. Many individuals may be able to benefit from lower intensity interventions that do not require a highly trained mental health professional, such as guided self-help; treatment manuals are ideal for this format. The books may be used on their own, or the treatment protocol may be integrated into an interactive computer or web-based program with features such as video clips and regular progress assessments. These approaches increase the availability of high-quality treatment services while also reducing cost and system burden.

See also: Behavior Therapy: Background, Basic Principles, and Early History; Counseling and Psychotherapy: Ethnic and Cultural Differences; Evidence-Based Psychotherapy; Group Psychotherapy, Clinical Psychology of; Managed Care and Psychotherapy in the United States; Multimodal Therapy in Clinical Psychology; Narrative Exposure Therapy; Psychological Treatment, Effectiveness of; Psychological Treatments Delivered Nontraditionally; Psychological Treatments, Empirically Supported; Psychological Treatments: Randomized Controlled Trials; Psychotherapy Integration; Psychotherapy Process Research; Therapist–Patient Relationship.

Bibliography

Addis, M.A., 1997. Evaluating the treatment manual as a means of disseminating empirically validated psychotherapies. Clinical Psychology: Science and Practice 4, 1–11.

Addis, M.A., Krasnow, A.D., 2000. A national survey of practicing psychologists' attitudes toward psychotherapy treatment manuals. Journal of Consulting and Clinical Psychology 68, 331–339.

Addis, M.E., Wade, W.A., Hatgis, C., 1999. Barriers to dissemination of evidence based practices: addressing practitioners' concerns about manual based psychotherapies. Clinical Psychology: Science and Practice 6, 430–441.

Antoni, M.H., Ironson, G., Schneiderman, N., 2007. Cognitive-behavioral Stress Management. Oxford University Press, New York.

Apple, R.F., Agras, W.S., 2007. Overcoming Your Eating Disorders: a Cognitive-behavioral Therapy Approach for Bulimia Nervosa and Binge-eating Disorder. Oxford University Press, New York.

Barlow, D.H., 1996. Health-care policy, psychotherapy research, and the future of psychotherapy. American Psychologist 51, 1050–1058.

Barlow, D.H., 2004. Psychological treatments. American Psychologist 59, 869–878.

Barlow, D.H., Boswell, J.F., Thompson-Hollands, J., 2013. Eysenck, Strupp and 50 years of psychotherapy research: a personal perspective. Psychotherapy 50, 77–87.

Barlow, D.H., Levitt, J.T., Bufka, L.F., 1999. The dissemination of empirically supported treatments: a view to the future. Behaviour Research and Therapy 37, 5147–5162.

Barlow, D.H., Farchione, T.J., Fairholme, C.P., Ellard, K.K., Boisseau, C.L., Allen, L.B., Ehrenreich-May, J., 2011. Unified Protocol for Transdiagnostic Treatment of Emotional Disorders: Therapist Guide. Oxford University Press, New York.

Bellack, A.S., Mueser, K.T., Gingerich, S., Agresta, J., 2004. Social Skills Training for Schizophrenia, Second Edition: A Step-by-step Guide. The Guilford Press, New York.

Bergin, A.E., Strupp, H.H., 1972. Changing Frontiers in the Science of Psychotherapy. Aldine Atherton, Inc, Chicago.

Bologna, N.C., Barlow, D.H., Hollon, S.D., Mitchell, J.E., Huppert, J.D., 1998. Behavioral health treatment redesign in managed care settings. Clinical Psychology: Science and Practice 5, 94–114.

Calhoun, K.S., Moras, K., Pilkonis, P.A., Rehm, L.P., 1998. Empirically supported treatments: Implications for training. Journal of Consulting and Clinical Psychology 66, 151–162.

Carroll, K.M., Nuro, K.F., 2002. One size cannot fit all: a stage model for psychotherapy manual development. Clinical Psychology: Science and Practice 9, 396–406.

Chambless, D.L., Ollendick, T.H., 2001. Empirically supported psychological interventions: controversies and evidence. Annual Review of Psychology 52, 685–716.

Clark, D.M., Layard, R., Smithies, R., Richards, D.A., Suckling, R., Wright, B., 2009. Improving access to psychological therapy: Initial evaluation of two UK demonstration sites. Behaviour Research and Therapy 47, 910–920.

Craske, M.G., Barlow, D.H., 2006. Mastery of Your Anxiety and Panic: Therapist Guide. Oxford University Press, New York.

Craske, M.G., Maidenberg, E., Bystrisky, A., 1995. Brief cognitive behavioral versus non-directive therapy for panic disorder. Journal of Behaviour Therapy and Experimental Psychiatry 26, I13–20.

Creed, T.A., Kendall, P.C., 2005. Therapist alliance-building behavior within a cognitive-behavioral treatment for anxiety in youth. Journal of Consulting and Clinical Psychology 73, 498–505.

Daley, D.C., Marlatt, G.A., 2006. Overcoming Your Alcohol or Drug Problem: Effective Recovery Strategies. Oxford University Press, New York.

Foa, E., Hembree, E., Rothbaum, B., 2007. Prolonged Exposure Therapy for PTSD: Emotional Processing of Traumatic Experiences, Therapist Guide. Oxford University Press, New York.

Foa, E.B., Yadin, E., Lichner, T.K., 2012. Exposure and Response (Ritual) Prevention for Obsessive-Compulsive Disorder: Therapist Guide. Oxford University Press, New York.

Forsyth, J.P., Eifert, G.H., 2008. The Mindfulness and Acceptance Workbook for Anxiety: A Guide to Breaking Free from Anxiety, Phobias, and Worry Using Acceptance and Commitment Therapy. New Harbinger Publications, Oakland, CA.

Gibbons, M.B., Crits-Christoph, P., Levinson, J., Gladis, M., Siqueland, L., Barber, J.P., Elkin, I., 2002. Therapist interventions in the interpersonal and cognitive therapy sessions of the treatment of depression collaborative research program. American Journal of Psychotherapy 56, 3–26.

Gilson, M., Freeman, A., Yates, M.J., Freeman, S.M., 2009. Overcoming Depression: A Cognitive Therapy Approach. Oxford University Press, New York.

Hope, D.A., Heimberg, R.G., Turk, C.L., 2010. Managing social anxiety. In: Therapist Guide, second ed. Oxford University Press, New York.

Jensen-Doss, A., Hawley, K.M., Lopez, M., Osterberg, L., 2009. Using evidence-based treatments: the experiences of youth providers working under a mandate. Professional Psychology: Research and Practice 40, 417–424.

Karlin, B.E., Ruzek, J.I., Chard, K.M., Eftekhari, A., Monson, C.M., Hembree, E.A., Resick, P.A., Foa, E.B., 2010. Dissemination of evidence-based psychological treatments for posttraumatic stress disorder in the Veterans Health Administration. Journal of Traumatic Stress 23, 663–673.

Kazdin, A.E., Blase, S.L., 2011. Rebooting psychotherapy research and practice to reduce the burden of mental illness. Perspectives on Psychological Science 6, 21–37.

Kazdin, A.E., Rabbitt, S.M., 2013. Novel models for delivering mental health services and reducing the burdens of mental illness. Clinical Psychological Science 1, 170–191.

Kendall, P.C., Chu, B., 2000. Retrospective self-reports of therapist flexibility in a manual-based treatment for youths with anxiety disorders. Journal of Clinical Child Psychology 29, 209–220.

Kendall, P.C., Chu, B., Gifford, A., Hayes, C., Nauta, M., 1998. Breathing life into a manual: flexibility and creativity with manual-based treatments. Cognitive and Behavioral Practice 5, 177–198.

Kroenke, K., Spitzer, R.L., Williams, J.B.W., Monahan, P.O., Lowe, B., 2007. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Annals of Internal Medicine 146, 317–327.

Lock, J., Le Grange, D., Russell, G., 2012. Treatment Manual for Anorexia Nervosa: A Family-based Approach, second ed. The Guilford Press, New York.

Luborsky, L., DeRubeis, R.J., 1984. The use of psychotherapy treatment manuals: a small revolution in psychotherapy research style. Clinical Psychology Review 4, 5–14.

Luborsky, L., Singer, B., Luborsky, L., 1975. Comparative studies of psychotherapies: is it true that "everyone has won and all must have prizes"? Archives of General Psychiatry 32, 955–1008.

May, P.R.A., 1973. Research in psychotherapy and psychoanalysis. International Journal of Psychiatry 1, 78–86.

McHugh, R.K., Barlow, D.H., 2010. The dissemination and implementation of evidence-based psychological treatments: a review of current efforts. American Psychologist 65, 73–84.

McHugh, R.K., Barlow, D.H., 2012. Dissemination and Implementation of Evidence-based Psychological Interventions. Oxford University Press, New York.

Miklowitz, D.J., 2010. Bipolar Disorder, Second Edition: A Family-focused Treatment. The Guilford Press, New York.

Mynors-Wallis, L.M., Gath, D.H., Day, A., Baker, F., 2000. Randomised controlled trial of problem solving treatment, antidepressant medication, and combined treatment for major depression in primary care. British Medical Journal 320, 26–30.

Nathan, P.E., Gorman, J.M., 2007. A Guide to Treatments that Work, third ed. Oxford University Press, New York.

Nelson, T.D., Steele, R.G., Mize, J.A., 2006. Practitioner attitudes towards evidence-based practice: themes and challenges. Administration and Policy in Mental Health and Mental Health Services Research 33, 398–409.

Nezu, A.M., Nezu, C.M., D'Zurilla, T., 2012. Problem-solving Therapy: A Treatment Manual. Springer Publishing Company, New York.

O'Connor, K.J., Ammen, S., 2012. Play Therapy Treatment Planning and Interventions: The Ecosystemic Model and Workbook, second ed. Academic Press, Waltham, MA.

Paul, G.L., 1967. Strategy of outcome research in psychotherapy. Journal of Consulting Psychology 31, 109–118.

Perepletchikova, F., Treat, T.A., Kazdin, A.E., 2007. Treatment integrity in psychotherapy research: analysis of the studies and examination of the associated factors. Journal of Consulting and Clinical Psychology 75, 829–841.

Resick, P.A., Monson, C.M., Chard, K.M., 2007. Cognitive Processing Therapy: Veteran/Military Version. Department of Veterans Affairs, Washington DC.

Roth, A.D., Pilling, S., 2007. The competences required to deliver effective cognitive and behavioural therapy for people with depression and with anxiety disorders. Retrieved from: http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_078537.

Strosahl, K., 1998. The dissemination of manual-based psychotherapies in managed care: promises, problems, and prospects. Clinical Psychology: Science and Practice 5, 382–386.

Strupp, H.H., 1964. The outcome problem in psychotherapy: a rejoinder. Psychotherapy: Theory, Research and Practice 1, 101.

Wilson, G.T., 1996. Manual-based treatments: the clinical application of research findings. Behavior Research and Therapy 34, 295–314.

Woods, D.W., Twohig, M.P., 2008. Trichotillomania: An ACT-enhanced Behavior Therapy Approach. Oxford University Press, New York.

Māori Social Work

Oliver Helena Fiona Te Momo, The School of Māori Studies, Knowledge & Education (AKE), Massey University, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In Aotearoa/New Zealand, the term 'Māori Social Work' emerged from a bicultural approach to the delivery of social services. This term has gained recognition in academic forums and the community as Māori and non-Māori authors attempt to provide descriptions and definitions to deconstruct the meaning. Māori Social Work can be located in the Western worldview and the Māori worldview. Within these worldviews the term weaves together theory and practice. It represents formal and informal types of social service delivery and is embedded in the philosophies of matauranga Māori (Māori knowledge and Māori education). Māori Social Work is broad, holistic, and grounded in tikanga (customs) and reo (language).

Introduction

Māori Social Work is communal social work. Māori people have always exercised aroha (compassion, love, help, and support), awhi (helping people), kaitiakitanga (guardianship), mahi (work), manaakitanga (caring for people, living things, heaven, and earth), tautoko (support for another), and whakawhanaungatanga (relationships, getting to know each other, and building relationships). For Māori, these exercises are the foundations of communal life in Te Ao Māori (Māori world). Māori, the native aboriginal and Indigenous ethnicity to Aotearoa/New Zealand (Aotearoa), signed the Treaty of Waitangi (Treaty) with the British Crown in 1840. The term Māori did not exist, instead chiefs from hapū, the Indigenous name for clans, signed the Treaty with representatives of Queen Victoria from Britain. When these two different ethnicities, Māori and European (Pākēhā), signed the Treaty, this contract gave birth to a bicultural nation. Māori culture thrived and they represented the majority population, but the British dishonored their part of the Treaty. Britain colonized Aotearoa through legislation, education, religion, diseases, and force to overthrow and subjugate Māori people and assume power.

From the colonization of Aotearoa, a bicultural view of the world developed. These were the Eurocentric colonial world, referred to as the Western world, and the world of the native aboriginal that was colonized, the Māori world. Subsequently, the two worldviews coexisted for over a century. Understanding the Māori worldview broadens a person's knowledge of Indigenous social work. A century later, the importance of the Māori people and the Treaty became recognized in bicultural social work practices in Aotearoa. This article explains Māori Social Work through a Māori lens and discusses the impact from Western worldviews to Māori knowledge, practice, and the delivery of social services that shaped Māori Social Work.

Western Worldview

Western worldviews of social work dominate the literature. The International Federation of Social Work (IFSW) and the International Association of Schools of Social Work say social work is a profession and an academic discipline. This Western worldview defines social work as promoting.

> social change, problem solving in human relationships and the empowerment and liberation of people to enhance well-being. Utilising theories of human behaviour and social systems, social work intervenes at the points where people interact with their environments. Principles of human rights and social justice are fundamental to social work.
>
> (IFSW, 2012, p. 1)

The definitions provided a singular view and neglected an Indigenous worldview of social work giving Western views power over Indigenous ones. The inclusive and supportive language used in this definition fashions knowledge to be homogeneous and portrays a universal worldview of social work. Sewpaul and Jones (2004) argue that the definition does not purport to provide a universal worldview:

> The extent to which schools of social work meet the global standards will depend on the developmental needs of any given country/region and the developmental status of the profession in any given country, as determined by unique historical, sociopolitical, economic, and cultural context.
>
> (Sewpaul and Jones, 2004, p. 226)

In a country where Indigenous people have been colonized, the definition is accepted by the majority as the indicator for measuring quality standards attached to social workers and social work despite local reality being different. The approach to dominate the discourse can be considered colonization by words applied in a subtle manner. It can be reinforced through legislation, government policies, academic success, rules, and regulations. This reinforcement can be seen in the Social Workers Registration Act 2003 and supported by the Social Workers Registration Board that "maintains a focus on the introduction of mandatory registration, which was one of the principal recommendations resulting from their 2011/12 review of the Act" (New Zealand Parliament, 2013, p. 3).

Traditionally, social work for Māori did not exist, instead Māori exercised values such as manaaki (to care for), aroha, kaitiakitanga, and rangatiratanga (leadership) when working

with tangata (people). In the last decades, the notion of Māori Social Work has emerged and values such as manaaki, aroha, kaitiakitanga, and rangatiratanga have reflected the common interpretations of social work that Māori exercised mainly as volunteers. "For many Māori, voluntary activity (acts of service to the people of the whānau, hapū and Iwi) was seen as an expression of the Māori philosophical cornerstones of collective consciousness, collective wellbeing, and collective responsibility" (Gray, 2002, p. 41). Gray (2002) stipulated that notions of 'voluntary work' had no traditional equivalent and providing services to people were an overarching philosophy in Māori families that entwined the values of voluntary work and social work. This is shown in the results of a research on voluntary activity by the New Zealand Office for the Community and Voluntary Sector.

> Māori were more likely than non-Māori to have been involved in unpaid activities outside the household….The fact that Māori women undertake more unpaid work outside the home than Māori men has been attributed to particular gendered roles that are ascribed to women on the marae.
>
> (Office for the Community and Voluntary Sector, 2007, p. 3)

Professional social work, according to Aotearoa NZ Association of Social Work (ANZASW) (2008, p. 2), "is focused on problem solving and change. As such, social workers are change agents in society and in the lives of the individuals, families and communities they serve." Kuther (2013, p. 1) describes a professional social worker as a person that helps "manage their daily lives, understand and adapt to illness, disability, and death, and obtain social services, such as health care, government assistance, and legal aid." Helping people can be a challenging task if the worldview one sees is not Western. Hart (2010) identifies this challenge and concurs that

> in any society there is a dominant worldview that is held by most members of that society…Alternative worldviews do exist, but they are not usually held by a majority of a society….work with Indigenous peoples will often require us to act outside of the dominant worldview found in social work internationally and particularly in fourth world territories.
>
> (Hart, 2010, p. 2)

A research report, *Puao-te-Ata-Tu*, produced results that government agents working in the Department of Social Welfare were racist and the quality of service to Māori was poor (Ministerial Advisory Committee on a Māori Perspective for the Department of Social Welfare, 1986). The research prompted a Minister of Parliament to review the Department of Social Welfare and adopt the notion of involving whānau when working with an individual client. The Children, Young Persons and Their Families (CYPF) Act 1989 was introduced that led to the creation of the Family Group Conference (FGC). "Social workers responded to the challenges of *Puao-te-Ata-tu* by developing practices that encouraged a greater degree of family/whānau involvement both in decision-making and with respect to hapū and extended family kinship care" (Connelly, 2004, p. 2).

FGC is a meeting between a young person who has offended his/her whānau, victims, and other people. The other people can represent various professionals like the police, counselor, a social worker, or youth advocate, who contribute to discussions with the whānau, victim, and youth. The formal meeting is organized by government agents with three main stages. Stage one is time for the professionals to inform whānau of their issues and concerns. In stage two, whānau are given time to discuss the information they have received and develop a plan to address the issues or concerns. In stage three, the plan is presented to the professionals, victims, and others and modified to suit the situation with feedback/contributions by all those involved (Child, Youth and Family, 2013).

Understanding Indigenous worldviews, Hart (2010, p. 1) argues, is important for the development of social work theory and social work practice. Accordingly, "our worldviews affect our belief systems, decision making, assumptions, and modes of problem solving." From a Māori worldview, values and practices are grounded on matauranga and individuals are always part of a collective that share their problems in life. Individual confidentiality rights gave way to whānau rights meaning the social worker could share with whānau the social issues of the client. The use of the FGC process changed social work with Māori clients.

Academic institutions produce predominantly Western worldviews of social work. Dumbrill and Green (2008) discuss how these institutions have changed once they realized their oppressive impact upon Indigenous people from the delivery of educational content that failed to acknowledge the value of Indigenous knowledge. "Consequently, social work academies, particularly in Canada, the United States, Central and South America, as well as Australia and New Zealand, are attempting to include Indigenous knowledge in the curricula" (Dumbrill and Green, 2008, p. 490). They argue that academic institutions need to be

> sensitive to the ways [in which] European knowledge dominates the academy and [become] open to disrupting this domination. Disruption is crucial because, despite a commitment to diversity and inclusion, most social work academies continue to teach from a Eurocentric perspective in a manner that perpetuates the colonization of not only Indigenous people and knowledge, but all other peoples and knowledge that falls outside the dominant European paradigm.
>
> (Dumbrill and Green, 2008, p. 490)

Determining who is a social worker, from an Indigenous view, is not solely dictated by academic institutions or governments. Neither should professionals and registered social workers be the only experts of social work practice for Indigenous people. Dumbrill and Green (2008, p. 500) assert that academic qualifications like a doctorate do not signal an automatic ability to "understand or teach Indigenous knowledge nor does it signify the ability to evaluate student's learning in this area. Even those with PhDs based in feminist or post-modern epistemologies cannot be relied on to understand or teach Indigenous knowledge." They say the origin of these academic qualifications for Europeans can limit their understanding of Indigenous people, and suggest "for

Indigenous communities and thinkers, not European thinkers no matter how progressive they might be, to determine the Indigenous knowledge to be taught and the standards to be applied when evaluating this learning," placing control of Indigenous knowledge in Indigenous hands. Dumbrill and Green (2008, p. 500) explain that their position is not to deter Aboriginal scholars from being high achievers in academia, but to stipulate that "a PhD is not the only criteria within academia that should count towards who is recognized as a Professor with knowledge that has value." Subsequently, community experts that may not possess formal qualifications, or be considered as professionals, but acknowledged for their work with Indigenous communities, can enter this knowledge realm. Bell (2006, p. 16) shows how the discipline of social work for Māori can assist in bridging many communities. She says "if a facilitated process of decolonisation has a role to play in healing the effects of colonisation amongst Māori and other indigenous peoples, then social work could be a context in which this facilitated process of decolonisation could be accessed."

Māori Worldview

The Māori worldview locates knowledge in Māori tikanga and reo. Maintaining Māori tikanga and reo in social work is a complex undertaking because it requires many attempts, to compete against, change or challenge, the mindset of people that hold fast to the Western worldview. Pohatu (2003, p. 2) says, "the ability and willingness to construct, apply and defend Māori Social Work options" can be scary and exciting with liberating possibilities. He uses a Māori proverb to explain his intellectual position.

> Māori world-views it is proposed here, demonstrate the 'ka hao te rangatahi' angle of the whakatauki (valued and wise cultural sayings)…in this context is Te Ao Māori (Māori thinking and its world-views) and what it offers social work theory and practice, especially in Aotearoa.…in its dialectical role, is consequently interpreted here as recognising the limitations that exist within current social work theory, practice and so their educational learning pathways. The need to respond proactively to the social work reality within Aotearoa is the constant companion in this dialogue.
> (Pohatu, 2003, p. 1)

Pohatu (2003) shows how language educates, informs, influences behavior, describes, and shapes the actions of people as a cultural norm. "Simple words used daily within social work environments, exemplify this force. The words, 'client' and 'customer' have their own bodies of knowledge and rationalizations" (Pohatu, 2003, p. 1) and relationships are connected to conduct, and the thought processes between those in need, and those who provide a service for them. He informs Māori people to look at their behavior, the language used, and the ethical implications of this work. Māori language, to him, is the kaitiaki (guardian) and the spoken word is an energy that permeates through a person with entry points to deeper readings because "Man cannot tune in so to speak when he is incapable of responding to the vibrations of the language" (Pohatu, 2003, p. 2). Ruwhiu (1999, p. 58) also speaks of how words are used in social services. He says the term Māori "is often used interchangeably with tangata whenua. What does this all mean for non-Māori social workers dealing with Māori clients? Simply, to be prepared to mediate in the mouth of the tiger." In addition, he adds, "the foundation stones of social work partnership, resistance, and activism for both Māori and non-Māori social workers have been refined on the raging fires of Te Tiriti o Waitangi" (Ruwhiu, 1999, p. 54), the Treaty of Waitangi.

Hollis-English (2012) illustrates how the knowledge from the writings of Pohatu (2003) explains Māori Social Work theory. Pohatu's use of the word 'Āta' is drawn out specifically to show the Māori theoretical position of social work. She does this to make it clear that nurturing and growing relationships are useful to a social worker. Hollis-English (2012, p. 47) communicates "social services as a transformative approach for Kaupapa Māori initiatives. It is about building and looking after relationships and can support practitioners with understanding relationships and wellbeing when interacting with Māori." She comments that Āta described by Pōhatu is a "behavioral and theoretical strategy for building and maintaining relationships. This principle is particularly relevant to the analysis of Māori social work practice and experiences as it helps to understand how these domains function and interconnect" (Hollis-English, 2012, p. 47).

Māori Social Work is not separated or divorced from the politics and public issues in Aotearoa. For example, a surge in social problems like domestic violence, criminal activity, and child abuse automatically leads to the media overemphasizing/sensationalizing Māori involvement in these activities (Merchant, 2010). This type of media coverage can spread negative interpretations of Māori causing the mainstream population to act inappropriately toward them. History gives numerous examples of this inappropriate behavior (Merchant, 2010). In turn, Māori become skeptical about professionals and disengage causing a breakdown of relationships with social workers.

Walsh-Tapiata (2008, p. 115) argues, "cultural legitimacy of Māori knowledge and values are being transposed into everyday practice." She is encouraged by Indigenous writers leading the transformation of social work practice. Institutions that teach social work, she considers, are "starting to see a trickle of students who come from such backgrounds, and an increasing number of Māori practitioners who have a strong cultural identity are effecting changes in practice across a range of the helping professions" (Walsh-Tapiata, 2008, p. 115). Māori social workers, according to Hollis (2006), identify with Te Ao Māori "and are guided by its worldview" (Hollis, 2006, p. 39), that is, closely associated with the Treaty of Waitangi. This relationship between social work practice and the Treaty of Waitangi is shared by Walsh-Mooney (2009, p. 69), who reinforces the notion that Māori social workers "are informed by a Māori worldview."

Māori Social Work

ANZASW Social Work Registration Board developed a bicultural policy in their Code of Ethics making it distinctive.

In 1993, the ANZASW set up a committee in Aotearoa that managed complaints from people and organizations, and established a Code of Ethics and a Bicultural Code of Practice. "The Bicultural Code of Practice continues to be incorporated in the ANZASW Code of Ethics in recognition of the Treaty relationship" (ANZASW, 2013, p. 2). This development "attempts to model an ideal treaty relationship. It does so, not only through being bilingually written, but also by developing core values, one of which is about ensuring members are committed" (Hollis-English, 2012, p. 40), and thereby making "social service legislation, structures, organisations and social work practice grounded in the Articles of Te Tiriti o Waitangi" (Aotearoa Association of Social Workers, 2008, p. 1). From an academic position, Māori professionalism in social work is promoted.

Professionalism and registrations influence Māori Social Work. Walsh-Tapiata (2008, p. 114) considers that the Social Work Registration Act (2003) presents "a new set of issues for Māori social workers, many of whom have worked much of their lives in their communities but who may not have a recognized social work qualification." Māori social workers enroll in courses offered by academic institutions that are designed to meet the requirements of the Act. This enrollment challenges the values of voluntary work and social obligations that Māori community workers predominantly undertake without being educated or registered. The government funds a range of social service organizations like Charitable Trust, Incorporated Societies, Māori organizations, and Iwi organizations to deliver social services to Māori. "Māori organizations can become established under the Māori Trust Boards Act, the Māori Community Development Act and Te Ture Whenua Act, which differentiates them from the realm of incorporated societies and charitable trusts" (Hollis-English, 2012, p. 69). Eligibility for funding can influence the functions of a Māori organization to deliver social work.

The insertion of Māori Social Work values must be heralded when placed within legislation, models of practice, and standards of competency for social workers "to practice with Māori clients. This legislation has increased the number of Māori social workers becoming members of the Aotearoa Association of Social Work (ANZASW)" (Walsh-Tapiata, 2008, p. 114). Subsequently, the Tangata Whenua Social Workers Association as a separate body to ANZASW was formed and established a platform to discuss Indigenous issues relating to social work. This Association provided a Māori worldview of the social work profession.

> Rangatiratanga has always been the aspiration of New Zealand's indigenous social workers. The development of 'by Māori for Māori' services has grown progressively over the past twenty years. Māori practice models, Māori frameworks, Māori fields of practice and Iwi/Māori social services are all assertions of rangatiratanga. An Association of Social Workers that commits to support Māori practitioners is a natural progression of the aspiration.
>
> (Tangata Whenua Social Workers Association, 2013, p. 1)

The vision of tangata whenua practitioners to provide advice that is adopted in social work policy became an accomplished dream. "The need to develop an ethical framework, practice standards, competencies and professional development opportunities that acknowledge mohiotanga (knowledge), matauranga (education), maramatanga (understanding), matatautanga (proficiency and fluency) and marautanga (curriculum and syllabus) provided the impetus for the establishment of the Tangata Whenua Social Workers Association" (Tangata Whenua Social Workers Association, 2013, p. 1).

Legislation can impact negatively on Māori Social Work. This is seen in the policies, rules, and regulations regarding the care of children for families of Māori ethnicity. Barret and Connolly-Stone (1998) critically analyze the CYPF Act 1989. They highlight the hypocrisy of the legislation with regard to the Treaty because it mentions "the special needs of Māori, but avoids creating Treaty obligations for Government by not mentioning the Treaty" (Barret and Connolly-Stone, 1998, p. 40).

> The CYPF Act requires that families and Whānau, Hapū, Iwi and family groups be involved in deciding the best ways for addressing care and protection needs or dealing with offending behaviour. It is the first legislation that defines family as any person who has a kinship relationship to a child or young person or any person with whom the child or young person has formed a significant bond. In the case of Māori children and young persons, the Act recognises that family groups can include all members of a Hapū or Iwi.
>
> (Department of Social Welfare, 1997, p. 1)

In 2003, a Chief Executive Officer (CEO) of an Iwi social service agency was interviewed by the Ministry of Justice. The commentary showed a lack of skilled professionals who understand matauranga Māori and the impact of legislation on Māori Social Work. Without this skilled comprehension, social work professionals are limited to the services they can provide Māori.

> The philosophy of the CEO, and the philosophical base of the organisation are clearly grounded in the Maori worldview described in the literature search. Whakapapa is the basis for Whānau engaging with the Iwi Social Service, and children who whakapapa to this Iwi are seen as belonging to the Whānau, Hapū and Iwi. Whānau, Hapū and Iwi are responsible for the care of the children. A fundamental concern for Iwi Social Service providers is that these principles, which are referred to in the Children, Young Persons, and Their Families Act 1989, are incorporated into other Family Court law. Consistency at this level should be supported by trained professionals and these principles applied in an informed way.
>
> (Ministry of Justice, 2002, p. 2)

Understanding Māori Social Work involves entering a new realm for many social workers. Pohatu (2003), Walsh-Tapiata (2008), Walsh-Mooney (2009), and Hollis-English (2012) identify mātauranga Māori as common words associated to Māori Social Work. Mātauranga Māori is a broad concept that covers tangible and intangible knowledge and cannot be compartmentalized into a basic definition. Rather, in the social work discipline it expands to describe the spirit and soul of the social worker and client. "This includes tikanga practices that Māori social workers have developed and applied to social work situations for the benefit of Whānau

Māori" (Hollis-English, 2012, pp. 16, 17). Hollis-English makes direct references to Māori Social Work to suggest, "knowledge that has come out of Te Ao Māori, the various contemporary methods and models that have been developed, and the ability to adapt these to the various Māori families with Aotearoa/New Zealand" (Hollis-English, 2012, p. 17).

Whakawhanaungatanga is entrenched in Māori Social Work. "For Māori families it is necessary to work collaboratively with tribal and Māori organizations who often have the insider networks and practices necessary to work effectively with their own people" (Walsh-Tapiata, 2008, p.112). Social workers can find certain behavior unacceptable, like a Māori client agreeing to everything suggested and entering agreements without understanding the punishment incurred from failure to comply. Whakawhanaungatanga can deter this behavior and enable the social worker and client(s) to get to know each other to generate positive results. When social workers work collaboratively with tribal and Māori organizations, the outcomes of success for the Māori clients are greater. Whakawhanaungatanga can assist in the prevention of an offense, a lessening of fear, and the fashioning of an agreement that the whānau and social workers are more likely to achieve. In Māori Social Work, whānau and the client are entwined and not viewed as independent of each other. Rather, the client is the whānau, and the whānau is the client as all members of the family are affected, one way or another, with the concern at hand that has brought the social worker(s) in front of them.

Māori Social Work embraces mahi aroha (working for compassion and love). This notion, mahi aroha, is strong in the community and can be delivered by professionals in paid positions or volunteers who are unpaid. Hollis-English (2012, p. 69) says, "government changes such as the employment of Māori staff and the CYPF Act were also affecting NGOs, Māori had different views of accountability and what volunteering entailed." In addition, Māori operated social services within the Māori organizations from whānau, hapū, and Marae (tribal meeting place). These organizations are less formal, very active in the community, and hold more autonomy over the delivery of social services. In the Māori world, "the underlying motivation to participate in unpaid work was the knowledge that to work the land that fed and housed the Whānau required unpaid labor. This knowledge spiritually bonded the members to the concepts of reciprocity and mutual responsibility" (Te Momo, 2003, p. 3). Tennant (2001) stipulates, when Indigenous voices began to be heard, the contribution by Māori voluntary organizations to the delivery of social services became noticed. As a consequence, government recognized the capacity of these organizations and a "shift in resources to Iwi and Māori-run services grew, in part, from a sense of historical injustice. In this context, Māori groups are not simply competitors for government support, but Treaty partners" (Tennant, 2001, p. 157).

Whānau are imperative to the health and well-being of social work clients. Hollis-English (2012, p. 196) purports, "Whānau as a collective group is/are instrumental in shaping Māori Social Work development and the lives of Māori social workers and second that it is a key priority area of Māori that work in the social work field." Adding to this field of social service delivery was Te Momo's (2003) earlier writings on Māori volunteerism:

> For Māori, participating in voluntary associations strengthened the new types of Māori communities and provided a way to address social problems by taking on the responsibilities to care for families. However, the vehicle for this early strength, the integration of Pakeha values and ways of working may also be seen to have contributed to the undermining of Māori ways of doing things that manifested Māori identity and values through Māori ways of working.
> (Te Momo, 2003, p. 6)

Whānau Ora, family well-being, is a founding principle that has underpinned Māori Social Work. Māori philosophers and practitioners like Sir Mason Durie, and Māori members of parliament, for example, the Honorable Tariana Turia, battled to create a policy named Whānau Ora. This policy impacted directly on social service organizations that accessed funding to deliver a social service to Māori and non-Māori. Research was initiated to "bring together government, providers and whānau, as the model provides for whānau having the opportunity to participate and voice their expectations from service delivery; and providers supported to design and deliver quality holistic services to whānau" (Turia, 2010, p. 4). Te Puni Kōkiri (2011, p. 1) comments that Whānau Ora, launched in 2010, was "an inclusive approach to providing services and opportunities to families across New Zealand. It aims to empower families as a whole, rather than focusing separately on individual family members and their problems." From 350 providers, a group of 25 was selected to deliver the services.

Māori Social Work resides in Māori development. Durie (2003) interprets Māori development to be holistic and have three core areas of development: social, political, and economic. Throughout extensive investigation of Māori health and well-being that produced many publications, Durie described how the three core areas overlapped, influenced, impacted, hindered, and improved the lives of Māori people. He continues to advocate for whānau to take control of life and live as Māori. The Māori Social Work practice operates in these core areas, for example, Māori require financial resources (economic development initiatives), to make informed decisions (political development initiatives), and live a healthy life (social development initiatives).

A social worker is unable to provide adequate social service to Māori without understanding their worldview. The reality, in the social work profession, is that Westerners need to be aware of how Māori development, matauranga Māori, and Māori Social Work are interrelated. Consequently, it is no longer acceptable for social workers to allege that they are ignorant of Māori culture, make excuses for not learning matauranga Māori, or decide to dismiss the vibrancy of Māori Social Work Theory.

Conclusion

Māori Social Work theory and practice is grounded in Māori principles. The practice will continue to evolve and reflect the

changes to government policy and Registration Board requirements. At the heart of Māori Social Work are the values aroha, awhi, kaitiakitanga, mahi, manaakitanga, and tautoko that are located in Te Ao Māori. Therefore, Māori would continue to provide social services regardless of money ensuring their tikanga and reo are not lost to the capitalist values of the Western world.

Understanding the history of Aotearoa assists social workers, work with Māori people. It is imperative to read the writings of an Indigenous population and build one's knowledge especially if a social worker is going to deliver a service to Indigenous people. Social workers must be aware of the impact from internal and external factors that occur in the social work profession, and determine ways to alleviate Māori people from the oppression they could be experiencing. This can be achieved through good community, tribal, and whānau relationships.

Whakawhanaungatanga is Māori Social Work in action. The positive relationships between agencies, institutions, communities, whānau, hapū, and Iwi are fundamental. Maintaining strong connections to the community, using tikanga and reo, are essential for constructive Māori Social Work. Also, realizing that whānau are the core to healing Māori families ensures Māori Social Work keeps true to Te Ao Māori because whānau are with a client before and after the social workers' job is completed. Therefore, applying a whanaungatanga technique is useful to the social worker, client, and their whānau.

Māori Social Work is unique to Aotearoa. New Zealanders, Māori and non-Māori, can take pride in knowing that this uniqueness sets their theory and practice apart from other countries. It provides an alternative application to social work literature, thought, and behavior that is not universal or global. That space has already been filled by Western knowledge. Instead, Māori Social Work can contribute to the growing discussions voiced by Indigenous people across the world and provide a guide for them to share their theories and practices. In conclusion, this article provided a range of descriptions to explain Māori social work.

See also: Indigenous Social Work; Neocolonialism and Social Work; Social Work Theory.

Bibliography

Aotearoa New Zealand Association of Social Workers, 2008. Submission on Proposed Changes to ACC Definitions. ANZASW, Christchurch.

Aotearoa New Zealand Association of Social Workers, 2013. Background of the Codes of Ethics. http://anzasw.org.nz/social_work_practice/topics/show/157-background-of-the-code-of-ethics (accessed 16.1.13.).

Barrett, M., Connolly-Stone, K., 1998. The Treaty of Waitangi and social policy. Social Policy Journal of New Zealand. 11, 29–47.

Bell, H.S., 2006. Exiting the Matrix: Colonisation, Decolonisation and Social Work in Aotearoa: Voices of Ngati Raukawa Ki Te Tonga Kaimahi Whānau. Master's thesis. Massey University, Palmerston North.

Child Youth and Family, 2013. Family Group Conferences. Available from: http://www.cyf.govt.nz/youth-justice/family-group-conferences.html (accessed 28.06.13.).

Connolly, M., 2004. A Perspective on the Origins of Family Group Conferencing. American Humane FGDM Issues in Brief. Retrievable from: http://www.americanhumane.org/assets/docs/protecting-children/PC-fgdm-ib-fgc-origins.pdf (accessed 24.06.13.).

Department of Social Welfare, 1997. Iwi Social Services Pack. Department of Social Welfare, Wellington, NZ.

Dumbrill, G.C., Green, J., August 2008. Indigenous knowledge in the social work academy. Social Work Education: the International Journal 27 (5), 489–503.

Durie, M., 2003. Nga Kahui Pou. Launching Māori Futures. Huia Publishers, Wellington.

Gray, M.M., 2002. A kaupapa Māori framework for volunteering. In: Proceedings of Volunteering NZ Conference, Otautahi Christchurch, Inspiring Volunteering, Whakamana Te Kaupapa Aroha Ki Te Takata a Rohe Maria Hand. Department of Internal Affairs and Volunteering New Zealand, Wellington, pp. 41–43.

Hart, M.A., 2010. Indigenous worldviews, knowledge and research: the development of an Indigenous research paradigm. Journal of Indigenous Voices in Social Work 1 (1), 1–16.

Hollis, A., 2006. Puao-te-Ata-tu and Māori Social Work Methods. Thesis, Master of Arts. University of Otago, Department of Social Work and Community Development, Otago.

Hollis-English, A. Māori Social Workers: Experiences within Social Service Organisations, Doctorate thesis. University of Otago, Department of Social Work and Community Development (Otago), Dunedin, unpublished.

International Federation of Social Work, 2012. Social work definition adopted by the IFSW General meeting in Montréal, Canada, July 2000. Also Agreed in GA with IASSW in Montreal 2000. http://ifsw.org/policies/definition-of-social-work/ (accessed 15.1.13.).

Kuther, T., 2013. What Is a Social Worker?. http://gradschool.about.com/od/alliedhealth/g/socialworkerdef.htm (accessed 14.1.13.).

Merchant, R. Who Are Abusing Our Children? An Exploratory Study on Reflections on Child Abuse by Media Commentators, Master's thesis. Massey University, Department of Social Work and Social Policy, Auckland, unpublished.

Ministerial Advisory Committee on a Māori Perspective for the Department of Social Welfare, 1986. Puao-te-ata-tu – Day Break: the Report of the Ministerial Advisory Committee on a Māori Perspective for the Department of Social Welfare. New Zealand. Department of Social Welfare, Wellington.

Ministry of Justice, 2002. Interview with an Iwi Social Service Provider. http://www.justice.govt.nz/publications/publications-archived/2002/guardianship-custody-and-access-maori-perspectives-and-experiences-august-2002/interview-with-an-iwi-social-service-provider (accessed 16.1.13.).

New Zealand Parliament, 2013. 2011/12 Financial Review of the Social Workers Registration Board. NZ Government, Wellington, NZ.

Office for the Community and Voluntary Sector, 2007. Mahi Aroha. Māori Perspectives on Volunteering and Cultural Obligations. Office for the Community and Voluntary Sector, NZ Government, Wellington.

Pohatu, T., 2003. Āta: Growing Respectful Relationships. http://www.kaupapamaori.com/assets//ata.pdf (accessed 14.1.13.).

Pohatu, T., 2004. Āta: growing respectful relationships. He Pukenga Korero. A Journal of Māori Studies 8, 1–8.

Pohatu, T. Māori Worldviews: Source of Innovative Social Work Choices. (Te Wānanga O Aotearoa, Manukau), unpublished manuscript.

Ruwhiu, L. Te Puāwaitanga O Te Ihi Me Te Wehi: the Politics of Māori Social Policy Development, Doctoral thesis. Massey University, Department of Social Work and Social Policy, Palmerston North, unpublished.

Sewpaul, V., Jones, D., 2004. Global standard for social work education and training. Social Work Education Volume 23 (5), 493–513.

Tangata Whenua Association Social Workers Association, 2009. Tangata Whenua Association Social Workers Association. Available from: http://twswa.org.nz/ (accessed 28.6.13.).

Te Momo, O.H.F. Demystifying a Relationship between Voluntary Work and Māori, Doctoral thesis. University of Waikato, Department of Development Studies, Hamilton, unpublished.

Tennant, M., 2001. Governments and voluntary sector welfare: historians' perspectives. Social Policy Journal of New Zealand (17), 147–160.

Te Puni Kōkiri, 2011. Whānau Ora in 2010. Putanga 21 Kohi-Tātea – Hui-Tanguru 2011. Kōkiri. Ministry of Māori Development, Te Puni Kōkiri, Wellington.

Turia, T., 2010. The Power to Define. Hon Tariana Turia. Minister responsible for whānau Ora. Associate Minister of Social Development and Employment Wednesday 1 December 2010, Māori Association of Social Sciences Conference – Ma Tau Rourou, Ma Taku Rourou. School of Languages and Social Sciences AUT University, AUT University, Auckland.

Walsh-Mooney, H. The Value of Rapport in Rangatahi Māori Mental Health: a Māori Social Work Perspective, Master thesis. Massey University, Department of Social Work and Social Policy, Palmerston North, unpublished.

Walsh-Tapiata, W., 2008. The past, the present, and the future: the New Zealand indigenous experience of social work. In: Gray, M., Coates, J., Yellow Bird, M. (Eds.), Indigenous Social Work around the World: towards Culturally Relevant Education and Practice. Ashgate Publishing Limited, England, pp. 107–115.

Relevant Websites

http://anzasw.org.nz/ – Aotearoa New Zealand Association of Social Workers. (accessed 16.01.13. and 27.06.13.).
www.minedu.govt.nz – Ministry of Education. (accessed 17.01.13. and 24.06.13.).
http://www.health.govt.nz/ – Ministry of Health. (accessed 16.01.13. and 24.06.13.).
http://www.justice.govt.nz/ – Ministry of Justice. (accessed 17.01.13. and 27.06.13.).
http://www.msd.govt.nz/ – Ministry of Social Development. (accessed 16.01.13. and 27.06.13.).
http://www.swrb.govt.nz/ – Social Work Registration Board. (accessed 16.01.13. and 27.06.13.).
http://twswa.org.nz/ – Tangata Whenua Social Workers Association. (accessed 16.01.13. and 27.06.13.).
http://www.tpk.govt.nz/en/ – Te Puni Kōkiri. (accessed 16.01.13. and 27.06.13.).

Marital Quality and Health

Deborah Carr, Institute for Health, Health Care Policy & Aging Research, Rutgers University, New Brunswick, NJ, USA
Tetyana Pudrovska, Population Research Center, University of Texas-Austin, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The protective effects of marriage on health are widely documented, yet recent research shows that these advantages are limited to high quality marriages. We describe the ways that marital quality affects a range of physical and mental health outcomes, with attention to gender, race, and life course differences therein. We also document how the effects of widowhood and divorce on health are moderated by the quality of the lost marriage. We highlight avenues for future research, including investigations of the linkages between relationship quality and health in cohabiting unions, same-sex partnerships, and remarriages.

The protective effects of marriage on health and well-being are widely documented. Studies dating back to Emile Durkheim's classic *Suicide* (1897) show that married persons enjoy better physical and mental health, including a lower risk of suicide, relative to their unmarried counterparts. Yet in recent decades, research has shown persuasively that all marriages are not equally beneficial, and that high quality marriages are particularly protective for health, well-being, and longevity – although the strength of this effect may vary by gender, race, and life course stage. In this article, we begin by defining 'marital quality' and 'health,' and show how these concepts are measured in empirical studies. Next, we describe how overall patterns of marital quality vary based on gender, race, and life course stage. We then describe key theoretical and methodological explanations for the protective effects of marital quality on health and well-being, and review recent empirical studies documenting statistical associations, with a focus on subgroup differences in the strengths of these patterns. Importantly, scholars have argued that one reason why marriage may be more beneficial to the health of men versus women, and whites versus blacks, is that men and whites typically report higher quality marital relations than their counterparts, and thus reap larger rewards. We also describe why and how the health-depleting effects of marital dissolution (i.e., divorce and widowhood) vary based on the quality of the dissolved marriage. Finally, we identify avenues for future research, and highlight the ways that research on marital quality and health can inform policies and practices to enhance well-being over the life course.

Marital Quality and Health: Concepts, Measures, and Methods

Marital Quality

Marital quality is a broad construct that encompasses a range of positive and negative marital interactions and perceptions. Its components may include interactions such as activities shared by the spouses, disagreements, and marital problems, as well as one's feelings toward their partner and level of satisfaction with the relationship. Some scholars go so far as to argue that divorce is the ultimate indicator of a poor quality marriage. Most empirical studies of marital quality focus on the specific dimensions of *marital happiness* or *satisfaction*, which are typically measured by asking an individual either a single question or a series of questions assessing how happy or satisfied one is with the marriage. Although satisfaction and happiness are conceptually similar, researchers draw the distinction that happiness taps one's feelings or emotions about the marriage, whereas satisfaction reflects a cognitive appraisal of one's relationship relative to some standard (e.g., what one hoped for in marriage, what others' marriages are like) (Campbell et al., 1976).

Most research on marital quality has focused on positively worded items such as "How happy are you in your marriage?" yet in recent years scholars have also assessed negatively worded items such as "my partner can get on my nerves," as well as the frequency of arguments and one's perceived likelihood that their marriage will dissolve. Researchers also may ask respondents to indicate how often they experience or how strongly they agree that a particular marital attribute characterizes their relationship. Items encompassing positive attributes typically refer to the amount of love, affection, support, and understanding one receives from one's spouse, whereas negative attributes include conflicts or excessive demands imposed by one's spouse. Scholars increasingly understand that positive and negative aspects of marriage are not opposite ends of a single pole; rather, positive and negative sentiments and interactions can coexist in a single relationship. For example, researchers have developed the measure of *marital ambivalence*, which encompasses high levels of positive and negative perceptions of one's spouse. This stands in stark contrast with *marital indifference*, which is marked by very low levels of both positive and negative sentiments (Fincham and Linfield, 1997).

Based on citations alone, the most widely used measure of marital quality is the Dyadic Adjustment Scale (Spanier, 1976), which comprises the subdimensions of troublesome dyadic differences, interpersonal tensions and personal anxiety, dyadic satisfaction, dyadic cohesion, and consensus on matters of importance to dyadic functioning (Spanier, 1976: p. 17). Other widely used measures include the Short Marital Adjustment Test (Locke and Wallace, 1959), and single-item global

measures such as "Taking things all together, how would you describe your marriage?"

Health

Most studies of marital quality and health rely on large random-sample surveys that obtain self-reported measures of physical and mental health, as well as the marital quality items described above. Researchers can then explore statistical associations between one's marital quality appraisals and self-reported health measures in order to understand linkages between the two. Health can be defined as the "state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" (World Health Organization, 2006). Consistent with this definition, studies of marital quality and health cast a wide net, linking marital characteristics to physical health outcomes including mortality risk and risk of particular diseases (e.g., heart disease, diabetes, and cancer) and health symptoms (e.g., high blood pressure, sleep problems); mental health outcomes such as depressive symptoms, anxiety, and substance use; and measures of social well-being like life satisfaction, psychological well-being, and self-esteem (Proulx et al., 2007).

One of the most important advances in the past two decades has been linking marital quality to physiological markers of health called 'biomarkers.' A growing number of population-based surveys now obtain biological indicators that provide information on one's cardiovascular functioning (e.g., blood pressure readings), endocrine systems (e.g., cortisol levels), and immune function (e.g., natural killer cell lysis); each of these biomarkers is associated with the development of particular diseases (Robles and Kiecolt-Glaser, 2003). Major surveys in the United States such as the Midlife in the United States (MIDUS), National Social Life, Health, and Aging (NSHAP), and Wisconsin Longitudinal Study (WLS) obtain detailed self-reports of one's marriage, physical health, and mental health, as well as biological measures such as heart rate and levels of stress hormones associated with the development of chronic diseases.

Mounting research also explores marital quality and health in laboratory settings. In such studies, researchers might induce a marital spat, and then explore whether the induced threat to marital quality affects physiological markers associated with illness, such as blood pressure, cortisol levels, or the speed with which a laboratory-induced injury will heal. For example, a team of researchers at Ohio State University brought into their laboratory 42 healthy married couples, whose ages ranged from 20 to 80. When the study started, the researchers gave each spouse a small (though relatively painless) cut on their forearm. Some study participants were then told to discuss a potentially contentious topic with their spouse, such as in-law woes or financial worries. The researchers found that the couples with the most heated arguments had significantly slower wound-healing, evidenced by how quickly the small forearm cut healed. Most subjects healed within 5 days, but those who had a heated discussion with their spouse took 1 full day more to heal. The researchers also obtained information on how the couple interacted during the argument. They videotaped the arguments and rated how 'hostile' the exchanges were; those couples rated as being hostile toward one another took fully 7 days to heal (Kiecolt-Glaser et al., 2005). Taken together, these studies reveal the broad range of measures and methods used to identify linkages between marital quality and health.

Subgroup Differences in Marital Quality

Before further delving into linkages between marital quality and health, it is important to first review social factors that affect marital quality. Patterns vary slightly based on the samples and methods used, yet most studies conclude that men, whites, and newlyweds report higher levels of marital quality than women, blacks, and longer-married persons, respectively. The former two differences, in particular, may partly explain why marriage is generally more protective for the health of men and whites, relative to blacks and women; those who experience the most rewarding marriages will reap the most profound health benefits. We briefly describe subgroup differences in marital quality, and potential explanations for these differentials.

Gender

One of the most widely documented findings in marital quality research is that women report lower levels of marital quality and satisfaction than men, although the magnitude of these differences is modest (Jackson et al., 2014). The main explanations given for women's slightly lower marital quality ratings relative to men's are: women's lower levels of power in marital relations; their elevated risk of experiencing domestic violence or mistreatment; the disproportionate burden that women bear regarding child care, housework, and other tasks essential to maintaining a household; their greater emotional investment in sustaining a successful marriage; and their greater willingness to acknowledge and address problems in the marriage (Jackson et al., 2014).

Race and Ethnicity

Surprisingly few studies have explored race and ethnic differences in marital quality, with most of this work focusing on black–white differences. Most studies conclude that blacks report higher levels of marital conflict and strain, higher rates of divorce, and lower levels of marital satisfaction than their white counterparts (Broman, 2005). The main explanations for these gaps focus on behaviors within the marriage, where blacks report more frequent fighting, substance use, and infidelity than their white counterparts (Broman, 2005). Blacks' elevated risk of socioeconomic adversities also may take a toll on one's marriage; stressors like unemployment, discrimination, and financial strain may contribute to marital tensions (Corra et al., 2009). Few studies focus on marital quality among Latinos, yet there is mounting evidence that Mexican-Americans report significantly higher levels of marital quality than blacks, although both subgroups experience elevated risk of socioeconomic adversities. Mexican-Americans' relatively high quality marriages are attributed to a familistic pro-marriage culture that supports and sustains marriage in many Latino communities (Bulanda and Brown, 2006). The high quality marriages documented among Mexican-Americans also may

contribute to the well-known 'Latino paradox' in health; Mexicans report relatively good health, despite their poor economic status compared to whites – perhaps due to the health-enhancing benefits received by their strong marriages.

Marital Duration

Marital quality changes with the duration of marriage, where satisfaction waxes and wanes over time, as expectations and social roles change (Brown et al., 2013). The family life cycle explanation suggests that marital satisfaction is high in the very early stages of a marriage (e.g., 'the honeymoon stage'), and then declines as the spouses face increasing demands with work, childrearing, and home-making. As these stressors diminish over time, marital quality tends to increase in the later years of marriage, once children have left the family home. This is also described as a U-shaped curve (Orbuch et al., 1996). By contrast, a second conceptual model called the disillusionment model presumes that newlyweds hold onto idealized views of their partners, yet after the honeymoon stage passes, spouses may become disappointed, disillusioned, or upset with one another. Whereas the life cycle model posits a U-shaped curve, the disillusionment model presumes that marital quality will either continue to decline, or the marriage will be terminated. Finally, the enduring dynamics model or maintenance hypothesis counters that while unrealistic expectations may be dashed in the early stages of marriage and marital quality may dip after the initial heady years, quality remains relatively steady thereafter, as the partners adjust their expectations to the realities of their lives (Brown et al., 2013).

Marital Quality and Health: Theoretical and Methodological Explanations

Why and how does marital quality affect health? Three main classes of explanations are offered: (1) social explanations emphasize the emotional support and health-enhancing assistance exchanged in high quality marriages (e.g., Umberson et al., 2006); (2) biological explanations focus on physiological responses to close and nurturing relationships (e.g., Robles and Kiecolt-Glaser, 2003); and (3) methodological explanations emphasize the concept of 'social selection,' where those with better health are predisposed to experiencing more satisfying and enduring marriages (e.g., Butterworth and Rodgers, 2006).

Social Support and Integration

High quality social relationships, in general, and marital relations, in particular, are a well-documented source of positive health (Robles et al., 2014). A high quality marriage can provide the sense of security and emotional support that helps to bolster mental health. Happily married persons also enjoy more satisfying and frequent sexual relations, which provide physical and emotional health benefits (Galinsky and Waite, 2014).

Marriage, especially high quality marriage, is also considered a source of 'social control' (Umberson et al., 2006). Spouses who love and care about one another will encourage the adoption of healthy behaviors and the loss of unhealthy ones. Husbands and wives may encourage each other to eat nutritious meals, to take their daily medications, to eschew or limit their smoking and alcohol consumption, and to engage in health-enhancing physical activities together. By contrast, persons in poor quality marriages are more likely to engage in risky health behaviors. Persons in unhappy marriages exhibit poor eating habits, erratic sleep patterns, and higher rates of smoking, alcohol use, and nonmedical use of prescription medications (Miller et al., 2013).

A high quality marriage also can protect against or 'buffer' the potentially health-depleting effects of stressors like unemployment, work strains, and illness (Cohen and Wills, 1985). For example, several studies of older married couples found that persons with high levels of physical disability, functional impairment, and impaired vision were far less likely to be depressed if they enjoyed high quality marriages, yet were more likely to be depressed if they were in marriages marked by frequent arguments and disagreements (Bookwala, 2011). Thus, marital quality can be a powerful resource that protects against mental health declines, especially as couples face the physical challenges of aging.

Biological Pathways

One of the most important scholarly advances of the past two decades has been the exploration of the physiological pathways through which marital interactions 'get under our skin' and affect physical health (Ryff and Singer, 1998: p. 214). The pathways most often considered are those that progress through multiple physiological systems, including cardiovascular, endocrine, immune, metabolic, and sympathetic nervous systems (Robles et al., 2014).

For example, laboratory studies show that maintaining physical contact and closeness with one's spouse when placed under a stressful experimental condition leads to decreases in blood pressure and heart rate, and an increase in the hormone oxytocin. Oxytocin weakens the potentially harmful impact of stress on cardiovascular and endocrine systems (Light et al., 2005). Similarly, in another laboratory study, women who held their husband's hand while experiencing painful electric shocks reported less pain, and showed decreased activity in the brain region responsible for directing the body's stress response. By contrast, women who held either a stranger's hand or no one's hand (control group) did not evidence such positive physiological responses, underscoring the distinctive effects of the marital relationship (Coan et al., 2006).

Negative interactions, such as marital strain and hostility, also compromise immune system responses (Robles and Kiecolt-Glaser, 2003). The immune system protects the body from bacteria and viruses by releasing antibodies and white blood cells into the circulatory system. In one experimental study, happily married couples were prompted to argue about a 'real problem' that they disagreed about. Spouses who showed supportive and constructive behaviors during the argument evidenced reduced stress hormone levels, with steeper effects among wives than husbands. Similarly, a prospective study of older persons in long-term marriages marked by higher levels of conflict and frequent negative behaviors evidenced significantly worse immune system

responses, compared to persons in marriages marked by low conflict and negative behaviors. Taken together, these studies reveal that couples who show the least negativity and most positivity in their marital relations enjoy the healthiest immune, cardiovascular, and endocrine system responses.

Social Selection

Thus far, the theories and data we have reviewed suggest that marital happiness can protect health, whereas marital conflict can hurt. Yet a mounting body of research suggests that marital quality and health may be mutually influential, and that poor physical and mental health may impede both one's own and one's spouse's assessment of marital quality. Physical health problems may put a damper on a married couple's sexual relationship, and may also place difficult caregiving demands on the healthier partner, thus compromising marital quality (Pinquart and Sorensen, 2011). Health problems may also affect one's ability to maintain steady work, which in turn may place financial strains on the couple. Mental health problems also may compromise a marriage. An emotionally volatile or depressed spouse may not be able to provide their partner the love, attention, and support they desire. Likewise, a spouse with a substance use disorder in particular may be likely to abuse their partner emotionally or physically, and may not contribute effectively to the daily household chores and activities that are at the centerpiece of a good marriage. Consistent with the notion that health problems may suppress marital quality, several studies have documented an elevated risk of divorce in couples where one or both spouses have a series physical or mental health condition (e.g., Butterworth and Rodgers, 2006).

Marital Quality and Health: Subgroup Differences

Gender

Does marital quality have more powerful effects on the health and well-being of women or men? Research is equivocal, with clinical and laboratory-based studies generally finding greater female vulnerability to marital strain, and population-based studies detecting few if any gender differences (Umberson and Williams, 2005). For example, some studies show that women experience greater physiological arousal in response to laboratory-induced arguments than do men, and this greater arousal, in turn, may have deleterious consequences for their physical health (Robles and Kiecolt-Glaser, 2003). Similarly, prospective studies of congestive heart failure patients show that marital quality is a stronger predictor of survival among women than men (Coyne et al., 2001). By contrast, survey-based studies and a recent meta-analysis (Robles et al., 2014) find relatively weak evidence of gender differences.

To the extent that gender differences exist, scholars have posited several explanations for the stronger linkages among women. First, interpersonal relationships, especially marriage, are considered more highly salient to the self-concept of women – whereas work-related accomplishments are considered more salient to the identities of men. As such, women may be more aware of and responsive to a relationship's ups and downs, and may evidence greater physiological responses to stressors related to their marriage (Robles et al., 2014). Second, some argue that women have lower social status than men in society, and this power imbalance follows men and women into marriage. Low status, in turn, is associated with greater levels of stress reactivity; thus, women may react more strongly to marital strains and suffer the health consequences thereof (Wanic and Kulik, 2011).

Race

Surprisingly little research has explored whether the ties between marital quality and health vary by race and ethnicity. This is an important omission. Some scholars have proposed that the well-documented black–white gap in nearly every health outcome ranging from mortality to cardiovascular disease to diabetes is partially due to the fact that blacks are less likely than whites both to be married, and to maintain high quality marriages. However, little is known about whether a high quality marriage provides richer health benefits to blacks versus whites.

Some scholars have argued that high quality marriage may be less protective (and strained marriages less deleterious) to the health of blacks versus whites because marriage may be less salient to blacks than whites. For example, blacks' social networks include a higher proportion of persons from their religious community and extended family, relative to whites. These networks may provide many of the supports that might otherwise be provided by spouses (Levin et al., 1994). However, one recent review of the literature concluded that marital quality figures equally into the health of blacks and whites (Blackman et al., 2005).

Life Course Stage

Mounting research shows that the protective effects of marital quality on health vary over the life course. In particular, the deleterious effects of marital strain on health intensify with age (Umberson et al., 2006). These findings are consistent with the core themes of socioemotional selectivity theory; this perspective holds that as adults grow older and their time horizons constrict, their closest relationships (such as marriage) become more salient to them than their other more distant social ties (Carstensen, 1991). Moreover, age-related biological vulnerabilities also explain these patterns; immunological impairment increases with age, and the stress associated with marital strain may further accelerate aging of the immune system (Kiecolt-Glaser and Glaser, 2001). Finally, cumulative advantage perspectives argue that the harmful consequences of exposure to stress may accumulate over time, rendering unhappily married older adults more vulnerable from the 'wear and tear' that they have experienced over the life course (Dannefer, 2003).

Marital Dissolution and Health: Does Marital Quality Matter?

Research consistently shows that widowed and divorced persons have poorer physical and mental health than their married counterparts, with effects documented for

diverse outcomes including self-rated health, mortality risk, cardiovascular disease, smoking-related cancers, functional limitations, chronic conditions, depression risk, and problem drinking (Hughes and Waite, 2009; Zhang and Hayward, 2006).

However, the health-depleting effects of marital dissolution are contingent upon the quality of the marriage that has been lost. Although early studies presumed that the stress surrounding spousal death or divorce would overwhelm one's ability to cope and thus impair one's health, more recent work recognizes that one's marital 'role history' must be taken into consideration. Sociologist Blair Wheaton (1990) observed that exiting a social role that provided relatively few rewards, such as a conflicted and unhappy marriage, may not lead to distress but may instead be a source of mental and physical uplift.

Empirical studies in generally support Wheaton's assertions. Several studies have tracked the health of persons who dissolved troubled marriages, and found physical and mental health gains among those who exited bad marriages, yet health decrements among those who remained married (Hawkins and Booth, 2005; Williams and Umberson, 2004). The authors concluded that dissolving an unhappy marriage may be more protective to one's well-being than staying married. However, other studies have found that exiting a troubled marriage does not enhance one's mental health, and in fact, may be worse than remaining in the strained marriage (e.g., Kalmijn and Monden, 2006). The authors did not argue that individuals should remain in problematic marriages; rather, many of the strains of a problematic marriage may persist after the dissolution, particularly when ex-spouses must negotiate difficult custody or financial issues.

Marital dissolutions due to spousal death are distinct from those due to divorce, as widowhood is almost always entered into involuntarily, whereas divorce is typically desired by at least one of the two partners. However, research on widowhood – like divorce – shows that the emotional toll of dissolution varies based on the quality of the late marriage. For example, Carr et al. (2000) investigated whether bereaved older adults' psychological adjustment to loss varied based on the levels of warmth, conflict, emotional dependence, and instrumental dependence the survivor had experienced in the marriage. They found that persons in very close and warm relationships experienced elevated grief symptoms, as they were exiting a social role that had provided rich emotional rewards.

Conclusions and Future Directions

Marital quality is protective for adult well-being because high quality marriages provide emotional and instrumental support as well as physiological rewards that enhance physical and mental health. Yet researchers continue to grapple with unanswered questions regarding marital quality and health. First, research on marriage and health generally shows that first marriages are more protective than are higher-order marriages (i.e., remarriages), cohabiting unions, and long-term same-sex relationships. One explanation offered for these patterns is that different types of unions provide different socioemotional benefits that, in turn, protect one's health and well-being. However, few studies recognize that relationship quality may vary widely even within particular types of unions; we urge researchers to identify the specific aspects of marital interactions in remarriages, same-sex unions, cohabiting unions, and dating relationships that are protective for health.

A second limitation of extant research is that most studies focus on the associations between one individual's marital quality assessments and his or her health, with little attention paid to the partner's views of the marriage. Husbands' and wives' reports of marital quality vary widely, typically yielding correlations of just 0.50 to 0.60 (Carr and Boerner, 2009). As such, researchers are increasingly interested in 'crossover' or 'partner' effects; this research explores whether one partner's reports of marital satisfaction affect the other spouse's health and well-being. For example, an unhappily married spouse may be less motivated to provide his or her partner with the emotional support and caregiving that enhances the partner's health. The development of statistical techniques like actor–partner independence models (Kenny et al., 2006), as well as the collection of couple-level data in large-scale sample surveys like the NSHAP, WLS, and Disability and Use of Time (DUST) daily diary supplement to the Panel Study of Income Dynamics, now enable researchers to explore how one's own health is linked to one's spouse's level of satisfaction and happiness in the marriage (Carr et al., 2014).

Understanding the complex ways that marital interactions affect health (and vice versa) is important from a public policy perspective. In the early 2000s, policy makers believed that being married was protective to the well-being of spouses and their children, thus President George H.W. Bush's administration supported social programs to encourage marriage. Yet in recent years, scholars have documented that not all marriage is good marriage, and that high quality marriages are protective for health. This research has triggered investment in programs aimed at enhancing communication, closeness, and cooperation in marriages (Staton, 2009). Interventions to facilitate communication, minimize conflict, and uphold high quality marriage may be an essential step to protecting the health of adults in the United States.

See also: Aging and Health in Old Age; Heterosexuality; Marriage; Motivation: Life Course and Sociological Perspectives; Personality and Marriage; Sexuality Over the Life Course; Social Relationships in Adulthood.

Bibliography

Blackman, L., Clayton, O., Glenn, N., Malone-Colon, L., Roberts, A., 2005. The Consequences of Marriage for African Americans: A Comprehensive Literature Survey. Institute for American Values, New York.

Bookwala, J., 2011. Marital quality as a moderator of the effects of poor vision. Journal of Gerontology Series B: Psychological Sciences and Social Sciences 66B, S605–S616.

Broman, C.L., 2005. Marital quality among blacks and whites. Journal of Family Issues 26, 431–434.

Brown, E., Orbuch, T.L., Bauermeister, J.A., McKinley, B., 2013. Marital well-being over time among black and white Americans: the first seven years. Journal of African American Studies 17, 290–307.

Bulanda, J., Brown, S., 2006. Race-ethnic differences in marital quality and divorce. Social Science Research 36, 945–967.

Butterworth, P., Rodgers, B., 2006. Concordance in the mental health of spouses: analysis of a large national household panel survey. Psychological Medicine 36, 685–697.
Campbell, A., Converse, P., Rogers, W., 1976. The Quality of American Life. Russell Sage Foundation, New York.
Carr, D., Boerner, K., 2009. Do spousal discrepancies in marital quality assessments affect psychological adjustment to widowhood? Journal of Marriage and Family 71, 495–509.
Carr, D., House, J.S., Kessler, R.C., Nesse, R.M., Sonnega, J., Wortman, C., 2000. Marital quality and psychological adjustment to widowhood among older adults: a longitudinal analysis. Journals of Gerontology Series B: Psychological Sciences and Social Sciences 55B, S197–S207.
Carr, D., Freedman, V.A., Cornman, J.C., Schwarz, N., 2014. Happy marriage, happy life? Marital quality and subjective well-being in later life. Journal of Marriage and Family 76, 930–948.
Carstensen, L., 1991. Socioemotional selectivity theory: social activity in life-span context. Annual Review of Gerontology and Geriatrics 11, 195–217.
Coan, J.A., Schaefer, H.S., Davidson, R.J., 2006. Lending a hand: social regulation of the neural response to threat. Psychological Science 17, 1032–1039.
Cohen, S., Wills, T.A., 1985. Stress, social support and the buffering hypothesis. Psychological Bulletin 98, 10–57.
Corra, M., Carter, S.K., Carter, J.S., Knox, D., 2009. Trends in marital happiness by gender and race, 1973 to 2006. Journal of Family Issues 30, 1379–1404.
Coyne, J.C., Rohrbaugh, M.J., Shoham, V., Cranford, J.A., Nicklas, J.M., Sonnega, J., 2001. Prognostic importance of marital quality for survival of congestive heart failure. American Journal of Cardiology 88, 526–529.
Dannefer, D., 2003. Cumulative advantage, and the life course: cross-fertilizing age and social science knowledge. Journals of Gerontology Series B: Psychological Sciences and Social Sciences 58B, S327–S337.
Durkheim, E., 1897[1951]. Suicide: A Study in Sociology. The Free Press, New York.
Fincham, F.D., Linfield, K.J., 1997. A new look at marital quality: can spouses feel positive and negative about their marriage? Journal of Family Psychology 11, 489–502.
Galinsky, A.M., Waite, L.J., 2014. Sexual activity and psychological health as mediators of the relationship between physical health and marital quality. Journals of Gerontology: Social Sciences. http://dx.doi.org/10.1093/geronb/gbt165.
Hawkins, D., Booth, A., 2005. Unhappily ever after: effects of long-term, low-quality marriages on well-being. Social Forces 84, 445–465.
Hughes, M.E., Waite, L.J., 2009. Marital biography and health at midlife. Journal of Health and Social Behavior 50, 344–358.
Jackson, J.B., Miller, R.B., Oka, M., Henry, R.G., 2014. Gender differences in marital satisfaction: a meta-analysis. Journal of Marriage and Family 76, 105–129.
Kalmijn, M., Monden, C.W., 2006. Are the negative effects of divorce on well-being dependent on marital quality? Journal of Marriage and Family 68, 1197–1213.
Kenny, D.A., Kashy, D.A., Cook, W.L., 2006. Dyadic Data Analysis. Guilford, New York.
Kiecolt-Glaser, J.K., Glaser, R., 2001. Stress and immunity: age enhances the risks. Current Directions in Psychological Science 10, 18–21.
Kiecolt-Glaser, J.K., Loving, T.J., Stowell, J.R., Malarkey, W.B., Lemeshow, S., Dickinson, S.L., Glaser, R., 2005. Hostile marital interactions, proinflammatory cytokine production, and wound healing. Archives of General Psychiatry 62, 1377–1384.
Levin, J.S., Taylor, R.J., Chatters, L.M., 1994. Race and gender differences in religiosity among older adults: findings from four national surveys. Journal of Gerontology Series B: Psychological Sciences and Social Sciences 49B, S137–S145.
Light, K.C., Grewen, K.M., Amico, J.A., 2005. More frequent partner hugs and higher oxytocin levels are linked to lower blood pressure and heart rate in premenopausal women. Biological Psychology 69, 5–21.
Locke, H., Wallace, K., 1959. Short marital-adjustment and prediction tests: their reliability and validity. Marriage and Family Living 21, 251–255.
Miller, R.B., Hollist, C.S., Olson, J., Law, D., 2013. Marital quality and health over 20 years: a growth curve analysis. Journal of Marriage and Family 75, 667–680.
Orbuch, T.L., House, J., Mero, R.P., Webster, P.F., 1996. Marital quality over the life course. Social Psychology Quarterly 9, 162–171.
Pinquart, M., Sorensen, S., 2011. Spouses, adult children, and children-in-law as caregivers of older adults: a meta-analytic comparison. Psychology and Aging 26, 1–14.
Proulx, C.M., Helms, H.M., Buehler, C., 2007. Marital quality and personal well-being: a meta-analysis. Journal of Marriage and Family 69, 576–593.
Robles, T.F., Kiecolt-Glaser, J.K., 2003. The physiology of marriage: pathways to health. Physiology & Behavior 79, 409–416.
Robles, T.F., Slatcher, R.B., Trombell, J.M., McGinn, M.M., 2014. Marital quality and health: a meta-analytic review. Psychological Bulletin 140, 140–187.
Ryff, C.D., Singer, B., 1998. The contours of positive human health. Psychological Inquiry 9, 1–28.
Spanier, G.B., 1976. Measuring dyadic adjustment: new scales for assessing the quality of marriage and similar dyads. Journal of Marriage and the Family 38, 15–28.
Staton, J., 2009. Making the Connection between Healthy Marriage and Health Outcomes: What the Research Says. Retrieved from the National Healthy Marriage Resource Center website: http://www.healthymarriageinfo.org/resource-detail/index.aspx?rid=3371.
Umberson, D., Williams, K., 2005. Marital quality, health, and aging: gender equity? Journal of Gerontology Series B: Psychological Sciences and Social Sciences 60B (Special Issue II), 109–112.
Umberson, D., Williams, K., Powers, D.A., Liu, H., Needham, B., 2006. You make me sick: marital quality and health over the life course. Journal of Health and Social Behavior 47, 1–16.
Wanic, R., Kulik, J., 2011. Toward an understanding of gender differences in the impact of marital conflict on health. Sex Roles 65, 297–312.
Wheaton, B., 1990. Life transitions, role histories, and mental health. American Sociological Review 55, 209–223.
Williams, K., Umberson, D., 2004. Marital status, marital transitions, and health: a gendered life course perspective. Journal of Health and Social Behavior 45, 81–98.
World Health Organization (WHO), 2006. Constitution of the WHO. Author, New York.
Zhang, Z., Hayward, M., 2006. Gender, the marital life course, and cardiovascular disease in late midlife. Journal of Marriage and Family 68, 639–657.

Relevant Websites

The following data resources include rich measures on physical health, mental health, and marital quality:

http://www.isr.umich.edu/acl/ – Americans Changing Lives (ACL).
http://psidonline.isr.umich.edu/DUST/dust09_UserGuide.pdf – Disability and Use of Time Supplement to the PSID.
http://midus.wisc.edu/ – Midlife Development in the United States (MIDUS).
http://www.norc.org/Research/Projects/Pages/national-social-life-health-and-aging-project.aspx – National Social Life, Health and Aging Project (NSHAP).
http://www.ssc.wisc.edu/wlsresearch/ – Wisconsin Longitudinal Study (WLS).

Marital Status, Divorce, and Child Development

Colleen M Carr and Sharlene A Wolchik, Arizona State University, Phoenix, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A significant proportion of children and adolescents experience parental divorce, and much research has focused on the impact of divorce on children's adjustment outcomes. This article reviews data on the prevalence of divorce and changes in divorce rates over time, summarizes results of literature on children's postdivorce adjustment outcomes, discusses societal costs of parental divorce, outlines theories of the effects of parental divorce, and describes factors associated with children's responses to parental divorce. The article concludes with a discussion of interventions shown to prevent or reduce children's postdivorce adjustment problems.

Different Societal Views of Divorce

To understand the impact of divorce on children's adjustment problems and well being, it is valuable to consider diversity in cultural attitudes toward divorce. Generally, individualistic (typically Western) societies have more favorable attitudes toward divorce than do collectivist societies. In individualistic cultures like the United States, it is normative for people to opt out of marriages that no longer enhance their self-interest or enable high levels of self-discovery and self-disclosure. Despite divorce being a common experience in the United States, it has long been associated with stigma. This stigma is reflected in the research in the 1960 and 1970s, in which divorced families were described as 'broken' and married families were described as 'ideal' or 'intact' families. Less stigma is associated with divorce in some European countries, like Scandinavian countries, which have liberal attitudes toward divorce and diverse family structures. In collectivist cultures like China, divorce is seen as violating traditional Confucian values that emphasize the centrality of family and interpersonal harmony. In these cultures, people tend to adhere to tradition and social convention, which often includes respecting parents' preferences for partner selection and religious rules regarding maintenance of marriage. In such societies, individuals are less likely than in other cultures to divorce even when the marriage is unsatisfying. Although divorce is still associated with intense stigma and shame in China, recently, the divorce rate has increased in regions that have experienced industrialization and modernization.

Demographics of Divorce

Different cultural attitudes about divorce are reflected in wide variation in divorce rates across the world. Crude divorce rates, which reflect the number of divorces per 1000 people, are higher in the United States than in other nations. In 2011, the crude divorce rate in the United States was 3.6, compared with rates ranging from 2.0 to 2.3 in Japan, France, Germany, and the United Kingdom. Crude divorce rates are typically lower in less developed countries. For example, rates are 0.8 in Mexico, 2.0 in China, and range from 1.4 to 1.9 in most reporting African nations.

Crude divorce rates are not an ideal method of assessing divorce rates because they do not account for the proportion of populations consisting of children and single adults. An alternative method is to consider the percentage of marriages that end in divorce. Recent estimates show that in some countries, the rate of marriages ending in divorce is almost as high if not higher than in the United States, where 53% of marriages end in divorce. For example, 55% of marriages in France, 47% in the United Kingdom, and 49% of marriages in Germany end in divorce. Rates are lower in Japan (36%), China (22%), and Mexico (15%). Although divorce rates vary among families of different demographic profiles, between 30–40% and 48% of American and British children, respectively, are expected to live in a divorced, single-parent home by age 16. Recent data from Australia revealed that only 25% of young adults reported having experienced parental divorce in childhood. Not only is divorce a common experience for children across nations, but often the divorce rate for remarriages is higher than for first marriages. Thus, a sizable number of children who experience parental divorce will experience more than one transition in family structure.

Risks Conferred by Parental Divorce

The majority of research examining children's responses to parental divorce has been conducted in the United States, and the current review summarizes this research as well as the less extensive research conducted in other countries. Commonly, researchers have compared outcomes at a single point in time of children who have experienced parental divorce with those of children who live with both of their biological parents. Some researchers have followed children at multiple points after divorce or assessed children's outcomes before and after divorce. A few researchers have assessed whether children's outcomes improve through participation in preventive interventions designed to change factors associated with children's postdivorce outcomes.

Improved research methodology in recent decades has allowed researchers to adequately control for variables that influence parental divorce and children's adjustment outcomes. For example, fixed-effect models account for time invariant sources of heterogeneity, genetically informed studies

account for the influence of shared genes, and randomized experimental trials of preventive interventions account for variables that are naturally correlated in nonexperimental studies (e.g., maternal personality, preexisting economic stress). Results from research using these methods have highlighted the great variability in children's responses and identified aspects of the circumstances faced by families before and after the actual marital dissolution that affect children's post-divorce adjustment outcomes.

For most children and adolescents, the separation period itself causes considerable stress, and during the first year or two after separation, many children experience distress, anxiety, anger, and disbelief about the divorce. However, most children 'bounce back' during the first couple of years after the divorce.

Although research in the United States shows that 70–75% of children who experience parental divorce do *not* have clinically significant mental health problems (e.g., Hetherington et al., 1998), there is compelling evidence demonstrating that divorce increases risk for mental health problems in childhood and adolescence. The most comprehensive picture of how children fare after parental divorce is provided by results of *meta*-analysis, a quantitative method of literature review in which measures of central tendency (e.g., the mean) and variation of outcomes are aggregated from multiple studies. *Meta*-analyses of 92 studies published mostly in the United States through the 1980s (Amato and Keith, 1991) and 67 studies published in the 1990s (Amato, 2001) showed that children in divorced families were more likely to exhibit externalizing problems, such as conduct disorders, delinquency, and impulsive behavior than children in two-parent families. Although the largest differences were found for externalizing problems, children from divorced families were also more prone to experience internalizing problems, such as depression and anxiety, than children from married families. These differences were consistent across children's age and gender and were slightly larger for more recent studies. It is important to note that the differences in adjustment outcomes of children from divorced versus married families were small to moderate in magnitude.

Similar results have been found cross-culturally. A large-scale study in Norway revealed that although most children who experienced parental divorce fared fine, a subgroup had more externalizing problems and difficulties with peers and, to a lesser extent, more internalizing problems than youths from two-parent families. Similarly, research in Turkey, where 20% of marriages end in divorce, has shown that girls whose parents divorced experienced higher depression levels than children in married families, and both boys and girls from divorced families exhibited more psychological problems than youths in married families. Furthermore, research in China has found that, compared with children with married parents, children who experienced parental divorce exhibited more behavioral problems and increased rates of anxiety and depression.

Research also shows that children whose parents divorce perform less well academically, score lower on achievement tests, and are more likely to drop out of school than children with married parents. For example, analyses of 10 U.S. national probability samples revealed school dropout rates of 31% among adolescents who experienced parental divorce compared with 13% in nondivorced families (McLanahan, 1999). Similarly, the Longitudinal Study of Australian Youth found that 22% of youths from single-parent and divorced families did not complete high school, compared with 13% from two-parent families (Curtis and McMillan, 2008). Furthermore, research in Scandinavia showed that youths in divorced families achieved lower grade point averages than their peers with married parents. And, research in China has found that parental divorce is associated with decreased school performance.

Children from divorced homes are more likely to engage in risky behaviors, such as substance use and early sexual activity. Adolescents from divorced families in the United States initiate alcohol use earlier and report higher alcohol, marijuana, tobacco, and illicit drug use than their peers from two-parent families. Similarly, researchers have found that Scandinavian youths who experienced parental divorce were more likely to smoke, consume alcohol, and use illicit drugs than their peers in two-parent families. Research also shows that adolescents whose parents are separated or divorced engaged in intercourse at younger ages than those whose parents are married. Compared with adolescents whose parents remained together or divorced later in their childhood, adolescents in the United States whose parents divorced when they were 5 years or below were at particularly high risk for engaging in sexual activity by age 16 (Donahue et al., 2010). Early parental separation is also associated with a greater number of sexual partners in adolescence. Furthermore, adolescents from divorced families in the United States experience teen birth rates of 33% compared with 11% among adolescents from nondivorced families (McLanahan, 1999). Similarly, research shows that adolescent females in divorced families in New Zealand are more likely to experience teenage pregnancy than their peers in married families.

Parental divorce might either initiate new difficulties or exacerbate existing ones. Some studies that have assessed children prior to parental divorce have shown that problem behaviors are evident well before marital dissolution. In addition, some studies show that differences between children who have and have not experienced parental divorce are diminished after controlling for predivorce functioning. It is important to note, however, that several other studies show a significant increase in problems among offspring of divorced parents after divorce after controlling for predivorce factors.

As noted earlier, although *most* children do not experience significant lasting problems following parental divorce, negative effects of parental divorce extend into adulthood for a substantial minority. Several prospective studies with epidemiologic samples have revealed an association between parental divorce and significant increases in mental health problems, substance abuse, mental health service use, and psychiatric hospitalization in adulthood. For example, Chase-Lansdale et al. (1995) reported that the odds of exhibiting clinically significant mental health problems at age 23 increased by 39% as a function of parental divorce after controlling for relevant predivorce variables like emotional problems, academic achievement, and socioeconomic status. More recently, the World Mental Health Survey revealed that after controlling for age, country, and type of disorder, parental divorce was associated with elevated rates of onset of mental

health disorders in adulthood in high-, middle-, and low-income countries (odds ratio 1.1–1.3) (Kessler et al., 2010).

Research also shows that the effects of parental divorce in adulthood extend beyond mental health problems. Many studies provide evidence that parental divorce in childhood is related to less success in key developmental tasks of young adulthood: education, work, and romantic relationships. Adults who experienced divorce in childhood have lower educational and occupational attainment and more economic and employment problems. Also, adults who experienced parental divorce in childhood experience poorer marital quality and are more likely to divorce, a phenomenon referred to as the 'intergenerational transmission of divorce.' Furthermore, parental divorce in childhood is associated with increased physical health problems in adulthood. Illustratively, a 70-year prospective study found that experiencing parental divorce before age 21 was related to a 44% increase in mortality risk, controlling for personality in childhood and individuals' own divorces (Schwartz et al., 1995).

Summary

Overall, research indicates that a substantial minority of children who experience parental divorce are at increased risk for a variety of difficulties in multiple domains of functioning. Specifically, in childhood and adolescence, youths who experience parental divorce are at heightened risk for internalizing problems, externalizing problems, reduced academic achievement, school dropout, substance use, and early sexual behavior. Adults who experienced parental divorce in childhood are at increased risk for experiencing mental health disorders, physical health problems, reduced educational and occupational achievement, poorer marital quality, and divorce.

Societal Costs of Parental Divorce

The divorce process and its immediate ramifications often place economic demands not only on the divorcing individuals, but also on society. Divorce-related policies vary tremendously among countries, with some nations, including many European nations, providing aid to single mothers that is intended to offset the negative economic impact of divorce. For instance, some states fund generous public childcare programs, which enable single mothers to more easily reenter the workforce. Nations that provide high levels of social welfare, like Belgium, Denmark, and the United Kingdom, indirectly pay high costs for parental divorce by funding such systems.

The long-term effects of parental divorce on youths' adjustment outcomes also have negative economic effects. Although the typical magnitude of differences in adjustment outcomes between individuals who experienced parental divorce in childhood and those who did not is not large, the short- and long-term outcomes associated with parental divorce, such as internalizing problems, externalizing problems, and substance use problems, have high societal costs. For instance, children's externalizing problems like antisocial behavior, aggression, and delinquency are related to unemployment and criminality in adulthood, which cost the United States between $335 and $350 billion annually (Miller, 2004). Also, internalizing problems, such as depression, cost the United States over $83 billion in medical services, lost lifetime earnings due to suicide, and lost workplace productivity (Greenberg et al., 2003).

The concept of population attributable fraction (proportion of an outcome due to a risk factor or percentage of cases that could be prevented by removing the factor or its consequences (PAF)) provides an important perspective on the public health significance of parental divorce. Results from a nationally representative survey of adults in the United States revealed that the PAF of parental divorce for drug dependence is 23%, suggesting that almost one-quarter of cases of drug dependence could be eliminated by reducing the negative effects of parental divorce.

Given that the effects of parental divorce are similar cross-culturally, we would expect the economic costs of divorce to be high worldwide, particularly among nations with high divorce rates. For instance, evidence from the Netherlands, where 43% of marriages end in divorce, showed that parental separation or divorce is one of 10 major adverse childhood experiences associated with a variety of emotional problems, health risks, disease burden, sexual health issues, disability, and healthcare costs that have significant public health burden.

Theories of Children's Responses to Parental Divorce

There are two main theoretical perspectives for understanding the risks conferred by parental divorce. The genetic model focuses on the genetic transmission of personality traits. The transactional model suggests that the constellation of stressful events associated with separation and divorce adversely affects children's adjustment outcomes.

The passive genetic model posits that personality traits of parents contribute to both divorce and children's adjustment problems. According to this view, some parents have genetic predispositions for traits like aggression and antisocial behavior that increase the risk of family conflict and divorce. Furthermore, offspring of these parents inherit traits that increase the probability of developing adjustment problems, such as externalizing problems. Results of twin and adoption studies have shown that although some of the outcomes associated with parental divorce, such as internalizing problems, might be due to genetic factors, the majority of children's postdivorce adjustment problems, including externalizing problems and substance use, cannot be explained by passive genetic inheritance.

In the transactional model of divorce, changes in family relationships and environmental circumstances before and after separation and divorce, rather than divorce itself, are viewed as critical determinants of children's postdivorce adjustment problems. Factors that are present prior to divorce include characteristics of parents themselves, their parenting behaviors, and economic conditions. Research indicates that parents who divorce are more likely to have a history of alcoholism, depression, antisocial behavior, and economic problems. Couples who divorce also have poorer conflict resolution

skills, higher levels of negativity, and poorer parenting skills prior to the divorce than married couples.

Stressful events that occur during separation and divorce and the availability of protective factors are seen as influencing the extent to which children are adversely affected by parental divorce. The stressful changes that occur for many children include increases in interparental conflict and parental depression, reductions in contact with the nonresidential parent and relatives, changes in residence, and reductions of material resources. These changes influence adjustment problems to the extent that they are mitigated by interpersonal protective factors, such as high-quality parent–child relationships, and intrapersonal protective factors, such as effective coping skills. For example, high-quality parent–child relationships might moderate, or reduce the strength of, the relation between interparental conflict and adjustment problems.

From this perspective, divorce can lead to an unfolding of failures to resolve developmental tasks and can increase susceptibility to mental health problems and impaired competencies that can persist over developmental stages. Parenting is viewed as playing a central role in facilitating children's successful adaptation. Furthermore, the skills and resources that are developed in successful resolution of developmental tasks, such as effective coping and academic success, are viewed as important tools when children face challenges in subsequent developmental periods. Dynamic interactions between children's resources and their developing competencies can lead to a positive or negative cascade of adaptive functioning in multiple domains over time. For example, high-quality parent–child relationships might lead to lower levels of externalizing problems, which might increase the likelihood of positive interactions with mothers, fathers, teachers, and peers, which could lead to further reductions in externalizing problems and improvements in other aspects of functioning, such as social competence. Alternatively, low-quality parent–child relationships might lead to higher levels of externalizing problems, which might increase the likelihood of negative interactions with mothers, fathers, teachers, and peers, which could lead to further increases in externalizing problems and impairments in social competence.

Summary

Although some of the difficulties exhibited by children after divorce are likely due in part to a genetic predisposition for developing these problems, environmental, intrapersonal, and interpersonal factors also likely influence children's post-divorce adjustment outcomes. In the next section, we discuss factors that have been shown consistently to be related to children's postdivorce adjustment outcomes.

Understanding Variability in Children's Postdivorce Adjustment Outcomes

In the context of the significant public health burden associated with parental divorce, variability in children's outcomes has prompted researchers to examine factors affecting children's postdivorce outcomes. This research has guided the development of interventions by identifying factors to target for change.

Stressful Events

As discussed above, parental divorce involves a multitude of potential stressful changes and disruptions in children's social and physical environments that occur before, during, and after parental separation. For example, interparental conflict often begins many months or years prior to separation. Conflict can escalate during the processes of separation and divorce as parents negotiate new parenting-related issues, such as residential and visitation arrangements and child support.

Children, parents, and clinicians who work with children from divorced families agree that the most stressful divorce-related experiences are those that occur in the nuclear family. These include verbal and/or physical interparental conflict, derogation of one parent by another, and being blamed for the divorce. Other experiences that children perceive as highly stressful include hearing criticism of their parents by relatives or neighbors, having to give up material possessions, and being exposed to maternal psychological distress.

Most children experience changes in their relationships with their parents during the divorce process. Although there has been an increase in joint residential arrangements over the past couple of decades, about 80% of children in divorced families in the United States live primarily with their mothers. Similarly, research on separated and divorced parents shows that 84% of mothers in Australia identified themselves as the primary residential parent, compared with 17% of fathers. Thus, parental divorce often means that children lose daily contact with fathers. Australian data show that half of children who live apart from one of their parents, typically fathers, see their nonresidential parent at least every other week. Statistics from the United Kingdom show that at least half of children whose parents separated or divorced had weekly contact with their nonresidential parent, although fewer than 30% had weekly overnight stays with their nonresidential parent. Research shows that for many children, contact with nonresidential fathers decreases substantially over time. For most children, the quality of the father–child relationship also changes following separation as fathers adapt to new parenting responsibilities and both children and father adjust to new living arrangements. Specifically, researchers have found that many children feel less close to their fathers and children after the divorce and fathers have difficulty providing consistent discipline and setting limits for their children.

Separation and divorce also affect children's relationships with their mothers. Primary residential mothers report high levels of stress as they negotiate the tasks of adjusting to the dissolution of marriage and managing additional household and childcare responsibilities. Juggling these demands increases mothers' risk for psychological distress, which has been linked to disruptions in parenting. After separation and divorce, mothers are often less supportive and affectionate and less consistent and effective in their discipline strategies.

Many newly divorced parents experience a decrease in family income. Researchers in Australia have found that even with increases in government income support and increased job opportunities for women in recent decades, women and

children in particular continue to be significantly negatively impacted by postdivorce financial hardships. Reductions in standard of living after divorce often lead to stressful changes for children, including having fewer material resources, moving to smaller homes, changing neighborhoods, and changing schools. These transitions can involve the loss of relationships with friends, teachers, and other important members of children's social support networks. In some nations, like Scandinavian welfare states, families are less financially impacted by divorce. However, research suggests that children in these societies do not fare better following divorce than children from societies that provide little or no financial support to families postdivorce. This finding suggests that financial stressors are only one factor influencing children's postdivorce outcomes.

For many children, parental divorce is followed by additional family transitions. Most adults remarry following divorce, and the median time between divorce and remarriage is 4–5 years in the United States (U.S. Census Bureau Marriage and Divorce, http://www.census.gov/). Half of these newly formed households include children under age 18. Furthermore, many children become part of new families through cohabitation rather than remarriage. Although remarriage or cohabitation can increase the financial and emotional support provided to the residential parent, remarriage or cohabitation alters the nature of children's relationships with their biological parents and often involves a complex family reconfiguration. For children, this includes the addition of a stepparent, as well as other step–kin relations with stepsiblings, grandparents, and other extended family and, in some cases, half-siblings.

The frequency of divorce-related changes relates positively to internalizing and externalizing problems and negatively to self-concept, social competence, and academic competence. Children exposed to more divorce-related stressors also show maladaptive cognitions, including a reduced sense of their ability to affect their environments, than those exposed to fewer divorce-related stressors.

Interparental Conflict

Although a large body of research has shown that interparental conflict is one of the most damaging aspects of divorce, research has shown that children in families characterized by intense, chronic marital conflict and/or violence are better off in the long run if their parents divorce. Although many children whose parents divorce have been exposed to high levels of interparental conflict prior to divorce, in approximately 25% of divorcing families, parents exhibited minimal conflict before separating.

Conflict is a complex construct, and interparental conflict's effect on children's adjustment outcomes depends on several factors, such as intensity and whether or not the child witnesses the conflict. Intensity of conflict is positively related to infants' and toddlers' anxiety and insecure attachment and to young children's distress and insecurity. In older children and adolescents, intensity of conflict is positively related to increases in internalizing problems and externalizing problems. Overt hostility between parents, which can involve slapping, screaming, and derisive behavior, is more strongly associated with externalizing problems and internalizing problems than passive–aggressive hostile behavior between parents. However, passive–aggressive conflict and subtle behaviors that put children in the middle can also be problematic and are linked to children's internalizing symptoms like depression, anxiety, and withdrawal.

Parenting

Most of the research on relations between parenting and children's postdivorce adjustment outcomes has focused on mother–child relationships. Father–child relationships after divorce are understudied, and most researchers who have considered fathers have studied nonresidential fathers.

There is strong, consistent correlational support for the protective effect of high-quality mother–child relationships. Mother–child relationships characterized by warmth, supportiveness, effective problem-solving skills, positive communication, and low levels of conflict and negativity are consistently associated with children's lower mental health problems and positive social adaptation outcomes, including higher self-esteem and better academic performance following divorce.

The results of experimental trials of preventive interventions have also shown that intervention-induced increases in mother–child relationship quality account for reductions in children's mental health problems. Because randomization to an experimental versus control condition makes it unlikely that third variables (e.g., shared genes, economic hardship) account for the relations between program-induced changes in relationship quality and changes in children's mental health problems, inferences about the causal nature of the relations between mother–child relationship quality and adjustment outcomes are strengthened over what can be inferred from correlational studies.

Correlational research has provided consistent support for the association between fewer postdivorce adjustment problems and use of effective discipline, which entails parents consistently enforcing age-appropriate rules, expectations, and rule compliance. Experimental trials of preventive interventions for divorced families have shown that program-induced improvements in mothers' use of effective discipline accounted for program effects on externalizing problems, delinquency, and academic performance.

Parental monitoring is an aspect of parenting that becomes especially relevant in adolescence given the normative increases in independence and time spent with peers. Correlational research has shown that adolescents whose mothers engage in higher levels of monitoring have lower levels of externalizing problems, substance use, and academic difficulties than those whose mothers engage in lower levels of monitoring. Experimental research with divorced families has shown that program-induced changes in parental monitoring accounted for program effects on adolescent substance use 6 years after participation for youths who had high levels of risk for developing adjustment problems at program entry.

Most of the research on the father's role in children's postdivorce adjustment has focused on whether the amount of father–child contact affects children's adjustment after divorce. Findings from this line of research have been inconclusive. Over the past couple of decades, researchers have used a more

fine-grained approach to identify aspects of postdivorce father–child relationships that matter for children's outcomes. Results of *meta*-analyses have shown that feelings of closeness; authoritative parenting, which entails a balance of encouraging children's independence and enforcing limits; and maintenance of appropriate expectations were significantly related to greater academic success and lower levels of externalizing problems and internalizing problems after divorce. Also, results of a *meta*-analysis that examined children's adjustment outcomes in sole- and joint-custody arrangements found that children in joint custody were better adjusted than those in sole custody and no less well adjusted than children living with married parents, supporting the importance of both parents remaining involved in parenting following divorce.

Most researchers have examined the relations between children's relationships with *either* mothers or fathers and children's postdivorce adjustment without considering their interplay. The few studies to examine the joint effects of parenting of mothers and fathers have shown that high-quality, responsive parenting by fathers *and* mothers was significantly related to lower mental health problems after accounting for the quality of the relationship with the other parent. Researchers who have compared adjustment outcomes of children who had a positive relationship with one parent, both parents, or neither parent have found that having a positive relationship with either the mother or father was associated with fewer mental health problems compared with not having a close relationship with either parent.

Child Factors

Researchers have examined whether age and gender predict postdivorce adjustment outcomes. Some studies, particularly older studies or those examining conduct problems, show greater adjustment problems for boys than girls. However, *meta*-analyses show that divorce was associated with adjustment problems for both boys and girls. Although *meta*-analyses show that children in middle childhood and adolescence appear to be more negatively impacted by divorce than preschoolers or college students, far fewer studies have examined postdivorce adjustment in these latter two age groups, making it difficult to draw firm conclusions regarding the relation between age at the time of divorce and children's adaptation.

There is a growing literature that shows that how children interpret and cope with divorce-related stressors affects their adjustment. Children who interpret divorce-related stressors in a characteristically negative style or as threatening to their well being have more postdivorce adjustment problems than children who interpret these stressors in a less negative manner. Coping efficacy, a global belief that one can deal both with the demands made and emotions aroused by a situation, has been shown to relate to lower postdivorce internalizing problems. Also, children who use active coping strategies, such as problem solving or positive cognitive restructuring, experience fewer mental health problems than those who employ avoidant coping strategies, such as disengaging from potentially stressful situations. The adaptive value of specific coping strategies depends on context. For example, children might benefit from using cognitive restructuring in situations that are largely outside of their control, whereas problem solving might be adaptive for situations that are more controllable.

There is also evidence that children's beliefs about social relatedness, self-worth, and control predict postdivorce adjustment problems. For example, concerns about social relatedness, defined as fear of abandonment, predict higher levels of mental health problems. Also, higher levels of self-worth relate to lower levels of mental health problems, and higher levels of internal locus of control predict lower levels of mental health problems.

Summary

Researchers have identified several potentially modifiable environmental-, family-, and child-level factors that are correlated with children's postdivorce adjustment outcomes. Although most research has examined direct relations between individual factors and children's outcomes, a few researchers have examined additive and interactive effects of some of these factors. For example, several researchers have shown that high-quality mother–child relationships reduce the negative impact of divorce-related stressors. At high levels of divorce stressors, children with higher quality relationships have fewer internalizing problems and externalizing problems than children with lower quality relationships.

Interventions to Reduce or Prevent Children's Postdivorce Adjustment Problems

Despite the high prevalence of divorce and large body of research demonstrating that parental divorce is associated with negative outcomes for many children, few rigorous trials of programs for divorced families have been conducted. Most of the programs that have been evaluated involved children only, some focused on parents only, and others included separate components for children and mothers. While most interventions have been developed and implemented in the United States, some programs have been implemented in several nations, such as the Children of Divorce Intervention Program (Pedro-Carroll and Cowen, 1985), which has been widely utilized in the United States, Canada, New Zealand, and Australia. The programs that have been found to have positive effects on children's adjustment outcomes have targeted factors that have been shown to be significantly related to children's postdivorce adjustment.

The most widely available preventive interventions are relatively brief (6–12 sessions), child-focused programs that are typically provided in school settings. Such interventions provide emotional support and teach adaptive coping and control beliefs. Several programs have been evaluated in experimental or quasiexperimental trials and have been shown to lead to improvements in multiple aspects of postdivorce adjustment outcomes, including internalizing problems, externalizing problems, self-concept, and competence. Program effects have been maintained 1–2 years after program completion.

Preventive programs that target primary residential parents have focused on teaching effective parenting skills to enhance the quality of parent–child relationships, increase the use of

effective discipline, and help parents reduce children's exposure to divorce-related stressors, such as interparental conflict. Experimental trials have shown such programs to have positive short-term effects on mental health outcomes and positive long-term effects on mental disorder, substance use disorder, frequency of alcohol, marijuana, and other drug use, delinquency, arrests, risky sexual behavior, academic performance, and self-esteem. Some research has shown positive program effects into young adulthood, and both short- and long-term improvements have been shown to be accounted for by intervention-induced improvements in mother–child relationship quality, effective discipline, and parental monitoring.

Programs that included components for parents and children were based on the premise that targeting empirically supported factors that are primarily controlled by children (e.g., coping) as well as those controlled by parents (e.g., effective discipline) would lead to additive effects. Surprisingly, research on dual-component programs involving mothers and children has *not* shown that more program components are better.

Summary

Several trials have shown that programs that targeted empirically supported family- or child-level factors improved multiple aspects of children's short- and long-term postdivorce adjustment outcomes. These findings have important public health implications given that their widespread implementation could significantly reduce the public health burden of parental divorce.

Conclusion

A sizable number of children, particularly in the United States and other developed countries, experience parental separation or divorce. Although most youths whose parents divorce do not have significant problems, a substantial minority are at increased risk for a variety of both short- and long-term problems. These problems include mental health, physical health, and substance use problems, lower educational and occupational achievement, and difficulties in interpersonal relationships. Research has identified potentially modifiable factors that are significantly correlated with postdivorce adjustment, such as parent–child relationship quality and coping efforts. Several interventions that target these factors have been evaluated in rigorous experimental trials and have been shown to improve children's postdivorce adjustment outcomes.

See also: Behavioral Problems, Effects of Parenting and Family Structure on; Family Theory: Economics of Marriage and Divorce; Repartnering and Stepchildren.

Bibliography

Amato, P.R., 2001. Children of divorce in the 1990's: an update of the Amato and Keith (1991) meta-analysis. Journal of Family Psychology 15, 355–370.

Amato, P.R., Keith, B., 1991. Parental divorce and the well-being of children: a meta -analysis. Psychological Bulletin 110 (1), 26.

Chase-Lansdale, P.L., Cherlin, A.J., Kiernan, K.E., 1995. The long-term effects of parental divorce on the mental health of young adults: a developmental perspective. Child Development 66 (6), 1614–1634.

Curtis, D.D., McMillan, J., 2008. School Non-completers: Profiles and Initial Destinations. Longitudinal Surveys of Australian Youth Research Report 74.

Donahue, K.L., D'Onofrio, B.M., Bates, J.E., Lansford, J.E., Dodge, K.A., Pettit, G.S., 2010. Early exposure to parents' relationship instability: implications for sexual behavior and depression in adolescence. Journal of Adolescent Health 47 (6), 547–554.

Greenberg, P.E., Kessler, R.C., Birnbaum, H.G., Leong, S.A., Lowe, S.W., Berglund, P.A., Corey-Lisle, P.K., 2003. The economic burden of depression in the United States: how did it change between 1990 and 2000? Journal of Clinical Psychiatry 64 (12), 1465–1475.

Hetherington, E.M., Bridges, M., Insabella, G.M., 1998. What matters? what does not? Five perspectives on the association between marital transitions and children's adjustment. American Psychologist 53 (2), 167.

Kessler, R.C., McLaughlin, K.A., Green, J.G., Gruber, M.J., Sampson, N.A., Zaslavsky, A.M., Williams, D.R., 2010. Childhood adversities and adult psychopathology in the WHO world mental health surveys. The British Journal of Psychiatry 197 (5), 378–385.

McLanahan, S., 1999. Father Absence and the Welfare of Children. Coping with Divorce, Single Parenting, and Remarriage, pp. 117–145.

Miller, T.R., 2004. The social costs of adolescent problem behavior. In: Biglan, A., Brennan, P.A., Foster, S.L., Holder, H.D. (Eds.), Helping Adolescents at Risk: Prevention of Multiple Problem Behaviors. Guilford Press, New York, pp. 31–56.

Pedro-Carroll, J.L., Cowen, E.L., 1985. The children of divorce intervention program: an investigation of the efficacy of a school-based prevention program. Journal of Consulting and Clinical Psychology 53 (5), 603.

Schwartz, J.E., Friedman, H.S., Tucker, J.S., Tomlinson-Keasey, C., Wingard, D.L., Criqui, M.H., 1995. Sociodemographic and psychosocial factors in childhood as predictors of adult mortality. American Journal of Public Health 85 (9), 1237–1245.

Relevant Websites

http://www.cdc.gov/nchs/ – Centers for Disease Control and Prevention.
http://www.census.gov – United States Census Bureau.

Maritime Networks, History of

Gordon Boyce, University of Newcastle, Callaghan, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the dynamics and forms of networks used in world shipping from the sixteenth century to the present day. It begins with a theoretical exposition of network structures, ethical standards, and internal functions in order to develop a basis for analyzing how these cooperative constructs were employed by British, European, American, and Japanese shipping enterprises. The article also explores the impact of technical and legal changes, especially the recent introduction of containerization, new shipbuilding methods, logistics, and computerization.

A network is an informal alliance of decision-makers (people or firms) that seek to further their mutual interests through cooperation. Networks are arguably the oldest of all frameworks used for organizing complex business activities. While other structures, such as partnerships and companies, are also of long lineage and widely employed, network-based constructs have persisted. They have been especially prevalent in certain industries, including shipping, merchant trading, finance, insurance, and more recently in film-making, raw material exploitation, engineering, and construction. In some sectors, networks proved to be of temporary relevance, especially in early developmental stages, and were superseded by other forms, usually the company. However, the network should not be considered as a transient business construct. Historically, it has proved to be a durable and distinctive structure (Powell, 1990). Moreover, in specific industries and historical contexts, it has been ubiquitous. Entrepreneurs from many countries have employed it, and indeed the network has been very evident in transnational business.

Broadly considered, there are three generic structures that can be used to effect economic exchanges. G.B. Richardson (1972) suggested that these constructs can be arranged along a spectrum, with the first, markets, and the second, hierarchy, positioned at opposite ends. In between these polar extremes, there resides an array of cooperative frameworks that generate varying levels of interdependence; they include long-term contracts, franchising agreements, joint ventures, strategic alliances, and the subject of this article, networks. Each of these three generic exchange-facilitating institutional frameworks possesses distinctive coordinating properties. At one end of Richardson's spectrum, market transactions are often depicted as supportive of one-off transactions between anonymous parties all of whom have complete information regarding considerations relevant to the exchange in question. Market forces of supply and demand generated by many such participants acting in a self-interested way to maximize their own private benefit determine the level of prices, which are the key bits of information upon which dealers act. At the other extreme, hierarchies (corporate or governmental) coordinate internal exchanges by means of commands issued by a legitimate authority residing at the apex of the structure. Information gathered and processed internally through various levels of the organization is the basis of decision-making. In contrast to both markets and hierarchies, exchanges within cooperative frameworks are shaped by the mutual interdependence of participants. In the short-term, these actors relinquish a measure of their freedom to follow the self-interest that actuates market dealers, in order to gain greater benefits through long-term cooperation. Within these frameworks there is no absolute authority that directs other members, as is the case in hierarchies, although there are often differentials in the power or influence among cooperators. For decision-making purposes, cooperative constructs rely on a mixture of interpersonal/organizational knowledge and preferential intelligence available only to members (see below). Networks share these basic characteristics of cooperative frameworks, but they exhibit their own distinctive properties.

Network interaction is based upon reciprocity; one member does another a favor on the expectation that it will be eventually returned. There is no preconceived time-frame for the payback, and the favor returned may not necessarily be precisely of equal benefit to the original one. Thus, reciprocity may or may not be symmetric. Over the long-term, however, the 'gives' and 'takes' are sufficiently equivalent to sustain enduring association. Underpinning reciprocity is trust, which is not reflective of a naive faith in others but rather what Mark Casson (1997) described as "a warranted belief that someone will honour their obligations, not merely because of material incentives, but out of moral commitment". Thus, network members possess sufficient interpersonal knowledge to judge the likely pattern of each other's future behavior. To these considerations we can add participants' recognition that an economic actor can be trusted because 'reputational deterrence' exists. The idea here is that trust in large measure arises as someone acquires a reputation for behaving in a consistently honorable way over a long period of time. Because renown is subject to rapid dissipation when one, who has hitherto been trustworthy, unexpectedly behaves opportunistically, with the consequence that future social and economic transacting possibilities may be severely curtailed, and may be difficult to recover, persons of high standing will be deterred from committing overt and guilefully self-interested acts. 'End game' strategies apart, participants infer that the greater the social, economic, and public capital an individual has accumulated, the more powerful the deterrent effect. Moreover, the reputations of members tend to become shared as a result of long-term affiliation and consistent behavior. For this reason, membership is exclusive. Participants need to safeguard their personal and collective interests from opportunists.

Alvesson and Lindkvist (1993) observed three broad types of network as determined by the nature of exclusive affiliation. First, 'economic cooperative' networks coordinate business activities alone and rely primarily upon mutual pecuniary gains to bind their members. Second, 'socially integrative' forms arise from some ethnic, religious, or cultural commonality among participants. Third, and perhaps the most basic, is the 'blood kinship' network which is founded on an extended family. Yet, in practice, networks often rely on combinations of family ties, common social/cultural affiliation, and shared business interests.

The significance of these bonds lies in their capacity to enhance particular forms of information exchange. As informal communicating systems, networks reduce uncertainty associated with transacting by conveying interpersonal knowledge that generates higher levels of trust than those prevailing in markets or hierarchies. In turn, trust sustains forthright communication that reduces transacting costs which arise from the necessity of building safeguards against being cheated. Thus, network members can quickly negotiate contracts, they can rely upon network connections and affiliation to confirm that the terms of the agreement are fulfilled, and they can depend upon noneconomic ties to provide a network 'culture' which by mobilizing feelings of inclusion deters opportunistic behavior.

With these strong foundations, network associates can share risks and assemble complex, growth-supporting sets of resources that one entrepreneur acting on their own might not be able to mobilize. The resource collection usually consists of capital and physical assets, but of greater significance it also includes preferential information and accumulated knowledge. Exchanges involving the latter two intangibles are notoriously problematic because it is difficult to determine a fair price without prior disclosure, and once information/knowledge has been revealed it can be used without paying compensation. Based on trust and expectations of reciprocity, networks have advantages in surmounting these problems and lowering the cost of making what otherwise might be prohibitively expensive transactions.

As trust-based communicating systems, networks preferentially circulate information that is 'thick' and 'rich', that is intelligence that has been preprocessed to some degree, is grounded in a context understood by each member, and enhanced by discussion. This type of valuable information contrasts starkly with the 'raw' publically available data that swirls around markets and is freely accessible by all economic actors. Unlike hierarchies where security concerns may limit the breadth of transmissions or cause content to be filtered, networks sustain 'free flowing' information that is complete and unfiltered. These reasons explain why networks are sometimes considered to be more flexible and dynamic than hierarchies. However, certain members of these cooperative constructs may occupy 'nodal' points within the web of communication channels and thereby have information-related advantages relative to other participants. Such figures derive power and what theorists call 'centrality' from their positions, and they are thus well placed to perform intermediary roles in distributing intelligence and arranging deals.

The nature of specific network channels also shape the quality of transmissions. A distinction is often drawn between 'weak' and 'strong' ties (Granovetter, 1973). The latter are based on common affiliation: members so joined observe shared routines that guide their interaction and engage in highly contextualized exchanges based on intimately shared experience that sustains a collective memory. Participants linked by strong ties know what each other know and can fairly accurately anticipate how one another will react to new stimuli. Thus, strong ties support rapid communicating and transacting as well as deep learning. They also muster a degree of anticipatory capability. In contrast, members connected by weak ties are not closely bound cognitively or in terms of shared experience. A weak tie arises when one member has an alternative set of information or a different way of processing intelligence perhaps because they come from a more distant social/culture background or have been educated in a particular way. As a result, these figures may furnish new data or singular analytical capacity, either of which may stimulate the recognition of new business opportunities. Network longevity is determined by many factors, but replenishing the constituent knowledge/information base by forging new weak ties is often important in generating enduring dynamism. Conversely, the collective memory enjoyed by participants bound by strong ties, being founded on common cognitive and behavioral conditioning, may create rigidity when completely new conditions arise within the external business environment. 'In grown' networks may fail to modify their processes or prove incapable of securing the new skills needed to adapt. All of these factors help to explain the rise and decline of specific maritime networks.

British Shipping

The discussion begins with British maritime networks, not because this cooperative structure originated in that country – indeed it arose in ancient times and was used by nearly every nation – but rather because Britain was the world's foremost shipping and trading nation for such an extended period and because the relevant literature is so well-developed. From the sixteenth and seventeenth centuries, British private overseas merchant trading businesses relied on blood kinship networks to support their activities and over time developed larger social integrative forms to secure resources from wider ranging religious, social/cultural, and geographically determined links. Moreover, despite their 'corporate' vestment, the great chartered companies – British, Dutch, and French – which enjoyed national monopolies in the geographically circumscribed trades, often employed networks of the economic cooperative form to harness the required resources and to control far flung operations. Before occupational specialization unfolded on a wide basis, freelance merchants and those associated with chartered entities traded and shipped goods. Both types of business worked across considerable distance and, without high-speed communication technology, firms in these sectors were compelled to delegate many functions to persons who had first hand information about conditions at points far from headquarters. As a result, personal knowledge, established bonds of trust, and participative incentives, all mobilized through network structures, were used to bind the interests of principal owners resident in the metropolis and ship masters, traders, and agents working abroad. In terms of resource

mobilization, even though a rich merchant could afford to buy a vessel independently, he attracted funds from allies in order to spread risk. (A reliable insurance market did not appear until 1771 when the principal supporters of Lloyds reformed its operation.) Uncertainty was pervasive: apart from the challenges of maintaining control, shipwreck, war, piracy, and unpredictable markets posed serious external threats. As communication systems, these early networks supplied information needed to reduce risks arising from internal sources and to assess the likely consequences of dangers posed from outside the venture. Finally, the network provided a mechanism for distributing residual and uncontrollable risk across its membership.

The traditional form of vessel ownership – the precise origins are not known – in which the capital invested in a ship was divided into 64 shares, allowed for significant risk spreading (Davis, 1962). The so-called 64-share system constituted a partnership in which investors, led by a managing owner (see below), were legally considered as tenants in common, and could sell or mortgage their shares without consulting other shareholders. Moreover, should an investor die, their shares passed to their estate and did not revert back to the partnership which continued to operate. (In contrast, land-based partnerships were dissolved if one member passed away, and individual partners could not mortgage or dispose of their shares without the agreement of all other owners.) Shipping came under the jurisdiction of Admiralty Law which deemed the vessel, not individual partners, as the responsible party, so that maritime investors enjoyed a limitation on their financial exposure. (The liability of landward partners was unlimited; each was responsible for debts incurred by the partnership and even other partners, to the full extent of their personal assets: 'to their last shilling and last acre'.) Individually the '64' owners (64thers) were only accountable for their share of any trading loss the vessel might sustain – profits were distributed in 64 proportions – and for their portion of any unpaid instalments due to shipbuilders. Thus, those who organized 64thers still had to attract solid and respectable investors, even though they faced less onerous internal risks than partners in land-bound business.

After the passage of the Bubble Act (1719), which severely restricted the formation of joint-stock companies, investment in shipping flowed through a legal loophole that made passive participation a relatively safe proposition. This interstice arose from recognition of the comparatively large capital requirements of ocean transport, the attendant risks, and its strategic importance. As a result, managing owners who could extend their networks and acquired sufficient renown had significant advantages over entrepreneurs in other sectors in attracting large amounts of capital from passive investors who came from a wide range of social, occupational, and geographic connections.

Within these larger constructs, ownership and management became somewhat separate, in part because occupational specialization grew from the late eighteenth into the early nineteenth centuries. Although the pace and extent varied, the increasing scale of markets supported larger enterprises, accentuated the generation of esoteric knowledge as well as skills which generated economies of scale in their application, and encouraged entrepreneurs to focus on a specific business activity. Gradually, general merchant shipowning traders concentrated on vessel management, trading operations, or finance. They often maintained portfolios of diverse interests and continued to reap economies of scope, but increasingly they directed their active endeavor toward a single business.

In shipping, managing owners, or managing agents, as they were often called, specialized in acquiring, outfitting, maintaining, and deploying vessels in trades in which they had some knowledge advantage. (Ship masters who accumulated sufficient funds also took up this occupation.) Increasingly, these figures managed ships that operated in a growing market for oceanic transport services. Their main interest was in carrying goods owned by others, although they sometimes arranged consignments of their own to load up their vessels.

The size and diversity of networks varied from port to port and changed across time, but most of them were configured according to the following representation. The 64er-based networks, as well as those that supported private joint-stock companies, chartered entities, and other ownership structures that emerged in the nineteenth century (see below), may be conceived of as a target consisting of three concentric circles which represent knowledge of diminishing quality as the distance from the bullseye increases. At the center, resided the managing agent who possessed superior information relative to other network investors, and it was this advantage that qualified him to occupy this key nodal position. His knowledge, track record, and reputation enabled him to contract with other investors who lacked first-hand knowledge of the venture's prospects and timely intelligence regarding its performance week to week or month by month.

The next circle, called 'the active core', included figures from some maritime-related occupation which provided them with the means of directly supporting the shipowning enterprise. Moreover, their decision to invest was informed by the specialized knowledge they accumulated in their main business and sustained by operating intelligence they gleaned through working directly with the managing agent. Thus, following the principle of reciprocity so important to network dynamics, sail makers, rope manufacturers, chandlers, provisioners, insurance brokers, and merchants committed funds on the expectation that their principal business would receive support from the managing agent's shipping venture, and in turn this working relationship provided them with some impression of its performance in addition to the data included in voyage reports issued by the managing agent. (For example, an insurance agent obtained some idea of the quality of the master and the extent to which a vessel was properly maintained from the quotes he received for coverage and by subsequently monitoring the frequency and nature of claims.) The geographic proximity of the managing agent might also give active members impressions of his reputation, his personal and business circumstances, and any wider associations he might have. A ship master bought or earned 64ths that served respectively as a bond or an incentive, and his reputation for safe navigation and prudent dealings at distant ports conveyed important impressions to other investors. (Indeed, so significant was the master as a factor determining risk, that Lloyd's compiled a register that recorded their performance to inform brokers' rate quoting decisions.) If the venture involved a new

vessel, its builder usually became the most important active constituent in the supporting networks. He had preferential information relative to all other members, including the managing agent, regarding the quality of the vessel's construction and thus his investment in 64ths represented a form of bond or guarantee. (So close and enduring did shipowner–shipbuilder links become in the steamship age that shared cognitive frames developed to facilitate their ongoing transactions (Boyce, 2003).) Often, builders took up sizable holdings which signaled their confidence to all other network constituents, and these men were well-known and highly respected in their port communities. Securing the support of a prominent shipbuilder was often the first step in the managing agent's network building endeavor for it brought reputation-by-association, which was subsequently enhanced as other active participants joined in.

When an 'active core' possessed substantial collective stature, it was relatively easy to attract a 'passive fringe' of rentier-type investors. These members possessed comparatively low quality information to guide their fund allocation activities, since they usually had no connection with a maritime business through which they could obtain impressions of prospects and performance. Instead, they relied on family or social ties with an active constituent, reputational assessment of the managing owner and his allies, and local knowledge. They monitored a venture over time with the aid the accounts that managing agents circulated at the end of each voyage and informal communication with he or other key members. Those included in the 'passive fringe' often consisted of wealthy farmers, local manufacturers of non-marine products, professionals, women, and the ubiquitous 'gentleman'. Shipping networks exhibiting this basic architecture proliferated and endured through the nineteenth century and into the twentieth because they furnished an effective framework for arranging contracts based on knowledge.

A Networked World

Between 1850 and 1914, world trade grew rapidly and production became more specialized geographically as a result of the spread of liberal trading policies, monetary stability based on the gold standard, population growth, the installation of a worldwide telegraph network, and the increasing mechanization of land – and sea-ward transportation systems. By the end of this period, 'the first global economy' had matured. Michael Miller (2012) argued that worldwide trading and shipping operations became enveloped in three networks centered on major European ports. First, hinterland connections radiated outward from the major ports of Britain and continental Europe. These links collected manufactured goods for export to developing regions and distributing primary products imported from colonies and areas of recent settlement. Second, merchant enterprise created global webs that organized the flow of commodities, consumer goods, and capital equipment, and in many instances these businesses played important roles in funnelling investment funds to productive opportunities overseas (also see Jones, 2000). Finally, shipowners, especially those operating liner services, created the oceanic transport systems needed to move people and goods around the globe. The economic structures that operated within these three sets of connections were network-based: port communities were interlaced with informal links, merchant enterprises were web-like cooperative constructs, and, of course, merchant shipping relied on its traditional information and resource gathering frameworks that became larger and increasingly complex.

The operation of these three interrelated systems was conditioned by the same commercial culture. This has been termed 'gentlemanly capitalism' and it grew from the transference of aristocratic patterns of behavior and standards of conduct to the business arena as a result of increasingly intimate interactions between members of the British ruling elite and commercial interests, especially bankers and financiers (Cain and Hopkins, 1993a,b). It was based on traditional conceptions of trust and honor, a common understanding of fairness and compromise, and recognition of mutual interdependence. As social and economic mobility unfolded over time, new occupational groups, including first merchants, then shipowners and others linked to maritime endeavor, and later yet major industrialists, came to observe the basic tenets of gentlemanly capitalism. Moreover, maritime entrepreneurs around the world followed suit. Although Miller detected the concurrent operation of other perspectives and mind sets at work, gentlemanly capitalism provided a central and vital cultural foundation for arranging complex transactions across the first global economy.

Gentlemanly capitalism also shaped behavior in exchange facilitating ways across the commercial infrastructure that sustained maritime business. A series of major business centers arose across Europe – Hamburg, Antwerp, Rotterdam, and lesser ones like Le Havre and Marseilles – but London emerged as the largest and most important of all. Its rise as the center of world trade, finance, and insurance was tied to the growth of specialized markets for goods and services. London became a veritable emporium where traders in specific commodities gathered first in seventeenth century coffee houses and then later in grand purpose-built venues to make transactions. The origins of the London Metal Exchange can be traced to the Jerusalem coffee house, the Virginia and Maryland coffee house provided the foundation for the Baltic Exchange, the world's center for commodity trading and ship hiring, and most famous of all Edward Lloyd's establishment in Tower St. developed into Lloyd's of London where insurance coverage was arranged across the globe. London hosted the Royal Exchange, the Stock Exchange, and a plethora of specialist commodity markets that set prices and standards for worldwide trade. Within these transacting arenas, networks also arose as individuals gained sound reputations and attracted allied dealers (Boyce, 2012). Norms and ethical standards associated with gentlemanly capitalism found expression in the motto of the Baltic Exchange, "our word is our bond" and that of Lloyds 'sine praejudicio' (without prejudice). Aided by this ethos, these markets working in combination constituted a vital commercial infrastructure for worldwide trading and shipping enterprise and the efficacy of this cultural edifice is reflected in the role it played in helping to sustain maritime networks during a prolonged period of salient technological and legal change.

Changes in Technology, Legal Regulations, and Organization

Adam Kirkaldy (1914) argued that the application of steam power to merchant shipping, a process in which Britain led the way, unfolded at a gradual pace between 1820 and 1890, but it induced fundamental changes in legal and financial structures within the sector. Steam propulsion made it possible to form liner services that involved regular calls at specified ports according to a fixed timetable. To run this type of business, shipowners had to make major investments in fleets of expensive vessels suited to cargo flows generated from the individual ports they visited. In addition, they had to develop complex support services. As a result, fixed costs increased significantly, and these businesses sought to attract high value general cargo that could sustain high freight rates through fast and reliable service. (High valued general cargo imposed onerous inventory costs upon merchants who were thus willing to pay more for its transport.) Kirkaldy observed that Britain's subsidized mail and passenger lines which appeared between 1830 and 1850 employed the joint-stock company and later became limited liability firms following the passage of general incorporation laws between 1855 and 1862. After 1860, private partnerships also shifted to limited liability enterprises to support liner operations. In the tramping sector, based on the carriage of full loads of low value bulk commodities between any ports nominated by merchants, single-ship limited firms appeared in large numbers and, in the 1870s, precipitated a speculative frenzy supported by investors from low and middling occupations. Yet, long-standing reputable shipowners retained the 64th system for several decades, in part because they and their allies believed that it conveyed valuable signals of personal commitment and probity. Moreover, contrary to Kirkaldy's assessment, these technical, legal, and operational developments did not undermine the traditional network foundations of maritime enterprise in Britain and continental Europe.

The early subsidized liner enterprises were formed by bands of allies, well known to each other, who pooled the required capital, physical assets, and business connections; most can be classified as economic cooperative networks that sustained subsequent growth by attracting a passive fringe of investors. Privately supported limited liability liner firms were configured so that the control of managing agents' families was consolidated and included active participants within a large social-integrative construct. Over time, these enterprises brought in outside rentier investors usually by offering them securities with restricted voting rights. In neither case did ownership and management become decisively divorced and administrative tasks did not require the creation of hierarchies. Within tramping, blood kinship and social integrative networks persisted and, like the liner firms, they developed passive fringes. (Traditional 64th managers, like Walter Runciman and James Knott, often developed close relations with nonactive participants; what nowadays is called 'investor relations' were almost intimate in these cases.) In both parts of the industry providers of support services were included in network cores, giving them an economic cooperative flavor, but social ties provided their real foundation. Principals and agents apprenticed each others' sons and nephews, and marriage ties often furnished additional adhesion. Network dynamics, norms, and cognitive frames extended across a wider group of allies beneath the surface of new legal forms which were manipulated to preserve private control (Boyce, 1992).

After the shipping cycle peaked at abnormally high levels in 1900–01 as a result of the Boer War combined with a surge in world trade, changed environmental conditions again caused network adjustment. A prolonged recession (1902–11), increasing foreign competition, and the consolidation of the conference system (a form of economic cooperative network that liner firms used to contain competition) lay behind this development. Most tramp owners continued building up their fleets and many consolidated their single ship firms into a limited liability company, but they still relied mainly on home port networks. However, a few did establish offices in London to tap into the capital's dense webs of information and other key ports from which they drew custom. The main liner operators enlarged their local networks by allying with figures of wider geographic and occupational affiliation. Lord Inverclyde (Cunard) and Lord Inchcape (P&O) established ties with government, while shipowners who exploited new specialist trades (fruit, oil, and refrigerated meat carrying) built new links with cargo suppliers and distributors. Investment bonds appeared between shipping and shipbuilding, steel, engineering, and coal firms – those formed by Sir Christopher Furness (later Lord Furness) around 1900 being the most noteworthy – and some large shipbuilders, especially Harland and Wolff, established close long-term contractual ties with their clients. In these ways, the extent and nature of network linkages shifted to secure access to new sources of information and emerging bases of specialized knowledge needed to sustain growth.

After world trade revived strongly in 1911, the structure of British shipping changed again as five major corporate groupings appeared. These enterprises – the so-called Big Five: Cunard, P&O, Furness Withy, Ellerman, and the Royal Mail – were built on the holding company structure, consisting of one or more parent entities that controlled and sometimes managed stables of subsidiary shipping, industrial, and service units. This development was supported by high profits and encouraged by the increased availability of acquisition targets as the first generation of steamship entrepreneurs retired or passed away. Growth by take-over continued during the World War I as u-boat losses, constrained access to shipbuilding berths and Excess Profit Duties encouraged acquisitions of shipping firms to secure tonnage and widespread backward extension of ownership ties with heavy industry. The big groups became interorganizational networks in which parent company nodes occupied positions of increased centrality relative to growing numbers newly acquired subsidiaries. For each of the Big Five, take-over patterns usually followed previously established communication lines and channels of interpersonal knowledge. Both defensive and offensive considerations propelled growth. Thus, crises in managerial succession exposed long-standing interfirm links which dominant firms safeguarded by acquiring the faltering firm, while other companies with valuable goodwill, knowledge, and contacts attracted bids from the majors. To fund these acquisitions, the main groups employed outside finance, especially debt, on a much larger scale and their leaders, all

men of growing economic and social stature, built stronger links with banks. Yet, the great shipowners juggled financial obligations with a widening array of capital suppliers in ways that preserved their control and freedom of action. As a result, underlying network structures remained intact: Some groups, like Cunard were sustained by interfamily ties of the social integrative type, while others were controlled by a single family and developed features of the economic cooperative network.

The Emergence of New Challengers and the Decline of British Maritime Networks

On the eve of the World War I, British shipowners commanded about 40% of the world's merchant fleet (Sturmey, 1962). Other countries attained less formidable but still significant positions in the industry with the aid of network structures. However, most other nations lacked the huge reserves of family-based wealth needed to found purely blood kinship networks and therefore relied on other economic agents to assist growth.

In Germany, the second largest shipping nation by 1914, merchants and other members of port communities worked together to pioneer steam services. After 1900, the great German lines, HAPAG and Norddeutscher Lloyd, forged supporting alliances with industry, government, and major banks. In contrast, Woerman, who was a specialist operator continued to rely on merchant links to build his business. Government aid in a variety of forms was vital to the expansion of Japan's merchant navy which ranked third in the world (Davies, 2008). The United States had built up a position as the foremost shipping country in the age of sail with the support of port-centered family networks, but after the Civil War these constructs had been destabilized by social and cultural turbulence and by a shift in business opportunity in favor of landward pursuits with the development of the West. America was a minor player before 1914 because it lacked any comparative advantage in steam shipping, and the few enterprises that arose in the oceanic transport sector were developed by corporate groupings consisting of railroads and banks. At the turn of the century, J.P. Morgan attempted to push the US to the forefront by merging several leading British and American lines under the banner of the International Mercantile Marine Corporation and aligning this new entity with German and Dutch shipping firms as well as Belfast shipbuilder Harland & Wolff. However, the combine was overcapitalized and struggled to earn sufficient profits with the result that it fell apart after World War I. In contrast, Greek, Spanish, and Scandinavian shipping employed family and social/cultural networks to build up their merchant navies before 1914.

World War I disrupted established shipping routes and changed patterns of world trade. America and Japan emerged as more important focal points of international commerce, while Russia largely withdrew from the world economy. Debate has developed as to whether the global economy that existed before 1914 retreated in the interwar period, or whether the internationalization of activity continued but in different directions, geographically and in terms of commodity trades. Brink Lindsey (2001) argued in favor of retreat whereas Michael Miller (2012) supported the view that ongoing development occurred but in new forms. It is true that the transatlantic passenger business declined when the US began restricting immigration, but streams of migrants from Eastern Europe and from within Asia arose to create new demand. Some long-established cargo liner routes declined or stagnated but new opportunities arose after the opening of the Panama Canal and some expanding South American trades provided increased employment for shipping. Britain's coal exports declined as new sources of supply emerged and as oil became an increasingly important substitute fuel. These developments reduced the volume of outbound cargoes for tramp ships leaving Europe and deprived many of two-way trading opportunities.

Adjustment to these shifting conditions was hindered by persistent worldwide oversupply of shipping capacity in relation to the volume of goods requiring transport. Despite enormous war losses, between 1913 and 1920 the global fleet grew by about 25% to 51.6 million GRT, as American shipyards, in particular ramped up production during the conflict. By 1939, it had reached 65.6 million GRT. Meanwhile, by 1924 the volume of seaborne trade returned to its prewar level and continued growing until it fell heavily in 1929 with the onset of the Great Depression. By 1939, trade has regained the level attained in 1929. Excess capacity dogged the industry even though the British, Japanese, French, and German governments introduced scrap and build subsidies to reduce the supply of obsolete tonnage.

Nevertheless, the period witnessed important shifts in the composition of the world fleet. Germany displayed remarkable resilience by reconstructing its trading networks after the nearly complete destruction of its merchant navy during the war. Japan made enormous strides as family-owned companies took advantage of government support and help from trading companies, industrial enterprises, and financial institutions affiliated with the Zaibatsu (family-controlled holding companies which resembled a cross between all three network forms identified by Alvesson and Lindkvist (Davies, 2008)). From this institutional foundation, Japanese shipping developed new routes and broke into European dominated conferences on a wider scale. Greek shipowners also reformed their networks and penetrated a larger range of deep sea tramping trades. Norway also enlarged its tramping business and became a major force in tanker operations. By 1939, the British fleet was still the world's largest but it had declined absolutely and relatively, accounting for 27.3% of the total. America came second with 13.6% but much of its shipping was government-owned surplus tonnage held in strategic reserve. Japan ranked third (8.5%), followed by Norway (7.3%), Germany (6.9%), and Italy (5.2%).

World War II witnessed further massive tonnage losses. After 1945, Britain's fleet ranked second in the world at 18 million GRT, behind a greatly expanded US merchant navy, which though it totaled nearly 41 million GRT, or 56% of the world total, was largely sold off to private owners and mothballed as a strategic reserve. The German, Japanese, and Greek fleets were virtually destroyed, while Norway's was reduced by half.

World trade underwent profound changes in the postwar era. With the reconstruction of Japan and Europe, a new form of globalization appeared. Decolonization broke apart the

Europe-centered routes which had previously connected dependencies with the main imperial nations. Increasingly, global exchange was concentrated between three main blocs: the US Europe, and east and South Asia. These key areas had subsidiary connections to South America, Africa, and Australasia. However, the extent to which the world was enveloped within a uniform 'global' system of exchange is subject to debate. Rugman (2003) argued that most investment and trade has been focused on pairs of the three main cores rather than embracing all of these areas of concentration and their satellite networks. Nevertheless, the nature and form of connectedness across the world changed significantly, and several factors played key roles in precipitating the shift.

Fundamental technical and institutional developments unfolded concurrently. First, vessel size increased to unforeseeable proportions as economies of scale were exposed by new modular construction techniques, more efficient propulsion systems, and a steep rise in the volume of goods requiring transport. These dynamics affected the tanker sector first, with particular poignancy when the Suez Canal was closed in 1956 and again in 1967, and then spread to the bulk trades where new ship designs eroded the market for traditional tramping services. Giant tankers and bulkers carried oil, iron ore, coal, and other basic commodities over greater distances, in larger quantities, and at lower unit cost than ever before. Other large specialized vessels appeared especially liquefied natural gas (LNG) ships, chemical transports, and car carriers. The passenger trades disappeared by the mid-1970s as jet aircraft of progressively greater size took to the air, but after a lag, the cruising industry blossomed and propelled by scale effects built large numbers of palatial ships of great size.

The general cargo or liner sector was completely transformed by containerization and the application of computer technology. Here to, the pursuit of economies of scale was paramount and their extent was accentuated by rapid growth in high-value manufactured goods that had been the preserve of the traditional cargo liners. Shipping items in boxes also transformed ports as facilities in city centers could no longer accommodate the huge volume of shipments being delivered and mustered for loading. Quays, storage yards, and craneage appeared at river mouths and new coastal sites that could accommodate massive container ships. Liverpool and London, both traditional centers of the general cargo trade, declined but other, long-established ports, like Rotterdam, adapted. New global container hubs appeared in what were previously secondary centers such as Hong Kong, Oakland, and Singapore. Containerization spawned intermodal transport that combined sea, road, and rail carriage with the aid of computer tracking and the application of logistics.

Institutionally, free trade policies promulgated by international organizations and supported by intercountry agreements encouraged rapid growth in global exchange. The deregulation of investment also spurred wider economic development. While flags of convenience had been exploited previously, they came into much wider use after 1945. Greek and American operators especially took advantage of the lower crewing costs and lax regulatory prescriptions these devices brought. All of these developments made national identity problematic within the new globalizing context.

Together, these changes challenged traditional maritime networks and created openings for new cooperative structures led by entrepreneurs from different backgrounds. Greece's long-standing island-based alliances were confronted by newcomers who rapidly built up large enterprises. In the tanker sector, Aristotle Onassis and Stavros Niarchos forged interorganizational connections with shipyards, financial institutions, and customers to achieve astonishing rates of growth. Firms that had been minor entities quickly achieved prominence in specialized sectors, for example the chemical carrying business run by Odfjells (Norway). Newcomers like Ted Arison and Sammy Ofer helped to transform cruising, while containerization was driven by men, like Malcolm McLean and Bruce Seaton, who came from outside the shipping industry. Denmark's major firms built conglomerate ties, and Norwegian owners formed long-term financial and chartering contracts to expand their businesses. The Japanese reconfigured their zaibatsu ties into vertically aligned keiretsu and pursued dynamic growth strategies through these interorganizational links. Entrants from South Korea, Taiwan, Singapore, and Hong Kong also created large networked enterprises. Despite their origins and principal operating bases, these entrepreneurs created multilateral affiliations that can be described as global rather than merely national.

Meanwhile, traditional liner firms quickly responded to containerization by forming national consortia, or economic cooperative networks to harness the resources needed to build the expensive fleets required in the new operating environment. However, these constructs were overwhelmed by the pace of change and replaced by international groupings. From the 1960s, network ties exhibited forms and a supporting culture that differed markedly from those that had previously sustained traditional European maritime endeavor. Now, competitive vigor was more overt (the decline of the conference system was symptomatic of a new dynamic), new skills based on computerized logistics became paramount, a university education assumed greater value than the old style on-the job and in-the-firm training processes, and the relevance of interpersonal relations and face-to-face contact declined. The networked basis of maritime business did not disappear, but it was utterly transformed.

Britain's shipping industry entered a period of precipitous decline in the 1970s. The private foundations of the sector had been undermined by an onerous postwar income tax regime, punitive inheritance levies, and new laws that destabilized privately owned companies. Labor governments had nationalized related industries – coal, steel, and shipbuilding – that had hitherto assisted shipping. New labor regulations hurt productivity and sapped investor confidence. Decolonization and the spread of discriminating legislation by the new governments of former colonies destroyed some of Britain's long-standing routes and dislocated the old trading companies that had worked in tandem with national vessel operators. Most fundamental, perhaps, were social and cultural changes that eroded the traditional culture upon which the maritime networks depended. As the relevance of gentlemanly capitalism receded, British shipowners found it increasingly difficult to extend their networks by forming new weak links that could have brought the new skills and information needed to exploit new opportunities.

Long-standing communication channels withered and patterns of socialization that had sustained family business and networking activities were confounded. Many aging industry leaders were discouraged and demoralized by changes unfolding from within the business arena and from the external environment (Boyce, 2010). Younger men from different educational and national backgrounds configured new cooperative constructs to exploit emerging opportunities.

Bibliography

Alvesson, M., Lindkvist, L., 1993. Transaction costs, clans, and corporate culture. Journal of Management Studies 30, 427–452.

Boyce, Gordon H., 1992. Sixty-fourthers, syndicates, and stock promotions: information flows and fundraising techniques used by British shipowners between 1870 and 1914. Journal of Economic History 52, 181–205.

Boyce, Gordon H., 1995. Information, Mediation, and Institutional Development: The Rise of Large-Scale Enterprise in British Shipping, 1870–1919. Manchester University Press, Manchester.

Boyce, Gordon H., 2003. Network knowledge and network routines. Business History 45, 52–76.

Boyce, Gordon H., 2010. Language and culture in a Liverpool merchant family firm, 1870–1950. Business History Review 84, 1–26.

Boyce, Gordon H., 2011. Edward Bates & Sons, 1897–1915. Tramping operations in recession and recovery. International Journal of Maritime History XXIII (1), 13–50.

Boyce, Gordon H., 2012. The development of commercial infrastructure for world shipping. In: Harlaftis, G., Tengold, S., Valdaliso, J. (Eds.), The World's Key Industry: History and Economics of International Shipping. Palgrave Macmillan, London, pp. 206–223.

Cain, P.J., Hopkins, A.G., 1993a. British Imperialism: Innovation and Expansion, 1688–1914. Longman, London.

Cain, P.J., Hopkins, A.G., 1993b. British Imperialism: Crisis and Deconstruction, 1914–1990. Longman, London.

Casson, Mark, 1997. Information and Organization a New Perspective on the Theory of the Firm. Clarendon Press, Oxford.

Davies, Peter N., 2008. A guide to the emergence of Japan's modern shipping industries. In: Fischer, Lewis R., Lange, Even (Eds.), International Merchant Shipping in the Nineteenth and Twentieth Centuries the Comparative Dimension. International Maritime Economic History Association, St John's Newfoundland, pp. 105–124.

Davis, Ralph, 1962. The Rise of the English Shipping Industry in the Seventeenth and Eighteenth Centuries. Macmillan, London.

Granovetter, M.S., 1973. The strength of weak ties. American Journal of Sociology 78, 1360–1380.

Harlaftis, Gelina, 1996. A History of Greek Owned Shipping. Routledge, Basingstoke.

Jones, Geoffrey, 2000. Merchants to Multinational: British Trading Companies in the Nineteenth and Twentieth Centuries. Oxford University Press, Oxford.

Kirkaldy, Adam W., 1914. British Shipping Its History, Organization, and Importance. London.

Lindsey, Brink, 2001. The decline and fall of the first global economy. Reason, 43–51.

Miller, Michael, 2012. Europe and the Maritime World: A Twentieth-Century History. Cambridge University Press, Cambridge.

Powell, Walter W., 1990. Neither market nor hierarchy: network forms of organization. Research in Organizational Behaviour 12, 295–336.

Richardson, Gordon, 1972. The organization of industry. Economic Journal 82, 883–896.

Rugman, Alan, 2003. Regional strategy and the demise of globalization. Journal of International Management 9, 409–417.

Starkey, David J., Harlaftis, Gelina (Eds.), 1998. Global Markets: The Internationalization of the Sea Transport Industries since 1850. International Maritime Economic History Association, St John's Newfoundland.

Sturmey, Stanley G., 1962. British Shipping and World Competition. Athlone Press, London.

Tengold, Stig, 2000. The Shipping Crisis of the 1970s: Causes, Effects, and Implication for Norwegian Shipping (Ph.D. dissertation). Norwegian School of Economics and Business Administration, Bergen.

Valdaliso, Jesus M., 2002. The rise of specialist firms in Spanish shipping and their strategies of growth, 1860–1930. Business History Review 74, 268–300.

Wray, William, 1984. Mitsubishi and the NYK, 1870–1914. Harvard University Press, Cambridge, MA.

Yui, Tsunehiko, Nakagawa, Keiichiro (Eds.), 1985. Business History of Shipping. University of Tokyo Press, Tokyo.

Markedness in Linguistics

Edwin Battistella, Southern Oregon University, Ashland, OR, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E. Battistella, volume 13, pp. 8945–8948, © 2001, Elsevier Ltd.

Abstract

Originating in the Prague School structuralism of Roman Jakobson and Nikolai Trubetzkoy, the concept of markedness – asymmetry in linguistic structure – was proposed as a means of characterizing binary oppositions in phonology and soon after in grammar, meaning, and culture.

Today, markedness ranges over a variety of formal and typological senses and applications, but is not without those who question the coherence of the term.

'Markedness' refers to an asymmetry relation between elements of linguistic or conceptual structure. In a marked–unmarked relation, one term of an opposition or of a set is taken to be the dominant one while the other(s) is taken to be secondary. The dominant term is known as the unmarked term and the secondary one(s) is marked. Linguistic markedness ranges over phonological, grammatical, and semantic oppositions and constructions. The concept has been extended to the analysis of cultural relationships as well, with background or matrix cultural categories (right-handedness, sightedness, etc.) being taken as unmarked.

Background in the Prague School

As Henning Andersen (1989) has noted, the idea of asymmetry in linguistics predated the actual coining of the terms 'marked' and 'unmarked' in the 1930s, going back to work in the early 1800s. The modern concept of markedness originated in the Prague School structuralism of Roman Jakobson and Nikolai Trubetzkoy, and was proposed as a means of characterizing binary oppositions. In Prague School structuralism, both sound and meaning were analyzed into systems of binary distinctive features representing sound and meaning as the presence or absence of relational properties or features. Binarism suggests symmetry and equivalence in linguistic analysis; markedness added to that the idea of hierarchy and value. The fact that distinctions were characterized as the presence versus the absence of some property implied an evaluation of any opposition into a mark and its absence. Thus, for example, phonological oppositions such as nasal versus nonnasal were defined as the presence versus the absence of nasality; nasality was marked, nonnasality unmarked. Grammatical oppositions such as singular versus plural were defined as the presence versus absence of plurality.

Oppositions manifest asymmetry in several ways, most notably for Trubetzkoy in terms of the diagnostic of neutralization, whereby the unmarked term of a phonological opposition was thought to be that which occurred in positions where the contrast was neutralized (that is, where only one term of the opposition could occur). In his 1932 article "Structure of the Russian verb," Jakobson extended the concept of asymmetry to a view of grammatical meaning in which the marked element 'announces the existence of [some meaning] A' while the unmarked element 'does not announce the existence of A, i.e., does not state whether A is present or not.' This meant that an opposition like past versus nonpast characterized three interpretations: past, present (as the opposite of past), and timelessness (the neutralization of the tense opposition). More generally, any semantic opposition involved three meanings: two unmarked and one marked. Thus, old can refer to the opposition of young or it can refer to age itself, as in "How old are you?" (="What is your age?").

The breadth of the unmarked terms was extended by Jakobson to oppositions that were not inherently binary. In a 1972 article, Jakobson described language by saying "every single constituent of a linguistic system is built on an opposition of two logical contradictories: the presence of an attribute ('markedness') in contraposition to its absence ('unmarkedness')." Jakobson also suggested a role for markedness in language acquisition and loss, in his 1941 monograph *Child Language, Aphasia, and Universals of Language*. There Jakobson proposed a relationship between a universal feature hierarchy, the acquisition of sound distinctions by children and the loss of sound distinctions during aphasia. Today few linguists or aphasiologists would adopt Jakobson's positions literally, though many view them as useful tendencies (see Caramazza and Zurif, 1978).

Semiotically oriented work in the Prague School tradition has often viewed markedness as a means to investigate the isomorphism of form and meaning. Henning Andersen and Michael Shapiro, for example, have fruitfully linked markedness to semiotic universals and to the work of Charles S. Peirce. The work of Cornelius van Schooneveld, Edna Andrews, Rodney Sangster, Yishai Tobin, and others had foregrounded the notion of 'semantic invariance' (understood as the different general meanings reflected in the contextual specific meanings of features). The most well-known semantic invariants are those in Jakobson's analysis of the Russian case and verbs systems which attempt to ground semantics in spatial relations and in the relations among participants in the overall speech event (for recent work on invariance, see Waugh and Rudy, 1991).

Typological Universals

Markedness has also been applied to typological and functional approaches to language, influenced by Joseph Greenberg's 1966 book *Language Universals*. Greenberg surveyed the

correlates and diagnostics of markedness proposed by Jakobson, Trubetzkoy, and others in an attempt to find parallel diagnostic criteria across levels of language. Noting that unmarked grammatical categories in a language show characteristics correlated to the greater frequency of a category vis-à-vis its opposite, Greenberg took frequency to be the primary determining factor of markedness in grammar and suggested that unmarked categories could be determined by 'the frequency of association of things in the real world.' In addition to emphasizing the role of frequency in determining markedness within a language, Greenberg applied frequency cross-linguistically, taking unmarked categories to be those that are unmarked in a wide number of languages. However, while frequency is an appealing diagnostic, it is also problematic because a phenomenon may be infrequent cross-linguistically but predominant in a particular language (a much-used example is preposition stranding in English, which appears to be marked cross-linguistically but unmarked in English). See Pustet (2009) for an attempt to revisit the criterion of frequency and to explain markedness in terms of Zipf's Laws.

Universals have also been proposed on the basis of implicational laws, with a category taken as generally marked if every language that has the marked category also has the unmarked one but not vice versa. Other approaches to universal markedness relations focus on functional economic and iconic motivations, tying recurring asymmetries to properties of communication channels and communication events. Croft (1990), for example, notes that asymmetries among linguistic elements may be explainable in terms of economy of form, in terms of iconism between the structure of language and conceptualization of the world.

Comrie (1983) treats markedness in terms of human interaction with others and with the world rather than as a purely formal property of language. Similarly, Givón (1990) suggests that markedness is related to cognitive complexity – 'in terms of attention, mental effort, or processing time.' Mayerthaler, Dressler, and other 'naturalists' view markedness relations in terms of the ways in which extralinguistic principles of perceptibility and psychological efficiency determine what is natural in language. Mayerthaler (1988), for example, defines unmarked as "in agreement with the typical attributes of the speaker." In other words, 'up' and 'front' are unmarked concepts because we perceive thing in terms of growing up and looking frontward, rather than the opposites (see also Bybee, 2010).

Markedness in Generative Grammar

Markedness entered generative linguistic theory through Chomsky and Halle's *The Sound Pattern of English*. Following the Prague School view, phonological features were assumed to be part of a universal phonetic vocabulary. In addition, features were also incorporated into what was known as the 'evaluation metric,' a central concept in early versions of generative grammar. The evaluation metric was a means of selecting the most highly valued grammar capable of describing the facts of a language. In *The Sound Pattern of English*, the value of a grammar was the inverse of the number of features required in that grammar. However, Chomsky and Halle realized that their initial approach to phonological features made implausible rules and segment inventories as highly valued as natural ones. To deal with this, they introduced markedness as a means of encoding the intrinsic content of features. Chomsky and Halle proposed a set of 39 universal marking conventions, tying naturalness to evaluation. In their system, the unmarked value of a feature was cost-free with respect to the evaluation metric, while the marked values were counted by the metric. Segment inventories could also be evaluated according to the number of marked features required to characterize them; sets of lexical items could be evaluated for markedness based on the number of marked features they entailed; and a concept of 'linking' reduced the number of features required to characterize processes corresponding to the markedness conventions. However, the use of phonological markedness as part of the evaluation metric was never able to fully account for the fact that some features are more likely than others or for the fact that phonological systems must have a certain minimal complexity and symmetry (see Kean, 1980 for an early extension of the featural markedness theory to address these issues).

In syntax, markedness as feature-evaluation did not receive the same attention that it did in phonology. In Chomsky's work, markedness evolved to a view of unmarked properties as an innate preference structure based first in constraints and later in parameters of universal grammar. Beginning in the l970s, Chomsky attempted to greatly restrict the class of possible grammars, suggesting a relative rather than an absolute interpretation of conditions on rules. In such an approach, rules apply in accordance with the conditions unless specifically complicated to evade them, with Chomsky noting in his article "On Wh-movement" that "The logic of this approach is essentially that of the theory of markedness." In their 1977 article "Filters and control," Chomsky and Howard Lasnik extended this to view markedness as part of a theory of 'core grammar':

> We will assume that [Universal Grammar] in not an 'undifferentiated' system, but rather incorporates something analogous to a 'theory of markedness'. Specifically, there is a theory of core grammar with highly restricted options, limited expressive power, and a few parameters. Systems that fall within core grammar constitute 'the unmarked case'; we may think of them as optimal in terms of the evaluation metric. An actual language is determined by fixing the parameters of core grammar and then adding rules or conditions, using much richer resources, … These added properties of grammars we may think of as the syntactic analogue of irregular verbs.
>
> 430

What is unmarked is the entire core grammar of a language, which is then supplemented by a marked periphery. In later versions of his theories, Chomsky has explored various positions on markedness, for example, associating the unmarked with a child's initial hypothesis. In a 1986 work, he summarized his view as follows:

> The distinction between core and periphery leaves us with three notions of markedness: core versus periphery, internal to the core, and internal to the periphery. The second has to do with the way parameters are set in the absence of evidence. As for the third, there

> are, no doubt, significant regularities even in departures from the core principles (for example, in irregular verb morphology in English), and it may be that peripheral constructions are related to the core in systematic ways, say by relaxing certain conditions of core grammar.

Other researchers in generative syntax have tried to flesh out aspects of markedness, attempting to operationalize the idea of the cost of a marked syntactic construction or integrate markedness into the theory of parameters. Various approaches have been suggested, including tying markedness to an ordering of parameter values defined by subset relations among sentences, connecting it to types of triggering experiences or to default learning strategies, and using markedness to establish a universal syntactic feature hierarchy. Optimality theory (OT) approaches emerging in the 1990s also incorporate markedness. In such approaches, the surface forms of language arise from the interaction between conflicting constraints as underlying representations are mapped to outputs. OT models encode markedness in the constraints and their ranking (see Woolford, 2007).

Phonological Markedness in OT

OT has also reenergized the study of markedness in phonology. Paul Smolensky, one of the founders, characterized markedness this way in his article "Harmonic completeness, local conjunction, and feature domains":

> The working hypothesis is that 'less marked' can be formalized in OT as **more harmonic** according to markedness constraints, and that universal markedness relationships such as '[−φ] is less marked than [+φ]' are encoded in UG via universally fixed rankings.
>
> 2006: 36

Smolensky notes too that "Less marked elements are less active" in terms of phonological processes though he notes that this is 'far from a crisp result' (37). Initially, the markedness constraints of OT assigned violation-marks to output candidates in isolation – without reference to the input or to other candidate outputs. John McCarthy (2003) proposed an alternative, more global approach in which markedness constraints compare the candidate under evaluation with another candidate, the one most faithful to the constraints: a candidate output might thus contain a marked structure also present in the fully faithful candidate or it might contain a marked structure not present in that candidate output, and the two approaches yield different constellations of results. Paul de Lacy (2002, 2006) has explored how segmental, sonority, and tonal hierarchies combine with prosodic structures to constrain phonological outputs and has proposed that the markedness constraints fall into two semi-antagonistic sets: one that seeks to eliminate marked elements in the hierarchy and one that seeks to preserve them. These antagonistic sets interact to produce varying markedness in different situations: asymmetries in place neutralization, assimilation, deletion, coalescence, and vowel reduction.

Rice (2007) argues "Natural markedness criteria including language-particular and cross-linguistic frequency, ease of articulation, perceptual salience, and likely historical and social factors as well work together to determine the tendencies toward certain phonetic outputs" (96). She offers a critique of the diagnostic of neutralization arguing instead for a more refined view of the neutralization, epenthesis, together with situations in which marked features are preserved (such as some types of assimilation), which together can help to account for cross-linguistic variation. She writes "While two languages with identical surface contrasts within a class may differ in how they pattern with respect to submergence-of-the-unmarked diagnostics, this variation is not without limit: it is constrained, with only certain features demonstrating unmarked patterning in a given set of contrasts" (2007: 93). And in a somewhat different approach Elizabeth Hume (2005, 2011) has argued that the basis of phonological markedness lies in the cognitive process of expectation, quantified probabilistically. Expectation is defined according to the linguistic and/or extralinguistic context of a pattern and is tied to the individual speaker. Hume's approach thus reverses the usual approach of deducing markedness from universals to predict both individual and cross-linguistic markedness from the patterns up.

Second Language Acquisition

Some researchers – both generative and typological – have applied markedness to second language acquisition theory, treating it as an inherent learning hierarchy which reflects the sequence in which constructions are acquired, the difficulty of acquiring certain constructions, and the transferability of rules across languages (see Rutherford, 1982 for an early survey). As Gregg (1993) points out, in order for markedness to be useful as a potential explanation of second language (L2) acquisition, "it is necessary to define markedness in such a way as to connect it with learning mechanisms within the individual learner." Fred Eckman (2004) has proposed that markedness plays an especially crucial role in L2 phenomena which cannot be explained in terms of either transfer from the learner's native language or in terms of target language (his example involves the occurrence of resumptive pronouns in the relative clauses of learners of Swedish as an L2 which can be explained in terms of the Accessibilty Hierarchy proposed by Keenan and Comrie (1977); for a phonological example, see Ellen Broselow et al. (1998).

Markedness and Code Switching

Carol Myers-Scotton (1983, 1993) has suggested a markedness model of variation in linguistic code choices – code switching. The premise is that participants in conversation understand all code choices as unmarked or marked in reference to the norms of their speech community (1983: 115). She proposed that speakers choose forms to index "the set of rights and obligations which [they] wish to be in force between speaker and addressee for the current exchange" (1993: 113). The unmarked choice of form is associated with the social norms while marked language constitutes a speaker's negotiation for a different social norm. Myers-Scotton sees code switching as

having four functions: (1) code switching may be a series of unmarked code choices, with each switch unmarked choice because of shifting social or conversational context; (2) code switching may be the unmarked choice between bilingual peers; (3) code switching may be a marked choice showing noncompliance with some norm; and (4) code switching may be an exploration to find the unmarked code when it is not evident in a context (Myers-Scotton, 1993). The model treats the more frequently used code as the unmarked, matrix language and the less frequently used code as the marked embedded language. As Myers-Scotton notes, the matrix language-embedded language model is not just about linguistic code switching, but extends to any situation in which a speaker or writer manages different styles or registers. The papers in *Codes and Consequences* (1998), discussed below, apply the markedness model to style shifting more broadly.

Cultural Markedness

A main theme of markedness is the information content and information value of an element. Thus many researchers have seen markedness as an encoding of that which is unusual or informative. Conceptual familiarity with cultural norms provided by familiar categories creates a ground against which marked categories provide a figure, opening the way for markedness to be applied to cultural and social categorization. In the 1930s Jakobson had already suggested applying markedness to any opposition, explicitly mentioning such pairs as *life/death*, *liberty/bondage*, *sin/virtue*, and *holiday/working day*. Waugh (1982) extended this idea to note that oppositions like *male/female*, *white/black*, *sighted/blind*, *hearing/deaf*, *heterosexual/homosexual*, *right/left*, *fertility/barrenness*, *clothed/nude*, and *spoken language/written language* can all be treated as unmarked/marked cultural oppositions. And McCawley (1985) suggested that aspects of belief systems and even scientific theorizing could be analyzed as systems in which some assumptions are culturally unmarked. Interrogating our culture in this way prompts us to view our cultural architecture in terms of implicit assumptions. We refer to *poets* and *writers*, for example, establishing an opposition between the two (and implying perhaps that a poet is not a writer). And terms like *homeless* (as opposed to the logically possible but nonexistent word *homeful*) imply that having a home is the linguistically unmarked case.

The papers in Myers-Scotton (1998) demonstrate ways in which intentionality is conveyed by adopting systems of stylistically and culturally marked choices, as opposed to mere isolated examples. Such approaches hold promise as a descriptive technique for domains beyond language, and recent work has applied markedness to music, myth, and facial expression: among other things (see Hatten, 1994, 2004; Liszka, 1989). In his analysis of musical styles and television genres, for example, Hatten identifies marked and unmarked musical oppositions (such as minor vs. major modes) and demonstrates their correlation with theme (such as tragic vs. nontragic expression) and styles (Classical vs. Baroque). Following Hatten, Ron Rodman used markedness to model changing aesthetic norms in television police dramas, looking at the spread of jazz into the musical genres previously dominated by military themes. Markedness serves as the innovative aspect of style growth or style change. James Liszka has applied markedness to elementary forms of narration, specifically the folktale and the myth, shedding new light on Levi-Strauss's structuralism with a notion of transvaluation based on C.S. Peirce's concept of the interpretant: transvaluation is an interpretant which revaluates the markedness of a sign. Liszka and others have also investigated facial expression in terms of marked and unmarked oppositions (see Carpenter et al., 2011 for a recent study).

Against Markedness

As is evident, the more widely markedness has been used, the more difficult it is to treat as a coherent, unified notion. There has been an inevitable semantic drift as markedness has been applied to different grammatical domains (semantics, phonology, syntax), theoretical approaches (generative linguistics, functional linguistics), and extralinguistic concerns (code switching, cultural analysis). Does markedness remain a coherent notion? And can it be reduced to other concepts? Martin Haspelmath (2006), for example, distinguishes a dozen senses of markedness, which he groups into four larger classes: markedness as complexity, markedness as difficulty, markedness as abnormality, and markedness as a multidimensional correlation among various diagnostics. He argues that the term 'markedness' is superfluous, and can be replaced by less ambiguous concepts such as frequency asymmetries, phonetic difficulty, and pragmatic inferences.

I tend to prefer the view of markedness as a global semiotic principle, a means of exploring the parallels among the various markednesses. However, there is no denying that markedness has been extended and reshaped over the past century and reflects a range of loosely connected theoretical approaches for naturalness, language universals, defaults, and preferences in language acquisition and use. What connects various approaches is a concern for the evaluation of linguistic structure, though the details of how markedness is determined and what its implications and diagnostics are vary widely, as we have seen. (*See* Generative Grammar; Semiotics; Prototype Theory in Linguistics. For book-length discussions of markedness see Battistella, 1990, 1996; Andrews, 1990; de Lacy, 2006; Hume, to appear.)

See also: Generative Grammar; Iconicity in Linguistics; Prototype Theory in Linguistics; Semiotics.

Bibliography

Andersen, H., 1989. Markedness – the first 150 years. In: Tomic (Ed.), Markedness in Synchrony and Diachrony. Mouton de Gruyter, Berlin.

Andrews, E., 1990. Markedness Theory: The Union of Asymmetry and Semiosis in Language. Duke University Press, Durham, NC.

Archangeli, D., 1997. Optimality theory: an introduction to linguistics in the 1990s. In: Archangeli, D., Langendoen, D.T. (Eds.), Optimality Theory: An Overview. Blackwell, Malden, MA.

Battistella, E., 1990. Markedness: The Evaluative Superstructure of Language. SUNY Press, Albany, NY.
Battistella, E., 1996. The Logic of Markedness. Oxford, NY.
Broselow, E., Chen, S.-I., Wang, C., 1998. The emergence of the unmarked in second language phonology. Studies in Second Language Acquisition 20, 261–280.
Bybee, J., 2010. Markedness: iconicity, economy, and frequency. In: Song, J.J. (Ed.), The Oxford Handbook of Linguistic Typology. Oxford, NY, pp. 131–147.
Calabrese, A., 1995. A constraint-based theory of phonological markedness and simplification procedures. Linguistic Inquiry 26, 373–463.
Caramazza, A., Zurif, B., 1978. Language Acquisition and Language Breakdown. The Johns Hopkins University Press, Baltimore, MD.
Carpenter, M., Liebal, K., Tomasello, M., 2011. Young children's understanding of markedness in non-verbal communication. Journal of Child Language 38 (4), 888–903.
Chomsky, N., 1977. On Wh-Movement. In: Culicover, P., Wasow, T., Akmajian, A. (Eds.), Formal Syntax. Academic Press, NY.
Chomsky, N., 1986. Knowledge of Language: Its Nature, Origin and Use. Praeger, NY.
Chomsky, N., Halle, M., 1968. The Sound Pattern of English. Harper and Row, NY.
Chomsky, N., Lasnik, H., 1977. Filters and control. Linguistic Inquiry 8.3, 425–504.
Croft, W., 1990. Typology and Universals. Cambridge University Press, Cambridge.
Comrie, B., 1983. Switch-Reference in Huichol: A Typological Study. In: Haiman, J., Munro, P. (Eds.), Switch-Reference and Universal Grammar. John Benjamins, B. V., Amsterdam, pp. 17–37.
Eckman, F., 2004. Optimality Theory, markedness, and second language syntax: The case of resumptive pronouns in relative clauses. Studies in Phonetics, Phonology, Morphology 10, 89–110.
Givón, T., 1990. Syntax: A Functional-Typological Introduction, vol. 2. John Benjamins, Amsterdam.
Greenberg, J., 1966. Language Universals. Mouton, The Hague.
Gregg, K., 1993. Second language acquisition: history and theory. In: Asher, R.E. (Ed.-in-Chief), Encyclopedia of Language and Linguistics, Pergamon Press, Oxford, 3720–3726.
Haspelmath, M., 2006. Against markedness (and what to replace it with). Journal of Linguistics 42, 25–70.
Hatten, R., 1994. Musical Meaning in Beethoven: Markedness, Correlation, and Interpretation. Indiana University Press, Bloomington, IN.
Hatten, R., 2004. Interpreting Musical Gestures, Topics, and Tropes: Mozart, Beethoven, Schubert. Indiana University Press, Bloomington, IN.
Hume, E., 2005. Deconstructing markedness: a predictability-based approach. In: Berkeley, CA, USA: Annual Meeting of the Berkeley Linguistics Society, 13–16 Feb 2004. In Proceedings, pp. 182–198.
Hume, E., 2011. Markedness. In: van Oostendorp, M., Ewen, C.J., Hume, E., Rice, K. (Eds.), The Blackwell Companion to Phonology, vol. 1. Wiley-Blackwell, Malden, pp. 79–106.
Hume, E. Phonological Markedness. Key Topics in Phonology Series. Cambridge UP, Cambridge, to appear.
Jakobson, R., 1984. The structure of the russian verb. In: Waugh, L., Halle, N. (Eds.), Russian and Slavic Grammar Studies, 1931–1981. Mouton, Berlin (1932, Zur Struktur des russischen Verbums, Charisteria Gvilelmo Mathesio gvinqvaqenario a discipulis et Circuli Lingvistici Pragensis sodalibus oblata, Prague).
Jakobson, R., 1968. Child Language, Aphasia and Phonological Universals. Mouton, The Hague (1941, Kindersprache, Aphasie und aligemeine Lautgesetze, Almquist & Wiksell, Uppsala).
Jakobson, R., 1972. Verbal communication. Scientific American 227, 72–80.
Kean, M.-L., 1980. The Theory of Markedness in Generative Grammar. Indiana University Linguistics Club, Bloomington, IN.
Keenan, E., Comrie, B., 1977. Noun Phrase Accessibility and Universal Grammar. Linguistic Inquiry 8, 63–99.
de Lacy, P., 2002. The formal expression of markedness (Doctoral dissertation), UMass Rutgers Optimality Archive 542, 28.
de Lacy, P., 2006. Markedness: Reduction and Preservation in Phonology. Cambridge University Press, Cambridge.
Liszka, J., 1989. The Semiotic of Myth. Indiana University Press, Bloomington.
Mayerthaler, W., 1988. Morphological Naturalness. Karoma, Ann Arbor, MI (1981 Morphologische Natürlichkeit, Akademische Verlagsgesellschaft Athenaion, Wiesbaden).
McCarthy, J., 2003. Comparative markedness. Theoretical Linguistics 29, 1–51.
McCawley, J., 1985. Kuhnian paradigms as systems of markedness conventions. In: Makkai, A., Melby, A. (Eds.), Linguistics and Philosophy: Studies in Honor of Rulon S. Wells. John Benjamins, Amsterdam.
Myers Scotton, C., 1983. The negotiation of identities in conversation: a theory of markedness and code choice. International Journal of the Sociology of Language 44, 115–136.
Myers-Scotton, C., 1993. Social Motivations for Codeswitching: Evidence from Africa. Clarendon Press, Oxford.
Myers-Scotton, C. (Ed.), 1998. Codes and Consequences: Choosing Linguistic Varieties. Oxford, NY.
Pustet, R., 2009. On the correlation between discourse frequency and structural complexity in markedness theory. Discourse [on line] 5.
Rice, K., 2007. Markedness in phonology. In: de Lacy, P. (Ed.), The Cambridge Handbook of Phonology. Cambridge University Press, Cambridge, pp. 79–97.
Rodman, R., 2010. Tuning in: American Narrative Television Music. Oxford, NY.
Rutherford, W., 1982. Markedness in Second Language Acquisition. Language Learning 32.1, 85–108.
Smolensky, P., 2006. Optimality in phonology II: harmonic completeness, local constraint conjunction, and feature-domain markedness. In: Smolensky, P., Legendre, G. (Eds.), The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar, vol. 2. MIT Press, pp. 590–716, 27–160.
Waugh, L., 1982. Marked and unmarked: a choice between unequals in semiotic structure. Semiotica 38, 299–318.
Waugh, L., Rudy, S. (Eds.), 1991. New Vistas in Grammar: Invariance and Variation. John Benjamins, Amsterdam.
Woolford, E., 2007. Introduction to OT syntax. Phonological Studies 10, 119–134.

Market and Nonmarket Allocation

Debra Satz, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

While markets are pervasive forms of social organization, there are many goods and services that we do not distribute through markets. For example, people buy, sell, and trade cars and shirts, but market exchanges of votes, babies, and legal justice are widely condemned. The market status of other goods – such as health care, human kidneys, and sex – is contested. What reasons can be given in favor of using and refraining from using markets? Many debates in the twentieth century have centered on this question. Typically, the debates over when and where a market is appropriate have focused on efficiency. This article will review the myriad reasons to use or refrain from using markets – including not only reasons of efficiency but also distributive justice and the effects of markets on democratic institutions and culture. This article also examines ways of enriching the list of distributive alternatives beyond the two poles of centralized plan and market such as gifts and lotteries.

While markets are pervasive forms of social organization, there are many goods and services that we do not distribute through markets. For example, people buy, sell, and trade cars and shirts, but market exchanges of votes, babies, and legal justice are widely condemned. The market status of other goods – such as health care, human kidneys, and sex – is contested. What reasons can be given in favor of using and refraining from using markets? Many debates in the twentieth century have centered on this question. Typically, the debates over when and where a market is appropriate have focused on efficiency. This article will review the myriad reasons to use or refrain from using markets – including not only reasons of efficiency but also distributive justice and the effects of markets on democratic institutions and culture. This article also examines ways of enriching the list of distributive alternatives beyond the two poles of centralized plan and market such as gifts and lotteries.

What Is a Market?

Modern economic textbooks rarely offer a definition of what precisely constitutes a market. The Shorter Oxford English Dictionary defines a market as "a meeting or gathering place for the purchase and sale of provisions or livestock" or as the "action or business of buying and selling." However, markets are not merely places, nor are they best understood in terms of episodic individual transactions. Functioning markets actually require many conditions that need to be in place. For example, there must be enough exchanges so that people can use price signals to coordinate their behavior, there must be background property rules, there must be mechanisms for information to flow, and there must be the curtailment of monopolies (McMillan, 2002). Many of the conditions that working markets require are dependent on state action. For this reason, it is a mistake to see market and state as simple contrasts.

Markets are mechanisms used for the production, circulation, and valuation of many goods and services. Their desirability rests on three of their noteworthy features: they are efficient, impersonal, and oriented to individual preference and choice. Let us consider these features in more detail.

Since rational individuals will only exchange goods and services when they have something to gain, markets will be efficient – purging the economy of the less desirable goods and moving the trading parties to preferred positions. Economists refer to this property of markets in terms of 'Pareto optimality.' A state of affairs is Pareto optimal when no one can be made better off (i.e., their utility increased) without someone else being made worse off (i.e., their utility decreased). An important result in economics, the so-called 'fundamental theorem of welfare economics,' shows that in a world in which everyone could trade everything – including futures, uncertainty, and so on – the allocation of resources would be Pareto optimal. This theorem formalizes the earlier conjectures of Adam Smith and others, that participation in competitive markets would lead rational self-interested individuals to produce, as an unintended consequence of their private actions, collectively good results.

Markets are also impersonal, suitable to regulating the interactions of strangers. They abstract away from particular features of a participant, such as his race, religion, or sexual orientation and focus only on his assets – what is being offered to buy or sell. The impersonality of market relations means that each individual is indifferent to his trading partners at any time. He has no precontractual obligations to those he trades with and is free to consider only his own interests in pursuing an exchange.

In market exchange, the parties also express their own individual preferences, pursuing their own sense of what is good, independent of what the state or others value. Markets are thereby conducive to individual freedom: they accommodate individual preferences without passing judgment upon them. By decentralizing decision making, markets give individuals a significant measure of freedom over their circumstances. Furthermore, one is free to walk away from any exchange – the same deal can be struck elsewhere, as it were. Albert Hirschman (1970) has termed this latter feature the power of 'exit,' to distinguish it from the power of 'voice' (i.e., speaking up).

A Contrast

The modern, neoclassical conception of economics, developed in the late nineteenth and early twentieth centuries by

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63048-9

Edgeworth, Robbins, and Jevons, treats the market as something obvious and simple. In the standard neoclassical general equilibrium model of a market economy formulated by Leon Walras (1926/1954), commodities are identical, the market is concentrated at a single point in space, individuals have complete information about the exchange commodity, and exchange is instantaneous. There are, accordingly, no costs in enforcing agreements and no important distinctions among kinds of markets.

The neoclassical treatment of markets has produced important mathematical results such as the fundamental theorem, and also allowed for the development of powerful models that can characterize the equilibrium properties of an entire economy. However, it depends on viewing markets as economic mechanisms whose operation can be understood in abstraction from the particular kinds of goods being exchanged and the relationships between the participants. In the neoclassical conception of markets, it does not matter whether what is being traded is oranges, sexual services, or credit – the nature of market exchange is seen as the same in each case. The participants are considered interchangeable, and, given the opportunities each has for exit, it is assumed that no party can exercise significant power over another. In Paul Samuelson's words, "in a perfectly competitive market, it does not matter who hires whom: so let labor hire 'capital'" (Samuelson, 1957, p. 894). The result is that the market is viewed as a private realm, whose smooth operation leaves no room for political intervention: intervention is justified only when markets in some goods fail to be established or are inefficient.

By contrast, the classical economists – among them Adam Smith, David Ricardo, and Karl Marx – offered distinct theories of the operation of markets in different domains and focused on the opposing interests of the three main social classes – landlords, capitalists, and workers – who came together to trade and between whom the produced social wealth was to be distributed. Not only were markets viewed as political, in the sense that they depended on legal property rights, but it was also recognized that their functioning in certain realms inevitably raised questions relevant to the structure of public life. For example, the classical economists noted that the market between buyers and sellers of goods such as oranges differs importantly from the market between buyers and sellers of human labor power. This is because the latter kind of market involves a good that has the highly unusual status of being embodied in a human being. While we do not have to worry about the noneconomic effects of a market on fruits, we do need to worry about whether a particular market in human labor power produces passivity, alienation, greed, or indifference to others.

The classical economists saw the marketplace as a political and cultural as well as an economic institution, where differing interests wielded power and where some types of exchanges had constitutive effects on their participants. Adam Smith (1776/1937) worried not only about the tendency of the owners of capital to collude and thus block market competition but also about the effects of a fragmented and monotonous division of labor on the ability of workers to participate in political life.

Criteria for Assessing the Use of Markets

The most famous modern debates about the legitimacy of market allocation versus its alternatives have focused on its efficiency. Consider, for example, the debate between socialist economist Oskar Lange and conservative economist Friedrich Hayek, which took place during the 1930s, when the Soviet experiment was underway. Lange and Hayek debated whether a noncapitalist economy with few market structures could experience economic growth. Lange argued that an ideal planned economy could allocate resources in a more optimal manner than a perfectly competitive market economy. In rebuttal, Hayek pointed out that planned economies have a tendency toward bureaucratic sclerosis, a lack of accountability and high information processing costs.

Each side raised important considerations. Each side had a point. However, from many perspectives their debate about markets can be criticized as narrow and conceptually impoverished. The debate not only tended to operate as if there were just two choices – command and control centralized planning and the system of unregulated markets – but also posed the issue about market or nonmarket choice solely in terms of efficiency.

If we see the market as a complex heterogeneous institution, which brings people into different kinds of relations with each other for different exchanges, then we can also see why the use of markets can raise other concerns besides efficiency. From this broader perspective, not only efficiency but also considerations of fairness as well as the effects of the market on the political structure of power and on human well-being are relevant to determining when and where markets should allocate goods and services and where other institutions should be used.

Efficiency

Through the use of prices, markets indicate what millions of goods and services are worth to sellers and buyers. They thereby function to apportion resources efficiently: they signal to sellers what to produce, to consumers what to buy, and to capitalists where to invest their wealth. The continual adjustments of supply and demand, registered in changing prices, allow markets to 'clear' inventory. When inventory is cleared, there is no excess demand or supply: supply equals demand at some price. The efficiency benefits that markets produce constitutes an important prima facie case for their inclusion in an economy. Furthermore, at least to date, we know of no mechanism for inducing innovation on an economy-wide basis that is as successful as market competition.

In certain circumstances, however, markets can fail to be efficient. While in the model of an ideal market, the transaction costs involved in exchanges are taken to be zero, and each participant is assumed to have perfect information and acts solely on the basis of his or her rational self-interest, it is well recognized that these assumptions do not always hold in reality. Thus, the standard theorems connecting markets and efficiency are usually coupled with the identification of 'market failures,' circumstances in which the efficiency effects of markets are either inoperable or eroded. For example, markets fail when and where certain effects of an exchange – which economists call 'externalities' – are not fully absorbed by the exchanging parties

and are passed on to third parties. In some circumstances, these external costs constitute public bads: the defacement of the landscape, pollution, and urban decay are notable examples. Markets also fail where there are natural monopolies, or technologies that give rise to economies of scale, thus making only monopolistic or oligopolistic firms viable.

In the last several decades, work within mainstream economics has shown that market failure may be more widespread than previously supposed. This work emphasizes the ubiquity of imperfect and asymmetric information held by the exchanging parties (Akerlof, 1970) as well as the fact that in many exchanges the parties cannot easily control one another's behavior. (Consider the costs of supervision that employers pay to make sure that their workers are actually working.) Economic models now take into account the incentives agents have to manipulate information and outcomes in order to press their own advantages. Thus, agents may act as if they were willing to pay less for goods and services than they really are, or threaten to withdraw from a trade entirely. It is now recognized that people can act in ways that prevent markets from clearing even in competitive equilibrium – there will be persistent excess supply or excess demand. This is because if one party wants to have power over the other, it can achieve this by giving the other something to lose, so that it now costs something to 'exit.' Thus, because not all prospective borrowers can find loans and not all those who seek jobs can find them, those who do have a strong interest in complying with, respectively, the banks or owners (Stiglitz and Weiss, 1981). The power that banks thus acquire over their creditors and owners acquire over their workers gives these markets a political character. One might say that this recent work emphasizing strategic action, incomplete information, and transaction costs gives mathematical precision to some of the earlier insights of the classical economists (Bowles and Gintis, 1993).

In an important sense, externalities are everywhere. Almost any exchange in an interdependent society is likely to impose a cost on third parties. High-rise apartment towers block the sun of neighboring houses, cars bring congestion, and noise circulates. We need some guidance, then, as to which externalities are serious enough to merit intervention in a market. This is a question for ethics and politics, not economics.

Various solutions to market failure have been proposed including state intervention in the marketplace, expansion in the number of markets (e.g., allowing externalities themselves to be priced by the market), and the use of nonmarket alternatives such as bargaining and negotiation. The point of such solutions is to restore efficiency.

In thinking about the ways that markets contribute to efficiency, we should note that the actual efficiency outcomes of the market depend on the initial distribution of endowments and assets. This initial distribution is not itself part of the definition of Pareto optimality. A state of the world can be Pareto optimal, even if its original distribution was based on theft and fraud or was otherwise unfair. In order to justify market outcomes, or to use Pareto optimality normatively, the theory of market efficiency needs to be supplemented with a theory of distributive justice. Some theories of distributive justice focus on the legitimacy of the initial distribution of assets while others focus on the legitimacy of the distributive inequalities that the market produces.

A related point is that efficiency interpreted as Pareto optimality has only modest moral content. This is true for several reasons. First, Pareto optimality deals exclusively with efficiency and pays no attention to distribution. A Pareto optimal state can thus be a state of social misery for some. Simply consider the case in which the utility of the downtrodden cannot be raised without lowering the utility of the millionaires. Second, the Pareto criteria dispenses with interpersonal comparisons, so that we cannot compare the contribution a meal makes to the poor man's utility with that it makes to a millionaire's (Sen, 1987).

These considerations bring home the point that efficiency is not the only value we have reason to care about. Even if certain markets are efficient, we may sometimes have good reason to block them or to further regulate them. The fact that a market in political votes might be efficient does not suffice to justify such a market.

Freedom

Market allocation is often defended in terms of freedom. Proponents cite a range of important effects that markets have on an individual's ability to develop and exercise the capacity for liberal freedom. Each of these effects, they argue, depends on leaving the market domain largely cordoned off from political interference. Markets (1) present agents with the opportunity to choose between a large array of alternatives, (2) provide incentives for agents to anticipate the results of their choices and thus foster instrumental rationality, (3) decentralize decision making, giving to agents alone the power to buy and sell things and services without requiring them to ask anyone else's permission or take anyone else's values into account, (4) place limits on the viability of coercive social relationships by providing (relatively) unimpeded avenues for exit, (5) decentralize information, (6) (may) enhance individuals' sense of responsibility for their choices and preferences, (7) allow people to practice and try out various alternatives, (8) create the material wealth that is a precondition for the possibility of having decent alternatives, and (9) allow liberal individuals to cooperate without having to agree on a large range of questions of value (Friedman, 1962; Nozick, 1974).

Liberal theories that assign substantial weight to individual freedom thus allot a central role for market allocation, pointing to the market realm as a place where the capacity for individual choice – indeed, where the liberal individual – is developed. As liberal theorists are likely to emphasize, respect for markets in goods and services can be an important way of respecting individual (and divergent) conceptions of value. In a market system, there is no preordained pattern of value to which individuals must conform and exchange gives to individuals the freedom to pursue distinct aims. In fact, the market leaves people free not to choose to sell or buy a particular good on the market.

The liberal theory of freedom is essentially (although not entirely) a negative one, emphasizing the space markets provide to individuals to pursue their private ends free from external intrusions by the state or other people. This liberal theory linking freedom and markets has been criticized for not attending to the preconditions for its own realization. These

preconditions include the existence of some realms that are protected from the market (Polanyi, 1971). For example, most liberals recognize that it is incompatible with liberal freedom to allow a person to sell herself into slavery; free people require rights to consultation, self-judgment, and control over the conditions in which they act. The formation of a market agent also requires education: there is no 'infans economicus' who is born with the developed ability to choose in a market. Some theorists have further argued that the securing of liberal freedoms requires the guaranteed provision of a minimal level of basic goods such as health care, food, and housing and may therefore require the regulation or blocking of certain markets (Satz, 2010).

Human Flourishing

Market allocation has often been defended on the grounds that markets, by satisfying people's preferences, contribute directly to human happiness and flourishing. In addition, by stimulating economic growth, markets have the potential to eradicate the poverty and hardship that is everywhere the enemy of a decent quality of life. Certainly, the market makes important contributions to the satisfaction of human wants. However, in considering the effectiveness of the market in maximizing overall human well-being, several considerations need to be taken into account. First, satisfaction of the subjective preferences of an individual agent may not provide that agent with well-being. A lot depends on the content and nature of his or her preferences. Agents may have preferences for things that do not contribute to their overall good, preferences that are mere accommodations to their circumstances, preferences they themselves wish they did not have, preferences based on false information, and preferences that are instilled in them by the market. In some such examples, the case for regulation can be made on paternalistic grounds, excluding exchanges born of desperation (such as selling oneself into slavery), lack of information, and imprudence (Sunstein, 1991).

Second, the market recognizes only those preferences that can be backed up by an ability to pay. The poor man's preferences for a good meal are invisible. Indeed, the notion of 'preference' itself does not distinguish between kinds of preferences, only degrees of strength. Thus, the market cannot distinguish between urgent human needs (the poor man's desire for nourishment) and mere whims (the rich man's desire for a fine burgundy), but urgent needs are more important to meet than mere wants (Scanlon, 1975). Third, there are important sources of satisfaction that do not go through the market: public goods that, by their nature, do not go through the market and interpersonal relations that may be spoiled if they become part of the market's transactional network. So care must be taken to avoid ignoring those preferences that are not expressed on the market.

Fourth, empirical studies point out that the contribution of material wealth to a person's well-being tapers off above a certain level of income. Money correlates with happiness, but the correlation diminishes with additional income.

Some critics have claimed that markets promote an inferior form of human flourishing. This criticism has a wide interpretation and a narrow interpretation. On the wide interpretation, markets everywhere shape and socialize people in the wrong way. They are taken to do this by promoting selfish behavior, making people more materialistic, and dulling people to important distinctions of value to which they should be responsive.

This worry about markets is sometimes posed in terms of the metaphor of infection – that market norms and relations will spill over and contaminate nonmarket realms such as friendship and love. Thus, it has been alleged that markets erode our appreciation of the true value of other people, since they lead us to think of goods and people as exchangeable items. This wide interpretation of the market's negative effects on human flourishing has only weak social scientific support. There is little evidence that people are greedier in market societies than they were in peasant economies, that they devalue love and friendship, or that they are now less likely to engage in moral behavior than in the past (Lane, 1991).

A narrower interpretation of this humanistic criticism is that some (but not all) markets have bad feedback effects on particular human behaviors. Studies have shown, for example, that economics and business students (who presumably are more likely to govern their behavior by the axioms of neoclassical economics than are literature or sociology students) are uniquely uncooperative in solving collective action problems (Marwell and Ames, 1981). Other studies have pointed to similar 'crowding out effects' – cases in which the appeal to a market incentive crowds out the other reasons a person has for performing the action. Bruno Frey and his colleagues found that people were less willing to host a noxious waste facility after monetary compensation was offered than they were without the money. In some cases, when individuals are civically minded, using price incentives will backfire and not increase supply. Crowding out therefore has implications for efficiency as well as for human character and flourishing (Frey and Oberholzer-Gee, 1997).

Richard Titmuss (1971) had earlier claimed that allowing blood to be sold and bought as a commodity would drive out altruistic blood donation by turning the 'gift of life' into the monetary equivalent of 50 dollars. People would become less willing to give blood freely as market distribution became more prevalent, for their gift loses its benevolent meaning. (Blood quality would also be lower since altruistic donors have no incentive to lie about the quality of their blood, while commercial donors clearly do.)

Titmuss aimed to show that markets have broader incentive effects than economists had previously supposed. His account also shows that markets can undermine flourishing if they lead us to fail to act from reasons we would prefer. If this is so, then sometimes market regulation – including prohibition of certain exchanges – may even shore up liberal freedom by allowing us to get the choice that we most want. (Consider the role of minimum wage laws and prohibitions on vote selling from shrinking people's opportunity sets over time.)

The market's negative feedback on motivation was the concern of diverse thinkers such as Adam Smith, Karl Marx (1977), and John Stuart Mill (1970), who each worried that labor markets that rendered workers as mere appendages to machines would cripple their capacity for independence and political participation. Each thinker noticed that labor market exchanges had a constitutive effect on the parties to the exchange: workers who engaged in servile, menial, and degrading work were likely to be servile citizens.

Where markets have significant feedback effects on human motivation, there is an additional possible case for social regulation. It is circular to justify markets on the grounds that they maximally satisfy individual preferences in cases where those preferences are themselves shaped by the market. At the very least, recognition of the endogenous nature of some preferences leaves the question of institutional design open since there may be other mechanisms that generate and satisfy more preferences.

The Nature of the Goods Exchanged

Liberals have traditionally argued that many of the problems that unregulated markets raise with respect to the values of efficiency, justice, freedom, and well-being can be attenuated through state intervention. We can redistribute income and regulate consumption through taxation, publicly provide for social goods such as schools and roads, and establish a minimum level of health care for all citizens. Other social institutions that, in turn, encourage values such as honesty, reciprocity, and trust can be called upon to support markets.

There is a different criticism of the market that cannot be addressed through market regulation but requires that the use of the market be blocked. This criticism focuses not on the negative consequences of market exchanges but instead on the ways that markets undermine the intrinsic nature of certain goods. The theorists who make this criticism reject the use of markets in certain domains and for certain goods categorically.

Markets are often taken to be neutral means of exchange. By this proponents mean that the market does not distinguish between my ethical valuation of a good and yours. I may value my Bible infinitely while you assign it a dollar price, but the market is compatible with both of our understandings of this good. The market responds only to effective demand, not to the reasons people have for wanting things. When the market puts a price on a thing, it leaves the ways that people value it untouched.

In reality, this view of market neutrality is overstated. Some goods are changed or destroyed by being put up for sale. The most obvious examples of this phenomenon are love or friendship. A person who thought that they could buy my friendship would simply not know what it means to be a friend. A proposal to buy love, prizes, honors, friends, or divine grace is conceptually incoherent: it is the nature of these things that they cannot be bought.

Various arguments have been given to show that many goods are not properly conceived of as market commodities. Some of these arguments depend on the idea that certain goods are degraded and corrupted when sold, or that some sales are inherently alienating to, or exploitative of, the seller (Anderson, 1993; Sandel, 2012; Walzer, 1983). Candidate goods include sex, reproductive labor, adoption, military service, health care, prisons, and political goods such as votes and influence.

Consider sex and reproductive labor. Some feminists have argued that markets are inappropriate for exchanging goods that place people in personal relations with each other. They argue that when sexual relations are governed by the market, such as in prostitution, not only is an inferior good produced since commercial sex is not the same as sex given out of love but also the dignity and freedom of the prostitute is compromised. Because practices such as prostitution are alleged to allow women's bodies to be controlled by others, these arguments conclude that women's freedom and dignity requires that their sexual and reproductive capacities remain market-inalienable (Radin, 1996).

Given the relatively poor economic opportunities open to many women, one may doubt whether paid pregnancy or prostitution is comparatively worse for women's freedom than the alternatives. For example, is prostitution more constraining than extremely low-paid, monotonous assembly-line work? Does commercial surrogacy really place more extensive control over a woman's body than other acceptable labor contracts such as those given to professional athletes? To answer these questions may be more a matter of looking at the empirical consequences of markets in sex and reproduction, and the history of women's roles in the economy and the family, than of examining the intrinsic nature of these goods. The charge that prostitution undermines dignity is different. We cannot ascertain whether a person's dignity is compromised simply by looking at her options.

Of course, even if the corruption argument is true with respect to a certain good, that does not by itself tell us whether or not we want to block markets in that good. In a liberal society we give people the right to make inferior choices, provided that they do not do harm to others.

Democracy

A special case is often made for regulating or banning markets so as to ensure that democratic institutions are supported. This case has several components: (1) a theory of those political goods that must be provided to everyone equally in a democracy – goods such as votes, political rights, and liberties; (2) a theory of the economic preconditions of democracy; and (3) a theory of the kind of institutions that foster the development of people likely to support democratic institutions and function effectively in a democratic environment. It is easy to see that democracy is incompatible with property qualifications on the right to vote, slavery, and vote selling. Current debates center on whether democracy requires greater regulation of the political process including public funding of elections and campaigns, whether democracy requires greater regulation of the economy to allow for 'voice,' and whether expanding the market in education would enhance or inhibit the development of democratic citizens.

Regulation vs. Alternatives

Arguments that purport to show that markets in some good do not promote freedom, happiness, well-being, or some other value do not directly enable us to conclude that such exchanges should be blocked. Even if markets interfered with or failed to promote certain values, interference with them might be worse overall from the point of view of those same values. Suppose, for example, that the only feasible institutional alternatives to markets all involved significant amounts of coercion, or suppose that a legal ban on a market simply drove the exchanges underground in a thriving black market in which the participants suffered additional harms (Satz, 2010). An

example of this effect can be seen in the case of the legal ban on addictive drugs such as heroin and cocaine. Conservatives frequently cite the costs of 'overregulated' markets such as rent control and limits on parental choice with respect to education. Therefore, a full assessment of market allocation must take into account the alternatives and the likely consequences of different interventions.

Other Forms of Market Regulation

Most egalitarians endorse the use of markets combined with redistributive taxation to redress the inequalities that market processes often produce (Dworkin, 2000). In this way the market's efficiency is allowed to function, and then transfers occur after the fact. (Of course, taxation itself has effects on the market's efficiency, as rational agents will change their behavior to avoid or diminish their own payments.) The important egalitarian goal is to ensure that the operation of markets do not undermine society's commitment to treat all of its citizens fairly.

Other forms of market regulation besides ex post tax and transfer mechanisms have been suggested. These include ex-ante restrictions on the transferability of income and wealth, as in ownership requirements that stipulate that owners must be active participants in the community in which their property is located, and restrictions as to what people can do to the things they can sell, as in historical building codes.

Nonmarket Allocation

There are important alternative mechanisms for distribution that do not rely on a market. In the twentieth century, the most important alternative allocation mechanism has been government provision. Government decision has allocated such goods to individuals as citizenship; social security; infrastructure such as roads, highways, and bridges; and education up to early adulthood.

Many allocative decisions are shaped neither by the government nor by the market. These include distribution through gift, lottery, merit, the intrafamily regulation of work and income, and other principles such as seniority and need. Consider a few cases.

In the United States, the lottery has been used to allocate resources literally bearing on issues of life and death. Military service, especially during wartime, has often been selected for by lottery; the alternative has been to allow people to volunteer. Lotteries have also been used to distribute licenses for low-power television and radio stations by the Federal Communications Commission, determine access to immigration, and distribute admission to magnet and charter schools.

College admission, election to professional associations, and the award of prizes have been governed neither by market norms nor by state decree but by selective principles of merit and desert. Need has been central in allocating organs for transplantation.

In each of these cases, one can compare the allocation achieved with that which would likely have been achieved through a market, in terms of the broad list of criteria specified above.

See also: Hierarchies and Markets; Market Structure and Performance.

Bibliography

Akerlof, G.A., 1970. The market for lemons: quality uncertainty and the market mechanism. Quarterly Journal of Economics 84, 488–500.
Anderson, E., 1993. Value in Ethics and Economics. Harvard University Press, Cambridge, MA.
Bowles, S., Gintis, H., 1993. The revenge of homo economicus: contested exchange and the revival of political economy. Journal of Economic Perspectives 7 (1), 83–102.
Dworkin, R., 2000. Sovereign Virtue. Harvard University Press, Cambridge, MA.
Frey, B., Oberholzer-Gee, F., 1997. The cost of price incentives: an empirical analysis of motivation crowding out. American Economic Review 87 (4), 746–755.
Friedman, M., 1962. Capitalism and Freedom. Chicago University Press, Chicago.
Hirschman, A.O., 1970. Exit, Voice and Loyalty: Responses to Decline in Firms, Organizations and States. Harvard University Press, Cambridge, MA.
Lane, R.E., 1991. The Market Experience. Cambridge University Press, Cambridge, UK.
Marwell, G., Ames, R., 1981. Economists free ride. Does anyone else? Experiments on the provision of public goods IV. Journal of Public Economics 15 (3), 295–310.
Marx, K., 1977. Capital, vol. 1. Vintage Books, New York.
McMillan, J., 2002. Reinventing the Bazaar: A Natural History of Markets. Norton, New York.
Mill, J.S., 1970. Principles of Political Economy. Penguin Books, London.
Nozick, R., 1974. Anarchy, State and Utopia. Blackwell, Oxford, UK.
Polanyi, K., 1971. The Great Transformation. Beacon Press, Boston.
Radin, M.J., 1996. Contested Commodities. Harvard University Press, Cambridge, MA.
Samuelson, P., 1957. Wages and interests: a dissection of Marxian economics. American Economic Review 47, 884–912.
Sandel, M., 2012. What Money Can't Buy: The Moral Limits of Markets. Farrar, Straus and Giroux, New York.
Satz, D., 2010. Why Some Things Should Not Be for Sale: The Moral Limits of Markets. Oxford University Press, Oxford, USA.
Scanlon, T., 1975. Preference and urgency. Journal of Philosophy 72 (19), 655–669.
Sen, A., 1987. On Ethics and Economics. Blackwell, Oxford, UK.
Smith, A., 1776/1937. The Wealth of Nations. Modern Library, New York.
Stiglitz, J., Weiss, A., 1981. Credit rationing in markets with imperfect information. American Economic Review 71 (June), 393–410.
Sunstein, C., 1991. Politics and preferences. Philosophy and Public Affairs 20, 3–34.
Titmuss, R., 1971. The Gift Relationship: From Human Blood to Social Policy. Pantheon Books, New York.
Walrus, L., 1926/1954. In: Jaffe, W. (Ed.), Elements of Pure Economics. Richard D Irwin, Homewood, IL.
Walzer, M., 1983. Spheres of Justice. Basic Books, New York.

Market Geographies

Jean-Claude Thill, University of North Carolina at Charlotte, Charlotte, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A commonly accepted definition of the concept of *market area* is that of the territory (or section of the geographic space) over which the entirety or majority of sales of a given good or service supplier takes place. It is the territorial expression of the supplier's clientele. The market area is much more than an ex post map depiction of a purely aspatial microeconomic process that defines customer–supplier relationships. In his 1870s work, the German economist Launhardt is credited for unveiling the novel idea that suppliers' success to sell their output in a market economy is shaped by considerations of transportation cost and position of competing suppliers one relative to another in a geographic space. The theory of market areas that grows out of his work emphasizes demand-oriented processes in shaping the spatial organization of economic systems and population settlements. Launhardt's pioneering ideas gave life to several research traditions in the social sciences. One emphasizes microeconomic processes and it is manifested in the seminal work of Christaller, Lösch, Greenhut, and others. A second research orientation is dedicated to the geometric characterization of market areas under alternative spatial behavioral and structural contexts. Third, the prescriptive approach looks at market areas as they ought to be delineated according to some system optimization criteria. Finally, revolving around the implied delineation of market areas through the social physics principle of spatial interaction, the fourth tradition has become a pillar of modern business geographics and spatial business intelligence. This article traces the roots of the concept of market area, the development of the four traditions outlined above, and discusses the relevance of the concept to modern social and behavioral sciences.

Market Area as Territorial Expression

A market is the set of actual or potential customers of a given good or service supplier. These consumers may be segmented according to any of a number of sociodemographic dimensions, including their age, gender, lifestyle, as well as their location relative to the supplier's own geographic position, to the position of other suppliers, and to the position of other consumers. The territory (or section of the geographic space) over which the entirety or majority of sales of a given good or service supplier takes place constitutes this supplier's market area. The same concept is also known as 'trade area,' 'service area,' 'hinterland,' or 'catchment area' across various fields of social and behavioral sciences.

In modern market-oriented economic systems, in which production and consumption are uncoupled, consumption (demand) is rather ubiquitous in geographic space (at home, at work, on the road, etc.), while production and distribution functions (supply) are limited to a few locales. The incongruity in the spatial organization of demand and supply is mediated by the mobility of economic agents or the availability of delivery services: market places emerge when consumers are mobile; home delivery requires suppliers to reach out to their customers; and the age-old institution of periodic markets emerges in sociocultural environments where production and consumption agents are mobile. In all cases, the concept of market area serves to capture the notion of territorial expression of a supplier's clientele. It applies to retailers and service marketers as disparate as grocers and other retailers (Applebaum, 1966; Allaway et al., 1994), health care providers (Fortney et al., 1999; Frech et al., 2004), and library branches (Jue et al., 1999; Donnelly, 2014). Sometimes, access to the service is highly regulated, as in the case of some public services where jurisdictional considerations predetermine the shape and geographic extent of market areas (school districts, for instance) (Yeates, 1963; Dost, 1968).

Goods and service providers are known to seek common locations in geographic space so as to share externalities created locally by agglomeration or clustering. Towns throughout the world, many cities, spontaneous urban retail and business districts, even planned shopping centers, owe their existence to such agglomerative forces. Because it is commonplace for many or all businesses belonging to these geographic clusters to share the same consumer bases, the narrow definition of market area given in the previous paragraph can be extended from single providers of goods and service providers to spatial clusters of them.

Why firms locate where they do is the central question of location theory. An early school of thought pioneered by von Thünen approached this question as a least-cost problem. On the other hand, the market-area analysis school emphasizes demand-side considerations by framing the locational decision(s) of the firm as a search for site(s) that command the largest number of customers. It is a straightforward consequence that, for a given spatial distribution of identical consumers, the size and shape of a firm's market area are indicative of the success of the firm at that site. However, the same does not hold once it is recognized that all households within a market area do not have the same disposable income. In fact, market geographies are then better encapsulated as a reflection of income geographies than of sheer population counts. The market-area analysis school provides the economic underpinnings for a substantial theoretical body of knowledge on the processes of spatial organization of businesses in consumer-oriented industries (Greenhut, 1956), and of market places, retail districts, and of settlement centers into hierarchical systems (Christaller, 1933; Berry and Parr, 1988). The market-area approach continues to provide the tenets for applied research in retail location analysis (Ghosh and McLafferty, 1987), public facility planning (Thisse and Zoller, 1983; Church and Murray, 2009), and spatial and territorial planning (Von Böventer, 1965). Interestingly, it remains

 http://dx.doi.org/10.1016/B978-0-08-097086-8.72031-9

a fundamental frame of reference in an era where much information on marketed products and services is obtained through electronic means (including social media), and where much of the trade is carried out electronically (e-commerce).

Economic Underpinnings

It was remarked above that a firm's market area is the territorial expression of the firm's clientele. However, the market area is far more than an ex post map depiction of a purely aspatial microeconomic process that defines customer–supplier relationships. In his 1880s work, the German economist Launhardt (1882) is credited for unveiling the novel idea that a firm's success in selling its output in a market economy is shaped by considerations of transportation cost and of position of competing firms, one relative to another in the geographic space. In fact, it has only recently been established that the very same ideas were in fact articulated four decades earlier by the German economist Rau (1968) in a little known letter to his colleague Quintelet.

Let us consider initially the simplified possible case, that of a single firm supplying a single good to a population of identical consumers scattered over a certain geographic territory. If all consumers are assumed to have an identical demand curve that is elastic to the good's 'real price' (defined as the sale price plus transportation cost), individual demand drops with distance from the firm location. The monopolistic firm's market area extends outward to the circular perimeter where demand falls to zero. This distance is known as the range of the good. The market beyond this point remains unserved by the firm, a situation that may raise serious social equity concerns if basic needs of the population remain unmet as a result. For the sake of increasing its sales and reducing its unit production cost, the firm may decide to serve consumers outside the normal market area. It may accomplish this in several ways. One approach involves absorbing part of the real price applied to more distant consumers. Any spatial pricing system aimed at accomplishing this goal subsidizes remote consumers and discriminates against nearby buyers. Similar economic circumstances are commonplace when several firms at different locations compete for consumers. Since firms have more control over their nearby customers, they may seek to discriminate against them, whereas consumers at the margin of market areas may enjoy lower prices ensuing interfirm price competition.

When a second firm whose geographic position is different from that of the first one supplies the same good, the situation becomes more complex. Each firm commands sales in the same circular market area if they are located far enough from each other (i.e., over twice the range of the good). The two firms enjoy monopolistic profits that may entice new firms to start business. New competitors can be expected to locate in the interstitial space between monopolistic market areas so as to command the largest possible volume of business and avoid competition. This leads us to the alternative condition where firms find themselves close enough to each other to create competition at the edge of their respective monopolistic market areas. The introduction of a competitor establishes another type of boundary line where real prices are equal. The economic significance of this boundary is considerable because it delineates the market areas of competing firms, and consequently determines their sales and profits. Remarkably, as indicated by Isard (1956), this boundary is optimal and socially efficient. Working with Launhardt's (1882) assumption that firms do not charge the same sale price, the boundary line is a hyperbolic curve defined by the equation of the difference of sale prices to the difference of transportation costs to firm locations. The American economist Fetter, who independently discovered the same economic principle, coined the phrase 'economic law of market areas' in 1924.

Figure 1 illustrates the law of market areas under two different conditions. When firm B charges a sale price higher than that of firm A, firm A has a cost advantage over B that translates into a larger market area. The market boundary is hyperbolic and closer to B than to A. In the special case where firms charge identical sale prices, neither competitor has a special advantage. The two market areas are equal and are separated by a boundary line that is a degenerated hyperbola equidistant from the two firm locations.

As indicated above, free entry of firms affects the size and shape of market areas. Christaller (1933) and Lösch (1954) argued that free entry of new firms would compress the circular market areas defined by the range into hexagons forming a lattice covering the entire territory and that firms position themselves at even distances from each other. Geometrically speaking, these hexagonal market areas are proximal areas, also known as Thiessen polygons (or Voronoi diagrams), drawn by intersecting the perpendicular bisectors of lines connecting adjacent firm locations (Okabe et al., 2000). These geometric constructs are fundamental elements of geographic information science. When firm locations are not perfectly equidistant, Thiessen polygons are distorted, but their economic interpretation is unaffected. This happens to be a very general case in practice due to departures from simplifying theoretical assumptions, such as when population (and therefore demand) is not evenly distributed across the geographic space.

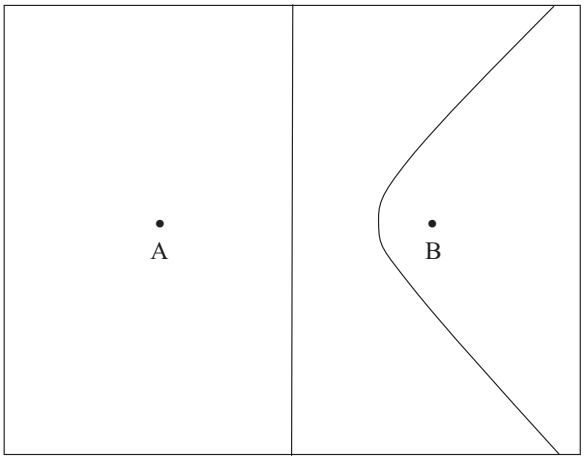

Figure 1 This diagram represents the market areas of two firms, A and B. When A and B charge equal sale price, market areas are separated by the vertical straight line. The hyperbolic curve separates market areas when B charges a price higher than A.

The theory of spatial organization advanced by Christaller and Lösch was also extended to encompass multiple goods – from the most to the least ubiquitous – to which corresponded multiple overlapping layers of hexagonal market areas. One area of deficiency of this theory was to underrate mutual interdependencies between markets for different goods and their territorial expression, the market areas. Demand-side linkages between markets are particularly meaningful to market-driven industries. One such linkage is created by consumers' effort to reduce the transactional cost of shopping by means of combining the purchase of several goods on the same shopping trip. Multipurpose shopping, as this practice is called, reduces the geographic dispersion of firms where these goods are purchased. Because of enlarged market areas, sites where a larger range of goods is marketed are consequently placed at an advantage over others.

Spatial Analysis

The previous section introduced the possibility of polygonal, hyperbolic, and circular market areas. The question arises now what shapes are possible and what conditions bring them about. Since market areas constitute a model of the geography of demand, knowledge of geometric and topologic properties of market areas (Hanjoul et al., 1988) may shed some light on various issues of interest to urban and regional scholars such as the life cycle of business districts, people's modality of usage of the urban space, or the quality of geographic accessibility to amenities (social services, health, and cultural facilities, for instance).

It turns out that market areas may take any of a large number of different shapes depending on how the perceived cost of traveling over space is related to the physical distance. For instance, the hyperbolic areas mentioned earlier are related to a linear transportation cost function. When transportation cost increases as the square root of distance (i.e., the cost of the extra mile traveled drops as the trip gets longer), market areas are elliptic. Circles arise when transportation costs follow the log-transform of distance. In contrast to hyperbolae, circular and elliptic market areas reveal that the firm charging the higher sale price manages to serve consumers in the direction opposite to its main competitor only if they are located rather close. In these conditions, distant consumers in fact have an economic advantage in bypassing the nearby firm to patronize a more distant but cheaper firm. This economic behavior stands in direct contradiction with common intuition on the matter, yet it can be proved to result from the decreasing marginal cost of transportation. The same fails to hold for a transportation cost that is linear in distance or increases more than proportionately with distance (the latter condition being very unlikely in practice). It follows that firms cannot automatically treat consumers that are distant from competitors as a 'safe' market. Determining how secure a geographic segment of the market is depends not only on where the segment lies with respect to firms but also on the structure of transportation costs.

Another interesting property of geographic markets is the inclusion of a firm location in its own market area, that is, whether a firm serves in the direct vicinity of its physical location. Intuition suggests that a firm, whatever the economic and geographic environment it is operating in, ought to be able to competitively serve consumers at its door steps. Mathematical arguments can be used to show that this expectation is sound in most prevailing conditions, namely, when transportation costs conform to a concave function of physical distance. In the unlikely event of transportation cost rising at an increasing rate, however, the firm charging a higher sale price would fail to serve nearby consumers if it is close enough to its main competitor. Quite understandably, this situation translates into greater vulnerability to competitive pressures of the market place.

Prescriptive Market Areas

As pointed out earlier, the distribution function of a market economy fulfills the essential role of bringing the demand to clear with supply. As demand and supply are grounded in specific locations, and as the clearing process involves expenses on shipping between locations, the allocation between supply and demand points in space is a pressing consideration of most cost-conscious organizations. This process can be achieved with efficient use of resources according to what has been termed the 'transportation problem.' The spatial pattern of allocation from supply points to demand points derived from the transportation problem is prescriptive in that it conforms to an overarching objective of optimization of shipping costs across the system. It may not reflect the patterns that would emerge spontaneously if individual economic and social agents were to decide on this matter, a view point that is more consistent with the perspective presented in the next section of this article.

Various sorts of organizations resort to delineating service areas on the basis of the prescriptive flows between localized agents in the spatial system. This includes corporate entities serving a specific customer base through a network of facilities, like warehouses. These warehouse networks are constantly assessed by their corporate parent in terms of their capacity, number, and service area configuration so as to maintain system-wide efficiency. Public agencies in charge of siting facilities (e.g., libraries, schools, police stations) also follow similar approaches. In many cases, market areas exhibit a strong tendency to be compact and contiguous regions as they are born from the principle of geographic accessibility maximization. It is, however, not a required property that they be compact or contiguous. In fact, entire classes of problems are notorious for their departure from the ideal of compact market areas. Capacitated facility planning problems differ from others because capacity constraints are imposed on the ability of facilities to accommodate demand. When the constraint is closely binding, nearby customers may be assigned to a more remote facility in the interest of satisfying the constraint, but a trade-off occurs as a more distant group of customers may find themselves assigned to this facility because it is small enough to not exceed the capacity constraint. A broader class of so-called location-allocation models subsumes problems that take the location of facilities as given and seek only to distribute flows between supply and demand points.

Behavioral Viewpoint

The perspective outlined in Section Economic Underpinnings may be crucial to foster our conceptual understanding of market areas, but too often it fails the test of operational dependability for the simple reason that it assumes that a locale falls entirely into one or another market area. Business practitioners and applied social scientists have argued that a firm's ability to command consumers does not stop at the boundary of its market area, but rather that that ability extends past that boundary and into the areas imputed to other firms. The argument continues as follows. Because of a host of other factors explaining people's selection of a firm and because of interpersonal variations in preferences, demand is best viewed as the outcome of a probabilistic process. As a result, markets are contestable to a degree that is determined by the degree of interpenetration of firms' respective drawing power on consumers. Firms that are seen to be close substitutes by a certain segment of consumers (consumers with a similar socioeconomic profile and/or living in the same neighborhood) would also be engaged in fierce competition for this segment and their market shares would be pretty much on par. Conversely, when cross-elasticities between firms are low (limited substitution), the local market is rather captive.

The dominant paradigm to frame a more behaviorally grounded view of spatial competition and the structuring of market areas is that of spatial interaction, according to which one's attraction to a firm (i.e., likelihood to buy from this firm) is directly related to the attractiveness of this firm and inversely related to the travel cost to it.

In this view, a firm's market area is recast as the territory where attraction to this firm is greater than to any other firm. Since the 1930s, the approach has been known as Reilly's law of retail gravitation (Reilly, 1929; Huff, 1962) by analogy with Isaac Newton's law of planetary gravitation. Of course, the operational concept of market area is richer than its strict economic counterpart discussed earlier. All the geometric and topological properties of market areas (Section Spatial Analysis) are still pertinent under the spatial interaction paradigm. The intensity of demand in and out of the market area adds an important dimension to the characterization of market areas. The principle of spatial interaction generates a pattern of decreasing demand intensity away from the firm location, the rate of which depends on the structure of transportation costs and on interfirm competition.

The spatial interaction approach to spatial market analysis recognizes that firm attractiveness may be a multifaceted notion shaped by price, service, and merchandising properties of the firm. It also recognizes that perceived attractiveness is what influences firm choices and that sociodemographics is a powerful mold of individual perceptions. The capability to segment consumers by their sociodemographic characteristics becomes handy when the study area is highly segregated socioeconomically. It is now commonplace to carry market analysis within the interactive computer environment offered by geographic information systems. Automated numerical and database manipulations provide an essential toolbox for spatially enabled market analysis, also known as 'business geographics.' This data analytical capability has dramatically changed market geographies. Rather small segments of customers living in a certain district of a city and sharing similar tastes, preferences, and aspirations can now really be identified and their demand can be met locally. As the social, cultural, and ethnic mosaic of our cities has changed and as the city's geography has morphed, so has the geography of markets and of retailing.

In contrast to other market analysis approaches, the spatial interaction framework emphasizes that location matters. The geographic space and the modalities of how one reaches one location from another can be richly portrayed in a spatial interaction model if the appropriate data are used. This approach has freed scientists from assuming that inhabited spaces are as devoid of features as the flat and uniform surface of the Euclidean space. Spatial separation between a residential neighborhood and a movie theater can be accounted for by the drive time, even as it varies with the time of the day in response to traffic conditions. In such a study, Yue et al. (2012) use fine-grained Global Positioning System taxi traces to calibrate a spatial interaction model of market area delineation. This application is in line with recent trends in spatial information analysis to rely on large and highly disaggregated data sets (so-called big data) to better contextualize the behaviors embedded in the data. Another trend in this respect is based on the recognition that space and time have substitution relationships that can be exploited to compensate deficiencies of choice options. Therefore, the ability to reach potential destinations is to be defined in the space–time context and may be advantageously extended to a hybrid space–time space (Shaw and Yu, 2009).

Concluding Remarks

Consumerism and compelling marketing practices have replaced mass merchandising with ever more finely defined customer segments with their own wants and needs. Where sale price and travel cost used to be the driving forces, markets have become more fluid, uncertain, and varied than ever before. It should not be too much of a surprise therefore that the marriage of spatial interaction and conventional, microeconomic market-area analysis has enhanced our grasp of contemporary spatial market systems. The greater propensity of firms to form spatial clusters can be ascribed to the disappearance of exclusive market areas stipulated by the spatial interaction paradigm (de Palma et al., 1985). Concomitantly, a countervailing transformation may manifest itself in the form of greater differentiation of firms from each other (Thill, 1992) to better cater to the diversity of consumer segments.

In these times where information technologies are quickly redefining the relationships between demand and supply in postmodern economies, the question naturally arises whether electronic commerce (e-commerce) will bring the demise of the notion of market area. This postulate is in line with Cairncross's 1997 provocative title 'The Death of Distance.' While no definite answer can be advanced at this time, partial arguments can be articulated as follows. As a rapidly increasing portion of postmodern economies is siphoned through e-commerce, more commerce takes on a new organizational structure wherein spatial relationships and territoriality may play a lesser role. New commercial relationships contribute to the emergence of

new geographies, namely, the virtual geography of information and communication flows on the network of computer servers, the geography of social media users, and so on. It has been argued that the notion of market area is not tied to the existence of a physical space and that it has bearing in the space framed by firm characteristics. In the same vein, the notion can be extended to the topological network of the information superhighway. Additionally, even the most optimistic projections on the expansion of e-commerce do not predict that it will entirely replace traditional commerce. Hence, in all likelihood, face-to-face commerce and associated geographic markets will subsist. This is not to say that spatially defined markets will remain unaltered by the digital revolution. Location-based services are slowly but surely leaving a mark in the market place, but they also enable new forms of selective sharing of information based on locational proximities (i.e., accessibility), thus reestablishing the importance of location and geography in human behaviors where it appeared to be receding not too long ago. If evolution of the past decade gives any indication of future trends, the notion of market area will continue evolving in response to technology-enabled practices of targeted marketing toward increasingly entwined fields of interaction.

See also: Consumption, Geography of; Economic Geography; Geographic Information Systems; Information Society, Geography of; Location Theory; Spatial Analysis in Geography; Spatial Interaction; Spatial Optimization Models; Time-Space in Geography; Transportation Geography.

Bibliography

Allaway, A.W., Black, W.C., Richard, M.D., Mason, J.B., 1994. Evolution of a retail market area: an event-history model of spatial diffusion. Economic Geography 70, 23–40.

Applebaum, W., 1966. Methods for determining store trade areas, market penetration, and potential sales. Journal of Marketing Research 3, 127–141.

Berry, B.L.J., Parr, J.B., 1988. Market Centers and Retail Location. Theory and Applications. Prentice Hall, Englewood Cliffs, NJ.

Cairncross, F., 1997. The Death of Distance: How the Communications Revolution Is Changing Our Lives. Harvard Business School Press, Cambridge, MA.

Christaller, W., 1933. Die Zentralen Orte in Suddeutschland. Gustav Fischer Verlag, Jena, Germany.

Church, R.L., Murray, A.L., 2009. Business Site Selection, Location Analysis, and GIS. Wiley, Hoboken, NJ.

Donnelly, F.P., 2014. The geographic distribution of United States public libraries: an analysis of locations and service areas. Journal of Librarianship and Information Science 46 (2), 110–129.

Dost, J., 1968. Reassessment of the school location problem: a multifunctional role for the school in the urban environment. Annals of Regional Science 2, 136–141.

Fetter, F.A., 1924. The economic law of market areas. Quarterly Journal of Economics 39, 520–529.

Fortney, J.C., Lancaster, A.E., Owen, R.R., Zhang, M., 1999. Geographic market areas for psychiatric and medical outpatient treatment. Journal of Behavioral Health Services and Research 25, 108–116.

Frech III, H.E., Langenfeld, J., McCluer, R.F., 2004. Elzinga-Hogarty tests and alternative approaches for market share calculations in hospital markets. Antitrust Law Journal 71 (3), 921–947.

Ghosh, A., McLafferty, S.L., 1987. Location Strategies for Retail and Service Firms. Lexington Books, Lexington, MA.

Greenhut, M.L., 1956. Plant Location in Theory and Practice: The Economics of Space. University of North Carolina Press, Chapel Hill, NC.

Hanjoul, P., Beguin, H., Thill, J.-C., 1988. Theoretical Market Areas under Euclidean Distance. Institute of Mathematical Geography, Ann Arbor, MI.

Huff, D.L., 1962. Determination of Intra-urban Retail Trade Areas. School of Business Administration, University of California, Los Angeles, CA.

Isard, W., 1956. Location and Space-economy. MIT Press, Cambridge, MA.

Jue, D.K., Koontz, C.M., Magpantay, J.A., Lance, K.C., Seidl, A.M., 1999. Using public libraries to provide access for individuals in poverty: a nationwide analysis of library market areas using a geographic information system. Library and Information Science Research 21, 299–325.

Launhardt, W., 1882. Kommerzielle trassierung der verkehrswege. Zeitsschrift fur Architektur und Ingenieurswesen 18, 515–534.

Lösch, A., 1954. The Economics of Location. Yale University Press, New Haven, CT.

Okabe, A., Boots, B., Sugihara, K., Chiu, S.N., 2000. Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. Wiley, Chichester, UK.

de Palma, A., Ginsburgh, V., Papageorgiou, Y.Y., Thisse, J.-F., 1985. The principle of minimum differentiation holds under sufficient heterogeneity. Econometrica 53, 767–781.

Rau, K.H., 1968. Report to the bulletin of the royal Society. In: Baumol, W.J., Goldfeld, S.M. (Eds.), Precursors in Mathematical Economics: An Anthology. London School of Economics and Political Science, London, pp. 181–182.

Reilly, W.J., 1929. Methods for the Study of Retail Relationships. University of Texas Bulletin 2944, Austin, TX.

Shaw, S.-L., Yu, H., 2009. A GIS-based time-geographic approach of studying individual activities and interactions in a hybrid physical-virtual space. Journal of Transport Geography 17 (2), 141–149.

Thill, J.-C., 1992. Competitive strategies for multi-establishment firms. Economic Geography 68 (3), 290–309.

Thisse, J.-F., Zoller, H.G., 1983. Locational Analysis of Public Facilities. North-Holland, Amsterdam, The Netherlands.

Von Böventer, E., 1965. Spatial organization theory as a basis for regional planning. Journal of the American Institute of Planners 30, 90–100.

Yeates, F., 1963. Hinterland delimitation: a distance minimizing approach. The Professional Geographer 15 (6), 7–10.

Yue, Y., Wang, H., Hu, B., Li, Q., Li, Y., Yeh, A.G.O., 2012. Exploratory calibration of a spatial interaction model using taxi GPS trajectories. Computers, Environment and Urban Systems 36, 140–153.

Market Structure and Performance

Paul B Ellickson, University of Rochester, Rochester, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Sutton, volume 14, pp. 9211–9216, © 2001, Elsevier Ltd.

Abstract

What determines structure and profitability across markets for goods and services? Understanding the relationship between market structure and performance is critical for determining effective economic policy governing antitrust, intellectual property, industry regulation, and international trade. Economists have analyzed these issues for many decades. An initial descriptive literature has given way to a more formalized game-theoretic approach. One stream of literature paints on a large canvas, identifying strong mechanisms that can hold across the broad run of industries. The other focuses on specific industries, building detailed models tailored to key institutional details. Both have yielded important insights and raised additional questions and directions for future research. Recent work promises to unite both streams under a common methodological framework.

Introduction

Why are some markets dominated by a handful of firms, even at a global scale, while others remain consistently fragmented? Why are some firms able to charge prices that far exceed their costs of production, while other firms can barely break even? Does competition drive firms to be more productive? Questions like these have occupied industrial organization (IO) economists – microeconomists that study markets characterized by imperfect competition – for many decades. One way of framing these questions is to focus on the connection between market structure and market performance. Do more concentrated markets facilitate greater collusion, yield higher prices, and/or experience fewer productivity gains? Can firms make strategic investments in R&D or advertising that yield market structures that are more conducive to collusion? Because economic policy is often implemented at a level that impacts many industries at the same time, it is useful to develop answers and insights that apply to the broad run of industries. At the same time, more detailed or 'local' predictions may be needed to understand, for example, the impact of a particular merger in a specific industry.

Perhaps not surprisingly, economists have approached these questions from both perspectives. Using the structure of game theory, the *bounds approach* focuses on identifying strong mechanisms that apply under the most general of conditions, while the *structural method* uses the game-theoretic structure to focus on more narrow predictions tailored to specific industries. The following short survey provides a brief overview of the evolution of these literatures. I begin with a review of an earlier descriptive literature that identified many of the key stylized facts that are still being grappled with today. I then highlight the limitations of this early analysis as identified by a more formal game-theoretic approach to modeling industrial structure. In response, two separate but complementary streams of literature have emerged. I first discuss the bounds approach, which emphasizes broad cross-sectional predictions that hold across many industries, and then the structural method, which focuses on detailed institutional settings in which stronger modeling assumptions are appropriate. I conclude with a brief summary of a recent literature that combines features of both approaches.

Definitions and Empirical Measures

To set the stage, it is useful to first provide some definitions, along with the empirical measures used to quantify both market structure and market performance. *Performance* generally refers to how competitive (or efficient) an industry is or, more broadly, how successful it is at delivering benefits to consumers. The main focus is on competition, and its connection to the profitability of firms. Competitiveness can be captured by the degree of market power, which refers to a firm's ability to set price above the cost of producing a single additional unit. A natural measure of market power is the price-cost margin $\left(p-c, \text{or } \frac{p-c}{p}\right)$. A share-weighted average of price-cost margins yields the more general Lerner index of market power for the industry as a whole, providing an overall measure of competitiveness. Although some authors prefer to calculate a rate of return on investment (to get more directly at profits), issues of measurement and interpretation can introduce significant noise into this calculation (Fisher and McGowan, 1983), leading many researchers to turn away from these and other accounting-based constructs.

Market structure refers to the identity and relative sizes of the firms that comprise an industry (and, in the case of multiproduct firms, the products they produce). Often, it can be fully captured by a list of each firm and their respective market shares (i.e., their share of total industry revenue). In industries with a large number of firms, or when relying on published summary statistics, employing a function of these shares to summarize market structure can be more practical. The most common such function is the k-firm concentration ratio, which reports the combined share of the largest k firms in the industry. Typical values of k are four and eight, though many such measures may be reported (unfortunately, the share of the largest firm is often masked for confidentiality reasons). The sum of the squared market shares, called the Herfindahl index,

provide a single measure of the *overall* structure that is often employed in merger analysis. Critically, any measure of market structure requires a careful definition of the relevant market, a task that can be complex in certain cases (e.g., does a 'smart' phone belong in the same market as a digital camera?) and plays an important role in antitrust analysis.

The Structure-Conduct-Performance Paradigm

Systematic efforts to describe the relationship between market structure and market performance began with the development of the Structure-Conduct-Performance (SCP) paradigm, a framework most closely associated with Edward Mason and Joseph Bain. Although it has been sharply critiqued in the ensuing decades, the SCP literature served to systematically identify and highlight many of the stylized facts and empirical regularities that motivate the modern, game-theoretic approach. While too vast to adequately summarize here, Schmalensee (1989) provides a comprehensive review of the SCP literature and the stylized facts it uncovered.

The SCP framework posited a one-way chain of causation running from industry structure (firm concentration) to firm conduct (pricing) to market performance (profitability, innovation). The central claim was that high concentration would lead to high prices and thus high profits. Recognizing that free entry by competing firms ought to quickly erode these profits, significant barriers to entry would then be required to maintain a concentrated structure. In his earliest work, Bain associated these entry barriers with fixed economies of scale in production (i.e., falling average costs), which could arguably be taken as exogenous to the firm's decision-making problem (i.e., predetermined). The validity of this exogeneity assumption will play a critical role in what follows. Analyzing a cross-section of US manufacturing industries, Bain (1956) documented a significant correlation between scale economies and concentration, consistent with the framework's predictions. The rapid increase in available data sets facilitated the production of many such studies throughout the 1960s and 1970s.

At the same time, Bain and later contributors to the SCP literature identified other industries that did not feature these production economies yet still remained concentrated due to intensive investment in advertising and/or research and development (R&D). For example, the soft drink industry has remained remarkably concentrated for many decades despite little required investment in fixed costs of production. Instead, there is an intense rivalry in advertising outlays required to effectively build and maintain a national brand. The shift of focus to these new sources of scale economies led to two problems. First, the empirical results here were more mixed, particularly in the case of R&D, which exhibited little systematic correlation with market structure. More critically, many economists argued that advertising and R&D should more appropriately be thought of as decision variables of the firm, not as predetermined features of the business environment. Addressing this concern is a key focus of the modern, game-theoretic literature.

The SCP literature was also successful in identifying a second strong empirical regularity: the impact of increased competition on prices (and price-cost margins). The hypothesis that a fall in market concentration leads to a fall in prices and price-cost margins found strong empirical support, as well as solid theoretical underpinnings. This conjecture provides another important jumping off point for the more recent literature. Unfortunately, the relationship between market structure and *profitability* proved much more vexing and the empirical evidence decidedly mixed. Measurement and interpretation issues plagued the accounting numbers used to quantify various measures of profit or rate of return (Fisher and McGowan, 1983). Perhaps more damning were issues of interpretation.

The SCP approach takes structure as given, implying that markets that are both concentrated and profitable must be in some way collusive or noncompetitive. An alternative viewpoint, most closely associated with economists from the Chicago School (Harold Demsetz and Sam Peltzman), reverses the order. To understand their concern, consider two industries, one in which all firms share the same capabilities and a second where firms exhibit sharp differences in efficiency. The industry with heterogeneous firms will likely end up more concentrated – with the efficient firms garnering a larger market share – and more profitable (at least for the leading firms, since free entry will only drive the profits of the marginal firm(s) to zero). However, in this case, the interpretation is clearly pro-competitive. Moreover, the relevant empirical evidence has tended to support the efficiency argument.

The Demsetz critique was emblematic of a deeper conceptual problem with the SCP framework. The assumption of a one-way chain of causation from structure to performance ignores the possibility that, for example, perceptions of conduct and performance might instead drive the entry decisions that determine structure. Consider a market in which firms perceive a fierce degree of price competition upon entry. Anticipating a limited scope for sustained profitability, a smaller number of firms might enter than would have had they perceived a more collusive economic environment (e.g., accommodation by a well-functioning cartel). Thus, policies that foster competition (e.g., tougher antitrust policy) might actually lead to *more* concentrated market structures. This surprising conclusion is a central point of the bounds approach discussed below. Broadly speaking, while the SCP approach was successful in identifying a large number of empirical regularities, it failed to address the complex role of strategic interaction in determining firm conduct. These interactions play a key role in the game-theoretic approach discussed next.

The Game-Theoretic Approach to Market Structure

The formalism of game theory provided IO economists with a rigorous framework with which to examine the subtle interconnections between firm strategies, market structure, and market outcomes. Additional motivation for differences in firm shares and fortunes were found in the strategic asymmetries arising from the opportunity to move first, the ability to 'learn by doing', or the incentive to exploit 'network effects' across interconnected markets. However, more than a decade of careful research brought the sobering conclusion that robust results that could apply to the broad range of industries were

few and far between. Seemingly arbitrary features of the game setup often lead to vastly different equilibrium outcomes. Critically, many results necessarily turn on such difficult to proxy factors as a firm's beliefs about its rivals' actions or the exact timing of investment decisions, as well as other inherently unobservable constructs such as the number and identities of potential entrants. On the one hand, the flexibility of the game-theoretic framework allows it to rationalize a wide spectrum of outcomes. On the other, a wide range of models can explain the same specific outcome, undermining a key assumption in most statistical models (a one-to-one mapping from parameters to outcomes) and thereby complicating statistical inference. IO economists faced a trade-off between constructing highly specialized models tailored to a particular setting, or opting for less-sharp predictions that might hold across a broad range of settings.

It is important to note that this ambiguity is a natural feature of the economic setting, not a failure of the game-theoretic approach. Realistic games often have multiple equilibria, arising from different beliefs or the vagaries of implicit coordination. More generally, the outcomes of formal models often depend delicately on assumptions that are difficult to observe or verify (e.g., Do players move simultaneously or sequentially? Is competition best modeled as occurring in prices or quantities?). This lack of information is an intrinsic feature of the economic environment. Many of the constructs that we would like to condition on (e.g., costs) are either unobserved or challenging to control for. In response, the modern game-theoretic empirical literature has moved in two separate, yet complementary directions. One embraces the ambiguity; the other places structure upon it.

The first, which is now known as the bounds approach, continues the SCP tradition in seeking to identify predictions that hold across a broad range of industries. To do so, it favors mechanisms that are robust to a wide range of modeling choices and depend only on observable features of the economic environment. The second approach, typically referred to as the structural method, focuses instead on single industries in isolation, where more specific modeling assumptions can be brought to bear, allowing unobserved elements to either be recovered from the assumptions of the model or controlled for in the statistical design. The most recent work combines elements of both.

The Bounds Approach

The bounds approach is most closely associated with the seminal work of John Sutton (1991, 1998), his students, and coauthors. This section draws heavily on Sutton's exposition. The goal of the bounds literature is to identify a number of 'strong mechanisms' that characterize equilibrium outcomes across a broad range of industry settings. The idea is to invoke as few assumptions as possible, building only from the most strongly confirmed empirical regularities, and deriving testable implications that hold irrespective of arbitrary or difficult-to-verify modeling choices. In practice, this involves placing restrictions on outcomes, rather than strategies. By constructing, painting on so large a canvas will limit the sharpness of predictions.

Sutton starts from a general game-theoretic setup in which a large number of firms (beginning life as potential entrants to an industry) compete in a series of stages, culminating in a final, product-market competition stage that determines industry and firm profits. In the simplest version, there are just two stages: an entry/investment stage and a price competition stage. Firms are strategic and forward-looking in their choices and investments, anticipating how their actions (and those of their rivals) will influence market outcomes. Unlike the SCP framework, entry is endogenous. For example, firms that do not foresee earning nonnegative returns in the product market will choose not to enter, reflecting a key component of the Demsetz viewpoint. In this setting, the outcomes that may survive as equilibria must satisfy two fundamental properties: *viability* and *stability*. Viability ensures that all firms that choose to enter the industry expect to earn nonnegative profits. This condition eliminates cases in which all firms choose to enter. Stability requires that any firm that chose not to enter expected to lose money (i.e., earn negative profits). Note that these are standard properties that any pure strategy, Nash equilibrium must satisfy within this class of models.

Based on these two properties, Sutton establishes a series of strong results or 'robust mechanisms' characterizing equilibrium market structure. The first applies to industries in which advertising and R&D play no major role in competition (i.e., the types of industries with which Bain was originally concerned). These are referred to as exogenous sunk cost markets or Type 1 industries. What happens to market structure as market size expands in these settings? Sutton shows that, in these exogenous sunk cost markets, as market size increases, the minimum level of concentration that can be supported as an equilibrium configuration decreases asymptotically to zero. Phrased differently, the *lower bound* to concentration decreases to zero as market size expands. Note that this does not imply that every market will experience a monotonic decrease in concentration as market size expands – this is a bound not a functional relationship. In some cases, markets may remain concentrated indefinitely. This lack of precision is a product of the theory's breadth: although we can construct examples in which the profitable opportunities that arise as market size expands are systematically capture by the largest firms, we cannot *rule out* equilibria in which markets become more fragmented as markets grow in size. For this reason, this is sometimes called a 'fragmentation' result. Note that we will find the opposite conclusion when we shift attention to advertising and R&D intensive industries below.

Sutton's second 'robust mechanism' concerns the relationship between the shape and position of this lower bound and the degree or 'toughness' of competition that prevails in a given market. The degree of competition determines the shape of the function that aggregates the decisions made by all firms in the investment stage(s) and produces the vector of profits arising in the competition stage. This shape is influenced by decisions such as whether to model competition as taking place in quantities (so-called Cournot competition, which is considered somewhat collusive) or prices (so-called Bertrand competition, which is viewed as more competitive) and environmental factors like the degree of antitrust enforcement and the ease

of transportation. What happens to concentration or, more precisely, the lower bound to concentration, as competition becomes tougher? Consider a shift from a more collusive setting to one that is more competitive. For example, suppose the regulating authority increases the penalty (or strengthens enforcement) associated with anticartel laws. This change would imply an increase in the toughness of competition, as firms would presumably now find it more difficult to collude. Sutton's second result establishes that an increase in the toughness of price competition will lead to an upward shift in the lower bound to concentration, although the lower bound will still tend toward zero as market size expands. This shift is driven by the viability condition noted above: a return to equilibrium will require some firms to exit or consolidate, thereby raising the equilibrium level of concentration for a given market size. Note that this has important policy implications: efforts to increase competition can lead to a more concentrated market structure. This is likely not what one might first expect and illustrates the power of the game-theoretic perspective.

Both results have garnered strong empirical support. Sutton (1991) provides a wide-ranging, multi-industry, cross-national analysis that includes both formal statistical tests and compelling case studies. More recently, Sutton (2007) provides a comprehensive overview of the related literature, which includes many empirical studies that complement the bounds approach.

Perhaps the most provocative of Sutton's results concerns those industries in which advertising or R&D plays a critical role. Owing to the strategic nature of these investments, these 'Type 2' industries are referred to as *endogenous* sunk cost markets. In these settings, Sutton instead finds that concentration is bounded below, breaking the fragmentation result described above. As market size expands, the share of the largest firm remains strictly positive. The formal result can be summarized as follows. Let $S\pi$ denote the profit earned in the final stage of competition, where S represents the size of the market and π a function capturing all previous investment and pricing decisions. Assuming marginal costs are flat and markets increase in size through successive replication (meaning larger markets are no more heterogeneous than their smaller counterparts), suppose that for some constants $a > 0$ and $k > 0$, a firm that spends k times as much on fixed costs as any rival can earn at least aS in the final stage, there then exists a lower bound to concentration that is independent of the size of the market. Thus, unlike the previous exogenous fixed cost case, the importance of R&D or advertising places a sharp limit on the structure of markets that can survive in equilibrium. Fragmentation cannot occur here. Larger markets are still more profitable than small, but as market size expands the incentive to increase the level of fixed investments rises accordingly. In particular, larger markets will induce greater investment by at least some firms, eliminating the indefinite fall in concentration observed in the Type 1 markets. Sutton therefore provides a compelling reason for why some industries are consistently dominated by a handful of firms: the escalating nature of their required sunk investments. Sutton (1991) provides strong empirical support for this nonconvergence result across a broad range of industries.

Applications and Extensions of the Bounds Approach

Several authors have applied Sutton's framework in a variety of institutional settings, often extending the scope of what are considered endogenous sunk investments. For example, Ellickson (2007) uses the bounds approach to analyze competition in the US supermarket industry. While relatively unconcentrated at the national level, the grocery industry is dominated at the local (city) level by 3–4 chain firms, an empirical regularity that holds for geographic markets both large and small. While R&D and advertising play almost no role in this industry, Ellickson argues that supermarkets invest competitively in information technology aimed at increasing the number of products offered in each store. These investments limit the number of firms that can profitably compete. A key challenge is identifying independent markets of varying size with which to test this nonconvergence result. Ellickson argues that, on the supply side, relatively independent grocery markets can be defined at the level of the distribution center, and uses variation in the size of these 'distribution markets' to identify the lower bound to concentration. He finds that the share of the largest firm is bounded below at a limiting level of over 12% while, in a comparison industry in which costs are exogenous (barber shops and hair salons), that level falls to under 0.5%.

Sutton recognized the role of independent submarkets in establishing robust predictions. The challenge is particularly acute in the case of R&D, where intense spending along a narrow research trajectory can lead to both high levels of investment and deceptively low levels of market concentration in industries where markets are defined too broadly. In Sutton (1998), he expands his earlier analysis to include settings in which R&D leads to multiple, independent trajectories. Once again taking a bounds approach, he provides a method for identifying the types of R&D intensive industries that will remain persistently concentrated. Finally, he also develops a bound prediction that applies to the long-standing debate over the size distribution of firms, providing a game-theoretic rational for what was historically a purely statistical literature.

Structural Methods

Reacting to both the inherent ambiguities embedded in the game-theoretic approach and the limitations on what we can clearly observe and quantify as empirical researchers, Sutton's bounds approach is based on making as few assumptions as possible. In contrast, the more dominant structural approach to empirical analysis adopts a more focused perspective, emphasizing single industry settings in which stronger assumptions can be employed and (ideally) validated. In so doing, it can deliver far more precise predictions. It also makes explicit the assumptions underlying identification of causal effects, allowing the reader to judge for themselves the appropriateness of various modeling choices. Furthermore, by uncovering the economic primitives that drive agents' strategic choices, alternative (counterfactual) scenarios can be evaluated, including large changes to economic policy that may be well outside what is observed in the data. Of course, the validity of these predictions relies on the realism of the

proposed model, which must be carefully tailored to the specific institutional setting. Consequently, the structural approach focuses expressly on specific industries, where detailed data and institutional knowledge can mitigate some of the concerns regarding unobservable, or poorly measured, covariates and justify the stronger assumptions needed to sharpen our predictions. When carefully executed, structural methods can yield insights that are not available from simple descriptive analysis.

While the earliest structural work attempted to estimate a 'conduct parameter' characterizing firm behavior, the modern approach assumes a form of conduct and uses detailed modeling assumptions to recover difficult-to-measure constructs like the variable and fixed costs of production. For example, to evaluate a merger in the soft drink industry, a detailed demand system would be estimated. By assuming a form particular of conduct (e.g., Nash in prices competition among differentiated firms), estimates of the variable cost of production can be recovered (eliminating the need to rely on the accounting measures criticized earlier). With these estimates in hand, a fully specified model of competition can then be used to simulate outcomes under different ownership structures, quantifying the impact on both consumers and producers (e.g., Nevo, 2000). Thus, the impact of the merger can be evaluated prior to its consummation, thereby guiding optimal policy decisions. The entry of new products or changes in other regulatory policies can be analyzed similarly. These methods now play a key role in merger analysis.

While structural methods have been used to address a broad range of questions in IO, the two areas closest in spirit to the market structure and performance agenda are the 'static entry' literature and the recent work on dynamic games. Both streams of literature include a substantial methodological component, as researchers work through the implications that the aforementioned informational incompleteness has on statistical inference, while tackling the daunting computational burden of estimating multiagent models. In keeping with the applied focus of this overview, I will emphasize the more policy-related issues addressed.

The static entry literature began with the seminal work of Bresnahan and Reiss (1990), which focused on the strategic entry decisions of small retail firms in isolated rural markets (the emphasis on isolated markets helps address the issue of market definition noted earlier). Bresnahan and Reiss examined the relationship between market size and the number of entrants to the market, attempting to isolate the impact of price competition on firm profits. They reasoned that if total industry profits decrease in the number of entrants and firms must pay a fixed and sunk cost to enter the market, then the market size that could accommodate two firms would need to be more than twice as large as a market that could only accommodate one (holding other drivers of profitability fixed). In particular, if competition drives down profits, the number of firms that enter a market should grow less than proportionately with the size of that market. Note that this provides an indirect test of the price-competition mechanism identified under the SCP paradigm and taken as a primitive in the bounds approach.

By focusing on the number of entrants (as opposed to their identities), they were also able to finesse some issues of incompleteness associated with multiple equilibria. In a series of papers, Bresnahan and Reiss documented a strong price-competition effect that tapered off once 3–5 firms were present in a market. This tapering off also provides support for the intuition that collusive behavior may be harder to sustain with a larger number of participants, a common assumption in many treatments of tacit collusion. Recognizing that the less than proportionate increase in firms might instead reflect a version of the Demsetz critique, Berry (1992) extended the Bresnahan and Reiss approach to accommodate heterogeneity in firms. Berry reasoned that if the more efficient firms enter first, it would then make sense that later entrants would require a proportionately larger market (greater demand) over which to spread their relatively higher cost outlays. Focusing on the airline industry, Berry continued to find a strong price-competition effect despite these differences in efficiency. However, in keeping with a central theme of the bounds approach, this sharper prediction required a stronger assumption: a specific mechanism for selecting among equilibria (with sharp differences across firms, the number of entrants may no longer be unique as it was with Bresnahan and Reiss). Subsequent authors have extended the framework to include additional sources of product differentiation (Mazzeo, 2002), network economies (Jia, 2008), and alternative information structures (Seim, 2006).

Up to now, the focus has been entirely on cross-sectional implications of what are mostly static models of firm competition and industry structure. Clearly, most industries are dynamic, with structures that continue to evolve over time. Concepts like 'learning by doing' and price skimming are inherently dynamic, while the payoffs to research and development are often realized several years in the future. Fortunately, in the 1970s and 1980s, detailed firm-level census data increasingly became available, allowing researchers to document industry evolution at an extremely disaggregated level. The nature of this evolution has important implications for issues like first mover advantages, industry shakeouts, learning, and innovation. What emerged was clear evidence of heterogeneity across both firms and industries, with entry and exit often occurring simultaneously. To begin to address these issues, Ariel Pakes and his coauthors have developed a flexible framework for analyzing firm dynamics from both a theoretical and empirical perspective. This framework, originally formulated by Ericson and Pakes (1995) but extended in many directions by subsequent authors, can accommodate simultaneous entry and exit, strategic investment, imperfect competition, and both firm- and market-level heterogeneity. While this research agenda is still in its early stages, its application has blossomed in the past few years due to recent advances in estimation and computational techniques (Doraszelski and Pakes, 2007).

Goettler and Gordon (2011) provide a nice illustration of how this framework can be applied to examine the relationship between market structure and performance. Up to now, we have emphasized the impact of market structure on firm profits, but structure should also influence other outcomes like innovation and productivity. In particular, we might ask if competition spurs productivity? Or does a monopolist, who is in an ideal position to fully appropriate the fruits of their labor, have greater incentives to innovate? Goettler and Gordon

develop a structural model of competition between Intel and AMD in the production of microchips (CPUs). Entering a long-standing debate between adherents of the viewpoints offered by Kenneth Arrow and Joseph Schumpeter, the authors ask what would happen to innovation (and prices) if AMD were to cease to exist. Consistent with the Schumpeter argument, they find that the rate of innovation would be 4.2% higher in AMD's absence. However, consumer surplus would also fall by 4.2% due to the higher prices that would be charged by Intel. Whether these insights might apply in other settings is an open question, but a similar approach could be used to analyze other industries.

A Structural Bounds Literature

Both the bounds and structural IO literature are motivated by the twin limitations of our ability to predict outcomes (via formal game-theoretic models) and the degree to which we observe all the relevant institutional details and statistical covariates that drive observed structure. This model 'incompleteness' arises from limited information on the part of the researcher regarding the correct modeling assumptions, the relevant covariates to include in a regression, the underlying equilibrium selection mechanism, and the key institutional details that govern firm behavior. As structural researchers seek to limit the number of assumptions they must invoke to conduct analysis, the two strands of literature start to merge. In particular, the pioneering work of Charles Manski and Elie Tamer regarding the econometric treatment of such incomplete models is now being used to tackle questions of market structure and performance. However, rather than starting from the most general setting (a la Sutton), the approach here is to begin from the particular and move to the more general. For example, in an extension of the Berry model of airline competition, Ciliberto and Tamer (2009) are able to relax the requirement of an explicit equilibrium selection mechanism in a model with heterogeneous firms, delivering on the promise of more detailed insights with fewer assumptions. Similar methods might relax the requirements on assumed conduct even further (to allow bargaining or matching models to be considered as well).

Conclusions

The relationship between market structure and market performance has occupied IO economists for several decades. Providing concrete insights is important for determining effective policy toward antitrust, intellectual property, regulation, and international trade. An initial descriptive literature has given way to a more formalized game-theoretic approach to tackling these questions. One stream of literature paints on a large canvas, identifying strong mechanisms that can hold across the broad run of industries. The other focuses on the more narrow picture, building detailed models tailored to specific industrial contexts. Both have yielded important insights and raised additional questions and directions for future research. Recent work promises to unite both streams under a common methodological framework.

See also: Cooperative Game Theory; Game Theory: Noncooperative Games; Game Theory; Global Developments, Markets, and Well-being; Hierarchies and Markets; Markets: Anthropological Aspects; Profits.

Bibliography

Bain, J.S., 1956. Barriers to New Competition. Harvard University Press, Cambridge, MA.
Berry, S., 1992. Estimation of a model of entry in the airline industry. Econometrica 60, 889–917.
Bresnahan, T.F., Reiss, P.C., 1990. Entry in monopoly markets. Review of Economic Studies 57, 531–553.
Ciliberto, F., Tamer, E., 2009. Market structure and multiple equilibria in airline markets. Econometrica 77, 1791–1828.
Doraszelski, U., Pakes, A., 2007. A framework for applied dynamic analysis in IO. In: Armstrong, M., Porter, R. (Eds.), Handbook of Industrial Organization, vol. 3. North-Holland, Amsterdam, pp. 1887–1966.
Ellickson, P.B., 2007. Does Sutton apply to supermarkets? Rand Journal of Economics 38, 43–59.
Ericson, R., Pakes, A., 1995. Markov perfect industry dynamics: a framework for empirical work. Review of Economic Studies 62, 53–82.
Fisher, F.M., McGowan, J.J., 1983. On the misuse of accounting rate of return to infer monopoly profits. American Economic Review 73, 82–97.
Goettler, R.L., Gordon, B.R., 2011. Does AMD spur intel to innovate more? Journal of Political Economy 119, 1141–1200.
Jia, P., 2008. What happens when wal-mart comes to town: an empirical analysis of the discount retail industry. Econometrica 76, 1263–1316.
Mazzeo, M., 2002. Product choice and oligopoly market structure. Rand Journal of Economics 33, 221–242.
Nevo, A., 2000. Mergers with differentiated products: the case of the ready-to-eat cereal industry. Rand Journal of Economics 31, 395–421.
Schmalensee, R., 1989. Inter-industry differences of structure and performance. In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 2. North-Holland, Amsterdam, pp. 951–1009.
Seim, K., 2006. An empirical model of firm entry with endogenous product-type choices. Rand Journal of Economics 37, 619–640.
Sutton, J., 1991. Sunk Cost and Market Structure. MIT Press, Cambridge, MA.
Sutton, J., 1998. Technology and Market Structure. MIT Press, Cambridge, MA.
Sutton, J., 2007. Market structure: theory and evidence. In: Armstrong, M., Porter, R. (Eds.), Handbook of Industrial Organization, vol. 3. North-Holland, Amsterdam, pp. 2301–2368.

Marketing Strategies

Enrico Valdani and Alessandro Arbore, Bocconi University, Milan, Italy

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous article by E. Valdani, volume 14, pp. 9216–9221, © 2001, Elsevier Ltd.

Abstract

Marketing strategies are reviewed both in an external (i.e., 'Market Power School') and internal (i.e., 'Resource-based View') perspective. Focus will be put on marketing strategies based on market-driving capabilities and resources. A model of competitive strategies will then be presented in a circular perspective, considering the shifting nature of what are called movement, imitation, and position games.

Introduction

Marketing strategies pursued by firms are aimed at obtaining a sustainable competitive advantage over competitors. In a marketing perspective, firms obtain a competitive advantage when they are able to identify, create, and communicate reasons of preference for their customers, with respect to their competitors.

In this article, marketing strategies will be reviewed both in an external and internal perspective. Focus will be put on marketing strategies based on market-driving capabilities and resources. A model of competitive strategies will then be presented in a circular perspective, considering the shifting nature of movement, imitation, and position games. Some final points about marketing strategies in the third millennium will be presented, with the proposition of a new, proactive firm, able to compete in different games, with different strategies and weapons, like trust, creativity, sensing, integration, and networking capabilities.

Marketing Strategies in an External and Internal Perspective

Marketing strategies can be viewed in two different ways: (1) with an external emphasis, typical of the 'Market Power School,' which dominated the 1980s; this school draws on the Harvard tradition (Mason, 1939; Bain, 1956) in industrial economics studies; however, Porter (1980, 1985) is the most important author within this stream of thought and (2) with an internal emphasis, typical of the Resource-based View of the Strategy, which dominated from the 1990s. Authors like Wernerfelt (1984), Rumelt (1984, 1987), Barney (1986), Dierickx and Cool (1989), Itami (1987), Grant (1991), Peteraf (1993), and Prahalad and Hamel (1990, 1994) have been among the most important contributors within this stream of thought.

1. The 'Market Power School' puts its emphasis on the analysis of the existing environment, following the 'Structure-Conduct-Performance' paradigm of industrial organization. The normative prescriptions of the Market Power School, particularly the three generic strategies suggested by Porter (1980, 1985) – differentiation, cost leadership, and niche strategies of focalization – depend on the external analysis of the environment, conducted mainly with the five-forces model. The five-forces model is an analytical tool developed by Porter (1980, 1985) which helps to individuate the attractiveness of an industry, by analyzing not only the competitors actually in the arena, but also the forces of the potential entrants, the producers of substitutes, and the influences of customers and suppliers. Therefore, the 'Market Power School' considers strategy primarily as positioning a firm in a given industry context; the focus is on the external analysis of the existing industry and on its attractiveness. The strategic process should determine a 'fit' between the given distinctive strengths of a firm and the given opportunities offered by the external environment.

In synthesis, the Market Power School considers the general strategic process, at the corporate level, as the allocation of the right mix of resources to the various Strategic Business Units, adopting a financial portfolio view, and the competitive and marketing strategies, at the business level, adopting a product-based view of competition.

2. The Resource-based View of the Strategy, rather than on the external environment, puts its emphasis on the unique endowment of resources, assets, competencies, and capabilities of each firm. By creating and continuously incrementing its resource endowment, a firm becomes different from another. So the unique endowment of each firm is at the roots of its competitive advantage, whose sustainability is higher, the higher are the barriers raised to imitability. Marketing strategies are not concerned with the exploitation of the existing resources, but with the continuous internal commitment to increase the resource endowment. A clear example of a resource that should be increased, and not just exploited, is brand equity (Aaker, 1984, 1991).

Resources may be either tangible or intangible; it has been widely recognized that marketing strategies based on intangible resources generate hard to imitate competitive advantages. Typical marketing resources are trust and knowledge; both are intangible and strongly linked with customer conviction (Reicheld, 2003, 2006), which is the beginning and the end of every marketing strategy. At the beginning, customer conviction is the key because a delighted customer, promoting our value proposition, is the very aim of any successful growth strategy.

At the end, it is the most important result of each marketing strategy, because the long-term survival of a firm depends on the level of satisfaction, loyalty, and trust of its customers. So analysis, measurement, and improvement of the customer experience are among the most important activities for marketing strategies in a resource-based perspective (Valdani, 2009).

Marketing Strategies in a Resource-Based-View Perspective

In recent years, firms tend to be living in a new competitive landscape, which is characterized by structural changes mostly related to the new digital environment. This leads to a tremendous empowerment of the customer, as well as to new opportunities for collaboration within the entire value system of the company, from suppliers to business partners.

At the same time, these changes inevitably lead to a dramatic increase of the competitive pressure, continuing a trend started in the 1990s and effectively described by the expression 'hypercompetition' (D'Aveni, 1994; Valdani, 1995, 2000).

As a matter of fact, in the current turbulent environment, normative prescriptions based on static industry analysis tend to become useless. As Grant (2010) puts it, when the external environment is in a state of flux, the firm's own resources and capabilities may be a much more stable basis on which to define its identity. In the resource-based perspective, marketing strategies recognize the logical and temporal dominance of resources upon products. Products are considered only as the expression of the internal resources of the firm. Moreover, resources are not product-specific, but they can lead to the achievement of competitive advantage in a wide range of products and services. For example, the exceptional resources and capabilities of Apple, in terms of market sensing, customer experience, and brand identity, allowed this company to replicate and enjoy a competitive advantage in a broad set of new markets, beyond the traditional computing battlefield.

In this perspective, it is possible to state that firms need to adopt a proactive approach to their marketing strategies, leveraging their market-driving capabilities in order to reach a double aim: to manage the present and to prepare for the future at the same time (Abell, 1993; Valdani and Arbore, 2012). This is a so-called dual strategy. To manage the present means developing capabilities of relationship management, competitive intelligence, and process reengineering in a customer-centered perspective. To prepare for the future means developing the managerial capabilities of destroying the existing common wisdom, to imagine, enact, and create a new competitive environment, regenerating strategies and reinventing industries and their borders.

Marketing Strategies in a Circular Model

It is possible to describe the marketing strategies of a firm by adopting the circular model described in Figure 1. The model shows the three possible states of nature of the competitive arena. They are: the game of position, the game of movement, and the game of imitation.

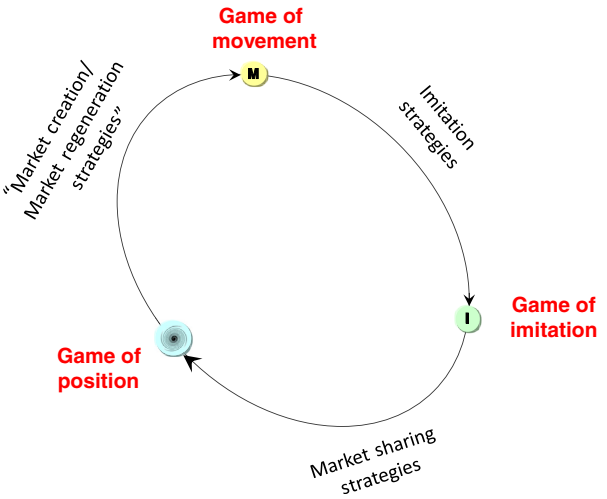

Figure 1 Marketing strategies in a circular perspective. Adapted from Valdani, E., Arbore, A., 2012. Competitive Strategies: Managing the Present and Planning the Future. Palgrave Macmillan, Hampshire, England.

Marketing Strategies in the Game of Position

During a game of position, the environment is quite stable, the product life cycle is generally in its maturity phase, and both customer needs and competitors are well defined. General marketing logics in a game of position are those of market sharing, aimed at increasing a firm's market share by reducing the competitors' one: aggressive marketing spending or pricing strategies are very important in these conditions. This, of course, makes the game of position a 'win–lose' scenario.

On 9 July 2007, for example, Sony Corporation issued a press release announcing that it was slashing the PlayStation3 price tag by $100, which was 17% less than the introductory price set in November 2006. At the same time, the company publicized the launch of a model with more memory at the previous price. The same day at a press conference in Tokyo, a Microsoft executive hastened to state that the company would not lower Xbox 360 prices. Nonetheless, on 7 Aug 2007 the Redmond firm announced a $50 price cut on the US market, to be followed two weeks later by a €50 reduction in Europe (approximately 13%). Similar moves by both companies would replay in October, November, and December. And again until the launch of the new versions. The strategic evolution in this competitive arena suggests that Sony and Microsoft were then playing what we called a Game of Position. These maneuvers typically cause a drop in profitability, since direct, frontal competition forces the companies to offer more and more for less.

Here the firms must effectively and efficiently manage all their processes and their offering in terms of brand, price, communication, promotion and distribution policies, potentially to the ultimate point of market commoditization. In sum, since the sources of the competitive advantage in a Game of

Position are well-known, any potential advantage is often a negligible one: it is all about pitting strength against strength. As a result, in sectors that enjoy no legal barriers to protect from competition, this scenario tends to evolve toward a so-called final frontier of value (D'Aveni, 1994), that is, a situation derived from the notion of perfect competition found in classic economic theory, where long-term profits are nil (Figure 2).

Marketing Strategies in the Game of Movement

In order to avoid the evolution to the 'final frontier of value,' firms pursue strategies of market regeneration and market creation, that is, what we call games of movement. In a proactive perspective, these companies increment, leverage, and stretch their market-driving capabilities and resources. By doing so, they create new markets or reconfigure radically the existing ones. These maneuvers usually require also networking strategies: in fact, a game of movement is often successful only in a coevolutionary perspective, putting together resources and competences of many companies.

"Give me a place to stand on, and I will move the Earth" said Archimedes. To initiate a Game of Movement, the "places to stand on" are as many as marketing managers can imagine. Indeed, disruptive levers are not exclusively or necessarily related to technology or research and development, as widely acknowledged and accepted evidence reminds us. For example, the Netflix's game of movement, which defeated Blockbuster, was just based on *different* channels (the Internet and the Postal service) and a *different* pricing scheme (monthly flat fee, instead of pay-per-rent). It is not anymore about doing more for less, like in the game of position, but doing things a different way. This is about winning without fighting, that is, avoiding a direct attack. This is the story of companies introducing new value propositions, or even new business models, like Ikea, Zara and low cost airlines.

The specificity of the game of Movement requires an array of competitive weapons, based on new knowledge generation, information, creativity, and firm competency integration at various levels: inside and outside the organization, and more and more often, also outside the borders of one's industry. The successful story of functional foods like Danacol (Danone's dairy drink that lowers cholesterol) is an example of this kind: a food company exploring the borders of the pharmaceutical industry.

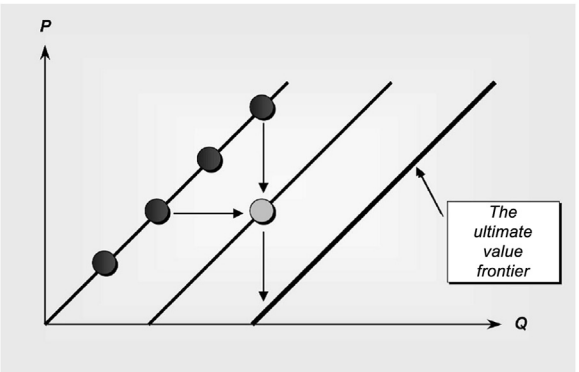

Figure 2 The ultimate value frontier. Reproduced from Valdani, E., 1995. Marketing Strategico. Etas Libri, Milan, Italy.

The arsenal of new competitive weapons is vital for developing, applying and disseminating innovation, in order to educate the market about innovation, to develop industrial infrastructure, to set new standards, and to make a quick entrance or exit from any given market. Last, but not least, these strategies require the ability of a firm to overcome important barriers to innovation, both externally (i.e., customer inertia) and internally (i.e., managerial inertia and technical specialization).

Marketing Strategies in the Game of Imitation

Successful innovations are typically followed by imitation. Follower firms, pursuing imitation strategies, transform a game of movement into a game of imitation. PlayStation Move and Microsoft Kinect, for example, are just the answers to the successful game of movement started by Nintendo Wii, with its innovative controllers.

To win the game of Imitation, companies must arm themselves with the ability to quickly learn what they need to know to follow the innovator, and to tap a stock of resources that is up to the task. At one end of the spectrum, imitation can produce an exact copy (see the capsule controversy for Nespresso espresso makers between Sara Lee and Nestlé). At the opposite end, imitation can be incremental, as in the Microsoft Excel worksheet, an incremental imitation of the pioneering Lotus 1,2,3. In this case, the imitation is about "doing the same things better" rather than simply "doing the same things the same way." Finally, if the original game of movement opened a completely new market, the first mover could actually welcome imitation: in fact, imitating companies would increase the overall marketing and communication efforts to support the growth of the industry.

Referring back to our previous example, by overturning existing orthodoxies the first PlayStation started a successful game of Movement. Soon, other companies followed this same strategy with the game of Imitation, starting with Microsoft. In a few years' time, with PS3 and Xbox 360 versions, the two rivals found themselves in the midst of a brutal position game, in other words, a direct, head-to-head battle once again.

Summing up, the duration of our circular model does not depend solely on the scope of innovation, but also on other players' action and reaction time, that is, again, on firms' competences and resources.

Conclusions: Marketing Strategies in the New Millennium

In the third millennium firms must be able to evolve from an era of hypercompetition to one of imagination.

Marketing strategies in an era of imagination are successful only if firms:

1. adopt a resource-based perspective;
2. focus on customer's experience, pursuing customer loyalty through 'customer's delight,' that is, exceeding customer expectations;

3. continuously regenerate their stock of market-driving capabilities, in order to avoid the trap of the 'final frontier of value' and shift continuously from games of position to games of movement;
4. coevolve with other firms, institutions, and customers, networking with them using the tremendous potentials of the new digital environments.

In order to win the competitive games of the third millennium, firms must develop marketing strategies using weapons completely different from the past. These new weapons are intangible resources and capabilities like trust, identity-signaling, creativity, integration, and networking abilities. Obviously, new weapons can only be managed by new warriors, that is, new proactive firms able to do the following:

- to deal with events and consequences of events that are yet to occur, as did Toyota when it successfully launched the Prius as far back as 1997, when ecofriendly issues were still far in the future;
- to plan for the future while governing the present, as do the most farsighted petroleum companies, which are working to reinvest today's profits from black gold to create and quickly develop new competencies in new markets;
- to redefine the destiny of the industry where they operate, as did Starbucks with its coffee shops;
- to discover new solutions to customer needs, by thinking out of the box, beyond the short-term and beyond tradition, as did Gore-Tex with fabrics that are both waterproof and 'breathable' at the same time;
- to create value by nurturing entrepreneurial spirit, teamwork, and the discipline of creativity, as does 3M, by continuously launching an amazing number of new products;
- to learn the art of coevolution, as does Apple, which owes its success to an impressive network of partners and application developers, together with customers in love with the brand;
- to reinvent itself and its markets, as Google is striving to do by turning the online search sector into an infinite editorial and advertising market.

All in all, marketing strategies will recognize the circular nature of competition: a never-ending challenge.

See also: Consumer Psychology; Corporate Finance: Financial Control; Corporate Governance; Corporate Law: Publicly Traded Corporations; International Business; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; Monetary Policy Since the 2007–2009 Financial Crisis; Transaction Costs and Property Rights.

Bibliography

Aaker, D.A., 1984. Strategic Marketing Management. Wiley, New York.
Aaker, D.A., 1991. Managing Brand Equity: Capitalising on the Value of a Brand Name. Free Press, New York.
Abell, D., 1993. Managing Dual Strategy. Free Press, New York.
Bain, J., 1956. Barriers to New Competition. Harvard University Press, Cambridge, MA.
Barney, J.S., 1986. Strategic factor markets: expectations, luck and business strategy. Management Science 32, 1230–1241.
D'Aveni, R., 1994. Hypercompetition. Free Press, New York.
Dierickx, I., Cool, K., 1989. Asset stock accumulation and sustainability of competitive advantage. Management Science 35, 1504–1511.
Grant, R.M., 2010. Contemporary Strategy Analysis. Concepts, Techniques, Applications, seventh ed. John Wiley & Sons, Oxford, UK.
Itami, H., 1987. Mobilizing Invisible Assets. Harvard University Press, Cambridge, MA.
Mason, E.S., 1939. Price and production policies of large scale enterprises. American Economic Review 29, 61–74.
Peteraf, M.A., 1993. The cornerstones of competitive advantage: a resource based view. Strategic Management Journal 14, 179–191.
Porter, M.E., 1980. Competitive Strategy: Techniques for Analyzing Industries and Competitors. Free Press, New York.
Porter, M.E., 1985. Competitive Advantage: Creating and Sustaining Superior Performance. Free Press, New York.
Prahalad, C.K., Hamel, G., 1994. Competing for the Future. Harvard Business School Press, Boston.
Reicheld, F., 2003. One number you need to grow. Harvard Business Review 81 (12), 46–54.
Reicheld, F., 2006. The Ultimate Question: Driving Good Profits and True Growth. Harvard Business School Press, Boston.
Rumelt, R., 1984. Towards a strategic theory of the firm. In: Lamb, R.B. (Ed.), Competitive Strategic Management. Prentice Hall, Englewood Cliffs, NJ.
Rumelt, R.P., 1987. Theory, strategy and entrepreneurship. In: Teece, D.J. (Ed.), The Competitive Challenge. Ballinger, Cambridge, MA.
Valdani, E., 1995. Marketing Strategico. Etas Libri, Milan, Italy.
Valdani, E., 2000. L'impresa Proattiva. McGraw-Hill, Milan, Italy.
Valdani, E., 2009. Cliente & Service Management. Egea, Milan, Italy.
Valdani, E., Arbore, A., 2012. Competitive Strategies: Managing the Present and Planning the Future. Palgrave Macmillan, Hampshire, England.
Wernerfelt, B., 1984. The resource-based view of the firm. Strategic Management Journal 5 (2), 171–180.

Markets: Anthropological Aspects

Theodore Bestor, Cornell University, Ithaca, NY, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9227–9231, © 2001, Elsevier Ltd.

Abstract

Anthropological perspectives on markets, both in historical and contemporary societies, focus on the varied institutional, political, social, and cultural forces that create and sustain market activity, and the ways in which economic life itself is inherently embedded in social and cultural relationships, that are often ignored in conventional economic analyses of market behavior. Anthropological interests in markets, often pursued through ethnographic research, examine commonalities in the principles and practices of exchange in subsistence and peasant societies as well as in centers of advanced capitalist finance. At the same time, questions about the nature of exchange and markets in societies at varied levels of complexity link ethnographic concerns with contemporary markets to historical interest in the development of markets and their impact on social and cultural development.

Markets are so routinely regarded as fundamentally economic institutions that long-standing and quite varied anthropological perspectives on them are often overlooked. Anthropological attention focuses on patterns of individual and small-group exchange relationships within specific markets, on institutional structures that organize markets, and on the social, political, and spatial hierarchies through which markets link social classes, ethnic groups, or regional societies into larger systems. Anthropological studies of markets analyze them as nodes of complex social processes and generators of cultural activity as well as realms for economic exchange. Anthropologists' interests in markets, therefore, are partially distinct from – although certainly overlapping with – the concerns of economists.

Markets and Marketplaces

The term 'market' is inherently ambiguous. Abstractly, 'market' refers to exchange organized around principles such as 'price' or 'supply-and-demand.' 'Market' may also refer to specific social relationships and frameworks through which economic transactions take place. Markets, in the first sense, are networks of economic processes and transactions which may occur without specific locations or spatial boundaries for the transactional universe. In the second sense, markets are social institutions, often located in geographically distinct places, which encompass specific social, legal, and political processes that enable economic transactions, but also extend far beyond them. Plattner (1989) makes a useful distinction:

> 'market' [is] the social institution of exchanges where prices or exchange equivalencies exist. 'Marketplace' refers to these interactions in a customary time and place. … A market can exist without being localized in a marketplace, but it is hard to imagine a marketplace without some sort of institutions governing exchanges. (p. 171)

Marketplaces embody a localized set of social institutions, social actors, property rights, products, transactional relationships, trade practices, and cultural meanings framed by a wide variety of factors including, but not limited to, 'purely economic' or 'market' forces. Of course, anthropological analyses of markets in the first sense are often ethnographically focused on marketplaces in the second sense.

Anthropological approaches to markets sometimes focus on the formal properties of exchange systems as frameworks for organizing behavior, relying on quantitative analyses of exchange relationships. However, anthropologists generally place such analyses within wider ethnographic contexts that see marketplaces as specific locations and social frameworks, characterized not only by economic exchanges in and among them, but also by their equally vital roles as arenas for cultural activity and political expression, nodes in flows of information, landmarks of historical and ritual significance, and centers of civic participation where diverse social, economic, ethnic, and cultural groups combine, collide, cooperate, collude, compete, and clash. Anthropological and sociological analyses emphasize this 'embeddedness' of markets in ongoing patterns of social organization and cultural meaning (Polanyi et al., 1957; Granovetter, 1985); that is, economic behavior is not analyzed as an autonomous sphere of human activity, but as inseparably intertwined with a wide variety of social, political, ritual, and other cultural behaviors, institutions, and beliefs.

Marketplaces as Ethnographic Sites

Marketplaces vary enormously. They differ according to the local, regional, or global scope of production, distribution, or consumption of the goods and services they trade. Retail and wholesale markets are structured quite differently around the distinct activities and social roles of consumers, producers, and traders. Some markets handle physical commodities; others trade intangible financial assets. Many marketplaces are permanent, but in some societies markets are periodic, held at regular or irregular intervals, sometimes as one stop along regional circuits for peddlers who visit specific sites on a fixed cycle. (The tiny handful of ethnographic studies of markets listed in the bibliography give only a glimpse of the wide

variety of anthropological analyses of markets, their internal structures, and their wider social and cultural contexts: e.g., Acheson, Bestor, Bohannan, Clark, Cohen, Geertz, Hertz, Meillassoux, Mintz, Polanyi, Plattner, Roseberry, Skinner, Smith, Trager.)

Many small-scale markets are socially embedded in communities, where producers and consumers deal face-to-face over the vegetables, chickens, or bolts of cloth that are the stuff of daily life, whether in a peasant community, an urban bazaar, or a farmer's market in a middle-class suburb. Local markets, as well as much more specialized ones such as urban wholesale markets of professional traders, are often organized around complex, multistranded relationships that intertwine gender, ethnicity, class, and kinship, as well as economic role.

Other very different kinds of markets (not marketplaces) embody diffuse, impersonal (and perhaps anonymous) ties among trade partners, such as in 'spot markets' where economic actors interact only through a one-time transaction, as in many real estate markets, labor markets, and global commodity markets for things such as sugar, coffee, or rubber.

Long-distance trade – both in exotic products and mundane commodities – may pass through highly specialized marketplaces that coordinate a regional or a global industry, such as Dutch flower auctions. Some long-distance markets are organized around small tightly knit communities of professional traders who transact business within networks of trust built on ethnic solidarity, such as New York's diamond exchanges or Ibadan's Hausa traders.

Markets exist along continua between the most informal sectors of society and the most highly regulated. Some markets are organized through informal or quasi-formal institutions (open and above board, but outside the realm of legal or political attention), while others are 'gray,' 'black,' or entirely illegal. Other specialized markets for professional traders are organized within tightly regulated institutional frameworks that govern access, terms of exchange, reporting requirements, or public accountability; some examples include stock markets, commodity exchanges, and other financial markets. Whether informal or formal, the frameworks of regulation that encompass the smooth functioning of any market usually mix self-regulating mechanisms created by market participants themselves with those imposed by political or legal authorities. The social, institutional construction of trade and markets is evident in the widely varied price mechanisms – bartering, bidding, haggling, setting posted prices, or negotiating contracts, as well as discounts, rebates, or kickbacks – that are established in various markets, reflecting and shaping very different balances of market power among buyers and sellers.

Markets as Institutions

Anthropologists focus ethnographically on the social structure of markets as institutional systems, the transactional behavior of market participants, and networks among trade partners or among markets (Plattner, 1985). They emphasize how economic activity is embedded in social institutions and relationships which structure solutions to economic problems (sometimes conceptualized as 'transaction costs' (Acheson, 1994)). These costs, which any market or enterprise inevitably faces, are not only direct overhead expenses on specific exchanges. More generally, the economic and social costs of exchange include those of establishing trust and reliability among trade partners, soliciting or extending credit, guaranteeing stable sources of supply, enforcing compliance with agreements, recruiting labor, distributing profits, monitoring employees, obtaining information on market conditions, creating or enforcing property rights, managing risk, and so forth (Geertz, 1978).

Various patterns of social structure that enable markets to form and economic transactions to occur are often conceptualized – by anthropologists influenced by institutional economics and sociology – in terms of 'governance structures,' the institutional structures that organize, constrain, and coordinate economic activities, that sanction some behaviors and provide incentives for others. Different governance structures – different forms of market relationships, different forms of business organization – provide different solutions to the challenges of achieving social and economic integration over the 'transaction costs' that all economic institutions must bear. Governance structures are, therefore, social institutions and systems of norms familiar to anthropologists in many other contexts and subject to similar kinds of social and cultural analyses.

Governance structures range along a theoretical continuum, from 'market governance' to 'governance by hierarchy.' In the former, an economic actor relies on the competitive forces of a spot market to obtain the goods, services, and trustworthiness it requires; in the latter, an economic actor controls goods, services, personnel, and reliability through direct ownership and administrative fiat (as in a vertically integrated industrial corporation). Midway between these extremes is governance by 'relational' or 'obligational' contracting, in which partners in ongoing exchange relationships agree, formally or informally, to do business with one another over time, relying on the strength of personal ties (such as trust) to overcome problems that may arise in the relationship because not all the terms and circumstances of trade are (or can be) specified ahead of time.

Anthropological analyses of markets often focus on the social and cultural patterns sustaining this middle form of governance, such as frameworks of self-regulation, the management of common property, the structural relationships between producers and buyers, the disposition of market power, or the political dynamics of trading communities. Other studies examine the creation of personal ties of trust and reciprocal obligation, and the microlevel transactional behavior among individual traders and other market participants. Some studies place market relationships within a broader cultural milieu of interpersonal interactions; still others examine negotiating, bargaining, or haggling as a transactional game through which traders form what Plattner calls 'equilibrating relationships.' Implicit and explicit decisionmaking models used in daily trade have been collected ethnographically to illustrate how economizing strategies are embedded in culturally or socially

specific contexts. Information costs and transactions costs have been analyzed within what Geertz (1978) calls 'bazaar economies.'

Markets and Urban Life

Throughout history, cities and markets have sustained each other, the former providing location, demand, and social context for the latter; the latter providing sustenance, profit, and cultural verve to the former. In many places, towns and marketplaces grew up together, with marketplaces as centers of economic and social life (and eventually political life as well) and as the institutions through which towns and cities were linked to their hinterlands and to other communities. Markets mediate connections and conflicts among very different segments of an economy or a society: across divisions between rural and urban, peasant and capitalist, premodern and modern, colonized and colonizing, or informal and formal sectors. These mediating roles have been examined in urban markets (in the diffuse sense, as in labor markets) and marketplaces (in the more specific sense, of geographically situated hierarchies of trade (Meillassoux, 1971)). Within commercialized economies, distribution channels and markets that connect large-scale business to small-scale family firms are another example of market linkages across social and economic sectors.

Market hierarchies have themselves been a major topic of study. Central place theory analyzes the spatial distribution of markets within hierarchies of settlements ('central places'), and within anthropology has been applied to peasant marketing systems and to interrelationships among urban markets. Alignments of trading patterns within market systems have been shown to be important indicators of a wide variety of other social, political, administrative, and ritual aspects of local, regional, and national organization. Also known as 'regional analysis,' this approach was developed in anthropology by Skinner's ethnographic and historical research on China (Skinner, 1977), and by extensive studies in Meso-America and elsewhere (Smith, 1976).

The cultural environment of trade and marketplaces is also a central aspect of urban life. In the repertoire of crucial social relationships and roles filled by urbanites, Hannerz (1980) includes 'provisioning relationships,' which necessarily involve people in exchange, within and outside of markets. In a spatial sense, marketplaces can be analyzed as a distinctive kind of urban place with economic as well as many political, social, and ecclesiastical functions, but most anthropologists situate marketplaces not simply in spatial terms but within the wider cultural milieu of a society's values, norms, and texture of relationships (Geertz, 1979).

Markets do not just organize sources of supply; they also satisfy (or create) demand and desire, as stages upon which consumption is rehearsed and displayed. Many studies of consumption take Bourdieu's perspectives on 'taste,' 'distinction,' and 'cultural capital' as points of departure for examining the cultural force of markets in shaping contemporary urban life. Studies of the logic of market capitalism and how it permeates one's experience of shopping and consumption echo Simmel's perspectives on market mentality as the quintessential condition of urban life. So too, did Redfield and Singer's earlier formulations of 'the cultural role of cities,' which placed the marketplace at the heart of the 'heterogenetic city,' a city that links itself to a wider world and, in the process, transforms the city, the rural hinterlands with which the city is interdependent, and society at large.

Markets and Globalization

More recently, sweeping transnational economic, political, and social forces have eroded the separability of societies and perhaps have disestablished the primacy of cities as nodes of exchange, but have accentuated the importance of markets. Since the early 1990s, proponents of globalization as a distinct type of social transformation have emphasized markets (in their deterritorialized sense) as central to this change. Globalization, in this view, is the expansion and ultra- integration of markets on an unprecedented scale, creating markets that incorporate societies, social sectors, and commodities that had formerly resisted or been outside market spheres of exchange. This process presumably has weakened nation-states, and has been facilitated by the increasing speed (or fluidity) of communications, transportation, media, and flows of goods, financial assets, and people, all sustained and accelerated by major technological breakthroughs in electronic media and information processing.

Of course, new patterns of global integration formed around markets are themselves nothing new. Anthropologists and sociologists have examined trade and market hierarchies as they establish linkages throughout what Immanuel Wallerstein conceptualized as the 'world system.' Within this expansion of Western European societies to incorporate most of the globe into their spheres of economic, political, and military hegemony, markets have been critical organizing principles for economic, social, political, and cultural phenomena on regional and national as well as global scales. The political economy of the contemporary world system can be seen through complex networks of 'commodity chains': the links, stages, and hands through which a product passes as it is transformed, fabricated, and distributed between ultimate producers and ultimate consumers (Gereffi and Korzeniewicz, 1994). Such chains connect far-flung components of the 'global factory,' the international division of labor among societies whose specialized niches in the world economy may center on resource extraction, low-cost fabrication, transportation efficiencies, or highly developed consumer markets.

Commodity chains are useful for understanding the widely dispersed industrial production characteristic of contemporary transnational trade as well as the fluidity of the circulation of agricultural and other commodities between producer societies on the global periphery and consumer societies of the global core. Sugar, for example, stands as a commodity at the intersection of imperialism, the colonization of the Caribbean as a region of plantation-slave societies, and Western European industrialization (Mintz, 1985). The contemporary coffee trade in the Western

Hemisphere reflects North American hyperconsumerism and the skillful marketing efforts of mass-market 'boutiques' for 'yuppie coffee,' which form new structural linkages between producer and consumer societies within systems of cultural symbolism and identity based on consumption style (Roseberry, 1996).

Against a backdrop of globalizing markets for commodities, people, assets, and images, Appadurai (1996) proposes that contemporary ebbs and flows of transnational culture be conceptualized as 'ethno- scapes,' 'technoscapes,' 'finanscapes,' 'mediascapes,' and 'ideoscapes.' Very roughly, these refer to the complicated tides and undertows of people(s), of technology, of capital, of media representations, and of political ideologies that concurrently link and divide regions of the globe. These 'scapes' resemble diverse and intersecting markets that exchange items and images across the globe, and across domains, creating value and the possibility of exchange through unexpected juxtaposition and disjuncture. Appadurai's vision of global integration (or disintegration) implies a deterritorialized world in which *place* matters little, but the fluidity of exchange is everything. These loosely coupled domains are organized around particular processes (migration, investment, representation, or consumption), and a varied repertoire of influences may travel quickly, in many directions almost simultaneously, across these domains. The center or disseminator of influence on one 'scape' may be simultaneously the periphery or recipient of influence on another. This decentralized, deterritorialized global culture is a world of many markets (broadly conceived) but few marketplaces, few specific central places of interaction. In contrast, Hannerz (1996) locates these processes (or markets) in world cities, nodes that receive, coordinate, and disseminate cultural content and creativity to locally synthesize global elements into diverse, multiple patterns of regional culture.

The trends that created this 'global cultural supermarket' (to use Stuart Hall's phrase) involve markets more than just metaphorically. The commodification of human bodies in global flows of guest workers, sex workers, and refugees involves markets of hope, desire, and misery. Global industry's ability to shift production from place to place has created markets for 'pollution credits' among different jurisdictions. The development of deep-sea diving equipment has reinvigorated debates over ownership of the seabed, and has vastly extended property rights regimes over global oceans. Electronic media have created commercial forms – online auction sites, for example – that raise familiar questions about trade partnerships, rules of trade, and principles of reliability. Global financial markets can now operate 24 hours a day, further transcending spatial and temporal (and hence political) limitations on their operations. The digitization of information into electronic form has created new forms of property rights (patents and copyrights) based on the value of – and the exchange rights inherent in – retrieval systems rather than underlying information. Common cultural heritage has been converted to 'content,' a commodity with which to fill information delivery systems. Biotechnology has created new crop species, and hence questions about the ownership of both new and existing species and the digitized information on human and other genomes as they are decoded. Increasing exploitation of natural resources raises questions about the common property rights of nations and communities over ecosystems, as well as about the ownership and marketability of indigenous knowledge about local biota, itself a source of valuable data for bioprospecting.

The fundamental issues of anthropological interest in markets and exchange (what is property? who can own it? how can it be exchanged?) are not issues simply of trying to understand small-scale, isolated societies in antiquity and the recent past. How markets are constituted, who has access to them, and how they affect the social order as a whole continue to be current issues, affected by the transformations now taking place on a global scale, creating new integrations of local and transnational market systems centered around new forms of property and new media of exchange. Anthropological analyses of these markets address cultural and social issues as fundamental as those raised by analyses of traditional patterns of exchange in peasant marketing systems. Anthropological interest in markets will continue to focus on emerging practices of capitalism as a global cultural and social system.

Bibliography

Acheson, J. (Ed.), 1994. Anthropology and Institutional Economics. Monographs in Economic Anthropology, no. 12, Society for Economic Anthropology. University Press of America, Lanham, MD.

Appadurai, A., 1996. Modernity at Large: Cultural Dimensions of Globalization. University of Minnesota Press, Minneapolis, MN.

Bestor, T.C., 1999. Wholesale sushi: Culture and commodity in Tokyo's Tsukiji Market. In: Low, S.M. (Ed.), Theorizing the City: The New Urban Anthropology Reader. Rutgers University Press, New Brunswick, NJ.

Bestor, T.C., 2001. Tokyo's Marketplace: Culture and Trade in the Tsukiji Wholesale Fish Market. University of California Press, Berkeley, CA.

Bohannan, P., 1955. Some principles of exchange and investment among the Tiv. American Anthropologist 57, 60–70.

Clark, G., 1994. Onions Are My Husband: Survival and Accumulation by West African Market Women. University of Chicago Press, Chicago.

Cohen, A., 1969. Custom and Politics in Urban Africa. University of California Press, Berkeley, CA.

Geertz, C., 1978. Bazaar economy: Information and search in peasant marketing. American Economic Review 68 (2), 28–32.

Geertz, C., 1979. Suq: The bazaar economy of Sefrou. In: Geertz, C., Geertz, H., Rosen, L. (Eds.), Meaning and Order in Moroccan Society. Cambridge University Press, New York, pp. 123–314.

Gereffi, G., Korzeniewicz, M. (Eds.), 1994. Commodity Chains and Global Capitalism. Greenwood Press, Westport, CT.

Granovetter, M., 1985. Economic action and social structure: The problem of embeddedness. American Journal of Sociology 91 (3), 481–510.

Hannerz, U., 1980. Exploring the City: Inquiries Toward an Urban Anthropology. Columbia University Press, New York.

Hannerz, U., 1996. Transnational Connections: Culture, People, Places. Routledge, London.

Hertz, E., 1998. The Trading Crowd: An Ethnography of the Shanghai Stock Market. Cambridge University Press, New York.

Meillassoux, C. (Ed.), 1971. The Development of Indigenous Trade and Markets in West Africa. Oxford University Press, London.

Mintz, S., 1961. Pratik: Haitian personal economic relationships. Proceedings of Annual Meeting American Ethnological Society.

Mintz, S., 1985. Sweetness and Power: The Place of Sugar in Modern History. Viking, New York.

Plattner, S. (Ed.), 1985. Markets and Marketing. Monographs in Economic Anthropology, no. 4. Society for Economic Anthropology. University Press of America, Lanham, MD.

Plattner, S. (Ed.), 1989. Economic Anthropology. Stanford University Press, Stanford, CA.
Polanyi, K., Arensberg, C.W., Pearson, H.W. (Eds.), 1957. Trade and Markets in the Early Empires: Economics in History and Theory. The Free Press, Glencoe, IL.
Roseberry, W., 1996. The rise of yuppie coffees and the reimagination of class in the United States. American Anthropologist 98 (4), 762–775.
Sahlins, M., 1972. Stone Age Economicss. Aldine-Atherton, Chicago, IL.
Skinner, G.W. (Ed.), 1977. The City in Late Imperial China. Stanford University Press, Stanford, CA.
Smith, C.A. (Ed.), 1976. Regional Analysis, vols. 1 & 2. Academic Press, New York.
Trager, L., 1981. Customers and creditors: Variations in economic personalism in a Nigerian marketing system. Ethnology 20 (2), 133–146.

Markets: Artistic and Cultural

Victoria D Alexander, University of Surrey, Guildford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article gives a brief overview of the development of markets for the sale of fine arts, the development of nonprofit organizations for the distribution of cultural forms, and the growth of for-profit cultural industries. Issues that affect both nonprofit cultural organizations and the profit-seeking cultural industries are discussed. Key concepts involve the effects on art and culture of uncertainty, the laws of supply and demand, and the responsiveness of markets to buyers' and sellers' desires. Cultural markets are currently in flux due to changes wrought by the Internet.

Introduction

A market is a forum for exchange between buyers and sellers. In the artistic and cultural realms, the term refers to a system for the production and distribution of the arts or culture. Such systems differ from a purely atomistic market in that actors (buyers, sellers, and intermediaries) include both organizations and individuals. All major cultural markets include institutional players in the form of profit-seeking businesses or nonprofit agencies (Caves, 2002).

Many approaches to artistic and cultural markets draw strong distinctions between high and popular culture. High culture, such as opera, theater, symphony, and the visual arts, is often distributed by nonprofit organizations, and is said to be oriented toward the artist. Popular culture, by contrast, is often distributed by profit-making organizations and is said to be oriented toward the consumer (see Gans, 1974). Changes in society and in our understanding of culture have blurred the distinction between fine and popular, and some fine arts are notably distributed through profit-seeking means. The laws of supply and demand apply regardless of whether the organizations involved are profit-seeking. Characteristics of the production system – levels of uncertainty, ambiguity, and risk; market concentration and the level of competition; the ratios of buyers, sellers, and intermediaries – shape not only factors such as the price of an object or performance, but also content of the art itself (Crane, 1992).

History

White and White's (1993[1965]) important work, a sociological study of the transformation of the French art world, shows that the market for fine arts emerged at the end of the nineteenth century. Previous systems in the arts included patronage, guild, and academic systems, which supported the production of art, but it was not until this time that the practice of exchanging of artworks for money in a marketplace became common. By the latter half of the nineteenth century, the French academic system had been successful for more than two centuries. It failed when the number of aspiring artists in Paris vastly outnumbered the rewards (prizes and academic appointments) available through the academy. The surplus artists wanted an outlet for their work and a means to make a living, so they turned to dealers, people who would sell their works directly to the public. In this way, White and White argue that the dealer–critic system emerged to supplant the academic one. Critics were a key component in this new system, as they worked with dealers and artists to create an aesthetic system that justified the new forms of painting. In conveying these ideas to the public, critics helped to create demand for an artist's work.

Proper markets for the popular arts also emerged, more or less, at the end of the nineteenth and the beginning of the twentieth century (Starr, 2004). These cultural markets arose as the cultural industries developed. Initially criticized, most notably by the Frankfurt School (Adorno, 1991), for promoting lowest-common-denominator pabulum, cultural industries grew, along with other industries, in their ability to capture niche markets and to create value from innovation (Lampel et al., 2000; Lopes, 1992).

White and White demonstrate that the system through which French painting was produced influenced the content of the artworks themselves. The academic system rewarded individual paintings exhibited in salons. The dealer–critic system, on the other hand, had different requirements. Dealers needed to sell a continuing stream of paintings, and as a result, had an interest in differentiating their artists' products by referring to each artist's unique style. Paintings were purchased by middle-class buyers, as well as more wealthy patrons, and so were more likely to be hung in houses than in museums. This meant that small, easel paintings were preferred to monumental works and that decorative scenes of landscape and everyday life were preferred to more complicated allegorical or historical works. Each of these requirements, along with a number of others, dovetailed with the new movement in painting, Impressionism. Though clearly there is more to artistic change than just the particular mechanisms that reward artists, White and White demonstrate the influence that these institutional arrangements had on aesthetic possibilities.

In an important article, Peterson and Berger (1975) showed how concentration in the music industry affected the innovation and diversity of hit songs. Indeed, the production of culture approach (Peterson and Anand, 2004) has strongly demonstrated effects of various market arrangements on aesthetic possibilities and the content of art.

In a classic study of market effects on artistic and cultural organizations, the economists Baumol and Bowen (1966) focused on the performing arts. Their much cited finding is that over time the performing arts become increasingly expensive relative to the cost of living, because the number of personnel needed for performances – the musicians in an orchestra, say, or actors in a play – is difficult to reduce. As a result, the performing arts do not see the productivity gains realizable in manufacturing industries. Performing arts, like other service industries, face increasingly expensive payroll bills. Thus, the performing arts always struggle to find external funding, charge higher ticket prices, or suffer cuts.

The studies by Baumol and Bowen (1966), Peterson and Berger (1975), and White and White (1993[1965]) were important in setting the stage for serious study of the distribution mechanisms and economics of the arts. Since their studies were published, there has been a convergence of interest in artistic and cultural markets in a variety of disciplines. Scholarship in this area accelerated through the 1980s (Frey and Pommerehne, 1989; Heilbrun and Gray, 1993), and explicit attention to artistic and cultural markets has received scholarly attention since the turn of the millennium (Caves, 2002; Hesmondhalgh, 2007).

Uncertainty

In an influential article, Hirsch (1972) looked at the distribution of popular arts by profit-seeking organizations in cultural industries. Focusing on popular books and records, he demonstrated that cultural industries are characterized by uncertainty. Record companies and publishers wished to make a profit, but they were not sure which books and albums would succeed. They responded to this situation in a variety of ways, most notably through the strategy of overproduction. They produced many more albums and books than they could possibly sell, with the full knowledge that most would not break even. A small proportion would earn a modest return, and a very few would be phenomenally successful. These last, the blockbusters (or hits), generate enough income to cover the losses on the failures and to make a profit for the company. A second strategy taken by firms was to look at what had been recently successful and to overproduce cultural objects in the same genre or format. In this way, change in cultural industries is often trendy, following fads and fashion.

A key concept in Hirsch's work is the idea of a gatekeeper, a person, or an organization who allows only certain objects into a system. Gatekeepers are important for the content of artwork, as their guesses on what will succeed, their personal tastes, and even their unquestioned cultural assumptions can affect what they accept and reject (Griswold, 1992). Gatekeepers do not directly mold artworks, but in deciding what works to produce or distribute, they determine the range and therefore the content of the artwork that actually reaches the public.

Art Worlds

The work of sociologist Howard Becker is extremely important in cultural sociology. Becker's (1982) key idea is that art is created in rich social networks called 'art worlds.' The characteristics of art worlds create a pattern of constraint and possibility that affects the production and distribution of art, and therefore the artists and their works. Becker's discussion of markets falls under his consideration of distribution systems. He suggested that distribution systems matter, especially to artistic reputation. What is not distributed cannot gain recognition or be remembered. He identifies three key mechanisms by which the arts are supported. The first, self-support, where an artist relies on a day job, a spouse, or an inheritance for monetary needs and produces art 'on the side,' may be the most purely atomistic market in arts distribution, or it may lack any market element. As Becker notes, self-supported art is often not distributed at all, let alone widely. Patronage – whereby a buyer commissions art, either a specific work or a part of the artist's general output – is the second mechanism, and has a clear market component. Patronage sounds genteel, and requires that only one buyer be pleased. It is a form of support that can foster innovation – if the buyer values this – but clearly only within the bounds of the buyer's taste and inclinations.

The third mechanism is public sale, and Becker discusses three types of public sale. Dealers act as intermediaries between artists and buyers where the artwork is a physical object (a painting, sculpture, or installation). Impresarios mediate between artists and audiences in the performing arts (theater, symphony, dance, or experimental performance art). Culture industries are comprised of organizations that link artists who produce works that exist in multiple copies (books, records, movies, film, and television shows) to a large number of consumers.

The 'High Culture' Model

DiMaggio (1982, 1992) shows that, during the late nineteenth and early twentieth century in America, art forms that are considered 'high culture' came to be distributed via nonprofit organizations. As a legal form, nonprofit organizations are constituted 'for the public good,' and because they offer public benefits, they are entitled to tax relief. DiMaggio (1982) shows that, in the late nineteenth century, 'cultural entrepreneurs' founded nonprofit orchestras and art museums in order to advance their interests by embedding their own cultural forms into a situation that conferred both a favorable tax regime and prestige. In the early twentieth century, opera and ballet companies and certain theaters incorporated as nonprofit organizations (DiMaggio, 1992). Nonprofit performing arts organizations offer an important outlet, a market we could say, for the sale of live performance. While museums act as repositories of the visual arts, rather than venues for exchange of artworks, they also provide an opportunity to experience art. Importantly, nonprofit organizations allow champions of each art form to create a boundary between a high-status version of the art form and its more popular and commercial relatives. In contrast, a more pure marketplace works against distinctions in culture, as it "drives entrepreneurs to elide aesthetic distinctions in order to create larger audiences and discourages canon formation by providing incentives for presenters to differentiate their products" (DiMaggio, 1992: 43–44).

The Marketization of Art Museums

Museums play a number of key roles in the visual art world; notably, they are the venue in which most ordinary viewers confront original art. American art museums gain income from grants for special exhibitions, especially from corporations and government agencies, alongside traditional patronage. In addition, American museums rely on earned income through general admission charges, gate receipts for special shows, and the sale of merchandise, books, and souvenirs. This funding situation has important implications for the art that is displayed in museums and for museum audiences (Alexander, 1996). Interestingly, art museums across the globe are facing similar pressures, as they are increasingly 'privatized' (Boorsma et al., 1999), that is, moved away from direct state support to a 'mixed economy' of fund-raising from private sources and earned income, most of which is linked, directly or indirectly, to the marketplace. The term marketization is used to indicate that a type of culture once produced and distributed in nonmarket settings has moved into market (or market-approximating) ones. Broadly, the term also indicates that a neoliberal way of thinking has pervaded settings that used to be 'pure.'

Marketized cultural organizations, such as museums, must pay attention to audiences. Not only do audiences bring in money directly in terms of entrance fees and sales, they also influence project funding. Both corporations and government agencies are concerned with the size of the audience their funded projects might attract (Alexander, 1996). This has led museums to operate, in essence, as public sales organizations, in that their exhibitions must appeal to the people who will 'buy' them. Museum curators remain committed to scholarly exhibitions and are less interested in popular ones, but the latter are easier to fund and the public likes them, so curators try hard to mount shows that are both scholarly and popular.

An interesting change has taken place in art museums over the last few decades: a shift in the basic assumptions of those in charge of museums, from a more scholarly approach to a more managerial one. Collecting, conservation, and connoisseurship remain mainstays of art museums. But a focus on business-world concerns – marketing a product, value for money, and the like – has pervaded museums. Some, particularly those with a business or economics background, see this in a positive light, as evidence that museums have entered the modern world and will finally be managed in a sensible way. Others, especially those vested in the older values system, see this as a corruption of museums, debasing the art and bringing crass considerations like the bottom line to a sacred arena where monetary concerns should be left aside.

Visual Art Markets

In the visual arts, separate markets exist for consecrated works (old masters and more recent artists with strong reputations), the contemporary avant-garde, contemporary decorative/figurative art, and art made by anonymous masters from developing countries (tribal art). An international market exists for each of these types of art, except the decorative arts, which is the most varied market and is sold by dealers in national, regional, or local markets, (Plattner, 1996; O'Neil, 2008) or even in such outlets as craft fairs and street stalls (Paulsen and Staggs, 2005).

Unlike most artistic markets, the market for consecrated works (those by 'old masters') is characterized by undersupply – there are fewer works of art than buyers wanting them. As a result, such works are often sold at auction where competing buyers can bid up prices to extraordinary heights (Thornton, 2008). But prices can differ substantially from the estimates of the auction houses, because social expectations and the dynamics of the auction floor, not just the laws of supply and demand, affect bids (Smith, 1990). These factors add a degree of uncertainty to the secondary (auction) market (Thompson, 2008).

Consecrated works are almost always pieces that are being 'resold' by an owner, which is why the market is called the 'secondary' art market. Artists in the secondary market have already been market tested during their previous sales history. Uncertainty is higher in the avant-garde setting, in the 'primary' market where art is sold via dealerships, both for artists and dealers, as new artist are, obviously, untested in the marketplace. The prestigious avant-garde markets are concentrated in a few world-class cities, notably New York, London, and Paris. A problem for dealers – the same faced by cultural industry executives – is to choose artists who will succeed. There are huge numbers of artists hoping to be discovered and shown, and it is difficult to predict those whom posterity will remember. The solution is to promote a relatively large number of artists as compared with the numbers who will succeed (a strategy of 'overproducing'). However, relatively few artists break through even to the level of success indicated by representation by an important dealer. As with other art markets, the avant-garde market is characterized by a 'winner-take-all' reward system where a few superstars receive the lion's share of the returns (see Rosen, 1981). This means a few contemporary artists will be phenomenally successful and become rich, a modest number will just scrape a living and most will have to supplement their incomes elsewhere.

Dealers may also promote artists who work in styles similar to those that have recently been successful. Unlike cultural industry executives, they do not profess this mimicking strategy, as it goes against central tenets of avant-garde art, namely, the sanctity of artistic choice, free from the interference of dealers. Velthuis (2005) points out that art dealers do not openly discuss prices, as to do so would undermine the pretense that art is completely separated from the marketplace and commercial concerns. Art dealers, nevertheless, are clearly commercial and profit-seeking.

The oversupply of artists leads to suppressed prices for art in general, and art dealers need to set prices in the absence of objective criteria for doing so. Clearly dealers have an incentive to limit market entry and to keep prices at a certain level. Their techniques include discouraging their artists from underselling the dealer, convincing potential buyers that gallery prices for emerging artists are a fair expression of their future potential, and excluding buyers who might 'flip' a work at auction and thereby undermine the dealer's pricing system. Velthuis (2005) discusses the many ways that the laws of supply and demand are subverted in the international visual arts market.

The largest, headline-grabbing sums in the visual arts belong to the old master and avant-garde markets. Figurative

and decorative arts are devalued relative to avant-garde art. The aesthetic system, prestigious dealers, and relative centralization that provide entry barriers in the avant-garde world do not exist in the decentralized figurative art market. As a result, prices are lower in this market. However, as Martorella (1990) demonstrates, corporations wishing to buy prestigious art also prefer noncontroversial, figurative art. Their presence in the art market has helped legitimize some forms of figurative art. American corporations' interest in collecting regional artists has also reinforced the decentralization in the figurative art market.

The market for tribal art (previously referred to as 'primitive' art) is also smaller and attracts lower purchase prices. The operation of this market raises questions about the exploitation of artists from developing countries by dealers and patrons in the first world (Price, 1989). Today, third-world masters are usually named and their work given the title of art, rather than cast anonymously as an expression of tribal visions. Issues such as cultural imperialism are less salient in markets for avant-garde and figurative art (though certainly not nonexistent), but play a key role in markets for other types of culture, notably film and television.

Moulin (1987) argues that the economic value of an artwork and its artistic value are inseparable and synonymous. An economist would argue, then, that the market was efficient in that it properly prices works based on aesthetic value. But it is also possible to argue that the economic value, rather than reflecting aesthetic value, creates it by attracting the attention of art critics and the cultural cognoscenti and setting in motion the theorizing necessary to underpin an artist's style.

As Menger (1999) demonstrates for a wide variety of artists in different cultural industries, more talent exists than do outlets for it. As a result, artistic careers are risky for artists. Most support themselves economically with a second job. The situation, as with other contingent labor markets (where employers hire workers only when work is available) provides a great deal of flexibility to dealers, and the other cultural organizations that rely on contingent labor, but this arrangement shifts costs from the organization to the individual.

Can Culture Survive the Marketplace?

An old but still important question is how market systems, as opposed to other forms of support, affect culture. Given the pervasive oversupply of artworks and artists, it follows that in artistic and cultural markets sellers must attract buyers. Art that no one wants simply cannot be supported by the market. Moreover, markets imply profits. In this case, not only must buyers exist, but also there must be enough of them for sales to cover production and overhead costs and to leave a surplus. The larger the audience needed, the more popular the product must be, and as a corollary, the lower the common denominator. Indeed, as cultural critics for more than a century have lamented the degraded nature of mass culture, contemporary critics worry about the declining state of art.

Not-for-profit enterprises may seem a good alternative to commercial industries, but the need to attract an audience is not limited to for-profit enterprises. Not-for-profit organizations must also reach sufficient audiences while facing an inexorable upward spiral of costs, as Baumol and Bowen showed nearly 50 years ago. These pressures, along with the privatization of arts funding, lead nonprofit enterprises to approximate profit-seeking systems in all but their requirement to return dividends to shareholders.

DiMaggio (1986) argues that large, established arts institutions can survive the marketplace. With proven repertoires of established works well accepted by the public, they can generate significant earned income. They may subsidize an innovative program with known crowd-pleasers, such as *The Nutcracker*, *The Messiah*, or Monet's *Water Lilies*. In addition to their ability to attract the 'middlebrow,' they have the organizational size and expertise to raise funds for such projects.

Since innovative, experimental works may be least likely to survive the marketplace, some have argued that there is a societal interest in subsidizing a supply of unpopular (or rather, not yet popular) artists through government funding. While generous state sponsorship can have a positive effect on innovation and diversity in the arts, not to mention making the lives of artists easier, it is worth noting that results depend on national politics (Alexander and Rueschemeyer, 2005). For instance, art in Soviet societies was strongly supported by the state – and highly constrained, and political controversies in the US served to narrow the type of art and artists sponsored by the federal government.

New Technologies

A key point of uncertainty for the future of artistic and cultural markets is the influence of new technologies, especially the Internet. The music industry, for instance, has been transformed by the digital age, and new business models are in the offing (Kusek and Leonhard, 2005). Gatekeepers within cultural industry systems have been powerful intermediaries – and barriers – between artists and the public. Today, artists reach out to the public via the Internet. There have been notable successes via this 'direct' route, but gatekeepers, of one sort or another, are likely to remain important, if only by providing help to consumers (buyers) in sifting through nearly endless possibilities for arts products and experiences.

Anderson (2006) describes the new cultural economy in which mass markets aimed toward 'hits' is joined, or even supplanted, by the 'long tail' in which cultural products attracting very small audiences can be marketed. In other words, technology drives niche marketing to the extreme, allowing profitable sales of even a few copies of a cultural product. While there is not much money in each sale in the long tail, the number of sales can add up to be enormously profitable. Amazon takes a long-tail approach, and makes more than half of its book sale profits from titles not sold in most mass-market bookstores (Anderson, 2006). Apple's iTunes and streaming services, such as Spotify and Netflix, also aim to make profits in the long tail.

Other cultural institutions are increasingly making use of digital technology. Most museums provide Web pages with digitized images from their collection, and Google's Cultural Institute and its associated Google Art Project allow viewers to take digital tours, via Street View technology, of galleries

in the world's great museums. Performing arts organizations offer 'trailers' of upcoming work on their home pages. The Metropolitan Opera simulcasts high-definition recordings of live performances to cinemas across the globe. While it is unclear what directions these changes will take nonprofit organizations and cultural industries, the implications for arts markets, not to mention artists, artworks, and audiences, are potentially enormous.

Summary

The scope for research in the area of artistic and cultural markets is vast. Even leaving aside artistic labor markets, the topic encompasses such areas as the international market for visual arts; auctions of contemporary works of art and of works by old masters; the funding dynamics of such arts organizations as museums or symphonies; the Hollywood film production and distribution network; film production elsewhere in the world, such as India's Bollywood; book publishing; television broadcast; the international flow of programing; music; radio; digital games; and much more from across the fine and popular arts. Tourism, visits to both heritage sites and entertainment centers like Disneyland, could be considered part of the global cultural market. Further, many consumer products, from soft drinks to jeans, are items of popular culture, and it would be easy to include an array of consumer markets as part of the cultural marketplace. Indeed, some theorists of consumerism would argue that all forms of contemporary purchasing belong to the cultural realm.

This article adopted a more limited view of the huge array of artistic and cultural markets, focusing on markets for the distribution of fine and popular arts. Although it could only scratch the surface of the subject, it highlighted the key characteristics of the means by which the fine and popular arts are distributed to consumers, either through cultural industries or nonprofit enterprise. Some more specific examples from the visual arts were provided. Most of the effects of the market on visual arts have complements in other artistic and cultural markets, though specific details will naturally vary from one art form to another. The main line of agreement among scholars is that, in contemporary Western societies, the arts are profoundly shaped by their market characteristics. Market and marketized settings in art and culture are more responsive to the desires of the buyers than those of the sellers; consequently, artists who do not gear their work toward audiences do not sell well, and cultural organizations that do not offer popular fare, or that are unable to tap into specialized, niche markets, also pay the price.

See also: Artists, Competition and Markets; Avant-Garde Art and Artists; Cultural Mediators and Gatekeepers; Cultural Policy Regimes in Western Europe; Culture and Economy; Culture and Institutional Logics; Culture, Production of: Prospects for the Twenty-First Century; Culture, Production of; Internet and Culture; Organizations and Culture; Symbolic Boundaries.

Bibliography

Adorno, T.W., 1991. The Culture Industry: Selected Essays on Mass Culture. Routledge, London.
Alexander, V.D., 1996. Pictures at an exhibition: conflicting pressures in museums and the display of art. American Journal of Sociology 101, 797–839.
Alexander, V.D., Rueschemeyer, M., 2005. Art and the State: The Visual Arts in Comparative Perspective. Palgrave Macmillan, London.
Anderson, C., 2006. The Long Tail: How Endless Choice Is Creating Unlimited Demand. Random House, London.
Boorsma, P.B., van Hemel, A., van der Wielen, N. (Eds.), 1999. Privatization and Culture: Experiences in the Arts, Heritage and Cultural Industries in Europe. Kluwer, Amsterdam.
Baumol, W.J., Bowen, W.G., 1966. Performing Arts: The Economic Dilemma. Twentieth Century Fund, New York.
Becker, H.S., 1982. Art Worlds. University of California Press, Berkeley, CA.
Caves, R.E., 2002. Creative Industries: Contracts between Art and Commerce. Harvard University Press, London.
Crane, D., 1992. The Production of Culture. Sage, London.
DiMaggio, P.J., 1982. Cultural entrepreneurship in nineteenth-century Boston: the creation of an organizational base for high culture in America. Media. Culture and Society 4, 33–50.
DiMaggio, P.J. (Ed.), 1986. Nonprofit Enterprise in the Arts. Oxford University Press, New York.
DiMaggio, P.J., 1992. Cultural boundaries and structural change: the extension of the high culture model to theater, opera and dance, 1900–1940. In: Lamont, M., Fournier, M. (Eds.), Cultivating Differences. University of Chicago Press, Chicago, IL, pp. 21–57.
Frey, B.S., Pommerehne, W.W., 1989. Muses and Markets: Explorations in the Economics of the Arts. Basil Blackwell, Oxford.
Gans, H.J., 1974. Popular Culture and High Culture. Basic Books, New York.
Griswold, W., 1992. The writing on the mud wall: Nigerian novels and the imaginary village. American Sociological Review 57, 709–724.
Heilbrun, J., Gray, C.M., 1993. The Economics of Art and Culture: An American Perspective. Cambridge University Press, New York.
Hesmondhalgh, D., 2007. The Cultural Industries, second ed. Sage, London.
Hirsch, P.M., 1972. Processing fads and fashions: an organization-set analysis of cultural industry systems. American Journal of Sociology 77, 639–659.
Kusek, D., Leonhard, G., 2005. The Future of Music: Manifesto for the Digital Music Revolution. Berklee Press, Boston, MA.
Lampel, J., Lant, T., Shamsie, J., 2000. Balancing act: learning from organizing practices in cultural industries. Organization Science 11, 263–269.
Lopes, P.D., 1992. Innovation and diversity in the popular music industry, 1969–1990. American Journal of Sociology 57, 56–71.
Martorella, R., 1990. Corporate Art. Rutgers University Press, New Brunswick, NJ.
Menger, P.M., 1999. Artistic labor markets and careers. Annual Review of Sociology 25, 541–574.
Moulin, R., 1987. The French Art Market: A Sociological View (A. Goldhammer, Trans.). Rutgers University Press, New Brunswick, NJ.
O'Neil, K.M., 2008. Bringing art to market: the diversity of pricing styles in a local art market. Poetics 36, 94–113.
Paulsen, K.E., Staggs, K., 2005. Constraint and reproduction in an amateur craft institution: the conservative logic of the county fair. Poetics 33, 134–155.
Peterson, R.A., Anand, N., 2004. The production of culture perspective. Annual Review of Sociology 30, 311–334.
Peterson, R.A., Berger, D.G., 1975. Cycles in symbol production: the case of popular music. American Journal of Sociology 40, 158–173.
Plattner, S., 1996. High Art Down Home: An Economic Ethnography of a Local Art Market. University of Chicago Press, Chicago, IL.
Price, S., 1989. Primitive Art in Civilized Places. University of Chicago Press, Chicago, IL.
Rosen, S., 1981. The economics of superstars. American Economic Review 75, 845–858.
Smith, C.W., 1990. Auctions: The Social Construction of Value. University of California Press, Berkeley, CA.
Starr, P., 2004. The Creation of the Media: Political Origins of Modern Communications. Basic Books, New York.
Thompson, D., 2008. The $12 Million Stuffed Shark: The Curious Economics of Contemporary Art and Auction Houses. Aurum, London.
Thornton, S., 2008. Seven Days in the Art World. Granta, London.
Velthuis, O., 2005. Talking Prices: Symbolic Meanings of Prices on the Market for Contemporary Art. Princeton University Press, Princeton, NJ.
White, H.C., White, C.A., 1993[1965]. Canvasses and Careers: Institutional Change in the French Painting World. University of Chicago Press, Chicago, NJ.

Markov Chain Monte Carlo Methods

Peter Müller, Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 14, pp. 9236–9240, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Markov chain Monte Carlo (MCMC) methods use computer simulation of Markov chains in the parameter space. The Markov chains are defined in such a way that the posterior distribution in the given statistical inference problem is the asymptotic distribution. This allows use of ergodic averages to approximate the desired posterior expectations. Several standard approaches to define such Markov chains exist, including Gibbs sampling, Metropolis–Hastings, and reversible jump. Using these algorithms it is possible to implement posterior simulation in essentially any problem which allows pointwise evaluation of the prior distribution and likelihood function.

Introduction

In Bayesian statistics the posterior distribution $p(\psi|y)$ contains all relevant information on the unknown parameters ψ given the observed data y. All statistical inference can be deduced from the posterior distribution by reporting appropriate summaries. This typically takes the form of evaluating integrals

$$J = \int f(\psi)p(\psi|y)d\psi \qquad [1]$$

of some function $f(\psi)$ with respect to the posterior distribution. For example, point estimates for unknown parameters are given by the posterior means, that is, $f(\psi) = \psi$; prediction for future data $\tilde{y}$ is based on the posterior predictive distribution $p(\tilde{y}|y) = \int p(\tilde{y}|\psi, y)p(\psi|y)d\psi$, that is, $f(\psi) = p(\tilde{y}|\psi, y)$, etc. The problem is that these integrals are usually impossible to evaluate analytically. And when the parameter is multidimensional, even numerical methods may fail.

Since the 1980s a barrage of literature has appeared concerned with the evaluation of such integrals by methods collectively known as Markov chain Monte Carlo (MCMC) simulation. The underlying rationale of MCMC is to set up a Markov chain in ψ with ergodic distribution $p(\psi|y)$. Starting with some initial state $\psi^{(0)}$ we simulate M transitions under this Markov chain and record the simulated states $\psi^{(j)}$, $j = 1, \ldots, M$. The ergodic sample average

$$\widehat{J} = \frac{1}{M}\sum_{j=1}^{M} f\left(\psi^{(j)}\right) \qquad [2]$$

converges to the desired integral J (subject to some technical conditions), that is, $\hat{J}$ provides an approximate evaluation. The art of MCMC is to set up a suitable Markov chain with the desired posterior as stationary distribution and to judge when to stop simulation, that is, to diagnose when the chain has practically converged.

In many standard problems it turns out to be surprisingly easy to define a Markov chain with the desired stationary distribution. We will review the most important approaches in this article (for general principle of Monte Carlo simulation, including independent Monte Carlo simulation, see Monte Carlo Methods and Bayesian Computation: Overview).

The Gibbs Sampler

Consider a variance components model $y_{ij} = \theta_i + e_{ij}$, $i = 1, \ldots, K$ and $j = 1, \ldots, J$, for data y_{ij} from K groups with J observations in each group. Assume independent normal errors $e_{ij} \sim N(0, \sigma_\theta^2)$ and a normal random effects model $\theta_i \sim N(\mu, \sigma_\theta^2)$ (Gelfand et al.1990). We assume that $\theta = (\theta_1, \ldots, \theta_k)$, (μ, σ_θ^2) and σ_e^2 are *a priori* independent with $p(\sigma_\theta^2) = IG(a_1, b_1), p(\mu|\sigma_\theta^2) = N(\mu_0, \sigma_\theta^2)$, and $p(\sigma_e^2) = IG(a_2, b_2)$. Here we use $N(m, s^2)$ to indicate a normal distribution with moments m, s, and $IG(a, b)$ to indicate an inverse gamma distribution with parameters a and b. Let $y = (y_{ij}, i = 1, \ldots, K, j = 1, \ldots, J)$ denote the data vector. It can be shown that the conditional posterior distributions $p(\sigma_\theta^2|y, \mu, \theta, \sigma_e^2)$ and $p(\sigma_e^2|y, \mu, \theta, \sigma_\theta^2)$ are inverse gamma distributions, and $p(\mu|y, \theta, \sigma_\theta^2, \sigma_e^2)$, and $p(\theta|y, \mu, \theta, \sigma_\theta^2, \sigma_e^2)$ are normal distributions.

To estimate posterior moments of the type (1) we define a Markov chain in $\psi = (\mu, \theta, \sigma_e^2, \sigma_\theta^2)$. Denote with $\psi^{(t)} = (\mu^{(t)}, \theta^{(t)}, \sigma_e^{2(t)}, \sigma_\theta^{2(t)})$ the state vector of the Markov chain after t transitions. Given the nature of a Markov chain, all we need to define is the transition probability, that is, given a current value for $\psi^{(t)}$, we need to generate a new value $\psi^{(t+1)}$. We do so by sampling from the complete conditional posterior distributions for $\mu, \sigma_e^2, \sigma_\theta^2$ and θ

a. $\mu^{(t+1)} \sim p(\mu|y, \theta^{(t)}, \sigma_e^{2(t)}, \sigma_\theta^{2(t)})$
b. $\theta^{(t+1)} \sim p(\theta|y, \mu^{(t+1)}, \sigma_e^{2(t)}, \sigma_\theta^{2(t)})$
c. $\sigma_e^{2(t+1)} \sim p(\sigma_e^2|y, \mu^{(t+1)}, \sigma_e^{(t+1)}, \sigma_\theta^{2(t)})$
d. $\sigma_\theta^{2(t+1)} \sim p(\sigma_\theta^2|y, \mu^{(t+1)}, \theta^{(t+1)}, \sigma_e^{2(t+1)})$

Steps (a) through (d) define a Markov chain $\psi^{(t)}$ which converges to $p(\mu, \theta, \sigma_e^2, \sigma_\theta^2|y)$, as desired. Ergodic averages of the type $\widehat{J} = 1/M \sum f(\psi^{(t)})$ provide numerical evaluations of any desired posterior integral J.

The described Markov chain Monte Carlo simulation is a special case of a Gibbs sampler. In general, let $\psi = (\psi_1, \ldots, \psi_p)$ denote the parameter vector. The Gibbs sampler proceeds by iteratively, for $j = 1, \ldots, p$, generating from the conditional posterior distributions

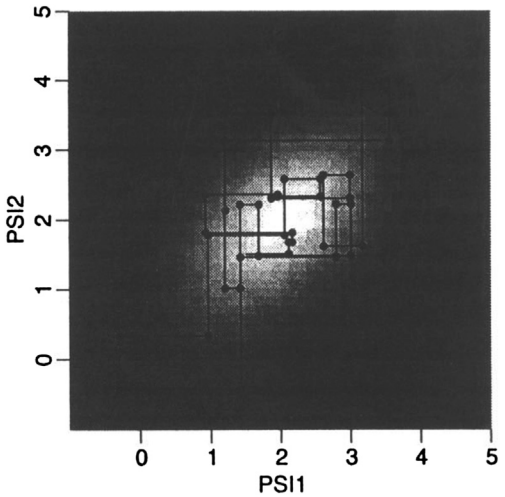

Figure 1 Gibbs sampler. The grey shades show a bivariate posterior distribution $p(\psi_1, \psi_2|y)$. The connected points show the parameter values $\psi^{(t)}$ generated in $M = 40$ transitions of the MCMC simulation. The transition probabilities in the Gibbs sampler are the full conditional posterior distributions (3), leading to the piecewise horizontal and vertical trajectories seen in the figure. Each horizontal line segment corresponds to generating a new value $\psi_1^{(t+1)} \sim p(\psi_1|y, \psi_2^{(t)})$. Each vertical line segment corresponds to generating $\psi_2^{(t+1)} \sim p(\psi_2|y, \psi_1^{(t+1)})$.

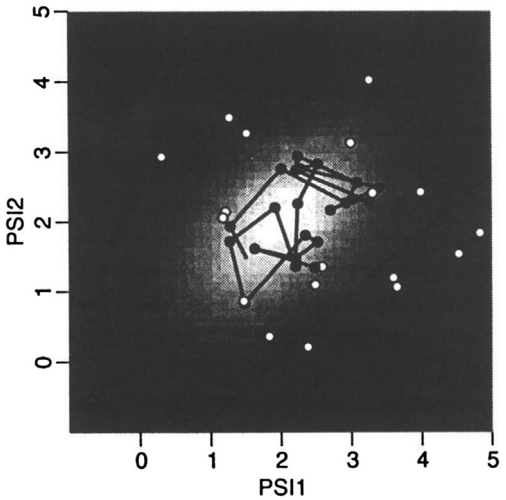

Figure 2 Metropolis sampler. The grey shades show a posterior distribution $p(\psi|y)$. The connected solid points show the parameter values $\psi^{(t)}$ generated in $M = 40$ transitions of a Metropolis chain with bivariate normal proposals $\tilde{\psi} \sim N(\psi, 0.75I)$, where I denotes the 2×2 unit matrix. The empty circles show generated proposals $\tilde{\psi}$ which were rejected using the acceptance probabilities (4). Compare with Figure 1 which shows a Gibbs sampler with an equal number of transitions.

$$\psi_j^{(t+1)} \sim p\left(\psi_j|\psi_1^{(t+1)}, \ldots, \psi_{j-1}^{(t+1)}, \psi_{j+1}^{(t)}, \ldots, \psi_p^{(t)}, y\right) \qquad [3]$$

If practicable it is advisable to generate from higher dimensional conditionals.

Figure 1 illustrates the Gibbs sampling algorithm. The figure shows simulated parameter values for a hypothetical bivariate posterior distribution $p(\psi_1, \psi_2|y)$.

The seminal paper by Gelfand and Smith (1990) and the companion paper by Gelfand et al. (1990) popularized the Gibbs sampler for posterior simulation in a wide class of important problems. Many earlier papers used essentially the same method in specific problems. For example, a special case of the Gibbs sampler occurs in problems with missing data. In many problems, the actually observed data y can be augmented by missing data z in such a way that simulation from $p(\psi|y, z)$ and $p(z|\psi, y)$ can be implemented in computationally efficient ways, even when simulation from the original posterior distribution $p(\psi|y)$ is difficult. Tanner and Wing (1987) propose what is essentially a Gibbs sampler for the augmented posterior distribution $p(\psi, z|y)$. Geman and Geman (1984) proposed the Gibbs sampler for posterior simulation in a spatial model with a Markov random field prior.

The Metropolis–Hastings Algorithm

The Gibbs sampler owes some of its success and popularity to the fact that in many statistical models the complete conditional posterior distributions $p(\psi_j|\psi_i, i \neq j, y)$ take the form of some well-known distributions, allowing efficient random variate generation. But there remain many important applications where this is not the case, requiring alternative MCMC schemes. Possibly the most generic such scheme is the Metropolis scheme (Metropolis et al. 1953). Consider generating from a posterior distribution $p(\psi|y)$. Denote with ψ the current state of the Markov chain. One transition is defined by the following steps:

a. Generate a proposal $\tilde{\psi}$ from some proposal generating distribution $q(\tilde{\psi}|\psi)$. The choice of the proposal distribution $q(\cdot)$ is discussed below.
b. Compute

$$a(\psi, \tilde{\psi}) = \min\left\{1, \frac{p(\tilde{\psi}|y)}{p(\psi|y)} \cdot \frac{q(\psi|\tilde{\psi})}{q(\tilde{\psi}|\psi)}\right\} \qquad [4]$$

c. With probability a replace ψ with the proposal $\tilde{\psi}$. Otherwise, leave ψ unchanged.

Figure 2 illustrates the algorithm. The figure shows the proposals $\tilde{\psi}$ and the (accepted) states $\psi^{(t)}$ for the first 40 iterations of a Metropolis chain simulation for a hyptothetical bivariate posterior distribution. The choice of the proposal distribution $q(\tilde{\psi}|\psi)$ is essentially arbitrary, subject only to some technical constraints. Using a symmetric proposal distribution with $q(\tilde{\psi}|\psi) = q(\psi|\tilde{\psi})$, for example a normal centered at ψ, has the practical advantage that the ratio of proposal distributions $q(\psi|\tilde{\psi})/q(\tilde{\psi}|\psi)$ cancels out of the expression for $a(\cdot)$. Often *Metropolis chain* is used to refer to this special case only. Another practically interesting variation is the use of an independent probing distribution $q(\tilde{\psi})$, that is, the proposal is independent of the current state. Tierney (1994) refers to such algorithms as *independence chains*. Hastings (1970) proposes a larger class of similar algorithms based on a more general expression for the acceptance probability. Chib and Greenberg (1995) give a tutorial introduction to the Metropolis–Hastings algorithm.

Convergence

The use of integral estimate (2) requires the verification of two conditions related to convergence.

First, the chain has to theoretically, that is, for $M \to \infty$, converge to the desired posterior distribution. Second, even if convergence for $M \to \infty$ is established, we need a convergence diagnostic to decide when we can terminate simulations in a practical implementation.

Tierney (1994, Theorem 1) shows convergence (in total variation norm) under three conditions: irreducibility, aperiodicity, and invariance.

The Markov chains which are used in MCMC schemes generally use a continuous state space, that is, $\psi^{(t)}$ is a real valued vector. For such continuous state spaces the notion of irreducibility is formally defined as π-irreducibility, with respect to some measure π on the state space. For the purpose of the present discussion we only consider $\pi(\psi) = p(\psi|y)$, that is, π denotes the desired stationary distribution. A Markov chain is π-irreducible if for any state ψ and any set B of states with $\pi(B) > 0$ there exists an integer $n \geq 1$ such that in n iterations the chain can with positive probability make a transition from ψ to some state in B.

Invariance refers to the property that if we start with a state vector generated from the desired posterior distribution, that is, $\psi^{(t)} \sim \pi$, then a further transition in the Markov chain leaves the marginal sampling distribution of ψ unchanged, that is, $\psi^{(t+1)} \sim \pi$.

The Gibbs sampler and the Metropolis–Hastings scheme define Markov chains which by construction are invariant with respect to the desired posterior distribution. Irreducibility and aperiodicity need to be verified, but are usually not a problem. However, sometimes MCMC implementations suffer from practical violations of irreducibility. There might be some subsets of the parameter space which are such that once the Markov chain simulation enters this set it is very unlikely to leave this subset again within any reasonable number of iterations. Such situations occur, for example, in independence chains if the proposal distribution $q(\tilde{\psi})$ has thinner tails than the desired posterior $\pi(\psi)$. The acceptance probabilities, Eqn. [4], include the ratios $\pi(\psi)/q(\psi)$. Assume the chain has generated a parameter value ψ far out in the tail, with very large ratio $\pi(\psi)/q(\psi)$. The chain will then reject any proposed move until a new proposal $\tilde{\psi}$ equally far out in the tail is generated.

Practically more important than establishing theoretical convergence is to recognize practical convergence, that is, to judge when sufficiently many transitions M have been simulated to obtain ergodic averages $\hat{J}$ close to the desired posterior expectations J. The simplest procedure is to plot the trajectories $\psi^{(t)}$ against iteration number t and judge convergence if an informal visual inspection of the plot does not reveal obvious trends. **Figure 3** shows a typical trajectory. Several more formal convergence diagnostics have been proposed in the recent literature. Gelman and Rubin (1992) propose to consider several independent parallel runs of the MCMC simulation. Convergence is diagnosed if the differences of $\hat{J}$ across the parallel runs are within a reasonable range. Gelman and Rubin (1992) formalize this with an ANOVA type statistic. Geweke (1992) proposes to compare an ergodic average based on early simulations (say the first 10 percent of the iterations) with an ergodic average based on later iterations (say the last 50 percent). Under convergence the two ergodic averages should be approximately equal. Using an approximate sample standard deviation based on spectral density estimates allows a formal test.

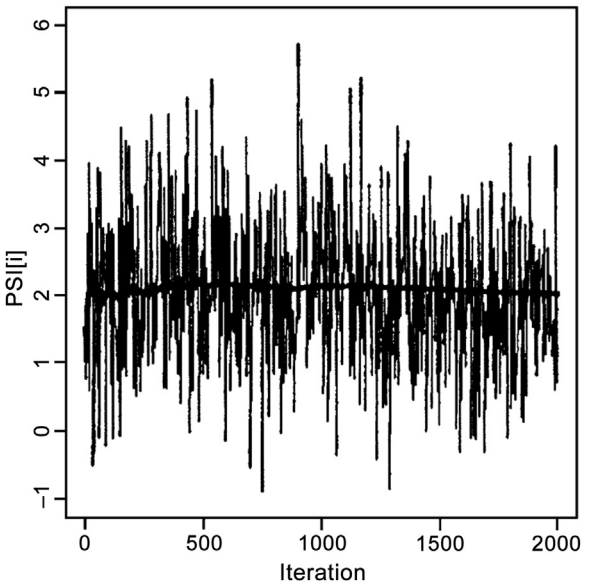

Figure 3 Convergence. The figure plots the first 2000 steps for the Gibbs sampler shown in Figure 1. The thin black curve plots $\psi_1^{(t)}$ against iteration t. The thick grey curve plots $\widehat{J}_t = 1/t\sum_{j=1}^{t}\psi_1^{(t)}$. After about 500 iterations the estimated posterior mean has practically converged. See the text for a discussion of more formal convergence diagnostics.

These and other convergence diagnostics are discussed in Best et al. (1995) and implemented in the public domain software BOA described there.

Limitations and Further Reading

The Gibbs sampler and the Metropolis–Hastings chain implicitely require a fixed dimension parameter space, that is, the dimension of ψ must not change across different values. This excludes, for example, a regression model with an unknown number of co-variates. In other words, the Gibbs sampler or the Metropolis–Hastings algorithm can not be used for model selection.

Several recent monographs provide more complete reviews of MCMC methods. Tanner (1996) provides an introduction including related schemes such as importance sampling. Assuming basic familiarity with the algorithms, Gilks et al. (1996) discuss Markov chain Monte Carlo simulation in the context of important statistical models. Gamerman (1997) and Robert and Casella (1999) review alternative algorithms and related theory.

See also: Markov Models and Social Analysis; Monte Carlo Methods and Bayesian Computation: Overview.

Bibliography

Best, N.G., Cowles, M.K., Vines, S.K., 1995. Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output. Version 0.3. MRC Biostatistics Unit, Cambridge, UK.

Carlin, B.P., Chib, S., 1995. Bayesian model choice via Markov chain Monte Carlo. Journal of the Royal Statistical Society B 57, 473–484.

Cauchemez, S., Carrat, F., Viboud, C., Valleron, A.J., Boelle, P.Y., 2004. A Bayesian Mcmc Approach to Study Transmission of Influenza: Application to Household Longitudinal Data. Statistics in medicine 23 (22), 3469–3487.

Chernozhukov, Victor, Hong, Han, 2003. An Mcmc Approach to Classical Estimation. Journal of Econometrics 115 (2), 293–346.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis–Hastings algorithm. The American Statistician 49, 327–335.

Eraker, Bjørn, 2001. MCMC Analysis of Diffusion Models with Application to Finance. Journal of Business & Economic Statistics 19 (2), 177–191.

Gamerman, D., 1997. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Chapman and Hall, London.

Gelfand, A.E., Smith, A.F.M., 1990. Sampling based approaches to calculating marginal densities. Journal of the American Statistical Association 85, 398–409.

Gelfand, A.E., Hills, S.E., Racine–Poon, A., Smith, A.F.M., 1990. Illustration of Bayesian inference in normal data models using Gibbs sampling. Journal of the American Statistical Association 85, 972–985.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7, 475–573.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (6), 721–740.

Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 4. Oxford University Press, Oxford, UK, pp. 169–194.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Markov Chain Monte Carlo in Practice. Chapman and Hall, London.

Green, P., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. Journal of Chemical Physics 21, 1087–1091.

Nylander, Johan A.A., Wilgenbusch, James C., Warren, Dan L., Swofford, David L., 2008. AWTY (Are We There Yet?): A System for Graphical Exploration of MCMC Convergence in Bayesian Phylogenetics. Bioinformatics 24 (4), 581–583.

Plummer, Martyn, Best, Nicky, Cowles, Kate, Vines, Karen, 2006. Coda: Convergence Diagnosis and Output Analysis for MCMC. R news 6 (1), 7–11.

Robert, C.P., Casella, G., 1999. Monte Carlo Statistical Methods. Springer-Verlag, New York.

Tanner, M.A., Wing, H.W., 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–550.

Tanner, M.A., 1996. Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, third ed. Springer-Verlag, New York.

Tierney, L., 1994. Markov chains for exploring posterior distributions. Annals of Statistics 22, 1701–1728.

Markov Decision Processes

Michael L Littman, Brown University, Providence, RI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Markov decision processes (MDPs), an established framework for specifying task environments for decision-making agents, are valuable in the behavioral sciences as a formal description of sequential problems faced by animals and people. This article provides a mathematical definition of MDPs, an example application, an introduction to algorithms for finding optimal policies in MDPs, and a list of useful extensions to the basic model.

Markov decision processes (MDPs) model decision making in discrete, stochastic, sequential environments (*see* Decision and Choice: Sequential Decision Making). The essence of the model is that a decision maker, or agent (*see* Autonomous Agents), inhabits an environment, which changes state randomly in response to action choices made by the decision maker. The state of the environment affects the immediate reward obtained by the agent, as well as the probabilities of future state transitions. The agent's objective is to select actions to maximize a long-term measure of total reward. Efficient algorithms for MDPs based on dynamic programing and linear programing, and more recently on compact representations, enable large planning problems from artificial intelligence, operations research, economics, robotics, and the behavioral sciences to be modeled and solved.

This article describes MDPs, an example application, algorithms for finding optimal policies in MDPs, and useful extensions to the basic model.

MDPs are attractive models of decision making and planning. Because of their more general reward and state-transition structure, MDPs extend standard planning models from artificial intelligence. Interesting environments can be expressed as MDPs. Relatively efficient algorithms exist for computing optimal decision strategies. Using current computer technology, these algorithms scale to MDPs with on the order of multiple millions of states.

Animal-Behavior Example

For a concrete example of an MDP model, consider this fictitious animal-behavior example. A flock of sheep grazes in a meadow. At any given moment, the meadow can have abundant food (s_1), adequate food (s_2), scarce food (s_3), or no food (s_4). Each day, the flock can decide to graze (a_1), search locally for another meadow (a_2), or search more globally for another meadow (a_3). The matrices in Tables 1 and 2 define the MDP. Search action a_2 has the advantage of being relatively cheap, but the disadvantage of not changing the state very much.

Roughly speaking, the sheep should graze when there is adequate food, and otherwise search for a better meadow. But when should they search and what search strategy should they use? The next sections give a formal definition of MDPs and describe how to solve them, providing a precise answer to this question.

Definitions

An MDP is a tuple $M = (S, A, T, R, \beta)$, where

- S is a finite set of states of the environment;
- A is a finite set of actions;
- T: $S \times A \rightarrow \Pi(S)$ is the state-transition function, giving for each state and agent action a probability distribution over states ($T(s, a, s')$ is the probability of ending in state s', given that the agent starts in state s and takes action a);
- $R : S \times A \rightarrow \mathscr{R}$ is the reward function, giving the expected immediate reward gained by the agent for taking each action in each state
  ($R(s, a)$ is the expected reward for taking action a in state s); and
- $0 < \beta < 1$ is a discount factor).

A policy is a description of the behavior of an agent. Because of their simple structure, this article considers only stationary policies, such as π:$S \rightarrow A$. These policies specify, for each state, an action to be taken should the agent find itself in that state.

The total expected discounted reward for a policy π from some initial state is a measure of its long-run success. It is defined as the expected value of the discounted sum of immediate rewards. To be more concrete, let $V^{\pi}(s)$ be the total expected discounted reward for starting in state s and executing policy π

$$V^{\pi}(s) = R(s, \pi(s)) + \beta \sum_{s' \in S} T(s, \pi(s), s') V^{\pi}(s'). \qquad [1]$$

The value function for policy π is the unique solution of this set of simultaneous linear equations, one for each state s. The

Table 1 Reward function and state transition function for the graze action in the fictitious sheep grazing example

$R(s, a)$	a_1	a_2	a_3
s_1	5	−1	−5
s_2	5	−1	−5
s_3	2	−1	−5
s_4	0	−1	−5

$T(s, a_1, s')$	s_1	s_2	s_3	s_4
s_1	0.6	0.3	0.1	0.0
s_2	0.0	0.4	0.4	0.2
s_3	0.0	0.0	0.4	0.6
s_4	0.0	0.0	0.0	1.0

Table 2 State transition functions for the search actions in the fictitious sheep grazing example

	s_1	s_2	s_3	s_4
$T(s, a_2, s')$				
s_1	0.5	0.4	0.1	0.0
s_2	0.3	0.5	0.2	0.0
s_3	0.0	0.4	0.4	0.2
s_4	0.0	0.0	0.4	0.6
$T(s, a_3, s')$				
s_1	0.1	0.4	0.3	0.2
s_2	0.1	0.4	0.3	0.2
s_3	0.1	0.4	0.3	0.2
s_4	0.1	0.4	0.3	0.2

system of linear equations can be solved using any of a number of standard techniques from linear algebra.

Equation [1] derives a value function from a policy. Reversing this, the greedy policy with respect to value function V, π_v, is defined as:

$$\pi_v(s) = \underset{\alpha}{\operatorname{argmax}}\left[R(s,a) + \beta \sum_{s' \in S} T(s,a,s')V(s')\right]. \qquad [2]$$

This policy is obtained by taking the action in each state with the best one-step look-ahead value according to V.

Given an initial state s, the goal is to execute the policy π that maximizes $V^{\pi}(s)$. Howard (1960) showed that a policy π^* exists that is simultaneously optimal for every starting state. The value function for this policy, written V^*, is defined by the set of equations:

$$V^*(s) = \max_{\alpha}\left[R(s,a) + \beta \sum_{s' \in S} T(s,a,s')V^*(s')\right]. \qquad [3]$$

Any greedy policy with respect to this value function is optimal. Thus, the goal of many MDP solution algorithms is finding V^* or a good approximation.

Algorithms for Solving Markov Decision Processes

Solving an MDP means finding the policy π^* that maximizes the total expected discounted reward. Although many approaches have been studied, the simplest and most general is value iteration.

Value iteration treats the equality in eqn [3] as an assignment. It starts off with an arbitrary approximation to V^*, say a value function V that assigns 0 to all states. Over successive iterations, this approximation is improved by using the previous value of V combined with one-step look-ahead to get a more refined estimate. The algorithm terminates when the maximum difference between two successive value functions is less than some predetermined values ($\varepsilon > 0$), at which time the greedy policy with respect to V is returned. The algorithm is guaranteed to terminate, and, if ε is set small enough, the returned policy will be optimal.

Although easy to implement, value iteration can be very slow to converge in some MDPs. For additional computational cost per iteration, policy iteration, due to Howard (1960), results in considerably more rapid covergence in practice. Like value iteration, policy iteration computes a sequence of value functions, V. In policy iteration, V is the value function for the greedy policy with respect to the value function V' from the previous iteration: $V = V^{\pi}_{v'}$. This assignment can be done by applying eqn [2] and then solving eqn [1].

Table 3 Sequence of value functions generated by value iteration in the fictitious sheep grazing example

Iteration	s_1	s_2	s_3	s_4
0	0.000	0.000	0.000	0.000
1	5.000	5.000	2.000	0.000
2	9.230	7.520	2.720	0.000
3	12.259	8.686	2.979	0.000
4	14.234	9.200	3.200	0.073
⋮				
86	19.480	12.502	6.502	3.670

Policy iteration finds the optimal value function and policy after a finite number of iterations. There are a finite number of policies, and each iteration with a nonoptimal policy finds a policy that is a strict improvement over the previous one.

Returning to the animal-behavior example given earlier, value iteration produces the sequence of value function given in Table 3. The table shows the result of the first four iterations of the algorithm and the result of the 86th iteration. After 86 iterations, the value function has converged to within machine precision. From the final value function, the optimal policy is to graze (a_1) in states s_1 and s_2, local search (a_2) in state s_3, and global search (a_3) in state s_4. Value iteration finds this policy after only four iterations, but convergence to the optimal value function takes considerably longer: for accuracy out to six digits past the decimal point, 170 iterations are required. In contrast, policy iteration steps through only one suboptimal value function (for the policy that always chooses a_1) before obtaining the optimal policy and value function and terminating.

As the discount factor β gets close to one, value iteration can require exponentially many iterations to converge. It is possible, although not proven, that policy iteration can also take an exponential number of iterations. The only approach to solving MDPs that has been shown to terminate after a polynomial amount of computation is linear programing (*see* Algorithmic Complexity). Littman et al. (1995) survey computational complexity issues in MDPs.

Related Models and Extensions

Although the type of MDP described in this article is very general and has been extensively studied, other closely related models are also quite useful.

- *Markov games*. In this model, state transitions and rewards are controlled by multiple agents instead of a single agent. Agents select actions to maximize their personal total expected discounted reward. The zero-sum version of this model (*see* Game Theory), first introduced by Shapley (1953), can be solved via variations of value iteration and policy iteration.

- *Partially observable MDPs*. This model is a hybrid of MDPs with hidden Markov models in which the environment's state is only visible via stochastically generated, and possibly ambiguous, 'observations.' Smallwood and Sondik (1973) provide one of the first algorithmic treatments of this model. Both value iteration and policy-iteration algorithms have been studied. A tutorial for behavioral scientists can be found in Littman (2009).
- *Continuous states, actions, and time*. Robotics problems, for example, feature continuous state spaces (e.g., arm angle) and action spaces (e.g., joint acceleration) and must run in real time (*see* Dynamic Decision Making). Puterman (1994) surveys some relevant results.
- *Other objective functions*. The problems of maximizing average reward per step, total expected reward (undiscounted), and worst-case reward all share important features with the total expected discounted reward model described here, although each introduces its own complications.
- *Propositional representations*. Many environments can be best described using states or actions in attribute-value form (e.g., HEALTH = good, AGE = young, SEASON = winter). Boutilier et al. (1999) survey propositional representations of MDPs and attempt to take advantage of these representations to speed the solution process.
- *Value-function approximation*. Another direction is the use of compact parameterized function approximators such as neural networks to store value functions. Bertsekas and Tsitsiklis (1996) dubbed this approach 'neuro-dynamic programing,' and it has been successfully applied to some very difficult sequential decision problems such as helicopter control and move selection in board games.
- *Reinforcement learning*. The MDP model has been embraced by reinforcement-learning researchers as a model of environments faced by agents learning from delayed feedback (Sutton and Barto, 1998). Behavioral and neuroscience applications of this idea abound.

The MDP literature is quite substantial. The early work of Bellman (1957) and Howard (1960) put the field on a firm footing. Later work synthesized the existing material and added new results. Textbooks by Puterman (1994) and Bertsekas (1995) give in-depth summaries. Contributions have been made by researchers in operations research, electrical engineering, dynamic programing, complexity theory, economics, algorithm analysis, artificial intelligence, and the behavioral sciences (see Chap. 1 of Puterman (1994) for a summary). Existing theory and algorithms have been applied successfully to real-world problems in a wide array of environments, and renewed attention in recent years has greatly expanded the range of problems to which this formalism can be applied.

See also: Decision Making, Psychology of; Dynamic Decision Making; Optimal Control Theory.

Bibliography

Bellman, R., 1957. Dynamic Programming. Princeton University Press, Princeton, NJ.
Bertsekas, D.P., 1995. Dynamic Programming and Optimal Control, vols. 1 and 2. Athena Scientific, Belmont, MA.
Bertsekas, D.P., Tsitsiklis, J.N., 1996. Neuro-Dynamic Programming. Athena Scientific, Belmont, MA.
Boutilier, C., Dean, T., Hanks, S., 1999. Decision-theoretic planning: structural assumptions and computational leverage. Journal of Artificial Intelligence Research 11, 1–94.
Howard, R.A., 1960. Dynamic Programming and Markov Processes. MIT Press, Cambridge, MA.
Littman, Michael L., 2009. A tutorial on partially observable Markov decision processes. Journal of Mathematical Psychology 53 (3), 119–125.
Littman, M.L., Dean, T.L., Kaelbling, L.P., 1995. On the Complexity of Solving Markov Decision Problems. In: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI-95). Morgan Kaufmann Inc., San Francisco, pp. 394–402.
Puterman, M.L., 1994. Markov Decision Processes—Discrete Stochastic Dynamic Programming. Wiley, New York.
Shapley, L., 1953. Stochastic games. Proceedings of the National Academy of Sciences of the United States of America 39, 1095–1100.
Smallwood, R.D., Sondik, E.J., 1973. The optimal control of partially observable Markov processes over a finite horizon. Operations Research 21, 1071–1088.
Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.

Markov Models and Social Analysis

Lynne Billard, University of Georgia, Athens, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Many social science processes such as consumer behavior patterns, mobility patterns, friendship formations, networks, voting patterns, environmental management such as patient movements between hospitals are best analyzed by Markov models. Four specific examples are examined. Discrete time processes are called Markov chains, and continuous time processes are Markov processes. The basic models are described and some properties are provided.

Introduction

First-order Markov models are appropriate for analyzing social systems when the individual under study (person, system, unit, etc.) can be characterized as being in one of S (say) clearly identifiable states, and when the transition probabilities of moving from one state to another are dependent only on the current state (and not on any of the previous states; rth order models are those for which this dependence rests on the r most current states). Discrete time models give rise to Markov chains. Some basic properties are presented for time-homogeneous (stationary) and time-nonhomogeneous chains along with some inference procedures (maximum likelihood and empirical Bayes estimation). Also considered are continuous time Markov chains and Markov processes. Some results for the linear birth and death models (probabilities, holding times, etc.) are provided. Other related models such as the mover–stayer, mixed Markov, semi-Markov, and interactive Markov models are described briefly.

There are a plethora of social science processes that are best analyzed through the use of some form of Markov model. In a study of consumer behavior, individuals may switch from one brand of a product to another brand. Note that a 'change' here includes staying with the current brand, as well as the possibility of returning to a previously used brand. This switching can be 'random' over time, or the individual/customer may in time settle on a particular brand, or may even move away entirely from all brands to a different product. Likewise, voters may change their preferences for candidates from one party to another. Equivalently, workers may change jobs from one sector of the workforce to another; or the focus may be on intergenerational occupational patterns. Households, or individuals, may move across differing income levels or across various social classes. Credit card holders may move in and out of debt levels.

In a different direction, understanding (and predicting) bed occupancy rates in a hospital can be modeled by Markov processes. Instead of the movement of patients through the stages of illness recovery, the question of interest may be movement up the company's hierarchical promotion ladder(s) or on progress through states (or lack thereof, as in recidivism) in rehabilitation programs (as in health, criminology, etc.). Friendship formation and more generally network analyses and their various manifestations also fall under the Markov model rubric. Bartholomew (1983), among others, provides a compact summary and references for a wide array of applications.

A common thread in all such processes is that at a particular time, t say, the individual under study (i.e., person, system, etc.) is characterized as being in one of a number of clearly identifiable states, s say, $s = 1,\ldots, S$, where S is the total number of states. Another thread is that the state currently occupied by an individual depends on the state occupied at one or more previous times; that is, there is an underlying dependence between the observations. The data record these states for a set of individuals at times $t_0, t_1,\ldots, t_T$ where $T \geq 1$ and t_0 is the initial time. Time itself can be a discrete or continuous variable, leading to Markov chains or Markov processes, respectively (named after the Russian mathematician Andrei Andreevich Markov, 1856–1922, who developed fundamental probabilist results for dependent processes). When time is discrete, the data may typically be thought of as occurring at times $t = 0, 1,\ldots, T$ (with an implied assumption that changes from one time t to the next $(t + 1)$ are not dependent on the length of time between such measurements; this assumption is not always valid, see Section Finite State Markov Chains (*see* Longitudinal Data: Event–History Analysis in Discrete Time).

A first-order Markov chain, or simply a Markov chain, is characterized by the probabilities of moving from state i at time $t - 1$ to state j at time t, $p_{ij}(t)$, with

$$\begin{aligned} p_{ij}(t) &= P\left(\text{State } s = j \text{ at } t \middle| \text{State } s = i \text{ at } t-1\right) \\ &= P(s = j \text{ at } t | s = s_0 \text{ at } t = 0, \ldots, s = s_{t-2} \\ &\qquad \text{at } t-2,\ s = i \text{ at } t-1) \end{aligned} \quad [1]$$

that is, the transition probability depends only on its most recent state and not on the states occupied at other previous times. More generally, an rth order Markov chain is one for which the transition probabilities depend on the r most recent states. Except where specified, this article assumes a first-order chain as governed by [1]. These transition probabilities may be independent of t with $p_{ij}(t) \equiv p_{ij}$, which is the stationary (time-homogeneous) case, or may be a function of t giving rise to nonstationary (time-inhomogeneous) Markov chains. (Note that, for some users, stationarity defines the steady state or equilibrium probabilities obtained after a process has settled down; see Section Discrete Time Models: Markov Chains.) Homogeneity (or conversely, heterogeneity) refers typically to the fact that the transition probabilities are the same (or not the same) for all individuals in the study. Continuous time models

take two broad forms. The first is a finite state Markov chain based on eqn [1]. The other form falls under the heading of Markov processes characterized by transition rates $\lambda_{ij}(t)$ of moving from state i at time t to state j at the infinitesimal time later $t + \Delta t$ also independent of earlier transitions, i.e.,

$$P(s = j \text{ at } t + \Delta t | s = i \text{ at } t) = \lambda_{ij}(t) + o(\Delta t) \qquad [2]$$

where $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$. As for the Markov chain, these rates can be homogeneous (i.e., time-independent with $\lambda_{ij}(t) = \lambda_{ij}$) or nonhomogeneous. In these cases, it is more usual to focus attention on estimating the numbers in each state at time t rather than the transition probabilities of a single individual moving from one state to another; see Section Markov Processes. However, recently, a new approach was introduced by Billard and Dayananda (2014) in which, instead of focusing on the numbers in each state, the waiting time spent in each successive state was derived from which probabilities of movement to later states could be obtained.

Section Some Examples provides four examples of Markov models from the social science literature. The attention then shifts to the basic models and some of their properties, along with aspects of parameter estimation described in Sections Discrete Time Models: Markov Chains and Continuous Time Models: Markov Processes for discrete and continuous time models, respectively. Some extensions, such as mover–stayer models, mixed Markov models, semi-Markov models, and interactive Markov chains, are also covered briefly in Section Some Other Models.

Some Examples

Brand Loyalty

Manufacturers, and equally retailers, will always be interested in whether or not the users of their products remain loyal to a particular product and/or switch to another product, and if so, what it is that entices the loyalty/switch that occurs. Is this the effect of advertising, product quality, peer pressure, or what? Draper and Nolin (1964), Lawrence (1966) and Chatfield (1973), among others, first studied this behavior as a Markov chain. Suppose there are three products ($S = 3$): Brand A ($s = 1$), Brand B ($s = 2$), and Brand C ($s = 3$). Suppose the transition probabilities p_{ij}, $i, j = 1, 2, 3$, are as shown in **Table 1**, where it is assumed implicitly that these probabilities are constant over time. These p_{ij}'s represent the probability that an individual who bought one brand ($s = i$) at time ($t - 1$) will buy another brand ($s = j$) the next time (t). For example, the probability that a person who bought Brand A and then bought Brand B the next time is $p_{12} = 0.18$. The diagonal elements p_{ss}, $s = 1, 2, 3$, represent product loyalty; e.g., the probability that an individual bought Brand C both last time and this time is $p_{33} = 0.64$. Typically, in practice, these transition probabilities are unknown, and so must be estimated from data. Estimation procedures are discussed in Section Discrete Time Models: Markov Chains. Once these probabilities are known (or estimated), there are numerous questions of interest (e.g., what is the long-run brand loyalty over many time periods?) that can be answered; some of these will be presented in Section Discrete Time Models: Markov Chains.

Table 1 Brand-switching transition probabilities

		Time t		
	p_{ij}	*A*	*B*	*C*
Time ($t - 1$)	A	0.72	0.18	0.10
	B	0.09	0.85	0.06
	C	0.21	0.15	0.64

Occupational Mobility

Occupational mobility, or more broadly social mobility, can be modeled by a Markov chain. The data of **Table 2** display the probabilities associated with the son's (first) occupation and the occupation of the father. There are $S = 5$ possible occupations: self-employed and salaried professionals ($s = 1$); proprietors, clerical workers, and retail sellers ($s = 2$); craftsmen in manufacturing, construction, and other industries ($s = 3$); service workers, operatives, and laborers ($s = 4$); and farm managers and workers ($s = 5$). Thus, p_{ii}, $i = 1, \ldots, 5$, represents the probability of a son following his father's footsteps by going into the same occupation; e.g., the probability a son follows his father into farming is $p_{55} = 0.39$. The p_{ij}, $i \neq j$ represent the probabilities that the son goes into a different occupation; e.g., the probability a son goes into a self-employed position while his father worked as a construction worker is $p_{21} = 0.32$. The Biblarz and Raftery (1993) study was actually concerned with whether occupational mobility varied across race and/or across types of family disruption. Thus, instead of **Table 2**, there is a corresponding table for each of the demographic categories considered. Although it is beyond the scope of this article, logistic regression methods were used subsequently to study what factors most influenced occupational mobility, with the transition probabilities p_{ij} being the starting point.

In this example, changes in time t correspond to generational changes. Had the interest focused on occupational mobility of individuals over time (let us say, over decades), then data of the type of **Table 2** pertain but where now, e.g., $p_{31} = 0.19$ is the probability that an individual working in construction has moved into a self-employed or salaried professional position one decade later.

Instead of occupational mobility, states may represent social classes so that social mobility questions can be studied.

Table 2 Occupational mobility transition probabilities

Father's occupation	*Son's occupation*				
	1	*2*	*3*	*4*	*5*
1	0.48	0.18	0.11	0.22	0.01
2	0.32	0.24	0.11	0.31	0.02
3	0.19	0.16	0.21	0.42	0.02
4	0.13	0.15	0.13	0.55	0.04
5	0.09	0.08	0.09	0.35	0.39

Adapted from Biblarz, T.J., Raftery, A.E., 1993. The effects of family disruption on social mobility. American Sociological Review 58, 97–109.

Or, in a study of economic mobility, states would represent levels of economic status, be these states representing financial assets (say), credit-worthiness (say), income levels, or whatever economic factor is the focus. In all cases, the probabilities of movement from one state to another take the form of Table 2.

Voter Patterns

Anderson (1954), Goodman (1962), and Brown and Payne (1986), among numerous other authors, have used Markov chain models to study voting behavioral patterns. In Goodman's study, 455 potential voters were asked in six consecutive months which party's candidate would receive their vote, be that Republican (R, $s = 1$), Democrat (D, $s = 2$) or do not know (U, $s = 3$). Table 3(a) gives the numbers of those respondents who indicated in May how they would vote, and again what those responses were in June. Thus, for example, of the 445 subjects interviewed, 11 of those who in May did not know for whom they would vote had by June decided to vote for the Republican candidate. Or, equivalently, from the (estimated) transition probabilities of Table 3(b), 83% of those who in May planned to vote Democrat were still planning in June to vote for the Democrat candidate.

Malaria

Let us consider Cohen and Singer's (1979) study of malaria. Each individual was identified at time t as being in one of $S = 4$ states, corresponding to uninfected ($s = 1$), singly infected with *Plasmodium malariae* ($s = 2$), singly infected with *Plasmodium falciparum* ($s = 3$), or jointly infected with both *Plasmodium* species ($s = 4$). Eight surveys were conducted at regular (10-week) intervals, i.e., $T = 7$. The study also separately considered the transitions into and out of the four disease states, for each of seven age groups. Thus, for each age group, there is a 4×4 transition probability matrix (tpm) for each transition from time $t = t_u$ to $t = t_v > t_u$, of the type shown in Table 4 (extracted from Cohen and Singer, 1979), where t_u, $t_v = 0, \ldots, 7$. It is assumed that infection can occur at any time t and not just at those times surveyed, thus making this a continuous time Markov chain (in contrast to the previous examples, all of which were discrete time chains). In a different direction, Cohen and Singer (1979) also studied the malaria epidemic as a Markov process with transitions governed by eqn [2].

Table 3 Voting patterns

		(a)					(b)		
		June					June		
	#'s	R	D	U	Totals	p_{ij}	R	D	U
May	R	125	5	16	146	R	0.86	0.03	0.11
	D	7	106	15	128	D	0.05	0.83	0.12
	U	11	18	142	171	U	0.06	0.11	0.83
Total		143	129	173	445				

Table 4 Malaria disease states transition probabilities

		Time t			
	$\hat{p}_{ij}$	1	2	3	4
Time ($t-1$)	1	0.64	0.03	0.32	0.01
	2	0.66	0.06	0.28	0.00
	3	0.39	0.01	0.53	0.07
	4	0.38	0.05	0.47	0.10

Discrete Time Models: Markov Chains

Stationary Markov Chains

Let $\mathbf{P}$ be an $S \times S$ matrix of transition probabilities $p_{ij}(t)$ (defined in eqn [1]), and let $p_{ij}(t) = p_{ij}$, i.e., the Markov chain is stationary. Note that $\sum_{j=1}^{S} p_{ij} = 1$. Let $p_s^{(0)}$ be the probability that an individual is in the initial state s at $t = 0$, $s = 1, \ldots, S$ and write $\mathbf{p}^{(0)}$ as the $S \times 1$ vector whose elements are $p_s^{(0)}$. The matrix $\mathbf{P}$ is called a tpm and $\mathbf{p}^{(0)}$ is the initial probability vector. The Markov chain is defined completely by $\mathbf{P}$ and $\mathbf{p}^{(0)}$ (*see* Probability Theory: Formal). Some basic results and definitions include the following.

1. The n-step-ahead probability matrix $\mathbf{P}^{(n)} = (p_{ij}^{(n)})$, $i, j = 1, \ldots, S$, and the n-step-ahead probability vector $\mathbf{p}^{(n)} = (p_i^{(n)})$, $i = 1, \ldots, S$, satisfy

$$\mathbf{P}^{(n)} = \mathbf{P}^n, \quad \mathbf{p}^{(n)} = \mathbf{p}^{(0)}\mathbf{P}^n \qquad [3]$$

where $\mathbf{P}^n$ implies matrix multiplication of $\mathbf{P}$ to the nth power in the usual way; and where the elements of $\mathbf{P}^{(n)}$ and $\mathbf{p}^{(n)}$, respectively, are $p_{ij}^{(n)}$, which is the probability that an individual in state i at time t is in state j at time $(t+n)$, and $p_i^{(n)}$, which is the probability the individual is in state i after n times units.

To illustrate, Table 5 provides the transition probabilities in $\mathbf{P}^{(2)}$ representing the probabilities associated with mobility between a grandfather's and a grandson's occupation, and were obtained from the $\mathbf{P}$ of Table 2. Thus, for example, the probability that both the grandfather and grandson are in farming is 0.1704. Note, however, that this says nothing about the father's occupation. Indeed, this term was calculated from

$$p_{55}^{(2)} = (0.09)(0.01) + (0.08)(0.02) + (0.09)(0.02) + (0.35)(0.04) + (0.39)(0.39) = 0.1704$$

Table 5 Two-generational transition probabilities

	Grandson's occupation				
Grandfather's occupation	1	2	3	4	5
1	0.3384	0.1810	0.1252	0.3321	0.0233
2	0.2934	0.1809	0.1268	0.3685	0.0304
3	0.2387	0.1708	0.1390	0.4176	0.0339
4	0.2102	0.1659	0.1332	0.4462	0.0445
5	0.1665	0.1335	0.1182	0.4114	0.1704

that is, the $p_{55}^{(2)}$ takes into account all possible occupations that the father might have had. The probability that all three generations were farmers is $(0.39)^2 = 0.1521$.

2. A state j is *accessible* from state i if and only if from some n, $p_{ij}^{(n)} > 0$. Note that i and j are not necessarily accessible in one step, but are accessible at some time $n > 0$. If no n exists for which $p_{ij}^{(n)} > 0$, then it is never possible to reach j from i. Two states, i and j, *communicate* with each other if each is accessible to the other.
3. If all states $\{1,\ldots, S\}$ communicate with each other, the chain is said to be *irreducible*.
4. If there exists an n such that $p_{ij}^{(n)} > 0$ for all i and j, then the chain is *ergodic*.
5. For an irreducible aperiodic ergodic chain, there exists a stationary distribution π, which satisfies $\pi = \pi\mathbf{P}$, with elements π_s representing the long-run proportion of times state s is visited and with $\sum \pi_s = 1$.

To illustrate with the brand-switching data of Table 1, it follows that $\pi = (0.299, 0.530, 0.171)$; i.e., in the long run, 29.9% of customers purchase Brand A, and likewise 53% (17.1%) selected Brand B (C).

There are innumerable other properties (*see* Probability Theory: Formal) that govern a Markov chain, providing answers to a wide range of questions of interest, such as whether or not a particular state or group of states is transient (i.e., eventual return is certain), absorbing (i.e., once reached cannot be exited), the expected number of visits to a state, and so forth. Details can be found in a variety of basic texts. All these properties evolve from a knowledge of the tpm $\mathbf{P}$ and the initial probability vector $\mathbf{p}^{(0)}$. Therefore, attention will now be directed to estimating these entities.

Statistical Inference-Classical Results

Suppose the data consist of N individuals, for each of whom it is known what state, X_{kt} say, that the kth person was in at each of the $t = 0, 1,\ldots, T$ time points recorded, $k = 1,\ldots, N$. For example, the X_{kt} may represent the brand of a product used by individual k at time t. It is assumed that $N > 1$, giving rise to the so-called panel studies. Let $n_s(0)$ be the number of individuals in state s at time $t = 0$. The $n_s(0)$ are assumed to be nonrandom; when they are random, the same results pertain except that the $n_s(0)$ are replaced by their expectations, $s = 1,\ldots, S$. Let $n_{ij}(t)$ be the number of individuals who moved from state i at time $t-1$ to state j at time t; and let $n_{ij} = \sum_{t=1}^{T} n_{ij}(t)$ be the total number of transitions from states i to j.

Then, the initial probabilities $p_s^{(0)}$, $s = 1,\ldots,S$, are estimated by

$$\widehat{p}_s^{(0)} = n_s(0) \Big/ \sum_{s=1}^{S} n_s(0). \qquad [4]$$

The maximum likelihood estimator (mle) of p_{ij}, $i, j = 1,\ldots, S$, is

$$\widehat{p}_{ij} = n_{ij} \Big/ \sum_{j=1}^{S} n_{ij} = \sum_{t=1}^{T} n_{ij}(t) \Big/ \sum_{t=1}^{T}\sum_{j=1}^{S} n_{ij}(t). \qquad [5]$$

When the Markov chain is not homogeneous, the mle of the transition probabilities $p_{ij}(t)$ is

$$\widehat{p}_{ij}(t) = n_{ij}(t)/n_{i.}(t), \quad t = 1,\ldots,T, \qquad [6]$$

where $n_{i.}(t) = \sum_{j=1}^{S} n_{ij}(t)$.

The transition probabilities for voter behavior from May to June given in Table 3(b) were estimated using eqn [5]. Estimated transition probabilities for each of the $T = 5$ transition times can be found by applying eqn [6] to Goodman's (1962) data (not shown).

For properties of these estimators and their original derivation, see Anderson and Goodman (1957). They also derived test statistics for key hypothesis testing and model validation situations. Further, these methods can be treated as special cases of log-linear models (see, e.g., Bishop et al., 1975) (*see* Multivariate Analysis: Discrete Variables (Overview)).

Empirical Bayes Estimation of P

When information from other similar data sets exists, that information can be incorporated into the estimation of the particular tpm $\mathbf{P}$ under discussion, by calculating an empirical Bayes estimator of $\mathbf{P}$. Not only does this utilize all the available information; the technique is especially useful when sample sizes are small, data are sparse, and/or mle's tend to take the extreme value of 0 or 1 for some p_{ij}. Thus, in a brand-switching scenario, data are available for each $l = 1,\ldots, L$ other locations (cities, stores, population sectors, etc.), and the location of particular interest is location $l = L + 1$. In effect, the p_{ij}'s of the tpm $\mathbf{P}_{L+1}$ for location $(L + 1)$ are assumed to have an underlying distribution whose unknown parameters are estimated by the data from the $\{l = 1,\ldots, L\}$ other chains; and thence through the use of Bayes' theorem and the data generated at the location $(L + 1)$, the tpm $\mathbf{P}_{L+1}$ can be estimated.

In this case, there are data values generated for $n_{ij}(t)$, n_{ij}, $n_s(0)$, $\mathbf{P}$, etc., for each $l = 1,\ldots, L, L + 1$. Then, the empirical Bayes estimate of the element p_{ij} of $\mathbf{P}_{L+1}$ is

$$\widehat{p}_{\mathrm{EB};ij} = \left(n_{ij}^* + \widehat{\rho}_{ij}\right) \Big/ \left(n_{i.}^* + \widehat{\rho}_{i+}\right), \quad i,j = 1,\ldots, S, \qquad [7]$$

where $n_{ij}^* \equiv n_{ij}$ and $n_{i.}^* \equiv n_i$ for location $L + 1$ and where $\widehat{\rho}_{i+} = \sum_{j=1}^{S} \widehat{\rho}_{ij}$. The $\widehat{\rho}_{ij}$ can be obtained by method of moments or maximum likelihood techniques. The method of moments estimate gives

$$\widehat{\rho}_{ij} = \overline{M}_{ij}\overline{X}_{ij} \big/ \left[\overline{M}_{ij}\left(1 - \overline{M}_{ij}\right) - \overline{X}_{ij}\right], \quad i, j = 1,\ldots,S, \qquad [8]$$

where

$$\overline{M}_{ij} = \frac{1}{L}\sum_{l=1}^{L} \overline{M}_{ijl}, \quad \overline{X}_{ij} = \frac{1}{L}\sum_{l=1}^{L} X_{ijl}$$

with

$$\overline{M}_{ijl} = n_{ijl}/n_{i.l}, \quad X_{ijl} = M_{ijl}\left(1 - M_{ijl}\right), \quad l = 1,\ldots,L.$$

The maximum likelihood estimate of $\widehat{\rho}_{ij}$ (and hence $\widehat{\rho}_{i+}$) are those values, which maximize the likelihood function

$$L\left(\rho_{ij}\middle|n_{ij}\right) \propto \frac{\Gamma(\rho_{i+})}{\Gamma\left(\rho_{ij} + n_{i..}\right)} \prod_{j=1}^{S} \frac{\Gamma(\rho_{ij} + n_{ij.})}{\Gamma\left(\rho_{ij}\right)} \qquad [9]$$

where $n_{ij.} = \sum_{l=1}^{L} n_{ijl}$, $n_{i..} = \sum_{j=1}^{S}\sum_{l=1}^{L} n_{ijl}$ and $\Gamma(\bullet)$ is the usual gamma function. Details of the derivation of the results

[7–9] can be found in Meshkani and Billard (1992), along with properties of these estimators and an example. When there is only a single Markov chain at each site (i.e., $N = 1$), the $\widehat{\rho}_{ij}$ and $\widehat{\rho}_{i+}$ are estimated differently; see Billard and Meshkani (1995).

Continuous Time Models: Markov Processes

Finite State Markov Chains

The Markov chains discussed in Section Discrete Time Models: Markov Chains were discussed in the context of discrete time. When there is a natural unit of time for which the data of a Markov chain process are collected, such as week, year, generational, etc., use of the discrete time model is satisfactory. If there is no natural time unit, then use of the discrete time model could be problematical. For example, for a given tpm $\mathbf{P}$, does the corresponding tpm for 1 year later equal $\mathbf{P}^{(2)}$, or $\mathbf{P}^{(12)}$, for a time unit of 6 months, or 1 month, respectively? Clearly, the answers are different though the time period – 1 year later – is the same. Such difficulties can be circumvented by using a continuous time Markov chain; see Singer and Spilerman (1975), and Shorrocks (1978). When time is continuous, the concepts elucidated for the discrete time chains carry over though the notation and some of the mathematics differ.

Let X_t represent the state of the chain at time t, and let

$$p_{ij}(t) = P(X_{t+t_0} = j | X_{t_0} = i) \qquad [10]$$

be the elements of the $S \times S$ matrix $\mathbf{P}(t)$ with $\sum_{j=1}^{S} p_{ij}(t) = 1$ for all $i = 1,\ldots, S$. Then, for continuous times t_1 and $t_2 \geq 0$, the $\mathbf{P}(t)$ satisfy

$$\mathbf{P}(t_1 + t_2) = \mathbf{P}(t_1)\mathbf{P}(t_2) \qquad [11]$$

or, equivalently,

$$p_{ij}(t_1 + t_2) = \sum_{k=1}^{S} p_{ik}(t_1)p_{kj}(t_2). \qquad [12]$$

The eqns [11] and [12] are called the Chapman–Kolmogorov equations. These equations are used to show that the matrix $\mathbf{P}(t)$ satisfies the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{P}(t) = \mathbf{Q}\mathbf{P}(t) \qquad [13]$$

where $\mathbf{Q}$ is an $S \times S$ matrix whose elements q_{ij} satisfy

$$\begin{aligned} q_{ij} &= \lim_{\Delta t \to 0} p_{ij}(\Delta t)/\Delta t, \quad i \neq j, \\ q_{ii} &= \lim_{\Delta t \to \infty} [1 - p_{ii}(\Delta t)]/\Delta t, \quad i = j, \end{aligned} \qquad [14]$$

and are called *intensities*. It follows that

$$q_{ii} = -\sum_{i=1, i\neq j} q_{ij} \equiv -q_i, \text{ say.} \qquad [15]$$

Then, $(-q_{ij}/q_i)$ represents the conditional probability that an individual in state i moves to state j given a transition occurred, and $(-1/q_i)$ is the average time an individual stays in state i.

Eqn [13] can be solved to give

$$\mathbf{P}(t) = \mathbf{e}^{\mathbf{Q}t} = \sum_{n=0}^{\infty} \mathbf{Q}^n t^n / n! = \mathbf{U}\mathbf{A}(t)\mathbf{U}^{-1} \qquad [16]$$

where $\mathbf{A}(t) = (\mathrm{diag}(e^{\lambda_i t}))$, $i = 1,\ldots, S$, is the $S \times S$ diagonal matrix with $\{\lambda_i\}$ being the eigenvalues of $\mathbf{Q}$, and where $\mathbf{U}$ is the $S \times S$ matrix of the eigenvectors of $\mathbf{Q}$.

Conditions (on the $p_{ij}(t)$ values), most notably embeddability and identifiability problems, which allow the solution in eqn [16] to exist have received considerable study. Singer and Spilerman's (1976) lengthy article provides a nice summary as well as extensive insights. In a series of papers (see, e.g., Stasny, 1986), movement in and out of the labor force entirely embodies these principles (where the $S = 3$ states represent those who are: employed, unemployed, and not in the labor force). Geweke et al. (1986) expand upon these difficulties within the context of mobility models. Some conditions include the following, where $\widehat{\mathbf{P}} = (\widehat{p}_{ij})$ is the observed estimate of $\mathbf{P}$ obtained by the methods of Section Discrete Time Models: Markov Chains.

1. If $S = 2$, $\widehat{p}_{ij} \geq 0$, $\widehat{p}_{i1} + \widehat{p}_{i2} = 1$, $i = 1, 2$, if and only if $\widehat{p}_{11} + \widehat{p}_{22} > 1$.
2. If $\widehat{p}_{ij}(t) = 0(\neq 0)$, then $\widehat{p}_{ij}^{(n)} = 0(\neq 0)$, for all (respectively, any) integers n.
3. The determinant $|\mathbf{P}| > 0$.
4. Other than the first eigenvalue ($\lambda_1 = 1$ always) no other λ_i equals one and any negative λ_i has even multiplicity.

Markov Processes

Markov processes are continuous time Markov models based on eqn [2], where the focus is on the number of individuals in a given state at time t (rather than the transitions probabilities). The two simplest such models are the so-called Poisson process and the birth and death models. Let $X_s(t)$ be the number of individuals in state s at time t, $s = 1,\ldots, S$.

The possible transitions can be visualized as movements in and out of compartments

$$\begin{aligned} &|\overline{s = 0}| \to \ldots \rightleftarrows |\overline{s = i}| \rightleftarrows \ldots \\ &\rightleftarrows |\overline{s = j}| \rightleftarrows \ldots \to |\overline{s = S}| \end{aligned} \qquad [17]$$

i and j at transition rates $\lambda_{ij}(t)$; see eqn [17]. It is assumed that two or more transitions do not occur simultaneously, and that transitions in different time intervals occur independently. Typically, the number of compartments S is finite. In many applications, the total size of the population $\sum_{s=1}^{S} X_s(t) = N$ is bounded at N. When N is unknown (or unbounded) as in a pure birth process (say, which has $S = 1$ state corresponding to the size of the population represented by the number of births), then in the context of [17], the 'source' of new births can be thought of as transitions from a hypothetical state $s = 0$. Similarly, a hypothetical state $s = S^* = S + 1$ can be incorporated if necessary. Thus, for a birth and death process, $S = 1$, but births correspond to transitions from an $s = 0$ state and deaths reflect transitions from $s = 1$ to a state $s = S^* = 2$. Markov processes are governed by a system of differential-difference equations obtained from the basic Chapman–Kolmogorov

relations [12]. To illustrate, for a birth and death process, these equations become

$$\frac{\mathrm{d}}{\mathrm{d}t}p_x(t) = -(\lambda_x + \mu_x)p_x(t) + \lambda_{x-1}p_{x-1}(t) + \mu_{x+1}p_{x+1}(t), \quad x = 1, 2, \ldots, \qquad [18]$$

$$\frac{\mathrm{d}}{\mathrm{d}t}p_0(t) = -\mu_0 p_0(t) + \mu_1 p_1(t), \quad x = 0,$$

where

$$p_x(t) = P[X(t) = x|X(0) = a] \qquad [19]$$

and where λ_x and μ_x are the transition rates of births and deaths, respectively, given the population is of size $X(t) = x$. The initial conditions are $p_a(0) = 1$ and $p_x(0) = 0$ for $x \neq a$.

When the birth and death rates are linearly proportional to the population size, i.e., when $\lambda_x = \lambda x$ and $\mu_x = \mu x$, say, then the solution is, for $a = 1$, and $\lambda \neq \mu$,

$$p_x(t) = \begin{cases} (1-\alpha)(1-\beta)\beta^{x-1}, & x \geq 1, \\ \alpha, & x = 0, \end{cases} \qquad [20]$$

where

$$\alpha = \mu(\delta - 1)/(\lambda\delta - \mu), \quad \beta = \lambda(\delta - 1)/(\lambda\delta - \mu)$$

$$\delta = \exp[(\lambda - \mu)t].$$

Thus, when $x \geq 1$, the population size is modeled by a geometric distribution.

When $\lambda = \mu$ and $a = 1$,

$$p_x(t) = \begin{cases} (\lambda t)^{x-1}/(1+\lambda t)^{x+1}, & x \geq 1, \\ (\lambda t)/(1+\lambda t), & x = 0. \end{cases} \qquad [21]$$

In the particular case that the birth rate is a constant (i.e., $\lambda_x = \lambda$) and the death rate is zero (i.e., $\mu_x = 0$), the solution of eqn [18] gives

$$p_x(t) = P[X(t) = x] = e^{-\lambda t}(\lambda t)^x/x!, \quad x = 0, 1, \ldots, \qquad [22]$$

where, without loss of generality, $a = 0$. This is the Poisson process. An implication of these results is that the waiting time between occurrences of events is exponentially distributed. Thus, in the Poisson process, the average waiting time between events (whose total number follows a Poisson distribution) is $(1/\lambda)$.

In the linear birth and death process, the average size of the population at time t is

$$E\{X(t)\} = \begin{cases} a\delta, & \lambda \neq \mu, \\ a, & \lambda = \mu, \end{cases} \qquad [23]$$

with variance

$$\mathrm{Var}\{X(t)\} = \begin{cases} a\delta(\delta-1)(\lambda+\mu)/(\lambda-\mu), & \lambda \neq \mu, \\ 2a\lambda t, & \lambda = \mu. \end{cases} \qquad [24]$$

The average waiting time for a birth, or a death, to occur given that a transition does occur is $\lambda_x/(\lambda_x + \mu_x)$, or $\mu_x/(\lambda_x + \mu_x)$, respectively. The probability of eventual extinction is 1 if $\lambda \leq \mu$, and $(\mu/\lambda)^a$ if $\lambda > \mu$.

Corresponding results for a pure linear birth (or death) process can be obtained by substituting $\mu = 0$, (or $\lambda = 0$, respectively), in eqns [18], [20], [23], and [24].

When there are S states, the system of eqn [18] is generalized as follows. Let $\mathbf{X}(t) = ((X_1(t), \ldots, X_S(t))$ be the random variable representing the number of individuals in each state $s = 1, \ldots, S$, at time t. Let $p_{\mathbf{x}}(t) = P(\mathbf{X}(t) = \mathbf{x}|\mathbf{X}(0) = \mathbf{a})$, where $\mathbf{x} = (x_1, \ldots, x_s)$ is a specific value of $\mathbf{X}(t)$. Write $x_{ij} = (x_1, \ldots, x_i + 1, \ldots, x_j - 1, \ldots, x_s)$, i.e., $x_i \equiv x_i + 1$ and $x_j \equiv x_j - 1$ and all other $x_k \equiv x_k$; and likewise for the $\lambda_{ij}(t)$. Then, the Markov process is governed by the system of differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t}p_x(t) = -\sum_{i,j=1}^{S} \lambda_{ij}(t)p_x(t) + \sum_{i,j=1, i\neq j} \lambda_{i+1,j-1}p_{x_{ij}}(t) \qquad [25]$$

for x in the S-dimensional space of possible values. Typically, there exist some boundary conditions on one or more of the values x_s, $s = 1, \ldots, S$, which necessitate a separate equation for that x_s value. This occurred in the linear birth and death process modeled in eqn [18], when $x = 0$. Models for which $S > 1$ are sometimes called interactive Markov models.

In general, except when the transition rates $\lambda_{ij}(t)$ are both independent of time and linear (in one only of i and j, but not in both, such as when, e.g., $\lambda_{ij}(t) = \lambda i$ but not λij), the existence and derivation of solutions to eqn [25] are nontrivial tasks. However, when all the transitions across states flow in the same direction, i.e., when in [2], $\lambda_{ij}(t) = 0$ for all $j < i$ to give the so-called right-shift process (or, equivalently, a left-shift process if $\lambda_{ij}(t) = 0$ for all $i < j$), then the recursion theorem of Severo (1969) can be applied successfully to obtain the underlying state probabilities from which all other properties can be obtained. Although written in the context of the progression of HIV to AIDS, Billard and Zhao (1994) give a general solution for time-dependent transitions moving to the next state, $j = i + 1$. Nevertheless, the solutions are still computationally intensive.

Unfortunately, since many applications have nonlinear transition rates, these difficulties cannot be dismissed lightly. For example, the spread of epidemics and rumors is modeled invariably with nonlinear rates. For simplicity, suppose $S = 3$ with states $s = 1$, $s = 2$, and $s = 3$ corresponding, respectively, to the number of susceptibles, infectives, and removals (either by recovery, death, immunity, etc.). Then, the rate that a susceptible person becomes infected is proportional to both the number of susceptible people and the number of infective people; thus $\lambda_{12}(t) = \beta x_1 x_2$ (where β is the infection rate), and is a nonlinear function of the underlying variables. More generally, there may be several stages of infection and/or removals, leading to an $S > 3$. The spread of rumors and news propagation corresponds to infection spread. Likewise, friendship formations, and indeed network systems and population modeling more generally, typically have nonlinear rates and are modeled as Markov processes.

If it is known that ultimately (as time t extends to ∞) steady-state conditions prevail, then, the derivatives in eqn [18] or [25] become $\mathrm{d}p_x(t)/\mathrm{d}t \equiv 0$ with $P_x(t) \equiv P_x$. The resulting equations can then be solved quite easily to obtain expressions for p_x. Notice these do not provide information about the process at specific times t before the steady-state conditions occur.

Some authors have attempted to work with the deterministic versions of the stochastic eqn [25]. In this case, instead of the probabilities $p_x(t)$, the $\mathbf{X}(t)$ are set as fixed rather than

random variables, and these $\mathbf{X}(t)$ replace the $p_x(t)$ in eqn [25]. Great care is needed here. While the solution to the deterministic equations equals the expected value of the stochastic variable when the λ_{ij} are linear rates, the equivalence does *not* hold for nonlinear transition rates. However, the deterministic result may provide a reasonable approximation when the population size increases to infinity ($N \to \infty$) for a fixed time t. This does not hold in the reverse case (as $t \to \infty$ for fixed N). Bartholomew (1984) and Lehoczky (1980) provide an excellent discussion of this issue.

Some Other Models

It can happen for a specific social process and/or a particular data set that a unique feature exists, which precludes the direct application of the basic Markov model as described thus far. This gives rise to models, which are still Markovian in nature but have an added characteristic.

One such model is the mover–stayer model, in which the population consists of two subgroups, the movers and the stayers. The movers are individuals who move (with nonzero probability) between the states according to a Markov chain process, while the stayers are those who always stay in their initial state. Thus, if θ is the proportion of stayers in state s, $s = 1,\ldots, S$, then the transition probability of moving from state i at t to state j at $t + n$ is

$$p_{ij}^{*} = \theta_i \delta_{ij} + \left(1 - \theta_i\right) p_{ij}^{(n)} \qquad [26]$$

where $p_{ij}^{(n)}$ is the n-step-ahead transition probability (found from eqn [3]) and δ_{ij} is the Kronecker delta ($\delta_{ij} = 0$ if $i \neq j$, and 1 if $i = j$). The $p_{ij}^{(n)}$ are estimated by the methods of Section Discrete Time Models: Markov Chains. The mle estimator of θ_s is

$$\widehat{\theta}_s = 1 - \left[1 - n_s \widehat{p}_{ss}^{T} \big/ n_s(0)\right] \Big/ \left(1 - \widehat{p}_{ss}^{T}\right) \qquad [27]$$

where $\widehat{p}_{ss}$ is the mle estimator of p_{ss}, $n_s(0)$ is the number of individuals in states s at time $t = 0$, and n_s is the number of individuals (stayers) who were in state s for all times $t = 0, 1,\ldots, T$; see Frydman (1984). Sampson (1990) considers a restricted model and illustrates the model with data describing movement of men across seven employment sectors.

While the mover–stayer model accommodates some heterogeneity, more generally there may be K subpopulations. The underlying process consists of a mixture of K Markov chains giving rise to a mixed Markov chain model. Thus, a mover–stayer model is the special case that $K = 2$, and that one of the tpm's is the identity matrix $\mathbf{I}$. Vogel (1996) considers a mixture of $K = 2$ populations, those who do and those who do not enter into extensive plea bargaining in the criminal setting. In a mixed model, individuals can only move within their own subpopulation. When they are able to move out of one subpopulation into another, the model becomes a latent Markov chain model. See Langeheine and Pol (1990) for a detailed discussion of these models.

A semi-Markov model is one that is specified by both the transition probabilities ($\mathbf{P}$ as in a Markov chain) and the distribution (specifically the parameters) of the length of time R_{ij} an individual stays in a state i before moving to state j, $i, j = 1,\ldots, S$ (which need not equal the geometric waiting time distribution of the standard Markov chain). Note that transitions and waiting times depend only on the current state and are conditionally independent of previously visited states; see, e.g., Weiss et al. (1982).

Conclusion

This article has described briefly the basic principles underlying Markov models. Markov models are particularly useful to describe a wide variety of behavior such as consumer behavior patterns, mobility patterns, friendship formations, networks, voting patterns, environmental management (e.g., patient movement between hospital stations), etc., among innumerable other applications in the social sciences. Four particular examples have been provided herein. The article also included brief summaries of how data governed by these Markov processes can be assessed and modeled. The texts by Bartholomew (1981, 1982) and Bartholomew and Forbes (1979), along with the review articles by Anderson (1954), Goodman (1962), Singer and Spilerman (1976), and Cohen and Singer (1979), among others, provide more extensive details of these models and their applications.

See also: AIDS: Acquired Immune-Deficiency Syndrome; Consumer Psychology; Data Bases and Statistical Systems: Crime; Fertility and Social Networks; Friendship During Adolescence and Cultural Variations; Friendship During Childhood and Cultural Variations; Markov Processes in Learning Spaces; Mathematical Models in Geography; Mortality and the HIV/AIDS Epidemic; Mortality, Epidemiological, and Health Transitions; Multistate Transition Models in Demography; Network Analysis; Network Models of Tasks; Population Dynamics: Mathematic Models of Population, Development, and Natural Resources; Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems; Probability Theory: Formal; Religion and Health: Clinical Considerations and Applications; Social Epidemiology; Social Integration, Social Networks, and Health; Social Network Analysis; Social Network Models: Statistical; Stochastic Models.

Bibliography

Anderson, T.W., 1954. Probability models for analyzing time changes in attitudes. In: Lazarsfeld, P.F. (Ed.), Mathematical Thinking in the Social Sciences. The Free Press, Glencoe, IL, pp. 17–66.

Anderson, T.W., Goodman, L.A., 1957. Statistical inference about Markov chains. Annals of Mathematical Statistics 28, 89–110.

Bartholomew, D.J., 1981. Mathematical Methods in Social Science, Guidebook 1. Wiley, Chichester, UK.

Bartholomew, D.J., 1982. Stochastic Models for Social Processes, third ed. Wiley, Chichester, UK.

Bartholomew, D.J., 1983. Some recent developments in social statistics. International Statistical Review 51, 1–9.

Bartholomew, D.J., 1984. Recent developments in nonlinear stochastic modelling of social processes. Canadian Journal of Statistics 12, 39–52.

Bartholomew, D.J., Forbes, A.F., 1979. Statistical Techniques for Manpower Planning. Wiley, Chichester, UK.

Biblarz, T.J., Raftery, A.E., 1993. The effects of family disruption on social mobility. American Sociological Review 58, 97–109.
Billard, L., Dayananda, P.W.A., 2014. A multi-stage compartmental model for HIV-infected individuals I: waiting time approach. Mathematical Biosciences 249, 102–109.
Billard, L., Meshkani, M.R., 1995. Estimation of a stationary Markov chain. Journal of the American Statistical Association 90, 307–315.
Billard, L., Zhao, Z., 1994. Nonhomogeneous Markov models for the AIDS epidemic. Journal of the Royal Statistical Society, B 56, 673–686.
Bishop, Y.M.M., Fienberg, S.E., Holland, P.W., 1975. Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge, MA.
Brown, P.J., Payne, C.D., 1986. Aggregate data, ecological regression, and voting transitions. Journal of the American Statistical Association 81, 452–460.
Chatfield, C., 1973. Statistical inference regarding Markov chain models. Applied Statistics 22, 7–20.
Cohen, J.E., Singer, B., 1979. Malaria in Nigeria: constrained continuous-time Markov models for discrete-time longitudinal data on human mixed-species infections. In: Levin, S.A. (Ed.), Some Mathematical Questions in Biology. American Mathematical Society, Washington, DC, pp. 69–133.
Draper, J.E., Nolin, L.H., 1964. A Markov chain analysis of brand preferences. Journal of Advertising Research 4, 33–38.
Frydman, H., 1984. Maximum likelihood estimation in the mover-stayer model. Journal of the American Statistical Association 79, 632–638.
Goodman, L.A., 1962. Statistical methods for analyzing processes of change. American Journal of Sociology 68, 57–78.
Geweke, J., Marshall, R.C., Zarkin, G.A., 1986. Mobility indices in continuous time Markov chains. Econometrica 54, 1407–1423.
Langeheine, R., Van de Pol, F., 1990. A unifying framework for Markov modeling in discrete space and discrete time. Sociological Methods and Research 18, 416–441.
Lawrence, R.J., 1966. Models of consumer purchasing behaviour. Applied Statistics 15, 216–233.
Lehoczky, J.P., 1980. Approximations for interactive Markov chains in discrete and continuous time. Journal of Mathematical Sociology 7, 139–157.
Meshkani, M.R., Billard, L., 1992. Empirical Bayes estimators for a finite Markov chain. Biometrika 79, 185–193.
Sampson, M., 1990. A Markov chain model for unskilled workers and the highly mobile. Journal of the American Statistical Association 85, 177–180.
Severo, N.C., 1969. A recursion theorem on solving differential-difference equations and applications to some stochastic processes. Journal of Applied Probability 6, 673–681.
Shorrocks, A.F., 1978. The measurement of mobility. Econometrica 46, 1013–1024.
Singer, B., Spilerman, S., 1975. Identifying structural parameters of social processes using fragmentary data. Bulletin International Statistical Institute 46, 681–697.
Singer, B., Spilerman, S., 1976. The representation of social processes by Markov models. American Journal of Sociology 82, 1–54.
Stasny, E.A., 1986. Estimating gross flows using panel data with nonresponse: an example from the Canadian labor force survey. Journal of the American Statistical Association 81, 42–47.
Vogel, M.E., 1996. The negotiated guilty plea: vacancies as an alternative to the caseload pressure explanation. Journal of Mathematical Sociology 21, 241–288.
Weiss, E.N., Cohen, M.A., Hershey, J.C., 1982. An iterative estimation and validation procedure for specification of semi-Markov models with application to hospital flow. Operations Research 30, 1082–1104.

Markov Processes in Learning Spaces

Jean-Claude Falmagne, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A learning space is a family of sets satisfying certain conditions. The sets are called knowledge states. In the most common application of these concepts, a knowledge state is a set of items of knowledge, such as all the concepts mastered by a student in some part of mathematics. Two applications of Markov processes to learning spaces are described. The most important one is the core component of an assessment system designed to uncover the knowledge state of the student by systematic questioning. A learning space can also be used as a teaching instrument. The student masters the items one by one, progressively increasing the size of his or her knowledge state. The second Markov process is a stochastic mechanism tracking the gradual learning of an item.

A *knowledge structure* is a family $\mathscr{K}$ of subsets of a basic set Q of *items* of knowledge. The set Q is called the *domain* of the knowledge space. The sets in $\mathscr{K}$ are the *knowledge state*, or simply the *states*. It is assumed that both the empty set $\varnothing$ and the domain Q are states. A knowledge structure is *finite* if its domain is finite.

As our typical example, we take Q to be the set of all the types of problems forming a curriculum in beginning algebra. Each state K in $\mathscr{K}$ represents a possible competence of a student in beginning algebra: the student in state K has mastered all the items in K and none of the items in Q, K. A student in state $\varnothing$ does not know anything in beginning algebra; a student in state Q has mastered the full curriculum in that subject. A knowledge structure is a *knowledge space* if it is closed under finite union, that is, for any two subsets K and L of Q, if both K and L are in $\mathscr{K}$, then so is $K \cup L$. A *learning space* is a special case of a knowledge space playing an important role in education. This concept is discussed in Sections The Learning Spaces and A Miniature Example.

Two Markov-type processes are used in the practical application of these concepts. One is the core component of the *assessment engine* designed to uncover the knowledge state of a student by systematic questioning (see Section The Assessment Engine). The other is a Markov mechanism monitoring the gradual learning of an item by a student (see Section The Learning Engine). Section Evaluating the Efficiency of the System reviews some methods used to evaluate the efficiency of the algorithms. Background material and references are covered in Section Background and References.

The Learning Spaces

An important special case of knowledge structures are the *learning spaces*. A knowledge structure $\mathscr{K}$ on a domain Q is a *learning space* if it satisfies the following two principles:

[L1] Learning Smoothness. If K and L are two states such that $K \subset L$, then there exists a finite chain of states $K_0 = K \subset K_1 \subset \ldots \subset K_n = L$ such that $|K_{i+1}, K_i| = 1$, for $0 \leq i \leq n-1$.

In words: *Stepwise learning is possible. Any knowledge state can be reached from a less advanced state by mastering the items one by one.*

[L2] Learning Consistency. If K and L are two states satisfying $K \subset L$ and $q \in Q, K$ is an item such that $K \cup \{q\} \in K$, then $L \cup \{q\} \in K$.

In words: *If a student is capable of mastering an item, then a student in a more advanced state may already have mastered this item. Otherwise, that more advanced student must also be capable of mastering it.*

However obvious these two principles may seem from an educational standpoint, they put severe constraints not only on the structure, but also on the choice of the items and the manner of writing them.

Note that a learning space is necessarily a finite knowledge structure. Indeed, because both $\varnothing$ and Q are states, principle [L1] requires that there be a finite chain $K_0 = \varnothing \subset K_1 \subset \ldots \subset K_m = Q$, with $|K_{i+1}, K_i| = 1$, for $0 \leq i \leq m-1$. This implies that $|Q| = m$. It can be shown that any learning space is a knowledge space, that is, it is closed under union.

While actual knowledge structures are typically very large, numbering millions of states, they are in fact relatively small in comparison with the set of all subsets of the domain. In the case of our example of beginning algebra, the curriculum contains about 350 types of algebra problems that a student must master. This means that the size of the domain Q is 350. So, the number of subsets of Q is 2^{350}. The proportion of such subsets that are states is minuscule.

The two principles [L1] and [L2] have important consequences. A critical one is that any state can be represented accurately by its two *fringes*, which are typically much smaller sets. So, the fringes are an efficient summary of a state.

The *outer fringe* of a state K in a learning space $\mathscr{K}$ is the set K^{o} of all items in Q, K such that, for any q in K^{o}, the set $K \cup \{q\}$ is also a state of $\mathscr{K}$. Such a state $K \cup \{q\}$ is thus a possible immediate successor of the state K in a learning sequence. Assuming that the learning space is an accurate representation of the cognitive organization of the items, we can interpret the outer fringe of a state K as the set of all those items that a student in that state is *ready to learn*.

The *inner fringe* of a nonempty state K is the set K^{I} of all items q in K such that the set $K\{q\}$ is also a state of $\mathscr{K}$. Such a state $K\{q\}$ is thus a possible immediate predecessor of the state K. The outer fringe and the inner fringe of a state are thus corresponding concepts.

A Miniature Example

We illustrate the concepts of learning space and outer fringe by an example using five items in beginning algebra, labeled *a*, *c*, *g*, *h*, and *i*, which are described in Table 1. These five items form a miniature learning space that is part of the considerably larger learning space currently used by several hundred thousand students in the United States and abroad (see Section Background and References). This miniature learning space is pictured by the graph of Figure 1.

Each ellipse of this graph represents one of the 16 knowledge states in that learning space. The blank circle at the top stands for the empty state. Learning proceeds from top to bottom, the items being mastered successively. So, the arrows represent the precedence relation between the states. The items *a* and *c* form the *outer fringe* of the state pictured by the red ellipse. This state contains the items *g*, *h*, and *i*. The outer fringe is made of the items that the student is *ready to learn*. The *inner fringe* of that state is made of the items *g*, *h*, and *i*. Each of these items could be the last one that the student has mastered. In

Table 1 Five items in beginning algebra with an instance

Item	*One sample instance*[a]
a. Quotients of expressions involving exponents	Simplify the following: $\frac{a^4 b^5}{5a^6 b}$
c. Plotting a point in the coordinate plane using a virtual pencil on a Cartesian graph	Using the pencil, plot the point $(-4, 7)$.
g. Multiplication of a decimal by a whole number	Multiply. $\begin{array}{r} 2.95 \\ \times \quad 5 \\ \hline \end{array}$
h. Integer addition (introductory problem)	$-4 + 4 = \square \quad -7 + 5 = \square \quad -3 + (-4) = \square$
i. Equivalent fractions	Fill in the blank to make the two fractions equivalent: $\frac{3}{5} = \frac{\square}{15}$

[a]An instance is a particular case of an item.

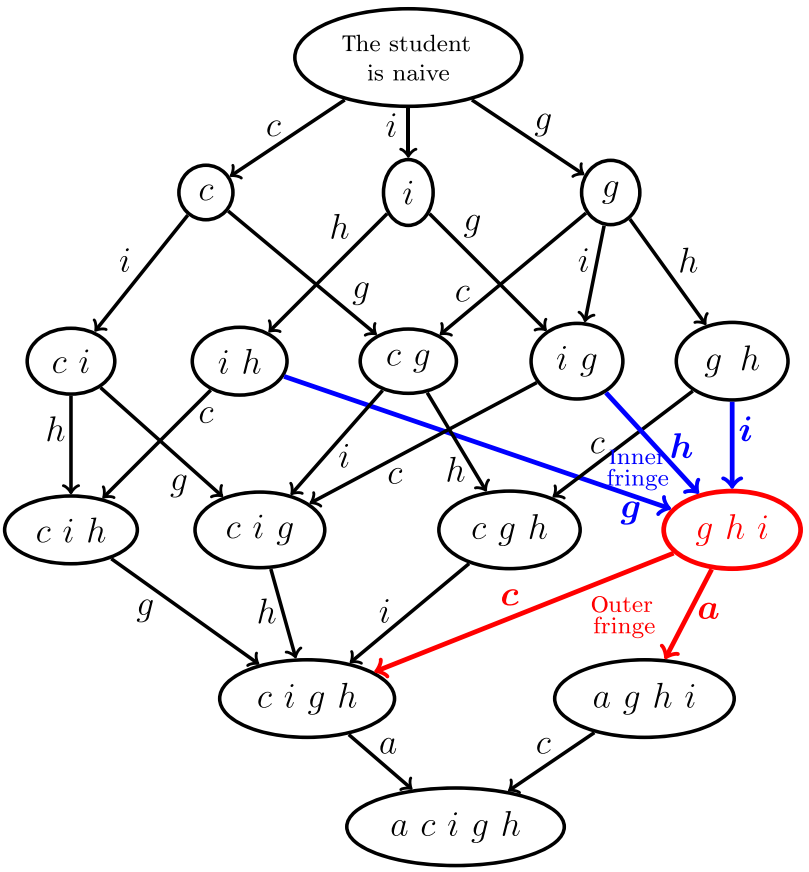

Figure 1 Miniature learning space on the items *a*, *c*, *i*, *g*, and *h*.

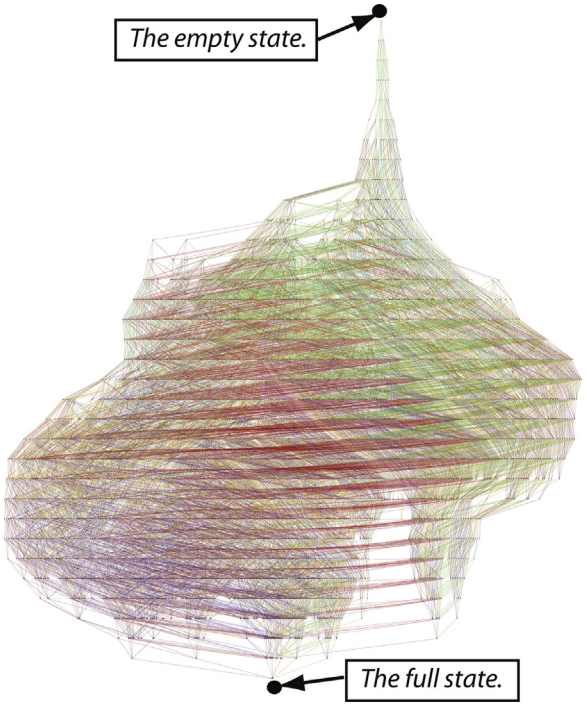

Figure 2 Graph of the learning space of 50 items in beginning algebra, in the style of the five-item graph of **Figure 1**. Every point in the graph represents a state. The colors of the lines linking the states correspond to the various subtopics of beginning algebra. The graph of **Figure 1** is a small portion of this graph, with its circles represented by points here. Drawing the full graph of the 350 items of beginning algebra is impractical. Reproduced from figures in Chapters 1 and 2 of Falmagne et al. (2013), with permission of the publisher.

this simple case, the inner fringe is the same as the whole state, but this is rarely the case in realistic learning spaces.

The five items are listed in Table 1, together with one exemplary *instance* of each item. An instance is a special case of an item. Typically, there are thousands of instances for each item. Whenever an item is used in an assessment, one instance is randomly selected and proposed to the student. These five items are part of a collection of 350 items in beginning algebra.

Drawing the same type of graph for the whole of beginning algebra, which contains about 350 items, is not only daunting but also actually impractical. Some idea of the difficulty is conveyed by the graph of Figure 2, which represents the learning space of just 50 items in beginning algebra.

Yet, the assessment engine and the learning engines of the system are capable of efficiently managing such enormous structures, and in fact the much larger structure of the full beginning algebra and similar scholarly curricula.

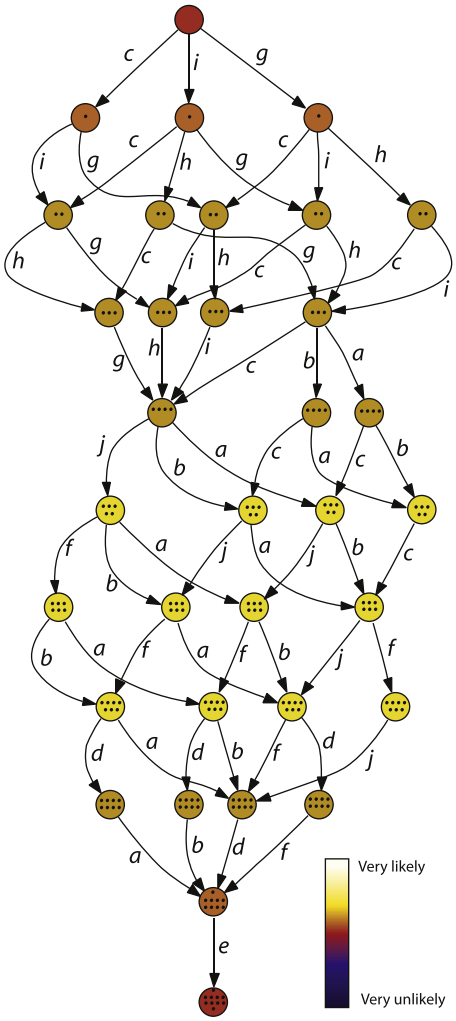

Figure 4 Initial probabilities of all the states at the beginning of the assessment. The likelihood is marked by the shading of the circles representing the states. Bluish black means very unlikely, and bright yellow means very likely. Reproduced from figures in Chapters 1 and 2 of Falmagne et al. (2013), with permission of the publisher.

The Assessment Engine

The task of the assessment engine is to uncover, by efficient questioning, the knowledge state of a student under examination. As the number of items in a typical learning space is large, only a small part of the items can be asked, in practice, in any assessment. So, the successive questions must be cleverly chosen. The procedure follows a scheme outlined in Figure 3.

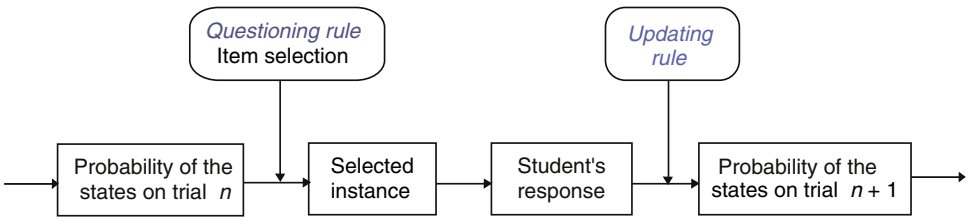

Figure 3 Diagram of the transitions in the assessment procedure. Four successive operations are involved: the selection of a maximally informative item, the choice of an instance, the recording of the student's response and its verification, and the updating of the likelihood distribution.

It has two main components: the *questioning rule* and the *updating rule*.

At the outset of the assessment, each knowledge state is assigned a certain initial probability, which may depend upon the school year of the student, if it is known, or on some other information. These initial probabilities represent how much the assessment engine knows about the student before any question is asked. They play no role in the final result of the assessment but may be helpful in shortening it.

These initial probabilities are pictured on the graph on the left by colors varying from reddish to yellow. Except that the ellipses are circles here, this graph is similar to that of Figure 1, which is in fact part of it. So, each circle represents a state. This graph represents a learning space on the 10 items *a*, *b*, …, *j*, five of which are described in Table 1. The color of each circle represents the probability of the corresponding state. The full color scale indicated at the bottom of the graph goes from bluish black for very unlikely to bright yellow for very likely. The somewhat lighter colors of some circles show that the system thinks the student might have mastered about five to seven items.

The system then asks an instance of the first item, which is chosen to be as informative as possible, on the basis of the current probabilities of the states. The adopted rule is to ask the student an item that has about a 50% chance of eliciting a correct response, on the basis of the current probabilities.

Suppose that the first item asked is *a*, and the student's response is correct. The updating rule will then increase the probabilities of all the states containing *a* and decrease the probabilities of all those states not containing *a*. This is pictured by the leftmost graph of Figure 5 in which the colors of the circles are brightened and darkened accordingly. If the next item asked is *f* and the response is incorrect, the updating rule will then decrease the probabilities of the state containing *f* and increase the probabilities of the states not containing *f*. The middle graph pictures the result of such an updating. At the end, the situation may be as pictured by the graph on the right. Only one state remains with a very high probability. The system will then choose the state containing *a*, *b*, *g*, *h*, and *i* as being the current state of the student.

We now give the formal description of the stochastic process. For details, see Falmagne and Doignon (2011, Chapter 13).

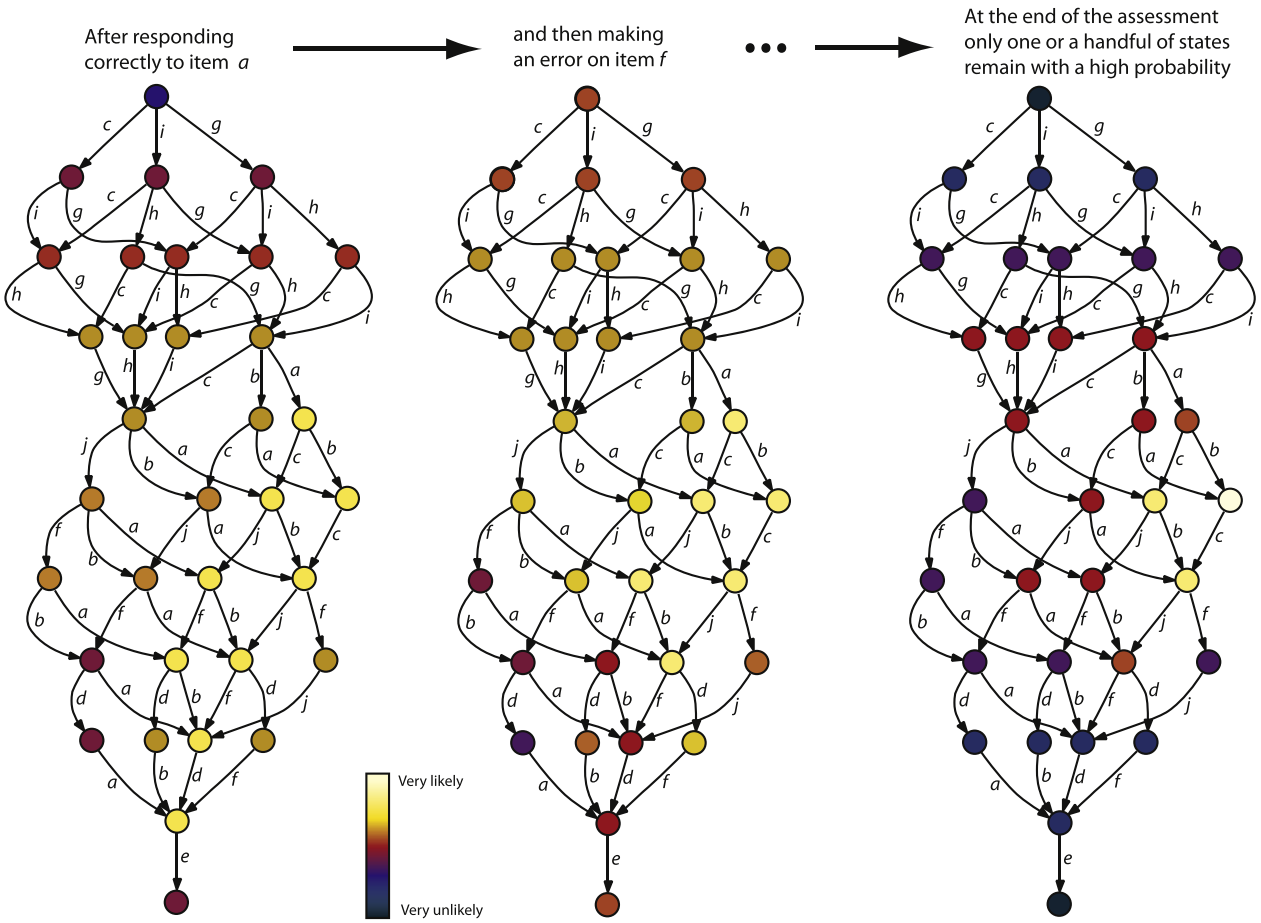

Figure 5 On the left, the effect of a correct response to item *a* on the state's probabilities is marked by brightening the yellow shades for states containing *a* and darkening those shades for those states not containing *a*. The graph in the middle depicts the subsequent effect of an incorrect response to item *f*. The graph on the right represents a possible situation at the end of the assessment. The assessment engine will pick the state containing *a*, *b*, *g*, *h*, and *i* as being the current state of the student. Reproduced from figures in Chapters 1 and 2 of Falmagne et al. (2013), with permission of the publisher.

Basic Concepts

Given a learning space $\mathscr{K}$ on a domain Q, any assessment is a realization of a Markovian stochastic process. We use the following notation:

n	the step number, or *trial* number, $n = 1, 2, \ldots$;
$\mathscr{K}_q$	the subcollection of $\mathscr{K}$ containing all the states containing q;
Λ_+	the set of the positive probability distributions on $\mathscr{K}$;
$\mathbf{L}_n$	a random probability distribution on $\mathscr{K}$; we have $\mathbf{L}_n = L_n \in \Lambda_+$ (so $L_n > 0$) if L_n is the probability distribution on $\mathscr{K}$ at the beginning of trial n;
$\mathbf{L}_n(K)$	a random variable measuring the probability of state K on trial n;
$\mathbf{Q}_n$	a random variable representing the question asked on trial n; we have $\mathbf{Q}_n = q \in Q$ if q is the question asked on trial n;
$\mathbf{R}_n$	a random variable coding the response on trial n: $\mathbf{R}_n = \begin{cases} 1 & \text{if the response is correct} \\ 0 & \text{otherwise} \end{cases}$
$\mathbf{W}_n$	the random history of the process from trial 1 to trial n;
ι_A	the indicator function of a set A: $\iota_A(q) = \begin{cases} 1 & \text{if } q \in A \\ 0 & \text{if } q \notin A \end{cases}$;
$\zeta_{q,r}$	with $1 < \zeta_{q,r}$ for $q \in Q$, $r = 0, 1$ a family of parameters specifying the updating operator (see Axiom [U] following).

The process begins, on trial 1, by setting $\mathbf{L}_1 = L \in \Lambda_+$. So, the initial probability distribution is the same for any realization. Any further trial $n > 1$ begins with a value $L_n \in \Lambda_+$ of the random distribution $\mathbf{L}_n$ updated as a function of the event on trial $n-1$. We write for any $\mathscr{F} \subseteq \mathscr{K}$,

$$L_n(\mathscr{F}) = \sum_{K \in \mathscr{F}} L_n(K). \qquad [1]$$

Three general axioms specify the stochastic process ($\mathbf{L}_n$, $\mathbf{Q}_n$, $\mathbf{R}_n$). The version of these axioms given below is an important special case.

[U] **Updating Rule.** We have $\mathbb{P}(\mathbf{L}_1 = L) = 1$, and for any positive integer n, with $\mathbf{L}_n = L_n$, $\mathbf{Q}_n = q$, $\mathbf{R}_n = r$, and

$$\zeta_{q,r}^K = \begin{cases} 1 & \text{if } \iota_K(q) \neq r, \\ \zeta_{q,r} & \text{if } \iota_K(q) = r, \end{cases} \qquad [2]$$

we have

$$L_{n+1}(K) = \frac{\zeta_{q,r}^K L_n(K)}{\sum_{K' \in \mathscr{K}} \zeta_{q,r}^{K'} L_n(K')}. \qquad [3]$$

This updating rule is called *multiplicative with parameters* $\zeta_{q,r}$.

[Q] **Questioning Rule.** For all $q \in Q$ and all positive integers n,

$$\mathbb{P}(\mathbf{Q}_n = q | \mathbf{L}_n, \mathbf{W}_{n-1}) = \frac{\iota_{S(L_n)}(q)}{|S(L_n)|}$$

where $S(L_n)$ is the subset of Q containing all those questions q minimizing

$$|2L_n(\mathscr{K}_q) - 1|.$$

Under this questioning rule, which is called *half-split*, we must have $\mathbf{Q}_n \in S(L_n)$ with a probability equal to 1. The questions in the set $S(L_n)$ are then chosen with equal probability.

[R] **Response Rule.** For all positive integers n,

$$\mathbb{P}(\mathbf{R}_n = \iota_{K_0}(q) | \mathbf{Q}_n = q, \mathbf{L}_n, \mathbf{W}_{n-1}) = 1$$

where K_0 is the *latent state* representing the set of all the items mastered by the student, that is, the state that must be uncovered by the Markovian procedure.

So, if the item selected by the process belongs to the latent state K_0, the probability of a correct response is equal to 1. In the more realistic versions of this axiom used in practice, one additional parameter is used that specifies the probability of a *careless error*. It may also be necessary to introduce a *lucky guess* parameter, for example, if a multiple choice paradigm is used.

Some Key Results

These results follow from axioms [U], [Q] and [R]. (For proofs, see Chapter 13 in Falmagne and Doignon, 2011.)

1. The updating operator specified by eqn [3] in axiom [U] is essentially a Bayesian operator. This can be shown by an appropriate transformation of the equation.
2. This updating operator is *permutable*. This term is used in the functional equations literature (Aczél, 1966) to designate a function F satisfying the equation

$$F(F(x, y), z) = F(F(x, z), y).$$

This property is essential because it means that the order of the questions has no import on the result of the assessment.

3. The stochastic process ($\mathbf{L}_n$) is Markovian.
4. The stochastic process ($\mathbf{L}_n$) converges to the latent state K_0 in the sense that

$$\mathbf{L}_n(K_0) \xrightarrow{\text{a.s.}} 1$$

(in which 'a.s.' means 'almost surely').

Remark. Another Markovian assessment procedure was also developed, which is based on a very different principle (Falmagne and Doignon, 1988b). The stochastic process is a finite Markov chain, the states of which are subsets of the domain Q of the learning space.

The Learning Engine

An important mathematical consequence of the two principles [L1] and [L2] defining a learning space is that, at the end of an assessment, the knowledge state is completely specified by its two fringes. In practice, the student is only given the list of items in the outer fringe of his or her knowledge state. These are the items that the student is *ready to learn*. This list can be shorter if the *intermediate objectives* feature of the system is used, which permits the teacher to establish a preferred progression of major topics. In such a case, the system gives the students only the part of the outer fringe that is consistent with the next intermediate objective.

The Random Walk Tracking Mechanism

In any event, the student chooses the next item to learn. This initiates the *Markovian Tracking Mechanism* illustrated by Figure 6. This system guides the student's progress. It monitors the sequence of successes and failures of the student in attempting to solve the successive instances of the item chosen via a random walk.

The student enters the random walk at the red point and is given an instance of the problem. In case of a failure to solve that instance, the student is given an explanation $\boxed{E}$ of the solution. A new instance of the problem is then presented. When the student gives a correct response to an instance, a new instance is also given. The random walk evolves until hitting one of the two absorbing barrier states represented by the blue and black dots in Figure 6. If the blue dot state is reached, the system decides that the problem is learned and the student leaves the random walk. The knowledge state of the student is then adjusted upward. On the other end, if the random walk hits the black dot state, the system decides that the item cannot be in the outer fringe of the student. The student leaves the random walk and either the knowledge state is reevaluated or a new assessment is triggered (depending on the cases). Because the items proposed to the students are picked in the outer fringe of the student's state, the proportion of failure is small in practice (see #2 in Section Evaluating the Efficiency of the System).

Evaluating the Efficiency of the System

Two evaluation methods have been used:

1. The extra question. In each assessment, an extra question is asked that is not used to determine the knowledge state. The student's response to the extra question can be predicted from the uncovered knowledge states. The correlation between the student's actual response to the extra question and the response predicted from the knowledge state uncovered is noticeably higher than similar statistics obtained in standardized tests.
2. The learning success probability. One may reasonably ask: what is the probability that the student successfully masters an item chosen in the outer fringe? In other words, what is the probability that the sequential tracking mechanism pictured in Figure 6 has the student end up at the blue exit (success) point?

The estimated median probability is .92. This value has been estimated on the basis of a very large number of learning occasions (see Section Background and References). What is meant by 'learning occasion' is that the student picks an item in the outer fringe, gets an instance, and starts learning. Statistics show that in later assessments, a student is capable of solving an instance of learned item in about 75% of the cases, which is high considering that different instances of the same item are typically used, and that careless errors are unavoidable. These numbers are based on the analysis of hundreds of thousands of learning occasions and assessments.

A third possibility may come to mind. We can mimic standardized testing and transform the state into a numerical score. This can be done simply by counting the number of items in the state uncovered by the assessment. The objection is that, in most cases, there are many thousand states having the same number of items. Collapsing all these states into the same numerical score results in a huge loss of information. Still, when such a collapse is performed, the predictive power of the numerical score is better than that of a standardized test such as the S.A.T. or the A.C.T. (see Ahlgren and Harper, 2011).

Background and References

The monograph by Falmagne and Doignon (2011) comprehensively presents the theory of knowledge spaces, and their important special case, the learning spaces. The

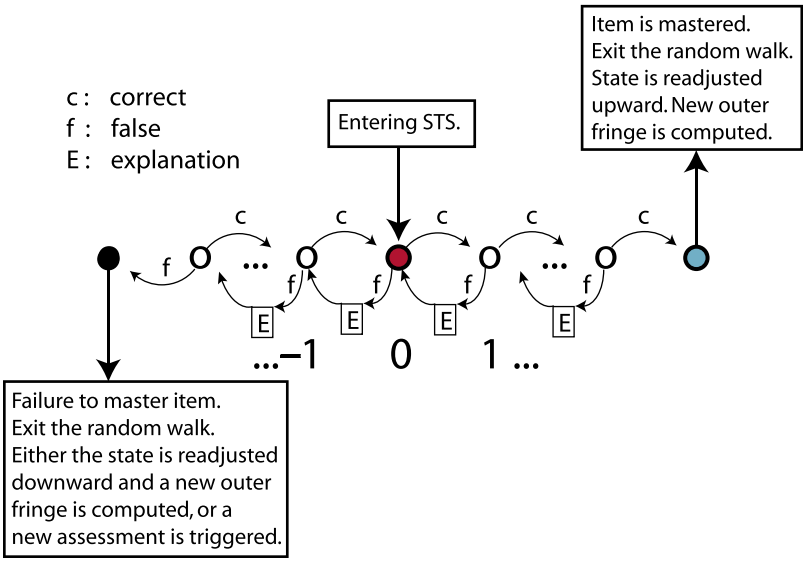

Figure 6 The random walk tracking mechanism with two absorbing states represented by the blue and the black dots. Reproduced from figures in Chapters 1 and 2 of Falmagne et al. (2013), with permission of the publisher.

theoretical developments of knowledge space theory were supported by several National Science Foundation (NSF) grants and one Department of Defense grant to Falmagne, first at New York University and later at the University of California, Irvine (UCI). The mathematical results supporting the stochastic assessment mechanism outlined in this article were published in Falmagne and Doignon (1988a). The basic idea of the updating algorithm was prompted by the Markovian learning models initiated by W.K. Estes (1950), and by R.R. Bush and F. Mosteller (1951). More specifically, the updating operator specified by eqn [3] was inspired by the *beta learning model* of Luce (1959). Other relevant references are Bush and Mosteller (1955), Estes and Suppes (1957), and Suppes and Ginsburg (1963). A general study of the Markov processes underlying all these learning models is presented in Norman (1972). For a general presentation of the mathematical theory of Markov processes, see Bharucha-Reid (1960). A description of some of the applications of knowledge spaces and some further theoretical development is contained in Falmagne et al. (2013).

The development of the educational software implementing the assessment engine and the learning engine on the Internet was supported by a large NSF grant to Falmagne at UCI. This software covers K12 mathematics and basic chemistry. It belongs to UCI and was licensed to the Aleks Corporation. ALEKS is an acronym for Assessment and LEarning in Knowledge Spaces. It is also the name of the software itself. Since its inception in 1999, the system has been used by millions of students in more than four thousand K12 schools, colleges, and universities not only in the United States, but also in Europe, Australia, and Latin America (the Aleks system is bilingual English–Spanish).

See also: Bayesian Theory, History, Applications, and Contemporary Directions; Cognition: Knowledge Spaces; Decision and Choice: Luce's Choice Axiom; Diffusion and Random Walk Processes; Knowledge Representation; Learning: Mathematical Learning Theory, History of; Markov Decision Processes; Psychometrics: Combinatorial Data Analysis; Psychometrics: Preference Models with Latent Variables; Psychometrics.

Bibliography

Aczél, J., 1966. Lectures on Functional Equations and Their Applications. Academic Press, New York.

Ahlgren, A., Harper, M., 2011. Assessment and placement through Calculus I at the University of Illinois. Notices of the American Mathematical Society 58 (10), 1460–1461.

Bharucha-Reid, A.T., 1960. Elements of the Theory of Markov Processes and Their Applications. Originally published by MacGraw-Hill Book Company, New York-Toronto-London. Reprinted by Dover Publications, Inc., Mineola, NY, 1997. xii+468 pp. ISBN: 0-486-69539-5 60-01 (60Jxx).

Bush, R.R., Mosteller, F., 1951. A mathematical model for simple learning. Psychological Review 58, 313–323.

Bush, R.R., Mosteller, F., 1955. Stochastic Models for Learning. John Wiley & Sons, New York.

Estes, W.K., 1950. Toward a statistical theory of learning. Psychological Review 57, 94–107.

Estes, W.K., Suppes, P., 1957. Foundations of linear models. In: Bush, R.R., Estes, W.K. (Eds.), Studies in Mathematical Learning Theory. Stanford University Press, Stanford, CA, pp. 137–179.

Falmagne, J.-Cl., Albert, D., Doble, C., Eppstein, D., Hu, X., 2013. Knowledge Spaces: Applications in Education. Springer, Heidelberg.

Falmagne, J.-Cl., Doignon, J.-P., 1988a. A class of stochastic procedures for the assessment of knowledge. British Journal of Mathematical and Statistical Psychology 41, 1–23.

Falmagne, J.-Cl., Doignon, J.-P., 1988b. A Markovian procedure for assessing the state of a system. Journal of Mathematical Psychology 32, 232–258.

Falmagne, J.-Cl., Doignon, J.-P., 2011. Learning Spaces, Interdisciplinary Mathematics Series. Springer, Heidelberg.

Luce, R.D., 1959. Individual Choice Behavior: A Theoretical Analysis. Wiley, New York.

Norman, F., 1972. Markov Processes and Learning Models. Academic Press, New York.

Suppes, P., Ginsburg, R., 1963. A fundamental property of all-or-none models, binomial distribution of responses prior to conditioning, with application to concept formation in children. Psychological Review 70, 139–171.

Maroons in Anthropology

Richard Price, College of William & Marry, Williamsburg, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Maroons – self-liberated slaves and their descendants – still form semi-independent communities in several parts of the Americas, for example, Suriname, French Guiana, Jamaica, Colombia, Belize, and Brazil. As the most isolated of Afro-Americans, they have since the 1920s been an important focus of anthropological research, contributing to theoretical debates about slave resistance, the heritage of Africa in the Americas, the process of creolization, and the nature of historical knowledge among nonliterate peoples.

Maroons – self-liberated slaves and their descendants – still form semi-independent communities in several parts of the Americas, for example, Suriname, French Guiana, Jamaica, Colombia, Belize, and Brazil. As the most isolated of Afro-Americans, they have since the 1920s been an important focus of anthropological research, contributing to theoretical debates about slave resistance, the heritage of Africa in the Americas, the process of creolization, and the nature of historical knowledge among nonliterate peoples.

Maroons and Their Communities

The English word 'Maroon' derives from Spanish *cimarrón* – itself based on an Arawakan (Taino) Indian root. *Cimarrón* originally referred to domestic cattle that had taken to the hills in Hispaniola, and soon after it was applied to American Indian slaves who had escaped from the Spaniards. By the end of the 1530s, the word was being used primarily to refer to Afro-American runaways. Communities formed by Maroons dotted the fringes of plantation America, from Brazil to Florida, from Peru to Texas. Usually called *palenques* in the Spanish colonies and *mocambos* or *quilombos* in Brazil, they ranged from tiny bands that survived less than a year to powerful states encompassing thousands of members that lasted for generations or even centuries.

Marronage became a response to enslavement in Africa as well, in the wake of the internecine wars fostered by the Atlantic slave trade. Fugitives from the chaos and violence that dominated much of West Central Africa, including runaways from slave coffles during the seventeenth- and eighteenth-century expansion of the trade, frequently founded Maroon settlements in marginal areas. Some had populations of several hundreds, lived by raiding more settled groups, or even fought against their former masters. But all remained sufficiently caught up in the political ebb and flow of the era that they did not maintain separate Maroon identities for more than a generation or two. During the eighteenth century, fugitives from the trade along the Upper Guinea coast also formed Maroon communities, sometimes on islands just offshore, but mass recapture after months or years was the norm.

In the Americas, the meaning and attractiveness of marronage differed for enslaved people in different social positions, varying with their perceptions of themselves and their situations, which were influenced by diverse factors such as their places of birth, the periods of time they had been in the New World, their task assignments as slaves, their family responsibilities, and the particular treatments they were receiving from overseers or masters, as well as more general considerations such as the proportions of blacks to whites in the region, the proportions of freedmen in the population, the natures of available terrain in which to establish communities, and the opportunities for manumission. Many Maroons, particularly men, escaped during their first hours or days in the Americas. Enslaved Africans who had already spent some time in the New World seem to have been less prone to flight. But Creole slaves who were particularly acculturated, who had learned the ways of the plantation best, seem to have been highly represented among runaways, often escaping to urban areas where they could pass as free because of their independent skills and ability to speak the colonial language.

The large number of detailed newspaper advertisements for runaway slaves placed by masters attests to the high level of planter concern, while at the same time affording the critical historian an important set of sources for establishing the profiles of Maroons, which varied significantly by historical period and country. Individual Maroons fled not only to the hinterlands – many, especially skilled slaves, escaped to urban centers and successfully melted into the population of freedmen – but also became maritime Maroons, fleeing by fishing boat or other vessel across international borders. In Haiti, Maroons played a signal role as catalysts in the Haitian Revolution (1791) that created the first nation in the Americas in which all citizens were free.

During the past several decades, anthropological fieldwork has underlined the strength of historical consciousness among the descendants of these rebel slaves and the dynamism and originality of their cultural institutions. Meanwhile, historical scholarship on Maroons has flourished, as new research has done much to dispel the myth of the docile slave. Marronage represented a major form of slave resistance, whether accomplished by lone individuals, by small groups, or in great collective rebellions. Throughout the Americas, Maroon communities stood out as an heroic challenge to white authority, as the living proof of the existence of a slave consciousness that refused to be limited by the whites' conception or manipulation of it. It is no accident that throughout the Caribbean today, the historical Maroon – often mythologized into a larger-than-life figure – has

become a touchstone of identity for the region's writers, artists, and intellectuals, the ultimate symbol of resistance and the fight for freedom.

Maroons and their communities can be seen to hold a special significance for the study of Afro-American societies. For while they were, from one perspective, the antithesis of all that slavery stood for; they were also a widespread and embarrassingly visible part of these systems. Just as the very nature of plantation slavery implied violence and resistance, the wilderness setting of early New World plantations made marronage and the existence of organized Maroon communities a ubiquitous reality.

Planters generally tolerated *petit marronage* – truancy with temporary goals such as visiting a friend or lover on a neighboring plantation. But in most slave-holding colonies, the most brutal punishments – amputation of a leg, castration, suspension from a meat hook through the ribs, slow roasting to death – were reserved for long-term, recidivist Maroons, and in many cases these were quickly written into law. Marronage on the grand scale, with individual fugitives banding together to create communities, struck directly at the foundations of the plantation system, presenting military and economic threats that often taxed the colonists to their very limits. Maroon communities, whether hidden near the fringes of the plantations or deep in the forest, periodically raided plantations for firearms, tools, and enslaved women, often permitting families that had formed during slavery to be reunited in freedom.

To be viable, Maroon communities had to be inaccessible, and villages were typically located in remote, inhospitable areas. In the southern United States, isolated swamps were a favorite setting, and Maroons often became part of Native American communities. In Jamaica, some of the most famous Maroon groups lived in the intricately accidented 'cockpit country,' where water and good soil are scarce but deep canyons and limestone sinkholes abound. In the Guianas, seemingly impenetrable jungles provided Maroons with a safe haven.

Successful Maroon communities learned quickly to turn the harshness of their immediate surroundings to their own advantage for purposes of concealment and defense. Paths leading to the villages were carefully disguised, and much use was made of false trails replete with dangerous booby traps. In the Guianas, villages set in swamps were approachable only by an underwater path, with other, false, paths carefully mined with pointed spikes or leading only to fatal quagmires or quicksand. In many regions, man traps, or even dog traps, were used in village defenses.

Throughout the hemisphere, Maroons developed extraordinary skills in guerrilla warfare. To the bewilderment of their colonial enemies, who attempted to employ rigid and conventional tactics learned on the open battlefields of Europe, these highly adaptable and mobile warriors took maximum advantage of confined environments, striking and withdrawing with great rapidity, making extensive use of ambushes to catch their adversaries in cross fire, fighting only when and where they chose, depending on reliable intelligence networks among non-Maroons (both slaves and white settlers), and often communicating by drums and horns.

In many cases, the beleaguered colonists were eventually forced to sue their former slaves for peace. In Brazil, Colombia, Cuba, Ecuador, Hispaniola, Jamaica, Mexico, Panama, Peru, Suriname, and Venezuela, for example, the whites reluctantly offered treaties granting Maroon communities their freedom, recognizing their territorial integrity, and making some provision for meeting their economic needs, in return for an end to hostilities toward the plantations and an agreement to return future runaways. Of course, many Maroon societies were crushed by massive force of arms, and even when treaties were proposed they were sometimes refused or quickly violated. Nevertheless, new Maroon communities seemed to appear almost as quickly as the old ones were exterminated, and they remained, from a colonial perspective, the 'chronic plague' and 'gangrene' of many plantation societies right up to final emancipation.

Today, the relationships between Maroon communities and the nation-states in which they live vary broadly. In much of Brazil or even Jamaica, communities remain proud of their Maroon heritage but, to an outsider, seem little different from the non-Maroon rural population that surrounds them. In contrast, along the rivers of Suriname, Maroons live in villages that outsiders often compare to precolonial Africa, where everything from language, dress, kinship, transportation, and laws are unique Maroon creations that instantly differentiate the population from others in the country. The tension between Maroon identity and national identity has become acute for many younger people who have moved from traditional territories to cities or even to foreign countries. Yet in some areas, such as the Saamaka villages along the Suriname river, both Maroon culture and identity remain vibrant.

African Origins, New World Creativity

The initial Maroons in any New World colony hailed from a wide range of societies in West and Central Africa – at the outset, they shared neither language nor other major aspects of culture. Their collective task, once off in the forests or mountains or swamplands, was nothing less than to create new communities and institutions, largely via a process of inter-African cultural syncretism. Those scholars, mainly anthropologists, who have examined contemporary Maroon life most closely seem to agree that such societies are often uncannily 'African' in feeling but at the same time largely devoid of directly transplanted systems. However 'African' in character, no Maroon social, political, religious, or aesthetic *system* can be reliably traced to a specific African ethnic provenience – they reveal rather their hybrid composition, forged in the early meeting of peoples of diverse African, European, and Amerindian origins in the dynamic setting of the New World (Price, 2008).

The political system of the great seventeenth-century Brazilian Maroon kingdom of Palmares, for example, which R.K. Kent has characterized as an 'African' state, "did not derive from a particular Central African model, but from several" (in Price, 1996:188). In the development of the kinship system of the Ndyuka Maroons of Suriname, writes André Köbben, "undoubtedly their West-African heritage played a part … [and] the influence of the matrilineal Akan tribes is unmistakable, but so is that of patrilineal tribes … [and there are] significant differences between the Akan and Ndyuka

matrilineal systems" (in Price, 1996:324). Historical and anthropological research has revealed that the magnificent woodcarving of the Suriname Maroons, long considered 'an African art in the Americas' on the basis of formal resemblances, is in fact a fundamentally new, Afro-American art 'for which it would be pointless to seek the origin through direct transmission of any particular African style' (in Price, 1996:29). Detailed investigations – both in museums and in the field – of a range of cultural phenomena among the Saamaka Maroons of Suriname have confirmed the dynamic, creative processes that continue to animate these societies (Price and Price, 1999; Price, 2008).

Maroon cultures do possess direct and sometimes spectacular continuities from particular African peoples, from military techniques for defense to recipes for warding off sorcery. These are, however, of the same type as those that can be found, albeit with lesser frequency, in Afro-American communities throughout the hemisphere. In stressing these isolated African 'retentions,' there is a danger of neglecting cultural continuities of a more significant kind. Roger Bastide divided Afro-American religions into those he considered 'preserved' or 'canned' – like Brazilian *Candomblé* – and those that he considered 'alive' – like Haitian *Vaudou*. The former, he argued, manifest a kind of 'defense mechanism' or 'cultural fossilization,' a fear that any small change may bring on the end, while the latter are more secure of their future and freer to adapt to the changing needs of their adherents (in Price, 1996:29). Indeed, tenacious fidelity to 'African' forms seems, in many cases, to indicate a culture finally having lost meaningful touch with the vital African past. Certainly, one of the most striking features of West and Central African cultural systems is their internal dynamism, their ability to grow and change. The cultural uniqueness of the more developed Maroon societies (e.g., those in Suriname) rests firmly on their fidelity to 'African' cultural principles at these deeper levels – whether aesthetic, political, religious, or domestic – rather than on the frequency of their isolated 'retentions.' With a rare freedom to extrapolate ideas from a variety of African societies and adapt them to changing circumstances, Maroon groups included (and continue to include today) what are in many respects at once the most meaningfully African and the most truly 'alive' and culturally dynamic of all Afro-American cultures (Price, 2008).

Six Cases

The most famous Maroon societies are Palmares in Brazil, Palenque de San Basilio in Colombia, the communities in Esmeraldas, Ecuador, the Mexican Maroons of San Lorenzo de los Negros, the Maroons of Jamaica, and the Saamaka and Ndyuka Maroons of Suriname. (For further readings on these communities, see Price, 1996.)

Palmares, in northeastern Brazil, flourished for a full century before its final defeat by a large Portuguese army in 1694–95. A close-knit federation of villages made up of some 11 000 people, largely Angolan-born Africans; it also included Brazilian Indians, poor whites, and Portuguese deserters. During the seventeenth century, wave after wave of Dutch and then Portuguese armies was sent out against Palmares, which resisted with the help of its complex political and social organization, under the leadership of Ganga Zumba and, during its final years, the legendary Zumbi. Knowledge of the internal affairs of Palmares remains limited, based as it is on soldiers' reports, the testimony of a captive under torture, official documents, archaeological work, and the like. But as a modern symbol of black (and anticolonial) heroism, Palmares continues to evoke strong emotions in Brazil.

Palenque de San Basilio boasts a history stretching back to the seventeenth century. In recent years, historians, anthropologists, and linguists – working in collaboration with Palenqueros – have uncovered a great deal about continuities and changes in the life of these early Colombian freedom fighters, whose proud descendants still occupy the town today.

The Maroon history of Esmeraldas, on the Pacific coast of Ecuador, and neighboring Barbacoas in Colombia, began in the early sixteenth century, when Spanish ships carrying African slaves from Panama to Guayaquil and Lima were wrecked amidst strong currents. A number of slave survivors sought freedom in the unconquered interior, where they allied with indigenous peoples. In the 1580s, having beaten back military expeditions sent to capture them, several Maroon leaders traveled to Quito to make peace with the Spanish. In 1599, a famous portrait of one such leader, Don Francisco de Arobe, and his two sons, was commissioned and sent to Philip III of Spain to commemorate these negotiations. Later, these Maroons were joined by rebel slaves from highland *haciendas* and lowland mines and formed a number of free towns, even as slavery continued nearby.

San Lorenzo de los Negros, in Veracruz on the Caribbean coast of Mexico, is probably the best known of the seventeenth-century Maroon towns in Mexico. Under their leader Yanga, these Maroons attempted to make peace as early as 1608, but it was not till 1630, after years of intermittent warfare, that the Viceroy and the crown finally agreed to establish the town of free Maroons.

The Jamaica Maroons, who continue to live in two main groups centered in Accompong (in the hills above Montego Bay) and in Moore Town (deep in the Blue Mountains), maintain strong traditions about their days as freedom fighters, when the former group was led by Cudjoe and the latter by the redoubtable woman warrior Nanny (whose likeness now graces the Jamaican 500 dollar bill). Two centuries of scholarship, some of it written by Maroons themselves, offers diverse windows on the ways these men and women managed to build a vibrant culture within the confines of a relatively small island.

The Suriname Maroons now constitute the most fully documented case of how former slaves built new societies and cultures, under conditions of extreme deprivation, in the Americas and how they developed and maintained semi-independent societies that persist in the present. From their late seventeenth-century origins and the details of their wars and treaty making to their current struggles with multinational mining and timber companies, much is now known about these peoples' achievements, in large part because of the extensive recent collaboration by Saamaka and Ndyuka Maroons with anthropologists. Today, Suriname Maroons – who number some 122 000 people – live in the interior of the country, in and around the capital Paramaribo, and in neighboring French Guiana as well as in the Netherlands. The relevant bibliographic resources now number in the

thousands – useful points of entry are Price and Price (1999) and Thoden van Velzen and van Wetering (2004).

Anthropological Issues

Since Melville and Frances Herskovits's first foreign fieldwork, with Saamaka Maroons during the summers of 1928 and 1929, Maroons have moved to the center of anthropological debates, from the origins of Creole languages and the 'accuracy' of oral history to the nature of the African heritage in the Americas and the very definition of Afro-American anthropology.

The Saamaka fieldwork played a signal role in Melville Herskovits's career, exercising – in the words of his widow – "a profound influence on his thinking. … In the Guiana bush… he saw, as he often told his students, nearly all of sub-Saharan Africa represented, from what is now Mali to Loango and into the Congo – and the Loango chief who came to our base camp [in Saamaka territory] invoked both the Great God of the Akan of the Gold Coast, Nyankompon, and the Bantu Zambi" (Herskovits, 1966:vii–viii). During that Maroon research, and on the voyages to and from Suriname, Herskovits developed a comparative vision of Afro-American studies (and of the Caribbean as 'a social laboratory') that would fuel his own prodigious work, and that of countless students, for decades to come. The seeds of his interest in exploring the African provenience of New World Negroes, in establishing an historical-ethnographic baseline in Africa for the study of New World developments, and in analyzing processes of cultural change (such as 'acculturation,' 'reinterpretation,' and 'syncretism') were all sown during this early fieldwork.

The expansion and critique of Herskovits' work, especially the controversial essay by Sidney W. Mintz and Richard Price (1992) and the reactions to it, have kept Maroons at the center of debate among Afro-Americanist anthropologists and historians (Price, 2006). Indeed, David Scott argues that the Saamaka Maroons have by now become "a sort of anthropological metonym … providing the exemplary arena in which to argue out certain anthropological claims about a discursive domain called Afro-America" (1991:269).

Much of the recent anthropological research among Maroons has focused on historiography and historical consciousness – how Maroons themselves conceptualize and transmit knowledge about the past – and it has privileged the voices of individual Maroon historians. *First-Time* presented oral historical accounts by Saamakas of the early years of their society's formation, 300 years ago, including their rebellions from slavery and their battles against the colonists (Price, 2002 (orig. 1983)). *Alabi's World* supplemented these Saamaka accounts with archival materials from German missionaries and Dutch administrators who lived in Saamaka villages beginning right after the 1762 peace treaties with the colonists, directly juxtaposing ever more complex written and oral materials about the same events (Price, 1990). *True-Born Maroons* combined oral and archival materials, with an emphasis on the perspective of Maroons themselves, for Jamaica, uncovering a remarkable number of African practices as well as a great deal of early New World creolization, involving language, religious practices, and much else (Bilby, 2005). A recent history in Dutch of the first century of Ndyuka Maroons presents alternating chapters based on oral and archival materials, creating a rich historical picture of politics and social life in this Suriname Maroon society (Thoden van Velzen and Hoogbergen, 2011).

Taken together, these studies helped pioneer the complementary use of oral history and archival research in anthropology more generally. They have caused significant chinks in traditional historians' reluctance to take oral data as seriously as written materials. Not surprisingly, this type of work has occasionally raised the charge of 'verificationism' from post-modernist scholars (Scott, 1991) but the particular kinds of attention it shows to the politics of memory – the strategic use of archival materials to complement Maroon narratives – can be defended as a decisive and progressive political act, one intended in part to foster social justice, rather than being judged naive or politically retrograde (Price, 2006).

Recent anthropological research has also focused on the ongoing debate about 'creolization' and the extent to which (and the ways in which) Maroons and other African-Americans have drawn, on the one hand, upon their diverse African heritages and, on the other hand, upon their New World experiences and their own creativity in building their cultural worlds. Researchers have been asking how, in each New World colony, enslaved Africans, coming from a variety of nations and languages, became African-Americans. Scholars have found that to begin to explore these processes, across the many regions where Africans were landed as captive laborers, one must first ask: How 'ethnically' homogeneous (or heterogeneous) were the Africans who arrived in a particular New World locality – in other words, to what extent was there a clearly dominant group – and what were the cultural consequences? How quickly and in what ways did they and their African-American offspring begin thinking and acting as members of new communities? In what ways did the African arrivants choose – and to what extent were they able – to continue particular ways of thinking and doing things that came from the Old World? What did 'Africa' (or its subregions and peoples) mean at different times to African arrivants and their descendants? How did the various demographic profiles and social conditions of particular New World settings encourage or inhibit these cultural processes? In short, what is it that made what Haitian anthropologist Michel-Rolph Trouillot calls 'the miracle of creolization' possible, across the Americas, over and over again?

Maroon societies play a privileged role in such discussion, in part owing to their relative isolation and, in the case of Suriname, their persistence into the present. The most detailed ethnography of creolization to date (Price, 2008) analyzes extensive data from esoteric Maroon languages, as well as other ritual materials, to demonstrate the complexity of these issues from both an ethnohistorical perspective and that of Maroons themselves. Eric Hobsbawm, commenting on the significance of this kind of work in the more general context of the social sciences, notes that "Maroon societies raise fundamental questions. How do casual collections of fugitives of widely different origins, possessing nothing in common but the experience of transportation in slave ships and of plantation slavery, come to form structured communities? How, one might say more generally, are societies founded from scratch? What exactly did or could

such refugee communities … derive from the old continent?" (1990:46). Questions such as these are sure to keep students of Maroon societies engaged in active research for many years to come.

Since the turn of the twenty-first century, Maroons have taken center stage in the struggle for territory and sovereignty within nation-states, particularly in Brazil and in Suriname. Following promulgation of the new constitution of Brazil in 1988 (and subsequent reforms), the descendants of historical *quilombos* (which originally meant 'Maroon communities') have, under certain conditions, had the right to claim collective title to the lands on which their ancestors lived. In the process of the ensuing legal battles, the definition of *quilombo* has expanded to include almost any landless black Brazilian community – rural or urban – whose members can claim a history of resistance to state power – not just descendants of Maroons. Today, many young Brazilian anthropologists are trained explicitly to help such *quilombo*-descendant communities in their quest for title to lands and related rights.

In 1990s' Suriname, the Saamaka Maroons – some 55 000 people – found the lands that their ancestors had lived on since the late seventeenth century invaded by Chinese loggers and other multinational mining and extraction companies that had received permits from the national government. Suriname's constitution grants ownership of the forest, which is inhabited by a number of indigenous peoples as well as six Maroon peoples, to the State, which during that decade began doling out mining and logging concessions to foreign multinationals. Organizing themselves and obtaining legal aid, the Saamaka People fought a decade-long legal battle with the Republic of Suriname. This struggle culminated in a landmark 2007 judgment by the Inter-American Court for Human Rights (the case of *Saramaka People v Suriname*), requiring the nation to change its laws to grant the Saamaka people collective title to their traditional territory as well as considerable sovereignty over it – a judgment that establishes jurisprudence for indigenous peoples and Maroons throughout the Americas. Richard Price served as advisor and expert witness for the Saamaka during this case, providing an example of advocacy in anthropology. The case and its significance are discussed in detail in Price (2011).

See also: Advocacy in Anthropology; Anthropology and History; Colonialism, Anthropology of; Creolization: Sociocultural Aspects; Human Rights, Anthropology of; South America: Sociocultural Aspects.

Bibliography

Bilby, Kenneth M., 2005. True-Born Maroons. University Press of Florida, Gainesville, FL.

Herskovits, F.S., 1966. Introduction. In: Herskovits, F.S. (Ed.), The New World Negro. Indiana University Press, Bloomington, IN, pp. vii–xi.

Hobsbawm, E.J., 6 Dec 1990. Escaped Slaves of the Forest. New York Review of Books, pp. 46–48.

Mintz, S.W., Price, R., 1992. The Birth of African-American Culture: An Anthropological Approach. Beacon Press, Boston (Original work published 1973), MA.

Price, R., 1990. Alabi's World. Johns Hopkins University Press, Baltimore.

Price, R. (Ed.), 1996. Maroon Societies: Rebel Slave Communities in the Americas, third ed. Johns Hopkins University Press, Baltimore.

Price, R., 2002. First-Time: The Historical Vision of an African American People, second ed. University of Chicago Press, Chicago.

Price, R., 2006. On the miracle of creolization. In: Yelvington, K.A. (Ed.), Afro-Atlantic Dialogues: Anthropology in the Diaspora. SAR Press, Santa Fe, pp. 113–145.

Price, R., 2008. Travels with Tooy: History, Memory, and the African American Imagination. University of Chicago Press, Chicago.

Price, R., 2011. Rainforest Warriors: Human Rights on Trial. University of Pennsylvania Press, Philadelphia.

Price, S., Price, R., 1999. Maroon Arts: Cultural Vitality in the African Diaspora. Beacon Press, Boston, MA.

Richard Price, www.richandsally.net.

Scott, D., 1991. That event, this memory: notes on the anthropology of African diasporas in the New World. Diaspora 1 (3), 261–284.

Thoden van Velzen, H.U.E., Hoogbergen, W., 2011. Een zwarte vrijstaat in Suriname: De Okaanse samenleving in de achttiende eeuw. KITLV Press, Leiden.

Thoden van Velzen, H.U.E., van Wetering, W., 2004. In the Shadow of the Oracle: Religion as Politics in a Suriname Maroon Society. Waveland, Long Grove.

Marr, David (1945–80)

Shimon Edelman, Cornell University, Ithaca, NY, USA
Lucia M Vaina, Harvard Medical School, Boston University, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

David Courtnay Marr was born in 1945 in Essex, England. Marr's dissertation, written at Trinity College, Cambridge and published between 1969 and 1971, presented a theory of mammalian brain function, parts of which remain relevant to the present day, despite vast advances in neurobiology in the past decades. In 1973, Marr joined the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology, where he was made a tenured full professor in 1980. Marr died in November 1980, of leukemia. His highly influential book, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, which has redefined and revitalized the study of human and machine vision, was published posthumously, in 1982, with a new edition appearing in 2010.

David Courtnay Marr was born on 19 January 1945 in Essex, England. He attended Rugby, the English public school, on a scholarship and went on to Trinity College, Cambridge. By 1966, he obtained his BS and MS in mathematics and proceeded to work on his doctorate in theoretical neuroscience, under the supervision of Giles Brindley. Having studied the literature for a year, Marr commenced writing his dissertation. The results, published in the form of three journal papers between 1969 and 1971, amounted to a theory of mammalian brain function, parts of which remain relevant to the present day, despite vast advances in neurobiology in the past four decades. Marr's theory was formulated in rigorous terms, yet was sufficiently concrete to be examined in view of the then available anatomical and physiological data. Between 1971 and 1972, Marr's attention shifted from general theory of the brain to the study of vision. In 1973, he joined the Artificial Intelligence Laboratory (AI Lab) at the Massachusetts Institute of Technology (MIT) as a visiting scientist, taking on a faculty appointment in the Department of Psychology in 1977, where he was made a tenured full professor in 1980. In the winter of 1978, he was diagnosed with acute leukemia. David Marr died on 17 November 1980, in Cambridge, Massachusetts. His highly influential book, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, which has redefined and revitalized the study of human and machine vision, was published posthumously, in 1982. A new edition (MIT Press, 2010) includes an introduction by Shimon Ullman and a postface by Tomaso Poggio, two of Marr's closest collaborators.

A Theory of the Brain

Marr's initial work in neuroscience combined high-level theoretical speculation with meticulous synthesis of the anatomical data available at the time. The question he chose to address is the *nec plus ultra* of neuroscience: what is it that the brain does? Marr proposed a definite answer to this question for three major brain structures: archicortex (the phylogenetically older part of the cerebral cortex), cerebellum, and neocortex. The three answers complement each other, rallying around the idea that the brain's central function is statistical pattern recognition and association, carried out in a very high-dimensional space of 'elemental' features. The basic building block of all three theories is a *codon*, or a subset of features, with which there may be associated a cell, wired so as to fire in the presence of that particular codon.

In the first paper, Marr proposed that the cerebellum's task is to learn the motor skills involved in performing actions and maintaining posture (Marr, 1969). The Purkinje cells in the cerebellar cortex, presumably implementing the codon representation, associate (through synaptic modification) a particular action with the context in which it is performed. Subsequently, the context alone causes the Purkinje cell to fire, which in turn precipitates the next elemental movement. Although the cerebellum is now seen as only a part of the brain-action learning and control system (which involves, notably, a number of functional loops linking it with the cortex, the hippocampus, the basal ganglia, and the thalamus), Marr's ideas regarding its computational role are still considered essentially correct.

The next paper (Marr, 1970) extended the codon theory to encompass a more general kind of statistical concept learning, which he assessed as "capable of serving many of the aspects of the brain's function" (the vagueness of this aspect of the theory would lead him soon to abandon this approach, which, as he realized all along, was "once removed from the description of any task the cerebrum might perform"). How can a mere handful of techniques for organizing information (such as the codon representation) support a general theory of the brain function? Marr's views in this matter were profoundly realist, and were based on a postulate of "the prevalence in the world of a particular kind of statistical redundancy, which is characterized by a 'Fundamental Hypothesis,'" stating that "Where instances of a particular collection of intrinsic properties (i.e., properties already, diagnosed from sensory information) tend to be grouped such that if some are present, most are, then other useful properties are likely to exist, which generalize over such instances. Further, properties often are grouped in this way" (Marr, 1970, pp. 150–151). This worldview and the ideas that it engendered presaged much of the later work by others on neural network models of brain function, which invoke the intuition of learning as optimization ('mountain climbing') in an underlying probabilistic representation space.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61085-1

A model at whose core is the tallying of probabilities of events needs an extensive memory of a special kind, allowing retrieval based on the content, rather than the storage location, of the items. Marr's third theoretical paper considered the hippocampus as a candidate for fulfilling this function (Marr, 1971). In analyzing the memory capacity and the recall characteristics of the hippocampus, Marr integrated abstract mathematical (combinatorial) constraints on the representational capabilities of codons with concrete data derived from the latest anatomical and electrophysiological studies. The paper postulated the involvement in learning of synaptic connections modifiable by experience – a notion originating with the work of Donald Hebb (*see* Hebb, Donald Olding (1904–1985)) in the late-1940s and discussed by Marr's mentor Brindley in a 1969 paper. Marr provided a mathematical proof of efficient partial content-based recall by his model and offered a functional interpretation of many anatomical structures in the hippocampus, along with concrete testable predictions. Many of these (such as the existence in the hippocampus of experience-modifiable synapses) were subsequently corroborated (see the reviews in Vaina, 1990).

The MIT Period

A consummation of this three-prolonged effort to develop an integrated mathematical–neurobiological understanding of the brain would in any case have earned Marr a prominent place in a gallery, spanning two and a half centuries (from John Locke (*see* Locke, John (1632–1704)) to Kenneth Craik) of British Empiricism, the epistemological stance perennially popular with neuroscientists. As it were, having abandoned the high-theory road soon after the publication of the hippocampus paper, Marr went on to make his major contribution to the understanding of the brain by essentially inventing a field and a mode of inquiry: *computational neuroscience*. By 1972, the focus of his thinking in theoretical neurobiology shifted away from abstract theories of entire brain systems, following a realization that without an understanding of specific tasks and mechanisms – the issues from which his earlier theories were 'once removed' – any general theory would be glaringly incomplete.

Marr first expressed these views in public at an informal workshop on brain theory, organized in 1972 at the Boston University by Benjamin Kaminer. In his opening remarks, he suggested an 'inverse square law' for theoretical research, according to which the value of a study varies inversely with the square of its generality – an assessment that favors top-down reasoning anchored in functional (computational) understanding, along with bottom-up work grounded in an understanding of the mechanism, but not theories derived from intuition, or models built on second-hand data.

The new methodological stance developed by Marr following the shift in his views is summarized in a remarkably lucid and concise form in a two-page book review in *Science*, titled "Approaches to Biological Information Processing" (Marr, 1975). By that time, Marr came to believe firmly that the field of biological information processing had not yet accrued an empirical basis sufficient for guiding and supporting a principled search for a general theory. Remarking that the brain may turn out to admit "of no general theories except ones so unspecific as to have only descriptive and not predictive powers" – a concern echoed in one of his last papers (Marr, 1981) – he proceeded to mount a formidable critique of the most common of the theories circulated in the early 1970s, such as catastrophe theory and neural nets (the current popularity of dynamical systems approaches and of connectionism, taken along with the integration of Marr's critical views into the mainstream theoretical neurobiology, should fascinate any student of the history of ideas).

The main grounds for his argument, which was further shaped by an intensive and fruitful interaction with Tomaso Poggio (Marr and Poggio, 1977), were provided by an observation that subsequently grew into a central legacy of Marr's career: the understanding of any information processing system is incomplete without insight into the problems it faces, and without a notion of the form that possible solutions to these problems can take. Marr and Poggio termed these two levels of understanding *computational* and *algorithmic*, placing them above the third, *implementational*, level, which, in the study of the brain, refers to the neuroanatomy and neurophysiology of the mechanisms of perception, cognition, and action.

Upon joining the MIT AI Lab, Marr embarked on a vigorous research program seeking computational insights into the working of the visual system, and putting them to the test of implementation as computer models. Marr's thinking in the transitional stage, at which he treated computational results on par with neurobiological findings, is exemplified by the paper on the estimation of lightness in the primate retina (Marr, 1974); subsequently, much more weight was given in his work to top-down, computational-theory considerations. This last period in Marr's work is epitomized by the theory of binocular stereopsis, developed in collaboration with Poggio, and presented in a series of ground-breaking papers (Marr and Poggio, 1976, 1979). At that time, Marr also worked on low-level image representation (Marr, 1976; Marr and Hildreth, 1980), and on shape and action categorization (Marr and Nishihara, 1978; Marr and Vaina, 1982). Marr's book, *Vision*, written during the last months of his life, is as much a summary of the views of what came to be known as the MIT school of computational neuroscience, as it is a personal credo and a list of achievements of the second part of Marr's scientific endeavor, which lasted from about 1972 to 1980.

Legacy

The blend of insight, mathematical rigor, and deep knowledge of neurobiology that characterizes Marr's work is reminiscent of the style of such titans of neuroscience as Warren McCulloch – except that McCulloch's most lasting results were produced in collaboration with a mathematician (Walter Pitts), whereas Marr did his own mathematics. A decade after his quest was cut short, it has been claimed both that Marr is cited more than he is understood (Willshaw and

Buckingham, 1990), and that his influence permeates theoretical neurobiology more than what one would guess from counting citations (McNaughton, 1990). Still, contributors to the mainstream journals in neurobiology now routinely refer to the 'computations' carried out by the brain, and the most exciting developments are those prompted or accompanied by computational theories and models.

In computer vision (a branch of artificial intelligence), the influence of Marr's ideas has been complicated by the dominance of the top-down interpretation of his methodology: proceeding from a notion of what needs to be done toward the possible solutions. For some time, Marr's school was identified with the adherents of a particular computational theory of vision, which claims that constructing an internal model of the world is a prerequisite for carrying out any visual task. The accumulation of findings to the contrary in neurobiology and in the behavioral sciences gradually brought to the fore the possibility that vision does not require geometric reconstruction. This encouraged researchers to seek alternative theories, some of which employ concepts and techniques that did not exist in the 1970s, or were not known to the scholars of vision at the time. These new ideas, in turn, are making their way into neuroscience, as envisaged by Marr. Interestingly, many such ideas complete a conceptual circle by reconnecting with Marr's own earlier work represented by the trio of papers summarizing his dissertation. Thus, the machine-learning approaches that have become indispensable in modeling cognition can be seen as related to the probabilistic methods invoked by Marr in connection with codon-based statistical learning and inference that he developed.

On a more general level, Marr's work provided a solid proof that a good theory in behavior and brain sciences need not have to trade-off mathematical rigor for faithfulness to specific findings. More importantly, it emphasized the role of explanation over and above mere curve-fitting, making it legitimate to ask *why* a particular brain process is taking place, and not merely what differential equation can describe it or what the neural circuit implementing it looks like.

See also: Cognitive Neuroscience; Computational Neuroscience; Concept Learning and Representation: Models; Information Processing Architectures: Fundamental Issues; Neural Plasticity; Neural Representations of Objects; Visual Perception, Neural Basis of.

Bibliography

Marr, D., 1969. A theory of cerebellar cortex. Journal of Physiology London 202, 437–470.
Marr, D., 1970. A theory for cerebral neocortex. Proceedings of the Royal Society of London B 176, 161–234.
Marr, D., 1971. Simple memory: a theory for archicortex. Philosophical Transactions of the Royal Society of London B 262, 23–81.
Marr, D., 1974. The computation of lightness by the primate retina. Vision Research 14, 1377–1388.
Marr, D., 1975. Approaches to biological information processing. Science 190, 875–876.
Marr, D., 1976. Early processing of visual information. Philosophical Transactions of the Royal Society of London B 275, 483–524.
Marr, D., 1981. Artificial intelligence a personal view. In: Haugeland, J. (Ed.), Mind Design. MIT Press, Cambridge, MA, pp. 129–142. Chap. 4.
Marr, D., 1982. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. W.H. Freeman, San Francisco, CA.
Marr, D., Hildreth, E., 1980. Theory of edge detection. Proceedings of the Royal Society of London B 207, 187–217.
Marr, D., Nishihara, H.K., 1978. Representation and recognition of the spatial organization of three dimensional structure. Proceedings of the Royal Society of London B 200, 269–294.
Marr, D., Poggio, T., 1976. Cooperative computation of stereo disparity. Science 194, 283–287.
Marr, D., Poggio, T., 1977. From understanding computation to understanding neural circuitry. Neurosciences Research Program Bulletin 15, 470–491.
Marr, D., Poggio, T., 1979. A computational theory of human stereo vision. Proceedings of the Royal Society of London B 204, 301–328.
Marr, D., Vaina, L.M., 1982. Representation and recognition of the movements of shapes. Proceedings of the Royal Society of London B 214, 501–524.
McNaughton, B.L., 1990. Commentary on simple memory: a theory of the archicortex. In: Vaina, L.M. (Ed.), From the Retina to the Neocortex: Selected Papers of David Marr. Birkhauser, Boston, MA, pp. 121–128.
Vaina, L.M. (Ed.), 1990. From the Retina to the Neocortex: Selected Papers of David Marr. Birkhauser, Boston, MA.
Willshaw, D.J., Buckingham, J.T., 1990. An assessment of Marr's theory of the hippocampus as a temporary memory store. Philosophical Transactions of the Royal Society of London B 329, 205–210.

Marriage

Perveez Mody, King's College, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J. Bestard-Camps, volume 14, pp. 9258–9261, © 2001, Elsevier Ltd.

Abstract

Marriage as a central framing device within anthropological theory has proved illusory because of the lack of a universal definitional basis on which to peg its significance. This article seeks to explore how marriage, as an anthropological object, has transformed through time alongside important changes within anthropology and especially with regards the anthropological study of kinship. It argues that marriage cannot be plausibly isolated from other social institutions and explores it as a web of meaning and effects that are open to anthropological interpretation.

Since the late nineteenth century, marriage in its ritual and regulative forms has come to be seen as a crucial aspect of the subject that anthropologists regard as most traditionally close to their hearts: the study of kinship (Morgan, 1997[1870]). Marriage has certainly been recognized as central in understanding bonds of kinship, for exploring the nature of these bonds and their effects upon the persons participating within relations effected by marriage and for expressing exchange, intimacy, sociality, and care. However, marriage as a central framing device within anthropological theory has proved illusory because of the lack of a universal definitional basis on which to peg its significance (though, from a postmodern perspective, we can probably reassure ourselves that this need not be desirable nor indeed, may it serve as the deformity it was often made out to be). Rivière, for instance, has argued successfully that 'marriage' cannot be treated as an isolable phenomena (1971), and Strathern has extended this caution in an ethnographically detailed account that serves to challenge what the subject matter of marriage, and marriage exchanges might be (1984). This historic complexity notwithstanding, globally and popularly, marriage continues to be thought of as a socially approved union that links sexual access to rules concerning the care, legitimacy, and social affiliation of children. Underpinning this popular (nonanthropological) consensus about what 'marriage' is about, is the noticeable presence in contemporary Western states of online and actual world 'marriage publics' who rally around particular conceptions of marriage as natural/cultural; or a 'universal human right' or defined as the 'right to found a family' (Borneman, 1996: p. 216). Marriage, as an anthropological object, thus sits at an interesting confluence of kinship studies (including those of gender, intimacy, commodification, and migration), politics and religion, and the study of legal anthropology that seeks to critique and understand the meanings, contexts, exchanges, and relationships that attach and make manifest 'rights.' Despite the approach of this article that explores the place of marriage within kinship theory, one could certainly argue with Leach that marriage cannot be plausibly separated out from other social institutions (1961) but that it forms an inextricable web of meaning and effects that are open to anthropological interpretation.

Definition of Marriage

The 1951 Edition of the Royal Anthropological Institute's 'Notes and Queries' defined marriage in the following way: "Marriage is a union between a man and a woman such that children born to the woman are recognized legitimate offspring of both parents" (1951: p. 110). Subsequent attempts along these lines to delineate a cross-culturally valid definition of the rights and responsibilities that constitute marriage (for example, defined as conferring rights to sexual access of women) have similarly come to grief (see Comments section following Bell, 1997). As we can see, older anthropological texts were not chary of declaring marriage to be about heterosexual unions (monogamous or polygamous), bearing offspring to swell the family and lineage. Ten years later, the British anthropologist Edmund Leach took a much more capacious view of marriage defining it as 'a bundle of rights,' which might include 'several distinguishable subtypes of institution' such as those that establish legal parenthood (fatherhood/motherhood), monopolize the spouse's sexuality, generate rights in the spouse's labour and property, establish a fund to care for offspring to the marriage, and/or establish a socially significant relation of affinity and 'possibly many more similar rights' (1961). The crucial aspect here is that no single set of rights or roles were regarded as central to a universal definition of marriage, because Leach did not consider such a universal definition of marriage to be viable. Leach's definition hinged upon legality and legitimacy, but nevertheless chimed with those like Needham who urged that 'marriage' as an analytic device be treated as an 'odd-job word' more likely to mislead rather than yield cross-cultural comparison, and worse still, of 'no use at all in analysis' (1971: pp. 7–8). In some ways, anthropological definitions of marriage were either driven by its ritual significance (e.g., 'marriage by capture,' 'group marriage,' 'polygynous marriage,' see for instance McLennan, 2012[1865]) or its regulative importance (marriage as a rite conferring legal rights) with little consideration of how people actually experienced the processes of marrying and making kin.

In recent years, feminist anthropologists and queer theorists have critiqued anthropological studies of marriage (most obviously, Levi-Strauss' ambitious volume *The Elementary*

Structures of Kinship) by reminding us that it is by no means obvious, natural, or universal that marriage involves a heterosexual couple, the exchange of wives by groups of men, nor indeed should we assume marriage to be the principle regulative force or loci of kinship solidarity and care (Rubin, 1975; Borneman, 2001). However, the critique of the cross-cultural variability through which marriage manifests itself was already widely in evidence. Evans-Pritchard had argued that the institution of 'woman marriage' showed that fatherhood among the Nuer was a status conferred to the person in whose name cattle bridewealth was given for the mother; so in this sense, a woman could be both a husband and the father to her wife's child if cattle were given in her name (1951: p. 107). Evans-Pritchard's account of boy-marriage among Azande warriors and *bagburu* lesbian 'friendships' among Zande women in polygamous marriages demonstrated the institutional marital forms of 'libidinous plasticity' that dealt with situations of shortage of heterosexual partners through same-sex relations that were crucially, he argued, *called* marriage (1970). The present attempts to legitimate and legislate for same-sex marriage in many parts of Euro-America, seek to institutionalize nonheterosexual marriages through legal means. Indeed, as Kath Weston has argued in her ground-breaking ethnography of gay and lesbian kinship in California (1997), the conflation of kinship with ties of biology and the consanguineous family has, in recent decades, depleted our anthropological imagination of the multitudinous ways in which kinship itself is made and *not* given (Weston, 1991), even if anthropologists have observed that this (re)making is surprisingly often along normative lines (Bonaccorso, 2009).

Marriage and Kinship

Over half a century ago, marriage and reproduction through parenthood (and sibblingship) were seen as centrally constitutive of kinship. The 'elementary family' of Radcliffe-Brown (1952), imagined to be constituted by two parents and their children were seen to constitute the proper subject of study that we call kinship, though as Bohannan subsequently argued, the matricentric family (constituted of a woman and her children) was far more "universal and more elementary than the nuclear family" (Bohannan, 1963: p. 73).

Marriage, in this schema was designated by the kinship outcome of 'affinity' with affines being the persons to whom one was related through marriage. We should note here an interesting point raised by the French anthropologist Dumont in his appraisal of the role of marriage in two competing anthropological theories of kinship: that of descent (whose proponents were of the British structural-functionalist school) and alliance (whose proponent was the French Structuralist Levi-Strauss). When we refer in English to the kinship ties of blood and marriage, we use the words 'relations of kinship *and* affinity' because the term kinship is broadly understood to be based on consanguineous ties. By contrast, in French the word *parenté* encompasses *both* consanguinity (relations of shared blood), and affinity (relations by alliance or marriage). Consequently, to demonstrate the importance of marriage to a study of kinship, anthropologists writing in English were self-consciously using the words kinship and affinity (to designate ties of blood and marriage) whereas the French *parenté* sufficed for both. Dumont argued from all this that the peripheral place of marriage in British Structural-functionalists descent theory, and its centrality to alliance theory among French Structuralists, may in fact be related to the linguistic peculiarity that requires an elaboration to the term 'kinship' to be paired with 'affinity' in English and that encompasses ties of blood and affinity through the word *parenté* in French (2006 [1971]: p. 3). In descent theory, descent groups import women to reproduce the group (through reproduction). The function of marriage is to perpetuate the lineage system and marriages between groups preclude further marriages for some generations. In alliance theory, the exogamic groups were ordered to organize marriage and one marriage generates a systemic alliance between groups – the marriage tie being a structural element in the system of alliance and reciprocity on a wider social scale. 'Marriage alliance' refers to the repetition of intermarriage between larger or smaller groups, typically descent groups. The concept of marriage alliance was developed by Levi-Strauss to deal with societies that practice various forms of cross-cousin marriages. He argued that such repeated exchange of women (by men) through marriage created ties of mutual interdependence among groups; marriage was a quintessential system of exchange.

The reciprocity and exchange that drives alliance theory were argued by feminist anthropologists to be "reckoned in female flesh" (Rubin, 1975: p. 182). Gayle Rubin pointed out that both British Structural-functionalism and French Structuralism were united in their uncritical view of the rules of gender division between male and female, and the 'obligatory heterosexuality' that ensued from both theories of marriage and consequently, that came to dominate the study of kinship (1975: p. 182). Anthropologists have gone to great lengths to identify the ways in which marital choice is constrained (through the veritable universality of some form or the other of the incest taboo as well as through a range of kinship practices that enjoin marriage in one direction and restrain it in others). It is also clear that in many societies, a high degree of constraint in terms of marital choice may frequently coexist with a relatively acquiescent attitude to premarital and extramarital sex (see Evans-Pritchard, 1951; Gell, 1996; Boellstorff, 2005). In this vein, while lauding the work of feminist anthropologists in their critique of biological/genealogical determinism in kinship studies (the idea that "kinship includes large universes of people whose relatedness may or may not have anything to do with notions of blood"), Bodenhorn argues that the study of marriage and kinship needs to be much more attentive to the 'instability' of categories such as race (2013). Given its alliance potential, marriage has always served as exemplary in its 'transboundary capacities,' and Bodenhorn draws attention to the ways in which the modern invention of categories such as 'race' and 'ethnicity' make 'intermarriage' a marked (rather than an unmarked) relationship (2013: pp. 132–133). This marking has a peculiarly 'muting effect' on the many crossings that have always served as a widespread basis for marriage and making kinship.

Marital Gifts

Anthropologists have frequently regarded marriage in so-called 'traditional' societies as an exchange between two communities rather than between two individuals, and consequently it is a public social event that has the potential to alter the relationship, property, and status between those two groups. From a structuralist perspective, marriage payments 'are conceived as elements within total exchange systems' and they are not considered interesting independent objects of investigation (Comaroff, 1980: p. 30). So while the act of a marriage is expressed in a ritual or a ceremony, it invariably also transfers 'goods, rights, and people'; with the end of marriages marked by the relocation of these (Simpson, 1994: p. 832). Two of the 'classic' forms of presentation given at marriage are called bridewealth and dowry and the difference between the two are characterized by the direction taken by the flow of gifts. Dowry is the name given to property passing in principle from the bride's parents to the bride, though it is often the case that the dowry goes directly to the bride's in-laws. Bridewealth on the other hand is property that goes from the groom or the kin of the groom to the kin of the bride. Among the Nuer, for instance, the exchange of bridewealth from groom's kin to bride's kin in the form of cattle serves to secure paternity for any children born to the marriage of the man in whose name the cattle had been given. They are, as Evans-Pritchard argued, "the children of the cattle" (1951: p. 98). The anthropological commonplace that bridewealth flows in 'the opposite direction of the bride,' while dowry moves with her conceals a number of interesting assumptions about the nature of marriage, not least that women move in true Levi-Straussian fashion between male kin and that their movements can be juxtaposed against the movement of marital gifts between these kin. Gift giving at marriage is a vitally important kinship-based economic activity and in the past, in China for instance, assembly of the wealth necessary for generating bridewealth could tie up the economic activities of household members for decades at a time (Yan, 2005). Osella and Osella provide a detailed exploration of a shift among Izhavas in Kerala from low-status 'poor people's' cross-cousin alliances toward hypergamous unions (preferably to Gulf-employed grooms) that are set apart by high prestige payments of large dowries that transform dowry into "groom price" (2000: pp. 81–116). Significantly, among Mawri men in Dogondoutchi, a Muslim Hausa-speaking town in Niger, Africa, 'nonmarriage [was] not an option' and young men felt an enormous pressure to demonstrate their marital eligibility through assembling suitable amounts of bridewealth or *sadaki* in order to find a wife (Masquelier, 2005: p. 59). Men unable to acquire suitable *sadaki* were accused of being idle (literally, staying in the room or *daki*) and unable to 'forge a productive masculinity through proper marriage.' Masquelier argues that in Dogondoutchi, "[t]o win a young girl's heart, a boy had to ply her with costly fripperies" (1995: p. 74). Marital gift giving can also be seen as a "serial form of family division" (Yan, 2005) wherein economic wealth is distributed successively depending on the gender of the child who has married and the prevailing traditions of gift giving. However, anthropologists have been slow to investigate the full complexity and transformation of marital gifts as they often change with new economic circumstances and constraints, coalesce or get distributed throughout the life of the couple. Notably, Yan describes such a process in Xiajia, a village in rural China, where he witnessed a process in which three separate categories of marital gifts were converted into *ganzhe* (cash) and eventually into *daganzhe* ('grand conversion'), a unified category of bridewealth that was demanded directly by the bride. Marital gifts in terms of land or houses can therefore serve as a significant productive base for a couple and can be a valued 'conjugal fund' that they may draw upon over the course of their lives together.

Marriage, Roles, and Change

An increasing interest in intimacy and the nature of marital relationships has drawn attention to the ways in which earlier anthropological theories viewed consent in marriage. This was broadly viewed as a correlate of social forms so that high degrees of parental control over spouse selection indicated that the society as a whole valued lineal solidarity (with marriage the key functional mechanism for ensuring this). As Winch argues, in societies that value the extended family system: "Marriage will … be considered as the focus of the process of family formation, and mate selection as one of its most problematic features" (1968: p. 1). The evolutionist argument proceeds to assume that with changing family forms and greater nuclearization and gender equality, family members allow the couple greater choice and, left to generate their own marital criteria, the young and the nubile seek out partners with whom they share an attraction, romance, or love (Winch, 1968; Giddens, 1992). However, Robbins provides a fascinating counterpoint with ethnographic contextualization of the place of the 'nakedly spoken will' ('calling the name') for women instigating their own marriages among the Urapmin of Papua New Guinea (2004).

A rather different, and pioneering thesis on the roles of men and women *within* marriages came from Elizabeth Bott's network analysis of 20 working-class London families, focusing on the extent to which husbands and wives segregated their lives and roles from each other. Bott showed that differences in terms of roles (i.e., 'reciprocally expected behavior') between heterosexual couples depended strongly on "patterns of social relationships with and among persons *outside* the family" (1957: p. 219). The thesis of the study was that the social world to which an urban family was related was often quite dispersed. This was in contrast to close-knit communities where a family interacts with individuals who are themselves in close contact with each other and this generates shared norms and expectations about marital roles. Consequently, the more dispersed the social network of a family, the more impaired the couple's ability to look outside the marriage for help and support, thus reducing the exercise of external normative pressures upon the couple who are forced to be more self-reliant. Bott argued, "… if husband and wife come to marriage with loose-knit networks, or if their networks become loose-knit after marriage, they must seek in each other some of the emotional satisfaction and help with familial tasks that couples in close-knit networks can get from outsiders. Joint organisation becomes more necessary for the success of the

family as an enterprise" (1957: p. 218). Bott's network analysis showed that conjugal roles were less marked when conjugal interdependence was an emotional and social necessity. While this work never did become central to the anthropology of marriage, it was pioneering in network analysis and demonstrates the potential for innovation in thinking about the role of marriage within changing social structures.

In the same measure by which studies of marriage in the mid- to late-twentieth century came under fire for placing an undue emphasis on the structural rules and constraints of marriage viewed from the perspectives of corporate groups, more recent anthropological work has sought to explore the cultural contexts in which parental marital arrangement, self-choosing, and constraint express themselves through processes of marriage making (de Munck, 1996; Ahearn, 2001; Hart, 2007; Mody, 2008; Fuller and Narasimhan, 2008). This emphasis on the processual and agentive aspects of marriage mark the emergence of new possibilities for anthropological studies, promising to restore marriage to the forefront of kinship theory. The emphasis on the agency and choice of individuals (acting as such) in marriage is also stressed by the Sociologist Giddens in his work *The Transformations of Intimacy* (1992). Giddens characterizes shifts in conjugal intimacy as an outcome of a veritable global movement in which the so-called 'democratization of society' is accompanied by a concomitant stress on the equality between men and women, of the need for personal fulfillment in marriage and on the increasing recognition that when these needs are no longer fulfilled, then the marriages dissolve and end in divorce.

Marriage and the Creative Potential of Divorce

A more nuanced and historical view of marriage comes from Pat Thane who argued that premarital intimacy has been a feature of the courtship practices of large swathes of the population since at least the mid-eighteenth century in Britain, and that a similar picture emerges on high rates of nonmarriage among men and women bringing up children in the same time period; Thane argues that the period 1945–70 is exceptional in its regard for marriage and conjugal togetherness rather than the norm (Thane, 2010). In the context of discourses on freedom and enterprise in Euro-American life, Strathern has argued that the enterprise culture heralds 'enterprising kinship' in which the possibilities of expanded personal choice (not least in marriage) play a significant role in transforming intimate, kinship, and caring relationships (Strathern, 1992). Indeed, the anthropologist Bob Simpson has argued that concomitant effects of the ubiquity of divorce and remarriage in contemporary (Euro-American) society is a 'distinctive weft which runs through the fabric of society' connecting households, children, and property in new and complex kinship relations (1994: p. 832). Decoupling in the form of divorce (the legal mechanism sanctioning separation), is accompanied by new couplings and relationships legitimated in marriage, and Simpson argues that the resultant 'unclear family' (as juxtaposed to the nuclear family) allows us to refocus our anthropological gaze upon the creative possibilities of divorce to restructure and reorder interpersonal kinship and economic relations.

The anthropologist Jonathan Parry takes to task the Giddensian theory that an increasing stress on equality and conjugal intimacy *necessarily* leads to higher divorce rates by drawing on his detailed ethnographic study of low-caste Satnamis (descendents of Chamar untouchable converts) in Chhattisgarh (central India). Following Leach, Parry is careful to distinguish between jural and conjugal stability; he argues that people can remain married to each other even if they haven't seen each other for years. Parry argues that while the general image of caste society is one of jural stability in terms of very low divorce rates, customary forms of divorce and remarriage are common at the lower end of the caste hierarchy and among *adivasi* or tribal groups in India. For instance, among the Girasias of Rajasthan, 50% of household heads had been previously married at least once (Unnithan-Kumar, 1997). Questioning the unremitting stereotype about the stability of Indian arranged marriages with his own ethnographic material, Parry argues that "Certain pockets of the country sustain divorce rates quite comparable to contemporary California" (2001: p. 786).

Parry found that marriages among his older Satnami informants were much more fluid than caste-Hindu marriages, and that almost half of all Satnami primary marriages had been terminated by divorce. By high caste-standards, Chhattisgarhi women were immodest; as one of Parry's female informants told him: "a new husband keeps a woman feeling young. Only when she gets old and tired that she settles for what she has got." Satnami women working as laborers commonly embarked upon trysts and elaborate extramarital affairs where commonplace and widespread. What Parry found was that among the younger generation of Satnamis, there was a genuine shift in values away from the free and easy marital ways of their parental generation. In general (and contrary to expectations) Parry found that his younger Satnami informants had a lower rate of divorce than the over-45s. This squared with his findings that among a certain segment of the working class who had availed the benefits of public sector (stable) employment in the Bhilai Steel Plant where many of Parry's informants worked, the figure for divorce in the same age cohort was even lower. In other words, the younger generation, particularly those in more stable forms of employment were rejecting the older styles of casual marriage, divorce, and remarriage punctuated by freer sexual couplings at all stages and were thus keen to emphasize the *indissolubility* of their marital bonds (2001: p. 817). As Parry saw it, the trend for younger Satnamis was away from the systematic (or 'cold-brained' as his informants put it) intimacies and romances accompanying high rates of divorce toward marriages that in Parry's words were more stable and carried a much heavier "emotional freight" (2001: p. 816). Parry's further insight was that the stability of these marriages came at an ideological cost – young Satnami women were far less likely to go out to work, and their increasing domestication as housewives undoubtedly limited their sexual opportunities as well as increasing their economic dependency upon their husbands. The economic (and sexual) independence of their

mothers' generation was thus traded for an emotionally secure, but economically unequal notion of marriage.

Marriage and Love

Parry's argument about marriage and divorce for younger Satnamis in Chhattisgarh runs contrary to Gidden's idea that the increasing prevalence of 'love' as a motive for marriage is tied to economic independence, and that it allows couples to break free from the kinship and familial obligations (and economic dependencies) they might have sustained in the past (see de Munck, 1996). An outcome of the resulting interest in marriage as a process has been an attention to forms of intimacy and love as socially derived. The anthropologist de Munck found that love was a 'deep motivation' for his Muslim informants in the Tamil Sri Lankan village of Kutali, but interestingly, he argued that it was utterly compatible with and indeed, helped to sustain arranged marriages between cross-cousins; thus love was a factor that was serving to perpetuate a particular kinship form of marriage that has been fast disappearing the world over (Fuller and Narasimhan, 2008). Marriages in Kutali were mediated by family members, with *marikars* (mosque officials) playing a central role in helping negotiate the appropriate dowry to be paid. Dowry in this agrarian community always included a house, some paddy land, and some money. 'Love' played no part in the dowry negotiations and the bride and groom were 'publicly backgrounded,' and expected to appear submissive to the authority of the elders, even when the marriage was known to be the outcome of a love affair that had come to light. When such affairs were discovered, the *marikars* sought to 'punish' the couple by invoking Islamic law or (Shari'a) and pressuring both families to agree to a marriage.

What is interesting in de Munck's work is the way in which he shows how love in Kutali has evolved in such a way that it is almost directed toward the most appropriate spouse, one's cross-cousin. So, for instance, boys in the village explained that it was embarrassing to bathe in the river in front of one's 'sister' (real or classificatory) but pleasurable to bathe with a female cross-cousin or *mathini* (1996: p. 708). Furthermore, male circumcision rites provided an 'institutionalized setting' for cross-cousins to recognize themselves as potential marital partners, with their *mathinis* (female cross-cousins) expected to visit them at home as they recuperated and enquire after their health with gifts of sweets.

The sweets were known as *kaum*, a slang word for vagina. Thus, the giving of *kaum* and the enquiry into the health of the 'friend' was interpreted by de Munck as a form of 'institutionalized foreplay.' De Munck says, "The absence of sanctions against public interaction between opposite-sex cross-cousins and the obvious intent of the post-circumcision visits help direct the course of romantic love toward someone from the culturally approved social category" (1996: p. 708). Of 18 newlywed couples that he worked with, 14 declared that they had *anpu* (love) before marriage was discussed, and of these, 10 love marriages were between individuals related as cross-cousins (1996: p. 709). Thus love served as a vital means through which the social structure was maintained and through which the flow of love and sexual desire was directed. In a different vein, my own work (Mody, 2008) with love-marriage couples in Delhi (India) showed that when couples' love became publicized they were frequently forced to inhabit a space of 'not-community,' beyond the pale of social relations and meaningful kinship connections and economic support. Consequently, maximum effort was expended to try and keep romantic 'love' from sight, working on the premise that given time (not infrequently, many years) of dutiful and loving kinship love and care, intransigent parents would eventually come around and agree to arrange a marriage of their children to each other. Here, what was sought was parental consent and its stamp of approval upon a visibly arranged marriage, and it was this flow of public marital legitimacy (concealing private love) that was powerfully deployed to domesticate love marriages in urban India. Given the centrality of consent in current debates about marriage (for instance, 'child marriage,' 'forced marriage') and the ways in which people can be characterized as objects (for example, in discourses on trafficking) or "subjects willingly agreeing to be objects" (Van Baal in Strathern, 1984), marriage continues to provoke profound anthropological questions on transnational migration, kinship, reciprocity, gender and sexuality to ethics, self-making, and care.

See also: Exchange in Anthropology; Family: The Anthropology of the Concept and Its History; Gender: Anthropological Aspects; Kinship in Anthropology; Matrifocality; Queer Anthropology; Sociality In Anthropology.

Bibliography

Ahearn, Laura, 2001. Invitations to Love: Literacy, Love Letters, and Social Change in Nepal. University of Michigan Press, Michigan, MI.

Bell, Duncan, April 1997. Defining marriage & legitimacy. Current Anthropology 38 (2), 237–253.

Bodenhorn, Barbara, 2013. On the road again: movement, marriage, mestizaje, and the race of kinship. In: Cannell, F., McKinnon, S. (Eds.), Vital Relations: Modernity and the Persistent Life of Kinship. SAR Press, Santa Fe, NM.

Boellstorff, T., 2005. The Gay Archipelago: Sexuality and Nation in Indonesia. Princeton University Press, Princeton, NJ.

Bohannan, Paul, 1963. Social Anthropology. Holt, Rinehart and Winston, New York.

Bonaccorso, Monica, 2009. Conceiving Kinship: Assisted Conception, Procreation and Family in Southern Europe. Berghahn, Oxford.

Borneman, J., May 1996. Until death do us part: marriage/death in anthropological discourse. American Ethnologist 23 (2), 215–235.

Borneman, J., 2001. On caring and being cared for: displacing marriage, kinship, gender and sexuality. In: Faubion, J.D. (Ed.), The Ethics of Kinship: Ethnographic Inquiries. Rowman and Littlefield, Lanham & Oxford.

Bott, Elizabeth, 1957. Family and Social Network: Roles, Norms, and External Relationships in Ordinary Urban Families. Tavistock Publications Ltd, London.

Carsten, Janet, 2000. Cultures of Relatedness: New Approaches to the Study of Kinship. Cambridge University Press, Cambridge.

Comaroff, John L., 1980. The Meaning of Marriage Payments. Academic Press, London.

Dumont, Louis, Parkin, R., 2006[1971]. (Trans.). Introduction to Two Theories in Social Anthropology: Descent Groups and Marriage Alliance. Berghann Press, London.

Evans-Pritchard, E., 1951. Kinship and Marriage among the Nuer. Clarendon Press, Oxford.

Evans-Pritchard, E., December 1970. Sexual inversion among the Azande. American Anthropologist, New Series 72 (6), 1428–1434.

Fuller, Chris, Narasimhan, Harpriya, 2008. Companionate marriage in India: the changing marriage system in a middle-class Brahman subcaste. Journal of the Royal Anthropological Institute 14, 736–754.

Gell, Alfred, September 1996. Armour, connaissance et dissimulation (C. Rouslin, Trans.). Terrain: Revue d'ethnologie de l'Europe (27).
Giddens, A., 1992. The Transformation of Intimacy: Sexuality, Love and Eroticism in Modern Societies. Polity, Cambridge.
Hart, Kimerly, 2007. Love by arrangement: the ambiguity of 'spousal choice' in a Turkish village. Journal of the Royal Anthropological Institute, (N.S.) 13, 345–362.
Kulick, Don, 2005. Four hundred thousand Swedish perverts. GLQ: A Journal of Lesbian and Gay Studies 11 (2), 205–235.
Leach, Edmund R., December 1955. Polyandry, inheritance and the definition of marriage. Man 55, 182–186.
Levi-Strauss, Claude, 1969[1949]. The Elementary Structures of Kinship. Beacon Press, Boston, MA.
Masquelier, Adeline, 2005. The Scorpion's sting: youth, marriage and the struggle for social maturity in Niger. Journal of the Royal Anthropological Institute, (N.S.) 11, 59–83.
McLennan, John, 2012[1865]. Primitive Marriage: An Inquiry into the Origin of the Form of Capture in Marriage Ceremonies. Forgotten Books.
Mody, Perveez, 2008. The Intimate State: Love-Marriage and the Law in Delhi. Routledge, Delhi.
Morgan, Lewis H., 1997[1870]. Systems of Consanguinity and Affinity of the Human Family. University of Nebraska, Lincoln, NE.
de Munck, Victor C., November 1996. Love and marriage in a Sri Lankan muslim community: toward a reevaluation of Dravidian marriage practices. American Ethnologist 23 (4), 698–716.
Needham, Rodney, 1971. Remarks on the analysis of kinship & marriage. In: Rethinking Kinship & Marriage. Tavistock Publications, London.
Osella, Filippo, Osella, Caroline, 2000. Social Mobility in Kerala: Modernity and Identity in Conflict. Pluto Press, London.
Parry, J.P., October 2001. Ankalu's errant wife: sex, marriage and industry in contemporary Chhattisgarh. Modern Asian Studies 35 (4), 783–820.
Radcliffe-Brown, A.R., 1952. Structure and Function in Primitive Society. Cohen & West, London.
Rivière, Peter, 1971. Marriage: a reassessment. In: Needham, R. (Ed.), Re-thinking Kinship and Marriage. Tavistock, London, pp. 57–74.
Robbins, Joel, 2004. Becoming Sinners: Christianity and Moral Torment in a Papua New Guinea Society. University of California Press, Berkeley, CA.
Royal Anthropological Institute, 1951. Notes and Queries on Anthropology, sixth ed. Routledge and Kegan Paul, London.
Rubin, Gayle, 1975. The traffic in women: notes on the 'political economy' of sex. In: Reiter, Rayna (Ed.), Toward an Anthropology of Women. Monthly Review Press, New York.
Simpson, Bob, December 1994. Bringing the 'unclear' family into focus: divorce and re-marriage in contemporary Britain. Man (Ns) 29 (4), 831–851.
Strathern, Marilyn, 1984. Marriage exchanges: a Melanesian comment. Annual Review of Anthropology 13, 41–73.
Strathern, Marilyn, 1992. Enterprising kinship: consumer choice and the new reproductive technologies. In: Heelas, P., Morris, P. (Eds.), The Values of the Enterprise Culture. Routledge, London.
Thane, Pat, 2010. Happy Families? History and Family Policy. British Academy Policy Centre, London.
Unnithan-Kumar, Maya, 1997. Identity, Gender and Poverty: New Perspectives on Caste and Tribe in Rajasthan. Berghahn Books, Oxford.
Weston, Kath, 1991. Families We Choose: Lesbians, Gays, Kinship. Columbia University Press, New York.
Winch, Robert, 1968. Marriage. In: Sills, David (Ed.), International Encyclopedia of Social Sciences, vol. 10. The Macmillan Company and the Free Press, pp. 1–8.
Yan, Yunxiang, 2005. The individual and transformation of bridewealth in rural North China. Journal of the Royal Anthropological Institute, (N.S.) 11, 637–658.

Marriage and the Dual-Career Family

Sara Raley, McDaniel College, Westminster, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by L.J. Waite, volume 14, pp. 9261–9265, © 2001, Elsevier Ltd.

Abstract

Marriage and dual-career families are two aspects of work and family that have changed dramatically over the past half century. In addition to the changing definition of marriage, there have also been shifts in the timing and stability of marriages since midcentury. Expectations associated with the roles of wives and husbands have also shifted to make way for the dual-earner family where both spouses are employed in market work. Dual-career families are those where both partners are pursuing careers, typically full time and often with children. This article will explore the changes and implications associated with both marriage and dual-career families.

Changes in Marriage

The legal definition of marriage as well as the rights and benefits associated with marriage have been a source of debate over the past few decades in the United States and other Western nations. Marriage has historically been defined as both a legal and religious union between a man and a woman. The Civil Rights Movement in the 1960s ultimately led to the legalization of interracial marriage and contemporary debates about marriage have focused on whether same-sex couples should be granted the right to marry. Recently, some Western countries have begun to legalize same-sex marriages, including some states in the United States. As these debates progress, the attitudes toward marriage have generally gotten more liberal. Internationally, the acceptance of homosexuality is most pervasive in more secular countries (Pew Research Center, 2013).

The primary reason why the conflicts over who is allowed to marry are so intense is because the legal, social, and economic benefits of marriage are numerous. Proponents of same-sex marriage argue, for example, that there are over 1000 benefits to marriage not available to single people or people in civil unions (an alternative to marriages offered to same-sex couples in many states). Further, some scholars and policy makers argue that marriage offers so many benefits to its members that marriage should be encouraged through policy. Some of the more obvious benefits to marriage include the ability of partners to specialize, pool risk in the case of economic shortfall (in the case of dual-earning spouses), and companionship. Detailed comparisons of married individuals to never married, widowed, divorced, and cohabiting individuals has yielded further insights into how marriage results in greater earnings, wealth health, life span, and even sexual satisfaction for those who marry (Waite, 1995).

Other scholars note that the so-called benefits to marriage are actually selection effects. Those who are the healthiest and have the most earning potential are more desirable marriage mates and thus more likely to enter into marriage. Further, whatever benefits that may arise from marriage itself, and not simply from selection effects, are bestowed unevenly to men and women, with men gaining marginally more from marriage than women. These scholars also argue that policies aimed at promoting marriage – particularly to low-income populations who are less likely to marry – are problematic because research on low-income individuals, particularly single parents, does not show that these individuals do not want to marry or are not aware of the benefits that may be associated with marriage. On the contrary, most people with low income aspire to get married but feel constrained by limited job opportunities and financial instability. They understand that economic stress strains relationships and do not want their marriages to end in divorce (Edin and Kefalas, 2005). As a result, some have concluded that if societies are concerned about marriage rates because they deem marriage desirable then policies should focus on improving the economic infrastructure of communities with low marriage rates if they are to be successful (or happen at all).

Marriage rates, in fact, are at an all-time low in recorded history in the United States at 31.1 – roughly 31 marriages per 1000 unmarried women. Marriage rates were high at midcentury (90.2) when 65% of women aged 15 years and over were married (Cruz, 2013). Still, most people marry or aspire to marry in the United States. Sixty-one percent of those who were not married would like to be someday and the husband–wife household is still the most common type of family household.

Marriage rates have declined partly because of the large shifts in the timing and stability of marriage. The extent to which the timing of marriage has changed over time, however, depends on the starting point of the analysis. If the starting point is the beginning of the twentieth century, then it would seem as if there is only a slight increase in the timing of the first marriage, but if the starting point is midcentury, the shifts appear more dramatic. Median age at first marriage in 1890 was 26.5 years for men and 23.6 years for women compared with 24.0 for men and 20.5 for women in 1950. Only by the 1990s did the median age at first marriage for men and women approach the 1890 levels (Elliott et al., 2012). Since the 1990s, the median age at marriage has continued to increase. In the United States, it is now in the late twenties (approaching age 30 years for men), and over age 30 years in some Western European countries. The continued increase in women's and men's median age at first marriage beyond the turn-of-century levels is likely connected to women's pursuit of educational and careers opportunities not available to them in earlier generations as well as expectations that women be economically independent from men (Martin, 2004). This is partially reflected in the higher age at first marriage among college-educated and post-college educated men and women.

Trends in marriage rates also vary widely by race. Rates are significantly lower among blacks such that about approximately a quarter of black men and women over the age of 35 years have never been married relative to roughly 10% of whites (Elliott et al., 2012). From the beginning of the twentieth century through midcentury, black men and women were getting married at younger ages than white men and women, respectively. By 1950, the median age at marriage for whites and blacks converged and the median age at first marriage for black women and men began to rise above the ages for whites in the decades that followed. At present, the median age at first marriage for black women is around 30.0 years compared with 26.4 years for white women and 30.7 for black men compared with 27.8 years for white men, reflecting a smaller gender disparity in age at first marriage for blacks than white (Elliott et al., 2012).

Cohabitation

While marriage rates have declined, cohabitation rates have increased. It is tempting to assume that cohabitation has become a substitute for marriage, but cohabitation in the United States is more aptly characterized as a precursor to marriage, although cohabitating unions are increasingly diverse and complex. Still, most marriages in the United States are preceded by cohabitation and there are only a handful of countries where the civil benefits associated with cohabitation rival marriage, and thus cohabitation has become a substitute for marriage.

Individuals may also cohabit in the wake of a divorce. Although divorce rates have stabilized since the 1980s, they are still relatively high in most Western nations with the United States and Eastern European countries having some of the highest rates in the world. Comparatively, rates were lower in places like Libya and Mongolia. Data on remarriages is limited, but evidence suggests that most people who get divorced in the United States get remarried, indicating that marriage is still a popular institution. Still, remarriage rates – like marriage rates – have declined (Sweeney, 2010).

Declining marriage rates have led some to question the extent to which marriage is still a powerful institution, and moreover, whether it has undergone a process of deinstitutionalization, meaning the expectations and norms associated with marriage have weakened or become unclear (Burgess and Locke, 1945; Cherlin, 2004). Marriage as a status symbol is still very strong, but what it means to be married is more complicated and varied than in previous generations. Cherlin (2004) argues that marriages moved from being institutional to companionate to ultimately individualized marriages emphasizing personal development, emotional satisfaction, and flexible roles rather than simply sacrificing for one's partner.

Increases in Women's Employment

It is difficult to say whether the changing norms associated with marriage are a result of changing behavior or changing behaviors resulted in changing norms, but women's increased employment and career aspirations are often noted as an explanation for delayed marriage, nonmarriage, and even divorce. The explanation is that women's increased employment leads to greater economic independence and thus reduced economic benefits to marriage (Becker, 1981; McLanahan and Casper, 1995). High rates of women's employment also make singlehood more attractive to women and divorce more feasible options, as women are less financially dependent on men. Changing norms seem to be an incomplete explanation for delayed marriage, however, given that women with attractive economic resources are not the least likely to marry and, in fact, may be more likely to marry (Oppenheimer and Lew, 1995). The period of rising employment has also been marked by slight declines in the earnings of men, particularly young men, which makes marriage less economically attractive for women (Oppenheimer, 1994).

Since the first half of the twentieth century, a period when married women faced discriminatory attitudes and policies limiting their employment opportunities, women's employment has increased steadily, particularly among married women and married mothers (Goldin, 1990). In 1978, 56% of all women aged 25–54 years were employed and this increased to 71% in 1998, or a 15.5% increase. When the universe is restricted to married women, there was an increase of 17.4% between 1978 when employment rates were 51.4% for all married women and 68.8% in 1998. The largest increase at 20.1% was observed for married mothers with preschool-aged children between 1978 and 1998 when rates rose from 38.1% employed to 58.2%. Even though most women were employed by the end of the 1990s, including most married women and most mothers with young children, rates of full-time employment (especially full-time, year-round employment) were much lower, at about half of all women and roughly a third of married mothers with children under age 6 years (Cohen and Bianchi, 1999). Trends in married women's employment rose to some of the highest levels in countries like Sweden where the rates ranged between 80 and 90% (Spain and Bianchi, 1996). These shifts in women's employment – particularly married women's employment – were so sweeping across most Western nations that they are best described as a revolution with few if any signs of reversing.

Division of Labor inside the Home

The movement away from the male breadwinner–female homemaker model to the dual-earner model from the 1970s to present was swift. In 1970, close to 90% of couples had conventional work–family arrangements where the husband was the sole or primary provider. By 2001, most (64%) couples still followed conventional earning arrangements where husbands were still the sole or major provider, but those where wives shared equally in providing income made up nearly a quarter of couples and couples where wives were primary or sole earners were 12% of all couples (Raley et al., 2006). Although the shift was less striking among working class and minority families, who had higher employment rates than middle class and white families in earlier eras, a similar pattern was observed (Bianchi, 1981).

As married women, especially mothers, became part of the labor force, sweeping changes were happening with work inside the home as well. Given that women were now spending

substantially more time in market work than previous generations, and that time is finite, employed women had to reduce the time they devoted to nonmarket work. In fact, the increase in time spent in market work is mostly accounted for by a reduction in time spent in housework (Bianchi, 2000). The reduction in time spent in housework is one that is observed among both employed and unemployed women, perhaps reflecting a decreased emphasis on spotless houses, the increase in time-saving devices to increase efficiency of completing household tasks, or the ability to outsource household tasks.

The gendered division of labor in the home – particularly in terms of housework – receives a lot of attention from scholars because it is seen as an indicator of gender equity to examine how much men and women in dual-career couples share in the additional unpaid work required to maintain a household. Time use studies of housework show that the reduction in housework among wives during the period of rising women's employment cannot be explained by a comparable increase in housework among married men, although married men's housework has steadily increased since the 1960s. Men's household contributions are still lower than women's, but they are closer than they were in the past largely due to declines in women's time rather than increases in men's time (Raley et al., 2012). There are a number of factors that seem to be associated with the proportion of housework that husbands do. For example, husbands' hours increase as they spend fewer hours in the labor market, as their wives spend more time in the labor market, and when they have more gender egalitarian attitudes (Coltrane, 1996).

The Second Shift: Strains of Dual-Career Families with Children

The dramatic increase in married mothers' movement into the labor force between the 1960s and 1970s sparked alarm that mothers were reducing their time with children to accommodate their time spent working outside the home. Fears of neglected, 'latchkey' children who were harmed by too many hours in day care and not enough time with their mother abounded. Yet, careful studies of parents' time use over the period of rising maternal employment (1960–90) suggested that time with children remained stable (Bianchi, 2000; Bianchi et al., 2006).

As employed married mothers stretched their schedules to make sure children were getting the time and attention they need, concern arose that mothers were carrying an unequal burden of the household maintenance compared to their husbands. Arlie Hochschild (1989) coined the term 'second shift' to describe the additional shift of housework and child care that mothers employed full time with young children worked after completing their first shift of market work. In her landmark qualitative study, Hochschild (1989) found that the men in these households did not work a lengthy second shift and that their contributions to the daily management of the household were marginal compared to what their wives did. Further, the men often overestimated their contributions or thought they were more equal than they actually were. Both the husbands and wives in these couples reported marital strain, sleep deprivation, and extreme feelings of time pressure, although feelings were more acute for women than for men.

Subsequent examinations of the total workloads (family work + market work) of men and women in dual-earner couples have found women's workloads to be slightly higher or similar to men (Bianchi et al., 2006; Milkie et al., 2009). Although women do far more nonmarket work than men on average, men tend to spend more time in the labor market, even when both members of the couple are employed full time. Gender disparities are larger for dual-earner couples with children, and these disparities are largest when both partners work full time and have preschool-aged children (Milkie et al., 2009). This population of mothers experiences the largest 'second shift,' and also feels extreme time pressures trying to do it all. Although they do not appear to get much less sleep than their spouses or compared with mothers in other family types, they experience less leisure than their husbands and mothers in other types of households.

Time Pressure and Total Hours of Family Work

Despite the large influx of women into the labor force, the average number of hours employed adults work has remained relatively stable over time. This seems difficult to reconcile with the increased sense of time pressure that so many Americans as well as Europeans report (Robinson and Godbey, 1997). One explanation for this is that work hours have become more bifurcated over time with more people working long hours, particularly those in dual-career marriages, and more people working short hours. Hence, examining average work hours across all families may be misleading. Second, the time deficit is more easily understood as one that families rather than individuals experience. Dual-career couples spend a total of 80–100 h a week in market work relative to single-earner couples who spend about half of that in market work and thus have more time available as a family unit, for nonmarket work and leisure (Jacobs and Gerson, 2004). At the same time, dual-career and dual-earner families may be better able to afford time-saving devices and outsource nonmarket work to protect leisure time as well as time with children.

Work–Life Balance

The seemingly increased pace of life for dual-career families has shifted more attention on the ability of husbands and wives to balance their work and family lives. This is quite challenging in an era where both spouses are not only employed, but also pursuing careers that might require long hours, extended time away from home in the form of frequent travel, and irregular or rotating shifts. Indeed, the most financially lucrative and high-status careers often carry some or all of the aforementioned characteristics, as do many low-status positions with fewer benefits (Presser, 1994). While work life is demanding for both working and middle class families, contemporary family life is also complex when marriages are less stable and the expectations to invest lots of time in both marriages as well as children are very high. Juggling the often competing demands of work and family life in a dual-career marriage is a source of great stress for dual-career families, especially those with children. The past two decades marked a period of

landmark legislation in the United States to address the strains felt by working families with the passage of the Federal and Medical Leave Act in 1993, which provides employees up to 12 weeks unpaid, job-protected leave in the event of a variety of pressing family or medical situations. Since that time, there have been efforts to expand that legislation as well as introduce additional legislation that would ease work and family strains, including universal health care adopted by many other Western nations. Western European nations, particularly the Nordic countries, have some of the most family-friendly policies, including more stringent regulations on employment hours, paid mandatory parental leave, and high-status part time work with substantial health and retirement benefits. Those countries also have greater occupational sex segregation than the United States (Gornick and Meyers, 2005).

Despite the movement toward gender parity in both market and nonmarket work, the implications of combining work and family responsibilities are quite different for men than for women. Given that responsibility for tending the home and caring for children still rests predominantly on women's shoulders, and that the expectations for investing lots of time in children are high, it is not surprising that wives tend to make more sacrifices in their careers for family life than husbands do (Bianchi et al., 2006; Hays, 1996). In fact, if anything, combining marriage, parenthood, and a career is beneficial for men (Nock, 1998; Waite, 1995). Married men tend to earn more and have higher status jobs relative to unmarried men. Further, the birth of children results in longer work hours and greater earnings for men (Nomaguchi and Milkie, 2003). The pattern of greater earnings and status for married men characterizes most of the developed world (Schoeni, 1995). Husbands, particularly upon the birth of a child, may have a greater motivation to assume the role of provider, whereas wives may also desire or feel pressure to help support their husbands in the provider role (Hays, 1996). The persistent gender gap in pay in both the United States and developed countries also helps to perpetuate a model where husbands remain primary breadwinners and wives' careers are secondary. Heavily investing in one's career also seems to pay off for men by making them more attractive mates and once they are married, their marriages are more stable (Bergstrom and Schoeni, 1996).

In contrast to married men, married women's career attainments generally do not result in greater marital stability or earnings, especially when they are mothers. Most women reduce their employment hours when they have children and only about a third of women with very young children work full time, year round (Bianchi et al., 2006). This, in turn, results in reduced earnings in what has been coined the 'motherhood wage penalty' of a 7% reduction in earnings for each child, with larger penalties for married women than unmarried women. The motherhood wage penalty is an additional reduction in earnings to mothers beyond what would be predicted with the lost job experience (Budig and England, 2001). When married women do not have children, they do a greater share of the nonmarket work, even when they are employed similar hours as their husbands (Sayer et al., 2009). Hence their penalty can be conceptualized in workloads rather than earnings. Marriage and parenthood may therefore enhance a man's career attainment, while weakening a woman's. In light of the fact that most married couples are dual earners and most of these couples have children, this means that most dual-earner couples are not in practice dual-career couples where the career trajectories are equally valued or supported. Most dual-career marriages are those where men's careers are primary and women's are secondary (Becker and Moen, 1999).

Earning Arrangements among Dual-Career Couples

Although it is still normative for men to be the primary earners even among dual-earner couples, trends over the past 40 years suggest that this model may have an uncertain future. In about a quarter of couples, wives contribute to family income at equal levels as their husbands, more than double the proportion of equal earners in 1970. Wives who are primary earners have also increased dramatically over the past four decades, although rates are still relatively low, especially if a strict standard of primary breadwinning is applied rather than crude measures where wives earn only a nominal amount more than their husbands (Raley et al., 2006). The trends are almost entirely linear, with an increasing trend toward greater diversity in earning arrangements and arrangements where wives have greater earning power.

Earning arrangements have a curious relationship with housework and the care of children. One might expect that as wives have greater earning power in the relationship and/or less time available to do housework or care for children relative to their husbands, that husbands' time spent in family work would steadily increase. Instead, there is a curvilinear relationship between earnings and housework whereby wives with much higher earnings than their husbands often do as much or more housework than their husbands who earn much less. The earning arrangements where men tend to do the most housework are those where husbands and wives are equal earners (Bittman et al., 2003). Similarly, men in dual-career couples with children do more of the routine, or the day-to-day child care when their wives have greater earning power when compared with men whose wives have less earning power. At the same time, dual-career wives who earn much more than their husbands still spend substantial amount of time in child care, typically surpassing that of their husbands. In other words, their husbands do more child care than other husbands, but still not more than their wives (Raley et al., 2012).

Work Schedules among Dual-Career Couples

The movement toward a service economy in many Western countries has had major implications for work schedules, and consequently, working families in particular. The service economy demands more workers be available for evening, night, and weekend shifts. In over 25% of dual-earner couples in the United States, at least one spouse works a nondaytime shift (Presser, 1994). Shifts are also increasingly variable, meaning workers may have fewer windows of time that are consistently available for family or leisure. Moreover, many dual-earner couples are dealing with a situation where their work schedule does not overlap with that of their spouse. When couples do not work at the same time, they have limited time to spend with each other or together as a family with children if one spouse is

at home while the other is at work. With nonoverlapping or rotating shifts, family dinner may have to be replaced with a family breakfast or forgone altogether (Presser, 1989, 1994).

Although there are negative implications for shift work among dual-career couples in the way of reduced couple time, decreased marital stability, and even negative health implications such as irregular or interrupted sleeping rhythms, there are also implications for greater gender equality (Presser, 1994; Wight et al., 2008). The more time that husbands are at work while their wives are simultaneously working outside the home, the more likely husbands are to spend time doing housework. In particular, they are more likely to do female-typed tasks, such as cooking and cleaning. Similarly, when wives are at home during the same period of time that their husbands are working, wives also spend more time doing housework traditionally done by women. This implies that when wives are not available to do certain housework tasks, their husbands will do them. An economy that increasingly requires irregular, nonstandard, and rotating shifts may indirectly result in a narrowed gender gap among dual-career couples. In dual-career couples where both spouses work nonstandard schedules and are professionals or managers, the gender gap in housework is even narrower than among other nonstandard workers (Presser, 1989, 1994).

Dual-career couples who work nonstandard hours and have children may face even greater time strains on their marriage than other dual-career couples. For example, married parents who work nonstandard hours get less time with their spouses than married parents who work standard schedules. They also get less sleep. At the same time, they may have more one-on-one time with their children, and father involvement may be greater (Wight et al., 2008).

Conclusion

The meaning of marriage has shifted in recent years as same-sex couples fight for the right to marry and people look to marriage as an institution for both companionship and personal development as opposed to self-sacrifice. People are delaying and leaving marriage at higher rates than prior decades. These shifts are connected to a changing economy and changing beliefs about the appropriate roles of men and women, resulting in a now normative family type of the dual-career couple. The great diversity among dual-career couples in work scheduling and earning arrangements suggests a movement toward greater gender equity in married men's and women's allocation of time and resources.

See also: Assortative Mating in the Marriage Market; Cohabitation: United States; Divorce, Sociology of.

Bibliography

Becker, Gary, 1981. A Treatise on the Family. University of Chicago, Chicago, IL.

Becker, Penny, Moen, Phyllis, 1999. Scaling back: dual-earner couples' work–family strategies. Journal of Marriage and the Family 61, 995–1007.

Bergstrom, Ted, Schoeni, Robert, 1996. Income prospects and age at marriage. Journal of Population Economics 9, 115–130.

Bianchi, Suzanne, 1981. Household Composition and Racial Inequality. Rutgers University Press, New Brunswick, NJ.

Bianchi, Suzanne, 2000. Maternal employment and time with children: dramatic change or surprising continuity? Demography 37 (4), 401–414.

Bianchi, Suzanne, Robinson, John, Milkie, Melissa, 2006. Changing Rhythms of American Family Life. ASA Rose Series. Russell Sage, New York.

Bittman, Michael, England, Paula, Folbre, Nancy, Sayer, Liana, Matheson, George, 2003. When does gender trump money? Bargaining and time in household work. American Journal of Sociology 109 (1), 186–214.

Budig, Michelle, England, Paula, 2001. The wage penalty for motherhood. American Sociological Review 66 (2), 204–225.

Burgess, Ernest, Locke, Harvey, 1945. The Family: from Institution to Companionship. American Book, New York.

Cherlin, Andrew, 2004. The deinstitutionalization of American marriage. Journal of Marriage and Family 66 (4), 848–861.

Cohen, Philip, Bianchi, Suzanne, 1999. Marriage, children, and women's employment: what do we know? Monthly Labor Review 122, 22–31.

Coltrane, Scott, 1996. Family Man: Fatherhood, Housework, and Gender Equity. Oxford University Press, New York.

Cruz, Julissa, 2013. Marriage: More than a Century of Change (FP-13-13). National Center for Family and Marriage Research. (online) Available from: http://ncfmr.bgsu.edu/pdf/family_profiles/file131529.pdf (15 August 2013).

Edin, Kathryn, Kefalas, Maria, 2005. Promises I Can Keep: Why Poor Women Put Motherhood before Marriage. University of California Press, Berkeley and Los Angeles.

Elliott, Diana, Krivickas, Kristy, Brault, Matthew, 2012. U.S. Census Bureau Historical Marriage Trends from 1890–2010: A Focus on Race Differences. SEHSD Working Paper Number 2012-12. (online) Available from: http://www.census.gov/hhes/socdemo/marriage/data/acs/ElliottetalPAA2012paper.pdf (15 August 2013).

Goldin, Claudia, 1990. Understanding the Gender Gap: An Economic History of American Women. Oxford University Press, New York.

Gornick, Janet, Meyers, Marcia, 2005. Families that Work: Policies for Reconciling Parenthood and Employment. Russell Sage, New York.

Hays, Sharon, 1996. The Cultural Contradictions of Motherhood. Yale University Press, New Haven.

Hochschild, Arlie, 1989. The Second Shift. Penguin, New York.

Jacobs, Jerry, Gerson, Kathleen, 2004. The Time Divide: Work, Family, and Gender Inequality. Harvard University Press, Cambridge, MA.

Martin, Steve, 2004. Women's education and family timing: outcomes and trends associated with age at marriage and first birth. In: Neckerman, K. (Ed.), Social Inequality. Russell Sage Foundation, New York, pp. 70–118.

McLanahan, Sarah, Casper, Lynne, 1995. Growing diversity and inequality in the american family. In: Ren, Farley (Ed.), State of the Union: America in the 1990s, vol. 2. Russell Sage, New York, pp. 1–46.

Milkie, Melissa, Raley, Sara, Bianchi, Suzanne, 2009. Taking on the second shift: time allocations and time pressures of U.S. Mothers and fathers with preschoolers. Social Forces 88 (2), 487–517.

Nock, Steven, 1998. Marriage in Men's Lives. Oxford University Press, New York.

Nomaguchi, Kei, Milkie, Melissa, 2003. Costs and rewards of children: the effects of becoming a parent on adults' lives. Journal of Marriage and Family 65 (2), 356–374.

Oppenheimer, V.K., 1994. Women's rising employment and the future of the family in industrial societies. Population and Development Review 20, 293–342.

Oppenheimer, Valerie, Lew, Vivian, 1995. American marriage formation in the 1980s: how important was women's economic independence? In: Mason, Karen Oppenheim, Jensen, An-Magritt (Eds.), Gender and Family Change in Industrialized Countries. Clarendon Press, Oxford, UK, pp. 105–138.

Pew Research Center, 2013. The Global Divide on Homosexuality. Pew Research Center, Washington, DC.

Presser, Harriet, 1989. Can we make time for children? The economy, work schedules, and child care. Demography 26, 523–543.

Presser, Harriet, 1994. Employment schedule among dual-earner spouses and the division of household labor by gender. American Sociological Review 59, 348–364.

Raley, Sara, Mattingly, Marybeth, Bianchi, Suzanne, 2006. How dual are dual-income couples? Documenting change from 1970–2001. Journal of Marriage and Family 68 (1), 11–28.

Raley, Sara, Bianchi, Suzanne, Wang, Rong, 2012. When do fathers care? Mothers' economic contribution and fathers' involvement in Childcare. American Journal of Sociology 117 (5), 1422–1459.

Robinson, John, Godbey, Geoffrey, 1997. Time for Life: The Surprising Ways Americans Use Their Time. Pennsylvania State University Press, University Park, PA.

Sayer, Liana, England, Paula, Bittman, Michael, Bianchi, Suzanne, 2009. How long is the second (Plus first) shift? Gender differences in paid, unpaid, and total work time

in Australia and the United States. Journal of Comparative Family Studies 40 (4), 523–544.
Schoeni, Robert, 1995. Marital status and earnings in developed countries. Journal of Population Economics 8, 351–359.
Spain, Daphne, Bianchi, Suzanne, 1996. Balancing Act: Motherhood, Marriage and Employment among American Women. Russell Sage, New York.
Sweeney, Megan, 2010. Remarriage and stepfamilies: strategic sites for family scholarship in the 21st century. Journal of Marriage and Family 72 (3), 667–684.
Waite, Linda, 1995. Does marriage matter? Demography 32, 483–508.
Wight, Vanessa, Raley, Sara, Bianchi, Suzanne, 2008. Time for children, one's spouse and oneself among parents who work nonstandard hours. Social Forces 87 (1), 2–28.

Relevant Websites

http://workfamily.sas.upenn.edu/ – Work and Family Researchers Network.
http://nationalmarriageproject.org/ – University of Virginia – The National Marriage Project.
http://www.timeuse.org/ Univ of Oxford, Dept of Sociology – Centre for Time use Research.

Marriage Systems, Evolution of

Laura Fortunato, Institute of Cognitive and Evolutionary Anthropology, University of Oxford, Oxford, UK; and Santa Fe Institute, Santa Fe, NM, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A marriage system is the set of rules and norms that regulate reproduction in a given human society. This article provides an overview of key concepts and general themes in the study of their evolution. The focus is on the number of spouses allowed (i.e., whether marriage is monogamous or polygamous), and the social and ecological factors associated with this aspect of the marriage system.

Definition of the Subject

A marriage system is the set of rules and norms that regulate reproduction in a given human society. This is closely linked to the set of rules and norms that shape interactions based on relatedness, the kinship system. For example, some societies draw a distinction between 'cross' and 'parallel' cousins. Your cross-cousins are the offspring of your parents' opposite-sex siblings (i.e., the children of your father's sisters and of your mother's brothers), your parallel cousins are the offspring of your parents' same-sex siblings (i.e., the children of your father's brothers and of your mother's sisters). The kinship system of one of these societies may emphasize social ties among cross-cousins over those among parallel cousins. The marriage system may stipulate marriage between cross-cousins as the preferred or prescribed form.

Both relatedness and reproduction play a role in, and are shaped by, the evolutionary process. Hence, the study of kinship and marriage systems figures prominently in biological approaches to human behavior. At the same time, it is a core topic in social and cultural anthropology. This article focuses on research on marriage systems at the interface of the two fields. A complementary treatment of kinship systems can be found in Evolution of Kinship.

Broadly, the rules and norms encoded in the marriage system pertain to two domains. One deals with definition of the socially acceptable marital arrangements: for example, whom one is allowed or not allowed to marry, whether particular categories of individuals are preferred or prescribed marriage partners, and whether multiple spouses are allowed. The other domain deals with the flow of resources at marriage. This includes the transfer of material property (e.g., dowry, where the bride's family transfers land and/or other resources to the newlywed), the provision of labor (e.g., bride-service, where the groom assists the bride's family in productive activities), and the movement of individuals (e.g., virilocal residence, where the newlywed take up residence with or near the groom's kin).

Both domains have been approached from an evolutionary perspective (see reviews and discussion in Low, 2003, 2007). Rather than attempting an exhaustive review of the literature, my aim here is to outline the general principles derived from biology that are used to frame questions about the evolution of marriage systems. To this end, I restrict the focus of the article to the rules and norms that regulate the number of spouses allowed (i.e., whether marriage is monogamous or polygamous) and their environmental correlates, both social and ecological.

Key Concepts

I begin by introducing some key concepts in the study of marriage systems, with the aim of 'unpacking' the relationship between marriage, mating, and reproduction.

What Is Marriage?

The definition of marriage system given above is intentionally vague. Clearly, this is tied to the definition of marriage itself; because of the diversity of family arrangements across human societies, development of a definition that applies universally (and that is universally accepted as such!) has proved surprisingly difficult (Holy, 1996, pp. 48–50).

For the purposes of this article marriage can be defined as a socially recognized union, normatively endorsed, between two or more individuals. This may involve (1) restrictions on sexual relations, typically allowed exclusively between spouses, and (2) the investment of resources in any children associated with the union. Resources may be 'social' (e.g., political status) or physical (e.g., material property), and they may be invested by the spouses themselves and/or by their kin. Where the resources involved are substantial, marriage may be linked to the notion of legitimacy, which serves to channel them to 'appropriate' individuals.

In these key aspects marriage is distinct from mating, irrespective of whether and to what degree pre- and/or extramarital sexual relations are tolerated (Parkin, 1997, pp. 39, 40; discussed further below). Evolutionary approaches have tended to view different marriage systems as alternative reproductive strategies; therefore, this article focuses exclusively on heterosexual marriage. The implications of homosexual marriage for anthropological theory are discussed in Levine (2008).

From a cross-specific perspective, marital unions with parental investment are analogous to nonhuman mating partnerships where mates cooperate in providing resources (e.g., food, protection) to the offspring. There are some peculiarities, however (Irons, 1983, p. 190). In addition to

cooperating in offspring provisioning, marital partners may engage in an extended pattern of reciprocal provisioning. This usually involves a division of labor between the sexes in productive activities: thus, the marital union provides access to resources that are controlled or produced by the opposite sex. Alongside reciprocal provisioning, marital partners entertain relationships of cooperation and reciprocity with several other adult individuals, while maintaining a relatively stable sexually exclusive bond with each other. In nonhuman species mate provisioning tends to be associated with copulation (e.g., courtship feeding), and transfers of resources in this context typically flow from males to females rather than reciprocally (Stevens and Gilby, 2004).

A further peculiar aspect of marital unions, compared to nonhuman mating partnerships, is the potential direct involvement of third parties (Low, 2003, 2007). For example, parents may be actively involved in selecting a spouse for their offspring; together with other kin, they may exert control over the timing of the union. By contrast, nonhuman mating partnerships are typically the outcome of direct interactions between two or more individuals. Of course, wider networks of interactions play a role in the formation of mating partnerships in any social species: for example, coalitionary alliances between males or within-group status-ranking may affect the sequence and timing of the partnerships. This, however, is distinct from the direct involvement of third parties in the arrangement of marital unions in humans.

In the evolutionary literature these peculiarities have often been portrayed as cultural 'frills' attached to the human mating system. In one view, the marriage system is simply a set of rules and norms grafted onto the mating system (e.g., Low, 2003, 2007). Another view draws on the distinction, developed for nonhuman species, between the social mating system and the sexual/genetic one. The former captures the pattern of social living arrangements (e.g., shared use of territory), the latter captures the network of sexual interactions and the resulting genetic relationships. In this view, the marriage system is analogous to the social mating system (e.g., Reichard, 2003, p. 4). As I discuss below, these and related approaches fail to account for several nuances in the interplay between marriage, mating, and reproduction.

Classification of Marriage Systems

In classifying marriage systems, a useful starting point is the distinction between two major exclusive types, based on the number of concurrent spouses allowed: 'monogamy' (individuals may be married to only one spouse at any one time) and 'polygamy' (individuals may be married to multiple spouses at any one time). Within polygamy, 'polyandry' and 'polygyny' refer to multiple concurrent spouses allowed for women and for men, respectively.

These labels derive from combination of the Greek words for single, sole (*mónos*) or many (*polús*) with the words for marriage (*gámos*), man (*anér*), or woman (*gunḗ*). Thus, polygamy is a generic term that captures any form of non-monogamous marriage, even though in the popular press it is commonly used to indicate the marriage of a man to multiple wives instead. In what follows, my use of the labels reflects the technical definitions given above, consistent with their etymology.

Classification of marriage systems as monogamous, polyandrous, or polygynous has proved particularly fruitful. This is likely because the relevant rules are readily observed, even in the absence of individual-level data on marital status; consequently, ethnographies typically include information on this aspect of the marriage system. In the Standard Cross-Cultural Sample (Murdock and White, 1969), a sample of 186 societies that is broadly representative of human cultural variation as reflected in the ethnographic record, classification of marriage systems into exclusive types yields 31 societies (17%) coded as monogamous, 2 (1%) as polyandrous, and 153 (82%) as polygynous (see discussion below). **Figure 1** shows their geographical distribution.

This classification captures variation along the synchronic dimension; additionally, there is variation along the diachronic dimension, which deals with whether individuals are allowed multiple nonconcurrent spouses over the course of their lifetime (Harrell, 1997, p. 22). In practice, the diachronic dimension relates to whether divorced and/or widowed individuals are allowed to remarry, which is not universally the case. For example, remarriage may be allowed for only one of the sexes (typically men).

Somewhat confusingly, in the evolutionary literature a marriage system that enforces monogamy but allows multiple nonconcurrent spouses is known as both 'successive monogamy' and 'successive polygamy'; the terms 'serial' and 'sequential' are also used (Mealey, 2000, p. 321). This confusion reflects the tension between the usefulness of a simple classification scheme, and its limitations in describing the relationship between the rules encoded in the marriage system, actual behavior, and the reproductive implications of each. Evolutionary approaches have been especially prone to confusion: with their focus on reproduction, they have often drawn inferences about mating behavior from information on the marriage system, which can be problematic. I return to this issue below; first, it is useful to clear up two additional potential points of confusion with the classification scheme outlined above.

One point of confusion results from the opposition of monogamy and polygamy as exclusive types. This does not imply that the two marriage systems are 'symmetric' with respect to the marital arrangements allowed: while polygamous unions are forbidden in a monogamous system, the reverse is not true; that is, monogamous unions are of course allowed in a polygamous system (discussed below).

The other point of confusion is that the terms 'polygyny' and 'polyandry' are not used consistently in the anthropological literature. Polygyny generally refers to systems where multiple concurrent spouses are allowed for men but not for women, whereas polyandry is often used for systems where multiple concurrent spouses are allowed for women, even when they are not forbidden for men. The rationale appears to be that in these so-called polyandrous societies multiple concurrent spouses are customary for women, but they are only occasional occurrences for men (discussed below). Terms that more accurately describe these scenarios ('polygynandry' or, better, 'polyandrogyny') are used rarely and inconsistently in anthropology.

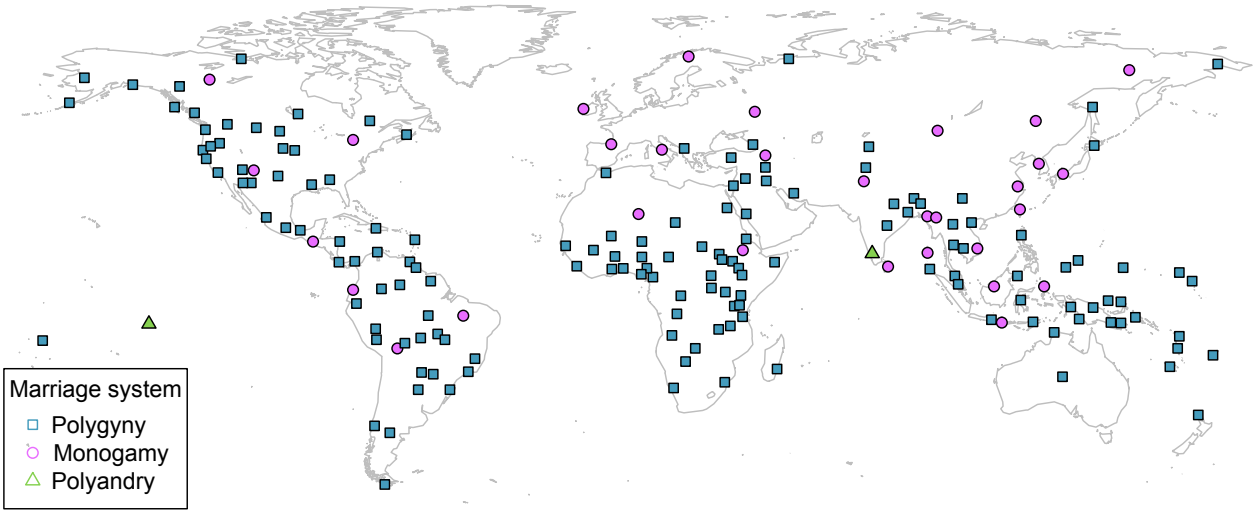

Figure 1 Distribution of marriage systems across societies in the Standard Cross-Cultural Sample (Murdock, G.P., White, D.R., 1969. Standard cross-cultural sample. Ethnology 8 (4), 329–369.). Data from variable 79 (White, D.R., Burton, M.L., Divale, W.T., Gray, J.P., Korotayev, A., Khaltourina, D., n.d. Standard Cross-Cultural Codes. From: http://eclectic.ss.uci.edu/drwhite/courses/SCCCodes.htm (retrieved 31.08.07.)). This variable captures the social composition of the prevailing form of family organization, with monogamy defined in terms of the prohibition or disapproval of polygyny (Murdock, G.P., Wilson, S.F., 1972. Settlement patterns and community organization: cross-cultural codes 3. Ethnology 11 (3), 254–295.). The two categories for polygynous marriage are collapsed (limited polygyny, i.e., less than 20% of men married to plural wives, and general polygyny, i.e., more than 20% of men married to plural wives).

Marriage, Mating, and Reproduction

We can now return to the issue we touched on in the preceding two sections, of how to best characterize the interplay between marriage, mating, and reproduction. A general framework includes four nested levels. At one end is the marriage system, defined above as the set of rules and norms that regulate reproduction. This maps onto the actual pattern of marital arrangements. In turn, this maps onto the actual pattern of mating interactions, both within and outside of marriage. A final distinction is between mating behavior and the genetic level: not all sexual interactions result in the production of children, through a host of mechanisms ranging from sperm competition (*see* Human Mate Choice, Evolution of) to contraception.

We can illustrate this general framework with reference to two fictitious societies. In one society men are allowed multiple concurrent spouses, but remarriage is forbidden (the marriage system); in any case, only a small fraction of men typically engages in plural concurrent unions (the pattern of marital arrangements). In the other society monogamous marriage is enforced, but remarriage is allowed for both men and women (the marriage system); as a result, a large fraction of the population engages in plural nonconcurrent unions (the pattern of marital arrangements). In both societies mating is effectively polygamous: on top of the pattern of marital arrangements, pre- and extramarital sexual relationships are common (the pattern of mating interactions). Finally, reproduction tends to be monogamous, because it is largely confined to sexual interactions within marriage, through the selective use of contraception (the genetic level).

In both cases, knowledge of the marriage system provides limited insight into mating behavior and its reproductive implications. This is why it is problematic to use information on the marriage system to draw inferences about other levels in the framework. Unfortunately, this practice is, to some degree, inevitable. An accurate description of variation in all relevant behaviors would require individual-level genetic data coupled with long-term behavioral observations; these could then be summarized into indices suitable for comparison within and across populations. For obvious practical and ethical reasons, such data are not readily available. As a consequence, studies into the evolution of marriage systems largely rely on typological classifications like the one outlined above, based on the number of concurrent spouses allowed.

Where individual-level data are available (e.g., estimates of reproductive success), these are usually based on self-reports or historical records (e.g., documenting the number of spouses and children), rather than on genetic assessment.

In sum, capturing in a systematic manner the diversity of possible marital and mating arrangements, within and across societies, is riddled with conceptual and empirical difficulties. It will be useful to keep in mind some of the caveats outlined here as we turn to the discussion of marriage system evolution.

Evolutionary Perspectives on Marriage Systems

The 1960s and 1970s saw the development of a conceptual framework for application of evolutionary principles to the study of behavior. This framework was enthusiastically, and in some cases perhaps uncritically, extended to our species. In particular, the framework promised to situate human behavior within a broad cross-specific perspective, while at the same explaining its variation within and across groups. Current evolutionary approaches to human behavior derive

from this framework, in a more or less direct line of descent (see discussion in Laland and Brown, 2011).

Several of the general themes that run through the literature on the evolution of marriage systems developed in this context, grounded in the logic of asymmetry in investment between the sexes (Trivers, 1972). This is the idea that in species with sexual reproduction the sex difference in gamete size leads to divergence in behavioral strategies. By definition, males produce smaller gametes than females. The result of this initial difference in investment is a typical pattern where females care for the offspring, while males seek out additional mates; females are selective in their choice of partners, whereas males are less discerning. This paradigm is currently being revised within biology, to move away from the standard behavioral roles and mating strategies attributed to the two sexes (see discussion in Kokko and Jennions, 2012). The revisions are trickling from biology into anthropology (e.g., Brown et al., 2009; Schacht et al., 2014), but stereotypical assumptions about the behavior of men and women derived from the 'classical' paradigm continue to dominate.

Broadly, studies of the evolution of marriage systems address two sets of questions, relating to variation within and across societies. Where polygamous marriage is allowed, what determines whether an individual marries one or more spouses? And what determines whether the marriage system is monogamous or polygamous in the first place? A third aspect of interest lies at the interface between the two sets of questions, namely the relationship between individual-level outcomes and the rules and norms encoded in the marriage system. For example, does the marriage system emerge through a 'bottom-up' process (e.g., it reflects the prevalent strategy within a group) or through a 'top-down' process (e.g., it reflects the strategy of dominant individuals within the group)? While important, this aspect remains largely unexplored in the evolutionary literature.

Below I outline general themes that run through this literature, with a final word of caution. Although explanations are often phrased in terms of motivations, preferences, choice, strategies, and so on, they do not rest on conscious decision-making by individuals occupied with weighing the fitness outcomes of the different options open to them. Rather, individuals act 'as if' their behavior were driven by such reckoning, possibly mediated by conscious decision-making operating in concert with, or against, the expectation of inclusive fitness maximization. A related issue is that an individual's choice or strategy need not reflect her or his own motivations or preferences: as noted above, it may be the outcome of some degree of 'negotiation' with third parties, such as parents or other kin.

Polygyny as the 'Default' Mating and Marriage System

The central idea derived from the classical paradigm of behavioral biology is that in humans, as in other mammals, polygyny is the 'default' mating system (see recent reviews in Low, 2003, 2007). Men, like males in other species, seek to increase the number of matings, often through intrasexual competition. Women's strategies, like those of females in other species, revolve around intersexual choice, possibly in response to the pattern of resource distribution.

If the marriage system is viewed simply as the cultural amplification of the mating system, a 'baseline' of polygynous mating conveniently explains the prevalence of polygynous marriage systems across societies (82% in the Standard Cross-Cultural Sample; **Figure 1**). The underlying narrative is that men will pursue plural mating as their preferred strategy if individual circumstances allow, and a polygynous marriage system best accommodates this tendency for men to seek polygynous mating whenever possible. The actual pattern of marital arrangements within a polygynous marriage system is the result of variation in individual circumstances.

Two sets of observations are typically used to corroborate this narrative. First, in societies that allow polygynous marriage, polygynously married men tend to be better off, reproductively, than their monogamously married peers. Second, across historical and geographical contexts, dominant men (i.e., those in control of social and/or physical resources, such as political status and material wealth) have directed substantial amounts of effort into the acquisition of multiple mates, in some cases (although not exclusively) through polygynous marriage (Betzig, 1982, 1986). Assuming that dominance enables men to 'get their way,' this is taken to indicate that polygynous mating, possibly through polygynous marriage, is what men ultimately and universally aspire to.

However, this view rests on an overly simplified approach to the relationship between marriage, mating, and reproduction. The underlying assumption is that reproductive success (number of offspring) increases linearly with mating success (number of mates), but this is not necessarily the case (Brown et al., 2009). For example, where males contribute heavily to parental investment, reproductive success increases less than linearly with mating success. The limited evidence available for humans, reviewed in Brown et al. (2009), shows that the relationship between mating and reproductive success varies across populations, for both men and women. This suggests that we should not expect men to always benefit from, and hence seek, polygynous mating.

With reference to the nested framework outlined above, a further difficulty arises from conflation of the marriage and mating levels. Even where men do benefit, reproductively, from polygynous mating, this does not imply that they should always seek polygynous marriage as a result. As we shall see below, in some ecological and social settings polygynous marriage may entail costs for men, which can be avoided through alternative arrangements (Fortunato and Archetti, 2010; Fortunato, 2012). In these contexts, the optimal strategy for men may involve non-polygynous marriage, possibly coupled with polygynous mating (e.g., through extramarital sexual relations).

Finally, recent empirical evidence shows that plural marriage does not necessarily result in long-term reproductive benefits for men. Among the horticultural Pimbwe of Tanzania, there is no evidence of a positive association between number of wives and reproductive success. In this contemporary group polygynous marriage is allowed but uncommon, and plural marriage is typically the result of remarriage following divorce (Borgerhoff Mulder, 2009). In a preindustrial agricultural

Finnish population, number of wives was positively associated with number of children, but it had no effect on number of grandchildren. In this historical group monogamous marriage was enforced, and plural marriage was exclusively the result of remarriage following death of one's spouse (Pettay et al., 2014).

Taken together, these theoretical and empirical insights point to a more nuanced approach to the strategies, motivations, and preferences of men and women than has been generally assumed in the evolutionary literature. As we shall see in the following sections, the view that polygynous marriage best accommodates men's aspiration to mate polygynously has had several repercussions for the study of both polygynous and non-polygynous marriage systems. Somewhat paradoxically perhaps, the strategies, motivations, and preferences of women have received most attention in the context of polygynous marriage – why would a woman marry a married man? Further, non-polygynous systems have typically been approached from the perspective of men, who are viewed as 'relinquishing' polygyny in these systems, possibly driven by peculiarities of the ecological and social environments. As a consequence, explanations have focused on the reproductive implications of non-polygynous marriage systems for men, to demonstrate that relinquishing polygyny 'makes sense,' evolutionarily, from their perspective in these environments. Less attention has been paid to the reproductive implications of non-polygynous marriage systems for women, and to potential conflict and feedback between male and female strategies (Borgerhoff Mulder, 2009; Borgerhoff Mulder and Rauch, 2009).

Polygynous Marriage as Resource Defense Polygyny

Emlen and Oring (1977) introduced a classification of animal mating systems based on the ability of individuals to monopolize mates. One aspect of the classification captures the means through which individuals succeed in this endeavor. For example, in 'female defense polygyny' (also known as 'harem defense polygyny') males control access to mating opportunities directly, by guarding groups of females from potential competitors. By contrast, in 'resource defense polygyny' males exert control indirectly, by monopolizing resources essential to female reproduction (e.g., food or breeding sites).

This classification has found extensive application in the study of marriage system evolution (e.g., Flinn and Low, 1986; Marlowe, 2000). Broadly, the resource defense polygyny model applies where men acquire multiple wives by controlling social resources (e.g., age, political authority) or physical ones (e.g., material wealth). The distinction with the female defense polygyny model is somewhat blurred, however: men's ability to control access to women directly (e.g., in the harems of despotic rulers; Betzig, 1982, 1986) is strongly mediated by their ability to appropriate resources.

Across societies, the degree to which men vary in controlling resources tends to vary with the subsistence system. The resource defense polygyny model applies straightforwardly in non-foraging societies, in which the subsistence system involves the accumulation of physical resources and their transfer across generations, and material wealth may be used directly to acquire a bride (i.e., through bride-wealth, where the groom's family transfers livestock and/or other property to the bride's family). In foraging societies the model may apply where men vary in the degree of control over food-extraction sites (Sellen and Hruschka, 2004) or over social resources (e.g., where age is the basis for access to wives; see discussion in Irons, 1983, pp. 180–185). However, in the majority of these societies there is limited variation across men in the degree of resource control, and the extent of polygynous marriage is therefore limited (e.g., the prerogative of particularly successful hunters). Alexander et al. (1979) labeled this system 'ecologically imposed monogamy' (to distinguish it from 'socially imposed monogamy,' discussed below; note that monogamous marriage is enforced in the latter, but typically not in the former).

The Polygyny Threshold Model, Female Choice, and Male Coercion

Applied to marriage system evolution, the resource defense polygyny model is typically linked to the polygyny threshold model, even though the two do not overlap entirely (see Emlen and Oring, 1977). The polygyny threshold model (Orians, 1969) was developed to explain the occurrence of territorial polygyny in birds, a mating system in which multiple females nest in the territory occupied by their mate (Searcy and Yasukawa, 1989). It applies to situations where males occupy territories of varying quality, and a female's reproductive success increases with the quality of her mate's territory, but it decreases if she has to share resources (e.g., male care, food, nest sites) with other females. Thus, the female incurs a cost by mating with an already-mated male, in the sense that she would be better off mating with the same male but as his only mate.

The polygyny threshold model specifies under what conditions it pays for the female to mate with an already-mated male despite this cost. For example, a female may be better off as the second mate of an already-mated male on a high-quality territory, than as the only mate of an unmated male on a territory of considerably lower quality. In this example, the polygyny threshold is the minimum difference in territory quality between the two males to make the polygynous arrangement advantageous for the female over the monogamous arrangement. According to the model, the occurrence of territorial polygyny in birds indicates that females are sufficiently compensated for the cost of polygynous mating through access to better quality territories (Searcy and Yasukawa, 1989).

Extended to polygynous marriage, this line of reasoning suggests that women may be better off marrying a man who is already married, if he controls sufficiently better resources (in terms of quality and/or quantity) than any of the available unmarried men. A common interpretation, introduced by Chisholm and Burbank (1991), focuses on the reproductive outcomes of women in different marital arrangements within a population. Specifically, if the wives of polygynous husbands do at least as well as the wives of monogamous husbands, this is taken to indicate that polygynous marriage is the result of female choice: the potential costs of sharing resources with co-wives are compensated by access to better resources. If they do

worse, then polygynous marriage is viewed as the result of male coercion: women are 'forced' into polygynous arrangements against their best interest and suffer costs. In this context, male coercion does not involve physical coercion: for example, women may have to tolerate co-wives because they do not have access to other men, or because social norms prevent them from leaving their husbands.

While pervasive in the study of marriage system evolution (e.g., see Lawson and Uggla, 2014, for a recent review), this interpretation rests on a distortion of the framework developed by Searcy and Yasukawa (1989) to explain the occurrence of territorial polygyny in birds. In the original formulation of the framework, under male coercion males are able to physically force females into polygynous arrangements (e.g., through herding or by preventing females from moving off, as occurs in some species of ungulates). By contrast, female choice encompasses a hierarchy of models, some involving reproductive costs for females in polygynous arrangements (e.g., female-biased sex ratios leading to unavailability of unmated males), and others involving no costs (e.g., females recouping the costs of sharing with the male's other mates, through access to better resources). In other words, the polygyny threshold model is only one of a suite of female choice models in the Searcy and Yasukawa (1989) framework. Therefore, it is incorrect to seek to discriminate between male coercion and female choice models based on evidence for or against compensation of the costs of polygynous arrangements to females (an issue discussed already extensively by Searcy and Yasukawa, 1989 and, later, by Borgerhoff Mulder, 1990 in the context of marriage systems).

In short, in the Searcy and Yasukawa (1989) framework costs of polygyny to females are compatible with both male coercion and female choice; in the re-formulation of the framework by Chisholm and Burbank (1991) costs of polygyny to females point exclusively to male coercion. An unfortunate consequence of the conceptual distortion introduced by Chisholm and Burbank (1991) is undue emphasis on male coercion, which is commonly invoked to explain worse outcomes for the wives of polygynous men, compared to the wives of monogamous men. This may partly explain why we have limited understanding of the motivations, or constraints, that lead women into polygynous arrangements, in the face of substantial empirical work focusing on women's outcomes in polygynous marriage systems.

On a related note, the motivations and preferences of men are also poorly understood. This is not surprising, given the emphasis in the evolutionary literature on polygyny as the default mating and marriage system, possibly compounded by a bias in empirical work toward sub-Saharan African societies (Winking et al., 2013). In these societies, polygynous marriage is often associated with prestige, wealth, and status for men; further, multiple wives and their children can be an economic asset, contributing labor to productive activities (Holy, 1996, pp. 63–66). Albeit limited, recent empirical evidence for the forager-horticulturalist Tsimane of Bolivia suggests that similar attitudes are not universal (Winking et al., 2013). In this contemporary population polygynous marriage is allowed but exceedingly rare (and typically sororal, i.e., co-wives are sisters), and both men and women find it less preferable than monogamous marriage. Monogamously married men have no aspiration toward polygyny, and they seem unable to identify any benefits associated with it. Lack of resources is cited as the chief difficulty that arises in polygynous arrangements, followed only distantly by conflict between co-wives.

Monogamous Marriage as Socially Imposed Monogamy

As we saw above, monogamous marriage is enforced in a small but nontrivial fraction of societies (17% in the Standard Cross-Cultural Sample; **Figure 1**). Within an evolutionary framework, explaining the origin of this system against a background of default polygyny presents a challenge. Why would dominant men devise, or allow, a marriage system that removes their prerogative to multiple wives? It seems reasonable to assume that, throughout history, these men have influenced, and to some degree directed, the development of rules and norms encoded in kinship and marriage systems to suit their own interests.

One proposal is that the emergence, maintenance, and later diffusion of such a system across societies were driven by inter-group competition: dominant men had most to gain from success in competition with other groups. In particular, Alexander et al. (1979) proposed that the enforcement of monogamous marriage in stratified societies served to reduce within-group competition among men for access to potential wives, thereby increasing the competitive ability of these societies in between-group conflict. In other words, monogamous marriage emerged as a mechanism of reproductive leveling that served to suppress differences in reproductive success among men. Such a mechanism was a prerequisite for the rise, and persistence, of large nation states. Alexander et al. (1979) termed this system 'socially imposed monogamy' (as distinct from 'ecologically imposed monogamy,' discussed above).

While intuitive, this proposal is not easily reconciled with broad empirical patterns (Fortunato and Archetti, 2010). First, the emergence of monogamous marriage long predates the establishment of large nation states (Herlihy, 1995; Scheidel, 2009). Additionally, the patterned present-day distribution of this marriage system across human societies, suggestive of an association with large nation states, may be the result of an early origin in the history of Indo-European-speaking societies, coupled with diffusion to descendant groups and, later, with processes such as imposition on other groups through conquest (Fortunato et al., 2006; Fortunato, 2008, 2011).

Second, the proposal rests on the key assumption that a monogamous marriage system leads to a substantial reduction in the variance in male reproductive success compared to a polygynous marriage system; in the classical paradigm of behavioral biology, variance in reproductive success underlies competition for access to mates. However, the limited comparative data on male reproductive success available for a sample of societies with monogamous and polygynous marriage systems fail to support this assumption (Fortunato and Archetti, 2010). Consistently, detailed analysis of individual-level data for a preindustrial agricultural Finnish population shows that variance in male reproductive success, and the associated strength of intrasexual competition, can be substantial even in a monogamous marriage system, driven by moderate rates of remarriage following widowerhood (Courtiol et al., 2012).

Finally, recent theoretical revisions of the classical paradigm of behavioral biology (reviewed in Kokko and Jennions, 2012) challenge another key assumption underlying the proposal, namely that the relative 'shortage' of potential wives caused by polygynous marriage inevitably results in increased within-group competition among men. These revisions suggest that male mate competition may be less intense if there is a relative shortage of females available for mating, not more (see discussion in Schacht et al., 2014): why invest effort in competing for an additional mate, if your chances of securing one are low?. In line with the theoretical developments, comparative data on the opportunity for sexual selection (a standardized measure of variance in reproductive success) across a sample of small-scale, nonindustrial societies suggest that the degree of mating competition among men is lower where the sex ratio of the mating pool is more male biased (Schacht et al., 2014).

In sum, if intergroup conflict did indeed play a role in the emergence, maintenance, and later diffusion of monogamous marriage systems, a more nuanced proposal is required that can account for broad empirical patterns, alongside recent theoretical developments.

Monogamous Marriage as 'Monogamous Inheritance'

An alternative proposal explains the emergence of monogamous marriage as a form of 'monogamous inheritance' – effectively a mechanism to channel a man's physical resources to the children of a single wife. In particular, Fortunato and Archetti (2010) investigated whether monogamous marriage can be understood as the outcome of strategic male and female behaviors linked to the allocation of resources to the next generation. Counterintuitively, where this arrangement produces the highest reproductive returns on the investment of resources into future generations, dominant men have most to gain from a monogamous marriage system, because they control the most resources. Further, to the extent that polygynous mating is tolerated alongside monogamous marriage, this proposal does not hinge on reduction in the variance in male reproductive success through monogamous marriage. In this respect, it can better accommodate the distinction between the marriage and mating levels than, for example, the proposal that monogamous marriage is a mechanism of reproductive leveling. At the same time, the theoretical model underlying this proposal (described in Fortunato and Archetti, 2010; Fortunato, 2012) does not consider variation among men in the degree to which they control resources; thus, this factor remains unexplored.

Fortunato and Archetti (2010) identified two conditions that may make monogamous marriage advantageous over alternative arrangements. First, monogamous marriage coupled with unigeniture (i.e., one individual inherits all or most of the resources) can be used to avoid partitioning resources among multiple heirs. For example, in monogamy with primogeniture the wife's first-born child (usually the first-born son) is the designated heir, to the full or partial exclusion of all other children. In polygynous marriage unigeniture typically applies within sets of children by the same mother. For example, in polygyny with primogeniture the first-born son of each wife may be the designated heir among his full siblings. Linking back to the discussion above of the motivations and preferences of women, the polygynous arrangement reflects how each wife (or her kin on her behalf) may require a share of the husband's resources as a condition of marriage. In other words, women may refuse marriage to a married man if inheritance were exclusively by the offspring of the first wife (Fortunato, 2012). As we shall see below, this structural difference in inheritance between monogamous and polygynous arrangements suggests a link to the type of resources transferred, and therefore to the subsistence system.

The second condition that may make monogamous marriage advantageous over alternative arrangements is that women may grant higher fidelity to husbands who marry monogamously over those who marry polygynously. If this is the case, a man's potential reproductive gain from polygynous marriage (i.e., the production of children by multiple wives) may be offset by a loss in terms of reduced paternity, to the point that he may be better off marrying monogamously. In this case, both spouses stand to gain from monogamous marriage: the husband benefits from investing resources 'safely' in those individuals in the next generation who provide the potentially highest reproductive returns (i.e., his wife's children), and the wife benefits from exclusive investment of her husband's resources in her own children.

This proposal is compatible with broad patterns in the history and geographical distribution of marriage systems – namely, why monogamy has prevailed across societies of Europe and Asia practicing intensive agriculture (i.e., plow or irrigation agriculture) and first emerged in these regions, whereas polygyny has prevailed across sub-Saharan African societies with subsistence systems based on extensive agriculture (i.e., horticulture, or hoe agriculture) or pastoralism (Goody, 1976). In intensive agriculture high productivity results in scarcity of land, and division of estates among multiple heirs over the generations leads to plots of land that are too small to sustain a household, effectively depleting the value of the estates. In horticulture scarcity of land is unlikely because productivity is low, and in pastoralism the reproduction of livestock acts as a buffer against scarcity. As a result, productivity is constrained by ownership of land in intensive agriculture, and by availability of labor in horticulture and pastoralism. This difference in the type of resources transferred, itself linked to the subsistence system, may have driven the emergence of monogamous marriage as a mechanism of 'monogamous inheritance' in the agrarian societies of Europe and Asia.

Several predictions can be derived from the proposal, for testing against data in the archaeological, historical, and ethnographic records (Fortunato and Archetti, 2010). For example, we expect to find evidence for an early origin of monogamous marriage, linked to the development of intensive modes of production, and possibly coupled with norms promoting fidelity (e.g., ideologies of honor, or emphasis on female virginity at marriage). This prediction is compatible with the broad historical patterns discussed above with reference to socially imposed monogamy. Further, we expect to see evidence of transitions from a polygynous to a monogamous marriage system (e.g., decreasing rates of polygynous marriage) linked to declines in availability of land, as one of a host of mechanisms

that serve to limit the number of heirs in a given household. Finally, in the ethnographic record we expect to find an association between monogamous marriage and inheritance of resources by a man's wife's offspring (Fortunato, 2009). Note that, for the most part, these predictions relate to the first condition identified by Fortunato and Archetti (2010). Predictions relating to the second condition may be harder to validate, given the difficulties involved in obtaining reliable comparative data on paternity rates, for both (pre-)historical and contemporary populations.

Polyandrous Marriage as a Response to Environmental Challenges

As we saw above, only 1% of societies in the Standard Cross-Cultural Sample are coded as polyandrous (**Figure 1**). This is likely an underestimate, however, reflecting a bias in anthropology toward a narrowly focused set of classic ethnographic descriptions (Starkweather and Hames, 2012). With one exception (the historical population of the Marquesas Islands in the South Pacific; see discussion in Harrell, 1997, pp. 267–271), these descriptions pertain to societies in the Himalayan regions of India, Nepal, and Tibet (discussed below). Starkweather and Hames (2012) recently identified several additional cases in the ethnographic record, noting that they have received limited attention in the study of marriage system evolution.

According to Starkweather and Hames (2012), two key differences between 'classical' and 'nonclassical' polyandrous marriage systems are that the incidence of polyandrous unions tends to be considerably higher in the former, and that polygyny alongside polyandry is found in the former but not in the latter. Recall that terms that more accurately describe the co-occurrence of the two forms of plural marriage ('polyandrogyny' and 'polygynandry') are used rarely and inconsistently in anthropology. Below I use the term 'polyandrogyny' for systems where both sexes vary in number of spouses, but women are more variable than men (Shuster and Wade, 2003, p. 161).

Our understanding of marriage systems where plural concurrent husbands are allowed is rather limited, perhaps as a result of these and other biases (e.g., the near-exclusive focus on outcomes for men). Broadly, this phenomenon is viewed as a response to peculiar social and/or ecological conditions, such as an unbalanced sex ratio leading to severe scarcity of marriageable men or a circumscribed, harsh ecology requiring the pooling of male labor (see reviews in Levine and Silk, 1997; Starkweather and Hames, 2012).

For example, a particular combination of ecological conditions may explain why some societies in the Himalayan regions of India, Nepal, and Tibet resort to fraternal polyandrogynous marriage, where one set of brothers is married to a single wife (see review and discussion in Barrett et al., 2002, pp. 224–233). As in other intensive agricultural societies of Europe and Asia, the risk of land scarcity is high (in this case through low carrying capacity rather than through increased productivity and subsequent population growth). At the same time, the land requires high labor input and yields low productivity, as in the extensive agricultural societies of sub-Saharan Africa (Fortunato, 2009, p. 38). In this ecological context, fraternal polyandrogynous marriage may serve two functions (Goody, 1976, 1990). The first is to limit the number of heirs to an estate: the 'monomarital principle' (Goldstein, 1971) limits the number of wives whose children have claims to the estate. The second function is to concentrate the labor of a set of brothers into the same household, ensuring an adequate workforce against the low productivity of the land.

From an evolutionary perspective, if this arrangement is indeed a response to harsh ecological conditions, a man in a polyandrous or polyandrogynous union is expected to do better, reproductively, than his peers in monogamous or polygynous unions (taking into account his own reproductive outcomes and those of his brothers). However, tests of this prediction are inconclusive, marred by practical difficulties such as the relatively small numbers of brothers at a given 'rank' within a household and uncertainty in the assignment of paternity among them (see review in Barrett et al., 2002, pp. 224–233). Conceptually, a further difficulty arises from interpretation of the data in an evolutionary context. If a key function of polyandrous/polyandrogynous arrangements is to limit the number of heirs, in the interest of long-term lineage survival (i.e., by keeping the family farm intact), the focus on the reproductive outcomes of individual men is misplaced. Analysis should focus on the reproductive implications for men (and women) as parents, rather than on those for men as husbands (Barrett et al., 2002, p. 229). Further insight may derive from research into the factors associated with the dissolution of polyandrous/polyandrogynous households when outside options for men become available, such as employment opportunities not linked to the family farm (e.g., Levine and Silk, 1997; Haddix, 2001).

An additional open question relates to broader patterns of variation across societies in the Himalayan region faced with what appear to be similar ecological challenges: not all groups resort to plural concurrent marriage for women or to the same set of inheritance norms. This has led to the suggestion that historical contingency may have played a greater role than ecological factors in shaping the observed distribution of marriage systems in the region (Berreman, 1978). In sum, of all the marriage systems those involving plural concurrent husbands are the least well understood.

Concluding Remarks

This article has introduced key concepts and general themes in the study of marriage system evolution. The overview has revealed some broad trends. Marriage systems were initially viewed as the cultural 'analog' of mating systems; as a result, early work focused on their ecological correlates (e.g., amount of rainfall or parasite load), failing however to uncover clear patterns (see reviews in Low, 2003, 2007). The realization that marriage may have more to do with the transfer of social and/or physical resources across generations than with mating and the production of children is shifting the focus toward the interplay of the marriage system with a broader range of ecological and social factors.

At the outset, this body of work was firmly grounded in the classical paradigm of behavioral biology, with its emphasis on asymmetries in investment between the sexes. Recent revisions of the paradigm are slowly being incorporated into evolutionary

approaches to marriage systems, but classical views continue to dominate (e.g., a bias toward the reproductive outcomes for men; Borgerhoff Mulder, 2009). These revisions suggest a more nuanced approach to conflict and to feedback between male and female strategies, and their coevolution with environmental factors.

See also: Comparative Method in Evolutionary Studies; Human Behavioral Ecology; Human Mate Choice, Evolution of; Kin Selection; Kinship, Evolution of; Sociobiology: Overview.

Bibliography

Alexander, R.D., Hoogland, J.L., Howard, R.D., Noonan, K.M., Sherman, P.W., 1979. Sexual dimorphisms and breeding systems in pinnipeds, ungulates, primates, and humans. In: Chagnon, N.A., Irons, W. (Eds.), Evolutionary Biology and Human Social Behavior: An Anthropological Perspective. Duxbury Press, North Scituate, MA, pp. 402–435 (Chapter 15).

Barrett, L., Dunbar, R., Lycett, J., 2002. Human Evolutionary Psychology. Palgrave, Basingstoke.

Berreman, G.D., 1978. Ecology, demography and domestic strategies in the western Himalayas. Journal of Anthropological Research 34 (3), 326–368.

Betzig, L.L., 1982. Despotism and differential reproduction: a cross-cultural correlation of conflict asymmetry, hierarchy, and degree of polygyny. Ethology and Sociobiology 3 (4), 209–221.

Betzig, L.L., 1986. Despotism and Differential Reproduction: A Darwinian View of History. Aldine, New York.

Borgerhoff Mulder, M., 1990. Kipsigis women's preferences for wealthy men: evidence for female choice in mammals? Behavioral Ecology and Sociobiology 27 (4), 255–264.

Borgerhoff Mulder, M., 2009. Serial monogamy as polygyny or polyandry? Marriage in the Tanzanian Pimbwe. Human Nature 20 (2), 130–150.

Borgerhoff Mulder, M., Rauch, K.L., 2009. Sexual conflict in humans: variations and solutions. Evolutionary Anthropology 18 (5), 201–214.

Brown, G.R., Laland, K.N., Borgerhoff Mulder, M., 2009. Bateman's principles and human sex roles. Trends in Ecology & Evolution 24 (6), 297–304.

Chisholm, J.S., Burbank, V.K., 1991. Monogamy and polygyny in Southeast Arnhem Land: male coercion and female choice. Ethology and Sociobiology 12 (4), 291–313.

Courtiol, A., Pettay, J.E., Jokela, M., Rotkirch, A., Lummaa, V., 2012. Natural and sexual selection in a monogamous historical human population. Proceedings of the National Academy of Sciences of the United States of America 109 (21), 8044–8049.

Emlen, S.T., Oring, L.W., 1977. Ecology, sexual selection, and the evolution of mating systems. Science 197 (4300), 215–223.

Flinn, M.V., Low, B.S., 1986. Resource distribution, social competition, and mating patterns in human societies. In: Rubenstein, D.I., Wrangham, R.W. (Eds.), Ecological Aspects of Social Evolution: Birds and Mammals. Princeton University Press, Princeton, NJ, pp. 217–243 (Chapter 11).

Fortunato, L., 2008. A phylogenetic approach to the history of cultural practices. In: Allen, N.J., Callan, H., Dunbar, R., James, W. (Eds.), Early Human Kinship: From Sex to Social Reproduction. Blackwell Publishing Ltd., Oxford, pp. 189–199 (Chapter 11).

Fortunato, L., 2009. The Evolution of Kinship and Marriage Systems. Doctor of Philosophy in Anthropology. University College London. Available from: http://discovery.ucl.ac.uk/18995/.

Fortunato, L., 2011. Reconstructing the history of marriage strategies in Indo-European-speaking societies: monogamy and polygyny. Human Biology 83 (1), 87–105.

Fortunato, L., 2012. The evolution of matrilineal kinship organization. Proceedings of the Royal Society B: Biological Sciences 279 (1749), 4939–4945.

Fortunato, L., Archetti, M., 2010. Evolution of monogamous marriage by maximization of inclusive fitness. Journal of Evolutionary Biology 23 (1), 149–156.

Fortunato, L., Holden, C., Mace, R., 2006. From bridewealth to dowry? A Bayesian estimation of ancestral states of marriage transfers in Indo-European groups. Human Nature 17 (4), 355–376.

Goldstein, M.C., 1971. Stratification, polyandry, and family structure in central Tibet. South-Western Journal of Anthropology 27 (1), 64–74.

Goody, J., 1976. Production and Reproduction: A Comparative Study of the Domestic Domain. Cambridge University Press, Cambridge.

Goody, J., 1990. The Oriental, the Ancient and the Primitive: Systems of Marriage and the Family in the Pre-industrial Societies of Eurasia. Cambridge University Press, Cambridge.

Haddix, K.A., 2001. Leaving your wife and your brothers: when polyandrous marriages fall apart. Evolution and Human Behavior 22 (1), 47–60.

Harrell, S., 1997. Human Families. Westview Press, Boulder, CO.

Herlihy, D., 1995. Biology and history: the triumph of monogamy. Journal of Interdisciplinary History 25 (4), 571–583.

Holy, L., 1996. Anthropological Perspectives on Kinship. Pluto Press, London.

Irons, W., 1983. Human female reproductive strategies. In: Wasser, S.K. (Ed.), Social Behavior of Female Vertebrates. Academic Press, New York, pp. 169–213 (Chapter 7).

Kokko, H., Jennions, M.D., 2012. Sex differences in parental care. In: Royle, N.J., Smiseth, P.T., Kölliker, M. (Eds.), The Evolution of Parental Care. Oxford University Press, Oxford, pp. 101–116 (Chapter 6).

Laland, K.N., Brown, G.R., 2011. Sense and Nonsense: Evolutionary Perspectives on Human Behaviour, second ed. Oxford University Press, Oxford.

Lawson, D.W., Uggla, C., 2014. Family structure and health in the developing world: what can evolutionary anthropology contribute to population health science? In: Gibson, M.A., Lawson, D.W. (Eds.), Applied Evolutionary Anthropology: Darwinian Approaches to Contemporary World Issues. Springer, New York, pp. 85–118 (Chapter 5).

Levine, N.E., 2008. Alternative kinship, marriage, and reproduction. Annual Review of Anthropology 37 (1), 375–389.

Levine, N.E., Silk, J.B., 1997. Why polyandry fails: sources of instability in polyandrous marriages. Current Anthropology 38 (3), 375–398.

Low, B.S., 2003. In: Reichard, Boesch (Eds.), Ecological and Social Complexities in Human Monogamy (Chapter 11).

Low, B.S., 2007. Ecological and socio-cultural impacts on mating and marriage systems. In: Dunbar, R.I.M., Barrett, L. (Eds.), Oxford Handbook of Evolutionary Psychology. Oxford University Press, Oxford, pp. 449–462 (Chapter 30).

Marlowe, F., 2000. Paternal investment and the human mating system. Behavioural Processes 51 (1–3), 45–61.

Mealey, L., 2000. Sex Differences: Development and Evolutionary Strategies. Academic Press, San Diego, CA.

Murdock, G.P., White, D.R., 1969. Standard cross-cultural sample. Ethnology 8 (4), 329–369.

Murdock, G.P., Wilson, S.F., 1972. Settlement patterns and community organization: cross-cultural codes 3. Ethnology 11 (3), 254–295.

Orians, G.H., 1969. On the evolution of mating systems in birds and mammals. The American Naturalist 103 (934), 589–603.

Parkin, R., 1997. Kinship: An Introduction to Basic Concepts. Blackwell Publishers Ltd., Oxford.

Pettay, J.E., Rotkirch, A., Courtiol, A., Jokela, M., Lummaa, V., 2014. Effects of remarriage after widowhood on long-term fitness in a monogamous historical human population. Behavioral Ecology and Sociobiology 68 (1), 135–143.

Reichard, U.H., 2003. In: Reichard, Boesch (Eds.), Monogamy: Past and Present, pp. 3–26 (Chapter 1).

Reichard, U.H., Boesch, C., 2003. Monogamy: Mating Strategies and Partnerships in Birds, Humans and Other Mammals. Cambridge University Press.

Schacht, R., Rauch, K.L., Borgerhoff Mulder, M., 2014. Too many men: the violence problem? Trends in Ecology & Evolution 29 (4), 214–222.

Scheidel, W., 2009. A peculiar institution? Greco-Roman monogamy in global context. The History of the Family 14 (3), 280–291.

Searcy, W.A., Yasukawa, K., 1989. Alternative models of territorial polygyny in birds. The American Naturalist 134 (3), 323–343.

Sellen, D.W., Hruschka, D.J., 2004. Extracted-food resource-defense polygyny in native Western North American societies at contact. Current Anthropology 45 (5), 707–714.

Shuster, S.M., Wade, M.J., 2003. Mating Systems and Strategies. Princeton University Press, Princeton, NJ.

Starkweather, K.E., Hames, R., 2012. A survey of non-classical polyandry. Human Nature 23 (2), 149–172.

Stevens, J.R., Gilby, I.C., 2004. A conceptual framework for nonkin food sharing: timing and currency of benefits. Animal Behaviour 67 (4), 603–614.

Trivers, R.L., 1972. Parental investment and sexual selection. In: Campbell, B. (Ed.), Sexual Selection and the Descent of Man 1871–1971. Heinemann, London, pp. 136–179 (Chapter 7).

White, D.R., Burton, M.L., Divale, W.T., Gray, J.P., Korotayev, A., Khaltourina, D., n.d. Standard Cross-Cultural Codes. From: http://eclectic.ss.uci.edu/drwhite/courses/SCCCodes.htm (retrieved 31.08.07.).

Winking, J., Stieglitz, J., Kurten, J., Kaplan, H., Gurven, M., 2013. Polygyny among the Tsimane of Bolivia: an improved method for testing the polygyny-fertility hypothesis. Proceedings of the Royal Society B: Biological Sciences 280 (1756).

Marshall, Alfred (1842–1924)

John K Whitaker, University of Virginia, Charlottesville, VA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9270–9276, © 2001, Elsevier Ltd.

Abstract

Founder of the 'Cambridge School of Economics' and mentor of John Maynard Keynes, Marshall was one of the leading economists of his era. After studying mathematics at Cambridge University in the UK he taught economics at Cambridge, Bristol, and Oxford before returning to Cambridge as professor in 1885, retiring in 1908. His influential *Principles of Economics* (1890) set out his main contributions to the theories of value and distribution. Especially associated with his name are the concepts of partial equilibrium analysis, consumer surplus, short-period versus long-period analysis, external versus internal scale economies, derived demand, and demand elasticity, all still in general use. He remained in the deductive tradition of the British classical school, seeking to improve rather than overthrow it, emphasizing the need to use deduction cautiously, with close attention to the facts of the situation being analyzed. Influenced by evolutionary ideas he stressed the importance of changes in the human factor. He published too late to be an originator of the neoclassical or marginalist revolution commencing in the early 1870s but was prominent in extending marginalism to distribution theory. Two further works, *Industry and Trade* (1919) and *Money Credit and Commerce* (1923) came too late to have much impact on economic thought.

Alfred Marshall, one of the leading economists of his era, was the progenitor of the 'Cambridge School of Economics,' which rose to world prominence in the interwar period. His *Principles of Economics* (1890) was a seminal work in the development of economic thought and introduced tools still prominent in the practice of applied economics.

The Man

Marshall was born in Bermondsey, a London suburb, on 26 July 1842, second son of William Marshall, a clerk at the Bank of England, and his wife Rebecca, née Oliver. He grew up in London, attending the venerable Merchant Taylors' School where he revealed considerable intellectual power and an aptitude for mathematics. Entering Cambridge University, he pursued the prestigious mathematical tripos, graduating in 1865 in the exalted position of 'second wrangler' – a result guaranteeing election as a fellow of his college, St. John's. His interests soon turned from mathematics to questions of human and social existence, and in 1868 he became a lecturer in moral sciences at St. John's. The moral science tripos covered philosophy, logic, and political economy. He lectured on political economy more from necessity than choice. But, perceiving that human and social advance would be possible only if economic constraints became less oppressive he soon settled on the subject as the focus for his life's work and was to be influential in renaming it 'economics.'

He rapidly became Cambridge's dominant teacher of the subject. While publishing little, his early manuscripts (reproduced in Whitaker, 1975) demonstrate that by the early 1870s he had developed many of the distinctive theoretical ideas later associated with his name. He also strove to attain a realistic grasp of the economic world both past and present. Reluctant to produce a theoretical monograph devoid of concrete application, he commenced a volume on international trade questions, visiting North America in the summer of 1875 for fieldwork. Although largely completed and accepted for publication, the project was abandoned in the late 1870s. But in 1879 H. Sidgwick privately printed two of its appendices for circulation in Cambridge and a few copies reached economists elsewhere. Known as *The Pure Theory of Foreign Trade* and *The Pure Theory of Domestic Values* these remarkable pieces were the high watermark of Marshall's achievement in pure economic theory but were not widely known or accessible in his lifetime. (Whitaker, 1975 provides the fullest version together with surviving portions of the text of the abandoned volume.)

In 1876 Marshall became engaged to Mary Paley, a Newnham lecturer, who had attended his early lectures for women and informally sat the 1874 moral sciences tripos examination. The two were married on 17 August 1877. Marriage required Marshall to resign his fellowship and he found a new livelihood as first principal of University College, Bristol. Here both Marshall and his wife taught economics. Meanwhile, he had taken a hand in the writing of a primer for extension classes, started by Mary Paley before their engagement. It increasingly became his book and appeared under their joint names as *The Economics of Industry* (1879). While ill-judged for its intended audience, it contained much novel material, including a theory of income distribution on marginal productivity lines. Its original features helped establish Marshall's reputation amongst economists, especially abroad.

The fledgling Bristol college lacked financial support, and the tasks of administration and fund raising proved uncongenial to Marshall, anxious as he was to develop and publish his own ideas. Ill health, diagnosed as kidney stones, constrained him further and he resigned with relief in July 1881, spending much of the next year traveling in Europe, with an extended stay in Palermo where composition of his *Principles* began.

The following academic year was again spent in Bristol, where the pair taught economics, after which Marshall – who had become a protégé of B. Jowett, the eminent Master of Balliol – moved to Balliol College, Oxford, as lecturer to the probationers for the Indian Civil Service. The move to Oxford had been viewed as permanent with good prospects for elevation, but the sudden death of H. Fawcett – professor of political

economy at Cambridge since 1863 – offered the possibility for a return to Cambridge in an assured, independent, and well-paid position. Marshall had already come to be recognized, after the 1882 death of W. S. Jevons, as the UK's leading economist and he took up the Cambridge chair in early 1885, retaining it until 1908 when he voluntarily retired at age 65.

Fawcett had carried the professorship lightly, combining it with an active political career. But Cambridge was changing rapidly and expanding its commitment to modern subjects. With the lapsing of celibacy requirements, new-model professors and lecturers devoted themselves to the tasks of teaching and advancing knowledge. Marshall gave much effort over the years to lecturing and advising students, urging the more promising to follow the high calling of the economist. Despite frequent disappointments, he did manage to inspire a number of students who were to be prominent in the development of British economics, among whom A. C. Pigou and J. M. Keynes were conspicuous. From the outset, Marshall strove as professor to expand Cambridge's scope for economic teaching, but with only modest success. He eventually secured in 1903 a new specialized tripos in economics and politics, but few resources were provided for it. The new degree only became firmly established after his retirement, but the seed he planted reached full flowering in the 'Cambridge School' of the interwar period.

The 1880s saw the most copious flow of occasional publications by Marshall as well as his important evidence to the Gold and Silver Commission of 1888–89. (Most of his occasional writings are collected in Pigou 1925, which includes a virtually complete, bibliography. Materials concerning his participation in government enquiries are collected in Keynes, 1926 and Groenewegen, 1996.) *Principles*, on which he had been working since 1881, at last appeared to widespread acclaim in 1890. It was a success not only among academic peers at home and abroad but also with an educated public, gratified by its undogmatic and conciliatory tone and its evident concern for human well-being. Marshall's intention had been to follow *Principles* – initially subtitled Volume I – with a companion volume devoted to the applied branches of economics, an idea eventually abandoned as unworkable after intermittent efforts over the next decade or more. Meanwhile considerable effort went into revising *Principles*, which was to go through eight editions in his lifetime, the last being in 1920. (Guillebaud, 1961 reprints the last edition with variant passages from earlier editions.) The copious rewritings and frequent rearrangements did not mark significant changes in Marshall's views and tended to impair the work's vigor and overall clarity. A condensation of *Principles* appeared as *Elements of Economics of Industry* in 1892, replacing the earlier work written with his wife, which he had come to dislike intensely. (Like the earlier work it included a treatment of trades unions omitted from *Principles*.)

Marshall's next book, *Industry and Trade*, only appeared in 1919, long after his retirement. It was started in 1904 in the aftermath of the UK's intense controversy over the desirability of maintaining free trade and was initially conceived as a short topical book on the issue of the day. But it soon developed into a more ambitious attempt to study the interrelationships between distinctive national conditions of industry and patterns of international trade. A motivating force was Marshall's deep concern for the UK's seemingly parlous economic future. Progress on the book was slow and delayed by war. The volume appearing in 1919 was only a fragment of the complete scheme with a vague promise, never fulfilled, of a continuation. An applied work, rich in contemporary detail but lacking in focus, it had a respectful reception but little impact. It nevertheless remains a remarkable if neglected work, displaying full vigor and wisdom of mind and much acumen about economic arrangements.

Marshall's last book, *Money Credit and Commerce*, was assembled with unprecedented speed, appearing in 1923. It brought together material, often written decades earlier, on money, the business cycle, and international trade and payments. Here the 50-year-old graphical analysis of *The Pure Theory of Foreign Trade* was published at last. Unfortunately, waning powers precluded a new synthesis, but the book has incidental merits.

Marshall died on 13 July 1924 at Balliol Croft, his Cambridge home since 1886. His wife was to outlive him by almost 20 years. She taught economics at Newnham for many years after their return to Cambridge, but any intellectual ambitions of her own were subordinated to the task of relieving her husband of everyday vexations.

Balliol Croft was the center about which Marshall's life had revolved after 1886. Under pleas of weak health he only left reluctantly as professional or social duties required. However, summers were normally spent in lodgings away from Cambridge – usually on the south coast of England or in the Tyrol – where he could write free of everyday distractions. But he was not a recluse. Balliol Croft saw many visitors and there were frequent 'at homes' for dispensing advice to students. While not relishing academic dispute, he strove to keep in touch with the real world by quizzing all in active life who came in his way: businessmen, trades unionists, social workers, and so on. Factory visits and inspections of working class life were also valued.

Marshall's was a complex and contradictory personality, which it is now difficult to assess. Prone to fuss over details, and sometimes lacking common sense, he was nevertheless inspirational, especially to the young, in his commitment to the advancement of his subject and his intense conviction of its importance for society's well-being. He had remarkable conversational powers and an impressive ability to speak extempore. Above all, his subtle and wide-ranging mind saw complexities that eluded others and made it impossible for him to communicate his thought in a cut-and-dried way. His economic writings both benefit and suffer from this characteristic. They benefit from a breadth of vision and a richness of unexpected and thought-provoking insights. They suffer from a lack of full coherence and clear articulation, making it frustratingly difficult to grasp and describe his thought comprehensively. Marshall himself struggled incessantly with the daunting task of transferring his thought to the printed page, a task further complicated by an undue sensitivity to criticism and an incessant urge to rewrite already adequate texts.

The Background to Marshall's Economics

British economics was at a critical juncture when Marshall came to it around 1870. The ideas of the classical economists,

especially J. S. Mill whose *Principles* (1848) was still dominant, seemed out of tune with the needs and spirit of the age. The publication of Jevons's *Theory* (1871) was a notable harbinger of the coming transformation from classical to neoclassical economics, which – while maintaining the deductive tradition of the classicals – was to emphasize the role of demand, the assumption of optimizing behavior, and the use of mathematical formalization. The British deductive tradition itself was coming under increasing attack from historically minded economists, especially in Germany, who sought to rely on induction and case studies as the best means of augmenting economic knowledge. More generally, Darwinian ideas on evolution were challenging traditional certainties, while there was increasing social concern over the growing political power of the working classes and the hard plight of the poorest.

Marshall was strongly influenced by this intellectual milieu. His initial economic studies centered on Mill's rendition of classical theory but his mathematical background, allied with his early discovery of the neglected work of A. A. Cournot (1838), pushed him towards a mathematical – mainly geometrical – mode of argument. In contrast to Jevons he saw no need to overthrow the inherited classical system and only sought to close its gaps while preserving its important insights. He remained firmly in the British deductive tradition but mistrusted extended deductions based on oversimplified formalizations. He sought to guide deduction by maintaining close contact with observation and history and by allowing fully for disturbing causes. Although somewhat sympathetic to the aims of the historical economists, he emphasized that observation without the aid of a coherent analytical framework or 'organon' could yield little useful knowledge. Influenced especially by the views of Herbert Spencer, he placed great emphasis on the way human character was formed by the experiences of home, workplace, and leisure activities. This opened the possibility of gradual change in human character by the intergenerational transfer of acquired traits, a process more Lamarckian than Darwinian. On the other hand, he strongly resisted the idea that human nature could be transformed by drastic social changes. He attached great importance to the gradual improvement of character and abilities, especially in the lower ranks of society. His ideals here were perhaps parochially reflective of the middle class mores of his era but they were heartfelt. Economic progress was to him not an end in itself but a means to the improvement of mankind. Faced by a choice between the two he would perhaps have chosen human over economic gains, but he fortunately saw them as complementing and reinforcing each other. He declined to base economics on an artificial 'economic man' and sought to deal with human nature in the round, giving recognition to the constraints of duty and the striving for recognition among professional peers. While conceding that the economist as such has no special standing as adviser on normative matters, he did not let this restrain him and was indeed something of a preacher.

Theoretical Contributions

Marshall's mature contributions to the theory of value and distribution build upon his earlier treatments in *The Pure Theory of Domestic Values* (1879) and *The Economics of Industry* (1879 jointly with M. P. Marshall) and are to be found in books 3, 5, and 6 of all editions of *Principles* after the first.

Book 3 of *Principles* deals with consumers' demand. The focus is on the quantity of a particular good demanded by a consumer at a given price and the 'consumer's surplus' consequently obtained. This surplus arises from the fact that a consumer's demand price for an additional unit falls as successive units are acquired and equals purchase price only for the final or marginal unit bought. Marshall's derivation of an individual's demand curve, contingent on tastes, income, and other prices, and based on the assumption of diminishing marginal utility, provided a simple if limited foundation for a downward sloping market demand curve for a good, relating its price to aggregate quantity demanded. It remains a staple of elementary pedagogy, as does his related concept of demand elasticity. The consumer's surplus notion seems to have been an independent rediscovery of a neglected idea developed by the French engineer J. Dupuit in the 1840s. Marshall was well aware of the limitations of welfare arguments based on this individual-specific monetary measure of benefit. The individual's marginal utility of money must remain constant, and interpersonal aggregation requires strong assumptions. Nevertheless, the concept continues to be widely used in applied economics.

Book 5 of *Principles* brings together market demand and supply for a commodity, taking for granted the broader economic context – an approach known as 'partial equilibrium' analysis which is closely associated with Marshall although hardly original to him. The treatment of supply is restricted to the extreme cases of monopoly and thoroughgoing competition, although the latter is often more akin to modern theory's imperfect competition than to its perfect competition. (*Industry and Trade* (1919) was to probe interestingly into the intermediate case of trusts, cartels, and 'conditional monopolies' always threatened by potential entrants.) The treatment of monopoly does not advance significantly beyond Cournot, apart from a remarkable brief treatment of socially optimal pricing by a publicly owned monopoly. The treatment of supply in the competitive case is more notable. There are two distinct lines of innovation. 'Period analysis' recognizes that quantities of some productive inputs may not be alterable in a limited time so that the industry's supply conditions become contingent upon the length of the period to which the analysis is meant to apply. Moreover, payments to any type of input will be determined by the cost of securing its use to the industry only if there is enough time to vary freely the quantity employed. Otherwise, payment to such inputs has the character of a rent – termed a 'quasi rent' – determined by the product's price rather than helping to determine this price as a necessary element in production cost.

The other notable feature is the explicit consideration of scale economies in production. External economies are due to the growth of the size of the industry as a whole rather than the sizes of its constituent firms. They reflect osmosis of worker skills, increased opportunities for specialization, and so on. Although somewhat elusive, the idea of such economies remains significant in modern economics. Internal economies are due to a growth in firm size that permits more efficient large-scale production methods to be utilized. While often plausible, a persistent ability of larger firms to undercut smaller firms threatens the viability of atomistic competition.

Marshall's solution, implied in his 'representative firm' concept, was to combine market and organizational restrictions on rapid firm expansion with a life cycle for the firm due to a waning of entrepreneurial drive on the part of the founder's successors. Such an argument applies more to the Victorian family firm than to the modern joint stock company, which was to take the limelight only in *Industry and Trade* (1919). The modern 'representative agent' concept owes little to Marshall beyond the name.

Scale economies under competition permit the supply price at which any quantity of a good is supplied to the market to fall with quantity, just as the corresponding demand price does. This admits the possibility that there may be more than one market-equilibrium quantity at which supply price equals demand price. The so-called Marshallian adjustment process through which market equilibrium is attained is very much in the classical tradition deriving from A. Smith. Price adjusts to clear the market of whatever quantity is made available, thus settling at the demand price for that quantity. Then the quantity supplied increases (decreases) as the supply price at the initial quantity is below (above) this market price, so that super-normal (infra-normal) profits are being made. An equilibrium quantity is unstable if there is divergence from it.

The possibility that there may be more than one market equilibrium that is stable for small perturbations is sufficient to disprove a claim that a competitive equilibrium must maximize social welfare. Marshall was clear that the most that could be claimed was what came to be called Pareto optimality. But he showed in a remarkable argument that even this might not be true, demonstrating that a social gain may be obtained by taxing an industry whose supply price rises with quantity in order to subsidize an industry having falling supply price. His analysis of this and other welfare issues drew heavily on the consumer-surplus concept and foreshadowed much in the 'new welfare economics' emerging in the 1930s.

Partial equilibrium analysis was extended in book 5 of *Principles* to cover a group of closely interrelated markets, as exemplified by a set of goods jointly demanded or supplied. More general is the interaction between the markets for an output and its several inputs. Here Marshall introduced the idea of derived demand – the demand for inputs being derived from the demand for the output they jointly produce – fruitfully characterizing the determinants of the elasticity of derived demand for any one input. Even if the number of goods involved is large, the analysis of such multi-market cases remains partial equilibrium in nature since an economy-wide setting that is little influenced by what happens in the markets analyzed is still tacitly assumed.

Only in *Principles* book 6 did Marshall turn to a consideration of the entire economy and its complex mutual interactions between value, production, and income distribution. His approach was macroeconomic, centering on the concept of the 'national dividend' or national income. This was seen as the combined product of the various factors of production, each unit of which claimed a share equal to its marginal product. Increased quantity of a homogenous factor would lower its marginal product but raise the national dividend, while the absolute or relative shares accruing to the entire factor could rise or fall.

The 'law of substitution' in production and consumption (essentially cost minimization and utility maximization), together with the equality of prices to all-inclusive costs required for competitive equilibrium, ultimately settle equilibrium prices for all goods and factors: also the quantity of each good produced, together with the quantities of the various factors used to produce it. None of this is described with any formality. It remains more a general vision than a precise theoretical statement that might rival the general-equilibrium theory of L. Walras. What book 6 effectively does, however, is to integrate a neoclassical marginal-productivity approach to the demand for factors of production with a classical approach to their supply (extensively developed in *Principles* book 4). This results in a real cost explanation of factor price determination in the long period. If factor supplies were permanently fixed, then factor prices would be determined wholly by the derived demand for their services, and the cost of any particular use would only be the opportunity cost of potential earnings in the best alternative use. But if, as Marshall supposes, continual replacement flows are needed just to maintain factor supplies then factor prices must exactly remunerate the real cost of eliciting these replacement supplies while still satisfying the derived-demand or marginal-productivity condition.

For Marshall, capital, labor of all kinds including managerial labor, and even business organization when firms have finite lives, all require replacement flows to maintain their levels. The case of labor is particularly complicated in the long period. Parents often make decisions for their children about occupation and acquisition of human capital, while the expected return to investment in human capital provides inadequate surety for borrowing to undertake it, being both uncertain and incapable of ownership transfer. A further complication is the 'economy of high wages' which results when increased wages improve worker efficiency by boosting physical or mental well-being, a possibility to which Marshall attached considerable importance. Indeed, his treatment of labor markets is one of the most valuable features of *Principles* book 6.

Apart from those embodied in *Principles*, Marshall's theoretical contributions are limited. His foreign trade analysis (1879) with its ingenious offer curves was indeed an important advance on Mill's treatment, but highly restricted circulation limited its impact. On money, he had a masterly grasp of British monetary tradition but added little that was novel. His most significant innovation was to replace the velocity of circulation with a demand for money function, a change which already appeared in an early manuscript and was to remain influential in the Cambridge monetary tradition leading up to Keynes's *General Theory* (1936).

Policy Views

Marshall was a proponent of economic freedom and free trade but not a doctrinaire one or an advocate of extreme *laissez faire*. He saw government as having important functions in regulating market forces and remedying or alleviating the problems they produced. Redistributive taxation was acceptable, even desirable, if it did not threaten the security of property or incentives to effort. For the well-to-do striving to keep up with their neighbors, or for professionals seeking the acclaim of their peers, Marshall saw absolute incomes as largely symbolic and felt that taxation which preserved income relativities in the middle and upper

classes would not have serious disincentive effects on effort, saving, or risk-taking. Town planning and provision of recreational facilities were needed for the well-being of the urban working classes, the young especially, while a government safety net was needed for those in abject poverty, although private charity could be more discriminating in the incentives it offered and in distinguishing the deserving from the undeserving. To 'administrative socialism' – proposals that government take over the active management of most industries – he objected violently, however. He foresaw that it would result in widespread bureaucratic lethargy and an undermining of the freedom to experiment and innovate essential to economic progress and society's future well-being. Increasingly he came to feel that the UK's future prosperity must rest on the chivalric competition of those in control of private business, freed from intrusive government regulation and from enervating protection against foreign competition. He took relatively lightly the possibility that cartels, combines and monopolies might come to dominate, viewing them as continually undermined by economic change. His early hopes that involvement in trades unions and cooperative societies would induce among workers an appreciation of and sympathy for the complexities of business management waned and he came to fear organized labor as an ossifying factor second only to the heavy hand of government and itself a distinct threat to the UK's international competitiveness. But, after the sobering experience of war and its expanded government control of the economy, *Industry and Trade* (1919) was to place more reliance on conscious industry-wide collaboration between management, labor, and government.

Education had a vital role in raising the working classes by improving abilities, skills, and ambitions, as workers, parents, or citizens. Schools and technical colleges, provided or sponsored by the government, were crucial, but Marshall stressed that much practical education occurred in the workplace: also in the home, where the duties of mothers and daughters were urged and extolled.

On monetary arrangements, Marshall's preference was for a managed paper currency operating within an international gold-exchange standard. His proposal in the 1880s for 'symmetalism' – use of a fixed-quantity combination of gold and silver as the monetary base – was only offered as a superior alternative to the clamor of the time for fixed-price-ratio bimetallism. His preferred goal for monetary policy was to maintain fixity of nominal input prices rather than output prices. Productivity increase would thus yield declining output prices. To palliate the effects of unforeseen price-level changes he proposed voluntary indexation of private contracts by means of a price index or 'tabular standard' published by the government.

Place in Economic Thought

Marshall can claim some subjective, but limited objective, originality as a contributor to the opening round of the so-called neoclassical revolution, initiated in the 1870s by the writings of Jevons, Walras, and C. Menger, although quite fully anticipated in the then largely forgotten writings of Cournot, Dupuit, H. H. Gossen, and others. He has stronger claims to have been an important contributor to the second round, culminating only in the 1890s, which extended marginalism from value to distribution by means of the marginal-productivity concept. Here, he claimed to have been influenced early on by the work of J. H. von Thunen, although the evidence for this is meager. But the lasting impact of Marshall arises less from transforming ideas than from the provision of serviceable tools which still remain in everyday use: partial-equilibrium analysis, consumer's surplus, demand elasticity, derived demand, long-period versus short- period analysis, and so on. These ideas remain prominent in applied work and pedagogy even if they no longer feature prominently at the frontiers of theoretical research. They perhaps serve as mute testimony in defense of Marshall's distrust of extended abstract theorizing and his preference for being roughly relevant rather than precisely and elegantly irrelevant.

But Marshall's impact on the development of economics came as much, if not more, from his influence on his students and his role in the professionalization, or perhaps better academicization, of economics in the UK as from the impact of his ideas. His *Principles* (1890) was prominent for 30 years or more in the training of economists in English speaking countries, while his students took over many of the new economic posts arising in the UK as higher education expanded, or else became prominent in government service. His struggles to establish economics more firmly in Cambridge bore later fruit in the influential 'Cambridge School,' although his intellectual hold on it had greatly weakened by the 1930s. Anxious that economists should present a public facade of unity and cumulative progress, he was led to present neoclassical ideas as more evolutionary than revolutionary and to take a conciliatory tone in methodological disputes, especially in dealing with the claims of the German historical school. While not an active promoter of, or participant in, professional organizations of economists, he placed his ecumenical imprint on the British Economic Association, founded in 1890 and subsequently renamed the Royal Economic Society.

Equilibrium analysis based on fixed conditions and fixed human response-patterns was only a first step for Marshall, the beginning rather than the end of economic knowledge. His evolutionist leanings led him to decry such analysis as merely mechanical and to look for approaches that were more biological in nature and which recognized the mutability and path-dependence of human character and institutions. His interest in such approaches was not to be much developed and remains somewhat enigmatic. Of particular interest in this connection is his analysis of the intertwined emergence of free enterprise and economic thought moved from *Principles* book 1 to *Principles* Appendices A and B after edition 4. Modern mainstream economics, although considerably indebted to Marshall, has hardly adopted his broader agenda, which indeed seems to have had little influence on the work of his students. Today's economic heterodoxy thus can often find both target and support in Marshall's work.

Further Reading

Keynes's biographical memoir (in Pigou, 1925) is not to be missed, and there is now a scholarly full-scale biography (Groenewegen, 1995) and a comprehensive edition of

correspondence and related documents (Whitaker, 1996). There is no satisfactory monograph on Marshall's economic thought but Tullberg (1990) and Whitaker (1990) provide assessments by various authors on the centenary of *Principles*. Wood (1982, 1996) provides a convenient but undiscriminating reprint in eight volumes of pertinent articles from academic journals.

See also: Agriculture, Economics of; Cost–Benefit Analysis; Economics, History of; Income Distribution and Inequality: Measurement Issues; Industrialization; Market Structure and Performance; Mill, John Stuart (1806–73); Monetary Policy Since the 2007–2009 Financial Crisis; Regulation, Economic Theory of; State and Local Taxation; Wage Differential and Structure.

Bibliography

Cournot, A.A., 1838. Recherches sur les Principes Mathématiques de la Théorie des Richesses. Hachette, Paris [1897 Researches into the Mathematical Principles of the Theory of Wealth. Macmillan, New York].

Groenewegen, P., 1995. A Soaring Eagle: Alfred Marshall 1842–1924. Edward Elgar, Aldershot, UK.

Groenewegen, P. (Ed.), 1996. Official Papers of Alfred Marshall; A Supplement. Cambridge University Press, Cambridge, UK.

Guillebaud, C.W. (Ed.), 1961. Alfred Marshall's Principles of Economics, ninth (Variorum) ed. Macmillan, London.

Jevons, W.S., 1871. The Theory of Political Economy. Macmillan, London.

Keynes, J.M. (Ed.), 1926. Official Papers of Alfred Marshall. Macmillan, London.

Keynes, J.M., 1936. The General Theory of Employment Interest and Money. Macmillan, London.

Marshall, A., 1879. The Pure Theory of Foreign Trade, The Pure Theory of Domestic Values. Privately printed [reproduced in Whitaker 1975].

Marshall, A., 1890. Principles of Economics, vol. 1. Macmillan, London.

Marshall, A., 1892. Elements of Economics of Industry. Macmillan, London.

Marshall, A., 1919. Industry and Trade. Macmillan, London.

Marshall, A., 1923. Money Credit and Commerce. Macmillan, London.

Marshall, A., Marshall, M.P., 1879. The Economics of Industry. Macmillan, London.

Mill, J.S., 1848. Principles of Political Economy. Parker, London.

Pigou, A.C. (Ed.), 1925. Memorials of Alfred Marshall. Macmillan, London.

Tullberg, R.McW. (Ed.), 1990. Alfred Marshall in Retrospect. Edward Elgar, Aldershot, UK.

Whitaker, J.K. (Ed.), 1975. The Early Economic Writings of Alfred Marshall. Macmillan, London, pp. 1867–1890.

Whitaker, J.K. (Ed.), 1990. Centenary Essays on Alfred Marshall. Cambridge University Press, Cambridge, UK.

Whitaker, J.K. (Ed.), 1996. The Correspondence of Alfred Marshall, Economist. Cambridge University Press, Cambridge, UK.

Wood, J.C. (Ed.), 1982. Alfred Marshall: Critical Assessments. Croom Helm, London.

Wood, J.C., 1996. Alfred Marshall: Critical Assessments, Second Series. Routledge, London.

Marshall, Thomas Humphrey (1893–1981)

Yasemin N Soysal, University of Essex, Colchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

T.H. Marshall's name is ubiquitous in citizenship scholarship. This recount of T.H. Marshall's accomplishments moves from a brief acknowledgment of him as an academic and citizen to his sociology of citizenship and its students and critics, and ends with a commentary on citizenship today and Marshall's influence on its theory. It is the tension between heightened expectations about the transforming capacity of citizenship and continuing inequalities in society that makes Marshall's work still relevant at the beginning of the twenty-first century.

As one of the most prominent British social and political theorists in the twentieth century, T.H. Marshall's work has significantly shaped the literature on citizenship. The never fading relevance of T.H. Marshall's work is evident in the vast literature on citizenship, written in the last three decades, either in response to Marshall's own writings or as a critique of his work. With each reimagination of citizenship, we are left with no choice but return to Marshall's original formulations.

This recount of T.H. Marshall's accomplishments moves from a brief acknowledgment of him as an academic and citizen to his sociology of citizenship and its students and critics, and ends with a commentary on citizenship today and Marshall's influence on its theory.

T.H. Marshall as an Academic and Citizen

If not disrupted by an imprisonment in a German civilian internment camp during World War I, T.H. Marshall would have followed the familiar path of English upper-middle class career-pattern – from boarding school to Oxbridge, with a likely career in the Foreign Service. Instead, equipped with a stronger 'social awareness and commitment,' Marshall opted out for sociology at the London School of Economics (LSE), where he had a distinctive career and provided a 'distinguished role model for aspirant sociologists' (Halsey, 1996; Smith, 1996). Marshall taught comparative social institutions at the LSE, with a firm interest in sociological explanations of social change and sociology's potential in creating that social change (Lockwood, 1974). It is this commitment to social issues and change, both as an academic and as a public intellectual, David Lockwood argued, made Marshall's work a par with those classical texts that paved the way for modern sociology. This commitment is evident in Marshall's combination of academic life with public service. He worked in the Foreign Office Research Department during World War II; served as educational adviser to the British High Commissioner in Germany from 1949 to 1950; and took the Directorship of Social Sciences in UNESCO when he retired from the LSE in 1956.

To accentuate Marshall's lifelong identification with sociology is no overstatement. Contemplating on his 4 years of imprisonment in Germany in an autobiographical piece, he recalls that in most of his letters to home he wrote evidence of 'a growing sociological curiosity about what was happening' to him and around. Marshall goes on to say that "if only [he] had been a sociologist then, what an opportunity this would have given [him] to study, by participant observation and any other survey method, the emergence of a structured society from a miscellaneous individuals by chance in a confined space, and its subsequent fortunes" (Marshall, 1973: 89). It is as if the sociologist in him made intelligible, and bearable, 'the world around' him and 'the remarkable social experiment' he had endured for 4 years. Marshall left the prison 'deeply affected' by the experience and with a strong resolve to start an academic career.

As a sociologist, Marshall contributed significantly to the study of social stratification and social policy. His seminal piece on citizenship, *Citizenship and social class*, established the many themes that have occupied our scholarly and intellectual agendas since then and for years to come, never diminishing in their import and urgency (the essay was first given as a lecture in 1949 at the University of Cambridge and then published in 1950). His collected essays *Sociology at the Crossroads* (first published in 1963; the American edition with a foreword by S.M. Lipset appeared under the title *Class, citizenship and social development* in 1964) and his text *Social policy* (1965) are still required readings in many countries and many languages. In the 1990s, Marshall's ideas have come to fore, not only in the excess academic literature on citizenship, but also in politics as A.H. Halsey (1996) asserts, with the transformation of the social democratic welfare state away from its 'socialistic' strands to a position of 'welfare capitalism.'

His was half-a-century of remarkable achievement in writing on and advocating for just citizenship and policy (Smith, 1996). In this lies his resonance in the advent of an age anticipating tumultuous shifts in the order of citizenship as we know it.

T.H. Marshall's Sociology of Citizenship

Marshall's was an evolutionary sociology grounded in economic history. *Citizenship and social class* combines elemental aspects of both of these disciplines effectively. His interest in equalities and inequalities, and thus, citizenship was shaped through his study of postfeudal England as a PhD student in Cambridge, as well as his day-to-day experiences with 'class' during his interment in Germany and during his brief encounter with politics as a Labor candidate in 1922.

Marshall's work on citizenship is an attempt at 'grand sociology.' It attempts to trace a 'developmental process' – the process of the development of citizenship rights. Marshall did this based on the reading of the British case, but nevertheless with the aim of chartering a more general map for the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61087-5

progressive development of rights. Each citizenship right built upon each other and labored the next one. The development of civil, political, and social rights not only corresponded to the eighteenth, nineteenth, and twentieth centuries in a sequential order, but also each one was the necessary condition for the other. Of this account of advancement of citizenship, Anthony Rees writes: "… [In the 1950s] Marshall almost single-handedly revived the notion of citizenship, and disseminated a particular view of it so successfully that it came to be seen (at least in England) as the only possible account" (1996: 3).

Marshall's interest in explaining citizenship rights has a purpose. It stems from his conceptual and political concerns for class and class inequalities. Marshall concedes that citizenship rights, in their earlier stages, were necessary for the flourish and maintenance of capitalism, a system of inequality in itself. The civil rights of individuals, particularly the right 'to own property and to conclude valid contracts,' were the necessary foundations for a market economy. But it is the inequalities generated by the capitalist market that the subsequent citizenship rights were contrived to alleviate. Here enter social rights as corrective and as a unifying force.

For Marshall, social rights constitute the 'inevitable capstone' of citizenship development. Substantive social entitlements and rights guaranteed by the welfare state should prevent social and economic exclusions that the earlier provisions of civil and political rights could not. This redefinition of citizenship, from a minimum of legal and political rights to a substantive body of social entitlements, is what made Marshall's lasting intellectual contribution (Somers, 1993; Rees, 1996).

And it is this conception of modern citizenship that has been a major inspiration and challenge to those whose intellectual concerns lie in democracy, equality, and social conflict. On both sides of the Atlantic, the difficult relationship between citizenship and social class dominated the scholarly agendas of 1950s and 1960s (Reinhard Bendix, Ralph Dahrendorf, A.H. Halsey, Seymour Martin Lipset, David Lockwood, and Peter Townsend). The question that underlies this early literature, whether the equalizing principles of citizenship can, and do, eliminate the inequalities of social class and incorporate the excluded into the national collectivity as equal members, still occupies a prime place in the sociological imagination, if only asked differently to include other excluded bodies of citizens (women, and diverse contingents of cultural, sexual, racial, and ethnic minorities).

Marshall saw modern citizenship as a product of a larger societal transformation through which "[d]ifferential status, associated with class, function and family was replaced by the single uniform status of citizenship, which provided the foundation of equality on which the structure of inequality could be built" (1964: 87–88). It is this tension between the expected transforming capacity of citizenship and the continuing inequalities of the modern society that makes Marshall's work still relevant.

T.H. Marshall and Citizenship Today

The concept of citizenship has lived a major intellectual and political renaissance in the aftermath of the political upheavals of the 1990s, mainly in response to the rebirth of new nation-states and the restructuring of the relationship between states and their citizens. These developments have laid bare the differentials in membership and right-bearing, and exposed the limits of citizenship. The theoretical repository set by Marshall's oeuvre has been the point of departure for the new bourgeoning interest in citizenship. At the current state of citizenship scholarship, it is unavailing to expect a publication that does not have a reference to Marshall's *Citizenship and social class*. Whether critically appraised or endorsed, what is striking is how much the recent conceptualizations of citizenship still owe to that of Marshall's.

Three main lines of criticism warrant consideration if only to reveal the influence of Marshall's work on the current undertakings on citizenship. Not so surprisingly, the most scrutinized aspect of Marshall's sociology of citizenship is its evolutionary character. Marshall's periodization of the rise of citizenship has come under attack for not allowing for alternative paths of progression. Moving from different analytical concerns and focusing on different empirical cases, research in historical sociology provides comparative contrast to Marshall's sequence (Turner, 1986; Mann, 1987; Barbalet, 1988; Somers, 1993). In this body of scholarship, attention to the existing political regimes, the strategies of the ruling elites and the ruled classes, and the varying institutional arrangements and patterns of associational and participatory forms, at both local and national levels, sanctions the redefinition of parameters and trajectory of citizenship. The path to citizenship, as it is evinced by these critical inquiries, is nowhere single and predetermined. Particularly revealing is the movement and importance that social rights displayed in the citizenship schemes of other countries. Bismarckian social reforms, for example, were introduced as substitute for full political citizenship (Mann, 1987). In Germany, not only the social rights preceded political rights, but also the imperative relationship between the two sets of rights that Marshall postulated based on the British case was not there.

Even the British case may not fit neatly with Marshall's periodization, it is argued. The empirical reanalysis of the English citizenship practices in the seventeenth and eighteenth centuries reveals that the emergence of national labor markets and development of civil rights happened in the company of extensive 'claims to and practices of political and social citizenship.' These claims and practices were then, under right conditions, transformed into explicit rights for English working class communities (Somers, 1993). Said otherwise, political and social citizenship materialized in varying intensity but significantly before the time frame predicted by Marshall's scheme.

Here, a second line of criticism to Marshall surfaces; the lack of agency and struggle in Marshall's explanatory framework. His explanation of citizenship rights relies on the dynamic tension between citizenship and social class in the framework of capitalist market development. Marshall's stages of citizenship progress with the formative demands of each emerging social class, but this class agency remains implicit in the main (Somers, 1993). In much of the scholarship that follows Marshall's steps class agency, or agency as such, does not get a proper treatment either. In this literature, citizenship is conceptualized as a personal status or legal category, which

guarantees equal rights and exact obligations for individuals in national polities. Inasmuch as the scholarship on citizenship moves away from this legalistic/formalistic take on rights and membership, they reveal a much more dynamic understanding of the citizenship development. The active agency is then reconstituted in the analysis of the evolution and elaboration of citizenship as a complex set of rights, statuses, and practices.

An approach which privileges the agency and participatory practices, for instance, effectively unearths women's place in the realm of citizenship, a social group that does not have visibility in the Marshallian framework. In the United States, despite being denied the franchise, women groups were able to mobilize successfully for mother's pensions and other legislation in the Progressive era (Skocpol, 1992). This finding not only reverses Marshall's sequential analysis, but also remedies the accounts of citizenship, which omit the formative role of women and their movements in the historic progress of citizenship.

Rights formation is not only a result of contradictions between the egalitarian principles of citizenship and capitalist society, as Marshall's account articulates (Barbalet, 1988). Citizenship is actively created by the participation of people in interaction with formal institutions and power structures (Turner, 1986, 1990). With this revision, that is, the expansion of Marshallian model to include 'citizenship from below,' the citizen enters the public space as an actor, 'an active bearer of claims,' and partakes in popular struggles to acquire rights. The resulting model thus resolutely incorporates the agency and mobilization of societal groups into the developmental framework cultivated by Marshall and his exponents.

The third line of criticism takes Marshall's framework to test at the end of the millennium. Marshall wrote his account of citizenship when national character of the state and welfare regimes was taken for granted. Even though Marshall was not so much attentive to the formation of the 'national' content of citizenship, nevertheless, he was aware that "the citizenship whose history [he wished to trace was], by definition, national" (1964: 72). The extension of socioeconomic rights to the previously excluded groups of people, such as the working class, facilitates their incorporation into the 'common culture' of the national collective. Marshall does not elaborate on this collective, neither does he take it to task to define what holds this collective together and what defines the boundaries of this collective.

A necessary corrective to the study of citizenship, brought to fore through empirical and theoretical treatments, has been to redefine citizenship from simply a legal status to a body of membership forms and practices (Brubaker, 1992; Soysal, 1994; Kymlicka, 1995). This shift in focus has brought with it a heightened and productive debate, with attention to historical and contemporary processes of boundary making, and the sources of inclusion and exclusion. As revealed by this debate, citizenship, which has evolved to encompass 'national' collectivities, variously following distinct paths and forming peculiar traditions, is no longer singularly located within national boundaries. Neither the distribution of membership rights relies on 'national' belonging as the indispensable criterion. Nor the citizenship practices are delimited by the boundaries of the nation-state within which the citizens reside. This *fin de siècle* reconstitution of citizenship implies a multiplicity of membership. In the new terrain of citizenship, different categories of members are accorded with differentiated set of rights – thus breaching the principle of uniform citizenship rights and accentuating disparities between members. Also in the new citizenship claims to rights and citizenship expand beyond the conventional modalities of political, civil, and social to embrace cultural, sexual, ecological, and global citizenships. The claims are increasingly advanced, and legitimated, by appeals to group differences but on the basis of universalistic human rights, as codified in transnational treaties and conventions. And the collectives and individuals organize and mobilize their claims not only at local and national but also transnational levels, targeting a variety of governing bodies (such as the European Court of Human Rights).

More recently, the manifest predominance of human rights frameworks but also their faltering reach has renewed questions about the relationship between rights and social justice – a long-standing concern in Marshall's scholarship. The debates in the newly invigorated sociology of rights revisit the relationship between citizenship and human rights, forcing to the forefront such issues as recognition, solidarity, and substantive enjoyment of rights (for various treatments see Somers, 2008; Nash, 2009; Walby, 2009; Morris, 2010). Much of these debates treat the relationship between citizenship and human rights contentious. Often the universal and natural character of (globally sanctioned) human rights is juxtaposed with the particularistic and political nature of (nation-state based) citizenship. The potential of human rights for furthering effective justice is questioned, on the grounds that they lack institutionalized enforcement mechanisms beyond the nation-state itself and they lack the solidarity to which national citizenship historically laid claim. While such juxtapositions inevitably warrant paying closer attention to the entangled practice of human rights and citizenship in the postwar period, at the same time, they underscore the broader analytical question of 'what rights do' (Soysal, 2012).

This is where T.H. Marshall reenters into our sociological conversations. For Marshall, rights in themselves did not have any inherent quality. Rights, he argued, are means for inclusion, which could be guaranteed only with their grounding in effective membership institutions – in other words the very social and institutional arrangements in which the projected citizenship of persons is recognized and the very possibility of a right exists. It is exactly these arrangements that are under threat today, laying bare the vulnerabilities of individuals, citizens, and noncitizens alike. Drawing on T.H. Marshall, Somers and Roberts (2008: 388) insightfully remind us that rights exist at 'multiple registers': as normative aspirations, codification and doctrine, and mechanisms and institutions of implementation. Human rights chart along all three of these registers; so do citizenship rights. Thinking along these three registers, we are able see that even 'the most formalized rights-bearing subject is incomplete' and citizenship is an 'incompletely theorized contract' between the state and the individual (Sassen, 2006: chapter 6). The increasing normative and cognitive authority accorded to human rights in national and transnational conventions in turn exposes the limits and 'incompleteness' of citizenship

projects, revealing them as contingent accomplishments, and opening up possibilities for expansion. It is this dynamic tension between the normative standards and practice of rights that gives much significance to the institution of citizenship and its potential for justice today, as ever before.

As his critiques have pointed out, the developments at the beginning of the twenty-first century project a significantly different topography of citizenship than the one analyzed and conceptualized by Marshall. As citizenship, and its associated rights, expands and differentiates within and without the nation-state, one may be impatient to register the infirmities and imperfections of Marshall's account. One may promptly find him guilty of not providing all the right answers. However, at the end, an eventual return to Marshall is only inevitable and fitting. His grand narrative of citizenship paves the way for lucid analyses of the new terrains of citizenship that confront us. Through conversations and argumentation with Marshall, the formations of new citizenship become manifestly tangible and substantive. And ultimately it is his preoccupation as a scholar and citizen that motivates our continuing interest in citizenship – that inequalities, independent of their sources and forms, are still part of our scholarly and societal agendas, and still take us, the sociologists, to task.

See also: Citizenship, Historical Development of; Social Inequality in History (Stratification and Classes); Welfare State, History of.

Bibliogarphy

Barbalet, J.M., 1988. Citizenship. Open University Press, Milton Keynes.

Brubaker, R., 1992. Citizenship and Nationhood in France and Germany. Harvard University Press, Cambridge.

Halsey, A.H., 1996. T.H. Marshall and ethical socialism. In: Bulmer, M., Rees, A.M. (Eds.), Citizenship Today: The Contemporary Relevance of T.H. Marshall. UCL Press, London.

Kymlicka, W., 1995. Multicultural Citizenship: A Liberal Theory of Minority Rights. Oxford University Press, Oxford.

Lockwood, D., 1974. For T.H. Marshall. Sociology 8, 363–367.

Mann, M., 1987. Ruling class strategies and citizenship. Sociology 21, 339–354.

Marshall, T.H., 1950. Citizenship and Social Class and Other Essays. Cambridge University Press, Cambridge.

Marshall, T.H., 1963. Sociology at the Crossroads and Other Essays. Heinemann, London.

Marshall, T.H., 1964. Class, Citizenship and Social Development. Doubleday, Garden City, New York [with an introduction by S.M. Lipset].

Marshall, T.H., 1965. Social Policy. Hutchinson, London [3rd revised edition appeared in 1970].

Marshall, T.H., 1973. A British sociological career. International Social Science Journal 25, 88–100.

Morris, L., 2010. Asylum, Welfare and the Cosmopolitan Ideal: A Sociology of Rights. Routledge Cavendish, London.

Nash, K., 2009. Between citizenship and human rights. Sociology 43, 1067–1083.

Rees, A., 1996. T.H. Marshall and the progress of citizenship. In: Bulmer, M., Rees, A.M. (Eds.), Citizenship Today: The Contemporary Relevance of T.H. Marshall. UCL Press, London.

Sassen, S., 2006. Territory, Authority, Rights: From Medieval to Global Assemblages. Princeton University Press, Princeton.

Skocpol, T., 1992. Protecting Soldiers and Mothers: The Political Origins of Social Policy in the United States. Harvard University Press, Cambridge.

Smith, J.H., 1996. Forward. In: Bulmer, M., Rees, A.M. (Eds.), Citizenship Today: The Contemporary Relevance of T.H. Marshall. UCL Press, London.

Somers, M., 1993. Citizenship and the place of the public sphere: law, community, and political culture in the transition to democracy. American Sociological Review 58, 587–620.

Somers, M., 2008. Genealogies of Citizenship: Markets, Statelessness, and the Right to Have Rights. Cambridge University Press, Cambridge.

Somers, M., Roberts, C.N.J., 2008. Toward a new sociology of rights: a genealogy of "buried bodies" of citizenship and human rights. Annual Review of Law and Science 4, 385–425.

Soysal, Nuhoğlu Y., 1994. Limits of Citizenship: Migrants and Postnational Membership in Europe. Chicago University Press, Chicago.

Soysal, Nuhoğlu Y., 2012. Citizenship, immigration, and the European social project: rights and obligations of individuality. The British Journal of Sociology 63, 1–21.

Turner, B.S., 1986. Citizenship and Capitalism: The Debate over Reformism. Allen & Unwin, London.

Turner, B.S., 1990. Outline of a theory of citizenship. Sociology 24, 189–217.

Walby, S., 2009. Globalization and Inequalities Complexity and Contested Modernities. Sage, London.

Martingales

René L Schilling, Institut für Mathematische Stochastik, Technische Universität Dresden, Dresden, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A martingale is the mathematical description of a fair game: The expected net gain or loss from further play, independent of the history, is 0. The rigorous mathematical definition involves conditional expectation. Following earlier work by Paul Lévy and Jean Ville, Joseph Leo Doob developed in the 1940s and 1950s basic martingale theory and many of its applications. Martingales allowed one to study, for the first time, the behavior of sums and sequences of random variables which are not independent. Martingale theory is one of the cornerstones of modern mathematical probability theory with wide-ranging applications in stochastic analysis and mathematical finance.

The notion of *martingale* is intimately connected with that of a *fair game* where the player and the bank have equal chances of winning. In the eighteenth and nineteenth centuries, a martingale meant a gambling strategy; in its simplest form a player doubles his stake if he lost in the previous round, and stops playing otherwise. Under the assumption that the player can bet indefinitely high stakes, the losses are given by a geometric progression and, after possibly arbitrarily high losses, there is a guaranteed net win of the *initial* stake. This is equivalent to the so-called *St. Petersburg paradox* (Gorroochurn, 2012). Usually this strategy fails due to limited wealth of the player, time constraints, and limits on the stakes on the side of the casino: The mean loss (prior to winning) is infinite.

In mathematical terms, the fair game underlying a simple martingale is a sum of independent, identically distributed Bernoulli random variables $Y_1, Y_2, \ldots$ such that $\mathbb{P}(Y_n = -1) = \mathbb{P}(Y_n = 1) = \frac{1}{2}$ (e.g., representing the nth toss of a fair coin); $Y_n = 1$ returns twice your stake, while $Y_n = -1$ means you lose all. Denoting by $e_1 = 1$ and $e_n = e_n(Y_1, \ldots, Y_{n-1})$, $n = 2, 3, \ldots$, the stake for the nth game, the total net winnings are $X_0 = 0$ and $X_n = e_1Y_1 + \ldots + e_nY_n$. One easily sees that the conditional expectation (Schilling, 2005; Schilling and Partzsch, 2014; Williams, 1991) satisfies

$$\mathbb{E}(X_{n+1}|Y_1, \ldots, Y_n) = X_n \qquad [1]$$

Since the events $(Y_i = \pm 1)$ have strictly positive probability, this abstract conditional expectation still has its classical meaning as

$$\mathbb{E}(X_{n+1}|Y_1 = y_1, \ldots, Y_n = y_n) = x_n$$

where the y_i, $i = 1, \ldots, n$, are the outcomes of the rounds $1, \ldots, n$, and $x_n = e_1y_1 + e_2(y_1)y_2 + \ldots + e_n(y_1, \ldots, y_{n-1})y_n$ is the net win after the nth round. Iterating eqn [1] yields

$$\mathbb{E}(X_{n+k}|Y_1, \ldots, Y_n) = X_n \quad k, n = 1, 2, \ldots \qquad [2]$$

The simple martingale strategy corresponds to the stakes $e_1(-1, -1, \ldots, -1) = 2^{n-1}$ and 0 in all other cases. In particular, the expected gains are

$$\mathbb{E}X_n = \mathbb{E}X_{n-1} = \ldots = \mathbb{E}X_0 = 0$$

Conditional expectation can be interpreted as the best mean-square predictor for the future win X_{n+k} conditional on the past outcomes $Y_1, \ldots, Y_n$. Since the stakes for the next round depend only on the past results, i.e., e_{n+1} is a function of $Y_1, \ldots, Y_n$, the stakes are previsible. The martingale property guarantees that the best prediction is always the current (time n) level of winnings. In this sense, a martingale describes a fair game.

It is not possible to change the character of a game by a previsible gambling strategy: A fair game will stay fair no matter which gambling strategy is being used. This follows immediately from the fact that in the discussion above we have made no further assumptions on the gambling strategy, i.e., the stakes $e_1, e_2, \ldots$, than previsibility. The same assertion is true for favorable [unfavorable] games that are submartingales [supermartingales].

Definition

The *general definition of a martingale* is as follows: Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, an index set I equipped with a partial order $s \le t$ and an increasing family of σ-algebras (a *filtration*) $\mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{F}$ if $s \le t$, then the real-valued random variables $(X_t)_{t\in I}$ are a *martingale* with respect to $(\mathcal{F}_t)_{t\in I}$, if each X_t is $\mathcal{F}_t$ measurable, $\mathbb{E}|X_t| < \infty$ for all $t \in I$, and

$$\mathbb{E}(X_t|\mathcal{F}_s) = X_s \quad \text{for all} \quad s, t \in I, s \le t \qquad [3]$$

if we replace '=' by '$\ge$' [or '$\le$'], we get a sub- [or super-] martingale.

The expectation $m(t) = \mathbb{E}X_t$ of a martingale is constant, while those of a sub-/supermartingale are increasing/decreasing functions of t.

A *downward*, *backwards* or *reversed martingale* is a martingale where the family of σ-algebras $(\mathcal{F}_t)_{t\in I}$ decreases as t increases. This means that $(X_{-t}, \mathcal{F}_{-t})_{t\in -I}$ is a martingale (with the natural order of $-I$) and all martingale results hold for backwards martingales. Therefore, there is no need to distinguish between martingales and backwards martingales. If I has a smallest element, i_0, then $-I$ has a largest element and the backwards martingale is right-closed, i.e., $X_{-t} = \mathbb{E}(X_{i_0}|\mathcal{F}_{-t})$ for all $t \in -I$. Backwards martingales were first studied in Doob (1953).

The notion of a (sub-/super-) martingale essentially depends on the given stochastic basis $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\ge 0}, \mathbb{P})$. If $(X_t)_{t\in I}$ is a martingale for the filtration $(\mathcal{F}_t)_{t\in I}$, it is also a martingale for the minimal σ-algebras $\mathcal{F}_t^X = \sigma(X_s : s \le t, s \in I)$

which is, by definition, the smallest σ-algebra such that all X_s, $s \le t$, are measurable. A further reduction is not possible.

On the other hand, enlarging the filtration usually results in the loss of the martingale property: The σ-algebra $\mathcal{F}_t$ has a natural interpretation as 'the information until time t' and enlarging $\mathcal{F}_t$ means increasing the available information (e.g., because of insider knowledge or cheating) which immediately leads to a nonfair situation.

In eqns [1] and [2] one uses the σ-algebra $\sigma(Y_1, \dots Y_n)$ which is generated by $Y_1, \dots Y_n$. Changing to the smaller σ-algebras $\sigma(X_1, \dots X_n)$, one would still get a martingale, since both $\sigma(Y_1, \dots Y_n)$ and $\sigma(X_1, \dots X_n)$ contain the essential information on the game. Indeed, X_n is a function of the $Y_1, \dots Y_n$, while

$$e_{n+1}(Y_1, \dots, Y_n)Y_{n+1} = X_{n+1} - X_n$$

this allows one to represent $Y_1, Y_2, \dots$ – on the sets where e_1, $e_2, \dots$ are strictly positive–recursively in terms of $X_1, X_2, \dots$.

For simplicity, we will assume that $I \subset [0, \infty]$. Integer indices are usually denoted by m or n, real indices by s, t.

Examples

In the examples $\mathcal{F}_t^X = \sigma(X_s : s \le t, s \in I)$ denotes the minimal filtration of the process $(X_t)_{t \in I}$.

1. (Sum of independent random variables) $X_n = Y_1 + \dots + Y_n$, $X_0 = 0$, where $(Y_n)_{n \in \mathbb{N}}$ are real-valued, independent, mean-zero random variables, is a martingale with filtration $\mathcal{F}_n^X = \mathcal{F}_n^Y$. If $\mathbb{E}Y_n \ge 0$ (≤ 0) for all n, then $(X_n)_{n \in \mathbb{N}_0}$ is a submartingale [supermartingale].
   If $\mathbb{E}(X_n^2) < \infty$ for all $n \in \mathbb{N}_0$, then $(X_n^2)_{n \in \mathbb{N}_0}$ is a submartingale and $(X_n^2 - \mathbb{E}(X_n^2))_{n \in \mathbb{N}_0}$ is a martingale with filtration $(\mathcal{F}_n^X)_{n \in \mathbb{N}_0}$.
2. (de Moivre's martingale) Let $(Y_n)_{n \in \mathbb{N}}$ be independent identically distributed random variables such that $\mathbb{P}(Y_n = 1) = p$ and $\mathbb{P}(Y_n = -1) = q = 1 - p$. Then $(X_n)_{n \in \mathbb{N}_0}$ with $X_0 := 1$ and $X_n := (q/p)^{Y_1 + \dots + Y_n}$ is a martingale for the filtration $(\mathcal{F}_n^Y)_{n \in \mathbb{N}}$.
3. (Wald martingale) Let $(Z_n)_{n \in \mathbb{N}}$ be nonnegative independent random variables with finite means $\mu_n = \mathbb{E}Z_n > 0$. Then $(X_n)_{n \in \mathbb{N}_0}$ with $X_0 := 1$ and $X_n := \prod_{i=1}^n \mu_i^{-1} Z_i$ is a martingale for the filtration $(\mathcal{F}_n^Z)_{n \in \mathbb{N}_0}$.
4. (Pólya's urn) An urn contains X_0 blue and Y_0 red balls. Draw randomly a ball from the urn. If a blue (resp. red) ball is drawn, replace the ball and add c more blue (resp. red) balls to the urn. Denote by X_n and Y_n the number of blue and red balls in the urn after the nth round. Then $(V_n)_{n \in \mathbb{N}_0}$ with $V_n = X_n/(X_n + Y_n)$ is a martingale for the filtration $\mathcal{F}_n := \sigma(X_i, Y_i : i \le n)$. (Pólya invented this scheme in order to model the spread of contagious diseases.)
5. Let $(X_t)_{t \ge 0}$ be a stochastic process with independent increments such that $m(t) = \mathbb{E}X_t$ exists. It is a martingale (with filtration$(\mathcal{F}_t^X)_{t \ge 0}$) if, and only if, $m(t)$ is constant; it is a submartingale if, and only if, $m(t)$ is increasing.
   The process $(X_t - m(t))_{t \ge 0}$ is a martingale. In particular, one-dimensional Brownian motion $(W_t)_{t \ge 0}$ and the compensated Poisson process $(N_t - \lambda t)_{t \ge 0}$ with intensity $\lambda > 0$ are martingales.
6. (Exponential martingale) Let $(X_t)_{t \ge 0}$ be a real-valued Lévy process (a process with independent and fixed increments which is continuous in probability). Then $(\exp(i\xi X_t)/\mathbb{E}e^{i\xi X_t})_{t \ge 0}$, $\xi \in \mathbb{R}$, is a martingale for the filtration $(\mathcal{F}_t^X)_{t \ge 0}$.
7. Let $(X_n)_{n \in \mathbb{N}_0}$ be a Markov chain with countable state space E and transition matrix $\mathrm{P} = (\mathrm{p}_{ik})_{i,k \in E}$. Then, for any $u : E \to \mathbb{R}$, $(u(X_n) - \sum_{m=0}^{n-1}(\mathrm{id} - P)u(X_m))_{n \in \mathbb{N}_0}$ is a martingale with filtration $(\mathcal{F}_n^X)_{n \in \mathbb{N}_0}$.
   If u is invariant for P, i.e., $Pu \equiv u$, then $(u(X_n))_{n \in \mathbb{N}_0}$ is again a martingale.
8. Let $(X_t)_{t \ge 0}$ be a Markov process with transition semigroup $(P_t)_{t \ge 0}$ and generator A. Then $(u(X_t) - \int_0^t Au(X_s)\mathrm{d}s)_{t \ge 0}$ is for any $u : \mathbb{R} \to \mathbb{R}$ from the domain of A a martingale with filtration $(\mathcal{F}_t^X)_{t \ge 0}$.
   If u is A-harmonic, i.e., $Au \equiv 0$, then $(u(X_t))_{t \ge 0}$ is again a martingale.
9. (Lévy martingale, closed martingale) Let $(\mathcal{F}_t)_{t \ge 0}$ be an increasing family of σ-algebras in $\mathcal{F}$ and X be a real-valued random variable with $\mathbb{E}|X| < \infty$. Then $(X_t)_{t \ge 0}$ where $X_t := \mathbb{E}(X|\mathcal{F}_t)$ is a uniformly integrable martingale. Since $X = X_\infty$ is the right endpoint, this martingale is often called a *(right-)closed* martingale.
10. Let $\mathbb{Q}$ be a further probability measure on $(\Omega, \mathcal{F})$, let $(\mathcal{F}_t)_{t \ge 0}$ be an increasing family of σ-algebras in $\mathcal{F}$ and denoted by $\mathbb{Q}_t$ and $\mathbb{P}_t$ the restrictions of the measures $\mathbb{Q}$ and $\mathbb{P}$ to $\mathcal{F}_t$. The *likelihood ratio* $L_t := \mathrm{d}\mathbb{Q}_t/\mathrm{d}\mathbb{P}_t, \quad t \ge 0$, is a nonnegative martingale.

Properties

(Schilling, 2005; Schilling and Partzsch, 2014; Revuz and Yor, 1999; Rogers and Williams, 1987/94; Williams, 1991; Chow and Teicher, 1997) (Sub-)martingales have good convergence properties: If $\sup_{t \in I}\mathbb{E}|X_t| < \infty$ and $t \to t_0 \in I$, then $X_t \to X_{t_0}$ almost surely; if $(X_t)_{t \in I}$ is a uniformly integrable (sub-)martingale, this convergence holds also in L^1-sense. In particular, all uniformly integrable martingales are of the form of Example 9 with $X = \lim_{t \to \infty} X_t$, and vice versa. A sufficient condition for uniform integrability is that $\sup_{t \in I}\mathbb{E}(|X_t|^p) < \infty$ for some $p > 1$. The convergence properties make martingale techniques powerful tools to get càdlàg (i.e., right-continuous with finite left limits) modifications of continuous-time stochastic processes. In the sequel, we will only consider càdlàg modifications.

If $(X_t)_{t \in I}$ is a martingale and $\phi : \mathbb{R} \to \mathbb{R}$ a convex function with $\mathbb{E}|\phi(X_t)| < \infty$ for all $t \in I$, then $(\phi(X_t))_{t \in I}$ is a submartingale. For (positive sub-)martingales we have *Doob's maximal inequalities*

$$\mathbb{E}\left(\sup_{t \in I} |X_t|^p\right) \le c_p \sup_{t \in I} \mathbb{E}\left(|X_t|^p\right)$$

with $c_p = (p/(p-1))^p$ and $1 < p < \infty$.

Martingale convergence theorems were the first convergence results for sums and sequences of random variables which are *not* independent. Even nowadays, martingale techniques are among the most successful methods to deal with dependent random variables. For example the Lindeberg–Feller central limit

theorem admits a martingale version: Let $F_0 := \{\emptyset, \Omega\}$, $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a filtration and $(X_n)_{n\in\mathbb{N}_0}$ be a sequence of L^2-martingale differences (i.e., $\mathbb{E}(X_n^2) < \infty$ and $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = 0$ for all $n \in \mathbb{N}_0$), set $\sigma_n^2 = \mathbb{E}(X_n^2)$ and $s_n^2 = \sigma_1^2 + \ldots + \sigma_n^2$. If $(X_n)_{n\in\mathbb{N}_0}$ satisfies the Lindeberg condition and

$$\sum_{i=1}^{n} \mathbb{E}\left|\mathbb{E}\left(X_i^2 \middle| \mathcal{F}_{i-1}\right) - \sigma_i^2\right| = o(s_n)$$

then $s_n^{-1}\sum_{i=1}^n X_i$ converges in distribution to a standard normal random variable. An analogous martingale version exists for the *Berry–Esseen theorem*.

A *stopping time* (also: Markov time or optional time) is a random time $\sigma : \Omega \to I \cup \{\infty\}$ that is adapted to a given filtration $(\mathcal{F}_t)_{t\in I}$ in the sense that $\{\omega : \sigma(\omega) \le t\} \in \mathcal{F}_t$ for all $t \in I$; this means that it is enough to know X_s for all $s \le t$, $s, t \in I$, in order to decide whether $\sigma \le t$ or not. If $(X_t)_{t\in I}$ is a uniformly integrable martingale and $\sigma \le \tau$ are stopping times with values in I, then eqn [3] still holds for random times $\sigma \le \tau < \infty$ (*optional stopping*); in particular, the stopped process $X_t^\tau := X_{\min(t,\tau)}$ is a martingale both for $(\mathcal{F}_t)_{t\in I}$ and $(\mathcal{F}_{\min(t,\tau)})_{t\in I}$ (*optional sampling*).

The fundamental relationship between martingales and submartingales is the *Doob–Meyer decomposition*: If $(X_t)_{t\ge 0}$ is a càdlàg submartingale of *class D* (i.e., the family $\{X_\tau : \tau$ finite stopping time$\}$ is uniformly integrable), then there exists a unique (up to indistinguishability), increasing, previsible (in a continuous-time setting *previsible* means that the process $(A_t)_{t\ge 0}$ is measurable with respect to the σ-algebra generated by all left-continuous adapted processes. In particular, any left-continuous process $(Y_t)_{t\ge 0}$ such that Y_t is $\mathcal{F}_t$ measurable is previsible) process $(A_t)_{t\ge 0}$, $A_0 = 0$, and a uniformly integrable martingale $(M_t)_{t\ge 0}$, $M_0 = 0$, such that $X_t = X_0 + A_t + M_t$. Without the assumption of *class D*, we still have a unique decomposition into an increasing previsible process A_t and a *local* martingale M_t. Local martingales are processes for which there is a sequence of stopping times τ_n converging almost surely to infinity as $n \to \infty$ and such that $(M_t^{\tau_n})_{t\ge 0}$ are, for each $n \in \mathbb{N}$, martingales.

Examples (7) and (8) point to the connection of martingales with potential theory. Recall that the Laplacian $\frac{1}{2}\Delta$ is the generator of Brownian motion $(W_t)_{t\ge 0}$. Thus, a (classical) harmonic function $u \in C^2(\mathbb{R}^n)$ satisfies $\Delta u \equiv 0$ and $(u(W_t))_{t\ge 0}$ is a martingale. Similarly a subharmonic function u produces a submartingale. The Doob–Meyer decomposition is the stochastic analogue of the Riesz decomposition (any superharmonic function can be written as the sum of a harmonic function and a potential) from potential theory.

Martingale theory and the Doob–Meyer decomposition were instrumental for the development of a general theory of stochastic (Itô) integrals. Let $(M_t)_{t\ge 0}$ be an L^2-martingale, i.e., $\mathbb{E}(|M_t|^2) < \infty$ for all $t \ge 0$. By the Doob–Meyer decomposition there is a unique increasing previsible process $\langle M\rangle_t$ such that $M_t^2 - \langle M\rangle_t$ is a martingale; $\langle M\rangle_t$ is called the (*previsible*) *compensator* or *angle bracket* of $(M_t)_{t\ge 0}$. If $t \mapsto M_t$ is almost surely continuous, then there is a simple formula for the angle bracket: Denoting by Π a partition $0 = t_0 < t_1 < \ldots < t_n = t$ of $[0, t]$ with mesh $|\Pi| = \max_{1\le i\le n}(t_i - t_{i-1})$, then

$$\langle M\rangle_t = \lim_{|\Pi|\to 0} \sum_{i=1}^{n} (M_{t_i} - M_{t_{i-1}})^2$$

(the limit is in the sense of local uniform convergence in probability). For a martingale with jumps $\Delta M_s = M_s - M_{s-}$, $M_{s-} = \lim_{r<s, r\to s} M_r$, and continuous part M_t^c one has

$$\sum_{s\le t} (\Delta M_s)^2 + \langle M^c\rangle_t = \lim_{|\Pi|\to 0} \sum_{i=1}^{n} (M_{t_i} - M_{t_{i-1}})^2$$

For instance, one-dimensional Brownian motion $(W_t)_{t\ge 0}$ satisfies $\langle W\rangle_t = t$, and for a Poisson process $(N_t)_{t\ge 0}$ of intensity λ and the associated martingale $M_t = N_t - \lambda t$ one has $\langle M\rangle_t = \lambda t$. In fact, *Lévy's martingale characterization of Brownian motion* says that a one-dimensional Brownian motion is the only (local) martingale in continuous time and with continuous paths such that $\langle W\rangle_t = t$ and $W_0 = 0$. A result by Doeblin, Dambis, and Dubins–Schwarz tells us that every L^2-martingale $(M_t)_{t\ge 0}$ with continuous paths can be embedded into a (suitable) Brownian motion by $M_t = M_0 + W_{\langle M\rangle_t}$.

In the guise of Itô's isometry (see below) quadratic variations can be used to define the stochastic integral driven by a martingale $(M_t)_{t\ge 0}$ with filtration $(\mathcal{F}_t)_{t\ge 0}$. A *simple* or *elementary process* $(Y_t)_{t\ge 0}$ is a bounded càdlàg process with finitely many values

$$Y_t(\omega) = \sum_{i=0}^{n-1} \eta_i(\omega) 1_{[t_i, t_{i+1})}(t), \quad t \ge 0, \omega \in \Omega$$

where $t_0 = 0 < t_1 < \ldots t_n <$ and η_i is a $\mathcal{F}_{t_i}$ measurable random variable. Then

$$\begin{aligned} Y\cdot M_t &= \int_0^t Y_{s-}\,\mathrm{d}M_s \\ &:= \sum_{i=0}^{n-1} \eta_i \left(M_{\min(t_{i+1},\, t)} - M_{\min(t_i,\, t)}\right) \end{aligned}$$

defines the Itô integral and $(Y\cdot M_t)_{t\ge 0}$ is an L^2-martingale; its compensator $(\langle Y\cdot M\rangle_t)_{t\ge 0}$ is given by $\int_0^t |Y_{s-}|^2 \mathrm{d}\langle M\rangle_s$. The compensator property yields the *Itô isometry*

$$\mathbb{E}\left\{\left(\int_0^t Y_{s-}\,\mathrm{d}M_s\right)^2\right\} = \mathbb{E}\int_0^t |Y_{s-}|^2 \mathrm{d}\langle M\rangle_s$$

which allows one to extend the stochastic integral by continuity to all càdlàg integrands such that Y_t is $\mathcal{F}_t$ measurable, $t \ge 0$, and $\mathbb{E}\int_0^T |Y_{t-}|^2 \mathrm{d}\langle M\rangle_t < \infty$.

The development of stochastic analysis gave rise to further generalizations of martingales: Local martingales and semimartingales (Revuz and Yor, 1999; Rogers and Williams, 1987/94). Martingales also turned out to be a powerful tool in characterizing and constructing Markov processes. The key role in these *martingale problem* approaches (Stroock and Varadhan, 1997; Ethier and Kurtz, 1986; Böttcher et al., 2013) is the observation that the martingales appearing in Examples 7 and 8 characterize the underlying Markov chains and processes.

Applications

Martingales have found many important applications also outside of probability theory.

One of the most prominent applications is in mathematical finance in the theory of *arbitrage* and *risk-neutral pricing* of

financial assets (Bingham and Kiesel, 1998). In a market, an arbitrage opportunity is a trading strategy that has a positive probability of winning and zero probability of losing money. It is a money-making machine, the ultimate unfair game. On the other hand, a martingale represents a fair game and will not allow for arbitrage. Since one can always invest in riskless interest-bearing bonds, it is clear that one has to discount prices by a numéraire, e.g., to divide asset prices by the price of the riskless bond. A *martingale measure* is a probability measure $\mathbb{Q}$ such that the discounted asset prices form a martingale. In a real market, no player (with their own beliefs, expressed by the player's own probability measure $\mathbb{P}$, sometimes called the real-world probability measure) expects the asset prices to be martingales since no one gains by betting on a martingale. On the other hand, in many situations there is an equivalent (with respect to $\mathbb{P}$) martingale measure $\mathbb{Q}$ such that the discounted asset prices become martingales under the conditional expectations induced by $\mathbb{Q}$. By *the fundamental theorem of asset pricing* the existence of a martingale measure guarantees that there is no arbitrage. Moreover, the market is complete (i.e., each derivative can be 'hedged' or 'replicated' by a trading strategy) if the martingale measure is unique. In a discrete-time setting the existence of an equivalent martingale measure is even equivalent to an arbitrage-free complete market. In the classical Black–Scholes model where the asset prices are modeled by a geometric Brownian motion, it is always possible to find a martingale measure $\mathbb{Q}$ with the help of Girsanov's theorem. In order to calculate the fair price of a derivative (e.g., a share option) one performs the following steps: First one determines the pay-off of the derivative, then one constructs an equivalent martingale measure $\mathbb{Q}$, finally the discounted price process of the derivative is the expectation (under $\mathbb{Q}$) of its payoff; since everything is, by construction, a martingale, no arbitrage is possible.

Within mathematics, martingales turn out to be useful in the study of almost everywhere convergence. Typical applications include the *geometry of Banach spaces* and *Littlewood–Paley theory*. An unconditional basis in a Banach space B is a basis $(e_n)_{n\in\mathbb{N}}$ such that every rearrangement of the series $\sum_n a_n e_n$ converges to the same limit $b \in B$. The spaces $L^p((0,\ 1),\ \mathrm{d}x)$ have the Haar functions as unconditional basis if, and only if, $1 < p < \infty$. This can be checked by martingale arguments involving the Burkholder–Davis–Gundy inequalities and so-called martingale difference sequences: $(e_n)_{n\in\mathbb{N}}$ is a martingale difference sequence if $\mathbb{E}(e_{n+1}|\sigma(e_1, \ldots, e_n)) = 0$. In fact, the Haar functions are an unconditional basis for every rearrangement-invariant function space which is isomorphic to a Banach space with an unconditional basis. This is the basis for many structure and embedding results for Banach spaces.

In the theory of vector measures, the L^1-martingale convergence theorem (for Banach space valued martingales) is equivalent to the Radon-Nikodým property of a Banach space B, i.e., every function f: $[0,\ 1] \to B$ of bounded variation is almost everywhere differentiable or, equivalently, every separable subspace of B has a separable dual.

If a function space has an orthonormal basis consisting of martingale differences, one can often use the pointwise martingale convergence theorem and maximal inequalities to obtain pointwise almost everywhere convergence results (Schilling, 2005: Chapter 24). In harmonic analysis (Stein, 1993, pp. 184–192) there are deep connections with (maximal) inequalities for the Littlewood–Paley functions and the Hilbert transform leading to proofs of Carleson's famous result that the Fourier series of a square-integrable function converges almost everywhere. In the theory of bounded mean oscillation (BMO) spaces, martingale maximal inequalities are most powerful tools. For example, one can prove the John–Nirenberg estimates for BMO spaces and Fefferman's theorem: BMO is the dual of the Hardy space H^1.

Historical Notes

Jean Ville (1936, 1939) introduced the term 'martingale' into probability. In his 1936 definition he considered only martingales of the form $X_n = e_1Y_1 + \ldots + e_nY_n$ (as in the second paragraph of this article) which he generalized in the monograph 1939 both in the time-discrete and time-continuous case. Ville used martingales to discuss *the concept of a collective, upon which many mathematicians found the theory of probability*. [...] *This leads* [...] *to a new definition of collective* (Doob, 1939). Ville's approach to von Mises' theory of collectives interprets the complete irregularity in a random sequence as the impossibility of a successful gambling system of 'martingale' type (in the classical sense of the first paragraph). In fact, von Mises (1932, pp. 157–158) already contains elements of this line of thought, and martingales are explicitly mentioned – however, as gambling systems only. The first general definition of martingales using conditional probabilities and the first ever martingale convergence theorem (*Lévy's 0–1 law*) can be found in Lévy (1937, pp. 122–123 (definition), p. 129, Theorème 41). Ville's 1939 definition follows (Lévy, 1937) and it is very close to the first 'modern' definition in Doob (1940) who calls eqn [3] *condition* ε (seemingly using Lévy's slightly obscure diction). It was mainly Doob who pioneered the theory and applications of martingales and his groundbreaking monograph (Doob, 1953) contains most of the basic results on martingales. Subsequently, major contributions to the theory, in particular in the connection with stochastic calculus, are mainly due to the French and Japanese schools of probability, see Meyer (1966), Dellacherie and Meyer (1980), Kunita and Watanabe (1967), and Ikeda and Watanabe (1981).

Etymology

According to the *Oxford English Dictionary* (OED) www.oed.com (Accessed online 20 July 2014.) the etymological origin of the word martingale is unclear. Most likely it derives from *martengala*, feminine of martegal/martengal, an inhabitant of Martigues (a small town in south-eastern France, département Bouches-du-Rhône). The word martingale has been used in the following senses:

1. *Horse riding*. A strap or harness preventing the horse from rearing or throwing its head back.
2. *Nautical*. Any part of a rigging strengthening the jib boom and bowsprit of a ship.

3. A dog collar made of two loops. When the animal tries to pull, the loop around the dog's neck contracts without the choking effect of a slip collar.
4. *Clothing*. (a) A loose half-belt or strap on the back of a garment. (b) A pair of trousers.
5. *Games*. A gambling system as described in the introduction to this article.
6. *Mathematics*. A stochastic process as described in this article.

The earliest use is in French as in 1 and 4(b). The OED states that "[t]he application may arise from a belief that the inhabitants of Martigues, a remote town, were eccentric and naive; hence also the application to an apparently foolish system of gambling. Sense two, however, is probably attributable to the former importance of Martigues as a port and ship-building center." An in-depth discussion on the origins of the word martingale can be found in Mansuy (2009).

See also: Diffusion and Random Walk Processes; Fourier Analysis; Markov Chain Monte Carlo Methods; Markov Decision Processes; Markov Models and Social Analysis; Probability Theory: Formal; Probability Theory: Interpretations; Queues; Time Series: Advanced Methods.

Bibliography

Bingham, N.H., Kiesel, R., 1998. Risk-Neutral Valuation. Springer, London.
Böttcher, B., Schilling, R.L., Wang, J., 2013. Lévy Matters III – Lévy-Type Processes: Construction, Approximation and Path Properties. In: Lecture Notes in Mathematics, vol. 2099. Springer, Cham.
Chow, Y.S., Teicher, H., 1997. Probability Theory. Independence, Interchangeability, Martingales, third ed. Springer, New York.
Dellacherie, C., Meyer, P.-A., 1980. Probabilités et Potentiel. Chapitres V à VIII: Théorie des Martingales. In: Actualités Scientifiques et Industrielles, vol. 1385. Hermann, Paris (English Translation: (1982) Probabilities and Potentials B: Theory of Martingales. Elsevier, Amsterdam.).
Doob, J.L., 1939. Book review of (Ville, 1939). Bulletin of the American Mathematical Society 45, 824.
Doob, J.L., 1940. Regularity properties of certain families of chance variables. Transactions of the American Mathematical Society 47, 455–486.
Doob, J.L., 1953. Stochastic Processes. John Wiley & Sons, New York.
Ethier, S.N., Kurtz, T.G., 1986. Markov Processes. Characterization and Convergence. John Wiley & Sons, New York.
Gorroochurn, P., 2012. Classic Problems of Probability. John Wiley & Sons, Hoboken, NJ.
Ikeda, N., Watanabe, S., 1981. Stochastic Differential Equations and Diffusion Processes. North-Holland, Amsterdam.
Kunita, H., Watanabe, S., 1967. On square integrable martingales. Nagoya Mathematical Journal 30, 209–245.
Lévy, P., 1937. Théorie de l'Addition des Variables Aléatoires. In: Monographies des Probabilités, vol. 1. Gauthier–Villars, Paris.
Mansuy, R., 2009. The origins of the word "martingale". Journal Eléctronique d'Histoire des Probabilités et de la Statistique 5 (1), 1–10.
Meyer, P.-A., 1966. Probabilités et Potentiel. In: Actualités Scientifiques et Industrielles, vol. 1318. Hermann, Paris (English Translation: (1966) Probability and Potentials. Blaisdell, Waltham, MA.).
Revuz, D., Yor, M., 1999. Continuous Martingales and Brownian Motion, third ed. Springer, Berlin.
Rogers, L.C.G., Williams, D., 1987/94. Diffusions, Markov Processes and Martingales, vols. 1, 2. John Wiley & Sons, Chichester.
Schilling, R.L., 2005. Measures, Integrals and Martingales. Cambridge University Press, Cambridge.
Schilling, R.L., Partzsch, L., 2014. Brownian Motion, second ed. de Gruyter, Berlin.
Stein, E.M., 1993. Harmonic Analysis. Real-Variable Methods, Orthogonality, and Oscillatory Integrals. Princeton University Press, Princeton, NJ.
Stroock, D.W., Varadhan, S.R.S., 1997. Multidimensional Diffusion Processes. Springer, Berlin.
Ville, J., 1936. Sur la notion de collectif. Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris) 203, 26–27.
Ville, J., 1939. Étude Critique de la Notion de Collectif. In: Monographies des Probabilités, vol. 3. Gauthier–Villars, Paris.
von Mises, R., 1932. Théorie des probabilités. Fondements et applications. Annales de l'Institut Henri Poincaré 3, 137–190.
Williams, D., 1991. Probability with Martingales. Cambridge University Press, Cambridge.

Relevant Websites

www.jehps.net, in particular www.jehps.net/juin2009/Mansuy.pdf – Journal Eléctronique d'Histoire des Probabilités et de la Statistique.
www.oed.com – Oxford English Dictionary (OED).

Maruyama, Masao (1914–96)

Andrew E Barshay, University of California Berkeley, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Maruyama Masao described himself as a "man with no talent other than scholarship" whose 'main business' was the study of East Asian, especially Japanese, political thought. As such he belongs among the most influential Japanese historians of the twentieth century. But the history he lived had also compelled him to open what he called a 'night stall' as a political scientist, activist intellectual, and perhaps the preeminent imaginer of democracy in Japan since 1945. Characterized variously as a modernist, national democrat, or simply as a liberal, Maruyama elicited both inspiration and criticism from across the ideological spectrum.

Introduction

Maruyama Masao described himself as a "man with no talent other than scholarship" (quoted in Tsuzuki, 2013: p. 76) whose 'main business' was the study of East Asian, especially Japanese, political thought. On that basis alone, he would number among the most influential Japanese historians of the twentieth century. But the history he lived had also compelled him to open what he called a 'night stall' as a political scientist, activist intellectual, and perhaps the preeminent imaginer of democracy in Japan since 1945. Characterized variously as a modernist, national democrat, or simply as a liberal, Maruyama belonged to no party. As an opinion leader he formed an 'invisible school,' or 'spiritual community' with generations of readers (Tsuzuki, 2013: p. 236), eliciting both inspiration and criticism from across the ideological spectrum.

Maruyama through 1945

Maruyama Masao was born in Osaka in 1914. His father, Maruyama Kanji, was a prominent journalist with the Osaka *Asahi* and *Mainichi* newspapers, an admirer of British institutions, and a critic both of the political domination by the 'Satsuma–Chōshū clique' and abuse by the military of its protected position under the Meiji Constitution of 1889. The Maruyama household was frequented by Hasegawa Nyozekan (1875–1969), whose sharp critiques of the 'metaphysical view of the state' were widely read. On his mother's side – she was herself a devout Buddhist – Maruyama's family included Inoue Kiroku, a journalist associated with the nationalist Seikyōsha group. In 1921, at the age of 7, Maruyama moved with his family to the Yotsuya district of Tokyo, and one of his first compositions was an account of the Kanto earthquake of 1923 that destroyed much of the city. The era's liberal atmosphere, in short, was an element of personal and family experience – as was an undercurrent of social and political volatility.

Maruyama was educated at the apex of the imperial system: at First Higher School (1931–34), then as an undergraduate (1934–37), and finally as a graduate assistant (1937–40) in the Law Faculty of Tokyo Imperial University. He would go on to hold the position of assistant professor (1940–50) and professor (1950–71) in that faculty.

At university, Maruyama studied with Nanbara Shigeru (1889–1974), a historian of Western political philosophy. The encounter was triply significant. With Nanbara, Maruyama read Hegel, and (despite his mentor's criticisms) was himself 'Hegelianized.' Through Nanbara, Maruyama seems to have absorbed the ethos of Uchimura Kanzō's Mukyōkai, the anti-institutionalist Non-Church wing of Japanese Protestantism. Maruyama's insistence on the inviolability of conscience and physical personhood, rejection of patriarchalism, and discomfort with 'master–disciple' relations in scholarship reflect Mukyōkai concerns. Most decisively, Nanbara pressed Maruyama not to follow him into the study of Western political thought, but to engage Japan's Tokugawa legacy from the viewpoint of Western social science.

This confrontation marked Maruyama's – somewhat countercyclical – passage to intellectual maturity. He had internalized the perspectives of the so-called Lectures Faction of Japanese Marxism, which emphasized the persistence of 'semifeudalism' in Japanese society and called for the completion of the 'bourgeois revolution' that the Meiji Restoration had begun. But his reading of the German neo-Kantians (already considered passé) convinced him that Marxist historians were insufficiently conscious of the critical principles that structured their narratives, tending instead to dissolve their judgments – as to 'progress,' 'reaction,' 'stasis,' etc. – into the events themselves.

Still, Maruyama had become a historian in a Hegelian mode. History moved through evolutionary stages, in a discernible direction. But it was also 'cunning': the meaning of history, including that of his own society, had to be wrested from events through methodologically informed intellectual struggle. Neither Japanese tradition nor its Marxist antagonists provided a satisfactory method of analysis.

Karl Mannheim's *Ideology and Utopia* (1936) brought a crucial breakthrough. Here Maruyama encountered the idea of a conditioned 'mode of thought' that could be shared by even sharply antagonistic intellectual groupings; this freed him from narrowly biographical or genealogical approaches to intellectual history. Along with it came the idea of 'frames of reference' through which social groups apprehended and gave meaning to reality: the 'Five Relations' of Confucianism were a cardinal example. As with Marxists, Mannheim held that in concert with shifts in the social base, these frames of reference

also shift. But change, he stressed, was not real until it had been apprehended via a frame of reference. Not the Marxian base but the totality of social life and thought made up reality.

Of signal importance was Mannheim's insight into movements of 'return,' which was a major theme both in mid-Tokugawa thought and in Maruyama's own day. In such movements, a group of intellectuals who have begun to be 'cast up by the social process' of emerging modernity "takes refuge in the past and attempts to find there an epoch or society in which an extinct form of reality-transcendence dominated the world, and through this romantic reconstruction … seeks to spiritualize the present" (Mannheim, 1936: p. 259). Maruyama was able to see that Tokugawa movements of 'return,' particularly Motoori Norinaga's nativism (*kokugaku*), depended upon categories and modes of thought that had ostensibly been 'overcome' – in this case the sinophile Ancient Learning of Ogyū Sorai (Maruyama, 1978: p. 18). In this way, such movements provided a vital insight into how social groups thought through the great transition, via absolutism, from feudal to modern society. Underlying Maruyama's strong response to Mannheim was a shared value position. To think through the transition from feudalism to modernity, for Mannheim, was to wrestle with the democratization of society and thought. And although 'democratization' does not appear as the explicit concern of Maruyama's work until after 1945, it is hard to read even his early writings except in such terms.

Thus, even as Japan plunged into total war, Maruyama defined his position as the conscious, and again countercyclical, affirmation of modernity in the face of powerful tendencies in Japanese thought and politics that were bent on rejecting or 'overcoming' it. Modernity denoted a mode of thought marked by the irreducible difference between what is and what ought to be. This consciousness of difference, and of necessary tension, gave Maruyama a critical stance flexible enough both to resist the emotional undertow of 'national polity' thought (which identified Japan with the 'unbroken' tradition of imperial rule) and to engage Marxism as an intellectual system. This 'interiority' or self-possession formed the normative core of Maruyama's notion of subjectivity (*shutaisei*). To be modern was, in short, to exercise a transcendent, universal critical faculty in a particular social/national totality. A modern or 'open' society, in turn, was one where the free and untrammeled association of individuals sets the norm, where politics was the space in which conflict was confronted and resolved in an ongoing history (as Hegel put it) of 'progress in the consciousness of freedom.' Human society and institutions, therefore, were neither natural nor metaphysically guaranteed. Rather, they were vital fictions worked out by actors with some degree of self-consciousness. The modern consciousness did not contrast fiction to reality or truth, but rather negated any metaphysically guaranteed order in which norm and nature, the cosmic and human spheres, mirrored and confirmed one another.

Here we recognize the core argument of Maruyama's first major work, *Studies in the Intellectual History of Tokugawa Japan*. Written and published separately in the *Kokka Gakkai Zasshi* between 1940 and 1944, and completed literally moments before his departure for military service, they appeared as a book in 1952. Maruyama had provided for Tokugawa intellectual history a narrative of the breakup of the neo-Confucian, 'natural order' orthodoxy of Zhu Xi under the external pressure of social and economic contradictions and the internal dissolution of its 'continuative mode of thought.' The negation of orthodoxy came in the form of Ogyū Sorai's paradoxical 'logic of invention.' Sorai as a Machiavelli–Hobbes *manqué* had called upon the shogun Yoshimune to 'reinvent' along pristine feudal lines a rapidly commercializing social order, thus seeking to 'produce nature' by the absolutist 'logic of invention.' "In order to solve the crisis of feudal society, Sorai set out to destroy the theory of natural order, but instead brought forth a demon whose actions he was unable to control" (Maruyama, 1974: pp. 222, 238). To have discovered in Sorai an 'autonomous personality' and a 'discoverer of politics,' and to have shown that "even the most monolithic of systems contains within it the inherent necessity for its own collapse" represented nothing less, Maruyama declared, than "the salvation of my soul" (Maruyama, 1952: p. 8).

Yet the 'demon' of this inventive logic was destined to be frustrated. Japanese history, then and thereafter, appeared to Maruyama as a succession of thwarted breakthroughs to universality. Japan could have, but did not, become a fully modern, democratic nation-state. "The degree to which the European [*Yōroppateki naru mono*] makes its way into the realm of the spirit," he wrote in 1943, "provides a barometer for gauging the total phenomenon of Japanese modernization." By this measure, Japan after the Meiji Restoration had assimilated the material techniques of the West, but had yet to undergo a spiritual revolution. Politically, the potentially liberating 'logic of invention' had been short-circuited into a ruthlessly pragmatic *raison d'état* and trapped, so to speak, in institutions of state closed to public scrutiny. In society as a whole, the senses had been liberated, but the nation had yet truly to confront the 'European spirit,' and instead persisted in treating European thought as a fetish. Japan's intellectual world had therefore developed – Maruyama quotes the émigré philosopher Karl Löwith – into a 'two story house' with no staircase to connect the floors (Maruyama, [1943]1976: pp. 119, 127–132). As yet, Maruyama wrote, at the first sign of absorption, all the trappings of Western thought and culture could simply be shaken off. But he believed this to be a deeply embedded fallacy, one that had led 'spiritual' Japan to embark on a catastrophic attempt to purge itself and Asia of 'materialism.' Far from claiming that Japan was called to 'overcome' modernity, Maruyama believed that modernity, as a conscious participation by individuals in the struggle, on the social plane, of 'ought' with 'is,' constituted a 'not yet' to be struggled for.

It certainly appears that in his work up to 1945, Maruyama was idealizing the West and denigrating Japan for its historical failures and 'lack' of a true Western consciousness. Understandably, this criticism followed him for most of his professional life. As Maruyama himself remarked some years later:

> If I am told, "You're idealizing the European past and treating it as universal," I can only answer: That is exactly the case. This isn't because there are no universal elements in other cultures, of course. But I acknowledge that in my thinking I rely on abstractions from European culture. I consider it to be a universal legacy to humankind [*jinrui fuhen no isan*]. I firmly believe this. And I want more and more to study it.
>
> Maruyama and Tsurumi, 1969: p. 93

Maruyama's strong affirmation of the universality within European culture came at a time in Japan when an already myopic and defensive exceptionalism had taken its most brutal form; clearly his confrontation with this exceptionalism marked his life in profound ways. Moreover one cannot read Maruyama in context without sensing an overriding critical and national concern. His work, in other words, was nationalist in intent: hence his lifelong attraction to Fukuzawa Yukichi (1835–1901), the great publicist and prophet of Japan's nineteenth century modernization. But as it was for Fukuzawa, for Maruyama too this nationalism was meant to be relative, not solipsistic – nationalism as a freely elected affinity and for that reason an instrument of genuine universalization.

Maruyama after 1945

At War's end, Maruyama was stationed in Hiroshima, attached to an army unit that monitored shipping in Japanese waters; his mother died on the day of Japan's surrender. Intellectually, he emerged from the War as a 'protestant': he had written in covert protest of an unfree present, and continued thereafter to be concerned with overcoming obstacles to freedom and modernity in Japan. It was the typical claim, and the central problem, of his 'protestantism' that such obstacles lie most thickly on the ground of Japanese tradition itself. In this sense he was, as many have described him, a modernist without apology.

With Japan under occupation, Maruyama struggled to adjust to the tragicomic feeling of having to "learn the ABCs of modern civilization from General Douglas MacArthur" (Maruyama, [1945]1976: p. 188). Yet more than anything else, this was liberation, albeit unearned. Now was the time to place the language of conscious, engaged, mass democratic action at the center of Japan's political discourse. In this effort, the institutional, intellectual, rhetorical, and psychological dimensions of Maruyama's public work were related – as was true for anyone active at the time – with unprecedented dynamism.

Negation was the first task. The assignment of 'war responsibility' required more than an attack on the 'military cliques' or *zaibatsu*. To expose the social roots of war, it was necessary to destroy the *ancien régime* in the domain of thought and ideology, above all the "tenacious familial [*dōzokudanteki*] social structure and its ideology, which is the hothouse of the old nationalism" (Maruyama, [1951]1969: p. 152; Maruyama, 1966: p. 168). Never had Maruyama felt so vindicated as by this opportunity for public expression: as a student at First Higher he had been arrested, jailed, beaten, deprived of personal papers, and placed under surveillance for attending a legal gathering of the Society for the Study of Materialism. It was an experience that left him in tears of chagrin at his inability to utter one word of protest. The Peace Preservation Law of which he had run afoul symbolized to him the legal paranoia of the *tennōsei* or 'emperor system,' and now the entire apparatus was being exposed to public scrutiny.

With the publication, in the new opinion journal *Sekai*, of 'Theory and Psychology of Ultranationalism'('Chōkokkashugi no ronri to shinri') in May 1946, Maruyama appeared – 'like a comet,' as the formula had it – to illuminate the devastated intellectual landscape. His attempt to render visible the "invisible net" of *tennōsei* ideology was "as influential as it is possible for a printed work to be" (Tsuzuki, 2013: p. 114), and it seems to have had a virtually physical impact on readers. To begin with, Maruyama's description of 'power dwarfing' (*kenryoku no waishōka*) captured a salient aspect of the public posture of key civilian and military elites, who habitually minimized their personal responsibility while emphasizing their lofty position of proximity to the source of ultimate value. It was their very closeness to the emperor that granted them the privilege of being dwarfed. Also penetrating were the observations concerning the 'transfer of oppression,' which recalled the common experience of many Japanese who found themselves projecting outward and downward the recurring sense of anger and violation that was inseparable from the workings of imperial Japan's familistic hierarchy.

Most important was Maruyama's jarring insight into the 'exteriorization of morality' and its consequences: the 'pathological' identification of the petty self with expansive state power, and along with it the severely compromised capacity to resist authority, or to live according to a moral standard internal to the personality. In this feature of the imperial state, the fact that it "trampled with shod feet on the mind [*seishin*] of every single person" (Maruyama, [1979]1995: vol. 11, p. 150). Maruyama saw the core of 'ultra' nationalism, and its decisive difference from the 'neutral' stance ostensibly assumed by the states of the West concerning moral values. The entire structure of the argument (juxtaposing an actual Japan to an ideal or abstracted West) may seem problematic today. The essay's power derived from the fact that Maruyama was writing of truly public matters, of the common experience of his countrymen. Everyone 'knew' what he was saying; he was trying to provide not merely description but concepts for the analysis of this experience. In this sense, Maruyama's essay opened a new phase in public and social science discourse, and in short order was followed up by major analyses of Japan's wartime leadership, the career of fascism in Japan and the contemporary world, and the theoretical background and prospects of Japanese nationalism (see Maruyama, 1966).

Thus far negation: but what did Maruyama seek thereby to construct, to affirm? The answer could now be stated openly: democracy. Despite the powerful presence of 'America' as occupier and as a model state and society, however, it was increasingly clear that democracy could not be imposed satisfactorily from without. Writing in the bitter aftermath of MacArthur's banning of the 1947 general strike, Maruyama defined the task as follows:

> It has now been made our task to accomplish what the Meiji restoration was unable to carry through: that of completing the democratic revolution. We are being pressed to confront the problem of human freedom itself. The situation facing Japan, of course, is not such that it can be resolved simply by rejoining the orthodox lineage of modern freedom. The bearer of "freedom" is no longer the "citizen" as conceived by liberals since the time of Locke; but must rather be the broad working masses with workers and farmers at the core. Even then the issue is not merely the sensual liberation of the masses, but rather how and how thoroughly the masses are to acquire a new normative consciousness.
>
> Maruyama, [1947]1976: p. 305

Clearly, Maruyama was a self-conscious heir of the European, and Meiji, enlightenment. For the most part he addressed his educated audiences through media of limited reach – university and public lectures, appearances before community and educational organizations, publication in opinion journals rather than mass-circulation dailies, etc. Although he was involved in typical 'enlightenment' projects such as Kibe Tatsuji's Shomin Daigaku ('People's University'), Maruyama's professional identity remained closely tied to Tokyo University, where he held the chair in East Asian political thought until his retirement in 1971. His organizational commitments were also small-scale. After 1945 he joined a number of groups, among them the Seinen Bunka Kaigi (Youth Culture Conference), Nijusseiki Kenkyūjo (Twentieth Century Research Institute), Shisō no Kagaku Kenkyūkai (Institute for the Science of Thought), Kenpō Mondai Chōsakai (Constitutional Problems Survey Group), Heiwa Mondai Danwakai (Peace Problems Discussion Group), and so on. Some of these memberships lapsed quickly; others were kept up. Maruyama never became a party ideologue of any kind on a national level, remaining instead an independent member of the left.

Maruyama, in short, relied implicitly on the diffusion and 'translation' of his ideas into the workaday world via a greatly expanding student and white-collar stratum. Indeed, he had come to recognize that enlightenment by and for the academically pedigreed could only go so far in rooting democracy. This was above all to be an age of the masses, of organized capital and labor – including mental labor. As was the case for the entire Japanese, left throughout the 1950s, for Maruyama too the relation of democracy to socialism – the issue of revolution – became a pressing issue.

Underlying the question of any political commitment, however, was the intellectual (and moral) problem of 'attitude determination.' Maruyama could not embrace the notion of 'class subjectivity' championed in the late 1940s by Marxists such as Umemoto Katsumi. This threatened to degenerate into formalism: the 'party of the proletariat' could not presume to define a 'new normative consciousness.' Ill at ease with such formalism, and contemptuous of wartime 'apostates' who had now returned aggressively to the leftist fold, Maruyama initially refused direct political engagement. Although he rejected as elitist and reactionary the 'old' Japanese liberalism, which had been uncomfortable with any notion of mass politics, class – or social – conflict, Maruyama's political commitments immediately after 1945 were 'diagonal' at best.

The incarnation of the Cold War in Japan, however, forced him into a higher-profile stance. In 1950, self-proclaimed 'realists,' including the Tokyo University historian Hayashi Kentarō, called for Japanese intellectuals to choose sides in the ideological struggle against communism, and in support of a bilateral rather than comprehensive peace treaty. Maruyama demurred. He did so for a number of philosophical and practical reasons, which he expressed in his open 'Letter to a Certain Liberal' (Aru jiyūshugisha e no tegami; Maruyama, [1950]1966).

First, it ran against the grain of Maruyama's self-described 'high-level pragmatism' to be compelled by external forces to subscribe to the rigid metaphysics of Cold War theology. By doing so, one precluded in principle the formation of political positions that would respond to, and help define, the objective demands of a given situation. He refused the 'liberal's' demand that he foreswear any support to the communist party, newly chastened by Cominform's criticism and reeling from the so-called Red Purge. We need, he said, to recognize the "paradox that the establishment of the autonomous human person moves forward via the energy provided by *groups* that, in comparison with Western societies, *stand relatively speaking on the left*" (Maruyama, [1950]1966: p. 513). Maruyama would aid those tendencies that promoted the modernization of Japanese society, and oppose those that did not.

Japan's 1950s, in fact, brought a decade of serious and consequential conflict over the politics of modernization. Along with Shimizu Ikutarō, Maruyama was recognized as an intellectual godfather of the national movement against continuation of the US-Japan Security Treaty (ANPO) that ushered in the end of that decade. In the ANPO protests, Maruyama saw an engaged public – students, white-collar workers, housewives – embodying the ideals of democratic subjectivity he had sought to translate to the Japanese body politic. It does appear that the 'new normative consciousness' he had called for in 1947 had taken hold:

> How do we conceive of politics?… It's a matter of habit, of seeing politics as activity that, even if it makes up only a small part of what we do, day in and day out, still occupies that place *consistently*, as part of carrying out the commonest of our social obligations. More so than any other ideology, however magnificent, or any system, however formally equipped it may be, democracy depends on this kind of thinking. This is its true foundation. Here in Japan we may have lacked the tradition of direct democracy practiced in the Greek city-states. But instead we have as our intellectual heritage the splendid concept of "homespun [lay] Buddhism." Put in contemporary terms, this means the political activity of non-professional politicians. You don't have to "take the political tonsure" in order to get involved in politics. When politics is left to politicians and Dietmen—people whose purpose is politics per se, or to groups such as parties who approach it in the same way—from that moment democracy begins to die.
>
> Maruyama, [1960]1966: p. 182

The 1960 protest against the continuation of (a revised) ANPO was in part bruising political struggle, in part collective effervescence. One of its legacies was undoubtedly the birth of the 'citizen' – a part-time politician out of necessity – as a force in Japanese politics. But while the movement brought down the government of the arch-technocrat and inveterate anti-communist Kishi Nobusuke, it failed in its chief aim. Together with the defeat of union-led strikes at the Miike coal mines, the ANPO struggles paved the way for a successful 'economistic' settlement that in effect redefined Japanese democracy. The institutional framework laid out in the new constitution remained, along with the democratic rhetoric of that period. But under a deepening conservative hegemony, its practical core would no longer be formed by social mobilization and broad political criticism. Instead, it entailed consumer participation in an expanding economy driven by "large-scale mergers and the development of designated contracting systems" (Ishida, 1984: p. 198). The anti-Treaty demonstrations looked to be an exceptional, albeit vital episode, while the meaning of the postwar 'enlightenment' grew clouded.

The Problem of 'Japanese Thought'

Maruyama's key text from these years is *Japanese Thought* (*Nihon no shisō*, 1961). Consisting of essays written between 1957 and 1959, it seems to prefigure the ambiguities of the entire ANPO experience, giving full vent to the agonistic strain in Maruyama's thought yet revealing nuances not present in his immediate postwar writings.

The essence of Japanese thought, Maruyama argues, is to have no essence. Japanese tradition has no axis analogous to Christianity in the West, or Confucianism in China. It is a tradition without structure in which history becomes the nostalgic welling up of national memory, something "akin to the attachment felt by a mother toward a child she has lost" (Nakajima, 1968: p. 84). Yet the intellectual landscape is far from empty. Japanese thinkers have accumulated an extraordinary 'stock' of discrete ideas to draw on for their own purposes. What is lacking is an "absolute being or a 'Way' that logically and normatively orders the world in its own characteristic manner." In its absence, Maruyama argues, 'Japanese thought' formed a tenacious pattern by which 'faith in felt reality' (*jikkan shinkō*) and 'faith in theory' (*riron shinkō*) operated as functional equivalents, cutting off the dialectic of mutual negation, testing, or correction that Maruyama, from his transcendent position, finds to be essential to the 'normal' operations of the socialized intellect.

Maruyama was not principally concerned with accounting for the origins of Japan's nonaxiality. He begins *in medias res*, looking back across the great but not far distant rupture of 1945. How did 'tradition,' as mobilized under the modern emperor system, actually function? In the wake of defeat, Maruyama would have answered that it functioned in a uniformly oppressive manner. By the time of *Nihon no shisō*, he has softened considerably his attitude toward 'tradition,' which he discusses in terms of an ethic of 'being' versus 'doing.' Now, rather than call for the elimination of the ethic of 'being,' he urges its incorporation in a synthesis of traditional and modern styles of life and thought. Echoing Max Weber, Maruyama finds that 'being' is functional in the sphere of art and culture: indeed it stands in immediate need of protection from an obsessive 'modern' concern with producing results. As he observed elsewhere, "*The Tale of Genji*, *Hamlet*, and *Faust* cannot be shot off as rockets" (Maruyama, 1969: p. 310). At the same time, he finds Japanese society lacking in a political ethic of 'doing,' where results – the rationalization of means according to publicly determined ends – do indeed matter. By applying criteria of instrumental rationality where they did not belong, Japan had become hypermodern in one sphere, and insufficiently modern in another.

But how does this relate to 'Japanese thought'? The answer lies in the totality of elements, premodern, modern, and hypermodern, and their particular distribution or valences within the social system; it is in the whole, rather than in any of its parts, that the Japanese mode of thought operates. And as Maruyama shows, there are definable historical patterns to this operation, which now can, and must, be challenged. The ideology of *kokutai* – the unbroken line of imperial rule over Japan – had legitimated rapid institutional modernization in the name of the monarch. At the same time, it claimed that the monarchy was the supreme embodiment and validator of the entire system of patriarchal, factional, and personalistic human relationships that characterized village society. The 'emperor system' in this sense was the most powerful and consequential 'invented' tradition of modern Japan. It provided the institutional and linguistic frame through which all other 'inventions' received their official imprimatur. The task now was to differentiate, to overcome the deep-seated 'need' for the all-encompassing legitimating frame that was at once so comforting and suffocating. But could Japanese people live without it? Could they overcome 'Japanese thought'? Maruyama was uncertain. In an enduring metaphor, he noted the tendency of organizations of all kinds in Japan to form themselves into 'octopus pots' (*takotsubo*) in which the group's activities, ideas, and values become largely self-referential, the authority of the group self-justifying, and socially isolating. This includes 'modern' organizations – political, business, and technical as well. Parallel to this is the increasing sophistication, centralization, uniformity, and pervasiveness of the mass media. The result of their interaction is a society in which lateral ties between groups of all kinds are weak; groups form images of themselves and the 'world' that go unchallenged and 'walk on their own.' In the absence of such ties and of genuine feedback throughout society, critical and democratic consciousness is fragmented and blunted. It is precisely in this mix of hypermodernity in one sector with premodernity in another that Maruyama saw functional continuity with *tennōsei*, albeit sans *tennō* in the prewar mode.

Counterposed to the 'octopus pot' was the *sasara*, a bamboo whisk, with a shared base that is finely separated toward the tip. The finely separated tip connotes articulation, differentiation, specialization, and pluralism as a value. But how is the 'shared base' of the *sasara* to be understood? Does it refer to a single nation or ethnicity with an intuited cultural code? It would seem not. Speaking of the development of the European social sciences, Maruyama refers almost enviously to a 'long common tradition' linking Greece, the middle ages, and the Renaissance with the modern emergence of highly specialized disciplines; the comprehensive systems (or at least catholic concerns) of 'Hegel, Stein, and Marx, or Bentham and Comte' that were typical of the early nineteenth century give way by its end to the specialized sciences familiar today. It was these latter – the 'finely separated tips' of the *sasara* – that had been cut off from their roots and transplanted to Japan. The shared root, in other words, was a history of dialectic, argument, and contestation. Without this history, the disciplines of social science in Japan were 'academic' from the beginning, and even more susceptible than in Europe to self-enclosure. More fatefully, the 'academy' itself was presumed to exist for the service of a state-defined public good.

Nihon no shisō, Maruyama asserts, is also his own 'self-criticism' from the point of view of "radical intellectual aristocratism [*kizokushugi*] linked interiorly with a radical democracy": he looks forward to a *sasara*-type society, one in which the image of 'Karl Marx reading Hölderlin' would capture the intellectual life of its people (Maruyama, 1961: pp. 179–80). For all that it seems to set out an uncompromising task of mental revolution, and therefore remain true to agonistic modernism, *Nihon no shisō* is good-natured and surprisingly upbeat in tone. It points to a modernism in process of differentiation, one making its peace, via the categories of aesthetics – including linguistic play – with nationality. But

would that be a slippery slope? Would modernism come to valorize its moment as one of modernity achieved, pressing forward to the presentation of Japan as a new model of development? Or would modernism retain the capacity for 'utopian' resistance to the status quo, even at the risk of marginalization? What, in short, would be the intellectual consequences for Japan's democracy of coming to terms with nationality?

The balance, Maruyama seems to have feared, was tipping in favor of the 'octopus pot' rather than the 'bamboo whisk.' More and more, Japanese society lacked the 'feedback of counter communication from the periphery,' without which it would corrupt (Maruyama, 1969: p. 347). Yet he had little to say about the concrete identity of the 'periphery,' the content, or social character of the feedback it generated. Indeed he appeared to pin his hopes on intellectual migrants to the periphery as producers of that vital 'counter communication.' Perhaps that should not be surprising. Maruyama's image of Japan's 'periphery' was one of traditionalism and irrational attachments. Although he had argued that democracy had to become 'irrationalized' – or deintellectualized – this still left unanswered the question of how the center is to be 'reintellectualized' from below. And because a backward and traditionalistic periphery was, in effect, a structural requirement of his thinking, Maruyama could never answer such a question.

Maruyama, then, remained a protestant. For him, democracy was priesthood of all believers, a 'permanent revolution.' Its enemy was complacency or 'blind faith' in ideology or institutions. No value, freedom (or democracy) included, could be made 'real' simply by being divested of utopian elements. 'Realism' as the ideological affirmation of the status quo, and the identification of the established order with democracy as such, in effect killed democracy in its own name. Thus Maruyama answered a nationalist critic of postwar Japan's 'occupation' or 'sham' democracy: "As for my own choice in the matter: Rather than opt for the 'reality' of the empire of Japan, I'll put my money on the 'sham' of postwar democracy" (Maruyama, 1966: p. 585). But who was betting along with him?

Shifting Ground: The Later Maruyama

The major writings of Maruyama's later years – 'later' beginning piecemeal by the late 1950s and definitively so by the early 1970s – were almost entirely historical. Increasingly he explored the problem of 'continuity in change' in Japanese political, ethical, and historical consciousness, using the metaphors, successively, of 'prototype,' 'ancient substrate,' and 'basso ostinato' as he refined his description of how they worked. Much as the seminal ethnographer Yanagita Kunio pursued the original Japanese ever outward to the archipelagic periphery, Maruyama pushed the sources of 'Japanese thought' further and further back in time. "A century after Nietzsche babbled about the death of God," Maruyama observed, in its absolutization of the 'now,' Western historicism had in fact drawn strikingly close to the Japanese scene. "It may be," he wrote, "that the 'continuity in change' that distinguishes our historical consciousness has, in *that aspect* as well, been a factor in placing Japan at the forefront of world's most advanced countries. Should we regard this paradox as yet another instance of 'the cunning of reason,' or as a comedy that is rapidly approaching its conclusion?" (Maruyama, 1972: p. 41).

For those who recalled his Mannheimian early work and acute analyses of political psychology, the later Maruyama was disappointing: ahistorical and essentialist. Yet it seems clear that he wished to hasten the 'comedy's' end, even if at the price of intellectual uncertainty and anxiety. The forms of political, historical, and ethical consciousness that he came to discuss in terms of sudden 'welling up' from the historical (not social) depths, or of a submerged musical line breaking through as the dominant 'theme,' are no more than the reverse of the forms of political subjectivity that he hoped would emerge from, and promote, the process of democratic revolution in postwar Japan. It was the 'universal' that gave meaning to the 'particular,' movement to stasis, and utopia to 'reality.'

In the end, Maruyama Masao was a thinker of the democratic nation-state and, increasingly, of a unitary, intransigent culture that tormented him intellectually. At the level of conceptualization he was unquestionably 'complicit' in both, thinking through and against the nation-state, through and against the national culture. What does 'complicity' mean? It is a major problem today and a theme of much recent criticism – including postcolonial criticism of Maruyama. As Fred Dallmayr has noted, in our times opposition to the ostensibly genetic (biology, race, and nature) has required a dualism whose other antipode is 'spirit.' The options for or against spirit, for or against a metaphysics of subjectivity, are "all 'frightfully impure and contaminated.' One always has to ask oneself which option or 'type of complicity' is less grievous" (Dallmayr, 1993: pp. 35–36). Maruyama sought to contest the sway of 'nature' over 'nation,' to free it for the 'spirit' he called democracy. He had a lively capacity for concrete analysis that he applied to the pathologies of revolution and counterrevolution – phenomena, that is, of repressive and mobilized political essentialism. But he was not a postmodern thinker; he retained in his thought the constitutive dualism of subject to object and the belief, shared with Harold Laski, that within the framework of a revolutionized polity the 'best self' of each individual could be brought to realization. In this or that aspect – such as his notion of revolution as massive (and rare) value dissensus – one can certainly find hints of a Foucaultian perspective. But it is with hints that he stops. Perhaps Maruyama did not push deeply enough in considering the problems – the pathologies – of 'normal' modernity itself. The question now, however, is how the inevitable problem of complicity is to be faced by those who would place themselves beyond the horizon of the modern.

See also: Berlin, Isaiah (1909–97); Dahrendorf, Ralf (1929–2009); Democracy, History of; Democratic Theory; Freedom, History of; Habermas, Jürgen (1929–); Historical Thought and Historiography: East Asia; Intellectual History; Intellectuals, Sociology of; Japan: Sociocultural Aspects; Mannheim, Karl (1893–1947); Meiji Revolution, The; Modernity; Modernization; Nations and Nation-States in History; Political Thought, History of; Schmitt, Carl (1888–1985); Social Science and Universities; Translating Sociology into Japanese, Chinese and Korean Languages.

Bibliography

Works by Maruyama Masao

Maruyama, M., 1943. Asō Yoshiteru, Kinsei Nihon tetsugakushi o yomu. In: Maruyama, 1976.
Maruyama, M., 1945. Kindaiteki shii. In: Maruyama, 1976.
Maruyama, M., 1946. Theory and Psychology of Ultranationalism. In: Maruyama, 1969.
Maruyama, M., 1947. Nihon ni okeru jiyū ishiki no keisei to tokushitsu. In: Maruyama, 1976.
Maruyama, M., 1950. Aru jiyūshugisha e no tegami. In: Maruyama, 1966.
Maruyama, M., 1951. Nationalism in Japan: Its Theoretical Background and Prospects. In: Maruyama, 1969.
Maruyama, M., 1952. Nihon seiji shisōshi kenkyū. Tokyo Daigaku Shuppankai, Tokyo.
Maruyama, M., 1960. Gendai ni okeru taido kettei. In: Maruyama, 1966.
Maruyama, M., 1961. Nihon no shisō. Iwanami Shinsho, Tokyo.
Maruyama, M., 1965. Patterns of individuation and the case of Japan: a conceptual scheme. In: Jansen, M. (Ed.), Changing Japanese Attitudes toward Modernization. Princeton University Press, Princeton.
Maruyama, M., 1966. Gendai seiji no shisō to kōdō, rev. and exp. ed. Miraisha, Tokyo.
Maruyama, M., 1969. Thought and Behavior in Modern Japanese Politics. Oxford University Press, New York.
Maruyama, M., 1972. Rekishi ishiki no 'kosō. In: Maruyama, M. (Ed.), Rekishi Shisō Shū. Chikuma Shobō, Tokyo.
Maruyama, M., 1974. Studies in the Intellectual History of Tokugawa Japan. Princeton University Press/University of Tokyo Press, Princeton and Tokyo.
Maruyama, M., 1976. Senchū to sengo no aida, 1936–1957. Misuzu Shobō, Tokyo.
Maruyama, M., September 1978. Shisōshi no hōhō o mosaku shite—Hitotsu no kaisō. Nagoya Daigaku Hōsei Ronshū, 1–31.
Maruyama, M., 1979. Nihon shisōshi ni okeru 'kosō' no mondai, vol. 11. In: Maruyama, 1995.
Maruyama, M., 1982. Kōei no ichi kara. Miraisha, Tokyo.
Maruyama, M., 1992. Chūsei to hangyaku. Chikuma Shobō, Tokyo.
Maruyama, M., 1995. Maruyama Masao Shū, 16 vols. Iwanami Shoten, Tokyo.
Maruyama, M., Tsurumi, S., 1969. Fuhen genri no tachiba. In: Tsurumi, S. (Ed.), Kataritsugu sengoshi, vol. 1. Shisō no Kagakusha, Tokyo.

Other References

Barshay, A., 2007. The Social Sciences in Modern Japan: The Marxian and Modernist Traditions. University of California Press, Berkeley.
Dallmayr, F., 1993. The Other Heidegger. Cornell University Press.
Fukuda, K., 2000. Maruyama Masao to sono jidai. Iwanami, Tokyo. Booklet No. 522.
Hiraishi, N., December 2004. Maruyama Masao, national democrat. Social Science Japan, 11–13.
Ishida, T., 1984. Nihon no shakai kagaku. Tokyo Daigaku Shuppankai, Tokyo.
Mannheim, K., 1936. Ideology and Utopia. Harcourt Brace Jovanovich, New York.
Nakajima, M., 1968. Sengo shisōshi nyūmon. Ushio, Tokyo.
Tsuzuki, T., 2013. Maruyama Masao e no michi annai. Yoshida Shoten, Tokyo.

Marx, Karl (1818–83)

Terrell Carver, University of Bristol, Bristol, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Karl Marx (1818–83) developed a distinctive 'new materialism' through which human history was understood as a succession of economic modes of production on which legal, political, moral, and intellectual systems were dependent, rather than the reverse. Science (*Wissenschaft*) was thus a practical activity undertaken in circumstances that were necessarily historical and social, presuming the ubiquity of class struggles and the communist project to transcend them. Marxism, originating with his friend and sometime collaborator Friedrich Engels (1820–95), promoted the controversial summarization of Marx's work in terms of a 'materialist conception' of history founded on 'dialectical' laws of nature, history, and thought.

For Marx science (*Wissenschaft*) (Terms denoting a unified conception of the sciences derived from the natural philosophy tradition and promoted by the German idealist and Hegelian philosophies. Marx's 'new materialist' version of this was opposed to empiricist reductionism of knowledge and truth to ahistorical and individualistic accounts of perception. His 'outlook' presumed that knowledge of any kind is practical, social, historical, and political in origins and usage.) was itself a unity, and natural science was an activity within, rather than a model for, the human social world. For him the master science was political economy because it dealt with the way that industry and technology had transformed humanity. His lifetime project was a critique of that science, informed by his understanding of communism (term denoting a long tradition which emphasizes community-sharing of goods and services, and other cooperative strategies in social relationships, to minimize conflict and corruption in society; Marx, Engels, and others revived this in the early to mid-nineteenth century to democratize not only political institutions but also the economic structures and practices through which goods and service are produced and consumed using modern industrial means) as the political movement to emancipate the working class. Friedrich Engels (1820–95), V.I. Lenin (1870–1924), and other commentators ascribed to Marx the discovery of important scientific laws of human history and capitalist production, attributed to him a unique 'dialectical' scientific method, and promulgated a 'scientific socialism' (term coined by Engels to summarize Marx's 'new materialism' as a union of anticapitalist but proindustrial politics, founded on a 'dialectical' science derived from both Hegelian logical reasoning and the factual character of the physical sciences) in his name. The extent to which posthumous 'materialisms' in social and natural science were an accurate reflection of Marx's own views has been debated within and outside the Marxist tradition, and these debates have contributed importantly to the philosophy of science. Marx and Marxism (term not commonly used until the 1880s and 1890s after the widespread circulation of Engels's summarizations of Marx's 'outlook' as a 'materialist conception' of history founded on 'dialectical laws' of nature, history, and thought; as a philosophical, political, and social scientific movement it was essentially contested, and by the early 1900s it had sparked a doctrinal conflict between orthodox and 'revisionist' versions) have themselves been reinterpreted separately and together in the light of later developments in twentieth- and twenty-first-century philosophies of science and social science. In empirical terms almost any social scientific study of social class in the context of industrial, 'emerging,' or developing economies and societies owes an important debt to Marx's emphasis on the social and political consequences of class struggle. (Term denoting a distinctive consequence of Marx's 'new materialism' through which human history was understood as a succession of economic modes of production. From this perspective social gradations were importantly coincident with a division of society into 'haves' and 'have-nots' – those who benefited disproportionately from surpluses and luxuries, and those who worked to produce these things, but were denied access to them. This was because of social, political, legal, religious, or other structures of oppression, against which there was persistent but more or less 'veiled' civil war of resistance. Marx and Engels predicted that a communist and specifically proletarian or industrial workers' revolutionary victory would institute a new and classless social and political order.)

Marx was hailed at his death as a 'man of science' who had made two great discoveries in but one lifetime: 'the law of development of human history' and 'the special law of motion governing the present-day capitalist mode of production' (Engels, 1989: pp. 467–468). In this graveside eulogy Marx's longtime friend and sometime coauthor Engels likened the first of these great discoveries, previously christened 'the materialist interpretation of history' (term coined by Engels to summarize Marx's 'conception' of history as one of successive economic 'modes of production' and concomitant class struggles to be transcended by a communist social revolution utilizing the productive powers of modern industry. Its relationship to social science has been controversial throughout the twentieth-century and remains so today) (Engels, 1980: p. 470), to the formulation by Charles Darwin (1809–82) of 'the law of development of organic nature' (Engels, 1989: p. 467). In other overviews of Marx's work, Engels linked Marx's critical work on political economy and capitalist society with a dialectical method in science derived, so he said, from the philosophy of G.W.F. Hegel (1770–1831). Engels wrote that "Marx was … the only one who could undertake the work of extracting from the Hegelian logic the kernel which comprises Hegel's real discoveries" (Engels, 1980: pp. 474–475). Moreover for Engels,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61088-7

science as conceived by Marx was fully compatible with, and indeed an integral part of, a revolutionary political outlook and practice: "Science was for Marx a historically dynamic, revolutionary force. … His real mission in life was to contribute, in one way or another, to the overthrow of capitalist society" (Engels, 1980: p. 475).

Famously Marx himself denied that he was a 'Marxist' (Engels, 1992: p. 356), whatever that may have meant to himself, and to his friends and enemies while he was alive. While the origins of Marxism lie in Engels's accounts of Marx's life and works from 1859 onward, and in Marx's own published works and articles, Marxism as a science and as a political movement dates from the years after Marx's death in 1883. For over a century it has been an intellectual and political force to be reckoned with, though the initial presumption of methodological and doctrinal unity, implied in the reference to an origin in one man and his works, inevitably began to break down. On both the intellectual and political fronts, the current age is one of post-Marxism.

In another influential overview of Marx's intellectual career, Lenin suggested that Marx combined German philosophy, English political economy, and French socialism and revolutionary doctrines in a unique way (Lenin, 1964: p. 50). Most intellectual biographies proceed chronologically and thematically through this list. However, in setting out what distinguishes Marx as a social scientist, and what contributions Marxism has made to social science, it is more productive to take the direct approach and to outline the characteristics of the Marxian synthesis. This will necessitate some contextual discussion of what science and philosophy meant in Marx's time and milieu, and what sort of political strategies and actions were possible for a committed socialist when even constitutional politics was radical and subversive.

Science and Philosophy

Marx was born on 5 May 1818 in Trier in the Rhineland. His family were converted or at least nonpracticing Jews, and he had a classical and liberal education. This took him to the faculties of law and philosophy at Bonn and Berlin universities in the mid-1830s to the early 1840s, preparing him, rather inadvertently, for a career as a progressive journalist and political radical. This was a time of political reaction in the German states when liberals, especially students, resisted the authoritarian doctrines and practices of self-styled autocrats. Liberals argued instead for representative and responsible government that would protect individual rights to free expression and institute a socioeconomic system favorable to individual mobility. While this may sound familiar enough to present-day readers, the political dialogue of Marx's time now seems laden with obscure philosophical and religious issues. This was partly the result of heavy government censorship imposed on overtly political pronouncements, and partly the result of the requirements of the debate: what facts and arguments would sustain a doctrine of popular sovereignty against the reigning doctrines of absolutist rule by divine right?

The Hegelian idealist philosophy within which these coded political debates took place was itself controversial in relation to the Trinitarian truths of the Christian confession, and this had indeed been recently and notoriously explored by the historian D.F. Strauss (1808–74) in his *Life of Jesus* (1835), and by the philosopher Ludwig Feuerbach (1804–72) in his *Essence of Christianity* (1841, 2nd edn., 1843). Marx was swift to draw the radical conclusion of atheism from these writings, and to align himself with a 'new' materialism rooted in human experience, which he defined as economically productive and historically changing. In the 1830s and 1840s, and particularly in the German-speaking world, 'science' as *Wissenschaft* referred to knowledge in the broadest sense systematically presented. Moreover the tradition of natural philosophy was central to this conception of science, which did not readily admit any great distinction in method between natural and human subjects of study. Science as *Wissenschaft* relied on scholastic concepts and commonsensical observation, rather than on mathematical abstraction and testable hypotheses.

Hegelian idealist philosophy was in any case utterly hostile to a materialist ontology of matter-in-motion and an empiricist epistemology through which truth is ascribed to concepts in so far as they coincide with human sensory experience of the material world. The later nineteenth century saw the rise of such materialist accounts of the natural sciences as empirical modes of study. Subsequently the development in the early twentieth century of the modern social and behavioral sciences was modeled on these physical scientific presuppositions and practices. In any case since the seventeenth and eighteenth centuries English-language philosophy has been hostile to philosophical idealism, and very closely aligned with philosophical empiricism. It follows that Marx's early works, and the overarching conception of science extending through his later works, can be quite difficult to understand from a later vantage point in the twenty-first century. Certainly the reader must be very careful in noting any claims in Marx about what science is supposed to be, so that his thought can be understood without anachronism.

In so far as Marx's notions of science have been influential, a conception of science rooted in idealist *Wissenschaft* has thus survived and flourished. This conception may not cover all conceptions of science within Marxism, nor is it necessarily exactly aligned with those that claim to be Hegelian in origin. Rather it is distinguished by these features:

1. a focus on human ideas as the medium through which science itself exists, rather than a view that knowledge about humanity must in some sense mirror knowledge of the physical world;
2. a presumption that throughout history the human focus on what is true and meaningful will change, rather than a presumption that truths are fixed by realities outside of human experience;
3. a view of science as a form of practical knowledge developed for, and incorporated within, a comprehensive range of human activities, rather than as a realm of abstract truth sought for its own sake by objective observers and politically neutral scientists; and
4. a political presumption that a critical evaluation of present practice, including practices of power and politics, is a necessary condition for productive science, rather than a presumption that proper science is necessarily unconcerned with its own political context, or a view that it simply does not have one worthy of notice.

Natural Science and Social Science

Marx's critique of political economy was a lifetime project pursued in political and intellectual terms that are somewhat unfamiliar today and require careful explication. His ideas and concerns surfaced in his early journalism, published in 1842 when Prussian censorship was comparatively relaxed. They can also be traced in his notebooks and articles from the following 2 years, during which time his newspaper was closed down, and as he says, he 'retired into his study.' Marx had only limited contact at this time with the social science of political economy, and in common with other German readers, he viewed writers such as Adam Smith (1723–90), Adam Ferguson (1723–1816), and Sir James Steuart (1712–80) through the medium of Hegel's synthetic and critical philosophizing on this subject, particularly in the *Philosophy of Right* (1821). This work was subjected to a lengthy critique in 1843, unpublished by Marx in his lifetime, and by 1844 he had also completed a set of 'Economic and Philosophical Manuscripts,' again largely unpublished until 1932. While there is of course controversy over the relationship between these works and those that Marx wrote later in life, there are demonstrable continuities in terms of text and project that confirm the early 1840s as the period when Marx's critique of the economic categories was truly launched.

Politically the young Marx aimed to produce a critique of political economy informed by his own version of some of the principles of socialism and communism that were then current. These were mostly French in origin, and derived from the writings of Henri Saint-Simon (1760–1825), Charles Fourier (1772–1837), and Étienne Cabet (1788–1856). Marx was also influenced by more contemporary writers such as Moses Hess (1812–75), who identified (rather hopefully) the newly emerging industrial poor as prime movers in socialist politics. At this juncture the relationship between socialism as a political doctrine, and social science as the pursuit of knowledge about society, may seem tenuous. However, in the intellectual and political context of the time the very notion that 'society' needed a scientific investigation was itself subversive and 'socialistic.' The established monarchical order was content by contrast to rely on Biblical fundamentalism, 'commonsensical' conservatism, and vigorous repression of any 'free thinking' that might stir up issues concerning social class and political power. This free thinking was exactly what both liberalism and socialism were about, and the two constituted a potent alliance at the time, albeit with many in-built tensions around social class, political participation, and economic goals.

After the revolutionary events of 1848–49 Marx went into exile in England, and during the remainder of his life, his political involvements were necessarily rather problematic in that context. As he himself noted, London was an excellent center for studying both the social science of political economy and the industrial and commercial life that it theorized. While the English tradition in political economy (actually more Scottish than English) was empiricist in its assumptions and methods (rather than idealist, as was German tradition), it was at this stage not rigorously mathematical or deductive. This paradigm shift occurred in the 1870s following the work of William Stanley Jevons (1835–82) and other pioneers of the 'marginalist' revolution.

Central to Marx's critique was the work of David Ricardo (1772–1823), whose *On the Principles of Political Economy and Taxation* (1817, 3rd edn., 1821) is classically descriptive and deductive in a mode of natural philosophy not far removed from Aristotelian conceptions. Ricardo defined and explicated the origins of wealth in terms of commercial value, relating this to the human labor involved in manufacture, and assuming the distribution of wages, profit, and rent to the 'three great classes' that necessarily constitute modern societies. Marx's critique of this influential view was both political, arguing that class-divided societies are but a transitional phenomenon, and technical, arguing that Ricardo's concept of labor could not account for the accumulation of profit through the exchange of value. This critique was published in stages, most notably *Capital*, Vol. 1, (1867) and has been subsequently enlarged with edited versions of articles for succeeding volumes and studies.

Marx died on 14 March 1883 in London of natural causes, after 35 years of exile from his native Rhenish Prussia. While other '48 ers' managed to settle their differences with the regime and to regain their civil rights, Marx made only infrequent nonpolitical visits to Germany, and was resolved to maintain his family home in England. He had married Jenny von Westphalen in 1843, and by 1855 they had three surviving daughters, having lost two sons and another daughter (Franziska) to childhood illnesses, certainly exacerbated by harsh financial circumstances. Since 1962 it has been claimed that Marx was the father of Frederick Demuth, the illegitimate son of the family housemaid, but the alleged evidence for this is highly suspect. Marx's wife and daughter (Jenny) both predeceased him by some months, and his two surviving daughters (Laura and Eleanor) both took their own lives some decades later. Against the obvious gloom and despair of family life in the political wilderness, there are memoirs and recollections extant of a happy and cultivated family life, of Marx's evident and loving affection for his wife and daughters, and his intense grief at the death of his two sons, Edgar and Guido.

Marx and Engels

Any interpretation of Marx is complicated by the fact that from 1844 until the end of his life he worked closely with Friedrich Engels. Though they wrote only three major works together, they engaged in an extensive and well-preserved correspondence (more than 7000 letters between them are currently extant). They also worked in tandem with political groups, particularly those promoting international cooperation among socialists, such as the International Working Men's Association (1864–76). Engels was a very considerable author and publicist before he met Marx, and was at that time the more famous of the two. He was also an autodidact, having availed himself of lectures and student discussions at Berlin University while he was in the army on Prussian national service. Engels was already a participant in the coded politics of post-Hegelian philosophy, and in the German appropriation of 'English' political economy, publishing an outline 'Critique' in 1844 that influenced Marx significantly. Engels also published an analytical and descriptive account of industrial poverty in *The Condition of the Working Class in England* (1845), based on

published sources and the testimony of his own eyes in Manchester.

As indicated above Engels was a notable publicist and popularizer in connection with Marx's work, and the biographical, intellectual, and political context developed in his accounts from 1859 onward set a profound though not unquestioned model for interpreting Marx. By the later 1860s it is evident from Engels's own works that he himself was increasingly influenced by the development of chemistry and physics as rigorous and mathematically precise sciences, closely linked with industrial processes and with the formulation of a philosophy of science. This eventually flowered as positivism, the view that truth of any kind can only follow from the presuppositions and protocols of a singular scientific method, and that the archetypes for this were physics and chemistry as they progressed to increasingly elegant and powerful theorizations and applications.

Engels's grand project in the 1870s was pursued in his own published works until his death in 1895. This was a synthesis of Hegelian 'dialectics,' which he admired for its ability to cope with change and contradiction, with the natural scientific materialism, on a positivist model, that was proving so successful in the laboratory and the factory. Engels formulated three laws of 'dialectics': unity of opposites, negation of the negation, transformation of quantity into quality (Engels, 1987: pp. 111–132). His view was that the discoveries of the natural sciences could be decoupled from the logical atomism and reductionism that empiricism usually implied, and linked instead to unifying 'dialectical' laws of nature, history, and logic itself. This then would unify and transcend materialism and idealism, science and philosophy, practical truth and logical judgment.

The extent to which Marx came to endorse this system, or even to lean in this direction at all, is keenly debated, and the evidence – textual and biographical – is deeply ambiguous. One of the chief difficulties here is achieving an understanding of Marx and his work that is not already deeply influenced by the canons of interpretation set by Engels, as only in that way could these questions be resolved. Moreover Engels was not only Marx's first biographer, but also the biographer of their relationship, as well as legatee and editor of Marx's articles, and authoritative voice on 'Marx.' Engels was the founder of what was becoming, largely through his own efforts at popularizing and developing Marx's ideas, a body of 'dialectical' thought and socialist practice known as 'Marxism.' Chief among the terms he used to promote these views was 'scientific socialism,' coined during the late 1870s and widely publicized.

Engels and Marxism

Science from Engels's perspective has been far more influential within socialist practice, and also within the developing social sciences, than any conception of science based on a fresh reading of Marx's works. This is unsurprising, given the biographical and political circumstances, and the philosophies of science and social science current from the 1890s through the 1920s. Engels coined the term 'materialist interpretation of history' (later denominated 'historical materialism') and also the notion of 'materialist dialectics' (later reformulated by one of his intellectual successors, Georgii Plekhanov (1856–1918), as 'dialectical materialism').

In this way Marxism was generally expounded as an outlook that integrated philosophical truth and political practice. This was done by presenting history, that is, human society in a developmental sequence, as an instance of scientific knowledge. This scientific knowledge, in turn, was said to be consistent with unifying 'dialectical' laws of nature and logic. Those laws were themselves reflections of the only cosmic reality, that of matter-in-motion, as studied by natural scientists. In that way for Engels, as for Marxists, social science was but a branch of the natural sciences and of philosophical logic, provided that these systems of knowledge were properly understood as unified by a 'great basic process' that was itself dialectical, and so properly captured in Engels's three laws.

In practice Marxist social science was typically organized by the notion that 'history is the history of class struggles,' laid out by Marx and Engels in their *Communist Manifesto* of 1848, and illustrated in the abbreviated account of history as successive 'modes of production' offered by Marx in his Preface to *A Contribution to the Critique of Political Economy* (1859), briefly footnoted in *Capital*, Vol. 1. Marxists defined contemporary political struggle as necessarily and profoundly a struggle between the owners of capital (the bourgeoisie or commercial classes) and the proletarians or workers (those who have nothing to sell but their labor), so their methodology was not merely historical but immediately strategic.

Tactically there were of course numerous difficulties in identifying classes and class-fractions and in finding or rejecting coalition partners for working-class interests. Indeed it was also difficult to define the short- and longer-term interests of the any group of workers at all, particularly with respect to religion, nationalism, and numerous other political expressions of identity and purpose. However, in terms of social science there is no doubt that Marxists were among the leaders in studying and publicizing industrial poverty, class divisions in society, and social protest and revolution. While pre-Marxist political economy was not entirely devoid of a concern for industrial workers as human beings and as political agents, it was not at all comfortable with the idea of any 'more or less veiled civil war' (Marx and Engels, 1976: p. 495). Marxist social science inverted the rather functionalist assumptions of traditional political economy and declared that such warfare was an inevitable stage on the way to human emancipation.

Under communism greater wealth would be produced and distributed in ways far more humane than any money-system of commercial exchange could ever allow. While communism was not itself theorized in any detail, either by Marx or Engels or by subsequent Marxists, the class struggle was itself documented and pursued with single-minded thoroughness. Despite the claims of a dialectical unity of all the social, natural, and logical sciences, this work was persuasive and respected to the extent that it was empirical and evidential. Any claim that Marxism genuinely established a social science different in principles and character from 'bourgeois' science founders on a point of incommensurability (i.e., conventional social scientists did not regard it as science at all because of the 'dialectical' claims, whatever the results) or on a point of disjunction (i.e., the 'dialectical' claims about social science were not the basis of the science that Marxists actually did,

whatever their protestations). In practice self-avowed Marxists focused on class division in industrial societies and characteristically did a social science that claimed to be both scientific and political; conversely any social science preoccupied with working-class poverty was often suspected of Marxist political intent that would bias any scientific results. Clearly, different conceptions of science, and of its relationship with politics, have been operative within the social sciences; Marxist social science highlights and largely constitutes this area of debate.

Orthodoxies and Revisionisms

Marxism became an orthodoxy once it had been challenged, most notably by 'revisionists' in the later 1890s. While this was initially a political challenge concerning the nature and timing of socialist revolution, further doubts were raised concerning the dialectical philosophy of the sciences espoused by Engels and the specific propositions of Marx's critique of political economy. By the 1920s György Lukács could write that orthodoxy in Marxism refers exclusively to method. This position reflected the success of the marginalist revolution in economics, which isolated Marx's critique of political economy as irrelevant, given that traditional political economy had been effectively superseded. It also reflected the success of more sympathetic attempts to adapt Marx's economic work on capitalism to a form of economics that would at least be intelligible to the mainstream. And most of all it reflected a process of rethinking Marx that was just beginning.

This process was aided by the first attempts to produce Marx's published and unpublished works in accessible form (the Marx–Engels *Gesamtausgabe* of the 1920s and 1930s, edited initially by D.B. Ryanzanov). Engels was of course in the edition as well, and the production of his works alongside Marx's inevitably led to comparisons and questions concerning the consistency of their views, particularly with regard to Engels's claims concerning dialectics and science. These questions of interpretation could be answered either way, but most readers were encouraged to see a coincidence of views between the two and to keep the assumptions of orthodoxy intact. This was unsurprising given the political impetus to unity, and to validating a Marxist tradition, that successive scholarly and popular editions of 'Marx and Engels' reinforced. Engels himself, in his foundational 1859 review of Marx's *A Contribution to the Critique of Political Economy* stressed the uniqueness and power of Marx's method even at that early date, suggesting that Marx used both a historical and a logical method and that the two were consistent and coincident.

As the working-class uprisings of the 1920s were suppressed, and as the vacuity of Stalinist thought became evident, 'western' Marxism became more theoretical, more concerned with philosophical and methodological issues, more focused on intellectuals than political leaders, and more academically situated in universities and research institutes, most famously the Institute for Social Research at Frankfurt (in the 1930s and then in a post-Nazi diaspora). The grander claims of Engelsian synthesis were dropped in favor of multidisciplinary inquiry (taking in psychoanalysis, aesthetics, popular culture, imperialism, ethics, and morality, for example). Explanatory reductionism to class (meaning a person's relationship to the means of production) was replaced by a multifactor approach to causation, and an interpretative turn toward diagnosis and persuasion. This latter took place within a 're-Hegelianizing' of Marx, levering him out of an Engelsian frame of scientific causation, explanation, and prediction, and emphasizing his work in conceptualizing long-term phenomena and overarching structures in social development.

This signals an adaptive process whereby Marx has been successively reinterpreted for the social sciences as the intellectual climate, particularly in philosophy of science, has changed. Engels cast Marx in the shadow of Darwin's 'discoveries' and of a positivist account of natural science (doing considerable violence to Hegel and Darwin in the process). Later Marxists reinterpreted Marx in the light of neo-Kantian epistemologies and ethical doctrines, and in the light of the burgeoning empirical social sciences. After World War II Marxists such as Louis Althusser (1918–90) adapted Marx to structuralism, producing an account of history and social development that reduced human agency to a mere effect of economically driven and only partly predictable large-scale social forces. Althusser proposed to distinguish a nonscientific from a scientific Marx, and Marx was said to have attained this intellectually by making an epistemological 'break' and so expunging from his thought any trace of Hegelian idealism. Althusser's project was never satisfactorily completed and is noteworthy for its method, that of an intense and large-scale rereading of Marx, always attempting to fit his ideas into a binarized frame, for example, nonscience versus science, humanist versus materialist.

Perhaps unsurprisingly the contrary reading, that the apparently Hegelian, 'humanist' Marx of the early works is the valuable one, gathered strength at about the same time. Marx's 'Economic and Philosophical Manuscripts' of 1844 were published substantially for the first time in French and English in the late 1950s, and attracted considerable attention from scholars and readers who were intellectually or temperamentally repelled by the aridities of orthodox dialectics. Any social science presupposes various facts or truths about human beings (or 'man,' as he appeared in those days before feminist critique). The 'humanist' Marx was also a reflection of dissatisfaction with behaviorist models that pictured humans as reactive systems, responding to stimuli in predictable ways. The 1844 articles could be read in isolation from any other works by Marx, and frequently were. They appeared to be self-contained meditations on human nature and history, delineating a story of necessary and progressive interactions with nature leading ultimately to the high productivity of modern industry and the discoveries in the natural sciences that made this possible.

Marx's early articles present 'man' as an inherently social being, necessarily engaged in a sensuous 'metabolism' with nature, which is effectively an external but essential part of the body. Labor, in Marx's view, is not merely a vital activity for satisfying needs (as with animals), but rather in the case of humans a 'free conscious' activity that takes human life itself to be an object and fashions things 'according to the laws of beauty,' Humans produce even when free from physical need, and only truly produce when production is freely undertaken.

This claim is meant to have descriptive force, distinguishing self-conscious, and self-reflexive activity undertaken by humans from the life activities undertaken by animals, said by Marx to be merely self-identical with what they do. Humans, in his view, produce themselves in different ways, as different kinds of people (intellectually and physically) in different productive cultures. Communism, in this scheme, is a way of making this self-constructing process more explicit in terms of egalitarian decision making in society.

The inverse of communism that Marx portrays in his articles is the contemporary world of 'alienated' labor, in which the worker is estranged in terms of possession and power from the products of labor, other workers, nature, society at large, and from the quintessential human species-activity, namely creative work. This philosophical analysis sets a context for the consideration of private property as the contemporary system through which production and exchange take place, and Marx offers his account of 'man' as an alternative to the schemes developed in classic texts of political economy, most notably the highly individualized yet commercially minded hunters and fishermen portrayed in the work of Adam Smith (1723–90).

'Alienation' (Term used by Marx in early writings but not wholly abandoned later, when more specific terminology was employed. It characterizes work-relations within which workers and products, workers and productive activities, social relations among workers, and work as the defining human 'species-activity,' are all experienced in ways that are at odds with or separated from 'truly human' modes of fulfillment. Alienation was the hallmark of the 'humanist' revival of Marx in the early 1960s, and has been controversially deployed to distinguish a 'young' or 'early' Marx from a 'mature' Marx of the later works.) is used as a general term by Marx for detailing the specifics of class division in commercial societies, and in the early articles he endeavored to spell these out with respect to workers, and also to nonworkers (though he spent little time on this). The thrust of Marx's argument was toward the overthrow of capitalist society as in no one's interest as a human being. This was precisely because material and psychological deprivation contradicted the inherent potential of the human species-being. Notwithstanding the political and metaphysical tinge to Marx's exposition, alienation functioned in the 1960s and 1970s as an empirical catchall concept for sociological and psychological research into and evaluation of modern industrial processes.

Perhaps the most significant development in Marxist social science in the later twentieth century has been the discovery and appropriation of the work of Antonio Gramsci (1891–1937), an Italian communist trade union activist imprisoned by the Fascist government from 1926 to 1935. During his years in prison he composed a lengthy set of political notebooks, which gradually became available after World War II. His views have a distinct affinity with those of the early Lukács and other 'Hegelian' Marxists, at least to the extent that he departed decisively from the supposed certainties and mechanistic models of dialectical materialism. While he did not have access to Marx's early articles, again there is an affinity with the style of thought and political thrust of these works.

Gramsci was in any case facing the problem that a proletarian revolution, declared inevitable by orthodox Marxism, did not seem to be happening in a steady historical development. Far from class position translating easily into political radicalism, his analysis revealed that cultural conservatism was deeply ingrained in peasants and workers, and that inculcating class-conscious political activism would necessitate an ideological counteroffensive to this already existing 'hegemony.' His work inspired a Marxist social science deeply concerned with tracing the flow of power in society into channels that were not reducible, or not directly reducible, to class position and economic circumstances. Ideas, ideologies, and institutions all became much more important in constructing an analysis of contemporary class politics and a set of political strategies and tactics. While this may have seemed to deemphasize Marx's lifelong project – a critique of the economic categories – it actually mirrored the methodology and content of his writings on contemporary politics, notably *The Class Struggles in France* (1850) and *The Eighteenth Brumaire of Louis Bonaparte* (1852). This Gramscian reading of Marx, then, is another way that he has been reread and reinterpreted, taking a fresh look at which of his texts is most relevant to the times.

For the social sciences Marx and Marxism offer a rich and varied tradition of inquiry and debate, posing difficult questions about objectivity and neutrality in a social science, about what sort of knowledge counts as scientific, and about what sort of intellectual and practical activity science as a whole actually is. These questions may be answered in terms of contextualized readings of Marx, Engels, Marxism, and later interpreters. How directly those contextualizations map onto current circumstances is very much a matter of judgment by those interested in the social and behavioral sciences today, but there is no doubt that many would set a positive value on what a rereading of this tradition can contribute in the present. Indeed it is likely that most educators would count at least a knowledge of this tradition to be essential for any social scientist, whatever their interests. This itself is a tribute to the wide-ranging and as we would say today 'interdisciplinary' value of the legacy left by Marx.

Marx's major works, including the *Economic and Philosophical Manuscripts*, *The German Ideology*, the *Communist Manifesto*, *The Eighteenth Brumaire of Louis Bonaparte*, the *Grundrisse*, the three volumes of *Capital*, and the three volumes of *Theories of Surplus Value*, have all been published in numerous editions and translations and are widely available in English.

The definitive edition of the complete works of Marx and Engels published in their original languages is the renewed *Marx–Engels Gesamtausgabe* (known as MEGA). This was begun in 1975 under communist party sponsorship and published by Dietz Verlag in East Berlin. It is now being continued by an international consortium (IMES) based at the Berlin–Brandenburg Academy of Social Sciences and is being published by Akademie Verlag. The series includes all published and unpublished works, articles, letters to and from Marx and Engels, third party letters, and considerable bibliographical and contextual background in the notes and apparatus. Current plans are to complete the series in approximately 114 volumes within the next decade.

A vast and scholarly selection from the Marx–Engels materials is published in English in a set of 50 volumes, begun in 1975 and completed in 2004. This is the *Collected Works* from Lawrence & Wishart (London).

See also: Bourgeoisie and Middle Classes, History of; Capitalism; Class: Social; Economics, History of; Field Theory; Marxism in Contemporary Sociology; Marxism–Leninism: The Ideology of Twentieth-Century Communism; Socialism: Historical Aspects; Socialism; Working Classes, History of.

Bibliography

Carver, T., 1989. Friedrich Engels: His Life and Thought. Macmillan, Basingstoke, UK.

Engels, F., 1980. [Review of] Karl Marx, a contribution to the critique of political economy (1859). In: Marx, K., Engels, F. (Eds.), Collected Works, vol. 16. Lawrence and Wishart, London, pp. 465–477.

Engels, F., 1987. Anti-dühring. In: Marx, K., Engels, F. (Eds.), Collected Works, vol. 25. Lawrence and Wishart, London, pp. 5–309.

Engels, F., 1989. Karl Marx's funeral (1883). In: Marx, K., Engels, F. (Eds.), Collected Works, vol. 24. Lawrence and Wishart, London, pp. 467–471.

Engels, F., 1992. Letter to Eduard Bernstein, 2–3 November 1882. In: Marx, K., Engels, F. (Eds.), Collected Works, vol. 46. Lawrence and Wishart, London, pp. 353–358.

Gamble, A., Marsh, D., Tant, T. (Eds.), 1999. Marxism and Social Science. Macmillan, Basingstoke, UK.

Lenin, V.I., 1964. Karl Marx: a brief biographical sketch with an exposition of Marxism (1918). In: Lenin, V.I. (Ed.), Collected Works, fourth ed., vol. 21. Foreign Languages Publishing House, Moscow, pp. 43–91.

McLellan, D., 1973. Karl Marx: His Life and Thought. Macmillan, London.

Marx, K., Engels, F., 1976. Communist manifesto. In: Marx, K., Engels, F. (Eds.), Collected Works, vol. 6. Lawrence and Wishart, London, pp. 477–519.

Thomas, P., 2008. Marxism and Scientific Socialism: From Engels to Althusser. Routledge, Milton Park.

Marxism in Contemporary Sociology

Matt Vidal, King's College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Burawoy, E.O. Wright, volume 14, pp. 9295–9298, © 2001, Elsevier Ltd.

Abstract

This article treats marxism as an ongoing and evolving research program based on a number of questions, categories of analysis, and theses specified by Karl Marx. Marxism constitutes a vibrant research program that spans all of the social science disciplines and many of the humanities. This article briefly discusses the assimilation of marxism into sociology and the use of marxism within sociology, before turning to the project of building the marxist research program, with a particular emphasis on contributions by sociologists. The research program can be summarized in terms of seven theses, which fall under three theoretical clusters: (1) the social construction of reality, (2) historical materialism, and (3) the contradictory and problematic reproduction of capitalism.

Karl Marx's work combined a distinctive political philosophy, building on the body of Western philosophy from Plato to Hegel, with a theory of political economy building on and singularly advancing the entire corpus of political economy to his day. The result was a grand theory of human history that provided foundational building blocks for the emerging discipline of sociology.

In the twenty-first century, marxism continues to guide vibrant research traditions in sociology, political economy (e.g., Baran and Sweezy, 1966; Bowles and Gintis, 1976; Gordon, 1976; Foley, 2000; Lebowitz, 2003; Duménil and Lévy, 2004; Brenner, 2005; Kliman, 2007; Foster and Magdoff, 2009; Moseley, 2009; Lapavitsas, 2011; Amin, 2013; Aglietta, [1979] 2000), political science (e.g., Wood, 2003; Panitch and Gindin, 2004; Callinicos, 2009a), political theory (e.g., Ollman, 1977; Callinicos, 2009b), philosophy (e.g., Lukács, 1971; Bhaskar, 1979; Mészáros, 1995; Hardt and Negri, 2000; Lazzarato, 2006; Smith, 2013), management (e.g., Adler, 2007; Böhm and Land, 2009), economic geography (e.g., Harvey, 1982), ecology (e.g., Foster, 1999; Clark and York, 2005), history (e.g., Anderson, 1974; Wallerstein, 1974; Brenner, 1976; Hobsbawm, 1994), literary criticism (e.g., Eagleton, 1976; Jameson, 1991), and media studies (e.g., Caraway, 2011). As evidence of such vibrancy, one need only consider the remarkable range of academic journals that were founded by marxists and continue to provide outlets for marxist social science and humanities:

- *Science and Society: A Journal of Marxist Thought and Analysis* (1936–present)
- *Monthly Review: An Independent Socialist Magazine* (1949–present)
- *New Left Review* (1960–present)
- *Socialist Register* (1964–present)
- *Telos* (1968–present)
- *Antipode: A Radical Journal of Geography* (1969–present)
- *Critical Sociology* (1969–present)
- *Capital & Class* (1977–present)
- *International Socialism: A Quarterly Journal of Socialist Theory* (1978–present)
- *Socialism and Democracy* (1985–present)
- *Rethinking Marxism: A Journal of Economics, Culture & Society* (1988–present)
- *Capitalism, Nature, Socialism* (1988–present)
- *Journal of World-Systems Research* (1995–present)
- *Cultural Logic: An Electronic Journal of Marxist Theory and Practice* (1997–present)
- *Historical Materialism* (1997–present)
- *International Socialist Review* (1997–present)
- *The International Marxist-Humanist* (2000–present)
- *Mediations: Journal of the Marxist Literary Group* (2007–present)
- *New Proposals: Journal of Marxism and Interdisciplinary Inquiry* (2007–present)

As the foregoing might suggest, it is difficult to specify what would distinguish 'marxist sociology' from other forms of academic marxism. Marx's research program, known as historical materialism, was explicitly conceived as a unified theory covering all of the modern social science disciplines and many within the humanities. Many subsequent marxists have – even when trained within particular disciplines – continued to work within the fundamentally transdisciplinary historical materialist research program. At the same time, many core marxist questions and categories have become assimilated into sociology, and marxist sociologists have often explicitly used marxism to engage disciplinary questions addressed by competing or complementary theoretical frameworks. This article briefly discusses the assimilation of marxism into sociology and the use of marxism within sociology, before turning to the project of building the marxist research program, with a particular emphasis on contributions by sociologists.

Assimilating Marxism into Sociology

Most sociologists, even those relatively unsympathetic to the theoretical and political agenda of marxism, recognize that the marxist tradition has been the source of a central theoretical current within sociology. Marx, Weber, and Durkheim are widely considered to be the founding fathers of sociology. Durkheim is identified with norms and problems of social integration, Weber with rationalization and the cultural sociology of meaningful action, and Marx with class, power, and conflict. Studies of politics routinely borrow from the marxist tradition a concern with the influence of business on the state, broader economic constraints on state action, and the class bases of political parties and political mobilization. The sociology of

work has been largely defined by marxist concerns with the problem of extracting effort from workers, managerial control systems, worker resistance, and the impact of technology on skills. The marxist concept of contradictions is used in discussions of social change. Theories of social conflict are influenced by the core marxist idea that conflicts are generated by structurally based social cleavages, not simply subjective identities. Marx's discussions of the relation between material social relations, ideology, and consciousness played a fundamental role in the development of the theory that reality is socially constructed (Berger and Luckmann, 1966). In these and many other ways, themes integral to the marxist tradition have been imported into the mainstream of sociological scholarship, frequently losing any explicit link to their marxist pedigree in the process.

Using Marxism within Sociology

Many marxists have engaged in a more self-conscious, more ambitious use of marxist ideas within sociology. Here the idea is to take the conceptual framework elaborated within the marxist tradition – mode of production, exploitation, the labor process, class structure, class struggle, class consciousness, the state, ideology, revolution, and so on – and to use these concepts to understand a wide array of sociological problems. The goal is for marxist concepts to challenge the mainstream with an alternative set of explanations and predictions, rather than lose their identity by being absorbed into the mainstream. In studies of political elites, the claim is not just that every society has a 'governing elite,' entry to which is based ultimately on merit, but that ownership of the means of production provides the material basis for a relatively cohesive and powerful 'ruling class' (e.g., Bottomore, 1964). Similarly, in studies of the state, the claim is not just that there may be business influence and economic constraints on state policies, but that class power sets fundamental limits on the range of possibilities of state action (e.g., Miliband, 1969; Offe, 1984). In studies of corporations, the claim is that the diffusion of ownership among shareholders has not led to a separation of ownership from control, hence destroying the old class structure of workers and owners, but that groups of proprietary owners are still able to realize their corporate objectives by maintaining controlling shares across corporations and investment banks (e.g., Zeitlin, 1974). In studies of work organization and technology, the claim is not just that employers face a problem of gaining cooperation from workers and getting them to perform with adequate effort, but that the antagonism of interests between workers and employers is a fundamental property of work organization in capitalist economies (e.g., Braverman, 1974). In studies of inequality, the claim is not just that class is one of the salient dimensions of social inequality in industrial societies, but that class relations structure fundamental features of the system of inequality (e.g., Wright, 1985).

Building the Marxist Research Program

To understand the tasks involved in building the marxist research program, it is necessary to briefly lay out the core substantive arguments of the marxist tradition. Such an exercise will always be contested because Marx's voluminous *oeuvre* was evolving and unfinished. Marx's thought evolved over the three decades in which he wrote and many of his writings consist in unfinished and unpublished manuscripts, often in fragments. Additionally, while Marx typically mixed academic analysis and political writing in the same texts, some of these – perhaps most notably *The Communist Manifesto* – were heavily political documents that clearly presented extreme simplifications of his extensively elaborated and deeply nuanced theoretical analysis (the Penguin versions of *Capital, vols. 1–3* total 2851 pages and *The Grundrisse* 904 pages).

This article thus treats marxism as an ongoing and evolving research program based on a number of questions, categories of analysis, and theses specified by Marx. The research program can be summarized in terms of seven theses, which fall under three theoretical clusters: (1) the social construction of reality, (2) historical materialism, and (3) the contradictory and problematic reproduction of capitalism.

The Social Construction of Reality

Marx developed a basic sociological theory of social structure and human agency, perhaps the earliest version of a theory of the social construction of reality. The core of this theory can be stated in terms of one ontological presupposition and two theses.

Ontological presupposition: The social world consists of surface appearances and underlying reality, in which the former typically mystify the latter.

Plato introduced the argument – most famously in his allegory of the cave – that reality is not what we observe directly but exists in nonmaterial, abstract forms. Marx accepted the distinction between appearance and reality, but inverted Plato's idealism, positing that underlying reality consists of material social relations (Wolff, 1988). Appearances, in Marx's version, are manifestations of material reality, but distorted ones. In Marx's ([1894]1981) own terms, "All science would be superfluous if the outward appearance and the essence of things directly coincided." The classical economists focused only on exchange, misrepresenting the market as a new realm of equal exchange and missing the fundamental importance of 'the hidden abode of production.' Their theories posit a world of individuals and things (commodities), thus failing to appreciate how social relations constitute individuals and produce commodities. "The individual and isolated hunter fisherman, with whom Smith and Ricardo begin," Marx ([1857–1858]2000: p. 83) observed, "belongs among the unimaginative conceits of the eighteenth-century Robinsonades." In order to flesh out this presupposition, to develop it into a scientific theory, Marx articulated a thesis on the social construction of reality, which is most easily understood in terms of two subtheses.

Thesis 1a: Humans produce the social world, including individual consciousness of it, through labor.

Thesis 1b: Human agency drives historical change, but individual consciousness is shaped by institutional context.

In a preface to an otherwise less-influential work, Marx ([1859]1977) wrote that social consciousness in any society 'corresponds' to the political and ideological 'superstructure,'

which 'arises' on top of the relations of production. This has often been read to suggest that the economic structure of society determines the political and cultural superstructure, but a wider reading of Marx shows that for him, even the 'material forces of production' were seen as a social and historical product (Nicolaus, 1968). Elsewhere Marx wrote forcefully that in the act of labor, understood in its broadest sense, humans produce not simply material wealth but themselves as human subjects, and therefore society as a whole (Marx and Engels, [1846]1996). Read through this lens, the distinction between base and superstructure is an analytical distinction of the process by which the social world – politics, culture, ideology – is produced *with and through* the core of human activity, labor.

Thesis 1a is best understood, then, as the proposition that cultural understandings and political institutions are *coproduced* with human economic activity. While consciousness and culture (and politics) correspond to the economic structure of society, then, this should not be reduced to an argument that particular material relations mechanically generate particular forms of thought, let alone specific contents of thought. The corollary argument (Thesis 1b) is that as the ideological forms – ways of understanding the world, conceptual frameworks, and cognitive institutions – are transmitted across generations, they come to preexist particular individuals, who are socialized in the context of existing ideologies (see Thesis 5 below for an elaboration in terms of class structure and consciousness). As Marx wrote in a justly famous passage:

> Men make their own history, but they do not make it just as they please; they do not make it under circumstances chosen by themselves, but under circumstances directly encountered, given and transmitted from the past. The tradition of all the dead generations weighs like a nightmare on the brain of the living.
>
> Marx, [1852]1994

Marx's theory of social constructionism includes the additional thesis that social relations are often constructed in inherently contradictory ways.

Thesis 2: The material social relations that humans produce are often constituted, in part, by dynamic contradictions.

As he did with Plato's idealist distinction between appearance and essence, Marx also inverted Hegel's idealist notion that thought develops through the progressive working out of the inherently contradictory nature of categories. For Marx, structural contradictions – internal antagonisms – exist within the material relations that humans socially construct. Specifically, a structural contradiction exists where two social entities are constituted as part of a mutually interdependent relationship and these two entities are potentially in conflict (Callinicos, 2009b). Where structural contradictions exists social relations are said to be dialectical, that is, each social entity in the social relation has causal impacts on the other. The two primary contradictions of any society (Thesis 3 below) are between (1) the forces and relations of production and (2) classes. These primary contradictions provide a fundamental source of dynamism in human society (Thesis 4) and give rise to a number of secondary contradictions (Thesis 7).

Historical Materialism

Marx's theory of history consists of three core theses. The first thesis is most easily understood in terms of three intimately connected subtheses.

Thesis 3a: All existing societies have been fundamentally structured by a mode of production, constituted by the dialectical interaction between forces of production and relations of production.

Thesis 3b: Primary contradiction I: The relations of production have historically constituted an antagonistic class structure objectively based on the relations of producers and nonproducers.

Thesis 3c: Primary contradiction II: The forces of production will tend to develop and will increasingly conflict with the relations of production.

Human history develops in terms of modes of production. The forces of production include the means of production (technology and raw materials) and labor power. The relations of production refer to (1) the relations of ownership over the means of production and (2) the mode of appropriation of surplus labor of direct producers (slaves, serfs, wage laborers) from nonproducing owners (slave owners, feudal lords, capitalists). The class structure is objectively determined by the two aspects of the relations of production, which vary across modes of production: (1) ownership of the means of production, which is a fundamental source of power in society, and (2) relations between direct producers and nonproducers in society.

Under feudalism, serfs owned some means of production (subsistence farming tools and perhaps a small amount of land) but not enough land to be self-sufficient. Land-owning feudal lords were the nonproducing class, which appropriated surplus labor from serfs in the form of rent, based on extra-economic feudal obligations. Under capitalism, there are a number of ways in which classes become internally differentiated and contradictory class locations emerge (Wright, 1985). The basic class structure remains one in which workers have been dispossessed of any means of production and capitalists, as nonproducing owners, extract surplus labor of wage laborers, nonowning direct producers, through the economic mechanism of the wage. The latter is a mystified relation which appears to compensate for an entire period of labor but actually compensates only for the labor necessary to reproduce the workers (Thesis 6; Burawoy, 1979; Marx, [1867] 1990).

Thesis 3c submits that the forces of production have a tendency to develop, due to attempts of direct producers to adopt labor-saving innovations (Callinicos, 2009b), which, eventually, will come into conflict with existing relations of production, leading to an intensification of this fundamental contradiction (see Thesis 7). Ultimately, pressure will be put on the existing class structure as its fettering of the productive forces becomes increasingly problematic, and as emergent classes are brought forth due to the changing economic structure of society.

Thesis 4: The intensification of the primary contradictions will force attempts to mitigate or displace them, the outcome of which will depend on the state of the class struggle.

So-called orthodox marxism is defined by the expectation that the intensification of the primary contradictions will

inevitably lead to the overthrow of the bourgeoisie by the proletariat and the establishment of a socialist society (for the most sophisticated articulation of this, see Cohen, 1978). This position is based on a narrow reading of Marx found in a political pamphlet (Marx and Engels, [1848]1948) and crisp but brief preface to a later work (Marx, [1859]1977). A less deterministic reading is as follows. The contradiction between the forces and relations of production has an inherent tendency to intensify. There may be various forms of temporary class compromise, but the class struggle will tend to develop a number of associated secondary contradictions rooted in the primary contradictions that will produce crises (Thesis 7a) and thus require attempts to mitigate or displace the crisis tendencies (Thesis 7b). Such displacement or mitigation may take many forms through various spatial and institutional fixes (Harvey, 1982; Aglietta, [1979]2000), but these will always be temporary as the primary structural contradictions can never been completely eliminated. Actual outcomes will ultimately depend on the class struggle (Brenner 1976; Chibber 2011), fought on the terrains of politics and ideology as well as economic relations. The role of ideology is central to marxist class analysis.

Thesis 5: The relationship between objective class relations and individual consciousness is mediated by ideology.

Marx wrote about ideology in at least two different ways. The first is the notion of false consciousness. Here the distinction is between a real, scientific grasp of reality, versus a distorted or incorrect, ideological understanding. However, the more influential reading has focused on Marx's second conception of ideology as the entire medium of consciousness and meaningfulness. This second reading has been developed in two primary, and not inconsistent, ways. The Italian theorist Antonio Gramsci developed ideology in terms of the concept of hegemony, referring to the political and moral leadership by which the capitalist class incorporates the working class into society. Hegemony is an attempt to present the interests of the ruling class as the universal interests of society but, crucially, to be effective it must have a material basis in the 'concrete coordination of interests' between the ruling class and subordinate groups (Gramsci, 1999[1929–1935]: pp. 182, 161). The French philosopher Louis Althusser ([1970]2001) argued that ideology constitutes individuals as conscious actors and patterns their behavior through discourse, conceptual frames, categories, and images.

All ideologies have some basis in material social relations, although ideology itself has discursive as well as nondiscursive determinants (Therborn, 1980). Language is the medium through which material things are represented, but language is multireferential: different meanings can be constructed around the same material relations, such as the market and the labor process, although some representations will be more distorted (in the sense of being 'one-sided') than others (Hall, 1986). Thus, although class position does not determine one's consciousness, ideologies – including hegemonic class ideologies – provide representations of material structures, including class structure. Class ideologies interact in complex ways with nonclass ideologies, most importantly gender, religion, nationality, and race. Material relations of patriarchy, for instance, give rise to gender ideologies. Capital accumulation becomes accommodated to patriarchal relations in a way that tends to perpetuate the latter, although in modified form – for instance, in terms of the early- to mid-twentieth century family model of male breadwinning in the labor market and female child rearing and homemaking (Hartmann, 1979). Understanding these complex social processes requires careful attention to ideology and its material bases. Indeed, it has been argued that non-marxist theories of the intersectionality between class, gender, and race have amounted to an endless series of descriptive empirical studies of their combined effects in specific contexts, without explaining the underlying sources of such inequalities (Gimenez, 2001). Gender, race, and nationality surely have their own material bases, but they interact with class in a macroinstitutional system where class power remains fundamental insofar as it is economic power, however much gender/race/nation modify this and are irreducible to it (on nationality and ethnicity, see Anderson, 2010).

The Contradictory and Problematic Reproduction of Capitalism

The contradictory nature of capitalism and its problematic reproduction can be stated in terms of two theses, the second of which consists of two subtheses.

Thesis 6: The output of capitalist production consists of commodities, whose underlying values – which can take the form of both use-values and exchange-values – are objectively determined by the minimum socially necessary labor time required to produce them, with the primary source of profit being the surplus labor of workers.

This proposition develops the ontological assumption about the distinction between appearance and reality into a complex and often-misunderstood thesis on the value forms of capitalism. Capitalist societies produce wealth in the form of commodities, which have use-value (when consumed) and exchange-value (when traded for other use-values). Exchange-value is necessarily the form in which value is expressed (and appears). Capitalist production is not driven by consumption (use-values), as Adam Smith and his followers would have it, but by capitalists' desire for accumulation (exchange-values). Capitalists pay workers only the equivalent of their necessary labor time, that is, the amount of time necessary for a worker to produce his or her own livelihood according to the basic standard of living of the era. Any and all additional working time – surplus labor time – is the source of profit. The underlying value of a commodity is determined by the labor time socially necessary to produce it: "the labor-time required to produce any use-value under the conditions of production normal for a given society and with the average degree of skill and intensity of labor prevalent in that society" (Marx [1867] 1990: p. 129).

Although marxist value theory is seen by many critics primarily as a theory of price determination, it is more fundamentally a social structural theory of how labor is valued under capitalism. Marx used the example of the power loom, which, he estimated, cut by half the labor time necessary to weave a certain quantity of yarn into cloth. Capitalist competition thus forced wages down to the new level of socially necessary labor-time, and as a result cut prices accordingly. However, the critics have succeeded in narrowing marxist value

theory to a technical question of the relation of money prices to labor values, which is mathematically complex and contested, and has led some marxists to reject the thesis, and many left academics to reject marxism altogether. Yet, in Marx's own work the question of prices was secondary to the larger theory of value relations and the fetishism of commodities. Commodities *appear* to have an intrinsic value in terms of their money price, independent of any social relations. But the world of money prices is a very bizarre world – the value of commodities was represented by classical political economy in terms of the amount of labor congealed within them – which mystifies the underlying social relations that value labor and produce profit (Wolff, 1988).

In terms of prices, socially necessary labor-time sets limits on the movement of prices; since they may fluctuate around underling values for a number of reasons, the relationship between labor-values and product-prices is probabilistic. Marxist economists continue to defend and elaborate the labor theory of value in terms of the relation between values and prices (e.g., Foley, 2000; Kliman, 2007), although these developments are beyond the scope of this article. In terms of the broader labor theory of value, the distinction between use-value and exchange-value – rooted in the contradiction between the forces and relations of production – is the basis for a number of secondary contradictions, to which we now turn.

Thesis 7a: Capitalism is a fundamentally dynamic political economic system that is progressive insofar as it greatly enhances human productive capacity, but has inherent tendencies toward disequilibrium and crisis.

Thesis 7b: Capitalism thus requires spatial and institutional fixes to secure temporary periods of stability and ensure its (expanded) reproduction.

It is often forgotten by critics that Marx was among the first scholars to note and praise the fundamental dynamism and progressivity of capitalism, in particular the extent to which it enhances the productive capacity of human society. While the orthodox marxist position predicts the inevitable decline of capitalism, a more nuanced, less teleological formulation is that (Thesis 7a) due to its inherent contradictions, capitalism is systemically characterized by disequilibrium and regular episodes of crisis and institutional decay. The corollary (Thesis 7b) is that the contradictory social relations of capitalism require various spatial and institutional fixes for their reproduction (Harvey, 1982; Aglietta, [1979]2000). Marx identified four crisis tendencies within capitalism.

The central crisis tendency is toward overaccumulation, which arises from the contradictions between the forces and relations of production (and between use-value and exchange-value). Competing capitalists tend to develop the forces of production without regard for the limits of the market (Clarke, 1988). Fueled by credit, capitalists will attempt to overcome the limit of the market by expanding production, and some will be successful in expanding it by making cheaper products or creating new ones, but the economy will always hit a new limit as innovations are generalized. This results in periodic crises of over-accumulation – a surplus of capital with a lack of sufficient outlets for investment – in which capital is devalued, capacity destroyed, and unemployment increased through bankruptcies, mergers, and acquisitions (Harvey, 1982; Foster and Magdoff, 2009). Such crises set the stage for renewed profitability and accumulation. This primary contradiction also generates secondary contradictions in the labor process. The labor process, cooperation to create a product, is part of the productive forces. The valorization process, the extraction of surplus labor to produce a profit, is part of the relations of production. While valorization pressures in the form of competition among capitalists stimulate the development of the productive forces, they also have detrimental effects on the labor process, because capitalists may have to sacrifice some efficiency in order to ensure control of labor (Gordon, 1976), and on the workforce, because the pursuit of profit takes place without regard for worker well-being (Adler, 2007).

Capitalism generates three additional crisis tendencies: a declining profit rate, underconsumptionism, and metabolic rift. First, as capitalists seek to compete by substituting machines for labor (the sole source of new value), they put downward pressure on profit rates (Kliman, 2012). This also generates a secondary contradiction in the labor market, as the systematic substitution of capital for labor means that technical change destroys more jobs than it creates, making paid work increasingly unable to provide the basis for improving the livelihood of the whole population (Aronowitz and DiFazio, 1994). Further downward pressure on the profit rate may be generated if labor succeeds in its demands for higher wages, as happened across the OECD in the late 1960s following the postwar class compromise. However, when capital gains the upper hand in open class struggle, as happened following the late 1960s profit squeeze, the second crisis tendency of underconsumptionism will become manifest. These crisis tendencies lead capitalists and capitalist state managers to adopt a number of strategies to offset or displace them. The role of the state is critical because market forces alone cannot secure the reproduction of capitalism, which requires extra-economic forms of regulation (Jessop, 2002). However, this gives rise to a secondary contradiction between the national form of the state the inherently global nature of capitalist accumulation, leading, for instance, to various well-known crises of Keynesianism (Clarke, 1988).

Among the oldest strategies to overcome the limit of the market and displace crises is the spatial fix of expanding markets. Indeed, capitalism has always been a world economy, in which states in the core of the world economy appropriate surplus from peripheral states, with semiperipheral states providing a form of political stability (Wallerstein, 1979). The history of capitalism as a world system has occurred in four systemic cycles of accumulation, in which a hegemonic state leads in terms of a material expansion and then shifts to a financial expansion, signaling its decline and the rise of a new hegemon (Arrighi, 1999). The Genoese cycle ran from the fifteenth to the early seventeenth century, the Dutch cycle from the late sixteenth through to the eighteenth, the British cycle from the latter half of the eighteenth through to the early twentieth century, and the US cycle from the late nineteenth century through its current phase of financialization.

Financialization in the US since the 1970s – the increasing role of financial activities in generating profit – has grown out of the general condition of overaccumulation: a surplus of

capital and a lack of sufficient outlets for productive investment (Foster and Magdoff, 2009). This, in turn, is the fundamental basis of the 2007–08 financial crisis, as rampant speculation in the face of declining investment opportunities in the real economy led to the development of a raft of esoteric financial instruments, a housing bubble, and predatory mortgage lending. These crisis tendencies were contained in the US and Western Europe from 1945 to around 1970 through a fordist institutional infrastructure (Jessop, 2002; Aglietta, [1979] 2000), consisting of the Bretton Woods system – fixed exchange rates and capital controls, which allowed national policy autonomy for Keynesian welfare states – along with states internally structured in terms of producer-driven supply chains and oligopolistic competition in core sectors. This fordist structural context allowed various national high-wage class comprises (Vidal, 2013). The institutional fix of the fordist accumulation regime was based in the ability of national states to limit the downward pressure exerted on wages by global competition. As this fix eroded under financialization, internationalization, and neoliberalism, the contradictions of capitalism have reasserted themselves with a vengeance. There is no clear fix to the twin problems of rising inequality and stagnant growth affecting the OECD economies.

The fourth crisis tendency is ecological. Marx discussed ecological sustainability in terms of a metabolic relation between humans and nature, arguing that large-scale capitalist agriculture created a metabolic rift between humans and the soil, as capitalism was unable to maintain sufficient recycling of soil nutrients (Foster, 1999). The rift is most developed today in terms of the carbon cycle, which has been ruptured by the accumulation of CO_2 in the atmosphere, overwhelming the capacity of natural sinks to absorb the additional carbon produced under capitalist production (Foster et al., 2011). It remains unclear how the ecological rift can be fixed within a growth-driven, capitalist economy.

See also: Alienation, Sociology of; Capitalism: Global; Class Consciousness; Class: Social; Marx, Karl (1818–83); Socialism; Sociological Theory.

Bibliography

Adler, Paul S., 2007. The future of critical management studies: a Paleo-Marxist critique of labour process theory. Organization Studies 28, 1313–1345.

Aglietta, Michel, [1979]2000. A Theory of Capitalist Regulation: The US Experience. Verso Classics, London.

Althusser, Louis, [1970]2001. Ideology and ideological state apparatuses. In: Althusser, L. (Ed.), Lenin and Philosophy and Other Essays. Monthly Review Press, New York.

Amin, Samir, 2013. The Implosion of Contemporary Capitalism. Monthly Review Press, New York.

Anderson, Kevin, 2010. Marx at the Margins: On Nationalism, Ethnicity, and Non-Western Societies. University of Chicago Press, Chicago.

Anderson, Perry, 1974. Lineages of the Absolutist State. New Left Books, London.

Aronowitz, Stanley, DiFazio, William, 1994. The Jobless Future: Sci-Tech and the Dogma of Work. University of Minnesota Press, Minneapolis.

Arrighi, Giovanni, 1999. The Long Twentieth Century. Verso, London.

Baran, Paul A., Sweezy, Paul M., 1966. Baran and Sweezy's Monopoly Capital. Monthly Review Press, New York.

Berger, Peter L., Luckmann, Tomas, 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.

Bhaskar, Roy, 1979. The Possibility of Naturalism: A Philosophical Critique of the Contemporary Human Sciences. Humanities Press, Atlantic Highlands, NJ.

Böhm, Steffen, Land, Chris, 2009. No measure for culture? Value in the new economy. Capital & Class 33, 75–98.

Bottomore, Thomas B., 1964. Elites and Society. Penguin, Harmondsworth, England.

Bowles, Samuel, Gintis, Herbert, 1976. Schooling in Capitalist America: Educational Reform and the Contradictions of Economic Life. Basic Books, New York.

Braverman, Harry, 1974. Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century. Monthly Review Press, New York.

Brenner, Robert, 1976. Agrarian class structure and economic development in pre-industrial Europe. Past and Present 70, 30–75.

Brenner, Robert, 2005. The Economics of Global Turbulence. Verso, London.

Burawoy, Michael, 1979. Manufacturing Consent: Changes in the Labor Process Under Monopoly Capitalism. University of Chicago Press, Chicago.

Callinicos, Alex, 2009a. Imperialism and Global Political Economy. Polity, Cambridge, UK.

Callinicos, Alex, 2009b. Making History: Agency, Structure, and Change in Social Theory. Haymarket, Chicago.

Caraway, Brett, 2011. Audience labor in the new media environment: a Marxian revisiting of the audience commodity. Media, Culture & Society 33, 693–708.

Chibber, Vivek, 2011. What is living and what is dead in the Marxist theory of history. Historical Materialism 19.2, 60–91.

Clark, Brett, York, Richard, 2005. Carbon metabolism: global capitalism, climate change, and the biospheric rift. Theory and Society 34, 391–428.

Clarke, Simon, 1988. Keynesianism, Monetarism and the Crisis of the State. Edward Elgar, Aldershot, UK.

Cohen, Gerald A., 1978. Karl Marx's Theory of History: A Defence. Princeton University Press, Princeton.

Duménil, Gérard, Lévy, Dominique, 2004. Capital Resurgent: Roots of the Neoliberal Revolution. Harvard University Press, Cambridge, MA.

Eagleton, Terry, 1976. Marxism and Literary Criticism. University of California Press, Berkeley.

Foley, Duncan K., 2000. Recent developments in the labor theory of value. Review of Radical Political Economics 32, 1–39.

Foster, John B., 1999. Marx's theory of metabolic rift: classical foundations for environmental sociology. American Journal of Sociology 105, 366–405.

Foster, John B., Clark, Brett, York, Richard, 2011. The Ecological Rift: Capitalism's War on the Earth. Monthly Review Press, New York.

Foster, John B., Magdoff, Fred, 2009. The Great Financial Crisis: Causes and Consequences. Monthly Review Press, New York.

Gimenez, Martha E., 2001. Marxism, and class, gender, and race: rethinking the trilogy. Race, Gender & Class 8, 23–33.

Gordon, David M., 1976. Capitalist efficiency and socialist efficiency. Monthly Review 28, 19–39.

Gramsci, Antonio, 1999[1929–1935]. Selections from the Prison Notebooks (Q. Hoare, G.N. Smith, Trans.). International Publishers, New York.

Hall, Stuart, 1986. The problem of ideology: marxism without guarantees. Journal of Communication Inquiry 10, 28–44.

Hardt, Michael, Negri, Antonio, 2000. Empire. Harvard University Press, Cambridge, MA.

Hartmann, Heidi I., 1979. The unhappy marriage of marxism and feminism: towards a more progressive union. Capital & Class 3, 1–33.

Harvey, David, 1982. The Limits to Capital. Basil Blackwell, Oxford.

Hobsbawm, Eric, 1994. Age of Extremes: The Short Twentieth Century: 1914–1991. Abacus, London.

Jameson, Fredric, 1991. Postmodernism, or, the Cultural Logic of Late Capitalism. Duke University Press, Durham.

Jessop, Bob, 2002. The Future of the Capitalist State. Polity Press, Cambridge, UK.

Kliman, Andrew, 2007. Reclaiming Marx's 'Capital': A Refutation of the Myth of Inconsistency. Lexington Books, Lanham, MD.

Kliman, Andrew, 2012. The Failure of Capitalist Production: Underlying Causes of the Great Recession. Pluto Press, London.

Lapavitsas, Costas, 2011. Theorizing financialization. Work, Employment and Society 25, 611–626.

Lazzarato, Maurizio, 2006. Immaterial labor. In: Virno, P., Hardy, M. (Eds.), Radical Thought in Italy: A Potential Politics. University of Minnesota Press, Minneapolis.

Lebowitz, Michael A., 2003. Beyond Capital: Marx's Political Economy of the Working Class. Palgrave, Houndmills, UK.

Lukács, Georg, 1971. History and Class Consciousness: Studies in Marxist Dialectics. MIT Press, Cambridge, MA.

Marx, Karl, [1852]1994. The Eighteenth Brumaire of Louis Bonaparte. International Publishers, New York.
Marx, Karl, [1857–1858]2000. Grundrisse. Penguin, London.
Marx, Karl, [1859]1977. A Contribution to the Critique of Political Economy. Progress Publishers, Moscow.
Marx, Karl, [1867]1990. Capital, vol. 1 (B. Fowkes, Trans.). Penguin, London.
Marx, Karl, [1894]1981. Capital, vol. 3. Penguin, London.
Marx, Karl, Engels, Frederick, [1846]1996. The German Ideology. International Publishers, New York.
Marx, Karl, Engels, Frederick, [1848]1948. The Communist Manifesto. International Publishers, New York.
Mészáros, István, 1995. Beyond Capital: Toward a Theory of Transition. NYU Press, New York.
Miliband, Ralph, 1969. The State in Capitalist Society. Basic Books, New York.
Moseley, Fred, 2009. The US economic crisis: causes and solutions. International Socialist Review 64.
Nicolaus, Martin, 1968. The unknown Marx. New Left Review I/48, 41–61.
Offe, Claus, 1984. The Contradictions of the Welfare State. MIT Press, Cambridge, MA.
Ollman, Bertell, 1977. Alienation: Marx's Conception of Man in Capitalist Society. Cambridge University Press, Cambridge, UK.
Panitch, Leo, Gindin, Sam, 2004. Global Capitalism and American Empire. Merlin, London.
Smith, Tony, 2013. The 'general intellect' in the grundrisse and beyond. Historical Materialism 21, 235–255.
Therborn, Göran, 1980. The Ideology of Power and the Power of Ideology. Verso, London.
Vidal, Matt, 2013. Postfordism as a dysfunctional accumulation regime: a comparative analysis of the USA, the UK and Germany. Work, Employment and Society 27, 451–471.
Wallerstein, Immanuel, 1974. The Modern World System I: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century. Academic Press, San Diego.
Wallerstein, Immanuel, 1979. The Capitalist World-Economy. Cambridge University Press, Cambridge, UK.
Wolff, Robert P., 1988. Moneybags Must Be So Lucky: On the Literary Structure of Capital. University of Massachusetts Press, Amherst.
Wood, Ellen M., 2003. Empire of Capital. Verso, London.
Wright, Erik O., 1985. Classes. Verso, London.
Zeitlin, Maurice, 1974. Corporate ownership and control: the large corporations and the capitalist class. American Journal of Sociology 79, 1073–1119.

Marxism–Leninism: The Ideology of Twentieth-Century Communism

William J Morgan, University of Nottingham, Nottingham, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by S.E. Hanson, volume 14, pp. 9298–9302, © 2001, Elsevier Ltd.

Abstract

The article considers Marxism–Leninism as a political ideology, its foundations as developed by Karl Marx (1818–83) and F. Engels (1820–95), its characteristics as established by V.I. Lenin (1870–1924), and its consolidation by J.V. Stalin (1879–1953) between 1922 and 1953 through 'socialism in one country' and the Communist International (Comintern). The article also considers critics of Marxism–Leninism, the transition from Comintern to the Soviet agencies of the Cold War, and support for national liberation movements. The final section considers the condemnation of Stalin by the Communist Party of the Soviet Union, attempts at revisionism, the impact of Gorbachev, the collapse of the Soviet Union, and the end of Marxist–Leninism as a dynamic ideology.

Introduction

Marxism–Leninism was the official ideology of the former Union of Soviet Socialist Republics (USSR) and by extension of the international communist movement during the twentieth century. The purpose of this article is to set out succinctly the origins of Marxism–Leninism as a political ideology, its foundations in the philosophy of history developed by Karl Marx (1818–83) and his collaborator F. Engels (1820–95), its characteristic features in revolutionary political theory and practice as established by V.I. Lenin (1870–1924), and its consolidation and operation by J.V. Stalin (1879–1953), effectively from 1922 until 1953. This was achieved through the application of the fortress doctrine of 'socialism in one country,' and through the paradoxical tool of the Communist International (Comintern), which organized and controlled foreign communist parties according to the political needs of the Soviet Union. The article also considers revolutionary critics of Marxism–Leninism and controversies that affected the application of the ideology in practice. The transition from the Comintern to the Communist Information Bureau (Cominform) and other Soviet agencies of the Cold War, including support for national liberation movements, is also considered. There is a final section on the condemnation of Stalin by the Communist Party of the Soviet Union (CPSU), attempts at revisionism and reform, the impact of Gorbachev, the collapse of the Soviet Union, and the end of Marxism–Leninism as a dynamic ideology.

Marx and Marxism

It is clearly necessary to begin with the philosophy of history developed by Karl Marx. This has given rise to a voluminous literature of analysis and comment, as well as to bitter ideological dispute. Marxism is based on the philosophical theory of dialectical materialism. Its fundamental principles are well known and will only be summarized here. In his early writings, Marx developed the concept of *alienation* with its roots in the labor process under capitalism. According to this system, the mass of the population was compelled by private ownership of production to work as employed labor or to be unemployed and destitute. Marx believed that such economic relations were fundamental to social and political relations. This is explained through his famous, if subsequently contentious, *base and superstructure* metaphor. In a classic passage in the *Preface to a Contribution to the Critique of Political Economy* written in 1859, Marx said: "The mode of production of material life conditions the social, political and intellectual life process in general. It is not the consciousness of men that determines their being, but, on the contrary, their social being that determines their consciousness" (Marx and Engels, 1968: 181). In other words, as the components of the economic base change, so does its influence on political, social, and cultural institutions, such as the family and the state, which comprise the superstructure. The contention is whether the relationship between base and superstructure is simple and deterministic or one that is complex and dialectic. Marx was, fundamentally, a theorist of capitalism rather than socialism (about which, logically and empirically, he could say very little). In *Capital* (1990), his most detailed work, he analyzed the structure of capitalism, elaborating a labor theory of value, a theory of capital accumulation, and a theory of commodity fetishism, with their implications for his earlier concept of human alienation.

Marx did not complete his planned study, but believed that in *Capital* he had shown that the capitalist system of ownership and production contained the seeds of its own destruction, and that economic and social revolution must follow once this process of decay was complete. The key to Marx's analysis is not the poverty of the proletariat in absolute terms, especially given the productive possibilities of technical advance, but its exploitation in the production of surplus value, over the disposal of which it had no control. This is important given revisionist debates. This analysis was, according to Marx, the key to understanding historical change and as he and Friedrich Engels declared in the opening pages of *The Communist Manifesto*: "The history of all existing society is the history of class struggles" (Marx and Engels, 1986: 79). Thus, slave societies had given way to feudal relations that had been replaced by capitalist employment relations. These, Marx predicted, would give way to socialist ownership and employment relations,

which would be brought about through proletarian or worker revolution and the expropriation of the capitalist class.

Each economic and social transition depended on active human intervention. As Marx said in his celebrated eleventh thesis *On Feuerbach*, while hitherto: "The philosophers have only *interpreted* the world in various ways, the point however is to *change* it" (Marx and Engels, 1968: 30). In another famous passage, this time from *The 18th Brumaire of Louis Bonaparte* (1852), Marx had recognized: "Men make their own history, but they do not make it just as they please; they do not make it under circumstances chosen by themselves, but under circumstances, directly found, given and transmitted from the past" (Marx and Engels, 1968: 96). Marx and Engels had developed a revolutionary praxis aimed at liberating humanity from its passive and alienated existence. As Engels said at the graveside of his friend and comrade: "For Marx was before all else a revolutionist…. Fighting was his element" (Marx and Engels, 1968: 430).

However, the exploratory and essentially dynamic nature of Marx's work meant that it was not suitable for use as a simple textbook of revolutionary organization. According to Engels, in an 1882 letter to Eduard Bernstein (1850–1938), Marx's response to such use of his work was to say: "If anything is certain, it is that I myself am not a Marxist." Despite this, a vulgarized version of his thought became influential among socialist parties in Europe, and particularly in the powerful German Social Democratic Party under the leadership of Karl Kautsky (1854–1938), the so-called 'Pope' of Marxism (S. Hook cited in Morgan, 2003: 71). This was challenged by a revisionist wing of the German Social Democratic Party led by Bernstein, which emphasized the moral values of socialism rather than historical materialism. It also claimed cogently that in the advanced economies, such as Western Europe and North America, workers were being integrated peacefully into economy and society through parliamentary democracy, trade unionism, state welfare, technological advance, and a rising standard of living. The collapse of the Second (Socialist) International with the War in 1914 demonstrated the absence of effective solidarity among the workers of the world.

Lenin and Leninism

The Russian revolutions of 1905 and 1917 had a considerable impact on the socialist and communist movements internationally, not least by distinguishing between them. During the nineteenth century, Russia had experienced a variety of revolutionary socialist and anarchist movements, which were held in check by Tsarist autocracy. The emergence of a constitutional monarchy between 1900 and 1914 gave promise that Russia might develop economically and socially along Western European lines (McKean, 1977). In this context and with the failure of revolution in 1905, the Russian Social Democratic Party split over its ideological interpretation of Marxism. The *Mensheviki* (minority) argued that Russia could not expect a proletarian revolution until it had completed each historical stage of economic and social development as described by Marx. The *Bolsheviki* (majority), led by V.I. Lenin, argued that power could be secured on *behalf* of the proletariat through the so-called vanguard leadership of a disciplined and revolutionary communist party, organized according to what was effectively the military principle of *democratic centralism* in which decisions were taken by an authoritative leadership and carried out by disciplined rank and file followers.

This was Lenin's theory of the 'vanguard party,' the essentials of which are found in the classic pamphlet *What Is To Be Done? Painful questions of our movement*. It is a cogent handbook of revolutionary organization, a polemic written specifically for the struggle within the Russian Social Democratic Party, and first published in Germany in 1902. Its title echoed that of the political novel published in 1864 by the Russian radical N.G. Chernyshevsky (1828–89). Lenin argued that, as Western Europe and North America showed, the workers' movements were developing spontaneously as reformist organizations, exemplified by trade unions with their objectives of improving pay and conditions; achieving favorable legislation regarding welfare, health and safety, hours of work, and employment contracts; and supporting labor candidates for local and national political assemblies. The trade or occupational basis and reformist character of such workers' organizations also had the effect of emphasizing sectionalism, even the social stratification of the working class, which undermined solidarity and class consciousness. The consequence was the increasing incorporation of the workers into what Antonio Gramsci (1891–1937) was later described as the *hegemonic* system of welfare capitalism and parliamentary democracy. Lenin's revolutionary political solution, developed with the particular circumstances of Russia in mind, was that of the 'vanguard' communist party described above. It would be comprised of Marxist intellectuals and militants, whose task was to develop among the working masses a revolutionary rather than a *mere* trade union consciousness.

Lenin was to elaborate this in *Left-Wing Communism: An Infantile Disorder* (Lenin, 1968: 572–585) in which he argued that trade union and similar working class organizations were objectives to be secured in an ideological war of position. This again anticipated the work of Antonio Gramsci. It is shown also in Lenin's attitude toward cultural and educational development, exemplified in the ruthless and successful polemics against A.A. Bogdanov, *aka* Malinovsky (1873–1928), the chief communist advocate of an autonomous proletarian culture to be developed through the *Proletcult* movement. The quasi-philosophical disputes in which Lenin engaged should also be noted in this context, notably *Materialism and Empirio-criticism* of 1908. This was essentially a polemic directed against Bogdanov's ultraleft ideology and his claims to leadership of the Russian revolutionary socialists, rather than a systematic statement of Lenin's philosophical position. Lenin had set out his political ideas on cultural issues as early as 1905 in *Party Organization and Party Literature*, where he argued: "Literature must become … 'a cog and a screw' set in motion by the entire politically-conscious vanguard of the entire working class" (Lenin, 1971: 22–27). In his final article, published in March 1923, he focused on the obstacles to socialist development posed by an inadequate education system and a low level of general culture. He insisted on the maintenance of a vanguard élite when he argued: "Better fewer, but better. We must follow that rule" (Lenin cited in Morgan, 2003: 140).

It has been argued that the formation and installation in power of such a party oligarchy was Lenin's key contribution to communist political ideology, but one "… which has no solid foundation in the writings of Marx" (Shapiro cited in Morgan, 2003: 128). Such a party oligarchy would be responsible for administering the *dictatorship of the proletariat* through a control of economic base and ideological superstructure achieved by the monopoly of state power. This was elaborated further in *The State and Revolution* published crucially in 1917 (Lenin, 1968: 263–348). Lenin had developed a plan for revolutionary organization based upon his conspiratorial experience of politics in the repressive circumstances of Tsarist autocracy. In this sense, he followed Russian revolutionary practice. However, he was innovative in his objective of not simply destroying the Tsarist state, but of replacing it with a proletarian state, controlled by a vanguard communist party which would justify itself according to the later Stalinist slogan of 'Party and People are One!'

Lenin's other significant political contribution to communist ideology was his theory of imperialism. This was set out in *Imperialism: the Highest Stage of Capitalism* published in 1916 (Lenin, 1968: 169–262). Lenin argued that imperialism was a consequence of the increasing dominance of finance capital in the economy and of the subsequent need for capital to be moved in search of profits, increasingly difficult to obtain because of the relative stagnation of national economies. This led to multinational capital trusts and competition for world resources and markets. This had been anticipated by Marx and Engels in *The Communist Manifesto* when they wrote: "The need of a constantly expanding market for its products chases the bourgeoisie over the whole surface of the globe. It must nestle everywhere, settle everywhere, establish connexions everywhere" (Marx and Engels, 1986: 83). It is recognized that Lenin's analysis of imperialism as monopoly capitalism, and therefore capitalism in its final stage, was derived theoretically from Marx and empirically from Rudolf Hilferding (1877–1941), another Marxist, and from the English political economist J.A. Hobson (1858–1940), who had pioneered the economic concepts of underconsumption and unemployment.

The debate about imperialism is a complex one and continues within current international political economy. Here, we focus on the impact Lenin's theory of imperialism had on shaping Marxism–Leninism as the ideology of the Soviet Union and of the twentieth-century communist movement. Its key importance was that it allowed him to link revolutionary struggle among 'the toiling masses' of the undeveloped countries with that which he anticipated among the proletariat of the developed world, thus uniting in revolutionary consciousness both peasant and proletarian. This is important considering the uneven development of the Russian Empire, which became the USSR. It may be regarded as another distinct contribution by Lenin to the ideology of the communist movement in the twentieth century.

Leninism as a variant of Marxism was criticized severely by both anticapitalist revolutionaries and liberal democrats, most especially for its development of an élite vanguard Communist Party. For example, Rosa Luxemburg (1871–1919), the Polish communist based in Germany, differed from Lenin on both his interpretation of imperialism and his definition of the revolutionary party. She came to the orthodox Marxist conclusion that capital would only collapse once it had completed its own trajectory of growth. She maintained also that revolution could not be accomplished by command, but would come about through a spontaneous mass action that the united working class would be forced by worsening economic circumstances to undertake. As for party organization and the conduct of politics, she predicted that once the Bolsheviks had eliminated multiparty democracy and internal communist opposition, the so-called dictatorship of the proletariat would degenerate into the dictatorship of a faction and, ultimately, of that of an individual (Luxemburg, 1961). The anarchist Peter Kropotkin (1842–1921), following Mikhail Bakunin (1814–76), the arch opponent of Marx in the First International, considered Marxism, and its Leninist variant in particular, a centralizing authoritarian ideology. In *Communism and Anarchy*, Kropotkin argued that State communism was a contradiction in terms as: "… without freedom the whole structure of co-operation would collapse" (Kropotkin cited in Morgan, 2003: 129). Indeed, it has been argued by Schapiro that the autocratic rigidity of communism in the Soviet Union (and by extension to the international communist movement) can be traced to the socialist ideology and organizational methods first set out by Lenin as early as the period between 1899 and 1902 (Schapiro cited in Morgan, 2003: 128).

Stalin and Marxism–Leninism

The events of the Russian Revolutions of February and October 1917 are very well known. In 1918, the name Russian Communist Party (Bolshevik) was adopted at Lenin's initiative. It later became the CPSU, which in 1919 reinforced the organizational principle of democratic centralism by setting up a Politburo with its own Secretariat above the Central Committee of the Party. The state was administered in parallel fashion through a Council of People's Commissars appointed by the CPSU from among its own membership. In practice, the CPSU Secretariat controlled the appointment of Party officials to key appointments throughout the state apparatus, which put Stalin, as the General Secretary, in an exceptionally powerful position. The tolerance of debate and dissent within the Communist Party eroded rapidly, especially in the circumstances of the Civil War in Russia. The mutiny of sailors at Kronstadt was crushed in 1921 and at the 10th Communist Party Congress in the same year, factions such as the Workers' Opposition were banned. The treatment of the cultural heresy of the Proletcult has already been noted. By 1922, all other political parties had been banned.

Lenin had developed the vanguard party and led its seizure of power in Russia, an undeveloped and culturally backward country, with the aim of imposing the conditions for socialism. He saw the revolutionary potential of the Russian peasants and attempted to unite them with the proletarian socialist revolution. As we have noted, this strategy was criticized by such as Rosa Luxemburg and Georgi V. Plekhanov (1856–1918), who regarded themselves as orthodox Marxists. They believed that the conditions for socialism required certain necessary stages of economic development and could not be imposed as Lenin

believed. In practice, the contradictions of Lenin's policy resulted in the long-term maintenance of a one-party dictatorship that came under the control of the ruthless political organizer Joseph V. Stalin (1879–1953). In his massive book *Leninism* (1940a), Stalin describes the basic ideology and program, which he described formally as Marxism–Leninism, to be followed by the Communist Party of the Soviet Union under his direction.

In the pamphlet *Foundations of Leninism* (1940b), Stalin claimed that Lenin, although the follower of Marx and Engels, unlike them "… pursued his activities in the period of developed imperialism … when the proletarian revolution had already triumphed in one country, had smashed bourgeois democracy and had ushered in the era of proletarian democracy, the era of the Soviets" (Stalin, 1940b: 10). The basics of Marxism–Leninism were in place by the time of Lenin's death in 1924. The 'toiling masses,' proletarians and peasants, were to be united in a revolutionary political consciousness developed through the leadership of a communist party of militant intellectuals recruited for their political discipline, knowledge, and capacity for socialist theory. The revolution was to be accomplished in two stages. First, a 'dictatorship of the proletariat,' managed by the élite 'vanguard' communist party, would suppress counterrevolution, and ensure that natural economic resources and the means of production and distribution were in common ownership. Finally, communism would be achieved in a classless society in which Party and State would have 'withered away.'

How this would be achieved in practice remained inchoate and there continued criticism and dissension both from foreign communists and within the CPSU. Rosa Luxemburg has been mentioned, but she was murdered by reactionaries in 1919. Karl Korsch, another prominent German communist, and Georgy Lukács, a leading member of the Hungarian Communist Party, set out their concerns in exceptional theoretical works that attempted to recover the original dialectical and Hegelian philosophical basis of Marxism. Korsch's *Marxism and Philosophy* was published in 1923 and Lukács' *Marxism and Class Consciousness: Studies in Marxist dialectics*, which also contained an essay on 'The Marxism of Rosa Luxemburg,' in 1922. They aroused savage criticism from the Communist International or Comintern (see below) and in practice proved to be of little political importance, although reappearing as inspirations for the New Left in the 1960s. Korsch, a persistent critic of the Soviet Union and the Comintern under Stalin, was expelled from the German Communist Party in April 1926 and in 1936 went to the United States, where he died in 1961. Lukács, on the other hand, recanted his heresy and remained a functionary of the Comintern and a lifelong Marxist–Leninist. It should be mentioned that the Italian communist Antonio Gramsci, particularly his concept of hegemony, is often linked with Korsch and the early Lukács, and again was later an inspiration for the New Left. But Gramsci was in practice a convinced Marxist–Leninist charged by the Comintern with the task of Bolshevizing the Communist Party of Italy (PCI). Moreover, he was imprisoned in Italy from 1926 until shortly before his death in 1937. His work on culture, hegemony, and politics found in *The Prison Notebooks* (1971), although now very influential, only came into prominence some years after World War II.

Marxism–Leninism as Orthodoxy

After 1924, Marxist–Leninist ideology as interpreted by Stalin was consolidated as Communist Party orthodoxy both in the Soviet Union and through the Comintern. Stalin was well placed as Lenin's political successor to deal with internal Communist Party opposition to his authority. The main challenge came from Leon Trotsky (1879–1940), who had prestige because of the role he had played during the Kronstadt Mutiny (1921), in the signing of the Treaty of Brest-Litovsk (1918), and in the Russian Civil War (1918–21). Although in 1904 Trotsky had been a Menshevik, anticipating the fears of Rosa Luxemburg about the dangers of the Leninist party, by 1924 his political record marked him as a disciplined and ruthless Marxist–Leninist. He was however without a strong base of support in the CPSU and was unable to secure the publication of Lenin's so-called 'Last Testament,' dictated in December 1922, which had expressed doubts of Stalin's fitness to lead the Party. Trotsky was also in disagreement with Stalin over 'socialism in one country.' He believed that socialism in Russia could not survive without revolution in the advanced industrial countries, especially in Western Europe. He argued for 'Permanent Revolution' toward this objective in which the Comintern would of necessity have a crucial role. By 1925 Trotsky was removed as Commissar of War, in 1927 was expelled from the CPSU, and in 1928 was deported from Russia. He continued opposition to Stalin and to Stalinism from exile through a formidable output of polemics, but was murdered by a Stalinist agent in 1940. Trotsky's influence continued throughout the twentieth century, not least among the various components of the New Left.

By 1931, the Stalinist program based on this Marxist–Leninist ideology was virtually unchallenged within the Soviet Union. It regarded the working masses as so much raw human resource material, a commodity that could be shaped and used according to the central plan determined by the Communist Party bureaucracy. This reality was emphasized by Stalin in an address to the graduates of the Red Army Academy on 14 May 1935, when he said: "Cadres decide everything. That is the main thing now" (Stalin cited in Morgan, 2003: 143). The result in the Soviet Union, and indeed wherever else Marxist–Leninist communist parties came to power, was a state socialism, which claimed to be building toward an Utopian future. The means of production and of the allocation of wealth were state-owned, utilized through a central plan drawn up by the Communist Party and administered by the State. Thus, the economic base was said to be commanded on *behalf* of the working masses.

The superstructure of society, law, religion, culture, education, and opinion was commanded in a similar way by Party and State. For example, the Soviet State practised militant atheism, achieving an accommodation with the Orthodox Church only in the national emergency of the Nazi invasion of 1941. Marxism–Leninism was the core ideology in which the cadres of the Communist Party were trained in the numbers its rule required. Such cadres formed a bureaucratic apparatus, in Russian *nomenklatura*, according to the parallel structures of the Party and State. Their operational principle was 'democratic centralism'; and this determined policy and practice from the highest Party and State organs to local production units and

house committees, and strengthened the hegemony of the Communist Party over everyday Soviet life.

The economic objective was the building of 'socialism in one country.' Stalin introduced a succession of *Five Year Plans* between 1928 and 1942 and collectivized agriculture, which consolidated the command economy in the Soviet Union. This, together with the Soviet Union's immense natural resources, made it one of the largest industrial powers in the world. The average annual growth rate for these years has been estimated at 10%, which is exceptionally high. The emphasis was on building the foundation for further industrial growth accomplished; using State power ruthlessly; through a massive increase in the labor force; and at the expense of agriculture, private enterprise, and consumption and general welfare. It is probable that without the industrialization of the Five Year Plans, the Soviet Union would not have been able to defend itself against the Nazi invasion of 1941, with consequences for the outcome of World War II in Europe.

In *Foundations of Leninism*, Stalin had declared confidently: "The combination of the Russian revolutionary sweep with American efficiency is the essence of Leninism in Party and state work" (Stalin, 1940b: 112). There were two further Five Year Plans after 1948, which concentrated on making good the industrial losses of the War years, again at the expense of consumption and general welfare. This is not the place to describe or analyze the enormities that accompanied this process of forced industrialization and of modernization through labor conscription, the destruction of a legal system independent of the administrative apparatus, and a repressive ideological hegemony leading to political purges and labor camps, again enforced by the State and its police system. This is now very well documented, for instance, by Robert Conquest (1990) and, from the inside by one of its countless victims, Alexander Solzhenitsyn (1988). The withering away of the Party and the State had become a lost horizon, an Utopian chimera of history.

Marxism–Leninism and International Communism

World War I was seen by Marxist–Leninists as proof of Lenin's *Imperialism: the Highest Stage of Capitalism* (1916). It was argued that the War would lead to the destruction of capitalism through international proletarian revolution. This echoed the final words of *The Communist Manifesto*: "Working Men of All Counties Unite!" Such unity would, it was believed, lead to a world socialist system, based on equality and fraternity, which would agree on the basic issues concerning resource allocation and be free of national prejudices, racism, and militarism. Lenin argued that the Second (Socialist) International had betrayed socialism when it collapsed on the outbreak of War and he was alarmed at efforts to revive it in 1919. In response, he organized the Communist or Third International (Comintern) in March 1919 with the object of promoting world revolution. As Stalin claimed later, the struggle against "the opportunism of the Second International" was to be "one of the most important tasks of Leninism" (Stalin, 1940: 11).

The Comintern was established as a single world communist movement with the national communist parties forming its affiliated 'sections' according to Twenty-one Conditions for membership. Essentially, each section was required to organize on the same Marxist–Leninist ideological basis as the CPSU, the dominant party. Indeed, as the post-War prospects of revolution elsewhere than in Russia receded, so the strategy and tactics of the Comintern reflected the internal struggles and purges of the CPSU. For example, in 1927 Trotsky was inevitably expelled from its Executive Committee when expelled from the CPSU. The national communist party sections were also expected to adapt to often drastic changes in direction according to the perceived aims of Soviet foreign policy. The swings from the *Popular Front* of 1935 to the *Nazi-Soviet Non-Aggression Pact* of 1939 and back again with the German invasion of Russia in 1941 are egregious examples. The Comintern was dissolved in May 1943 in the context of the Allied coalition against Nazi Germany.

After World War II, the Red Army's occupation saw Marxist–Leninist communist parties installed in power in much of eastern and central Europe. These were 'satellite' parties of the CPSU, with the exception of socialist Yugoslavia and Albania. The Comintern was replaced by the Cominform in 1947, with the *Council for Mutual Economic Assistance* in 1949 and, of course, the military alliance of the *Warsaw Pact* in 1955. These provided a common platform of politics, economics, and security for the members of the Soviet bloc. It had an influence extending beyond Europe to the developing world, for example, in socialist Cuba and Vietnam, and as the Soviet Union tried to engage with so-called nonaligned countries. However, once again the essential aim was to serve Soviet foreign policy, rather than the goal of world revolution. These organizations were themselves dissolved in the 1990s with the collapse of the Soviet bloc and of the Soviet Union itself.

However, during the Cold War with the capitalist West, Soviet Marxist–Leninist ideology was often accompanied by the use of force, as in Hungary in 1956, in Czechoslovakia in 1968, and in Afghanistan in 1979; and through proxy conflicts such as those in Korea, Cuba, Angola, Mozambique, South Africa, and Vietnam. Indeed, Marxism–Leninism had a very influential role throughout the old Third World during the twentieth century. It was an inspiration for national liberation movements, for example, in Latin America, especially Cuba, where Fidel Castro (1926) and Ernesto 'Che' Guevara (1928–67) became revolutionary icons; in Chile, where the Communist Party, supported by prominent intellectuals such as the poet and Nobel Prize laureate Pablo Neruda (1904–73), was a key component of the Popular Unity government of the Marxist and Socialist Party leader Salvador Allende (1908–73) deposed by military coup d'état in 1973; in Southeast Asia, where the Vietnamese communists, led by another revolutionary icon Ho Chi Minh (1890–1969), inflicted decisive military and political defeats on the French and later on the Americans; and in South Africa, where the Communist Party, led by other iconic leaders such as Joe Slovo (1926–95) and even it is claimed Nelson Mandela (1918–2014), allied with the African National Congress in the antiapartheid struggle.

Despite these significant local successes, ideologically there was a retreat from belief in the inevitability of imperialist war as capitalism showed it could adapt to the end of empire. There was also nuclear weapons technology, which invalidated possible gains from direct armed conflict. Marxism–Leninism

turned to the theory of state monopoly capitalism as an alternative, but this implied that Lenin's theory of imperialism had been overtaken by events. The goal of world revolution was also to become a lost horizon of Marxism–Leninism, another Utopian chimera of history, whatever official Soviet texts may have said (Volkov, 1982).

The *Eurocommunism* phenomenon of the 1970s and 1980s is a further example. This was an attempt, notably by the Italian Communist Party (PCI) and the Spanish Communist Party (PCE), to develop a strategy relevant to contemporary Western Europe and not tied to the Marxist–Leninist orthodoxy of the CPSU. The phenomenon was influenced by a growing interest in the Gramscian concept of hegemony. Ideologically, it symbolized the growing divisions between communist revisionists and doctrinaire Marxist–Leninists, and was mentioned by Mikhail Gorbachev as an influence on his own political thinking. However, the large French Communist Party (PCF), apart from a brief period, and a number of smaller parties remained opposed to Eurocommunism and stayed in accord with the CPSU until the end of the USSR.

Mao Zedong Thought and Marxism–Leninism

Given the continuing impact of the Communist Revolution and People's Liberation in China in 1949, it is necessary to say something briefly about the relationship between Mao Zedong Thought and Marxism–Leninism. It should be noted that there are several points of difference between the Chinese and the Russian revolutions. The former was overwhelmingly a peasant revolution, successful only after a long guerrilla war against both foreign (imperialist Japanese) and domestic (nationalist *Kuomintang*) enemies. In such circumstances, Mao Zedong (1893–1976), the leader of the Chinese Communist Party (CCP) laid down a *mass-line*, which argued that all power and authority is derived from the will of the people and must be endorsed by them, through mass meetings and popular demonstrations. It emphasized also that officials and intellectuals, especially those within the Communist Party, should be monitored consistently and called to account regularly. This would prevent the growth of privileged hierarchies and ensure that revolutionary momentum was sustained. The ideological essentials of this *mass-line* were set out in Mao Zedong's now famous *Talks at the Yenan Forum on Art and Literature* in May 1941 (Mao, 1954: 63–93). They anticipated dramatic later Chinese revolutionary episodes such as the *Great Leap Forward* (1958–59) and especially the *Great Proletarian Cultural Revolution* (1965–69).

That said, Mao Zedong constructed the CCP according to Marxist–Leninist organizational principles and considered himself to be applying the basic tenets of the ideology to the specific conditions in which the CCP found itself. One is reminded of Marx's comment in *The 18th Brumaire of Louis Napoleon* cited above. Mao Zedong was not fundamentally a sophisticated theoretician, but he was a revolutionary tactician and a ruthless political mobilizer of genius, who inspired Maoist or quasi-Maoist groups elsewhere such as the *Naxalites* in India, the *Sendoro Luminoso* or Shining Path in Peru, and the *Khmer Rouge* in Cambodia. It has been said that Marx provided the philosophical basis of communism, Lenin its political organization, and Mao Zedong the guerrilla warfare strategy necessary for an extended revolutionary war. However, there is now virtually nothing left in contemporary China of Maoism with its égalitarian rejection of bourgeois values, other than memories of the enormities of the Cultural Revolution, although the People's Republic itself remains Mao Zedong's achievement. The CCP retains its Marxist–Leninist organizational structure of democratic centralism, but in the service of what is essentially an authoritarian and indeed nationalistic state capitalism.

Conclusion

The outcome of World War II raised hopes that the system of economic and of ideological controls, developed during forced industrialization and total war, might give way to something more open and flexible. In practice, aggravated by the onset of the Cold War, Marxist–Leninist ideology continued to hold that the world had two camps, that of the capitalist imperialists and that of the socialists and democrats who were building communism. In the Soviet Union, this was reasserted to ensure ideological conformity in literature, music, the creative arts, education, and the sciences. It should be noted that the last had for pragmatic reasons been relatively exempt from such ideological control. This campaign, intended to control those capable of constructing ideological alternatives, was enforced by A.A. Zhdanov (1896–1948), probably on the specific instruction of Stalin himself, and was known in Russian as the *Zhdanovschina*. The most notorious example was the official support for Trofim D. Lysenko (1898–1976) and his ideological version of genetics. Although this campaign eased following the death of Stalin in 1953, such ideological direction of the intelligentsia became associated increasingly with efforts to control prominent dissidents such as the authors and Nobel Prize winners Boris L. Pasternak (1890–1960), Alexander I. Solzhenitzyn (1918–2008), and the nuclear physicist Andrei D. Sakharov (1921–89). This may be compared with the triumphal reception accorded to the award of the 1965 Nobel Prize for Literature to the socialist realist novelist Mikhail A. Sholokhov (1905–84).

The long post-War retreat of Marxism–Leninism as an ideology may perhaps be dated from the occasion of the condemnation of Stalin by Nikita S. Khruschev (1894–1971) through his so-called 'secret speech' *On the Cult of Personality and Its Consequences* to a closed session of the 20th Party Congress of the Communist Party of the Soviet Union on 25 February 1956. The speech drew attention to Stalin's purges of the Soviet military and Communist Party cadres, especially in the period leading up to World War II. Khrushchev explained these as a consequence of Stalin's personality cult and as a betrayal of Marxism–Leninism, rather than as something inherent in the system. It was prompted by struggles within the CPSU, but also marked the beginning of a long ideological thaw and the eventual breakup of the Soviet system. The speech became widely known, although it was only published officially in 1989.

The years subsequent to Khrushchev saw the Marxist–Leninist system come increasingly under strain, both within the Soviet Union and externally, with the period of Leonid Brezhnev (1906–82) especially stagnant, the *Brezhnev Doctrine*

(1968) being essentially defensive. However, revisionists believed the system could still be revived from within. In March 1985, Mikhail S. Gorbachev (1931) was selected as CPSU General Secretary, becoming the first president of the Soviet Union in 1990 (when he also won the Nobel Peace Prize). Gorbachev attempted reform through a twin policy of *perestroika* (restructuring) and *glasnost* (openness). It was supported by a revisionist Party intelligentsia and encouraged by a slowly emerging *civil society* (a concept anathema to doctrinaire Marxism–Leninism).

It proved too little, too late as the Soviet system, together with those of its satellites in eastern and central Europe, imploded. It is important to note that this process had been long and drawn-out as a reservoir of support for the Marxist–Leninist system persisted, as did its outposts elsewhere in the world. Nevertheless, on 26 December 1991, the USSR was dissolved formally and Marxism–Leninism was effectively moribund as a political movement. However, its influence internationally was considerable as the history of the twentieth century shows. The question whether Marxism continues to be influential is quite another matter as, for example, the continuing use of dependency theory and of historical materialism in the debates of international political economy also shows.

See also: Analytical Marxism; Cold War, The; Cold War: Impact on Social and Behavioral Sciences; Counter-Insurgencies and the Social Sciences; Marx, Karl (1818–83); Marxism in Contemporary Sociology; National Liberation Movements; Russian Revolution, The.

Bibliography

Brown, A., 2009. The Rise and Fall of Communism. Bodley Head, London.
Brown, A., 2004. The Demise of Marxism–Leninism in Russia. Palgrave Macmillan, Basingstoke, England.
Bukharin, N., Preobrazhensky, E., 1969. The ABC of Communism (E.H. Carr, Introduction). Penguin Books, Harmandsworth, England.
Carew-Hunt, R.N., 1969. The Theory and Practice of Communism: An Introduction. Penguin Books, Harmandsworth, England.
Chin, S.S.K., 1979. The Thought of Mao Tse-Tung (A.H.Y. Lin, Trans.). Centre of Asian Studies, University of Hong Kong, Hong Kong.
Claudin, F., 1975. The Communist Movement: From Comintern to Cominform. Penguin, Harmandsworth, England.
Conquest, R., 1990. The Great Terror: A Re-assessment. Hutchinson, London.
Evans, A.B., 1993. Soviet Marxism–Leninism: The Decline of an Ideology. Praeger, Westport, CT, USA.
Gramsci, A., 1971. Selections from the Prison Notebooks (Q. Hoare, G. Nowell Smith, Ed. and Trans.). Lawrence and Wishart, London.
Harding, N., 1971. Lenin's Political Thought. Macmillan, London.
Kolakowski, L., 2005. Main Currents of Marxism: The Founders; the Golden Age; the Breakdown. W.W. Norton and Company, New York and London.
Lenin, V.I., 1927. Materialism and Empirio-criticism. Martin Lawrence, London.
Lenin, V.I., 1968. Selected Works. Progress Publishers, Moscow.
Lenin, V.I., 1971. 'Party Organisation and Party Literature': Socialist Realism in Literature and Art. Progress Publishers, Moscow, pp. 22–27.
Lenin, V.I., 1988. What Is to Be Done? (R. Service, Introduction) Penguin Books, London.
Luxemburg, R., 1961. The Russian Revolution and Leninism or Marxism? University of Michigan Press, Ann Arbor, Michigan, USA.
Marx, K., Engels, F., 1968. Selected Works. Lawrence and Wishart, London.
Marx, K., Engels, F., 1986. The Communist Manifesto (A.J.P. Taylor, Introduction). Penguin Books, Harmandsworth, England.
Marx, K., 1990 (E. Mandel, Introduction). Capital: A Critique of Political Economy, vol. 1. Penguin Books in Association with New Left Review, London.
Mattick, P., 1978. Anti-Bolshevik Communism. Merlin Press, London.
Mao, T., 1954. Selected Works, vol. 3. Lawrence and Wishart, London.
McKean, R.B., 1977. The Russian Constitutional Monarchy, 1907–17. The Historical Association, London.
Morgan, W.J., 2003. Communists on Education and Culture 1848–1948. Palgrave Macmillan, Basingstoke, England and New York, USA.
Service, R., 2007. Comrades: Communism; a World History. Pan Macmillan, London.
Solzhenitsyn, A., 1988. The Gulag Archipelago 1918–1956 (Abridged in one volume). Collins-Harvill, London.
Spring, D.W. (Ed.), 1991. The Impact of Gorbachev. Pinter Publishers, London.
Stalin, J.V., 1940a. Leninism. Lawrence and Wishart, London.
Stalin, J.V., 1940b. Foundations of Leninism. Lawrence and Wishart, London.
Volkov, G.N. (Ed.), 1982. The Basics of Marxist–Leninist Theory. Progress Publishers, Moscow, Union of Soviet Socialist Republics.

Marxist Approaches to Social Work

Tom Vickers, Northumbria University, Newcastle, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores the diversity within Marxist approaches to social work, outlining key aspects of Marxism and the implications for social work under capitalist states, in situations where states break down, and under socialist states. Within each of these contexts, key examples are explored to demonstrate the range of approaches taken. The article argues that there has been a neglect in the English-language literature of Marxist approaches to social work under socialism, and suggests that recent moves to open up the definition of social work through the concept of 'popular social work' offer one approach to begin filling this gap.

Introduction

English-language writing on Marxist social work has generally been limited to the 'Radical Social Work' (RSW) tradition (e.g., Corrigan and Leonard, 1978; Reisch and Andrews, 2001; Lavalette, 2011) and has neglected much of the diversity within Marxism including the experiences of socialist countries (the limited English-language literature includes Oanh, 2002; Strug, 2006; Yip, 2005). A comprehensive coverage of Marxist approaches to social work is beyond the scope of this article. Instead, the author outlines the key features of Marxist understandings of society and the implications for social work under capitalist states, in situations where states break down, and under socialist states. Within each of these contexts, key examples are explored, following Marxism's attention to concrete conditions. The article is written from a Marxist, specifically Leninist, perspective. The author has attempted to give a fair account of other trends, and be open about his standpoint to support a critical engagement.

Marxist Understandings of Society

> Marxism is many things…in a state of flux and development, and is subject to highly divergent interpretations…Marxism is not simply a theory: it is a political practice which confronts capitalism with an alternative model of a social order. (Corrigan and Leonard, 1978: pp. xiii–xiv)

Marxism is unfinished, containing theory and practice in a dialectical unity, 'praxis' (Lenin, [1895–1916]1972: p. 103). Marxism presents a systemic view that can help social workers relate the personal to the social to address the root causes of people's problems, an aim shared by some feminist and anti-racist approaches.

Dialectical Materialism

The Marxist analytic method, dialectical materialism, involves an iterative movement from holistic and concrete living phenomena to a 'number of determinant, abstract, general relations' and from there back to a more complex understanding of the living whole (Marx, [1857]1973: pp. 100–102). Marxism focuses on the shaping of consciousness by experience, shaped in turn by the way society is organized, with its most basic roots in processes of production and reproduction. Marxism's central premises are that people must produce to survive, that the satisfaction of needs leads to further needs, that people reproduce not only themselves but their species, and that all this activity is organized socially, depending on the means of production available. Dialectical materialism is distinguished from crude materialism by its acknowledgment that history is determined by not only the economic class struggle, but also "the legal, political, religious, artistic or philosophic – in short, ideological forms in which men (sic) become conscious of this conflict and fight it out" (Marx, [1859]1971: p. 21), with ideas impacting back on the future development of the material basis of society. In one example of the application of this to social work, Corrigan and Leonard (1978: pp. 129–132) suggest that migrant families within Britain can be better understood in the context of the relations of production in their country of origin, and how these interact with relations of production within Britain. Such understandings enhance agency by developing an understanding of how structural conditions shape and limit individual behavior.

Class Struggle

Marxism is not the only approach focusing on class, but Ferguson (in Lavalette, 2011: pp. 128–129) distinguishes Marxism by its attention to inequalities in the ownership of the means of production, as opposed to simply in distribution, and exploitation as opposed to simply domination, with the contradictions and consequent struggle between exploiters and exploited driving the development of society.

In some cases, racialized, gendered, and other forms of oppression have been neglected by Marxists, subsumed within a homogenous view of the working class that is implicitly white, male, heterosexual, in paid employment, and able-bodied, and critiques of this neglect have contributed to more complex understandings of class (e.g., Dixon et al. in Craig et al., 1982; Meagher and Tett in Cooke and Shaw, 1996). However, this neglect is not fundamental to Marxism. Anderson (2010) shows that Marx himself devoted considerable attention to 'race,' and many Marxists have

continued in this tradition. For example, Chandler Owen, the recipient of one of the first social work fellowships in the United States, linked capitalism to racism, supported unionization of black labor, and was the leader of the Political Council, which was the only African-American political group to come out in support of women's suffrage (Reisch and Andrews, 2001: pp. 44–45). In a long-running debate within Marxism, Clough (1992) argues that the sidelining of 'race' and gender within some approaches to Marxism reflects the narrow interests of a minority of predominantly white, male, and 'skilled' workers in capitalist countries, who are afforded relative privileges on the basis of their countries' imperialist position in the global economy, receiving super profits from the exploitation of the workers and resources of oppressed countries (Callinicos (2009) gives an opposing view).

The Politics of Social Work

Marxism suggests that social work is inherently political:

> conventional practice is fully political, whether or not its politics are acknowledged. The ends it serves, however, are conservative ones. Radicals do not seek to introduce politics into an apolitical situation. Rather we mean to challenge the politics of compliance and to introduce the politics of resistance and change. (Galper, 1980: pp. 10–11)

For example, Lavalette and Ioakimidis (2011: pp. 142–146) point to the political role of the US-backed international expansion of social work in Western Europe and Asia following the Second World War as part of the Marshall plan, "the goal of which was to make the countries 'safe' for Western capitalism and establish a social and political bulwark against Communist expansion," first through the United Nations Reconstruction and Rehabilitation Agency and then the American Rehabilitation Missions. Similarly, the first professional training programs for Nicaraguan social workers were established in the early 1960s, following the Cuban Revolution "to promote a model of development in the region intended to reduce social discontent and prevent further such revolutions" (Wilson and Hernández in Lavalette and Ioakimidis, 2011: p. 77). In such a context, for social workers to fail to consciously adopt a political position leads to tacit complicity with the status quo, which serves the interests of the dominant class.

The State

Marxists differ in their understanding of the capitalist state, its capacity for reform, and if it must be abolished then what should replace it. They agree in viewing the class character of the state as important in creating the conditions under which social work operates. Lenin ([1917]1972: p. 8) defines the state as an institution emerging in the midst of class struggle as a means of holding in check irreconcilable antagonisms to stabilize the system. In this view, the state emerges as the organ of the most powerful, economically dominant class, who, through the exercise of a state apparatus specifically tailored to its needs, maintains itself as the politically dominant class within a given geographical territory (Lenin, [1917]1972: pp. 13–14). Another important Marxist theorist of the state, Gramsci ([1929–1935]1982), analyzed the dialectical relationship between consent and coercion in the modern capitalist state, and has been particularly influential within Western academia. Disagreements with the Leninist conception of the state typically concern arguments that the British state enjoys 'relative autonomy' from capitalist pressures; is neutral with regard to different sections of society; is fought over by competing interest groups; or is not coherent, with different interests served by different sections of the state. The latter argument has been associated historically with the Fabians in the British Labour Party, who differentiate between a 'good' side of the capitalist state, including social services, health, education, and nationalized industries, and a 'bad' side, including defense, law, and order, and aid to private industry (LEWRG, 1980: pp. 52–53).

Among Marxists who understand the capitalist state as irredeemably antiworking class, there is a divergence between 'libertarian' or 'autonomous' trends, which reject any kind of state, and those who argue a socialist state is needed to facilitate a transition to a classless society. Within the latter, there is a further division as to what constitutes socialism, with different Marxists categorizing the same states as 'socialist,' 'state capitalist,' or 'degenerate/deformed workers' states' (Galper, 1980: pp. 29–39). The class character of the capitalist and socialist states implies a need for a different kind of social work under each. This is connected to

- Socialism's claims to overcome the root causes of the social problems social work seeks to address rather than simply manage their consequences (Strug, 2006);
- The role of socialist state, political and 'mass' organizations in tackling the same issues social work seeks to address (Oanh, 2002);
- Contradictions between professional ideals of individual autonomy and socialist ideals of collective political leadership (Ngai, 1996).

It is a point of contention in the historiography of social work whether social work was unnecessary in countries attempting to build socialism because the problems it sought to address did not exist, at least not to the same extent as under capitalism, or whether the existence of those problems was simply denied by political leaderships lacking in reflexivity, or whether social work continued under socialism, but in different forms and with different personnel (Ngai, 1996: p. 293). The reality may be a mixture of all the three, in different measures in different countries at different times.

Social Work under Capitalism

Because Marxism poses a threat to the capitalist state, for social workers to openly declare their Marxism may jeopardize their employment and with it their ability to practice Marxist social work. Although Marxist social work academics often have greater freedom than practitioners, there has still been a tendency to avoid explicitly discussing their underlying Marxist method. The *Critical and Radical Social Work* journal

launched in 2013 is typical in not mentioning Marxism in its title or aims despite being edited by committed Marxists. The most public face of Marxist social work under capitalism has been as an influential strand within the RSW tradition. In the United States, Reisch and Andrews (2001) trace RSW's history back through the Rank and File Movement of the 1930s, led by Marxists such as Bertha Reynolds, who "warned that unless New Deal policies moved beyond 'offering palliatives to assuage the miseries of poverty and racism,' social workers would do little more than 'carry out the designs of the ruling class and victimize clients'" (Reynolds cited in Reisch and Andrews, 2001: p. 79). Similar histories exist in other countries, such as the East London Federation of Suffragettes, founded in Britain in 1914 by the communist Sylvia Pankhurst, which organized cost price restaurants, nurseries, and other services alongside political agitation. In the 1970s, economic crisis increased antagonisms between the state and service users in some countries, prompting a growth in RSW including a practitioner movement around the *CaseCon* manifesto and magazine in Britain; organizations such as 'Inside Welfare' in Australia; publications such as Galper (1980) in the United States; and Bailey and Brake (1975) and Corrigan and Leonard (1978) in Britain. The history of RSW is contested, seen as either dying out with the defeat of working-class struggles in the 1980s (Langan and Lee, 1989) or developing into antiracist, black, and feminist approaches in response to Marxists' failure to respond adequately to racism, sexism, and other forms of oppression that intersect with class (Dominelli, 1986, 1988; Dominelli and McLeod, 1989). In the early twenty-first century, RSW returned to prominence as a distinct approach, with publications such as Lavalette (2011) and the founding of the Social Work Action Network in Britain, connected to similar organizations in countries as diverse as Hungary, Japan, and the United States (Lavalette and Ferguson, 2011). The discussion here will focus on Britain and the United States, both because they have a more extensive English-language RSW literature and because of their impact on RSW internationally. In Britain, a complicated history between social work and community work led to a distinct strand of 'radical community work' (e.g., Craig et al., 1982), part of social work in many countries, and this branch of the tradition offers important contributions that are included here. The author focuses on the more or less explicitly and self-consciously Marxist elements within RSW, while acknowledging Marxism's wider impact, including within service user movements that have developed alongside RSW (Beresford in Lavalette, 2011: pp. 96–97), and recognizing that many social work writers draw on Marxism without explicitly discussing Marxist theory.

The central problem driving the development of RSW is the contradictory character of capitalist state welfare, simultaneously ensuring a healthy and compliant workforce, managing the behavior of the working class (including attempts to undermine resistance to exploitation), and also providing important material resources for working-class people that often represent concessions won from the capitalist class through hard struggle. Jones and Novak (1999: pp. 79–80) point out, as many others within RSW have done, that state social work interventions are almost exclusively targeted at the poorest in society, who have 'the most marginal if any relationship to waged labor' and who are therefore regarded by the capitalist class as a surplus population to be managed. Stevenson (1978) suggests that while elements of service and control are inherent in all areas of 'human services,' including social work, the balance is different for different groups, with the example that

> the service aspect is dominant in New York City schools when teachers teach white middle-class students. The control aspect is dominant when they teach black and Latino working-class students. (Stevenson, 1978: p. 459)

In a concise article that moves beyond many of the classic social work texts in its treatment of the political economy of welfare work, Stevenson (1978: pp. 457–459) argues that 'human services' labor represents production of a special kind, involving direct contact between producer and consumer that opens up possibilities for solidarity through human contact but operating in a context where there is a material incentive for human service workers to maintain dependency among service users to ensure continuing demand for their 'product.' While Marxists are not alone in arguing that dependency is avoidable, Marxist approaches offer an explanation of the tendencies within capitalist state welfare that pull social workers in contradictory directions, to side with service users in a struggle to overturn the basis of their oppression, requiring at times a willingness to confront the capitalist state, and to cooperate with the state in managing service users' responses to their oppression within limits that will not disrupt the fundamental structures of capitalist society.

There has been a divide within Marxist social workers over how to respond to the contradictions outlined above, centered on whether social work is simply a mechanism of social control, with little or no potential to contribute to progressive change, or whether there is potential to subvert ruling class priorities for social work and develop progressive practice within the capitalist state. RSW has predominantly represented the latter perspective, informed by an understanding of the state as 'semiautonomous' (LEWRG, 1980), containing contradictions and reflecting the class struggle and the balance of forces at a particular point, and therefore capable of shifting its priorities in response to pressure from a strong and organized working class. This implies spaces within the state, created by the class struggle, for socialist practice, which can in turn support further advances in the class struggle. RSW theorists such as Corrigan and Leonard (1978: p. 106) and Jones and Novak (1999: pp. 154–155) point out that compared to civil servants or employees of large state bureaucracies, social workers, and other welfare professionals have a degree of autonomy in their day-to-day work that makes it difficult for the state to monitor and control them, even if this autonomy has been reduced in many countries through increasing bureaucratization of social work and closer supervision (Garrett, 2009).

Some Marxists argue the aims of RSW should not focus on solving individuals' problems, but on highlighting the contradictions that they demonstrate and raising political consciousness about their wider causes as a basis for forming

alliances that could achieve structural change (Cooke in Cooke and Shaw, 1996: pp. 6–25). This approach calls for education alongside organization, recognizing that "the development of a fuller political analysis can only be achieved by the introduction of systematic reflection and more structured educational opportunities within the community work process" (Cooke in Cooke and Shaw, 1996: pp. 20–21). Others argue that there is not necessarily a contradiction between Marxist social work and meeting immediate needs, but rather that RSW needs to link the personal and immediate with the collective and the long term:

> If counselling is required, it must be provided. But if…counselling…fails…to link…temporary and partial solutions with the larger social transformation that is required for realistic solutions, then it is extremely limited, at best, and deceptive and repressive, at worst. (Galper, 1980: pp. 12–13)

Following this approach, standard processes of referral might be reinterpreted to include workers "investigating appropriate political resources in the community and determining their relevance to the particular issues faced by those with whom we work" (Galper, 1980: p. 137). RSW advocates have argued that as well as more overtly political activities, such as political education and organizing, social work also has an important role to play in supporting people to sustain their involvement in collective struggle. For example, Bailey and Brake (1975: pp. 9–10) argue for the need to help people overcome the 'psychological damage' that may result from resisting capitalist hegemony. Baldock (in Craig et al., 1982: p. 30) draws on work with single parents to argue that RSW can foster strong mutual caring to support political activity by sections of the working class who are normally excluded from political activity by the pressures of their personal circumstances.

Questions regarding the relative importance of workplace and community struggles, and the relationship between them, have been contested within RSW. This reflects a contradiction between the emphasis on trade union struggles within dominant trends in the broader Marxist movement in countries such as Britain, Canada, and the United States, and the daily contact between social workers and sections of the working class facing oppression outside the workplace, and at times engaged in significant struggle there, such as the movement against the Community Charge, known popularly as the 'Poll Tax,' in Britain in the late 1980s and 1990s. Corrigan and Leonard (1978: pp. 142–147) are typical of the dominant trend in the wider Marxist left in these countries when they interpret an alliance between social workers and the working class as meaning "the trade unions must come into the forefront of all those involved in campaigns against the welfare cuts at a local and national level," because "they represent the only powerful force that has struggled and will continue to struggle against the cuts in the welfare state." The authors acknowledge the existence of only two trends within Marxism on the question of trade unions, both of which assert that trade unions must be the preeminent working-class organizational form, and differ only on the emphasis placed on the rank and file versus the leadership. This view neglects the work of Marxist theorists such as Lenin ([1917]1972) and Gramsci ([1929–1935]1982) who have written extensively on the limitations of trade union politics and the potential for their co-option by the ruling class. Corrigan and Leonard note the limitations of unions, which "by themselves (they) are not revolutionary or potentially revolutionary organisations of the working class," but argue this has a historical and cultural, but not material, basis, and can therefore be changed through a long period of persuasion (pp. 142–147; for an alternative view on British trade unions see Clough, 1992). Stevenson (1978) points to the dual character of human service workers' role as both oppressed workers and agents of oppression toward other working-class people, which creates the potential for a material divide between the immediate material interests of social workers represented by their trade unions and the interests of other sections of the working class using their services, so that

> certain struggles of progressive elements among human service workers (i.e. those around traditional bread-and-butter issues such as higher salaries and better working conditions, as well as more fundamental struggles around democratization and control) may not have any bearing on their role as oppressors. (Stevenson, 1978: p. 460)

In a recent example, in June 2013, Britain's biggest public sector union, Unison, issued a circular to branches instructing its members to implement a cut to rent payments for social housing tenants assessed as having one or more unused bedrooms, popularly dubbed the 'bedroom tax,' lest Unison members jeopardize their employment.

Some RSW writers place greater emphasis on organizing outside unions. Galper (1980: pp. 198–212) considers possibilities for organizing with other social workers inside and outside unions, where possible with other workers sharing 'socialist commitments.' Cooke (in Cooke and Shaw, 1996: p. 11) argues that, although building links with the trade union movement and Labour Party were central to the aims of many radical community workers in Britain in the 1970s, this was often not achieved in practice, and cites factors including the often 'lukewarm response from Trade Union representatives' and the fact that many community campaigns of the period were forced to confront Labor governments and labor-controlled local authorities. Fleetwood and Lambert (in Craig et al., 1982: pp. 48–58) discuss the disconnect between socialist community workers' experience of housing struggles, in particular from 1975, and the orthodoxy of socialist practice in Britain, which "scorned nonworkplace struggles" and "exhorted activists to link up with trade unions and trades councils in the muscle of the labor movement," even in a period where "the very existence…of a labour movement can be seriously doubted" (Fleetwood and Lambert in Craig et al., 1982: p. 49). Accusations of neglect of racism and sexism also contributed to the fragmentation of struggles in and against the state. Fleetwood and Lambert argue that the lack of engagement between socialist community workers and

community struggles lead to the left critique of the limits of spontaneous community action becoming a self-fulfilling prophesy, since

> divorced from socialism, denied legitimacy by the labour movement, such work has tended to be small-scale, cautious, uncoordinated and individualistic. (Fleetwood and Lambert in Craig et al., 1982: p. 50)

Against this, the authors propose a form of socialist practice that

> starts and grows with the experience of people in struggle…developing techniques, organisational forms and relationships which recognize the personal barriers (constructed by capitalist society) to a class consciousness…it will entail people to do extraordinary things, to dress up, sing songs, perform antics in council chambers, to travel unprecedented distances. (Fleetwood and Lambert in Craig et al., 1982: p. 57)

This history of debates within RSW can be drawn upon to inform social work practice today.

Social Work Where the State Breaks Down

Lavalette and Ioakimidis (2011) use the term 'popular social work,' encompassing welfare activities provided by communities and social movements, including situations where states break down or are unable to cope, whether as a result of war, revolution, or natural disaster. There is a rich tradition of welfare services as part of anticapitalist struggles, dating back to Karl Marx, who organized committees to support fellow German refugees after his arrival in London in 1848. The most significant international efforts include the International Organisation for the Support of Fighters for the Revolution (MOPR), known in many countries as International Red Aid, founded in 1922 (Schilde in Hering and Waaldijk, 2003: p. 140), and Cuba's internationalism in education, sport, and health care since 1959 (Yaffe, 2009). MOPR supported communists imprisoned, injured, or killed in the course of political activity, together with their families, through "legal counselling, social welfare for prisoners (including cash, clothing and food for their families), children's homes, support for campaigns directed at the liberation of communist political prisoners and support for political refugees" (Schilde in Hering and Waaldijk, 2003: p. 144). Red Aid activities varied from country to country, and sometimes went far beyond support for political prisoners, for example, in the early 1920s the Swiss section provided free contraceptives and advice (Hering in Hering and Waaldijk, 2003: p. 91).

Welfare activities were also organized by communists during periods of revolutionary warfare, for example,

- from the 1930s the "patriotic and revolutionary Vietnamese tried to build networks of youth, students, workers (horse-cart drivers, carpenters, shoe-makers, porters) in the form of 'red relief services' to serve the poor and provide mutual assistance" (Oanh, 2002: p. 85);
- from 1940 to 1944 in Greece, under Nazi occupation the communist-led EAM (*Ethniko Apeleftherotiko Metopo*, trans. National Liberation Front) developed welfare provision based on grassroots democracy and solidarity, encompassing "the fight for survival, popular administration and 'holistic development'" (Ioakimidis in Lavalette and Ioakimidis, 2011: pp. 115–132).

Social Work under Socialism

The author uses the term socialism to represent a method for organizing society rather than a blueprint for utopia, characterized by social ownership of the means of production, a high degree of democratic participation, and prioritization of quality of life and conditions for each individual to flourish (Galper, 1980: pp. 29–39). The characterization of countries as socialist is widely contested, but has included countries in Africa, Asia, Europe, and America (Galper, 1980: pp. 29–30). The limited literature that exists on social work under socialism shows the diversity of approaches and suggests there would be much to be gained from further research in this area, illustrated by the following brief case studies of the changing role of social work in China and Cuba.

In China, professional social work was abolished in the 1950s along with other social sciences that were judged unsuited to socialism because they developed under capitalism. Ngai (1996: p. 292) argues that professional social work's reemergence in the 1980s was driven by social problems consequent on China's transition to a market economy. During the socialist period, social work took fundamentally different forms within the context of a wider revolutionary process aimed at advancing human welfare. Ngai (1996: p. 293) describes a range of organizations playing 'social work service' roles, including the Ministry of Civil Affairs and mass organizations such as the All-China Federation of Trades Unions, the Chinese Communist Youth League, the All-China Federation of Youth and the All-China Federation of Women. There is also a strong resonance between social work and health care developments during the Cultural Revolution from 1964, which aimed to generalize medical knowledge to the whole population through the 'barefoot doctors' movement, equipping people with the skills to care for themselves and one another, alongside advances in professionalized specialist care. In mental health care, there was a movement in the early years of the Revolution away from institutionalization and medical approaches to a range of individual and collective approaches focusing on rehabilitation wherever possible, the abolition of restraint and combining Western and Chinese practices developed under capitalism with new methods including therapy through productive work and other activities (Yip, 2005: p. 108). During the Cultural Revolution, the political character of psychiatric treatment was intensified, with mental illnesses viewed as inseparable from the damage done to individuals by capitalism. Forms of treatment included mutual support between patients; 'self-reliance,' meaning that patients should be actively engaged in investigating their conditions and understanding their treatment; political education classes, led by a patient member of the group; and open self- and mutual-criticism. Critics argue that psychiatric

confinement was also used during this period as a political weapon against opponents of the Revolution's leadership (Yip, 2005: pp. 108–110).

In Cuba, in the first few years after the 1959 revolution, mass organizations took on social work roles. For example, the Cuban Federation of Women described its members as 'empirical social workers,' who

> facilitated the entry of women into the labor market, promoted their economic, political and social involvement with the Revolution, and organized community members for participation in major educational and public health initiatives. (Strug, 2006: pp. 751–752)

When professional social work was reintroduced, through technical training institutes from 1973, it was as assistants to health care workers (Strug, 2006: p. 752). In the 1990s, the extreme hardships of the 'special period' following the collapse of the Soviet Union, the tightening US blockade, and social problems fueled by the reintroduction of tourism such as drugs and prostitution, led social workers to target particularly at-risk sections of society, including pregnant women and older people. These social workers became increasingly involved in the People's Councils that were intended to act as a bridge between municipalities and local communities, 'comprising community delegates, mass organizations, and administrative entities,' while also maintaining strong links to medical practitioners and the Communist Party (Strug, 2006: pp. 753–755). Social workers engaged in community organizing, built community members' capacity to participate in the People's Councils, and advocated for specialist services for at-risk groups. Strug (2006) cites an interview with a Cuban social work educator who described how these practice developments drove educational reforms, including the creation of a 6-year social work degree program within the Department of Sociology at the University of Havana in 1998:

> We had to change our way of training social workers. We needed an educational program that would teach them how to work with Committees for the Defense of the Revolution and the Federation of Cuban Women [mass organizations], delegates [local elected officials], school teachers, and out-of-school-kids. (Strug, 2006: p. 755)

This was followed by the creation of paraprofessional social work schools to train out-of-school and unemployed youth, who had grown up during the special period, to work in Cuba's poorest neighborhoods, according to one social work student to become 'friends of the family, knocking on doors not touched by the Revolution,' as part of a 'Battle of Ideas' that aimed to strengthen Cuba 'economically, socially, and ideologically' (Strug, 2006: p. 757). Strug (2006) places Cuban social work in the context of socialism, reflecting

> the fact that the central government in Cuba has for half a century promoted collective goals and community ties, and has involved local self-government in the management of neighborhood affairs. (Strug, 2006: p. 758)

Bertera (2003) describes Cuban work with older people that combines health, social work, and geriatric professionals, integrated within the community and sustained by 'broad community participation,' not as an act of charity but as 'organized community action' in concert with the state. Voluntary work is widespread and is seen as a social duty, 'one of the ideological pillars of society,' and a resource on which social workers draw in their work (Bertera, 2003: p. 319). This demonstrates a role for social workers as organizers of people to meet one another's needs as part of a socialist process together with the state, contrasting with social work under capitalism to resist class attacks on working-class people by the state or to mitigate their most damaging effects.

Conclusion

Marxism emphasizes that realizing subjective agency, for practitioners and services users, requires analysis of the constraints imposed by objective conditions. Under capitalism, social work can facilitate social change, through political education, building alliances between oppressed groups, and helping people to cope with the alienation and exploitation caused by capitalism while they struggle to change oppressive structures. Under socialism, a radical shift in approach is needed for social work to retain its relevance, as RSW's goals of organizing society to meet human needs are no longer marginal or oppositional, but are pursued by a socialist state and political and mass organizations tailored to that purpose. At the same time, professional autonomy threatens to compete with the political leadership of the revolution. Approaches to social work and welfare in countries where Marxist-led revolutions have attempted to build socialism offer a rich source of experience, but are underdocumented in English. Recent steps toward broadening what constitutes social work, through the concept of 'popular social work,' represent a step forward in comparing social work approaches under such radically different contexts as capitalist states, states that are breaking down or fundamentally contested, and socialist states.

See also: Community Work; Disasters in Social, Cultural, and Political Context; Neoliberalism; Radical Social Work; Social Work in China; Social Work in Eastern Europe: Surviving Neoliberalism after the 'Iron Curtain'; Social Work in Russia; The Global Agenda for Social Work and Social Development; Trade Union and Social Work in the Age of Austerity: Lessons from Canada.

Bibliography

Anderson, K.B., 2010. Marx at the Margins: On Nationalism, Ethnicity, and Non-western Societies. University of Chicago Press, Chicago, IL.

Bailey, R., Brake, M. (Eds.), 1975. Radical Social Work. Edward Arnold, London.

Bertera, E.M., 2003. Social services for the aged in Cuba. International Social Work 46 (3), 313–321.

Callinicos, A., 2009. Imperialism and the Global Political Economy. Polity, Cambridge.

Clough, R., 1992. Labour: A Party Fit for Imperialism. Larkin Publications, London.

Cooke, I., Shaw, M. (Eds.), 1996. Radical Community Work: Perpsectives from Practice in Scotland. Moray House Institute of Education, Edinburgh.
Corrigan, P., Leonard, P., 1978. Social Work Practice under Capitalism: A Marxist Approach. MacMillan, London.
Craig, G., Derricourt, N., Loney, M. (Eds.), 1982. Community Work and the State: Towards a Radical Practice. Routledge and Kegan Paul, London.
Dominelli, L., 1986. Love and Wages. Novata Press, Norwich.
Dominelli, L., 1988. Anti-racist Social Work. BASW-Macmillan, London.
Dominelli, L., McLeod, E., 1989. Feminist Social Work. Macmillan Education, Oxford.
Galper, J., 1980. Social Work Practice: A Radical Perspective. Prentice-Hall, Englewood Cliffs, NJ.
Garrett, P.M., 2009. Marx and 'modernization': reading capital as social critique and inspiration for social work resistance to neoliberalization. Journal of Social Work 9 (2), 199–221.
Gramsci, A., [1929–1935]1982. Selections from the Prison Notebooks. Lawrence and Wishart, London.
Hering, S., Waaldijk, B. (Eds.), 2003. History of Social Work in Europe (1900–1960): Female Pioneers and Their Influence on the Development of International Social Organizations. GWV Fachverlage GmbH, Wiesbaden.
Jones, C., Novak, T., 1999. Poverty, Welfare and the Disciplinary State. Routledge, London.
Langan, M., Lee, P. (Eds.), 1989. Radical Social Work Today. Unwin Hyman, London.
Lavalette, M. (Ed.), 2011. Radical Social Work Today: Social Work at the Crossroads. The Policy Press, Bristol.
Lavalette, M., Ferguson, I. (Eds.), 2011. International Social Work and the Radical Tradition. Venture Press, Birmingham.
Lavalette, M., Ioakimidis, V. (Eds.), 2011. Social Work in Extremes: Lessons for Social Work Internationally. The Policy Press, Bristol.
Lenin, V.I., [1895–1916]1972. Philosophical Notebooks. Lawrence and Wishart, London.
Lenin, V.I., [1917]1972. The State and Revolution. Progress Publishers, Moscow.
LEWRG, London Edinburgh Weekend Return Group, 1980. In and against the State. Pluto Press, London.
Marx, K., [1857]1973. Grundrisse. Penguin, London.
Marx, K., [1859]1971. A Contribution to the Critique of Political Economy. Lawrence and Wishart, London.
Ngai, N.P., 1996. Revival of social work education in China. International Social Work 39 (3), 289–300.
Oanh, N.T., 2002. Historical development and characteristics of social work in today's Vietnam. International Journal of Social Welfare 11 (1), 84–91.
Reisch, M., Andrews, J., 2001. The Road Not Taken: A History of Radical Social Work in the United States. Brunner-Routledge, Philadelphia, PA.
Stevenson, G., 1978. Social relations of production and consumption in the human service occupations. International Journal of Health Services 8 (3), 453–463.
Strug, D., 2006. Community-oriented social work in Cuba: government response to emerging social problems. Social Work Education 25 (7), 749–762.
Yaffe, H., 2009. Cuban development: inspiration for the Bolivarian Alliance for the Americas (ALBA). Journal of Iberian and Latin American Research 15 (2), 145–162.
Yip, K.S., 2005. An historical review of the mental health services in the People's Republic of China. International Journal of Social Psychiatry 51 (2), 106–118.

Relevant Websites

http://www.radical.org.uk/barefoot/ – Barefoot Social Worker list of radical social work texts (accessed December 2012).
http://marxists.org/ – Marxist Internet Archive (accessed December 2012).
http://www.rankandfiler.net – The Rank-and-Filer, political analysis for radical social service workers (accessed December 2012).
http://www.socialworkfuture.org/ – Social Work Action Network (SWAN) (accessed December 2012).
http://www.redandgreen.org/Cuba.htm – Social Work and Social Policy in Cuba (accessed December 2012).

Marxist Archaeology

Antonio Gilman, Encino, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the principal currents of explicitly Marxist thought that have arisen in archaeological investigations: the application of Engels's evolutionary stages in Soviet archaeology, the contributions of Gordon Childe, the critique of the ecological functionalism and scientism of the Anglo-American New Archaeology, and the social archaeology developed Latin America, Italy, and Spain.

Marxist archaeology is taken here to be the use of the principles of historical materialism as developed by Marx and Engels to analyze the dynamics of the past societies whose material remains archaeologists study. Such a definition must take into account, of course, Marxism's political importance, which has made its use as a guiding historical theory range from mandatory to proscribed, with various degrees of toleration in between. Archaeology is an expensive and cumbersome activity that requires extensive funding and official permits, so that political controls are relatively easily imposed, with a corresponding effect on how archaeologists present themselves. Completely uncensored, archaeologists may decide (in the interests not of prudence, but of persuasiveness) to omit explicit references to Marx and Engels with respect to matters on which Marx and Engels had nothing direct to say (i.e., almost all archaeological cases). Indeed, enough of Marxist thought has become assimilated into the mainstream of the social sciences that many archaeologists may be unaware of the derivation of their ideas. Marxist archaeology is, therefore, what archaeologists who say they are using Marxist principles do.

Soviet Marxist Archaeology

The first broad application of a Marxist approach to archaeology occurred in the Soviet Union. This took the form of applying the ethnological stages of Savagery, Barbarism, and Civilization that Engels (1972[1884]) had adapted from Lewis Henry Morgan's *Ancient Society* (1964[1877]) to the archaeological evidence from the Soviet territories. During the 1920s, most archaeologists carried on with traditional descriptive studies much as before, on the assumption that "because they were studying the history of material culture, their work accorded sufficiently with the materialist perspective of the new social and political order" (Trigger, 1989). During the 'Third Period' of collectivization and forced industrialization, however, orthodoxy along the lines indicated by Engels and Morgan was imposed from the center: unenthusiastic scholars were dismissed (or worse), regional associations promoting the study, and preservation of local monuments and antiquities were disbanded for glorifying the past, and so on (Klejn, 2012).

The result of this imposition, however, was the first systematic, large-scale sociological reading of prehistoric evidence, one that did not fail to render interesting and useful interpretations and to stimulate innovative research. Reading Engels into the preliterate past of the Soviet Union required a broadly functionalist approach that integrated all aspects of the archaeological record. As Trigger (1989) points out, the concrete results achieved in the 1930s in many ways anticipated those of the Anglo-American New Archaeology of the 1960s (a neo-evolutionist processual movement also rooted in Morgan by way of Leslie White (1959)). At the same time, however, the high political stakes of Marxist theorizing, and the intensely hierarchical institutionalization of archaeology under Stalin impeded critical analysis of Engels's propositions in the light of the emerging evidence, a critique all the more necessary given the provisional character of Engels's work. On the basis of the new, up-to-date ethnology of Morgan, Engels had been willing to revise extensively the initial reflections on primitive society and the development of the state that he and Marx had developed in works such as the *Formen* (Marx, 1965[1857–58]) and the *Anti-Dühring* (Engels, 1934[1877]). In the Soviet Union, however, the unilineal stage theory formulated by N.I. Marr, the first director of the State Academy of History and Material Culture, was not formally abandoned until Stalin himself denounced it in 1950, 16 years after Marr's death. Soviet Marxist archaeology was the product of a relatively small group of Bolshevist enthusiasts, so the post-Stalin relaxation of political controls led, as might be expected, less to the development of a critical Marxist archaeology than to a concentration on practical work in which such theorizing as did take place concentrated largely on explaining the diversity of the Soviet Union's archaeological record. This focus on 'ethnogenesis' had a greater affinity to the nationalist archaeology of Gustav Kossina than to the theoretical work of Marx and Engels (Shnirelman, 1995; *see* Archaeology, Politics of; Ethnic Identity/Ethnicity and Archaeology).

Childe

Outside the Soviet Union, the first archaeologist to develop Marxist perspectives systematically was V. Gordon Childe (Trigger, 1980). As professor of prehistory at Edinburgh in the 1920s, Childe (1925, 1928, 1929) had developed a synthesis of Kossinna's paleoethnological concerns with Oskar Montelius's comparative chronological approach on a scale that encompassed both Europe and the Near East. Childe was a man of the Left (he had been active in the Labour Party of his native Australia) and he took keen professional and political interest in emergent Soviet archaeology. The Soviet experiment showed

him how he could fit the prehistory he had elucidated into a larger evolutionary whole. He would recast the transitions from Savagery to Barbarism and from Barbarism to Civilization as the Neolithic and Urban Revolutions, respectively (Childe, 1936, 1942), stressing the dynamic interplay of production and social organization. At the same time, he saw no reason to accept the unilinealism of the Soviet school. Instead, he anticipated late twentieth-century world-system approaches by seeing Europe as peripheral to and dependent on the Near East (Childe, 1958). Finally, he grounded his interpretations in an explicitly realist epistemology (Childe, 1956) in which knowledge is socially constructed, but long-lasting to the extent that it is practical. Precisely because he adjusted the details of his historical reconstructions to the best evidence of his day, many of Childe's specific contentions are dated, but his consistent view that historical changes are rooted in the social relations of production continues to inspire research (e.g., Wailes, 1996).

'Western' Marxist Archaeology

Childe's eminence (and British tolerance of what could be regarded as eccentricity) permitted him to be a solitary exponent of Marxist prehistory in the West during the first part of the Cold War. A broader avowal of Marxist approaches to prehistory would wait until a generation of archaeologists, trained during the university-based radicalism of the late 1960s and early 1970s, attained professional independence. In the United States and Great Britain this renewed, explicit interest in Marxism arose in the context of dissatisfaction with certain aspects of the by-then dominant New Archaeology. This form of neo-evolutionism shared with Marxism many of the same concerns with the prevailing historicist theory and practice and so was susceptible in turn to a Marxist critique. (For detailed treatment with references of the New Archaeology, *see* Evolutionary Approaches in Archaeology; History of Archaeology; Theory in Archaeology; Archaeology and Science Studies.) This critique addresses both of the principal strands of the New Archaeology.

First, the New Archaeology was committed to cultural ecology as an explanatory strategy. Cultures were seen as analytically isolable, functionally integrated, homeostatic regulatory mechanisms in which change occurs due to demographic or environmental pressures external to the cultural system itself. Where small-scale, egalitarian societies are concerned, cultural ecology is not incompatible with Marxism (e.g., Marx, 1965), although attention to the internal tensions (contradictions) of kinship-organized societies may improve upon an innocent systems functionalism (e.g., Gilman, 1984; Saitta, 1997). In societies that exhibit substantial inequalities, however, notions that their privileged elites act as higher-order regulators of complex adaptive systems are wide open to critique from a Marxist perspective. Inasmuch as complex societies are driven by internal competition, the winners of which set themselves above and exploit the mass of the population, they cannot be adequately understood as systems in which an adaptive equilibrium benefits all their members (e.g., Gilman, 1981; Zagarell, 1986; Brumfiel, 1992). Furthermore, complex societies rarely develop in isolation: their class inequalities emerge in the context of external trade and warfare, so that analogs of the world-system approach to capitalism (e.g., Frank, 1967) can be applied to ancient cases (e.g., Frankenstein and Rowlands, 1978; Kohl, 1989; Frank, 1993). Finally, against the New Archaeological view that a society's ideational system consists of the information regulating its adaptive response to the environment, archaeologists have appealed to a Marxist view of ideology as means by which competing classes and factions legitimate their interests so as to control and manipulate others (e.g., Patterson, 1986; De Marrais et al., 1996; Bernbeck and McGuire, 2011). These critiques have been so effective that they have been absorbed into the mainstream of Anglo-American archeological thought.

Second, the New Archaeology insisted upon emulating the methods of the natural sciences by adopting logical–positivist research strategies. This scientism presents inherent difficulties (Wylie, 2000) and is clearly incompatible with a Marxist view of ideology, according to which archaeology would be a body of knowledge developed by social actors in accordance with their interests. Archaeologists have used Marxist ideas to show how the theory and practice of the discipline has been influenced not (or not just) by value-free scientific objectives, but by the social and economic constraints of its setting (Patterson, 1995). Thus, the pressures of sponsorship (e.g., Gero and Root, 1995) would generate archaeological models (e.g., Keene, 1983) and popularizations (e.g., Leone, 1981) that reinforce the necessity of present-day capitalist social relations by showing that they existed in the past. The argument that archaeological claims about knowledge reflect the interests of those who advance those claims can, however, be only a short step away from a self-defeating relativism that sees Marxism as an approach to understanding history that can only be judged on its own terms (e.g., Saitta, 1983).

In addition to providing a critique of nonhistorical materialism, Marxist archaeologists have sought to use the material record to examine the operation of capitalism (e.g., Leone and Parker, 1999), with particular attention to how it can illustrate the lives of the oppressed: such as racial minorities (Orser, 2007), miners (McGuire, 2008), and mill workers (Mrozowski et al., 1996). Study of the recent history and more remote past of the indigenous populations that stood in the way of the modern world system has, of course, been the central object of the prehistoric archaeology of North America and Australia. With some exceptions (e.g., Muller, 1997; Trigger, 1985), these have been uninformed by Marxist perspectives. Inevitably perhaps, the concrete reconstructions of the past arrived at by archaeologists (overwhelmingly of European descent) often conflict with the accounts of their own history developed by members of present-day First Nations. Sympathy for the victims of capitalism is an essential part of Marxism, but when scholars on the left (e.g., McGuire, 1992, p. 243) urge archaeologists to validate indigenous accounts of the past, they veer into a relativism that is more akin to Romantic nationalism than to Marx's Enlightenment rationalism.

'Southern' Marxist Archaeology

In recent years, the most systematic construction of a Marxist archaeology has occurred within the 'social archaeology' current in Latin America (e.g., Lumbreras, 1974; Sanoja and

Vargas, 1978; Bate, 1998; cf Patterson, 1994). In these areas as in post–World War II Italy or post-Franco Spain, Marxism has the prestige of being associated with the opposition to a capitalism operating under authoritarian political auspices. It differs critically from the 'Western' Marxist archaeology in that it is not so much a critique of New Archaeological processualism as a substitute for it. In Latin America, the dominant archaeological paradigm is the traditional historicism that preceded the New Archaeology in the United States and Britain, an historicism oriented toward an ethnopopulist emphasis on the indigenous roots of the modern nations (Vargas, 1995) and/or toward the development of archaeological resources for education and tourism. Archaeologists politically committed to the Left in Latin America have (where their presence has been tolerated) used a Marxist framework to build a processual alternative to the predominant historicism, one that does not suffer from the limitations of the adaptationist, ahistorical processualism developed in the US. Classic Marxist concepts, such as 'mode of production' and 'social formation' are combined with novel constructs such as '*modo de vida*' (mode of life) to develop a systematized basis for archaeological research that has none of the postprocessual distaste for an evolutionist grand narrative. In the 1960s, the Italian archaeologists of the Dialoghi di Archeologia group developed similar approaches in a Gramscian vein (e.g., Bianchi Bandinelli, 1961). In Italy, as in Latin America, institutional constraints have prevented this theory from being put into widespread practice; however, in Spain the democratization of the late 1970s was associated with expanded employment and funding in archaeology. This process in turn has led to active implementation of a variety of historical materialist research strategies (e.g., Ruiz et al., 1986; Vicent, 1998; Nocete, 2001; Lull, 2007; Criado Boado, 2012).

Conclusions

The societies that most archaeologists are concerned with were not at the center of Marx's attention and almost no information about them was available in his and Engels's day. To create a Marxist archaeology one must instead critically and creatively deploy the central themes of Marxist thinking. Archaeology is a discipline that requires institutional support, and Marxism, as a critical approach to historical inquiry, is not easily or fruitfully institutionalized. At the same time, archaeological evaluation of critical Marxist variables, such as property relations ("it is always the relationship of the owners … to the direct producers … which reveals the innermost secret, the hidden basis of the entire social structure" [Marx, 1967[1894]), is not a straightforward task (Gilman, 1997). As a result, Marxist interpretations of the archaeological records are only beginning to be developed.

See also: Anthropology and History; Archaeology, Politics of; Archaeology, Theory in; Cultural Evolution: Overview; Ethnic Identity and Ethnicity in Archaeology; Ethnogenesis in Anthropology; Field Theory; Functionalism in Anthropology; Historical Archaeology; Historicism; Marx, Karl (1818–83); Marxism in Contemporary Sociology; Marxist Geography.

Bibliography

Bate, L.F., 1998. El Proceso de Investigación en Arqueología. Crítica, Barcelona.
Bernbeck, R., McGuire, R.H. (Eds.), 2011. Ideologies in Archaeology. University of Arizona Press, Tucson.
Bianchi Bandinelli, R., 1961. Archeogia e Cultura. Ricciardi, Milan.
Brumfiel, E.M., 1992. Breaking and entering the ecosystem: gender, class, and faction steal the show. American Anthropologist 94, 551–567.
Childe, V.G., 1925. The Dawn of European Civilization. Kegan Paul, London.
Childe, V.G., 1928. The Most Ancient East. Kegan Paul, London.
Childe, V.G., 1929. The Danube in Prehistory. Oxford University Press, Oxford.
Childe, V.G., 1936. Man Makes Himself. Watts, London.
Childe, V.G., 1942. What Happened in History. Penguin, Harmondsworth.
Childe, V.G., 1958. The Prehistory of European Society. Penguin, Harmondsworth.
Childe, V.G., 1956. Society and Knowledge. George Allen & Unwin, London.
Criado Boado, F., 2012. Arqueológicas: La Razón Perdida. Edicions Bellaterra, Barcelona.
De Marrais, E., Castillo, L.J., Earle, T., 1996. Ideology, materialization, and power strategies. Current Anthropology 37, 15–31.
Engels, F., 1934. Herr Eugen Dühring's Revolution in Science. Charles H. Kerr, Chicago.
Engels, F., 1972. The Origin of the Family, Private Property and the State. International Publishers, New York.
Frank, A.G., 1967. Capitalism and Underdevelopment in Latin America: Historical Studies of Chile and Brazil. Monthly Review Press, New York.
Frank, A.G., 1993. Bronze Age world system cycles. Current Anthropology 34, 383–429.
Frankenstein, S., Rowlands, M.J., 1978. The internal structure and regional context of early Iron Age society in south-western Germany. Bulletin of the Institute of Archaeology, University of London 15, 73–112.
Gero, J., Root, D., 1995. Public presentations and private concerns: archaeology in the pages of National Geographic. In: Gathercole, P., Lowenthal, D. (Eds.), The Politics of the Past. Unwin Hyman, London, pp. 19–37.
Gilman, A., 1981. The development of social stratification in Bronze Age Europe. Current Anthropology 22, 1–23.
Gilman, A., 1984. Explaining the upper Palaeolithic revolution. In: Spriggs, M. (Ed.), Marxist Perspectives in Archaeology. Cambridge University Press, Cambridge, pp. 115–126.
Gilman, A., 1997. Como valorar los sistemas de propiedad a partir de los datos arqueológicos. Trabajos de Prehistoria 54 (2), 81–92.
Keene, A.S., 1983. Biology, behavior, and borrowing: a critical examination of optimal foraging theory in archaeology. In: Moore, J.A., Keene, A.S. (Eds.), Archaeological Hammers and Theories. Academic Press, New York, pp. 137–155.
Klejn, L.S., 2012. Soviet Archaelogy: Trends, Schools, and History. Oxford University Press, Corby.
Kohl, P.L., 1989. The use and abuse of world systems theory: the case of the pristine West Asian state. In: Lamberg-Karlovsky, C.C. (Ed.), Archaeological Thought in America. Cambridge University Press, Cambridge, pp. 218–240.
Leone, M.P., 1981. The relationship between artifacts and the public in outdoor history museums. Annals of the New York Academy of Sciences 376, 301–313.
Leone, M.P., Parker Jr., P.B., 1999. Historical Archaeologies of Capitalism. Kluwer Academic, New York.
Lull, V., 2007. Los Objetos Distinguidos: La Arqueología como Excusa. Edicions Bellaterra, Barcelona.
Lumbreras, L., 1974. La Arqueología como Ciencia Social. Histar, Lima.
Marx, K., 1965. Pre-Capitalist Economic Formations. International Publishers, New York.
Marx, K., 1967. Capital. In: The Process of Capitalist Production as a Whole, vol. 3. International Publishers, New York.
McGuire, R.H., 1992. A Marxist Archaeology. Academic Press, San Diego.
McGuire, R.H., 2008. Archaeology as Radical Political Action. University of California Press, Berkeley.
Morgan, L.H., 1877. Ancient Society. Holt, New York.
Mrozowski, S., Ziesing, G.H., Beaudry, M., 1996. Living on the Boott: Historical Archaeology of the Boott Mills Boardinghouses. University of Massachusetts Press, Amherst.
Muller, J., 1997. Mississippian Political Economy. Plenum Press, New York.
Nocete, F., 2001. Tercer Milenio antes de Nuestra Era: Relaciones y Contradicciones Centro/Periferia en el Valle del Guadalquivir. Edicions Bellaterra, Barcelona.
Orser, C.E., 2007. The Archaeology of Race and Racialization in Historic America. University of Florida Press, Gainesville.

Patterson, T.C., 1986. Ideology, class formation, and resistance in the Inka state. Critique of Anthropology 6 (1), 75–85.
Patterson, T.C., 1994. Social archaeology in Latin America: an appreciation. American Antiquity 59, 531–537.
Patterson, T.C., 1995. Toward a Social History of Archaeology in the United States. Harcourt Brace College Publishers, Fort Worth.
Ruiz, A., Molinos, M., Hornos, F., 1986. Arqueología en Jaén (Reflexiones desde un Proyecto Arqueológico No Inocente). Diputación Provincial de Jaén, Jaén.
Saitta, D.J., 1983. The poverty of philosophy in archaeology. In: Moore, J.A., Keene, A.S. (Eds.), Archaeological Hammers and Theories. Academic Press, New York, pp. 299–304.
Saitta, D.J., 1997. Power, labor, and the dynamics of change in Chacoan political economy. American Antiquity 62, 7–26.
Sanoja, M., Vargas, I., 1978. Antiguas Formaciones y Modos de Producción Venezolanos. Monte Avila, Caracas.
Shnirelman, V., 1995. From internationalism to nationalism: forgotten pages of Soviet archaeology in the 1930s and 1940s. In: Kohl, P.L., Fawcett, C. (Eds.), Nationalism, Politics, and the Practice of Archaeology. Cambridge University Press, Cambridge, pp. 120–138.
Trigger, B.G., 1980. Gordon Childe: Revolutions in Prehistory. Thames and Hudson, London.
Trigger, B.G., 1985. Natives and Newcomers. McGill/Queen's University Press, Kingston.
Trigger, B.G., 1989. A History of Archaeological Thought. Cambridge University Press, Cambridge.
Vargas, I., 1995. The perception of history and archaeology in Latin America: a theoretical approach. In: Schmidt, P.R., Patterson, T.C. (Eds.), Making Alternative Histories: The Practice of Archaeology and History in Non-Western Settings. School of American Research Press, Santa Fe, pp. 47–67.
Vicent, J.M., 1998. La prehistoria del modo de producción tributario. Hispania 58, 827–843.
Wailes, B. (Ed.), 1996. Craft Specialization and Social Evolution: In Memory of Gordon Childe. University Museum of Archaeology and Anthropology, Philadelphia.
White, L., 1959. The Evolution of Culture. McGraw-Hill, New York.
Wylie, A., 2000. Questions of evidence, legitimacy, and the (dis)union of science. American Antiquity 65, 227–237.
Zagarell, A., 1986. Structural discontinuity: a critical factor in the emergence of primary and secondary states. Dialectical Anthropology 10, 155–177.

Marxist Geography

Richard Peet, Graduate School of Geography, Clark University, Worcester, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Marxist geography came into existence as part of the radicalization of the social sciences during the 1960s and 1970s. It sees crises in society's relations with space and nature emanating from the deep structure of capitalism. These crises threaten human existence. Marxist geographers believe in social transformation to a new kind of rational, democratic society.

Geographical issues, like globalization, environmental destruction, and geopolitical contention lie at the center of social concern. Many scholars in the social and behavioral sciences have assumed deeply critical stances with regard to the origin of crises in the structures of capitalist societies. Marxist geography is one of these critical analyses. Marxist geography agrees with conventional geography that the focus of geographical study should lie on the relations between society and the natural environment including, in the term 'environment,' the veneer of earth space, the place where society's multiple effects on nature are concentrated (*see* Geography: Overview; Nature–Society in Geography). Marxist geographers argue that capitalism, as an entire way of life, has inherently destructive effects on nature and space. Marxist geographers argue that crisis is more severe in regions where capitalism is most developed, organizationally, technically, culturally, and economically. For Marxist geographers, ending environmental crises means ending the basic cause of the problem, and this in turn entails transforming capitalist societies. This radical intention differentiates Marxist geography, analytically, and politically, from conventional social and environmental studies and analyses.

History of Marxist Geography

The radical movement in geography began as part of a culture of political opposition that coalesced in the mid- to late 1960s. The leading issues of the time involved problems like inequality, civil rights, racism, sexism, and the environment, all brought into greater contention by geopolitical issues like the Vietnam War. Articles dealing with socially relevant geographical topics began to appear in the discipline's mainstream journals. Until that time, geography had mainly engaged in a conservative intellectual discourse, whether in the form of regional description, or the quantitative analysis of spatial distributions of cities and industries. In 1969, *Antipode: A Radical Journal of Geography* was founded, specifically to publish the new kinds of critical work. Early issues of the journal dealt with urban and regional poverty, discrimination against women and minority groups, unequal spatial access to social services, advocacy planning, Third World development, and similar topics. For several years radical geography used conventional philosophies and analytical techniques for studying this new range of more socially engaged issues. Only gradually did socially relevant radical geography, armed with an eclectic critical politics, turn into a more disciplined Marxist geography, aimed at the achievement of democratic socialism (Peet, 1977).

The transformation of radical geography into Marxist geography was informed by a key text, *Social Justice and the City*, written by Harvey (1973). This series of essays records the author's movement from a liberal, critical position, focused on moral issues of urban social justice, to a revolutionary position, based in a philosophy of historical materialism. As scholars and activists began to read Marx, in part stimulated by Harvey's work, in part propelled by the prevailing culture of protest, and in part compelled by the deficiencies of conventional reasoning, geographical thought was transformed, in content and depth, moving from an obsession with the empirical, the descriptive, and the quantitative, to an equally committed obsession with the philosophical, the social theoretical, and the politically engaged. The last 30 years of the twentieth century saw radical and Marxist geography emerging from the backwaters of a remote, conservative discipline to assume a place of prominence in contemporary critical social science as a whole (Peet, 1998). With the onset of economic recession in 2007–08, Marxism and Marxist geography contributed compelling analyses of the structural origins of systemic crises, resuming a central position in the discipline.

Historical and Geographical Materialism

The philosophical position adopted by Marxist geographers, often termed geographical materialism, is derived from the Marxist philosophy of historical materialism. Karl Marx and Friedrich Engels were followers of the German idealist philosopher G.W.F. Hegel. In Hegelian philosophy, events in the real world were determined by the prior thoughts of a World Spirit in a separate spiritual realm – the material realm was some kind of manifestation, or condensation, of transcendent consciousness. Marx and Engels eventually turned this belief back on its material base: consciousness, they claimed, was the product of matter, rather than its origin. Marx retained from Hegel's idealism dialectical conceptions like development through contradiction. However, this development was rethought in terms of real, materialist dialectics, rather than the dialectics of thought, as with struggles between classes, or the contradictory relations between society and nature. In a dialectical, geographical materialism, the natural and social

 http://dx.doi.org/10.1016/B978-0-08-097086-8.72032-0

worlds are not seen as entities remaining eternally the same, but as interacting processes capable of rapid, and indeed dramatic, change.

Geographical materialism follows historical materialism, beginning its account of history with the natural bases of life, or the environments inhabited by groups of people, and with the modification of nature through human action. As with other animals, the assimilation of natural materials ('metabolism') was "the everlasting nature-imposed condition of human existence … common to all forms of society in which human beings live" (Marx, 1976: p. 290). Marx argued that human survival had to be earned by applying labor to the natural world: people are primarily workers. Human consciousness and intentionality result from a life process dominated by labor, and are not gifts from a benevolent God to his earthly sons and daughters. Hence, experience in the production and reproduction of human existence is interpreted to yield ideas that construct consciousness. As consciousness develops, as with science and technology, the productive forces employed in the social construction of continued existence improve, in the sense of becoming more powerful (but also potentially more destructive). This liberates human life from the immediate constraints of nature. Yet 'liberation' occurs through a contradictory process of social exploitation and political domination. The modern world is headed for problems from its inception.

For Marx, the social relations entered into during the production of the material basis of life (i.e., the labor process) have an inherently contradictory aspect. In particular, Marx focuses on social relations between a ruling class that owns the means of production – the raw materials, machines, buildings, infrastructures, and technologies that physically produce goods and services – and a working class, deprived of independent means of livelihood, and forced to work for the capitalist class to earn wages. Class relations between owners and workers, or between capital and labor, are based in the extraction of surplus labor time, in a process Marx calls 'exploitation.' Societies are exploitative when owners extract uncompensated labor time from direct producers. Used in such an exploitative way, the money that hires labor to produce value becomes capital (hence capital is formed by a social labor process rather than resulting from private entrepreneurship), and social relations are necessarily contradictory, in the sense that one class lives off another. At the same time, competitive relations among capitalists compel the reinvestment of capital in more efficient means of production. Hence, exploitation and competition are the social relations compelling capitalism's tremendous economic growth. For Marx, the exploitation process was an arena of intense social struggle, with the owning, ruling class using a combination of economic, political, and ideological force to ensure their dominance, and the working class resisting through overt means like social organization and rebellion, and hidden means, like reluctant compliance.

In such a context of competition, exploitation, and struggle, Marxists theorize that social consciousness has to be diverted into ideological forms. Primarily, ideologies rationalize and legitimate exploitation, but they also support competition as efficient, and excuse environmental destruction as necessary for economic growth and continued employment. Some Marxists have extended the notion of ideology deep into culture, claiming that even common sense is a form of ideology (Gramsci, 1971; Althusser and Balibar, 1968). In Marxism, even the democratic state is controlled by the rich – for example through buying elections and lobbying in the contemporary capitalist state. The total ensemble of social relations, forces of production, ideological forms of consciousness, and political institutions that made up any social system was termed, by Marx, the 'mode of production.' For Marx, history takes the form of a succession of modes of production: gathering and hunting societies, organized by primitive communal relations; agricultural societies, organized through tribal social structures; state societies, organized by payments of tribute, taxes, and rents; and finally industrial societies organized by capitalist social relations of competition and exploitation (Marx, 1976).

Social Relations with Nature

Following this analysis, Marxist geographers, as dialectical, geographical materialists, see advanced societies, armed with powerful technologies, organized by competitive and exploitative social relations, having disastrously destructive effects on society, culture, and nature. For Marxist geographers, development is a contradictory movement: the production of social surplus began a process of emancipation from natural constraints; yet emancipation was accompanied by internal social differentiation marked by institutions of domination like the state, patriarchy, and the market. The precise resolution of the contradiction between emancipation and domination forms the specific kind of society. The bourgeois revolution, the formative social process of capitalism, introduced a class society in which economic growth occurs not for the sake of meeting needs (food, shelter, essential services), but for purposes of making profit and accumulating capital by the ruling elite. While all societies have socially mediated relations with nature, capitalism differs in its higher degree of complexity: under capitalism, natural relations are determined by blind forces of profit and competition, rather than deliberative, democratic, human intention. Under the abstract dictatorship of competition, capital must expand continuously, stalking the earth in search of raw materials, appending nature to the production process. This argument leads some Marxist geographers (like Smith, 1984) to the conclusion that nature is now 'socially produced' in the sense that no original relation with nature is left unaltered, and no living thing unaffected. Yet the production of nature universalizes contradiction as capitalism creates new barriers – scarcity of resources, nuclear technology, and pollution of the entire earth environment. This suggests, for Marxist geographers, the 'natural inevitability' of fundamental, social transformation, a new form of society spurred by the 'revenge of nature' (e.g., the greenhouse effect).

Mode of Production and Space

Marxist geography also looks at the relations between modes of production and space. In some of this work, especially in structural Marxist geography following the ideas of Althusser

(Althusser and Balibar, 1968), space is seen as a 'level' or 'instance' in a mode of production. Each aspect of social life is theorized as having a particular relation with environmental space, each mode of production creates distinct spatial arrangements, and successions of modes of production alter the landscape in any given space – so medieval churches lie within four-lane highways leading out to super-shopping centers. The themes of social practice, consciousness, and space were developed originally by the Marxist philosopher and sociologist, Lefebvre (1991). He proposed an analysis of the active production of space to bring the various kinds of space and the modalities of their genesis within a single theory. Lefebvre's thesis, simply put, was that social space is a social product: every society, each mode of production, produces its own space. Space consists of natural and social objects (locations) and their relations (networks and pathways). Yet while space contains things, it is often thought of the relations between them. For him, people are confronted by many, interpenetrated social spaces superimposed one on top of the other, what he calls a 'hypercomplexity,' with each fragment of space reflecting, yet hiding, many social relationships (see also Soja, 1989, 1996).

A coherent version of this idea was developed by the urban sociologist Castells (1977). For Castells, any social form, like space, can be understood in terms of the historical articulation of several modes of production. By mode of production, Castells means the particular combinations of economic, politico-institutional, and ideological systems present in a social formation, with the economy determining the mode of the production of space in the last instance. In terms of this theory one can 'read' the resulting space according to the economic, political, and ideological systems that formed it. An economic reading concentrates on the spatial expressions, or 'realizations,' of production (factories, offices, etc.), consumption (housing, sociocultural institutions), and exchange (means of circulating goods and ideas). The politico-juridical system structures institutional space according to processes of integration, repression, domination, and regulation emanating from the state apparatus. Under the ideological system, space is charged with cultural meaning, its built forms and spatial arrangements being articulated one with another into a symbolic landscape. For Castells (1977: p. 127), the social organization of space can be understood in terms of the determination of spatial forms by each of the elements of the three instances (economic, politico-juridical, ideological), by the structural combination of the three (with economy determining which instance is dominant in shaping a particular space), but also (empirically and historically) by the persistence of spatial forms created by earlier social structures, articulated with the new forms, in interaction with the natural environment.

The intention behind this kind of work was to create a Marxist 'structural science of space' – that is an organized explanation of the structural order of the forces and relations that impinge on space. The idea was not how economy is 'reflected' in space, but how economy arranges the political, cultural, and social organization of space or, in a more sophisticated statement, how an economy is made up from its various spatial instances. Yet this kind of scientific abstraction sometimes gave structural work on space a quality of unreality, as though the contents of space were laid down by driverless social paving machines passing over it. Also the return effects of space on society are minimal and mechanical in some structuralist approaches. Basically the approach lacked the active mediation of human agency, people as individuals, organized by social relations into families, classes, and institutions, and informed by different kinds of consciousness (Gregory and Urry, 1985).

Spatial Relations

Marxist theories see the relations between societies – often termed 'spatial relations' by Marxist geographers – in terms of exploitation, conquest, and domination. During the 1970s, a series of articles drew out the spatial implications of Marxist theory in a 'capital logic' school of spatial thought. This kind of Marxist geographic theory began with contradictions in the historical dynamic of capitalism and moved to manifestations of these contradictions in space. Spatial relations between environmentally embedded social formations are thought of as modifying, and even transforming, their internal social, economic, and political contents. Thus, the geography of capitalism is composed of unevenly and differently developing social formations. In such approaches, interregional social relations are often theorized as spatial components of a more general internal–external dialectic. For Harvey (1975), the Marxist theory of economic growth under capitalism puts capital accumulation at the center of a dynamic and inevitably expansionary mode of production. This dynamism encounters barriers (in labor, means of production, markets) which precipitate crises (chronic unemployment, realization crises, etc.) with the eventual result of a shift in the capital accumulation process to a higher plane (new social wants, new markets, etc.). Newly 'rational' locational patterns and improvements in transportation and communication are an inevitable, necessary part of capital accumulation. But there is a spatial aspect to contradiction: in overcoming spatial barriers and 'annihilating space with time,' spatial structures are created which ultimately act as barriers to further accumulation. This is particularly the case when capitalism comes to rely on immobile, fixed capital (i.e., means of production fixed in space) rather than more mobile variable capital (i.e., labor). Then capital has to negotiate a path between preserving the value of past investments in the built environment (the urban landscape, for example) and destroying these investments to make new room for accumulation (urban renewal). The virtue of Marx's location theory, dispersed though it may be in textual fragments, lies in the way space can be integrated into "fundamental insights into the production of value and the dynamics of accumulation" (Harvey, 1975: p. 13).

In terms of geographical theory this dynamic theory of contradictions contrasts with conventional (bourgeois) location theory employing an equilibrium analysis of optimal spatial configurations: as Harvey claimed (1975: p. 13):

> The Marxian theory … commences with the dynamics of accumulation and seeks to derive out of this analysis certain necessities with respect to geographical structures. The landscape which capitalism creates is also seen as the locus of contradiction and tension, rather

> than as an expression of harmonious equilibrium … The Marxian theory teaches us how to relate theoretically, accumulation and the transformation of spatial structures and ultimately … the reciprocal relationships between geography and history.

Thus, the internal and external dimensions of space are linked to each other and with capital accumulation. Revolutions in the productive forces, the increasing scale of production, and the concentration and centralization of capital are paralleled by urban agglomeration in a widening international capitalist space.

Smith (1984) argued that different conceptions of space were produced by different types of human activities; hence there was a history of theories of space forming part of the history of the human experience. The increasing scale of human intervention in nature, with advance of the forces of production, is seen as the historical-material basis of a theoretical bifurcation of absolute, natural, and physical space (the world of physical and natural phenomena) from relative and social space (the humanly constituted field of social events). For Smith, the development of the forces of production transforms absolute (natural) space into relative (social) space in a contradictory movement. Development emancipates society from the constraints of natural space, leading toward 'equalization,' but only by sinking capital into certain spaces, producing 'differentiation' and a relativized space. These contradictory movements determine the specific forms of capitalist space as synthetic outcome: "Space is neither leveled out of existence nor infinitely differentiated. Rather the pattern that results is one of 'uneven development'" (Smith, 1984: p. 90). Integrating the geographical notion of 'scales' (urban, national, global) as levels of specificity, with temporal notions of rhythms, cycles, and long waves, Smith proposes a 'seesaw' dynamic of uneven development: "capital attempts to see-saw from a developed to an underdeveloped area, then at a later point back to the first-area which is by now underdeveloped" (Smith, 1984: p. 149) – for example the return of some manufacturing to deindustrialized regions, or at the urban level, where mobile capital destroys, then gentrifies, inner city neighborhoods. For Smith, the real resolution between equalization and differentiation can be achieved only by global political cooperation among working-class people.

The culminating triumph in this series of creative exegeses came with Harvey's *The Limits to Capital* (1982), a reading of Marx's *Capital* organized around three versions of a general theory of capitalist crisis. The first version sees crises of overaccumulation – more capital produced than opportunities for using it – resolved through violent episodes of the destruction of value, times when capitalists turn on each other, yet workers end up paying the social costs (Harvey, 1982: pp. 190–203). The second version examines the expression of productive crises in money and finance, speculative booms and busts which require the intervention of the state: in an internationalized economy, crises take the form of interstate competition over shifting the effects of devaluation, with national policies to export the effects of crises, and imperialism, neocolonialism, and even war as potential 'solutions' (Harvey, 1982: pp. 324–329). The third and last version integrates geographically uneven development: devaluations of capital are systematized into the continuous restructuring of space through interregional competition. Harvey focuses on 'switching crises' in the movement of capital from one place to another and crises in the hierarchy of capitalist institutions, so that problems which originate in the chaos and confusion of local events build upward into global crises. For Harvey, the basic rhythm of crisis involves regional economies, or more broadly regional class alliances, periodically alleviating internal crises through transformations in external relations in a spatial 'fix' which again has imperialism as its integral component – hence, "an interpretation of interimperialist wars as constitutive moments in the dynamics of accumulation" (Harvey, 1982: p. 443). What better reason, he asks, could there be for declaring that capitalism's time has passed, and that there is a need for a saner mode of production?

What difference does theoretical elaboration make for understanding contemporary sociospatial issues? Take for example the emergence of an intensified process of globalization in the late twentieth and early twenty-first centuries. Conventional explanations of globalization stress the development of transportation and communication technologies as means of overcoming the frictions of space – hence a global, rather than a national, economy. And clearly this is part of the story. But an overemphasis on technology disguises an unequal process of the accumulation and extension of power. Power means control, by a person or an institution, over the minds, livelihoods, and beliefs of others. Power accumulates into systems. So the term 'geography of power' refers to the concentration of power in a few spaces that control a world of distant others. With globalization, new kinds of power system have arrived on the world scene. A few financial centers control the global economy, and crises caused by speculation on the part of the financial elite become world crises adversely affecting everyone. Power has increasingly been accumulated at the global level by governance institutions – the G7/G8, the European Union, the Bretton Woods Institutions, and the United Nations. These institutions control the economies and livelihoods of people all over the world through policy devices like structural adjustment, the 'conditionalities' accompanying loans or debt relief, annual inspections by teams of experts, and similar strategies. For Marxist geography, globalization is analyzed through types of power accumulation – economic, ideological, and political – and the flows of power over global space. Marxist geography also participates in popular resistances to concentrated power by social movements and organic intellectuals (Peet, 2007, 2011).

Politics of Space and Nature

Marxist geographers argue, in general, that the social inequalities endemic to capitalist societies result in spatial inequalities, regional and local differences within and between societies. The capitalist world is split between countries that have far more production and consumption than their peoples need, so that a range of problems appear, from overweight children to excessive pollution, and countries that have too little, with their own range of problems, such as starving children and extreme vulnerability to natural disasters like earthquakes and floods. Some Marxist theories of space, based in dependency and world systems theories (Frank, 1969; Wallerstein, 1974),

argue that overdevelopment in the First World is predicated on underdevelopment of the Third World. Uneven spatial development also brings entirely different degrees of economic pressure to bear on the natural world, so that environmental crises appear unevenly as they accumulate into global catastrophes. This interaction between Marxist geography's spatial and environmental dimensions provides the theoretical basis for a range of radical politics linking development with nature. Marxist spatial theorists advocate evening out the development process so that basic needs can be met everywhere. Marxist natural theorists advocate producing and consuming less, while doing this sustainably. Both require fundamental changes in social organization, especially in the ways societies relate over space, and the way nature is treated and necessarily altered during the social production of existence.

Social scientists in the Western tradition of Marxism favor rational and democratic ways of controlling society's environmental and spatial relations. The notion of state ownership of productive resources has always been far weaker in the Western tradition of Marxism than in the Soviet or Eastern tradition. Marxists believing in all-powerful states have long been countered by others believing just as strongly in workers' self-management, cooperative organizations, participatory democracy, and democratic planning. Marxists essentially believe in extending democracy from the political sphere into all aspects of social life, especially the productive and reproductive spheres – that is, democratic control over factories, offices, schools, and families. Marxists think that people directly responsible for their fundamental social and economic activities can reach collective decisions that are more rational and farsighted than those reached by political and economic elites distanced from direct contact with real social practices. In this way democratic socialism is seen by Marxist geographers as the final culmination of the Enlightenment project of rationality, equality, and liberation.

See also: Field Theory; Marxism in Contemporary Sociology; Resource Geography.

Bibliography

Althusser, L., Balibar, E., 1968. Reading Capital. New Left Books, London.
Castells, M., 1977. The Urban Question: A Marxist Approach (A. Sheridan, Trans.). MIT Press, Cambridge, MA.
Frank, A.G., 1969. Capitalism and Underdevelopment in Latin America. Monthly Review Press, New York.
Gramsci, A., 1971. Selections from the Prison Notebooks. International Publishers, New York.
Gregory, D., Urry, J., 1985. Social Relations and Spatial Structures. St. Martins Press, New York.
Harvey, D., 1973. Social Justice and the City. The Johns Hopkins University Press, Baltimore, MD.
Harvey, D., 1975. The geography of capitalist accumulation: a reconstruction of the Marxian theory. Antipode 7, 9–21.
Harvey, D., 1982. The Limits to Capital. University of Chicago Press, Chicago.
Harvey, D., 1989. The Condition of Modernity: An Inquiry into the Origins of Cultural Change. Oxford, UK, Blackwell.
Lefebvre, H., 1991. The Production of Space (D. Nicholson-Smith, Trans.). Oxford, UK, Blackwell.
Marx, K., 1976. Capital: A Critique of Political Economy, vol. 1. (B. Fowkes, Trans.). Penguin, Harmondsworth, UK.
Peet, R., 1977. Radical Geography. Maaroufa Press, Chicago.
Peet, R., 1998. Modern Geographical Thought. Oxford, UK: Blackwell.
Peet, R., 2007. Geography of Power: Making Global Economic Policy. Zed Books, London.
Peet, R., 2011. Contradictions of Finance Capitalism Monthly Review 63 (7), 18–32.
Smith, N., 1984. Uneven Development: Nature, Capital and the Production of Space. Oxford, UK, Blackwell.
Soja, E., 1989. Postmodern Geographies: The Reassertion of Space in Critical Social Theory. Verso, London.
Soja, E., 1996. Thirdspace: Journeys to Los Angeles and Other Real-and-imagined Places. Oxford, UK, Blackwell.
Wallerstein, I., 1974. The Modern World System, vol. 1. Academic Press, New York.

Masculinity and Social Work Practice

Bob Pease, Deakin University, Geelong, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Approaches to working with men will be shaped by how we understand masculinity and men's lives. This article explores the challenges and directions for profeminist practice with men in social work by locating an understanding of men's lives in the context of patriarchy and the divisions of class, race, sexuality, and other forms of social inequality. The article concludes by outlining the implications in practice, of a profeminist approach to men as gendered workers in social work, engaging men as individuals and in men's groups, and working for structural change at the community and societal levels.

Introduction

Masculinity and the place of men in social work has until recently been a neglected area in social work writing (Christie, 2001a; Featherstone et al., 2007; Pease and Camilleri, 2001; Pringle, 1995). There is extensive social work literature that presents men and masculinities in an implicit and untheorized way, with mainstream social work texts rarely addressing men as gendered clients.

Feminists in social work have drawn attention to the prevalence of sexist attitudes and practices in social welfare and, during the last 25 years, feminist social workers have developed interventions and policies aimed at overcoming sexism (e.g., Day and Langan, 1992; Dominelli, 2002; White, 2006). Feminist practice with women has thus made significant contribution to critical social work practice. While the feminist social work literature has not said much about working with men, they emphasized the need for feminist thinking to infuse work with men. Building upon these early discussions, Cavanagh and Cree (1996) published the first feminist book that systematically examined feminist practice with men.

It is not surprising that attention has been directed toward women's issues in combating sexism in the welfare and human services sector. The majority of both social workers and welfare clients are women. However, it is increasingly recognized that addressing sexism necessitates an understanding of men as well as women. All social workers will have contact with men. Furthermore, the majority of the concerns that women bring to social workers are connected to their relationships with men.

Since the 1990s, a small number of social work texts have been published to fill this gap in human services literature (Christie, 2001a; Featherstone et al., 2007; Pease and Camilleri, 2001; Pringle, 1995; Scourfield, 2003). These books address diverse issues related to men, masculinities, and social work. This short article endeavors to do justice to the complexities involved.

Men in a 'Women's Profession'

A continuing issue facing social work is the gender imbalance of the human services workforce. Evidence suggests that the proportion of women in social work in North America (Sakamoto et al., 2008), the United Kingdom (GSCC Report, 2010), and Australia (Healy and Lonne, 2010) is increasing.

The framing of social work as a caring profession and the association of caring with women is one of the barriers that deters some men from considering social work as a legitimate career option. There also seems to be some societal ambivalence about men as carers, especially if this caring involves physical and personal care (McLean, 2003).

While social work has been historically regarded as a 'women's profession' due to the female majority in its workforce, predominantly female clientele, and association with caring, many of the values associated with hegemonic masculinity such as emotional distance, rationality, and technical expertise also characterize social work (Pringle, 1995). Orme (2009) argues that social work practice is predominantly a rational-technical project that reflects dominant forms of masculinity.

Women's majority status in social work leads to social work being referred to as a female-dominated profession (Simpson, 2004). However, the larger numbers of women in social work do not necessarily translate into female leadership in the profession (Scourfield, 2003). Women in social work still work within the patriarchal structures of employing organizations and may at times serve patriarchal interests, even if they are actively challenging them.

McPhail (2004) challenges the representation of social work as a female-dominated profession because despite constituting the majority of members, women do not control the profession. Curriculum content is still dominated by male social work theorists; and quantitative and masculinist approaches to research still predominate, at least in the West. McPhail argues that framing social work as a female-dominated profession hides sexism and male prejudice within social work. In her view, it is more accurate to refer to social work as a "male-dominated female majority profession."

Men are overrepresented in both management and higher status jobs in the social services. Studies of social workers demonstrate that women are less likely than men to undertake managerial roles in human services and earn less money than men in comparable positions (Bent-Goodley and Sarnoff, 2008; Sakamoto et al., 2008).

These inequalities are reflected in various forms of horizontal and vertical occupational segregation in social work (Sakamoto et al., 2008). Notwithstanding the influence of

feminist ideas on social work, the social work profession retains patriarchal, gendered practices and these gendered practices can be understood as representing an unequal gender regime and provide the context in which social work with men needs to be located.

Masculinity and the Issues Facing Men in Social Work

Issues associated with masculinity underpin the problems that men present in human service organizations. The general consequences of traditional masculinity that have been identified include emotional inexpressiveness, poor health, distant fathering, family breakdown, stress associated with competitiveness and overwork, the overrepresentation of men in substance abuse, high risk behaviors, homelessness, suicide, and criminality (Lichtenberg, 1999).

Various writers have suggested that men will change when they consider these costs of masculinity for themselves. However, the focus on the problems facing men has placed too much emphasis on the costs for men associated with patriarchal relations and given insufficient attention to the costs of these male practices to women and children. In this regard, Cowburn and Pengelly (1999: p. 198) make the distinction between "men with problems" and "men as problems." The latter category includes rape and sexual assault, violence against women in the home, sexual harassment, and sexual misconduct as a group norm.

The privileging of men does not mean that men do not experience pain in their lives. Men can be privileged and miserable simultaneously. Kaufman (1993) argues that men's lives involve both power and pain and that much of men's pain arises from men's power and privilege to constitute what he calls "the contradictory experiences of men's power." Thus, the patriarchal dividend is not totally successful in advantaging men because men experience emotional and physical costs associated with their dominant position (Whitehead, 2002). Connell (1995) acknowledges that for men, there are disadvantages associated with gendered privilege. Connell (1995) posits that many of these costs are by-products of the advantages men gain from the patriarchal dividend.

When the costs of masculinity are documented in popular books about men, these are rarely framed as the consequences of men's advantages. Populist writers of books about men talk about a 'crisis in masculinity' as men find their traditional privileges and symbolic power being eroded (e.g., see Biddulph, 2008).

Several writers have expressed concern about the ways in which men's physical and mental health issues position men as the 'new disadvantaged' (e.g., Connell, 2000; Whitehead, 2002). Whitehead (2002) argues that "the male crisis discourse" distorts the connections between hegemonic masculinity and men's health. The idea of masculinity in crisis may itself be a strategy enacted by men to reinforce male power because it focuses on men's pain rather than their privilege. Connell (2000) points out that most of the costs associated with patriarchy for men are not necessarily experienced by the men who gain most of the benefits. Many of the current health policies and programs for men, for example, fail to recognize the social and economic context of men's lives and the impact of class and race divisions on their health (Pease, 2009b).

To develop a critical framework for practice with men, issues facing men have to be conceptualized. These are confusing and unsettling times for many men as they respond to changes in women's lives and women's demands for gender equality. To make sense of their confusion it is important to understand men's experiences within the context of the patriarchal structures in society and their relationship to class, race, and gender regimes. Men and women who work with men in social work should have an analysis of the social construction of masculinities and an understanding of how the forces that construct dominant masculinities embed men and women in relations of dominance and subordination that limit their potential to be in partnership with each other (Dominelli and McLeod, 1989).

To the extent that the social construction of masculinity is ignored, it blocks insight into the real trouble in men's lives. Furthermore, if men do not grasp gender as socially constructed, then feminist critiques of patriarchy, dominant masculinity, and abusive male behaviors will be felt by men at a deeply personal level.

Theorizing Men and Masculinities

Approaches to working with men will be shaped by understandings of masculinity and men's lives. Depending upon whether masculinity is framed primarily from a biological, psychological, or sociological perspective, it will impact the ways of understanding the issues facing men. It will influence whether the focus is more on men's pain or men's privilege (Featherstone et al., 2007). Work with men has also been differentiated in terms of whether it has been supportive of feminism and gender equality (e.g., Featherstone et al., 2007; Pease and Camilleri, 2001) or whether it is nonfeminist (or even antifeminist) and oriented toward reproducing traditional gender roles and inequalities (e.g., Glicken, 2005).

There is currently an enormous volume of literature on men and masculinities. Theoretical approaches have ranged from sociobiological, psychoanalytical, Jungian, and sex role approaches, through to materialist and discursive approaches. I do not provide an overview of the different theories, but focus specifically on theorizing men's gender privileges and situated dominance in relation to women.

In spite of the vast scholarship on men and masculinities published in the last 20 years and the developing awareness that challenges traditional thinking about male dominance, the 'common sense' approach to understanding men continues to invoke the language of biological differences. Claimed biological differences between men and women continue to legitimize male dominance and gender inequality.

Biological and essentialist perspectives on men and masculinities tend to legitimize male privilege. Stephen Goldberg's (1973) *The Inevitability of Patriarchy* is perhaps the best historical example of this argument. Goldberg argues that differences in concentrations of particular hormones give men an 'aggression advantage' over women. He believes this 'aggression advantage' means that men will inevitably

dominate women, as "the hormonal makes the social inevitable" (Goldberg, 1973: p. 49).

The central argument here is that if inequality is based upon some natural order, opposition is futile. In this view, there is no point in trying to equalize the genders because patriarchy is an inevitable product of biology. Nor is there any point in changing the basis of gender relations because it will only upset the natural order. Consequently, the feminist vision of a gender equal society is doomed to failure. This sociobiological analysis is 'alive and well' in much contemporary popular writing about men (Biddulph, 2008).

Psychological and sex role perspectives on men and masculinities posit that male behavior results from customary expectations that are internalized by men. In the 1970s, men were encouraged to "break…out of the straight-jacket of sex roles" (Farrell, 1975: p. 8). This theory was updated in the 1980s (Pleck, 1987); and from the 1990s it was reframed as gender role theory (Shimanoff, 2009) An implication of this approach is that men can transform themselves without reference to wider social processes, the male role being something men could dispose of, allowing the human being in the man to emerge.

One of the major limitations of sex-role theory (and the gender role theory that followed it) is that it underemphasizes the economic and political power that men exercise over women. Male and female roles are seen to be equal, thus enabling men and women to engage in a common cause against sex-role oppression. What is consistently missing in sex-role theory is the extent to which men's gender identities are based upon a struggle for social power. There is no evidence that liberating men from the traditional male sex-role will lead to men relinquishing their privilege and social power.

Sociological perspectives on men and masculinities emphasize that gender and masculinity are socially constructed throughout life. Following Connell (2000), it is most useful to understand men and masculinities as involving six key dimensions: (1) multiple masculinities that arise from different cultures, different historical periods, and different social divisions between men; (2) different positions reflected in these multiple masculinities in relation to power, with some forms of masculinity being hegemonic and dominant, while other masculinities are marginalized and subordinated; (3) institutionalized masculinities embedded in organizational structures and in the wider culture as well as being located within individual men; (4) embodied masculinities that are represented physically in how men engage with the world; (5) masculinities produced through the actions of individual men; and (6) fluid masculinities that change in relation to the reconstructive efforts of progressive men and in response to changes in the wider society. Within this theoretical context, Connell (1995) identifies four forms of masculinity: hegemonic, complicit, marginalized, and subordinate.

Hegemonic masculinity is the culturally dominant form of masculinity that is manifest in a range of different settings. Such masculinity is idealized and promoted as a desirable attainment for boys and young men to strive toward. It is presented as heterosexual, aggressive, authoritative, and courageous (Connell, 1995). The manliness of men and boys are judged by their ability to measure up to this normative notion. Sporting prowess, the role of men as providers, work status, and power over women are key signifiers of this form of masculinity. Hegemonic masculinity establishes the foundations for men's violence against women. Many writers have noted the close links between exalted masculinity and the legitimization of violence, especially in response to any form of threat (Messershmidt, 2000).

Although the majority of men do not adhere to hegemonic masculinity, it represents the most valued form of masculinity and all men are positioned in relation to it. It does not have to correspond to the lives of the majority of men to hold power over the male concept of being a man.

The majority of men engage in what Connell (1995) refers to as complicit masculinity. Complicit masculinities refer to those men who do not meet the normative standard of hegemonic masculinity but benefit from it in various ways. The concept of complicit masculinity is useful in understanding how men, who do not regard themselves as oppressors, reproduce male dominance and privilege.

Complicit masculinities also maintain the structures and ideologies that reproduce male violence. While the majority of men may not engage in excessive forms of hegemonic masculinity, they do not challenge patriarchy and male privilege that support this form of dominance, because they adhere to the dominant discourse and benefit from it.

Marginalized and subordinate masculinities are useful in understanding the relationship between gender and other dimensions of stratification, such as class, race, and sexuality. Connell (2000) uses these concepts to illustrate how diverse masculinities are marked by hierarchy and exclusion. The resultant hierarchy of masculinities means that men do not benefit equally from male privilege.

Diversifying Men and Masculinities

A new stage in which variations among men are seen as central to the understanding of men's lives has begun. It focuses not on masculinity as a singular term but explores masculinities. Men are as socially diverse as women and this diversity entails differences between men in relation to class, ethnicity, age, sexuality, bodily facility, religion, world views, parental/marital status, occupation, and propensity for violence.

Differences are also found across cultures and through historical time. The discourse about 'masculinity' is constructed out of 5% of the world's population of men, in one region of the world, at one moment in history. Ethnographic work in different cultures indicates how nonwestern masculinities can be very different from the western norm (Connell, 2000).

Anthropology has studied men's practices in nonwestern contexts. Some important studies are: Hertdt's (1981) study of the boy-inseminating rituals among the Sambia of Papua New Guinea; Silberschmidt's (1991) study of the changing gender roles in Kenya; Gutmann's (2003) edited papers on studies of masculinity in Latin America; and a study by Karim (1995) on gender relations in Asia, which considers the changing subjectivities and practices of men.

Cornwall and Lindesfarne (1994) emphasize the need to add the critical study of men's practices to attempts to

understand gender regimes in various parts of the world. Hofstede et al. (1998) examine different countries through how they mirror dimensions of masculinity and femininity, and construct a typology that measures the differences between what they call feminine and masculine societies. Gilmore (1990) provides one of the early comprehensive pieces of research on how men in different cultures perceive and experience manhood.

As a heterogeneous group, men do not benefit equally from gender domination. Issues of race, sexuality, class, disability, and age and the interaction between them complicate men's gender dominance in social work. Intersectional analyses can reveal experiences of both privilege and subordination. Multiple forms of social difference shape both unities and differences among men. These analyses provide a framework for designing interventions with men, which take into account their different placements in relation to privilege and oppression (Pease, 2009a).

Developing a Profeminist Sensibility among Men in Social Work

Men *can* change in the direction of feminism. They have choices as to whether they accept patriarchy or work collectively against it. Before men can organize collectively, they must transform their subjectivities and practices.

Although men cannot escape their material position in patriarchy, men can change their ideological and discursive position by drawing upon feminist standpoint theory (Harding, 2004). The advantage of standpoint theory is that it relates to both structural location and the discursive construction of subjectivity, to distinguish between a 'traditional man's standpoint' and a 'profeminist man's standpoint.'

A traditional man's standpoint is based on the privileges and powers that men have and it excludes the perspective of women. A profeminist man's standpoint involves an ability to be critical of man's traditional position in society and its contribution to the inequality of women and the development of an ethical and moral commitment to address inequality and discrimination because it causes harm (May, 1998).

Men can change their subjectivities and practices to constitute a profeminist man's standpoint. The process of change is itself a requirement in formulating a profeminist man's standpoint. Men have to change their vantage point if they want to see the world from a different position and this entails more than just a theoretical shift. It requires men actively to engage in profeminist struggles in both the private and public arenas.

Profeminism involves a sense of responsibility for a man's own sexism as well as other men's sexism, and a commitment to work with women to end men's violence (Pease, 2009a). It acknowledges that men benefit from the oppression of women, drawing men's attention to the privileges men receive as men and the harmful effects these privileges have on women. Profeminist men also recognize that sexism has an impact on men as well as women. Systemic male dominance dehumanizes men too, as evidenced in stress-related illnesses and emotional inexpressiveness.

Over 25 years ago, Tolman et al. (1986: pp. 66–73) articulated 11 principles for a profeminist commitment among men in social work: develop a historical, contextual understanding of women's experience; be responsible for themselves and other men; redefine masculinity; accept women's scrutiny without holding them responsible; support the efforts of women without interfering; struggle against racism and class consciousness; overcome homophobia and heterosexism; work against violence in all its forms; do not set up a false dichotomy between ones self and other men; act at individual, interpersonal, and organizational levels; and attend to process and product. They noted that very few men had been proactive in promoting feminist values such as these in social work and that men had contributed very little to ending sexism (Tolman et al., 1986). Written in 2013, this article reveals that little has changed. Nevertheless, men in social work have an important part to play in challenging sexism.

Encouraging the development of profeminist practice by men in social work is a natural expression of antioppressive practice with regard to gender. While antioppressive practice generally, with its focus on challenging oppression, is an appropriate framework within which to examine antisexist and profeminist practice by men (Cavanagh and Cree, 1996), Pease (2006) argues that antioppressive practice has been slow to encourage practitioners who are members of dominant groups to examine their own privileged positions.

To what extent then do men in social work recognize their gender privilege and to what extent are they willing to give up the privileges and power of their position? If men are to play an antioppressive role in relation to gender in social work, they will need to embrace a commitment to profeminist practice. In the context of wider antioppressive theory and practice, social work education needs to develop specific knowledge and skills to inform a profeminist commitment by men in social work. If men gain unearned benefits from their presence in social work, they have a responsibility to challenge the basis of those entitlements (Pringle, 1995).

Working with Men in Social Work

Pringle (1995) identifies four levels at which male social workers should engage in antioppressive practice: one's own behavior, individual work, group work, and structural change at the community and societal levels. This remains an useful framework to unravel the various dimensions of social work practice with men.

Working on One's Own Social Location

The starting point for profeminist practice for men, involves critical reflection upon the privileges associated with his individual position in the gender order and the gendered division of labor within social work. Christie (2001a) has stressed the importance of men as social workers, critically reflecting upon and challenging their gendered privilege and the gendered nature of their work. Pringle (2001: p. 45) similarly argues that male social workers "have a particular responsibility to engage in challenging those oppressive

structures of power from which they may benefit directly." This requires men in social work to monitor and challenge their own behavior and recognize the contradictions and dilemmas arising from their position.

Individual Work with Men

Men often want to change but they do not want to relinquish their power. A profeminist approach encourages men to rethink their power. It involves what Cree (2001: p. 161) calls 'critical engagement' – how to "build open connection with men while at the same time not being seen to condone their behaviour or attitudes."

Many of the beliefs men hold are the causes of the troubles in their lives. Thus, the starting point for work with men is to assess their beliefs (Russell, 1995). What beliefs does the man hold about masculinity? What are the sources of these beliefs? What are the potential harmful effects of these beliefs? How are these beliefs associated with the difficulties that the man is experiencing?

Male socialization leads to individual beliefs that can promote abusive behaviors (Russell, 1995). Men need to be helped by male social workers to acknowledge their tendencies to act oppressively and they should be assisted to devize strategies for avoiding those situations and changing their behavior. While men changing themselves will not in itself address the structural components of gender oppression, they should nevertheless be encouraged to develop wider repertoires of behavior and models of masculinity that are not associated with violence, control, and objectification (Pringle, 1995).

Group Work with Men

Antipatriarchal consciousness-raising group work with men can clarify the social dimensions and historical shifts in masculinities. Such consciousness-raising in men's groups can provide a link between personal experiences and the wider social context of men's lives. Men can come to understand their own sexist behavior and develop emotional support from other men to encourage their antisexism.

Profeminist group work emphasizes the importance of working in groups with men who oppress and disempower others, where male prejudices and oppressive behaviors can be challenged (Pringle, 1995). As Hearn (2001) comments, profeminist group work models educate men about the oppressiveness of their beliefs and behaviors and involve them in analyzing their use of power and control tactics, to enable them to move toward more equal relationships with women, within the limitations imposed by the structurally unequal gender relations within patriarchy.

Creating Change at Community and Societal Levels

Feminist and profeminist writers in social work emphasize that personal change does not go far enough and they advocate the importance of connecting personal change with cultural and structural changes (Cree, 2001). Work with men should be one aspect of a broader strategy for changing unequal power relations between women and men.

For example, men need to become more actively involved as partners with women in social movements against violence toward women (Pease, 2008). However, as the involvement of men in the prevention of male violence has become more widely supported, I have become concerned that the tensions, dilemmas and dangers of such involvement have not been adequately addressed. These dangers include: reducing funding for women's services and programs, weakening their feminist orientation, silencing women, taking over their campaign, colluding with violent men, gaining more praise, and failing to earn women's trust (Pease, 2008).

When developing alliances between women and men, particular conditions have to be met and specific principles adopted, to address the potential problems associated with such partnerships (Pease, 2008). The theoretical premises underpinning men's involvement in the prevention of violence has to be based upon feminist theory and critical scholarship on men and masculinities. This might ensure that involving men in the prevention of male violence does not replicate the same structures and processes that reproduce the violence being challenged.

Conclusion

Not examining the social construction of masculinities in social work leads to failure in acknowledging how masculinities impact women's lives. To the extent that issues of men and masculinity are discussed in social work education, the focus tends to be on the costs of sexism for men and how such costs may encourage men to be more responsive to feminism.

It is important that social work with men locates men's lives within an understanding of patriarchal gender relations and that such work is committed to transforming those gender relations. While the gendered belief systems of men are only a part of the processes that perpetuate gendered inequality, they are amenable to intervention at the individual and collective level. Challenging the abusive beliefs of men makes an important contribution to respectful and egalitarian practices (Featherstone et al., 2007).

If more men are to be recruited into social work, the debate about the gendered nature of caring work has to be reignited within the profession and in social work education. The social work professionals and social work educators need to interrogate the gendered assumptions about nurturing, caring, intimacy, and love which are embedded within social work. Ascribing caring to women and femininity reproduces patriarchal discourses by devaluing women's practices in social work and legitimizing men's avoidance of caring (Camilleri and Jones, 2001). The challenge for the profession is to encourage men and women in social work to understand how affective inequalities in the delivery and the acceptance of care and love, reproduce inequalities in economic, political, and social relations (Lynch et al., 2009). Thus, fostering caring and nurturing masculinities is an essential part of promoting gender equality in the profession and wider society.

There are progressive and important roles for men in social work in promoting a more egalitarian profession. These require clarity about what constitutes 'good practice' for men and awareness of the dangers and problems, as well as the

possibilities, associated with increasing male presence in social work.

Some men in social work do not subscribe to dominant forms of masculinity. Social work as a profession needs more men who are critical of hegemonic masculinity and who are aware of the gendered injustices associated with male privilege and unearned male entitlements. I have argued that if men in social work are not actively promoting profeminist practices, they will reproduce an unequal gender regime within the profession.

See also: Fathers and Fathering: Social Work Perspectives; Gender History; Gender and Place; Gendered Care and Social Work Practice; Masculinity, Scripts of; Social Work Theory.

Bibliography

Biddulph, S., 2008. Raising Boys: Why Boys are Different and How to Help Them Become Happy and Well-balanced Men. Celestial Arts, Berkeley, CA.

Bent-Goodley, T., Sarnoff, S., 2008. The role and status of women in social work education: past and future considerations. Journal of Social Work Education 44 (1), 1–8.

Camilleri, P., Jones, P., 2001. Doing 'women's work'? Men, masculinity and caring. In: Pease, B., Camilleri, P. (Eds.), Working with Men in the Human Services. Allen and Unwin, Sydney, pp. 25–33.

Cavanagh, K., Cree, V. (Eds.), 1996. Working with Men: Feminism and Social Work. Routledge, London, pp. 45–64.

Christie, A. (Ed.), 2001a. Men and Social Work: Theories and Practices. Palgrave, London.

Connell, R., 1995. Masculinities. Allen and Unwin, Sydney.

Connell, R., 2000. The Men and the Boys. Allen and Unwin, Sydney.

Cornwall, A., Lindesfarne, N., 1994. Introduction. In: Cornwall, A., Lindesfarne, N. (Eds.), Dislocating Masculinity: Comparative Ethnographies. Routledge, London, pp. 1–11.

Cowburn, M., Pengelly, H., 1999. Values and processes in groupwork with men. In: Wild, J. (Ed.), Working with Men for Change. UCL Press, London, pp. 197–206.

Cree, V., 2001. Men and masculinities in social work education. In: Christie, A. (Ed.), Men and Social Work: Theories and Practices. Palgrave, London, pp. 147–163.

Day, L., Langan, M. (Eds.), 1992. Women, Oppression and Social Work. Milton Park, Routledge, Oxon.

Dominelli, L., 2002. Feminist Social Work Theory and Practice. Palgrave, Houndmills, Basingstoke.

Dominelli, L., McLeod, E., 1989. Feminist Social Work. Macmillan, London.

Farrell, W., 1975. The Liberated Man. Bantam, New York.

Featherstone, B., Rivett, M., Scourfield, J., 2007. Working with Men in Health and Social Care. Sage, London.

General Social Care Council, 2010. A Report on Social Work Education in England 2009–2010. London.

Gilmore, D., 1990. Manhood in the Making: Cultural Concepts of Masculinity. Yale University Press, New Haven.

Glicken, M., 2005. Working with Troubled Men: A Contemporary Practitioner's Guide. Lawrence Erlbaum, Mahwah, NJ.

Goldberg, S., 1973. The Inevitability of Patriarchy. William Morrow, New York.

Gutmann, M. (Ed.), 2003. Changing Men and Masculinities in Latin America. Duke University Press, Durham.

Harding, S. (Ed.), 2004. The Feminist Standpoint Theory Reader. Routledge, New York.

Healy, K., Lonne, B., 2010. The Social Work and Human Services Workforce. Report from a National Study of Education, Training and Workforce Needs. Australian Learning and Teaching Council, Strawberry Hills, NSW.

Hearn, J., 2001. Men, social work and men's violence towards women. In: Christie, A. (Ed.), Men and Social Work: Theories and Practice. Palgrave, London, pp. 63–86.

Herdt, G., 1981. Guardians of the Flute. McGraw-Hill, New York.

Hofstede, G., Arindall, W., Best, D., De Mooij, M., Hoppe, M., Van de Vliert, E., Van Rossum, J., Verweji, J., Williams, J., 1998. Masculinity and Femininity: The Taboo Dimension of National Cultures. Sage, Thousand Oaks, CA.

Karim, W. (Ed.), 1995. 'Male' and 'Female' in Developing South East Asia. Berg Publishers, Oxford.

Kaufman, M., 1993. Cracking the Armour: Power, Pain and the Lives of Men. Viking, Toronto.

Lichtenberg, P., 1999. Men: Overview, in Encyclopedia of Social Work, nineteenth ed. National Association of Social Workers, Washington DC. pp. 1691–1697.

Lynch, K., Baker, J., Cantillon, S., Walsh, J., 2009. Which equalities matter? The place of affective equality in egalitarian thinking. In: Lynch, K., Baker, J., Lyons, M. (Eds.), Affective Equality: Love, Care and Injustice. Palgave Macmillan, Houndmills, Basingstoke, pp. 12–34.

May, L., 1998. A progressive male standpoint. In: Digby, T. (Ed.), Men Doing Feminsm. Routledge, New York, pp. 337–354.

McLean, J., 2003. Men as minority: men employed in statutory social care work. Journal of Social Work 3, 45–68.

McPhail, B., 2004. Setting the record straight: social work is not a female-dominated profession. Social Work 49 (2), 323–326.

Messerschmidt, J., 2000. Nine Lives: Adolescent Masculinities, the Body and Violence. Westview Press, Boulder.

Orme, J., 2009. Feminist social work. In: Gray, M., Webb, S. (Eds.), Social Work: Theories and Methods. Sage, London, pp. 65–75.

Pease, B., 2006. Gendering men: implications for gender equality. In: Varanka, J., Narhinen, A., Siukola, R. (Eds.), Men and Gender Equality: Towards Progressive Policies. Ministry of Social Affairs and Health, Helinski, Finland, pp. 39–47.

Pease, B., 2008. Engaging Men in Violence Prevention: Exploring the Tensions, Dilemmas and Possibilities. Australian Domestic and Family Violence Clearinghouse. Issues Paper 17, August.

Pease, B., 2009a. Challenges and directions for profeminist practice with men. In: Allan, J., Briskman, L., Pease, B. (Eds.), Critical Social Work: Theories and Practices for a Socially Just World, second ed. Allen and Unwin, Sydney, pp. 160–174.

Pease, B., 2009b. Racialised masculinities and the health of immigrant and refugee men. In: Broom, A., Tovey, P. (Eds.), Men's Health: Body, Identity and Social Context. Wiley-Blackwell, Chichester, West Sussex, pp. 182–201.

Pease, B., Camilleri, P. (Eds.), 2001. Working with Men in the Human Services. Allen and Unwin, Sydney.

Pleck, J., 1987. The male sex role. Journal of Social Issues 32, 155–164.

Pringle, K., 1995. Men, Masculinities and Social Welfare. UCL Press, London.

Pringle, K., 2001. Men in social work: the double edge. In: Christie, A. (Ed.), Men and Social Work: Theories and Practice. Palgrave, London.

Russell, M., 1995. Confronting Abusive Beliefs: Group Treatment for Abusive Men. Sage, Thousand Oaks, CA.

Sakamoto, I., McPhail, B., Anastas, J., Colarossi, L., 2008. Status of women in social work education. Journal of Social Work Education 44 (1), 37–62.

Scourfield, J., 2003. Gender and Child Protection. Palgrave, Houndmills, Basingstoke.

Shimanoff, S., 2009. Gender role theory. In: Littlejohn, S., Foss, K. (Eds.), Encyclopedia of Communications Theory. Sage, Thousand Oaks, CA, pp. 434–437.

Silberschmidt, M., 1991. Rethinking Men and Gender Relations. Centre for Development Research, Copenhagen.

Simpson, R., 2004. Masculinity at work: the experiences of men in female dominated occupations. Work, Employment and Society 18 (2), 349–368.

Tolman, R., Mowry, D., Jones, L., Brekke, J., 1986. Developing a profeminist commitment among men in social work. In: Van Den Bergh, N., Cooper, L. (Eds.), Feminist Visions for Social Work. National Association of Social Workers, New York, pp. 61–79.

White, V., 2006. The State of Feminist Social Work. Routledge, London.

Whitehead, S., 2002. Men and Masculinities: Key Themes and New Directions. Polity Press, Cambridge.

Masculinity, Scripts of

Elizabeth Nagel, Rachel Kalish, and Michael Kimmel, Stony Brook University, Stony Brook, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores contemporary scripts of masculinity. Scripting theory allows for behaviors to be examined in regard to individual actors, interactions between individuals, and also how these are influenced by culture more broadly. Each of these levels of scripting is discussed independently as it relates to the construction of masculine scripts. Although these scripts guide individuals in their social interactions, it is necessary to understand how multiple scripts emerge. Here, the ways in which masculine scripts are constrained by social structures to produce particular masculinities are examined in regard to age, sexual orientation, class, and race.

Scripting Theory

Scripting theory was developed by Gagnon and Simon in the late 1960s and early 1970s. This theory has had an incomparable impact on the field of gender and sexuality studies within sociology. Simon and Gagnon looked for explanations of human sexual behavior that went beyond the biological. They write, "Unlike most sociologists of a generation earlier than us, our preference was to yield as little as possible to the explanatory power of biology in any of its guises" (2003: p. 492). As students of the Chicago School, they were well versed in the notion of the self, and applied this train of thought to gender. In so doing, they foreshadowed the movement toward social constructionism that was to come, first applying aspects of social learning to gender. The shift from social learning to social construction implied more agency on the part of the actor, as well as a recognition or rejection of social norms and cultural narratives around behavior. Simon and Gagnon saw the enactment of gender as central to one's identity, and were able to see the linkages between private acts and public expectations, positing that gender was learned through interaction with others, and also within the large-scale cultural narratives that proscribe ways to behave.

The underlying ideology of sexual scripts, the idea that meanings are generated through social interaction, paved the way for a variety of work, not only in sexuality studies within sociology, but in gender studies across disciplines. The most significant of these subsequent trends is the idea of gender as a performance. As Kimmel writes

> This effort to understand both the multiple meanings of sex and sex's centrality in the construction of identity set Gagnon and Simon's work apart as revolutionary; indeed, their breakthrough book *Sexual Conduct* (1973) heralded the new paradigm from which all subsequent readings of sexuality in the social sciences and humanities have sprung. Without Gagnon and Simon, there could not have been a Judith Butler. (Kimmel, 2007: p. ix)

Levels of Scripting Theory

Simon and Gagnon recognized that cultural context tells individuals ('actors') what is appropriate to do, when, and with whom. They also recognized that interpersonal interaction shapes behavior, so that when one person violates a social norm or expectation, they are not likely to repeat the act in the same way. They did also honor the personal choices and fantasies of the actor (the intrapsychic); however, they see these as rooted in social norms and expectations instead of individual biology. These three interrelated levels of scripts allow for gender and sexuality to be fluid and change over time and across cultural contexts. One of the major contributions of their work is that they accounted for differences in gender, foreshadowing the move of modern gender theory, feminist theory, and queer theory toward social constructionism.

The cultural element of scripts describes the large-scale messages that exist in different societies about gender. There is a set of cultural messages that proscribe behavior in a general context, as well as sets of messages that are more specific to different contexts, subcultures, and historical time periods. These messages are ubiquitous, but because specific messages may contradict one another, their utility as general guidelines for behavior can be problematic.

Many of the cultural messages about gender and appropriately gendered behaviors are simultaneously messages about (hetero)sexuality, as the enactment of (hetero)sexuality is an important element of the masculine script. These cultural narratives set up gender as if their categories were mutually exclusive, so that any behavior, body, or actor that is not explicitly masculine must be somehow feminized, and behavior that is not adequately feminine is denigrated as masculine, or just inappropriate.

One of the enduring cultural messages, albeit rooted in biology, indicates that men are geared to seek out multiple sexual partners, while women are supposed to guard their fertility and therefore be more selective about when and with whom they procreate. As these messages endure throughout decades, they translate into the often-repeated ideology, which dictates that males should seek sex and women should seek love (McGinty et al., 2007). This translates into a large-scale acceptance of male pleasure-seeking behavior, as well as providing a rationale that neutralizes such behavior: it is often excused as 'boys being boys.' Moreover, the ubiquity of such messages reinforces the masculine script as an enactment of heterosexuality.

The interpersonal element of gender scripts is also significant in that the communication among peers is used to deploy new

meanings or evaluations of behaviors. The opinions of the peer group are instrumental in shaping the way an actor behaves, as well as how one evaluates the behavior of others. It is within the interpersonal context that a label, such as 'fag,' is potentially devastating for young boys (Pascoe, 2007), as it is a public statement about an unsuccessful performance of masculinity. If a young male sees a peer labeled as 'fag,' he is less likely to engage in a similar behavior. In this way, the peer group is a significant regulator of behavior (Roseneil, 2005) and simultaneously becomes a site where one can earn or lose status (Kimmel, 2008; Kalish and Kimmel, 2011; Armstrong et al., 2009, 2012; Arnold, 2010; McGinty et al., 2007; Paul, 2006).

The third element of sexual script theory, the intrapsychic, is the least-theorized (Plummer, 2005, 2007) aspect of scripting. Relating to the individual fantasies, thoughts, and sensations that affect gender and sexuality, the intrapsychic level is significant because it allows for the introduction of pleasure. Pleasure, or corporeal sensation, is an important element of sexuality, and it affects how choices are inscripted on the body (Powell, 2008). Powell (2008) sees this as an embodied gender practice: "In using this term, I refer to the ways in which gendered norms and discourse are enacted through the body in everyday practice; in thoughts, feelings, desires, and responses, in a way that is not always subject to individual recognition and change" (p. 170). Following the work of Pierre Bourdieu, Powell (2008) posits that this embodied gender practice is the product of the culmination of social construction and social structure inscribed on the bodies of actors, a sexual habitus, if you will. While recognizing the intrapsychic as another site in which gendered messages influence conduct, scholars (Powell, 2008; McNay, 2000) also see the possibility for the creation, or renegotiation of new gendered selves.

As a theory, sexual scripts have remained fluid, and have been extended to explain sexual behavior in a variety of contexts. Indeed, the benefit of the three interrelated levels of sexual scripts (cultural, interpersonal, and intrapsychic) allows for different scripts to emerge in different historical and cultural contexts.

Critiques

While the theory of sexual scripts has remained viable over the decades since *Sexual Conduct* came out, it has been challenged. Perhaps the biggest challenge to sexual scripts is that it does not account for the corporeal. The intrapsychic level of scripting allows for behavior to be generated from fantasy, but the script theory does not explicitly or implicitly discuss the meanings of feelings, activities, and sensations of the body. As the sociology of the body has developed, this criticism has gained momentum. Plummer (2005) is not suggesting that script theory itself does not allow for involvement of the bodily reactions to sex, rather that much work in sexuality has taken the actual bodies that perform sex for granted. The corporeal body is important in guiding decision making, even if the actors themselves are not able to situate those corporeal feelings into the larger social and cultural context promoting sexuality, or to a lesser extent, to be able to interpret those feelings as being mediated by their peers and the culture in which they live.

Another challenge to scripting theory, acknowledged by Simon and Gagnon themselves, is that while the three levels of sexual scripts are interrelated, and one level may offer a context through which the others may be viewed, the three levels (cultural, interpersonal, and intrapsychic) do not offer 'the power to provide a comprehensive understanding of the other' (Simon and Gagnon, 2003: p. 493). Again, more recent work on specific cultural settings such as fraternities or colleges offers some insight into how structural forces affect behavior, but more work could be done to explain how these forces impact the interpersonal and intrapsychic implications for the actors.

A third challenge to sexual scripts is a more postmodern reading, calling for attention to structural inequalities and their impact on sexual scripts (Jackson, 2007), which we have attempted to address above. According to Stevi Jackson:

> The locations we find ourselves in are shaped by material social circumstances and our location within gendered, classed, racialized, and sexual hierarchies; our definitions of those situations and, in particular, whether our definition "works" for us depends on our social and cultural resources. If we take seriously the idea that our sexual lives are simply one facet of our broader social lives, then we must recognize that our sexual selves are shaped within profoundly unequal local, national, and global social orders. (Jackson, 2007: p. 13).

Hence, structural inequalities cannot be ignored when looking at how social forces affect individual behaviors. In an increasingly global world, where large-scale inequalities persist even as information flows rapidly, factors affecting the sexual script for one actor will change for another actor with access to more resources, information, technology, social recognition, and support.

Specific Contexts and Scripts of Masculinities

Masculinities across the Life Course

Due to their recursive relationship with cultural narratives about appropriate behaviors, scripts of masculinity change over the life course. As a male ages, his enactment and understanding of the masculine script will be reinforced by various sources of socialization: at early ages, the family is an important regulator of gendered behavior; through adolescence and emerging adulthood, the peer group becomes more important, and throughout adulthood and middle age, both distal and proximal sources can reinforce behavior.

Regulation of gender enactment starts early, sometimes before birth, as parents will project notions of gender to the behaviors of an infant, characterizing the instinctual grasp of a finger as strength. Parents regulate the enactment of gender in their children by rewarding appropriately gendered behavior and by punishing or correcting other behavior. Young boys learn the script of masculinity early on, encouraged to engage in rough play and physical activity, while being steered away from expressive toys. Most importantly, boys are taught that they are not girls, and they learn that to be associated with anything feminine is bad, for example, as they are instructed not to cry or show emotion. In childhood, the masculine script emphasizes difference from the feminine.

Schools cement gendered scripts by segregating boys from girls in gym classes, and by guiding boys into different, more physical, activities from their female peers, reiterating the emphasis on difference, and separation from anything feminine. During adolescence, peers become a significant reinforcement of the masculine script. It is during adolescence that the masculine script begins to emphasize the enactment of heterosexuality. In adolescence, the distance between boys and girls grows, and within the male peer group, heterosexual performance is emphasized (Pascoe, 2007). Also during this time, the enactment of successful masculinity becomes more public, and aggression, dominance, and violence are valued, as is the denigration of anything feminine.

In emerging adulthood, the masculine script underscores the performance of heterosexuality, and begins to stress dominance over anything feminine; so successful male performances will not only show a difference from females, but also display dominance over females sexually, as male sexual aggression is normalized and condoned. Also during this time, views about appropriate behavior emerge, highlighting the double standard regarding sexuality, where males will be rewarded for having multiple sex partners (i.e., positively termed 'a player'), while females will be scorned for similar behavior (i.e., with the label of 'slut') (Hamilton and Armstrong, 2009).

Emergent adulthood is an important liminal stage, where males are affected by the combination of scripts geared toward adolescent males, as well as those geared toward older men. In this time period, we see some elements of the adult masculine script emerge, such as the emphasis on responsibility and the status enhancing the effect of conspicuous consumption. Due to this, some young men may feel a sense of pressure in terms of keeping up with their peers in regard to both sexual accomplishment with women as well as with an accumulation of material goods that display wealth and status.

Because the cultural narratives, which shape gendered scripts, are themselves changing across time, culture, social class, and geographic region, we can see different scripts for different types of masculinities. However, many elements of hegemonic masculinity endure over time. In the words of Erving Goffman (1963, cited in Kimmel, 2000: p. 11), the successful performance of masculinity involves a man who is: "a young, married, white, urban, northern, heterosexual, Protestant, father, of college education, fully employed, of good complexion, weight, and height, and a recent record in sports". The emphasis on bodily strength, athleticism, and gainful employment are all central to the successful enactment of masculinity, even in more specific scripts of masculinity 50 years after this quote was originally written.

Masculine Sexual Scripts

The heterosexual masculine script is quite similar to the hegemonic masculine script, prizing conquest of women and an insatiable libido. Within this script, males are rewarded for sexual prowess, and they are able to earn status with their peers for being able to 'get' highly prized sexual partners. The ability to engage in sex acts without emotion is also emphasized in the masculine script, as emotion itself is typically associated with femininity. This script is largely devoid of meaningful interpersonal relations, as hegemonic heterosexual masculinity rewards sexual activity, and sexual behavior becomes a means for establishing status with male peers more than a means by which to establish a meaningful connection with female partners (Kalish and Kimmel, 2011; Kimmel, 2008).

Despite the fact that the appropriate intimate sexual relationships for men are with women, their appropriate social relationships are with other men. This social distance from women reinforces the idea that women are culturally and socially inferior to men. Thus, men's social lives are focused largely around their relationships with other men through sports, social clubs, fraternities, and other all-male institutions (Flood, 2008; Kimmel, 1996). These homosocial relationships put men at constant risk for that intimacy to be seen as sexual rather than as purely social. In order to maintain their masculine identity, men must socialize with other men, but do so in a manner that puts them above suspicion of homosexuality by relying on regular displays of homophobia to ensure that their masculinity is maintained (Kimmel, 1996).

The script for homosexual masculinity both draws from and simultaneously serves as a challenge to hegemonic masculinity. Although homosexuality is frequently tied to femininity, homosexual masculine scripts often reject femininity, asserting their masculinity despite their sexual orientation. Similar to the script for heterosexual masculinity, the script for gay men emphasizes their sexual prowess. Additionally, the script for homosexual masculinity stresses the importance of physical appearance, especially that of one's partner (Klinkenberg and Rose, 1994). Like their heterosexual peers (Kimmel, 2008; Kalish and Kimmel, 2011), gay men gain status by having an attractive partner. Also in accordance with the masculine script for heterosexual men, sex is prioritized over emotion. However, the script for gay men allows for more emotionality than heterosexual men, as well as some encouragement of more expressive emotions. Particularly, increased concern regarding safer sex practices is stressed more in the script for gay men than their heterosexual peers (Klinkenberg and Rose, 1994).

Class-Based Masculinities

Class plays an important role in the ways in which particular men enact masculine scripts. While all men may aspire to the hegemonic ideal of masculinity, not all men can realize that ideal in the same ways. As wealth is directly implicated in realizing that ideal, men, particularly in the lower or working class must often rework their enactments of the masculine script in ways that reinforce their own masculinity as another legitimate display of masculinity. As such, men are structurally constrained in their choice, an enactment of masculine scripts.

One of the dominant features of masculinity, particularly for white males, has been the role of provider and breadwinner, and one of the primary challenges to white working-class masculinity involves the threat to that role that has occurred with the increasing numbers of women and minorities entering the workforce (Kimmel, 1996; Morgan, 2005). These changes directly affect men's ability to serve as the provider in their families in several ways. First, the entrance of both women and minorities has led to a decline in wages, frustrating men's ability to be the sole provider in the family. Secondly, increased competition for jobs creates hostility between men for scarce jobs. In this process, working-class

men, especially white working-class men, often turn to blaming other minorities for their inabilities to find work or to find work that will support their families. They see themselves as hard-working men who are being deprived of the opportunity to support their families. Even those who are employed often feel frustrated with low-wage jobs that provide them with little authority, autonomy, or social status (Morgan, 2005; Weis, 2006).

The frustration of lack of autonomy in the workplace can carry over to the home, where men feel that they are able to exercise their control (Morgan, 2005). Even though working-class men may be dissatisfied with their ability to fulfill the demands of the provider role, they are still the heads of their families. This allows working-class men assert their dominance in the home, reinforcing the separation between the masculine and the feminine. Even in situations where the woman in the home works, her work is generally seen as secondary to his work and his role as provider (Morgan, 2005).

In addition to reifying the differences between themselves and women in the home, working-class men may also draw on sexism and homophobia to effeminize other men. Working-class jobs often involve hard, physical labor, and working-class men claim this as a superior form of masculinity to middle- and upper-class men, whose jobs require little physical labor, leaving their bodies softer and more effeminate. Men's distancing themselves from middle- and upper-class men in this way reframes aspects of masculinity to conform to a working-class script (Kimmel, 1996; Morgan, 2005; Pyke, 1996).

Racialized Masculinities

Just as there are structural constraints to masculine scripts based on class, race/ethnicity is another social context, which constrains and produces alternate masculine scripts. Race is often implicated in the discussion of class, but also serves as a unique context in which men learn masculine scripts. It is important to remember that these scripts are not only racialized, but localized as well. Here, the American context specifically is considered.

African-American masculinity has been uniquely affected by the social history of the United States. While some minority racial and ethnic groups have seen greatly increased acceptance into US society, African-Americans, particularly African-American men, still face a great deal of social discrimination, and are limited to particular scripts of masculinity. It is important to recognize that most scholarship on African-American men speaks about impoverished urban African-American men, who face multiple intersecting forms of oppression (Collins, 2004).

Acting as a provider is central to African-American men's perceptions of masculinity, however, African-American men face additional challenges in being able to accomplish this goal (Collins, 2004; Payne, 2006). Although this may seem similar to the problem of white working-class males, African-American men are faced with additional structural barriers to accomplishing this ideal, often competing for the same jobs as white men, while they lack the same skills, education, and must contend with enduring racism (Collins, 2004). These problems often create a conflict for the African-American men, specifically in their romantic relationships. Facing structural inequality, alongside criticism at home for not fulfilling their obligations to the family, creates a great deal of tension between African-American men and women (Franklin, 2001). Here, African-American men and women are placed in opposition to one another, and in order to avoid seen as being controlled by their women; men must exert their control over their women to assert their male dominance. Although these may not be the only possible outcomes, these are the culturally prescribed scripts available to African-American men (Collins, 2004).

Asian-American men have also developed a script to respond to stigmatization by the larger culture. Asian-American men have been stigmatized as both effeminate and sexually impotent by the larger society, denying their ability to fulfill the script of hegemonic masculinity, while simultaneously being upheld as a 'model minority.' In response to these labels, Asian-American masculinity has developed a less rigid masculine script that challenges these cultural messages to varying degrees. Asian-American men explicitly challenge the notion that they are somehow sexually inferior, instead enacting an active and adventurous sexual script. They also challenge the more traditional masculine script by engaging in traditionally feminine tasks, such as housework, but incorporating these tasks into their own masculine scripts. In this way, the Asian-American masculine script is more egalitarian and flexible than other masculine scripts (Chua and Fujino, 1999; Shek, 2006).

Conclusion

Masculine scripts are embedded within layers of social context. While it is possible to tease apart these layers analytically, men experience their gender, race, class, and sexuality simultaneously and in particular social situations. The position of these identities against one another and the uncertainty of social situations in general can create ambiguity for men regarding which script is appropriate in a particular situation.

Though much of the description here is of cultural scripts, these directly impact interpersonal scripts. Cultural messages inform our everyday interpersonal interactions in the broadest sense. We rely on these cultural scripts to provide us general guidelines for action, but it is through our interpersonal interactions that we refine and redefine our interpersonal scripts. Some of these interpersonal scripts exist as general guidelines of social interactions with others in particular social settings, while the interpersonal scripts that we develop in more intimate settings may be highly particularized (Gagnon and Simon, 2005; Simon and Gagnon, 1984, 2003).

At the level of the intrapsychic, cultural scripts shape the hopes, desires, and expectations that each of us carry into our interactions with others. Of course, these cultural messages are internalized in unique and highly individual ways; individuals may embrace these messages wholeheartedly, reject them completely, or, more often, interpret them in their own ways. But in social interactions, people can not only consider their personal desires, individuals must manage these intrapsychic scripts in their interactions with others, choosing when, how, and if they will reveal their personal desires to others who may have conflicting aspirations (Gagnon and Simon, 2005; Simon and Gagnon, 1984, 2003).

Scripting theory allows us to examine each of these levels, and the ways that they interact with one another to gain a greater understanding of the meanings individuals bring into

social situations and how meanings are created through those interactions. It allows us to understand that while masculinities may be socially constrained, they are simultaneously being constructed through social interactions. This creation of meaning through interaction means that these scripts are not static, but constantly being challenged and redefined, leaving open the possibility for alternative masculine scripts to emerge.

See also: Adolescent Sexual Risk; Femininities; Heterosexual Expression at Older Ages; Hooking up and Casual Sex; Intimate Partner Violence; Life Course: Sociological Aspects; Sexual Orientation: Historical and Social Construction; Sexuality Over the Life Course; Sexuality, Theories of.

Bibliography

Armstrong, E.A., England, P., Fogarty, A.C.K., 2009. Orgasm in college hookups and relationships. In: Risman, B. (Ed.), Families as They Really Are. WW Norton, New York, pp. 362–377.

Armstrong, E.A., England, P., Fogarty, A.C.K., 2012. Accounting for women's orgasm and sexual enjoyment in college hookups and relationships. American Sociological Review 77 (3), 435–462.

Arnold, K.D., 2010. College student development and the hook up culture. Journal of College and Character 11 (4), 1–10.

Chua, P., Fujino, D.C., 1999. Negotiating new Asian-American masculinities: attitudes and gender expectations. The Journal of Men's Studies 7 (3), 391–413.

Collins, P.H., 2004. Black Sexual Politics: African Americans, Gender, and the New Racism. Routledge, New York.

Flood, M., 2008. Men, sex, and homosociality: how bonds between men shape their sexual relations with women. Men and Masculinities 10 (3), 339–359.

Franklin, C.W., 2001. Black male-black female conflict: individually caused and culturally nurtured. In: Kimmel, M.S., Messner, M.A. (Eds.), Men's Lives. Allyn and Bacon, Boston, pp. 139–154.

Gagnon, J.H., Simon, W., 2005. Sexual Conduct: The Social Sources of Human Sexuality, second ed. Aldine Transaction, New Brunswick.

Goffman, E., 1963. Stigma: Notes on the Management of Spoiled Identity. Prentice Hall, New Jersey.

Hamilton, L., Armstrong, E.A., 2009. Gendered sexuality in young adulthood: double binds and flawed options. Gender and Society 23 (5), 589–616.

Jackson, S., 2007. The sexual self in late modernity. In: Kimmel, M. (Ed.), The Sexual Self: The Construction of Sexual Scripts. Vanderbilt University Press, Nashville, pp. 3–15.

Kalish, R., Kimmel, M., 2011. Hooking up: hot hetero sex, or the new numb normative? Australian Feminist Studies 26 (67), 137–151.

Kimmel, M.S., 1996. Manhood in America: A Cultural History. Free Press, New York.

Kimmel, M.S., 2000. The Gendered Society. Oxford University Press, New York.

Kimmel, M.S., 2007. The Sexual Self: The Construction of Sexual Scripts. Vanderbilt University Press, Nashville.

Kimmel, M.S., 2008. Guyland: The Perilous World Where Boys Become Men. Harper, New York.

Klinkenberg, D., Rose, S., 1994. Dating scripts of gay men and lesbians. Journal of Homosexuality 26 (4), 23–35.

McGinty, K., Knox, D., Zusman, M.E., 2007. "Friends with benefits" women want "friends", men want "benefits". College Student Journal 41 (4), 1128–1131.

McNay, L., 2000. Gender and Agency: Reconfiguring the Subject in Feminist and Social Theory. Polity Press, Cambridge.

Morgan, D., 2005. Class and masculinity. In: Kimmel, M.S., Hearn, J., Connell, R. (Eds.), Handbook of Studies on Men & Masculinities. Sage, Thousand Oaks, pp. 165–177.

Pascoe, C.J., 2007. Dude, You're a Fag: Masculinity and Sexuality in High School. University of California Press, California.

Paul, E.L., 2006. Beer goggles, catching feelings, and the walk of shame: the myths and realities of the hookup experience. In: Kirkpatrick, Duck, Foley (Eds.), Relating Difficulty: The Processes of Constructing and Managing Difficult Interaction. Lawrence Erlbaum Associates, Mahwah, NJ.

Powell, A., 2008. Amor fati? Gender habitus and young people's negotiation of (hetero) sexual consent. Journal of Sociology 44 (2), 167–184.

Plummer, K., 2005. Foreward: permanence and change: sexual conduct thirty years on. In: Gagnon, J.H., Simon, W. (Eds.), Sexual Conduct: The Social Sources of Human Sexuality, second ed. Aldine Transaction, New Brunswick.

Plummer, K., 2007. Queers, bodies, and postmodern sexualities: a note on revisiting the "sexual" in symbolic interactionism. In: Kimmel, M. (Ed.), The Sexual Self: The Construction of Sexual Scripts. Vanderbilt University Press, Nashville, pp. 16–30.

Payne, Y.A., 2006. "A gangster and a gentleman": how street life-oriented, U.S.-born African men negotiate issues of survival in relation to their masculinity. Men and Masculinities 8 (3), 288–297.

Pyke, K., 1996. Class-based masculinities: the interdependence of gender, class, and interpersonal power. Gender and Society 10 (5), 527–549.

Roseneil, S., 2005. Living and loving beyond the boundaries of heteronorm: personal relationships in the 21st century. In: McKie, L., Cunningham-Burley, S., McKendrick, J. (Eds.), Families in Society: Boundaries and Relationships. Policy Press, Bristol, pp. 241–258.

Shek, Y.L., 2006. Asian American masculinity: a review of the literature. The Journal of Men's Studies 14, 379–391.

Simon, W., Gagnon, J.H., 1984. Sexual scripts. Society 22, 53–60.

Simon, W., Gagnon, J.H., 2003. Sexual scripts: origins, influences and changes. Qualitative Sociology 26 (4), 491–497.

Weis, L., 2006. Masculinity, whiteness, and the new economy: an exploration of privilege and loss. Men and Masculinities 8 (3), 262–272.

Mass Communication: Normative Frameworks

John D Peters, University of Iowa, Iowa City, IA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

From ancient times, mass communication has always raised normative questions. Most ancient views distrusted mass communication for misleading people from God or truth. Modern views, especially liberalism, endorse mass communication as a means of discovering truth and reaching consensus, but there are many dissents as well. The article examines the ancient and modern frameworks on the ethics and politics of the mass media. Mass communication, the public dissemination of symbols that are in principle addressed to everyone, has always aroused controversy, not only about specific policies and practices, but about larger philosophical questions concerning morality, politics, and art.

Distrust of Mass Communication

Despite its twentieth-century ring, mass communication is as old and fiercely debated as civilization itself. Mass communication will always occupy a symbolically charged spot in any social order, whether its medium of communication be pyramids or stained glass, newspapers or television, word of mouth or the Internet, and whether its aim be to keep subjects in awe, believers in the fold, or citizens informed. The lack of discrimination in audiences and the easy transgression of social boundaries are two key objections to it. Over the past century, such media of mass communication as dime novels, comics and radio, television, cinema, videogames, and the Internet have been blamed for about every ill imaginable, from the degradation of taste to the corruption of youth and the decline of democracy. Though such attacks are motivated by genuine worries arising from modern conditions, they have intellectual roots that reach into the Judeo–Christian and Greco–Roman past.

Normative frameworks with ancient roots are very much alive in both academic and popular debates about the social meaning of mass communication. They often act today more as moral intuitions about what a social order should expect from its public forms of communication than as systematically developed doctrines. Since such intuitions, at their worst, scapegoat mass media and thus invite simplistic causal attributions for complicated social problems, much social research on the effects of mass communication has explicitly contested older moralistic or normative ideas about mass communication. It would be a mistake, however, to assume that such ideas are of little interest to the social sciences. First, these larger frameworks provide the moral and political resources that drive the enterprise of social science as a quest for truth, humane social organization, and political progress. Second, many age-old topics are very much on the academic and public agenda today, such as worries about the role of the image in politics, the socialization and education of children, the possibility of global understanding amid radically incommensurable political and religious views, the ownership of information, or the place and purpose of the arts (see Table 1).

Iconoclasm

What a visually dominated society implies for politics, cognition, and education is a recent concern with a long past. We might start a genealogy of the distrust of the bewitching power

Table 1 Normative frameworks for mass communication

Doctrine	*A characteristic good*	*A characteristic evil*	*Representative theorist*
Iconoclasm	Representational purity	Idolatry	Moses
Platonism	Reality	Appearance	Plato
Christianity	Compassion	Callousness	Augustine
Stoicism	Cosmopolitanism	Anxiety	Marcus Aurelius
Elitism	Quality	Vulgarity	Arnold, Adorno
Republicanism	Public virtue	Corruption	Cicero, Machiavelli
Liberalism	Free inquiry	Censorship	Locke, Jefferson
Right-wing critics (nineteenth century)	Excellence, tradition	Leveling herd-mentality	Carlyle
Left-wing critics (nineteenth century)	Human fulfillment	Domination	Marx
Liberal critics (nineteenth century)	Expressive variety	Tyranny of majority	Mill, Tocqueville
New deal	Public interest	Irresponsibility	Hutchins commission
Social democracy	Public service	Commercialism	Habermas
Neo-liberalism	Free markets	Regulation	Hayek, Coase
Postmodernism	Free play	Hierarchy	Lyotard, Baudrillard
Cultural Marxism	Empowerment	Dominant ideology	Gramsci, Stuart Hall
Communitarianism	Participation	Apathy, disengagement	Dewey, Charles Taylor

 http://dx.doi.org/10.1016/B978-0-08-097086-8.95002-5

of images atop Mount Sinai. The second of the 10 Mosaic commandments proscribes the making of graven images and of mimetic visual arts in general. Ever since, iconoclasm (the smashing of images) has proven to be an enduring theme in the history of thinking about public life (Goux, 1978). Moses was specifically concerned with the proper representation of deity, but the intuition that images mislead hearts away from the 'true' vision or way of life has long had a more general purchase. Monotheism's gesture of casting away the false gods is an early ancestor of the modern scientific idea of sorting truth from error. In cultures influenced by the three ethical monotheisms of Judaism, Christianity, and Islam, notions of idolatry and simulation are still used to describe communications and their dangers. Depictions that are taken to be insulting of sacred things can still cause worldwide outrage and create calls for silence or censorship, such as the Muhammad cartoons controversy started by the Danish newspaper *Jyllands-Posten*. Such calls rest upon the controversial viewpoint that some things should not be represented. From a liberal point of view, such an idea is censorship. But the bitterness of the fights we continue to see and a culture of paparazzi photos and online pornography remind us that restraint in depiction can be a useful lesson. The principled refusal to look can stem not only from cowardice, but from tact and civility, and offers a rich vein for considering ethical dilemmas in media ethics and image ethics (Dayan, 2006).

Fabrications and the Corruption of Youth

Fear of the seduction of youth dates at least to Plato's *Republic*. Plato famously banished poets from his utopia, a gesture archetypal for normative thinking on communication since in at least two respects. First, Plato considered poets imitators who multiply appearances rather than realities in a world that already suffers from an overpopulation of copies. Just so, distrust of media veracity is widespread today, whether in the routine suspicion that some of the news and all of so-called 'reality TV' is fabricated or the far more sinister inkling that the Holocaust was a carefully orchestrated media conspiracy. Second, poets, claims Plato, have a bad influence on the young, even if their falsehoods can be pedagogically useful. Plato's ideas about socialization continue to resonate in debates about mass media, from comics and cinema to popular music and computers. The notion that a just society must patrol public representations appears in the long history of regulating depictions of sexuality and violence in film, and in various content ratings systems for film, television, and popular music. Though some modern interpreters of Plato, most notably Sir Karl Popper, read the *Republic* as the fountainhead of authoritarian thinking, few would deny that mass communication aimed at children deserves special care or that the portrayal of 'reality' is a thorny matter of public importance.

Responsibility to Distant Suffering

Both Stoicism and Christianity teach the idea of moral responsibility to the entire world. The Stoics speak of cosmopolitanism, or literally, world citizenship, and the Christians speak of a love that can encompass the entire human family. In the Christian notion of the Gospel, or literally, the good news, there is the ambition, in principle, of dissemination to a worldwide body of communicants. The hope for a great community on a global scale remains a potent political and moral project today, however much it may be stripped of its religious or philosophical content. Here the features of mass communication that usually evoke suspicion are valued. Its lack of discrimination in recipients and ability to cross social barriers make it potentially an inclusive and just form of communication.

Mass media are prime instigators of responses to distant suffering (Chouliaraki, 2006; Silverstone, 2007). Television, advertising, and documentary film all inherit the long colonial and humanitarian legacy of inspiring care for the wretched of the earth, seen at a distance. Some suggest that foreign policy is partly dictated by the location of television cameras. Victims whose fates can be made graphically visible to influential audiences may, in this argument, be more likely to receive political, humanitarian, or military aid than others who languish beyond the reach of the image. Others suggest that representations of human suffering do not deepen but rather deaden moral sensibilities, as news organizations, fighting for shrinking audiences, push the limits on shock and outrage by packaging suffering for popular titillation in ever more gruesome ways. The question is whether the aestheticization of suffering serves to inspire ethical action or to divert it. In any case, the ability of the media to alert the public to global crisis, to arouse humanitarian interest in those who suffer and to build cross-national bonds of solidarity – or hostility – is clearly one of the most important sites of debate in the politics and ethics of the media today. The Christian and Stoic heritage suggests that the representation of one's fellow humans is an ethical responsibility of the first order. Mass communication cannot act in a vacuum: nothing less in this view than global brotherhood and sisterhood is at stake.

The Degradation of Art

Finally, a tradition with a complex genealogy extending to early modern thinkers such as Montaigne and Pascal or the Roman satirists criticizes mass communication for its cheapness and low level of taste. Often frankly elitist in its condemnation of vulgar taste, commercial motivation, and formulaic structure, the high-art tradition wants communication to be an intense and concentrated aesthetic experience. It is perhaps false to say that elitism offers a normative framework for mass communication at all, since mass communication itself – the fact of widely diffused and openly addressed forms of culture – often seems the prime target. Matthew Arnold's dictum that genuine culture consists of 'the best that is known and thought in the world' or the Frankfurt School's fulminations against the 'culture industry' are two instances of the viewpoint that the arts can be morally and socially redemptive when appreciated in the proper, often rarefied, conditions, but catastrophic when they circulate in a bored and restless mass society. Here art's power prevails only when spread among the few.

An affirmative case for art in mass communication is found in the republican tradition. For Cicero, who stands at the tradition's head, rhetoric is not simply an art for inflating weak arguments and wooing crowds; it is an authentic technique of political deliberation and inspiration. Republicans, often

looking to classical Athens, also argue that theater can be a healthy part of political life, an opportunity for the collective consideration of the most important problems facing the community, not just a tool of deception or manipulation. To be sure, a republican such as Machiavelli has a bad reputation on the uses of rhetoric and theater, but he also understands communicative artistry as the very stuff of public life. Whereas many democratic theorists have distrusted the theater for breeding inauthenticity, fostering fantasies, and silencing the people, republican theorists insist that rhetoric and drama can have a legitimate place in a just polity (Hariman, 1995).

Though concern for aesthetic taste and mass culture was the very heart of the last century's debates about media, culture, and society from the 1920s through the 1980s, changes in education, society, technology, and taste have since weakened the widespread confidence that a canon of works exists whose quality is beyond dispute. High-brow taste has become increasingly omnivorous and elite art hybrid, jumbling heretofore segregated aesthetic levels. Even so, the sense that the products of commercial culture, produced to fill the many-channel environment of cable television and the almost endless bandwidth of the Internet, are often appalling and that mass communication could theoretically be a thing of beauty are not likely to disappear from the agenda (Gans, 1999). The critique of popular taste is not just snobbery about one's neighbors; it can also be a genuine insight about the way society is organized.

In sum, several of the most important intellectual sources in the western world – Hebrew, Greek, Roman, and Christian – have contributed a number of views about mass communication, all of which converge on the notion that the public dissemination of symbols is of such religious, social, moral, and aesthetic weight that it can never be left without some kind of control or restriction. Whether seen as dangerous or valuable, mass communication is held by all these traditions to be powerful indeed.

Faith in Openness: Liberalism and Its Critics

As opposed to the ancient nervousness about dangerous messages, most modern normative thinking about mass communication developed from or in critical dialog with the Anglo–American tradition, central to which is the proposition that public culture serves society best when left to its own devices, without meddling from crown or church. Liberalism, as the tradition has become known owing to its ideal of liberty, emerged from the writings of English Protestant thinkers in the seventeenth and eighteenth centuries, and has spread widely to become the official doctrine not only of communication policy, but of many governments throughout the modern world (where it is often honored in the breach.) Mass communication, liberalism argues, may threaten extant powers or opinions, but is never truly dangerous.

Distinctive in liberal thought is the special place given to freedom of communication, considered not simply one of many valuable liberties, but as fundamental to all others. The marketplace, whether of goods or ideas, is considered a refuge against the power of the state or church. The market's openness, equality, and lack of fakery are held to make it a model of the nature of symbolic exchange as well. Liberals conceive of liberty as the absence of constraint. They abhor censorship. Public toleration of ideas held to be repellent or dangerous is, for liberals, the price paid for liberty and an expression of the faith that truth, in the long run, will triumph. Open discussion, like the open market, is held to be self-correcting. The press, said Thomas Jefferson, 'is the best instrument for enlightening the mind of man' and liberal thinkers often give newspapers – and later mass media in general – a privileged mission in civil society, sometimes to the neglect of other institutions. (The press, as is often pointed out, is the only nongovernmental institution mentioned in the US Constitution.) Liberals dream of the press as a 'fourth estate,' an unofficial branch of government, an all-seeing monitor of society and politics generally, and the public's lifeline of political truth.

The liberal framework, in sum, values openness, criticism, diversity, and liberty in mass communication; enjoins an attitude of vigilance, tolerance, self-control, public curiosity, and rationality on the part of citizens; and believes in a knowable world and in the ultimate fruitfulness of open discussion for societal progress and discovering truth. It owes much to the Enlightenment confidence that scientific progress will banish all shadows and advance society. Because of their strong faith in reason and progress, liberals often dismiss as censorious and no longer relevant the older frameworks discussed earlier (Peters, 2005).

Nineteenth-Century Criticisms and Alternatives

The history of modern philosophical norms of mass communication is largely that of the development and criticism of liberalism. In the course of the nineteenth century, liberalism came under criticism from the right, the left, and from within. From the political right came the charge that liberalism's refusal to identify any positive moral good besides liberty was destructive of tradition and public spirit. Its love of freedom splintered civil society either into private egos pursuing their own interests or herds of bovine citizens haplessly searching for leadership. Such critiques urged that the communication structures shaping public life should not be left to shift for themselves but must be actively cultivated so that the 'best men' and ideas can prevail, and that discrimination can mark out zones of taste.

From within, liberal critics of liberalism such as Tocqueville and John Stuart Mill noted the potential of the liberal trust in public opinion as the ultimate arbiter of political and social life to breed conformity and 'soft' forms of despotism. Tocqueville called for deepened forms of community involvement in voluntary associations, and Mill for a more robust openness and courage in thought and discussion. Both saw the dangers of a tyranny of the majority and the unintended consequences of social atomization. Tocqueville and Mill did see the state and church as impediments to liberty, but their great contribution was to discover subtler forms of democratic degradation – in culture, thought, opinion, and everyday life.

From the left came an attack on the liberal conception of a self-regulating marketplace of goods and ideas. For Marx, most notably, to speak of liberty in the abstract was misguided because it neglected the historical and economic structures that limit the real experience of free expression. Thinkers in Marxist

and socialist traditions have been trenchant critics of the claim that unrestrained discussion will automatically yield truth rather than ideology. Quite like right-wing critics of liberalism, they argue that freedom must mean more than the absence of constraint. Freedom of the press is wonderful, the old quip goes, if you happen to own one. For leftist critics, the liberal equation of trade and discussion helps reinforce massive inequalities in the power to propagate ideas. Freedom of communication must, in this view, include meaningful chances for people to act and speak in public, a point made more recently by the ways that liberal ideas can exclude on the basis of gender, race, class, sexuality, ability, and religion. To achieve the ideal of authentically free communication, citizens need access to the means of communication. More than freedom from censorship, this ideal hinges on people playing meaningful roles in the making of history and in shaping their own political economic fates.

Twentieth-Century Developments

In the twentieth century, fascism emerged from the right-leaning critique of liberalism. European social democracy and the American New Deal developed as left-leaning correctives, with the state in each case assuming an affirmative role in subsidizing social services not delivered by the market, including new mass communication services. In the USA, where the role of the state was weaker than in Europe, broadcasting became a commercial enterprise with light federal regulation. Even so, American broadcasting in theory served the 'public interest, convenience, or necessity.' Though rarely enforced, the public ownership of the airwaves, including the right of listeners to reply and to have programming that met their needs, was found constitutional in several important Supreme Court cases in the USA in the postwar decades.

In Europe, broadcasting developed as an organ of the state, with a different history in each nation (*see* Public Broadcasting). Whereas media in the USA operated in a capitalist system with occasional governmental oversight, in Europe and elsewhere, government or quasi-government institutions were developed to provide information and entertainment to the entire nation. The British Broadcasting Corporation, long seen as the international flagship of public-service broadcasting, rested on such principles as universal service, quality programing, national integration, and rejection of creeping commercialism (often read as Americanism). Public-service broadcasting was one part of the welfare state's program of providing citizens with the common necessities of life (Curran, 2011).

Important theoretical treatises on mass communication in the twentieth century reflect a belief that mass communication should serve the public and not simply be free from constraint. In the USA, an important proposal for reform was *A Free and Responsible Press* (1947), better known as the Hutchins Commission. Its title already tells the story: freedom is not enough; the press must actively contribute to the commonweal and stand responsible for its effects. Highly critical of sensationalistic and shallow reporting, which it traced to the economic conditions that required media institutions to grab rather than educate audiences, the report did not, however, call for the state to intervene in the business of public information. Its answer was professional self-regulation within capitalism. Many journalists received the report with hostility, though the project of private ownership combined with public responsibility prevailed in the best postwar American journalism, both print and broadcast such as news produced by NBC (National Broadcasting Corporation) and CBS (Columbia Broadcasting Service) in their heyday. Here the normative vision combines the classic liberal doctrine of freedom from state interference with an affirmative concern for civic duty.

Perhaps the most important theoretical treatise on the normative grounds of mass communication in the second half of the twentieth century was the German philosopher Jürgen Habermas's *Structural Transformation of the Public Sphere* (1962). It articulates the European sense that mass communication is a social good too precious to be left to the market or the state. Notable about this book is its historical and social contextualization of the project of rational public communication. Habermas's notion of the public sphere is a rethinking of the liberal ideal of publicity (now often called 'transparency') for radical democratic purposes. Whether interpersonal or mass, any act of communication discloses, Habermas believes, the potential of persuasion without violence. Habermas's thinking about mass communication turns on the norm of uncoerced discussion in which the only power is that of the better argument. Rationality and participation are his requirements for just communication; he does not believe that the culture industries (mass media) will do the job. Though his practical alternative is not always clear, his aim is to deepen social democratic ideas and practices of mass communication. Habermas's critics have made many improvements in his model (Calhoun, 1992). But so has Habermas himself, most recently exploring the thorny problem of religion in a secular public sphere.

With the globalization of the world economy and the deregulation of the media in the USA, Europe, and elsewhere in the past decades, classic liberal ideas, 'neo-liberal' or 'market liberal' ideas, have come roaring back in the realm of mass communication. Talk of liberty and the free market has been persuasively used to justify regulating companies producing news and other content for outlets of mass communication. Neo-liberal arguments that have been made for the more traditional media businesses of news and entertainment have generally emphasized one side of liberal doctrine (the lack of state regulation of the market of goods and ideas) over the other (the creation of free, open, and diverse public fora). In arguments made for new or digital media businesses, the opposite is generally the case. A technolibertarian view is the reigning ideology for Internet companies such as Google (Turner, 2006) which argue for the absolute freedom of information (while also making billions from tracking online behavior and choices). Theorists of the cultural and legal implications of the digital domain grapple with a liberal legacy that has been put under duress by new technical and political conditions (Boyle, 2008).

Other Criticisms

A long tradition of social research on the production of news shows how the liberal dream can serve as a legitimating ideology for professional journalists and mask the social reality

of newswork. The sociology of news production shows that journalists follow formulaic routines and organizational cultures that shape their output more decisively than does a fearless quest for truth. The actual content and practices of journalism seem relatively unaffected by liberal ideas of the press. Exalted rhetoric about the free press coexists with a long history of sensationalism that dates to eighteenth-century broadsheets, if not earlier. The production of news is a crucial and fascinating tangle in contemporary politics and culture (Schudson, 2011).

Postmodernism offers another philosophical attack. The liberal model of communication, postmodernists argue, is simplistic, missing the instability of language and the deeply rooted impediments to communication as traditionally conceived. Further, in an age rife with visual rhetorics, censorship should not be the chief worry. Public consciousness can be manipulated not only through the denial of information but from too much information, or information selectively presented or 'spun.' For postmodernists, liberalism's fear of coercion from church or state leaves it unprepared for the simulations and dream worlds that stem from the market or the psyche itself. Not only suppression but seduction shapes the public agenda.

A similar point is made by cultural Marxists: instead of overt coercion, power largely works through the co-optation of consent or the creation of 'hegemony' (*see* Hegemony and Cultural Resistance). The liberal public sphere can serve as a forum of pleasurable domination – a view anticipated, from a different political position, by Tocqueville and Mill. In contrast to the liberal emphasis on dissemination, Marxist-inspired cultural studies often focus on reception: how audiences read messages against the grain of dominant ideologies (Chen and Morley, 1996). Here, the liberal emphasis on active citizens is mirrored, but with a richer, more politicized account of the receiving end of mass communication. As in classical Marxism, neo-Marxists insist on looking at the results, not just the liberties, of communication.

A similar critical approach is found among theorists of race, gender, and sexuality, who show the hidden injuries of the liberal dream of ignoring bodily difference and call, instead, for a more plural politics based on a recognition of the plurality and fluidity of identity and the key role played not only by the dominant public but by 'counterpublics' as well (Warner, 2005).

Communitarians, descendents of nineteenth-century rightist and centrist critics of liberal egotism, have a revisionist program for mass communication. They often assail television as we know it for its alleged stupefactions and sapping of social capital, and in turn call for new forms of news gathering and discussion to invigorate community ties and public involvement. For communitarians, a chief ill of contemporary life is the alienation of ordinary people from public communication or the self-enclosure of citizens inside narcissistic bubbles generated by ever more selective forms of personal news filtering. (In contrast to many who herald the Internet as a beacon of liberty, some libertarians view it as a means to add the process of social fracturing into ever more minute taste publics.) It becomes the task of the press, in this view, among other agencies of mass communication, to draw citizens into the common good and to bring them together into fora in which they bump elbows with strangers. The 'public journalism' movement, which reflects communitarian ideas, calls for the participation of all citizens in open and public dialog. Its normative vision of mass communication is not the liberal free dispersion of ideas but a civic conversation of all with all about the important issues of the day (Glasser, 1999).

Finally, some fault liberalism for its overly rational emphasis. The American philosopher John Dewey, sometimes a communitarian hero, argued that one fault of the liberal conception of mass communication was its dryness. For him, drama in its diverse forms is a key resource in modern society. He believed in an aesthetic education of humans and citizens. Similarly, recent advocates of public service broadcasting endorse entertainment as more than frivolity; it can play a role in widening sympathies for fellow citizens, allowing a sort of communion among strangers. These theorists take artistry in mass communication to be a means of expanded social imagination, rather like the republican tradition. They contest the liberal commitment to the left brain, as it were, of mass communication: the privilege of news in public life. Fiction, they argue, can have as important a civic role as fact (*see* Entertainment).

Whither Media Ethics?

Media ethics has largely developed as a field concerned with the practices of professionals, especially the analysis of hard cases that face journalists (deception, disclosure of sources, the blurry borders of objectivity, etc.). The sheer range and diversity of frameworks reflecting on mass communication and its larger social, political, moral, spiritual, pedagogical, economic, cultural, and aesthetic implications suggest a richer agenda for media ethics. Questions about the tenor of public life, the experience of childhood, the relief of suffering, the possibilities of artistic expression, or the texture of political life in an age of user-generated content, for instance, might take a rightful place in debates about mass communication next to questions about liberty, information, and ideology. Though various normative frameworks may be mutually incompatible in their principles, emphases, or conclusions, each one has something to offer.

See also: Broadcasting: Regulation; Freedom of the Press; Mass Media: Introduction and Schools of Thought; Media Imperialism; News: General; Norms; Violence and Media.

Bibliography

Boyle, J., 2008. The Public Domain. Yale University Press, New Haven, CT.
Calhoun, C. (Ed.), 1992. Habermas and the Public Sphere. MIT Press, Cambridge, MA.
Chen, K., Morley, D. (Eds.), 1996. Stuart Hall: Critical Dialogues in Cultural Studies. Routledge, London.
Chouliaraki, L., 2006. The Spectatorship of Suffering. Sage, London.
Curran, J., 2011. Media and Democracy. Routledge, New York.
Dayan, D. (Ed.), 2006. La terreur spectacle. Boeck, Paris.
Gans, H., 1999. High Culture and Popular Culture. Basic, New York.
Glasser, T.L. (Ed.), 1999. The Idea of Public Journalism. Guilford Press, New York.
Goux, J.-J., 1978. Les Iconoclastes. Seuil, Paris.

Habermas, J., 1962. Strukturwandel der Öffentlichkeit. Luchterhand, Neuwied (1989 Structural Transformation of the Public Sphere, Trans.). MIT Press, Cambridge, MA.
Hariman, R.D., 1995. Political Style: The Artistry of Power. University of Chicago Press, Chicago, IL.
Hutchins, R., 1947. A Free and Responsible Press. University of Chicago Press, Chicago, IL.
Keane, J., 2009. The Life and Death of Democracy. Norton, New York.
Peters, J.D., 2005. Courting the Abyss: Free Speech and the Liberal Tradition. University of Chicago Press, Chicago, IL.
Schudson, M., 2011. The Sociology of News. Norton, New York.
Silverstone, R., 2007. Media and Morality. Polity, Cambridge.
Turner, F., 2006. From Counterculture to Cyberculture. University of Chicago Press, Chicago, IL.
Warner, M., 2005. Publics and Counterpublics. Zone, New York.

Mass Imprisonment and Its Consequences

Holly Foster, Texas A&M University, College Station, TX, USA
John Hagan, Northwestern University, Evanston, IL, USA; and American Bar Foundation, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A social exclusion framework is used to provide a synopsis of literature around the collateral consequences of mass imprisonment for individuals and adult family members. Connections between imprisonment and life course consequences of economic attainments, homelessness, mental health, social disability functioning, subjective status, and political involvement are reviewed. A recent work on mediating mechanisms is identified. Effects on spouses and caregivers are further reviewed in terms of maternal well-being and financial supports. Suggestions for future research on other aspects of spillover effects and mediating mechanisms are made, along with the need for research on the collateral consequences of women's imprisonment.

Mass imprisonment refers to the social and historical increase and concentration of imprisonment in the United States since the 1970s primarily among disadvantaged minority males (Garland, 2001; Pettit and western, 2004; Western, 2006). The most recent statistics show some decreases in state prison populations over the past 3 years but increases in federal prisons (Carson and Golinelli, 2013). However, the major overall trend is of an overwhelming increase in incarceration rates, where the prison population has more than tripled between 1997 and 2007 (Garland, 2001; Pew Center on the States, 2008). Men comprise 93% of prison population (West et al., 2010). Rates indicate that females are one of the fastest growing segments of incarcerated groups (Kruttschnitt, 2010). At present, male imprisonment rates exceed those of females by 14 times (Guerino et al., 2011). In addition to gender disparities, incarceration is disproportionately concentrated by race and class, with young black males with low levels of education having the highest rates reflecting social inequality in society (Pettit and Western, 2004; Western, 2006). Among men, black non-Hispanic males have an imprisonment rate that is seven times higher than for white non-Hispanic males (Guerino et al., 2011). Pronounced racial disparities are also found among females where black non-Hispanic females have an imprisonment rate three times that of white non-Hispanic females (Guerino et al., 2011). Furthermore, in the past decade there has been a growing Hispanic population in prison, where one in every six Hispanic males and one in every 45 Hispanic females born today can expect to go to prison in his or her lifetime. These are double the estimated rates for non-Hispanic whites (Mauer and King, 2007).

Social exclusion is a consistent theme in the work on collateral consequences of incarceration in the life course. Social exclusion implies being 'shut out' from conventional society (Micklewright, 2002). This may include disengagement from key social institutions (Walker and Walker, 1997) as well as broader societal disconnection fostered by health limitations (Aber et al., 2002). Much of the collateral consequences research on mass imprisonment identifies exclusion among former prisoners in roles as workers, citizens, and spouses (Pettit and Western, 2004; Western, 2006; Manza and Uggen, 2006). Furthermore, the majority of adult male and female prisoners are parents (Glaze and Maruschack, 2008). Another consequence of mass incarceration is the social exclusion or disconnection of those affected by incarceration, in the form of spillover effects (Comfort, 2007), including prisoners' families, from major societal institutions (Foster and Hagan, 2007; Murray, 2007). Pettit (2012) identifies a further form of social exclusion of incarcerated groups, and especially disadvantaged African-American males, from many data collection efforts, rendering this group largely invisible to contemporary social science and social policy efforts. We focus on forms of social exclusion and related social and health consequences for adults of mass incarceration in this article.

The literature on the consequences of mass imprisonment most often focuses on adult men given this societal concentration by gender; however, women's incarceration also has various important consequences (Kruttschnitt, 2010; Hagan and Foster, 2012). We therefore attend to research on the incarceration of both men and women, to the extent the research literature allows.

In this article, we review recent evidence pertaining to individual outcomes and the influences of one's own incarceration and its effects on adult family members. Our article begins to identify trends and areas of future research from this developing literature in terms of mediators and moderators of incarceration influences on various life domains in one's own life and of other adults. A critical area of further research concerns parental incarceration influences on children, including their social exclusion as well (Foster and Hagan, 2007; Murray, 2007). Thus we also refer the reader to key overviews that will hopefully stimulate and prioritize further research (Hagan and Dinovitzer, 1999; Comfort, 2007; Murray and Farrington, 2008; Murray et al., 2012; Wakefield and Uggen, 2010; Wildeman and Muller, 2012; Wildeman, 2010).

Individual Consequences

This section reviews the effects of incarceration on a host of life course outcomes for individuals. These include economic attainments, homelessness, mental health, social disability, subjective status, physical health, and political involvement.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45068-3

We highlight some of the most recent work that identifies factors that mediate the association between incarceration and various outcomes. These mediating factors are of particular policy significance, as they suggest the mechanisms through which the problem effects of incarceration are transmitted, and therefore may be socially modified.

Holzer (2009) provides a comprehensive review of studies on the effects of incarceration in the economic realm on employment and earnings. He attends to effects on employers (the demand side) and effects on workers (the supply side). Evidence is fairly consistent on the demand side, showing that incarceration decreases employer demand for those with criminal records (e.g., Pager, 2007). On the supply side, often using data from large-scale survey studies, there is some mixed evidence, but overall this research shows negative effects of incarceration on the earnings and employment of individuals who have been incarcerated compared to those who have not (e.g., Western, 2006). Consequences include lower wage levels as well as lower wage growth. They also include fewer weeks of employment per year and annual earnings (Western, 2006).

Western (2006) further details some of the macroeconomic consequences in American society of mass imprisonment. Central among these are persistent white–black wage inequalities that are found over time when calculations take into account the hidden penal population in the population. When calculations omit the penal population, wage inequalities are seen as decreasing between whites and blacks. When the penal population is included, heightened disparities in men's economic well-being are evident by race and ethnicity. Pettit (2012) further highlights the consequences of the social exclusion of penal populations from official statistics.

Also, recent research on former inmates expands information on racial differences in the labor market consequences of mass imprisonment over the life course (Lyons and Pettit, 2011). African-Americans are disproportionately exposed to imprisonment. Furthermore, African-Americans are more adversely affected by imprisonment spells in terms of hourly wage growth trajectories. While wage trajectories increase for young men after imprisonment, they grow at a slower pace for African-Americans than whites. The evidence thus supports wage divergence by race following imprisonment and compounded disadvantage for African-Americans (Lyons and Pettit, 2011). Future work with women can expand on these findings.

Further investigating linkages between incarceration and social exclusion in the labor market, Pager (2003, 2007) examines whether a criminal record operates as a negative credential that decreases the likelihood of legitimate paid employment. Using a field experiment with an in-person audit methodology in Milwaukee, Wisconsin, Pager sent pairs of testers to real-world entry-level job openings. She used pairs of African-Americans and white testers, all of whom were college students, who were similar within pairings with the exception of a criminal record that was assigned to them. This study showed that for African-American and white subjects, those with a criminal record received fewer callbacks. Furthermore, African-American subjects *without* a criminal record were similarly likely to receive a callback from employers as white subjects *with* a criminal record. African-American males with a criminal record received even fewer callbacks than white males with a criminal record. These results point both to the exclusion of those with a criminal record from paid employment, but also the broader and enduring significance of race in employment discrimination in hiring decisions.

Subsequent research furthers this line of work with African-American, Hispanic, and non-Hispanic white males in New York City (Pager et al., 2009). One component of the experimental audit study in NYC with male subjects applying for entry-level employment positions revealed that a white male *with* a criminal record was equally likely to be hired as a Latino or African-American male *without* a criminal record – similar to the results of Pager (2007) but now further inclusive of Latinos. These results advance our understanding of stigma and racial discrimination in hiring decisions.

The majority of research of incarceration effects in the economic realm uses male samples. However, one study with incarcerated females by Cho and Lalonde (2008) finds little effect on women's employment outcomes. As Cox (2012) observes, these analyses may benefit from taking into account the type of release women have following incarceration, as inmates discharged with community supervision have to maintain employment as part of their release. This may have led to the unexpectedly negligible effects observed. As study cited in Cox (2012) shows in a correspondence test study that previously incarcerated women were 31% less likely to receive a positive response from employers than women who were not incarcerated. Another cite to a study in Cox (2012) building on the work of Pager with men (2003, 2009), reports that African-American women were least likely to receive a positive response from employers. More work is needed on economic consequences to examine female incarceration in a range of sample conditions.

Adding to this literature on economic exclusion is recent work on associations between incarceration and housing insecurity. Using Fragile Families data on a large sample of urban men, a positive association is found between incarceration and housing insecurity (two times higher among the ever-incarcerated than nonincarcerated) as well as homelessness (four times as high among ever-incarcerated than nonincarcerated) (Geller and Curtis, 2011). Furthermore, postincarceration earnings constitute a mechanism through which incarceration is linked to housing insecurity and homelessness. Higher levels of earnings reduce housing insecurity. This analysis suggests a mechanism by which housing problems can be reduced among those with prison experiences in the form of postincarceration earnings.

In the realm of social relationships, research with incarcerated compared to nonincarcerated men has investigated marital relationships and domestic violence. Incarcerated men are less likely than their nonincarcerated counterparts to be married, but are just as likely to be fathers (Western, 2006). The incapacitation effect of imprisonment is also evident in the higher likelihood of divorce for those who have been incarcerated compared to those who have not been (Western, 2006). Marriage has been identified in life course research as a protective factor against criminal activity (Sampson and Laub, 1993[1995]; Laub et al., 1998; Laub and Sampson, 2003). Marriage also reduces the probability of domestic violence among both drug and violent offenders (Western, 2006). Further analyses may reveal whether it is high-quality relationships that decrease the chances of domestic violence (Western, 2006).

Further work examines the robust association between incarceration and marital dissolution among men (Lopoo and Western, 2005; Western, 2006). Massoglia et al. (2011) examine stigma and duration of separation as explanations for the relationship between incarceration and marital dissolution using NLSY79 data on adult males. They find the initial association between ever being incarcerated and marital dissolution is explained by the duration of incarceration. Each year of incarceration increases the odds of divorce or separation by 32%. In comparative analyses with a nonstigmatized group, they also find that the duration of deployment affects the likelihood of divorce among male military personnel. These results combine to support separation arguments. The authors suggest that added investigation of how these mechanisms work in lives of children could reveal the mechanism for intergenerational influences warranting further exploration.

Research on incarceration and social exclusion spans the social institutions of connections to the labor market and families. This literature further encompasses individual health, which may constitute a pathway to societal exclusion. A number of studies link incarceration to mental health problems. Turney et al. (2012) found an association between recent and current incarceration and depression, net of prior incarceration using longitudinal data on men from the Fragile Families and Child Wellbeing Study. Using stress process theory, this article considers incarceration as a primary stressor and socioeconomic status and family functioning as secondary stressors. The latter are considered as mediators of the incarceration and depression relationship among men. Results support a mediational model where current and recent incarceration influences on depression are reduced by socioeconomic factors, in particular, changes in employment and homelessness, and family functioning, in terms of changes in relationship quality with the child's mother. These results suggest pathways through which incarceration affects depression among men.

Among former prisoners, Turney et al. (2013a) examine the role of stigma and discrimination in the incarceration and mental health relationship. They find criminal record discrimination and race and ethnic discrimination both lead to psychological distress as experienced in a sample of recently released formerly incarcerated men in California. These results suggest that in this previously incarcerated sample, discrimination is a pathway through which incarceration is associated with more psychological distress.

Schnittker et al. (2012) further examine the link between incarceration and psychiatric disorders in a national probability sample survey of adults. They find a particularly robust association between incarceration and mood disorders, net of multiple controls. They further investigate whether psychiatric disorders mediate the relationship between incarceration and disability. They find that psychiatric conditions, and particularly mood disorders, mediate this association. This work suggests the potential value of attention by policy analysts to mood disorders in reintegration efforts as well as research on social exclusion.

Several studies support a role of stigmatizing processes as another pathway from incarceration to social exclusion. These studies include social psychological and physical health outcomes. With national data on adult males, Schnittker and Bacak (2013) find that incarceration decreases subjective social status, or one's perception of one's standing in the community, net of a range of controls. This finding is consistent with experiencing incarceration as a stigmatizing process. Furthermore, Wildeman and Muller (2012) comprehensively review linkages between incarceration and health. In particular, incarceration is associated with worse physical health after release, including both mortality and morbidity (e.g., infectious diseases).

Several studies on the incarceration–health relationship highlight the role of stigma and stress-related diseases among other mediators (Wildeman and Muller, 2012). Schnittker and John (2007) find with a large community sample of males and females, across modeling strategies, that those with a longer history of incarceration have more severe health limitations. Some mediation of incarceration history effects is provided by schooling, income, and marital status but not drug use. Massoglia (2008) reports in a broad community survey that incarceration is positively associated with medically diagnosed illnesses strongly related to stress (e.g., hypertension, heart problems, and chronic lung disease) but not illnesses unrelated to stress (arthritis, cancer, and diabetes). He further finds positive associations between incarceration and stress-related self-reported health problems (e.g., chest pain, heart problems, general health problems) as well as infectious diseases (e.g., urinary tract infections and hepatitis or TB). Finally, a recent study addresses spillover effects on health with survey data from NYC, where variability in neighborhood-level adult incarceration rates range from 2.1 to 12.8% (Frank et al., 2013). This research finds that neighborhood-level incarceration rates are associated with individual asthma prevalence, net of individual incarceration history (Frank et al., 2013). However, this macroeffect is explained when individual sociodemographic covariates are added to the model. The positive individual effect of incarceration history on asthma held net of covariates. Incarceration is therefore linked to asthma prevalence among adults.

Finally, an important realm of social exclusion involves political disconnection. One of the definitive studies in this area by Manza and Uggen (2006) finds that among a sample of young adults, being arrested and being in prison or jail, there was a decrease in the likelihood of voting compared to those who have never been arrested. Furthermore, those who vote are less likely to report a range of offenses. Voting privileges may therefore constitute a pathway to social inclusion and civic reintegration that puts former felons back into contact with a major societal institution and fosters desistance from criminal activity.

Together, this work with adults on economics, housing insecurity, relationships, mental health, physical health problems, and political participation suggests persistent associations between incarceration experiences and various measures of adult social exclusion. Given the pronounced gender disparities in imprisonment, a lot of the work on incarceration and social exclusion among adults is conducted with male samples. Some research uses community samples with males and females and other studies examine relationships between incarceration and a range of outcomes with females. However, fewer studies have been conducted with community samples of women. Much of the recent mediational work identifying socially modifiable

mechanisms pertains to work with adult males. More work on mediators is needed with incarcerated women. Future work needs to identify if mediators are generic or gendered in nature.

Family Consequences

This section builds on the above body of work on the social exclusion of the formerly or currently incarcerated to consider social exclusion among their families. We focus here on collateral consequences of mass imprisonment of spouses, families, and caregivers. A number of comprehensive reviews cover recent research on parental incarceration influences on children (Hagan and Dinovitzer, 1999; Comfort, 2007; Murray and Farrington, 2008; Murray et al., 2012; Wakefield and Uggen, 2010; Wildeman and Muller, 2012). Research further connects paternal incarceration to children's social exclusion (Foster and Hagan, 2007), an area that needs additional investigation with attention to maternal influences and life course stages.

Comfort (2007) details the myriad spillover effects extending from a loved one's incarceration on members of his family, including spouses and significant others. Comfort's empirical research (2008) details a process of 'secondary prisonization' encountered by the women attached to men serving time in prison. The women encounter the rules and regimentation of the prison, even in their own homes waiting for calls from inmates and in their letter writing process, as well as in visitation rituals. Although these women have committed no crimes, they intimately feel the reach of the prison in their everyday lives through their connections to men who are incarcerated. Foster and Hagan (2014) detail the support role that women (e.g., mothers, aunts, sisters) also play in the lives of incarcerated women, particularly as this involves child caregiving.

Several recent studies find effects of male incarceration on women's social exclusion from mainstream society. These include effects of male partner imprisonment and mother's well-being and social support. Wildeman et al. (2012) find that recent and distal paternal incarceration is positively associated with maternal depression and maternal life dissatisfaction. The association of recent incarceration with both paternal and maternal outcomes holds net of a multitude of controls, including maternal and paternal characteristics and a lagged dependent variable. Changes in material hardship and changes in mother's parenting stress explain the association between recent paternal incarceration and maternal depression. Changes in relationships and parenting stress mediate associations between recent paternal incarceration and maternal life dissatisfaction. These mediating pathways illuminate how male partner incarceration affects maternal outcomes.

The influence of male incarceration on women's physical health is also seen at the macrolevel in terms of African-American and white disparity in AIDS infection rates. Johnson and Raphael (2009) show that higher incarceration rates among African-American males explain the racial disparity in AIDS infection rates among females. Lee and Wildeman (2013) further theorize the health effects of male imprisonment on female cardiovascular disease. They identify three mechanisms for future empirical research on this anticipated connection: socioeconomic status, family functioning, and stress and mental health problems.

Turney et al. (2012) find that recent, but not current, paternal incarceration reduces the amount of perceived in-kind (e.g., emergency childcare, housing assistance) and small and large financial support available to mothers in the Fragile Families and Child Wellbeing data set, net of numerous controls. Paternal incarceration therefore contributes to social support disparities among families. Thus recent paternal incarceration undermines the in-kind and financial support of mothers, contributing to the isolation of women from social support resources. These findings suggest that better societal supports are needed in the lives of women experiencing male partner incarceration.

In one of the few studies to look at the effects of parental incarceration on caregivers, Turanovic et al. (2012) use a rich and unique data source to examine the influence of both maternal and paternal incarceration on the caregivers of children of incarcerated parents in Arizona ($n = 100$) with information from a broader study they are conducting. Qualitative responses of caregivers are thematically coded and indicate that 58% of respondents experienced a negative effect of parental incarceration while 22% experienced no change and 20% experienced positive changes. Caregivers were more likely to report negative effects concerning paternal than maternal imprisonment. Negative effects included severe financial strains, childcare strains, and emotional effects. Those reporting positive effects referred to parents who had inconsistently been involved in children's lives prior to imprisonment. Those reporting no changes were referring to inconsistently involved parents in the lives of children before imprisonment. Thus the majority of effects experienced by caregivers were in the form of additional financial, childcare, and emotional strains. These kinds of strains were apparent also in the lives of women partners of incarcerated men in the other studies reviewed on familial effects.

Several studies show effects of paternal incarceration on strained family material conditions, similar to what Turanovic et al. (2012) find among caregivers of children of incarcerated parents. Social exclusion further includes economic aspects of disadvantage and deprivation, with significant intergenerational implications (Kahn and Kamerman, 2002). Using the Fragile Families and Child Wellbeing longitudinal data resource, Sugie (2012) finds that recent paternal incarceration but not earlier incarceration is associated with familial receipt of food stamps and the Medicaid/State Children's Health Insurance Program (SCHIP). However, neither prior nor recent paternal incarceration was associated with Temporary Aid to Needy Families (TANF) assistance. Sugie (2012) notes that food stamps and Medicaid/SCHIP are more inclusive support programs for families than TANF. It may be that families experiencing paternal incarceration are prevented from accessing the full array of government programs for material assistance.

Further research highlights the financial strains experienced by families with incarcerated fathers (Schwartz-Soicher et al., 2011). This research examines material hardship or the extent to which families face difficulties in meeting their basic needs (e.g., could not pay rent or mortgage, evicted from home due to nonpayment of rent or mortgage). Across a variety of modeling strategies, these results show that paternal incarceration increases family material hardship. One of the pathways identified suggests that diminished economic resources result from paternal incarceration leading to family hardship. Geller et al.

(2011) further show that paternal incarceration reduces father's material contributions to families, and one route by which this effect is transmitted is through relationship structure. Cohabitating fathers tend to contribute more to children's financial support.

This article has provided a comprehensive examination of the collateral consequences of incarceration for individuals. These are important and attest to a wide range of effects including multiple dimensions of social exclusion. Our contribution then detailed the consequences of incarceration for other family members, a domain that begins to capture 'spillover effects.' Further work on both topics is needed, including work on mediators and moderators of incarceration effects for individuals as well as additional work on spillover effects. In particular, important future research will involve sustained attention to the collateral consequences of mass incarceration in terms of spillover effects on neighborhoods and communities (Clear, 2007; Sampson and Loeffler, 2010), schools (Hagan and Foster, 2012), and children (Murray and Farrington, 2008; Hagan and Dinovitzer, 1999; Wildeman and Muller, 2012).

Acknowledgment

We would like to thank the National Science Foundation, Law and Social Sciences Program for support for this work through award #: SES-1228345.

Bibliography

Aber, Lawrence J., Gershoff, Elizabeth, Brooks-Gunn, Jeanne, 2002. Social exclusion of children in the United States: identifying potential indicators. In: Kahn, Alfred, Kamerman, Sheila (Eds.), Beyond Child Poverty: The Social Exclusion of Children. The Institute for Child and Family Policy at Columbia University, New York, pp. 247–286.

Carson, E. Ann, Golinelli, Daniela, 2013. Prisoners in 2012 – Advance Counts. U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics. #NCJ 242467.

Charles, Kerwin K., Luoh, Ming Ching, 2010. Male incarceration, the marriage market, and female outcomes. The Review of Economics and Statistics 92, 614–627.

Cho, Rosa M., Lalonde, Robert J., 2008. The impact of incarceration in a state prison on the employment prospects of women. Journal of Quantitative Criminology 24, 243–265.

Clear, Todd R., 2007. Imprisoning Communities: How Mass Incarceration Makes Disadvantaged Neighborhoods Worse. Oxford University Press, New York.

Cox, Robynn J.A., 2012. The impact of mass imprisonment on the lives of African American women. Review of Black Political Economy 39, 203–212.

Comfort, Megan, 2007. Punishment beyond the legal offender. Annual Review of Law and Social Science 3, 271–296.

Frank, Joseph W., Hong, Clemens S., Subramanian, S.V., Wang, Emily A., 2013. Neighborhood incarceration rate and asthma prevalence in New York city: a multilevel approach. American Journal of Public Health 103, e38–e44.

Foster, Holly, Hagan, John, 2007. Incarceration and intergenerational social exclusion. Social Problems 54, 399–433.

Foster, Holly, Hagan, John, 2014. The costs of incarcerating mothers and non-mothers: the gendered distribution of family care and human rights. Journal of Gender, Race and Justice 17, 257–278.

Garland, David, 2001. Introduction: the meaning of mass imprisonment. In: Garland, David (Ed.), Mass Imprisonment: Social Causes and Consequences. Sage, Thousand Oaks, CA, pp. 1–3.

Glaze, Lauren E., Maruschak, Laura M., 2008. Parents in Prison and Their Minor Children. U.S. Department of Justice, Washington, DC. #NCJ 222984.

Geller, Amanda, Curtis, Marah, 2011. A sort of homecoming: incarceration and the housing security of urban men. Social Science Research 40, 1196–1213.

Geller, Amanda, Garfinkel, Irwin, Western, Bruce, 2011. Paternal incarceration and support for children in fragile families. Demography 48, 25–47.

Guerino, Paul, Harrison, Paige M., Sabol, William J., 2011. Prisoners in 2010. U.S. Department of Justice, Office of Justice Programs. Bureau of Justice Statistics, Washington, DC. #NCJ 236096.

Hagan, John, Dinovitzer, Ronit, 1999. Collateral consequences of imprisonment for children, communities and prisoners. Crime & Justice 26, 121–162.

Hagan, John, Foster, Holly, 2012. Children of the American prison generation: the paradoxical "spillover" school effects of incarcerating mothers. Law & Society Review 46, 37–69.

Holzer, Harry J., 2009. Collateral costs: effects of incarceration on employment and earnings among young workers. In: Raphael, Steven, Stoll, Michael A. (Eds.), Do Prisons Make Us Safer? The Benefits and Costs of the Prison Boom. Russell Sage Foundation, New York, pp. 239–265.

Johnson, Rucker C., Raphael, Steven, 2009. The effects of male incarceration dynamics on acquired immune deficiency syndrome infection rates among African American women and men. Journal of Law and Economics 52, 251–293.

Kahn, Alfred J., Kamerman, Sheila B., 2002. Social exclusion: a better way to think about childhood deprivation? In: Kahn, Alfred, Kamerman, Sheila (Eds.), Beyond Child Poverty: The Social Exclusion of Children. The Institute for Child and Family Policy at Columbia University, New York, pp. 13–36.

Kruttschnitt, Candace, 2010. The paradox of women's imprisonment. Daedalus Summer, 32–42.

Laub, John H., Sampson, Robert J., 2003. Shared Beginnings, Divergent Lives: Delinquent Boys to Age 70. Harvard University Press, Cambridge, MA.

Laub, John H., Nagin, Daniel S., Sampson, Robert J., 1998. Trajectories of change in criminal offending: good marriages and the desistance process. American Sociological Review 63, 225–238.

Lee, Hedwig, Wildeman, Christopher, 2013. Things fall apart: health consequences of mass imprisonment for African American women. Review of Black Political Economy 40, 39–52.

Lopoo, Leonard M., Western, Bruce, 2005. Incarceration and the formation and stability of marital unions. Journal of Marriage and Family 67 (3), 721–734.

Lyons, Christopher J., Pettit, Becky, 2011. Compounded disadvantage: race, incarceration, and wage growth. Social Problems 58, 257–280.

Manza, Jeff, Uggen, Christopher, 2006. Locked Out: Felon Disenfranchisement and American Democracy. Oxford University Press, New York.

Massoglia, Michael, 2008. Incarceration as exposure: the prison, infectious disease, and other stress-related illnesses. Journal of Health and Social Behavior 49, 56–71.

Massoglia, Michael, Remster, Brianna, King, Ryan D., 2011. Stigma or separation? Understanding the incarceration-divorce relationship. Social Forces 90, 133–155.

Mauer, Marc, King, Ryan S., 2007. Uneven Justice: State Rates of Incarceration by Race and Ethnicity. The Sentencing Project, Washington, DC.

Micklewright, John, 2002. Social exclusion and children: a European view for a U.S. debate. In: Kahn, Alfred, Kamerman, Sheila (Eds.), Beyond Child Poverty: The Social Exclusion of Children. The Institute for Child and Family Policy at Columbia University, New York, pp. 89–129.

Murray, Joseph, 2007. The cycle of punishment: social exclusion of prisoners and their children. Criminology and Criminal Justice 7, 55–81.

Murray, Joseph, Farrington, David P., 2008. The effects of parental imprisonment on children. Crime & Justice 37, 133–206.

Murray, Joseph, Farrington, David P., Sekol, Ivana, 2012. Children's antisocial behavior, mental health, drug use, and educational performance after parental incarceration: a systematic review and meta-analysis. Psychological Bulletin 138, 175–210.

Pager, Devah, 2003. The mark of a criminal record. American Journal of Sociology 108, 937–975.

Pager, Devah, 2007. Marked: Race, Crime, and Finding Work in an Era of Mass Incarceration. University of Chicago Press, Chicago.

Pager, Devah, Western, Bruce, Bonikowski, Bart, 2009. Discrimination in a low-wage labor market: a field experiment. American Sociological Review 74, 777–799.

Pettit, Becky, Western, Bruce, 2004. Mass imprisonment and the life course. American Sociological Review 69, 151–169.

Pettit, Becky, 2012. Invisible Men: Mass Incarceration and the Myth of Black Progress. Russell Sage Foundation, New York.

Pew Center on the States, 2008. One in 100: Behind Bars in America 2008.

Sampson, Robert J., Loeffler, Charles, 2010. Punishment's place: the local concentration of mass incarceration. Daedalus Summer, 20–31.

Sampson, Robert J., Laub, John H., 1993[1995]. Crime in the Making: Pathways and Turning Points through Life. Harvard University Press, Cambridge, MA.

Schwartz-Soicher, Ofira, Geller, Amanda, Garfinkel, Irwin, 2011. The effect of paternal incarceration on material hardship. Social Services Review 85, 447–473.

Schnittker, Jason, Massoglia, Michael, Uggen, Christopher, 2012. Out and down: incarceration and psychiatric disorders. Journal of Health and Social Behavior 53, 448–464.

Schnittker, Jason, Bacak, Valerio, 2013. A mark of disgrace or a badge of honor? Subjective status among former inmates. Social Problems 60, 234–254.

Schnittker, Jason, John, Andrea, 2007. Enduring stigma: the long-term effects of incarceration on health. Journal of Health and Social Behavior 48, 115–130.

Sugie, Naomi F., 2012. Punishment and welfare: paternal incarceration and families' receipt of public assistance. Social Forces 90, 1403–1427.

Turanovic, Jillian J., Rodriguez, Nancy, Pratt, Travis C., 2012. The collateral consequences of incarceration revisited: a qualitative analysis of the effects on caregivers of children of incarcerated parents. Criminology 50, 913–959.

Turney, Kristin, Lee, Hedwig, Comfort, Megan, 2013a. Discrimination and psychological distress among recently released male prisoners. American Journal of Men's Health. http://dx.doi.org/10.1177/15579883/3484056.

Turney, Kristin, Schnittker, Jason, Wildeman, Christopher, 2012. Those they leave behind: paternal incarceration and maternal instrumental support. Journal of Marriage and Family 74, 1149–1165.

Turney, Kristin, Wildeman, Christopher, Schnittker, Jason, 2013b. As fathers and felons: explaining the effects of current and recent incarceration on major depression. Journal of Health and Social Behavior 53, 465–481.

Walker, Alan, Walker, Carol, 1997. Britain Divided: The Growth of Social Exclusion in the 1980s and 1990s. Child Poverty Action Group, London, UK.

Wakefield, Sara, Uggen, Christopher, 2010. Incarceration and stratification. Annual Review of Sociology 36, 387–406.

West, Heather, Sabol, William J., Greenman, Sarah J., 2010. Prisoners in 2009. Bureau of Justice Statistics Bulletin. U.S. Department of Justice. Office of Justice Programs.Bureau of Justice Statistics, Washington, DC. #NCJ 231675.

Western, Bruce, 2006. Punishment and Inequality in America. Russell Sage Foundation, New York.

Wildeman, Christopher, 2010. Parental incarceration and children's physically aggressive behaviors: evidence from the fragile families and child wellbeing study. Social Forces 89, 285–310.

Wildeman, Christopher, Muller, Christopher, 2012. Mass imprisonment and inequality in health and family life. Annual Review of Law and Social Science 8, 11–30.

Wildeman, Christopher, Schnittker, Jason, Turney, Kristin, 2012. Despair by association? The mental health of mothers with children by recently incarcerated fathers. American Sociological Review 77, 216–243.

Mass Killings and Genocide, Psychology of

Ervin Staub, University of Massachusetts, Amherst, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes influences that lead to violence as well as avenues to prevention and reconciliation. The influences include social conditions, the psychological reactions of groups of people to them, and the societal processes they generate. They include cultural characteristics, such as a history of devaluation of a group, past victimization and woundedness, and overly strong respect for authority. Great violence evolves progressively, making prevention possible. The article discusses avenues to prevention, reconciliation as a way of preventing of violence and renewed violence, and the role of active bystanders in prevention.

The primary starting points or instigators for genocide, mass killing, and other mass violence are (1) difficult conditions of life in a society, such as economic problems, political conflict, and great societal changes; (2) intense conflict between subordinate and dominant groups or conflicts involving vital interests; (3) material self-interest; and (4) war, this already intense violence, which makes genocide more likely when other instigators are also present. The first two instigators activate basic psychological needs in people. The resulting psychological and social processes develop in part as means of fulfilling these needs. These processes include individuals turning to a group for identity, scapegoating, and the adoption of destructive ideologies that offer the vision of a better future but also identify enemies.

As the group begins to harm members of another group an evolution follows, with changes in individuals and the group that allow and is likely to lead to increasing violence. The passivity of internal and external bystanders, outside groups and nations, encourages perpetrators. Once violence has begun, actions by the community of nations to halt its further evolution are essential. These can include intense diplomatic efforts by powerful actors, but at this late point often require military action. But genuine prevention means early actions, when conditions that indicate the likelihood of violence are present. It requires actions by both 'internal' and 'external' bystanders.

To understand and ultimately prevent mass violence, like the Holocaust, the genocide of the Armenians, the genocide in Rwanda, the 'autogenocide' in Cambodia, the mass killings in Argentina and Bosnia, and many others, the social conditions, cultural characteristics, psychological processes of groups and individuals that lead to it must be identified. Important questions and issues include What are the motives of perpetrators, how do these motives evolve, how do inhibitions against killing decline? What are the starting points or instigating conditions in particular instances? What characteristics of the cultures and social processes contribute? What is the psychology of perpetrators and bystanders?

An essential source of groups turning against other groups is the human proclivity to differentiate between 'us' and 'them' and the tendency to devalue 'them.' How can this tendency be mitigated? The passivity of bystanders, ranging from individuals to nations, encourages perpetrators. How can preventive responses be promoted? The conception of origins that follows is based primarily on Staub (1989); of prevention on Staub (2011) (see also Charny, 1999; Fein, 2007; Hamburg, 2007, and references below). (**Table 1** shows influences leading to and processes in mass violence, as well as avenues to prevention.)

Instigators of Collective Violence

These are conditions in a society or in a group's relationship to another group that have great impact on people. They give rise to psychological reactions in individuals and whole groups of people, and actions and events in a society or social group that lead the group to turn against another group, often a subgroup of the society.

Difficult Life Conditions

One starting point for mass violence is difficult conditions in a society, such as severe economic problems, great political disorganization and confusion, rapid social changes, and their combinations. These conditions often frustrate the fulfillment of core material needs, for food, shelter, and so on, but probably even more universally they frustrate basic psychological needs – for security, a positive identity, feelings of effectiveness and control, positive connections to people or a group, and a comprehension of reality.

Group Conflict

One type of conflict involves 'vital' interests. Even though these conflicts have objective elements, such as some territory both sides need for living space, the psychological elements are powerful. They include attachment to territory that is part of the groups' identity; mutual devaluation, distrust, and fear; the experience of assault on one's identity (Zartman et al., 2012); at times fear for the group's and one's own survival; and unfulfilled or frustrated basic psychological needs. These elements make the conflict especially difficult to resolve. Another type of conflict is between dominant and subordinate groups in a society. It is demands by the subordinate group for greater rights that can start active violence between the groups that may end in mass killing or genocide (Fein, 1993).

Table 1 The origins of violence between groups, prevention and Reconciliation[a]

A: Starting points	
Difficult life conditions – economic deterioration, political disorganization, great social changes	
Conflict between groups – material, psychological	
War, revolution	
Self-interest as a motivation (e.g., acquiring the resources of an indigenous group)	
Consequences of these starting points:	
The frustration of basic psychological needs	
Individuals turning to a group for identity and support	
Scapegoating	
(Destructive) ideologies	
B: History, culture, and current practices	
Elements contributing to violence (←):	**Elements of prevention (→):**
Devaluation of the other	Humanizing the other
Destructive, exclusive ideology	Constructive, inclusive ideology
Unhealed psychological wounds	Healing of past wounds
Uncritical respect for authority	Moderate respect for authority
Monolithic society	Pluralism (in values, access)
Unjust societal arrangements	Just social arrangements
Lack of freedom – Repression	Eliminating repression
Lack of truth about past events	Complex truth – Moving toward shared narratives/shared history
Passive bystanders	Active bystanders
C: Continuous processes	
The evolution of harm doing (changes in perpetrators, bystanders, institutions, social norms, and culture)	
Reciprocity in hostile actions and violence	*Reciprocity in positive actions*
The ongoing influence and role of leaders, followers, bystanders	

[a]Extended version of Table 4.1 in Staub, E., 2011. Overcoming Evil: Genocide, Violent Conflict and Terrorism. Oxford University Press, New York.

Like difficult life conditions, conflicts between groups can frustrate material needs and also tend to frustrate basic psychological needs. Difficult life conditions often go together with and intensify the impact of group conflict. Dominant groups, faced by demands from a subordinate group, often protect not only their rights and privileges, but also their security and identity, as well as their comprehension of reality, which includes their view of the 'right' social arrangements. They create and hold on to legitimizing ideologies, views of the world and justifications of their privilege. Such ideologies may stress the intelligence, hard work, moral character, and traditions of the privileged group.

Self-Interest

When a subgroup of society is greatly devalued (see below), a superior group may engage in mass killing to advance its interests. Mass killing or genocide of indigenous peoples has often been in part due to the desire to gain land or develop resources where these groups have lived (Hitchcock and Twedt, 1997).

War (and Violent Revolution)

When violence evolves and becomes intense, it can fairly easily spread from its original targets to other targets as well. When the conditions that tend to lead to violence against a subgroup of society exist, such as intense devaluation, a destructive ideology, or past conflict, the violence of war or revolution (Melson, 1992) can spread to such a target. War can also be a cover under which violence against a subgroup of society is perpetrated.

Turning against the Other

Groups could respond to instigating conditions by cooperative efforts to improve conditions or by resolving conflict through negotiation and mutual concessions. Instead, frequently a complex of psychological and social processes arises that turn the group against another and ultimately lead to violence. Individuals turn to a group for identity and security; people elevate their group by devaluing or harming others (Tajfel, 1978); they scapegoat another group for life problems or blame the other for the conflict; ideologies are adopted that offer a vision of a better life (nationalism, communism, Nazism, social equality in Cambodia, Hutu power in Rwanda, and so on), but also identify enemies who must be 'dealt with' in order to fulfill the ideology.

The Evolution of Destructiveness

The group and its members begin to take actions that harm the other group and its members. This begins an evolution. Individuals and whole groups 'learn by doing.' As they harm others, the beliefs and values of the perpetrators and the whole society they are part of change. The processes through which this happens include just world thinking, the resolution of cognitive dissonance, and adaptation. Just world thinking is the belief that

the world is a just place and those who suffer must have somehow deserved their suffering. Cognitive dissonance is the result of a discrepancy between beliefs and actions. To reduce this discrepancy, people devalue further those they are harming and change their views of what is moral in relation to the victim group. They may start with disengagement from their moral values in relation to victims (Bandura, 1999). Perpetrators, and even bystanders, move on to moral exclusion, excluding the victimized group and its members from the moral realm, the realm in which moral values and standards apply (Staub, 1989, 2011; Opotow, 2011). They also replace moral values that protect other people's welfare with values such as obedience to authority or loyalty to the group. Many of them progress to moral transformation. Killing members of the devalued group becomes the right thing for them to do, a reversal of morality (Staub, 1989, 2011). As the evolution progresses individuals change, the norms of group behavior change, new institutions are created to serve violence (for example, paramilitary groups and government offices serving persecution and destruction).

Contributing Cultural Characteristics

Certain characteristics of a culture make it more likely that, in response to instigation, the psychological and social processes that initiate group violence arise.

Cultural Devaluation

One of these is a history of devaluation of another group or subgroup of society. Such devaluation can vary not only in intensity but also in its nature: the other is lazy; of limited intelligence; manipulative, morally bad; dangerous, an enemy that intends to destroy one's own group. A devalued group that does relatively well – its members have good jobs and some standing in the society – is an especially likely victim. Sometimes two groups develop intense, mutual hostility, which may be called an ideology of antagonism (Staub, 1989, 2011). Seeing the other as their enemy, and themselves as an enemy of the other, becomes part of each group's identity. This makes group violence more likely.

Overly Strong Respect for Authority in a Society

This makes it especially difficult to constructively deal with instigating conditions. Accustomed to being led, people are more likely to turn to leaders, even if they lead in destructive directions, and ideological groups. They are unlikely to oppose it when their group increasingly harms another group. They are less likely to oppose destructive leaders and more likely to follow direct orders to engage in violence.

Monolithic (and Autocratic) vs Pluralistic (and Democratic) Societies

The more varied are the values in a society, the greater the extent that all groups can participate in societal processes, the less likely is the evolution toward mass violence. People will be more likely to oppose harmful, violent policies and practices. Democracies (Rummel, 1994), especially mature ones (Staub, 1999) that are pluralistic and have a well-developed civic culture, are unlikely to engage in genocide.

Unhealed Wounds of a Group due to Past Victimization or Suffering

Without healing from past victimization, the group and its members will feel diminished, vulnerable, and see the world as a very dangerous place. At times of difficulty or in the face of conflict, they may engage in what they think of as necessary self-defense. But, instead, this could be the perpetration of violence on others (Staub, 1998, 2011).

Other Cultural Characteristics

A history of aggression as a means of resolving conflict, as well as certain group self-concepts – a sense of vulnerability, or a feeling of superiority that is frustrated by events, or the combination of the two – make violence also more likely.

The Role of Bystanders

The passivity of bystanders greatly encourages perpetrators. It helps them believe that what they are doing is right. Unfortunately, bystanders are often passive, and go along with a system that is becoming increasingly destructive (Barnett, 1999; Charny, 1999; Staub, 1989, 2011). Internal bystanders, members of the population, often go along with or participate in discrimination. As a result, just like perpetrators, they change. Like the perpetrators, bystanders, as members of the same society, have also learned to devalue the victims. They are also affected by instigating conditions. It is difficult for them to oppose their group, especially in difficult times and in an authority-oriented society. To reduce their guilt and their empathy, which makes them suffer, bystanders often distance themselves from victims. Over time, some become perpetrators (Lifton, 1986).

External bystanders, outside groups, and other nations, also usually remain passive, continue with business as usual, or even support perpetrators. Nations do not see themselves as moral agents. They use national interest, defined as wealth, power, and influence, as their guiding value. When they have ties to another country, they tend to support the leaders, not a persecuted group. This has been changing to some degree, especially, in principle and to a limited degree in practice, with new humanitarian laws, a new international court to try violators, the United Nation's general assembly's resolution of the Responsibility to Protect (the responsibility of the international community to protect the citizens of a country if its own government does not). However, in spite of more preventive interventions by the UN (Fein, 2007), there is no system of enforcement and often the world has remained inactive in the face of conditions indicating the likelihood of violence and even in the face of seemingly impending, or already ongoing violence.

The Role of Leaders

It is the inclinations of a population, the result of the joining of culture and instigating conditions that to a substantial degree

create the possibility and likelihood of mass murder. The people select, or are willing to follow, leaders who respond to their inclinations and fulfill their needs.

But leaders and the elite have an important role in shaping and influencing events. They scapegoat and offer destructive ideologies, and use propaganda to intensify negative images and hostility. They create institutions, such as media and paramilitary groups that promote or serve violence. Often destructive leaders are seen as acting only to gain support or enhance their power. But leaders are also members of their society, impacted by life conditions and group conflict, and, at least in part, act out of the motives and inclinations described above.

Other Influences

One additional influence is a sudden shift in government, combined with 'state failure,' the new government failing to deal effectively with problems that face the society (Melson, 1992). An ongoing war, especially a civil war, also adds to the probability of mass killing or genocide. Economic interconnections between a country and other countries make genocide and mass killing less likely (Harff, 2003).

Halting Persecution and Violence

Once violence against a group has become intense, halting it requires action by nations and the community of nations. Early warning is important, but it only has value if it leads to action. Often, as in the case of the genocide in Rwanda in 1994 (des Forges, 1999), when information about impending violence is available, the international community does not respond. For this to change requires changes in values, and actions by citizens leading governments to build institutions that activate responses by the community of nations. Institutions within governments that have the responsibility for prevention, and connections between these, the UN, and regional organizations are needed.

Appropriate actions include diplomatic efforts: to warn perpetrators, as well as to offer mediation and incentives to stop violence. These must be accompanied, as needed, by withholding aid, by sanctions and boycotts – ideally designed to affect leaders and elites – and the use of force if necessary. To prevent rather than halt violence, to avoid crises, and the need to use force requires early preventive actions when conditions exist that are likely to lead to mass violence (Carnegie Commission on the Prevention of Deadly Conflict, 1997; Hamburg, 2007; Staub, 1999, 2011).

Prevention and Reconciliation

Effective prevention often requires joint effort by internal and external bystanders. External actors can be instrumental in activating internal actors. Unfortunately, the international community rarely responds to indicators of *potential* violence (Hamburg, 2007; Staub, 2011).

Reconciliation between hostile groups can prevent violence in the first place; reconciliation after violence can prevent renewed violence (Staub, 2011).

Creating positive connections between groups, significant contact between members of hostile groups in shared efforts on behalf of joint goals, helps people overcome past devaluation and hostility (Pettigrew and Tropp, 2006). Coming to understand the other's history and culture can also contribute. Democratization (Rummel, 1994), which creates pluralism, moderates respect for authority, and tends to lessen differences in power and privilege, is important for prevention.

Helping to heal both previously victimized groups and members of perpetrator groups – who are wounded, at the very least, by their own violent actions or by their inaction in the face of their group's violence – reduces the likelihood of new or renewed violence. Testimonies, commemorations, especially if they also focus on a positive future, are among processes of healing. Establishing the truth about what happened in cases of prior collective violence, and groups moving toward a shared history, in place of conflicting narratives about who did what, also helps with reconciliation and peace. Assumptions of responsibility and expressions of regret by perpetrators, which are usually difficult to bring about (or mutually, when violence was mutual) further healing, forgiveness, and reconciliation. The punishment of especially responsible perpetrators (but not revenge on a whole group) is important (Staub, 2011). Improving people's economic situation, through equitable economic development of a country, can also diminish the likelihood of violence.

Understanding the influences that lead to violence between groups, as well as avenues to prevention and reconciliation, can have important positive effects. They can increase both resistance to instigating conditions and openness to reconciliation. They can motivate active bystandership in the service of prevention and building a peaceful society. Such understanding can be promoted through workshops and trainings, as well as through educational radio programs, as this was done in Rwanda (Staub, 2011; Staub et al., 2010). Finally, but importantly, the positive socialization of children, the development of inclusive caring, caring that extends beyond the group, and moral courage that enables people to resist destructive influences, are also essential for long-term prevention (Staub, 2005, in press). (See **Table 1** for an overview of approaches to prevention.)

Future Directions

This article has reviewed influences that lead to varied forms of mass violence. Further research ought to consider whether specific forms of violence, such as government persecution, conquest, revolution, civil war, and others, which may ultimately lead to mass killing, also have specific or unique determinants (Totten et al., 1997). Testing our capacity to predict group violence is important. So is the development of techniques to help groups heal and reconcile (Agger and Jensen, 1996; Staub, 2011). Creating positive bystandership by nations and nongovernmental organizations is essential for prevention. Citizen involvement is required for this. When this happens, the effects of different types of bystander behavior need to be assessed. To create a less-violent world, the development of knowledge in this realm and its application has to go hand in hand.

See also: Aggression, Social Psychology of; Belief in a Just World; Cognitive Dissonance; Dehumanization; Identity and Identification, Social Psychology of; Intergroup Relations; Obedience: Social Psychological Perspectives; Prosocial Behavior and Empathy; Social Movements: A Social Psychological Perspective; Social Psychology; Xenophobia: Social Psychological Aspects.

Bibliography

Agger, I., Jensen, S., 1996. Trauma and Recovery under State Terrorism. Zed, London.

Bandura, A., 1999. Moral disengagement in the preparation of inhumanities. Personality and Social Psychology Review 3, 193–209.

Barnett, V.J., 1999. Bystanders: Conscience and Complicity during the Holocaust. Greenwood Press, Westport, CT.

Carnegie Commission on the Prevention of Deadly Conflict, 1997. Preventing Deadly Conflict: Final Report. Carnegie Corporation of New York, New York.

Charny, I.W. (Ed.), 1999. Encyclopedia of Genocide, vols. 1 and 2. ABC-CLIO, Santa Barbara, CA.

Fein, H., 1993. Accounting for genocide after 1945: theories and some findings. International Journal of Group Rights 1, 79–106.

Fein, H., 2007. Human Rights and Wrongs: Slavery, Terror, Genocide. Paradigm Publishers, Boulder, CO.

des Forges, A., 1999. Leave None to Tell the Story: Genocide in Rwanda. Human Rights Watch, New York.

Hamburg, D., 2007. Preventing Genocide: Practical Steps toward Early Detection and Effective Action. Paradigm Publishers, Boulder, CO.

Harff, B., 2003. No lessons learned from the Holocaust? Assessing risks of genocide and political mass murder since 1955. American Political Science Review 97 (1), 57–73.

Hitchcock, R.K., Twedt, T.M., 1997. Physical and cultural genocide of various indigenous peoples. In: Totten, S., Parsons, W.S., Charny, I.W. (Eds.), Century of Genocide: Eyewitness Accounts and Critical Views. Garland Publishing, New York.

Lifton, R.J., 1986. The Nazi Doctors: Medical Killing and the Psychology of Genocide. Basic Books, New York.

Melson, R., 1992. Revolution and Genocide. University of Chicago Press, Chicago.

Opotow, S., 2011. Reconciliation in times of impunity: challenges for social justice. Social Justice Research 14 (2), 149–170.

Pettigrew, T., Tropp, L., 2006. A meta-analytic test of intergroup contact theory. Journal of Personality and Social Psychology 90, 751–783.

Rummel, R.J., 1994. Death by government. Transaction, NB, Canada.

Staub, E., 1989. The Roots of Evil: The Origins of Genocide and Other Group Violence. Cambridge University Press, New York.

Staub, E., 1998. Breaking the cycle of genocidal violence: healing and reconciliation. In: Harvey, J. (Ed.), Perspectives on Loss: A Source Book. Taylor & Francis, Washington, DC.

Staub, E., 1999. The origins and prevention of genocide and other group violence. Peace and Conflict: Journal of Peace Psychology 5, 303–336.

Staub, E., 2005. The roots of goodness: the fulfillment of basic human needs and the development of caring, helping and nonaggression, inclusive caring, moral courage, active bystandership, and altruism born of suffering. In: Carlo, G., Edwards, C. (Eds.), Moral Motivation through the Life Span: Theory, Research, Applications. Nebraska University Press, Lincoln, pp. 33–73. Nebraska Symposium on Motivation.

Staub, E., 2011. Overcoming Evil: Genocide, Violent Conflict and Terrorism. Oxford University Press, New York.

Staub, E. The Roots of Goodness and Resistance to Evil: Inclusive Caring, Moral Courage, Altruism Born of Suffering, Active Bystandership and Heroism. Oxford University Press, New York, in press.

Staub, E., Pearlman, L.A., Bilali, R., 2010. Understanding the roots and impact of violence and psychological recovery as avenues to reconciliation after mass violence and intractable conflict: applications to national leaders, journalists, community groups, public education through radio, and children. In: Salomon, G., Cairns, E. (Eds.), Handbook of Peace Education. Psychology Press, New York, pp. 269–287.

Tajfel, H., 1978. Social categorization, social identity and social comparison. In: Tajfel, H. (Ed.), Diferentiation between Social Groups. Academic Press, London, pp. 61–76.

Totten, S., Parsons, W.S., Charny, I.W. (Eds.), 1997. Century of Genocide: Eyewitness Accounts and Critical Views. Garland Publishing, New York.

Zartman, I.W., Antsey, M., Meers, P., 2012. The Slippery Slope to Genocide: Reducing Conflict and Preventing Mass Murder. Oxford University Press, New York.

Mass Media and Cultural Identity

Philip Schlesinger, University of Stirling, Stirling, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9341–9346, © 2001, Elsevier Ltd.

Abstract

Mass media are media of communication such as the printed press, cinema, radio, and television that address large and diverse audiences. Cultural identity refers to a set of qualities attributed to given populations, often thought of as static but which change through time. Mass media have played a key historical role in shaping national cultures and therefore a range of cultural identities. The press, cinema, radio, and television have successively impacted on cultural and national identities within the international system of nation-states. An account is offered of Europe's century-old concern about 'Americanization' through Hollywood films or US television programs. The Third World's interest in a New World Information and Communication Order to redress the unbalanced global flow of news and entertainment is discussed. Media imperialism is considered in relation to the defense of cultural identity and national sovereignty. An account is given of the trade dispute between the USA and the EU over the cultural status of audiovisual products, showing the connection to ideas of cultural identity. Finally, examples of relationships among mass media, language, and cultural identity are presented. The article concludes by noting that 'globalization' will not eliminate the diversity of cultural identities.

The term 'mass media' refers to media of communication such as the printed press, cinema, radio, and television which have characteristically addressed large and diverse audiences. The term 'cultural identity' refers to the attribution of a given set of qualities to a particular population. A cultural identity is not static and eternal but rather changes through time. Cultural collectivities commonly think of their identities in terms of how they differ from others. Distinguishing 'us' from 'them' is therefore central to groups' cultural self-identification. In this article, the relationship between mass media and debates about cultural identity is examined.

How Mass Media Relate to Cultural Identity

Throughout the history of mass communication, there has been a continuing interest in the relationships between mass media and cultural identities. This has been driven by the general, underlying assumption that different media variously form, influence, or shape collective identities as such. How mass media might, at various times, specifically affect the formation and maintenance of national cultures and identities has occupied center stage in this continuing discussion. National identities normally embrace a range of cultural identities based on ethnic, religious, and linguistic diversity. Nationalists, however, often wish to portray national identity as based in a single culture. Nation-states remain, for the most part, the political frameworks in which claims about cultural diversity and expressions of cultural identity within a given society are pursued. Nevertheless, cultural identities may transcend the state system, as for instance in the case of systems of belief that unite members of the world religions. Diasporic communities also share cultural identities despite belonging to different nations and holding diverse citizenships (Schlesinger, 1991, Hall and du Gay, 1996).

Mass Media and National Identities

From a historical perspective, mass media have been an important part of the nation-building process. Successively, the press, the cinema, radio, and television have been invoked as shapers of collective consciousness, as the bearers of a collectivity's culture, or as objects of policy for shaping collective identities. To the extent that the mass media have contributed to the construction of national cultures, they have also necessarily played a role in the creation of national identities, which are still the most important forms of modern cultural identity in the new millennium.

In the latter half of the nineteenth century, the development of a popular press, nationally distributed magazines, and news agencies making use of the telegraph reinforced the creation of national communication systems in countries such as the USA, the UK, and France. Mass communication developed intensively within the boundaries of national states and through widespread, cross-class consumption, it contributed to the creation of common space for public debate, a shared political agenda, and that sense of collective belonging that characterizes national identities (Carey, 1989). Mass communication has therefore played an increasingly central part in the formation of national political cultures. The latter part of the nineteenth century also saw the first major steps in the development of what would come to be a worldwide communications capacity carried through the spread of telegraphy and subsequently telephony in part in line with European imperial expansion. These nineteenth-century beginnings have proved to be the foundation stones of the developing global communications infrastructure which has become steadily more relevant for debate about the development of transnational cultural identities at the end of the twentieth century and at the start of the twenty- first.

Audiovisual Culture

The rapid growth of the cinema before and up to the First World War inaugurated an era in which the moving image ultimately came into its own as the globally dominant cultural form. From its inception, the cinema developed as a national cultural institution. It provided images of the people and,

through stardom, of popular icons of identification that played into the elaboration of national identities by representing a culture. At the same time, however, the early cinema was also an international medium that addressed diverse audiences across the boundaries of nation-states. This duality has made the relationship between the moving image and the maintenance of cultural identity within the territorial boundaries of the nation-state an especially potent issue for many countries.

Since World War I, the role of Hollywood in the international cinematic marketplace has been of prime importance. What has accompanied this primacy has been a widespread concern, long articulated in Europe by politicians and cultural elites from the 1920s onwards, about the impact of 'Americanization' – the exporting of US cultural values, tastes, attitudes, and products – on national cultures and identities. The most vociferous critics have seen the transnational flow of the moving image as an instrument of a US government policy intent on global cultural domination. Other detractors have seen 'Americanization' as the unintended outcome of a successful model of production that invites competition as a mode of self-defense.

Given its salience, it is hardly surprising that throughout the twentieth century 'Americanization' has repeatedly come to the center of public debate. Looked at from the vantage point of elites, it has been identified with the reshaping of popular culture and seen as having a negative impact on actions, beliefs, and identities, especially among young people and the working class. For such groups, American cultural forms have been a key source of pleasure: they have embodied modernity and an escape from a dominant version of national culture.

Although Hollywood's output has often been the principal point of attack, there has also been a much more diffuse concern among national elites with the effects of consumerism on the moral fabric of their societies. So, for instance, jazz and rock, fast food and Coca Cola, shopping malls and coffee bars, advertising and blue jeans have each, at times, come to represent the threatening and transformative ubiquity of America to disconcerted moral guardians. The cultural challenge constituted by 'Americanization,' moreover, has also gone deeper into the social fabric to encompass the challenges represented by the economic model of free-market capitalism. This has long been seen as inimical to the social solidarity ideally aspired to by European welfare states. Not surprisingly, then, the history of US market dominance and cultural anxiety in recipient countries has played into intercultural relations across the Atlantic. It also shaped the wider, global debate about cultural and media imperialism.

Alongside the rise of the cinema, the development of nationally based radio broadcasting in the inter-War years and through to the 1950s ensured the dissemination of a range of genres – drama, music, news, comedy, sport – that addressed segments of nationally constituted audiences. Radio domesticated national culture, defined national public spheres, and ensured that cultural identities became deeply mass mediated in many respects.

The audiences for radio and, later, television shared a public world which persisted for most of the twentieth century. The classic vehicle for explicitly pursuing the role of televising the national culture from the 1950s onwards was public service broadcasting, of which the British Broadcasting Corporation (BBC) has been the key model, broadly replicated in many countries (Scannell and Cardiff, 1991).

Before its networks and channels began to fragment and specialize with the diversification of distribution systems, general broadcasting offered a container of common mediated experiences to its audiences and articulated these with other forms and sites of cultural consumption. Formally, through the education system, learners would encounter variants of the national folk history and absorb key literary reference points that in combination offered marks of collective identity and thereby constituted collective memories. More informally, outside the sphere of broadcasting, whether as citizens or consumers, people encountered the national cultural gamut in all its variety. This included recorded and live music, popular comedy, theatrical performances, and national and local sporting events. State ceremonies and religious celebrations punctuated the calendar. By both complementing and re-engineering these cultural forms into its own output, broadcasting invited listeners and, later, viewers into a new relationship of social communication. It offered shared ways of speaking and looking that became an integral part of everyday life. It invented schedules that structured temporal patterns and genres that shaped cultural expectations.

Although its transmissions are no respecters of political or cultural boundaries, for the most part radio has been experienced by listeners as a national medium. However, it has always played an important transnational role in wartime propaganda, whether in the hot war of World War II or during the Cold War from the late 1940s until the turn of the 1990s. And music radio has always had strong cross-border attractions, sustaining international communities of taste.

Besides public service systems, commercial systems of broadcasting such as the US networks – ABC, CBS, NBC – both in their heyday and since, have also had a profound impact on the mediation of national culture by way of providing a wide range of televisual forms. US programs have been pre-eminent in defining the international experience of television for many countries. National television systems, and latterly subscription-based movie channels, moreover, have been important vehicles for the dissemination of cinema, especially of Hollywood's output, which has thereby reached a global audience alongside distribution via the theatrical circuit and through the sale and rental of video recordings.

As the ways in which television was distributed rapidly diversified in the 1980s and 1990s – with the growth of cable and satellite broadcasting, alongside the original terrestrial systems – the relations between cultural identity and the medium became more complex. The national audience's experience of television consumption began to fragment. These changes ushered in the crisis of public service broadcasting in many countries. Fragmentation of the national television experience is likely to be further accentuated as digital signals increasingly replace analogue transmissions, and as the range of channels on offer multiplies rapidly in the most developed parts of the world. It is also going to be affected in increasing measure by the delivery of radio, television, and music over the Internet, as these will impact on the range of available choices and bypass broadcasting schedules.

Cultural Identity, Mass Media and Cultural Imperialism

The debate about the impact of mass media on cultural identity increasingly rose up the policy and academic agendas in the 1970s. This coincided with the new interest being taken in the cultural impact of films and television programs that were crossing state frontiers. Shaped by the political contours of the Cold War, in which the strategic interests of the capitalist West were both symbolically, and in reality, opposed to those of the communist East, the debate was greatly stimulated by the postcolonial and postimperial interest in the influence exercised over the rest of the world of cultural production deriving from the main capitalist states (Tomlinson, 1991).

International Debate on Cultural Identity

Discussion centered on the work of the United Nations Educational, Social and Cultural Organization (UNESCO) and on that of the critical writings of media scholars such as Schiller (1969) and Mattelart (1979). In the early 1970s, the first major studies of international television flows were published and academic debate about the significance of audiovisual trade intensified. In Latin America, which became the key locus for critical thought, arguments centered on the economic and cultural dependence of Third World states on those of the developed First World and the perceived threat posed to indigenous cultures and identities by popular media and consumerism.

UNESCO's interventions stimulated a critique of US cultural and media imperialism. The role of journalism (not least US-, UK-, and French-owned international news agencies) was central to this debate. The news agencies were seen as exercising monopolistic control over news flows and as screening out 'south-south dialogue.' Arguments centered on the worldwide impact of the consumption of Western cinematic and televisual fiction as well as that of news. To counter media imperialism, its critics proposed that a New World Information and Communication Order (NWICO) be instituted. The idea of such an international order implied the existence of a transnational communicative space that should and could be regulated in the interests of more equal dialogue. The defense of national cultures in weak and dependent nations and the pursuit of equity and balance in the flow of reports and images in international cultural trade reached its high-water mark in the UNESCO-supported MacBride report (MacBride, 1980). This debate prefigured later attempts to discuss the conditions for creating an international public sphere of open debate (Habermas, 1997).

Supranationalism and Cultural Identity

The defense of national cultural identities by nation- states has remained on the international media research agenda, and not just in the Third World. The European Union (EU) has been a key test case for thinking about the relationships between mass media and cultural identity. The EU is the outcome of a half-century-long process dating from World War II, in which previously warring states have sought to develop a framework for political and economic integration. By the end of the twentieth century, 15 European states were bound together by treaty as a trading bloc in the global economy, with a growing line waiting for accession. Official thinking in the EU has assumed a strong causal connection between media consumption and cultural identity. The desired shaping of a new, 'European' cultural identity has been linked to the formation of a supranational political public sphere and a common citizenship.

EU policy-makers' views on imports of US audiovisual products have strikingly echoed arguments aired in the NWICO debate: from a cultural perspective, the flow of films and television programs has been represented as a threat to European identity. As noted earlier, ever since World War I, political and cultural elites in Europe have been concerned about the impact of American films on their national cultures and identities (Jarvie, 1992). The original national focus has been easily transposed into official, supranational preoccupations about 'European culture.'

The cultural case for domestic film-making has been that home-made moving images represent the nation and invite its members to see themselves as part of a community with a common culture. As television broadcasting became a cultural force across Europe from the 1960s on, official and industrial worries about US dominance of the box office extended to the small screen. There has also always been a major economic interest in limiting imports and safeguarding jobs at home. Hence, the EU successfully negotiated the 'cultural exclusion' of audiovisual services from the final settlement of the General Agreement on Trade and Tariffs (GATT) in 1993. The USA did not accept this principle and the issue remains on the agenda of the successor World Trade Organization.

The EU has found it impossible so far to create a single European cultural marketplace for the moving image. The aspirations have stretched no further, and reasonably so. That is because in the case of radio, television (imported entertainment aside), and the press, cultural preferences have remained predominantly national and regional, following the contours of states and languages. There is no reason to think that this is going to change, even though English is increasingly widely diffused as a second language.

Despite their political convergence, European states still offer a bulwark against widespread cultural homogenization. Publishing, the press, radio, and national education systems continue to sustain diversity and difference in each country. Perhaps it is not so surprising, then, despite some relatively modest investment in financing cross-national film and television coproductions, in supporting producer training and distribution, and in enforcing the provisions of the EU's single-market legislation, that there is little evidence of the growth of a major transnational audience for European audiovisual products. Both national and transnational patterns of cultural consumption co-exist. So, at the start of the new millennium, despite 15 years of policy intervention, Hollywood still provided the common audiovisual diet, with national cinema an also-ran everywhere apart from France, where cultural protectionism had been most vigorously enforced.

Although a supranational polity, the EU is based on nation-states that seek to maintain their distinctive national cultures. The EU's integration process has stimulated a reassertion of regional identities, thereby undercutting the hegemony of nation-states. In some instances, these regions are also 'nations without states' such as Catalonia and the Basque Country in

Spain, and Scotland and Wales in the UK. There, regionalism is inescapably connected with the protection of national and cultural identity. Particular importance has been attached to the role of indigenous mass media in sustaining distinct national cultural identities within the wider state.

Mass Media, Language, and Cultural Identity

National media systems have been crucial disseminators of standardized vernaculars. Anderson (1991) has suggested that the convergence of capitalism and printing created the conditions for the dissemination of secular vernaculars in the early modern period. 'Print language' allowed national languages to be disseminated through markets for books and newspapers, creating a common culture. Audio and audiovisual media have performed a similar role in reinforcing the national idiom. For instance, it was not until the post-World War II period that the majority of Italians came to speak standard Italian, and this was substantially due to the role of radio and television broadcasting, reinforcing and complementing the earlier influence of the cinema. Similarly, in Latin America, it was the mass media that first gave audiences images of themselves that they could connect to their everyday life experiences, offering them an entry point into the national culture and the political public sphere. The mass media also provided the idioms of a common culture. Key instances were the Mexican cinema of the 1930s to 1950s, Argentinean radio drama from the 1920s to the 1940s, and popular journalism in Chile from the 1920s to the 1950s (Martín-Barbero, 1993).

Although mass media have had a homogenizing impact within national boundaries, in some places they have been used to ensure language diversification. This has depended on the political efficacy of language activists seeking access to the media. Success or failure has depended on the status of the language concerned, the amount of political support for broadcasting or otherwise disseminating a language, whether a language campaign has been effectively led, and the state authorities' willingness to encourage linguistic diversity.

Many European states regulate their broadcasting systems to ensure that minority and 'lesser-used' languages are assured of airtime and thereby encourage and sustain a diversity of cultural expression and identities. The policies of individual states have been supported since 1992 by a European Convention on regional or minority languages and in several cases – notably in the Basque Country and Catalonia in Spain, Wales in the UK, and among the Sami of Scandinavia – minority-language broadcasting appears to have been successful in sustaining their language communities. However, Europe is by no means unique. Movements aimed at securing both public support for, and state endorsement of, indigenous radio and television production have been active in Australasia and the Americas (Browne, 1996).

Canada has been an especially illuminating case of the complex inter-relations between linguistic and cultural identity considerations as the Canadian state has fought its identity wars on two fronts. Externally, the country has had to define itself politically and culturally against the USA. Internally, Canada has had to deal with the continuing debate over the status of Quebec, where claims to sovereignty have long been argued in terms of a distinctive national cultural identity (Collins, 1990).

By the middle of the twentieth century, the advent of television extended the existing US media penetration of English-speaking Canada by print and radio. The presence of American media has long shaped Canadian policy-makers' understanding of broadcasting as essential to the maintenance of national identity and cultural sovereignty. As commercial television developed alongside the public sector, governments tried to regulate the system to ensure that Canadian content was carried by Canadian broadcasters, alongside highly popular US material. Consequently, in line with this protectionist stance, when Canada negotiated the North American Free Trade Agreement (NAFTA) with the USA in 1993, it sought successfully to ensure the 'exception' of its cultural industries from the treaty (McAnany and Wilkinson, 1996). This paralleled the EU's approach to GATT, for similar reasons.

While the establishment of public service broadcasting in Canada was in part a response to the commercial model prevailing south of the border, it also took the form of reinforcing the country's cultural and linguistic fault-line between Anglophones and Francophones. Although Francophone broadcasting has sustained the national cultural identity of French speakers, this has run counter to the Canadian federal government's attempts to create an overarching political and cultural identity.

Such tensions have also been experienced in the heartland of global media production, the USA. The major migratory impact of Hispanics has transformed the cultures of states such as California and Florida and has provided the market conditions for the development of Spanish-language television channels with a national reach. This assertion of cultural and linguistic difference has provoked debate about the unique position of the English language in the USA and how multiculturalism relates to a common national identity.

Globalization and New Cultural Identities

In the twenty-first century, the impact of transnational and global changes on what are still largely nation-state-bound systems of communication is likely to increase (Castells, 1997). Ideas of communication sovereignty rooted in the heyday of the nation-state are increasingly difficult to sustain in the face of cross-national media mergers, the development of supranational polities such as the EU, and the rapid, if unequal, diffusion of access to the Internet with its capability of making global communicative connections. Although the worldwide flow of information and cultural products has now permeated the boundaries of states in unprecedented ways, such 'globalization' does not mean that the world is becoming uniform. On the contrary, the reshaping of communicative spaces worldwide means that new cultural identities will emerge continually within, and across, the existing international system of nation- states.

See also: Cultural History; Culture, Sociology of; Entertainment; Film and Video Industry; Identity and Identification; International Communication: History; Mass Media, Political

Economy of; Mass Media, Representations In; Mass Media: Introduction and Schools of Thought; Media Imperialism; Media and Child Development; Public Broadcasting; Public Sphere and the Media.

Bibliography

Anderson, B., 1991. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London.

Bauman, Z., 1998. Globalization: The Human Consequences. Polity Press, Cambridge, UK.

Browne, D.R., 1996. Electronic Media and Indigenous Peoples: a Voice of Our Own? Iowa State University Press, Ames, IA.

Carey, J.W., 1989. Communication as Culture: Essays on Media and Society. Unwin Hyman, Boston.

Castells, M., 1997. The Power of Identity. Blackwell, Malden, MA.

Collins, R., 1990. Culture, Communication and National Identity: the Case of Canadian Television. University of Toronto Press, Toronto, Canada.

Friedman, J., 1994. Cultural Identity and Global Process. Sage, London.

Habermas, J., 1997. Between Facts and Norms. Polity Press, Cambridge, UK.

Hall, S., du Gay, P. (Eds.), 1996. Questions of Cultural Identity. Sage, London.

Jarviel, 1992. Hollywood's Overseas Campaign: the North Atlantic Movie Trade, 1920–1950. Cambridge University Press, New York.

MacBride, S., 1980. Many Voices: One World. Kogan Page, London.

McAnany, E.G., Wilkinson, K.T. (Eds.), 1996. Mass Media and Free Trade: NAFTA and the Cultural Industries. University of Texas Press, Austin, TX.

Martín-Barbero, M., 1993. Communication, Culture and Hegemony: from the Media to Mediations. Sage, London.

Mattelart, A., 1979. Multinational Corporations and the Control of Culture. Harvester Press, Brighton, UK.

Morley, D., Robins, K., 1995. Spaces of Identity: Global Media, Electronic Landscapes and Cultural Boundaries. Routledge, London.

Scannell, P., Cardiff, D., 1991. A Social History of British Broadcasting. Blackwell, Oxford, UK.

Schlesinger, P., 1991. Political Violence and Collective Identities. Sage, London. Media, State and Nation.

Schiller, H., 1969. Mass Communication and American Empire. A. M. Kelley, New York.

Tomlinson, J., 1991. Cultural Imperialism. Pinter, London.

Mass Media and Politics

Walt Borges, University of North Texas – Dallas, Dallas, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since the emergence of political science as a research discipline, the influence of mass media on political behavior and political institutions has been well debated, with numerous conflicting and complementary theories proposed to explain how the Fourth Estate reports, interprets, analyzes, and contextualizes information about government and public affairs. In the late twentieth and early twenty-first centuries, technological advances in computers and communication, the reorganization of commercial media operations, and the use of communication professionals skilled in marketing and message manipulation have triggered reconsideration of the roles and functions of mass news media in a world of multiplying news sources and fragmented audiences.

Behavioral Effects and Institutions

From the attempts of sociologists to explain the appeal of totalitarian regimes through their use of mass media, to current efforts to understand the political effects of a worldwide network of interpersonal relationships through Internet and wireless technologies, the study of politics and mass media has developed through investigations of two sets of relationships. One branch of research explores mass media effects on the behaviors of political participation, electoral choice, and policy choice. Another branch examines the impact of the mass media, specifically news media and organizations, as a set of political institutions within a democratic framework. Both effects and institutional approaches are framed within democratic theory, which recognizes the need for citizen knowledge of public affairs and government as a foundation for effective political choices. An understanding of public affairs, diverse political interests, and the general structure of the political system provides the basis for electoral choices involving candidates, efforts to influence policy choices, and citizen participation that ultimately endows legitimacy on a political system.

Theoretical Origins

At the beginning of the twentieth century, mass media could be defined entirely by the concept of the press. The high-speed printing press made possible the production of large numbers of local newspapers. While the telegraph and telephone had enabled accelerated news gathering over great distances in the United States, Europe, and the rest of the world, the production of news in the United States was almost exclusively local. The newspapers were almost exclusively locally owned, and the families and corporations that operated these newspapers were integrated economically, socially, and politically into their communities. The exception was national and international news cooperatives that circulated stories of broader geographical reach and national and international scope.

Real concern about the political effects of mass media first emerged with the use of radio propaganda during the First World War and the post-war creation of radio networks both in the United States and abroad. Of particular concern to sociologists and democratic scholars was whether the use of radio and other mass media in Nazi Germany and the Soviet Union allowed regimes controlling those media to manipulate their populations toward totalitarian ends. The result was a new interest among social scientists in how mass media functioned and how it affected the democratic process. Walter Lippmann (1922) pioneered this connection in his book on public opinion, but it was Harold Lasswell's formulations in the 1930s that sketched the five-point analytical framework for communications. Lasswell's analytical construct (1969) defined the message sender, message content, the message recipients, the communications channel, and the effects as key elements in the communications process.

The initial supposition of public opinion- and propaganda-based studies was that the mass media or message sender injected a political message into an audience of passive message recipients, with effects that varied by recipient according to internal interaction of the message with existing beliefs and attitudes within the individual. These interactions can produce different understanding of the same message. This hypodermic model of communication describes a one-step flow of communication between the mass media and its audiences that predicts powerful effects of the news media.

Mass Media and the Flow of Information

The problems with that model were revealed by a series of elections studies conducted by the Columbia sociologists working with Paul Lazarsfeld. In their study of Erie County, Ohio, in the 1940 presidential election, Lazarsfeld et al. (1944) questioned a panel of respondents seven times, finding that only a few voters changed their minds during the campaign. Lazarsfeld et al. found that the changes had far less to do with campaign news and campaign communications than it had to do with the influence of family and friends. Similar findings in a 1948 study of Elmira, New York, during the 1948 presidential campaign (Berelson et al., 1954) dented the idea that media coverage of presidential campaigns could persuade large numbers of voters. In response, Katz and Lazarsfeld (1955) developed a new model of a two-step flow of communication to emphasize the importance of interpersonal communications. The two-step flow of communication begins with messages disseminated through the mass media. However, rather than being directly received by an audience of

individuals who are attentive, the messages are received by a layer of opinion leaders who are interested and engaged in public affairs. Opinion leaders absorb the messages from the mass media, recast and reinterpret the messages, and through personal connections, pass them along to an audience that is often distracted, unaware, or uninterested in political matters.

The two-step flow model has been criticized from several angles. First, from its formulation, there was little empirical evidence supporting personal contact as the major step in the transmission of ideas and opinion about public affairs (Gitlin, 1978). This lack of empirical evidence led Robinson (1976) to point out that the hierarchal downward flow of information in the two-step model did not reflect research findings that opinion leaders were often engaging in discussions among themselves. Not only were there horizontal cross flows of information and discussion, in some cases evidence supported a direct one-step communication of information from the mass media to non-discussant audiences. The lack of evidence also led Bennett and Manheim (2006) to conclude that mass media has direct rather than mediated effects on its audiences, a conclusion that reverberates in the fragmented media marketplace of the early twenty-first century. With a plethora of channels for delivering news and political messages outside the control and mediation of the mass media, candidates, parties, officeholders, government agencies, and interest groups are often original, exogenous sources for public affairs information and opinion, communicating directly with members of the general public through live video feeds, Internet connections, and interactive cell phone applications. With the recent emergence of punditry-based public affairs programming, the mass media itself appears in the role of a message mediator rather than a message sender.

Diana Mutz (1998) proposed another theoretical challenge to the two-step process. In *Impersonal Influence*, Mutz drew a dichotomy between the pre-television studies of mass media effects and the larger national scope of the broadcast and cable television era. She rejuvenated a key role for the mass media in the political and public opinion process. Mutz argues that the mass media is the primary source for individuals' perceptions of mass collectives, that is, the polity. In other words, where personal relationships may influence personal attitudes and beliefs far more than mass media, the mass media nonetheless influences sociotropic perceptions of the political environment and informs each individual what others are thinking about.

The arguments for a one-step flow of information from the mass media to audiences re-emerged with changes in communications technology and the resulting expansion of news sources and news interpreters. While television news still follows a traditional model of delivering information and context, live video feeds over the Internet or a cell phone place images of events directly before audiences, often without context or interpretation, and with immediate emotional effect. Moreover, the Internet, social networks, and cell phone communications make it easier than in the past for political actors to bypass the news media entirely and send "blast" or targeted information to individuals and groups.

Yet the modern information environment also provides footing for the two-step model: blogs, text feeds, and social network sites providing a broad-based horizontal flow of information among friends and acquaintances. These mediating discussions take place outside the traditional structured news gathering and deliberative forums of the mass media, and their content is unconstrained by professional journalism standards of confirmation, objectivity, and balance.

Audience-Centered Effects

Another upshot of the Columbia school's electoral research was the development of the theory that mass media produced only minimal effects on citizens. This finding attained wide acceptance after the publication of Joseph Klapper's exhaustive review of communications studies (Klapper, 1960). While assuaging fears that media was not unduly influencing American politics, the persistence of the minimal effects conclusion constrained research on the nexus of politics and the news media into the 1990s. Subsequently, much research was conducted to explain why the news media had so little influence on the political choices of individuals. One answer to that problem was the development of theories of selective exposure which argued that individuals would selectively expose themselves to mass media that agreed with their existing beliefs and predisposition. Klapper used this behavior to explain the minimal effects of communication, just as Berelson et al. (1954) used selective exposure to explain why political campaign communications transmitted through the mass media mainly reinforce and mobilize existing preferences rather than persuade the individual voter. Sears and Freedman (1967) contested these interpretations. In a review of previous studies, they argued that there was some evidence for *de facto* selectivity on the part of the mass media audience, in large part because the general audience comprised an overrepresentation of persons sympathetic to the views being expressed. Put another way, the social environment of an individual is likely to provide reinforcement leading to *de facto* selectivity in choosing news sources. On the other hand, there has been little research support for the hypothesis that individuals seek to avoid conflicting viewpoints and information. However, German experimental researchers (Fischer et al., 2005) found that when limits were placed on the amount of information that could be searched for or when subjects received hints of information scarcity, they became more selective in their exposure to information. The researchers concluded that information scarcity emphasized need for quality information and led subjects to select supporting information.

Selective exposure in campaign settings was examined by Iyengar et al. (2008) with an experimental design during the 2000 U.S. presidential campaign, which found varying support for selective exposure hypotheses. One hypothesis, namely the anticipated agreement hypothesis, suggested that people seek out information about candidates with whom they expect to agree and avoid information about candidates with whom they anticipate disagreement. The second hypothesis called the issue public hypothesis predicts that voters seek information on policy issues of great personal importance. In the presidential election context of 2000, the researchers found that Republicans preferred to access information about Republican candidate George W. Bush, but Democrats failed to prefer information about Democratic candidate Al Gore. However, there was strong evidence for the issue public hypothesis. This

finding ties in with a more pragmatic theory of news selection. Uses and gratification theories suggest that an individual follows particular news stories and political messaging that will prove useful or gratifying to the individual (Garramone, 1985; see also Graber, 2006; Sears and Freedman, 1967).

The impact of mass media on individuals may also be lessened by the lack of interest and attention, or for that matter, the accumulation of political knowledge. Beginning with Converse (1964), scholars have recognized that the inattentive and uninterested comprise a significant portion of the American electorate, a finding that is replicated to a somewhat lesser degrees in election studies in other nations. John Zaller (1992), while constructing a general theory of opinion change and mass communication, offers two additional assertions. First, the likelihood of receiving a political message depends on an individual's general information about politics. People with greater levels of information are more likely to expose themselves to political communication and to understand the messages and information received. The second observation is that people resist political messages that are inconsistent with their political predispositions as long as they have context information sufficient to recognize the inconsistencies. The greater an individual's attentiveness to public affairs coverage in the mass media, the greater the likelihood that the individual will resist political messages. Elisabeth Noelle-Neumann (1974) proposed another effect on public opinion resulting from the mass dissemination of opinion. The spiral of silence theory suggests that individuals who perceive their opinions in the minority will be pressured to remain silent or to express a majority opinion. Noelle-Neumann emphasized that individuals find it more important to avoid isolating themselves socially than they think it is to express oneself. She also recognized a dynamic element in the spiral of silence; if an individual perceives his or her viewpoints not just as dominant, but also to be gaining acceptance, the individual will be willing to express those opinions publicly. Noelle-Neumann argued that when mass media disseminates opinion throughout the society, minority viewpoints are increasingly excluded and the majority opinion acquires increasing attention and legitimacy. The mass media's role in this spiral can be extended a little farther: the mass media disseminates the most common perspectives in order to appeal to consumers and sell its products, legitimizing the majority opinion, and providing an environment in which holders of minority opinions perceive a hostile environment for expressing their views. For more than 20 years following its publication, the spiral of silence was considered one of the most influential theories of public opinion formation, but testing of the theory and its hypotheses has not focused on the dissemination of dominant opinions by the mass media. A review of more than 15 published studies based on the theory provided inconclusive evidence of the effect of majority opinions in survey-based studies (Glynn et al., 1997), prompting a recommendation that experimental designs might provide better tests of the theory.

Media-Centered Political Behavior

Much of the research on politics and mass media reflects a concern for the mass media's roles. Traditional interpretations of the mass media suggest three main roles in public and political affairs: the collection of and dissemination of news and information about public affairs, which may include analysis of that information and the setting of the news in context; the provision of a forum in which to debate public affairs and political choices; and the provision of a watchdog or whistleblower function to alert citizens to the failings of their government and political institutions. Critiques of the former role often function on normative concerns over whether news organizations and reporters can maintain objectivity and balance in the presentation of news, but this concern has faded with the creation of organizations that emphasize niche audience in a limited segment of the political spectrum rather than the general audience. The plethora of news media outlets that emerged with the development of cable and satellite TV providers and Internet news operations has greatly enhanced opportunities for the mass media to serve as a forum for public debate, although that debate is sometimes one sided within a single news entity. (For example, in the United States, both Fox News and MSNBC reflect discernible bias in their presentation of news and commentary. In the United States and other countries, public television channels often attempt to maintain a more balanced alternative to private news operations.)

A larger question is the extent to which mass media determine the content of the public affairs debate. Bernard Cohen's study of the press and U.S. foreign policy (1963) was among the first to recognize that where the press might fail to persuade people about what to think, the mass media did indeed tell its readers what to think about. Cohen noted that editors often believed that their publications were simply serving the preferences of an attentive, expert audience, which limited the scope of their conclusions. McCombs and Shaw (1972) were not so constrained in correlating local and national news coverage with the problems that uncommitted voters around Chapel Hill, N.C., identified as needing attention. Although the study's strong correlations seemed to establish the importance of mass media to establishing the political agenda, subsequent research has weakened the findings, showing that context, personal experiences, levels of political discussion, and television news watching (Iyengar and Kinder, 1987), all impact the ability of the mass media to set political agendas for individuals. But so do self-interest, group membership, and ideology. The agenda-setting function for mass media is enhanced by gate-keeping behavior by editors and news directors, who determine what news is of interest to their subscribers, viewers, and other audiences.

Mass media also exercise a form of personalized agenda setting in elections. Primary elections eliminate the partisan cues that many voters use as judgmental shortcuts (i.e., heuristics) in general elections, hence it falls to the mass media to define the roles for particular offices and cast candidates in those roles (Graber, 2006). The mass media also define the winners and losers of election contests in terms of pre-election expectations (Bartels, 1985). Thus, a winning candidate in a primary election who fails to achieve victory by an expected large margin may be perceived to have 'lost' the contest. Similarly, a second candidate who lost the election but performed better than predicted may be perceived as a 'winner.' Both these approached to news coverage comport with the preference of the news media to

cover dramatic horse-race aspects of elections and image-based events rather than complex and visually unappealing policy debates and performance evaluations that motivate the choices of many voters.

The mass media also act on public perceptions and evaluations of issues and candidates through priming. Priming occurs when the mass media focus on a limited set of issues, policies, and characteristics, establishing a framework of judgmental standards for individuals. If newscasts and news articles focus on national defense issues or a crisis in the national economy, an evaluative framework is provided for viewers and readers (Iyengar and Kinder, 1987). The flip side of the priming coin is the selection made by mass media to frame particular issues in a particular perspective. When small numbers of broadcast news operations dominate national news dissemination, framing by one news entity often coincides with that of the others, as each media operation competes to attract a larger share of the same mass audience. With the emergence of lucrative niche news operations and the explosion of expert blogs, there is no lack of frames from which news consumers may choose.

Of some concern to partisans and researchers alike is whether various media outlets derive a particular slant or bias in their presentation of news. For most of the twentieth century, the most credible news organizations in the United States and around the world were regarded as having sublimated the political preferences of the editors, publishers, and owners in their news columns to the reportorial standards of objectivity, fairness, and balance. While editorial pages framed content on the editorial and opinion pages with the particular interests of editors, publishers and owners, reporters, and news editors were expected to maintain a relatively objective and balanced approach to news despite their individual political preferences. Prior to the emergence in the 1990s of the punditry model of news shows on cable television, most print and broadcast news operations maintained strict separation of the commentary and news functions.

Groseclose and Milyo (2005) recognized this division in generating a measure of news bias that focuses on the reporting of news. Excluding editorial page, opinion page, and external commentary from their study, the researchers found a liberal bias in the news columns of 20 national news organizations, including some that are generally regarded as conservative in their presentations. Media bias was defined as the disconnect between the measured political ideologies of average American voters and the reported news content. These ideological differences were measured comparatively by analyzing the citations of advocacy organizations and think tanks used by reporters and members of Congress. The researchers used standard ideological ratings of legislators to define the ideologies of advocacy organizations and think tanks, then extrapolated the ideology of the sources to generate ideological scores for the news organizations that used them. The central ideology of the voting public was extrapolated as a function of the political ideology of their representatives. Groseclose (2012) later expanded the initial study in a normative book, *Left Turn*, estimating that biased reporting altered the political perceptions of American voters with the effect of adding 8 to 10 percentage points to Democratic vote share in a typical election.

Significant scholarly response to Groseclose's book prompted a symposium discussion of his work in an American Political Science Association journal (Nyhan et al., 2012). Much of the commentary praised the creative attempt to model bias and internal validity of Groseclose's approach, but questioned his foundational assumptions and the failure of the theoretical models and counterfactuals to capture political reality. In addition to these valid critiques, the book failed to consider the substantial differences in the purposes of legislators and journalists (policy advocacy as opposed to informing the public and increasing readership) and a demonstrated profound misunderstanding of how reporters gather and use expert sources on deadline (i.e., there is satisficing behavior at work and sources with effective media response teams will be contacted).

Gentzkow and Shapiro (2010) also examined the concept of media slant among American newspapers and, like Groseclose and Milyo, found evidence of bias in news organizations. But Gentzkow and Shapiro concluded that the viewpoints of the newspapers reflected less of the particular ideological bias of the newspaper owners than an attempt to satisfy what was perceived as the publication's audience. A publication's slant was the result of the desire to promote a reputation of accuracy and sustaining existing readership while drawing in new readers. Thus, slant was more a matter of economic incentive than an effort to promote particular parties or ideologies.

Studies of media bias and coverage in the twenty-first century are based on text analytics, the now automated practice of analyzing the text of news broadcasts and publications to determine what is being discussed, by whom, for whom, and with what evidence. Analyzing text is not new, but what had been a slow process involving graduate student and volunteer coders volunteers is now the province of automated algorithm-based scanning of text, sound bites, and video clips. These algorithm-based programs were initially conceived as a tool for marketing and business intelligence, but their design has involved several top quantitative social scientists, including Harvard's Gary King. In addition to counting hits of words and phrases, the new programs are able to evaluate context and meaning to generate relevant data. The new content-based data may lead to substantial revisions of communications and political theories in the next decades.

Changes in Mass Media Ownership

Since the 1970s, changes in the commercial structures of news organizations and their parent corporations have led to vertically and horizontally integrated media corporations in which unprofitable news operations are tolerated for their contributions to public service and community involvement. Moving from a model of individual or family ownership of mass media to a centralized corporate conglomerate model has subjected many news organizations to corporate demands that they become profit centers, a goal that is often hard to achieve. For most news organizations, advertising dollars accumulated in the entertainment portfolio – not the news and public affairs portfolio – and news operations were often subsidized by income in other divisions of the media corporation. International affairs journalism, traditionally conducted by

newspapers and broadcast news organizations by posting a series of correspondents and stringers around the world, soon faced drastic cuts as rising operating costs of newsprint and nonlocal staff combined with falling local advertising revenues, which were now going to television stations or to Internet advertising. Political advertising was no exception, with expensive broadcast advertising being replaced by cheaper, narrowly targeted Internet and cable advertisements (*see* Political Advertising).

But the late twentieth century also saw the emergence of a new model of profitable media corporation that has the potential to affect political news. Where the standard mass media strategy was to appeal to all elements of the audience in offering political news, Rupert Murdoch's News Corporation and Fox News broadcast and cable news operations pioneered the idea that niche journalism – a punditry-based approach to television news that appealed to a particular right-wing point of view – could be profitable. In the twenty-first century, the Fox News model was proving to be more profitable than news networks that remained focused on the general audience.

Impacts of Technology: Television and Cable News

Foundational theories on the relationships between mass media and politics were laid down in the 1940s and 1950s, but the explosion of communication technologies from the 1980s into the early twenty-first century invites revision of existing theory in a more complex and immediate communications environment. From 1960 on, the emergence of broadcast and cable television, the creation of the Internet, and cell phone and other wireless technologies have prompted reconsideration of theories about how the mass media influences politics.

With the ascendency of television news after 1960 as the main source for many citizens in industrialized democracies, the role of the mass media in political systems changed. No longer were journalists hunter gatherers of news, foraging among party and executive branch leaders, legislators, lobbyists, bureaucrats, campaign consultants, and interest groups. Television news operations became co-producers of political news (Gurevitch et al., 2009). Television moved reporters from observers of the political environment to constructors of a new political reality. The visual nature of the medium also transformed the nature of electoral campaigns from a primary focus on complex and difficult issue debates and incumbent performance evaluations to a simple emphasis on the construction of candidate images. Image and symbolism, always an element of campaign flyers and mailers, achieved an immediacy with the live and video coverage of campaign events and photo opportunities. Many candidates and campaigns, understanding the visual limits of lengthy discussions of policy choices, opted instead for image over substance. If a picture is worth a thousand words, an image is equally valuable because the opposition cannot hold a debate with it. One example of this behavior is the scripting of American political conventions: mobilizing the party faithful and constructing a party platform is no longer the aim. Rather, the convention is constructed as an opportunity to garner free and unanswered media coverage that promotes candidate and party image, including the presentations of sometimes false images of the diversity of the party.

The emergence of broadcast news reduced selective exposure to party propaganda in both parliamentary and presidential democracies, established nonpartisan norms of news presentation, and enlarged the audience for political news (Blumler and Kavanagh, 1999). But a 'third age' of mass media was marked by the increasing popularity of cable television, which avoided the limited broadcast bandwidth problem and allowed for expanded content and channels. Channel multiplication led to audience fragmentation: no longer did the bulk of national populations share a common information stream. Much of the news audience for television in the days of the three major channels was inadvertent, composed of audiences waiting for a particular sitcom or entertainment show (Bennett and Iyengar, 2008). With the multiplication of available channels and the presence of specialized news channels, the inadvertent audience ceased to exist. Rather, individual viewers in the age of cable news were empowered to select news that fit their particular partisan or ideological biases. The creation of an expanded Internet compounded the abundance of political news and the sources from which that news could be gleaned.

The new communications technologies also altered the relationship between parties and candidates. For one thing, the emergence in the United States of the concept of the permanent campaign – i.e., politics is about individuals and parties taking and holding power, not about effective governance of the community – stresses the importance of partisan action and reaction, necessitating that parties and candidates keep a weather eye on the news media and how to obtain constant coverage of party and candidate positions. The legislative or management record of parties and candidates matters less than the record of media-disseminated statements in support of party or ideology. At the same time, the computer and new media technologies have altered traditional roles of the party as providers of communications expertise, voter information, and campaign logistics. This is most prevalent in presidential systems like the United States, where candidates are often chosen in party primaries and may propose platforms that are at odds with the party on certain items. The availability of computerized voter and primary election records allows candidates and officeholders to purchase such information cheaply and directly from the government, and selectively target voters, all without the help of the party apparatus. A well-funded candidate may speak directly to voters before primaries through carefully targeted television, direct mail, and email campaigns, all without the constraint of their party. Parties have little choice but to fall in behind the positions of their leading candidates.

Brave New World: Politics and the Internet

The development of Internet from a network of research institutions to a ubiquitous worldwide web used by citizens around the world prompted even more change for politicians. Citizens gained news opportunities to search for and immediately access news from a vast range of sources, act as citizen watchdogs over the excesses of corporations, governments, and international organizations, participate in political discussions and organize around political interests. But there is a downside: the unregulated diversity of political voices on the Internet

allows for wide dissemination of extremist views and promotes the ability of extremists to link with like-minded individuals (Mutz and Young, 2011). The Internet offered candidates yet another channel of communications, and one that could be used not just to target supporting voters cheaply. The fundraising successes of Howard Dean's, 2004 campaign suggested to candidates and political parties that the Internet could also be a powerful tool for mobilizing small contributors behind candidates, and within 4 years the Obama presidential campaign was campaigning and fundraising with online technologies coordinated with traditional campaign functions. The Obama campaign may be the exemplar for future campaigns in developed democracies, with campaigns utilizing marketing and computer-assisted marketing analysis to identify target audience, followed by a multichannel delivery of messages to the targeted audiences. The need for communications-savvy staff already has led campaigns to professionalize by hiring partisan consultants to direct campaign communications across a multitude of channels. Government agencies and officeholders soon followed the campaigns into upgrading the communications function and hiring a professional communications staff.

The Internet offers several challenges for research on dissemination of political news and commentary. One pertinent question is whether the new media – cable news, talk radio, and blogs – have focused on their niche audiences to the exclusion of alternative viewpoints and balance. Baum and Groeling (2008) found that the new media provided more partisan filtering of the news than news services like Reuters and The Associated Press.

A key question is whether the expanding availability of news sources and political commentary on the Internet undermines the socializing role of the traditional mass media, a question that is linked to selective exposure in a high-choice media environment. Although most Internet public affairs news sites have small audiences compared to traditional print, broadcast, and cable news organizations, many of which now have a presence on the Internet, the focus on these fragmented smaller audiences raises concerns that there is less common perspective on public affairs than in the past (Putnam, 2000, see also Norris, 2001). In the United States, this translates to concern whether the fragmented audience is composed of large numbers of individual who selectively expose themselves to news sources that reinforce their partisan and ideological preferences. Such behavior could give the mass media a role in explaining why the American electorate is becoming more polarized, an assertion that is itself hotly contested.

Much has also been made of the Internet as a new political tool for mobilizing a sense of freedom and democracy within authoritarian states. The Arab Spring uprisings of 2011 provided some grist for the arguments that the Internet provided the technology for breaking the hold of authoritarian regimes, which largely control the national and local mass media as a first priority. In Tunisia and Egypt, the communication costs of collective protests actions was largely reduced by the use of Facebook and Twitter to mobilize protesters and communicate among activists. There is also some evidence that the preference falsification that kept citizens living under authoritarian regimes from realizing the extent of opposition to the regime was undermined by exchanges on social media sites such as Facebook (Lynch, 2011).

Thinking Ahead

Among the normative questions that remain unanswered for all democracies is how well the mass media serve the democratic roles of providing information, forums for discussions, and a watchdog on the excesses of government corruption and secrecy. Among the democracies, the general answer is that the mass media, government, parties, and interest groups provide an imperfect market for ideas on how to govern society. One strategy for improving that performance is to provide enhanced openness and transparency for government records and governmental decision making. Another line of attack by critics of the mass media is to call for new regulations that will at least preserve access to public affairs programming and news in mass media where economic costs or limited bandwidth results in less coverage of public affairs, and may have dire consequences for public support of governmental systems and governments.

Although much has been made of the Americanization of campaigns and political communication through the use of American political advertising and marketing techniques, there is a substantial question of how well the theoretical pillars apply across a developing world comprising diverse political cultures and systems. The overlap of such communication strategies in different cultures is likely to produce a variety of results.

See also: Agenda Setting, Media Effects on; Attitudes, Political and Public Opinion; Communication Research and Media Studies, History of; Mass Media: Introduction and Schools of Thought; Media Effects; New Media and Democracy in the Arab World; Political Advertising; Political Communication; Social Science and the Media.

Bibliography

Bartels, L.M., 1985. Expectations and preferences in presidential nominating campaigns. American Political Science Review 79, 804–815.

Baum, M.A., Groeling, T., 2008. New media and the polarization of American political discourse. Political Communication 25, 345–365.

Bennett, W.L., Iyengar, S., 2008. The new era of minimal effects? The changing foundations of political communication. Journal of Communications 58, 707–731.

Bennett, W.L., Manheim, J.B., 2006. The one-step flow of communication. Annals of the American Academy of Political and Social Science 608, 213–232.

Berelson, B.R., Lazarsfeld, P.F., McPhee, W.N., 1954. Voting: A Study of Opinion Formation in a Presidential Campaign. University of Chicago Press, Chicago.

Blumler, J.G., Kavanagh, D., 1999. The third age of political communication: influences and features. Political Communication 16, 209–231.

Cohen, B.C., 1963. The Press and Foreign Policy. Princeton University Press, Princeton.

Converse, P.E., 1964. The nature of belief systems in mass publics. In: Apter, D. (Ed.), Ideology and Discontent. Free Press, New York.

Cornfield, M., 2005. The Internet and Campaign 2004: a look back at the campaigners. Pew Research Internet Project. http://www.pewinternet.org/files/old-media/Files/Reports/2005/Cornfield_commentary.pdf.pdf (accessed 14.09.11.).

Fischer, P.E., Frey, J.D., Schulz-Hardt, S., 2005. Selective exposure to information: the impact of information limits. European Journal of Social Psychology 35, 469–492.

Garramone, G.M., 1985. Motivation and political information processing: extending the gratifications approach. In: Kraus, S., Perloff, R. (Eds.), Mass Media and Political Thought. Sage, Beverly Hills, CA, pp. 201–222.

Gentzkow, M., Shapiro, J.M., 2010. What drives media slant? Evidence from US daily newpapers. Econometrica 75, 35–71.

Gitlin, T., 1978. Media sociology: the dominant paradigm. Theory and Society 6, 205–253.
Glynn, C.T., Hayes, A.F., Shanahan, J., 1997. Perceived support for one's opinions and willingness to speak out: a meta-analysis of survey studies on the "Spiral of silence." Public Opinion Quarterly 61, 452–463.
Graber, D.A., 2006. Mass Media and American Politics, seventh ed. CQ Press, Washington, DC.
Groseclose, T., 2012. Left Turn: How Liberal Media Bias Distorts the American Mind. St. Martin's Press, New York.
Groseclose, T., Milyo, J., 2005. A measure of media bias. The Quarterly Journal of Economics 120, 1191–1237.
Gurevitch, M., Coleman, S., Blumler, J.G., 2009. Political communication – old and new Media relationships. Annals of the American Academy of Political and Social Science 625, 164–181.
Iyengar, S., Hahn, S., Krosnick, J.A., Walker, J., 2008. Selective exposure to campaign communication: the role of anticipate agreement and issue public membership. Journal of Politics 70, 186–200.
Iyengar, S., Kinder, D., 1987. News that Matters: Television and America Opinion. University of Chicago Press, Chicago.
Katz, E., Lazarsfeld, P.F., 1955. Personal Influence: The Part Played by People in the Flow of Mass Communications. Free Press, New York.
Klapper, J.T., 1960. The Effects of Mass Communications. Free Press, Glencoe, IL.
Lasswell, H.D., 1969. The structure and function of communication in society. In: Schramm, W. (Ed.), Mass Communications. University of Illinois Press, Urbanna.
Lazarsfeld, P., Berelson, B., Gaudet, H., 1944. The People's Choice. Columbia University Press, New York.
Lippman, W., 1922. Public Opinion. Harcourt Brace, New York.
Lynch, M., 2011. After Egypt: the limits and promise of online challenges to the authoritarian Arab state. Perspectives on Politics 9, 301–310.
McCombs, M.D., Shaw, D.E., 1972. The agenda setting function of mass media. Public Opinion Quarterly 36, 176–187.
Mutz, D., 1998. Impersonal Influence: How Perceptions of Mass Collectives Affect Political Attitudes. Cambridge University Press, New York.
Mutz, D., Young, L., 2011. Communication and public opinion: plus ca change? Public Opinion Quarterly 75, 1018–1044.
Norris, Pippa, 2001. Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide. Cambridge University Press, New York.
Nyhan, B., McCarty, N., Gross, J.H., Shalizi, C.R., Gelman, A., Rosenblum, N.L., Jamieson, K.H., 2012. Does the U.S. media have a liberal bias? A discussion of Tim Groseclose's "Left turn: how Liberal Media bias distorts the american mind." Perspectives on Politics 10, 767–785.
Putnam, R.D., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon and Schuster, New York.
Robinson, J.P., 1976. Interpersonal influence in election campaigns: two-step flow hypothesis. Public Opinion Quarterly 40, 304–319.
Sears, D.O., Freedman, J.L., 1967. Selective exposure to information: a critical review. Public Opinion Quarterly 31, 194–213.
Zaller, J.R., 1992. The Nature and Origins of Mass Opinion. Cambridge University Press, New York.

Relevant Websites

http://www.washingtonpost.com/blogs/monkey-cage/ – The Monkey Cage.
http://www.pewresearch.org/ – Pew Research Center.
http://www.people-press.org/ – Pew Research Center for the People & the Press.
http://www.pewinternet.org/ – Pew Research Internet Project.

Mass Media and Sports

David Rowe, Institute for Culture and Society, University of Western Sydney, Penrith, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mass media and sports have become increasingly entwined in modernity, now forming what can be called the media sports cultural complex. While television, the most powerful contemporary medium, has made sports increasingly accessible to audiences, there is scholarly concern that it now dominates and even degrades sports. Political economic analysis has, therefore, been an influential approach in this scholarly field, but it is also important to engage with the dimensions of sports that recognize how it works in diverse ways as culture. Therefore, analysis of media sports meanings, ideologies, and uses is important in an increasingly globalized and digitized environment.

From the vantage point of the twenty-first century, it is increasingly difficult to conceive of professional sports existing independently of the mass media, and of the mass media continuing to operate without a substantial infusion of sports content. Once social and cultural institutions that were not only autonomous but also competitive and mutually suspicious, mass media and sports in modernity have intermeshed so extensively as to now constitute what can be described as the 'media sports cultural complex' (Rowe, 2004). This concept builds on earlier conceptions of institutional and industrial interpenetration of mass media and sports to embrace the much greater, multidirectional flow of mediated sports texts, genres, practices and meanings in local, national, transnational, and global contexts.

There are two enduring, principal problematics addressed by social and behavioral scientists regarding the relationship between mass media and sports. The first derives from significant concern about the power of the media over other institutions, including through the process of mediatization by which disparate sociocultural phenomena are not just represented by the media but changed, and even transformed, by media representation. In the case of sports, which has a strong historical thread of romanticism as an autonomous domain beyond everyday politics and economics, there is specific analytical anxiety that the mass media – especially television, since the second half of the twentieth century – has become dominant in sports and consequently to have degraded it. The second, related principal problematic concerns the mediated politics of sport and, specifically, how the media can, in ideological terms, harness sports to particular, often negative social ends. Thus, media sports are often analyzed and critiqued as complicit in the dissemination and reproduction of social class inequality, sexism, racism, xenophobia, ethnocentrism, homophobia, and so on. A long-standing area of concern is that the media offer a bread and circuses distraction of whole populations, thereby obstructing challenges to prevailing structures of power. Such concerns persist despite major changes to the field of mass media and sports, including technological changes that may be said to have empowered previously passive media sport consumers. In addressing these issues from a social and behavioral sciences perspective, this article begins with a sociohistorical explanation of the convergence of institutions of media and sports, and elaborates on the analytical utility of the concept of the media sports cultural complex.

The Rise of the Media Sports Cultural Complex

The institutions of the mass media and sports developed in parallel as a consequence of modernity's transformative impact on both the reach and speed of communication, and the prominence and organization of physical play. The mass media emerged, in both public (state) and private (commercial) manifestations, as the key institutionalized communicator of political discourse, bearer of news, provider of entertainment, and vector of 'promotional' messages of various kinds. As the commercial mass media developed, so did advertising as a major income stream that sustained the burgeoning media apparatus. For example, the economics of commercial free-to-air television and radio were, from the first, dependent on their capacity to capture large, reliable audiences for exposure to advertised products and services. Broadcast ratings, as a measure of audience size and penetration, could be converted into measurable advertising and promotional value for which the media, as both audience generator and gatekeeper, would be well compensated by commercial entities. Technological innovations, such as more efficient printing presses and satellite signal transmission, created an ever-growing media space where systematically stimulated audience demand required enhanced media supply. The proliferation of media channels made 'premium' content even more valuable, and hence created greater pressure to provide compelling news and popular entertainment in an information and image market that was becoming more complex and competitive. Crucially, sports was well suited to fulfilling this task.

While physical play has been a feature of all known societies for millennia, the specific social institution of sports – with its binding rules, carefully calculated scoring, regular competitive events, industrial provision of space, apparel and equipment, and large-scale operation at national and international levels – is a phenomenon that only began to take full shape in the nineteenth century. Sports emerged as an important collective manifestation of popular leisure

and pleasure (Boyle and Haynes, 2009), developing (first in Britain, which, not coincidentally, was also the pioneering industrial and capitalist power) from intermittent forms of 'folk' physical play on feast days and holidays into codified disciplines in which participants were either lovers of the game (amateurs) or skillful paid-performers entertaining paying spectators (professionals). As sports became more commercialized and, crucially, mediatized, professional sports progressively overshadowed its amateur forms. As the process advanced, the British imprint on sports weakened in the face of the American influence on sports media (especially television), promotion and marketing. It was in the United States that the commercial potential of sports as a mediated phenomenon was most comprehensively grasped.

Sports began to commend itself to the media as an ideal source of content that would draw in large audience cohorts, able to flow across the entire spectrum of news, entertainment, and public affairs in a manner that has now made 'mediasport' pervasive in contemporary culture (Wenner, 1998). The immediacy of sports, involving the unscripted unfolding of events in real time, invited reportage (news) while, at the same time, being captured and packaged by the media (entertainment) as content for display. This content produced more news (in both hard and soft forms) flowing from this content display which, in turn, stimulated more coverage and interest in future-mediated sport content, and so on. It is here, at the intersection of the media's need for content and sports' desire for profile that the genesis of the media sports cultural complex can be found.

It seems curious given the intensive entanglement of mass media and sports that they were initially, mutually cautious in establishing a framework for cooperation. The electronic media, in particular, were initially skeptical that sports could sustain its audience appeal after its novelty value wore off, while various sports organizations were concerned that armchair listening and viewing would reduce paid attendance at enclosed sporting events and so undermine their economic sustainability. As this relationship came to be viewed as more symbiosis than competition, the pivotal bargain struck over broadcast rights – as the single largest element of the sports economy – has both endured and matured. Sports, then, while a major staple subject of the press, radio, and the Internet, has formed its closest and most lucrative alliance to date with television. This article, as a consequence, pays particular attention to the television–sports nexus, which has had a profound impact on the economics, politics, and social dynamics of contemporary culture on an increasingly global scale (Billings, 2011). Sports is not only a prominent genre (or set of genres) of media content, but has much more extensive sociocultural resonance (as discussed later).

In general terms, the two principal forces behind the growth of media sports have been nation building and market development, with the balance and rate of development varying according to social, historical, and spatial context. In Britain, for example, a strong impulse to nurture national culture has been evident through the state-fostered pioneering radio and television representation of great national sports events, like the Grand National and Derby horse races and the soccer FA Cup Final, by the public British Broadcasting Corporation. These were prime instances of the mass media communicating the nation to itself, and seeking to convey a sense of a seamless notion of British (especially English) identity (Whannel, 1992). Contrastingly, in the United States, a much larger, newer nation in which the public organization of broadcasting was subordinated to commercial impulses, the market potential of mass media sports was, in the first instance, more speedily realized. More recently, as mediated sports has become a progressively global phenomenon (Rowe, 2011), sociohistorical differences have produced many local variations in other continents and nations, although still within limits imposed by its formative agents in the Anglophone West. A key reason for sports' rapid global diffusion is its extraordinary capacity to appear in many forms across a variety of national and international contexts, simultaneously reproducing, adapting, reshaping, and extending the media sports cultural complex.

Media Sports Forms, Genres, and Texts

Sports is especially serviceable to the mass media because of its extraordinary adaptability to a range of media formats, including live broadcast action, dramatic still photography, short and lengthy print stories and, increasingly, Web-based exchanges between journalists and fans. Broad developments in media technologies, genres, organizations, and audiences have molded sports' media representation and use, but it has also been influential in shaping the media landscape in nonsports areas. The initial coverage of sports in the mass media – which can be traced back to the early eighteenth century – was print based. In the absence of instantaneous electronic sports news production, the prime purpose of the mass media was to report significant sports events and to record their outcomes in a manner that would attract and sustain substantial readerships. With sports developing as a popular pursuit in the late nineteenth century, sports reports became more elaborate, with detailed descriptions of play and the poetic elaboration of event atmosphere. Prominent organized sports was routinely discussed in the media, and sports events were treated prospectively in stirring anticipation of athletic encounters, which were then minutely analyzed and subjected to comparative quantitative measures of scores and performance patterns. By this means, sports was both advertised in the media and, in drawing audiences to its coverage, used as a vehicle for advertising other products. Sports illustrations, action and portrait photography, and sports-infused advertisements (including those using prominent sports people to endorse products) helped to develop the visual aesthetics of media sports, which emphasized the capture of dramatic action in the frozen moment of the still photograph (later made more vibrant by the replacement of black-and-white by color), as well the noble, perhaps poignant athlete portrait that drew on and reinforced sports' heroic mythology and iconography.

The coming of the electronic media brought unprecedented potential for instantaneous media sports coverage that even accelerated print production schedules could not match. From

the 1920s onward, radio delivered real-time coverage and ambient sounds to growing audiences for soccer, boxing, baseball, cricket, and other sports, creating a new cultural practice of home-based and then mobile (including motor car-based) radio listening to live sports. Descriptive and analytical summary, combined with actual recorded and even synthesized sound, gave some sense of having been there for distant, dispersed but interested sports followers. But it was broadcast television that had the unrivaled capacity to simulate plausibly the experience of physically attending sports contests. Although at its most experientially intense and economically valuable in live broadcast form, sports television could also, through delayed telecasts, replays, highlights, and so on, overcome some of the limitations of temporality by evoking a sense of simultaneity and/or drama long after sports contests had concluded in real time.

As television technology developed, the use of multiple cameras, instant replays, slow (and, later, super slow) motion, split screening, data call-up, and other innovations made the remote-viewing experience in some regards superior to that of actual attendance at the event (leading to the rather paradoxical insertion of large television screens into sports stadia). It is for this reason that live TV is the most prized and valuable form of mass media sports. While orthodox broadcast sports television is currently dominant, live action is increasingly moving onto other screens – notably computers and mobile devices – while broadcast and Web-based technologies are also converging in ways that are transforming how, where, and in what technological–textual combination mediated sports is experienced (Hutchins and Rowe, 2012).

Live TV sports is a perishable form, but its media life can be preserved and extended by acting as a reference point for many other types of programing, including sports magazine shows; previews, *postmortems*, and retrospectives; quiz and chat programs; and documentaries and sport-based dramas supplementing regular sports news bulletins and updates. Confronted by television's capacity to create many different types of sports text, the print media adapted by utilizing the reflective space provided by the written form to expand the sports pages in newspapers (often including large daily sports supplements) and to produce many general (the best known of which is *Sports Illustrated*) and specialist (such as *Baseball Digest* and *Hoop*) sports magazines, all of which are accompanied by sports photography (sometimes in the form of sports photo essays). At the same time, sports organizations and celebrities are covered extensively in general news, not least by means of proliferating sports celebrity scandals and the more routine use of the gossip column. When the many novels and films that deal centrally with sports are included, as well as its key role in the promotion, advertising, and endorsement of products and services from soft drink and leisure wear to life insurance and fitness programs, the degree to which mass media sports has become both ubiquitous and multifunctional can be fully appreciated. Each new media technological innovation has generated novel possibilities for sports, such as multimedia Web sites, sports video games such as *Madden NFL* and *Tiger Woods PGA Tour*, and more physically interactive sports simulation games such as *Wii*.

Indeed, the imprint of sports can be seen in the emphasis on liveness, competition, and partisanship in so-called reality television, and in other genres that rely on uncertainty of outcome and audience (both copresent and distant) involvement. This sportslike quality can be found in popular, though disparate television programs like *Survivor*, *Big Brother*, *The Biggest Loser*, *Beat the Star*, *The Chase*, *The Whole 19 Yards*, *Wipeout*, and *Iron Chef*. Each of these programs, in their different ways, contains elements that can be found in mediated sports contests, thereby suggesting that contemporary popular culture is becoming to a degree 'sportified' by drawing on the habitual (and so largely unconscious) cueing of audience expectations of contest, spectacle and, crucially, a quantifiable result that in most cases determines unambiguous winners and losers. The success of the mass media in insinuating sports into the fabric of everyday sociocultural life – even for those who are hostile or indifferent to it – has raised significant questions about its social and ideological ramifications, and provoked keen debate and substantial scholarship (for example, Raney and Bryant, 2006) about what it means to live in societies where sports has passed, in barely two centuries, from rough play in rural communities to rationalized, highly capitalized, and globalized cultural spectacle.

Analytical Approaches to Mass Media and Sports

Until comparatively recently, social and behavioral scientists have tended to neglect mass media and sports as a suitable area for systematic research and scholarship, preferring to concentrate on the formal politics of the news media or the corporeal dimensions of sports. Where sports was addressed as parts of mediated entertainment, it was usually not given the close analytical attention of, for example, film and television drama and documentary, but treated more generically as one form of commercialized popular culture merging with another. However, as sports became increasingly cultural visible and economically integral to the media, scholars in traditional disciplines (such as sociology, psychology, economics, and philosophy) and new interdisciplinary areas (like media studies, cultural studies, sport studies, and leisure studies) began to engage more intensively with it from the mid-1980s onward. The field has been shaped by three principal concerns: How is media sports produced, what does media sports mean, and how do audiences receive and use media sports? For example, great global and national-mediated festivals of sports, with their reliance on a combination of deep belief and spectacular ritual, began to be seen as a powerful form of secular religion under late modernity or as offering a particularly powerful symbolic narrative of common identities and expressions of difference. However, as noted, a more powerful initial strand in mass media and sports research, drawing in part on the Corinthian ideal of noble amateur sport or on its original status as uncommodified folk physical play, involved a critique that commerce, especially in its embrace of television, has degraded, colonized and corrupted sports, and exploited its audiences.

Although there have been various changes to the mass media–sports relationship over the last 30 years – for example, the increasing involvement of umbrella sports organizations and clubs in producing their own media content – this political economic perspective critiquing corporate media power

remains pivotal in much media sports social and behavioral science. Particular emphasis is placed on the media's capacity to modify sports to make it more suitable for requirements such as advertising breaks and audience retention. Thus, the introduction of tie-breaks in tennis; shorter cricket contests like one-day and Twenty20 matches; longer time-outs and the shot clock in basketball; the designated hitter in baseball; penalty shoot-outs in soccer; more breaks for commercials in gridiron; minimal scrum rules in rugby league; and the speeding up of rucks and mauls in rugby union can be attributed wholly or partially to the drive to make sports more telegenic and/or commercially valuable. The political economic critique of mass media sports has, therefore, emphasized the intensification of the media–sports nexus and the large-scale growth of the media sports cultural complex. The sale of live sports broadcast and other media rights, and the development of global media sports corporations, have focused scholarly attention on the conditions of production of mass media sports.

Live TV sports is able to bring global audiences simultaneously to the screen in massive numbers. For example, the estimated cumulative television audience for the 17 days of the 2008 Beijing Olympics was 4.7 billion viewers, exceeding those of the Athens 2004 (3.9 billion) and Sydney 2000 (3.1 billion) Games. Although such numbers are often exaggerated and misused for promotional and commercial purposes – for example, the claim that 70% of the world's population watched some of the Beijing Olympics does not take into account repeat rather than unique viewing – they nonetheless represent impressive evidence that advertisers and sponsors can find enormous audiences through sports. Thus, they are willing to pay for the privilege to broadcast sports and be conspicuously displayed at mediated sports events. Although television is still the dominant sports medium in the second decade of the twenty-first century, technological development in the media field has produced important audiences involving other platforms and devices. For example, at the 2010 FIFA (Fédération Internationale de Football Association) World Cup in South Africa, over 220 000 people followed the official body's Twitter feed, while its Web site FIFA.com attracted over 250 million visits, including 410 million page views in a single day. As such media sports viewing increases, so does the cost of acquiring its rights, with the US TV broadcast rights for the summer Olympics paid by the National Broadcasting Company (NBC) network rising from US$72 million in 1980 (before the landmark 1984 Los Angeles Games, which in their then-unprecedented commitment to commercialism were characterized as The Hamburger Olympics) to US$894 million for the 2008 Olympics. Here, the professionalization of the Olympics, which once prized its amateurism, is starkly revealed, with its governing body, the International Olympic Committee (IOC), developing and assiduously protecting the global commercial value of its sporting product. Such commodification permeating both organizations and athletes, from a critical political economic perspective, increasingly parallels the culture of commercial media corporations whose aggressive competition for sports rights is matched only by their defense of their acquired intellectual property in the form of exclusive mediated sports.

Media sports political economists, therefore, monitor closely the power of media (and, increasingly following the rise of the Internet, telecommunications companies, and Internet service providers) corporations who bring sports spectacles to a substantial proportion of the world's citizenry. For example, NBC (jointly owned by Comcast and General Electric) is in competition for prime TV sports assets with other corporations such as the 21st Century Fox-owned FOX network (controlled by the world's most powerful and, perhaps, notorious contemporary media sports proprietor, Rupert Murdoch) and the Entertainment and Sports Programing Network, which is owned by the Walt Disney Company and the Hearst Corporation. This competition for mediated sports content goes far beyond mega sports spectacles like the Olympics and the FIFA World Cup. Large local, regional, national, and (with the growth of satellite pay TV and online sports) inter/transnational audiences can be attracted by regular, long-form sports competitions within single countries. Hence, for example, the United States' most powerful sports league, the National Football League (NFL), received US$20.4 billion for free-to-air and pay television rights in its most recent contract, with the value of the 2014–22 broadcast rights substantially higher at almost $40 billion. The not-inconsiderable value of nonbroadcast rights, as well as of the broadcasts of the League's own NFL Network, indicates the contemporary capacity to convert premium sports into high-level media capital.

Sports is crucial to free-to-air television in generating advertising revenue, while for pay television it is the most important content generator of subscriptions, even overshadowing major, first-run films (which lack the immediacy of real-time mediated sports). Of particular concern to political economists of sports is the question of access to mass media sports for the citizenry in general. In many countries (the USA being a signal exception), sports was first shown, and for many years dominated by, public service television as a key aspect of nation building through common cultural ritual. Although commercial free-to-air television subsequently supplemented and, frequently, came to displace and dominate public sports television, its relatively open platform ensured exposure to most sports events of national importance and cultural significance, although this arrangement also required often-resented exposure to advertising messages and hyped up commentary and presentation. When subscription television began to outbid both public and commercial free-to-air broadcasters for sports rights, a barrier to access was immediately introduced on the basis of socio-economic class/income. In countries such as the UK, pay television from the early 1990s onward emerged as the most powerful media force in sports. Indeed, it was exclusive sports rights that were crucial to its establishment in the UK, with the Rupert Murdoch-controlled BSkyB satellite service's securing of the broadcast rights to English Premier League (EPL) soccer in 1992 saving the nascent subscription service, turning a struggling pay television company into the most profitable of all British media inside a decade.

But the siphoning of the best content in Britain's most popular sport from free-to-air television created some public and government resistance, and the analytical interest of media sports scholars, with anxiety about the next step in the direction of commodity logic – the ownership and control of leading sports clubs by the media companies that broadcast

them – rejected there at the advent of the twenty-first century. Although the convergence of sports, business and media interests is common in other parts of the world – for example, ownership of European soccer teams A.C. Milan and Paris St Germain by the (former Italian Prime Minister) Silvio Berlusconi-controlled Mediaset and Canal Plus respectively, and the (now relinquished) ownership of the Los Angeles Dodgers baseball team by (the then) News Corporation – there was uneasiness in Britain about the same vested interest being on both sides of the bargaining table. Thus, a 1999 takeover bid by News Corporation of one of the world's most famous sports brands, Manchester United Football Club, was blocked on grounds of both diminishing competition in the broadcast market and with regard to a wider public interest test of exacerbating inequalities between association football clubs in Britain. The EPL's provision of highly skewed broadcast rights revenues via BSkyB to only 20 out of 92 professional football clubs has created an unprecedented economic divide within British football. This belated social-democratic resistance to the extension of the power of media and of other commercial corporations over sports was recognition by the state of the groundswell of anticoncentration sentiment among grassroots sports fans and their representative bodies (as well as a response to pressure by commercial competitors of BSkyB).

In this instance, the economic power of media capital, the responsibilities of the state, and the civil societal place of sports are simultaneously the concern of the approach to mass media and sports that emphasizes sociopolitical conflicts caused by the intensification of commodity logic within the media sports cultural complex. While this inquiry sprang first within national media sport systems, that framework is increasingly inadequate given the development of a transnational/global mediated sports regime. There is increasing saturation of mature Western media sports markets, leading to the globalization of major sports clubs brands in soccer, basketball, baseball, gridiron, and so on, and especially targeting the fastest growing economies of the populous Asia Pacific region (Rowe, 2011). Here, the image appeal of astutely marketed Western sports often overshadows local sports competitions in a manner that recalls long-standing analyses of media and cultural imperialism. These developments, alongside the growth of mass media and sports research and scholarship in non-Western contexts, have begun to break down the Western dominance of the academic field in a manner mirroring that of the media sports field itself. Social and behavioral sciences work on mass media and sports is required to be less-nationally insular, and the international field is gradually becoming less-fixated on the West, and more open to analytical interventions from places where the media sports landscape is rather different (Bernstein and Blain, 2003).

Despite the unquestionable contribution of the political economic approach to the social analysis of mass media and sports, it is not equipped comprehensively to explain the sociocultural processes by which some forms of mediated sports take hold in particular contexts, while others do not, nor does it usually provide the necessarily detailed interpretive analysis of the meanings and ideologies of mediated sports texts and their interaction with diverse audience cohorts. In short, the ways in which mass media sports works as culture and connects with a range of aspects of sociality needs to be understood through a wider research agenda and other methods. In addressing these questions and shifting focus from the more orthodox economic aspects of the production and consumption of mass media sports, culturalist approaches (incorporating both Cultural Sociology and Cultural Studies) have been influential in attending to media sports as a significant aspect of the fabric and politics of everyday life (Horne, 2006). In seeking to divine the meanings and uses of media sports, the identities it helps to construct and reinforce, and the social fault lines and points of resistance and incorporation that it lays bare, this broad approach has not so much sought to repudiate political economic approaches as to deter abstract, one-dimensional conceptions of media sports markets, producer and consumers.

Culturalist analyses of media sports (through interdisciplinary combinations of elements of ethnography, sociolinguistics, semiotics, and so on) seek to be open to the complexities, contingencies, and uncertainties of common cultural experience. For example, given the burgeoning of global sports spectacles and infrastructures, mediated sports is a highly suggestive test case for globalization theory. The extent to which global sports may be emerging, or the ways in which local or indigenous sports may survive or adapt to the changes precipitated by commercial media sports, are important concerns. The failure of predominantly American sports like baseball and gridiron to become truly global despite heavy promotion by media sports corporations, for instance, or the corresponding slowness of the world game, soccer, to consolidate itself commercially at the highest level in the US despite that country's hosting in 1994 of the massive media spectacle of the World Cup, demonstrate how social and cultural groups do not simply accept – and may even actively resist – the media sports made commercially available to them for their consumption. Important to the dispersion and establishment of sports around the world are the specific histories of the relationships between different countries. For example, international cricket developed in the Indian subcontinent, sub-Saharan Africa, Australasia, and the Caribbean according to the southerly trajectory of the British Empire. Until the end of the twentieth century, cricket was dominated by white former imperial and colonial powers, but the growth of the Indian media sports market has been so marked as to shift the international balance of the game in a decidedly postcolonial direction. In the twenty-first century, India does not only control the International Cricket Council, but has produced a new kind of media sports spectacle in the Indian Premier League that incorporates the visual sumptuousness and high-energy musical routines of Bollywood cinema (Rowe, 2011).

The relationships between media sports texts and audiences are made available for analysis in more open, less culturally pessimistic ways by those scholars who emphasize that media sports texts have multiple meanings (Brookes, 2002), and that audiences can be active and agentic. Instead of seeing sports *aficionados* as inevitably victimized and exploited, then, greater allowance is made in culturalist approaches for how audiences might actively experience and negotiate their encounters with the sports media. For example, the pleasures that women may take in viewing male sports can be viewed from this perspective as a legitimate leisure choice, rather than as an inevitable sign

of patriarchal domination and commercial media domination. Similarly, the a priori assumption that the media sports gaze of fans is inevitably controlled by capitalist interests for their own acquisitive ends can be challenged by examining the ways in which sports fans cooperate to assert their own power or act in ways that are quite distinct from the projections of sports promoters and marketers despite the full armory of media persuasion devices at their disposal.

There is a danger, however, that such discoveries of empowered, active media sports audiences may exaggerate the degrees of freedom of choice of the socially disadvantaged and act as a smoke screen for persistent social inequality. For this reason, it is necessary for continuing social scientific scrutiny both of the density of representations of human subjects in mass media sports (who is most visible) and of their differential social quality (how are they depicted). A striking illustration of this point is the conventional neglect of women's sports in the print and electronic media, such that it remains stubbornly less than 5% of total sports coverage in even the most mature media sports markets, and which is compounded further for women of color, while the sexualization of the bodies of sportswomen is still routine media practice (Scraton and Flintoff, 2002). Hence the role of mass media sports in the stereotyping, marginalization, and disempowerment of subaltern groups can be approached through a judicious combination of quantitative (including content analysis and audience statistics appraisal) and qualitative method (such as ethnographic studies of media sports audiences and textual analysis of media sports imagery). Forms of research and analysis with a cultural emphasis, when informed by a political economic perspective, remain the most promising current lines of inquiry into mass media sports.

These approaches (summarized for the purposes of clarity in **Table 1**) come together productively in the critical interpretive tradition that explores the relationship between major forms of social inequality and media sports representation. A key case in point is the rather anomalous celebration of African-American masculinity in sports like basketball in the context of persistent racism and racial disadvantage in the USA (Carrington, 2010). Others include the continuing ascription of sporting and related capacities, both positive and negative, according to race, ethnicity, or through the racialization of Muslim athletes after the 2001 destruction of the World Trade Center. Another fault line that is available for critical interpretive analysis is the media representation of sexuality in sport (Miller, 2001), including the frequent treatment of athlete homosexuality in the language of scandal. Indeed, the intermittent eruption of mediated sports scandals in general provides rich, varied opportunities for scholars to engage with both the substantive question and the media's (including, but not limited to, the specialist sports media's) handling of it. These pretexts for analysis range from scandals and controversies that are institutionally serious (such as corruption in the IOC and FIFA, or China's treatment of Tibetan human rights in the lead-up to hosting the Olympics) to those that are pruriently voyeuristic (like the infidelities of golfer Tiger Woods) but which, nonetheless, can help stimulate debates over issues concerning gender or the manufacture of celebrity image. That such divergent avenues of inquiry exist attests above all to the sheer scale, complexity, and dynamism of the media sports cultural complex.

Table 1 Comparison of political economic and culturalist analytical approaches to mass media and sports

Political economic	*Culturalist*
Economic ownership and control of sport and media	Symbolic meaning and use of sport and media
Quantitative method	Qualitative method
Social class position	Multiple identities
Centralized model of power	Diffused model of power
Domination and manipulation of media sport audiences	Contestation and negotiation by media sport audiences

New Media, Future Sports, and Research Trajectories

Mass media and sports are now firmly established as a focus for many social and behavioral scientists, whose number grows alongside their subject. The period of relative stability that characterized the previous three decades has ended with the coming of digitization and media convergence, and what can be called 'networked media sport' (Hutchins and Rowe, 2012). As the dominance of broadcast television is eroding with the proliferation of screens and platforms, it is now necessary for researchers, like media sports audiences, to be flexible, agile and dedicated to multitasking. Data collection is becoming more difficult. Recent market-based research has claimed that the Internet is now the preferred media sports platform among 18- to 35-year-olds in Europe. But the widespread practice of polymedia use, including social media, is revealing a more complex media sports attention economy in which singular technologies and platforms operate in different media permutations. Under these circumstances, it is not only difficult to separate sports from their representation (as mediatization takes further hold), but the emergence of 'produsers,' such as citizen journalists, is breaking down the distinction between media sports producers and consumers, as well as between professionals, hobbyists, and amateurs. These trends have prompted increasingly vigorous debates between social scientists over the power of media sports corporations that have been contrastingly adjudged to be in either complete control of media sports consumption or mere service providers subject to the whims of sovereign consumers. Therefore, there is continuing contestation over the connection between mediated sports and cultural citizenship rights as more premium sports television is acquired exclusively by subscription-based platforms at the expense of public and commercial free-to-air audiovisual services. There is also research and analytical exchange over the extent to which media sports producers are being subjected to surveillance so that they can be more effectively tracked and targeted in their consumption behavior, as well as being exploited as unpaid labor that produces much of the additional content of sports Web sites that are owned by private companies, and rely on the generation of traffic for their profitability.

The explosion of the quantity and range of mediated sports, and the inextricable nature of the media–sports nexus, means

that the task of analyzing their content, meanings, and uses, and their corresponding link to sociocultural structures and practices, is commensurately enlarged. This is the case not only when global media sports spectacles carry a profusion of images of nations and bodies (Tomlinson and Young, 2006), but applies to the seemingly mundane news reports that insinuate sports into the daily stock of events selected by the media as worthy of coverage, as well as to the deployment of sports metaphors in advertisements and of sportslike competitive frameworks in very different media genres. Such academic work requires the full spectrum of social and behavioral science research methods ranging from quantitative content analysis to ethnography, from political economic inquiry to discourse analysis. Without this breadth of research, the media sports cultural complex will be perversely inescapable and yet inexplicable.

See also: Audiences, Media; Celebrity; Entertainment; Mass Media and Cultural Identity; Mass Media, Political Economy of; Media Events.

Bibliography

Bernstein, Alina, Blain, Neil (Eds.), 2003. Sport, Media, Culture: Global and Local Dimensions. Frank Cass, London.

Billings, Andrew C. (Ed.), 2011. Sports Media: Transformation, Integration, Consumption. Routledge, New York.

Boyle, Raymond, Haynes, Richard, 2009. Power Play: Sport, the Media & Popular Culture, second ed. Edinburgh University Press, Edinburgh.

Brookes, Rod, 2002. Representing Sport. Arnold, London.

Carrington, Ben, 2010. Race, Sport and Politics: The Sporting Black Diaspora. Sage, London.

Horne, John, 2006. Sport in Consumer Culture. Palgrave Macmillan, Basingstoke, UK.

Hutchins, Brett, Rowe, David (Eds.), 2012. Digital Media Sport: Power and Technology in the Network Society. Routledge, New York.

Miller, Toby, 2001. Sportsex. Temple University Press, Philadelphia.

Raney, A.A., Bryant, J. (Eds.), 2006. Handbook of Sports and Media. Lawrence Erlbaum, Mahwah, NJ.

Rowe, David, 2004. Sport, Culture and the Media: The Unruly Trinity, second ed. Open University Press, Buckingham, UK.

Rowe, David, 2011. Global Media Sport: Flows, Forms and Futures. Bloomsbury Academic, London and New York.

Scraton, Sheila, Flintoff, Anne (Eds.), 2002. Gender and Sport: A Reader. Routledge, London and New York.

Tomlinson, Alan, Young, Christopher (Eds.), 2006. National Identity and Global Sports Events: Culture, Politics, and Spectacle in the Olympics and the Football World Cup. State University of New York Press, New York.

Wenner, Lawrence A. (Ed.), 1998. MediaSport: Cultural Sensibilities and Sport in the Media Age. Routledge, New York.

Whannel, Garry, 1992. Fields in Vision: Television Sport and Cultural Transformation. Routledge, London.

Whannel, Garry, 2001. Media Sport Stars: Masculinities and Moralities. Routledge, London.

Relevant Websites

http://www.la84foundation.org/4sl/over_frmst.htm – LA84 Foundation Sports Library.
http://www.playthegame.org/ – Play the Game.

Mass Media: Introduction and Schools of Thought

Peter Dahlgren, Lund University, Lund, Sweden

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9350–9358, © 2001, Elsevier Ltd.

Abstract

After a brief discussion on the varying nomenclature applied to the field, this article discusses the two basic orientations: academic and applied. From there, the discussion takes up the sprawling character of the field, underscoring that the field can be conceptually organized in a variety of ways. The main part of the discussion consists of a presentation of four major schools of thought. These are defined as: (1) the mainstream perspective, which derives from the postwar US social sciences; (2) the critical tradition, which comprises several different strands; (3) the culturalist approach, which also blends elements from various intellectual sources; and (4) policy horizons, which deals with central issues such as ownership structures, financing and regulation. The article concludes with brief reflections on possible future developments of the field.

The term 'the mass media' refers to institutions that via technologies of largely one-way communication, reach audiences that consist of large numbers of people. The book can be seen as the first mass medium and consequently it could be argued that mass media emerged with the development of printing as a medium at the end of the fifteenth century. However, the concept of the mass media is normally associated with the development of modern, industrial society in the West, beginning in the early nineteenth century. In this period, newspapers joined the book as an important mass medium; later in the century, the telegraph became an important point-to-point media technology that had important consequences for the mass media via its conjunction with the press. In the twentieth century, film, radio, phonograms, and then television and video were added. The growth of the media has of course always been conditioned by prevailing historical circumstances (Winston, 1998) and how they are to be best used remains at the core of ongoing media policy issues. The spread of the mass media has proceeded in an accelerating manner and today they constituent a central feature of modern society. From the micro-level of individuals in everyday life to the macro-level of societal institutions and processes, including the global setting, modernity is inexorably intertwined with the development of the mass media (Thompson, 1995).

The Field

While research about the mass media can be traced to the early decades of the twentieth century, mass media studies first arose as an academic field during the 1950s, establishing itself institutionally in the following decade, first in the US and then elsewhere. Emerging from several disciplines in the social sciences and humanities, the field launched its own academic departments, journals, associations, and conferences. In the US today, most colleges and universities have departments offering courses and programs about the mass media, a pattern that is increasingly manifested in other countries as well. The field today is broad and eclectic, reflecting both the influences of its earlier intellectual origins, as well as the newer developments in the media's evolution and changing socio-cultural significance.

Mass media studies manifest a heterogeneous collection of theories, methods, and general intellectual profiles (see McQuail, 1994 for a useful overview; Boyd-Barrett and Newbold (1995) provide an extensive collection of key texts in the field). Academic departments go under a variety of names, emphasizing different angles on the field, e.g., in the US, many departments have their origins in the older schools of journalism, others developed out of speech communication and rhetoric, others from sociology or literature, and so forth. Programs in journalism studies and training are often coordinated with more general mass media studies, while film history and film studies more generally often tend to maintain a separate institutional identity. Particularly in the UK, the field of mass media at universities is often linked with cultural studies.

Nomenclature

The labeling of the field varies to some extent. In some academic departments, the dominant thematic is 'communication,' and mass media studies, treated as 'mass communication' may be juxtaposed to such areas as interpersonal communication, organizational communication, and intercultural communication. In the last decade or so, particularly in the light of newer developments in digital media technology that make possible more interactive modes of communication, there is a tendency to drop the adjective 'mass' and simply refer to media studies. This inclination is also prompted by changes in social theory: the term 'mass media' is seen as echoing older theories of 'mass society,' a perspective that today appears out of step with the complexities of contemporary social realities.

Two Basic Orientations

As a field of teaching and research, mass media studies tends to encompass two basic orientations. On the one hand is academic research and a critical liberal arts perspective on the media and their relationship to society and culture, often aimed at answering scientific questions about their institutional operations, their output, and their audiences. On the other hand one finds an emphasis on applied knowledge. This can be geared to the many operational questions facing media

industries, to various organizational actors that make use of the media for their own purposes (corporations, political parties, interest groups, etc.), and to occupational practices within the media. Such occupational practices include production skills within the different media, as well as professional training, in particular for journalism. Other occupational orientations include public relations, advertising, and media management. The lines between the academic and the applied orientations are by no means fixed; the interplay between the two can be both productive as well as problematic. Outside the academy there is considerable media research done from commercial horizons, for example market research, audience measurement, and studies on the effects of advertising. Also, political actors have increasingly been making use of media consultants for strategic purposes.

Organizing Logics

While mass media studies is a rather sprawling field, there are several organizing logics that give it some structure and coherence. These logics are not mutually exclusive, nor do they have equivalent status. However, they can serve as partial maps of the research terrain. The first such logic derives simply from specifying a particular medium. Many people in the field will identify themselves for example as researchers of television, radio, or the press. Others will specialize still further and focus, for example, on television history or the television industry. Many research studies, however, may encompass more than one medium, for example in a case of the analysis of news coverage.

A second logic is to emphasize the different elements of the communication process: a basic model of communication will often use the elements 'sender,' 'message,' and 'receiver.' Translated into the context of the mass media, this yields, respectively, studies of media institutions, media output, and media audiences. Research on media institutions examines the conditions, both internal and external, that shape the way they function. This can encompass for instance questions about regulation, political influence, and organizational routines. The study of media output includes the form, content, and modes of representation. Often specific genres, particularly television genres such as soap operas and talk shows, become the object of sustained research attention. Studies of media audiences can include analyses of the impact of specific output, the uses which people make of it, and the interpretations they derive from it.

Third, specific topics or themes provide ongoing structure to the field. These themes can be broad, such as media effects, or (more specifically) media effects on children. Many themes link media studies with particular topics on the wider social research agenda, such as gender and ethnicity; health communication, media and social movements, and media and sports, are also examples of such specialization. Some themes derive instead from particular theoretical and/or methodological developments, for example, agenda setting, media events, or media uses and gratification, and can be applied to a variety of empirical areas.

A fourth organizing logic derives from the still-remaining links between mass media and the 'parent' disciplines; such links serve to focus research (and connect researchers) from particular disciplinary perspectives. For example, political communication is a subfield that interfaces with political science, often incorporating public opinion studies; the psychology of communication examines cognitive processes in relation to mass media. Often, however, within the overall field of mass media studies, intellectual lineage is not made explicit, given that the field has developed a distinct identity over the course of several decades.

Finally, weaving through the above logics are sets of intellectual traditions that very loosely structure the field into schools of thought (see below). It should be emphasized that 'school' is used loosely here; it points to tendencies within the field, not unified factions with clear demarcations. These schools basically reflect differing theoretical and methodological orientations; they point to general intellectual dispositions. The relationships among them have evolved over time, and the development of the field can be understood to a considerable extent by looking at the emergence, interface, and transitions of these schools of thought.

While the exact labeling may be differ in historical accounts of the field, the following constitute the main schools of thought within mass media research: the mainstream perspective, the critical tradition, the culturalist approach, and the policy horizon.

The Mainstream Perspective

The systematic study of the mass media arose in the early decades of the twentieth century, and the mainstream perspective today can be understood as the intellectual descendants of the pre-World War II pioneers. While the mainstream perspective encompasses a wide array of research orientations and is continually developing new approaches, it is united in its basic adherence to traditional social scientific methodologies and approaches to theory. Its intellectual underpinnings are found chiefly within sociology, psychology, and social psychology, and it constitutes the still-dominant tradition of empirical research in mass communication.

Origins

Its origins can be found in nineteenth century sociological conceptions about the crowd and the mass; early research efforts and attempts to develop theory were geared toward among other things the theme of propaganda, with the experiences of World War I as a point of reference. In the 1920s, the Chicago School of sociology, with its German-trained prominent member Robert E. Park, developed a strongly qualitative, ethnographic micro-sociology of mass media uses and impact, within a generally progressive and reformist social vision. Walter Lippman's (1922) theories about public opinion were an important element in conceptually linking the mass media with the processes of democracy.

The political scientist Harold Lasswell, in the 1930s under President Roosevelt's New Deal, saw new opportunities for studying public opinion and propaganda. His formula of 'who says what to whom through what channel and with what effects' served as a framework for media studies for many years. In psychology, Kurt Lewin's psychology of personality and of group

dynamics foreshadowed analyses of the social processes of media influence. After the war, Carl Hovland and his associates (1953) further developed behaviorist perspectives on persuasion.

Interest in the media grew in step with their expanding role. The imperatives of the wars, the importance of the media in political processes, concern about the media's impact on psychology, emotions, and behavior, as well as increasing commercial concerns of the media themselves stimulated research. After World War II, functionalist sociological theory had become the dominant paradigm for fledgling media research, backed up by developments in methodological techniques. Quantitative content analysis, aimed at mapping the manifest dimensions of media messages, became the norm for studying output. Studies of audiences' media preferences, and, most importantly, effects of the media on audiences, began to blossom. Surveys became an established approach; later the strategy was developed to repeat surveys with the same respondents to yield panel studies over time (see Rosengren and Windahl, 1989).

A key person in the USA at this time was the Austrian-born Paul Lazarsfeld, who was active in ushering in a new phase of mass-media research. Lazarsfeld began working for the commercial media in the 1930s, and he viewed the media research tradition he helped to develop as 'administrative' research. This was geared to serve the strategic purposes of media institutions. Gradually, a more society-wide perspective emerged, that helped shape a media-research orientation that was in harmony with the consensus assumptions about society. Together with Elihu Katz (Katz and Lazarsfeld 1955), Lazarsfeld helped introduce the notion of how the impact of the media is filtered through social interaction with others, giving rise to the concept of the two-step flow of communication. Wilbur Schramm (1970), in the 1950s and 1960s, did much to synthesize the diverse elements and promote the identity of mass media research as a distinct field of inquiry.

The Golden Age: Effects Research and Content Analysis

From about 1950 to the early 1970s, the mainstream perspective emerged, solidified, and remained essentially unchallenged. This was its golden age; since then the heterogeneity of the field has increased. During the 1950s, concern with the effects of the media continued to grow in the USA, especially in the light of television's reach into just about every home. The questions around media effects were not merely scientific: there was popular concern about violence and the media and about portrayals of sexuality, and, later on, about pornography. The media industries landed in a seemingly paradoxical position of trying on the one hand to soothe popular opinion by asserting that such negative effects were minimal, while on the other hand assuring the advertising industry of the effectiveness of media messages. There was also a malaise about 'media manipulation' in the realm of politics, as the question of just how much the media shape attitudes and opinions took on increasing concern. A landmark book in this context was Joseph Klapper's (1960) *The Effects of Mass Communication*, which argued for a 'minimal effects' position.

Within the media effects research tradition, the nature of the 'effects' investigated can vary considerably: on beliefs, attitudes, knowledge, or behavior; intentional or unintentional; long-term or short-term; and so forth. In the USA, the Surgeon General's massive, multivolume report (1972) was perceived to be inconclusive by many, but what it did emphasize was the importance of so-called intervening variables. That is to say, the effects of the mass media must be understood as being qualified by such factors as age, gender, education, social circumstances, audience contexts, and so forth. Generally the drift in media effects research has been away from the earlier psychologism toward increasing socio-cultural complexity. Concern with the effects of mass media continues today, even if the task of identifying and isolating specific media output and linking it to particular effects becomes all the more difficult with the increasing mediatization of society.

The mainstream perspective has been no less interested in the nature of media output. Journalistic output is one of the major areas in which content analysis has been applied, where bias in reporting has been studied by measuring, for example, the visibility of opposing candidates during election campaigns, and coding degrees of positive and negative valence. More encompassing studies have been done where researchers have attempted to elucidate important dimensions of journalistic coverage or even in fictional portrayals that are normal or routine, yet may remain invisible precisely because of their everyday, taken-for-granted character. Measurements have been developed, for example, for tallying acts of violence in television programs. The ambitious Cultural Indicators project of George Gerbner (1973) and his colleagues sought to illuminate the symbolic world of TV fiction in the USA, mapping not only violence, but the more general norms and values manifested in the programs.

Within the tradition of so-called agenda-setting research (see Dearing and Rogers, 1996 for an overview), news and current affairs in the mass media are studied and then compared to audiences' understandings of, for instance, what are the major issues of the day. The comparison allows researchers to draw conclusions about the extent to which the media set the political agenda; generally the findings are that the media do have this initiative, but it has been found that audiences, or rather the public, do not always simply mirror the concerns of media coverage.

Tensions in the Field

The mainstream perspective continues to develop, partly in response to new phenomena in the media, e.g., we witness today considerable attention to the transnational character of the media and a growing research concern with political advertising (Bennett, 1996). However, innovation in theory also generates renewal. Thus, in the 1970s the uses and gratifications tradition (Blumler and Katz, 1974) arose and posed a set of complementary – if not fully competitive – questions to the effects tradition. Basically, instead of asking what the media do to people, the newer orientation asked what people do with the media and what satisfactions do they thus derive. This shift points to an ongoing key tension within the field, namely the question of the relative power of the media and the degree of autonomy of the audience. The view of minimal effects, in keeping with the intellectual origins of functionalism, liberal democracy, and the needs of commercial media had strong support but did not go unchallenged. Indeed, critics began

claiming that the theories and methods embedded in the mainstream perspective structured a status-quo perspective toward media and society (Gitlin, 1978). In 1983 and again in 1993, the leading US journal in the field, *The Journal of Communication*, had themed issues on the tensions within the field (Gerbner, 1983, Levy and Gurevitch, 1993).

It should also be noted that the prevailing conception of the process of communication at work within the mainstream perspective, operating for the most part implicitly, was that of the transfer of messages or, more generally, of information. This view has its origins in, among other things, Shannon and Weaver's (1949) cybernetic view of communication. Gradually this basic model was augmented by contributions from cognitive psychology, giving rise to more elaborate notions of human information processing. Yet, other voices, with an anchoring in the humanities and cultural theory, were challenging such views, arguing for the centrality of meaning for the process of communication. Both the critical tradition and then the culturalist approach began to coalesce and to offer competing paradigms of mass media research.

The Critical Tradition

During the late 1960s, critical traditions from various strands of neo-Marxism were beginning to make themselves felt within the social sciences, and by the early 1970s they were manifesting themselves in mass media studies. There are several intertwining intellectual strands here: political economy, the critique of ideology and public sphere theory.

Political Economy

Though the political economy of mass media need not a *priori* be based in the Marxian tradition, within media studies the term does most often connote this lineage rather than, say, neoclassical political economy or conventional economics. The emphasis in Marxian political economy of the media is often on the links between economics and the social, political, and cultural dimensions of modern life; pluralist or consensus models of society are generally rejected as inaccurate. A recurring thematic is the tension between on the one hand, the capitalist logic of media development and operations, and on the other, concerns for the public interest and democracy. The political economy of the media does not anticipate the elimination of commercial imperatives or market forces, but rather seeks to promote an understanding of where and how regulatory initiatives can establish optimal balances between private interest and the public good. The work of the Canadian scholar Dallas Smythe (1977) was an important element in the development of this tradition; among contemporary researchers Nicholas Garnham (1990) has further expended the perspective, which has in turn been systematized in Mosco (1996). Within this perspective, the emphasis is most often on the institutional arrangements of the media: ownership, control and regulation, and their consequences. The political economy of media structures at the global level has increasingly come under research scrutiny (e.g., Herman and McChesney, 1996).

A variant of the political economy tradition has focused on media imperialism, in particular how the combination of US corporate and political interests over the years have made use of the media abroad and the consequences that this has had for people in other countries, particularly the developing nations. The work of Herbert Schiller (1976) has been central here. Media flows from the West to 'the Rest' were plotted (Nordenstreng and Varis 1974), including fiction, other entertainment, journalism, and also the role of international advertising (Mattelart et al., 1984). Unesco became a political forum and research promoter for a New International World Information Order (1980). Such research was framed by historical circumstance: the bipolar East-West situation of the Cold War shaped US media strategies abroad in ways that had impact on the North-South global divide. Since the collapse of communism, the North-South socio-economic gap has continued to grow, and though the transnationalization of the mass media is now often framed analytically within paradigms of globalization, critical attention is still addressed to global media disparities.

The Critique of Ideology

Another strand of critical endeavor attends in more detailed ways to media representations, analyzing them from the standpoint of their ideological dimensions. The intellectual roots for the critique of ideology derive not only from Marx, but also from the structural neo-Marxism of Althusser, as well as the more culturally oriented versions of the Frankfurt School's Critical Theory and of Gramsci. Methodologies and epistemologies were at times in conflict, not only between the mainstream perspective and the critical tradition, but also within the critical tradition. Versions both of cultural Marxism and of structural Marxism – the latter making use of Freudian theory adapted via Lacan – became incorporated into the development of cultural studies (see below). A landmark effort in the critique of ideology in media studies was the work of the Glasgow Media Group (1976), which made explicit use of traditional social science methodologies of content analysis to make its points about the class bias of the British Broadcasting Corporation's journalism.

Ideology as such is a slippery concept, and the earlier assumptions about false consciousness could not hold up to critical scrutiny and increasingly sophisticated epistemology. Gradually, the critique of ideology began to make use of the hermeneutic tradition, treating ideology as an aspect of the interpretation of meaning in the reproduction of relations of social domination. The critique of ideology began to blend with cultural theory, signaling a growing entwinement of the critical and culturalist schools. Efforts to analyze cultural hegemony, in both its domestic and transnational manifestations, strove to relate media representations and the political economy of the media to specific social circumstances. Also, the growing presence of feminism within the critical tradition began to render the exclusive concern with social class as the foundation of ideology untenable: ideology could manifest gendered social relation, not least in the media. Gradually, feminist perspectives began to take hold in media research (van Zoonen 1994). Also, race and ethnicity also became foundations for the critique of ideology in the mass media (Gandy 1998).

Public Sphere Theory

Discussions of democracy and the media are increasingly framed within the concept of the public sphere, a concept associated with Jurgen Habermas (1989, see also Calhoun 1992). In schematic terms, a public sphere is understood as a constellation of institutional spaces that permit the circulation of information and ideas. These spaces, in which the mass media figure prominently, also serve to foster the development and expression of political views among citizens as well as to facilitate communicative links between citizens and the power holders of society. Beginning in the 1970s, media researchers in Europe and later, the USA, began using the concept to pursue critical analyses of mass media, examining the factors that impede the public sphere. The concept has many parallels with the liberal notion of the 'marketplace of ideas' and similar metaphors, and today it has entered into more mainstream usage, where the problems of journalism are discussed and researched; its Frankfurt School origins often remain in the background.

The Culturalist Approach

Within mass media research, contributions from the humanities and from qualitative, interpretive social science had long remained in the shadow of the dominant social scientific methods. Studies based in literature, history, rhetoric, or anthropology were very much the exception. Marshall McLuhan's (1964) work, with a historical view of media technology, was an exception, but it had more impact outside the university than within academic media research.

Cultural Studies

In the 1960s British Cultural Studies was in its formative years, and in the following decade, under the leadership of Stuart Hall (Hall et al., 1980) had begun to impact on media research (Turner 1996). The eclectic synthesis of intellectual ingredients in cultural studies – including neo-Marxism, structuralism, feminism, psychoanalysis, semiotics, and ethnography – was applied not only to the media but a range of topics having to do with contemporary culture and society. In particular, popular culture has been an ever-present topic of research and debate. Cultural studies has among other things shifted the focus of issues around popular culture (and/versus high culture) from earlier largely aesthetic concerns to questions about social relations and power; taste becomes socially contextualized. There has also been a 'rehabilitation' of popular culture, where 'the popular' is seen as no longer so clearly demarcated from high culture or, alternatively, is defined in a manner to give it political relevance (Fiske, 1989). Other research takes the ways in which the media, via advertising and general lifestyle imagery, are central to contemporary consumer culture, not least on the global level (Featherstone, 1995).

Cultural studies has grown into a heterogeneous, multidisciplinary field in its own right, with contributions from currents such as postmodernism and postcolonialism, among others. Today studies of the media are only a small part of its vast concerns. Cultural studies can be seen as part of a larger 'cultural turn' within the human sciences in recent decades, but today it is clearly the dominant culturalist tendency, even if there is a lack of consensus about both its core and its boundaries. As cultural studies has expanded and become a global academic phenomenon mingling with many other academic traditions, the critical character of its earlier years, where issues of social and semiotic power were thematized, has not always remained evident. For example, the emphasis on the multiplicity of interpretations of popular culture – at times augmented by postmodern perspectives – has tended to undercut arguments about ideology.

The Centrality of Meaning

In culturalist mass media studies two themes from the humanities and qualitatively oriented social sciences have come to assume central positions: sense-making and social constructionism. Research has come to deal with questions of how media representations convey meaning, how media audiences make sense of these representations, and how media-derived meanings further circulate in society. Hall's (1980) model of how meaning is encoded by the media and decoded by audiences became an important organizing concept for much of such work. Studies of reception – how audiences interpret what they encounter in the media – became an important subspecialty within mass media research; the work of Dave Morley (1992) and Ien Ang (1985) are significant contributions in this tradition. Generally, as the media landscape becomes more multidimensional and theorizing about the audience becomes more ambitious, the frontlines of audience research manifest both a convergence of different research approaches and a growing complexity. (e.g., Hay et al.,1996)

If semiotics and hermeneutics were the major text-analytic tradition within cultural studies, today they have become augmented and partly replaced by innovations from various other currents such as linguistics and cognitive psychology. These newer efforts have coalesced to generate several versions of what is now termed discourse analysis (van Dijk, 1997) that is increasingly used in the elucidation of meaning in media texts.

Active Subjects

In culturalist approaches to the mass media, the emphasis on the production of meaning and social constructionism tends to privilege versions of the active subject. This perspective is manifest not least in a strand of research that has underscored the audience's sense-making, though debates have emerged (again) as to the degree of semiotic power that should be accorded the audience relative to the media (Curran, 1990). Research in some quarters has linked audiences' interpretation of the media as forms of cultural resistance, and tried to delineate interpretations that run counter to hegemonic trends; Radway's (1984) study of women's uses of romance novels for a variety of purposes is a compelling example. More generally, theories of subjectivity have been mobilized to study how individuals construct identities via their active appropriation of media; feminist contributions have been particularly visible here. From the standpoint of collectivities,

mass media and cultural identity has become a growing domain of research, not least in transnational contexts (Morley and Robins, 1995).

The Policy Horizon

It could be argued that the policy horizon is not a school of thought, but rather a thematic organizing logic encompassing diverse perspectives. However, the centrality of policy in the shaping of media development and the distinctiveness of its concerns speak for its status as a school.

Media Systems and Norms

Situated in the force fields between economics, politics, and technology, the way the media are owned, organized, and financed, and the way that they operate have much to do with the vicissitudes of the policy process. Policy is of course shaped by the specific interests and actors involved, such as the state, commercial media institutions, the advertising industry, media production organizations, citizens groups, and other representatives of the public interest. One line of policy research has addressed the historically anchored traditions in political culture and in philosophies of the public good that buttress prevailing normative frameworks in regard to the media. These frameworks are generally accepted but often given competing interpretations. Freedom of the press, for example, has had a particularly strong status in Western liberal democracies, though how it is to be best applied in a concrete case may be contested. Historically, and today globally, there are competing models for media systems, including authoritarian, communist, and social responsibility versions that build upon differing sets of norms.

Regulation and Deregulation

Another line of research has been more engaged at the frontlines of ongoing policy issues and battles, and often seeks explicitly to intervene in the policy process. The growing concentration and conglomerization of media ownership has been charted by liberal critics in the USA (Bagdikian, 1997), and elsewhere; generally much of the political economy perspective has fed into policy research. Over the past two decades, the increasing commercialization of media industries, coupled with a political climate shaped by neoliberalism, has contributed to a massive deregulation of the media, giving market forces ever greater free play and repositioning audiences less as citizens or as a public and more as consumers. This has been especially true in the realm of broadcasting regulation, and in Western Europe public broadcasting has undergone a radical realignment as it has lost its monopoly position (Tracey, 1998). For over two decades the Euromedia Research Group (1998) has studied these developments. Confronted by commercial channels borne first by cable and satellite systems, and then terrestrial stations, public broadcasters in Europe and elsewhere find that their *raison d'etre* has to be redefined as they look to the future. Public broadcasting emerged as a part of the project of the nation-state, and its current circumstances can be seen as part of the dilemmas facing the nation-state in the era of globalization. At the transnational level, the regulation of international communication becomes all the more important – and complex.

The global media landscape today is dominated by fewer than two dozen conglomerate giants who increasingly control the production of media output and the distribution process, as well as the technical infrastructure. Technical advancements promote a convergence between the technologies of broadcasting, telecommunications, and the computer. Institutionally, major companies in the computer and Internet field have begun to merge with the giants of the traditional mass media. This evokes profound policy issues of diversity, accountability, and democracy (McChesney, 1999).

Looking Ahead

The media are obviously in transition, and research is in the process of adapting itself to the emerging realities. The increasing on-line presence of the traditional media, as well as other forms of functional integration technical convergence, suggest that the field of mass media research will in particular increasingly take on a digital focus. There will be growing attention to such areas as the information society, computers and society, human-computer interface, and electronic democracy. Digitalization has prompted a flurry of research from many disciplines, including social theory (Castells, 2000) and thus it can be argued that the field of mass media research will increasingly blend with initiatives from other directions. This may involve some erosion of its status as a separate field, but the enhanced attention being paid to the media will no doubt increase our knowledge and understanding of them.

See also: Agenda Setting, Media Effects On; Broadcasting: Regulation; Communication, Twostep Flow of; Mass Media and Cultural Identity; Mass Media, Political Economy of; Mass Media, Representations In; Media Effects on Children; Media Effects; Media Events; Media and History; Printing as a Medium; Radio as a Medium; Television: General; Television: History; Violence and Media.

Bibliography

Ang, I., 1985. Watching 'Dallas'. Methuen, London.
Bagdikian, B., 1997. The Media Monopoly. Beacon Press, Boston.
Bennett, L., 1996. The Governing Crisis: Media, Money, and Marketing in American Elections, second ed. St. Martins, New York.
Blumler, J.G., Katz, E. (Eds.), 1974. The Uses of Mass Communication. Sage, Beverly Hills, CA.
Boyd-Barrett, O., Newbold, C. (Eds.), 1995. Approaches to Media. Arnold, London.
Calhoun, C. (Ed.), 1992. Habermas and the Public Sphere. MIT Press, Cambridge, MA.
Castells, M., 2000. The Rise of the Network Society, second ed. Blackwell, London.
Curran, J., 1990. The new revisionism in mass communication research: A reappraisal. European Journal of Communication 5, 135–164.
Dearing, J.W., Rogers, E., 1996. Agenda-setting. Sage, London.
Euromedia Research Group, 1998. Media Policy: Convergence, Concentration and Commerce. Sage, London.
Featherstone, M., 1995. Undoing Culture: Globalization, Postmodernity and Identity. Sage, London.
Fiske, J., 1989. Understanding Popular Culture. Unwin, London.
Gandy, O., 1998. Communication and Race. Arnold, London.
Garnham, N., 1990. Capitalism and Communication. Sage, London.

Gerbner, G., 1973. Cultural indicators – the third voice. In: Gerbner, G., Gross, L., Melody, W. (Eds.), Communications Technology and Social Policy. Wiley, New York, pp. 553–573.
Gerbner, G. (Ed.), 1983. Ferment in the field. Journal of Communication 33.
Gitlin, T., 1978. Media sociology: the dominant paradigm. Theory and Society 6, 205–253.
Glasgow Media Group, 1976. Bad News. Routledge and Kegan Paul, London.
Habermas, J., 1989. The Structural Transformation of the Public Sphere. Polity Press, Cambridge, UK.
Hall, S., 1980. Encoding and decoding in the television discourse. In: Hall, et al. (Eds.), Culture, Media Language. Hutchinson, London, pp. 197–208.
Hall, S., et al. (Eds.), 1980. Culture, Media Language. Hutchinson, London.
Hay, J., Grossberg, L., Wartella, E. (Eds.), 1996. The Audience and its Landscape. Westview Press, Boulder, CO.
Herman, E., McChesney, R., 1996. The Global Media. Cassell, London.
Hovland, C., Janis, I., Kelly, H., 1953. Communication and Persuasion. Yale University Press, New Haven, CT.
Katz, E., Lazarsfeld, P.F., 1955. The People's Choice. Free Press, New York.
Klapper, J., 1960. The Effects of Mass Communication. Free Press, New York.
Levy, M., Gurevitch, M., 1983. The future of the field. Journal of Communication 43.
Lippman, W., 1922. Public Opinion. Harcourt Brace, New York.
MacBride, S., 1989. Many Voices, One World. UNESCO, Paris.
Mattelart, A., Mattelart, M., Delcourt, X., 1984. International Image Markets. Comedia, London.
McChesney, R., 2000. Rich Media, Poor Democracy: Communication Politics in Dubious Times. University of Illinois Press, Urbana, IL.
McLuhan, M., 1964. Understanding Media. McGraw-Hill, New York.
McQuail, D., 1994. Mass Communication Theory: An Introduction. Sage, London.
Morley, D., 1992. Television Audiences and Cultural Studies. Routledge, London.
Morley, D., Robins, K., 1995. Spaces of Identity: Global Media, Electronic Landscapes and Cultural Boundaries. Routledge, London.
Mosco, V., 1996. The Political Economy of Communication. Sage, London.
Nordenstren, K., Varis, T., 1974. Television Traffic – A One-way Street? UNESCO, Paris.
Radway, J., 1984. Reading the Romance. University of North Carolina Press, Chapel Hill, NC.
Rosengren, K.E., Windahl, S., 1989. Media Matter. Ablex, Norwood, NJ.
Schiller, H., 1976. Communication and Cultural Domination. Sharpe, New York.
Schramm, W. (Ed.), 1970. The Process and Effects of Mass Communication. University of Illinois Press, Urbana, IL.
Shannon, C., Weaver, W. (Eds.), 1949. The Mathematical Theory of Communication. University of Illinois Press, Urbana, IL.
Smythe, D., 1977. Communications: Blindspot of western Marxism. Canadian Journal of Political and Social Theory 1, 120–127.
Surgeon General's Scientific Advisory Committee, 1972. Television and Growing Up: The Impact of Televised Violence. GPO, Washington, DC.
Thompson, J., 1995. The Media and Modernity. Polity Press, Cambridge, UK.
Tracey, M., 1998. The Decline and Fall of Public Service Broadcasting. Oxford University Press, New York.
Turner, G., 1996. British Cultural Studies, second ed. Routledge, London.
van Dijk, T., 1997. Discourse Studies: A Multidisciplinary Study. Sage, London.
van Zoonen, L., 1994. Feminist Media Theory. Sage, London.
Winston, B., 1998. Media, Technology and Society: A History from the Printing Press to the Superhighway. Routledge, London.

Mass Media, Political Economy of

Graham Murdock, Loughborough University, Loughborough, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Political economy is an intellectual project, and effort to develop a model of social order and change. Political economy set out to understand how capitalism worked and assess its consequences for the organization of social and political life. The central question that a political economy approach asks is how public communications' embedded in a capitalist economy based on private enterprise, affect its ability to provide disinterested information, plural viewpoints, open debate, and communal culture required by democratic theory? The central assumptions of the perspective and its questions is at the center of this article.

The Birth of Political Economy

Political economy first emerged as an intellectual project in the eighteenth century as part of the Enlightenment's concerted effort to replace religious and metaphysical accounts of the world with rational, man-made, models of order and change. As a building block in what was later to become the 'social sciences,' political economy set out to take the new machinery of capitalism apart, find out how it worked, and assess its consequences for the organization of social and political life. Following the revolutionary upheavals in America and France, these connections came to center more and more on the material and cultural conditions required to secure a democratic social order in which people were no longer subjects of a monarch or an autocratic ruler, but were autonomous citizens. This new status conferred rights to participate fully in social and political life and elect the representatives making the laws by which they consented to the obligation to balance individual advancement against contributions to collective life. Envisioning a society of citizens lay at the heart of the Enlightenment's emancipatory project of banishing ignorance, fear, and exploitation and promoting the rational pursuit of liberty, equality, and fraternity. This ideal continues to provide political economy with its moral touchstone positioning many of its practitioners as critics of capitalism's failure to reconcile possessive individualism with the common good.

Capitalism, Democracy, and Communications

It was immediately clear to many observers, that the emerging system of public communications, centered at that time on the daily press, had a pivotal role to play in securing a democratic social order. First, it provided the most effective way of disseminating the information and knowledge citizens needed to understand the problems facing society and how these might be tackled. Secondly, it offered a communal space in which alternative interpretations of the situation and competing policies could be tested against the available evidence. Thirdly, by overcoming geographical separation it helped construct the nation-state as an imagined community of citizens as well as an administrative apparatus.

By the mid-nineteenth century however, it was equally evident that communication systems also played a pivotal role in organizing the new industrial forms of capitalism. First, they provided the infrastructural supports that allowed firms to track and coordinate activities that were becoming ever more diverse and dispersed. Secondly, they provided new outlets for display advertising that was increasingly central to the task of matching supply to demand. Thirdly, they offered enticing opportunities for capitalist enterprises in their own right. As migrants moved into the cities to work in the new factories and offices, the demand for popular reading matter and later for recorded music and cinema, grew exponentially.

The central question that these developments posed for political economy was deceptively simple: 'How did public communications' increasing integration into a capitalist economy based on private enterprise affect its ability to provide the disinterested information, plural viewpoints, open debate, and communal culture required by democratic theory?

A Free Market in Ideas

For the first half of the nineteenth century most commentators argued that the best guarantee of reliable information and open democratic debate was a free market in ideas and arguments, in which a plurality of publishers competed for readers with minimum interference from the state. There were good reasons to identify the state as the principal enemy of press freedom. In England, the government had introduced an array of compulsory levies on the press (including taxes on the paper used and the advertisements) in an effort to make radical publications economically unviable. The battle to repeal these 'taxes on knowledge' (as they were known) lasted until 1869, when the last remaining provision was abolished. However, it was at exactly this point, when the nominal freedom to enter the newspaper market was extended, that the introduction of new production technologies (improved presses, mechanized composition, and new graphic capabilities) rapidly raised the price of market entry and began to concentrate press ownership in the hands of wealthy individuals. This development placed a sizable question mark against the press's ability to fulfill its promise of acting as a watchdog on abuses of power. In their pamphlet of 1846, *The German*

Ideology, Karl Marx and Frederick Engels had famously argued that the capitalist class's control of economic resources allowed them to 'regulate the production and distribution of the ideas of their age' and manipulate public culture in their own interests. Witnessing the rise of the new 'press barons,' even commentators who rejected revolutionary politics began to wonder if Marx might not be at least partly right. Most of the old style proprietors had owned a single title, which they often edited themselves. The new press owners commanded chains of titles stretched across key geographical markets. Moreover, the increasing importance of advertising seemed to many observers, to be extending the influence of capital still further. Drives to maximize readership among the urban working class had led to cuts in cover prices and the spread of the 'penny press.' Shortfalls in income were made up by advertising revenues, altering the previous relations between paid publicity and editorial matter. As the English commentator, Thomas Bowles, noted in 1884: "Newspapers … profess to sell news and give advertisements to boot. What they really do is sell publicity … and give news to boot."

The emergence of a press system funded primarily by advertising whose ownership was increasingly concentrated in fewer and fewer hands challenged the ideal of public communications in the service of democracy in several ways. Firstly, it raised the possibility that proprietors might use the titles they owned to promote their own political enthusiasms or business interests while denying space to opponents and critics. Secondly, dependence on advertising revenue opened newspapers to editorial pressure from large advertisers wishing to promote their activities or censor unfavorable publicity. Thirdly, the growth of advertising content privileged the speech of commerce and squeezed the space available to other voices. Fourthly, because advertisers required access to either mass readerships or wealthy niche markets, publications catering to marginalized minorities or less-affluent groups became ever more precarious or unviable. Surveying the situation in the United States in 1910, Edward Ross, reluctantly concluded that 'the commercial news-medium does not adequately meet the needs of democratic citizenship.' Many of his contemporaries agreed, but saw the entrenched identification of a free press with a free market as an immovable barrier to reform. As Delos Wilcox lamented in 1900: "The American people must dearly love the freedom of the press, or we should have heard before now much talk of government control or operation of the newspaper" (Wilcox, 1900).

The growing tensions between the communication system's centrality as simultaneously a resource for capitalist expansion and a support for democratic culture prompted two reactions. Some analysts set out to develop a political economy of public communication, measuring corporate performance against the requirements of democratic theory. Others, like Alfred Marshall in his *Principles of Economics* (1890), announced the death of political economy and birth of 'economics.'

Political Economy and Economics

By dropping the adjective 'political,' economists measured their distance from political economy in three ways. Firstly, where political economists set out to trace the general consequences of shifts in economic organization for social and cultural lives, economists focused on the working of 'the economy' defined as a separate, bounded, sphere. Secondly, because economists saw economic behavior as the product of a universalistic individualist psychology rather than of socially constructed institutions, they generally avoided critical reflection on prevailing institutional arrangements and possible alternatives to them. They saw the economy as a set of interlinked markets, and not as a system of power. Thirdly, they defined their role in technical terms and detached themselves from political economy's engagement with moral philosophy. They were concerned with efficiency rather than justice, with individuals as consumers rather than citizens. The field of mass media research is still deeply marked by this division.

Media economists are primarily interested in the peculiarities of the communication industries as industries and tend to see regulated markets as the least worst way of organizing them. In contrast, critical political economists of media have tended to focus on market systems' failures and distortions in the cultural sphere and to champion various forms of public intervention. They begin with the possession of core forms of capital (both material and symbolic) rather than with the dynamics of exchange.

Institutionalizing the Public Interest

In 1808, Thomas Jefferson's Treasury Secretary, Albert Gallatin, argued strongly for federal control of America's developing road and canal system on the grounds that they would befit everyone by uniting 'the most remote quarters' of the nation, both physically and imaginatively. Later observers saw the same potential in the telegraph and the telephone, arguing that these new media markets were 'natural monopolies' because a single operator was likely than competing firms to optimize both economic efficiency and social benefits. This argument was widely accepted, but institutionalized in different ways. In the United States the task was assigned to a private company, American Telephone and Telegraph (AT&T), charged with providing a universal service that was both geographically comprehensive and priced at a level that made it accessible to all, and regulated (from 1934) by the Federal Communications Commission (FCC). In contrast, in Europe the task of providing universal service was assigned to state monopoly Post, Telegraph and Telephone (PTT) organizations owned and operated by the state. This same division between public enterprise and federally regulated private companies also characterized the institutionalization of broadcasting.

Unlike cinema seats or paperback books which are commodities purchased for personal consumption, broadcast services are public goods. They are available to anyone with suitable receiving equipment, and any number of people can listen or view at the same time without interfering with each other's enjoyment. This characteristic provided the technical rationale for the key tenet of universality. The institutional forms that broadcasting assumed in the early 1920s however were profoundly shaped by two contexts. Firstly, the British experience of managing strategic resources, such as timber in World War I, had convinced at least some economists that public corporations were the most appropriate way of

addressing the limited space in the radio spectrum available for broadcasting. Secondly, the social unrest that had followed the end of the War, the victory of the Bolsheviks in Russia, and the extension of the franchise to women for the first time led many commentators to see a public monopoly as the best way of securing responsible and rational participation in mass democracy in uncertain and volatile times. Accordingly, after careful deliberation, in 1926 the British government opted to transform the commercial monopoly initially given to the British Broadcasting Company (operated by a cartel of radio manufacturers) into a public service monopoly, the British Broadcasting Corporation, forbidden to take advertising, funded by a compulsory license fee on set ownership, and charged with providing informational and cultural resources for citizenship, as defined by political and cultural elites. In the United States, in contrast, radio was widely seen from the outset as a business opportunity. Its promise of widespread and instantaneous access to homes made it an attractive advertising medium in a context where a mass consumer system was expanding rapidly. Despite dissenting voices and a fierce rearguard action by educational and other public interest groups, broadcasting became primarily a privately owned advertising supported system, oriented to maximizing audiences by providing popular entertainment, with the FCC allocating (and revoking) broadcast licenses and instituting backstop regulation to ensure a minimum degree of access to air-time for minority viewpoints.

These contrasting models of broadcasting as public service and commercial enterprise, communication for citizenship and communication for consumption, provided the basic templates for many other countries, although a number eventually settled for a mixed broadcasting economy, with public and commercial sectors running in tandem.

Consolidation and Critique

Broadcasting and telecommunications were exceptions, however. The key organizations in the other major mass media sectors – the press, cinema, recorded music, and advertising – remained mostly privately owned and minimally regulated, consolidating the potential control that corporations could exercise over the organization of public knowledge and culture. Critical observers responded by reaffirming and extending the main themes of earlier commentaries. Some, like Upton Sinclair (1920) in his attack on the American commercial press, *The Brass Check*, returned to the potential abuses of owner and advertiser power. Others, notably Danielian in his 1939 study, *The AT&T*, turned their attention to the operation of power in conditions of regulated monopoly. In Britain, Klingender and Legg, observing the monopolistic tendencies in the country's cinema industry and the growing power of the Hollywood studios, produced the first detailed political economy of a major mass media sector, *The Money Behind the Screen* (1937). But the most comprehensive contribution to establishing a general political economy of media came from two German scholars, Theodore Adorno and Max Horkheimer, who were strongly influenced by Marx, and whose 1944 essay, 'The Culture Industry,' sketched out a general model of the possible relations between the increasing commodification and industrialization of cultural production and what they saw as the narrowing possibilities for creativity and critical expression in the public sphere.

Most of these early contributions came from commentators who worked either outside the universities or on the edges of academia. It was not until the expansion of the university system in the 1960s and the rapid development of communications and cultural studies as academic specialisms in their own right that the political economy of the mass media found a secure institutional base within the academy.

Political Economy and Cultural Analysis

This new security led to a significant growth in academic research, but tenured faculty did not enjoy a monopoly of insight; and commentators working outside the universities continued to make important interventions in debates. Although this renewed interest in the relations between culture and economy ranged widely, it tended to coalesce around several key strands.

Firstly, the rise of multimedia conglomerates with interests spanning a range of public communications sectors breathed new life into long-standing questions about the possible links between patterns of ownership, structures of corporate control, and the range and diversity of cultural production, and produced a significant volume of research charting the new configurations. Debates around the implications of these emerging structures also found an audience outside the academy through books like Ben Bagdikian's *The Media Monopoly* (2000).

Secondly, analysts attempted to refine the general idea of 'the cultural industries' by investigating the contrasting ways production processes were organized within major media sectors and tracing their implications for the range and forms of cultural activity. The contributions of French scholars led by Patrice Flichy (1980) and Bernard Miege et al. (1986) were particularly significant in establishing this line of research.

Thirdly, researchers returned to political economy's central concern with the constitution of complex democracies by way of Jurgen Habermas's influential idea of 'the public sphere' as a space of open debate situated between the state and capital and relatively independent of both. This led a number of critical analysts in Britain and Europe to reply to commentators calling for an enlarged market sector in television with a vigorous defence of public service broadcasting as the cornerstone of a mediated public sphere.

Fourthly, scholars moved beyond the issues raised by the changing conditions of national media to explore the recomposition of the emerging transnational communications system. Just as the last of the former colonial territories achieved political independence and emerged as autonomous nation states, critics led by Herert Schiller (1969) argued that they were being incorporated into a new imperial system, based not on the annexation of their territories but on the colonization of their cultures by the aggressive consumerism promoted by the major American media and entertainment corporations.

Fifth, encouraged by the rapid growth of the underground press, community radio, and a range of other alternative and oppositional media from the mid-1960s onward, critical

researchers began to explore the political economy of nonmainstream media.

These various strands of work attracted widespread criticism from cultural analysts working in other intellectual traditions who accused political economists of reducing cultural life to economic forces and denying the agency of audiences and consumers. This criticism is misplaced. Political economy does not deny that consumption may be marked by creativity, improvisation, and opposition but its core concern is with how the underlying dynamics of economic life shape communicative practice by allocating the material and symbolic resources required for action and how the resulting asymmetries in their control and distribution structure both the conditions of creativity in cultural production and the conditions of access and participation in cultural consumption. Audience may be active but they are also commodities traded between advertisers and commercial media.

Marketizing Media

As noted earlier, political economy is particularly concerned with the way that asymmetries in the distribution and control of core resources for cultural action impinge on the vitality of participatory democracy. In pursuing this argument, critical political analysts start from the material and cultural resources required for full citizenship. This leads them to evaluate the performance of public communications system in terms of their promotion of cultural rights-to information, knowledge, representation, and respect and their provision of a cultural commons hospitable to the negotiation of unities that incorporate diversity. This is in marked contrast to those economists and mainstream political economists who start from the conditions required to maximize market choices and address individuals as consumers of goods rather than as participants in a moral community.

Since 1980, propelled by a revival of neoliberal economic thinking the market has increasingly become the preferred mechanism for organizing public communicative activity and the touchstone for evaluating the performance of those cultural institutions that remain within the public sector. This general process of marketization (Murdock and Wasko, 2007) has four major dimensions:

1. *Privatization*. Media organizations that were formerly part of the public sector and supported by public money have been sold to private investors and converted into private enterprises. Examples include BT, Britain's former PTT and now one of the leading corporations in the national economy, and the former French public service television channel, TF1.
2. *Liberalization* involves opening markets that were previously dominated by one or two operators (monopolies and duopolies) to competition from new entrants. Examples include the liberalization of telecommunications in the United States with the breakup of AT&T and the ending of its historic monopoly, and the introduction of commercial cable and satellite television services into European broadcasting markets previously monopolized by state-funded public service channels.
3. *Regearing of the regulatory regimes* by shifting toward 'lighter touch' regulation that allows private companies more freedom of action in key areas such as corporate acquisitions and the amount and type of advertising carried, and removes or weakens established public interest requirements.
4. *Corporatization* involves reducing the real level of public subsidy and encouraging or impelling public institutions to raise more money from market-based activities, moving them closer to the strategies and rationales of the corporate sector. The BBC's decisions to maximize the commercial returns on its programming, to develop new initiatives with private sector partners, and to launch its own commercial enterprises, provide a good example of this process in action.

Taken together these intersecting processes of marketization have had three major effects on the ecology of public communications. First, they have consolidated and extended the economic power of the major multimedia corporations by opening up new fields for enterprises (Fitzgerald, 2011). Secondly, they have eroded the relative countervailing power of public institutions. Thirdly, they have progressively detached the core sectors of public communications from the project of providing everyone with the basic cultural resources required for full citizenship and moved them toward the satisfaction of plural consumer demands. As market segments have proliferated, the cultural commons has contracted.

Globalizations

The process of globalization is not confined to the major capitalist economies. With the collapse of the Soviet Union, the expansion of markets in the world's two most populous countries, China and India, and the almost universal embrace of market disciplines in the rest of the world, it has become the dynamic principle underlying economic globalization. This process has had contradictory consequences. Firstly, it has extended the geographical reach of the major western media conglomerates prompting a revival of the cultural imperialism thesis. Secondly, it has expanded opportunities for national entrepreneurs creating new forms of competition and collaboration between global, regional, and local companies. Thirdly, it has raised new questions about the relations between capital and the state in the cultural sphere, dramatized most forcefully by the increasing economic power of China and the endemic tensions between the 'party line and the bottom line' (Zhao, 2008).

Digitalizations

The present media systems is also being transformed by the fundamental technological shift, from analogue to digital communications systems, and the emergence of new spaces of expression and enterprise, led by the Internet. Four dimensions of this shift are of particular interest to political economy. First, the growing concentration of control over the Internet in the hands of established media majors and the new communications conglomerates, typified by Google, coupled with the centrality of advertising finance has revivified long-standing concerns around the impact of ownership and commercialization and the role of regulation. Secondly, the commercial appropriation of ideas and material provided by users, also the routine collection, processing, and the sale of their personal

information has prompted have renewed interest in the commodification of audience activity and the shifting relations between work and play. Thirdly, digitalization has major consequences for public cultural institutions such as museums, libraries, and public service broadcasters, furnishing opportunities to extend the range of materials and services they offer, and to forge new relations with their constituencies. Fourthly, the explosion of vernacular expression and horizontal peer-to-peer exchange on the Internet has revived interest in the dynamics and sustainability of non-market gift economics based on reciprocity and their possible role in creating a new public sphere of deliberation and participation capable of reinvigorating democratic ideals and processes.

Analyzing the shifting relations between the cultural economies of commodities, public goods and gifts, and the ways they are being shaped and reshaped by the intersecting forces of marketization, globalization, and digitalization not only presents the political economy of communications with a major challenge but also presents a major opportunity to reaffirm the continuing relevance and centrality of its core concern with the constitution of democratic life.

See also: Civil Society, Concept and History of; Critical Theory; Cultural Mediators and Gatekeepers; Marxism in Contemporary Sociology; Public Sphere and the Media.

Bibliography

Bagdikian, B.H., 2000. Media Monopoly, sixth ed. Beacon Press, Boston, MA.
Barnouw, E., et al., 1997. Conglomeration and the Media. The New Press, New York.
Danielian, H.R., 1939. The AT&T. Vanguard, New York.
Fitzgerald, S.W., 2011. Corporations and Cultural Industries: Time Warner, Bertelsmann and News Corporation. Rowman and Littlefield, Boulder, CO.
Flicy, P., 1980. Les Industries de Imaginaire: Pour une Analyse Economique des Media. Universitaires de Grenoble, Grenoble, France.
Garnham, N., 2000. Emancipation, the Media and Modernity: Arguments about the Media and Social Theory. Oxford University Press, Oxford, UK.
Golding, P., Murdock, G., 2000. Culture, communication and political economy. In: Curran, J., Gurevitch, M. (Eds.), Mass Media and Society, third ed. Arnold, London, pp. 70–92.
Horkheimer, M., Adorno, T.W., 1973[1944]. The culture industry: enlightenment as mass deception. In: Horkheimer, M., Adorno, T.W. (Eds.), Dialectic of Enlightenment. Allen Lane, London.
Klingender, F.D., Legg, S., 1937. The Money behind the Screen. Lawrence and Wishart, London.
Marshall, A., 1890. Principles of Economics: An Introductory Volume. Macmillan, London.
Marx, K., Engles, F., 1970[1846]. The German Ideology (Edited and Introduced by C.J. Arthur). Lawrence and Wishart, London.
Mattelart, A., 1979. Multinational Corporations and the Control of Culture. Humanities Press, NJ.
McChesney, R.W., 1999. Rich Media, Poor Democracy: Communication Politics in Dubious Times. University of Illinois Press, Urbana, IL.
Miege, B., Pajon, P., Salaun, J.-M., 1986. L'Industrialisation de L'Audiovisuel. Editions Aubier, Paris.
Mosco, V., 1996. The Political Economy of Communication. Sage, London.
Murdock, G., Wasko, J. (Eds.), 2007. Media in the Age of Marketisation. Hampton Press, Cresskill, NJ.
Schiller, H., 1969. Mass Communication and American Empire. A M Kelley, New York.
Schiller, H., 1989. Culture Inc: The Corporate Takeover of Public Expression. Oxford University Press, New York.
Sinclair, U., 1920. The Brass Check: A Study of American Journalism. Published by the author. Pasadena, CA.
Smythe, D., Guback, T. (Eds.), 1994. Counterclockwise: Perspectives on Communication. Westview Press, Boulder, CO.
Wasko, J., Murdock, G., Sousa, H. (Eds.), 2011. The Handbook of Political Economy of Communication. Wiley-Blackwell, Oxford.
Wilcox, D.F., July 1900. The American newspaper: a study in social psychology. Annals of the American Academy of Political and Social Science 56–92.
Zhao, Y., 2008. Communication in China: Political Economy, Power and Conflict. Rowman and Littlefield, Boulder, CO.

Mass Media, Representations in

Robert Entman, North Carolina State University, Raleigh, NC, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9363–9368, © 2001, Elsevier Ltd.

Abstract

Representation involves making something perceivable to an audience not in its physical presence. Inevitably partial, representation in the media requires framing, the selection and highlighting of aspects to stand for the whole. Scholars of media representation probe the combination of absences and presences in media texts, the meanings this total package likely conveys to audiences, and the ramifications of patterns in representation for power and belief in a society. Following a conceptual discussion, this article explores three major scholarly issues: how representations in the media reflect and affect the social status of group members such as women and blacks; how the representations might influence a mass audience's preferences in a particular election or policy dispute; and the ways media representations might influence elites and officials. The article suggests that media representations help define a society's culture, reflecting and reinforcing dominant cultural values and status hierarchies, yet also, at the margin, revealing and stimulating cultural change.

Representation involves making something perceivable to an audience that is not in its physical presence. Often the item represented – an idea, an ideology, an interest – possesses no tangible physical embodiment so representation means representing it in a new symbolic form or context. Other times a physical antecedent exists, such as a person, a widely recognized social group, an institution, or a place. Even these objects, when represented in the media, typically carry ideological or cultural meanings. There is now an extensive literature on media representations of social and political groups and institutions, and on their implications for the distribution of status, wealth, and power.

No representation is comprehensively accurate. All media representations inevitably omit some aspects of the item represented. Indeed in most cases only a tiny fragment of the actual idea, person, or group will be perceivable to the audience. Audiences may and often do fill in much of the absence themselves, drawing on their pre-existing mental theories or schemas. Studying representation means probing the combination of absences and presences, the meanings this total package is likely to convey to audiences, and the ramifications of these patterns of representation for power and belief in a society.

It is also vital to understand that there is no one-to-one correspondence between representation in the media text and the representation of the text and its signals as inferences drawn in the minds of audiences. The message that a communicator intentionally encodes in a media text may get decoded in ways that contradict or bear little relation to the communicator's purposes. Because representation involves both encoding and decoding, the conscious intentions of the communicator may have surprisingly little relevance to individuals' reactions and thus to the political and social impacts of media representation. This point is sometimes lost in popular and even academic studies of the media's influence, which frequently take individual media personnel to task for what analysts assume will be the effects of the representations. Vidmar and Rokeach (1974) provided a classic example of the slip between intentions and influence in their study of the popular American television program 'All in the Family.' To illustrate the absurdity of ethnic prejudice, Norman Lear, the show's creator, made its central character, Archie Bunker, a laughably ignorant, outspoken bigot. Yet despite the ridicule heaped upon Archie and his bigotry in most episodes, Vidmar and Rokeach revealed, the show merely reinforced viewers' predispositions. Far from changing their minds or even experiencing discomfort, prejudiced viewers saw Archie's tirades as vindication of their own sentiments.

Communication scholars studying media representation have typically done so to discern likely media influence. But cultural studies analysts and others working in the postmodern mode found surprising allies among those within the positivist and behavioralist traditions in stressing that individuals' interpretations of a given text vary so markedly that studying texts in isolation from their audiences produces flawed accounts of media representation. The major work of representation goes on within each thinking individual's brain rather than in the text, they say. Therefore, much of the scholarly focus in the 1990s shifted towards exploring the audience's interpretations rather than the media text by itself.

This article assumes that the representations in the text influence without determining individuals' sentiments. Beyond this, at the aggregate level, the pattern of representations in a society's mass media illustrate and help define its culture, reinforcing dominant cultural values yet also revealing and stimulating cultural change. These assumptions do not require a one-to-one correspondence between text and mind or text and culture, but do insist that the correlations are often significant. The conflict over just how significant – that is, how powerful media representations are – runs through media scholarship from its earliest days and continues in the most recent work. To cite ideal, polar types, some take the position that individuals make up their own minds, independently reconfiguring anything represented in the media to find meanings that suit their own experiences, knowledge, psychic needs, and goals. Others assert that in 'telling people what to think about,' the media can significantly and sometimes decisively influence an audience's attitudes and preferences. Between these two poles lies a diverse array of theorists who believe an audience's thinking is both influenced by and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.95043-8

influences the representations that media transmit, with the degree of media impact depending on many environmental and individual differences.

One of the first major bodies of social scientific research on mass communication explored the representation of violence in movies and its impacts on youth. World War II spawned much research on propaganda – on media texts and their uses of symbols and language to generate potential shifts in mass attitudes. The next two decades saw a dry spell as scholars generalized from this and other early empirical research into media effects, concluding that any consequences of media representation were of minimal significance, limited largely to reinforcing existing predispositions. Advances in conceptualization and in analytical techniques challenged this stance and led to a renaissance in scholarship on media influence.

The more recent social scientific scholarship tends to run in two channels, one focusing on journalism and the news, the other on television and film entertainment (with occasional attention to magazines, radio, popular music, and novels). However, scholars toward the end of the century were increasingly recognizing the blurring of this always-problematic distinction between entertainment and news. Not only was news becoming more 'entertaining,' more shaped by strategies for audience appeal traditionally used by entertainment – in response to rising competition from commercial broadcasters, websites, print media, and other sources – but entertainment offerings on the US broadcast networks were becoming increasingly topical. A resulting hybrid dubbed 'infotainment' occupied increasing portions of the broadcast airwaves and of daily newspapers as well, in the US and elsewhere. As a result of this dissolution of once-distinct lines, an essay on media representation must not limit itself to the traditional news media, even though the bulk of social scientific research has probed the news.

Framing

Representation is always partial, always requires selection of aspects to stand for the whole (otherwise we would be speaking of presentation, not representation). Since framing is the process of selecting, highlighting, and suppressing, it is the means through which most media representation takes place. Framing has organized an increasing portion of the literature on media representation. Most of this research rests on the proposition that framing devices are worth studying because they affect (without necessarily determining) the audience's thoughts and perhaps actions. Framing can be defined as selecting aspects of a perceived reality and making them more salient in a communicating text, thereby promoting a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation.

The classic experiments of Kahneman and Tversky show that framing the same basic event or problem differently can have marked effects on an audience's understanding of the item represented. For instance, Kahneman and Tversky (1984) gave subjects in one group a choice between a disease prevention program, A, that would 'save' 200 out of 600 people, and B, which would have a one-third probability that all 600 would be 'saved.' Then, using a matched group, Kahneman and Tversky reversed the frame so that option C meant '400 people will die' and option D offered a one-third probability that 'nobody will die.' Option C was identical to A, as was D to B, excepting only that A and B were framed in terms of 'saving' and C and D in terms of 'dying.' Yet in the first experimental group, 72% of subjects chose A, whereas only 22% in the second group chose the identical option C – presumably because C was framed to highlight death instead of lives saved.

Kahneman and Tversky's experiments typically involved small changes in wording bringing about significant changes in response (apparently because the participants in the experiments thought about the matter differently, focusing their minds on different aspects of it). Media studies tend to document far more global differences in framing, involving not just word choice but disparities in visual portrayals, aspects of the item emphasized, sources relied upon, people quoted, and more. If differences of a single word can cue markedly different responses, the multidimensional differences in media frames often demonstrated by scholars implicitly suggest considerable media influence.

For example, US media developed starkly different frames for two incidents in which civilian airliners were shot down by military forces, resulting in nearly 300 deaths each. In the first incident, the downing of a Korean Air Lines jet by the Soviet Union in 1983, adjectives and nouns, dramatic graphics, attributions of agency and other aspects of the text emphasized the suffering of the victims and the Soviets' moral guilt. This frame fit with then-president Ronald Reagan's political agenda of heightening Americans' distrust of the Soviet Union and polls showed that the overwhelming majority of Americans accepted Reagan's interpretation of the incident. By contrast, when US forces shot down an Iranian passenger jet in 1988, the media framed the event as a regrettable accident traceable to understandable human error and deficiencies in technology. Language eschewed moral judgments and the victims received scant attention. Here too large majorities of Americans in surveys accepted the Reagan administration's framing – this time exculpatory rather than condemnatory. Public acceptance of the way the media represented the incidents in turn had significant political and policy consequences (Entman, 1991).

Nonetheless, framing is far from all-powerful. William Gamson's studies of 'collective action frames' have shown how individuals, interpersonal discussion, and media frames interact. Studying issues including nuclear power and affirmative action, Gamson (1992) found that audience members take cues from news reports, but, depending on their individual thinking, group discussions, and the nature of the issue, can draw from the media-supplied frames new ways of understanding the phenomenon reported. For example, journalists did not construct the dominant news frames and thus the representation of nuclear power to stimulate antinuclear activism. Nonetheless, audience members gathered by Gamson in focus groups were sometimes able through collective deliberation to reframe the issue in a manner that might have moved some to action.

An understanding of the bidirectional influence of media representations and audience thinking immediately cautions against inferring any single meaning or effect from a particular media text. The same text can represent or 'mean' different things to different people in different circumstances – including

different historical periods. Thus the words 'racial equality' represented one concept during the 1950s and 1960s – equal treatment and equal rights under the law for black Americans. More recently, the words appear to connote something quite different for many at least in the US, something akin to radical redistribution of income, wealth, and opportunity from the majority Whites to minority groups.

Despite these cautions, the bulk of research on media representations appears to assume that one basic interpretation often – though certainly not always – dominates, at least among the members of any specific culture at a particular time. In fact, one definition of a culture might be sharing cognitive habits and schematic mental linkages that promote similar responses to an attitude object among the individuals who constitute its membership. The (typically implicit) assumption that representations in the media have a 'preferred' or 'privileged' interpretation or meaning in a given culture structures much of the remaining research reviewed here.

To generalize broadly (an inevitability in a literature synthesis covering such a wide area of research), media research explores three aspects of representation corresponding to three kinds of effects (often presumed rather than empirically demonstrated): the ways representation reflects and affects the general social status of group members such as women and blacks; the ways media representations might affect a mass audience's political preferences in a particular dispute, say in an election or a debate over legislative proposals; and the ways media representations might influence political elites and government officials.

Status Reinforcement or Enhancement

Students of media representation frequently examine depictions of groups and how they might reflect, reinforce, or challenge existing status hierarchies. For this purpose, a wide range of human classifications has been invoked, from race and ethnicity to gender, sexual orientation, age, and language groupings.

In this research tradition a major thrust focuses on ethnic groups. For the most part, the research suggests that media images reflect and reinforce dominant status judgments. That is, in framing activities involving members of lower-status groups, the media highlight aspects likely to reinforce perceptions among the dominant group's members of their difference from and superiority to outgroups. Representations tend to carry explicit or implicit judgments of the outgroups on dimensions of morality, intellectual capacity, industriousness, lawfulness, and other valued traits. But because media images are in constant flux and often contain contradictory elements, they also provide cognitive resources for challenges to status markers.

Of particular concern to many researchers has been a pattern whereby the media link crime and dangerous deviance to people of color and especially black males (Entman and Rojecki, 2000). Experimental studies repeatedly have suggested that these representations engender cognitive and emotional effects on white audiences that affect the exercise and distribution of power – in this case heightening support for punitive prison sentences or diminishing support for white candidates when political advertisements associate them with black criminality (Gilliam and Iyengar, 2000).

Studies of depictions of 'foreigners,' by which is usually meant nations and peoples other than those with white, European ancestry, also fit into this mode. For example, news media coverage of humanitarian crises in Africa, though typically sympathetic to those suffering, manages to endorse racial status hierarchies that place whites at the peak of the pyramid. Thus, Western news media made Bosnian whites appear in a variety of ways more human, more worthy of help, than news of the genocide going on simultaneously in Rwanda (Myers et al., 1996). The same appears true in entertainment. For example, a study of Brazilian TV star Xuxa suggests that her celebrity (and earnings) derived largely from her conformity to North American and North European ideals of blonde sexuality and beauty (Simpson, 1998) – even in an ethnically heterogeneous country priding itself on multiracial tolerance. This white cultural preference reflects continuity with a long history in which European and US mass media implicitly derogated black people and valorized whites and the qualities typically associated with whiteness.

In a similar vein, gender representations in the media are found to register cultural change on some dimensions but to endorse traditional patriarchy and gender roles on others. Thus we see women actors portraying business executives in advertisements, and single women living happy lives without obsessing over men. But, although women's sports now receive media attention that gives the lie to assumptions that women are inevitably passive, gentle, timid and the like, closer study suggests the continued impact of gender stereotypes and a preference for males. For instance, research on portrayal of women's sports in Sweden shows that female athletes and their pursuits are implicitly derogated in comparison to males – over 90% of the time devoted to sports in a sample of Swedish television programming covered male athletics (Koivula, 1999). Beyond this, women still drive the plot in far fewer movies and television programs than men, and still serve as sex objects more often than males. Women's sexuality indeed appears more than ever to be pictured in unrealizable body types that have been linked to an increase in anorexia and other eating disorders. Thus a study of Australian adolescent girls finds a strong influence of media consumption on tendencies to aspire toward unhealthily thin body types, findings replicated in other countries (Wertheim et al., 1997). The studies of female body images typically find that many audience members absorb the preferred meanings and integrate them into their own self-images. These findings challenge any claim that mass audiences always engage in subversive readings of media representations.

On the other hand, Fiske (1992) argues that despite the traditional patriarchal and commercial values inherent in the pop star Madonna's representations of femininity, young women frequently found reinforcement for nontraditional roles and values through their own counter-readings of Madonna's music and videos. Certainly the growth in sheer numerical representation of blacks, women, and other traditionally less-empowered groups in nontraditional roles in Hollywood movies, music videos, and television dramas has both reflected and reinforced a degree of movement toward a more egalitarian status hierarchy.

Looking at status on the dimension of sexual orientation, Joshua Gamson (1999) argues that although the confessional talk shows that became popular in the US during the 1990s, such as 'Jerry Springer' and 'Maury Povich,' seem on the surface to belittle anyone who deviates from the middle class, heterosexual norm, the programs actually offer a more complex ideological brew. Sometimes the shows construct morality plays that censure gay, lesbian, or transsexual guests, but other times the intolerant snob, ivory tower expert, or sexual bigot assumes the role of villain. Arguably, says Gamson, gays and others living outside the bounds of dominant cultural values find a more welcoming home for self-representation in the 'trash' talk shows than in the rest of the mainstream media. In this sense, tabloid TV, driven as it may be by crass commercial goals, debasing as it may appear to public discourse, may open up new space in the public sphere for persons, ideas, and values previously denied much representation in either entertainment or news.

Media Representation and Political and Policy Disputes

Moving along a continuum from media representation's impacts on mass publics to impacts on the political linkages between them and elites, we turn now to representation in news coverage of current policy and political contests. Here too the research suggests media most often operate to reinforce structures of power that privilege native born white men.

Thus, for example, female political candidates have been shown to face disadvantages against male candidates. They often receive less news coverage than males, and the news agenda seems to follow male candidates' issue choices most closely. On the other hand, women candidates may benefit from being associated with positive sex-role stereotypes (Kahn, 1994). With a black candidate in the race, campaign news tends to emphasize candidates' racial identities and to frame elections as contingent on bloc voting by whites and blacks (Reeves, 1997), and this demonstrably handicaps African American candidates.

The representation of nonelectoral dissent against powerful institutions has also received careful scrutiny. Research reveals a tendency for media to accept moderate, polite dissidents who work within the system and make conventional, incremental demands, but to slant decisively against groups that seek more radical objectives, undermining their legitimacy with the public and thus reducing the need for elites to respond positively to their goals (Gitlin, 1980).

A subject of much dispute in the US has been representations of parties and candidates in national elections, with some scholars alleging systematic biases favoring one party and others denying these exist. Additional studies explore other kinds of bias than partisanship, suggesting the possible impacts on voting turnout or civic orientations. For example, US journalists increasingly infused their representations of candidates and campaigns with cynicism during the latter decades of the twentieth century, and this appeared to encourage political distrust and withdrawal among audiences.

Alongside such research there also arose theoretical essays that challenged the notion of bias by questioning the legitimacy of its presumed opposite, 'objectivity.' The notion that media can represent people and events in comprehensively accurate and politically neutral ways appears to be a powerful professional ideal for journalists. Nonetheless, scholarship almost universally denies the proposition and asserts that in the course of manufacturing the news, individuals and media institutions must inevitably embed certain kinds of biases (that is, preferred meanings or politically consequential omissions) into their reports. Although US studies, at least, fail to uncover systematic partisan biases, other forms of bias have been analyzed extensively, such as a bias for reporting political conflict over agreement, and political process over policy substance.

Elite Impacts

The newest and least developed area of research examines the impacts of media representations on elites. In democratic systems, elites are under pressure to respond to 'public opinion,' but it turns out they have few reliable ways to discover what it is on any given issue under debate in the legislature. Surveys are only sporadically useful in trying to gauge the sentiments of a legislator's constituency. Thus, media representations of 'public opinion' to officials become highly consequential. Political elites can only respond to the public thinking that they perceive, and they get much of their data on this from media reports. Moreover, these reports can become self-fulfilling, because they also may influence the public itself. In this view the media become a representative vehicle in the more specific political sense of forging ties (imperfect to be sure) between mass public and elites, by representing 'public opinion' to elites – by making abstractions, here the public's purported preferences and priorities, perceivable so elites can act on that information.

The process becomes even more complicated because elites contending to control the outcome of policy disputes frequently resort to manipulating the media. This may help create impressions of public support or opposition, impressions that augment or derail the momentum of policy proposals. Yet another twist is that media representations have themselves become a contentious political issue – an issue that is itself subject to biased representation. In the US, surveys clearly reveal that increasing majorities over the last few election cycles of the twentieth century came to believe that the media exhibit a liberal bias (Watts et al., 1999). Yet, as noted earlier, empirical studies fail to demonstrate a systematic ideological slant in the news. The public's perception of bias appears rooted in news coverage of news-making itself, and in stories quoting politicians, party elites, and think tank intellectuals who have made purported media bias a theme of their speeches and writings. Watts et al., (1999: p. 144) find that in campaign coverage, 'conservative candidates, party officials, and supporters have dominated the discourse with allegations of liberally slanted news content.' The claims of left-leaning bias 'overwhelmingly' outnumber those of conservative bias – in the 1996 campaign by 91 to 9%. Thus there appears to have been a bias in the media's representation of points of view about the media's ideological representations – ironically, conservatives may have succeeded in biasing the US media's representations of media bias. The topic of media representation has itself entered the arena of political contestation and strategy,

although the effects on the media, or on power in politics and society, are not yet clear.

Conclusion

The spread of new media and many new sources of mediated information and images raises questions about the nature of the system of media representation and scholars' ability to generalize about it. The Internet is freely open to almost any representation one might imagine, and its images are readily and simultaneously available to audiences around the globe. The impact of this wider circulation of more diverse images and words could be profound, disrupting the past ability of mainstream media representations to set agendas and shape dominant interpretations. On the one hand, the Internet puts more potential power over individuals' internal representations and thoughts into the hands of each individual, at least those who devote sufficient time to surfing the Web and seeking new views. The seemingly significant media effects discussed here – on group status, on the outcome of elections and other disputes, or on elite behavior – could diminish. On the other hand, this very decline of a homogenizing mass media has some scholars worried that the public sphere will dissipate, only to be replaced by a cacophonous set of specialized subpublics that communicate not with each other but only within their own boundaries. In the future, the very notion of common culture and nationhood forged by public attention to common media of mass communication could come under pressure.

Bibliography

Entman, R.M., 1991. Framing United States coverage of international news: Contrasts in narratives of the KAL and Iran Air incidents. Journal of Communication 41 (4), 6–27.

Entman, R.M., Rojecki, A., 2000. The Black Image in the White Mind: Media and Race in America. University of Chicago Press, Chicago.

Fiske, J., 1992. British cultural studies. In: Allen, R.C. (Ed.), Channels of Discourse, Reassembled, 2nd edn. University of North Carolina Press, Chapel Hill, NC.

Gamson, J., 1999. Freaks Talk Back: Tabloid Talk Shows and Sexual Nonconformity. University of Chicago Press, Chicago.

Gamson, W.A., 1992. Talking Politics. Cambridge University Press, New York.

Gilliam, F.D., Iyengar, S., 2000. Prime suspects: The influence of local television news on the viewing public. American Journal of Political Science 44, 560–573.

Gitlin, T., 1980. The Whole World is Watching: Mass Media in the Making and Unmaking of the New Left. University of California Press, Berkeley, CA.

Kahn, K.F., 1994. Does gender make a difference? An experimental examination of sex stereotypes and press patterns in statewide campaigns, coverage. American Journal of Political Science 38, 162–195.

Kahneman, D., Tversky, A., 1984. Choices, values, and frames. American Psychologist 39, 341–350.

Koivula, N., 1999. Gender stereotyping in televised media sport coverage. Sex Roles 41, 589–604.

Myers, G., Klak, T., Koehl, T., 1996. The inscription of difference: news coverage of the conflicts in Rwanda and Bosnia. Political Geography 15, 21–46.

Reeves, K., 1997. Voting Hopes or Fears?: White Voters, Black Candidates & Racial Politics in America. Oxford University Press, New York.

Simpson, A., 1998. Representing racial difference: Brazil's Xuxa at the televisual border. Studies in Latin American Popular Culture 17, 197–222.

Vidmar, N., Rokeach, M., 1974. Archie Bunker's bigotry: A study in selective perception and exposure. Journal of Communication 24 (1), 36–47.

Watts, M.D., Domke, D., Shah, D.V., Fan, D.P., 1999. Elite cues and media bias in presidential campaigns. Communication Research 26, 144–175.

Wertheim, E.H., Paxton, S.J., Schutz, H.K., Muir, S.L., 1997. Why do adolescent girls watch their weight? An interview study examining sociocultural pressures to be thin. Journal of Psychosomatic Research 42, 345–355.

Mass Society: History of the Concept

Salvador Giner, University of Barcelona, Barcelona, Spain

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9368–9372, © 2001, Elsevier Ltd.

Abstract

The history of the notion and theory of mass society is explored, from its ancient roots in Greek and Roman political thought about 'the many' and 'the few,' until modern times. Theories about the plebs, demagoguery and revolutionary crowds in thinkers such as Plato, Cicero and Hobbes are shown to bear upon later conceptions. The modern idea of a leveled, homogenous society developed by Tocqueville in the nineteenth century, combined with the fears about 'the masses' expressed both by 'crowd psychologists' (Le Bon, Freud) and conservative writers, are seen as the precursors of the contemporary mass society conception, as it appears in its mature form in Ortega's *Revolt of the Masses* (who introduced the notion of 'mass man') and in the works of Mannheim, who introduced the expression 'mass society.' The diverse ramifications of the mass society conception into the contemporary theories of 'mass culture,' 'mass politics,' the 'homogenization' of society, 'mass comsumption,' and the 'depersonalization' of modern human beings (Riesman's notion of the 'outer-directed personality,' Marcuse's 'one-dimensional man') and other, closely-related, interpretations are also analyzed. The widespread persistence of the mass society conception in contemporary thought, often subsumed into other interpretations of modernity, is explored.

'Mass society' is a notion central to the assumption that modern, advanced societies possess the following features: a growing internal homogeneity, a combination of elite and bureaucratic control over the majority of the population (the so-called 'masses'), a specific kind of predominant culture ('mass culture,' linked to the 'mass media'), and an illiberal form of politics ('mass politics' and 'mass parties'). They are also said to reflect a new stage in the development of the industrial economy through 'mass production' and 'mass consumption.' 'Mass society' frequently assumes that, in such advanced societies, a certain type of human personality, the so-called 'mass man,' is proliferating and becoming omnipresent. This mass man is thought to embody the mindless uniformity of social conditions and moral attitudes.

The Twentieth Century Origin of the Concept

In its contemporary sense, the term 'mass society' (*Massengesellschaft*) was coined by Karl Mannheim in Germany in 1935. However, the first general description of a mass society, as it was to emerge in the social sciences and in the sphere of social ideas in the post-World War II period, was first put forward by the philosopher José Ortega in 1929, in his influential essay *The Revolt of the Masses*. Ortega also launched the expression 'mass man' (*hombre masa* in Spanish), a notion intimately related to that of mass society. These contributions from Mannheim and Ortega were preceded by other notions, such as 'mass production,' but most of the 'mass society' expressions appeared after these had become accepted terms. Of these, 'mass politics,' 'mass culture,' 'massification' – a translation of the German *Vermassung* – and 'mass consumption' are the best known. Still others, which did not always include the qualifying word 'mass,' were often used as synonyms of mass society or equivalents for it: for example, David Riesman's 'lonely crowd' and his concept of 'outer directed personality' which largely overlaps with standard descriptions of 'mass man.'

Despite the twentieth century origin of the expression (as well as of the set of concepts to which it soon became inextricably linked), both the notion of mass society and the arguments (or theories) behind it are much older. Thus the idea of a mass society contains some conceptions (and prejudices) about the nature of human society which are truly ancient, while some of the most elaborate interpretations of the world it claims to describe hark back to classical nineteenth century thinkers, such as Alexis de Tocqueville, to several pre-World War I philosophers and sociologists, and to contributions made by the so-called 'crowd psychologists' of the same period.

The Classical Origins of The 'Mass Society' Conception

The Ancient Roots of the Concept

The set of problems destined to occupy the minds of modern theorists of mass society are clearly related to those which inspired the ideas of the earliest conservative critics of democracy. The latter were concerned that the extension of citizenship and equality to the majority of the people making up the body politic might also entail the access of common, ignorant, vulgar, and politically and legally untrained men to public office. Public office, they thought, ought to be the preserve of legally trained, well educated, and wise citizens. Although Plato in the *Republic* put forward a scheme for the selection of the ablest and best citizens for high office (both women and men) on an egalitarian basis – through equal educational opportunities for all – most conservative observers at the time feared that all actual democracies would easily degenerate into mob rule.

Later, some historians, such as the Greek Polybius who lived under Roman rule, suggested that humankind underwent a cyclical historical process and that periods of democracy

quickly and inevitably degenerated into what he called 'ochlocracy,' that is, mob rule and social chaos. Saint Augustine for his part was probably the first to use the word 'mass' for the majority of the people, by referring to those incapable of saving their own souls as the *massa damnata*. He explicitly distinguished an elite – those to be saved, to use his Christian term – and the rest, the undistinguished, united in their failure to free themselves from sin.

Thus the association between crowd behavior, unruly assemblies, emotional and irrational meetings, on the one hand, and the general breakdown of social order, on the other, appeared at the earliest period in the history of Western political thought. Similarly the people's supposedly innate vulgarity and lack of moral fiber was contrasted with the refinement, creativity, courage and distinction, and individuality of the select and responsible few. The dangers of demagoguery – irresponsible leadership of the *plebs* or populace, often coming from isolated members of the elite – were also mentioned by numerous political thinkers during the classical period – from Aristotle to Cicero – in terms not altogether different from those used by some representatives of the modern theory of mass politics.

Classical Underpinnings from Hobbes to Tocqueville

Neither medieval nor Renaissance thought permitted much further development of the seminal ideas about the majority and its relationship to privileged and powerful minorities, much less the rise of a conception of the social world in which egalitarianism was seen as posing a danger to freedom and a source of human bondage. The social upheaval of the seventeenth century, however, inspired Thomas Hobbes in *Leviathan* and especially in his reflections on the disturbances of the English Civil War, the *Behemoth*, to return to ancient themes about the pernicious results of upsetting the hierarchical order of society by demagoguery and mob rule. Political obedience to the sovereign by the many might entail strains and injustice, but anarchy led to much greater evils, both for the body politic and for the preservation of civilization and human life. In *Leviathan*, by simplifying the structure of a viable body politic and making it quite homogenous under the power of a single sovereign, Hobbes pointed to some of the features and problems of government to which theorists of mass politics (i.e., of the political dimension of mass societies) would turn their attention in the twentieth century.

In the history of the concept of mass society, it is worth noting that neither the first fully-fledged interpretation of such phenomenon (Ortega, 1929) nor the first general presentation of anything approaching the notion of a mass society, Alexis de Tocqueville's *Democracy in America* actually used the expression itself. Although Tocqueville's study (1835 and 1840) did not describe the rise of a mass society, his analysis of the dynamics of a liberal, individualistic and democratic polity, clearly pointed out the trends (both manifest and 'subterranean,' to use his characteristic expression) that would eventually lead to it: thus the pressure of a growing 'equality of conditions' in the modern world would make people more similar to each other and therefore erase the distinctions between would-be unique individuals. The powerful ideology of egalitarianism would also have similar repercussions: concessions to majority wishes made by politicians would produce a more mediocre kind of democracy; conformity, rather than orderly dissent, would undermine the vitality of the new culture of the age; hedonism, rather than striving for higher forms of culture and ideals, would inspire the behavior of a selfish citizenry. In a word, a world of lonely, aimless, vulgar pleasure-seeking and disoriented citizens could eventually – though not necessarily – arise from the general trends to be perceived in the age of democracy.

Conservatism and 'Fear of the Masses'

Tocqueville's contribution is important because, for the first time, and almost explicitly, he made a distinction between crowd and mob behavior, on the one hand, and the increasingly uniform structure of advanced democratic societies, on the other. Nevertheless, in the decades that followed, observers tended to concentrate on crowd and mass behavior rather more than on the homogenization (or 'massification' as it later was called) of the entire society.

This interest in crowd behavior came, almost entirely, from the conservative quarter. In their eyes, the progress of democracy during the nineteenth century, especially in the great urban settings of the West, would bring many people – i.e., particularly industrial workers – into the streets to confront governments or to defy public order with demonstrations, strikes, and public meetings. An entire series of observers – mostly the so-called crowd psychologists – produced studies and speculated about the irrational, emotional, and manipulable behavior of the masses. Gustave Le Bon's *The Psychology of Crowds* (1896) is the best-known, though many other similar studies were produced, especially in Italy and England. A best-seller for its time, Le Bon's study brought the concept a step closer to the identification of crowd behavior with the general structure of society, already described by some at the time as 'the age of the masses.' A later essay by Ortega bears a title that still refers to the 'masses' rather than to an 'age' or 'society.' Meanwhile Oswald Spengler's *Decline of the West* ends with a somber and catastrophic view of the future of civilization precisely in terms of an 'age of the masses,' the 'barbarians from below.' For his part, Sigmund Freud wrote one of his most significant studies by developing critically some of Le Bon's ideas and intimated, like Le Bon and Spengler, that the future would belong to the rising masses and their dangerous manipulators. The masses were, by definition, incapable of rational and analytical thought and were also undemocratic in every way.

An alternative, democratic view, of the majority had difficulty emerging. At a given moment, anarchist and socialist thinkers began to accept the notion of 'the masses' in order to refer to the people at large (industrial and peasant workers together) in an entirely different vein to that of conservatives. Yet, with the rise of Bolshevism and, later, Stalinism, the 'masses' came to refer abstractly to a given population as controlled by the single party (the 'vanguard of the proletariat') and its 'cadres.' As such, the 'masses' were only assigned a subordinate role of acquiescence and obedience to the monopolistic party, instead of the autonomy characteristic of the people making up a truly free civil society. It cannot be said that, during the period of (Stalinist) Communist

predominance in large parts of the world (1917–89), observers within the narrow bounds of the official Marxist–Leninist ideology showed any confidence on the 'masses.' Thus, paradoxically, their views coincided with the pessimistic liberals' view of the same period.

Some Marxists or Neo-Marxists would eventually come to accept the term 'mass society,' which they applied mostly to advanced capitalist and 'imperialist' countries. Nevertheless, Western Marxists who escaped the discipline of their Stalinist counterparts often incorporated the concept of mass society to their analyses, as we shall presently see.

The Rise of the Modern Theory of Mass Society

Ortega and Mannheim

As pointed out above, the 'standard' interpretation of the modern world in terms of a mass society found in Ortega's *The Revolt of the Masses* represented the first general synthesis. His essay also included a number of original ideas (e.g., specialism as a form of 'barbarism,' or the notion that 'mass man' is not necessarily linked to the lower classes or to the proletariat). The anxieties and preoccupations of liberals (Tocqueville, John Stuart Mill), philosophers of civilization's decline (Spengler, Burckhardt), crowd and mass psychologists (Le Bon, Freud) elegantly and forcefully converged with Ortega's considerations but did not use the expression. Mannheim's notion of 'society of the masses,' more easily translated into 'mass society,' completed Ortega's vision, with his greater insistence on the bureaucratic and industrial structure of such universe. This was developed in his 1935 *Mensch and Gesellschaft im Zeitalter des Umbaus* (*Man and Society in the Age of Reconstruction* 1940). Mannheim, less conservative than Ortega, put forward possible solutions in a reformist, social democratic spirit.

Translated into English by Edward Shils – himself later a critical analyst of the notion – Mannheim's essay opened the way for the widespread use of the expression 'mass society' in the English-speaking world. In Spanish, many writers retained the plural (*sociedad de masas*), but this has not always been the case in other languages, such as French (*société de masse*) or Italian (*società di massa*; also, for mass politics, *regime di massa*; while Ortega's *hombre masa* was translated into Italian as *uomo di massa*).

Predominance and Expansion of the Mass Society Conception

As Daniel Bell was to point out in his *The End of Ideology* (1960), the mass society conception of advanced societies had become the predominant general vision of the world at the time amongst critics and thoughtful observers, save for the then widespread Marxist views about capitalism and its future. It remained so for a long time, although ramifications, schools of thought, and rival interpretations would also appear. This is not the place for a description nor a classification of such interpretations, however, we should mention that, alongside more conservative (in some cases, overtly pessimistic) views about the rise of a mass society in the framework of Western industrial advanced societies, there emerged some observers who embraced the notion from a left-wing or radical stance. C. W. Mills' attacks against the supposed degeneration of American society and democracy in his two studies *White Collar* (1951) and *The Power Elite* (1956) were typically made from a libertarian position against the development of a mass society on American soil (as a consequence of mass marketing, political manipulation, the undemocratic elite control over the citizenry by a 'military-industrial complex,' etc.)

Likewise, the widely influential neo-Marxist Frankfurt School, on both sides of the Atlantic, embraced the mass society interpretation and had a vast influence upon a wide educated public. Some of its products (*see* Adorno's *The Authoritarian Personality* (1950), which included a re-working of the mass man notion) were significant contributions in terms of social research; other studies, such as Horkheimer and Adorno's 1944 *Dialectic of Enlightenment* (English translation 1972) included a relatively sophisticated 'hegelianization' of the mass society theory. Yet others from the same school – e.g., Marcuse's *One-dimensional Man*, 1964 – popularized the notions of mass society and mass man for a large and quite anxious audience with less philosophical rigor.

More significant perhaps than any division of the notion and theory of mass society into a conservative and a radical version (from the early 1950s to the mid-1970s, by which time it had become truly dominant) was the development of 'specialized' branches of the conception. Two stood out. Firstly, the study of mass politics (further divided into speculations and research on totalitarian mass politics, and the study of pluralistic and democratic mass politics and parties). Secondly, considerable interest was aroused by what Shils had called 'the culture of mass society' (1975a, 1975b). Mass culture and media studies, sociological or otherwise as well as philosophical essays, proliferated and soon became a highly developed field of research and analysis. Most of the debates about the nature and dynamics of mass society took place (or often, raged) within this new field.

Accommodation and Persistence of the Concept

The successful emergence of other general concepts for referring to advanced societies in the most general terms (for instance 'post-industrial societies') either relegated or, in some cases, even drove out the use of the term 'mass society' from its frequent use by the mid-1970s. It did not disappear, however. The mass society conception was anything but a fashion, as its credentials in social philosophy, political theory, and the rigorous critique of civilization demonstrate. The predominant tendency, from the 1980s till the first years of the twenty-first century has seen the mass society notion finding accommodation within a number of other general descriptive concepts (and corresponding interpretations). Thus, along with such concepts as 'post-industrial societies,' we find 'post-capitalist' societies, 'corporate' societies, 'consumer' societies, 'information' societies, and several others.

These expressions are not mutually exclusive. Rather, they refer to different facets or dimensions of the same highly complex phenomenon. To give one example, the widespread interest in the development of corporatism, neocorporatism, and the so-called 'corporate societies' in the 1970s and 1980s demonstrated that theories and interpretations produced by students of these phenomena were often highly compatible

with the mass society theory. The latter's emphasis on administrative control, the rise of large transnational corporations, and anonymous social forces seemed to complement theories of corporatism.

The situation at the turn of the twenty-first century seems to indicate a certain decline, although a number of examples can be produced to demonstrate its persistence. Of all its several branches, the concept, research field, and theories of mass culture and the mass media have not lost – the contrary seems to be the case – any ground at all, either in the West or elsewhere.

More recent general interpretations, notoriously the 'globalization' literature, are not only compatible with the classical tenets of the mass society interpretation but are partly, it can be easily claimed, extensions of it. The notion of mass society always included a view of the expansion, interdependence, and general globalization of the supposed mass characteristics of late industrial Western civilization. The idea of world convergence of mass societies was explicitly built into it.

See also: Authoritarian Personality; Communism, History of; Crowds in History; Mannheim, Karl (1893–1947); Tocqueville, Alexis de (1805–59); Totalitarianism.

Bibliography

Adorno, T.W., 1950. The Authoritarian Personality. Harper Row, New York.
Arendt, H., 1958. The Origins of Totalitarianism. Meridian Books, New York.
Bell, D., 1960. The End of Ideology. Free Press, Glencoe, IL.
Chakhotin, S., 1940. The Rape of the Masses. Routledge, London.
Giner, S., 1976. Mass Society. Academic Press, New York. Martin Robertson, London.
Horkheimer, T.W., Adorno, T., 1972. Dialectic of Enlightenment. Herder & Herder, New York.
Kornhauser, W., 1959. The Politics of Mass Society. Free Press, Glencoe, NY.
Le Bon, G., 1896. The Psychology of Crowds. Alcan, Paris.
Mannheim, K., 1940. Man and Society in the Age of Reconstruction. Kegan Paul, London.
Marcuse, H., 1964. One-dimensional Man. Beacon Press, Boston.
Mills, C.W., 1951. White Collar, The American Middle Classes. Oxford University Press, New York.
Mills, C.W., 1956. The Power Elite. Oxford University Press, New York.
Olson, P. (Ed.), 1964. America as a Mass Society. Free Press, Glencoe, NY.
Ortega y Gasset, J., 1929. La rebelin de las masas. Revista de Occidente, Madrid.
Rosenberg, B., White, D.M. (Eds.), 1957. Mass Culture. Free Press, Glencoe, NY.
Shils, E., 1975a. The stratification system of mass society. In: Shils, E. (Ed.), Center and Periphery: Essays in Macrosociology. University of Chicago Press, Chicago, pp. 304–316.
Shils, E., 1975b. The theory of mass society. In: Shils, E. (Ed.), Center and Periphery: Essays in Macrosociology. University of Chicago Press, Chicago, pp. 91–110.
Vidich, A., Bensman, J., 1960. Small Town in Mass Society. Doubleday, Garden City, NJ.

Mass Trauma: Psychopathological Effects across the Life Span

Frank Neuner, Claudia Catani, and Verena Ertl, Bielefeld University, Bielefeld, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Around the world, many populations are affected by mass trauma events such as natural disasters, wars, or terrorist attacks. Such large-scale events threaten the mental health of adults and children. The major mental health consequences of mass trauma are high prevalence rates of post-traumatic stress disorder and depression. For children, a broad range of maladaptation may result from the exposure to mass trauma events. Research from different such catastrophes has identified risk and resilience factors as well as intervention and prevention programs that meet the mental health needs of the affected populations.

Mass Trauma as a Global Phenomenon

Mass trauma refers to extremely adverse events and conditions that affect the well-being and mental health of large groups or even of whole populations. This relatively new concept allows summarizing and comparing findings on the psychological consequences of large-scale man-made and natural disasters such as wars, terrorist attacks, earthquakes, and floods. For example, in the last two decades, different research groups have gained considerable knowledge about risk factors for psychological disorders of adults and impaired development of children who live in a war-torn country such as Liberia (Johnson et al., 2008), survived the 2004 tsunami disaster in Asia (Neuner et al., 2006), or were affected by the 9/11 terror attack in the US (e.g., Perlman et al., 2011). The goal of mass trauma research is to identify common pathological reactions to such events and to identify risk factors and mechanisms that lead to mental health impairment. In addition, mass trauma research aims to develop and test programs for the prevention of dysfunctional development and the treatment of psychological disorders that result from the events.

While each catastrophe can be perceived as a regional event, a comprehensive understanding of mass trauma phenomena requires the analysis of differences and commonalities of reactions to extreme adversities across the world. Typically, mass trauma events are global phenomena as they raise international attention and assistance and are influenced by international processes such as the world trade of resources or extreme weather events associated with climate change. Low- as well as high-income countries of all continents have been equally affected by natural disasters such as floods or bush fires during the last decade. However, by far the highest number of casualties occurred in low-resource countries, while the highest material damage was caused in high-income countries (International Federation of Red Cross and Red Crescent Societies, 2013). At the same time, the 24 wars and 12 armed conflicts that took place in 2012 were restricted to low- and middle-income countries in the Middle East, Africa, and Asia. Still, most of these conflicts were associated with global economic or security interests. The term mass trauma also gives consideration to the fact that different types of catastrophes often occur simultaneously. For example, the tsunami disaster has affected different war-torn countries in Asia and terrorist attacks commonly occur during armed conflicts.

Psychological Consequences of Mass Trauma Events in Adults

Mass trauma events typically involve the destruction of the existential basis of a large number of the affected families. Whole populations may be forced to flee from their homes, which involves the destruction of social networks and roles. Life after a disaster or during a war involves states of continuous insecurity and dependency. Mass trauma events threaten the subsidence and survival of many people at once. While all of these different components of adversities present extreme stressors for the affected individuals, research on natural disasters as well as on wars has revealed that the exposure to events that involve a direct threat to the life and limb (like witnessing other people dying or experiencing an attack on a village) is the most detrimental aspect of mass trauma (Norris et al., 2002; Steel et al., 2009). Such events are referred to as potentially traumatic events as they can cause post-traumatic stress disorder (PTSD), which is, next to depression, the most common psychological disorder in the aftermath of a mass trauma event. PTSD is characterized by persistent intrusive memories of the traumatic event that keep recurring in the form of vivid images, thoughts, or nightmares, as well as avoidance of reminders of the event and a general state of hyperarousal and anxiety. Different studies have found a wide range of prevalence rates of PTSD following a disaster, and it is likely that methodological limitations including inadequate sampling and assessment explain much of the variance across studies (Steel et al., 2009). Sound studies have found PTSD and major depressive disorder in up to one-third of highly exposed survivors of natural disasters (North and Pfefferbaum, 2013) and up to almost one-half of survivors of war (e.g., Johnson et al., 2008; Karunakara et al., 2004), while far smaller rates of psychopathology were found for the typically less severely exposed populations living in high-income countries (Perlman et al., 2011).

As noted above, the most consistent and strongest risk factor for a trauma-related disorder is the amount of exposure to potentially traumatic events. In the context of mass trauma, several studies have identified a linear dose–effect relationship between traumatic events and probability of PTSD (e.g., Neuner et al., 2004). Next to the direct exposure to potentially traumatic events, different studies have identified a range of other risk factors for a pathological mass trauma reaction, including female gender, preexisting psychopathology, lacking

social support, and secondary adversities resulting from the event such as loss of property, poverty, and an increase of community crimes and threats (Norris, 2005; North and Pfefferbaum, 2013). Although spontaneous recovery from initial trauma reactions is common, longitudinal studies of disaster victims (e.g., Zhang et al., 2013) indicate a chronic course of PTSD in about one-half of those who had developed PTSD.

Psychological Consequences of Mass Trauma in Children and Adolescents

Children and adolescents belong to the most vulnerable groups in the context of mass trauma. Surveys with children growing up in war-torn areas have found prevalence rates up to 75% for PTSD (Catani et al., 2008; Thabet and Vostanis, 2000) and rates of major depression ranging up to 30% (e.g., Thabet et al., 2004). Natural disasters seem to have a similarly detrimental effect on the mental health of children (Neria et al., 2008; Weems et al., 2007). After the Asian tsunami in 2004, studies with affected children in Thailand and Sri Lanka reported high prevalence rates of PTSD, major depression, and suicidal ideation (Neuner et al., 2006; Wickrama and Kaspar, 2007).

Although the vast majority of data on the psychological impact of children's exposure to natural disasters origin from developed countries (Norris et al., 2002), such disasters typically have a much stronger impact on children's development and adjustment in low-income countries. Mass trauma events in low-resource countries are often associated with additional adverse events and conditions such as armed conflict, community and family violence, extreme poverty, and a lack of resources for health provisioning, or even the co-occurrence of mass adversities. For example, the 2004 tsunami catastrophe severely affected regions in Sri Lanka that suffered from an ongoing civil war (Catani et al., 2008). Children are particularly sensitive for the accumulation of stressors that occur in such contexts, and the detrimental developmental effects are not restricted to specific mental health diagnoses such as PTSD and depression. A broader and multifaceted set of outcomes needs to be taken into account when studying children affected by mass trauma. For example, unspecific internalizing as well as externalizing behavior problems have been associated with exposure to natural disasters as well as to war and conflict (Betancourt et al., 2014; McLaughlin et al., 2009).

A Risk-and-Protection Perspective on Children's Psychological Well-Being and Functioning in the Aftermath of Disaster

Many earlier studies and reviews on the effect of extreme adversity presented dose–response gradients to illustrate the detrimental impact of the accumulation traumas on children's psychological well-being. In fact, across diverse countries and disaster types, the severity and frequency of exposure to disaster-related traumatic events was one of the main predictors of psychopathology. For instance, a study with survivors of the Hurricane Mitch found that the level of the Hurricane impact was the strongest predictor of PTSD symptoms in adolescents (Goenjian et al., 2001). In war-torn populations, a linear relationship between exposure and trauma-related disorders has been confirmed in several studies with children exposed to mass trauma (Catani et al., 2008).

However, recent observations indicate that the mere assumption of a direct causal association between mass disaster and trauma-related impairment may be oversimplified. Beyond the individual level of analysis, and following Bronfenbrenner's bioecological system theory (Bronfenbrenner, 1979), it is likely that disaster-related trauma does not affect only isolated individuals but also communities and families. Impairments in family and community functioning may, in turn, impair child adaptation and development through secondary adversities such as parental loss, poverty, homelessness, and community and family violence (Pynoos et al., 1999). In line with this thinking, studies in postconflict settings have revealed elevated levels of family violence against children and adolescents (Catani et al., 2008; Saile et al., 2014). In addition, war and co-occurring family violence have unique effects on symptoms of PTSD as well as on child adaptation (Catani et al., 2010). These findings indicate that it is important to consider the multiplicity of adversities as moderators of impairment (Panter-Brick et al., 2011). Furthermore, a developmental mass trauma perspective needs to identify potentially protective factors within a risk-and-resilience framework (Betancourt and Khan, 2008; Masten and Narayan, 2012). Protective factors or assets consist of individual characteristics of the child's developing biosystem but may also be found on the family as well as community level or in the societal and cultural context (Norris and Stevens, 2007). Given the high impact of the parent–child relationship, the family might not only act as stressor in case of family violence but also foster children's resilience through care and warmth (Masten and Narayan, 2012). Low parental care and high levels of authoritarian and intrusive parenting (overprotection) have been associated with poor psychological adjustment (Enns et al., 2002). There is initial evidence that this relationship might also be valid in war-torn populations on top of the mass trauma context. Betancourt and coworkers (Betancourt et al., 2013) found a correlation between low parental care and high rates of internalizing behavior problems in Sierra Leonean children exposed to the war.

Apart from risk and protective factors related to the family system, potential moderators on other ecological levels might potentiate or buffer the effects of mass trauma on children's mental health. With respect to the individual level, age and gender of the child, specific disabilities, and differences in personality, e.g., self-control or coping skills have been discussed (for a review, see Masten and Narayan, 2012). In recent years, the study of biological variables has become a major research focus, and genetic moderators have been discussed that might potentiate the developmental risk in the context of trauma and disaster (Kim-Cohen and Turkewitz, 2012).

Interventions

Mass trauma events call for humanitarian assistance on different levels. The survivors of disasters and wars often present with the most urgent needs such as food, clean water,

and medical assistance. On the background of such salient shortages, mental health care is usually not perceived as a priority. However, different researchers have pointed to the importance of preventing and curing long-term mental health impairment that may interfere with functioning of individuals and communities. A particular challenge is that mass trauma mainly affects low- and middle-income countries with insufficient mental health service and limited availability of specialized professionals. Nevertheless, some first attempts have been made to develop and test prevention and intervention programs and principles, and some of these programs have shown promising effects in evaluation research (see below).

Mass trauma mental health programs vary with regard to their level of specialization (from universal to highly specialized services), the targeted ecological context (individual, family, school, community), and the timing in relation to the occurring mass trauma (immediate, intermediate, or long term). Universal interventions usually focus on more distal levels like schools and communities, and are often implemented early after incidents of mass trauma with the goal of preventing chronic pathological reactions. On the other hand, more specialized therapeutic interventions are located at more proximal levels, i.e., at the individual and family level. They target specific disorders rather than general distress and deal with identified cases that present with mental health disorders.

Different expert groups (e.g., Hobfoll et al., 2007) have formulated general and universal principles for the evidence-informed mental health response to mass trauma. The authors emphasize the promotion of conditions that are associated with recovery from trauma such as a sense of safety, calming, a sense of self and community efficacy, connectedness, and hope (Hobfoll et al., 2007). The idea is to create an environment that increases resilience and reduces risk factors across specific and unspecific interventions.

School-Based Programs

More specific preventive programs have been developed to support a healthy development of children affected by mass trauma. Most of these programs are school-based and aim at strengthening resilience and reducing symptoms of trauma-related distress in children with and without PTSD. School-based interventions are usually composed of modules such as psychoeducation, sociodrama, movement/dance, playful group cohesion activities, cognitive-behavioral stress inoculation techniques, as well as, in some programs, limited trauma exposure. Prominent examples of such programs are Classroom-Based Intervention (Macy et al., 2003), ERASE-Stress (Gelkopf and Berger, 2009), Overshadowing the Threat of Terrorism (Berger et al., 2007), and Teaching Recovery Techniques (Smith et al., 1999). Systematic evaluations of classroom-based programs in the context of mass trauma have been carried out in schools of Nepal, Indonesia, Sri Lanka, Palestine, and Uganda (Ager et al., 2011; Layne et al., 2008; Tol et al., 2008). The effects on psychological symptoms have often been moderate, but stronger effects have been found on resilience or protective factors, like prosocial behavior and hope (Barry et al., 2013). However, caution is warranted in the widespread implementation of classroom-based interventions in the context of mass disasters, as at least one study has shown that they might actually be harmful for some children as they interfere with natural recovery of the most severely affected children (Tol et al., 2012). An innovative new approach is the attempt to psychologically inoculate children against trauma reactions even before a mass trauma event occurs. Such a procedure is feasible in a conflict context with foreseeable escalation that will bring about threat in near future such as attacks on villages in Israel. Wolmer and colleagues (Wolmer et al., 2011) implemented a variant of stress inoculation training that teaches children techniques to cope with feelings of fear in a threatening situation. The authors show that this program might indeed prevent the onset of PTSD and stress reactions among children effectively when an attack occurs.

Group Psychotherapy

In contrast to classroom-based programs, other authors have adapted established forms of group psychotherapy to the context of mass trauma, in particular for severely affected children and adolescents living in war regions. This attempt requires the identification of single children with symptoms of PTSD or depression before treatment. For example, Bolton et al. (2007) investigated a group variant of interpersonal therapy (IPT) for depression in Northern Ugandan youths. IPT targets four interpersonal areas: (1) grief, (2) interpersonal disputes, (3) role transitions, and (4) interpersonal deficits. However, the effects of group IPT were limited to the reduction of depression symptoms among girls. Recently, McMullen and coworker (McMullen et al., 2013) evaluated a mixed individual and group-based variant of Trauma-Focused Cognitive-Behavioral Therapy (TF-CBT) for former male child soldiers and war-affected boys in the Democratic Republic of Congo (DRC). It was implemented by expert clinicians and included, next to various unspecific methods for stress management and relaxation, specific trauma-focused procedures such as creating a trauma narrative and cognitive processing. The authors showed a significant reduction of PTSD symptoms and comorbid distress. The same group (O'Callaghan et al., 2013) replicated this study with sexually exploited and war-affected girls in the DRC with local social workers as therapists and found promising effects. As a group treatment option for adults, Bass et al. (2013) have adapted and evaluated Cognitive Processing Therapy (CPT) for adult women who were sexually violated in the context of the civil war in the DRC (Bass et al., 2013). CPT aims to modify false beliefs and shattered assumptions related to the traumatic experience in group discussions. CPT effectively reduced anxiety and depression symptoms. Trained local assistants carried out the treatment.

Individual-Level Treatment

Outside the context of mass disasters, different variants of trauma-focused individual psychotherapy have been identified as the first-choice treatments of trauma-related disorders. All these approaches include a confrontation with the trauma memory in the form of the discussion of the meaning of the event or imaginative exposure to the memory, often

through a narrative reconstruction of the traumatic event. Trauma-focused therapies have been adapted to the context of mass trauma and evaluated. Grief and Trauma Intervention (Salloum and Overstreet, 2008) is one such approach for children that includes training in coping skills as well as the creation of a coherent trauma narrative through telling the story of the traumatic events. It has been applied in individual as well as in a group format, including a single individual session for exposure to the worst event. In a dismantling study of the effective ingredients of this treatment with disaster victims, a variant that did not include the trauma narration showed smaller effect sizes than the original treatment, which confirms the importance of the confrontation component of trauma treatment for children affected by mass disasters (Salloum and Overstreet, 2012). A well-established variant of trauma-focused therapy that had been specifically developed for victims of war and torture is Narrative Exposure Therapy (NET; Schauer et al., 2011). This treatment focuses exclusively on the processing of the trauma memory through documenting the life story of the survivors in detail. While NET has been originally developed for adults, it has been adapted to children as KIDNET. Both, the child and the adult version of this short treatment (4–12 sessions) have been tested in different populations (Robjant and Fazel, 2010) including refugees and war-affected populations. As individual treatment, NET might be more costly than a group- or classroom-based approach. However, the efficacy of this treatment, the superiority in comparison to different control conditions, and the feasibility in different contexts including unsafe regions have been well documented.

While current findings on developmental mass trauma reactions indicate the importance of family coping, the study of mass trauma programs that target the family level is still in its infancy. Testing the effects of a brief psychoeducation intervention offered to groups of parents living in a postconflict setting, Jordans and colleagues (Jordans et al., 2013) found beneficial effects on conduct problems among boys. Dybdahl et al. (2001) have found small positive effects of a mother empowerment program on mothers' mental health and the children's psychological functioning.

Taking together the evidence of mass trauma intervention and prevention, the findings argue for a stepped-care approach to prevent and alleviate psychological symptoms in children, adolescents, and adults. The key for any such approach is a systematic screening and triage to assign the affected children and adults to interventions that meet their demands. While the majority of survivors of mass traumas present with low to moderate levels of symptoms and distress and do not require any intervention, children presenting with moderate distress and low-level psychopathology should be assigned to school-based approaches that aim to strengthen their resources and coping skills. However, specialized clinical trauma-focused interventions (e.g., TF-CBT, NET) are required for those who fulfill criteria for mental health disorders, in particular PTSD. At the same time, North and Pfefferbaum (2013) have emphasized that disasters can also foster the reoccurrence and worsening of preexisting psychopathology, which will require a broad and functional mental health-care system next to special trauma interventions. The evidence clearly shows that, even for specialized trauma-focused psychotherapy, the employment of trained lay counselors or lay therapists within a care system that allows supervision and referral is an effective alternative in the absence of clinical professionals.

See also: Armed Conflict and Social Work: Significance and Implications; Community Resilience to Disasters; Disasters in Social, Cultural and Political Context; Hazards and Disaster Research in Contemporary Anthropology; Narrative Exposure Therapy; Natural Disasters: Health-Related Aspects; Post-Traumatic Stress Disorder; Post-Traumatic Stress in Social Work; Psychological Treatments: Randomized Controlled Trials; Torture and its Consequences, Psychology of; War, Political Violence, and Effects on Children.

Bibliography

Ager, A., Akesson, B., Stark, L., Flouri, E., Okot, B., McCollister, F., Boothby, N., 2011. The impact of the school-based psychosocial structured activities (PSSA) program on conflict-affected children in northern Uganda. Journal of Child Psychol Psychiatry Allied Disciplines 52 (11), 1124–1133. http://dx.doi.org/10.1111/j.1469-7610.2011.02407.x.

Barenbaum, J., Ruchkin, V., Schwab-Stone, M., 2004. The psychosocial aspects of children exposed to war: practice and policy initiatives. Journal of Child Psychology Psychiatry 45 (1), 41–62.

Barry, M.M., Clarke, A.M., Jenkins, R., Patel, V., 2013. A systematic review of the effectiveness of mental health promotion interventions for young people in low and middle income countries. BMC Public Health 13, 835. http://dx.doi.org/10.1186/1471-2458-13-835.

Bass, J.K., Annan, J., McIvor Murray, S., Kaysen, D., Griffiths, S., Cetinoglu, T., Bolton, P.A., 2013. Controlled trial of psychotherapy for Congolese survivors of sexual violence. The New England Journal of Medicine 368 (23), 2182–2191. http://dx.doi.org/10.1056/NEJMoa1211853.

Berger, R., Pat-Horenczyk, R., Gelkopf, M., 2007. School-based intervention for prevention and treatment of elementary-students' terror-related distress in Israel: a quasi-randomized controlled trial. Journal of Trauma Stress 20 (4), 541–551. http://dx.doi.org/10.1002/jts.20225.

Betancourt, T.S., Khan, K.T., 2008. The mental health of children affected by armed conflict: protective processes and pathways to resilience. International Review of Psychiatry 20 (3), 317–328.

Betancourt, T.S., McBain, R., Newnham, E.A., Brennan, R.T., 2014. Context matters: community characteristics and mental health among war-affected youth in Sierra Leone. Journal of Child Psychology Psychiatry and Allied Disciplines 55, 217–226. http://dx.doi.org/10.1111/jcpp.12131.

Betancourt, T.S., McBain, R., Newnham, E.A., Brennan, R.T., 2013. Trajectories of internalizing problems in war-affected Sierra Leonean youth: examining conflict and postconflict factors. Child Development 84, 455–470. http://dx.doi.org/10.1111/j.1467-8624.2012.01861.x.

Bolton, P., Bass, J., Betancourt, T., Speelman, L., Onyango, G., Clougherty, K.F., Verdeli, H., 2007. Interventions for depression symptoms among adolescent survivors of war and displacement in northern Uganda: a randomized controlled trial. JAMA: The Journal of the American Medical Association 298 (5), 519–527. http://dx.doi.org/10.1001/jama.298.5.519.

Bronfenbrenner, U., 1979. The Ecology of Human Development: Experiments by Nature and Design. Harvard University Press, Cambridge, MA.

Carrion, V.G., Weems, C.F., Bradley, T., 2010. Natural disasters and the neurodevelopmental response to trauma in childhood: a brief overview and call to action. Future Neurology 5 (5), 667–674.

Catani, C., Gewirtz, A.H., Wieling, E., Schauer, E., Elbert, T., Neuner, F., 2010. Tsunami, war, and cumulative risk in the lives of Sri Lankan schoolchildren. Child Development 81 (4), 1176–1191.

Catani, C., Jacob, N., Schauer, E., Kohila, M., Neuner, F., 2008. Family violence, war, and natural disasters: a study of the effect of extreme stress on children's mental health in Sri Lanka. BMC Psychiatry 8, 33.

Dybdahl, R., 2001. Children and mothers in war: an outcome study of a psychosocial intervention program. Child Development 72 (4), 1214–1230. http://dx.doi.org/10.1111/1467-8624.00343.

Enns, M.W., Cox, B.J., Clara, I., 2002. Parental bonding and adult psychopathology: results from the US National Comorbidity Survey. Psychological Medicine 32 (6), 997–1008.

Gelkopf, M., Berger, R., 2009. A school-based, teacher-mediated prevention program (erase-stress) for reducing terror-related traumatic reactions in Israeli youth: a quasi-randomized controlled trial. Journal of Child Psychology and Psychiatry, and Allied Disciplines 50 (8), 962–971. http://dx.doi.org/10.1111/j.1469-7610.2008.02021.x.

Goenjian, A.K., Molina, L., Steinberg, A.M., Fairbanks, L.A., Alvarez, M.L., Goenjian, H.A., et al., 2001. Posttraumatic stress and depressive reactions among Nicaraguan adolescents after hurricane Mitch. The American Journal of Psychiatry 158 (5), 788–794.

Green, B.L., 1996. Traumatic stress and disaster: mental health effects and factors influencing adaptation. International Review of Psychiatry 2, 177–210.

Hobfoll, S.E., Watson, P., Bell, C.C., Bryant, R.A., Brymer, M.J., Friedman, M.J., Ursano, R.J., 2007. Five essential elements of immediate and mid-term mass trauma intervention: empirical evidence. Psychiatry 70 (4), 283–315.

International Federation of Red Cross and Red Crescent Societies, 2013. World Disasters Report 2013. International Federation of Red Cross and Red Crescent Societies, Geneva. Retrieved from Google Scholar.

Johnson, K., Asher, J., Rosborough, S., Raja, A., Panjabi, R., Beadling, C., Lawry, L., 2008. Association of combatant status and sexual violence with health and mental health outcomes in postconflict Liberia. JAMA: The Journal of the American Medical Association 300 (6), 676–690. http://dx.doi.org/10.1001/jama.300.6.676.

Jordans, M.J., Tol, W.A., Ndayisaba, A., Komproe, I.H., 2013. A controlled evaluation of a brief parenting psychoeducation intervention in Burundi. Social Psychiatry and Psychiatric Epidemiology 48 (11), 1851–1859. http://dx.doi.org/10.1007/s00127-012-0630-6.

Karunakara, U.K., Neuner, F., Schauer, M., Singh, K., Hill, K., Elbert, T., Burnha, G., 2004. Traumatic events and symptoms of post-traumatic stress disorder amongst Sudanese nationals, refugees and Ugandans in the West Nile. African Health Sciences 4 (2), 83–93.

Kim-Cohen, J., Turkewitz, R., 2012. Resilience and measured gene-environment interactions. Development and Psychopathology 24 (4), 1297–1306.

Layne, C.M., Saltzman, W.R., Poppleton, L., Burlingame, G.M., Pasalić, A., Duraković, E., Pynoos, R.S., 2008. Effectiveness of a school-based group psychotherapy program for war-exposed adolescents: a randomized controlled trial. Journal of the American Academy of Child and Adolescent Psychiatry 47 (9), 1048–1062. http://dx.doi.org/10.1097/CHI.0b013e31817eecae.

Macy, R.D., Macy, D.J., Gross, S.I., Brighton, P., 2003. Healing in familiar settings: support for children and youth in the classroom and community. New Directions for Youth Development 98, 51–79. http://dx.doi.org/10.1002/yd.44.

Masten, A.S., Narayan, A.J., 2012. Child development in the context of disaster, war, and terrorism: pathways of risk and resilience. Annual Review of Psychology 63, 227–257.

McLaughlin, K.A., Fairbank, J.A., Gruber, M.J., Jones, R.T., Lakoma, M.D., Pfefferbaum, B., et al., 2009. Serious emotional disturbance among youths exposed to Hurricane Katrina 2 years postdisaster. Journal of the American Academy of Child and Adolescent Psychiatry 48 (11), 1069–1078.

McMullen, J., O'Callaghan, P., Shannon, C., Black, A., Eakin, J., 2013. Group trauma-focused cognitive-behavioural therapy with former child soldiers and other war-affected boys in the DR Congo: a randomised controlled trial. Journal of Child Psychology and Psychiatry, and Allied Disciplines 54 (11), 1231–1241. http://dx.doi.org/10.1111/jcpp.12094.

Neria, Y., Nandi, A., Galea, S., 2008. Post-traumatic stress disorder following disasters: a systematic review. Psychological Medicine 38 (4), 467–480.

Neuner, F., Schauer, E., Catani, C., Ruf, M., Elbert, T., 2006. Post-tsunami stress: a study of posttraumatic stress disorder in children living in three severely affected regions in Sri Lanka. Journal of Trauma Stress 19 (3), 339–347.

Neuner, F., Schauer, M., Karunakara, U., Klaschik, C., Robert, C., Elbert, T., 2004. Psychological trauma and evidence for enhanced vulnerability for posttraumatic stress disorder through previous trauma among West Nile refugees. BMC Psychiatry 4, 34.

Norris, F.H., 2005. Psychosocial Consequences of Natural Disasters in Developing Countries: What Does Past Research Tell Us about the Potential Effects of the 2004 Tsunami? Dept. of Veterans Affairs. Retrieved from Google Scholar.

Norris, F.H., Stevens, S.P., 2007. Community resilience and the principles of mass trauma intervention. Psychiatry: Interpersonal and Biological Processes 70 (4), 320–328. Retrieved from Google Scholar.

Norris, F.H., Friedman, M.J., Watson, P.J., Byrne, C.M., Diaz, E., Kaniasty, K., 2002. 60,000 disaster victims speak: Part I. An empirical review of the empirical literature, 1981-2001. Psychiatry 65 (3), 207–239.

North, C.S., Pfefferbaum, B., 2013. Mental health response to community disasters: a systematic review. JAMA: The Journal of the American Medical Association 310 (5), 507–518. http://dx.doi.org/10.1001/jama.2013.107799.

O'Callaghan, P., McMullen, J., Shannon, C., Rafferty, H., Black, A., 2013. A randomized controlled trial of trauma-focused cognitive behavioral therapy for sexually exploited, war-affected Congolese girls. Journal of the American Academy of Child and Adolescent Psychiatry 52 (4), 359–369. http://dx.doi.org/10.1016/j.jaac.2013.01.013.

Panter-Brick, C., Goodman, A., Tol, W., Eggerman, M., 2011. Mental health and childhood adversities: a longitudinal study in Kabul, Afghanistan. Journal of the American Academy of Child and Adolescent Psychiatry 50 (4), 349–363.

Perlman, S.E., Friedman, S., Galea, S., Nair, H.P., Eros-Sarnyai, M., Stellman, S.D., Greene, C.M., 2011. Short-term and medium-term health effects of 9/11. Lancet 378 (9794), 925–934. http://dx.doi.org/10.1016/S0140-6736(11)60967-7.

Pynoos, R.S., Steinberg, A.M., Piacentini, J.C., 1999. A developmental psychopathology model of childhood traumatic stress and intersection with anxiety disorders. Biological Psychiatry 46 (11), 1542–1554.

Robjant, K., Fazel, M., 2010. The emerging evidence for narrative exposure therapy: a review. Clinical Psychology Review 30 (8), 1030–1039. http://dx.doi.org/10.1016/j.cpr.2010.07.004.

Saile, R., Ertl, V., Neuner, F., Catani, C., 2014. Does war contribute to family violence against children? Findings from a two-generational multi-informant study in Northern Uganda. Child Abuse & Neglect 38 (1), 135–146. http://dx.doi.org/10.1016/j.chiabu.2013.10.007.

Salloum, A., Overstreet, S., 2008. Evaluation of individual and group grief and trauma interventions for children post disaster. Journal of Clinical Child and Adolescent Psychology 37 (3), 495–507. http://dx.doi.org/10.1080/15374410802148194.

Salloum, A., Overstreet, S., 2012. Grief and trauma intervention for children after disaster: exploring coping skills versus trauma narration. Behaviour Research and Therapy 50 (3), 169–179. http://dx.doi.org/10.1016/j.brat.2012.01.001.

Schauer, M., Neuner, F., Elbert, T., 2011. Narrative Exposure Therapy, second ed. Hogrefe, Seattle.

Smith, P., Dyregrov, A., Yule, W., 1999. Children and disaster: teaching recovery techniques. Children and War Foundation, Bergen, Norway.

Steel, Z., Chey, T., Silove, D., Marnane, C., Bryant, R.A., van Ommeren, M., 2009. Association of torture and other potentially traumatic events with mental health outcomes among populations exposed to mass conflict and displacement: a systematic review and meta-analysis. JAMA: The Journal of the American Medical Association 302 (5), 537–549. http://dx.doi.org/10.1001/jama.2009.1132.

Thabet, A.A., Vostanis, P., 2000. Post traumatic stress disorder reactions in children of war: a longitudinal study. Child Abuse & Neglect 24 (2), 291–298.

Thabet, A.A., Abed, Y., Vostanis, P., 2004. Comorbidity of PTSD and depression among refugee children during war conflict. Journal of Child Psychology Psychiatry 45 (3), 533–542.

Tol, W.A., Komproe, I.H., Jordans, M.J., Vallipuram, A., Sipsma, H., Sivayokan, S., DE Jong, J.T., 2012. Outcomes and moderators of a preventive school-based mental health intervention for children affected by war in Sri Lanka: a cluster randomized trial. World Psychiatry: Official Journal of the World Psychiatric Association (WPA) 11 (2), 114–122.

Tol, W.A., Komproe, I.H., Susanty, D., Jordans, M.J., Macy, R.D., De Jong, J.T., 2008. School-based mental health intervention for children affected by political violence in Indonesia: a cluster randomized trial. JAMA: The Journal of the American Medical Association 300 (6), 655–662. http://dx.doi.org/10.1001/jama.300.6.655.

Weems, C.F., Pina, A.A., Costa, N.M., Watts, S.E., Taylor, L.K., Cannon, M.F., 2007. Predisaster trait anxiety and negative affect predict posttraumatic stress in youths after Hurricane Katrina. Journal of Consulting and Clinical Psychology 75 (1), 154.

Wickrama, K.A., Kaspar, V., 2007. Family context of mental health risk in Tsunami-exposed adolescents: findings from a pilot study in Sri Lanka. Social Science & Medicine 64 (3), 713–723.

Wolmer, L., Hamiel, D., Laor, N., 2011. Preventing children's posttraumatic stress after disaster with teacher-based intervention: a controlled study. Journal of Clinical Child and Adolescent Psychology 50 (4), 340–348.

Zhang, G., North, C.S., Narayanan, P., Kim, Y.S., Thielman, S., Pfefferbaum, B., 2013. The course of postdisaster psychiatric disorders in directly exposed civilians after the US embassy bombing in Nairobi, Kenya: a follow-up study. Social Psychiatry and Psychiatric Epidemiology 48 (2), 195–203. http://dx.doi.org/10.1007/s00127-012-0535-4.

Mastery Learning

Thomas R Guskey, College of Education, University of Kentucky, Lexington, KY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Over the last four decades, few programs have been implemented as broadly or evaluated as thoroughly as those associated with mastery learning. This article describes how mastery learning originated and the essential elements involved in its implementation. It discusses the improvements in student learning that typically result from the use of mastery learning and how this strategy provides practical solutions to a variety of persistent instructional problems. Finally it explains the common misinterpretations of mastery learning and summarizes the results of research on its effects.

Over the last four decades, few programs have been implemented as broadly or evaluated as thoroughly as those associated with mastery learning. Programs based on mastery learning principles operate today in nations throughout the world and at every level of education. When compared to traditionally taught classes, students in mastery learning classes consistently have been shown to learn better, reach higher levels of achievement, and develop greater confidence in their ability to learn and in themselves as learners (Anderson, 1994; Guskey and Pigott, 1988; Klecker and Chapman, 2008; Kulik et al., 1990a; Miles, 2010).

This article describes how mastery learning originated and the essential elements involved in its implementation. It discusses the improvements in student learning that typically result from the use of mastery learning and how this strategy provides practical solutions to a variety of persistent instructional problems. Finally it explains the common misinterpretations of mastery learning and summarizes the results of research on its effects.

John B. Carroll's 'Model for School Learning'

Although the basic tenets of mastery learning can be traced to such early educators as Comenius, Pestalozzi, and Herbart, modern versions were inspired by a 1963 article by John B. Carroll of Harvard University entitled, 'A Model for School Learning.' In his article, Carroll challenged long-held perceptions of student *aptitude*. He pointed out that student aptitude traditionally had been viewed as the *level* to which a child could learn a particular subject. Children with high aptitude would be able to learn the most complex aspects of that subject, while those with low aptitude would be able to learn only the most basic elements. When aptitude is viewed in this way, children are seen as either good learners (high aptitude) or poor learners (low aptitude) with regard to the subject.

Carroll argued, however, that student aptitude more accurately reflects an indicator of *learning rate*. That is, all children have the potential to learn quite well, but differ primarily in terms of the time they require to do so. Some children are able to learn a subject very quickly while others may take much longer. When aptitude is viewed as an indicator of learning rate, children are seen not simply as good and poor learners, but rather as fast and slow learners.

Carroll then proposed a model for school learning based on this alternative view of aptitude. He believed that if each child was allowed the time needed to learn a subject to some criterion level, and if the child spent that time appropriately, then the child probably would attain the specified level of achievement. But if not enough time were allowed or if the child did not spend the time required, then the child would learn much less. The degree of learning attained by a child, therefore, could be expressed by the following simple equation:

$$\text{Degree of learning} = f\left(\frac{\text{time spent}}{\text{time needed}}\right)$$

In other words, degree of learning is a function of the time a child actually spends on learning, relative to the time he or she needs to spend. If the time spent was equal to time needed, the learning would be complete and the equation would equal 1. If the time spent was less than the time needed, however, the learning would be incomplete by that proportion.

Carroll further identified the factors that, he believed, influenced the time spent and the time needed. He argued that both of these elements were affected by characteristics of the learner and by characteristics of the instruction. Specifically, he believed that the time spent was determined by a child's *perseverance* and the *opportunity to learn*. Perseverance is simply the amount of time a child is willing to spend actively engaged in learning. Opportunity to learn is the classroom time allotted to the learning. In other words, time spent is determined by the child's persistence at a learning task and the amount of learning time provided. Time needed, on the other hand, Carroll believed was determined by the child's *learning rate* for that subject, the *quality of the instruction*, and the child's *ability to understand the instruction*. Specifically:

$$\text{Degree of learning} = f\left(\frac{\begin{array}{l} * \text{ perseverance} \\ * \text{ opportunity to learn} \end{array}}{\begin{array}{l} * \text{ learning rate} \\ * \text{ quality of instruction} \\ * \text{ ability to understand the instruction} \end{array}}\right)$$

Again, a child's learning rate is a measure of the time required by the child to learn the concepts or skills under ideal instructional conditions. If the quality of the instruction was high, then the child would readily understand it and would

probably need little time to learn. If the quality of the instruction was not as high, however, then the child would have greater difficulty understanding and would require much more time to learn. In other words, the quality of the instruction and the child's ability to understand the instruction interact to determine how much time is needed for the child to learn the concepts or skills.

Carroll's article made a significant contribution to learning theory. It set forth new guidelines for research into the concept of aptitude and identified specific factors that influence learning in school settings. His ideas about learning rate also prompted the development of numerous 'individualized instruction' programs that allowed students to progress through a series of learning units at their own, self-determined pace. Two of the best known of these 'continuous progress' programs were Individually Prescribed Instruction, developed at the University of Pittsburgh (Glaser, 1966) and Individually Guided Education, developed at the University of Wisconsin (Klausmeier et al., 1968). Carroll himself, however, did not address the problem of how to provide sufficient time or how to improve instructional quality. These issues were left unresolved.

Benjamin S. Bloom's 'Learning for Mastery'

Benjamin S. Bloom, Distinguished Service Professor at the University of Chicago, was impressed by the optimism of Carroll's perspective on learners and particularly by the idea that students differ in terms of the *time required* for learning rather than their *ability to learn*. If aptitude was indeed predictive of the time a child would require to learn, Bloom believed it should be possible to set the degree of learning expected of each child at some mastery performance level. Then by attending to the instructional variables under teachers' control – the opportunity to learn and the quality of the instruction – teachers should be able to ensure that every child attain that specified level.

To determine how this might be practically achieved, Bloom first considered how teaching and learning take place in typical group-based classroom settings. He observed that most teachers begin by dividing the concepts and skills that they want students to learn into smaller learning units. These units are usually sequentially ordered and often correspond to the chapters in the textbook used in teaching. Teachers then teach the unit concepts to all students in the same way, provide all students with the same amount of time to learn, and evaluate students' learning at the end with some form of test or assessment. The few students for whom the instructional methods and time were ideal learn excellently and perform well on the unit assessment. The largest number of students for whom the methods and time were only moderately appropriate learns less well. And students for whom the instruction and time were inappropriate due to differences in their backgrounds or learning styles, learn very little and perform poorly on the unit assessment. In other words, little variation in the teaching resulted in great variation in student learning. Under these conditions the pattern of student achievement was similar to the normal curve distribution shown in Figure 1.

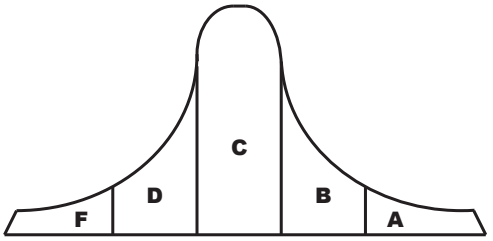

Figure 1 Distribution of achievement in traditional classrooms.

To attain better results and *reduce* this variation in student achievement, Bloom reasoned that we would have to *increase* variation in the teaching. In other words, because students varied in their learning styles and aptitudes, teachers must diversify and differentiate instruction to better meet students' individual learning needs. The challenge was to find practical ways to do this within the context of group-based classrooms so that *all* students learn well.

In searching for such a strategy, Bloom drew primarily from two sources of evidence. First, he considered the ideal teaching and learning situation in which an excellent tutor is paired with each student. He was particularly influenced by the work of early pioneers in individualized instruction, especially Washburne (1922) and his Winnetka Plan, and Morrison (1926) and his University of Chicago Laboratory School experiments. In examining this evidence, Bloom sought to determine what crucial elements in one-to-one tutoring and individualized instruction could be transferred to group-based classroom settings.

Second, Bloom looked at studies of the learning strategies of academically successful students, especially the work of Dollard and Miller (1950). From this research he tried to identify the activities of high-achieving students in group-based classrooms that distinguish them from their less successful classmates.

Bloom saw value in teachers' traditional practice of organizing the concepts and skills they want students to learn into learning units. He also considered it important for teachers to assess student learning at the end of each unit. But the classroom assessments most teachers used seemed to do little more than show for whom their initial instruction was and was not appropriate.

Bloom believed that a far better approach would be for teachers to use their classroom assessments as learning tools, and then to follow those assessments with a *feedback and corrective* procedure. In other words, instead of using assessments only as evaluation devices that mark the end of each unit, Bloom recommended using them as part of the instructional process to *diagnose* individual learning difficulties (feedback) and to *prescribe* remediation procedures (correctives).

This is precisely what takes place when an excellent tutor works with an individual student. If the student makes an error, the tutor first points out the error (feedback) and then follows up with further explanation and clarification (correctives) to ensure the student's understanding. Similarly, academically successful students typically follow up the mistakes they make on quizzes and assessments. They ask the teacher about the items they missed, look up the answer in the textbook or other

resources, or rework the problem or task so that they do not repeat those errors.

With this in mind, Bloom outlined an instructional strategy to make use of this feedback and corrective procedure, labeling it 'learning for mastery' (Bloom, 1968), and later shortening it to simply 'mastery learning' (Bloom, 1971a). With this strategy, teachers first organize the concepts and skills they want students to learn into learning units that typically involve about a week or two of instructional time. Following initial instruction on the unit, teachers administer a brief quiz or assessment based on the unit's learning goals. Instead of signifying the end of the unit, however, this assessment's purpose is to give students information, or 'feedback,' on their learning. To emphasize this new purpose Bloom suggested calling it a *formative assessment*, meaning 'to inform or provide information.' A formative assessment identifies for students precisely what they have learned well to that point, and what they need to learn better (Bloom et al., 1971).

Paired with each formative assessment are specific 'corrective' activities for students to use in correcting their learning difficulties. Most teachers match these 'correctives' to each item or set of prompts within the assessment so that students need work on only those concepts or skills not yet mastered. In other words, the correctives are 'individualized.' They may point out additional sources of information on a particular concept, such as page numbers in the textbook or workbook where the concept is discussed. They may identify alternative learning resources such as different textbooks, learning activities, alternative materials, CDs, digital lessons, or web-based instructional materials (DeWeese and Randolph, 2011). Or they may simply suggest sources of additional practice, such as study guides, computer exercises, independent or guided practice activities, or collaborative group activities.

With the feedback and corrective information gained from the formative assessment, each student has a detailed prescription of what more needs to be done to master the concepts or skills from the unit. This 'just-in-time' correction prevents minor learning difficulties from accumulating and becoming major learning problems. It also gives teachers a practical means to vary and differentiate their instruction in order to better meet students' individual learning needs. As a result, many more students learn well, master the important learning goals in each unit, and gain the necessary prerequisites for success in subsequent units.

When students complete their corrective activities after a class period or two, Bloom recommended they take a *second* formative assessment. This second, 'parallel' assessment covers the same concepts and skills as the first, but is composed of slightly different problems or questions, and serves two important purposes. First, it verifies whether or not the correctives were successful in helping students overcome their individual learning difficulties. Second, it offers students a second chance at success and, hence, has powerful motivational value (Changeiywo et al., 2011).

Some students, of course, will perform well on the first assessment, demonstrating that they have mastered the unit concepts and skills. The teacher's initial instruction was highly appropriate for these students and they have no need of corrective work. To ensure their continued learning progress, Bloom recommended that teachers provide these students with special 'enrichment' or 'extension' activities to broaden their learning experiences. Enrichment activities are often self-selected by students and might involve special projects or reports; digital academic games; or any variety of complex, problem-solving tasks. Figure 2 illustrates this instructional sequence.

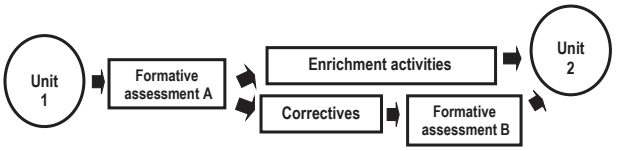

Figure 2 The mastery learning instructional process.

Through this process of formative classroom assessment, combined with the systematic correction of individual learning difficulties, Bloom believed all students could be provided with a more appropriate quality of instruction than is possible under more traditional approaches to teaching. As a result, nearly all might be expected to learn well and truly master the unit concepts or learning goals (Bloom, 1976). This, in turn, would drastically reduce the variation in students' achievement levels, eliminate achievement gaps, and yield a distribution of achievement more like that shown in Figure 3.

In describing mastery learning, however, Bloom emphasized that reducing variation in students' achievement does not imply making all students the same. Even under these more favorable learning conditions, some students undoubtedly will learn more than others, especially those involved in enrichment activities. But by recognizing relevant, individual differences among students and then altering instruction to better meet their diverse learning needs, Bloom believed the variation among students in how well they learn specific concepts or master a set of articulated learning goals could eventually reach a 'vanishing point' (Bloom, 1971b). In other words, *all* students would be helped to learn well the knowledge and skills prescribed in the curriculum. As a result, gaps in the achievement of different groups of students would be closed (Guskey, 2007).

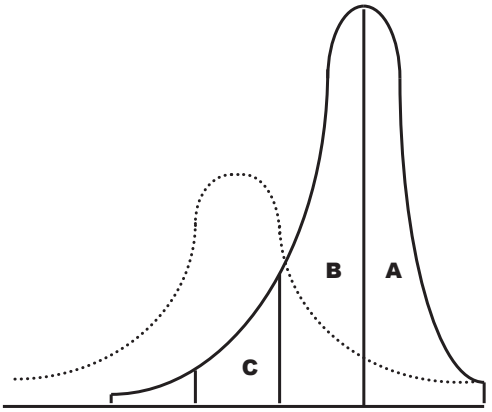

Figure 3 Distribution of achievement in mastery learning classrooms.

The Essential Elements of Mastery Learning

After Benjamin Bloom described his ideas, several of his students took up the task of clarifying mastery learning instructional strategies (e.g., Block et al., 1989) and numerous programs based on mastery learning principles sprung up in schools throughout the United States and around the world (see Postlethwaite and Haggarty, 1998; Reezigt and Weide, 1990; Wu, 1994; Yildiran, 2006; Yildiran and Aydin, 2005). Although differing from setting to setting, those programs true to Bloom's ideas include two essential elements: (1) the *feedback, corrective, and enrichment process*; and (2) *instructional alignment* (Guskey, 1997a).

Feedback, Correctives, and Enrichment

Teachers who use mastery learning provide students with frequent and specific *feedback* on their learning progress through regular, formative classroom assessments. This feedback is both diagnostic and prescriptive. It reinforces precisely what students were expected to learn, identifies what they learned well, and describes what needs to be learned better. The US National Council of Teachers of Mathematics (NCTM) emphasizes this same element in its latest iteration of the standards for school mathematics. To overcome inequities in mathematics instruction, NCTM stresses the use of assessments that support learning and provide useful information to both teachers and students (NCTM, 2000).

Feedback alone, however, does little to help students improve their learning. Significant improvement requires that feedback be paired with *correctives*: activities that offer guidance and direction to students on how to remedy their learning problems. Because of students' individual differences, no single method of instruction works best for all. To help every student learn well, therefore, teachers must differentiate their instruction, both in their initial teaching and especially through the corrective activities (Bloom, 1976). In other words, teachers must *increase* variation in their teaching to *decrease* variation in results.

To be effective, correctives must be qualitatively different from the initial teaching. They must provide students who need it with an alternative approach and additional time to learn. The best correctives present concepts differently and involve students in learning differently than did the initial instruction. They incorporate different learning styles, learning modalities, or types of intelligence. Although developing effective correctives can prove challenging, many schools find that providing teachers with time to work collaboratively, sharing ideas, materials, and expertise, greatly facilitate the process (Guskey, 2008).

Most applications of mastery learning also include *enrichment* or *extension* activities for students who master the unit concepts from the initial teaching. As described above, enrichment activities offer students exciting opportunities to broaden and expand their learning. They reward students for their learning success and challenge them to go further. Many teachers draw from activities developed for gifted and talented students when planning enrichment activities, both to simplify implementation tasks and to guarantee these students a high-quality learning experience.

Teachers implement the feedback, corrective, and enrichment process in a variety of ways. Many use short, paper-and-pencil quizzes as formative assessments to give students feedback on their learning progress. But formative assessments also can take the form of essays, compositions, projects, reports, performance tasks, skill demonstrations, oral presentations, or any device used to gain evidence on students' learning progress. In essence, teachers adapt the format of their formative assessments to match their instructional goals.

Following a formative assessment, some teachers divide the class into separate corrective and enrichment groups. While the teacher directs corrective activities, guaranteeing that all students who need the extra time and assistance take part, the other students work on self-selected, independent enrichment activities. Other teachers pair with colleagues and use a team-teaching approach. While one teacher oversees corrective activities, the other monitors enrichments. Still other teachers use cooperative learning activities in which students work together in teams to ensure all reach the mastery level. Since students have their own personal scores on the formative assessment, individual accountability is assured. Offering the entire team special recognition or credit if all members attain mastery on the second formative assessment encourages group responsibility.

Feedback, corrective, and enrichment procedures are crucial to mastery learning, for it is through these procedures that mastery learning differentiates and individualizes instruction. In every learning unit, students who need extended time and opportunity to remedy learning problems receive these through the correctives. Students who learn quickly and find the initial instruction highly appropriate have opportunities to extend their learning through enrichment. As a result, all students experience more favorable learning conditions and more appropriate, higher-quality instruction (Bloom, 1977; Guskey, 2010).

Instructional Alignment

While feedback, correctives, and enrichment are important, they alone do not constitute mastery learning. To be truly effective, Bloom stressed they must be combined with the second essential element of mastery learning: *instructional alignment*. Reducing variation in student learning and closing achievement gaps require clarity and consistency among all instructional components (Bloom, 1971a).

Bloom believed three major components compose the teaching and learning process. To begin there must be specific ideas about what students are expected to learn and be able to do, that is, *learning goals or standards*. Next comes *instruction* that, ideally, results in *proficient learners* – students who have learned well and whose proficiency can be assessed through some form of assessment or evaluation. Mastery learning adds a feedback and corrective component, allowing teachers to determine for whom their initial instruction was appropriate and for whom an alternative approach may be needed.

Although essentially neutral with regard to what is taught, how it is taught, and how learning is assessed or evaluated, mastery learning requires consistency or alignment among these instructional components, as shown in **Figure 4**. For example, if students are expected to learn higher-level skills such as those involved in making applications, solving complex problems, or developing thoughtful analyses, mastery learning stipulates that instructional activities must be planned

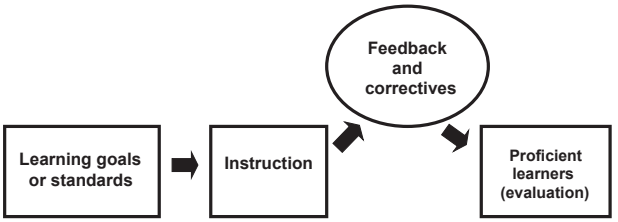

Figure 4 Major components in the teaching and learning process.

to give students opportunities to practice and actively engage in those skills. It also requires that students be given specific feedback on how well they have learned those skills, coupled with directions on how to correct any learning errors. Finally, procedures for assessing or evaluating students' learning should reflect those higher-level skills as well.

To ensure alignment among instructional components, teachers must make a number of crucial decisions. First, they need to decide what concepts or skills are most important for students to learn and most central to students' understanding. Teachers must determine, for example, if they want students to learn only basic skills, or if they want students to develop higher-level skills and more complex cognitive processes. Second, teachers need to decide what evidence best reflects students' mastery of those basic or higher-level skills. Critics sometimes challenge teachers' abilities to make these crucial decisions. But, in essence, teachers at all levels already make these decisions in conducting regular classroom activities. Every time they administer an assessment, grade a paper, or evaluate students' learning, teachers communicate to students what is most important to learn. Using mastery learning simply compels teachers to make these decisions more thoughtfully, intentionally, and purposefully.

Misinterpretations of Mastery Learning

Shortly after Bloom presented his ideas on mastery learning, misinterpretations began to occur. Some early attempts to apply mastery learning, for example, were based on narrow and inaccurate understandings of Bloom's theory. These efforts focused only on low-level cognitive skills, attempted to break learning down into small segments, and insisted that students 'master' each segment before being permitted to move on. Teachers were regarded in these programs as little more than managers of materials and record-keepers of student progress. Unfortunately, similar misinterpretations of mastery learning continue today.

Nowhere in Bloom's writing, however, can this kind of narrowness and rigidity be found. In fact, Bloom emphasized quite the opposite. He considered thoughtful and reflective teachers vital to the successful implementation of mastery learning and continually stressed flexibility in its application. In his earliest description of the process Bloom wrote:

> There are many alternative strategies for mastery learning. Each strategy must find some way of dealing with individual differences in learners through some means of relating the instruction to the needs and characteristics of the learners. … The nongraded school is one attempt to provide an organizational structure that permits and encourages mastery learning.
>
> Bloom, 1968, pp. 7–8

Bloom further emphasized his belief that instruction in mastery learning classrooms should focus on higher-level learning goals, not simply basic skills. He noted:

> I find great emphasis on problem solving, applications of principles, analytical skills, and creativity. Such higher mental processes are emphasized because this type of learning enables the individual to relate his or her learning to the many problems he or she encounters in day-to-day living. These abilities are stressed because they are retained and utilized long after the individual has forgotten the detailed specifics of the subject matter taught in the schools. These abilities are regarded as one set of essential characteristics needed to continue learning and to cope with a rapidly changing world.
>
> Bloom, 1978, p. 578

Modern research studies have shown mastery learning to be particularly effective when applied to instruction focusing on higher-level learning goals such as problem solving, drawing inferences, deductive reasoning, and creative expression (Arredondo and Block, 1990; Kozlovsky, 1990; Mevarech, 1981; Zimmerman and Dibenedetto, 2008). When well implemented, the process helps teachers improve student learning and close achievement gaps in a broad range of learning goals from basic skills to highly complex cognitive processes.

Another misinterpretation comes from educators who believe that the constraint of limited class time will inhibit efforts to implement mastery learning (Guskey, 1997a). They assume that the introduction of feedback, corrective, and enrichment procedures will reduce the amount of material teachers will be able to cover. In other words, teachers will have to sacrifice *coverage* for the sake of *mastery*. But as discussed earlier, minor alterations in instructional pacing typically resolve this concern.

Early mastery learning units usually do require more time. Teachers who allow class time for students to complete corrective activities often find themselves 'behind' other teachers who teach in more traditional ways after the first two or three units. But once students become familiar with the process, mastery learning teachers generally pick up their pace. Since students in mastery learning classes spend a larger portion of their time actively engaged in learning, they make more rapid progress than students in more traditionally taught classes (Fitzpatrick, 1985). As students catch on to the process, they also do better on first formative assessments. With fewer students involved in correctives and less corrective work needed, teachers reduce the class time allocated to corrective activities. And because mastery learning students learn well the concepts and skills from early units, they are better prepared for later, more advanced units. Instruction in later units can therefore be more rapid and include fewer review activities. Most teachers discover that with slight adjustments in the pacing of their instruction – slightly more time spent in early units but less time in later ones – they can cover just as much material using mastery learning, and in

some cases more, as they were able to using more traditional approaches to instruction (Block, 1983).

Research Results and Implications

Teachers who implement mastery learning generally find that it requires only modest changes in their instructional procedures. Excellent teachers use many aspects of mastery learning in their classes already, and others discover that the process blends well with their current teaching strategies. This makes mastery learning particularly attractive to teachers at all levels, especially considering the difficulties associated with new approaches that require major changes in teaching procedures.

Despite the modest nature of these changes, however, extensive research evidence shows that the use of mastery learning can have exceptionally positive effects on student learning (Block et al., 1989; Guskey and Pigott, 1988; Miles, 2010). Evidence gathered in Africa (Changeiywo et al., 2011), Asia (Wu, 1994; Yildiran and Aydin, 2005), Australia (Chan, 1981), Europe (Langeheine, 1992; Mevarech, 1985; Reezigt and Weide, 1990), South America (Cabezon, 1984), and the United States (Walberg, 1990) shows the careful and systematic application of these elements can lead to significant improvements in student learning outcomes. Some researchers even suggest that the superiority of Japanese students in international comparisons of achievement in mathematics operations and problem solving may be due largely to the widespread use in Japan of instructional practices similar to mastery learning (Nakajima, 2012; Waddington, 1995).

Long-term investigations have yielded similarly impressive results. A study by Whiting et al. (1995), for example, representing 18 years of data gathered from over 7000 high school students showed mastery learning to have remarkably positive influence on students' test scores and grade-point averages as well as their attitudes toward school and learning. Another field experiment conducted in elementary and middle school classrooms showed that the implementation of mastery learning led to significantly positive increases in students' academic achievement and their self-confidence (Anderson et al., 1992). Even more impressive, a comprehensive, meta-analysis review of the research on mastery learning by Kulik et al. (1990a) concluded:

> We recently reviewed meta-analyses in nearly 40 different areas of educational research (J. Kulik and Kulik, 1989). Few educational treatments of any sort were consistently associated with achievement effects as large as those produced by mastery learning. … In evaluation after evaluation, mastery programs have produced impressive gains.
>
> Kulik et al., 1990a, p. 292

Similarly, in *Research on Educational Innovations*, Ellis and Fouts (1993) conclude: "the research on mastery learning is about as strong as one can find in the annals of educational investigation. Study after study indicates the superiority of mastery learning over traditional methods of instruction" (p. 112).

Research evidence also shows that the positive effects of mastery learning are not limited to cognitive or achievement outcomes. The process also yields improvements in students' confidence in learning situations, school attendance rates, involvement in class sessions, attitudes toward learning, and a variety of other affective measures (Block et al., 1989; Changeiywo et al., 2011; Guskey and Pigott, 1988; Whiting and Render, 1987). This multidimensional impact has been referred to as the 'multiplier effect' of mastery learning and makes it one of the most cost-effective means of educational improvement.

It should be noted that one review of the research on mastery learning, contrary to all other reviews, indicated that the process had essentially no effect on student achievement (Slavin, 1987). This finding surprised not only scholars familiar with the vast research literature on mastery learning showing it to yield very positive results, but also large numbers of practitioners who had experienced its positive impact first hand. A close inspection of this review shows, however, that it was conducted using techniques of questionable validity (Hiebert, 1987), employed capricious selection criteria (Kulik et al., 1990b), reported results in a biased manner (Bloom, 1987; Walberg, 1988), and drew conclusions not substantiated by the evidence presented (Guskey, 1987, 1988). Most importantly, three much more extensive and methodologically sound reviews published since (Guskey and Pigott, 1988; Kulik et al., 1990a; Miles, 2010) have verified mastery learning's consistently positive impact on a broad range of student learning outcomes and, in one case (i.e., Kulik et al., 1990b), showed clearly the distorted nature of this earlier report.

Conclusion

Researchers today generally recognize the value of the essential elements of mastery learning and their importance in effective teaching at any level of education. As a result, fewer studies are being conducted on the mastery learning process, per se. Instead, researchers are looking for ways to enhance results further, adding additional elements to the mastery learning process that positively contribute to student learning in hopes of attaining even more impressive gains (Bloom, 1984). Recent work on the integration of mastery learning with other innovative strategies appears especially promising (Guskey, 1997b; Guskey and Jung, 2011).

In his later writing Bloom, too, described exciting work on other ideas designed to attain results even more positive than those typically achieved with mastery learning (Bloom, 1984). These ideas stemmed from the work of two of Bloom's doctoral students, Anania (1981) and Burke (1983), who compared student learning under three different instructional conditions. The first was *conventional instruction* in which students were taught in group-based classes that included about 30 students and where periodic assessments were given mainly for the purposes of grading. The second was *mastery learning*, where students again were taught in group-based classes of about 30 students but were administered regular 'formative' assessments for feedback, followed by individualized corrective instruction and parallel second assessments to determine the extent to which they mastered specific learning goals. The third was *tutoring*, where a good tutor was paired with each student. Under tutoring students were also administered regular formative assessments, along with

corrective procedures and parallel second assessments, although the need for corrective work under tutoring was usually quite small.

The differences in students' final achievement under these three conditions were striking. Using the standard deviation (sigma) of the control (conventional) class as the measure of difference, Bloom's students discovered that:

> The average student under tutoring was about two standard deviations above the average of the control class (the average tutored students was above 98% of the students in the control class). The average student under mastery learning was about one standard deviation above the average of the control class (the average mastery learning student was above 84% of the students in the control class). … Thus under the best learning conditions we can devise (tutoring), the average student is 2 sigma above the average control student taught under conventional group methods of instruction.
>
> Bloom, 1984, p. 4

Bloom referred to this as the '2 Sigma Problem':

> The tutoring process demonstrates that *most* students do have the potential to reach this high level of learning. I believe an important task of research and instruction is to seek ways of accomplishing this under more practical and realistic conditions than the one-to-one tutoring, which is too costly for most societies to bear on a large scale. This is the '2 Sigma' problem. Can researchers and teachers devise teaching-learning conditions that will enable the majority of students under *group instruction* to attain levels of achievement that can at present be reached only under tutoring conditions?
>
> Bloom, 1984, pp. 4–5)

Bloom believed that attaining this high level of achievement would probably require more than just improvements in the quality of group instruction. Researchers and teachers might also need to find ways of improving students' learning processes, the curriculum and instructional materials, the home environmental support of students' school learning, and providing a focus on higher-level thinking skills. Nevertheless, Bloom remained convinced that careful attention to the elements of mastery learning would allow educators at all levels to make great strides in their efforts to reduce the variation in student achievement, close achievement gaps, and help all children to learn excellently.

See also: Academic Engagement; Motivation and Actions, Psychology of; Motivation, Learning, and Instruction; School Achievement: Motivational Determinants and Processes; School Burnout and Engagement: Lessons from a Longitudinal Study in Finland; Schooling: Impact on Cognitive and Motivational Development; Self-Regulated Learning: Theories, Measures, and Outcomes; Test Anxiety and Academic Achievement.

Bibliography

Anania, J., 1981. The Effects of Quality of Instruction on the Cognitive and Affective Learning of Students (Unpublished doctoral dissertation). University of Chicago, Chicago, IL.

Anderson, S.A., 1994. Synthesis of Research on Mastery Learning (ERIC Document Reproduction Service No. ED 382 567).

Anderson, S., Barrett, C., Huston, M., Lay, L., Myr, G., Sexton, D., Watson, B., 1992. A Mastery Learning Experiment (Technical Report). Yale Public Schools, Yale, MI.

Arredondo, D.E., Block, J.H., 1990. Recognizing the connections between thinking skills and mastery learning. Educational Leadership 47 (5), 4–10.

Block, J.H., 1983. Learning rates and mastery learning. Outcomes 2 (3), 18–23.

Block, J.H., Efthim, H.E., Burns, R.B., 1989. Building Effective Mastery Learning Schools. Longman, New York, NY.

Bloom, B.S., 1968. Learning for mastery. Evaluation Comment 1 (2), 1–12.

Bloom, B.S., 1971a. Mastery learning. In: Block, J.H. (Ed.), Mastery Learning: Theory and Practice. Holt, Rinehart & Winston, New York, NY, pp. 47–63.

Bloom, B.S., 1971b. Individual Differences in School Achievement: A Vanishing Point? Phi Delta Kappan International, Bloomington, IN.

Bloom, B.S., 1976. Human Characteristics and School Learning. McGraw-Hill, New York, NY.

Bloom, B.S., 1977. Favorable learning conditions for all. Teacher 95 (3), 22–28.

Bloom, B.S., 1978. New views of the learner: implications for instruction and curriculum. Educational Leadership 35 (7), 563–576.

Bloom, B.S., 1984. The 2 sigma problem: the search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher 13 (6), 4–16.

Bloom, B.S., 1987. A response to Slavin's mastery learning reconsidered. Review of Educational Research 57 (4), 507–508.

Bloom, B.S., 1988. Helping all children learn in elementary school and beyond. Principal 67 (4), 12–17.

Bloom, B.S., Hastings, J.T., Madaus, G., 1971. Handbook on Formative and Summative Evaluation of Student Learning. McGraw-Hill, New York, NY.

Burke, A.J., 1983. Students' Potential for Learning Contrasted under Tutorial and Group Approaches to Instruction (Unpublished doctoral dissertation). University of Chicago, Chicago, IL.

Cabezon, E., 1984. The Effects of Marked Changes in Student Achievement Patterns on the Students, Their Teachers, and Their Parents: The Chilean Case (Unpublished doctoral dissertation). University of Chicago, Chicago, IL.

Carroll, J.B., 1963. A model for school learning. Teachers College Record 64, 723–733.

Chan, K.S., 1981. The Interaction of Aptitude with Mastery versus Non-mastery Instruction: Effects on Reading Comprehension of Grade Three Students (Unpublished doctoral dissertation). University of Western Australia, Perth, Australia.

Changeiywo, J.M., Wambugu, P.W., Wachanga, S.W., 2011. Investigations of students' motivation toward learning secondary school physics through mastery learning approach. International Journal of Science and Mathematics Education 9 (6), 1333–1350.

DeWeese, S.V., Randolph, J.J., February 2011. Effective Use of Correctives in Mastery Learning. Paper presented at the Association of Teacher Educators National Conference, Orlando, FL.

Dollard, J., Miller, N.E., 1950. Personality and Psychotherapy. McGraw-Hill, New York, NY.

Ellis, A.K., Fouts, J.T., 1993. Research on Educational Innovations. Eye on Education, Inc, Princeton Junction, NJ.

Fitzpatrick, K.A., April 1985. Group-Based Mastery Learning: A Robin Hood Approach to Instruction?. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Glaser, R., 1966. The Program for Individually Prescribed Instruction. University of Pittsburgh, Pittsburgh, PA.

Guskey, T.R., 1987. Rethinking mastery learning reconsidered. Review of Educational Research 57 (2), 225–229.

Guskey, T.R., 1988. Response to Slavin: who defines best? Educational Leadership 46 (2), 26–27.

Guskey, T.R., 1997a. Implementing Mastery Learning, second ed. Wadsworth, Belmont, CA.

Guskey, T.R., 1997b. Putting it all together: integrating educational innovations. In: Caldwell, S.J. (Ed.), Professional Development in Learning-Centered Schools. National Staff Development Council, Oxford, OH, pp. 130–149.

Guskey, T.R., 2007. Closing achievement gaps: revisiting Benjamin S. Bloom's "learning for mastery". Journal of Advanced Academics 19 (1), 8–31.

Guskey, T.R., 2008. The rest of the story. Educational Leadership 65 (4), 28–35.

Guskey, T.R., 2010. Lessons of mastery learning. Educational Leadership 68 (2), 52–57.

Guskey, T.R., Jung, L.A., 2011. Response to intervention and mastery learning: tracing roots and seeking common ground. The Clearing House 84 (6), 249–255.

Guskey, T.R., Pigott, T.D., 1988. Research on group-based mastery learning programs: a meta-analysis. Journal of Educational Research 81, 197–216.

Hiebert, E.H., 1987. The context of instruction and student learning: an examination of Slavin's assumptions. Review of Educational Research 57, 337–340.

Klausmeier, H.J., et al., 1968. Individually Guided Education in the Multi-Unit School: Guidelines for Implementation. Office of Education (DHEW), Bureau of Research, Washington, DC.

Klecker, B.M., Chapman, A., November 2008. Advocating the Implementation of Mastery Learning in Higher Education to Increase Student Learning and Retention. Paper presented at the annual meeting of the Mid-South Educational Research Association, Knoxville, TN.

Kozlovsky, J.D., 1990. Integrating thinking skills and mastery learning in Baltimore County. Educational Leadership 47 (5), 6.

Kulik, C.C., Kulik, J.A., Bangert-Drowns, R.L., 1990a. Effectiveness of mastery learning programs: a meta-analysis. Review of Educational Research 60, 265–299.

Kulik, J.A., Kulik, C.C., 1989. Meta-analysis in education. International Journal of Educational Research 13 (2), 221–340.

Kulik, J.A., Kulik, C.C., Bangert-Drowns, R.L., 1990b. Is there better evidence on mastery learning? A response to Slavin. Review of Educational Research 60, 303–307.

Langeheine, R., April 1992. State Mastery Learning: Dynamic Models for Longitudinal Data. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Levine, D.U., et al., 1985. Improving Student Achievement through Mastery Learning Programs. Jossey-Bass, San Francisco, CA.

Mevarech, Z.R., April 1981. Attaining Mastery on Higher Cognitive Achievement. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Mevarech, Z.R., 1985. The effects of cooperative mastery learning strategies on mathematical achievement. Journal of Educational Research 78 (6), 372–377.

Miles, K., 2010. Mastery Learning and Academic Achievement (Unpublished doctoral dissertation). Walden University, Philadelphia, PA.

Morrison, H.C., 1926. The Practice of Teaching in the Secondary School. University of Chicago Press, Chicago, IL.

Nakajima, A., 2012. A powerful influence on Japanese education. In: Guskey, T.R. (Ed.), Benjamin S. Bloom: Portraits of an Educator, second ed. Rowman & Littlefield Education, Lanham, MD, pp. 109–111.

National Council of Teachers of Mathematics, 2000. Principles and Standards for School Mathematics. Author, Reston, VA. Available: http://standards.nctm.org/document/index.htm.

Postlethwaite, K., Haggarty, L., 1998. Towards effective and transferable learning in secondary school: the development of an approach based on mastery learning. British Educational Research Journal 24 (3), 333–353.

Reezigt, B.J., Weide, M.G., April 1990. The Effects of Group-based Mastery Learning on Language and Arithmetic Achievement and Attitudes in Primary Education in the Netherlands. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Slavin, R.E., 1987. Mastery learning reconsidered. Review of Educational Research 57 (2), 175–213.

Waddington, T., April 1995. Why Mastery Matters. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Walberg, H.J., 1988. Response to Slavin: what's the best evidence? Educational Leadership 46 (2), 28.

Walberg, H.J., 1990. Productive teaching and instruction: assessing the knowledge base. Phi Delta Kappan 71, 470–478.

Washburne, C.W., 1922. Educational measurements as a key to individualizing instruction and promotions. Journal of Educational Research 5 (1), 195–206.

Whiting, B., Render, G.F., 1987. Cognitive and affective outcomes of mastery learning. A review of sixteen semesters. The Clearing House 60 (6), 276–280.

Whiting, B., Van Burgh, J.W., Render, G.F., 1995. Mastery Learning in the Classroom. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wu, W.Y., April 1994. Mastery Learning in Hong Kong: Challenges and Prospects. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Yildiran, G., 2006. Multicultural Applications of Mastery Learning. Faculty of Education, Bogazici University, Istanbul, Turkey.

Yildiran, G., Aydin, E., 2005. The effects of mastery learning and cooperative, competitive, and individualistic learning environment organizations on achievement and attitudes in mathematics. Research in Mathematical Education: Journal of the Korean Society of Mathematical Education Series D 9 (1), 55–72.

Zimmerman, B.J., Dibenedetto, M.K., 2008. Mastery learning and assessment: implications for students and teachers in an era of high stakes testing. Psychology in the Schools 45 (3), 206–216.

Material Flow Analysis

Heinz Schandl, Commonwealth Scientific and Industrial Research Organization, Canberra, ACT, Australia
Anke Schaffartzik, Institute of Social Ecology, Alpen-Adria University, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

All societies depend on natural resources to feed, house, and transport people and to produce the goods and services everyday life depends upon. The magnitude of the physical interaction between society and nature may be measured through material flow accounting. The accounts measure the amount of materials extracted domestically – biomass, fossil fuels, metal ores, and construction materials – as well as traded materials and also measure the disposal of waste and emissions to the environment by taking a whole life cycle perspective. The accounts measure yearly material flows in tonnes and provide an information base about the scale of natural resource use that underpins human development and economic prosperity. Material flow accounts have become part of the system of integrated economy–environmental accounts and inform modern environmental policy making that understands environmental outcomes are greatly related to the process of consumption and production. Indicators based on material flow accounts play a major role in the environmental policy process and figure prominently in international debates about sustainable consumption and production, resource efficiency and decoupling, and the green economy.

Introduction

The study of material flow accounting (MFA) focuses on the natural resource requirements of national economies, specific economic activities (such as construction and housing, transport, and mobility), or geographical units such as cities. MFA accounts for the input of primary materials – biomass, fossil fuels, metal ores, and minerals – and semimanufactures and final goods into economic activities. MFA also accounts for the outputs of economic systems including final goods for export, waste, and emissions. MFA often conceptualizes the economic system as a 'black box.' There are, however, accounting strategies for material flows within economic systems available as well.

Material flow accounts are complementary to economic accounts and measure tonnes of materials that are associated with production and consumption processes. The accounts measure the physical implications of economic systems and allow links to be established between economic activities and the natural resource base and natural environment they depend upon. Material flow accounts can therefore be used as tools in both environmental and economic policy. Material flow accounts have grown in political importance over the past decade in response to converging environmental and economic pressures such as climate change, food security, rising energy costs, and the global financial crisis of 2008.

MFA has a sound theoretical base in the theory of social-ecological systems and the concept of social metabolism and industrial metabolism (Ayres and Simonis, 1994). MFA operationalizes the concept of metabolism, provides an understanding of the amounts of natural resources used in different economies at different times in history, and has a multitude of application scales – global, regional, national, subnational, urban, businesses, and households – and a variety of policy applications in the context of integrated economic and environmental policy making.

History of MFA

For a long time, there has been recognition that all economic activities depend on the timely availability and affordability of natural resources. Major economic powers such as the Soviet Union and the United States of America have established accounts of their natural resource base and resource requirements and the supply systems that deliver those resources. The most prominent example is the Paley report of the 1950s (Paley, 1952) comparing the resource base of the United States and its allies to those of the Soviet Union.

Early research that pioneered a physical accounting approach in economics includes the work of Nicholas Georgescu-Roegen, Herman Daly, and Robert Ayres. Georgescu-Roegen was one of the first economists who assessed the interplay of economic activity and the natural environment within the concept of thermodynamics (Georgescu-Roegen, 1971). Herman Daly introduced the notion of scale referring to the total volume of resource flows: the matter–energy throughput taken from the environment as low-entropy resources and returned to the environment as high-entropy waste. Scale, according to Daly, is relative to environmental carrying capacity and is often ignored by mainstream economics (Daly, 1992). Robert Ayres published an early material flow account for the United States that set an analytical standard for the next generation of MFA studies (Kneese et al., 1970).

National material flow accounts were pioneered in Germany, Japan, and Austria in the early 1990s (Fischer-Kowalski and Hüttler, 1998). The World Resources Institute published two groundbreaking comparative studies "Resource Flows: The Material Basis of Industrial Economies" (Adriaanse et al., 1997) and "The Weight of Nations" (Matthews et al., 2000) that highlighted differences in methodology and triggered a process of methodological harmonization and standardization. The European statistical office – EUROSTAT – has since been fine-tuning methods and has produced several iterations of the

MFA methods guidebook (Eurostat, 2012). The OECD, supported by a political decision of the leadership of the group of eight (G8) at Sea Island in 2008, identified the policy implications of MFA data and indicators in a series of reports and further supported the process of standardization of material flow and resource productivity information (OECD, 2008). Most recently, MFA was integrated into the system of environmental–economic accounting helping to establish a sound base for environmental accounts at national, regional, and global levels (UN, 2014).

The sustainable development discourse that emerged in the 1990s profoundly changed the way in which the policy community viewed environmental problems. The new sustainability discourse helped establish the notion of decoupling of economic activities from environmental pressures and impacts. In essence, sustainable development suggested a new path to development in which the exploitation of natural resources, the direction of investments, the orientation of technological development, and institutional change are all in harmony and enhance the current and future potential of socioeconomic systems to meet human needs and aspirations.

As a consequence of the new sustainable development discourse, many national governments and international organizations such as the United Nations have moved from a conservationist approach in environmental politics to a new understanding of the social and economic drivers that cause environmental pressures. MFA measures a specific aspect of environmental pressures and allows progress in resource efficiency and decoupling to be assessed. This new policy orientation has inspired a large number of global and regional studies into natural resource use and economic development.

Scientifically, MFA is used as a data organization strategy and an analytical tool in the research fields of industrial ecology and ecological economics, but also figures prominently in physical environmental history (Krausmann et al., 2008) and environmental sociology (Fischer-Kowalski and Haberl, 1997).

Accounting Standards and Headline Indicators

MFA assembles data on material use, often at the national level. The accounting is based on the concept of industrial metabolism and starts from a clear definition of system boundaries between economy and nature (and national borders between economies) and an assessment of physical stocks, for example, buildings and transport infrastructure, but also livestock animals and people. It measures the yearly flows of materials required to fuel and maintain existing stocks, to replace stock that has become depreciated, and to build new stock. The accounts provide data on material inputs into, the changes in material stock within, and the material outputs in the form of goods for exports or discharge of waste and emissions to the domestic environment of an economy. MFA covers all solid, gaseous, and liquid materials that are mobilized in economic activities with the exception of bulk water and air. Water is usually treated in a separate water account and air is dealt with as a memorandum item when closing the material balance of an economy.

The unit of measurement of an MFA is most commonly (metric) tonnes. Inputs are distinguished by their origin, i.e., whether they are extracted domestically (domestic extraction, DE) or are imported goods from other countries (imports, IM and exports, EX). Primary material inputs are most commonly grouped into four main aggregate material groups: biomass (renewable), and fossil fuels, metal ores, and industrial and construction minerals (nonrenewable). The outputs are distinguished by gateways through which they leave the economy and include goods exported, waste to landfill, and emissions to air and water. Material flow accounts aim for close complementarity with economic accounts and include materials at the commodification stage, i.e., when a natural resource becomes a primary material and a market value is attached. In order to provide a complete picture of the physical interaction of the economy with the environment, this complementarity to economic accounts cannot always be upheld and the accounts include a number of inputs that have no market value attached. This is the case for biomass grazed and scavenged by livestock, crop residues that are used economically in the context of a farm, the waste rock component of metal ores, and sand and gravel that is extracted within the confines of a construction company on the premises of the company to mention the most important categories. Overall, material flows that have no market value attached make up 20% of global materials extraction.

Materials that are mobilized in the course of economic activities but do not enter the production process as such are classified as unused extraction. They may include earth movement in mining and agriculture, by-catch in fisheries, and by-harvest of timber and are usually left where they occur. These flows are classified as unused DE (**Figure 1**).

A full material flow dataset hence comprises all economically relevant material exchanges and enables the construction of a set of indicators. These are increasingly used by national governments and businesses. The standard headline indicators derived from an MFA account include the following:

Direct material input (DMI), which measures the overall material input into a national economy and is calculated as the sum of DE and imports (DMI = DE + IM).

The *physical trade balance (PTB)* is a measure of net trade and is calculated as the difference between imports and exports. The PTB is the reverse of the monetary terms of trade

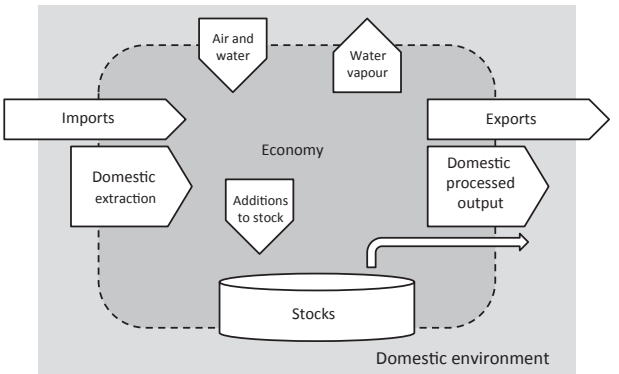

Figure 1 The material cycle.

and shows whether a country is a net importer of materials and hence depending on the world market for primary materials to satisfy the physical requirements of its economy, or a net exporter supplying primary resources to global markets (PTB = IM – EX).

Domestic material consumption (DMC) is a measure of apparent consumption of an economy and is calculated by adding imports and subtracting exports from DE of materials (DMC = DE + IM – EX or DMC = DE + PTB).

Domestic processed output (DPO) provides a comprehensive measure of all materials that leave an economy at any stage. Included in DPO are emissions to air, industrial and household waste deposited in landfill, material loads in wastewater, and materials dispersed into the environment as a result of product use.

Net additions to stock (NAS) measures the yearly addition to the stock of buildings, transport infrastructure, productive capacity, and durable consumer goods. It subtracts materials that are discarded from gross additions to stock (NAS = DMI – DPO – EX).

MFA accounts are a first step toward a satellite account for materials to the system of national accounts (SNA) and provide a rich analytical database and indicators to inform modern, integrated economic and environmentally minded policy decisions.

The complementarity of MFA and national accounts allows material efficiency (ME) indicators to be calculated. The ME is usually expressed as material intensity, i.e., the amount of material required per unit of economic output. It is measured by dividing DMC by gross domestic product (GDP). Efficiency can also be expressed as material productivity (GDP/DMC) which is the inverse of material intensity.

Most studies and policy reports to date have focused on these indicators and hence have presented a production-focused perspective on the material requirements, waste and emissions of economies. In the context of global concern for resource supply security and global fairness a consumption-based approach to material use has been introduced, to complement other consumption-based indicator systems such as those for energy use (Wiedmann, 2009) and carbon emissions (Hertwich and Peters, 2009) which have pioneered consumption-based analysis. When countries import goods, the primary resources required to produce those goods remain in the country of origin. At the same time, when a country exports goods the primary resources required to produce the good remain in the country. Traded goods are measured at the time when they cross the boundary between countries (direct imports and exports) but they are related to an indirect material requirement referred to as the raw material equivalent (RME). RME refers to the primary materials that were required to produce the goods and can be measured with various analytical techniques (Munoz et al., 2009).

More recently, the notion of *material footprint (MF)* of consumption was introduced (Wiedmann et al., 2013). The MF of an economy is accounted for through the attribution of global material extraction to final consumption in a country and requires a global, multiregional input–output framework as an analytical tool. Alternatively, the upstream and downstream requirements of traded goods could also be measured by employing life cycle analysis coefficients REF. Both approaches have been tested but are still considered immature with regard to the robustness and reliability of results. Technically, MF can be expressed as DE plus the RME of imported goods, RME_{IM} minus the RME of exported goods, RME_{EX} ($MF = DE + RME_{IM} - RME_{EX}$).

The MF indicator allows the establishment of a revised ME indicator which is sensitive to the burden shifting that occurs between nations importing and exporting natural resources. Material intensity is now measured by dividing MF by GDP.

Applications of MFA

Most commonly, material flow accounts are established for national economies. These accounts complement economic accounts and report progress in the ME of nations. Such accounts now exist for most OECD countries and a number of developing countries. While early studies focused on descriptive aspects of a country's material use, the level of analytical sophistication has continued to increase. More recent studies discuss the social and economic drivers of material use, look at the economic structure of a country and the role the country plays in the world economy, analyze trade patterns and their material underpinnings, and unpack policy implications of material flow accounts.

Material flow accounts often conceptualize the economy as a black box and focus on those inputs and outputs for which data are more easily accessible than for material exchanges within an economy. As a consequence, material flow accounts for specific economic activities have been rare. To gain a good understanding of the material interrelationships within a national economy requires a full physical input–output table (PIOT) of that economy. Establishing such a table is very data and time intensive and for this reason PIOTs are often outdated when they are published. PIOTs do, however, exist for Denmark and Germany (Weisz and Duchin, 2006) and provide in-depth insight into the material interactions within those economies.

Material flow accounts for cities have been another focus of the research community and many case studies have become accessible recently. Urban material flow accounts also have a long history and go all the way back to the early work of Stephen Boyden (Boyden et al., 1981) and Abel Wolman (Wolman, 1965). Accounting for urban metabolism using similar accounting strategies to national MFA has become a very active area of research more recently (Weisz and Steinberger, 2010).

During the past two decades, material supply and demand relationships in the world economy have become ever more complex and have introduced a new focus in MFA, looking at upstream and downstream material requirements for traded goods.

Globalization of production and consumption processes has increased the spatial disconnect between primary natural resource supply and centers of demand and has added complexity to societies' metabolism. In many economies, a proportion of resources extracted domestically is used for the production of traded goods while at the same time high-end processed goods with substantial amounts of embodied materials and energy are imported. Standard MFA does not

distinguish whether resources are extracted to satisfy domestic or foreign final demand and the structure of production and the origin of material inputs remains unknown within the 'black box' that represents an economy. Environmentally extended input–output analysis (EE-IOA) based on the work of economist Wassily Leontief (Leontief, 1970) supports the analysis of production structure and intra-industry relationships within the 'black box.'

MFA and EE-IOA cover the flows of material, energy, and money within the economy, but not accumulated physical and monetary capital or the stocks of buildings, infrastructure, productive capital, and durable consumer goods which play a central role in the economy, and in consumption and everyday life. They provide physical services such as shelter, mobility, and industrial capacity, representing large accumulations of material and investment. They are usually long lived and their depreciation time determines the amount of material required to maintain or replace them. Physical stock also determines the speed at which new technologies can penetrate markets, and there are technical properties that link service provision with energy and material throughput. Understanding the dynamic properties of physical stock in use is the subject of dynamic stock modeling (Muller, 2006).

MFA has also underpinned a variety of local studies, mostly in developing countries REF looking into the current metabolic transition, and historical studies of social metabolism providing long-time accounts (centuries) of industrial development and changing natural resource bases (Fischer-Kowalski et al., 2011). MFA data and indicators have assisted in empirically exploring the nonlinear relationship between material and energy use and human development (Steinberger and Roberts, 2010).

Policy Use

The information provided by material flow accounts is increasingly being used by the policy and business communities. Most notably, material flow accounts have become an essential part of environmental satellite accounts to the SNA and offer an alternative to the green accounting approaches of the 1970s and 1980s (Bartelmus, 2003). Material flow data show the quantity of natural resources required per unit of output in a national economy, for a certain economic sector or activity, and provide insight into the eco-efficiency of the production process in different countries and different economic sectors. They allow for benchmarking and target setting and for monitoring progress in resource efficiency and waste minimization. They also form an important part of triple bottom line accounting in businesses and industries (Foran et al., 2005).

Material flow accounts allow the creation of a set of pressure indicators to monitor the physical flow of natural resources and waste and emissions between economy and environment. Pressure indicators allow for a swift policy response, even before pressures have led to changes in the state of the environment and are causing visible impacts. This allows much shorter response times for environmental policy making with regard to major environmental problems including the depletion of natural resources, climate change, biodiversity loss, and problems related to growing amounts of household and industrial waste.

Decoupling has become a signature concept of modern integrated economic–environmental policy (UNEP, 2011a) and refers to the ability to expand the economy while reducing material and energy use, as well as waste and emissions. This may occur through the adoption of new technologies and innovations in major systems of provision including agriculture and food, construction and housing, and transport and mobility but also in heavy industries such as iron and steel, cement, and paper production. Decoupling is based on the idea of achieving more human well-being with less environmental disruption. The potential for efficiency gains is very large in many sectors of the economy, sometimes up to 80% (Von Weizsaecker et al., 2009).

The new policy concepts of sustainable consumption and production, resource efficiency and waste minimization, and investing in a green and low carbon economy depend on information about material flows to set objectives and targets, monitor progress, and evaluate policy outcomes. Many countries and regional bodies have developed their national strategies and framework policies to deliver human development and economic growth while minimizing natural resource depletion and environmental degradation. Examples include the European Union Resource Strategy supporting a resource-efficient Europe (EU, 2011) and the material flow accounts delivered by member countries and published by the European statistical office to monitor progress, the Chinese law to promote a circular economy (UNEP, 2011b), and the Japanese policy framework for a Sound Material Cycle Society REF which includes a set of material flow-related targets including material productivity of the economy, the recycling rate, and the amount of final waste disposal (Takiguchi and Takemoto, 2008).

The economic context for material use has changed profoundly. Economic development in the twentieth century was fueled by affordable natural resources, and prices were declining during most of the century. Since 2000, the price of most natural resources has been rising and price volatility has increased because of the vast resource demand of many developing economies, especially China, and supply capacities have fallen behind demand. In this new economic context, national governments and business leaders require an extended set of information for decision making beyond monetary economic indicators. Material flow accounts and indicators provide this information.

See also: Ecological Economics; Environmental Sociology; Industrial Ecology; Sustainable Production and Consumption.

Bibliography

Adriaanse, A., Bringezu, S., Hammond, A., Moriguchi, Y., Rodenburg, E., Rogich, D., Schuetz, H., 1997. Resource Flows: The Material Basis of Industrial Economies. World Resources Institute, Washington, DC.

Ayres, R.U., Simonis, U.E., 1994. Industrial Metabolism. Restructuring for Sustainable Development. United Nations University Press, Tokyo and New York.

Bartelmus, P., 2003. Dematerialization and capital maintenance: two sides of the sustainability coin. Ecological Economics 46, 61–81.

Boyden, S., Millar, S., Newcombe, K., 1981. The Ecology of a City and its People; The Case of Hong Kong. Australian National University Press, Canberra.
Daly, H.E., 1992. Allocation, distribution, and scale: towards an economics that is efficient, just, and sustainable. Ecological Economics 6, 185–193.
EU, 2011. In: European Commission (Ed.), A Resource Efficient Europe – Flagship Initiative under the Europe 2020 Strategy. Brussels.
Eurostat, 2012. Economy-wide Material Flow Accounts (EW-mfa). Compilation Guide 2012. European Statistical Office, Luxembourg.
Fischer-Kowalski, M., Haberl, H., 1997. Tons, joules, and money: modes of production and their sustainability problems. Society and Natural Resources 10, 61–85.
Fischer-Kowalski, M., Hüttler, W., 1998. Society's metabolism. Journal of Industrial Ecology 2, 107–136.
Fischer-Kowalski, M., Singh, S.J., Lauk, C., Remesch, A., Ringhofer, L., Grunbuhel, C.M., 2011. Sociometabolic transitions in subsistence communities: Boserup revisited in four comparative case studies. Human Ecology Review 18, 147–158.
Foran, B., Lenzen, M., Dey, C., Bilek, M., 2005. Integrating sustainable chain management with triple bottom line accounting. Ecological Economics 52, 143–157.
Georgescu-Roegen, N., 1971. The Entropy Law and the Economic Process. Harvard University Press, Cambridge, MA.
Hertwich, E.G., Peters, G.P., 2009. Carbon footprint of nations: a global, trade-linked analysis. Environmental Sciences and Technology 43, 6414–6420.
Kneese, A.V., Ayres, R.U., D'Arge, R.C., 1970. Economics and the Environment. A Materials Balance Approach. Baltimore.
Krausmann, F., Fischer-Kowalski, M., Schandl, H., Eisenmenger, N., 2008. The global sociometabolic transition. Journal of Industrial Ecology 12, 637–656.
Leontief, W., 1970. Environmental repercussions and economic structure. Input–output approach. Review of Economics and Statistics 52, 262–271.
Matthews, E., Bringezu, S., Fischer-Kowalski, M., Huettler, W., Kleijn, R., Moriguchi, Y., Ottke, C., Rodenburg, E., Rogich, D., Schandl, H., Schuetz, H., Van der Voet, E., Weisz, H., 2000. The Weight of Nations. Material Outflows from Industrial Economies. World Resources Institute, Washington, DC.
Muller, D.B., 2006. Stock dynamics for forecasting material flows – case study for housing in The Netherlands. Ecological Economics 59, 142–156.
Munoz, P., Giljum, S., Roca, J., 2009. The raw material equivalents of international trade. Empirical Evidence for Latin America. Journal of Industrial Ecology 13, 881–897.
OECD, 2008. Measuring Material Flows and Resource Productivity. Synthesis report. OECD, Paris.
Paley, W.S., 1952. Resources for Freedom: Report of the President's Materials Policy Commission. United States Government Printing, Washington, DC.
Steinberger, J.K., Roberts, J.T., 2010. From constraint to sufficiency: the decoupling of energy and carbon from human needs, 1975–2005. Ecological Economics 70, 425–433.
Takiguchi, H., Takemoto, K., 2008. Japanese 3R policies based on material flow analysis. Journal of Industrial Ecology 12, 792–798.
UN, 2014. System of Environmental-Economic Accounting 2012. Central Framework. United Nations, New York.
UNEP, 2011a. Decoupling Natural Resource Use and Environmental Impacts from Economic Growth. United Nations Environment Programme, Nairobi.
UNEP, 2011b. Resource Efficiency: Economics and Outlook for Asia and the Pacific. CSIRO Publishing, Melbourne.
Von Weizsaecker, E., Hargroves, K., Smith, M., Desha, C., Stasinopoulos, P., 2009. Factor Five. Transforming the Global Economy through 80% Improvements in Resource Productivity. Earthscan, London and Sterling, VA.
Weisz, H., Duchin, F., 2006. Physical and monetary input–output analysis: what makes the difference? Ecological Economics 57, 534–541.
Wiedmann, T., 2009. A first empirical comparison of energy Footprints embodied in trade - MRIO versus PLUM. Ecological Economics 68, 1975–1990.
Weisz, H., Steinberger, J.K., 2010. Reducing energy and material flows in cities. Current Opinion in Environmental Sustainability 2, 185–192.
Wiedmann, T.O., Schandl, H., Lenzen, M., Moran, D., Suh, S., West, J., Kanemoto, K., 2013. The material footprint of nations. Proceedings of the National Academy of Sciences.
Wolman, A., 1965. The metabolism of cities. Scientific American, 179–190.

Relevant Websites

http://is4ie.org/ – International Society of Industrial Ecology.
http://www.unep.org/resourcepanel/ – UNEP International Resource Panel.

Materiality and Culture

Stuart McLean, University of Minnesota, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Materiality' and 'culture' are among the most discussed terms in contemporary social and cultural theory. This article offers a broad introductory overview of current debates both within the discipline of anthropology and beyond and situates these briefly with reference to anthropology's and social thought's entanglements with longer-term histories of, for example, capitalism and European colonial expansion.

How might we begin to describe and understand the relationship of collective life to the material universe in which it unfolds? How is the materiality of, for example, bodies, artifacts, and environments implicated in formations of power, identity, and belonging? Is the material world significant only insofar as it is interpreted and represented by humans, or should we think of matter as an autonomous power, possessed of its own determinations and agentive capacities and thus as capable of intervening in human affairs in ways that challenge or refuse social scientific explanation?

The terms 'materiality' and 'culture' have been the subject of some of the most intense and wide-ranging debates in contemporary social and cultural theory, drawing on a broad range of intellectual influences, including Marxism, structuralism and semiotics, philosophy, human geography, biology, and physics (Coole and Frost, 2010; Dolphijn and van der Tuin, 2012). In anthropology, a renewed attentiveness to the material world has been fostered too by, among other things, science studies, especially the approach some times referred to as Actor Network Theory, with its call to include other-than-human presences and agencies in accounts of social life, a call to which contemporary ecological crises are seen as lending a new urgency (Latour, 1993, 2013). At the same time, anthropological theorizing developed on the basis of ethnographic research conducted in (predominantly) non-Western contexts, notably Melanesia, has drawn attention to the variety of ways in which humans and material things are interconnected, suggesting the possibility of understanding personhood and agency as 'fractal' or 'distributed' (composed of relationships extending across time and space) rather than simply individuated (Gell, 1998; Strathern, 1988; Wagner, 1991). There remain, however, profound disagreements among many of the parties to these discussions. Indeed, it may be that 'culture' and 'materiality' necessarily exist in a relationship of tension. On the one hand, the concept of 'culture' in anthropology has often been understood as that which defines the distinctiveness of humans as a species, as well as the differences among human groups, whether this is understood (for example) with reference to language, to the use of symbols more generally, or to the broad array of beliefs, values, and behaviors that, according to Tylor, humans acquired by virtue of their participation in a social group (Kroeber and Kluckhohn, 1952; Tylor, 1903). 'Materiality,' of course, can be taken as referring either to particular material things or to the material being of the world in general. As such, it might be understood as pertaining not to what sets humans apart but to what they share inescapably both with other life forms and with material entities of all kinds. For example, the human beings who, as culture-bearers, have featured as anthropology's self-designated object of study lead lives that are inextricably entangled with a variety of material worldly presences. These might include, on the one hand, features of the physical environment such as trees, rivers, mountains, oceans, or a range of humanly manufactured artifacts, including cooking utensils, items of clothing, buildings, ritual objects, automobiles, computers, and cell phones. At the same time, human beings are themselves material entities, that is, bodies composed of cells and molecules. As such, they eat, drink, sweat, urinate and defecate, fall sick, become sexually aroused, grow old, and eventually die and decompose, whereupon their constituent substance is redistributed and recombined with that of their surroundings. Materiality and culture are, therefore, neither straightforwardly identical nor unambiguously distinct. Rather, they remain for anthropology inextricably yet unresolvedly entangled.

Despite this seeming centrality to anthropology's self-designated subject matter, however, it has often been suggested that materiality has emerged only in recent decades as a focus of sustained anthropological inquiry. Certainly, the period since the 1980s has witnessed a striking proliferation of works of anthropological scholarship specifically focused on aspects of the material world, from fashion to communication technologies to home decoration. Yet, this rapid expansion of interest can be contrasted with a relative lack of interest on anthropology's part throughout most of the twentieth century in the variegated materialities of social life. This lack of interest has sometimes been explained with reference to the divisions between 'Mind' and 'Matter' or between 'Nature' on the one hand and 'Society' or 'Culture' on the other that have often been taken to characterize modern, Western thought. Some commentators have suggested that these dualisms can be traced to the so-called 'scientific revolution' in seventeenth-century Europe and the resultant division of academic labor between the natural sciences on the one hand and the social sciences on the other (Latour, 1993; Shapin and Schaffer, 1985). Others have drawn attention to the influence of the protestant Reformation's denial of the efficacy of religious relics and the reality of transubstantiation as a repudiation of the agency of material things and their relegation to a more passive role as

bearers of humanly assigned 'symbolic' meanings (Keane, 2007). Yet others have suggested that anthropology's longstanding disinclination to engage materiality on its own terms must be traced back further to the beginnings of Western thought, in particular to the overturning of pre-Socratic cosmologies (like that of Heraclitus of Ephesus) based on material flows and transformations with a new emphasis on intelligible form as the basis of identity, knowledge, and classification, the latter view finding one of its most influential expressions in Aristotle's Physics, where it is argued that acts of creation, of making, involve precisely the imposition of form, *morphe*, on inchoate matter, *hyle* (Harrison, 1993; Ingold, 2011). These different periodizations nonetheless concur in seeing anthropology's neglect of the material world as symptomatic of a more long-standing and deep-seated Western intellectual prejudice and in arguing for a contemporary reengagement with materiality as a way of overcoming this. Indeed, a number of commentators have noted that recent work in material culture studies has often echoed the idioms of postcolonial theory in arguing for the 'emancipation' of material things from what is seen as a hitherto one-sided emphasis on culture, history, society, or the human subject (Fowles, 2010). Many contemporary debates about the status of material things and processes can be understood as centering precisely on the degree to which different conceptual and methodological approaches can be seen to have succeeded in achieving such an emancipation by supplanting the entrenched dichotomies of nature and culture and mind and matter against which such a project is staked.

Nonetheless, contemporary anthropology's awakened interest in materiality is not without historical precedents. Indeed, the emergence of anthropology as a discipline is, arguably, closely intertwined with the transformations of the material world brought about by the parallel histories of European colonial expansion and the development of industrial capitalism. For example, the nineteenth-century heyday of European imperialism that coincided with the beginnings of anthropology as an academic discipline also witnessed the influx into Europe and North America of quantities of non-Western artifacts from Europe's onetime colonies. These in turn came to form the basis of major ethnographic collections, like those of the Museum of Natural History (New York), the British Museum (London), the Pitt-Rivers Museum (Oxford), the Musée de l'Homme (Paris), the Ethnological Museum (Berlin), or the Russian Museum of Ethnography (St. Petersburg) and were extensively discussed in the writings of many of the founding figures of anthropology, like Franz Boas (Boas, 1983). Many of these same artifacts, exhibited in museums or sold in the open market would prove to be a source of inspiration for pioneers of artistic modernism like Braque and Picasso. As such, they would also facilitate a sharing of influences between anthropology early twentieth-century literary and artistic avant-gardes, as in Paris in the 1920s and 1930s. The resultant moment of 'ethnographic surrealism' (as James Clifford would later describe it) was one that brought together anthropologists like Alfred Métraux, Marcel Griaule, and the young Clause Lévi-Strauss and figures on the fringes of the Surrealist movement like Georges Bataille, in part on the basis of a shared curiosity about material artifacts (Clifford, 1981; Wilcken, 2010). It was this same milieu that also produced one of the classic works of economic anthropology, Marcel Mauss's Essay on the Gift (1925) a work exploring the mutual entanglement of human persons and objects of ritual exchange (mainly in the context of Pacific island societies) and often cited as an important influence on contemporary material culture studies (Mauss, 2000). The focus on artifacts characteristic of much of the nineteenth and early twentieth-century anthropology was, nonetheless, to a large degree displaced from the 1920s onward by the rise of fieldwork-based anthropology and, in particular, the approaches known as functionalism and structural functionalism. Rather than being treated as a potential source of cultural insights (as in the writings of Boas and others) material objects were often henceforward to be relegated to a merely utilitarian significance or else disregarded altogether in pursuit of more abstractly conceived understandings of social structure (Tilley et al., 2006a,b). More recent scholarship, however, has sometimes returned to reconsider the fate of colonial era and later artifacts and the transformations to which they were subject in traversing the different contexts of, for example, religious or ritual use, commercial exchange and display in museums or galleries (e.g., Thomas, 1991; Price, 1989).

Another important antecedent of present-day material culture studies is to be found in the writings of Karl Marx. Marx's influence on the social sciences is a varied and wide-ranging one (Bloch, 2010; Diamond, 1979). Nonetheless, two aspects of his work in particular seem especially pertinent to contemporary engagements with materiality and culture. The first of these is the version of materialism (sometimes referred to as 'historical materialism') that Marx developed through his engagements with the philosophical legacy of Hegel on the one hand and the history and political economy of industrial capitalism on the other. Marx's materialism was one that emphasized productive human activity and sought to understand social institutions and forms of consciousness on the basis of their derivation from modes of human interaction with the material world. In this respect, he distinguishes his own approach from what he takes to be the passive or 'contemplative' materialism of contemporaries like Feuerbach (Marx, 1978). Marx's writings are, therefore, concerned primarily with the productive transformation of materials by human labor rather than with the ontological status of materials themselves or the possibility that they might lead a life independent of the humans who make use of them. A second aspect of Marx's thought that is frequently referenced in contemporary debates about materiality and culture is his discussion of the fate of humanly manufactured objects under capitalist conditions of production. One aspect of the alienation experienced by workers under capitalism in Marx's view consists in their inability to recognize themselves in the products of their own labor, which are appropriated and sold for a profit by the owners of the means of production. Under capitalism, Marx suggests, objects are, increasingly, produced as commodities. That is, they are produced in order to be transferred to others by way of market exchange and to derive their 'exchange value' precisely from their relation to other commodities and

to the money form as a universal equivalent capable of expressing the market values of diverse products. In this respect, Marx's principal interest is, arguably, in what commodities have in common with each other as objects to be exchanged. Consideration of the sensuous particularity of objects is, rather, reserved for the discussion of 'use values' – that is, of the capacity of particular things to meet particular needs on the basis of their specific material properties. In contrast, later ethnographic studies of the social life of commodities have often introduced more extended consideration of the material and symbolic differences between different kinds of objects of exchange. Sometimes this has meant focusing on a particular commodity and its movement through an emerging capitalist world system (Mintz, 1986). Sometimes it has involved exploring the ways in which particular objects can be understood as shifting in and out of commodity status, for example, by being, bought, resold, given as gifts or incorporated into private collections, making it possible, so the argument goes, to consider gift-based and commodity-based economies as potentially contemporaneous and coexistent rather than successive moments of a linear historical teleology (Appadurai, 1988).

In recent decades, such a shift of emphasis from production to consumption has also been characteristic of much work done under the name of the interdisciplinary field of material culture studies, the development and rapid expansion of which is attested by the establishment of specialized academic programs, like the one at University College London and specialized publications, like the *Journal of Material Culture* (edited out of UCL and founded in 1996). Material culture studies is, however, a wide-ranging and heterogeneous field, practitioners of which have, at various times, defined themselves in terms both of breaks from and continuities with anthropology's past, acknowledging some antecedents (including Mauss and late nineteenth and early twentieth-century studies of ethnographic artifacts), while pointedly disavowing others. Structuralist influenced approaches, for instance, had previously sought to extend consideration of the social meaning of manufactured objects by locating them in a larger system of relations from which their individual significances are understood to derive. Lévi-Strauss himself collected quantities of indigenous artifacts during his fieldwork in Brazil (1935–39) and criticized functionalist approaches like that of Malinowski for their neglect of material culture. Several of Lévi-Strauss's own published works were devoted to the analysis of material artifacts, for example his study of the Northwest Coast Native American masks that he first encountered in New York's Museum of Natural History (Lévi-Strauss, 1982; Wilcken, 2010). From the 1960s onward, Lévi-Strauss's writings would in turn influence a number of studies of the consumer culture of Western, capitalist societies, such as Roland Barthes account of the 'fashion system' (Barthes, 1990). Nonetheless, despite obvious commonalities of subject matter, more recent work in material culture studies has often been explicitly critical of structuralist approaches, not least for their indebtedness to the linguistic theories of Ferdinand de Saussure (1885–1913). In particular, structuralistic accounts of material objects have been criticized for their emphasis on symbolic meaning as a relational effect of disembodied sign systems and their consequent neglect of the material specificities of the objects under consideration (Saussure, 1998). Webb Keane, for example, argues that Saussure's understanding of the linguistic sign, with its insistence on the arbitrary and conventional relationship between signifiers and signifieds, represents a version of what Bruno Latour has called the work of 'purification,' reflecting a characteristically modernistic desire to definitively separate meanings and materialities, humans, and nonhumans, to be accomplished in part by denying the agency of material things. Keane criticizes such a separation as being expressive of a characteristically modern, Western 'semiotic ideology.' Instead, he argues for a view of material objects as existing radically independently of human interpretations and actions and, therefore, capable of intervening on human affairs to unforeseen and surprising effect. As an alternative to Saussure, he proposes a renewed engagement with the semiotic theories of the American philosopher Charles Sanders Peirce (1839–1914), including the latter's insistence on the nonarbitrary relationship of certain signs to their referents, notably icons and indices, understood as linked to their referents on the basis of resemblance and physical contiguity, respectively (Keane, 2007; Latour, 1993; Peirce, 2011).

Alfred Gell too, in his widely cited *Art and Agency*, explicitly rejects the extension of Saussure's formulation of the 'arbitrary nature of the sign' into material culture studies on the grounds that such a move represents an 'over generalization of linguistic semiotics.' Developing a theory of art based not on aesthetics but on an analysis of the effects that artworks can be seen to produce, Gell argues, famously, that material things as well as human subjects can be understood as possessing agency not only insofar as they produce worldly effects but also because they can be viewed on that basis as indexing a prior intention to produce such effects (Gell, 1998). Subsequent commentators on Gell's work, however, have often expressed doubt as to the degree to which his account does, in fact, grant agency to material things in their own right. James Leach, for example, suggests that, for Gell, the intentions indexed by things are, ultimately, human ones, while Daniel Miller, another influential contemporary theorist of material culture, characterizes Gell's theory or art as one of 'natural anthropomorphism' in which the primary point of reference remains the humans behind the artifacts (Leach, 2007; Miller, 2005). Miller himself emphasizes rather the interactions between subjects and objects and the ways in which they can be understood to reciprocally constitute one another. In a series of studies of shopping and consumption, Miller invokes not Marx but Hegel to argue that such practices can be understood in terms of the externalization of human powers and intentions in the guise of institutions like the economy and the overcoming of such alienation through practices of consumption. In this respect, Miller differs from a number of other contemporary commentators on Marx and Hegel, notably the philosopher Slavoj Žižek, for whom Marx's historical materialism remains definitively grounded in the thought of Hegel (Žižek, 2013). Instead, rejecting Marx's vision of revolutionary change, Miller argues that this two-way movement of the externalization and reappropriation of human powers is a continuous and open-ended process rather something that will, eventually, be definitively resolved through the revolutionary transformation of capitalist relations of production (Miller, 2001). Miller's emphasis on the human appropriation

of material objects has been criticized by his colleague Christopher Pinney, who argues for an engagement with materiality beyond the assumed complementarity of subject and object and the human-centered understandings of context and history that have tended to go with it. Instead, Pinney invokes the work of two thinkers associated with the Marxist-influenced Frankfurt School, Theodor Adorno and Siegfried Kracauer (who, in his later writings, draws explicitly on the work of Lévi-Strauss) to argue for a 'nonhomogeneity' of time and an excess or 'torque' of materiality that can never be fully assimilated to the closure of 'culture' or 'history' (Pinney, 2005).

Another influential current in contemporary material culture studies has called upon the resources of phenomenology to explore the varieties of human sensory engagement in the material world. Drawing upon the philosophical work of, among others, Husserl, Heidegger, and Merleau-Ponty, such approaches seek to overcome received dualisms of mind and matter by taking as their starting point the experience of a human subject understood as always already embodied in the world. This approach has proven particularly fruitful in studies relating to landscape, like those of the British archaeologist Christopher Tilley, who has drawn on phenomenological literature to explore sensorial engagements with archaeological sites, particularly prehistoric structures where there are no supporting documentary sources to provide access to the intentions or cultural self-understandings of builders, occupants, and users. Instead, Tilley suggests that the sensing body can provide important clues to the interpretation of such sites by way of, for example, the experiential contrast between open and closed spaces or the mutual visibility of built structures within a particular landscape (Tilley, 1994, 2004).

However, if the burgeoning interdisciplinary field of material culture studies has been an exceptionally productive one in recent decades, it should also be emphasized that the range of approaches it represents does not by any means exhaust contemporary discussions of materiality and culture. For example, despite the emphasis of much phenemenologically inspired work in material culture studies on sensory experience as a way of dispensing with assumed distinctions between subjects and objects, this has often entailed little or no sustained engagement with the ways in which the biological and cognitive sciences have sought to understand the sensing body as a material entity. Indeed, phenomenological perspectives have sometimes been criticized by scholars working outside the field of material culture studies for presenting themselves as a more empathetic and humanistic alternative to the supposed reductionism of scientific accounts of the body and thus as failing to acknowledge or respond to the intellectual challenges that the latter might be seen to pose (Colebrook, 2010). It is worth noting, however, that a number of anthropologists influenced by phenomenology have recently begun to engage more explicitly with current research in the neurological and physiological sciences (e.g., Downey and Lende, 2013; Marchand, 2010).

Among anthropologists Tim Ingold has criticized material culture studies both for failing to overcome the separation of material realities and mental representations (continuing to treat them rather as distinct realms that need to be brought into relation through the work of analysis) and for its predilection for the study of readymade artifacts such as consumer objects rather than processes of making. Closer attention to such processes, he suggests, might yield an approach to the material world based not on discrete objects and the collective or personal significances with which they are supposedly endowed, but rather on material substances and their transformations, whether at the hands of human artists or craftspersons or through the actions of other than human agents such as meteorological processes. The term 'materiality' is, Ingold suggests, "too abstract to adequately convey the density, dynamism and variousness of materials, which are," he argues, "better understood with reference to a more expansively conceived ecology of 'life' encompassing a range of human and other than human presences" (Ingold, 2011). Ingold's arguments have been echoed more recently by Eduardo Kohn, in calling for an 'anthropology beyond the human' addressed to an 'ecology of selves' encompassing not only human subjects but also animals, plants, spirits, and entities of other kinds. Like Keane, Kohn suggests that, rather than Saussurean linguistics, Peircean semiotics, with its openness to semiotic modalities other than the symbolic, may be a more suggestive basis for engaging the nonhuman world (Kohn, 2013).

Thinkers seeking an alternative to phenomenology's emphasis on a sensing, experiencing human subject have often found inspiration in the work of the philosopher Gilles Deleuze (including his collaborations with the activist and psychoanalyst Félix Guattari). Deleuze's prioritization of becoming and the impersonal movement of 'A Life' of which both subjectivity and the individuated organism are understood as transitory products have proved suggestive not least in their traversing of familiar distinctions between the human and the nonhuman, the material and the ideal, the theoretical and the practical and have exercised a direct influence on, for example, Latour and Ingold. At the same time, Deleuze's commitment to the irreducibility of difference, refusing any appeal to transcendence or a hierarchy of being, combined with his insistence on the reality of concepts and the imperative to continuously create new ones offers, arguably, a strong provocation to rethink the terms on which anthropology engages with other humans and nonhumans (Deleuze, 1994, 2001; Jensen and Rödje 2010). In this respect, Deleuze's thought has been a significant influence on what has sometimes been called the 'ontological turn' in recent anthropological theory, along with the work of, among others, Marilyn Strathern (United Kingdom), Roy Wagner (United States), and Eduardo Viveiros de Castro (Brazil). Rejecting what they take to be a false and parochially Western distinction between a material world understood as given (or 'nature') and different peoples' variegated representations of it ('culture'), anthropologists associated with the ontological turn have argued for the need to 'take seriously' the concepts generated by the people anthropologists study. Rather than treating such concepts as representations (or misrepresentations) of a world assumed to be external to them, they argue (like Deleuze) that such concepts are efficacious worldly entities that contribute to the production of particular realities. One of the most sustained applications of this approach to the study of the material world is to be found in the volume *Thinking through Things* (Henare et al., 2007). Although referencing Latour's

account of the simultaneously purifying and hybridizing work of the 'modern constitution,' contributors to the volume also distance themselves from Latour's own universalizing ontological claims regarding the ubiquity of networks by arguing for a plurality of ontologies and for a methodology of anthropological engagement with the material world that might allow these to emerge. Rather than appealing to assumed social or cultural 'contexts' as a basis for understanding material artifacts and their implications in collective life, they suggest that the ethnographic study of 'things' (a term preferred to the allegedly more theoretically laden 'objects') might allow the latter to dictate the terms of their own analysis. Hence, for example, Martin Holbraad's account of the *aché* powder that Cuban *Ifá* diviners both use and talk about as constituting their divinatory power argues that *aché* should be understood as presenting a challenge to anthropology to revise its own theoretical lexicon by moving beyond the assumed Western opposition between concepts and things insofar as *aché* powder is simultaneously and irreducibly both a material substance and a power of divination (Holbraad, 2007). In a more recent essay, Holbraad argues that his earlier approach does not go far enough, concentrating as it does on the diviners' own definitions of *aché* rather than on the material qualities of the power itself (e.g., its motility and particulate composition) as directly informing its conceptualizations by its human users (Holbraad, 2011).

If the ontological turn argues for an understanding of materiality and artifactuality as variously emergent within different ontologies and world-making projects, other approaches have sought to depart from representational thinking by emphasizing the material world as a presence that potentially exceeds and disrupts human understandings of it. Jane Bennett, for example, has argued for a 'vital' materialism refusing any clear-cut separation between 'inert' matter and life. Drawing eclectically on science studies, complexity theory, and the philosophies of Spinoza, Nietzsche, and Deleuze, Bennett envisions a world in continuous becoming, populated by heterogeneous entities that are themselves emergent within the differentiating, world-forming flows of a single matter energy. A sensitivity to and an appreciation of the vitality of matter, she suggests, has potential political consequences insofar as it is likely, for example, to deter destructive and purely instrumental human engagements with the natural environment. At the same time (and in contrast to some phenomenologically inspired work) Bennett urges scholars in humanities and social sciences to pay closer attention to the new accounts of the material world being produced by the contemporary sciences, readings of literary, philosophical, and other texts being interspersed in her work with discussions of recent developments in fields such as evolutionary biology and nanotechnology (Bennett, 2001, 2010).

If one of Bennett's stated aims is to promote more sensitive and sustainable human engagements with the material world, other recent theorists have elaborated a vision of materiality as more radically autonomous of and indifferent to human projects and purposes. An important influence here has been the work of Georges Bataille (1897–1962), whose writings ranged over anthropology, economics, literature, and surrealism. Like some more recent and canonical theorists of material culture, Bataille draws upon the writings of Hegel, but does so not in order to resolve the dualism of mind or spirit and matter but to insist upon and accentuate it. Bataille's reading of Hegel (influenced in turn by the lectures of the Russian émigré Alexander Kojève) emphasizes not Spirit's cumulative realization of its own freedom but rather the unsurpassability of the negative, which Bataille identifies specifically as an autonomous power of matter, reasserting itself continually in the face of humanity's efforts to transcend it and as such suggesting the possibility of a 'virulent materialism' not subservient to any higher, 'spiritual' authority. Bataille's work can be understood as a series of attempts to articulate this power of matter through a variety of genres and media, including experimental essays, pseudonymously published pornographic fictions, reflections on Aztec human sacrifice, and European prehistoric art and through his uncompleted trilogy *The Accursed Share*, in which he sets out an alternative science of political economy (a 'General Economy') founded not upon assumed scarcity but on excess, expenditure, and the prodigality of matter-energy, understood as manifesting itself in the guise of superabundance if considered on a planetary or cosmic scale rather than a 'restricted' human one (Bataille, 1985, 1988; Kojève, 1980).

In the decades, since his death, Bataille's writings have proved intermittently influential in anthropology as in other fields, for example the work of the British philosopher Nick Land, who emphasizes the militantly antihumanist character of Bataille's project (e.g., Taussig, 1995; Ochoa, 2010; Land, 1991). More recently, Bataille (and Land's account of him) has been the subject of renewed attention by the philosophers and theorists associated with what has sometimes been called speculative materialism or speculative realism. The latter, encompassing a topically and stylistically diverse range of projects, has been characterized by an insistence on the imperative to think and describe a world understood as existing radically independently of its being given to a knowing, perceiving human subject. For example, the 'theory fiction' practiced by the (now US-based) Iranian thinker Reza Negarestani echoes Bataille's fictions and literary essays by splicing together narrative and philosophical reflection to evoke a vision of the contemporary Middle East and the oil that flows from its subterranean depths as animate, sentient entities, capable of interfering in a variety of ways with projects of human history-making (Negarestani, 2008). Quentin Meillasoux, meanwhile, avails of a more conventionally philosophical idiom to argue both for a renewed engagement with what he terms the 'ancestral' or the 'Great Outdoors' – that is, with a material world understood as existing prior to and autonomously of human thought and perception. The sciences, he suggests – in particular evolutionary biology and geology – present a challenge to philosophy precisely because they purport to describe such a world, making reference to events such as the emergence of terrestrial life that necessarily predate human subjectivity and, thus, the possibility of philosophical reflection (Meillasoux, 2009). In contrast, the work of Graham Harman, who has been credited with coining the phrase 'object-oriented philosophy' (or ontology), has been characterized by a closer and more sustained engagement with objects as such, which, for Harman, are not necessarily material artifacts but can comprise a variety of other entities, including philosophical movements. Indeed, Harman, has been explicitly critical of the writings of Deleuze and those influenced by him for dissolving the density and

opacity of objects into what they assume to be an underlying flux of becoming and of Latour's actor-network theory for subordinating objects to the relations that supposedly constitute them (Harman, 2009). Instead, Harman argues that objects are characterized by their capacity both to withdraw from engagement with humans and with each other and to enter into relations with one another without reliance on any mediating human presence (Harman, 2002). Harman, in turn, has been criticized by the literary scholar and cultural theorist Steven Shaviro for failing to provide a persuasive account of change, in particular of how objects (of what ever kind) come into existence (Shaviro, 2011). It is perhaps speculative materialism's rejection of the view that the material world is inescapably mediated by a human subject, language or culture that presents the greatest challenge to anthropology. Indeed, what Meillasoux, Harman, and others ask us to entertain is, precisely, the possibility of a world without humans. As a result, perhaps (and despite the wider currency their work has attained in other fields such as cultural geography and philosophy), many of these thinkers have been only occasionally cited in recent anthropological literature (e.g., McLean, 2013), although the work of Harman (including his book length study of Bruno Latour) has attracted attention in science studies and archaeology (e.g., Graves-Brown et al., 2013). Nonetheless, the debates initiated by speculative realism seem likely to have a significant influence on future research and writing.

The above inevitably represents a selective sampling of past and present debates rather than a comprehensive overview. It should be apparent nonetheless that discussions of materiality and culture encompass a broad, heterogeneous, and often fiercely contested terrain, on which a variety of methodologies and theoretical perspectives have been brought to bear. If the relationship between materiality and culture remains central to anthropology's understandings both of its subject matter and of itself as an academic discipline, it is a relationship that can perhaps never be definitively resolved, but is likely to remain rather a productive source of tensions and disagreements. Materiality, considered in relation to culture offers not the reassurance of a ground or a refuge of certitude but a constant source of new questions – questions that are of fundamental importance not only to the theory and practice of anthropology but also, more broadly, to human beings' understanding of themselves, their place in the world, and the relations they maintain with other entities of all kinds.

See also: Craft Production, Anthropology of; Diffusion: Anthropological Aspects; Everyday Life, Anthropology of; Exchange in Anthropology; Fetishism; Infrastructures of Society, Anthropology of; Modes of Production; Potlatch in Anthropology; Property: Anthropological Aspects; Sociality In Anthropology.

Bibliography

Appadurai, Arjun (Ed.), 1988. The Social Life of Things: Commodities in Cultural Perspective. Cambridge University Press, Cambridge.

Barthes, Roland, 1990[1967]. The Fashion System (Matthew Ward, Richard Howard, Trans.). University of California Press, Berkeley.

Bataille, Georges, 1985[1930]. Base materialism and Gnosticism. In: Allan Stoekl (Ed.), (Allan Stoekl, R. Carol, Donald M. Lovitt, Leslie Jr., Trans.), Visions of Excess: Selected Writings 1927–1939. University of Minnesota Press, Minneapolis, pp. 45–52.

Bataille, Georges, 1988[1967]. The Accursed Share: An Essay on General Economy (Robert Hurley, Trans.). Zone Books, New York.

Bataille, Georges, 1990[1955]. Hegel, death and sacrifice (Jonathan Strauss, Trans.). Yale French Studies 78, 9–28.

Bennett, Jane, 2001. The Enchantment of Modern Life: Attachments, Crossings, Ethics. Princeton University Press, Princeton.

Bennett, Jane, 2010. Vibrant Matter: Apolitical Ecology of Things. Duke University Press, Durham, NC.

Bloch, Maurice, 2010. Marxism and Anthropology: The History of a Relationship. Routledge, London.

Boas, Franz, 1983[1927]. Primitive Art. Dover Publications, New York.

Clifford, James, 1981. Ethnographic surrealism. Comparative Studies in Society and History 23 (4), 539–564.

Colebrook, Claire, 2010. Deleuze and the Meaning of Life. Continuum, London.

Coole, Diana, Frost, Samantha (Eds.), 2010. New Materialisms: Ontology, Agency, Politics. Duke University Press, Durham, NC.

Deleuze, Gilles, 1994[1968]. Difference and Repetition (Paul Patton, Trans.). Columbia University Press, New York.

Deleuze, Gilles, 2001[1995]. Pure Immanence: Essays on a Life (Anne Boyman, Trans.). Zone Books, New York.

Diamond, Stanley (Ed.), 1979. Toward a Marxist Anthropology. Monton Publisher, The Hague.

Dolphijn, Rick, van der Tuin, iris, 2012. New Materialisms: Interviews and Cartographies. Open Humanities Press, Ann Arbor.

Downey, Gerg, Lende, Daniel H., 2013. The Encultured Brain: An Introduction to Neoroanthropology. MIT Press, Cambridge, MA.

Fowles, Severin, 2010. People without things. In: Bille, Mikkel, Hastrup, Frida, Sørenson, Tim Flor (Eds.), An Anthropology of Absence: Materializations of Transcendence and Loss. Springer, New York, pp. 23–41.

Gell, Alfred, 1998. Art and Agency: An Anthropological Theory. Oxford University Press, Oxford.

Graves-Brown, Paul, Harrison, Rodney, Piccini, Angela (Eds.), 2013. The Oxford Handbook of the Archaeology of the Contemporary World. Oxford University Press, Oxford.

Harman, Graham, 2002. Tool Being: Heidegger and the Metaphysics of Objects. Open Court, Chicago.

Harman, Graham, 2009. Prince of Networks: Bruno Latour and Metaphysics. re.press, Melbourne.

Harrison, Robert Pogue, 1993. Forests: The Shadow of Civilization. University of Chicago Press, Chicago.

Henare, Amiria, Holbraad, Martin, Wastel, Sari (Eds.), 2007. Thinking through Things: Theorizing Artifacts Ethnographically. Routledge, London.

Holbraad, Martin, 2007. The power of powder: multiplicity and motion in the divinatory cosmology of Cuban Ifá (or *mana*, again). In: Henare, A., Holbraad, M., Wastel, S. (Eds.), Thinking through Things, pp. 189–225.

Holbraad, Martin, 2011. Can the Thing Speak? Open Anthropology Coperative Press, Working Papers Series #7.

Ingold, Tim, 2011. Being Alive: Essays on Movement, Knowledge and Description. Routledge, London.

Jensen, Casper Bruun, Rödje, Kjetil, 2010. Introduction. In: Jensen, C., Rödje, K. (Eds.), Deleuzian Intersections: Science, Technology, Anthropology. Berghahn Books, New York, pp. 1–35.

Keane, Webb, 2007. Christian Moderns: Freedom and Fetish in the Mission Encounter. University of California Press, Berkeley.

Kohn, Eduardo, 2013. How Forests Think: Toward an Anthropology Beyond the Human. University of California Press, Berkeley.

Kojève, Alexandre, 1980[1947]. Introduction to the Reading of Hegel: Lectures on the Phenomenology of Spirit (James H. Nichols Jr., Trans.). Cornell University Press, Ithaca.

Kroeber, Alfred, Kluckhohn, Clyde, 1952. Culture: A Critical Review of Concepts and Definitions. In: Papers of the Peabody Museum of American Archaeology and Ethnology, vol. XLVII, No. 1. Harvard University.

Land, Nick, 1991. The Thirst for Annihilation: Georges Bataille and Virulent Nihilism. Routledge, London.

Latour, Bruno, 1993[1991]. We Have Never Been Modern (Cathetrine Porter, Trans.). Harvard University Press, Cambridge, MA.

Latour, Bruno, 2013[2012]. An Inquiry into Modes of Existence: An Anthropology of the Moderns (Catherine Porter, Trans.). Harvard University Press, Cambridge, MA.
Leach, James, 2007. Differentiation and encompassment: a critique of Alfred Gell's theory of the abduction of creativity. In: Henare, A., Holbraad, M., Wastel, S. (Eds.), Thinking Through Things. Routledge, New York, pp. 167–188.
Lévi-Strauss, Claude, 1982[1975]. The Way of the Masks (Sylvia Modelski, Trans.). University of Washington Press, Seattle.
Marchand, Trevor H.J. (Ed.), 2010. Making knowledge: explorations of the indissoluble relation between mind, body and environment. Special Issue of the Journal of the Royal Anthropological Institute. Wiley-Blackwell, Oxford.
Marx, Karl, 1978[1845]. Theses on Feuerbach. In: Tucker, Robert (Ed.), The Marx Engels Reader. W.W. Norton and Co., New York, pp. 143–145.
Marx, Karl, 1990[1867]. Capital (Ben Fowkes, Trans.), vol. 1. Penguin Books, London.
Mauss, Marcel, 2000[1925]. The Gift: The Form and Reason for Exchange in Archaic Societies (W.D. Halls, Trans.). W. W. Norton and Co., New York.
McLean, Stuart, 2013. "Seaweeds and limpets will grow on our gravestones": on islands, time, death and inhuman materialities. In: Willerslev, Rane, Christiansen, Dorthe Refslund (Eds.), Taming Time, Timing Death: Social Technologies and Ritual. Ashgate, London, pp. 17–39.
Meillasoux, Quentin, 2009[2006]. After Finitude: An Essay on the Necessity of Contingency (Ray Brassier, Trans.). Continuum, New York.
Miller, Daniel, 2001. The Dialectics of Shopping. University of Chicago Press, Chicago.
Miller, Daniel, 2005. Materiality: an Introduction. In: Miller, Daniel (Ed.), Materiality. Duke University Press, Durham, NC, pp. 1–50.
Mintz, Sidney, 1986. Sweetness and Power: The Place of Sugar in Modern History. Penguin Books, New York.
Negarestani, Reza, 2008. Cyclonopedia: Complicity with Anonymous Materials. re.press, Melbourne.
Ochoa, Todd Ramón, 2010. Society of the Dead: Quita Manaquita and Palo Praise in Cuba. University of California Press, Berkeley.
Peirce, Charles Sanders, 2011[1940]. Philosophical Writings of Peirce. Dover Publications, New York.
Pinney, Christopher, 2005. Things happen: or, from which moment does that object come. In: Miller, Daniel (Ed.), Materiality. Duke University Press, Durham, NC, pp. 256–272.
Price, Sally, 1989. Primitive Art in Civilized Places. University of Chicago Press, Chicago.
Saussure, Ferdinand de, 1998[1972]. Course in General Linguistics (Roy Harris, Trans.). Open Court Classics, Chicago.
Shapin, Stephen, Schaffer, Simon, 1985. Leviathan and the Air Pump: Hobbes, Boyle and the Experimental Life. Princeton University Press, Princeton.
Shaviro, Steven, 2011. The actual volcano: whitehead, harman and the problem of relations. In: Bryant, Levi, Snricek, Nick, Harman, Graham (Eds.), The Speculative Turn: Continental Materialism and Realism. re.press, Melbourne, pp. 279–290.
Strathern, Marilyn, 1988. The Gender of the Gift: Problems with Women and Problems with Society in Melanesia. University of California Press, Berkeley.
Taussig, Michael, 1995. The sun gives without receiving: an old story. Comparative Studies in Society and History 37 (2), 368–398.
Thomas, Nicholas, 1991. Entangled Objects: Exchange, Material Culture and Colonialism in the Pacific. Harvard University Press, Cambridge, MA.
Tilley, Christopher, 1994. A Phenomenology of Landscape: Places, Paths and Monuments. Berg, Oxford.
Tilley, Christopher, 2004. The Materiality of Stone: Explorations in Landscape Phenomenology, vol. 1. Berg, Oxford.
Tilley, Christopher, Keane, Webb, Küchler, Susanne, Spyer, Patricia, Rowlands, Michael, 2006a. Introduction. In: Tilley, C., Keane, W., Küchler, S., Spyer, P., Rowlands, M. (Eds.), Handbook of Material Culture, pp. 1–6.
Tilley, Christopher, Keane, Webb, Küchler, Susanne, Rowlands, Mike, Spyer, Patricia (Eds.), 2006b. Handbook of Material Culture. Sage Publications, London.
Tylor, Edward Burnett, 1903[1871]. Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Language Art and Custom, 2 vols. John Murray, London.
Wagner, Roy, 1991. The fractal person. In: Strathern, Marilyn, Godelier, Maurice (Eds.), Big Men and Great Men: Personifications of Power in Melanesia. Cambridge University Press, Cambridge, pp. 183–202.
Wilcken, Patrick, 2010. Claude Lévi-strauss: The Poet in the Laboratory. The Penguin Press, New York.
Žižek, Slavoj, 2013. Less Than Nothing: Hegel and the Shadow of Dialectical Materialism. Verso, London.

Mathematical Concepts during Early Childhood Across Cultures, Development of

Inge Schwank, Faculty of Mathematics and Natural Sciences, University of Cologne, Cologne, Germany
Elisabeth Schwank, Institute for Cognitive Mathematics, University of Osnabrueck, Osnabrueck, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Approaches to acquiring comprehension of mathematical concepts and the linked ability of reasoning differ across history and culture. Research shows that a process view should be preferred to the more widely spread object view in order to benefit the formation of sustainable early mathematical concepts. Opportunities for cognitive development in the context of early mathematics education are among other things facilitated by mathematical playworlds that take the broad mathematical background seriously. The multifarious uses and possible resulting benefits of action-based mathematical playworlds are displayed within this contribution.

Introduction

Mathematics is basically about mathematical truths outlasting time. Mathematical concepts should fit together smoothly; furthermore mathematical truth needs reasoning: reasoning tools and reasoning objects. Everyday life does not tell much about mathematics unless one starts thinking mathematically. Look at the diagonal's length d of a square with the given length of s on each side (Figure 1).

There is no way to see, read off, or directly guess either the precise length of d in terms of numbers (the word 'number' refers to natural numbers) or the numerical relationship between d and s. Successful mathematical reasoning may lead to a mathematical equation, and in case of a specific number for s to a mathematical numerical value. Given $s = 1$ cm the number in question is something close to 1.4142135623730950488016887242096 98 with even more, actually an endless, number of digits to follow after the last 8. In fact, applying the well-known theorem of Pythagoras, which goes back for thousands of years, and using the knowledge about irrational numbers, which emerged only much later the argumentation easily follows: $d = \sqrt{2}$ cm.

The central question is, How to introduce early mathematical ideas to young children so that early mathematical reasoning is actually involved? Figure 2 shows a well-known approach for preschoolers to deal with a number setting. Obviously, without adult guidance focusing on numerical relations true mathematical understanding is unlikely to emerge. In order to reach this goal the use of special material suggests itself. Fortunately, mathematics is rule based. Rules are also part of games. One promising approach to our question is developing suitable mathematical playworlds. (Mathematical playworlds: Various action-based games strongly related to mathematical concepts used in early mathematical education in order to enhance the mathematical reasoning and comprehension as well as the cognitive capacity of children.) Thereby it is possible to demonstrate mathematical ideas and their relationships in a way everyday objects cannot achieve. Sticking to the rules or exploring what is happening when rules are changed is also part of nonmathematical games as a natural process of socialization.

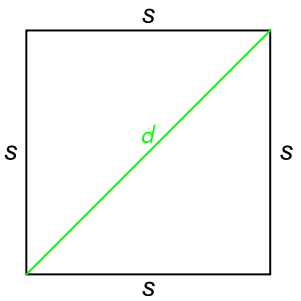

Figure 1 Given the length s in a square, what length is d? Also think of it in case of $s = 1$ cm.

Figure 2 Preschooler arranges beads on poles in a configuration from left 1 to right 10 beads.

Development of Mathematical Reasoning

Little is yet known about the development of mathematical reasoning. It is largely considered progressive to examine the so-called mental number line within various contexts. Investigating the interaction of language and cognition is widely spread, also in case of numerical cognition. Bender and Beller even go as far as to assume that "there may be no other domain in the field of cognitive sciences where it is so obvious that language (i.e., the verbal numeration system) affects cognition (i.e., mental arithmetic)" (2008: 215; see also Bender and Beller, 2011). Accordingly, investigating the link between the different cognitive tools used as number words and the underlying mathematical comprehension can be considered especially insightful.

It is often regarded reasonable to exert influence on children in a way that they first develop the number word sequence and learn counting afterward. Also, the five counting principles (Gelman and Gallistel, 1978) are often referred to. However, they lack in emphasizing the special significance of counting being particularly about managing numeric relations within a number space. The works of Starkey et al. point out that the focus in early childhood mathematics should be on building on the informal mathematical knowledge by creating, inviting situations within the framework of kindergarten/preschool and parental care instead of emphasizing formal knowledge too early (Starkey and Cooper, 1980; Starkey and Klein, 2008).

Predicative- and Functional-Logical Reasoning

Our long-term research about the development of mathematical thinking includes examining some basic principles of logical thinking. In this regard we adopted one of Spearman's task formats (Spearman, 1904) and specified it in terms of two different types of inferential logical reasoning:

- Predicative-logical reasoning: A type of inferential logical thinking that establishes relations by focusing on similar or identical components of objects.
- Functional-logical reasoning: A type of inferential logical thinking that establishes relations by thinking in action sequences and achievable effects by means of what differences in objects can be converted into one another.

Depending on the way of thinking the orientation in the world, types of basic sources for getting insight and construction of a world view are not the same.

Let us examine one of our matrices (Figure 3): In the bottom right corner there is a figure missing. Which one might fit in? Why?

Cognitive procedures allow establishing relations between the figures: specific figure components come to the fore while other components remain apparitional.

Mental Constructions with Accents

Predicative-logical reasoning establishes relationships by focusing on similar or identical components. Concerning the given task, in the first row the upper and lower lines 'belong together' but the side lines do not fit the pattern (Figure 4). This categorization generates order and structure. Continuing this cognitive appraisal results in a figure with lines from the bottom row (upper and lower) and the right-hand column (sides) respectively.

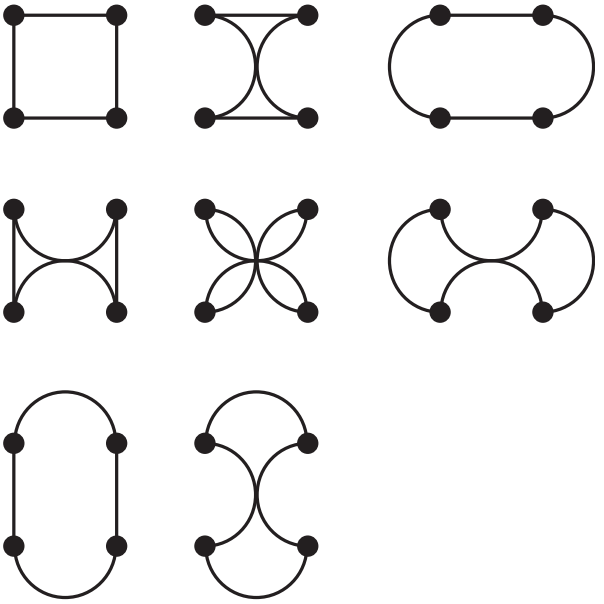

Figure 3 Task example from the Qualitative Diagnostic Instrument for Predicative-logical vs Functional-logical Thinking (QuaDiPF).

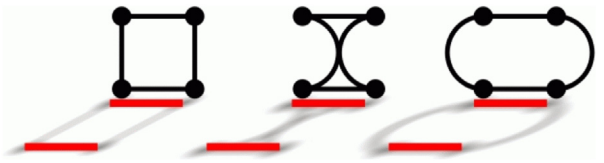

Figure 4 Mental construction with Predicative-logical focus.

Functional-logical reasoning establishes relations by thinking in action sequences and achievable effects by means of which differences can be transformed into one another (Figure 5). Concerning the given task, in the first row the side lines can be pulled inward and subsequently outward. This mentally creates a dynamic view of alterable figures. Continuing this cognitive appraisal results in a figure with the upper, lower, and side lines pulled outward. Actor-bound reasoning is required in order to produce result figures (Figure 6).

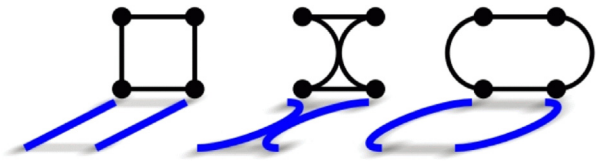

Figure 5 Mental construction with Functional-logical focus.

Figure 6 Importance of action-oriented reasoning in case of Functional-logical thinking.

In the context of two PhD projects, one run in Indonesia by Marpaung and the other in China by Xu, it turned out that predicative-logical and functional-logical approaches also occur in non-Western cultures (see e.g., Schwank, 1999, 2003).

In line with findings from psychological studies including several differing approaches and experimental setups (e.g., Kimura, 2000; Ceci and Williams, 2007), findings of our research show more boys scoring higher in functional-logical reasoning than girls. In early mathematical education which is strongly coined by arithmetical content, these differences in logic-based orientation should be kept in mind. Calculating means handling numbers. Numbers are being changed during calculating. Being able to understand numbers as arising from calculation processes is beneficial.

Reference Systems – M^4-Model

Interrelations play an important role. How ideas are generated is not clear yet, but reference systems seem to hold a key role since ancient times. One example for such a reference system is language. Another example is specific technology, invented by mankind for certain events.

Due to their innate nature technological systems can be used for tackling other problems while their originally intended purpose retreats to the background.

Let us have a look at the following example:

> Bells in China go back about 5,000 years. … [Particular bells] would once have been part of a set of either nine or fourteen. Each would have been a different size, and would produce two different tones, depending on where it was struck [from the outside with a hammer]. … But the role of bells in China could go far beyond the musical. To produce perfect tones they had to have absolutely standardized shapes, and the consistency of these shapes meant that the bells could also be used to measure standard volumes. And as the amount of bronze in each one was also carefully controlled, they could just as well provide standard weights. So a set of bells in ancient China could also serve as a sort of local weights and measures office, bringing harmony to commerce as well as society. (MacGregor, 2010: 193)

At first the goal was producing sounds with hollow articles based on the new technology of bronze work (Figure 7: Bronze Yong Bell).

The importance in this undertaking lay not so much in the individual sound as in the perfect harmonization of the various sounds generated by different bells. By succeeding in this task a system allowing a perfect attunement between the individual elements was developed which could then be used for other tasks like measuring volume or weight.

Using this normative structure it became possible to handle more precise indications of size, e.g., this volume is twice as much as that, this weight is four times as much as that one. The mentally possible continuation of building bells allows applying the bells' relations to situations for which actual bells could never have been built. The idea of manufacturability and thus usability of many further bells permits mentally reaching beyond reality. The sizes and their relations become elements of thought no longer relying on existing bells. However, the generated idea of construction originates from the original universe of bells. Therefore the system of bells is both source and root of this mental construction and consequently inseparably associated with it.

Figure 7 Chinese Bronze Yong Bell.

The basic idea of using reference systems can also be introduced into early mathematics education. The M^4-Model displays important structural connections (Figure 8). Initially, neither mathematical symbols nor mathematical terminology is known to the learner. The point is to generate mathematical ideas, made possible by dealing with mathematical idea-related events. This most important connection between these compartments is highlighted in purple within the model. Thereby these events should be envisioned just like the bells used in the example above.

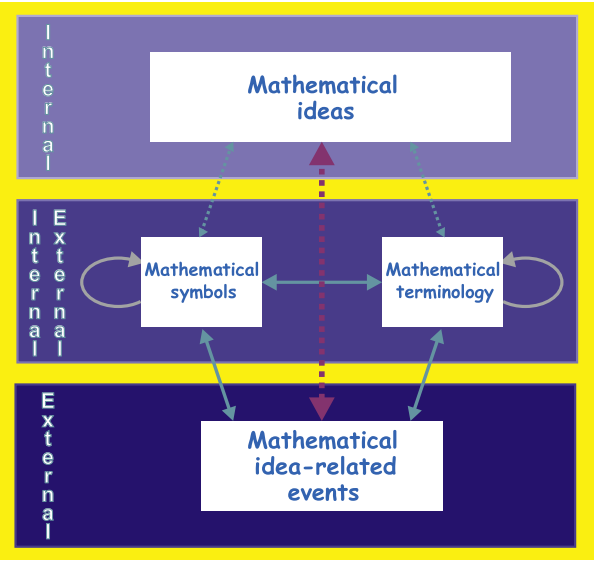

Figure 8 M^4-Model.

Preschoolers' gaining of mathematical insight is substantially dependent on to what extent the mathematical playworlds provided as cognitive stimulation actually contain mathematical elements. There is a lack of appreciation for playworlds developing mathematical thinking. The fundamental experience of relating things to one another – e.g., by means of number-related events – is of great importance to foster mathematical-logical thinking and reasoning. We assume this to be a formative fundamental experience that profoundly contributes to handling mathematics with pleasure.

Understanding Numbers

N^4-Model

Breaking the M^4-Model down into the number world results in the N^4-Model (Figure 9). (Based on Dehaene's triple-code model, the following four interacting compartments are displayed: number ideas, number symbols, number terminology, and number idea-related events. To do the requirements of mathematics education justice, the connection number ideas–number idea-related events is of major significance in the field of early childhood mathematics.)

The N^4-Model resembles Dehaene's triple-code model (1992). 'Number symbols' and 'number terminology' match Dehaene's 'visual Arabic number form' and 'auditory verbal word frame' but as opposed to Dehaene's "analogue magnitude representation" here 'number ideas' are suggested. Numbers and their multifarious relations are of great importance in terms of construction principles. Therefore, ideas or ways to arrange number relations are needed. The main difference is the addition of a fourth compartment: the 'number idea-related events' that is connected to the number ideas (connection marked in purple).

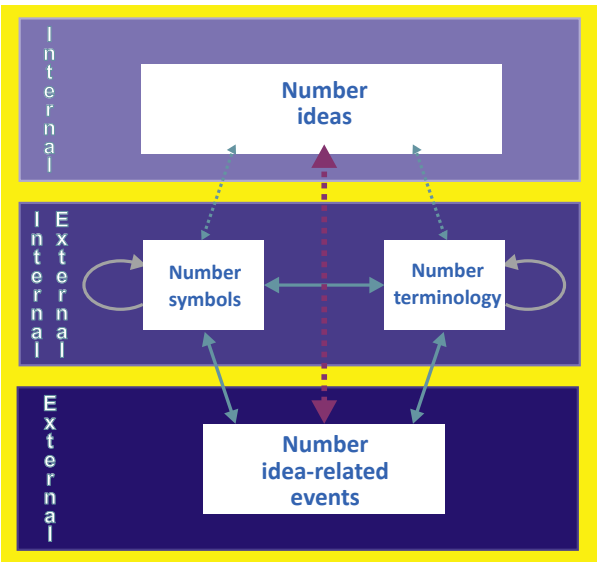

Figure 9 N^4-Model.

Internal and External Representation of Numbers

Number ideas express that numbers are mental constructions. Numbers cannot be seen or touched. Internal and external representations of numbers are necessary. Number terminology and number symbols form a part of these representations. The meaningfulness begins with analyzing the number idea-related events.

Number Terminology–Number Words

When jumping a couple of times or placing objects on a table saying 'one, two, three, …' number names are used in communication – the focus being on numerosity. There are huge differences among history and culture as to which words emerge and are used for numerical contexts (e.g., Ifrah, 1998). Presumably the starting point was the development of expressions not containing basic counting principles such as both, couple, pair, brace, score, gross, several, some, any, and so forth, depending on the specific situation within different cultures and languages.

Number Symbols

The success of mathematics is closely linked to the fact that mathematicians invent their own language to communicate about their newly conceptualized and deeply elaborated mathematical objects. Throughout the millennia people succeeded in developing a sign system for numbers facilitating symbolic calculation: the decimal place-value system. Due to the decimal system's inbuilt functionality only the 10 numerals 0–9 are needed to display any of the infinite numbers.

Fundamental Number Meaning

The significance of all, not only natural numbers, is undoubted (e.g., Dantzig, 1930; Brainerd, 1979). The scientific explanation of what the so naturally seeming numbers are about occurred at a relatively late point in mathematical history (e.g., Dedekind, 1969/1887).

As a matter of fact, every number itself is the result of a calculation. The basic meaning of a number is how often it has been counted forward by one, starting with zero, e.g.,

$$0 + 1 + 1 + 1 = 3 \quad \text{or} \quad 0 + 1 + 1 + 1 + 1 = 4$$

It should be beard in mind that it is not the number of 'ones'/'1' that count but the number of 'counting forward'/'+1.'

The key role in learning to understand numbers is not so much taken on by static situations than by dynamic events. Using not only number word sequences and especially not doing so too early would be beneficial in early childhood mathematics. The emphasis should be on cases of repetition. Therefore, simple verbalizations are needed for events such as something being added, e.g., 'one more, one more, one more.' The importance of a linguistic attendance of early mathematical concept formation certainly needs to be further elaborated.

Proper Names–Number Names

During unreflected everyday life, using number words as proper names, e.g., while head counting attendant children,

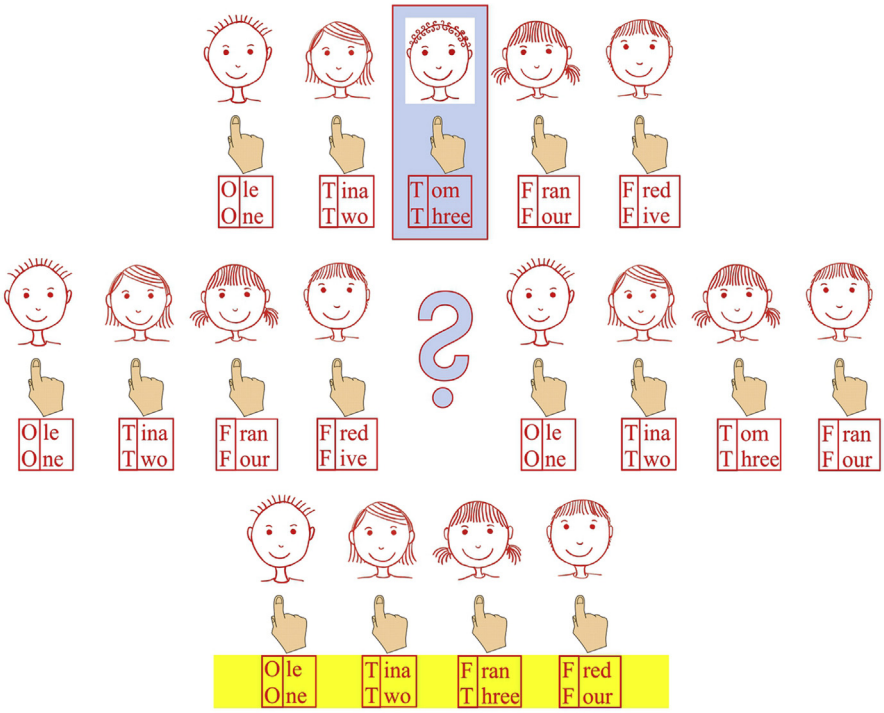

Figure 10 Important difference in the usage of proper names versus number names.

easily happens. However, numeral words are not proper names. Let us consider the head counting of five children in comparison to calling them by name (Figure 10). What happens, when Tom leaves the row?

The assignment of names does not change, but the assignment of numeral words does with all children following Tom. In this respect, only the assignment of words marked in yellow is appropriate. In order to learn the content of number words, environments of events are much more needed than environments of denomination.

Object View–Process View

In case of object view, counting refers to objects, e.g., plates on a table; whereas in case of process view, to incidents, e.g., sequences of action conducted by someone involved. (Most commonly spread, counting applies to measuring the number of elements of a set of objects. When learning calculation, an understanding of counting as engaging in actions of greater importance. The first case refers to object view, the second to process view.)

When plates are to be placed on a table, they have to be taken from somewhere. In the object world there arises the problem of a subtraction happening in the cupboard from where the plates are taken and an addition happening on the table the plates are put onto. In the process world it is a person's potential to carry out an action. Therefore an environment is needed in which this action can take place. It is true that the order of elements is not important during counting. This finding of invariance of order became known as one of the five counting principles by Gelman and Gallistel (1978). However, it may not be overlooked that this counting principle is not a basic principle useful for comprehending calculation. Every number has its fixed position in the number space. Hence, four always comes after three. With the help of the process view the number space can be mentally captured by means of an acting person in his specific environment.

Focusing on actions and changes as opposed to objects is challenging. Consequently it is beneficial to promote children's functional-logical view.

Let us have a look at the situation of playing a game of dice (e.g., ludo) with children. First, it is necessary to discuss with the children that the quantity of the diced dots states how many times one may jump forward with the counter. The goal is to count the jumps and not the pads that lend themselves to counting as well. This settles the question as to where the focus during the counting process should be: on the action executed by the counter.

This process shows, if children are unable to understand counting as counting forward by one, starting with zero, their approach to calculating is hindered. Furthermore, if no attention is paid to finding situations in which children experience counting forwards as well as counting backwards, their approach to subtraction is hindered in comparison to addition.

Building and Strengthening a Culture of Reasoning

For many years we have been running an academic meeting point for early enhancement in mathematics education called 'mathemagic.' The meeting point's goal is a holistic support and enhancement of early mathematics education, respectively. Particular emphasis is placed on children needing support or especially talented children and facilitating each child's individual cognitive approach to mathematics.

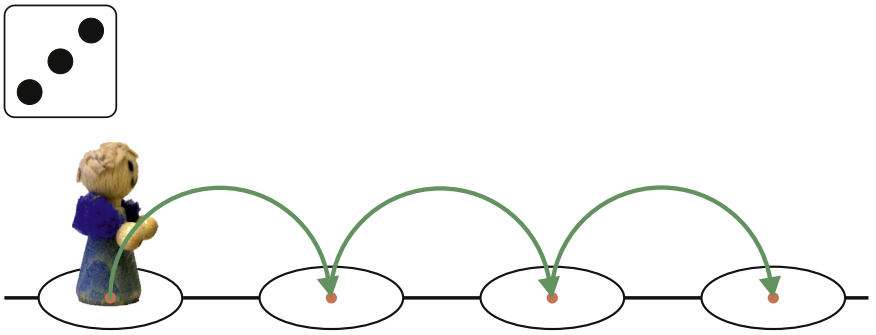

Figure 11 A three is rolled. This means, the counter may jump forward three times – thereby visiting all four pads.

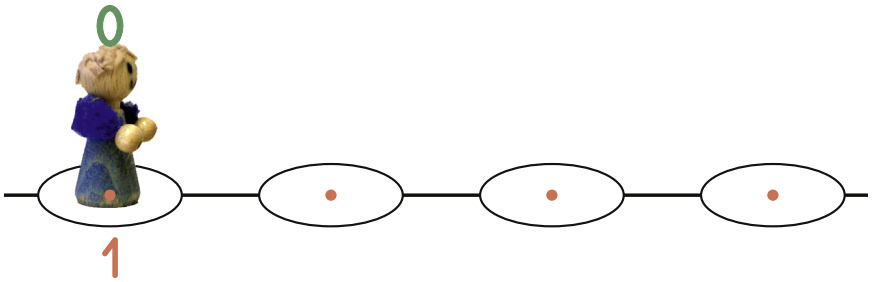

Figure 12 Initially, the counter is on the first pad. So far, it did not jump yet, i.e., zero times.

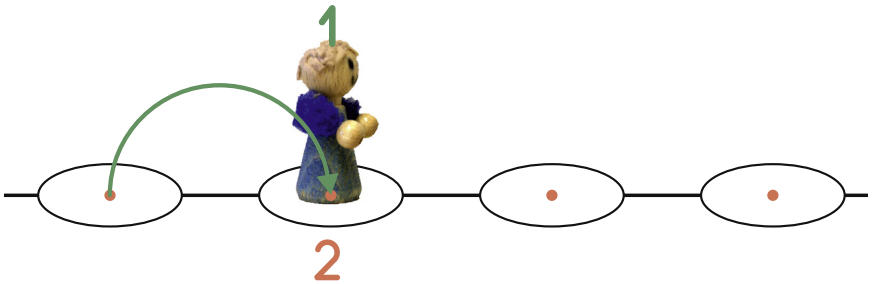

Figure 13 The counter jumps from the first to the second pad. It jumped once.

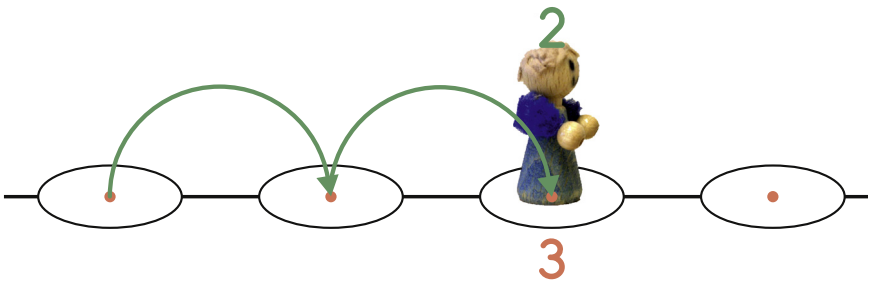

Figure 14 The counter jumps one more time. In total it has now jumped twice. It jumped from the second to the third pad. If pads were counted, one could stop now – matching the three dots on the die. This is a typical mistake when focusing on an object view instead of a process view.

Capacities of Mathematically Talented Children

One important undertaking of the meeting post was hosting an annual Mathematics Olympiad in which 2000 third graders from approximately 90 primary schools have taken part during the last 13 years. The Olympic tasks were designed to challenge children to give reasons for their solutions. Children's responses to the example 'Tiffany's tricky number task' (Figure 16) vary from no answer, a single numeral example without any reason (e.g., 'If $4 - 2 = 2$, then the task

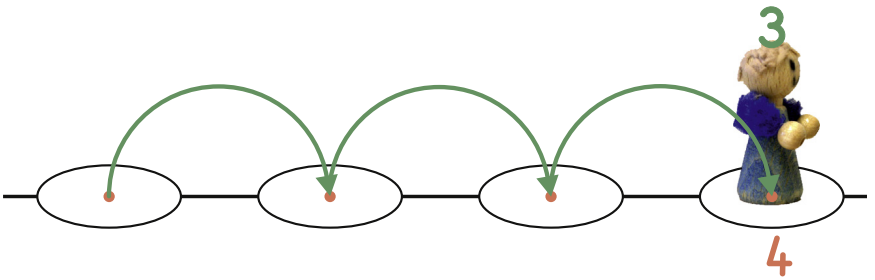

Figure 15 The counter jumps one more time. In total, it has now jumped thrice. This time, it jumped from the third to the fourth pad. By performing three jumps according to the diced dots, the move is finished.

fits.'), stating several possible numeral examples without special result relation (e.g., '4 and 2, 6 and 4, 3 and 1, 2 and 0, etc.') or with stronger relation to the construction of the result (e.g., 'Tiffany could: $3 - 1 = 2$, $4 - 2 = 2$ and doing until ∞ – nearly $\infty = 2$. Tiffany could not: $3 - 2 = 2$ or $5 - 100 = 2$ and $5 + 0 = 2$.') to complete conceptual linguistic examples (e.g., 'All, as long as one of them is two numbers higher' or even 'The numbers always have to be apart as much as the result is.'). Only in the last case the relational nature of the numbers is successfully described in a way that allows capturing all (!) possible number pairs that satisfy Tiffany's first equation.

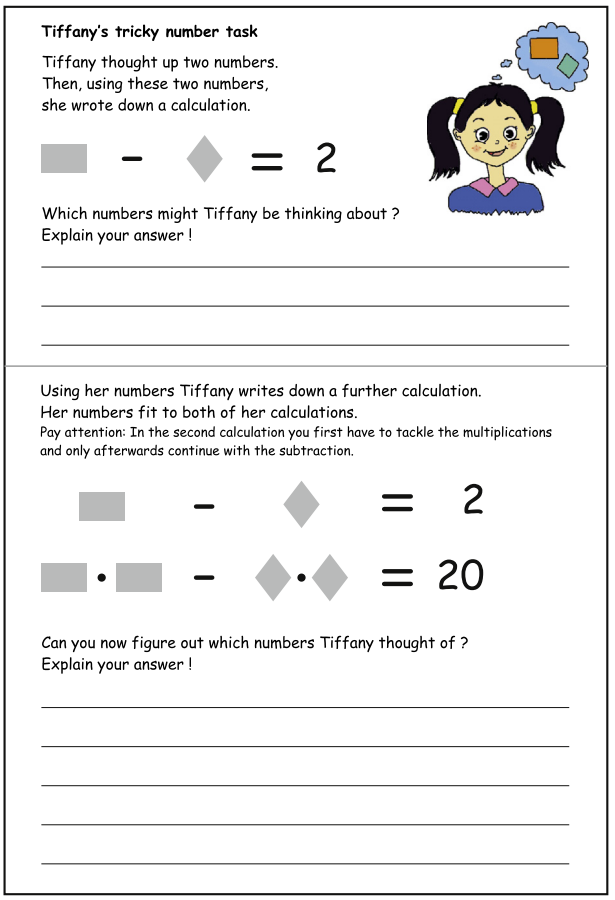

Figure 16 Task example from the Meeting Point's Mathematics Olympiad for third graders.

This displays the profound argumentation skills achievable by mathematically talented children. However, there still is a demand for initiation, nurture, and maintenance of an improved culture of reasoning, already with young children. In formal education, more importance has to be attached to not just reciting and testing number sentences, but on fathoming the connections of validity of these number sentences.

Mathematical Playworlds Boost Reasoning

The art in teaching an understanding of numbers lies in providing children with adequate arithmetic experiences based on counting. Psychological studies indicate that already infants can differ between one, two, or three objects, but are unable to differ between approximately four to six objects. Also, infants notice when one object is added to another or when one of two objects is taken away (e.g., Geary, 2007).

Therefore the first interesting number in terms of counting onwards is four. Numbers are systematic; so children should become acquainted with numbers in a systematic environment.

Event-Related Number Space Orientation

In the ENSO (event-related number space orientation) playworld (Figure 17), it is possible to explore the number range from zero to nine. This suits the later to be learned numeral diction of the decimal system. Cover sheets are used to focus on playing in a certain area, e.g., in the beginning just the number range from zero to four.

The ENSO playworld is inhabited by Hoppies and their animal companions who all travel by jumping. The number space is created by means of 10 places with ropes serving the purpose of holding balls. There are no balls at the starting space; no jump has been carried out yet. From this point onwards, there is always one more jump (basis for addition) or one jump back (basis for subtraction) possible. The balls on which a counter is standing can be counted, but also – and more importantly – the number of jumps necessary to reach a certain destination. Alongside the ropes runs a wooden plate serving as a path where jumps are also possible.

One Hoppy could e.g., start with zero balls and do three jumps. Simultaneously, an animal companion could do three jumps as well, also starting with zero, but on the path. If all goes

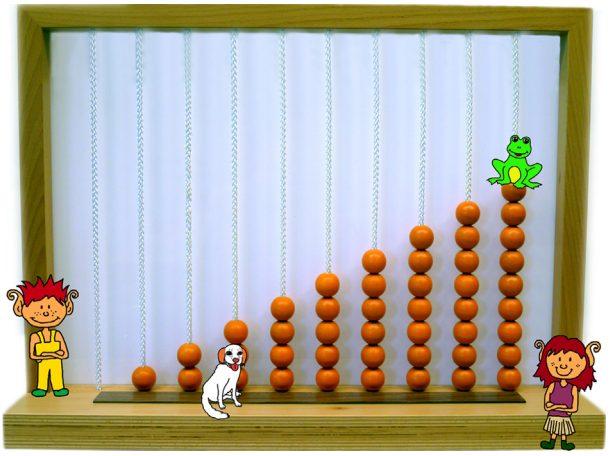

Figure 17 Mathematical playworld 'ENSO' (event-related number space orientation).

Figure 18 Number line from 0 to 3.

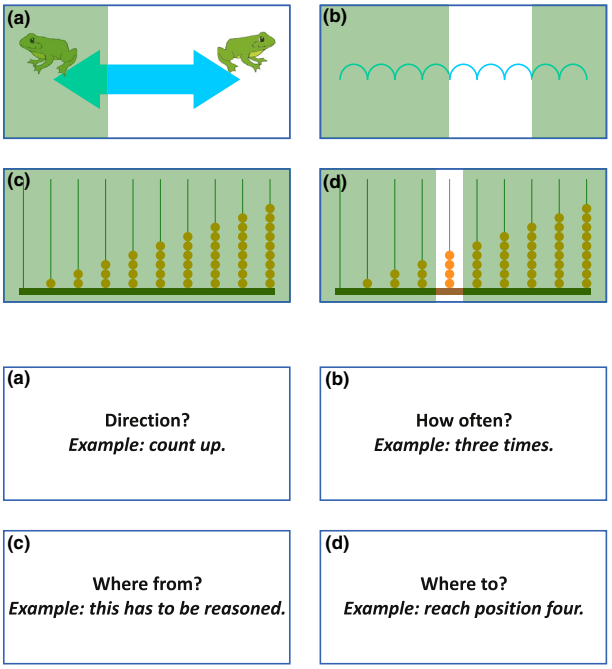

Figure 19 Logic game 'TooFP' (three out of four panels). (Example: x+3=4; 'Where from' (c) is unknown).

well Hoppy and his friend can wave at each other because their destinations fit together. This procedure can lead to an understanding of the number line. Also questions such as why it makes sense to count 'just' three when there are four lines can be discussed (Figure 18). One clever thing about the ENSO world is that the path itself does not contain lines that could be seen as countable objects and thereby hindering the process view.

Number Range Extension 'Zero–Four'

When jumping moves in the number range from zero to three are flexibly conducted – forward and backward, starting at different places – an extension suggests itself. The jumping game is enlarged by executing three jumps from the start and then one more jump. This process is analogical to the familiar situation of executing one more jump after having done two jumps. At first, a word is missing: The word for expressing how many jumps are necessary to reach the new destination from the starting point. The new word associated with this series of events is 'four.'

Components of Calculations are Connected

Various game actions are possible within the ENSO world. A game involving several number relations at once is the logic game 'TooFP': Three out of Four Panels.

The basic idea is: 'Define three of four panels. Draw conclusions' (Figure 19).

By means of the upper four display panels (Figure 19), movements of the counters in the ENSO playworld can be described or prompted. Sliding plates (displayed in the upper drawings through the green shaded areas) allow settling agreements. The lower four arrays verbalize the reading of the panels.

In the given example (Figure 19) the task is: $x + 3 = 4$ as shown by the area marked in white and the green shaded area, respectively.

The 'direction' of movement as shown by the panels has been set: 'count up' (a). Also, it has been set 'how often' the move should be performed: 'three times' (b). The starting point ('where from' (c)) is unknown and has to be found out.

Lastly, the final position is defined ('where to'(d)). In the example, it is 'four.'

By dint of this information it can be reasoned where the counter came from. The definitions also allow exploring possibilities and impossibilities. This kind of reasoning boosts the logical intellectual power. One advantage of mathematical playworlds lies in the possibility to discuss mathematical situations long before the script needed in formal calculation has been learned. At first, children are able to do a lot more mathematical thinking than they could do mathematical writing. This potential should be used for the children's benefit.

Spiral Stairs of Calculation

A further investigation of the decimal system can be achieved with the help of the mathematical playworld called spiral stairs of calculation (SSC), where the counters can move within the number space from 0 to 19 (Figure 20).

Due to the circularly arranged poles with balls, in the inner circle from 0 to 9, in the outer circle from 0 to 19, strategic counting operations can be considered. A counter on the inner circle could e.g., start on three balls and do two jumps upwards to reach five balls. Now it can contact its animal friend on the outer circle that starts on 13 balls and also jumps upwards twice and tell him on which number of balls his

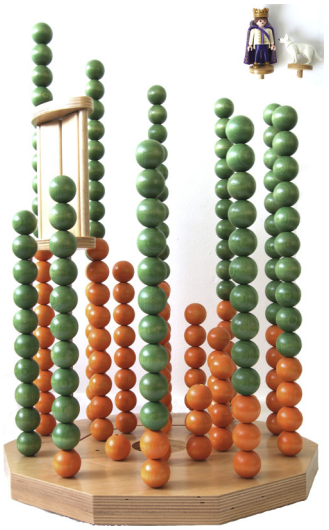

Figure 20 Mathematical playworld 'SSC' (spiral stairs of calculation).

move will end, on 15. By means of the calculation $3 + 2 = 5$ and the underlying decimal construction the result of the calculation $13 + 2 = 15$ can easily be deduced. Structural features of the decimal system are elucidated by the color of the balls. The balls in the inner circle are orange, just as the balls from 0 to 9 on the outer circle. The following 10 balls on the outer circle are green though. Further 'easy' calculations where the result can 'easily' be deduced can be looked for and found. Examples could be games for the situations $4 + 10$ (just jump at the corresponding place on the outer circle) or 13 – 3 (just jump on the outer pole with no orange balls) and many more.

The inner circle of the SSC can already be used with preschoolers. In the following section we are going to demonstrate two characteristic discussions from the context of our play lessons.

Calculations Are Connected

In the present situation, several poles have already been completely equipped (Figure 21). In the up to now last step it has been found out that four more balls would fit nicely to the five balls that already have been placed on the nine-ball pole ($5 + 4 = 9$). Now five balls are handed out for the eight-ball pole, the task being to consider, how many further balls can be added.

A 5-year-old preschooler explains in this playworld where no formal notation is used that the answer to '$5 + 3 = 8$' must be 'three.' The reason being: during the beforehand discussion of the task '$5 + 4 = 9$' it had turned out that the missing number of balls is 'four.' As an introduction to its statement the preschooler points out 'I knew that!' after the correct answer 'three' had already been given by another child at the playing table. His inference is simple: For nine balls only one more ball and so with eight balls only three balls have to be added. Without having the eight balls before his eyes the child has an overview in the mathematical playworld that '$9 = 5 + 4$,' '$9 = 8 + 1$,' therefore '$8 = 5 + 3$.'

Figure 21 Filling the spiral stairs of calculation using certain stages.

Does It Make Sense to Think of Zero as an Even Number?

Zero deserves special attention because its understanding is fundamental for the use of the decimal place-value system. One counts from 0 to 9 per place and then starts all over again. In the SSC there is a fixed space for zero balls: the place where the space for balls is one ball smaller than the space for balls at 1 ball, just like the space for balls at 1 ball is smaller than the space at 2 balls. In this way the use of zero can be implemented in the arrangement of thoughts. While walking the SSC downwards, connumerating the place can easily be denominated, as well as it should be denominated when walking upwards. And of course, a counter standing on the starting place, has not jumped yet, i.e., it jumped zero times.

In a game situation with preschoolers the discussion arose as to where in the SSC it would be possible to place a foam-rubber slate so that the number of balls under and over the slate is equal (the guiding question being 'Where else is this possible?'). Firstly, the poles with 1–9 balls were successfully investigated (Figure 22), afterward the space for zero balls was given consideration.

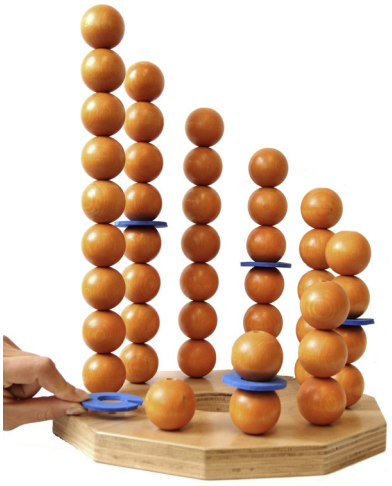

Figure 22 Exploring some properties of numbers using blue foam rubbers.

Group A:		
Teacher:	I tell you there is another place…	
John:	Yes, I know one…	
Teacher:	Which one, John?	
John:	The zero, *because* zero is at the bottom…and at the top.	[Laughs.] [Gestures at the starting place and for the lower zero balls.] [Gestures at the starting place for the upper zero balls.]

Group B:		
Teacher:	I tell you, that…	
Rita:	Ah!	[Raises her hand.]
Teacher:	…it works *again*.	
Rita:	I know it…	
Nora:	The zero, *because* then everyone gets nothing.	[Nora gets a foam-rubber slate from the teacher.]
Teacher:	Rita, what was it, you wanted to say?	
Rita:	Zero.	
Teacher:	Why?	
Rita:	*Because*, uhm, then there is nothing on either side.	

Similar to Rita's argumentation is Tina's from Group C, but she makes a remarkable addition:

Tina:	Uhm…*because*, *because*, *because* with zero, there, there, there is nothing on each side. *So* it is equal.

The apodosis 'So it is equal' nicely shows that Tina can work with 'nothing' as a number.

Mathematical ideas are grounded on reason and argument. By means of mathematical playworlds, children are introduced to applying a connected series of statements to establish mathematical meaning. Thereby they get a better understanding of the realm of numbers.

Concerning early childhood mathematics, the question is how children can learn the usage of and the necessity for words to articulate reasons and in which kind of situations this may be learned. Internationally there still is not very much knowledge about these questions, so research is dearly needed.

Place-Value System: Mathematical Playworlds as a Mediating Tool for Understanding Ideas Embedded in Mathematical Symbolism

Understanding the functioning of the place-value system that builds the foundation for the common notation of numbers and the algorithms for the basic calculation operations learned at primary school is a cognitive challenge. In our days, this notational system is used almost worldwide to operate with natural numbers – on paper, using electronic calculation devices or last but not least the computers. Children of cultures like China or Japan seem to have some advantages in getting familiar with this mathematical concept because their native language fits better than others to the logical construction of this rule-based symbolic notion (e.g., Miller et al., 2005). In Western countries, major problems of kids' usage of the place-value system in calculation tasks are discussed (e.g., Ashlock, 2010; Fuson et al., 1997; Haylock and Cockburn, 2013; Kamii, 1999). Children e.g., do have difficulties to connect the written symbols with objects (Figure 23 left) or to use calculation methods such as the addition algorithm (Figure 23 right).

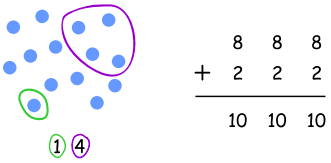

Figure 23 Example of possible difficulties when learning the decimal place-value system.

Usually people suggest improving children's thinking in structures of 10 to understand the place-value system and therefore using material similar to base-ten blocks or ten-sticks and dots to help children. For this purpose, the instructions for multidigit addition and multidigit subtraction in vertical layout are strongly based on carrying and borrowing together with the principles of exchange and decomposition. Little 1s come in, e.g., in case of exchanging 10 tens to 1 hundred, then a little 1 is to be written at an appropriate place in the hundreds column. From a mathematical viewpoint the usage of material like base-ten blocks has to be put into question. Little 1s do not make sense. 1 hundred, of course is much greater than 1 ten or 1 unit. It has to be noted that the decimal system solely allows the digits from 0 to 9 at each place. Exchange and borrowing are not provided for in terms of the notation. The decimal system is a counting device in which counting forward and backward is noted in a special way (e.g., Schwank, 2013). Its functionality is displayed in the following mathematical playworlds.

Stellanian Accounting System for Their Balance of Starthaler

The first Stellanians are Hella, Tella, and Ulla (Figure 24).

They are responsible for the Stellanian accounting system using special rung ladders for recording Starthaler (Figure 25).

There are rules for how to use this particular accounting system

1. Each of the three Stellanians has its own ladder. None of them may leave their ladder during a recording process.
   a. Jump from rung to rung with the Stellanians. How often can each of them jump from the bottom to the top? And back again?

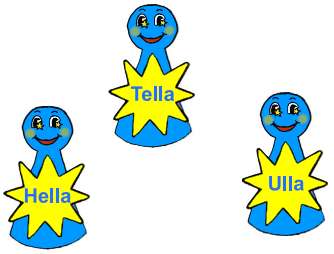

Figure 24 The Stellanians Ulla, Tella, and Hella.

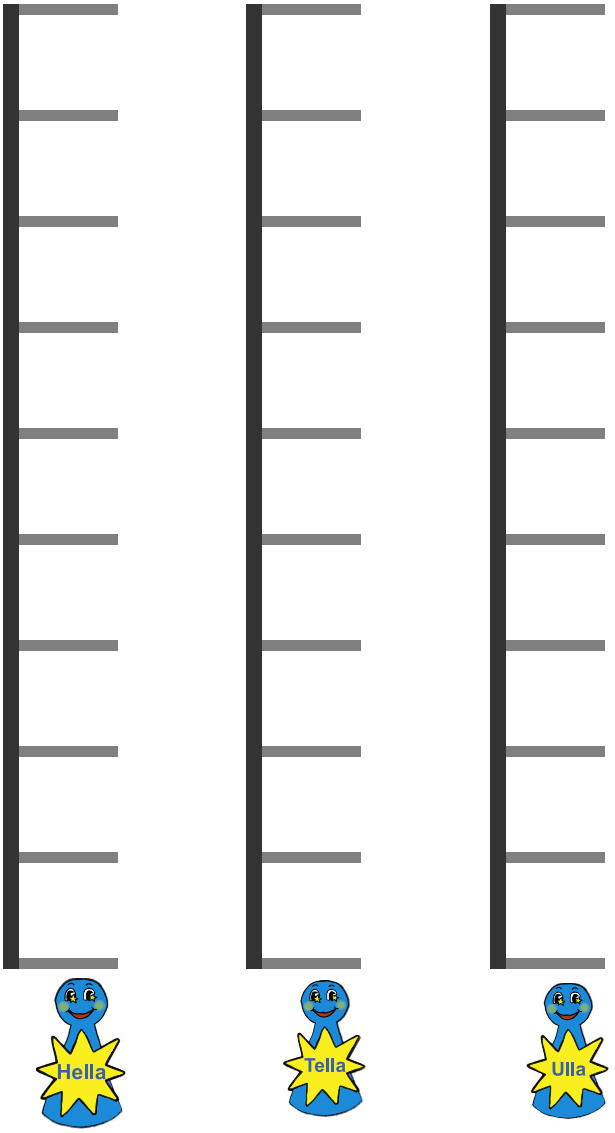

Figure 25 Ulla, Tella, and Hella positioned at their rung ladders.

2. If Ulla jumps one rung upward this means the balance increases by one Starthaler. If Ulla jumps one rung downward, the balance decreases by one Starthaler.
   a. What balance can Ulla display?
   b. Which changes can Ulla record?

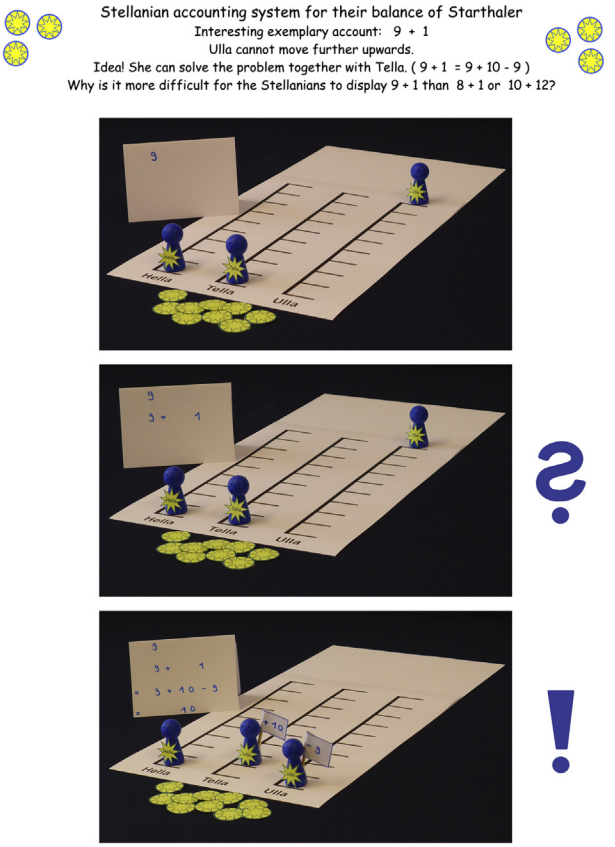

Figure 26 How can Ulla and Tella cooperate to solve the task 9+1?

3. Tella's case is similar to Ulla's. But when she jumps from rung to rung, the balance increases or decreases by 10 Starthaler.
   a. What balance can Tella display?
   b. Which changes can Tella record?
4. In Hella's case the jumps are worth 100 Starthaler!
   a. What balance can Hella display?
   b. Which changes can Hella record?
5. The three of them can work together!
   a. What balance can they display together?
   b. For recording changes together they have to think and cooperate carefully.
      – Find cases of change in the amount of Starthalers that they may record together easily.
      – If the amount of 23 Starthaler is increasing by 14 Starthaler the recording is rather simple, right? But if the amount of 9 Starthaler is increasing by 1 Starthaler, there is a problem, is not it? (Figure 26)
      – If the amount of 578 Starthaler is decreasing by 467 Starthaler, this is a great loss of Starthaler, but the recording is easily done. Why? But if the amount or 10 Starthaler is decreasing by 1 Starthaler, the loss of Starthaler is only little, but the recording turns out to be challenging. Why?

The rung ladders provide an overall system for displaying all numbers from 0 to 999 (Figure 27: left). The transition to decimal script is achieved by labeling the individual rungs,

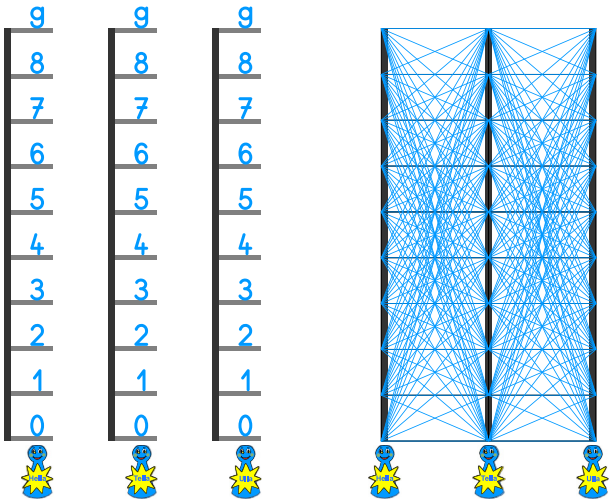

Figure 27 Possible numbers that Ulla, Tella and Hella can display (from 000 to 999).

e.g., the lowermost rungs are labeled 0. This is the Stellanian's starting position where no jump has been carried out yet. All further rungs on a ladder are labeled according to the number sequence, ending with 9 as label for the last rung. From the lowermost to the uppermost rung nine jumps are necessary.

The combinations in which Hella, Tella, and Ulla can take a seat on a rung add up to 1000 possibilities (Figure 27: right). The blue lines display that Tella has 10 options to choose a rung for herself according to every rung that Ulla has taken a seat on. This means, they have on the whole 100 possibilities. Now Hella, for her part, can choose a rung for herself according to each of these 100 possibilities, resulting in a total of 1000 possibilities.

The comprehension of numbers is based on the thorough immersion in the functioning of the place-value system by mastering the mental arrangement of the Stellanians on their rung ladders (Figure 28).

In advanced lessons, using the lesson format of dialogic learning (Gallin, 2012), children can experiment with changing the number of rungs. They can convince themselves that the number of rungs is variable: It is possible to choose from two rungs to any higher number. The already known principles of the decimal representation of numbers as well as the principles of decimal calculations function analogously.

Conclusion

Obviously, teaching preschoolers to know that the quality of being an even number can be attributed to zero is not the point. Nor is it that people should know the number to express the length of the diagonal in the introductory example up to 10 or more decimal places.

Nonetheless, we have advisedly chosen these examples. The goal was presenting something mathematical that adults familiar with everyday mathematical situations do not simply take for granted. Hence, there is a need for educated inferential argumentation. Concerning mathematical learning processes, thinking theoretically through

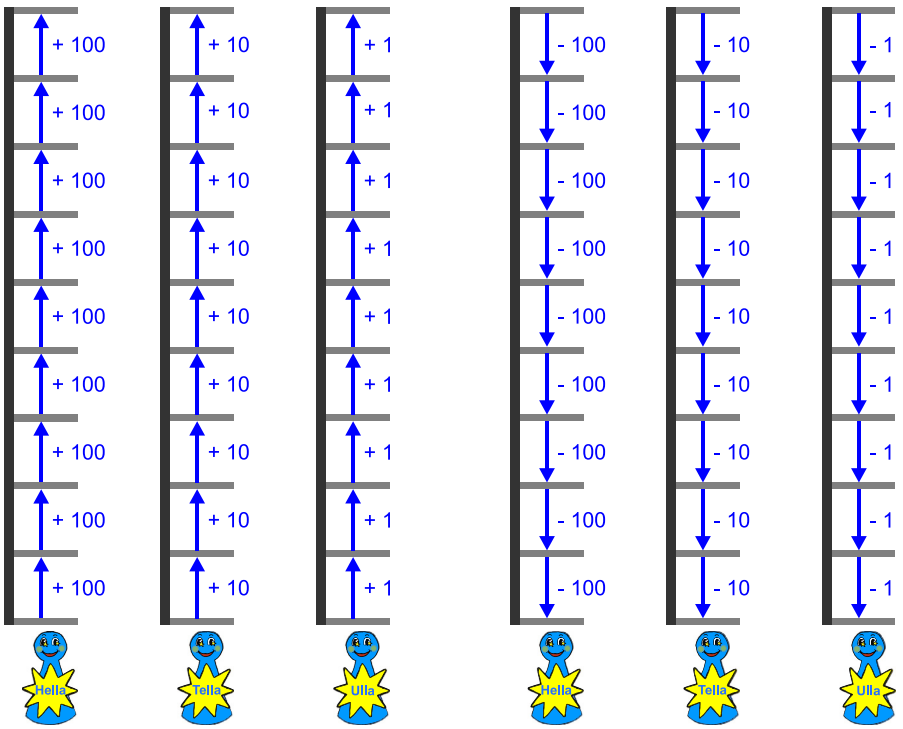

Figure 28 Possible plus and minus actions of the three Stellanians.

a problem by means of logical reasoning is essential. Children need occasions to get familiar with mathematical concepts in a reflective way in order to develop their mathematical thinking.

We are convinced that – just as children are obviously able to learn any language as their mother tongue – children can, in principle, also learn the universally valid ways of expressing elementary mathematics. From the viewpoint of cognition, this learning opportunity is initially independent of cultural affinities. The question is to what extent children are conceded to educational opportunities and to what extent these are professional.

This kind of professionalization certainly asks for further research. For the future it would be eligible to run an increased amount of national as well as cross-cultural studies that help exploring ways in which young children can acquire the ability of mathematical reasoning and thereby develop mathematical concepts. For that matter the focus should be not so much on learning and thinking results but in learning and thinking processes. Moreover, further methods and procedures to positively influence mathematical reasoning are to be developed. One promising starting point might be providing learning environments rich in mathematical content such as mathematical playworlds.

Acknowledgment

We are grateful to our meeting point's main donator Stiftung Stahlwerk Georgsmarienhütte.

See also: Cognitive Development During Infancy and Early Childhood across Cultures; Concept Formation in Childhood: Cross-Cultural Perspectives; Logical and Hypothetical Reasoning in Adolescence, Development of; Problem Solving During Infancy and Early Childhood, Development of; Scientific Concepts During Childhood, Development of.

Bibliography

Ashlock, R.B., 2010. Error Patterns in Computation, tenth ed. Allyn & Bacon, Boston.

Bender, A., Beller, S., 2008. The Limits of Counting: Numerical Cognition Between Evolution and Culture. Science 319, 213–215.

Bender, A., Beller, S., 2011. Cultural variation in numeration systems and their mapping onto the mental number line. Journal of Cross-Cultural Psychology 42 (4), 579–597.

Brainerd, C., 1979. The Origins of the Number Concept. Praeger, New York.

Ceci, S.J., Williams, W.M. (Eds.), 2007. Why Aren't More Women in Science: Top Researchers Debate the Evidence. US: American Psychological Association, Washington, DC.

Dantzig, T., 1930. Number, the Language of Science. The Macmillan Company, New York.

Dedekind, R., 1969/1887. Was sind und was sollen die Zahlen? Vieweg, Braunschweig. The nature and meaning of numbers. http://www.gutenberg.org/files/21016/21016-pdf.pdf (accessed 26.07.14.).

Dehaene, S., 1992. Varieties of numerical abilities. Cognition 44, 1–40.

Fuson, K., Smith, S.T., Lo Cicereo, A.M., 1997. Supporting Latino first graders' ten-structured thinking in urban classrooms. Journal for Research in Mathematics Education 28 (6), 738–766.

Geary, D.C. 2007. Development of mathematical understanding. In: Kuhl, D., Siegler, R.S. (Vol. Eds.), Cognition, Perception, and Language, vol. 2. In: Damon, W. (Gen. Ed.) Handbook of Child Psychology. John Wiley & Sons, New York.

Gallin, P., 2012. Dialogic learning – from an educational concept to daily classroom teaching. In: Baptist, P., Raab, D. (Eds.), Implementing Inquiry in Mathematics Education. Bayreuth University Press, Bayreuth.

Gelman, R., Gallistel, C., 1978. The Child's Understanding of Number. Harvard University Press, Cambridge, MA.

Haylock, D., Cockburn, A., 2013. Understanding Mathematics for Young Children, fourth ed. SAGE, London.

Ifrah, G., 1998. Universal History of Numbers: From Prehistory to the Invention of the Computer. The Harvill Press, London.

Justus, C., 1996. Numeracy and the German upper decades. Journal of Indo-European Studies 24, 45–80.

Kamii, C., 1999. Young Children Reinvent Arithmetic, second ed. Teacher College Press, New York.

Kimura, D., 2000. Sex and Cognition. MIT Press, Cambridge.

MacGregor, N., 2010. A History of the World in 100 Objects. Penguin, New York.

Miller, K.F., Kelly, M., Zhou, X., 2005. Learning mathematics in China and the United States. In: Campbell, J.I.D. (Ed.), Handbook of Mathematical Cognition. Psychology Press, New York, pp. 163–178.

Schwank, I., 1993. On the analysis of cognitive structures in algorithmic thinking. Journal of Mathematical Behaviour 12 (2), 209–231.

Schwank, I., 1999. On predicative versus functional cognitive structures. In: Schwank, I. (Ed.), European Research in Mathematics Education, vol. II. Forschungsinstitut für Mathematikdidaktik, Osnabrück, pp. 84–96.

Schwank, I., 2003. Einführung in funktionales und prädikatives Denken. In: Schwank, I. (Ed.), Zentralblatt für Didaktik der Mathematik, vol. 35(3), pp. 70–78.

Schwank, I., 2013. Einblicke in Mathematische Spielwelten – Mathematisch-logisches Denken Will Erlernt Werden. Deutsch-Chinesischer Reader. Forschungsinstitut für Mathematikdidaktik, Osnabrück. 认识数学的游戏世界-数学的逻辑思维是可以学习的. (Insights into Mathematical Playworlds – Mathematical Logical Thinking Wants to be Learned. German-Chinese Reader.).

Spearman, C., 1904. General intelligence, objectively determined and measured. American Journal of Psychology 15, 201–293.

Starkey, P., Cooper, R.G., 1980. Perception of numbers by human infants. Science 210, 1033–1034.

Starkey, P., Klein, A., 2008. Sociocultural influences on young children's mathematical knowledge. In: Saracho, O., Spodek, B. (Eds.), Contemporary Perspectives on Mathematics in Early Childhood Education. IAP/Information Age, Charlotte, NC, pp. 253–276.

Mathematical Models in Geography

Sergio J Rey, GeoDa Center for Geospatial Analysis and Computation, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores the role of mathematical modeling within human geography. The focus is thematic in nature providing an overview of a highly diverse literature. The approaches to modeling in geography are organized as a typology that classifies models according to genre and applications. Key methodological challenges that distinguish geographic modeling from that practiced in other social science disciplines are discussed. These include issues related to spatial dependence, changes in geographical scale as well as the question of local contextual effects, and their treatment in geographical models. This article ends with some comments about the future prospects for geographical modeling.

Introduction

This article explores the use of mathematical modeling in the field of human geography. The focus is on providing an overview of the definition of modeling in human geography, main research directions, methodological and theoretical issues, and a view of the future prospects of modeling. By necessity, the focus is broad rather than in-depth, and thematic rather than technical.

The article is organized as follows. Definitions of the term model are first briefly reviewed to provide context; next a typology of models in geography describes different types of models and their main uses. Attention is then directed to a number of important challenges in geographic modeling. The article ends with a view of the future for modeling in human geography.

Definitions

The term model has taken on multiple definitions within the field of human geography. Moreover, as the discipline has experienced paradigmatic shifts, the status, roles, and definitions of models have also evolved and varied. As in many social sciences, theories and models within geography are closely connected. Indeed, many social science theories could be described as models as they are often composed of a set of interrelated hypotheses rather than empirically validated laws (Cadwallader, 1985).

A common view of a model within geography is a simplified representation of some aspect of reality. The representation abstracts the central components of interest to allow for an investigation of their role in the process under study, while the aspects less central to the process are omitted from the model. This representation can be articulated in one of the three geographic model forms: (1) iconic, where representation is based on a reduction in scale, such as a classroom globe; (2) analogue, relying on both changes in scale and transformation of the properties of the system, as in a map; and (3) mathematical, in which formal symbols and equations are used to represent the system. The focus here is on this last form of geographical model.

As knowledge is developed through an ongoing interaction between theory and observation, mathematical models play a critical role in this process, serving as an empirical counterpart of a theory which is compared against reality. This conventional view of a geographic model might be characterized as the 'model of a process' perspective. There is a second, somewhat more subtle but equally important, view of the term model in geography, and this is the 'model as a process' perspective. Models are not simply about the world but are also about our knowledge about the world (Haines-Young, 1989). In this way, the process of modeling can be seen to serve three useful purposes. (1) It forces the researcher to be explicit concerning theoretical constructs, hypotheses, and other aspects of their understanding of a system. (2) Formalization of a problem in an empirical model can be of great utility in revealing complex indirect effects that are not articulated by the theory. (3) Models can assist in policy analysis where there is often a gap between existing substantive theory and the real world that policy makers are confronted with.

A Typology of Geographic Models

Given the integrative nature of geography as a social science, a wide variety of models currently exist within the discipline. Human geographers have employed a vast array of models based on different mathematical frameworks to study processes at widely ranging spatial scales for a variety of purposes. Consequently, a cataloging of these models is clearly beyond the scope of this article. Instead a generalized structure is presented in which human geographic models can be categorized according to three dimensions: (1) spatial scale, (2) mathematical framework, and (3) applications.

Spatial Scale

One important dimension of spatial analysis in human geography concerns the spatial scale of the study. Human geographers have developed mathematical models for scales ranging from the spatial decision-making of individuals, to the spatial structure of housing values across neighborhoods in a city, to patterns of uneven development of regions

within a national setting, and to studies of international trade flows.

Microscale

Microscale geographical models are those that represent a spatial system at its elemental or individualistic level. *Agent-based models* (ABMs) represent the behaviors and interactions of individual actors located in space. Different types of agents are included in these models depending on the context. For example, in an ABM exploring residential search processes in an urban setting, the agents might be households stratified by household size and income, who make decisions based on preassigned rules that may reflect underlying microeconomic principles (Ettema, 2011). Through an agent's interactions with other agents in bidding for housing units, complex patterns of market outcomes can emerge.

Closely related to ABMs are *cellular automata* (*CA*) *models* (Clarke et al., 2007). CA models are similar in spirit to ABMs since they are individualistic in their representation; however, the individual units are typically taken to be cells or discrete spatial zones in a CA model rather than human individuals. The focus is typically on how the cells transition between land uses, such as developed or undeveloped, over time given simple rules, for example:

$$S^l_{i,t+1} = \begin{cases} 1 & \text{if} \quad \sum\limits_{j \in \omega_i} S^l_{j,t} \geq \alpha_l \\ 0 & \text{otherwise} \end{cases}$$

where $S^l_{i,t+1}$ is an indicator variable specifying that the land use class for cell i at time period $t + 1$ is set to l if a sufficient number (α_l) of its neighbors (the set ω_i) were in class l in the previous period.

Mesoscale

ABM and CA models can be viewed as the latest in the evolution of models designed to study the human spatial structure of cities. Earlier models were more aggregate in their representation of cities reflecting their reliance on early urban economic and land use theory where representative consumers maximized their utility in deciding where to consume housing subject to a budget constraint (Alonso, 1964). These models were predicated on an assumed central business district (CBD), serving as the point of maximum access for the surrounding urban area, and increasing transportation costs with movement away from the CBD.

While the origins of this monocentric model are found in urban economics, geographers have played an important role in developing extensions of this classic framework to the case multi- or polycentric city models (Odland, 1978). In these models, the city may have several nodes of concentrated economic activity, and the articulation of the role of accessibility in determining land use patterns in such a context results in a highly varied landscape. More recently, geographers have continued the discretization of cities by adopting novel fractal geometric approaches to study urban form and structure (Batty and Longley, 1994).

Macroscale

Closely related to the work on internal urban spatial structure have been efforts to develop models of urban systems and their evolution. An urban system consists of the complete set of urban areas and cities within either a region or nation. A central concern in this literature has been the size distribution of cities within a system (Carroll, 1982). These models consider the population sizes and rankings of a system's cities through the use of the so-called rank–size curves, which take the following general form:

$$\ln(r_i) = \alpha + \beta \ln(p_i)$$

where $\ln(r_i)$ is the natural logarithm of the rank of city i with respect to size and p_i is the city i's population. A large number of empirical studies have found that a value for β close to 1 well describes many developed national city-size distributions. As Fujita et al. (1999) have noted, part of the fascination with the rank–size rule is that it comes as close to an empirical law as any relationship in the social sciences.

The work on city-size distributions has been paralleled by an extensive body of modeling on regional development, growth, and inequality dynamics, where the focus is on variation in standards of living across regions and over time. One branch of this literature has adopted modeling frameworks from international growth theory to the case of convergence between regions within a single country (Rey and Le Gallo, 2009).

Other regional models focus on the economic structure and performance of urban and regional units. Economic base models posit that a city or regional economy can be divided into two components. The first component consists of those activities that produce for markets outside the region, which are thus classified as the 'export base.' The second component, referred to as the local sector, consists of activities that produce for internal markets and provide goods and services for local residents and businesses. The central idea in these models is that local service employment is derived from the expansion of the export sector, which is the basis for economic growth in the region. An extensive literature has evolved around the issues of implementing and applying economic base analysis (Krikelas, 1992).

Regional input–output models can be viewed as an expansion of the economic base construct to consider a more detailed set of interindustry linkages within the urban or regional economy (Dewhurst et al., 1991). Like economic base models, the earliest regional input–output models focused on a single regional economy; however, input–output modelers have also extended these frameworks to examine the nature of interregional interindustry interactions. These efforts have provided rich insights into the nature of modern space economies.

Spatial Scale Issues

A long-standing issue in spatial modeling is the so-called modifiable areal unit problem that arises when inferences about the nature of a phenomenon change with the spatial scale used to study the phenomenon. This has been illustrated in studies of the correlates of voting behavior in which the relationship between age and propensity to vote for a particular party has been demonstrated to vary both in strength and direction as the analysis moved from the individual level, to census block, county, and state.

A related challenge facing geographic modelers surrounds the relationships between processes that operate at different scales. For example, there is a growing need to integrate mesolevel models of urban land use with macromodels of a regional economy so that the interaction between dwindling land supply and aggregate regional economic growth can be analyzed and better understood. On a broader level, there is also a pressing need to analyze the relationships between regionally focused human-induced modification of the physical environment and global climate change.

Mathematical Framework

The second way one can categorize and group different types of models within human geography is to focus on their underlying mathematical framework. Here distinctions can be made based on three dimensions: (1) stochastic or deterministic model, (2) spatial interactions between units taken into account, and (3) framework integrated in that it treats multiple subsystems or is it focused on a single process/system.

Deterministic versus Stochastic Models

A deterministic model is one in which the values for the dependent variables of the system are completely determined by the parameters of the model. In contrast, stochastic, or probabilistic, models introduce randomness in such a way that the outcomes of the model can be viewed as probability distributions rather than unique values.

Deterministic models have the advantage of often being amenable to mathematical analysis, as in the case of using ordinary differential equations to model the geographic spread of infectious disease within human populations. Simulation of these models through parameter and input adjustment can provide important insights into the quantitative and qualitative features of epidemics (Sattenspiel, 2009).

The mathematical clarity offered by deterministic models can, however, come at a cost of decreased realism. Because of this, most of the mathematical modeling done in human geography is stochastic. These models provide a framework to represent spatial uncertainty in mathematical modeling that is well recognized as an important issue in its own right (Goodchild and Zhang, 2002). That uncertainty can surround a number of dimensions of the modeling effort including data uncertainty and model specification uncertainty, parameter uncertainty, among others. Stochastic models can also provide estimates of confidence intervals around point estimates and forecasts.

Global or Local, Multi- or Interspatial

A central concern in human geographic modeling is the interaction across space between the observation units. This reflects Tobler's *First Law of Geography* that "everything is related to everything else, but near things are more related than distant things." The process of spatial interaction is therefore a key mechanism for creating the spatial structures underlying human societies. At the same time, spatial interaction can also be influenced by the underlying spatial structure of the society. Spatial interaction models are used to study a broad spectrum of human spatial behavior that takes the form of movement or communication over space (Fotheringham and O'Kelly, 1989). Some of the most common types of spatial interaction, which have been modeled, include migration patterns, commuting, commodity flows, and retail shopping behavior.

Spatial interaction models illustrate how different locations can become functionally interdependent. This can be seen by considering a simple interaction model:

$$T_{i,j} = \frac{P_i P_j}{d_{i,j}^{\beta}}$$

where $T_{i,j}$, the degree of spatial interaction (e.g., migration, trade flows, air travel) between places i and j is directly proportional to the sizes of the population at the original location P_i and the destination P_j, and inversely proportional to the distance between these locations $d_{i,j}$. The strength of the decline in interaction with increasing separation is captured by β, the distance decay coefficient.

In models such as these, the spatial interactions appear as dependent variables in that predictions between each pair of locations i and j are the outcomes of the model. Spatial interactions can also appear in human geographic models as explanatory variables as well. In the spatial econometrics literature (Anselin, 2000), an example is the spatial lag specification used in the regional convergence literature:

$$y_i = \rho \sum_j W_{i,j} y_j + X_i \beta + \varepsilon_i$$

where y_i is the growth rate for regional economy i, which is related to traditional determinants of growth X_i as well as the weighted average of growth rates in neighboring regions reflected in the term $\sum_j W_{i,j} y_j$. A final issue related to spatial interaction is the proper specification of contextual effects in developing models of socioeconomic systems. Geographers are faced with the twin challenges of building models that capture general systematic properties while also capturing the richness of spatially explicit variation in many social phenomena. At first glance, these two challenges appear to be mutually exclusive as the search for empirical regularities would seem to rule out local variations and nuances. However, an active branch of geographical modeling has focused on the so-called 'local spatial analysis' in which methods are being designed to identify 'hot spots' and deviations from overall general patterns. This opens the possibility of integrating both local and global types of modeling.

Integrated versus Isolated

The final genre in this classification is found in integrated human geographic models. A model can be classified as integrated in a number of ways. *Substantively* integrated models encompass a number of different activities within the same analytical framework, recognizing the interdependencies between, for example, regional labor markets, housing markets, demographic processes, transportation networks, and environmental systems (Briassoulis, 1986).

Spatially integrated models provide a comprehensive treatment of one or more geographical dimensions of a process. For example, in multiregional economic models, the interactions between different regional economies are treated as endogenous to the system. A second important class of spatially integrated models are those that focus on the relationships

between different levels within a spatial hierarchy, such as global, national, regional, and urban (Jin and Wilson, 1993).

The final type, *methodologically* integrated models combine different modules, each with a potentially different mathematical structure within the larger comprehensive framework, so as to achieve modularity and flexibility (Rey, 2000). All types of integrated models are enjoying increasing popularity with the increased interest in interdisciplinary problem solving.

Applications of Models in Human Geography

Broadly, human geography models are used for (1) statistical investigation, (2) optimization, or (3) simulation and forecasting.

Statistical investigation is perhaps the most common, often used method for hypothesis testing. For example, human geographers have used specialized regression models to investigate issues such as urban decentralization, regional income inequality, historical voting patterns, and socioeconomic aspects of deforestation. What distinguishes these types of investigations from those in other social sciences is that the spatial dimensions assume a central focus.

Two broad classes of statistical modeling in geography are recognized. The first is exploratory, where the focus is on the identification of spatial pattern, testing for spatial clustering, and departures from random sampling. Models and methods for exploratory spatial data analysis have attracted a great deal of attention (Getis, 1999).

The second approach to statistical modeling is known as confirmatory spatial data analysis. Here the field of spatial econometrics has played a central role in developing new methods to deal with the treatment of spatial autocorrelation and spatial structure in regression models (Anselin, 2000). Spatial autocorrelation presents a fundamental problem to the application of traditional (i.e., a-spatial) regression methods, given that their application for inferential purposes is based on an assumption of random sampling, which is rendered invalid in the presence of spatial autocorrelation.

Optimization models are used for normative purposes to identify the optimal solution for a given problem. All optimization models have a quantity that is to be maximized (or minimized) as reflected in an objective function, for example, in the case of the well-known transportation model:

$$\text{Minimize } L = \sum_{i=1}^{n} \sum_{j=1}^{n} T_{i,j} Q_{i,j}$$

subject to:

$$D_j \leq \sum_{i}^{n} Q_{i,j} \forall j$$

$$S_i \geq \sum_{j}^{n} Q_{i,j} \forall i$$

$$T_{i,j} Q_{i,j} \geq 0 \forall i, j$$

where the objective is to minimize total transportation costs. The costs are a function of the unit transport cost, $T_{i,j}$ and the number of units shipped between location i and j, $Q_{i,j}$. The solution has to respect a number of constraints such that, the amount shipped to each region satisfies demand in that region D_j, the amount shipped from a region is no greater than the supply in the region S_i, and all costs are nonnegative.

Several types of optimization models have been developed within human geography. *Location-allocation models* (Gosh and Rushton, 1987) typically involve a set of consumers distributed over a given area and a second set of facilities (stores, public services) to serve them. These models are used to allocate the facilities to best serve the customers. Optimization methods developed by geographers have also been used to delineate regions for specific objectives, such as school and political zone redistricting.

Simulation and forecasting models are used to study and compare the responses of a system to policy or environmental changes. Examples of these applications can be found in the areas of integrated environmental human analysis (Knight, 2000) and the analysis of trade policy impacts on regional economies (Warf and Randall, 1994).

Prospects and Challenges

Prospects

Recent developments, within and outside human geography, bode well for the future development of modeling. Within the discipline, modeling is now well established as an important area of research, and its role in legitimizing human geography as a discipline is widely accepted.

Among the social sciences, the importance of space has grown in recognition. The rediscovery of the geographical dimensions of many social problems has given rise to some new research agendas within these disciplines. Scholars in other disciplines are turning to geography and geographers as they address this new terrain. Coupled with the demand for interdisciplinary approaches, the opportunities for geographic modeling to contribute to more sophisticated analysis in social science research are bound to increase. Human geographers active in the development and application of geographic modeling will have an important role to play in strengthening human geography's claim as an integrating discipline within the social sciences.

Challenges

In order for geography to fully realize this claim, two challenges will have to be addressed:

- space–time modeling; and
- computational complexity

Space–Time Modeling

As geographic modeling becomes more widely used in other social sciences, the need to consider temporal as well as spatial dimensions of processes will grow. As argued by Cressie (1993: p. 24), a purely spatial model typically has no causative component but rather reflects the aggregation of short-term effects over a fixed time period for which data are available. Static, or purely cross-sectional, models are those in which time is not explicitly incorporated in the sense that no time subscript

formally appears in the model, only spatial coordinates are used to index the observations. Implicitly, of course, the models are solved for the moment in time in which data used to calibrate the model were obtained. This is the case for models in which cross-sectional observations are obtained for different geographic locations. Underlying such models is the assumption that the system under study has reached some sort of equilibrium, which is a necessary condition to enable parameter estimation.

In dynamic models, time is given an explicit representation. Values for the dependent and explanatory variables are now indexed by spatial as well as temporal coordinates. Different forms of dynamic representation arise from alternative ways of linking observations across space and time. In a comparative statics approach, different inputs are specified for a model at each time step, leading to different equilibria, and a comparison of these different solutions can reveal insights about how the system adapts to changes in inputs.

The second class of dynamic models in human geography adopts a disequilibrium perspective in which the focus is on how shocks to the system result in movements away from equilibria and the time path involved in the evolution of the system. In most human geographic models, time is specified as discrete, using data observed at regular fixed time intervals (e.g., annual, daily, etc.); however, in some instances, continuous time modeling has been adopted to study human travel behavior (Brockmann et al., 2006).

At present, there is a rich and mature body of methods for dealing with time series, and a modest but growing set of methods for spatial series. Methods for dealing with both dimensions simultaneously within the same model face significant challenges due to the disparity in different types of space–time data and the associated substantive problems that can be considered (Goodchild, 2013). A prominent type of space–time data is tracking data, in which a series of point observations in space is recorded at different time intervals to study human movement, for example. A second form of space–time data compares time sequences of remote sensing data to detect land-use change. Still other forms of space–time data underlie the CA and ABMs described earlier, and polygon coverage data consisting of longitudinal observations on areal units such as census tracts over multiple time steps. Distinct differences among these data types have hindered the development of a unified approach to space–time modeling in human geography.

Computational Complexity

Technological changes have shaped some of the methodological issues that modelers face. In the past, discussions concerning data availability focused on the dearth of good data sets to base models on. Today, advances in data collection technologies (i.e., remote sensing, geographic information system, global positioning systems) have shifted the discussion to problems of too much data. Traditional geographic modeling techniques are ill-suited for these enormous data sets and novel new approaches will be required.

These new approaches will reflect the growing importance of the so-called *fourth paradigm of science* (Tolle et al., 2011) in which the use of models in human geography will shift from their traditional roles in theory development and testing, to other more exploratory purposes. Here, the application of exploratory spatial data analysis and data mining to data streams is used to uncover new empirical patterns that may lead to the new types of questions, and ultimately hypothesis generation and theory building. The distinguishing characteristic here is that the science and modeling effort is data instigated rather than driven by a priori theoretical constructs.

Coupled with the data avalanche induced by new geospatial technologies has been the rise of *geocomputation* as an approach within human geography involving the application of high-performance computing technologies to analyze digital representations of human activity in a geospatial frame of reference (Miller, 2008). Together with advances in cyber-infrastructure for geospatial analysis (Wang, 2010), human geographers can now develop models that incorporate parallelism, adaptive systems, and emergence that were simply not possible in earlier generations of computational infrastructure, but that now may provide more realistic representations of human geographic systems.

Finally, while the new geospatial technologies have fueled a resurgence of interest in geographical modeling and analysis, these technologies have also redefined the types of questions researchers are addressing. A promising development in this regard has been the rise of open-source software or shared software development among a research community (Rey, 2009). Prominent examples of open-source projects supporting spatial modeling in human geography are the Python Spatial Analysis Library (Rey and Anselin, 2010) and the R project (Bivand and Neteler, 2000). Beyond facilitating the dissemination of advances in spatial modeling and analysis methods through free software, open source has also had major impacts on pedagogy in training human geographic modelers.

See also: Location Theory; Regional Science; Spatial Autocorrelation; Spatial Data; Spatial Optimization Models; Spatial Sampling.

Bibliography

Alonso, W., 1964. Location and Land Use. Harvard University Press, Cambridge.

Anselin, L., 2000. Spatial econometrics. In: Baltagi, B. (Ed.), Companion to Econometrics. Blackwell, London, 310–330.

Batty, M., Longley, P., 1994. Fractal Cities: A Geometry of Form and Function. Academic Press, New York.

Bivand, R., Neteler, M., 2000. Open source geocomputation: using the R data analysis language integrated with GRASS GIS and PostgreSQL data base systems. In: Proceedings of the 5th International Conference on GeoComputation.

Briassoulis, H., 1986. Integrated economic-environmental-policy modeling at the regional and multiregional levels: methodological characteristics and issues. Growth and Change 17, 22–34.

Brockmann, D., Hufnagel, L., Geisel, T., 2006. The scaling laws of human travel. Nature 439 (7075), 462–465.

Cadwallader, M., 1985. Analytical Urban Geography. Prentice-Hall, Englewood Cliffs.

Carroll, G., 1982. National city size distributions: what do we know after 67 years of research? Progress in Human Geography 6, 1–34.

Clarke, K., Gazulis, N., Dietzel, C., Goldstein, N., 2007. A decade of SLEUTHing: lessons learned from applications of a cellular automaton land use change model. In: Fisher, P. (Ed.), Classics in IJGIS: Twenty Years of the International Journal of Geographical Information Science and Systems. CRC Press, Boca Raton, pp. 413–427.

Cressie, N., 1993. Statistics for Spatial Data. Wiley, New York.

Dewhurst, J., Hewings, G., West, R., 1991. Regional Input-Output Modeling: New Developments and Interpretations. Avebury, Aldershot.

Ettema, D., 2011. A multi-agent model of urban processes: modelling relocation processes and price setting in housing markets. Computers, Environment and Urban Systems 35 (1), 1–11.
Fotheringham, A.S., O'Kelly, M., 1989. Spatial Interaction Models: Formulations and Applications. Kluwer Academic Publishers, Dordrecht.
Fujita, M., Krugman, P., Venables, A.J., 1999. The Spatial Economy: Cities, Regions, and International Trade. MIT, Cambridge.
Getis, A., 1999. Spatial statistics. In: Longley, P., Goodchild, M., Maquire, D., Rhind, D. (Eds.), Geographical Information Systems. Wiley, New York.
Goodchild, M., Zhang, J., 2002. Uncertainty in Geographical Information. Taylor & Francis, London.
Goodchild, M.F., 2013. Prospects for a space–time GIS. Annals of the Association of American Geographers 103 (5), 1072–1077.
Gosh, A., Rushton, G., 1987. Spatial Analysis and Location-Allocation Models. Van Nostrand Reinhold, New York.
Haines-Young, R., 1989. Modelling geographical knowledge. In: Macmillan, B. (Ed.), Remodelling Geography. Basil Blackwell, London, pp. 22–39.
Jin, Y., Wilson, A., 1993. Generation of integrated multispatial input-output models of cities GIMIMoc 1: initial step. Papers in Regional Science 72, 351–368.
Knight, C., 2000. Regional assessment. In: Encyclopedia of Global Change. Oxford University Press, New York.
Krikelas, A.C., 1992. Why regions grow: a review of research on the economic base model. Economic Review, 16–29.
Miller, H.J., 2008. Geocomputation. In: Fotheringham, A.S., Rogerson, P.A. (Eds.), The SAGE Handbook of Spatial Analysis. SAGE Publications, pp. 397–418.
Odland, J., 1978. The conditions for multi-center cities. Economic Geography 54, 234–244.
Rey, S.J., 2000. Integrated regional econometric + input–output modeling: issues and opportunities. Papers in Regional Science 79, 271–292.
Rey, S.J., 2009. Show me the code: spatial analysis and open source. Journal of Geographical Systems 11, 191–207.
Rey, S.J., Anselin, L., 2010. PySAL: a Python library of spatial analytical methods. In: Fischer, M.M., Getis, A. (Eds.), Handbook of Applied Spatial Analysis. Springer, Berlin, pp. 175–193.
Rey, S.J., Le Gallo, J., 2009. Spatial analysis of economic convergence. In: Mills, T.C., Patterson, K. (Eds.), Handbook of Econometrics, Applied Econometrics, vol. II. Palgrave Macmillan, New York, pp. 1251–1290.
Sattenspiel, L., 2009. The Geographic Spread of Infectious Diseases: Models and Applications. Princeton University Press.
Tolle, K.M., Tansley, D., Hey, A.J., 2011. The fourth paradigm: data-intensive scientific discovery [point of view]. Proceedings of the IEEE 99 (8), 1334–1337.
Wang, S., 2010. A cyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis. Annals of the Association of American Geographers 100 (3), 535–557.
Warf, B., Randall, J., 1994. The U.S.–Canada free trade agreement: impacts on U.S. States and Canadian Provinces. International Regional Science Review 17, 99–119.

Mathematical Models in Philosophy of Science

Zoltan Domotor, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents an overview of the most important foundational aspects of mathematical models used in science. The importance of mathematical models is widely recognized for their essential role in the acquisition and organization of scientific knowledge, and their success in representing the structure, mechanism, and behavior of real-world systems is undeniable. However, what are these models, in virtue of what do they represent, what is the basis for choosing one particular model over another, and how are models validated? These are some of the conundrums contemporary philosophy of science tries to solve. Answers to these and other problems are given in the general frameworks of *equationally* vs. *structurally* conceived mathematical models and their symmetry relationships. We illustrate several major types of models, their modifications and structural enrichments on a concrete extended example, known for its significant intuitive content.

On the most fundamental level, the basic questions contemporary philosophy of science asks and tries to answer about mathematical models used in science are as follows:

1. *What are mathematical models*? There are many issues buried in this question, known as the 'ontological puzzle': are models equations, formal systems, set-theoretic or category-theoretic structures, fictional entities, all of the above, or something else (e.g., 'mediators' between theory and the real world, autonomous tools of investigation, reconstructions of or formal stand-ins for so-called target systems)?
2. *In virtue of what do mathematical models represent the chosen aspects of a natural system, phenomenon, mechanism, or process under study – henceforth referred to as the target system – and how do model builders know whether or not a mathematical model correctly represents*? These are very hard semantic and epistemic questions, since the success of mathematical modeling in science is undeniable, but the nature of linkages between models and systems, underlying the models' success, is difficult to pin down. Einstein (1954, page 233) puts his finger on this central problem in his famous remark: "*As far as the propositions of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality*." Applied mathematicians, scientists, and systems engineers have always held the view that most propositions of an adequate mathematical model refer to reality in some way, because without such reference, the presence of mathematics in scientific representations would be unintelligible. A major problem is that a mathematical model-based 'representation' is generally not a variant of copying or mirroring.
3. *In what ways can mathematical models represent*? Mathematical methods form a huge part of scientific and engineering practice. A natural system under study may be represented by several different mathematical models, and conversely, a given mathematical model often represents many different (possibly unrelated) systems, both within and across various scientific disciplines. A large part of contemporary natural and social science is concerned with investigating the temporal and spatial changes in target systems. There are many ways in which models can and do represent such changes. To illustrate the *pluralism* in dynamical model types, recall the choices between deterministic and stochastic mathematical models, continuous-time and discrete-time models, process-based and agent-based models, classical and relativistic models, and so forth. Applications of mathematics in systematic studies of stereotype empirical systems tend to involve *parametrized families*, *hierarchies*, or *networks* of models.
4. *How are models validated, what is the basis for choosing one particular model over another, and when is a model acceptable*? Validation is an empirical procedure intended to determine the degree to which a mathematical model under consideration is an accurate representation of its target system within its domain of intended application. As products of model developers' cognitive work, mathematical models are not independent of human understanding. Nevertheless, their validity rests solely on their ability to predict experimental outcomes to within antecedently specified errors (irrespective of agreement with common sense and intuition) and on the degree of coherence with other accepted models.
5. *In virtue of what do mathematical models provide information about their target systems*? This question concerns the *epistemology* of models. We know that models usually represent their target systems in ways that uphold the basic principles of knowledge acquisition about them. However, according to the philosophers of several different realist persuasions, a mathematical model can serve as a reliable source of information about the target system only if it correctly represents the target system's structure, function, or behavior under consideration, that is, if researchers want to extract information from a mathematical model, they have to come to terms with how the model represents.

Having identified the basic questions about models studied by philosophers of science, we next summarize several corresponding answers. We gather them in one place largely to standardize the terminology. It should come as no surprise to the reader that there are several important questions to which philosophers of science do not have satisfactory answers.

Debates Over Mathematical Models in Philosophy of Science

Recently, philosophers of science have developed a large variety of approaches and solutions to the questions itemized earlier. We begin by reviewing some of them here and set the scene for a more detailed study of mathematical models, their construction, and application.

Mathematical Models

The question 'What are mathematical models?' has two prominent answers, which considerably simplify the inflationary use of the term 'model' in science:

1. In the first instance, there is the syntactic *equational* answer, advocated by logical empiricists in the 1920s and retained by most philosophers of science as the so-called *received view* or *standard conception* until the mid-1960s. The idea is to define mathematical models in terms of *systems* of (most commonly equational) *axioms* of some sort, modulo logical equivalence, stated in a (semi) formal language. According to the standard conception, equations together with chosen constraints and initial conditions are used in deriving numerous predictions that must then be compared with independently obtained evidential data, gained through measurement and experimentation. Repeated prediction–data comparisons form an integral part of model verification and confirmation. If the calculated predictions of a model under study and independently obtained data agree to within a reasonable error, then model users can be said to have a proper grounding for their growing degrees of confidence that the model correctly represents the target system. A model is *confirmed* when its observable predictions are borne out, and only highly confirmed models should be accepted by model users. From an empiricist perspective, a mathematical model is *empirically adequate* provided that *all* of its predictions are borne out. Unfortunately, a model user may never be in a position to *know* whether the model under study is empirically adequate. An inadequate model is usually reformulated by appropriately changing the assumptions under which it was constructed.

   The logical essence of real-world applications of models in the *empiricist received view* sense is succinctly caught in the following inference scheme:

$$\frac{\begin{array}{c}\text{Equations \& Init. Cond.}\\ (\text{Equations \& Init. Cond.}) \rightarrow \text{Prediction}\end{array}}{\therefore \text{Prediction}}$$

   Simply, empirically or contingently true equations combined with initial conditions together with mathematically sound calculations of predictions from these equations and initial conditions yield empirically true predictions that can be compared with experimental data. The model is viewed as inadequate if the agreement between prediction and independently obtained data lies outside of a reasonable estimate of error.

   The basic example of a model in the received view sense is the state equation $P \cdot V = R \cdot T$ of an ideal gas enclosed in a vessel. The real-valued macroscopic quantity P in the equation denotes gas pressure, quantity V stands for the vessel's volume, quantity T refers to the absolute temperature of gas in the vessel, and R is the gas constant, ensuring the equation's dimensional homogeneity. For another example of a model from the perspective of the standard conception consider the now well-known Lotka–Volterra differential equations $\frac{dx}{dt} = ax - xy$ and $\frac{dy}{dt} = xy - by$ of predator–prey interactions. A simple regression model in statistics is defined by a linear matrix equation of the form $Y = \mathbf{A}X + Z$, where Y is a (normally distributed) random variable, $\mathbf{A}$ denotes a coefficient matrix comprising constants, and Z stands for an independent error random variable with normal distribution. In theoretical physics, systems theory, and mathematical economics, the most common examples of mature mathematical models (in the received view sense) are special families of ordinary, partial differential, or stochastic differential equations, involving selected real-valued quantities and parameters, and usually derived from first principles and idealized target system specifications. In some cases, it is possible to solve these equations analytically, and in analytically intractable situations suitable discretization algorithms are necessary to obtain approximate computer-generated solutions.

2. There is an alternative *semantic* or *structuralist* answer to the question about the nature of mathematical models, first proposed by Suppes (1957, last Chapter) and Beth (1961), and further developed by those who followed them. The semantic view treats mathematical models as set-theoretic structures of some sort – typically sets with some kind of additional structure (e.g., relations and operations thereon).

   In the heydays of semantic model theory, Suppes (1988) and others argued that a scientific theory should be specified by the class of those set-theoretic structures in which the (equational) axioms of the theory are true. Accordingly, whenever the classical view specifies a model by a body of nonlogical axioms or other descriptive means, Suppes introduces a corresponding set-theoretical predicate whose extension coincides with the class of those set-theoretic models that satisfy the axioms. The choice of set theory as the best formal tool for representing theories is amply justified by its universal character. On this view, scientific models are nothing but appropriate classes of set-theoretic structures of the same similarity type. In fact, one can proceed a step further and consider *hierarchies* of models, starting from low-level *data* models and closing with high-level *theoretical* models. All this can be extended to networks of models in which the traditional notion of a scientific theory (i.e., a deductively closed system of empirical propositions) disappears. For Suppes, the meaning of the concept of model is the same in both mathematics and the empirical sciences.

   Scientific models tend to involve intricate mixtures of linguistic and mathematical elements. Typically, equations or other stylized syntactic items are used to *describe* the modeler's set-theoretic structure, intended to represent the target system in some idealized or approximate manner.

According to the structural empiricists, a workable conception of *validation of models of target systems* relies on (1) constructing *data models*, i.e., mathematical structures (matrices, histograms, trajectories, signal curves, etc.) that encode raw data from target system measurements in an idealized and rectified manner and (2) structure-preserving embeddings of these data models into the system's model. Accordingly, a model *correctly represents* its target system provided that all available data models are faithfully embeddable into it. We summarize the *structural empiricist* idea of data-based model validation by the following block diagram:

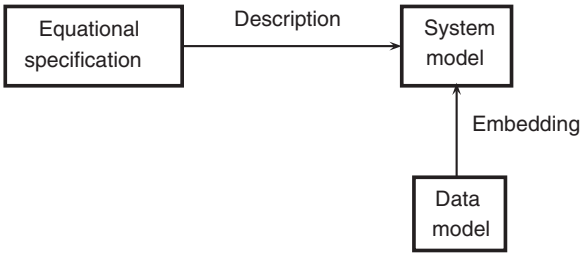

The great classical example of a set-theoretical predicate is given by the class of Kolmogorov's probability models. A set-theoretic triple $\langle \Omega, \mathscr{B}, \mathbf{P} \rangle$ is said to be a *probability model* provided that (1) Ω is a nonempty set, called the *sample space*; (2) $\mathscr{B}$ is a Boolean sigma algebra of subsets of Ω, interpreted as the algebra of events; and (3) $\mathbf{P}$ is a countably additive probability measure on $\mathscr{B}$. Here the customary domain of application of a theory is taken care of by the choice of a concrete probability model $\langle \Omega, \mathscr{B}, \mathbf{P} \rangle$. Differential manifolds of relativity theory, directed graphs constructed for kinship relations, and normal-form game-theoretic structures provide additional examples of structuralist-type models.

Models and Representations

In response to the question 'In virtue of what do mathematical models represent?' an often-discussed metaphysical answer is that mathematical models *represent* in virtue of nature itself being essentially mathematical and logical. This goes well with the contention that causal models represent causal relationships realized by the target system in a one-to-one fashion. The semantic view, held (among others) by Van Fraassen (1994), offers the following explicit *structuralist* account of representation:

> A mathematical model $\mathfrak{M}$ represents the target system S provided that $\mathfrak{M}$ is structurally isomorphic to S and $\mathfrak{M}$ is actually intended by the model builder to represent S.

A model $\mathfrak{M}$ is said to be *isomorphic* to the object of modeling (target system) S if the following conditions hold: (1) each basic constituent of S stands in a one-to-one relationship with a uniquely associated element in $\mathfrak{M}$ and of course conversely, (2) to every quantitative property of the target system S there corresponds a unique real number-valued function ear-marked by the model, and (3) each quantitative property relationship in S has a unique counterpart relationship in the model $\mathfrak{M}$.

A *structural realist* view of the model-target system relationship is outlined in the following block diagram:

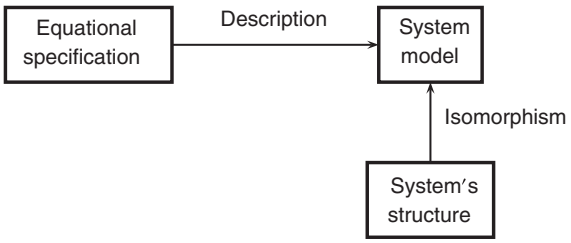

For fairly obvious reasons (stemming from various idealizations, abstractions, and approximations), mathematical models constructed and used in science and engineering seldom meet the unreasonably strong structuralist definition of representation; the isomorphism requirement is usually relaxed in several ways by passing to a partial isomorphism or homomorphism of some sort.

In contemporary philosophy of science, structuralism is a leading approach to models and modeling, and it has several competing variants. The so-called *ontic* structuralists hold the view that mind-independent structure is all there is 'out there.' In more detail, according to French and Ladyman (2003), natural systems, phenomena, and processes should be conceptualized in purely structural terms. Any nonstructural understanding of the nature of empirical objects is metaphysical and unwarranted by the physical science itself. Since according to ontic structuralists systems and phenomena do *possess* structure but they themselves are *not* structure, models can be meaningfully thought of as representing their target systems via partial isomorphisms.

Another approach, called *epistemic* structuralism, defends the view that there is more to the world than its structure (e.g., objects, particles, events, and minds with natures and essences), but only structures are *knowable*. Structural *empiricism* argues that the structure of data and observables can be known, but denies that science should aim at knowing more. In brief, a target system's mathematical model is correct provided that the system's data structures are faithfully embeddable into the model. Thus, the model of a target system is *empirically adequate* if it correctly captures the structure of all system-generated data, made accessible via embedding maps.

Giere (1988, Chap. 3) and his followers argue that none of the above-mentioned mappings between models and target systems can capture the notion of representation. In their view, a model represents a target system just in case it is *similar* to the system. This substantially weakened requirement allows models to be only approximately the same as their target system. Unfortunately, a similarity relation-based account of representation does not fare any better than the structuralist isomorphism requirement, since in real-world modeling situations the notion of similarity applies to practically anything, until the modeler provides an explicit list of all its relevant aspects and degrees – an unworkable task in practice. More importantly, mathematical models tend to possess extra structure that their target systems do not have (e.g., the assumption of infinitely many agents in economic models or infinitely many gas atoms in

thermodynamic models), and conversely, for reasons of tractability, not all structures instantiated by a natural system are included in its model. This issue comes up also in the context of isomorphism, discussed earlier. Most philosophers share the view that the problem of representation requires further study.

Pluralism in Model Construction

Aside from technical tractability and approximation issues, there is no limit to the ways in which mathematical models may be constructed. Excluding the cases of unmanageable complexity and uninformative simplicity, models may employ any mathematical structure whatsoever, including deterministic, probabilistic, and stochastic variants and involving discrete, computable, measurable, continuous, or smooth real- or complex-valued quantities.

Conventional *top-down* model building (e.g., in physics and systems engineering) is characteristically based on first principles of an established scientific theory and constraints provided by the target system's idealized specification. Often from the simple 'initial' models more complex ones can be developed by gradually removing certain idealizations and abstractions. However, small increments in de-idealization or modeling realism are often paid for by significant increases of hard-to-manage mathematical complexity.

Since in biology, economics, and the social sciences in general, there is limited access to universal laws and empirical regularities, the styles of model construction vary widely. In particular, *bottom-up* or *data-driven* (e.g., time series generated) models tend to represent a relatively narrow spectrum of phenomena, while broadly intended models often suffer from low degrees of realism. In these contexts, the point of using a model is mostly to reveal its implications, explore its limits, and treat it as a guide in the development of epistemically more powerful models.

Validation of Mathematical Models

Although some mathematical models are built for general scientific research purposes (e.g., fictional, simulation, and 'toy' models), applications demand that models be validated by comparisons of their predictions with measurement data and results of other (valid) models. Model validation is a process of determining the degree to which a model under study is an accurate representation of its target system from the perspectives of its intended use and domain of application.

There are several problems that occupy the center stage in philosophical discussions of validation. First, since in view of idealizations a mathematical model is rarely correct in all subregions of its theoretically stipulated domain of application, in general, mathematical models cannot be proved valid. Second, in the case of complex models it would be too costly and time-consuming to determine whether or not the model is valid over its entire intended domain of application. Verifications are usually conducted until there is sufficient confidence that the model works well within its envisioned domain of application. Third, determining that the model is adequate in the context of several empirical tests does not guarantee that it is valid everywhere in its applicable domain. Model users' confidence that the model correctly represents the system is available only in degrees of probability. Unthoughtful model acceptance may to lead to two kinds of errors: (1) rejecting a good model, prompted by validation based on incorrect or biased data, or (2) accepting a bad model, endorsed by incomplete validation. Fourth, validation is often limited by the available class of observational data that are far from sufficient to showing that the model is correct in its entire range of application. Last, it is not true that good predictions can only be obtained from a model that is causally sound. Even near-perfect prediction–observation agreements over a wide range of observations may turn out to be a mere coincidence. Conversely, even good models can make bad predictions, as illustrated by Batitsky and Domotor (2007) in discussing models that are intended to represent the *chaotic* behaviors of simple target systems. In addition to robustness and sensitivity analyses and establishing as many connections between the model under study and its target system as possible, validation should also investigate the compatibility relations between the model and the target system's other accepted models.

In applications, it is far from sufficient to know that a dynamical model is *statistically* valid. The reason is that a statistically valid theoretical model may turn out to be *dynamically* (or qualitatively) invalid, meaning that it may fail to represent correctly the dynamical invariants of the target system (including equilibrium states, periodic, aperiodic, chaotic, or strange attractors, and their basins) under various choices of parameter values. Note that even if a dynamical model provides highly accurate predictions in its tested regions of intended use, it does not necessarily follow that the model is dynamically valid, since it may include spurious dynamics in untested regions that are also of interest. The target system's spectrum of actual behaviors remains largely unknown from the perspective of its representing model. Further observation and theoretical research may be needed to obtain a better knowledge about the dynamical invariants and characteristics of the system. Additional problems arise when the model fails to be valid in most parts of its domain of application. Deficiencies in the model may be traced to at least three starting conditions: (1) incorrectly derived equations (e.g., exclusion of decisive higher degrees of nonlinearity), (2) a coarse-grained choice of time and/or space discretization parameters that misses their pathological sensitivity to infinitesimal changes, and (3) reliance on data samples that are contaminated with gross errors.

Learning from Models

Models can serve as reliable vehicles of learning about target systems only if they correctly represent these systems. Thus, model-based knowledge acquisition about systems is intimately tied to the quality of representation. What is of particular interest here to philosophers of science is how a transfer of knowledge from a correctly representing model to its target system is achieved.

We have now surveyed several answers to the philosophers' basic questions about models. After these preliminaries, we now turn to a more detailed study of different types of mathematical models, furnished with a wide variety of algebraic and geometric structures. The underlying objective is

a characterization of the fundamental *duality* between equational and structuralist conceptions of mathematical models that most modern philosophers of science failed to recognize, and instead they continue to defend the idea that mathematical models are basically set-theoretic or category-theoretic structures. Unfortunately, the idea that mathematical models in science are structures that exist independently of and are unrelated to equational characterizations does not sit well with several contemporary mathematical results.

Symmetry Relationship Between Equational and Structuralist Conceptions of Mathematical Models

In the first part of the previous section, we drew the reader's attention to two opposing views about the nature of mathematical models. There is, however, a different approach to models that is even more important. In this alternative approach, the main idea is *unification* via *symmetry* or *duality*, based on a less familiar one-to-one correspondence between the classical (algebraic differential) equation-based and (geometric) structure-based conceptions of models. This unified view of models is attractive, because it is respectful of the algebraic and geometric aspects of models in a complementary fashion. Although this type of algebraic–geometric correspondence can be formalized in an abstract category-theoretic setting, we prefer a concrete illustration furnished by various pendulum models. We choose pendulum models as our running example because (1) most readers are intuitively familiar with the physical behavior of pendulums, (2) they are of historic and basic importance to philosophers of science, (3) it is possible to extend a pendulum model's domain of application by gradual de-idealization and structural enrichment, and (4) pendulum models are available for various physical modifications of classical pendulums, including coupled and double pendulums, extensible and forced pendulums, pendulums in a viscous fluid, relativistic and quantum pendulums, and so forth.

Recall that a simple undamped planar pendulum is classically modeled by the autonomous, deterministic, second-order nonlinear differential equation

$$\frac{d^2\theta}{dt^2} + \frac{g}{\ell}\sin\theta = 0$$

in which the time-dependent *unknown* quantity θ represents the pendulum's continuously varying angular displacement, measured from its downward vertical to the *arm* pivoting without friction on a shaft, to which a *bob* of tiny size with mass m is attached at its swinging end. Constant g denotes the homogeneous local gravitational acceleration acting on the pendulum's bob downward and parameter ℓ captures the length of the pendulum's idealized weightless and perfectly rigid arm. These coefficients are needed for individuating the target pendulum (modulo duplicate relationships) in its *classical* gravitational ambience. The foregoing nonlinear pendulum equation is derived from Newton's second law of motion and the pendulum's accompanying *idealizing scenario*, involving significant idealizations of friction, torque, resistance, and elasticity. We mention in passing that in addition to the pendulum, the nonlinear equation above also represents the motions of many other physical systems. However, the accompanying physical interpretations are different.

The first term in the equation above encodes inertia, and the second term stands for the impact of gravity. Recall that the pendulum's arm is suspended from a pivot point around which it oscillates or rotates in a vertical plane without surface resistance and forcing, so that there are no extra *additive* terms in the equation for these causes. Recall that generally a model is a simplified mathematical representation of a real-world system of interest for designated scientific purposes. Although the target system has many important features, not all of them can and should be included in the model for reasons of tractability and limited epistemic import. Those features that are included often involve drastic idealizations known to be empirically false.

Solutions to the foregoing nonlinear pendulum equation are continuously twice differentiable, time-dependent, real-valued functions of the form $\theta : \mathbb{T} \to \mathbb{S}$ (intended to capture the pendulum bob's time-varying angular positions in all regimes), where $\mathbb{T}$ denotes a designated continuum-time domain, usually isomorphic to the additive group of reals, and the less familiar notation $\mathbb{S} =_{df} \mathbb{R}_{/2\pi\mathbb{Z}} = \mathbb{R} \bmod 2\pi$ is reserved for the circle group of reals, whose underlying set is isomorphic to the half-open interval $[0,2\pi]$. In this way, a given solution θ assigns to each time instant t a unique angle $\theta(t) = \xi$ in $\mathbb{S}$, interpreted as the bob's angular position at t. Technically, all exact (nonelementary) solutions of the pendulum equation are specified analytically by Jacobian elliptic sine functions, and the period is expressed in terms of the so-called complete elliptic integral of the first kind. Therefore, for any given initial condition the corresponding solution can only be obtained numerically, albeit with arbitrary accuracy, using rapid computer calculations. An alternative approach is a passage to *discrete-time difference* equations of the form

$$\theta_{n+1} - 2\theta_n + \theta_{n-1} + \frac{g}{l}(\Delta t)^2 \sin\theta_n = 0$$

where Δt is the length of the time interval, serving as the elementary step of the difference scheme, and for any natural number n, value $\theta_n =_{df} \theta(n\Delta t)$ refers to the bob's approximate angular position at time $n\Delta t$. To improve the accuracy of calculations, modelers should choose very small values of Δt.

To link the *classical* pendulum equation and its solutions to a corresponding *structural* description of the pendulum's dynamical behavior, we must reduce the original equation to a pair of first-order autonomous differential equations of the form

$$\frac{d\theta}{dt} = v \qquad [1]$$

$$\frac{dv}{dt} = -\frac{g}{\ell}\sin\theta \qquad [2]$$

Evidently, their solutions come in pairs of functions that are conveniently combined by set-theoretic pairing into a single solution of the form $\theta; v : \mathbb{T} \to \mathbb{S} \times \mathbb{R}$ that assigns to each instant of time t a unique angular position-angular velocity pair $[\theta;v](t) = (\theta(t), v(t))$, interpreted as the pendulum's *state* of motion at t. We now have the mathematical

means at hand to encode the pendulum's physical behavior *geometrically*. In particular, the pendulum's two-dimensional cylinder-shaped topological *state space* $\mathbb{S} \times \mathbb{R}$ (generated by the circle $\mathbb{S}$) and certain smooth curves therein, corresponding to state-parametrized solutions and forming the target system's so-called *phase portrait*, provide a complete *geometric* as opposed to an *algebraic* understanding of the pendulum's possible behaviors.

We turn now from the solutions of the foregoing equations to a uniquely associated continuously differentiable *dynamical transition map* of the form

$$\mathsf{T} : \mathbb{T} \times (\mathbb{S} \times \mathbb{R}) \rightarrow \mathbb{S} \times \mathbb{R}$$

defined by $\mathsf{T}_t(\xi, v) =_{df} \langle \theta_{\xi,v}(t), v_{\xi,v}(t) \rangle$ for all time instants t, angles ξ, and velocities v. We may regard $\mathsf{T}_t(\xi, v)$ as a formal encoding of the pendulum's 'final' state into which the 'initial' state (ξ, v) evolves after time t. Here it is important to notice that, although there is a plentiful supply of solutions to the pendulum equations, they have to be individuated by states. For example, in view of the equality $\langle \theta_{\xi,v}(0), v_{\xi,v}(0) \rangle = (\xi, v)$, the state (ξ, v) singles out the solution $\theta_{\xi,v}; v_{\xi,v}$ (used in the definition above) and therefore the *trajectory* or *orbit* $\{\mathsf{T}_t(\xi, v) | t \in \mathbb{T}\}$ of (ξ, v) in $\mathbb{S} \times \mathbb{R}$. Trajectories densely fill the entire cylindrical state space and serve as prominent geometric tools in the study of pendulum motions.

It is easy to verify that the induced time-indexed *dynamical maps*

$$\mathsf{T}_t : \mathbb{S} \times \mathbb{R} \rightarrow \mathbb{S} \times \mathbb{R}$$

satisfy the following functional equations for all time instants t and t' in $\mathbb{T}$:

1. *Identity property*: $\mathsf{T}_0 = \mathsf{I}$ (identity transformation) and
2. *Group property*: $\mathsf{T}_{t+t'} = \mathsf{T}_t \cdot \mathsf{T}_{t'}$.

Dynamical maps completely determine the time evolution of the pendulum's motion in the following way: if we know that the pendulum is in state (ξ, v) at time $t = 0$, then it is in state $\mathsf{T}_t(\xi, v)$ at time t. Needless to add, in general, knowledge of T_t is only approximate. As the reader may have already realized, we now have a canonical structuralist *dynamical model* $\langle \mathbb{S} \times \mathbb{R}, \mathsf{T} \rangle$ of the classical pendulum – a semantic counterpart or dual of the pendulum equations. In general, dynamical models are given by an underlying (topological, measurable, etc.) state space and a time-indexed family of transformations (dynamical maps) thereon, satisfying the functional equations.

An impressively large class of deterministic dynamical models arises from autonomous systems of ordinary or partial differential equations by simple state-parametrizations of their solution spaces. Conversely, since $\mathsf{T}_t(\xi, v)$ is a state-parametrized solution of the pendulum equations, it automatically determines these equations. In general, the (vector) differential equation

$$\frac{\mathrm{d}\mathsf{T}_t}{\mathrm{d}t} = \mathbf{A}\mathsf{T}_t$$

corresponding to a *dynamical model* $\langle \mathbb{S} \times \mathbb{R}, \mathsf{T} \rangle$ is reconstructed from the analytic properties of T_t and those of a suitable operator **A**. For additional details, see Engel, Nagel (2000), and the references therein.

Structuralists argue that the particular system of classical equations that generates the dynamical transition map is of secondary concern. In a dynamical model, comprising a state space and dynamical maps, each state space point is an abstract, information-bearing encoding of the target system's physical mode of being at a given time. Assuming that the system under consideration has a state structure, its adequate state space model comes with a state space comprising points that are presumed to contain *complete information* about the past history of the system, relevant to its future behavior. Thus, if the target system's instantaneous physical state – encoded by a point in its representing state space – is known, then the system's subsequent temporal evolution can be predicted with the help of the model's dynamical maps, without any additional knowledge of what has happened to the system. The dynamical map need not be continuous or smooth. It can also be measurable, computable, discrete, and local (i.e., only partially defined). The foregoing structuralist definition of a deterministic dynamical model fits well into the conceptual framework of category theory that provides a formal *adjointness* account of interplay between dynamical models and received-view equations.

An obvious question now arises: what is duality in the case of probability models? Thanks to the so-called Riesz and spectral representation theorems, the question is not hard to answer. Each probability model $\langle \Omega, \mathscr{B}, \mathbf{P} \rangle$ comes with its von Neumann algebra $\mathfrak{L}_\infty(\Omega, \mathscr{B}, \mathbf{P})$ of bounded real-valued random variables of the form $X : \Omega \rightarrow \mathbb{R}$, modulo **P**-measure zero – meaning that random variables X and X' are viewed as equal whenever the equality $\mathbf{P}(\{\omega | X(\omega) \neq X'(\omega)\}) = 0$ holds. And every abstractly given commutative von Neumann algebra is isomorphic to an algebra of this type. Thus, as common in statistics, one can start with a probability model and then switch to the induced algebra of random variables, their expectations, and equations between them or proceed in the reverse direction.

We digress briefly to consider data models. Since experimental data and numerical analysts' calculations tend to come only with limited degrees accuracy (due to measurement errors and round-off operations), it is necessary to consider selected *finite* configurations of states of a parent dynamical model that are conveniently viewed as the underlying *discrete state spaces* of data models. In the spirit of state space discretization algorithms, a region of interest (e.g., a basin of attraction of the model's phase portrait) in the model's state space is subdivided into cells of appropriate mesh size. Each cell comes with a designated representative state, usually calculated by suitable averaging. Dynamical maps on discrete state spaces are then determined by temporal transitions between cell representatives. Discretized dynamical models associated with state space regions provide the formal meeting ground for measurement results and calculated predictions. From a structural empiricist point of view, data models are finitary variants of discretized dynamical models, intended to validate the parent model. Large families of compatible data models can be used to 'cover' the parent dynamical model's domain of

intended application, as illustrated in the following simplified block diagram:

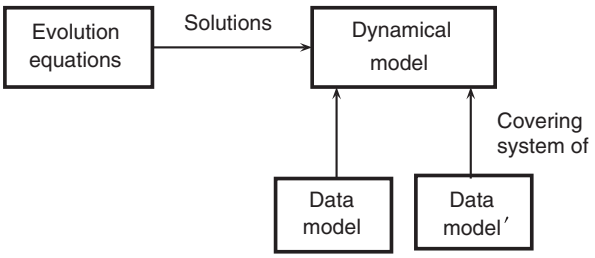

If the *covering system* of data models obtained from various measurements locally agrees with the parent dynamical model's domain of application, that is to say, each data model in the covering system is embeddable into the parent model to within antecedently specified errors, then the parent model is said to be *empirically adequate* relative to the cover. Anyone who is familiar with the chaotic behavior of certain systems quickly realizes that covering chaotic regimes by deterministic data models is ineffective and a passage to carefully designed statistical data models (e.g., built from histograms) becomes necessary.

At this stage, we widen the structuralist point of view somewhat. Models in contemporary mathematical physics are based on the notion of *observable*, intended to encode a target system's quantitative property of interest. In simple situations, model builders assume that (1) observables are real-valued functions, possessing values that can be approximately ascertained by reading an appropriately chosen measuring instrument and (2) their algebraic nature is best expressed by a suitable quantity algebra. Thus, in our running example instead of describing the pendulum's possible motions geometrically in terms of state space trajectories, we turn to a more effective quantitative description of pendulum dynamics, using an associated algebra of observables. The all-important observable quantity is the pendulum's total energy (Hamiltonian) function $\mathsf{E} : \mathbb{S} \times \mathbb{R} \to \mathbb{R}$, defined by the sum of kinetic and potential energy functions

$$\mathsf{E}(\xi, v) \underset{df}{=} \frac{m\ell^2}{2} v^2 + mg\ell(1 - \cos\xi)$$

at each pendulum state (ξ, v). The total energy function is truly significant because (1) the pendulum equation can be reconstructed from it by Hamiltonian differentiation and (2) each curve of the phase portrait corresponds to a different energy value and thus to a different initial condition. Furthermore, it is well-known that contingently on the possible values E of the energy function E with respect to its *critical* level $2mg\ell$ (required for the bob to reach the top point on the vertical), there are three major types of pendulum motion that the experimenter can observe:

1. If $E < 2mg\ell$, then there is insufficient energy available and the pendulum's bob is performing simple oscillations around the minimum. Larger energy values E specify correspondingly larger closed (quasi) elliptical orbits in the pendulum's state space.
2. If $E > 2\,mg\ell$, then the bob makes complete unidirectional rotations ad infinitum. These large energy values individuate open orbits.
3. If $E = 2\,mg\ell$, then the pendulum's bob has exactly the energy needed to reach the top vertical position, and it will remain in this unstable position until the slightest perturbation sends it into one of the two trajectories intersecting at this point. The associated 'border' trajectory $\{(\xi, v | \mathsf{E}(\xi, v)) = 2mgl\}$ is called a *separatrix*, since it separates the two previous regimes (oscillations and rotations) and accordingly divides the state space into disjoint regions of closed and open orbits. From a physical standpoint, separatrix is the trajectory in $\mathbb{S} \times \mathbb{R}$ that encodes the bob's motion with exactly enough energy for it to reach the top unstable (hyperbolic) fixed point, taking an infinite amount of time to get there.

These kinds of considerations lead to the conclusion that equations in a differential paradigm are necessary for formal derivations and calculations, while their semantic counterparts are crucial in geometric interpretations and embodiment of physical content.

Before leaving this section, we make a few technical remarks about observables and their dynamical models. It is easy to see that the bob's velocity $\boldsymbol{v} : \mathbb{S} \times \mathbb{R} \to \mathbb{R}$ is also an observable quantity, if viewed as the second projection map, defined by $\boldsymbol{v}(\xi, v) = v$. A very similar approach applies also to angular displacement $\boldsymbol{\theta}$. Since these and other observables can be added and multiplied, they automatically form an algebra. Let $\mathbf{C}(\mathbb{S} \times \mathbb{R})$ be the algebra of all continuous observables on the cylindric topological state space $\mathbb{S} \times \mathbb{R}$. Conceptually, the topological structure of the state space is now traded for an algebraic structure in a unified manner. Since angular position, angular velocity, and other measurements are seldom perfectly accurate, we need a second structural component that handles all uncertainties associated with information acquisition about quantities. Let $\mathbf{S}(\mathbb{S} \times \mathbb{R})$ be the convex space of all normalized positive linear functionals of the form $\varepsilon : \mathbf{C}(\mathbb{S} \times \mathbb{R}) \to \mathbb{R}$. Since for any quantity Q the value $\varepsilon(\mathsf{Q})$ of ε is the *expectation* of Q, it makes good sense to interpret the abstract functional ε as the target system's *statistical state*. All this is based on the so-called spectral theorem that states that each statistical state has a unique expectation integral representation relative to a probability measure on some measurable space. In sum, quantity algebras and associated convex spaces of statistical states allow model builders to construct mathematical models in an admittedly abstract albeit highly unified manner. More details on applications of quantity algebras to measurement theory may be found in Domotor and Batitsky (2009).

In particular, each structuralist *topological state space dynamical model* $\langle \mathbb{S} \times \mathbb{R}, \mathsf{T} \rangle$ induces an *algebraic dynamical model* $\langle \mathbf{C}(\mathbb{S} \times \mathbb{R}), \mathsf{T}^* \rangle$ of observables, defined by the operator composition $\mathsf{T}^*_t(\mathsf{Q}) \underset{df}{=} \mathsf{Q} \circ \mathsf{T}_t$ for all quantities Q and time instants t. Concretely, in view of $\mathsf{E}_t(\xi_0, v_0) = \mathsf{E}(\xi_t, v_t)$, where $(\xi_t, v_t) = \mathsf{T}_t(\xi_0, v_0)$, the energy quantity can now be considered as time dependent and evolving. In the same spirit, we can also consider a statistical dynamical model induced on statistical states $\mathbf{S}(\mathbb{S} \times \mathbb{R})$.

The passage from an underlying state space dynamical model to the algebraic dynamical model of quantities becomes necessary when the modeler can no longer ignore the ways in which measurements are performed on the systems under study. When an observable is measured, its values are obtained only with a certain probability. Unfortunately, the structuralists' state space dynamical model represents the target system's behavior in itself, without providing the necessary conceptual ingredients for physical measurement and accompanying statistical properties. Discretized algebraic dynamical models of quantities and associated algebraic data models are based on algebraic dynamical models induced by *discrete* state space dynamical models. Topological state space dynamical models suffer from a fundamental inadequacy in the case of *chaotic* target systems, since they focus only on the deterministic side of behavior and the statistical aspects are ignored. However, in these complex scenarios only statistical methods are effective. We shall not give any further details. The technical aspects of these scenarios are studied by Lasota and Mackey (1994).

Structural Enrichments, Perturbations, and Other Modifications of Mathematical Models

In applications of mathematics to real-world problems, there is a frequent requirement of significant flexibility in mutating models by various combinations, de-idealizations, and structural enrichments. As an illustrative background, we shall continue to use the running example of pendulums.

We start with two examples involving a combination and mutation of nonlinear pendulum models. First, consider a double pendulum. It consists of one pendulum attached to another, so that it has two arms and two balls, where the second arm is attached to the first arm's ball. In the case of large displacements, a double pendulum is famously chaotic, but for small oscillations it behaves in simple predictable ways. Double pendulum equations are derived from Newton's second law of motion and simple background conditions, dictated by the idealizing scenario. Structuralists represent the double pendulum by the product dynamical model $\langle \mathbb{S}^2 \times \mathbb{R}^2, \mathsf{T} \otimes \mathsf{T}' \rangle$ with state space $\mathbb{S}^2 \times \mathbb{R}^2 \cong (\mathbb{S} \times \mathbb{R}) \times (\mathbb{S} \times \mathbb{R})$ and a product dynamical map $\mathsf{T} \otimes \mathsf{T}'$ obtained from the constituent dynamical maps of each pendulum and the coupling relation, responsible for chaotic regimes.

Our other example addresses de-idealization cases of pendulum damping and forcing by various mechanisms. Elementary treatments usually assume that the damping torque is directly proportional to the angular velocity. This so-called *linear damping* is captured quite simply by adding an extra term to the original nonlinear pendulum equation, so that now we have

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \alpha\frac{\mathrm{d}\theta}{\mathrm{d}t} + \frac{g}{\ell}\sin\theta = 0$$

where the damping coefficient α in the linear damping term is assumed to be positive. In order to achieve a better representation of friction, quadratic and other terms may be added to the initial pendulum equation, making the computation of approximate solutions increasingly harder. Pendulums driven by constant or periodic external forces and their combinations have mathematical significance in adding extra complexity to model construction. The question now arises as to whether or not less idealized, and hence more complete and more accurate models are decisively better to work with than simpler, less complicated models. As we have already indicated, model accuracy and overall adequacy is not achievable by a simple increase in structural complexity. The reasons are buried in inaccuracies in background conditions, limited knowledge of too many parameter values, and escalating round-off errors in calculating predictions based on complex models. Since complex models tend to lead to resource-bounded and hence significantly less accurate results than those obtained from simpler models, their complexity must be commensurate with background information available in their applications.

In many applications, the same basic pendulum differential equation drops out of a wholly different idealizing scenario of, say, a coupled mass-spring system, consisting of a (point) mass attached to a spring at one end, where the other end of the spring is tied to a fixed frame. Other prominent examples of systems, whose dynamical behavior is also modeled by 'pendulum equations', include electric circuits, chemical reactions, and interacting biological populations with oscillatory behaviors. Of course, in each application, the indeterminate and individuating parameters are interpreted differently.

A whole new field of mathematical model construction is opened up with the introduction of relativistic and quantum aspects of a target system. In the special case of a simple pendulum, we already know that the solution space of the nonlinear pendulum equation includes many relativistically forbidden solutions that encode physically impossible behaviors – individuated by the pendulum's *superluminal* angular velocities. The key idea here is that in the absence of complete knowledge of the model's empirical *domain of applicability*, a researcher may be in an epistemic danger of trying to physically interpret (in terms of behavior) some of the solutions – countenanced by the model, which are actually meaningless in the physical world.

To remedy the situation, the foregoing classical simple pendulum model must be extended to the *relativistic* case (studied, e.g., by Erkal (2000)), having the form

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \left(1 - \left(\frac{\ell}{c}\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2\right)\frac{g}{\ell}\sin\theta = 0$$

where parameter c denotes the speed of light. Because formal models usually possess limited empirical domains of applicability, in comprehensive treatments of target system behavior several different, closely related models may become necessary.

Real-world systems tend to have many representing models, and these models often possess a surplus content, which supports their mutations and extensions in unexpected ways. For example, in the presence of randomness or 'noise', a classical *stochastic pendulum equation* provides the most appropriate representation of randomly perturbed motion, observed (say) on a pendulum attached to a vibrating ceiling. A typical model of the behavior of simple pendulums affected by 'noise' is

presented by the following pair of stochastic differential equations:

$$\frac{d\Theta}{dt} = \Upsilon \qquad [3]$$

$$\frac{d\Upsilon}{dt} = -\frac{g}{\ell}\sin\Theta + \epsilon\Upsilon\frac{dW}{dt} \qquad [4]$$

where Θ denotes the 'unknown' stochastic angular position, i.e., a time-dependent angle-valued random variable; Υ symbolizes the 'unknown' stochastic angular velocity, i.e., a time-dependent real-valued random variable; the noise parameter $\epsilon > 0$ is intended to capture the strength of random noise; and W is a stochastic (e.g., Wiener) process, encoding the pendulum's random erratic motion on a microscale. In this model, realizations of trajectories, representing the pendulum's motion in a sample, are jagged in proportion to the value of noise parameter ε. Putting the earlier described constructions together we envision the construction of stochastic dynamical models for double pendulums, relativistic pendulums, and so forth.

We can also consider the case in which the stochastic component has a parametric perturbation form, captured by the stochastic differential equation

$$\frac{d^2\Theta}{dt^2} + \left(\frac{g}{\ell} + \epsilon W\right)\sin\Theta = 0$$

in which the 'noise' affects the period of oscillation.

Importantly, although in recent years efforts have been made to come to grips with the conceptual side of structuralist dynamical models, practically all stochastic models are presented in an equational form.

In conclusion, in the absence of other options it is too early to tell which of the two conceptions (i.e., equational vs. structural) of mathematical models will prove more important in the future development of models in science. However, it is clear already that the duality between equations and solution-induced structures has significantly deepened our understanding of the nature of mathematical models, their construction, and validation.

See also: Axiomatic Theories; Logical Positivism and Logical Empiricism; Mathematical Psychology; Measurement Theory: Conjoint Analysis Applications; Structuralism.

Bibliography

Batitsky, V., Domotor, Z., 2007. When good theories make bad predictions. Synthese 157, 79–103.

Beth, E., 1961. Semantics of physical theories. In: Freudenthal, H. (Ed.), The Concept and the Role of the Model in Mathematics and Natural and Social Sciences. Reidel, Dordrecht.

Domotor, Z., Batitsky, V., 2009. Measurement, information channels, and discretization: exploring the links. Measurement Science Review 9, 134–161.

Einstein, A., 1954. Ideas and Opinions. Crown Publishers, New York.

Engel, K.-J., Nagel, R., 2000. One-Parameter Semigroups for Linear Evolution Equations: Graduate Texts in Mathematics 194. Springer, New York.

Erkal, C., 2000. The simple pendulum: a relativistic revisit. European Journal of Physics 21, 377–384.

French, S., Ladyman, J., 2003. Remodeling structural realism: quantum physics and the metaphysics of structure. Synthese 136, 31–65.

Giere, R.N., 1988. Explaining Science. Chicago University Press, Chicago, IL.

Lasota, A., Mackey, M.C., 1994. Chaos, Fractals and Noise: Stochastic Aspects of Dynamics, second ed. Springer-Verlag, New York.

Suppes, P., 1957. Introduction to Logic. Van Nostrand, Princeton, NJ.

Suppes, P., 1988. Scientific Structures and Their Representation. Preliminary Version. Stanford University, CA.

Van Fraassen, B.C., 1994. Interpretation in science; science as interpretation. In: Hilgevoord, J. (Ed.), Physics and Our View of the World. Cambridge, New York, pp. 169–187.

Mathematical Psychology

Jean-Claude Falmagne and Michael D Lee, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article reviews the field from a historical perspective, starting from the works of Fechner and Thurstone, and outlining the basic theoretical concepts in four traditional areas: learning, psychophysics, measurement and choice, and response times. More recent topics that have emerged from these areas are also briefly mentioned together with their most prominent contributors.

Introduction

Mathematics has been used in psychology for a long time, and for different purposes. When William James (1890/1950) writes,

$$\text{Self-esteem} = \frac{\text{Success}}{\text{Pretensions}}$$

he is using mathematical notation metaphorically. Because no method is provided to measure the three variables, the equation cannot be taken literally. James means to convey, by a dramatic formula, the idea that if your pretensions increase without a corresponding increase of your success, your self-esteem will suffer. Such usages of mathematics, which Miller (1964) calls 'discursive,' can be found in psychological discourse at least since Aristotle. Despite their historical interest, we shall not review such metaphorical uses of mathematics here (see instead, Boring, 1950, or Miller, 1964).

In this article, we reserve the term *mathematical psychology* to the elaboration and the testing of mathematical theories and models for behavioral data. In principle, such a theory entails an economical representation of a particular set of data in mathematical terms, where 'economical' means that the number of free parameters of the theory is substantially smaller that the number of degrees of freedom in the data. In that sense, mathematical psychology plays for behavioral data the role that mathematical physics or mathematical biology play for physics or biology, respectively. In the best cases, the theory is cast in probabilistic terms and is testable by standard statistical methods. The large majority of such mathematical theories for behavioral data have emerged from four partially overlapping traditional fields: psychophysics, learning, choice, and response times. Each field will be outlined from the standpoint of the prominent mathematical models that have been proposed. Other topics to which mathematical psychologists have devoted much work are also mentioned.

The Precursors: Fechner and Thurstone

Gustav Theodor Fechner (1801–87; see Boring, 1950) was by training an experimental physicist with a strong mathematical background. While his interests were diverse and his contributions many, ranging from experimental physics to philosophy, we only consider him here as the founder of psychophysics. His main purpose was the measurement of 'sensation' in a manner imitating the methods used for the fundamental scales of physics, such as length or mass. Because 'sensation' could not be measured directly, Fechner proposed to evaluate the difference between the 'sensations' evoked by two stimuli by the difficulty of discriminating between them.

To be more specific, we introduce some notation. We write x, y, and so forth for positive real numbers representing physical intensities measured on some ratio scale, such as sound pressure level. Let $P(x, y)$ be the probability that stimulus x is judged to be louder than stimulus y. Cast in modern terms (cf Luce and Galanter, 1963; Falmagne, 1985/2002) Fechner's idea amounts to finding a real valued function u defined on the set of physical intensities, and a function F (both strictly increasing and continuous) such that

$$P(x, y) = F[u(x) - u(y)]. \qquad [1]$$

This equation is supposed to hold for all pairs of stimuli (x, y) such that $0 < P(x, y) < 1$ (subjectively, x is close to y). If such an equation holds, then the function u can be regarded as a candidate scale for the measurement of 'sensation' in the sense of Fechner. A priori, it is by no means clear that a scale u satisfying [1] necessarily exists. A necessary condition is the so-called *Quadruple Condition*: $P(x, y) \leq P(x', y') \Leftrightarrow P(x, x') \leq P(y, y')$, where the equivalence is assumed to hold whenever the four probabilities are defined. In the literature, the problem of constructing such a scale, or of finding sufficient conditions for its existence, has come to be labeled *Fechner's problem*. An axiomatic discussion of Fechner's problem can be found in Falmagne (1985/2002; see also Krantz et al., 1971).

For various reasons, partly traditional, psychophysicists often prefer to collect their data in terms of discrimination thresholds, which can be obtained as follows from the discrimination probabilities. We define a *sensitivity function* $\xi : (y, v) \mapsto \xi_v(y)$ by the equivalence: $\xi(y, v) = x \Leftrightarrow P(x, y) = v$. The *Weber function* $\Delta : (y, v) \mapsto \Delta_v(y)$, (y) is then defined by the equation $\Delta_v(y) = \xi_v(y) - \xi_{0.5}(y)$. (Weber, E.E. (1795–1878), a colleague of Fechner, professor of anatomy and physiology at Leipzig.). A representation of these fundamental concepts of classic psychophysics is given in **Figure 1**, in a special case where $P(y, y) = 0.5$. For each value of y, the S-shaped function $x \mapsto P(x, y)$ is called a *psychometric function*.

Starting with Weber himself, the Weber function Δ has been investigated experimentally for many sensory continua. In practice, $\Delta_v(y)$ is estimated by stochastic approximation

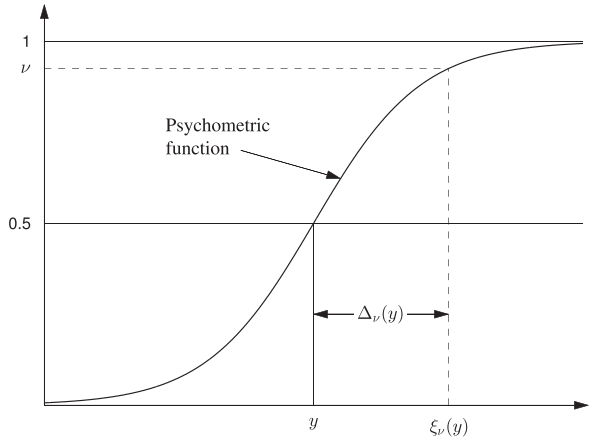

Figure 1 The Weber function Δ in the case where $P(y, y) = 0.5$ for all stimuli y; thus $\xi_{0.5}(y) = y$. The S-shape function $x \mapsto P(x, y)$ is referred to as a *psychometric function*.

(cf Robbins and Monro, 1951; Wasan, 1969; Levitt, 1970) for one or a few *criterion* values of the discrimination probability ν and for many values of the stimulus y. A typical finding is that, for stimulus values in the midrange of the sensory continuum and for some criterion values ν, $\Delta_\nu(y)$ grows approximately linearly with y, according to the equation:

$$\Delta_\nu(y) = yC(\nu), \qquad [2]$$

where C is a function depending on the criterion. The label *Weber's law* is attached to this equation. It is easily shown that [2] is equivalent to the homogeneity equation $P(\lambda x, \lambda y) = P(x, y)$, $(\lambda > 0)$, which in turn leads to $P(x, y) = H(\log x - \log y)$, with $H(s) = P(e^s, 1)$. This means that the function u in [1] has the form

$$u(y) = A \log y + B, \qquad [3]$$

where the constants $A > 0$ and B arise from uniqueness considerations for both u and F. Equation [3] has been dubbed *Fechner's law*. Our discussion indicates that, in the framework of (reffechone), Weber's law and Fechner's law are equivalent. Much of psychophysics evolved from Fechner's ideas.

The most durable contribution of Leon Louis Thurstone (1887–1955) to mathematical psychology is his *Law of Comparative Judgments* (Thurstone, 1927a, b; see Bock and Jones, 1968), a cornerstone of binary choice theory closely related to [1] of Fechner's problem. Thurstone supposed that a subject confronted with a choice between two alternatives x and y (where x, y are arbitrary labels and do not necessarily represent numerical values), makes a decision by comparing the sampled values of two random variables $\mathbf{U}_x$ and $\mathbf{U}_y$ associated to the alternatives. Suppose that these random variables are independent and normally distributed, with means $\mu(x)$, $\mu(y)$, and variances $\sigma^2(x)$, $\sigma^2(y)$, respectively. Denoting by Φ the distribution function of a standard normal random variable and with $\mathbb{P}$ standing for the probability measure, this leads to

$$P(x, y) = \mathbb{P}(\mathbf{U}_x > \mathbf{U}_y) = \Phi\left(\frac{\mu(y) - \mu(y)}{\sqrt{\sigma^2(x)^2 + \sigma^2(y)^2}}\right). \qquad [4]$$

Assuming that all the random variables have the same variance $\sigma^2 = \alpha^2/2$, we obtain

$$P(x, y) = \Phi(u(x) - u(y)), \qquad [5]$$

with $u(x) = \mu(x)/\alpha$ and $F = \Phi$, a special case of [1]. Equations [4] and [5] are called *Case III* and *Case V* of the *Law of Comparative Judgments*, respectively. Thurstone's model has been widely applied in psychophysics and choice. Other models in the same vein have been proposed by various researchers (e.g., Luce, 1959; Luce and Suppes, 1965). Thurstone's other important contributions are in learning theory (e.g., Thurstone, 1919), and especially Multiple-Factor Analysis (e.g., Thurstone, 1947).

The Beginning: Mathematical Learning Theory

Two articles mark the beginning of mathematical psychology as a distinguished research field: one by Estes (1950), *Toward a Statistical Theory of Learning*, and another by Bush and Mosteller (1951), *A Mathematical Model for Simple Learning*. These works were the highlights of a movement spearheaded by R. R. Bush and R. D. Luce at the University of Pennsylvania, and R. C. Atkinson, W. K. Estes, and P. Suppes at Stanford, which set out to formalize mathematical learning theories in terms of stochastic processes, and especially, Markov processes (cf Bharucha-Reid, 1960; Parzen, 1994). There were good reasons for such a development at that time. The previous decade had been plagued by fruitless controversies concerning the basic mechanisms of learning. While a considerable experimental literature on learning was available (cf Hilgard, 1943), the statistical tools in use for the analysis of the data were poor and the prominent theories ambiguous. When mathematics was used, it was metaphorically. Moreover, the scope of the theories was ambitious, covering a vast class of experimental situations loosely connected to each other conceptually. A typical example of such an endeavor is Hull (1943). By contrast, the Markov models developed by Bush, Estes, and their followers were designed for specific experimental situations. Under the influence of Suppes, a philosopher of science from Stanford who played a major role in the development of mathematical psychology, these models were often stated axiomatically. As a consequence, the predictions of the models could in many cases be derived by straightforward mathematical arguments. Two classes of models were investigated.

Finite State Markov Chains

In these models, the basic idea is that the subject's responses in a learning experiment are the reflection of some internal states, which coincide with the states of a finite Markov chain. The transitions of the Markov chain are governed by the triple of events (stimulus, response, reinforcement) occurring on each trial. In many situations, the number of states of the Markov chain is small. We sketch a simple case of the so-called *one-element model*, in which the chain has only two states that we denote by N and K. Intuitively, N stands for the 'naive' state, and K for the 'cognizant' state. On each trial of the experiment, the subject is presented with a stimulus and has to provide a response, which is either labeled as C (correct) or

as F (false). For example, suppose that the subject has to identify a rule used to classify some cards into two piles. Say, a drawing on each card has either one or two lines, which can be either both straight or both curved, and either both vertical or both horizontal. The subject must discover that all the cards with curved lines, and only those, go on the left pile. The subject is told on each trial whether the chosen pile is the correct one.

In words, the four axioms of the model are as follows: [M1] The subject always begins the experiment in the naive state N; [M2] the probability of a transition from the naive state N to the cognizant state K is equal to some parameter $0 < \theta < 1$, constant over trials, regardless of the subject's response; [M3] in state N, the probability of a correct placement is equal to a parameter $0 < \alpha < 1$, constant over trials; [M4] in state K, the response is always correct. The derivation of the model can either be based on a two-state Markov chain with state space $\{N, K\}$, or on a three-state Markov chain with state space $\{(N, C), (N, F), (K, C)\}$ (where N, K denote the subject's states and C, F the responses). In either case, the derivations are straightforward. Notice that, from axioms [M1] and [M2], the trial number of the occurrences of the first K state has a geometric distribution with parameter θ. Writing $\mathbf{S}_n$ and $\mathbf{R}_n$ for the cognitive state and the response provided on trial $n = 0, 1, \ldots$ (thus, $\mathbf{S}_n = N, K$ and $\mathbf{R}_n = C, F$), we get easily, for $n = 0, 1, \ldots$,

$$\begin{aligned} &\mathbb{P}(\mathbf{S}_n = N) = (1-\theta)^n, \\ &\mathbb{P}(\mathbf{R}_n = C|\mathbf{S}_n = N) = \alpha \\ &\mathbb{P}(\mathbf{R}_n = C|\mathbf{S}_n = K) = 1 \\ &\mathbb{P}(\mathbf{R}_n = C) = 1-(1-\theta)^n(1-\alpha), \\ &\text{which implies, with } p_n = \mathbb{P}(\mathbf{R}_n = C), \\ &p_{n+1} = (1-\theta)p_n + \theta. \end{aligned} \qquad [6]$$

A strong prediction of this model is that the number of correct responses recorded before an error occurring on trial number $(n+1)$ should be binomially distributed with parameters α and n (i.e., all the subjects' responses have been generated by the naive state). This prediction is surprisingly difficult to reject, at least in some situations (Suppes and Ginsburg, 1963).

Linear Operators Models

To facilitate the comparison, we take the same concept identification experiment as above, involving the classification of cards into two piles, and we consider a simple representative of this class of models. As before, we denote by $\mathbf{R}_n$ the response on trial n. We take as the sample space the set of all sequences $(\mathbf{R}_0, \mathbf{R}_1, \ldots, \mathbf{R}_n, \ldots)$, with $\mathbf{R}_n = C, F$ for $n = 0, 1, \ldots$. We also define

$$p_{w,0} = \mathbb{P}(\mathbf{R}_0 = C),$$

$$p_{w,n} = \mathbb{P}(\mathbf{R}_n = C|\mathbf{R}_{n-1}, \ldots, \mathbf{R}_0), \quad n = 1, 2, \ldots.$$

Let $0 < \theta < 1$ be a parameter. The axiom of the model has two cases, for $n = 0, 1, \ldots$,

$$[\text{L1}] \quad p_{w,n+1} = (1-\theta)\, p_{wn} + \theta \quad \text{if } \mathbf{R}_{n-1} = F,$$

$$[\text{L2}] \quad p_{w,n+1} = p_{wn} \quad \text{if } \mathbf{R}_{n-1} = C.$$

Thus, the model has the two parameters $p_{w,0}$ and θ, and learning occurs only when false responses are provided. As in the case of the finite Markov chains models, many predictions could be derived from such models, which could then be tested on the type of learning data traditionally collected by the experimenters. For both classes of models, the results of such enterprises were often quite successful. Nevertheless, the interest for such models waned during the 1960s, at least for learning situations, because researchers gradually realized that their simple mechanisms were far too primitive to capture all the intricacies revealed by more sophisticated analyses of learning data (see especially Yellott, 1969). General presentations of this topic can be found in Atkinson and Estes (1963) or Atkinson et al. (1965) for the finite state Markov learning models, and in Estes and Suppes (1959) or Sternberg (1963) for the linear operator models. A mathematical discussion of Markov processes for learning models is contained in Norman (1972).

Despite the partial failure of these models to provide a satisfactory detailed explanation of traditional learning data, their role was nevertheless essential in the introduction of modern probability theory (in particular stochastic processes) and axiomatic methods in theoretical psychology, and in promoting the emergence of mathematical psychology as a field of research. In recent years, a renewed interest in learning theory has appeared from the part of some economists.

Psychophysics

The main research topics in psychophysics can be traced back to the ideas of Fechner and Thurstone outlined in the first section of this article. Fechner's method of measuring sensation is indirect and based on the difficulty of discriminating between two stimuli. Under the impetus of S. S. Stevens, a psychologist from Harvard, different experimental methods for 'scaling' sensation became popular.

Direct Scaling Methods

In the case of the *magnitude estimation* method, Stevens (1957) asked subjects to make direct numerical judgments of the intensities of stimuli. For example, a subject may be presented with a pure tone of some intensity x presented binaurally, and would be required to estimate the magnitude of the tone on a scale from 1 to 100. Typically, each subject would only be asked to provide one or a couple of such estimations, and the data of many subjects would be combined into an average or median result, which we denote by $\phi(x)$. In many cases, these results would be fit reasonably well by the so-called *Power law* $\phi(x) = \alpha x^\beta$ or such variants as $\phi(x) = \alpha x^\beta + \gamma$ and $\phi(x) = \alpha(x-\gamma)^\beta$. In the *cross modality matching* method, the subject is presented with a stimulus from one sensory continuum (e.g., loudness), and is required to match its intensity with that of some other stimulus, from a different sensory continuum (e.g., brightness). Power laws were often also obtained in these situations. In the context of a discussion concerning the measurement of sensation, the difference of forms between [3] and the Power law was deemed important. While not much mathematical theorizing was involved in any particular application of these ideas, a real challenge was offered by the need to construct a comprehensive theory linking all important aspects of

psychophysical methods and data. The most ambitious effort in this direction is due to Krantz (1972). For a slightly different approach to direct scaling, see Anderson (1981).

Functional Equation Methods

Because the data from psychophysical experiments are typically noisy, the theoretician may be reluctant to make specific assumptions regarding the form of some functions entering in the equations of a model. An example of such a model is [1], in which the functions u and F are not specified a priori. In such cases, the equations themselves may sometimes specify the functions implicitly. For instance, if we assume that both Weber's law and [1] hold, then Fechner's law must also hold, that is, the function u must be logarithmic, as in [3]. Many more difficult cases have been analyzed (see Aczel et al., 2000; for the application of functional equation methods in the behavioral sciences).

Signal Detection Theory

Response strategies are often available to a subject in a psychophysical experiment. Consider a situation in which the subject must detect a low intensity stimulus presented over a background noise. On some trials, just the background noise is presented. The subject may have a bias to respond 'yes' on some trials even though no clear detection occurred. This phenomenon prevents a straightforward analysis of the data because some successful 'yes' responses may be due to lucky guesses. A number of 'signal detection' theories have been designed for parsing out the subject's response strategy from the data. The key idea is to manipulate the subject's strategy by systematically varying the *payoff matrix*, that is, the system of rewards and penalties given to the subject for his or her responses. These fall in four categories: correct detection or 'hit'; correct rejection; incorrect detection or 'false alarm'; and incorrect rejection or 'miss.' An example of a payoff matrix is displayed in **Table 1**, in which the subject collects four monetary units in the case of a correct detection and loses one such unit in the case of a false alarm.

For any payoff matrix θ, we denote by $p_s(\theta)$ and $p_n(\theta)$ the probabilities of a correct detection and of a false alarm, respectively. Varying the payoff matrix θ over conditions yields estimates of points $(p_s(\theta), p_s(\theta))$ in the unit square. It is assumed that these points lie on a receiver-operator-characteristic (ROC) curve representing the graph of some ROC function $\rho : p_n(\theta) \mapsto p_s(\theta)$. The function ρ is typically assumed to be continuous and strictly increasing. The basic notion is that the subject's strategy varies along the ROC curves, while the discriminating ability varies across these curves. The following basic random variable model illustrates this interpretation.

Table 1 An example of payoff matrix. The subject collects four monetary units in the case of a correct detection (or hit)

		Responses	
		Yes	*No*
Stimulus	Yes	4 (Hit)	−2 (Miss)
	No	−1 (False alarm)	3 (Correct rejection)

Suppose that to each stimulus s is attached a random variable $\mathbf{U}_s$ representing the effect of the stimulus on the subject sensory system. Similarly, let $\mathbf{U}_n$ be a random variable representing the effect of the noise on that system. The random variables $\mathbf{U}_s$ and $\mathbf{U}_n$ are assumed to be independent. We also suppose that the subject responds 'yes' whenever some threshold λ_θ, (depending on the payoff matrix θ), is exceeded. We obtain the two equations

$$p_s(\theta) = \mathbb{P}(\mathbf{U}_s > \lambda_\theta), \quad p_n(\theta) = \mathbb{P}(\mathbf{U}_n > \lambda_\theta). \qquad [7]$$

The combined effects of detection ability and strategy on the subject's performance can be disentangled in this model, however. Under some general continuity and monotonicity conditions and because $\mathbf{U}_s$ and $\mathbf{U}_n$ are independent, we get

$$p(\mathbf{U}_s > \mathbf{U}_n) = \int_{-\infty}^{\infty} \mathbb{P}(\mathbf{U}_s > \lambda)\mathrm{d}\mathbb{P}(\mathbf{U}_n \leq \lambda) = \int_0^1 \rho(p)\mathrm{d}p, \quad [8]$$

with p the ROC function and after changing variable from λ to $p_n(\lambda) = p$. Thus, for a fixed pair (s, n), the area under the ROC curve, which does not depend upon the subject's strategy, is a measure of the probability that $\mathbf{U}_s$ exceeds $\mathbf{U}_n$. Note that [8] remains true under any arbitrary continuous strictly increasing transformation of the random variables. For practical reasons, specific hypotheses are often made on the distributions of these random variables, which are (in most cases) assumed to be Gaussian, with expectations $\mu_s = E(\mathbf{U}_s)$ and $\mu_n = E(\mathbf{U}_n)$, and a common variance equal to 1. Replotting the ROC curves in (standard) normal–normal coordinates, we see that each replotted ROC curve is a straight line with a slope equal to 1 and an intercept equal to $\mu_s - \mu_n$.

Obviously, this model is closely related to Thurstone's Law of Comparative Judgments. Using derivations similar to those leading to [4] and [5] and defining $d'(s,n) = \mu_s - \mu_n$ we obtain $\mathbb{P}(\mathbf{U}_s > \mathbf{U}_n) = \Phi(d'(s,n)/\sqrt{2})$, an equation linking the basic signal detectability index d' and the area under the ROC curve. The index d' has become a standard tool not only in sensory psychology but also in other fields where the paradigm is suitable and the subject's guessing strategy is of concern. Multidimensional versions of this Gaussian signal detection model have been developed. Various other models have also been considered for such data, involving either different assumptions on the distributions of the random variables $\mathbf{U}_s$ and $\mathbf{U}_n$, or even completely different models, such as 'threshold' models (Krantz, 1969). Presentations of this topic can be found in Green and Swets (1974) and MacMillan and Creelman (2004).

Mathematical models for 'multidimensional' psychophysics were also developed. One approach takes the guise of 'geometric representations of perceptual phenomena,' which is the title of a seminal volume on the topic (Luce et al., 1995; see, in particular Indow, 1995). Another approach emphasizes psychological dissimilarity as the foundational concept (Dzhafarov, 2011).

Measurement and Choice

Because they were preoccupied with the scientific bases of their discipline, a number of mathematical psychologists have

devoted considerable efforts to the elucidation of the foundation of measurement theory, that is, the set of principles governing the use of numbers in the statement and discussion of scientific facts and theories. An account of the results can be found in the three volumes of *Foundation of Measurement* (Krantz et al., 1971; Luce et al., 1990; Suppes et al., 1989).

The literature on choice theory is extensive. Contributors come from diverse fields including mathematical psychology, but also microeconomics, political science, and business, for example, the latter two concerned with the study of voter or consumer choice. Early on, the literature was dominated by Thurstone's Law of Comparative Judgments, which still remains influential. Many other situations and models have been analyzed, however, and we only give a few pointers here. A generalization of Thurstone model is obtained by dropping the assumption of normality of the random variables $\mathbf{U}_x$ and $\mathbf{U}_y$ in the last part of [4]. Despite many attempts, the problem of characterizing this model in terms of conditions on the binary choice probabilities, posed by Block and Marschak (1960), is still unsolved. In other words, we do not know which set of necessary and sufficient conditions on the choice probabilities $P(x, y)$ guarantee the existence of the random variables $\mathbf{U}_x$, $\mathbf{U}_y$ satisfying the first part of [4] for all x and y in the choice set. A number of partial results have been obtained, however.

In the *multiple choice paradigm*, the subject is presented with a subset Y of a basic finite set $\mathscr{X}$ of objects and is required to select one of the objects in Y. We denote by $P(x; Y)$ the probability of selecting x in Y. By abuse of notation, we also write $P(X;Y) = \sum_{x \in X} P(x;Y)$. Suppose that $P > 0$. The *Choice Axiom*, proposed by Luce (1959), states that, for all $Z \subseteq Y \subseteq W \subseteq \mathscr{X}$, we have

$$P(Z;Y)P(Y;W) = P(Z;W). \qquad [9]$$

Defining the function $v: x \mapsto P(x; \mathscr{X})$, [9] yields immediately $P(x;Y) = v(x)[\sum_{y \in Y} v(y)]^{-1}$, for all $Y \subseteq \mathscr{X}$ and $x \in Y$. This model plays an important role in the literature. In the binary case, it has an interpretation in terms of random variables as in the Thurstone model, but these random variables, rather than being Gaussian, have a negative exponential distribution.

In the general case of such a random variable model for the multiple choice paradigm, we simply suppose that to each x in $\mathscr{X}$ is attached a random variable $\mathbf{U}_x$ such that, for all subsets X of $\mathscr{X}$ and all x in X, we have

$$P(x;X) = \mathbb{P}\left(\mathbf{U}_x = \max\{\mathbf{U}_y | y \in X\}\right). \qquad [10]$$

The general characterization problem for this model has been solved by Falmagne (1978) who states necessary and sufficient conditions for the existence of the random variables satisfying [10]. His article also contains a uniqueness result. As in the binary case, specific assumptions can be made on the distributions of these random variables.

Other models, based on different principles, have also been proposed for the multiple choice paradigm. For example, in the *elimination by aspects* model, due to Tversky (1972), a subject's choice of some object x in a set X is regarded as resulting from an implicit Markovian-type process gradually narrowing down the acceptable possibilities. For reviews of probabilistic choice models see Luce and Suppes (1965) or Suppes et al. (1989). A sample of some of the results can be found in Marley (1997).

Response Time Models

The time or latency of a response has been used as a behavioral index of the sensory or mental processes involved in the task since the inception of experimental psychology in the nineteenth century. Many mathematical models are based on Donders's idea that the observed response time is a sum of a number of unobservable components including at least a sensory, a decision, and a motor response part (Donders, 1868/1969). These models make various assumptions on the distributions of the component times, which are often taken to be independent. McGill (1963), for instance, assumes that the component times are all distributed exponentially and independently, with possibly different parameters, so that their sum is distributed as a general gamma random variable.

Another category of models is grounded on the assumption that the observed response results from an unobservable accumulation of evidence with absorbing boundaries. These models are known as sequential sampling models (Busemeyer and Rapoport, 1988; Laming, 1968; Link and Heath, 1975; Ratcliff, 1978; Vickers, 1979), and are based on sequential sampling methods from statistics, including the sequential probability ratio test (Wald and Wolfowitz, 1948). As psychological models, sequential sampling processes assume the latent accumulation of information or evidence from a stimulus, based on a series of samples. When the accumulated evidence reaches a boundary or threshold, the decision corresponding to that boundary is made, and the number of samples taken provides a measure of the time taken to make the decision. This means the models generate a joint distribution over both decisions and response times. There are also some theoretical accounts of confidence that can be related to the dynamics of sequential sampling processes, and so allow this third behavioral variable to be modeled (Vickers, 1979; Pleskac and Busemeyer, 2010). In some reaction time situations, the successive stimuli follow each other in a fast-paced sequence. In such cases, prominent sequential effects appear. Markovian models explaining such sequential effect have been developed (Falmagne, 1965; Falmagne et al., 1975).

Within the general sequential sampling framework, there are many models and model classes making different assumptions about the way evidence is accumulated, and the form of the boundaries. Random-walk or drift diffusion models (Ratcliff, 1978; Ratcliff and McKoon, 2008) assume a single tally is maintained; race or accumulator models (Smith and Vickers, 1988; Vickers, 1970) maintain separate tallies for each alternative decision; and ballistic models (Brown and Heathcote, 2008) also maintain separate tallies but without stochastic variability. In most models, the sampling process that generates evidence from a stimulus is assumed to be homogeneous, and boundaries are also assumed to be constant throughout the decision-making process. There are, however, some exceptions Smith (2000), especially considering more general utilities incorporating deadlines or time pressure (Frazier and Yu, 2008), or considering mechanisms for the learning or self-regulation of the boundaries over trials (Busemeyer and Myung, 1992; Simen et al., 2006; Vickers, 1979).

Sequential sampling models were at first mostly applied to simple visual and perceptual decision-making phenomena, but have also found application in two other areas. One of these is in modeling higher-order cognitive processes such as categorization (Nosofsky and Palmeri, 1997) and judgment and preference phenomena, especially through decision field theory (Busemeyer and Townsend, 1993). The other is in combining behavioral and neuroscientific data relating to the time course of simple decisions (Gold and Shadlen, 2007; Smith and Ratcliff, 2004). There is recent work on hierarchical extensions of the model, to account for individual differences in subjects, differences between stimuli, and other sources of variation beyond the level of a single decision trial (Rouder et al., 2003; Vandekerckhove et al., in press).

Early work in the modeling of response times relied on analytic results, making the comparison of models to data, and the estimation of parameters from data tractable. Computational approaches have played a progressively more important role, but there continues to be important foundational mathematical development (Navarro and Fuss, 2009; Smith, 2000). Luce (1986) is a basic reference for the models and their applications, and Smith (2000) and Bogacz et al. (2006) provide reviews of many of the relevant mathematical and statistical results.

Other Topics

From these four traditional areas, research in mathematical psychology has grown to include a wide variety of subjects. Current research includes many aspects of perception, cognition, memory, decision making, and more generally information processing (cf Dosher and Sperling, 1998). In some cases, the models can be seen as more or less direct descendants of those proposed by earlier researchers in the field. The multinomial process tree models (Batchelder and Riefer, 1999), for instance, is in the spirit of the Markovian models of the learning theorists.

However, as with response time modeling, the advent of powerful computers also gave rise to different types of models for which the predictions could be obtained by simulation, rather than by mathematical derivation. Representative of this trend are the parallel distributed processing or connectionist models (Rumelhart and McClelland, 1986), scaling and clustering models of stimulus representation (Shepard, 1980; Navarro and Griffiths, 2008), and a wide range of cognitive process models, including especially models of category learning (Ashby and Maddox, 2005; Kruschke, 2008; Nosofsky, 1992) and memory (Clark and Gronlund, 1996; Norman et al., 2008). Despite the computational emphasis, there continue to be important mathematical results for some of these models (Myung et al., 2007; Navarro, 2005).

Most recently, Bayesian methods have had a rapid and widespread influence over the areas traditionally studied by mathematical psychology. There are at least three types of Bayesian influence on the field (Lee, 2011). One type involves applying Bayesian statistics for data analysis, and to compare and evaluate models (Pitt et al., 2002; Rouder et al., 2009; Kruschke, 2011). Another influence is as a framework for extending cognitive process model accounts of complicated behavioral data, allowing the incorporation of individual differences, stimulus variability, covariate information, and other hierarchical and latent mixture structure (Lee, 2008; Rouder et al., 2007). The third influence Bayesian methods have had is as a theoretical metaphor for the mind, to contrast with alternatives like the information processing or connectionist metaphors. The Bayesian view treats the mind as applying Bayesian inference to sparse and noisy data to learn about richly structured mental hypothesis spaces (Chater et al., 2006). This approach has risen quickly in prominence, and many Bayesian models have been developed across a wide range of phenomena, including especially generalization, concept learning, and inductive reasoning (Anderson, 1991; Kemp and Tenenbaum, 2008; Griffiths and Tenenbaum, 2009; Tenenbaum and Griffiths, 2001).

Finally, we mention *psychometrics*. The research in psychometrics concerns the elaboration of statistical models and techniques for the analysis of test results. As suggested by the term, the main objective is the assignment of one or more numbers to a subject for the purpose of measuring some mental or physical traits. In principle, such a topic could be regarded as part of our subject. For historical reasons, however, this line of work has remained separate from mathematical psychology. The research on knowledge spaces is an alternative to psychometrics. Instead of measuring a person's numerical aptitude in a subject, it uses stochastic algorithms to uncover the knowledge state of a person, which is the set of all concepts mastered by the person in that subject (Falmagne and Doignon, 2011).

The Journals, the Researchers, the Society

The research results in mathematical psychology are mostly published in specialized journals such as the *Journal of Mathematical Psychology*, *Mathematical Social Sciences*, *Psychometrika*, *Econometrica*, and *Mathematiques, Informatique et Sciences Humaines*. Some of the work also appears in mainstream publications: *Psychological Review* or *Psychonomic Bulletin & Review*. Early on, the research was typically produced by psychologists, and the work often had a strong experimental component. The last couple of decades other researchers became interested in the field, coming especially from economics, applied mathematics, and machine learning.

The Society for Mathematical Psychology (SMP) was founded in 1979. The society manages the *Journal of Mathematical Psychology* and organizes a yearly meeting gathering several hundred participants coming from all over the world. The European Mathematical Psychology Group (EMPG) is an informal association of about one hundred scientists meeting every summer in some European university. The first meeting was in 1971.

See also: Bayesian Theory, History, Applications, and Contemporary Directions; Cognition: Knowledge Spaces; Cognitive Development: Mathematics Learning and Instruction; Cognitive Modeling: Connectionist Approaches; Computational

Approaches to Model Evaluation; Decision and Choice: Economic Psychology; Decision and Choice: Luce's Choice Axiom; Decision and Choice: Random Utility Models of Choice and Response Time; Decision and Choice: Sequential Decision Making; Diffusion and Random Walk Processes; Discrete-State Models of Information Processing; Dynamic Decision Making; Information Theory; Learning: Mathematical Learning Theory, History of; Markov Decision Processes; Mathematical Psychology: History; Measurement Theory: History and Philosophy; Measurement, Representational Theory of; Memory Models: Quantitative; Model Testing and Selection, Theory of; Psychophysical Theory and Laws, History of; Signal Detection Theory, History of; Signal Detection Theory.

Bibliography

Aczel, J., Falmagne, J., Luce, R., 2000. Functional equations in the behavioral sciences. Japonica Mathematica 52, 469–512.

Anderson, J., 1991. The adaptive nature of human categorization. Psychological Review 98, 409–429.

Anderson, N., 1981. Foundations of Information Integration Theory. Academic Press, New York.

Ashby, F., Maddox, W., 2005. Human category learning. Annual Review of Psychology 56, 149–178.

Atkinson, R., Bower, G., Crothers, E., 1965. An Introduction to Mathematical Learning Theory. Wiley, New York.

Atkinson, R., Estes, W., 1963. Stimulus sampling theory. In: Luce, R.D., Bush, R.R., Galanter, E. (Eds.), Handbook of Mathematical Psychology, vol. 2. Wiley, New York, pp. 121–268.

Batchelder, W., Riefer, D., 1999. Theoretical and empirical review of multinomial process tree modeling. Psychonomic Bulletin & Review 6, 57–86.

Bharucha-Reid, A., 1960. Elements of the Theory of Markov Processes and Their Applications. McGraw-Hill, New York.

Block, H., Marschak, J., 1960. Random ordering and stochastic theories of responses. In: Olkin, I., Ghurye, S., Hoeffding, W., Madow, W., Mann, H. (Eds.), Contributions to Probability and Statistics. Stanford University Press, Stanford, CA, pp. 97–132.

Bock, R., Jones, L., 1968. The Measurement and Prediction of Judgement and Choice. Holden-Day, San Francisco, CA.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., Cohen, J., 2006. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced choice tasks. Psychological Review 113, 700–765.

Boring, E., 1950. A History of Experimental Psychology. Prentice Hall, Englewood Cliffs, NJ.

Brown, S., Heathcote, A., 2008. The simplest complete model of choice response time: linear ballistic accumulation. Cognitive Psychology 57, 153–178.

Busemeyer, J., Myung, I., 1992. An adaptive approach to human decision making: learning theory, decision theory, and human performance. Journal of Experimental Psychology: General 121, 177–194.

Busemeyer, J., Rapoport, A., 1988. Psychological models of deferred decision making. Journal of Mathematical Psychology 32, 91–134.

Busemeyer, J., Townsend, J., 1993. Decision field theory: a dynamic-cognitive approach to decision making. Psychological Review 100, 432–459.

Bush, R., Mosteller, F., 1951. A mathematical model for simple learning. Psychological Review 58, 313–323.

Chater, N., Tenenbaum, J., Yuille, A., 2006. Probabilistic models of cognition: conceptual foundations. Trends in Cognitive Sciences 10, 287–291.

Clark, S., Gronlund, S., 1996. Global matching models of recognition memory: How the models match the data. Psychonomic Bulletin & Review 3, 37–60.

Donders, F., 1868/1969. On the speed of mental processes. Acta Psychologica 30, 412–431. (Translated by W. G. Koster).

Dosher, B., Sperling, G., 1998. A century of information processing theory: vision, attention and memory. In: Hochberg, J., Cutting, J.E. (Eds.), Handbook of Perception and Cognition at Century's End: History, Philosophy, Theory. Academic Press, San Diego, CA, pp. 199–252.

Dzhafarov, E., 2011. Mathematical foundations of universal Fechnerian scaling. In: Berglund, B., Rossi, G.B., Townsend, J.T., Pendrill, L. (Eds.), Measurements with Persons. Psychology Press, New York, pp. 185–210.

Estes, W., 1950. Toward a statistical theory of learning. Psychological Review 57, 94–107.

Estes, W., Suppes, P., 1959. Foundations of linear models. In: Bush, R., Estes, W. (Eds.), Studies in Mathematical Learning Theory. Stanford University Press, Stanford, CA, pp. 137–179.

Falmagne, J.C., 1965. Stochastic models for choice reaction time with application to experimental results. Journal of Mathematical Psychology 2, 77–124.

Falmagne, J.C., 1978. A representation theorem for finite random scale systems. Journal of Mathematical Psychology 18, 52–72.

Falmagne, J.C., 1985/2002. Elements of Psychophysical Theory. Oxford University Press, New York. (Reprinted in 2002).

Falmagne, J.C., Cohen, S., Dwivedi, A., 1975. Two-choice reactions as an ordered memory scanning process. In: Rabbit, P., Dornic, S. (Eds.), Attention and Performance V. Academic Press, pp. 296–344.

Falmagne, J.C., Doignon, J.P., 2011. Learning spaces. In: Interdisciplinary Mathematics Series. Springer, Heidelberg.

Frazier, P., Yu, A., 2008. Sequential hypothesis testing under stochastic deadlines. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (Eds.), Advances in Neural Information Processing Systems, vol. 20. MIT Press, Cambridge, MA, pp. 465–472.

Gold, J., Shadlen, M., 2007. The neural basis of decision making. Annual Review of Neuroscience 30, 535–574.

Green, D., Swets, J., 1974. Signal Detection Theory and Psychophysics. Krieger, New York.

Griffiths, T., Tenenbaum, J., 2009. Theory-based causal induction. Psychological Review 116, 661–716.

Hilgard, E., 1943. Theories of Learning, second ed. Appleton Century Crofts, New York.

Hull, C., 1943. Principles of Behavior. Appleton Century Crofts, New York.

Indow, T., 1995. Psychophysical scaling: scientific and practical applications. In: Luce, R.D., D'Zmura, M., Hoffman, D.D., Iverson, G.J., Romney, A.K. (Eds.), Geometric Representations of Perceptual Phenomena. Papers in Honor of Tarow Indow for his 70th Birthday. Erlbaum, Mahwah, NJ, pp. 1–28.

James, W., 1890/1950. The Principles of Psychology, vol. I. Holt. (Reprinted by Dover Publications).

Kemp, C., Tenenbaum, J., 2008. The discovery of structural form. Proceedings of the National Academy of Sciences 105, 10687–10692.

Krantz, D., 1969. Threshold theories of signal detection. Psychological Review 76, 308–324.

Krantz, D., 1972. A theory of magnitude estimation and cross modality matching. Journal of Mathematical Psychology 9, 168–199.

Krantz, D., Luce, R., Suppes, P., Tversky, A., 1971. Foundations of measurement. In: Additive and Polynomial Representations, vol. 1. Academic Press, New York.

Kruschke, J., 2008. Models of categorization. In: Sun, R. (Ed.), The Cambridge Handbook of Computational Psychology. Cambridge University Press, New York, pp. 267–301.

Kruschke, J., 2011. Doing Bayesian Data Analysis: A Tutorial with R and BUGS. Academic Press/Elsevier.

Laming, D., 1968. Information Theory of Choice Reaction Time. Academic Press, London.

Lee, M., 2008. Three case studies in the Bayesian analysis of cognitive models. Psychonomic Bulletin & Review 15, 1–15.

Lee, M., 2011. How cognitive modeling can benefit from hierarchical Bayesian models. Journal of Mathematical Psychology 55, 1–7.

Levitt, H., 1970. Transformed up-down methods in psychoacoustics. Journal of the Acoustical Society of America 49, 467–476.

Link, S., Heath, R., 1975. A sequential theory of psychological discrimination. Psychometrika 40, 77–105.

Luce, R., 1959. Individual Choice Behavior. Wiley, New York.

Luce, R., 1986. Response Times: Their Role in Inferring Elementary Mental Organization. Oxford University Press, New York.

Luce, R., D'Zmura, T., Hoffman, D., Iverson, G.J., Romney, A.K. (Eds.), 1995, Geometric Representations of Perceptual Phenomena. Papers in Honor of Tarow Indow for his 70th Birthday. Erlbaum, Mahwah, NJ.

Luce, R., Galanter, E., 1963. Discrimination. In: Luce, R., Bush, R., Galanter, E. (Eds.), Handbook of Mathematical Psychology, vol. 1. Wiley, New York, pp. 191–244.

Luce, R., Krantz, D., Suppes, P., Tversky, A., 1990. Foundations of measurement. In: Representation, Axiomatization, and Invariance, vol. 3. Academic Press, San Diego, CA.

Luce, R., Suppes, P., 1965. Preference, utility and subjective probability. In: Luce, R.D., Bush, R.R., Galanter, E. (Eds.), Handbook of Mathematical Psychology, vol. 3. Wiley, pp. 252–410.

MacMillan, N., Creelman, C., 2004. Detection Theory: A User's Guide, second ed. Erlbaum, Hillsdale, NJ.

Marley, A. (Ed.), 1997. Choice, Decision and Measurement. Papers in Honor of R. Duncan Luce. Erlbaum, Mahwah, NJ.

McGill, W., 1963. Stochastic latency mechanisms. In: Luce, R.D., Bush, R.R., Galanter, E. (Eds.), Handbook of Mathematical Psychology, vol. 1. Wiley, New York, pp. 309–360.
Miller, G., 1964. Mathematics and Psychology. McGraw-Hill, New York.
Myung, J., Montenegro, M., Pitt, M., 2007. Analytic expressions for BCD-MEM models of recognition memory. Journal of Mathematical Psychology 51, 198–204.
Navarro, D., 2005. Analyzing the RULEX model of category learning. Journal of Mathematical Psychology 49, 259–275.
Navarro, D., Fuss, I., 2009. Fast and accurate calculations for first-passage times in Wiener diffusion models. Journal of Mathematical Psychology 53, 222–230.
Navarro, D., Griffiths, T., 2008. Latent features in similarity judgment: a nonparametric Bayesian approach. Neural Computation 20, 2597–2628.
Norman, F., 1972. Markov Processes and Learning Models. Academic Press, New York.
Norman, K., Detre, G., Polyn, S., 2008. Computational models of episodic memory. In: Sun, R. (Ed.), The Cambridge Handbook of Computational Psychology. Cambridge University Press, New York, pp. 189–224.
Nosofsky, R., 1992. Similarity scaling and cognitive process models. Annual Review of Psychology 43, 25–53.
Nosofsky, R., Palmeri, T., 1997. An exemplar-based random walk model of speeded classification. Psychological Review 104, 266–300.
Parzen, E., 1994. Stochastic Processes. Holden-Day, San Francisco, CA.
Pitt, M., Myung, I., Zhang, S., 2002. Toward a method of selecting among computational models of cognition. Psychological Review 109, 472–491.
Pleskac, T., Busemeyer, J., 2010. Two-stage dynamic signal detection: a theory of confidence, choice, and response time. Psychological Review 117, 864–901.
Ratcliff, R., 1978. A theory of memory retrieval. Psychological Review 85, 59–108.
Ratcliff, R., McKoon, G., 2008. The diffusion decision model: theory and data for two-choice decision tasks. Neural Computation 20, 873–922.
Robbins, H., Monro, S., 1951. A stochastic approximation method. The Annals of Mathematical Statistics 22, 400–407.
Rouder, J., Lu, J., Speckman, P., Sun, D., Morey, R., Naveh-Benjamin, M., 2007. Signal detection models with random participant and item effects. Psychometrika 72, 621–642.
Rouder, J., Speckman, P., Sun, D., Morey, R., Iverson, G., 2009. Bayesian t-tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin & Review 16, 225–237.
Rouder, J., Sun, D., Speckman, P., Lu, J., Zhou, D., 2003. A hierarchical Bayesian statistical framework for response time distributions. Psychometrika 68, 589–606.
Rumelhart, D., McClelland, J. (Eds.), 1986, Parallel Distributed Processing. Exploration in the Microstructure of Cognition, vol. 1. MIT Press, Cambridge, MA.
Shepard, R., 1980. Multidimensional scaling, tree-fitting, and clustering. Science 214, 390–398.
Simen, P., Cohen, J., Holmes, P., 2006. Rapid decision threshold modulation by reward rate in a neural network. Neural Networks 19, 1013–1026.
Smith, P., 2000. Stochastic dynamic models of response time and accuracy: a foundational primer. Journal of Mathematical Psychology 44, 408–463.
Smith, P., Ratcliff, R., 2004. The psychology and neurobiology of simple decisions. Trends in Neurosciences 27, 161–168.
Smith, P., Vickers, D., 1988. The accumulator model of two-choice discrimination. Journal of Mathematical Psychology 32, 135–168.
Sternberg, S., 1963. Stochastic learning theory. In: Luce, R.D., Bush, R., Galanter, E. (Eds.), Handbook of Mathematical Psychology, vol. 2. Wiley, New York.
Stevens, S., 1957. On the psychophysical law. Psychological Review 64, 153–181.
Suppes, P., Ginsburg, R., 1963. A fundamental property of all-or-none models, binomial distribution of responses prior to conditioning, with application to concept formation in children. Psychological Review 70, 139–171.
Suppes, P., Krantz, D., Luce, R., Tversky, A., 1989. Foundations of Measurement, vol. 2. Academic Press, San Diego, CA.
Tenenbaum, J., Griffiths, T., 2001. Generalization, similarity, and Bayesian inference. Behavioral and Brain Sciences 24, 629–640.
Thurstone, L., 1919. The learning curve equation. Psychological Monographs 26, 1–51.
Thurstone, L., 1927a. A law of comparative judgement. Psychological Review 34, 273–286.
Thurstone, L., 1927b. Psychological analysis. American Journal of Psychology 38, 368–389.
Thurstone, L., 1947. Multiple-Factor Analysis. University of Chicago Press.
Tversky, A., 1972. Elimination by aspects: a theory of choice. Psychological Review 79, 281–299.
Vandekerckhove, J., Tuerlinckx, F., Lee, M. Hierarchical diffusion models for two-choice response time. Psychological Methods, in press.
Vickers, D., 1970. Evidence for an accumulator model of psychophysical discrimination. Ergonomics 13, 37–58.
Vickers, D., 1979. Decision Processes in Visual Perception. Academic Press, New York, NY.
Wald, A., Wolfowitz, J., 1948. Optimal character of the sequential probability ratio test. Annals of Mathematical Statistics 19, 326–339.
Wasan, M., 1969. Stochastic approximation. Cambridge Tracts in Mathematics and Mathematical Physics 58.
Yellott Jr, J., 1969. Probability learning with non contingent success. Journal of Mathematical Psychology 6, 541–575.

Relevant Website

http://www.mathpsych.org – The Society for Mathematical Psychology.

Mathematical Psychology: History

William H Batchelder, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W.K. Estes, volume 14, pp. 9412–9416, © 2001, Elsevier Ltd.

Abstract

The application of mathematics to certain problems within the field of psychology dates back to at least the seventeenth century. This article reviews some of these early applications, most of which either involve theories for experimental phenomena or statistical methods for measuring individual differences. These later applications led to the field of psychometrics in the 1930s; and the former led to the field of mathematical psychology in the 1950s, and both fields are active today. Early mathematical psychology was characterized by testable, formal theories in the areas of learning and memory, perception and psychophysics, choice and decision making, language and thinking, and measurement and scaling; and these areas still characterize the field today.

Introduction

The scientific use of mathematics to understand empirical observations has played a role in psychological theory at least as early as the seventeenth century; however, the term *mathematical psychology* does not simply refer to the scientific application of mathematics to psychology. Instead, mathematical psychology refers to a field of psychology established in the late 1950s by a small group of interdisciplinary scholars who were dissatisfied with the state of theory in experimental psychology at that time. These scholars had a vision for new ways to develop testable psychological theories that were informed by recent theoretical developments in modern mathematics and the information sciences. The field rapidly grew in the 1960s, and it is quite active today with contributions to all the areas of theoretical psychology.

The immediate groundwork for the development of mathematical psychology occurred with a flurry of papers and books in the 1950s that provided information-theoretic underpinnings for psychological theory in several research areas of psychology, notably: (1) learning and memory, (2) perception and psychophysics, (3) choice and decision making, (4) language and thinking, and (5) measurement and scaling. Many of these papers were reprinted in two volumes, titled *Readings in Mathematical Psychology* edited by Luce et al. (1963a, 1965a). At the same time, three volumes of the *Handbook of Mathematical Psychology* appeared, also edited by Luce et al. (1963b,c; 1965b), each with several long chapters representing new developments in the five research areas of the *Readings* volumes. In 1964, the *Journal of Mathematical Psychology* was started, and a year later *The British Journal of Statistical Psychology* changed its name to *The British Journal of Mathematical and Statistical Psychology*. Shortly thereafter, an annual conference on mathematical psychology was held in 1968 in the United States that eventually led to the formation of the Society for Mathematical Psychology. Also an annual conference on mathematical psychology started in Europe in 1971, and this led to the European Mathematical Psychology Group (EMPG). In addition, a growing number of universities offered graduate programs in mathematical psychology. With these developments, the establishment of mathematical psychology as an accepted field of psychology was complete (*see* Markov Decision Processes).

The remainder of this article will present some of the key mathematical developments in psychology that preceded the establishment of mathematical psychology as a formal field of psychology. A few of these developments occurred in the seventeenth and eighteenth centuries, before psychology became an accepted academic area of study; however, most occurred in the nineteenth century and in the early twentieth century. In the later part of the nineteenth century, psychology became recognized as a separate academic area of study in Germany and England. The main emphasis of psychology was quite different in these two countries, and this was reflected in their mathematical approaches. The German approach emphasized experimental psychology with a goal to understanding the common psychological processes that all humans share (normative psychology), where as the English focus was on measuring differences between humans on their psychological traits (differential psychology). In this time frame, the mathematical developments in psychology reflected these two approaches; however, the impetus that led to the creation of mathematical psychology was much more aligned with the experimental approach.

Mathematics and Psychology before the Nineteenth Century

Although the scientific use of mathematics before the nineteenth century was substantial, it was mostly seen in the areas of physics and chemistry. However, in that period a few important mathematical applications occurred in areas related to psychology, especially in the area of choice and decision making. Most modern theories of decision making assume that when making a choice between two or more probabilistic options, the decision maker considers and integrates in some way both the subjective probabilities of the options and their values (e.g., utilities) to the decision maker. The root of this idea goes back at least to Blaise Pascal (1623–62) in *Pensées*, a book constructed and published posthumously from his notes in 1670. In his notes, he

considers how a person facing a compulsory wager should choose between two incompatible, uncertain options when there are compelling arguments that either option may be the more likely. He argues that in such a case, the decision maker must consider the consequences of each option, or in other words, weigh the gains and losses. He goes on to apply his approach to a hypothetical wager on whether or not to believe that (a Christian) God exists. This situation has become known as Pascal's wager, and in his analysis of the consequences of each option, he sees a clear reason to select the option that God exists based on the supposed infinite positive value of eternal life compared with the minor negative value resulting from the embarrassment of being incorrect.

In the field of memory research, the scientist and inventor Robert Hooke (1635–1703) read a paper to the Royal Society of London in 1682 detailing a geometric model of human memory (Hintzman, 2003). The model appears to foreshadow many of the topics seen in the modern theories of memory; for example, explanations of encoding, retrieval, priming, and forgetting. In fact some of the aspects of Hooke's theory lead to mathematical laws; for example, Hintzman shows how one can derive a power function for forgetting based on the model, namely $s(l) = cl^{-2/3}$, where $s(l)$ is the memory strength after l time units since a memory was originally acquired and c is a constant depending on the nature of the memory. The shape of this forgetting function comes quite close to current studies of retention. Unfortunately, Hooke did not run any experiments, and apart from a few studies in perception and psychophysics in the eighteenth century discussed later, developers of mathematical theories in psychology did not conduct experiments to support their theories until the middle of the nineteenth century.

In the eighteenth century, additional grounding in decision making is attributable to Nicholas Bernoulli (1687–1759) and Daniel Bernoulli (1700–82). In 1713, Nicholas Bernoulli proposed what later became known as the St. Petersburg Paradox (Martin, 2011). The paradox concerns a lottery where you bet a sum of money and then flip a fair coin repeatedly until the first head appears. Your payoff for the bet depends on the number of flips needed to get the first head. In particular, if n is the number of flips until the first head appears, then you receive 2^{n-1} monitory units. Bernoulli shows that the expected value of this game is infinite, $\sum (½)^n 2^{n-1} = \infty$; therefore, he observed that one ought to be willing to pay any price to play the game. Of course it is obvious that a person would not pay a large amount to play, so this is where the paradox appears. A solution to the paradox was offered in 1738 by Daniel Bernoulli, who introduced the idea that a decision maker should consider utility rather than price, and this distinction has become an important component of modern decision-making theories. He pointed out that a given amount of money has different utilities depending on the needs of the decision maker, and that for all decision makers, the utility of money has decreasing marginal utility. He conjectured that the function relating price to utility was logarithmic, and that the initial wealth of the player must also be considered. Later others, including Laplace (1749–1827), proposed modifications to Daniel Bernoulli's analysis; however, the net effect of the St. Petersburg paradox was to provide a foundation for the critical variables in determining 'rational' decision-making behavior.

Considerations involving probability were also evident in the establishment of Bayesian reasoning in the eighteenth century. This development was due to the Reverend Thomas Bayes (1701–61), and to his important work, which was presented posthumously to the Royal Academy in 1763 in "An Essay Toward Solving a Problem in the Doctrine of Chances" (Bayes and Price, 1763). Although in the essay, he does not solve a problem directly connected with psychology, he did provide a new approach to statistical thinking that continues to have a great influence today in the statistical sciences.

Earlier work with probabilities had shown how to calculate the probability of an outcome given an explicit probability distribution; however, the key innovation in Bayes' work was to solve the so-called inverse probability problem, namely that given an outcome what can be said about the underlying probability that generated it. In particular, he showed a logically reasonable way to compare the probability of an event when one has no prior information versus when one has information consisting of the number of occurrences of the event out of n independent tries. Using his logic, he derived the beta posterior distribution for the uniform-binomial problem. In modern terms, this problem assumes that in the case that one knows nothing about the probability of the event, the prior probability p is uniformly distributed, $\pi(p) = 1$, $0 \leq p \leq 1$, and when one has knowledge of, say, s occurrences of the event out of n trials, then the posterior probability of p is given by:

$$g(p|s) = \frac{(n+2)!}{(s+1)(n-s+1)!} p^s (1-p)^{n-s}, \qquad [1]$$

with a posterior mean of $(s+1)/(n+2)$. Using Bayes' own reasoning, it is a small step to generalize eqn [1] to handle other probability distributions and types of evidence. Today, Bayesian statistics is a major area of statistical theory, and in mathematical psychology Bayesian ideas are increasingly influencing not only the structure of cognitive models but also the methods for estimating and evaluating them when confronted with data.

Other areas of psychology before the eighteenth century received very few applications of mathematics. Many historians of science attribute this lack to the influential philosopher, Immanuel Kant (1724–1804), whose writings are often interpreted as expressing the view that psychology cannot be a mathematical, experimental science (e.g., Hergenhahn, 2009). For example, in the *Metaphysical Foundations of Natural Science* of 1786, Kant points out that every natural science is essentially grounded in mathematics, and he goes on to argue that mathematics cannot be applied to the phenomenon of the internal senses. In an interesting essay, Sturm (2006) argues that Kant meant his remarks only to reflect the view that a science of psychology cannot be grounded on data provided by introspective (self) reports. Sturm goes on to show that in the period of Kant's writing, there were a number of ongoing experimental studies in the areas of perception and psychophysics, some known to Kant.

Examples were studies of visual acuity as a function of distance from a light source, brightness and color discrimination, and the durations of the after effects of visual, auditory, and tactile sensory experiences; and some of these studies included mathematical relationships between variables.

Mathematics and Psychology in the Nineteenth Century

Psychology achieved status as a separate academic field in the late nineteenth century after many experimental and theoretical developments. The transformation of psychology into a recognized university subject was due to research developments mostly in Germany and England. These developments led to quite different ways of using mathematics in psychological theory. In particular, the German approach emphasized the experimental method, where behavior is observed as a function of stimulus conditions manipulated by an experimenter. The primary goal of the experimental method is to develop theoretical laws of cognition that apply to all humans. On the other hand, the English approach emphasized how humans differ, and because these differences are due to both the genetic makeup and the environmental experiences of the individual, the experimental method is not of much use. Instead, the English approach involved developing and applying correlational methods to measure individual differences in psychological traits such as intelligence, attitudes, or personality traits. Cronbach (1957) has a nice discussion of the differences between the experimental and correlational approaches to psychological theory, and the next two sections discuss mathematical applications characteristic of these two approaches in the nineteenth century.

Mathematical Developments in German Psychology

In the case of Germany, most historians select Wilhelm Wundt (1832–1920) and the establishment of his experimental psychology laboratory at Leipzig, Germany in 1879 as the start of academic psychology in Germany (e.g., Hergenhahn, 2009). Before Wundt's laboratory was started, the German tradition was underway in the 1800s with major work in neural and sensory processes, psychophysics, attention, and choice response time. This work provided the main underpinnings of what was to become the field of experimental psychology. The early mathematical developments in German psychology were mostly in the areas of sensory processes and psychophysics, areas that enjoyed a connection to physiology and physics; however, mathematical developments in other more cognitive areas such as memory and choice reaction time occurred later in the century.

The area of psychophysics concerns the relationship between physical stimuli and the sensations that they create. The groundwork for psychophysics was established by the physiologist, Ernst Weber (1795–1878), who created the first experimentally validated mathematical law of the mind. Based on extensive experimental studies across a wide range of stimulus modalities, Weber noted that the smallest detectable increase in the intensity of a physical stimulus depends on its magnitude. This work led him to formulate the famous Weber's law, namely that the additional physical intensity needed to perceive a noticeable increase in sensation is directly proportional to the physical intensity, that is:

$$\Delta I = kI, \qquad [2]$$

where ΔI is the just noticeable increment in stimulus intensity I, and k is a constant that depends on the sensory modality. Now we know that Weber's law only holds strictly for a limited range of physical intensities, but the fact that it holds across a variety of modalities makes it an important theoretical milestone.

Gustav Fechner (1801–87), a physicist and mathematician, contributed many mathematical ideas in psychological theory connecting physical properties of stimuli to the sensations and perceptions they create. Much of this work was published in a two-volume work, *Elemente der Psychophysik* (*Elements of Psychophysics*). His most notable contribution was that he utilized Weber's law along with other mathematical assumptions to create his famous psychophysical law:

$$S = k \log I, \qquad [3]$$

namely that the perceived magnitude of a sensation S is functionally related to the logarithm of the corresponding physical intensity I that produced it. Fechner's work led to the creation of a large research field of psychology known as psychophysics, and many additional mathematical developments in this area followed on his ground-breaking work (*see* Markov Decision Processes and Psychophysical Theory and Laws, History of for details on Weber's and Fechner's contributions).

Like Fechner, Hermann von Helmholtz (1821–94) was another physicist and mathematician who made major contributions to experimental psychology. Among his many accomplishments outside of physics was his development of the resonance theory of hearing and the Young–Helmholtz theory of color vision. Both of these theories were invested in explaining how the human perceiver could detect a huge range of sound frequencies or colors with many fewer types of neurons. Also very important was his experimental work discovering that the speed of neural conduction was about ten times slower than the speed of sound, and is therefore capable of measurement.

Capitalizing on Helmholtz's discovery about the measurable rate of neural conduction, Franciscus Cornelius Donders (1818–89) devised an experimental approach along with a mathematical model to measure the amount of time that the mind needs to carry out elementary operations like perceiving a simple stimulus, discriminating between two simple stimuli, and making a simple choice. His mathematical ideas led to the so-called subtraction method. The essence of this method involves setting up two experimental conditions involving speeded responses, where one task is identical to the other except that it involves an additional elementary mental act. Then Donders' logic is to subtract the response time it takes to perform the simpler task from the response time to perform the more complex task, and the difference is then inferred to be the time that the mind needs to perform the additional act. Although the accuracy of his early measurements was not great, he is credited with

starting the area of mental chronometry, which is one of the main ways that experimental psychologists learn about mental information-processing architecture.

While most of the applications of mathematics in psychology in the German tradition were in the areas of sensation and perception, there were a few mathematical developments in the more cognitive areas of the mind as well. The most noted example was due to Johann Friedrich Herbart (1776–1841). Herbart developed an extensive mathematical system for understanding mental activity in terms of the interplay of combinations of ideas competing with each other to reach consciousness. The system was based on deterministic, mechanical laws, and at least some implications of his theory had the potential to be tested by experiment. Unfortunately, Herbart, who was a strong advocate of Kant's philosophy, interpreted Kant's writing about psychology as indicating that experimental psychology was impossible (see the previous section), and so unfortunately Herbart never tried to conduct experiments to test his theory (e.g., Boudewijnse et al., 1999).

Mathematical theories of mental activity that were connected to experimentation occurred much later in the nineteenth century. For example, Hermann Ebbinghaus (1850–1909) was the first psychologist to conduct rigorous experimental studies of memory. Ebbinghaus conducted numerous serial list-learning experiments on himself that verified some of the philosophical theories of memory association that trace back to Aristotle. Among his mathematical formulations is his famous exponential decay curve for forgetting in long-term memory. In fact, Ebbinghaus attributed his reading of Fechner's *Elemente der Psychophysik* with the ideas that led him to formulate his forgetting curve. Just as Fechner showed that increases in physical stimulus intensity exhibit decreasing marginal returns on sensation, Ebbinghaus showed that increases in time since study exhibit decreasing amounts of marginal forgetting.

Mathematical Developments in English Psychology

The English tradition in psychology can be traced to the work of Charles Darwin (1809–82) and especially research that followed in the spirit of his work. Darwin's theory of evolution was grounded on several principles, not the least of which was the fact that individuals differ in their physical traits. It was the work of Darwin's cousin, Sir Francis Galton (1822–1911), that set the stage for the emergence of English psychology. Galton's psychological concerns were mostly centered on ways to measure individual differences in a variety of personality and cognitive traits. Galton's interests in individual differences led him to establish a laboratory in 1884 where he studied differences in both physical and psychological traits in over 9000 individuals. The mathematical developments that came out of the English approach to psychology all involved the development of statistical theory. In fact, Galton is perhaps the principle figure behind the fact that England became a center for the development of statistical theory in the century after his birth (Stigler, 1986).

Galton was a pioneer in introducing the bivariate normal distribution to model the relationship between two measurable traits from the same individual. In this work, he introduced the concept of regression to the mean, which refers to the phenomenon that if an observation of a variable is extreme on its first measurement, then it is likely to be closer to the mean on its second measurement. Galton also laid the groundwork for the concept of the correlation coefficient as the measurement of the degree of statistical connection between two traits across a series of individuals. One impetus for this development was the so-called *Galton's problem*, which refers to the difficulty in drawing conclusions from cross-cultural measurements because they exhibit autocorrelation.

The final development of a measure of correlation was the Pearson product moment correlation coefficient, due to the statistician and mathematician Karl Pearson (1857–1936), who was a protégé of Galton. At the population level, the correlation between two random variables, X and Y, representing two traits of an individual is given by:

$$\rho(X,Y) = \frac{E\{[X-E(X)][Y-E(Y)]\}}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}}, \qquad [4]$$

where $E(\bullet)$ is the expectation operator and $\mathrm{var}(X) = E[(X-E(X)]^2$.

The correlation coefficient at the level of a sample from a population became the central tool in the early studies of individual differences because of its many desirable properties such as its scale, $-1 \le \rho(X, Y) \le 1$, its role in providing best linear fits to scatter plots of individual trait variations, and the fact that ρ^2 is a measure of the variance in one trait 'explained' by the values of another trait.

Galton was quite interested in the nature of measurement error, an area of little interest except when regarded as a nuisance factor to experimental psychologists in the German tradition. He invented a model of measurement error known as Galton's box (quincunx) to illustrate the central limit theorem. Basically, the box is a physical device where a number of identical small balls can be dropped into a small hole at the top, and after falling down through a number of pins, they are collected into bins at the bottom of the box. Random error leads to an approximately normal distribution of the bin counts.

In 1873, Galton began an interest in the extinction of family names, and he proposed a stochastic branching process to examine this phenomenon. The process is defined by a sequence of random variables $\{X_n | n \ge 0\}$, where $X_0 \equiv 1$, and X_n is the population size of the nth generation. Galton sets up a recursion expressing X_n in terms of the offspring of members of the population on the previous generation, namely:

$$X_n = \sum_{j=1}^{X_{n-1}} Z_j, \qquad [5]$$

where $\{Z_j | j \ge 1\}$ is a family of independent and identically distributed random variables each of which represents the number of male offspring of a married man. Soon thereafter Henry Watson (1827–1903) formalized the process and provided a solution to the probability that the population will eventually die out, and this process, known as the Galton–Watson branching process, has been employed in a number of applications in the social and biological sciences.

Of particular interest to Galton and others in the English tradition was the study of intelligence. Galton's early interest

in intelligence led to a book in 1869, *Heredity Genius*, which attempted to study the distribution of genius in families. Galton's study of intelligence was thwarted by the lack of formal tests of intelligence; however, James McKeen Cattell (1860–1944), a research fellow under Galton, first began to develop measures of intelligence based on sensory acuity tests. By the end of the nineteenth century, Alfred Binet (1857–1911) and others had begun to develop tests of intelligence along with formal scoring procedures that were based more on cognitive than sensory abilities.

Roots of Mathematical Psychology in the Twentieth Century

By the start of the twentieth century, sophisticated statistical methods for the analysis of test data were developed. The key player in this work was Charles Spearman (1863–1945), an English psychologist who received a PhD in experimental psychology from Wundt, but later became fascinated by the writings of Galton, especially on intelligence. From this interest, Spearman (1904) invented a major statistical method used to study individual differences called factor analysis. He examined the pattern of correlation coefficients among assessments of academic achievement for a group of school children and found an orderly pattern that was discernible despite statistical error that became known as Spearman's law of tetrads. Suppose X_i, $i = 1, \ldots, n$ are random variables representing the scores of an individual on n tests. Then at the population level, the tetrad law states that all distinct sets of four tests, X_i, X_j, X_k, X_l:

$$\rho(X_i, X_j)\rho(X_k, X_l) = \rho(X_i, X_l)\rho(X_k, X_j). \qquad [6]$$

The tetrad law leads to a two-factor theory of intelligence for the scores on the battery of tests, namely that the score of an individual on a particular test depends on aspects unique to the particular test as well as a general intelligence factor that belongs to the individual test taker.

Early in the twentieth century, both the German and the English approaches to psychology were combined into academic psychology in the United States, which became the main center of psychological research in the world. In this period, mathematical applications in psychology were dominated by the English approach, and as a consequence they were mostly of a statistical character. One well-known drawback of factor analysis and other methods involving correlation is the inability to draw causal conclusions from patterns of correlations. Several approaches to this problem were proposed, such as the initial development of structural equations modeling (SEM) by the geneticist Sewall Wright (1889–1988), which appeared in Wright (1921). Subsequently, SEM was formalized into a standard approach to infer causal patterns from correlational data, and it is popular in psychology today. Another important statistical development was Sir Ronald Fisher's development of analysis of variance (ANOVA) in the 1920s. ANOVA remains one of the standard statistical tools used both in experimental and differential psychology. Another example is the further development of factor analysis by Leon Thurstone (1887–1955), who invented the common factor model. The approach in Thurstone (1935) allows one to explain observed test scores by multiple latent variables, which can be estimated by statistical methods based on the singular valued decomposition of the correlation matrix.

Work of this sort led to the formal field of *psychometrics*, and most mathematically sophisticated psychologists in this period were psychometricians. The Psychometrics Society started in 1935, and the flagship journal, *Psychometrika*, started a year later. As stated on the cover of the early issues of *Psychometrika*, the mission of the Society and its journal was "devoted to the development of psychology as a quantitative rational science" (Green, 1986). Nevertheless, despite the inclusiveness suggested by this statement, most of the publications in *Psychometrika* involved statistical methods in the spirit of the work initially begun in the English tradition by Galton and his followers.

Mathematical psychology as a formal field of psychology did not emerge as a development within psychometrics, and instead, as explained in the introduction to this article, it was begun by researchers seeking new mathematical theories for experimental phenomena in five main subareas of experimental psychology listed there. However, due to lack of a natural journal outlet, some of the early papers in mathematical psychology were published in *Psychometrika*. The remainder of this section will consist of five subsections that will briefly describe several of the key papers that came to form the central areas of research in mathematical psychology. For further discussion of these and related papers, other articles in this Encyclopedia can be consulted.

Mathematical Learning Theory

As the 1950s began, experimental psychology was dominated by behavioral learning theory, so it is not surprising that much of the early work in mathematical psychology concerned mathematical models of the learning process. A typical learning experiment involves an organism learning to select appropriate responses to experimentally presented stimuli over repeated trials. This experimental structure was a natural for stochastic modeling, and two influential approaches in this spirit were developed. One was due to the psychologist William K. Estes (1919–2011), which led to the development of stimulus sampling theory; and the other approach was due to a collaboration between a physicist, Robert R. Bush (1920–72), and a statistician, Frederick Mosteller (1916–2006), which led to linear operator models. The work in these two approaches was initiated by two seminal papers that appeared in *Psychological Review*: Estes (1950) and Bush and Mosteller (1951). Both of these papers showed how one could develop stochastic models that could successfully predict the detailed statistics of the learning process in simple situations. Very shortly after these two seminal publications, mathematical learning theory became a central topic in mathematical psychology (e.g., Bush and Estes, 1959).

Even though mathematical learning theory was one of the immediate progenitors of mathematical psychology, there were intellectual predecessors to this work in the 1900s. For one, Thurstone (1930) proposed a mathematical theory of the learning curve that is remarkably modern in its foundations. Basically Thurstone regarded a learning task as consisting of

a collection of acts, some that succeed and some that fail. He develops the mathematics of a formal urn-sampling scheme, where acts are drawn at random. If an act that fails is drawn, it is eliminated with a certain probability, and so learning consists of increasing the proportion of successful acts, which is in turn a function of the model's parameters. In addition, Thurstone has equations that lead to a decrease in recurrence of unsuccessful acts and an increase in successful acts. However, unlike the work of Estes and Bush and Mosteller, Thurstone did not validate his theory with extensive statistical analysis of learning data.

Another predecessor to mathematical learning theory in the 1950s was the work of the major learning theorist, Clark Hull (1884–1952). Hull's (1943) book is full of postulates and theorems that predict the quantitative details of an organism's behavior in a learning situation. Despite the liberal use of mathematical notation in this work, it has been generally accepted that it was a failure in that it did not provide any solid measurement scales for implicit variables like drive and incentive. Indeed, it was partly as a reaction to the failed goals of Hull's work, that the work of Estes and Bush and Mosteller was initiated.

Perception and Psychophysics

In the area of psychophysics, a major advance took place in the 1950s that became a central aspect of mathematical psychology. This advance involved new mathematical models for how human observers detect signals imbedded in noise. These models were developed from ideas already worked out in the field of electrical engineering known as signal processing. The transfer of this knowledge to psychophysics occurred in two seminal papers: Tanner and Swets (1954) and Swets et al. (1955). Consider the special case of auditory signal detection. The authors discussed a simple auditory signal detection experiment in which a sample of random, 'white noise' is delivered to the subject on each of a series of discrete trials, sometimes along with a signal consisting of a pure tone of a fixed frequency. The subject's task is to respond 'yes' or 'no' on each trial depending on whether or not they think the signal tone was present. Such responses can be classified into 'hits,' correctly detecting a signal, 'false alarms,' incorrectly calling a noise trial a signal trial, 'misses,' failing to detect a signal, and 'correct rejections,' not detecting a signal on a noise trial. In the earlier work in signal processing, optimal statistical detectors had been derived based on the fact that in the physical setup, the statistical properties of signal and noise trials could be represented by probability distributions. However, in the psychophysical setting, the impact of the stimuli on the brain precluded the derivation of an optimal human detector. Nevertheless, human performance in the signal detection task could be compared with the optimal detector in a way to shed light on the underlying psychophysical processes in simple auditory perception.

Soon thereafter, signal detection theory began to be used in other, nonsensory areas of cognition. For example, Egan (1958) in an unpublished but highly cited technical report noticed that the structure of the signal detection task is identical to a number of other detection tasks that do not involve mathematically describable stimuli – for example, simple recognition memory – where the subject must detect whether or not a test item was on a previous study list. Egan's insight triggered a cottage industry in new cognitive models for recognition memory, and this continues to the present time as perhaps the largest area of formal modeling in mathematical psychology.

Choice and Decision Making

Of central importance in the area of decision making was the publication of *Theory of Games and Economic Behavior* by Von Neumann and Morgenstern (1944). The scope of the theory includes not only social interactions that involve conflicts of interest but also all situations in which individuals or groups make choices among actions or strategies whose outcomes involve uncertainty and risk. Although the work was central to economic theory, it has provided the formal framework within which psychological theories of human decision theories have been developed. A little over a decade later, a related book on game theory appeared, *Games and Decisions: Introduction and Critical Survey* (1957) by R. Duncan Luce (1925–2012) and Howard Raiffa that contributed to attracting psychologists as well as economists to the subject.

Another influential book in establishing mathematical psychology in the area of choice and decision making was *Individual Choice Behavior: A Theoretical Analysis* by Luce (1959). In it, Luce proposed the so-called choice axiom for probabilistic choice behavior, where a chooser must select one of the choice alternatives from a proffered choice set. One of the key ideas is to represent choice alternatives as members of a master set, $T = \{a_i | 1 \leq i \leq n\}$, and the axiom supposes that choice from different contexts (subsets of the master set) behaves like conditional probability behaves. In particular, if each alternative in a choice set has a positive probability of being chosen from any choice set that includes it, then the axiom states that:

$$\forall\ A \subseteq S \subseteq T,\ \ \mathbf{P}_T(A) = \mathbf{P}_S(A)\mathbf{P}_T(S), \qquad [7]$$

where, for example, $\mathbf{P}_S(A)$ is the probability that if a choice set S is presented the chosen alternative is in the set A. The axiom has many implications; for example, it is possible to show from eqn [7] that there exists a function $v(a)$ from the master choice set into the positive reals such that:

$$\forall\ a \in S \subseteq T,\ \ \mathbf{P}_S(\{a\}) = \frac{v(a)}{\sum_{b \in S} v(b)}. \qquad [8]$$

Equation [8] became known as Luce's ratio rule seen later in many applications in cognitive models of memory and categorization.

In addition, when the positive probability condition is removed from the choice axiom, allowing some alternatives to have zero probability of being chosen in certain contexts, there are many more theoretical results that can be derived from the choice axiom. In fact the book shows applications of the choice axiom in learning theory, information theory, psychophysics, and utility scaling. Of course the axiom as expressed in eqn [7] is too simple to handle complex choice situations

involving similarities among choice objects as well as context effects; however, it illustrates one of the central approaches in early mathematical psychology, namely, to capture in detail simple, nontrivial situations as a building block for more complex systems.

In the case that only choice sets of cardinality two are presented to the subject, one has what is called a paired-comparison experiment, and eqn [8] reduces to a logistic statistical model developed earlier by the statisticians Bradley and Terry (1952), and it is now called the BTL model. Even earlier than that, Thurstone (1927) developed a paired-comparison model based on Gaussian rather than logistic distributions. The BTL and Thurstone paired-comparison models became the building blocks, respectively, of logit and probit analysis used in the social sciences, especially econometrics, to analyze binary response data. Another area involving choice that became important in mathematical psychology was the effort to explain the time it takes to make simple choices. Since the 1950s, mathematical psychologists have developed a number of mathematical models for predicting the response times to perform various simple choice tasks, such as deciding which of two objects is larger, which tone has just sounded, or whether or not a string of letters is a word. On the mathematical side, this work was facilitated by earlier work in stochastic processes, such as the identity for sums of random variables developed by Abraham Wald (1902–50) and the mathematical developments of Brownian motion diffusion processes by Norbert Weiner (1894–1964) (e.g., Feller, 1957). Mathematical psychologists have used these and other mathematical results to regard a simple choice between two objects to be an accumulation of evidence for each governed by a diffusion process. This work represents a culmination of the work discussed earlier by Helmholtz and Donders concerning the timing of nervous and mental events.

Language and Thinking

In the areas of language and thinking, major progress was made possible by developments in the information sciences, in particular the formal theory of computation known as automata theory as well as the early design of digital computers. In the case of language, Chomsky (1959) developed the first formal models of grammar, and tied the theory of grammars to the computational hierarchy in automata theory. In the area of thinking, Newell et al. (1959) first published a computer model of human problem solving that was developed in the decade to follow into a major computer simulation project by Newell and Simon called the *Human Problem Solver*. These two projects received a lot of coverage in the *Handbook of Mathematical Psychology* discussed in the introduction to this article, and there were initial hopes that these areas would become basic subareas of mathematical psychology. However, as future work in these areas continued, it turned out not to be closely associated with mathematical psychology as it developed after the 1960s.

Measurement and Scaling

As mentioned earlier, before the 1950s, the development of approaches to the measurement of psychological quantities was the province of psychometricians. This work emphasized the structure of measurement error and the development of high-power statistical methods that enabled a researcher to analyze error-prone response data and retrieve its underlying structure. However, little of the psychometric work concerned the structure of the psychological quantities being measured or the cognitive factors behind the generation of the response data. There were three important new theoretical approaches to measurement and scaling in the 1950s that became strongly associated with mathematical psychology. The first was to attempt to understand at a foundational level the relationship between the structure of data and the structure of numerical systems that represent the data. This area was initiated in psychology with Suppes (1951) and Scott and Suppes (1958). These early efforts approached the measurement problem by representing the empirical situation as a relational structure with axioms that jointly reflect the measurement operations at the empirical level. Suitable axioms then permit a homomorphism from the empirical relational structure into a suitable numerical relational structure, where justified computations can be carried out. Others soon joined this effort, and it led to the area known as foundations of measurement that became one of the main research areas in mathematical psychology (e.g., Krantz et al., 1971).

The second new theoretical advance in the areas of measurement and scaling was to construct formal cognitive models of the respondent in a measurement situation rather than statistical data processing algorithms. The leader in this work was Clyde H. Coombs (1912–88). One example is Coombs (1950), where he formalized the one-dimensional unfolding model for a situation where respondents rank order their preferences for a series of items. The idea is that both the items and the respondents are simultaneously scaled on a single dimension, and each respondent ranks the objects in order of increasing distance from the placement of the respondent. Unfolding theory was extended to multiple dimensions, and it has become an often-preferred scaling model in the social and behavioral sciences. Coombs early contributions to model-based scaling methodology gave rise to his major work in this area (Coombs, 1964). A third major breakthrough in measurement and scaling in the 1950s was the development of a new form of the psychophysical scaling function along with new methods of collecting psychophysical data. As mentioned earlier, a major advance in psychology was Fechner's derivation of the logarithmic psychophysical function from Weber's law and mathematical considerations in eqn [3]. Weber used psychophysical paradigms where a subject would decide which of two stimuli was perceived to be more intense. In a major theoretical advance, Stevens (1957) changed Weber's procedure and began to ask subjects to make direct numerical estimates of the magnitude of the perceived intensities of sensory stimuli, such as weights, loudness, and time duration. In many cases, Steven's data were fit closely by the so-called Steven's power law:

$$\Psi(I) = \alpha I^{\beta}, \qquad [9]$$

where I is a physical intensity and $\alpha > 0$, $\beta \in \mathrm{Re}$ are constants that depend both on the subject and the psychophysical continuum. Several mathematical psychologists, who sought

to discover the foundational assumptions behind the law, pursued Stevens' contribution to psychophysics later. In addition, Stevens' approach was developed into a measurement technique that led to wide use in many areas of the social and behavioral sciences.

Conclusion

This article has examined some of the history of how mathematics has been applied in psychology from the seventeenth century to the mid-1950s when mathematical psychology was developed into a formal field of psychology. There have been two main themes in this history, one characterized by mathematically based theories of psychological properties of humans in general, and the other characterized by statistically based methods of measuring individual differences in psychological traits. The first is exemplified by developments in German experimental psychology in the nineteenth century in areas such as choice response time, memory, perception, and psychophysics; and the second is due to developments at approximately the same time in English differential psychology in areas such as assessing individual differences in intelligence, attitudes, and personality traits. It is fair to say that the research areas of the field of mathematical psychology come mostly from the experimental approach, and the mathematical concepts that are used are usually drawn from mathematical developments in modern mathematics and information science, mostly developed in the first half of the twentieth century. The usual approach in early mathematical psychology was to develop simple, clear theoretical ideas that could handle empirical observations in simple experiments. These mathematical approaches allowed one to state their theory precisely with testable assumptions. This approach still characterizes the field today, and almost every area of theoretical psychology has by now been positively affected by the work of mathematical psychologists.

See also: Axiomatic Theories; Bayesian Theory, History, Applications, and Contemporary Directions; Diffusion and Random Walk Processes; Discrete-State Models of Information Processing; Game Theory; Information Theory; Learning: Mathematical Learning Theory, History of; Markov Decision Processes; Memory Models: Quantitative; Psychometrics; Psychophysical Theory and Laws, History of; Signal Detection Theory, History of; Stochastic Models.

Bibliography

Bayes, T., Price, M., 1763. An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London 53, 370–418.
Boudewijnse, G.J., Murray, D.J., Bandomir, C.A., 1999. The fate of Herbart's mathematical psychology. History of Psychology 4, 107–132.
Bradley, R.A., Terry, M.F., 1952. Rank analysis of incomplete block designs, 1. The method of paired-comparisons. Biometrika 39, 324–345.
Bush, R.R., Estes, W.K. (Eds.), 1959. Studies in Mathematical Learning Theory. Stanford University Press, Stanford.
Bush, R.R., Mosteller, F., 1951. A mathematical model for simple learning. Psychological Review 58, 413–423.
Chomsky, N., 1959. On certain formal properties of grammar. Information and Control 2, 137–167.
Coombs, C.H., 1950. Psychological scaling without a unit of measurement. Psychological Review 57, 145–158.
Coombs, C.H., 1964. A Theory of Data. Wiley, New York.
Cronbach, L.J., 1957. The two disciplines of scientific psychology. American Psychologist 12, 671–684.
Egan, J.P., 1958. Recognition Memory and the Operating Characteristics (Tech. Note AFCRC-TN-58-51). Hearing and Communication Laboratory, Indiana University, Bloomington, IN.
Estes, W.K., 1950. Toward a statistical theory of learning. Psychological Review 57, 94–107.
Feller, W., 1957. An Introduction to Probability Theory and Its Application, vol. 1. Wiley, New York.
Green, B.F., 1986. Models, computers and policies. Psychometrika 51, 65–68.
Hergenhahn, B.R., 2009. An Introduction to the History of Psychology, sixth ed. Wadsworth, Belmont, CA.
Hintzman, D.L., 2003. Robert Hooke's model of memory. Psychometric Bulletin and Review 10, 3–14.
Hull, C., 1943. Principles of Behavior. Appleton-Century-Crofts, New York.
Krantz, D., Luce, R.D., Suppes, P., Tversky, A., 1971. Foundations of measurement. In: Additive and Polynomial Representations, vol. 1. Academic Press, New York.
Luce, R.D., 1959. Individual Choice Behavior: A Theoretical Analysis. Wiley, New York.
Luce, R.D., Bush, R.R., Galanter, E.G. (Eds.), 1963a. Readings in Mathematical Psychology, vol. I. Wiley, New York.
Luce, R.D., Bush, R.R., Galanter, E.G. (Eds.), 1963b. Handbook of Mathematical Psychology, vol. I. Wiley, New York.
Luce, R.D., Bush, R.R., Galanter, E.G. (Eds.), 1963c. Handbook of Mathematical Psychology, vol. II. Wiley, New York.
Luce, R.D., Bush, R.R., Galanter, E.G. (Eds.), 1965a. Readings in Mathematical Psychology, vol. II. Wiley, New York.
Luce, R.D., Bush, R.R., Galanter, E.G. (Eds.), 1965b. Handbook of Mathematical Psychology, vol. III. Wiley, New York.
Luce, R.D., Raiffa, H., 1957. Games and Decisions: Introduction and Critical Survey. Wiley, New York.
Martin, R., 2011. The St. Petersburg paradox. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/archives/win2011/entries/paradox-stpetersburg/.
Newell, A., Shaw, J.C., Simon, H.A., 1959. Report on a general problem-solving program. In: Proceedings of the International Conference on Information Processing, pp. 156–264.
Scott, D., Suppes, P., 1958. Foundational aspects of theories of measurement. Journal of Symbolic Logic 23, 113–128.
Spearman, C.E., 1904. 'General intelligence' objectively determined and measured. American Journal of Psychology 15, 72–101.
Stevens, S.S., 1957. On the psychophysical law. Psychological Review 64, 153–181.
Stigler, S.M., 1986. The History of Statistics: The Measurement of Uncertainty before 1900. Harvard University Press, Cambridge, MA.
Sturm, T., 2006. Is there a problem with mathematical psychology in the eighteenth century? A fresh look at Kant's old argument. Journal of the History of Behavioral Science 42, 353–377.
Suppes, P., 1951. A set of independent axioms for extensive quantities. Portugaliae Mathematica 10, 163–172.
Swets, J.A., Tanner Jr., W.P., Birdsall, T.G., 1955. Decision processes in perception. Psychological Review 68, 301–340.
Tanner Jr., W.P., Swets, J.A., 1954. A decision-making theory of visual detection. Psychological Review 61, 401–409.
Thurstone, L.L., 1927. A law of comparative judgment. Psychological Review 34, 273–286.
Thurstone, L.L., 1930. The learning function. Journal of General Psychology 3, 469–493.
Thurstone, L.L., 1935. The Vector of Mind. University of Chicago press, Chicago, IL.
Von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ.
Wright, S., 1921. Correlation and causation. Journal of Agricultural Research 20, 557–585.

Mathematics Education

Lieven Verschaffel, Center for Instructional Psychology and Technology, Leuven, Belgium
Fien Depaepe, Center for Educational Policy, Innovation and Teacher Training, Leuven, Belgium
Erik De Corte, Center for Instructional Psychology and Technology, Leuven, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article first describes the nature of mathematics as an intellectual domain and how several general theoretical perspectives in educational psychology as well as mathematics-specific theories have influenced and still influence research on mathematics education. Subsequently, we review the current research-based views on (1) mathematical competence, (2) learning mathematics, and (3) teaching mathematics.

During the past decades mathematics education has evolved from a terrain operated on by various groups of scholars – psychologists, mathematicians, and educational scientists – with very different theoretical and methodological backgrounds, research goals, and with little or no mutual interaction in an interdisciplinary field of study in which not only the scholars with above-mentioned disciplinary backgrounds, but also historians, philosophers, sociologists, anthropologists, and, more recently, cognitive and educational neuroscientists, actively cooperate (De Corte and Verschaffel, 2006; Sriraman and English, 2010). Thus, to a considerable extent, the past decades have witnessed the emergence of research in mathematics education as a field of study in its own right, even as it maintains strong symbiotic links with other disciplines. Of these other disciplines, psychology and mathematics themselves remained, of course, preeminent, but, increasingly, an important role is being played by the other above-mentioned disciplines, reflecting the increasing recognition of the 'situatedness' of mathematical thinking and learning in social, cultural, and historical contexts (De Corte and Verschaffel, 2006) as well as its physical 'embodiment' (Lakoff and Núñez, 1997).

In the first section we briefly reflect on mathematics as an intellectual domain. Thereafter, we introduce several general as well as domain-specific theoretical perspectives that have historically shaped today's thinking about mathematics education. In the main body of the article, we review the current research-based view on three major aspects of mathematics education: (1) mathematical competence, (2) learning mathematics, and (3) teaching mathematics.

For a more complete overview of these topics, we refer to the following recent volumes in which the vast body of theoretical reflection and empirical research in mathematics education has been extensively reviewed: *Mathematics Education as a Research Domain* (Sierpinska and Kilpatrick, 1999), *Handbook of Research on the Psychology of Mathematics Education* (Gutteriez and Boero, 2006), *Second Handbook of Research on Mathematics Teaching and Learning* (Lester, 2007), and *Theories of Mathematics Education* (Sriraman and English, 2010).

Mathematics

When people refer to mathematics, they mostly refer to formal academic or professional mathematics – that is, the scholarly study of quantity, space, and structure, by seeking out patterns, formulating new conjectures, and establishing truth by rigorous deduction from appropriately chosen axioms and definitions, or the simplified version of it they learnt at school. But mathematics also involves the more informal kinds of human practices related to number, space, and patterns that are present in human activities like counting, locating, measuring, designing, playing, and explaining.

The answer to the question 'What is mathematics?' has changed throughout history and will continue to change. Very broadly, the answer during the twentieth century has evolved from 'the science of number and space' to a more abstract characterization such as 'the science of patterns.' However, others would give more weight to 'problem solving' or the 'modeling' aspects of the real world, and the quest for a single phrase that could encapsulate an intellectual enterprise with so many facets seems pointless. Two other elements of development in the conception of what mathematics essentially is, are: (1) the shift away from the view of mathematics as the epitome of 'certain knowledge' and toward a conceptualization as an activity grounded in human practices and (2) the increasing use of computers in mathematical research, which has resulted in a greater emphasis on mathematics as an experimental science and forced reconsideration of the nature of proof.

Researchers attempting to describe what makes mathematics significantly different from other intellectual or scientific domains point to its 'dual nature': mathematics as a descriptor of a perceived reality, but at the same time, as an autonomous abstract construction. This duality is acknowledged in the distinction between applied mathematics (the branch of mathematics concerned with application of mathematical knowledge to other fields such as natural science, engineering, medicine, and social sciences) and pure mathematics (mathematics for its own sake, without having any application in mind). These two different and complementary aspects of mathematics, the practical and theoretical, have earned the subject a place at the center of education throughout history. Some other typical features of mathematics are: its hierarchically structured development on both the cultural and individual planes, and the intrinsic inaccessibility of mathematical (thinking) objects, implying that their conceptual acquisition necessarily passes through the acquisition of one or more semiotic representations.

Even though there is little doubt that mathematics is a unique and integral intellectual domain, it is also possible to look at it as an aggregation of subdomains that differ in terms of their most characteristic methods, styles of thinking, and forms of representation, and which therefore may need different 'local' theories of expertise, learning, instruction, and assessment. These subdomains are typically represented in leading handbooks on mathematics education as edited by Gutteriez and Boero (2006) and Lester (2007), which include chapters on arithmetic, algebra, geometry, and statistics and probability.

Theoretical Views on Mathematics Education

One way to categorize the theoretical orientations of the research and analyses of the nature of mathematical cognition, how it is learnt, and how this learning can be optimally enhanced, is by referring to the broad paradigmatic traditions within psychology and education in general: the behaviorist/empiricist, the cognitive/rationalist, and the situative/pragmatist–sociohistoric tradition (Greeno et al., 1996). These general theories have largely influenced, and continue to influence, the research field of mathematics education.

The behaviorist/empiricist tradition includes associationism, behaviorism, and connectionism. According to this view, mathematical knowledge and skills are the result of an accumulation of acquired associations and skills. A classic example of this tradition is Thorndike, who applied his associationist view on learning (as differential strengthening of associations by reinforcement) to the learning of arithmetic. Although this tradition plays a minor role in contemporary research, its ideas still permeate into current instructional practices: for example, in drill-and-practice approaches. Conceptions of mathematics as providing correct answers to well-defined mathematics tasks, of learning mathematics as incremental (with errors to be avoided or immediately stamped out), and of mathematics teaching as the reinforcement of mathematically correct responses, remain prevalent in many current instructional practices and, as such, represent the legacy of this first tradition.

The cognitive/rationalist perspective involves research traditions such as Gestalt psychology, symbolic information processing, and constructivism. These traditions conceive mathematical knowledge as universal mental entities located in individual learners, such as cognitive schemes and procedural rules, and define learning mathematics as changes in these universal mental schemes and rules. Typical representatives of this tradition are Piaget, who theorized mathematical development in terms of the stepwise acquisition of increasingly complex logicomathematical structures; Wertheimer, who emphasized the insightful detection and exploitation of structure in mathematical thinking and learning; and information-processing theorists such as Simon, Anderson, and Siegler, who performed fine-grained analyses of the cognitive structures and processes involved in the acquisition and use of key mathematical concepts and skills. This perspective has also had a clear impact on the practice of mathematics education: performing detailed analyses of the concepts and skills needed to perform certain mathematical tasks, looking at mathematical learning processes rather than learning outcomes, paying ample attention to the role of prior knowledge, valuing conceptual understanding, and problem solving besides procedural fluency are reminiscent of this second perspective.

The situative/pragmatist–sociohistoric perspective entails traditions such as ethnography, anthropology, and situation theory. This third framework embraces all those theories that view mathematics learning as a reorganization of activity that accompanies the integration of an individual learner within a 'community of practice.' Having its origins in Vygotsky's work, this third tradition emerged strongly in the late 1980s in reaction to the then dominant cognitive view of mathematics learning as a highly individual and purely mentalistic process of knowledge and skill acquisition occurring in the learner. In contrast to this cognitivist view, the situated perspective stresses that mathematics learning is enacted essentially in interaction with social and cultural contexts and artifacts, and especially through participation in cultural activities and contexts. Whereas in the cognitivist view, mathematics is seen as a universal knowledge system, the situative/pragmatist–sociohistoric perspective assumes that mathematics differs according to the setting in which it has been developed and in which it is practiced. The influence of this perspective on mathematics education is reflected in attempts to implement elements of successful out-of-school learning settings into school mathematics, and increased attention to the sociocultural and affective side of mathematics learning and teaching.

However, research in mathematics education of the past decades is not merely, not even primarily, theoretically framed in terms of the three above-mentioned general traditions. Of the many theories that have been developed over the past decades and are being used within the field of mathematics education, the dominant ones are 'home-grown theories.' Such theories put the specificity and integrity of the domain at the center, with no hesitation, whatsoever, to borrow ideas and techniques from other disciplines. But they have little or no interest in the possible relevance of the theory outside the field of mathematics education. Therefore, since its emergence as a field in its own right, mathematics education has witnessed many such home-grown theories of mathematical cognition, learning, and teaching, such as Brownell's (1945) theory of what he called 'meaningful mathematics,' emphasizing the understanding of mathematical relationships and the ability to think quantitatively, Skemp's (1979) theoretical distinction between 'instrumental' and 'relational understanding' of mathematics, Van Hiele's (1986) theory of how children learn geometry, Pólya's (1945) analysis of the role of heuristic processes in mathematical problem solving, Freudenthal's (1983) view on mathematics (education) as a human activity, which formed the basis of the Dutch model of 'realistic mathematics education,' and Fischbein's (1987) theory of the intuitive sources of mathematical thinking – to mention just few 'giants' of the history of research in mathematics education. Although well versed in the theories and methods of psychology, most of them were quite critical of its limited application to the domain of mathematical cognition, learning, and teaching, and emphasized that (the psychology of) mathematics education had to develop its own questions, and theoretical and methodological perspectives (Verschaffel and Greer, 2013).

In the course of development, tensions and perspectival differences between mathematics educators on the one hand

and psychologists, educational scientists, and mathematicians on the other hand, have continued. The books edited by Sierpinska and Kilpatrick (1999) and Sriraman and English (2010), for instance, make clear that the field has far from settled into a phase of 'normal science' (in the Kuhnian sense). But rather it has entered a period of considerable complexity and diversity in theoretical perspectives, experimental methodologies, and reflections, not only on the fundamental questions of what mathematics is and how it can be learnt and taught, but also of the equally fundamental question what is mathematics (education) education *for*?

Mathematical Competence

While for a long time it was quite common to think of mathematical competence as the mastery of specific knowledge and skills, it is now generally agreed that it involves a broad set of competencies including (De Corte and Verschaffel, 2006):

- A well-organized and flexibly accessible domain-specific knowledge base involving facts, symbols, conventions, definitions, formulas, algorithms, concepts, and rules that constitute the contents of mathematics as a subject-matter field.
- Heuristic methods, that is, search strategies for problem solving that do not guarantee but significantly increase the probability of finding the correct solution because they induce a systematic approach to the task (e.g., thinking of an analogous problem, decomposing a problem into subgoals, and visualizing the problem using a diagram or a drawing).
- Metaknowledge, which involves knowledge about one's cognitive functioning (metacognitive knowledge, e.g., believing that one's cognitive potential can be developed and improved through learning and effort) and knowledge about one's motivation and emotions that can be used to deliberately improve volitional efficiency (e.g., becoming aware of one's fear or failure when confronted with a complex mathematical task or problem).
- Self-regulatory skills, which embrace skills relating to the self-regulation of one's cognitive processes (metacognitive skills or cognitive self-regulation; e.g., planning and monitoring one's problem-solving processes) and skills for regulating one's volitional processes/activities (metavolitional skills or volitional self-regulation; e.g., keeping up one's attention and motivation to solve a given problem).
- Positive beliefs about oneself in relation to mathematical learning and problem solving (self-efficacy beliefs), about the social context in which mathematical activities take place, and about mathematics and mathematical learning and problem solving.

This broad conception of mathematical competence is nicely reflected in the characterization of 'mathematical proficiency' in the report of the U.S. National Research Council (2001), *Adding it up*, wherein proficiency is defined in terms of five strongly interwoven strands, namely:

- conceptual understanding, comprehension of mathematical concepts, operations, and relations;
- procedural fluency, skill in carrying out procedures flexibly, accurately, efficiently, and appropriately;
- strategic competence, the ability to formulate, represent, and solve mathematical problems;
- adaptive reasoning, the capacity for logical thought reflection, explanation, and justification; and
- productive disposition, a habitual inclination to see mathematics as sensible, useful, and worthwhile discipline, coupled with a belief in diligence and one's efficacy.

It is certainly useful to distinguish these different categories of components, but it is also important to realize that in expert mathematical cognition they are applied integratively and interactively. Actually, research in mathematical subdomains, such as whole number arithmetic, algebra, geometry, and statistics, contains numerous illustrations of the importance and interdependency of these different strands of proficiency (see, e.g., Gutteriez and Boero, 2006; Lester, 2007).

Learning Mathematics

In line with the above broad conception of mathematical competence, it is now quite generally agreed that mathematics learning is not a one-dimensional phenomenon; as it proceeds, each strand of mathematical proficiency develops in interaction with the others. Moreover, research over the past decades has resulted in the view of mathematics learning as active and cumulative construction of mathematical proficiency – constrained by the structural and functional properties of the human cognitive architecture, and by the form of the human body more generally (i.e., embodied cognition), and situated in a given community of learners and a broader sociocultural context.

The view of learning as an active and constructive process is now the common ground among mathematics educators and is supported by substantial empirical evidence. Essentially, mathematical competence is seen as the result of the mindful and effortful involvement of the learner in the processes of knowledge and skill acquisition in interaction with the environment, rather than the result of direct information transmission. This involvement is well illustrated in children's invented idiosyncratic but accurate procedures for solving arithmetic problems in school as well as out-of-school contexts. For instance, some children seem to find out by themselves that subtraction problems of the type $82 - 78 = \ldots$ can be efficiently solved by a subtraction-by-addition strategy, whereby the result is found by determining how much should be added to the smaller number to arrive at the larger one. However, the constructive nature of learning is also expressed in a negative way in misconceptions and in defective or 'buggy' procedures that many learners acquire. Examples of misconceptions are: "multiplication always makes bigger whereas division always makes smaller" and "every number has a number immediately preceding and following it." Illustrations of defective procedures are in multidigit subtraction, subtracting in each column the smaller digit from the larger, and in addition of fractions, simply adding the nominators and the denominators, that many learners acquire.

The constructive nature of mathematics learning involves that it is also cumulative. Indeed, students build new

knowledge and skills on the basis of what they already know and can do. This pivotal role of prior knowledge can affect learning in both positive and negative ways. There is, for instance, ample evidence that children's thinking about number relies on certain core principles or presuppositions that are based on their experience with natural numbers. This facilitates some kinds of learning but inhibits others. For example, children's early understanding of natural number and its properties strongly supports their conception of multiplication as repeated addition (Fischbein, 1987). But at the same time it leads to the misconceptions, mentioned in the preceding paragraph, that multiplication always leads to a bigger result, or that every number has a number immediately preceding or following it (Vosniadou and Verschaffel, 2004). The view of learning as a constructive activity implies also its self-regulatory nature: learners have to assume control over their own learning, thinking, and problem-solving activities (De Corte et al., 2010).

Furthermore, mathematics learning is constrained by the human cognitive architecture and functioning, that is, the manner in which our cognitive structures are organized in memory and our information-processing system works. Most treatments of the human cognitive architecture use the sensory memory/working memory/long-term memory model as their base. It is generally accepted that the characteristics of working memory and long-term memory and their functional relations (e.g., the fact that, when processing novel information, working memory is very limited in duration and in capacity) determine the quantity and quality of what will actually be learnt. For instance, it is well known that low working memory is a barrier to the learning of both simple calculations and higher level problem solving. It has also been shown that children with mathematics disability have memory deficits that are directly responsible for their disability. These memory problems prevent them from memorizing and/or recalling arithmetic facts, which leads to difficulties executing calculation procedures, or create problems with the temporary storage of numbers in working memory, which is detrimental for doing complex mental arithmetic or problem solving.

It is also increasingly accepted that (mathematical) cognition is shaped by aspects of the body, such as the motor system, the perceptual system, and how the body is functioning in the environment. This theoretical perspective of 'embodied cognition,' which emphasizes that (mathematical) cognition is situated and made comprehensible by the shared biology and fundamental bodily experiences of human beings, entails a reconceptualization of cognition and mathematics itself, and may also have important consequences for mathematics education. According to Lakoff and Núñez (1997), abstract concepts, such as 'balancing' a system of simultaneous equations, are conceptual extensions of the image schemata involved in the bodily experience of 'balance.' As Lakoff and Núñez (1997: p. 30) state, "much of what is 'abstract' in mathematics (…) concerns coordination of meanings and sense making based on common image-schemata and forms of metaphorical thought. Abstract reasoning and cognition are thus genuine embodied processes." At a more basic level, various studies have pointed to finger counting as prime example of embodied cognition.

Finally, mathematics learning – and mathematics as such – are not only situated in a person's body, but also in the broader sociocultural environment, which means that it is enacted essentially within social and cultural contexts and in interaction with other people and artifacts, especially through participation in sociocultural activities and communities. Thus, mathematics educators have now broadly embraced the view that mathematics learning is not a purely solitaire and mental activity; rather thinking and learning processes are distributed over the individual learner, his/her partners in the environment, and available representational tools (e.g., abacus) and technological resources (e.g., computer and interactive whiteboard). The outcomes of a large series of ethno-mathematical studies of the informal calculation procedures, problem-solving strategies, and learning mechanisms of particular groups of adults involved in everyday cultural practices, such as tailoring, weaving, carpentry, navigating, grocery, packing, and cooking, has largely contributed to the popularity of this viewpoint. Although the situated nature of doing and learning mathematics has been documented especially well within such out-of-school contexts, it is obvious that this situatedness also applies to mathematics learning *in school*. For instance, the numerous examples of students' 'suspension of sense-making' when doing school word problems (leading to computationally based solutions to absurd problems such as in "There are 26 sheep and 10 goats on a ship. How old is the captain?" Answer: $26 - 10 = 16$) can be considered as evidence for the importance of the situatedness of mathematical cognition and learning (Verschaffel et al., 2000). From an educational perspective, the situated view particularly implies the importance of interaction and collaboration in mathematics learning as well as of the importance of a proper understanding of the role of tools in mathematical thinking and learning.

Teaching Mathematics

For a long time, mathematical learners worldwide received school-based instruction according to a 'mechanistic' or 'skills approach' – an instructional approach that is based on the view that mathematics learning is a highly individual activity consisting mainly of absorbing and memorizing a fixed body of fragmented and decontextualized knowledge and procedural skills directly transmitted by the teacher. Baroody (2003: p. 17) describes it as follows:

> The skills approach, with its focus on memorization of basic skills by rote (…), is based on the assumption that mathematical knowledge is simply a collection of socially useful information (facts, rules, formulas, and procedures). Except for the mathematically gifted, children are viewed as "empty vessels," largely helpless, and perhaps incapable of understanding many aspects of school mathematics. The aim of math instruction is to inform them about how to do mathematics. Proponents of the skills approach (e.g., the "back-to-basics" movement) believe that the most efficient way to transmit information is by means of direct instruction and practice – without taking the time to promote understanding about the whys of procedures. The teacher directs all aspects of the learning process to ensure efficient instruction (transmittal and mastery) of basic skills.

In the skills approach, a teacher who wants his or her pupils to master the algorithm for long division demonstrates the

subsequent steps of the algorithm on the blackboard several times, and then instructs the pupils to copy these steps on long lists of similar problems and provides corrective feedback until the pupils master this algorithmic procedure, without explaining where the algorithm comes from, why it works, and how to choose between this algorithm and other computational methods (such as mental arithmetic or estimation). Against the background of the above-mentioned view that mathematical proficiency comprises the integrated mastery of various kinds of competencies and that mathematics learning is an active, cumulative, and situated process of knowledge building and skill acquisition, many mathematics educators have criticized this traditional 'mechanistic' or 'skills approach' and proposed, implemented, and tested reform-based instructional approaches, which aim at the mastery of all aspects of mathematical proficiency and which are characterized by more conceptually-, inquiry- and/or problem-based forms of learning.

For example, Lampert (1990) has designed a learning environment that aims at promoting meaning construction and understanding of multiplication in fourth graders by connecting and integrating principled conceptual knowledge (e.g., the principles of additive and multiplicative composition, associativity, commutativity, and the distributive property of multiplication over addition) with their computational skills. She starts from familiar and realistic problems to allow children to use and explore their informal prior knowledge, and practices collaborative instruction whereby she engages in cooperative work and discussion with the whole class. Students are solicited to propose and invent alternative solutions to problems that are then discussed, including their explanation and justification. It is obvious that this learning environment also embodies a classroom climate and culture that differs fundamentally from what is typical of traditional mathematics lessons.

A second and more comprehensive example is Realistic Mathematics Education (RME) developed in the Netherlands. RME was the Dutch answer to the internationally felt need, in the late 1960s and early 1970s, to reform mathematics education. It was a reaction to both the then prevailing 'mechanistic' or 'skills approach' to mathematics education, and the structuralistic 'New Math' movement that was expected to dominate the (Western) math educational world in those days (De Corte and Verschaffel, 2006; Treffers, 1987). RME conceives mathematics learning essentially as doing mathematics starting from the study of phenomena in the real world as topics for mathematical modeling, and resulting in the reinvention of mathematical knowledge. Based on this fundamental conception of doing mathematics, the design of 'realistic' learning environments is guided by a set of five interrelated principles: (1) learning mathematics is a constructive activity, (2) progressing toward higher levels of abstraction, (3) encouraging students' free production and reflection, (4) learning through social interaction and cooperation, and (5) interconnecting knowledge components and skills (Treffers, 1987). Of course, these general instructional design principles of RME have to be refined and adapted to specific mathematical topics and educational contexts, before they become operational and can be put to an empirical test. Typical examples of such local RME theories have been developed for the teaching of arithmetic algorithms, fractions, and basic statistics, and have been implemented and tested in the context of 'design experiments' (Wittmann, 1995), which have reported results that are, in general, quite positive for the RME approach.

A third example is Schoenfeld's (1985) instructional approach for teaching problem solving to college students. Starting from the idea (developed on the basis of several detailed observational studies) that using heuristics in conjunction with metacognitive executive control skills and appropriate beliefs is an essential feature of expert problem solving, Schoenfeld developed a teaching approach that focuses on the heuristic, metacognitive, and affective aspects of mathematical problem solving. Besides extensive use of modeling by the teacher to demonstrate how an expert applies these various elements of mathematical expertise, and whole class discussions, ample use is made of small-group problem solving, whereby the teacher acts as a consultant, regularly asking questions such as: what are you doing, why are you doing this, and how will it help to find the final solution? The results of this instructional approach were very positive, in the sense that both the problem-solving performance and processes of students who had followed his course became more expert-like.

During the past decades, similar efforts to reform mathematics education, from elementary school to university, have been undertaken in many other countries, all of which embody to some degree the above-mentioned ideas about what mathematics essentially is and how it can be optimally learnt and taught, such as, the conception of mathematics as human activity, the constructivist view of learning, the crucial role of students' prior – informal as well as formal – knowledge, the orientation toward understanding and problem solving, the importance of social interaction and collaboration in doing and learning mathematics, the need to embed mathematics learning into realistic contexts, and to create a genuine mathematics classroom culture. Generally speaking, the researchers of these studies report promising results that demonstrate that these reform-based learning environments, if properly implemented, can lead to fundamental changes in the way children acquire mathematical knowledge, skills, and beliefs, and make them more autonomous and more motivated mathematical learners and problem solvers.

However, the evaluation of these reform-based frameworks, approaches, and materials is not univocally positive. Critics have argued that convincing empirical evidence in favor of this reform-based approach is lacking (because of methodological shortcomings in these design experiments), that there are indications that gains in certain domains and aspects of mathematics learning are paralleled with losses in other domains, and that the wider implementation of the envisioned reform-based ideas and materials and methods would require competencies from teachers – mathematical as well as didactical – that go far beyond what they currently possess (Baroody, 2003; KNAW, 2009).

During the past decade, these attacks against these reform-based views on and curricula and textbooks for mathematics education have been most obvious in the United States, under the (unfortunate) metaphor of Math Wars (Schoenfeld, 2004), also showing recent signs of export to other countries. Interestingly, the debate about the content and methods of mathematics education is accompanied by a deep disagreement

about how the merits of different approaches to mathematics education should be assessed. The National Mathematics Advisory Panel Report (United States Department of Education, 2008), for instance, almost completely eliminated mathematics education research from its consideration of 'what works' by adopting rigid methodological criteria that largely ruled out studies other than large-scale randomized controlled trials. In various sharp reactions the community of mathematics educators (e.g., Cobb and Jackson, 2008; Greer, 2012) objected to this attempt to conform mathematics educators to problematically narrow criteria for rigorous designs, reliable and valid assessment instruments, and significant results (Sriraman and English, 2010).

However, most mathematics educators by no means reject experimental and quantitatively oriented methodologies, as long as they are combined and balanced with other methods that do respect the complex and multi-dimensional nature of mathematical thinking and learning, and of the educational and broader cultural environments in which it occurs (Cobb and Jackson, 2008). Moreover, it would be naive to expect that empirical research alone – of whatever type it is – will yield a clear and conclusive answer as to what kind of instructional approach is most valuable for teaching mathematics, not only because of the complexity of the field or the diversity of its subdomains, but also because of the serious differences in beliefs about and valuations of what mathematics is after all and why it is important for learners and for people and society in general (Greer, 2012).

See also: Education (Primary and Secondary Schools) and Gender; Environments for Education; Instructional Psychology; Science Education; Spatial Cognition.

Bibliography

Baroody, A.J., 2003. The development of adaptive expertise and flexibility: the integration of conceptual and procedural knowledge. In: Baroody, A.J., Dowker, A. (Eds.), The Development of Arithmetic Concepts and Skills. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 1–34.

Brownell, W.A., 1945. When is arithmetic meaningful? Journal of Educational Research 38, 481–498.

Cobb, P., Jackson, K., 2008. The consequences of experimentalism in formulating recommendations for policy and practice in mathematics education. Educational Researcher 37, 573–581.

De Corte, E., Mason, L., Depaepe, F., Verschaffel, L., 2010. Self-regulation of mathematical knowledge and skills. In: Zimmerman, B., Schunk, D. (Eds.), Handbook of Self-regulation of Learning and Performance. Routledge/Taylor & Francis, New York, pp. 155–172.

De Corte, E., Verschaffel, L., 2006. Mathematical thinking and learning. In: Damon, W., Lerner, R., Sigel, I., Renninger, A. (Eds.), Handbook of Child Psychology, Child psychology in practice, vol. 4. Wiley, New York, pp. 103–152.

Fischbein, E., 1987. Intuition in Science and Mathematics: An Educational Approach. Reidel, Dordrecht.

Freudenthal, H., 1983. Didactical Phenomenology of Mathematical Structures. Reidel, Dordrecht, The Netherlands.

Greeno, J.G., Collins, A.M., Resnick, L.B., 1996. Cognition and learning. In: Berliner, D.C., Calfee, R.C. (Eds.), Handbook of Educational Psychology. Macmillan, New York, pp. 15–46.

Greer, B., 2012. The USA mathematics advisory panel. In: Skovsmose, O., Greer, B. (Eds.), Opening the Cage: Critique and Politics of Mathematics Education. Sense Publishers, Rotterdam, The Netherlands, pp. 107–124.

Gutteriez, A., Boero, P. (Eds.), 2006. Handbook of Research on the Psychology of Mathematics Education. Past, Present, and Future. Sense Publishers, Rotterdam.

Koninklijke Nederlandse Akademie van Wetenschappen, 2009. Rekenonderwijs op de basisschool. Analyse en sleutels tot verbetering [Mathematics Education in the Elementary School. Analysis and Keys for Improvement]. Koninklijke Nederlandse Akademie van Wetenschappen, Amsterdam.

Lakoff, G., Nuñez, R., 1997. Where Mathematics Comes from: How the Embodied Mind Brings Mathematics into Being. Basic Books, New York.

Lampert, M., 1990. Knowing, doing, and teaching multiplication. Cognition and Instruction 3, 305–342.

Lester Jr., F.K., 2007. Second Handbook of Research on Mathematics Teaching and Learning. Information Age Publishing, Greenwich, CT.

National Research Council, 2001. Adding it up: Helping Children Learn Mathematics. National Academy Press, Washington, DC.

Pólya, G., 1945. How to Solve it. Princeton University Press, Princeton, NJ.

Schoenfeld, A.H., 1985. Mathematical Problem Solving. Academic Press, New York.

Schoenfeld, A.H., 2004. The math wars. Educational Policy 18, 253–286.

Sierpinska, A., Kilpatrick, J. (Eds.), 1999. Mathematics Education as a Research Domain: A Search for Identity. Kluwer, Dordrecht, The Netherlands.

Skemp, R.R., 1979. Intelligence, Learning, and Action. Wiley, Chichester, England.

Sriraman, B., English, L. (Eds.), 2010. Theories of Mathematics Education. (Advances in Mathematics Education Series.). Springer, Berlin.

Steffe, L.P., Nesher, P., Cobb, P., Goldin, G.A., Greer, B. (Eds.), 1996. Theories of Mathematical Learning. Erlbaum, Mahwah, NJ.

Treffers, A., 1987. Three Dimensions. A Model of Goal and Theory Description in Mathematics Instruction. The Wiskobas Project. D. Reidel Publishing, Dordrecht, The Netherlands.

United States Department of Education, 2008. Foundations for success. The final report of the National Mathematics Advisory Panel. Retrieved 17.01.09 from: http://www.ed.gov/about/bdscomm/list/mathpanel/index.html.

Van Hiele, P.M., 1986. Structure and Insight: A Theory of Mathematics Education. Academic Press, Orlando, FL.

Verschaffel, L., Greer, B., 2013. Domain-specific strategies and models: mathematics education. In: Spector, J.M., Merrill, M.D., Elen, J., Bishop, M.J. (Eds.), Handbook of Research on Educational Communications and Technology, fourth ed. Springer Academic, New York, pp. 553–563.

Verschaffel, L., Greer, B., De Corte, E., 2000. Making Sense of Word Problems. Swets and Zeitlinger, Lisse, The Netherlands.

Vosniadou, S., Verschaffel, L., 2004. Extending the conceptual change approach to mathematics learning and teaching. Learning and Instruction 14, 445–451.

Wittmann, E. Ch, 1995. Mathematics education as a design science. Educational Studies in Mathematics 29, 355–374.

Relevant Websites

http://www.fi.uu.nl/fisme/en/ – Freudenthal Institute.
http://www.mathunion.org/ICMI/ – International Commission on Mathematics Instruction.
http://www.nctm.org/ – National Council of Teachers of Mathematics.

Matrifocality

Raymond T Smith, University of Chicago, Chicago, IL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9416–9418, © 2001, Elsevier Ltd.

Abstract

The term 'matrifocal family' was introduced in the 1950s to analyze the process through which Afro–Guyanese women, in their role as mothers, became the focus of affective ties while men were increasingly marginal to the complex of domestic relations, but not necessarily absent from them. Its use spread, first in the Caribbean and then globally, though it has too often been confused with female-headed households. Worldwide changes in the status of women, the nature of conjugal relations, patterns of child-rearing, and the configuration of household structures, suggest the continuing relevance of the term 'matrifocality' as an analytical concept taking note of the relation among class, status, gender, conjugal, and domestic relations. The term has achieved a much wider popular currency and is now used to refer to almost any situation in which women are prominent.

Introduction: Definition and Derivation of the Term

Although there is no generally agreed upon definition of matrifocality, and indeed the term has been applied so variously that its meaning is quite vague, for the purposes of this article it is defined as follows: 'matrifocality is a property of kinship systems where the complex of affective ties among mother and children assumes a structural prominence because of the diminution (but not disappearance) of male authority in domestic relations.'

The term 'matrifocal family' was introduced to solve a problem of empirical description as well as to suggest an alternative theoretical view of the domestic organization of rural African-Americans in the then British Guiana (Smith, 1956). After considering (and rejecting) such terms as 'matricentral' and 'matriarchal' to refer to the internal relations among members of households, many of which were female headed (but certainly not all), the term 'matrifocal' was used as follows:

> The household group tends to be matri-focal in the sense that a woman in the status of 'mother' is usually the *de facto* leader of the group, and conversely the husband-father, although *de jure* head of the household group (if present), is usually marginal to the complex of internal relationships of the group. By 'marginal' we mean that he associates relatively infrequently with the other members of the group, and is on the fringe of the effective ties which bind the group together. (Smith, 1956)

In the last sentence the words "effective ties" is an undetected printer's error; in the manuscript it reads "affective ties" which was, and is, the author's intention.

The term was quickly taken up and widely used in Caribbean studies though not consistently or without debate; it frequently was confused with 'female-headed.' Since the 1950s both the use and the meaning of matrifocality has been extended beyond the Caribbean as it tracked the worldwide revolution in gender relations, in marriage patterns, and in domestic relations. Its appeal as a less precise, but more evocative, expression has value in pointing to the changes in kinship practices and gender relations in the context of broader socioeconomic and political changes, but for the analysis of those changes something closer to its original meaning is necessary.

The Term 'Matrifocal' and the Context of Its Introduction

Beginning in the 1930s an upsurge of labor unrest and nationalism in the Caribbean colonies led to policies of 'development and welfare.' Family life became a central issue in those policies, on the assumption that anything differing from middle-class European family ideals was unsuited to the modern world. Sociology had developed quite complex theories purporting to show that the isolated nuclear family of man, wife, and children was not only the most highly evolved, but also the most adapted to modern industrial society with its complex division of labor and high degree of social mobility based on male occupations. These ideas converged with the assumption – the false assumption – that a stable nuclear family based upon legal marriage is natural, being found in all human societies (*see* Family: The Anthropology of the Concept and its History). Given these assumptions, households containing women with children by several fathers, unstable conjugal relations, conjugal partners living together without being legally married, were all considered to be deviant, abnormal, or pathological.

More technical anthropological analysis had shown that counting descent through females did not alter male dominance; however, matrilineal descent, that is, the inheritance or transmission of power and economic resources from a man to his sister's sons (rather than to his own sons), could lead to unstable marriage and complicated domestic organization when the contrary pulls of marriage and parenthood on the one hand, and loyalty to the lineage on the other, precipitated an array of varying domestic arrangements (see Fortes, 1949).

The analysis of matrilineal systems was important for countering the ethnocentric assumptions of nuclear family universality and paternal authority patterns; unfortunately, it had little influence on the emerging discussions of matrifocality. A widely cited publication was a paper entitled *A Survey of the Consanguine or Matrifocal Family* in which the author defined the matrifocal family as 'a co-residential kinship group which

includes no regularly present male in the role of husband-father' (Kunstadter, 1963). Given this definition, it is not surprising that the author classifies extreme matrilineal cases where women and children live with, and under the jurisdiction of, the woman's brother, as 'matrifocal,' but it also excludes cases where a husband father is actually present, and abandons the criterion specified by Smith (1956) who later stated that, the term 'matrifocal' – specifically intended to convey that it is women in their role as mothers who come to be the focus of relationships, rather than head of the household as such – the nuclear family is both ideally normal, and a real stage in the development of practically all domestic groups (Smith, 1973). Kunstadter also ignored the developmental cycle of domestic relations, pioneered by Fortes in his study of Ashanti. It was central to the original concept of matrifocality because it analyzed the changing focus of domestic authority and cohesion as children are born and mature. Kunstadter's paper, because of its place of publication, its title, its simplicity, and apparent authority, became the most widely cited source on matrifocality, drastically modifying its meaning by equating it with a static definition of female-headed households.

Family Policy and the Search for the Causes of Poverty

In European and North American societies it had long been assumed that poverty and disorganized families were at once the cause and the result of each other. These assumptions were given vigorous new life, and a decided racial slant, by the declaration in 1965 of a 'War on Poverty' in the US, when it became evident that the material well-being of the African-American population actually had deteriorated in spite of revolutionary civil rights legislation. A report by the United States Department of Labor (1965) purported to find the cause of continuing social deprivation in family structure, and more specifically in the 'matriarchal' pattern of Negro families in which the roles of husbands and wives are reversed. The term matrifocal had been introduced precisely to avoid this kind of false reading of the data, but both critics and defenders of this report managed to negate that original meaning, moving toward Kunstadter's view and equating it with matriarchal or female headed.

It is ironic that most of the suggested cures for poverty require a return to traditional gender roles (on the assumption that a stable nuclear family based upon marriage is both natural and necessary for the adequate socialization of children), while at the same time gender roles themselves were undergoing a radical transformation as women of all classes increasingly entered the labor force, bore children out of wedlock, and frequently raised them in households without a male head.

Matrifocality and Feminism

It was inevitable that the term would be taken up in the new wave of feminism that has done so much to transform societies in the latter quarter of the twentieth century. The idea of a stage of human evolution in which women were dominant is not new. The Swiss jurist Bachofen (1861) took Herodotus's descriptions of the matriarchal system of the Lycians as its point of departure, but contributed to nineteenth-century speculations about the development of patriarchy out of primitive promiscuity through a stage of matriarchy (*see* Kinship in Anthropology). The same kind of research into myth and religion has been employed by recent theorists of cognitive or symbolic archaeology, most notably Marija Gimbutas and her followers. Admirers and opponents alike of Gimbutas's theory of a woman-centered stage of European social development have made free use of matrifocality in their commentaries, referring to a matrifocal stage in the development of European civilization (Marler, 1997). This kind of generalized use has spread far outside anthropological circles, sometimes being used to refer to almost any female focussed activity, whether motherhood is involved or not. The term has also found its way into discussions of animal behavior.

The Analytic Dimension of Matrifocality

Matrifocality has been identified by anthropologists in societies too numerous to list here, and including such varied cases as Java, Portugal, Thailand, Italy, South Africa, and Brazil, as well as the *locus classicus* of the Caribbean and urban North America. Many have used the broader definition of female-headed or female-dominated households and pointed to the economic difficulties faced by husband-fathers as a presumed cause of matrifocality, thus, implicitly (if not explicitly) characterizing such families as abnormal forms of the nuclear family. However, all raise the theoretical issue of the relations among class, status, gender, and the family. Matrifocality is never an isolated condition or a simple cultural trait; it must always be considered in the context of the entire social system of which it is a part.

One of the first studies to address that issue was the study of marriage in the context of race and class in nineteenth century Cuba (Martinez-Alier, 1974). She showed the effect upon domestic organization of the marginalization of colored men and women in a society where racial purity and legal marriage defined social status and social honor. In later work Smith (1987) argued that the matrifocal family in the Anglophone Caribbean is an integral part of the complex status system that emerged out of the *slave* regime, where men of higher status were able to marry equals and enter nonlegal unions (coresidential or visiting) with women of lower status, thus institutionalizing a dual marriage system that operated at all levels of the social system. It is not poverty that produces matrifocal families but the combination of societal norms, status, and property considerations stressing nuclear families based on legal marriage, coexisting with nonlegal unions, illegitimacy, and matrifocal families at all levels of the hierarchy for those whose unions do not serve to conserve either status or significant property. Another example is O'Neill's demonstration that in rural Portugal, illegitimacy and matrifocal families are closely related to the status and land-owning structure of local communities (O'Neill, 1987).

Bibliography

Bachofen, J.J., 1861. Das Mutterrecht. Krais and Hoffmann, Stuttgart (1967 Myth, Religion, and Mother Right: Selected Writings of J J Bachofen. Princeton University Press, Princeton, NJ.

Fortes, M., 1949. Time and social structure: an Ashanti case study. In: Fortes, M. (Ed.), Social Structure: Essays Presented to Radcliffe-Brown. Clarendon Press, Oxford, UK.

Kunstadter, P., 1963. A survey of the consanguine or matrifocal family. American Anthropologist 65, 56–66.

Marler, J. (Ed.), 1997. From the Realm of the Ancestors: An Anthology in Honor of Marija Gimbutas. Knowledge, Ideas & Trends, Manchester, CT.

Martinez-Alier, V., 1974. Marriage, Class, and Colour in Nineteenth-Century Cuba: A Study of Racial Attitudes and Sexual Values in a Slave Society. Cambridge University Press, London and New York (reprinted as Stolcke, V., 1989. Marriage, Class, and Colour in Nineteenth-Century Cuba: A Study of Racial Attitudes and Sexual Values in a Slave Society, second ed. University of Michigan Press, Ann Arbor, MI).

O'Neill, B.J., 1987. Social Inequality in a Portuguese Hamlet: Land, Late Marriage, and Bastardy, 1870–1978. Cambridge University Press, New York.

Smith, R.T., 1956. The Negro Family in British Guiana: Family Structure and Social Status in the Villages. Routledge & Kegan Paul, London (the full text of this book is now available, with the kind permission of Routledge, online at. http://home.uchicago.edu/~rts1/first.htm/).

Smith, R.T., 1973. The matrifocal family. In: Goody, J. (Ed.), The Character of Kinship. Cambridge University Press, Cambridge, UK.

Smith, R.T., 1987. Hierarchy and the dual marriage system in West Indian society. In: Collier, J.F., Yanagisako, S.J. (Eds.), Gender and Kinship: Essays Toward a Unified Analysis. Stanford University Press, Stanford, CA.

United States Department of Labor, 1965. The Negro Family: The Case for National Action. Office of Policy Planning and Research, US Department of Labor, Washington, DC.

Mauss, Marcel (1872–1950)

Victor Karady, CNRS, Paris, France

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9419–9421, © 2001, Elsevier Ltd.

Abstract

Marcel Mauss (1872–1950) was the most influential social scientist of the French School of Sociology headed by his uncle Emile Durkheim. Though an armchair scholar, having been teaching all his life at Parisian institutions of erudition (*Ecole Pratique des Hautes Etudes* – since 1901, *Collège de France* – after 1931), he is generally considered as the founding father of contemporary French social antropology. His work comprises epoch-making studies on comparative religion (Sacrifice, Magic, Eskimo rituals), social cohesion (especially the *Essay on Gift*, 1925), the arbitrary and multifunctional nature of cultural facts ('total social phenomena'), the prevalence of social determinism on individual behaviour and the social origin of mental categories (time, space, the self, etc.).

The father figure of contemporary social and cultural anthropology in France, Mauss was among the most influential social scientists of the early twentieth century. His scholarly activities and achievements were closely connected to the birth, organization, and coming to power in the French academe of the 'Sociological School' headed by Emile Durkheim and gathering among other luminaries Célestin Bouglé, Maurice Halbwachs, Robert Hertz, Henri Hubert, or Francois Simiand, the intellectual heritage of which is still being largely claimed, discussed, and studied worldwide (see *Durkheimian Studies*, Oxford) as one of the major foundation acts of modern social sciences proper.

Mauss was born on 10 May 1872 in Epinal (Vosges county) in a middle class Jewish family long established in Eastern France and having fully adopted the pattern of French Liberal patriotism. His personal career followed in many respects the rise of the first generation of famous French–Jewish intellectuals represented by Bergson, Durkheim, or Lucien Lévy–Bruhl. He shared with them their Republican commitment, notably as a young activist engaged in the battle in defense of Captain Dreyfus, as one of the earliest contributors (reporting on the cooperative movement) to the socialist journal *L'Humanité* founded by Jean Jaurès and, in the interwar years, as a dedicated expert of the (anti-Bolshevik) French socialist party (notably in matters financial and those related to the stock exchange). Mauss died on 10 February 1950 in Paris.

With formal university training in philosophy (a quasi must for would-be social scientists in his time), Mauss skipped the then almost inescapable stint of secondary school teaching, but also the stages of a normal academic career in the network ofthe Faculties of Letters. Several study trips and stays in Britain, Holland, and Germany enabled him to commit himself from the outstart to research and the training of scholars. He was first appointed as a lecturer on archaic religions (*The Religion of Peoples Without Civilization*) at the 5th Section (Sciences of Religion) of the *Ecole Pratique des Hautes Etudes* (1901), a unique agency housed in, but independent from, the Sorbonne and aiming at the instruction of specialists in various fields of erudition. He combined later (1926) these courses with seminars on 'descriptive ethnology' at the newly founded *Institut d'Ethnologie* under the aegis of the Sorbonne, only to occupy the chair of sociology at the *Collège de France* (1931), an even more prestigious institution of higher learning. While lecturing there he also shared in a Socratic manner with a number of (often mature) students the insights of a uniquely ingenious armchair anthropologist, until he was forced to early retirement under the German occupation in 1941. His positions in these both central, but academically marginal Parisian institutions helped him to achieve a rather atypical career of a scholar exclusively dedicated to research on exotic societies without visiting any of them and producing neither a thesis, nor even books proper, but, instead, a number of substantial studies which – though highly specialized by their thematic focus – refer equally to various tribal civilizations of Australia, Amerindia, Africa, and Oceania as well as to a vast variety of relevant 'social facts' of the Hindu, Hebrew, or Muslim world.

His education, intellectual development, and academic options were directed and supervised by Durkheim, his uncle and life long tutor. Their works – though intimately linked – appear to be considerably divergent. With the benefit of historical hindsight Mauss is credited nowadays as a less theory-bound, more inspiring and heuristically rewarding ancestor for contemporary successors of the 'French School of Sociology,' his personal impact extending upon many significant anthropologists, sociologists, folklorists, political scientists, social psychologists, or even human geographers or social historians starting a career in France after the First World War, especially after 1945. His association with Durkheim took a decisive form by his participation in the empirical elaboration of databanks for Durkheim's *Suicide* (1897) and continued, more importantly, with the editorial work of the *Année sociologique* (12 volumes between 1898 and 1913), the central scholarly organ of the 'Sociological School.' This quasi yearly survey of relevant publications embodied the first major attempt since August Comte (see *Comte, Auguste (1798–1857)*) to organize empirically the burgeoning social sciences into an inter-related system of disciplines. In charge of a large sector of the *Année* covering 'religious sociology,' Mauss became the author of one-quarter, approximately, of the 10 000 odd pages of review articles published in the *Année* under Durkheim, and took over the direction of the *Année* – transformed in the interwar years into the irregularly published *Annales sociologiques* – following the uncle's untimely death in 1917.

In the division of labor of what became the 'French School of Sociology' Mauss was, from the outset, dealt out the sector or

archaic civilizations, especially that of 'primitive religion,' strategic field for the experimentation of Durkheim's basic methodological precepts. In this view indeed the main 'social functions' appear to be apprehensible in a straightforward form in archaic societies, so much so that they may inform the interpretation of more complex societal realities as well. Thus, tribal societies – not without ambiguity – can be equated to the simplest, the most essential, and the historically earliest patterns of social organization. To boot, following the 'religious turn' occurring around 1896 in Durkheim's scholarly evolution, religion was considered as central among 'social representations' or objectivations of 'collective consciousness,' instrumental in the integration of archaic and, by implication, of all social formations. Hence the epistemological priority granted to the study of 'elementary religion' and the allocation of such pursuits to sociology proper as a master discipline, whatever the degree of development, power, extension, or complexity of societies concerned could be. Thus most of Mauss' work centered on comparative problems related to belief systems, rituals, and mental habits, and other collective practices serving as a foundation for 'social cohesion' in extra-European and pre-industrial civilizations, could qualify for being sociological, without much reference to anthropology, ethnology, ethnography, or folklore; such terms being practically absent from the topical listing of 'branches of social sciences' in the *Année*.

Mauss' major early studies are the outcome of his brotherly collaboration with Henri Hubert, the ancient historian and archeologist of the 'Sociological School.' They deal with the social functions of '*Sacrifice*' (1899) (biblographic details of Mauss' studies are listed by years of publication at the end of the third voume of his *Oeuvres* (pp. 642–94) and propose 'A general theory of magic' (1904), to which Mauss added a number of personal essays, especially those concerning 'The sources of magic power in Australian societies' (1904) and the first part of his unpublished (and unpresented!) doctoral dissertation on 'Prayer' (1909). Some of these essays will be published in their common book *(Mélanges d'histoire des religions*, 1909), where the authors offer in their introduction the outline of a sociological theory of religion. Denominational systems of faith are social facts and should hence be put on the agenda of sociology. They divide the life world of societies into a sacred and a secular (profane) sphere. The former is invested with essential collective values, so much so that most if not all moral customs, legal, educational, artistic, technical, and scientific practices display religious foundations. Mauss' first ever review article discussed already a book on 'Religion and the origin of penal law' (1896). In the light of his and his companions' comparative studies concerning 'religious representations' (including totemism, 'positive and negative rites' and their mythical rationalizations, the Melanesian notion of *mana* as the paradigm of socio-religious force, the religious origins of values – like that of money, etc.), religion emerges as a fundamental 'social function,' even if secularized modern societies tend to replace it by functional equivalents (like patriotic rituals) to bring about a desired degree of integration.

Mauss will be less systematic (some would say, less dogmatic) than his uncle in the generalization and theoretical exploitation of observations gained from the study of archaic religions. But he shares with Durkheim the conviction that the social prevails over the individual among norms commanding human behavior and, more specifically, that sociohistorical models are the sources of mental categories epitomized in the 'reason' or 'logos' of Western philosophy, permitting the intelligence of time, space, causality, genres, force, etc., though this approach was often condemned as 'sociocentrism' or 'sociological imperialism' by contemporary critics. In concrete terms Mauss' main interest lay, on the one hand, in the exploration of interconnections between collective practices and ideas, in the manner how mental elaborations respond to the organizational patterns of societies contributing thus to their cohesion and, on the other hand, in the exemplification of the arbitrary nature of cultural facts which become meaningful in particular social configurations only. The first focus would lead to decisive methodological insights for the sociology (or the social history) of knowledge. The latter would inspire developments in structuralist social anthropology.

Major statements about the covariation and the fundamental integrity of various fields of collective conduct are made in the essay of comparative anthropology published together with Durkheim on *Primitive social classifications* (1903) and in a *Study of social morphology*, produced with the collaboration of Mauss' disciple Henri Beuchat on *Seasonal Changes of Eskimo Societies* (1906).

In the much cited (and not less criticized) first study a vast array of evidence, drawn from a number of extremely different Australian and American societies, is presented and analyzed in order to illustrate the functional unity of societies (as totalities) in which mental habits, religious ideas, and even basic categories guiding the perception of environment and other – residential, economic, technical – practices are demonstrably interconnected. Thus, 'the classification of things reproduces that of people,' as exemplified in many tribal civilizations where membership in phratries, matrimonial classes, clans, gender, or age groups serves as a principle for the organization of religious and other social activities as well as for the interpretation of reality in 'logical classes.' The spatial situation of the tribe is a basic model for categories of orientation, like cardinal points. Hence the intimation, that categories of reasoning, instead of following universal patterns, are sociohistorical products and even modern science owes some of its principles to primitive categorizations. Such results will later encourage Mauss to affirm the pretensions of 'social anthropology to be once in the position of substituting itself to philosophy by drafting the history of the human mind.'

The study of Eskimo life offers a more empirical evidence for a case of covariation in time of the material substratum – morphology – of societies ('the mass, density, shape, and composition of human groups') and the moral, religious, and legal aspects of collective existence. The approach is, exceptionally, not comparative – though references are made to other exotic and even European societies too – only to be more closely focused on a privileged example where even the pattern of settlement of families concerned (gathering in winter, dispersion in summer) differs with the seasons, together with the intensity of social interactions, religious activities, and kinship relations. Winter represents the high point of social togetherness accompanied by festivities, the unity of residence of partner families, and various forms of economic, sexual, and

emotional exchange among their members, generating in them a strong sense of political and moral integrity.

In the second part of his career after World War I Mauss' work appears to be thematically more diversified without losing its initial foci centered upon the ways and means by which society determines individual behavior and the conditions of collective 'solidarity' or 'cohesion,' the latter – while realized thanks to contingent or 'arbitrary' assets – forming an operational whole called 'culture' or 'civilization.' This helps to restate the historical and ethnographically localized nature of mental categories. These problems and ideas inform some major studies published by Mauss in his elderly years like 'Relations between psychology and sociology' (1924), 'The physical impact of the idea of death suggested by society' (1926), 'The problem of civilisations' (1929), 'Bodily techniques' (1935), 'A category of the mind, the concept of person, that of the self' (1938). But the masterpiece emerging from that period remains the famous comparative essay *On Gift, Form and Reason of Exchange in Archaic Societies* (1925).

'The gift' is a vast, though fragmentary accomplishment of a program to study primitive economic systems where much of the exchange of goods is carried out by apparent donations accompanied by other – mostly symbolic or ritual – services, which make it a 'total social phenomenon.' The main focus of the study is what some North American Indians name *potlatch*, a system of 'total exchange' of 'agonistic' nature, whereby rivalry between groups, relationships of prestige and force, forms of social inequality and alliance are expressed, confirmed and enforced. The exchange in question may equally imply the circulation of goods, honors, courtesy, rituals, women, etc., and its essential principle is the obligation to offer and accept as well as return gifts, even if the things given hold purely symbolic value only, since many acts of donation consist of the ritualized destruction of a maximum of goods – which is the very meaning of the *potlatch*. Destruction of valuables is in such systems an essential source of power and prestige by the demonstration of one's possessions and the capacity to dispense with them on an apparently voluntary, but in fact strictly compulsory basis. Survivals of such arrangements, proper to several archaic societies, can be traced in the code of behavior of industrial societies as well. Modern contractual trading relations are still often completed or accompanied by moral customs of inescapable reciprocities grounded in gifts, sacrifice, symbolic grants, acts of generosity, the generalization and institutionalization of which (possibly in the regime of social welfare) appears in Mauss' view as a desirable development. He will extend the scope of the study of 'total social facts' in several essays, especially in one on 'Joking relationships' (1926), socially organized delivery of jokes and insults among kins, destined to strengthen the cohesion of the clan by the enforcement of verbal reciprocities.

Mauss' scholarly achievement has left a considerable mark on the social sciences in France and elsewhere. In contemporary French sociological tradition he is considered – alongside may be with Maurice Halbwachs – as the most 'legitimate' ancestor from the 'Sociological School.' The historic importance of his work has been selectively and critically received but warmly appraised by men like Pierre Bourdieu (who is holding a chair of sociology illustrated by Mauss at the *Collège de France*), Roger Caillois, Georges Dumézil, Louis Dumont, Georges Gurvitch, Michel Leiris, Claude Lévi-Strauss (who succeeded Mauss at the *Collège de France*).

See also: Anthropology, History of; Belief, Anthropology of; Classification, Understandings of in the Social Sciences; Collective Beliefs, Sociological Explanation of; Communes, Anthropology of; Community and Society: History of the Concepts; Durkheim, Emile (1858–1917); Exchange in Anthropology; Exchange: Social; Folklore; Functionalism in Anthropology; Habit: History of the Concept; Individuals in Society: History of the Concept; Magic, Anthropology of; Potlatch in Anthropology; Religion, Family, and Kinship; Religion, History of; Religion, Phenomenology of; Religion, Sociology of; Religion: Evolution and Development; Religion: Morality and Social Control; Ritual; Sacrifice; Social History; Sociology, History of; Solidarity, Sociology of; Solidarity: History of the Concept; Totemism; Tribe.

Bibliography

Fournier, M., 1994. Marcel Mauss. Fayard, Paris.
Karsenti, B., 1997. L'homme total. Sociologie, anthropologie et philosophie chez Marcel Mauss. Presses Universitaires de France, Paris.
James, W., Allen, N.J. (Eds.), 1998. Marcel Mauss, A Centenary Tribute. Berghahn Books, New York.
Mauss, M., 1950. In: Sociology et anthropologie. Presses Universitaires de France, Paris.
Mauss, M., 1968–1969. Oeuvres, vols 3. Éditions de Minuit, Paris.
Mauss, M., 1997. Écrits politiques. Fayard, Paris.
Mauss, M., 1998. Lettres écrites à Marcel Mauss. Presses Universitaires de France, Paris.

McDonaldization

George Ritzer and Chih-Chin Chen, University of Maryland, College Park, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

McDonaldization is Ritzer's theoretical adaptation of Max Weber's theory of rationalization. While Weber uses the bureaucracy as the model for the process of rationalization, Ritzer uses the fast-food restaurant, especially McDonald's, as the contemporary paradigm of the process. He defines McDonaldization as "the process by which the principles of the fast-food restaurant are coming to dominate more and more sectors of American society as well as of the rest of the world."

The McDonaldization of Society

The process of McDonaldization as well as the following principles that lie at its base have, in various ways and forms, found their way into most organizations and institutions in American society and as well as into many other parts of the world (Ritzer, 2010):

Efficiency: is defined as the optimum method of getting from one point to another. Emblematic of efficiency is McDonald's drive-through that reduces to the minimum the necessary time spent on ordering food and getting the needed service. Efficiency is accomplished through a series of modifications in the system designed to accomplish a wide range of ends. The efficiency of the McDonaldized system has its roots in Weber's bureaucracy, Ford's assembly line, and Taylor's scientific management. This culminated in a highly efficient method of accomplishing virtually any task. An example outside the fast-food industry is the McDonaldized educational system, which seeks to produce test results in the most efficient manner possible through the use of multiple choice exams and scantron grading system.

Calculability: involves an emphasis on the quantitative aspects of products and services such as size, price, and speed. McDonald's does not tout its products because of their exquisite quality or distinctive nature, but emphasizes their size. The best example is the 'Big Mac.' Like McDonald's, other highly McDonaldized companies such as IKEA and Walmart compete on the basis of low price, rather than on the high quality of goods and services provided.

Predictability: is the assurance that products, services, and settings will be very similar, if not entirely identical, over time and in all locales. There are no surprises or shocks in these settings and in the products offered. Instead, there is the safety and comfort of experiencing the same processes and products over and over.

Control: is the power exerted over both customers and employees of McDonaldized settings. Such control is accomplished by, for example, a material structure (e.g., the drive-through) that restricts actions available to consumers, scripted behavior on the part of employees, and a strict control on work processes. Interaction between consumers and employees is highly restricted and controlled. The possibility of any unwanted human elements occurring is thereby reduced.

These principles make McDonaldization a major aspect of, and force in, the process of globalization. McDonaldization is a key example of what Ritzer (2003, 2007) calls "the globalization of nothing." In this context, nothing is any social form that is centrally conceived, centrally controlled, and lacking in distinctive content. McDonald's is an example of 'nothing' since its 35 000 restaurants around the world are conceived and controlled by headquarters in Illinois and every McDonald's follows the same basic model.

The Irrationality of Rationality

Ritzer's critique of McDonaldization is not to be found in the four principles per se, but rather mainly in the idea of *the irrationality of rationality*. McDonaldization has provided many benefits for management, but many disadvantages for the consumers. The efficiency for restaurants often results in inefficiency for customers. McDonald's customers must often wait in line or stuck in traffic jams in the drive-through lane. Walmart and Costco customers must often buy things in bulk and some of it often proves unnecessary and inefficient to transport and store. Automated phone services often keep callers waiting and transferred from one extension to another, when callers could have received much more efficient service from a trained customer service representative. Seemingly, efficient methods and processes often prove inefficient for customers, and sometimes even for organizations and their employees.

McDonaldization does not necessarily produce the cheapest products and services for customers. It greatly reduces the manufacturing, distribution, and the processing cost for McDonaldized organizations, but this reduction is not fully reflected in the sale prices. A meal for four in a fast-food restaurant is priced much higher than its cost of production, to say nothing of what its nutritional value would suggest. The same amount of money spent on ingredients for home cooking would produce a much more wholesome meal. Disney World, a highly McDonaldized entertainment setting, is becoming more and more expensive, while the actual satisfaction seems to diminish. Some years ago, a journalist calculated that due to, among other things, long drives and waits in the park, his 5-day trip to Disney World resulted in a grand total of 6 h and 47 min of actual fun at the cost of $261 per hour (it would be much more expensive today) (Garfield, 1991).

In a McDonaldized society, more and more work is being performed by the consumers. For example, they take on the work of bank tellers when using ATMs, the work of waiters and buspersons when ordering, eating, and cleaning up in the fast-food restaurants, and the work of customer service representatives when working their way through the automated phone directories. Customers are forced to perform more and more

work, but they are not paid for their labor. Ritzer and Jurgenson (2010; Ritzer et al., 2012) have urged a theoretical shift in scholars' mind-set from a focus on either producer or consumer to one that focuses on consumers who work (as well as producers who consume). In other words, the focus should be on 'prosumers.' In order to understand the true nature and the social logic hidden in the structure of economic activities, we have to look at economic activity in its totality that involves an intertwined and codependent fusion of production and consumption.

Following another aspect of Weber's work, McDonaldized settings can also be seen as disenchanted (Ritzer, 2010). That is they have lost, if they ever had, their magic. Every process, object, and interaction is spelled out, structured, and routinized to the minutest detail and followed closely by both employees and customers. The excitement of the unknown, the mystery of the unpredictable, and people's idiosyncratic interactions are controlled or eliminated by McDonaldized systems. Those who manage McDonaldized settings care less about how good and enjoyable their services and goods are, but focus more on such things as the number of customers served and of the products sold. McDonaldization's pursuit of quantity has reduced or eliminated its ability to create magical experiences through, for example, the pursuit of authenticity and quality.

McDonaldization tends to dehumanize those involved in the process. It tends to turn employees into nonthinking robots who mindlessly follow scripts and customers into little more than livestock fed on a food conveyor belt. Both are transformed into cogs and wheels of the machine, which is a McDonaldized system. Neither is allowed much ability to think and act independently or to deviate much from the scripts. The fantasy (or nightmare) McDonaldization constructs for its customers is one of mechanical dystopia in which everything and everyone is routinized, standardized, and preprogrammed. In McDonaldized systems, one can choose to be livestock or robot, but never the human of free will and agency.

McDonaldization has produced "Islands of the living dead" (Ritzer, 2003), as opposed to Weber's 'iron cage.' McDonaldization has transformed each setting into an individual island of McDonaldized processes and rules, but there is no overarching bureaucratized society with an iron cage that structures all settings to operate in the same way. The efficiency produced in one setting may not provide the same level of efficiency in a different setting. Thus, people move from island to island (e.g., fast-food restaurant to Walmart) and are led to act like the 'living dead' in each setting. Governed by the principles of McDonaldization, each island deadens people by restricting their freedom to think and act independently. The only true freedom available to people is in the non- or less-McDonaldized spaces between islands of the living dead.

The Advantages of McDonaldization

In spite of its numerous problems, McDonaldization has its advantages. It has helped to produce a monumental abundance for society. A wide variety of goods and services have become accessible to portions of society that found it difficult or impossible to obtain them in the past, at least at affordable prices. For example, the supply of medicine and other essentials for survival are now provided to more people than ever, and this is made possible by McDonaldizing the processes and organizations involved in the production of pharmaceuticals. The speed by which goods and services are delivered has increased significantly (e.g., FedEx). The convenience of efficiency also appeals to people who are working longer and have less time to spare. McDonaldized settings' uniform treatment of employees and customers also decreases discrimination based on race, sex, and class.

Theorizing and Studying McDonaldization

A number of other theories relate to the theory of McDonaldization. Some are similar, some are antithetical, and some extend the ideas into fields unrelated to consumption. Benjamin Barber's (1992, 2010) "Jihad vs McWorld" applied the idea of McWorld to the field of political science. He explores the struggle between McDonaldized process and entities, 'McWorld,' and indigenous and local entities, 'Jihad.' While Jihads in various places attempt to retain their local distinctiveness and authenticity, McWorld eliminates differences, leaving the same predictable product and ideologies everywhere. Barber attributes the forces behind McWorld and its global expansion to neoliberal economic forces and active corporate control. Barber believes that McWorld will eventually win out over Jihad and homogenize the entire planet due to its financial and political advantages. Ritzer (2003, 2007) proposed a similar duality "something vs nothing" to explain the tension in the globalization process. While something involves forms locally conceived and controlled with distinctive content, nothing, as pointed out above, encompasses forms that are centrally conceived and controlled, and that lack distinctive content. Ritzer argues that something is rapidly losing ground in the face of the spread of nothing in a process he calls "grobalization of nothing," but he also leaves the room for a dialectical relationship with the "glocalization of something." This offers some hope in an otherwise gloomy view of a world increasingly dominated by the nothingness associated, especially, with McDonaldized System.

Some approaches such as "Starbuckization," "IKEAization," and "eBayization" offer alternative explanations and paradigms to deal with the changing landscape of society. Starbuckization offers a softened image of McDonaldization with theatrical performances making the customers feel welcome to stay, rather than being harried and rushed in McDonald's. IKEAization emphasizes the transfer of work from employees to customers, like McDonaldization's "putting customers to work." eBayization offers an alternative picture of market-directed activity (rather than McDonaldization's centrally controlled system), a variety of products (as opposed to McDonald's limited and predetermined menu), and unpredictable offerings (in contrast to McDonald's predictable products). Even though all of these processes offer interesting and useful perspectives, their structure and leading principles continue to rely heavily on the principles of McDonaldization. Starbucks' success is built on the premise of selling coffee in a McDonaldized grab-n-go fashion; it would never be able to sustain itself on the revenue from the relatively few sit-in customers. IKEA's products are highly McDonaldized and can even be thought of

as pushing the control over customers beyond the confines of its geospatial settings, as customers are likely to assemble furniture at home in accord with instruction provided by IKEA. eBay's variety and unpredictability are made possible by utilizing a highly McDonaldized infrastructure and interface for all listings and sales.

Many scholars have studied McDonaldization empirically. Drane (2002, 2012) examined the ways in which the church adopts the principles of McDonaldization. Hayes (2006) studied higher education and its degree of McDonaldization in, for example, the increase in the student–faculty ratio and the pumping out of more and more graduates. Timimi's (2010) work on children's mental health describes the McDonaldized process we have developed to deal with children's aberrant behavior. The expansionistic and increasingly universal nature of McDonaldization has led to many cross-cultural studies (Ram, 2004; Caldwell, 2004; Iqbal, 2011; Heslop, 2011; Peak, 2012).

Critiquing and Resisting McDonaldization

There are a number of criticisms of McDonaldization and critiques of how well the theory depicts and explains social reality. Many of the authors in Smart (1999) have critiqued the theory on empirical grounds. For example, McDonaldization does not mesh well with the current academic discourse on the postmodern society's transition to a post-Fordist mode of production. McDonaldization produces identical objects with a Fordist mode of production, and thus is not in accord with the differentiation and distinction postmodern consumers seek in today's commodities and services. Second, the highly predictable and secure 'comfort' McDonaldization is supposed to provide runs counter to the increasing risk and unpredictability of contemporary risk society (Beck, 1992/2006). McDonaldization is also heavily criticized for being a school of elitist thought in accord with Horkheimer and Adorno's (1969) critique of popular culture. Other authors in the Smart volume criticize Ritzer for oversimplification and neglecting the nuanced operation and management maneuvering of local units on the ground, as well as the advantages of the corporate system that enables the smooth operation of massive entities like McDonald's or Walmart.

There are many examples of resistance to the expansion of both McDonald's restaurants and McDonaldization. Some (such as McLibel) directly target McDonald's by pointing out its environmental, cultural, and health threats to society. Others choose to tackle the process itself and the ideology behind it. The Slow Food movement is the most famous organization seeking to protect and promote traditional ways of food production, preparation, and dining from the threat of McDonaldized food items and restaurants. Less noticeable struggles in local communities are reported by the media from time to time.

McDonaldization and the Internet

The promise of large-scale de-McDonaldization can be glimpsed in the realm of the Internet. The advent of Web 2.0 shifts much content production from a centralized source to the users. Wikipedia, Facebook, Craiglist, Twitter, etc., are all examples of de-McDonaldization. Much of their content runs counter to the four principles of McDonaldization (Jurgenson, 2010). They are often inefficient (requiring large investments of time and resources), unquantifiable, unpredictable, and less controlled. However, we should not overlook the fact that they are made possible by highly efficient, calculable, predictable, and highly controlled programming language and digital infrastructure. This points us to a very different picture in which both McDonaldization and de-McDonaldization exist simultaneously, as de-McDonaldized content accumulates on an increasingly McDonaldized platforms.

See also: Craft Production, Anthropology of; Culture and Economy; Customer Satisfaction; Ethical Consumers; Generations in History; Happiness and Work; Indigenous Management Styles; Industrialization, Typologies and History of; Innovation; Job Design and Evaluation: Organizational Aspects; Managerial and Organizational Cognition; Modes of Production; Organizational Climate; Strategic Adaptation; Wellbeing and Burnout in the Workplace: Organizational Causes and Consequences; Work, History of; Workplace Stress.

Bibliography

Barber, B., 1992. Jihad vs McWorld. The Atlantic Monthly 269 (3), 53–65.
Barber, B., 2010. Jihad vs McWorld. Transworld Digital.
Beck, U., 1992. Risk Society: Towards a New Modernity, vol. 17. SAGE Publications Limited.
Caldwell, M.L., 2004. Domesticating the French fry McDonald's and consumerism in Moscow. Journal of Consumer Culture 4 (1), 5–26.
Drane, J.W., 2002, 2012. The McDonaldization of the Church: Consumer Culture and the Church's Future. Smyth & Helwys Pub.
Garfield, Bob, 7 July 1991. How I Spent (and Spent and Spent) My Disney Vacation. Washington Post/Outlook, p. B5.
Hayes, D., 2006. The McDonaldization of Higher Education. lap.
Heslop, R., 2011. The British police service: professionalisation or'McDonaldization'? International Journal of Police Science & Management 13 (4), 312–321.
Horkheimer, M., Adorno, T., 1969. Dialectic of Enlightenment. Continuum, New York.
Iqbal, Z., 2011. McDonaldization, Islamic teachings, and funerary practices in Kuwait. OMEGA – Journal of Death and Dying 63 (1), 95–112.
Jurgenson, N., 2010. The de-McDonaldization of the internet. In: McDonaldization: The Reader, third ed, Thousand Oaks, CA, pp. 159–171.
Jurgenson, N., Ritzer, G., 2012. The Internet, web 2.0, and beyond. In: The Wiley-Blackwell Companion to Sociology, pp. 626–648.
Peak, B., 2012. The McDonaldization of Political Campaigning in Ireland. Doctoral dissertation. Dublin City University. School of Communication.
Ram, U., 2004. Glocommodification: how the global consumes the local-McDonald's in Israel. Current Sociology 52 (1), 11–31.
Ritzer, G., 1993/2010. The McDonaldization of Society 6. Pine Forge Press/Sage Publications, Thousand Oaks, CA.
Ritzer, G., 2003, 2007. The Globalization of Nothing. Pine Forge Press, Thousand Oaks, CA.
Ritzer, G., 2003. Islands of the living dead, the social geography of McDonaldization. American Behavioral Scientist 47 (2), 119–136.
Ritzer, G., 2010. Enchanting a Disenchanted World, third ed. Pine Forge Press, Thousand Oaks, CA.
Ritzer, G., Jurgenson, N., 2010. Production, consumption, prosumption the nature of capitalism in the age of the digital 'prosumer'. Journal of Consumer Culture 10 (1), 13–36.
Ritzer, G., Dean, P., Jurgenson, N., 2012. The coming of age of the prosumer. American Behavioral Scientist 56 (4), 379–398.
Timimi, S., 2010. The McDonaldization of childhood: children's mental health in neo-liberal market cultures. Transcultural Psychiatry 47 (5), 686–706.

Mead, George Herbert (1863–1931)

Daniel R Huebner, The University of North Carolina at Greensboro, Greensboro, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

American pragmatist philosopher George Herbert Mead (1863–1931) has had a major influence in the social sciences through his examinations of the processes of symbolic communication and the genesis of the social self. He developed his philosophy in dialogue with unique influences from empirical science, critical philosophy, and the progressive social reforms in which he was heavily involved during his eventful career. Although his ideas were known to colleagues and students at the University of Chicago in the early twentieth century, his subsequent influence has largely been through volumes compiled from disparate sources and published after his death.

George Herbert Mead (1863–1931) was an American philosopher, considered to be one of the classical representatives of pragmatism, along with Charles S. Peirce, William James, and John Dewey. For Mead, Dewey was a particularly important interlocutor and colleague. Mead's ideas have had a major influence in the behavioral and social sciences, especially in sociology and social psychology, where he is best known for his theory of the development and process of human symbolic communication and for his account of the social nature and genesis of the 'self.' These notions have been central to the formation of the 'symbolic interactionism' perspective in sociology and have led Mead to be considered a foundational theorist in that discipline. The following sections examine Mead's philosophy in the context of his education, teaching, and social reform efforts and discuss the dominant interpretations made of his work.

Life and Context

Early Life and Education

George Herbert Mead was born to Rev. Hiram Mead and Elizabeth Storrs (*née* Billings) Mead in the village of South Hadley, Massachusetts on 27 February 1863. Hiram Mead came from a large, educated New England family, and several of his siblings also pursued Protestant religious callings. Hiram was the pastor of the First Congregationalist Church of South Hadley when George was born, and he remained in that position until appointed to the chair of Sacred Rhetoric and Pastoral Theology at Oberlin Theological Seminary in 1869, at which time the family moved to Ohio. Elizabeth Mead also came from a large, educated New England family, which including Elizabeth's twin sister Harriet. She was educated at Ipswich Female Seminary and taught in secondary schools prior to her marriage. After Hiram's death in 1881, Elizabeth began teaching English composition at Oberlin College and, later, Abbot Academy, before becoming president of Mount Holyoke College in 1890. Documentation on George Mead's early life is somewhat sparse, but it is clear that he and his older sister Alice were raised in an environment that highly valued broad education with emphasis on religion and service. The biographical information in this article draws primarily upon work by Joas (1985), Cook (1993), and Huebner (2014).

After attending the 'preparatory' department connected with Oberlin College, George Mead attended Oberlin from 1879 to 1883, graduating with a bachelor's degree in 'Philosophy and the Arts.' He became close friends with classmate Henry Northrup Castle in the last 2 years of college, bonding over mutual interests in philosophy and literature. As in many private American colleges of the nineteenth century, Oberlin's humanities faculty taught an 'Intuitionist' philosophy that emphasized the freedom of the will and the intuitive perception and experience of the world. As Mead and Castle were exposed to the critical, materialist doctrines of modern scientific researchers from their science classes, however, they became increasingly skeptical about this philosophy. They also coedited and contributed to the student news and literary paper, the *Oberlin Review*, during their final year of college, 1882–83. When Henry's older sister, Helen Kingsbury Castle, came to Oberlin to attend school that year, Mead met his future wife for the first time, although they would not marry for almost another decade. Helen and Henry were the youngest children of one of the most prominent American settler families in the (then independent) Kingdom of Hawaii. Their parents had arrived in the islands with early American missionaries, but quickly took up business for themselves in sugar production, shipping, and general merchandise. As a result, Helen and Henry were heirs to the massive company Castle & Cooke, a fact that would play a major role in Mead's life.

After college, Mead took a series of jobs to help support himself and his mother while he struggled to find a meaningful direction for his adult life. From 1883 to 1887, he taught at an elementary school, worked on railroad surveying crews, and tutored young men preparing for college. During this period, he wrestled with an increasing agnosticism, resulting at least in part from his independent study of critical philosophy and modern physical and evolutionary science. He considered possible careers, including starting a preparatory school or a literary magazine or pursuing some ameliorative work in the Christian ministry. However, Mead was too conflicted to pursue a religious calling and he did not have the resources for the others. Meanwhile Henry Castle, who had worked for a period in his brother's law office in Honolulu and had

attended courses in Germany, decided to attend graduate school at Harvard in 1886. Henry's letters to Mead, and his trip to visit Henry that year, seem to have helped Mead to decide on pursuit of graduate education in philosophy.

Mead attended Harvard beginning in the fall of 1887 as an advanced undergraduate (Harvard did not accept his degree from Oberlin). He roomed with Henry and took courses with Josiah Royce, George Herbert Palmer, and others. Although documentation is sparse, Mead's views on the self, symbolic communication, the history of science, and social progress seem to evince influences from Royce's idealist philosophy. At the end of the first year, Mead took the oral examination for an honors degree, and his performance impressed foundational psychologist and pragmatist philosopher William James. Although Mead did not take courses with James, he was invited to tutor James' son at their country house in the summer of 1888, and he received encouragement from James to apply for fellowships to pursue further graduate study in Germany.

As a result, Mead began coursework at the University of Leipzig in the winter 1888–89 semester, where he studied with Wilhelm Wundt, among others, and again roomed with Henry Castle. Although Wundt was perhaps the leading psychologist in Germany and directed what is often considered the first psychological laboratory, Mead took only his philosophy course on 'metaphysics.' After that semester, Mead transferred to the University of Berlin (Friedrich-Wilhelms-Universität). The circumstances of Mead's move are not entirely clear, but recent work has shown that William James recommended Mead to Hermann Ebbinghaus, Berlin's premier experimental psychologist. By this time, Mead had decided to focus his study on 'physiological psychology' because, according to Castle, this was a topic upon which he could pursue critical inquiry without fear of 'anathema and excommunication' from the 'all-potent Evangelicalism' of American Protestantism.

At Berlin, Mead became a laboratory assistant in Ebbinghaus's laboratory, working on psychophysical experiments. He took advanced courses in anatomy and physiology with Wilhelm Waldeyer and Hermann Munk that focused on recent discoveries about the central nervous system. From Wilhelm Dilthey, he took courses in ethics and the history of philosophy and planned to write a dissertation on the construction of the human conception of space through the cooperation of the hand and eye. Dilthey was a leading epistemologist of the human sciences, arguing that descriptive and interpretative methods were more adequate to the study of the meaningful structure of human experience than were natural scientific experiments and explanations. This period appears to have had a major impact on Mead's subsequent work, introducing him to the philosophical questions entailed in scientific discovery and raising questions about the nature of human experience and consciousness. He also became interested in the social democratic labor movement, the developmental psychology of children, and the psychological treatment of ethics during his time in Germany.

Professional Career and Social Activism

Prior to completing his degree at Berlin, Mead was offered a position as instructor in philosophy at the University of Michigan beginning in fall 1891 by the department's chair, John Dewey. Although he maintained hopes of completing his Berlin degree *in absentia*, Mead never obtained a graduate diploma. During his last year in Germany, Mead had become engaged to Helen Castle. They married in Berlin on 1 October 1891, just before traveling to Ann Arbor, Michigan for George's first semester there. Mead's position required him to teach courses in physiological psychology and the history of philosophy, including recent scientific developments such as evolutionary theory. Mead also took charge of the new psychological laboratory, where he prepared specimens of neurological tissue, performed psychological experiments, observed behaviors of microscopic and complex animals, and considered evidence from pathological psychology. On 30 November 1892, Helen Mead gave birth to their only child, Henry Castle Albert Mead, named for his uncles Henry Castle and Rev. Albert Swing (the husband of George's sister, Alice).

Mead stayed at the University of Michigan for 3 years, the last as assistant professor. During that time, Dewey and Mead, along with instructor Alfred H. Lloyd, formulated a unique 'functional psychology' perspective that brought the new findings in biology and psychology to bear on the problems of idealist philosophy. They were critical of theories of a mental 'substance' or definite mental structures, and instead proposed to examine fundamental philosophical and psychological phenomena such as consciousness and emotion in terms of the normal, physiological, and social processes of action in which they were implicated.

In 1894, Dewey was offered the chair in philosophy at the newly formed University of Chicago, and as part of his acceptance he negotiated an offer for Mead. Chicago was one of the new comprehensive research universities that developed in the United States in the late nineteenth century, largely following the German university model. At Chicago, they became part of a close-knit cohort of innovative faculty, and they formed early ties with the social and biological sciences. Until Dewey left in 1904–05 for Columbia University, the philosophy department also contained the faculties of psychology and education, and it administered the psychology laboratory and experimental pedagogical schools.

Throughout his 36 years at the University of Chicago, Mead offered a wide variety of courses. He taught a large number on the history of philosophy and the history of science, including all periods but especially ancient Greek thought and all fields and major authors of modern philosophy. His classes on the epistemology and methodology of the physical, psychological, and social sciences also evince his interest in the philosophical problems of science. When he moved to Chicago he ceased offering courses in physiological psychology and gave up directorship of a psychological laboratory, but in exchange he began offering what may be the earliest courses on 'comparative psychology' in any American college, from 1894 through 1911. In these courses, he examined the development of intelligence by tracing the increasingly complex forms of control of environments by biological organisms. During that period, Mead also actively observed and offered suggestions on the studies of animal behavior conducted by students in the psychology laboratory.

Beginning in 1900, he taught 'social psychology,' first intermittently, and then regularly after 1908. From 1918 on, this course was reformulated as 'Advanced Social Psychology' to reflect the growing literature in the field and the newly instated introductory social psychology courses in Chicago's

psychology and sociology departments. This was the course that became famous after his death through the book *Mind, Self, and Society* (Mead, 2015[1934]). After 1913, he taught periodic seminars on 'social consciousness' and other related social psychological questions, and from 1905 to 1910 he taught courses on the philosophy of education and the teaching of science. In the last decade of his career, he became increasingly interested in the controversies arising from relativity theory and contemporary physics, and he taught several advanced courses on these issues. In addition to his courses at the University of Chicago and other lectures at local professional schools of social workers, Mead gave over 200 public speeches. Indeed, he was undoubtedly known more widely during his lifetime for his public reform efforts than for his contributions to professional philosophy or social thought.

During the late nineteenth and early twentieth centuries, Chicago was rapidly expanding and industrializing, and it was the scene of major contentions over labor rights, immigration, race relations, public education, and political corruption. The city was a center of the American Progressive Movement, and George and Helen Mead were heavily involved in these social reform efforts. The Meads fostered a relationship with Jane Addams and the 'Hull House' social settlement in their first months in Chicago. They advocated publicly on behalf of women's suffrage, the rights of immigrants and minorities, birth control education, compensation for prison labor, development of juvenile justice institutions, and increased opportunities for education and employment of women. In 1910–11, when the garment workers of Chicago went on strike over working conditions, Helen met with the striking women and donated to their support while George worked as part of the semiofficial 'Citizen's Committee' to develop recommendations for arbitration.

The Meads were involved in many efforts to democratize education in Chicago and elsewhere. In addition, George Mead helped found the 'Chicago Physiological School,' an early experimental school dedicated to the study of children with disabilities; the 'Juvenile Psychopathic Institute' (later the 'Institute for Juvenile Research') dedicated to the study of juvenile delinquency; and the University of Chicago 'Laboratory School,' an experimental school dedicated to participatory pedagogical methods that make use of the spontaneous interests and activities of students. In these and other endeavors, Mead was an advocate of coordinating systematic data regarding education and social conditions. He helped develop plans for a municipal 'bureau of social research' that would coordinate the data collected from various public and private agencies concerning the citizens of Chicago, an effort that ultimately resulted in creation of the city's Public Welfare Bureau.

The Meads also promoted many of these same efforts in Hawaii, where Helen's family members were prominent officials and philanthropists, and where the Meads made 13 extended stays together from 1897 to 1924. The Castles were strong supporters of the revolution against the hereditary monarch in 1893 that established the short-lived 'Republic of Hawaii' and the subsequent US Territory of Hawaii. On several occasions, Mead accompanied the Territorial Governor and other officials on hiking tours of the islands, examining the conditions of native and immigrant life, agriculture, and geology, and he participated in public discussions on US annexation, public education, and municipal governance. He even officially represented the Territory as their delegate to the 'National Farm Land Congress' in 1909. The Mead's also helped arrange extended stays in Hawaii for Jane Addams, John and Alice Dewey, and progressive activist Harriet Park Thomas (first wife of sociologist W.I. Thomas), among others.

George Mead's influence in social reform was probably at its peak during the 1910s, when he was the president of the City Club of Chicago and held official positions in the University of Chicago Settlement League, the Immigrants' Protective League, and the Vocational Supervision League. During this period, he helped prepare reports on Chicago's public school system, vocational education efforts, public library, and the living and working conditions for people living in its 'packinghouse' district. Although the Meads were early supporters of the 'Woman's Peace Party' that argued against the United States entering World War I, by late 1917 George Mead began to advocate intervention in order to combat perceived imperialism and promote democratic institutions. In this turn, he was like many American intellectuals, including President Woodrow Wilson. During this period, Mead's published work sought to define the acceptable grounds for pacifism and conscientious objection to war, the social psychological origins of hostility, the possibilities for peaceful expansion of social consciousness, and the uses of nationalism for the promotion of infrastructure-building and reform.

The Meads' son Henry became a foreign correspondent for the Chicago Daily News stationed in Berlin during the early months of World War I. After a period back in the United States, Henry volunteered for the military and was sent as an officer to France. In August 1918, he was seriously wounded by shrapnel in battle and spent several weeks in a military hospital while his family tensely awaited word. Henry's wife, Irene Tufts Mead (daughter of George's colleague James H. Tufts), volunteered to aid wounded women and children in France during the war as well. During the same time, Mead was in charge of supervising the curriculum for several midwestern states of a newly founded civics education program called the 'War Issues Course' instituted in Student Army Training Corps throughout the country.

After the war, the Meads were members of the League to Enforce Peace and prominent advocates for adoption of the Covenant of the League of Nations. This advocacy brought controversy, as George Mead was accused by US Senator Medill McCormick of improperly using his position at the City Club to promote a partisan, personal belief. Accounts of Mead's intellectual biography have often pointed to the trauma of World War I as a major turning point in Mead's public persona (Cook, 1993; Deegan, 2008; Joas, 1985). Although he quit some of his more prominent posts in the aftermath, Mead remained active in the Vocational Supervision League and the Chicago Council on Foreign Relations until near the end of his life. This career-spanning dedication to social reform work clearly influenced his understanding of democratic governance, the nature of scientific inquiry, and social consciousness.

Late in his career, Mead began to receive professional recognition. In 1929, he taught courses at the University of

California, Berkeley as the honorary visiting Mills Professor of Philosophy, and in 1930, he gave the Carus Lectures, considered one of the highest honors awarded by the American Philosophical Association. The honor had been bestowed only twice before, to John Dewey and A.O. Lovejoy, and included publication of the lectures as a book by Open Court Publishing as part of the award.

At the same time, however, his late life was witness to illness and frustration. Helen Mead died on Christmas day 1929 following a period of declining health. The following year, Mead reluctantly accepted the interim chairmanship of the Chicago philosophy department when his longtime colleague, James H. Tufts, retired earlier than expected. He was faced with addressing the demands of the new university president, Robert M. Hutchins, who mishandled appointments by questioning the acumen of current philosophy faculty and attempting to appoint his acquaintances rather than the faculty's own recommended candidates at salaries higher than the sitting philosophy faculty (Cook, 1993). Mead quit his position in protest along with several junior faculty members in late 1930, and his action caused considerable controversy when it became public in February 1931. He planned to accept John Dewey's invitation to teach at Columbia University the following fall. However, he fell seriously ill and was forced to give up teaching the remainder of his courses at Chicago at the end of February 1931. He died on 26 April 1931 at the age of 68, in the Chicago home of his son and daughter-in-law.

The Work

The center of Mead's philosophy is typically considered his social psychology, which begins not with individual organisms but with an ongoing 'social process' consisting of humans cooperating in the accomplishment of a 'social act.' Drawing from Charles Darwin's analysis of the 'expression of emotions' in animals and Wilhelm Wundt's notion of 'gesture,' Mead proposed that organisms learn to respond to the movements of one another in ways that anticipate the further courses of action. When an organism responds to the early phases of another's action as if it indicated the complete act, that movement becomes a gesture or 'truncated act' that may serve as a 'sign' for the act as a whole. When organisms adjust back and forth to one another's gestures in guiding the course of action, there is a 'conversation of gestures.' In developing this theory, Mead sought not to presuppose an innate self-indicative faculty in the mind and instead to show that if humans appear to have this self-reflective or symbol-using capacity, then it must result from the interaction of fundamental organic features of the human species and the features of social action.

Gestures can come to have the same 'meaning' for all of the participants in an act when the participants learn to respond to their own gestures in the same way as those of others. This happens primarily through the 'vocal gesture,' because individuals can hear and respond to their own vocalizations in much the same way that they hear and respond to one another. As a result, what had been merely gestural signs become 'significant symbols' to the extent their utterance indicates the same set of actions for all those involved.

For Mead, the development of this symbolic capacity has enormous consequences. It allows humans to internalize and anticipate not only the action of others, but also their own action. Humans can respond virtually (that is, within the nervous system of the individual, rather than through external response) to gestures and can check or inhibit responses by the influence of other gestures. Indeed, Mead thought that this 'internalized conversation of gestures' constituted the fundamental process of rational thought. Because an individual's partners in social action have this same anthropological capacity, they can develop binding, reciprocal expectations regarding one another's behavior, and all those who share common meanings for some set of gestures constitute a 'universe of discourse.' The development of this symbolic communication fundamentally transforms the possibilities of social organization, by allowing for rational anticipation and planning, complex functional interdependence of social groups, and the possibility of progressive social change.

Mead proposed that humans 'take the role of the other' when they anticipate another's behavior. This concept is often misunderstood to mean that humans imaginatively project themselves into one another's places, but Mead explicitly sought to avoid a theory that relied on such mental projection or empathy. Additionally, individuals progressively develop the capacity to take the role of others toward their own actions, and thus they develop a consciousness of 'self' by treating their own action as a social object. In conduct, the self has two phases: the 'I' and the 'me.' The 'I' is the impulsive aspect of the self, which acts spontaneously in response to present situations. The 'me' is the internalization of the expectations of others through the course of experience, and serves to evaluate and structure the spontaneous responses of the 'I.' Each unique person or group that serves as a significant 'other' for the individual contributes to the internalization of a different 'me.' The integration of a unitary self requires the synthesis of these various responses in previous situations to numerous others into a coherent center of action and evaluation, and hence it is both a practical and moral problem for the individual.

Developmentally, the formation of the self happens through a series of stages. The earliest stage discussed by Mead is 'play,' in which a child acts out a course of conduct in a series of roles. The child takes on a new role when he or she responds to his or her own action in previous roles. This play is how the child practices and learns the situational expectations of various roles. In the 'game' stage, the child begins to anticipate and guide his or her conduct not by taking on various roles, but by the organization of one's responses to the expectations of several others into one coherent role. This organization is guided by the functions that the various others adopt within a group activity, such as the 'positions' in a baseball game, and learning to play the game means being able to anticipate the actions of the other players in relation to one's own position and respond appropriately.

The furthest stage in the formation of the self is the development of the 'generalized other' that occurs as the individual is able to orient his or her action toward the goals or values of the group action in which they are involved and to respond to those goals or values as a coherent 'other.' In this way, individuals develop the ability to evaluate and react to the group as a whole. Individual responsibility for one's actions and for the

direction of social action comes as the individual is able to identify that action with the self, and in that sense to be 'self-reflective.' Individuals can develop ever broader 'generalized others' to the extent that they orient their conduct toward more encompassing social action processes beyond the immediate groups and situations of which they are a part. This theory was key to Mead's advocacy of inclusive democratic participation, the development of international cooperation, and pragmatic ethics more generally.

The principles of Mead's social psychology are also central to the other aspects of his philosophy. For example, the capacity to manipulate the surrounding environment (especially by use of hands) combined with symbolic and gestural communication allows humans to create and reconstruct a common physical environment of objects and to represent different perspectives on that environment. These capacities are key to the philosophy of nature and to the formation of scientific knowledge. The experiences of individual scientists are essential to the progress of modern science: each scientist begins with some socially shared conception of the nature of the objects under study, but when his or her experience testing the objects is inconsistent with that conception the investigator must reconstruct the object in a way that is consistent with all new and old observations. In this sense, they act as fully developed selves responding to and enlarging a community. In Mead's last few years, he worked to apply these notions to metaphysical or 'cosmological' questions of relativity, emphasizing the objective location of different 'perspectives' in nature and their fundamental 'sociality,' as well as the novel or creative emergence of both past and future from the present.

Influence of Mead's Thought

During Mead's life he was known widely for his reform work, and his philosophical work received significant attention almost exclusively from his closest colleagues and students at the University of Chicago in philosophy, psychology, and the social sciences. He was an important part of the so-called Chicago School of American pragmatist philosophy associated with John Dewey that was closely tied with the development of functionalist psychology, institutionalist economics, behavioral political science, empirical sociology, and progressive pedagogy. From 1920 to 1931, his 'Advanced Social Psychology' courses received additional attention from University of Chicago sociology students, because his former student Ellsworth Faris began teaching an 'introductory' social psychology course that instructed sociologists in Mead's ideas and served as an informal prerequisite for Mead's advanced course. Faris also wrote a large number of articles and reviews in sociology publications, which argued that Mead's philosophy offered fundamental insights for sociology.

These efforts notably influenced students of the so-called Chicago School of sociology to learn aspects of Mead's social psychology. Herbert Blumer, who was a student at the University of Chicago from 1923 to 1928, has no doubt had the largest influence on the subsequent interpretations of Mead in sociology. He was also the person who took over teaching Mead's 'Advanced Social Psychology' course when Mead died. Blumer (1969) drew heavily from Mead in developing 'Symbolic Interactionismor,' and as a result Mead came to be taught and understood as the quintessential progenitor of this perspective. As Symbolic Interactionism and related phenomenological perspectives have been incorporated into mainstream sociology, Mead has become recognized as a classical sociological theorist.

Also important to the development of Mead's reputation has been the reception of his work in post–World War II social theory, especially in the work of German authors Arnold Gehlen, Hans Joas, Jürgen Habermas, and Axel Honneth. Joas (1985), in particular, considers Mead a central figure linking American pragmatism to continental philosophy and offering novel solutions to the problems of critical philosophy. Habermas' influential *Theory of Communicative Action* and other writings treat Mead's philosophy as a crucial resource in the formulation of his 'communicative action' framework and 'discourse ethics.'

The predominant source material for most commentaries on Mead's ideas has been a series of books published after Mead's death, and especially *Mind, Self, and Society: From the Standpoint of a Social Behaviorist* (Mead, 2015[1934]), which alone accounts for the vast majority of citations to Mead since the mid-twentieth century. Within days of Mead's death, members of his family began consulting with his colleagues and former students on the possibility of publishing a memorial or legacy for him. Although Mead had written several major journal articles and a large number of book reviews, paper abstracts, newspaper articles, editorial notes, committee reports, encyclopedia entries, and conference proceedings, he had never written a book-length work. From 1932 to 1938, four major works resulted from these efforts and were published under Mead's name: *The Philosophy of the Present* (Mead, 1932) based on the transcript of Mead's well-received 'Carus Lectures'; *Mind, Self, and Society* based on edited transcripts and student notes from Mead's social psychology courses; *Movements of Thought in the Nineteenth Century* (Mead, 1936) based on sets of notes from a course of the same name; and *The Philosophy of the Act* (Mead, 1938) compiled from manuscript fragments written by Mead in the last decade of his life and fragments from various student notes and papers. Since the 1960s, several other publications attributed to Mead have appeared, including manuscript papers and speeches that Mead gave and students' notes from several more of his courses in social psychology, philosophy of education, and history of philosophy. And beginning with the *Selected Writings* (Mead, 1964), there have been several attempts to publish in an accessible format his most important published articles and papers.

The volumes of posthumously published materials have been a frequent source of controversy, because of the lack of transparency in how they were constructed and because of the seemingly idiosyncratic representation they give of Mead's philosophy. For example, the term 'social behaviorist,' used to characterize Mead's work, appears only in *Mind, Self, and Society* and was added to the text by the volume's editor, Charles Morris (later known for his work in 'semiotics'). More generally, the order and internal structure of the volumes do not directly reflect the logical structure or development of Mead's philosophy. Instead, they are heavily influenced by the editorial processes through which the volumes were composed. In this light, some criticisms of Mead, such as

those that point out his supposed lack of focus on emotions or on social structure, result at least in part from the popularity of these posthumously constructed volumes rather than Mead's own body of published works, public speeches, and classroom teachings.

As interest in classical American pragmatism revived in the late twentieth century, Mead began to receive renewed critical attention from philosophers, historians, and others. With this increased attention to Mead from several fields and the criticisms of dominant interpretations of Mead has also come a growing critical literature on Mead's intellectual biography, the historical context of his ideas, and the formation of his intellectual legacy, including a four-volume collection of 'critical assessments' (Hamilton, 1992). Most recently, Mead's work has found new audiences in some branches of science studies, cognitive science, and ecology. However, Mead's ideas have continued to have their largest and most consistent influence on the discipline of sociology.

See also: Action, Theories of Social; Dewey, John (1859–1952); Dilthey, Wilhelm (1833–1911); Habermas, Jürgen (1929–); Interactionism, Symbolic; James, William (1842–1910); Pragmatism: Philosophical Aspects; Pragmatist Social Thought, History of; Self: History of the Concept; Self: Philosophical Aspects; Social Psychology, Sociological; Sociology, History of; Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Blumer, Herbert, 1969. Symbolic Interactionism: Perspective and Method. Prentice-Hall, Englewood Cliffs, NJ.

Cook, Gary A., 1993. George Herbert Mead: The Making of a Social Pragmatist. University of Illinois Press, Urbana and Chicago.

Deegan, Mary Jo, 2008. Self, War, & Society: George Herbert Mead's Macrosociology. Transaction Publishers, New Brunswick, NJ.

Hamilton, Peter (Ed.), 1992. George Herbert Mead: Critical Assessments, 4 vols. Routledge, London and New York.

Huebner, Daniel, 2014. Becoming Mead: The Social Process of Academic Knowledge. University of Chicago Press, Chicago.

Joas, Hans, 1985. G.H. Mead: A Contemporary Reexamination of His Thought (R. Meyer, Trans., second ed. 1997). MIT Press, Cambridge, MA.

Mead, George H., 1932. The Philosophy of the Present. In: Murphy, Arthur E. (Ed.). Open Court Publishing Co., Chicago and La Salle, IL.

Mead, George H., 1936. Movements of Thought in the Nineteenth Century. In: Moore, Merritt H. (Ed.). University of Chicago Press, Chicago.

Mead, George H., 1938. The Philosophy of the Act. In: Morris, Charles W., Brewster, John M., Dunham, Albert M., Miller, David L. (Eds.). University of Chicago Press, Chicago.

Mead, George H., 1964. Selected Writings. In: Reck, Andrew J. (Ed.). University of Chicago Press, Chicago.

Mead, George H., 2015. Mind, Self, and Society: From the Standpoint of a Behaviorist. In: Huebner, Daniel R., Joas, Hans (Eds.), revised ed. University of Chicago Press, Chicago. Originally edited by Charles W. Morris.

Relevant Website

http://www.brocku.ca/MeadProject – 'Mead Project 2.0'.

Mead, Margaret (1901–78)

Margaret M Caffrey, University of Memphis, Memphis, TN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article gives details on the life and professional career of American cultural anthropologist Margaret Mead, at the time of her death in 1978 the most famous anthropologist in the United States. It also outlines and follows what is known of the Mead–Freeman controversy, New Zealand anthropologist Derek Freeman's severe critiquing of her Samoan fieldwork and of *Coming of Age in Samoa* after her death. The controversy lasted an intense 18 years until Freeman's own death in 2001. That year was also the centennial of Margaret Mead's birth, and seems to have marked a turn away from the controversy and more interest in pursuing other lines of Margaret Mead scholarship.

Margaret Mead, noted twentieth-century American cultural anthropologist, was born on 16 December 1901, in Philadelphia, Pennsylvania, to Emily Fogg Mead, sociologist and social activist, and Edward Sherwood Mead, economist and professor at the Wharton School of Finance and Commerce at the University of Pennsylvania. The oldest child, she later had a brother and three sisters. She grew up in a household that had its share of troubles – a younger sister died at 18 months, her father and mother once came close to divorce. But it was also a household open to new ideas and social change. Her mother was an ardent advocate for the vote for women; her paternal grandmother, Martha Ramsay Mead, who lived with them, was, along with her mother, interested in experimenting with the latest ideas in education. Her father, a more conservative progressive, also acted as devil's advocate in lively debates over issues at mealtimes.

Her education was a combination of home schooling and local schools. She attended Buckingham Friends School and graduated from Doylestown High School, both in Bucks County, outside of Philadelphia. She spent her first year of college at DePauw University in Indiana, but found herself an outsider there. She transferred to Barnard, a highly regarded women's college in New York City, where she made lifelong friends and from which she graduated in the spring of 1923.

In the fall after graduation she married Luther Cressman, a young man she had met in Bucks County and been engaged to since leaving for DePauw. He was a college graduate and a graduate of General Theological Seminary in New York, which he attended while she was at Barnard. For both, the ideal marriage was an equal partnership.

They attended graduate school at Columbia University in New York City together, he in sociology, she receiving her master's degree in psychology, then her PhD in anthropology. At Columbia she was a student of Franz Boas, the preeminent anthropologist in the United States, and part of a distinguished anthropological community that included Melville Herskovits, Ruth Bunzel, Esther Goldfrank, and Ruth Benedict, who began as a mentor, and through the graduate school years became a colleague, close friend, and lover. A bisexual, throughout her life Mead developed relationships with both men and women. During this time she also met anthropologists from outside New York, of whom Edward Sapir was the most important. They had a brief relationship that ended while she was in Samoa.

She defended her dissertation, a piece of library research titled, *An Inquiry into the Question of Cultural Stability in Polynesia*, in the spring of 1925 (published in 1928). After accepting a position as an assistant curator at the American Museum of Natural History in New York upon her return, that fall she set off for her first field trip to American Samoa in the Pacific. She was gone altogether 9 months, while Luther studied in Europe. The result of that field trip was her book, *Coming of Age in Samoa* (1928), which showed Americans there could be a different system of morality in which premarital sex was accepted, and adolescence itself was not necessarily a time of 'storm and stress.' This book became a best-seller and began the process, which, by the time of her death, made her the best-known anthropologist in the United States. With this book she began a pattern she continued on later field trips, of writing a popular book for the general public, and also a scholarly monograph. In this case the monograph was *Social Organization of Manu'a* (1930). Both anthropologist Robert Redfield and Ralph Linton gave it favorable reviews. Redfield wrote, "The documentation that was lacking in *Coming of Age in Samoa* is given in this book. It is a monograph of exceptional merit" (Redfield, 1931).

On the trip home, she met New Zealander Reo Fortune, on his way to Britain to study psychology. They fell in love and over the course of 2 years of internal struggle she and Luther decided to divorce. In 1928, she married Reo, who had changed his field of study to anthropology and had already taken his first field trip to the Dobu people in the Pacific. They immediately embarked on their first field trip together, to the Manus people of the Admiralty Islands in the Pacific. This resulted in her book, *Growing Up in New Guinea* (1930), a study of the socialization and education of Manus children. The monograph paired with this popular book was *Kinship in the Admiralty Islands* (1934).

She took two other field trips with Fortune. The first, a summer trip to the American West, resulted in her book *The Changing Culture of an Indian Tribe* (1932), which studied assimilation among the Omaha Indians. The second, a return to the Pacific, this time the mainland of New Guinea, resulted in *Sex and Temperament* (1935), a comparative study of three New Guinea peoples, the Arapesh, Mundugumor, and Tchambuli, which gave evidence for the first time of the constructed nature of gender systems. Prior to this, gender had been viewed as biologically innate. The monograph that

accompanied this popular book ultimately became the five-volume series called *The Mountain Arapesh* (1938–49). During this field trip, problems in their marriage came to a head, resulting in Mead and Fortune's divorce on their return from the field. In 1936, Mead married British anthropologist Gregory Bateson, whom she had met and fallen in love with in New Guinea, and they set out for their first field trip together to Bali in the Pacific, which resulted in their coauthored book, *Balinese Character: A Photographic Analysis* (1942). After this field trip, their daughter and her only child, Mary Catherine Bateson, was born in 1939.

During World War II, Mead served as Executive Secretary of the National Research Council's Committee on Food Habits, which did serious research on nutrition planning and cultural food problems, such as how to help people used to rice use white flour. She wrote a pioneering popular book, *And Keep Your Powder Dry* (1942), the first by an anthropologist to attempt to use anthropological insights and models to deal with a nonindigenous culture, that of the United States as a whole.

After the war ended, she became involved with Ruth Benedict in a project studying nonindigenous cultures around the world titled Research in Contemporary Cultures, popularly known as National Culture studies. When Benedict died in 1948, Mead took over as director of the project and brought it to a conclusion in 1952. Sometime in the late 1940s, after Benedict's death, Mead also began a long-term relationship with anthropologist Rhoda Metraux, which lasted until the end of her life. In 1949, using information from the eight cultures she had done fieldwork in, she wrote *Male and Female: A Study of the Sexes in a Changing World*. In 1950, she and Gregory Bateson divorced, the war having drawn them apart.

In the 1950s, Mead decided to revisit the Manus in the Pacific, to see what a difference 25 years had made for them. The result was *New Lives for Old: Cultural Transformation – Manus, 1928–53* (1956). This began a pattern of revisiting old fieldwork sites that continued almost until her death that resulted in several other shorter publications.

Mead wrote many other books and articles after World War II. Increasingly she emphasized working with colleagues outside anthropology for interdisciplinary insights, such as psychoanalyst Martha Wolfenstein, with whom she edited *Childhood in Contemporary Cultures* (1955) and with photographer Ken Heyman in *Family* (1965), or those within anthropology willing to do interdisciplinary work, such as Ray Birdwhistell who studied kinesics, or body language as it was popularly known. She examined topics such as how communication processes affect cultural change. She studied ekistics, the science of human settlements.

In 1961, she agreed to write columns for *Redbook Magazine*, a popular woman's magazine in the United States. These began to appear in 1962. This, coupled with her many appearances on television talk shows and radio, along with her lectures around the United States and throughout the world, made her a celebrity. She retired as a full curator at the American Museum of Natural History in 1969.

In 1970, she published *Culture and Commitment: A Study of the Generation Gap*, one of her most successful popular books after World War II, and in 1971 published a conversation between herself and African-American author James Baldwin, titled *A Rap on Race*. Her autobiography, *Blackberry Winter*, appeared in 1972. Still the most popular anthropologist in America, she developed pancreatic cancer and died in New York City, on 15 November 1978, at the age of 76.

Mead was most noted in the profession of anthropology as a founder and leader of the American 'Culture and Personality' school of anthropology, which, in the 1930s and 1940s, tried to meld interdisciplinary, particularly psychological, ideas and methodologies with those of anthropology. Mead pioneered the use of psychological testing among indigenous populations, worked with psychologists and psychoanalysts to develop or test theories, and fostered a wide variety of interdisciplinary connections throughout her life. After World War II, she also became known as an advocate of anthropological studies of nonindigenous peoples. She was recognized for her pioneering use of audio–visual techniques in the field, and for her continuing efforts to develop field training and new techniques for use in the field.

During her lifetime, Mead played a leadership role in anthropology and science. She was elected to membership in the National Academy of Sciences in 1975, and that same year became president of the American Association for the Advancement of Science. She had been a founder and president of the Society for Applied Anthropology (1940), and president of the World Federation of Mental Health (1956–57), and the American Anthropological Association (1960). She received over 40 awards during her lifetime, including the Viking Medal in general anthropology and, after her death, the Presidential Medal of Freedom.

In American culture and to the world at large, at the time of her death Mead was a celebrity, called in a *New York Times* editorial, 'the grandmother of the world.' She was listened to as an expert on child-rearing, the family, education, ecology, community building, and women's issues. She coined the phrase 'postmenopausal zest,' for the renewal and refocusing of women's energies after menopause. She was not afraid to be controversial, advocating trial marriage, the decriminalization of marijuana, and women's rights in the 1960s and 1970s.

But she had her critics, both during her lifetime and after. In her long career, Mead was criticized as well as lauded by other anthropologists for her popularizing efforts and for her interdisciplinary initiatives. Feminist Betty Friedan, in her 1963 book, *The Feminine Mystique*, accused Mead of falling into biological determinism in her book, *Male and Female*, and retarding the cause of women's rights through the 1950s. By the 1970s, educated New Guineans were beginning to look at her fieldwork critically as a product of colonial thinking. This was part of a general movement among former colonial peoples concerning anthropologists who had done fieldwork in their countries. By the 1990s, Mead's work, along with all the classic ethnographies, was being seriously critiqued by both indigenous scholars and Western scholars – in her case, in New Guinea she was criticized for a pro-Western stance in a time of colonialism and imperialism (Foerstel and Gilliam, 1992). Questions of implicit racism in her work were raised as well (Newman, 1999, 2004).

But the quality of her anthropological fieldwork was not seriously questioned until 5 years after her death. Then began what is now called the Mead–Freeman controversy, one of the longest standing intellectual controversies within academia, lasting 18 intense years until Freeman's death in 2001. There seems to be no one willing to keep his side going, and histories

of the controversy are now beginning to appear (Shankman, 2009). In his (1983) book, *Margaret Mead and Samoa: The Making and Unmaking of an Anthropological Myth*, Freeman attacked Mead's fieldwork in *Coming of Age in Samoa*, accusing her of total misunderstanding of Samoan culture and of describing Samoan culture and customs inaccurately. He depicted her as a young, naive girl deceived by Samoans who lied to her, or made up stories. He suggested that, as a young woman eager to succeed in the eyes of her mentor, Franz Boas, she skewed her work to prove his theories. While some took his side (e.g., Appell, 1984), Freeman's accusations stunned American anthropologists, many of whom critiqued Freeman and defended Mead. For example, American anthropologist Lowell Holmes, who had restudied Ta'u, a village where Mead had worked in American Samoa, while he had his own critiques of her work, saw her study overall as generally sound for its time and place (Holmes, 1983, 1987). Several researchers saw Samoan culture as containing elements of both Freeman's and Mead's studies (e.g., Shore, 1983; Scheper-Hughes, 1984; Feinberg, 1988). Others felt that both had overgeneralized (e.g., Côté, 1994; Orans, 1996). Many noted the differences between Western Samoa, where Freeman had done his fieldwork in the 1940s and 1960s, and American Samoa, where Mead did hers in the 1920s, and the role of the ethnographers' gender, age, and personality (e.g., Ember, 1985; Holmes, 1987; Côté, 1994). Several also noted Freeman's selective quoting and quoting out of context from other anthropologists, other writers on Samoa, and from Mead's work, which misrepresented these people's positions (e.g., Weiner, 1983; Holmes, 1987; Leacock, 1988, 1992; Shankman, 1996, 2009).

The argument continued for so long, partly because Freeman's accusations tapped into deeper questions anthropologists were especially wrestling with by the late 1980s and 1990s, and writing about Mead's or Freeman's work was a way to focus on them. One of these issues was a reevaluation of the roles of biology and culture within anthropology and social science. Freeman was seen as championing biological determinism and attacking Boas and Mead as cultural determinists (McDowell, 1991: C21; Levy, 1983; Strathern, 1983; Leacock, 1992). Another was the question whether anthropology was indeed a science or belonged instead in the humanities. Freeman's definitions of science and objectivity as he questioned Mead's work, in turn were questioned, and gave others an opportunity to reflect on the nature of science, and of objectivity, and on the nature of truth (e.g., Ember, 1985; Patience and Smith, 1986). Mead and Freeman also became part of a debate on ethnographic authority, and the nature of anthropology. The usefulness of ethnographies came under fire by the mid-1980s, as they were seen as artificial constructs, someone's idea of a different people, not necessarily the truth about a people (Clifford and Marcus, 1986). By the mid-1980s, other anthropologists had done fieldwork in Samoa: Lowell Holmes, James Côté, Martin Orans, and Paul Shankman. They criticized Mead, and found her work dated, but not, as Freeman did, totally wrong. Freeman had no room for a relative view of fieldwork. He believed there could only be one truth, not many, and his was the correct vision.

Thus the controversy also continued partly because Freeman worked to keep it alive. From 1983 through the next 18 years until his death in 2001, Freeman wrote at least 20 replies to his critics in academic and other journals, promoted a movie that espoused his views, *Margaret Mead and Samoa* (1988) and in 1996 promoted a play about himself by Australian playwright David Williamson, titled *Heretic*. Then in 1999 his second book on Mead came out, *The Fateful Hoaxing of Margaret Mead* (1999), in which his major argument was that Mead got her material on sexuality from two young girls who lied to her. They may have lied, or shaded the truth to Freeman, but it does not matter because other evidence in Mead's papers at the Library of Congress shows that she got her information from a variety of sources, not just the two girls highlighted in Freeman's book. At least one letter to Boas at the time shows that she knew of the potential of people lying to her.

Because she was such a celebrity in American society by the time of her death, the American mass media picked up the controversy quickly and spread it widely around the country. The first story actually came out in the *New York Times* 2 months before the book was published. Responses to Freeman's accusations lagged far behind the message that he had proved her wrong. The critique of Freeman's accusations largely took place in academic publications not read by the public (Shankman, 2009). Today, after almost 30 years, most anthropologists see her work as professional, if limited by time and place, but much of the general public ascribes to Freeman's view that he proved her wrong.

The year that Freeman died, 2001, was also the centennial of Margaret Mead's birth. Events were planned and carried out from the Library of Congress to California. This allowed the continuation of a reevaluation of Mead both personally and professionally in the following years, and the scholarship has generally ceased to focus on the Mead–Freeman controversy. The relationship between Ruth Benedict and Mead, privately and as colleagues learning from each other, became a subject of serious investigation (Lapsley, 1999; Banner, 2003). A collection of articles concerning Mead and Benedict focusing on 'feminism, race, and imperial visions' came out a year later (Banner and Janiewski, 2004). A volume of Mead's selected letters revealed hidden aspects of Mead's life and the people important to her (Caffrey and Francis, 2006). Another book examined Mead's celebrity and the images throughout her life which helped her maintain it; her role as an American Icon and as a 'representative figure,' meaning more to people in American culture than the sum of her own achievements during her lifetime (Lutkehaus, 2008). Another direction that the scholarship has taken so far is to review Mead's anthropological works from the 1920s to 1930s for what they tell readers about American culture rather than the culture of a non-American people (Molloy, 2008).

Mead left a legacy that includes the ideas of the relativity of adolescence and sexual mores that differ from culture to culture; the relative shape of the family, ranging from the nuclear, to widely extended by blood and adoption; the idea of the cultural construction of gender, of masculinity and femininity, as categories of human identity; and the connection between child-rearing practices and the shape of a culture.

See also: Anthropology, History of; Anthropology: Overview; Benedict, Ruth (1887–1948); Boas, Franz (1858–1942); Childhood: Anthropological Aspects; Everyday Life, Anthropology of; Gender: Anthropological Aspects; Kinship in

Anthropology; Oceania, Sociocultural Overviews: Polynesia and Micronesia.

Bibliography

Appell, George N., 1984. Freeman's refutation of Mead's coming of age in Samoa: the implication for anthropological inquiry. The Eastern Anthropologist 37, 183–214.

Banner, Lois, 2003. Intertwined Lives: Margaret Mead, Ruth Benedict, and Their Circle. Alfred A. Knopf, New York.

Banner, Lois, Janiewski, Dolores (Eds.), 2004. Reading Benedict, Reading Mead: Feminism, Race, and Imperial Visions. The Johns Hopkins University Press, Baltimore.

Caffrey, Margaret M., Francis, Patricia (Eds.), 2006. To Cherish the Life of the World: Selected Letters of Margaret Mead. Basic Books, New York.

Clifford, James, Marcus, George E., 1986. Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley, CA.

Côté, James, 1994. Adolescent Storm and Stress: An Evaluation of the Mead–Freeman Controversy. L. Eribaum Associates, Hillsdale, NJ.

Ember, Melvin, 1985. Evidence and Science in Anthropology: Reflections on the Freeman-Mead Controversy. American Anthropologist 87, 906–910.

Feinberg, Richard, 1988. Margaret Mead and Samoa: coming of age in factor fiction. American Anthropologist 90, 656–663.

Foerstel, Lenora, Gilliam, Angela (Eds.), 1992. Confronting the Margaret Mead Legacy: Scholarship, Empire, and the South Pacific. Temple University Press, Philadelphia.

Freeman, Derek, 1983. Margaret Mead and Samoa: The Making and Unmaking of an Anthropological Myth. Harvard University Press, Cambridge, MA.

Freeman, Derek, 1999. The Fateful Hoaxing of Margaret Mead: A Historical Analysis of Her Samoan Research. Westview Press, Boulder, CO.

Friedan, Betty, 1963. The Feminine Mystique. W.W. Norton & Company, Inc., New York.

Gordon, Joan (Ed.), 1976. Margaret Mead: The Complete Bibliography, 1925–75. Mouton, The Hague.

Heimans, Frank, 1988. Margaret Mead and Samoa [videorecording, 52 min.]. Brighton Video, New York.

Holmes, Lowell D., 1983. A tale of two studies. American Anthropologist 85, 929–935.

Holmes, Lowell D., 1987. Quest for the Real Samoa: The Mead/Freeman Controversy & Beyond. Bergin & Garvey Publishers, South Hadley, MA.

Howard, Jane, 1984. Margaret Mead: A Life. Simon and Schuster, New York.

Jacknis, Ira, 1988. Margaret Mead and Gregory Bateson in Bali: their use of photography and film. Cultural Anthropology 3, 160–177.

Lakoff, Andrew, 1996. Freezing time: Margaret Mead's diagnostic photography. Visual Anthropology Review 12, 1–18.

Lapsley, Hilary, 1999. Margaret Mead and Ruth Benedict: The Kinship of Women. University of Massachusetts Press, Amherst, MA.

Leacock, Eleanor, 1988. Anthropologists in search of a culture: Margaret Mead, Derek Freeman, and all the rest of us. Central Issues in Anthropology 8, 3–23 (1992. In: Foerstal, L., Gilliam, A. (Eds.), Confronting the Margaret Mead Legacy. Temple University Press, Philadelphia, PA).

Leacock, Eleanor, 1992. Anthropologists in search of a culture: Margaret Mead, Derek Freeman, and all the rest of us. In: Foerstel, Lenora, Gilliam, Angela (Eds.), Confronting the Margaret Mead Legacy: Scholarship, Empire, and the South Pacific. Temple University Press, Philadelphia, pp. 3–32.

Levy, Robert, 1983. The Attack on Mead. Science 220, 829–832.

Lutkehaus, Nancy, 2008. Margaret Mead: The Creation of an American Icon. Princeton University Press, Princeton.

McDowell, Nancy, 1991. The Mundugumor: From the Field Notes of Margaret Mead and Reo Fortune. Smithsonian Institution Press, Washington, DC.

Molloy, Maureen, 2008. On Creating a Useable Culture: Margaret Mead and the Emergence of American Cosmopolitanism. The University of Hawaii Press, Honolulu.

Newman, Louise, 1999. White Women's Rights: The Racial Origins of Feminism. Oxford University Press, New York.

Newman, Louise, 2004. Coming of Age, but Not in Samoa: Reflections on Margaret Mead's Legacy for Western Liberal Feminism in Dolores Janiewski and Lois Banner. In: Janiewski, Dolores E., Banner, Lois W. (Eds.), Reading Benedict, Reading Mead: Feminism, Race, and Imperial Visions. The Johns Hopkins Press, Baltimore.

Orans, Martin, 1996. Not Even Wrong: Margaret Mead, Derek Freeman, and the Samoans. Chandler and Sharp Publishers, Novato, CA.

Patience, Allan, Smith, Joseph, 1986. Derek Freeman and Samoa: the making and unmaking of a Biobehavioral myth. American Anthropologist 88, 157–162.

Redfield, Robert, 1931. Review of social organization of Manu'a. American Journal of Sociology 37 (3), 511–512.

Scheper-Hughes, Nancy, 1984. The Margaret Mead controversy: culture, biology, and anthropological inquiry. Human Organization 43, 85–93.

Shankman, Paul, 1996. The history of sexual conduct in Samoa and the Mead/Freeman controversy. American Anthropologist 98, 555–567.

Shankman, Paul, 2009. The Trashing of Margaret Mead: Anatomy of an Anthropological Controversy. The University of Wisconsin Press, Madison, WI.

Shore, Bradd, 1983. Paradox regained: Freeman's Margaret Mead and Samoa. American Anthropologist 85, 935–944.

Strathern, Marilyn, 1983. The punishment of Margaret Mead. Canberra Anthropology 6, 70–79.

Weiner, Annette B., 1983. Ethnographic determinism: Samoa and the Margaret Mead controversy. American Anthropologist 85, 909–919.

Meaning and Rule Following

Hans-Johann Glock, Institute of Philosophy, University of Zurich, Zürich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

According to a venerable tradition in philosophy and linguistics, expressions have meaning through being subject to conventions or rules. This claim has become a central topic of contemporary philosophy of language and mind in the wake of Wittgenstein and Kripke, largely because the normativity of meaning is regarded as a serious challenge to naturalism. One reaction to this challenge is to deny that the normativity of meaning is genuine. While there are 'semantic principles' specifying conditions for the correct application of expressions, these are either not genuinely normative or they are not in fact constitutive of meaning. But this dilemma can be defused if one distinguishes different dimensions of semantic normativity and pays attention to different types of mistakes that can afflict linguistic behavior. In particular, one needs to keep apart norms of truth and norms of meaning, regulative, and constitutive rules, and to appreciate the normative dimensions of the notion of correctness. The final section discusses the role of communally *shared* rules or conventions for language, in particular the age-old problem of circularity: how could such conventions be essential to language, given that the latter appears prerequisite for establishing and communicating conventions in the first place.

The idea that language is a *rule-guided activity* is as old as the study of language. It must not be confused with two other views about language involving normativity. We should distinguish:

supremacism: specific natural languages are superior to others, as in the idea that some languages are 'logical,' complex, or advanced, whereas others are 'primitive,' or in myths about a primordial or 'divine' language.

prescriptivism: one variety of a particular natural language (standard or written language) is inherently superior to all others and ought to be adopted by the whole community.

normativism: any form of language, including slang, dialects or even idiolects, is guided by rules, and it is the task of linguistics to identify the rules actually in force in a linguistic community.

Supremacism has been discredited by comparative linguistics (Sapir, 1921), and prescriptivism has few contemporary adherents (e.g., Dummett, 1993). By contrast, normativism continues to dominate among philosophers and linguists. The topic is not just pivotal to an adequate understanding of language; it epitomizes two fundamental controversies in contemporary thought.

The first is the conflict between *naturalists* and *normativists*. Naturalists believe that there is no genuine knowledge outside natural science (epistemological naturalism), and that the spatiotemporal phenomena studied by natural science exhaust what there is (ontological naturalism). Normativists have objected that logic, knowledge, and language depend on norms that cannot be captured by the factual descriptions and causal explanations of natural science.

The second is the conflict between *individualism* and *communitarianism*. According to individualists, the fundamental linguistic phenomenon is the idiolect, the language of an individual. According to communitarians, the fundamental phenomenon is that of a public language guided by shared conventions. Communitarians treat a language as a social practice or institution, which is more than the sum of individual speakers or idiolects. Both debates have profound repercussions for the behavioral and social sciences. Natural languages are the most striking and fundamental characteristics of human as opposed to animal communities. Whether they have an irreducible normative dimension is central to the question of what distinguishes us from animals. Furthermore, individualism is hospitable to the idea of an extremely fragmented society. We not only hold diverse beliefs and values, we need not even share a common language. Communitarians, by contrast, insist that we must share at least a linguistic medium in which to express conflicting viewpoints. And because they treat language as an intersubjective practice, they acknowledge and deplore phenomena such as 'language death' (Evans, 2010).

It is customary to distinguish different types of linguistic rules, following the orthodox division of semiotics: *syntactic* rules concern the combinatorial possibilities of linguistic expressions, in particular word arrangement within sentences; *semantic* or *lexical* rules determine the meaning or sense of words and sentences; *pragmatic* rules concern the use of words in communication, e.g., the speech-acts we perform by uttering sentences. This article focuses on semantic normativity.

Historical Background

Ever since Aristotle, there has been a broad consensus that words mean what they do as a result of human convention (*nomos*) rather than nature (*physis*). This consensus is epitomized by Peirce's distinction between three different types of signs – indices based on causal relations, icons based on similarity, and symbols based on conventions – and by Grice's (1989) dichotomy of natural and nonnatural meaning, as in 'these clouds mean rain' versus '*sesquipedalius* means "1 and ½ feet" in Latin.' The normative aspects of nonnatural meaning came to the fore through the later Wittgenstein. Wittgenstein attacked referentialist accounts. The meaning of an expression is not an object it stands for, but 'its use in the language' (1953: Section 43).

At the same time, Wittgenstein (like Chomsky afterwards) criticized behavioristic and causal theories of meaning (Ogden and Richards, 1923). These theories also explain the meaning of an expression by reference to its use. But the operative conception of use is purely causal. The meaning of a word is equated with the conditions, which *cause* a speaker to utter it and/or the *effects* that such utterances have on hearers.

Wittgenstein's alternative is a normative conception of use. The meaning of an utterance depends not on how it is actually understood, but on how *it is to be understood* within a linguistic community. What is semantically relevant about the use of an expression is neither the causes nor the effects of its utterance, nor the intentions with which speakers utter it, but its *correct* use. The linguistic meaning of expressions depends on general *rules*. These rules provide standards for the correct use of expressions, and thereby determine what it makes sense to say with the help of these expressions.

Kripke (1982; anticipated by Fogelin, 1976) portrays Wittgenstein as constructing a 'skeptical paradox' that questions the possibility of distinguishing between following and violating a rule, and thereby casts doubt on the very phenomenon of meaning. Furthermore, Kripke adopts a *communitarian* reading according to which Wittgenstein regarded rule following and language as inherently social. Kripke's tentative interpretation provoked hostile reactions from scholars (Baker and Hacker, 1984b; cp Kusch, 2006). The skeptical interpretation of Wittgenstein is wrong and the communitarin interpretation contentious at best (Glock, 1996). Nevertheless, there has been an extensive debate about 'Kripkenstein' on rule following, which is now conducted largely in blissful disregard for Wittgenstein's own writings (Miller and Wright, 2002). As regards the *normativity of meaning*, moreover, Kripke propounded ideas with Wittgensteinian roots and great substantive import.

The Kripkensteinian Paradox

According to Kripke's paradox, there is no set of facts that determines what rule R a speaker S is following in using an expression *e* and hence what S means by e. For all facts about the use of *e* by S up to a time t_1 are compatible not just with R but also with R′. I have been using '+' in the customary fashion, judging, for example, that $7+5=12$, $19+24=43$, etc. Accordingly, I appear to follow the rules of addition, and to mean the *plus* function by '+.' Inevitably, however, there are infinitely many sums that I have never performed. Suppose that I have never performed an operation '$x+y$' with x or y greater than 56 and am asked to calculate '$68+57$.' Everything about my use of '+' up to now is compatible with the possibility that I have been using '+' to mean the *quus* function, which is defined as follows:

- x quus $y = x+y$ for $x, y < 57$,
- x quus $y = 5$, in all other cases.

It is tempting to think that I meant plus in virtue of interpreting '$68+57$' as an instruction of the following kind: count out a set of 68 items; count out a set of 57 items; combine the two sets; count out the result. But what makes it the case that by 'counting,' I meant counting rather than quounting, an operation, which coincides with counting for numbers smaller than 57, yet yields 5 for all higher numbers? We have merely replaced a problem concerning a rule involving '+' by one concerning a rule involving 'count.' By pain of regress, rule following cannot be explained through invoking rules on how to apply rules.

The paradox questions neither the arithmetic truth that $68+57=125$ nor the reliability of my memory. At issue is not whether I can *know* that I meant plus by '+'; it is rather what facts, if any, *constitute* my following one rule rather than another and hence my meaning plus rather than quus. The threat is one of semantic *nihilism* rather than skepticism; it is directed against the very ideas of rule following and meaning. In the first instance, at stake is only my having meant plus by '+' *so far*. It must be assumed that at present I mean plus by '+,' since otherwise the paradox cannot even be formulated. However, if there are no facts constituting my having meant plus in the past, then there cannot be facts constituting my meaning plus at present. The paradox concerns the phenomenon of a speaker meaning something by *e* – *speaker meaning* – rather than the *lexical* or *literal* meaning of *e* in a shared language. But even if one rejects the Gricean project of reducing literal to speaker meaning, there *could not be* such a thing as literal meaning unless members of a linguistic community could, individually, mean something by their utterances. Finally, the threat extends beyond the meaning of linguistic expressions to the 'content' of thinking, i.e., to what subjects believe, desire, etc. If nothing constitutes S meaning *plus* by '+' in saying '$7+5=12$,' then nothing constitutes S *judging* that 7 *plus* 5 *equals* 12 rather than that 7 *quus* 5 *equals* 12.

For Kripke, the only escape is an indirect 'skeptical solution.' No fact constitutes following or violating a rule in using e and hence meaning one thing rather than another by e. Fortunately, there is a surrogate. The correct answer to the question 'What is $68+57$?' is '125' and '5' is an incorrect answer in the following sense: the former is accepted and the latter is rejected by the linguistic community of which I am a member. Consequently, rule following, meaning and thought are *inherently social*: an individual outside of the context of a community can neither speak meaningfully, nor think, nor follow rules of other kinds. Communal approval is not a condition under which statements about what (semantic) rules a speaker or a community is following are *objectively true*, but it is a condition under which such statements can be *warrantedly asserted*.

Whatever the merits of communitarianism, however, it proffers no solution of the paradox, skeptical, or other. For one thing, one cannot rule out a priori that a whole community might misapply a rule. For another, if nothing constitutes my meaning plus by '+,' nothing constitutes the members of my community meaning approval rather than disapproval by saying e.g., 'Correct!.' An 'antirealist' semantics revolving around assertion conditions is just as vulnerable to nihilistic vagaries as a realist semantics based on truth conditions.

Nihilism about meaning and content is patently self-refuting. If it were correct, it could neither be articulated nor entertained. This does not absolve us, however, from pinpointing where precisely nihilism goes astray. The most popular avenue has been to invoke the idea of a *disposition*. Following (understanding and accepting) a rule is a potentiality. I mean plus by virtue of the fact that I would

have answered '125' to the question. But dispositionalism cannot do justice to the normative character of rules in general and meaning in particular. Dispositions can sustain predictions about what *will* occur, yet not statements about what *should* occur or what is *correct* (Waismann, 1965: Chapters 7, 18; Kripke, 1982: 22–37).

To say, for instance, that '125' is the correct answer is not to say that it is the one, which a computer or I are predisposed to give. Computers can malfunction and I might make a mistake. Rather, we judge by reference to the rule whether the computer has broken down, or whether I am inclined to give the *right* answer. If '+' is to mean plus in a language, then something must count as the correct answer, irrespective of what individual speakers are inclined to answer. After all, speakers do not just use words, they also misuse them. Even if *I* am to mean plus by '+,' irrespective of lexical meaning, something must count as the answer *I should proffer*, irrespective of my disposition. Otherwise, I could not make a mistake relative to my understanding, since the answer I actually give is the one I am disposed to give.

There have been four distinct reactions to Kripkenstein. The first is *reductionism*. It tries to account for normativity in purely descriptive terms acceptable to naturalism. The most popular avenue has been to modify dispositionalism. It insists that the correct answer is the one I am predisposed to give *under ideal conditions*. Thus one might characterize the semantically constitutive dispositions as those, which are not subject to interference like disease or malfunction, or as those which are in line with the biological function and hence, the evolutionary history of the pertinent cognitive system (Millikan, 1990). Yet no set of conditions, however, propitious to an individual's performance or crucial to the emergence of the pertinent cognitive system, can *guarantee* that an answer is correct. Unless, that is, conditions *count as ideal* only if they engender the *correct* answer. In that case, however, the explanation of what constitutes correctness is circular.

The second reaction harks back to Wittgenstein. The 'paradox' according to which 'no course of action could be determined by a rule' is based on a 'misunderstanding'; it shows not that the rule leaves its applications undetermined, as nihilism has it, but that "there is an understanding of the rule which is *not* an interpretation, but is exhibited in what we call 'obeying the rule' and 'going against it' in actual cases" (1953: Sections 202, 189). There is no potential 'gap' between a rule and its application. Rather, the relation between the two is internal, i.e., constitutive of the relata. *Plus* simply *is* the function that yields 125 for the arguments 68 and 57. And to understand 'plus' simply *is* to know that 125 is the correct answer (Baker and Hacker, 1984b; Glock, 1996: 323–329). This reaction faces a further skeptical question: What makes it the case that I am following the *plus* rule (internally related as it is to its applications), as opposed to the *quus* rule (Hattiangadi, 2007: 172–177). And it must meet that challenge without appeal either to mere communal dispositions or treating rules as mysterious platonic entities that somehow contain all of their infinitely many applications.

A third reaction may be capable of evading this challenge. It maintains that the paradox arises from a misguided reductionist project of specifying nonnormative, purely natural facts that rule following, meaning, and thinking consist in (McDowell, 1984). These phenomena are (rooted in) human practices, but they are *sui generis*. While they can be elucidated by pointing out their connection to other normative phenomena such as explanation and understanding, they cannot be reduced to phenomena of a more fundamental, natural kind.

Finally, there is the option of denying that the normativity of meaning is genuine in the first place. It has become increasingly popular, especially among naturalists trying to defuse the challenge posed by normativity. Drawing on recent work by Davidson and Chomsky, antinormativists have questioned received views about the connection between meaning and rules. The remainder of this article is devoted to the lively ensuing debate.

Dimensions of Semantic Normativity

Norms in the relevant sense differ from statements of mere regularities or natural laws in three interrelated respects.

First, they do not (merely) describe or explain, but set standards, in particular, standards against which behavior can be assessed as correct or incorrect. If planets turn out to follow paths other than elliptical ones, it is Kepler's Laws rather than the movement of the planets which needs revision. By contrast, if I drive on the left-hand side it is my driving rather than the Traffic code, which is at fault. The direction of fit is from world (action) to mind (norm).

Second, although rule following presupposes a regularity in behavior, there is a difference between rule-guided and merely regular behavior. If an agent follows a rule in Φ-ing, the rule must be part of her reason for Φ-ing, and more specifically, for Φ-ing in a particular manner. Rule following is, therefore, a type of intentional or goal-directed behavior.

Third, 'to follow a rule' is an achievement verb. There is a difference between an agent actually following a rule and her merely believing that she does.

The first feature is equally present in desires, and all three are shared by commands, orders, and requests. A fourth feature sets norms apart, namely that they are inherently general. Although norms are commonly restricted to certain types of subject and types of situation (context sensitive), they govern an unlimited multiplicity of occasions (occasion neutral).

Finally, there is a difference between a rule and its linguistic expression – a rule formulation. A rule need not be expressed linguistically, and one and the same rule may be expressed in different ways. In short, a norm is a general standard that provides a reason for action and against which actions can be assessed as correct or incorrect.

Within the debate over the 'normativity of meaning,' we must distinguish various normativity claims. Three of them concern linguistic meaning.

Bare Normativity of Meaning

e is meaningful ⇒ [conceptually entails] there are conditions for the correct use of *e*.

If an expression *e* is meaningful, then there must be conditions for its correct application. "Suppose the expression 'green' means *green*. It follows immediately that the expression 'green' applies correctly only to these things (the green ones) and not to those (the non-greens). The fact that the expression means something implies, that is, a whole set of *normative* truths about my behaviour with that expression: namely, that my use of it is correct in application to certain objects and not in application to others" (Boghossian, 1989: 148).

Rule-Based Normativity of Meaning

e is meaningful ⇒ there are rules for the use of *e*.

Historically 'norm' means roughly the same as *rule*. However, 'normativity' has also been employed more widely, for all phenomena that can be assessed for *correctness* according to a *standard*, whether or not this involves rules, understood as standards guiding action. According to rule-based normativity of meaning (RNM), meaning requires rules of use, since the correctness conditions of bare normativity of meaning (BNM) are laid down by such rules.

Prescriptive Normativity of Meaning

e is meaningful ⇒ there are prescriptions for *e*'s use.

Prescriptive normativity of meaning (PNM) is stronger than RNM, since not all rules prescribe what one ought to do.

In contemporary debates, normativity claims that concern the meaning of linguistic expressions are intertwined with those that concern the *conten*t – the concepts or thoughts – expressed by their employment. One of them is Dummett's (1981: 300) proposal that the concept of assertion entails the existence of certain rules, notably that an assertion should aim to be true.

Normativity of Content: Assertion

a asserts that *p* ⇒ there are rules for asserting that *p*.

The other concerns the content of beliefs rather than assertoric utterances.

Normativity of Content: Belief

a believes that *p* ⇒ there are rules for asserting that *p*.

Accordingly, our general topic divides as follows (**Table 1**):

Table 1 Aspects of semantic normativity

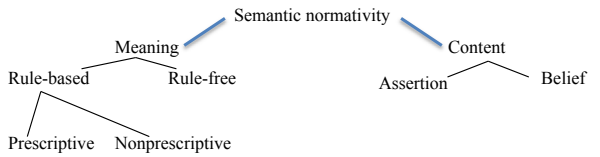

Finally, the normativity of meaning is connected to the issue whether meaning is a matter of conventions, i.e., of rules that are arbitrary and shared between members of a speech community.

Conventionality of Meaning

e is meaningful ⇒ there are conventions for using *e*.

Normativity, Truth, and Meaning (BNM)

Most antinormativists grant BNM. That an expression has a specific meaning implies that there are conditions for its *correct* use. At the same time, they confront normativism with a *dilemma*: either these semantic principles are *not really normative*; or they are *not in fact constitutive of meaning*, i.e., they do not follow from an expression being meaningful (Glüer and Wikforss, 2009). BNM is supposed to fall under the first horn; RNM and especially PNM under the second.

A popular way of capturing the idea that general terms require conditions of correct application is

$$(\mathrm{BNM}_1)\ e \text{ means } F \Rightarrow \forall x\ (\text{it is correct to apply } e \text{ to } x \leftrightarrow x \text{ is } f)$$

(where '*e*' is a general term, '*F*' gives its meaning, and '*f*' is that feature in virtue of which *e* applies to an object *x*). The consequent of BNM_1 provides the scheme for a 'semantic principle' of the kind employed in formal semantics. The first antinormativist objection runs: 'correct' in (BNM_1) is not genuinely normative, since it is simply a place holder for 'true'; 'it is correct to apply *e* to *x*,' in other words, boils down to *e holds true of x* (Boghossian, 2005; Hattiangadi, 2007: 51–61).

One cannot counter this objection by pointing out that semantic principles *give rise* to norms. The same holds for descriptive statements. For one thing, they can give rise to technical norms or rules, given additional premises of a normative kind. That compact fluorescent lamps consume less energy than standard lightbulbs may give rise to norms requiring the fitting of compact fluorescent lamps, given a norm requiring energy savings. For another, one can criticize statements and actions by reference to empirical statements no less than by reference to semantic principles.

At the same time, by dint of its connection with truth, holding true is what Parfitt (1997) calls 'normatively relevant.' A proposition being true (or a predicate holding true of something) can have implications for what it is correct to believe or do, implications that depend on fundamental features of human thought and action rather than specific normative premises. We need to modify, i.e., 'correct' our beliefs and activities in the light of what is the case, of things being thus-and-so. This, in turn, reflects the fact that how things is essential to how we should pursue our goals (Horwich, 1998: 44–46; Glock, 2003: 131–136).

Furthermore, the antinormativist objection assumes that semantic norms would have to be norms of truth, norms to the effect that one ought to think and state what is true. That assumption has also been accepted by most normativists (e.g., Blackburn, 1984: 281–282). But the contrast between correct and incorrect applies to language in a variety of ways, and their connection to the notion of meaning differs. In particular, one

can use a word meaningfully, in a sentence with a sense that expresses a thought, without that thought being *either* true *or* justified. This difference between truth and meaning emerges when we consider different types of mistakes that people can make in using words. First, in the run up to a by-election to the House of Commons a candidate once said,

1. I know that I cannot win this seat.

There was nothing wrong with (1) either linguistically or factually. Yet she made a dreadful mistake, because it is not politic to admit to such dim prospects. Next consider Tony Blair's statement in the House of Commons,

2. Iraq is capable of deploying weapons of mass destruction within 45 min of an order by Saddam Hussein.

There is nothing wrong linguistically with (2), yet it is both unjustified and false. Now consider George W. Bush's

3. You teach a child to read, and he or her will be able to pass a literacy test.

Here we are confronted with a linguistic mistake. But it is of a syntactic rather than semantic kind. In fact, once we correct for that error, (3) is analytic, at least on a certain understanding of 'teach' and 'literacy.' In spite of involving a solecism, therefore, this example shows Bush at the very height of his semantic powers. That is more than can be said of

4. I am mindful not only of preserving executive powers for myself, but for my predecessors as well.

Here Bush's travails are semantic. The statement (4) betokens not lack of factual knowledge, but an insufficient grasp of the meaning of at least one of the two terms 'preserving' and 'predecessors.'

There is a fundamental contrast between saying something false or unjustified – as in (2) – and saying something meaningless – as in (4). Some uses of words are mistaken *solely* because of *what these words mean*, irrespective of any other facts, syntactic rules, or social expectations. Conversely, one can apply a word in a way, which is *semantically correct* – based on a proper understanding of its meaning – without applying it correctly in the sense of saying something *true*, namely if one errs about pertinent facts (but see Whiting, 2013).

The meaning of *e* is determined by a rule, which specifies the conditions *f* under which *e* applies to *x*. Whether a speaker S violates that rule depends not on whether *x* actually satisfies *f* – that is a matter of fact/knowledge rather than meaning/understanding; it depends instead on whether S applies or withholds *e* according to whether *x* satisfies *f*.

For Wittgenstein, whether a sentence expresses a rule does not depend on its linguistic form, but on whether it has a normative function, either in general or on a given occasion of utterance. Rule formulations need not contain a deontic verb. And formulations of linguistic rules need not be metalinguistic statements that mention the expressions for which they provide a rule. By these lights, all explanations of meaning are semantic rules.

5. The term 'drake' means *male duck*.
6. A drake is a male duck.

Both (5) and (6) are standardly used to explain what the term 'drake' means. Because they function as explanations, they also have a normative role: we can criticize and justify our use of 'drake' by reference to them.

Facts of Use, Correctness, and Normativity (BNM)

Focusing on the explanations we actually employ in teaching, justifying, and criticizing the use of expressions (dictionary explanations included) avoids serious distortions (Glock, 2003: Chapter 5). If the normativity of meaning is instead discussed by reference to formal semantic principles the latter must at least be suitable as explanations of meaning. For 'drake,' such a principle might run:

7. It is correct to apply 'drake' to *x* iff *x* is a male duck.

Antinormativists concede

$$(\mathrm{BNM}_2)\ \text{'drake' means what it does} \Rightarrow (7)$$

At the same time they deny that (7) has a bona fide normative status, a first objection is that the alleged semantic norms boil down to mere 'descriptive facts' concerning linguistic behavior (Glüer, 2001). But we must distinguish between (7), which standardly expresses a norm, and something like

8. In English, it is correct to apply 'drake' to an object *x* iff *x* is a male duck.

(8) be used to invoke a norm within the Anglophone community. Yet it can also be used as an empirical 'norm-proposition' to the effect that this community follows that norm (von Wright, 1963: viii, 108). In that capacity, (8) states a fact about the behavior of Anglophones. It does so, however, only because the Anglophone speech community accepts a certain norm, a standard for the correct use of 'drake.' That is to say, (8) is a true factual statement about a norm precisely because (7) expresses a norm.

One could insist, however, that (8) overstates the normativist case (see Davidson, 2005: 109–112). All we really have is

9. Among a certain population, most people tend to apply 'drake' to an object *x* only if *x* is a male duck.

(9) is indeed bereft of normative implications. But there is more to our practice than statistical regularity. We not only tend to apply 'drake' to male ducks. We also explain the term as a label for male ducks, and we criticize and justify our use of the term by reference to statements like (5)–(7). Applying 'drake' to *x* on the grounds that *x* is a male duck or a female fox counts as (semantically) incorrect, a sign of deviant understanding.

It is widely assumed that notions like 'correct' or 'mistaken' are normative (Blackburn, 1984: 286–287), even among those skeptical about the normativity of meaning (Boghossian, 2003: 35; Fodor, 2008: 205). A second antinormativist objection questions this assumption.

A serviceable explanation of correctness runs as follows: (COR) *x* is correct iff *x* is in accordance with an acknowledged standard.

This standard may be part of a specific system of evaluation appropriate to *x*, as when we speak of correct academic

dress. Or it may be a general standard expressing certain demands, e.g., on prudential or moral conduct. Either way, there must be a standard specifying conditions under which *x* counts as correct.

At this juncture it is tempting to argue that calling *x* correct is merely to state that it satisfies these conditions, and hence that the concept has no normative implications after all. However, in characterizing *x* as correct we do not just ascribe to *x* the *features* specified in the relevant standard. We also characterize *x* as *meeting a standard* of positive evaluation. The term 'correct' is not purely descriptive or factual, but *evaluative*. What is more, unlike other evaluative terms, e.g., 'good,' it carries the further implication that failure to meet this standard is not just undesirable, but provides grounds for *intervention*. After all, the verb 'to correct' means to set right, rectify, or amend. Applying *terms used in the standard* may be purely descriptive, but applying the term 'correct' is *not*. And to call something incorrect is not only to evaluate it negatively; if something is incorrect, then one has grounds for *correcting* it. No normative force attaches to the statement that *x* is not wearing a cap and gown. But given the standards of correct academic dress, not wearing them on an official occasion constitutes being dressed incorrectly, and is therefore liable to be corrected.

This lesson carries over to semantic normativity. According to antinormativists, "To say that some use of a term is 'correct' is … merely to describe it in a certain way—in light of the norm or standard set by the meaning of the term" (Hattiangadi, 2006: 225). To categorize something by reference to the defining features of a drake has no normative implications per se. To deny that an object *x* possesses the defining features of a drake, for instance, is not to characterize *x* as fit for correction. At the same time, the 'norm or standard set by the meaning' of 'drake' makes reference to these features, just as the standard for correct academic dress makes reference to cap and gown. And to categorize applying 'drake' to *x* on the grounds that *x* is a female fox as incorrect has direct normative implications, since it characterizes that application as fit for correction.

More generally, to assess an assertion or belief as false is to present it as fit for epistemic correction. Equally, to assess an utterance as meaningless, nonsensical or a malapropism is to present it as fit for semantic correction. To be sure, on particular occasions a speaker may aim to make false assertions, to commit linguistic mistakes, or to utter nonsense. In that case, what she does may be correct relative to a different standard – a prudential or moral one, for example. If that standard is more important in the situation, what she does may also be correct in an 'all things considered' sense. But this does not detract from the fact that she is violating semantic or epistemic norms. In semantic principles like (7), 'correct' is not used in the all things considered sense. Conversely, the general norms of reason presupposed in that usage are not constitutive of the meaning of specific terms (although they feature in the semantic norms governing terms like 'reasonable' or 'rational').

Meaning, Rules, and Prescriptions (RNM, PNM)

Even if (7) has a normative status vis-à-vis linguistic performances, RNM may be wrong in portraying it as a genuine rule (Stainton, 1996: 139–142; Glüer and Pagin, 1999). After all, it does not oblige or even entreat us to apply the term 'drake' to an object in a particular situation.

However, there is also the 'definitory aspect of rules' (Baker and Hacker, 1984a: 259). Some rules do not prescribe a certain form of behavior, but instead lay down what a thing must be like to satisfy a certain description. Prime examples are the beloved EU norms concerning, for example, what qualifies as chocolate or the rules of games like chess. These norms lay down *what counts as* chocolate or as castling, without requesting anyone to produce or buy chocolate, or to castle.

This point is connected to a well-known distinction between *regulative* and *constitutive* rules (Searle, 1969). Regulative rules advise one on how to pursue in an optimal fashion an activity that can be specified independently of the rule. Constitutive rules are partly definitive of the activity in question. What goes on at dinner parties can be described as eating independently of the rules of etiquette. By contrast, what goes on when people move certain pieces across a board cannot be described as castling, checking, mating, or, more generally, as playing chess, independently of rules like 'The king cannot castle out of a check' or "The king is mate if it is in check in such a way that no move will remove it from being in check." This contrasts with a regulative rule of chess like 'Do not relinquish control of the center.' Even someone who disregards all these regulative rules does not cease to play chess, he will just play chess badly.

Constitutive rules – especially those of legal systems and games – make up a paradigmatic type of rule. But do semantic principles qualify? It might be thought that someone who violates a constitutive rule of chess is no longer playing chess, while someone who violates the alleged constitutive rules of English is still speaking English. This qualm is dubious on both grounds. It is implausible to insist that a game involving a single incorrect move no longer counts as chess. Conversely, although someone who commits a mistake may still speak English, the *mistake itself* does not count as English. It is quite common to react to a serious mistake by protesting 'That's not English.' And someone who commits many such mistakes can be said not to be speaking English. What is true is that the margin of tolerance is greater and less well defined in the case of natural languages than in the case of chess. There is a simple though important reason for this: the constitutive rules of natural languages do not form a precise and canonized system.

A final worry is that semantic principles lack the action-guiding force of genuine rules. The latter, it is often maintained, have to be deontic statements about what one *ought to do* (Hattiangadi, 2006; Boghossian, 2003). As we have seen, however, by laying down conditions under which terms of a language can be applied, semantic principles *also* lay down conditions under which certain utterances are *correct*. For instance, by specifying what counts as a drake, explanations like (5) or (6) *also lay down* what counts as using the term 'drake' in a correct, in the sense of meaningful, way – a point which comes out in statements like (7). Such statements support *prohibitions*. We can justify as

10. You ought not to apply 'drake' to a female fox or a male duck.

by reference to (7). More generally, we often say

11. You shouldn't call that …, because it's not ….

Finally, even among regulative rules, there are not just *prescriptive* 'norms of obligation' (von Wright, 1963) or 'mandatory norms' (Raz, 1978) of the form

(MN) x ought to ϕ in Condition C

There are also 'permissive norms' of the form

(PN) x is entitled to ϕ in condition C

Entitlements presuppose a background of prohibitions. Such a background is provided by injunctions like (10) and (11), even though the immediate sanctions for violating them may be confined to correction and ridicule. Some theorists have suggested that permissions reduce to obligations, either because they are the mere absence of obligations or because they are obligations on others not to interfere. The second claim does not pose a threat to regarding permissions as rules, since it merely relocates their normative trajectory. And the first ignores two points. We can invoke permissive norms in defense of Φ-ing, and this goes beyond merely averring that there is no proscription on Φ-ing. Furthermore, we can adduce these norms to explain our Φ-ing even in the absence of any challenge.

It is common to employ explanations like (5)–(6) as permissive norms. I can appeal to a definition like (5) in defending or explaining my use of 'drake.' If challenged why I call a male duck a drake I can defend myself by saying that 'drake' means *male duck* (Glock, 2005: 234–235; Whiting, 2009). Furthermore, explanations of meaning also function as transformation or inference rules. Thus Wittgenstein urged that ostensive definitions are substitution rules. 'This ☞ is red,' for instance, entitles one to pass from 'My bike has this ☞ color' to 'My bike is red' (1974: 88–91, 202).

The idea that transformation rules are constitutive of meaning was developed by inferential role semantics (Sellars, 1963; Brandom, 1994). The statement (5) licenses transitions between empirical propositions, e.g., from 'This creature is a male duck' to 'This creature is a drake' and vice versa. In other words, we can appeal to (5) in support of an inference like,

12. Donald is a male duck, therefore Donald is a drake.

And this inference is licensed by the semantic rule alone, independently of facts.

A final criticism of PNM revolves around the idea that genuine norms must be or imply *categorical imperatives*. Unlike merely hypothetical imperatives, such imperatives are independent of any desires or goals that the addressee might have. Now, a speaker can violate semantic prohibitions like (10) by lying. Yet this only shows that they are genuine rules, something one can obey or fail to obey. What disqualifies semantic prohibitions from being categorical is that they can be overridden by other rules *in certain situations*. Kant notwithstanding, however, this arguably holds for *all* imperatives.

In any event, that an imperative is merely hypothetical does *not* entail that it lacks normative force altogether. The opposite view is fueled by the conviction that a genuine 'ought' must indicate an 'all things considered' obligation. Yet in our normative discourse we recognize the possibility of bona fide yet prima facie or *pro tanto* obligations. We frequently engage in deliberations like,

13. On the one hand I ought to Φ, on the other I should Ψ in cases in which we cannot both Φ and Ψ.

RNM would be threatened if injunctions like (10) could be overridden simply by 'a desire to lie or deceive' (Hattiangadi, 2006: 231–232; Glüer and Wikforss, 2009: 2). But lying is something one ought not to do, since it undermines the trust essential to human communication and interaction (Glock, 2005: 237–238). Although the prohibition to lie can be overridden by more pressing obligations in certain situations, its validity does *not* depend on any desire or decision on the speaker's part. One is not entitled to lie simply because one feels like it.

Normativity of Content (NCA, NCB)

The issue of lying is also pertinent to normativity of content: assertion (NCA). A rule that is often regarded as implied by the possibility of assertion

14. $\forall p$ (one ought to assert that p only if p)

But to assert what one believes or knows to be false does not straightforwardly undermine the very possibility of assertion (Boghossian, 2003: 39). Otherwise lies could not be assertions. On the other hand, a norm can be constitutive of a practice in spite of being occasionally violated. From the fact that something can *sometimes* be the case, it does not follow that it can *always* be the case.

In this vein, Aquinas and Kant contended that a practice of uniformly insincere assertions cannot be consistently conceived. One possible rationale runs as follows. A speaker can only assert something by uttering 'p' if she *presents herself* as thereby expressing something she believes. Statements like 'p, but I do not believe that p' or 'I assert that p, but I do not believe that p' are self-defeating. Furthermore, spare a thought for a speaker who persistently says that p without believing that p. If this were common knowledge, i.e., if it were known to everyone and also known to everyone – including the speaker – that it was known to everyone, her utterances would no longer count as assertions. And if this scenario extended to all speakers, the practice of assertion would collapse.

Dummett has further maintained that there is a convention whereby in making an assertion speakers aim at truth, just as there is a convention according to which in making a move in chess players aim at the stipulated point of the game, e.g., checkmating in chess (1981: 296–298). But one must not take the analogy between a language and a game too far. Although we complain about insincere assertions, our complaint makes no reference to the constitutive rules of a specific language. We do not say "In French one doesn't lie, whereas in English it's ok", as we might say, "In draughts, one doesn't mate, whereas in chess one does" (Schroeder, 1998: Section 21).

There is a crucial difference between NCA and normativity of content: belief (NCB). The question 'Should I believe the truth' cannot even arise, whereas the question 'Should I tell the truth' often does. If the person accepts that a belief is false, she thereby ceases to have the belief. But does this mark a special kind of normativity or instead

cast doubt on NCB (see Boghossian, 2003; Zalabardo, 2013)?

According to a traditional principle, ought implies can. It must be within the agent's power to do what she ought to do, but *also* within her power to decide *against* doing it. But it is not within my power to decide to believe something that I take to be false (Williams, 1973: Chapter 9; cp McHugh, 2012). Whereas I have an obligation to refrain from asserting what I take to be false, I cannot have this kind of obligation vis-à-vis my beliefs.

Admittedly, NCB can be taken in a weaker sense. That we ought to Φ can simply mean that we can be held *responsible* for not Φ-ing. This notion applies beyond actions that one can refrain from performing. Under certain circumstances we can be held responsible for our character traits and moods, even though these are not actions that we can deliberately perform or refrain from performing. In a similar fashion, we can be held responsible for our beliefs. But this responsibility can only underpin 'subjective' norms such as

15. ∀p (one ought to believe that p only if it is reasonable to believe that p)

not 'objective' norms. For we cannot be held responsible for believing something that is false, provided our belief is reasonable, e.g., based on convincing evidence. Whereas truth is decided by what is the case, the responsibility we have for our beliefs reaches no further than the best available reasons we have for them. A belief must allow of being 'objectively' assessed for truth. But only the way we heed the reasons for or against a belief is subject to the type of action-guiding rules that apply to assertion.

Meaning and Convention (CM)

For present purposes, conventions are social rules that are arbitrary (Glock, 2010). Expressions do not have an intrinsic meaning. Rather, what they mean is determined by how they are used. As a result, what a linguistic expression means is arbitrary in the following respects:

- In different languages, different words are used to mean one and the same thing.
- In some languages the same linguistic form is used to mean different things (e.g., 'link' in English and German).
- We could, in principle, use words to mean something different from what they in fact mean.

The debate between individualists and communitarians turns on whether arbitrary semantic principles must be shared between speaker and hearer, or, more generally, between the members of a linguistic community. Individualist opponents of the conventionality of meaning (CM) argue as follows: if the meaning of an expression *e* in a language L is determined or constituted by the use speakers make of *e*, the meaning of *e* cannot at the same time prescribe a certain use to the speakers of L. Accordingly, the idea that the use of *e* is governed by shared lexical norms is a lapse into the Platonist picture according to which "the meanings of words are magically independent of the speaker's intentions" (Davidson, 1990: 310).

But one must distinguish between the use an individual speaker makes of *e*, and the use that the linguistic community makes of it. *Communal* use may *constitute* meaning, while *individual* use is *responsible* to it. From this perspective, the existence of lexical norms and their independence from the utterances and intentions of individual speakers is no more mysterious than the existence of legal norms and their independence from the acts and motives of individual agents.

However, while an appeal to conventions can explain legal norms, by pain of regress it cannot explain language in the same manner, because articulating conventions presupposes language (Russell, 1921: 190; Davidson, 1984: 280). Nevertheless, given that there is a difference between a rule and its linguistic formulation, there can be *implicit* rules that guide our communicative behavior without having been enunciated or being consciously consulted (Brandom, 1994: Chapter 1; Glock, 2003: 92–95, 247–248). Language probably arose when cooperative activities put a premium on the intentional and regular use of signals, a communicative practice that gradually became subject to normative reactions and sanctions (Tomasello, 2008). But it remains to be shown how this process can be conceptualized and explained without assuming that the subjects possess mental faculties of a kind that in turn presuppose language.

Turning from the phylogenesis of language to its nature, individualists standardly grant that we tend to speak as others do, while insisting that this is for extrinsic social reasons that are inessential to linguistic communication. Knowledge of shared rules is neither necessary nor sufficient for successful linguistic communication. For malapropisms, spoonerisms, and similar infelicities are not merely common, but commonly understood (Davidson, 2005: Chapters 8–9). In appropriate circumstances, we have no difficulty in understanding Mrs Malaprop's remark "A nice derangement of epitaphs" as meaning "a nice arrangement of epithets." The question remains whether such understanding does not rely on a background of shared conventions after all. For one thing, from the fact that people can deviate sometimes, it does not follow that they could always deviate. It is difficult to see how we might all be constantly committing malapropisms without communication breaking down. For another, while we are often, albeit by no means always, capable of understanding semantic infelicities, this may be limited to recognizable and, in some way, intelligible deviations from a shared norm. Finally, in the case of malapropisms, what we understand is what speakers *meant to say*, not what *they actually said* or the literal meaning of the expressions they uttered.

Individualists point out that communication can succeed even between individuals who do not share a language, even passively. In interactions with foreigners, we sometimes communicate by way of facial expressions, gestures, and intonation. The moot question is whether there is or could be communication by *language* without semantic conventions, i.e., whether linguistic communication could be based solely on a mutual process of ingenious mutual interpretation. Our concept of language is not sufficiently clear cut to provide a definite answer. But verbal communication without conventions, and hence without literal meaning, would definitely be incapable of expressing the fine-grained thoughts we can express in natural languages. This is crucial to the unique

role such languages play for human beings. They allow us to express and communicate complex thoughts even to strangers, people of whose beliefs, intentions, and desires we are largely ignorant. Semantic conventions may not be part of the essence of linguistic communication, yet nevertheless required by the anthropological function language has for our species of social primates. Finally, the debate about shared intentionality (Gilbert, 1996) suggests that obligations can arise from the structure of a practice that we participate in. By these lights, an obligation to speak as others do can result from our participation in the practice of linguistic communication, a practice we engage in willy-nilly, as a result of enculturation.

See also: Conventions and Norms: Philosophical Aspects; Intentionality and Rationality; Linguistic Turn and Discourse Analysis in History; Methodological Individualism: Philosophical Aspects; Naturalism; Wittgenstein, Ludwig (1889–1951).

Bibliography

Baker, G., Hacker, P.M.S., 1984a. Language, Sense and Nonsense. Blackwell, Oxford.
Baker, G., Hacker, P.M.S., 1984b. Scepticism, Rules and Language. Blackwell, Oxford.
Blackburn, S., 1984. The individual strikes back. Synthese 58, 281–301.
Boghossian, P., 1989. The rule following considerations. Mind 98, 507–549. Reprinted in: Miller, A., Wright, C. (Eds.), 2002, pp. 141–187.
Boghossian, P., 2003. The normativity of content. Philosophical Perspectives 13, 31–45.
Boghossian, P., 2005. Is meaning normative? In: Nimtz, C., Beckermann, A. (Eds.), Philosophy – Science – Scientific Philosophy. Mentis, Paderborn, pp. 205–218.
Brandom, R., 1994. Making it Explicit. Harvard University Press, Cambridge, MA.
Davidson, D., 1984. Inquiries into Truth and Interpretation. Oxford University Press, Oxford.
Davidson, D., 1990. The structure and content of truth. Journal of Philosophy 87, 297–328.
Davidson, D., 2005. Truth, Language and History. Oxford University Press, Oxford.
Dummett, M., 1981. Frege: Philosophy of Language. Duckworth, London.
Dummett, M., 1993. Grammar and Style. Duckworth, London.
Evans, N., 2010. Dying Words. Wiley, Oxford.
Fodor, J., 2008. LOT 2. Oxford University Press, Oxford.
Fogelin, R., 1976. Wittgenstein, second ed. Routledge, London. 1987.
Gilbert, M., 1996. Living Together. Rowman & Littlefield, Lanham.
Glock, H.-J., 1996. A Wittgenstein Dictionary. Blackwell, Oxford.
Glock, H.-J., 2003. Quine and Davidson on Language, Thought and Reality. Cambridge University Press, Cambridge.
Glock, H.-J., 2005. The normativity of meaning made simple. In: Nimtz, C., Beckermann, A. (Eds.), Philosophy – Science – Scientific Philosophy. Mentis, Paderborn, pp. 219–241.
Glock, H.-J., 2010. Does language require conventions? In: Frascolla, P., Marconi, D., Voltolini, A. (Eds.), Wittgenstein: Mind, Meaning and Metaphilosophy. Palgrave Macmillan, Basingstoke, pp. 85–112.
Glüer, K., 2001. Dreams and nightmares. Conventions, norms, and meaning in Davidson's philosophy of language. In: Kotatko, P., Pagin, P., Segal, G. (Eds.), Interpreting Davidson. CSLI Publications, Stanford, CA, pp. 53–74.
Glüer, K., Pagin, P., 1999. Rules of meaning and practical reasoning. Synthese 118, 207–227.
Glüer, K., Wikforss, A., 2009. The normativity of meaning and content. In: Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/meaning-normativity/.
Grice, H.P., 1989. Studies in the Way of Words. Harvard University Press, Cambridge, MA.
Hattiangadi, A., 2006. Is meaning normative? Mind and Language 21, 220–240.
Hattiangadi, A., 2007. Oughts and Thoughts. Rule-Following and the Normativity of Content. Oxford University Press, Oxford.
Horwich, P., 1998. Truth. Oxford University Press, Oxford.
Kripke, S., 1982. Wittgenstein on Rules and Private Language. Blackwell, Oxford.
Kusch, M., 2006. A Sceptical Guide to Meaning and Rules. Acumen, Chesham.
McDowell, J., 1984. Wittgenstein on following a rule. Synthese 58, 325–363.
McHugh, C., 2012. Epistemic deontology and voluntariness. Erkenntnis 77, 65–94.
Miller, A., Wright, C. (Eds.), 2002. Rule-Following and Meaning. Acumen, Chesham.
Millikan, R., 1990. Truth rules, hoverflies, and the Kripke-Wittgenstein paradox. Philosophical Review 99, 325–353.
Ogden, C.K., Richards, I.A., 1923. The Meaning of Meaning. Kegan Paul, London.
Parfitt, D., 1997. Reasons and motivation. Proceedings of the Aristotelian Society 71.
Raz, J., 1978. Reasons for action, decisions and norms. In: Raz, J. (Ed.), Practical Reasoning. Oxford University Press, Oxford, pp. 128–143.
Russell, B., 1921. The Analysis of Mind. Allen and Unwin, London.
Sapir, E., 1921. Language. An Introduction to the Study of Speech. Harcourt Brace, New York.
Schroeder, S., 1998. Das Privatsprachenargument. Schöningh, Paderborn.
Searle, J., 1969. Speech Acts. Cambridge University Press, Cambridge.
Sellars, W., 1963. Some reflections on language games. In: Sellars, W. (Ed.), Science, Perception and Reality. Routledge & Kegan Paul, London, pp. 321–358.
Stainton, R.J., 1996. Philosophical Perspectives on Language. Broadview, Peterborough.
Tomasello, M., 2008. Origins of Human Communication. MIT Press, Cambridge, MA.
Waismann, F., 1965. The Principles of Linguistic Philosophy. Macmillan, London.
Whiting, D., 2009. Is meaning fraught with ought? Pacific Philosophical Quarterly 90, 535–555.
Whiting, D., 2013. What is the normativity of meaning? Inquiry 56.
Williams, B., 1973. Problems of the Self. Cambridge University Press, Cambridge.
Wittgenstein, L., 1953. Philosophische Untersuchungen. Philosophical Investigations (G.E.M. Anscombe, Trans.). Blackwell, Oxford.
Wittgenstein, L., 1974. Philosophical Grammar. Blackwell, Oxford.
von Wright, G.H., 1963. Norm and Action. Routledge & Kegan Paul, London.
Zalabardo, J. (Ed.), 2013. The Aim of Belief. Special Issue of Teorema XXXXII/3.

Relevant Website

http://philpapers.org/browse/normativity-of-meaning-and-content – PhilPapers Page on Normativity of Meaning and Content.

Measurement Error Models

Alicia L Carriquiry, Iowa State University, Ames, IA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 14, pp. 9435–9442, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

In many areas of application, response variables are modeled in terms of one or more covariates that are not directly observable. Instead, substitute variables that can be observed are used in place of the covariates of interest in the model. Because these substitute variables measure the original covariates with error, we use the term *measurement error models* to describe those models. In the simple additive case, measurement error has the effect of masking or *attenuating* the relationship between the response variable and the covariate measured with error. A correction for attenuation can be carried out only if replicate observations are available, or if information about the measurement error can be obtained from other sources. Results derived for the simple case can be extended, but not in a straightforward manner, to the multivariate regression case. Measurement error problems in a nonlinear setting have also been addressed in recent years. Logistic regression models with measurement error in one or more of the covariates have received much attention since the 1980s.

The term *measurement error model* is used to denote a regression model, either linear or nonlinear, where at least one of the covariates or predictors is observed with error. If x_j denotes the value of the covariate for the ith sample unit, then x_i is unobserved and we observe instead $X_i = f(x_i, u_i)$, where u_i is known as the *measurement error*. The observed (or *indicator*) variable X_i is assumed to be associated to the unobserved (or *latent*) variable x_i via the function f; the form of this association defines the different types of measurement error models described in the literature. For a comprehensive bibliography on this literature, see Fuller (1987) and Carroll (1995).

In the simple additive measurement error model, measurement error has the effect of obscuring the relationship between the response variable and the covariate that is measured with error. This effect is sometimes known as *attenuation*, as in the linear regression case, the ordinary least squares estimator of the regression coefficient associated to the variable that is measured with error is biased towards zero (e.g., Fuller, 1987).

The next section describes the effect of additive measurement error in the simple regression case, and notes the relationship between this topic and those in other entries. Subsequent sections describe extensions to the multivariate and nonlinear regression cases. The article ends with a brief historical overview of the development of measurement error models.

Simple Linear Regression and Measurement Error

The simple linear regression model (*see* Linear Hypothesis: Regression (Basics)) is given by

$$Y_i = \beta_0 + \beta_1 x_i + e_i \qquad [1]$$

where Y_i is the response observed for the ith sample item, x_i is the explanatory variable (or covariate) measured for the ith sample item, β_0 and β_1 are unknown regression coefficients, and e_i is random error typically assumed to be distributed as a normal random variable with mean 0 and σ_e^2. Here we talk about *functional models* when the explanatory variable x is assumed to be a fixed value, and about *structural models* when instead x is assumed to be a random draw from some distribution. Typically, we choose $x \sim N(\mu_x, \sigma_x^2)$ (e.g., Fuller, 1987). [Carroll (1995), however, makes a different distinction between functional and structural models. They refer to *functional modeling* when the variable x is either fixed or random but minimal assumptions are made about its distribution, and to *structural modeling* when a parametric form is chosen for the distribution of x. Carroll (1995) argues that when a consistent parameter estimate under the structural model formulation can be found, it tends to be robust to misspecification of the distribution of x.]

It is well known that the ordinary least squares estimator (*see* Linear Hypothesis: Regression (Basics)) of β_1

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad [2]$$

is unbiased and has smallest variance among linear unbiased estimators. If $x \sim N(\mu_x, \sigma_x^2)$ and $Cov(x, e) = 0$, then the estimator $\widehat{\beta}_1$ in (2) is the maximum likelihood estimator of β_1 (*see* Estimation: Point and Interval).

Models for Measurement Error

Suppose now that we cannot observe the explanatory variable x_i directly. Instead, we observe the *manifest* or *indicator* variable X_i which is a noisy measurement of x_i. We distinguish two different types of measurement error models depending on the relationship between X_i and x_i:

1. Error models, discussed in this entry, where we model the distribution of X given x.
2. Regression calibration models, where the distribution of x given X is of interest.

Within the class of error models, we distinguish two variants: *classical additive error models*, and *error calibration models*. The *classical additive error model* establishes that the observed variable X is an unbiased measure of x. That is

$$X = x + u \qquad [3]$$

where $u \sim (0, \sigma_u^2)$. As an example, consider the problem of estimating the relationship between serum cholesterol levels and usual intake of saturated fat for a sample of individuals. Suppose further that intake of cholesterol for an individual is measured by observing the individual's intake over a randomly chosen 24-hour period. Because daily nutrient intake varies from day to day for an individual, the intake measured on any one day is a noisy measurement of the individual's usual or habitual intake of the nutrient.

We talk about *error calibration models* when the observed variable is a *biased* measurement of the variable of interest. In this case, we use a regression model to associate the two variables.

$$X = \alpha_0 \alpha_1 x + u \qquad [4]$$

where as before, $E(u) = 0$, but now $E(X) = \alpha_0 + \alpha_1 x$. As an example, consider the problem of assessing the habitual daily alcohol consumption for a sample of individuals. Typically, individuals in the sample are asked to record their alcohol consumption over a short period, for example over a week. The mean of the observed alcohol intakes over the seven days cannot be assumed to be an unbiased estimator of the usual daily alcohol consumption, as there is evidence (see Freedman et al., 1991; Carroll, 1995) that individuals tend to under-report the amount of alcohol they drink. Suppose in addition that the under-reporting is not correlated with alcohol consumption. In this case, the measurement X is biased for x and must be corrected. If information about relationship (4) is available, the measurement X can be *calibrated* by using $\alpha_1^{-1}(X - \alpha_0)$.

We briefly discuss *regression calibration models* later in this article.

The Effect of Measurement Error

Consider the simple example of a linear regression with additive measurement error in the covariate, as in expressions (1) and (3). Further, assume that the measurement error is independent of the true measurements x and of the errors e. If $X = x + u, x \sim (\mu_x, \sigma_x^2)$ and $u \sim (0, \sigma_u^2)$ then the regression coefficient estimated using the noisy measurements is biased towards zero. If $\widehat{\gamma}_1$ is the estimated regression coefficient obtained from using X in the regression equation, then

$$E(\widehat{\gamma}_1) = \lambda \beta_1 \qquad [5]$$

where β_1 is the regression coefficient associated to x and λ is called a *reliability ratio* (e.g., Fuller, 1987, Sect. 1.1; Carroll, 1995, Sect. 2.2.1), computed as

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \qquad [6]$$

Clearly, $\lambda < 1$, and approaches 1 as the variance of the measurement error goes to zero. Therefore, ordinary least squares estimation in the model with X instead of x produces an estimator of the regression coefficient that is biased towards zero, or is *attenuated*.

Measurement error has in addition an effect on the variability of points around the estimated regression line. Note that under model (1) with measurement error as in (3),

$$\mathrm{var}(Y|X) = \sigma_e^2 + \frac{\beta_1^2 \sigma_u^2 \sigma_x^2}{\sigma_x^2 + \sigma_u^2} \qquad [7]$$

Thus measurement error has the effect of increasing the noise about the regression line (Carroll, 1995).

A simple simulated example

To illustrate the effect of measurement error on regression estimation, we constructed 10 responses $Y = \beta_0 + \beta_1 x + e$, where $\beta_0 = 0$, $\beta_1 = 2$, $e \sim N(0, 1)$, and $x \sim N(1, 1)$. We then constructed a noisy measurement of x by adding measurement error u where $u \sim N(0, 3)$. We fitted two simple linear regression models, using x and using X as the covariate in the model. The data, together with the two fitted regression lines are shown in Figure 1. In the figure, the dark dots represent the true measurements, while the diamonds represent the noisy measurements. Note that the regression line fitted to the noisy data is attenuated towards zero. In fact, in this simulated example, the true attenuation coefficient or reliability ratio is only $\frac{1}{4}$. Attenuation of the slope is greatly dependent on the simple additive error model. In other cases, the ordinary least squares estimator of the regression coefficient may *overestimate* the true slope. Consider, for example, the case where X is a *biased* measurement of x, and in addition, the measurement error is correlated with the equation error. That is, X is as in eqn [4] and Corr $(e, u) = \rho_{eu}$. It can be shown that

$$E\{\widehat{\gamma}_1\} = \frac{\beta_1^2 \alpha_1 \sigma_x^2 + \rho_{eu} \sigma_e \sigma_u}{\alpha_1^2 \sigma_x^2 + \sigma_u^2} \qquad [8]$$

where, as in eqn [5], $\widehat{\gamma}_1$ denotes the estimated regression coefficient obtained by using X in the model, and β_1 is the true

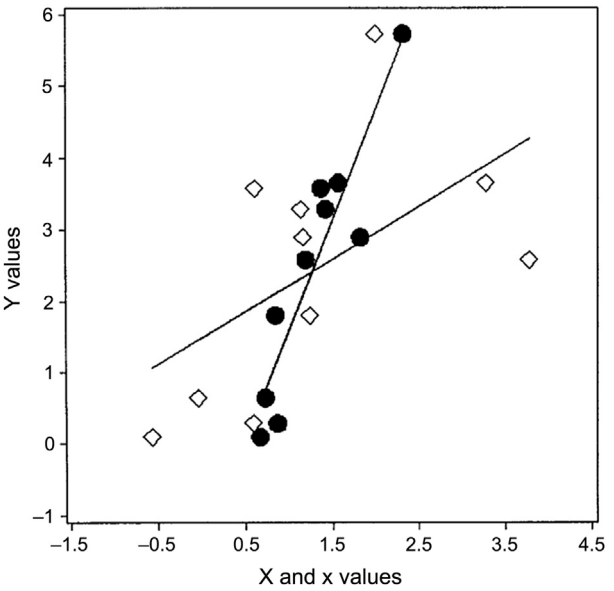

Figure 1 Simulated example. Black circles represent observations without measurement error. Diamonds represent the observations subject to measurement error in the covariate. The steeper line is the fitted regression line in the absence of measurement error, while the flatter line shows the fitted regression line in the presence of measurement error.

slope. Notice that, depending on the values of α_1 and of ρ_{eu}, $\widehat{\gamma}_1$ may actually over-estimate β_1.

Multiple Linear Regression and Measurement Error

Extension of results presented in Section Simple Linear Regression and Measurement Error to the case of multiple predictors is not straightforward, even if only one of the predictors is subject to measurement error and the measurement error model is as in eqn [3].

Let

$$Y = \beta_0 + \beta_1 x + \beta_w' \mathbf{w} + e \qquad [9]$$

where now $\mathbf{w}$ is a $p \times 1$ vector of predictors that are measured without error, β_w is the vector of regression coefficients associated to $\mathbf{w}$, and x is measured with simple additive measurement error as in eqn [3]. Furthermore, let $\sigma^2_{x|w}$ denote the variance of the residuals in the regression of x on $\mathbf{w}$. If model (9) is fitted using X in place of x, it can be shown (e.g., Gleser et al., 1987) that, as in eqn [5], the estimated regression coefficient for X is $\widehat{\gamma}_1 = \lambda\beta_1$, where

$$\lambda = \frac{\sigma^2_{x|w}}{\sigma^2_{x|w} + \sigma^2_u}. \qquad [10]$$

Notice that when x and $\mathbf{w}$ are uncorrelated, λ in eqn [10] reduces to eqn [6]. Gleser et al. (1987) show that when one of the covariates is measured with error, the regression coefficients associated to $\mathbf{w}$ that are measured without error are also biased, unless x and $\mathbf{w}$ are uncorrelated.

If now $\mathbf{x}$ is a $q \times 1$ vector of covariates measured with error, and $\mathbf{w}$ is as above, then the ordinary least squares estimates have expectation.

$$E\begin{bmatrix}\widehat{\gamma}_x \\ \widehat{\gamma}_w\end{bmatrix} = \begin{bmatrix}\Sigma_{xx} + \Sigma_{uu} & \Sigma_{xw} \\ \Sigma_{wx} & \Sigma_{ww}\end{bmatrix}^{-1} \times \left\{\begin{bmatrix}\Sigma_{xx} & \Sigma_{xw} \\ \Sigma_{wx} & \Sigma_{ww}\end{bmatrix}\begin{bmatrix}\beta_x \\ \beta_w\end{bmatrix} + \begin{bmatrix}\Sigma_{ue} \\ 0\end{bmatrix}\right\}. \qquad [11]$$

The ordinary least squares estimate of the regression coefficients will be unbiased only when $\sum_{uu} = 0$ and $\sum_{ue} = 0$.

Estimation in Measurement Error Models

Consider first the simple model (1) with measurement error as in eqn [3]. Assume that $(x, e, u)' \sim N[(\mu_x, 0, 0)' diag(\sigma^2_x, \sigma^2_e, \sigma^2_u)]$, and suppose that σ^2_u is known. Under this model, the vector, $(\mathbf{Y}, \mathbf{X})$ is distributed as a bivariate normal vector, and therefore, the sample first- and second-order moments form a set of sufficient statistics for the parameters. Fuller (1987, Sect. 1.2.1) and Carroll (1995, Sect. 2.3.1) derive *method of moments* (*see* Estimation: Point and Interval) parameter estimates that can be obtained from a one-to-one mapping from the minimal sufficient statistics to the vector $(\widehat{\mu}_x, \widehat{\sigma}^2_x, \widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2_e)$. For example,

$$\begin{aligned}\widehat{\beta}_0 &= Y - \widehat{\beta}_1 \overline{X} \\ \widehat{\beta}_1 &= \left(s^2_x - \sigma^2_u\right)^{-1} s_{XY}\end{aligned} \qquad [12]$$

where $\overline{Y}$ and $\overline{X}$ are the sample means of Y and X, and s^2_X band s_{XY} are the sample variance of X and the sample covariance between X and Y, respectively. Fuller (1987) shows that when σ^2_u is known and as $n \to \infty$,

$$n^{\frac{1}{2}}\left[\left(\widehat{\beta}_0 - \beta_0\right), \left(\widehat{\beta}_1, \beta_1\right)\right]' \xrightarrow{\ell} N\left\{(0,0)'\mathbf{T}\right\} \qquad [13]$$

where $(\widehat{\beta}_0 - \beta_1)$ are as in (12) and $\mathbf{T}$ is a positive definite matrix (see Fuller, 1987, Sect. 1.2.2).

Suppose now that rather than assuming that σ^2_u is known, we consider the case where the *relative* sizes of the variances $\sigma_u^{-2}\sigma^2_e$ are known. We use δ to denote the ratio of variances. Under the assumption of normality of $(x, e, u)'$ as above, method of moments estimators for $(\beta_0, \beta_1, \sigma^2_u)$ are

$$\begin{aligned}\widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X} \\ \widehat{\beta}_1 &= \frac{s^2_Y - \delta s^2_X + \left[\left(s^2_Y - \delta s^2_x\right)^2 + 4\delta s^2_{XY}\right]^{\frac{1}{2}}}{2 s_{XY}} \\ \widehat{\sigma}^2_u &= \frac{s^2_Y - \delta s^2_X + \left[\left(s^2_Y - \delta s^2_x\right)^2 + 4\delta s^2_{xY}\right]^{\frac{1}{2}}}{2\delta}\end{aligned} \qquad [14]$$

where for $\widehat{\sigma}^2_u$ we choose the smallest root of the quadratic expression above, so that $\widehat{\sigma}^2_u < s^2_X$. Fuller (1987, Theorem 1.3.1) shows that, for $\theta = (\beta_0, \beta_1, \sigma^2_u)'$, and $\widehat{\theta} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2_u)'$ (as given in eqn [14]), as $n \to \infty$

$$n^{\frac{1}{2}}\left(\widehat{\theta} - \theta\right) \xrightarrow{L} N\left(0, \mathbf{\Gamma}\right) \qquad [15]$$

for $\mathbf{\Gamma}$ a positive definite covariance matrix. Asymptotic normality of $(\widehat{\beta}_0, \widehat{\beta}_1)$ holds even if the distribution of x is not normal.

The estimators above can be extended to the case where a vector $\mathbf{x}$ is measured with error. While in the simple linear regression case results apply in the functional and in the structural model framework, in the multivariate case we must distinguish between models in which $\mathbf{x}$ is assumed to be random and those in which $\mathbf{x}$ is assumed to be fixed. We consider first the *structural model* set-up, where $\mathbf{x}$ is taken to be random. In this case, the maximum likelihood estimators of model parameters derived under the normal model can be shown to be consistent and asymptotically normal under a wide range of assumptions.

Let

$$\begin{aligned}Y &= \beta_0 + \mathrm{x}\beta_1 + e \\ X &= \mathrm{x} + \mathrm{u}\end{aligned} \qquad [16]$$

where

$$\begin{bmatrix}\mathrm{x} \\ \mathrm{e} \\ \mathrm{u}\end{bmatrix} \sim N\left(\begin{bmatrix}\mu_x \\ 0 \\ 0\end{bmatrix}, \begin{bmatrix}\Sigma_{xx} & 0 & 0 \\ 0 & \sigma^2_e & \Sigma_{eu} \\ 0 & \Sigma_{ue} & \Sigma_{uu}\end{bmatrix}\right) \qquad [17]$$

Here, $\sum_{uu}$ and $\sum_{eu}$ are known. The maximum likelihood estimators of (β_0, β_1) are

$$\begin{aligned}\widehat{\beta}_0 &= \overline{Y} - \overline{X}\widehat{\beta}_1 \\ \widehat{\beta}_1 &= \left(S_{XX} - \Sigma_{uu}\right)^{-1}\left(S_{XY} - \Sigma_{ue}\right)\end{aligned} \qquad [18]$$

where $\mathbf{S}_{XX}$ is the sample covariance matrix of $\mathbf{X}$ and $\mathbf{S}_{XY}$ is the covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$. A modified maximum likelihood estimator of $(\beta_0, \beta\mathrm{a}_1)$ for the case where $\sum_{uu}$ and $\sum_{eu}$ are

unknown and must be estimated from the data is given by Fuller (1987, Sect. 2.2).

If the vector $\mathbf{x}$ is taken to be fixed, rather than random, the method of maximum likelihood fails to produce consistent estimators of the parameters in the model. This is because the vector of unknown model parameters (β_0, β_1, $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$, $\sigma^2{}_e$) is n + 3-dimensional, but the sample size is only n. In this case, then, the number of parameters increases with sample size. Estimation of model parameters in these cases has been discussed by, for example, Morton (1981).

Consider now model (16) with assumptions as in (17), and suppose that we do not know $\sum_{uu}$ and $\sum_{eu}$ as above. Instead, let $\boldsymbol{\varepsilon} = (e, \mathbf{u})' \sim N\,(\mathbf{0}, \sum_{\varepsilon}\sum_{\varepsilon})$, where $\sum_{\varepsilon\varepsilon} = \boldsymbol{\Omega}_{\varepsilon\varepsilon}\,\sigma^2$, and $\boldsymbol{\Omega}_{\varepsilon\varepsilon}$ is known. In this case, the maximum likelihood estimators of β_1 and σ^2 are

$$\begin{aligned}\widehat{\beta}_1 &= \left(\overline{\mathrm{X}}'\overline{\mathrm{X}} - \widehat{\lambda}\Omega_{uu}\right)^{-1}\left(\overline{\mathrm{X}}'\overline{\mathrm{Y}} - \widehat{\lambda}\Omega_{ue}\right)\\ \widehat{\sigma}^2 &= (q+1)^{-1}\widehat{\lambda}\end{aligned} \qquad [19]$$

where $\overline{\mathrm{X}}'\overline{\mathrm{X}}$ and $\overline{\mathrm{X}}'\overline{\mathrm{Y}}$ are averages taken over the n sample observations, and $\widehat{\lambda}$ is the smallest root of $\left|\overline{Z}'\overline{Z} - \lambda\Omega_{\varepsilon\varepsilon}\right|$ with $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$. The estimator of β_1 is the same whether $\mathbf{x}$ is fixed or random. Under relatively mild assumptions, the estimator $\widehat{\beta}_1$ is distributed, in the limit, as a normal random vector (see Fuller, 1987, Theorem 2.3.2).

Related Methods

Regression with Instrumental Variables

In the preceding sections we have assumed implicitly that when the variance of the measurement error is not known, it can be estimated from data at hand. Sometimes, however, it is not possible to obtain an estimate of the measurement error variance. Instead, we assume that we have available another variable, which we denote Z, and which is known to be correlated with x. If Z is independent of e and u, and Z is correlated with x, we say that Z is an *instrumental variable* (*see* Instrumental Variables in Statistics and Econometrics). Fuller (1987) presents an in-depth discussion of instrumental variable estimation in linear models (Sect. 1.4 and 2.4).

He shows that method of moments estimators for the regression coefficient can be obtained by noticing that the first and second sample moments of the vector (Y, X, Z) are the set of minimal sufficient statistics for the model parameters. The resulting estimators are

$$\begin{aligned}\widehat{\beta}_1 &= s_{XY}^{-1}s_{YZ}\\ \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1\overline{X}\end{aligned} \qquad [20]$$

where s_{XZ} and s_{YZ} are the sample covariances of the covariate and the response variables with the instrumental variable, respectively. Because the sample moments are consistent estimators of the population moments, the estimators in (20) are consistent estimators of β_0 and β_1. Notice, however, that if the assumption $\sigma_{xZ} \neq 0$ is violated, the denominator in the expression for $\widehat{\beta}_1$ above estimates zero and the estimator is therefore not consistent.

Amemiya (1985, 1990a, 1990b), Carroll and Stefanski (1994) and Buzas and Stefanski (1996) discuss instrumental variable estimation in nonlinear models.

Factor Analysis

Factor analysis is heavily used in psychology, sociology, business, and economics (*see* Factor Analysis and Latent Variable Models in Personality Psychology). We consider a model similar to the model used for estimating an instrumental variable, with a few additional assumptions. The simplest version of a factor model is

$$\begin{aligned}Y_1 &= \beta_{01} + \beta_{11}x + e_1\\ Y_2 &= \beta_{02} + \beta_{12}x + e_2\\ X &= x + u\end{aligned} \qquad [21]$$

Here, (Y_1, Y_2, X) are observed. We assume that the coefficients associated to x are nonzero, and that $(x, e_1, e_2, u)' \sim N[(\mu_x, 0, 0, 0)', \mathrm{diag}\,(\sigma_x^2, \sigma_{e1}^2, \sigma_{e2}^2, \sigma_u^2)]$.

In factor analysis, the unobservable variable x is denoted *common factor* or *latent factor*. For example, x might represent intelligence, and the observed variables (Y_1, Y_2, X) might represent the individual's scores in three appropriate tests. The ratio of the variance of x to the variance of the response variable is called the *communality* of the response variable. For example, the communality of Y_1 above is calculated as

$$\lambda_1^2 = \frac{\beta_1^2\sigma_x^2}{\beta_1^2\sigma_x^2 + \sigma_{e1}^2} \qquad [22]$$

Note that λ_1^2 is equal to the reliability ratio that was presented in Section Simple Linear Regression and Measurement Error.

Estimators for β_{11} and for β_{12} can be obtained using the same methods that were used for estimating instrumental variables in Section Regression with Instrumental Variables. For example, $\widehat{\beta}_{11}$ is the instrumental variable estimator of β_{11} under the model $Y_1 = \beta_{01} + \beta_{11}x + e$ using Y_2 as an instrumental variable. Similarly, an estimator for β_{12} is obtained using Y_1 as an instrumental variable.

For an extensive discussion of factor analysis, refer to Jöreskog (1981) and Fuller (1987).

Regression Calibration

Suppose that for each sample item we observe Y, X, and $\mathbf{w}$, where $X = x + u$, and $\mathbf{w}$ is a vector of covariates measured without error. The idea behind regression calibration is simple. Using information from X and $\mathbf{w}$, estimate the regression of x on (X, $\mathbf{w}$). Then fit the regression model for Y using $E\{x|X, \mathbf{w}\}$ using standard methods. The standard errors of the parameters in the regression model for Y must be adjusted to account for the fact that x is estimated rather than observed.

Estimation of the regression of x on (X, $\mathbf{w}$) is straightforward when replicate observations are available. Suppose that there are m replicate measurements of x for each individual, and that $\overline{X}$ is their mean. In that case, the *best linear estimator* of x is $E\{x|\overline{X}, w\}$, where

$$E\{x|\overline{X}, w\} = \mu_x + \begin{pmatrix}\sigma_x^2\\ \Sigma_{wx}\end{pmatrix} \times \begin{bmatrix}\sigma_x^2 + \sigma_u^2/m & \Sigma_{xw}\\ \Sigma_{xw}' & \Sigma_{ww}\end{bmatrix}^{-1}\begin{pmatrix}\overline{X} - \mu_x\\ \mathbf{w} - \mu_w\end{pmatrix} \qquad [23]$$

Moment estimators for all unknown parameters in the expression above can be obtained as long as replicate

observations are available and an estimator for the unknown measurement error variance can be computed. If replicate observations are not available, it is possible to use an external estimate of the measurement error variance.

Regression calibration was proposed by Gleser (1990). Carroll (1995, Chap. 3) discusses regression calibration in nonlinear models.

An alternative approach called SIMEX (for simulation extrapolation) that is well suited for the additive measurement error case was proposed by Cook and Stefanski (1994).

Nonlinear Models and Measurement Error

Carroll (1995, Chap. 6) presents an in-depth discussion of the problem of estimating the parameters in a nonlinear model when one or more of the covariates in the model are measured with error. Most of their discussion assumes that the measurement error is additive and normal, but some results are presented for more general cases. Other references include Amemiya (1990), Gleser (1990), and Roeder et al. (1996).

Carroll (1995) presents two approaches for estimating the parameters in the nonlinear model subject to measurement error, denoted *conditional score* and *corrected score* methods, that are based on an *estimating equations* approach. In either case, the resulting estimators do not depend on the distribution of x.

The *conditional score* method is derived for the case of the canonical *generalized linear model* (McCullagh and Nelder, 1989) (*see* Linear Hypothesis). Without loss of generality we discuss the case where both x and w are scalar valued. The conditional density of Y given (x, w) is given by

$$f(y|x, w, \Theta) = \exp\left\{\frac{y\eta - D(\eta)}{\phi} + c(y, \phi)\right\} \qquad [24]$$

(expression 6.4 in Carroll, 1995), where $\eta = \beta_0 + \beta_1 x + \beta_2 w$ is called a *natural parameter*, $\Theta = (\beta_0, \beta_1, \beta_2, \varphi)$ is the unknown parameter to be estimated, and where the first and second moments of the distribution of Y are proportional to the first and second derivatives of $D(\eta)$ with respect to η. The form of η, $D(\eta)$ φ and $c(y, \varphi)$ depends on the model choice.

If x were observed, then Θ is estimated by solving the following equations

$$\sum_{i=1}^{n}\left\{Y_i - D^{(1)}(\eta_i)\right\}\begin{pmatrix}1\\ w_i\\ x_i\end{pmatrix} = 0$$

$$\sum_{i=1}^{n}\left[\left(\frac{n-1}{n}\right)\phi - \frac{\left(Y_i - D^{(i)}(\eta_i)\right)^2}{D^{(2)}(\eta_i)}\right] = 0 \qquad [25]$$

where $D^{(k)}$ (η) is the kth order partial derivative of D (η) with respect to η.

We now assume that x is measured with normal additive measurement error with variance σ_u^2. A sufficient statistic Δ can be found for x (Stefanski and Carroll, 1987) so that the distribution of Y conditional on (Δ, w) has the same form as the density given in (24). The form of the functions D (η), η, φ and c (y, φ) must be modified accordingly (see, for example, Carroll, 1995, p. 127). Therefore, this suggests that an estimator for Θ when x is observed with measurement error can be obtained by solving a set of equations similar to eqn [25].

The approach described above can be used only in the case of canonical generalized linear models. An approach applicable to the broader class of generalized linear regression models is the *corrected score* method, also derived under an estimating equations framework. Suppose, first, that x is observable, and that an estimator of Θ is obtained by solving

$$\sum_{i=1}^{n}\xi(Y_i, x_i, w_i, \Theta) = 0 \qquad [26]$$

where ξ is a likelihood score from the model when x is observed without measurement error. Suppose now that we can find a function $\xi * (Y, X, w_i, \Theta, \sigma_u^2)$, such that

$$E\{\xi * (Y, X, w_i, \Theta)|Y, x, w\} = \xi(Y, x, w, \Theta) \qquad [27]$$

for all (Y, x, w). An estimate for the unknown parameter Θ can then be obtained by solving eqn [26] where the corrected score $\xi*$ is used in place of ξ. Corrected score functions that satisfy eqn [27] do not always exist, and when they do exist they can be very difficult to find. Stefanski (1989) derived the corrected score functions for several models that are commonly used in practice.

A Brief Historical Overview

The problem of fitting a simple linear regression model in the presence of measurement error was first considered by Adcock (1877). Adcock proposed an estimator for the slope in the regression line that accounted for the measurement error in the covariate for the special case where $\sigma^2{}_a = \sigma^2{}_e$. The extension to the p-variate case was presented by Pearson (1901).

It was not until the mid-1930s that the terms *errors in variables models* and *measurement error models* were coined, and a systematic study of these types of model was undertaken. The pace of research in this area picked up in the late 1940s, with papers by Lindley (1947), Neyman and Scott (1951), and Kiefel and Wolfowitz (1956). These authors addressed issues of identifiability and consistency in measurement error models, and proposed various approaches for parameter estimation in those models. Lindley (1947) presented maximum likelihood estimators for the slope in a simple linear regression model with additive measurement error for the functional model. He concluded that, when the two variance components are unknown, the method of maximum likelihood cannot be used for estimating β_1, as the resulting estimator must satisfy the relation $\widehat{\beta}_1, \widehat{\sigma}_u^2 = \widehat{\sigma}_e^2$. Solari (1969) later showed that this unexpected result could be explained by noticing that the maximum likelihood solution was a saddle-point and not a maximum of the likelihood function. Lindley (1947) also showed that the generalized least squares estimator of the slope is the maximum likelihood estimator when the ratio of variances is assumed known.

Extensions of the simple linear regression model to the multiple regression model case, with one or more covariates measured with error were presented in the late 1950s, 1960s and 1970s. Excellent early in-depth reviews of the literature in measurement error models are given by Madansky (1959), and Morton (1981).

In the early 1990s, Schmidt and Rosner (1993) proposed Bayesian approaches for estimating the parameters in the regression model in the presence of measurement error. More recently, Robins and Rodnitzky (1995) proposed a semi-parametric method that is more robust than the maximum likelihood estimator of the slope in structural models when the distributional assumptions are not correct. Carroll et al. (1997), however, showed that the Robins and Rodnitzky estimator can be very inefficient relative to the parametric estimation when the distributional assumptions are correct. Roeder et al. (1996) and Carroll et al. (1999) have argued that the sensitivity of the maximum likelihood estimator to model specification can be greatly reduced by using flexible parametric models for the measurement error. In particular, the authors suggest that a normal mixture model be used as the flexible model. In the 1999 paper, Carroll et al. (1999) proceed from a Bayesian viewpoint, and use Markov chain Monte Carlo methods to obtain approximations to the marginal posterior distributions of interest.

The literature on measurement error models is vast, and we have cited a very small proportion of the published books or manuscripts. In addition to the review articles listed above, the text books by Fuller (1987) and Carroll (1995) are widely considered to be the best references for linear and nonlinear measurement error models, respectively. The recent manuscript by Stefanski (2000) cites several papers that have appeared in the literature since 1995.

Bibliography

Adcock, R.J., 1877. Note on the method of least squares. The Analyst 4, 183–184.

Amemiya, Y., 1990. Two stage instrumental variable estimators for the nonlinear errors in variables model. Journal of Econometrics 44, 311–332.

Buzas, J.S., Stefanski, L.A., 1996. Instrumental variable estimation in generalized linear measurement error models. Journal of the American Statistical Association 91, 999–1006.

Carroll, R.J., Stefanski, L.A., 1994. Measurement error, instrumental variables, and corrections for attenuation with applications to metaanalyses. Statistics in Medicine 13, 1265–1282.

Carroll, R.J., Freedman, L.S., Pee, D., 1997. Design aspects of calibration studies in nutrition, with analysis of missing data in linear measurement error. Biometrics 53, 1440–1457.

Carroll, R.J., Roeder, K., Wasserman, L., 1999. Flexible parametric measurement error models. Biometrics 55, 44–54.

Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M., 2006. Measurement error in nonlinear models: a modern perspective. CRC press, Boca Raton, FL.

Carroll, R.J., 1995. Measurement Error in Nonlinear Models. Chapman and Hall, London.

Cook, J., Stefanski, L.A., 1994. Simulation extrapolation estimation in parametric measurement error models. Journal of the American Statistical Association 89, 1314–1328.

Freedman, L.S., Carroll, R.J., Wax, Y., 1991. Estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake. American Journal of Epidemiology 134, 510–520.

Fuller, W.A., 1987. Measurement Error Models. Wiley, New York.

Gleser, L.J., Carroll, R.J., Gallo, P.P., 1987. The limiting distribution of least squares in an errors-in-variable regression model. Annals of Statistics 15, 220–233.

Gleser, L.J., 1985. A note on Dolby G. R. unreplicated ultrastructural model. Biometrika 72, 117–124.

Gleser, L.J., 1990. Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. In: Brown, P.J., Fuller, W.A. (Eds.), Statistical Analysis of Measurement Error Models and Applications. Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference. American Mathematical Society, Rhode Island, RI, pp. 99–114.

Grewal, R., Cote, J.A., Baumgartner, H., 2004. Multicollinearity and measurement error in structural equation models: Implications for theory testing. Marketing Science 23 (4), 519–529.

Jarvis, C.B., MacKenzie, S.B., Podsakoff, P.M., 2003. A critical review of construct indicators and measurement model misspecification in marketing and consumer research. Journal of consumer research 30 (2), 199–218.

Jöreskog, K.G., 1981. Structural analysis of covariance structures. Scandinavian Journal of Statistics 8, 65–92.

Kiefer, J., Wolfowitz, J., 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Annals of Mathematical Statistics 27, 887–906.

Lindley, D.V., 1947. Regression lines and the linear functional relationship. Journal of the Royal Statistical Society, Supplement 9, 218–244.

Madansky, A., 1959. The fitting of straight lines when both variables are subject to error. Journal of the American Statistical Association 54, 173–205.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman and Hall, London.

Morton, R., 1981. Efficiency of estimating equations and the use of pivots. Biometrika 68, 227–233.

Neyman, J., Scott, E.L., 1951. On certain methods of estimating the linear structural relations. Annals of Mathematical Statistics 22, 352–361.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2, 559–572.

Robins, J.M., Rodnitzky, A., 1995. Semi-parametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association 90, 122–129.

Roeder, K., Carroll, R.J., Lindsay, B.G., 1996. A semi-parametric mixture approach to case-control studies with errors in covariables. Journal of the American Statistical Association 91, 722–732.

Rosner, B., Spiegelman, D., Willett, W.C., 1990. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. American Journal of Epidemiology 132, 734–745.

Schmidt, C.H., Rosner, B., 1993. A Bayesian approach to logistic regression models having measurement error following a mixture distribution. Statistics in Medicine 12, 1141–1153.

Solari, M.E., 1969. The 'maximum likelihood solution' to the problem of estimating a linear functional relationship. Journal of the Royal Statistical Society B 31, 372–375.

Stefanski, L.A., Carroll, R.J., 1987. Conditional scores and optimal scores in generalized linear measurement error models. Biometrika 74, 703–716.

Stefanski, L.A., 1989. Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. Communications in Statistics, Series A 18, 4335–4358.

Stefanski, L.A., 2000. Measurement error models. Journal of the American Statistical Association 95, 1353–1357.

Measurement, Representational Theory of

Reinhard Suck, Universitat Osnabrueck, Osnabrueck, Austria

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9442–9448, © 2001, Elsevier Ltd.

Abstract

The principles of representational measurement theory are expounded in historical and contemporary perspective, including a preliminary formulation of representation, uniqueness, and meaningfulness. The whole article is based on the concept of an ordered relational structure. The definition and representation theorems of extensive and difference structures are presented and discussed. These are ordered relational structures with operations or difference like relations defined on them. Stevens's classification of scale type (i.e., the classical ratio, interval, and ordinal scales) and its modern development in the context of the automorphism group of a measurement structure are dealt with, leading naturally to the problem of meaningfulness of numerical statements. Finally problems and applications of the representational theory of measurement are pointed out.

The representational theory of measurement – RTM for short – is a sort of 'interface' between a qualitative, empirical, or material domain of research and the use of numbers to quantify the empirical observations and to draw inferences from the data. Roughly, the main tenets of RTM consist of the following steps (for definition of the main concepts *see* Ordered Relational Structures):

1. Describe the situation under scrutiny by a qualitative relational structure, consisting of a base set A and a finite number of relations. Among these relations is usually an order relation and sometimes special elements such as units. These entities are the primitives of the structure. They must have an empirical identification. The properties of this qualitative relational structure are stated as conditions or axioms on its primitives, which are true statements about the structure, at least in the intended empirical identification.
2. Verify by empirical procedures (experiments, observations) the truth of those conditions that are testable (which should be almost all of them) and find arguments (plausibility, consequences of failure) for those that are not testable.
3. Find a structure with a numerical base set (a subset of $\mathbb{R}$ or $\mathbb{R}^n$) and find homomorphism between both structures. Finding a homomorphism may consist of providing a representation theorem, which essentially, states the existence of at least one homomorphism.
4. Find all homomorphisms between the qualitative and the numerical structure. This step amounts to proving a uniqueness theorem, which in most cases yields a procedure to calculate any homomorphism from a given one.
5. Clarify on the basis of the uniqueness properties of the set of homomorphisms which numerical statements about the empirical domain are meaningful and which are not. Meaningful statements can be used for further analysis, such as statistical tests.

This theory is thoroughly described in the three volumes *Foundations of Measurement* (see Krantz et al., 1971; Suppes et al., 1989; Luce et al., 1990). Other comprehensive expositions, differing slightly in emphasis and philosophical argumentation, are Pfanzagl (1971), Roberts (1979), and Narens (1985).

History and Contemporary a Formulation of Measurement Within RTM

The above kind of formalization of measurement goes back to Helmholtz (1887). Hölder (1901) was the first to prove a representation theorem in the above sense. One version or other of Hölder's theorem lies at the heart of almost all representation theorems proved in the following years up to the present time. It did not, however, stimulate much research until the late 1950s. The 1930s and 1940s saw an ongoing debate about the measurability of sensory variables, which seemed to some a prerequisite to laying a firm foundation for psychology, not unlike that of most of the natural sciences. In this debate, the introduction of scale types by Stevens (1946) was a cornerstone. The search for a precise formulation led to the model-theoretic formulation of Scott and Suppes (1958). This framework was extended in the three volumes *Foundations of Measurement* (Krantz et al., 1971; Suppes et al., 1989; Luce et al., 1990) and in the other books cited above.

In its present form, RTM consists of two relational structures $\mathscr{A}$ and $\mathscr{N}$. The base set of $\mathscr{N}$ is some subset of the reals or a real vector space. The objects which are empirically identified with the elements of A, the base set of $\mathscr{A}$, can be measured if it is possible to give or construct a homomorphism φ between the two structures. In most cases the structure $\mathscr{A}$ contains among its relations a weak order $\lesssim$. Some authors distinguish functions and constants such as a unit element from the proper relations. In our notation, these variants are all subsumed in $(R_i,\ i \in I)$. This sequence of relations can even be infinite. Sometimes it is necessary to endow the structures with several base sets and relations among them. This case can also be formulated within the present scheme, although it looks a bit clumsy.

The following are three of the major issues that confront RTM:

Representation: Does there exist a homomorphism between two structures, the qualitative and the numerical one, i.e., is there a mapping which preserves structure? This question is concerned with the measurability in principle of the variables under scrutiny.

Uniqueness: How can one describe the set of all representations of a given structure? This topic includes questions about the precise definition of scale type.

Meaningfulness: Which assertions about a measurement structure make sense, which do not? What is the meaning of numerical operations performed on the scale values? These questions are relevant to the applicability of statistical tests and to the content of descriptive indices.

The representation question is answered in RTM by formulating representation theorems; the next section gives two prominent examples. The uniqueness problem was initially dealt with by describing the set of all homomorphisms, this kind of result is usually part of the respective representation theorems (see Section 'Representation Theorems'). However, the modern theory differs in several aspects from this approach (see Section 'Scale Type and Modern Theory: Automorphisms of Measurement Structures'). The discussion of meaningfulness seems to have gone way beyond RTM. Several notions of invariance and meaningfulness have been proposed and analyzed, partly under the headline of dimensional analysis. We give but a short overview in Section Meaningfulness.

Representation Theorems

The procedure outlined in the introductory paragraph and the formal definition is exemplified in this section by presenting a few representation and uniqueness theorems and by discussing examples. Basically, there are a few general properties that can be exploited to produce relational structures which can be successfully mapped into a numerical structure. Primarily these are ordered structures. The empirical identification of the elements of the base set and the order relation is in most cases rather straightforward. The other relations of the structures are operations, or are derived from the comparison of intervals, and the property of conjointness (*see* Measurement Theory: Conjoint). In the present article we focus on extensive structures and difference structures.

Extensive Measurement

Our first structure was also historically the first to be considered. It goes back to Helmholtz (1887). It postulates a qualitative ordering and a concatenation operation, i.e., a binary operation $\circ$: $A \times A \to A$ which resembles the operation of addition of real numbers.

Definition 1. Let A be a nonempty set, $\lesssim$ a binary relation on A, and $\circ$ a closed operation on A, i.e., for all elements a, $b \in A$ there exists an element $c \in A$ such that $a \circ b = c$. The triple $(A, \lesssim, \circ)$ is a closed extensive structure if the following conditions (E1), (E2), (E3), and (E4) are satisfied for all a, b, c, $d \in A$; if in addition (E5) is satisfied it is a positive structure.

(E1) $\lesssim$ is a weak order on A, i.e., it satisfies $a \lesssim b$ and $b \lesssim c$ implies $a \lesssim c$, and $a \lesssim b$ or $b \lesssim a$.
(E2) $a \circ (b \circ c) \sim (a \circ b) \circ c$.
(E3) $a \lesssim b$ iff $a \circ c \lesssim b \circ c$ iff $c \circ a \lesssim c \circ b$.
(E4) If $a \prec b$, then for any c, d there exists a positive integer n such that $na \circ c \lesssim nb \circ d$, where na is defined inductively by $1a = a$ and $(n + 1)a = na \circ a$.
(E5) $a \prec a \circ b$.

The example of weights in Section 'Representation Theorems' of the article *Ordered Relational Structures* can be regarded as such a structure if one is willing to accept certain idealizations. It formalizes the measurement of mass by comparing objects on a pan balance. Another obvious example is the measurement of length, also modulo several idealizations.

The following theorem not only guarantees the existence of a numerical representation (a homomorphism into the reals) but also states that this existence implies the conditions of Definition 1. It reads as follows:

Theorem 1. Let $(A, \lesssim, \circ)$ be a system with $\lesssim$ a binary relation and $\circ$ a binary operation on A. Then $(A, \lesssim, o)$ is a closed extensive structure if and only if there exists a function φ: $A \to \mathbb{R}$ such that for all a, $b \in A$

1. $a \lesssim b$ iff $\varphi(a) \leq \varphi(b)$;
2. $\varphi(a \circ b) = \varphi(a) + \varphi(b)$.

Another function φ': $A \to \mathbb{R}$ satisfies these two properties iff there is a constant $r \in \mathbb{R}$ such that $\varphi' = r\varphi$. The structure is positive iff $\varphi(a) > 0$.

The first part is the existence theorem, the second part the uniqueness theorem. The latter can be rephrased as: A closed concatenation structure allows the measurement of the objects on a ratio scale.

This result has been extended in several ways. One point of improvement consists of dropping the assumption of closedness of the operation. A closed operation forces the possibility to concatenate arbitrary large and arbitrary small elements, an assumption that might be too strong in practice. Krantz et al. (1971) give sufficient conditions when the operation is not necessarily closed. Other modifications are concerned with minimal and maximal elements in the structure, such as the velocity of light in relativistic velocity concatenation. Recently, this problem has been generalized by regarding 'singular points' in the structure, i.e., elements that behave differently with respect to $\circ$ than do the rest of the elements.

Difference Measurement

While extensive operations abound in physics, they are rarely encountered in social sciences. However, the comparison of intervals is often used in psychological experiments. For instance, one can ask subjects whether the political system of two nations a, b is more similar to each other than that of the nations c, d. In economics the probability of choosing one alternative over another is used as an index of the size of a utility interval. In psychophysics, the method of bisection requires an observer to choose a stimulus whose value on some psychological scale such as loudness or brightness is between two given stimuli, such that the ensuing sensory differences are equal. This procedure assumes that people are able to compare intervals. Further examples are described in Chapter 4 of Krantz et al. (1971).

The following definition assumes a weak order $\lesssim$ on a subset A^* of pairs. For simplicity, the reader might assume that exactly one of the pairs ab and ba is in A^* for all a, $b \in A$.

The definition and theorem are more general, but this is the most important case.

Definition 2. Let A be a nonempty set, A^* a nonempty subset of the Cartesian product $A \times A$, and $\lesssim$ a binary relation on A^*. The triple $(A, A^*, \lesssim)$ is a positive difference structure if for all $a, b, c, d, a', b', c', a_1, a_2, \ldots \in A$.

(D1) $\lesssim$ is a weak order on A^*.
(D2) if $ab, bc \in A^*$, then $ac \in A^*$.
(D3) If $ab, bc \in A^*$, then $ab, bc \prec ac$.
(D4) If $ab, bc, a'b', b'c' \in A^*$, $ab \lesssim a'b'$, and $bc \lesssim b'c'$, then $ac \lesssim a'c'$.
(D5) If $ab, cd \in A^*$ and $ab \succ cd$ then there exist $d', d'' \in A$ such that $ad', d'b, ad'', d''b \in A^*$ and $ad' \sim cd \sim d''b$.
(D6) If $a_1, a_2, \ldots$ is a strictly bounded standard sequence, i.e., $a_i\, a_{i+1} \in A^*$ and $a_i a_{i+1} \sim a_1 a_2$ for $i = 1, 2,$. and for some $d'd'' \in A^*$ holds $a_1 a_i \prec d'd''$, then it is finite.

Theorem 2. Let $(A, A^*, \lesssim)$ be a positive difference structure. Then there exists a function $\psi: A^* \to \mathbb{R}^+$ such that for all $a, b, c, d \in A$

1. if $ab, cd \in A^*$, then $ab \lesssim cd$ iff $\psi(ab) \leq \psi(cd)$;
2. if $ab, bc \in A^*$, then $\psi(ac) = \psi(ab) + \psi(bc)$

If ψ' is another function with these properties, then there is some $r > 0$ such that $\psi' = r \cdot \psi$.

If for all $a \neq b$, $a, b \in A$ either ab or ba in A, then there exists $\varphi: A \to \mathbb{R}$ such that for all $ab \in A^*$, $\psi(ab) = \varphi(b) - \varphi(a)$. If φ' has the same property, then for some $s \in \mathbb{R}$, $\varphi' = \varphi + s$.

The uniqueness assertion of this theorem roughly says that the intervals, i.e., the elements of A^*, are measured on a ratio scale, while the elements of A are measured on an interval scale.

Other difference structures, called algebraic-difference structures, have been developed, allowing both for positive and negative intervals (Krantz et al., 1971). Even the case of two simultaneous orderings of intervals compatible in a well-defined sense has been considered. The representation theorems guarantee the measurement of one ordering by differences of the representing function, while the other is represented by their ratios. A modification of this theorem was used by Suck (1998) for a measurement theoretic characterization of the exponential distribution.

Comments on the Axioms

Some remarks are in order regarding the nature of the different axioms. Some describe particular properties of the structures, such as the order axioms, or associativity (E2) in Definition 1, or (D4) in Definition 2. They are called *structural* and are all testable for given elements of A. Another class of axioms are the solvability conditions such as (D5). They postulate the existence of certain elements, given a few others, such that a given 'equation' is solved. Often a consequence of this kind of axiom is that the structure becomes infinite. Characteristically, a solvability axiom is often a part of conditions that are sufficient for numerical representations; however, it is not a necessary condition. Testing it is a delicate matter if one does not find elements d', d'' in (D5), then their nonexistence is in general not proved.

A third class of conditions are Archimedean axioms, with Definition 1 (E4) and Definition 2 (D6) being examples. They are necessary for a representation with real numbers, because $\mathbb{R}$ satisfies a similar property (for all $x > y > 0$ there is an integer n such that $ny > x$). In ordered relational structures, the corresponding formulation employs the notion of standard sequences. These are increasing or decreasing sequences of elements of A, which are equally spaced in some sense specific for the structure. Examples are the sequences in (E4) and (D6).

Standard sequences are also very important for the construction of the homomorphisms in the representation theorems. This fact becomes clear if one studies the proofs of results such as Theorems 1 and 2. The key idea is as follows: a first standard sequence $a_1, a_2, \ldots$ is chosen. An element $a \in A$ is enclosed in $a_k \lesssim a \prec a_{k+1}$. The value k is taken as a first approximation to $\varphi(a)$. Then a second standard sequence $b_1, b_2, \ldots$ is constructed with each $b_{2i} = a_i$, i.e., a new element is introduced between two consecutive elements of the first sequence. If now $b_i \lesssim a \prec b_{i+1}$ then $i/2$ is the second approximation to the desired $\varphi(a)$. Proceeding in this way, a sequence of approximations is constructed. If this sequence can be shown to converge, then one has a value $\varphi(a)$. This φ can in many cases be shown to be one homomorphism. Other homomorphisms drop out by beginning with a different standard sequence. The uniqueness statements are derived from observing how these sequences are interrelated. This kind of proof mimics in some sense the actual procedures used in physics to measure objects. It is constructive and, depending on how many approximating steps are employed, controls the accuracy of the measurement.

Scale Type

Stevens (1946) introduced the concept of 'scale type' in response to a vivid debate on the possibility of evaluating subjective sensory events quantitatively. Much of the development of the present form of RTM in the 1950 and 1960s was influenced by the endeavor to give a sound foundation to Stevens' classical scale types, namely nominal, ordinal, interval, and ratio scales.

Roughly, the idea is to characterize scale type by describing the set of all homomorphisms of a representation in a uniqueness theorem. Such a description can be performed by giving the set of admissible transformations of a given homomorphism. Thus, if any strictly increasing transformation of a representation is admissible, then we have an ordinal scale if transformations of the form $x \to rx + s$ with $r > 0$ are admissible, we have an interval scale, and with $x \to rx$ we have a ratio scale. If necessary, one can define a nominal scale as one with all injective mappings (*see* Ordered Relational Structures) as admissible transformations.

Over time this simple theory needed several modifications. Log-interval scales and difference scales were introduced, mainly because the choice of the numerical structure was recognized to be highly arbitrary. Furthermore, not all scales could be given a clear place in this scheme. Roberts and Franke (1976) found that the scale type of semiorders (*see* Partial Orders) was 'irregular' compared with the previously considered types. This discussion was soon linked with invariance considerations and the meaningfulness problem. The scale type question was finally resolved by the work of Louis Narens (e.g., Narens, 1981a,b), which led to the consideration of the automorphism group and the degree of homogeneity and

uniqueness of an ordered structure (*see* Ordered Relational Structures). The next section discusses this theory.

Modern Theory: Automorphisms of Measurement Structures

The article on *Ordered Relational Structures* presents the homogeneity-uniqueness definition of scale type. Using automorphisms instead of homomorphisms and isomorphisms in this context frees the scale type and meaningfulness discussion from the arbitrariness of the choice of the representing structure. Furthermore, instead of applying an admissible transformation to a given homomorphism, one can apply an automorphism to the structure and then use the homomorphism. This way the possible representations drop out more naturally.

The first major success of this approach, which has been championed primarily by Narens, was the successful characterization of scale types (M, N) on the reals in the Alper/Narens Theorem. The type $(1, 1)$ corresponds to a ratio scale, $(2, 2)$ is an interval scale, $(1, 2)$ is something in between, and (∞, ∞) is an ordinal scale. Other scale types do not exist if the homomorphism is onto a real interval. This deep result explains why science does not employ scales of a type intermediate between the interval and ordinal type. However, if the range of the homomorphism is the rational numbers, Cameron (1989) and Macpherson (1996) showed that all types (M, N) with $1 \leq M \leq N$ are possible.

The main problem is now to classify structures by scale type. In the investigation of this question a subset of the automorphism group, namely the translations, plays a crucial role. An automorphism $\tau \in \mathrm{Aut}_{\mathscr{A}}$ is a *translation* if $\tau(a) \neq a$ for all $a \in A$ or if it is the identity map id_A. Luce and Narens (1994) formulate three major questions about translations which arose from the analysis of the proof of the Alper/Narens theorem:

1. Do the translations form a group?
2. Can the translations be ordered such that they satisfy an Archimedean property?
3. Are the translations 1-homogeneous?

If all these questions are answered affirmatively, a numerical representation can be constructed (see Luce and Narens, 1994).

Thus, a major problem is to find structural conditions that imply these three properties of the translations. But this seems to be a difficult problem, and not very much is known about it (see Luce, 2001). However, this shift of focus in the scale type discussion has already proved fruitful and has considerably increased the understanding of the philosophical problems of measurement. Possibly, final answers require a simultaneous solution of the meaningfulness problem which we discuss next.

Meaningfulness

In this section a short overview of the various aspects of the meaningfulness problem is given.

The 'meaning' of words or statements used to be a much debated issue in analytical philosophy in the first half of the twentieth century, as witnessed, for example, in the work of Carnap and others. In the same vein, and probably influenced by the philosophical discussion, the meaningfulness problem has been of interest from the very beginning of RTM. It is just the meaning of statements involving numerical values which is being debated. To give an example, consider the two statements:

1. The ratio between yesterday's maximum temperature and today's is 2.
2. The ratio between the difference of yesterday's maximum and minimum temperature and today's difference of the same values is 2.

Clearly, Statement (1) depends on the employed temperature scale. If it is correct using a Fahrenheit scale it is wrong in a Celsius scale, and vice versa (except for a few particular values). The second statement remains true or false in each temperature scale which is used nowadays. As a result, (1) seems meaningless, while (2) has a chance to be meaningful.

In its first formulation, the meaningfulness problem was closely linked to the uniqueness of scales and to scale type. Roughly, one seemed to settle with the formulation: a statement involving numerical scale values is meaningful if its truth value (i.e., true or false) does not change when an admissible transformation is applied to any of the scale values occurring in its formulation.

This approach falls short for several reasons, among which is the detection of irregular scales by Roberts and Franke (1976). More advanced theories of meaningfulness distinguish several concepts of invariance, such as reference and structure invariance. By the former it is understood that a relation is invariant under permissible transformations, while by the latter it is invariant under endomorphisms, i.e., homomorphisms of the structure into itself. These invariance properties were shown to provide necessary conditions on numerical relations for empirical meaningfulness. By contrast, the definition of a relation using the primitives of the structure may influence its meaningfulness. This is an intricate question connecting the meaningfulness problem to questions of definability.

Another aspect is dimensional analysis, which is a powerful tool in physics. The idea of this theory is to consider an equation involving the values of several different measurement homomorphisms, such as $f(x_1, \ldots, x_k) = 0$. If admissible transformations are applied to the scale values, the postulation of in-variance narrows down the possibilities for f. in most cases, this argument leads to a functional equation which, when solved, gives the form of the physical law underlying the original equation. Luce (1959) applied this principle to psychophysical laws. Roberts and Rosenbaum (1986) undertook a systematic investigation of this technique, and Narens and Mausfeld (1992) presented a theory on the relationship of the psychological and the physical in psychophysics resulting in what they call an 'equivalence principal'. This method turns out to be a tool to sort out quantitative formulae and statements that have psychological significance. The role of dimensional analysis to find or discuss 'lawful' expressions is elaborated at a more advanced level in Luce (1978) and Chapter 22.7 of Luce et al. (1990).

One of the issues of the meaningfulness discussion was the question of meaningful statistics. In it an answer is sought to the question of which statistical procedures are applicable to which kind of data. For example, one can look at the test statistic of a *t*-test. It consists of a ratio of a difference of mean values and a standard deviation (or a 'mixed' standard deviation calculated from two samples). It is easy to see that this value is invariant under affine transformations of the data, but may vary considerably if the data are transformed monotonically. Does this fact imply that the *t*-test is applicable to interval scales and not to ordinal scales? Opinions on this question seem to differ widely. Luce et al. (1990) give an overview of what RTM can contribute to this question. Their point is that it is in general not the employed statistical value (e.g., the mean value) itself which might be inappropriate, but rather the numerical relations involving this value. Therefore, a difference has to be made between the meaningfulness of hypotheses on populations and the meaningfulness of statistics used for the description of samples.

A completely new way of looking at the meaningfulness issue was recently suggested by Narens (2002). He seems to turn the tables, in the sense that RTM is justified in terms of a new formulation of meaningfulness. In this formulation one still captures the domain of interest in a qualitative relational structure, but instead of deriving meaningfulness from invariance properties of this structure, one assumes a theory of meaningfulness in terms of which the concept of a meaningful scale can be formulated. In his paper it is shown that for each (in this sense) meaningful scale there exists a mathematical representing structure such that the scale consists of homomorphisms between these structures. Thus, the meaningful problem is solved by giving it a more fundamental place in the scientific process. Future discussion and elaboration of this point of view will decide on its viability.

Problems of RTM

Philosophical debate regarding the foundations of RTM continues – see, for instance, critiques by Michell (1986), Niederée (1987, 1992, 1994), and Dzhafarov (1995). The main points of criticism are analyzed and discussed in Luce and Narens (1994).

A major driving force of RTM has been the idea that applied scientists would test the axioms of representation theorems, and if the fit was satisfactory, they would construct a scale accordingly, using the uniqueness part for a justification of the operations performed in the course of the research on the topic. This program, if it was a program, seems not yet to have been implemented. Others would say it has failed. There are very few attempts to test axioms for this purpose. Luce's (2000) book on utility theory is an exception and may point a way of renewing this program.

A major obstacle to an application of RTM in the way described is the lack of a satisfactory error theory of measurement within RTM. To what extent must an axiom be violated to be sure it does not hold? To give an example: even for physical length measurement, one can observe many instances of violations of the testable axioms of Theorem 1 (see Section 'Extensive Measurement'). This will necessarily happen if the investigated objects differ only very little in length, less than the accuracy of the measuring device.

There have been numerous attempts to extend the classical 'deterministic' RTM to 'nondeterministic' versions by adding a probabilistic structure. However, none has provided a practical device for a theory of measurement error within RTM.

Another critical issue is the almost ubiquitous use of infinite structures. More than that the most elegant and useful results, such as the Alper/Narens Theorem, are obtained for continuum structures, i.e., ordered structures which satisfy Dedekind completeness (all infima and suprema exist, *see* Ordered Relational Structures). It can be surmised that a powerful error theory calls for mappings on a continuum.

The idea of embedding structures in Dedekind complete ones using extension procedures such as Dedekind cuts works in some cases but is intricate in others (*see* Ordered Relational Structures and Partial Orders). In what sense finite data sets are approximating such structures is a problem of theoretical importance for which not very much is known as yet. Sommer and Suppes (1997) proposes a critical approach to the use of the continuum in this context.

There are more results needed on classification by scale type. In particular, structural equivalents of homogeneity and uniqueness of automorphisms are not yet known. By structural equivalents, we mean conditions formulated within the structure and not within the automorphism group. Likewise, it is unclear how properties of the translations can be formulated or derived within $\mathscr{A}$.

Finally, the meaningfulness question is one of the outstanding open problems of RTM. The recent approach of Narens (submitted) mentioned in Section 'Meaningfulness' appears completely new from the perspective of RTM, and it might possess the quality of a Copernicanean revolution.

Applications of RTM

It has already been mentioned that axiom testing is only rarely performed. But there are such investigations, mostly in connection with conjoint measurement (*see* Measurement Theory: Conjoint Analysis Applications). Much of the research on color vision, spatial vision, and other psychophysical questions uses representations of measurement structures. The benefit of RTM in such investigations is often theoretical: it provides a firm basis for formulating hypotheses and performing tests. As an example we cite Heller (1997) on binocular space perception and the elaborate mathematical development ensuing from this paper (see Aczél et al., 1999).

An interesting application of RTM can be found in dimensional analysis and the derivation of numerical laws in physics and psychophysics. Chapter 10 of Krantz et al. (1971) is the paradigm for this kind of application. Since then many results have been published. We cite one important article: Narens and Mausfeld (1992).

Finally, we mention the application of measurement theory in decision making. Expected utility theory, its subjective counterparts, e.g., Prospect theory, rank dependent utility (Luce, 2000), etc all rely to a great extent on results from RTM. This tendency is kept to the present day and is

likely to remain so for some time. Luce (2000) is an outstanding recent example. The papers Luce and Narens (1994) and Luce (1995) give an in-depth overview of the general discussion.

See also: Mathematical Psychology; Measurement Theory: Conjoint Analysis Applications; Order Statistics; Ordered Relational Structures; Statistical Analysis, Special Problems of: Outliers.

Bibliography

Aczél, J., Baros, Z., Heller, J., Ng, C.T., 1999. Functional equations in binocular space perception. Journal of Mathematical Psychology 43, 71–101.

Cameron, P.J., 1989. Groups of order-automorphisms of the rationals with prescribed scale type. Journal of Mathematical Psychology 33, 163–171.

Cecconi, P., Franceschini, F., Galetto, M., January 2006. Measurements, evaluations and preferences: a scheme of classification according to the representational theory. Measurement 39 (1), 1–11.

Dzhafarov, E.N., 1995. Empirical meaningfulness, measurement-dependent constants, and dimensional analysis. In: Luce, R.D., D'Zmura, M., Hoffman, D., Iverson, G.J., Romney, A.K. (Eds.), Geometric Representations of Perceptual Phenomena: Papers in Honor of Tarow Indow on His 70th Birthday. Erlbaum, Mahwah, NJ.

Heller, J., 1997. On the psychophysics of binocular space perception. Journal of Mathematical Psychology 41, 29–43.

Helmholtz, H., 1887. Zählen und Messen erkenntnistheoretisch betrachtet. Philosophische Aufsatze Edward Zeller gewidmet, Leipzig. Reprinted in Gesammelte Abhandlungen, vol. 3, pp. 356–391(1895). English translation by Bryan, C.L., 1930. Counting and Measuring. Van Nostrand, Princeton, NJ.

Hölder, O., 1901. Die Axiome der Quantitat und die Lehre vom Mass. In: Berichte uber die Verhandlungen der Koniglich Sachsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-physische Klasse, vol. 53, 1–64.

Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A., 1971. Foundations of Measurement, vol. 1. Academic Press, New York.

Kyngdon, Andrew, February 2008. The Rasch model from the perspective of the representational theory of measurement. Theory and Psychology 18 (1), 89–109.

Luce, R.D., 1959. On the possible psychophysical laws. Psychological Review 66, 81–95.

Luce, R.D., 1978. Dimensionally invariant numerical laws correspond to meaningful qualitative relations. Philosophy of Science 45, 1–16.

Luce, R.D., 1995. Four tensions concerning mathematical modeling in psychology. Annual Review of Psychology 46, 1–26.

Luce, R.D., 2000. Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches. Erlbaum, Mahwah, NJ.

Luce, R.D., Narens, L., 1994. Fifteen problems concerning the representational theory of measurement. In: Humphreys, P. (Ed.), Patrick Suppes: Scientific Philosopher, vol. 2. Kluwer Dordrecht, The Netherlands, pp. 219–249.

Luce, R.D., Krantz, D.H., Suppes, P., Tversky, A., 1990. Foundations of Measurement, vol. 3. Academic Press, New York.

Luce, R.D., 2001. Conditions equivalent to unit representations of ordered relational structures. Journal of Mathematical Psychology 45 (1), 81–98.

Macpherson, D., 1996. Sharply multiply homogeneous permutation groups, and national scale types. Forum Mathematicum 8, 501–507.

Michell, J., 1986. Measurement scales and statistics: a clash of paradigms. Psychological Bulletin 100, 398–407.

Narens, L., 1981a. A general theory of ratio scalability with remarks about the measurement – theoretic concept of meaningfulness. Theory and Decision 13, 1–70.

Narens, L., 1981b. On the scales of measurement. Journal of Mathematical Psychology 24, 249–275.

Narens, L., 1985. Abstract Measurement Theory. MIT Press, Cambridge, MA.

Narens, L., Mausfeld, R., 1992. On the relationship of the psychological and the physical in psychophysics. Psychological Review 99, 467–479.

Narens, L., 2002. A meaningful justification for the representational theory of measurement. Journal of Mathematical Psychology 46 (6), 746–768.

Niederee, R., 1987. On the reference to real numbers in fundamental measurement: a model-theoretic approach. In: Roskam, E.E., Suck, R. (Eds.), Progress in Mathematical Psychology, vol. 1. North-Holland, Amsterdam, pp. 3–23.

Niederee, R., 1992. What do numbers measure? A new approach to fundamental measurement. Mathematical Social Sciences 24, 237–276.

Niederee, R., 1994. There is more to measurement than just measurement: measurement theory, symmetry, and substantive theorizing. Journal of Mathematical Psychology 38, 527–594.

Pfanzagl, J., 1971. Theory of Measurement. Physika, Würzburg, Germany.

Roberts, F.S., 1979. Measurement Theory with Applications to Decision Making, Utility, and the Social Sciences. Addison–Wesley, Reading, MA.

Roberts, F.S., Franke, C.H., 1976. On the theory of uniqueness in measurement. Journal of Mathematical Psychology 14, 211–218.

Roberts, F.S., Rosenbaum, Z., 1986. Scale type, meaningfulness, and the possible psychophysical laws. Mathematical Social Sciences 12, 77–95.

Rossi, Giovanni Battista, January 2006. A probabilistic theory of measurement. Measurement 39 (1), 34–50.

Scott, D., Suppes, P., 1958. Foundational aspects of theories of measurement. Journal of Symbolic Logic 23, 113–128.

Sommer, R., Suppes, P., 1997. Dispensing with the continuum. Journal of Mathematical Psychology 41, 3–10.

Stevens, S.S., 1946. On the theory of scales of measurement. Science 103, 677–680.

Suck, R., 1998. A qualitative characterization of the exponential distribution. Journal of Mathematical Psychology 42, 418–431.

Suppes, P., Krantz, D.H., Luce, R.D., Tversky, A., 1989. Foundations of Measurement: Geometrical, Threshold, and Probabilistic Representations, vol. 2. Academic Press, New York.

Measurement Theory: Conjoint

Peter Fishburn, AT&T Laboratories, Florham Park, NJ, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9448–9451, © 2001, Elsevier Ltd.

Abstract

The theory of conjoint measurement concerns quantitative measurement of a qualitative aspect of objects that are characterized by a number of attributes. The qualitative aspect is often a subjective notion of value or preference, and its quantitative counterpart is a sum of numerical values or utilities over the attributes. If object x is characterized by levels x_1, x_2, and x_3 of three attributes, its numerical valuation is $u_1(x_1) + u_2(x_2) + u_3(x_3)$, where ui is a utility function for attribute i. A particular model of conjoint measurement for this case is based on axioms or assumptions about preferences between objects which imply the existence of u_1, u_2, and u_3 such that, for any two objects x and y, x is preferred to y if and only if $u_1(x_1) + u_2(x_2) + u_3(x_3) > u_1(y_1) + u_2(y_2) + u_3(y_3)$. The axioms make assumptions about ordering properties for preferences and tradeoffs among attributes that lead to the additive utility representation. A variety of axiomatic schemes arise from different attribute structures and different problem settings.

The theory of conjoint measurement is a branch of measurement theory in which objects described by several attributes are assigned numerical values that preserve an ordering of the objects by a qualitative relation such as preference or likelihood. Numerical values of objects usually are presumed to equal the sum of values assigned separately to each attribute. A conjoint measurement theory consists of the structure assumed for objects and their attributes, assumptions about the qualitative relation among objects, a numerical representation that connects the qualitative ordering of objects to numerical values, and a description of all sets of numerical values that satisfy the representation.

Objects and Attributes

The objects are members $x, y, \ldots$ of a subset X of the Cartesian product $X_1 \times X_2 \times \ldots \times X_n$ of $n \geq 2$ other sets referred to as attributes, factors, or criteria. When x_i is the member of attribute X_i associated with object x, $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$, … Any set of entities that share a common set of attributes could serve as X, including consumption bundles (each attribute is a good, x_i is the amount of attribute i in bundle x), investment portfolios (stocks, bonds), houses (location, style, size, cost), candidates for a position (education, experience, skills, quality of references), and a host of other possibilities.

A particular theory specifies the structure of the X_i and X. Structural aspects are the number of attributes, whether X equals or is a proper subset of $X_1 \times \ldots \times X_n$, and the nature of each X_i (finite, infinite but discrete, a continuum). One theory assumes that n is any integer greater than 1, X is an arbitrary subset of $X_1 \times \ldots \times X_n$, and every X_i is a finite set. Another assumes that $n = 2$, $X = X_1 \times X_2$, and each X_i, is a continuum. Additional restrictions on structure may be imposed by preference axioms.

Preferences and Axioms

Qualitative comparisons between objects in X are described by a binary relation $\succsim$ on X and induced relations $\succ$ and $\sim$ defined by $x \succ y$ if $x \succsim y$ and not $(y \succsim x)$, and $x \sim y$ if $x \succsim y$ and $y \succsim x$. When preference is primitive, $x \succsim y$ signifies that x is at least as preferable as y, $x \succ y$ means that x is strictly preferred to y, and $x \sim y$ denotes equally preferred objects. When objects are uncertain events, $x \succ y$ if x is judged to be at least as likely to occur as y, $x \succ y$ if x is judged to be more probable than y, and $x \sim y$ if x and y are considered equally likely.

Conditions imposed on the behavior of $\succsim$ on X are referred to as axioms. Although some axioms describe intuitively rational or reasonable behaviors, they are empirically refutable assumptions.

Ordering Axioms

It is often assumed that $\succsim$ is a weak order, that is, that $\succsim$ is complete (for all $x, y \in X : x \succsim y$ or $y \succsim x$) and transitive (for all $x, y, z \in X :$ if $x \succsim y$ and $y \succsim z$ then $x \succsim z$). When $\succsim$ is a weak order, $\sim$ is an equivalence relation (reflexive: $x \sim x$; symmetric: $x \sim y \Rightarrow y \sim x$; transitive) that partitions X into equivalence classes, with $x \sim y$ if and only if x and y are in the same class. A less common but more realistic axiom presumes only that $\succ$ is asymmetric $(x \succ y \Rightarrow \text{not}(y \succ x))$ and transitive, in which case $\sim$ is not necessarily transitive, as when $x \sim y$, $y \sim z$, and $x \succ z$. Still less common is the axiom which assumes that $\succsim$ is complete but not necessarily transitive.

Independence and Cancellation Conditions

Independence and cancellation conditions are needed for the additive utility representation in the next section. The most basic independence condition assumes that, for each i, if $x_i = z_i$ and $y_i = w_i$, and $x_j = y_j$ and $z_j = w_j$ for all $j \neq i$, then $x \succsim y \Leftrightarrow z \succsim w$. When $\succsim$ is a weak order on $X = X_1 \times \ldots \times X_n$, this implies that the relation $\succsim_i$ on X_i defined by

$$x_i \succsim_i y_i \text{ if } x \succsim y \text{ whenever } x_j = y_j \text{ for all } j \neq i \qquad [1]$$

is a weak order. Moreover, if $x_i \succsim_i y_i$ for all i, then transitivity implies $x \succsim y$, with $x \succ y$ if $x_i \succ_i y_i$ for at least one i.

A more general first-order cancellation axiom asserts that if $\{x_i, z_i\} = \{y_i, w_i\}$ for all i, then $x \succsim y \Leftrightarrow w \succsim z$. This extends to the general cancellation condition which says that if $m \geq 2$, if

$x^1, \ldots, x^m$, $y^1, \ldots, y^m \in X$, and if $y_i^1, \ldots, y_i^m$ is a permutation of $x_i^1, \ldots, x_i^m$ for every i, then it is not true that $x^j \succsim y^j$ for $j = 1, \ldots, m$ with $x^j \succ y^j$ for some j. The second-order cancellation axiom is the special case of general cancellation for $m = 3$. An instance for $n = 2$ is : it is false that $(x_1, x_2) \succsim (y_1, z_2), (y_1, y_2) \succsim (z_1, x_2)$, and $(z_1, z_2) \succ (x_1, y_2)$. Its $\sim$ version, whose conclusion is $\{x^1 \sim y^1, x^2 \sim y^2\} \Rightarrow x^3 \sim y^3$, illustrates a notion of higher-order tradeoffs and is sometimes referred to as the Thomsen condition.

Solvability and Archimean Axioms

Solvability axioms posit the existence of attribute values that allow compensating tradeoffs and thereby impose further restrictions on the X_i and X. An example for $n = 2$ is: for every $(x_1, x_2) \in X$ and every $y_1 \in X_1$, there is a $y_2 \in X_2$ for which $(y_1, y_2) \sim (x_1, x_2)$.

Archimedean axioms, sometimes referred to as continuity conditions, are technical conditions for infinite structures that ensure the existence of real-valued utilities in representations of $\succsim$. When each X_i is a nondegenerate real interval and $\succ$ is the lexicographic order defined by $(x_1, x_2) \succ (y_1, y_2)$ if $x_1 > y_1$ or $(x_1 = y_1, x_2 > y_2)$, there is no real-valued function u on X that satisfies $x \succ y \Leftrightarrow u(x_1, x_2) > u(y_1, y_2)$. Archimedean axioms forbid this type of noncompensatory dominance structure.

Additive Representations

The basic additive conjoint representation of $\succsim$ on X consists of a real-valued function u_i on X_i for $i = 1, \ldots, n$ such that, for all $y \in X$

$$x \succsim y \Leftrightarrow \sum_{i=1}^{n} u_i(x_i) \geq \sum_{i=1}^{n} u_i(y_i) \qquad [2]$$

An alternative is the multiplicative representation $x \succsim y \Leftrightarrow f_1(x_1) \ldots f_n(x_n) \geq f_1(y_1) \ldots f_n(y_n)$, where $f_i = e^{u_i}$. Other algebraic combinations of attribute utilities are discussed in Krantz et al. (1971) under polynomial measurement. Additional restrictions may be imposed on the u_i of eqn [2] for representations of qualitative probability.

Uniqueness

Let $v_i = au_i + b_i$ with $a > 0$ and $b_1, \ldots, b_n$ any real numbers. When eqn [2] holds, it holds also for $v_1, \ldots, v_n$ in place of $u_1, \ldots, u_n$. When no other transformations of the u_i satisfy eqn [2], the attribute utilities for the additive representation are unique up to similar positive affine transformations. The common $a > 0$ for the transformations preserves scale commensurability of the attribute utility functions. When X is finite, more general transformations are possible unless enough $\sim$ comparisons hold to force uniqueness up to similar positive affine transformations (Fishburn and Roberts, 1989).

Specific Theories

When X is finite, eqn [2] reduces to a finite number of linear inequalities with coefficients in $\{1, 0, -1\}$ which have a solution if and only if $\succsim$ satisfies weak order and general cancellation (Kraft et al., 1959; Fishburn, 1970; Krantz et al., 1971).

When the X_i are infinite, $X = X_1 \times \ldots \times X_n$, and $n \geq 3$ weak order, first-order cancellation, and an Archimedean axiom lead to u_i for eqn [2] that are unique up to similar positive affine transformations. A topological approach (Debreu, 1960; Fishburn, 1970; Wakker, 1989) also endows X with special topological structure, and an algebraic approach (Fishburn, 1970; Krantz et al., 1971) uses a solvability condition. Similar approaches apply for $n = 2$, where a second-order cancellation condition is also required.

Another weak-order theory considers a denumerable number of attributes that refer to successive time periods and involve notions of discounting and impatience (Fishburn, 1970). The relaxation of weak order where $\succ$ is asymmetric and transitive leads to the one-way representation

$$x \succ y \Rightarrow \sum_{i=1}^{n} u_i(x_i) > \sum_{i=1}^{n} u_i(y_i) \qquad [3]$$

for partially ordered preferences (Fishburn, 1970). Transitivity can be omitted without affecting additively as shown by the nontransitive additive conjoint representation

$$x \succsim y \Leftrightarrow \sum_{i=1}^{n} \phi_i(x_i, y_i) \geq 0 \qquad [4]$$

in which ϕ_i is a skew-symmetric $(\phi_i(a, b) + \phi_i(b, a) = 0)$ real-valued function on $X_i \times X_i$ (Fishburn, 1991).

Conjoint Measurement and Social Choice

Ideas from conjoint measurement have been applied to social choice theory (Fishburn, 1973) under the following interpretations. Each i indexes one of n voters, and $\succsim_i$ denotes the preference relation of voter i. Unlike eqn [1], the $\succsim_i$ are primitive relations not derived from $\succsim$, and $\succsim$ takes the role of a social preference relation which is a function of the voter preference profile $(\succsim_1, \ldots, \succsim_n)$. Conditions of order, fairness, positive responsiveness, and so forth, can then be proposed to relate $\succsim$ to realizations of $(\succsim_1, \ldots, \succsim_n)$. Some sets of conditions lead to additive representations for simple majority, weighted majority, and more general rules, whereas others are collectively incompatible (Arrow, 1963) and result in impossibility theorems.

Decision Under Risk and Uncertainty

Conjoint measurement is used extensively for multi-attribute outcomes in decision under risk and uncertainty (Keeney and Raiffa, 1976). Consider the expected-utility representation

$$p \succsim q \Leftrightarrow \sum_{x \in X} p(x)u(x) \geq \sum_{x \in X} q(x)u(x) \qquad [5]$$

where p and q are simple probability distributions on $X = X_1 \times \ldots \times X_n$ and u on X is unique up to a positive affine transformation $au + b$, $a > 0$. Let p_i be the marginal distribution on X_i of p. The independence condition which says that $p \sim q$ whenever $p_i = q_i$ for all i implies an additive decomposition $u(x_1, \ldots, x_n) = u_1(x_1) + \ldots + u_n(x_n)$ for outcome utilities (Fishburn, 1970: p. 149). A less-restrictive condition says that, for every i, the preference order over marginal distributions on

X_i at fixed values of the other attributes is independent of those fixed values. This implies either the additive decomposition or a multilinear form which under somewhat stronger conditions can be expressed as

$$ku(x_1, \ldots, x_n) + 1 = [ku_1(x_1) + 1] \ldots [ku_n(x_n) + 1] \qquad [6]$$

where k is a nonzero constant. A number of other decompositions of u are also described in Keeney and Raiffa (1976).

See also: Measurement Theory: Conjoint Analysis Applications; Probability Theory: Interpretations.

Bibliography

Arrow, K.J., 1963. Social Choice and Individual Values, second ed. Wiley, New York.

Debreu, G., 1960. Topological methods in cardinal utility theory. In: Arrow, K.J., Karlin, S., Suppes, P. (Eds.), Mathematical Methods in the Social Sciences, 1959. Stanford University Press, Stanford, CA.

Fishburn, P.C., 1970. Utility Theory for Decision Making. Wiley, New York.

Fishburn, P.C., 1973. The Theory of Social Choice. Princeton University Press, Princeton, NJ.

Fishburn, P.C., 1991. Nontransitive additive conjoint measurement. Journal of Mathematical Psychology 35, 1–40.

Fishburn, P.C., Roberts, F.S., 1989. Uniqueness in finite measurement. In: Roberts, F.S. (Ed.), Applications of Combinatorics and Graph Theory in the Biological and Social Sciences. Springer-Verlag, Berlin.

Keeney, R.L., Raiffa, H., 1976. Decisions with Multiple Objectives: Preferences and Value Tradeoffs. Wiley, New York.

Kraft, C.H., Pratt, J.W., Seidenberg, A., 1959. Intuitive probability on finite sets. Annals of Mathematical Statistics 30, 408–419.

Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A., 1971. Foundations of Measurement, vol. 1. Academic Press, New York.

Wakker, P.P., 1989. Additive Representations of Preferences. Kluwer, Dordrecht, Holland.

Measurement Theory: Conjoint Analysis Applications

Jordan J Louviere, University of Sydney, NSW, Australia

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 4, pp. 2565–2568, © 2001, Elsevier Ltd.

Abstract

This article reviews the evolution and development of conjoint analysis (CA) applications and their similarities, differences, strengths, and weaknesses. Traditionally, CA used experimental designs to vary levels of attributes/features hypothesized to underlie decisions and to create combinations of factor levels (e.g., a bus trip of 30 min with service every 15 min at $2, or a trip of 15 min at $2.75). Subjects evaluate each designed combination holistically in category rating or ranking tasks; their responses are decomposed by statistical analysis, typically using general linear statistical models. More recently, methods based on random utility theory have been developed and applied. These new methods use experimental design techniques not only to vary attributes and levels but also to create choice sets from which subjects select one option. Choices are decomposed using discrete multivariate statistical models. Very recently, theoretical links between choices in experiments and in real settings have been derived and investigated, and several empirical tests suggest that these links hold to a close first approximation. Additionally, comparisons of experimental and other 'stated' choices suggest that these theoretical links can be extended to many preference data source comparisons.

Introduction

This article provides a general discussion of conjoint analysis (CA) applications and identifies and reviews principal contributions, streams of research, and applications within the overall field. CA is a generic term for methods developed to model and measure preferences and trade-offs that share several common features:

(a) independent variables or 'attributes' that explain preferences are identified and defined (e.g., bus travel times and fares; MHz and prices of PCs);
(b) attribute 'levels' relevant to the research are assigned (e.g., 30, 60 min travel time or $1, $3 fare for buses; 600, 733 MHz and $2500, $2850 prices for PCs);
(c) statistical design techniques are used to combine the attribute levels and create descriptions of options (e.g., 15 min travel time for $2 or 733 MHz for $2850);
(d) a sample of individuals evaluate all or a subset of the descriptions;
(e) individual- or group-level models are estimated from the evaluation data (e.g., transport modes or PC options); and
(f) the estimated models are used to measure subjective quantities (e.g., the subjective value of 600 vs. 733 MHz or implied willingness to pay to save 30 min travel time).

Despite commonalities, CA methods also differ in experimental designs, response tasks, model form(s), and measurement of subjective quantities. For example, initial CA applications were based on an axiomatic theory of the behavior of a ranking of all possible combinations of a discrete set of attribute levels (e.g., Krantz and Tversky, 1971). All possible combinations of attribute levels are given by complete factorial enumeration, such that if $A_1(I)$, $A_2(J)$, …, $A_n(N)$ represent the number of levels of attributes 1, 2, …, N, then all combinations are given by the expression $I \times J \times \ldots \times N$. Thus, in a full factorial experiment the total combinations equal the product of the levels of each attribute. For example, if the attributes (levels) are bus fare (3), travel time (3), and service frequency (3), then there are $3 \times 3 \times 3$ (or 3^3) $= 27$ possible bus descriptions.

If a subject ranks all designed descriptions in order of preference (or another subjective dimension), and certain axiomatic conditions are satisfied, the ranking can be represented by an algebraic model defined by marginal and joint attribute effects. For example, for two or more factors, a ranking of a factorial set of options is additive in the attribute effects, if, and only if, the effects of all attributes on the ranking are independent of one another and there are no joint attribute effects (i.e., only marginal effects apply). That is, an additive process underlying a ranking of all combinations of, say, the ith, jth, and kth levels of attributes A_{1i}, A_{2j}, and A_{3k}, respectively, can be expressed as:

$$V_{ijk} = c + \upsilon(A_{1i}) + \upsilon(A_{2j}) + \upsilon(A_{3k}) \qquad [1]$$

where V_{ijk} is the overall value of attribute level combination ijk, c is a constant that sets the response scale origin and $\upsilon(A_{1i})$, $\upsilon(A_{2j})$, and $\upsilon(A_{3k})$ are values of the ith, jth, and kth levels of each attribute.

Statistically, if the underlying process is additive, all attribute interactions will be nonsignificant. If the underlying process is multiplicative, the model is as follows:

$$V_{ijk} = c^* + \upsilon(A_{1i})\upsilon(A_{2j})\upsilon(A_{3k}) \qquad [2]$$

where all terms were previously defined, except c^*, a constant to set the scale origin. Equation [2] requires all attribute main effects and interactions to be significant. Early applications lacked error theories, hence used goodness-of-fit tests to evaluate and compare models.

Full factorial designs proved too large, so ways to reduce numbers of combinations were proposed, such as fractional or other partial factorial designs, ranking all pairs of attribute levels or using incomplete block designs to select attributes and levels (e.g., Green, 1974). Most applications rely on non replicated fractional factorials that cannot be used to test model

forms and produce few observations relative to parameters for individuals (Louviere, 1988). However one can estimate individual models if response, design, and model assumptions are satisfied (e.g., equal interval response scales, insignificant interactions, independent and identically distributed normal errors).

A new paradigm emerged when McFadden (1974) developed computationally tractable extensions of random utility theory (RUT) to the multiple-choice case. This provided ways to estimate and test choice models with a formal theory of errors based on RUT. RUT assumes an unobservable value, U_i, for each of $i = 1, 2, \ldots, I$ options that is a sum of observable (explainable) and random (unexplainable) components:

$$U_i = V_i + \varepsilon_i \quad [3]$$

where V_i and ε_i are, respectively, observable and random components. RUT assumes that individuals try to choose their preferred options, but the random component implies that choices are inherently stochastic from a researcher's perspective: that is, the probability that a subject will choose an option from a set of options is predictable, not the exact choice. If $P(i|C)$ is the probability of choosing option i in set C, and Max = maximum operator, then

$$P(i|C) = P[(V_i + \varepsilon_i) > \mathrm{Max}(V_j + \varepsilon_j), \text{ for all } j \text{ options in } C] \quad [4]$$

Different discrete choice models are derived by making assumptions about the distributions of the ε's. For example, McFadden (1974) derived the multinomial logit model (MNL) by assuming that the ε's were independently and identically distributed Gumbel random variates. MNL choice probabilities have a simple closed-form, but if the ε's are non-independently and identically distributed normals, the nonclosed form multinomial probit (MNP) model results.

Louviere and Woodworth (1983) integrated RUT-based choice models with CA, and proposed ways to design choice experiments consistent with RUT-based choice models. This choice-based (CB) approach designs attribute level combinations (to estimate evaluation/utility functions) and selects choice sets in which they appear (to satisfy properties of choice models), and the error theory is used to select and test model forms. The CB approach typically estimates models from samples of subjects because individuals provide too few observations to satisfy asymptotic theory and obtain reliable model estimates.

Choice experiments mimic real choice situations. For example, to model choice of travel mode to work, one designs experiments to offer commuters sets of mode choices; to model choice of health insurance plan, one designs experiments that offer subjects sets of health insurance options. Thus, the CB approach designs attribute level combinations to describe sets of options (say M of them). If $M = 1$, subjects evaluate each option separately, making Mbinary choices (e.g., will/will not choose); if $M > 1$, subjects evaluate sets of multiple options (say S of them), making S multiple choices (Louviere and Woodworth, 1983; Louviere et al. 2000).

Significant progress has been made in linking behavior in choice experiments and real environments. Specifically, in RUT models the magnitude (or 'scale') of the parameters are inversely proportional to choice variability. For example, if we denote the scale as λ, it can be defined as follows for MNL models (Ben-Akiva and Lerman, 1985 : 104–105): $\lambda^2 = \pi^2/6\sigma_\varepsilon^2$ (π is the natural constant; σ_ε^2 = error variance). In empirical applications, analysts estimate the quantities $\lambda\beta_k$, not β_k (β_k is a k-element parameter vector associated with design matrix, X_{ki}, and the ith choice option).

Thus, λ is not identified in any one data source, and typically is set equal to one for convenience. Ben-Akiva and Morikawa (1990) showed that ratios of λ's can be identified and estimated from two or more choice data sources; hence, if experimental and actual choices differ only in response variability, parameters estimated from each will be proportional. Empirical tests in several fields suggest that parameter proportionality often holds to a close first approximation (see Louviere et al. 1999).

This property of RUT-based choice models allows many sources of preference and choice data to be compared, including traditional CA data. For example, one can transform CA ranks into implied choice sets by 'exploding' subjects' complete or partial orders into one or more choice sets (Louviere et al. 2000). That is, for a ranking of M options, the first is preferred to the other M–1, the second to the remaining M–2, etc; hence M rankings yields M–1 choice sets (if A, B, C, D are ranked 1, 2, 3, 4 the first set contains A, B, C, D with A chosen; the second set contains B, C, D with B chosen; etc). Such explosions transform preference responses into discrete choices that can be compared using RUT-based statistical models (e.g., Louviere et al. 2000). Recently Bayesian methods have been used to estimate individual-level parameters (Wedel et al. 1999), but the pros and cons of these methods in CA and CB applications have not been fully explored.

Statistical properties of choice experiments also remain elusive. In particular, humans interact with CA and choice experiments and particular designs can impact response variability. Statistically efficient designs may require many choices, which may increase choice variability; increases in choice variability may offset higher statistical design efficiency. Hence, research into behavioral impacts of designs and effects of non design issues like task layouts, formats, response tasks, etc, on choice variability also are needed.

See also: Measurement Theory: Conjoint; Measurement Theory: History and Philosophy; Psychometrics: Preference Models with Latent Variables; Psychometrics; Test Theory: Applied Probabilistic Measurement Structures.

Bibliography

Anderson, N.H., 1970. Functional measurement and psychophysical judgment. Psychological Review 77, 153–177.

Ben-Akiva, M.E., Lerman, S., 1985. Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press, Cambridge, MA.

Ben-Akiva, M., Morikawa, T., 1990. Estimation of switching models from revealed preferences and stated intentions. Transportation Research A 24A, 485–495.

Green, P.E., 1974. On the design of choice experiments involving multifactor alternatives. Journal of Consumer Research 1, 61–68.

Hensher, D.A., Louviere, J.J., Swait, J., 1999. Combining sources of preference data. Journal of Econometrics 89, 197–221.

Krantz, D.H., Tversky, A., 1971. Conjoint measurement analysis of composition rules in psychology. Psychological Review 78, 151–169.

Louviere, J.J., 1988. Analyzing decision making: metric conjoint analysis. Sage University Papers Series in Quantitative Applications in the Social Sciences, Newbury Park, CA. No. 67.

Louviere, J.J., Woodworth, G.G., 1983. Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data. Journal of Marketing Research 20, 350–367.

Louviere, J.J., Hensher, D.A., Swait, J., 2000. Stated Choice Methods: Analysis and Applications in Marketing, Transportation and Environmental Valuation. Cambridge University Press, Cambridge, UK.

Louviere, J., Meyer, R., Bunch, D., Carson, R., Delleart, B., Hanemann, M., Hensher, D., Irwin, J., 1999. Combining sources of preference data for modelling complex decision processes. Marketing Letters 10 (3), 205–218.

Luce, R.D., 1959. Individual Choice Behavior. Wiley, New York.

McFadden, D., 1974. Conditional logit analysis of qualitative choice behaviour. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.

Wedel, W., Kamakura, W., Arora, N., Bemmaor, A., Chiang, J., Elrod, T., Johnson, R., Lenk, P., Neslin, S., Poulsen, C.S., 1999. Discrete and continuous representation of unobserved heterogeneity in choice modelling. Marketing Letters 10 (3), 219–232.

Measurement Theory: History and Philosophy

Joel Michell, University of Sydney, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Measurement theory is the attempt to understand scientific measurement. It is at least as old as Euclid's *Elements*, where concepts basic to the classical understanding are defined. At first the classical concept applied only to extensive quantities, but in the Middle Ages it was expanded to include intensive quantities. It was further developed to include the concepts of unbounded, continuous quantity and fundamental, derived, and conjoint measurement, all of which raised questions about practices called 'psychological measurement.' The classical concept was later challenged by representational and operational concepts, but it still dominates within the paradigmatically quantitative sciences.

Measurement theory is the attempt to understand scientific measurement. Quantitative concepts and methods infuse the entire corpus of modern physical science. The presumption that physics is the methodological touchstone for science led psychologists to propose their own, special methods intended to measure psychological attributes. After a century or more, the issue of whether these methods deliver measurement, in the scientific sense, remains open. Critical scrutiny of psychological methods requires a theory of measurement matching the realist, cognitive aspirations of science. Such a theory evolved over the past two and a half millennia in tandem with developments in quantitative physical science, but has been influenced also by nonrealist philosophical fashions. The vicissitudes of this evolution constitutes the subject matter of the history and philosophy of measurement theory. The following sketch tracks some main lines of this evolution.

Measurement Theory and the Development of Quantitative Physics

Extensive Magnitudes and Real Numbers

The practice of measuring quantitative properties by counting units was a ubiquitous feature of all early human societies (Morley and Renfrew, 2010). Out of this practice, the problem of incommensurability became clear to ancient Greek scientists. Book V of Euclid's *Elements* (Heath, 1908), compiled around 300 BC, resolved this problem by redefining the fundamental metrological concept of ratio. The resulting understanding of measurement remains standard today.

The ancient Greeks distinguished discrete quantity (e.g., the size of a football crowd) from continuous quantity (e.g., the length of a stadium), measures in each case being assessed via numbers of appropriately defined units. Number, then, was what is now called 'whole number.' The problem of incommensurability is that two different magnitudes of a continuous quantity may be such that the size of one relative to that of the other does not equal a ratio of whole numbers (e.g., the lengths of the side and diagonal of a square). This seemed to imply that some magnitudes are not measurable relative to some units.

The solution to this problem lay in liberalizing the concept of ratio. First, two magnitudes are "said to have a ratio to one another which are capable, when multiplied, of exceeding one another" (Heath, 1908, p. 114). That is, any two lengths, say L_1 and L_2, where $L_1 \leq L_2$, have a ratio because there exists a whole number, n, such that $nL_1 > L_2$. Then, any two lengths (call them L_i and L_j) are said to be in the same ratio as another pair (L_h and L_k) if and only if for each pair of whole numbers, n and m,

$nL_i < mL_j$ if and only if $nL_h < mL_k$;
$nL_i = mL_j$ if and only if $nL_h = mL_k$; and
$nL_i > mL_j$ if and only if $nL_h > mL_k$.

This definition allows incommensurable pairs of lengths to stand in the same ratio, and while this ratio does not equal a ratio of two whole numbers, the ratio of any such incommensurable pair, L_i and L_j, is uniquely specified by the class of all whole-numbered ratios, n/m, where $nL_i \leq mL_j$. Of course, the ancients did not recognize such classes as identifying numbers, as Dedekind (1872) did when he later defined the concept of a real number, but Dedekind's definition was deeply indebted to Euclid (Stein, 1990). As far as measurement is concerned, when a magnitude is incommensurable with the unit employed, the ratio between them always falls between ratios of whole numbers and, so, can always be estimated to any level of accuracy. This liberalized concept of ratio generalizes to any continuous quantity admitting multiples of magnitudes, which the ancients took to include extensive quantities such as weight, time, and the geometric magnitudes. It implies the classical concept of measurement, viz., that measurement is the estimation of the ratio between a magnitude of a quantity and a unit of the same quantity.

Intensive Magnitudes and the Concept of Continuous Quantity

The ancients recognized that nonextensive attributes (e.g., temperature) admitted of different degrees and sometimes discussed these in terms of ratios (e.g., Aristotle (1941)). However, medieval philosophers like Duns Scotus (Cross, 1998) seem to have been the first to seriously consider the idea that some nonextensive attributes might actually possess the same kind of structure as extensive quantities and, so, be measurable in exactly the same sense.

Euclid's *Elements* (Definitions 3 and 4 of Book V) taught that the class of magnitudes constituting an extensive quantity are homogeneous in the sense of being all of the same kind, but had used the idea of extensive multiplication as the criterion of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.43062-X

homogeneity. Duns Scotus, thinking in terms of changes from 1° or intensity of a quality to another, understood homogeneity in relation to differences: the degrees constituting a quality are homogeneous when the differences between them are all of the same kind and do not differ qualitatively, and the attribute is quantitative if the degrees differ from one another only quantitatively. Such an attribute is an intensive quantity, and while its internal quantitative structure is not manifest via operations of extensive addition and so it is not measurable by procedures analogous to extensive quantities, it is nonetheless measurable, in principle, because, as Oresme put it, "whatever ratio is found to exist between intensity and intensity, … a similar ratio is found to exist between line and line and vice versa" (Clagget, 1968, p. 167).

While medieval philosophers successfully freed the concept of quantity from its ancient exclusive attachment to extensive attributes, they were never successful in finding ways to measure attributes that they thought of as intensive quantities. Attempts to expand the range of measurable attributes were more successful from the Scientific Revolution of the seventeenth century onward, when science became more closely identified with quantitative methods. Such attempts raised another problem: it was one thing to introduce the possibility of intensive magnitudes theoretically; it was quite another to have evidence showing that a putative intensive quantity really was quantitative.

Most of the attributes found to be measurable from the seventeenth century onward are not extensive and are measured only indirectly via other quantities. For example, velocity is measured via distance and time, and density via mass and volume. The magnitudes measured in this derived way were thought of as quantitative attributes of the objects or systems involved, but ones for which no physical operation directly reflecting the attributes' underlying additive structure had yet been identified. Because quantitative physics was thought to provide a model for aspiring sciences, other disciplines aped its quantitative character. For example, Hutcheson (1725) proposed a quantitative moral psychology, but was criticized by Reid for "applying measures to things that properly have not quantity" (1948, p. 717), a charge that was to become increasingly familiar in the nineteenth century as psychologists attempted to quantify mental attributes.

The problem of how to distinguish quantitative from nonquantitative attributes required for its solution an explicit specification of quantitative structure. Helmholtz (1887) made a first attempt, which in turn inspired Hölder (1901) to successfully define the concept of an unbounded, continuous quantitative attribute, that is, the kind of attribute presumed in quantitative physics. Letting Q stand for any attribute, $a, b, c, \ldots$, etc., for any magnitudes of Q, and allowing that the expression $a + b = c$ is true if and only if c is entirely composed of discrete parts a and b, then Q is an unbounded, continuous quantity if and only if

1. Given any two magnitudes, a and b of Q, one and only one of the following is true:
   (i) a is identical to b (i.e., $a = b$ and $b = a$);
   (ii) a is greater than b and b is less than a (i.e., $a > b$ and $b < a$); or
   (iii) b is greater than a and a is less than b (i.e., $b > a$ and $a < b$).
2. For every magnitude, a of Q, there exists a b in Q such that $b < a$.
3. For every pair of magnitudes, a and b in Q, there exists a magnitude, c in Q such that $a + b = c$.
4. For every pair of magnitudes, a and b in Q, $a + b > a$ and $a + b > b$.
5. For every pair of magnitudes, a and b in Q, if $a < b$, then there exist c and d in Q such that $a + c = b$ and $d + a = b$.
6. For every triple of magnitudes, a, b, and c in Q, $(a + b) + c = a + (b + c)$.
7. For every pair of classes, ϕ and ψ of magnitudes of Q, such that
   (i) each magnitude of Q belongs to one and only one of ϕ and ψ,
   (ii) neither ϕ nor ψ is empty, and
   (iii) every magnitude in ϕ is less than each magnitude in ψ,
   there exists a magnitude x in Q such that for every other magnitude x' in Q, if $x' < x$, then $x' \in \phi$ and if $x' > x$, then $x' \in \psi$ (depending on the particular case, x may belong to either class).

If an attribute has this kind of structure, the system of ratios of its magnitudes is isomorphic to the system of positive real numbers, which entails that each magnitude of the attribute may be measured by any other magnitude taken as the unit of measurement. Hence, Hölder's seven conditions adequately characterize the kind of measurable attributes found in the physical sciences and nicely rounds out the classical concept of measurement.

The issue could now be considered of the sort of evidence needed to confirm the hypothesis that some attribute is quantitative. If, relative to any attribute, it is possible to identify a concatenation operation directly reflecting the attribute's additive structure, then evidence of quantitative structure may be obtained. For example, rigid straight rods may be joined so that the resulting length equals the sum of the lengths concatenated. Campbell's (1920) term "fundamental measurement" is widely used to describe this way of establishing that an attribute is measurable.

Campbell also identified 'derived measurement.' For example, the ratio of mass to volume is a different constant for each different kind of substance, say, one constant for gold, another for silver, and so on. Taking such constants to identify magnitudes of a quantitative attribute is sound scientific reasoning. In the case of density, for example, the fact that such a correlation between constants and substances is always observed requires that each different substance possesses a different characteristic as its basis, and the fact that the constants are quantities means that this basis must likewise be quantitative. A similar line of thought applies to all cases of indirect quantification in physics.

Campbell (1920) thought that physics was, by definition, "the science of measurement" (p. 267) and concluded that evidence for quantitative structure was exhausted by his categories of fundamental and derived measurement. He did not know that Hölder (1901) had earlier given a counterexample to this view using a simple model. Consider the set of all intervals between points on a straight line. If all that is known of any two intervals were whether they span equal distances, would this be sufficient to prove that distance is a quantitative attribute?

Hölder proved yes, exploiting the fact that some intervals are completely nested within others and the distance spanned by any interval always equals the sum of the distances spanned by the discrete intervals nested within and constituting it. His result displays that there are ways over and above fundamental and derived measurement of showing attributes to be quantitative. Hence, the possibility of measuring nonphysical attributes (such as psychological attributes) cannot be ruled out a priori, as Campbell insisted.

Measurement of Psychological Attributes

There are no fundamentally measurable, extensive psychological attributes upon which to base the enterprise of psychological measurement. Thus, attempts to gain evidence that psychological attributes are quantitative cannot utilize Campbell's categories of fundamental and derived measurement. However, Hölder's counterexample illustrated the fact that alternatives are possible, and Luce and Tukey (1964) introduced one to the psychological literature, viz., conjoint measurement theory (see article Measurement Theory: Conjoint). This theory applies in circumstances where a first attribute is a noninteractive function of a second and third attribute, such that distinct levels of the second and third can be identified and classified as the same or different and levels of the first attribute at least weakly ordered. Since, in psychology, such classifications and orderings of levels of attributes are sometimes possible, this theory promised to resolve the issue of evidence regarding whether psychological attributes are quantitative. If the attributes involved are quantitative, then this may be discernable via trade-offs between the contributions of the second and third attributes to order relations between levels of the first. This theory also made explicit the need for such evidence in every case of quantification by revealing the fact that derived measurement in physics is never a matter of mere definition, as, for example, Carnap (1966) had suggested was the case for density measured via mass and volume, but always depends upon empirically testable propositions.

While conjoint measurement has been applied in psychology (see article Measurement Theory: Conjoint Analysis Applications), it is most often employed as a scaling device (on the assumption that the relevant attributes are quantitative), rather than as a diagnostic tool for testing this assumption. In the best case scenario, the theory of conjoint measurement would allow the conjecture that some attribute is quantitative to be demonstrated via construction of 'standard sequences' (Krantz et al., 1971). In this way, the character of the putative additive structure hypothesized within psychological attributes could be articulated. Otherwise, the quantitative structure presumed of such attributes remains an empty speculation.

From the time of the first serious attempts at psychological measurement (Fechner, 1860), reservations were expressed about whether such attributes possess additive structure or even satisfy the requirement of possessing the necessary kind of homogeneity (e.g., Tannery, 1875). As Tannery and others (e.g., von Kries, 1882) noted, if differences between degrees of an ordered attribute are of different kinds (i.e., not entirely homogeneous), then the attribute cannot be quantitative. This is because differences between its degrees are not naturally ordered and, so, cannot satisfy even Hölder's condition (1).

However, while this kind of objection was taken seriously in other disciplines (e.g., Keynes, 1921), in psychology it was never seriously considered and, instead, the 'psychometricians' fallacy' (Michell, 2009) was endorsed: viz., the conclusion that, if an attribute is ordered, then it must be quantitative (e.g., Titchener, 1905). Consequently, psychological 'measurement' (so-called) involves attributes that have never been shown to possess an additive structure and, in some cases, attributes that appear to lack the required kind of homogeneity (Michell, 2013).

Measurement Theory and Issues External to Quantitative Science

Philosophy of Mathematics

A potent external factor influencing measurement theory is the philosophy of mathematics. Hölder's conclusion of an isomorphism between ratios of magnitudes of a continuous quantitative attribute and the positive real numbers secured the view, popular at least since Newton (1728/1967) that the numbers are such ratios, a view articulated in detail by Frege (1903) and Whitehead (Whitehead and Russell, 1913). However, consistent with the modernist transformation that sought to separate the subject matter of mathematics from that of the empirical sciences (Gray, 2008), other philosophers attempted to divorce the concept of number from that of quantity (e.g., Russell, 1897). The view that dominated thinking in the twentieth century was that numbers are 'abstract entities,' that is, they exist only outside space and time (Bostock, 2009). As a result, measurement was no longer understood in the classical manner as the estimation of ratios of magnitudes because such ratios were no longer identified with real numbers. Instead, measurement was redefined as the representation of objects and relevant relations between them by numbers and numerical relations, the numbers themselves being understood as existing outside of the empirical context of measurement and only introduced as a tool (Russell, 1903; Campbell, 1920; Nagel, 1931; Ellis, 1966; Krantz et al., 1971; Suppes et al., 1989; Luce et al., 1990; Narens, 2002) (see article Measurement, Representational Theory of).

According to Luce and Narens (1994), the representational approach to the measurement of any attribute reduces to four steps:

First, an empirical structure, involving an observable set of objects and a finite number of empirical relations in which they stand, is identified.

Second, a set of axioms characterizing this empirical structure is specified. As far as possible these should be empirically testable.

Third, an abstract numerical structure is identified such that a set of many-to-one mappings (homomorphisms) between the empirical structure and this numerical structure can be proved from the axioms (the 'representation theorem').

Fourth, a specification of how the elements of this set of homomorphisms relate to one another is given, generally by identifying the class of mathematical functions that

includes all transformations of any one element of this set into all other elements (the 'uniqueness theorem'). This allows the distinctions between scale types, earlier popularized by Stevens (1951), to be precisely defined.

Expressed in this mathematically sophisticated version, the representational theory of measurement attempts to assimilate the concept of measurement to the theory of numerical representations. As has been pointed out, however (Mundy, 1994; Michell, 2007a), understood in representational terms, quantitative theories in physics would require something very like Hölder's (1901) concept of continuous quantity as the empirical structure to be numerically represented, which entails that something having the structure of the system of positive real numbers (i.e., the system of ratios of magnitudes) is already present in the empirical structure. The idea that such ratios of magnitudes are numbers is implicit within the classical theory of measurement (Michell, 1994). This kind of empirical realist view of number fits with the more general thesis that mathematics is the general science of structure (i.e., both empirically instantiated and empirically unrealized structures), and it avoids the difficulties encountered in attempting to construe numbers as 'abstract entities,' not the least of which is the issue of how, as such, they could ever be known by us. Recently, philosophies of mathematics that locate numbers in the empirical contexts of their application have reappeared (Franklin, 2009).

Philosophy of Science

The philosophy of science has also influenced measurement theory. In relation to the understanding of measurement in psychology, the most significant philosophical influence during the twentieth century was operationism (Bridgman, 1927). If, as Bridgman maintained, the meaning of any concept is the set of operations used to identify it, then measurement is the 'assignment of numerals to objects or events according to rules' as Stevens (1951, p. 1) defined it. Operationism was adopted because it seemed an *avant-garde* philosophy promising to bring psychology into line with physics, but ironically and despite its enduring popularity, this definition of measurement does not match that still standard in physics (Michell, 2007b). First, as measurement is understood in physics, it makes no sense to talk of measuring 'objects or events' unless it is attributes of such things that are being considered. Attributes (i.e., properties of things or relations between things) are always the object of measurement. Second, while 'numerals' (i.e., the names of or signs for numbers) are generally present in most acts of measurement, this is only because measurement involves reference to numbers as ratios of magnitudes. It is always numbers that are identified in measurement in the first instance and numerals, only secondarily. Third, in measurement, numerals are no more 'assigned' to anything than in, say, describing colors, color words are assigned to things. Just as a description via color words attempts to express a fact, so an act of measurement attempts to describe a quantitative fact, and making a description is never a matter of assigning one thing to another. Fourth, according to Stevens' definition, there is nothing that cannot in principle be 'measured' because numerals can always be assigned to anything according to some rule or other. However, according to the understanding current in physics, the only measurable attributes are those having the kind of structure that sustains ratios (i.e., quantitative structure). If the aim was to bring the understanding of measurement in psychology into line with that in physics, Stevens' proposed definition represents an intellectually self-defeating maneuver, which, because it encouraged the use of the term 'measurement' for practices not falling under that predicate, alienated psychology from the traditions of quantitative science.

Michell's (1999 and 2007b) works contain further material on the history and philosophy of measurement theory in psychology and social sciences.

See also: Measurement Theory: Conjoint Analysis Applications; Measurement Theory: Conjoint; Measurement, Representational Theory of.

Bibliography

Aristotle, 1941. On generation and corruption. In: McKeon, R. (Ed.), The Basic Works of Aristotle. Random House, New York.
Bostock, D., 2009. Philosophy of Mathematics: An Introduction. Wiley-Blackwell, Oxford.
Bridgman, P.W., 1927. The Logic of Modern Physics. Macmillan, New York.
Carnap, R., 1966. An Introduction to the Philosophy of Science. Dover, New York.
Campbell, N.R., 1920. Physics, the Elements. Cambridge University Press, Cambridge, UK.
Clagget, M. (Ed.), 1968. Nicole Oresme and the Medieval Geometry of Qualities and Motions. University of Wisconsin Press, Madison, WI.
Cross, R., 1998. The Physics of Duns Scotus: The Scientific Context of a Theological Vision. Clarendon Press, Oxford.
Dedekind, R., 1872. Stetigkeit Und Irrationale Zahlen. Vieweg, Braunschweig.
Ellis, B., 1966. Basic Concepts of Measurement. Cambridge University Press, Cambridge, UK.
Fechner, G.T., 1860. Elemente der Psychophysik. Breitkopf and Hartel, Leipzig.
Franklin, J., 2009. Aristotelian realism. In: Irvine, A. (Ed.), The Philosophy of Mathematics. Elsevier, Amsterdam.
Frege, G., 1903. Grundgesetze der Arithmetik. Georg Olms, Hildesheim.
Gray, J., 2008. Plato's Ghost: The Modernist Transformation of Mathematics. Princeton University Press, Princeton, NJ.
Heath, T.L., 1908. The Thirteen Books of Euclid's Elements, vol. 2. Cambridge University Press, Cambridge, UK.
Helmholtz, H von, 1887. Eduard Zeller zu senium füfzigjährigen Doktorjubiläum gewindmet. Fues' Verlag, Leipzig.
Hölder, O., 1901. Die Axiome der Quantität und die Lehre vom Mass, Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig. Mathematisch-Physische Klasse 53, 1–46.
Hutcheson, F., 1725. An Inquiry into the Original of our Ideas of Beauty and Virtue. Darby, London.
Keynes, J.M., 1921. A Treatise on Probability. Macmillan, London.
Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A., 1971. Foundations of Measurement, vol. 1. Academic Press, New York.
Luce, R.D., Krantz, D.H., Suppes, P., Tversky, A., 1990. Foundations of Measurement, vol. 3. Academic Press, San Diego, CA.
Luce, R.D., Narens, L., 1994. Fifteen problems concerning the representational theory of measurement. In: Humphreys, P. (Ed.), Patrick Suppes: Scientific Philosopher, vol. 2. Kluwer, Dordrecht.
Luce, R.D., Tukey, J.W., 1964. Simultaneous conjoint measurement: a new type of fundamental measurement. Journal of Mathematical Psychology 1, 1–27.
Michell, J., 1994. Numbers as quantitative relations and the traditional theory of measurement. British Journal for Philosophy of Science 45, 389–406.
Michell, J., 1999. Measurement in Psychology: Critical History of a Methodological Concept. Cambridge University Press, Cambridge, UK.
Michell, J., 2007a. Representational theory of measurement. In: Boumans, M. (Ed.), Measurement in Economics: A Handbook. Elsevier, Amsterdam.

Michell, J., 2007b. Measurement. In: Turner, S.P., Risjord, M.W. (Eds.), Philosophy of Anthropology and Sociology. Elsevier, Amsterdam.
Michell, J., 2009. The psychometricians' fallacy: too clever by half? British Journal of Mathematical and Statistical Psychology 62, 41–55.
Michell, J., 2013. Constructs, inferences, and mental measurement. New Ideas in Psychology 31, 13–21.
Morley, I., Renfrew, C., 2010. The Archaeology of Measurement: Comprehending Heaven, Earth and Time in Ancient Societies. Cambridge University Press, Cambridge, UK.
Mundy, B., 1994. Quantity, representation and geometry. In: Humphreys, P. (Ed.), Patrick Suppes: Scientific Philosopher, vol. 2. Kluwer, Dordrecht.
Nagel, E., 1931. Measurement. Erkenntnis 2, 313–333.
Narens, L., 2002. Theories of Meaningfulness. Lawrence Erlbaum, Mahwah, NJ.
Newton, 1728/1967. Universal arithmetic: or, a treatise of arithmetical composition and resolution. In: Whiteside, D.T. (Ed.), The Mathematical Works of Isaac Newton, vol. 2. Johnson Reprint Corp, New York.
Reid, T., 1748/1849. An essay on quantity. In: Hamilton, W. (Ed.), The Works of Thomas Reid. Maclachlan, Stewart & Co, Edinburgh, UK.
Russell, B., 1897. On the relations of number and quantity. Mind 6, 326–341.
Russell, B., 1903. Principles of Mathematics. Cambridge University Press, Cambridge, UK.
Stein, H., 1990. Eudoxos and Dedekind: on the ancient Greek theory of ratios and its relation to modern mathematics. Synthese 84, 163–211.
Stevens, S.S., 1951. Mathematics, measurement and psychophysics. In: Stevens, S.S. (Ed.), Handbook of Experimental Psychology. Wiley, New York.
Suppes, P., Krantz, D.H., Luce, R.D., Tversky, A., 1989. Foundations of Measurement, vol. 2. Academic Press, New York.
Tannery, J., 1875. A propos du logarithme des sensations. La Revue Scientifique 14, 876–877.
Titchener, E.B., 1905. Experimental Psychology: A Manual of Laboratory Practice. Macmillan, London.
von Kries, J., 1882. Über die Messung intensiver Grössen und über das sogenannte psychophysische Gesetz. Vierteljahrsschrift für wissenschaftliche Philosophie 6, 257–294.
Whitehead, A.N., Russell, B., 1913. Principia Mathematica. Cambridge University Press, Cambridge, UK.

Mechanism-Based Causal Analysis

Carly R Knight, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Over the past decades, 'causal mechanisms' have become an important part of social scientific explanation. Causal mechanisms now appear in annual review pieces, edited volumes, and occasional symposia. The publication of the *Oxford Handbook of Analytic Sociology* – a manifesto of the analytic paradigm that places mechanisms at the helm of its enterprise – is further evidence of this new focus. But what, precisely, are causal mechanisms? And how do they aid in explanation in the social sciences? This article investigates five major approaches to causal mechanisms toward the goal of identifying major points of consensus and contention. It further suggests that there are two distinct approaches to causal mechanisms: 'top-down' approaches that seek to generalize empirical events under widely instantiated causal patterns, and 'bottom-up' approaches that seek to disaggregate 'average causal effects' by opening up the 'black-boxes.'

Introduction: Mechanisms and Explanation in the Social Sciences

Over the past decades, 'causal mechanisms' have become an important part of social scientific explanation. Causal mechanisms now appear in annual review pieces, edited volumes, and occasional symposia. The publication of the *Oxford Handbook of Analytic Sociology* – a manifesto of the analytic paradigm that places mechanisms at the helm of its enterprise – is further evidence of this new focus. But what, precisely, are causal mechanisms?

There are plenty of answers to this question but no clear consensus. James Mahoney (2001), for instance, has cataloged some 24 variants. Doug McAdam, Sidney Tarrow, and Charles Tilly remark that "no conceptual, theoretical, or methodological consensus has so far emerged" (2008: 308). Similarly, Neil Gross notes that even though "a clear definition is an essential first step towards a sociological theory of mechanisms … confusion abounds as to what exactly a mechanism is" (2009: 359).

The rising appeal of mechanistic explanation on the one hand, and the intensifying disagreement over the nature of that enterprise on the other hand, has been a source of tension in the literature. Some have worried that this lack of coherence prevents theory building. Others have questioned the usefulness of this vague and contested explanatory tool. What is at stake in this debate, then, is whether causal mechanisms have the potential to be a foundation for a unified approach to social science, or whether debates over mechanisms are likely to remain unresolved.

This article of the literature argues that, to an extent, both conclusions are tenable. That is, underlying many of the competing approaches to causal mechanisms lies a core conceptual consensus upon which a unified approach to social science is being built. This consensus revolves around the notion of mechanisms as 'generalized causal chains that produce an effect.' Debates over mechanisms can thus be understood, as John Gerring (2007) argues, not as debates over the nature of mechanisms but rather as debates over which subtypes of mechanisms are important to social science. While competing research agendas drive scholars to privilege different mechanism subtypes, many (though, importantly, not all) of these competing accounts are largely compatible with the minimal concept of a mechanism provided above. This is a cause for optimism. To the extent that debates over mechanisms can be recast as elaborations of subtypes rather than as irreconcilable views on the very nature of mechanisms, mechanisms can be a basis for a unified social-scientific logic of inquiry.

However, there are at least two reasons that mechanism debates are unlikely to be wholly resolved anytime soon. The first is salutary. A plurality of mechanism accounts may simply be the product of the success of the concept. Because mechanisms are so broadly applicable, further specification is necessary, even for those accounts that have a unified conceptual core. That is, researchers must go beyond considering the nature of mechanisms and actually locate mechanisms in the empirical world. Mechanism subtypes give guidance; for instance, rational choice mechanisms provide a more analytically powerful theoretical lens than 'mechanisms' as a general concept. That these subtypes simultaneously provide theoretical guidance and generate theoretical tension means that we should neither expect nor desire a single mechanisms approach.

The second reason for the continuing debate over mechanisms reflects a deeper disagreement, however, and therefore raises the possibility that there are two irreconcilable approaches to mechanistic explanation. This is the case because not all approaches can be considered subtypes of a conceptual core that views mechanisms as generalized causal chains that produce effects. The source of disagreement is the requirement of 'generalized' explanation: While one notion of mechanisms relies upon a theory of explanation based on generalization or unification (i.e., mechanisms explain by subsuming a phenomenon under some widely instantiated, robust causal pattern), a second notion relies upon a view of explanation more akin to opening a black-box (i.e., mechanisms explain by identifying contextually specific chains of causation that may be neither robust nor generalizable). Theorists in the second camp would not require that mechanisms be generalizable. In fact, they would laud the explanatory content of the specific. By contrast, theorists in the first camp would hold that removing 'generalizable' from the concept violates the core premise of mechanistic explanation, and perhaps of explanation more generally. Whether differences between these camps implies

that there exist two distinct notions of mechanisms remains an open question, but it is one worth exploring as the literature on mechanisms progresses.

This article provides an extended review of the literature on mechanisms and an assessment of its main points of contention. This article proceeds as follows: Part II presents a review of five major theoretical accounts of causal mechanisms; Part III assesses the principal areas of contention; and Part IV concludes.

The Many Meanings of Mechanisms

Before delving into the various approaches to mechanisms, it will be helpful to ask whether it is worth searching for conceptual consensus. How big a problem is it to have different accounts of mechanisms? Might there not be a benefit to using ambiguous concepts? After all, it is perhaps no more realistic or desirable to come up with a unified definition of a causal mechanism than it is to come up with a unified definition of democracy or culture. Moreover, as stated above, there are good reasons for disharmony. Mechanisms operate differently in different domains: Because biological, historical, and psychological mechanisms refer to different content, we should not expect to come up with a universally satisfying conceptual definition (Kuorikoski, 2009). Distinct definitions allow for the mutual recognition of our similar projects and the cross-fertilization of ideas, without the impossible requirement that theory building come after a clear definition has been settled upon. Second, the fact that there are many approaches to mechanisms may overstate the degree of confusion in the field. While the debate over mechanisms may evince deep disagreement, if put upon to provide a mechanism for a given relationship, most social scientists will understand what they are being asked to do.

However, the fact that there are good reasons for theoretical proliferation does not mean that the literature on mechanisms should stop exploring whether various approaches rest upon a shared conceptual foundation. Quite the opposite. Over the last few decades, debates over mechanisms have served as one of the important arenas in which social science has considered, stipulated, and debated its goals. Clarity on these goals is essential to justifying the kinds of endeavors social science is engaged in and for articulating – to ourselves as much as to outsiders – why the sort of information it provides is important. Calls for clarity on mechanisms are in essence calls for clarity on social science's understanding of its explanatory endeavors and thus are crucial to the development of a coherent logic of inquiry. For this, we need not be worried that there are different subtypes of mechanisms, but we certainly should be concerned if it turns out that mechanisms evince a deeper form of conceptual plurality in which the single term masks distinct approaches to explanation. Below, I explore five major accounts of mechanisms to analyze the extent to which mechanisms evince theoretical plurality in their approaches to explanation, as well as the extent to which these different approaches are engaged in a single, coherent endeavor.

What Mechanisms Are Not: The Hempelian View

Many accounts of mechanisms see themselves as responding to a view of the social sciences as engaged in the endeavor of discovering 'social laws.' Articulated by Carl Hempel, this view, known as the covering-law model of explanation, argued that the explanation of an event consisted in developing its description into a series of premises, at least one of which contained a lawlike statement (Hempel and Oppenheim, 1948). For example, I can explain why my door is difficult to open in the summer by deriving the following premises: (1) my door is made of wood; (2) wood expands in heat; and (3) summer is hot. Hempel contended that explanations of human behavior were fundamentally similar to these sorts of physical phenomena. Logical positivists (with whom Hempel is typically associated, though he rejected that term) thus root explanation in the identification of constant conjunctions of events. They reject the notion that knowledge could come from anything other than direct sensory experience and consequently take the antirealist stance of rejecting the existence of theoretical or unobservable entities.

Hempel's covering law approach proved influential but critics soon rejected it. They argued that explanation by 'constant conjunction' provides 'no understanding' (Bunge, 1997), a criticism perhaps even more damning than the charge that social laws, as opposed to physical laws, did not, in fact, exist (Hedström and Swedberg, 1998). Most importantly, Hempel's antirealist stance was anathema to social scientists who considered their object of study, society, to be both a real and a theoretical entity (Steinmetz, 1998).

In short, the 'covering law' approach to explanation has now been largely rejected by social theorists. Nevertheless, most contemporary accounts of mechanisms have this view in mind when articulating the explanatory advantages of the mechanistic approach.

Critical Realism

The critical realist approach to social science, associated with Roy Bhaskar, identifies the very purpose of social science as the search for causal mechanisms. Rejecting the antirealism of both logical positivism and postmodernism, Bhaskar motivates the existence of causal mechanisms by asking the question: 'What must be true about the world in order for science to be intelligible?' (Bhaskar, 1978; Steinmetz, 1998). Bhaskar's answer is that the scientific method indicates the existence of real causal mechanisms. First, empirical generalities in the real world are a rarity. Instead, science has to engineer 'constant conjunctions' in controlled, laboratory experiments ('closed systems'). Because constant conjunctions are artificial productions, they should not, as positivism argues, form the basis of explanation. However, even though constant conjunctions must be engineered, the relationships identified within closed systems successfully predict events outside the laboratory (within 'open systems'). This forces us to also reject postmodernism's claim that the objects of inquiry in science are mere theoretical creations, without real referents. In short, in order for science to be intelligible, there must exist real causal forces that are ontologically distinct from the events that they generate.

Bhaskar (1998[1979]) distinguishes mechanisms from the empirical world by differentiating the world into three hierarchically arranged levels: the 'real,' the 'actual,' and 'events.' Mechanisms refer to a reality separable from both empirical

actuality and perception. They could therefore be at work in a particular context even if they don't produce anything we can observe in the realm of the 'actual.' To see how this is the case, consider the psychological mechanism of an anger management problem: Other mechanisms, such as conformism to societal norms or social ties may mitigate the manifestation of the anger management problem, but this does not negate the presence of the 'real' psychological process underneath.

Bhaskar's causal mechanisms are general and abstract, in the sense that they refer to processes that will be at work in both closed and open systems, and thus exist apart from the context of either the experiment or the empirical world. Finally, for Bhaskar, social scientific explanations would necessarily be different from scientific explanations due to social science's unique objects of inquiry. That is, social structures exist in a meaningful world of concepts and must necessarily engage in hermeneutics to make action intelligible.

Critical realism proved a powerful rejoinder to positivism. Its influence remains evident in nearly all contemporary realist approaches to mechanisms: indeed, while discussions over mechanisms do not always invoke the critical realist approach, the critical realist critique of logical deductivism is often an implicit and essential part of their formulations.

Jon Elster: Mechanisms as 'Cogs and Wheels'

Like Bhaskar, Jon Elster is intent on demonstrating the limitations of the covering law approach. (Elster's views have shifted considerably over the years, and here I focus on his later thought (Elster, 1998, 2007).) Elster argues that general laws are unsatisfactory not because this approach is inappropriate to social reality in any deep, metaphysical way, but rather because uncertainty permeates the social world, undermining the discovery of laws. While ideally we would form laws of the kind 'if p, then q,' uncertainty in initial conditions prevents us from stipulating what 'p' will entail. Elster's concern is that, if we accepted Hempel's account, the lack of identifiable universal laws would result in a purely descriptive social science. Mechanisms provide a solution to this problem by identifying intermediary-level conjectural propositions that explain why 'if p, sometimes q.' As Elster states: 'The "plea for mechanisms" is not an argument against lawlike explanations, only against the idea that when such explanations fail – which they usually do, we must fall back on narrative and description' (49).

On Elster's account, if sufficient evidence could be brought to bear, such that the uncertainty associated with mechanisms could be resolved, mechanisms would, in fact, become laws (Elster, 1989; Persson, 2012). Elster's position thus characterizes mechanisms as an epistemic tool (a feature of our state of knowledge), rather than as a real thing that exists in the world. It therefore constitutes a rupture with realist approaches. Furthermore, his contention that mechanisms be 'easily recognizable,' 'brute' empirical patterns violates the realist position that mechanisms must be conceptually separate from the empirical world. Some theorists associate Elster's focus on empirical generalities and the preference for lawlike explanation with a brand of logical positivism (Norkus, 2005). However, Elster's discussion and examples of mechanisms typically do not feature explanation by deduction nor do they always manifest in empirical regularities (Elster, 1998).

Few mechanism accounts draw upon these more controversial aspects of Elster's definition – the preference for lawlike generalities and the antirealist language (Little, 2013). Instead, Elster's account has primarily been influential due to its intuitive definition ('cogs and wheels') and the incisive delineation of the benefits of mechanisms. However, another aspect of Elster's advantages has been both influential and controversial: the fact that he firmly rooted mechanisms in the reductive strategy of methodological individualism.

As a methodological individualist, Elster holds that: (1) social entities do not exist independently of individuals and (2) social science explanations must cite the individual-level behaviors that give rise to some phenomena (the 'micro-foundations'). Some scholars have argued that it is a mistake to root a definition of mechanisms in methodological individualism (Tilly, 2001), while others have adopted aspects of this brand of methodological individualism (Hedström and Bearman, 2009). However, the rationale for requiring that the conceptual apparatus of mechanisms should reference individual action remains unclear. Most likely, Elster's commitment to methodological individualism has shaped his account of mechanisms, rather than vice versa. If one accepts that explanation must refer to micro-foundations, then so must explanations via causal mechanisms.

Robert Merton and Analytic Sociology

This section combines a discussion of Robert Merton's explication of mechanisms with Hedström and Bearman's discussion of Analytic Sociology, since the latter draws heavily on the former. Though Merton did not develop a full theory of mechanisms, he provided an early and influential definition. Merton (1968) defined social mechanism as 'social processes having designated consequences for designated parts of the social structure.' Social mechanisms and middle-range theories were Merton's antidote for both grand theory and empiricism. Merton anticipated that a more fruitful sociological approach would engage with limited phenomena as these provided an easier route to theoretical advancement. While Merton's critique against grand theory has been powerful and his ideal examples of social mechanisms influential, his account has been criticized as a vague foundation for a positive approach to middle-range theory building (Hedström and Bearman, 2009).

Analytic Sociology begins where Merton left off by further explicating social mechanisms and middle-range theories and placing these concepts at the core of sociological explanation (Hedström and Bearman, 2009). Drawing upon Machamer et al. (2000), they define mechanisms "as constellation of entities and activities that are organized such that they regularly bring about a particular type of outcome" (5). Furthermore, they argue that statistical association is insufficient for explanation: statistical techniques should provide the starting point but not the full content of explanation. For Hedström and Bearman, mechanisms must satisfy two requirements. Firstly, they should be 'semigeneral' in that they should 'make no reference to time, place, identity of actors, content of beliefs, or types of actions' (7). The purpose of abstracting away from

concrete context is that the mechanism should be applicable across a wide range of phenomena. Secondly, mechanisms should refer to the microlevel processes of individual action. For Analytic Sociology, Elster's brand of methodological individualism is softened into 'structural individualism' which gives a nod to the importance of social structure, but maintains that an inextricable link exists between explanation and the elaboration of micro-foundations.

There has been some confusion as to whether Analytic Sociology refers to a type of sociological research or whether it is meant to be a stronger argument for what proper explanation entails (with the implication that explanation not following this approach would be inadequate) (Santoro, 2012). To the extent that it is the latter, there has been pushback to Analytic Sociology's requirement that mechanisms should reference individual-level behavior (Tilly, 2001; Keith Sawyer, 2004).

Finally, further clarification could also be provided as to what 'semigeneral' mechanisms entail. Does a call for semigeneral mechanisms that abstract from place and time imply that mechanisms are, like Hempelian laws, universal? If many of the concepts that govern social life are in flux, then (1) is it indeed more explanatory for mechanisms to be stipulated at this level of abstraction?; and (2) how would a sensitivity to hermeneutics work within this approach? As Analytic Sociology continues to identify its objectives, these ambiguities are likely to require further elaboration.

Mechanisms in Quantitative Methodology

Calls for mechanistic analysis among quantitative researchers hardly represent a single, monolithic view. Nevertheless, a few commonalities can be drawn from some of the more prominent discussions of the estimation of causal mechanisms. Orienting itself against a 'black-box' statistical approach whose goal is the estimation of 'average effects,' this literature argues for the appeal of mechanisms on theoretical as well as epistemological grounds. Average causal effects in and of themselves may not be theoretically interesting parameters (Heckman, 2005). For instance, in the event of heterogenous treatment effects (in which an effect differs by subpopulation), the average causal effect offers only a weighted average, obscuring underlying variation. Morgan and Winship (2007) motivate the need for mechanisms by arguing that "social scientists desire explanations for outcomes that can be modified in straightforward ways when populations shift or contexts change in ways that are separable from the fundamental conditional average causal effects" (224).

There are two points here that differ from the emphasis often provided by theoretical mechanistic accounts. First, it is said that mechanisms, as well as causes, depend upon their social situation; thus, it may be more explanatory to have fine-grained detail about the mechanism's dependence upon context. Second, a key benefit of mechanistic analysis is prediction: knowing how a treatment works will provide information that is transferrable to new contexts.

A separate argument for causal mechanisms concerns their epistemic value. Estimating causal mechanisms may provide a better test of social scientific theories than does the estimation of average causal effects (Morgan and Winship, 2007; Imai et al., 2011). In these cases, the test for mechanisms consists in identifying a variable that represents the mechanistic pathway and in disaggregating the causal effect into direct and indirect pathways (VanderWeele, 2009). Furthermore, estimating mechanisms can provide information that is important for intervention, as mechanisms identify the generative processes that link cause and effect. Barbara Reskin makes this point explicitly in her call for further attention to the mechanisms that give rise to ascriptive inequality (2003). Through mechanisms 'we stand to gain the opportunity to meaningfully contribute to social policy' (17). Finally, mechanisms are useful because they bring additional support to ensure that a statistical relationship is indeed causal (Pearl, 2000). As Pearl demonstrates with the 'Front Door Criterion,' total causal effects can be estimated by stipulating the intermediate variables that link cause and effect.

Overall, the quantitative literature is more concerned with estimating mechanisms than with stipulating their nature. Consequently, it tends to have a more permissive account of mechanisms, identifying them as links within a causal chain, without commitments to generalization, methodological individualism, or a particular theory of action.

Charles Tilly, Mechanisms, and Processes

Charles Tilly (2001), too, promotes mechanisms as an alternative to covering laws. However, Tilly's approach differs markedly from that of Elster and Analytic Sociology in that it does not require that mechanisms refer to individual action. Rather, Tilly argues that individual level 'cognitive mechanisms' are only one of three types of mechanisms (the others being environmental and relational mechanisms).

Tilly identifies the purpose of explanation as unraveling 'episodes,' such as revolution or war, into the mechanisms and processes that generate them. Mechanisms are differentiated from 'processes' in that the former are defined as a 'delimited class of events that change relations among specified sets of elements in identical or closely similar ways over a variety of situations' (25). Processes are essentially concatenations of mechanisms.

Tilly's account provides a view on mechanisms that captures many common aspects of major approaches without the commitment to methodological individualism and with room for relational approaches to social action. Furthermore, his typology of relational, environmental, and cognitive mechanisms is a helpful tool for assessing the sorts of mechanisms that would be present in a full mechanism-account (though Tilly does not argue that mechanism schema must refer to all three types).

However, Tilly's differentiation between mechanisms and processes raises additional questions: if processes are composed of mechanisms, then must mechanisms be 'rock bottom' theoretical entities in that they themselves cannot be further unraveled into submechanisms? A more general point widely applicable to the mechanisms literature concerns the required level of abstraction, such that mechanisms are general enough to be robust across time and place but specific enough to explain the particular 'episode' in question.

Peripheral Debates and the Conceptual Core

These accounts are only a small slice of the literature on mechanisms, which itself is, by all indications, a growth industry. Lest a reader get caught in the weeds, it is helpful to identify some major points of debate. Below, I briefly review three sources of disagreement among the approaches discussed above: micro-foundations, objects of inquiry, and observability.

Micro-Foundations

Many prominent mechanism accounts, including those of Elster and analytic sociologists, posit that mechanisms must be rooted in methodological individualism. In most instantiations of methodological individualism, the fact that social structures can be made explicit in terms of individuals and their relations requires that any acceptable explanation must reference this level of action. As Hedström and Berman explain 'actions are important because all the things that interest us as sociologists are the intended or unintended outcomes of individuals' actions' (9).

However, Tilly (2001) and others have argued that there is little reason for this requirement: 'That intellectual strategy has the advantage of remaining close to the main line of political science explanations and the disadvantage of ignoring a wide range of significant cause–effect connections' (25). For instance, Little (1992) argues that causal–structural explanations, such as the alliance structures among European powers prior to World War II, are perfectly valid without referencing individual-level action. In other words, the position of ontological individualism, the metaphysical position that social facts are composed of individualistic facts, does not entail explanatory individualism, the position that causal explanations must reference those individualistic facts. Moreover, it is often the case that lower-level compositional explanations will provide less satisfying explanations than macro-level ones, since lower-level mechanisms can bring in causally irrelevant material that may be too heterogenous to be explanatory (Jepperson and Meyer, 2011). For example, recognizing that all physical structures are composed of individual molecules does not necessitate that all chemical explanations reference those molecules.

Individual-level mechanisms may certainly be preferred for disciplinary reasons, or may be a part of the research agenda of Analytic Sociology, but these are arguments about the subtypes of mechanisms that we want to identify, rather than arguments about the nature of the mechanisms themselves. As Stinchcombe (1991) argues, levels of analysis should therefore be an empirical, not a doctrinal, issue.

Objects of Inquiry

Some writers have sought to identify mechanisms with a particular theory of action. Neil Gross (2009), for example, views mechanisms as being 'composed of chains or aggregations of actors confronting problem situations and mobilizing more or less habitual responses' (368). Gross' definition draws on his pragmatist orientation that argues that cultural systems and meanings are (1) primary to action, and (2) not reducible to beliefs. This is important for Gross in that Merton and those who draw upon him, do not always include belief systems in their account of action. For instance, Merton's mechanism of 'self-fulfilling prophecies' does not include reference to what it means for actors to hold a particular belief, or the necessary conditions for them to act upon that belief. Other theorists take a very different approach. Elster (2007) identifies mechanism with rational choice. Andrew Abbott (2007), meanwhile, argues that relations, as opposed to individuals, are the important sociological objects of inquiry. (Abbott rejects the mechanistic approach due to its individualist orientation but argues for explanation largely in line with a more permissive account of mechanisms.)

However, it is possible to argue for a certain theoretical approach to mechanisms while not requiring that mechanisms be by their nature rooted in that approach. The explanatory appeal of identifying mechanisms of a certain subtype – pragmatist, rational choice, or relational – is not compromised by recognizing that the nature of mechanisms need not consist in these features.

Observability

Finally, many theorists argue that mechanisms should be considered unobservable theoretical entities. As Gerring (2007) notes, "for many writers, the hallmark of causal mechanisms is their unobservability and hence their hypothetical quality" (166). For example, Mahoney (2001) defines mechanisms as 'posited relations or processes that the researcher imagines to exist; they do not refer to any particular set of empirical conditions' (581); while George and Bennett, quoted in Gerring (2007), argue that 'no matter how far we push the border between the observable and the unobservable, some irreducibly unobservable aspect of causal mechanisms remains' (143). For many, this position is related to the causal realist notion that mechanisms must be differentiated from the causal regularities they generate and cannot be reduced to them.

Still other researchers argue that mechanisms should not be characterized as unobservable; rather, they are simply 'difficult to observe' (McAdam et al., 2008: 308). In fact, for many empirical researchers, the fact that mechanisms produce empirical consequences makes it easier to identify causal relationships by providing evidence of the links between events (Elster, 2007; Pearl, 2000).

Gerring (2007) characterizes this as a debate over positivism, or specifically, the extent to which models should be realistic. There are certainly positivist/empiricist elements to the debate (i.e., Elster's stipulation that mechanisms are easily observable), and yet causal mechanisms tend to presuppose a realist view that sharply contrasts with strict positivism. Rather than entailing some fundamental disagreement over positivism, debates over observability may be reconcilable to the extent that observablists accept that mechanisms are not reducible to empirical variables and unobservablists accept that mechanisms entail empirical consequences that, at least theoretically, could be identified.

A Conceptual Core?

These differences do not necessarily threaten a coherent mechanisms account. Rather, they may indicate different

subtypes of a concept on which there is a basic consensus. In fact, despite their differences, the accounts we have considered largely agree on the following points: Firstly, mechanisms provide an important alternative to covering laws; secondly, mechanisms provide justification for a realistic approach to the social world; and thirdly, mechanisms provide an intuitive description of causal relationships that focus on processes.

In defining this consensus, Gerring (2007) proposes that various approaches to mechanisms rest upon the core concept of a mechanism as 'the pathway or process by which an effect is produced or a purpose is accomplished' (178). Gerring notes that elaborations on this core are peripheral, properly regarded as 'arguments about mechanisms rather than the definition of mechanism' (178). When mechanisms are defined simply as chains of causation, many debates over mechanisms can constructively be reconceived as debates over mechanism subtypes.

However, while mechanisms can productively be thought of as simple causal pathways, there is a worry that this definition masks two fundamentally different views – it is questionable whether theorists would agree that just any process is a causal mechanism.

The first of these views about mechanisms is based on an understanding of explanation that depends upon generalization or unification. Mechanisms are said to be explanatory insofar as they organize knowledge, either into middle-range theories (Merton, 1968) or universal processes (Elster, 1998) that are widely instantiated. These views thus draw on a top-down approach to explanation, whereby we explain an event by showing that it is an instance of a general pattern. On this view, explanation by causal mechanisms requires that mechanisms are robust processes. Philip Kitcher (1989) presents the unificationist view:

> The heart of the unification approach is that we cannot make sense of the notion of a basic mechanism apart from a systemization of the world in which as many consequences as possible are traced to the prediction of as small a number of basic mechanisms as possible.

While many mechanistic approaches in the social sciences are less specific about the requirement of unification, a commitment to generalization remains. A single historical process, for example, would not count as a causal mechanism with any sort of explanatory value.

The second view about mechanisms is fundamentally different. It corresponds to a 'bottom-up' approach in which explanation is said to consist in identifying the chains of causal relations, complete with details of their context, which link a cause and an effect. Glennan (2002) refers to this as 'actual sequence explanation,' in which explanation consists in listing the contextual or historical particulars of a causal chain. Empirical researchers who orient themselves against black-box statistics typically espouse this view: whereas black-box approaches view causality as a relation at a distance of counterfactual dependence, mechanisms, on the other hand, view causality as a productive process where the links between cause and effect are identified. This 'bottom-up' view of mechanisms is consistent with theories that define mechanisms as causal chains, without the requirement that they refer to generalizable processes (Woodward, 2002; Bunge, 2004).

The epistemic goals of the 'bottom-up' or 'actual-sequence' view are often distinct from that of the 'top-down' view. For the 'bottom-up' view, part of the epistemic value of mechanisms consists in identifying information that is helpful (at least theoretically) for intervention or prediction. For example, one might want to know the mechanism that links a specific job training program to employment outcomes in order to predict the program's effect in new contexts, or to identify whether there are other efficacious ways to bring about the desired outcome. By 'opening the black-box,' mechanisms provide causal information that is better equipped than average causal effects in identifying the social levers that allow for change (Reskin, 2005; Morgan and Winship, 2007). In sum, 'bottom-up' explains by providing detail, while 'top-down' explains by achieving abstraction.

These two approaches are not necessarily mutually exclusive: after all, digging into a causal relationship may identify a mechanism that is an instance of a generalizable causal pattern. Despite this, it seems that the kind of explanation that undergirds each approach is distinct. That is, the different emphasis on disaggregation as opposed to aggregation may indicate that there is no conceptual core that can intelligibly serve as a foundation for both approaches. It is out of the scope of this article to provide a definitive answer; yet I hope that identifying this possible crux serves as a small step toward clarification.

Conclusion

This article has discussed five major approaches to causal mechanisms. I have argued that many seemingly divergent views rest upon a core conceptual consensus, one that orients social-scientific explanation away from constant conjunction and toward process identification. At the same time, the literature on mechanisms does contain points of deep contention. This article has raised the question of whether there are two distinct approaches to causal mechanisms, each with its own explanatory agenda. While it was not in the space of this article to fully investigate this question, identifying potential points of consensus and conflict is an important and necessary step toward conceptual clarity.

See also: Causal Counterfactuals in Social Science Research; Causal Inference: Overview; Causation: Physical, Mental, and Social; Causes and Laws: Philosophical Aspects; Explanation: Conceptions in the Social Sciences; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Scientific Explanation.

Bibliography

Abbott, Andrew, 2007. Mechanisms and relations. Sociologica 2, 1–22.
Bhaskar, Roy, 1978. On the possibility of social scientific knowledge and the limits of naturalism. Journal for the Theory of Social Behaviour 8 (1), 1–28.
Bhaskar, RoyA., 1998[1979]. The Possibility of Naturalism. Routledge, London.
Bunge, Mario, 1997. Mechanism and explanation. Philosophy of the Social Sciences 27 (4), 410–465.

Bunge, Mario, 2004. How does it work?: the search for explanatory mechanisms. Philosophy of the Social Sciences 34 (2), 182–210.
Elster, Jon, 1989. Nuts and Bolts for the Social Sciences. Cambridge University Press, Cambridge.
Elster, Jon, 1998. A plea for mechanisms. In: Hedström, Peter, Swedberg, Richard (Eds.), Social Mechanisms: An Analytical Approach to Social Theory. Cambridge University Press, Cambridge.
Elster, Jon, 2007. Explaining Social Behavior. Cambridge University Press, Cambridge.
Gerring, John, 2007. The mechanismic worldview: thinking inside the box. British Journal of Political Science 38 (01).
Glennan, Stuart, 2002. Rethinking mechanistic explanation. Philosophy of Science 69 (S3), S342–S353.
Gross, Neil, 2009. A pragmatist theory of social mechanisms. American Sociological Review 74 (3), 358–379.
Heckman, James, 2005. The scientific model of causality. Sociological Methodology 35, 1–97.
Hedström, Peter, Bearman, Peter, 2009. The Oxford Handbook of Analytic Sociology. Oxford University Press, Oxford.
Hedström, Peter, Swedberg, Richard (Eds.), 1998. Social Mechanisms: An Analytical Approach to Social Theory. Cambridge University Press, Cambridge.
Hempel, Carl, Oppenheim, Paul, 1948. Studies in the logic of explanation. Philosophy of Science 15 (2), 135–175.
Imai, Kosuke, Keele, Luke, Tingley, Dustin, Yamamoto, Teppei, 2011. Unpacking the black box of causality: learning about causal mechanisms from experimental and observational studies. American Political Science Review 105 (04), 765–789.
Jepperson, Ronald, Meyer, John W., 2011. Multiple levels of analysis and the limitations of methodological individualisms. Sociological Theory 29 (1), 54–73.
Kitcher, Philip, 1989. Explanatory unification and the causal structure of the world. In: Kitcher, Salmon (Eds.), Scientific Explanation: Minnesota Studies in the Philosophy of Science, vol. 13. University of Minnesota Press, Minneapolis, pp. 410–505.
Kuorikoski, Jaakko, 2009. Two concepts of mechanism: componential causal system and abstract form of interaction. International Studies in the Philosophy of Science 23, 143–160.
Little, Daniel, 1992. Jon Elster. In: Samuels, Warren (Ed.), New Horizons in Economic Thought: Appraisals of Leading Economists. Edward Elgar Publishing.
Little, Daniel, 2012. Analytical sociology and the rest of sociology. Sociologica 6 (1).
Little, Daniel, 2013. Social mechanisms and scientific realism. Social Epistemology and Reply Collective, 1–5.
Machamer, Peter, Darden, Lindley, Craver, Carl, 2000. Thinking about mechanisms. Philosophy of Science, 1–25.
Mahoney, James, 2001. Review essay: beyond correlational analysis: recent innovations in theory and method. Sociological Forum 16 (3), 575–593.
McAdam, Doug, Tarrow, Sidney, Tilly, Charles, 2008. Methods for measuring mechanisms of contention. Qualitative Sociology 31 (4), 307–331.
Merton, Robert, 1968. Social Theory and Social Structure. Simon and Schuster.
Morgan, Stephen L., Winship, Christopher, 2007. Counterfactuals and Causal Inference: Methods and Principles for Social Research. Cambridge University Press, New York.
Norkus, Zenonas, 2005. Mechanisms as miracle makers? The rise and inconsistencies of the 'mechanismic approach' in social science and history. History and Theory.
Pearl, Judea, 2000. Causality: Models, Reasoning and Inference. Cambridge University Press, New York.
Persson, Johannes, 2012. Mechanistic explanation in social contexts: Elster and the problem of local scientific growth. Social Epistemology 26 (1), 105–114.
Reskin, Barbara, 2005. Including mechanisms in our models of ascriptive inequality. In: Handbook of Employment Discrimination Research.
Santoro, M., 2012. The whole and the parts. or: is analytical sociology analytical enough about sociology, and itself? Sociologica 6 (1).
Keith Sawyer, R., 2004. The mechanisms of emergence. Philosophy of the Social Sciences 34 (2), 260–282.
Steinmetz, George, 1998. Critical realism and historical sociology. A review article. Comparative Studies in Society and History 40 (01), 170–186.
Stinchcombe, Arthur L., 1991. The conditions of fruitfulness of theorizing about mechanisms in social science. Philosophy of the Social Sciences 21 (3), 367–388.
Tilly, Charles, 2001. Mechanisms in political processes. Annual Review of Political Science 4 (1), 21–41.
VanderWeele, Tyler J., 2009. Mediation and mechanism. European Journal of Epidemiology 24 (5), 217–224.
Woodward, Jim, 2002. What is a mechanism? A counterfactual account. Philosophy of Science.

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF
JAMES D. WRIGHT
University of Central Florida, Orlando, FL, USA

VOLUME 15

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser–Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 15

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 15 *xlv*

VOLUME 15

Media and Child Development 1
Jessica T Piotrowski, Helen GM Vossen, and Patti M Valkenburg

Media and History 11
Brendan Dooley

Media and Social Movements 19
Kimberley Voss

Media, Crime and Crime Waves 23
Kevin Buckler

Media Effects 29
Don W Stacks, Zongchao Cathy Li, and Cylor Spaulding

Media Effects on Children 35
Marie-Louise Mares and Valerie Kretz

Media Ethics 42
Adam Drozdek

Media Events 48
Eric W Rothenbuhler

Media Imperialism 55
Annabelle Sreberny

Media Talk Shows 60
Andrew Tolson

Mediating Variable 64
David P MacKinnon

Mediation, Arbitration, and Alternative Dispute Resolution (ADR) 70
Carrie J Menkel-Meadow

Mediation, Statistical 75
Xiaolu Wang

Medical Education 81
Fred CJ Stevens and Cees PM van der Vleuten

Medical Geography 86
Jonathan D Mayer and Alex Wagner

Medical Profession 92
David Coburn

Medical Sociology 98
David Mechanic

Medicalization: Sociological and Anthropological Perspectives 105
Peter Conrad and Meredith Bergey

Medicine and Health for Sexual Minorities 110
Judith A Houck

Medicine, History of Western 118
Alfons Labisch

Meiji Revolution, The 126
Masato Miyachi

Meltzoff's "Like Me" Theory 131
Daina Crafa

Memmi, Albert (1920–) 135
Judith Roumani

Memory Binding 140
Richard J Allen

Memory Development During Infancy and Early Childhood across Cultures 147
Harlene Hayne, Kana Imuta, and Damian Scarf

Memory in the Bee 155
Randolf Menzel

Memory Models: Quantitative 160
Bennet Murdock

Men's Health 166
Richard G Owens

Mental Causation 172
Jens Harbecke

Mental Disorders, Epidemiology of 179
Katja Beesdo-Baum and Hans-Ulrich Wittchen

Mental Health 187
Traolach Brugha

Mental Health and Social Work 193
Di Bailey

Mental Health, Mentalism and Sanism 200
Jennifer M Poole and Tania Jivraj

Mental Illness, Etiology of 204
Peter Tyrer

Mental Illness, Genetics of 209
James J Rucker

Mental Illness: Family and Patient Organizations in Europe 216
JH Henderson and R Ramsay

Mental Imagery: Visual Cognition 220
Zaira Cattaneo and Juha Silvanto

Mentally III, Public Attitudes towards 228
Dinesh Bhugra, Sara Evans-Lacko, and W Cutter

Mentoring: Supervision of Young Researchers 235
Rachelle Hollander

Meritocracy and Tokenism 239
Stephen C Wright and Gregory D Boese

Merton, Robert K (1910–2003) 246
Christian Fleck

Mesoamerica, Archaeology of 252
Lane F Fargher

Meta-Analysis, History of 266
David J Roelfs

Meta-Analysis: Theory 272
Larry V Hedges and Arend M Kuyper

Metacognitive Development: Educational Implications 282
Wolfgang Schneider

Metaphor, Role in Social Thought: History of the Concept 289
David Lambourn

Method of Moments and Generalized Method of Moments 294
Lars P Hansen

Methodological Individualism in Sociology 302
Emmanuel Picavet

Methodological Individualism: Philosophical Aspects 308
Pierre Demeulenaere

Methodological Issues in the Study of Sexuality 314
Stuart Michaels

Methodology of the History of the Social and Behavioral Sciences 319
Christian Fleck and Christian Dayé

Mexican Revolution, The 326
Thomas Rath

Microeconomics, History of 331
Paola Tubaro

Microevolution 338
Francisco M Salzano

Microsimulation in Demographic Research 343
Emilio Zagheni

Middle Ages, The 347
Gábor Klaniczay

Middle East, North Africa: Sociocultural Overview 354
Seteney Shami

Middle Eastern and North African Legal Traditions 360
Lena Salaymeh

Midlife Psychological Development 369
Mathias Allemand

Migrants and Migration in Europe 376
David Coleman

Migration and Development 389
Ronald Skeldon

Migration and Fertility 397
Emilio A Parrado

Migration and Health 407
Sunshine Rote and Kyriakos Markides

Migration and Social Capital 411
Nadia Y Flores-Yeffal

Migration and Social Work 416
Usha George

Migration: Anthropological Perspectives 422
Caroline B Brettell

Migration: Cultural Aspects 427
Marcelo M Suárez-Orozco

Migration, Economics of 433
George J Borjas

Migration, Ethnicity, Aging and Social Work Practice 440
Naina Patel

Migration History 446
KJ Bade

Migration: Motivations 452
Douglas S Massey

Migration Out of Europe 457
Timothy J Hatton

Migration to and from Latin America 462
Douglas S Massey

Migration, Theory of 466
Douglas S Massey

Migrations, Colonizations, and Diasporas in Archaeology 472
Marcel Otte

Militaries and Militarization, Anthropology of 483
Andrew Bickford

Militarism 490
Anna Stavrianakis

Military and Disaster Psychiatry 495
Neil Greenberg

Military Geography 501
Rachel Woodward

Military History 506
Hew FA Strachan

Military Sociology 511
Mary K Kniskern and David R Segal

Military, War, and Politics 518
Jan W Honig

Mill, John Stuart (1806–73) 524
Iain Hampsher-Monk

Millennialism 529
Bryan R Wilson

Mills, Charles Wright (1916–62) 533
A Javier Treviño

Mind, Theories of 539
Fulvia Castelli

Mind–Body Dualism 545
Rom Harré

Mindfulness 549
Katherine Gardhouse and Zindel Segal

Mineral Resources, Limits to: The Case of Peak Oil 554
Ugo Bardi

Minimum Wages 561
William Wascher

Minnesota Multiphasic Personality Inventory 566
James N Butcher

Minorities 569
Richard T Schaefer

Miracles, Cultural Logic of 575
Paolo Parigi

Mirror Neurons, Theory of 582
Michael C Corballis

Misdemeanor Offenses 589
DeAnn Kalich, Margot Hasha, and Craig J Forsyth

Mismatch on the Labor Market 595
Kostas Mavromaras and Peter J Sloane

Missing Data 602
Roderick J Little and Donald B Rubin

Missing Girls: A Globalizing Issue 608
Christophe Z Guilmoto

Mission: Proselytization and Conversion 614
Brian Stanley

Mixed Methods 618
Abbas Tashakkori, Charles Teddlie, and Burke Johnson

Mixture Models in Statistics 624
Geoffrey J McLachlan

Mobile Communication 629
Rich Ling

Mobility and Travel Activity Patterns 636
Joe Weber and Mei-Po Kwan

Mobility, Social 640
Walter Müller and Reinhard Pollak

Model Selection and Model Averaging 647
Gerda Claeskens and Maarten Jansen

Model Testing and Selection, Theory of 653
Karl Christoph Klauer

Models of Influence in Networks 659
Noah E Friedkin

Models of Power in Exchange Networks 667
John Skvoretz

Moderator Variable: Methodology 672
Charles M Judd

Modern Management: Origins and Development 675
John-Christopher Spender

Modernity 682
Bernard Valade

Modernity: Anthropological Aspects 687
Aihwa Ong

Modernity: History of the Concept 692
Gurminder K Bhambra

Modernization 697
Manfred G Schmidt

Modernization and Modernity in History 700
Paul Nolte

Modernization, Sociological Theories of 707
Sergei Gavrov and Igor Klyukanov

Modes of Production 714
Donald L Donham

Modularity and Organizations 718
Carliss Y Baldwin

Monasticism: West and East 724
William JF Keenan

Monetary Policy since the 2007–2009 Financial Crisis 727
Benjamin M Friedman

Money and Finances, Psychology of 735
Eva Jonas, Jochim Hansen, Sabine Walper, and Dieter Frey

Money, Sociology of 741
Frederick F Wherry

Money: Anthropological Aspects 746
Bill Maurer

Montague Grammar 750
Barbara H Partee

Monte Carlo Methods and Bayesian Computation: Importance Sampling 758
Roman Liesenfeld and Jean-Francois Richard

Monte Carlo Methods and Bayesian Computation: Overview 763
Siddhartha Chib

Moore Jr, Barrington (1913–2005) 768
Dennis Smith

Moral Development and Education 775
Larry Nucci

Moral Development, Cultural Differences in 781
Klaus Helkama and Florencia M Sortheix

Moral Development, Theories of 788
Marc Jambon and Judith G Smetana

Moral Entrepreneur 796
Marianne E de Campo

Moral Panics 800
James M Jasper

Moral Reasoning in Psychology 803
Elliot Turiel

Morality 'East' and 'West': Cultural Concerns 806
Kwang-Kuo Hwang

Morality: Evolution of 811
Steve Stewart-Williams

Morphology in Linguistics 819
Antonio Fábregas

Morphophonology 826
Hans Basbøll

Mortality and the HIV/AIDS Epidemic 834
Sara Hertog and Cheryl Sawyer

Mortality, Biodemography of 844
Anatoliy I Yashin and Eric Stallard

Mortality Crossover 848
Irma T Elo

Mortality Differentials: Selection and Causation 851
Noreen Goldman

Mortality, Epidemiological, and Health Transitions 857
Graziella Caselli

Mortality of the Oldest-Old 863
Vladimir Canudas-Romo and Adam Lenart

Mortality: The Great Historical Decline 868
Tommy Bengtsson

Mother-Child Communication: Cultural Differences 874
Carolin Demuth

Motherhood 881
Karen J Swift

Motivation and Actions, Psychology of 887
Peter M Gollwitzer and Gabriele Oettingen

Motivation and Emotion: Reflections on the History of an Attributional Approach 894
Bernard Weiner

Motivation in Australian Aboriginal Populations 899
Rhonda G Craven and Alexander Seeshing Yeung

Motivation in Clinical Interventions 907
Daniel W McNeil and Laura L Quentin

Motivation in Youth Sport and Physical Activity: Developmental Perspectives 914
Maureen R Weiss and Alison C Phillips

Motivation, Familial Influences on 921
Sandra D Simpkins and Jennifer A Fredricks

Motivation, Learning, and Instruction 928
Eric M Anderman and DeLeon Gray

Motivation: History of the Concept 936
Peter M Gollwitzer and Gabriele Oettingen

Motivation: Life Course and Sociological Perspectives 940
Steven Hitlin, Matthew A Andersson, and Glen H Elder, Jr

Motivational Development, Systems Theory of 946
Marianne E Schneider

Motivational Development: Cross-Cultural Perspectives 951
Jan Hofer and Holger Busch

Motor Control Models: Learning and Performance 957
Pietro G Morasso

Motor Cortex 965
Steven P Wise

Motor Skills: Development in Infancy and Early Childhood 971
Cibelle KMR Formiga and Maria BM Linhares

This page intentionally left blank

CONTRIBUTORS TO VOLUME 15

Mathias Allemand
University of Zurich, Zurich, Switzerland

Richard J. Allen
University of Leeds, Leeds, UK

Eric M. Anderman
The Ohio State University, Columbus, OH, USA

Matthew A. Andersson
Yale University, New Haven, CT, USA

K.J. Bade
Osnabrveck, Germany

Di Bailey
School of Social Sciences, Nottingham Trent University, Nottingham, UK

Carliss Y. Baldwin
Harvard Business School, Boston, MA, USA

Ugo Bardi
Università di Firenze, Firenze, Italy

Hans Basbøll
University of Southern Denmark, Odense M, Denmark

Katja Beesdo-Baum
Institute of Clinical Psychology and Psychotherapy, Behavioral Epidemiology, Technische Universität Dresden, Dresden, Germany

Tommy Bengtsson
Lund University, Lund, Sweden

Meredith Bergey
Brandeis University, Waltham, MA, USA

Gurminder K. Bhambra
University of Warwick, Warwick, UK

Dinesh Bhugra
King's College London, London, UK

Andrew Bickford
George Mason University, Fairfax, VA, USA

Gregory D. Boese
Simon Fraser University, Burnaby, BC, Canada

George J. Borjas
Harvard Kennedy School, Harvard University, Cambridge, MA, USA

Caroline B. Brettell
Southern Methodist University, Dallas, TX, USA

Traolach Brugha
College of Medicine, Biological Sciences and Psychology, University of Leicester, Leicester, UK

Kevin Buckler
University of Texas at Brownsville, Brownsville, TX, USA

Holger Busch
University of Trier, Trier, Germany

James N. Butcher
University of Minnesota, Minneapolis, MN, USA

Vladimir Canudas-Romo
Max-Planck Odense Center on the Biodemography of Aging, University of Southern Denmark, Odense C, Denmark

Graziella Caselli
Sapienza University of Rome, Rome, Italy

Fulvia Castelli
University of Pavia, Pavia, Italy

Zongchao Cathy Li
School of Communication, University of Miami, Coral Gables, FL, USA

Zaira Cattaneo
University of Milano-Bicocca, Milano, Italy

Siddhartha Chib
Washington University, St. Louis, MO, USA

Gerda Claeskens
ORStat and Leuven Statistics Research Center, KU Leuven, Leuven, Belgium

David Coburn
University of Toronto, ON, Canada; and University of Victoria, Victoria, BC, Canada

David Coleman
University of Oxford, Oxford, UK

Peter Conrad
Brandeis University, Waltham, MA, USA

Michael C. Corballis
School of Psychology, University of Auckland, Auckland, New Zealand

Daina Crafa
Douglas Mental Health Institute, McGill University, Montreal, QC, Canada

Rhonda G. Craven
Institute for Positive Psychology and Education, Australian Catholic University, Strathfield, NSW, Australia

W. Cutter
South London and Maudsley NHS Trust, Denmark Hill, London, UK

Christian Dayé
University of Graz, Graz, Austria

Marianne E. de Campo
Faculty of Public Administration, Berlin School of Economics and Law, Berlin, Germany

Pierre Demeulenaere
University of Paris-Sorbonne, Paris, France

Carolin Demuth
Aalborg Universitet, Aalborg, Denmark

Donald L. Donham
University of California, Davis, CA, USA

Brendan Dooley
CACSSS/Graduate School, University College Cork, Cork, Ireland

Adam Drozdek
Duquesne University, Pittsburgh, PA, USA

Glen H. Elder
Carolina Population Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Irma T. Elo
Population Studies Center, University of Pennsylvania, Philadelphia, PA, USA

Sara Evans-Lacko
King's College London, London, UK

Antonio Fábregas
University of Tromsø, Tromsø, Norway

Lane F. Fargher
Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional — Unidad Mérida, Yucatán, México

Christian Fleck
University of Graz, Graz, Austria

Nadia Y. Flores-Yeffal
Texas Tech University, Lubbock, TX, USA

Cibelle K.M.R. Formiga
University of Goiás State, Goiás, Brazil

Craig J. Forsyth
University of Louisiana, Lafayette, Lafayette, LA, USA

Jennifer A. Fredricks
Human Development Department, Connecticut College, New London, CT, USA

Dieter Frey
Ludwig-Maximilians-Universität, Munich, Germany

Noah E. Friedkin
University of California, Santa Barbara, CA, USA

Benjamin M. Friedman
Harvard University, Cambridge, MA, USA

Katherine Gardhouse
University of Toronto Scarborough, Toronto, ON, Canada

Sergei Gavrov
Moscow State University of Design and Technology, Moscow, Russia; and Moscow City Pedagogical University, Moscow, Russia

Usha George
Ryerson University, Toronto, ON, Canada

Noreen Goldman
Princeton University, Princeton, NJ, USA

Peter M. Gollwitzer
University of Konstanz, Konstanz, Germany; and New York University, New York, NY, USA

DeLeon Gray
North Carolina State University, Raleigh, NC, USA

Neil Greenberg
King's College London, London, UK

Christophe Z. Guilmoto
IRD/CEPED, Paris, France

Iain Hampsher-Monk
University of Exeter, Exeter, UK

Jochim Hansen
University of Salzburg, Salzburg, Austria

Lars P. Hansen
University of Chicago, Chicago, IL, USA

Jens Harbecke
Witten/Herdecke University, Witten, Germany

Rom Harré
Georgetown University, Washington, DC, USA; and Linacre College, Oxford, UK

Margot Hasha
University of Louisiana, Lafayette, Lafayette, LA, USA

Timothy J. Hatton
University of Essex, Colchester, UK; and Research School of Economics, College of Business and Economics, Australian National University, Canberra, ACT, Australia

Harlene Hayne
University of Otago, Dunedin, New Zealand

Larry V. Hedges
Northwestern University, Evanston, IL, USA

Klaus Helkama
University of Helsinki, Helsinki, Finland

J.H. Henderson
St Thomas' Hospital, London, UK

Sara Hertog
Population Division, United Nations Department of Economic and Social Affairs, New York, NY, USA

Steven Hitlin
University of Iowa, Iowa City, IA, USA

Jan Hofer
University of Trier, Trier, Germany

Rachelle Hollander
National Academy of Engineering, Washington, DC, USA

Jan W. Honig
King's College London, London, UK

Judith A. Houck
University of Wisconsin—Madison, Madison, WI, USA

Kwang-Kuo Hwang
National Taiwan University, Taipei, Taiwan

Kana Imuta
University of Queensland, St Lucia, QLD, Australia

Marc Jambon
University of Rochester, Rochester, NY, USA

Maarten Jansen
Université libre de Bruxelles, Brussels, Belgium

James M. Jasper
New York, NY, USA

Tania Jivraj
School of Social Work, Ryerson University, Toronto, ON, Canada

Burke Johnson
University of South Alabama, Mobile, AL, USA

Eva Jonas
University of Salzburg, Salzburg, Austria

Charles M. Judd
University of Colorado, Boulder, CO, USA

DeAnn Kalich
University of Louisiana, Lafayette, Lafayette, LA, USA

William J.F. Keenan
Nottingham Trent University, Nottingham, UK

Gábor Klaniczay
Medieval Studies Department, Central European University, Budapest, Hungary

Karl Christoph Klauer
Albert-Ludwigs-Universität Freiburg, Freiburg, Germany

Igor Klyukanov
Eastern Washington University, Cheney, WA, USA

Mary K. Kniskern
University of Maryland, College Park, MD, USA

Valerie Kretz
University of Wisconsin—Madison, Madison, WI, USA

Arend M Kuyper
Northwestern University, Evanston, IL, USA

Mei-Po Kwan
University of Illinois at Urbana-Champaign, Champaign, IL, USA

Alfons Labisch
Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

David Lambourn
University of Warwick, Coventry, UK

Adam Lenart
Max-Planck Odense Center on the Biodemography of Aging, University of Southern Denmark, Odense C, Denmark

Roman Liesenfeld
Eherhard Karls University, Tubingen, Germany

Rich Ling
Nanyang Technological University, Singapore

Maria B.M. Linhares
School of Medicine at Ribeirão Preto, University of São Paulo, São Paulo, Brazil

Roderick J. Little
University of Michigan, Ann Arbor, MI, USA

David P. MacKinnon
Arizona State University, Tempe, AZ, USA

Marie-Louise Mares
University of Wisconsin—Madison, Madison, WI, USA

Kyriakos Markides
The University of Texas Medical Branch at Galveston, Galveston, TX, USA

Douglas S. Massey
Office of Population Research, Princeton University, Princeton, NJ, USA

Bill Maurer
University of California Irvine, Irvine, CA, USA

Kostas Mavromaras
National Institute of Labour Studies, Flinders University, Adelaide, SA, Australia and IZA

Jonathan D. Mayer
University of Washington, Seattle, WA, USA

Geoffrey J. McLachlan
University of Queensland, Brisbane, QLD, Australia

Daniel W. McNeil
West Virginia University, Morgantown, WV, USA

David Mechanic
Institute for Health, Health Care Policy and Aging Research, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

Carrie J. Menkel-Meadow
University of California, Irvine, CA, USA

Randolf Menzel
Institut für Biologie, Freie Universität Berlin, Germany

Stuart Michaels
Academic Research Centers, NORC at the University of Chicago, Chicago, IL, USA

Masato Miyachi
University of Tokyo, Tokyo, Japan

Pietro G. Morasso
Italian Institute of Technology, Genoa, Italy

Walter Müller
Mannheim Center for European Social Research, University of Mannheim, Mannheim, Germany

Bennet Murdock
University of Toronto, Toronto, ON, Canada

Paul Nolte
International University, Bremen, Germany

Larry Nucci
Graduate School of Education, University of California, Berkele, CA, USA

Gabriele Oettingen
University of Hamburg, Hamburg, Germany; and New York University, New York, NY, USA

Aihwa Ong
University of California, Berkeley, CA, USA

Marcel Otte
University of Liège, Liège, Belgium

Richard G. Owens
School of Psychology, University of Auckland, Auckland, New Zealand

Paolo Parigi
Stanford University, Stanford, CA, USA

Emilio A. Parrado
Population Studies Center, The University of Pennsylvania, Philadelphia, PA, USA

Barbara H. Partee
University of Massachusetts Amherst, Amherst, MA, USA

Naina Patel
Ageing and Migration, Lucinges, France

Alison C. Phillips
School of Kinesiology, University of Minnesota, Minneapolis, MN, USA

Emmanuel Picavet
Université Paris-I Panthéon-Sorbonne, Paris, France

Jessica T. Piotrowski
University of Amsterdam, Amsterdam, The Netherlands

Reinhard Pollak
Freie Universität Berlin, Berlin, Germany; and WZB Berlin Social Science Center, Berlin, Germany

Jennifer M. Poole
School of Social Work, Ryerson University, Toronto, ON, Canada

Laura L. Quentin
West Virginia University, Morgantown, WV, USA

R. Ramsay
Edinburgh, UK

Thomas Rath
University College London, London, UK

Jean-Francois Richard
The University of Pittsburgh, Pittsburgh, PA, USA

David J. Roelfs
University of Louisville, Louisville, KY, USA

Sunshine Rote
The University of Texas Medical Branch at Galveston, Galveston, TX, USA

Eric W. Rothenbuhler
School of Communications, Webster University, Webster Groves, MO, USA

Judith Roumani
Writer and Editor, Sephardic Horizons, *Potomac, MD, USA*

Donald B. Rubin
Harvard University, Cambridge, MA, USA

James J. Rucker
The Institute of Psychiatry, Kings College London, London, UK

Lena Salaymeh
University of California (UC) Berkeley School of Law, Berkeley, CA, USA

Francisco M. Salzano
Federal University of Rio Grande do Sul, Porto Alegre, Brazil

Cheryl Sawyer
Population Division, United Nations Department of Economic and Social Affairs, New York, NY, USA

Damian Scarf
University of Otago, Dunedin, New Zealand

Richard T. Schaefer
DePaul University, Chicago, IL, USA

Manfred G. Schmidt
Ruprecht Karls Universität Heidelberg, Heidelberg, Germany

Marianne E. Schneider
University of Zurich, Zurich, Switzerland

Wolfgang Schneider
Department of Psychology, University of Würzburg, Würzburg, Germany

David R. Segal
University of Maryland, College Park, MD, USA

Zindel Segal
University of Toronto Scarborough, Toronto, ON, Canada

Seteney Shami
Social Science Research Council, Brooklyn, NY, USA; and The Arab Council for the Social Sciences, Beirut, Lebanon

Juha Silvanto
Faculty of Science and Technology, University of Westminster, London, UK

Sandra D. Simpkins
T. Denny Sanford School of Social and Family Dynamics, Arizona State University, Tempe, AZ, USA

Ronald Skeldon
University of Sussex, Brighton, UK; and Maastricht University, Maastricht, The Netherlands

John Skvoretz
University of South Florida, Tampa, FL, USA

Peter J. Sloane
National Institute of Labour Studies, Flinders University, Adelaide, SA, Australia and IZA

Judith G. Smetana
University of Rochester, Rochester, NY, USA

Dennis Smith
Loughborough University, Loughborough, UK

Florencia M. Sortheix
University of Helsinki, Helsinki, Finland

Cylor Spaulding
College of Fine Arts & Communication, Towson University, Towson, MD, USA

John-Christopher Spender
Kozminski University, Warszawa, Poland

Annabelle Sreberny
Leicester University, Leicester, UK

Don W. Stacks
School of Communication, University of Miami, Coral Gables, FL, USA

Eric Stallard
Center for Population Health and Aging, Duke University, Durham, NC, USA

Brian Stanley
Westminster College, Cambridge, UK

Anna Stavrianakis
University of Sussex, Brighton, UK

Fred C.J. Stevens
Maastricht University, Maastricht, The Netherlands

Steve Stewart-Williams
Swansea University, Swansea, UK

Hew F.A. Strachan
University of Glasgow, Glasgow, UK

Marcelo M. Suárez-Orozco
Graduate School of Education and Information Studies, University of California, Los Angeles, CA, USA

Karen J. Swift
York University, Toronto, ON, Canada

Abbas Tashakkori
University of North Texas, Denton, TX, USA

Charles Teddlie
Louisiana State University, Baton Rouge, LA, USA

Andrew Tolson
University of Leicester, Leicester, UK

A. Javier Treviño
Wheaton College, MA, USA

Paola Tubaro
University of Greenwich, Greenwich, UK

Elliot Turiel
Graduate School of Education, University of California, Berkeley, CA, USA

Peter Tyrer
Centre for Mental Health, Imperial College, London, UK

Bernard Valade
Paris Descartes University, Paris, France

Patti M. Valkenburg
University of Amsterdam, Amsterdam, The Netherlands

Cees P.M. van der Vleuten
Maastricht University, Maastricht, The Netherlands

Kimberley Voss
Nicholson School of Communication, University of Central Florida, Orlando, FL, USA

Helen G.M. Vossen
University of Amsterdam, Amsterdam, The Netherlands

Alex Wagner
University of Washington, Seattle, WA, USA

Sabine Walper
Deutsches Jugendinstitut, Munich, Germany

Xiaolu Wang
Columbia University, New York, NY, USA

William Wascher
Board of Governors of the Federal Reserve System, Washington, DC, USA

Joe Weber
University of Alabama, Tuscaloosa, AL, USA

Bernard Weiner
University of California, Los Angeles, CA, USA

Maureen R. Weiss
School of Kinesiology, University of Minnesota, Minneapolis, MN, USA

Frederick F. Wherry
Yale University, New Haven, CT, USA

Bryan R. Wilson
University of Oxford, Oxford, UK

Steven P. Wise
Olschefskie Institute for Neurobiology of Knowledge, Potomac, MD, USA; Edmond and Lily Safra International Institute of Neurosciences of Natal, Natal, Brazil

Hans-Ulrich Wittchen
Institute of Clinical Psychology and Psychotherapy, Technische Universität Dresden, Dresden, Germany

Rachel Woodward
Newcastle University, Newcastle upon Tyne, UK

Stephen C. Wright
Simon Fraser University, Burnaby, BC, Canada

Anatoliy I. Yashin
Center for Population Health and Aging, Duke University, Durham, NC, USA

Alexander Seeshing Yeung
Institute for Positive Psychology and Education, Australian Catholic University, Strathfield, NSW, Australia

Emilio Zagheni
University of Washington, Seattle, WA, USA

Media and Child Development

Jessica T Piotrowski, Helen GM Vossen, and Patti M Valkenburg, University of Amsterdam, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Groebel, volume 14, pp. 9458–9463, © 2001, Elsevier Ltd.

Abstract

Decades of research have shown that the relationship between media and childhood is not unidirectional but reciprocal. In this article, both directions of the media–child development relationship are presented. We discuss how child development predisposes children's media use and preferences by reviewing key developmental characteristics of four periods across childhood, the typical media use during each period, and then hypothesizing how these characteristics influence media use and preferences. Next, we present an overview on the negative and positive influences of media that are most prominent in the public debate. We conclude by identifying directions for future research.

Children's media has become a global, multibillion dollar business. From the beginnings of life, children in the twenty-first century are developing in front of a screen (Wartella and Robb, 2011). Infants now have a wide array of DVDs available for them. The early childhood television market has fierce competition as marketers try to reach this financially profitable audience through programming, advertisements, and licensing deals. Children in middle childhood are now playing video games at rates higher than ever before, while 80% of adolescents are now using social network sites such as Facebook to connect with peers (Lenhart et al., 2011). Media entertainment has become so pervasive in children's domestic and school environments that cultural commenters have alluded to this as the 'entertainization' of childhood (Gabler, 1998).

Given the proliferation of entertainment media, it is important for researchers, health practitioners, and public policy advocates to clearly understand the current media landscape of children. This includes understanding the intricate relationship between media and child development. Decades of research into children and the media have shown that the relationship between media and childhood is not unidirectional but reciprocal. In this article, we address both directions of the media–child development relationship. In the first part, we discuss how child development predisposes children's media use and preferences. In the second part, we discuss how media use influences child development. We conclude by identifying gaps in the literature that are critical for future research.

How Development Influences Children's Media Use and Preferences

The media children use, and in particular their media preferences, are predicted – in a large part – by their developmental capabilities. Researchers have suggested that young children have a preference for media that can at least be partly incorporated into their existing framework and that they show less preference for extremely simple or extremely complex stimuli (e.g., Anderson and Lorch, 1983). This moderate-discrepancy hypothesis predicts that at any given age, a moderate level of stimulus complexity is preferred and that this level increases as the child matures (Valkenburg and Cantor, 2000). First incorporated into Helson's adaptation level theory on sensory habituation (Helson, 1948), the moderate-discrepancy hypothesis has proven to be a plausible explanation for behavioral and sensory regulation (e.g., Berlyne's general theory of curiosity, Berlyne, 1960) as well as a viable explanation for why the media preferences of children in various age groups differ so greatly. After all, the perceived simplicity and complexity of media content changes dramatically as children mature. Media content that is only moderately discrepant and therefore attractive to 2-year-olds may be overly simple and thus unattractive to 6-year-olds.

As such, in order to understand children's media usage and preferences, it is important to understand how children typically develop. To that end, we present a brief review of key developmental characteristics of four periods across childhood (i.e., infancy and toddlerhood, early childhood, middle childhood, and adolescence) and hypothesize how these characteristics influence children's media use and preferences. We also present statistics regarding the amount of media that is typically consumed during each period. In this article, we mainly present data from American media use studies. However, the media use patterns presented here are similar to other industrialized countries as well (Beentjes et al., 2001; Szybist, 2011).

Infants and Toddlers

This age group involves children between birth and 2 years. In the first 2 years of life, children exhibit significant growth as they move from relying on basic reflexes to engaging in more purposeful behaviors around 18 months of age (Bukatko, 2007). When children are around 4 months of age, they begin to exhibit some interest in television watching. This onset is not accidental. At 4 months of age, the social smile emerges: infants start to laugh when they see or hear things they like. At the same time, their ability to detect cues in their environment has matured as has their ability to locate a sound in space by turning their head or eyes in the direction of the sound (Field et al., 1980).

In the first year of life, the orienting reflex mainly predetermines attention. Infants are typically interested in sudden

and novel sounds and movements along with bright colors. In the second year, attention becomes somewhat less influenced by novelty and more influenced by relevant and interesting content. By 18 months of age, their first problem-solving skills (means-end behavior) emerge, and with this, a preference for solving simple puzzles and playing simple games also emerges. Infants and toddlers have limited language skills, although they begin to experience a word spurt of nearly 20 new words per week around 18 months of age (Bukatko, 2007). At this point, they enjoy verbally labeling familiar objects (boat, house, train) and creatures (dog, cat) that they recognize when presented with a storybook or audiovisual media.

Media use for this age group is a rather controversial topic. Beginning with the introduction of *Baby Einstein* in the late 1990s, media for the infant and toddler demographic have become an enormous international industry (Christakis, 2009). And yet, many health care practitioners argue that we do not yet know enough about the (potentially negative) influence of media on this audience (e.g., Christakis, 2009). These and other similar concerns have brought about an emergence of public policy initiatives. For example, in the United States, the American Academy of Pediatrics discourages screen media exposure in the first 2 years of life (Council on Communications and Media, 2011), while France has taken the step of banning programming directed at infants. Yet, despite these efforts, recent data show that the average child under 1 year of age is exposed to nearly 1 h of television daily, and by 2 years of age, these rates are closer to 2 h of daily exposure. To put this in a historical context, in 1971, the average age at which children began watching television was about 4 years old. As discussed, today it is 4 months (Christakis, 2009).

Despite the relatively contentious debate on whether or not media is appropriate for this young audience, we do know that infants and toddlers exhibit very distinct media preferences. For example, this young audience prefers content that relies heavily on music and song. We also know that content with salient formal features is best able to hold children's attention. This preference wanes slightly as children become older toddlers (around 18 months old) and is replaced by an interest in simple narratives. Given their limited cognitive and verbal skills, content that relies on slow pacing, familiar contexts, incorporates significant repetition, provides opportunities for verbal labeling, and incorporates simple characters is also appealing for this group.

Early Childhood

Early childhood covers children between 3 and 6 years old. During this period of time, children typically demonstrate egocentrism (i.e., an inability to separate their own perspective from the perspectives of others) as well as perceptual boundedness (i.e., tendency to focus on immediately perceptible attributes of an object, whereas other types of information are ignored) and centration (i.e., tendency to focus on an individual, striking feature of an object, to exclude other less striking information) (Bukatko, 2007). Early childhood is characterized by an inability to distinguish between fantasy and reality. Although as of 18 months, most children are able to distinguish fantasy from reality in their own imaginary play, they are less able to do so when they need to judge the reality status of objects and stimuli in their environment, including the media.

Early childhood is also characterized by rapid emotional developments (Thomas, 2005). Newborns can produce facial expressions associated with primary emotions such as joy, sadness, and anger. However, at the onset of early childhood, children start to experience so-called conscious emotions, such as envy, guilt, and embarrassment. Conscious emotions differ from primary emotions in that they require perspective-taking skills (e.g., the notion that someone is disappointed with you). By age 4 years, children are able to understand these more complex conscious emotions in others. However, their detection of emotions is still perceptually bounded. Children in this age group are only able to recognize emotions based on external cues (e.g., crying; sad face) and not based on less perceptual cues.

Today's early childhood media market is one of the most competitive and crowded markets across the media landscape. Media use reports indicate that children between 3 and 6 years old spend nearly 3 h per day using media (Common Sense Media, 2011). Younger children (3 years) spend the majority of their media time using television. As children get older (4–6 years), we see that they also begin to add video games to their daily media diet (approximately 30 min per day). Unsurprisingly, nearly no time is spent engaging in social media use or other forms of interactive media use (e.g., Internet use, text messaging) during early childhood.

During early childhood, attention to media content increases dramatically (Valkenburg and Vroone, 2004). This reflects the rapid increase in children's information processing skills and their improved vocabulary. However, their cognitive and emotional capacity is still limited. Children in early childhood lack experience and semantic knowledge (e.g., grass is green; a chair has four legs), which can make processing new media content quite challenging. As a result, they typically prefer media content that relies on slow pacing, incorporates repetition, and contains simple characters in familiar contexts. These preferences may change at the end of early childhood when children, especially boys, become more attracted to fast-paced media, more complicated characters, and adventurous contexts.

The perceptual boundedness that is characteristic of early childhood also induces children to judge media characters on perceptually salient features, such as their appearance and perceptually visible motives and emotions. And, because of their inability to distinguish fantasy from reality in media, fantasy characters are just as attractive and engaging as real-life characters. By the time they are 3 years old, children start to make statements indicating attachment to television personalities. But they can be just as much or even more attracted to animals or cartoon characters than to real-life characters. Children in this age group are also very impressed by 'special effects' such as a character vanishing into a puff of smoke.

Middle Childhood

Middle childhood involves children between 7 and 11 years old. The most important difference between early and middle

childhood is that children in middle childhood begin formal schooling. During this stage, thought processes become more mature and adultlike. Although children are still unable to engage in abstract and hypothetical thinking, they are now able to think logically and solve problems in a logical fashion so long as the problems apply to concrete events or objects. Children can now clearly distinguish between fantasy and reality, and as a result, their admiration for cartoon characters decreases dramatically. Emotional development continues to improve in sophistication. Children are now better able to understand their own emotions and the emotions of others. Relatedly, children are increasingly able to understand the perspective of others. The combination of formal school entry and understanding other's perspectives results in an important focus on peer relationships.

Children in middle childhood continue to incorporate media into their daily lives at relatively high rates. In the beginning of middle childhood, we see estimates similar to early childhood with approximately 3 h per day spent using media – most of which is predominantly television (about 40 min with games, Common Sense Media, 2011). However, around age 8 years, children spend nearly 4 h per day viewing television, 1 h per day playing video games, and 30 min per day using social media sites (Rideout et al., 2010). Although television viewing is certainly increasing throughout middle childhood, these estimates highlight an increased interest and use of interactive media.

Along with an increased interest in interactive media, there are several other changes in middle childhood that influence media preferences. Beginning at age 6 years, children lose their interest for slow-paced educational programs such as *Sesame Street* and develop a preference for entertainment media that is faster and more complex. They also begin to develop an interest for action and violence. For this reason, there is an abundance of media entertainment developed for this age group that contains action, violence, and binary characters (i.e., good/bad, male/female). During middle childhood, children begin to 'decentrate' and develop an interest in detail. As a result, they prefer more complicated media characters and more sophisticated plotlines. That said, given that they still struggle with abstract thinking, they prefer content that focuses on concrete problems and objects (e.g., the quest for a treasure; saving a princess). Although children now understand the difference between fantasy and reality, not all children are equally interested in realistic content at this age. Girls typically develop an interest in real-life entertainment, whereas many boys continue to be attracted to fantasy content (e.g., *Transformers*, *Power Rangers*). There is also a change in character preference. While early childhood was characterized by a focus on character appearance, the increased emotional development of middle childhood results in a greater focus on the psychological characteristics of media characters. Children now enjoy characters with which they can psychologically identify, for example, characters with an attractive sense of humor.

Adolescence

Adolescence, which covers the period between 12 and 18 years old, is accompanied by rapid physical, psychological, and cognitive changes (Bukatko, 2007). Physical changes include both rising hormone levels (McClintock and Herdt, 1996) as well as brain development (Blakemore and Choudhury, 2006). In particular, the prefrontal cortex – an area of the brain responsible for decision making, planning, inhibiting inappropriate behavior, and understanding social relations – is developing dramatically. These physical changes also lead to psychological changes. For example, adolescence is often characterized by mood swings, more risk-taking behavior, and increasing sexual interest. Alongside these physical and psychological changes, cognitive development improves greatly and reaches an adult level at age 15 years. Adolescents are able to think abstractly and in a more sophisticated manner, and now possess the capacity for higher order reasoning.

Egocentrism reemerges during this period as adolescents become self-absorbed (i.e., viewing things from only their point of view) and idealistic (Thomas, 2005). This makes adolescents highly critical of authority figures (e.g., parents and teachers). Also as a result of this egocentrism, an imaginary audience may develop, that is, the belief by adolescents that others are constantly watching and thinking about them. Peer relationships are quite important to adolescents, and as they begin to establish their own unique identity, the imaginary audience perspective results in them looking to their peers for advice and support. Lastly, this egocentrism often leads adolescents to feel as though their own experiences are unique, which sometimes can result in risk-taking behaviors.

Media represent an important component of adolescents' daily life. Adolescents are spending more than 8 h per day using some form of media (Rideout et al., 2010). While television reigns supreme in terms of time (i.e., nearly 5 h), adolescents are spending nearly 3 h per day playing video games and using the Internet (e.g., social network sites, instant messaging, Rideout et al., 2010). Given the increased importance of peers during this period, it makes sense that adolescents would turn to the Internet – particularly social media – to foster and maintain these connections.

The physical, psychological, and cognitive changes that occur during adolescence strongly influence media preferences. First, because of the ability to think in an abstract manner, adolescents prefer more complex media compared to children in middle childhood. In particular, they prefer media that is developed for an adult audience. Horror movies, programs with an element of fear (e.g., *Jackass*), risk taking, sarcasm, and other forms of crude humor are specifically appealing to this audience. Second, perhaps as a result of their increased idealistic views, adolescents also prefer media with social lessons (e.g., solutions to social problems) and enjoy watching older characters that they can admire. Lastly, adolescents are particularly interested in using media that enable them to connect with friends as well as exercise self-expression and identify formation (e.g., social network sites such as Facebook, content creation sites such as YouTube, Greenhow et al., 2009).

The Effects of Media on Child Development

Given the developmental limitations of young audiences, the amount of time that children spend using media, and the constant evolution of media, it is unsurprising that many researchers are concerned about the potential influence of

media. In fact, these concerns have fueled more than 50 years of research on the role that media play in children's lives (e.g., Pecora et al., 2007). Beginning in the 1950s, much of this research focused on the negative effects of media on child development – with particular attention to the role of media violence on aggression. By the 1970s, with the emergence of *Sesame Street*, researchers also began to ask whether media could positively affect children's development. Although a complete review on the positive and negative influences of media is beyond the scope of this article, we present an overview of the effects of media that are most prominent in the public debate. We begin with outlining several negative outcomes associated with media use including aggressive behavior, attention-deficit hyperactivity disorder (ADHD)-related behaviors, sexual socialization, sexual risk behavior, materialism, parent–child conflict, and childhood overweight. We then present several positive outcomes associated with media use including prosocial behavior, school readiness, core subject knowledge, creativity, and imagination.

Negative Effects

Aggressive Behavior

When it comes to the effects of media on children, the influence of violent media on aggressive behavior has received the most research attention. Hundreds of correlational, experimental, and longitudinal studies have evaluated whether and how violent media exposure is related to subsequent aggression. These studies have typically looked at the influence of violent media exposure in middle childhood and adolescence (e.g., Gentile et al., 2004; Huesmann et al., 2003), although some studies have also investigated violent media effects in early childhood (e.g., Christakis and Zimmerman, 2007). The majority of studies have evaluated the influence of violent television content (e.g., Bushman and Huesmann, 2001) on aggression, although in recent years the influence of violent interactive media (e.g., video games and computer games) has received increased research attention (e.g., Anderson and Dill, 2000).

While some studies have suggested no relationship between violent media exposure and aggression (e.g., Ferguson et al., 2011), most demonstrate that watching violent television or playing violent video games is positively related to aggressive behavior and may even lead to aggressive behavior later in life (e.g., Huesmann et al., 2003). Meta-analytic research supports this argument. Paik and Comstock (1994) conducted a meta-analysis to ascertain the effect size associated with violent movies and violent television programs, whereas Anderson and Bushman (2001) were interested in the effect size associated with violent video games. Both analyses suggested small to moderate effects (i.e., $r = 0.31$ and 0.19, respectively) of violent media on aggressive behavior. Moderator analyses also indicate that the effects of media violence are stronger for boys than for girls and stronger for younger than for older children (Paik and Comstock, 1994).

Several theories have been used to explain how violent media content can increase aggressive behavior. Two of the most widely used theories are social cognitive theory (Bandura, 2001) and desensitization theory. According to social cognitive theory, children learn behavior not only through their own experiences but also by observing others. When children see others behave in a certain way and subsequently being punished or rewarded for this behavior, children learn which behaviors are socially acceptable and which ones are not. These social cognitions then help guide the actual behavior of children. Desensitization theory, on the other hand, states that repeated exposure to media violence leads to a gradual habituation in response to displays of aggression. Over time and with increased exposure to media violence, children become more accustomed to aggressive behavior, which subsequently impacts moral judgments and behaviors.

In light of the large body of research on violent media exposure and aggression, most media researchers today accept that violent media exposure plays a role in children's aggressive tendencies. Many of these researchers also argue that both social cognitive theory or desensitization theory are fair explanations for this relationship. Yet, these same researchers acknowledge that there remain many unanswered questions. For example, we know little about which children are most susceptible to violent media exposure and also how family or peer environments may moderate this relationship. Research that seeks to identify who is most susceptible to the influence of violent media is a critical area for future research.

ADHD-Related Behaviors

ADHD is a behavioral disorder characterized by elevated levels of inattentiveness, hyperactivity, and impulsivity (American Psychiatric Association [DSM-IV-TR], 2000). Although ADHD is a clinical diagnosis, many authors considered ADHD as the extreme end on a continuum of behaviors. In recent decades, there has been an increase in the number of children diagnosed with ADHD (e.g., Akinbami et al., 2011) and receiving pharmacological treatment for ADHD-related behaviors (e.g., Hollingworth et al., 2011).

There have been recurrent claims that specific features of entertainment media may cause children's ADHD-related behaviors (Christakis, 2009; Nigg, 2009). Most of the claims focus upon children in early and middle childhood, as these are the periods in which ADHD diagnoses typically occur. While the research investigating these claims remains limited, existing research supports the argument that time spent with media (television and video games) may play a role in subsequent ADHD and ADHD-related behavior (Christakis and Zimmerman, 2007; Miller et al., 2007). For example, in longitudinal research, Swing et al. (2010) found that television and game exposure was related to greater attention problems in a middle childhood sample. Other researchers have gone beyond global exposure amounts to look at the relationship between specific media content and ADHD-related behaviors. This research has shown that violent media content is related to more attention problems (Kronenberger et al., 2005) and impulsivity (Anderson and Maguire, 1978).

To date, most of the effects of entertainment media use on ADHD-related behaviors have been attributed to the violent content of media. An important hypothesis associated with this is the arousal-habituation hypothesis. This hypothesis posits that violent media negatively affect children's basic arousal level, which may subsequently lead to attention problems, hyperactivity, and impulsivity (Huizinga et al., 2013).

Although the existing literature intimates a positive relationship between (violent) media exposure and ADHD-related

behaviors, research in this field is still in its infancy and there remain many gaps in our understanding of this relationship. For example, most existing work has mainly focused on the direct relationship between media use and ADHD-related behaviors. Future research should investigate underlying mechanisms and test whether and if the arousal-habituation hypothesis sufficiently explains this relationship. More research on the effect of fast-paced media on ADHD-related behaviors is also needed in order to differentiate between the effect of fast-paced media and violent media. Finally, evaluating the role of individual differences is an important direction for future work.

Sexual Socialization and Sexual Risk Behavior

Sexual socialization – that is, the process by which knowledge, attitudes, and values about sexuality are acquired (Ward, 2003), typically begins during adolescence. The media are consistently acknowledged as an important agent of adolescents' sexual socialization (see Ward, 2003 for a review). Adolescents are heavy users of media and the media they consume frequently include sexual content. As a result, they are exposed to an abundance of sexual content in television, movies, magazines, and the Internet – and this content has caused many researchers to ask what the effects of such exposure might be.

To better understand how sexual content is presented in the media, researchers have conducted content analyses of frequently used media such as television and the Internet. A content analysis from the United States, for example, indicated that two out of three programs on television (aside from sports, news and children's programs) contain some sexual content, whereas only 1% of these programs place primary emphasis on sexual risks or responsibilities (Kunkel et al., 2007). Furthermore, sexually explicit material (SEM) is easily accessed on the Internet. In online SEM, sex is often portrayed as a merely physical, self-indulgent activity between uncommitted partners with few negative consequences (e.g., Mehta, 2001). This portrayal of sex in the media is thought to increase sexual risk behavior in adolescents. Empirical research has in fact found that exposure to sexual media content can result in earlier initiation of sexual behavior (e.g., Brown et al., 2006; Collins et al., 2004), increased sexual risk behavior such as unprotected sex (Peter and Valkenburg, 2011), and objectification of women (Peter and Valkenburg, 2007).

The negative effects of sexual media content on adolescents' sexual behavior are typically explained by cultivation theory (Gerbner et al., 2002). Cultivation theory states that heavy media users, when compared to light media users, are more likely to believe that the real world is similar to the messages portrayed in the media. Therefore, if casual sex is often portrayed as not having any negative consequences in the media, heavy media users are more likely to believe this message.

The research to date certainly supports researchers' efforts to understand how sexual media content is related to adolescents' sexual behavior. However, most of the research has been conducted within relatively homogenous populations (e.g., American youth, Dutch youth). Considering that attitudes toward sex and sexual behaviors are dependent on culture and religiosity (Marston and King, 2006), efforts to compare media effects across different countries would be worthwhile. Furthermore, efforts to identify explanatory factors, such as perceived vulnerability or perceived norms of sexual behaviors, would provide important information to researchers interested in understanding the process of how sexual media affect adolescents as well as ways to circumvent this process.

Materialism, Parent–Child Conflict, and Childhood Overweight

As the children's media environment has become increasingly commercialized, advertising has become an important aspect of children's media exposure. More than ever, advertisers now view children as an important and relevant target group. And advertisers now rely on a host of technologies to reach these young audiences – ranging from traditional television commercials to branded Web sites and brand placement in video games. Along with this consistent growth in child-directed advertising, we have seen an increase in concerns associated with the negative influence of advertising on children's behavior and well-being. These concerns have predominantly focused on the negative influence of child-directed advertising on materialism, parent–child conflict, and childhood overweight. Research investigating these outcomes has typically looked at the effects of television advertising (e.g., Halford et al., 2004), although research has begun to address newer advertising formats found in video games and the Internet (e.g., Henry and Story, 2009).

In terms of materialism, scholars have argued that the inherent nature of advertising (i.e., it is designed to arouse desires for products that would otherwise not be salient) makes children more materialistic. Cross-sectional, experimental, and longitudinal studies have indeed demonstrated that advertising leads to an increase in materialism in children (Buijzen and Valkenburg, 2003; Greenberg and Brand, 1993; Opree et al., 2013). Furthermore, researchers have also argued that the exposure to advertisements leads to increased parent–child conflict because children will ask for the advertised products and parents do not want to comply. Indeed, studies have shown that advertising increases product requests and these product requests subsequently induce parent–child conflict (Buijzen and Valkenburg, 2003).

Finally, in recent years, advertising has been implicated as an important factor in children's likelihood to become overweight or obese. Researchers argue that the preponderance of junk food advertisements has created a generation of children who expect a heavy presence of these foods in their diets, and parents who opt against this practice are thought to act outside the norm (Jordan, 2007). As a result, advertising exposure is said to stimulate intake of high-calorie and low-nutrient food and beverages among children, which subsequently leads to overweight and obesity (food advertising effects hypothesis, Gantz et al., 2007). A growing body of empirical research supports this contention (e.g., Borzekowski and Robinson, 2001; see also Buijzen and Valkenburg, 2012).

The evidence for the negative influence of advertising on children is rather convincing. While more studies are needed to identify those children who are most susceptible to the negative effects of advertising, it is also important to recognize that advertising is part of today's consumer culture. Although one can imagine taking steps to limit children's exposure to violent media content or sexual media content, it is harder to imagine ways to limit children's exposure to advertising content. Perhaps as a result of this fact, scholars are now investigating

ways to help children become more resistant to advertising messages (e.g., by teaching advertising literacy) as well as identifying whether and how the power of advertising can be used to promote healthier outcomes in childhood (e.g., healthier eating).

Positive Effects

All too often, discussions surrounding the effects of media in childhood focus upon negative effects. And while accurate and certainly of concern, these negative effects do not present the entire picture. Far less attention has been paid to the role of developmentally appropriate media in supporting positive outcomes among children. Yet, as Fisch (2004) has observed, if we believe that children can learn negative lessons from media, it stands to reason that they can learn positive lessons too. To that end, we review the existing research on the positive impacts of media on prosocial behavior, school readiness skills, core subject knowledge, creativity, and imagination.

Prosocial Behavior

Prosocial behavior refers to a range of positive behaviors including positive interactions (e.g., friendly play or peaceful conflict resolutions), altruism (e.g., sharing, offering help), and behaviors that reduce stereotypes (Mares and Woodard, 2007). Researchers who believe that media can influence prosocial behavior argue that depictions of prosocial behavior in the media are more consistent with social norms than antisocial behavior, and as a result, imitations of these behaviors are more likely to be received positively than imitation of antisocial acts (Rushton, 1979).

Since the late 1970s, more than 30 studies have been conducted to identify whether and how prosocial media content might contribute to prosocial behavior. These studies have largely focused upon the relationship between prosocial television content and subsequent prosocial behavior (e.g., *Mister Rogers' Neighborhood*, Friedrich and Aletha Huston, 1973). A meta-analysis by Mares and Woodard (2007) revealed that children who watched more prosocial television content exhibited significantly more prosocial behavior. The overall effect size ($r = 0.27$) resulting from this meta-analysis is similar to that found in meta-analyses of the relationship between violent television content and subsequent aggression. Television is therefore just as prone to fostering prosocial behavior as it is to fostering aggression (Mares and Woodard, 2007). The relationship reported by Mares and Woodward was strongest when prosocial behaviors were explicitly modeled in the television show. Moreover, children's age moderated this relationship such that the effect of prosocial content increased sharply during early childhood, peaked at age 7 years, and then declined until and throughout the teen years. Mares and Woodard (2007) posit that the peak in middle childhood suggests that younger children may lack the cognitive ability to fully understand prosocial acts on television. More research is needed to understand why the relationship weakens after middle childhood. A possible explanation is that this decrease in effect is due to a dearth of developmentally appropriate prosocial television content available for older children.

Recent studies also suggest that interactive media that are designed to support prosocial behavior do work. In a series of studies conducted in three countries across three different age groups (i.e., middle childhood, adolescence, adults), researchers found robust evidence to support the argument that playing video games in which characters model prosocial behavior does increase short- and long-term prosocial behaviors (Gentile et al., 2009). Similar findings have been found by other scholars investigating the relationship between prosocial video game play and subsequent prosocial behavior, particularly in middle childhood and adolescence (Saleem et al., 2012).

Overall, the literature indicates that exposure to media that explicitly depict prosocial behavior does subsequently result in children enacting this behavior. This effect is most typically explained through Bandura's social cognitive theory (Bandura, 2001) or via the general learning model (GLM) (Gentile et al., 2009). Both models focus on the imitation of modeled behaviors, although the GLM pays particular attention to affective, arousal, and cognitive effects during media use. These theories have helped researchers identify why prosocial media content can be beneficial for children, yet there remain many unanswered questions. In particular, more research is needed to identify best practices when creating prosocial media content for children. We also know little about the role of individual differences in determining who is exposed to prosocial content and how individual differences impact processing of this content.

School Readiness

School readiness refers to children's abilities to engage in and benefit from formal schooling. This concept encompasses early academic skills as well as social and emotional development. The majority of research on media and school readiness has focused on the role of television in early childhood. More than 1000 studies, for example, have examined the influence of the television program *Sesame Street* on children's school readiness. The earliest of these studies (Ball and Bogatz, 1973; Ball et al., 1974) was a pair of experiments testing the effectiveness of the first two seasons of the program. Results indicated that *Sesame Street* viewers showed significant gains across a host of school readiness skills. These effects have been replicated and extended in numerous correlational, experimental, and longitudinal studies since then – both with the American version of *Sesame Street* as well as international coproductions of the program (Fisch, 2004). In fact, a recent meta-analysis on the effects of the international coproductions of *Sesame Street* revealed significant positive effects of exposure to the program on children's school readiness skills (Mares and Pan, 2013). *Sesame Street*, however, is only one example of a program that has been shown to support school readiness. There are many others including *Barney & Friends* (Singer and Singer, 1994), *Blue's Clues* (Anderson et al., 2000), and *Pinky Dinky Doo* (Linebarger and Piotrowski, 2009). What these shows, and others like them, have in common is that they were developed with the explicit intent to support school readiness skills. Building upon formative and summative research, these programs merge best pedagogical practices with an understanding of the needs and preferences of young audiences.

Compared to television, evidence to support the benefits of interactive media on school readiness remains in its infancy. The available research indicates that developmentally

appropriate interactive media designed with the explicit intent to support school readiness skills are effective (see Lieberman et al., 2009). However, more work is needed to understand how the benefits of interactivity can be capitalized upon to support school readiness skills.

The mechanisms to explain how media affect school readiness skills are less frequently discussed in the literature. Some researchers proscribe to the capacity model (Fisch, 2004) – a model of children's learning from television (and, to some extent, interactive media) that focuses upon children's allocation of working memory to media content, while others focus on modeling-based theories such as social cognitive theory (Bandura, 2001). The notable theoretical consistency across studies investigating school readiness is the recognition that young children are active viewers who use the formal features embedded within media to guide their attention and comprehension (Anderson and Lorch, 1983). However, much more research is needed to understand the theoretical mechanisms to explain when, and for whom, media can support school readiness.

Core Subject Knowledge

Core subject knowledge refers to a broad array of information and skills that children are introduced to during formal schooling including literacy, mathematics, and science. The majority of research looking at the relationships between media use and core subject knowledge has focused on the role of television in middle childhood. For example, experimental research with the literacy-based program *Between the Lions* found that daily exposure to the program, over the course of 4 weeks, resulted in significant literacy gains (Linebarger et al., 2004). Similarly, experimental research with the math-focused television show *Cyberchase* revealed positive results on children's mathematical problem-solving skills (Fisch, 2003). And research looking at the influence of television on science skills has shown that television programs designed to support science knowledge (e.g., *The Magic School Bus* and *Bill Nye the Science Guy*) do support viewers understanding of scientific concepts as well as how to engage in the process of scientific discovery (ARC Consulting LLC, 1995; Rockman et al., 1996). Like school readiness, these and other similar findings illustrate that developmentally appropriate media content designed with the explicit intent to support children's core subject knowledge is effective.

The available research on interactive media also suggests that interactive media can support core knowledge. For example, children who actively use the Web site corollary of the program *Cyberchase* demonstrated gains in their mathematical problem-solving skills – particularly when used in combination with the television program (Fisch et al., 2010). Similarly, electronic storybooks can support children's reading comprehension (Doty et al., 2001) and phonics skills (McKenna et al., 2003) and using interactive media to demonstrate scientific concepts results in better concept comprehension when compared to standard lessons (Ardac and Akaygun, 2004).

Despite the growing knowledge base on the role of media in supporting children's core subject knowledge, the mechanisms to explain how media support these skills are less known. As with school readiness, the capacity model (Fisch, 2004) and social cognitive theory (Bandura, 2001) emerge as the two most prevalent explanations for these effects. However, the paucity of theoretical research on why and how media can affect core knowledge skills highlights a critical area for future research.

Creativity and Imagination

Creativity and imagination are multifaceted concepts that are operationalized in different ways throughout the literature. Creativity is most typically defined as the ability to generate many different novel or unusual ideas (Valkenburg and van der Voort, 1994), and is commonly evaluated during middle childhood and adolescence. Imagination is typically divided into two subcategories: (1) imaginative play, which is defined as play in which children transcend reality by acting 'as if,' and (2) daydreaming, which is defined as a state of consciousness in which attention turns to thoughts or images that are based on memory (Valkenburg and Calvert, 2012). Imaginative play typically occurs during early childhood years while daydreaming is more likely to occur in middle childhood and throughout adolescence.

Surprisingly, little research has been done to evaluate whether media can support creativity and imagination during childhood. In the early 1990s, Valkenburg and van der Voort conducted a pair of reviews to identify whether television supports or hinders children's creativity and imagination (Valkenburg and van der Voort, 1994; van der Voort and Valkenburg, 1994). The results of these reviews suggested that television viewing in general hinders children's imaginative play and creativity. However, television programs that are designed to stimulate creativity and imagination may actually do so (see Valkenburg and Calvert, 2012). For example, research with the television program *Barney* (a program designed specifically to stimulate children's imagination) indicated that, after 2 weeks of viewing the program, viewers demonstrated more imaginative play than their nonviewing peers (Singer and Singer, 1998). Longitudinal evidence also revealed that viewing educational television during early childhood was associated with increased creativity during adolescence (Anderson et al., 2001).

Just as research on television is limited, research on the role of interactive media in creativity and imagination is equally scarce. Given the contingent feedback that is characteristic of interactive media, and given that such feedback has been shown to support innovation, scholars have wondered why interactive media have not taken a more prominent role in training children to be creative (Fishman, 2004; see also Singer and Singer, 2011). Moreover, in an era where children can design avatars and various trial selves to interact in cyberspace, where they can create their own videos (e.g., YouTube) and share them with others, where they can play games that allow them to construct their own homes, cities, landscapes (e.g., Sim), the potential benefits of interactive media on creativity and imagination seem vast. And yet, presently, we know very little. Correlational evidence suggests that video game playing can support creativity in middle childhood and adolescence (see also Calvert and Valkenburg, 2013; Jackson et al., 2012), but causal evidence to support this finding does not yet exist. Future efforts to understand whether and how interactive media can support children's creativity and imagination are warranted.

Conclusion

Media use and child development are reciprocally linked processes. This article has demonstrated this reciprocal relationship by illustrating how development plays a predictive role in children's media use and preferences as well as how media use subsequently impacts development in both positive and negative ways. Our article on the influence of child development on media usage and preferences highlights the consistent presence of television in children's lives across developmental stages, while at the same time showing that content preferences change dramatically as children get older. Children's preference for interactive media also increases with age. While the very youngest children are not yet engaging with interactive media, children in early and middle childhood are incorporating video games into their daily life, and by adolescence, social media use is daily norm. As technology continues to evolve, interactive media are likely to make inroads into children's lives at younger and younger ages. It will be important for researchers to track these changes in use and preference as well as identify how these changes affect children's lives.

This article also highlights how media can affect child development, making the clear point that quality is much more important than quantity. While it is certainly true that a healthy media diet is one that limits the amount of media children are exposed to, it is also critical to realize that a healthy media diet must pay attention to what children are consuming. Violent content, sexual content, and advertising content are reasons for parental concern and public policy initiatives. Yet on the contrary, developmentally appropriate media content designed with the explicit intent to support healthy outcomes in children can be beneficial for children's cognitive, emotional, and social developments. Efforts to help families achieve a healthy media diet that balances both quantity and quality are worthwhile.

Our article further identifies three important gaps in the literature. First, if one of our goals is to understand how media affect children, then our methods should reflect this. While laboratory experiments will certainly continue to provide important information regarding causality, longitudinal research designs are needed to help researchers assess the long-term consequences of children's media use without sacrificing external validity.

Second, all too often, the potential underlying mechanisms to explain why media affect children are ignored. In some cases, this results from the use of theories that do not carefully explicate the role of specific mediating and moderating variables (e.g., social cognitive theory, see Valkenburg and Peter, 2013). In other cases, researchers theorize about underlying mechanisms but do not provide empirical support to substantiate these claims (e.g., arousal-habituation hypothesis, also see discussion in Potter, 2011). As a result, after more than 50 years of research, although we can say with some confidence that violent media exposure is related to subsequent aggression, it remains largely unclear "*which* effects of *which* kind of media violence on *which* types of aggressive behavior can be attributed to cognitive, physiological, or emotional underlying processes" (Valkenburg and Peter, 2013). Efforts to investigate how media affect children, rather than simply investigating whether they affect children, will make important contributions to the literature.

And finally, researchers must pay more attention to conditional media effects. This article shows that media can have positive and negative effects on children, but it is still largely unclear for whom these effects occur. The most important question for future research is to investigate which children are highly susceptible to positive effects, which are highly susceptible to negative effects, and which are highly susceptible to both. Only by comparing the susceptibles to the insusceptibles can we obtain a true understanding of the role of media in child development.

Acknowledgment

This article was supported by a grant to the third author from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no [AdG09 249488-ENTCHILD].

See also: Communication Research and Media Studies, History of; Educational Media: Potentials for Learning; Internet and Social Media: Anthropological Aspects; Lifelong Learning and Its Support with New Technologies; Mass Media and Cultural Identity; Mass Media, Representations In; Media Effects on Children; Media Effects; Social Media; Social Science and the Media; Teens, Gender, and Self-Presentation in Social Media; Violence and Media; Visual Images in the Media.

Bibliography

Akinbami, O.J., Liu, X., Pastor, P.N., Reuben, C.A., 2011. Attention Deficit Hyperactivity Disorder Among Children Aged 5–17 Years in the United States, 1998–2009. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Atlanta, GA.

Anderson, C., Maguire, T., 1978. The effect of TV viewing on the educational performance of elementary school children. Alberta Journal of Educational Research 24 (3), 156–163.

Anderson, C.A., Bushman, B.J., 2001. Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: a meta-analytic review of the scientific literature. Psychological Science 12 (5), 353–359. http://dx.doi.org/10.1111/1467-9280.00366.

Anderson, C.A., Dill, K.E., 2000. Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. Journal of Personality and Social Psychology 78 (4), 772. http://dx.doi.org/10.1037//0022-3514.78.4.772.

Anderson, D.R., Bryant, J., Wilder, A., Santomero, A., Williams, M., Crawley, A.M., 2000. Researching Blue's Clues: viewing behavior and impact. Media Psychology 2 (2), 179–194. http://dx.doi.org/10.1207/S1532785XMEP0202_4.

Anderson, D.R., Huston, A.C., Schmitt, K.L., Linebarger, D.L., Wright, J.C., 2001. Early childhood television viewing and adolescent behavior: the recontact study. Monographs of the Society for Research in Child Development 66 (1), 1–143.

Anderson, D.R., Lorch, E.P., 1983. Looking at television: action or reaction. In: Bryant, J., Anderson, D.R. (Eds.), Children's Understanding of Television: Research on Attention and Comprehension, first ed. Academic, San Diego, CA, pp. 1–33.

ARC Consulting LLC, 1995. Research Findings: The Magic School Bus (No. Volume 1, Executive Summary and Volume 2, Comprehensive Report). Author, New York, NY.

Ardac, D., Akaygun, S., 2004. Effectiveness of multimedia-based instruction that emphasizes molecular representations on students' understanding of chemical change. Journal of Research in Science Teaching 41 (4), 317–337. http://dx.doi.org/10.1002/tea.20005.

Ball, S., Bogatz, G.A., 1973. Reading with Television: An Evaluation of the Electric Company (No. 1). Educational Testing Service, Princeton, NJ.

Ball, S., Bogatz, G.A., Kazarow, K.M., Rubin, D.B., 1974. Reading with Television: A Follow-up Evaluation of the Electric Company (No. 2). Educational Testing Service, Princeton, NJ.

Bandura, A., 2001. Social cognitive theory of mass communication. Media Psychology 3, 265–299. http://dx.doi.org/10.1207/S1532785XMEP0303_03.

Beentjes, J.W.J., Koolstra, C.M., Marseille, N., van der Voort, T.H.A., 2001. Children's use of different media: for how long and why? In: Livingstone, S., Bovill, M. (Eds.), Children and Their Changing Media Environment: A European Comparative Study. Lawrence Erlbaum, Mahwah, NJ, pp. 85–112.

Berlyne, D.E., 1960. Conflict, Arousal, and Curiosity. McGraw-Hill, New York.

Blakemore, S., Choudhury, S., 2006. Development of the adolescent brain: implications for executive function and social cognition. Journal of Child Psychology and Psychiatry 47 (3–4), 296–312. http://dx.doi.org/10.1111/j.1469-7610.2006.01611.x.

Borzekowski, D.L., Robinson, T.N., 2001. The 30-second effect: an experiment revealing the impact of television commercials on food preferences of preschoolers. Journal of the American Dietetic Association 101 (1), 42–46. http://dx.doi.org/10.1016/S0002-8223(01)00012-8.

Brown, J.D., L'Engle, K.L., Pardun, C.J., Guo, G., Kenneavy, K., Jackson, C., 2006. Sexy media matter: exposure to sexual content in music, movies, television, and magazines predicts black and white adolescents' sexual behavior. Pediatrics 117 (4), 1018–1027. http://dx.doi.org/10.1542/peds.2005-1406.

Buijzen, M., Valkenburg, P.M., 2003. The unintended effects of television advertising: a parent-child survey. Communication Research 30 (5), 483–503. http://dx.doi.org/10.1177/0093650203256361.

Buijzen, M., Valkenburg, P.M., 2012. The intended and unintended effects of advertising on children. In: Scharrer, E. (Ed.), The International Encyclopedia of Media Studies, first ed. Oxford, UK: Blackwell. http://dx.doi.org/10.1002/9781444361506.wbiems132.

Bukatko, D., 2007. Child and Adolescent Development: A Chronological Approach, first ed. Cengage Learning, Florence, KY.

Bushman, B.J., Huesmann, L.R., 2001. Effects of televised violence on aggression. In: Singer, D.G., Singer, J.L. (Eds.), Handbook of Children and the Media, first ed. Sage Publications, Thousand Oaks, CA, pp. 223–254. http://dx.doi.org/10.1001/archpedi.160.4.348.

Calvert, S.L., Valkenburg, P.M., 2013. The influence of television, video games and the internet on children's creativity. In: Taylor, M. (Ed.), Oxford handbook of the Development of Imagination, first ed. Oxford University Press, New York, pp. 438–450.

Christakis, D.A., 2009. The effects of infant media usage: what do we know and what should we learn? Acta Paediatrica 98, 8–16. http://dx.doi.org/10.1111/j.1651-2227.2008.01027.x.

Christakis, D.A., Zimmerman, F.J., 2007. Violent television viewing during preschool is associated with antisocial behavior during school age. Pediatrics 120 (5), 993–999. http://dx.doi.org/10.1542/peds.2006-3244.

Collins, R.L., Elliott, M.N., Berry, S.H., Kanouse, D.E., Kunkel, D., Hunter, S.B., Miu, A., 2004. Watching sex on television predicts adolescent initiation of sexual behavior. Pediatrics 114 (3), 280–289. http://dx.doi.org/10.1542/peds.2003-1065-L.

Common Sense Media, 2011. Zero to Eight: Children's Media Use in America (No. 1). Common Sense Media, San Francisco, CA.

Council on Communications and Media, 2011. Media use by children younger than 2 years. Pediatrics. http://dx.doi.org/10.1542/peds.2011-1753.

Doty, D.E., Popplewell, S.R., Byers, G.O., 2001. Interactive CD-ROM storybooks and young readers' reading comprehension. Journal of Research on Computing in Education 33, 374–384.

Ferguson, C.J., San Miguel, C., Garza, A., Jerabeck, J.M., 2011. A longitudinal test of video game violence influences on dating and aggression: a 3-year longitudinal study of adolescents. Journal of Psychiatric Research 26 (2), 141–146. http://dx.doi.org/10.1016/j.jpsychires.2011.10.014.

Field, J., Muir, D., Pilon, R., Sinclair, M., Dodwell, P., 1980. Infants' orientation to lateral sounds from birth to three months. Child Development 51 (1), 295–298.

Fisch, S.M., 2003. The Impact of Cyberchase on Children's Mathematical Problem Solving (No. 1). MediaKidz Research & Consulting, Teaneck, NJ.

Fisch, S.M., 2004. Children's Learning from Educational Television. Lawrence Erlbaum, Mahwah, NJ.

Fisch, S.M., Lesh, R., Motoki, E., Crespo, S., Melfi, V., 2010. Children's Learning from Multiple Media in Informal Mathematics Education (No. 1). MediaKidz Research & Consulting, Teaneck, NJ.

Fishman, B., 2004. Creating a framework for research on systematic technology innovations. Journal of the Learning Sciences 13 (1), 43–76. http://dx.doi.org/10.1207/s15327809jls1301_3.

Friedrich, L.K., Aletha Huston, S., 1973. Aggressive and prosocial television programs and the natural behavior of preschool children. Monographs of the Society for Research in Child Development 38 (4), 1–64. http://dx.doi.org/10.2307/1165725.

Gabler, N., 1998. Life, the Movie: How Entertainment Conquered Reality. Random House, New York.

Gantz, W., Schwartz, N., Angelini, J.R., Rideout, V., 2007. Food for Thought: Television Food Advertising to Children in the United States (No. 1). The Henry J. Kaiser Family Foundation, Washington, DC.

Gentile, D.A., Anderson, C.A., Yukawa, S., Ihori, N., Saleem, M., Ming, Lim Kam, Sakamoto, A., 2009. The effects of prosocial video games on prosocial behaviors: international evidence from correlational, longitudinal, and experimental studies. Personality and Social Psychology Bulletin 35 (6), 752–763. http://dx.doi.org/10.1177/0146167209333045.

Gentile, D.A., Lynch, P.J., Linder, J.R., Walsh, D.A., 2004. The effects of violent video game habits on adolescent hostility, aggressive behaviors, and school performance. Journal of Adolescence 27 (1), 5–22. http://dx.doi.org/10.1016/j.adolescence.2003.10.002.

Gerbner, G., Gross, L., Morgan, M., Signorielli, N., Shanahan, J., 2002. Growing up with television: cultivation processes. In: Bryant, J., Zillman, D. (Eds.), Media Effects: Advances in Theory and Research. Lawrence Erlbaum, Mahwah, NJ, pp. 43–67.

Greenberg, B.S., Brand, J.E., 1993. Television news and advertising in schools: the "Channel one" controversy. Journal of Communication 43 (1), 143–151. http://dx.doi.org/10.1111/j.1460-2466.1993.tb01252.x.

Greenhow, C., Robelia, B., Hughes, J.E., 2009. Learning, teaching, and scholarship in a digital age web 2.0 and classroom research: what path should we take now? Educational Researcher 38 (4), 246–259. http://dx.doi.org/10.3102/0013189X09336671.

Halford, J.C., Gillespie, J., Brown, V., Pontin, E.E., Dovey, T.M., 2004. Effect of television advertisements for foods on food consumption in children. Appetite 42 (2), 221–225. http://dx.doi.org/10.1016/j.appet.2003.11.006.

Helson, H., 1948. Adaptation-level as a basis for a quantitative theory of frames of reference. Psychological Review 55 (6), 297–313. http://dx.doi.org/10.1037/h0056721.

Henry, A.E., Story, M., 2009. Food and beverage brands that market to children and adolescents on the internet: a content analysis of branded web sites. Journal of Nutrition Education and Behavior 41 (5), 353–359. http://dx.doi.org/10.1016/j.jneb.2008.08.004.

Hollingworth, S.A., Nissen, L.M., Stathis, S.S., Siskind, D.J., Varghese, J.M., Scott, J.G., 2011. Australian national trends in stimulant dispensing: 2002–2009. Australian and New Zealand Journal of Psychiatry 45 (4), 332–336. http://dx.doi.org/10.3109/00048674.2010.543413.

Huesmann, L.R., Moise-Titus, J., Podolski, C.L., Eron, L.D., 2003. Longitudinal relations between children's exposure to TV violence and their aggressive and violent behavior in young adulthood. Developmental Psychology 39, 201–221. http://dx.doi.org/10.1037/0012-1649.39.2.201.

Huizinga, M., Nikkelen, S.W.C., Valkenburg, P.M., 2013. Childrens' media use and its relation to attention, hyperactivity, and impulsivity. In: Lemish, D. (Ed.), The Routledge International Handbook on Children, Adolescents and Media, first ed. Taylor & Francis Ltd, London, p. 179.

Jackson, L.A., Witt, E.A., Games, A.I., Fitzgerald, H.E., von Eye, A., Zhao, Y., 2012. Information technology use and creativity: findings from the children and technology project. Computers in Human Behavior 28 (2), 370–376. http://dx.doi.org/10.1016/j.chb.2011.10.006.

Jordan, A.B., 2007. Heavy television viewing and childhood obesity. Journal of Children and Media 1 (1), 45–54. http://dx.doi.org/10.1080/17482790601005124.

Kronenberger, W.G., Mathews, V.P., Dunn, D.W., Wang, Y., Wood, E.A., Giauque, A.L., Li, T., 2005. Media violence exposure and executive functioning in aggressive and control adolescents. Journal of Clinical Psychology 61 (6), 725–737. http://dx.doi.org/10.1002/jclp.20022.

Kunkel, D., Farrar, K.M., Eyal, K., Biely, E., Donnerstein, E., Rideout, V., 2007. Sexual socialization messages on entertainment television: comparing content trends 1997-2002. Media Psychology 9 (3), 595–622. http://dx.doi.org/10.1080/15213260701283210.

Lenhart, A., Madden, M., Smith, A., Purcell, K., Zickuhr, K., Rainie, L., 2011. Teens, Kindness, and Cruelty on Social Network Sites: How American Teens Navigate the New World of "digital Citizenship" (No. 1). Pew Research Center's Internet & American Life Project, Washington, DC.

Lieberman, D.A., Bates, C.H., So, J., 2009. Young children's learning with digital media. Computers in Schools 26 (4), 271. http://dx.doi.org/10.1080/07380560903360194.

Linebarger, D.L., Kosanic, A.Z., Greenwood, C.R., Doku, N.S., 2004. Effects of viewing the television program between the lions on the emergent literacy skills of young children. Journal of Educational Psychology 96 (2), 297–308. http://dx.doi.org/10.1037/0022-0663.96.2.297.

Linebarger, D.L., Piotrowski, J.T., 2009. TV as storyteller: how exposure to television narratives impacts at-risk preschoolers' story knowledge and narrative skills. British Journal of Developmental Psychology 27, 47–69. http://dx.doi.org/10.1348/026151008X400445.

Mares, M., Pan, Z., 2013. Effects of Sesame street: a meta-analysis of children's learning in 15 countries. Journal of Applied Developmental Psychology 34, 140–151. http://dx.doi.org/10.1016/j.appdev.2013.01.001.

Mares, M., Woodard, E., 2007. Positive effects of television on children's social interaction: a meta-analysis. In: Preiss, R.W., Gayle, B.M., Burrell, N., Allen, M., Bryant, J. (Eds.), Mass Media Effects Research: Advances through Meta-analysis. Lawrence Erlbaum, Mahwah, NJ, pp. 281–300. http://dx.doi.org/10.1207/S1532785XMEP0703_4.

Marston, C., King, E., 2006. Factors that shape young people's sexual behaviour: a systematic review. The Lancet 368 (9547), 1581–1586. http://dx.doi.org/10.1016/S0140-6736(06)69662-1.

McClintock, M.K., Herdt, G., 1996. Rethinking puberty: the development of sexual attraction. Current Directions in Psychological Science 5 (6), 178–183.

McKenna, M.C., Reinking, D., Bradley, B.A., 2003. The effects of electronic trade books on the decoding growth of beginning readers. In: Joshi, R.M., Leong, C.K., Kaczmarek, B.L.K. (Eds.), Literacy Acquisition: The Role of Phonology, Morphology, and Orthography, first ed. IOS Press, Amsterdam, pp. 193–202.

Mehta, M.D., 2001. Pornography in usenet: a study of 9,800 randomly selected images. Cyberpsychology & Behavior 4 (6), 695–703. http://dx.doi.org/10.1089/109493101753376641.

Miller, C.J., Marks, D.J., Miller, S.R., Berwid, O.G., Kera, E.C., Santra, A., Halperin, J.M., 2007. Brief report: television viewing and risk for attention problems in preschool children. Journal of Pediatric Psychology 32 (4), 448–452.

Nigg, J.T., 2009. What Causes ADHD? Understanding What Goes Wrong and Why, first ed. Guilford Press, New York.

Opree, S.J., Buijzen, M., Van Reijmersdal, E.A., Valkenburg, P.M., 2013. Children's advertising exposure, advertised product desire, and materialism: a longitudinal study. Communication Research. Advance online publication. http://dx.doi.org/10.1177/0093650213479129.

Paik, H., Comstock, G., 1994. The effects of television violence on antisocial behavior: a meta-analysis. Communication Research 21, 516–546. http://dx.doi.org/10.1177/009365094021004004.

Pecora, N., Murray, J.P., Wartella, E. (Eds.), 2007. Children and Television: Fifty Years of Research. Lawrence Erlbaum, Mahwah, NJ.

Peter, J., Valkenburg, P.M., 2007. Adolescents' exposure to a sexualized media environment and notions of women as sex objects. Sex Roles 56, 381–395. http://dx.doi.org/10.1007/s11199-006-9176-y.

Peter, J., Valkenburg, P.M., 2011. The use of sexually explicit internet material and its antecedents: a longitudinal comparison of adolescents and adults. Archives of Sexual Behavior 40 (5), 1015–1025. http://dx.doi.org/10.1007/s10508-010-9644-x.

Potter, W.J., 2011. Conceptualizing mass media effect. Journal of Communication 61 (5), 896–915. http://dx.doi.org/10.1111/j.1460-2466.2011.01586.x.

Rideout, V., Foehr, U.G., Roberts, D.F., 2010. Generation M2: Media in the Lives of 8- to 18-Year Olds (No. 1). Henry J. Kaiser Family Foundation, Menlo Park, CA.

Rockman, et al., 1996. Evaluation of the Bill Nye the Science Guy Television Series and Outreach: A Report to KCTS-tv (No. 1). San Francisco, CA. http://www.rockman.com/about/index.php.

Rushton, J.P., 1979. Effects of prosocial television and film material on the behavior of viewers. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology. Academic Press, New York, pp. 321–351.

Saleem, M., Anderson, C.A., Gentile, D.A., 2012. Effects of prosocial, neutral, and violent video games on children's helpful and hurtful behaviors. Aggressive Behavior 38 (4), 281–287. http://dx.doi.org/10.1002/ab.21428.

Singer, D.G., Singer, J.L., 2011. Make-believe play, imagination, and creativity: links to children's media exposure. In: Calvert, S.L., Wilson, B.J. (Eds.), The Handbook of Children, Media, and Development, first ed. Wiley-Blackwell, UK, pp. 290–308.

Singer, J.L., Singer, D.G., 1994. "Barney & Friends" as Education and Entertainment: Phase 2. Can Children Learn through Preschool Exposure to "Barney & Friends?" (No. 2). Yale University Family Television Research and Consultation Center, New Haven, CT.

Singer, J.L., Singer, D.G., 1998. *Barney and friends* as entertainment and education: evaluating the quality and effectiveness of a television series for preschool children. In: Asamen, J.K., Berry, G. (Eds.), Research Paradigms, Television, and Social Behavior. Sage, Beverly Hills, CA, pp. 305–367.

Swing, E.L., Gentile, D.A., Anderson, C.A., Walsh, D.A., 2010. Television and video game exposure and the development of attention problems. Pediatrics 126 (2), 214–221. http://dx.doi.org/10.1542/peds.2009-1508.

Szybist, J., 2011. Is TV on the decrease? Televizion 24, 26.

Thomas, R.M., 2005. Comparing Theories of Child Development, sixth ed. Thomson Wadsworth, Belmont, CA.

Valkenburg, P.M., Calvert, S.L., 2012. Media and the child's developing imagination. In: Singer, D.G., Singer, J.L. (Eds.), Handbook of Children and the Media, second ed. Sage, Thousand Oaks, CA, pp. 157–170.

Valkenburg, P.M., Cantor, J., 2000. Children's likes and dislikes of entertainment programs. In: Zillman, D., Vorderer, P. (Eds.), Media Entertainment: The Psychology of Its Appeal, first ed. Erlbaum, Hillsdale, NJ, pp. 135–152.

Valkenburg, P.M., Peter, J., 2013. Five challenges for the future of media-effects research. International Journal of Communication 7, 197–215. http://ijoc.org/index.php/ijoc/article/view/1962/849.

Valkenburg, P.M., van der Voort, T.H.A., 1994. Influence of TV on daydreaming and creative imagination: a review of research. Psychological Bulletin 116 (2), 316–339. http://dx.doi.org/10.1037/0033-2909.116.2.316.

Valkenburg, P.M., Vroone, M., 2004. Developmental changes in infants' and toddlers' attention to television entertainment. Communication Research 31, 288–311. http://dx.doi.org/10.1177/0093650204263435.

van der Voort, T.H.A., Valkenburg, P.M., 1994. Television's impact on fantasy play: a review of research. Developmental Review 14 (1), 27–51. http://dx.doi.org/10.1037//0033-2909.116.2.316.

Ward, L.M., 2003. Understanding the role of entertainment media in the sexual socialization of American youth: a review of empirical research. Developmental Review 23 (3), 347–388. http://dx.doi.org/10.1016/S0273-2297(03)00013-3.

Wartella, E., Robb, M., 2011. Historical and recurring concerns about children's use of the mass media. In: Calvert, S.L., Wilson, B.J. (Eds.), The Handbook of Children, Media, and Development, first ed. Wiley-Blackwell, UK, pp. 7–26.

Media and History

Brendan Dooley, CACSSS/Graduate School, University College Cork, Cork, Ireland

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by M. Ferro, volume 14, pp. 9463–9468, © 2001, Elsevier Ltd.

Abstract

Media history considers the historical dimension of communicating information, knowledge, and values to a broad audience. The numerous approaches to media history share an interest in understanding the impact such structures have had on societies around the globe, the particular forms they have taken, and the dynamics of historical change. Each of these areas is the subject of a significant body of theoretical and empirical work, with many intersections and overlaps, taking examples from the various media, their spatial contexts, and development over time.

Introduction

Media history considers the historical dimension of communicating information, knowledge, and values to a broad audience. Although the term 'media' came into use only in the 1920s to denote the structures of such communication, media history takes account of the period at least from the advent of the handpress in the fifteenth century, and some interpretations include the scriptoria, oral traditions, and wall paintings of medieval times, delving occasionally into ancient and prehistory as well. The numerous approaches to media history share an interest in understanding the impact such structures have had on societies around the globe, the particular forms they have taken, and the dynamics of historical change. Each of these areas is the subject of a significant body of theoretical and empirical work, with many intersections and overlaps, taking examples from the various media, their spatial contexts, and development over time.

Major Trends

The terminus post quem for any media history would be the emergence of writing in several distinct civilizations around the globe some five millennia ago. By decoupling speaker and utterance, lifting language out of a determinate time and space, this episode began a shift in human history, still under way, which has to a large degree shaped the world we share. In the Renaissance age of cultural, economic, and demographical renewal, communication was set on a new footing by the application of movable type, first by Johannes Gutenberg in Mainz, Germany (1439). The first actual printed books, largely modeled on the manuscripts they were intended to replace, were not just luxury commodities, though they mostly circulated in the upper social echelons. Fly sheets and pamphlets reached a broader audience, and over the following centuries a wide variety of popular publications vastly increased the range of the written word and the drawn image.

Almost contemporary with the printing press, manuscript newsletters began circulating in multiple copies along the public mail routes between the major cities of Europe of which some of the best known are those eventually collected by the Augsburg-based Fugger family of merchants. Two to six pages in length, usually anonymous, and distinguished only by place name and date, appearing at regular weekly or biweekly intervals, they bore multiple stories regarding the economic, political, cultural, and religious news of the day. The first newspapers, printed in Strasbourg by Johannes Carolus from around 1605, simply provided word-for-word transcriptions of newsletter material. By 1650, there was already a daily newspaper in Leipzig, and by 1702, in London too (*The Daily Courant*). Over the course of the seventeenth century, in many places newspapers became increasingly differentiated from newsletters in the amount of critical comment they published.

The variety magazine, another seventeenth-century invention, increased the scope for regular publication of communications of all sorts. The new genre emerged directly from earlier media for exchanging cultural news, such as book catalogs and handbills, as well as manuscript letters, still valid conduits between scholars and amateurs alike. From these occasional media to the regular weekly or monthly emission, with a promise of indefinite continuation, of volumes concerning every aspect of cultural activity interesting to amateurs and specialists alike, was not a very great step and the first to take it was Denis de Sallo in his Paris *Journal des sçavans*, published from 1665. Soon afterward came the Roman *Giornale dei letterati* directed by the mathematician Michelangelo Ricci and collaborators. More along the lines of a record of academic work were the *Philosophical Transactions* of the Royal Society, published from 1662.

Important cultural movements thrived on a wider distribution of the products of the press while the printing industry profited in turn. In this light, the European Enlightenment could in some sense be said to be a media phenomenon, indeed, according to Robert Darnton (1979), analyzing the publishing career of Charles-Joseph Panckoucke, it was a business. A veritable explosion of journal publishing brought the ideas of the Enlightenment to larger numbers of readers while the printing trades benefited from advertisement of their wares in the pages of the literary press. Among the new genres developed and diffused in this context were the so-called 'moral weeklies,' notably the German-language *Discourse der Mahlern* (1721–23) published by Johann Jakob Bodmer and Johann Jakob Breitinger, and, directed particularly to female readers, *Die vernünftigen Tadlerinnen* of Johann Christoph Gottsched (1725–1727) all partly stemming from the French

Mercure galant (1672–1714) of Jean Donneau de Visé and others, and especially from the *Spectator* (1711–14) of Richard Steele and Joseph Addison. A similar case could be made for Romanticism as a movement, with necessary modifications to the type and volume of the media involved.

Political movements, too, were accelerated and to some degree shaped by the available media, from the English Civil War to the Masaniello Revolt in Naples, from the French Revolution to the Revolutions of 1848, from the various nationalisms to the Russian Revolution, from the Communist Revolution to the Cultural Revolution in China, from the student revolts of the 1960s to the breakup of the Eastern bloc, all of which have become better understood by a media history perspective. The World Trade Organization protests and the Arab Spring stand as demonstrations of how new generations have mastered the mobilizing possibilities of media-enabled discontent.

Across the centuries, the change-inducing potential of the media has inspired various forms of control exercised by church and state, sometimes in tandem, largely aimed at enforcing a social, cultural, religious, and political status quo. On the one hand, controls and prohibitions could be deployed in such a way as to allow distribution of certain notions among elites, whose very exclusiveness was defined in part by such privileged access; on the other hand, prohibitions could create sites of conflict, where the state or a ruling group could direct selected acts of violence against individuals whose inferiority or exclusion was marked by their vulnerability to such treatment. Among the first in early modern times to articulate the innovative potential of a wider circulation of knowledge was Tommaso Garzoni (1588) in his eulogy of the printing press: "This is the art that indicates the fools, that exposes the arrogant, that makes known those who are learned …." John Milton famously advocated press freedom, although the idea was not taken seriously until the Enlightenment, becoming enshrined first in the United States Bill of Rights in March 1789 and a few months later by the French National Assembly in the Declaration of the Rights of Man and Citizen. Meanwhile, such was the quantity and distribution of the products of the printing press that thinkers like Gaetano Filangieri (1780–1785) referred to public opinion as a 'tribunal.' From the outset, media potential and the imperatives of church, state, and powerful pressure groups have more or less run along a collision course, and mutatis mutandis, they still do today.

To media history may also belong various types of popular entertainment, from music performances to prose theater to opera, which in the nineteenth century attracted a mass following. Indeed, live performance has in certain periods been singled out for special forms of censorship because of the supposed mimetic power. New technology would eventually bring all such entertainments into the homes of millions.

Increasing literacy and increasing demand were met in the nineteenth century by the application of steam power to the printing press (fundamental contributions made between 1802 and 1818) and the use of cheap wood pulp for paper (from the 1840s). Already in the first years of the century, Hegel remarked that perusing the daily morning newspaper had become nothing less than a secular ritual joining readers in a simultaneous act, even if the events read about were no longer fresh. Photography, between 1800 and 1839, introduced the diffusion potential of images, eventually realized in cinematography (1873–96). Radio (1887–95) introduced the diffusion potential of telepresence, realized by television (1897–1925).

As the economic potential of media productivity came to be recognized, from the nineteenth to the twentieth century, the first media conglomerates were born. William Randolph Hearst (d. 1951) is said to have been the first media mogul, owner of dozens of newspapers across the United States and a movie studio as well. The second largest media conglomerate currently in the world in terms of revenue is the Disney Company, outgrowth of a concern founded by Walt and Roy Disney in 1923 and listed on the New York Stock Exchange since 1991. The largest is Comcast, with over 100 000 employees and a predominant influence in virtually every market and genre, owner of cable networks, news and sports programs, a movie studio (Universal), media-related theme parks, and even, in part, a professional basketball team. Among the largest in the Asian market is the Sony Corporation, including books and electronics, which developed from a firm begun in Tokyo in 1946 by Masaru Ibuka. The power of multinational media giants to defy local norms or at least local tabus in the gathering and publication of information about individuals or things regarded as newsworthy, while propounding particular political and social visions held by their owners, has most recently been discussed in the case of Australian media magnate Rupert Murdoch.

The Internet enters the realm of the media in a series of information and communication technology (ICT) developments going from ARPANET to the World Wide Web (1969–89) and begins to assume a predominant role. The advent of cloud computing and media in streaming motivates new approaches in industries that once thrived by the sale of tangible commodities such as disks, tapes, and further back in time, records and even cylinders, while new media, social media, and open media authorship suggest new possibilities for media productivity, whose significance may only be imagined. Interconnectvity enables scientific and scholarly collaboration while introducing new challenges. With governments increasingly pursuing online activity due to the potential for policing and corporations interested for purposes of, movements for open information, open source, open data, and the like have raised an increasingly insistent voice in favor of public interests, however vaguely defined. E-democracy defines a movement to utilize ICT resources for instituting consensus-based societies motivated by greater citizen participation than is customary in current representative systems (Macintosh, A., 2006).

The so-called 'digital divide' defines the boundary between areas of the Earth reached by the latest advances in ICT and those left behind. Whether a 'knowledge society' (in the terminology of Peter Drucker, 1969) or an 'information society' (Daniel Bell, 1976) is possible, where the electronic media do not reach, may be a matter of concern for overall global development.

Historical Change

Theories of media change date at least to Giambattista Vico (1725), who attempted to draw conclusions regarding the

recurrent cycles experienced by human societies from time to time and found that the communication of information played a key role. Reading back into the earliest periods of ancient society, he posited an initial 'poetic' stage of primitive cruelty without rational thought or articulate speech, where authority was believed to proceed directly from divine fiat. In the subsequent 'heroic' or aristocratic stage, political discourse was intentionally concealed from the purview of the general population and kept as a sort of professional monopoly among the ruling elite. "Naturally continuing to practice religious customs, they religiously continued to keep the laws mysterious and secret (this secrecy being the soul and life of aristocratic states)." Discussing political interests only among themselves, the aristocrats decided generally for the good of the territory because of their private interests as proprietors. In the 'human' stage, by contrast, everyone was required to know the law and to judge private utility in relation to society. "Naturally open, generous and magnanimous (being commanded by the multitude, who naturally understand natural equity) … [the nations] went on to make public what had been secret." Such, he believed, had happened in Europe within recent memory.

Modern media historians usually speak of three fundamental developments between early and modern times, namely, the advent of the handpress; the application of new technologies, including steam power, cinematography, radio, and television; and most recently, the emergence of ubiquitous computing and the Internet. Whether such developments can plausibly be denominated as 'revolutions' is among the issues that concern a body of theoretical work regarding historical change. In this connection, Schumpeter's notion of 'creative destruction' and Thomas Kuehn's notion of paradigm shift may be relevant, even though not originally intended to apply to media history.

Proponents of the 'revolution' thesis include Marshall McLuhan (1962) and, more recently, Frederic Barbier (2000). Peter Burke and Asa Briggs (2002) sound a cautionary note, eschewing any metahistorical pattern-seeking, in the light of the stops and starts, the movements back and forth, the geographical inconsistency, and the subcultures of nonconformity that will fail to be noticed in any period if a metahistorical perspective is rigorously applied. Adrian Johns (1998) and Andrew Pettegree (2010), analyzing the specific case of print, emphasize the aspects of continuity with the use of script, questioning the fixity usually attributed to print.

Some change theorists, such as Lewis Mumford (1934), adopt a form of technological determinism ultimately harking back to Karl Marx, although in the version of David Landes (1969), this approach supports an account drawing attention to the advantages of free-market capitalism. Curiosity- and demand-driven inventions, he maintains, have built the Western industries that provided the goods characteristic of a modern lifestyle. In later work, Landes (1998) complicated his earlier model of economic development by adding the effects of variables such as culture, religion, geography, and so forth to explain inequalities and unequal outcomes within the world economies. Joel Mokyer (1990) claims to bring an evolutionary and systems-analytic perspective to the theory of technological change. Technological systems, he notes, like all systems, must have some built-in stability, but when resistance to change prevents creativity or the acceptance of innovations, this may help explain some of the differences in media development.

More along the lines of a sociological determinism, Niklas Luhmann (2012) posits, "Societal communication produces various media/forms depending on what problem is to be solved." For instance, printing was an extension of the rationalization of book production already begun in the Middle Ages and basically tended to allow societies to take full advantage of the system-differentiation potential unleashed by the advent of writing. The age of electricity freed communication still further by placing energy production in a separate realm from the communication of messages. Finally, computer-assisted communication tends in the direction of a world society.

Theories of Impact

That the media, in each period, have been agents of change is self-evident; therefore, analytical distinctions need to be made. Fernand Braudel (1981) remarked that the printing press, for instance, "expanded and invigorated everything." Exactly which changes are to be credited to which media in a given period and how the change was experienced and by whom are the chief concerns of a body of theoretical and empirical work regarding media impact. Elizabeth Eisenstein (1979) has made a suggestive case for the role of print in triggering the Renaissance, the Protestant Reformation, and the so-called Scientific Revolution. Subsequent work has attempted to validate such claims in the light of other plausible causes for the observed phenomena.

A highly influential analysis, sometimes known as the liberal thesis, suggests that media across the ages have ultimately contributed to the emergence of democratic societies and a responsible electorate (Lasswell, 1948). Whether a clear line of development can indeed be traced from the media of the past to those in modern democracies is open to debate. Since the 1980s, the liberal thesis has largely been superseded by public sphere theory, especially in the form expressed by Jurgen Habermas ([1962]1989). This theory is particularly concerned with the emancipatory role of the press in given periods, sometimes pointing to the benefits accrued to particular social strata or classes. According to Habermas, between the seventeenth and eighteenth centuries in Europe, wherever there was relatively free access to information, "men making public use of their reason" among the emerging bourgeoisie, tending to the formation of a collective will, replaced a previous regime of public expressions in regard to power and authority confined to official celebrations demonstrating ('representing') the position of the monarch and court vis-à-vis the rest of the populations. Variations on this pattern may occur from place to place due to the presence or absence of a free press and a substantial bourgeoisie. Some proponents of this theory have been more sanguine than Habermas was about whether the 'bourgeois public sphere' (his terminology) was a wholly satisfactory outcome. A feminist version of the theory draws attention to the emancipatory role of the media in certain periods viewed from a gender perspective (e.g., Gelbart, 1987).

Criticisms of such views have drawn attention to the inapplicability of the category of 'representative' public sphere, in

particular early modern contexts (Michel Reinders, 2010), or to the different local and regional timing of the transition away from the 'representative.' More general criticisms have questioned the focus on the news media as the main constituent of a public sphere, pointing out the many other interaction contexts that may contribute (Bloemendal and Van Dixhoorn, 2010). In addition, when information becomes a commodity, industrial strategies and political agendas may produce manipulations and distortions disruptive to collective will formation. Public spheres, whatever and wherever they may be, surely contain as much knowledge as nonsense, as much reason as unreason, and as much information as propaganda so that equally predominant effects may be to reinforce prejudices and spread pseudoknowledge, when not simply narcotizing by a surfeit of information. Some works question whether women have ever indeed fully participated in the public sphere as defined in the literature (Magalhães et al., 2012). Political philosopher John Rawls (1993), on the other hand, rather than the emancipatory effect of abstracted reason in the ideal situation imagined by most public sphere theories, posited instead, as the true sphere of reasonable citizenship, the real interchange of ideas in the semipublic spaces of associational life within civil society, tending to resolve issues between real persons and among groups, which occurs from day to day apparently without recourse to coercion, organized or otherwise, not to be confused with the more restricted realm of public decision making within the institutions.

Habermas (1975) and Manuel Castells (1996) have both pointed to the legitimation crisis of late capitalism, whereby the public sphere, fragmented by the encroachment of political, economic, and social systems, no longer provides that area of public reason which obtained at the dawn of the mass media. Consequently, public reason retreats from open discussion of common issues of public concern into NGOs and voluntary organizations. Castells, for instance, points to the paradox of increasing globalization and aggregation in the realm of mass media and communications accompanied, as though in reaction, by an increasing trend toward reinforcement of local cultural affiliations, real or invented.

This would be the place to point out that in certain regimes, perhaps all regimes, the media have been said to be instrumentalized by elites, to a greater or lesser extent, for the purpose of coaxing the wider public into behaving in ways that promote the economic or political welfare of such elites. Propaganda theory attempts, either for purpose of practical use (Bernays, 1952) or for purpose of understanding (Lasswell, 1971; Doob, 1935; more recently Chomsky, 1988) or both, to identify the most effective means of utilizing the media for mass persuasion. Some critics point out that propaganda theory inevitably falls short at the moment when it fails to recognize that propagandists and their publics normally belong to the same society and may indeed be impelled by the same social forces that one side is pretending to direct. Jacques Ellul ([1962]1973) defines propaganda as the effect of a technological society embracing the entire person and tending to total integration. According to this analysis, "Propaganda stops us from feeling that things in society are oppressive and persuades us to submit with good grace." Enzensberger cautions against ignoring the manipulative potential of every type of media, insisting on the ways in which a 'consciousness industry' reinforces the essential class basis of advanced capitalism (Enzensberger, 1974).

Previous predictions about the imminent demise of television due to an advancing Internet have failed to account for the resilience of a medium that has evolved in part through on-demand programming to accommodate an increasing range of tastes and preferences on the part of viewers. More to the point have been the attacks on quality of content. Frankfurt School critics Horkheimer and Adorno ([1969] 2002) famously decried, "the impoverishment of aesthetic matter so drastically, that by tomorrow the thinly veiled identity of all industrial culture products can come triumphantly out into the open, derisively fulfilling the Wagnerian dream of the Gesamtkunstwerk, the fusion of all the arts in one work." In their view, matured on the basis of the observation of thousands of hours of programming, interesting only for its very banality, "The alliance of word, image, and music is all the more perfect than in Tristan because the sensuous elements which all approvingly reflect the surface of social reality are in principle embodied in the same technical process, the unity of which becomes its distinctive content." The only corrective could be a facetious one: "Television points the way to a development which might easily enough force the Warner Brothers into what would certainly be the unwelcome position of serious musicians and cultural conservatives." That such criticisms, quite apart from the evident hostility to a certain strand of German late-Romanticism, now appear somewhat quaintly elitist may only be due to shifting definitions of what is good taste and what is not.

If public sphere theory seems to be mainly concerned with the political aspect of the media, general systems theory at least has the benefit of taking a larger portion of the contents of the media under its purview. Such is the case because, at the most abstract level, this theory deals with societies' advancement toward ever-increasing structural complexity, suggesting a special role of the media in achieving that. Social networks, so the theory goes, set continuously shifting end points for themselves during the course of their self-reproduction and the media tend to help incorporate new information – the so-called 'feedback loops.' An advantage of this view, according to a few of its proponents, is that in this way all the behavior of collectivities – rational and irrational – can be examined equivalently, without the hazards to objectivity involved in labeling the irrational as deviant (Morin, 1977–80).

Much thought has been devoted to whether digital communications have fundamentally changed the way we think, teach, and write. Holding to a minimal effect are those, such as Robert Darnton (2009), who point out that a supposed shift to extensive and varied bibliographies, processed discontinuously, partially, and with varying degrees of attention, was already noticed in the beginning of print culture. Holding to a transformative effect are those who suggest an inevitable correlation between certain evident characteristics of the digital media and certain probable configurations that seem to turn up in the realms of thought and expression. The 'new literacy' (Coiro and Dobler, 2007) offers new possibilities to be realized by acquiring new skills.

Especially the takeoff of social media and the impact of dematerialization, reembodiment, sharing, co-construction of the self, and distributed memory have engaged a rethinking

of 1980s notions about the 'extended self' in media socio-psychology (Belk, 2013). Moreover, the prevalence of cross-border blogging has provided new insights for understanding the construction of national and ethnic identities (Rulyova and Zagibalov, 2012).

Theories of Structure and Content

The media gain cultural significance as much by the contents of their particular products as by the effect they have on audiences. They are differentiated by the ways in which they exercise these effects. Harold A. Innis (1950) characterized 'time-based media' as those, such as clay tablets or inscriptions on stone, meant to carry messages across the ages but not necessarily to the multitudes. These he contrasted with 'space-based media' capable of bearing more ephemeral messages to a mass audience. Rarely found in their pure types, he saw the first as characteristic of community-based, tradition-bound societies, while the second favored rapid change, administrative centralization, and empire.

Marshall McLuhan (1967) famously maintained, "the medium is the message." By this he meant that the tools of communication themselves, apart from the message content, profoundly shape an age. As for his own age, he was not as optimistic as were some technophiles; indeed, he remained much dismayed about the all-absorbing effects of the modern media on the individual psyche. When he said radio, telephone, and television "eliminate time and space factors in human association," "creating involvement in depth" (from Understanding Media, McLuhan, 1964), he said so with profound misgivings.

McLuhan remains a highly controversial figure. Nonetheless, media history continues to be intensely concerned with how the characteristics of particular media have impacted upon the transmission of ideas. In principle, the contents of books, films, broadcasts, and so forth can hardly be separated from these as media. However, when the content intersects with other discourses extraneous to media history per se, the more or less separate histories of literature, rhetoric, art, performance studies, film studies, and so forth may be called into play, and so also the various critical movements associated with literary deconstruction, postmodernism, post-structuralism, postcolonial studies, gender studies, queer theory, and so forth, none of which constitutes a media theory per se.

Regis Debray (2000) notes that a particular medium delivers its content not as a simple container but as an element within a system, setting the boundary conditions for the way in which such content may be received. To the special study of the 'frameworks that administer belief in that content,' he gave the name 'mediology.'

Late twentieth-century television critics such as Raymond Williams (1974) and Pierre Bourdieu (1998) have drawn attention to the structural features that have supposedly prevented the media from fulfilling their promise by limiting viewer autonomy and channeling responses. Paddy Scannell (2009) has more recently noted that the advent of on-demand programming and the density of the media landscape vitiate such criticisms, adding that the endless 'now' generated by the emerging global connectivity effectively overcomes "the time-space constraints of older transport and communication systems."

That film theory actually begins with Henri Bergson, as is sometimes said, can only be partly true, as his *Creative Evolution* (Bergson, 1911) uses the cinema mainly as a metaphor for human consciousness in general and the epistemological and physiological principles of his *Matter and Memory* (Bergson, 1912) have mostly been incorporated into media studies (by Gilles Deleuze, 1986–89) ex post facto. Walter Benjamin's view (Benjamin, [1936]1968) was arguably more influential, focusing on the psychosocial implications of the art of cinematography in the light of the role assigned to culture in the Marxian model of historical development. An age of industrialism and proletarian advance, he argued, had spawned a technology-based art form, transcending traditional values, associated with the uniqueness or 'aura' of a work, such as had predominated in the ages of the great 'artists.' The new art of cinema would not only destroy the old forms and concepts in a spirit attuned to 'the universal equality of things' but it would produce new expressions more attuned to modern tastes.

The realistic tendency in film was the object of numerous contributions to film criticism ranging from Rudolf Arnheim and Béla Balázs to Siegfried Kracauer to André Bazin. A new turn emerged in the *Cahiers du Cinema*, edited in Paris from 1951 by Bazin and others, when the contributing cinematographer, François Truffaut, propounded the auteur theory, a background for the so-called Nouvelle Vague cinema of Truffaut and Jean-Luc Godard. Structuralism, psychoanalysis, and Marxism became the magazine's dominant tendencies in the 1960s, and these entered Anglophone film theory largely by way of the journal *Screen*, published from 1969 by the UK Society for Education in Film and Television.

In recent times, widespread amateur engagement with digital media, according to Laura Rascaroli (2009) and others, has opened a new chapter not only in film history but in the history of self-expression and ego-documentation, as well as the textualizing, illustrating, narrating, and recording of personal and group experience, essentially amounting to a new practice of the media.

Research Methods

Media history has furnished compelling objects of analysis to legions of researchers, dating at least to the advent of media criticism in the sixteenth century (Garzoni, 1590). A media historiography extending from Gutenberg to modern times has experimented with the application of a wide array of empirical approaches and theoretical perspectives rooted in numerous disciplinary specializations. As far as audiences are concerned, literacy statistics exist only from the 1880s (Louis Maggiolo), based at first on signatures applied to official documents. Studies of the history of reading begin arguably with Daniel Mornet (1933) working with probate records from the late eighteenth century. Connecting particular audiences with particular media has been notoriously difficult, and more or less sophisticated semiotic approaches, drawing conclusions from media content, have continued to prevail ever since the advent of audience surveys, due in part to the absence of other kinds of evidence for audience response in the early periods.

Such, with many variations, were the approaches of the interdisciplinary Frankfurt School of social research, including Horkheimer and Adorno. The first real surveys are said to be the Payne Fund Studies (1929–33) concerning the influence of cinema on children. At the close of World War II, Robert K. Merton (1946) and his team combined empirical research with theoretical insights to study the effects of radio broadcasting on the sale of war bonds. Machine-assisted analysis has permitted survey-taking and quantification on an unprecedented scale.

A major problem for the retrospective application of some of the available social theories is that they have been elaborated in epochs when the conditions of the media were far different, and if used incautiously, may add more obscurity than clarity to the understanding of the past. An example is propaganda theory, whose use needs to take into account that previous media systems had far less potential for controlling, distributing, and excluding information than they have since the advent of modern technology, and explaining precisely how a particular propagandistic 'message' is 'diffused' by an early modern 'propagandist' (as the theories require) is notoriously difficult.

The tremendous quantities of data to be analyzed by the media historian, in the effort to grasp the changes and uniformities over time, have encouraged engagement with digital approaches. Vast databases have been created, and the texts and images from the various sorts of media have been marked up for digital processing. Numerous methods of digital content analysis have been deployed of which a prominent example, at least for text-based media, has been *General Inquirer: A Computer Approach to Content Analysis* (P.J. Stone et al., 1966). A method of qualitative analysis has been elaborated by Philipp Mayring (seventh ed., 2000). More recently, corpus-based studies, employing data mining techniques and linguistic analysis, have gained currency (on which the treatise by Andrew Hardie and Tony McEnery (2011) is particularly illuminating) not only for understanding the construction of texts but also for assessing style, sources, movement, and so forth in a comparative framework.

A particular challenge of recent research on print media, especially in regard to the more remote periods, has been to find ways of examining the contents of large numbers of publications distributed across various spatial, temporal, and linguistic boundaries in the effort to understand culturally based transmutations and transformations, as texts travel from place to place. The linguistic challenges alone have inspired attempts, such as by Roberto Valdeón, to join media history and translation studies (Valdeón, 2007).

As the products of visual and audiovisual media proliferate and the commercial demand for tools enabling management of personal and public archives becomes more pressing, much attention has been devoted to developing reliable automatic and semiautomatic systems of image and video retrieval. These systems are highly relevant to media history. Audiovisual media, such as television programming, are now analyzed retrospectively utilizing such systems, for instance, by the Key Visuals project of Peter Ludes (2001) with the collaboration of Otthein Herzog between Jacobs University and the University of Bremen, aimed at recovering the repertory of key visuals diffused to audiences from time to time around the globe, especially within news programs. The result may be a new understanding of the cross-cultural patterns and variations in the presentation of events, persons, and objects of interest to viewers in a given period. Media images, Ludes has said, "in their specific multimodal compositions, create collective (audio-) visual memories. Thereby they enhance mutual expectations and the coordination of social actions." Another challenge would regard the vast amount of visual material, not yet annotated or marked up in any way, now stored on the Web for which new approaches hold out the promise for future retrieval for purposes of analysis and study.

As the line between public and private communications keeps being redrawn, the definition of media will shift and so will the definition of what is of interest to media history. Huge amounts of personal data and commentary voluntarily vouchsafed by users and collected by social media companies, Facebook, etc., shared online and utilized for focused advertizing campaigns and the like constitute an unmatched source for understanding cultural change, social integration, identity formation, political alignment, gender attitudes, etc., so far accessible to researchers only piecemeal due to privacy protections and corporate interests. Government databases compiled from massive collection of public and private media sources originating among citizens and nonnationals alike, onshore and offshore, sanctioned by committees whose deliberations are off-limits to the public, may one day provide a key source of media history. So far, private research normally gains access to similar data only by accident or by stealth, as formal requests are routinely denied for reasons of national security. For instance, trillions of hours of random cellular telephone conversations of personages ranging from the insignificant to the significant, including heads of state and high-ranking military officials, digitized and stored, no one knows for how long or according to what classified criteria, could yield, by advanced tools of voice recognition and concept indexing, important results in terms of opinion tracking and, even with the most primitive of tools, key pieces of evidence explaining the processes of industrial decision making or the conundrums of international politics. Finally, highly reserved internal media documenting the day-to-day workings of the government agencies themselves, possibly as voluminous as all the media publicly available to private citizens if not more (Galison, 2004), are currently leaving a media trail for researchers of the future, detailing the words and deeds of persons currently protected from scrutiny by state secrets legislation supposing that no more sophisticated bulwarks or privileged shredding agreements are put in place to shield them from the curious. Meanwhile, penultimate versions of technology spun off from national security agencies engaged in compiling information gleaned from high-level surveillance of millions of homes and offices, now commercialized as Sqrrl, offering "secure database software for Big Data applications," may be of some use in analyzing some of the masses of data that are currently available to researchers, especially considering that it "supports a wide variety of analytics including search, SQL, and graph."

The need to archive and manage what is and what will be known of the media becomes all the more urgent, as the history of the media is ultimately the history of us.

See also: Absolutism, History of; Advertising: General; Ancient Greece and Rome, History of; Attitude: History of Concept; Attitudes, Political and Public Opinion; Audience Measurement; Audiences, Media; Bourgeoisie and Middle Classes, History of; Broadcasting: Regulation; Burke, Peter (1937–); Capitalism; Chomsky, Noam (1928–); Civil Society and the Public Sphere: History of the Concept; Communication and Transportation, History of; Community and Media; Computer Mediated Communication; Consumption, History of; Content Analysis; Corpus Linguistics; Cultural History; Cultural Production in Networks; Culture and Economy; Culture and Institutional Logics; Culture and Networks; Democratic Theory; Discourse Studies; English Revolution, The; Enlightenment; Entertainment; Film History; Film and Video Industry; First Language Acquisition, Developmental Psychology of; French Revolution, The; Globalization and World Culture; Habermas, Jürgen (1929–); Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; Human Language, Evolution of; Information Society, Geography of; Information Society; Internet and Culture; Journalism; Journals, Scholarly; Knowledge Society, History of; Kuhn, Thomas S. (1922–96); Liberalism: Historical Aspects; Literacy and Illiteracy, History of; Literacy and Orality; Mannheim, Karl (1893–1947); Marx, Karl (1818–83); Mass Communication: Normative Frameworks; Mass Media and Cultural Identity; Mass Media and Sports; Mass Media, Political Economy of; Mass Media, Representations in; Media Effects; Media Events; Media and Social Movements; Media, Crime and Crime Waves; Merton, Robert K (1910–2003); Modernization and Modernity in History; Modernization, Sociological Theories of; Networks and Cultural Consumption; Public Opinion: Social Attitudes; Public Sphere and the Media; Public Sphere: Eighteenth-Century History; Public Sphere: Nineteenth and Twentieth-Century History; Publishing as Medium; Radio as a Medium; Rawls, John (1921–2002); Reformation and Confessionalization; Renaissance; Schumpeter, Joseph A. (1883–1950); Social Justice: Historical and Theoretical Considerations; Social Media; Social System; Television: General; Visual Images in the Media.

Bibliography

Adorno, T.W., Horkheimer, M., [1969] 2002. Dialectic of Enlightenment (E. Jephcott, Trans.). Stanford UP, Stanford.

Allen, J.S., 1981. Popular French Romanticism: Authors Readers, and Books in the 19th Century. Syracuse University Press, Syracuse, New York.

Barbier, F., Bertho-Lavenir, C., 2000. Histoire des médias: de Diderot à Internet, second éd. rev. Armand Colin, Paris.

Baron, S., Dooley, B. (Eds.), 2001. The Politics of Information in Early Modern Europe. Routledge, New York.

Belk, R.W., 2013. Extended self in a digital world. Journal of Consumer Research 40 (3), 477–500.

Bell, D., 1976. The Coming of Post-Industrial Society. Basic Books, New York.

Benjamin, W., [1936]1968. In: Arendt, H., Bassett, C. Harcourt (Eds.), The Work of Art in the Age of Mechanical Reproduction. Brace and World, New York.

Bergson, H.-L., 1911. Creative Evolution (A. Mitchell, Trans.). Macmillan, New York.

Bergson, H.-L., 1912. Matter and Memory (N.M. Paul, W.S. Palmer, Trans.). Macmillan, New York.

Bernays, E.L., 1952. Public Relations. University of Oklahoma Press, Norman. Blackwell, Oxford.

Bloemendal, J., van Dixhoorn, A., 2010. 'De scharpheit van een gladde tong': Literaire teksten en publieke opinievorming in de vroegmoderne Nederlanden. Bijdragen en Mededelingen betreffende de Geschiedenis der Nederlanden 125, 3–28.

Bourdieu, P., 1998. On Television. New Press, New York.

Böning, H., 2002. Periodische Presse, Kommunikation und Aufklärung: Hamburg und Altona als Beispiel. Ed. Lumière, Bremen.

Braudel, F., 1981. Civilization and Capitalism, XV–XVIII Century. In: The Structures of Everyday Life, vol. 1. Harper and Row, New York.

Burke, P., Briggs, A., 2002. A Social History of the Media: From Gutenberg to the Internet. Polity, Cambridge. Cambridge University Press, Cambridge.

Castells, M., 1996. The Information Age: Economy, Society and Culture.

Chomsky, N., Herman, E.S., 1988. Manufacturing Consent: The Political Economy of the Mass Media. Pantheon Books, New York.

Coiro, J., Dobler, E., 2007. Exploring the Online Reading Comprehension Strategies Used by Sixth-Grade Skilled Readers to Search for and Locate Information on the Internet. Reading Research Quarterly 42 (2), 214–257.

Darnton, R., 1979. The Business of Enlightenment: A Publishing History of the Encyclopédie 1775–1800. Harvard University Press, Cambridge, MA.

Darnton, R., 2009. The Case for Books: Past, Present, and Future. Public Affairs, New York.

Debray, R., 2000. Transmitting Culture (E. Rauth, Trans.). Columbia University Press, New York.

Deleuze, G., 1986–89. Cinema, 2 vols. University of Minnesota Press, Minneapolis.

Doob, L., 1935. Propaganda: Its Psychology and Technique. New York.

Dooley, B. (Ed.), 2010. The Dissemination of News and the Emergence of 'Contemporaneity' in Early Modern Europe. Ashgate, Farnham, Surrey, UK.

Dooley, B., 1999. The Social History of Skepticism: Experience and Doubt in Early Modern Culture. Johns Hopkins University Press, Baltimore.

Drucker, P.F., 1969. The Age of Discontinuity: Guidelines to Our Changing Society. Harper & Row, New York.

Eisenstein, E.L., 1979. The Printing Press as an Agent of Change, 2 vols. Cambridge University Press, New York.

Ellul, J., [1962]1973. Propaganda (K. Kellen, J. Lerner, English Trans.), New York.

Enzensberger, H.M., 1974. The Consciousness Industry (S. Hood, Trans.). Seabury Press, New York.

Filangieri, G., 1780–1785. La scienza della legislazione, 7 vols. Raimondi, Naples.

Galison, P., 2004. Removing knowledge. Critical Inquiry 31 (1), 229–243.

Garzoni, T., 1588. La piazza universale di tutte le professioni del mondo. Gio. Battista Somasco, Venice.

Gelbart, N.R., 1987. Feminine and Opposition Journalism in Old Regime France: Le Journal des dames. University of California Press, Berkeley.

Habermas, J., 1975. Legitimation Crisis (T. McCarthy, Trans.). Beacon Press, Boston.

Habermas, J., 1981. The Theory of Communicative Action, 2 vols. (T.A. McCarthy, Trans.). Beacon Press, Boston, MA.

Habermas, J., [1962]1989. The Structural Transformation of the Public Sphere (T. Burger, F. Lawrence, Trans.). M.I.T. Press, Cambridge, MA.

Infelise, M., 2005. Prima dei giornali. Alle origini della pubblica informazione. Laterza, Bari.

Infelise, M., 2009. I libri proibiti da Gutenberg all'Encyclopédie. Laterza, Bari.

Innis, H.A., 1950. Empire and Communications. University of Toronto Press, Toronto.

Johns, A., 1998. The Nature of the Book: Print and Knowledge in the Making. University of Chicago Press, Chicago.

Landes, D.S., 1969. The Unbound Prometheus: Technological Change and Industrial Development in Western Europe from 1750 to the Present. Cambridge University Press, Cambridge.

Landes, D.S., 1998. The Wealth and Poverty of Nations: Why Some Are So Rich and Some So Poor. W.W. Norton, New York.

Lasswell, H.D., 1948. Attention structure and social structure. In: Bryson, L. (Ed.), The Communication of Ideas. Harper, New York, pp. 247–248.

Lasswell, H.D., 1971. Propaganda Technique in World War I. M.I.T. Press, Cambridge, MA.

Lazarsfeld, P.F., Merton, R.K., 1948. Mass communication, popular taste, and organized social action. In: Bryson, L. (Ed.), The Communication of Ideas. Harper, New York, pp. 95–118.

Ludes, P., 2001. Multimedia und Multi-Moderne: Schlüsselbilder: Fernsehnachrichten und World Wide Web – Medienzivilisierung in der Europäischen Währungsunion. VS Verlag für Sozialwissenschaften, Wiesbaden.

Luhmann, N., 2000. The Reality of the Mass Media (K. Cross, Trans.). Stanford University Press, Stanford.

Luhmann, N., 2012. Theory of Society, 2 vols. (R. Barrett, Trans.). Stanford University Press, Stanford.

Macintosh, A., 2006. Characterizing e-participation in policy-making. In: Proceedings of the Thirty-seventh Annual Hawaii International Conference on System Sciences (HICSS-37); Big Island, Hawaii, January 5th–8th. IEEE Computer Soc., Los Alamitos, CA.

Magalhães, S.I., Cerqueira, C., Bernardo, M., 2012. Media and the (im)permeability of public sphere to gender. In: Nunes da Costa, M. (Ed.), Democracia. Mass Media e Esfera Pública. Democracy, Mass Media and Public Sphere. Edições Húmus, Vila Nova de Famalicão, pp. 35–52.
Maggiolo, L., 1880-83. Pouillé scolaire, ou Inventaire des écoles dans la paroisse et annexes de l'ancien diocèse de…: avant 1789, de 1789 à 1833, 3 vols. Berger-Levrault, Nancy.
Mayring, P., 2000. Qualitative Inhaltsanalyse. Grundlagen und Techniken, seventh ed. Deutscher Studien Verlag, Weinheim.
McEnery, T., Hardie, A., 2011. Corpus Linguistics. Method, Theory and Practice. Cambridge University Press, Cambridge.
McLuhan, M., 1962. The Gutenberg Galaxy: The Making of Typographic Man. Signet Books, New York.
McLuhan, M., 1964. Understanding Media: The Extensions of Man. Signet Books, New York.
McLuhan, M., 1967. The Medium Is the Massage: An Inventory of Effects. Bantam Books, New York.
Merton, R.K., 1946. In: Lazarsfeld, P. (Ed.), Mass Persuasion. Free Press, New York.
Mokyr, J., 1990. The Lever of Riches: Technological Creativity and Economic Progress. Oxford University Press, Oxford.
Morin, E., 1977–80. La methode, 3 vols. Paris.
Mornet, D., 1933. Les origines intellectuelles de la Révolution française, (1715–1787). A. Colin, Paris.
Mumford, L., 1934. Technics and Civilization. Harcourt, Brace and Co., New York.
Ong, W.J., 2002. Orality and Literacy: The Technologizing of the Word, second ed. Routledge, New York.
Pettegree, A., 2010. The Book in the Renaissance. Yale University Press, New Haven.
Rascaroli, L., 2009. The Personal Camera: Subjective Cinema and the Essay Film. Wallflower Press, New York.
Rawls, J., 1993. Political Liberalism. Columbia University Press, New York.
Raymond, J., 1996. The Invention of the Newspaper: English Newsbooks, 1641–1649. Oxford University Press, Oxford.
Reinders, M., 2010. Gedrukte chaos. Populisme en moord in het Rampjaar 1672 (Amsterdam, 2010).
Rulyova, N., Zagibalov, T., 2012. Blogging the other: construction of national identities in the blogosphere. Europe-Asia Studies 64 (8), 1524–1545.
Scannell, P., 2009. The dialectic of time and television. Annals of the American Academy of Political and Social Science 625, 219–235.
Stone, P.J., et al., 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press, Cambridge, MA.
Valdeón, R.A., 2007. Political and sexist bias in news translation: two case studies. Trans 11, 231–243.
Vico, G.B., 1725. Principi di una scienza nuova. F. Mosca, Naples.
Wilke, J., 1999. Mediengeschichte der Bundesrepublik Deutschland. Böhlau, Köln.
Williams, R., 1974. Television: Technology and Cultural Form. Schocken Books, New York.

Media and Social Movements

Kimberley Voss, Nicholson School of Communication, University of Central Florida, Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W.A. Gamson, volume 14, pp. 9468–9472, © 2001, Elsevier Ltd.

Abstract

Contemporary social movements engage in a complicated and uneasy dance with the mass media. Historically, movements have typically had their own media, controlled by movement participants and directed at the people they hoped to rally to their cause. Where mass media were once peripheral to understanding movements, in the last half-century, with the rise of television, they have become central. It is no longer possible to understand the actions of contemporary movements without understanding the nature of the movement–mass media transaction.

This article introduces the reader to the critical role of mass media for social movements as sites of contest and indicator of broader cultural changes in the civil society. The article describes the role of mass media in the social construction of meaning that shapes language use and political consciousness of individuals in society. The concept of media structural opportunities is presented and discussed, and the strategic implications of the constant interaction of social movements with media are discussed.

Contemporary social movements engage in a complicated and uneasy dance with the mass media. Historically, movements have typically had their own media, controlled by movement participants, and directed at the people they hoped to rally to their cause. Where mass media were once peripheral to understanding movements, in the last half-century, with the rise of television, they have become central. It is no longer possible to understand the actions of contemporary movements without understanding the nature of the movement–mass media transaction.

Social movements are fields of collective actors maintaining a sustained challenge to institutional practices, policies, and power arrangements and/or to cultural codes by mobilizing people for collective action, at least some of which is contentious. Many of the above terms require some brief elaboration. Collective actors (generally called social movement organizations) are typically formal with designated officers but may also include more informal advocacy networks, often grouped around some publication or think tank. The field of actors has some means of sustaining the challenge over time; a one-time march or rally does not make a social movement. Challenges may have a strong cultural component as well as representing efforts to change institutions and policies. In democracies, there are institutionalized and prescribed ways of citizens exercising influence on policy using the legal system, electoral system, and the peaceful petitioning of public officials ('lobbying'). Unless at least one actor in the relevant field is engaged in other, more contentious forms of collective action – be it rallies, demonstrations, publicity campaigns aimed at embarrassing opponents, boycotts, strikes, violence, or various other types of unruly behavior – it does not qualify as a social movement. The role of social media has changed the role of social movements. It has changed the hierarchy of the media landscape as citizen journalists have emerged and stoked the fires of a social movement. Most notable in the intersection of social media and social movements was the Arab Spring. Facebook and Twitter became as influential as the mainstream media.

Role of the Mass Media

Social movement organizations typically make use of direct media in the form of newsletters to members and, increasingly, e-mail. In addition, there are often alternative media read by many of their constituents that report sympathetically on their activities. However, they cannot afford to ignore general audience mass media for several reasons. First, the mass media provide a master forum in the sense that the players in every other forum also use the mass media, either as participants or as part of the gallery. Among the various forums of public discourse, the mass media provide the most generally available and shared set of cultural tools. Social movement organizations must assume that their own constituents are part of the mass media gallery and the messages their would-be supporters hear cannot be ignored, no matter how extensive the movement's own alternative media may be.

Second, mass media forums are *the* major site of contest politically in part because all of the potential or actual sponsors of meaning – be they authorities, members, or challengers – *assume* pervasive influence (whether justified or not). The mass media often become the critical gallery for discourse carried on in other forums, with success measured by whether a speech in a legislative body or other venue is featured prominently in elite newspapers or in television news coverage.

Finally, the mass media forum is not simply a site of contest and indicator of broader cultural changes in the civil society. The meanings constructed at this site influence changes in language use and political consciousness in the workplace and other settings in which people go about the public part of their daily lives. When a cultural code is being challenged successfully, changes in the media forum both signal and spread the change. To have one's preferred understanding of a policy issue increase significantly in mass media forums is both an important outcome in itself and carries a strong promise of a ripple effect.

It is useful to think of mass media output in terms of the relative prominence of competing *frames*. A frame is a central organizing idea for making sense of relevant events, suggesting

what is at issue. 'Media frames,' Gitlin (1980, p. 7) writes, "largely unspoken and unacknowledged, organize the world both for journalists who report it and, in some important degree, for us who rely on their reports." Movement organizations compete with others as sponsors of frames, attempting to promote the career of their favorite through a variety of tangible activities – speech-making, pamphlet writing, demonstrations, advertising, interviews with journalists, and the like.

Media Opportunity Structure

Social movement organizations must compete with other sponsors of frames – public officials, political parties, corporations, vested interest groups, and journalists themselves. Mass media norms and practices can sometimes facilitate movement success in this contest but on balance tend to present a series of obstacles that put them at a relative disadvantage. This media opportunity structure is one subpart of the larger political opportunity structure in which movements must operate in attempting to shape an effective symbolic strategy.

Movement Opportunities

Social movements often make good copy for the media. They provide drama conflict, and action, colorful copy, and photo opportunities. Because visual material puts a higher premium on spectacle, television is more likely than print media to emphasize it. Spectacle means drama and confrontation, emotional events involving people who have fire in the belly, who are extravagant and unpredictable. This puts a high premium on novelty, on costume, and on confrontation.

Violent action has most of these media-valued elements. Fire in the belly is fine but fire on the ground photographs better. Burning buildings and tires make better television than peaceful vigils and orderly marches. Visual spectacle is high in entertainment value. When news is a vehicle for procuring an audience to sell to advertisers, then one needs to worry about people tuning out. The media opportunity structure provides an incentive for action strategies that provide strong visual material.

Movement Constraints

Social movement organizations typically need to overcome four major obstacles in media norms and practices. The focus here is on the American news media but to the extent that the news media in other countries are similar, the same obstacles will be present. First, journalists have a strong tendency to give official frames the benefit of the doubt. In many cases, the assumptions of these officials are taken for granted; but even when they are challenged by social movement organizations, it is these challengers who carry the burden of proof. A weaker form of this argument is that journalists make official frames the starting point for discussing an issue.

Various observers have noted how subtly and unconsciously this process operates. Halberstam (1979, p. 414) describes how former television news anchor Walter Cronkite's concern with avoiding controversy led to his acceptance of the assumptions underlying official packages. "To him, editorializing was going against the government. He had little awareness, nor did his employers want him to, of the editorializing which he did automatically by unconsciously going along with the government's position."

A second disadvantage for social movement organizations stems from the daily news routines of journalists. These news routines bring reporters into regular – sometimes even daily – contact with sources representing public agencies, political parties, corporations, and other well-established groups. They rarely include regular contact with groups who are challenging the dominant interpretations of events. Social movement organizations, then, lack the routine access to journalists that are characteristic of their rivals.

Third, other media norms and practices in the United States – particularly the *balance norm* – favor certain sponsors of alternative frames. In news accounts, interpretation is generally provided through quotations and balance is provided by quoting spokespersons with competing views. The balance norm is vague and the practices it gives rise to favor certain spokespersons over others. Organized opposition is a necessary condition for activating the norm. Once invoked, there is a strong tendency to reduce controversy to two competing positions – an official one and the alternative (if there is one) sponsored by the most vested member of the polity. In many cases, the critics may share the same unstated common frame as officials, differing only on secondary issues of the most effective means of achieving particular outcomes.

The balance norm, however, is rarely interpreted to include challenger frames, even when no other alternative is available. Tuchman (1978, p. 112) argues that balance in television news in the United States 'means in practice that Republicans may rebut Democrats and vice-versa' but that 'supposedly illegitimate challengers' are rarely offered the opportunity to criticize government statements. Instead, she suggests, "reporters search for an 'establishment critic' or for a 'responsible spokesman' whom they have themselves created or promoted to a position of prominence."

Finally, there is a culture clash between the more pragmatic and cynical subculture of American journalists and the more idealistic and righteous subculture typical of movement organizations (see Eliasoph, 1998). Movements seem to demand unreasonable and unrealistic things and often have a righteousness that is unappealing to those who are living with the inevitable compromises of daily life. Movements hector people and call them to account. This means that internal movement conflicts and peccadilloes will have a special fascination for journalists, giving them an opportunity to even the score from their standpoint. As Gamson and Wolfsfeld (1993, p. 120) put it, "The fall of the righteous is a favored media story wherever it can be found, and movements offer a happy hunting ground."

Strategic Dilemmas of Movements

These characteristics of the media opportunity structure create a series of dilemmas for the symbolic strategies of social movement organizations. Movement–media communication is, to quote Gamson and Wolfsfeld (1993, p. 119), "like

a conversation between a monolingual and a bilingual speaker." The media speak mainstreamese, and movements are tempted to adopt this language since journalists are prone to misunderstand or never hear the alternate language and its underlying ideas. Movement activists often feel that something has been lost in translation. One can accept the dominant cultural codes and not challenge what is taken for granted, but for many movements this would surrender fundamental aspects of their *raison d'e tre*.

There is, then, a fundamental ambivalence and, for some, an estrangement between movements and media. Movement activists tend to view mainstream media not as autonomous and neutral actors but as agents and handmaidens of dominant groups. The media carry the cultural codes being challenged, maintaining and reproducing them. In this sense, they are a target as much as a medium of communication. But they are also the latter since one tries to speak *through* the media rather than to them. This dual media role is the central problematic from the movement standpoint and gives rise to the following dilemmas.

Publicity as Validation

When demonstrators chant, 'the whole world is watching,' it means that they matter, that they are making history. The media spotlight validates the fact that the movement is an important player. Conversely, a demonstration with no media coverage at all is a nonevent, unlikely to have any positive influence either on mobilizing followers or changing the target. No news is bad news.

Given the lack of routine access, getting attention is a necessary step in promoting a preferred frame. But the very tactics that are employed to garner this attention can detract from the preferred frame one is sponsoring. Members of the club enter the media forum through the front door when they choose, are treated with respect, and given the benefit of the doubt. Movement organizations must find some gimmick or act of disorder to force their way in. But entering in this manner, they risk being framed as crazies or fanatics and the preferred frame they are promoting may be obscured in the process. "Those who dress up in costume to be admitted to the media's party will not be allowed to change before being photographed," write Gamson and Wolfsfeld (1993, p. 122).

Weak Control Over Followers

Movement organizations face the additional dilemma that, even if they engage in strategic planning, their weak control over the actions of their followers may make it impossible to implement their plan. The media influence internal movement leadership by certifying some people or groups and ignoring others. Some media-designated leaders may have had few followers before their anointment but with their media-generated celebrity, they soon find followers. As Gitlin (1980) argues, it is precisely those leaders who are attached to followers only through their media image and are unaccountable to the rank and file who are likely to advocate the extravagant and dramatic actions that generate good media copy. This symbiotic relationship can undercut the carefully planned media strategies of more accountable leaders.

Measuring Success

A movement strategy can be judged successful if it solves two problems: gaining *media standing* and increasing the *relative prominence of its preferred frame* in mass media discourse.

Media Standing

In legal discourse, standing refers to the right of a person or group to challenge in a judicial forum the conduct of another, especially with regard to governmental conduct. The rules for according legal standing have been anything but fixed and clear; rather than a matter of clear definition, legal standing is a battle ground.

By analogy, media standing is also contested terrain. In news accounts, it refers to gaining the status of a regular media source whose interpretations are directly quoted. Note that standing is not the same as being covered or mentioned in the news; a group may be in the news in the sense that it is described or criticized but has no opportunity to provide interpretation and meaning to the events in which it is involved. Standing refers to a group being treated as an agent, not merely as an object being discussed by others.

From the standpoint of most journalists who are attempting to be 'objective,' the granting of standing is anything but arbitrary. Sources are selected, in this view, because they speak as or for serious players in any given policy domain: individuals or groups who have enough political power to make a potential difference in what happens. Standing is a measure of achieved cultural power.

Journalists operating in a news forum try to reflect their perceptions of who the key players are, but in practice, they are influenced by various other factors in choosing sources and quotes. In cultural contests, sources are often chosen because they are seen as representing a particular perspective. Rather than being seen as representative in the typical sense, they are chosen as prototypes who represent a particular cultural tendency in a compelling and dramatic way. In this sense, standing still reflects a journalistic political judgment about which cultural movements make a difference or are serious players.

Frame Prominence

If one charts mass media coverage of some issue domain over time, frames and their associated idea elements and symbols will ebb and flow in prominence. Success in gaining new advantages in cultural terms is measured by changes in the relative prominence of the movement's preferred frames compared to antagonistic or rival frames. This measure of success can be extended to other non-news forums as well. Frames can be extracted from cartoons, films, advertising, and entertainment as readily as from news accounts. The prominence of preferred movement frames can be assessed over time in such forums in the same way as in news forums.

Using these two measures, four media outcomes can be defined for movements. *Full response* means that the movement organization receives both media standing and a significant increase in the prominence of its preferred frame. *Collapse* means it receives neither standing nor increased prominence

for its preferred frame. *Co-optation* means that the group receives media standing but no significant increase in its preferred frame. Finally, *preemption* means that the challenger's preferred frame has significantly increased in media prominence in spite of the absence of media standing for its sponsor.

See also: Information Society; Mass Communication: Normative Frameworks; Mass Media and Cultural Identity; Media Ethics; Media Imperialism; Media and History; Political Communication; Popular Culture; Social Movements, History of: General.

Bibliography

Eliasoph, N., 1998. Avoiding Politics. Cambridge University Press, Cambridge, UK.
Epstein, E.J., 1973. News from Nowhere, first ed. Random House, New York.
Gamson, W.A., Wolfsfeld, G., 1993. Movements and media as interacting systems. The Annals of the American Academy of Political and Social Science 528, 114–125.
Gans, H.J., 1979. Deciding What's News. Pantheon Books, New York.
Gitlin, T., 1980. The Whole World Is Watching. University of California Press, Berkeley, CA.
Halberstam, D., 1979. The Powers that Be. Knopf, New York.
Hallin, D.C., Mancini, P., 1984. Political structure and representational form in US and Italian television news. Theory and Society 13, 829–850.
Howard, P.N., Parks, M.R., 2012. Social media and political change: capacity, constraint and consequences. Journal of Communication 2, 359–362.
Iyengar, S., Kinder, D.R., 1987. News that Matters. University of Chicago Press, Chicago, IL.
Lim, M., 2012. Clicks, cabs and coffee houses: social media and oppositional movements in Egypt, 2004–2011. Journal of Communication 13, 231–248.
Morley, D., 1980. The 'Nationwide' Audience. British Film Institute, London.
Paletz, D.L., Entman, R.M., 1981. Media, Power, Politics. The Free Press, New York.

Media, Crime and Crime Waves

Kevin Buckler, University of Texas at Brownsville, Brownsville, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article proceeds by first establishing market model tendencies (over the public sphere model) as constraining factors that predominately dictate all news media decision making. The article then discusses a second constraining factor, the need to develop news content in an easy-to-use and easily understood format and structure, which leads media organizations to develop news themes. The result is a news selection process that facilitates the development of media-generated crime waves focusing typically on highly sensational crimes. The sporadic reporting of sensational crimes enables the development of moral panics.

Turn on the local news, watch for 5–10 min, and chances are high that the viewer will see at least one news item, probably more, that addresses some crime issue or reports on a criminal incident that has occurred. Flip the channel over to the national news, and also watch that for 5–10 min; the chances are also high, although somewhat lower than with local news, that the viewer sees a news item focusing on crime. Change the station yet again, this time to one of the multitude of 24-h news channels – Fox, CNN, MSNBC, and the like – and it is an almost certainty that one topic or more will deal with some matter of crime and criminal justice. If the person doesn't prefer one of the many news options offered by television and, instead, prefers print media – whether the handheld variety or a news item from the Internet – the chances are still incredibly likely that the person will be exposed to news on crime. It is additionally noteworthy that the typical news item is focused on a certain group of crimes: the sensational, and the type that affirms our worst fears about the world we occupy and our worst fears about the people who occupy the world with us.

After (or, for some, before) getting our daily dose of crime news, most of us venture out into the 'real world.' Walk up to any person or group of people, strike up a conversation about crime, and observe where the discussion goes. Ask the person or the group what crime is feared most. Chances are likely that the dialogue is dominated by talk of crimes of violence, stranger-based crime, a celebrity suspect or victim, a heinous crime, or a crime committed against a particularly vulnerable population (women or children), all of which can best be described as statistical anomalies in reference to true crime patterns and trends. Why is crime news is such a focus of news media? More to the point, why is it that sensational crimes are the focus of news media? Furthermore, why are Americans so predisposed to think of the most violent, most sensational types of offenses in discussions about crime?

This article provides insight on these questions. This article first discusses the market and public sphere models of news media production. The article then contends that in the world of crime news, the market model is predominant. Market imperatives heavily influence crime news selection processes. Organizational imperatives such as the need to package news into themes frequently create a false impression that an objective increase in a certain (marketable) offense is occurring, when, in fact, the only objective increase is in media reporting of the event type. Market imperatives coupled with the organizational realities of news production often create a sense that some individual, group of individuals, condition, or episode is a threat to societal interests (a moral panic) and contributes substantially to a culture of fear.

Market Demand versus the Public Good

A complete conceptual understanding of media behavior regarding crime coverage necessitates a discussion of media within a larger sociopolitical framework. Several notable scholars have argued that in the United States, the language and practices of business have had a pronounced impact on social institutions that, arguably, should be resistant to a dominance of business norms (Currie, 1991; Messner and Rosenfeld, 1994; Polanyi, 1957), and that mass media have not been immune from this trend (McManus, 1994; Postman, 2005). One scholar, in documenting a change in news media in New Zealand toward business-oriented principles, dubbed the shift the 'Americanisation' of news (Atkinson, 1994). Contemporary media producers consistently balance two competing ideas: media as a mechanism to facilitate civic and democratic participation of citizens in social and political discourse, and media as a tool to generate profits for corporate shareholders. David Croteau and William Hoynes (2001) discuss the tension between these two objectives in their analysis of the 'market' and 'public sphere' dichotomy.

In the market model of media, the media organization is viewed as a private company that is, at a basic level, simply engaging in the activity necessary to sell a product. As such, the media organization operating in accordance with the market model views the general public as consumers who have tastes and preferences in terms of the type of news item, the content of the news item, and the style in which it is delivered. A news organization fully implementing the market model is concerned with identifying news stories that are popular among readers or viewers, focusing on the aspects of the news item that are most stimulating to the reader or viewer, and using presentation structures that are engaging and entertaining. What is deemed to be in the interest of the general public is that which is popular among the news organization's 'consumers.' Additionally, it is noteworthy that attempts to regulate the

media industry are viewed by the market model advocate as an interference with normal market processes that should be influenced only by supply and demand.

In contrast, the public sphere model conceptualizes the role of the media organization as a public resource that, while having commitments to shareholders, also is obligated to serve the public. Proponents of the public sphere model suggest that the primary purpose of news media is to engage citizens through information, education, and social integration. With this model, the behavior of media is quite different. Mass media are expected to deliver diverse, substantive content through varied structures, even if the content and structures are not completely popular with the general public. The public sphere model measures success of the mass-media organization not in terms of the level of profits generated by the organization but, instead, based on whether the media company is meeting its obligations to citizens. Advocates of this model view mass-media industry regulation as essential to protect the public interest.

Market Demand Impacts on News Content and Delivery

In considering the current state of news media coverage of crime, it is difficult to obtain evidence of influence of the public sphere model. Indications of the primacy of the market model are in abundant supply (McManus, 1994). There is strong evidence that the market model impacts media decision making in terms of media selection of crime news as important relative to other categories of news items (a between-category effect). More simply stated, news producers value crime news relative to many other types of events as options for news coverage. Content analyses of general media coverage show that news items categorized as crime news occupy an elevated hierarchal position compared to other types of news items (Lipshultz and Hilt, 2002). Also, a wealth of empirical work suggests that if one considers all of the possible criminal events or all of the possible criminal events within a particular crime type that a media organization may choose to cover (e.g., assault, rape, or homicide), certain offenses have aspects that make them more appealing to news media than other criminal events (a within-crime category effect).

While both the between-crime category effect and the within-crime category effect are important and will be discussed in this article, it is the within-crime category effect that merits the most attention. Many of the practical examples of the between-crime category effect are intuitive owing to our collective sense of crime seriousness. But when the focus is on one category of crime, the factors that a journalist uses to distinguish the market worth of one offense over another may be less commonsensical. The most powerful research technique to examine factors associated with assessments of newsworthiness within a particular crime category is to begin with a list of cases and collect news coverage data on each case. Studies utilizing this approach have measured newsworthiness in the following ways: (1) whether coverage appeared at all, (2) the prominence of the coverage (time or space devoted to the case, and where the coverage was placed in the overall news day reporting), and (3) the overall intensity score (accounting for several different indicators of newsworthiness).

Scholars studying the within-crime category effect of the market model have presented arguments and theories to explain why a particular crime news item is deemed more 'newsworthy' than other crime news items. Most of the theorizing about newsworthiness assessments has been in the form of identification of the factors that increase the likelihood of coverage of a particular criminal event. Pat Doyle of the *New York Daily News* (Chermak, 1995) suggested that a good human interest murder case is one that involves an influential or well-known figure as a victim or suspect, a victim who is capable of being presented as an overmatched target in some way, and a set of circumstances that are particularly shocking or brutal, involve multiple victims or offenders, and/or have a narrative that is capable of creating a sense of mystery or drama. Similarly, Steve Chermak (1995) asserted that journalistic assessments of newsworthiness are governed by the violent or heinous nature of the offense, demographic characteristics of the victim and offender, uniqueness of the event, qualities of the incident producers (the news agency and staff), and geographic location of the criminal event.

Sensing the lack of a concise theoretical approach to the study of newsworthiness of criminal events, David Pritchard and Karen Hughes (1997) organized the various factors identified by prior scholarship into a theoretical approach that focuses attention on the degree to which a particular homicide event deviates from the normal homicide. Although their focus is on homicide events, it is appropriate to extend their theory to understanding media coverage of other within-crime category cases. They theorized that the greater the degree of statistical deviance, normative deviance, status deviance, and cultural deviance from the normal homicide case, the more likely that news media will deem the event newsworthy. First, with respect to 'statistical deviance,' Pritchard and Hughes (1997) suggested that homicide events that are abnormal purely from a statistical standpoint are more likely to be deemed newsworthy and garner coverage. Homicides that involve a gangland component, relative to the total number of homicides in any given year in the United States, are rare; therefore, based on the notion of statistical deviance from the norm, these homicides are likely considered newsworthy by journalistic standards. Statistical deviance also has applicability to between-crime category analyses; for instance, because murder or nonnegligent homicide is such a small portion of the overall number of violent crimes committed, a murder is more likely to receive coverage.

Second, Pritchard and Hughes (1997) argued that 'normative deviance' matters in how journalists assess the newsworthiness of a homicide event. Normative deviance, at its most basic level, acknowledges differential treatment of cases in statutory law and suggests that statutory provisions legitimize differences within a given crime category and, in turn, impact journalistic assessments of newsworthiness. A primary statutorily legitimized aspect of the criminal code that sets forth principles of differential treatment concerns *mens rea* and punishment. With respect to homicide, statutory law allows for harsher punishment where the crime is committed knowingly (spousal murder with intent to collect

on life insurance) compared to a homicide resulting from reckless actions (reckless driving leading to another's death). Additionally, harsher criminal sentences are permissible when a crime involves aggravating factors and less harsh sentences are allowable when mitigating factors are present. Because these are differences that have been normalized through codification into the criminal code, those crimes involving situational aspects that are treated as more serious in the criminal code are more likely to be deemed newsworthy by a typical journalist.

Third, Pritchard and Hughes (1997) defined 'status deviance' as the extent to which a person or group of persons is different based on established social benchmarks. Because a wealthy White male, for instance, is of higher social status than a poor Hispanic woman, a crime committed against a wealthy White man is deemed more newsworthy. This is the one class of 'deviance' that is not fully explored by Pritchard and Hughes (1997) in the sense that they fail to completely account for how the social statuses of the victim and the offender interact in newsworthiness assessments. It seems that the gap between the social status of the victim and that of the offender matters tremendously, that is, whether the victim's social status is higher than the offender's or the offender's social status is higher than the victim's. The potential variations of this status deviance are much too numerous to examine here, but it suffices to say that the notion of status deviance is more complex in its form than is examined in the work of Pritchard and Hughes (1997).

Last, Pritchard and Hughes (1997) advanced the concept of 'cultural deviance' as an explanation of journalistic assessments of newsworthiness. They considered a crime to be more culturally deviant when it is considered unclean, unhealthy, or perverted relative to mainstream social norms. Crimes involving a linkage to drugs or crimes committed against a particularly vulnerable population (e.g., an elderly or child victim) are examples of culturally deviant offenses. Although there is a subjective element to the notion of cultural deviance, it is important to consider that Pritchard and Hughes (1997) conceptualized cultural deviance as something to be considered relative to the typical, reasonable person. In essence, any offense where, upon hearing of the circumstances of the crime, the average person experiences a negative emotional state (anger or desire for revenge) can be considered culturally deviant. Pritchard and Hughes (1997) reported data that supported their four-item deviance framework. Their analysis of 100 homicides committed in Milwaukee, Wisconsin, found that homicides involving White persons (status deviance), female victims (statistical deviance), or elderly or young victims (cultural deviance and normative deviance) received the most extensive coverage in the *Milwaukee Sentinel* and the *Milwaukee Journal*.

The work of sociologist Richard J. Lundman (2003) addressed the limitations discussed in this article concerning status deviance. He sought a more comprehensive framework for understanding how the demographic qualities of the victim and offender impact journalistic assessments of newsworthiness. Lundman (2003) argued that race- and gender-based typification (or stereotyping) provides journalists with ready-made and easily understood 'scripts' for potential crime news events. Race- and gender-based stereotypes provide the media professional with a beginning conceptual framework to use in filtering the substantial number of criminal events for possible coverage to determine those events that will receive coverage. The more closely a particular crime is to fitting an existing race- or gender-based stereotype, the more likely that the crime will be deemed newsworthy and marketable.

The rationale of this argument is that stereotypes about race and gender are easily understood in the market audience owing to their commonality and acceptance among the populace. According to Lundman (2003), audiences are likely to reject, as confusing, stories that fail to conform to common gender and race stereotypes. Lundman (2003) pointed to two illustrations of common stereotypes. He suggested that homicides involving an African-American offender and a White victim are easily framed in the context of White fear of minority offenders. He further claimed that crimes that involve a male offender and a female victim are easily presented as either 'male sexism' or an illustration of 'male aggression and female submission.' Empirical studies have supported this expanded status deviance framework (Buckler and Travis, 2005; Lundman, 2003).

Barry Glassner (1999) also identified a theoretical factor (not expressed by the other work cited here) to account for media decisions to report events. He suggested that events that can be framed as having a larger social significance owing to their ability to communicate an underlying social crisis are more likely to be viewed as newsworthy and marketable. Although Glassner (1999) did not expressly write that these events are also ones that can be viewed as universally impacting a large portion of the population, the implication is certainly evident. Glassner's (1999) discussion centered on media reporting of workplace violence as an 'epidemic,' the Halloween candy 'panic,' and sex offenders targeting child victims, to name but a few of the illustrations. The underlying implication is that most people are employed and work in a setting away from home that places them in contact with others and most people have or know children who trick-or-treat or could potentially fall prey to child sex offenders. Glassner (1999) used data to show how these illustrations are largely social crises that are more creations of news media than true threats to personal safety.

News Themes and the Crime Wave

The Market Demand versus the Public Good and Market Demand Impacts on News Content and Delivery sections established that mass news production is heavily oriented in the direction of the market model. This overemphasis on market imperatives creates a climate wherein news organizations overreport crime items that are statistically, culturally, and normatively deviant, and dedicate attention to criminal events that reinforce gender and race-based stereotypes. Furthermore, events that may be framed as incidents that threaten the moral fabric and normal functions of important social institutions are seen as more newsworthy. Most often, however, it is not the reporting of an event, in isolation, that is problematic. After all, it is rather common knowledge that news media have a thirst for the sensational and abnormal

event. The adages "If it bleeds, it leads," and "If a dog bites a man it's not news, but if a man bites a dog, it is news," are general reference points used to describe the tendencies of mass media. Instead, the reporting of isolated events as if they represent a trend or an increased prevalence of some crime type misrepresents reality and misleads the public.

Mark Fishman (1978) referred to this process of reporting isolated events as evidence of a broader social ill as a news theme. A 'news theme' is defined as nothing more than a broad idea or theme that is common to a set of individual news items and that holds together these isolated news items in a package that is easily presentable and understood. He suggested that the organizational demands of news reporting place tremendous pressure on journalists and editors to package the news in a readily usable and easily interpretable format. The concept of a news theme is evident in terms of how a news organization arranges news items in a single day's news cycle, or with respect to how news items are presented over a longer period of time (e.g., a week, a month, or a year).

Fishman's (1978) analysis of news reporting by a network in New York City provides an illustration of a media-generated news theme. He observed that the station covered the following news stories in their reporting of the day's news:

- Police apprehend juveniles who mugged an elderly couple in Queens.
- Police and citizens in Queens meet to discuss crimes against the elderly.
- A feature segment on the Senior Citizens Robbery Unit.
- Police seize guns and drugs intended for warring gangs.
- Two members of a youth gang are arrested for robbery at knifepoint.
- An ROTC cadet is arrested in the stabbing to death of another cadet.
- An audit of NYC finds police have been mishandling funds.
- The NYC police union is working on its contract at the same time that laid-off firemen and subway cops will be rehired.

In this news segment, there are three broad themes communicated to the viewer: crime against the elderly, crime committed by young people, and city budgetary issues. Key to understanding the concept of a news theme is that the probability that any of these individual events is reported increases substantially because the events have the potential to connect to other events that, taken together, are capable of conveying a message concerning a broader issue. Fishman (1978) further asserted that news media reliance on generating news themes creates an unintended consequence that is antithetical to the purpose of journalism: in organizing information in this way, the journalist's search for knowledge produces an outcome of not fully knowing about the isolated events because, in the processing of the packaging, the isolated news items are devoid of true context.

Fishman's (1978) analysis also provided evidence that a news theme developed over the course of one day's news cycle can take on a dynamic life of its own, influencing media reporting for weeks or months. His study showed that the theme of crime committed against the elderly was evident in news reports in New York City for a 7-week period. Prior to this substantial wave of media coverage, the news organizations in New York City had covered crimes against the elderly at a pace of one news item every 2 weeks. The 'crime against the elderly wave' was, however, completely a media-generated phenomenon. Official data were analyzed and showed clearly that the news media's increased attention to the issue did not correspond with any real increase in this type of crime. Thus, many of the perceptions that citizens may experience from their consumption of news are based purely on media behavior, not on any real trend. In this regard, a 'crime wave' is defined by Fishman (1978) as little more than sporadic social awareness of crime. Any particular conclusion drawn concerning crime prevalence or social context that is based on news media depictions is likely most appropriately characterized as "things of the mind" (p. 42).

From Market Tendencies and News Themes to Moral Panic

The organizational constraint on news organizations in the form of market pressures and the need to package news in a format and structure that are easily presentable and understood by citizens have major consequences. These organizational demands have the potential of generating panic among the populace that creates pressures on political leaders to respond. The work of several scholars, including Stanley Cohen (1972), Barry Glassner (1999), and Russell E. Ward (2002), has discussed the processes and intricacies of panic. Cohen (1972) coined the term 'moral panic' to describe a social circumstance where a "condition, episode, person, or group of persons emerges to become defined as a threat to societal values and interests." He also suggested that a moral panic occurs when there is a gap between the actual and perceived threat, and the gap needs to be explained. Ward (2002: p. 466) further elaborated on the idea by adding that moral panic is characterized by "rapid and intense emotional fervor toward an issue that media and other social control agents call to public attention." Glassner's (1999) examination of "The Culture of Fear" extended the understanding of the development of panic to include others that journalists rely on in communicating information to the public. Noting, "Professional narrators play an important role too in transforming something implausible into something believable," Glassner (1999: p. 206) illuminated the importance of credible-sounding experts in giving credence to facts suggesting that fear is justified.

Cohen's (1972) "moral panic" concept was developed from his study of events that occurred in Canton, England, on Easter in 1964. Young adults and adolescents were hanging out on a street in town when a rumor began to circulate that a local bartender had refused to serve some young people. A fight eventually ensued, and several young people raced their scooters along the city streets. The incident culminated in the arrests of around 100 youths by police for very minor charges. The incident received coverage by several prominent newspapers in England, the United States, Canada, Australia, South Africa, and the European continent. The story also led to a news theme that was maintained for about a 3-year period in England. The press coverage used terms and phrases such as 'orgy of destruction,' 'battle,' 'smeared with blood and violence,' and 'screaming mob' to describe the behavior

of the youth involved. The news media distorted the amount of property damage and the number of youths on scooters.

Cohen (1972) analyzed not the behavior of the youth involved but the reaction to the event by media and officials. In developing the concept of moral panic, Cohen (1972) started with the basic presupposition that behavior of an individual or group is defined by media and officials in a sensational and stylized manner to communicate that the behavior represents a threat to existing societal values. The behavior is, in some instances, novel, but, in many cases, it is behavior that has long existed but is receiving more attention. In defining the behavior, the news media convey a very simplistic narrative steeped in stereotypical assumptions. It is at this point that media, officials, political figures, special interest groups, and experts can do and say things that prevent the panic from resonating, or further developing. If these groups, instead, do things to further advance the panic, members of these groups generally begin a process of prescribing diagnoses and recommending solutions. And, often, the diagnoses or solutions are an exaggerated response that is disproportionate to the realities of the event, condition, or episode, or they reduce further the civil rights of certain groups in society who are, quite often, already marginalized.

Gary Potter and Victor Kappeler (2006) argue that the response of the media, the US government, and the public to drugs is one of the most powerful moral panics in the history of the nation. While not specifically referencing the work of Cohen or using the term 'moral panic,' investigative journalist Dan Baum (1996), in his book *Smoke and Mirrors: The War on Drugs and the Politics of Failure*, provides powerful insight supportive of the contention that this constitutes a moral panic. Craig Reinarman and Ceres Duskin (1992) analyzed the *Washington Post* story of, and public reaction to, a fabricated 8-year-old heroin addict named 'Jimmy,' pointing out that the editorial staff of the paper and officials in the Washington, DC, area failed to question some of the most basic assumptions made in the story and, instead, gave credence to a stereotyped version of drug use. Ted Chiricos (1998) documented that the crack cocaine 'epidemic' in the 1990s was a moral panic. He showed that the attention dedicated to cocaine by media occurred at a time when use was actually declining. He attributed much of the 'panic' to the dominant theme of the coverage at the time: that the problem was spreading to previously 'safe' areas and that the issue would inflict children.

Although drug panics have been among the most prominent of moral panics, this is not to downplay the incidence of other moral panics. Scholars have identified a number of moral panics. Kathyrn J. Fox (2012) argued that the interaction between local and national news contributes to the sustainment of a moral panic concerning sex offenders that has transformed into an attack on judicial sentencing discretion. Ronald Weitzer (2007) concluded that many of the claims appearing in news media and made by officials and other experts concerning prostitution are unsubstantiated. Donna Killingbeck (2001) showed that the news media constructed a moral panic focused on school violence. Joel Best and May M. Hutchinson (1996) analyzed the presentation of gang initiation rites in media discourse and showed that media depictions invoke images of conspiratorial deviance to a degree not present in actual gang initiations. Joel Best (1991) presented evidence that news media reporting of highway shootings contributed to a short-lived moral panic.

A potential contributing factor to the origination and maintenance of moral panics is the reliance of news media on official sources of information. Kevin Buckler et al. (2008) have characterized the media and official crime sources as being 'coupled,' meaning that both the media organization (profit motive) and the government agency (public relations motive) have an interest in the coproduction of crime news to the public. These researchers also presented data showing that in two major newspapers – the *New York Times* and the *Washington Post* – quotes from or references to official sources of information appeared much more than quotes from or references to professors and nonacademic researchers with much less of an interest in furtherance of the crime item to the point that a moral panic develops. Additionally, concerning is that other scholars have reported data showing that the type of source quoted or referenced is related to the balance of information or the ideological slant of the story (Beckett, 1995; Welch et al., 1998).

Future Directions and Conclusions

This article has broadly addressed the issue of media, crime, and crime waves. Crime waves are presented as media creations resulting from sporadic coverage of isolated events that journalists perceive as marketable and that are covered as if they are evidence of some underlying trend or social ill. Future research should address several questions that remain unexplored and should examine several of the assumptions that journalists seem to make in selecting news items for coverage. One of these assumptions concerns what journalists perceive to be newsworthy and marketable, especially as it pertains to crime in general as a preferred news item for consumption and the status deviance concept identified by Pritchard and Hughes (1997). No firm body of scholarship has questioned in a systematic and fundamental way the assumptions made by journalists concerning what the public wants and, by extension, what is marketable. In essence, no body of work has questioned, as it pertains to crime coverage, whether the public is consuming the product because it is what is being produced, or whether, given viable options, the public would choose something else. Research has also yet to fully consider how changing US demographic characteristics will impact journalistic assessments with respect to status deviance and interactions between victim and offender characteristics.

Another very important avenue for future research concerns how the new 24/7 media age will impact, on the local, state, national, and global levels, coverage concerning news themes, the generation of crime waves, and the sustainment of crime waves into moral panics. The recent work of Kathyrn J. Fox (2012) is particularly instructive. She discussed how the interaction between local and national coverages sustained a moral panic focused on sex offenders. Local news was reinforced by national coverage and produced a panic that attributed blame to the 'lenient' sentencing practices of

judges. The potential exists that the new media age presents the opportunity for crime news themes to last longer, generating more powerful crime waves and sustainable moral panics.

See also: Crime Trends and Debates; Crime, Sociology of; Crime: Knowledge about and Prevalence; Fear of Crime; Homicide.

Bibliography

Atkinson, J., 1994. The "Americanisation" of one network news. Australasian Journal of American Studies 13, 1–26.

Baum, D., 1996. Smoke and Mirrors: The War on Drugs and the Politics of Failure. Little, Brown, New York.

Beckett, K., 1995. Media depictions of drug abuse: the impact of official sources. Research in Political Sociology 7, 161–182.

Best, J., 1991. Road warriors' on 'hair-trigger highways': cultural resources and the media's construction of the 1987 freeway shootings problem. Sociological Inquiry 61, 327–345.

Best, J., Hutchinson, M.M., 1996. The gang initiation rite as a motif in contemporary crime discourse. Justice Quarterly 13 (3), 383–404.

Buckler, K.G., Griffin, T., Travis, L.F., 2008. Criminologists as 'benchwarmers': the exclusion of eggheads from crime news discourse. Southwest Journal of Criminal Justice 5, 4–21.

Buckler, K.G., Travis, L.F., 2005. Assessing the newsworthiness of homicide events: an analysis of coverage in the *Houston Chronicle*. Journal of Criminal Justice and Popular Culture 15, 1–33.

Chermak, S., 1995. Victims in the News: Crime and the American News Media. Westview Press, Boulder, CO.

Chiricos, T., 1998. The media, moral panics, and the politics of crime control. In: Cole, G.F., Gertz, M.G. (Eds.), The Criminal Justice System: Politics and Policies. Wadsworth, Belmont, CA.

Cohen, S., 1972. Folk Devils and Moral Panics. MacGibbon and Kee, London.

Croteau, D., Hoynes, W., 2001. The Business of Media: Corporate Media and the Public Interest. Pine Forge Press, Thousand Oaks, CA.

Currie, E., 1991. Crime and the market society: from bad to worse in the nineties. Dissent (Spring), 254–259.

Fishman, M., 1978. Crime waves as ideology. Social Problems 25 (5), 531–543.

Fox, K.J., 2012. Incurable sex offenders, lousy judges, and the media: moral panic sustenance in the age of new media. American Journal of Criminal Justice 38, 160–181.

Glassner, B., 1999. The Culture of Fear. Why Americans are Afraid of the Wrong Things: Crime, Drugs, Minorities, Teen Moms, Killer Kids, Mutant Microbes, Plane Crashes, Road Rage, and So Much More. Basic Books, New York.

Killingbeck, D., 2001. The role of television news in the construction of school violence as a 'moral panic'. Journal of Criminal Justice and Popular Culture 8, 186–202.

Lipschultz, J.H., Hilt, M.L., 2002. Crime and Local Television News: Dramatic, Breaking, and Live from the Scene. Lawrence Erlbaum, Mahwah, NJ.

Lundman, R.J., 2003. The newsworthiness and selection bias in news about murder: comparative and relative effects of novelty and race and gender typifications on newspaper coverage of homicide. Sociological Forum 18, 357–386.

McManus, J.H., 1994. Market-driven Journalism. Sage, Thousand Oaks, CA.

Messner, S.F., Rosenfeld, R., 1994. Crime and the American Dream. Wadsworth, Belmont, CA.

Polanyi, K., 1957. The Great Transformation: The Political and Economic Origins of Our Time. Beacon, Boston, MA.

Postman, N., 2005. Amusing Ourselves to Death: Public Discourse in the New Age of Show Business. Penguin Books, London.

Potter, G.W., Kappeler, V.E. (Eds.), 2006. Constructing Crime: Perspectives in Making News and Social Problems. Waveland Press, Long Grove, IL.

Pritchard, D., Hughes, K., 1997. Patterns of deviance in crime news. Journal of Communication 47, 49–67.

Reinarman, C., Duskin, C., 1992. Dominant ideology and drugs in the media. International Journal of Drug Policy 3, 6–15.

Ward, R.E., 2002. Fan violence: social problem or moral panic? Aggression and Violent Behavior 7, 453–475.

Weitzer, R., 2007. The social construction of sex trafficking: ideology and institutionalization of a moral crusade. Politics & Society 35, 447–475.

Welch, M., Fenwick, M., Roberts, M., 1998. State managers, intellectuals, and the media: a content analysis of ideology in experts quotes' in feature newspaper articles on crime. Justice Quarterly 15, 219–241.

Media Effects

Don W Stacks and Zongchao Cathy Li, School of Communication, University of Miami, Coral Gables, FL, USA
Cylor Spaulding, College of Fine Arts & Communication, Towson University, Towson, MD, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E. Katz, volume 14, pp. 9472–9479, © 2001, Elsevier Ltd.

Abstract

This article examines the impact of the media on audiences and society from two perspectives. The *traditional* approach examines how different media outlets (e.g., newspapers, radio, and television) impact on how audiences perceive the information contained in the outlet or how the audiences use a particular outlet through the media. The *contemporary* approach examines how the media are *used by others* to change attitudes and beliefs. Both traditional and contemporary approaches to media effects help to better understand how the media – and those who use it – influence people and society.

The question of whether the media really have an effect on some group of people has been studied for some time. Trying to define what 'media effects' are is not simple in today's increasingly complex media milieu, but typically definitions are focused on either what the media *does* to an audience or *how* that audience uses the media. One of the earliest media effects models was the 'hypodermic' model that posited that the media could affect changes in attitudes, beliefs, and ultimately behavior. The model, however, failed to take into account individual beliefs and uses of media. Later models have focused on how dependent people are on the media for either information or specific uses of the media. We classify these models as 'traditional.'

A second approach to media effect models, with a more persuasive and less passive overtone, has come into play. This phase, which we classify as 'contemporary,' focuses on specific models that help to understand how the audience can be persuaded into changes in attitude, belief, or action. While the traditional models have focused on the *outlets* – newspapers, magazines, radio, television, and now the social media – the contemporary models focus on *mediating factors* in the *outcome* of the media effects (Stacks, 2010; Stacks and Michaelson, 2011).

Traditional Media Effects Models

Agenda Setting

The idea of agenda setting can be traced back in the 1920s when Water Lippmann (1992), journalist and social commentator, argued in his *Public Opinion* that people were not capable of directly experiencing the bigger world, thus had to rely on the images and messages constructed by news media to form perceptions. Bernard Cohen (1963) refined Lippmann's ideas by pointing out that the media do not tell people what to think, but what to think about. Maxwell McCombs and Donald Shaw (1972) first put this idea to empirical test by comparing news media agenda and public agenda during the 1968 US presidential election. Their study found strong correlations between the prominent issues of the news media and the leading public issues. This initial effort established what is known as the agenda setting theory, since which more than 400 empirical investigations have been conducted and published.

Although agenda setting originated as political communication studies, its research scopes and applications extend beyond political settings to a wide range of public issues. Furthermore, researchers have examined the agenda-setting role of news media from the 'attributes' perspective, as compared to the traditional agenda of 'issues.' Often referred to as a second level of agenda setting, the attributes agenda setting emphasizes on the prominent characteristics and traits of objects, which composed and defined the agenda. While the first level of agenda setting is the transmission of object salience, the second level is the transmission of attribute salience (Valenzuela and McCombs, 2009).

Attribute agenda setting converges with framing theory in that they both draw attention to how particular attributes dominate – or frame – the way certain topics are addressed among publics. A media frame is a "central organizing idea for news content that supplies a context and suggests what the issue is through the use of selection, emphasis, exclusion, and elaboration" (Tankard et al., 1991, p. 3). This definition of framing converges with attribute agenda setting by emphasizing how the media picture certain topics through the news contents. Another approach of framing diverges from agenda setting by defining frames as overall principles that become socially accepted over time (Valenzuela and McCombs, 2009). The similarities and differences between agenda setting and framing are among the most discussed and investigated theoretical connections in the communication field.

Meantime, as the evidence for mass media's impact on setting public agenda accumulated, by the 1980s, scholars began to ask the question that "If the press sets the public agenda, who sets the media agenda?" (McCombs, 2005, p. 548) Three elements were prompted as key influencers for media agenda: the traditions and norms of journalism professions, daily interactions among news organizations, and news sources and their interactions with media (McCombs, 2005).

In addition, a large body of research has looked into the consequences of agenda-setting effects on attitudes and behavior. Among all, scholars have identified three distinct consequences: forming opinions, priming opinions by emphasizing on particular issues, and shaping opinions by emphasizing on particular attributes; these consequences further lead to behavior changes.

Cultivation

Cultivation theory examines the long-term effects of television viewing on viewers' conceptions of social reality. Cultivation analysis initiated as part of the Cultural Indicators Project founded by George Gerbner in the late 1960s. The central hypothesis of cultivation analysis is that people who spend more time watching television are more likely to perceive the real world in a way as most commonly depicted in television messages, as compared to those who watch less television, but are otherwise comparable in major demographic features (Gerbner, 1969).

Television is the most pervasive and constant learning system in society. Cultivation analysis approaches television as a message system with aggregate and repetitive patterns of images, the contents of which were absorbed by viewers over long period of time. Different from the then traditional mass communication research, which focused on the effects of individual messages, episodes, series, and genres on immediate attitudes and behavior changes, cultivation analysis emphasizes the general and pervasive outcomes of cumulative television exposure (Signorielli and Morgan, 2009).

Findings of nearly four decades of cultivation research on television message systems have unveiled a 'distorted' reality: exaggerated crime rates, overrepresentation of violence, gender-role stereotyping, nontraditional family composition, the 'mean world syndrome,' and so on (Morgan et al., 2009). This misshaped, yet repetitive represented 'reality,' gradually dominates the worldview of heavy viewers, who rely on television as their main source of beliefs and values.

As a primary storyteller in our society, television not only impacts on individual ideologies, but also influences society as a whole as the fundamental manifestation of the mainstream culture. Among heavy television viewers, the relatively common outlooks and values cultivated by television become the dominant or mainstream culture of that society, despite individual differences. The notion of *mainstreaming* refers to "television viewing may reduce or override differences in perspectives and behavior which stem from other social, cultural, and demographic influences" (Signorielli and Morgan, 2009, p. 113).

Some of the major aspects in cultivation studies have examined television's influences in areas such as violence, gender-role stereotypes, health, family, and most recently, online gaming and psychological health implications (Signorielli and Morgan, 2009).

Spiral of Silence

The Spiral of Silence model was raised by German researcher Elisabeth Noelle-Neumann (1984), through her observations and following studies of public opinion of the 1965 political election in the Federal Republic of Germany. According to Noelle-Neumann, Spiral of Silence is a process that "observations made in one context (the mass media) spread to another and encourage people either to proclaim their views or to swallow them and keep quiet until, in a spiraling process, the one view dominated the public scene and the other disappeared from public awareness as its adherents became mute" (Noelle-Neumann, 1984, p. 5).

Based on the assumption that individuals have opinions, the Spiral of Silence model is also built upon the contention that individuals have an innate fear of social isolation. To publicly express an idea that is unpopular or against social norms is incompatible and risky. As a result, individuals constantly monitor their environment, looking for cues that form the majority opinions (Salmon and Glynn, 2009). Mass media, in this sense, function as the predominant source of cues that construct the 'majority' side of an issue or opinions, through the use of repeated and reinforced messages and images.

While agenda setting theory focuses on media effects on the micro-level, the Spiral of Silence model is more concerned with the macro-level, long-term consequences of the public agenda. Consequently, those who hold the minority viewpoints become more reluctant to express their opinions. Over time, as fewer people speak out the unpopular sentiments, such opinions slowly fade away and eventually disappear from public discussions. The majority opinions, on the other hand, increasingly gain confidence and become more pervasive through society, eventually becoming social norms and policies.

The principles of Spiral of Silence have been massively applied in political campaigns. Political parties or advocacy groups strive to increase through the media their stances on issues to strengthen the perception of confidence. Different strategies are employed to demonstrate the magnitude of support to attract media coverage, which reinforce the impression that the party or advocacy group's viewpoints are dominant and winning.

The Spiral of Silence model, however, is somewhat controversial in that the whole model has never been subjected to a comprehensive empirical test (Salmon and Glynn, 2009). A good part of this may be because that the Spiral of Silence is a slow-going process, which can take up to decades to emerge. Empirical testing of the whole model will be extremely time-consuming and expensive. Yet, researchers are continuing to inquire into this phenomenon by testing through manageable time frames of its effect on society.

Uses and Gratifications

Uses and gratifications theory is based on the assumption that individuals consciously choose media and content to fulfill their specific needs. Compared with other traditional media effects research, which examined the communication process from the communicator's end, uses and gratifications take the perspective of audience use of the media (Windahl, 1981). It offers an alternative framework of media effects studies in that it asserts that individuals differ in their motives, attitudes, and behaviors related to media consumption.

In 1974, Katz, Blumler, and Gurevitch summarized the theoretical foundation of uses and gratifications research as focusing on: "(1) the social and psychological origins of (2) needs, which generate (3) expectations of (4) the mass media or other sources, which lead to (5) differential patterns of media exposure (or engagement in other activities), resulting in (6) need gratifications and (7) other consequences, perhaps mostly unintended ones" (p. 20).

The current view of uses and gratifications is grounded in five assumptions:

1. Communication behavior, including media selection and usage, is goal oriented, purposive, and motivated.
2. Audiences actively initiate the selection and use of media to satisfy needs or desires.
3. Social and psychological factors mediate behavior.
4. Media compete with other communication forms, such as interpersonal communication, to satisfy audiences' needs.
5. People are usually, but not always, more influential than media in this process (Rubin, 2009).

Early uses and gratifications research commenced in the 1940s, when several studies investigated why people used certain media formats, such as listening to radios (Papacharissi, 2009). In the early 1970s, researchers started to identify the motivations of media usage, and how media were used to gratify social and psychological needs (Rubin, 2009). For example, McQuail et al. (1972) observed and categorized that people sought gratifications from watching television for *diversion* (to escape and seek emotional release), *personal relationships* (companionship and social utility), *personal identity* (self-expression, reality exploration, and value fulfillment) and *surveillance* (to acquire news and information).

Extant uses and gratification research has examined media usage and consequences in various formats, including soap operas, news programs, VCR, talk radio, cable TV, magazine, channel surfing, tabloid reading, reality TV, and religious television (Papacharissi, 2009). Rubin (2009) summarizes current research areas, as generally following seven directions: (1) media-use motivations and association with attitudes and behaviors; (2) motivations across media and with new media; (3) social and psychological circumstances of media use; (4) gratifications sought and obtained; (5) variations in motives and exposure in affecting outcomes; (6) theoretical extensions with other communication perspectives; and (7) motivation measurement.

Contemporary Media Effects Models

Traditional models such as agenda-setting have served the academic community well and provided a useful framework for analysis; however, like many models, older ones may not always be able to anticipate newer aspects of media and its effects, and have, in some cases, generated conflicting results. Such changes in the media and people's accessibility to and understanding of it, often result in the need for newer models and means of analyzing media effects. Many of the more traditional means of understanding them may not take into account concepts and variables that scholars now consider to be part of the process of analyzing media effects; thus, a discussion of the Elaboration Likelihood Model (ELM) and the concept of Third-Party Endorsement is important to gain a full understanding of the scholarship surrounding the concept of media effects.

Elaboration Likelihood Model

Among the more prominent and important of the newer theories to explain media effects is the Elaboration Likelihood Model or ELM. Initially proposed by Richard E. Petty and John Cacioppo (1986, 1996), the ELM is principally concerned with explaining the way people process *persuasive* messages and how that causes their attitudes to shift over time. Before the creation of the ELM, much of the research on attitude change had found contradictory results, so the ELM was proposed as a way to explain these conflicts. The model focuses on the processes by which the source, message, and receiver variables influence attitude shifts in people, and it suggests that the change may occur either through what is termed the central or peripheral processing routes (Booth-Butterfield and Welbourne, 2002). By examining these two internal processes in message receivers, the ELM can anticipate and predict variables which impact the effectiveness of message persuasion in the audience. Additionally, the ELM anticipates that people have a motivation and desire to hold correct attitudes but can neither process nor ignore every persuasive argument they encounter (Cacioppo et al., 1986).

The advantage of using the ELM approach to media effects comes from its ability to gauge an audience's general knowledge and experiences and, through some of the more traditional media effects models, explain how the media – where most people perceive it as credible information sources and receive gratifications through its use – impact on society. In addition, adding the knowledge of how the media set the agendas and influence minority voices in a culture, the ELM provides a source-based approach to media effects, most often through advertising and public relations endeavors.

In the ELM, *central route processing* refers to the conscious and purposeful consideration of an argument based on an individual's knowledge and experiences, and the idea that they are generating positive or negative thoughts as a result of the communication (Petty et al., 2009). Part of this process is the consideration as to whether or not the information being communicated is worth considering in the first place. If the information is determined to be worthy, then it will travel through central route processing. As Petty et al. (2009) note, the kind of information judged to be important will depend on both the situation and the person; for example, one person may find a moralistic argument on an issue worth considering, while another might be more likely to think about a legal argument. Any change in attitude as a result of central route processing will be subject to internal scrutiny by the individual, so sound, well-constructed messages should create positive thoughts about the particular issue, resulting in the message being perceived as more persuasive, while the converse is also true (Booth-Butterfield and Welbourne, 2002). Additionally, the ELM anticipates that individuals who engage in extensive thinking about the issue presented to them will have a strong association between their attitudes and behaviors, lending even more importance to the active cognitive processing that accompanies central route processing (Cacioppo et al., 1986).

If information is not processed via the central route, it may be processed in the *peripheral route*. The peripheral route does not subject the information to as much careful consideration and thoughtful analysis as does the central route. In this way, the peripheral route can be considered to be a shortcut in cognitive processing, which people employ if they have little interest or ability in processing issue-relevant messages (Petty et al., 2009). As a result, attitude change based on peripheral route processing will take place because it is either associated positively or

negatively with outside cues that may or may not have any relation to the actual message, not as a result of careful consideration of the message (Petty et al., 1983).

An example of an outside cue would be the environment where the individual received the message – if they were comfortable they might process the message in a more favorable way. Another example would be the use of third-party expert to deliver the message with whom the audience has a strong favorable or unfavorable opinion, which could cause the receiver to make a decision on the message without careful consideration. In addition to using outside cues as a catalyst for decision making, an audience member may not consider all of the arguments presented, and may consider only one or two of six arguments presented, resulting in a less thoughtful decision (Booth-Butterfield and Welbourne, 2002).

Since peripheral process decisions are not based on a careful and comprehensive consideration of arguments, generally, attitudes formed as a result are not as strong as those formed via the central route, making them shorter in duration, easier to change, and less predictive of behavior (Booth-Butterfield and Welbourne, 2002). This idea that the simple cue processes underlying the peripheral process can dissipate over time has given rise to a variety of theorized effects impacting attitude generation, including the 'sleeper effect,' which implies that the influence of a discounting cue given for a message will decrease over time, allowing the message to increase in effectiveness (Petty et al., 2009). In this case, the message would have to be carefully constructed initially and then the discounting cue would be peripherally noted. The idea of peripheral processing causing a sleeper effect is reinforced by applying the differential decay hypothesis, which states that the effect will occur if the impact of the discounting cue dissipates more quickly than the original message's (Pratkanis et al., 1988). However, the order of the source cue and the discounting cue are important, since the discounting cue preceding the source cue will not typically result in a sleeper effect (Pratkanis et al., 1988).

ELM shares many qualities with another model developed around the same time. The *Heuristic-systematic model* (HSM) was originally proposed by Shelley Chaiken (1980, 1987), and it similarly proposes that people process messages in two ways: heuristically and systematically. Information processed systematically will be examined for its relevance to the person's tasks and the validity of the information will be evaluated based on prior knowledge of the issue, while people processing information heuristically will focus on simple cues and rules that associate a position with validity (Todorov et al., 2002). Systematic processing then is more deliberate and conscious, while heuristic processing tends to be more automatic and unconscious (Chaiken et al., 1989). The HSM assumes the primary goal of receivers is "… to assess the validity of the persuasive messages they encounter … and that both heuristic and systematic processing occur in the service of this goal" (Chaiken et al., 1989, p. 214). Though these processing ideas are similar to those of the ELM, the HSM holds that the methods of processing can occur concurrently. Additionally, the HSM anticipates three motivational factors which can impact a person's engagement in processing a message: desire for accuracy in judgment, defense of valid judgments, and the desire to impress others (Todorov et al., 2002). Though both HSM and ELM consider similar information in evaluating messages, the analysis of the motivations and method of process selection differ in each model, and the idea of concurrently running processes can alter the interpretation of the data.

With regard to understanding media effects, both the ELM and HSM can be useful frameworks for understanding how media can persuade audiences. By understanding motivating factors for audience members, media can work to tailor programming and messages for central or peripheral route processing, depending on the outcome they desire. Understanding which techniques are the most effective in persuading audiences, however, requires a significant investment on the part of the message originator to research the motivations and standpoint epistemology of the audience they are reaching; this is true for politicians and others seeking to use media as a persuasive tool as well.

Third-Party Endorsement

Though third-party endorsement is a relatively new concept, it is important to understanding any discussion about media effects. As it relates to media effects, the concept of third-party endorsement refers to the idea that media are supporting or endorsing the stories they report on, thus giving approval to the companies, events, people, or other concepts portrayed in the story. This approach also argues that because the source (e.g., journalist or best friend) is known to be objective and trustworthy, he or she will lend a significantly higher amount of credibility to the idea or product being reported on than if the story was delivered directly by the backer of the product or concept directly, as in an advertisement (Stacks and Michaelson, 2009). Much of the research on third-party endorsement pertains to establishing the driving force behind interpersonal advocacy of an issue, belief, company, product, and so forth leading to behavior change.

The perceived endorsement of a product, company, issue, or event by the press is really an implied endorsement, since there is rarely an explicit promotion of a product, company, or cause in a news media story. Implied endorsement refers to the idea that a source endorses the product in the story simply because he or she is covering it in one case and advocating it from media knowledge in the other case. Key to understanding this idea is the concept of intent, and it can be argued that the selection of a media story demonstrates intent to endorse on the part of the story or advocate, whether consciously or unconsciously (Hallahan, 1999). Even if this were the case, however, Kirk Hallahan (1999) notes that few studies have found large increases in audience support for issues or causes that have been explicitly supported by the media, as editorial pages in newspapers do. Even extensive, favorable product reviews, which would essentially be a positive third-party endorsement, do not increase the number of audience members expressing purchase intent toward a product as much as those exposed to an ad (Wang and Lin, 2011). The researchers, however, note that for the endorsement to be effective, it must cater to that audience's *desired communication needs*, as would be predicted by earlier media effect models.

When discussing media effects and third-party endorsement, source credibility is crucially important. In traditional mass media research, journalistic credibility is often cited as the main reason why third-party endorsement is successful in

driving interpersonal advocacy and audience behavior. Studies have found that these claims are true in some instances, but that result depends significantly on other variables associated with the message and source. Both Patricia Anderson and Sharon M. Abbott (1985) and Shelley Chaiken and Durairaj Maheswaran (1994), for instance found that the level of involvement audience members had with an issue had a direct impact on how strongly source credibility impacted their attitude toward the issue. Hallahan (1999) posits that what are perceived as third-party endorser effects might, in fact, be misclassified and the result of an audience predisposed to see media reporting favorably, which would make the consequential positive message reception a direct result of source bias. More importantly, Glen T. Cameron (1994) found that despite consciously recognizing the credibility of news organizations, people's memories generally do not retain a distinction between editorial and advertising with any great accuracy. Moderating this effect, though, is the finding that retention of information varies by media source (Cameron, 1994).

Given the variable impact that memory can have on message content, here, again, the idea of the differential decay hypothesis and sleeper effect is important. To understand the influence of third-party endorsement on persuasion and attitude, message retention is a key element. A low credibility source experiences less decay over time than a high credibility one does, provided the high credibility source had less initial impact (Cook et al., 1979). However minimally significant the impact of third-party endorsement is in persuading a general audience, the impact of third-party endorsement by media to a specific, engaged audience should not be overlooked. Some specific audiences may be more apt to be influenced by certain types of media coverage (e.g., product reviews) than others, so drawing overall conclusions as to the influence of third-party endorsement may be misleading, as it can be considerably more influential on specific audiences.

New Media

New media have brought changes and challenges to traditional media effects theories and models. Broadcast television and newspapers are no longer the only information source to which the public has access. Digital video devices, hand-held mobile wireless devices, cable and satellite TV, and Internet have increasingly become accessible and dominant in American families. The public has more choices of what to watch and read and the themes and cultures are presented by various media formats in a more diverse way as well. In addition, the trend of globalization also enables different cultures and values around the world to be heard and respected. All these new developments pose challenges to the cultivation effects of television. Although issues such as sex and violence, stereotypes, and mainstreaming are still predominant in the mass media, the cultivation effects may be less dominant as people rely more on other information options and become more open-minded.

Moreover, social media could bring revolutionary changes to the traditional media effects theories as well. Researchers have identified various motivations and gratifications for social media consumption as distinct from traditional media use (e.g., Wang et al., 2012; Muntinga et al., 2011). The distinct features of social media, such as user-generated contents, easy access, fast and wide dissemination, allow for exchanging information and expressing diverse opinions. Recent studies have found that social media have impact on the traditional media agenda (e.g., Schiffer, 2006). Eventually, some of the personal agendas widely discussed online may be picked up by the traditional media and become the public agenda. Media agenda in the new media age will be more diverse as well.

In addition, the anonymity and security offered by the new media platform offer minority groups the opportunity to form online communities to express their opinions and seek supports. Their voices are louder as they feel less isolated but more confident. Eventually, as traditional media may pick up contents from the online groups as a news source, the spiral of silence model may not be applicable.

Summary

The impact of the media on people, audiences, societies, and cultures cannot be overstated. Two of the traditional media effects models have focused on the media outlets' ability to set the agenda for discussion or change of attitudes and ultimately behavior. Two traditional models, however, focus on the decisions made by people on how that media is used or how it impacts on their ability to advocate different points of view, often leading to societal changes over time. Contemporary media effects models focus more on the media's capacity to influence through messaging and implied source endorsement of some object, whether it be an issue, a company, or simply an opinion expressed by others. The traditional models provide much insight into what the media do while the contemporary models focus more on how those media effects are obtained.

See also: Mass Media and Cultural Identity; Mass Media and Sports; Media Effects on Children; Religion and New Media; Social Media.

Bibliography

Anderson, P., Abbott, S.M., 1985. Comparing infomercial and commercial on a newspaper cable television channel. In: Houston, M.J., Lutz, R.J. (Eds.), Marketing Communications—theory and Research. American Marketing Association, Chicago, pp. 140–144.

Booth-Butterfield, S., Welbourne, J., 2002. The elaboration likelihood model. In: Dillard, J.P., Pfau, M. (Eds.), The Persuasion Handbook: Developments in Theory and Practice. Sage Publications, Thousand Oaks, pp. 155–173.

Cacioppo, J.T., Petty, R.E., Kao, C.F., Rodriguez, R., 1986. Central and peripheral routes to persuasion: an individual difference perspective. Journal of Personality and Social Psychology 51 (5), 1032–1043.

Cameron, G.T., 1994. Does publicity outperform advertising? An experimental test of the third-party endorsement. Journal of Public Relations Research 6 (3), 185–207.

Chaiken, S., 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. Journal of Personality and Social Psychology 39 (5), 752–766.

Chaiken, S., 1987. The heuristic model of persuasion. In: Zanna, M.P., Olson, J.M., Herman, C.P. (Eds.), Social Influence: The Ontario Symposium. Lawrence Erlbaum, Hillsdale, NJ.

Chaiken, S., Maheswaran, D., 1994. Heuristic processing can bias systematic processing: effects of source credibility, argument ambiguity and task importance on attitude judgment. Journal of Personality and Social Psychology 66 (3), 460–473.

Chaiken, S., Liberman, A., Eagly, A.H., 1989. Heuristic and systematic information processing within and beyond the persuasion context. In: Uleman, J.S., Bargh, J.A. (Eds.), Unintended Thought. Guilford Press, New York, pp. 212–252.

Cohen, B.C., 1963. The Press and Foreign Policy. Princeton University Press, Princeton, NJ.

Cook, T.D., Gruder, C.L., Hennigan, K.M., Flay, B.R., 1979. History of the sleeper effect: some logical pitfalls in accepting the null hypothesis. Psychological Bulletin 86 (4), 662–679.

Gerbner, G., 1969. Dimensions of violence in television drama. In: Baker, R.K., Ball, S.J. (Eds.), Violence in the Media. US Government Printing Office, Washington, DC, pp. 311–340. Staff report to the National Commission on the Causes and Prevention of Violence.

Hallahan, K., 1999. No, Virginia, it's not true what they say about publicity's "implied third-party endorsement" effect. Public Relations Review 25 (3), 331–350.

Katz, E., Blumler, J.G., Gurevitch, M., 1974. Utilization of mass communication by the individual. In: Blumler, J.G., Katz, E. (Eds.), The Uses of Mass Communications: Current Perspectives on Gratifications Research. Sage, Beverly Hills, CA, pp. 19–32.

Lippmann, W., 1992. Public Opinion. Macmillan, New York.

McCombs, M., 2005. A look at agenda-setting: past, present and future. Journalism Studies 6 (4).

McCombs, M., Shaw, D.L., 1972. The agenda-setting function of mass media. Public Opinion Quarterly 36 (2), 176–187.

McQuail, D., Blumler, J.G., Brown, J.R., 1972. The television audience: a revised perspective. In: McQuail, D. (Ed.), Sociology of Mass Communications. Penguin, Middlesex, England, pp. 135–165.

Morgan, M., Shanahan, J., Signorielli, N., 2009. Growing up with television: cultivation processes. In: Bryant, J., Oliver, M.B. (Eds.), Media Effects: Advances in Theory and Research, third ed. Routledge, New York, pp. 34–49.

Muntinga, D.G., Moorman, M., Smit, E.G., 2011. Introducing COBRAs: exploring motivations for brand-related social media use. International Journal of Advertising 30 (1), 13–46.

Noelle-Neumann, E., 1984. The Spiral of Silence: Public Opinion, Our Social Skin. University of Chicago Press, Chicago.

Papacharissi, Z., 2009. Uses and gratifications. In: Stacks, D.W., Salwen, M.B. (Eds.), An Integrated Approach to Communication Theory and Research, second ed. Routledge, New York, pp. 137–152.

Petty, R.E., Cacioppo, J.T., 1986. Communication and Persuasion: Central and Peripheral Routes to Attitude Change. Springer-Verlag, New York.

Petty, R.E., Cacioppo, J.T., 1996. Attitudes and Persuasion: Classic and Contemporary Approaches. Westview Press, Boulder.

Petty, R.E., Briñol, P., Priester, J.R., 2009. Mass media attitude change: implications of the elaboration likelihood model of persuasion. In: Bryant, J., Oliver, M.B. (Eds.), Media Effects: Advances in Theory and Research, third ed. Routledge, New York, pp. 125–164.

Petty, R.E., Cacioppo, J.T., Schumann, D., 1983. Central and peripheral routes to advertising effectiveness: the moderating role of involvement. Journal of Consumer Research 10 (2), 135–146.

Pratkanis, A., Greenwald, A.G., Leippe, M.R., Baumgardner, M.H., 1988. In search of reliable persuasion effects: III. The sleeper effect is dead. Long live the sleeper effect. Journal of Personality and Social Psychology 54 (2), 203–218.

Rubin, A.M., 2009. Uses-and-gratifications perspective on media effects. In: Bryant, J., Oliver, M.B. (Eds.), Media Effects: Advances in Theory and Research, third ed. Routledge, New York, pp. 165–184.

Salmon, C.T., Glynn, C.J., 2009. Spiral of silence: communication and public opinion as social control. In: Stacks, D.W., Salwen, M.B. (Eds.), An integrated approach to communication theory and research, second ed. Routledge, New York, pp. 153–168.

Schiffer, A.J., 2006. Blogswarms and press norms: news coverage of the downing street memo controversy. Journalism and Mass Communication Quarterly 83, 494–510.

Signorielli, N., Morgan, M., 2009. Cultivation analysis: research and practice. In: Stacks, D.W., Salwen, M.B. (Eds.), An Integrated Approach to Communication Theory and Research, second ed. Routledge, New York, pp. 106–136.

Stacks, D.W., 2010. Primer of Public Relations Research, second ed. Guilford, New York.

Stacks, D.W., Michaelson, D., 2009. Exploring the comparative communications effectiveness of advertising and public relations: a replication and extension of prior experiments. Public Relations Journal 3 (3), 1–20.

Stacks, D.W., Michaelson, D., 2011. A Practitioner's Guide to Public Relations Research, Measurement, and Evaluation. Business Expert Press, New York.

Tankard, J.W., Handerson, L., Sillberman, J., Bliss, K., Ghanem, S., 1991. Media Frames: Approaches to Conceptualization and Measurement. Paper Presented at the Annual Convention of the Association for Education in Journalism and Mass Communication, Boston, MA.

Todorov, A., Chaiken, S., Henderson, M., 2002. The heuristic-systematic model of social information processing. In: Dillard, J.P., Pfau, M. (Eds.), The Persuasion Handbook. Sage Publications, Thousand Oaks, pp. 195–212.

Valenzuela, Z., McCombs, M., 2009. The agenda-setting role of the news media. In: Stacks, D.W., Salwen, M.B. (Eds.), An Integrated Approach to Communication Theory and Research, second ed. Routledge, New York, pp. 90–105.

Wang, A., Lin, C., 2011. Effects of content class with endorsement and information relevancy on purchase intention. Management Research Review 34 (4), 417–435.

Wang, Z., Tchernev, J.M., Solloway, T., 2012. A dynamic longitudinal examination of social media use, needs, and gratifications among college students. Computers in Human Behavior 28 (5), 1829–1839.

Windahl, S., 1981. Uses and gratifications at the crossroads. In: Wilhoit, G.C. (Ed.), Mass Communication Review Yearbook II. Sage Publications, Beverly Hills, CA, pp. 174–185.

Media Effects on Children

Marie-Louise Mares and Valerie Kretz, University of Wisconsin–Madison, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S.L. Calvert, volume 14, pp. 9479–9483, © 2001, Elsevier Ltd.

Abstract

Media effects on children as identified by resarchers may broadly be categorized into cognitive, emotional, social, and health-related outcomes. While much of this research focuses on negative outcomes like fear, aggression, materialism, and obesity, media may have positive effects on children as well such as learning and prosocial attitudes. At different ages, children are affected by media in different ways. Likewise, the most successful strategies for mitigating negative outcomes vary with children's ages.

Despite the rapid and dramatic changes in the media landscape over the past 30 years, international studies continue to suggest that children in countries as diverse as India, Indonesia, France, Switzerland, and the United States spend most of their media time with television. Thus, despite the relatively new concerns that have arisen about children's privacy and vulnerability in online interactions, much of the research and concern about media effects on children continues to revolve around a core set of outcomes. These may be categorized into cognitive/educational outcomes, emotional outcomes, social outcomes, and health-related outcomes.

Cognitive Effects of Exposure

Effects of Media on Infant Development

One genre that has emerged since the mid-1990s consists of television or DVD programing targeted to infants. Such material is often focused on language development or preliteracy and prenumeracy skills. A rapidly expanding set of experimental findings indicates that although children under 2 can learn from audiovisual representations, they learn substantially better from live interactions (the 'video deficit'). In various studies, 12-month-olds to 3-year-olds were less able to imitate novel actions seen on a video than the same actions demonstrated (in exactly the same way) by an experimenter. Recent research suggests that young children's learning from video content is improved if there are more demonstrations or longer demonstrations, if the narration is more explicit and detailed, or if familiar, liked characters (such as Elmo, from *Sesame Street*) are used rather than unfamiliar characters.

Studies that examine the outcomes of long-term exposure to educational DVDs or TV programs in the first 2 years find little evidence of positive effects and some evidence of negative relationships between language development and exposure to specific programs or to baby DVDs. One explanation for these findings may be that exposure to children's programing displaces or disrupts play–time activities and parent–child interactions that foster cognitive development. Indeed, studies suggest that such disruptions also take place when children are given electronic versions of toys rather than traditional, nonelectronic versions (e.g., an electronic book vs a board book). Research also points to negative outcomes of having adult-targeted television on in the background – in various studies, children were less focused in their play and parents were less responsive when an adult program was on. Consistent with this, a longitudinal study of 60 infants found that early exposure to background adult television (e.g., at age 1) is associated with poorer performance on measures of learning outcomes (e.g., preliteracy and numeracy) at age 4 (Barr et al., 2010). Overall, research thus far points to little benefit from electronic media use or exposure in the first 2 years of life, although there is considerable evidence of positive effects of hearing stories read aloud by caregivers. For more detail on this literature, see Kirkorian et al. (2008).

Development of Comprehension of Media Content

Even after infancy, children differ from adults in their cognitive development (including working memory capacity and attentional resources), their knowledge of the world, their understanding of how stories typically develop, and in the way they reason, make connections, and form moral judgments. Not surprisingly, research suggests that they differ from adults in the ways that they interpret and remember the stories that are read to them or that they watch on TV or in films. Although, like adults, they remember more central events of a story than peripheral events, they remember far fewer of those central events than adults do even if watching child-targeted programs. Moreover, they are less likely to focus on the characters' goals and motives, and have more difficulty connecting parts of the story to make inferences about why things occur or what will happen next. Making sense of mediated stories is made more complicated by the fact that children are also learning about the nature of each medium (e.g., about the characteristics of different genres of programing or the cues used to signal flashbacks and dreams).

As part of their developing understanding of the nature of TV programing, children change with age in their judgments about the factuality and realism of TV content. In some studies, preschool viewers responded as though they believed that the depicted characters and objects were literally inside the television set, asserting that characters could hear and see the viewers. In other studies, preschool and kindergarten children appeared to regard television as a 'magic window' onto a world where real events were taking place and where subsequent events continued to unfold and characters continue to live their lives even when the program was over. Only gradually, across

the elementary school years, do children become more adult like in their judgments about realism of programs or stories (Wright et al., 1994).

Children also develop in their ability to distinguish ads from programs and their understanding of the purpose of ads. Considerable research suggests that throughout the elementary school years, children show limited understanding of the persuasive intent of advertisers to sell products – the most common response in various studies is that ads are there to provide information about products (see Owen et al., 2007).

Effects of Educational Media Use in Childhood

Despite developmental differences in comprehension, there is substantial evidence of positive effects of exposure to educational television content among 2- to 5-year-olds. Most of the research (both in the United States and elsewhere) has focused on the effects of exposure to *Sesame Street* or the various coproductions of the program which are specifically created for a particular country or region. Overall, the research suggests that exposure to the program is associated with higher scores on literacy, numeracy, and other early learning outcomes, even for children in low- and middle-income regions or in families that are experiencing various stressors such as poverty, conflict, or maternal depression. Longitudinal research conducted in the United States suggests that early viewing (e.g., at ages 2–3) is associated with greater school readiness at age 5. In one follow-up study, early viewing of *Sesame Street* even predicted positive outcomes in high school, presumably because it initiated trajectories of school success and patterns of media preferences that continued to shape learning into adolescence (Anderson et al., 2001). In addition to the research on *Sesame Street*, there is also evidence of early learning from *Blue's Clues* (assessed across a range of early skills such as sequencing, patterning, relational concepts) and *Dora the Explorer* (divergent thinking), both of which have been aired in numerous countries beyond the United States where they were originally created.

There is less evidence of learning from educational television programing targeted toward school-aged children. Although various studies conducted in the United States have documented that children can learn from specific programs (e.g., those that teach science content), research also suggests that as children grow older and start attending school, their parents are less likely to encourage viewing of educational programs (or, indeed, even to regulate the content of exposure) and children are less likely to choose to watch educational programs of their own accord. In part, this may reflect a resistance to more learning after many hours spent in formal education each day. In one study conducted in the United States, 8- to 11-year-olds evaluated a program somewhat more negatively when told that it was educational than without such labeling.

A meta-analysis of 32 studies of learning from interactive computer games or simulations reported that (averaged across preschool, elementary, and high school students) significantly more learning took place when students used interactive media than when they had traditional classroom instruction only (Vogel et al., 2006). However, there are various studies in which there were no additional benefits of interactive media compared with traditional instruction. For example, Rosas et al. (2003) studied over 1200 first- and second-grade students from lower income regions of Chile. They found that those who were assigned to play math and literacy video games daily for 12 weeks in school performed no better than other children in the same school who did not play the games (both groups performed better than children at other, distant schools). Overall, the current evidence points to small, positive effects of educational, interactive media use, but the research base is still small enough that it is unclear what makes some programs more effective than others.

Relationships between Media Use and Academic Outcomes

Analyses of several large data sets measuring students' academic achievement, gathered in multiple countries, generally indicate a negative relationship between amount of television viewing and children's performance on standardized tests. This relationship has been observed across a variety of outcomes, including math, reading, writing, science, and social studies. However, there is also evidence that the relationship is qualified somewhat by other factors.

One key factor is the age of the child. Razel (2001) conducted a meta-analysis of six large studies (one involving standardized testing of students in 12 countries, another with students in 39 countries) providing a total sample of over 1 million students in elementary, middle, and high school. He reported that among the youngest children in the studies (those aged 9 and younger) academic achievement was highest among those who watched approximately 2 h of TV per day – watching less was associated with slightly lower achievement, and watching much more was associated with considerably lower performance. With increasing age, the optimal amount of viewing decreased substantially so that more of the viewing range was associated with decreased performance. In fact, across the various data sets Razel analyzed, the majority of children of all age groups watched more than the optimal amount of TV, and their achievement was lower than those of children who watched less.

A second factor that influences the relationship between amount of viewing and academic achievement is the socioeconomic status (SES) of the child's family, probably because SES is linked to the availability of other education-relevant resources such as books, magazines, tutoring, and other extracurricular activities. Analyses of standardized reading, writing, and math scores gathered in the United States with approximately 500 000 children suggest that the negative relationship between viewing and achievement is stronger for children from high SES families than for children from lower SES families.

Various explanations have been given for the inverse relationship between viewing and achievement. One is that viewing displaces educationally relevant activities such as homework or reading. There is some modest support for this hypothesis. In addition, there is also some evidence that viewing while studying or doing homework (something more probable among heavy viewers) reduces the amount of attention paid to the primary task and thereby reduces the quality of the work. Moreover, there is some evidence that heavy TV viewing is associated with reduced mental effort during reading.

What about academic outcomes and use of other media? In a longitudinal study of two Australian cohorts of children followed from 0- to 5-year-olds and from 4- to 9-year-olds, there was some indication that children's Internet access and frequency of computer use, as well as their frequency of reading print materials, predicted their subsequent vocabulary and language scores. A similar positive association was observed in the United States in a study of approximately 400 third-grade students – those who had access to and used a computer at home scored higher on standardized achievement tests, even after controls for parents' education, children's other media use, and their time spent reading and doing homework.

Emotional Effects of Media Use in Childhood

Children's cognitive and social development affect how they respond emotionally to the images, sounds, and narratives they see and hear. One area of research, conducted primarily by Joanne Cantor and her colleagues, has focused on developmental changes in fear responses to media content.

Fear Responses

The results of numerous studies suggest that young children (aged roughly 3–7) tend to be scared or disturbed by visually grotesque images or scary sounds, such as those often found in fantasy content, even if the actual threat implied by the narrative is relatively minimal. Material that does not contain visual or auditory distortions, such as scenes focusing on a character's anxiety or scenes in which abstract threats are conveyed via words, tend to disturb younger children relatively less. Age differences in fear responses to news stories are consistent with this pattern – in several studies, kindergarten children were much more likely to be frightened by stories of natural disasters (typically shown with scenes of destruction) than by stories of stranger violence (which typically do not have disturbing visual images). In contrast, older children, aged roughly 8–11 are no longer as focused on the perceptually striking characteristics of a situation, can understand more abstract threats, and respond empathically to the emotional displays of the characters. For a review of this research, see Cantor (2011).

Several studies of adults' recollections of media content that frightened them at some earlier point suggest that emotional responses can linger long past the initial exposure – in one study, a quarter of the sample of undergraduate students reported that they still felt residual anxiety about material they had seen as a child or adult, and a subset of those reported being unable or reluctant to engage in activities connected with the events depicted (e.g., swimming in a lake or pool having seen the shark scenes from *Jaws*).

Developmental Changes in Coping Responses

Experimental and survey research suggests that the types of strategies that help young viewers cope with frightening or upsetting material also vary with age. For children under 8, avoidance, distraction, and physical comforts such as hugging a parent or toy or getting a snack appear to be more effective than verbal reassurances that the content is not real or that there is little chance of such events happening to them. Such assurances require the ability to reassess material (a relatively difficult task, particularly under conditions of arousal and cognitive load associated with viewing something frightening) and an understanding of what probabilistic terms such as 'really unlikely' mean. As children grow older, they are more likely to have these abilities and to be able to derive comfort from verbal reassurances. An additional strategy for reducing ongoing negative affect after media exposure involves gradually desensitizing children to images, objects, or sounds related to the frightening material, under safe, nonthreatening conditions.

Future Topics: Children's Emotional Responses

There is much less research on the role of media content in fostering positive affect or enjoyment among young viewers. Research has noted, for example, that preschool viewers are often actively engaged in responding to characters on the screen, but what types of content make children feel happy or amused and to what degree the effects vary by age and by medium remains substantially understudied. Moreover, there is virtually no research on the emotional implications of children's and preteens' online interactions (whether via social networking sites, game sites, or YouTube comments). Research in England by Buckingham and his associates has explored the role of popular media content in shaping children's play and the ways in which children watch and discuss TV programs. Further work, both qualitative and quantitative, is needed to examine still further the emotional outcomes of these moments of connection around media content.

Effects of Media Use on Children's Social and Moral Development

For the past 60 years, the primary focus of concern about children's media use has been whether it makes them more hostile and aggressive and increases their chances of becoming violent adults. A parallel but much smaller body of research has focused on whether, and under what conditions, there may be prosocial outcomes of media use.

Over the years, various theories have sought to explain short- and long-term effects of exposure on viewers' attitudes and behaviors. Although most were developed within the context of media violence research, they also help explain effects of other types of content and predict other outcomes besides physical violence. These theories are reviewed elsewhere (*see* Media Effects). We proceed by briefly reviewing a relatively recent theoretical account – the General Learning Model – which integrates the key elements of earlier accounts. We then consider what evidence there is of effects of violent or prosocial content on young viewers, as predicted by the model.

Theorizing How Effects May Occur

The General Learning Model (Buckley and Anderson, 2006) proposes that media-based experiences contribute to users' knowledge structures, including their person schemata

(i.e., typical characteristics of people or groups of people) and their behavioral scripts (expectations of how people behave in particular situations). Furthermore, the theory proposes that these knowledge structures can contain links to affective states evoked by the initial experience (e.g., anger, fear, warmth) and information about what emotions are typical or appropriate in a given situation. When activated, these interconnected components (person schemata, behavioral scripts, affect, and affective knowledge) may then interact with other factors (personality, values, long-term goals, etc.) to influence people's appraisals and reactions.

How does this apply to young viewers' responses to violent or aggressive content? In the short term, exposure to violent, hostile content (whether in TV programing, in films, in music, or in video games) increases the probability that aggressive thoughts and feelings are activated and tends to increase arousal. These thoughts and feelings, combined with arousal, increase the probability that young viewers will behave aggressively. In the long term, repeated exposure to violence appears to alter viewers' beliefs and attitudes about aggression, decreasing their tendency to notice and respond to real-world pain and suffering and increasing their tendency to interpret social situations in hostile ways. These in turn appear to increase the probability that heavy consumers of media violence will act in hostile or aggressive ways.

Effects of Media Violence on Children's Physical Aggression

Bushman and Huesmann (2006) conducted a meta-analysis of 431 studies related to the effects of media violence, of which 264 focused specifically on children (for a total sample of 50 312 children aged 18 or younger). Studies were included if they assessed the impact of violent content in TV programs, films, video games, music, or comic books. Across this sizable body of research, both experimental and self-selected exposure to media violence were associated with negative outcomes: more aggressive behaviors, thoughts, and feelings, as well reduced helping behavior. The authors also noted that although short-term effects (measured experimentally) were somewhat stronger for adults than for children, long-term relationships between self-selected exposure and subsequent aggression (in some studies, measured years later) were stronger for exposure during childhood than during adulthood.

In one major study (Huesmann et al., 2003), over 400 US 6- to 9-year-olds were surveyed about their exposure to and interpretations of TV violence. Fifteen years later (i.e., when the participants were aged 21–23), their levels of aggression were assessed through several means: (1) interviews with the participant, (2) interviews with a spouse, close friend, or significant other, and (3) by examining criminal records. Childhood exposure to TV violence predicted individuals' physical aggression as adults. These relationships between childhood exposure and adult aggression were stronger for those who, as children, had perceived TV violence to be realistic and who had identified with same-sex aggressive TV heroes and heroines.

Effects of TV Content on Children's Non-Physical Aggression

The Huesman et al. (2003) longitudinal study of media violence also found that among women in the sample childhood exposure to TV violence predicted nonphysical aggressive acts in adulthood (e.g., trying to get someone into trouble). Other studies conducted in England and the United States have also found relationships between children's media use and their tendency to engage in various forms of nonphysical aggression (often referred to as indirect, relational, or social aggression). These types of aggression involve intentionally harming someone's social standing, often by manipulative acts such as gossiping, spreading rumors, or exclusion. In various studies, including a recent longitudinal study by Gentile et al. (2011), preteens or young teens who watched more programs featuring these forms of interpersonal hostility were more likely to engage in those behaviors themselves. There is some indication that such relationships are more likely to be observed for females than for males.

Effects of Media Racial and Gender Stereotyping

Researchers who have documented racial and gender biases in media depictions of reality have long speculated about the implications for children's feelings, beliefs, and attitudes. Content analyses of TV and magazine content in the United States continue to document underrepresentation and stereotyped representations of females and non-Anglo ethnic/racial groups. At the same time, surveys of non-Anglo children in the United States indicate a strong desire to see representations of their own group and an awareness of racial bias in news and entertainment depictions (Children Now, 1999).

Taken together, these patterns may help explain the findings of a longitudinal study conducted by Martins and Harrison (2012) with over 400 children aged 7–12 in the United States. They found that television exposure in second to fourth grade was associated with decreased self-esteem 1 year later among Anglo and African American girls and African American boys (i.e., those groups that tend to be underrepresented and misrepresented on US TV programing). In contrast, television viewing was associated with an increase in self-esteem among Anglo-American boys. These relationships were significant even after controlling for age, body satisfaction, and baseline self-esteem.

Media content may also shape children's broader beliefs about social roles, such as gender roles. Oppliger (2007) conducted a meta-analysis of 31 studies conducted in a variety of countries, examining the relationship between media use (including TV, magazine, and film use) and gender role stereotyping. Measures of stereotyping included judgments about the appropriateness of various occupations for men and women, the appropriateness of playing with sex-typed toys, estimates of the number of men and women in various occupations, and behavioral measures such as the child's own play behaviors, reported performance of gender-normative chores, and imitation of counter-stereotypical role models. Across the subset of studies dealing specifically with children, there was a small but significant relationship between exposure and stereotype-consistent beliefs, attitudes, and behaviors.

Effects of Media Use on Children's Materialism

Research has also examined the effects of exposure to advertising and other media content on children's levels of

materialism (desire for objects, beliefs that personal worth is derived from material possessions). A review of the literature by Buijzen and Valkenburg (2003) found weak to moderate relationships between children's exposure to advertising and their levels of materialism (assessed by responses to standardized items about the importance of material goods) and similar relationships between exposure to advertising and levels of parent–child conflict. However, they did not find any indication that exposure to advertising made children more unhappy. Other research has found relationships between children's exposure to advertising and their recognition of brands, their desire for advertised products, and their requests to parents for the products advertised.

Prosocial Effects of Media Exposure

Although most research has focused on negative outcomes, studies also suggests that media content may be used to foster prosocial attitudes and behaviors among young viewers, encouraging them to interact with others in friendly positive ways. A recent meta-analysis of research on the effects of *Sesame Street* coproductions in 15 non-US countries found a small but significant positive effect of exposure to the program on children's attitudes toward social out-groups, including racial, ethnic, and religious out-groups (Mares and Pan, 2013). Other research suggests that prosocial effects on young audiences are most likely to occur when programs are viewed in a classroom setting where the teacher discusses the prosocial themes and/or encourages related activities. Correlational studies, conducted in the United States and in Europe, examining the relationships between self-selected exposure to prosocial TV content at home and prosocial attitudes and behaviors have tended to find relatively small relationships or no observable relationships at all.

One explanation is that those who watch prosocial television may often watch lots of antisocial content as well, given the prevalence of physical and nonphysical aggression in TV programing. An additional explanation is that young audiences have difficulty understanding the intended lessons and elaborating on them in ways that would lead to positive changes in attitudes and/or behaviors. In one study by Narvaez and her colleagues, children and undergraduates read stories in which the main character faced a moral dilemma (e.g., whether to return extra change), and then chose which of four choices best captured the moral theme of the story. The 8-year-olds in the study chose the right answer only 11% of the time and 10-year-olds only made the correct choice 45% of the time, compared with 91% for the undergraduates. Similar findings emerge in studies of children's comprehension of moral themes in TV or films.

Given these findings, it is not surprising that the most successful interventions involve strategies to lighten children's cognitive load and help them elaborate on the themes. For example, Friedrich and Stein (1973) found no effects of *Mister Rogers' Neighborhood* (a live action program emphasizing friendly interactions) when preschool-aged children watched the program at home. In contrast, there were significant changes among children who watched 12 episodes of *Mister Rogers' Neighborhood* over 4 weeks in preschool, accompanied by related activities. They persisted longer at tasks, were more likely to obey rules, and were more likely to delay gratification without protest, than those who watched other programs.

Effects of Media Use on Children's Health

One relevant theoretical perspective is the Integrated Model of Behavioral Prediction (Fishbein and Ajzen, 2010) which was previously known as the theory of reasoned action. This account proposes that the best predictors of behavior (e.g., smoking, sexual initiation) are the individual's behavioral intentions, and that these in turn are derived from the individual's attitudes toward the behavior, perceived social norms, and self-efficacy about the behavior. Media depictions may help shape all three of these constructs. Indeed, tests of this and related theoretical accounts show that media use is a contributing factor in a number of problematic health outcomes including early initiation of tobacco use, alcohol consumption, and sexual activity. However, the issue that has garnered most attention with regard to young children (rather than adolescents) is the role of media exposure in contributing to children's chances of becoming overweight or obese, particularly as global rates of childhood obesity have been rising.

Media Use and Children's Weight

A study of television advertisements from 23 countries in Europe and Asia found that foods advertised to children were low in nutritional value (e.g., sugary cereals, fast food, candy, snacks, and soft drinks). Similar findings have been reported in the United States. Food products are also advertised in TV programs and movies via product placement, and on websites targeting children that feature product-centered games, known as advergames. As with other forms of advertising, researchers have found that advergames primarily focus on foods that are high in sugar and fat.

Research suggests two primary ways in which media use may contribute to overeating and poor nutritional choices. First, studies indicate that children who are exposed to food advertising are more likely to prefer, request, and choose the advertised (less healthy) foods. Second, children are more likely to snack during and after television use. A third, hypothesized possibility is that media use may displace physical activity. However, studies consistently fail to support this hypothesis (see Marshall et al., 2004, for a meta-analysis).

One line of research has examined whether video games (such as those that involve dancing or simulate sports) may help children become more physically active and maintain a healthy weight. Thus far the research evidence is mixed. However, a study conducted by Maddison et al. (2011) in New Zealand found that using active video games over 6 months had a small but positive impact on body fat and body mass index in overweight and obese children as compared with a control group.

Media Use and Children's Body Image Concerns

Magazine, film, and television content all present images of thin-ideal bodies – even a content analysis of cartoons found that thin characters were associated with more positive traits

than overweight characters (Klein and Shiffman, 2005). Despite the evidence that media use is implicated in rates of childhood obesity, there is also some evidence that media use may contribute to disordered body image and eating, even in childhood. Both cross-sectional and longitudinal research with elementary and middle-school children in the United States has found relationships between children's television viewing and eating disorder symptomatology (e.g., Harrison and Hefner, 2006).

Mitigating Negative Effects of Exposure to Media Content

Given concerns over media effects on children, there have been various attempts to mitigate negative outcomes. At the national level, various countries such as the United States have restricted the amount of advertising during children's programing and required audiovisual separators between programs and ads. In other countries, like Sweden and Belgium, advertising during programing aimed at children aged 12 and younger is banned completely. Another strategy is to require advisories and content labels such as 'parental discretion advised' and 'PG-13: Parents Strongly Cautioned' that would help parents regulate their children's viewing. However, the results of several studies suggest that such advisories often have the unintended 'forbidden fruit' effect of making such content *more* attractive to children.

Other research (much of it conducted in the United States by Nathanson and her colleagues) has focused on the effects of caregiver coviewing, rules, and commentary about controversial media content, such as TV violence and sexual material. In some studies, children in coviewing-only or neutral-commentary conditions were no less likely to engage in aggressive play after viewing violent programing as those who heard praise for the violent actions depicted. In a correlational study, children whose parents regularly watched violent programing with them reported greater (rather than less) likelihood of using physical violence in various situations. One possible explanation is that coviewing without any overt criticism serves as a form of tacit endorsement of the content.

A second type of caregiver behavior, often labeled restrictive mediation, involves setting rules about the types or amount of content the child may watch. In various studies conducted in the United States and in Europe, parents were more likely to report using such strategies with younger children than with teens. Indeed, there is some indication that restriction is effective primarily with younger children and may cause a backlash among adolescents. Moreover, various researchers have noted the difficulty of effectively controlling children's exposure to violent media content given the prevalence and popularity of violence in all media forms.

The third strategy for mitigating negative effects is called active mediation (or instructive mediation or evaluative guidance). This approach involves making comments or asking questions that are intended to reduce negative effects. Several studies suggest that adult criticism of depicted violence may alter children's evaluations of the material or their aggressive dispositions after viewing. However, there is also some indication that the format of the commentary matters. In one study, simple, declarative statements (e.g., That's not how real people behave!) were more effective with 5- to 8-year-olds than questions (e.g., Do you think that's how real people behave?). For older children (9- to 12-year-olds) the question format was more effective than the evaluative statement, perhaps because such questions were less likely to create resentment by threatening the older child's sense of autonomy.

Conclusion

Because children are seen as vulnerable audiences, the focus of research is typically on negative outcomes. As this article indicates, such research generally yields small to moderate significant effects, sometimes enduring years later, as in the longitudinal study of media violence. Nonetheless, it is worth highlighting the evidence of educational and prosocial outcomes as well and the fact that these effects can endure for years too. Academics, producers, and policy makers should use current and future research to maximize positive outcomes and to minimize the more negative outcomes.

See also: Advertising: Effects; Audiences, Media; Computer Games and Academic Achievement; Mass Media, Representations in; Media Effects; Media and Child Development; Public Broadcasting; Violence and Media.

Bibliography

Anderson, Daniel R., Huston, Aletha C., Schmitt Kelly, L., Linebarger Deborah, L., Wright, John C., Larson, Reed, 2001. Early childhood television viewing and adolescent behavior: the recontact study. Monographs of the Society for Research in Child Development.

Barr, Rachel, Lauricella, Alexis, Zack, Elizabeth, Calvert, Sandra L., 2010. Infant and early childhood exposure to adult-directed and child-directed television programming: relations with cognitive skills at age four. Merrill-Palmer Quarterly 56, 21–48.

Buckley, K.E., Anderson, C.A., 2006. A theoretical model of the effects and consequences of playing video games. In: Vorderer, Peter, Bryant, Jennings (Eds.), Playing Video Games – Motives, Responses, and Consequences. LEA, Mahway, NJ, pp. 363–378.

Buijzen, Moniek, Valkenburg, Patti M., 2003. The effects of television advertising on materialism, parent–child conflict, and unhappiness: a review of research. Journal of Applied Developmental Psychology 24, 437–456.

Bushman, Brad J., Huesmann, L. Rowell, 2006. Short-term and long-term effects of violent media on aggression in children and adults. Archives of Pediatric and Adolescent Medical Research 160, 340–352.

Cantor, Joanne, 2011. The media and children's fears, anxieties, and sleep disturbances. In: Singer, Dorothy G., Singer, Jerome L. (Eds.), Handbook of Children and the Media, second ed. SAGE Publications, Thousand Oaks, CA, pp. 215–230.

Children Now, 1999. A Different World: Children's Perceptions of Race and Class in Media. Report downloaded from http://www.childrennow.org/index.php/learn/reports_and_research/article/228.

Fishbein, M., Ajzen, I., 2010. Predicting and Changing Behavior: The Reasoned Action Approach. Psychology Press, New York.

Friedrich, Lynette Kohn, Stein, Aletha Huston, 1973. Aggressive and prosocial television programs and the natural behavior of preschool children. Monographs of the Society for Research in Child Development 151, 1–64.

Gentile, Douglas A., Coyne, Sarah, Walsh, David A., 2011. Media violence, physical aggression, and relational aggression in school age children: a short-term longitudinal study. Aggressive Behavior 37, 193–206.

Harrison, Kristen, Hefner, Veronica, 2006. Media exposure, current and future body ideals, and disordered eating among preadolescent girls: a longitudinal panel study. Journal of Youth and Adolescence 35, 146–156.

Huesmann, L. Rowell, Moise-Titus, Jessica, Podolski, Cheryl-Lynn, Eron, Leonard D., 2003. Longitudinal relations between children's exposure to TV violence and their

aggressive and violent behavior in young adulthood: 1977–1992. Developmental Psychology 39 (2), 201–221.

Kirkorian, Heather L., Wartella, Ellen A., Anderson, Daniel R., 2008. Media and young children's learning. The Future of Children 18, 39–61.

Klein, Hugh, Shiffman, K.S., 2005. Thin is "in" and stout is "out": what animated cartoons tell viewers about body weight. Eating and Weight Disorders 10, 107–116.

Maddison, Ralph, Foley, Louise, Mhurchu, Cliona Ni, et al., 2011. Effects of active video games on body composition: a randomized controlled trial. The American Journal of Clinical Nutrition 94, 156–163.

Mares, Marie-Louise, Pan, Zhongdang, 2013. Effects of *Sesame Street*: a meta-analysis of learning in 15 countries. Journal of Applied Developmental Psychology 34, 140–151.

Marshall, Simon J., Biddle, Stuart J.H., Gorely, Trish, et al., 2004. Relationships between media use, body fatness and physical activity in children and youth: a meta-analysis. International Journal of Obesity 28, 1238–1246.

Martins, Nicole, Harrison, Kristen, 2012. Racial and gender differences in the relationship between children's television use and self-esteem: a longitudinal panel study. Communication Research 39, 338–357.

Narvaez, Darcia, Bentley, Jennifer, Gleason, Tracy, Samuels, Jay, 1998. Moral theme comprehension in third graders, fifth graders, and college students. Reading Psychology 19, 217–241.

Nathanson, Amy, 2004. Factual and evaluative approaches to modifying children's responses to violent television. Journal of Communication 54, 321–336.

Oppliger, Patrice, 2007. Effects of gender stereotyping on socialization. In: Preiss, Raymond W., Mae Gayle, Barbara, Burrell, Nancy, Allen, Mike, Bryant, Jennings (Eds.), Mass Media Effects Research: Advances through Meta-analysis. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 199–214.

Owen, Laura, Auty, Susan, Lewis, Charlie, Berridge, Damon, 2007. Children's understanding of advertising: an investigation using verbal and pictorially cued methods. Infant and Child Development 16, 617–628.

Razel, Micha, 2001. The complex model of television viewing and educational achievement. The Journal of Educational Research 94, 371–379.

Rosas, Ricardo, Nussbaum, Miguel, Cumsille, Patricio, et al., 2003. Beyond Nintendo: design and assessment of educational video games for first and second grade students. Computers & Education 40, 71–94.

Vogel, Jennifer J., Vogel, David S., Cannon-Bowers, Jan, et al., 2006. Computer gaming and interactive simulations for learning: a meta-analysis. Journal of Educational Computing Research 34, 229–243.

Wright, John C., Huston, Aletha C., Reitz, Alice L., Piemyat, Suwatchara, 1994. Young children's perceptions of television reality: determinants and developmental differences. Developmental Psychology 30, 229.

Media Ethics

Adam Drozdek, Duquesne University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Media ethics is a branch of ethics that addresses moral issues arising in connection with the acquisition, preparation, storage, presentation, dissemination, and reception of information through the means of mass media. Mass media include print media (newspapers, magazines, and books), recordings, motion pictures, and electronic media (radio, television, and the computer). Media ethics seeks to help the media practitioners resolve various moral problems arising in all the areas of media communications: journalism, advertising, public relations, and entertainment. The media exercise a strong and complex influence upon the perception and understanding of the world by the public and, consequently, upon shaping the personality of each individual and the interactions of individuals with one another. News and reportage, commercials and advertisements, soap operas, and films – all exert in the long run more or less subtle influence on people's views, choices, and behavior. Because of the ubiquity of the media and their growing presence, the ethical problems that the media practitioners face become increasingly important. The number of publications and conferences on media indicates that among the issues that concern the media people and the public the most important are truth in the media (objectivity and fairness), freedom and responsibility (self-regulatory mechanisms), privacy, and the quality of the media content (violence and pornography).

Media ethics is a branch of applied ethics that addresses moral issues arising in connection with the obtaining, preparation, storage, presentation, dissemination, and reception of information through the means of mass media. Mass media include print media (newspapers, magazines, and books), recordings, motion pictures, and electronic media (radio, television, and the computer); that is, media used to deliver messages to mass audiences through mass distribution of information. Media ethics seeks to help the media practitioners resolve various moral problems arising in all the areas of media communications: journalism, advertising, public relations (PR), and entertainment. It assesses media performance, evaluates standards that guide media practice, and defines best practices.

The media exercise a strong and complex influence upon the perception and understanding of the world by the public and, consequently, upon shaping the personality of each individual and the interactions of individuals with one another. News and reportage, commercials and advertisements, soap operas, and films – all exert in the long run more or less subtle influence on people's views, choices, and behavior. Because of the ubiquity of the media and their growing presence, the ethical problems that the media practitioners face become increasingly important.

There are a number of issues discussed in media ethics that include the truth and consistency of information, fairness, transparency, open access, objectivity, honesty, respect, and tolerance. However, the importance of these issues depends on the function and nature of the media. For example, the problem of objectivity is far more important in journalism than in entertainment or advertisement; moreover, the problem of objectivity is much more affected by the visual media than by print media. Media ethics has strong ties to communication ethics; however, media communication is always mediated, whereby it often becomes a one-way communication with no direct interaction, so that recipients frequently have no control over the type and quality of information they receive through the media. This raises the ethical problem of weighing the self-interests of those who control media against the public interest.

Truth and Objectivity

Traditionally, discussion concerning media ethics concentrated around journalistic ethics for which the problem of truth is of critical importance. Incidentally, the concern with truth along with other reasons were used to consider journalistic ethics as distinct from media ethics (Czarnecki, 2007) and even to propose replacing journalistic ethics with the ethics of individual journalists (Pleszczyński, 2007).

There are two reasons for the importance of truth. First, the media are the primary source of information in a democracy and reliable information is indispensable for a genuine democratic political system. The government must be accountable to the people it represents, and one means of ensuring this is access to information related to governmental policies and informed discussion about government. For this reason, the press is often viewed as the fourth estate that monitors the actions of the three governmental estates: the legislature, executive, and judiciary (Thomas B. Macaulay's concept). Second, the use of inaccurate information contributes to a decline in the autonomy of the individual. Deception is an exercise of self-interest; it makes making choices harder and undermines human freedom.

One facet of truth in journalism is objectivity; that is, the commitment to impartiality, to reflect the world as it is, without any bias or distortion. The journalist should be a neutral, detached observer who reports facts free from personal opinion. There are aspects of each event that are independent of the convictions of reporters. News reports attempt to capture the essence of an event and its underlying causes, and locate the event in a social, historical, and political context. However, perfect objectivity is impossible to achieve. First, there is a problem of selection. News is not just reporting

what recently happened; otherwise the most trivial events would qualify as news. It is the selection of the news based on its importance for the community as reinforced by showing these events in a larger context. The news is, implicitly or explicitly, always contextualized, presented in a particular cultural framework and social perspective. Drug trafficking can be presented as an understandable way to fight poverty, as a means of providing money to buy weapons to fight for a certain cause, or as an inexcusable way of enrichment at the cost of victims' health. The mere choice of terminology can reflect one's opinions; one warring party in a certain conflict can be called soldiers, warriors, combatants, insurgents, fighters, militants, partisans, or regarded as terrorists, guerrilla, rebels, mutineers, renegades, ragtag mob. Also, what is news depends very much on where the reporters and cameras are. In the television era, reporters are influenced by the existence of good pictures and this, in turn, determines how much time is assigned to a story.

Second, journalism striving for objectivity relies on an unrealistic view of the separation of facts and values. Perfect detachment and neutrality becomes an unreachable goal, i.e., reporting facts only ('just the facts') without any personal bias. For this reason, fairness usually becomes the goal as the best approximation of objectivity. Fairness, or balance, is evenhandedness in the treatment of a presented issue; it is the shedding of light on an issue from different angles. Its opposite, unfairness or bias, often means the misleading slanting of news; because it conflicts with the central journalistic duty to inform the public, bias should be avoided. Bias can be the result of faulty thinking (giving too much weight to eyewitness testimonies which are never purely objective; looking for confirmation and not falsification of a view; using unrepresentative samples; committing the *post hoc ergo propter hoc* error), the way of presenting the material (presenting news out of context; using emotionally charged language; focusing on the negative; using close-ups rather than medium-long shots), or the nature of the organization (insufficient time allotted for news presentation). However, it frequently turns out that avoiding bias, presenting material in the most nonpartisan way possible, is simply good for business and extends the scope of viewership or readership, since partisan programs and papers appeal almost exclusively to those sharing the same views. On the other hand, partisan publications, editorials, and opinion columns are expected to be partisan. Frequently, investigative journalists or activist journalists related to some social causes are committed to certain issues and are thus biased. However, this should not excuse them from basing their opinions on verified information, confirmed facts, and a reasoned presentation of their opinions.

Although perfect impartiality is unattainable, making explicit journalists' own prejudices and assumptions can bring them closer to the truth. If journalism did not make seeking the truth untainted by subjectivity its goal, it would severely undermine its credibility and usefulness. By assuming, after the postmodernist thinkers, that truth is just a social construct, the borderline between fact and fiction is blurred, whereby no news could be trusted.

This leads to pragmatic objectivity in which the interpretation of facts is included in reporting; however, interpretations should be tested by the best available journalistic standards (Ward, 2004).

It is often stressed that in the interest of objectivity, journalists should avoid, if possible, conflicts of interest and commitment. Not infrequently, story sources want to receive positive coverage and attempt to influence journalists through gifts, free tickets to events to be reviewed, free travel, etc. Some codes of ethics explicitly prohibit receiving such favors, but for financially struggling organizations such opportunities can be hard to pass over. Also, there may be a conflict of interest when journalists are members of, or sit on the boards of, organizations they cover. Sometimes even the appearance of such a conflict can be as damaging to the credibility of the media outlets as the conflict itself. However, if a conflict cannot be avoided, it should be acknowledged to the public.

The position of truth in the context of advertising is different to that in journalism. The role of advertising is to persuade people to do something: in commercial advertising, to buy a product, or, in political advertising, to accept a person or views. This requires the stressing of the strong points of the product and omitting the weaknesses. Because the only goal of advertisement is to promote a product or idea, it is one-sided and should not be expected to be balanced and complete. The public must be aware of this and treat the advertising accordingly: *caveat emptor*. Also, advertising is a persuasive, not rational, communication. As such, it wants to catch potential client's attention, and is not primarily concerned with the product's quality. But this does not mean that truth is irrelevant for advertising. Advertising should not be deceptive by suggesting nonexisting qualities of a product; it should avoid deceptive means, such as using staged testimonials. Furthermore, an advertisement should be clearly distinguished from a detached presentation of facts. This distinction is violated in infomercials where advertisement is presented as facts so that it may not be clear to the viewer or listener that an infomercial is really a persuasion method presented in a disguised way as an objective fact.

Similar statements can be made about PR, which is the management of communication between organizations and the public; frequently, this is done not only to provide the public with information about organizations, but also to create a positive public image of a particular organization, group, party, or person, often to gain public support to champion a cause. This is done by controlling the flow of information that is presented to the public. Persuading the public, winning it over to a cause, not truthfulness, is the goal. However, when PR managers resort to the use of misleading or manufactured information, their pronouncements can lose credibility in the eyes of the public and have the effect opposite to what was intended. Therefore, although bias is written into the essence of PR, it has to have a strong component of veracity, accuracy, and verifiability of information.

This element is particularly important in political advertising. Politicians must win public support to win an election, so they must have legitimacy obtained by favorable media visibility which stresses the positive and downplays the negative aspects of the politician's views and activity. This often means diverting the public's attention from important political and economic issues to what plays well in the media, particularly on television: polished demeanor, good looks, witty

ripostes. It is claimed that the voters are more interested in the trustworthiness of politicians than in their particular promises. Realization of this problem leads the politicians and their PR advisors to project the best image of these politicians through the media, to shape in public opinion an image of their great personality and character. For this reason, political PR advisors and consultants become increasingly important on the political scene.

Freedom and Responsibility

The freedom to publish opinions is defended with the argument that it is necessary for democracy, and that the discovery of the truth requires a free marketplace of ideas. However, others say, the powerful dominate the media, and consequently the competition of ideas is not genuine because the access to the media is unequal. Also, an argument is used that free speech fosters individual autonomy and development, but, on the other hand, it may encourage hate speech and obscenity.

Unqualified free speech, even in democratic society, is unattainable and it is legally restrained, e.g., in the interest of national security and public safety. There remains a problem of defining the situation when national security is endangered. It is sometimes maintained that the United States lost the war in Vietnam because the media coverage largely undermined the image of the military. Hence, media coverage of the British military operations in the Falklands (1982) and the United States operation in Iraq (1991) was restricted much more severely than necessary.

Most media practitioners realize that – notwithstanding legal issues – free speech must be qualified with responsibility. Journalists in a democracy have a responsibility to foster pluralism and diversity of political debate, and to refrain from perpetuating myths about different social groups. Also, the media must take into account the consequences of their actions. For instance, reporters can sometimes anticipate that their approach to news gathering may induce antisocial behavior: the camera can turn a demonstration into a riot, and a live coverage of a hijacking, which gives the hijackers free publicity, makes hard the release of the hostages. When their stories bring some harm, journalists defend themselves by saying that they just report a story. But news organizations are very willing to take the credit for positive consequences. News is created by journalists, so they need to be sensitive to possible influences their news may have. They need to take into account who may be harmed by their actions. This is particularly important in the context of citizens' privacy (*see* Privacy of Individuals in Social Research: Confidentiality).

Privacy

Journalists justify breaching someone's privacy by saying that the public has the right to know. This is an important argument, because it implies that the media are obligated to publicize some reports, whereas the freedom of the press merely permits such publicizing. However, the argument leaves unspecified what it is that the public has the right to know. Although the media, journalism in particular, ought to act in the public interest, it does not mean that they are obligated to fully satisfy the public's interest. Public interest is not tantamount to mere curiosity. The critics indicate that the media much too often publicize information to satisfy the public's curiosity, but refrain from publishing revelations about the media themselves; journalists routinely avoid stories that put other journalists in a negative light (Hausman, 1992).

The problem of privacy violation is particularly acute in the case of public figures. It is often said that public figures relinquish their right to privacy by the very fact that they become public figures; the loss of privacy is a price for being famous and thereby for high social status, wealth, power, etc. For this reason, all details of private lives of public figures can be published. It is claimed that in the case of politicians such details are important, because their personal qualities can interfere with their public duties and affect their decision making, especially under pressure. Also, the character of the people associated with a politician influences the politician's priorities and decisions. People elect a whole person, not only a public side; consequently, they should know about the private side as well. Hence, notwithstanding the motives of the media, they are acting on the public's behalf as a watchdog of the political arena, and, therefore, they have the right and responsibility to know the facts concerning people who shape political and economic policies of the country. In particular, the media should expose the hypocrisy of politicians who publicize goals and policies that are at variance with their own life. However, others say, in line with La Rochefoucauld's maxim, that hypocrisy is the homage which vice pays to virtue, and, hence, if positive policies are enacted in spite of violation of them by the enactors, so much the better for the policies. For this reason, a hasty exposure of politicians' vices can be counterproductive because it may lead to the demise of the politicians and, consequently, to the demise of their positive and desirable policies. Besides, an invasion of privacy discourages many people from entering an election race, which is detrimental to the democratic process. It is also said that to maintain their mental health, public persons should have time off from the sight of the public.

Privacy is violated by undercover investigation and the use of long-range cameras, electronic eavesdropping devices, or surreptitious recording. The use of such means is often justified as indispensable in the investigation of crime and corruption, public health dangers, and safety negligence because it may be the only available means to unearth the story, and it may contribute to greater objectivity of the story. However, the critics maintain that the use of such means may lead to a general erosion of trust, and discussion of the news obtained by questionable methods often focuses on these methods rather than on the news itself.

Self-Regulatory Mechanisms

The media recognize that some self-regulation is necessary to avoid an erosion of confidence and public demand for governmental regulation. Self-regulation aims at establishing ethical guidelines for professionals to properly address legitimate concerns, respond to complaints, and solve problems. There are several mechanisms used to develop and maintain responsibility of the media.

Codes of ethics mark the stage of maturity of a profession; they are an indication of good faith and thereby project a positive image of an organization or company and enhance its credibility. They also articulate values of this profession, provide a framework for self-regulation, and can help to define and direct the work of professionals by providing ethically relevant framework and setting moral standards. On the other hand, codes of ethics are not infrequently too general, vague, and lack specific guidance for particular situations; also, they often leave open the problem of implementing the enforcement mechanism. However, such codes are generally useful as an educational tool, particularly for novices in the particular profession. They can be used as a defense against doing something that violates them; they can be used by the public to see whether ethical standards are followed by professionals; also, they force professionals to reflect on their obligations when formulating the code.

Another mechanism is internal criticism by a neutral arbitrator within the organization to investigate and to respond to complaints – an ombudsman (Swedish for 'deputy') who can increase an organization's credibility, because the existence of an in-house critic signals to the public its sensitivity to public opinion and its willingness to modify the organization's practices. However, only a small percentage of media organizations have ombudsmen (cf The Organization of News Ombudsmen and its site, newsombudsmen.org).

Moreover, press councils are formed, with variable success, to debate complaints against the press (e.g., British Press Council since 1953, Der Deutsche Presserat since 1956 in Germany, National News Council in the United States in 1973–84; for a list of an ever-growing number of national and local press councils, see www.media-accountability.org and wanewscouncil.org). They are primarily composed of media professionals and their primary task is to deal with complaints concerning the maintenance of ethical standards of the profession.

Finally, education offered in classes on journalism ethics, conferences, meetings, and seminars explores issues of media ethics. This education is important in at least two respects. Not all journalists have majored in journalism and thus they may not have taken classes in professional ethics. Moreover, focusing on ethical issues in some meetings and seminars indicates that the ethical side of exercising journalistic profession is important for a particular media organization or company.

Quality of Media Content

The media practitioners frequently argue that they show and publish what the audience wants, which is determined democratically by conducting polls, and that the criticism of the media for low-quality programs is elitist. This argument is rebutted by the claim that the media respond only to the immediate interests of the public, and what goes beyond these interests is inevitably disregarded. Also, this way of determining interests has a tendency to confine them to the lowest tastes. The pursuit of higher ratings leads to trivialization and sensationalism of journalism ('yellow journalism') and entertainment. The proliferation of tabloid journalism is the result of a conviction that the media only reflect the public's tastes; interests in sensational and bizarre aspects of our culture is natural, and no harm results from satisfying them. However, it is claimed that the constant exposure to tabloid journalism diverts public attention from important social and political problems, and gives credibility to sensational aspects of human existence. Tabloid journalism has become mainstream because sensationalism is one of the most important elements shaping the news, particularly in television, which, by definition, gives rise to distortion. Visual media call attention to conflict, magnitude, the unusual, the sensational, and the spectacular. Television relies on brevity and swift transitions. Visual events (fires, crashes) gain prominent coverage on television; an explanation of a tax reform receives only a short sound bite. The public will eventually come to think that the sensational aspects are the norm.

The reliance on sensationalism is apparent in entertainment. For example, talk shows live on the voyeurism of the public and exhibitionism of the guests. The primary goal of exposing the unusual behavior or appearance of talk show guests is to entertain the public, not to help the guests. The shows break taboos, exhibit deviant behavior, and make it appear ordinary. They blur the line between private life and the public domain by moving the former into the latter with the help of willing participants who may not be entirely aware of, or simply do not care about this shift. These shows also feed on the viewers' willingness to watch others publicly reveal the most private matters even if these viewers would never make such revelations public.

Stressing the sensational manifests itself in the saturation of entertainment with violence and erotic content. Many media critics express a concern that watching violence stimulates violent behavior generally, and copying the watched behavior in particular, people who watch violence regularly become desensitized to its horror and may even develop a positive attitude to it. However, it is possible that pictures of violence increase the level of revulsion concerning its consequence, that it inhibits aggressive behavior, that is, it has a cathartic effect – Aristotle claimed that a tragedy with its depiction of 'serious actions' has a cathartic effect (*Poetics* vi).

The effect on children of watching television is probably the most researched subject today in media influence on people. Results of many research studies indicate that television, especially the depiction of violence, has a profound effect on children's behavior. This is also true for computer games, which have become increasingly saturated with violence. Also, action movies engage viewers' emotions and reduce them to primal instincts, leaving out intellect (Sloterdijk, 1993). However, some say, even granted that it may be difficult to establish empirically the existence of a causal link between viewing and future behavior, many actions of the media are predicated on the existence of such a link. For example, if there were no influence of the media on the public, no organization would invest any money for advertising. Therefore, if noncriminal behavior can be influenced, it is difficult to assume that the media have no impact on criminal behavior. Moreover, constant repetition is both a tool of education for acquiring and holding information, and of propaganda for getting across a political message. It would be remarkable if constant viewing of violence in movies or games did not have a similar effect.

The media often use rating systems to indicate the content of films and shows, but, some contend that this system is often used as an allurement rather than a warning, and the movie producers are tempted to include gratuitous violence, pornography, and profanity in their products to make them more appealing.

The media create popular culture, and in this culture they gravitate toward what is most exciting. Therefore, popularity seems to be a crucial factor in deciding the content of the media. But if popularity is a deciding factor, then there is no room for creativity and acceptance of new ideas. There is little room for real choices because the media become the subject of a form of Gresham's law which says that poor quality entertainment drives out high-quality entertainment, news, and educational programming (Fink, 1988).

The Internet

A distinctive feature of traditional media (press, movies, TV) is the rigid separation of media providers and consumers in which all control of the type and quantity of produced and transmitted information is in the hands of the providers and in which consumers are only passive spectators. The situation drastically changes in the context of the Internet where interaction of the two parties is fairly easily achievable. An online journalist can receive immediate feedback in the form of praise, criticism, correction, additional information, etc. from the reader and this feedback can very quickly appear in an online publication. Including the Internet in the picture can also make traditional media interactive. An online poll can be immediately presented on TV news; questions sent via e-mail can be instantaneously presented to a person interviewed on a TV or radio program.

In the context of such interaction, the roles of information providers and receivers can be blurred and even reversed. In traditional media, a journalist is fully in control of the information provided to the public. This may also be the case in online public/civic journalism. Frequently, however, a site mostly or entirely relies on the information that comes from the public (through regular mail, e-mail, comments posted online), and only the journalist supervising the site may decide which information to post and can also edit it to make it more readable. The journalist is still in control, but in the role of a controller, whereby, in a way, the ethical role of such a journalist-supervisor is even more pronounced than in traditional media: ethical guidelines should decide which information submitted by the public is posted and which should be rejected, if it is e.g., libelous, blasphemous, disrespectful, hurtful, or some other legitimate reason. Such a journalist may also encourage the public to provide some types of information, e.g., information that addresses public health issues (hard news) and not only where to groom a pet (soft news). In this way, civic journalism establishes very close connections between journalists and the public with the intention that the proactive role of the public can influence the political agenda of the community and thus enhance and shape democracy. However, the overarching ethical principles for journalists in the new media are very much the same as in the old: providing truthful information, promoting of the democratic process, respecting for differences, avoiding or at least minimizing harm, and the need for verification of information, even if methods may vary (prepublication verification is required by traditional media, online journalism leans toward postpublication verification).

Another distinctive feature of traditional media is the rigid separation between different types of media. With the Internet, a unification of media is possible. The Internet allows for an easy way of mixing the media. A text can be amplified by still images, audio and video clips, accessible through links added to the text. Links can also lead to other sites and other Web pages with relevant information. Moreover, since TV transmits programs digitally, they can easily be viewed on a laptop, tablet, or even a cell/mobile phone. Newspapers and scientific journals have increasingly made digital version accessible in the cyberspace, and thus the line between newspaper and TV becomes blurred to the point that the prospect of media convergence arises in which the capabilities of various media are merged into one supermedium (cf Phadke, 2008).

There is much more information readily available on the Internet than before the Internet era. This may make abiding to ethical principles more difficult. Because of the sheer amount of information, checking its validity and confirming facts can become an onerous and time-consuming task. Because of time pressure, particularly in cyberspace, some information can be presented without sufficient verification, possibly with a warning, so that the audience should decide how credible the information is. In this, responsibility for the credibility of information is, at least partially, shifted to the audience.

Media User Ethics

Enforcing the ethical standards in the international context of one large digital supermedium is well-nigh impossible, and thus the public should be much more aware of the potential dangers posed by this medium than before. Therefore, the following guidelines have been proposed to direct the consumer in its use. The consumer should (1) be alert and discriminating; (2) oppose censorship; (3) not unduly interfere with editorial independence; (4) guard against sexist and racist stereotyping; (5) look for alternative sources of information; (6) ask for a pluralist supply of information; (7) protect one's own privacy; (8) be a reliable source of information; (9) avoid participation in for-profit journalism; (10) demand accountability of the media (Hamelink, 1995). Such obligations concern the public as participants of mediated communication process in at least three respects: civic duty; economizing one's own schedule; and responsibility for the upbringing of children (Funiok, 2007). This calls for an active participation of the public in shaping the media so that they deliver their content in the way that enhances the wholesomeness of the social environment. Receiving this content is as important as not receiving it, which cuts to the time spent on being a participant of the media communication (watching TV or maintaining one's blog). Finally, parents are responsible for protecting children from potentially harmful yet easily accessible media content.

See also: Advertising: Effects; Freedom of the Press; Mass Media: Introduction and Schools of Thought; Media Effects on Children; Media Effects; Media Imperialism; Media and Child Development; Media and History; Political Advertising; Social Science and the Media; Violence and Media.

Bibliography

Alix, F.X., 1997. Une Éthique pour l'Information: de Gutenberg à Internet. L'Harmattan, Paris.

Christians, C.G., Rotzoll, K.B., Fackler, M., McKee, K.B., Woods, R.H., 2005. Media Ethics: Cases and Moral Reasoning, seventh ed. Pearson, Boston, MA.

Cohen, E.D., Elliott, D., 1997. Journalism Ethics: A Reference Handbook. ABC-Clio, Santa Barbara, CA.

Czarnecki, P., 2007. Etyka Mediów. Difin, Warszawa, Poland.

Day, L.A., 2005. Ethics in Media Communications: Cases and Controversies, fifth ed. Wadsworth, Belmont, CA.

Fink, C.C., 1988. Media Ethics: In the Newsroom and beyond. McGraw-Hill, New York.

Friend, C., Singer, J.B., 2007. Online Journalism Ethics. M.E. Sharpe, Armonk, NY.

Funiok, R., 2007. Medienethik: Verantwortung in der Mediengeselschaft. Kohlhammer, Stuttgart, Germany.

Hamelink, C., 1995. Ethics for media users. European Journal of Communication 10, 497–512.

Hausman, C., 1992. Crisis of Conscience: Perspectives on Journalism Ethics. HarperCollins, New York.

Hulin, A., Smith, J. (Eds.), 2008. The Media Self-Regulation Guidebook: All Questions and Answers. Office of the Representative on Freedom of the Media, Vienna.

Kieran, M. (Ed.), 1998. Media Ethics. Routledge, London.

Limburg, V.E., 1994. Electronic Media Ethics. Focal, Boston, MA.

Phadke, S.P., 2008. Media and Journalism Ethics. ABD Publishers, Jaipur, India.

Pigeat, H., 1997. Médias et Déontologie: Règles du Jeu ou Jeu sans Règles. PUF, Paris.

Pleszczyński, J., 2007. Etyka dziennikarska. Difin, Warszawa, Poland.

Schicha, Ch, Brosda, C. (Eds.), 2010. Handbuch Medienethik. VS Verlag für Sozialwissenschaften, Wiesbaden, Germany.

Sloterdijk, P., 1993. Medien-Zeit: Drei gegenwartsdiagnostische Versuche. Cantz, Stuttgart, Germany.

Ward, S.J.A., 2004. The Invention of Journalism Ethics: The Path to Objectivity and beyond. McGill-Queen's University Press, Montreal, QC.

Wiegerling, K., 1998. Medienethik. Metzler, Stuttgart, Germany.

Media Events

Eric W Rothenbuhler, School of Communications, Webster University, Webster Groves, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Media events' is a concept developed in communication and media studies literature in the 1980s and 1990s, primarily by Daniel Dayan, Elihu Katz, and their associates. It describes large, preplanned social or political events, usually of ceremonial form, often sponsored by the state or other traditional institutions, broadcast live on television and dominating other media, and attracting unusually large and attentive audiences. Inaugurations, state funerals, royal ceremonies, diplomatic visits, and the Olympic Games are typical examples, usually seen as socially integrative ceremonial television. Since the first writing on the topic, other authors have challenged the possibility of ceremony and community by media as well as proposed more attention to conflictive events, social change, historical context, and the relation of these events to the conduct of social power. The recent literature has been dominated by reconceptualizations motivated by more attention to historical context, to globalism, and to the apparent dominance in recent years of conflict, crisis, war, and international terrorism over integration, ceremony, and celebration.

Media events, as the term is used here, are large, planned events that interrupt normal media schedules and audience activities, obtaining special coverage and special attention. State funerals, royal weddings, diplomatic visits, legislative hearings, political inaugurals, and the Olympic Games are representative examples. In these cases an agency, independent of the media, sponsors an event, with a claim to historical importance. The media validate that claim by interrupting their normal schedules for live coverage. The audience delivers the final validation by granting the media event their special attention. When successful, such events have an integrating effect on societies, working as secular ritual. The media event is not only media coverage of something important, but a new social form in which the original event, the media coverage, and the activities of the audience at home unite in a form of dispersed ceremony. This raises a host of questions for analysts about the power of communication, the possibilities for participation at a distance, and the potential misuse of ceremonial forms and audiences, as well as inspiring criticism of the role these events play in reinforcing established institutions and authorities. These questions and concerns have produced vigorous debate and a literature full of proposals for reconceptualization and alternative perspectives on media events and their social roles.

The Paradigmatic Conception

The term 'media event,' of course, is used in a variety of ways. The paradigmatic conception of media events addressed here was developed by Elihu Katz, Daniel Dayan, and their collaborators and students in the 1980s and that early work appeared in a series of articles and dissertations. The concept received its most coherent, complete, and influential presentation in Dayan and Katz's (1992) *Media Events: The Live Broadcasting of History*. Following that work, 'media event' has become a technical term in communication and media theory, to describe what is essentially ritualized, ceremonial television: an interruption of the normal routine for a live broadcast from a remote location of a preplanned event organized independently of the media; the tone of the broadcast is more serious, ceremonial, or even reverential than normal, it attracts unusually large and attentive audiences, and it often has serious political and social consequences. The concept has been used in studies of a wide range of phenomena, from formal state events to media award shows to cult films and fan communities to disaster news and terrorism, as well as generating extensive and vigorous theoretical debate (e.g., Liebes and Curran, 1998; Couldry et al., 2010).

The early studies focused on Sadat's first trip to Jerusalem, the early travels of Pope John Paul II, the wedding of Prince Charles and Lady Diana, the televised Olympic Games, and historic events like the astronauts' walk on the moon and the funeral of President Kennedy. Examples from recent years include the first inauguration of President Obama, the funeral of Edward Kennedy, the Opening Ceremonies of the Chinese Olympic Games, the Royal Wedding of Prince William and Catherine Middleton, and the Diamond Jubilee of the Queen of England. These examples, with the explicit and intended centrality of ceremonial form and attitude, illustrate the paradigmatic form of the media event as an example of ritual communication and an updating of Durkheim's (1912/1995) analysis of the role of ritual and ceremony in modern society.

One of the early approaches to the study of media events melded Weber's method of ideal-type sociology with genre theory. Media events were seen as tending toward three generic forms: contest, conquest, and coronation, corresponding with Weber's three forms of rational, charismatic, and traditional authority (Katz and Dayan, 1986). Contests are modernist; they celebrate fair play, the empirical measure of results, and rational decision. Conquests, like Sadat's trip to Jerusalem, celebrate the heroic individual. Coronations, like inaugurations or royal events, celebrate tradition and continuity. This formula has been a focus of scholarly attention with repeated proposals to add conflict as a fourth term and other analyses amending and elaborating the three original terms. Dayan and Katz (1992) also explored the possibility that all ceremonial media events contain at least rudimentary elements of all three forms with one or another being elaborated in a given instance.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.95046-3

This is clear in the Olympic Games, for example, where the sports themselves are rationalist contests, the media coverage features conquest stories of charismatic heroes, and the whole event is opened, closed, and punctuated by carefully governed traditional ceremonies.

Other uses of the term media event deserve attention and help illustrate some of the key issues of concern and debate in the scholarly literature. Among the professionals who coordinate events for media coverage, a media event is anything arranged for the purposes of media coverage, such as a press conference, public appearance, or photo opportunity. These events are, by definition, designed to achieve the social purposes of their subjects or sponsors. Given the predominant role of the for-hire communication professional vis-à-vis the journalist working, at least in principle, under the ethical ideal of representing the public interest, many of these events are contrived, the journalists who cover them feel manipulated, and so will much of the audience.

Already, early in the television age, long before either the media or their audiences had grown so cynical as in the present, Daniel Boorstin (1961) had coined the term pseudo-event for phenomena that have no stature independent of media coverage. The press conference, the celebrity interview, the news release, the photo appearance, and so on would not exist were it not for the chance of media coverage, and usually have no substantive social consequences unless they are covered. To the extent the media and the public are giving their attention to pseudo-events, they are presumably distracted from more real or more consequential events elsewhere. This is an important critique of media performance and it touches on fundamental philosophical issues for any communication theory as well. Boorstin and the many other critics after him using arguments of this form distinguish, at least implicitly, between symbolic things, or communicative things, and things that would exist independent of communication. They then privilege the noncommunicative things, at least implicitly, in preferring communication to be about that which would exist independently. This misrecognizes and undervalues the importance of communication itself as an event of the world and especially the performative capacity of communication to bring about effects and consequences in the world (Peters and Rothenbuhler, 1989). In such familiar rituals as the wedding ceremony or the oath of office, for example, it is the performance of signs and symbols, the communication as such, that produces the effects of the ritual, and some of this capacity inheres in communication of all types and contexts (Carey, 1988; Rothenbuhler, 1998, 2006a).

We will return to these issues below, but first let us consider what phenomena of the observable world are captured by the paradigmatic conception of a ceremonial media event. From Boorstin's own era, the live television broadcast of the funeral of President John F. Kennedy and the surrounding events attracted an extraordinary audience. For extended periods of time it appears that a majority of all Americans were watching and hearing the same events at the same time on television. These were ceremonial events that followed a social crisis and preceded a return to normality. Research conducted at the time showed that huge numbers of people were attending to the television coverage, reporting that they thought it important, and rearranging their schedules to watch it, often in groups of family and friends. Many people also reported symptoms of stress, anxiety, and uncertainty in the days following the assassination, which receded to normal levels in the days and weeks following the televised funeral (Greenberg and Parker, 1965).

Similar phenomena were found in attendance to television coverage following the terrorist attacks on 11 September 2001 (Noll, 2003). Those events and their coverage lacked a central ceremony in any classic sense, of course, they were not preplanned, institutionally sponsored ceremonial media events in the standard form, nevertheless, they were replete with quasi-religious and ceremonial symbols and forms (Rothenbuhler, 2005). The crisis event was an anticeremonial, antisocial interruption of everyday life and was compensated for, and recovered from, via a counterbalancing extra attention to symbol and ceremony in the media coverage following the attack, from that first morning right on through the annual anniversary ceremonies (Rothenbuhler, 2006b).

To choose another example, seemingly from the opposite end of the spectrum of seriousness, every four years the broadcasting media interrupt their normal schedules for special coverage of the Olympic Games. This is a large, regularly reoccurring event, organized by agencies officially independent of the media. Much of the coverage is live and both broadcasters and audiences betray an ethical preference, in criticizing edited and tape-delayed coverage and packaged programming segments. Media agents emphasize the importance of the games within and beyond sports; they make frequent historical comparisons and provide coverage of preceding and concurrent social, political, and cultural events. The games attract an audience that is unusually large, heterogeneous, and attentive. They are more likely than normal television audience members to be in groups, to be eating and drinking, to have planned their viewing ahead of time, to have rearranged their schedule to accommodate the viewing, and to have visitors from outside the home (e.g., Rothenbuhler, 1988).

Therefore, although originating in the world of sports and games rather than politics, work, religion, or life and death, the broadcast Olympic Games also show evidence of seriousness. Although they have much of the pseudo-event about them – so much of the television presentation is designed only for presentation, or repackaged so radically for coverage – along with the many other examples of media events they illustrate another classic dictum of social analysis: If people define things as real, they will be real in their consequences.

Many of the events selected for case studies in this literature have continued to be remembered as peak moments in television and media history. In examples like the broadcasts of the funeral of President Kennedy, Anwar Sadat's first trip to Jerusalem, the early trips of Pope John Paul II, selected Senate hearings, the first inaugural of President Obama, the television coverage itself became ceremonial and the audience was recruited into a sense of participation at a distance. In that case the activities on the scene, the coverage, and the audience activities at home or wherever viewed, together constituted a new event, a media event that was constructed as and functioned as a ritual in mediated communication. As ritual communication, it had the performative capacity to create its own social consequences, to produce the realities it described.

Theoretical Implications and Conceptual Tensions

If the media event constitutes a ritual and it is not a false one, those facts have major implications for social theory more broadly – and the idea was not received without debate. Other strains of conceptual tension followed from a general suspicion of ritual and representation, the selection of events for study, an interest in applying the idea to examples of conflict and social change, and a concern for the influence, if not the power, of the media. We will consider each of these issues in turn.

The face-to-face context is conventionally presumed to be necessary for ritual and ceremony; such communicative products as television or newspaper coverage, commemorative art, or memoirs are considered by-products. The audience on the scene may be considered primary, and may participate in the ceremony as congregation or witness, while media audiences would be considered secondary spectators. In the media event, though, the happenings in the face-to-face context are only a part of the event, no more important than the coverage and the activities of the at-home audience. The audience on the scene of a state funeral media event, for example, plays a role on the stage of the event, while the audience at home becomes the official audience, or even a kind of dispersed congregation (Dayan et al., 1984; Katz and Dayan, 1985).

MacAloon (1984) denied this was possible. While analyzing the role of ritual in pulling live spectators at the Olympic Games into a communal order, he asserted that ritual and community by television were impossible. While the power of symbols is widely recognized and the capacity of ritual to create social realities is accepted by scholars, as well as anyone who has ever planned a wedding, the idea that television could be the medium of that power is discomfiting. Television is, after all, a technology of spectatorship and a medium of entertainment, on the one hand, and the rational, modernist principles of journalism on the other. In the one case it is not about reality, nor serious, and in the other it should be objectively reporting rather than creating.

Compare the media event with a broadcast religious ritual. The broadcast would not appear to be the ritual, any more than a painting of a scene could be mistaken for the scene. Yet, observation consistently shows that for the faithful the broadcast coverage can function as participation in the ritual and have many of the same effects that physical participation would have. Who is to tell the faithful that television viewing does not count, if their belief and behavior confirms that it is special in their experience, and if religious leaders encourage such behavior and confirm their feelings? While the Pope himself has on occasion blessed the television audience, research shows that both commercial entertainment dramas devoted to religious subjects and apparently exploitative, money-raising television preachers also inspire genuine faith and devotional activity in some of their audience. Who, watching television on the morning of 11 September 2001 did not feel that they had witnessed the collapse of the Twin Towers? Much of the television audience of Lady Diana's funeral also like the audience for President Kennedy's funeral 35 years earlier, and surely many in the smaller audience for Ted Kennedy's funeral more recently, acted as if and felt themselves to be participating in the mourning. Of course the audience at home does not get wet if it rains, but that may not matter for meanings and effects that can work through the dramatic performances of mediated communication. The implication is that the media event is a new form for the ritual and ceremony that Durkheim (1912/1995) argued provides the quasi-religious foundations of society.

The ceremonial aspect of media events and the social consequences of its ritual power are quite distinct from what is normally expected of news coverage. There is reason for concern that it could be misused, and that the attitudes and behaviors appropriate to ritual and ceremony might spread to other domains where they would not be appropriate. Becker (1995), Chaney (1983, 1986), Liebes (1998), and others have pointed to the dangers of ritualizing media coverage. Politicians will prefer ceremonial treatment as it implies their actions are that much more important, while giving journalists and the public reason to regard them reverentially rather than suspiciously. From the media point of view, interrupting normal programming for special coverage can give the audience special reason to watch, it can enhance the status of the journalists and their channel, and the excitement can be its own reward. There are pressures, then, to ritualize even small events such as the opening of a press conference. Simultaneously, we can expect media organizations to be on the lookout for any event large enough to warrant special coverage and the claim to being a participant in historic events. Such tendencies can short-circuit rational consideration and democratic process, and diffuse the resource of audience attention.

Consider also the performative power of the ritualized media event. If successful media events can produce social integration and loyalty to ideals, symbols, and traditions, any critical consideration of the desirability of that outcome is obviated. Media events are, by definition, sponsored by authorities, central powers, and traditional institutions with the cooperation of the mainstream, national media, and they may in turn have the power to create more loyalty to these already powerful institutions. This has attracted considerable critical attention. Similarly, as the case studies on which this literature is built were primarily selected to illustrate these very ideas – state funerals, royal weddings, and such – there have been many attempts to broaden the concept to include more popular, less authoritative events like sports matches, televised contests, award shows, and fan-organized events, on the one hand, and to include more conflictful, unofficial, or antiauthoritarian ones on the other hand.

The study of mediatization (Lundby, 2009) also casts a new light on media events. This line of research shows how the internal forms and logics of media institutions, genres, and technologies become influential, implicitly or explicitly, in other social spheres. Politics, to take the most familiar example, has come to be shaped by and for the media, taking on many forms familiar from news, entertainment, and celebrity culture. So too has contemporary religion and education been influenced by media logics. Children's toys are increasingly designed to fit the promotional logics of television; fast food restaurants and airport waiting rooms have become media outlets, among many other examples (see Hjarvard, 2008).

These studies of mediatization raise questions about the production of media events and the influence of the needs of television on the event itself. This question had precedent in the earliest days of television and it remains a difficult one to

sort (Lang and Lang, 1952; Katz and Dayan, 2003). The small screen and domestic setting of the television, for example, produces a preference for close-ups, thus faces. The need to tell a story, the need to attract and hold attention leads to a preference for emotional reactions on those faces and narrators who speculate about feelings. If these forms are used for socially important events, it could tend to produce a personalistic, emotional understanding of society and history – not what the theory of modern democracy would prefer. In the process of producing media events, then, there may be a mediatization of the social forms of a previously independent institution; the ceremonial order may then be reshaped for television. This can be readily observed in the Olympic Games over the last 50 years and one wonders if the Queen of England's Diamond Jubilee celebrations in 2012 was not the most television-ready of them all.

The historical influence of the media on traditional rituals and ceremonies is only one side of the process. Another aspect of mediatization in relation to media events is the tendency of the media to 'eventize' everything they can. The mass media are attention machines; they are designed for attracting, holding, and aggregating attention. Media professionals, then, prefer what is special, unique, important, big, and dynamic; they will work to develop those characteristics wherever possible. In this way routine news and everyday entertainment comes to be promoted as if it were a special event worthy of interrupting the routines of our lives, like the ceremonial media event.

In addition to the theoretical questions about the possibility of mediated ceremony and participation at a distance, and concern that the power of that is ripe for misuse, concern that the ritual form of the media event is not appropriate to news coverage, on the one hand, or that too much influence would go the other way and that media would reshape the forms and processes of previously independent institutions, there was also a concern that the original work on media events was too focused on integrative events, on traditional events, and socially positive outcomes.

Some of this appeared in calls to shift attention from ceremonial television to the live broadcasting of history. The implicit tension between those orientations was displayed by the title of Dayan and Katz's book, *Media Events: The Live Broadcasting of History*, and while the consistent emphasis of the text was on the ceremonial, for some readers what was most intriguing were the examples of television and media participating in historical events and processes of social change.

Scannell (1995) criticized the book for 'historical flatness,' for lack of attention to how the past was different. He proposed that the type of events under study were already fading away and that there is "something almost nostalgic about the kinds of event which the authors take as their focus of concern" (p. 155). The aura of ceremony and the charisma of authority, he proposed, were worn away by television and the electronic media, by the constant performance and display, which teach the audience to watch as critics rather than as citizens or congregants. Nevertheless, he points out, the moments of live coverage of highly charged, real world events do produce a sense of witnessing, and it is this historically situated experience that he argues deserves more attention. He would, in other words, emphasize what the media coverage of the Kennedy assassination and the funeral have in common, as important historic events unfolding in live broadcast coverage. The more ceremonial conception of media events emphasizes the differences between the funeral, as a planned ceremonial, and the news events leading to it. One view emphasizes the flow of historic experience, the other emphasizes the forms of the communication that punctuate that experience.

The emphasis on history over ceremony draws attention more toward change and conflict than tradition and integration. Others turned the media events framework on specifically conflictive events. Carey (1998) provided an analysis of televised political rituals of degradation and excommunication. The Senate confirmation hearings for Supreme Court nominee Robert Bork – which resulted in his rejection – is his primary example. The point of Carey's analysis is to demonstrate a type of media event devoted to "ceremonies at the social level of excommunication defining the permissible range of social discourse; and, at the individual level, of status degradation marking out the consequences for transgressing this range." These are, then, ceremonies of great social importance, and serious personal consequence, taking place also, like the more integrated forms of media events, in and by media.

Liebes (1998) proposed the disaster marathon as a quasi-genre of news coverage in which some of the thinking and work routines appropriate to media events are carried over into other coverage. Liebes' concern is with open-ended, live coverage of disastrous events – terrorist bombings, airplane crashes, and so on. In such cases, media workers turn over much of their normal control of coverage to unscripted events, available spokespeople, and person-in-the-street interviews. Emotional reaction prevails over analysis, with one emotional expression feeding the next and no one being able to bring the coverage to a sensible close. Liebes worries that such live-coverage marathons can highjack a representative democratic process, as the media coverage of emotional public reactions short-circuits the deliberative process – and actually coaches the public to respond more emotionally than thoughtfully.

One of the interesting features that Liebes essays on disaster marathons, has in common with earlier writing on media and ritual, is an underlying concern that what might be appropriate for one category of social life, ritual and ceremony, is being applied to another where it is not appropriate, in news coverage and politics.

This line of analysis was developed further in a series of articles on the news coverage of the terrorist attacks on 9/11 and subsequent political and popular responses (Blondheim and Liebes, 2002, 2003; Liebes and Blondheim, 2005). The argument was that the disaster marathon form of coverage has the unintended consequence of increasing the terror, and thus the terrorists' grip on public attention and political process. Others have examined terrorism as a kind of media event (Weimann, 1987; Weimann and Winn, 1994; Nossek, 2008) and Sumiala (2013) has turned the lens on news coverage of school shootings and other violent events, as well as celebrity deaths.

Parallel to these works on conflictive and negatively charged forms of media events has been a literature proposing a critical theory approach to the topic (e.g., Couldry, 2003). From this perspective, a media event sponsored by the state or some other

powerful institution and carried by capitalist media, that succeeds in recruiting large and attentive audiences, especially when willing to suspend disbelief and grant the values, symbols, and narratives of the event their credulity, calls for critical analysis. These scholars focus on such issues as how the sponsoring institution uses the event to maintain its social power, how the media use the event to further their own interests, including cementing their position as privileged communicators providing access to the symbolic social center, and how the communicative relationship with the audience is then distorted, portraying itself as a transparent channel when it is in fact manipulation. The point of emphasis among critical scholars varies, of course. For some, the key is that media events in modern society work as Bloch (1989) said about rituals of authority in traditional societies, as a distraction from the real. Others, Couldry (2003) especially, emphasize that media events are 'constructions, not expressions' and that the sense of social order, power, importance, and connection that they convey are their own self-reinforcing creations. This performative aspect, the capacity of the media event to produce realities it purports to describe, was already inherent in the theory and attention to the construction and performance of events was a thread in the earlier literature. Nevertheless, the emphasis Couldry and others laid on the role of the media event in producing and preserving social power is distinct.

Media Events in a Multichannel, Hyperpartisan, and Globalized World

The paradigmatic ceremonial television conception of media events was developed in the 1980s and early 1990s, in a world where one or a few national networks dominated television in most countries, where national communication systems were relatively closed off to each other, when there was hope for peace between the Israelis and Palestinians, hope for consensus among the parties in the Western democracies, the Cold War had ended, and terrorism, war, and conflict did not dominate the news and political discourse as they have come to do in the years since. The much changed conditions in which we share our globally connected, contentious, and violent world, and a seeming decline in the number and importance of media events of the traditional kind, produced new critiques of the original conception and to new thinking on the part of its authors as well. As Scannell (1995) proposed, historical thinking has over taken the more formalist, theoretical focus on ceremony.

Katz and Liebes (2007) offered a major reconsideration of media events, arguing that coverage of disaster, war, and terror had coalesced into a kind of disruptive media events genre of its own that was upstaging ceremonial events, thus displacing the integrative functions they had served. They proposed that the changing technologies and organization of the media had fragmented audiences, dispersed their attention, and made live coverage of purportedly special events routine. Media audiences worldwide had grown more cynical and disenchanted, less ready for ceremony and celebration, more prepared to believe bad news than good. And the global media system allows bad news from somewhere at most any time. The result was that "marathons of terror, natural disaster, and war have become established genres of mainstream television" (p. 163). This was a move to open up the concept of media events to consideration of major news events, historical context, and change, as well as conflict. It was well received by former critics of the media events theory. Less explicitly, it was a move away from the ceremonial conception and the forms of analysis that follow from that.

Dayan too has been involved in the reconceptualization of media events, motivated explicitly by greater attention to the changing historical context. Price and Dayan's (2008) *Owning the Olympics* collects a set of studies focused on preparations for the 2008 Summer Olympics in China and its worldwide broadcast, and presented as testing and updating the media events idea in the context of a changed world. Across the chapters collected there we see the Chinese government, its journalists, and people, the International Olympic Committee, sports federations, other nations and national media, journalists on the scene and elsewhere, professionals and amateurs, affiliated and independent, advertisers, and sponsors engaged in the push and shove for access and control of the media event stage, narrative, symbols, public attention, and interpretation. The backstage production of media events has been much understudied, and this book makes a valuable contribution while again shifting attention toward the daily flow of events, toward history, and away from ceremony. In fact, the book was completed before the Games opened and there is not one chapter on what happened then; the ceremonial has no place in the book, only the context of its preparation.

In concluding these studies Dayan (2008) makes a series of points about the changing context and nature of media events. First, the semantic context, the mood of the world is more conflictive in recent years than it was in those hopeful years when the first media events work was done; now, Dayan proposes, events produce division and conflict as much as they bring people together. The progression of the Olympic torch around the world before the Chinese games opened, for example, became the locus of repeated protests, conflict, and attempts at censorship. The syntactics of media events have changed too, as the elements of their form have been used throughout the media to make ever more instances of ordinary news or entertainment, a bit more eventful, and thus reducing the unique attractions of more traditional events like the Olympics. The pragmatics, then, have changed too as the call of a ceremonial media event to a 'we' has become profligate, with daily, hourly claims on loyalty we no longer have to give to the media. The call or the claim that we have a collective identity has become banal and lost its power. Dayan sees, then, as others do, a disenchanted world in which the magic of the ceremonial media event no longer has much power.

These articles have been widely interpreted as a major revision of the concept of a media event, or even an abandoning of it, claiming that media events no longer work as they once did.

There is an alternative view, though. Rothenbuhler (2010b) argues that ceremonial media events, disruptive events, integrative events, disaster marathons, conflict events, denigration rituals, and all of the other forms that have been proposed are usefully conceived as varying instances of ritual communication.

As instances of ritual communication each of these phenomena is communication according to form that enables participation in serious life, even when implicitly so. Ceremonial media events are explicit in following ritualized forms and in their claim to be doing, as well as portraying, serious things, such as in a presidential inauguration. News coverage of crises, disasters, terror, and war may not be intentionally or explicitly ritualized, but it often is nevertheless (Coman, 2005). The ritualized elements of this news coverage work in the same ways as such familiar and intentionally planned ritual communication forms as the wedding ceremony: They are both self-referencing, self-steering, reality constituting symbolic systems; they both tend to create the worlds they portray (Rothenbuhler, 2010a). The disaster marathon may have attracted, and deserved, critical attention from Leibes and other scholars for violating the rationalist principles of journalism and producing more fear, conflict, and division. Its production of those consequences, though, depends precisely on those formal elements, as shown when the coverage of terrorism becomes routinized (Liebes and Kampf, 2007). It is not only history, but the form of the 'ceremonial' that changes and produces changed outcomes.

Conclusion

In sum, media events represent an important social form of dispersed ceremony. It is a form that raises important questions. On the one hand are questions about the power of communication and the necessity of the face-to-face context. On the other hand are the host of questions arising from the incongruity of serious, quasi-religious social functions being served by media normally devoted to entertainment, news, and profit. It is a form that can be abused or inappropriately imitated by politicians, government, the media, and other institutional interests.

The study of media events has also generated productive intellectual debates and the concept has been adapted to a broader range of examples, including conflict and social change, as well as the originally studied integrative, ceremonial events, and has been reconsidered and revised to fit changing media systems and historical circumstances.

See also: Audiences, Media; Mass Media: Introduction and Schools of Thought; Political Advertising; Terror and the Internet.

Bibliography

Becker, K., 1995. Media and the ritual process. Media, Culture, and Society 17, 629–646.

Blondheim, M., Liebes, T., 2002. Live television's disaster marathon of September 11 and its subversive potential. Prometheus 20, 271–276.

Blondheim, M., Liebes, T., 2003. From disaster marathon to media event: live television's performance on September 11, 2001, and September 11, 2002. In: Noll, A.M. (Ed.), Crisis Communication: Lessons from September 11. Rowman & Littlefield, Lanham, MD.

Bloch, M., 1989. Ritual, History, and Power: Selected Papers in Anthropology. Athlone Press, London.

Boorstin, D., 1961. The Image: A Guide to Pseudo-Events in America. Artheneum, New York.

Carey, J.W., 1988. Communication as Culture: Essays on Media and Society. Unwin Hyman, Boston.

Carey, J.W., 1998. Political ritual on television: episodes in the history of shame, degradation, and excommunication. In: Liebes, T., Curran, J. (Eds.), Media, Ritual and Identity. Routledge, London, pp. 42–70.

Chaney, D., 1983. A symbolic mirror of ourselves: civic ritual in mass society. Media, Culture, and Society 5, 119–135.

Chaney, D., 1986. The symbolic form of ritual in mass communication. In: Golding, P., Murdock, G., Schlesinger, P. (Eds.), Communicating Politics: Mass Communications and the Political Process. Holmes & Meier, New York.

Coman, M., 2005. Cultural anthropology and mass media: a processual approach. In: Rothenbuhler, E.W., Coman, M. (Eds.), Media Anthropology. Sage, Thousand Oaks, CA, pp. 46–55.

Couldry, N., 2003. Media Rituals: A Critical Approach. Routledge, London.

Couldry, N., Hepp, A., Krotz, F. (Eds.), 2010. Media events in a Global Age. Routledge, London.

Dayan, D., 2008. Beyond media events: disenchantment, derailment, disruption. In: Price, M., Dayan, D. (Eds.), Owning the Olympics: Narratives of the New China. University of Michigan Press.

Dayan, D., Katz, E., 1992. Media Events: The Live Broadcasting of History. Harvard University Press, Cambridge.

Dayan, D., Katz, E., Kerns, P., 1984. Armchair pilgrimages: the trips of Pope John Paul II and their television public. On Film 13, 25–34.

Durkheim, É., 1995. The Elementary Forms of the Religious Life (K.E. Fields, Trans.). Free Press, New York (Original work published 1912).

Greenberg, B.S., Parker, E.B., 1965. The Kennedy Assassination and the American Public: Social Communication in Crisis. Stanford University Press, Stanford, CA.

Hjarvard, S., 2008. The mediatization of society: a theory of the media as agents of cultural change. Nordicom Review 29 (2), 105–134.

Katz, E., Dayan, D., 1985. Media events: on the experience of not being there. Religion 15, 305–314.

Katz, E., Dayan, D., 1986. Contests, conquests, coronations: on media events and their heroes. In: Graumann, C.F., Moscovici, S. (Eds.), Changing Conceptions of Leadership. Springer-Verlag, New York, pp. 135–144.

Katz, E., Dayan, D., 2003. The audience is a crowd, the crowd is a public: latter-day thoughts on Lang and Lang's "MacArthur Day in Chicago." In: Katz, E., Peters, J.D., Liebes, T., Orloff, A. (Eds.), Canonic Texts in Media Research. Polity Press.

Katz, E., Liebes, T., 2007. "No more peace!": how disaster, terror and war have upstaged media events. International Journal of Communication 1, 157–166.

Lang, K., Lang, G.E., 1952. The Unique Perspective of Television and Its Effect: A Pilot Study.

Liebes, T., 1998. Television's disaster marathons: a danger for democratic processes? In: Liebes, T., Curran, J. (Eds.), Media, Ritual and Identity. Routledge, London, pp. 71–84.

Liebes, T., Blondheim, M., 2005. Myths to the rescue: how live television intervenes in history. In: Rothenbuhler, E.W., Coman, M. (Eds.), Media anthropology. Sage, Thousand Oaks, CA, pp. 188–198.

Liebes, T., Curran, J. (Eds.), 1998. Media, Ritual, and Identity. Routledge, London.

Liebes, T., Kampf, Z., 2007. Routinizing terror: media coverage and public practices in Israel, 2000–2005. Press/Politics 12, 108–116.

Lundby, K. (Ed.), 2009. Mediatization: Concept, Changes, Consequences. Peter Lang, New York.

MacAloon, J.J., 1984. Olympic games and the theory of spectacle in modern societies. In: MacAloon, J.J. (Ed.), Rite, Drama, Festival, Spectacle: Rehearsals toward a Theory of Cultural Performance. Institute for the Study of Human Issues, Philadelphia, PA, pp. 241–280.

Noll, M.A. (Ed.), 2003. Crisis Communications: Lessons from September 11. Roman & Littlefield.

Nossek, H., 2008. "News media"-media events: terrorist acts as media events. Communications 33, 315–332.

Peters, J.D., Rothenbuhler, E.W., 1989. The reality of construction. In: Simons, H.W. (Ed.), Rhetoric in the Human Sciences. Sage, London, pp. 11–27.

Price, M., Dayan, D. (Eds.), 2008. Owning the Olympics: Narratives of the New China. University of Michigan Press.

Rothenbuhler, E.W., 1988. The living room celebration of the Olympic Games. Journal of Communication 38 (3), 61–81.

Rothenbuhler, E.W., 1998. Ritual Communication: From Everyday Conversation to Mediated Ceremony. Sage Publications, Thousand Oaks, CA.

Rothenbuhler, E.W., 2005. Ground zero, the firemen, and the symbolics of touch on 9/11 and after. In: Rothenbuhler, E.W., Coman, M. (Eds.), Media Anthropology. Sage, Thousand Oaks, CA, pp. 176–187.

Rothenbuhler, E.W., 2006a. Communication as ritual. In: Shepherd, G.J., St John, J., Striphas, T. (Eds.), Communication as…: Stances on Theory. Sage, pp. 13–21.

Rothenbuhler, E.W., 2006b. Déchirure symbolique et processus de réparation: Les témoins du 11 Septembre [Symbolic disorder and repair after 9/11]. In: Dayan, D. (Ed.), La Terreur Spectacle: Terrorisme et television. De Boeck, Brussells, pp. 199–209 and Institut National de l'audiovisuel, Paris.

Rothenbuhler, E.W., 2010a. From media events to ritual to communicative form. In: Couldry, N., Hepp, A., Krotz, F. (Eds.), Media Events in a Global Age. Routledge, London, pp. 61–75.

Rothenbuhler, E.W., 2010b. Media events in the age of terrorism and the internet. Revista Româna de Jurnalism si Comunicare 5 (2), 34–41.

Scannell, P., 1995. Media events. Media, Culture, and Society 17, 151–157.

Sumiala, J., 2013. Media and Ritual: Death, Community, and Everyday Life. Routledge, London.

Weimann, G., 1987. Media events: the case of international terrorism. Journal of Broadcasting and Electronic Media 31, 21–39.

Weimann, G., Winn, C., 1994. The Theater of Terror: Mass Media and International Terrorism. Longman, New York.

Media Imperialism

Annabelle Sreberny, Leicester University, Leicester, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9489–9494, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Media imperialism was a key critical interpretive paradigm in international communication in the 1960s. It prompted a considerable body of empirical research and was used in support of key policy concerns such as the New World Information and Communication Order. Yet as the transnational media environment develops, the paradigm is increasingly criticized from a number of vantage points: from historical process and effects, from production and flow, from audiences and effects, from genre, from media policy and regulation. As media industries around the world come of age and as globalization processes spread, including massive movements of people as well as of images, 'media imperialism' is of diminishing use in understanding contemporary cultural processes, although its underlying concern about the hegemony of Western, primarily US, cultural domination remains a live concern.

'Media imperialism' has been one of the most influential models in international communication. It has given rise to a significant body of empirical research, considerable theoretical debate, as well as support for international policy interventions about global communications imbalance. However it is increasingly subject to critique, particularly as a reflection of a particular moment of international media history that is now passing.

The Broader Paradigm of Cultural Imperialism

Media imperialism needs to be seen as a subset of the earlier and broader paradigm of 'cultural imperialism' that is most closely associated with Herbert Schiller (1969, 1976). This critical Marxisant paradigm itself developed out of broader critiques of the triumphalist paradigm of 'modernization' propounded in the late1950s and the 1960s predominantly by US theorists. This 'dominant' model proposed a single global process of modernization through a unilinear diffusion of Western technologies, social institutions, modes of living, and value systems to the eponymous 'Third World' (Sreberny-Mohammadi, 1996). Critical views, especially those advanced by dependency theorists such as Gunder-Frank, saw the process instead as the spread of inegalitarian capitalism from a few core advanced industrial countries to the periphery, the South. In this process, the West was supplied with raw materials and cheap labor as well as with markets for their manufactured goods. The spread of Western influence was seen as establishing an ideological bulwark against the much-feared spread of communism.

Cultural imperialism, Schiller argued, 'best describes the sum of the processes by which a society is brought into the modern world system and how its dominating stratum is attracted, pressured, forced, and sometimes bribed into shaping social institutions to correspond to, or even promote, the values and structures of the dominating center of the sytem' (1976: p. 9). For Schiller, the carriers of Western cultural influence included communication technologies themselves, which were not value-neutral instruments but imbued with capitalist values; language (evident as English takes over as the global linguistic currency from the varieties of colonial linguistic heritage); business practices; the genres as well as content of soap opera, blockbuster film, popular music, etc. Schiller's view was US-centric, and he saw media as central elements in the global expansion of capitalism centered on the US, fueled by advertising and consumerism. In the context of the Cold War, US policy makers' promotion of the 'free flow of information' – with no trade or legal barriers to the production or movement of mediated cultural products – undoubtedly helped the spread of US hegemony in ways that were consonant with, if not totally driven by, US foreign policy.

Media Imperialism

Media imperialism, a more focused subset of this model, was defined as 'the process whereby the ownership, structure, distribution or content of the media in any one country are singly or together subject to substantial external pressures from the media interests of any other country or countries without proportionate reciprocation of influence by the country so affected' (Boyd-Barrett, 1977: p. 117). It differentiated between the shape of the communications vehicle, a set of industrial arrangements, a body of values about best practice, and media content, although it is the latter claim that received most research attention.

It highlighted both the issues around 'cultural invasion' as well as a more general imbalance of power resources. In essence, the model argued that the world patterns of communication flow, both in density and direction, mirrored the system of domination in the international economic and political order, the control by the West over the rest.

Empirical Research

Part of the strength of the paradigm of media/cultural imperialism was its compelling clarity, which gripped many a critical academic imagination, and which suggested many fertile avenues of empirical research. It precipitated numerous studies that sought to examine the unidirectional flow of mediated

product from the West to the South. The 'one-way street' of television traffic was mapped by Nordenstreng and Varis (1974); film production and distribution was studied by Guback and Varis (1982); international news flow by the IAMCR/UNESCO study (Sreberny-Mohammadi et al., 1985). Scholars investigated the reach and dominant role of Western news agencies (Boyd-Barrett, 1980), supplemented latterly by study of the televisual news agencies (Paterson, 1998). Others showed that Third World media organizations were modeled on those of the earlier mother empires (Katz and Wedell, 1977). More recent studies detail the on-going transnational processes of conglomeratization and the increasing vertical and horizontal linkages of media corporations (Herman and McChesney, 1997).

Cultural imperialism and media imperialism were among the first models to take the global dynamics of media production and diffusion seriously. These paradigms posed important questions about cultural homogenization, about the diffusion of values such as consumerism and individualism, and about the possible impacts of Western media product on indigenous cultures in the South (Hamelink, 1983).

Political/Policy Struggles

These models also provided the conceptual and empirical framework for the NWICO (New World Information and Communication Order) debates which preoccupied UNESCO and the nonaligned movement through the 1970s until the USA, the UK, and Singapore walked out of UNESCO in 1984–5 (see Vincent and Galtung, 1992). The MacBride Commission gathered input from over 100 international experts, and in its final report, *Many Voices, One World* (UNESCO, 1980) challenged as inadequate the prevailing idea of 'free flow of information' since it functioned to privilege Western hegemony and cultural diffusion and to turn the South into mere consumers of Western news values, entertainment, and advertising images. Instead the report suggested the important qualification of a 'free and balanced' flow, which was adopted as UNESCO policy, although it defaulted from precise suggestions as to how, for example, international news coverage could be appropriately 'balanced.'

It is significant that UNESCO funded many of the early landmarks in critical international communication research yet from the mid-1980s, its focus and rhetoric has shifted toward a more neo-liberal position and a greater concern with issues around freedom and democratization than around balance/flow.

The debates about global media inequality and concern about the logics of conglomeratization and the threats to local cultures continue even in 2001 at the forums of the MacBride Round Tables; the demand for a Peoples Communication Charter and in the activities of the Cultural Environment Movement (see Voices 21, 1999). However, while the debates in the 1970s involved Third World policy makers, the critical voices are now mainly Western academics, while Southern politicians have become more involved in nitty-gritty decisions around deregulation, digitalization, and media convergence.

Critiques of Media Imperialism

Arguments From and About Historical Process and Effects

'Media imperialism' was a problematic argument both theoretically and empirically from the beginning, and by year 2001 it appears increasingly ossified. An argument that might have had some validity for some countries at a certain period of time became the dominant critical paradigm, operating as a form of political correctness by which critics were seen as apologists for the USA and its demand for a 'free flow international regime for trade in cultural products' (Sinclair et al., 1995: p. 6). By the 1980s when the global television and other media industries really changed shape, the 'critical' paradigm had become the 'orthodoxy' (Sinclair et al., 1995: p. 5).

Historically it is doubtless the case in most, especially technologically based, endeavors that the prime mover sets the model: in cars, in high-rise architecture, in media. Broadcasting did not develop with world domination in mind, even if some of its spread has been consonant with Western foreign policy interests. However, the model of state-controlled broadcasting which was adopted in much of the South hardly supported the 'free flow of information' of Western interests. It is also the case that subsequent media players 'catch up,' begin to produce their own content, and rework genres to fit their cultural habits. Already in the 1970s, Tunstall (1977) was arguing that Latin America's high level of television imports in the 1960s was a transitional phase while nationally based television systems were being established.

Empirically, it is mainly from the 1990s with the expansion of global satellite systems and development of infrastructure that much of the population of the rural South has actually been introduced to television. Thus, oddly enough, media could not have had the strong effects imputed by the model (nor by the modernization theorists like Lerner, 1958) because the televisual media actually didn't reach most people in the South until the satellite expansion of the late 1980s.

The crude adoption of 'media imperialism' arguments promoted a truncated historical analysis and tended to disregard earlier historical processes. It is probably the case that the cultural legacies of imperialism – the spread of Christianity, training in and use of European languages, establishment of systems of schooling and higher education, spread of administrative cultures – have had more enduring and far-reaching effect as carriers of Eurocentric modernization to the South than have the subsequent arrival of electronic media (Sreberny-Mohammadi, 1997; Bitterlli, 1989).

Additionally, many other industries, connected to but not reducible to the media industries, produce and market symbolic goods (fashion, foodstuffs, architecture and interior design, consumer durables) while the global tourism and travel industries literally transport millions of people a year. A focus on only one element of the contemporary production of culture cannot be read for the whole. Capitalism is also not organized and instituted through media, albeit that media content diffuses images of the goods, lifestyles, attitudes that are increasingly commodified and available for consumption in the capitalist market. In that sense, the media can support

processes of economic transformation and the profound shift from culture as 'practice' to culture as 'commodity,' and here reconnect to the bigger issues around modernization and development that tend to become submerged in the focus on media structures alone.

Arguments About Production and Flow

'Media imperialism' was a model developed at a time of comparative media scarcity and newly established broadcasting structures in the South. By the year 2001 there are many significant cultural industries in the South: Globo in Brazil, and Televisa in Mexico produce telenovelas; a huge multimedia complex near Cairo supports the production of Islamic soap operas, which Turkey also produces. And if the focus shifts away from television alone to include other cultural products, the diversity increases: Bollywood, the familiar label for the Indian film industry, is the Eastern challenge to Hollywood in the sheer number of film titles produced yearly, with the Asian diaspora constituting a sizeable audience. The Iranian and Chinese film industries are gaining global recognition and audiences. The marketing of 'world music' has helped the diffusion of Algerian, Senegalese, Cuban, and Brazilian contemporary music. The Indian Zee-TV is a powerful independent newscaster while Qatar's Al-Jazeera is revolutionizing factual programming in the Arab World.

Thus the model's metaphors of the 'one-way flow' down the 'one-way street' from the West to the rest (Nordenstreng and Varis, 1974) are challenged by more diverse production, and more complex images of the 'patchwork quilt' (Tracey, 1988). Southern media exports into the West have been dubbed 'reverse imperialism' while many of the new media mogul empires are not based in the West and Asian corporations and entrepreneurs (SONY, Masushita) have bought up segments of Western media.

Arguments About Audiences and Effects

'Media imperialism' also implied direct unmediated effects of television programming on audiences, turning them into Americanized consumers. It tended to ignore the particularities of local culture and their meaning systems as well as other processes of modernization (shifts in material production, migration and tourism, political institutionalization and democratization) that often operate alongside media development.

New approaches to the 'active audience' within media studies have forced a rethinking of international effects also. As with domestic viewing practices, analyses of the negotiation of meanings around US-produced programs such as Dallas by 'foreign' audiences show different audience viewing strategies including the imposition of local interpretive frames on 'foreign' media product (Liebes and Katz, 1990). There is also strong evidence that audiences prefer locally made mediated culture where and when that becomes available (Sepstrup and Goonasekura, 1994). The original logic of argument also depended on a rather ossified notion of 'national' culture, whereas even US product changes, so that the *Cosby Show* has become a popular, if rather inaccurate, symbol of changing race relations in the USA.

All this means that a single global product doesn't always work across cultural divides. While a film such as *Titanic* may have global reach, falling into that magical, much-desired and heavily marketed category of the 'global popular,' most media production is now tailored for local, that is, linguistically bounded, sometimes still nationally-bounded, cultural markets. The mantra is to 'think global, produce local.'

Regional linguistic-cultural markets are growing in economic significance as the 'cultural discount' (Hoskins and McFadyen, 1991) of local cultural preference kicks in. While this can produce regional exchanges of news and television programming, it can also produce localized versions of the old paradigm, as in the concern about India's cultural dominance over the subcontinent.

Arguments About Genres

Discussion about the nature and definition of media genres in relation to the South has perhaps not been fully engaged. The advent of an externally originated cultural form into a society has often caused controversy in, for example, fear about the impact of free verse on traditions of formal poetry or about the impact of the novel on non-Western writing cultures. Yet now Latin American magic realism and postcolonial family sagas crossing generations and national boundaries have transformed the very form of the novel. A debate once raged about whether Latin American telenovelas were a new form of television series, inflected with socially developmental themes, or whether they were simply soap operas with commercial promotional messages. Anthropologists have analyzed the rise of the Islamic soap opera (Abu-Lughod, 1993) and it seems clear that Southern media texts are not 'simply' copies of Western texts, but neither are they utterly novel genres either. Textual give and take, multiple sources of inspiration and derivation in media content, needs to be taken seriously, and attention paid to how forms evolve over time, both within developed television industries like the USA and the UK as well as within 'younger' media systems. For example, Zee-TV is not the same as MTV, although it may also show music videos.

Arguments About Media Policy and Regulation

As the global media environment changes with growing convergence between broadcasting and telecommunications, growing digitalization and the spread of the Internet, so there has also been a move from notions of imperialism to ideas of cultural hegemony. Such constructs are less overtly spatial and usefully deconstruct a singular and unproblematic 'West.' There have been fierce intra-Western disputes, most obviously the French disagreements about the deregulated flow of US film within GATT talks, or the differences and competition between US and UK modes of news construction. Parts of South East Asia may now be more concerned about the dominance of Indian cultural product than US; and national law in one Western country encounters a different set of rules in another, as in the US-French dispute about the sale of Nazi ephemera through the Yahoo gateway on the Net.

Arguments About the State and Democratization

The concern of much of the debate with transnational media domination by the West and the dominance of the market often precluded serious analysis of national processes, particularly the complex class formations in the South and the role of states. Thus, for example in Latin America, repressive military juntas have used media as key institutions of authoritarian government (Waisbord, 2000) while in rapidly modernizing South East Asian economies the ruling elites have aligned with global corporate interests, often against the needs of their own population. As popular movements for democratization and equality have grown inside such states, this has prompted greater analytic focus on the relations between transnational processes and national politics and media control, and how states and markets interact within different political and economic formations.

It is amongst the autocratic, sometimes religious, regimes of the Middle East and South East Asia that the argument about 'media imperialism' is still heard. The 'protection of indigenous culture' can be used as a weapon to prevent internal change and demands for democratization, and the 'outside' can be constructed solely as a predator involved in acts of rape against the feminized, vulnerable nation. Yet, in numerous Southern countries – Indonesia, Nigeria, Iran, Brazil – secular intellectuals, ethnic groups, and women are struggling for greater political openness and basic human rights as well as a media environment which facilitates the construction of an inclusive national public sphere. Sometimes the transnational space can act as one of solidarity and support for such struggles.

The Rise of 'Globalization'

Many of the early debates around cultural imperialism and media imperialism actually prefigure key strands of argument that were taken up by the rhetoric of globalization in the 1990s: the recognition of the collapse of space and time through communications technologies; the commodification of social life; the significance of cultural flows. But within the model of globalization is also a sense of collapse of the older, single-centered hegemon into a more disorganised post-Fordist set of global production processes (Lash and Urry, 1994); a more decentered and unstable political and cultural environment; and a variety of resistances to Western modernity with the rise of primordial religious and ethnic loyalities. The interpenetration of the global and local actually produces multiple formations of modernity.

'Cultural imperialism' tended to see cultures as singular, somewhat ossified, nationally bounded entities: 'American' television impacting on 'Iran,' for example. There is increased awareness of the loosening of the hyphen in the nation-state and thus the 'national culture' is better seen as a site of contest. These are not just issues for countries in the South, but for most countries around the world. In the UK, this is enacted politically with the devolution of Scotland and Wales, and in debates around immigration policy; and culturally by debates about the religious content of the millenial exhibition in the Dome, and about the representation of ethnic minorities in the media. National identity may be practiced daily through a variety of symbolic forms and organizational routines, but its actual content may be more fluid and disputed than ever.

Globalization highlights the manner in which pressures to modernize are being met with particular local orientations and practices. It can still be argued that Western values lie hidden behind a seemingly neutral 'universal' and abstract set of processes, but the precise way these are encountered, understood, and dealt with varies from country to country, even within the same region. Latin America presents considerable national variation as does the Arab world; hence, even an increased awareness of regionalism is insufficient. Among media academics, more particularism, more case studies, more nuance is required.

The real collapse of the three worlds of development has left a conceptual lacuna. It is hard to propose such totalizing models any longer. Rather, we need regional frames that take cultural specificities, local structures, political values, seriously. Globalization has exacerbated economic disparities among – but also within – countries and has not ended political repression. Contemporary 'flows' include massive movements of capital and of people, so others are no longer 'there' but 'here' and building transnational communities through various kinds of diasporic media.

Work on the media has also tended to be focused on a single medium, initially mainly television, now heavily focused on the Net; yet rather bizarre situations exist whereby entire newspapers are banned, for example in Tunisia, Egypt, and Iran, while Internet-based information circulates with impunity. Looking at only one element of a complex media environment can also lead to wrong conclusions.

A recent attempt to rework the 'media imperialism' model suggested the need to 'encompass neo-colonialisms of inter-ethnic, inter-cultural, inter-gender, inter-generational and inter-class relations' (Boyd-Barrett, 1998: p. 167) but instead helps us to recognize that in the end the model is simply about power imbalances, which do not require the 'imperial' stamp to be of importance to examine. Indeed the 'imperialism' obscures the differentiated processes of economic inequality and complex power dynamics. Access to cultural expression often involves complex internal struggles for power as well as pushes from the outside. As Tomlinson (1991) has argued, cultural imperialism always consisted of many different discourses; the on-going attempt to rewrap them into one through the trope of 'media imperialism' is an increasingly forlorn task. The world has changed and so must our language and our theoretical frames.

Conclusion

Perhaps the enduring cultural legacy of the USA lies in the development of a commercial model of television as a medium: using entertainment content to attract audiences who could be sold to advertisers (Sinclair et al., 1995). While this is still not a universal model and many states still try to control broadcasting, commercial pressures and availability of multiple forms of media distribution/reception such as satellite and cable mean the globalization of capitalism and the

manufacture of mediated culture are in the ascendant. In that sense, the focus on US-centric media diffusion obscures a far wider concern around the generic mass production of culture that is now flourishing globally.

Editor's Note

The domination of mass media and the larger culture by, at most, a few Western institutions continues to be a prominent research theme in media studies and mass communications research. Useful contemporary statements of the issues and current thinking include: Fuchs, 2010; Neyazi, 2010; Omoera and Ibagere, 2010; Dowd and Janssen, 2011.

See also: Advertising: Effects; British Cultural Studies; Globalization and World Culture; Media Effects on Children; Media Effects; Television: History.

Bibliography

Abu-Lughod, L., 1993. Finding a place for Islam: Egyptian television serials and the national interest. Public Culture 5 (3), 494–512.

Bitterlli, U., 1989. Cultures in Conflict: Encounters between European and Non-European Cultures 1492–1800. Polity Press, Cambridge, UK.

Boyd-Barrett, O., 1977. Media imperialism: Towards an international framework for the analysis of media systems. In: Curran, J., Gurevitch, M., Woolacott, J. (Eds.), Mass Communication and Society. Edward Arnold, London, pp. 116–135.

Boyd-Barrett, O., 1980. The International News Agencies. Constable, London.

Boyd-Barrett, O., 1998. Media imperialism reformulated. In: Thussu, D. (Ed.), Electronic Empires. Arnold, London, pp. 157–217.

Boyd-Barrett, O., Rantanen, T., 1998. The Globalization of News. Sage, London.

Dowd, T.J., Janssen, S., 2011. Globalization and diversity in cultural fields: comparative perspectives on television, music, and literature. American Behavioral Scientist 55 (5), 519–524.

Fuchs, C., 2010. New imperialism information and media imperialism? Global Media and Communication 6 (1), 33–60.

Galtung, J., Vincent, R., 1992. Global Glasnost:Toward a New World Information and Communication Order? Hampton Press, Cresskill, NJ.

Guback, T., Varis, T., 1982. Transnational communication and cultural industries. Reports and Papers on Mass Communication. UNESCO, Paris. No. 92.

Hamelink, C., 1983. Cultural Autonomy in Global Communications. Longman, New York.

Herman, E., McChesney, R., 1997. The Global Media. Cassell, London.

Hoskins, C., McFadyen, S., 1991. The US competitive advantage in the global television market: Is it sustainable in the new broadcasting environment? Canadian Journal of Communication 16 (2), 207–224.

Katz, E., Wedell, G., 1977. Broadcasting in the Third World. Harvard University Press, Cambridge, MA.

Lash, S., Urry, J., 1994. Economies of Signs and Space. Sage, London.

Leibes, T., Katz, E., 1990. The Export of Meaning. Oxford University Press, Oxford, UK.

Lerner, D., 1958. The Passing of Traditional Society. Free Press, New York.

Neyazi, T.A., 2010. Cultural imperialism or vernacular modernity? Hindi newspapers in a globalizing India. Media, Culture and Society 32 (6), 907–924.

Nordenstreng, K., Varis, T., 1974. Television traffic – a one-way street?. Reports and Papers on Mass Communication, No. 70 UNESCO, Paris.

Omoera, O.S., Ibagere, E., 2010. Revisiting media imperialism: a review of the Nigerian television experience. International Journal of Research and Review 5, 1–18.

Paterson, C., 1998. What makes news. In: Boyd-Barrett, O., Rantanen, T. (Eds.), The Globalization of News. Sage, London, pp. 177–190.

Peoples Communication Charter, 1999. http://www.tbsjournal.com/Archives/Spring99/Documents/Congress/Charter_intro/Charter/charter.html.

Schiller, H., 1969. Mass Communications and American Empire. A. M. Kelley, New York.

Schiller, H., 1976. Communication and Cultural Domination. International Arts and Sciences Press, White Plains, NY.

Sepstrup, P., Goonasekura, A., 1994. TV transnationalization: Europe and Asia. Reports and Papers on Mass Communication. UNESCO, Paris.

Sinclair, J., Jacka, E., Cunningham, S., 1995. New Patterns in Global Television: Peripheral Vision. Oxford University Press, New York.

Sreberny-Mohammadi, A., 1996. The global and the local in international communication. In: Curran, J., Gurevitch, M. (Eds.), Mass Media and Society. Arnold, London, pp. 177–203.

Sreberny-Mohammadi, A., 1997. The many cultural faces of imperialism. In: Golding, P., Harris, P. (Eds.), Beyond Cultural Imperialism. Sage, Thousand Oaks, CA, pp. 49–68.

Sreberny-Mohammadi, A., Nordenstreng, K., Stevenson, R., Ugboajah, F., 1985. Foreign news in the media: International reporting in 29 countries. Reports and Papers in Mass Communication. UNESCO, Paris. No. 90.

Tomlinson, J., 1991. Cultural Imperialism. Johns Hopkins University Press, Baltimore, MD.

Tracey, M., 1988. Popular culture and the economics of global television. Intermedia 16 (2).

Tunstall, J., 1977. The Media are American. Columbia University Press, New York.

Voices 21 1999 http://www.tbsjournal.com/Archives/Spring99/Documents/Congress/congress.html.

Waisbord, S., 2000. Media in South America: Between the rock of the state and the hard place of the market. In: Curran, J., Park, M. (Eds.), De-Westernizing Media Studies. Routledge, London, pp. 43–53.

UNESCO, 1980. Many Voices, One World. International Commission for the Study of Communication Problems. UNESCO, Paris.

Media Talk Shows

Andrew Tolson, University of Leicester, Leicester, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A.G.D. Leurdijk, volume 23, pp. 15437–15440, © 2001, Elsevier Ltd.

Abstract

There are different definitions of media talk show ranging from news and current affairs, to celebrity interviews and audience participation programs. A comprehensive account might also include talk radio formats, such as call-ins, from which the pioneering American TV talk show, *Donhaue*, was developed. This article distinguishes between five variants, or subgenres, of this type of TV talk shows: the audience discussion format, the focus on social identity, therapeutic talk shows, confrontation programs, and belligerent interrogation. Through these various formats the role of the 'ordinary person' has changed, from 'sociable' forms of audience participation to individuals becoming the recipients of authoritarian moral judgments.

Definition

The fact that most forms of electronic media, including some uses of the Internet, involving people talking, might lead to the conclusion that there is a wide range of media 'talk shows.' This is indeed how some discussions proceed, using the term to refer to genres as diverse as news and current affairs (particularly breakfast news programs and late night debates) and celebrity interviews (or 'chat' shows) as well as audience participation programs. Furthermore, there are of course media enterprises explicitly dedicated to 'talk,' such as talk radio stations structured around phone-ins. More broadly still in contemporary culture, 'talk' is promoted as having a certain moral value, as in the slogan "It's good to talk" adopted by the UK company British Telecom (and used in the title of an important book on this topic – see Cameron, 2000). This validation of talk is not only used as a means to sell mobile phones, but it also catered for, in various ways, by conventional broadcast media.

However a moment's reflection suggests that there are some distinctions to be made. Programs differ generically in terms of their location, their participants, and crucially by the role of the audience. A breakfast news program is studio based, with 'talk' taking the form of announcements, reports, and live 'two-way' interviews between anchors and fellow journalists. There is no live audience here. On late night debates and celebrity interview shows the participants are public figures, and though there is a live audience, opportunities for its participation are restricted and in most chat shows it is invisible. The formats for (usually daytime) audience participation programs do differ, but as the phrase suggests the role of the audience is less circumscribed, the participants are ordinary people and the setting is more theatrical, with the stage itself coming into focus, as 'backstage' areas are included. So even if the term 'talk show' has been used to cover current affairs debates (such as the long running BBC program *Question Time*) and celebrity interviews (such as NBC's *David Letterman Show*) as well as the daytime talk shows, it is nevertheless important to recognize the differences, and that increasingly, that the term is used exclusively to refer to the third of these formats.

(Pre)history

The broader context remains important however particularly for understanding the history of media talk shows. Here, there are two levels of analysis. Firstly, at the empirical level, it is instructive to trace the participation of 'ordinary' people in broadcasting. In both the USA and the UK this started with experimental radio programming in the 1930s. For example, Wayne Munson (1993) discusses the emergence of 'human interest' or 'vox pop' programs, as well as game shows, on American radio in the 1930s, as precursors to what was the first 'comedy audience participation' TV show, *Art Linkletter's House Party* (1952). In the UK, such audience participation was less extensive, but by the mid-1930s 'ordinary' (or working-class) people began to appear in radio documentaries and in what were known as 'character' programs. The vox pop approach was used in the BBC radio program *In Town Tonight* which featured its presenter, Michael Standing, 'standing on the corner' interviewing passers-by, defined, in the terminology of that era as 'the-man-in-the-street.'

On 24 March 1939 *The Radio Times* included a special supplement featuring Standing on its cover and an article entitled 'The Man in the Street is a veteran broadcaster.' This article provides a retrospective summary of the aforementioned documentaries, plus a radio series called *Harry Hopeful* which "brought all sorts and conditions of 'ordinary' men and women together in a happy party to entertain you at a common microphone." Clearly then the 'party' idea was not entirely absent from the BBC (even though contributions to *Harry Hopeful* were scripted) but the importance of this program is that it introduces a second level of historical analysis. As discussed by Paddy Scannell (1996) – see also Scannell and Cardiff (1991) – *Harry Hopeful* introduced to Britain a key principle of media talk, namely 'sociability' in which ordinary, conversational talk, for its own sake, is presented as entertainment. This relates critically to Scannell's theorization of 'broadcast talk,' that it is designed to appeal to overhearing audiences. Media talk, therefore, has a performative, and not just a communicative function, and this is a key point for understanding how talk shows have developed in the age of television.

TV Talk Shows

The evolution of the contemporary TV talk show can be seen as a combination of two developments, both discussed by Munson. The first development was the transition of 'audience participation' (or the use of 'ordinary' people as performers) from radio to TV. There was some overlap here between developments in the USA and the UK, largely because British commercial TV in the 1950s adopted many American formats. If *Art Linkletter's House Party* was the pioneer, this soon extended to a variety of game shows, quiz programs, and proto-reality TV (programs such as *Candid Camera*). However unique to America was the *Tonight* show which introduced an interesting variant of comedy talk involving its presenter Johnny Carson mingling with the studio audience. The second development was talk radio's use of phone-in programming which became popular in the USA in the 1960s. It was in 1967 that Phil Donahue introduced a phone-in program on WLWD-TV in Dayton Ohio, which then evolved into the TV talk show *Donahue*, by all accounts the pioneer of a new TV genre. By 1980 this show was a national phenomenon, syndicated to 215 TV stations, with a particular appeal to women, and a subsequent life span that lasted until 1996.

However, even within this relatively specific definition of what constitutes a 'talk show,' it is possible to identify five variations or subgenres. These range across a spectrum and have a 20-year history in which some of the longer running programs, such as *The Oprah Winfrey Show*, exhibit cross-over features. However, it is broadly possible to differentiate TV talk shows according to types of topic, the arrangement of studio settings, and the preferred forms of talk. The starting point for this survey is what has been termed the 'audience discussion format' (Haarman, 2001), based on the American model of the 'town hall' format, still used in some televised presidential election debates. Here social issues (juvenile crime, job discrimination, and police conduct) are debated by a collective 'public' consisting of ordinary people, experts, and sometimes politicians.

Crucially, in audience discussion programs there is no stage – participants are seated in the round, with no particularly privileged place for the experts, and everyone has access to a mobile microphone controlled by the host. For some commentators (Carpignano et al., 1990: p. 49), this represents a 'crisis of theatricality' creating a new kind of 'public sphere' which transforms the traditional separation between an audience and its (on stage) representatives. What it certainly achieves is a forum in which different forms of public discourse are in circulation and sometimes in conflict. In Sonia Livingstone and Peter Lunt's (1994) account, there are two major 'epistemologies' in play, where participants demonstrate different kinds of knowledge of the issues under discussion and where, because of the setting, no particular type of knowledge is inherently superior.

In this public sphere experts and politicians speak from institutional perspectives, on professional 'footings' (Goffman, 1981), whereas ordinary people speak the language of 'commonsense.' There are two particular manifestations of this difference. The first is 'entitlement': whereas the expert has a preestablished institutional entitlement to speak (and is introduced as such), the ordinary person has to establish his/her entitlement in the act of speaking (e.g., by referring to their occupation or marital status). This then leads to a second discursive feature, the personal narrative; the knowledge of the ordinary person is grounded in experience and one of the most effective ways of communicating this is through oral narrative or 'anecdotes' (Tolson, 2006). In audience discussion shows anecdotes have a significant rhetorical effect; when told well they are dramatic extended turns at speaking which lay claim to moral truths (cf Labov, 1972).

It is a short step from this use of personal narrative to a second variant of the talk show. This is where issues are selected for discussion primarily to evoke personal perspectives and/or the display of social identities. Issues such as race and gender discrimination, sexual preferences, unconventional domestic arrangements, and so on, cross the boundaries between public and private limits. Their discussion can facilitate 'coming out' in expressions of cultural diversity and 'identity politics.' In Jane Shattuc's (1997) account of the 'Oprahification of America' this involves a kind of therapeutic discourse which is not so much the individualized psychoanalytic 'talking cure' (despite the title of her book) as the puritan-American tradition of public testimonial, or 'testifying.' This seems to have much in common with what Donal Carbaugh (1988) describes in a study of *Donahue* as 'talking American'; that is the performance of 'self' in terms of 'being honest,' self-aware, and where 'good communication' entails the 'sharing' of experiences.

A third variant of the TV talk show is where its rationale is more explicitly therapeutic. Here the focus is not on issues or identities but on the personal problems of dysfunctional, or at least unhappy, individuals. The purpose of the show is then to encourage such individuals to acknowledge and confront their problems, and to commit to change. Again Jane Shattuc is clear that this is not so much a Freudian practice (it cannot be such on a TV show) as a form of 'rational emotive therapy' which is allied to consumer-oriented lifestyle choices. A classic example of this would be a problem such as 'comfort-eating' (or similar forms of addiction) where the prescription might be a more carefully controlled diet. As regards the talk show format, this shift to the therapeutic is signaled by three generic features. Firstly, program titles are typically expressed in the first person, from the point of view of a complainant. Secondly, the role of the professional expert has shifted to that of 'after-care' specialist, where that person might be part of a backstage 'support team.' Thirdly, on stage, the defining moment of such shows occurs when the dysfunctional individual, having acknowledged his/her problem, makes a public commitment to change. Often this involves the performance of a speech act (Austin, 1962) such as a public vow or promise, which is greeted by a round of audience applause (Tolson, 2006).

Mention of the 'on stage'/'backstage' arrangements in some therapeutic talk shows heralds a fourth key development. Gradually, and despite the interesting point about the transformation of 'theatricality' made by Carpignano et al., in some TV talk shows, stage-like seating arrangements, which separate performers from the audience, have been reintroduced. In therapeutic talk shows the audience remains seated in the round and microphones are passed to members to make comments. However guests, and sometimes presenters, occupy a separate, central, studio floor. In this subgenre, presenters interview guests, to formulate their problems for subsequent audience comment. Audiences too speak the language of

therapy and express their commitment to the values of 'good communication.'

Further development occurs when theatricality becomes a show's defining feature. As Shattuc describes it, this shift occurred in the USA in the mid-1990s when TV producers discovered a youth audience for TV talk shows. Here, the fact that the problems in therapeutic talk shows were often expressed in the form of complaints (made by one partner in a relationship to another) was now taken to a further level of staged confrontations between accusers and accused. Interviewing was now a residual feature, simply a matter of introducing complaints for protagonists to pursue, to argue their differences on stage and with members of the audience. Talk shows now staged conflicts between identities that were pre-established and fixed, rather than issues to be discussed or personal problems to be resolved.

On *The Jerry Springer Show* (widely distributed in Europe as well as the USA) such confrontations escalate into physical fights, with bouncers employed to restrain participants. The verbal exchanges too seemingly break boundaries of acceptable broadcast talk, with utterances bleeped out, creating the ironic situation of a 'talk show' where some of the 'talk' cannot be heard. Two sorts of argument have been made about this. Firstly, Lunt and Stenner (2005) have suggested that *Jerry Springer* too can be understood as a kind of therapeutic situation – where the exteriorizing of emotion ('acting out') provides a release for the inarticulate. The second argument, made in Shattuc's (1997) discussion of *Riki Lake* suggests that such talk shows are a form of melodrama – a modern version of a nineteenth century popular theatrical form (vaudeville or pantomime), with excessive displays of behavioral 'flooding out' served up for audiences to react with populist moral assessments (see also Myers, 2001).

In the final variant of a TV talk show, such assessments are expressed, not by the audience, but by the host. Whereas Jerry Springer supervises proceedings from a position of ironic distance (and stands with the studio audience), Jeremy Kyle advances from the audience to the edge of the stage, to deliver moralistic harangues to his hapless guests. The 'harangue' is a belligerent face-threatening act which has the force of an absolute moral truth, and is often supported on *The Jeremy Kyle Show* by 'evidence' in the form of DNA and lie detector tests. It is the climax to an extended sequence of aggressive interrogation and is always greeted by applause and cheering from an audience which otherwise does not contribute (see Tolson, 2013). The fact that many of the 'guests' that appear on *Jeremy Kyle* are (like *The Jerry Springer Show*) recognizably from the lower working class (where DNA tests for example are used to resolve paternity issues) is a matter for discussion below; but what it illustrates for the talk show genre is a situation where 'audience participation' has been reduced to responding to a performance in which the host/presenter is the star and the collective discussion of issues or problems has been entirely displaced by authoritarian moral judgments.

Evaluations and Critical Perspectives

Together with so-called 'reality television' (and the phenomena are closely related) TV talk shows have, in the past 20 years, received extensive commentary, both journalistic and academic. Some of the journalism has focused on the 'scandalous' aspects of some talk shows (revelations about fake guests and the alleged 'effects' of talk shows on pathological social behavior – see Gamson, 1998). In the opinion of some academics (Abt and Seesholtz, 1994) this can amount to a 'crisis in the social construction of reality.' More generally, much academic commentary has not only focused on the 'blurring of boundaries' in talk shows, particularly between public and private, but also in the challenge to social norms provided by the relentless 'display of deviance.' At times, refracted through the lens of journalistic sensation, this has seemed like a kind of 'moral panic' (Thompson, 1998). However it has also been interpreted more positively as a 'postmodern' challenge to conservative social categories (Munson, 1993).

We have already encountered the idea that some talk shows might constitute an alternative or 'oppositional' public sphere. Here the argument is that previously suppressed voices are given a public hearing that challenges "the dominant ideology conveyed by television" (Carpignano et al., 1990: p. 50). In this view an alternative politics 'emanating from social, personal, and environmental concerns' (what Anthony Giddens (1991) called 'life politics') becomes conceivable. Perhaps not surprisingly, this view has been particularly popular with some feminist academics who have argued for instance that *The Oprah Winfrey Show* 'scandalizes' the dominant culture in its tolerance, and even the encouragement of 'excessive' forms of communication (Masciarotte, 1991). There has been an extensive subsequent debate between 'positive/radical' and critical feminist perspectives (summarized by Tolson, 2001) with the latter perspective arguing that the language of therapy is unable to form political perspectives that effectively challenge social oppression. For Shattuc (1997) talk shows are not feminist because they fail to deal critically with patriarchy. For Gamson (1998), who writes from a gay perspective, TV talk shows represent a 'tightrope of visibility,' where sexual nonconformity becomes public precisely in the context of its representation as 'deviance.' Whatever kind of public sphere this might be, it is fraught with contradictions.

A critical perspective on talk shows also has to recognize the contexts of its industrialized production. Here again, Jane Shattuc provides useful data. As a staple of the mainstream TV industry for 25 years their growth is clearly related to three key developments: firstly the proliferation of TV channels, secondly the introduction of continuous daytime schedules, and thirdly the significance of 'demographics' in contemporary marketing. As Shattuc makes clear, this is cheap daytime television popular with a predominantly female (conceived as 'housewife,' though more latterly also 'youth') demographic. It is used to advertise cosmetics, personal care, and domestic products. However, that context in itself does not explain the particular format of the talk show (for daytime soaps also appeal to this audience and are used to advertise similar products). To the industrial history there needs to be added a critical/generic perspective, and this is provided by John Dovey (2000).

In his book *Freakshow*, Dovey charts the development of what he calls 'first person media.' This includes camcorder and 'reality' television as well as TV talk shows (and had he been

writing 5 years later he might have included some uses of the Internet such as YouTube). Importantly, he relates these developments to 'neoliberalism,' the dominant political and cultural ideology built around the central figure of the individual consumer. Such TV programming supports neoliberalism not only by advertising consumer goods, but also in its focus on self development (the therapeutic agenda of many talk shows) and the cultivation of 'lifestyle.' This ideology is promoted not only in advertising on commercial television, but also in the talk show's focus on personal development and self-realization.

Since the 1980s there has been an international narrative to neo-liberalism, developing from the deregulation of banking and spread of financial services to increasing levels of individual consumer debt. The consequent crisis of 2008, and the shift in most Western economies toward an age of austerity, has provided a context in which the more belligerent talk shows have flourished. This is not to say that talk shows have simply reflected that context, still less that they have discussed it, but rather that their theatrical display of belligerence resonates with a neo-conservative perspective on social deviance. If credit-fuelled consumerism was partly to blame for the crisis, this was most visibly represented by an unsustainable lower working class (in the UK 'chav') consumer lifestyle; and if the dominant response was to be cuts on public spending, where better to start than from dysfunctional families living on welfare? This class fraction has become the *dramatis personae* of the melodramatic and moralistic forms of talk shows which, transmitted as they are by the more popular commercial TV stations, reproduce a populist social agenda (Tolson, 2013). Accordingly, it is instructive to reconsider what these forms of broadcasting are doing now compared to the introduction of 'ordinary people' in 1930s radio, which by coincidence, following the great depression, was also an age of 'austerity.' Whereas the ordinary person then was given a sociable voice to be celebrated; today it seems, in the authoritarian discourse of *Jeremy Kyle*, the opposite is true.

See also: Audiences, Media; Mass Media and Cultural Identity; Media Effects.

Bibliography

Abt, V., Seesholtz, M., 1994. The shameless world of Phil, Sally and Oprah: television talk shows and the destruction of society. Journal of Popular Culture 28 (1), 171–191.

Austin, J.L., 1962. How to Do Things with Words. Oxford University Press, Oxford.

Cameron, D., 2000. Good to Talk? Living and Working in a Communication Culture. Sage Publications, London.

Carbaugh, D., 1988. Talking American: Cultural Discourses on Donahue. Ablex, Norwood, NJ.

Carpignano, P., Andersen, R., Aronowitz, S., Difazio, W., 1990. Chatter in the age of electronic reproduction: talk television and the 'public mind'. Social Text 25 (6), 33–55.

Dovey, J., 2000. Freakshow: First Person Media and Factual Television. Pluto Press, London.

Gamson, J., 1998. Freaks Talk Back: Tabloid Talk Shows and Sexual Nonconformity. University of Chicago Press, Chicago.

Giddens, A., 1991. Modernity and Self Identity. Polity Press, Cambridge.

Goffman, E., 1981. Forms of Talk. University of Pennsylvania Press, Philadelphia.

Haarman, L., 2001. Performing talk. In: Tolson, A. (Ed.), Television Talk Shows: Discourse, Performance, Spectacle. Lawrence Erlbaum Associates, Mahwah, NJ.

Labov, W., 1972. Language in the Inner City. University of Pennsylvania Press, Philadelphia.

Livingstone, S., Lunt, P., 1994. Talk on Television: Audience Participation and Public Debate. Routledge, London.

Lunt, P., Stenner, P., 2005. The Jerry Springer Show as an emotional public sphere. Media, Culture & Society 27, 59–81.

Masciarotte, G.-J., 1991. C'mon girl: Oprah Winfrey and the discourse of feminine talk. Genders 11, 81–110.

Munson, W., 1993. All Talk: The Talk Show in Media Culture. Temple University Press, Philadelphia.

Myers, G., 2001. I'm out of it; you guys argue: making an issue of it on the Jerry Springer Show. In: Tolson, A. (Ed.), Television Talk Shows: Discourse, Performance, Spectacle. Lawrence Erlbaum Associates, Mahwah, NJ.

Scannell, P., 1996. Radio and Television in Modern Life. Basil Blackwell, Oxford.

Scannell, P., Cardiff, D., 1991. A Social History of British Broadcasting. In: Serving the Nation, vol. 1. Basil Blackwell, Oxford.

Shattuc, J., 1997. The Talking Cure: TV Talk Shows and Women. Routledge, New York.

Thompson, K., 1998. Moral Panics. Routledge, London.

Tolson, A., 2001. Talking about talk: the academic debates. In: Tolson, A. (Ed.), Television Talk Shows: Discourse, Performance, Spectacle. Lawrence Erlbaum Associates, Mahwah, NJ.

Tolson, A., 2006. Media Talk: Spoken Discourse on TV and Radio. Edinburgh University Press, Edinburgh.

Tolson, A., 2013. Moments of truth: telling it like it is on The Jeremy Kyle Show. In: Lorenzo-Dus, N., Garces-Conejos Blitvitch, P. (Eds.), Real Talk: Reality Television and Discourse Analysis in Action. Palgrave Macmillan, Maidenhead, UK.

Mediating Variable

David P MacKinnon, Arizona State University, Tempe, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A mediating variable is intermediate in the causal sequence relating an independent variable to a dependent variable, such that the independent variable causes the mediating variable that causes the dependent variable. Theoretical mediating variables are common in the social and behavioral sciences and several examples are given in this articles. The major approaches to investigating mediation with statistical methods and experimental designs are outlined. New methods to investigate the role of confounding variables and approaches to improve causal inference from mediation studies are described. A program of experimental, statistical, and qualitative research for the investigation of mediating variables is emphasized.

A mediating variable explains and identifies the causal process underlying the relationship between two other variables. A mediating variable (M) is intermediate in the causal sequence relating an independent variable (X) to a dependent variable (Y), such that the independent variable causes the mediating variable and that, in turn, causes the dependent variable. Mediating variables are known as intervening variables or intermediate variables because they come between the independent and the dependent variables. Important aspects of mediating variables are their close link with theory and the potential that the mediating variables identified in one context may operate in a wide variety of contexts. This article starts with examples of mediating variables, outlines tests for mediating variables, and describes limitations and extensions of mediating variable models.

The Mediating Variable Effect

When quantified, the mediating variable effect is known as the mediated or indirect effect. It is called the indirect effect because it represents an effect of X on Y that is transmitted indirectly through the mediating variable. If some, but not all, of the effect of X on Y is transmitted through M, the effect is partially mediated because a direct effect of X on Y exists even after adjustment for the mediator.

The mediating variable effect differs from other third-variable effects used to understand the relationship between two variables. For a moderator or interaction variable, the relation of X on Y differs for different values of the moderator variable. Unlike a mediating variable, a moderator does not transmit the effect of X on Y. A confounding variable, when included in the investigation of the association between X and Y, changes the association between X and Y. A confounder changes the association between X and Y because it is related to both X and Y but not because it is in a causal sequence relating X to Y. Although the conceptual distinction between a confounder and a mediator is clear, it can be difficult to differentiate between them with actual data. The underlying causal sequence X to M to Y is the important aspect of the mediating variable.

Examples of Mediating Variables

Because the mediating variable elaborates the relationship between two variables by explaining why or how an effect occurs, it is of considerable importance in many fields. One of the first applications of the mediating variable was in psychological theories for how a stimulus is translated into a response. For example, habit strength (Hull, 1943) or appetite (Tolman, 1938) may intervene between stimulus of a maze (X) and response time (Y) to complete the maze. Later, a distinction was made between hypothetical mediating constructs representing processes or entities not directly observed and intervening variables that were observed measures of these hypothetical constructs (MacCorquodale and Meehl, 1948). Duncan (1966) established the use of mediating variables in sociology when he applied path analysis techniques developed earlier by Wright (1934) to examine causal mediating relations among variables. One of the first sociological examples was that father's socioeconomic status causes the mediating variable, child's educational achievement, which causes child's socioeconomic status. The effect of father's socioeconomic status on child's socioeconomic status was not entirely mediated by child's educational achievement, so a direct effect between father's and child's socioeconomic status also was included in the model. Decomposition of effects into direct and indirect effects is a common focus of sociological research.

Another application of the mediating variable is in chronic disease research (Schatzkin et al., 1990). Prospective studies examining the effects of etiological factors or prevention strategies on chronic disease require lengthy follow-up measurements because of the low frequency of occurrence and the slow development of the disease. As a result, researchers attempt to identify mediating variables on the causal pathway relating risk factors to disease. The mediating variables are called intermediate or surrogate endpoints. Using colon cancer as an example, the proliferation of epithelial cells in the large bowel occurs before colon cancer and is related causally to colon cancer. A study targeting cell proliferation rather than colon cancer requires less time and fewer subjects. The use of these endpoints as a surrogate for the ultimate endpoint is an active area of research with unique substantive and statistical aspects. The focus of surrogate endpoint research is typically a single intermediate variable rather than the multiple mediators common in sociology and other areas of social science. It often is expected that the relation between the surrogate and ultimate outcome is close to perfect.

Mediating variables are critical in the development and application of treatment and prevention programs. Programs

 http://dx.doi.org/10.1016/B978-0-08-097086-8.44037-7

are designed to change mediating variables hypothesized to be related causally to the dependent variable. It is assumed that the program causes change in the mediating variable, which, in turn, causes change in the dependent variable. Programs to prevent coronary heart disease target mediating behaviors, such as diet and smoking, and biological factors, such as cholesterol level and blood pressure. Social influence-based drug prevention programs are designed to increase skills to resist drug offers and engender norms less tolerant of drug use. Treatment programs for substance abuse target mediators, such as communication skills and social support, to prevent a relapse to drug abuse. Tobacco treatment programs seek to reduce craving for nicotine to reduce remission. Cognitive behavioral approaches to depression treatment focus on cognitive reinterpretation of life events.

Theories of health behavior and empirical relationships from prior research guide the selection of mediating variables for treatment and prevention programs. The selection of mediators based on their relation to the outcome variable has been called the conceptual or etiological theory of an intervention. An intervention based on established theory regarding mediating variables may be more likely to change the outcome measure, and the results provide a test of the theoretical basis of the intervention. Competing theories of the onset of drug abuse, for example, may suggest alternative mediators that can be tested in an experimental design. Prevention and treatment programs will cost less and have greater benefits if effective and ineffective mediating processes are identified. Unlike prevention, which serves to prevent problem behavior from occurring, treatment programs target mediators to stop a problem behavior from reoccurring. As a result, mediating variables selected for treatment may be different from those selected for prevention for several reasons. The delivery of prevention is typically in large groups of people in contrast to treatment that is delivered individually or in small groups. The source of treatment may include therapist actions, actions conducted by the client alone, actions conducted by the client as part of a group, pharmaceutical treatments, or changes in environment.

Mediating variables in prevention and treatment serve a different purpose than other applications of mediating variable methodology. In prevention and treatment, mediating variables are selected before the study and are fundamental to the prevention program because of their causal relationship to the dependent variable. In most other applications, the purpose of the mediating variable is to identify the processes that generated an effect, after the effect has been found. The approach of finding an explanation of an observed relation between variables can be traced to the work of Hyman (1955) and Lazarsfeld (1955). The design of interventions to target mediating variables is more recent and combines theory for which mediators are related to the outcome with theory for actions that will change the mediators as part of the intervention. Before an intervention study is conducted, researchers select mediators that are malleable during the intervention period that also are related causally to the outcome variable. In some cases, a mediator with a weaker relation to the outcome variable may be selected because it is easier to change by actions of the intervention.

Tests of the Mediated Effect

The single mediator model with symbols to represent relationships between variables is shown in Figure 1. The parameter estimates and standard errors from the following three equations provide the information for tests of the mediated effect (MacKinnon and Dwyer, 1993):

$$Y = \tau X + \varepsilon_1 \quad [1]$$

$$M = \alpha X + \varepsilon_2 \quad [2]$$

$$Y = \tau' X + \beta M + \varepsilon_3 \quad [3]$$

where Y is the dependent variable; X is the independent variable; M is the mediating variable; τ codes the relationship between the independent variable and the dependent variable; τ' is the coefficient relating the independent variable to the dependent variable adjusted for the effects of the mediating variable; α is the coefficient relating the independent variable to the mediating variable; β is the coefficient relating the mediating variable to the dependent variable adjusted for the independent variable; and ε_1, ε_2, and ε_3 code unexplained variability. The intercepts are not included to simplify the presentation. Two estimators of the mediated effect, $\alpha\beta$ and $\tau - \tau'$, are algebraically equivalent in ordinary regression but not in other analyses, such as multilevel and logistic regression.

Three major types of tests of the mediated effect use the information in the above regression models: (1) causal step tests, (2) difference in coefficients tests, and (3) product of coefficients tests. Methods to assess mediation based on causal steps entail tests of the different logical relationships among the three variables involved that must be true for a variable to be a mediator. The following sequence of causal steps described in Baron and Kenny (1986) is the most widely used method to assess mediation: (1) The independent variable (X) must affect the dependent variable (Y), τ in eqn [1], (2) the independent variable (X) must affect the mediator (M), α in eqn [2], and (3) the mediator must affect the dependent variable (Y) when the independent variable (X) is controlled, β in eqn [3]. The conceptual links between each necessary causal relationship and the statistical tests are clear in the causal step method. The causal step method, however, has no direct estimate of the mediated effect and standard error needed to construct confidence limits. The first requirement, a significant relationship between the independent and dependent variables excludes models where mediation exists, but the relationship between

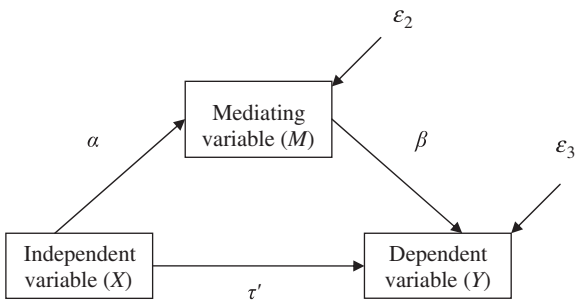

Figure 1 The mediating variable model.

the independent variable and the dependent variable is not significant. The first requirement is consistent with the origination of this approach in the explanatory model described by Hyman (1955) and Lazarsfeld (1955) with the goal of explaining an observed relation among variables.

The second method to test for mediation compares the relationship between the independent variable and the dependent variable before and after adjustment for the mediator. The method tests whether a third variable, here a mediator, significantly changes the relationship between two variables. The difference in the regression coefficients ($\tau - \tau'$) is an example of this approach. Formulas for the standard error of $\tau - \tau'$ can be applied to construct confidence limits for the mediated effect (Clogg et al., 1992). A drawback of the change in coefficient method is that it is conceptually more similar to a confounding variable than to a mediating variable.

The third method to test the significance of the mediated effect is based on the product of coefficients, which is more consistent with the causal sequence in mediation. The estimator of the mediated effect is $\alpha\beta$, the product of regression coefficients α and β. The most commonly used standard error is the first-order Taylor series solution for the product of two random variables derived by Sobel (1982) using the multivariate delta method, where σ_α and σ_β are the standard errors α and β, respectively:

$$\sigma^2_{\alpha\beta} = \alpha^2\sigma^2_\beta + \beta^2\sigma^2_\alpha \qquad [4]$$

The matrix formulas for the computation of this standard error are included in most covariance structure analysis software programs. Confidence limits for the mediated effect can be calculated using the standard error in eqn [4], and the result is then compared to a standard normal distribution to test for significance. The mediated effect divided by its standard error, $\alpha\beta/\sigma_{\alpha\beta}$, however, rarely follows a normal distribution. Methods based on the distribution of the product of two random variables provide more accurate confidence intervals and significance testing than assuming a normal distribution. Methods based on resampling, such as the percentile bootstrap, also provide more accurate tests because they do not require that the mediated effect has a normal distribution.

The product of coefficients methods provides an estimate of the mediated effect and the standard error of the mediated effect. In addition, the underlying model is a mediation model in which the mediated effect is the product of coefficients hypothesized to measure causal relationships. For an independent variable coding experimental assignment, the α estimate tests whether the manipulation successfully changed the mediating variable it was designed to change, and the β estimate tests whether the mediating variable is related to the dependent variable, as suggested by theory. There are several important assumptions of the product of coefficients method, including reliable and valid measures and the usual linear regression assumptions of independence of errors across equations. The interaction of X and M in eqn [2] could be added, which tests whether the relation of M and Y changes across levels of X. The test evaluates the extent to which the mediation process differs as a function of the value of the mediator (MacKinnon, 2008). The method also assumes that causal relations among variables are accurate, which is an active area of recent research, as described in the next section.

Causal Analysis of Mediating Variables

Methods based on the regression approach described thus far have been criticized based on causal analysis of the relationships among variables. For example, if X, M, and Y are measured simultaneously, other models (e.g., X is the mediator of the M to Y relationship or M and Y both cause X) would explain the data equally well, and it is not possible to distinguish these alternatives without more information (Spirtes et al., 1993). The case in which X represents random assignment to conditions improves causal interpretation of mediating variables (Holland, 1988; Robins and Greenland, 1992). Holland applied Rubin's (1974) causal model to a design in which students are randomized to one of two groups, either to a group receiving motivation to study or to a control group that did not receive motivation. The mediating process is that assignment to the motivation group affects the number of hours studied, which affects test performance. Under some assumptions, the typical regression coefficient for the group effect on test score, τ, and the group effect on number of hours studied, α, are valid estimators of the causal effect, primarily because of the randomization of units to treatment. The relationship between the mediating variable of the number of hours studied and test score is more problematic (e.g., Y may cause M) and the regression coefficient β is not an accurate estimator of the causal effect because this relationship is correlational, not the result of random assignment. The estimator τ' is not an accurate causal estimator of the direct effect. The missing information for the causal effects is whether the relationship between the number of hours studied and test score would have been different for subjects in the treatment group if they had instead participated in the control group. The problem with causal estimation of mediation effects described by Holland has been more formally described as an assumption of sequential ignorability. Sequential ignorability is the assumption that omitted variables do not bias the X to M relation and omitted variables do not bias the X to Y relation (Imai et al., 2010; Lynch et al., 2008; ten Have et al., 2007). With random assignment, omitted variable bias is balanced for the relation of X to M so that α is an estimator of a causal effect. The influence of omitted variables on the regression coefficient β cannot be interpreted as a causal effect even when X is randomized, because the value of M is not randomized to individuals but rather is self-selected by participants. The estimator, τ', is also not an accurate estimator of the causal effect of X on Y, controlling for M for the same reason. The limitations of interpreting the β coefficient as a causal effect have been described in several different substantive areas but with somewhat different nomenclature (Robins and Greenland, 1992; MacKinnon, 2008; Winship and Morgan, 1999). For example, omitted variable bias is the term used in the social sciences, while epidemiology and medicine call this confounder bias.

Holland (1988) has described a solution to estimate the causal effect of M on Y using an instrumental variables approach to estimating the β path based on assumptions of additivity, linear, and constant effects. A classic reference on the instrumental variable method is Angrist et al. (1996) who investigated the effect of Vietnam War service on health with random selection in the draft as the independent variable, serving in

Vietnam as the mediating variable, and health as the dependent variable. In the instrumental variable model, random selection in the draft (X) was used to predict serving in Vietnam (M), and the predicted values of war service were used to predict Y. Several assumptions required for the b path to have a causal interpretation often make it difficult to successfully apply instrumental variable methods (MacKinnon, 2008; Shadish et al., 2002). One assumption is that randomization of X leads to changes in M such that the stronger the relation of X to M the better the instrument; the ideal instrument has a correlation of one for X with M, which makes X indistinguishable from M. Because M can be considered randomized from its prediction by X, the influence of any omitted variable on the relation of M to Y is removed. A second important assumption, the exclusion restriction, is that M completely mediates the effect of X on Y such that when M is considered there is no relationship between X and Y. Complete mediation, however, may be unrealistic with real data from many research areas. The assumptions for the instrumental variable approach for mediation analysis were clarified and extended by Sobel (2008).

Another important method to address limitations of the mediation model is based on the classification of different possible response patterns for how X affects M and M affects Y. In principal stratification approaches, subsets of persons are specified based on their true relation for how M and Y could change in response to treatment X or if no treatment is received. For example, in a two-group treatment versus control study, there are four different types of hypothetical responses across X to M to Y as described by Jo (2008): (1) never-improvers or people whose mediator would not improve if they were in either the treatment or the control group; (2) forward-improvers, people whose mediator would improve only if they received the treatment; (3) backward-improvers, people whose mediator would improve only if they were in the control group; and (4) always-improvers, people whose mediator would improve if they were in either treatment group. These four types of people are determined outside the experiment so their classification is independent of their experimental assignment and as a result the omitted variables that confound the M to Y relation are ignorable. In most studies using principal stratification, a monotonicity assumption is made such that there are no participants for whom the mediator would improve for participants in the control condition because they are unexposed to the treatment strategies to improve the mediator. Given the three possible stratifications, the indirect effect is estimated by comparing means within and between these stratifications (Frangakis and Rubin, 2002; Jo, 2008). In these models, covariates often are used to provide information about the stratifications to identify parameters in the model, including propensity score methods to help define stratifications (Jo et al., 2011; Stuart et al., 2008).

The counterfactual interpretation of the mediation model has led to an important distinction between controlled and natural effects. Controlled effects are effects fixed to certain values, and natural effects are effects at observed or natural values obtained by research participants (Pearl, 2001, 2012; Robins and Greenland, 1992). Research questions may focus on a controlled effect such an effect that occurs when changing from no intervention to an intervention condition. There is no direct control of the mediator in most research studies so it is necessary to estimate effects at the natural value of the mediator. As a result, the natural indirect effect is the indirect effect at the value of M that the participant would have had if that participant was in the control condition and not exposed to an intervention. For the case of ordinary least squares regression, the controlled and natural direct effects are equivalent and provide a causal estimate of the indirect effect under the assumption of sequential ignorability (Imai et al., 2010; Pearl, 2001). For nonlinear models, such as logistic regression, the controlled and natural effects may differ at different values of variables in the model. Several alternative approaches are available to evaluate direct and indirect effects in nonlinear models, including a general mediation formula that is applicable across statistical models (Pearl, 2010).

Recent work has provided methods to evaluate the sensitivity of research results to violation of model assumptions, such as sequential ignorability. Sensitivity analyses for measurement error may be conducted by changing values of the error variance in a structural equation model and reestimating model estimates (MacKinnon, 2008). One way to investigate the sensitivity of model results to omitted variables is to assess how large a confounder effect on the M to Y relation must be to invalidate conclusions about mediation (Frank, 2000; Li et al., 2007; Lin et al., 1998; Rosenbaum, 2002). Another way to conduct sensitivity analyses for mediation would be to systematically increase the correlation between the errors in eqns [2] and [3], and evaluate how much the indirect effect estimates change when omitted variables related to both X and M would affect the results as described in Imai et al. (2010). VanderWeele (2008, 2010) has formalized methods to probe bias in mediation relations based on two parameters. One parameter corresponds to the relation of an unobserved binary confounder to Y. The second parameter corresponds to the difference in prevalence of the confounder variable, U, for people with the same value of M in the treatment and control groups. The extent to which a confounder affects the indirect effect estimate can be obtained by subtracting the biased value from the estimate of the indirect effect.

Historically, the method used to identify mediating processes is a program of experimental research designs in which each experiment provides increasing evidence to support a mediation relation. It is rare that one or even a few experiments provide enough information to demonstrate a true mediation relation (MacKinnon, 2008; Mark, 1986; West and Aiken, 1997). There are several experimental designs to assess mediation (MacKinnon, 2008; MacKinnon et al., 2010). The blockage design consists of an intervention to block the mediation process so that a relation observed in the usual condition, but not in the blockage condition, supports a mediating process. In an enhancement design, an experimental condition is designed to enhance the mediation process so that participants in this group would experience the greatest mediated effect. Another type of experimental design that has received more attention recently is the double randomization design, which uses one randomized study to evaluate the X to M relation and a second randomized study is used to evaluate the M to Y relation adjusting for X (MacKinnon, 2008; MacKinnon et al., 2010; Spencer et al., 2005). Other experimental designs focus on testing the consistency and specificity of mediation relations across different contexts, subgroups, and

measures of the mediating and outcome variables (MacKinnon et al., 2010). Even in experimental designs, estimation of the mediated effect is important. New developments in causal inference address how designs such as those described thus far may be combined with modern causal methods to obtain more accurate evidence for indirect effects (Imai et al., 2013). In particular, Imai et al. (2013) described several designs to overcome the sequential ignorability assumption.

More Complicated Mediation Models

The single mediator model is easily expanded to include a chain of mediating variables. In fact, most mediating variables are actually part of longer theoretical mediational chain (Cook and Campbell, 1979). For example, it is possible to measure each of the four constructs in a theoretical chain from exposure to a prevention program, to comprehension of the program, to short-term attitude change, to change in social norms, and to change in the dependent variable. Typically, researchers measure an overall social norms mediator rather than all mediators in the chain, even though a more detailed chain is theorized. It is a primary task of a mediation study to specify the chain of mediation to be examined, including mediating variables that are not so close theoretically to X and Y that they are the same variable as X and Y.

Longitudinal data provide evidence for temporal relations among variables. More detailed evidence for mediation can be obtained when X, M, and Y are measured repeatedly. For example, evidence for temporal relations among variables may be obtained by demonstrating that change in the mediator precedes and is related to subsequent change in the dependent variable (Cheong et al., 2003; Cole and Maxwell, 2003; MacKinnon, 2008). Longitudinal data also allow researchers to investigate mediation for within-person changes as well as between-person differences. For example, mediation can be modeled as X causes changes in M between baseline and the first follow-up, and the change in M between baseline and the first follow-up can be used to predict change in Y between the second and third follow-up, providing longitudinal evidence for mediation. The planning of longitudinal studies of mediation have the same challenges as other longitudinal studies, especially the determination of the optimal timing of when variables are related to each other and how to measure variables so that they have the best chance to detect these temporal relations. There are several different types of longitudinal models, including autoregressive models for relations between waves of measurement, with patterns consistent with measures at earlier times predicting a mediator at a later time that then predicts an outcome variable at a later time. Latent growth curve methods model growth in variables over time and relate the growth over time in the mediator and outcome. Latent change score models specify change models between waves and are particularly useful for modeling change at earlier waves on change at later waves. Another important longitudinal modeling approach attempts to classify people based on their pattern of change over time into patterns that are consistent or inconsistent with a mediation process (von Eye et al., 2009). Other approaches attempt to relate variables in the mediation model in continuous time.

Models with both mediators and moderators are often relevant in many research areas. For example, a mediation process may differ for males and females or may differ across ages. Models that combine moderation and mediation investigate the mediation process as well as differences in the process across these groups. The first moderation and mediation model would include the interaction between X and M in the eqn [2] such that the relation of M to Y differs across levels of Y. This interaction suggests that the mediated effect is different at different values of M and is an important interaction that can be routinely tested using the X and M interaction in eqn [2]. Other more complicated models with moderation and mediation are possible as well.

The single mediator methods can be extended for multiple mediators and multiple outcomes with correspondingly more mediated effects (Bollen, 1987). Multiple mediator models are justified because most independent variables have effects through multiple mediating processes. The true causal relationships are difficult to disentangle in this model because of the number of alternative relationships among variables. One solution to the problems inherent in the causal interpretation of more complex, as well as single mediator models, is to view the identification of mediating variables as a sustained research effort requiring a variety of experimental and nonexperimental approaches to identify mediating variables. The analysis of multiple mediators in one study informs the design of randomized experiments to contrast alternative mediating variables, leading to a refined understanding of mediating processes (MacKinnon, 2008; West and Aiken, 1997). Meta-analytical studies provide information about the consistency of mediating variable effects across many situations (Cook et al., 1992). Furthermore, the identification of mediating variables requires examination of additional sources, including ethnographic, historical, and clinical information (Cronbach, 1982).

Future Directions

Mediating variables will continue to play a major role in the social and behavioral sciences because of the need to understand how and why variables are related. In particular, the practical and theoretical benefits of mediating variables in prevention and treatment research should guide the development of effective interventions. Investigators will endeavor to find the information needed for the further application of detailed causal models outlined by Rubin (1974) and Pearl (2010) and related approaches to mediating variable models. Likely these causal inference approaches will be applied to more complicated mediation models, including experimental designs to investigate mediation. Accurate point and interval estimators of mediated effects will continue to be developed for various statistical methods, including categorical and longitudinal models. In addition to statistical approaches, sustained study of mediating variables will include information from a variety of sources, including history, journalism, and clinical experience. Such comprehensive efforts are necessary to determine whether a variable is truly intermediate in the causal sequence between two other variables.

See also: Field Observational Research in Anthropology and Sociology.

Bibliography

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables (with discussion). Journal of the American Statistical Association 91, 444–472.

Baron, R.M., Kenny, D.A., 1986. The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology 51, 1173–1182.

Bollen, K.A., 1987. Total direct and indirect effects in structural equation models. In: Clogg, C.C. (Ed.), Sociological Methodology. American Sociological Association, Washington DC, pp. 37–69.

Cheong, J., MacKinnon, D.P., Khoo, S.T., 2003. Investigation of mediational processes using parallel process latent growth curve modeling. Structural Equation Modeling 10, 238–262.

Clogg, C.C., Petkova, E., Shihadeh, E.S., 1992. Statistical methods for analyzing collapsibility in regression models. Journal of Educational Statistics 17 (1), 51–74.

Cole, D.A., Maxwell, S.E., 2003. Testing mediational models with longitudinal data: questions and tips in the use of structural equation modeling. Journal of Abnormal Psychology 112, 558–577.

Cook, T.D., Campbell, D.T., 1979. Quasi-experimentation: Design and Analysis Issues for Field Settings. Rand McNally, Chicago, IL.

Cook, T.D., Cooper, H., Cordray, D.S., Hartmann, H., Hedges, L.V., Light, R.J., Louis, T.A., Mosteller, F., 1992. Meta-analysis for Explanation. Sage, New York.

Cronbach, L.J., 1982. Designing Evaluations of Educational and Social Programs. Jossey-Bass, San Francisco, CA.

Duncan, O.D., 1966. Path analysis: sociological examples. American Journal of Sociology 72, 1–16.

Frangakis, C.E., Rubin, D.B., 2002. Principal stratification in causal inference. Biometrics 58, 21–29.

Frank, K.A., 2000. Impact of a confounding variable on a regression coefficient. Sociological Methods and Research 29, 147–194.

Holland, P.W., 1988. Causal inference, path analysis, and recursive structural equation models. Sociological Methodology 18, 449–484.

Hull, C.L., 1943. Principles of Behavior. Appleton-Century, New York.

Hyman, H.H., 1955. Survey Design and Analysis: Principles, Cases and Procedures. The Free Press, Glencoe, IL.

Imai, K., Keele, L., Yamamoto, T., 2010. Identification, inference, and sensitivity analysis for causal mediation effects. Statistical Science 25, 51–71.

Imai, K., Tingley, D., Yamamoto, T., 2013. Experimental designs for identifying causal mechanisms. Journal of the Royal Statistical Society: Series A (Statistics in Society) 176, 5–51.

Jo, B., 2008. Causal inference in randomized experiments with mediational processes. Psychological Methods 13, 314–336.

Jo, B., Stuart, L., MacKinnon, D.P., Vinokur, A., 2011. The use of propensity scores in mediation analysis. Multivariate Behavioral Research 46, 425–452.

Lazarsfeld, P.F., 1955. Interpretation of statistical relations as a research operation. In: Lazarsfeld, P.F., Rosenberg, M. (Eds.), The Language of Social Research: A Reader in the Methodology of Social Research. Free Press, Glencoe, IL, pp. 115–125.

Li, Y., Bienias, J., Bennett, D., 2007. Confounding in the estimation of mediation effects. Computational Statistics and Data Analysis 51, 3173–3186.

Lin, D.Y., Psaty, B.M., Kronmal, R.A., 1998. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics 54, 948–963.

Lynch, K.G., Cary, M., Gallop, R., ten Have, T.R., 2008. Causal mediation analyses for randomized trials. Health Services and Outcome Methodology 8, 57–76.

MacCorquodale, K., Meehl, P.E., 1948. On a distinction between hypothetical constructs and intervening variables. Psychological Review 55, 95–107.

MacKinnon, D.P., 2008. Introduction to Statistical Mediation Analysis. Erlbaum, Mahwah, NJ.

MacKinnon, D.P., Dwyer, J.H., 1993. Estimating mediated effects in prevention studies. Evaluation Review 17, 144–158.

MacKinnon, D.P., Pirlott, A.G., Moser, S., 2010. The Unbearable Lightness of B: Approaches to Improve Causal Interpretation of the M to Y Relation. Symposium presentation at the Society for Personality and Social Psychology Conference, Las Vegas, NV.

Mark, M.M., 1986. Validity typologies and the logic and practice of quasi-experimentation. In: Trochim, W.M.K. (Ed.), Advances in Quasi-experimental Design and Analysis. Jossey-Bass, San Francisco, CA, pp. 47–66.

Pearl, J., 2001. Direct and Indirect Effects. Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, CA. pp. 411–420.

Pearl, J., 2010. The foundations of causal inference. Sociological Methodology 40, 75–149.

Pearl, J., 2012. The causal mediation formula: a guide to the assessment of pathways and mechanisms. Prevention Science, in press 13, 426–436.

Robins, J.M., Greenland, S., 1992. Identifiability and exchangeability for direct and indirect effects. Epidemiology 3, 143–155.

Rosenbaum, P.R., 2002. Observational Studies. Springer, New York.

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688–701.

Schatzkin, A., Freedman, L.S., Schiffman, M.H., Dawsey, S.M., 1990. Validation of intermediate endpoints in cancer research. Journal of the National Cancer Institute 82, 1746–1752.

Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and Quasi-experimental Designs for Generalized Causal Inference. Houghton Mifflin Company, Boston, MA.

Sobel, M.E., 1982. Asymptotic confidence intervals for indirect effects in structural equation models. In: Leinhardt, S. (Ed.), Sociological Methodology, 1982. American Sociological Association, Washington, DC, pp. 290–312.

Sobel, M.E., 2008. Identification of causal parameters in randomized studies with mediating variables. Journal of Educational and Behavioral Statistics 33, 230–251.

Spencer, S.J., Zanna, M.P., Fong, G.T., 2005. Establishing a causal chain: why experiments are often more effective than meditational analyses in examining psychological processes. Journal of Personality and Social Psychology 89, 845–851.

Spirtes, P., Glymour, P., Scheines, R., 1993. Causation, Prediction, and Search. Springer-Verlag, New York.

Stuart, E.A., Perry, D.F., Le, H., Ialongo, N.S., 2008. Estimating intervention effects of prevention programs: accounting for noncompliance. Prevention Science 9, 288–298.

ten Have, T.R., Joffe, M.M., Lynch, K.G., Brown, G.K., Maisto, S.A., Beck, A.T., 2007. Causal mediation analysis with rank preserving models. Biometrics 63, 926–934.

Tolman, E.C., 1938. The determiners of behavior at a choice point. Psychological Review 45, 1–41.

VanderWeele, T.J., 2008. Simple relations between principal stratification and direct and indirect effects. Statistics and Probability Letters 78, 2957–2962.

VanderWeele, T.J., 2010. Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiology 21, 540–551.

von Eye, A., Mun, E.Y., Mair, P., 2009. What carries a mediation process?: configural analysis of mediation. Integration of Psychology and Behavior 43, 228–247.

West, S.G., Aiken, L.S., 1997. Toward understanding individual effects in multiple component prevention programs: design and analysis strategies. In: Bryant, K., Windle, M., West, S.G. (Eds.), The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research. American Psychological Association, Washington, DC, pp. 167–209.

Winship, C., Morgan, S.L., 1999. The estimation of causal effects from observational data. Annual Review of Sociology 25, 659–707.

Wright, S., 1934. The method of path coefficients. Annals of Mathematical Statistics 5, 161–215.

Mediation, Arbitration, and Alternative Dispute Resolution (ADR)

Carrie J Menkel-Meadow, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article defines and describes modern processes of dispute resolution beyond court adjudication, including negotiation, mediation (facilitated negotiation), arbitration (decisions by party chosen private dispute resolvers), and a variety of new hybrid forms of dispute resolution (e.g., med-arb, summary jury trials, public policy consensus building) that are used in both public and private disputes. The article reviews the history and theory behind these processes, outlining quantitative and qualitative reasons for their use, and then reviews a variety of controversies associated with their use, including the privatization of dispute resolution, difficulty in evaluating their effectiveness, and power imbalances in their use. The aspirational qualities of process pluralism (peaceful, better quality, and consent-based resolution of conflicts) are contrasted to the realities of their use and cooptation in some contexts, such as when powerful private and public parties dictate the form that dispute resolution takes, without true party consent.

Mediation, arbitration, and alternative dispute resolution (ADR) are processes used to resolve disputes, either within or outside the formal legal system, without adjudication or decision by a judge. More recently, the terms 'appropriate dispute resolution' and 'process pluralism' have been used to express the idea that different kinds of disputes, variable by subject matter type, parties involved, or location of the dispute or transaction, may require different kinds of processes – no one legal or informal dispute process can serve for all human disputing. Mediation is a process in which a third party (usually neutral and unbiased) facilitates a negotiated consensual agreement among parties, without rendering a formal decision. In arbitration, a single third party or a panel of arbitrators, most often chosen by the parties themselves, renders a decision, in terms less formal than a court, but often with a written award. ADR includes a variety of hybrid processes built on the basic processes of negotiation, mediation, and arbitration, now including med-arb, minitrials, summary jury/judge trials, and early neutral evaluation, which vary in providing opportunities for settlement of cases, advisory opinions, or decisions in both public (court-annexed) and private settings. This article defines the basic ADR terms and reviews the theory, history, controversies, and future of ADR processes.

As noted below, the full panoply of processes denominated under the rubric of ADR now includes a variety of primary and hybrid processes, with elements of dyadic negotiation, facilitative, advisory, and decisional action by a wide variety of third party neutrals, sometimes combined with each other to create new formats of dispute processing (*see* Negotiation and Bargaining: Role of Lawyers; International Arbitration; Litigation; Courts and Adjudication; Disputes, Social Construction and Transformation of; Legal Systems: Private; Lex Mercatoria; Legal Pluralism; Lawyers: Social Organization of the Profession; Judges; Paralawyers: Other Legal Occupations).

Definitions and Types of Processes

In an era characterized by a wide variety of processes for resolving disputes among individuals, organizations, and nations, process pluralism has become the norm in both formal disputing systems, like legal systems and courts, and in more informal, private settings, as in private contracts and transactions, family disputes, and internal organizational grievance systems. There are a number of factors that delimit the kinds of processes which parties may choose or may be ordered to use under rules of law, court, or contract.

The primary processes consist of individual action (self-help, avoidance), dyadic bargaining (negotiation), third-party-facilitated approaches (mediation), and third-party decisional formats (arbitration and adjudication). Hybrid or secondary processes combine elements of these processes and include med-arb (facilitated negotiation followed by decision), minitrials (shortened evidentiary proceedings followed by negotiation), summary jury/judge trials (use of mock jurors or judges to hear evidence and issue advisory verdicts to assist in negotiation, often conducted within the formal court system), and early neutral evaluation (third parties, usually lawyers or other experts, who hear arguments and evidence, and advise about the issues or values of the dispute, for purposes of facilitating a settlement or structuring the dispute process). Increasing judicial involvement in dispute settlement suggests that judicial, and often mandatory, settlement conferences are another form of hybrid dispute mechanism. Retired judges provide a hybrid form of arbitration or adjudication in private 'rent-a-judge' schemes that are sometimes authorized by particular states (in the United States, e.g., California provides for this process in its state constitution). In other countries, it is becoming more common for retired judges to offer these services, usually for private fees.

Dispute processes are also characterized by the extent to which they are voluntary and consensual (whether in predispute contract agreements, ADR *ex ante*, or voluntarily undertaken after the dispute ripens, ADR *ex post*), or whether they are mandated (by a predispute contract commitment) or by court rule or referral. The ideology that contributed to the founding of modern mediation urges that mediation should be entered into voluntarily and all agreements should be arrived at consensually (Menkel-Meadow, 1995a). Nevertheless, as courts have increasingly sought to 'manage' or reduce their caseloads and have looked to ADR processes as a means of diverting cases to other fora, even mediation may be mandated,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86083-3

although it is usually participation in, not substantive agreement, that is required.

The taxonomy of different dispute processes also differentiates between binding and nonbinding processes. Arbitration, for example, can be structured either way. Under some contractual and statutory schemes (such as the American Federal Arbitration Act), decisions by private arbitrators are final and binding on the parties, and subject to very limited court review, including only such claims as fraud, corruption of the arbitrator, or, in a few jurisdictions, serious errors of law or extreme 'miscarriages of justice.' Nonbinding processes, including nonbinding decisions in some arbitrations, allow appeals or follow-through to other processes, such as mediation or full trial. Many court-annexed arbitration programs, for example, allow a *de novo* trial following an arbitration if one party seeks it, often having to post a bond or deposit for costs. The process of mediation itself is nonbinding; as it is a consensual process, a party may exit at any time. On the other hand, once an agreement in mediation is reached, a binding contract may be signed, which will be enforceable in a court of law.

Finally, dispute processes are often subject to different requirements depending on whether they are used in private settings (by contract, in employment or other organizational settings) or in public arenas such as courts. Court-related or court-annexed ADR programs, now encompassing the full panoply of dispute processes, may be subject to greater legal regulation, including selection, training, and credentialing of the arbitrators or mediators, ethics, confidentiality, and conflicts of interest rules, as well as providing for greater immunity from legal liability.

ADR processes are often differentiated from each other by the degree of control the third party neutral has over the process (the rules of proceedings), the substance (decision, advice, or facilitation), and the formality of the proceeding (whether held in private or public settings, with or without formal rules of evidence, informal separate meetings, or caucuses with the parties, and with or without participation of more than the principal disputants). ADR processes are being applied increasingly to diverse kinds of conflicts, disputes, and transactions, some requiring expertise in the subject matter (such as scientific and policy disputes) and spawning new hybrid processes such as consensus building, which engage multiple parties in complex, multi-issue problem solving and draw on negotiation, mediation, and other nonadjudicative processes (Susskind et al., 1999).

Although there have been efforts to develop taxonomies or predictive factors for assignment of particular case types to particular processes (and some courts assign or prohibit certain case types in some categories of dispute resolution), for the most part these efforts 'to fit the forum to the fuss' (Sander and Goldberg, 1994) have been unsuccessful. Amenability of different cases to different processes just as often depends on the personalities of the disputants, parties, lawyers, and third party neutrals as on any particular case type characteristic.

Theory and History of ADR

The modern growth of arbitration, mediation, and other ADR processes can be attributed to at least two different animating concerns. On the one hand, scholars, practitioners, consumers, and advocates for justice in the 1960s and 1970s noted the lack of responsiveness of the formal judicial system and sought better 'quality' processes and outcomes for members of society seeking to resolve disputes with each other, with the government, or with private organizations. This strand of concern with the quality of dispute resolution processes sought deprofessionalization of judicial processes (a reduction of the lawyer monopoly over dispute representation), and greater access to more locally based institutions such as neighborhood justice centers, which used community members, as well as those with expertise in particular problems, with the hope of generating greater party participation in dispute resolution processes (Merry and Milner, 1993). Others sought better outcomes than those commonly provided by the formal justice system, which tend toward the binary, polarized results of litigation in which one party is declared a loser, while the other is, at least nominally, a winner. More flexible and party-controlled processes were believed to deliver the possibility of more creative, Pareto-optimal solutions which are geared to joint outcomes, reduction of harm or waste to as many parties as possible, improvement of long-term relationships, and greater responsiveness to the underlying needs and interests of the parties (Menkel-Meadow, 1984; Fisher et al., 1991). Some legal and ADR processes (like arbitration) are rule based, but other forms of ADR (negotiation and mediation) are thought to provide individualized solutions to problems, rather than generalized notions of 'justice.'

A second strand of argument contributing to the development of ADR was, however, more quantitatively or efficiency based. Judicial officers, including those at the top of the American and English justice systems, argued that the excessive cost and delay in the litigation system required devices that would divert cases from court and reduce case backlog, as well as provide other and more efficient ways of providing access to justice (Burger, 1976; Woolf, 1996). This efficiency-based impetus behind ADR encouraged both court-mandated programs (like court-annexed arbitration for cases with lower economic stakes) and contractual requirements to arbitrate any and all disputes arising from services and products (provided in banking, health care, consumer, securities, educational, and communication-based industries).

Modern ADR structures are related only loosely to their historical antecedents. In many countries, arbitration had its origins in private commercial arbitrations, outside of the formal court structure, and it was used principally by merchants when disputing with each other (Dezalay and Garth, 1996). In the United States, labor arbitration developed to secure 'labor peace,' as well as to develop a specialized substantive 'law of the shop floor' (Fuller, 1963).

Early use of mediation or conciliation occurred in some courts and communities seeking both to reduce caseloads and to provide more consensual agreements in ethnically or religiously homogeneous areas (Auerbach, 1983). Indeed, mediation and other consensually based processes are thought to work best in regimes where there are shared values, whether based on common ethnicity or communitarian or political values (Shapiro, 1981). In Asian and other nations with more communitarian and harmony-based cultures (as contrasted to more litigative or individualistic cultures), mediation is often

the preferred form of dispute resolution, but it too has been used for system or political regime purposes beyond resolving the disputes of the parties (Lubman, 1967). Thus, most political regimes have had to deal with both public and private forms of dispute resolution that often supplement, but sometimes challenge or compete with, each other. Mediation is now promoted in transborder disputes, where the applicable law may not be clear (see European Directive, 2008/52/EC; Steffek, 2011).

The introduction or revival of multiple forms of dispute resolution (including mediation, arbitration, ombuds, and conciliation) within the legal system probably dates to the 1976 conference on the "Causes of Popular Dissatisfaction with the Administration of Justice." The idea of a 'multidoor courthouse' was suggested in order to meet both the caseload needs of the judicial system and the 'quality of justice' needs of consumers in a rapidly growing arena of legally and culturally cognizable claims (Sander, 1976). More deeply contextualized study of the social transformation of conflicts into legally cognizable claims by a community of sociolegal scholars (Felstiner et al., 1980–1981), drawing on anthropological, sociological, political, and psychological insights, also contributed to the theoretical, as well as practical, significance of pluralism in disputing. More recently, dispute resolution has been championed by those who focus on facilitating deliberative democracy in the larger polity as well (Menkel-Meadow, 2011, 2012).

Applications

Each of the ADR processes has its own logic, purposes, and jurisprudential justifications. Mediation and conciliation are often used to improve communications between parties, especially those with preexisting relationships, to 'reorient the parties to each other' (Fuller, 1971) and to develop future oriented solutions to broadly defined conflicts. Arbitration, on the other hand, being more like adjudication (Fuller, 1963, 1978) is used more often to resolve definitively a concrete dispute about an event that has transpired and requires fact finding, interpretation of contractual terms, or application of legal principles.

These basic forms have been adapted to a number of subject areas and dispute sites. As regular use of these formats of dispute resolution becomes more common, mediation seems to be overtaking arbitration as a preferred method of dispute resolution (because of the ideology of party self-determination and the flexibility of agreements). Arbitration, still most commonly used in labor disputes, is now the method of choice in form contracts signed by consumers as well as merchants. In the United States, mandatory predispute contractual commitments to arbitrate a wide variety of consumer, health care, securities, and employment disputes have been sustained by the US Supreme Court against claims these clauses violate due process or other constitutional rights to trial for disputes. Similarly, thus far many European courts have held that assignment to some form of ADR (mediation usually, but in some cases, arbitration as well) does not violate Art. 6 of the European Convention on Human Rights, granting basic hearing rights for human rights cases. Arbitration has, thus far, been the mode of choice for resolving international commercial, investment, and trade disputes, such as in the public law settings of the World Trade Organization (WTO), the General Agreement on Tariffs and Trade (GATT), and the International Center for the Settlement of Investment Disputes (ICSID), and the more private international law settings of the International Chamber of Commerce or the London Court of Arbitration. Arbitration has also been deployed in new forms of disputes developing under both domestic and international intellectual property regimes. Various forms of mediation and arbitration are also increasingly being used to resolve transnational disputes of various kinds (political, economic, natural resource allocation, and ethnic violence) and are employed by international organizations such as the United Nations and the Organization of American States, as well as multinational trade and treaty groups (NAFTA, the European Union, and Mercosur) and nongovernmental organizations in human rights and other issue-related disputes (Greenberg et al., 2000). Arbitration (and occasionally other forms of dispute resolution) is now commonly used in sports disputes, such as for the Olympics and FIFA (the International Federation for Football (Soccer)), with a special tribunal to administer such dispute resolution, located in Lausanne, Switzerland.

Beginning in the United States, but now in use internationally, mass injury (class action) cases, both involving personal and property damages, have been allocated to ADR claims facilities, utilizing both arbitral and mediative forms of individual case processing. In legal regimes all over the world, family disputes are increasingly assigned to mediative processes, both for child custody and for support and maintenance issues. In many nations, this growth in family mediation has spurred the development of a new profession of mediators, drawn from social work or psychology, who sometimes compete with lawyers both in private practice and as court officers (Palmer and Roberts, 1998).

In many jurisdictions, such as the United States, the United Kingdom, Italy, and the Netherlands, among others, some form of referral to ADR is now required before a case may be tried. Increasingly, however, parties to particularly complex disputes, such as environmental, mass torts, or governmental budgeting, may convene their own ADR processes, with a third party neutral facilitating a new form of public participatory process that combines negotiation, fact-finding, mediation, and joint problem solving. Such consensus building processes have also been applied to the administrative tribunal processes of both rule-making and administrative adjudication in a new process called 'reg-neg' (negotiated rule-making or regulation).

Although ADR has been considered, until quite recently, principally an American alternative to courts, the use of ADR is spreading slowly around the world, being used to relieve court congestion, to provide expertise in various subject matter disputes (e.g., construction, labor matters, family law), to build transnational dispute systems for economic, human rights, and political issues, and to offer alternative justice systems where there is distrust of existing judicial institutions. Increasingly, many national legal systems (e.g., England and Scotland, Italy, Argentina) require some effort at ADR (usually mediation) before a case may proceed to trial in civil cases. There is increased education and training for dispute resolution professionals, including both lawyers and nonlawyers (social

workers, psychologists, accountants, engineers, and architects) in a wide variety of settings, both within universities and in trade schools and other programs. The use of ADR across borders and cultures raises complex questions about intercultural negotiations (Salacuse, 1998) and multijurisdictional sources of law or other principles for dispute resolution.

Controversies

The use of mediation, arbitration, and ADR processes, in lieu of more traditional adjudication, has not been without its controversies, which are reviewed briefly in this section (Menkel-Meadow, 1997).

Privatization of Jurisprudence

With the increased use of negotiated settlements, mediation, and private arbitration, there has been concern that fewer and fewer cases will be available in the public arena for the making of precedent (Fiss, 1984), and debate about and creation of rules and political values for the larger community will suffer (Luban, 1995). As settlements are conducted in private and often have confidentiality or secrecy clauses attached to them, others will not learn about wrongs committed by defendants, and information that might otherwise be discoverable will be shielded from public view. Settlements may be based on nonlegal criteria, threatening compliance with and enforcement of law. Claims are more likely to be individualized than collectivized.

Whether there is more privatization or secrecy in the settlement of legal disputes than at some previous time itself remains a subject of controversy as empirical studies document relatively stable rates of nonjudicial case terminations (at over 90 percent in many jurisdictions and across all types of disputes) (Kritzer, 1991).

Related concerns about the privatization of the judicial system include increased indirect state intervention in the affairs of the citizenry through more disputing institutions, at the same time that the exit of wealthier litigants gives them less stake in the quality and financing of public justice systems (Abel, 1982). The debate centers on whether dispute resolution systems can serve simultaneously the private interests of disputants before them and the polity's need for the articulation of publicly enforced norms and values (Menkel-Meadow, 1995b).

Inequalities of Bargaining Power

A number of critics have suggested that less powerful members of society, particularly those subordinated by race, ethnicity, class, or gender, will be disadvantaged disproportionately in ADR processes where there are no judges, formal rules, or, in some cases, legal representatives to protect the parties and advise them of their legal entitlements (Delgado et al., 1985; Grillo, 1990–1991). Responses from ADR theorists suggest that there is little empirical evidence that less advantaged individuals or groups necessarily fare better in the formal justice system, and that sophisticated mediators and arbitrators are indeed sensitive to power imbalances and can be trained to 'correct' for them without endangering their neutrality in the ADR process. Many private ADR organizations have begun developing standards for good practices and due process protocols to protect the parties and ensure the integrity of the process.

Evaluation and Empirical Verification of Effectiveness

There are few robust research findings with respect to the effectiveness of ADR in meeting its claimed advantages. Recent findings from studies of ADR in the American federal courts have been contradictory about whether arbitration, mediation, and some forms of early neutral evaluation decrease case processing time or costs, either for the parties or for the system. Preliminary studies from England demonstrate low usage of mediation schemes (Genn, 1999). Yet studies continue to demonstrate high satisfaction rates among users of arbitration and mediation programs (MacCoun et al., 1992), and higher compliance rates with mediated outcomes than traditional adjudication (McEwen and Maiman, 1986). In light of the variation in ADR programs, it is too early for there to be sufficient databases for accurate comparisons between processes (Menkel-Meadow, 2010).

Distortions and Deformations of ADR Processes

Within the nascent ADR profession there is concern that the early animating ideologies of ADR are being distorted by their assimilation into the conventional justice system. Within a movement that sought to deprofessionalize conflict resolution, there are now competing professional claims for control of standards, ethics, credentialing, and quality control between lawyers and nonlawyers. Processes like mediation that were conceived as voluntary and consensual are now being mandated by court rules and contracts. Processes that were supposed to be creative, flexible, and facilitative are becoming more rigid, rule and law based, and judicialized as more common law is created by courts about ADR, and more laws are passed by legislatures. The overall concern is that a set of processes developed to be alternative to the traditional judicial system are themselves being coopted within the traditional judicial process with its overwhelming adversary culture. Policy makers and practitioners in the field are concerned about whether a private market in ADR is good for 'disciplining' and competing with the public justice system or whether, on the other hand, there will be insufficient accountability within a private market of dispute resolution.

The Future of ADR

There is no question that the use of a variety of different processes to resolve individual, organizational, governmental, political, and international problems is continuing to expand. New hybrid forms of ADR (as in mediation on the Internet (Katsh and Rifkin, 2001) and public policy mediation and consensus building) are developing to help resolve new problems, with greater participation by more parties. Large organizations are creating their own internal dispute resolution systems. There are clear trends in favor of mediation and arbitration in the international arena, where globalization of

enterprises and governmental interests require creative and simple processes that are not overly attached to any one jurisdiction's substantive law. These systems can promote goals of efficiency, fairness, clarity, and legitimacy, particularly in regimes with underdeveloped formal legal systems. It is also clear that there is competition over who will control such processes, and which processes will dominate in which spheres of human disputing and deal-making. The likely result is that the creative pluralism and flexibility of ADR will be subject increasingly to its own forms of formality and regulation in an effort to keep its promises of efficiency, participation, better quality outcomes, and justice.

See also: Conflict: Anthropological Aspects; Dispute Resolutions in Economics; International Arbitration; Lex Mercatoria.

Bibliography

Abel, R., 1982. The contradictions of informal justice. In: Abel, R. (Ed.), The Politics of Informal Justice: The American Experience. Academic Press, New York.

Auerbach, J., 1983. Justice without Law? Resolving Disputes without Lawyers. Oxford University Press, New York.

Burger, W., 1976. Agenda for 2000 AD – need for systematic anticipation. Federal Rules Decisions 70, 92–94.

Delgado, R., et al., 1985. Fairness and formality: minimizing the risk of prejudice in alternative dispute resolution. Wisconsin Law Review 1985, 1359–1404.

Dezalay, Y., Garth, B., 1996. Dealing in Virtue: International Commercial Arbitration and the Construction of a Transnational Legal Order. University of Chicago Press, Chicago.

European Union, Directive on Certain Aspects of Mediation in Civil and Commercial Matters, Directive 2008/52/EC.

Felstiner, W., Abel, R., Sarat, A., 1980–1981. The emergence and transformation of disputes: naming, blaming and claiming. Law & Society Review 15, 631–654.

Fisher, R., Ury, W., Patton, B., 1991. Getting to Yes: Negotiating Agreement without Giving In, second ed. Viking Penguin, New York.

Fiss, O., 1984. Against settlement. Yale Law Journal 93, 1073–1090.

Fuller, L., 1963. Collective bargaining and the arbitrator. Wisconsin Law Review 18, 3–47.

Fuller, L., 1971. Mediation: its form and functions. Southern California Law Review 44, 305–339.

Fuller, L., 1978. The forms and limits of adjudication. Harvard Law Review 92, 353–409.

Genn, H., 1999. The Central London County Court Pilot Mediation Scheme: Final Report. Lord Chancellor's Department, London.

Greenberg, M.C., Barton, J., McGuinness, M.E., 2000. Words Over War: Mediation and Arbitration to Prevent Deadly Conflict. Rowman & Littlefield, Lanham, MD.

Grillo, T., 1990–1991. The mediation alternative: process dangers for women. Yale Law Journal 100, 1545–1610.

Katsh, E., Rifkin, J., 2001. Online Dispute Resolution: Resolving Conflicts in Cyberspace. Jossey-Bass Publishers, San Francisco.

Kritzer, H., 1991. Let's Make a Deal: Understanding the Negotiation Process in Ordinary Litigation. University of Wisconsin Press, Madison, WI.

Luban, D., 1995. Settlements and the erosion of the public realm. Georgetown Law Journal 83, 2619–2662.

Lubman, S., 1967. Mao and mediation: politics and dispute resolution in communist China. California Law Review 55, 1284–1359.

MacCoun, R., Lind, E.A., Tyler, T., 1992. Alternative dispute resolution in trial and appellate courts. In: Kagehiro, D.K., Laufer, W.S. (Eds.), Handbook of Psychology and Law. Springer-Verlag, New York.

McEwen, C.A., Maiman, R.J., 1986. The relative significance of disputing forum and dispute characteristics for outcome and compliance. Law & Society Review 20, 439–447.

Menkel-Meadow, C., 1984. Toward another view of legal negotiation: the structure of problem solving. UCLA Law Review 31, 754–842.

Menkel-Meadow, C., 1995a. The many ways of mediation: the transformation of traditions, ideologies, paradigms and practices. Negotiation Journal 11 (3), 217–242.

Menkel-Meadow, C., 1995b. Whose dispute is it anyway? A philosophical and democratic defense of settlement (in some cases). Georgetown Law Journal 83, 2663–2696.

Menkel-Meadow, C., 1997. When dispute resolution begets disputes of its own: conflicts among dispute professionals. UCLA Law Review 44, 1871–1933.

Menkel-Meadow, C., 2010. Dispute resolution. In: Cane, P., Kritzer, H. (Eds.), Oxford Handbook of Empirical Legal Research. Oxford University Press, Oxford.

Menkel-Meadow, C., 2011. Scaling up deliberative democracy as dispute resolution in health care reform: a work in progress. Law & Contemporary Problems 74 (3), 1–30.

Menkel-Meadow, C., 2012. Complex Dispute Resolution, vol. 1, Foundations of Dispute Resolution; vol. 2, Multi-Party Dispute Resolution, Democracy and Decision Making; vol. 3, International Dispute Resolution. Ashgate Press, Farnham, UK; Burlington, Vt.

Merry, S., Milner, N., 1993. The Possibility of Popular Justice: A Case Study of American Community Justice. University of Michigan Press, Ann Arbor, MI.

Palmer, M., Roberts, S., 1998. Dispute Processes: ADR and the Primary Forms of Decision Making. Butterworth, London.

Salacuse, J., 1998. Ten ways that culture affects negotiating style: some survey results. Negotiation Journal 14 (3), 221–240.

Sander, F., 1976. Varieties of dispute processing. Federal Rules Decisions 70, 111–134.

Sander, F., Goldberg, S., 1994. Fitting the forum to the fuss: a user friendly guide to selecting an ADR procedure. Negotiation Journal 10, 49–68.

Shapiro, M., 1981. Courts: A Comparative and Political Analysis. University of Chicago Press, Chicago.

Steffek, F., 2011. Mediation. In: Basedow, J., Hopt, K.J., Zimmermann, R., Stier, A. (Eds.), The Max Planck Encyclopedia of European Private Law. Oxford University Press, Oxford.

Susskind, L., McKearnan, S., Thomas-Larmer, J., 1999. The Consensus Building Handbook: A Comprehensive Guide to Reaching Agreement. Sage, Thousand Oaks, CA.

Woolf, H., 1996. Access to Justice: Final Report to the Lord Chancellor on the Civil Justice System. HMSO, London.

Mediation, Statistical

Xiaolu Wang, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The goal of causal mediation analysis is to investigate the extent to which the causal effect of a 'treatment' variable on an outcome variable is propagated through a third (or more) variable – the mediator of interest. The focus is to partition the total effect into an indirect effect and a direct effect. After defining various relevant causal effects using potential outcomes, four approaches of causal mediation analysis are addressed, including linear structural equation models, linear counterfactual models with potential errors, marginal structural models, and principal stratification.

The goal of causal mediation analysis is to investigate the extent to which the causal effect of a 'treatment' variable Z (the cause of interest) on an outcome variable Y is propagated through a third (or more) variable M – the mediator of interest (see Figure 1). The focus is not to estimate the total effect of treatment Z on outcome Y, but to partition the total effect into an indirect effect mediated by M and a direct effect not mediated by M. (Note that the direct and indirect effects are relative terms, depending on the choice of the mediator(s) of interest.) That is, it is the underlying causal mechanism (the causal pathways from Z to Y), rather than the simple overall causal effect of Z on Y, that is of primary interest.

Causal mediation analysis has been widely practiced in many social, behavioral, and other scientific disciplines, such as psychology, sociology, political science, and epidemiology (see MacKinnon, 2008). Moreover, mediation is the basis of causal inference in more complicated longitudinal and network settings. Thus, discussions in this article also have direct implications for those settings.

As an example, Rossi et al. (1980) conducted a randomized experiment to examine whether providing monetary support to recently released ex-convicts could reduce the rate of recidivism. Thousands of ex-convicts were randomized into a treatment group in which subjects either received money transfers ($Z=1$) or assigned to a control group without monetary support ($Z=0$). (This is a simplified description of the original experiment.) The outcome Y was whether the ex-convict got rearrested ($Y=1$) or not ($Y=0$). The authors found no total effect of money transfers on the rearrest rate in the experiment. In an attempt to uncover the causal mechanism that led to this zero total effect, Rossi et al. introduced a mediator – the number of hours each ex-convict worked per week after release. Their mediation analysis suggested that the supposed negative direct effect of money transfers on the rearrest rate had been canceled out by a positive indirect effect via the mediator. A (over)simplified version of the authors' argument was as follows. On the one hand, many ex-convicts get reinvolved in illegal (money-generating) activities soon after release because of financial difficulty. Thus, giving them money would reduce the incentive for them to reengage in illegal activities, which would lower the rearrest rate (a negative direct effect). On the other hand, money transfer would also reduce the incentive for them to work legally, resulting in fewer weekly working hours (M). Less time spent in legal employment means weaker social regulations, which would in turn increase recidivism (a counterbalancing positive effect). However, this argument ignited a heated debate on the validity of the methods used for mediation analysis (see Zeisel, 1982a,b; Rossi et al., 1982; Smith, 2013). The mediation framework of this experiment is used in this article in various places as an example.

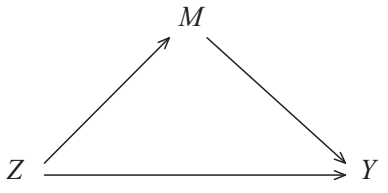

Figure 1 Three variable path diagram with mediator *M*.

Formal Definitions of Causal Effects

Holland (1988) and Robins and Greenland (1992) use the notation of potential outcomes (or counterfactuals) to formally define various causal effects relevant to mediation. Unless otherwise stated, the treatment, mediator, and outcome are assumed throughout to be either binary or continuous (this can be easily modified to accommodate more general cases). Let $M_i(z)$ denote the mediator value subject (or respondent) i would take on if i, possibly contrary to fact, received treatment level z. $Y_i(z, m)$ is the outcome value i would take on if i were assigned treatment level z and mediator level m. $Y_i[z, M_i(z)]$ is the outcome of i under treatment z, that is, $Y_i[z, M_i(z)] \equiv Y_i(z)$. $Y_i[z, M_i(z')]$ denotes i's outcome if i received treatment z while the mediator was fixed at the level which i would 'naturally' take on under treatment level z' different from z. For example, in the recidivism experiment, $M_i(1)$ is the weekly hours ex-convict i would work after release if he or she received money transfers; $Y_i(0, 40)$ is the outcome status of i (irrespective of being rearrested or not) if i did not receive money transfers and was 'forced' to work 40 h per week; $Y_i[1, M_i(1)]$ is i's outcome if the particular recently released ex-convict received money transfers; $Y_i[0, M_i(1)]$ is i's outcome if the individual did not receive money transfers and was 'forced' to work $M_i(1)$ hours per week, where $M_i(1)$ is the number of hours i would 'voluntarily' work if receiving money transfers. It is assumed throughout that the potential

outcomes of i depend only on the treatment status of i, *not* on the treatment status of any other subject j ($j \neq i$). Rubin (1980) calls this the stable unit treatment value assumption (SUTVA), which rules out between-subject interference.

In contrast, let Z_i, M_i, Y_i denote the actually realized values of subject i's treatment, mediator, and outcome, respectively, and $\mathbf{X}_i$ a vector of observed pretreatment covariates. Clearly, for binary treatment, $M_i = M_i(1) \times Z_i + M_i(1) \times (1 - Z_i)$, and $Y_i = Y_i[1, M_i(1)] \times Z_i + Y_i[0, M_i(0)] \times (1 - Z_i)$.

With potential outcomes, various causal effects at both the subject/unit and aggregate levels can be defined. For each subject i,

unit effect of Z on M:

$$M_i(z) - M_i(z') \qquad [1a]$$

(average) effect of Z on M:

$$\mathrm{E}\{M(z) - M(z')|\mathbf{X}\} \qquad [1b]$$

where the aggregate effect ([1b]) is obtained by averaging the unit effect ([1a]) over some subpopulation/subgroup defined by covariates $\mathbf{X}_i$ (subscript i is suppressed for all causal effects at the aggregate level for simplicity); If it is the average effect for the whole population/group that is of interest, the conditioning part (on $\mathbf{X}$) can be removed;

unit controlled direct effect (of Z on Y):

$$Y_i(z, m) - Y_i(z', m) \qquad [2a]$$

controlled direct effect (of Z on Y):

$$\mathrm{E}\{Y(z, m) - Y(z', m)|\mathbf{X}\} \qquad [2b]$$

unit controlled effect of M on Y:

$$Y_i(z, m) - Y_i(z, m') \qquad [3a]$$

controlled effect of M on Y:

$$\mathrm{E}\{Y(z, m) - Y(z, m')|\mathbf{X}\} \qquad [3b]$$

unit total effect (of Z on Y):

$$Y_i[z, M_i(z)] - Y_i[z', M_i(z')] \qquad [4a]$$

total effect (of Z on Y):

$$\mathrm{E}\{Y[z, M(z)] - Y[z', M(z')]|\mathbf{X}\} \qquad [4b]$$

unit pure direct effect (of Z on Y):

$$Y_i[z, M_i(z')] - Y_i[z', M_i(z')] \qquad [5a]$$

pure direct effect (of Z on Y):

$$\mathrm{E}\{Y[z, M(z')] - Y[z', M(z')]|\mathbf{X}\} \qquad [5b]$$

unit total indirect effect (of Z on Y):

$$Y_i[z, M_i(z)] - Y_i[z, M_i(z')] \qquad [6a]$$

total indirect effect (of Z on Y):

$$\mathrm{E}\{Y[z, M(z)] - Y[z, M(z')]|\mathbf{X}\} \qquad [6b]$$

For example, in the recidivism experiment, the unit controlled direct effect $Y_i(1, 30) - Y_i(0, 30)$ is the effect of receiving money ($z = 1$) versus not receiving money ($z = 0$) on rearrest for the *same* subject i, given that the mediator (weekly working hours) is fixed at 30. In contrast, the unit pure direct effect $Y_i[1, M_i(0)] - Y_i[0, M_i(0)]$ is the effect of receiving money versus not receiving on rearrest for subject i, given that the mediator is fixed at $M_i(0)$ – the weekly hours i would 'voluntarily' work if i did not receive money ($z = 0$). Note that for the controlled direct effect (CDE), $E\{Y(1, 30) - Y(0, 30)\}$, the mediator is uniformly fixed at the *same* level – 30 h per week – for *all* subjects, while for the pure direct effect (PDE), $E\{Y[1, M(0)] - Y[0, M(0)]\}$, the mediator is set at a subject-specific level $M_i(0)$ which usually varies across subjects.

Several remarks are in order. First, all the causal effects defined above are comparisons between pairs of potential outcomes of the *same* subject. However, in most empirical cases, only one of the two potential outcomes in comparison is observed for each subject. As a result, the unit level effects in general cannot be obtained (see Bertrand and Mullainathan, 2003 for an exception) – this has been described as the fundamental problem of causal inference. Second, the total effect ([4b] or [4a]) can be decomposed into the sum of PDE ([5b] or [5a]) and the total indirect effect (TIE) ([6b] or [6a]). See Wang and Sobel (2013) for a comprehensive review of other ways to decompose the total effect. Third, in general, CDE ([2b] or [2a]) is not equal to PDE ([5b] or [5a]), and thus cannot be used in the decomposition. Fifth, Pearl (2001) renames ([5b]) as 'natural direct effect,' and many authors now use this name instead. At last, the controlled effects (([2a] or [2b]) and ([3a] or [3b])) are often of greater interest in cases where both the treatment and the mediator can be manipulated in practice, while the PDE ([5a] or [5b]) and TIE ([6a] or [6b]) are of greater interest when the mediator cannot be manipulated through intervention.

Approach One: Linear Structural Equation Models

Baron and Kenny (1986) popularized the use of linear structural equation models (LSEMs) to study mediation. A typical LSEM is specified as

$$M_i = \alpha_0 + \alpha_1 Z_i + \boldsymbol{\alpha}'\mathbf{X}_i + \varepsilon_i \qquad [7a]$$

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 M_i + \boldsymbol{\beta}'\mathbf{X}_i + \eta_i \qquad [7b]$$

where the error terms satisfy $\mathrm{E}\{\varepsilon_i|Z_i, \mathbf{X}_i\} = 0$ and $\mathrm{E}\{\eta_i|Z_i, M_i, \mathbf{X}_i\} = 0$, and ε_i and η_i may or may not be correlated. The LSEM parameter α_1 is usually called the 'effect' of Z on M, β_1 the 'direct effect,' β_2 the 'effect' of M on Y, $\alpha_1\beta_2$ the 'indirect effect,' and $\beta_1 + \alpha_1\beta_2 \equiv \tau$ the 'total effect.'

However, Robins (1986) and Holland (1988) point out that, even if the LSEM ([7a] and [7b]) is correctly specified, these so-called 'effects' are in general not equal to their causal counterparts defined in the section 'Formal Definitions of Causal Effects' and thus cannot afford causal interpretations. For instance, with binary treatment Z, the LSEM parameter β_1, the so-called 'direct effect,' is equal to $\mathrm{E}\{Y|Z = 1, M = m, \mathbf{X} = \mathbf{x}\} - \mathrm{E}\{Y|Z = 0, M = m, \mathbf{X} = \mathbf{x}\}$. That is, the comparison is between two subpopulations/subgroups. The first includes those subjects with $\{Z = 1, M = m, \mathbf{X} = \mathbf{x}\}$, while the second includes those with $\{Z = 0, M = m, \mathbf{X} = \mathbf{x}\}$, and these two subgroups are generally not exchangeable. For example, in the recidivism experiment, the subgroup with $\{Z = 1, M = 30, \mathbf{X} = \mathbf{x}\}$ consists of ex-convicts who receive money

transfers and work 30 h week^{-1}, while the one with $\{Z=0, M=30, \mathbf{X}=\mathbf{x}\}$ includes those who do not receive money and work 30 h week^{-1}. These two subgroups may differ from each other systematically – e.g., those in the first subgroup could be on average poorer (or more diligent) than those in the second, as they, even though receive money transfers, still work the same number of hours as those in the second who do not receive money. Thus, such a comparison is problematic. In contrast, the causal counterpart, the CDE ([2b] and [2a]) or PDE ([5b] and [5a]), is defined by comparing each subject with him/herself. It then becomes clear that the LSEM parameters in general cannot be interpreted causally.

In order for the LSEM parameters to equal their causal counterparts, strong assumptions are needed:

$$Y_i(z,m), M_i(z) \perp Z_i | \mathbf{X}_i \qquad [8a]$$

$$Y_i(z,m) \perp M_i | Z_i, \mathbf{X}_i \qquad [8b]$$

Robins (1986) calls these the (sequential) ignorability assumptions. The substantive meaning of [8a] is that given covariates $\mathbf{X}_i$, treatment Z_i can be seen as randomly assigned; or, in observational contexts, $\mathbf{X}_i$ includes all confounders of both the *Z-Y* (treatment-outcome) and *Z-M* (treatment-mediator) relations. Condition [8b] means that given covariates $\mathbf{X}_i$ and treatment assignment Z_i, the mediator M_i behaves as if it is randomly assigned; or, $\mathbf{X}_i$ and Z_i include all confounders for the *M-Y* (mediator-outcome) relationship. If both [8a] and [8b] hold, the total effect ([4b]) $= \beta_1 + \alpha_1\beta_2 \equiv \tau$, the effect of Z on M ([1b]) $= \alpha_1$, the controlled effect of M on Y ([3b]) $= \beta_2$, CDE ([2b]) = PDE ([5b]) $= \beta_1$, and TIE ([6b]) $= \alpha_1\beta_2$.

For experimental studies in which only the treatment is randomized, [8a] automatically holds; however, [8b] in general does not hold unless $\mathbf{X}_i$ includes all confounders for the *M-Y* relationship. If [8a] holds but [8b] does not, as in the case of the recidivism experiment, and the LSEM ([7a] and [7b]) is employed, then only the 'total effect,' $\tau \equiv \beta_1 + \alpha_1\beta_2$, equals the total effect ([4b]), and the 'effect of Z on M' α_1 equals the effect of Z on M ([1b]), but the rest of the LSEM 'effects' do not equal their causal counterparts. (Note that as the outcome in the recidivism experiment is binary, logistic model is a better alternative to eqn [7b]; see VanderWeele and Vansteelandt, 2010 for details. However, this does not affect the argument made herein.) Thus, applying LSEM (or its variants) to those experiments in which only the treatment, but not the mediator, is randomized would generally give wrong and probably misleading results about the direct and indirect effects. Wang and Sobel (2013, p. 225) give a simulated example of such an experiment in which both the true direct and indirect effects are zero, but the correctly specified LSEM (with uncorrelated error terms ε_i and η_i) wrongly gives nonzero 'effects.'

There are two possible remedies. First, one could randomize both the treatment and mediator in an experiment. Randomizing the treatment solves the confounding problems for the *Z-M* and *Z-Y* relations, and randomizing the mediator removes confounding from the *M-Y* relation. Thus, sequential ignorability ([8a] and [8b]) is guaranteed to hold. Second, if the mediator cannot be randomized, one could collect information on as many *M-Y* confounders as possible, in the hope that these confounders included in $\mathbf{X}_i$ account for most of the *M-Y* confounding. In observational studies, even [8a] typically does not hold. In such cases, one should include into $\mathbf{X}_i$ all major confounders for not only the *Z-M*, *Z-Y*, but also the *M-Y* relations. Since one can never be sure if all the major confounders are included, sensitivity analysis is highly recommended (see VanderWeele, 2010).

Approach Two: Linear Counterfactual Models with Potential Errors

Holland (1988) and Sobel (2008) develop a linear counterfactual model (LCM) parallel to LSEM:

$$M_i(z) = \alpha_0^c + \alpha_1^c z + \boldsymbol{\alpha}_i^{c\prime\mathbf{X}} + \varepsilon_i(z) \qquad [9a]$$

$$Y_i(z,m) = \beta_0^c + \beta_1^c z + \beta_2^c m + \boldsymbol{\beta}_i^{c\prime\mathbf{X}} + \eta_i(z,m) \qquad [9b]$$

where $\mathrm{E}\{\varepsilon_i(z)|\mathbf{X}_i\} = 0$ and $\mathrm{E}\{\eta_i(z,m)|\mathbf{X}_i\} = 0$ for any fixed z and m; the superscript 'c' indicates these are parameters of a counterfactual model. There are several differences between the LCM ([9a] and [9b]) and the LSEM ([7a] and [7b]). First, the observed variables M_i and Y_i in the LSEM become potential outcomes in the LCM. Second, the observed treatment and mediator Z_i and M_i (as random variables) in the LSEM become 'fixed' variables z and m that can be conceptually manipulated. Third, the error terms ε_i and η_i in the LSEM turn into 'potential errors' $\varepsilon_i(z)$ and $\eta_i(z,m)$. It is straightforward to verify that α_1^c is the effect of Z on M ([1b]); β_1^c is the CDE ([2b]); β_2^c is the controlled effect of M on Y ([3b]). (Note that $\boldsymbol{\alpha}^c$ and $\boldsymbol{\beta}^c$ should not be given causal interpretation.)

In general, these LCM parameters cannot be identified, because not all potential outcomes $M_i(z)$ and $Y_i(z, m)$ can be observed for each subject i. However, if sequential ignorability ([8a] and [8b]) holds, it turns out that $\alpha_1^c = \alpha_1, \beta_1^c = \beta_1, \beta_2^c = \beta_2$, That is, the LCM parameters can be identified with the corresponding LSEM parameters. Moreover, the total effect ([4b]), denoted here by τ^c, can be shown to satisfy $\tau^c = (\beta_1^c + \alpha_1^c\beta_2^c) = (\beta_1 + \alpha_1\beta_2) \equiv \tau$; CDE ([2b]) = PDE ([5b]) $= \beta_1^c = \beta_1$, and TIE ([6b]) $= \alpha_1^c\beta_2^c = \alpha_1\beta_2$.

Because sequential ignorability ([8a] and [8b]) is very strong and cannot be met in most empirical settings, Holland (1988) instead assumes only [8a] and replaces [8b] with the following two assumptions:

$$\varepsilon_i(z) = \varepsilon_i, \eta_i(z,m) = \eta_i \quad \text{for all } z \text{ and } m \qquad [10]$$

$$\beta_1^c = 0, \quad \alpha_1^c \neq 0 \qquad [11]$$

If assumptions [8a], [10], and [11] hold, it can be shown that the LCM parameters satisfy $\tau^c = \alpha_1^c\beta_2^c$, $\tau^c = \tau$, and $\alpha_1^c = \alpha_1$. It follows that $\tau = \alpha_1\beta_2^c$, which leads to $\beta_2^c = \tau/\alpha_1$. Thus, all the LCM parameters are identified with the LSEM parameters, and in particular, the controlled effect of M on Y ([3b]), or β_2^c, is identified using the treatment Z as an instrumental variable (IV). Moreover, CDE ([2b]) = PDE ([5b]) $= \beta_1^c = 0$ (where the last equation is by assumption [11]), and TIE ([6b]) $= \tau^c = \tau = \alpha_1\beta_2$.

Assumption [10] requires that each unit effect defined in the section 'Formal Definitions of Causal Effects' remain absolutely the same across all subjects, and thus is called the constant unit effect assumption. This assumption is apparently substantively

unreasonable (even more so than sequential ignorability). Sobel (2008) replaces it with a weaker one:

$$\eta_i(z,m) - \eta_i(z',m') \perp M_i(z), M_i(z') | \mathbf{X}_i, \quad \text{for any } z, m, z', m' \qquad [12]$$

This assumption means that the *difference* between two potential outcomes, $Y_i(z, m) - Y_i(z', m')$, are independent of the *potential* mediators $M_i(z)$ and $M_i(z')$, given covariates $\mathbf{X}_i$. In contrast, [8b] assumes that the *absolute levels* of the potential outcomes are independent of the *observed* mediator, given $\mathbf{X}_i$. For examples in which assumption [12] may hold but ([8b]) may not, see Sobel (2008) and Lindquist and Sobel (2011). By replacing assumption [10] with [12], but still assuming [11] and [8a], one can get the same identification results as in the foregoing paragraph.

Sobel (2008) gives another assumption weaker than [12], under which (alongside [8a] and [11]) the controlled effect of M on Y, β_2^c, is identified in the same way, but the PDE and TIE are not identified. At last, assumption [11] requires $\beta_1^c = 0$, which is also strong. For situations in which $\beta_1^c \neq 0$, see Emsley et al. (2010).

Approach Three: Marginal Structural Models

Ignorability ([8b]) requires that the pretreatment covariates $\mathbf{X}_i$ include all confounders for the M-Y relation. However, there might be some posttreatment M-Y confounders that are not included in $\mathbf{X}_i$ and may or may not be affected by the treatment. This means [8a] holds but [8b] does not. Then replace [8b] with the following assumption,

$$Y_i(z,m) \perp M_i | Z_i, \mathbf{X}_i, \mathbf{W}_i \qquad [13]$$

where $\mathbf{W}_i$ is a vector of posttreatment confounders for the M-Y relationship. If [8a] and [13] hold and some variable(s) in $\mathbf{W}_i$ is affected by the treatment, none of the direct and indirect effects defined in the section 'Formal Definitions of Causal Effects' can be identified using LSEM (and regressions).

VanderWeele (2009) proposes an alternative using marginal structural models (MSM). Consider the group of MSM below:

$$\mathrm{E}\{Y(z,m)\} = \gamma_0^M + \gamma_1^M z + \gamma_2^M m \qquad [14]$$

$$\mathrm{E}\{M(z)|\mathbf{X}\} = \alpha_0^M + \alpha_1^M z + \boldsymbol{\alpha}^{M'}\mathbf{X} \qquad [15a]$$

$$\mathrm{E}\{Y(z,m)|\mathbf{X}\} = \beta_0^M + \beta_1^M z + \beta_2^M m + \boldsymbol{\beta}^{M'}\mathbf{X} \qquad [15b]$$

where the superscript 'M' indicates these are MSM parameters. For eqn [14], γ_1^M is the CDE ([2b]), and γ_2^M is the controlled effect of M on Y ([3b]). Since the LCM ([9a] and [9b]) implies the MSM ([15a] and [15b]), the MSM parameters have causal interpretations similar to their LCM counterparts. Under assumptions [8a] and [13], to identify and estimate the CDE and the controlled effect of M on Y, only eqn [14] is needed. However, to identify and estimate the PDE ([5b]) and TIE ([6b]), one needs eqns [15a] and [15b], and another assumption in addition to [8a] and [13]:

$$Y_i(z,m) \perp M_i(z') | \mathbf{X}_i \qquad [16]$$

which suggests that none of the posttreatment confounders for the M-Y relation in $\mathbf{W}_i$ is affected by the treatment Z_i. If [16] does not hold, PDE and TIE in general cannot be identified. However, if [8a] and [13] hold, the CDE can still be identified and estimated using eqn [14], but LSEM cannot be used, as mentioned above.

Unlike the case of the LCM in the section 'Approach Two: Linear Counterfactual Models with Potential Errors,' where the identification/estimation of the LCM parameters relies on the identification/estimation of the corresponding LSEM parameters, the estimation of the MSM is achieved through a procedure called inverse probability of treatment weighting (IPTW). To remove confounding, both the LSEM and LCM approaches include confounders $\mathbf{X}_i$ into the structural models as covariates. As a result, controlling for confounding and the estimation of parameters are accomplished simultaneously. However, for MSM with IPTW, confounding is dealt with in a separate step (IPTW) preceding the final estimation of model parameters. (Note that although $\mathbf{X}_i$ is included into the MSM ([15a] and [15b]), the purpose of doing so is not to adjust for confounding; see VanderWeele, 2009 for details.)

This two-stage procedure is as follows. First, pseudotreatment and control groups are created by properly weighting the original sample. The weights are calculated using both observed pre- and posttreatment confounders $\mathbf{X}_i$ and $\mathbf{W}_i$ such that, after weighting, the pseudotreatment and control groups become exchangeable (balanced) in terms of $\mathbf{X}_i$ and $\mathbf{W}_i$, while the casual effects in this pseudosample remain the same as in the original one. Thus, all confounding is removed in this stage. In the second stage, regular regression methods are applied to the pseudosample to estimate the MSM parameters. For more details concerning MSM and IPTW, see Robins (1999) and Robins et al. (2000). Using MSM and IPTW entails another assumption that both $\Pr\{Z=1|\mathbf{X}=\mathbf{x}\}$ and $\Pr\{M=1|\mathbf{W}=\mathbf{w}, Z=z, \mathbf{X}=\mathbf{x}\}$ cannot equal to either 1 or 0 for any given $\mathbf{x}_i$, z_i, and $\mathbf{w}_i$. If this is violated, structural nested models (SNM) can be used instead, for which see Robins et al. (1992), Hernán et al. (2005), and Ten Have et al. (2007).

All the aforementioned models (LSEM, LCM, and MSM) assume no treatment-by-mediator interaction on the outcome, but these can be readily extended to include an interaction term; see VanderWeele (2010), VanderWeele and Vansteelandt (2009), and Wang and Sobel (2013). Brumback et al. (2004) provide a general method of sensitivity analysis for MSM.

Approach Four: Principal Stratification

In previous sections, the mediator M_i is treated as some variable that can be (at least conceptually) manipulated. Frangakis and Rubin (2002) propose a different approach, called principal stratification (PS), which focuses on causal effects within subgroups/subpopulations defined by the potential mediators $M_i(z)$ and $M_i(z')$. Since conceptually the potential mediators share the same nature with pretreatment covariates $\mathbf{X}_i$ (except that the former are not observed before treatment and only partially observed after treatment), PS in some sense transforms the issue of mediation into one of moderation. Then,

the questions of interest would be like: What is the effect of Z on Y for those subjects with $\{M_i(z)=m,\ M_i(z')=m'\}$? The subpopulations/subgroups defined as such are called principal strata, and the effects defined within principal strata are called principal effects. A classic treatment of a special case of the PS approach can be found in Angrist et al. (1996). Rubin (2004, 2005) proposes definitions of direct and indirect effects within the PS framework. VanderWeele (2008) discusses the conceptual relations between Rubin's principal direct and indirect effects and the various direct and indirect effects defined in the section 'Formal Definitions of Causal Effects.'

To illustrate the rationale of PS, now return to the recidivism example. For ease of exposition, the mediator M_i – weekly working hours – is transformed into a binary variable (=1 if worked at all, 0 otherwise). In such a case, four principal strata can be defined. The first stratum consists of those ex-convicts with $\{M_i(0)=1,\ M_i(1)=1\}$, who would always work no matter whether they received money transfers or not, and thus may be called 'always workers.' The second stratum includes those with $\{M_i(0)=0,\ M_i(1)=0\}$, who would never work whether received money or not, and are called 'never workers.' The third consists of those with $\{M_i(0)=1,\ M_i(1)=0\}$, who would go to work if they did not receive money transfers but would not go to work if they did receive money transfers, and, following Angrist et al. (1996), are called 'defiers.' The fourth stratum includes those ex-convicts with $\{M_i(0)=0,\ M_i(1)=1\}$, who would not work if they did not receive money but would go to work if they did, and are called 'compliers.' However, as only one of the two potential mediators is observed for each subject, the principal membership cannot be completely identified without further restrictions. Of particular interest is the effect of money transfers Z on rearrest Y for 'never workers', that is,

$$\mathrm{E}\{Y[1,M(1)]-Y[0,M(0)]|M(0)=0,M(1)=0\} \qquad [17]$$

because this principal effect corresponds to the 'clean' direct effect of money transfer on rearrest 'uncontaminated' by the mediator (the potential working status is guaranteed to be 0 no matter whether received money transfers or not), though with reference to only for the subpopulation of 'never workers'.

To identify and estimate the principal effect ([17]), first note that ignorability ([8a]) automatically holds by virtue of randomization. Further, assume there is no complier:

$$M_i(0)\geq M_i(1) \qquad [18]$$

which is called monotonicity. Assumption [18] is in line with the theory that receiving money transfers is a disincentive to work: all those who would not work if they did not receive money transfers would become even more unwilling to work (and thus would still not work) if they did receive money transfers. Unlike Angrist et al. (1996) who make another assumption that the CDE ([2a] and [2b]) is zero – the exclusion restriction, here it is exactly the direct effect that is of interest. Therefore, exclusion restriction does not apply here).

Under assumption [18], one can obtain **Table 1** (observed mediator by observed treatment). Cell 1 of **Table 1** with $\{Z_i=0,\ M_i=0\}$ consists solely of never workers; cell 2 with $\{Z_i=0,\ M_i=1\}$ consists of always workers and defiers; cell 3 with $\{Z_i=1,\ M_i=0\}$ contains never workers and defiers; cell 4 with $\{Z_i=1,\ M_i=1\}$ includes only always workers.

Table 1 Principal strata composition in the recidivism example

	$M_i=0$ *(not worked)*	$M_i=1$ *(worked)*
$Z_i=0$ (no payment)	*Never workers*: $M(0)=0$ & $M(1)=0$ (cell 1)	*Always workers*: $M(0)=1$ & $M(1)=1$ *Defiers*: $M(0)=1$ & $M(1)=0$ (cell 2)
$Z_i=1$ (with payment)	*Never workers*: $M(1)=0$ & $M(0)=0$ *Defiers*: $M(1)=0$ & $M(0)=1$ (cell 3)	*Always workers*: $M(1)=1$ & $M(0)=1$ (cell 4)

Due to randomization, the 'never workers' in the treatment group (cell 1) are exchangeable with the never workers in the control (cell 3), and the former are fully visible. Then, if the never workers in cell 3 can be identified, the principal effect ([17]) is identified. To achieve this goal, one more assumption is needed:

$$Y_i[1,M_i(1)]\perp M_i(0)|\mathbf{X}_i \qquad [19]$$

Jo and Stuart (2009) call this the principal ignorability assumption (they provide a more general form). This assumption means that pretreatment covariates $\mathbf{X}_i$ are all that are needed to predict principal strata membership of the subjects.

If assumptions [8a], [18], and [19] hold, the principal effect ([17]) can be estimated by a three-step procedure which is very similar to propensity-score-based methods. First, use all subjects in cells 1 and 2 to fit a logistic regression in some functional form to predict $\Pr[M(0)=0|\mathbf{X}=\mathbf{x}]$, or equivalently $\Pr[M=0|Z=0,\ \mathbf{X}=\mathbf{x}]$, for each subject in cell 1 (never workers). These predicted probabilities are called principal scores by Hill et al. (2002). Second, use the same fitted regression to predict $\Pr[M(0)=0|\mathbf{X}=\mathbf{x}]$ for each subject in cell 3, which gives his/her principal score. Third, based on these two sets of principal scores for subjects in cells 1 and 3, respectively, one can use either matching or weighting methods to estimate the principal effect ([17]). For details of this estimating procedure, see Hill et al. (2002).

Angrist et al. (1996) show that if ignorability ([8a]), monotonicity $M_i(0)\leq M_i(1)$ (no defiers), and exclusion restriction hold, then the effect of Z on Y for compliers (also called the complier-average causal effect, CACE), $\mathrm{E}\{Y[1,\ M(1)]-Y[0,\ M(0)]|M(0)=0,\ M(1)=1\}$, is identified as the instrumental variable (IV) estimand: $\mathrm{E}\{Y[1,M(1)]-Y[0,M(0)]\}/\mathrm{E}[M(1)-M(0)]$, using treatment assignment as the instrument and assuming the denominator is nonzero.

In some cases, one might be interested in the causal effects within strata defined by a single potential mediator, e.g., $\mathrm{E}\{Y[1,\ M(1)]-Y[0,\ M(0)]|M(1)=0\}$. In the recidivism example, this is the effect of Z on Y for the subgroup consisting of both 'never workers' and 'defiers.' This approach is called single principal stratification; see Joffe et al. (2007) for details. In the context of noncompliance, Mattei et al. (2013) and Mealli and Pacini (2013) use Bayesian analysis to improve the identification of principal effects by utilizing the information of a secondary outcome associated with the primary outcome.

By far, the treatment and mediator are both assumed to be binary. Jin and Rubin (2008, 2009) discuss situations in which the mediator has more categories or is continuous and/or situations where the treatment can take on more values in the context of partial compliance.

A major shortcoming of the PS approach is that the principal strata cannot be observed before treatment. To solve this problem, Hill et al. (2002) propose to define strata with principal scores which are calculated using pretreatment variables. Thus, these principal score strata can be identified before treatment. See Wang and Sobel (2013) for a review of more general forms of principal (score) stratification.

To conclude, all the approaches discussed herein require very strong ignorability types of assumptions that are often implausible in empirical cases. There is no universally superior or inferior approach. The choice of approaches should be made on a case-by-case basis by examining which assumptions are likely to hold (or not) and the nature of the research questions one is asking. Sensitivity analysis is highly recommended. In cases where the type of data available does not permit valid causal mediation analysis, one may want to work on improving the research design, since a good research design is often the most fundamental and effective solution to problems of causal inference. For example, Bertrand and Mullainathan (2003) use résumé experiment to measure more than one potential outcomes for each experimental unit. Thus, no ignorability assumptions are needed. At last, the discussion here focuses mostly on parametric causal models of mediation (except the PS approach). For more nonparametric methods, see, for example, Hong and Nomi (2012) (although their causal model is largely nonparametric, the estimation procedure involves parameterization) and Pearl (2001).

See also: Analysis of Variance and Generalized Linear Models; Causal Inference: Overview; Graphical Markov Models: Overview; Longitudinal Causal Inference.

Bibliography

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91, 444–472.

Baron, R.M., Kenny, D.A., 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic and statistical considerations. Journal of Personality and Social Psychology 51, 1173–1182.

Bertrand, M., Mullainathan, S., 2003. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. The American Economic Review 94, 991–1013.

Brumback, B.A., Hernán, M.A., Haneuse, S.J.P.A., Robins, J.M., 2004. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. Statistics in Medicne 23, 749–767.

Emsley, R., Dunn, G., White, I.R., 2010. Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. Statistical Methods in Medical Research 19, 237–270.

Frangakis, C.E., Rubin, D.B., 2002. Principal stratification in causal inference. Biometrics 58, 21–29.

Hernán, M.A., Cole, S.R., Margolick, J.B., Cohen, M.H., Robins, J.M., 2005. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. Pharmacoepidemiology and Drug Safety 14, 477–491.

Hill, J., Waldfogel, J., Brooks-Gunn, J., 2002. Differential effects of high-quality child care. Journal of Policy Analysis and Management 21, 601–627.

Holland, P.W., 1988. Causal inference, path analysis, and recursive structural equation models (with discussion). Sociological Methodology 18, 449–493.

Hong, G., Nomi, T., 2012. Weighting methods for assessing policy effects mediated by peer change. Journal of Educational Effectiveness (special issue on the statistical approaches to studying mediator effects in education research) 5, 261–289.

Jin, H., Rubin, D.B., 2008. Principal stratification for causal inference with extended partial compliance. Journal of the American Statistical Association 103, 101–111.

Jin, H., Rubin, D.B., 2009. Public schools versus private schools: causal inference with partial compliance. Journal of Educational and Behavioral Statistics 34, 24–45.

Jo, B., Stuart, E.A., 2009. On the use of propensity scores in principal causal effect estimation. Statistics in Medicine 28, 2857–2875.

Joffe, M.M., Small, D., Hsu, C.-Y., 2007. Defining and estimating intervention effects for groups that will develop an auxiliary outcome. Statistical Science 22, 74–97.

Lindquist, M.A., Sobel, M.E., 2011. Graphical models, potential outcomes and causal inference: comment on Ramsey, Spirtes and Glymour. Neuroimage 57, 334–336.

MacKinnon, D.P., 2008. Introduction to Statistical Mediation Analysis. Lawrence Erlbaum Associates, New York.

Mattei, A., Li, F., Mealli, F., 2013. Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program. The Annals of Applied Statistics 7, 2336–2360.

Mealli, F., Pacini, B., 2013. Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. Journal of the American Statistical Association 108, 1120–1131.

Pearl, J., 2001. Direct and indirect effects. In: Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence. Morgan Kaufmann, San Francisco, CA, pp. 411–420.

Robins, J.M., 1986. A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. Mathematical Modeling 7, 1393–1512.

Robins, J.M., 1999. Association, causation and marginal structural models. Synthese 121, 151–179.

Robins, J.M., Blevins, D., Ritter, G., et al., 1992. G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. Epidemiology 3, 319–336.

Robins, J.M., Greenland, S., 1992. Identifiability and exchangeability of direct and indirect effects. Epidemiology 3, 143–155.

Robins, J.M., Hernán, M.A., Brumback, B., 2000. Marginal structural models and causal inference in epidemiology. Epidemiology 11, 550–560.

Rossi, P.H., Berk, R.A., Lenihan, K.J., 1980. Money, Work, and Crime: Experimental Evidence. Academic Press, New York.

Rossi, P.H., Berk, R.A., Lenihan, K.J., 1982. Saying it wrong with figures: a comment on Zeisel. American Journal of Sociology 88, 390–393.

Rubin, D.B., 1980. Discussion of 'randomization analysis of experimental data: the Fisher randomization test' by D. Basu. Journal of the American Statistical Association 75, 591–593.

Rubin, D.B., 2004. Direct and indirect effects via potential outcomes. Scandinavian Journal of Statistics 31, 161–170.

Rubin, D.B., 2005. Causal inference using potential outcomes: design, modeling, decisions. Journal of the American Statistical Association 100, 322–331.

Smith, H.L., 2013. Research design: toward a realistic role for causal analysis. In: Morgan, S.L. (Ed.), Handbook of Causal Analysis for Social Research. Springer, New York, pp. 45–73.

Sobel, M.E., 2008. Identification of causal parameters in randomized studies with mediating variables. Journal of Educational and Behavioral Statistics 33, 230–251.

Ten Have, T.R., Joffe, M.M., Lynch, K.G., et al., 2007. Causal mediation analyses with rank preserving models. Biometrics 63, 926–934.

VanderWeele, T.J., 2008. Simple relations between principal stratification and direct and indirect effects. Statistics and Probability Letters 78, 2957–2962.

VanderWeele, T.J., 2009. Marginal structural models for the estimation of direct and indirect effects. Epidemiology 20, 18–26.

VanderWeele, T.J., 2010. Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiology 21, 540–551.

VanderWeele, T.J., Vansteelandt, S., 2009. Conceptual issues concerning mediation, interventions and composition. Statistics and Its Interface 2, 457–468.

VanderWeele, T., Vansteelandt, S., 2010. Odds ratios for mediation analysis for a dichotomous outcome. American Journal of Epidemiology 172, 1339–1348.

Wang, X., Sobel, M.E., 2013. New perspectives on causal mediation analysis. In: Morgan, S.L. (Ed.), Handbook of Causal Analysis for Social Research. Springer, New York, pp. 215–242.

Zeisel, H., 1982a. Disagreement over the evaluation of a controlled experiment. American Journal of Sociology 88, 378–389.

Zeisel, H., 1982b. Hans Zeisel concludes the debate. American Journal of Sociology 88, 394–396.

Medical Education

Fred CJ Stevens and Cees PM van der Vleuten, Maastricht University, Maastricht, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes the development and current status of medical education as a professional field of practice and inquiry along two questions: first, how can we educate medical students in such a way that they will become active and permanent learners, ready to adapt to changing content and context of medical practices? Secondly, how can we socialize students into the profession of medicine so that they will become committed to its core principles and occupational values? Two developments are described that significantly contributed to answering these questions, namely problem-based learning and professionalism.

Within a period of less than four decades medical education has become a distinct and highly regarded professionalized field of practice and inquiry. National and international societies have been founded. Thousands of people attend medical education conferences all over the world. New textbooks and journal publications on how to teach and assess in medical education appear on a regular basis. At present, there are at least 20 different peer-reviewed international scientific journals on medical education, not to mention the many opportunities to publish in medical journals. Evidently, the coming into existence of medical education as a distinct and international professional field of expertise has taken place at a fascinating pace.

Not surprisingly, research in medical education overwhelmingly supports the training and preparation of future doctors for medical practice. A recent literature review shows that since the end of the 1980s clinical and communication skills, clinical clerkships, student assessment, and problem-based learning (PBL) were medical education's key themes of investigation (Rotgans, 2011). Publications of the past 25 years strongly focus on the individual student while the psychologist perspective dominates, typically emphasizing the study of individuals. However, medical education as a topic for development and inquiry has a longer history.

Since the appearance of the Flexner report in the United States in the early 1900s, medical education has been in an almost permanent state of change and development (Cooke et al., 2006). Abraham Flexner, a professional educator, was asked to conduct an assessment of all US and Canadian medical schools. His findings shocked the nation. In his report Flexner criticized the mediocre quality and profit motives of many schools and teachers in his country and the inadequacy of curricula and facilities (Flexner, 1910). Only a handful of medical schools (of 155 he visited) were capable of providing an adequate medical education. He therefore advised that, similar to the German system of medical education, all medical schools should become university-based, and that formal analytical reasoning, typical for the natural sciences, should be at the heart of the teaching of physicians.

Following up on the Flexner report the basic medical curriculum has primarily been built around the biological and physical sciences for at least half a century. Minimum standards and requirements for schools were permanently raised, resulting in substantial improvement in the quality of medical education. In light of Flexner's recommendations, schools became university affiliated and began to work toward a better connection between the preclinical and clinical phases. Medical graduates received clinical instruction in academic and affiliated hospitals and were taught by full-time scientists of the preclinical disciplines in modern, scientific, and biomedicine. In many schools this continues to be the standard model for the undergraduate teaching of physicians.

Notwithstanding these changes, curriculum development remained a problem, not only in the United States, but worldwide. By the 1960s, due to the growing complexity of medicine and according to (sub)specializations, everywhere curricula had become rather detailed and ultimately became unmanageable from an educational point of view. Students were required to take courses in too many (sub)disciplines and specialties, some with questionable relevance to their professional role and later career. At the end of the 1960s it seemed as if the most apparent change since the Flexner report was that the typical medical student had moved away from being an apprentice with more clinical experience than he would ever be able to understand, to a physician well trained in the basic sciences though with a rather limited and fragmented clinical experience. From the late 1960s on many efforts in curriculum development were therefore tuned toward bringing this more into balance with the support for content and curriculum development from behavioral and social scientists. These, together with reformist clinicians, ultimately brought new views on how to educate physicians and what to include in the curriculum, in line with individuals' future plans and changing societal demands of what a modern medical doctor should be.

A complete description of the many changes that have taken place in medical education in the last decades could probably fill a bookshelf. But two questions have been leading in these changes, first, how can we educate medical students in such a way that they will become active and permanent learners, ready to adapt to changing content and context of medical practices? Secondly, how can we socialize students into the profession of medicine so that they will become committed to the core principles and occupational values of the

practice of medicine? We describe two developments that significantly contributed to answering these questions, PBL and professionalism.

Problem-Based Learning

Perhaps the greatest milestone for the development of medical education in the past 50 years has been the introduction of PBL. In the late 1960s, PBL, started at McMaster University in Canada, was adopted in 1974 as the basic teaching model by the new medical faculty in Maastricht, the Netherlands, followed soon by the University of Newcastle, Australia, and the University of New Mexico in the United States. From there it was disseminated all over the world to numerous faculties in developed as well as developing countries. The acceptance and success of the new problem-based model was not happenstance. In its infancy, PBL as an educational strategy, and so its later alternates like self-directed learning, integrated learning, task-based learning, focal problems learning, community-based learning, and project-based learning were looked at with suspicion by the medical elite. But soon it proved to foster adult learning, self-responsibility of students, and early introduction of 'real' patient problems to students in a better way than any traditional teaching system up until that time. Instead of medical students having a rather passive role in their undergraduate education by attending and listening to lectures from the knowledge silos of the disciplines, in PBL they had to be active in shaping their own learning process throughout the curriculum. PBL was nothing less than a paradigm shift in students' learning and turned out to become the new trust in medical education. When PBL started to gain a foothold in medical schools it was seen as a paradigm shift and the solution to problems with traditional curricula of students being passive, only memorizing, and failing to learn how to apply and integrate basic knowledge into a clinical context. PBL was at the heart of changing this.

PBL came into existence at a time when there was an urgent need for a new type of doctor. This new doctor would not only be an expert in the somatic aspects of a disease, but also responsible for delivering integrated care, including in his/her diagnosis, treatment, and follow-up procedures, the insights of different knowledge domains including those of the social and behavioral sciences (Van Berkel et al., 2010). PBL was the new strategy toward the modernization of medical education. It appeared to be an appropriate response to the pressures of modernization to make a shift from the pedagogical lecturer-centered approach of education, to a more andragogy, self-directed, innovative adult learning approach (Dent and Harden, 2005; Harden, 2006). What began as an Anglo-Saxon model of teaching and learning soon spread throughout medical schools in Africa, the Pacific, Asia, as well as the Americas. Simply put, using PBL implied you were a modern school.

Notwithstanding its success and efforts of fine-tuning through the years, up to this day it has been quite difficult to establish its superiority in robust designs, nor has it been fully clarified what it exactly is that does the trick in PBL (Dochy et al., 2003; Norman and Schmidt, 1992). PBL studies in medical education still focus on this issue, and also on the question of how to align assessment procedures with the practice of PBL and to bridge the gap between the preclinical and clinical years.

Professionalism and Professional Socialization

The second question – how can we socialize students into the profession of medicine so that they will become committed to the core principles and occupational values of medical practice – is at the heart of medicine's modern-day professionalism movement. Professionalism and processes of socialization into medical school and the medical profession were already at the core of the work of sociologists from the 1950s to the 1970s, based on three landmark studies "The student physician" (Merton et al., 1957), "Boys in White" (Becker et al., 1961), and "Profession of medicine" (Freidson, 1970). Though these first two studies were primarily aimed at testing competing sociological theories (structural functionalism versus symbolic interactionism), they implicitly also signified a major arrival of medicine in its efforts to understand and improve the selection and education of physicians. For example, up to that time medical schools, in particular in the United States, selected students based on 'traits' they must possess to successfully enter and finish their studies. Such traits were also deemed crucial in the development of the currently *in vogue* concept of professionalism, referring to key attitudes and attributes every medical professional should possess (in particular altruism) (Hodges et al., 2011). Research evidenced that the professional education of physicians must also be seen as a continuous, long-term process of growth and development, not only focusing on discrete clinical skills and clinical competencies. Learning the 'rules of the game' and understanding the importance of becoming successfully socialized into the culture of medicine underscored that the attitudes and skills the student develops in medical school and during clerkship training and internships are important sources for his/her later performance as a physician. To this day, however, the question remains whether in the preparation of a physician education at medical school is of more importance than the later work environment. As Freidson (1970: p. 90) noted, a significant amount of behavior is situational and, "(…) people are constantly responding to the organized pressures of the situations they are in at a particular time (…) and what they do is more an outcome of the pressures of the situation they are in than of what they have earlier 'internalized.'" Hafferty et. al. (2014) coined the term 'chameleon socialization' indicating that modern students are not so much socialized into an occupational or organizational milieu but rather into the demands of the moment.

Such insights are vital to the growth and development of every medical student to become a competent, professional expert, and to the necessity of a physician's lifelong learning. Medical schools around the world are increasingly using portfolios as assessment tools, whereby students learn to reflect on their knowledge, skills, and attitudes (van der Vleuten and Dannefer, 2012). In a few schools portfolios have now been introduced as their primary assessment tool. With these portfolios it is the student's responsibility to show and to prove that she or he is to size in a blend of the different required

competencies of a physician – medical expert, professional, communicator, collaborator, manager, scholar, and health advocate (Frank, 2005; Frank and Danoff, 2007).

Today, there is little doubt that medical education comprises a complex program of socialization into the profession of medicine and its culture, and that the study of medical socialization, professionalism, and lifelong learning will continue to be important niches to further develop.

Medical Education as a Professional Field of Inquiry

For not too long the art of clinical practice was transmitted in a guilt-like fashion, driven by the concept of role modeling (Krause, 1996). Nowadays, this practice is carried out more systematically, in teams, according to a plan, grounded in scientific evidence and described, framed, and assessed in line with an index of formalized competencies (Frank, 2005; Frank and Danoff, 2007). Research in medical education contributed substantially to this. The setup of dedicated academic (research) departments of medical education in medical schools all over the world fostered the development of the professional field of medical education. By the time, these departments succeeded more and more in involving physicians and other health professionals and disciplines into medical education research and innovation. For those involved, medical education and the study and advancement of it became an essential part of their livelihood and academic career (Gallagher, 2001; Harden, 2006).

As is known from the sociology of the professions, the optimal cognitive basis for any professional discipline is one whose claimed tasks are sufficiently distinctive to allow the drawing of clear jurisdictional boundaries so that standards of competent performance can be established and controlled (Abott, 1988). Such tasks, however, must not be standardized entirely, because that would make discretionary professional judgment needless (Stevens et al., 2007). This delicate balance applies to the field of medical education as well. In focusing too narrowly on educational theory development, the medical educationalist risks his/her working relationships with the practice of medicine (Albert et al., 2007). However, educationalists may also lose perspective and theoretical grounding when they identify too closely with the medical world (Dimitroff and Davis, 1996). Medical education as a professional field of expertise has been in a continuous effort to keep that balance, occasionally with mixed results. Medical education research has been accused by practitioners of being not conclusive or practical. But where the insights from research are practical and conclusive, the application can nevertheless be limited (Broekkamp and Van Hout-Wolters, 2007). By some, medical education is still seen as an intellectual activity, focusing on how do teachers transmit ideas and/or how do students acquire them. Thus goals of professional performance and behavior would seem to be subordinate to goals of generating new ideas for medical education (Bloom, 1988). This does no justice to current medical education research. Through learning from many different disciplines and perspectives, medical education is now in the middle of an exciting process of theoretical, philosophical, and practical elaboration. While cognitive psychology, with its strong emphasis on the individual learner, dominated medical education research for many years, today the interest has shifted from how do students learn to how do students process knowledge, and toward the study of processes of constructing professional identities conducive to lifelong learning, and learning from and with others in authentic contexts (Mann, 2011). Sociocultural perspectives, which in recent years have become very popular in medical education, emphasize that learning and development is embedded in specific cultural matrices of histories, artifacts, procedures, signs, and symbols (Rogoff and Chavajay, 1995). The cultural matrix not only affects a person's outward behavior but also the internal mental processes generating that behavior. These perspectives, pointing to understanding behavior within sociocultural contexts, challenge our notions of learning, and of the course of individual development. In medical education this means that research is shifting from the individual medical student toward workplace learning and communities of practice, where newcomers become engaged, are transformed, and then, on their turn, transform the community in which they interact (Lave and Wenger, 1991).

Summarizing, at present there is not one theoretical perspective that dominates the field but medical educationalists use different approaches, admittedly still in a rather eclectic way, borrowing from theories and perspectives of sociologists, anthropologists, economists, organizational and management scientists, and so on. While most research in medical education is still effectiveness driven, seeking for solutions to practical problems in order to find what works or what does not, there is a move going on toward more discovery and curiosity-driven research ("why and how does it work") and from quantitative toward more qualitative or mixed methods research approaches (Cook et al., 2008; Rotgans, 2011). Insights and methods from other disciplines contribute to new understandings, for example that education requires continuity and that socialization for professional roles must be seen as closely interdependent with the structure of the situation in which it occurs and for which it prepares. More and more, students are considered active participants in their learning and socialization processes, while teachers are the facilitators to how to learn and what to learn. Fueled by the increasing attention for new and old topics such as team work, lifelong learning, professionalism, medical school culture, management, human development, and academic leadership, medical education awaits exciting times.

Medical Education for the Future: Old Wine in New Bottles?

Now, one hundred years after Flexner's report, are the questions – how can we improve medical education, and produce competent and compassionate physicians effectively and efficiently – sufficiently answered (Cook et al., 2008)? While modern medical education has established the inclusion of the continuous developments in biomedical science, more doubt exists about the preparation of physicians to serve the broader health needs of the society. Notwithstanding the many changes in content and structure of medical education through the years, in the centennial year of the Carnegie Foundation's

groundbreaking Flexner Report that radically changed medical education, the Foundation released with the report *Educating Physicians: A Call for Reform of Medical School and Residence* another call for reform to bring medical education to the next level of excellence (Cooke et al., 2006). Medical schools and residency programs are using modern teaching approaches, like simulations and standardized patients for instruction and assessment. Required competencies and learning objectives are defined, and small group teaching has become quite common. Yet, health care has changed too and new challenges have come to the front, requiring new pedagogies, forms of assessment, and different views on what is expected from the modern doctor. The awareness that, in many places, medical education is inflexible, long, and teacher-oriented is growing. Connections between formal and experiential learning are poor, and there is little opportunity for learners to observe and follow up patients over time. Moreover, students are less well trained in nonclinical roles, nor sufficiently prepared to serve the changing health needs of the society. Only in a limited number of schools across the world clinical and nonclinical experiences, and disciplinary knowledge domains, practical skills, and the teaching and understanding of attitudes and values are well integrated in a healthy blend (Cook et al., 2008).

Old wine in new bottles? – as similar concerns were also raised 50 years ago (Bloom, 1988). Not really. More and more medical educators are becoming aware of the need to prepare learners for the professional roles of tomorrow. Medical education today is more than ever seen as an endeavor contributing to the process of human development and the construction of professional identities. Research in medical education over the past 50 years has significantly contributed to this understanding. Shifts in perspectives have taken place. Medical educators have learned from the views of other disciplines, and are including different perspectives in their ways of dealing with teaching and learning. The insight has gained a foothold that learning takes place through and with others. Learning is now seen as a continuous, participative, and collaborative process that happens within a context, less and less as an individualized acquisition of knowledge and skills.

With this, medical education's role is not finalized, requires reflection and extension, and is becoming less neutral or value free in regard to the content of medical education it helps to produce. With the increasing call on medical educators to prepare doctors for the future the question is legitimate of what responsibility educationalists have in shaping this (Frenk et al., 2010). Similar to the situation 50 years ago, when PBL was expected to play a key role in shaping a new kind of doctor, medical education needs to play a substantive role again in helping to answer the needs of the society, contributing to the human development process of learners and socializing them into a broader concept of the content and culture of medicine (Bloom, 1988). Due to demographic developments, changing disease patterns, increasing health care inequities, health care needs of patients are changing everywhere in a rapid pace. These bring new challenges to the forefront. Graduates are ill equipped to deal with these. For example, while in all societies care at home and in the community needs more attention, in most medical curricula around the world the culture of medicine still means the culture of biomedicine and the hospital. Even now, when in medical education we speak about the importance of workplace learning this covertly means learning in the hospital. Already, for years, a global workforce misbalance exists of an oversupply of hospital specialists at the expense of generalists, primary care physicians, and physicians specialized in elderly care. So is medical education future-proof?

Medical education at the beginning of the twentieth century introduced and substantiated the scientific curriculum. Sixty years later, PBL was at the heart of instructional innovation in assisting to make the shift from memorizing medical knowledge to searching, analyzing, and integrating knowledge. Now another reform is happening, coined as transformative learning (Frenk et al., 2010). Having moved first from formative to informative learning, transformative learning aims, among others, at developing critical, reflective qualities in students in order to be able to synthesize information for decision making and to show leadership for adopting educational models so they can address different local health care priorities (Frenk et al., 2010). By helping to find ways of how to actively engage learners in the community, medical education could play a substantive role again in preparing medical students for the social, economic, and health care realities of tomorrow.

See also: Co-constructivism in Educational Theory and Practice; Communities of Practice; Education and Learning: Lifespan Perspectives; Educational Research, The Field of; Globalization in Education; Higher Education Research; Instructional Design; Learning Theories and Educational Paradigms; Teaching as a Profession: United States; Vygotsky's Theory of Human Development and New Approaches to Education.

Bibliography

Abott, A., 1988. The System of Professions. An Essay on the Division of Expert Labor. University of Chicago Press, Chicago.

Albert, M., Hodges, B., Regher, G., 2007. Research in medical education: balancing service and science. Advances in Health Sciences Education 12, 103–115.

Becker, H., Greer, B., Hughes, E.C., Strauss, A.L., 1961. Boys in White: Student Culture in Medical School. University of Chicago Press, Chicago.

Bloom, S.W., 1988. Structure and ideology in medical education: an analysis of resistance to change. Journal of Health & Social Behavior 29, 294–306.

Broekkamp, H., Van Hout-Wolters, B., 2007. The gap between educational research and practice: a literature review, symposium, and questionnaire. Educational Research and Evaluation 13, 203–220.

Cook, D.A., Bordage, G., Schmidt, H.G., 2008. Description, justification and clarification: a framework for classifying the purposes of research in medical education. Medical Education 42, 128–133.

Cooke, M., Irby, D.M., Sullivan, W., Ludmerer, K.M., 2006. American medical education 100 years after the Flexner report. New England Journal of Medicine 355, 1339–1343.

Dent, J.A., Harden, R.M. (Eds.), 2005. A Practical Guide for Medical Teachers. Elsevier, Edinburgh.

Dimitroff, A., Davis, W.K., 1996. Content analysis of research in undergraduate medical education. Academic Medicine 71, 60–67.

Dochy, F., Segers, M., Van Den Bossche, P., Gijbels, D., 2003. Effects of problem-based learning: a meta-analysis. Learning and Instruction 13, 533–568.

Flexner, A., 1910. Medical Education in the United States and Canada: A Report to the Carnegie Foundation for the Advancement of Teaching. Carnegie Foundation for the Advancement of Teaching, New York.

Frank, J.R., 2005. The CanMEDS 2005 Physician Competency Framework. The Royal College of Physicians and Surgeons of Canada, Ottawa.

Frank, J.R., Danoff, D., 2007. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. Medical Teacher 29, 642–647.

Freidson, E., 1970. Profession of Medicine: A Study of the Sociology of Applied Knowledge. Dodd, Mead, New York.

Frenk, J., Chen, L., Bhutta, Z.A., Cohen, J., Crisp, N., Evans, T., Fineberg, H., Garcia, P., Ke, Y., Kelley, P., Kistnasamy, B., Meleis, A., Naylor, D., Pablos-Mendez, A., Reddy, S., Scrimshaw, S., Sepulveda, J., Serwadda, D., Zurayk, H., 2010. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. The Lancet 376, 1923–1958.

Gallagher, E.B., 2001. Health, health care, and medical education in the Arab world. In: Cockerham, W.C. (Ed.), The Blackwell Companion to Medical Sociology.

Hafferty, F., Cruess, R.L., Cruess, S.R., 2014. Medical school socialization. In: The Wiley-Blackwell Encyclopedia of Health, Illness, Behavior and Society, pp. 1400–1043.

Harden, R.M., 2006. International medical education and future directions: a global perspective. Academic Medicine 81, S22–S29.

Hodges, B.D., Ginsburg, S., Cruess, R., Cruess, S., Delport, R., Hafferty, F., Ho, M.-J., Holmboe, E., Holtman, M., Ohbu, S., Rees, C., Ten Cate, O., Tsugawa, Y., Van Mook, W., Wass, V., Wilkinson, T., Wade, W., 2011. Assessment of professionalism: recommendations from the Ottawa 2010 conference. Medical Teacher 33, 354–363.

Krause, E.A., 1996. Death of the Guilds: Professions, States, and the Advancement of Capitalism, 1930 to the Present. Yale University Press, New Haven and London.

Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, Cambridge.

Mann, K.V., 2011. Theoretical perspectives in medical education: past experience and future possibilities. Medical Education 45, 60–68.

Merton, R.K., Reeder, L.G., Kendall, P.L., 1957. The Student Physician: Introductory Studies in the Sociology of Medical Education. Harvard University Press, Cambridge, MA.

Norman, G.R., Schmidt, H.G., 1992. The psychological basis of problem-based learning: a review of the evidence. Academic Medicine 67, 557–565.

Rogoff, B., Chavajay, P., 1995. What's becoming of research on the cultural basis of cognitive development? American Psychologist 50, 859–877.

Rotgans, J.I., 2011. The themes, institutions, and people of medical education research 1988-2010: content analysis of abstracts from six journals. Advances in Health Sciences Education 17, 515–527.

Stevens, F.C.J., Diederiks, J.P.M., Grit, F., Horst, F.V.D., 2007. Exclusive, idiosyncratic and collective expertise in the interprofessional arena: the case of optometry and eye care in The Netherlands. Sociology of Health and Illness 29, 481–496.

Van Berkel, H., Scherpbier, A., Hillen, H., van der Vleuten, C. (Eds.), 2010. Lessons from Problem-Based Learning. Oxford University Press, Oxford.

van der Vleuten, C.P., Dannefer, E.F., 2012. Towards a systems approach to assessment. Medical Teacher 185–186.

Medical Geography

Jonathan D Mayer and Alex Wagner, University of Washington, Seattle, WA, USA

Published by Elsevier Ltd.

Abstract

Medical geography concerns the relationships between humans and the environment that produce states of health and disease. It also concerns the relationships between health and the spatial structure of society. While medical geography still examines these relationships through understandings of culture, behavior, and techniques emphasizing human–environment interaction, novel concepts and theories are also beginning to be used by medical geographers. These include data from genetics, genomics, and proteomics; Big Data and large spatially referenced datasets; and the exposome, or the sum total of environmental exposure through an individual's life. Medical geography is a subfield not only of geography, but is also of public health. Spatial analysis and data, including GIS, have become essential techniques that are in great demand in academic schools of public health and departments of environmental health and toxicology; they are also techniques that are in great demand in public health agencies that are responsible for disease surveillance and control, health planning, and environmental toxicology. Medical geography is beginning to embrace multiple scales, ranging from the atomic to the macro-social and global scales.

Introduction

There are two ways of understanding medical geography. The first includes only research done by groups formally identified with departments of geography. The second is far more inclusive and includes all research that uses spatial methods and theories to understand health and disease. The latter has been recently reviewed (Mayer, 2009).

This presents a definitional dilemma. Medical geography is a subfield of geography and of public health – and indeed, of population-based medical research. One of the most important trends in medical geography is the increasing sophistication of medical geographers in epidemiology (including molecular epidemiology and genetic epidemiology), environmental health, exposure modeling, and biostatistics. Conversely, spatial concepts and methods – particularly GIS and spatial modeling – have become increasingly important in all of the more traditional fields of public health. Thus there is much ambiguity surrounding the term 'medical geography.' Many who were trained formally in medical geography have moved wholly or partly into other disciplines such as epidemiology, while others trained in a variety of fields including epidemiology and clinical medicine have moved into departments of geography. Restricting 'medical geography' to work that is done only by people with primary academic affiliations in departments of geography is unduly narrow, and misses the majority of research in what, intellectually, is clearly medical geography.

Recent developments in medical geography include the integration of political ecology with health, the nascent use of genetic and molecular methods, increasing use of more sophisticated GIScience and 'Big Data' spatial analysis of health and disease, and the use of social theory in understanding health. Equally important is the adoption of spatial methods to fields other than institutional geography. Major contemporary global and local problems, ranging from globalization and global health inequities to environmental change, have been highlighted. In the past two decades especially, noninfectious diseases such as cancers, heart diseases, and neurologic diseases have become as important to medical geography as are infectious diseases and their relations with the environment and culture.

The 1990s was a decade of change. Some argued that medical geography overemphasized the 'biomedical model' of disease. This critique was particularly strong in places where spatial methods were no longer regarded as 'cutting edge' or were considered obsolete, and where social theory was dominant; Australia–New Zealand and Britain were the prime movers in this development of what has come to be labeled 'health geography' though the critique of the biomedical model has been notable in the US as well. Ambiguity remains, though, since (1) the so-called biomedical model is protean and ill-defined and (2) multiple terms are used to denote medical geography, including the 'geography of health and disease,' 'epidemiologic geography,' 'public health geography,' and 'spatial epidemiology' (Mayer, 2009). Defining mainstream medical geography as being primarily limited to social theoretic approaches has had three main consequences: (1) much of value was learned and continues to be learned by including social theory; (2) many scientific medical geographers have switched the focus of their publications away from geographical journals to those in epidemiology and medicine; and (3) the majority of researchers who use spatial methods and concepts are not 'formal' geographers, most of whom are unaware of the internal geographic debates of the 1990s, and, if aware, consider them to be esoteric and lacking relevance for their research.

The Evolution of Medical Geography

The ecological approach to medical geography was dominant following World War II until spatial analytic techniques were introduced into medical geography in the 1960s, as the quantitative revolution spurred the growth and expansion of medical geography. Medical geography also embraced the developing approaches of health services research and the

delivery of care. Critical questions such as barriers to access, including geographic barriers to access, were brought into focus by the Great Society Programs and by ethical arguments surrounding the right to health care. This focus is broadly labeled as the geography of health care and is beyond the scope of this article.

The ecological imagination considers disease as the result of maladaptive relations between people, their behavior, and the environment (Meade, 1977). Much of this work has been based on the intellectual tradition that began with Jacques May (1958), a surgeon-turned medical geographer affiliated with the American Geographical Society. May wrote extensively about the ecology of disease and the geography of nutrition. His interest stemmed from his experiences and observations on the relations between culture, habitat, and disease in what was then called 'French Indochina.' Disease ecology considers the complex relations between the physical, sociocultural, and biological environments on the one hand, and the resulting patterns of disease on the other. May suggested that disease only occurs when there is coincidence in time and space of agent (pathogen) and host. The emphasis was almost entirely on infectious disease caused by pathogens such as viruses, bacteria, and protozoa, on the virulence of these agents, and on the susceptibility of hosts, those organisms that harbor the disease. Most medical geography has been concerned with human hosts, though clearly diseases among nonhuman organisms can also be analyzed geographically.

The 'One Health Movement' and EcoHealth, both of which have become major interdisciplinary movements in fields ranging from ecology to public health, hold that the division between human and animal diseases, and, indeed, between these and ecosystem health, is artificial and impedes both understanding and the attainment of sustainability in societies and nested ecosystems. Many of the roots of this view are found in ecological medical geography, in which diseases with animal reservoirs, such as African trypanosomiasis ('African sleeping sickness') or yellow fever were considered, both of which have animal reservoirs. The other roots include sustainability science and biological ecology. The One Health movement will undoubtedly be further incorporated into medical geography in the next few years (post-2014).

Most medical geographers, like epidemiologists, identify Hippocrates as their intellectual precursor. In addition to being a prominent Greek physician, he was also a major thinker in the relationships between the physical environment and health. In his essay, *Of Airs, Waters, and Places* (Dubos, 1965), Hippocrates noted the importance of climate, water, and other physical characteristics as influences on the health conditions of areas. Some of his environmental thinking has stood the test of time. Diseases as diverse as malaria, dengue, schistosomiasis, and arsenicosis have their roots in the natural environment and in human interaction with the environment.

Few advances were apparent in medical geography until the latter part of the eighteenth century and the nineteenth century, with the advent of the 'public health movement' in Europe and the Americas. This was not a monolithic movement, but importantly many thinkers and social activists saw the local environment as a determinant of public health. Much of the attention was naively devoted to cleanliness and the minimization of filth. The 'miasma theory of disease' held that foul areas and polluted conditions caused all diseases. Rodent control, water quality, and air quality were all major issues in the public health movement.

Some of this thinking was based on observation, and many diseases are associated with an abundance of rodents, poor water quality, and air pollution. Dr John Snow, a London anesthesiologist, is usually considered to be the major progenitor of epidemiology, medical geography, and environmental health. GIScience and cartography courses frequently begin with descriptions of Snow's conviction that the cholera epidemics in mid-nineteenth century London were caused by contaminated water, though he had no way of knowing that there was a microbial basis to the disease. A pleasing legend is that he mapped cholera cases, and noted that a water pump was at the center of the distribution. According to this narrative, Snow convinced local officials to remove the handle of the pump, and the epidemic ended, all because of his disease map. The truth is a bit different, though the simple narrative is still erroneously taught in epidemiology and subfields of geography. By the time the pump handle was removed, the epidemic was waning, a development that preceded Snow's map. He conducted an epidemiologic survey of the affected and unaffected population – effectively, a case-control study. He discovered that a disproportionate number of cases received their water from one private company, and that controls obtained water from another. This elevated risk led to the removal of the pump handle, which was associated with only one of the water suppliers. Snow subsequently mapped the distribution of cases, and the famous Broad Street pump was near the center of the distribution. He incorporated elements of what later became epidemiology and medical geography as one part of his analysis and civic activism (Paneth, 2004).

Ironically, and simultaneously with the increase in environmental understandings of disease was the fundamentally important discovery of the microbial basis of infectious disease. This discovery completely reoriented biological science to an entirely different scale, and away from the environment. The 'germ theory of disease' and the related doctrine of specific etiology exerted a profound and lasting influence on science and medicine that countered the major emphasis of geographic studies on the macro level. Pasteur and Koch discovered the microbial basis of infectious disease between approximately 1860 and 1885. Koch won the Nobel Prize in Physiology and Medicine in 1905 for his discoveries. It was not until the second half of the twentieth century, with the development and emphasis on the field of environmental health in public health, that there was recognition that both the micro and the macro scales are important.

The practical utility of disease mapping and medical geography was further realized during World Wars I and II, where again it became important to know what diseases would be encountered by troops who went abroad from their homelands. This knowledge was of obvious strategic utility; interestingly, it has received minimal consideration in histories of medical geography. Disease mapping thus played a role in both wars, as well as in the Vietnam War and the wars in Iraq and Afghanistan. It has practical utility in advising travelers about vaccination and in forming differential diagnoses in clinical medicine should somebody return from travel with symptoms that could lead to broad differential diagnoses (Mayer, 1989).

Following World War II, the major emphasis on medical geography was initially on the development of disease ecology, first based largely on the work of May and subsequently on that of Learmonth (1988) and others who also sought to understand the interplay of biological, cultural, and environmental factors in disease. The mobility of populations became another theme for investigation. Prothero (1965), for example, investigated the role of migration in malaria transmission. This research was influential in the formation of policy with the World Health Organization (WHO).

The 'quantitative revolution' in geography of the 1950s and 1960s provided an impetus to new techniques of studying disease and health care. Some of the earliest research using more sophisticated quantitative techniques was by researchers such as Gerald Pyle (1971) who investigated the spatial patterns of cancers, heart diseases, and strokes in Chicago. This research was also significant in that it represented some of the earliest research by geographers to relate the spatial patterns of diseases at a metropolitan level to urban structure and to emphasize the geographical patterns of noninfectious diseases.

Spread of Disease

Some of the earliest work in medical geography following the beginning of the quantitative revolution considered the mechanisms by which disease spreads from place to place. Though this is called 'diffusion' in geography because of Hagerstrand's research on the spatial diffusion of innovation, the term 'diffusion' is a misnomer. In the context of disease, 'diffusion' is not recognized. The simpler 'disease spread' is both correctly evocative and universally recognized. Early geographic work on disease spread includes Pyle's study of the spread of cholera in the USA in the nineteenth century (Pyle, 1969), which demonstrated that the spread of cholera reflected the change in regional structure and connectivity. As transportation systems developed and places became better connected from the early to the late nineteenth century, the pattern of spread of cholera changed from a distance-based 'contagious' diffusion pattern to one based more on the settlement hierarchy and therefore usually termed a 'hierarchical' pattern by geographers – another term that breeds misunderstanding in an interdisciplinary context. Since then, disease spread has become such an important theme in medical geography and public health that it is impossible to summarize the research here. In general, the research has become less descriptive and more predictive. It is difficult to discover an infectious disease whose spread has not been analyzed.

GIScience and the Spatial Analysis of Infectious Disease

Pyle's research on cholera was followed by the groundbreaking statistical and mathematical research by Haggett, Cliff, and colleagues on the spatial distribution of measles (Haggett, 2000). Subsequently, these early works led to more sophisticated analysis and modeling of the spatial variation of disease, facilitated by the development of GIS. Well-known to geographers and now considered to be an essential tool in public health, GIS allows the interactive display of many layers of spatial data, so that, for example, knowing the environmental requirements for the survival and propagation of a vector can lead to a precise display of areas where that vector might survive.

Infectious disease models have become more sophisticated in their ability to predict the incidence and prevalence of disease using a GIS model by exploiting such spatial statistical techniques as inverse distance weighting and kriging. These techniques are also valuable for interpolating population data, which is required in larger volumes as disease mapping is better able to represent the heterogeneous data involved in disease mapping (Tatem et al., 2012).

Predictive spatial and epidemiologic modeling has the potential to assign probabilities to where and when epidemics may arise. During an ongoing epidemic, the ultimate test will be to predict with accuracy which places are most vulnerable to disease spread. Such prediction can be essential in deciding where to focus health education campaigns, vaccination, and even quarantine. Advances are being made in this direction (e.g., Matrajt et al., 2013; Chao et al., 2011; Matrajt and Longini, 2010). During the World Trade Organization meeting in Seattle in 1999, real-time symptomatic data, called 'syndromic surveillance,' was made possible for local emergency departments. Subsequently, GIS has been incorporated into such systems on a limited basis, with the goal of improved public health surveillance that incorporates real-time display of spatial data. In some areas, it is possible to query the system for a symptom such as 'vomiting.' A real-time display of emergency department visits with patients who have vomited will result, with the ability to display spatial, temporal, and other epidemiologic patterns. Such developments should provide an early warning that an epidemic is occurring or that a bioterrorist event is in progress. Though it presents ethical dilemmas, effective disease surveillance and control is being facilitated by the ability to trace movement patterns based upon GPS embedded in mobile telephones; novel applications of this technology have been used in Africa and elsewhere as important inputs to tracking a particular epidemic.

Chronic Disease and the Environment

Chronic diseases have a long onset and remain present for a long time or relapse and remit over long periods. Causation is usually multifactorial. Unlike infectious diseases, it is not possible to point to one factor that gives rise to a chronic disease. The Global South faces a 'dual burden' of chronic and infectious diseases. While people may be at high risk for cholera or malaria, they are also at risk for cardiovascular disease and cancer as a result of environmental exposure, changing patterns of nutrition, and decreasing exercise.

Chronic diseases are influenced by both genetics and environment. 'Environment' is typically defined as that which is external to the body. Thus, in most writing in the health sciences 'environment' does not necessarily correspond with earth science and social science notions of environment. The oft-repeated statement that 90% of all cancers have an environmental component is easy to misinterpret. It does not mean that the source of 90% of all cancers has an ecosystem cause. In

epidemiology and environmental health, 'environment' may be used in both senses.

There is a strong need for standard nomenclature that includes the health sciences, the social sciences, and environmental studies. The dichotomy between chronic diseases and infectious diseases is turning out to be false, yet use of the terms as distinct persists in medical geography and elsewhere. In the Global North, where there is adequate access to antiretroviral therapy, HIV/AIDS can usually be effectively treated. AIDS has turned into a chronic infectious disease. Moreover, many of the chronic diseases have turned out to have infectious causes. Classic examples are HPV (human papillomavirus), which causes 90% of cervical cancer cases, and the bacterium *H. pylori*, which causes 90% of peptic ulcers. A more appropriate set of descriptive labels is 'infectious diseases' and 'apparently noninfectious diseases.'

Within medical geography, interest in chronic disease is centered on two foci: (1) understanding the environmental exposure pathways of a particular disease and (2) understanding the roles of social system attributes in introducing or perpetuating these exposure pathways. With respect to the former, medical geographers may work with advanced statistical techniques to better understand the variance in prevalence and incidence rates for a disease in a particular place. Alternatively, similar results may be obtained through epidemiology-based research designs that can help establish exposure pathways of interest with regard to a particular disease. These study designs include case-control, cohort/prospective cohort, and other standard designs.

'Big Data' has the potential to advance understandings of infectious and noninfectious diseases. Again, exposure pathways are key. We are approaching a time when actual exposure to environmental contaminants is possible, so that as an individual goes about for a day, exposure to noxious environments may be measured and patterns may then be analyzed using large spatial datasets. Big Data refers to databases with large volumes of different kinds of data. Research within the relational structures in these databases has the potential to resolve connections between disease and environmental, social, and genetic drivers. Subsequently, these connections will direct future research toward areas of higher potential impact on disease outcome. For example, a recent study in Kentucky between local pharmacies and patients has combined asthma patient data with spatial data. Subjects can use asthma inhalers that are attached to small sensors that transmit the location where each inhalation of medication is taken. In turn, such a novel application allowed the integration of patient location and environmental conditions, including meteorologic conditions (MacDonald, 2012).

Much of this analysis would not be possible without advances in the data storage and modeling capacity of GIS. For example, spatial data obtained through mobile devices can be brought into a GIS and analyzed against neighborhood attributes, demographic data, meteorological conditions, etc. The ability to analyze individual movements, as opposed to fixed locations such as individual addresses, has contributed to expanded definitions of 'neighborhood' and 'community,' which are important units of analysis in medical geography.

Taking advantage of the evolving character of medical data promises new advances in computing the geographic context of both infectious and chronic diseases. However, working with Big Data in medical geography will require advanced database management structures within geography, particularly GIScience. This developing framework parallels foundational work in environmental hazard and sustainability research, two areas where human health is becoming increasingly more apparent as an important domain for planning and impact analysis. The emphasis on electronic health records also has the potential to yield tremendous information in a context such as an urban slum in Ghana, where local health-care workers may not only record health information into a handheld unit that communicates with a server, but also where the data are geotagged. Such an application has the potential to improve vaccination rates and to contribute to preventive medicine through targeted health checks.

One key element in understanding individual risk of exposure to ozone and other environmental contaminants is information on a person's daily movement. GPS tracking in mobile devices, multilayer map displays, and a refined understanding of how an individual's surrounding shapes exposure are all advancements within geography that are expanding understanding of the health impacts of exposure to environmental contaminants.

New concepts like the exposome, defined as the measure of all the exposures of an individual over a lifetime and how those exposures relate to health, are aiding in these efforts. Not only is this concept helpful in predicting individual health impacts it also offers a useful metric for defining the health resiliency of a community. Environmental contaminants, which result from both natural and technological causes, are a growing field of interest within medical geography, an interest that parallels advances in the fields of epidemiology, toxicology, and public health. Such contaminants are transmitted through air, water, or soil, and can initiate, or exacerbate, existing cases of chronic disease.

Ozone, for example, is an airborne chemical formed *in situ* through primary precursors that may be a combination of naturally occurring elements and human-caused pollution. Asthma and COPD (chronic obstructive pulmonary disease) are particularly affected by pollutants since airborne contaminants cause inflammation in the respiratory system. Because of systemic inflammation, particulate and gaseous pollutions have been causally related to the development of atherosclerosis and myocardial infarctions ('heart attacks') and to serious arrhythmias that could be immediately life-threatening. These relationships have been demonstrated for people who are stuck in traffic in polluted conditions.

Social Environment

Medical geographers are also interested in the aspects of a social system that introduce or perpetuate an exposure pathway. Social epidemiologists demonstrated that neighborhood and local areas can themselves be either protective or pathogenic – and insight enabled by using multilevel or hierarchical models, in which outcome is a function of individual as well as areal-level data. This methodological development has been one of the major innovations in understanding since 1990. Theoretically, this understanding

is grounded in sociological, demographic, and urban geographic thoughts.

The insights that spatial analysis has provided in understanding patterns of cancers and cardiorespiratory diseases have been reviewed recently (Mayer, 2009). For diseases of uncertain cause, spatial analysis can be instrumental in identifying causes that covary spatially at the same scale as the disease. Such techniques have facilitated insights into the causes of asthma and other respiratory diseases, environmental etiologies of ischemic heart disease and stroke, and environmental contributors to neurologic disease such as multiple sclerosis (MS). All involve both genetic and environmental factors. Usually, small variations in the host genome are 'turned on' by environmental switches. Gene expression is thus partially determined by the environment, so that a gene that might predispose to MS becomes expressed only with the appropriate environmental exposure. This process is termed 'epigenetics' and is only now beginning to be incorporated into work done by medical geographers.

Political Ecology of Disease

Political ecology – the marriage of cultural ecology and political economy – has had a great impact in understanding the relationships between globalization, power, environmental modification, and their effects on individuals and groups. The advantages of applying political ecology to health and disease were first suggested in 1996 (Mayer, 1996). Disease ecology considers the interaction between environment, population, and behavior (Meade and Emch, 2010). However, disease ecology has not traditionally considered the roles of political and economic power at varying scales. In adding these themes to traditional disease ecology, one arrives at a political ecologic understanding of health, which fundamentally unites social explanations with environmental and behavioral explanations of disease. For example, the 1960s saw the completion of 'Big Dams' as parts of environmental control and economic development projects; two prominent examples are the Aswan High Dam in Egypt and the Akosombo ('Volta River') Dam in Ghana. In many ways, the completion of these dams was beneficial for all concerned, but the formation of large, upriver lakes have had major, unintended health consequences. Because of changing ecologies and greater human exposure to the lake, e.g., Lake Akosombo, for washing, bathing, recreation, fishing, and other economic activities, the prevalence of three major diseases increased upriver, and, to some extent, downriver. These are malaria, onchocerciasis, and schistosomiasis, which are all vector-borne infectious diseases. All are major burdens to health in Ghana generally, causing significant mortality and morbidity.

In one sense, the increases in morbidity and mortality due to the completion of the Akosombo Dam have been caused not only by environmental factors, or even human–environment interaction, but also by strong government influence and economic 'imperative' within the context of the global economy. The dam was constructed to support the well-being of one industry, bauxite, and to some extent, one corporation within that industry. It was strongly pushed by the government of Ghana, both to increase national wealth and to improve quality of life and human well-being. Without electrical power, it would be difficult for either of these to happen.

Conclusion

Medical geography and health geography are intertwined, and there is little advantage to attempting to differentiate between the two. The incorporation of social theory in the 1990s and of models other than the biomedical model in attempting to understand and explain both health and disease added new dimensions. At the same time, further development of biosocial medical geography has yielded insights into disease causation and the pathways of exposure to environments that have the potential to cause disease in populations. Moreover, genomics and molecular methods are beginning to affect medical geography, allowing the researcher to understand specific genetic structures of a pathogen, or of a host, and to understand the environmental 'switches' that result in the expression of genes. Both infectious and apparently noninfectious diseases result from the interaction of genetic and environmental factors. Exposure to environments that generate disease is caused by individual behavior, culture, belief systems, social structure, and many other factors. Geography emphasizes scale, and one of the potential contributions of medical geography is that it provides us with concepts and tools that facilitate integration of knowledge at scales ranging from the molecular to the macro and even global scales. Contemporary developments in spatial data and data structures, as well as genomics and proteomics, may well herald a new age that will yield novel explanations for disease causation. Traditional medical geography, when merged with these emerging theories, techniques, and methodologies, together can result in entirely novel explanations for why population health patterns have their own particular signatures. Ultimately, these developments have the potential to further human health and well-being.

See also: Diffusion: Geographical Aspects; Disability, Geography of; Geographic Information Systems; Geographies of Food: Politics, Place, Power Reprised; Globalization: Geographical Aspects; Health, Geography of; Nature and Society in Geography; Spatial Analysis in Geography; Spatial Data; Spatial–Temporal Modeling.

Bibliography

Chao, D.L., et al., 2011. Planning for the control of pandemic influenza A H1N1 in Los Angeles County and the United States. American Journal of Epidemiology 173, 1121–1130.

Dubos, R., 1965. Man Adapting. Yale University Press, New Haven, CT.

Haggett, P., 2000. The Geographical Structure of Epidemics. Oxford University Press, Oxford.

Learmonth, A., 1988. Disease Ecology: An Introduction. Basil Blackwell Ltd, Oxford.

MacDonald, C., 2012. Using big data to improve health. E: The Environmental Magazine 23, 14–15.

Matrajt, L., et al., 2013. Optimal vaccine allocation for the early mitigation of pandemic influenza. PLoS Computational Biology 9 (3), e1002964. http://dx.doi.org/10.1371/journal.pcbi. 1002964. Epub March 21, 2013.

Matrajt Jr., L., Longini, I.M., November 11, 2010. Optimizing vaccine allocation at different points in time during an epidemic. PLoS One 5 (11), e13767. http://dx.doi.org/10.1371/journal/pone.0013767.

May, J., 1958. The Ecology of Human Disease. MD Publications, New York.

Mayer, J.D., 1989. Clinically applied medical geography: its relevance for travel medicine. The Professional Geographer 41, 421–428.

Mayer, J.D., 1996. The political ecology of disease as a new focus for medical geography. Progress in Human Geography 20, 441–456.

Mayer, J.D., 2009. Medical geography. In: A Companion to Health and Medical Geography. Wiley-Blackwell, New York, pp. 33–54.

Meade, M.S., Emch, M., 2010. Medical Geography. Guilford Press, New York.

Meade, M.S., 1977. Medical geography as human ecology: the dimension of population movement. The Geographical Review 64, 379–393.

Paneth, N., 2004. Assessing the contributions of John Snow to epidemiology 150 years after the removal of the Broad Street pump handle. Epidemiology 15, 514–516.

Prothero, M., 1965. Migrants and Malaria. Longmans, London.

Pyle, G.F., 1969. The diffusion of cholera in the United States in the nineteenth century. Geographical Analysis 1, 59–75.

Pyle, G.F., 1971. Heart Disease, Cancer, and Stroke in Chicago. The University of Chicago, Department of Geography Research Paper Series, No. 134, Chicago.

Riley, D.L., et al., 2013. Neighbourhood walkability and physical activity among family members of people with heart disease who participated in a randomized controlled trial of a behavioural risk reduction intervention. Health & Place 21, 148–155.

Salehi, R., Ali, S. Harris, 2006. The social and political context of disease outbreaks: the case of SARS in Toronto. Canadian Public Policy 32, 373–385.

Tatem, A.J., et al., 2012. Mapping populations at risk: improving spatial demographic data for infectious disease modeling and metric derivation. Population Health Metrics 10, 8–21.

Wu, Xiaobing, et al., 2012. Multilevel and spatial analysis of syphilis in Shenzhen, China, to inform spatially targeted control measures. Sexually Transmitted Infections 88, 325–329.

Medical Profession

David Coburn, University of Toronto, ON, Canada; and University of Victoria, Victoria, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Medicine is a profession. Professions are assumed to be altruistic, self-governing in the public interest and having work autonomy. Medicine rose to a height of power in the middle of the twentieth century when it dominated health care, health policy, and societal definitions of health and illness. This medical dominance has been increasingly challenged by payers, other health occupations, medical industries, and patients. Explanations for a decline in medical power range from pluralistic interest group explanations, through theories of elite power, to arguments that medical power is always dependent on its congruence with the interests of powerful capitalist classes.

Social scientists analyze medicine as an occupation rather than writing biographies of great medical figures or histories of medical discoveries. Emphasis is on the ways in which the production and application of knowledge in helping prevent, cure, or care for illness is an organized social activity.

The term profession implies differences from other kinds of work such as job, position, trade or craft. As individuals, our work is partly determined by our family's social position and, in turn, strongly influences our own income, social status, and class location. As an occupation, medicine reflects the social structures of which it is a part and in turn influences these structures.

Professionals are assumed to be experts in an applied field and are generally of high income, prestige, and social position. Today the ever-increasing number of occupations claiming expertise all try to assume the title of 'profession' and its accompanying privileges. Medicine is often referred to as the archetype profession because it exemplifies the professional ideals of expertise, an altruistic rather than self-interest orientation, high work autonomy, and occupational self-government.

This article focuses mainly on medicine in the English-speaking countries. Yet, even among these nations, differing social structures and varying social science traditions produce somewhat contrasting visions of medicine's historical trajectory. In particular, the United States is an outlier in having less direct state and more direct corporate involvement in medicine. In most other nations, the state has had a much more influential role (Johnson et al., 1995).

Medicine as a Profession

The traditional professions have a history rooted in the craft guilds of the Middle Ages (Krause, 1996). By the nineteenth century, the professions promised a mode of work which was more autonomous, complex, and less alienating than the deadening factory systems of the industrial revolution. Freidson (2001) has pointed out that bureaucracy, business, and professions are contrasting work principles, bureaucracy characterized by hierarchy and routine, business by material gain within markets, and the professions by claimed expertise, self-government, and an altruistic orientation to their service. These characterizations are overdrawn but do provide a flavor of the ideals, if not the reality, of such types of work.

From the 1930s to 1960s, the social sciences largely accepted the professions' own view of themselves as ethical and self-governing in the public interest. Actual studies of medicine, however, indicated that the profession was not as rigorously self-regulating or altruistic as its practitioners claimed (Coburn, 1999; Freidson, 1970; Larkin, 1983; Larson, 1977; Light, 2010; Starr, 1982; Willis, 1989). By the latter part of the twentieth century, the medical profession was viewed as exploiting its monopoly powers in order to reduce competition, to secure higher incomes, to control or exclude competing occupations, and to protect, rather than correct, its own miscreants. The core of professional identity came to be defined as autonomous work and professionalization as a collective strategy for attaining and maintaining control over a set of tasks (Larson, 1977).

Sociologists began to focus on the power of medicine to orient medical care as much to its own interests as to the health needs of patients or populations, hence the term 'medical dominance' (Freidson, 1970). Dominance refers to the ability of medicine to control its own work, the activities of other health occupations, the way medical tasks are socially organized and delivered and even the very definitions of what constitutes health and human well-being. The term dominance, however, may be interpreted in strong (medicine is hegemonic) or weak (medicine is an intermediary power) versions leading to contrasting explanations for medical power. There was an increasing skepticism about professional altruism (Saks, 1995) which some, however, feel was overdrawn (Strong, 1979).

Medicine, a Heterogenous Profession in a Complex Service Industry

To speak of 'the medical profession' in contemporary times is to ignore the profession's internal and international variation. Within nations there are dozens of medical boards, associations, regulatory agencies, and specialty groupings. Public health, academic, and curative medicine have divergent aims and somewhat competing philosophies. Physicians work in a variety of organizations and settings which differentially shape their

work and policy orientations. Moreover, medicine is now practiced within highly complex divisions of labor in which the image of the doctor with the little black bag plays a part only as romantic history. Rather, there is a 'medical industrial' complex consisting of the pharmaceutical industries, hospital 'factories,' medical supply and equipment industries, and hundreds of nonphysician occupations and professions, from massage therapy through computerized medical imaging. The drug industries are powerful and increasingly expensive drivers of the health field, its research directions and applications. In a number of the developing societies, in addition to western-type medical personnel, there persist a variety of indigenous healers. This curative complex is attached to various kinds of private and public payment systems. In the world's richest 38 nations of the Organization for Economic Cooperation and Development (OECD), total expenditures on health average 9.5% of Gross Domestic Product (Australia 9.1; Canada 11.4; United Kingdom 9.6), with the United States being the high outlier at 17.6% yet having relatively poor average population health (OECD Health Data, 2012). In nations with government provided or subsidized care, health, along with education, consumes the largest proportion of government budgets hence the relationships between medicine and the state are critical.

In addition to formal curative care, there is a vast social complex consisting of health-related areas from nonorthodox or alternative healing to the diet, wellness, mental/emotional well-being, and fitness industries. Well-being, bodily appearance, and longevity are at the center of concern within all of the advanced nations. Modern societies are said to be 'medicalized' although the medical profession is only one of the many sources of this medicalization with the drug industries a major source. We are living in a 'risk society.' Almost any activity now carries with it an assessment of its possible threat to health. Moreover, in a secular and materialist society, human life is not viewed as preparation for an afterlife but as an end in itself. There is a desperate desire to delay death, avoid or control risks, and to appear and be, young and vigorous, or at least not old and decrepit. In all of these areas, medicine and its surrounding services have something to say. However, while we can speak of medicine, the state or other institutions none of these actors are unitary or impermeable. All of them are heterogenous and perfused with influence from 'the outside' by other actors in complex networks. For example, medicine and the state have symbiotic interrelationships and influence one another's internal functioning.

Medicine is always part of, rather than 'above' society. In the nineteenth and early twentieth centuries, medicine was overtly sexist reflecting the patriarchal social structures of the time. At various times, women, Jews, and blacks were prevented from entering medicine. In the nineteenth and early twentieth centuries, women patients were viewed as fragile victims of their own uteruses (Ehrenreich and English, 1972). Before and during World War II, the profession in Germany actively displayed antisemitism and physicians took part in the Holocaust. At all times in the modern world, the poor tend to be less well-represented in medical ranks than are the rich or middle classes. Under state command economies, medicine sometimes, as in psychiatry, directly served the purposes of the state in oppressing or subduing citizens. In Europe and North America, medicine early on was eugenicist and in accord with the norms of the time some physicians sterilized the mentally ill and those with low cognitive functioning. Later, tranquillizers often used by doctors to treat depression in women also served, it is claimed, to 'adjust' women to what many perceived as oppressive social roles. In numerous societies, while there have been many insightful and progressive individual medical practitioners, the organizations representing physicians have most frequently had conservative political views and have generally opposed such welfare state measures as state health care or universal health insurance measures. Given the tendency to reflect the interests, values, and regulatory regimes of the era in which it is embedded Foucaultians see medicine as part of the movement to control and regulate populations (Petersen and Bunton, 1997; Turner, 1995).

Internationally, medicine ranges from doctors using basic medical equipment to highly skilled experts employing complex computer assisted procedures. It encompasses practitioners who routinely accept bribes to high status doctors catering to rich clients. While some countries or regions have a physician oversupply others are underserviced. Internationally, 47% of the member nations of the World Health Organization (WHO) have less than one physician per 1000, many nations in Southern Africa have less than one per 10 000 population. At the other end of the spectrum, a number of nations have more than 40 per 10 000 physicians, including, for example, Austria, Cuba, Georgia, Kazakhstan, and the Russian Federation. The United States has 24, the United Kingdom 27, Australia 30, and Canada 20 physicians per 10 000 population. Despite the persistence of indigenous healing systems, globalization has been accompanied by an almost universal reverence for highly technical medicine to which elites around the world want access. Such a reverence for high-tech curative care constitutes a problem for poor nations.

Historical Development

In the second half of the nineteenth century, medicine was internally fragmented and fractious, with various competing healing modalities ranging from allopathy to homeopathy to eclectics. Medicine was composed almost entirely of males and faced with a variety of 'external' competitors from bone-setters to midwives within pluralistic healing systems. Many lay persons thought they knew as much about health and illness as did physicians. While anesthetics and aseptic techniques had helped to improve surgical procedures, medical therapies generally were not far removed from the bleeding, blistering, and purging prevalent earlier.

By the middle of the twentieth century, in less than 100 years, medicine had risen from its previously lowly position to one of social and cultural authority (Starr, 1982). Physicians controlled key health care institutions and were the acknowledged experts in what constituted disease and what should be done about it. Medicine first unified and then established its dominant position in health care (Larson, 1977). In mid-century, unlike in earlier times patients generally benefitted rather than suffered from visiting a doctor. Medical education was lengthy, rigorous, and science- and technology-based. Entrance into medical school was highly prized, competitive, and exclusive. Physicians had high status, unprecedented autonomy and authority, and high incomes.

Medicine heavily influenced definitions of health and illness and largely controlled both other health occupations and health policy.

The major trend for medicine has thus been that of an increasingly scientifically sophisticated, powerful, and efficacious profession. There are, however, dissenting views about the positive effects of orthodox medicine and its emphasis on curative therapies. Medicine has possible iatrogenic effects on the individual, the institutional, and the social level (Illich, 1975). Powerful technologies and drugs often have negative side effects, which threaten to overwhelm their positive influence. For example, the proliferating use of antibiotics creates drug-resistant organisms. Medicine still is more effective at treating injuries, trauma, and acute disease than it is at curing the rising tide of more chronic diseases in aging populations. A biologically based curative medicine individualizes the social causes of illness, hence leading to a de-emphasis on the social determinants of health and on preventive medicine and health maintenance. Professional power may rob the public of the inclination to view themselves as responsible or competent regarding their own health.

Regarding the aim of medicine, improved health, and well-being, social analysis indicates that more doctors and more of health care are not necessarily the major factors leading to the improved health of populations (Evans et al., 1994; McKeown, 1979; Szreter, 2002). Most recently, WHO has published a report on the Commission on the Social Determinants of Health Final Report (2008) which contends the ways people are brought up as children, the way they live and work, rather than health care, are the important determinants of population health. The WHO Report directly implicates unfair economic arrangements and 'bad politics' in producing health inequalities within or between nations. Within these new paradigms medicine and health care, focused largely on individual cure, are less intellectually salient than they once were although they are certainly not less socially or politically central.

Medical Dominance: Medicine, Central and Powerful yet Challenged

The chief focus of social scientists in the past few decades has been on the power of the medical profession to ensure that health care policy and procedures suit their own interests, that is, medical dominance. This area of interest, introduced most prominently by Freidson, was later directly or indirectly tackled in various nations (Coburn, 1999, 2006; Hafferty and McKinlay, 1993; Larkin, 1983; Light, 2010; Starr, 1982; Tousijn, 2002; Willis, 1989, 2006). However, the questions arose – if medicine was dominant within health care, if it indeed controlled its own work, other health occupations, health policy and the way societies viewed health, has medicine always been dominant and does medical dominance vary in different national contexts? With the advent, before and after World War II, of the rise of welfare states, national health systems and subsequent focuses on health care costs, a major trend of challenges to medical dominance and a possible sequence of 'the rise and fall' of medical power ensued (see special editions of Health and Social Behavior, 2010; Health Sociology Review, 2006; Health Politics, Policy and Law, 2004).

It is accepted that medicine was uniquely powerful in the mid-twentieth century and that there have been challenges to that power ever since although there are still disagreements about whether or not medicine was 'ultimately' dominant. Challenges have arisen from other health occupations struggling against medical control or scopes of practice, that is, 'turf wars,' by corporate and state payers regarding the costs of medical services, by patients over problems with medical self-regulation and discipline, by hospital complexes, and by industries, such as the pharmaceutical industry, which attempt to shape health care to their own interests. It is now felt that medicine, while still powerful, is somewhat less so than it was, although the degree of decline is more obvious regarding health policy than in some other areas. As Tousijn (2002) points out, declines in power may vary by the dimension of dominance one is examining. In what follows the historical rise of, and contemporary challenges to medicine are traced through a brief analysis of medical knowledge and practice. We then note varying explanations of such historical trajectories.

Medical Knowledge

Doctors in the eighteenth and nineteenth centuries claimed authority, not on the basis of medical discoveries, but because of their gentlemanly status. This authority was underwritten by attendance at elite universities and knowledge of classical learning. Medicine was a male middle-class enterprise. Women, minority religious or racial work groups, and the working class were largely excluded.

In the nineteenth and early twentieth centuries, medicine increasingly asserted authority because it claimed association with the rapidly rising prestige of science. But, this was a particular kind of medicine; a medicine based on individual biology and the germ theory gained ascendance over the public health medicine and various health-related political movements rather focused on the social roots of illness in poverty and poor living conditions.

Foucault emphasized the inseparability of knowledge and power. Control over the production of knowledge as well as credentialism are certainly two of the strategies used by medicine to gain and then maintain control over health care. The profession administers the production of knowledge in medical schools and health science complexes and gains control over it even though much of this research is carried out by basic scientists rather than physicians. Physicians have a monopoly over direct access to patients and to prescription drugs, crucial aspects of the research endeavor.

Recent developments have undermined the authority of medical claims to expertise. Physician and nonphysician experts alike question the efficacy of particular medical procedures. Increasing numbers of clinical epidemiologists and others, often at the behest of governments, are busy formulating what medical procedures work and what do not. Even individual clinical work can be routinized through medical guidelines or protocols. In this respect, governments, seeking to rationalize care, are permeating and shaping medicine through the coop-tation of medical expertise itself.

In a neoliberal era, the drive by private and public payers alike to control costs has led to attempts to 'rationalize' care.

The current drive to reform health care has weakened the links between medical knowledge and medical authority both regarding medical control over health care systems and over clinical work. Economists, planners, and managers claim more expertise in health care policy than do physicians. State initiated medical protocols and clinical guidelines shape what doctors do in their day-to-day work. Some of these controls are aided by managers 'using' some physicians in managerial or scientific roles to help control others. There may be a restratification within medicine in which physician managers and academic and scientific medical experts are somewhat at odds with day-to-day practitioners.

The relationship between medicine and the social sciences has itself changed. Originally, medicine used the social sciences to accomplish its own tasks. Typically, social scientists were asked to help understand why some patients did not 'comply with' medical regimes. But the social sciences eventually began to take a more critical approach to medicine by challenging the biomedical paradigm, through their focus on medical power and self-interests, and for their support of the view that health is determined as much or more by social factors as it is by health care. Social science analyses once 'in' medicine are now mostly 'of' medicine.

The foundations of medical knowledge are also being challenged by new views of science in which knowledge is viewed as being socially constructed rather than science directly reflecting 'nature' (Lupton, 2000). Biomedicine is not viewed as scientifically neutral but as a truth claim among others in contests over power. From this perspective, the profession is also conceived as part of state attempts to control populations (Petersen and Bunton, 1997; Turner, 1995).

Medical Practice

Until relatively recently, doctors worked by themselves within private markets for health care. In the eighteenth and nineteenth centuries, doctor's care was often a last resort and hospitals were largely a place for the poor to die. In any event, few could afford physician fees. With improved perceived efficacy and the rise of a market for health care, physicians emerged as the preeminent free-enterprise professionals and adopted a corresponding individualist ideology.

In the twentieth century, as part of the movement toward the welfare state, the provision of medical care became a public concern. Internationally, many forms of payment mechanism and systems developed, from health care systems as in Britain to national health insurance as in Canada to more private and fragmented systems as in the United States. These systems influence what physicians do, where, and for how much (Hafferty and McKinlay, 1993; Johnson et al., 1995). Though medicine resisted the intrusion of state or corporate power or tried, with some success, to shape it, everywhere there has been more external influence than there was previously over the nature, conditions, and quality of practice.

Medical autonomy is being undermined in both private and publicly financed systems. In public systems, once some form of control over physician costs has been instituted, the state tends to leave purely clinical matters in the hands of the profession. Ironically, private systems may be more challenging to medical power than public systems because corporations are less open to medical political manipulation than are politicians. Yet, general government health care policies, such as the almost universal limitations over the use of technologies or payment methods, have a profound impact on what individual physicians do at the clinical level. More generally, state-profession relationships may now be characterized as 'mesocorporatist.' That is, in a number of jurisdictions the state seeks to use medical organizations to control medicine itself. For example, medical regulatory organizations depend on state power to control the body of the profession itself which in turn seeks to use regulatory organizations to control physician behavior.

The heart of professional standing is self-regulation, that is, the implicit promise by professions that, in exchange for extensive autonomy, the occupation will 'govern itself in the public interest.' Yet, within the market and outside of it, by competition, by corporate fiat and/or by state regulation, medical practice is subject to more surveillance and regulation including through reformed medical institutions of self-government. There is a deepening division in many jurisdictions between the function of representing the profession, the role of professional associations, and that of licensing and regulation, the function of medical 'colleges' or governmental institutions. Sensational examples of physicians abusing patients forced some governments, as in the United Kingdom, to constrain professional self-regulation which had failed to prevent medical crimes or misconduct (Dixon-Woods et al., 2011). Those medically dominated self-regulatory organizations that remain, even if still staffed mainly by physicians, are more constrained by public or governmental direction, regulation, or representation than they were previously (Allsop, 2006; Coburn, 1999). Just as medicine once permeated state centers of power in order to control health policy, now there is a major flow in the other direction, states shape medicine through coopting key medical organizations. Internal fragmentation made more difficult the efforts of medical elites to mobilize the profession to face its political, economic, and social challenges.

The last third of the twentieth century witnessed the rise of neoliberalism and globalization accompanied by attacks on the role of the state in the provision of social, health, and welfare services and the decline or transformation of the welfare state. This period was quickly followed by economic crises and austerity. Health care systems became more controlled, rationalized, and privatized. Under neoliberal regimes as in Britain in the 1970s, the professions were viewed as monopoly conspiracies against public 'consumers.' Though in various areas of the world neoliberalism is in decline, today business and states are in an implicit or explicit coalition to control health care costs and hence the work of physicians. Medical practice has thus historically moved from a situation of patronage, in which clients controlled physicians, to one of collegiate or colleague control in a medically dominated system, to one in which doctor–patient relationships are mediated by (public or private) third parties (Johnson, 1972). It has been argued that new forms of 'governmentality' involve states attempting to instill in both patients and providers new ways of viewing the self, that is, characteristics of self-responsibility and autonomy, which enables governments to escape from

actually providing necessary health resources and which shifts the political blame for health problems onto citizens themselves (Dent, 2006).

Understanding Medicine and Changes in Medical Power

What do developments in the last century and a half tell us about the medical profession? There is more agreement regarding the general trends than there is about how to explain these. Medicine rose to power in the nineteenth and early twentieth centuries. By the mid-twentieth century, medicine was one of the most powerful actors within health and health care. Other health occupations were absorbed, subordinated, limited, or excluded from the official health care system (Willis, 1989). Powerful within health care systems, medicine had important effects on the societies in which it was not only embedded but also reflected trends within the society of which it was a part. Yet, medical power was always contested and medicine was never all-powerful.

Today, however, medicine faces challenges from 'countervailing powers' (Light, 2010). Patients use unorthodox forms of healing, demand particular forms of care, and complain about or sue physicians. Health care occupations, from nursing to physiotherapy to chiropractic increasingly seek autonomy, and chip away at the borders of the medical monopoly. A proliferating variety of forms of unorthodox healing, from shiatsu to chiropractic, claim legitimacy in the 'system of professions' (Abbott, 1988; Saks, 1995). States and corporations, public and private payers come to shape medical work. The contribution to population health of a technologically sophisticated medicine, as opposed to improvements in living conditions, income equality or the social determinants of health, on human well-being or the health of nations is not as unequivocal as it once appeared. Moreover, just when medicine is facing external challenge, it is weakened by internal fragmentation by specialty and gender and by a developing hierarchy between an academic and research elite and practitioners. The major issue for social science analysts has thus become to understand a historical trajectory of rise, consolidation, and challenge or possible decline of medical power.

Some social scientists have used terms such as proletarianization or corporatization (McKinlay and Stoeckle, 1988) to depict trends diminishing medical power beginning in the latter part of the twentieth century. Proletarianization refers to processes similar to those that affected workers in the nineteenth century as skilled workers became employees and their work came to be controlled, routinized, and fragmented. Medicine, previously an autonomous and powerful profession has become subject to state, corporate, and bureaucratic imperatives. Though the term proletarianization is overdrawn at the very least, doctors now have to answer for the results of their work if not the processes involved, that is they are subject to processes of 'responsibilization' or 'performativity' (Dent, 2006).

A common-sense explanation for medical power is that medicine became authoritative as it gained in scientific knowledge and became more effective in curing disease. However, while improving effectiveness, or at least the perception of such efficacy, was important (doctors have to be sought out by patients), historians have noted that medicine became cohesive and dominant before it became more efficacious (Wailoo, 2004). Efficacy based theories of the rise of medicine also have difficulty explaining recent weakening medical power, though changing disease patterns from the predominance of infectious to chronic disease may be important.

Freidson's emphasis on medical dominance in the 1970s stimulated broader debates about the reasons for the rise of medical power but had little to say about future challenges. Pluralist interest group theories, or forms of Weberian-based closure theory (Murphy, 1988; Witz, 1992) which stress medicines' own struggles for superiority, contrast with structurally oriented theories which emphasize the rules within which interest group struggles occur (Abbott, 1988; Murphy, 1988; Navarro, 1976, 1988; Witz, 1992) or which contend that medical power is always subordinate to the power of dominant classes (Navarro, 1976).

Closure theory describes the strategies of demarcation of occupational boundaries and of exclusion used by medicine in attaining and maintaining its dominance over other health occupations as well as the contemporary struggles by other occupations to escape medical control. Closure theory though focuses almost entirely on interprofessional relationships, only one dimension of medical dominance. By contrast, structural theories claim that medical power is fundamentally anchored in the congruence of its interests and ideologies with those of dominant elites, industries, classes, and the state. Medicine in this view is a powerful intermediate source of power but not dominant. Navarro (1988), for example, argues that medicine was never dominant nor is it now being routinized. Rather, medicine, and a particular form of individually focused curative medicine, was preferred by a rising industrial capitalist class opposed to political movements pointing to poverty and poor working conditions as the source of illness. Individualist-oriented explanations drew attention away from the social causes of disease hence served the interests of industrial capitalists. Medicine's contemporary decline is also viewed in class terms involving a divergence between dominant class or corporate interests in saving costs and those of medicine to retain control over care (Coburn, 1999). A dividing line does exist between those social analysts focusing on a pluralist or power elite version of medical history and those paying more attention to class and capitalism as a mode of production. In general, more actor-oriented views imply that medicine imposed biomedicine on society, while structural analyses, whether power elite or Marxist, focus more on the 'selection' of a particular type of medicine by prevailing structures of power, class, gender, or race.

Any perspective on contemporary developments has to come to terms with broader economic and political developments accompanying change from the neoliberalism of the last quarter of the twentieth century to the economic and political crises of the early twenty-first century. In the social science literature, perhaps significantly, the medical profession and the concept of professional dominance are not as prominent as they once were – economic and political crises have pushed the fate of medicine onto the back-burner. Such a loss in centrality may also be another signal of declining medical power.

Conclusions

Medicine is like other forms of work and subject to many of the same general pressures. Physicians are both active in shaping and reproducing particular social structures yet the profession cannot escape being subject to external powers. Most recent social studies of the medical profession are oriented to its power, to changes in that power and evince skepticism regarding professional claims of altruism. The focus is on the self-interests of the profession and its changing accommodation with powerful social classes, elites, or institutions. This view contrasts with perspectives which emphasize the many individual (Norman Bethune in China) and collective (Medicine Sans Frontieres) acts of heroism and self-sacrifice of physicians or groups of physicians. However, the social science emphasis on self-interests and power is a little one-sided, medicine is not only self-interested. Moreover, a health care controlled by private corporations, managers, or the state, may presage a worse future for patients than one in which medicine, or at least individual physicians, retain a 'relative autonomy.'

Medicine rose during a particular period of industrial capitalism and flourished, perhaps had its golden age, during the period of the development of the welfare state at a time when working class pressures were at the height. In that milieu, medicine could partially reconcile both service to others and its own self-interests. We are today in a new era, one which began with a greater emphasis on individualism and on markets and is now enmeshed in crises in which corporate elites or classes are clearly in control, though not unchallenged. The ramifications of these changes for medicine imply greater constraints on medical prerogatives and do not presage a return to medicine's previous heights.

Medicine today is revered yet envied, sought after but viewed with skepticism, scientific yet collectively saturated with self-interests and motives not pregiven by its science or service, with many altruistic practitioners yet apparently mercenary organizations. The challenge for medicine is to fulfill its professional promise of service to others in an era of austerity, profit-seeking, and individual materialism. The social sciences focused on medicine are today in somewhat of a theoretical doldrums, still obsessed with medical dominance and interprofessional rivalries, but facing the broader task of providing an understanding of medicine more adequate to recent societal transformations and contradictions.

See also: Health Care Markets: Theory and Empirical Evidence; Nonphysician Health Professionals; Occupational Health.

Bibliography

Abbott, A., 1988. The System of Professions. The University of Chicago Press, Chicago, Illinois.

Allsop, J., 2006. Medical dominance in a changing world: the UK case. Health Sociology Review 15 (5), 444–457.

Coburn, D., 1999. Phases of capitalism, welfare states, medical dominance and health care in Ontario. International Journal of Health Services 29 (4), 833–851.

Coburn, D., 2006. Medical dominance then and now: critical reflections. Health Sociology Review 15 (5), 432–442.

Commission on the Social Determinants of Health Final Report, 2008. Closing the Gap in a Generation: Health Equity through Action on the Social Determinants of Health. WHO, Geneva.

Dent, M., 2006. Disciplining the medical profession? Implications of patient choice for medical dominance. Health Sociology Review 15 (5), 458–468.

Dixon-Woods, M., Yeung, K., Bosk, C.L., 2011. Why is UK medicine no longer a self-regulating profession? The role of scandals involving 'bad apple' doctors. Social Science and Medicine 73 (10), 1452–1459.

Ehrenreich, B., English, D., 1972. Witches, Midwives and Nurses: A History of Women Healers. Feminist Press, New York.

Evans, R., Barer, M., Marmor, T. (Eds.), 1994. Why Are Some People Healthy and others Not? Aldine de Gruyter, New York.

Freidson, E., 1970. Profession of Medicine. Dodd, Mead and Co., New York.

Freidson, E., 2001. Professionalism, the Third Logic: On the Practice of Knowledge. University of Chicago Press, Chicago.

Hafferty, F.W., McKinlay, J.B. (Eds.), 1993. The Changing Medical Profession. Oxford University Press, New York.

Illich, I., 1975. Medical Nemesis: The Expropriation of Health. McClelland and Stewart and Marion Boyars, London.

Johnson, T., 1972. Professions and Power. Macmillan, London.

Johnson, T., Larkin, G., Saks, M. (Eds.), 1995. Health Professions and the State in Europe. Routledge, London and New York.

Krause, E.A., 1996. Death of the Guilds: Professions, States and the Advance of Capitalism. Yale University Press, New Haven, Connecticut.

Larkin, G.V., 1983. Occupational Monopoly and Modern Medicine. Tavistock, London.

Larson, M.S., 1977. The Rise of Professionalism. University of California Press, Berkeley, California.

Light, D.W., 2010. Health care professions, markets, and countervailing powers. In: Bird, C., Conrad, P., Timmermans, S. (Eds.), Handbook of Medical Sociology. Vanderbilt Press, Nashville, Tennessee.

Lupton, D., 2000. Social construction of medicine and the body. In: Albrecht, G.L., Fitzpatrick, R., Scrimshaw, S.C. (Eds.), Handbook of Social Studies in Health and Medicine. Sage, London, Thousand Oaks, California.

McKeown, T., 1979. The Role of Medicine: Dream, Mirage or Nemesis? Basil Blackwell, Oxford.

McKinlay, J.B., Stoeckle, J.D., 1988. Corporatization and the social transformation of doctoring. International Journal of Health Services 18, 191–205.

Murphy, R., 1988. Social Closure: The Theory of Monopolization and Exclusion. Clarendon Press, Oxford.

Navarro, V., 1976. Social class, political power and the state and their implications in medicine. Social Science and Medicine 10, 437–457.

Navarro, V., 1988. Professional dominance or proletarianization?: neither. The Milbank Quarterly 66 (Suppl. 2), 57–75.

Petersen, A., Bunton, R. (Eds.), 1997. Foucault, Health and Medicine. Routledge, London and New York.

Saks, M., 1995. Professions and the Public Interest: Medical Power, Altruism and Alternative Medicine. Routledge, London.

Starr, P., 1982. The Social Transformation of American Medicine. Basic Books, New York.

Strong, P.M., 1979. Sociological imperialism and the profession of medicine: a critical examination of the thesis of medical imperialism. Social Science and Medicine 13A, 199–215.

Szreter, S., 2002. Rethinking McKeown: the relationship between public health and social change. American Journal of Public Health 92 (5), 722–725.

Tousijn, W., 2002. Medical dominance in Italy: a partial decline. Social Science and Medicine 55, 733–741.

Turner, B.S., 1995. Medical Power and Social Knowledge, second ed. Sage, London.

Wailoo, K., 2004. Sovereignty and science: revisiting the role of science in the construction and erosion of medical dominance. Journal of Health Policy. Politics and Law 29 (4), 643–659.

Willis, E., 1989. Medical Dominance: The Division of Labour in Australian Health Care, second ed. George Allen and Unwin, Sydney (first ed., 1983).

Willis, E., 2006. Taking stock of medical dominance. Health Sociology Review 15 (5), 421–431.

Witz, A., 1992. Professions and Patriarchy. Routledge, London.

Medical Sociology

David Mechanic, Institute for Health, Health Care Policy and Aging Research, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Medical sociology builds on substantive areas of sociology. A dominant concern has been the distribution of health; illness and disability; services use; and class, race, ethnic, and age disparities. The field examines the organization and provision of service as shaped by culture, science and technology, and economic organization. The workforce, especially physicians, has been a core interest as has competition among professions for control over spheres of work. Other key areas include clinician–patient interaction and stress and coping. Attention focuses on the economic and ideological basis of U.S. health care, its exceptional expenditure levels and efforts to constrain costs.

Many individual sociologists in the first half of the twentieth century studied the distribution of illness, the plight of the sick poor, and physicians and other health care practitioners (Bloom, 2002), but medical sociology first emerged as a distinctive area of sociology in the 1950s, first in the United States, and later in many other countries. As a subdiscipline, it drew on the methodologies and middle-range theories of the three then-main areas of sociology: social organization, social psychology, and ecology and demography. Relevant subfields included social stratification, occupations and professions, political sociology, gender and race, complex organizations, deviant behavior, and socialization (Mechanic, 1978). Over the last half of the twentieth century, medical sociology came to share concepts and methods with related social and medical sciences including public health, health services research, medical economics, social epidemiology, health psychology, health geography, medical ethics, medical history and politics, and mass communications. As we proceed in the twenty-first century, the boundaries among the disciplines become more permeable and more blurred, and increasingly investigators study interconnections between social and social psychological factors and biological processes, including genetics, molecular biology, immunology, and psychophysiology.

Medical sociology from the 1960s in the United States developed rapidly, stimulated by the support of the National Institute of Mental Health (NIMH). The NIMH endorsed a public health approach to mental health issues and defined sociology as among the important basic sciences for mental health progress. The dominance of the NIMH funding helps to account for the disproportionate emphasis on mental health research in medical sociology's early development. By the late 1970s, medical sociology had become one of the largest sections of the American Sociological Association, exceeding 1000 members. In more recent decades, most institutes of the National Institutes of Health (NIH) have funded some medical sociological research and training.

Social Stratification and Health

A dominant theme in medical sociology has been the social distribution of health, illness, and disability and access to and use of relevant services. Much of the literature has focused on social class (usually conceptualized as socioeconomic status (SES)), gender, race and ethnicity, and age as associated with patterns of health and illness behavior, illness risk, distress and suffering, morbidity, mortality, and disability. On the services side, interest has focused on access to health care services, utilization of care, patient–clinician interaction, health outcomes, subjective well-being, and health-related quality of life.

Concerns with social stratification are now encompassed in the growing centrality of the field of population health (Mechanic, 2007), an interdisciplinary and cross-national endeavor to understand the determinants of health across geographic areas, nations, and cultures, with research focused on broad determinants, such as standards of living, poverty, inequalities, and globalization, to more microinfluences, such as nutrition, health behaviors, family structures, social networks, and social capital.

Socioeconomic Status and Health

Medical sociologists have given much effort to specify and test pathways through which social stratification, and especially SES, affects a range of health-related behavior, exposures to environmental risks, and mortality, morbidity, and other outcomes. SES, and its primary indicators, education, income, and occupation, consistently have been among the most powerful predictors of health care outcomes, such as maternal and infant mortality, chronic disease, disability, longevity, effective health behavior, and access to sophisticated and cutting-edge interventions. Much of the research in related fields as well as sociology has focused on specific risks and disadvantages correlated with SES as precursors to important health outcomes, but sociological work has persisted vigorously in the tradition of Durkheim seeking to elucidate distinctive sociological influences.

Link and Phelan's theory of social class as a 'fundamental cause' of health outcomes represents one such important effort (Phelan et al., 2010). They theorize that SES, and the power, money, prestige, information, and social connections associated with it, provide flexible resources that allow individuals and groups to be aware of health risks, live in environments and neighborhoods that protect against risk, take additional preventive actions when available, and have access

to high-quality interventions. SES is seen as fundamental in that while risk factors change over time, these flexible resources continue to provide advantage whatever the modifiable risks. Thus, they maintain, these resources allow control over a wide range of risks and diseases and are adaptable over time as risk factors and disease patterns change. Link, Phelan, and colleagues have derived a number of testable ideas from the theory of fundamental cause that provides greater explanatory power than risk factor theories that predominate in social epidemiology.

For example, Link, Phelan, and their collaborators observed that SES maintains its predictive effects over long time periods during which common risk factors change, that SES has little influence on nonpreventable disease, and that relevant health outcomes cannot be explained by measured intelligence, highly correlated with SES and commonly offered as an explanation for the predictive ability of SES.

Researchers have made efforts to establish which of the components of SES are most important for health. Although education, income, and occupation are highly associated, these factors are differentially important for varying health challenges. Education, for example, may be most important for the quality of parenting and protection of child health, while income is most important for quality of housing or neighborhood. In multivariate analyses, each additional component of social class commonly contributes additional explanatory power. From a sociological perspective, these indicators are proxies for theoretical conceptions of social class, which function more holistically to denote structural and hierarchical relationships.

Gender and Health

Gender as a major determinant of health encompasses the variations in biology, social roles, responsibilities, inclinations, normative behavior, styles of living, and interpersonal associations that differentiate the sexes. Sociologists have long been interested in the advantages in longevity that women in industrial nations have relative to men, their greater willingness to report illnesses and to use medical care, and many differences in illness trajectories. By way of illustration, I focus on two areas of persistent sociological interest, female advantage in longevity and the considerably greater prevalence of depression among women.

Women generally live longer in industrial countries than men with a pattern of favorable survival from birth onward. This is in contrast to some developing countries where males are much preferred and where selective abortion, differential feeding and nurturance, and even infanticide may favor males. Studies provide evidence for a biological advantage for women although many of the persistent sex differences can be explained by differential exposure to risk factors among men and women and stronger protective factors favoring women.

At the beginning of the twentieth century, average life expectancy in the United States was around 50 years with a female advantage of almost 3 years, but by 2006 expected length of life from birth had advanced to 75.1 years for men and 80.2 years for women. Similar patterns are evident in other industrial nations. In recent decades, however, advances in longevity in the United States while continuing, have not kept pace with many other industrial countries, and particularly for women. By 2007, expectation of life among American women (80.1), tied with Denmark for last place among 10 countries studied, ranging among the other eight nations from 86 in Japan to 81.7 in England and Wales. A recent study by the National Academy of Sciences of high-income countries, seeking to understand the relative American disadvantage, examined a range of factors to the extent that data were available, including smoking, obesity, physical activity, social networks, health care services received, hormone therapy, and inequality (Crimmins et al., 2010). The consequences of smoking which began earlier and more extensively in the United States was estimated to explain 78% of the female gap in life expectancy with other countries between 1950 and 2003, and 41% of the male gap. Although there is more uncertainty concerning obesity, data suggested that obesity in the United States also has had a significant role (Crimmins et al., 2010). Underlying these demographic analyses are the crucial roles of culture and social norms that have affected habit formation and its timing between men and women in different nations. Because smoking prevalence has fallen more sharply in the United States than other countries in recent years, we can anticipate the gap in smoking-related mortality to close in future years.

A large literature documents the higher prevalence of depression in women relative to men (in the range of 200–300%) in modern nations, and in many developing countries as well. The female excess is apparent by adolescence and continues throughout the life course. Although many hypotheses have been suggested and a variety of factors may be relevant, there is overwhelming evidence that social factors play a major causal role. A consistent pattern persists whether investigations examine depressive affect, clinical diagnoses of depression, help-seeking for depression, use of depression medications, or hospital inpatient admissions for depression. It appears that the excess prevalence of depression relative to men is balanced by excess prevalence of antisocial behavior, substance abuse, explosive behavior, and violence among men. These observations are consistent with a view that these two types of disorders, internalizing (intrapsychic pain) and externalizing (antisocial behavior), represent alternative modes of expressing distress and emotional tension and are not simply discrete entities each with independent causes.

Race and Ethnicity

Work in medical sociology has focused on disparities among the largest racial and ethnic subgroups in the United States: Caucasians, blacks, Hispanics, and Asians. Much of the literature rejects the construction of racial categories themselves and particularly those routinely used in national statistics and population studies. With greater understanding of population genetics, these conventional rubrics are seen as poor proxies for genetically distinctive subgroups. Moreover, the conventional categories themselves group heterogeneous and diverse subgroups that share little geographic propinquity or cross-breeding. For example, the Hispanic category spans diverse populations from Mexico, Cuba, Puerto Rico, Spain, and other Spanish-speaking cultures. Moreover, when focused on the U.S. population, added diversity comes from selective migration

that varies from one historical period to another such as the pre- and post-Mariel migrations from Cuba. Similarly, vast differences among blacks from varying regions of the world commonly are lumped into a single category. Occasional studies have the specificity and funding to examine the substantial subcategory variations, but most of the literature uses simple categorizations.

In the United States, race and ethnic categories are correlated substantially with SES. Because most categories mix high- and low-SES subgroups, and the mix often differs among geographic areas, there is limited consistency in research findings. The black population, especially in the United States, is an exception because of its protracted history of slavery, segregation and discrimination, and persistent poverty, resulting in consistent findings of high infant and adult mortality, chronic disease incidence, disability, and lesser access to high-quality health services. Despite efforts to diminish health care disparities, studies find poorer outcomes among blacks, even among those who have insurance coverage comparable with more privileged subgroups, attributable less to blatant prejudice and discrimination and more to how housing segregation of blacks funnel them to health facilities with less-sophisticated personnel and resources. Thus, differences in health and health care access are embedded in a larger configuration of poverty and geography (Sampson, 2012). An intriguing paradox relates to psychiatric disorders, which appear to be less prevalent among African Americans than other racial or ethnic groups despite their disadvantaged social position and unfavorable position on physical health indicators. Explanations vary.

A second ethnicity theme that elicits special interest is the relative favorable health status of Mexican Americans relative to Native Americans and other ethnic groups despite their disadvantaged socioeconomic circumstances. Conceptualized as the Hispanic paradox, considerable attention has been focused on alternative hypotheses involving selective migration, health-related behavior, family structure, and culture, among others. The immigrant advantage appears to dissipate in subsequent generations as native-born Mexican Americans acculturate to American and peer culture.

Age and the Life Course

Age, like gender, is important both as a biological and a social determinant of health outcomes over varying points in the life course. Age is strongly associated with the increased prevalence of chronic disease, particularly in later stages of the life course, although determinants often influence outcomes decades before illness manifestations, and even as early as prenatal and early life. Although increasing chronic disease in later life commonly has been viewed as a manifestation of aging, the common paradigm now distinguishes between the consequences of disease and the aging process.

The occurrence of chronic disease and disability in modern nations increasingly has been delayed by improved living standards, healthier lifestyles and behavior, and innovations in public health and medicine. This observation highlights the importance of the experience of varying birth cohorts who have different health outcomes at any given age reflecting the conditions affecting their social development, social and cultural changes, lifestyles and behavior, exposure to social and environmental trauma, and the vicissitudes of life under different historical, environmental, and political conditions. Although the trend in longevity generally has been upward in industrial countries, historical disruptions such as war and political disintegration, increased environmental threats, or changes in health behavior can reverse the pattern, as occurred in Russia after the fall of the Soviet Union. There is now speculation that changing social trajectories involving growing inequalities, environmental assaults, and changing health behavior (poor nutrition, lack of exercise, substance use, increasing obesity leading to diabetes and the like) could overtake medical progress and retard population health improvements. These observations make evident the extraordinary importance of culture, governance capability, and social capital in sustaining an upward course in population health (Kunitz, 2007).

Period effects in the evolution of populations are most difficult to model, including future threats such as war, thermonuclear accidents, climate change, terrorism, economic disruptions and depressions, deathly epidemics, or positive innovations that enhance opportunities for health advancement affecting aging and its quality. Most industrial nations, particularly European nations but also the United States, face aging populations with increasing proportions in the old–old age-groups in which frailty increases and cognition declines. All these nations face the economic challenge of maintaining a health and supportive safety net as the numbers of persons in their 80s and 90s continue to increase and as living arrangements and family structures change, making family caregiving more burdensome and less accessible. How varying nations respond to this growing population through support of alternative institutions depends on economic capacity, ideologies and values, and developing the appropriate institutional structures and workforce to respond meaningfully. It also will depend on the success of new innovations in computer applications, tele-medicine, robotics, nanotechnology, and increased knowledge and technology that will make frailty and dementia more manageable. The prediction that the mortality curve increasingly follows a rectangular pattern where successive cohorts experience a larger proportion of their life span free of disability before death remains to be substantiated but represents an aspiration in ensuring the quality of old age and the future abilities of nations to respond in a meaningful and humane fashion.

The Study of Health Care Organization and Provision

National systems of health care develop through their unique historical trajectories, national cultures, political systems and processes, and economic circumstances. Nevertheless, medical science and medical care, and the systems in which they are embedded, participate in worldwide networks and collaborations, sharing knowledge, technology, and organizational and economic innovations. In addition, common trends among modern nations in demography, urbanization, epidemiology of disease, rising public expectations, and the growing gap between medical possibilities and population demands, on the one hand, and the ability or will to fund them, on the other, pose many common problems for health care systems

(Mechanic and Rochefort, 1996). These problems include means of preventing illness and promoting health; organizing levels of health care and their interrelationships; expanding access to care; linking health and social services into more integrated patterns of care; improving the quality of services; monitoring system performance and health outcomes; regulating professional practice; ensuring the safety and efficacy of drugs and devices; and controlling costs and carrying out required rationing of care wisely and fairly. Most organizational studies of health systems bear on such problems as they relate to the organization of particular health care systems or subsystems. Although challenging, increasing efforts are made to study health systems comparatively across nations. Health knowledge, care provision, and its many complexities are part of a worldwide culture in which much learning goes on among nations, but ideologies and cultural differences often make rigorous comparisons difficult.

The Technological Imperative and Public Expectations

Among the most impressive medical developments in recent decades has been the growth of new technologies, devices and appliances, pharmaceuticals, and the like. Sophisticated advances in imaging, robotics, computer applications, nanotechnologies, prostheses, and more have fundamentally changed the potentialities and practice of medicine and surgery. Coincident with these developments has been the growth of large and powerful industries that have a powerful stake in how medical care is carried out and who spend many billions of dollars seeking to influence policy makers and regulators, physicians, other clinicians, and the general public. These developments are expensive and have contributed greatly to the challenge of increasing costs in all nations.

No nation is prepared to meet all clinician and patient demands for all possible interventions and every nation limits access and utilization in a manner consistent with its economic circumstances, political culture, demanding commercial and professional interests, and social ideologies. Most nations ration care more explicitly than in the United States depending on more centralized policy making on benefits, access, and provider payment, on the purchase and location of specialized facilities and equipment, and on cost–benefit analyses of the value of highly expensive treatments. In the United States, the concept of rationing is politically unacceptable, but rationing occurs in many ways, most obviously by making care and access dependent on insurance coverage, limiting the comprehensiveness of insurance, and the magnitude of copayment required of insured patients (Mechanic, 2006). Although many of the available treatment technologies are valuable in limited instances, the relatively high remuneration for procedural medicine encourages their use far more widely than justified. Industry hype and advertising increase public expectations of the value of technological interventions, and the marketing of new expensive modalities often contribute to misuse. Policy makers intervene cautiously on professional discretion and patient expectations, thus allowing unnecessary treatment and intractable growth in medical care costs. Many experts believe that overtreatment, waste, and fraud accounts for as much as a third to half of all care in the United States (Mechanic, 2006).

The United States is an outlier among nations in health expenditures spending double the proportion of gross domestic product (GDP) for health care than member countries of the Organisation for Economic Co-operation and Development , now approaching 18% of U.S. GDP. This is attributable to the U.S. commitment to a market-based system of care, the predominance of fee-for-service medicine, and much higher prices than elsewhere in the world for physician reimbursement, pharmaceuticals, and almost everything else. In 2010, the United States passed a major but highly controversial medical reform, the Affordable Care Act (ACA), which greatly expands insurance coverage and seeks to develop new ways to pay provider organizations and clinicians through financial incentives to provide high-quality and better coordinated care with less waste. Much attention has been given to elevating the role of primary medical care relative to specialist services, replacing fee-for-service medicine as much as possible with adjusted capitation approaches, including single payments for 'bundled services.' The legislation provides inducements for developing more integrated service provision through entities called accountable care organizations and medical homes by sharing savings derived from increased efficiency. Little about the ACA is certain at this time, such as the success of new organizations and incentives in reducing waste and lowering increases in the costs of medical care. Many of its other initiatives to prevent illness and promote health, to improve behavioral health care and integrate it better with general health care, to elevate primary care relative to specialty care, to build new medical care interdisciplinary teams, to improve chronic disease care, and to promote improved public health remain challenging. The legislation is extraordinarily inspirational, and if it resists repeal, it will require decades for full implementation. Among its provisions is a $10 billion fund to develop and study innovations. It leaves much for medical sociologists and other health services researchers to do.

The Changing Role of the Physician

Sociological work on the organization of care often has focused on the health professions – and particularly changing roles, the continuing struggle against dominance of physicians and their authority, intraorganizational networks and relationships, and the difficulties of integrating levels and sectors of care. There has been much interest in how incentives and other organizational characteristics affect processes and quality of care. Studies focus on a wide range of health organizations, including hospitals, health maintenance organizations (HMOs), large integrated multispecialty organizations, primary care practices, intensive care units, perinatal units, and nursing homes.

Much earlier sociological work focused on the medical profession and how it achieved dominance over medical work and public authority (Freidson, 1970; Starr, 1982). Although physicians remain extremely influential and highly trusted by the public, their dominance has eroded significantly since the 1970s. Sociological studies shifted from studying professional values and authority to more dynamic analyses of how professions and occupations compete to gain control over various spheres of work. With the growth of corporations,

including large insurers, health science centers, pharmaceutical and medical device industries, and a larger role for federal and state government – and as consumers gained power through disease advocacy organizations and access to information and communication channels through the Internet and other means – many strong competing voices have emerged on medical affairs and new coalitions seeking to influence policy. Associations and unions representing different groups in the large health workforce have helped to achieve greater independence and, in many cases, such groups (nurse practitioners, psychologists, or chiropractors) have gained recognition as independent practitioners and opportunities for reimbursement under private and government insurance programs.

The medical profession underwent extraordinary stratification with the proliferation of specialties and subspecialties and a continuing loss of relative status and income among those in primary medical care. Primary care became a disfavored career among graduates of U.S. medical schools and as primary care physicians were increasingly aggregated into larger medical groups, allowing fixed work hours and responsibilities, a greater proportion of women who were now flocking to medical school in proportions equal to men, were recruited disproportionally (Hoff, 2010). With the growth of the hospitalist who now increasingly manages inpatient care, and the economics of practice that made it inefficient for primary care physicians to manage a few hospital patients each day, primary care increasingly has become divorced from inpatient practice, which some see as 'dumbing down' the field. Increasingly, nurse practitioners, urgent care clinics, physician assistants, and behavioral health personnel compete for important components of primary care. Looking ahead the aspiration is to remake primary care within a comprehensive, integrated system of care that coordinates the wide range of services and specialists patients now use, that addresses chronic disease and its many aspects through teams, and that brings together the physical and behavioral care needed by patients with complex comorbidities. Much research now is focused on understanding how these organizations can be best organized and remain economically viable within the concepts of medical homes and accountable care organizations, and the potential for a more effective and meaningful primary care system.

The Intractable Cost Crisis and Managed Care

Managed care has a long history in Europe through general practitioner gatekeepers, remunerated by capitation, who controlled access to hospital and specialty care and other expensive care modalities and who had no financial incentive to overtreat patients. In the United States, however, fee-for-service predominated, and many studies showed how incentives in payment encouraged unnecessary services and cost. Particularly successful were some large multispecialty group organizations organized around prepayment like Kaiser-Permanente (K-P), an integrated health care program serving more than 8 million people. Studies found that such systems provided comparable or better care than private fee-for-service practice in their geographic areas at lower cost. As costs became an increasing concern, and with the failure of the Clinton health care plan in 1994, the private sector organized a variety of prepaid organizations and managed care approaches seeking to control costs. Organizing the K-P type of practice, sources of most of the favorable research, required considerable capital and organizational capacity, and such closed practices were resisted by many patients who did not want to leave ongoing physician relationships or be locked into a closed physician roster.

Insurers developed a variety of alternative models of managed care, most referred to as HMOs or network models that share some common strategies, such as the use of capitation payment and transfer of risk through organizational levels, the use of utilization review to limit the provision of 'unneeded' care, and new systems of constraints on physicians, including practice guidelines, disease management programs, and physician profiling. Most of these new organizations were virtual HMOs in that physicians typically provided care in their usual patient offices and, unlike K-P, served both capitated and fee-for-service patients. Insurers also rapidly developed preferred provider organizations (PPOs) that provided insured patients with a panel of physicians who agreed to provide services at the insurer's reimbursement rate with either no or modest copayment by the patient. When patients used physicians outside the panel, however, they were required to pay a larger proportion of the cost. Researchers found that although they rarely achieved cost and outcome advantages comparable to large integrated prepaid plans, they provided similar services to traditional fee-for-service medicine at reduced cost.

Many physicians and their patients disliked the requirement to get approvals from utilization reviewers and other types of explicit rationing that made it apparent at the point of service that interventions they and their doctors wanted could be denied by the insurer. Physicians also disliked the reduced payment for services and other directives and constraints and organized a strong backlash against many of these management practices. Although initially costs were substantially controlled under strict management, insurers backed down under pressure, litigation, and bad publicity and substantially reduced the strictness of reviews, what analysts came to call 'managed care lite'. The consequence, also, was less success at controlling increased costs.

Management limits were reduced in the case of medical and surgical services, but strict managed care continued in behavioral health care, often administered separately from medical or surgical coverage, by 'carving out' behavioral health and contracting with private managed behavioral health care organizations to manage these services.

In many ways, current efforts to move to new types of reimbursement like bundled payment and payment by episode is a return to the incentives of capitation, seeking to reduce the tendency to generate services to increase income. The aspiration is that by linking this to monitoring of quality, and paying physicians bonuses for good performance, the major disadvantage of capitation to incentivize withholding useful services will be mitigated. This remains to be studied as these new approaches are implemented.

Social Psychology of Health and Health Care

There is a long-standing interest in medical sociology and social psychology in microsocial processes. Such areas as doctor–patient interaction as well as the stress, coping, and social support

processes are perhaps the best integrated bodies of knowledge tied to common theoretical interests and methodological concerns. These interests are shared by other disciplines as well, such as psychosocial epidemiology and health psychology.

Clinician–Patient Interaction

Although the financial and organizational structures of care are crucial, for most patients, what counts most is their personal experience with their physicians, how much they believe they can trust them, the quality of care and caring they believe they are receiving, and their outcomes. There is now a massive literature on doctor–patient relationships, patient–clinician expectations, determinants of trust, quality of communication and instruction, and outcomes of shared decision making (Roter and Hall, 2006). Disparities in access, satisfaction, trust in medical care, and outcomes vary among race and ethnic groups, stimulating interest in understanding the sources of these gaps and whether similarities in background, culture, and attitudes among patients and clinicians mitigate these variations. Aspects of this concern include issues of cultural competency and medical literacy.

From patients' perspectives, paramount concerns are clinician competence and caring, and the extent they perceive clinicians are committed to their interests. Competence typically is assumed and generalized from the ways the clinician interacts, listens, and responds appropriately to questions, provides coherent instructions, and demonstrates interest in the patient's understanding. Microstudies of such interactions monitor such behavior as listening, responding to questions, providing information, interrupting patient communications, distracting behaviors, eye contact, and nonverbal communication.

Stress and the Coping Process

This health paradigm posits that stressful life events cause negative health outcomes when individuals or groups lack the counterharm resources necessary to deal with the stressors, such as ability to deal with the instrumental challenges and to successfully manage the emotions elicited (Mechanic, 1978). Social support buffers the effects of the stressor, often has a direct effect on outcomes, or both. Other important moderating effects include one's sense of personal efficacy embodied in measures of personal control, mastery, or low helplessness. Attributional factors also are important, such as whether the cause of the difficulty is seen as internal to oneself or due to independent events. Over the years, many models have elucidated pathways in the stress process and the relative influence of varying intervening or moderating factors. More recent studies link studies of the stress process with psychophysiological and immunological measures, and brain imaging, demonstrating the complex relations among perception, behavior, and biological processes.

Many important questions persist in the stress-coping literature despite large advances in measurement, data quality, and analysis. Much debate continues on the role of meaning, how it should best be measured, and its importance in response to potential stressors. It is difficult empirically to capture the fact that many stressful life events are not single occurrences but rather are part of a sequence of interconnected stressors with important feedback effects. In the area of social support, researchers are now more aware that people who wish to or attract support are different than those who do not, and that such selection is evident even early in life. Thus, it is important to distinguish between the effects of support from the personal attributes that draw support from others. There is also now greater attention to the fact that supportive relationships carry reciprocal obligations that can induce significant stress.

Social Networks and Social Influence

Social support at the individual level is a source of information, instrumental assistance, and affection and intimacy. An individual's location in a larger network of both strong and weak ties encompasses different opportunities and influences. Although earlier literature focused on primary group ties, which are a strong source of meaning, affection, and assistance, understanding the increased range of weak ties that links with broader perspectives and opportunities, such as the availability of a needed job, captures a much broader field of influence.

With the growth of improved data sources and the power of computing, it has become increasingly possible to map extensive social network and their influence on habits, preventive health behaviors, and a wide range of other social responses. Network influences have been shown to affect smoking patterns, weight gain, and obesity, and information on how to access varying services, including medical care (Christakis and Fowler, 2009). Opportunities for tracking complex networks suggest new ways of using social network theory and methodology to achieve desired health behavior and outcomes.

Neighborhood Influences

As population health perspectives have gained prominence in medical sociology and the importance of place has become more widely appreciated, research has increased on how neighborhoods shape health opportunities and threats. An obvious deterrent to disentangling the characteristics of people selected into neighborhoods from the impact of place itself are the continuing ways in which people select into and out of neighborhoods. Greater availability of longitudinal data and multilevel modeling make this challenge more tractable, and there is growing credible evidence of neighborhood effects that transcend the effects of the characteristics of people and their choices. There is also persuasive evidence of how population characteristics of neighborhoods, and especially race and poverty, and how residents perceive the order of their neighborhoods, are associated with various indicators of health and welfare (Sampson, 2012). Finally, an extraordinary randomized study, Moving to Opportunity (Ludwig et al., 2011), which provided opportunities for poor families to improve their housing by moving to housing in better neighborhoods, demonstrated both the complexity of the person–neighborhood interaction, and illustrated how such change potentially can improve child health status.

Disability and Quality of Life

In the past several decades, medical sociologists have contributed importantly to understanding determinants of disability

and how individual and social definitions of bodily impairments, and social constructions about the agency of persons with these impairments, can lead to widely varying outcomes dependent on social expectations, public attitudes, social norms, and the regulation of the built environment. Sociological work in this area contributed much of the theoretical basis that inspired U.S. legislation affecting people with disabilities and the societal resources made available.

Sociological work made evident that disability that commonly was seen as an incapacity of the person was more properly conceptualized as an interaction between an impairment and the environment. The extent of disability experienced was dependent on the housing, transportation, and other accommodations that either were facilitators or barriers to participation. Changes in attitudes, physical arrangements, and social expectations greatly expanded opportunities for people with impairments to participate in social life from work to athletics. Many studies documented how people with even the most severe physical and emotional impairments could participate in work and other activities to an extent hardly imaginable 50 years ago.

The past 50 years during which medical sociology developed as a major area of inquiry has seen a transformation in the health and aging of the population. On the one hand, people have lived longer and healthier lives and have had opportunities to participate more fully in social life. The health enterprise has expanded enormously, and although sociologists are concerned with increasing medicalization of ordinary problems of living, overdiagnosis, and overtreatment, there is little question that medical developments have contributed meaningfully to people's lives. On the other hand, growing rates of obesity, the large numbers of people with chronic disease, the growing prevalence of Alzheimer's disease and other dementias among the elderly, and increasing rates of autism and other developmental disorders, among many other problems, indicate that enormous scientific terrain remains to be plowed and many opportunities are yet to be explored for promoting health and welfare. As René Dubos sagaciously noted, the idea of health perfection is a mirage, and the environmental and social conditions we create inevitably bring new diseases and bodily assaults. At best, we achieve greater understanding that allows us to ameliorate threat and enhance the opportunity of each person and population group to achieve the highest quality of life in their circumstances.

See also: Class: Social; Disability: Sociological Aspects; Equality of Opportunity; Income Inequality; Organizations, Sociology of; Population Forecasts; Professions, Sociology of; Social Stratification

Bibliography

Bloom, Samuel W, 2002. The Word as Scalpel: A History of Medical Sociology. Oxford University Press, New York.

Christakis, Nicholas A., Fowler, James H., 2009. Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives. Little, Brown and Company, New York.

Crimmins, Eileen M., Preston, Samuel H., Cohen, Barney (Eds.), 2010. International Differences in Mortality at Older Ages: Dimensions and Sources. National Academies Press, Washington, DC.

Freidson, Eliot, 1970. Profession of Medicine: A Study of the Sociology of Applied Knowledge. Dodd, Mead, New York.

Hoff, Timothy, 2010. Practice under Pressure: Primary Care Physicians and Their Medicine in the Twenty-first Century. Rutgers University Press, New Brunswick, NJ.

Kunitz, Stephen J., 2007. The Health of Populations: General Theories and Particular Realities. Oxford University Press, New York.

Ludwig, Jens, Sanbonmatsu, Lisa, Gennetian, Lisa, et al., 2011. Neighborhoods, obesity and diabetes – a randomized social experiment. New England Journal of Medicine 365, 1509–1519.

Mechanic, David, 1978. Medical Sociology, second ed. Free Press, New York.

Mechanic, David, 2006. The Truth about Health Care: Why Reform Is Not Working in America. Rutgers University Press, New Brunswick.

Mechanic, David, 2007. Population health: challenges for science and society. Milbank Quarterly 85, 533–559.

Mechanic, David, Rochefort, David A., 1996. Comparative medical systems. Annual Review of Sociology 22, 239–270.

Phelan, Jo C., Link, Bruce G., Tehranifar, Parisa, 2010. Social conditions as fundamental causes of health inequalities: theory, evidence and policy implications. Journal of Health and Social Behavior 51 (Suppl. 1), S28–S40.

Roter, Debra L., Hall, Judith A., 2006. Doctors Talking with Patients/Patients Talking with Doctors: Improving Communication in Medical Visits, second ed. Praeger, Westport, Conn.

Sampson, Robert S., 2012. Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago Press, Chicago.

Starr, Paul, 1982. The Social Transformation of American Medicine. Basic Books, New York.

Medicalization: Sociological and Anthropological Perspectives

Peter Conrad and Meredith Bergey, Brandeis University, Waltham, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by M. Lock, volume 14, pp. 9534–9539, © 2001, Elsevier Ltd.

Abstract

Medicalization is the process by which nonmedical problems become defined and treated as medical problems often requiring medical treatment. The term medicalization first appeared in the sociology literature and focused on deviance, but it soon expanded to examine other human conditions. This article focuses on sociological and anthropological perspectives on the expansion of medicalization. This includes a review of the characteristics, origins, and consequences of medicalization, with sociology's emphasis on the emergence of medical categories and anthropology's emphasis on biomedicine's institutional power and cultural authority. Medicalization has numerous social consequences, including the pathologization of human differences and individualization of human problems while minimizing social and political context.

Sociological Roots

Medicalization means 'to make medical' or more specifically the process by which previously nonmedical problems become defined and treated as medical problems, usually as diseases or disorders. Social scientists and critics including Foucault (1965), Szasz (1970), and Illich (1976) criticized medicalization for its potential to serve as a mechanism of social control, although they did not use the term. Jesse Pitts (1968), in an earlier edition of a *Social Science Encyclopedia*, wrote specifically and largely positively about the medicalization of deviance, relating it to changing forms of social control. It was Zola's (1972) paper that became the touchstone for medicalization in sociology, where he too focused on the expansion of medical social control. Shortly thereafter, Conrad (1975, 1976) and Conrad and Schneider (1992) wrote specifically about various aspects of the medicalization of deviance. Research on medicalization expanded from a focus on deviance to a broader lens of life conditions and ailments. Since those early works literally thousands of papers have been written on medicalization (e.g., see Clarke et al., 2011; Conrad, 2007) by sociologists, anthropologists, historians, feminists, and medical scholars. Medicalization has become a common term in the social science and medical literature and to a degree in the public media.

The focus of medicalization is on the process by which human problems or conditions become defined and treated as medical problems. While some critics have used the term medicalization synonymously with 'overmedicalization' (too many diagnosed or medically treated), there is nothing inherent in the sociological analysis of medicalization that necessarily suggests an evaluation of medicalization. Rather, medicalization is a phenomenon that can be examined the way other 'zations,' such as industrialization, urbanization, or secularization can be analyzed – as a process to be understood in its own right (Conrad, 2013). The most common sociological perspective for examining medicalization is as a form of social constructionism (Conrad and Barker, 2010), although some analysts have suggested a broader perspective under the term biomedicalization (Clarke et al., 2011).

Characteristics of Medicalization

There are several characteristics of medicalization. (1) The definitional issue is central to medicalization; that is, how a problem is defined is key to what is to be done about it. (2) There are degrees of medicalization. For example, some problems are almost completely medicalized (e.g., schizophrenia, epilepsy), while others are only somewhat medicalized (e.g., Internet addiction, sexual addiction), with still others contested or somewhere in between (e.g., obesity, opiate addiction). Thus medicalization is more of a continuum than a binary either/or distinction. (3) Medicalized categories can expand or contract. For example, attention deficit hyperactivity disorder (ADHD) was for many years seen as a disorder of children but it is now seen as a lifespan disorder, including adult ADHD. Posttraumatic stress disorder emerged in the 1970s as a diagnostic category depicting Vietnam veterans and has expanded to include survivors of violence, sexual abuse, or natural disasters. Now it can even be applied to individuals who have witnessed violence or disaster. On the other hand, hysteria, which was a common medical category in the late nineteenth century, has almost disappeared as a diagnosis. (4) Physician involvement in medicalization is variable. With the medicalization of alcoholism, physicians were minimally involved; rather, the emergence of Alcoholics Anonymous was central to this process. With many other examples of medicalization, direct medical involvement is a key. (5) Medicalization is bidirectional; that is, there can be medicalization as well as demedicalization. The most common example of demedicalization has been homosexuality, which due to the 1974 vote of the American Psychiatric Association and subsequent changes became 'officially' demedicalized (Conrad, 2007). There is no question, however, that there has been much more medicalization than demedicalization in the past four decades. One recent study (Conrad et al., 2010) estimates

that medicalized disorders cost about $77 billion or about 3.9% of our health care expenditures, more than is spent on all of US public health.

Caveats about Medicalization

A few caveats about medicalization studies should be noted: (1) Some researchers focus on whether medicalized problems are 'real' diseases or whether they are socially invented medical categories. It is important for social scientists to remain agnostic on what is or is not a 'real' disease; a better question to ask is how did this problem become defined as a medical disorder and with what consequences? (2) Much of the popular and some of the academic writing on medicalization has been concerned with the 'overmedicalization' of society. While considerable research on medicalization does suggest potential overmedicalization is an issue to be considered, medicalization studies focus on the 'process' of medicalization and are not necessarily driven by concerns of overmedicalization. Sociologists study medicalization just like they study other social processes, focusing on the origins and consequences of the phenomena. It is probably likely in some cases that there has been an overmedicalization of some medical problems, while in others undermedicalization of a certain problem may be the case. (3) Occasionally, critics will use the term medicalization as an epithet or pejorative, with the assumption that medicalization is by its nature a 'bad' thing. But as we have emphasized, medicalization is not a normative process, in itself neither bad nor good. Historically, for instance, there are many examples of beneficial medicalization, such as epilepsy and, in important ways, childbirth. It is hard to imagine that most people would prefer to return to the era when epilepsy was seen as possession or a 'curse,' a sign of being bewitched and a justification for institutionalization; nor would we want to revisit a childbirth era in which maternal mortality was common and infant mortality was a regular birth outcome. That is not to say that feminists and others have not provided important analyses of some of the consequences of what might be called the unreflective and extreme medicalization of modern childbirth, such as the very high rate of caesarian sections in the United States and elsewhere (Morris, 2013).

Anthropological Perspectives

While sociologists may have pioneered studies in medicalization, other disciplines (e.g., history, bioethics, medicine) have made important contributions to the medicalization literature. Most notably, the number of studies and statements regarding medicalization has grown in the field of anthropology. This is especially the case in the subfield of medical anthropology – where considerable attention has been given to the ways in which disease and disease-like states are created (Cohen, 1998; Lock, 2001), and the effects of such processes on people's lived experience.

Together with sociologists, anthropologists in the 1970s and 1980s often described medicine as an institution of social control, with health professionals acting as the agents of such control (Lock and Nguyen, 2010: p. 70). Numerous anthropological studies depicted physicians as aggressive medicalizers (Lock, 2001) and great attention has been paid to "the power of clinicians who wield biomedical discourse and technologies in constituting illness, in 'making medical'" (Crowley-Matoka and True, 2012). A turn in recent decades toward critical medical anthropology emphasizes power relations that affect biomedicine, often critiquing biomedicine's dominance over alternative health systems, the power structures existing within biomedicine, and marginalized populations' lack of fair or equal access to biomedical options (Baer, 2004; Singer, 1989, 1990).

Anthropologists continue to explore how the modern medical profession amasses and uses its power and authority (Behague, 2009). Anthropological studies have provided important insights into how biomedical perspectives and treatments are applied to numerous aspects of human life, often in non-Western settings. They have also emphasized individual's responses, experiences, and meaning making with respect to processes of medicalization. Such emphases differ at times from a stronger tendency within sociology to explore the emergence and development of medical categories (e.g., diagnoses) as well as their application. In contrast to anthropological literature, much of the sociological research on medicalization also continues to focus on Western populations and groups.

Ethnographic research, the cornerstone of many anthropological as well as sociological studies, provides rich insights into the diverse responses – individual, family, and community – to medicalization. Ethnographic and other qualitative research approaches are well suited for capturing thick descriptions of people's perspectives and lived experiences (Geertz, 1973). Lock (2001: p. 482) describes how the complexity of responses to medicalization necessitates the use of an "approach that is sensitive to the situated exigencies of the everyday worlds of informants and their relationships within families, communities, and local political groupings." The following sections provide examples of some of anthropology's contributions to the medicalization literature.

Medicalization and Mental Health

Anthropological studies contribute to a growing body of scholarly literature on medicalization and mental health. Sociologists and anthropologists have made important contributions to understanding the medicalization of mental disorder (e.g., Behague, 2009; Conrad and Slodden, 2013; Horwitz, 2002; Horwitz and Wakefield, 2007; Lane, 2007) revealing medicine's widening jurisdiction over mental facets of life (e.g., behaviors related to attention, sadness, shyness). For example, reflecting upon the American Psychiatric Association's designation of grief as a diagnostic category in the diagnostic and statistical manual–fifth edition, Kleinman (2012) cautions against turning ordinary bereavement into an appropriate target of medical intervention. He describes the medical profession's ability to reframe experience in ways that can influence culture change, for example, around death.

Alongside sociologists and historians, anthropologists emphasize how medical models of causation ignore sociopolitical context and individualize aspects of mental health

(Behague, 2009; Conrad and Slodden, 2013; Horwitz and Wakefield, 2007). Scheper-Hughes (1992), for example, showing how hunger and related distress, or *nervos*, in Northeast Brazil, were medicalized and treated through medical interventions such as pharmaceuticals rather than through social and political reform.

Debates have also emerged within anthropology and sociology on medicalization and psychiatry about how to move beyond a polarity between scientific objectivism and social constructionism (Behague, 2009). Some scholars point out that research grounded within social constructionist frameworks, while providing important insights into subjective experience, dislocate such subjectivity from broader contextual factors (Behague, 2009; Butt, 2002). Such views reflect more general issues in the field about a need to explore both micro- and macrolevel contexts, individual experiences and social relations, local contexts as well as broader political and economic forces (e.g., Scheper-Hughes and Lock, 1986; Singer, 1989; Taussig, 1980).

Medicalization and Women's Health

The anthropological literature is also replete with studies of the medicalization of women's bodies and health (e.g., fertility, pregnancy, childbirth, menopause) (see Behague, 2002; Davis-Floyd and Sargent, 1997; Ivry, 2010). Hunter de Bessa (2006: pp. 221–222) notes, for example, that "aspects of reproduction once thought to be primarily political, personal, or social in nature have come to be seen as medical problems – problems that are thought to be best solved by physicians through medical practice and management." Anthropologists have explored how (e.g., via cesarean section or natural birth), where (e.g., in a hospital or at home), and when (e.g., prolonged pregnancy and induced labor) women give birth, as well as the ideological and sociocultural frames which may influence such decisions (e.g., Behague, 2002; Davis-Floyd, 2003; Donner, 2003; Westfall and Benoit, 2004). They have studied women's fertility (e.g., Becker, 2000; Ivry, 2010), including research on domestic and transnational assisted reproductive technologies and their markets (e.g., Franklin, 2011; Inhorn and Gurtin, 2011; Inhorn and Patrizio, 2012; Morgan and Michaels, 1999). They have also described how resistance to medicalization contributed to the emergence of the home birth movement and the use of alternative therapies (Davis-Floyd, 1993, 2003; Lock and Nguyen, 2010).

An important line of anthropological research connects feminist critiques of the biomedical profession's power, particularly over women's health. Medicine has been described as a deeply patriarchal institution that views the female body as a vessel upon which to intervene with various – often reproductive – technologies (Davis-Floyd, 2003; Hunter de Bessa, 2006; Lock and Nguyen, 2010). Feminist anthropologists have drawn attention to the potential negative consequences that medical technological advancements and 'an ethos of intervention' can have on women's health and human rights (Hunter de Bessa, 2006). Some scholars have also cautioned against depicting women in passive roles – arguing that such depictions perpetuate the very kind of assumption that feminists have been trying to combat (Lock and Nguyen, 2010).

Anthropological studies describe women's diverse responses to medicalization (Hunter de Bessa, 2006) including showing that women are not always victims or passive recipients of medicalization (Lock, 2001; Hunter de Bessa, 2006). Martin's (1987) study of women in Baltimore, Maryland, revealed that they may resist and form meanings that vary from those which are dominant in the medical profession. Women may also embrace medicalization and medical technology as enabling, particularly with respect to infertility (Becker, 2000; Kielmann, 1998) or use technology to seek some degree of advancement or independence for themselves (Lock, 2001; Nahman, 2011; Pande, 2010). Women's behaviors may stem from a desire to satisfy, or respond to pressure from others (e.g., family members, women's groups, their community) (Lock, 2001). Individual responses to biomedical technologies are often pragmatic; they are based upon what individuals perceive to be in their own, their families' or their communities' best interests (Donner, 2003; Lock and Kaufert, 1998; Lock, 2001). Anthropological studies also emphasize the role that ideology, irrespective of medical necessity, may play in shaping reactions to medicalization (Behague, 2002; Donner, 2003) and the intricate relationships between ideology and forms of social inequality which may shape responses (Behague, 2002). Such insights shed light on the complexity of responses to medicalization and complicate our notions of resistance, agency, and autonomy (Lock, 2001).

Social Consequences of Medicalization

Both anthropologists and sociologists have described various social consequences of medicalization, including how, by extending the sick role over a broader range of human conditions, medicalization can excuse an individual from social roles and expectations (e.g., related to school, work, caregiving) (Lock, 2001; Scheper-Hughes and Lock, 1986). As Lock (2001) indicates, at times, medicalization can function to 'exculpate individuals from responsibility for being sick.' Along with historians and sociologists, anthropologists have also shown how, with medicalization, attention may be deflected away from the sociopolitical and economic context that influences matters related to health, prevention, and treatment (Lock, 2001; Marshall, 2005).

Anthropological researchers suggest that subjectivity and the reporting of symptoms are 'subsumed into medical pathologies and standard deviations from medical norms,' while attention and responsibility are placed on individuals and their bodies (481). Sociologists, focusing on the emergence and treatment of diagnoses such as ADHD, mild depression, obesity, and menopause (see Conrad, 2007; Horwitz and Wakefield, 2007), have expressed concerns about the pathologization of everything, that is, turning all human differences into diagnoses and treatable disorders.

Much anthropological research on medicalization explores the perspective and lived experience of individuals whose health may be affected by processes of medicalization. For example, Wentzell's (2013) ethnographic research

revealed how older Mexican men resisted a medical model of aging and its associated antiaging interventions by rejecting the use of erectile dysfunction drugs. Men framed decreased sexual activity as a means of fulfilling socially valued gender norms.

Anthropological studies of physicians often focus upon their hegemony over knowledge without exploring their subjective experiences. Several recent studies mark exceptions to this trend. For example, Crowley-Matoka and True's (2012) ethnographic study of clinician's perspectives suggests that incomplete and ineffective medicalization of pain and its management may lend to disempowerment – for both patients and clinicians. Ivry (2010) describes how in negotiations between medical professionals and rabbinic authorities over matters of fertility assistance, religious law and authority could be mystified by both parties in such a way that led many physicians to feel helpless. Such works call into question well-established notions about power and control in the medical profession.

Additional Uses of Medicalization

In a few cases, the term medicalization itself has been used to mean more medical involvement, attention, or concern as in a particular area. While these works use the term medicalization, as in 'The Medicalization of Anthropology: A Critical Perspective on the Critical-Clinical Debate' (Morgan, 1990) or The Medicalization of Cyberspace' (Miah and Rich, 2008) they use the term to mean the increasing medical concerns among anthropologists or the expanding medical attention on the Internet. Speaking more specifically of the subfield of medical anthropology, Browner (1999) describes a 'medicalization of medical anthropology,' by which she means increased attention within the field to the study of 'entities as they are conceptualized by biomedicine.' Browner describes how medical anthropologists are increasingly taking jobs in the United States, not in basic research, but in work that is based in facilities that provide medical care. Such trends need not necessarily be positive or negative, she notes, but there are potential substantive and methodological consequences, which may result from a field that is increasingly medicalized. Browner (1999) cautions against uncritical acceptance of biomedical explanations and categories for illness, and points to potential challenges associated with aligning hallmark anthropological research methods (e.g., ethnography) with 'a circumscribed biomedical condition.' These uses of the term medicalization, however, are different than the analysis of the development of medical categories or the use of medicalized conceptions presented in this article.

The term medicalization has in the past four decades become an important concept in the social scientific studies of medicine. An increasing number of researchers in the fields of sociology and anthropology, as well as history, bioethics, and medicine are both using the term and conducting research in the area. Given the continuing expansion of medicine in the twenty-first century and the exportation of Western medical categories (i.e., diagnoses, treatments) as part of the globalization of biomedicine, one can predict that medicalization will continue to expand and be of major interest to social scientists.

See also: Body, History of; Foucault, Michel (1926–84); Gender History; Illich, Ivan (1926–2002); Inequality, Social; Medicine, History of Western; Plagues and Diseases in History; Professions and Professionalization, History of; Social Constructivism.

Bibliography

Baer, H., 2004. Toward an Integrative Medicine: Merging Alternative Therapies with Biomedicine. AltaMira Press, Oxford.

Becker, G., 2000. The Elusive Embryo: How Women and Men Approach New Reproductive Technologies. University of California Press, Berkeley.

Behague, D., 2002. Beyond the simple economics of cesarean section birthing: women's resistance to social inequality. Culture, Medicine, and Psychiatry 26, 473–507.

Behague, D., 2009. Psychiatry and politics in Pelotas, Brazil: the equivocal quality of conduct disorder and related diagnoses. Medical Anthropology Quarterly 23 (4), 455–482.

Browner, C.H., 1999. On the medicalization of medical anthropology. Medical Anthropology Quarterly 13 (2), 135–140.

Butt, L., 2002. The suffering stranger: medical anthropology and international mortality. Medical Anthropology 21 (1), 1–24.

Clarke, A.E., Shim, J.K., Mamo, L., Fosket, J.R., Fishman, J.R., 2011. Biomedicalization: a theoretical and substantive introduction. In: Clarke, A.E., Mamo, L., Fosket, J.R., Fishman, J.R., Shim, J.K. (Eds.), Biomedicalization: Technoscience, Health, and Illness in the U.S. Duke University Press, Durham, pp. 1–47.

Cohen, L., 1998. No Aging in India: Alzheimer's, the Bad Family, and Other Modern Things. University of California Press, Berkeley.

Conrad, P., 1975. The discovery of hyperkinesis: notes on the medicalization of deviant behavior. Social Problems 23, 12–21.

Conrad, P., 1976. Identifying Hyperactive Children: The Medicalization of Deviant Behavior. D.C. Heath, Lexington, MA.

Conrad, P., 2007. The Medicalization of Society: On the Transformation of Human Conditions into Treatable Disorders. Johns Hopkins University Press, Baltimore.

Conrad, P., 2013. Medicalization: changing contours, characteristics, and contexts. In: Cockerham, W.C. (Ed.), Health Sociology on the Move. Springer, New York.

Conrad, P., Barker, K., 2010. The social construction of illness: key insights and policy implications. Journal of Health and Social Behavior 51 (S), 567–579.

Conrad, P., Mackie, T., Mehrotra, A., 2010. Estimating the costs of medicalization. Social Science and Medicine 70, 1943–1947.

Conrad, P., Schneider, J.W., 1992. Deviance and Medicalization: From Badness to Sickness (expanded ed.). Temple University Press, Philadelphia.

Conrad, P., Slodden, C., 2013. The medicalization of mental disorder. In: Aneshensel, C., Phelan, J. (Eds.), Handbook of the Sociology of Mental Health, second ed. Springer, New York.

Crowley-Matoka, M., True, G., 2012. No one wants to be the candy man: ambivalent medicalization and clinician subjectivity in pain management. Cultural Anthropology 27 (4), 689–712.

Davis-Floyd, R., 1993. Hospital birth as a technocratic rite of passage. Mothering 67, 68.

Davis-Floyd, R.E., 2003. Birth as an American Rite of Passage, second ed. University of California Press, Berkeley.

Davis-Floyd, R.E., Sargent, C.F. (Eds.), 1997. Childbirth and Authoritative Knowledge. Cross-cultural Perspectives. University of California Press, Berkeley.

Donner, H., 2003. The place of birth: childbearing and kinship in Calcutta middle-class families. Medical Anthropology 22, 303–341.

Foucault, M., 1965. Madness and Civilization. Bantam Books, New York.

Franklin, S., 2011. Not a flat world: the future of cross-border reproductive care. Reproductive BioMedicine. Online 23, 814–816.

Geertz, C., 1973. Interpretation of Culture. Basic Books, New York.

Horwitz, A.V., 2002. Creating Mental Illness. The University of Chicago Press, Chicago.

Horwitz, A.V., Wakefield, J., 2007. The Loss of Sadness: How Psychiatry Transformed Normal Sorrow into Depressive Disorder. Oxford University Press, New York.

Hunter de Bessa, G., 2006. Medicalization, reproductive agency, and the desire for surgical sterilization among low-income women in urban Brazil. Medical Anthropology 25, 221–263.

Illich, I., 1976. Medical Nemesis. Random House, New York.
Inhorn, M.C., Gurtin, Z., 2011. Cross-border reproductive care: a future research agenda. Reproductive Biomedicine Online 23, 665–676.
Inhorn, M.C., Patrizio, P., 2012. The global landscape of cross-border reproductive care: twenty key findings for the new millennium. Current Opinion in Obstetrics and Gynecology 24 (3), 158–163.
Ivry, T., 2010. Kosher medicine and medicalized halacha: an exploration of triadic relations among Israeli rabbis, doctors, and infertility patients. American Ethnologist 37 (4), 662–680.
Kielmann, K., 1998. Barren ground: contesting identities of infertile women in Pemba, Tanzia. In: Lock, M., Kaufert, P.A. (Eds.), Pragmatic Women and Body Politics. Cambridge University Press, Cambridge, pp. 127–163.
Kleinman, A., 2012. Culture, bereavement, and psychiatry. Lancet 379 (9816), 608–609.
Lane, C., 2007. Shyness: How Normal Behavior Became a Sickness. Yale University Press, New Haven.
Lock, M., 2001. Medicalization: cultural concerns. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Kluwer Academic Publishers, Dordrecht.
Lock, M., Kaufert, P., 1998. Pragmatic Women and Body Politics. Cambridge University Press, Cambridge.
Lock, M., Nguyen, V., 2010. An Anthropology of Biomedicine. Wiley-Blackwell, West Sussex.
Marshall, W.E., 2005. AIDS, race and the limits of science. Social Science and Medicine 60, 2515–2525.
Martin, E., 1987. The Woman in the Body: A Cultural Analysis of Reproduction. Beacon Press, Boston.
Miah, A., Rich, E., 2008. The Medicalization of Cyberspace. Routledge, London.
Morgan, L.M., 1990. The medicalization of anthropology: a critical perspective on the critical-clinical debate. Social Science and Medicine 30 (9), 945–950.
Morgan, L.M., Michaels, M.W., 1999. Introduction: the fetal imperative. In: Morgan, L.M., Michaels, M.W. (Eds.), Fetal Subjects, Feminist Positions. University of Pennsylvania Press, Philadelphia.
Morris, T., 2013. Cut It Out: The C-section Epidemic in America. NYU Press, New York.
Nahman, M., 2011. Reverse traffic: interesting inequalities in human egg donation. Reproductive BioMedicine Online 23, 626–633.
Pande, A., 2010. Commercial surrogacy in India: manufacturing a perfect mother-worker. Signs 35 (4), 969–992.
Pitts, J.R., 1968. Social control: the concept. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences, vol. 14. Macmillan and The Free Press, New York, pp. 381–396.
Scheper-Hughes, N., 1992. Death without Weeping: The Violence of Everyday Life in Brazil. University of California Press, Berkeley.
Scheper-Hughes, N., Lock, M., 1986. Speaking "truth" to illness: metaphors, reification, and a pedagogy for patients. Medical Anthropology Quarterly 17 (5), 137–140.
Singer, M., 1989. The coming of age of critical medical anthropology. Social Science and Medicine 28 (11), 1193–1203.
Singer, M., 1990. Reinventing medical anthropology: toward a critical realignment. Social Science and Medicine 30 (2), 179–187.
Szasz, T.S., 1970. The Manufacture of Madness: A Comparative Study of the Inquisition and the Mental Health Movement. Harper and Row, New York.
Taussig, M.T., 1980. Reification and the consciousness of the patient. Social Science and Medicine 14B, 3–13.
Wentzell, E., 2013. Aging respectably by rejecting medicalization: Mexican men's reasons for not using erectile dysfunction drugs. Medical Anthropology Quarterly 27 (1), 3–22.
Westfall, R.E., Benoit, C., 2004. The rhetoric of "natural" in natural childbirth: childbearing women's perspectives on prolonged pregnancy and induction of labor. Social Science and Medicine 59, 1397–1408.
Zola, I.K., 1972. Medicine as an institution of social control. The Sociological Review 20, 487–504.

Medicine and Health for Sexual Minorities

Judith A Houck, University of Wisconsin–Madison, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Over the course of the last 40 years, lesbian health has been an evolving concept. Emerging from the alleged neglect of lesbians in the women's health movement, lesbian health has become an institutional specialty, a research subject, and a social movement. Although lesbian health has gained a great deal of attention, just who counts as a lesbian remains in flux.

In 2010, the US Department of Health and Human Services (DHHS) articulated its national health goals for the next 10 years in its Healthy People 2020 program. One of its new initiatives was the goal of improving the "health, safety, well-being of lesbian, gay, bisexual, and transgender (LGBT) individuals." Justifying this focus, the government cited research demonstrating that "LGBT individuals face health disparities linked to societal stigma, discrimination, and denial of their civil and human rights." It maintained that discrimination led to "high rates of psychiatric disorders, substance abuse, and suicide." Efforts to improve LGBT health focused both on improving the clinical climate and the larger social and cultural environment for LGBT individuals. (HealthyPeople.gov. www.healthypeople.gov/2020 (accessed 25.06.13.).)

The inclusion of LGBT health needs in Healthy People 2020 is the culmination of 40 years of activism to better identify and meet the changing health needs of gay, lesbian, and transgendered people. Its inclusion, however, presumes a questionable intellectual coherence to LGBT health needs, and obscures the distinct paths the members of the LGBT communities have forged as they struggled to gain attention to their particular health needs.

This article focuses on the history of 'lesbian health' in its many guises: as an area of unmet need, an institutional specialty, a research subject, and a social movement. It shows how 'lesbian health' emerged out of two overlapping social movements, the women's health movement and the gay liberation movement. It also demonstrates that 'lesbian health' has always been a category with fuzzy boundaries, shifting content, and complex politics. This history also highlights several concerns still at play under the rubric of lesbian health. Who is the target for lesbian health interventions? Is lesbian health best understood as a subset of women's health, LGBT health, or as a separate entity? How has the changing social climate influenced lesbian health needs? Are lesbian health issues largely social and economic, the consequences of discrimination, or are they primarily embodied? The roots of many of these questions emerged with the concept in the early 1970s and remain unresolved in the twenty-first century. The inclusion of lesbian health as a Healthy People 2020 goal demonstrates its emergence and tenacity as a pressing issue, but its history highlights its shifting scope.

Radical Health Movements: The Context for Lesbian Health

The long 1960s and 1970s are well known for a variety of social movements that demanded a profound reorganization of the social and cultural fabric of the United States. These included the civil rights movements, the student movement, and the antiwar movement. The women's rights movement and the women's liberation movement emerged out of this widespread social ferment.

In broad strokes, the women's rights movement sought to equalize the educational, professional, and political opportunities for women. It tended to focus on change within the system, promoting for example, the Equal Rights Amendment, Title IX, and equal pay provisions. On the other hand, the women's liberation movement, identifying patriarchy and sexism as the primary oppressors of women, called for a more profound reconceptualization and reorganization of society. Believing that the Personal was Political, the women's liberation movement promoted consciousness raising, zap actions, and female-centered alternatives to male-dominated institutions. The issue of sexual identity in these branches of the women's movements was sometimes fraught. Although lesbians actively participated in both the rights and liberation movements (indeed, many lesbians were created through these feminist movements), lesbians frequently felt their concerns were overlooked in the agenda of straight (often white) women. (Echols, Alice, 1989. Daring to Be Bad: Radical Feminism in America, 1967–1975. University of Minnesota Press, Minneapolis; Enke, Anne, 2007. Finding the Movement: Sexuality, Contested Space and Feminist Activism. Duke University Press, Durham; Radicalesbians, 1972/1992. Woman-identified woman. In: Jay, Karla, Young, Allen (Eds.), Out of the Closets: Voices of Gay Liberation, second ed. NYU Press, New York, pp. 172–181.)

The gay liberation movement similarly emerged from these civil rights, antiwar, and student movements. In the late 1960s, the older homophile movement was supplanted by a distinctly different movement that rejected the politics of respectability, instead demanding recognition of sexual and gendered behaviors outside the norms of the middle-class, white mainstream. This gay liberation movement did not dismiss the homophile movement's focus on the rights of gay and lesbian people, but it created an in-your-face demand for cultural

recognition. Gay liberation celebrated sexual freedom, encouraged 'gay pride,' and demanded respect for gender nonconformity. In gay pride parades, openly gay community establishments, and artistic creations, gay men and lesbians fought for their rights by living out loud.

The boundaries between the women's liberation movement and this gay liberation movement were not clear, particularly when it came to lesbians. Some lesbian activists believed that their oppression was grounded in sexism, heterosexism, and homophobia, so they simultaneously worked through both movements. Similarly, lesbian feminists, working both inside and outside of feminism, shaped the agenda of the women's liberation movement in significant ways.

The women's movement and the gay and lesbian movement both had factions that focused on health. From safe and legal abortions to convenient and effective methods of birth control, from natural childbirth to sexual self-determination, feminists insisted that the ability to make decisions about their bodies provided the essential foundation for the larger liberation of women. By the late 1960s, this attention to women's bodies had coalesced into a women's health movement. (For general histories of the women's health movement, see Ruzek, Sheryl Burt, 1968. The Women's Health Movement: Feminist Alternatives to Medical Control. Praeger Publishers, New York; Kline, Wendy, 2010. Bodies of Knowledge: Sexuality, Reproduction, and Women's Health in the Second Wave. University of Chicago Press, Chicago; Morgan, Sandra, 2002. Into Our Own Hands: The Women's Health Movement in the United States, 1969–1990. Rutgers University Press, Brunswick.)

In general, health feminists identified sexism and patriarchal dominance as the root of women's oppression. (These activists were not blind, however, to the role of racism, capitalism, and imperialism in securing and maintaining women's oppression.) To combat the sexism and oppression that harmed women's health, health feminists were generally guided by three principles. First, health feminist insisted that women rather than men – whether in their roles as physicians, legislators, or spouses – should be in charge of women's bodies. Women should decide for themselves how to control their fertility, express their sexuality, and cope with their illnesses. Second, the movement insisted that women had the right to participate actively in the doctor–patient relationship, rejecting the patronizing and paternal cultural and professional assumption that 'the (male) doctor knows best.' Health feminists demanded access to information about their bodies, healthy and ill, and to participate fully in their health care decisions. Finally, health activists argued that physicians had taken over many bodily events (menstruation, childbirth, and fertility control) that should not be considered 'medical' in nature. Women active in the movement sought to return these women's health matters to women's hands.

The self-help gynecology movement, one strain of the larger women's health movement, exemplified these principles. The movement, founded in 1969 when Carol Downer demonstrated how women could see their own cervices using a mirror, a speculum, and a flashlight, promoted women's health. They toured nationally showing women how to see their interiors; they established self-help clinics where women gathered regularly to learn about their own and each others bodies; they published books and pamphlets based upon their own research; they established reproductive health clinics staffed largely with lay women; they campaigned in multiple venues (legislative, professional, and medical) to return control of women's bodies to women.

The gay and lesbian health movement emerged in the 1970s to combat "[the] hostility of mainstream health services and practitioners toward the issue of homosexuality in general and to homosexual men and women, in particular." Perhaps inspired by the successful 1973 campaign to remove homosexuality from the Diagnostic and Statistical Manual of Mental Disorders, gay and lesbians health care providers created professional organizations to increase gay visibility and to combat homophobia. The Gay Nurses' Alliance (1973) paved the way for other health professionals, including public health workers and pharmacists. In 1976, activists created the National Gay Health Coalition (later the National Coalition for LGBT Health) as an umbrella group, which became the "nexus of gay health activism." (Deyton, Bopper, Lear, Walter, 1988. A brief history of the gay/lesbian health movement in the U.S.A. In: Shernoff, Michael, Scott, William A. (Eds.), The Source Book on Lesbian/Gay Health Care, second ed. National Lesbian/Gay Health Foundation, Washington, D.C., pp. 15, 18.)

Activists, working at both the local and national levels, also attempted to meet the health needs of gay men, and to a lesser degree, lesbians, by establishing health clinics. As the human immunodeficiency virus infection/acquired immunodeficiency syndrome (HIV/AIDS) crisis began in 1981, much of the movement's attention focused on the epidemic and the critical health needs of gay men. Of the 11 gay and lesbian-focused clinics extant in 1992, 3 were founded in response to the AIDS epidemic; 2 others were founded as sexually transmitted disease (STD) clinics for gay men. (Untitled grant proposal, National Alliance of Lesbian and Gay Health Clinics, n.d. (1992).) Lesbians played important roles in the HIV/AIDS crisis, but it was often, although not exclusively, understood as work by lesbians for their gay male compatriots.

'Lesbian health' was formed within and emerged out of these larger social movements. Lesbians' health needs overlapped with women's health needs, and lesbians surely benefited from and participated in the efforts to combat sexism in medicine. Lesbians also contributed to efforts to combat homophobia in the health professions and to provide health care to gay men and lesbians. Neither women's health nor gay and lesbian health movements, however, put the health needs of lesbians first; lingering heterosexism and sexism helped foment a more targeted focus on lesbian health both within these movements and independent of them.

Creating Lesbian Health

Women's health issues brought to feminism many who had never before identified with the women's movement. Many of these new recruits were lesbians. In her history of the lesbian health movement, Risa Denenberg claims that many of the leaders of the women's health movement were lesbians. The proportion of lesbians in the movement seems to have varied significantly geographically. (Denenberg, Risa, 1997. A history of the lesbian health movement. In: White, Jocelyn, Martinez,

Marissa C. (Eds.), The Lesbian Health Book: Caring For Ourselves. Seal Press, Seattle.) Other women came to identify as lesbians after they became involved in the women's movement. Although many lesbians were involved in the women's health movement, lesbians frequently decried the alleged heterosexist, reproductive focus of the movement, and wondered what was in it for them. A lesbian from Iowa City explained the problem in 1975. In 'Is There Life After Abortions,' Freddy Brass lamented "do we have any energy left for ourselves after spending the day as a patient advocate for women getting abortions (or birth control or babies)? Do we even know what there is to be done for ourselves?" She understood the importance of securing reproductive services for all women, but she wanted to turn her attention to "… lesbian health care, lesbian health research, and how to acquire the technology to care for ourselves and do that research." This desire to focus on lesbian health encouraged Brass to leave the women's health movement. (Is There Life After Abortions? Lesbian Connection 1, 12 (November 1975). On concerns regarding the heterosexist nature of the self-help movement, see also Second Women-Controlled Women's Health Center Conference. Monthly Extract 3, 5 (December 1974/January 1975). For a similar critique, see I Am a Lesbian Health Care Worker, n.d. unpublished. Wayne State University, Walter P. Reuther Library, DFWHCC, Box 5, File 11: Lesbian Self-Help.) Lesbian health emerged from the critiques by Brass and others primarily as an unarticulated need that was neglected by the very movement dedicated to reconceiving women's health care. In this critique, lesbian health was separate from (and unmet by) women's health, and it therefore required and deserved its own particular focus. It was unclear from this critique, however, just what lesbian health might include.

Unlike Brass and others, some lesbians and straight health activists denied that the women's health movement was either heterosexist or homophobic, arguing instead that it ultimately benefited all women. In the oft-cited 1973 publication 'Lesbian Health Care,' Frances (Francie) Hornstein, a lesbian living in Iowa City, made the case. She was active in the Women's Liberation Front and hoped to transform her interest in women's health issues into feminist action. Her first efforts, bringing feminist sensibilities into "a local free clinic" were frustrated by the pervasive sexism even in this "alternative" setting. Still, she vowed to serve the health needs of women. When Carol Downer and Lorraine Rothman came to Iowa to demonstrate the 'Self-Help Clinic,' Hornstein witnessed "the most revolutionary concept in the area of women's health care." As she gained information and watched women learning from each other, she "saw very clearly just how we were going to create women's medicine!" In the aftermath of the demonstration, Hornstein and three of her Iowa City compatriots packed up and moved to Los Angeles to work with Downer and Rothman. (F[rances] Hornstein, 1973/1974. Lesbian Health Care. Feminist Women's Health Center, Los Angeles. pp. 1, 2.)

After arriving in Los Angeles, however, they began to hear grumblings from some lesbians, who asked, in effect, "Why would three nice lesbians like you want to work so hard for straight women?" Hornstein was shocked to hear the movement framed in such heterosexist terms. She condemned the lesbian critique as "foggy, irrational thinking." She countered: "The whole idea that we, as lesbians were being used by heterosexual women to do the work that would benefit only straight women was very narrow thinking…both in an immediate practical sense as well as in a wider political sense." Most importantly, however, she noted that abortion and contraception – the ability to control fertility – were "vital to all of us who are Feminists in light of the use of women's bodies by men for their purposes—from rape to population control—two issues that affect every one of us, regardless of our sexuality." Finally, she denied that reproductive health only affected straight women by reminding her readers that many lesbians have children and many others hope to. In sum, she insisted that "the myth that the women's health movement is for heterosexual women must be destroyed," because she believed that the movement's efforts to wrest control of women's bodies away from men benefited straight and lesbian women equally. For Hornstein, then, lesbian health was rightly and unambiguously a subset of women's health and thus addressed through the women's health movement. (Hornstein, Lesbian Health Care, 3.)

Many lesbians agreed with Hornstein and eagerly embraced the women's health movement, especially the self-help gynecology impulse that one claimed to be "…much like a course in self defense." "For as lesbians, we are the most likely to be intimidated, patronized, and literally ostracized by a male-dominated medical establishment which knows nothing—and cares even less—about the kinds of health care lesbians need and want." (Riley, Kathy, December 1974. Self-defense and health care. Lavender Woman 3, 5.) Other lesbians organized lesbians-only self-help events. In September 1976, for example, the Atlanta Lesbian Feminist Alliance held a potluck dinner and a self-help clinic. (Self-Help Clinic, ALFA Newsletter, September 1976, 6.) Lesbians in other cities (e.g., Berkeley and Chicago) created their own self-help groups. (Monthly Extract 5 (September/October 1976) and 2 (April/May 1973).) Presumably in response to demand, other health centers, many of them feminist, also held regular clinics devoted to meeting the health needs of well lesbians. (See, for example, 'Womyn's Night' every Tuesday at St. Marks Clinic, NYC sponsored by the Lesbian Health Collective. 'Lesbian Health Collective, St. Mark's Clinic', WomaNews, 1 (April 1980), 6. 'Lesbian Well-Woman Clinic', Feminist Women's Health Center, Atlanta, GA, n.d. Emma Goldman Clinic for Women in Iowa City also had monthly gynecology clinics for lesbians. Klein, Paula, 10 September 1980. Health Needs Assessment in a Lesbian Community. Emma Goldman Clinic (Iowa City, Iowa) records, Iowa Women's Archives, The University of Iowa Libraries, Iowa City, Box 14.)

These lesbian-centered clinics sometimes failed to attract crowds of enthusiastic or curious lesbians, however. The Emma Goldman Clinic for Women in Iowa City, for example, attempted to overcome the perception that the clinic had a heterosexual bias by sponsoring monthly gynecological clinics for lesbians, but lesbians failed to show up and the clinic ended. (Emma Goldman Clinic (Iowa City, Iowa) records, Iowa Women's Archives, The University of Iowa Libraries, Iowa City, Box 14, Minutes of Lesbian Health Committee. Minutes of 26 April 1976 meeting.)

The tensions around sexual identity within the women's health movement were also evident in *Our Bodies Ourselves* (*OBOS*). Probably because none of the original members of

the Boston Women's Health Book Collective identified as gay, lesbian's health needs (and sexuality) received short shrift in the original 1970 newsprint edition (*Women and Their Bodies*) and in the 1971 New England Free Press edition (*OBOS*). For the 1973 Simon and Schuster edition, the collective recruited women to write a chapter by and for lesbians, but *In Amerika They Call Us Dykes* suggests an unhappy collaboration. As the lesbian authors wrote: "We had no connection with the group that was writing the rest of the book … and in fact we disagreed, and still do, with many of their opinions." (Cited in Kline, *Bodies of Knowledge*, 37.) That the flagship women's health publication had to farm out its lesbian content to a separate organization challenges the claim that the women's health movement could fully meet the needs of lesbians. Over time, *OBOS* developed more lesbian-specific content and a critique of homophobia, but as of 1982, a lesbian reader still complained, "Only about 1/3 of the book applies to me." (Cited in Kline, *Bodies of Knowledge*, 38.)

By the mid-1970s, the lesbian feminist movement, separate from the lesbian health movement, had identified health issues that disproportionately affected the lesbian community. Homophobia and its relation to alcoholism was understood as a particular problem. In 1975, for example, the East Lansing, Michigan-based *Lesbian Connection* published a three-part series on 'Lesbians and Alcoholism.' These articles identified the "modern tragedy of lesbian alcoholism," claiming that one in seven lesbians was an alcoholic. 'Kristi,' the author of the articles blamed both homophobia, for condemning lesbians as "queers and perverts," and the gay bar scene, as the "only place where lesbians can be lesbians." In these articles, the author explained why lesbians might suffer disproportionately from alcoholism. "Not only are we oppressed by strait [sic] society, but we are also trapped by a subculture that condones drinking, and further ensnared by an institution that definitely encourages it." (Kristi, February 1975. Lesbians and alcoholism. Lesbian Connection 1, 4. See Kristi, March 1975. Lesbians and alcoholism: part II. Lesbian Connection 1, 6–8; see also Are you a lesbian alcoholic? Lesbian Connection 1, 18 (December 1974); and Catherine, May 1975. An address to alcoholic lesbians by a drunken Dyke. Lesbian Connection 1, 18–20.) These statements from the lesbian feminist movement highlight the possible health consequences of homophobia and demonstrate the need for a feminist critique of medical sexism to also be a critique of homophobia.

The women's health movement helped to create and define the category of lesbian health in a variety of ways. Emerging as a critique of the women's health movement, lesbian health was first identified as an absence, as a category of neglect, a symbol of the heterosexist assumptions of the women's health movement. In response, representatives of the women's health movement, both gay and straight, insisted that lesbians, because they were women, benefited from the movement's efforts to wrest control of women's bodies from men, constructing lesbian health as the beneficiary of the broader improvements in women's health created by the feminist health movement. Finally, health activists understood that lesbians might want to explore their bodies, sexuality, and health needs with other lesbians, so they established lesbian-centered events and clinics. In this way, the health movement constructed lesbian health as a category of service offered to lesbians. The larger women's health movement, however, did not place homophobia at the center of their critique of medicine. For the most part, that critique came from activists aligned with the lesbian feminist movement outside the parameters of the women's health movement.

Establishing Lesbian Health Institutions

Although some feminist health clinics offered lesbian nights and lesbian self-help clinics, these gestures were not enough for many lesbians. Uncomfortable with the homophobia of mainstream medicine and the heterosexism of some feminist alternatives, lesbians were more likely to skip routine health exams and to feel uncomfortable during health visits. In order to bring health care to underserved lesbians, a few lesbian feminists founded lesbian health clinics. These clinics provided an alternative to straight-focused women's clinics and to the male focused gay clinics that frequently treated lesbian health as an afterthought.

An early lesbian health clinic was developed in New York City. In 1974, Joan Waitkevicz and several other lesbian activists founded the St. Marks Women's Health Collective, which aimed to provide health care "by and for lesbians." Waitkevicz, a physician and a lesbian, was moved to offer lesbian-focused services because an acquaintance of hers had developed inoperable cervical cancer "because she didn't feel she could go to a regular doctor for a Pap test." (Labaton, Vivien, September/October 1998. Joan Waitkevicz: Putting Lesbian Health in Focus, Ms., 80.)

Another lesbian clinic was located in San Francisco. Lyon Martin Women's Health Services (named for early lesbian activists Del Martin and Phyllis Lyon) was conceived and founded by Patricia Robertson, a resident in obstetrics and gynecology; Sherron Mills, a nurse practitioner; and Alana Schilling, a medical administrator, all at the University of California San Francisco Medical School, to meet the needs of lesbians in San Francisco. When the clinic first opened in 1979, it was a small, makeshift production, operating two nights a week out of borrowed space with an all-volunteer staff. Every night the staff would cart in the patient charts, medical supplies, and educational materials to the clinic, and carry them away again at the end of the shift. The tenuous nature of the clinic, however, did not discourage clients. In 1980, working at full capacity, it served more than 500 patients. Believing that it could serve even more women with more space, the clinic opened its own offices in the Mission district in San Francisco in 1983.

After this move, many poor, monolingual Spanish-speaking women from the neighborhood, who were desperate to find a heath care facility that accepted Medi-Cal, claimed Lyon Martin as their own. As long as the majority of the clients remained lesbian identified, most of the clinic stakeholders were happy to serve any women in need of health care services. This take-all-comers philosophy was tested, however, when a new (1989) executive director, Marj Plumb, hoping to expand the clinic and expand it in a multicultural way, applied for and received funding for Lyon Martin to provide health and social services to HIV-infected women. While this expanded mission did bring in more women of color, many of whom desperately

needed quality health care, it diminished the percentage of women at the clinic who identified as lesbian. This led to a recurrent conflict about the mission and services of the clinic. Over the years since, various stakeholders – staff, clients, board members, funders, and community members – have debated the meaning of a 'lesbian health clinic.' Nevertheless, the clinic has thus far weathered these conflicts and continues (in 2013) to offer health care to the lesbian community of San Francisco.

When these clinics opened, they attracted patients because they offered a sensitive, experience-based alternative to the heterosexism so prevalent in the medical encounter. They could not however, claim special expertise in lesbian health issues since research on lesbian health barely existed in the 1970s. (Stevens, Patricia E., 1992. Lesbian health care research: a review of the literature from 1970–1990. Health Care for Women International 13, 91–120.)

Lesbians as a Topic of Health Research

Although it was accused of neglecting lesbian health needs, the women's health movement actually spearheaded efforts to research lesbian health concerns. Frances Hornstein's 1973 landmark piece 'Lesbian Health Care' focused on how venereal diseases and vaginal infections could be passed between women. Based on 2 years of feminist research at self-help clinics and with trusted medical professionals, Hornstein noted the rarity of STDs among lesbians but that some genital infections (e.g., warts and lice) might be passed between women.

Significantly for lesbian health, Hornstein admitted that not all women who identified as lesbians "relate sexually exclusively with women"; that some occasionally had sex with men, and that these women were at risk for the "serious health hazard[s]" of heterosexuality. This discussion highlights the slipperiness of the category of lesbian and how difficult it was to isolate lesbian health needs. Lesbians did experience venereal disease even if their status as lesbian might theoretically protect them from the perils of heterosexual sex. The lived experience of lesbians often included heterosexual practices. (Hornstein. Lesbian Health Care 4, 5.)

Other groups of activists quickly followed Hornstein with their own efforts to investigate and publicize the health needs of lesbians. (These groups included the Chico Feminist Women's Health Center; the Women's Community Health Center in Cambridge, Massachusetts; and the Emma Goldman Clinic for Women. Paula Klein and Suzanne Vilmain, Self-health for Lesbian Women, Emma Goldman Clinic for Women, 1978. Lesbian Self-help Group, Health Care for Lesbians, Chico Feminist Women's Health Center, n.d.; undated letter from Lesbian Pamphlet Group of the Women's Community Health Center to 'sisters'.) The most significant and comprehensive of these was the 1979 publication of *Lesbian Health Matters!* (1979) by the Santa Cruz Women's Health Collective. (For an overview of the state of lesbian health by mainstream and alternative publications see O'Donnell, Mary, 1978. Lesbian health care: issues and literature. Science for the People 10, 8–19.) *Lesbian Health Matters!* sought to consolidate health information specific to lesbians, provide information without heterosexist bias on various topics (e.g., menopause, alcoholism, self-care, and 'gynecological health'), educate providers about the health needs of lesbians, and build "strength and solidarity among lesbians." (O'Donnell, Mary, Pollock, Kater, Leoffler, Val, Saunders, Ziesel, 1979. Lesbian Health Matters! A Santa Cruz Women's Health Collective Publication, Santa Cruz. 5.) The pamphlet addressed issues of particular concern to lesbians including alternative fertilization and issues shared by straight and gay women (e.g., vaginitis). The book mixed the practical (avoid coffee while battling a bladder infection) with the political (lesbians use of alcohol to escape homophobia). In the end, the book identified the need for more research about lesbian health. The authors "suspect[ed]," for example, that certain health conditions were more common (breast cancer, alcoholism) or less common (syphilis, cervical cancer) in lesbians than in straight women, but they called for more research to provide "more accurate answers." (Lesbian Health Matters! 45.)

At the same time as the researching and publishing of *Lesbian Health Matters!*, other lesbians were also making lesbian health a research priority. These included a handful of lesbian physicians, psychologists, and medical students who were generally open at work and unafraid of the professional consequences of studying lesbians. (See, for example, Susan Johnson, an ob/gyn resident at the University of Iowa Medical School and Patricia Robertson, ob/gyn resident at University of California San Francisco Medical School.) These researchers did not see themselves as lesbian health activists; instead, they were meeting an unmet need of their community and advancing their professional careers. Nevertheless, these health care providers assumed significant risk in order to help identify the health status and health needs of lesbians.

By the 1980s, the medical literature on lesbian health, based almost exclusively on survey data, converged on several general claims: (1) lesbians frequently avoided medical professionals; (2) lesbians frequently withheld their lesbian identity from their health care providers; (3) many lesbians reported mistreatment from physicians after revealing their identity; (4) particular health risks or advantages of lesbians remained largely unstudied; and (5) the rates of alcoholism among lesbians appeared to be significantly higher (five to seven times higher) than in heterosexual women. These overarching findings demonstrated the need for further research into the health status of lesbians and the health consequences of homophobia. (See, for example, Johnson, Susan R., Guenther, Susan M., Laube, Douglas W., Keetel, William C., 1981. Factors influencing lesbian gynecologic care: a preliminary study. American Journal of Obstetrics and Gynecology 140, 20–28; Johnson, Susan R., Smith, Elaine M., Guenther, Susan M., 1987. Comparison of gynecologic health care problems between lesbians and bisexual women: a survey of 2345 women. The Journal of Reproductive Medicine 32, 805–811; Robertson, P., Schachter, J., 1981. Failure to identify venereal disease in a lesbian population. Sexually Transmitted Diseases 8, 75–76.)

Other survey research came out of a more explicitly activist impulse, often but not exclusively, supported by the fledgling gay and lesbian health movement. The most significant of these survey studies was the National Lesbian Health Care Survey (NLHCS) published in 1988. The idea for the NLHCS began in the late 1970s when activists in the nascent gay and lesbian

health movement noted the paucity of information on lesbian health. After securing sponsorship from the newly created National Gay Health Education Foundation – later known as the National Lesbian and Gay Health Foundation – and funding from the Msp Foundation, Caitlin Ryan, Bernice Goodman, and Judith Bradford set to work on an ambitious survey of lesbians across the country. The survey gathered and analyzed data from 1925 lesbians from every US state and several territories. This survey, published in 1988, remained the "most extensive study of lesbians" published through 2001. (Ryan, Caitlin, Bradford, Judith, 1999. Conducting the national lesbian health care survey: first of its kind. Journal of the Gay and Lesbian Medical Association 3, 91.)

This survey discovered and documented several things. Depression or sadness was the most common health problem; lesbians and straight women shared many health problems (e.g., arthritis, menstrual cramps, and allergies); nearly one-third of the lesbians had been pregnant and another third hoped to be; financial barriers kept many lesbians from getting the health care they needed; many lesbians resented the heterosexist attitude of their health care providers. In addition to these findings, the published report insisted that the real health status of lesbians would only improve through the broader social acceptance of lesbianism. It concluded: "unless service providers are able to view lesbianism as an acceptable lifestyle and not a symptom of mental illness or deviance, they risk delivering services that will be harmful rather than helpful or will prevent lesbians from accessing them." (Ryan, Caitlyn, Bradford, Judith, 1988. The National lesbian health care survey: an overview. In: Shernoff Michael, Scott, William A. (Eds.), The Sourcebook on Lesbian/Gay Health Care, second ed. National Lesbian/Gay Health Foundation, Washington, D.C., p. 39. See also Ryan, Caitlyn, Bradford, Judith, 1988. The National Lesbian Health Care Survey: Final Report. National Lesbian and Gay Health Fund.)

For most of the research in the 1970s and 1980s, the definition of 'lesbian' was generally unanalyzed. For the most part, lesbians qualified as lesbians for research purposes if they said they were. (Bisexuals, however they were defined, were sometimes considered in the same category as lesbians, sometimes in the category of straight women.) The research surveys were generally circulated through friend networks or distributed at various lesbian cultural sites including bars, musical festivals, and bookstores. Due to this method of data collection, the 'lesbian' studied was overwhelmingly white, middle class, and young. It captured those who identified with a cultural identity; it did not attempt to locate and survey those who participated in particular behaviors. Further, because cultural factors including race, ethnicity, and religion influence the acceptance of the label, using lesbian identity to recruit subjects probably underrepresented the women likely to have felt the effects of homophobia most intensely.

Building a Lesbian Health Movement

With the publication of the NLHCS, there was a flurry of activity on lesbian health matters. Grassroots lesbians were demanding more attention for their particular needs at gay community health centers. Gay civil rights groups included lesbian health as one of their concerns. (Warchafsky, Lynn, 1992. Lesbian Health Needs Assessment: Los Angeles, Los Angeles Gay and Lesbian Community Center; Michigan Organization for Human Rights, August 1991. Michigan Lesbian Health Survey, MOHR Special Report.) Lesbian health became the subject of books aimed at both medical and popular audiences. (See, for example, White, Jocelyn, Martinez, Marissa C., (Eds.), 1997. The Lesbian Health Book: Caring For Ourselves. Seal Press, Seattle; Stern, Phyllis Noerager (Ed.), 1993. Lesbian Health: What Are the Issues. Taylor and Francis, Washington, D.C.; Ponticelli, Christy M., 1995. Gateways to Improving Lesbian Health and Health Care: Opening Doors. Haworth Press, York.) Individual lesbians, touched by health issues, created foundations to address lesbian health issues. (Mautner Project for Lesbians with Cancer founded in 1989 is the most important of these private philanthropic ventures. In 2013, the Mautner project, which had abandoned its exclusive focus on cancer and had expanded its mission to include bisexual and transgender individuals, merged with Walker Whitman Health.) By the early 1990s, activists and observers considered these varied impulses a lesbian health movement. (See, for example, Denenberg, A History of the Women's Health Movement.)

One of the major milestones of this movement was attracting federal attention and dollars to the cause of lesbian health. In 1993, lesbian health activists met with Secretary of Health and Human Services Donna Shalala to request more federal attention to the health needs of LGBT individuals. In February 1994, lesbian and bisexual activists met in a Lesbian Health Roundtable to formalize recommendations to the DHHS and to establish a lesbian health agenda. (Solarz, Andrea L., (Ed.), 1999. Lesbian Health: Current Assessment and Directions for the Future. National Academy Press, Washington, D.C. IOM, viii.) These recommendations included the expansion of federally funded research on lesbian health issues. The Institutes of Medicine (IOM) took up this challenge, and in July 1997, it convened the Committee on Lesbian Health Research Priorities. The Committee was charged with assessing the extant research on the physical and mental health issues of lesbian, reviewing the methodological challenges involved in the research, and proposing a research agenda for the future. *Lesbian Health: Current Assessment and Directions for the Future* published in 1999 by the IOM was the path-breaking product of this workshop.

The IOM report highlighted the problems with the then-current research on lesbian health. The report, for example, cited research that demonstrated that lesbians have higher rates of some high-risk behaviors (e.g., smoking and problem drinking) and lower rates of some protective behaviors (e.g., pregnancy and oral contraceptive use) and thus claimed that lesbians were at increased risk (compared to women in general) of certain conditions including breast cancer and heart disease. The report cautioned, however, that no epidemiological evidence verified that lesbians truly suffered disproportionately from either breast cancer or heart disease. The report was willing to concede that lesbians, by virtue of having less heterosexual intercourse were "probably" at decreased risk for cervical cancer.

The report noted that the current reliance on survey data rather than on population data contributed to the current

uncertainty about the true health profile of lesbians. The report also pointed to the ambiguity in the category of lesbian itself. To illustrate the problem, the report cited a study that separated 'lesbian' into three possible characteristics: current same-sex desire, current identity as lesbian or bisexual, and same-sex behavior since age 18 years. The study showed that less than 16% of lesbians claimed all three aspects. This study highlighted the methodological incoherence plaguing the category 'lesbian' and called into question the legitimacy of comparing the extant literature on lesbian health.

Through questioning the existing literature, the existence of the IOM report fueled the visibility of the lesbian health movement and resulted in professional legitimacy for lesbian health research and other projects. In 1999, for example, the American Public Health Association called for more research on sexual orientation and gender identity. In June 2001, the *American Journal of Public Health* devoted its publication to LGBT public health. (Most, but not all, of the articles in this issue, however, lumped all the constituent populations of this acronym together, masking the unique aspects of lesbian health.) More researchers felt emboldened and inspired to undertake lesbian health research.

Some of this new research provided the population-based perspective called for in the IOM report. For example, a 2000 article published in the *Archives of Family Medicine* claimed to be the first study of lesbian health based on population data. The study confirmed earlier work that showed that lesbian and bisexual women were more likely than heterosexual women to have poor health behaviors (especially smoking and problem drinking), worse access to health care, and lower rates of satisfaction with their health care providers. (Diamant, Allison L., Wold, Cheryl, Spritzer, Karen, Gelberg, Lillian, 2000. Health behaviors, health status, and access to and use of health care: a population-based study of lesbian, bisexual, and heterosexual women. Archives of Family Medicine 9, 1043–1051.)

Who counted as a lesbian remained a problem in this new research. Population-based studies avoided some of the issues with survey data. It still often relied on self-identity as lesbian, however, and lesbian was increasingly recognized as a fuzzy category. In order to broaden the pool of research subjects, some researchers began calling for research on women who participated in same-sex sexual behaviors (women who have sex with women) rather than lesbians.

This new vocabulary and approach significantly expanded the breadth of the women whose interests were allegedly being served by the lesbian health agenda. It did not, however, clarify the scope of category of lesbian health. A review of the literature between 2000 and 2006 found that "sexual orientation was measured in more than 100 different ways" and that the results of the different classification schemes resulted in different understandings of the health consequences at play. (Bauer, Greta R., Jairam, Jennifer A., 2008. Are lesbians really women who have sex with women (WSW)? Methodological concerns in measuring sexual orientation in health research. Women and Health 48, 385.) Consequently, although the research on lesbian health increased, the ambiguity over who this category included remains unresolved.

The ambiguity over who counts as a lesbian in terms of research and access to health services has only become more obvious with the increasing number of lesbians who have come to identify as transgendered (or transsexual, transmen, Female-to-Male (FTM), *trans, or genderqueer). Over the last 20 or so years, transsexuals and transgender individuals have been an increasing visible and vocal part of the LGBT community. The health needs of transgender individuals are assuredly not identical to the needs of lesbians. Nevertheless, because many trans people emerge from the lesbian community and many remain identified with it, (Rafkin, Louise. Straddlng Sexes. SFGate.com (accessed 31.01.12.).) lesbian health concerns must now include the medical control of the embodied transition (testosterone use, surgery, surgical follow-up), as well as promoting respectful and sensitive general health care. Perhaps more importantly, the category of lesbian now includes people with varied embodied and psychological enactments of gender and sex. Looking out for the health needs of lesbians includes making sure that men with vaginas, women with penises, and lesbians who reject binary gender identities have a place to receive competent and compassionate health care without condescension or judgment. As trans-activist Leslie Feinberg put it, attention to the health needs of trans people "is life or death."

If the identity of the lesbian in lesbian health is and always has been unclear, the goals of the lesbian health movement have been and remain similarly uncertain. Health professionals clamor for more health research focused on lesbians. Lesbians of color demand a sustained analysis of the role of racism in the lesbian health movement and the health effects of racism on lesbian lives. Some lesbian health activists are committed to serving women with HIV/AIDS. Others focus on bringing reproductive technologies to lesbians. Homelessness, hunger, and poverty plague the lives of patients at lesbian clinics, but poverty has failed to emerge as a central concern of the lesbian health movement. By 2013, the movement has succeeded in gathering support and concern for lesbian health matters. It remains to be seen how the movement will decide which lesbians and which health matters matter most.

See also: Feminist Theory: Radical Lesbian; Gay and Lesbian Movements; Heterosexism and Homophobia; Lesbian, Gay, Bisexual and Transgender Aging; Lesbians: Social and Economic Situation; Sexism; Sexual Minorities and Sexual Citizenship; Sexual Minority Families of Choice and Family Networks: We Are Family; Sexual Orientation: Historical and Social Construction; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Diamant, Allison L., Wold, Cheryl, Spritzer, Karen, Gelberg, Lillian, 2000. Health behaviors, health status, and access to and use of health care: a population-based study of lesbian, bisexual, and heterosexual women. Archives of Family Medicine 9, 1043–1051.

Dibble, Suzanne, Robertson, Patricia, 2010. Lesbian Health 101: A Clinician's Guide. UCSF Nursing Press, San Francisco.

Goldhammer, Hilary, Mayer, Kenneth H., Potter, Jennifer, Makadon, Harvey J., 2010. The Fenway Guide to Lesbian, Gay, Bisexual and Transgender Health. American College of Physicians, Philadelphia.

Hornstein, Frances, 1973/1974. Lesbian Health Care. Feminist Women's Health Center, Los Angeles.

Johnson, Susan R., Guenther, Susan M., Laube, Douglas W., Keetel, William C., 1981. Factors influencing lesbian gynecologic care: a preliminary study. American Journal of Obstetrics and Gynecology 140, 20–28.

Kline, Wendy, 2010. Bodies of Knowledge: Sexuality, Reproduction, and Women's Health in the Second Wave. University of Chicago Press, Chicago.

Michigan Organization for Human Rights, August 1991. Michigan Lesbian Health Survey, MOHR Special Report.

O'Donnell, Mary, Pollock, Kater, Leoffler, Val, Saunders, Ziesel, 1979. Lesbian Health Matters! A Santa Cruz Women's Health Collective Publication, Santa Cruz.

Ponticelli, Christy M., 1995. Gateways to Improving Lesbian Health and Health Care: Opening Doors. Haworth Press, New York.

Robertson, P., Schachter, J., 1981. Failure to identify venereal disease in a lesbian population. Sexually Transmitted Diseases 8, 75–76.

Ryan, Caitlyn, Bradford, Judith, 1988. The National Lesbian Health Care Survey: Final Report. National Lesbian and Gay Health Fund, Washington, DC.

Shernoff, Michael, Scott, William A. (Eds.), 1988. The Source Book on Lesbian/Gay Health Care, second ed. National Lesbian/Gay Health Foundation, Washington, D.C.

Stern, Phyllis Noerager (Ed.), 1993. Lesbian Health: What Are the Issues. Taylor and Francis, Washington, D.C.

Stevens, Patricia E., 1992. Lesbian health care research: a review of the literature from 1970–1990. Health Care for Women International 13, 91–120.

White, Jocelyn, Martinez, Marissa C. (Eds.), 1997. The Lesbian Health Book: Caring for Ourselves. Seal Press, Seattle.

Medicine, History of Western

Alfons Labisch, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

At all times and in all cultures, attempts have been made and still are made, to cure diseases and alleviate suffering. In medicine in its narrow sense, the physician's appropriate action is derived from medico-scientific knowledge. Thus medicine developed as soon as diseases came to be explained 'according to nature' and were treated with scientific rationality. In the European context this took place in the fourth century BC in ancient Greece. We speak of 'medicine' and 'physician' only in contexts where diseases are not primarily seen in magical, mythical, religious, or purely empirical references. From ancient scientific beginnings in humoral pathology, different historical concepts of medicine have been generated via iatro-astrology, iatro-physics, iatro-chemistry, and iatro-morphology to the modern concept of iatro-technology. Concepts of medicine comprise a continuous causal interrelation between a particular physiology, a particular pathology, and its deducible therapy. This inner perspective of the scientific and conceptual history of Western medicine is broadened by an outer perspective. Guided by socio-historical, historico-sociological, and anthropological approaches, the institutional and organizational history of Western medicine is focused. In recent decades the leading concept of iatro-technological medicine is in transition to the molecular concept of medicine. The scientific medicine of the late nineteenth and early twentieth centuries was the reproductive rearguard of industrialization. By contrast, the medicine of the late twentieth and early twenty-first centuries is a forerunner of the ongoing scientific, economic, and social change. The transition to molecular medicine will reshape the concept of the human body and thus change individual life and the social world accordingly.

At all times and in all cultures attempts have been made to cure diseases and alleviate suffering. Healing methods nowadays range from experience-based medical treatment within the family and the lay-world to the theories and methods of experts. Medicine in a narrow sense refers to 'scientific' knowledge. Such special knowledge is distinct from other forms of knowledge following culturally characteristic criteria. The physician's appropriate action is derived from medico-scientific knowledge: We may only speak of 'medicine' and the 'physician' when diseases are not primarily seen in magical, mythical, religious, or empirical contexts. Medicine developed as soon as diseases came to be explained 'according to nature' and were treated with scientific rationality. Historically, this took place in the fourth century BC in ancient Greece. Scientific medicine generated subsequently from these ancient beginnings. This concept of modern Western medicine has come to be the leading concept of medicine world-wide. At the same time, there have always been alternative understandings and methods of medicine, which will also be mentioned briefly. Concepts of medicine comprise a continuous causal interrelation between a particular physiology, a particular pathology, and its deducible therapy. The inner perspective of the scientific and conceptual history of Western medicine (=1) will afterward be broadened by an outer perspective guided by *socio-historical*, *historico-sociological*, and *anthropological approaches* (=2). As we face a paradigmatic change of the concept of medicine in recent decades, the transition from the iatro-technological to the molecular concept of medicine will be analyzed profoundly (=3).

The Scientific and Conceptual History of Modern Western Medicine

The magical-mythical art of healing in indigenous cultures has been part of a world in which the human relationships both toward nature and toward societal structures have not been understood as independent in their own right. In a society of hunters and gatherers and in the early agriculturally determined cultures, diseases were the results of violated taboos, offended demons, or neglected ancestors. Therapies aimed at restoring the good old order: exorcism was practised, demons and ancestors were pacified and with this, the spiritual, physical, and social order were restored. The animistic-demonic medicine presented a closed rational system of thinking and acting. As the trepanations (perforations of the skull, which were disseminated world-wide) indicate, fairly complex surgical procedures were performed within this magical-mythical thinking.

Magical conceptions of the world not only had their effects on notions of disease and therapeutic methods of prescientific times but iatro-magic ('iatro' from ancient Greek '*iatros*': the physician) influenced medical theory and even practical medical action far into the seventeenth century AD. Even today, 'simile-magic' (e.g., ginseng-roots) or 'singularity-magic' (e.g., charms, symbols, spells) determine popular medicine, nature healing, and the so-called complementary medicine.

A religious world order characterized the highly developed medicine of the archaic cultures of Mesopotamia, Egypt, and Asia Minor. This order was guaranteed by (national) deities and the (divine) ruler. Diseases were seen as conscious or unconscious violations of the world order with which the healers were familiar. These healers were characterized by practical as well as prophetic skills. Thus physician and priest are hardly distinguishable. There was an abundance of empirical practices both in magical-mythical medicine and in the archaic art of healing. The one criterion that clearly distinguishes the latter from medicine in the narrow sense was that the entire art of healing remained metaphysically oriented. Even those treatments that worked also in the modern sense, were in their time assigned to a magical-mythical, i.e., religious domain. This also applied to rules of hygiene such as sanitary and dietary laws.

It was Hippocrates of Kos (*c*.460–375 BC) who took the decisive step toward scientific medicine in Classical Antiquity. Hippocrates and his school emancipated medicine from religious concepts. Diseases came to be explained 'according to nature.' Natural philosophical elements (fire, water, air, soil) and their qualities (warm, humid, cold, dry) were combined with the teachings of the body fluids (blood, phlegm, yellow and black bile) to become humoral pathology. Apart from this concept, ancient medicine developed in its progress a large spectrum of theories and methods in association with the respective philosophical schools. The spectrum ranged from the pathology of the body fluids (humoral pathology, humoralism) to the pathology of solid body parts (solid pathology). This included (functional) anatomy (including vivisections), surgery (e.g., vascular ligatures), and pharmacy (The Greek Herbal of Dioscorides Pedanius was written in the first century BC). The Alexandrine medical schools (Herophilus of Chalcedon, *c*.330–250 BC; and Erasistratus of Kos, *c*.320–245 BC) followed empirical, partly experimental concepts in the modern sense. With this, almost all the possible patterns of modern medical thinking were *in nuce* already put into practice, or at least, anticipated in Classical Antiquity.

Galen of Pergamum (AD 130–200) subsumed theory and practice of ancient medicine under a humoral pathological view. The authoritative Hippocratic–Galenian medicine entered Persian and Arabian medicine via Byzantium (e.g., Oribasius of Pergamum, *c*.325–400 BC, and Alexander of Tralles, *c*.525–600 BC). As a result of Islamization, Persian-Arabian medicine spread eastward to South East Asia and westward to the Iberian Peninsula. In the heyday of the Arabian-Islamic culture (AD seventh to twelfth centuries), ancient medicine was transformed into practice-oriented compendia in a process of independent scientific work. Far beyond the High Middle Ages, the 'Canon Medicinae' of Ibn Sina Avicenna (AD 980–1037) became the handbook even of Western European physicians.

For approximately 1500 years, humoralism was the dominating medical concept from India to Europe. The original theory of the four body fluids, qualities, and elements was constantly expanded: by the four temperaments (sanguine, phlegmatic, choleric, melancholic), by the four seasons, the 12 signs of the zodiac, then later by the four evangelists, certain musical scales etc. This concept allowed the physician to put the patient with his/her respective symptoms into the center of a comprehensive model: Continuously, the powers of the macrocosm and the microcosm mutually influenced each other within the individual. By means of thorough anamnesis, prognoses, and dietary instructions ('contraria contrariis'; allopathy) physicians supported the healing powers of nature ('vis medicatrix naturae'; 'medicus curat, natura sanat'), a model which (though according to the principle of 'similia similibus') is still valid today in homeopathy.

With the political decline of the Western Roman Empire during the fifth and sixth centuries, a Christian art of healing came to the fore in Western Europe. In this, the traditions of ancient practice (Cassiodor, AD 487–583) mingled with a magic-influenced popular art of healing and Christian cosmology to create the world of *Christus Medicus*. Such a theological-soteriological concept of medicine was able to take up the ancient cult of Asclepius. In early Christian and medieval 'iatro-theology,' diseases were interpreted as the result of sin or as a divine act of Providence. This still continues to have an effect on today's religious practices (votive tablets, intercessions, pilgrimages). In her medical writings (e.g., Physica; Causae et Curae), Hildegard of Bingen (AD 1098–1179) left abundant evidence of the medieval art of healing.

The so-called *monastic medicine* or *presalernitarian medicine* (AD sixth to twelfth centuries) was superseded in Southern Italy and Spain by the adoption of ancient traditions from Arabian medicine. With Salerno in the eleventh century (Constantinus Africanus, AD 1018–87) and Toledo in the twelfth century (Gerhard of Cremona, AD 1114–87 et al.), the era of scientifically oriented medicine taught at universities began (Northern Italy; Montpellier; Paris). Hippocratic–Galenian medicine became the all-embracing and generally binding concept.

Neoplatonism and astrology (e.g., Marsilius Ficino, 1433–99) contributed to shaping Hippocratic–Galenian medicine into an all-encompassing model of microcosm/macrocosm during the Renaissance. Though initially not directly critical of the authorities but of corrupted textual transmission, modern medicine has been gradually developing since the Renaissance. This development was furthered by numerous text discoveries (e.g., Celsus, approx. AD 40; De Re Medicina, discovered in AD 1455). *Andreas Vesal* (AD 1514–64) in his secular anatomical work 'De Humani Corporis Fabrica' (AD 1543) wanted to rid the old authorities of bad transmissions. More and more, medicine based itself on its own experience, on its own theories and on its own experiments ('experimenta ac ratio'). *Theophrast of Hohenheim*, known by the name of *Paracelsus* (AD 1493/94–1541) introduced a chemistry-oriented way of thinking into medicine, while René Descartes (AD 1596–1650) contributed a mathematical-mechanical thought pattern; both were complemented by *Francis Bacon's* (AD 1561–1626) empirical thinking. The human body was released from religious cosmology and gradually became the outstanding subject of scientific, i.e., (in the first place) anatomical examinations.

A secular step was taken at the end of the eighteenth and the beginning of nineteenth centuries in English and French hospitals: the focus of medicine and the physicians turned from the patient to his/her disease. The *sick man disappeared from medical cosmology*. Seventeenth century attempts at scientific systematization led to extensive nosologies (Thomas Sydenham, AD 1624–89) in medicine. The Hippocratic–Galenian symptomatology, which was directed at the individual patient, was replaced by described syndromes. These syndromes became detached from the individual sick person who turned into a mere case. With this, a scientific-experimental field of separate research of disease entities opened up for medicine.

The man–machine model of *iatro-physics* and the man–reagent model of *iatro-chemistry* finally introduced modern scientific thinking into medicine. As medicine was being progressively freed from any religious or metaphysical reasoning, the question of the source of life was raised. It was answered by dynamic concepts of disease, by psycho-dynamism and bio-dynamism. These concepts finally led, via Brownianism, to modern psychopathology and psychotherapy. In contrast, iatro-morphology was built on the visible and observable. Georgio Baglivi (AD 1668–1708) paved the way from humoral to solid pathology. Pathology improved from

Giovanni Battista Morgagni's (AD 1682–1771) organ pathology via *Francois-Xavier Bichat's* (AD 1771–1802) tissue pathology to the functional cellular pathology of *Rudolf Virchow* (AD 1821–1902). Cellular pathology established the physiological and pathological fundamentals of scientific medicine in a modern sense.

The actual step toward modern scientific medicine was taken in the late eighteenth and early nineteenth centuries. Initially chemical and physical, then biological thinking – which was gradually mathematically-statistically tested – has been determining the theories of medicine ever since. Deliberately reproducible (research) results led to a 'triumphal victory' of modern medicine in the second half of the nineteenth century. The development of anesthesia (nitric oxide AD 1844/46; AD 1846 ether; AD 1847 chloroform); antisepsis (Joseph Lister, AD 1827–1912); asepsis (AD 1886/1891: Ernst von Bergmann, AD 1836–1907; Curt Schimmelbusch, AD 1860–95); immunology; serum therapy; and the diagnostic (AD 1895/96) and therapeutic (AD 1896) possibilities of X-rays must be acknowledged for the surgical disciplines. With acetylsalicylic acid, the first chemically representable fever-reducing drug was introduced into medicine in 1873. Along with chemistry, industry with its economic concerns, became involved in medicine.

Modern scientific medicine follows the *iatro-technical concept*; pathophysiology, causal analysis of distinguishable disease entities, objectifying and gauging methods, and causal therapy characterize such medicine. The driving force of this scientifically and technically oriented medicine is its endeavor to put physiology, pathology, and therapy on a scientific footing; first chemistry, then physics became central guiding sciences; laboratories and experiments were introduced even into clinical practice. With the animal model, bacteriology launched a biological experimental situation. From the start, the *iatro-technical concept* created specific tensions in medicine: On the one hand, the scientific fundaments which aim at general insights get into the maelstrom of the physician's need for action (consequence: the 'autistic-indisciplined thinking' and acting; Eugen Bleuler, AD 1857–1915). On the other hand, medical action directed at the sick individual patient as a subject gets caught in the maelstrom of generalized technical applications and scientifically justifiable therapies (consequence: the '*therapeutic nihilism*' of the Second Viennese School; Josef Skoda, AD 1805–81). As a result, since the 1840s, both patients and medical practitioners have been turning to and from homeopathy and nature healing over and over again.

The Institutional and Organizational History of Western Medicine

Medicine is a science built neither on pure knowledge nor on mere empiricism. It is rather an 'action-based science' (Handlungswissenschaft), directed at the sick individual. As already formulated in Hippocrates' first aphorism, the patient and his/her environment are granted equal importance in medicine. Therefore, a scientific and conceptional internal perspective can only partly explain the significance of medicine. From a historical perspective, diachronic questions are necessary, directed at medicine from outside, i.e., from society. Consequently, a comprehensive history of medicine also demands socio-historical, historico-sociological and anthropological perspectives.

The dialectics of scientific knowledge and practical experience has been part of medicine ever since. Due to the category of experience, *history in medicine* has presented a significant argument since the existence of the earliest written evidence (such as: 'On Ancient Medicine' in the Corpus Hippocraticum). In the course of making medicine 'scientific' during the nineteenth century, the historical argument simply disappeared from scientific reasoning. As a well aimed counter-reaction against the purely scientific-technical concept, the history of medicine was reintroduced into the medical curriculum in the early twentieth century. Far into the 1990s, the history of medicine was considered in medical faculties and schools to be the substantial representative of the *humane aspects of medicine*. Just as much as objects and methods of the historical sciences began to expand in the late twentieth century, *medicine also became an object of general history*, i.e., especially of social and cultural history (*see* Social History; Environmental History; Body, History of; Gender History; etc.).

Similar to the steadily growing specialization of the scientific and clinical aspects of medicine, the sciences that specialized more and more dealt with the aspects of medical practice in the interaction of doctor and patient and the psychological and social environment of this interaction. Since the 1970s a broad canon of disciplines developed besides the classical history of medicine, so that a group of 'humanities in medicine' emerged. These can be summarized as: medical history, medical sociology, medical psychology, psychosomatic medicine as a special subject, economy of health care etc. With the increasing technical possibilities of modern medicine, especially at the beginning and the end of life, questions of an ethical nature got ever greater importance both in theory and in the clinic. Since the 1990s, the ethics of medicine has grown into its own as a theoretical and practical subject. Due to the different tasks to be addressed – which partly refer to the current problems of legitimization and action, and partly to the debate about contents and methods in historiography – it is useful to distinguish between a 'history of medicine' and a 'history in medicine.' 'History of medicine' can be conceived as a part of general history – 'history in medicine' is a part of the theory of medical action in its historic settings.

Within the doctor–patient encounter, patients receive special attention due to their special need for help, as do doctors due to their specific need for action. From Antiquity to the beginnings of scientific-technical medicine, the conceptual world of medicine and patients' lay interpretations have overlapped in large areas. Humoral pathological medicine lives on in popular medicine (Volksmedizin), in nature healing, and in alternative medicine ('rheumatism,' 'blood poisoning,' 'diets,' etc.). Due to the scientific concept, the world of medicine and that of the patient drifted apart. Since the eighteenth century, the physician as an expert began to face the patient as a lay person. At the same time, medicine came to hold the monopoly of definition and action over the human body. In the socio-genesis of health as a social good and value, medicine and the physicians were endowed with an exclusive task: They

were given a special position in the society. The historical and sociological discussion of the 1970s called this process 'the professionalization of medicine' (*see* Professions and Professionalization, History of).

The medical marketplace has always been characterized by enormous competition. Because of that, the healing professions have almost always depended on their clients, the patients. In medieval and early modern times, the patronage system made the academically trained physician subject to his socially superior patron–client. This was only inverted with the gradual scientification of medicine in the eighteenth century. With the compulsory linkage to scientific medicine, 'orthodox medicine (Schulmedizin)' has developed since the 1840s. Orthodox medicine was confronted with traditional or modern alternatives as a kind of 'outsider medicine'; healers who applied these methods were both internally and externally discriminated against as quack doctors. Industrial societies with their various forms of *social security systems* and their general claim of social inclusion, opened up for medicine a new clientele of the working class and the lower economic strata for medicine (*see* Poverty in History; Welfare State, History of; Social Inequality in History (Stratification and Classes)).

A significant institution of medical help both in the Western and in the Islamic tradition is the hospital: Hospitals (from Latin 'hospes,' meaning 'host') were established partly for pilgrims and partly for anybody in need of help, in order to fulfill the Christian ideal of 'caritas.' When it migrated to the orient along with ancient medicine, the hospital became the place of training for physicians in Islam, around the turn of the millennium. Within the European tradition, the charitably oriented hospital changed only in the late eighteenth century: The modern hospital (German: *Krankenhaus*) came to be a place to which only sick people were admitted, with the purpose of discharging them after a well-directed treatment within a calculable period of time. Such was initially the responsibility of proficient 'nursing.' Around the same time, selected hospitals had become places of practical training for future physicians. Scientific-medical research has also been conducted in these university hospitals from the 1840s onwards. Only as a result of antisepsis, medicine entered the General Hospitals in the 1860s. This was the beginning of the history of modern nursing, of scientifically oriented medical training, of the many medical disciplines, of the modern Medical Departments and Schools, and of the hospitals which only then became medically oriented.

On first sight, *diseases* embody the biological reality in the encounter between physician and patient. As the examples of leprosy or plagues show (*see* Plagues and Diseases in History), there are neither in one period nor over the periods, standardized terms to label diseases. Therefore it is hardly possible to determine with certainty which diseases were referred to in reports handed down to us (e.g., the so-called 'plague' of Thucydides). However, culturally interpreted diseases induce the encounter between physician and patient. It is therefore the sick human being who accounts for the existence of medicine. Over and above the conceptual and institutional interpretations, diseases have always been experienced as crises in human existence. In different cultures and eras sick people and their environments interpreted diseases always with reference to their metaphysical reasoning. Diseases are given a cultural meaning as a divine trial or as a retaliation for violated principles of life and nature. The scientific-analytical notion of disease in the iatro-technical concept is not suitable for taking the patient's quest for meaning into account. The chronically ill, the terminally ill, and the dying, therefore, present a constant challenge for scientific-technical medicine – including its necessarily inherent thoughts of progress. The result is a continuous push toward *alternative and complementary medicine as well as toward 'lay' and 'self' help*.

From an historical and sociological point of view, even the *interpretations of health* take place in a repetitive processes. In antiquity, health resulted from the philosophical ideal of the beautiful and good human being (ancient Greek: *'kalokagathia'*). The Islam and the Christian Middle Ages assigned the health of the body to religious categories. In the sciences of early modern times, the nature-bound human body came to be understood as an objectified body (The dichotomy of German: 'Leib/Körper' is a differentiation in philosophical anthropology which is hardly possible to translate into English: 'Leib' means the self-given presence of the body; 'Körper' means the scientifically perceived and shaped part of body as a 'living corpse'). In early rationalism, Gottfried Wilhelm Leibniz (AD 1646–1716) elevated medicine into a means of ordering private and public life according to scientific rules: the order of the soul, the order of the body, and the order of society corresponded with piety, health, and justice; they had to be guaranteed by the church, by medicine, and the judiciary. In critical rationalism, reason, physics, and morality were declared a unity (cp. Christoph Wilhelm Hufeland (AD 1762–1836), Macrobiotics). In the early nineteenth century, the interpretations of health joined the sciences. The question of the meaning of health was excluded from medicine. Within the iatro-technical concept, health as a chemical, physical, or statistical norm (percussion, auscultation, thermometer, etc.) became a negative definition of scientific notions of disease. Those values and ideas of order which were indispensably associated with the notion of health seem to have disappeared during the scientific-rational progress.

With the scientific notion of health and disease, modern medicine was finally granted the monopoly of interpretation and action for a rational design of the biological human resources in modern societies. Disease is the cause for medical action, health is its aim; via the societal value of health medical knowledge and medical action, it is integrated into the societal surroundings. The definition of health mediates in its valid interpretation of a particular civilization between the individual, the social, and the 'natural' nature of mankind. Thus, health integrates the 'natural' biological basis, the social basis, and the individuality – the 'I' and 'me' of human existence. *In terms of philosophical anthropology*, the valid notions of health mediate between the outer world/nature (= 'Aussenwelt'), the social world/society (= 'Mitwelt') and the inner world/individuality (= 'Innenwelt') with regard to their body/living corpse. The place and the scope of medical action in society are defined via the interpretation and the effect of the valid definitions of health.

To the same extent to which the human relationship toward nature becomes more scientific and leads to a scientific–technological civilization, more and more aspects become scientific within the combined interplay of the body as a part

of oneself ('Leib'), and the body as an object ('Körper'). The 'homo hygienicus,' i.e., mankind, who define themselves and are defined by others in terms of medical standards, are typical examples of the paradox of modernity: The autonomously progressing discoveries of scientifically oriented medicine liberated the person, the communities and society from the task of having to account for the values and norms of their bodily existence. From this position, the individual and public 'rationalization of the body,' as was stated by Max Weber (1864–1920), amongst others, at the turn of the century, could begin their widely welcomed and from all sides demanded 'victory march of medicine.' With the end of the belief in progress, the disadvantages of this process have been perceived since the 1970s. So in historical, sociological, and anthropological debates, medicine turned into an 'institution of social control.' '*Medicalization*' – as Michel Foucault (1926–84), amongst many others, pointed out from the early seventies onwards – became the focus of a fundamental critique on the penetrating social impact of modern medicine (*see* Body, History of; Gender History).

The doctor–patient relationship is no anthropological constant. Rather, this relationship is an integral part of the economic and social organization of a community. This is especially true for public medicine and the health sciences. Early forms of the modern public health service developed at the end of the thirteenth and the beginning of the fourteenth century, in North Italian cities. The driving forces were the endemic plagues, which kept coming back at regular intervals after the Great Plague of 1347–51 (*see* Plagues and Diseases in History). In early modern times those public health measures evolved and became regular institutions of the cities of commerce: General regulations which also had health effects (such as food inspection and market orders etc.); municipal supervision of those practicing medicine; city hospitals for the infirm, and others for special isolation (leprosaria, plague hospitals, etc.); the beginnings of the municipal doctor and surgeon service.

When the territorial states were thoroughly consolidated on the legal and administrative level, the medical and sanitary supervision unfolded in the late seventeenth and early eighteenth centuries. During Enlightened Absolutism, the 'medical police' (German: 'Medicinische Polizey'; e.g., Johann Peter Frank, 1745–1821) generated a public medicine which was linked to administrative and political goals. Within the framework of mercantilism, one intention in calculating power was to increase the population. The paternalistic welfare state toward the end of Absolutism developed 'public health' (German: Staatsarzneikunde) as part of its 'populating policy' (German: 'Peuplierungs-Politik'). This is where the special, still heavily disputed relationship between state, *medicine and women* (*see* Gender History) begins: to the extent to which women were made the 'human capital' of a well-directed population policy, their role came to be substantially determined by medicine (medicalised midwifery, medicalization of birth and childhood, medicalization of housekeeping, child-rearing, etc.)

Through the bourgeois revolutions the concept of the nation as sovereign made itself the object of public health. Besides the idea of a qualitative evaluation of the population the thought evolved to calculate 'the value of a human being in terms of money.' In a combined act of statistics, epidemiology, physics, geography, metereology, etc. and driven by a paternalistic idea of welfare, the first modern health sciences came into being. The nineteenth century cholera pandemics accelerated this development. With the beginning of industrialization, a comprehensive health care was promoted due to the recognition of the 'public value' of health. The modern health sciences finally defined, in a biological chain of causes and effects, a closed circle of human hygiene: the environment as conditional hygiene of Michel Levy (AD 1809–72), Jules Guérin (AD 1801–86), Edmund Parkes (AD 1819–76) or Max von Pettenkofer (AD 1818–1901); the concept of microbiology or bacteriology as infectious hygiene pioneered by Louis Pasteur (AD 1822–95) or Robert Koch (AD 1843–1910); the dynamic relation of exposition and disposition as constitutional hygiene; the classification of forthcoming human life through racial hygiene and eugenics; and finally the concept of health and society as social hygiene – all these combined approaches took hold of basically all areas of human existence in the late nineteenth and early twentieth centuries. This is especially true for the unborn life. At this point the historical reflection on public health care, medical statistics and epidemiology, health economics, health laws, and health systems etc. begins.

The idea of obligatory collective health has been developed since the turn of the twentieth century. In the *biologistic ideology of National Socialism* it was molded into a highly rational model of future-oriented health. The goal of *National Socialist health care* was a 'genetically healthy' and 'racially pure' population of 'Arian-German blood.' This genetically and racially cleansed 'Volk' would be fit to survive the secular 'war of races.' The goal of obligatory national – which really meant: racial – health was located beyond all individual rights. This model of a *totalitarian program of health care* was based categorically on exclusion. It was carried out in decisive areas: at first sterilization (at least 350 000 victims, all of them Germans), then 'euthanasia' (over 100 000 victims, including so-called 'wild euthanasia,' probably more than 250 000 victims; all of them Germans), and finally the determination of 'Völkerparasiten' during the Holocaust (approximately 6 million victims, approximately 180 000 Germans).

Medicine in National Socialism asks for special reference here, as it is historically unique in the world. Basically, there is always inherently an excluding function in medicine. In the daily routine of medicine this becomes apparent wherever decisions have to be made on scarce goods and services in short supply (e.g., transplantation medicine, assessment and social medicine, allocation of medical services). As to their forcefully carried out eugenic actions the National Socialists themselves referred to the examples of other countries, especially the United States. Indeed there had been, prior to 1933 and after 1945, an 'Internationale' of eugenicists. The National Socialist model of health serves as a historical example of the conflicting nature of the rationalization process on which modern times are based – a process which is irreversible. National Socialism is therefore to be seen as the janus-faced flipside of the modern era, which is just as much obsessed with progress as it is blinded by it. This means that National Socialism and medicine in the National Socialist era are not only a particular phenomena of German history. They are rather a problem of our times

and thereby a problem of all societies, which have built on modernism.

The Recent Transition to Molecular Medicine

In the 1980s the concept of a natural-scientific and iatrotechnological medicine began to shift into the concept of molecular medicine. The transition to molecular medicine has important implications for medicine and health care in general. A concept of medicine is a self-containing compound of a physiology, a pathology and a therapy. The concept of natural-scientific medicine and the consequent iatrotechnological approach – starting with Johannes Müller (AD 1801–58) and continued by his famous disciples such as Emil du Bois-Reymond (AD 1818–96), Hermann (of) Helmholtz (AD 1821–94), Rudolf Virchow (AD 1821–1902) or Ernst Haeckel (AD 1834–1919) – regarded the natural sciences as exclusive referential disciplines of modern medicine. Diseases appear to be incidents of a machinery that, in principle, can be repaired by chemical or physical, and technological interventions. From the 1980s and intensifying in the 1990s, the reference disciplines of chemistry and physics were replaced by biology, and especially by genetics. The necessary background technology is electronic data processing, which has accelerated research significantly, e.g., during the Human Genome Project, and in fact has even been the factor that made research possible in the first place. The classical natural sciences will not disappear, but they vanish into the second row, in order to clarify the novel issues raised by molecular genetics.

The transition to molecular medicine was – as always in such secular processes – accompanied by mistrust and misunderstanding. At the turn of the twenty-first century, even some doctors did not admit that a change was ongoing, at the same level of importance as the cellular pathology of Rudolf Virchow to medicine in the 1850s or the microbiology of Louis Pasteur and the bacteriology of Robert Koch to medicine in the 1870s and 1880s. In molecular medicine the cell is not the smallest explanatory unit, which can optionally be influenced chemically or physically or removed surgically or by strong radiation. The cell rather appears as a separate universe of significant data storage, messages, processes, products, and features. These highly complex events of the cell are controlled by a self-regulating process of genes (genomics), proteins (proteomics), and metabolism (metabolimics) that are in constant interchange with the requirements given from the outside (epi-genetics). Molecular biology and the genotype take the place of the former morphology and phenotype of features and processes. Two examples may illustrate this fundamental change. In molecular physiology mice lacking the gene for hemoglobin and thus lacking red blood cells, do not differ in their phenotype from mice that possess hemoglobin. Inflammation in rheumatoid arthritis has been previously treated with steroids and nonsteroidal, antiinflammatory drugs. The body is flooded with these drugs, in order to access the areas where the drugs must discharge their chemical effects. Of course, the drugs will work in many other areas, too. In molecular antirheumatic treatment the so-called biologicals are given. These drugs imitate human proteins and stop the destructive anti-autoimmune process. Thus, the disease is interrupted in its causal process. The intervention with TNF-alpha antagonists, however, is not specific to rheumatic arthritis, and therefore also can lead to serious side effects. This was noticed in the development of this therapy. As a result, rheumatologists have recently learned to treat the so-called 'nonresponders,' a group of patients particularly difficult to medicate.

Molecular medicine is the medicine that investigates biochemical processes from messages encoded in genes referring to the respective production processes up to the products of the cell – in particular nucleic acids and proteins – for the physiology and pathology of organisms, in order to use the results for an appropriate therapy. A person's genetic repertoire is a singular entity. Mankind is identified by its genes as individuals – today genetic fingerprints are customarily recognized in paternity or criminological studies (restriction enzymes, noncoding regions of DNA). Molecular medicine is a medicine that leads to an individualized, and personalized treatment. Ultimately drugs have to be prepared for the individual patient. This is already done in certain areas of medicine: for example, in hemat-oncology cytostatics are individually related to the specific surface marker of the individual patient. This explains the progress that has been achieved, particularly in children suffering from blood cancers.

The question that now arises is whether, when, and in what form molecular medicine gets relevance in health care and in society in general. All half-true, sensationalist, and imagined concerns against molecular medicine have to be excluded in advance. Cloning or genetic enhancement of humans are beyond biological possibility – if there should be some biological pathways, the results are unpredictable. These evil utopias reflect, however, the public fears and concern, which accompany this recent explosion of knowledge in medicine, just as it was in the case of cellular pathology in the 1850s or in bacteriology in the 1880s – impassioned conflicts that we have stored in (medical) history, and that we will similarly store in history in respect of molecular medicine, when it is taken for granted. These discussions are to be taken seriously, because they reveal a genetically transparent world. But these discussions also obscure the tasks arising from this new level of explanation of medicine. The connection to the classic discussions, especially of public health care, is obvious: wherever the genetic component of diseases seems to be known, since the development of eugenics in the 1880s, always a discussion arises about public interventions in medicine. Attached to molecular medicine are hopes and fears, which a society has to deal with openly. Individual rights, property rights on knowledge and ignorance, and the determination of the genetic self, sets the limits.

At this point individual and public medicine should be distinguished once more – on the one hand the individual and personal encounter between doctor and patient, on the other hand the public, mostly administrative, work of public medicine. Under the aspect of an anticipative health care – be it individual or public – molecular medicine reveals a new dimension. Risk assessments are carried out by statistical tests. These statistically calculated risks are always related to collectives, never to individuals. Molecular medicine differs insofar, as specific risks for a specified group of diseases can be clearly identified biologically – for example the recognition of the BRCA-1 or BRCA-2 genes as risks for developing breast

cancer. As a result, two new risk groups have to be differentiated. Those that do not have these genes are still a risk group for breast cancer. Those who carry the gene, represent a risk group with a biologically manifest, higher probability, (but not total) risk of developing this cancer. Those that carry these particular dangerous genes, have the possibility of preventing the disease risk resolutely. These preventive measures may be early surgery – such as in hereditary colon cancer (HNPCC, Lynch-Syndrome) or hereditary thyroid cancer (MEP2); a permanent monitoring, generally with BRCA; certain rules, for instance provision of increased protection against radiation in BRCA II; and upto a fatal waiting – as is the case of Huntington's disease. Generally, the new phenomenon of the 'healthy ill' emerges – the seemingly healthy patient who is exposed by hereditary features to an increased risk of disease and thus may have to live for decades facing this deleterious knowledge. This results in many other issues, such as the obligation to know and the right not to know at the individual level or risk-calculation in health or life insurances at a public level.

The task of a molecular medically oriented public health care would be to identify risk groups, to provide safe preventive measures, and to offer appropriate diagnosis and treatment if possible. This idea reveals the dilemma of a molecular, medically oriented, public health care. The basis of any effective public action of molecular medicine is a comprehensive genetic screening, or at least a systematic molecular epidemiology. The data should be stored in biobanks and made available to doctors and institutions of health care and the patient, respectively. Will this find public acceptance in view of the risk of data abuse, the possible violation of the right to personal data, and the chances of genetic identity leading to personal identifiability? Important for the consequences to organized medicine is the question of whether the data collected have, *a priori*, a clinical character. The collected data may directly lead to therapeutic interventions. This raises the question whether 'public health genetics' is a form of public health care or affiliated to curative medicine. A historic example is the transformation of the public mother and child care of the 1920s to individual peri-natal medicine since the 1960s. This raises the question of whether curative medicine has the perception and forms of public intervention. Public medicine can go out to vulnerable groups, while curative medicine is dependent on the demand of patients.

So on the one hand, there is the need to organize disease prevention, health promotion, and risk reduction based on molecular biological susceptibility to certain diseases. On the other hand, new forms of intervention are required for this new concept of public health genomics and protection, which unite the various areas of individualized medicine of gene testing, special interventions and preventive measures, to the necessary mass data and applied public health care measures. The facilities and measures that have been recently established, primarily take place at the level of research and the dissemination of available knowledge on international and national networks. Public interventions based on molecular medicine, such as in diabetology or in rheumatology, are as yet cohort studies or limited, research projects. Permanent forms of medical intervention of molecular public health care or of public health genetics have not yet been observed.

Conclusion and Outlook: Western Medicine and Molecular Health

In the narrow scientific-practical sense, medicine is determined by the dialectics of medical knowledge and medical action. Such dialectics have accompanied medicine since its first historical steps to its scientific-rational establishment in Greek Antiquity. So the controversy between 'dogmatists' and 'empiricists,' between 'medics' and 'doctors,' has characterized the positions during the entire history of medicine. It is the patient who gives these dialectics a direction: The patient's need for help constitutes medicine. Due to the patient and the special situation in which the patient and the doctor meet, medicine possesses a domain distinct from that of other sciences: In the encounter between doctor and patient medicine faces the necessity to take action. With this, the doctor–patient encounter is placed within a historical context. This is the starting point of historical reflection as an indispensable aspect, both of a 'history in medicine' and 'a history of medicine.'

In recent decades, mankind has experienced a fundamental change in all aspects of human life. During the eighteenth and nineteenth centuries the production of power shifted from humans and animals to machinery, with the then unforseeable consequences of industrialization. Today the production and distribution of data is being transferred from humans to machines, still with unpredictable consequences, (which finally will probably be even more dramatic than industrialization). Medicine is also covered by these processes. The position and the role of medicine within society is changing fundamentally. The scientific medicine of the late nineteenth and early twentieth centuries was the reproductive rearguard of industrialization. By contrast, the medicine of the late twentieth and early twenty-first centuries is a forerunner of scientific, economic, and social change. Since the 1980s, molecular biology has become the referential discipline of medicine. Thus it has become possible to diagnose diseases in their genetic informational principles and their molecular mechanisms. The molecular transition means a secular change in the concept of medicine. The molecular–biological thought pattern transfers the process of health and disease into the informational bases of the production and function of proteins. With that, genetics and information-processing in the cellular and extra-cellular realm, step into the limelight.

The historical analysis has shown to what extent medical knowledge and medical action affect society. Thus, experimental hygiene and bacteriology have cleansed and pasteurized conditions and behavior since the middle of the nineteenth century. Since the 1960s the pharmacological possibilities of contraception have altered the self-definition and the role of women fundamentally. Molecular genetics and clinical reproductive medicine already show their effects on the generative behavior of mankind (*in-vitro* and *in-vivo* fertilization, pre-implantation diagnostics etc.). The transition to molecular medicine will also change individual life and the social world: Molecular medicine leads to a new image of humans. The medicine of the last decades followed the model of 'exposition': diseases had an external origin (e.g., bacteria, viruses, stress factors, etc.). Molecular medicine shifts the etiology from exposition (i.e., the outside world of all human beings) to disposition and therefore to the (genetic) inner

world of the individual. Predispositions for certain diseases are internalized as an individual (biological) fate. The result is the 'genetization' of individual life and of its subsequent societal acts (e.g., life and health insurance). The interpretation of health including the notion of 'normality' of the body, will, in substance, be linked with knowledge of genetically determined dispositions. Such a genetic life world will be as 'real' for human beings of the coming decades as the bacteriological life world had been 'real' for the last century. The 'genetization of society' is the current aspect of the general 'medicalization' process and the continually progressing rationalization of corporeality.

See also: Academic Self-Concept and Achievement; Body, History of; Environmental History; Gender History; Plagues and Diseases in History; Poverty in History; Social History; Social Inequality in History (Stratification and Classes); Welfare State, History of.

Bibliography

Basic Information and Classical Reading

Bynum, W.F., Porter, R. (Eds.), 1993. Companion Encyclopedia of the History of Medicine, 2 vols. London/New York.

Duffin, Jacalyn, 2005. Clio in the Clinic: History in Medical Practice. Oxford Univ. Press, Oxford, u.a.

Duffin, Jacalyn, 1999. History of Medicine. A Scandalously Short Introduction, Toronto second revised an enlarged ed. 2010. Univ. of Toronto Press, Toronto.

Garrison, F.H., 1929. An Introduction to the History of Medicine, with Medical Chronology, Suggestions for Study and Bibliographic Data, fourth ed. Philadelphia/London.

Huisman, Frank, Warner, John Harley (Hrsg.), 2004. Locating Medical History: The Stories and Their Meanings. JH Univ. Press, Baltimore.

Kiple, K.F. (Ed.), 1993. The Cambridge World History of Human Disease. Cambridge.

Norman, J.M. (Ed.), 1991. Morton's Medical Bibliography. An Annotated Check-List of Texts Illustrating the History of Medicine, fifth ed. (Garrison and Morton). Aldershot.

Porter, R. (Ed.), 1996. The Cambridge Illustrated History of Medicine. Cambridge.

Rothschuh, K.E., 1978. Konzepte der Medizin in Vergangenheit und Gegenwart. Stuttgart.

Schott, H., 1993. Die Chronik der Medizin. Gütersloh, München (Reprint: 1997, 2000; Translations: Hungarian: Budapest 1993; Spanish: Barcelona 1994; Polish: Warsaw 1994; Czech: Prague 1994).

Shryock, R.H., 1936/1979. The Development of Modern Medicine. An Interpretation of the Social and Scientific Factors Involved. University of Wisconsin Press, Madison, Philadelphia, ND.

Sigerist, H.E., 1951–1961. A History of Medicine, vol. I: Primitive and Archaic Medicine; vol. II: Early Greek. Hindu and Persian Medicine, New York.

Information Sources and Contemporary Research

Bibliography of the History of Medicine 1ff., 1965ff.

Bulletin of the History of Medicine 1ff., 1933ff.

Current Work in the History of Medicine. An International Bibliography 1ff., 1954ff.

Isis. An International Review devoted to the History of Science and its Cultural Influences 1ff., 1912ff.

Journal of the History of Medicine and allied Sciences 1ff., 1946ff.

Medizinhistorisches Journal. Internationale Vierteljahresschrift zur Wissenschaftsgeschichte, Mainz, 1ff., 1966ff.

Social History of Medicine, Oxford 1ff., 1988ff.

Meiji Revolution, The

Masato Miyachi, University of Tokyo, Tokyo, Japan

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 14, pp. 9545–9549, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

The Meiji Revolution between 1853 and 1877 saw the overthrow of Japan's military government by the Tokugawa shoguns presents and the restoration of the imperial rule of the Emperor. At the same time it has been seen as the rapid period of modernization in Japan and its entrance into the international scene as a modern commercial state. As such, the feudal state made up of *shogun* and *daimyo*, which had existed for more than 260 years, was destroyed, and it is in this sense that a genuine revolution can be identified. The revolution occurred in three stages: the first stage, the overthrow of the Bakufu (1853–68); the second stage, the restoration of Imperial Rule (1868–71); and the third stage, the centralized nation-state (1871–77).

Meiji is the era name adopted in 1868 by the new government that overthrew Japan's previous military government of the Tokugawa shoguns and readopted the emperor as the country's sovereign. Since 1853, when Matthew C. Perry, Commander of the American Squadron to the China Seas and Japan, first put military pressure on Japan to open up its borders to foreign commerce, the country had been quickly incorporated into the world's capitalist system. For the small feudal state that Japan had been up to that moment, the problem was to find a way to survive as an independent country and to decide on the type of state that would be able to deal with world capitalism and the international relations of East Asia. We shall call the great political and social changes that were the result of the political struggle, which developed around of this problem the Meiji Revolution. The political struggle can be divided into three stages: the first stage, the overthrow of the Bakufu (1853–68); the second stage, the restoration of Imperial Rule (1868–71); and the third stage, the centralized nation-state (1871–77). The political group, which took the leadership in the last stage, was the part of the samurai class that was in the process of being transformed into the bureaucracy of the new nation-state. We shall consider the year 1877, when the last rebellion of the samurai class was suppressed, the year of its completion.

The Meiji Revolution (1853–77)

Meiji ('Enlightened Government') is the era name adopted in 1868 by the new government that overthrew Japan's previous military government of the Tokugawa *shoguns* (*bakufu*) and readopted the emperor as the country's sovereign. Since 1853, when Matthew C. Perry, Commander of the American Squadron to the China Seas and Japan, first put military pressure on Japan to open up its borders to foreign intercourse, the country had been quickly incorporated into the world's capitalist system. For the small feudal state that Japan had been up to that moment, the problem was to find a way to survive as an independent country and to decide on the type of state that would be able to deal with world capitalism and the international relations of East Asia. The great political and social changes that were the result of the political struggle, which developed around this problem, we shall call here the Meiji Revolution. The year 1877, when the last rebellion of the samurai class was suppressed, we shall consider the year of its completion. The political struggle can be divided into three stages.

The First Stage: The Overthrow of the Bakufu (1853–68)

The *bakufu's* authority was greatly damaged by the fact that it had been forced to open up the country to intercourse with foreign powers. In order to stabilize its political foundation, the *bakufu* asked the *daimyo*, or feudal lords of Japan, for their opinions and groped toward achieving some kind of consensus among the *daimyo* class as a whole. On the other hand, it wanted to clarify the legitimacy of its own national government by strengthening its connection with the emperor in Kyoto, who nominally still was Japan's monarch. Moreover, the *bakufu* began to employ military drills in the Western-style and made a start to build its own navy tutored by naval officers from Holland. However, when the news that a joint English and French squadron, which had just defeated China in the Second Opium War was about to visit Japan in 1858, the *bakufu* buckled under this military pressure to sign treaties allowing free trade with the United States, Holland, Russia, England, and France. What is more, these treaties were unequal treaties in that they included clauses on extraterritoriality and fixed tariffs that denied Japan its legal and tax autonomy. The *bakufu* sought ratification of the treaties from Emperor Komei, but the emperor refused on the ground that Japan would cease to exist as an independent state if this attitude of acquiescence to the demands of the foreigners were to persist. As a result, the discussion about Japan's existence as a nation-state became divided between those supporting the *bakufu* and those backing the position of the emperor.

As a result of the opening, in 1859, of the harbors of Yokohama, Nagasaki, and Hakodate to free trade, Japan was quickly absorbed into the world's capitalist system. Prices rose precipitously, and the markets of the country were thrown into confusion. Voices denouncing the signing of the treaties and

the *bakufu*'s policy of submission to the foreigners arose not only among the members of the ruling class, but also among well-to-do farmers and merchants, and so became a protest of the Japanese people as a whole. As a result, the *bakufu* lost confidence in its ability to rule the country. By contrast, the idea of a government that would take a hard line toward the foreigners and make the *bakufu* cooperate with that effort was gaining in popularity among those supporting Emperor Komei, who continued to refuse to ratify the treaties. Powerful *daimyo*, who, in their turn, were thinking of taking up the right to govern the country themselves, were also found supporting this idea.

Warriors of the faction that proposed radical military reforms wanted to throw out the foreigners even more badly. They were of the opinion that there was no military advantage to be gained just by superficially adopting Western-style drilling methods, while keeping the old feudal system linking military ranks with family background. They stressed that Japan should aim to modernize its army while engaged in actual battle and so they wanted to use war to accomplish internal reform. Thus, in June of 1863, radicals from the domain of Choshu first attacked foreign shipping in the Straits of Shimonoseki in order to force the shogun to take military action against the foreigners, for the country's leadership in such matters was considered to be his responsibility. However, this type of behavior by the radical warriors, which showed no regard for the feudal order, frightened Emperor Komei and the highest officials of his court. They proceeded, in September of 1863, to chase the Choshu warriors from Kyoto with the help of the warriors from the Aizu domain in northeastern Japan, which traditionally supported the *bakufu*.

On the other hand, the Great Powers of Europe and the United States, which had concluded treaties with Japan, would not allow a native movement that aimed to break away from world capitalism by military force. Therefore, they decided to complete Japan's incorporation into the world's capitalist system by dealing its radical faction some harsh blows. In August of 1863, an English naval squadron attacked the home base of the Satsuma at Kagoshima. Next, in September of 1864, a joint naval force of English, French, American, and Dutch warships attacked the strategic port of Shimonoseki in Choshu. This fief had been declared an 'Enemy of the Imperial Court' by Emperor Komei, because its soldiers had entered the capital against his will. Moreover, in November of 1865, the joint naval force of these four nations penetrated Osaka Bay, and obtained from Emperor Komei, through military pressure, the ratification of the treaties. Thus, Satsuma and Choshu both painfully experienced Western military power at first hand. They were, therefore, also the first to understand that, in order to resist foreign pressure, it would be necessary to bundle the power of the entire Japanese people through the establishment of a unified nation-state. They became convinced that, as a first step to accomplish this, they should take immediate and drastic measures to reform their military forces.

In Satsuma, the old system based on lineage was abandoned and a new homogeneous army of samurai warriors established. In Choshu, lower ranking samurai, farmers, and townsmen joined to form an armed militia or *shotai*. The *bakufu*, on the other hand, used the opportunity offered by the fact that Choshu had been named an 'Enemy of the Imperial Court' and the domain's defeat at Shimonoseki to try and reestablish its authority. It succeeded in co-opting the Court and in destroying the lateral alliance of the powerful *daimyo*. In 1866, with the sanction of the Emperor, the *bakufu* organized an expedition to punish Choshu, but the domain had just concluded a military alliance with Satsuma and with its newly organized militia it was able to defeat the army of the *bakufu*.

The Second Stage: The Restoration of Imperial Rule (1868–71)

In a *coup d'etat* of January 1868, samurai from Satsuma, convinced that it would be impossible to establish a strong nation-state under the leadership of the Tokugawa, together with Tosa warriors and members of the Imperial Court, who were concerned with the loss of prestige the Imperial House had suffered because of its recent connection with the *bakufu*, took over power and established a government based on the restoration of imperial rule (*osei fukko*). Warriors from Satsuma, Choshu (which had obtained a cancellation of the order declaring it to be an 'Enemy of the Imperial Court'), and Tosa defeated the *bakufu*'s forces near Kyoto in the same month. By means of a civil war, which continued until June of 1869, the restoration government was able, with the cooperation from the domains, to eradicate the power of the old *bakufu* and to seize the lands, which had been under its jurisdiction. The Tokugawa family was reduced to nothing more than being one of many *daimyo*. The restoration government made Emperor Meiji, who had been born in 1852 and had succeeded to the throne after the death of his father Komei in 1866, the country's new sovereign. Instead of the slogan *Sonno joi* ('Revere the Emperor and Expel the Barbarian') through which it had come to power, the new government now stressed Japan's national power with a new slogan *Bankoku Taiji* ('Holding Out Against All Nations of the World'). Also, in order to obtain support for the new government, the new leaders criticized the autocratic nature of the old *bakufu*, professed its respect for public opinion, and established a deliberative assembly the members of which were elected by the domains.

During the one-and-a-half-year long civil war, which claimed more than 10 000 lives, the idea that birth should determine social standing was abandoned in the domains of the various *daimyo*, which also Westernized their military forces and accepted the participation of rich farmers and merchants in their local governments. For these men, it was natural to expect that the army of the new Japanese nation-state would consist of the combined military forces of the domains. But to the restoration government, it seemed that trying to take a hard line against the foreign powers in this manner might spiral the power of the local samurai out of control. Conversely, the leaders of the domains attacked the weakness of the government's foreign policy. The effort to stabilize a hard-line government based on a consensus and with the cooperation of the domains was abandoned. In August of 1871, after several 1000 elite warriors from the three domains of Satsuma, Choshu, and Tosa, had gathered in Tokyo, the government took the drastic measure of abolishing the domains and establishing prefectures, saying that "The existence of the domains hurts the realization of the *Bankoku Taiji* policy,"

The Third Stage: The Centralized Nation-State (1871–77)

Therefore, the problem of establishing this was felt to be an urgent matter to be accomplished through the cooperation of the three domains mentioned above. These, together with the domain of Hizen, formed the backbone of the new government. In December of the same year, Iwakura Tomomi (the Court), Okubo Toshimichi (Satsuma), and Kido Koin (Choshu) and others were sent as the representatives of the new government on an embassy to the United States and Europe in order to revise the unequal treaties. Saigo Takamori (Satsuma) was the most important man of the caretaker government remaining behind in Japan and he forcefully pursued the legal and institutional reforms that would make the Great Powers agree to a revision of the treaties. In order to accomplish this, the armament factories and the ships built and owned for the sake of modernization by the *bakufu* and the different domains were brought under government control, students were sent overseas, and many able foreigners were employed. Ten percent of the original assessed income of their former domain was given as personal property to the *daimyo*, and the courtiers (even those of the lowest ranks) were assured an income at the same level as the former *daimyo*. The government, however, aimed at building a nation-state on the model of those in Europe and America, which required an enormous outlay of funds. For this reason it did not relieve the tax burden required of the farmers. Also, there was no agreement between the four domains on the question of how to deal with the warrior class. Even though their stipends had been guaranteed, from 1873 an army draft had been started from among all layers of the population in order to be able to suppress internal rebellions. Therefore, it became necessary to face the problem of the position of the samurai in the new society.

When the caretaker government in Japan realized that, is the short term, a revision of the treaties was out of the question, it strengthened its intention to display the energy and prestige of the new Japanese nation-state within East Asia. The Japanese government wanted to assure its prestige in an East Asian world that seemed dominated by continuing concessions to the Great Powers of Europe and the United States since the defeat of the Qing in 1842. It hoped in this way to heighten the prestige of the new nation-state at home and decided it could use the samurai class to project its military power. Already in 1871, during negotiations with the Qing the Japanese had started to build new international relationships in East Asia. When aborigines from Taiwan harmed and killed some inhabitants of the Ryukyu Islands, which traditionally paid tribute to both China and Japan, this was used as an excuse to plan a military expedition to Taiwan. Moreover, the government considered the attitude of Korea, which had refused, since 1868, to negotiate with Japan because its diplomatic form was no longer that used by the Tokugawa *bakufu*, an affront to its national prestige. For this reason, in August of 1873, the caretaker government decided to send an embassy with the message that it was ready to go to war over this issue. However, the chief ambassador of the delegation that had been sent to Europe and the United States, Iwakura, came back to Japan in September of that year and forcefully opposed this idea. He was supported by that part of the samurai class that was in the process of being transformed into the bureaucracy of the new nation-state.

Because Emperor Meiji canceled the decision to send the embassy, Saigo and Eto Shinpei (Hizen), who had been the main members of the caretaker government, resigned from the government. The Satsuma and Tosa warriors, who were living in Tokyo as part of the Imperial Household guards, seceded from this force, and in February of 1874, the samurai of Hizen rebelled under the leadership of Eto. The Meiji government, now mainly led by Iwakura and Okubo, used the warships it had collected and the newly installed nationwide telegraph network to the maximum extent possible and subdued the insurrection within a short time, thus preventing it from spreading. However, in order to win over the armed forces and its officer corps, which both still mainly consisted of samurai, the government proceeded with the invasion of Taiwan in April 1874. Ignoring China's protest that this was an infringement of its national sovereignty and showing its readiness to go to war with the Qing, the government showed that it was able to take a hard line. When the Qing recognized, at the very last moment in October 1874, that the Taiwan expedition had been a 'heroic act,' the Japanese leadership learned it need not respect the power of the Qing. On the other hand, from the attitude China had taken on the Taiwan issue, it was possible to forecast that, in case of a Japanese dispute with Korea, the Qing, even though it was officially Korea's suzerain, would not come to its rescue. Moreover, it was becoming clear that if Japan engaged in military ventures in East Asia the Western powers were likely to remain neutral.

On the national level, this success earned the government the loyalty of its officer corps and, across the board, the pace toward a system of general conscription was accelerated in order to prepare for war. For the first time, it seemed possible to replace the samurai class and establish bureaucratic control over the army. In February 1876, Japan imposed, through military force, an unequal treaty on Korea, which had kept itself isolated in a manner similar to Japan before Perry's arrival. Also, with the treaty concluded with Russia in May 1875, by which Sachalin was exchanged for the Kuril islands, the government succeeded in solving a border dispute, which had existed since the last years of the eighteenth century.

So between 1874 and 1876, the Meiji government gave permanent shape to Japan's borders and reorganized its international relations with East Asia. In this way, it put an end to the general feeling among the population of Japan, existing since the 1840s, that the state was in a crisis which might lead to its dissolution. The only problem left was the revision of the unequal treaties with the Great Powers of Europe and the United States.

With the national prestige the government harvested from these developments, it decided to embark on some radical internal reforms. These reforms were meant to create the conditions necessary to establish capitalism in Japan, which, in turn, would enable the country to respond adequately to world capitalism. One of these conditions was to revise the land tax and to establish the right to private ownership of land without diminishing the national tax income as a whole, and the government meant to accomplish this within 1876. Another was to convert the stipends of the former *daimyo*, the courtiers, and the samurai into government bonds

(August 1876). This latter reform was no rejection of feudalism, but rather a decision to change it into a tool for the adoption of capitalism.

The land tax reform called forth a movement that opposed the fixation of the taxes at the high levels of feudalism. This opposition was led by rich farmers, and peasant rebellions exploded, one after another, in virtually all regions of Japan. At the same time, from October 1876, a revolt by samurai broke out that reached its peak with the Satsuma rebellion of February 1877. In January of that year, therefore, the government lowered the land tax from 3% to 2.5% of the land value and also diminished the local taxes. This drove a wedge between the peasant movement and the samurai rebellion and allowed the government, after a violent struggle lasting 8 months, to defeat, with its conscription army of 60 000 men, the 40 000 samurai of Satsuma. As a result, the movement to incorporate the samurai class into the structure of the nation-state was extinguished and the power of the new state was no longer challenged.

However, after its defeat, the samurai class held the government to its original promise to respect public opinion and it became the foremost power in the struggle for the opening of the National Diet to all classes of Japanese society. In this struggle, the samurai, allied with the rich farmers and merchants who stressed that there could be no taxation without representation, opposed the increasingly authoritarian Meiji government. And so the struggle, which had centered on the manner in which national power was to be wielded, now transformed itself into one for democratic rights for everyone.

Historiography of the Meiji Revolution

These events were, until the second decade of the twentieth century, considered to be the epochal restoration of direct imperial rule, accomplished by the imperial house with the reverent support of the Japanese people. When Japanese capitalism reached a dead end during the 1920s, with its repressive system at home and its increasingly aggressive policies abroad, Japanese social scientists, and Marxists, in particular, started to link the unholy trinity of monopoly capital dominating the country's political system, its half-feudal land ownership, and its imperially indoctrinated army and bureaucratic leadership with these events that were now known as the Meiji Restoration. This train of thought divided itself into two schools, one maintaining that the Meiji Restoration was an incomplete bourgeois revolution and the other that it was nothing but the establishment of absolutism. After World War II, and especially during the 1960s, the prewar interpretations were considered to be flawed because they neglected the concrete facts of the international situation around Japan at the time of the revolution and had been narrowly focused on the events as one stage in the development of the Japanese nation by itself. According to this thinking, foreign pressure in the middle of the nineteenth century forced the realization that if Japan wanted to keep its independence, it needed to cooperate in achieving a thorough overhaul of the state. Because a bourgeois revolution was not necessarily an indispensable ingredient for the success of the capitalist system transplanted to Japan in the second half of the nineteenth century, the revolution compromised and merged with the feudal elements of Japanese society.

However, today the positions are divided as follows. One group of historians stresses the fact that the transition to modern society was comparatively easy because of the high levels of social and economic development in Japanese society during the second half of the Edo period (1600–1867). Another group of historians considers the Meiji Restoration a popular or a nationalist revolution, ignoring such problems as feudalism, class, and the status of power holders. While recognizing a certain level of development in premodern Japan, a third group maintains that the Meiji Restoration was the formation of a centralized imperial state, which aimed to establish capitalism on a basis of support for the old nobility and the *daimyo* class of Tokugawa society. During this process of introducing world capitalism into Japan, the state itself went through several transformations.

Issues of the Debate on the Meiji Revolution

The Meiji Revolution presents, by its very nature, a complicated problem. On the one hand, it should be called a 'restoration of imperial rule' because of the efforts that were made to revive the structure of the imperial state that had centralized power in Japan more than 1200 years earlier. Until 1945, the position of the emperor as the supreme wielder of power in the modern Japanese state seemed to be fixed, and this in itself continued to confront and suppress all political thought stressing the sovereignty of the Japanese people. On the other hand, in the sense that the feudal state made up of *shogun* and *daimyo*, which had existed for more than 260 years, was destroyed, it can be called a sort of revolution. This Asian-style nation-state, which had existed since the beginning of the seventeenth century, could not have gradually reorganized and modernized itself from above without the overwhelming military and economic pressures from world capitalism of the second half of the nineteenth century.

In Japan, the men responsible for these revolutionary changes did not come from the bourgeoisie as in the case of Western-style revolutions. During Japan's closed country policy, a new class of rich farmers and merchants had, to a certain degree, emerged with the formation of a national market that involved the whole country, but not the outside world. Their nationalism formed the basis of and fueled the changes that took place. However, the leaders of the Meiji period were a group of men, who had their roots in the warrior class and had changed themselves into a literary and military bureaucracy while treating the state as their private property. By 1877, they had gained complete control of the Japanese state with the emperor at its head. However, without the support of the rich farmers, merchants, and ex-samurai, political and social stability still eluded them. This had to wait until the 1890s, when a constitution had been put in place and a parliament was established so that other groups might have some say in the running of the government. In the very broadest sense, Japan's Meiji Revolution ranks with Germany, Italy, and Russia as one of the latecomers in the drive toward modernization. Further generalizing seems to be meaningless, and this author would like to stress that the revolution creating

a nation-state in response to foreign pressure was probably the result of the specific circumstances existing in the second half of the nineteenth century in East Asia.

See also: East Asian Studies: Politics; Ethnic Cleansing, History of; Feudalism; Imperialism, History of; Japan: Sociocultural Aspects; Modernization and Modernity in History; Nationalism, Historical Aspects of: East Asia; Nations and Nation-States in History; Parliaments, History of; Racism, History of; Revolutions, History of.

Bibliography

Akamatsu, P., 2010. Meiji 1868: Revolution and Counter-Revolution in Japan. Routledge, London.

Beasley, W.G., 1972. The Meiji Restoration. Standard University Press, Stanford, CA.

Blomström, M., La, C.S.J., 2006. Institutional Change in Japan. Routledge, London.

Hane, M., 2001. Modern Japan: A Historical Survey. Westview Press, Boulder, CO.

Jansen, M.B., 1961. Sakamoto Ryoma and the Meiji Restoration. Princeton University Press, Princeton.

Jansen, M.B., 2000. The Making of Modern Japan. Belknap Press of Harvard University Press, Cambridge, MA.

Norman, E.H., 1940. Japan's Emergence as a Modern State: Political and Economic Problems of the Meiji Period. International Secretariat of the Institute of Pacific Relations, New York.

Swale, A., 2009. The Meiji Restoration: Monarchism, Mass Communication and Conservative Revolution. Palgrave Macmillan, Basingstoke, UK.

Tsutsui, W.M., 2007. A Companion to Japanese History. Blackwell Pub, Malden, MA.

Walthall, H., 1999. The Weak Body of a Useless Woman: Matsuo Taseko and the Meiji Restoration. University of Chicago Press, Chicago.

Waters, N.L., 1983. Japan's Local Pragmatists: The Transition from Bakumatsu to Meiji in the Kawasaki Region. Council on East Asian Studies, Harvard University, Cambridge, MA.

Meltzoff's "Like Me" Theory

Daina Crafa, Douglas Mental Health Institute, McGill University, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The 'Like Me' theory, proposed by American psychologist Andrew Meltzoff, suggests that infants understand the behaviors of others through their own body movements. This process, called self–other mapping, allows infants to learn new actions and to interpret the intentions of others. This system of representation appears to be innate and has implications for human social development and empathy as well as related neural processes, such as mirror neurons, and disorders, such as autism.

Theoretical Background

The 'Like Me' theory was proposed by contemporary American psychologist Andrew Meltzoff as a description of human social learning. This theory was developed based on a series of experiments, which demonstrated that infants represent the actions of others as analogous to their own actions. These experiments also show that infants assume these analogous actions accompanied by analogous psychological states, including similar perceptions, emotions, and intentions (Meltzoff, 2007a).

The 'Like Me' theory suggests that the actions of others are cross-modally mapped onto the infants' own bodily movements (Meltzoff, 2007b), meaning that when the actions of others are observed they are represented to the infant through the infants' *body schema* (Gallagher and Meltzoff, 1996). According to the 'Like Me' theory, this process, which is commonly termed *self–other mapping*, acts as the foundation for interpreting others. Through self–other mapping of actions, accompanying psychological states are also consequently mapped. Through mapping the actions and psychological states of others onto oneself, the infant gains access to interpreting these actions and states (Meltzoff, 2007b).

The interpretation of these actions and states are considered to be *supramodal*, meaning that the observation and execution of actions share a common framework. In other words, mapping the actions of the other onto the body schema of the self gives the infant the ability not only to interpret these actions but also to execute them. To put this idea even more simply, self–other mapping gives the infant the ability to simultaneously interpret and imitate the actions of others. Through these abilities, infants understand other people as being like themselves.

This 'Like Me' understanding of others is theorized to give rise to more complex processes, such as social learning and empathy. For example, as earlier research by Meltzoff (1995) demonstrates, children as young as 18 months old are able to infer an adult's intended actions by watching them make failed attempts. After viewing failed attempts, children were observed pursuing the intended goal by taking other actions rather than imitating the adult's mistakes. In contrast, when adults successfully accomplished their intended goal, the children imitated the observed actions. Observation combined with the 'Like Me' understanding of others essentially allows infants to learn possible outcomes and consequences of their own actions by watching others.

Such findings support one of the central claims of the 'Like Me' theory: that cross-modal mapping of the action of others drives child development and social learning, at least in part. Through social observation, children can use a combination of action imitation and modification to pursue successful achievement of the intentions of others (and presumably also of themselves). Invocation of the intentions of others is thought to be an early reflection of the human drive to compete and is proposed as propelling human development by motivating the acquisition of new skills (Meltzoff, 2007a).

Meltzoff proposes that the 'Like Me' framework is already present the first day of life, corresponding with the first observed imitative acts (Meltzoff and Moore, 1977). Given its early emergence in neonates and subsequent ability to drive development, this framework is proposed as "the starting point for social cognition, not its culmination" (Meltzoff, 2007a: p. 126).

Experimental Background

Early Experiments

In the 1970s to the 1980s, the early experimental predecessors of the 'Like Me' theory focused on infant imitation during the first hours and months of life. Perhaps the most famous of these experiments is a 1977 study that Meltzoff and Moore published in *Science*, in which they showed still images of infants imitating the facial expressions of adult researchers (see **Figure 1**). In this experiment, infants aged 12–21 days saw a researcher displaying three randomly ordered facial expressions: tongue protrusion, open mouth, and lip protrusion. Infants were observed imitating the modeled behaviors, and it was then theorized that infants could model gestures supramodally. A second study by Meltzoff and Moore, published in 1983, corroborated these findings by demonstrating that infants as young as 0.7–71 h old could imitate tongue protrusion and mouth opening. The early age of these behaviors rule out possibilities that they were actions previously learned at home, and support the thesis that infants are born with the innate ability to represent and imitate the visual actions of others. This research ultimately provides a historical and theoretical foundation for the 'Like Me' theory.

Testing the 'Like Me' Theory

The 'Like Me' theory was tested from three different angles, each building on top of its predecessor (Meltzoff, 2007a). First, it was tested whether infants perceived the *actions* of others as

Figure 1 Still images of infants imitating an adult experimenter, originally published in Meltzoff and Moore, 1977. Imitation of facial and manual gestures by human neonates. Science 198, 75. (Reprinted with permission.)

equivalent to their own. Second, it was tested whether infants perceived the *perceptions* of others as equivalent to their own. Third, it was tested whether infants perceived the *actions* of others as *relevant to or informative* for their own.

Acts 'Like Me'

In the first experiment, 14-month-old infants were placed across from two adults. The two adults mimicked the infants' behaviors, either in real time or with a short delay. This created the appearance that one adult was copying the infant's behaviors, while the second adult appeared to be performing similar behaviors to the infant but not identical. Infants' behaviors were evaluated to see which adult they preferred, based on who they looked at longer and smiled at more. In line with the initial hypotheses, this study demonstrated that infants preferred the researcher who acted the most like themselves. These findings were interpreted as demonstrating that the infants have an awareness of their own actions and actions that are 'Like Me.' Implicitly, this suggests that they have a system for representation the actions of another person enough to discriminate between similar and dissimilar actions. It also demonstrates that these action representations are temporally locked to the initial actions.

Perceives 'Like Me'

In the second experiment, it was tested whether infants understood that another person needed to have an unobstructed line of sight in order to see something. To test this, infants watched an adult turn his or her head while having the eyes either open or closed. The head movements corresponded with an object movement. Infants were tested to see whether they looked or pointed at the adult or followed the adult's gaze to the object. Infants looked longer at the object and pointed more often when the adult's eyes were open. As a follow-up experiment, infants were given time to play with the blindfold or headband, and some infants were briefly blindfolded with either a normal blindfold or one that had a window cut out so that they could still see. Then, adults wore either a blindfold or a headband in their respective normal fashions and moved their head in the direction of an object. Infants who had previous experience with an opaque blindfold showed different behaviors compared to those who had not. This study demonstrated that infants who had their vision obscured by a blindfold understood that the adults wearing a blindfold could not see. Together, these two experiments show that the 'Like Me' framework is developed through experience with the external world and that the perception of others are compared to these past experiences through the 'Like Me' analogy.

Experiences 'Like Me'

In the third experiment, infants were situated so that they could eavesdrop on an interaction between two adults. One adult was given an object to manipulate, while the second adult would either respond neutrally or angrily when the first did. Later, infants were given the same object – and sometimes the second adult was looking at them during this manipulation, while other times they had their backs turned. Infants were significantly more likely to manipulate the object if the adult was not angry or was not looking. These findings demonstrate that the infants interpreted the experiences of the first adult as relevant to themselves and applicable to their current situation. It also suggests that the 'Like Me' analogy does more than drive development through action imitation and revision – it also provides informational resources that can be applied to future situations.

Related Theoretical Work

In 1996, many of the principles of the 'Like Me' theory were applied and extended in a theoretical article authored by

Shaun Gallagher, a philosopher well known for his work on embodiment theory, and Andrew Meltzoff. This collaboration examined the human developmental emergence of *body schema* – the spatial representation of the body – and *body image* – the visual representation of the body. Movement and planned actions rely on body schema, which essentially informs one about where the body is extended in space and enables one to move accordingly. In contrast, body image is how one thinks the body looks while performing these actions (e.g., fat or thin, elegant or awkward).

Previous theorists proposed that body schema does not emerge until later in life by claiming that *aplasic phantom limbs* – the sensation that congenitally missing limbs are present – does not affect young children. Moreover, this schema is not completely tied to body image, since a visibly missing limb can be felt. Applying the 'Like Me' principles, Gallagher and Meltzoff (1996) determined that body schema is fundamental for understanding more complex phenomena, such as the self–other mapping. They hold that body schema provides an individual and extended point of reference for the actions of others to be mapped onto. In this respect, early development of body schema would allow children to map the actions of others cross-modally onto their own, whereas later development would theoretically preclude this ability. Thus, they argue that early emergence of imitation less than an hour after birth requires body schema as a prerequisite. This suggests that contrary to previous theories, body schema is innate. Body image is also necessary for imitation and therefore is likely also an innate ability. The neonate must be able to understand that when they see a researcher sticking out their tongue, this visual appearance (i.e., the body image) corresponds to the act of sticking out their own tongue (i.e., body schema). Although body image may not be completely tied to body schema, these frameworks are functionally intertwined during the act of imitation.

Applications and Implications

Mirror Neurons

Mirror neurons are proposed as one possible explanation for the 'Like Me' framework (Meltzoff, 2007a). *Mirror neurons* are cells in the frontoparietal circuits that specialize in sensorimotor integration. These cells become active when a person acts as well as when a person observes others acting. These neurons are hypothesized to provide a basis for empathy as well as for self–other mapping. By responding identically to actions performed by oneself and others, they are thought to establish a biological link between the behaviors of self and other. Although they have directly been observed in macaque monkeys, they have only been indirectly observed in humans and their function remains theoretical (Iacoboni and Dapretto, 2006).

Culture

Imitation appears to play a large role in sociocultural learning and has been implicated in the emergence in cross-cultural behavioral differences (Losin et al., 2009; Wörmann et al., 2012; Crafa and Nagel, in press). These differences emerge during early childhood, with culture-specific behaviors such as level of comfort around strangers and when to say 'thank you,' emerging during infancy (Keller and Otto, 2009; Wörmann et al., 2012). As the 'Like Me' theory explains, imitation drives child development and sociocultural learning. Children learn by observing, imitating, and improving upon behaviors they see their parents and other adults performing. Cross-cultural differences in the adult behaviors are also observed by children, and are therefore integrated into the child's framework (Crafa and Nagel, in press).

Cultural imitation may be partially influenced by shared cultures and appearance. Buttelmann et al. (2013) demonstrated that people were more likely to imitate others of the same cultural and racial background. Losin et al. (2012) also showed differential brain activity when imitating own race compared with other race people. This activity appeared to correspond with reported levels of familiarity with the other race, and these findings potentially suggest that familiarity is a key factor in cultural learning.

Disorders

Imitation and self–other mapping may also explain cross-cultural differences in the symptoms and behaviors corresponding with many psychiatric disorders. Crafa and Nagel (in press) propose that patients may exhibit different behaviors across cultures because (1) they are learning different cultural behaviors and (2) they may experiencing disruptions to functions underlying imitation or self–other mapping. According to this theory, actions and psychological states are not mapped exhaustively. Instead, only actions that are analogous ('Like Me') to a person are mapped. The underlying fodder for this 'Like Me' framework may be predetermined by biological predispositions and previous social experiences. Anomalies among these predispositions – such as atypical epigenetic or neural processes – may lead to differences in the fundamental self and therefore in the 'Like Me' framework that the other is mapped onto.

Autism

Autism spectrum disorder is hypothesized to emerge from a disruption to self–other mapping and the 'Like Me' framework. This idea was first proposed by Meltzoff (2007a) as one possible implication of the 'Like Me' theory. Lombardo et al. (2010) tested this proposal by examining neural processes associated with self–other mapping. They found that neurons in self-information processing areas of the brain behaved differently in people with autism diagnoses compared with people without autism. These disruptions potentially prevent or alter the mapping of representations of others onto the self in autism.

Counter Theories

In contrast with the proposal that self–other mapping acts as a foundation for interpreting others, other theorists suggest that interpretation of others' intentionality occurs prior to any cognitive representation. Tomasello et al. (2005) claim that

shared intentionality – the ability to share actions and goals with another person – gives rise to other cognitive representations, including the interpretation of another person as being 'Like Me.' According to this theory, many animals are able to understand the intentions of others, but humans uniquely can collaborate in the execution of these intentions. This ability to collaborate comes from two abilities that converge during early childhood: (1) the understanding that others are "animate, goal-directed, and intentional agents," and (2) the "motivation to share emotions, experience, and activities with other persons" (Tomasello et al., 2005: p. 675). Through a series of experiments, Tomasello et al. (2005) demonstrate that these two streams of abilities emerge during childhood at different rates. When they converge, around the age of 14 months, infants begin to display shared intentionality. Once developed, mental representations of others can then be constructed. This theory is fundamentally opposed to the 'Like Me' theory, by claiming that representations of others emerge later in infancy.

Research Prospects

With a few exceptions, research surrounding the 'Like Me' theory has focused on behavioral evidence during early childhood and a few neuroimaging studies in adults. Despite theorized relationships with mirror neurons and other neural events, the brain processes directly relating to the 'Like Me' phenomena have not been widely studied in infants or children. 'Like Me' behaviors have also not been studied in children diagnosed with autism, and such study could potentially benefit this research area as well.

See also: Autism Spectrum Disorders; Goal-Oriented Behaviour Among Human and Non-Human Primates; Imitation During Infancy and Early Childhood; Infant Gaze Behaviors; Intentionality During Infancy and Early Childhood, Development of; Joint Attention During Infancy and Early Childhood Across Cultures; Mirror Neurons, Theory of; Neonatal Imitation; Self in Culture: Early Development; Self-Development in Childhood and Adolescence; Social Referencing During Infancy and Early Childhood across Cultures; Theory of Mind During Infancy and Early Childhood Across Cultures, Development of.

Bibliography

Buttelmann, D., Zmyj, N., Daum, M., Carpenter, M., 2013. Selective imitation of in-group over out-group members in 14-month-old infants. Child Development 84, 422–428.

Crafa, D., Nagel, S., in press. Traces of culture: the feedback loop between brain, behavior, and disorder. Transcultural Psychiatry.

Gallagher, S., Meltzoff, A.N., 1996. The earliest sense of self and others: Merleau-Ponty and recent developmental studies. Philosophical Psychology 9, 211–233.

Iacoboni, M., Dapretto, M., 2006. The mirror neuron system and the consequences of its dysfunction. Nature Reviews Neuroscience 7, 942–951.

Keller, H., Otto, H., 2009. The cultural socialization of emotion regulation during infancy. Journal of Cross-Cultural Psychology 40, 996–1011.

Lombardo, M.V., Chakrabarti, B., Bullmore, E.T., Sadek, S.A., Pasco, G., Wheelwright, S.J., et al., 2010. Atypical neural self-representation in autism. Brain 133, 611–624.

Losin, E.A.R., Dapretto, M., Iacoboni, M., 2009. Culture in the mind's mirror: how anthropology and neuroscience can inform a model of the neural substrate for cultural imitative learning. Progress in Brain Research 178, 175–190.

Losin, E.A., Iacoboni, M., Martin, A., Cross, K.A., Dapretto, M., 2012. Race modulates neural activity during imitation. Neuroimage 59, 3594–3603.

Meltzoff, A.N., 2007a. "Like Me": a foundation for social cognition. Developmental Science 10, 126–134.

Meltzoff, A.N., 2007b. The "Like Me" framework for recognizing and becoming an intentional agent. Acta Psychologica 124, 26–43.

Meltzoff, A.N., 1995. Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. Developmental Psychology 31, 838–850.

Meltzoff, A.N., Moore, M.K., 1983. Newborn infants imitate adult facial gestures. Child Development 54, 702–709.

Meltzoff, A.N., Moore, M.K., 1977. Imitation of facial and manual gestures by human neonates. Science 198, 74–78.

Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H., 2005. Understanding and sharing intentions: the origins of cultural cognition. Behavioral and Brain Sciences 28, 675–735.

Wörmann, V., Holodynski, M., Kärtner, J., Keller, H., 2012. A cross-cultural comparison of the development of the social smile: a longitudinal study of maternal and infant imitation in 6- and 12-week-old infants. Infant Behavior and Development 35, 335–347.

Memmi, Albert (1920–)

Judith Roumani, Writer and Editor, *Sephardic Horizons*, Potomac, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Born in Tunis in 1920, into a poor Jewish family, Memmi participated in the world of the colonizer and that of the colonized, a member of a subject minority within a colonized country. Memmi experienced a Nazi labor camp and worked for postwar Tunisian independence. His fiction and sociological essays extrapolate and theorize on the basis of his lived experience. Best known for *Pillar of Salt* (1953) and *The Colonizer and the Colonized* (1957), he has been a tireless advocate for the oppressed and analyst of personal and collective situations of dependency, domination, and racism.

Biographical Facts

Albert Memmi was born in 1920 in the city of Tunis, when Tunisia was a French Protectorate, into a Jewish family that was relatively poor, living (he implies) in a small alley off the slightly larger streets and alleys of the Jewish quarter of Tunis, *la Hara*, on the border of this neighborhood and those of other ethnic groups. Little remains of this quarter today as it was demolished under the guise of slum clearance in the 1950s. His father, who owned a small leather-working shop and with one assistant made harnesses and shoes, was from a family that might have had some Italian Jewish connections; and his mother came from a humble Jewish family that had always been in Tunisia. In other words, his father may have come from the *gorni* or Livornese, while his mother was from the *touansa* or indigenous Jews, two groups that hardly mixed until the twentieth century. The family spoke the Judeo-Arabic of Tunisia and Memmi's accounts of his childhood emphasize the relative deprivation he grew up in. For example (as recounted by Alexandre, the main character in his first novel), though he had to wear hand-me-down clothes, when he grew out of them they were passed on in turn – minus the buttons – to an even more deprived younger relative or neighbor. He recounts not having pocket money, not having enough lunch and snack money for school. Once he lent his two pennies for lunch to a rich boy who spent them on a chocolate bar that he then threw away, to Alexandre's (and obviously Albert Memmi's own) disgust and jealousy. Since the Muslim world was already in flux, the family also lived in fear of attacks on the *Hara* or Jewish quarter by neighboring Arabs. He attended the local Alliance Israélite Universelle school, where he picked up French and had the good fortune to win a scholarship to attend the prestigious Lycée Carnot high school of Tunis. There he had to confront profound cultural contradictions; for though inclined to study philosophy as one of its top students, as a sensitive teenager he was constantly wounded by the wealthy and non-Jewish students' lack of understanding of his position, poised uncomfortably between three civilizations: the North African/Middle Eastern, the Jewish, and the French.

Following the *lycée*, Memmi obtained an assistantship at the school in order to study philosophy in a correspondence course from the University of Algiers. During the war, under Nazi occupation between November 1942 and May 1943, like other male Tunisian Jews, he was rounded up and sent to a labor camp on the front where the Jews were forced to dig trenches against the Allied advance, under very dangerous circumstances. His novelized memoir, *La Statue de sel* (*The Pillar of Salt*), recounts a dramatic escape and trek to safety with a group of other prisoners and an attempt to join the Free French army. Though initially they did not want Jews, it seems that Memmi did eventually succeed in joining.

His later career is also rife with contradictions, which provide much material for the principled and reasoned stances he took in his writings and in his life. After leaving for Paris and the Sorbonne and graduating in philosophy in 1949, he returned to Tunis as an instructor at the technical college, together with his non-Jewish, French wife. His first novel, published in Paris in 1953, likewise caused a *succès de scandale* for its uncompromisingly realistic portrayal of Tunisian life. He also opened in Tunis an institute of psychology, whose patients must have provided him with much material for his subsequent novels and analytical works. Memmi actively supported Tunisian independence and was one of the founders of a weekly magazine, which was published "au service de la lutte nationale" ("in the service of the national struggle"), originally named *L'Action* (Action) and later *Jeune Afrique* (Young Africa).

After Tunisian independence, though, like many Jewish professionals, Memmi began to feel out of place in the new nation, which was attempting to define itself as a Muslim state; in 1956 at the age of about 36 (there has been some controversy as to exactly when he left, either in 1956 or 1957, but Memmi has informed me that he left in 1956. See my *Albert Memmi*, p. 42) saying that he felt his work in and for Tunisia was done, left again for Paris. A period of intense work began as he joined the Centre National de Recherches Scientifiques, and at the Ecole Pratique des Hautes Etudes headed a team of researchers. In 1970 he became a Professor of Sociology at the University of Nanterre; he has been associated with the Centre Universitaire d'Etudes Juives, was Vice-President of the Pen Club of France, and Director of the Centre de Recherches et d'Etudes sur la Dépendance. Thus, in combining philosophy, sociology, and psychology, as well as earning a reputation as a novelist, Memmi has always melded several disciplines in his search for truth. His commitment to Tunisia did not end with his second departure, and in 1984 he was awarded the highest honor of the Tunisian state, the Prix de Carthage, awarded personally by President Bourguiba himself. Memmi also supports idea that the Jewish people have a right

to self-determination (he describes himself as a "Zionist of the Left") and has visited Israel several times, a stance that probably has cost him some friends. In recent years he has traveled less, due to his late wife's illness, but in 2010 he visited Morocco to preside over a major conference on Jewish and Muslim interactions in the Maghreb, 'Migrations, identité et modernité au Maghreb.' Now in his nineties, despite recent bouts with illness and an accident, he continues to participate actively in the world of ideas. Memmi's experience has spanned the major traumas of twentieth century history (from anti-Semitism and the Holocaust to the struggle of a colonized nation for independence) leading him to diagnose the main sources of sociological and psychological distress and take courageous positions in favor of the oppressed, the colonized, the dominated, and those subject to racism.

Description of Major Works

Memmi's work naturally falls into two main groups, his fiction and poetry on one hand, and his sociological and philosophical writings, on the other. Despite the diversity of genres, his work as a whole has had a strong coherence over the years. It has reflected his opposition to the negative effects of colonialism, oppression, and racism and his support of attempts to achieve wholeness in the face of the fragmenting forces of cultural, political, and psychological schisms. He has been a builder of bridges, both within the individual, among individuals, and between individuals and social realities. He opened French minds to the idea that the French language could convey an entirely different mentality, world view, and aesthetic, those of the Maghreb. Always skeptical of received ideas and pieties, in later years through aphoristic writings, he advocated the view that personal happiness was nevertheless possible. These threads can be found running through the entire body of his fiction, poetry, and philosophical and sociological writing.

Since Memmi's first published book was a novel, we begin with his fiction. Writing *La Statue de sel* (*The Pillar of Salt*, 1953) was a psychological imperative for its author, who thought of it as a matter of survival. Writing was, for the young man who had just emerged from a forced labor camp and left for France, "a means of exorcism" (Yetiv, *L'Aliénation*: p. 160). Memmi believed that literature had a therapeutic function: by confessing, the writer frees himself from his obsessions. In Memmi's first novel, confession is as important as the aesthetic and the imaginary. However, as I have pointed out elsewhere, what appears to be autobiography in his novels sometimes has to be taken with a grain of salt, as the author deliberately does not tell us where biography ends and fiction begins. (See my 'Memmi's Introduction to History: *Le Désert* as Folktale, Chronicle and Biography,' *Philological Quarterly* 61:2 (Spring 1982), 193–207, esp. note 7, p. 205. Also Lia Nicole Brozgal, *Against Autobiography: Albert Memmi and the Production of Theory*, Univ. of Nebraska Press, Lincoln, 2013.) The success of Memmi's first novel was followed by a number of others, most of them also set in the Tunis of his young adulthood, such as *Agar* (Hagar, translated as *Strangers*, 1955); *Le Scorpion ou la Confession imaginaire* (*The Scorpion or the Imaginary Confession*, 1969); *Le Désert ou la Vie et les aventures de Jubaïr Ouali El-Mammi* (*The Desert or the Life and Adventures of Jubaïr Ouali El-Mammi*, 1977), set in medieval times, as the autobiography of an imaginary ancestor; *Le Pharaon* (*The Pharaoh*, 1988), set in postcolonial Tunis; and *Nomade immobile* (The Immobile Nomad, 2000). Memmi published one essay on literary theory, advocating different-colored typefaces for different voices within a novel: *L'Ecriture colorée, ou Je vous aime en rouge* (*Writing in Colors, or I Love you in Red*, 1986). His novels were never published in colored typefaces but he did achieve different typefaces for different voices or personae in *Le Scorpion* and *Le Désert*. He has also published short fiction, such as *Térésa et autres femmes* (2004) (Teresa and Other Women, 2006), and poems in various journals and one volume, *Le Mirliton du ciel* (*The Reed Pipe of the Sky*, 1985). These poems are very personal, with simple and homely language, relating on the one hand to relationships with old friends in Tunis and the traditional cycle of the Jewish year, and on the other hand including a number of love poems.

As he once admitted a little regretfully, he has been better known as an incisive essayist than as a belletrist. Memmi's early sociological writings brought him some notoriety in an age when the major colonial powers, particularly France, were still reluctant to set free their colonized subjects. His first nonfictional work, *Le Portrait du colonisé précédé du Portrait du colonisateur* (*Portrait of the Colonized Preceded by Portrait of the Colonizer*, 1957) became as well known as *La Statue de sel*. It was one of the first books to portray the psychological malaise of the colonized; over the years Memmi's name was linked with those of Octave Mannoni, Frantz Fanon, Aimé Césaire, and Jean-Paul Sartre, as one of the principal theorists of the colonial situation and as an advocate of decolonization. Memmi's theoretical writing in the field predated that of Fanon, since the latter's *Les Damnés de la terre* (*The Wretched of the Earth*) did not appear until 1961 (Fanon had previously published *Peaux noires, masques blancs* (Paris: Seuil, 1952). Memmi criticized Fanon's hypothesis of the *necessity* of violence as a purifying mechanism for new nations achieving independence. See '*La Vie impossible de Frantz Fanon*,' *Esprit* 9 (Sept. 1971, 248–273).

In his *Portrait du colonisé précédé du Portrait du colonisateur* (in the English edition, *The Colonizer and the Colonized*) Memmi described with great vividness the meaning of the negative image imposed on the colonized individual by the colonizer. Like the existentialists, he believed in the necessity of portraying concrete situations, and therefore concentrated on the experience of the individual. Lacking independent national institutions, the colonized man became simply the negative aspect of all the positive characteristics of the colonizer (himself deformed by the relationship) and the colonized man came to see himself in this same negative image. The colonized man ultimately attempts a 'recovery of self,' which is the refusal of the image imposed on him by the colonizer (p. 156) and eventually passionately rebels against the colonizer's culture. Such ideas, which were original and revolutionary in the 1950s, today have entered the general currency. Memmi followed this with several works in the vein of 'portraits,' giving us *Portrait d'un juif* (Portrait of a Jew, 1962) and his much later *Portrait du décolonisé arabo-musulman et de quelques autres* (Portrait of the Arab-Muslim Decolonized and of Several Others), titled in the English edition, *Decolonization and the Decolonized* (2004a). He described his method as a 'back-and-forth' between

a concrete situation that he knows intimately, where he is on sure ground having often lived it himself, and a generalizing hypothesis, a stylization that he constantly compares with the lived experience. Other writings, such as *L'Homme Dominé* (*Dominated Man*, 1968), describe the dominated and the dominators, in relations between masters and servants or men and women; and relations of dependency e.g., in *La Dépendance* (*Dependence*, 1979) either between two people who are dependent on each other (as in a marriage) or an individual's dependency on some perhaps unhealthy habit, such as drinking or smoking e.g., *Le Buveur et l'amoureux* (The Drinker and the Lover, 1998). In relations between Jews and Arabs, Memmi analyzed the psychologically unhealthy relationship of 'dhimmitude' (the subjection of a religious minority under Islam) and the need for Jews to have their own independent state in order to break this relationship, in *Juifs et arabes* (*Jews and Arabs*, 1974). Thus psychology and sociology have merged in his work. He has also analyzed racism and its manifestations, coining the term 'heterophobia' (hatred or fear of someone who is different). See *Le Racisme: Description, definition, traitement* (*Racism: Description, Definition, Treatment*, 1982). In later years, Memmi transitioned to a less critical stance, largely emphasizing the positive. He published three books on happiness, all in the 1990s; a *Dictionnaire critique à l'usage des incrédules* (*Critical Dictionary for the Use of Non-Believers*, 2000); his most recent book is his 'ethical will' *Testament insolent* (*Insolent Testament*, 2009).

Discussion of His Lasting Impact

Memmi's startlingly original ideas, emerging from painful personal experience and an uncompromising philosophical search for truth, have over the decades entered so fully into the mainstream of accepted thought in several fields that today they no longer strike us with their newness: it takes an effort to appreciate their originality. He was one of the first, together with Frantz Fanon, to diagnose the psychological malaise of the colonized as simultaneous with that of the colonizer (1957). In 1993, Edmond Jouve described the impact of Memmi's work when it appeared over 30 years before.: "Nous étions jeunes, et l'Algérie était en guerre. En vain, nous cherchions à comprendre. Et voice que … un nom s'inscrivit dans notre ciel: celui d'Albert Memmi. Et qu'un livre imposa à nous, qui nous délivra de nos peurs: *Portrait du colonisé*" ("We were young and Algeria was at war. We were trying in vain to understand. And then a name appeared: Albert Memmi, and a book came to us and delivered us of our fears: the *Portrait of the Colonized*.") (Jouve: p. 7). Memmi followed up his experience of anti-Semitism during the war years with one of his penetrating psychological portraits, the *Portrait d'un Juif* (*Portrait of a Jew*, 1962), which besides reflecting his lived experience is said to be influenced by Sartre's writing on anti-Semitism; this work was followed by the *Libération du juif* (*Liberation of the Jew*, 1966). In the field of Middle East studies or Maghrebian studies, the hypothesis of cultural fusion between Jews and Arabs originated with Memmi, as he described himself as an 'Arab-Jew' in *Juifs et arabes* (*Jews and Arabs*, 1974). He has been one of the few Jewish intellectuals who have been nevertheless welcome in some Maghrebian countries, such as Tunisia and Morocco; his earning the highest award of the Tunisian state in 1984 attests to a certain acceptance. Today, the widely used term 'Arab-Jew' is often applied to Jews who were obliged to leave Arab countries upon independence, but who still have some residual affinity or nostalgia for the culture of their countries of birth.

Recent approaches have valued Memmi's contributions as a theorist more than as a committed participant in sociological debates.

Memmi's contribution to the field of studies in racism is incontrovertible. Not long after arriving in Paris for the second time, he collaborated on a team effort, based on a questionnaire, and resulting in a book, *Les Français et le racisme* (*The French and Racism*, 1965). This book and a later one (*Le racisme*, 1982) also attempted a theory of racism which has informed similar studies to this day. Memmi was the originator of the term 'heterophobia' which covers racism and a broader range of the fear of those who are different from us in various ways. His *Dictionnaire critique* (Critical Dictionary, 2002) defines the term, the origins, and the psychology of racism. To note a difference is not a sign of racism, unless it is done in a negative way, he says. His brief definition is "racism is the general and definitive valorization of real or imaginary differences, to the advantage of the accuser and to the detriment of the victim, in order to legitimize aggression or privilege" (p. 285). Memmi sees the origin of racism in a fear of those who are different, causing preventive aggression. He also notes that racism and heterophobia are all the more tenacious because, as well as excluding the other, they also serve to reinforce our own membership in a particular group (p. 286). One can trace the influence of Memmi's thought (as well as of other thinkers, of course) in today's legal and institutional emphasis on the importance of encouraging tolerance of diversity in the workplace and on campus still sorely in short supply in both Western and non-Western countries.

The theory of dependency is also indebted to Memmi's thought. From *L'Homme dominé* (*Dominated Man*, 1973) through *La Dépendance* (*Dependence*, 1979) and to later works such as *Le Buveur et l'amoureux* (*The Drinker and the Lover*, 1998), Memmi elaborates on his definition of dependency as a constraining relationship (real or ideal) with another person, an object, a group or an institution, which the dependent individual more or less accepts, and which meets a need of the dependent person (my paraphrase from *La Dépendance* (Paris: Gallimard, 1979), p. 212). The concept of dependency as something undesirable has mellowed in his later writing where he discusses the need for mutual dependency within the couple. His later writings on happiness, often aphorisms, emphasize consideration and compromise toward those we love. Like many rational philosophers, Michel de Montaigne in particular, Memmi has tended, especially in his later works, to advocate the 'golden mean,' and moderation in all things, as a form of practical wisdom. Mutual dependency is thus linked to the idea of moderation.

Perhaps Memmi is still best known for his theory of colonialism – the toxic relationship which deforms both sides (the colonizer and the colonized, both those holding power and those who lack power) culturally, psychologically, and socially. The book had broad influence in the Third World, and even among 'dependent peoples' in Europe (Jouve: p. 9). Memmi extrapolated from his personal experience as a member of

a minority subjected by both the colonizers and the colonized in Tunisia, to a systematic analysis of the position of oppressed peoples in colonized countries – his book appeared at a time when many colonized peoples were achieving or struggling toward independence. He extended his well-known analysis of 1957 in a much later (2004) discussion of the 'decolonized,' depicting how the formerly colonized, even many years later, are still tainted by the original imbalanced and exploitative relationship. He described links within the new nations between poverty, corruption, and tyranny that may have offended some of his former allies. He also identified population shifts (migrations) and culture clashes between new and old nations, leading to a turn to political and religious myths, fundamentalism, and violence. As always, Memmi continued to advocate humanism and moderation in these struggles for identity. This work was received less enthusiastically than usual, due perhaps to readers' misapprehensions of what he was trying to say.

Memmi's thought has also had some influence in the area of language. From an earlier prediction of the death of the French language in new nations that France had formerly colonized, he came to recognize by the 1980s as a practical matter, the existence of the linguistic community of francophonie (Francophonia), which recognized that the French language no longer depended on the hegemony of France. His own work exemplifies this. Besides bringing a new reality to the literary consciousness of the metropolis from the 1950s onward through his fiction, he also edited a series of anthologies of francophone literature of the Maghreb. He has been recognized for this role with the Grand Prix de la Francophonie in 2004, awarded by the Académie Française for his work as a whole.

Memmi has also engaged in literary experiment in several of his novels. *Le Scorpion* has many layers and many voices, and Memmi's original plan, to differentiate between them by using different-colored printing, has been replaced by a series of different typefaces. However, he took up the subject again in a later volume entitled *L'écriture colorée*. This would seem to be an excellent idea for novels with multiple voices, such as *Le Scorpion* and *Le Désert*. He also favors frame stories, such as the discovery of manuscripts in the drawer belonging to the missing novelist, in *Le Scorpion*, or the aged prince telling his life story to Tamerlane in *Le Désert*: these devices constitute a modern version of the frame story of the *Thousand and One Nights*. The ambiguity between fictional narrator and the author's autobiography, which has often confused readers and critics of Memmi's fiction, also undermines the apparent verisimilitude of his novels and is decidedly disconcerting in a deliberately modern or postmodern way.

As a modern, engaged, taboo-breaking, secular writer, uncompromising in his willingness to discuss uncomfortable truths, Memmi could hardly be expected to be a supporter of tradition. Yet he has championed the traditional oral culture of the environment that he grew up in, the humorous tales of Djoha, tales of epic heroes, moral tales, folk wisdom, proverbs, a culture that was shared by the poorer Jews and Arabs in Tunisia. The multilayered novel *Le Scorpion* includes some references to Kabbalah. His novel *Le Désert*, set in medieval North Africa, the story of a deposed prince from a Jewish kingdom in southern Morocco, a fictional ancestor of the author, embodies many of these traditional elements and a kind of balanced and moderated wisdom that owes something to European philosophers and something to the age-old values of Tunisia that he grew up with. The structure of the novel resembles the frame story of Arabic literature. On his writing desk Memmi keeps several objects that remind him of traditional life in Tunisia: a piece of leather (recalling for him his artisan father's profession as a saddlemaker), scented amber, a Roman coin from Carthage. These he calls his *petite patrie portative* (little portable homeland), the objects that link him to his origins and inspire him to go forward to the ends of the world.

Select Bibliography

Several of Memmi's books in English translation are now available in eBook editions from Plunkett Lake Press, www.plunkettlakepress.com.

See also: Colonization and Colonialism, History of; Diversity in Organizations; Happiness: History of the Concept; Migration History; National Liberation Movements; Near Middle East/North Africa Studies: Culture; Postcoloniality; Prejudice in Society: Psychological Perspectives; Prejudice in Society: Sociological Perspectives; Racism, History of; Racism: Social Psychological Perspectives; Social Dominance Orientation.

Bibliography

Primary Sources

Memmi, A., 1953. La Statue de sel. Corrêa, Paris. (E. Roditi, Trans., 1992). The Pillar of Salt. Beacon, Boston.
Memmi, A., 1955. Agar. Rpt. Gallimard, Paris. (B. Rhys, Trans., 1984/1960). Strangers. Orion, New York.
Memmi, A., 1957. Portrait du colonisé precede du portrait du colonisateur. Pref. Jean-Paul Sartre. 1973. Payot, Paris. (H. Greenfield, Trans., 1995). The Colonizer and the Colonized. Beacon, Boston.
Memmi, A., 1962. Portrait d'un juif. Gallimard, Paris. (E. Abbot, Trans., 1962). Portrait of a Jew. Orion, New York.
Memmi, A., 1964. La Double leçon de Freud. In: Bakan, D. (Ed.), Freud et la tradition mystique juive. Payot, Paris, pp. 249–262. Included in 1968, 1973. L'Homme dominé. Payot, Paris. Dominated Man: Notes toward a Portrait. 1968. (E. Levieux, Trans). Orion, New York.
Memmi, A., 1965. Les Français et le racism: Enquête en collaboration. Payot, Paris.
Memmi, A., 1966. La Libération du juif. Gallimard, Paris. (J. Hyun, Trans., 1966). The Liberation of the Jew. Orion, New York.
Memmi, A., 1969. Le Scorpion ou la confession imaginaire. Gallimard, Paris. (E. Levieux, Trans., 1971). The Scorpion. Grossman, New York.
Memmi, A., 1971. La Vie impossible de Frantz Fanon. Esprit 39 (406), 248–273.
Memmi, A., 1974. Le Personnage de Jeha dans la litterature orale des arabes et des juifs. Ben Zvi Institute, Jerusalem.
Memmi, A., 1974. Juifs et arabes. Gallimard, Paris. Eleanor Levieux, Trans. 1975. Jews and Arabs. J.P. O'Hara, Chicago.
Memmi, A., 1975. Un Entretien avec Robert Davies. L'Etincelle, Quebec.
Memmi, A., 1976. La Terre intérieure: Entretien avec Victor Malka. Gallimard, Paris.
Memmi, A., 1977. Le Désert, ou La Vie et les aventures de Jubaïr Ouali El-Mammi. Gallimard, Paris.
Memmi, A., 1979. La Dépendance: Esquisse pour un portrait du dépendant. Gallimard, Paris. (P. Facey, Trans., 1984). Dependence. Beacon, Boston.
Memmi, A., 1982. Le Racisme: Description, définition, traitement. Gallimard, Paris. (S. Martinot, Trans., 2000). Racism. University of Minnesota Press, Minneapolis.

Memmi, A., 1985. Ce que je crois. Grasset, Paris.
Memmi, A., 1986. L'Ecriture coloré, ou Je vous aime en rouge: Essai sur une dimension nouvelle de l'écriture, la couleur. Périple, Paris.
Memmi, A., 1988. Le Pharaon: Roman. Julliard, Paris.
Memmi, A., 1992. Bonheurs: 52 semaines. Arléa, Paris.
Memmi, A., 1998. Le Buveur et l'amoureux. Arléa, Paris.
Memmi, A., 2000. Le nomade immobile: récit. Arléa, Paris.
Memmi, A., 2002. Dictionnaire critique a l'usage des incredules. Felin, Paris.
Memmi, A., 2004a. Portrait du décolonisé arabo-musulmane et de quelques autres. Gallimard, Paris. (R. Bononno, Trans., 2006). Decolonization and the decolonized. University of Minnesota Press, Minneapolis.
Memmi, A., 2004b. Térésa et autres femmes: récits. Félin, Paris.
Memmi, A., 2009. Testament insolent. Odile Jacob, Paris.

Secondary Sources

Brozgal, L.N., 2013. Against Autobiography: Albert Memmi and the Production of Theory. University of Nebraska Press, Lincoln.
Dugas, G., 1984. Albert Memmi: Ecrivain de la déchirure. Naaman, Sherbrooke.
Dugas, G., 2001. Albert Memmi: Du Malheur d'être juif au bonheur sépharade. Alliance Israélite Universelle, Paris.
El Houssi, M., 2004. Albert Memmi: L'Aveu, le plaidoyer. Bulzoni, Rome.
Jouve, E. (Ed.), 1993. Albert Memmi: Prophète de la decolonization. Agence de cooperation culturelle et technique, Paris.
Ohana, D., Sitbon, C., Mendelson, D., 2002. Lire Albert Memmi: déracinement, exil, identité. Editions Factuelles, Paris.
Roumani, J., 1982. Memmi's introduction to history: Le Désert as Folktale, Chronicle and Biography. Philological Quarterly 61 (2), 193–207.
Roumani, J., 1987. Albert Memmi. CELFAN Editions, Philadelphia.
Strike, J., 2003. Albert Memmi: Autobiographie et autographie. L'Harmattan, Paris.
Toumi, A.B., 2002. Maghreb divers: Langue française, langues parlées, littératures et représentations des maghrébins a partir d'Albert Memmi et de Kateb Yacine. Peter Lang, New York.
Yetiv, I., 1972. Le Thème de l'aliénation dans le roman maghrébin d'expression française, 1952–1956. CELEF, Sherbrooke.

Relevant Websites

www.examiner.com/article/issues-of-identity-albertmemmi – Examiner.com.
http://quod.lib.umich.edu/cgi/t/text/text-idx?cc=mqr;c=mqr;c=mqrarchive;idno=act2080.0046.326;rgn=main;view=text;xc=1;g=mqrg – Michigan Quarterly Review.
www.sephardichorizons.org/Volume1/Issue3/AlbertMemmi.html#_edn2 – Sephardic Horizons.

Memory Binding

Richard J Allen, University of Leeds, Leeds, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

How are the constituent features of elements within the environment bound together in memory to form integrated or associated representations? A body of research has started to emerge examining this broad question, with important implications for how we experience the world around us. The present article explores binding and relational processing across a range of contexts, taking a cognitive perspective in focusing on different forms of binding within working memory and relating this to long-term association formation, before considering how the brain might achieve memory binding.

The objects of memorial processing – visual and auditory stimuli in the environment, for example – typically consist not of meaningless and unidimensional features, but of multiple, often disparate elements, associated with context and current task goals. In the visual domain, for example, objects are made up of multiple configurations of shape, color, and size (e.g., fruit and vegetable selections in a supermarket). Similarly, when processing linguistic information, elements are normally embedded within the context of a meaningful flow of information, rather than as isolated fragments. In each case, elements of the experience must be bound together to form coherent representations in memory, in order to be understood and to potentially influence further action. More broadly, episodes in our lives are typically experienced as integrated occurrences in which information from across modalities and feature dimensions are brought together to form stable and meaningful experiences, rather than random and chaotic sensory noise. Each of these involves processes of *binding* in memory, that is, drawing together and integrating information from different environmental sources and the neural areas responsible for initially processing them.

For the purposes of the present work, binding is used as an overarching term to denote situations in which different forms of information are integrated together in memory. It is assumed that within this, multiple and separable forms of binding can be distinguished, each serving different contexts and types of element. As will be discussed, these are likely to involve differing cognitive processes and draw on distinct neural areas. Binding links to, and is sometimes used interchangeably with, terms such as chunking, grouping, association formation, and relational and conjunctive processing. For example, chunking can refer to an active, strategic method of linking elements together on the basis of prior semantic relatedness, summed meaning, or perceptual features (Miller, 1956). While all these terms refer to the identification or creation of connections between elements (with resulting gains to memory from information compression), it is probably not appropriate to collapse the meanings of these terms together or equate them simply with memory binding. However, such concepts are often relevant when discussing how binding might operate and help in placing such processes in the context of wider cognitive functions. Indeed, they may in some cases reflect fruitful ways of differentiating between subcategories of binding process. Therefore, in some cases they are assumed to reflect forms of binding (e.g., relational and conjunctive processing), while in others they bear relation to binding but are not functionally equivalent.

The question of how information binding is achieved has often been termed the 'binding problem,' primarily in the perception literature (Treisman and Gelade, 1980) but more recently in the context of working memory (WM) and long-term memory (LTM). It is often characterized as being problematic due to the different brain areas and parallel functions thought to be involved in initial processing of sensory information (e.g., form, location, audition, etc.). Solving this problem or, more likely, collection of problems involves adopting perspectives from across cognition psychology and neuroscience (among others) in exploring how these forms of information are processed and integrated, from low-level perceptual integration, through more active and effortful processing and manipulation in WM, up to the integration of rich tapestries of information to be held in LTM, as well as identifying ways in which these levels tie together. It is clearly highly informative to explore how information binding is achieved, as this provides insights into key issues such as the nature of representations in memory, processes of encoding, maintenance, and retrieval, the role of attention, the structure and function of memory, and also helps characterize how memory changes across the lifespan and the nature of deficits that exist in patients with different forms of brain injury or disease. Each of these issues will be briefly covered, in the context of information binding.

Binding in Working Memory

A major focus of recent research has explored how binding processes might operate within WM, that is, the system responsible for temporarily maintaining, processing, and manipulating information from the environment and relating it to prior knowledge. A number of WM models have been proposed and subsequently developed in the literature which, while differing in important ways (e.g., on existence of fractionable, domain-specific subsystems, and how WM and LTM may be related), each emphasize the limited capacity of WM, it's close links with perception, attention and LTM, and the central role of WM in wider cognitive function. Until recently, however, these models did not attempt to explicitly capture information binding. Newly emerging work indicates this may

be a crucial factor in determining limits to WM capacity, the nature of representational storage, and the success with which these representations are created and subsequently retained or lost over time.

Visuospatial Binding

As stated above, the issue of binding has been highlighted for some time in the area of visual perception, concerning how distributed neural codes representing elements within a visual scene are combined to support accurate perception of objects such as a red apple next to a green leaf, rather than a red leaf and a green apple (Treisman and Gelade, 1980). How then might the binding of visual and spatial features such as shape, color, orientation, and location be achieved in WM? This question is often assessed in controlled experiments by presenting participants with arrays of stimuli (often simple objects, to minimize impact of preexisting associative knowledge) and requiring various forms of recognition or recall response (see Figure 1).

One possibility is that, as in perceptual tasks, focused attention is required for features to be successfully and accurately bound together. It is indeed likely that one sense of focused attention, that is, spatial attention, is critical for binding in visual WM. This can be distinguished however from a more domain-general set of attentional resources involved in active, strategic processing, tied to the concept of a central executive. In a revision of the multicomponent model of WM, Baddeley (2000) suggested that active and attention-demanding binding was a key role played by the episodic buffer, a storage component claimed to be involved in binding information from different sources (e.g., separate phonological and visuospatial subsystems) and linking with LTM. However, while the episodic buffer (or an analogous component) may play a role as a passive recipient of prebound representations, research to date has failed to consistently support the notion of an attention-dependent binding process in visual WM. For example, Allen et al. (2006) examined memory for shapes, colors, and shape–color conjunctions while performing simple or more demanding (e.g., backward counting) verbal tasks (intended to disrupt availability of attentional resources) concurrently during presentation and a short retention interval. It was found that the conjunction conditions (measuring binding), in which participants were required to decide whether a color–shape combination had been present in a previously presented display (see Figure 1), were no more affected by this attentional manipulation than single feature memory (in which participants simply decided whether a test shape or color had been encountered). It appears then that, at least for simple unitized objects that are briefly encountered, binding may be achieved relatively automatically, with attention only important for processing constituent features (Baddeley et al., 2011). However, the constraints that apply to this automaticity are yet to be fully mapped out.

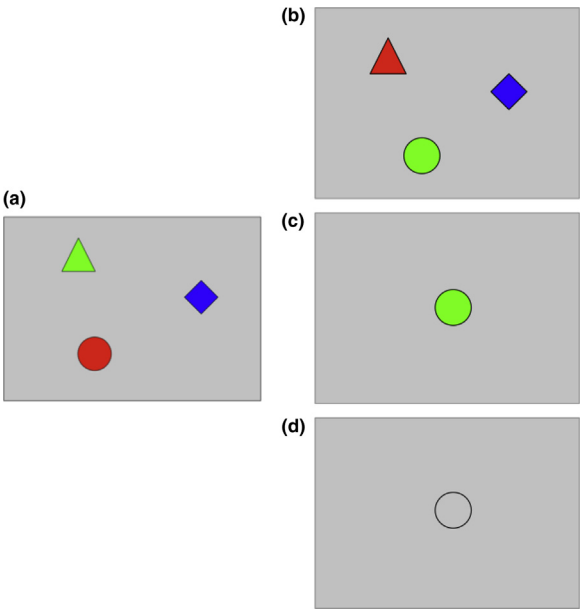

Figure 1 Illustration of typical presentation and test methods used to assess binding of color and shape in visual working memory. To-be-remembered array (a) is briefly presented, followed after a short delay by the test phase. Across different studies, this might involve (b) whole-display change detection recognition, requiring detection of any change in feature conjunctions (normally via key press or button box), (c) single probe recognition, requiring a decision on whether this combination had been present in the display, or (d), cued recall, requiring verbal recall of, for example, which color was presented with the test shape. Similar methods can be implemented to assess other forms of binding, for example, between feature and location. Note that other methods of presentation and assessment have also been utilized, and comparison with memory for constituent features must always be considered.

When unitized objects such as colored shapes are encountered, what is the nature of their representative format? In their seminal work, Luck and Vogel (1997) observed that change detection judgments for features within briefly presented visual arrays were as accurate when objects varied along four different feature dimensions (e.g., color, orientation, size, and presence of a gap), compared to a single dimension, with performance declining as number of distinct items increased. Based on this, Vogel et al. (2001) argued that visual WM representations are object based, with capacity limited (at around three bound objects; Cowan et al., 2012) by number of objects rather than number of features. In contrast, Wheeler and Treisman (2002) suggested that visual WM instead consists of separate feature stores, with the bindings between them being particularly vulnerable to loss caused by attentional withdrawal; thus, attention is important for maintenance of conjunctions between constituent parts. In line with this, Allen et al. (2006) found that bindings are fragile and liable to interference from subsequently encountered stimuli; memory for conjunctions was disproportionately worse when stimuli were presented sequentially, as opposed to simultaneously. However, a range of research has suggested this fragility to be independent of available attentional resources; adding a concurrent load or directing attention elsewhere in the visual scene in the retention delay between presentation and test disrupts performance, but generally not for binding any more than individual features (Morey and Bieler, 2012). Furthermore, Ueno et al. (2011) have demonstrated that binding is fragile even when subsequently encountered interfering stimuli are

to be ignored, indicating that it is not the active encoding of these stimuli into WM that causes loss of bound representations, but their basic presence in the environment and resulting perception.

Cowan et al. (2012) have recently suggested that bindings in memory can break down and lose features independently of each other, leaving partial representations that lack some of their original elements (e.g., we may remember that a *red* object was present, but not that it was a *triangle*). They examined memory for features and their bindings across a range of encoding and retrieval conditions, and found evidence of a fixed object limit (of around three items), but incomplete retention of attributes from these objects. Other studies (Bays et al., 2011) have also observed partial loss of an object's features from memory, and interpreted this instead as evidence for precision based as opposed to object-based constraints on WM, though this issue is somewhat outside the scope of the present review. Nevertheless, it is emerging that objects are not always retained in an all-or-none way, and partial feature loss can occur.

So far, discussion has focused on processing of visual items that are 'unitized,' that is, encountered as single unitary objects in the environment (as in **Figure 1**). It is informative to examine how attributes are bound in memory when they are not clearly part of the same object. The question of what constitutes an object is too broad for the current review, though it is worth noting that studies have highlighted principles such as boundedness, connectedness, temporal contiguity, and shared spatial location (Kahneman et al., 1992). In line with this, it is likely that the processes involved in binding vary depending on whether features are 'unitized' in the environment. For example, a number of studies have demonstrated that visual features are better remembered when presented within the same object and spatial location, as opposed to when they are spatially distinct. Similarly, binding between surface features such as color and their spatial locations, and binding within feature dimensions (e.g., between adjacent colors) may behave differently to unitized shape–color binding, for example, in terms of relative accuracy levels in recognition tasks and whether they show consistent decrements with healthy aging (see Section Neural Basis of Binding).

Related to this, a distinction has recently been drawn between *intrinsic* and *extrinsic* binding (Ecker et al., 2012). The former is equivalent to unitized binding, concerning information, that is, intrinsic to a perceptual unit (e.g., an object's color), while the latter relates to binding of contextual information that is part of the overall event (e.g., the color of another object present alongside the target). Ecker et al. observed that memory for shapes was only disrupted when task-irrelevant changes in color were intrinsic to the target (i.e., part of the same object), and not extrinsic (i.e., part of the scene background). In line with previous work (Baddeley et al., 2011), Ecker et al. (2012) argue that intrinsic binding occurs in an obligatory and automatic fashion, while binding of extrinsically related features does not and may instead require active and strategic encoding. This distinction may mirror those operating during perception and in long-term episodic memory, with binding of extrinsic and contextual details possibly representing a major constraint on WM capacity. Connected to this, work by Oberauer and colleagues (Oberauer et al., 2008) has suggested that such forms of binding and relational processing may be closely linked to individual differences in WM capacity, reasoning, and fluid intelligence. They argue that WM is itself a system primarily for constructing relational information in the form of temporary bindings between separate components of a scene or episode.

Cross-Domain and Cross-Modal Binding

Unitized binding appears to be relatively automatic (though fragile), while some other forms of visual binding, particularly those potentially characterized as extrinsic or relational in nature, may require additional processes and potentially distinct types of binding for features to be connected in memory. In addition, it is possible for information from different domains (e.g., verbal and spatial) and modalities (visual and auditory) to be bound in memory (e.g., remembering the location of captions placed on an image, or a group of people's names). These forms of binding represent a particularly strong test of the episodic buffer concept (Baddeley, 2000), in linking together separate information streams in WM. Other theoretical approaches to WM that do not emphasize independent subsystems do not require such a concept, though they still require further development in order to effectively capture these abilities.

There is somewhat less work on such forms of binding at present. Elsley and Parmentier (2009) examined the role of attentional control in letter–location binding in WM, and found that memory judgments concerning letters and locations were faster and more accurate when initially presented in conjunction (as opposed to being part of separate 'objects'), but this binding effect was considerably reduced when a demanding tone-judgment task was concurrently performed. On this basis, Elsley and Parmentier argued for an attention-demanding cross-domain binding process, in line with Prabhakaran et al.'s (2000) observation of greater prefrontal activation in trials involving integration of letter and location. This might suggest that binding across domains is more akin to extrinsic than intrinsic binding in terms of attentional demands, although more work is required to explore the constraints on automaticity in this regard. Nevertheless, these findings indicate that binding letters to spatial locations in memory can benefit memory for the constituent elements, perhaps through reducing memory load or enhancing access at retrieval. A similar finding has been demonstrated using verbal recall; Darling et al. (2012) demonstrated that verbal digit recall is improved when digits are presented within certain spatial configurations, as opposed to single locations. Importantly, this verbal–spatial binding was shown to be crucially dependent on whether these associations were already present in LTM, as the effect only emerged for digits presented in typical numerical keypad displays (as found on keyboards and telephone interfaces), and not when the digits were randomly reorganized within the same overall configuration. Thus, implicit verbal–spatial binding appears to support recall of verbal information only when the participant has preexisting knowledge of such associations.

While examining binding across domains, these studies still involve integration of information presented within the same modality (i.e., visual). What about binding of information across modalities (e.g., between vision and audition)? Participants do seem to be able to make recognition judgments concerning associations between sequentially presented visual stimuli (e.g., shapes) and aurally presented verbal labels (color names), as accurately as to unitized colored shapes (see Baddeley et al., 2011). Binding between audition and spatial location is also possible; we are able to localize sounds such as speech both in perception and memory. Maybery et al. (2009) observed consistent evidence for memory of the binding between aurally presented letters or nonverbal sounds and which of the locations they were presented from, noting greater persistence for such conjunctions than in visual feature binding (Allen et al., 2006). Further work is required to establish how such integration takes place, and whether additional processes are required above and beyond simple feature binding within modalities.

Verbal Binding

Although involving different representational forms and types of processing to those outlined so far, it is also possible to identify binding during purely verbal tasks. This may operate in a hierarchical fashion, with binding between components such as syllables and phonemes, between elements and their position in a sequence (potentially via temporal or ordinal context signals), between words in the formation of meaningful phrases and sentences, and between concepts in order to form mental models and representations of complex episodes.

The established observation that memory for meaningful sequences such as sentences is more accurate than for random lists of words may reflect a further form of information binding (though interchangeable use of the terms binding and chunking is more prevalent in this area). Baddeley et al. (2009) examined the contribution of attentional resources to sentence and word list processing using concurrent verbal and visuospatial tasks of varying levels of difficulty. It was concluded that use of linguistic structure and redundancy in the binding between elements of a sentence in order to form larger cohesive segments appears to develop relatively automatically; while demanding concurrent tasks had large effects on performance, they did not reduce the recall advantage for sentences over lists. This advantage necessarily involves access to linguistic knowledge concerning meaning and grammatical structure in LTM, with subsequent storage of bound segments or chunks potentially held in WM within a component such as the episodic buffer.

While Baddeley et al. (2009) examined within-sentence binding, Jefferies et al. (2004) explored binding across sentence boundaries, comparing memory for short sets of sentences that were either presented within the context of a meaningful narrative, or in a random, meaningless order. In relation to the impact on memory for random word lists, a concurrent visuospatial task (intended to load on attention) did have a significantly larger disruptive effect on unrelated sentence memory. However, this was not the case for related sentence sets presented in the context of a narrative. Thus, binding or chunking together meaningful concepts across sentence boundaries does seem to require additional attentional resources, but only when the connections between these segments need to be newly constructed and do not fit into a schematic narrative.

Binding and Association Formation in LTM

Discussion has focused primarily on binding over the short term, within working memory. As it is plausible to regard WM and LTM as separate constructs (Baddeley, 2000), how then might more robust, long-lasting representations be created, and how does this relate to the studies outlined above? In terms of verbal conjunctions, temporary bindings in WM may be translated into LTM-based associations, perhaps through a gradual process of Hebbian learning (Baddeley et al., 2009). A similar process was described for the binding of shape and color in visual memory by Logie et al. (2009), with repeatedly presented bindings gradually leading to longer-term residual traces of object representations, although this learning process was very fragile and liable to disruption from new presentations on later trials. Thus, while functionally independent, it is likely that binding processes and representations in WM and LTM are mutually supportive and to some extent share representational formats.

Hommel and Colzato (2009) distinguished between the formation of new, arbitrary binding in WM (e.g., *red-circle*) and familiar conjunctions in LTM (e.g., *red strawberry*). They do note, however, that in each case binding is automatic. Indeed, development of long-term associative memory representations can, in some cases, operate without specifically relying on attention. Naveh-Benjamin and colleagues (Naveh-Benjamin et al., 2003) have shown that associative memory for unrelated word pairs, nonword pairs, or word and presentation context pairings is not particularly affected by divided attention during encoding. This fits with certain forms of automatized binding and chunking in WM (Baddeley et al., 2011), though it may be that temporary chunking of previously unrelated verbal pairs in WM does still draw on attentional resources. Again, it is likely that key processes vary with the particular nature of the task and the form of binding being examined.

One consistent parallel emerging between binding in WM and LTM relates to the intrinsic–extrinsic distinction discussed by Ecker et al. (2012). Zimmer and Ecker (2010) have outlined a model of LTM that distinguishes between object tokens (containing intrinsic detail only) and episodic tokens (involving both intrinsic and contextual extrinsic information). Such a distinction has many similarities with that often drawn between conjunctive and relational integration (Moses and Ryan, 2006). Ecker et al. suggest that this may be fundamental across tasks ostensibly measuring perception, WM, and LTM. Indeed, unitization effects (i.e., superior memory for conjunctions of color and shape when part of the same object) have been observed in long-term recognition tasks that may be analogous to those reported in WM. Therefore, as in the case of WM, it may be that only LTM tasks requiring memory for extrinsic binding information (including context) are affected by divided attention

manipulations. How this explanation maps onto the forms of binding previously described remains to be explored.

Neural Basis of Binding

It is important, of course, to map out the neural as well as the cognitive underpinnings of memory binding. Discussion of how the brain achieves memorial binding has focused on two different questions: how and where. In terms of the former, Hommel and Colzato (2009) identify convergence and synchronization as two principles that have been particularly influential in describing how conjunctions are represented at the neural level. Convergence describes a hierarchical system in which lower level neurons code for simple features (e.g., shape or color) that project onto higher level neurons that code for conjunctions, which in turn may feed to higher levels for objects. As often noted, however, a system purely based on this principle would suffer from the problem of combinatorial explosion (i.e., there are too many possible feature combinations for each to be represented individually). Synchronization, in contrast, is the suggestion that neurons representing individual features achieve binding by firing in temporal synchrony (Vogel et al., 2001). Hommel and Colzato (2009) have suggested these principles may operate together, with convergence more likely to support memory for familiar feature conjunctions (e.g., yellow banana, red strawberry) while synchronization is more important for arbitrary conjunctions (e.g., yellow circle, red square).

Irrespective of function at the neural level, where in the brain might different forms of binding be based? It is established that many forms of relational or conjunctive processing rely on the hippocampus within the medial temporal lobes (Moses and Ryan, 2006; O'Reilly and Rudy, 2001). This certainly seems to be the case for LTM, with binding or associative memory between constituent elements of an episode being impaired in patients with hippocampal lesions and showing increased activation in studies using PET and fMRI, relative to single elements (Cohen et al., 1999). Thus, in line with the critical role it is thought to play in supporting episodic memory and recollective experience (including memory for contextual detail), the hippocampus is central for LTM binding. As hippocampal function is known to deteriorate with healthy aging, this may explain why older adults show an associative memory deficit in LTM (Naveh-Benjamin et al., 2003), displaying particular problems acquiring new associations between paired stimuli of different types (e.g., words, pictures, and faces). Similarly, Alzheimer's disease is associated with hippocampal degradation, and tasks measuring learning of new associates in LTM (such as the paired associates subtask of the CANTAB (Cambridge Neuropsychological Test Automated Battery), in which participants are required to learn sequences of shape–location pairs) appear to be a particularly sensitive method of identifying individuals with dementia.

The role of the hippocampus appears to be more compelling for extrinsic or relational binding, as opposed to intrinsic/conjunctive binding (Ecker et al., 2012; Moses and Ryan, 2006). Indeed, Ecker et al. argue that anterograde amnesia, typically associated with hippocampal damage, may not be an LTM impairment, but instead reflects a deficit in extrinsic binding. If this is the case, the hippocampus should also be crucial to certain forms of temporary binding in WM tasks (e.g., those requiring binding between objects, or to location), and this is indeed what has been reported across several studies using fMRI or examining patients with medial temporal lobe damage (Olson et al., 2006; Mitchell et al., 2000; though see Ryan and Cohen, 2004). Similarly, certain age-related deficits in binding between items and their spatial locations in WM tasks have also been observed (Mitchell et al., 2000), which may also at least partly reflect medial temporal lobe decline.

In contrast, other forms of binding in WM, for example, between shape and color, appears to be intact in individuals with selective hippocampal damage. One such patient, Jon, was shown to perform extremely well on a battery of tasks assessing WM for sequentially presented color–shape conjunctions of various forms (Baddeley et al., 2010). Evidence regarding older adults' performance on such tasks is mixed to date (Brown and Brockmole, 2010), though on balance also indicates that healthy aging is not consistently and robustly associated with particular problems in shape–color binding in the same way that long-term relational binding appears to be, and thus may operate independently of age-related hippocampal decline. Recent research is also emerging to indicate that measures of binding (e.g., between shape and color) within WM may also serve as particularly useful preclinical markers in detecting the onset of Alzheimer's disease (Parra et al., 2010) They examined a group of related individuals with a gene mutation (E280A in *presenilin-1*) that always leads to this disease, and observed deficits in a task requiring WM for shape–color binding in carriers of the mutation who were either diagnosed with Alzheimer's disease, or were asymptomatic. Healthy relatives without the gene mutation did not show this binding deficit. As already stressed, this impairment is likely separable from more established hippocampal-based LTM deficits, and may instead be attributable to a wider loss of brain connectivity in Alzheimer's disease that may be important in this task. Nevertheless, the emerging use of such tasks in the early detection of individuals with potential cognitive impairment is an example of how developments in methodology and theory can feed into beneficial impacts on society.

It is clear that areas of the prefrontal cortex are important across the majority of memory tasks, given this area's potentially diverse and critical roles in executive function and attentional control. What is less clear, however, is whether the prefrontal cortex might specifically have a function in information binding. As noted earlier, Prabhakaran et al. (2000) used fMRI during a task aimed at measuring letter and location memory and found evidence to suggest that the prefrontal cortex, specifically right BA 10, may be highly activated when letter–location conjunctions are processed. Similarly, in their study of young and older adults, Mitchell et al. (2000) observed increased activation in the same area of the prefrontal cortex (as well as hippocampal activity) in a task measuring shape–location binding in WM. On this basis, Mitchell et al. argued for a frontal-hippocampal binding circuit, though noted that the precise contribution of right BA 10 to binding remains unclear (as this area is different from

those typically identified as being involved in maintenance, manipulation, or sustained attention). In each of these cases, however, further research is required to corroborate and more closely specify the precise areas that may be involved.

A further area that may have a potentially important role in information binding is the parietal cortex. Shafritz et al. (2002) used fMRI to demonstrate increased activation in parietal areas normally involved in spatial attention (right superior parietal and anterior intraparietal), when binding conjunctions of shape and color, though only when location was available as a cue, thus supporting Treisman and Gelade's (1980) claims of a central role for location in visual binding. As WM may be closely tied to the products of perception, links can be drawn with observations of shape–color binding problems in perceptual tasks reported by patients with bilateral parietal lesions (Friedman-Hill et al., 1995). Indeed, areas of the parietal cortex including the intraparietal sulcus have been identified as being crucial for working memory storage more generally, possibly reflecting integration of information from within and across modalities as part of a frontoparietal network.

Overall, it is likely that different networks of brain areas are critically involved in memory binding, but as with patterns of findings from empirical research, this likely varies with features of the memory task, depending on stimuli, type of binding required, and relative loads on temporary storage and LTM, amongst other factors. As well as frontal-based control, the available evidence strongly suggests a crucial hippocampal involvement for LTM, though this is less clear for WM tasks, which may also (or alternatively) draw on parietal contributions for storage.

Summary

Binding, while used here as a single overarching concept, relates to disparate sets of processes within and across domains. Recent work has started to specify its nature and function in each case, and in doing so has provided important new insights and constraints for theories of memory. In addition, studies have started to explore ways in which memory binding varies with healthy cognitive aging, disease, and brain damage. In this way, developing a full picture of how information binding operates is potentially very useful not only for informing theory, but also for identifying, understanding, and potentially remediating cognitive impairment.

See also: Aging and Memory; Episodic Memory; Hippocampal Amnesia; Hippocampus and Related Structures; Short Term Memories, Theories of; Short-Term Memory: Psychological and Neural Aspects; Visual Memory, Psychology of; Working Memory, Neural Basis of; Working Memory, Psychology of.

Bibliography

Allen, R.J., Baddeley, A.D., Hitch, G.J., 2006. Is the binding of visual features in working memory resource-demanding? Journal of Experimental Psychology: General 135, 298–313.

Baddeley, A.D., 2000. The episodic buffer: a new component of working memory? Trends in Cognitive Sciences 4 (11), 417–423.

Baddeley, A.D., Hitch, G.J., Allen, R.J., 2009. Working memory and binding in sentence recall. Journal of Memory and Language 61, 438–456.

Baddeley, A.D., Allen, R.J., Vargha-Khadem, F., 2010. Is the hippocampus necessary for visual and verbal binding in working memory? Neuropsychologia 48, 1089–1095.

Baddeley, A.D., Allen, R.J., Hitch, G.J., 2011. Binding in visual working memory: the role of the episodic buffer. Neuropsychologia 49, 1393–1400.

Bays, P.M., Wu, E.Y., Husain, M., 2011. Storage and binding of object features in working memory. Neuropsychologia 49, 1622–1631.

Brown, L.A., Brockmole, J.R., 2010. The role of attention in binding visual features in working memory: evidence from cognitive ageing. The Quarterly Journal of Experimental Psychology 63, 2067–2079.

Cohen, N.J., Ryan, J., Hunt, C., Romine, L., Wszalek, T., Nash, C., 1999. Hippocampal system and declarative (relational) memory: summarizing the data from functional neuroimaging studies. Hippocampus 9, 83–98.

Cowan, N., Blume, C.L., Saults, J.S., 2012. Attention to attributes and objects in working memory. Journal of Experimental Psychology: Learning, Memory, & Cognition.

Darling, S., Allen, R.J., Campbell, A., Rattray, E., 2012. Visuospatial bootstrapping: long-term memory representations are necessary for implicit binding of verbal and visuospatial working memory. Psychonomic Bulletin & Review 19, 258–263.

Ecker, U.K.H., Maybery, M., Zimmer, H.D., 2012. Binding of intrinsic and extrinsic features in working memory. Journal of Experimental Psychology: General.

Elsley, J.V., Parmentier, F.B.R., 2009. Is verbal-spatial binding in working memory impaired by a concurrent memory load? The Quarterly Journal of Experimental Psychology 62 (9), 1696–1705.

Friedman-Hill, S.R., Robertson, L.C., Treisman, A., 1995. Parietal contributions to visual feature binding: evidence from a patient with bilateral lesions. Science 269, 853–855.

Hommel, B., Colzato, L.S., 2009. When an object is more than a binding of its features: evidence for two mechanisms of visual feature integration. Visual Cognition 17 (1/2), 120–140.

Jefferies, E., Lambon Ralph, M.A., Baddeley, A.D., 2004. Automatic and controlled processing in sentence recall: the role of long-term and working memory. Journal of Memory and Language 51, 623–643.

Kahneman, D., Treisman, A., Gibbs, B.J., 1992. The reviewing of object files: object-specific integration of information. Cognitive Psychology 24 (2), 175–219.

Logie, R.H., Brockmole, J.R., Vandenbroucke, A.R.E., 2009. Bound feature combinations in visual short-term memory are fragile but influence long-term learning. Visual Cognition 17, 160–179.

Luck, S.J., Vogel, E.K., 1997. The capacity of visual working memory for features and conjunctions. Nature 390, 279–281.

Maybery, M.T., Clissa, P.J., Parmentier, F.B.R., et al., 2009. Binding of verbal and spatial features in auditory working memory. Journal of Memory and Language 61, 112–133.

Miller, G.A., 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review 63, 81–97.

Mitchell, K.J., Johnson, M.K., Raye, C.L., D'Esposito, M., 2000. fMRI evidence of age-related hippocampal dysfunction in feature binding in working memory. Cognitive Brain Research 10, 197–206.

Morey, C.C., Bieler, M., 2012. Visual short-term memory always requires general attention. Psychonomic Bulletin & Review.

Moses, S.N., Ryan, J.D., 2006. A comparison and evaluation of the predictions of relational and conjunctive accounts of hippocampal function. Hippocampus 16, 43–65.

Naveh-Benjamin, M., Hussain, Z., Guez, J., Bar-On, M., 2003. Adult age differences in episodic memory: further support for an associative-deficit hypothesis. Journal of Experimental Psychology: Learning, Memory, & Cognition 29 (5), 826–837.

Oberauer, K., Süß, H.-M., Wilhelm, O., Wittmann, W.W., 2008. Which working memory functions predict intelligence? Intelligence 36, 641–652.

Olson, I.R., Page, K., Moore, K., Chatterjee, A., Verfaellie, M., 2006. Working memory for conjunctions relies on the medial temporal lobe. Journal of Neuroscience 26, 4596–4601.

O'Reilly, R.C., Rudy, J.W., 2001. Conjunctive representations in learning and memory: principles of cortical and hippocampal function. Psychological Review 108 (2), 311–345.

Parra, M.A., Abrahams, S., Logie, R.H., Mendez, L.G., Lopera, F., Della Sala, S., 2010. Visual short-term memory binding deficits in familial Alzheimer's disease. Brain 133, 2702–2713.

Prabhakaran, V., Narayanan, K., Zhao, Z., Gabrieli, J.D., 2000. Integration of diverse information in working memory within the frontal lobe. Nature Neuroscience 3, 85–90.

Ryan, J.D., Cohen, N.J., 2004. Processing and short-term retention of relational information in amnesia. Neuropsychologia 42, 497–511.

Shafritz, K.M., Gore, J.C., Marois, R., 2002. The role of the parietal cortex in visual feature binding. Proceedings of the National Academy of Sciences of the United States of America 99, 10917–10922.

Treisman, A.M., Gelade, G., 1980. A feature-integration theory of attention. Cognitive Psychology 12, 97–136.

Ueno, T., Allen, R.J., Baddeley, A.D., Hitch, G.J., Saito, S., 2011. Disruption of visual feature binding in working memory. Memory & Cognition 39, 12–23.

Vogel, E.K., Woodman, G.F., Luck, S.J., 2001. Storage of features, conjunctions, and objects in visual working memory. Journal of Experimental Psychology: Human Perception and Performance 27 (1), 92–114.

Wheeler, M.E., Treisman, A.M., 2002. Binding in short-term visual memory. Journal of Experimental Psychology: General 131, 48–64.

Zimmer, H.D., Ecker, U.K.H., 2010. Remembering perceptual features unequally bound in object and episodic tokens: neural mechanisms and their electrophysiological correlates. Neuroscience and Biobehavioral Reviews 34, 1066–1079.

Memory Development During Infancy and Early Childhood across Cultures

Harlene Hayne, University of Otago, Dunedin, New Zealand
Kana Imuta, University of Queensland, St Lucia, QLD, Australia
Damian Scarf, University of Otago, Dunedin, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Our autobiography is composed of memories from our past and plans for our future. Most adults' earliest autobiographical memories date from between 3 and 4 years of age and the absence of memories from before this age is referred to as infantile or childhood amnesia. This article traces memory development from the emergence of memory during infancy to the emergence of autobiographical memory in early childhood.

> This is the use of memory:
> For liberation – not less of love but expanding
> Of love beyond desire, and so liberation
> From the future as well as the past.
> T. S. Eliot (1943)

As T.S. Eliot (1943) notes, memory liberates us from being bound solely to the past, present, or future; it allows us to travel between these epochs in time and, in doing so, allows us to construct our autobiography. An autobiography is composed of not only memories from the past but also thoughts of the present and plans for the future. To adults, these aspects of memory are second nature but, when thinking about our personal past, just how far back can we go? And, if we can only go so far back, does that mean that at some point in our lives we did not have *memory*? The answer to the first question is clear; most adults' earliest autobiographical memories date from between 3 and 4 years of age (e.g., Davis et al., 2008; Jack and Hayne, 2007, 2010; Jack et al., 2012; Sheingold and Tenney, 1982). The almost universal absence of autobiographical memories from before this age is referred to as infantile or childhood amnesia (Freud, 1963). With respect to the second question, the answer is also clear. Although our early past cannot be recalled as adults, even very young infants are able to encode, store, and retrieve memories after significant delays (e.g., Barr et al., 1996; Hartshorn et al., 1998; Hayne et al., 2000, 1986). Having no *autobiographical memories* from a time when we once had *memory* may seem somewhat paradoxical; it simply reflects the complex nature of memory development. This article traces memory development from the emergence of explicit memory during infancy to the emergence of autobiographical memory in childhood.

What Is Memory?

There is general consensus that there is more than one type of memory (cf Roediger, 1984; Snodgrass, 1989). Typically, a distinction is made between explicit and implicit memory (Graf and Schacter, 1985), also referred to as declarative and nondeclarative memory (Squire, 1992). Essentially, theorists draw a distinction between memories that people are consciously aware of (i.e., declarative or explicit) and those that they are not (i.e., nondeclarative or implicit) (Hereafter we will use the terms implicit and explicit, as these are the terms most commonly used in the developmental literature (Hayne, 2004; Rovee-Collier et al., 2000).) This distinction was largely born out of experiments with people who became amnesic as a result of the removal of, or damage to, the medial temporal lobe (Scoville and Milner, 1957). These patients displayed profound amnesia for recent events and, highlighting the pivotal role of the hippocampus in memory, the degree of impairment was positively related to "… the extent of destruction to the hippocampal complex" (Scoville and Milner, 1957: p. 20). Critically, follow-up experiments with one patient established that he still displayed memory for simple perceptual and motor tasks (Corkin, 1968, 1984, 2002; Milner et al., 1968). The patient learned these tasks over several laboratory visits, suggesting that the learning from one visit carried over to the next, despite the fact that he could not recall his previous visit to the laboratory or the task that he performed. This clear memory dissociation was explained by appealing to a dual memory system, with the medial temporal lobe being critical for explicit memory, such as memory for daily life events, while implicit memory is supported by brain regions outside the medial temporal lobe (Cohen and Squire, 1980; Tulving and Schacter, 1990).

Schacter and Moscovitch (1984) were the first to apply this dual memory view to memory development, arguing that the implicit memory system is functional at birth and that the explicit memory system becomes functional between 8 and 10 months of age. Some authors have supported this view (Carver and Bauer, 2001; Richmond and Nelson, 2007), whereas others have argued that the implicit and explicit memory systems develop simultaneously rather than sequentially (Rovee-Collier et al., 2000), that the explicit memory system is preceded by a preexplicit system (Nelson, 1995), or that a dual memory system is not needed at all

to explain infant memory development (Rovee-Collier and Cuevas, 2009). While these authors disagree on exactly when during development the explicit memory system is functional, there is clear agreement that it is functional by 12 months of age.

Assessing Explicit Memory across Development

One obvious question is how can explicit memory be tracked across development? Young children's limited language skills mean that one cannot simply ask them to tell what they remember. Over the past several decades, developmental psychologists have come up with some ingenious tasks that allow young children to nonverbally convey their explicit memory skill. These tasks include both deferred (Meltzoff, 1995) and elicited (Bauer and Mandler, 1989) imitation paradigms, and conditioning tasks, such as the mobile conjugate (Rovee and Rovee, 1969) and train (Hartshorn and Rovee-Collier, 1997) reinforcement paradigms. While each of these paradigms has made a significant contribution to the study of memory development (see Hayne, 2004; Rovee-Collier, 1997; for reviews), they do require procedural changes when assessing memory in participants of different ages. For example, the complexity and number of actions in imitation tasks must be altered to equate for task difficulty across ages. Similarly, while the mobile conjugate paradigm is appropriate for use with infants up to 7 months of age, it must be substituted with the train paradigm to assess memory in older infants and young children.

Among developmental memory paradigms, the visual paired comparison (VPC, Berlyne, 1958; Fantz, 1956) paradigm is unique in that the exact same stimuli and procedure can be used with infants, children, and adults. In the VPC paradigm, participants are familiarized with a visual stimulus, and after a delay, they are presented with a pair of visual stimuli – the stimulus they viewed during familiarization and a novel stimulus (**Figure 1**). Memory is inferred if participants spend a greater proportion of time looking at the novel, relative to the familiar, stimulus. This preference for the novel stimulus is thought to be the result of participants recognizing the familiar stimulus and, consequently, directing more attention to processing the novel one. This interpretation is based on Sokolov's (1963) comparator model, which holds that participants attempt to match each stimulus they encounter to a stored representation. If a match is found, attention to that stimulus is inhibited, whereas if no match or a partial match is found, then attention is maintained. Despite the relative simplicity of the task, there is a growing body of neurophysiological and developmental work supporting the view that the VPC paradigm provides a measure of explicit memory. For example, patients with damage to the hippocampus show substantial deficits on the VPC paradigm when tested after a delay (McKee and Squire, 1993; Munoz et al., 2011; Pascalis et al., 2004). Furthermore, using Tulving's (1983) principle of experimental dissociation, researches have shown that the same variables that influence adults' performance on a wide range of explicit memory tasks also influence infants' performance on the VPC paradigm.

Explicit Memory Development

Although the VPC paradigm has historically been used with very young infants (Fantz, 1956), the authors have recently used it to trace the development of explicit memory during infancy and early childhood. All their work with the VPC paradigm was carried out using an apparatus that consists of a three-sided enclosure, covered in black felt, with two 21-inch monitors mounted to the back panel of the enclosure. The monitors are positioned 46 cm apart and a low-light video camera, also mounted to the back panel of the enclosure, is positioned between them to allow for off-line coding of infants' looking. In all the studies described here, specially designed stimuli, consisting of computer-generated three-dimensional cartoon faces – a blue mail-box shaped face, a yellow circular face, and a red square – were used (see **Figure 1**). During presentation, the faces either blinked their eyes or moved their mouth. Although some researchers have argued that the VPC paradigm may not be suitable for young children (Atkinson, 2000), the authors' use of animated stimuli has all but eliminated participant attrition due to a lack of interest. Over the past few years, this task has been used to study age-related changes in encoding, retention, and retrieval.

Encoding

Prior research has shown that the speed with which infants encode information increases considerably over the first 6 months of life (Davis and Rovee-Collier, 1983; Greco et al., 1986; Hill et al., 1988). Using the VPC paradigm, Morgan and Hayne (2006b) traced age-related changes in encoding time from infancy through early childhood. In their study, separate groups of 1- to 4-year-old children received either a 5- or 10-s familiarized period and were then tested either immediately or after a 24-h or 1-week delay. Note that the immediate test provides a pure measure of encoding. When the familiarization period was just 5 s, 1-year-old children did not display a novelty preference even when they were tested immediately after familiarization. This finding suggests that, for 1-year-olds, the 5-s familiarization was simply not sufficient for them to encode the stimulus. The 4-year-old children, on the other hand, exhibited a novelty preference when tested immediately and after the 24-h delay. When the familiarization period was 10 s, both the 1- and 4-year-old children exhibited retention when tested immediately, but only the 4-year-old children exhibited retention after the 24-h and 1-week delays.

The fact that the 1-year-old children displayed a significant novelty preference only when the familiarization period was 10 s, and that the novelty preference was only observed when they were tested immediately, suggests that 10 s may be the minimum amount of time 1-year-old children require to encode a stimulus on the VPC paradigm. By 4 years of age, however, a 5-s familiarization period is sufficient for encoding. The additional finding that 4-year-old children's retention increased when the familiarization time was increased from 5 to 10 s suggests that encoding is not an all-or-none phenomenon, a point that is touched on when discussing long-term retention.

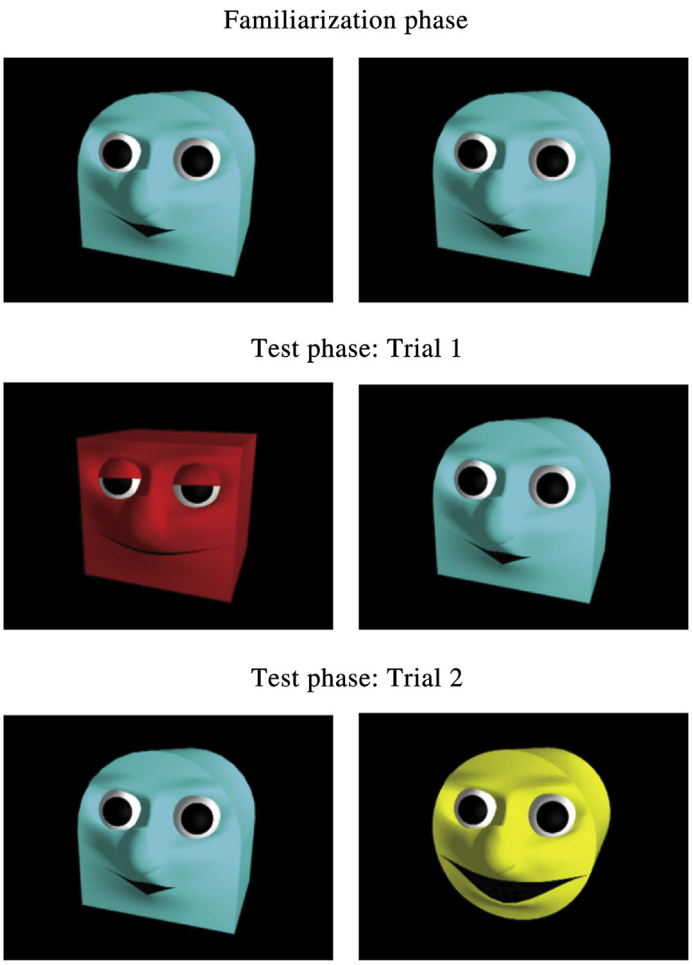

Figure 1 An example of an experimental trial. Top panel: familiarization phase in which identical stimuli are displayed on both screens. Middle panel: Trial 1 of the test phase, with a novel stimulus presented on one screen and the familiar stimulus presented on the other screen. Bottom panel: trial 2 of the test phase in which the right and left positions of the test stimuli are switched. The animated stimuli are provided in the multimedia annexes.

Long-Term Retention

The duration over which infants and young children can recall information increases steadily over the first 18 months of life (Greco et al., 1986; Hartshorn et al., 1998; Hill et al., 1988; Vander Linde et al., 1985). Using the VPC paradigm, Morgan and Hayne (2011) investigated whether this trend continues through to early childhood. In Experiment 1 of their study, 1-, 2-, and 3-year-old children were familiarized with one of the three cartoon faces for 10 s and were then tested either immediately or after a 24-h or 1-week delay. Consistent with Morgan and Hayne (2006b), 1-year-old children displayed a novelty preference when tested immediately, but not after the 24-h or 1-week delays. In contrast, 2-year-old children exhibited retention when tested immediately and after the 24-h delay, but not when tested after the 1-week delay; 3-year-old children exhibited retention across all delays. To test the limits of 3-year-old children's retention, a separate group was tested after a 1-month delay. After the 1-month delay, 3-year-olds did not display a novelty preference, suggesting that the limits of their retention on the VPC paradigm falls somewhere between 1 and 4 weeks. In Experiment 2 of their study, 4-year-old children were tested after the same delays as the 1-, 2-, and 3-year-old children, as well as after an additional 6-month delay. The 4-year-old children exhibited retention after all delays, including the 6-month delay. This sharp increase in retention between 3 and 4 years of age suggests that the linear increase in retention observed by Hartshorn et al. (1998) may not apply across all stages of development (**Figure 2**). Interestingly, the sharp increase in retention documented by Morgan and Hayne (2011) corresponds to the age at which many adults recall their first autobiographical memory.

Memory Retrieval

A large body of work has now shown that even once a memory has been forgotten, it can sometimes be restored by the presentation of a reminder (Campbell and Jaynes, 1966; Hartshorn and Rovee-Collier, 1997; Rovee-Collier et al., 1980; Spear and Parsons, 1976). For example, in the seminal demonstration of reminding in infants, Rovee-Collier et al. (1980) trained

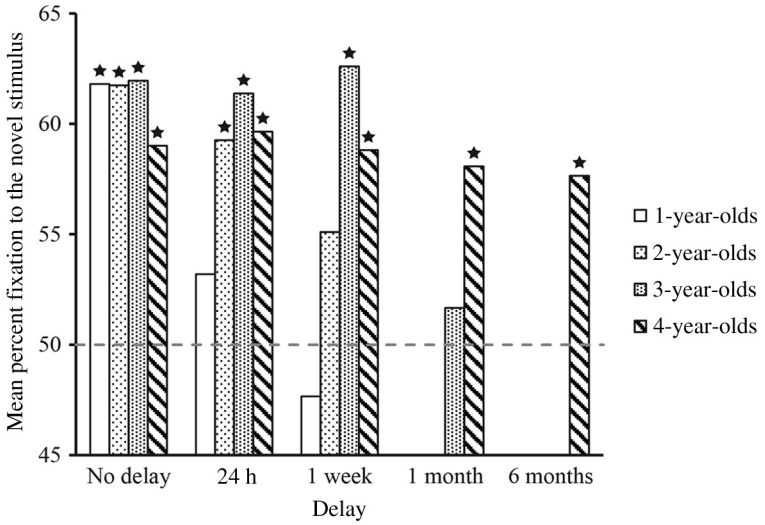

Figure 2 The performance of 1-, 2-, 3-, and 4-year-old children on the VPC paradigm as a function of delay.

3-month-old infants in the mobile conjugate reinforcement paradigm and tested them after a 13- or 27-day delay – retention intervals after which 3-month-olds typically fail to exhibit retention. Twenty-four hours prior to testing, infants were presented with a reminder consisting of a 3-min exposure to the training mobile. Importantly, during the reminder, the mobile was controlled by the experimenter rather than the infant, precluding the opportunity for any retraining during the reminder treatment. During the retention test, the infants in both the 13- and 27-day delay conditions exhibited excellent retention; in fact, their performance was virtually identical to that seen immediately following the conclusion of training 2–4 weeks earlier, clearly demonstrating that forgotten memories can be retrieved and expressed under the appropriate circumstances.

Extending the work previously conducted with young infants, Morgan and Hayne (2006a) assessed the effectiveness of a reminder treatment on retention by 1- and 2-year-old children using the VPC paradigm. In their initial experiments, children were assigned to a reminder group or one of three control groups: forgetting control, encoding control, or reminder control. Children in the reminder group received a 10-s familiarization period followed by a 1-week delay, a delay after which children in both age groups typically fail to show a novelty preference (Morgan and Hayne, 2006b, 2011). After the delay, children in the reminder condition received a brief reminder, consisting of a 1-s presentation of the familiarized stimulus, immediately followed by the test. Children in the forgetting control group received the same familiarization and test as children in the reminder condition, but did not receive the reminder treatment prior to testing. Children in the encoding control condition were tested immediately after familiarization (i.e., they did not experience a delay) and children in the reminder control condition received only the reminder treatment and test (i.e., they did not experience the initial 10-s familiarization period).

Consistent with prior research, 10 s was sufficient for 1- and 2-year-olds to encode the target stimulus, that is, children in the encoding control condition exhibited retention when they were tested immediately after familiarization. Also consistent with prior research, following a 10-s familiarization period, 1- and 2-year-old children exhibited forgetting after a 1-week delay, that is, children in the forgetting control condition did not exhibit a novelty preference during the test. In addition, exposure to the reminder alone, in the absence of prior familiarization yielded no evidence of memory during the test, that is, children in the reminder control condition did not display a novelty preference. With respect to the reminder condition, 2-year-old children exhibited a significant novelty preference after the 1-week delay, demonstrating that the brief reminder was sufficient to reactivate the forgotten memory. Although 1-year-old children did not exhibit retention under these familiarization and reminding conditions, they did exhibit retention if the duration of the original familiarization period or the duration of the reminder was increased.

Prior research has shown that while younger infants require the re-presentation of the original experimental stimuli or the original context in order for memory retrieval to be initiated (e.g., Morgan and Hayne, 2006a; Rovee-Collier et al., 1980), older children can exploit retrieval cues that are different from those they originally encountered (e.g., Hayne et al., 2000; Herbert and Hayne, 2000). This increased flexibility in memory retrieval allows older children (and adults) to access and use their memories under conditions that are different from those under which the memory was originally encoded. In addition to physical reminders such as objects and contexts that were part of the original event, verbal reminders are also powerful retrieval cues. When during development can children begin to exploit verbal retrieval cues to reactivate their forgotten memories? In the first study designed to address this question, Morgan and Hayne (2007) assessed the effect of verbal reminders on memory reactivation in young children. In their study, separate groups of 4-year-old children received a 5-s familiarization period in the VPC paradigm followed by a 2-week delay. Utilizing the method developed by Morgan and Hayne (2006a), children in the nonverbal reminder group were reexposed to the

familiarization stimulus for 1 s prior to testing. Children in the verbal reminder group received the same familiarization time and were tested after the same delay, but before entering the experimental room, they were provided with the following reminder: "*Can you remember coming here two weeks ago?*" followed by "*Can you remember what face you saw last time you came?*" Consistent with the previous study, forgetting, encoding, and reminder control conditions were also included.

Children in both the nonverbal and verbal reminder conditions displayed a significant novelty preference during the test. In addition, there was no difference in the degree of novelty preference displayed by the nonverbal and verbal reminder groups suggesting that at least at 4 years of age, verbal and nonverbal reminders are equally effective in reactivating the forgotten memory. To investigate whether there were age-related changes in children's ability to utilize verbal reminders, Imuta et al. (2013) tested 2-, 3-, and 4-year-old children using a paradigm comparable to that used by Morgan and Hayne (2007). Replicating Morgan and Hayne's (2007) finding, 4-year-old children displayed a significant novelty preference following the presentation of the verbal reminder. Furthermore, both the 2- and 3-year-old children displayed a significant novelty preference following the presentation of the verbal reminder (Imuta et al., 2013). This finding suggests that by at least 2 years of age, children can use another person's verbal cue to reactivate a forgotten memory.

The research reviewed thus far demonstrates that recognition memory continues to develop throughout early childhood. Although recognition memory is a critical component of the explicit memory system, it is just one form of explicit memory subserved by the hippocampus. Indeed, recent neurophysiological evidence suggests that the emergence and development of recognition memory, like other hippocampus-dependent memory processes, depends on the maturation of specific circuits within the hippocampus (Brun et al., 2002; Lavenex and Lavenex, 2013). The maturation of each individual circuit, and the emergence of their respective memory processes, culminates in the emergence of episodic memory (Brun et al., 2002; Lavenex and Lavenex, 2013). With respect to infantile amnesia, this developmental sequence is important because it demonstrates that the explicit memory system can be functional during early childhood, yet still be unable to support the formation and long-term retention of episodic memories. Because episodic memory is the type of memory on which our autobiography is built, the next section reviews its emergence and development during early childhood.

The Emergence of Episodic Memory

The term episodic memory was coined by Tulving (1972) and was originally defined solely in terms of the spatial (i.e., where) and temporal (i.e., when) components of an event (i.e., what). In order to capture the autobiographical nature of episodic memory, Tulving (1985) added the notion of autonoetic consciousness, a consciousness that endows one with the knowledge that what one remembers is a personal past episode rather than something one simply has information about (i.e., noetic knowledge). Obviously, defining episodic memory by the conscious experience of remembering, rather than the content (i.e., what, where, and when), presents a challenge for testing the construct in young children, whose verbal skills are still limited. To overcome this verbal obstacle, Tulving (2005) devised the 'spoon test.' The spoon test derives from an Estonian children's story, in which a little girl dreams about attending a friend's birthday party. Guests at the party are served the girl's favorite dessert; unbeknownst to the little girl, however, guests were required to bring their own spoon. Because she does not have a spoon, the little girl must stand by and watch as others enjoy the dessert. The next night, before going to bed, the girl tucks a spoon underneath her pillow, just in case she returns to the party in her dreams. According to Tulving (2005), placing the spoon underneath her pillow provides nonverbal evidence of the girl's autonoetic consciousness.

To investigate the emergence of episodic memory, Scarf et al. (2013) tested 3- and 4-year-old children on a novel spoon test. Children participated in two sessions separated by a delay. In Session 1, children were taken out to a sandbox and were told, "I saw a pirate around here earlier today and I think that he might have hidden his treasure in our sandbox. Can you help me find it?" The experimenter and the child then dug up a treasure chest only to find that it was locked. The experimenter then asked the child if he or she had a key and, when the child said no, the experimenter suggested that they go back inside. Twenty-four hours later, the child returned to the laboratory and was told that they were going to go back out to the sandbox but, before they went outside, they were allowed to choose one of three objects: a brightly colored ball, a key, and a brightly colored wind-up toy. The critical measure was whether children would correctly pick the key to take with them.

Of the 4-year-old children, 75% selected the key, a proportion significantly above chance (1 in 3, 33%). In contrast, the 3-year-old children performed at chance, with only 33% selecting the key. There are two possible reasons why the 3-year-old children failed to select the key. One possibility is that the single trial did not provide the 3-year-old children with enough time to encode the sandbox episode. A second possibility is that the 3-year-old children adequately encoded the sandbox episode, but that they failed to retain the memory across the delay. To determine the source of their difficulty, Scarf et al. (2013) tested separate groups of 3-year-old children either immediately after the sandbox episode or after a 15- or 30-min delay. Consistent with the view that 3-year-old children were able to encode the sandbox episode, when tested immediately, a significant number of 3-year-old children selected the key. Across the no-delay, 15-min delay, and 30-min delay conditions, there was a linear decrease in the proportion of children selecting the key such that at the 30-min delay, performance was at chance. This finding suggests that the source of difficulty for 3-year-old children on the spoon test is retaining the memory across significant delays. To test the limits of 4-year-old children's retention, a separate group was tested after a 1-week delay. Remarkably, 100% of the 4-year-old children selected the key after 1 week (**Figure 3**).

The fact that episodic memory emerges by 3 years of age, and that the long-term retention of explicit and episodic

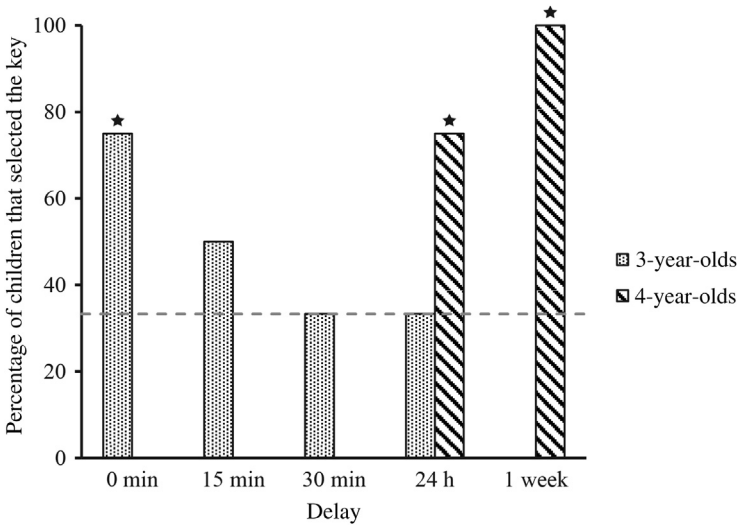

Figure 3 The performance of 3- and 4-year-old children on the spoon test as a function of delay.

memories develops considerably between 3 and 4 years of age, accords well with the average age of most adult's earliest memories. It is important to note, however, that although the seeds of episodic memory are clearly sown between 3 and 4 years of age, there are also significant individual differences in adult's earliest memories, suggesting that just the emergence of episodic memory per se does not explain the phenomenon of childhood amnesia entirely. The most consistent individual differences in the age of adults' earliest memories have been documented in cross-cultural studies and they provide, perhaps, the best evidence that the conversations that children have with adults are a causal factor in the development of autobiographical memory (Fivush et al., 1995; Nelson, 1990, 1993; Pillemer and White, 1989). New Zealand has the unique privilege of being home to the culture with the earliest memories of any culture studied thus far (Hayne and MacDonald, 2003; MacDonald et al., 2000). Māori are the native population of New Zealand and, on average, their first memories date from a few months before their third birthday. In contrast, New Zealand European's earliest memories, consistent with other European populations (Mullen, 1994; Mullen and Yi, 1995; Wang, 2001, 2006a,b), date from approximately $3^1/_2$ years of age (Hayne and MacDonald, 2003; MacDonald et al., 2000). The final section of this article explores how Māori cultural practices help adults encode and hold onto their earliest childhood memories.

Ngā Pūmahara o Ngā Tamarikitanga

The title of this section, *Ngā Pūmahara o Ngā Tamarikitanga*, is Māori for *Recollections of One's Childhood*. As noted above, Māori have the earliest memories of any culture studied to date, and this is a product of both how Māori parents recollect memories with their young children and also the types of memories they choose to recollect. Initial work comparing the way in which Māori and European New Zealand mothers talk to their children yielded the somewhat surprising finding that Māori mothers were less elaborative, when compared to New Zealand European mothers, when recollecting memories from the past (Hayne and MacDonald, 2003). Given the fact that Māori adults have significantly earlier memories than New Zealand European adults, this finding was inconsistent with work demonstrating that children of highly elaborative mothers generally have earlier first memories (Jack et al., 2009). From a cultural perspective, this finding was also surprising given that the Māori culture is renowned for its strong oral traditions and that their *whakapapa*, the oral transmission of family ancestry, is viewed as one of the richest in the world (Biggs, 1970).

One potential explanation for Hayne and MacDonald's (2003) finding is that the types of past events the parents were asked to talk about in the course of the research – recent events both the mother and child engaged in – were not representative of the types of events Māori parents generally talk about with their children. To investigate this possibility, Reese et al. (2008) compared Māori and New Zealand European mothers' elaborative style when they talked to their child about the child's birth and about a recent past event. Consistent with Hayne and MacDonald (2003), when discussing recent past events, New Zealand European mothers were more elaborative than Māori mothers. In contrast, when talking about the child's birth, Māori mothers were more elaborative, although only marginally so. Given that Māori and New Zealand European mothers are comparable in their general level of elaboration, another influence on Māori children's autobiographical memory development may be the content of their mother's elaborations. With respect to content, Reese et al. (2008) found that Māori and New Zealand European mothers talked with equal frequency about the people, places, and specific times of events, but that Māori mothers made more references to relative time and to internal states. Reference to

relative time and internal states may be especially important in children's autobiographical memory development. Young children have a great deal of difficulty with concepts of time, and the ability to organize episodic memories continues to develop well into early adolescence (Chen et al., 2013; Friedman, 2003). The use of relative time may help young children learn the chronological nature of autobiographical memory before they are able to master temporal concepts and it may also foster the ability to appreciate the order of events in individual episodic memories.

With respect to Māori mothers' references to internal states, outside of autobiographical memory development, there is a large body of work demonstrating that parent talk about mental states relates to young children's language and emotion understanding (e.g., Taumoepeau and Ruffman, 2006, 2008). Similarly, by referencing internal states, Māori mothers may contribute to children's developing sense of self. A sense of self is critical for both episodic and autobiographical memory, because it endows children with autonoetic awareness, allowing them to distinguish between memories they personally experienced (i.e., autonoetic) and those they did not (i.e., noetic). Although these accounts are speculative, in the absence of significantly greater levels of elaboration, the language used by Māori mothers when recollecting with their children, provides another potential explanation for Māori adults early autobiographical memories.

Summary

This article opened with a quote from T.S. Eliot's (1943) *Little Gidding* and it shall be ended on the same note. This article builds on earlier reviews by Rovee-Collier (1997) and Hayne (2004), providing an updated view on the current status of memory research in developmental psychology. Although empirical evidence for the principles of explicit memory development, the emergence of episodic memory, and the development of autobiographical memory has built over the past decade, the basic theoretical views have remained largely unchanged. With respect to the phenomenon of childhood amnesia, recent empirical work is providing us with a more fine-grained analysis of the phenomenon and how early theories may be integrated to provide a complete developmental account. In this context, T.S. Eliot's (1943) quote seems appropriate – the theories of childhood amnesia may not change, but each new finding makes the theories seem anew and in need of further exploration.

> We shall not cease from exploration
> And the end of all our exploring
> Will be to arrive where we started
> And know the place for the first time.
> T.S. Eliot (1943)

See also: Autobiographical Memory During Early Childhood and Cultural Variations; Autobiographical Memory; Childhood Memories; Early Memory Development and Cultural Variations; Episodic Memory; Hippocampal Amnesia; Memory Binding.

Bibliography

Atkinson, J., 2000. The Developing Visual Brain. Oxford University Press, New York.

Barr, R., Dowden, A., Hayne, H., 1996. Developmental changes in deferred imitation by 6-to 24-month-old infants. Infant Behavior and Development 19, 159–170.

Bauer, P.J., Mandler, J.M., 1989. One thing follows another: effects of temporal structure on 1-to 2-year-olds' recall of events. Developmental Psychology 25, 197–206.

Berlyne, D.B., 1958. The influence of the albedo and complexity of stimuli on visual fixation in the human infant. British Journal of Psychology 49, 315–318.

Biggs, B., 1970. The Maori People in the 1960's: A Symposium. The Maori language past and present. B. & J. Paul, Auckland, NZ, 65–84.

Brun, V.H., Otnæss, M.K., Molden, S., Steffenach, H.-A., Witter, M.P., Moser, M.-B., et al., 2002. Place cells and place recognition maintained by direct entorhinal-hippocampal circuitry. Science 296, 2243–2246.

Campbell, B.A., Jaynes, J., 1966. Reinstatement. Psychological Review 73, 478–480.

Carver, L.J., Bauer, P.J., 2001. The dawning of a past: the emergence of long-term explicit memory in infancy. Journal of Experimental Psychology: General 130, 726–745.

Chen, Y., McAnally, H.M., Reese, E., 2013. Development in the organization of episodic memories in middle childhood and adolescence. Frontiers in Behavioral Neuroscience 7, 1–9. http://dx.doi.org/10.3389/fnbeh.2013.00084.

Cohen, N.J., Squire, L.R., 1980. Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that. Science 210, 207–210.

Corkin, S., 1968. Acquisition of motor skill after bilateral medial temporal-lobe excision. Neuropsychologia 6, 255–265.

Corkin, S., 1984. Lasting consequences of bilateral medial temporal lobectomy: clinical course and experimental findings in HM. Paper presented at the Seminars in Neurology.

Corkin, S., 2002. What's new with the amnesic patient HM? Nature Reviews Neuroscience 3, 153–160.

Davis, J.M., Rovee-Collier, C.K., 1983. Alleviated forgetting of a learned contingency in 8-week-old infants. Developmental Psychology 19, 353–365.

Davis, N., Gross, J., Hayne, H., 2008. Defining the boundary of childhood amnesia. Memory 16, 465–474.

Eliot, T.S., 1943. Little Gidding. Faber & Faber.

Fantz, R.L., 1956. A method for studying early visual development. Perceptual and Motor Skills 6, 13–15.

Fivush, R., Haden, C., Adam, S., 1995. Structure and coherence of preschoolers' personal narratives over time: Implications for childhood amnesia. Journal of Experimental Child Psychology 60, 32–56.

Freud, S., 1963. New introductory lectures on psycho-analysis. In: Strachey, J. (Ed.), The Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 15. Hogarth Press, London, England, pp. 199–201.

Friedman, W.J., 2003. The development of a differentiated sense of the past and the future. Advances in Child Development and Behavior 31, 229–269.

Graf, P., Schacter, D.L., 1985. Implicit and explicit memory for new associations in normal and amnesic subjects. Journal of Experimental Psychology: Learning, Memory, and Cognition 11, 501.

Greco, C., Rovee-Collier, C., Hayne, H., Griesler, P., Earley, L., 1986. Ontogeny of early event memory: I. Forgetting and retrieval by 2-and 3-month-olds. Infant Behavior and Development 9, 441–460.

Hartshorn, K., Rovee-Collier, C., 1997. Infant learning and long-term memory at 6 months: a confirming analysis. Developmental Psychobiology 30, 71–85.

Hartshorn, K., Rovee-Collier, C., Gerhardstein, P., Bhatt, R.S., Wondoloski, T.L., Klein, P., et al., 1998. The ontogeny of long-term memory over the first year-and-a-half of life. Developmental Psychobiology 32, 69–89.

Hayne, H., 2004. Infant memory development: Implications for childhood amnesia. Developmental Review 24, 33–73.

Hayne, H., Boniface, J., Barr, R., 2000. The development of declarative memory in human infants: age-related changes in deferred imitation. Behavioral Neuroscience 114, 77.

Hayne, H., Greco, C., Earley, L., Griesler, P., Rovee-Collier, C., 1986. Ontogeny of early event memory: II. Encoding and retrieval by 2-and 3-month-olds. Infant Behavior and Development 9, 461–472.
Hayne, H., MacDonald, S., 2003. The socialization of autobiographical memory in children and adults: the roles of culture and gender. In: Fivush, R., Haden, C.D. (Eds.), Autobiographical Memory and the Construction of a Narrative Self: Developmental and Cultural Perspectives. Erlbaum, Mahwah, NJ, pp. 99–120.
Herbert, J., Hayne, H., 2000. Memory retrieval by 18–30-month-olds: age-related changes in representational flexibility. Developmental Psychology 36, 473–484.
Hill, W., Borovsky, D., Rovee-Collier, C., 1988. Continuities in infant memory development over the first half-year. Developmental Psychobiology 21, 43–62.
Imuta, K., Scarf, D., Hayne, H., 2013. The effect of verbal reminders on memory reactivation in 2-, 3-, and 4-year-old children. Developmental Psychology 49, 1058–1065. http://dx.doi.org/10.1037/a0029432.
Jack, F., Hayne, H., 2007. Eliciting adults' earliest memories: does it matter how we ask the question? Memory 15, 647–663.
Jack, F., Hayne, H., 2010. Childhood amnesia: empirical evidence for a two-stage phenomenon. Memory 18, 831–844.
Jack, F., MacDonald, S., Reese, E., Hayne, H., 2009. Maternal reminiscing style during early childhood predicts the age of adolescents' earliest memories. Child Development 80, 496–505.
Jack, F., Simcock, G., Hayne, H., 2012. Magic memories: young children's verbal recall after a 6-year delay. Child Development 83, 159–172.
Lavenex, P., Lavenex, P.B., 2013. Building hippocampal circuits to learn and remember: insights into the development of human memory. Behavioural Brain Research.
MacDonald, S., Uesiliana, K., Hayne, H., 2000. Cross-cultural and gender differences in childhood amnesia. Memory 8, 365–376.
McKee, R.D., Squire, L.R., 1993. On the development of declarative memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 19, 397.
Meltzoff, A.N., 1995. What infant memory tells us about infantile amnesia: long-term recall and deferred imitation. Journal of Experimental Child Psychology 59, 497–515.
Milner, B., Corkin, S., Teuber, H.-L., 1968. Further analysis of the hippocampal amnesic syndrome: 14-year follow-up study of HM. Neuropsychologia 6, 215–234.
Morgan, K., Hayne, H., 2006a. Age-related changes in memory reactivation by 1- and 2-year-old human infants. Developmental Psychobiology 48, 48–57.
Morgan, K., Hayne, H., 2006b. The effect of encoding time on retention by infants and young children. Infant Behavior and Development 29, 599–602.
Morgan, K., Hayne, H., 2007. Nonspecific verbal cues alleviate forgetting by young children. Developmental Science 10, 727–733.
Morgan, K., Hayne, H., 2011. Age-related changes in visual recognition memory during infancy and early childhood. Developmental Psychobiology 53, 157–165.
Mullen, M.K., 1994. Earliest recollections of childhood: a demographic analysis. Cognition 52, 55–79.
Mullen, M.K., Yi, S., 1995. The cultural context of talk about the past: Implications for the development of autobiographical memory. Cognitive Development 10, 407–419.
Munoz, M., Chadwick, M., Perez-Hernandez, E., Vargha-Khadem, F., Mishkin, M., 2011. Novelty preference in patients with developmental amnesia. Hippocampus 21, 1268–1276.
Nelson, C.A., 1995. The ontogeny of human memory: a cognitive neuroscience perspective. Developmental Psychology 31, 723.
Nelson, K., 1990. Remembering, forgetting, and childhood amnesia. In: Fivush, R., Hudson, J.A. (Eds.), Knowing and Remembering in Young Children. Cambridge University Press, Cambridge, pp. 301–316.
Nelson, K., 1993. The psychological and social origins of autobiographical memory. Psychological Science 4, 7–14.
Pascalis, O., Hunkin, N., Holdstock, J., Isaac, C., Mayes, A., 2004. Visual paired comparison performance is impaired in a patient with selective hippocampal lesions and relatively intact item recognition. Neuropsychologia 42, 1293–1300.
Pillemer, D.B., White, S.H., 1989. Childhood events recalled by children and adults. Advances in Child Development and Behavior 21, 297–340.
Reese, E., Hayne, H., MacDonald, S., 2008. Looking back to the future: Māori and Pakeha mother–child birth stories. Child Development 79, 114–125.
Richmond, J., Nelson, C.A., 2007. Accounting for change in declarative memory: a cognitive neuroscience perspective. Developmental Review 27, 349–373.
Roediger, H.L., 1984. Does current evidence from dissociation experiments favor the episodic/semantic distinction? Behavioral and Brain Sciences 7, 252–254.
Rovee, C.K., Rovee, D.T., 1969. Conjugate reinforcement of infant exploratory behavior. Journal of Experimental Child Psychology 8, 33–39.
Rovee-Collier, C., 1997. Dissociations in infant memory: rethinking the development of implicit and explicit memory. Psychological Review 104, 467–498.
Rovee-Collier, C., Cuevas, K., 2009. Multiple memory systems are unnecessary to account for infant memory development: an ecological model. Developmental Psychology 45, 160–174.
Rovee-Collier, C., Hayne, H., Colombo, M., 2000. The Development of Implicit and Explicit Memory, vol. 24. John Benjamins Publishing, Amsterdam, The Netherlands.
Rovee-Collier, C.K., Sullivan, M., Enright, M., Lucas, D., Fagen, J., 1980. Reactivation of infant memory. Science 208, 1159–1161.
Scarf, D., Gross, J., Colombo, M., Hayne, H., 2013. To have and to hold: episodic memory in 3- and 4-year-old children. Developmental Psychobiology 55, 125–132.
Schacter, D.L., Moscovitch, M., 1984. Infants, amnesics, and dissociable memory systems. In: Moscovitch, M. (Ed.), Advances in the Study of Communication and Affect, vol. 9. Springer, pp. 173–216.
Scoville, W.B., Milner, B., 1957. Loss of recent memory after bilateral hippocampal lesions. Journal of Neurology, Neurosurgery, & Psychiatry 20, 11–21.
Sheingold, K., Tenney, Y.J., 1982. Memory for a Salient Childhood Event. In: Neisser, U. (Ed.), Memory Observed. Cambridge University Press, New York, pp. 201–212.
Snodgrass, J.G., 1989. How many memory systems are there really?: some evidence from the picture fragment completion task. In: Isawa, C. (Ed.), Current Issues in Cognitive Processes: The Tulane Flowerree Symposium on Cognition. Erlbaum, Hillsdale, NJ, pp. 135–173.
Sokolov, E., 1963. Perception and the Conditioned Reflex. Pergamon Press, Oxford.
Spear, N., Parsons, P., 1976. Analysis of a reactivation treatment: ontogenetic determinants of alleviated forgetting. In: Medin, D.L., Roberts, W.A., Davis, R.T. (Eds.), Processes of Animal Memory. Erlbaum, Hillsdale, NJ, pp. 135–165.
Squire, L.R., 1992. Declarative and nondeclarative memory: multiple brain systems supporting learning and memory. Journal of Cognitive Neuroscience 4, 232–243.
Taumoepeau, M., Ruffman, T., 2006. Mother and infant talk about mental states relates to desire language and emotion understanding. Child Development 77, 465–481.
Taumoepeau, M., Ruffman, T., 2008. Stepping stones to others' minds: maternal talk relates to child mental state language and emotion understanding at 15, 24, and 33 months. Child Development 79, 284–302.
Tulving, E., 1972. Episodic and semantic memory. In: Tulving, E., Donaldson, W. (Eds.), Organization of Memory. Academic Press, New York, pp. 381–402.
Tulving, E., 1983. Elements of Episodic Memory. Oxford University Press, New York.
Tulving, E., 1985. Memory and consciousness. Canadian Psychology/Psychologie Canadienne 26, 1.
Tulving, E., 2005. Episodic memory and autonoesis: uniquely human. In: Terrace, H., Metcalfe, J. (Eds.), The Missing Link in Cognition. Oxford University Press, New York, pp. 4–56.
Tulving, E., Schacter, D.L., 1990. Priming and human memory systems. Science 247, 301–306.
Vander Linde, E., Morrongiello, B.A., Rovee-Collier, C., 1985. Determinants of retention in 8-week-old infants. Developmental Psychology 21, 601–613.
Wang, Q., 2001. Culture effects on adults' earliest childhood recollection and self-description: implications for the relation between memory and the self. Journal of Personality and Social Psychology 81, 220–233.
Wang, Q., 2006a. Earliest recollections of self and others in European American and Taiwanese young adults. Psychological Science 17, 708–714.
Wang, Q., 2006b. Relations of maternal style and child self-concept to autobiographical memories in Chinese, Chinese immigrant, and European American 3-year-olds. Child Development 77, 1794–1809.

Relevant Websites

http://www.otago.ac.nz/psychology/staff/harlenehayne.html – University of Otago, Department of Psychology.
http://www.otago.ac.nz/psychology/staff/otago025837.html – University of Otago, Department of Psychology.

Memory in the Bee

Randolf Menzel, Institut für Biologie, Freie Universität Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Honeybees contradict the notion that insects elicit relatively inflexible, stereotypical behaviors. Their social life, navigation skills, symboliclike communication, rich learning repertoire, and dynamic forms of memory are embodied in a brain containing only one million neurons. Elemental and nonelemental forms of associative learning allow for context-dependent forms of learning, concept formation, and rule extraction. Neural plasticity correlating with learning can be traced to single identified neurons and small neural nets. Memory is formed by five sequential and parallel phases that are linked by specific forms of consolidation.

Navigation and Communication in a Social Insect, The Honeybee

The social life of the honeybee colony forms the evolutionary framework for the individual animal's behavior and is crucial for the survival of each bee (von Frisch 1967; Lindauer 1967). Social bees are exposed to a large and unpredictable range of environmental conditions and thus learning plays an important role both in the social context and in individual adaptations. The location of the nest site needs to be learned relative to landmarks within the reference frame of the sun compass. Changes to the society's conditions have to be recognized and used to control individual behavior. Communication about potential nest sites and food, water, and resin resources involve transmitting symbolic signals indicating distance and direction to the location of the respective object. The behavioral routine of communication, called the waggle dance, is performed in the dark on the vertical comb and encodes direction to the location as measured during outbound flights relative to the sun azimuth in the movement of the waggle phase relative to gravity. Distance as measured visually during flight is encoded in the speed with which the dance rounds are performed. The following bees (recruits) learn about the location and the quality of the respective resource and apply this information on their outbound search flights. Since they integrate dance communicated information into their navigation memory of the environment and perform novel short cutting flights between experienced and communicated locations, it is concluded that their spatial memory has the structure of a cognitive map (Menzel et al., 2011).

Learning

Learning Under Natural Conditions

Most studies of learning and memory formation in bees have focused on operant learning in the context of food collection. When a searching bee discovers a source of sucrose solution (under natural conditions; nectar), it quickly learns to associate the surrounding visual and olfactory signals with the reward. Acquisition functions indicate that olfactory stimuli (e.g., floral odorants) are learned within one or two learning trials, colors within one to five learning trials, and black-and-white patterns from the fifth learning trial onward (Menzel, 1985). Multiple learning paradigms have been studied under such conditions. Reversal learning, overlearning, and multireversal learning are effective learning phenomena in the bee. Reward schedules (e.g., continuous, fixed-ratio, and fixed-interval reinforcement schedules) are learned well. Additional operant learning phenomena have been studied in the tradition of experimental psychology and are reported by Bitterman (1988).

Visual categorization has been studied in free-flying honeybees trained to different visual stimuli, e.g., pattern orientation or bilateral symmetry. Bees choose the appropriate patterns despite the novelty of the structural details. The cases of categorical learning studied so far (orientation and bilateral symmetry of patterns) have been interpreted to demonstrate that bees focus their attention on a *single feature at a time* (orientation or bilateral symmetry) to solve the problem. However, Stach et al. (2004) and Zhang et al. (2004) showed that bees not only focus on a single feature but also assemble different features to build a generic pattern representation, which could be used to respond appropriately to novel stimuli sharing such a basic layout. These results show that honeybees extract regularities in their visual environment and establish correspondences among correlated features such that they generate a large set of object descriptions from a finite set of elements. The authors suggest that configurational properties of the figures, in which specific features such as circular symmetry, angular periodicity, bilateral symmetry, and the presence of a horizontal, high-contrast edge, the horizon, are integrated, could help the bees to classify the different stimuli.

Rule learning presupposes positive transfer of an appropriate response from a known set to a novel set of stimuli. One example is the detection of sameness and of difference. These rules are uncovered through delayed-matching-to-sample (DMTS) and the delayed-non-matching-to-sample (DNMTS) experiments, respectively. DMTS and DNMTS are paradigms frequently used to characterize working memory. Honeybees foraging in a Y-maze learn both rules when trained to either pairs of colors or patterns (Giurfa et al., 2001). Bees trained with colors and presented in transfer tests with gratings that they have not experienced before solve the problem and choose the grating identical to the sample at the entrance to

the maze. Similarly, bees trained with the gratings and tested with colors in transfer tests also solve the problem and choose the novel colors corresponding to those of the sample grating at the maze entrance. Working memory for the sample underlying solving a DMTS task is around 5 s (Zhang et al., 2005). Moreover, bees trained in a DMTS task can learn to pay attention to one of two different samples presented successively in a flight tunnel (either to the first or to the second) and can transfer the learning of this sequence weight to novel samples (Zhang et al., 2005).

Contextual learning is a term widely used for describing conditional discrimination, which can be subsumed in several cases of the so-called occasion setting problem. In these conditions, a given stimulus, the occasion setter, informs the animal about the outcome of its choice. For example, bees were trained to fly to two feeding stations, one in the morning and one in the afternoon. Thus, depending on the time of day, bees flew in one direction or the other, showing that temporal information can act as the occasion setter for a sensory-motor routine. Olfactory cues can also act as occasion setters for the choice of appropriate navigation vectors. Bees trained to forage at two feeders, placed at different outdoor locations and carrying different scents, recall the locations of these food sites and fly to them when the respective scent was blown into the hive. These results indicate that familiar scents can trigger navigational memories in order to define appropriate context-dependent responses.

Classical conditioning of innate responses is a most convenient way to study the behavioral and neural mechanisms of associative learning. In hungry honeybees, a sucrose stimulus at the antennae reliably releases the extension of its proboscis (PER: proboscis extension response). Conditioning to olfactory or mechanical stimuli as a conditioned stimuli (CS) is very effective (Menzel and Giurfa, 2001). The associative nature of PER conditioning was established by demonstrating that only forward pairing of CS–unconditioned stimuli (US) sequences is effective, by various control groups (unpaired CS and US, CS- or US-only presentations), and by differential conditioning of the two olfactory stimuli (Bitterman et al., 1983). The predictive value of the CS depends on the reliability with which it is causally related to the US. In differential conditioning, the reversal to the initially unpaired stimulus is slower after more frequent unreinforced preexposures than after a lesser number of preexposures. Overshadowing and blocking experiments are used to characterize the informational content of the CS. It was concluded that attentiveness is not a limiting factor, and that the salience or associability of a stimulus is indeed an important factor.

In studies of nonelemental learning, paradigms have been developed for which the associative strength of a specific event, stimulus or reaction, is ambiguous and therefore cannot be predictive for solving a problem. In the honeybee, several recent studies have addressed the issue of elemental vs nonelemental learning, using visual training of free-flying animals or olfactory conditioning of harnessed animals. In both modalities, bees are able to solve a biconditional discrimination (AB+, CD+, AC−, BD−). This capacity demonstrates that both visual and olfactory compounds are learned under certain circumstances as entities different from the simple sum of their elements. Similarly, negative patterning experiments (A+, B+, AB−) have also been performed in both the visual and the olfactory modality. Such discrimination can only be explained if the compound AB is treated as being different from the simple sum of its elements. The results are interpreted according to a modified unique cue hypothesis (Lachnit et al., 2004).

Neural Correlates of Learning and Memory Formation

The neural correlates of learning and memory in bees can be traced to particular regions of the brain, and in some exceptional cases to unique identified neurons. Recording these neural correlates is possible under conditions when the animal associates an odor with sucrose reward and retrieves the memory later (PER conditioning paradigm, see above). This form of Pavlovian classical conditioning is a robust learning phenomenon of bees that are constraint to a holder and the brain is exposed to intra or extracellular recordings, Ca^{2+} imaging and manipulation of molecular targets.

Insect brains are rather small and equipped with neurons that are at least partially identifiable at the single-neuron level. The honeybee brain has a volume of about 1 mm^3 and contains approximately 950 000 neurons which are organized in well-structured neuropils. The small size of the brain allows to fully reconstruct in 3D the spatial arrangement of the neuropils with the resolution of the light microscope and to incorporate the connectivity of neural nets into a virtual atlas of the brain (http://www.neurobiologie.fu-berlin.de/beebrain). An essential structure for associative olfactory learning and memory storage is the paired mushroom body (MB). The MB is a high order integration site where processed sensory inputs of all modalities converge. Using the olfactory PER conditioning paradigm it was shown that the memory trace is distributed across the olfactory pathway involving the antennal lobes (AL) and the MB. Reversible blocking of neural activity in the AL and the MB after a single olfactory learning trial indicated that the AL contributes to the consolidation of a memory trace only during the first 1–2 min after a single conditioning trial, and the MB over a period of up to 10 min. Furthermore, it was found that blocking neural activity in the AL could prevent formation of olfactory memory, but applying the same procedure to the MB could block memory formation only partially indicating that consolidation to lasting memory is distributed between these two neuropils and involves different time-dependent processes. Additional information about the localization and dynamic of the respective memory traces come from manipulation of cellular processes and recordings of neural activity (see below).

Neural substrates of learning and memory are expected to be related to plasticity of neural functions, and the search for such plasticity should ideally be performed on the single-neuron level under conditions when the animal learns, consolidates memory and retrieves it. For such measurements, intra and extracellular recordings have been successfully used to monitor neural processes during learning and memory retrieval at the single-neuron level. In two cases, the ventral unpaired median neuron 1 of the maxilary neuromere

(VUM_{mx1}) neuron and the pedunculus extrinsic neuron 1 (PE1) neuron in the bee brain, has it been possible to return to the same individual neuron in subsequent recordings from different animals exposed to associative learning schedules.

A particularly striking neuron in the bee brain is VUM_{mx1}, which was shown to serve the function of a value system (Hammer, 1993). Activity in this neuron constitutes the neuronal representation of the food reward in appetitive associative olfactory learning. Behavioral learning of an olfactory stimulus can be induced by substituting the sucrose reward in PER conditioning by an artificial depolarization of VUM_{mx1} immediately after olfactory stimulation (forward pairing). If depolarization precedes olfactory stimulation (backward pairing) no learning is observed. The same forward–backward effect is seen when sucrose is used as the reward under similar experimental conditions. Thus VUM_{mx1} constitutes the neural correlate of the US in associative olfactory learning.

Intracellular recording of VUM_{mx1} during differential conditioning to two odors, a forward-paired one (CS+), and a backward-paired one (CS−), leads to an enhanced response of VUM_{mx1} to CS+ and to an inhibitory response to CS−. Thus VUM_{mx1} learns about the forward-paired odor, a fact that can support second-odor conditioning, blocking, and the phenomena involved in reconsolidation (see below). The response of VUM_{mx1} to an expected reward is strongly reduced, and even inhibited during sucrose stimulation. The response to sucrose after CS− (no reward expected) remains normal. These properties provide the neural circuitry with an error signal for learning similar to what was found for dopamine neurons of the basal ganglion in the mammalian brain (Schultz, 2006).

It is yet unknown whether VUM_{mx1} itself or the other 14 identified VUM neurons of the subesophageal ganglion belong to a general modulatory pathway, which may also serve non-associative forms of plasticity and arousal. Furthermore, since VUM_{mx1} converges only with the olfactory pathway and not with other sensory pathways in the bee brain, it may be the specialized reward system for olfactory cues, and other value systems may exist for visual and other forms of learning.

A second single identified neuron with learning related plasticity is the PE1. PE1 reduces selectively its responses to CS+ during odor conditioning. PE1 receives predominantly excitatory input from Kenyon cells (MB intrinsic neurons) across the whole peduncle. Inhibitory input to PE1 can be traced to side branches of GABA-ir profiles of the protocerebral calycal tract (PCT) (Okada et al., 2007). PCT neurons both enhance or reduce their responses to the learned odor. It is concluded that PE1 receives input from the subpopulation of PCT neurons that express learning related enhancement providing an example of associative plasticity on the network level as a correlate of learning. Ultimately memory traces have to be localized in neural nets, and this is accomplished by Ca^{2+} imaging approaches, which also allow to identify the contributing neurons. Olfactory memory traces were localized both in the AL and the MB by Ca^{2+} imaging (Szyszka et al., 2008). It will be the task for future research to identify the populations of synapses, their distributions, and associative properties that contribute to the different forms of memory in the bee brain.

Memory

Even an insect with such a little brain consolidates memory via several sequential and parallel processes, recalls memory from different stores, and organizes memory according to the behavioral needs. In this respect its memory structures resembles features that can be considered to reflect basic components of memory formation and retrieval.

Sequential Memory Processing

A stable, lifelong memory is formed after even only a few learning trials in free-flying bees (Menzel, 1999). The retention curve after a single learning trial follows a biphasic time course, which indicates that memory and retrieval is at first very high, fades to a minimum 2–3 min after learning, rises again within the following 20 min, and stays high for a number of hours; this is true both in PER conditioning and for free-flying bees. The early memory phase in the minutes range is particularly susceptible to both extinction and reversal learning, whereas the consolidated memory is much more resistant. The sensitizing effect of the appetitive stimulus overlaps with the early high response level, indicating that, at least in PER conditioning, the nonassociative memory initiated by the US contributes substantially to the response probability immediately after the single learning trial. Amnestic treatments (cooling, narcosis, weak electroconvulsive brain stimulation), if applied within 2–3 min after the single-trial learning, erase the memory trace, indicating a susceptible short-term memory (STM). The memory initiated by a single conditioning trial consolidates within a few minutes so that it becomes resistant to retrograde amnestic treatments. The time course of retrograde amnestic sensitivity is independent of the kind of learning (operant color learning, olfactory PER conditioning), but depends on the particular structure of the brain where the treatment is applied. Several conditioning trials within STM make the memory trace immune to amnestic treatment.

Several forms of longer-lasting memory have been discerned on the basis of behavioral, biochemical, and pharmacological experiments performed with olfactory PER conditioning. The respective time courses and transitions between five forms of memory have been related to the temporal sequence during the foraging cycle. (1) Associative induction and early short-term memory (eSTM): eSTM covers the time window during which foraging bees can expect to be exposed to the same stimuli, since flowers most frequently occur in patches. No specific choices need to be performed at this time, and general arousal (depending on the strength of the US) will suffice to control whether the animal stays in the patch or postpones choices for a later time. (2) Late short-term memory (lSTM): the transition to the selective associative memory trace during lSTM is a rather slow process after a single learning trial lasting up to several minutes, and happens very quickly after multiple learning trials. Thus consolidation is both time- and event-dependent, where events have to be associative experiences and not just CS or US repetitions. The behavioral relevance of these findings for foraging behavior has been related to the temporal separation between intra and interpatch visits. (3) Midterm memory (MTM): at the beginning of MTM, behavior is controlled by consolidated, highly specific memory. At this stage, memory is more resistant to

extinction, conflicting information and elapsing time, and some information about the context dependencies may have already been stored. Under natural conditions bees have usually returned to the hive and departed on a new foraging bout within the time window of MTM. (4) Early long-term memory, and (5) late long-term memory: LTM requires multiple learning trials, indicating that specific information, which can be extracted only from multiple experiences (signal reliability, context dependence) controls the transfer to LTM.

Processes of memory consolidation bridge the five different forms of memory phases, and these processes can be partially interpreted on the level of cellular and neural events. Most information exists for olfactory reward conditioning in harnessed bees (Menzel et al., 2007).

Associative Induction and Early Short-Term Memory (eSTM)

An associative learning trial leads to associative induction and eSTM in the range of seconds. This memory is initially localized in the primary sensory neuropil, the antennal lobe, and is highly dominated by appetitive arousal and sensitization induced by the US: sucrose. Thus eSTM is restricted to the first synaptic connections in the sensory pathway. It is rather unspecific and imprecise. The associative component, which connects specific stimuli with motor actions, already exists at a low level, and develops later over a period of several minutes.

At the cellular level, stimulus association is reflected in the convergence of excitation of the pathways for the CS and the US. There are three neuroanatomical convergence sites of these pathways (the antennal lobe, AL, the lip region of the mushroom bodies, MB, and the lateral protocerebrum). Two of these sites can independently form an associative memory trace for an odor stimulus. The dynamics of these two traces are different: the AL establishes the trace quickly, the MB stepwise by a consolidation process in the range of minutes during the transition to a late form of STM (lSTM). The molecular substrates for the formation of the associative links are unknown. Neither a glutamate receptor, as shown for the hippocampus, nor adenylylcyclase, as proposed for *Aplysia*, appear to play a role in the bee. However, cyclic adenosine monophosphate (cAMP) upregulation of protein kinase A (PKA) during STM is necessary for memory transition to MTM and long-term memory (LTM), and an increase in PKA activity is specifically connected with associative trials, indicating PKA's role in consolidation to lasting memories established during consolidation (Müller, 2012).

Late Short-Term Memory (lSTM)

The transition to the associative memory trace during lSTM is a rather slow process after a single learning trial lasting up to several minutes, and a quick one after multiple learning trials. Thus, consolidation is both time- and event-dependent, where events must be associative experiences and not just CS or US repetitions. The MB appear to be selectively involved.

Midterm Memory (MTM)

MTM is physiologically characterized by a wave of protease-dependent protein kinase C (PKC) activity in the antennal lobe. It might be, therefore, that the primary sensory neuropil is a substrate for MTM. Since the MB appear to be the critical substrate for consolidation during lSTM, it has been speculated that the MTM trace in the AL is established under the guidance of the MB. The MB provides the information that relates the memory traces in the primary sensory neuropils to context stimuli across modalities. Output neurons of the MB feedback to the AL.

Long-Term Memory (LTM)

Consolidation into lasting memory depends not only on PKA activity but also on NO synthase activity, indicating that both cAMP and cyclic guanosine monophosphate (cGMP) are important second messengers for consolidation. Indeed, activating cAMP dependent PKC in the antennal lobe after a single learning trial leads to stable LTM (Müller, 2000). LTM is divided into two forms, an early LTM (eLTM, 1–3 days) characterized by prolonged PKC activity and translation dependence, and late LTM (lLTM, >3 days) depending on transcription. The transition from lSTM to both forms of LTM appears to be independent of MTM, because inhibiting the characteristic substrate of MTM (protease-dependent enhancement of PKC activity) does not prevent eLTM and lLTM being formed.

Structural changes in the connectivity between neurons have been proposed as the substrates for LTM in vertebrates and invertebrates and are believed to be at least one target of the translation and transcription dependence of LTM formation. In bees, direct evidence for LTM-related structural changes is lacking, but measurements of MB subcompartment volume indicate that more experienced bees have bigger volumes and more elaborate dendrite branches.

Glutamate Transmission and Memory Formation

Glutamate is the major excitatory transmitter in mammals and plays a central role in neuronal plasticity in vertebrates. Its function in the insect brain, however, is only poorly understood. Glutamate receptors have been identified in honeybees, and manipulation of glutamate function using pharmacological tools and RNAi supports a function in memory consolidation processes (Mussig et al., 2010). Uncaging glutamate immediately after a single-trial training in the MB improves the formation of an LTM, and thus mimics the effect of a strong training protocol. Uncaging glutamate shortly before weak training has no effect, pointing to a defined function of glutamate with respect to memory processing rather than learning. This function of glutamate is restricted to the MB circuitry, since glutamate release in the AL does not improve learning or retention. Taken together, glutamate action appears to be restricted to the MBs, where it contributes to induction of LTM in a time window shortly after training. Although it is still unclear whether this late memory phase is mechanistically identical with the LTM induced by a strong training (three-trial conditioning), the findings provide an example for parallel molecular processes underlying learning and memory formation *in vivo* occurring in different neural circuits.

In summary, the bee brain of merely 1 mm^3 and a total of 950 000 neurons, establishes multiple and distributed memory traces not very different from mammalian memory in the general temporal dynamics, characteristics of contents,

and cellular substrates. It may thus serve as a suitable model for the study of memory structure and formation.

Reconsolidation

When memory is retrieved from a remote store by exposing the animal to the learned stimulus without reinforcement (extinction trial) two processes are initiated: extinction learning, due to the fact that the stimulus originally associated with reinforcement is now presented without it, and reminder learning, which may either recruit the original stable memory about the learned stimulus and bring it back into an unstable form, or it initiates a new learning trial based on the fact that a learned stimulus also activates the reward system and thus initiates new learning (see above). Both extinction and reminder learning are followed by their respective consolidation processes: consolidation of an extinction memory and reconsolidation of the reminder memory. In the honeybee the balance between the extinction and reminder learning processes and their respective consolidation processes depends on the number of extinction trials. Extinction learning is dominant after two extinction trials, whereas reminder learning is expressed in a recovery from extinction and is induced by many (e.g., five) extinction trials. Consolidation of reminder learning (reconsolidation) depends on protein synthesis, which indicates that a new learning process is going on, as shown in the vertebrate studies (Stollhoff et al., 2005).

The results on reconsolidation in the honeybee can be explained by reference to the properties of the reward system in the bee brain, and do not require the assumption that reconsolidation is based on processes by making the consolidated memory trace again sensitive to novel consolidation (Eisenhardt and Menzel, 2007). It was found that the reward neuron VUM_{mx1} (Hammer, 1993) not only codes for appetitive reinforcement of odors, but also learns to respond to the conditioned odor stimulus. An extinction trial indeed activates the VUM neuron. This means that an extinction trial will lead to new appetitive learning without an external reinforcing stimulus but through in internal reinforcement signal from the VUM neuron. The data confirm this interpretation: two forms of learning can be separated, one based on the lack of external reinforcement (extinction learning) and the other based on internal reinforcement (reminder learning). These conditions are best captured by the 'internal reinforcement hypothesis,' which states that the balance between extinction and reminder learning depends on the relative strength of the respective reinforcements. The balance will be shifted to the internal reinforcement with more unrewarded presentations of the learned stimulus.

See also: Genes and Behavior: Lessons from the Fruit Fly; Genetic Approaches to Memory; Learning and Memory, Neural Basis of.

Bibliography

Bitterman, M.E., 1988. Vertebrate–invertebrate comparisons. NATO ASI Series Intelligence and Evolutionary Biology 17, 251–275.

Bitterman, M.E., et al., 1983. Classical conditioning of proboscis extension in honeybees (*Apis mellifera*). Journal of Comparative Psychology 97, 107–119.

Eisenhardt, D., Menzel, R., 2007. Extinction learning, reconsolidation and the internal reinforcer hypothesis. Neurobiology of Learning and Memory 87, 167–173.

Giurfa, M., et al., 2001. The concepts of 'sameness' and 'difference' in an insect. Nature 410, 930–933.

Hammer, M., 1993. An identified neuron mediates the unconditioned stimulus in associative olfactory learning in honeybees. Nature 366, 59–63.

Lachnit, H., Giurfa, M., Menzel, R., 2004. Odor processing in honeybees: is the whole equal to, more than, or different from the sum of its parts?. In: Slater, P.J.G., et al. (Eds.), Advances in the Study of Behavior, vol. 34. Elsevier, Amsterdam, pp. 241–264.

Lindauer, M., 1967. Recent advances in bee communication and orientation. Annual Review of Entomology 12, 439–470.

Menzel, R., 1985. Learning in honey bees in an ecological and behavioral context. In: Hölldobler, B., Lindauer, M. (Eds.), Experimental Behavioral Ecology and Sociobiology. Gustav Fischer Verlag, Stuttgart, pp. 55–74.

Menzel, R., 1999. Memory dynamics in the honeybee. Journal of Comparative Physiology A 185, 323–340.

Menzel, R., Brembs, B., Giurfa, M., 2007. Cognition in invertebrates. In: Kaas, J.H. (Ed.), Evolution of Nervous Systems, Evolution of Nervous Systems in Invertebrates, vol. II. Academic Press, Oxford, pp. 403–422.

Menzel, R., Giurfa, M., 2001. Cognitive architecture of a mini-brain: the honeybee. Trends in Cognitive Science 5, 62–71.

Menzel, R., Kirbach, A., Haass, W.-D., Fischer, B., Fuchs, J., Koblofsky, M., Lehmann, K., Reiter, L., Meyer, H., Nguyen, H., Jones, S., Norton, P., Greggers, U., 2011. A common frame of reference for learned and communicated vectors in honeybee navigation. Current Biology 21 (8), 645–650.

Müller, U., 2000. Prolonged activation of cAMP-dependent protein kinase during conditioning induces long-term memory in honeybees. Neuron 27, 159–168.

Müller, U., 2012. Molecular biology of learning and memory in the honey bee brain. In: Galizia, C.G., Eisenhardt, D., Giurfa, M. (Eds.), Honeybee Neurobiology and Behavior. Springer Verlag, Heidelberg, New York, pp. 409–422.

Mussig, L., et al., 2010. Acute disruption of the NMDA receptor subunit NR1 in the honeybee brain selectively impairs memory formation. Journal of Neuroscience 30, 7817–7825.

Okada, R., et al., 2007. Learning-related plasticity in PE1 and other mushroom body-extrinsic neurons in the honeybee brain. Journal of Neuroscience 27, 11736–11747.

Schultz, W., 2006. Behavioral theories and the neurophysiology of reward. Annual Review of Psychology 57, 87–115.

Stach, S., Benard, J., Giurfa, M., 2004. Local-feature assembling in visual pattern recognition and generalization in honeybees. Nature 429, 758–761.

Stollhoff, N., Menzel, R., Eisenhardt, D., 2005. Spontaneous recovery from extinction depends on the reconsolidation of the acquisition memory in an appetitive learning paradigm in the honeybee (*Apis mellifera*). Journal of Neuroscience 25, 4485–4492.

Szyszka, P., Galkin, A., Menzel, R., 2008. Associative and non-associative plasticity in Kenyon cells of the honeybee mushroom body. Frontiers in Systems Neuroscience 2, 1–10.

von Frisch, K. (1967). Harvard University Press, Cambridge, MA.

Zhang, S., et al., 2005. Visual working memory in decision making by honey bees. Proceedings of the National Academy of Sciences of the United States of America 102, 5250–5255.

Zhang, S.W., et al., 2004. Grouping of visual objects by honeybees. Journal of Experimental Biology 207, 3289–3298.

Memory Models: Quantitative

Bennet Murdock, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

After a brief introduction, a description is given of memory models, the information-processing view, and recent developments. Three specific memory models are discussed in greater detail: MINERVA, REM, and TODAM. The article ends with some applications to data and a look at other areas in the memory field.

Introduction

As has been said before, the study of learning and memory has a long history but a short past (empirically a long history and theoretically a short past). What many people do not realize is that over the last 150 years of research there has been a steady accumulation of data in the area of human learning and memory. Many years ago a physiological psychologist confided to me that our data was better (more reliable and robust) then his. This may have been a slight exaggeration, maybe he was only being polite, or maybe things have changed in the interim. Whatever, we have the right to be proud of our data.

The models or theory are another matter. We are relative novices in this area, and although many models can give accurate fits to selected data sets, they are not always as broad as they could be and the level of mathematical sophistication is not too high. However, this is gradually improving and the nature of the modeling seems to be changing from formal 'hypothetico-deductive' models (Hull, 1943) to computer-simulation models (Simon and Feigenbaum, 1964), and it can be argued that this is the more appropriate method to understand the wide variability that is perhaps the outstanding characteristic of human memory.

There are a number of 'textbooks' of human memory, research monographs that describe in breadth and depth the state of the field at the time they were written. My favorite examples are McGeoch (1932), Osgood (1953), Crowder (1976), Murdock (1962), Neath and Suprenant (2002), and Kahana and Caplan (2002). Introductory psychology textbooks (of which there are many) only begin to scratch the surface, giving you bit of the terminology and a few of the major empirical results, but they are usually quite superficial when it comes to models and theories.

This article is about quantitative memory models, and I can assure the reader that most of the experimental data analyzed in these memory models has stood the test of time. However, the models are more recent and constantly being revised. This should not bother the reader; Newton's model was quite useful until Einstein came along, and sometimes early versions of a model turn out to be just special cases of a more general model.

The memory models to be covered here fall into four categories corresponding to four different types of memory tasks: item recognition, paired associates, serial-order effects, and free recall. To make the models more comprehensible I will briefly explain what is involved in each of these four tasks. This is necessary because often the models are task specific; they only apply (rigorously) to one type of task, though of course they may generalize. In studies of item recognition the experimenter presents a list of items (words, pictures, whatever) to the subject(s) and then tests with old and new examples. Some items in the test list are 'old' (i.e., presented in the list) and some are 'new' (i.e., not presented in the list). With a yes-no procedure, all the subject has to say when a test list containing a mixture or old and new items is presented is 'old' or 'new,' but often a confidence judgment is used where the subject also has to indicate on some scale how confident he or she is. This scale can be one to three, one to five, one to ten, or whatever the experimenter chooses, but the range is seldom more than ten.

In studies of paired associates, lists of pairs (A–B, C–D, E–F, and so on) are shown (visual or auditory presentation) to the subject and then, usually in a scrambled order, each pair is tested in a random order one by one, usually in a forward order (C–?, A–?, F–?) or in some experiments in a backward order (D–?, B–?). There can be a single presentation (i.e., study then test) for one trial or 'short-term' or multiple presentations perhaps for a fixed number of trials or perhaps until some criterion (e.g., 75%; 100%, two consecutive perfect trials) is reached. This is testing recall. The first or second item in each pair is presented to the subject, who must then recall its mate (second or first item, respectively).

Sometimes with this paired-associate method pair recognition is tested. 'Intact' (e.g., C–D) or 'rearranged' (E–B) pairs are used in the test phase, and the subject has to say either 'old' or 'new.' This old–new judgment may be supplemented with confidence judgments. Either way (recall or recognition), there is only a single list. This 'single-list,' procedure is often extended to studies of pairs of related lists. These related lists may test for 'retroactive' or 'proactive' interference (Crowder, 1976). Retroactive interference (RI) can be symbolized by an A–B, A–C arrangement (first member same in the two lists, second member different), while proactive interference (PI) can be symbolized as A–B, C–B, so that the first member is different but the second member is the same. Thirty years of research under the aegis of interference theory (Crowder, 1976) was devoted to this topic (main effects and similarities and differences between RI and PI), and although the data are still as reliable as ever, interference theory has been incorporated into formal memory models and will not be discussed further.

In serial learning, a list of items (common words, nonsense syllables, poetry) is learned trial by trial, either with a study-test

 http://dx.doi.org/10.1016/B978-0-08-097086-8.43065-5

procedure or the method of anticipation. If the former is used, the entire list is presented and then recalled as best is possible, and this is repeated until some criterion is reached. In the method of anticipation, each item is presented and the subject is to anticipate what the next item will be. The next item is shown, and so on, trial after trial, until some learning criterion is reached. The dependent variable (measure of learning) is usually the number of trials to some criterion (all correct, some fraction, or some number of successive correct trials). This is the learning paradigm used by Ebbinghaus in his pioneering study of learning and memory (Ebbinghaus, 1913/1964). Free recall will be presented shortly.

Models

Formal (quantitative) memory models probably started with the Ebbinghaus forgetting curve. He summarized the results of an extensive set of experiments on serial learning, where he was both subject and experimenter, with what we would now call an exponential decay curve showing rapid forgetting at first and progressively slower forgetting as time passed. It was exponential because the rate of forgetting was roughly inversely proportional to the time passed between learning and test. This exponential forgetting curve may well have been the first formal quantitative description of forgetting.

After Ebbinghaus, serial learning was '*de rigueur*' in the early 1900s. It was to human learning and forgetting what maze learning was in rats. The emphasis in both cases was on accurate description of the learning or forgetting curves, and although it can be disparaged as mere 'curve fitting,' at least it was a start. These studies provided the data base for McGeoch's influential book on learning and memory (McGeoch, 1946), and this is probably where a theoretical dispute broke out. There were 'decay theorists' and 'interference theorists,' and McGeoch epitomized the interference point of view by stating 'iron rusts in time but the cause is oxidation.'

American psychologists took note and the 'interference' banner was carried by Benton Underwood at Northwestern University and Leo Postman at UC Berkeley. Their view of memory was the prevailing view until the early 1970s (Postman and Underwood, 1973). However, British psychologists held staunchly to the decay point of view with a rather more creative approach (see Bartlett's *War of the Ghosts*). Neither of these could be considered quantitative models of memory by any stretch of the imagination.

Then a fourth memory/learning paradigm appeared: free recall (Bousfield, 1953). This task was less structured then item recognition, paired associates, or serial-order learning. The task was simplicity itself: the experimenter would present a list of items (usually words) of some list length and at some presentation rate and ask the subject to recall as many items as possible in any order. Single-trial free recall was of course only one trial; free-recall 'learning' involved repeating the study list over again to the subject always in the same order with recall after each presentation, and this could be done for a fixed number of trials or until some criterion (one perfect, two consecutive perfect, etc.) was reached.

This method was popularized by Deese (1959), but the results of a more elaborate study by Murdock (1962) turned out to be quite influential. The results showed a brief (three-item) primacy effect, a more extensive (seven-item or so) recency effect, and a level asymptote connecting the two. The primacy and recency effect were essentially independent of list length (10, 15, or 20 words) and presentation rate ($2\ s^{-1}$, $1\ s^{-1}$, or 1 per 2 s), but the asymptote was approximately proportional to the total presentation time. The results could be characterized by a Gompertz double-exponential function. These results generated considerable interest in memory modelers, so much so that they were included along with the results from a short-term probe test as the two short-term memory data sets that any creditable model should be able to explain (Norman, 1970).

There is not space here to report all the data obtained from these four paradigms, or even a small subset. The interested reader is referred to Kahana and Caplan (2002) for the latest survey, but the reader should be warned that even his data presents only a small subset of all the data that have been reported in the literature in well over 100 years of research.

Information-Processing View

The next major step was probably an information-processing view (although it was not called that at the time), namely, Broadbent's (1958) Y-shaped tube. Broadbent considered the human information-processing system to be preceded by a 'limited-capacity' attentional system, which he characterized as a Y-shaped tube. Consider an experimental study of 'dichotic listening,' meaning that a subject was listening to two simultaneous streams of unrelated words presented over separate headphones to the two ears. That is, one stream was presented to the left ear and the other stream was presented to the right ear. Imagine the information from the two ears feeding into some hypothetical central processor that could map input into output or recall. Accuracy would depend upon how well this was done.

As it turned out, human performance (recall accuracy) was not too good, and if recall was delayed somewhat after the presentation ended, it was even worse. To explain these data, Broadbent expanded this metaphor to postulate that there was 'acid' in the tubes and if the information (conceived as metal balls resting in the tubes) stayed there too long before recall (which emptied the tubes), the acid would result in decay and performance (number of words recalled) would decrease. The Y-shaped tube metaphor was used to explain the fact that if the input streams to the two ears were reported sequentially, the channel reported second was always reported (recalled) less accurately than the channel reported first (Moray, 1959).

There is an important issue here. Long before Broadbent's work, 'decay' had been a popular notion in the study of human memory. Then the American psychologist John McGeoch took issue with the decay notion. The disagreement between McGeoch and Broadbent (interference vs decay) put American and British psychologists in opposing camps, and traces of this controversy still exist today.

Recent Developments

The dispute between interference and decay still goes on today but at a more sophisticated level. The big leap ahead in our

theoretical understanding occurred at Stanford University in the mid- to late 1960s under the aegis of the Institute for Mathematical Studies in the Social Sciences. Under the leadership of Professors William K. Estes and Patrick Suppes, there was a critical mass of very talented researchers and theorists, most notably and in alphabetical order, Richard Atkinson, Bob Bjork, Gordon Bower, Ed Crothers, James Greeno, Chizuko Izawa, Walter Kintsch, Steven Link, Richard Shiffrin, James Townsend, and Joe Young. Much of their early work was on the 'one-element model,' a Markov model for learning and memory. This model postulated all-or-none learning so a particular memory trace was in one of only two states: 'learned,' in which case it could be recalled, or 'unlearned,' in which case it could not be recalled. This one-element model had a short life; experiments by various members of this group (most notably Greeno, 1967) showed that this view of memory was untenable. Actually, there were some Gestalt-oriented psychologists (Rock, 1957) who did not give up easily and presented some experimental evidence supporting an all-or-none view, but it was not accompanied by any formal modeling attempt.

One remnant of this early work not only survived but became probably the single most widely quoted and influential quantitative memory model in the history of our field to date. This was the Atkinson and Shiffrin (1965) 'buffer model.' There is a rumor that one night Atkinson asked his graduate students over for an evening of beer and pretzels, and the agenda was to suggest how human memory might be like a computer. Computers (not laptops but mainframes and Digital Equipment Corporation laboratory computers such as the PDP-7 and the PDP-12) were starting to become popular, and maybe we could not only teach them to learn but also learn from them. The result is reputed to be the buffer model.

The buffer model assumed that (like a PDP-12) there was an accumulator (AC) or very small short-term memory and a 'multiplier/quotient' (MQ) that could use the data entry from the AC as 'grist for the mill' (i.e., number crunching or multiply and divide). The AC in the buffer model was a short-term limited capacity store that could hold a few items. The longer an item resided in this short-term store, the greater the chances that it would be transferred to a hypothetical long-term store. Thus there was a short-term memory (STM) and a long-term memory (LTM). This buffer model was basically a simulation model rather than an analytic model, although some explicit expressions were derived. It was able to explain or fit a large amount of experimental data (including the serial position effect of free recall, described earlier). By then the experimental literature had grown considerably since the time of Ebbinghaus and was highly reliable. A noted psychologist who worked with both physiological and psychological memory and learning data once told me that the human memory data were much more robust and reliable than the physiological data he had to work with. However, that was probably 30 years ago and the situation may have changed since then.

This buffer model was soon followed by a cognitive revolution at Stanford under the aegis of John Anderson, who was a graduate student working with Gordon Bower. An important component of this so-called 'cognitive revolution' was a book by Ulrich Neisser (Cognitive Psychology, 1967) that is still cited today. Important work was done both at Stanford and at Carnegie Mellon with Newall, Simon, and Shaw. After graduation from Stanford, John Anderson joined them and developed the Automatic Control of Thought (ACT) (Anderson et al., 1998). These models have been quite influential not only in psychology but also in education and the school system. They are a little off the beaten path that I am trying to follow and so they will not be discussed further here, but they should not be overlooked in any account of the development of our field.

The cognitive science approach (not ACT or ACT-R) received a body blow over the issue of 'catastrophic interference' in studies of retroactive inhibition (RI). As has been mentioned, there are two lists of paired associates to be learned. After the first A–B list has been learned then the second or A–D list has to be learned. Many experimental studies have shown that the second-list learning is gradual; slowly but surely the D responses of the second list replace the B responses of the first list. What happens in a connectionist model? As shown by Ratcliff (1990), the 'relearning' (A–D) list only takes a single trial or two. Thus connectionist models show 'catastrophic' (one- or two-trial) unlearning, which is quite different from what happens in experimental studies. Although there have been a few theoretically inspired attempts to provide a connectionist solution to this situation, the applecart had been overturned and connectionist theorists turned to other problems.

However, this does not mean that all interest in simulation models of memory was lost. Roughly at the same time, distributed (as opposed to localized) memory models appeared on the scene. The most widely tested were the Theory of Distributed Associative Memory (TODAM) model of Murdock (1983), after the Content Addressable Distributed Associative Memory (CADAM) model of Liepa (1977); MINERVA (Hintzman, 1984), named after the Greek goddess of poetry, medicine, wisdom, commerce, weaving, crafts, and magic; and the Search of Associative Memory (SAM) model (Gillund and Shiffrin, 1984). CADAM was really three models in one (for recognition, paired associates, and serial-order effects) whereas TODAM was a unified model that applied to all three.

CADAM/TODAM are simulation models while MINERVA/SAM are computational (analytic) models. CADAM/TODAM assume the underlying encoding/storage/retrieval processes, run the simulations, and calculate the predicted results from analyses of the simulated data. MINERVA/SAM make hypotheses about the underlying processes and mechanisms, derive analytic expressions from these assumptions, and then use these expressions to make predictions of the data. Rather than describe each model in detail, I will simply discuss their storage assumptions, retrieval assumptions, and response rules and try to give some indication of how they work. The predictions can be obtained mathematically from derivations or from computer simulation.

Overview

These memory models make different assumptions about encoding, storage, and retrieval, the three basic processes of short-term memory. CADAM/TODAM assume that features are represented by random samples from a normal distribution with mean zero and some variance. This assumption was used in the CADAM model by Liepa, who chose this assumption

because he told me that in quantum mechanics this assumption is standard, as it obviates the need for normalization later on. MINERVA assumes a ternary representation for the encoding of features, 1, 0, or −1, while SAM assumes a geometric representation. That is, each feature value is a random sample from a geometric distribution with some probability (p value). Features are encoded with probability p or not encoded with probability $(1 - p)$.

How are associations represented? In TODAM, an association is represented by the convolution of two A and B item vectors. Convolution ('Faltung' in German) is a particular way of combining two distributions; for details, see (Bracewell, 1965: Chapter 3). This convolution maintains the vector form so items and associations can be stored in a common memory vector. If circular convolution is used (Plate, 1995), items and associations are the same dimensionality. One can form a linear convolution by forming the outer product matrix and then summing the diagonals. A circular convolution can be formed by folding a linear convolution or using the Fourier transform directly (Bracewell, 1965).

In MINERVA, the A and B members of an A–B pair are simply concatenated and stacked end-to-end and zero denotes a missing or unknown feature. What about retrieval? In MINERVA, applying the probe (A item of an A–B pair or B member for backward associations) will retrieve the corresponding A member. Associative symmetry is guaranteed from the convolution-correlation identity (Murdock, 1983). This is important because there is strong evidence for associative symmetry; backward associations are as good as forward associations (Kahana and Caplan, 2002). In TODAM, associative symmetry is guaranteed because it follows from the convolution/correlation identity. So one way or another all three models can predict associative symmetry, and this seems to hold up in experimental situations.

In an experiment, subjects can recognize items that were or were not in the list of pairs (targets and lures, respectively). I do not know of any rigorous proofs of this, but suspect that even if this prediction does not flow naturally (as it does in TODAM), it would not be difficult for any of these models to predict associative symmetry. However, in TODAM all information is stored in a common memory store, so given the retrieval operation (correlation), the retrieved information is at best a noisy approximation to the target item. This approximation will vary with a number of factors (number of pairs in the list, number of features in each item vector, A and B item similarity both within and between). There are deblurring algorithms, that is, computational procedures that will find the best target (A or B item) for any A′ or B′ approximation (e.g., Hopfield, 1982). However, to my knowledge these have not been implemented in any of the models. Instead it is simply assumed that the best match will be the item retrieved, but how one goes from an approximation to the best match has not been specified.

MINERVA

Here is an account of how MINERVA (Hintzman, 1984) works in more detail. With stimulus presentation, the features of an item are from a short-term or 'primary' memory (PM) and copied to a long-term or 'secondary' memory (SM), and SM 'echoes' or returns a signal to PM; this is called 'echo intensity.' SM is a large collection of 'experiential records,' records of past experience. There are multiple copies in SM rather than 'stronger' copies of each individual experience.

When a probe goes to SM it is matched by all the SM traces and you get a 'summed similarity.' The resulting memory depends on summed similarity (S). Let $P(j)$ represent feature j of the probe and $T(i, j)$ be the corresponding features of memory trace i. Then the similarity of trace i to the probe j is computed as:

$$S(i) = \sum_{i=1}^{N} P(j)T^{(i,j)}/N$$

where N is the total number of features that are nonzero in either the probe or the trace. Activation of a trace or A is the cube of the similarity, so:

$$A(i) = S(i)^3$$

Raising the similarity to the third power increases the signal-to-noise ratio as it increases the number of poorly matching traces required to overshadow a trace that closely matches the probe. When a probe activates a trace in SM, information is returned in the echo. The echo is assumed to have two properties, intensity and content. If M is the total number of traces in memory, the intensity I is given by:

$$I = \sum_{i=1}^{M} A(i)$$

If there are no matching traces, than the expected value of I or $E(I) = 0$. Otherwise $E(I) \; ! = 0$. The variance of I depends on the number of matching traces. Frequency judgments and recognition judgments are assumed to be based on echo intensity. For the j-th constituent features the content of the echo $C(j)$ is:

$$C(j) = \sum_{i=1}^{M} A(i)T(i,j)$$

These C values can range from negative to positive and are assumed to be immediately available in PM. They form the basis for frequency judgments and recognition. A one-element version of MINERVA has also been able to give a good fit to some paired-associate learning data.

REM

REM (Retrieving Effectively from Memory) is an update of the SAM (Search of Associative Memory) model, which in turn was an update of the original Atkinson and Shiffrin buffer model. REM is designed primarily to deal with explicit episodic recognition memory. The model (or at least one version of it) is given in ten sets of linear equations specifying the likelihood ratios and their distributions. There is no room for a detailed discussion here; the interested reader is referred to the appendix of the original article (Shiffrin and Styvers, 1997). All we can do here is present the basic outline.

Like MINERVA, memory is assumed to consist of separate images, which are represented by a vector V of geometrically distributed features such that:

$$P[V = j] = (1 - g)^{j-1}g, j = 1, \ldots, \infty$$

For recognition, an error-prone copy of the item is stored, and for retrieval this probe is matched in parallel with all of the n words on the list. A Bayesian decision rule is used, and the odds are:

$$\Phi = \frac{1}{n}\lambda_j$$

The parameter lambda (λ) can be written in terms of the probabilities in three different ways (eqn [3], [4A], or [4B]), and these are used to generate the predictions of the model.

TODAM

TODAM is more a computer simulation model than a formal mathematical model. Unlike MINERVA and REM, TODAM is a distributed memory model where all information (item information and associative information) is stored in a common memory vector. The big difference between TODAM on the one hand and MINERVA and REM on the other hand is that the latter two are 'localist' models where information (items or associations) is stored in separate 'bins' (locations) as in a computer memory, while in TODAM everything is assumed to be stored in a common memory system and no search is required. As has been mentioned, it assumes direct access but, like MINERVA and REM, the retrieved information is 'noisy.' All three of these models focus on the same data base and each model can do a creditable job explaining some of the vast data in the literature but not all of it. Eventually the dust may settle (and I use the metaphor advisedly) and we may be able to say with some assurance that one model is better than the other two, or some other model is better than all of them, but that is a job for the future.

Applications to Data

There are many reliable empirical phenomena from studies of short-term memory and each of these three model has demonstrated the capacity to explain (a nonoverlapping) subset of them. One of the most important empirical effects is similarity. Similarity effects in human short-term memory have long been known. In a classic early study, Conrad (1959) found that acoustic similarity had a major effect on response accuracy in a simple memory-span task. It was a very elegant experiment; using a distinctive features analysis of his stimulus letters, Conrad showed a monotonic decline in the recall accuracy for single substitution errors (four letters out of five in a memory span task were correctly recalled but the first was not), and the greater the overlap in distinctive features the worse the performance. Neither TODAM nor SAM should have any trouble dealing with these similarity effects, but MINERVA, with ternary vectors, might have more trouble. However, if it were assumed that similarity was measured by the dot product or cosine, then MINERVA could probably deal with Conrad's similarity effects as well.

There are very many more reliable empirical effects in the study of short-term memory, but it is not possible to deal with any more of them here. A recent review of the field (Kahana and Caplan, 2002) presents a very comprehensive review and evaluation of these data and it will be interesting to see how these various models (and others yet to appear) can deal with them.

Other Areas

I have focused on these three particular models of human memory (TODAM, MINERVA, and REM) because they are relatively well developed. They permit some experimental resolution because they all can deal with the same broad range of data. There is yet a fourth area of short-term memory that should be mentioned if only briefly, and that is free recall. The free recall paradigm is very simple: present a short list of items (often common words) and ask the subjects to recall them in any order they wish. As has been noted, the classic result is a relatively extensive (seven or eight item) recency effect, a short (three or four item) primacy effect, and a middle asymptote (level part) that depends on the presentation rate and list length. The two major models here are the Temporal Context Model (TCM) (Howard and Kahana, 2002) and Donald Laming's work (Laming, 2010).

A different approach to associative learning has been provided in a series of studies and models in the Broadbent tradition by Burgess and Hitch (1999); these models assume item-to-position associations rather than item-to-item associations. There is the Dennis and Humphreys model (Dennis and Humphreys, 2001) and the Brain-State-in-a-BOX or SOB model in Australia by Farrell and Lewandowsky (2002).

Most prominent in the area of recognition memory is the random walk (diffusion model) of Ratcliff (1978). The Ratcliff model followed closely after Link's random model now incorporated in Link's WaveTheory model (Link, xxxx) with ideas that go back as far as Einstein. These are very different models of memory using a random-walk algorithm for the accumulation of information over time and have given very detailed (and accurate) predictions of accuracy and latency as a function of drift rate. Unfortunately there is simply no space here for an adequate coverage of any of these models, but the interested reader can track them down and see how they fit into the growing list of models of short-term memory and learning.

Finally, some time ago a very elegant (and widely ignored) model was suggested by Kanerva (1988), in which "many memory storage registers participate in a single write or read operation. Storage locations and their addresses are given from the start and only the contents are modifiable" (p. 61). Retrieval works on a best match principle. As the author says, "The mathematical feasibility of distributed storage is the main result of this study."

Conclusion

As has been noted, there are a host of reliable short-term memory effects and a fair number of quantitative models that have their various successes and failures. I think it is fair to say that at this point this is no convincing evidence that any one model is clearly superior to the others. However, as time passes, the evidence will come in and some resolution should result. While one may deplore the lack of resolution of these issues, perhaps we should be pleased by the developments that have occurred. Our models and theoretical concepts are still relatively primitive compared to those of the 'hard' sciences, but still we have come a long way since Ebbinghaus. Our learning and memory models are more comprehensive and sophisticated than ever before. Our goal should be to develop

mathematical and simulation models of human memory and learning that can do justice to the vast accumulation of experimental data that has been collected over the last 100 or more years of experimentation since Ebbinghaus.

See also: Diffusion and Random Walk Processes; Knowledge Representation; Reconstructive Memory, Psychology of; Signal Detection Theory, History of; Signal Detection Theory.

Bibliography

Anderson, J.R., Bothel, D., Lebierre, C., Mantessa, M., 1998. An integrated theory of list memory. Journal of Memory and Language 38, 341–380.

Atkinson, R.C., Shiffrin, R.M., 1965. Mathematical Models for Memory and Learning. Technical Report Number 79. Institute for Mathematical Studies in the Social Sciences, Stanford University.

Bousfield, W.A., 1953. The occurrence of clustering in the recall of randomly arranged associates. Journal of General Psychology 49, 229–240.

Bracewell, R., 1965. The Fourier Transform. McGrawHill, New York.

Broadbent, D.E., 1958. Perception and Communication. Pergamon Press, New York.

Burgess, N., Hitch, G.J., 1999. Memory for serial order: a network model of the phonological loop and its timing. Psychological Review 106, 551–581.

Conrad, R., 1959. Errors of immediate memory. British Journal of Psychology 50, 349–359.

Crowder, R.G., 1976. Principles of Learning and Memory. Lawrence Erlbaum Associates, Hillsdale, NJ.

Deese, J., 1959. On the prediction of occurrence of particular verbal intrusions in immediate recall. Journal of Experimental Psychology 58, 17–22.

Dennis, S., Humphreys, M.S., 2001. A context noise model of episodic word recognition. Psychological Review 108, 452–478.

Ebbinghaus, H., 1913/1964. Memory: A Contribution to Experimental Psychology. Teachers College, Columbia University (Reprinted by Dover), New York.

Farrell, S., Lewandowsky, S., 2002. An endogenous distributed model of ordering in serial recall. Psychonomic Bulletin & Review 9, 59–79.

Gillund, G., Shiffrin, R.M., 1984. A retrieval model for both recognition and recall. Psychological Review 91, 1–67.

Greeno, J.G., 1967. Paired-associate learning with short term retention: mathematical analysis and data regarding identification of parameters. Journal of Mathematical Psychology 4, 430–472.

Hintzman, D.L., 1984. MINERVA 2: a simulation model of human memory. Behavior Research Methods & Instrumentation 16, 96–101.

Hopfield, J.J., 1982. Neural networks and physical systems with emergent collective computation abilities. Proceedings of the National Academy of Science 79, 2554–2558.

Howard, M.W., Kahana, M.J., 2002. A distributed representation of temporal context. Journal of Mathematical Psychology 46, 269–299.

Hull, C.L., 1943. Principles of Behavior. Appleton-Century-Crofts, New York.

Kahana, M.J., Caplan, J.B., 2002. Associative symmetry in probed recall of serial lists. Memory & Cognition 30, 841–849.

Kanerva, P., 1988. Sparse Distributed Memory. MIT Press. pp. 1–145.

Laming, D.J., 2010. Serial position curves in free recall. Psychological Review, 93–133.

Liepa, P., 1977. Models of Content Addressable Distributed Associative Memory (Unpublished manuscript).

McGeoch, J.A., 1932. Forgetting and the law of disuse. Psychological Review 39, 352–370.

McGeoch, J.A., 1946. The Psychology of Human Learning. Longmans, Green, New York.

Moray, N., 1959. Attention in dichotic listening: affective cues and the influence of instructions. Quarterly Journal of Experimental Psychology 11, 56–60.

Murdock, B.B., 1962. The serial position effect of free recall. Journal of Experimental Psychology 64, 482–488.

Murdock, B.B., 1983. A distributed memory model for serial-order information. Psychological Review 90, 316–338.

Neath, I., Suprenant, A., 2002. Human Memory, second ed. Wadsworth.

Neisser, U., 1967. Cognitive Psychology. Appleton-Century-Crofts, New York.

Norman, D.A. (Ed.), 1970. Models of Human Memory. Academic Press, New York.

Osgood, C.E., 1953. Method and Theory in Experimental Psychology. Oxford University, New York.

Plate, T.A., 1995. Holographic reduced representations. IEEE Transactions on Neural Networks 6 (3), 623–641.

Postman, L., Underwood, B.J., 1973. Critical issues in interference theory. Memory & Cognition 1, 19–40.

Ratcliff, R., 1978. A theory of memory retrieval. Psychological Review 85, 59–108.

Ratcliff, R., 1990. Connectionist models of recognition memory. Psychological Review 97, 295–308.

Rock, I., 1957. The role of repetition in associative learning. American Journal of Psychology 70, 186–193.

Shiffrin, R.M., Styvers, M., 1997. A model for recognition memory: REM – retrieving effectively from memory. Psychonomic Bulletin & Review 4 (2), 145–166.

Simon, H.E., Feigenbaum, E.A., 1964. An information processing theory of similarity, familiarization, and meaningfulness in verbal learning. Journal of Verbal Learning and Verbal Behavior 3, 385–396.

Men's Health

Richard G Owens, School of Psychology, University of Auckland, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

It is increasingly being recognized that just as the demands of female gender-role expectations impact negatively on the health of women, so too do the demands of male gender-role expectations on men. It is widely recognized that life expectancy for men is, in most developed countries, lower than that of women, and such discrepancies can be linked, at least in part, to behaviors associated with 'being a man.' Morbidity, too, is affected by conformity to the expectations of hegemonic masculinity, with men differing markedly from women across a range of dimensions including emotional expression, body-image expectations, sexual behavior, and help-seeking. Ironically, stepping outside traditional male norms may also have adverse effects on health as such men become to a greater or lesser extent excluded from mainstream society. The suggestion, therefore, is that rather than seeking solutions at an individual level, there is a need for a wider societal revision of the specific expectations place upon both men and women.

Introduction

A number of factors have, over recent decades, led to increasing recognition of the extent to which socially determined gender roles impact on health. Much of the early work in this field, developed within a context of hegemonic masculinity, was concerned with ways in which social constructions of gender affected adversely the health of women (see e.g., Lee, 1998). It is perhaps natural that when attention is paid to issues such as gender, the main focus should be on those groups most obviously disadvantaged by these expectations, but recent years have seen increasing recognition not only of the ways in which men, as well as women, are disadvantaged by gender roles, but also of the fact that the interplay of male and female gendered expectations does not result in a zero-sum game. That is to say, it is increasingly apparent that deconstructing and challenging the expectations on one sex does not necessarily imply a disadvantageous outcome for the other. Indeed, close examination of the processes involved leads to the conclusion that the challenging of conventional expectations may prove beneficial to both men and women. This applies not only in terms of such dimensions as economics, but also, significantly, to the health of the individuals concerned.

From early and often piecemeal beginnings (e.g., Helgeson and Lepore, 1997), a picture has begun to emerge, which demonstrates the wide-ranging impact on men's health of their conformity (or otherwise) to perceived gender-role expectations. Such effects can be seen across a wide range of settings, including increased exposure to and risk of threats to health, decreased access to or take-up of protective and other health measures and increased mortality. Some of these have previously been described in various contexts including a special issue of the *Journal of Health Psychology* (May 2002) on men's health, and the publication of *The Psychology of Men's Health* (Lee and Owens, 2002). The present article thus has two main aims: (1) to present in concise form some of the arguments presented in earlier work and (2) to update and develop some of the material previously available.

Men and Mortality

In a world in which to be a male is often seen to be placed in a privileged position, it perhaps appears anomalous to note that in most developed countries, the life expectancy of men is substantially lower than that of women. By exactly how much women can expect to outlive men varies. In the United States, for example, men can expect to die around 6 years younger than women, in France nearly 8 years younger, and in the UK almost 5 years younger. The reasons underlying this are many and varied, and at various times, a range of mechanisms have been put forward. Not surprisingly the discrepancy has traditionally (see e.g., Madigan, 1957) been linked to biological or physiological factors. For example, Waldron (1995) attributes the bulk of the effect to factors such as the possible protective effect of estrogen on coronary heart disease (implying that women's risk of heart disease will, at least premenopausally, be lower than that of men). Certainly, there is evidence, which is consistent with such a suggestion; for example, in the United States 75% of those dying from heart disease below age of 65 years are men, and in the Framingham Study (Lerner and Kannel, 1986) women showed only around half the cardiac mortality of men. Clearly, however, this can be only part of the story. First, it should be noted that the male–female differences in biology have been around as long as the human race, but significant sex differences in mortality only started to appear toward the beginning of the twentieth century (Gonça et al., 1999). These biological differences of course are the same in all countries, yet as we have seen above there is considerable international variation in the size of the discrepancy, e.g., between France and the UK.

Second, close examination of the pattern of mortality differences reveals that for many developed countries, the number of surviving women exceeds that of surviving men from about age of 25 years, well before there is any significant impact of coronary heart disease on mortality. Indeed, plotting of the differences in mortality against age reveals that the discrepancy tends to peak at two points. One late in life, as may be expected under an assumption of the effects of

coronary heart disease, but another substantial (and, when corrected for raw mortality rates, greater) discrepancy between ages 20–35 years. It appears that in addition to any physiological process affecting such problems as heart disease, many more men than women are dying 'prematurely.' The discrepancy at these younger ages is more commonly ascribed to factors such as men being typically being involved in more dangerous occupations and activities, being greater risk takers, and so forth. It is a small step from such an observation to the acknowledgment that much of this can be linked to these activities being seen as, predominantly, 'typically male.' Thus in Australia, the three-quarters of homicides who are 'nonintimate' (i.e., do not involve the death of someone close such as a current or ex-partner) men are about three times as likely to be the victims as women (Pratt and Deosaransingh, 1997). Other causes of death show similar patterns, for example, men are three times as likely as women to be killed in motor-vehicle accidents (Li et al., 1998), largely reflecting men's greater exposure to risk (e.g., time spent traveling). Similarly, sporting injuries and fatal accidents are more common among men than women (see e.g., Williams et al., 1998). Notably, however, the idea that men are greater risk takers turns out to be somewhat more complex, and only true to the extent that men are more likely than women to be involved in higher risk activities. If one corrects for the greater extent of men's involvement than women's, then these discrepancies disappear or even, in some cases, reverse. Thus, it is not simply an issue that can be dismissed in terms of individual analyses and 'men doing stupid things,' but rather requires a broader societal analysis, which highlights the gender expectations, which socialize men toward these activities while simultaneously directing women away. Observations such as these have led some workers to investigate the role of factors such as patriarchal attitudes on mortality differences. In one such study (Stanistreet et al., 2005), 49% of the international variance in male–female mortality differences could be predicted from a country's rate of female homicides, which the authors used as a proxy measure for the degree of patriarchy. Thus, it was argued, the more patriarchal a society, the earlier men could expect to die relative to women.

Once the potential effect of gender-role expectations is recognized, of course, it also becomes apparent that there is likely to be an interaction between these and the biological factors to which the discrepancies have traditionally been attributed. In a society, for example, which regards smoking as an acceptable behavior for men but not for women we may expect decades later to find that deaths from heart disease and lung cancer, even with all other things being equal, will be higher among men. In passing, it is worth remembering that a common concern for those working in the population health field, of course, is the convergence of smoking rates for men and women; Harris (1983) notes that in the United States by 1920 over half the male population smoked, whereas it was not until 1950 that a third of the female population did so. Almost like an experimental multiple-baseline design (Sidman, 1960), the data show that changes in important outcome variables – in this case mortality differences – can be linked to earlier changes in societal factors – e.g., the acceptability of smoking across the sexes.

How Conforming to Gender Expectations May Adversely Affect Men's Health

Although it has been said that mortality is hard to beat when looking for a dependent variable, the reality is that there is more to health than simply being alive. We have seen already how mortality can be related to differences in societal expectations for men and women. It would be odd, if these effects on mortality were not also reflected in morbidity, and indeed much of the evidence supports the supposition that 'being a man' may be, in many respects, bad for one's health. Around the world, implicit or explicit exhortations imply that to be 'a man' is to be more than just an adult male. A Japanese story tells of a Samurai saying to his son "I would not have you a priest, of whom there are many, but a man, of whom there are few." In the UK, Rudyard Kipling's poem 'If,' voted as the nation's favorite, calls upon the man to "force your heart and nerve and sinew …" and to "… fill the unforgiving minute, with sixty seconds' worth of distance run …". A Chinese idiom "不到长城非好汉" roughly translates as "if you haven't climbed the Great Wall, you're not a real man." The 1980s saw the whole edifice of expectations of masculinity satirized in Bruce Feirstein's *Real Men Don't Eat Quiche*, a book which would have undoubtedly had very little success had the expectations parodied therein not been familiar to the reader.

Unfortunately, the kinds of characteristics that distinguish the 'real man' from his failing counterpart are often ones, which have a direct or indirect bearing on his health. Thus men are expected to display emotions differently to women; for men, external positive emotions such as love are less likely to be expressed, as are internal negative ones such as pain or symptoms of illness. By contrast, men are actively encouraged to express external negative emotions such as hostility. Numerous writers (e.g., Courtenay, 2003) have commented on how such restrictions on emotional expression might be expected to impact on health. This may occur directly – for example, two men, each expressing hostility may be a recipe for physical violence and subsequent injury. There may also be indirect effects; Pennebaker (1997) and his colleagues, for example, have demonstrated across a wide range of studies that the experience of disclosing negative experiences can have positive effects on elements of immune functioning, with the clear implication in the present context that men who are reluctant to do so will fare worse than those who do.

Other expectations placed upon the 'real man' can also be seen to impact on health. Research in women's health has highlighted the negative effects of the 'keep young and beautiful' expectation in terms of problems such as anorexia nervosa, but men too have expectations regarding appearance; in particular men, are expected to be physically strong and fit. While, this might at first glance be expected to be, from the point of health, a positive element in traditional male expectations, the same could be said of the requirement that women avoid carrying excess body fat. In both cases, unfortunately problems arise when the expectation is taken to extremes, or when extreme measures are taken to meet the expectation.

As far as men's physical strength and fitness is concerned, the most obvious example of taking the expectation to extremes is in the world of competitive bodybuilding (note that competitiveness itself is seen as a traditionally male

characteristic). One of the most notable pioneers of bodybuilding, of course, Charles Atlas, famously based his whole advertising campaign on the notion that it was bad to be a 'seven-stone weakling' who has 'sand kicked in his face,' and the crude strip cartoons that accompanied his campaign made it clear that the story of the weakling who became strong and physically overwhelmed the 'beach bully' was that of the quintessential hero.

Unfortunately, while the 'dynamic tension' taught by Charles Atlas was perfectly capable of producing a strong healthy body, subsequent years have taken the objective of bodybuilding so far beyond the Atlas physique that achieving a body which would win today's competitions requires distinctly unhealthy strategies. Prominent among these is the use of substances such as anabolic steroids, a risky behavior, which is by no means restricted to bodybuilding (note the disqualification of the 1988 Olympic 100m champion, Canadian Ben Johnson when it was found that he had taken banned steroids).

As with problems relating to emotional expression, the health risks of illicit steroid use may be both direct and indirect. Even under ideal conditions, extensive use of steroid medications may give rise to concern about health problems such as disturbed lipid metabolism, prostate enlargement and liver damage. In practice, given that the steroids are obtained and administered under conditions of secrecy, the direct effects are likely to be exacerbated by indirect health consequences including added risks from taking drugs whose exact provenance may be unknown (and therefore, may have problems such as being outdated, or being veterinary rather than human preparations) and associated behaviors such as needle sharing.

Should one be tempted to dismiss such problems as only applying to a tiny minority (only a percentage of bodybuilders who are themselves only a small percentage of the male population) it should be noted that the figures are nevertheless significant. Wroblewska (1997) in a review of published surveys found rates of anabolic steroid use ranging from around three-quarters of competitive bodybuilders to 4% of males in general. A Swedish study of 16–19 year students found that while only 0.4% of females had used steroids the figure rose to 2.7% for males. Clearly, the use of steroids is not only restricted to those in a few sports with low participation rates.

Concern with appearance, of course, is intimately linked to perceptions of one's own attractiveness, particularly one's attractiveness to potential sexual partners. Once again, however, we find that the social expectations on men are notably different to those on women. A common classroom exercise illustrates this well; students are asked to write down as many names they can think of to describe a woman who has casual sex with anyone to whom she's attracted, even if she hardly knows them. It takes only a short time to come up with a list of derogatory expressions such as 'slut,' 'slag,' 'tart,' and so on. On trying to do the same for a man who exhibits the same behavior, the most obvious name is 'James Bond.'

Once again, it might appear at first glance that such a double standard will be unequivocally advantageous for men, yet a closer examination reveals that it is far from unproblematic. The expectation that men will engage in indiscriminate sexual behavior, may bring a number of health concerns, most obviously by increasing the risk of exposure to sexually transmitted infections. While one might argue that in today's sexually aware modern world, the risk of such infections can be minimized or eliminated by practices such as condom use, the reality is that males are more reluctant than females to suggest condom use (Troth and Peterson, 2000) and more likely to meet criteria for high-risk activity (e.g., multiple partners, inconsistent use of condoms) than women (Newman and Zimmerman, 2000).

It is not only in terms of exposure to risk, however, that traditional male roles have a potentially negative effect on health. With the most gender-neutral society imaginable both males and females would still become sick. Unfortunately, the current gender-role expectations associated with being male not only conspire to heighten men's risk of illness, but also affect their involvement with the systems designed to deal with such illness.

A stereotype that demands men should be 'naturally' strong, resistant to disease, able to cope with pain, and unconcerned with minor problems suggests that they may benefit less from structures in society, which are designed to promote health and minimize the seriousness of disease. As expected the data indeed show, across a range of populations and ages, that men engage in more activities that are harmful to health (e.g., smoking, pursuing unhealthy diets, heavy drinking) than women (e.g., Uitenbroek et al., 1996; Holtzman et al., 2000). A considerable amount of evidence suggests that men use health care and screening services less than women (Schappert, 1999) even when the potentially biasing effects of gynecological services are taken into account (Stoverinck et al., 1996), although the evidence is by no means unequivocal, with a Scottish survey (Macintyre et al., 1999) finding no evidence for male–female differences in symptom reporting or health care. The picture appears complex, with the same research group finding that men were less likely to report a range of symptoms than women, but that when they did so they were just as likely to have consulted a doctor (Wyke et al., 1998). One area, however, where men's behavior appears to be consistently different to women's is the use of psychological services, with men much less likely than women to seek out psychological services for themselves (see e.g., Jorm, 1994). Such reluctance to make use of psychological services is, of course, consistent with a prescriptive male stereotype, which expects men to be strong, self-reliant, and emotionally restrained.

In terms of self-reliance, it is interesting to note that men and women may differ in terms of the health-related actions they take or the areas they neglect. Studies have reported that men are less likely than women to be interested in making healthy lifestyle changes (Gabhainn et al., 1999), and that when they do make health-related lifestyle changes these differ from those typically made by women. Thus, Welsh et al. (1998) in studying health-related attitudes of male and female students found that the former tended to focus on physical activity while the latter focused on diet. Similarly, men appear to know less about health and healthy lifestyles than women, with studies showing that popular media tend to direct such information toward wives and mothers (Lyons and Willott, 1999). While this may be seen as pragmatic, the result may be that the messages are less effective; when women attempt to control the health-related aspects of husbands' behavior this is likely to have less effect than self-directed change, and of course

such campaigns may have little impact on those men who are not in a close relationship with a woman. Indeed, the evidence over many years has consistently indicated that the health of men who are unmarried is worse than that of those who are. Similarly the loss of a partner, whether it be through divorce or bereavement, is associated with more negative changes in health behaviors for men than for women (Umberson, 1992; Clayton, 1990; Byrne et al., 1999). One implication of such findings is that if health campaigns and similar interventions are to succeed, they may need to have a degree of gender-specificity if they are to be maximally effective.

A picture thus emerges whereby aspects of traditional expectations within masculinity may be seen as having potentially serious negative effects on the health of men who subscribe to them, and moreover that the greater the extent to which men attempt to conform to the 'strong, unemotional, self-reliant' image, the greater the negative impact. Of course, few men subscribe unquestioningly to all aspects of this image, and some men will fail to meet, or will positively reject the expectations placed upon them. Ideally, one might hope that in doing so, the negative effects of the traditional male role would be attenuated or eliminated. In practice, however, as with men who subscribe to traditional expectations, the results of rejecting such a role may similarly be complex and problematic.

How Rejecting or Failing to Meet Gender Expectations May Adversely Affect Men's Health

In many societies, it remains the case that the overwhelming majority of men conform to some degree, at least, to traditional male expectations. Thus, in Australia around 90% of married men with children are in the paid workforce, while only 40% of women with children under 15 do so, supporting a traditional notion that men's primary role is to be a breadwinner, providing for the family, while women's is to provide nurturing and care for children. One outcome of this is a potential conflict between social expectation and the reality of family life. Thus, in countries where there is a contraction in the job market with consequent high unemployment, there is evidence that this impacts particularly on men. For example, Pritchard (1992) in a study of suicide rates in Europe during the 1970s and 1980s found that these were significantly correlated with the associated rates of unemployment.

For the men who do have jobs there may, particularly for fathers, be a perceived role conflict if they also take on domestic responsibilities – to a higher degree than do women in similar situations (Greenglass et al., 1988; Duxbury and Higgins, 1994). Such role conflict may not simply result from their own internalizations of traditional expectations, but also from others. Thus, an Australian study (Lobo and Watkins, 1995) found unemployed men who attempted to take on traditional 'women's work' within the house being met with resentment from wives who perceived their husbands as failing to keep 'their side of the bargain,' with associated impact on the quality of the marital relationship and relationships between fathers and children.

For some men, of course, their stepping outside a traditional role expectancy may be seen as global and irretrievable, particularly, in the case of men whose sexual preferences do not conform to the perceived norm of exclusive heterosexuality. Homosexual, bisexual, and transgender males are routinely perceived as deviant and threatening, to be excluded from 'normal' society. Such drastic exclusion continues despite decades of research challenging the essentialist nature of sexual preferences, the evidence that these may change over time, and the recognition even from the middle of the last century that a range of positions may exist between the stereotypes of 'exclusively homosexual' and 'exclusively heterosexual.' Ironically, the man who is attracted to other men may end up with a strong commitment to other aspects of traditional masculinity, as evidenced by such activities as spending hours building a hypermasculine body (possibly with the help of steroids, as mentioned earlier) and becoming involved in extensive sexual promiscuity (with the associated increased risk of contracting sexually transmitted infections). Despite subscribing heavily to such other masculine norms, however, the centrality of sexuality in Western society means that he remains an outcast, unacceptable in 'normal' society, with the associated stresses of either coping with rejection or maintaining secrecy about his sexual identity.

The picture that begins to emerge, therefore, is that, however, a man chooses to play the hand he is dealt, the expectations of traditional masculinity, whether he subscribes to them or not, are likely to impact negatively on his health in a range of ways. It is important, therefore, to recognize that the way forward involves not simply individual men choosing whether or not to subscribe to societal expectations, but a change in those expectations themselves. Such change may be slow, and will almost certainly need to arise from changes in a variety of areas – again, operating not simply at the individual but also at the broader societal level.

Societal Messages around Gender-Role Expectations

As things stand, the socialization of men and women into specific gender roles operates through a variety of routes. It has long been recognized that even before starting school, most children have little difficulty differentiating mothers' and fathers' household and occupational tasks (Hartley, 1961), with messages around the differing expectations apparent even in children's nursery rhymes. One such, for example, "Sugar and spice, and all things nice, that's what little girls are made of; slugs and snails and puppy dog's tails, that's what little boys are made of" illustrates clearly that girls are expected to be pleasant, clean and 'nice,' while for boys it is acceptable to be grubby and unpleasant – messages which soon translate into differences between what is acceptable play for boys and girls, and the extent to which boys and girls are differently encourages or discouraged from different forms of play, physical punishment, and so on.

Throughout life the receipt of such messages continues; in popular media such as television, it has been reported that male characters are considerably more likely to be obese, to smoke, to drink, and to participate in high-risk behaviors. Conversely, females are portrayed as having the highest health risks and being more likely to die, while men and boys who are risk-takers or victims of violence are portrayed as either

invulnerable to harm or likely to recover from such harm with marked rapidity (Courtenay, 2003). Magazines and television programs aimed at men are also, in the United States, more likely to advertise substances such as alcohol and tobacco (Klein et al., 1993). Observations such as these indicate the need for a substantial change in the societal messages and portrayals of male-appropriate and female-appropriate behavior, if the negative health impacts of gender roles are to be addressed.

Beyond Gender

Having said all of, which it is, of course, important to remember that while men's health is significantly affected by gender-role expectations, this is not the whole story, and the role of biology and anatomy cannot be neglected. Thus although the suggestion, noted earlier, that sex hormone levels might account for the bulk of male–female differences in life expectancy is probably an overstatement, there is no doubt that such factors will have an effect. Anatomy, similarly, will lead to differences in the health problems of men and women, irrespective of whether or not the two sexes have differing societal expectations placed upon them. Cancers such as testicular cancer, primarily affecting men while still young, and prostate cancer, more associated with advancing years, will continue to take their toll in the male population. Even so, however, it is perhaps also useful to note that gender roles and biology are intertwined. In many countries, for example, it is routine for women who attend for medical consultations to be offered, routinely, health checks such as cervical smears; it is much less common for men to be offered such things as Prostate-specific antigen checks to identify early signs of prostate cancer.

It is also important to note that the very biological processes involved in many diseases and health problems may themselves interact with the social expectations. Remaining with prostate cancer for the moment, there is some suggestion that the more cigarettes a man smokes, the higher his risk of prostate cancer; the heaviest of smokers have a 24–30% higher risk of death from prostate cancer than nonsmokers (Huncharek et al., 2010). Thus, as long as society perpetuates messages of masculinity that inflate men's risk of becoming smokers (e.g., risk-taking, invulnerability, etc.) there will continue to be an impact even on those diseases whose causes are dominated by biological factors.

Conclusions

Inevitably, the present article can do little more than provide a broad overview of issues relating to men's health and in particular the relationship between this and societal gender-role expectations. Nevertheless, it is clear that while any understanding of men's health inevitably requires an appreciation of anatomical, physiological, and other biological differences between men and women, such understanding is both insufficient to give a complete picture and to indicate appropriate routes for health improvement. The problem is exacerbated, and likely to continue so, by changes in the patterns of disease in modern society. Where, a century or so ago, things like acute infectious diseases (e.g., smallpox, influenza) accounted for a high proportion of deaths across both sexes, the development of technologies such as antibiotic therapy and immunization has led to the role of such diseases being dramatically reduced (or, in the case of smallpox, eliminated). The diseases of the modern world – cancers, heart disease, hypertension, and so on have much closer links to lifetime patterns of behavior. Moreover, the mortality associated with such diseases may similarly be closely linked to behavior, not just in terms of things like seeking out health care and the availability of screening for early detection, but also in terms of behaviors that influence the seriousness of a disease once it has developed – as we have seen in the example of prostate cancer and smoking. From birth, we are exposed to messages from a variety of sources, which prescribe certain activities and proscribe others, often on the basis of whether we are male or female. There is now ample evidence to link such societal conventions to a variety of health behaviors, only a handful of which it has been possible to outline in the present article. As far as men are concerned, the continued expectation that they will adopt certain lifestyles, behave in particular ways is likely to expose them to increased mortality and morbidity with respect to a variety of health problems. To the extent that they opt out of, or fail to meet such expectations, they are likely to experience feelings of stress and failure, again impacting on both physical and mental health. Looked at from such a perspective, any realistic attempt at improving the health status of men in today's society will need to take into account the demands placed on men by that society.

See also: Cardiac Disease, Coping with; Gender and Health Care; Gender and Physical Health; Sexual Risk Behaviors; Social Support and Recovery from Disease and Medical Procedures.

Bibliography

Byrne, G.J.A., Raphael, B., Arnold, E., 1999. Alcohol consumption and psychological distress in recently widowed older men. Australian and New Zealand Journal of Psychiatry 33, 740–747.

Clayton, P.J., 1990. Bereavement and depression. Journal of Clinical Psychiatry 51, 34–38.

Courtenay, W.H., 2003. Key determinants of the health and the well-being of men and boys. International Journal of Men's Health 2, 1–30.

Duxbury, L., Higgins, C., 1994. Interference between work and family: a status report on dual-career and dual-earner mothers and fathers. Employee Assistance Quarterly 9, 55–80.

Gonça, A., Tomassini, C., Vaupel, J.W., 1999. Pourquoi les femmes survivent aux hommes. La Recherche 322, 96–98.

Greenglass, E.R., Pantony, K.L., Burke, R.J., 1988. A gender-role perspective on role conflict, work stress and social support. Journal of Social Behavior and Personality 3, 317–328.

Gabhainn, S.N., Kelleher, C.C., Naughton, A.M., Carter, F., Flanagan, M., McGrath, M.J., 1999. Socio-demographic variations in perspectives on cardiovascular disease and associated risk factors. Health Education Research 14, 619–628.

Hartley, R.E., 1961. Sex-roles and urban youth: some developmental perspectives. Bulletin on Family Development 2, 1–12.

Harris, J.E., 1983. Cigarette smoking among successive birth cohorts of men and women in the United States during 1900–1980. Journal of the National Cancer Institute 71, 473–479.

Helgeson, V.S., Lepore, S.J., 1997. Men's adjustment to prostate cancer: the role of agency and unmitigated agency. Sex Roles 37, 251–267.

Holtzman, D., Powell-Griner, E., Bolen, J.C., Rhodes, L., 2000. State and sex-specific prevalence of selected characteristics: behavioral risk factor surveillance system 1996 and 1997. Morbidity and Mortality Weekly Report: CDC Surveillance Summaries 49 (6), 1–39.

Huncharek, M., Haddock, K.S., Reid, R., Kupelnick, B., 2010. Smoking as a risk factor for prostate cancer: a meta-analysis of 24 prospective cohort studies. American Journal of Public Health 100, 693–701.

Jorm, A.F., 1994. Characteristics of Australians who reported consulting a psychologist for a health problem: an analysis of data from the 1989–90 National Health Survey. Australian Psychologist 29, 212–215.

Klein, J.D., Brown, J.D., Childers, K.W., Oliveri, J., Porter, C., Dykers, C., 1993. Adolescents' risky behaviour and mass media use. Pediatrics 92, 24–31.

Lee, C.E., 1998. Women's Health; Psychological and Social Perspectives. Sage, London.

Lee, C.E., Owens, R.G., 2002. The Psychology of Men's Health. Open University Press, Buckingham.

Lerner, D.J., Kannel, W.B., 1986. Patterns of coronary heart disease morbidity and mortality in the sexes; a 26-year follow-up of the Framingham population. American Heart Journal 111, 383–390.

Li, G., Baker, S.P., Langlois, J.A., Kelen, G.D., 1998. Are female drivers safer? An application of the decomposition method. Epidemiology 9, 379–384.

Lobo, F., Watkins, G., 1995. Late career unemployment in the 1990s: its impact on the family. Journal of Family Studies 1, 103–111.

Lyons, A.C., Willott, S., 1999. From suet pudding to superhero; representations of men's health for women. Health 3, 283–302.

Macintyre, S., Ford, G., Hunt, K., 1999. Do women 'over-report' morbidity? Men's and women's responses to structured prompting on a standard question on long standing illness. Social Science and Medicine 36, 15–20.

Madigan, F.C., 1957. Are sex mortality differentials biologically caused? Millbank Memorial Fund Quarterly 25, 202–223.

Newman, P.A., Zimmerman, M.A., 2000. Gender differences in HIV-related sexual risk behaviour among urban African American youth: a multivariate approach. AIDS Education and Prevention 12, 308–325.

Pennebaker, J.W., 1997. Opening Up; The Healing Power of Expressing Emotions. Guilford Press, New York.

Pratt, C., Deosaransingh, K., 1997. Gender differences in homicide in Contra Costa County, California: 1982–1993. American Journal of Preventive Medicine 13, 19–24.

Pritchard, C., 1992. Is there a link between suicide in young men and unemployment? A comparison of the UK with other European Community countries. British Journal of Psychiatry 160, 750–756.

Schappert, S.M., November 1999. Ambulatory care visits to physician offices, hospital outpatient departments and emergency departments: United States 1997. In: Vital and Health Statistics – Series 13: Data from the National Health Survey, pp. 1–39.

Sidman, M., 1960. Tactics of Scientific Research. Basic Books, New York.

Stanistreet, D., Bambra, C., Scott-Samuel, A., 2005. Is patriarchy the source of men's higher mortality? Journal of Epidemiology and Community Health 59, 873–876.

Stoverinck, M.J., Lagro-Janssen, A.L., Weel, C.V., 1996. Sex differences in health problems, diagnostic testing and referral in primary care. Journal of Family Practice 43, 567–576.

Troth, A., Peterson, C.C., 2000. Factors predicting safe-sex talk and condom use in early sexual relationships. Health Communication 12, 195–218.

Uitenbroek, D.G., Kerekovska, A., Festchieva, N., 1996. Health lifestyle behaviour and socio-demographic characteristics: a study of Varna, Glasgow and Edinburgh. Social Science and Medicine 43, 367–377.

Umberson, D., 1992. Gender, marital status and the social control of health behaviour. Social Science and Medicine 34, 10–24.

Waldron, I., 1995. Contributions of biological and behavioural factors to changing sex differences in ischaemic heart disease mortality. In: Lopez, A.D., Caselli, G., Valkonen, T. (Eds.), Adult Mortality in Developed Countries; from Description to Explanation. Clarendon, Oxford.

Welsh, M.C., Robinson, T.L., Lindman, L.S., 1998. Sex differences in health attitudes and choice of health behaviors. Psychological Reports 83, 1161–1162.

Williams, J.M., Wright, P., Currie, C.E., Beattie, T.F., 1998. Sports related injuries in Scottish adolescents aged 11–15. British Journal of Sports Medicine 32, 291–296.

Wroblewska, A.M., 1997. Androgenic-anabolic steroids and body dysmorphia in young men. Journal of Psychosomatic Research 42, 225–234.

Wyke, S., Hunt, K., Ford, G., 1998. Gender differences in consulting a general practitioner for common symptoms of minor illness. Social Science and Medicine 46, 901–906.

Mental Causation

Jens Harbecke, Witten/Herdecke University, Witten, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article focuses on the philosophical problem of mental causation as it has been debated in mainstream metaphysics of the last 50 years. After a brief overview of the history of the problem, the assumptions constituting it are made explicit and shown to be inconsistent. In a second step, various arguments supporting canonical solutions to the problem are summarized, including interactionist, epiphenomenalist, identity theorist, and overdeterminationist approaches. The final section highlights the most recent developments of the debate by presenting a family of compatibilist approaches as well as interventionist theories of mental causation.

Introduction

In interpreting the behavior of others, we intuitively credit thoughts, beliefs, desires, emotions, memories, representations, and sensations causal powers on various kinds of observable occurrences, such as body movements, speech acts, and cooperative actions. For instance, we causally explain Suzy's walk to Dairy Queen by her desire to eat ice cream combined with her belief that she can buy it there. Moreover, we take it that Tamar's slowing down of the car was partly caused by her perception of a red stop sign on the road in front of her and her knowledge of its meaning.

Such mental causation speak is not limited to folk psychology. In clinical, cognitive, and developmental psychology, mental causal language is prevalent in the sense that the mind is described as an information-processing system that transforms stimuli inputs into representations. These representations, typically in conjunction with further informational data, are then processed into behavioral outputs.

A fundamental question in the context of such causal ascriptions is whether mental phenomena can in fact be causes of physical phenomena in the physical world, given that the mind is not physical itself in any obvious way. It is now widely believed that this question points to a philosophical problem: the 'problem of mental causation.' It is a philosophical problem in the sense that it is based on a set of assumptions, all of which enjoy credence in isolation, but which cannot all be true.

This article focuses on the philosophical problem of mental causation as it has been debated in mainstream metaphysics of the last 50 years. After a brief overview of the history of the problem, the assumptions constituting it are made explicit and shown to be inconsistent. In a second step, various arguments supporting canonical solutions to the problem are summarized, including interactionist, epiphenomenalist, identity theorist, and overdeterminationist approaches. The final section highlights the most recent developments of the debate by presenting a family of compatibilist approaches as well as interventionist theories of mental causation.

A Brief History

The problem of mental causation is a comparatively young philosophical problem in the sense that ancient and medieval philosophy did not consider questions of mental causality as genuinely problematic. To be sure, the relation of mind and body was already debated in Greek philosophy. Plato (427–347 BCE) in the *Republic* and the *Timaeus* offered an analysis of the human soul (*psyché*) as consisting of three parts all of which inhabit the human body. Two of these, the appetite and the spirit, are intimately connected to the body and die with it. Only the rational ingredient, or the *logos*, is immortal and persists even when the body is destroyed. The logos was conceived of as having causal effects on bodily movements and human actions by playing a part in deliberation and decision making. However, Plato does not seem to have viewed such soul-to-body causation as particularly problematic.

Aristotle (384–322 BCE) in his *De Anima* presented a view according to which nothing in the soul is wholly independent of the body as Plato thought. But neither is the soul a conglomerate of material atoms as Democritus believed. Rather, each soul essentially inheres in a body without being identical to it. When humans act or behave in some way, this is always the result of body and soul working together (cf. Caston, 2009, p. 318). From a modern perspective, the fact that Aristotle did not explain in detail how this 'working together' proceeds precisely is unsatisfactory. For instance, the theory leaves open whether human actions have sufficient bodily causes, or whether they only have sufficient causes that are partly constituted by the activities of body and soul.

The uncritical attitude toward the causal powers of the mental that characterized ancient and much of medieval philosophy of mind changed only when 'physicochemical' thinking gradually spread in early modern philosophy, notably through the works of René Descartes (1596–1650). By invoking a skeptical argument based on the observation of the uncertainty of the outside world, Descartes concluded that the mind and body must be quite distinct kinds of entities, namely two kinds of substance. Substances are traditionally described as entities that can exist all by themselves. In this sense, smiles and colors are not substances, but faces (that can smile) and iron rods (that can be brown) probably are.

Descartes' dualist ontology postulated a thinking substance (*res cogitans*) and an extended substance (*res extensa*). The philosopher did not seem to have been particularly bothered by questions about causal interactions between the substances until in 1643 Elizabeth of Bohemia in a now famous letter demanded an explanation from him of "(…) how man's soul,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63095-7

being only thinking substance, can determine animal spirits so as to cause voluntary actions?" (Adam and Tannery, 1904, III, p. 661). Pierre Gassendi in his *Objections V* remarked in a similar way, "How can there be effort directed against anything, or motions set up in it, unless there is mutual contact between what moves and what is moved? And how can there be contact without a body (…)?" (Cottingham et al., 1985, p. 237). Descartes responded in his *Passions of the Soul* of 1649 (Adam and Tannery, 1904, XI) by arguing that a suitable mediating instance can satisfactorily explain the interaction between different kinds of substances. The philosopher had identified the human pineal gland as the locus of this mediation.

Descartes' reply has convinced few philosophers ever since, but it may have been less ad hoc than it appears at first. Gottfried Wilhelm Leibniz (1646–1716) provided a famous reconstruction of Descartes' theory, according to which the latter was perhaps the first philosopher of mind to recognize a particular kind of conservation principle that strongly constrains the mediating model. This conservation principle contends that the quantity of energy and matter remains constant in any closed physical system. Consequently, causation between *res cogitans* and *res extensa* cannot be a matter of exchanging energy quanta. However, the principle allows for spontaneous alterations of the *vectorial direction* of energy quanta within closed physical systems, and it is here where Descartes saw the mind in action. However, according to Leibniz, Descartes was able to maintain this hypothesis only because "(…) in his times the natural law stating also the direction of matter is preserved was not yet accepted" (*Monadology*, §80, Leibniz, 1890, VI, pp. 620–621). Even if Leibniz was right about Descartes holding this view (cf. McLaughlin, 1993), the question remains how in the pineal gland a nonextended and nonlocalizable mental substance can get in contact with physical substances.

In the centuries following Descartes, virtually all prominent philosophers aimed to provide some answer or other to the problem of mental causation (cf. for instance Leibniz, 1890, 6; Spinoza, 1677, Section 2.1.2; Kant, 1910, 1:19–21; Schopenhauer, 1980, 3, pp.171–173; Huxley, 1893). However, it was only in the second half of the twentieth century that the question of how the mind can be causally active made its way into mainstream metaphysics (cf. Kim, 1997). The problem surfaced already in C.D. Broad's classic *The Mind and Its Place in Nature* of 1925, in Ludwig Wittgenstein's reflections on mentality and mental discourse, as well as in Gilbert Ryle's *The Concept of Mind* (1949). However, its modern metaphysical version was presented by Place (1956), Feigl (1958), and Smart (1959) only in the late 1950s. This version is less concerned with the logic of psychological explanations and focuses on metaphysical causal connections.

Finally, the standard formulation of the problem that still forms the basis of today's debate was proposed in the late 1960s by authors such as Malcolm (1968) and Goldman (1969), who construed the puzzle as a set of incompatible assumptions. The subsequent debate in the 1970s centered on functionalist theories of the mind (cf. Putnam, 1967, 1975; Fodor, 1974). Since functionalism defines mental phenomena as inherently causal and since it subscribes to a token-identity theory, this approach seemed to leave little room for the problem of mental causation. In the late 1970s, however, the feeling spread that also functionalism does not provide a satisfactory solution. Among those philosophers who expressed this worry against functionalism, Kim (1979, 1985) developed the most prominent attacks. From the 1980s onward, the literature on the problem of mental causation has mushroomed, offering a large collection of competing answers to the question of how the mind can have effects in a physical world. A small collection of authors representing the spectrum of positions developed in this late stage of the debate on mental causation is Baumgartner (2009, 2010), Bennett (2003), Bishop (2006), Block (1997, 2003), Bontly (2002), Campbell (1970), Crisp and Warfield (2001), Dardis (2008), Elder (2001), Fodor (1997), Harbecke (2008, 2010), Harbecke and Atmanspacher (2012), Kim (1998, 2003, 2005), Loewer (2007), MacDonald and MacDonald (1986), Marras (2007), Menzies (2008), Menzies and List (2010), Noordhof (1999), Papineau (2002), Raatikainen (2010), Schiffer (1987), Sosa (1984), Weslake (2009), Woodward (2008, 2011), and Yablo (1992a,b). As a result, the problem of mental causation is one of the most widely discussed problems in contemporary philosophy.

The Premises of the Problem

The standard formulation of the problem of mental causation construes it as being constituted by the following four plausible assumptions that, if conjoined, provably imply a contradiction (cf. Harbecke, 2010, pp. 73–74 for a rigorous formalization and a proof for the inconsistency).

(P1)	Some mental events cause physical events.
(P2)	Every physical event has a physical cause.
(P3)	Mental events are not identical to physical events.
(P4)	Physical events are not causally overdetermined.

The following deliberations make clear why the four assumptions all seem plausible in isolation, but are inconsistent with one another nevertheless. Assumption (P1) departs from observations as those mentioned above: Both in everyday and folk-psychological situations as well as in certain scientific contexts, we accept ascriptions of causal powers to mental states or events as highly plausible. Allowing minds to be causes is in fact at the very core of our self-understanding as humans. As a consequence, many of us are convinced that some mental events cause some physical events (premise (P1)).

However, in contrast to the ancient philosophers and early modern thinkers such as Descartes, we now have available overwhelming evidence for the truth of physical principles stating a conservation of mass, energy, and vectorial direction of movement in closed physical systems. Moreover, recent neurophysiological research suggests that bodily movements can often be reliably predicted from prior neural activities. In other words, in the physical world as we know it today, no gaps open up that nonphysical causes could fill. Insights of this kind suggest that any physical event has a sufficient physical cause (premise (P2)). Hence, unless mental phenomena are nothing but physical phenomena, they can only be redundant causes.

But mental phenomena seem to be very different from physical phenomena indeed. Even though most of us would presume our mental life to occur more or less in the same place as our brain activities, opening a skull is quite useless for isolating mental phenomena. All that can be discovered there is gray and white matter, or perhaps neural activities if appropriate imaging techniques are applied.

Moreover, mental states are intentional in the sense that they have content or are 'about' something else and have a nonlocal nature in this sense. Physical phenomena do not have this property of 'aboutness' or nonlocality. But even many purely local mental phenomena are not obviously identical to physical phenomena as the mental kinds that they are an instance of do not seem to be coextensional with any physical kind. For instance, at some future point in time it will probably be possible for a sophisticated robot to entertain the genuine thought that grass is green. However, even then the robot's physical state will differ radically from that of humans when they entertain the thought that grass is green.

Coextensionality is at least a necessary condition for an identity of kinds, and an identity of kinds is usually considered a necessary condition for an identity of the local phenomena. For instance, within a fine-grained model of events (cf. Kim, 1973), the instantiated kinds are part of the identity conditions of events. All of this suggests that mental events are not identical to physical events (premise (P3)).

With respect to causal redundancy, many of us can conceive of cases in which two independent causes are responsible for the same effect. For instance, when Otto was about to enter the road, both Bertha and Anna yelled 'STOP!', which caused him to halt exactly on the edge of the sidewalk (if Otto had not stopped, he would have been hit by a car passing by). Both Bertha's and Anna's yelling 'STOP!' was sufficient for Otto's halting, and neither of the two yells is easily declared as the 'true' cause. But whilst an occasional overdetermination of certain effects seems quite plausible, it is difficult to believe that certain physical effects are systematically overdetermined by physical and mental causes. In other words, physical effects are not systematically causally overdetermined (premise (P4)), and mental events cannot be redundant causes – contradicting what we inferred above. The contradiction shows that premises (P1)–(P4) actually constitute a philosophical problem.

Does the Problem Generalize?

It is now widely believed that the problem of mental causation generalizes in at least four senses (cf. Harbecke, 2008, Section 1.6). The first one concerns the observation that a wholly analogous causal problem can be constructed, for instance, for phenomena referred to by the biological and chemical sciences. Mental phenomena may be unique in the sense that they have content, are not 'visible,' and are possibly external so that the problem of mental causation receives independent intuitive force. However, just like mental phenomena, biological and chemical phenomena are nonfundamental, or 'higher level.' In light of the completeness of physics, it is not immediately clear how such higher level phenomena can be causally efficacious for physical events in general.

The second sense concerns a distinction of 'macro properties' and 'micro-based properties.' Whereas micro-based properties were sometimes believed to be somehow inferable from micro particles, such as the properties of a mineral stone are inferable from the properties of its crystal lattice, macro properties including functional properties were usually considered as less transparently related to micro properties. As a consequence, the problem of mental causation was sometimes believed to concern macro properties only. However, in recent years, it has been shown that, if micro-based properties exist at all and are not identical to micro properties, an analogous problem of micro-based properties emerges (cf. Bontly, 2002).

A third sense in which the problem generalizes points toward the fact that the problem (P1)–(P4) can be reformulated in various ways. For instance, no matter if the causal *relata* mentioned by the premises are events (as done in Section The Premises of the Problem), properties, facts, or states, a version of the problem of mental causation emerges.

The fourth sense concerns the theory of causation presupposed. It is usually assumed that the problem occurs under Humean theories of causation, as well as under process and power theories of causation. It is currently more controversial whether the problem can be identified also within an interventionist framework (cf. Section Interventionsim and Mental Causation).

Canonical Solutions to the Problem

The contradiction implied by (P1)–(P4) allows us to reject any one of the four assumptions through a *reductio* proof. However, deciding which of the assumptions is most likely false is tantamount to choosing a particular solution to the problem of mental causation for which some background argument is required (for an informal explication of these *reductio* proofs, cf. Sturgeon, 1998, pp. 414–432; Harbecke, 2008, 29–31). The four resulting general positions are described in the following subsections.

Rejecting (P1): Epiphenomenalism or Parallelism

There are actually two related theories of mental causation that are based on a rejection of (P1). The first one is an epiphenomenalist position. The second is a parallelism, sometimes also called a 'dual explanandum strategy.'

An epiphenomenon is a phenomenon that is instantiated alongside, or that co-occurs with, some primary phenomenon. In the context of mental causation, the term epiphenomenon is intended to describe phenomena that are caused, but that do not cause anything themselves. In this sense, epiphenomenalism is the view that mental states occur as effects of physical effects and perhaps even regularly alongside certain neural correlates. However, mental causation is illusory. Only the physical events correlated with mental events do real causal work.

One of the earliest explicit defenses of this position was developed by Huxley (1893), who claimed that human beings are nothing but 'conscious automata.' Huxley's view was influenced by the studies of a certain Dr. Mesnet of a French soldier who had suffered severe brain injuries during the Franco–Prussian War of 1870. The soldier was able to carry

out various everyday tasks such as washing and dressing himself, eating whatever was set before him, going to sleep at regular hours, etc. However, he seemed completely unaware of environmental influences such as noise and odor, even of pins run into his skin. As Huxley was not ready to reject dualism, he concluded on the basis of such observations that physical states outplay mental states with respect to bodily behavior.

In the twentieth century, most philosophers no longer believed in substance dualism. The more recent epiphenomenalist positions therefore focus on the causal efficacy of mental types. The central idea driving these positions is that mental events are identical to physical events, but these psychophysical events do not cause anything in virtue of the mental properties they instantiate (cf. Bieri, 1992; Walter 2008a,b; note that the rejected premise (P1) must then be rephrased as 'some mental properties are causally relevant for some physical properties,' and (P2)–(P4) must be rephrased accordingly). The arguments for this position point out that the physical types of the relevant events are already sufficient for the corresponding behavioral effects and the mental types are at best redundant. Others state that mental properties are essentially external, or holistic, in character. Since causation is a local matter, mental properties cannot have causal relevance (cf. Fodor, 1987, pp. 27–54).

A special kind of epiphenomenalist theory is based on an argument for the possibility of 'Zombies' and the existence of qualitative experiences. Zombies are creatures that are just like humans physiologically, but that do not enjoy any conscious experiences. If Zombies really are possible, then mind and body are separable in principle, and the mind is not identical to the physical. But also then the mind does not contribute anything to what goes on in the world and, hence, is epiphenomenal (cf. Kripke, 1972; Jackson, 1982, 1986; Chalmers, 1996).

Counterarguments against epiphenomenalist positions have emphasized that denying mental causation would falsify almost everything humans believe about themselves and their environment. Since such a massive misunderstanding is not plausible in light of the success of our explanations of other people's behavior as well as our own, the existence of mental causation should be considered an undeniable fact (cf. Baker, 1993; Burge, 1993). Moreover, epiphenomenalism is in conflict with the widely accepted Eleatic principle in metaphysics, which is meant to rule out extravagant ontologies. The principle demands that nothing exists that does not have a causal effect some time and somewhere. Hence, if epiphenomenalism denies the causal powers of the mental, it thereby denies the very existence of mental phenomena.

The above-mentioned parallelist, or dual explanandum, strategies typically deny that mental phenomena cause physical phenomena, but claim that mental phenomena only cause further mental or other nonphysical phenomena. An early version of parallelism is found in Leibniz' *Monadology*. In Leibniz' view, God has designed the world in a way that lets mental and physical substances develop in time without interaction. Just as two clocks displaying exactly the same time may appear to be magically connected, the mental and the physical run parallel in complete harmony with another, but psychophysical interaction is an illusion.

More recent approaches argue that mental causation takes place only within the level of the mental (cf. Gibbons, 2006) or that ascriptions of mental causes typically cite different entities as effects than do causal explanations of bodily behavior (cf. Hornsby, 1986). Some compatibilists are also drawn toward a more complex kind of parallelism (cf. Section New Compatibilism and Harbecke, 2008, Chapter 3). These modern theories reject mental substances and locate mental causation on a type level. A vibrant debate has developed over the question whether such approaches are viable (cf. Kim, 2003; Harbecke, 2010; Harbecke and Atmanspacher, 2012).

Rejecting (P2): Interactionism

Denying (P2), i.e., the causal completeness of the physical, has become known as an 'interactionist dualism,' under which the position defended by Descartes should be subsumed (cf. Section A Brief History). Today only few philosophers subscribe to substance dualism, among which are Swinburne (1994, 1997) and Foster (1991, 1994). One argument brought forward by these modern dualists points out that the folk perception of matter as impenetrable and persistent is false. The physical sciences of today say little about what material things are in themselves. Rather, physics states only relational properties of physical entities. As a consequence, it is in principle possible to construct two observationally equivalent theories of our world T_1 and T_2, where T_1 describes the nature of all fundamental particles as physical and T_2 makes the same predictions but describes some of the fundamental particles to be of a mental nature. Hence, substance dualism is reasonable even in light of modern science. A second argument for interactionism builds on the conceivability, and therefore possibility, of a situation in which a conscious being survives the destruction of its body. The possibility implies that, already when the body and mind coexist, they are nonidentical.

The two general arguments have drawn criticisms concerning their validity as well as their motivation. In particular, the road from conceivability to possibility has been called into question (cf. Hill, 1997). It should also be noted that, even if the argument for substance dualism is correct, it is still not clear why then mental causes should have an influence on the physical domain.

That it is possible to challenge (P2) even without subscribing to any kind of substance dualism has recently been argued for by authors such as Hendry et al. (2006), Bishop (2008), and Bishop and Atmanspacher (2011). Whether these scientific arguments against the completeness of physics provide good reasons for mental-to-physical causation is still subject to an ongoing debate.

Rejecting (P3): Identity Theories

The rejection of (P3) is a popular move made by identity theorists. The claim that not all mental events are physical has been held in combination with a type nonidentity, or 'nonreductionist' (Davidson, 1970, 1993; Fodor, 1974, 1997; MacDonald and MacDonald, 1986) as well as a type identity, or reductionist (Lewis, 1966; Kim, 1998) theory.

One version of a nonreductionist token-identity theory is anomalous monism as promoted by Davidson (1970, 1993).

Davidson's main claim is that there are no psychophysical laws since mental types are holistically individuated whilst physical types are not. And since lawhood is essential for causation, mental events can only cause physical events qua being physical. As a second version, functionalist theories of the mind usually point out that mental types are multiply realizable by, and therefore not identical to, physical kinds (cf. Putnam, 1967, 1975; Fodor, 1974). Only mental tokens are identical to physical tokens.

Critics of functionalism such as Honderich (1982) and Kim (1993a,b) have attacked nonreductionist identity theories on the basis of the insight that events always cause further events in virtue of some of their properties. But if mental events do not cause any physical events in virtue of their mental properties and only in virtue of their physical properties as Davidson claims, then the token version of the identity theory does not provide a convincing theory of mental causation despite its rejection of (P3). Moreover, functionalism has been attacked on the grounds that mental kinds are always screened off by physical kinds with respect to their alleged physical effects (cf. Kim, 1979, 1985, 1998, 2003).

In light of these conclusions, several authors have tried to demonstrate that, contrary to initial appearance, mental properties are reducible to physical properties. The idea is that a reduction always preserves the ontology as well as the causal relevance of mental properties. Most prominently, Kim (1998, 2005) has promoted a technique to identify mental with physical properties, which has become known as 'functional reduction.' The original idea goes back to Lewis and Ramsey (cf. Lewis, 1966, 1972, 1979). According to this strategy, mental properties should first be defined wholly in terms of their inputs and outputs; i.e., they should be functionally construed. In the second step, the physical properties that have exactly the same inputs and outputs should be identified. The third step consists of the identification of the mental and the physical properties on the basis of their functional resemblance.

If things go well, for every mental property, precisely one physical property satisfies the functional definition. However, since it may well seem possible that a long list of physical properties all satisfy the same functional description definitional of a single mental state, it is unclear whether Kim gets any further than the token-identity theorists. To answer to this worry, some authors have tried to develop ways of preserving the mental ontology even in light of multiple realization. For example, Bickle (1998) argues that sometimes even reduction on disjunctive properties can be ontology-preserving. In contrast, Esfeld and Sachse (2007) try to demonstrate that with the introduction of mental subtypes it is possible to attain bijective reductions.

Rejecting (P4): Overdeterminationism

By applying a *reductio ad absurdum* inference against (P4), one accepts an overdeterminationist position, according to which mental causes, if they are causes of physical events at all, are always redundant causes. One version of this account has been developed by Loewer (2007). The author accepts that every physical event has a sufficient physical cause, but argues that certain mental events are counterfactually necessary for some physical events. This conclusion is then taken to establish at least a 'light' version of mental causation. As a second example, Bennett (2003) has explicated a counterfactual test for genuine overdetermination, which is a causation of a single effect by two counterfactually independent causes. Since mental events are typically not counterfactually independent of the physical causes of their putative effects, the test shows that the overdetermination described in the context of mental causation is of an unproblematic kind and should therefore be accepted as real.

Other authors such as Mills (1996) and Sider (2003) have critically reviewed the metaphysical principle ruling out a genuine systematic overdetermination of physical events. Mills demonstrates that the determining role by mental events becomes visible in counterfactual contexts, whilst Sider shows that systematic overdetermination is not in itself an incoherent assumption.

Recent Developments

Two main topics have been at the focus of the recent debate on mental causation. The first one has been described as a 'new compatibilism' in the sense that it strives to make a psychophysical nonidentity and mental causality compatible. The second one has focused on the implications of the interventionist theory of causation for the case of mental causation. This section provides a brief overview on both topics.

New Compatibilism

The core idea that unifies several approaches under the term new compatibilism is that mental and physical phenomena are neither identical nor wholly distinct. The idea is implicit in Yablo (1992a, 1992b), where the relation between mental and physical events is described as 'metaphysical determination.' Pereboom and Kornblith (1991) have promoted the idea that mental phenomena are 'nonidentical but wholly constituted' by physical phenomena. The idea surfaces also in the works of Bennett (2003), Shoemaker (2001), and Harbecke (2008).

All of these contributions take the fact that mental and physical events are not wholly distinct to imply that they cannot causally compete with another. This result corresponds to intuitions we have about colors such as 'red' and 'crimson,' which are also very intimately related but do not compete for causal relevance. For instance, it would be quite absurd to claim that the redness of the flag was not the cause of the bull's chasing it, because the crimsonness of the flag was already sufficient for the chase. It may seem as though the new compatibilists reject (P4) and should therefore be described as overdeterminationists. This, however, is not the case for those new compatibilists who combine their view with a principle of causal proportionality. According to this principle, causes must resemble their effects in specificity and detail. Hence, complex microphysical events do not qualify as proportionate to macro events such as a walk to the ice cream store, stopping a car, etc. The intuition is that some mental events will in fact turn out proportionate to events of this kind. This result would mean that the new compatibilists either reject (P1) (cf. Section Rejecting (P1): Epiphenomenalism or Parallelism), or hold onto premises (P1)–(P4) but identify an ambiguity in

the premises with respect to the word "physical" (cf. Sturgeon, 1998).

Interventionism and Mental Causation

Some authors have argued that interventionist theories of causation shed new light on the problem of mental causation. According to interventionism, causal relations can be analyzed and assessed only with respect to a causal model. And there are good reasons to believe that mental events (represented as mental variables in a model) satisfy all the relevant causal criteria with respect to some physical events in some validated causal models. That is, in some models an independent intervention on a mental variable induces a change in a physical variable (cf. Menzies, 2008; Menzies and List, 2010; Weslake, 2009; Raatikainen, 2010; Woodward, 2008).

The problem with this observation seems to be that any model that qualifies mental variables as causes can be extended into a model in which the mental variables fail to be causes. This is due to the widely accepted assumption of a psychophysical supervenience (cf. Kim, 1984), according to which any mental variable necessarily depends on, and is determined by, some physical variable. In this case the requirement of an independent intervention on the mental variables is no longer satisfied (cf. Baumgartner, 2009, 2010). A potential reply would be that the original toolbox of interventionism was not designed to include metaphysically related variables into the model and that, consequently, the putative violation of the independent intervention requirement is vitiated (cf. Woodward, 2011). Whether this reply is acceptable or whether it is merely ad hoc is the topic of an ongoing debate.

See also: Functional Explanation: Philosophical Aspects; Rational Choice Explanation: Philosophical Aspects.

Bibliography

Adam, C., Tannery, P., 1904. Oeuvres de Descartes, vol. 12. J. Vrin, Paris.

Baker, L., 1993. Metaphysics and Mental Causation (See Heil and Mele (1993), pp. 75–95).

Baumgartner, M., 2009. Interventionist causal exclusion and non-reductive physicalism. International Studies in the Philosophy of Science 23 (2), 161–178.

Baumgartner, M., 2010. Interventionism and epiphenomenalism. Canadian Journal of Philosophy 40 (3), 359–383.

Bennett, K., 2003. Why the exclusion problem seems intractable, and how, just maybe, to tract it. Noûs 37 (3), 471–497.

Bickle, J., 1998. Psychoneural Reduction: The New Wave. MIT Press, Boston.

Bieri, P., 1992. Trying out epiphenomenalism. Erkenntnis 36 (3), 283–309.

Bishop, R., 2006. The hidden premises in the causal argument for physicalism. Analysis 66 (1), 44–52.

Bishop, R., 2008. Downward causation in fluid convection. Synthese 160 (2), 229–248.

Bishop, R., Atmanspacher, H., 2011. Causal closure of physics and free will. In: Kane, R. (Ed.), Oxford Handbook of Free Will. Oxford University Press, Oxford, pp. 101–114.

Block, N., 1997. Anti-reductionism slaps back. Noûs 31 (Supplement: Philosophical Perspectives, 11, Mind, Causation, and World, 107–132).

Block, N., 2003. Do causal powers drain away? Philosophy and Phenomenological Research 67 (1), 133–150.

Bontly, T., 2002. The supervenience argument generalizes. Philosophical Studies 109 (1), 75–96.

Burge, T., 1993. Mind–Body Causation and Explanatory Practice (See Heil and Mele (1993), pp. 97–120).

Campbell, K., 1970. Body and Mind. Doubleday Anchor Books, New York.

Caston, V., 2009. Aristotle's psychology. In: Gill, M., Pellegrin, P. (Eds.), A Companion to Ancient Philosophy. Wiley-Blackwell, Oxford, pp. 316–346.

Chalmers, D., 1996. The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press, USA.

Cottingham, J., Stoothoff, R., Murdoch, D., 1985. The Philosophical Writings of Descartes, vol. 2. Cambridge University Press, Cambridge.

Crisp, T., Warfield, T., 2001. Kim's master argument. Noûs 35 (2), 304–316.

Dardis, A., 2008. Mental Causation: The Mind–Body Problem. Columbia University Press, New York.

Davidson, D., 1993. Thinking Causes (See Heil and Mele (1993), pp. 3–18).

Davidson, D., 2001/1970. Mental events. In: Davidson, D. (Ed.), Essays on Actions and Events. Oxford University Press, New York, pp. 207–225.

Elder, C., 2001. Mental causation versus physical causation: no contest. Philosophical and Phenomenological Research 62 (1), 111–127.

Esfeld, M., Sachse, C., 2007. Theory reduction by means of functional sub-types. International Studies in the Philosophy of Science 21 (1), 1–17.

Feigl, H., 1967/1958. The Mental and the Physical: The Essay and a Postscript. University of Minnesota Press, Minneapolis.

Fodor, J., 1974. Special sciences (or: the disunity of science as a working hypothesis). Synthese 28 (2), 97–115.

Fodor, J., 1987. Psychosemantics: The Problem of Meaning in the Philosophy of Mind. MIT Press, Cambridge.

Fodor, J., 1997. Special sciences: still autonomous after all these years. Noûs 31 (Supplement: Philosophical Perspectives, 11, Mind, Causation, and World, 149–163).

Foster, J., 1991. The Immaterial Self: A Defence of the Cartesian Dualist Conception of the Mind. Routledge, London.

Foster, J., 1994. The token-identity thesis. In: Warner, R., Szubka, T. (Eds.), The Mind–Body Problem: A Guide to the Current Debate. Blackwell, Oxford, pp. 299–310.

Gibbons, J., 2006. Mental causation without downward causation. The Philosophical Review 115 (1), 79–103.

Goldman, A., 1969. The compatibility of mechanism and purpose. The Philosophical Review 78 (4), 468–482.

Harbecke, J., 2008. Mental Causation. Investigating the Mind's Powers in a Natural World. Ontos, Frankfurt a. M.

Harbecke, J., 2010. The problem of mental causation formalized. Mind and Matter 8 (1), 63–91.

Harbecke, J., Atmanspacher, H., 2012. Horizontal and vertical determination of mental and neural states. Journal of Theoretical and Philosophical Psychology 32 (3), 161–179.

Heil, J., Mele, A., 1993. Mental Causation. Oxford University Press, New York.

Hendry, R., Baird, D., Scerri, E., McIntyre, L., 2006. Is there downward causation in chemistry?. In: Philosophy of Chemistry: Synthesis of a New Discipline, Volume 242 of Boston Studies in the Philosophy of Science. Springer, Dordrecht, pp. 173–189.

Hill, C., 1997. Imaginability, conceivability, possibility and the mind–body problem. Philosophical Studies 87 (1), 61–85.

Hohwy, J., Kallestrup, J., 2008. Being Reduced: New Essays on Reduction, Explanation, and Causation. Oxford University Press, USA, New York.

Honderich, T., 1982. The argument for anomalous monism. Analysis 42 (1), 59–64.

Hornsby, J., 1986. Physicalist thinking and conceptions of behaviour. In: McDowell, J., Philip, P. (Eds.), Subject, Thought, and Context. Clarendon Press, pp. 95–115.

Huxley, T.H., 1893. On the hypothesis that animals are automata, and its history. In: Huxley, T.H. (Ed.), Collected Essays. Method and Results, vol. 1. Macmillan, London, pp. 195–250.

Jackson, F., 1982. Epiphenomenal qualia. The Philosophical Quarterly 32 (127), 127–136.

Jackson, F., 1986. What Mary didn't know. The Journal of Philosophy 83 (5), 291–295.

Kant, I., 1902–1910. Gesammelte Schriften. Reimer, Berlin.

Kim, J., 1973. Causation, nomic subsumption, and the concept of event. The Journal of Philosophy 70 (8), 217–236.

Kim, J., 1979. Causality, identity, and supervenience in the mind–body problem. Midwest Studies in Philosophy 4 (1), 31–49.

Kim, J., 1984. Concepts of supervenience. Philosophy and Phenomenological Research 45 (2), 153–176.

Kim, J., 1985. Supervenience, determination, and reduction. The Journal of Philosophy 82 (11), 616–618.

Kim, J., 1993a. Can Supervenience and Non-strict Laws Save Anomalous Monism? (See Heil and Mele (1993), pp. 19–26).

Kim, J., 1993b. The Nonreductivist's Troubles with Mental Causation (See Heil and Mele (1993), Chapter 11, pp. 23–42).

Kim, J., 1997. The mind–body problem: taking stock after forty years. Nous 31, 185–207.
Kim, J., 1998. Mind in a Physical World: An Essay on the Mind–Body Problem and Mental Causation. MIT Press, Cambridge.
Kim, J., 2003. Blocking causal drainage and other maintenance chores with mental causation. Philosophy and Phenomenological Research 67 (1), 151–176.
Kim, J., 2005. Physicalism, or Something Near Enough. Princeton University Press, Princeton.
Kripke, S., 1972. Naming and Necessity. Harvard University Press, Cambridge, MA.
Leibniz, G.W., 1875–1890. Saemtliche Schriften und Briefe. Akademie-Verlag, Berlin.
Lewis, D., 1966. An argument for the identity theory. The Journal of Philosophy 63 (1), 17–25.
Lewis, D., 1972. Psychophysical and theoretical identifications. Australasian Journal of Philosophy 50 (3), 249–258.
Lewis, D., 1979. How to define theoretical terms. The Journal of Philosophy 117 (13), 427–446.
Loewer, B., 2007. Mental causation, or something near enough. In: McLaughlin, B., Cohen, J. (Eds.), Contemporary Debates in Philosophy of Mind. Blackwell Publishers, Oxford, pp. 243–264.
MacDonald, C., MacDonald, G., 1986. Mental causes and explanation of action. The Philosophical Quarterly 36 (143), 145–158.
Malcolm, N., 1968. The conceivability of mechanism. The Philosophical Review 77 (1), 45–72.
Marras, A., 2007. Kim's supervenience argument and nonreductive physicalism. Erkenntnis 66 (3), 305–327.
McLaughlin, P., 1993. Descartes on mind–body interaction and the conservation of motion. The Philosophical Review 102 (2), 155–182.
Menzies, P., 2008. The Exclusion Problem, the Determination Relation, and Contrastive Causation (See Hohwy and Kallestrup (2008), New York, pp. 196–217).
Menzies, P., List, C., 2010. The causal autonomy of the special sciences. In: MacDonald, C., MacDonald, G. (Eds.), Emergence in Mind. Oxford University Press.
Mills, E., 1996. Interactionism and overdetermination. American Philosophical Quarterly 33 (1), 105–117.
Noordhof, P., 1999. Micro-based properties and the supervenience argument: a response to Kim. Proceedings of the Aristotelian Society 99, 109–114.
Papineau, D., 2002. Thinking about Consciousness. Oxford University Press, New York.
Pereboom, D., Kornblith, H., 1991. The metaphysics of irreducibility. Philosophical Studies 63 (2), 125–145.
Place, U.T., 1956. Is consciousness a brain process? British Journal of Psychology 47, 44–50.
Putnam, H., 1975. Philosophy and our mental life. In: Putnam, H. (Ed.), Mind, Language and Reality: Philosophical Papers, vol. 2. Cambridge University Press, Cambridge, pp. 291–303.
Putnam, H., 1980/1967. The nature of mental states. In: Block, N. (Ed.), Readings in Philosophy of Psychology, vol. 1. Harvard University Press, Cambridge, pp. 223–231.
Raatikainen, P., 2010. Causation, exclusion, and the special sciences. Erkenntnis 73 (3), 349–363.
Schiffer, S., 1987. Remnants of Meaning. MIT Press, Cambridge, MA.
Schopenhauer, A., 1980. Sämtliche Werke. Wissenschaftliche Buchgesellschaft, Darmstadt.
Shoemaker, S., 2001. Realization and mental causation. In: Gillett, C., Loewer, B. (Eds.), Physicalism and Its Discontents. Cambridge University Press, Cambridge/New York, pp. 74–98.
Sider, T., 2003. What's so bad about overdetermination? Philosophy and Phenomenological Research 67 (3), 719–726.
Smart, J., 1959. Sensations and brain processes. The Philosophical Review 68 (2), 141–156.
Sosa, E., 1984. Mind–body interaction and supervenient causation. Midwest Studies in Philosophy 9 (1), 271–281.
Spinoza, B. d, 1999(1677). Ethica Ordine Geometrico Demonstrata et in quinque Partes distincta. Meiner, Hamburg.
Sturgeon, S., 1998. Physicalism and overdetermination. Mind 107 (426), 411–431.
Swinburne, R., 1994. Body and soul. In: Warner, R., Szubka, T. (Eds.), The Mind–Body Problem: A Guide to the Current Debate. Blackwell, Oxford, pp. 311–316.
Swinburne, R., 1997. The Evolution of the Soul, revised ed. Clarendon Press, Oxford.
Walter, S., 2008a. Epiphenomenalism and the notion of causation. In: McLaughlin, B., Cohen, J. (Eds.), Gehirne und Personen. Ontos, Frankfurt a. M., pp. 227–236.
Walter, S., 2008b. Ist der Epiphaenomenalismus absurd? Ein neuer Blick auf eine tot geglaubte Position. Zeitschrift Fuer Philosophische Forschung 62, 93–110.
Weslake, B., 2009. Exclusion Excluded (unpublished). http://bweslake.s3.amazonaws.com/research/papers/weslake_exclusion.pdf.
Woodward, J., 2008. Mental Causation and Neural Mechanisms (See Hohwy and Kallestrup (2008), pp. 218–262).
Woodward, J., 2011. Interventionism and Causal Exclusion. http://philsci-archive.pitt.edu/8651/.
Yablo, S., 1992a. Cause and essence. Synthese 93 (3), 403–449.
Yablo, S., 1992b. Mental causation. The Philosophical Review 101 (2), 245–280.

Mental Disorders, Epidemiology of

Katja Beesdo-Baum, Institute of Clinical Psychology and Psychotherapy, Behavioral Epidemiology, Technische Universität Dresden, Dresden, Germany
Hans-Ulrich Wittchen, Institute of Clinical Psychology and Psychotherapy, Technische Universität Dresden, Dresden, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Epidemiology in the field of mental disorders is the study of their distribution in populations and of the risk factors associated with their onset and course. The article summarizes considerable advances in terms of number of studies, their scope, the degree of methodological sophistication, and their linkages to the allied disciplines and points to deficits with regard to longitudinal and causal analytic epidemiological studies. These designs are of core importance to understand the natural course, the longitudinal stability of subsymptoms, and comorbidity associations as well as core vulnerability and risk factors for the onset and persistence of mental disorders.

Epidemiology in the field of mental disorders is the study of their distribution in populations and of the risk factors associated with their onset and course. During the past decade, there have been considerable advances in terms of number of studies, their scope, the degree of methodological sophistication, and their linkages to allied disciplines. There is now increased national and international collaboration on the basis of more stringent methodologies and designs, which allow for powerful coordinated cross-national reanalyses with regard to prevalence, age of onset, disability, and demographic correlates (e.g., World Mental Health Survey Initiative (Kessler et al., 2011)). However, deficits remain, particularly, with regard to longitudinal and causal analytic epidemiological studies. These designs are of core importance to understand the natural course, the longitudinal stability of subsymptoms, and the nature and development of comorbid constellations as well as core vulnerability and risk factors for the onset and persistence of mental disorders. Such studies can also provide more convincing evidence for the clinical and nosological validation of mental disorders in terms of their prognostic value and stability, family and genetic findings, as well as laboratory findings, ultimately allowing for a sharper genotypic and phenotypic classification. Beyond the continuing need for descriptive epidemiological studies that play an increasingly important role in basic clinical and public health research, the key challenge of epidemiology will be to gain a deeper understanding of how multiple risk and vulnerability factors interact over the life span in producing one or multiple mental disorders over brief or longer periods of time.

Epidemiological methods provide tools to conduct research with regard to the distribution of mental disorders in populations, their impact, and their etiology, and serve as the basis of outcome studies, clinical trials, and public health research. Key requirements of epidemiological studies are (1) the definition of the target population (e.g., total population or representative samples of a region or a country, or representative fractions thereof); (2) explicit, reliable, and valid criteria for disorders or what constitutes a case (e.g., symptoms or syndromes); (3) explicit, reliable, and valid criteria for variables and factors that might be associated with a disease (i.e., gender, social class, genetic factors, infectious agents), and (4) use of epidemiological methods for measuring outcome occurrence (e.g., *prevalence*, defined as the proportion of individuals affected by a particular disorder at a specified time period or point in time, and *incidence*, defined as the number of new cases among persons in a population, who were initially free from the disorder and developed the disorder over a given period of time, such as a lifetime, 1 year, or 30 days), and for measuring associations (risk and protective factors), and impact (i.e., course of illness, associated impairments, and disability). Epidemiology can be divided further into two interrelated orientations and methodologies, namely, *descriptive epidemiology*, aiming at measuring the extent of mental disorders in the community, and *analytic epidemiology* (e.g., longitudinal cohort designs, case–control studies), that focuses on understanding the natural course and etiology of mental disorders. Epidemiology offers some unique and promising research strategies for describing and clarifying the nature, etiology, course, and outcome of mental disorders, because patients in treatment settings usually represent a small and highly selective segment of the full spectrum of mental disorders. Thus, findings for risk factors and prognosis might be biased by selection biases as well as the severity of the studied condition.

Originally, epidemiology and epidemiological research designs were developed to study chronic and infectious diseases, and have only been adapted slowly for use in mental and behavioral disorders since the 1950s, largely due to the initially controversial status of past diagnostic classification systems for mental disorders. The complex manifestations, poorly understood etiologies, and variability of the course of mental disorders, in addition to their previously low diagnostic reliability, were often difficult to capture in basic epidemiological designs. This difficulty was particularly true for studies involving one or two points in time (e.g., cohort studies). In addition, risk factors for mental disorders can be as difficult to conceptualize and assess as variables of relevance for mental disorders. Despite early difficulties, these problems have been partly overcome in recent years.

Strongly related to the introduction of more reliable and valid classification systems for mental disorders, based on explicit criteria and operationalized diagnoses in 1980 (APA,

Table 1 Epidemiology of mental disorders in the first decade of the twenty-first century

Progress	*Deficits*
Broad availability of large-scale general population studies (prevalence, sociodemographic correlates, comorbidity, age of onset)	Limited number of studies in children, adolescents, and the elderly
More comprehensive diagnostic spectrum (e.g., childhood and adolescent disorders, somatoform disorders, eating disorders, sleep disorders, comorbidity with somatic conditions)	Incomplete coverage of diagnoses (deficits: psychotic disorders, personality disorders)
Increasingly sophisticated sampling, design, and statistical procedures	Few prospective longitudinal, multiple wave studies
International collaboration (joint reanalyses on prevalence, risk factors, comorbidity)	Explorations of thresholds for and boundaries between disorders
Documentation of associated impairments and disabilities	Assessment of diagnosis-specific disability and impairment
Crude documentation of poor recognition and intervention	Service utilization and need assessment
Improvements in diagnostic instruments and contributions to psychopathology and diagnostic classification	Assessment of natural course and persistence, developmental pathways, and trajectories
Increasing number of prospective longitudinal and family genetic studies, mostly in children and adolescents: contributions to nosological research (identification of vulnerabilities and risk factors) and natural course	Linkage to genetic research and neurobiology, developmental continuity of adult mental disorders

1980) and the increasing availability of structured and standardized diagnostic assessment instruments, the past three decades witnessed an unprecedented progress in epidemiological research on mental disorders. The field has tremendously advanced in terms of the number of studies, their scope, the degree of methodological sophistication, and their linkages to allied disciplines such as psychology, neurobiology, and sociology (see **Table 1**). What started in the 1950s as a scientifically problematic and quite restricted area of 'psychiatric epidemiology' (Robins, 1992) has now reached firmer ground and opened up new perspectives.

Descriptive Epidemiological Findings

Starting with the landmark Epidemiological Catchment Area program in the early 1980s (Robins and Regier, 1991), increasingly sophisticated large-scale studies in the general population have made it evident that mental disorders are very frequent disorders of the brain, affecting more persons than not over their life course. These studies have also highlighted the correlates and the variability in the manifestations of disorders of emotion, cognition, physiology, and behavior, demonstrating that mental disorders are not as uniform as previously believed in terms of their risk factors, courses, outcome, associated disabilities, and impairments.

Prevalence and Correlates of Mental Disorders in the Community

Cross-sectional community surveys such as the US National Comorbidity Survey Replication (NCS-R) (Kessler et al., 2012) and similar studies in other countries around the world (see World Mental Health Survey Initiative, Kessler et al., 2011) have estimated the lifetime rates for mood, anxiety, impulse control, and substance use disorders alone to range between 20 and 48% with anxiety disorders being the most common group in most countries followed by mood disorders and substance use disorders (Wang et al., 2011). The projected estimated lifetime risk up to the age of 75 years is even higher with rates up to 56%. Prospective longitudinal studies that examine mental disorders in the same cohort at multiple time points provide evidence that cumulative rates are even considerably higher (Beesdo et al., 2010; Copeland et al., 2011; Moffitt et al., 2010). The 12-month prevalence estimates vary around 10–14% for anxiety, 5–8% for mood, and 2–4% for substance use disorders, in spite of differences in coverage of diagnoses, design, and cultural setting (Kessler et al., 2011; Wittchen et al., 2011).

Similar convergent evidence has also become available from these studies with regard to sociodemographic correlates and impairment. For example, (1) the majority of people affected by a mental disorder usually report the onset of their condition before adulthood, largely due to a predominant early onset of anxiety disorders (childhood or adolescence) and substance use disorders (in adolescence), whereas depressive disorders occur over the whole life span; (2) females are 2–3 times more frequently affected by various depressive and anxiety disorders; (3) males develop substance use and antisocial personality disorders more frequently; (4) low social class has not been confirmed consistently as being associated with the onset of mental disorder in all countries; and (5) the presence of mental disorders has been found to be associated consistently with increased rates of disability and impairment days, which vary by the type of disorder and comorbidity.

The World Mental Health Survey Initiative (Kessler et al., 2011) also revealed that even in developed countries only about half of those with the most severe disorders receive any treatment, irrespective of appropriateness in terms of type, dose, and duration. The vast majority of cases were cared for in the primary health care sector, with only few receiving treatment by mental health specialists. Overall, only about one-fourth of all cases in developed countries receive any minimally adequate treatment. A further disturbing finding is the fact that the majority of patients delay many years after the onset of their disorder before getting treatment. It is not entirely clear to what degree these low treatment rates are due to patients' poor help-seeking behavior, structural barriers in the health care system, or health service providers' lack of recognition and diagnostic skills. However, the fact that low treatment rates are not confined to countries that do not cover health care and treatment cost by insurance plans suggests that there are many reasons for this problem.

Another important finding of studies is that, even in community samples, a high proportion of subjects suffer from more than one mental disorder. In the NCS-R (Kessler et al., 2005a), the majority of people with one lifetime mental disorder had comorbid conditions as well; overall, 17% of the population had three or more lifetime conditions. Among those with a 12-month diagnosis, 22% had two and 23% had three or more diagnoses (Kessler et al., 2005b). Most frequent patterns of comorbidity were among anxiety, mood, and substance use disorders. The consistency of comorbidity findings across studies has made it clear that comorbidity is a fundamental characteristic of most mental disorders. Comorbidity has been shown to be neither simply an artifact of random association of frequent disorders nor influenced by help-seeking behavior or methodological aspects of the studies (Wittchen, 1996). Comorbidity has specific effects on the degree of suffering as well as the likelihood of seeking professional help. Research into the meaning and the implications of comorbidity has received considerable attention during the past decade resulting even in proposals for a new simplistic 'meta-structure' of mental disorders, which, however, insufficiently reflects the developmental variations in the expression of psychopathology (Wittchen et al., 2009).

Prevalence of Mental Disorders in Primary Care

Although large-scale comprehensive studies about the prevalence of the range of mental disorders in primary care are lacking, there is considerable evidence from international collaborative primary care studies on selected disorders (Üstün and Sartorius, 1995) that these disorders are highly prevalent in primary care settings around the world: point prevalence estimates for depressive disorder range between 6 and 11%, and those for anxiety disorders range between 4 and 10%, with some more recent studies indicating even higher rates (Ansseau et al., 2004). Such studies have also revealed that mental disorders in primary care are usually poorly recognized and rarely treated (Wittchen et al., 2002). Only every second case is recognized, and of those recognized, only one-third receives some type of state-of-the-art treatment.

The Burden of Mental Disorders

Along with the increased emphasis of epidemiological research on the role of general and diagnosis-specific disabilities and impairment, community studies have also been able to demonstrate the substantial burden that mental disorders have on the subjects' lives, social functioning, and interpersonal environment. Based on available epidemiological data, the Global Burden of Disease study (Murray and Lopez, 1996; Murray et al., 2012; Vos et al., 2012) showed that the burden of mental disorders has been underestimated greatly. Mental and behavioral disorders accounted for almost one-quarter of all years lived with disability (YLDs) and have increased by over one-third from 1990 to 2010. Major depressive disorders alone caused 63 million YLDs and ranked number 2 among all causes across all ages (**Figure 1**), and number one for ages 10–29 years. Anxiety disorders, drug use disorders, alcohol use disorders, schizophrenia, bipolar disorders, and dysthymia were also major causes of global YLDs. Given these numbers it is not surprising that a recent study on the cost of disorders of the brain in Europe in 2010 revealed that mental disorders result in tremendous costs (overall €461 billion in 2010), with mood disorders being the most costly condition of all disorders of the brain (Gustavsson et al., 2011). Indirect health care costs usually exceed direct health care costs for mental disorders because people are not able to work.

Progress and Unresolved Issues in Descriptive Epidemiology

Despite the fact that considerable progress has been made in the past two decades, descriptive epidemiological knowledge is

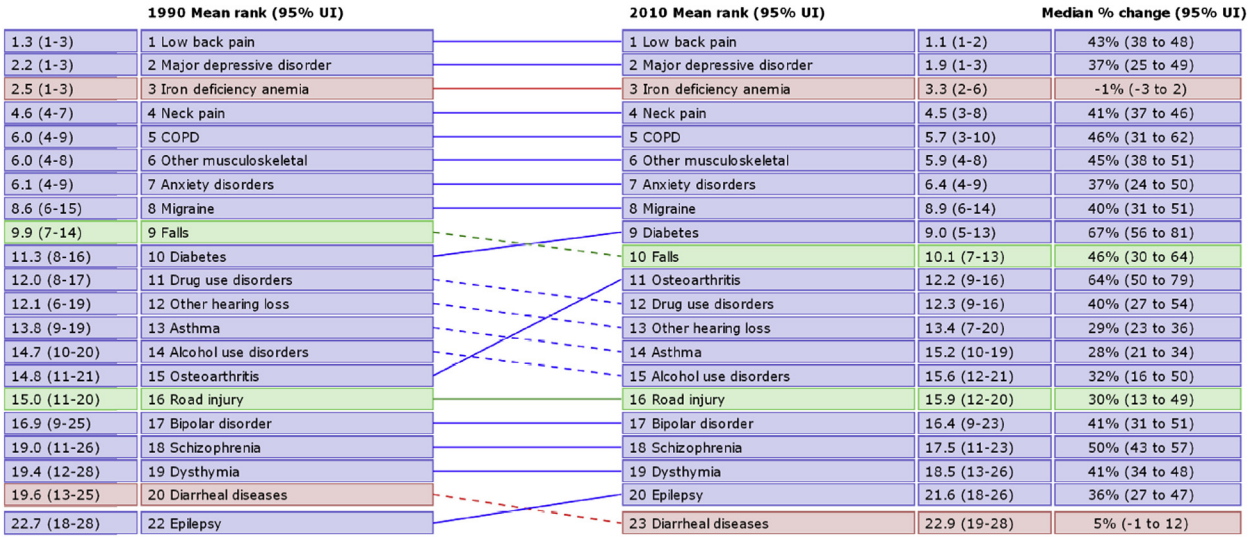

Figure 1 Global years lost to disability for the 20 most common causes in 1990 and 2010. COPD, chronic obstructive pulmonary disease. Source: 2013, Institute for Health Metrics and Evaluation. http://www.healthmetricsandevaluation.org/gbd/visualizations/gbd-arrow-diagram (accessed 19.06.13.).

still limited in various respects, given in particular that not all mental disorders have been investigated. Other areas with remaining deficits are (1) mental disorders in children, adolescents, and the elderly; (2) the dimensional and subthreshold characterization of mental disorders; (3) a more comprehensive identification of risk factors; (4) the range of associated disabilities; and (5) the use of health services with emphasis on national and cultural variations.

Coverage of Disorders

Epidemiological research has expanded in terms of the range of clinically significant psychopathological conditions examined, with increasingly more data becoming available for somatoform disorders, sleep disorders, and some forms of childhood disorders. Noteworthy deficits, however, still exist, particularly with regard to personality disorders and psychotic disorders. Each of these conditions merits considerable epidemiological attention from a public health, and even more from a pathogenetic perspective, because they have different onset and course characteristics, and are frequently comorbid. For example, somatoform disorders (i.e., pain disorders, hypochondriasis, somatization), which were shown to be highly prevalent conditions that start early in life and constitute a major burden on the health care system, have rarely been included in past community surveys. Similarly, sleep disorders, which rank high as a principal reason for primary care consultations, have only been studied in some community surveys. The unclear nosological status of these conditions may be responsible for this neglect. Obstacles to research on personality disorders and psychotic disorders are of a methodological nature because there are not sufficiently reliable and time-efficient assessment tools available.

For childhood conditions, the challenge lies in consensus regarding the appropriate choice of age-/developmental-stage-specific diagnostic assessments. Researchers must also concur on the degree to which multiple sources of information (i.e., parents and teachers) can be combined into one coherent strategy that mirrors the continuity from childhood to adolescent and adult mental disorders. Despite growing collaboration, there is still a remarkable division among investigators regarding epidemiological designs and methods used in childhood, adolescent, and adult mental disorder research. Inherent in this division is the issue of the developmental continuity of psychopathological features, which also touches partly on the ongoing controversy of dimensional versus categorical measures in epidemiological studies of mental disorders. Intensified research in this area is needed, especially because of evidence that most adult mental disorders begin in childhood or adolescence.

Subthreshold Conditions and Dimensional Approaches

With few exceptions, descriptive epidemiological evidence is based on a range of threshold diagnoses, assessed with diagnostic instruments, such as the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI) or versions thereof (Kessler and Üstün, 2004; WHO, 1990). The reliance on categorical threshold diagnoses carries substantial risks of artifactual explanations (such as in comorbidity analyses) and fails to acknowledge the dimensional nature of most expressions of psychopathology. The *Diagnostic and Statistical Manual* (DSM)-IV (APA, 1994), as the most frequently used classification system in research into mental disorders in the past 20 years, made only a few attempts to derive discrete categories that are mutually exclusive and lead to a single classification of an individual. In fact, the system was intended to stimulate further development of research on the thresholds for and boundaries among disorders (APA, 1994). DSM-5 (APA, 2013) attempts to better resemble the rest of medicine by including the use of dimensional approaches (Kupfer et al., 2013) and it remains to be shown by future research whether this approach is a more optimal source for modifying diagnostic systems and 'to carve nature at its joints' (Regier et al., 2009).

Assessment of Disability and Need of Treatment

Over recent years, increasing recognition has emerged that diagnoses of mental disorders themselves cannot appropriately answer questions about the need for care, service utilization, and treatment match. Such domains are thought to be determined largely by the individual's functioning status and disability. Epidemiological studies from the beginning of the twentieth century have begun to provide more coherent and comprehensive information about service-related issues. However, reliable characterizations of diagnosis-specific degrees of disability with regard to various social roles is still scarce, as is sufficiently detailed data about help-seeking behavior, service utilization, and services needs (Regier et al., 1998). The WHO has conducted collaborative, systematic, conceptual, and psychometric developmental work to design a generic and practical instrument to assess health and disability levels across all diseases, including mental disorders (the World Health Organization Disability Assessment Schedule 2.0) (Üstün et al., 2010). This instrument is based on the International Classification of Functioning, Disability and Health and can be used in the general population or in clinical settings to provide a better basis for need assessment in the area of treatment. In the context of the Global Burden of Disease study developed estimates such as YLDs (prevalence of the condition multiplied by the disability weight for that condition) and disability-adjusted life years (DALYs; a summary measure of YLDs and years of life lost (YLLs) due to premature mortality) (Murray et al., 2012; Vos et al., 2012) clearly indicate great challenges on the health care systems due to the rising burden from mental and behavioral disorders and the urgent need to develop effective and affordable strategies to deal with this rising burden.

Assessment Instruments and Diagnostic Classification

Population studies and methods-related epidemiological work have been instrumental in the improvement of diagnostic classification systems for mental disorders. Reliable symptom and diagnostic assessment instruments of mental disorders have been created for use in epidemiology and clinical research. This work has not only significantly influenced the content and structure of clinical instruments, such as the Structured Clinical Interview for DSM-IV (First et al., 1997) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) (Brugha et al.,

1999b), as well as nonclinical tools such as CIDI (WHO, 1990; Kessler and Üstün, 2004), but also played an important role in the revision processes of diagnostic classification systems.

Yet these conceptual models of mental disorders are not, and have never been, a paragon of elegance, and they have also not resulted in sufficiently neat and crisp classification systems that match basic research findings and clinical management and decision making. The introduction of these operationalized and descriptive manuals has resulted in greater diagnostic reliability and consistency in the use of diagnostic terms around the world. In particular, they have been a key prerequisite for epidemiological progress. However, major problems (i.e., thresholds, overlap, and comorbidity), which remain a source of significant dissatisfaction and controversy, will require extensive future work.

At the center of this agenda is the need for convincing clinical and nosological validation in terms of prognostic value and stability, family and genetic findings, and laboratory findings for almost all mental disorders, allowing a sharper genotypic and phenotypic classification. Current diagnostic classification manuals (DSM-5 and International Classification of Disease-10) deliberately do not contain mutually exclusive diagnostic categories in order to simulate research inquiries into diagnostic boundaries and thresholds – a valuable target for epidemiological research. Consensus is lacking on how to tailor appropriate psychopathological assessment instruments that are able to address such threshold issues appropriately. Further, despite the substantial scientific exploration and examination that went into instruments like the CIDI (WHO, 1990) and versions thereof (Kessler and Üstün, 2004) or the SCAN (Brugha et al., 1999b), some basic problems of reliability and validity inherent in the assessment of some mental disorders are yet unresolved.

At the center of discussion is no longer the traditional question of whether to go for categorical or dimensional, but rather to what degree and for which psychological conditions 'clinical judgment and probing' should be regarded as a mandatory core element. Empirical evidence needs to be gathered to determine which diagnostic domains and clinical instruments are superior to fully standardized instruments, such as the CIDI, which try to identify explicitly the latent variables behind the vagueness of clinical judgment. Progress in the resolution of this issue will also ultimately offer more appropriate strategies in resolving the 'gold-standard' question of the optimal strategy for validating epidemiological instruments (Brugha et al., 1999a; Wittchen et al., 1999).

Longitudinal Studies and Causal Analytic Epidemiology

Several costly large-scale prospective longitudinal studies have become available in various countries (e.g., Early Developmental Stages of Psychopathology Study, Great Smoky Mountain Study, Dunedin Multidisciplinary Health and Development Study) that have significantly contributed to improved knowledge about the patterns of incidence, the natural course, longitudinal stability of symptoms, and comorbidity associations, as well as vulnerability and risk factors for the onset and persistence of mental disorders particularly in children, adolescents, and young adults (overview in Beesdo et al., 2009). However, the 'causal risk factors' status (Kraemer et al., 1997) has not yet been established for most putative risk factors of mental disorders and there is limited insight into specific mechanisms and potential moderators and mediators (Kraemer et al., 2001).

There is strong evidence from studies with a longitudinal component that the majority of people affected with a mental disorder have had the first onset in childhood or adolescence, rarely remit spontaneously, and have a substantially increased risk for a number of temporarily secondary psychopathological complications (Beesdo et al., 2010; Copeland et al., 2009; Gregory et al., 2007). These developmental pathways – reflected in cross-sectional studies as co- or multimorbidity – are associated with an increasing degree of psychosocial dysfunctions and disability and seem to indicate the existence of critical trajectories for certain types of disorders that are potentially promising targets for interventions. Testing the effectiveness of specific interventions and the long-term outcomes on a broad, population-based level might not only be crucial from a public health perspective but also have significant value for improved causal models.

Prospective longitudinal studies have provided a substantial body of evidence with regard to partly diagnosis-specific and partly general risk factors for many disorders (e.g., Beesdo et al., 2010; Moffitt et al., 2007; Wittchen et al., 2000). Well-established risk factors include a family history of mental disorders, the effect of threatening or stressful disruptions in the individuals' immediate social environment, childhood loss, abuse, and trauma. There is also further evidence for more complex interactions: family genetic factors may enhance the effects of loss events, patterns of symptom progression from preceding early anxiety disorders relate to the risk of developing secondary depression, and certain childhood temperamental traits increase the risk of subsequent mental disorders. What remains, however, is the large challenge of completing the complex vulnerability–stress interaction, and understanding how vulnerability and risk factors from formerly competing paradigms interact with each other in specific disorders or groups of mental disorders.

Linking Epidemiology Closer to Clinical and Basic Neurobiological Research

The limitations of convenience samples from clinical and other settings for etiological and pathogenic research have been recognized. Limitations include the risks of artifactual findings, the over- and underestimation of effects, confoundation by comorbidity, and the impossibility of establishing risk factors for first onset of a disorder. Prospective longitudinal studies and causal analytic designs in representative population samples are of key importance to overcome these limitations. A requirement for such designs is a comprehensive evaluation of the epidemiological triad of host (i.e., genetic variables or temperamental predispositions), agent (life stage transitions or life stress), and environment (social processes or environmental agents). A considerable challenge for such studies is the identification of how and when certain environmental factors potentiate or protect against genetic and biological vulnerability factors. Neuroscience, and in particular, genetic research,

Table 2 Relevance of epidemiology of contemporary psychiatry

Examination of the generalizability of clinical samples
Identification of potential etiologic factors and correlates of psychiatric disorders
● Sex differences in psychiatric disorders
● Age-specific patterns of onset and offset
● Cohort effects
● Environmental agents (e.g., viral exposure, toxins, diet, and agents)
● Geographic patterns (e.g., cultural pattern of expression, risk factors, migration)
Genetic and family genetic studies in community samples
Estimation of attributable risk of neurobiological and genetic factors
Identifying core psychopathological processes over the life span
Identifying common and unique vulnerabilities and risk factors for disorders
Establishing of norms for biological and genetic markers
Testing the generalizability of biological markers in community samples

Source: Adapted from Merikangas, K.R., 1999. Editorial: the next decade of psychiatric epidemiology. International Journal of Methods in Psychiatric Research 8 (1), 1–5.

is likely to be essential for future understanding of the etiology and pathogenesis of mental disorders, and ultimately a better genotypic classification. The complexity of longitudinal designs, and the time and costs involved, are substantial; however, emerging findings begin to shape a better understanding of the relevant developmental pathways of specific mental disorders and their interrelationships among each other (comorbidity) (see **Table 2**).

Family and other genetic studies have demonstrated convincingly that an individual's genetic makeup is an important factor in determining vulnerability for almost all mental disorders (Merikangas, 1999; Schreier et al., 2008). Methods of genetic epidemiology for studying risk factors and etiology of familial diseases appear to be the most promising ways to unravel the complex mechanisms through which genes exert their influence. The integration of population genetics and epidemiology is critical for determining the attributable risk of particular deoxyribonucleic acid markers for disease as well as how environmental conditions increase or reduce expressions of genetic vulnerability. Such studies might be directed at states or traits conferring susceptibility. In the first decade of the twenty-first century, initial findings from such types of genetic epidemiological studies emerged identifying some of those genetic factors that place individuals at increased risks for disorders such as substance dependence, anxiety, and depression (Caspi et al., 2003; Polanczyk et al., 2009).

Mental Disorders Are the Core Health Challenge of the Twenty-First Century

The increasingly broad and methodologically improved epidemiological database that has become available over the past decade has allowed not only for considerably improved estimates but also for a more comprehensive picture with regard to the diagnostic spectrum and age groups covered. The resulting prevalence findings and estimates of the burden of disease underline impressively that mental disorders must be considered the major health challenge of the twenty-first century, at least in developed countries (Wittchen et al., 2011). Since the 1980s, quite comprehensive, interdisciplinary mental health systems of providers have emerged in most industrialized countries. However, given the size of the problem of mental disorders it is evident that no system could afford comprehensive treatments for all affected persons. Epidemiological data are a key prerequisite for identifying deficits and problems in health care systems, and offering guidance in service planning and resource allocation. At a time of increasing numbers of effective pharmacological and psychological treatments, competing provider models for mental disorders, and tighter health care budgets, epidemiology can be expected to gain further importance, especially in highlighting the efficacy of interventions (see **Table 3**).

Core elements of need assessment are a reliable definition of a disorder, clearly defined associated disabilities, and existing effective interventions (taking into account both their limitations and modes of delivery). Nevertheless, despite the existence of many effective psychological and drug treatments shown to prevent disability, relapse, chronicity, and suffering for many mental disorders, few epidemiological studies are available that answer with sufficient detail core questions such as how many anxiety or depressive disorders were treated by psychiatrists, psychotherapists, or other types of providers; how many were treated by medication or some form of psychological treatment, or even treated at in- or outpatient institutions; how many in need are treated appropriately; and how many remain untreated. Instead, mostly crude data regarding rates of service utilization and unmet service needs are available that emphasize predominantly the role of primary care physicians. Even though the quite limited role and poor recognition abilities of primary care physicians for mental disorders has been noted repeatedly throughout the world (Üstün and Sartorius, 1995; Wittchen et al., 2002), it seems that traditional psychiatric epidemiology still seems to favor this model strongly. More comprehensive service utilization and need assessment strategies for specialist treatment and other interventions are neglected. Another unfortunate deficit in this field is its strong emphasis on so-called 'major mental disorders,' such as psychotic disorders, as opposed to 'minor and neurotic disorders.' Although this oversimplified dichotomy has clearly outlived its usefulness and scientific justification since the 1980s, many health care utilization and needs surveys still overemphasize major morbidity, neglecting what are actually the most prevalent and persisting 'minor morbidity conditions,' which cause by far the greatest financial burden of all disorders (Murray et al., 2012; Vos et al., 2012). Current

Table 3 Types of questions related to comprehensive need assessment

Stepwise questions	*Epidemiological activity*
What is a case?	Prevalence, incidence
What do patients need?	Subjective/objective
Do patients get what they need?	Inappropriate/insufficient use and service delivery
How can structure and process of services/treatment be improved?	Evaluation of changes, stepwise planning of improvements

attempts that merely link diagnosis with measures of disability in order to improve need assessment (see above) might not solve this critical problem. Rather, they might result once again in an inappropriately strong emphasis on the most severely ill, neglecting those in earlier stages of their illness process who might profit most from modern treatment methods.

The challenge for epidemiology on a cross-national and international basis lies in a systematic comparison of such competing models. Development of appropriate assessment instruments for use in epidemiological studies to identify advantages and disadvantages of each of these perspectives, in terms of legal, cost, comprehensiveness, and effectiveness issues, has been identified as a necessary first step (Jagger et al., 1998). Current perspectives on this issue seem to overemphasize two search strategies: the development of reliable and valid measures of disability (Regier et al., 1998; Üstün et al., 2010), and the search for other 'marker' variables of those in greatest need or the 'most severe.' This perspective might fall too short. In search of improved approaches for comprehensive need assessment and evaluation, future epidemiological research should additionally emphasize: (1) a more comprehensive assessment of help-seeking behaviors that covers the full spectrum of professional providers in the respective country or region; (2) a wider coverage of types of interventions received, contingent on the availability of treatments in that country for that diagnosis; and (3) a detailed inquiry into perceived barriers to recognition and treatment.

Conclusion

In light of the ongoing rapid developments in neuroscience, clinical psychology, and psychiatry, as well as public health research, epidemiology can be expected to further play an increasingly important role in basic, clinical, and public health research of mental disorders. The key challenge will be to understand how multiple risk and vulnerability factors interact over the life span in producing a single or multiple mental disorders over brief or longer periods of time.

See also: Anxiety and Anxiety Disorders; Mental Illness, Etiology of; Mental Illness, Genetics of; Normality and Mental Health; Public Psychiatry and Public Health Mental Systems.

Bibliography

Ansseau, M., Dierick, M., Buntinkx, F., Cnockaert, P., deSmedt, J., vandenHaute, M., vanderMijnsbrugge, D., 2004. High prevalence of mental disorders in primary care. Journal of Affective Disorders 78, 49–55.

APA, 1980. Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Press, Washington, DC.

APA, 1994. Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Press, Washington, DC.

APA, 2013. Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Association, Arlington, VA.

Beesdo, K., Knappe, S., Pine, D.S., 2009. Anxiety and anxiety disorders in children and adolescents: developmental issues and implications for DSM-V. Psychiatric Clinics of North America 32 (3), 483–524.

Beesdo, K., Pine, D.S., Lieb, R., Wittchen, H.U., 2010. Incidence and risk patterns of anxiety and depressive disorders and categorization of generalized anxiety disorder. Archives of General Psychiatry 67 (1), 47–57.

Brugha, T.S., Bebbington, P.E., Jenkins, R., 1999a. A difference that matters: comparisons of structured and semi-structured psychiatric diagnostic interviews in the general population. Psychological Medicine 29 (5), 1013–1020.

Brugha, T.S., Nienhuis, F.J., Bagchi, D., Smith, J., Meltzer, H., 1999b. The survey form of SCAN: the feasibility of using experienced lay survey interviewers to administer a semi-structured systematic clinical assessment of psychotic and nonpsychotic disorders. Psychological Medicine 29, 703–711.

Caspi, A., Sugden, K., Moffitt, T.E., Taylor, A., Craig, I.W., Harrington, H., McClay, J., Mill, J., Martin, J., Braithwaite, A., Poulton, R., 2003. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science 301 (5631), 386–389.

Copeland, W., Shanahan, L., Costello, E.J., Angold, A., 2011. Cumulative prevalence of psychiatric disorders by young adulthood: a prospective cohort analysis from the great smoky mountains study. Journal of the American Academy of Child and Adolescent Psychiatry 50 (3), 252–261.

Copeland, W.E., Shanahan, L., Costello, J., Angold, A., 2009. Childhood and adolescent psychiatric disorders as predictors of young adult disorders. Archives of General Psychiatry 66 (7), 764–772.

First, M.B., Spitzer, R.L., Gibbon, M., Williams, J.B.W., 1997. User's Guide for the Structured Clinical Interview for DSM-iv Axis I Disorders (SCID-i). American Psychiatric Press, New York.

Gregory, A.M., Caspi, A., Moffitt, T.E., Koenen, K., Eley, T.C., Poulton, R., 2007. Juvenile mental health histories of adults with anxiety disorders. American Journal of Psychiatry 164 (2), 301–308.

Gustavsson, A., Svensson, M., Jacobi, F., Allgulander, C., Alonso, et al., 2011. Cost of disorders of the brain in Europe 2010. European Neuropsychopharmacology 21 (10), 718–779.

Jagger, C., Ritchie, K., Bronnum-Hansen, H., Deeg, D., Gispert, R., et al., 1998. Mental health expectancy – the European perspective: a synopsis of results presented at the conference of the European network for the calculation of health expectancies (Euro-REVES). Acta Psychiatrica Scandinavica 98 (2), 85–91.

Kessler, R.C., Aguilar-Gaxiola, S., Alonso, J., Chatterji, S., Lee, S., et al., 2011. The burden of mental disorders worldwide: results from the World Mental Health surveys. In: Galea, N.C.S. (Ed.), Population Mental Health: Evidence, Policy, and Public Health Practice. Routledge, Abingdon, pp. 9–37.

Kessler, R.C., Berglund, P., Demler, O., Jin, R., Merikangas, K.R., Walters, E.E., 2005a. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. Archives of General Psychiatry 62 (7), 593–602.

Kessler, R.C., Chiu, W.T., Demler, O., Walters, E.E., 2005b. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the national comorbidity survey replication. Archives of General Psychiatry 62, 617–627.

Kessler, R.C., Petukhova, M., Sampson, N.A., Zaslavsky, A.M., Wittchen, H.U., 2012. Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. International Journal of Methods in Psychiatric Research 21 (3), 169–184.

Kessler, R.C., Üstün, T.B., 2004. The world mental health (WMH) survey initiative version of the world health organization (WHO) composite international diagnostic interview (CIDI). International Journal of Methods in Psychiatric Research 13 (2), 93–11.

Kraemer, H.C., Kazdin, A.E., Offord, D.R., Kessler, R.C., Jensen, P.S., Kupfer, D.J., 1997. Coming to terms with the terms of risk. Archives of General Psychiatry 54 (4), 337–343.

Kraemer, H.C., Stice, E., Kazdin, A., Offord, D., Kupfer, D., 2001. How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. American Journal of Psychiatry 158, 848–856.

Kupfer, D.J., Kuhl, E.A., Regier, D.A., 2013. DSM-5-The future arrived. Jama-Journal of the American Medical Association 309 (16), 1691–1692.

Merikangas, K.R., 1999. Editorial: the next decade of psychiatric epidemiology. International Journal of Methods in Psychiatric Research 8 (1), 1–5.

Moffitt, T.E., Caspi, A., Harrington, H., Milne, B.J., Melchior, M., Goldberg, D., Poulton, R., 2007. Generalized anxiety disorder and depression: childhood risk factors in a birth cohort followed to age 32. Psychological Medicine 37, 441–452.

Moffitt, T.E., Caspi, A., Taylor, A., Kokaua, J., Milne, B.J., Polanczyk, G., Poulton, R., 2010. How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective versus retrospective ascertainment. Psychological Medicine 40 (6), 899–909.

Murray, C.J.L., Lopez, A.D. (Eds.), 1996. The Global Burden of Disease. A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries and Risk Factors in 1990 and Projected to 2020, I. Harvard School of Public Health on behalf of the World Health Organization and the World Bank, Cambridge, MA.

Murray, C.J.L., Vos, T., Lozano, R., Naghavi, M., Flaxman, A.D., et al., 2012. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the global burden of disease study 2010. Lancet 380 (9859), 2197–2223.

Polanczyk, G., Caspi, A., Williams, B., Price, T.S., Danese, A., et al., 2009. Protective effect of CRHR1 gene variants on the development of adult depression following childhood maltreatment replication and extension. Archives of General Psychiatry 66 (9), 978–985.

Regier, D.A., Kaelber, C.T., Rae, D.S., Farmer, M.E., Knauper, B., et al., 1998. Limitations of diagnostic criteria and assessment instruments for mental disorders - implications for research and policy. Archives of General Psychiatry 55 (2), 109–115.

Regier, D.A., Narrow, W.E., Kuhl, E.A., Kupfer, D.J., 2009. The conceptual development of DSM-V. American Journal of Psychiatry 166 (6), 645–650.

Robins, L., 1992. The future of psychiatric epidemiology. International Journal of Methods in Psychiatric Research 2, 1–3.

Robins, L.N., Regier, D.A. (Eds.), 1991. Psychiatric Disorders in America: The Epidemiologic Catchment Area Study. The Free Press, New York.

Schreier, A., Wittchen, H.U., Höfler, M., Lieb, R., 2008. Anxiety disorders in mothers and their children: prospective longitudinal community study. The British Journal of Psychiatry 129, 308–309.

Üstün, T.B., Kostanjsek, N., Chatterji, S., Rehm, J., 2010. Measuring Health and Disability: Manual for WHO Disability Assessment Schedule (WHODAS 2.0). World Health Organization, Geneva.

Üstün, T.B., Sartorius, N., 1995. Mental Illness in General Health Care: An International Study. John Wiley & Son, Chichester, NY.

Vos, T., Flaxman, A.D., Naghavi, M., Lozano, R., Michaud, C., et al., 2012. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the global burden of disease study 2010. Lancet 380 (9859), 2163–2196.

Wang, P.S., Aguilar-Gaxiola, S., Alonso, J., Lee, S., Schoenbaum, M., et al., 2011. Assessing mental disorders and service use across countries: the WHO World Mental Health Survey Initiative. In: Regier, D.A., Narrow, W.E., Kuhl, E.A., Kupfer, D.J. (Eds.), The Conceptual Evolution of DSM-5. American Psychiatric Publishing, Inc., Arlington, VA, pp. 231–266.

WHO, 1990. Composite International Diagnostic Interview (CIDI). World Health Organization, Geneva, Switzerland.

Wittchen, H.-U., 1996. Critical issues in the evaluation of comorbidity of psychiatric disorders. British Journal of Psychiatry 168 (Suppl. 30), 9–16.

Wittchen, H.-U., Beesdo, K., Gloster, A.T., 2009. A new meta-structure of mental disorders: a helpful step into the future or a harmful step back to the past? A commentary on the Andrews series papers. Psychological Medicine 39 (12), 2083–2089.

Wittchen, H.-U., Kessler, R.C., Beesdo, K., Krause, P., Höfler, M., Hoyer, J., 2002. Generalized anxiety and depression in primary care: prevalence, recognition and management. Journal of Clinical Psychiatry 63 (8), 24–34.

Wittchen, H.-U., Kessler, R.C., Pfister, H., Lieb, R., 2000. Why do people with anxiety disorders become depressed? A prospective-longitudinal community study. Acta Psychiatrica Scandinavica 102 (Suppl. 406), 14–23.

Wittchen, H.-U., Üstün, T.B., Kessler, R.C., 1999. Diagnosing mental disorders in the community. A difference that matters? Psychological Medicine 29 (5), 1021–1027.

Wittchen, H.U., Jacobi, F., Rehm, J., Gustavsson, C., Svensson, M., et al., 2011. The size and burden of mental disorders and other disorders of the brain in Europe 2010. European Neuropsychopharmacology 21, 655–679.

Relevant Websites

http://www.healthmetricsandevaluation.org/gbd – Institute for Health Metrics and Evaluation – Global Burden of Disease.

http://www.who.int/topics/global_burden_of_disease/en/ – WHO Global Burden of Disease.

http://www.hcp.med.harvard.edu/wmh/ – The World Mental Health Survey Initiative.

Mental Health

Traolach Brugha, College of Medicine, Biological Sciences and Psychology, University of Leicester, Leicester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term mental health sits alongside the larger topic of mental disorder with which it overlaps while retaining a wider concept of general, including physical health, well-being, positive functioning, and the development and maintenance of interpersonal relationship functioning. The promotion of mental health depends in part on the prevention of mental disorder, a vital but largely neglected topic for which there are now promising signs of progress.

Introduction and Concepts

Mental health has been defined in several related ways. It describes a level of psychological well-being, or an absence of a mental disorder; the World Health Organization defines mental health as "a state of well-being in which the individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community." Most recently, the field of 'global mental health' has emerged, which prioritizes equity in mental health for all people in the world. From the perspective of 'positive psychology' or 'holism,' mental health may include an individual's ability to enjoy life, and create a balance between life activities and efforts to achieve psychological resilience. Mental health can also be defined as an expression of emotions, and as signifying a successful adaptation to a range of demands. Mental well-being relates to a person's psychological functioning, life-satisfaction and ability to develop and maintain mutually benefiting relationships. Psychological well-being includes the ability to maintain a sense of autonomy, self-acceptance, personal growth, purpose in life and self esteem. Staying mentally healthy is more than treating or preventing mental illness (Tennant et al., 2007).

When originally introduced, the term 'mental health' was intended to be interpreted literally. Its originators were primarily motivated by the wish to promote mental health and tackle any obstacle to this aim (Bertolote, 2008). This article is devoted to an overview of understanding and evidence about mental health and how it can be achieved or sustained within human populations. We focus on the historical, conceptual, policy, population, and intervention evidence underpinning mental health. Historically, the term mental health owes its origins to the mental hygiene movement developed by Clifford Biers in nineteenth-century America (Bertolote, 2008) based on his personal experiences of mental hospital admissions. Its initial purpose was to promote early diagnosis and treatment and humane care for psychiatric patients but notably it expanded to the wider community, policy and government embracing "not a single patient, but a whole community; and it considered each member of that community as an individual whose mental and emotional status was determined by definite causative factors and whose compelling need was for prevention rather than cure." However, the eventual, gradual and welcome embrace of psychiatry and of mental disorder officially at national and international levels (for example, the World Health Organization) soon adopted the term 'mental health' with little more than a side reference to well-being, prevention and promotion.

Conceptually, it may also be unhelpful to keep apart the terms mental health and mental disorder. The trouble is that people perceive, think and fear, in categories and prefer the language of 'mental breakdown' to that of 'increasing anxiety or depression.' Evidence suggests that the two represent different ends of a smooth distribution (Melzer et al., 2002). At its very simplest, a depression score that is low or zero can be described as health, but at or above an agreed threshold as disorder. Guidelines on treating depression now advocate a stepped care approach in which investigation and intervention are tailored to severity, chronicity (and risk). This approach can be seen as replacing earlier binary guidance hinging on whether or not a patient meets an all or nothing set of diagnostic criteria, with all resources being devoted to the latter only.

Achieving (or sustaining) mental health thus means avoiding or preventing mental disorder and to do the first arguably could result in the second although this has not yet been shown. This requires an understanding of normal development and of how psychopathology in its different forms unfolds, which we can describe and measure reliably in psychiatry although it is less understood. Any definition must also consider the terms mental health, mental well-being, and mental illness or disorder. But research suggests that mental well-being is not simply part of the mental health disorder continuum. Thus, although similar factors are associated with mental disorder and mental well-being, models that incorporate both together do not suggest that they are equivalent (Weich et al., 2011).

Mental Well-Being

The term mental health is gradually being over shadowed by the term mental well-being (perhaps because well-being appeals to policymakers and politicians and because they have allowed the term mental health to mean absence of mental disorder). Mental well-being (see also Subjective Wellbeing, Psychology of) is not just the opposite of mental disorder; evidence suggests that both states can coexist in one person or population subgroup. The same factors have also been shown to be associated with both states (Weich et al., 2011). Mental well-being

may also be said to be a predictor of and predicted by mental health. Mental well-being can be defined as the positive end of a spectrum that includes the concepts 'mental health' and 'positive mental health' (Weich et al., 2011). There are two broad (and complementary) traditions in conceptualizing well-being: the 'hedonic' and 'eudaimonic' approaches. The hedonic approach emphasizes happiness (pleasant affect, life satisfaction) whereas the eudaimonic approach (with origins in Platonic philosophy) emphasizes optimal psychological and social functioning (or 'flourishing') (Weich et al., 2011).

Measurement of well-being has often relied on single questions assessing aspects of mental well-being, for example, questions on calmness and energy or on distinct factors derived from measures of mental symptoms, as in the General Health Questionnaire and the mental health component of the Short Form (SF-36 and SF-12) (Ware, 1993), which is also used to measure the related concept of quality of life. Thus components and elements of well-being measures can include mental health symptoms, principally of anxiety and depression, but possibly including somatic symptoms such as sleep problems, pain, fatigue, low drive, sexual functioning disturbances. There may also be positive well-being items like feeling healthy, energetic etc. Well-being measures tend to extend further to social functioning items, possibly activity items (work, leisure, community activities). But purpose designed measures are now being developed, which should foster better research. A significant development in the measurement of mental health and well-being outcomes is the Warwick-Edinburgh Mental Well-being Scale (Tennant et al., 2007).

Mental well-being is beginning to be regarded by national governments for its economic importance because higher national income has not brought the better *quality of life* that many expected (Layard, 2010). Gross Domestic Product and National Accounts measure economic well-being according to an internationally agreed system, but they are increasingly seen as measuring only part of societal well-being (http://www.statistics.gov.uk/hub/people-places/communities/societal-wellbeing). Governments in Europe are, therefore, beginning to monitor trends in well-being. Specific measures have been developed that are used in repeat general population surveys that also keep track of possible determinants of well-being, economic prosperity, education and health outcomes, progress in housing and welfare policies.

Physical and Mental Health

In this section on health, we need to ask is the distinction between mental and physical health valid? There are many indications that such a distinction often should be questioned: the growth of research into the biological including genetic underpinnings of mental disorder as if such disorders were physical; the idea that, for example, taking time off from work for a spell of depression should be viewed as no different than taking time off for back pain, a viral respiratory infection etc.; and the often quoted contemporary health policy statement 'no health without mental health' (Prince et al., 2007). It has long been recognized in primary care that mental health problems often present initially as physical complaints. Epidemiologists are increasingly demonstrating that throughout the world physical and mental health are interrelated in research also showing that mental disorders are more likely to be associated with severe disability than are chronic physical conditions (Scott et al., 2009). Interestingly in the last decade depression has overtaken back pain as the number one medical reason for claiming work absence benefits in England.

Mental Disorder and Disorders from a Neurodevelopmental Perspective

If mental health is defined in part as the absence of mental and behavioral disorders it is important to move toward a seamless approach that takes into account phases of development (of psychopathology) across the full life course (Kingdon et al., 2010). Readers new to the subject will notice the puzzling but distinctive difference between current childhood and adulthood classifications of mental and behavioral disorders. Although the same conditions that appear in child classifications, for example, conduct disorder and autism, are recognized in adulthood, in practice they either acquire a different term, for example, antisocial personality disorder, or are rebranded, for example, asperger syndrome, learning disability. Definitions and defining boundaries also shift making empirical comparison, learning, and understanding more difficult. The problem probably reflects a lack of consensus on what underlies mental and behavioral disorders. A simple biological analogy may assist us, based on the growth of a tree (or any plant; see panel).

Biological cause	*Abnormality in tree or other plant*	*Mental disorder analogy*	*Possible cause*
Genetic	Stunted growth	Huntington's Chorea	Trinucleotide repeats on the HTT gene located on short arm of chromosome 4
		Autism	Mechanism highly heritable, complex and yet to be identified, understood
Environmental challenge, infection	Characteristic abnormal appearance	Delirium	Septicemia
		Depression	Adverse life events, long-term difficulties, child sexual abuse
Unknown	New disease with characteristic appearance	Dementia	Well-recognized brain pathology and neuroimaging findings but cause unknown

If genetically abnormal, growth will be stunted although that may not become evident until maturity has been reached. The human analogy, such as intellectual disability or a specific neurodevelopmental disorders such as autism can continue

into old age. Disease may impinge at any time in the absence of previous abnormality (for example, conditions that usually present first from puberty onward such as functional psychosis, depression etc.). Damage may also be caused by environmental challenge such as extreme weather (in the human example, adverse life events or long-term difficulties such as debt, prolonged physical and mental abuse). Growth, development, and aging may also be affected by climate, nourishment (analogous to unemployment, isolation caused by racism, stigma, prolonged conflict). In common across these examples is the notion of healthy and interrupted development. Any example, therefore, demands a consideration of prior level of development unless the condition is a clearly recognized example of a late onset disorder such as dementia where prior development has little effect on later outcome. And yet it is unusual for professionals looking at an adult needing an assessment or treatment to devote time and effort to obtaining information about prior growth and development and prior environmental exposure (parenting and other childhood social and learning experiences). The same argument goes for research to be meaningful and to cast a light on complex causal mechanisms that could help maintain mental health and lead to prevention of disorder.

Mental Health and Resilience, Personality, Resources, Social Support Networks, and Relationships, Social Structures and Social Capital

As well as well-being measures extending to social functioning, relationship quality, work, leisure, community activities, there is also ample evidence interconnecting these separate factors with mental health and with individual characteristics, which must form part of any account of mental health (Brugha, 2003). The complexity of such connections can be illustrated by a number of research findings. Anger or irritability, a sign of depression, is often directed at others and must partly explain the isolation that results from depression. An extensive literature connects adverse life events with many different types of mental disorder and probably interacts with temperament and personality (Goodyer, 2002). Associations involving the structure and composition of social relationships, including social network research, has been consistently shown to be important in mental health over many decades of research.

Not all findings are in the expected direction, however. Structural characteristics of communities including the relatively newer concept of social capital, represented by local characteristics such as the way that people contribute to a community, local crime rates etc., have been found to add little if anything to associations with mental health and disorder that include individual, household and family factors (Weich et al., 2003). There is also a disappointing lack of experimental research testing hypotheses linking such social factors to mental health or disorder. There is also the curious observation that interpersonal communication and relationship quality seems to underlie some of the effectiveness of psychological therapies. A recent systematic review of psychological therapy treatment trials concluded that most of the effect of therapy for adult depression is realized by nonspecific factors, with the results suggesting that the contribution of specific effects (cognitive behavioral, psychodynamic etc.) is limited at best (Cuijpers et al., 2012). It is hardly surprising, therefore, if interpersonal relationships (in this example between client and therapist) play a fundamental part in sustaining mental health as well as in defining it.

The Epidemiology of Mental Health and Mental Disorder

While 'rates' of mental health have not been specifically addressed in community surveys and studies, there exists an extensive literature on the frequency of mental disorders from which much can be approximately inferred (see also Health Surveys; Epidemiology, Concepts and Methods of). Earlier, we conceptualized mental health (and disorder) as being a dimensional construct within which it is useful to identify typologies or syndromes and to set thresholds, which expert consensus agrees to recommend as the dividing line between normality and clinically significant psychopathology. An unusual but compelling suggestion has been to describe common mental disorder in the population in terms of mean scores with normal or skewed distributions (Melzer et al., 2002), but most studies focus mainly on the frequency of types of disorders, which seems to appeal to policymakers, with dimensionality or severity receiving relatively less attention (Demyttenaere et al., 2004). The international setting for synthesizing such categorical disease or morbidity frequency information is the Global Burden of Disease study (GBD) (Murray et al., 2012), which ranks neuropsychiatric disorders such as schizophrenia and depression consistently above most physical causes of disability.

So how common is mental health (and or mental disorder) and are rates declining or increasing? The proportion of the child and adult population free of mental disorder has perhaps been most closely represented in estimates of common mental disorder as a dimension or score (Goldberg and Huxley, 1992; Meltzer et al., 1995). These two closely related approaches suggest that between 75 and 84% of the adult population enjoy mental health (or have few if any symptoms of common mental disorder). However, if one goes on to factor in problems associated with drug and alcohol dependency or misuse and lifelong factors such as intellectual disabilities these estimates fall further.

Good news is not exciting and does not grab media attention. No matter how many times good survey designs are repeated, using consistent methodology, seemingly well informed media discussions persist in presuming that rates of disorder in childhood and adulthood are rising. But as the GBD syntheses and high quality UK, Netherlands and US studies have shown rates are largely stable over recent decades (Murray et al., 2012). We do not know why. It is possible that a mix of positive and negative factors have been operating over time to increase and decrease rates. But as mentioned earlier, increased use of treatments for the commoner depressive disorders has not lead to any reduction in rates of these conditions (Brugha et al., 2004; Kessler et al., 2005). This has lead to calls to focus more on prevention because preventive approaches have had a positive impact on other forms of morbidity, globally (Murray et al., 2012).

Interventions to Promote Mental Health and Well-Being – A Research Method Note

Assessment of well-being and mental health outcomes has been considered above. Nonexperimental, observational research, whether to discover (or confirm) social, psychological, and environmental or genetic profiles associated with health states cannot finally determine causality and the means to improve health outcomes. Only experimental research designs involving controlled exposure to interventions can achieve this step forward convincingly, the gold standard experimental design being the randomized controlled trial (RCT). Contrary to expectation, scientific activity devoted to conducting RCTs and the synthesis of the findings of multiple RCTs (systematic reviews) have not ignored well-being and mental health as outcomes of importance and interest. A search for the term 'wellbeing' on the world leading Cochrane Collaboration Web site yielded 96 'hits' on well-being (http://www.cochrane.org/search/site/wellbeing accessed 18.12.13.) and 243 on 'well-being.' Frequently targeted population subgroups were cancer patients, women in labor, the fetus, older children, adults in stressful environments. Just to pick out a couple of specific examples: a completed systematic review of flexible working conditions and their effects on employee health and well-being (the findings of which 'seem to indicate that flexibility in working patterns, which gives the worker more choice or control is likely to have positive effects on health and well-being'); a protocol under development for a systematic review of organizational interventions for improving well-being and reducing work-related stress in teachers; a completed systematic review of parental training for improving parental psychosocial health.

Mental Health Promotion and Prevention

At the heart of initial thinking on mental health was the idea that mental health is a value and a state to be taken note of, considered and given attention to. This set of values was intended to be applied not just to those who are well (not mentally ill) but also to those with mental disorder including those suffering from its most severe forms. What began as a much needed set of values and principles has not, however, been taken up in the form of scientifically supported evidence development and testing leading to the effective prevention of mental disorders (see also Prevention Research).

Well-being and emotional health is central to the mental health promotion movement (see also Health Education and Health Promotion). Various activities have been suggested as increasing well-being and mental health including, activity and expressive therapies, meditation, biofeedback, counseling, and psychotherapy.

Selected, Indicated and Universal Prevention

If achieving mental health is substantially about prevention of mental disorder, progress with the latter deserves effort and attention. Gains have been made in two areas: providing resources to families and small children, and identifying adults at high risk of developing mental disorder and providing early intervention similar to psychological treatment. There is some evidence that parenting programs improve parental psychosocial health and child outcomes (Sandler et al., 2011; http://summaries.cochrane.org/CD002020/parent-training-for-improving-parental-psychosocial-health). One of the most promising long term, experimentally designed programs was designed to promote resilience of parentally bereaved children and their bereaved surviving parent (Sandler et al., 2010).

Evidence for the possibility of preventing mental disorders in older children and adults has come mainly from clinical researchers expanding the range of techniques that are effective treatments, such as cognitive behavior therapy (CBT), to a wider range of individuals than those who would normally be treated, such as those at high risk of developing, for example, depression, in the near future (Munoz et al., 2012). Thus there are now several studies showing that children in the final years before leaving school who already exhibit symptoms of depression (but are not clinically depressed) do benefit from CBT approaches applied within schools at a group level. Targeting older school children at high risk of becoming cases of depression with school-based psychological prevention programs seems to be effective (Garber et al., 2009).

Prevention and Policy

Quite appropriately, the global mental health movement has prioritized the narrowing of the gap between the burden of mental disorder and the delivery of care and services particularly in less-developed countries of the world (Prince et al., 2007). When mentioned, it is acknowledged that there is little evidence on which to base prevention but that governments should follow the evidence that exists. But how well would this work? One country has made the prevention of depression in adulthood a matter of national policy, the Netherlands. However, although indicated prevention of depression is available for about 80% of the Dutch population at little or no cost, only a small proportion of those with subthreshold depression make use of these services (Cuijpers et al., 2010). Most ambulatory mental health services have a prevention department serving the local or regional community but only 1% of adults with subthreshold depression access these services, which make use of a group-based depression prevention intervention, 'Coping with Depression' (Lewinsohn, 1987) that has been shown to be effective. While, it is important to find ways of persuading more people to consider using these approaches different ways of delivering prevention interventions are needed that overcome possible barriers including stigma, denial, and hopelessness (Cuijpers et al., 2010).

Selected and indicated approaches, for example, targeting high-risk groups face a further limitation: the relatively few who benefit are unlikely to alter significantly population prevalence and thus societal burden. This has lead to the recommendation to position prevention services in primary care or to integrate prevention interventions in community-wide interventions (Cuijpers et al., 2010). RCT-based evidence for prevention, particularly covering the whole population, offers the hope of measurable declines in the prevalence of depression attributable to prevention programs. But universal prevention trials are extremely rare in the

mental health field partly because they require the involvement of substantial numbers of participants and are therefore very costly (Munoz et al., 2012). Particular windows of opportunity do exist when people face challenging life transitions and have normal high contact with experienced health professionals, for example, during pregnancy and after childbirth. At such times there is both increased actual or perceived risk and normal access to nonstigmatizing care. Encouraging evidence that such a phased psychological service reorganization, possibly lowering the risk of depression, could be acceptable and feasible comes from a cluster RCT (Morrell et al., 2009). Women who scored positively on a postnatal test for depression were given guided psychological support based on counseling or cognitive behavioral principles provided by their community health visitor (who had undergone additional psychological training). In a secondary analysis of the same RCT (Brugha et al., 2011), postnatal women at 4–6 weeks following childbirth testing negative on the Edinburgh Depression Scale (Cox and Holden, 1994) were less likely to develop later postnatal depression if attended by a health visitor trained to offer the additional psychological support. Development of symptoms of depression was also experimentally shown to be less likely where the health visitor had also evaluated and discussed with the mother her risk of depression 6–8 weeks after childbirth. The same research group has since pilot demonstrated that services for women during pregnancy provided by community midwives can undertake the necessary additional training and follow the same model of treatment and prevention, but large-scale trials are needed to test this promising approach fully.

Primary Care Mental Health

As the previous section shows, primary care is a key setting for understanding and promoting mental health. The same epidemiological surveys mentioned earlier also show that most forms of mental disorder are mild in severity and adults suffering from them are often in contact with primary health care services, rarely engaging with specialized services. While the accent is on the medical model of disorder detection and treatment, primary care physicians increasingly view mental health as they do physical health as a responsibility attached to their role. A sign that not all is as it should be is consistent evidence that a significant proportion of prescribing of psychotropic medication is being taken up by the mentally well part of the population (Brugha et al., 2004; Kessler et al., 2005), which seemed to occur following a period of rapid expansion in prescribing particularly of SSRI antidepressive medication. It is not known if this is a sign that individuals are being prescribed in order to prevent recurrence of previous depressive disorder, or if it is a sign of unnecessary prescribing of preparations that are increasingly known to be effective only for moderate to severe forms of depression, or represents dependency. It may also reflect demand from service users who should be encouraged and supported instead to address possible causes in their lives and to pursue lifestyle activities that are thought to promote well-being, physical and mental health.

Conclusion

The achievement of mental health can be an outcome in its own right but cannot be completely separated from considering mental disorder and its prevention and early treatment. Mental well-being is a term that is increasingly occupying the space originally intended by the term 'mental health,' which often is used as a nonstigmatizing term for mental disorder and for psychiatry. Far too little research is devoted to prevention of mental disorders thereby, promoting mental health but there are seeds of promise that need to be nourished.

See also: Depression, Pessimism, and Health; Functioning, Disability and Health, International Classification of; Health Surveys; Social Integration, Social Networks, and Health.

Bibliography

Bertolote, J., 2008. The roots of the concept of mental health. World Psychiatry 7 (2), 113–116.

Brugha, T.S., 2003. The effects of life events and social relationships on the course of major depression. Current Psychiatry Reports 5 (6), 431–438.

Brugha, T.S., Bebbington, P.E., Singleton, N., Melzer, D., Jenkins, R., Lewis, G., Farrell, M., Bhugra, D., Lee, A., Meltzer, H., 2004. Trends in service use and treatment for mental disorders in adults throughout great Britain. The British Journal of Psychiatry 185, 378–384.

Brugha, T.S., Morrell, C.J., Slade, P., Walters, S.J., 2011. Universal prevention of depression in women postnatally: cluster randomized trial evidence in primary care. Psychological Medicine 41 (4), 739–748.

Cox, J.L., Holden, J., 1994. Perinatal Psychiatry Use and Misuse of the Edinburgh Postnatal Depression Scale Gaskell. London.

Cuijpers, P., Driessen, E., Hollon, S.D., van, O.P., Barth, J., Andersson, G., 2012. The efficacy of non-directive supportive therapy for adult depression: a meta-analysis. Clinical Psychology Review 32 (4), 280–291.

Cuijpers, P., van, S.A., Warmerdam, L., van Rooy, M.J., 2010. Recruiting participants for interventions to prevent the onset of depressive disorders: possible ways to increase participation rates. BMC Health Services Research 10, 181.

Demyttenaere, K., Bruffaerts, R., Posada-Villa, J., Gasquet, I., Kovess, V., Lepine, J.P., Angermeyer, M.C., Bernert, S., De Girolamo, G., Morosini, P., Polidori, G., Kikkawa, T., Kawakami, N., Ono, Y., Takeshima, T., Uda, H., Karam, E.G., Fayyad, J.A., Karam, A.N., Mneimneh, Z.N., Medina-Mora, M.E., Borges, G., Lara, C., De Graaf, R., Ormel, J., Gureje, O., Shen, Y., Huang, Y., Zhang, M., Alonso, J., Haro, J.M., Vilagut, G., Bromet, E.J., Gluzman, S., Webb, C., Kessler, R.C., Merikangas, K.R., Anthony, J.C., Von Korff, M.R., Wang, P.S., Brugha, T.S., Aguilar-Gaxiola, S., Lee, S., Heeringa, S., Pennell, B.E., Zaslavsky, A.M., Ustun, T.B., Chatterji, S., 2004. Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. Journal of the American Medical Association 291 (21), 2581–2590.

Garber, J., Clarke, G.N., Weersing, V.R., Beardslee, W.R., Brent, D.A., Gladstone, T.R., DeBar, L.L., Lynch, F.L., D'Angelo, E., Hollon, S.D., Shamseddeen, W., Iyengar, S., 2009. Prevention of depression in at-risk adolescents: a randomized controlled trial. Journal of the American Medical Association 301 (21), 2215–2224.

Goldberg, D., Huxley, P., 1992. Common Mental Disorders. A Bio-social Model. Routledge, London.

Goodyer, I.M., 2002. Social adversity and mental functions in adolescents at high risk of psychopathology. Position paper and suggested framework for future research. British Journal of Psychiatry 181, 383–386.

Kessler, R.C., Demler, O., Frank, R.G., Olfson, M., Pincus, H.A., Walters, E.E., Wang, P., Wells, K.B., Zaslavsky, A.M., 2005. Prevalence and treatment of mental disorders, 1990 to 2003. The New England Journal of Medicine 352 (24), 2515–2523.

Kingdon, D., Afghan, S., Arnold, R., Faruqui, R., Friedman, T., Jones, I., Jones, P., Lloyd, K., Nicholls, D., O'Neill, T., Qurashi, I., Ramzan, A., Series, H., Staufenberg, E., Brugha, T., 2010. A diagnostic system using broad categories with clinically relevant specifiers: lessons for ICD-11. International Journal of Social Psychiatry 56 (4), 326–335.

Layard, R., 2010. Economics. Measuring subjective well-being. Science 327 (5965), 534–535.

Lewinsohn, P.M., 1987. The coping-with-depression course. In: Muñoz, R.F. (Ed.), Depression Prevention: Research Directions. Hemisphere, Washington, DC, pp. 159–170.

Meltzer, H., Gill, B., Petticrew, M., Hinds, K., 1995. OPCS Surveys of Psychiatric Morbidity in Great Britain. Report 1: The Prevalence of Psychiatric Morbidity among Adults Living in Private Households. HMSO, London.

Melzer, D., Tom, B.D., Brugha, T.S., Fryers, T.F., Meltzer, H., 2002. Common mental disorder symptom counts in populations: are there distinct case groups above epidemiological cutoffs? Psychological Medicine 32 (7), 1195–1201.

Morrell, C.J., Slade, P., Warner, R., Paley, G., Dixon, S., Walters, S.J., Brugha, T., Barkham, M., Parry, G.J., Nicholl, J., 2009. Clinical effectiveness of health visitor training in psychologically informed approaches for depression in postnatal women: pragmatic cluster randomised trial in primary care. BMJ Clinical Research Ed. 338, a3045.

Munoz, R.F., Beardslee, W.R., Leykin, Y., 2012. Major depression can be prevented. American Psychologist 67 (4), 285–295.

Murray, C.J., Vos, T., Lozano, R., et al., 2012. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 380 (9859), 2197–2223.

Prince, M., Patel, V., Saxena, S., Maj, M., Maselko, J., Phillips, M.R., Rahman, A., 2007. No health without mental health. Lancet 370 (9590), 859–877.

Sandler, I.N., Ma, Y., Tein, J.Y., Ayers, T.S., Wolchik, S., Kennedy, C., Millsap, R., 2010. Long-term effects of the family bereavement program on multiple indicators of grief in parentally bereaved children and adolescents. Journal of Consulting and Clinical Psychology 78 (2), 131–143.

Sandler, I.N., Schoenfelder, E.N., Wolchik, S.A., MacKinnon, D.P., 2011. Long-term impact of prevention programs to promote effective parenting: lasting effects but uncertain processes. Annual Review of Psychology 62, 299–329.

Scott, K.M., Von, K.M., Alonso, J., Angermeyer, M.C., Bromet, E., Fayyad, J., De, G.G., Demyttenaere, K., Gasquet, I., Gureje, O., Haro, J.M., He, Y., Kessler, R.C., Levinson, D., Medina Mora, M.E., Oakley, B.M., Ormel, J., Posada-Villa, J., Watanabe, M., Williams, D., 2009. Mental-physical co-morbidity and its relationship with disability: results from the World Mental Health Surveys. Psychological Medicine 39 (1), 33–43.

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., Parkinson, J., Secker, J., Stewart-Brown, S., 2007. The Warwick-Edinburgh mental well-being scale (WEMWBS): development and UK validation. Health and Quality of Life Outcomes 5, 63.

Ware, J.E., 1993. SF-36 Health Survey. Medical Outcomes Trust, Boston, MA.

Weich, S., Brugha, T., King, M., McManus, S., Bebbington, P., Jenkins, R., Cooper, C., McBride, O., Stewart-Brown, S., 2011. Mental well-being and mental illness: findings from the Adult Psychiatric Morbidity Survey for England 2007. British Journal of Psychiatry 199 (1), 23–28.

Weich, S., Holt, G., Twigg, L., Jones, K., Lewis, G., 2003. Geographic variation in the prevalence of common mental disorders in Britain: a multilevel investigation. American Journal of Epidemiology 157 (8), 730–737.

Mental Health and Social Work

Di Bailey, School of Social Sciences, Nottingham Trent University, Nottingham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article is about social work relating to people with mental health problems. It explores historically how the care and treatment of service users with mental health issues have developed. This exploration is informed by four key themes: the advances in psychiatric treatments, the ways in which mental health and recovery are understood, a changing landscape of policy and legislation, and the influence of those with lived experience of mental distress and those who care for them on the reform of the service delivery agenda. The article offers international examples to illustrate the above and considers the role of mental health social workers in respect of the dilemmas, challenges, and opportunities that arise from the work they do.

Introduction

The World Health Organization (WHO, 2001) defines mental health as a state of well-being in which an individual realizes their abilities, can cope with the normal stresses of everyday life, can work productively or fruitfully, and is able to make a contribution to his/her community (WHO, 2001: 1). In contrast, mental ill health includes mental health problems and strain, impaired functioning associated with distress, symptoms and diagnosable mental health disorders such as schizophrenia and depression. In terms of prevalence, the WHO (2000) reports over a third of people in most countries with symptoms meeting criteria for one or more common mental health problems. Not surprisingly then mental ill-health is a worldwide public health concern accounting for a significant proportion of the global burden of disease. It is influenced by individual and family, biological, social, ecological, and environmental factors (European Commission, 2005: 4).

Within this article a social work contribution is important because often associated with mental ill health are conditions of disadvantage, stigma and trauma reflecting the legacy of social work as a profession rooted in societies' struggles to deal with poverty and associated social crises. These beginnings in Western Europe and North America following the Industrial Revolution in the nineteenth century and continuing into the present day, for example, in South Asia in response to armed conflict and the tsunami of 2004 (Fernando and Weerackody, 2009).

In England, the history of mental health social work can be traced back to the 1800s when the wife of Alfred Meyer, a psychiatrist, visited her husband's patients to learn more about their home backgrounds; and was later called a 'social worker' (MIND – 29.04.2008).

Modern day mental health-related social work is practiced in some way and to some degree in the majority of countries and has seen some significant changes since its early beginnings in the last century. These changes have been influenced by advances in psychiatric treatments, the ways in which mental health and recovery are understood, a changing landscape of policy and legislation, and the influence of those with lived experience of mental distress and those who care for them on the reform of the service delivery agenda. It is to these four areas that this article now focuses in turn whilst recognizing that these areas are inextricably linked and exert mutual influence.

Advances in Psychiatric Treatments

While social workers are less concerned with diagnosing mental disorders compared with their medical and nursing colleagues it is important that they are able to assess how the signs and symptoms of distress may present and the treatment options that are likely to be made available to individuals. The legacy of medical treatments dates back to the Arabic empire between the tenth and thirteenth centuries with the first medical approach to madness pioneered in 'moristans,' which were a kind of forerunners to the asylums in Spain in the fifteenth century. Fernando (2010) explains how these European asylums evolved to be under religious jurisdiction, and were devoted to treating mental illnesses such as melancholia and conditions resembling obsessive-compulsive disorders.

Between the Middle Ages and the twentieth century, different conceptions of madness prevailed depending upon a society's belief systems and political persuasions, shifting from spiritual to physical explanations. Concomitantly care and treatment developed from the increasingly punitive and harsh regimes of the early madhouses and asylums to the more humane approach which saw asylums renamed as 'mental hospitals' and inmates as 'patients.' The asylum era of the early nineteenth century took off in Europe and colonial states with these mental hospitals paving the way for a greater reliance on the contribution of medical doctors to the treatment of the mentally disordered. By the late 1800s doctors openly recognized mental disorder as a form of illness, biological and pharmaceutical treatments dominated and the symptoms of schizophrenia were first described by Kraepelin in the 1890s.

Compared with England and Europe treatment for mental illness was less hospital-centered in the United States until the Second World War. With the increase in conditions that came to be termed 'shell shock' there was a greater realization of the contribution of social and environmental stressors to psychopathology which coincided with the development of social work and clinical psychology as professions. Lewis (1974) reported on the introduction of training for social workers in the US in the early part of the twentieth century.

Despite a global trend of deinstitutionalization and the upsurge in community care in the twentieth century many mental health professionals, service users, and their families would contend that the disease model characteristic of the institutional era continues to dominate treatment interventions even to the present day. Although from the early to mid-1900s, mental disorders were seen to arise from a more complex interplay between genetics, biological vulnerabilities, and stress a hierarchy of models for understanding mental distress prevailed with an associated hierarchy of treatment modalities (see **Figure 1** below). According to Bond and Lader (1996: 4) "The notion of an illness implies that there is a fundamental difference from normal, and a categorical rather than a dimensional approach is therefore preferred." Whilst the hierarchy of models suggests a more multidisciplinary approach to mental health care that came to the fore from the 1970s onward (see Bailey, 2012), social work interventions in the form of social and practical support and more psychosocial-oriented talking treatments were subsumed within an overall approach based on a premise that once the illness was medically treated the need for other forms of intervention would be negated as the signs and symptoms would improve. Historically such treatments have revolved around psychotropic medication and the use of electroconvulsive therapy (ECT) with psychological therapies playing a comparatively insignificant role.

This way of working, while acknowledging professional contributions other than medicine to interventions for mental ill health, has maintained the differential status of the other professions relative to psychiatry and set the scene for resistance to a more integrated interdisciplinary approach that has subsequently been deemed necessary to meet the diverse needs of service users in the twenty-first century.

Against this back-drop, mental health social workers need an understanding of the treatments for disorders spanning the neuroses (predominantly disorders of emotion with symptoms that are commonly recognized) to the more severe psychoses usually characterized by disorders of thought and/or perception. In some countries such as England, Australia, and the US, social workers may be called upon to assess whether such disorders are severe enough to pose a risk to the safety of individuals and/or others and as such warrant compulsory treatment in hospital although increasingly nowadays this can be enforced in the community.

Anxiety disorders may be generalized or specific like phobias and also include obsessive-compulsive disorder. They are usually treated using newer Selective Serotonin Reuptake Inhibitor drugs alongside talking treatments such as cognitive behavioral therapy (CBT). Posttraumatic stress disorder is typified by severe anxiety symptoms and flashbacks usually in response to some sort of significant life event, crisis or trauma. The relationship between childhood abuse and neglect and symptoms of mental ill health in adulthood (e.g., self-injury) is increasingly being seen as mediated through a type of post-traumatic stress response.

Mood disorders include those considered to stem from a reaction to a life event usually involving loss (e.g., symptoms of depression in response to bereavement) through to those where the cause seems elusive and is usually attributed to an internal biochemical imbalance. Severity varies from mild to major depression and bipolar disorder (also known as manic depression) is typified by abnormally high or pressured mood states often with associated thought disorder and problems with concentration/perception. Treatment varies depending upon an assessment of the individual's symptom profile and a range of interventions may be offered from antidepressant therapy with or without CBT, through to a combination of mood stabilizing drugs such as lithium preparations and carbamazepine augmented by antidepressants in small doses (Bond and Lader, 1996).

Treatment for psychotic disorders such as schizophrenia may also vary depending upon the extent to which symptoms such as hallucinations and delusions override those of withdrawal and lethargy (referred to as negative symptoms) which

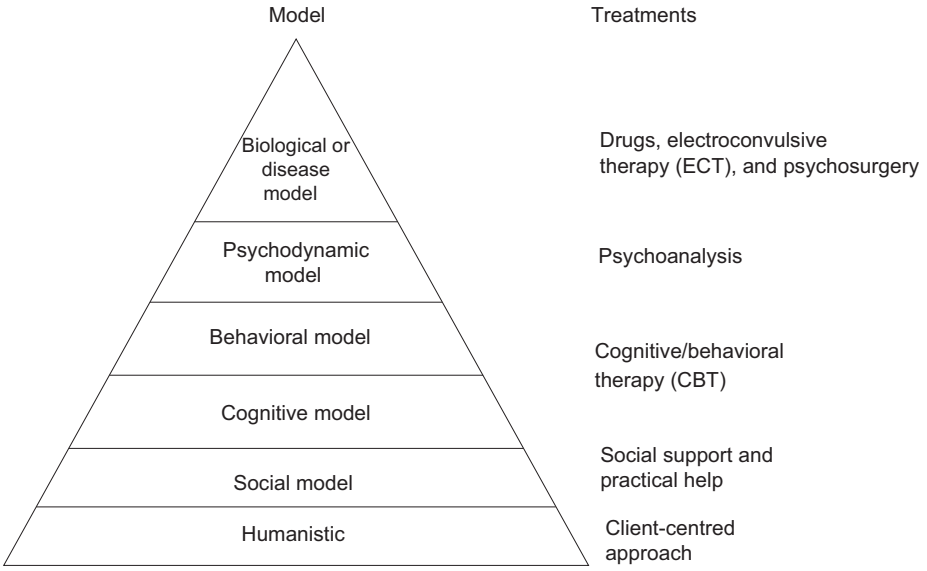

Figure 1 Hierarchy of models of mental distress in the mid-1900s.

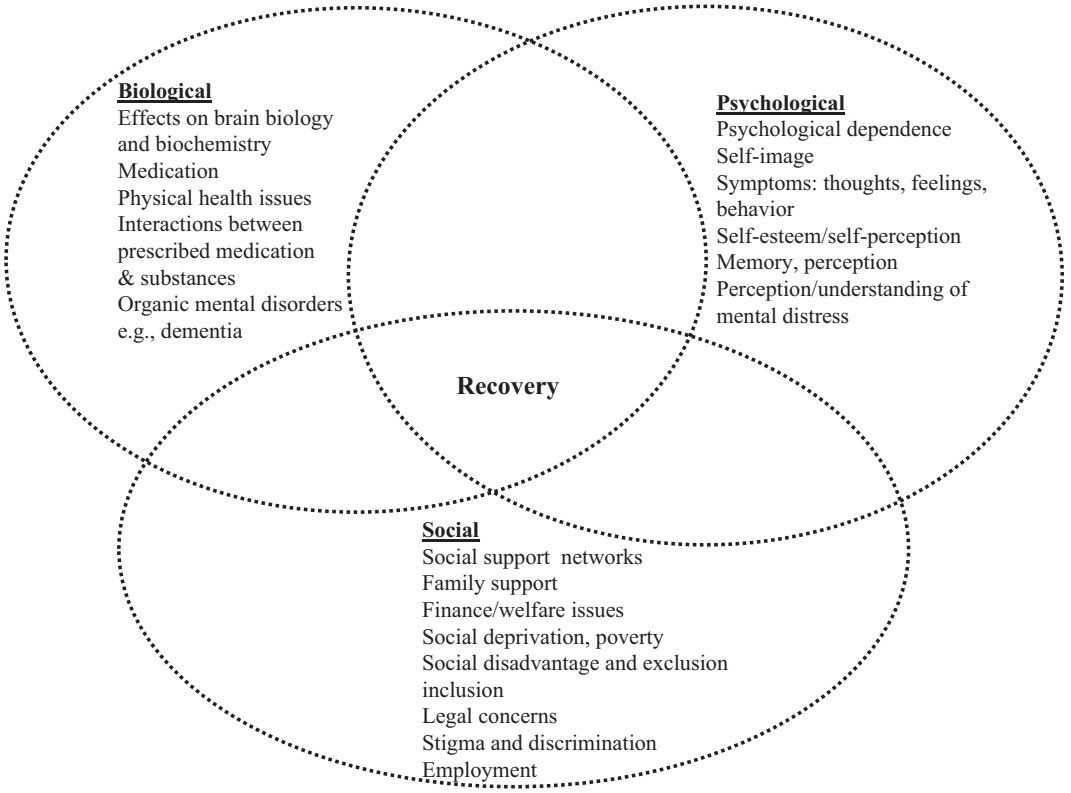

Figure 2 Recovery model of mental distress.

can look more similar to depression than psychosis. Treatment regimes have progressed from the major tranquilizers such as chlorpromazine to drugs such as clozapine which is now regarded as particularly effective for the treatment of these 'negative' type symptoms.

Increasingly, the importance of detecting the onset of psychoses early and using prevention techniques associated with the identification of prodromal symptoms has become the focus for modern day mental health care. Social workers in mental health are likely to find themselves helping service users and their caregivers to identify signs of deterioration in mental health before a full-blown psychotic episode develops. These 'relapse signatures' once identified can be especially important for individuals to learn how to manage their mental health more effectively (Nelson, 1997) and it is therefore important for social workers to be able to work with relapse signatures to prevent unnecessary hospitalization (Brown, 2009).

Diagnostic tools like the DSM-IV (in the US) and the ICD-10 in the UK are perhaps the most common but not the only classification systems that are used to help mental health professionals identify, through the process of an interview and assessment, the different types of mental disorders with which individuals may present.

Whilst the existence of such tools suggests that disorders fall neatly into specific categories this is often not the case. Eating disorders, for example, often present with associated symptoms of depression and anxiety and altered perception of body shape which can take on an almost delusional quality particularly in anorexia nervosa (Fernandez and Rothery, 2000). Personality disorders, characterized by abnormally rigid and maladaptive traits with associated behaviors that may be acutely dysfunctional or more chronic and difficult to manage also often present with symptoms of depression and overwhelming distress (Ladd and Churchill, 2012).

Where mental health social workers are particularly skilled as a result of their training and professional development is in recognizing the interplay of risk factors for mental ill health and in using individual and family-focused techniques such as CBT, systemic family therapy, and dialectical behavior therapy to intervene.

Mental Health and Recovery

The move toward more holistic frameworks for understanding why people's mental health deteriorates include the stress vulnerability model as mentioned above (Zubin and Spring, 1977), the biopsychosocial model (Engel, 1977, 1980), and the recovery model (Deegan, 1988; Anthony, 1993). The latter, in particular, reflects a recognition that modern day mental health practitioners increasingly need to adopt a 'being with' rather than 'doing to' approach when working with service users (Hinshelwood, 1998).

Interestingly, despite the advances in psychiatric treatments identified above long-term international studies of schizophrenia (Harrison et al., 2001; Jobe and Harrow, 2005) have found that medication makes little difference to symptoms in about a third of service users whilst another

third will report a reduction in symptomatology that may be accompanied by a regaining of social and psychological functioning, sometimes without any medication. According to Romme and Escher (1989) who have pioneered work on voice hearing in the Netherlands, neuroleptic medication has little effect on voices heard by individuals with a diagnosis of schizophrenia but do reduce the anxiety and distress often provoked by the voices.

Similarly, of people diagnosed with bipolar disorder, around half report a recovery from symptoms usually within a 6 week period with nearly all achieving this within 2 years (Tohen et al., 2003). Unfortunately, for service users with this type of disorder relapse tends to affect more than half with individuals.

What these studies suggest to mental health social workers is that there will be times when individuals with mental health problems are well and will be better able to develop coping strategies for managing their symptoms that manifest when they are ill. This in effect is the essence of the 'recovery model' which acknowledges that individuals may continue to experience symptoms of mental distress whilst retaining a high degree of control over their lives.

Seen as a process, rather than a state of being, effective 'recovery' draws from social, psychological, and medical explanation of mental health and illness, embracing these contributions from the different disciplinary groups to best effect and including the experiences of the person with the mental health problem and their family. This is perhaps most clearly articulated by Perkins and Repper (1998: 25) who assert that "different people adopt different models for understanding what has happened to them … organic constructions … psychological, social, religious or spiritual formulations. People have a right to define their own experiences for themselves and it is rarely helpful and more likely to be alienating-for the clinician to insist that their understanding is correct" (**Figure 2**).

The recovery model also allows for the influence of culture on understanding mental distress. For example, in African and Asian cultures individuals experiencing symptoms may be more likely to turn to traditional healers rather than access approaches that stem from mainstream Western psychiatry. The recovery approach also chimes with an increasing interest in the concept of well-being as a way of paying greater attention to how individuals attribute meaning to their mental health. According to Diener and Suh (2000), well-being is based upon the values and standards people ascribe to in their lives, their success or failure in achieving outcomes and pursuits which they personally seek and their ability to achieve fulfillment of their basic human needs.

Therefore, rather than assess how a service user's symptoms of mental ill health can be eradicated social workers need to be more concerned with what help and interventions can be mobilized to assist the service user manage their mental health needs. This stance supports risk management as a shared endeavor requiring more emotional closeness or 'hands on approach' (Ramon and Williams, 2005) with the social worker genuinely interested in the everyday concerns which matter most to the service user.

Bracken and Thomas (2004) identify this 'thinking outside' traditional models of mental ill health as 'post psychiatry' of the twenty-first century. They see approaches to recovery and well-being as inextricably linked with the concept of citizenship which they define as "being free from discrimination, exclusion and oppression" and "being able to define one's own identity and celebrate this in different ways" (Bracken and Thomas, 2004: 6). This is an important concept for mental health social workers to grasp as individuals often report that their experiences of discrimination and exclusion as a mental health service user are often more serious than the signs and symptoms of their mental distress that they cope with on a day to day basis. Social work interventions should, therefore, tailored toward countering discriminatory practice and promoting interventions that increase individuals' understanding of their mental health needs and how they might work collaboratively with mental health professionals to achieve positive outcomes.

Policy and Legislative Reform

In 2013, three-quarters of countries around the world have mental health legislation to prevent people with mental health problems being detained in hospital against their wishes. This trend of legislating for detention and the giving of treatment dates back in some countries to the early twentieth century. According to Unsworth (1987: 5), mental health law forms the foundations of the mental health care system which influences the 'intricate mechanisms of control, surveillance and discipline.'

For example, in England the 1930 Mental Treatment Act provided for voluntary treatment and psychiatric outpatient clinics. As mental disorder became increasingly medicalized as opposed to being seen as the manifestation of immoral or criminal conduct the Royal Commission on the Law Relating to Mental Illness and Mental Deficiency was set up and resulted in the Percy Report in 1957. The key themes of this report were that mental disorder should be regarded "in much the same way physical illness and disability" (para 5) and that mental hospitals should be run as nearly as possible like hospitals for physical disorders. This trend has been replicated in the way that mental hospital services evolved over the last century throughout the world.

The role of social workers in the assessment and treatment of individuals with mental health problems has been introduced at different times and in a range of ways in different countries. In 1913, three clinical social workers were introduced to the hospital in the United States and set the trend of clinical social work which today revolves around education, discharge planning and linking service users and their families to resources.

In England, it was the 1959 Mental Health Act that introduced the role of the Mental Welfare Office (MWO) as an applicant for the admission to hospital based on the recommendations of two medical practitioners. Eventually, MWOs were replaced by Approved Social Workers with the introduction of the 1983 Mental Health Act. In the UK, Approved Social Workers, and in the US, clinical social workers, require additional training and specialism in psychiatry in order to be able to fulfill their statutory responsibilities. In the US, clinical social workers are required to be licensed to practice in any state across the country, while in the UK, the registration of

Approved Social Workers has been undertaken at a local level by social services departments. Since the introduction of the 2007 Mental Health Act in England and Wales, other mental health professionals such as Community Psychiatric Nurses, Occupational Therapists, and Clinical Psychologists can take on the statutory role and remit historically limited to Approved Social Workers if suitably trained and experienced. The professional title has changed to an Approved Mental Health Professional to reflect that individuals fulfilling this role can now be drawn from a range of disciplinary backgrounds.

The conditions that relate to hospitalization or in some countries compulsory treatment in the community differ in the degree to which they require the presence of mental disorder, the need for treatment and or the assessment of risk in the form of a threat or danger to oneself or others. Salize (2006) identifies that in the existence of mental disorder *and* the existence of a threat or danger has to be present for involuntary admission in France or Germany whilst in England, Wales Portugal, and Finland the need for treatment must also be demonstrated. In Sweden and Italy, if mental disorder is evident alongside a need for treatment, this can be sufficient to warrant compulsory admission and the risks of threat or danger do not apply. Similarly, countries across the globe also differ in the ways in which they define mental disorder. Increasingly, the issue of mental capacity is defined separately and dealt with by related legislation to that which deals with mental illness. A significant change made by the Mental Health Act 2007 in England and Wales was that it introduced a singular definition of mental disorder as 'any disorder or disability of the mind' resulting in an abolition of the categories of mental illness, psychopathic disorder and mental impairment that were present in the 1983 Mental Health Act. The latter is now dealt with by the legislation relating to mental capacity.

This shift in the definition of mental disorder occurred alongside an extension of the definition of medical treatment in England and Wales to include: "nursing, psychological intervention and specialist mental health habilitation, rehabilitation and care" (Mental Health Act, 2007) which accords more closely with the recovery model of mental health outlined above (Bailey, 2012).

As well as setting out conditions under which individuals with mental health problems may be compulsorily treated mental health legislation also provides for a right of appeal either by the service user directly, indirectly by their caregivers or increasingly with the help and support of independent advocates. In this respect, mental health law begins to dovetail with the Human Rights Act of 1998 that incorporates the Convention for the Protection of Human Rights and Fundamental Freedoms usually referred to as the European Convention on Human Rights.

Article 5 of the European Convention on Human Rights deals with everyone's right to liberty and personal security and states that no one shall be deprived of their liberty except in certain circumstances and in accordance with a procedure set out in law. In Article 8, individuals have a right to respect for private and family life. The challenge for social workers in the field of mental health is to be able to make informed judgments, based on sound evidence where the individual rights of service users are superseded by the rights of the general public or other people who may be at risk from the service user's behavior, stemming from their mental health problem. This is never an easy decision-making process and one not to be undertaken lightly. Importantly, social workers need to be able to treat individuals with dignity and respect promoting choice wherever possible whilst recognizing that when people are unwell their choices may be limited. It is these fundamental values that inform best social work practice and which service users report are of most import to them.

Influence of People with Lived Experiences of Mental Ill Health and Those Who Care for Them on Service Development

According to Gosling (2010: 38), "Involvement is a two-way street – involvement is not just about us being involved in your (mental health professionals) lives – it is about you being involved in ours." In the UK, Gosling has been one of the key social activists in involving service users in the training of mental health social workers since the early 1980s.

The influence of individuals and their families on mental health care began once the programs of deinstitutionalization got under way in Western countries (Bailey, 2012). The survivor movement in the UK gathered momentum from the 1950s onward in response to individuals' experiences of psychiatric treatment. For example, a number of individuals began to report lasting memory impairments as a result of ECT and subsequently to speak out about the long-term consequences of their disability. What survivors of the psychiatric system seemed to have had in common was a collective experience of social oppression and discrimination being articulated more widely by the disabled peoples' movement in the UK as disability politics came more to the fore. Berresford and Wallcraft (1997), however, explain how the survivors' movement was in turn distinct, as a result of the mental health context. They also emphasized that the success of service user organizations such as Survivors Speak out in the UK was down to this being an organization that embraced people from a 'broad church' of backgrounds and political persuasions with no pressure to conform to a particular belief system.

Since the 1950s the number of service user groups has proliferated largely in response to policy changes such as the National Service Framework in England and similar policy initiatives in other countries (e.g., the National Mental Health Program in Norway). These developments have meant that the involvement of people with lived experiences and their families in the way that care and treatment is delivered is recognized and articulated as best practice at government level even though its implementation at a local level may remain fraught with complexities and a challenge to achieve.

The lobbying on behalf of those with mental health issues is increasingly undertaken on an international scale by a growing range of independent non-government organizations/charities. In 1946 the National Association of Mental Health in the UK joined other organizations to form MIND (a charity organization) in 1972. Basic Needs is an international mental health charity working in India, Ghana, and Vietnam since 1999, and SANE Australia (another charity organization) has

been in existence since 1986. These organizations undertake a range of functions on behalf of citizens and their families. These range from providing help lines and information to research into treatment interventions and campaigns to improve services and inform policy and legislative change. For example, MIND has recently responded to the current policy initiative of Improving Access to Psychological Therapies in the UK commenting that whilst greater access is welcomed waiting times in some areas are still problematic and regional variation of services is of increasing concern.

Other national service user groups and networks have played a key role in changing the way professionals think about and intervene with specific mental health issues. The National Hearing Voices network for example has redefined voice hearing as no-longer the debilitating condition associated primarily with schizophrenia that is was understood to be in the age of the asylum. Instead/with contributions from the likes of Romme and Escher in the Netherlands and Ron Coleman in the UK voice hearing has been demystified. Today it is understood as often a natural and short-term response to trauma or an experience which can be controlled and lived with using a range of cognitive and behavioral strategies. The premise of the hearing voices network is on the value of self help and peer support.

In the case of the latter, there is a growing body of literature testifying to its value for a range of other mental health conditions and influence on the creation of peer support roles in the mental health workforce, in America (Davidson et al., 1999), Canada (Canadian Foundation for Healthcare Improvement, 2010), and Australia (McKenzie, 2012). These roles can augment; in important ways the statutory role of mental health social workers as peer support workers can provide more hands on practical help to prevent crises and hospitalization. They can also provide valuable information about how service users and their caregivers are managing on a daily basis that in turn can assist mental health social workers make a judgment about whether compulsory treatment is necessary. For these reasons, peer support workers are becoming key members of multidisciplinary mental health teams such as Crisis Assessment and Treatment Teams in Australia and Assertive Outreach Teams in the UK.

Involvement in one's individual care package and more collectively with service user groups and organizations is reported to offer a wealth of benefits for service users such as a strong sense of self-worth and shared experience and understanding. In a systematic review by Crawford et al. (2002), 42 studies described the outcomes of involving service users and revealed that although some people were dissatisfied with the process of involvement, others welcomed the opportunity to participate, and their self-esteem improved as a result. In several of the 42 studies, staff also found the experience of involving those who used the service rewarding. Changes to services most frequently reported as a result of participation were the production of new or improved information for individuals about their care, alongside efforts to make services more accessible through improved appointment procedures, opening times or transport arrangements. In the same review Crawford et al. find examples of new services being commissioned at the request of service users. Other effects of involvement included more favorable staff attitudes and organizations becoming more 'open' to the concept of involvement.

However, Gosling (2010) would argue that 'involvement' is only a step toward service user emancipation that allows individuals to take power from the professionals rather than have it given. Stickley (2006) offers up projects such as the Service Users Research Group England (SURGE), User Focused Monitoring Schemes, and the Strategies for Living project at the Mental Health Foundation 2000 as examples of service users providing services to mainstream mental health services but not being managed by them. He argues that this places service user groups and organizations in a different position from which to negotiate collaboration and involvement.

The other group of individuals who have a role to play on influencing the development of mental health care and treatment are service users' caregivers. They can provide important information to mental health social workers to help in the decision-making process and have legal rights and responsibilities in relation to the care of their relatives. In England the National Service Framework (DH, 1999a) reemphasized the 'Care Programme Approach' (CPA) originally introduced in 1991, as the mechanism for care planning in mental health. The Framework stated explicitly how carers should be involved in the CPA (DH, 1999b) as a result of subsequent legislation (the Carers (Recognition and Services) Act 1995) which specifically outlines carers' entitlements generally not just in relation to mental health. In particular mental health social workers need to be especially concerned with the needs of young people caring for a relative with mental health issues as their rights are often enshrined in legislation that straddles that of adult mental health and child protection (Brown, 2009).

In some countries such as Scotland, Integrated Care Pathways have been developed (McQueen and Milloy, 2001) for service users and their caregivers as a guide to what each can expect to happen by certain stages of the assessment process. This kind of information can be especially troubling for young people who are often concerned what will happen to their relative if they talk openly about the types of mental health issues and risks they are dealing with.

Concluding Comments

The above account provides merely a flavor of the range and type of issues and opportunities facing social workers in mental health settings particularly given the complexity of the discipline and its international diversity. It remains for the author one of the most interesting and humbling areas of social work practice where the values of empathy, warmth and genuineness remain as fundamental. As the number of individuals with mental health issues grows globally the need for practitioners who can strike an effective balance between the preservation of human rights and the person-centered management of mental distress becomes ever more important. The landscape of mental health care in the twenty-first century requires that contribution from mental health social workers.

See also: Culture as a Determinant of Mental Health; Mental Health; Normality and Mental Health; Unemployment and Mental Health.

Bibliography

Anthony, W.A., 1993. Recovery from mental illness: the guiding vision of the mental health service system in the 1990s. Psychosocial Rehabilitation Journal 16 (4), 11–23.

Bailey, D., 2012. Interdisciplinary Working in Mental Health. Palgrave Macmillan, Basingstoke.

Beresford, P., Wallcraft, J., 1997. Psychiatric system survivors and emancipatory research: issues, overlaps and differences. In: Barnes, C., Mercer, G. (Eds.), Doing Disability Research. Disability Press, Leeds, pp. 66–87.

Bond, A.J., Lader, M.H., 1996. Understanding Drug Treatment in Mental Health Care. John Wiley and Sons, Chichester.

Bracken, P., Thomas, P., January/Feburary 2004. Post psychiatry is not another model! Openmind 125, 6–7.

Brown, R., 2009. The Approved Mental Health Professional's Guide to Mental Health Law, second ed. Learning Matters, Exeter.

Canadian Foundation for Healthcare Improvement, 2010. Ingredients for Successful Patient Engagement: The Power of Relationships. Canadian Foundation for Health Care Improvement, Ottowa.

Crawford, M.J., Rutter, D., Manley, C., Weaver, T., Bhui, K., Fulop, N., Tyrer, P., November 30, 2002. Systematic review of involving patients in the planning and development of health care. British Medical Journal 325, 1263–1265.

Davidson, L., Chinman, M., Kloos, B., Weingarten, R., Stayner, D., Tebes, J.K., 1999. Peer support among individuals with severe mental illness: a review of the evidence. Clinical Psychology Science and Practice 6, 165–187.

Deegan, P.E., 1988. Recovery: the lived experience of rehabilitation. Psychosocial Rehabilitation Journal 11 (4), 11–19.

Department of Health (DH), 1999a. National Service Framework for Mental Health: Modern Standards and Service Models. Department of Health, London.

Department of Health (DH), 1999b. Effective Care Coordination in Mental Health Services: Modernising the Care Programme Approach - A Policy Booklet. Department of Health, London.

Diener, E., Suh, E.M., 2000. Measuring subjective well-being to compare quality of life of cultures. In: Diener, E., Suh, E.M. (Eds.), Culture and Subjective Well-being. MIT Press, Cambridge, Massachusetts, pp. 3–12.

Engel, G.L., 1977. The need for a new medical model: a challenge for biomedicine. Science 8, 129–136.

Engel, G.L., 1980. The clinical application of the biopsychosocial model. American Journal of Psychiatry 137, 535–544.

European Commission, 2005. Green Paper – Improving the Mental Health of the Population. Towards a Strategy on Mental Health for the European Union. Brussels 14.10.2005, Com (2005) 484.

Fernandez, C., Rothery, D., 2000. Mental health problems in young people. In: Bailey, D. (Ed.), At the Core of Mental Health Practice. Pavilion, Brighton, pp. 61–79.

Fernando, S., 2010. Mental Health, Race and Culture, third ed. Palgrave Macmillan, Basingstoke.

Fernando, S., Weerackody, C., 2009. Challenges in developing community mental health services in Sri Lanka. Journal of Health Management 11 (1), 195–208.

Gosling, J., 2010. The ethos of involvement as a route to recovery. In: Weinstein, J. (Ed.), Mental Health Service User Involvement and Recovery. Jessica Kingsley, London, pp. 30–44.

Harrison, G., Hopper, K., Craig, T., 2001. Recovery from psychotic illness: a 15 year and 25-year international follow-up study. British Journal of Psychiatry 176 (6), 506–517.

Hinshelwood, R.D., 1998. Creatures of each other: some historical considerations of responsibility and care, and some present undercurrents. In: Foster, A., Roberts, V.Z. (Eds.), Managing Mental Health in the Community: Chaos and Containment in Community Care. Routledge, London, pp. 15–26.

Jobe, T.H., Harrow, M., 2005. Long-term outcome of patients with schizophrenia: a review. Canadian Journal of Psychiatry 50 (14), 892–900.

Ladd, P.D., Churchill, A.M., 2012. Person Centred Diagnosis and Treatment for Mental Health – A Model for Empowering Clients. Jessica Kingsley, London.

Lewis, N.D.C., 1974. American psychiatry from its beginnings to World War II. In: Arieti, S. (Ed.), American Handbook of Psychiatry, second ed., vol. 1. Basic Books, New York, pp. 28–42.

McKenzie, P., 2012. Single session peer work: a framework for peer support. In: Paper Presented at the Mental Health Services Annual Conference Cairns Australia, 21–24 August. www.peersupportvic.org/resources-directory.

McQueen, J., Milloy, S., 2001. Why clinical guidelines or care pathways in mental health? Journal of Integrated Care Pathways 5 (2), 44–53.

Nelson, H., 1997. Cognitive Behavioural Therapy with Schizophrenia: A Practice Manual. Stanley Thornes, Cheltenham.

Perkins, R., Repper, J., 1998. Principles of working with people who experience mental health problems. In: Brooker, C., Repper, J. (Eds.), Serious Mental Health Problems in the Community. Policy, Practice & Research. Balliere Tindall, London.

Ramon, S., Williams, J.E., 2005. Towards a conceptual framework: the meanings attached to the psychosocial, the promise and the problems. In: Ramon, S., Williams, J.E. (Eds.), Mental Health at the Crossroads: The Promise of the Psychosocial Approach. Ashgate, Aldershot, pp. 13–24.

Romme, M.A.T., Escher, A.D., 1989. Hearing voices. Schizophrenia Bulletin 15, 209–216.

Salize, J., 2006. European Mental Health Laws, Social Inclusion and Fundamental Rights. Central Institute of Mental Health, Mannheim, Germany.

Stickley, T., 2006. Should service user involvement be consigned to history? A critical realist perspective. Journal of Psychiatric and Mental Health Nursing 13 (5), 570–577.

Tohen, M., Zarate, C.A., Hennen, J., 2003. The McLean Harvard first episode Mania study: prediction or recovery and first occurrence. American Journal of Psychiatry 160 (12), 2099–2107.

Unsworth, C., 1987. The Politics of Mental Health Legislation. Clarendon Press, Oxford.

WHO, 2001. Strengthening Mental Health Promotion. World Health Organization, Geneva (Fact sheet, No. 220).

World Health Organization International Consortium in Psychiatric Epidemiology, 2000. Cross-national comparisons of the prevalence and correlates of mental disorders. Bulletin of the WHO 78 (4).

Zubin, J., Spring, B., 1977. Vulnerability: a new view on schizophrenia. Journal of Abnormal Psychology 86, 103–126.

Mental Health, Mentalism and Sanism

Jennifer M Poole and Tania Jivraj, School of Social Work, Ryerson University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mentalism or sanism is a devastating form of oppression, often leading to negative stereotyping and arguments that individuals with mental health histories are not fit to study, work, or enjoy basic human rights. Despite 50 years of scholarship and recent developments in policy and research, the terms are still rarely used, understood, or interrogated in the social sciences. Indeed, social work in particular has been so loyal to the medical model that mentalist/sanist aggressions, such as pathologizing, labeling, exclusion, and dismissal have become a typical part of best professional practice and education. In this article, we provide an overview, history, analysis of, and response to mentalism and sanism.

Introduction

In this article, we detail mentalism and sanism with respect to mental health, locating our review in social work scholarship to date. We detail the roots of mentalism, their connection to the movements of ex-patients, consumers, survivors, consumer/survivors, and Mad people. We also explore the effects of sanist practices in particular on social work and suggest possibilities that not only respond to sanism but those grounded in critical, structural theoretical work, that support a practice of anti-sanism.

Tracing the Terms

Social sciences scholarship, and in particular social work literature, has long been known for its anti-oppressive analyses of issues such as racism, heterosexism, and ableism. Yet another form of oppression has received comparatively little attention – namely, mentalism as coined by the late American mental health consumer activist Judi Chamberlin.

In her foundational article on the ex- (mental health) patients' movement (1990), Chamberlin argues that in the early days of organizing, ex-patients (also known as consumers, survivors, consumer/survivors, users, refusers, lunatics, crazies, or Mad) identified mentalism and sane chauvinism as a set of negative assumptions which most people, including ex-patients themselves, held about 'mental patients.' More recently, mentalism has been defined as the systematic subjugation of people who have received mental health diagnoses or treatment. Like racism, mentalism may result in blatant discrimination, but will be most commonly expressed in "multiple, small insults and indignities" known as "microaggressions" (Kalinowski and Risser, 2005: p. 1). These microaggressions will include low expectations and professional judgements that such individuals are helpless, incompetent, violent, and irrational (Chamberlin, 1990).

In some circles, mentalism is also known as sanism (Birnbaum, 1960; Perlin, 1992). As explained by Ingram (2011), sanism was coined by Morton Birnbaum, a graduate of Columbia Law School who also gained an MD from New York Medical School. Influenced by the work and friendship of black legal scholar Florynce Kennedy, in 1958 Birnbaum began to argue for the constitutional right to adequate treatment for patients in US state mental hospitals. After it had been rejected by 50 other publications, Birnbaum's article, 'The Right to Treatment,' which outlined sanism, was published by the *American Bar Association Journal* in 1960. In 1980, Michael Perlin, a professor at New York Law School, applied Birnbaum's work, to identify discriminatory behaviors in law to which he had become accustomed and has focused his prolific scholarship on this issue ever since.

In the first of his influential papers, Perlin defines sanism as "… a form of bigotry that respectable people can express in public" (1992: p. 375). He notes how deeply the roots of sanism are entwined with social notions of sin, evil, God's punishment, crime, and demons (1992: p. 388). This linking allows for the labeling of "sickness as reassurance that the other … is not like us" (p. 390), creating the conditions in which those with mental 'illness' have been made largely invisible and without power, institutionalized, and deprived of the right to vote or the right to be parents (p. 392). It follows that certain sanist myths about people with mental health issues and labels have flourished. According to Perlin, these include the notion that such individuals are different, erratic, weak, frightening, and unable to make decisions or participate in the public arena. In addition, there are common myths that declining or refusing medications is a predictor of future 'dangerousness,' that mental health issues can easily be identified by lay persons, that what is depicted in the media is the truth, and finally that using pejorative labels to describe such people, including crazy, psycho, or insane is not as problematic as other recognized hateful terms. As inspired by Kimberlé Williams Crenshaw's work on intersectionality, like Birnbaum, Perlin recognizes that sanism cannot be disentangled from other oppressions.

Current Conceptualizations

Putting Perlin's work into practice, The Coalition Against Sanist Attitudes (CASA, 2011) explains that sanism and its myths about people with mental health issues have flourished because of particular factors. These include the practice of rejecting evidence to the contrary, basing our sanist responses on 'gut feelings,' anecdotes portrayed by the media, and our history of segregating and distancing ourselves (emotions, thoughts, behaviors) from those labeled as 'different,' odd, or psychiatrically 'unwell.'

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28110-5

Another contributing factor is the dominance of the medical approach to mental health. The medical model is situated in a positivist or generally modernist paradigm. Born of enlightenment ideas which have subsequently privileged binaries, order, objectivity, de-contextualization, and the scientific approach to cure or recovery from illness, this paradigm has dominated western approaches to what is 'normal' or right with respect to social behavior, feelings, moods, actions, and desires. Anything outside this scientific norm is thus pathologized, categorized, and treated in accordance with medical/biological frameworks. With the best of intentions, this paradigm has seen the proliferation of conventional and (neuro)psychiatry, abnormal psychology, biological determinants of mental health, and binaries such as sane/insane and well/sick. All are centered on the individual as the site of the 'problem.'

In contrast, the theoretical roots of mentalism and sanism are aligned with a more structural, critical, or postconventional paradigm. Through these lenses, the focus changes to oppression, power, privilege, language, multiplicities of mood, emotion and ways of being, de-categorization, contextualization, connection, and the intersection of identity with multiple forms of oppression. In this theoretical place, we find mental health approaches that include but are not limited to feminism, antiracism and anticolonial work, critical race work, critical psychiatry, antipoverty work, anti-oppressive practice (AOP), social determinants of mental health approach, queer studies, narrative analysis, critical disability studies, sociology of the body, and, of course, Mad Studies.

Indeed, 21 years after Perlin's seminal piece on sanism, the conversation around mentalism and sanism has continued, especially in Mad Studies (see LeFrançois et al., 2013). For one such Mad scholar, mentalism speaks to the "mentalization or reduction of socially interpreted acts, feelings and thoughts to essential categories of mind, especially those not categorized as criminal" (Fabris, 2012: p. ix). Sanism is "the division of persons into 'mad' and 'sound,' especially based on an interpretation of ill logic, reason, or wisdom, such as in 'delusion'" (Fabris, 2012: p. ix). However, it is sanism (rather than mentalism) that appears to be more widely used within the Mad community (Diamond, 2012). It has also been broadened by Hamer (2011) to include 'state sanism' and 'institutional sanism.' Inspired by Michel Foucault, state sanism orders all citizens into normal and abnormal categories. Institutional sanism assumes that the abnormal are unpredictable and violent. Through psychiatry, medication is then justified in creating docile and compliant bodies of these abnormal subjects, thus lowering their 'threat' to society.

Stigma and Sanism

Despite this work on sanism, there has been far more interest in stigma, with various theoretical models, approaches, and levels carefully articulated by a wide range of researchers. Of note is Goffman's description of stigma as a deeply discrediting attribute (1963). Similarly, many have been influenced by the work of Link and Phelan (2001) who break the process of stigmatization down into labeling and stereotyping. Other authors have drawn our attention to the dangers of internalized stigma which makes individuals fearful of seeking help and support and the notion of public stigma when the general public agree with negative stereotypes (Corrigan and Shapiro, 2010). However, according to Holley et al. (2012), most of this kind of research focuses on the individual being stigmatized and not on the sanist power relations that make stigmatization possible in the first place.

For this reason, some argue for a research and practice focus on sanism rather than stigma. CASA (2011) explains that the focus must be on sanism because it is sanism which makes possible and gives rise to stigma. Similarly, Thornicroft (2006) and Sayce (1998) argue that a focus on stigma was always too limiting for not only does it perpetuate medical conceptions/language around mental health, but it minimizes the jagged reality of widespread rights abuse and oppression (or sanism) experienced by Mad individuals. Large and Ryan (2012) have argued the same with respect to the sanist belief in the dangerousness of Mad people. Finally, in a recent review of stigma and anti-oppressive social work practice, Holley et al. (2012) find the term stigma to be grossly inadequate. Citing Castro and Farmer's work on HIV (2005), they argue that "stigma is often the tip of the iceberg … and has frequently served as a means of giving short-shrift to powerful social inequalities … that are much harder to identify and conceptualize" (Holley et al., 2012: p. 54). Indeed, some argue that a singular focus on stigma whether in research, education, or the media is itself a sanist practice.

Sanism and Social Work

Given this, as well as the dominance of the medical model and widespread social investment in sanist myths, it follows that rare is the acknowledgment of mentalism or sanism in social work. With few exceptions (see Holley et al., 2012; LeFrancois, 2011a,b; Poole et al., 2012), it is often the case that in social work "offensive and injurious practices are integrated into everyday procedures to the point where we no longer recognize them as discrimination" (Kalinowski and Risser, 2005: p. 2).

For example, some have argued that social work needs to be protected from dangerous practices by social work students with mental health issues and indeed that acceptance to social work programs be delayed in these instances. Some feel that social work educators should not allow students with these issues to graduate at all, and in some jurisdictions, national debates continue around (un)suitability, mental health disability, and social work education. These debates make sense given that social work has long been loyal to the medical model of mental health, pushing other approaches to the periphery. However, it is not the only possible orientation for social work scholars interested in the study of sanism. Indeed, more productive have been the critical theoretical approaches known as anti-oppressive practice AOP (Dominelli, 2002), intersectionality (Crenshaw, 1989), and the social model of disability (Barnes, 1998).

AOP is the latest iteration of critical social work, an approach that includes but is not limited to feminist, structural, and antiracist social work. An AOP stance assumes that there are multiple forms of oppression, that oppression is tied to unequal power relations, and that critical reflection on these matters is paramount (Healy, 2005). In addition, with AOP the focus must be on creating spaces that promote service users'

empowerment and working in partnership with those users in practice, pedagogical, and research settings.

Also used by Perlin, the second approach is intersectionality (Crenshaw, 1989). Originated by critical race and legal theorist Kimberlé Williams Crenshaw while thinking through how gender and oppression interact in Black women's lives, she wrote

> Consider an analogy to traffic in an intersection, coming and going in all four directions. Discrimination, like traffic through an intersection, may flow in one direction, and it may flow in another…and sometimes, from all of them. Similarly, if a Black woman is harmed because she is in an intersection, her injury could result from sex discrimination or race discrimination […] But it is not always easy to reconstruct an accident: Sometimes the skid marks and the injuries simply indicate that they occurred simultaneously, frustrating efforts to determine which driver caused the harm.
>
> Crenshaw, 1989: p. 149

With respect to 'mental health,' it can be similarly impossible to see if the 'injuries' incurred by a Mad person result from sex or race discrimination, from poverty, and/or from sanism.

Finally, the social model of disability (Barnes, 1998) locates the difficulties that Mad people face not only in their issues, but also in pathologizing and sanist societies. Depending on social location, gender, race, class, and sexual orientation, these power relations make it possible for individuals acting 'differently' to be taken against their will to the hospital; for workers to be denied employment because of 'mental health' histories; or for Mad students to be excluded from social work programs.

Through all these lenses (and like Perlin), it is now possible to identify sanist behaviors to which social workers have become accustomed. For example, social work discourses on mental health have become rife with references to diagnosis, classification, and inappropriate behavior, and are heavily influenced by positivist, traditional forms of research that regularly exclude service users from the production of knowledge. In addition, many of these courses are organized around pathology, intervention, and working with clients to 'stay on their meds,' and many researchers approach madness from a clinical rather than a critical or antioppressive stance. Such practices have often led to a divisive 'us' and 'them' mentality in social work where 'we,' the rational, well, social work practitioners/educators decide on and distance ourselves from 'them,' the irrational, 'ill' users of 'mental health'/social work services. Indeed, Kalinowski and Risser (2005) detail that this splitting ensures a lower standard of service for users of mental health services, and it promotes the unjust attitude that any medication/residence/job/education that is substandard for social workers is fine for users. It has also resulted in what Buckles and colleagues (2008) call "an unwillingness to learn about or adopt terms such as consumer, survivor or Mad, a refusal to include professionals with psychiatric disabilities and finally, complacency with regard to advocacy and responsibility for public education" (p. 259).

Additionally, these lenses allow social work to see that "like other liberation struggles of oppressed people, the activism of former psychiatric patients has been frequently ignored and discredited" (Chamberlin, 1990: p. 1). Seldom do social workers learn that, since the late 1960s, people with mental health issues have been organizing and questioning medical treatments, rights abuses, and forced institutionalization (Rheaume, 2002), and that the struggle against sanist assumptions was seen as best accomplished in groups composed exclusively of people with lived experience (Chamberlin, 1990). Rare is the acknowledgment that many within the Mad Movement will not 'fix and fit' themselves. Additionally, seldom are social workers aware that consciousness raising has developed into the international Mad Pride movement, challenging conceptions of normal and 'well' while celebrating madness through writing, research, organizing as well as annual festivals, marches, lectures, and art shows around the world (see www.madprideto.com).

Responses to Sanism in Social Work

Some suggest that the next step in responding to sanism in social work is by involving the Mad community in social work education, practice, and research, especially in community-based and participatory ways. In one such foray (Poole et al., 2012; Reid and Poole, 2013), research participants identified a dearth of alternative or nonmedical course options on mental health in social work and ongoing institutional resistance to requested accommodations such as extended time to complete assignments. There were also multiple instances of how madness has been languaged in the classroom such as the indiscriminate use of the term 'crazy' and fears that self-identified Mad students may be a danger to the school and/or others.

To combat this kind of sanism, arguments have been made to shift the discourse swirling around mental health in social work from 'danger' to discrimination; from 'fears of a threat' to awareness of a benefit to the 'profession,' and from talk of physical and chemical restraints to one of rights and accessibility. Arguments have also been made to ground a social work response to sanism in rights work at the local, national, and international levels. If social work were to align with advocacy bodies such as the Ontario Human Rights Commission in Canada, social work would shift its focus to sanism, knowing that sanism continues to be deeply embedded in institutions including education. Social work would understand that an example of this may be the common practice of mandatory social work student placements on specific days and at specific times of the week-days that may conflict with the early morning affects of certain medications. Social work would also acknowledge that sanism takes place when policies, procedures, requirements, eligibility criteria, or qualifications may appear neutral but amount to discrimination such as the common practice of awarding scholarships only to students in full-time attendance, effectively excluding those Mad students who may only attend on a part-time basis.

Toward Antisanist Practice

At present, there is no established literature on anti-sanist practice. However, drawing on the aforementioned critical approaches as well as the scholarship on sanism, anti-racism, and anti-oppression, anti-sanism may be the way forward.

In social science education, such an anti-sanist approach may involve more support for critical mental health courses and those that detail the Mad Movement. It may include a shift toward offering alternatives to courses that privilege medicalized 'mental health' practices and heavy penalties for sanist language use in publishing, teaching, and policy. In social science research, an anti-sanist approach may focus not only on inclusion but direction by individuals with mental health histories. It may mean a shift away from stigma research and toward sanism scholarship and more support for Mad Studies as a developing field. It may also mean the examination of sanist myths held by research ethics boards concerned that Mad scholars are unfit to carry out their own inquiries.

Additionally, in practice, and specifically social work, anti-sanist practice may mean a dramatic reworking of how Mad people are supported. It may mean an acknowledgment of sanist myths at work in our own lives, how medicalization is sanist, and a refusal to separate or segregate madness from its intersection with race, gender, and sexuality. Anti-sanist practice may also mean that professionals are allowed to own their own madness without fear of retribution.

In summary, this short review of mentalism and sanism demonstrates that although the terms were coined decades ago, social workers and social scientists have been slow to take them up or apply them in their work, an exclusion which is itself sanist. However, dedicated scholarship by Perlin, members of the ex-patients, survivor and Mad Movements, as well as those seeking to go beyond stigma demonstrate that sanism in particular has more currency than ever before. With respect to social work, there is still much that is sanist in its education, practice, and research, and much promise exists in the possibility of antisanist practice.

See also: Mental Health; Normality and Mental Health; Unemployment and Mental Health.

Bibliography

Barnes, C., 1998. The social model of disability: a sociological phenomenon ignored by sociologists. The Disability Reader: Social Science Perspectives, 65–78.

Birnbaum, M., 1960. The right to treatment. American Bar Association Journal 46, 499.

Buckles, B., Brewer, E., Kerecman, J., Mildred, L., Ellis, A., Ryan, J., 2008. Beyond stigma and discrimination: challenges for social work practice in psychiatric rehabilitation and recovery. Journal of Social Work in Disability and Rehabilitation 7 (3/4), 232–283.

Chamberlin, J., 1990. The ex-patients' movement: where we've been and where we're going. Challenging the Therapeutic State – Special Issue 11 (3), 323–336.

Coalition Against Sanist Attitudes (CASA), 2011. Who Coined Sanism and What Is Sanism?. Retrieved on 15.02.11 from: http://www.freewebs.com/lindsaycasa/sanism.htm.

Corrigan, P., Shapiro, J.R., 2010. Measuring the impact of programs that challenge the public stigma of mental illness. Clinical Psychology Review 30 (8), 907–922.

Crenshaw, K.W., 1989. Demarginalizing the intersection of race and sex: a black feminist critique of anti-discrimination doctrine, feminist theory and anti-racist politics. University of Chicago Legal Forum 140, 139–167.

Diamond, S., 2012. Against the Medicalization of Humanity a Critical Ethnography of a Community Trying to Build a World Free of Sanism and Psychiatric Oppression (doctoral dissertation). University of Toronto.

Dominelli, L., 2002. Anti-oppressive Social Work: Theory and Practice. Palgrave Macmillan, London.

Fabris, E., 2012. Experiences Labelled Psychotic: A Settler's Autoethnography beyond Psychosic Narrative (doctoral dissertation). University of Toronto.

Goffman, I., 1963. Stigma. Spectrum, Englewood Cliffs, NJ.

Hamer, H.P., 2011. Inside the City Walls: Mental Health Service Users' Journeys towards Full Citizenship (Doctoral dissertation). University of Auckland.

Healy, K., 2005. Social Work Theories in Context: Creating Frameworks for Practice. Palgrave MacMillan, London.

Holley, L.C., Stromwall, L.K., Bashor, K.H., 2012. Reconceptualizing stigma toward a critical anti-oppression paradigm. Stigma Research and Action 2 (2), 51–61.

Ingram, R., 2011. Sanism in Theory and Practice. Presentation for Second Annual Critical Inquiries Workshop. British Columbia, Vancouver.

Kalinowski, C., Risser, P., 2005. Identifying and Overcoming Mentalism. InforMed Health Publishing & Training. Retrieved on 10.03.10 from: http://www.newmediaexplorer.org/sepp/Mentalism (pdf).

Large, M., Ryan, C.J., 2012. Sanism, stigma and the belief in dangerousness. Australian New Zealand Journal of Psychiatry 46 (11), 1099–1103.

Lefrancois, B., 2011a. Queering child and adolescent mental health services: the subversion of heteronormativity in practice. Children & Society. http://dx.doi.org/10.1111/j.1099-0860.2011.00371.x.

LeFrancois, B., 2011b. Supporting positive relationships in families where fathers suffer with mental distress. Critical Social Work 12 (1), 2–12.

LeFrançois, B.A., Menzies, B., Reaume, G. (Eds.), 2013. Mad Matters: A Critical Reader in Canadian Mad Studies. Canadian Scholars Press, Toronto.

Link, M.G., Phelan, J.C., 2001. Conceptualizing stigma. Annual Review of Sociology 27 (1), 363–385.

Mad Pride Toronto, 2013. A Celebration of Madness. Retrieved on 08.07.13 from: http://www.madprideto.com.

Perlin, M.L., 1992. On "sanism". Southern Methodist University Law Review 46, 373–407.

Poole, J.M., Jivraj, T., Arslanian, A., Bellows, K., Chiasson, S., Hakimy, H., Reid, J., 2012. Sanism, 'mental health', and social work/education: a review and call to action. Intersectionalities: A Global Journal of Social Work Analysis, Research, Polity, and Practice 1, 20–36.

Reid, J., Poole, J., 2013. Mad students in the social work classroom? Notes from the beginnings of an inquiry. Special Issue of the Journal of Progressive Human Services on the 'Experiences of Mental Health Consumers/survivors' 24 (3), 209–222.

Rheaume, J., 2002. Lunatic to patient to person: nomenclature in psychiatric history and the influence of patients' activist in North America. International Journal of Law and Psychiatry 25, 405–426.

Sayce, L., 1998. Stigma, discrimination and social exclusion: what's in a word? Journal of Mental Health 7 (4), 331–343.

Thornicroft, G., 2006. Actions Speak Louder … Tackling Discrimination against People with Mental Illness. Mental Health Foundation, London.

Mental Illness, Etiology of

Peter Tyrer, Centre for Mental Health, Imperial College, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mental illness rarely has a single cause. Most episodes of mental disorder are created by a combination of (1) genetic factors, (2) nongenetic constitutional factors that are largely formed through early life experiences, (3) later life experiences that may predispose the individual to mental illness (sometimes called vulnerability factors), and (4) recent life events that can trigger the timing of a mental illness. It is difficult, but not impossible, to determine the relevant proportions of each of these in an individual subject, and this is complicated by gene–environment interactions, which may also precipitate mental illness.

Mental illness, like most types of illness, is multifactorial in origin. It can be influenced by constitutional and genetic factors, early and recent environmental factors, and cultural and social elements. These are best described in chronological format, with those more distant described first and the most recent described last.

Early Causes

Genetic Influences

The advances made by genetics in the last twenty-five years of the twentieth century have been enormous, and are discussed elsewhere (see, e.g., Genetic Studies of Behavior: Methodology; Genetics of Complex Traits Through the Life Cycle; Mental Illness, Genetics of). It may appear that few of these advances have yet had much impact on the causation of mental illness, because there have been few dramatic advances comparable with that of the identification of the gene for Huntington's chorea in 1987 (Gilliam et al., 1987). However, there has been a steady growth in knowledge, best summarized as incremental mapping, which has helped to place the genetic contribution to a wide range of mental illness, in context.

The standard methods of studying the genetic component of causation of illness are the same in psychiatry as in other branches of medicine. The relative genetic risk of having, or developing, a particular mental disorder is determined by three research approaches: family, twin, and adoption studies. Family studies identify people with the disorder through questions (probands); the rates of disorder are identified in other family members and relatives, and are then compared with the rates in the general population. From these figures, it is possible to compute the numbers of people likely to develop the condition at some time in the future (the expectancy rate or morbid risk). A higher rate in immediate relatives compared with others and greater rates than the general population indicate a genetic contribution to the disorder.

However, environmental factors may also contribute to risk in family studies, and twin studies are better at distinguishing the genetic component. Monozygotic (MZ) and dizygotic (DZ) twin probands are compared with their co-twins for concordance rates (i.e., the rate of co-occurrence of the disorder in the other twin). If the ratio of the concordance rates for MZ twins (who are similar to clones as they have identical genes) and DZ twins is significantly greater than one, the disorder has a genetic component. Because there is often a bias in reporting disorder, it is preferable to derive probands and their co-twins from twin registers wherever possible (Puri and Tyrer, 1998).

Even with twin studies, the effect of environment cannot be completely separated since there is a tendency, fortunately exhibited less now than formerly, for MZ twins to be treated as though they were the same person and to identify psychologically with each other to a greater extent than DZ twins. Adoption studies overcome this effect, and when they are carried out with MZ twins, are seen to their best advantage. Other studies include adoptee approaches in which one or both of the biological parents have the disorder but the adoptive parents have not, so the rates of disorder in the adoptees can be compared with a control population, and adoptee family studies in which the adoptee is the proband and the rate of disorder determined in the biological and adoptive parents.

The results of these studies, in general, show that there is a strong genetic contribution to bipolar mood disorder and schizophrenia with MZ/DZ ratios of greater than three (Hauge et al., 1968; Gottesman and Shields, 1976), but this falls far short of that shown by a disorder such as Huntington's chorea. A more common finding is that illustrated by personality disorder, in which twin studies suggest that most personality traits are equally contributed to by genetic and environmental factors (Livesley et al., 1998) and several models have been postulated for the interaction between them (Kendler and Eaves, 1986). Now that the human genome has been identified, the techniques of molecular genetics are being used increasingly to identify specific genes associated with psychiatric disorder, but to date, only one clear disorder has been identified, fragile X syndrome, which affects males mainly and has clear locus on the X chromosome (Sutherland, 1979). However, a range of other disorders has been shown to be associated with chromosomal abnormalities, and approximately 40% of the causes of severe mental handicap (Down's syndrome, phenylketonuria, tuberous sclerosis, Hurler syndrome, Lesch–Nyhan syndrome, and Tay–Sachs disease) are known to be caused by identifiable chromosomal abnormalities that are potentially preventable by genetic counseling (Weatherall, 1991; *see* Mental Illness, Genetics of). Genetics is also helping greatly in the understanding of common mental illnesses. In the last 40 years, there has been a great tendency to split mental disorders into smaller and smaller groups but the

genetic evidence suggests that they could be lumped into much larger components, with bipolar disorder and schizophrenia, showing much more commonalities than differences (Craddock and Owen, 2005), and the same applying to anxiety and depression (Kendler et al., 1995).

Environmental Factors before and during Birth

The time immediately before, during, and shortly after birth is a very vulnerable one for the fetus or newborn infant. It is at these times that anoxia can lead to brain damage. The most obvious manifestation of this is the loss of brain cells leading to learning disability in adult life, and better obstetric care and supervision of babies with low birth weight reduces this risk (Illsley and Mitchell, 1984). There have also been a number of studies suggesting that the development of other mental illnesses, particularly schizophrenia and related disorders, is linked to obstetric difficulties (Geddes and Lawrie, 1995), but there is some doubt about the methodology of some of these studies. More careful matching of cases and controls in a larger sample, however, has failed to replicate these findings, apart from prolonged labor and emergency Caesarian section being more common in schizophrenic probands (Kendell et al., 2000), and there is also increasing evidence that neurodevelopmental processes in intellectual disability and schizophrenia are related (Owen, 2012).

There is also evidence that infectious diseases, particularly when contracted before birth, may be associated with mental disorder. Schizophrenia is more common in those born in the winter months (Hare et al., 1973) and the case has been made, not altogether satisfactorily, that this could be a consequence of maternal influenza in the winter months (Adams et al., 1993). Another infectious disease linked to mental illness is rubella, which is associated strongly with learning disability (due to microcephaly), if infection is contracted in the first trimester of pregnancy.

Early Years of Development

The hypothesis that mental illness is a consequence of problems in early development has been active for many years and is one of the fundamental components of psychoanalysis. This argues that psychological conflicts encountered in early development that remain unresolved are likely to surface later in life in distorted form that psychoanalysis can succeed in decoding. This is perhaps best encapsulated in the phrase 'give us a child to the age of five and let who will have him thereafter.' Another approach initiated by the work of John Bowlby is the possibility that the earliest stages of the infant's relationship with its mother can also lead to problems that arise again in adult life. This has led to the growth of attachment theory (Bowlby, 1988) and has also stimulated psychoanalytical approaches to treatment.

These difficulties are often not associated with overt psychological trauma, or, if they are, it is usually within the bounds of everyday experience. More severe trauma is associated with a different range of pathology. There are now many conditions that have been shown to be associated with severe psychological trauma in the early years of life, including borderline personality disorder, post-traumatic stress disorder, recurrent depressive episodes, a range of problems of sexual dysfunction, and also, in recent years, a strong link between early trauma and psychosis, including schizophrenia, in adult life (Read et al., 2005). Depressive illness in adult life is also more likely to occur in those who have suffered the loss of a parent through separation, divorce, or death (Harris et al., 1986). This is postulated to be due to increased vulnerability to adversity after adverse early experiences. The long gap between initial loss and the onset of depression has been studied by life events researchers, and the concept of 'brought forward time' has been introduced to compare it with the immediate response to a major life event.

Late Causes

Later Environmental Causes

There are a host of environmental causes of mental illness from late adolescence onward that, unlike the previous ones described earlier, are determined to a much greater extent by the individual concerned. The group of conditions showing this characteristic most prominently, are the substance misuse disorders, as absence of indulgence can never lead to disorder. Although there are likely to be genetic factors that influence the extent that people become addicted to a substance (Kendler et al., 1995), by far the most important factor is repeated exposure, and this is primarily an environmental issue. Alcohol is the most important of the substance misuse disorders, accounting for one-third of all mental disorders and having harmful effects on around 10% of the population per year (Kessler et al., 1994), a staggering figure however it is interpreted.

Important gene–environment interactions have been noted for depression, where those with the shorter allele of the serotonin transporter gene have been noted to be prone to depression in the context of adverse events (Caspi et al., 2003), and for antisocial behavior, with those having low activity for the monoamine oxidase A gene being more prone to antisocial behavior after abuse than those with greater gene expression (Fergusson et al., 2011).

Other important environmental causes are infections. These include viral, protozoal, and bacterial organisms, and most recently, prion infections, such as the human form of bovine spongiform encephalopathy, and result in mental disorder when they affect the brain. The acquired immune deficiency syndrome, neurosyphilis, and viral encephalitis are the most important. Trauma to the brain is also an important cause of personality change and functional handicap, including dementia. Whenever the oxygen supply to the brain is significantly impaired, there is the danger of neuronal death and consequent organic mental illness, primarily dementia. Similar problems can be created by metabolic disturbances but these are more likely to be reversible. Thus a range of medical illnesses: hypertension, atherosclerosis, renal failure, hepatic failure, thyrotoxicosis, myxedema, sarcoidoisis, autoimmune diseases, and Cushing's disease (or any medical illness treated with steroids) may all be associated with mental disorder, of which anxiety, depression, and dementia are the most common.

A more common environmental factor influencing the onset and nature of many common mental disorders is social deprivation. Epidemiological studies have consistently shown that almost all mental illnesses are more common in the people of lower than higher social classes (Hollingshead and Redlich, 1958; Kessler et al., 1994). The reasons for this are much more difficult to unravel, but include specific aspects of deprivation (poorer material circumstances – greater crime, traffic pollution, substandard housing, unemployment), less control over direction in life, and a higher rate of adverse life events (Boulton, 1998).

Recent Stress as a Precipitant of Mental Disorder

The notion that stress is the cause of much mental (and physical) disorder is a very old one. It has been researched heavily in recent years by those interested in the effect of life events on mental illness. Although some of the early causes of mental illness might be perceived as stressors that create mental illness, the general notion of stress as a recent precipitant is rather a distant one. Early experiences may create a vulnerability to mental illness that is then activated by a stressful experience. These are often combined in the stress–diathesis model of mental illness that places mental illness and stress on separate scales: low stress being associated with no mental illness, greater levels of stress creating illness in those who have a vulnerable diathesis, and very severe stress creating symptoms of mental illness in almost anybody.

There are three specific mental illnesses that are defined as stress related: acute stress disorders, adjustment reactions, and post-traumatic stress disorder. All of these present primarily with symptoms within the neurotic spectrum (e.g., anxiety, panic, depression, irritability, social withdrawal), and satisfy the common definition that they would not have occurred in the absence of a psychosocial stressor. Acute stress disorders can occur in almost anybody when under intolerable pressures (e.g., car accident) but resolve rapidly when the natural processes of restitution take place (usually within 24 h). Adjustment disorders are longer-lasting, either because the stress is persistent (e.g., stress in a relationship) or because it is more intense. Post-traumatic stress disorder is associated with stress that is beyond the range of normal experience (e.g., earthquake, observing homicide) and is associated with specific symptoms, including flashbacks of the traumatic event, nightmares, extreme avoidance of situations that evoke memories of the event as well as the range of symptoms shown in adjustment disorders with generally greater intensity (particularly depressive symptoms).

However, although these conditions are defined by the nature of the stressful experiences preceding them, there are many others in which stress is a provoking factor. These are specific codes for stressful events in both the *DSM* and *ICD* classifications (American Psychiatric Association, 2013; World Health Organization, 1992) and in a full diagnostic summary, the nature of any associated stresses should be noted. In research work, stresses are often quantified into life events, whose intensity depends both on their nature and the extent of their (contextual) threat (Brown and Harris, 1978) and there have been numerous studies that show that the onset of almost all mental illnesses is accompanied by a greater rate of life events than those without such illness. It is difficult to know to what extent, however, these could be regarded as causes of the illness in the true etiological sense, and it is more common to regard them as triggers or provoking factors. Thus, for example, it has been shown that exposure to high levels of critical expressed emotion in patients with schizophrenia is more likely to provoke relapse than if more normal expression of emotion is shown (Vaughn and Leff, 1976) and interventions to reduce such emotion in families may have a beneficial effect (Pharoah et al., 2000). Even with this evidence, it is not suggested that critical expressed emotion is the cause of schizophrenia; it is the specific combination of a specific type of stress and its relationship to a particular illness that makes its effects shown. This 'Achilles' heel' phenomenon, the activation of an illness by a particular combination of circumstances, is very common in mental illness, and gives support to the stress–diathesis model for many of the most common conditions encountered, particularly those in primary care.

The multifactorial nature of causation of mental illness is illustrated in the following case report, based on a real patient but with some important elements distorted to prevent identification.

Mrs A came as a refugee from a country in sub-Saharan Africa where three members of her family were killed in front of her by rebels involved in fighting a civil war. She had become withdrawn and depressed after this, but was persuaded to flee the country by relatives and came to another country (the UK), where at the time she was being seen clinically, her application to stay in the country was being considered. Her relatives were also concerned that at times she had become more seriously disturbed and believed that her mind had been taken over by alien beings that had entered her body after a native witch doctor in her country had cursed her.

On assessment, she was shy but cooperative and keen to please. Her family history revealed that her mother had been unwell after one of her children was born and believed her body had been possessed by spirits that were preventing her from looking after her baby. A niece had also killed herself by walking into a swamp when unwell and allegedly possessed. During her childhood, she did well at school but had times when she lost confidence in her own abilities and had always been more anxious than her brothers and sisters. In general, she had always had low self-esteem. Mental state examination revealed that she was persistently anxious with episodes of panic, particularly at night and when her husband was away from home (he worked variable hours). She also had vivid nightmares of her parents being killed in front of her by men wielding machetes, and was unable to have any sharp knives in her kitchen because they aroused such high levels of anxiety. She also had periods of quite deep depression when she felt that life was such a struggle that it would be easier to take her own life than continue to fight. Direct questioning about her alien experiences revealed she still believed that she had been cursed, as indeed had her whole family, and that twitching movements of her hands and arms indicated that she was still under external control. She was also very anxious that her application to remain in the country as a refugee might not be granted and kept asking the services to intervene on her behalf so she could stay.

A diagnosis of post-traumatic stress disorder, comorbid generalized anxiety disorder, and recurrent depressive episode was made, and an unspecified diagnosis in the schizophrenia group, probably persistent delusional disorder, was also made.

Considering the causes of this complex problem, which is representative of many others who present to psychiatric services, the following etiological factors were considered:

1. the family history of possession and being cursed, which could have a cultural explanation, but which might be better explained by a family history of schizophrenia, which Mrs A would be then more likely to have, because of its genetic component,
2. the extreme stress occasioned by seeing members of her family killed in front of her (one of the characteristic features of post-traumatic stress disorder),
3. her lifelong tendency of low self-esteem and tendency to be anxious, and
4. her current anxiety about the real possibility of being asked to leave the country if she did not achieve refugee status.

The etiology of these problems can be described in the form of a tree, in which the main trunk indicates her fundamental diatheses, her possible tendency to develop schizophrenia, and the likelihood that she has some of the features of personality disturbance in the anxious/fearful group (cluster C) (a mixture of genetic and constitutional factors). The branches include the relatively recent experiences of seeing her family killed, and the most recent, more characteristic of an adjustment disorder, and indicate the stress of waiting to hear whether she would be allowed to stay in the country.

In constructing this tree, we need to be aware that there is a natural tendency for people to attempt to find causal explanations for everything in life, and sometimes these may be quite wrong. However, most of them have face validity, because in the search to make sense of a chaotic world, success in finding explanations makes it more bearable and also easier to recall – a phenomenon nicely described as 'effort after meaning' by one of the earlier researchers on the subject (Bartlett, 1932). Sometimes it is more accurate to say that the cause of a particular mental illness is unknown (idiopathic) but it makes us more authoritative to pick on a trunk, branch, or leaf of the etiology tree and give a suitable explanation.

Summary

The perceptive reader will recognize from reading this article that there is a great deal of more ignorance than knowledge when it comes to summarizing the causes of mental illness. A major deficiency in our knowledge is the ability to detect specific mental illnesses in advance of their manifestation. A great deal of energy is currently going into those who have what is called 'at risk mental states' or prepsychotic conditions that are likely to progress to full expression without intervention. At present, we remain remarkably poor at predicting which of these will progress to psychosis if untreated, and those that will resolve spontaneously. Statistical evidence of risk is of very little use to the average clinician unless it is at a high level (i.e., over 50%) and we rarely get this degree of prediction in psychiatry.

See also: Bipolar (Manic-Depressive) Disorder; Depression, Pessimism, and Health; Depression; Differential Diagnosis in Psychiatry; Environmental Stress and Health; Gene–Environment Interactions in Health and Well-Being; Gene–Environment Interplay; Mental Disorders, Epidemiology of; Mental Health; Mental Illness, Genetics of; Post-Traumatic Stress Disorder; Post-Traumatic Stress in Social Work; Schizophrenia; Social Exclusion, Social Deprivation and Health; Social Integration, Social Networks, and Health; Stress, Coping and Health; Stress: Psychological Perspectives.

Bibliography

Adams, W., Kendell, R.E., Hare, E.H., Munk-Jorgensen, P., 1993. Epidemiological evidence that maternal influenza contributes to the aetiology of schizophrenia: an analysis of Scottish, English and Danish data. British Journal of Psychiatry 163, 522–534.

American Psychiatric Association (APA), 2013. DSM-5: Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Association, Washington, DC.

Bartlett, F.C., 1932. Remembering. Cambridge University Press, Cambridge, UK.

Birtchnell, J., 1972. The inter-relationship between social class, early parental death and mental illness. Psychological Medicine 2, 166–175.

Boulton, M., 1998. Sociology. In: Puri, B.K., Tyres, P.J. (Eds.), Sciences Basic to Psychiatry, second ed. Churchill Livingstone, Edinburgh, UK, pp. 327–335.

Bowlby, J., 1988. A Secure Base: Parent–Child Attachment and Healthy Human Development. Basic Books, New York.

Brown, G.W., Harris, T.O., 1978. Social Origins of Depression: A Study of Psychiatric Disorders in Women. Tavistock, London.

Caspi, A., Sugden, K., Moffitt, T.E., Taylor, A., Craig, I.W., Harrington, H., McClay, J., Mill, J., Martin, J., Braithwaite, A., Poulton, R., 2003. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science 301, 386–389.

Craddock, N., Owen, M.J., 2005. The beginning of the end for the Kraepelinian dichotomy. British Journal of Psychiatry 186, 364–366.

Fergusson, D.M., Boden, J.M., Horwood, L.J., Miller, A.L., Kennedy, M.A., 2011. MAOA, abuse exposure and antisocial behaviour: 30-year longitudinal study. British Journal of Psychiatry 198, 457–463.

Geddes, J.R., Lawrie, S.M., 1995. Obstetric complications and schizophrenia: a meta-analysis. British Journal of Psychiatry 167, 786–793.

Gilliam, T.C., Bucan, M., MacDonald, M.E., Zimmer, M., Haines, J.L., Cheng, S.V., Pohl, T.M., Meyers, R.H., Whaley, W.L., Allitto, B.A., 1987. A DNA segment encoding two genes very tightly linked to Huntington's disease. Science 238, 950–952.

Gottesman, I.I., Shields, J., 1976. Genetics of schizophrenia – critical review of recent adoption, twin and family studies – behavioral genetics perspectives. Schizophrenia Bulletin 2, 360–401.

Hamshere, M.L., Stergiakoulim, E., Langley, K., Martin, J., Holmans, P., Kent, L., Owen, M.J., Gill, M., Thapar, A., O'Donovan, M., Craddock, N., 2013. Shared polygenic contribution between childhood attention-deficit hyperactivity disorder and adult schizophrenia. British Journal of Psychiatry 203, 107–111.

Harris, T., Brown, C.W., Bifulco, A., 1986. Loss of parent in childhood and adult psychiatric disorder: the role of lack of adequate parental care. Psychological Medicine 16, 641–659.

Hauge, M., Harvard, B., Fischer, M., Gotlieb-Jewen, K., Juel-Nielsen, N., Raebild, I., Shapiro, R., Videbeck, T., 1968. The Danish twin register. Acta Geneticae Medicae et Gemellologiae 17, 315–332.

Hare, E.H., Price, J.S., Slater, E.T.O., 1973. Mental disorder and season of birth. Nature 241, 480.

Hollingshead, A.B., Redlich, F.C., 1958. Social Class and Mental Illness: A Community Study. Wiley, New York.

Illsley, R., Mitchell, R.J., 1984. Low Birth Weight: A Medical, Psychological, and Social Study. Wiley, Chichester, UK.

Kendell, R.E., McInnenny, K., Juszczak, E., Bain, M., 2000. Obstetric complications and schizophrenia: two case-control studies based on structured obstetric records. British Journal of Psychiatry 176, 516–522.

Kendler, K.S., Eaves, L.J., 1986. Models for the joint effects of genotype and environment on liability to psychiatric illness. American Journal of Psychiatry 143, 279–289.

Kendler, K.S., Walters, E.E., Neale, M.C., Kessler, R.C., Heath, A.C., Eaves, L.J., 1995. The structure of the genetic and environmental risk factors for six major psychiatric disorders in women. Phobia, generalized anxiety disorder, panic disorder, bulimia, major depression, and alcoholism. Archives of General Psychiatry 52, 374–383.

Kessler, R.C., McGonagle, K.A., Zhao, S., Nelson, C.B., Hughes, M., Eshleman, S., Wittchen, H.U., Kendler, K.S., 1994. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. Archives of General Psychiatry 51, 8–19.

Livesley, W.J., Jang, K.L., Vernon, P.A., 1998. Phenotypic and genetic structure of traits delineating personality disorder. Archives of General Psychiatry 55, 941–948.

Owen, M.J., 2012. Intellectual disability and major psychiatric disorders: a continuum of neurodevelopmental causality. British Journal of Psychiatry 200, 268–269.

Pharoah, F.M., Mari, J.J., Streiner, D., 2000. Family Intervention for Schizophrenia (Cochrane Review). Cochrane Database of Systematic Reviews, 2. Update Software. Cochrane Library Issue 2, Oxford, UK.

Puri, B., Tyrer, P., 1998. Sciences Basic to Psychiatry, second ed. Churchill Livingstone, Edinburgh, UK.

Read, J., van Os, J., Morrison, A.P., Ross, C.A., 2005. Childhood trauma, psychosis and schizophrenia: a literature review with theoretical and clinical implications. Acta Psychiatrica Scandinavica 112, 330–350.

Sutherland, G.R., 1979. Heritable fragile sites on human chromosomes II. Distribution, phenotypic effects and cytogenetics. Human Genetics 53, 136–148.

Vaughn, C., Leff, J., 1976. The measurement of expressed emotion in the families of psychiatric patients. British Journal of Social and Clinical Psychology 15, 157–165.

Weatherall, D.J., 1991. The New Genetics and Clinical Practice. Oxford University Press, Oxford, UK.

World Health Organization (WHO), 1992. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines. World Health Organisation, Geneva, Switzerland.

Mental Illness, Genetics of

James J Rucker, The Institute of Psychiatry, Kings College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W. Maier, volume 14, pp. 9669–9675, © 2001, Elsevier Ltd.

Abstract

Common mental disorders are seen to run in families. However the relative contribution of genetics and the environment is uncertain. While rare forms of learning disability, autism, and Alzheimer's dementia have been found to have single-gene causes, the vast majority of cases of common mental disorders, such as schizophrenia and bipolar disorder, exhibit a highly polygenic, complex genetic etiology with gene–environment interaction the rule rather than the exception. New genetic technologies and the completion of the human genome project have led to an explosion of new molecular genetic data; however, the relatively small contribution of each variant to relative risk for common mental disorders has meant that progress in identifying specific associated genetic variants has been slow. It appears highly likely that we all carry a degree of liability to develop common mental disorders, and that this liability is mostly explained by a large repertoire of evolutionarily stable common genetic variants. This concept of a broad liability throughout the population both humanizes these previously stigmatized conditions and also explains why the population prevalence of these disorders remains constant over time, despite the reduced fecundity conferred.

All common mental disorders (affective, psychotic, and anxiety disorders and addiction) aggregate in families due to genetic influence. The clinical phenotypes are heterogeneous without distinct boundaries. For each common mental disorder genome-wide association studies propose the involvement of multiple genes in addition to environmental factors. Modern molecular genetic techniques greatly facilitate the search for these susceptibility genes, although distinguishing contributory genomic variation from the statistical 'noise' of multiple association testing has proven challenging, requiring very large sample sizes.

Psychotic and affective disturbances were first described in the philosophical literature of ancient Greece under the concepts of mania and melancholia, which were considered as extremes of personality traits. In this epoch the familial inheritance of these disturbances had already been noted. This wisdom continued until the beginning of the nineteenth century, when these disturbances were first considered as medical diseases. During this century the new concept of degeneration stressed the accumulating effects of familial inheritance across generations and was applied to psychotic and emotional disorders. At the end of the nineteenth century alcoholism was additionally recognized as a major source of degeneration. Although the classical concepts of inheritance and degeneration did not specify the mechanisms of transgenerational transmission, both concepts prepared the basis for the eugenics movement starting in Anglo-Saxon countries. This ideology proposed methods of reproductive planning for increasing the health and welfare of humankind and for reducing the impact of diseases, criminality, and bad habits.

Twentieth Century Developments

Gregor Mendel was the first to postulate specific mechanisms of genetic transmission in plants. He was fortunate that he was working on traits that happened to be due to monogenic variation (that is, one mutation causing one variable trait). The relevance of this work was not immediately recognized. The rediscovery of the Mendelian law and the invention of the concept of genes in 1900 led to rapidly growing interest in the genetic hypothesis and its application to humans. Remarkably, the original concept of genes coding for physical and psychological properties was hypothetical and was developed without any knowledge of the physical existence or structure of genes (which became apparent only several decades later through the work of Watson and Crick). The early concept of genes already included the possibility that a gene could have multiple variants (mutations) with different functional consequences (polymorphism). Thus, the distinction between genetic and nongenetic influences became feasible. The familial transmission of patterns and disturbances of human behavior was also explored from a genetic perspective. The advantage of this was the development of systematic family and twin studies and their application to mental disorders, mainly in the 'Munich School' headed by E. Rüdin at the beginning of the twentieth century. Familial aggregation and genetic determination of most mental disorders was concluded.

These results were misinterpreted in the predominant 'one disease/one gene' perspective (monogenetic concept). This concept is valid mainly for rare genetic diseases, but not for common mental disorders. The eugenics movement in many countries was influenced by this misinterpretation, and served as the conceptual basis of programs to prohibit people affected from having children. Nazi Germany in particular established forced sterilization programs among patients with schizophrenia and affective disorders. This development is particularly surprising because it was noticed that many inherited behavioral traits did not conform to a Mendelian model of familial transmission. An etiology proposing many contributory genes (polygenic) to explain familial aggregation without Mendelian patterns of transmission was initially proposed in a paper by the British statistician Sir Ronald Fisher in 1918 (Fisher, 1918). However leading psychiatric geneticists

(e.g., E. Rüdin in Germany) appeared to be ignorant of these insights and continued to promulgate the monogenic, Mendelian origin of schizophrenia and manic-depressive illness. From the perspective of a polygenic etiology, the rationale of the eugenic movement becomes largely invalid, as prevention of affected groups from reproducing is unlikely to change the frequency of contributory genetic variants in the population.

Despite excellent empirical work, the genetic origin of mental disorders hypothesis did not receive much attention between the 1940s and 1970s. Several factors explain this lack of interest: first, the genetic hypothesis was seriously discredited by the morally disreputable eugenics movement focusing on these diseases; second, psychoanalytic thinking predominated in these decades in many countries (especially in the USA, until about 1970); third, the research tools for studying genetically influenced disorders without a Mendelian mode of transmission were very limited before the molecular-genetics era; and finally, tools for understanding complex interactions of genetic and environmental influences were more or less nonexistent in these times. Given this lack of genetic research potential the environmental causes of mental disorders were probably overemphasised.

Motivated by the emergence of biological psychiatry during the 1970s, family-genetic approaches to the etiology of mental disorders received growing attention. Family, twin, and adoption studies using newly developed epidemiological and biostatistical techniques were performed, exploring the etiology of, and the relationship between, different mental disorders. Particularly, the landmark Danish adoption studies by Kety and Goodwin changed the predominant assumption of the psychosocial origin of psychoses and alcoholism and motivated a nature versus nurture discussion that dominated the field for a decade. A refined arsenal of genetic–epidemiological methods was used to demonstrate the interaction between environmental and genetic factors as the etiological basis of common mental disorders. Gene–environment interactions appeared to be the rule rather than the exception. The results strongly argued against the oversimplification of the nature versus nurture debate.

Based on the discovery of the biochemical structure of chromosomes as long strings of DNA, and the process of gene expression as the production of functionally relevant proteins, the polymorphic nature of genes and the functional consequences thereof became apparent in the middle of the twentieth century. Genetic variability was found to occur as different variants of the gene itself (termed an allele), which might result in different gene products (proteins), or as variation in the promoter regions of genes, which influenced the degree of gene expression and consequently the quantity of the protein product. Within this framework, genetic causes of diseases could be identified at the level of DNA and also the protein product. Genetics thus became a basic tool for unraveling the pathophysiology of diseases and for the understanding and development of effective treatments.

The rapidly developing field of molecular genetics shifted the focus from the nature–nurture debate to the search for the genetic origin of diseases at the level of DNA variation. Stimulated by progress in unraveling the causes of monogenic diseases, extensive efforts are now underway to identify genetic variants of etiological relevance.

In the meantime, the genetic causes of most monogenic diseases have been detected, especially those characterized by severe or global impairment of cognitive function. Causal genes for rare forms of early-onset dementia of Alzheimer type have also been identified. These discoveries, however, appear to represent the 'low-hanging fruit.' More common mental disorders, such as schizophrenia and bipolar affective disorder, are increasingly seen as highly polygenic disorders where a spectrum of preexisting genetic liability to disorder exists in a complex interaction with environment (Uher, 2009). The completion and refinement of the human genome project and the subsequent HapMap project to identify and make available a database of human genomic variation has presented a unique opportunity to resolve the relative contribution of different types of genetic variation to common mental disorders.

The development of DNA microarrays, which represent a miniaturization of the Southern blotting technique developed in 1975, has been a particularly important contribution to the exponential expansion of new molecular genetic data seen in the early part of the twenty-first century. DNA microarrays allow the interrogation of the genome at thousands, or even millions, of known genomic locations on a single glass microscope slide. This has allowed massively parallel genomic association studies to be carried out on large numbers of samples at a cost effective rate. Microarrays using single nucleotide polymorphism (SNP) markers capitalize on the fact that SNPs represent unique genomic markers of known position within the reference human genome sequence that have only two possible forms (known as 'alleles') which, in a normal chromosome comprised of two chromatids, results in three possible variants – often arbitrarily denoted as 'AA,' 'AB,' and 'BB.' On a DNA microarray, each allele is tagged with a different flurophore that fluoresces a different color (usually red or green) depending on the genetic variation present within the sample being analyzed. In addition to these variant-specific microarrays, whole-genome sequencing arrays have been developed that allow the vast majority of the genome to be sequenced, base-by-base. Given the recent development of still other arrays that are able to interrogate epigenetic variation such as DNA methylation and microRNAs, the amount of genetic data is expanding at an exponential rate. The problem now arising is how this vast quantity of biological data can be meaningfully analyzed in common mental disorders and what benefits, if any, can be conferred at the clinical level.

The problem of multiple association testing presents a particularly thorny problem if, as is now expected, most of the genetic explanation for liability to common mental disorders arises from many genetic variants each contributing only a small increase in relative risk. Distinguishing true causal variants from the statistical 'noise' generated by multiple association tests may require sample sizes in the region of hundreds of thousands, or even millions, of samples. Rare variation may contribute substantially to risk in a small minority of cases, while a collection of common variants contribute toward risk in other cases, with still more cases having a contribution from both forms of variation. The picture in the majority of cases, it appears, will be very complicated.

Among the common mental disorders, schizophrenia and the affective disorders have been most intensively studied. The genetics of rare mental disorders, particularly early-onset

Alzheimer's disease and specific subtypes of mental retardation, have been extensively and very successfully studied; given the limitations of space these rare disorders are not included in this overview. The evidence related to epigenetic data, which will focus on analyses restricted to the DNA sequence itself, is still in its infancy at the time of writing and is also not covered in this review.

Specific Disorders

Schizophrenia

Among the common mental disorders, as mentioned in the preceding section, schizophrenia and the related psychoses have been most intensively studied from a genetic perspective. Schizophrenia is a disabling, lifelong disorder that can manifest at any age (lifetime prevalence 0.5–1%) but which is most often seen to develop in the second and third decades of life. The cardinal symptom is lack of insight, although delusions, hallucinations, thought disorder, and an amotivational or dysexecutive syndrome that results in disabling psychosocial decline and dysfunction (often called the 'negative' or 'deficit' syndrome) are also common. Schizophrenia is a very heterogeneous disorder, with no one symptom or sign being seen in all cases. Retrospectively it is often possible to identify subtle developmental problems in childhood, although these by no means predict the onset of the disorder. An extensive body of evidence suggests that schizophrenia represents a broad range of neurocognitive problems, which are themselves the result of subtle neurodevelopmental abnormalities resulting from a complex interaction between a polygenic liability to disease and precipitating and perpetuating environmental factors.

The familial basis of schizophrenia was first established early in the twentieth century. The risk to siblings of affected probands (proband being defined as the primary affected case in a family, or pedigree) is about 5–7%, resulting in a relative risk (lifetime risk in the group of interest divided by lifetime risk in the general population) of about 10. The risk among parents of probands with schizophrenia is reduced (3–5%) because schizophrenia confers a reduced fecundity (i.e., affected subjects have fewer children than random subjects in the general population). Currently, having a first-degree relative with schizophrenia shows the highest predictive power among all known risk factors for schizophrenia (*see* Schizophrenia).

Twin Studies

Twin studies explore the genetic impact on disease by utilizing the fact that monozygotic twins share 100% of their genes, while dizygotic twins share only 50%. By comparing the concordance rate among monozygotic and dizygotic twins, an estimate of the relative contribution of genes and environment can be made. A higher concordance rate for schizophrenia is consistently found in monozygotic compared to dizygotic twins (50% vs 10%). Estimates vary; however, this suggests a substantial genetic component (or heritability) of between 60 and 80%, but it also argues for the impact of non-genetic environmental factors (because if the disorder was entirely genetic in origin, the concordance rate between monozygotic twins would be 100%). By combining twin data with population prevalence rates, it is possible to distinguish three sources of etiological variance – genetic, non-genetic familial environment, and non-genetic individual-specific environment. There have been different conclusions from twin compared to adoption studies with regard to the relevance of familial environment for the emergence of schizophrenia; while this problem is unresolved, the relevance of strong genetic influences is unquestionable. Similarly, it is unquestionable that the inheritance patterns seen in schizophrenia are not Mendelian, but rather complex and reflective of a broad polygenic contribution.

Schizophrenia shares a series of features with other complex disorders:

1. Environmental factors operate in concert with genetic factors (evidenced by less than 100% concordance among monozygotic twins).
2. Penetrance of the phenotype is incomplete: the offspring of an unaffected identical twin of a proband (discordant monozygotic twins) have the same risk (10%) of transmitting the disorder to their children as monozygotic twins where both are affected.
3. The boundaries of the transmitted familial phenotype of schizophrenia are not distinct. Related syndromes and isolated symptoms are aggregative. That is, among the relatives of probands, characteristics of schizophrenia (that do not amount to a diagnosis) occur more often than expected by chance (e.g., attention deficit, abnormal patterns of neural evoked potentials, memory problems, and disturbances in pursuit eye movement) (Tsuang, 2000).

Molecular Genetics

The combination of the strong support for a genetic influence on schizophrenia from twin studies, the completion of the human genome project, and the ongoing cataloging of human genetic variation via the HapMap project, along with rapid advances in miniaturization of molecular genetic techniques, has resulted in a huge drive to discover the genetic basis of liability to the disorder.

Prior to the completion of the human genome project, the genetic linkage technique was used to probe the genome for regions linked to schizophrenia. This technique, enormously successful in identifying genomic regions associated with monogenic diseases was generally unsuccessful in schizophrenia (and many other complex disorders) largely because it can only detect variants of moderate or large effect. To this extent, however, a few families with highly transmissible (and very rare) forms of schizophrenia were identified, most notably a family in Scotland with a translocation between chromosomes 1 and 11, disrupting a gene that went on to be called *DISC1* (disrupted in schizophrenia 1), and now known to be involved with the regulation of neuronal axon and dendrite outgrowth. It is prescient and pertinent to note in this study that the clinical picture seen in affected individuals included psychiatric diagnoses other than schizophrenia.

The enthusiasm for genetic linkage at the time was largely misplaced, and represented not only the views of those who still believed that schizophrenia was caused by one or a few genes, but also because no other better technique for probing the genome in multiple locations existed at the time.

To this extent it was the development of the genomic microarray that revolutionized complex disease genetics in the early part of the twenty-first century. The microarray allows interrogation of the genome at vast numbers of unique genomic locations on a single glass slide and is far better suited to detect variants that may be contributing a relatively small amount of risk. The microarray is responsible for the initial explosion of new genetic data in complex disorders and the birth of the genome wide association study (GWAS). However with these new studies and new data came a difficult statistical problem – how to distinguish between a true variant of small effect and a variant that does not contribute to risk but is seen to be associated with disorder by chance because of the number of association tests being carried out (a statistical 'type 1 error'). The natural solution to such a problem is to set a statistical significance level that takes account of chance association in the thousands, or millions, of statistical tests being undertaken (widely cited in GWAS as the probability of a type 1 error of less than 5×10^{-8}). However in early GWAS studies, which did not utilize high sample numbers, this probably resulted in variants of small effect being lost within the statistical 'noise' generated by multiple association testing (a statistical 'type 2 error'). This is a problem of lack of statistical power, or the ability to detect an effect if one exists. The most obvious solution is to increase the number of case and control samples, however the mathematics involved revealed that the sample sizes of the early GWAS were insufficient to the extent of one, two, or even three orders of magnitude to detect an effect. This led to an all-to-familiar story of excitement over initial findings being quelled with disappointment over non-replication. This has led to an intensive global collaboration to combine sample sets and achieve adequate statistical power to detect variants with modest effects on relative risk, most notably resulting in the formation of the Psychiatric GWAS Consortium.

The largest and most technically proficient GWAS performed so far supports the theory of a polygenic contribution to liability for schizophrenia (Ripke et al., 2013). It is important to note that there still exists considerable variation between studies, and replication rates have, on the whole, been poor. The picture is in constant flux. In the Ripke et al. (2013) study cited above, eight replicated single nucleotide polymorphisms (SNPs) were associated with schizophrenia. Using approximate Bayesian polygenic analysis, the authors estimate that there are 8300 SNPs across the genome that contribute risk for schizophrenia, accounting for 50% of the total genetic contribution to liability. The most consistently replicated area of the genome contributing to liability to schizophrenia is the major histocompatibility complex on chromosome 6. This is a highly complex genomic area that is naturally prone to rearrangement, but it lends support to the notion that there may be an immunomodulatory component involved in the etiology of schizophrenia.

SNPs only represent one type of genetic variation (although associations seen with SNPs may 'tag' other types of variation). The role of copy number variants (CNVs) has also been investigated. CNVs refer to deletions or duplications of genomic material, usually less than 3 million base pairs in size (above this size, duplications and deletions of genomic material can usually be seen under the light of microscope in a standard cytogenetic analysis). Rare CNVs (occurring in less than 1 in 100 individuals) have been robustly associated with autism, hence attention has turned to the possible role these variants might have in schizophrenia. While the studies of copy number variation are in their infancy in comparison to SNPs, early evidence suggests that rare CNVs are seen at an increased frequency in schizophrenia, although the same variants are also seen in cases of autism, epilepsy, and learning disability, along with healthy controls (at a reduced frequency). This observation underlines the clinical heterogeneity that may result from observed genetic variation. This probably reflects both the conceptual 'distance' between observed genetic variation and clinical diagnostic categories and also the relative contribution of other genetic variants within a case that happens to contain a rare variant.

The most frequently observed CNV in cases of schizophrenia is a deletion of 3 million base pairs in the long arm of chromosome 22. This is an area of recent evolutionary rearrangement, and probably under high selection pressure. The deletion is also seen clinically in the velocardiofacial syndrome, DiGeorge syndrome, autism, affective disorders, and in the normal population. Again this underlines the clinical heterogeneity of any observed genetic variant and emphasizes the concept of genetic penetrance.

It can be concluded from this evidence that there is no single causal or major gene that explains the genetic liability to schizophrenia. Both rare and common variants contribute to genetic liability, which will vary between individuals. We probably all carry a degree of risk to developing schizophrenia, with this degree of risk approximately normally distributed among individuals in the population. In this sense, the disorder is a very human one, with the idea that we all carry a risk for the disorder serving, perhaps, as a basis for reducing stigmatization in mental disorders. Such an observation also means that the selective breeding policies employed by the eugenics movement would never have worked to breed out liability for schizophrenia (or any complex disorder), and also explains why the prevalence of schizophrenia has remained stable over time, even though sufferers show reduced fecundity. It also means that the underlying biology of the disorder is likely to be variable between individuals. This appears, at least partially, consistent with the clinical picture, in which no one sign or symptom is either necessary or sufficient to result in a diagnosis, and that the disorder can occur at any age. Similarly it explains why sub-threshold signs of the disorder are seen at an increased frequency in the relatives of probands, and why sporadic cases of schizophrenia are seen in families with no prior history of the disorder. Both common and rare variants contribute risk, but very few individuals will have genetic variants that are, individually, associated with any substantial risk for the disorder. The best example of a rare variant contributing to risk of schizophrenia is the 22q11 deletion, however even with this large deletion of many genes a spectrum of disorder exists, from normality to profound disability. It is likely that genetic variants act both individually and in concert to produce both a risk of, or a resilience to, any particular disorder, and the environment is also critical in determining whether this risk results in clinical manifestation of a diagnosis.

Affective Disorders

The affective disorders are characterized by single or recurrent bouts of clinically significant mania, depression, or both. Traditionally recurrent bouts of mania and an alternation between manic and depressive states have both been considered under the diagnosis of bipolar affective disorder (BPD) with single and recurrent depressive states without mania considered under the major depressive disorder (MDD) diagnosis. Manic states are characterized by lack of need for sleep, extreme irritability or elation, racing thoughts, pressure of speech and ideas, thought disorder, disinhibited and socially inappropriate behavior, grandiose delusions, and lack of insight. Depressive states are characterized by low mood, anhedonia, disturbed sleep and appetite, low energy levels with amotivation and psychomotor retardation, poor concentration and decision-making abilities, a strongly negative cognitive bias including suicidal thoughts and acts, prominent comorbid anxiety, and, in severe cases, depressive delusions and hallucinations. Insight is usually preserved in all but the most severe depressive states. Mixed affective states occasionally occur, as does a syndrome of recurrent distinct psychotic and affective episodes, which is often classified as schizoaffective disorder. The lifetime prevalence of BPD is approximately 0.5–1.5%, whereas this rises to estimates of between 4 and 17% for MDD, which is also seen more often in females. The wide range of estimates for lifetime prevalence in MDD is largely attributable to the clinical definition used, with milder forms of the condition much more prevalent than more severe types as well as having clear overlaps with the normal processes of grief, sadness, and loss.

Twin Studies

The familial aggregation of mood disorders has been noted for over 100 years; however, there are clear differences between BPD and MDD. Within formal family studies, the first-degree relatives of probands with MDD are about three times more likely to suffer from the disorder than controls, whereas this figure rises to about eight in BPD. Mean monozygotic concordance rates are 30–70% for unipolar depression and 50–80% for bipolar disorder. Dizygotic concordance rates are 10–30% for unipolar depression and 10–20% for bipolar disorder. The calculated heritability rates are 30–50% for unipolar depression and 50–80% for bipolar disorder, the figure calculated being highly dependent on the severity of sample analyzed, but even in the severest samples leaving considerable space for environmental risk factors (Sullivan et al., 2000). A graduation of risk is seen as genetic distance from the proband increases; however, the etiological and clinical heterogeneity of affective illness probably precludes estimates that are generalizable to a large proportion of sufferers. This can confound attempts to combine different samples of affective disorders together to study the genetic basis of liability to disorder, as even more clinical heterogeneity is introduced into an already notoriously heterogeneous group.

Interestingly, the familiality of affective illness is not bidirectional. The relatives of probands with BPD have a much higher risk of developing both BPD and MDD. However the relatives of probands with MDD do not carry a higher risk of developing BPD. This argues for a partial genetic overlap between MDD and BPD, but suggests that BPD in particular has a proportion of unique genetic liability. The picture is further complicated, by the fact that the relatives of probands with BPD have a higher chance of developing psychotic disorders such as schizophrenia, although not as high as relatives of probands with schizophrenia. Relatives with probands of schizophrenia have a higher chance of developing MDD, and the reverse also appears to be true (but to a much lesser extent). Put simply, there appears to be a complex overlapping genetic liability between clinical subtypes of mental disorders and there may be a common liability to mental disorder that transcends traditional diagnostic boundaries.

Molecular Genetics

Given the relatively high lifetime prevalence of unipolar depression in the general population, informative longitudinal studies exploring the relationship between different risk factors from various domains are feasible. A population-based prospective twin study (Kendler, 1990) proposed that genetic and environmental risk factors (such as early parental loss, perceived parental warmth in childhood and critical life events, and social support later on) interact. The risk factors occurring already in childhood also influence personality features (e.g., neuroticism) and coping strategies that also operate as risk factors for depression and mediate the early environmental and genetic sources for the final disease status. A particularly noteworthy example of a specific gene–environment interaction in MDD was demonstrated by Caspi et al. in a paper published in 2003 (Caspi et al., 2003). The results from the study suggested that those who carried a particular form of the promoter for the gene that encodes the serotonin reuptake transporter on presynaptic neurons (*SERT*) were more vulnerable to developing MDD in the presence of stressful life events, than those who had the other form of the promoter. A plausible result from many perspectives, this result has nonetheless been hotly debated for over 10 years, with different studies providing both replication and non-replication of the result. The evolving consensus suggests that the effect is probably real although the authors may have been subject to the so-called 'winner's curse' (the tendency toward overinflation of effect size in novel findings, for a variety of reasons).

Caspi et al.'s work is a good example of a 'candidate gene study.' Such studies start with a hypothesis (in this case that genetic variation in the serotonin transporter may moderate the liability to develop MDD in response to stressful life events) and tailor the experiment to provide data that will confirm or refute it. This is in particular contrast to GWAS, which are 'hypothesis free.' GWAS interrogate huge numbers of sites on the genome and test each for association with disease. Because they are 'genome-wide' experiments (i.e., testing variants from across the genome) they come with no prior assumption that any particular variant is more likely than any other to be associated. While this hypothesis-free approach is intuitively attractive because it frees research from the subjectively biased hypotheses of the researching team, statistical issues resulting from a lack of power to detect what are small individual effects of any one variant require sample sizes that are so huge that they test the limits of practicality.

To this extent, the results of GWAS in affective disorders have been less exciting than in psychotic disorders such as

schizophrenia. Again, this probably denotes the highly heterogeneous nature of the disorders, both at a clinical and biological level. In support of this, candidate gene studies in affective disorders have generally reported inconsistent results and linkage analysis was also inconclusive, with no replicated findings standing the test of time and scrutiny by the research community. GWAS in BPD have identified one SNP that replicates consistently, and others are showing signs of association as samples are combined into larger and larger datasets. The one robustly identified SNP is in the gene *CACNA1C*, which encodes a subunit of a major neuronal calcium channel found in the brain (Craddock and Sklar, 2013). This is of great interest, since calcium channel moderating drugs (such as the antiepileptic drug lamotrigine) are widely used to treat both BPD and MDD. This represents the first example of a plausible marriage between the clinical reality of the disorder and the underlying biology and genetics. However, while there was initial excitement over genome-wide significant findings in MDD, a recent mega-analysis with 9240 cases and 9519 controls, revealed no significant associations (Ripke et al., 2012). However even at this sample size, it is likely that the lack of findings is probably due to inadequate statistical power to detect an effect. In MDD, where the prevalence is relatively high and the clinical and genetic heterogeneity also high, sample sizes in the order of hundreds of thousands, or even millions, of individuals may be needed.

Conclusion

In 1967, Gottesman and Shields wrote, in an attempt to harmonize what, at that time, was a substantial split between the psychiatric and biometric view of schizophrenia:

> One of the possibilities not given sufficient attention involves positing a large proportion of cases as being polygenically determined. We should like to consider the merits of treating schizophrenia as a threshold character whose appearance is predictable from a diathesis-stress model …. Let us suppose that the diathesis is polygenically determined and that what is inherited is a constitutional predisposition to developing schizophrenia …. Although schizophrenia is necessarily viewed as an all-or-none character for record keeping purposes, clinical contact with preschizophrenics or 'recovered' cases shows clearly the artificiality of such a dichotomy …. Falconer has now made available a model and methods for the handling of disease incidences in the relatives of probands with threshold characters, i.e., diseases that appear to have an all-or-none manifestation but are in fact determined by an underlying graduation of some attribute really causing the disease. The latter attribute has been termed liability by Falconer and is intended to convey not only the individual's innate tendency to develop the disease (i.e., susceptibility), but also the environmental milieu to which he is exposed that makes him likely to develop the disease.
>
> Gottesman and Shields, 1967

The latest evidence suggests not only that their opinion was, broadly speaking, correct but that such an etiological framework may be applied to a broad range of common, complex disorders. As such, a reductionist approach probably has very little place in the consideration of the genetic basis for liability to psychiatric disorders. The replicable genetic associations with psychiatric disorders reported here represent the first steps on what is likely to be a long and difficult path to the understanding of etiology and pathogenesis. Effect sizes for individual variants are likely to be small, except for a small minority of cases that harbor rare variants that confer a relatively high degree of risk.

All common psychiatric disorders display a complex genetic heterogeneity, meaning that multiple partially overlapping sets of genes, themselves in a complex interaction with the environment, can predispose to similar syndromes that are currently indistinguishable on clinical grounds. There may be a common liability to psychiatric disorders in general, rather than individual disorders specifically. There still exists a vast conceptual gap between the clinical reality of common psychiatric disorders and the demonstrable, replicable science, but progress toward understanding the basis of genetic liability to these disorders is slowly being made. Whole genome sequencing efforts, which can detect still more variants that may contribute to risk or resilience are currently underway and firmer conclusions from this data should be available in the next decade or so. We should no longer expect the quest to uncover the genetic liability to complex disorders, especially mental disorders, to be simple.

In summary, the genetics of mental disorders has a long and sometimes controversial history. Currently, only rare variations of mental retardation and early onset Alzheimer's disease have been demonstrated to be classical genetic diseases with mutations at one gene locus causing the clinical syndrome (monogenic diseases). Most mental disorders are highly prevalent in the general population. Such common mental disorders share a series of features with common diseases such as hypertension, cardiac disease, and type 2 diabetes mellitus:

1. All common mental disorders are familial (i.e., relatives of patients are more likely to be affected with the same disorder than random subjects in the general population).
2. Twin studies performed for all common mental disorders point to a substantial genetic component, which is probably best conceptualized as a liability to develop the disorder in the presence of precipitating and perpetuating environmental factors.
3. The concordance rate among monozygotic twins is far from 100%; thus, the etiology is multifactorial with environmental as well as genetic factors contributing.
4. The phenotype as transmitted in families is variable between relatives; unlike monogenic diseases a clear distinction between affected and healthy individuals is impossible.
5. The familial aggregation of one disorder often goes together with the co-aggregation of another disorder (e.g., generalized anxiety disorder and depression); a common genetic liability may account for this co-aggregation in families.
6. The genetics of each common mental disorder is complex and associated with genetic heterogeneity; the genetics does not follow a clear Mendelian mode of transmission, but genome wide association studies argue for the contribution of multiple genes to each disorder. Consequently, the contribution of a single genetic variant is not causal, but only probabilistic (susceptibility genes influencing the risk of the disorder).

7. Multiple susceptibility genes for a disorder may either emerge from multiple monogenic, clinically indistinguishable subtypes or from the simultaneous and cumulative contribution of many genes. While the first possibility cannot be excluded definitively, the second possibility is substantially more plausible given the lack of Mendelian transmission. While the contributing genetic variants are likely to be rare under the first condition, the contrary should be the case under the second condition. Each of these common genetic variants is likely to be of ancient origin, dating back about 100 000 years. The functional consequence of each variant is presumably modest, which explains the persistence of the variants despite evolution by natural selection. In contrast, rare genetic variants causing monogenic diseases are more recent in origin, confer much reduced fecundity, and are therefore, presumably, under strong negative evolutionary selection pressure.
8. The first replicable glimpses of genetic variation associated with common, complex mental disorders have been established, but the picture is more complicated than many imagined and the technical challenges in unpicking the relative contribution of different genetic variants contributing to mental disorders are daunting. Slow but steady progress is being made, and advances in molecular genetic techniques over the past 20 years particularly have been spectacular. An explosion of new genetic data since the completion of the human genome project has led to both renewed hope and new challenges, in the field of psychiatric genetics.

See also: Behavioral Genetics: Psychological Perspectives; Brain Development: Ontogenetic Neurobiology of; Developmental Behavioral Genetics and Education; Epigenetic Inheritance; Epigenetics of Brain Plasticity and Behavior; Eugenics in Europe; Familial Studies: Genetic Inferences; Gene–Environment Interactions in Health and Well-Being; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Studies of Behavior: Methodology; Genetics, Disease, and Reproduction; Genetics: The New Genetics; Intelligence, Genetics of: Cognitive Abilities; Mental Disorders, Epidemiology of; Personality Differences and Development: Genetic and Environmental Contributions.

Bibliography

Caspi, A., et al., 2003. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science 301 (5631), 386–389.

Craddock, N., Sklar, P., 2013. Genetics of bipolar disorder. The Lancet 381 (9878), 1654–1662.

Fisher, R., 1918. The correlation between relatives on the supposition of mendelian inheritance. Transactions of the Royal Society of Edinburgh 52, 399–433.

Gottesman, I., Shields, J., 1967. A polygenic theory of schizophrenia. Proceedings of the National Academy of Sciences of the United States of America 58 (1), 199–205.

Kendler, K.S., 1990. Familial risk factors and the familial aggregation of psychiatric disorders. Psychological Medicine 20 (2), 311–319.

Ripke, S., et al., 2012. A mega-analysis of genome-wide association studies for major depressive disorder. Molecular Psychiatry 18 (4), 497–511.

Ripke, S., et al., 2013. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nature Publishing Group 45 (10), 1150–1159.

Sullivan, P.F., Neale, M.C., Kendler, K.S., 2000. Genetic epidemiology of major depression: review and meta-analysis. American Journal of Psychiatry 157 (10), 1552–1562.

Tsuang, M., 2000. Schizophrenia: genes and environment. Biological Psychiatry 47 (3), 210–220.

Uher, R., 2009. The role of genetic variation in the causation of mental illness: an evolution-informed framework. Molecular Psychiatry 14, 1072–1082.

Mental Illness: Family and Patient Organizations in Europe

JH Henderson, St Thomas' Hospital, London, UK
R Ramsay, Edinburgh, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9665–9669, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Organizations focusing on the concerns of users and families operate at a number of levels, global, national, and within individual countries. This article describes one regionally based organization, the European Network of (ex-)Users and Survivors of Psychiatry. Organizations work in a number of different ways, and have developed different models of linking with statutory services. The benefits of belonging to a family organization are reported, as well as the development of self-help groups.

Mental Health Policies and the Family

In November 1999 Dr Gro Harlem Brundtland, director general of the World Health Organization (WHO), launched the WHO's new global strategies for mental health. These strategies aim to ease the 'burden' of mental disorders and neurological diseases currently affecting about 400 million people, by improving the quality of care throughout the world, particularly in the developing countries.

WHO aims to achieve these objectives through a number of measures in both developing and developed countries. The organization plans to raise awareness of the relative importance of mental disorders as a major contributor to the global burden of disease in different groups, including health professionals and public health decision makers, and also among the general public. WHO will fight the social stigma, misconceptions, and discrimination associated with mental illness, as well as promote the human rights of people with a mental illness: "Very often and in many countries, individuals who are affected by neuropsychiatric disorders endure double suffering, namely from the conditions themselves and from the social stigma and discrimination attached to them" (WHO/67 Press Release, 21 November 1999 – Raising Awareness, Fighting Stigma, Improving Care). In December 2001, the 10th anniversary of the Principle for the Protection of Persons with Mental Illness and the Improvement of Mental Health Care adopted by the UN General Assembly in 1991, Dr Brundtland proposed launching measures to foster the implementation of these principles with, for example, an International Convention on the Rights of Persons with Mental Disorders.

Mental health promotion should go beyond simply achieving the absence of a mental disorder, but aim to improve mental well-being, a state in which individuals can realize their abilities, cope with the stresses of life, work productively, and make a positive contribution. WHO has identified poverty as a great obstacle to such mental well-being. In both mental health promotion and in treatment and care Dr Brundtland saw the family, often extended by the closest community network, as having a central role to play: "Much of the burden of caring for the mentally ill, or for prevention for those who are in danger of becoming ill, is left to the family…" (Office of Director General, WHO, 13 October 1999). The family needed more support and better information, in recognition of their work. She commended the 1995 'Barcelona Manifesto,' put forward by the European Union Federation of Family Associations of Mentally Ill People, which outlines the needs and perceived rights of the family of a person with mental illness, and would consider its principles in WHO's work.

At a regional level, within the European Union, the Council of Ministers of Health has also focused on mental health, in particular the promotion of mental health within the context of improving social inclusion. It has invited the European Commission to analyze the impact of community activities on mental health, for example, in the fields of education, youth policy, social affairs, and employment, and to consider the need to draw up – after consulting with member states – a proposal for Council recommendations on the promotion of mental health in the European Union.

User and Family Organizations Related to Mental Illness

There are organizations focusing on users and families' concerns operating at a number of levels: global, regional, national, and within individual countries. Some organizations are run by users for users, while others involve mental health professionals. At a global level, the World Association for Psychiatric Rehabilitation and the World Federation for Mental Health each has an interest in the perspective of users and carers.

An example of an user-run organization at a regional level is the European Network of (ex-)Users and Survivors of Psychiatry, which describes itself as "a regional initiative to give (ex-) users and survivors of psychiatric services a means to communicate, to exchange opinions, views, and experiences in order to support each other in the personal, political, and social struggle against expulsion, injustice, and stigma in our respective countries."

The history of the network goes back to 1990 when an initiative was taken in the Netherlands to form a network of associations of (former) psychiatric patients from various European countries. Since then the network has organized four

European conferences. At the last conference in Luxembourg in 1999, more than 90 delegates, all of them (ex-)users/survivors from 26 European countries, met and created an action plan for the coming years.

The aims and objectives of the European Network are:

1. The European Network is against any unilateral approach to, and stigmatization of, mental and emotional distress, madness, human suffering, and unconventional behavior.
2. The European Network should support (ex-)users/survivors' autonomy and responsibility in making their own decisions (self-determination).

Priorities for the network include the following:

1. Act against any kind of discrimination in society (both inside and outside the mental health care system) of people who have been subject to the psychiatric system.
2. Support the development of (ex-)user/survivor groups throughout Europe (with a particular emphasis on those countries where there are no existing organizations).
3. Create and support new alternatives to the psychiatric system and collect and share information on the existing ones.
4. Influence and try to change present treatment in psychiatry.

The European Network attempts to influence policy at a European level and maintains contacts with other international organizations active in the mental health field. The network collaborates with the WHO Regional Office for Europe, the European Union, the European Disability Forum, the International Labour Organization, Mental Health Europe/Santé Mental Europe (the former European Regional Council of the World Federation for Mental Health), and the Geneva Initiative on Psychiatry.

The network is a federal structure of national and local associations of (ex-)users and survivors and of mixed organizations with a significant (ex-)user/survivor membership. For countries where there are no such associations, exceptionally individual (ex-)user/survivor members may become members. The aim is for the network to be a grassroots, democratic, and fully (ex-)user/survivor-controlled organization. Through the membership of its member organizations the network represents several tens of thousands of (ex-)users/survivors from across Europe.

In the USA during the last 20 years the National Alliance for the Mentally Ill (NAMI) has grown, with a membership now of over 200 000 family members and service users. It operates at a national, state, and local level to 'provide the nation's voice on mental illness.' It represents family members and service users who seek more equitable services for people with a severe mental illness.

In Europe, voluntary organizations from a number of countries, which represent the relatives of people with a mental illness, came together in DeHaan, Belgium in 1990. A manifesto signed by family associations from 10 European countries led to the establishment of the European Federation of Associations of Families of Mentally Ill People in 1992. At their third congress held in Sweden in 1999, 22 member associations from 15 countries – directly representing 55 000 families – set out their aim to work toward the reduction of stigma and discrimination against people with a mental illness and to seek adoption by health professionals of the highest standards of good practice.

Function of User and Family Organizations Related to Mental Illness

There are some descriptions of individual groups and programs available. One group focuses on issues of loss and how to overcome it (Baxter and Diehl, 1998). The BRIDGES program and the Journey of Hope are peer-taught programs that offer education and support to people with a severe mental illness and their families. The aim is to validate the participants' sense of loss as normal, and provide a structure for a new sense of self, so that individuals and their families can move from isolation and loneliness to empowerment and reconnection with ordinary life.

Families of people with dementia may have particular needs given the progressive nature of the illness and its increasing impact on carers as a person's mental and also physical health declines. Support groups for carers allow members to trade information about the disease and daily care requirements (Wormstall et al., 1996). Around the globe there are many voluntary organizations dealing with mental health issues in general or specific psychiatric disorders such as depression, eating disorders, or suicide. Each of these have specific remit and national charity laws and regulations apply to them.

The Relationship between Family Organizations and Statutory Services

Members of family organizations may be dissatisfied with the care provided by statutory services. Members of the Dutch organization Ypsilon, which supports the relatives of people with a long-term psychotic illness, took part in a questionnaire study. They reported a lack of access to treatment professionals, a lack of information about their relative's illness, and a lack of family involvement in treatment planning (Schene and van Wijngaarden, 1995).

In some family organizations, health professionals may be more involved. In Massachusetts, Alliance for the Mentally Ill (AMI) provides eight family support groups, with a total average monthly attendance of 86 relatives and 25 professionals (Bouricius et al., 1994). Family members lead the groups, but they encourage health professionals to attend. Results of a questionnaire given to the relatives showed that a large majority found the attendance of professionals at the meetings to be helpful in accessing needed services. The researchers also reported that the support groups tend to improve services for people with a mental illness in the area.

Across the USA there is a wide variation in the contact between doctors training in psychiatry and psychoeducational programs involving families of people with a mental illness. In less than half of the psychiatry training schemes is NAMI formally involved. Barbee et al. (1991) give some examples of successful collaboration. In one program, an exercise in which trainee psychiatrists and family members change roles is recommended for its ability to improve communication between trainees and families.

Such programs may have some effect in reducing the delay between a person being diagnosed as having a mental health problem, and family members getting in touch with a support organization. A survey in Quebec of AMI members found that half of the members experienced a delay of over 2 years and only 10% were referred to the organization by a psychiatrist (Looper et al., 1998). The majority of respondents would have preferred earlier contact, and some approved of the more proactive methods developed by AMI-Quebec such as a telephone call after a relative's first admission to hospital.

Benefits of Involvement with a Family Organization

There is some evidence of family members gaining benefit from joining a support group. For example, in a study of 225 families with a relative with a mental illness, there were significant differences between those who participated in a support group and those who did not (Mannion et al., 1996). The support group participants were more likely to be a parent with a higher functioning relative who had been ill for a longer time. They reported less subjective burden and a greater use of adaptive coping mechanisms than did members of families who did not participate in a support group.

It is possible to demonstrate that there are some ingredients of a group that are particularly helpful. A US study compared two types of group, an interactive psychoeducational model and a support model offering nonstructured discussions (Kane et al., 1990). Each was held over four sessions for the families of people with a relative with schizophrenia or schizoaffective disorder. In the psychoeducational group, over three-quarters of the participants rated the quality of the information received as excellent, and 94% said they had received the kind of information they wanted. This compares with a quarter of the support group rating the information as excellent and none feeling that they had received the help they wanted. There was a lower level of depression in members of the psychoeducational group.

A study in Germany reported benefits for family members who joined a self-help group (Schulze-Mönking, 1994). In this case the families of men with a severe mental illness were most likely to join the group. Over the 2 years of the study, there was a tendency for the relatives of the families which participated to have a better outcome. Family members in the group developed more social contacts and reported fewer physical complaints.

As well as providing support and information about a relative's illness, these organizations can fulfill a number of other functions. For example, in the Circle Model, family members caring for a relative who is cognitively impaired are enabled to get some respite (Jansson et al., 1998). Equal numbers of family members and volunteers train together in a study circle, and once trained, the volunteers can replace the carer on a regular basis in the person's home. The relatives gained support from meeting others in the same position as themselves, and they also expressed feelings of security and relaxation in relation to their respite from the home situation. The volunteer caregivers also expressed satisfaction and appreciation of the knowledge they gained from the caregivers.

In many caregiving situations, an elderly parent is looking after a grown-up son or daughter at home or providing substantial amounts of support out of the home. One of the biggest worries for the carer may be what will happen after his or her death. The Planned Lifetime Assistance Network is available in some parts of the USA through NAMI. It aims to provide lifetime assistance to individuals with a mental health problem whose parents or other family carers are no longer alive, or are no longer able to provide care (Lefley and Hatfield, 1999).

Another area of growth outside the statutory sector is in self-help groups in the voluntary sector. Their focus is diverse and represents an interest in promoting self-help and support for people with a range of mental health problems including anxiety disorders, mood disorders, schizophrenia, and dementia. These groups are created and operate at a local, regional, and national level. Some have developed groupings at an international level whose main focus is to act as a lobby group and advocate for intergovernmental initiatives in policies around mental health.

Self-Help Groups

In the USA, Borkman (1997) has linked the development of contemporary self-help groups with the founding of Alcoholics Anonymous (AA) in 1935. AA is a model for other self-help groups, which are also nonhierarchical direct democracies that avoid advocacy but focus on providing support. The major growth of all these groups took place during and after the 1970s alongside the development of the civil rights and women's movements, both of which challenged bureaucracies and traditional authority.

To understand more about the people involved in self-help groups, Chamberlin et al. (1996), using a participatory action research paradigm, with an advisory committee of individuals who had used such a program, carried out a survey of participants in six representative programs. Respondents spent an average 15 h per week in their program and had been attending for almost 5 years. Members both received and gave help. Overall they reported that attending a self-help program had a salutary effect on their quality of life.

Other organizations controlled by consumers or survivors may have different aims, carrying out different activities. Trainor et al. (1997) define consumer/survivor organizations as being operated for and controlled and staffed by people who have used the mental health system. In Ontario, Canada, the Consumer/Survivor Development Initiative funds 36 such organizations. They carry out a range of activities, which include offering mutual support, cultural activities, advocacy skills training, and education for the public and professionals. Involvement in a self-help group leads to a drop in contact with services, with less time spent in inpatient care, and less contact with crisis services, suggesting that contact with such groups may help individuals handle difficulties in a different way.

In the UK information from members of different user groups suggested that these groups could have a role in ensuring that individual users' rights were respected (Barnes and Shardlow, 1997). They also enabled people with a mental health problem to improve the accountability of

services and supported their wider participation as citizens. Working with families and across social and health care is one way forward for delivering mental health services in the twenty-first century (Bhugra and Carlile, 2013). Resilience can be improved by involving families and other stakeholders (SPN, 2007; Friedli, 2009). Patients are at the center of the therapeutic interactions but they are surrounded by families, employers, society, and others and a truly multipronged approach is needed.

See also: Health Interventions, Community-based; Health-Related Support and Self-help Mutual Aid Groups; Mentally Ill, Public Attitudes towards; Public Psychiatry and Public Health Mental Systems; Social Support and Recovery from Disease and Medical Procedures.

Bibliography

Barbee, J.G., Kasten, A.M., Rosenson, M.K., 1991. Towards a new alliance: psychiatric residents and family support groups. Academic Psychiatry 15, 40–49.

Barnes, M., Shardlow, P., 1997. From passive recipient to active citizen: participation in mental health user groups. Journal of Mental Health UK 6, 289–300.

Baxter, E.A., Diehl, S., 1998. Emotional stages: consumers and family members recovering from the trauma of mental illness. Psychiatric Rehabilitation Journal 21, 349–355.

Bhugra, D., Carlile, A., 2013. Starting Today: The Future of Mental Health Services. Mental Health Foundation, London.

Borkman, T., 1997. A selective look at self-help groups in the United States. Health and Social Care in the Community 5, 357–364.

Bouricius, J.K., Kersten, E., Nagy, M., McCartney, P.L., Stein, R., 1994. Family support groups. AMI of western mass style. Innovations and Research 3, 33–40.

Chamberlin, J., Rogers, E.S., Ellison, M.L., 1996. Self-help programmes: a description of their characteristics and their members. Psychiatric Rehabilitation Journal 19 (3), 33–42.

Friedli, L., 2009. Mental Health, Resilience and Inequalities. WHO Europe, Copenhagen.

Jansson, W., Almberg, B., Grafström, M., Winblad, B., 1998. The circle model: support for relatives of people with dementia. International Journal of Geriatric Psychiatry 13, 674–681.

Kane, C.F., DiMartino, E., Jimenez, M., 1990. A comparison of short-term psycho-educational and support groups for relatives coping with chronic schizophrenia. Archives of Psychiatric Nursing 4, 343–353.

Lefley, H.P., Hatfield, A.B., 1999. Helping parental caregivers and mental health consumers cope with parental aging and loss. Psychiatric Services 50, 369–375.

Looper, K., Fielding, A., Latimer, E., Amir, E., 1998. Improving access to family support organizations: a member survey of the AMI-Quebec alliance for the mentally ill. Psychiatric Services 49, 1491–1492.

Mannion, E., Meisel, M., Solomon, P., Draine, J., 1996. A comparative analysis of families with mentally ill adult relatives: support group members versus non-members. Psychiatric Rehabilitation Journal 20 (1), 43–50.

Schene, A.H., van Wijngaarden, B., 1995. A survey of an organization for families of patients with serious mental illness in the Netherlands. Psychiatric Services 46, 807–813.

Schulze-Mönking, H., 1994. Self-help groups for families of schizophrenic patients: formation, development and therapeutic impact. Social Psychiatry and Psychiatric Epidemiology 29, 149–154.

SPN, 2007. Whose Recovery Is It Anyway? Social Perspectives Network, London.

Trainor, J., Shepherd, M., Boydell, K.M., Leff, A., Crawford, E., 1997. Beyond the service paradigm: the impact and implications of consumer/survivor initiatives. Psychiatric Rehabilitation Journal 21, 132–140.

Wormstall, H., Gunthner, A., Morawetz, C., Schmidt, W., 1996. Groups for family caregivers of patients with Alzheimer's disease in Germany. Nervenarzt 67, 751–756.

Mental Imagery: Visual Cognition

Zaira Cattaneo, University of Milano-Bicocca, Milano, Italy
Juha Silvanto, Faculty of Science and Technology, University of Westminster, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mental imagery can be defined as a quasi-perceptual experience occurring in the absence of perceptual input. The present article provides a review of the key processes involved in mental imagery, the relationship of imagery to working memory, and of the debate on the underlying format of mental images. We also review the functional significance of imagery in a range of cognitive processes, such as memory, creative thinking, reasoning, and problem solving. Finally, the brain basis of mental imagery and its overlap with the cortical regions involved in visual perception are discussed.

What Is Mental Imagery?

Mental imagery has been the subject of debate for millennia. Philosophers in Ancient Greece, who argued that "the soul never thinks without a mental image" (Aristotle, De anima 431a 15–20), described mental images as either impressions stamped on wax by thoughts and percepts (Plato, Theatatus 191c,d) or pictures painted in the soul by an inner artist (Plato, Philebus 39c). Mental imagery can be defined as a quasi-perceptual experience occurring in the absence of perceptual input (Rinck and Denis, 2004). It refers to internally generated conscious experiences and is often described as 'seeing in the mind's eye,' 'hearing in the head,' and 'imagining the feel of' (Kosslyn, 1994; Kosslyn et al., 2006). Mental images can be created from information held in long-term memory (LTM) or from recently presented sensory information.

One of the first modern empirical investigations on imagery was carried out in Edward B. Titchener's laboratory at Cornell University by one of his postgraduate students, Mary Cheves West Perky (1910). In her experiments, observers were asked to fixate on a glass screen on which they were required to project mental images of various objects. Unknown to the observers, visual stimuli were back-projected onto the screen while they were engaged in imagery. In almost all cases, the observers were unaware of these stimuli even though in the absence of imagery they were easily detected, indicating that mental images impaired the encoding of concurrently presented external sensory input. However, although the Perky effect is often described in terms of masking of visual input by mental imagery, one of its key features is the merging of internally generated percepts with externally presented information. For example, when participants were asked to imagine a banana in an upright position, the subliminal presentation of a banana in horizontal position modulated the participants' experience of their mental imagery such that they were more likely to report the orientation of the visual stimulus, rather than what they had been asked to imagine. A further fascinating example was provided by the studies of Sydney Segal (e.g., Segal and Gordon, 1969). In one of Segal's experiments, when observers were asked to imagine the New York skyline, they tended to describe imaging the skyline at sunset when a uniform red color was presented on the screen. This aspect of the Perky effect has been recently replicated in a psychophysical paradigm, and it is informative of how mental imagery can be distinct from memory content (Bona et al., 2013). Fundamentally, the Perky effect is theoretically important because the ability of visual input to merge with mental images indicates that the latter calls upon mechanisms that are engaged in the perception of externally presented visual information.

The Imagery Debate (the Analog-Propositional Debate): What Is the Representational Format of Mental Images?

The representational format of mental images has been perhaps the most hotly debated topic in the study of visual imagery. The debate has focused on two competing theoretical viewpoints: the *pictorial* (or *depictivist*) *theory* (Kosslyn, 1994; Kosslyn et al., 2006) and the *propositional* (or *description*) *theory* (Pylyshyn, 2003). According to the depictivist viewpoint (Kosslyn et al., 2006), mental images are representational brain states that are genuinely 'picture-like.' This view is supported by the classic experiments on mental rotation (e.g., Shepard and Metzler, 1971) and mental scanning (Denis and Kosslyn, 1999; see **Figure 1**). In the seminal study of Shepard and Metzler (1971), participants were asked to judge whether two figures that varied in their degree of rotation were the same or different. In order to perform the task, participants presumably needed to use imagery to rotate the figures such that they were in the same orientation. The key finding was that the time taken to perform the task correlated with the angle of rotation by which the two objects differed, consistent with the view that, if the representation is pictorial, rotation of a larger angle is more time-consuming. Similar results were obtained in mental scanning experiments, in which participants had to memorize locations on maps or figures and were asked to mentally scan the path from one location to another. The amount of time required was directly proportional to the distance between the two locations, as if the participants were scanning a physical map with their eyes (Borst and Kosslyn, 2008). These results led to the conclusion that mental images are 'picture-like.'

According to the propositional (or descriptive) view (Pylyshyn, 2003) there is nothing 'picture-like' in the underlying format in which mental images are represented. In this view, the conscious experience of mental imagery is an epiphenomenon

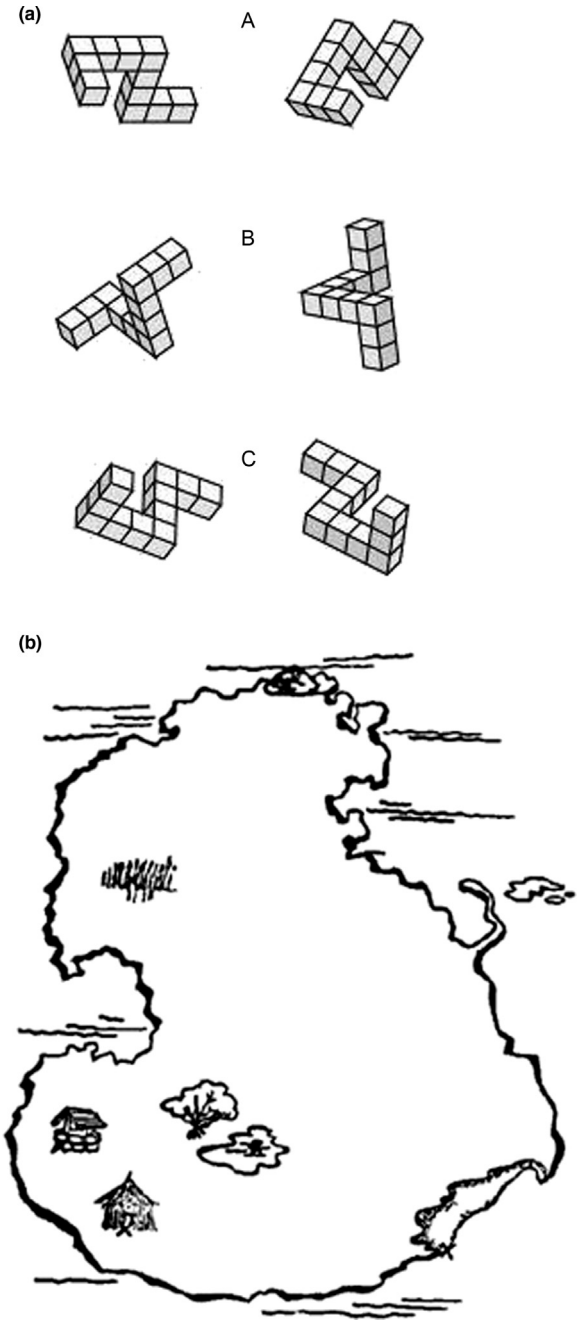

Figure 1 (a) A sample mental rotation task (from Shepard, R.N., Metzler, J., 1971. Mental rotation of three-dimensional objects. Science 171(972), 701–703). Reprinted with permission from AAAS. (b) Scanning mental maps (from Kosslyn, S.M., Ball, T.M., Reiser, B.J., 1978. Visual images preserve metric spatial information: evidence from studies of image scanning. Journal of Experimental Psychology: Human Perception and Performance 4 (1), 47–60). Reprinted with permission from APA.

since they are not indispensable for having an internal representation of an object or a scene. Rather, mental images involve the same forms of representation as reasoning in general; that is, propositional representations consisting of an abstract system of symbols, which can be used to express the contents of the mind. According to the propositional view of mental imagery, many findings used in support of the pictorial view can be accounted for in terms of participants' 'tacit knowledge' of the external visual world. Specifically, Pylyshyn argued that "[…] nearly all experimental findings cited in support of the picture theory can be more naturally explained by the hypothesis that when asked to imagine something, people ask themselves what it would be like to see it, and they then simulate as many aspects of this staged event as they can and as seem relevant. I refer to this explanation as the 'null hypothesis', because it makes no assumptions about format – it appeals only to the tacit knowledge that people have about how things tend to happen in the world, together with certain basic psychophysical skills" (Pylyshyn, 2003).

The imagery debate has been difficult to resolve; however, recent findings from neuroimaging studies have shed new light on the issue. These findings will be reviewed in the last section.

Mental Images Generated from Short-Term Memory versus Long-Term Memory

The classic work of George Sperling (1960) demonstrated that our perceptual system stores a large amount of visual information for approximately 500 ms beyond the presentation of the target stimuli. This memory storage is referred to as *iconic memory*, and it is sometimes confused with mental imagery. The key difference is that, in contrast to iconic memory traces that are somewhat automatic and closely reflect bottom-up visual input, mental images are thought to result from (wilful) top-down processes. Furthermore, mental images can be generated from information held either in LTM or from STM (short-term memory); in the latter case, briefly after the presentation of the to-be-imagined item. For example, participants may be asked to generate a mental image from verbal instructions (e.g., 'imagine an apple'), in which case the image needs to be retrieved from LTM and involves the transfer of verbal information into a visual representation (cf Baddeley, 2007). In contrast, when participants are asked to form an image from visual information (e.g., when they are presented with a picture of an apple and then asked to imagine it beyond the presentation of the visual cue), no transformation is required as the mental image can be based on the representation held in visual working memory (WM). The latter is likely to retain all the features of the cue on which the mental image is based, whereas the former could be conceptualized as a basic prototypical representation of an object. In this vein, a taxonomy of mental images has been suggested, with a distinction between *general* and *specific* images. When mental images are created from LTM, performance in mental scanning paradigms is inferior (reflected as slower reaction times) relative to images based on recently presented visual information. This may be because mental images generated from LTM are akin to 'degraded' iconic images. Thus slower scanning with images generated from LTM would reflect difference in 'vividness' between images generated from LTM and "vivid" images generated from recently encoded perceptual input. Also, mental images generated from LTM may be more fragmented than those generated based on STM content, with the scanning process filling in missing material (Borst and Kosslyn, 2008).

Imagery in Models of Working Memory

In Kosslyn's view (1994; Kosslyn et al., 2006) potential mental images are stored in LTM as 'deep representations,' defined as structural descriptions not directly accessible to awareness. These can be used to generate the actual depictive-analogical images in the so-called *visual buffer*. An analogy can be made with a computer program: data are saved in specific files (analogous to the 'deep representations' stored in LTM) from which pictures can be generated on the computer screen. In this view, the computer screen is analogous to the visual buffer, where information becomes available to consciousness and can be analyzed by the 'mind's eye.' Kossyln (1994; Kosslyn et al., 2006) suggested that visual imagery and visual perception draw on many of the same cognitive processes and representations, and recruit broadly the same brain areas, including the primary visual cortex (see below). Topographically organized visual areas would be grouped into a single functional structure, the visual buffer, representing the medium that supports mental images. Once formed, mental images fade in few hundred milliseconds, corresponding to the time required to initiate and carry out an eye movement (Kosslyn, 1994); thus, active maintenance is required for imagery to last sufficiently long for it to be inspected. In Kosslyn's view, this is accomplished by the reactivation of visual memory representations in an object properties-processing subsystem (Kosslyn et al., 2006).

Even with active maintenance, holding a mental image for more than a few seconds in the 'mind's eye' is challenging (Cocude and Denis, 1988). This may be because mental images rely on the same visual areas of the occipital cortex that underlie visual perception (Kosslyn et al., 2006) and, consequently, there is mutual interference between visual imagery and visual perception (e.g., Baddeley and Andrade, 2000; Quinn and McConnell, 2006). However, imagery can nevertheless be efficiently carried out while the visual system is receiving visual input and there are no consistent differences in the reported vividness of images experienced under 'eyes-open' and 'eyes-closed' conditions (Isaac and Marks, 1994). Image maintenance is thought to depend on attentional resources that can become rapidly depleted (Logie, 1995; Pearson et al., 2001); studies have shown mental images maintenance to highly tax general-purpose attentional resources involving the central executive component of the WM system (Pearson et al., 1996).

In a parallel line of research, Baddeley (2003, 2007) considered mental imagery as a function of the WM network (see also Logie, 1995). Baddeley proposed that "… working memory … temporarily maintains and stores information, supports human thought processes by providing an interface between perception, long-term memory and action" (Baddeley, 2003: p. 829). The multi-component model of WM emphasizes the operational capacity of the system (hence the name *working* memory), in contrast to previous models that conceptualized memory as primarily a storage facility. In Baddeley's original model, a key component of WM is the *visuospatial sketchpad*, dedicated to storage and maintenance of visuospatial information and mediating both the retention of perceptual information and mentally generated images. When a mental image is generated from visual input, it can be directly mapped into the *visuospatial sketchpad*. In contrast, when visual mental images need to be generated from verbal cues, retrieval from LTM with the aid of phonological loop (the *verbal* component of the WM system) is required (Baddeley, 2007). Cornoldi et al. (1998) distinguished between visual traces and generated images on the basis of the degree of attentional requirement, voluntary manipulation, rehearsal mechanisms, and capacity limitation associated with them.

Subsequent models, motivated by Baddeley's theory, have conceptualized the visuospatial WM system as a multi-component mental workspace, with distinct components playing key roles in visual imagery and visual short-term retention (VSWM; Logie, 1995, 2011). According to the VSWM model, the generation of mental images requires the activation of stored knowledge about the stimulus; in addition, executive functions involved in processing and manipulating the activated knowledge are necessary. Image generation requires an active store (visual buffer) that is proposed to use brain structures involved in visual perception, such as the primary visual areas in the posterior occipital cortex (Kosslyn et al., 2006; Kosslyn and Thompson, 2003). Because this active store uses resources involved in visual perception, it is sensitive to disruption from irrelevant visual noise. However, the actual retention of both visual information and mental images occurs in a passive VSTM store or visual cache (Logie, 1995). There is evidence in support of this view: once the mental image has been generated, it is not susceptible to irrelevant visual input (Borst et al., 2012, but see Bona et al., 2013).

Imagery and 'Visual' Cognition

Imagery has been shown to play a facilitatory role in a number of cognitive functions such as learning, memory, reasoning, mathematics, and creative thought. Thus, imagery is functionally beneficial in a range of cognitive functions. These will be reviewed below.

Learning and Memory

The beneficial contribution of mental imagery to memory has been known for centuries, and mnemonic techniques involving imagery were used already in the ancient times (Yates, 1966). As reported by Cicero (106–43 BC), it was the Greek poet Simonides (*c*.556–*c*.468 BC) who first discovered the powerful mnemonic properties of imagery. As the story goes, Simonides was attending a banquet in Thessaly in order to present a poem in praise of his host. After his speech, Simonides was called outside and, during his short absence, the roof of the banqueting hall suddenly collapsed. Simonides was able to identify the victims on the basis of where their bodies were found by creating from his memory a mental image of the diners sitting around the banqueting table. From this tragic event, Simonides concluded that the creation of a mental image in which the items to be remembered are linked to specific locations is an efficient way of facilitating memory (Cicero, *De Oratore*, II, lxxxvi); this technique is known as the *method of loci*. As described by Cicero and Quintilian (AD *c*.35–*c*.95), it was originally

employed by speakers to retrieve passages of their speech in the correct order. The English phrases 'in the first place' and 'in the second place' originate from this technique.

In modern psychology, the dual coding theory of memory proposed by Allan Paivio (1991) provides a theoretical framework for the mnemonic benefits of imagery. In this model, memory is comprised of two distinct systems with unique structural and functional properties. The first system consists of verbal representations and is based on language; the second derives from perceptual experience. Imagery can facilitate the recall of verbal material because the presentation of a word (e.g. apple) activates not only the linguistic memory trace, but also the visual memory related to the word. Consequently, concrete words are recalled more successfully than abstract ones because the former are more likely to activate a visual memory trace (Paivio, 1991). Numerous subsequent studies have shown that the concreteness of a piece of text is a key predictor of subsequent recall performance (e.g., Sadoski et al., 2000).

Reasoning and Mental Models

Mental imagery and reasoning are strongly interconnected (De Soto et al., 1965). Reasoning is a cognitive process in which new information is derived from statements provided to the participant. Two major theories have dominated the cognitive literature on deductive reasoning. The *mental logic theory* posits that the deductive processes are carried out by a system of linguistic rules sensitive to the logical form of the argument (Rips, 1994), whereas the *mental model theory* proposes that a visuo-spatial representation of the argument is constructed and evaluated (Johnson-Laird, 1994). According to the mental model theory, many problems can be solved in an analogical, 'pictorial' format: in fact, a mental model can be conceived as a representation *isomorphic* to the physical situation that it represents, and in which the inference simulates the physical processes being reasoned about. Imagery seems to be particularly important for *relational reasoning*, in which the problem is usually given by two statements (referred to as premises) and the task is to derive the conclusion that follows from these premises. A basic example of relational reasoning is 'If A is taller than B and B is taller than C, is C shorter than A?' A relational reasoning problem can be represented as a mental image where the conclusion can be 'read off' by inspecting the mental image (Goodwin and Johnson-Laird, 2005). According to this approach, the information given in the premise is integrated into a single spatial representation that usually takes the form of a linear array. This array is generally left-to-right oriented as a result of a cultural bias in visual scanning. As computational models suggest, the inspection process can be functionally described as a shift of spatial focus that checks the various cells of the spatial array and infers from the scan direction the relation between two objects in the array (Schlieder and Berendt, 1998). This type of representation can be used to derive novel conclusions even when the relation is temporal or abstract (Boroditsky, 2000; Goodwin and Johnson-Laird, 2005).

Although some studies have found positive effects of imageability on reasoning, others have failed to obtain such results. This discrepancy may result from the difference between *spatial imagery* and *visual imagery* (Knauff and May, 2006). *Visual imagery* refers to the representation of the visual appearance of an object, including its shape, color, and brightness, whereas *spatial imagery* refers to the representation of the spatial relationships between the parts of an object and the location and movement of the object in space. *Spatial* imagery has been proposed to be critical for reasoning while the generation of visual images can occur when the generation of the appropriate spatial mental model requires the retrieval of spatial information from a visual image (Knauff and May, 2006). In these situations, additional processes come into play, which can even impede the reasoning process (Knauff and Johnson-Laird, 2002). In short, the 'structure' of the reasoning process is likely to consist of spatially organized mental models that cannot be identified with visual images.

Mathematics and Numerical Cognition

It has been suggested that all mathematical tasks require *spatial* thinking. As in reasoning, the distinction between spatial and visual imagery is important (Hegarty and Kozhevnikov, 1999). The use of schematic images (i.e. images that encode the spatial relations described in a problem) has been found to be positively correlated with success in mathematical problem solving in sixth-grade students, whereas the use of pictorial images (i.e. images that encode the visual appearance of objects or people) have the opposite effect (Presmeg, 1997, 2006). Presmeg (2006) has identified five types of visual imagery that can be used by students to solve mathematical problems: (1) concrete pictorial imagery (pictures in the mind); (2) pattern imagery (pure relationships depicted in a visual-spatial scheme); (3) kinaesthetic imagery (hand movement and other gestures); (4) dynamic imagery (dynamic transformations of geometric figures); and (5) memory of formulas, wherein visualizers typically imagine a formula written on a blackboard or in their notebook. Only some forms of imagery facilitate performance and others may even be harmful (Presmeg, 1997, 2006). For instance, the use of concrete pictorial imagery may focus reasoning on irrelevant details that distract from the key elements of the problem. Pattern imagery (i.e., imagery where pure relationships are depicted in a visual-spatial scheme) may be the most useful type of imagery in mathematical reasoning because it represents the *relational* aspects of a problem and is better suited to abstraction and generalization (Presmeg, 1997, 2006). Data obtained with gifted students and with those affected by learning disabilities support this view. The former have been found to perform better in spatial visualization tasks than students with learning disabilities as well as average-performing students, and the use of schematic imagery in problem solving is positively correlated with performance on each spatial visualization measure (van Garderen, 2006).

Creative Thought

Mental imagery is known to have played an important role in the work of many great scientists, for example, in Galvani's discovery of 'animal electricity' (i.e. the electricity present in

a condition of disequilibrium between the interior and the exterior of excitable animal fibers) (Piccolino, 2008). However, most of the available positive evidence is based on self-reports of major figures in the history of science rather than on empirical studies. According to a meta-analysis by LeBoutillier and Marks (LeBoutillier and Marks, 2003), imaging capacity may account for only the 3% of the total variance in divergent thinking (i.e., creative thought) measures. Mental imagery has been widely investigated in psychology of sport, where its use has been demonstrated to be beneficial to improve athletes' performances (see Smith and Wright, 2008) and in the training of high-level manual skills. For example, imagery is frequently used by surgeons as preparation for operations (Sanders et al., 2008).

Imagery and Perception: Where in the Brain?

Neuroimaging techniques (such as positron emission tomography (PET), functional magnetic resonance imaging (fMRI)) and noninvasive brain stimulation techniques (such as transcranial magnetic stimulation (TMS)) have been instrumental in our understanding of cortical circuits involved in visual imagery. Several neuroimaging studies have found visual imagery to engage the primary visual cortex and extrastriate cortices (for reviews, see Kosslyn et al., 2001; Kosslyn and Thompson, 2003) and imagery vividness appears to be positively correlated with visual cortical activity (Cui et al., 2007). Using brain stimulation, it is possible to transiently impair the activity in specific brain areas and assess their causal role in cognitive tasks. Application of TMS over the primary visual cortex has been found to impair performance in visual imagery, demonstrating that activity in this region is functionally significant for visual imagery, rather than merely being epiphenomenal (Cattaneo et al., 2009). Furthermore, visual cortex excitability is increased by visual imagery, demonstrating that the engagement of imagery recruits the early visual cortex (Sparing et al., 2002; Cattaneo et al., 2011). A particularly direct demonstration of the overlap between visual perception and imagery comes from a study by Kreiman et al. (2000), in which neural activity was directly recorded from patients undergoing surgery. It was found that many of the same neurons were activated by visual presentation of stimuli as well as when the participants were asked to imagine these visual stimuli. This is, perhaps, the strongest single piece of evidence indicating an overlap between the neuronal basis of visual perception and imagery.

Further evidence has come from the use of the fMRI-adaptation paradigm, which is based on the reduction of the BOLD signal when a sensory stimulus is repeatedly presented. Yi et al. (2008) found that the effect of mentally visualizing a recent perceptual stimulus was comparable to its repeated visual presentation, suggesting that the same population of neurons in the visual cortex is involved in visual perception and mental imagery. Furthermore, category-specific visual areas were activated when participants imagined faces, houses, or chairs; thus the extrastriate processing of imagery is associated with similar functional specialization as visual perception (Cichy et al., 2012; Ishai et al., 2000; O'Craven and Kanwisher, 2000). Similarly, Goebel et al. (1998) found activity in perceptual motion area V5/MT when participants were required to imagine movements. Furthermore, the spatial extent of imagined objects has been found to match with the known spatial organization of early visual areas, with objects subtending smaller visual angles producing activity near the occipital pole (where stimuli falling in the central visual field are represented), and objects subtending larger visual angles inducing more anterior activity (where stimuli falling in the peripheral visual field are represented) (Kosslyn et al., 1995). Finally, behavioral cross-adaptation effects between imagery and visual perception have been observed in conjunction with brain stimulation, which have provided direct causal evidence for the recruitment of the same orientation channels in the early visual cortex (V1/V2) in both imagery and perception (Cattaneo et al., 2012).

According to Kosslyn (1994; Kosslyn et al., 2006), the overlap in cortical regions involved in imagery and perception reflects the involvement of the visual buffer in both processes. In Kosslyn's view, visual mental images arise through reactivation of the same cortical areas that process the relevant information during perception. Such reactivation mechanism of visual cortices would be possible by means of top-down projections from other (higher-level) cortical areas. Whether visual imagery involves primary visual cortices (V1) has been a controversial issue. Slotnick et al. (2005) reported that a high-resolution visual mental imagery task evoked topographically organized activity in striate cortex, but this was found only in half of the participants. Furthermore, V1 activity in imagery may reflect memory processes rather than imagery per se. Kaas et al. (2010) used a motion imagery paradigm in which participants were never presented with a stimulus identical to the one they had to imagine. The authors found increased activation in the motion-selective extrastriate area hMT+/V5, whereas critically, the BOLD signal in V1 was *decreased* during imagery – a result in accordance with the view that conscious visual perception is suppressed during mental imagery. Because in most imagery studies participants are instructed to imagine a manipulated version of a stimulus they have encountered during the experiment, the observed V1 activation might reflect memory for the stimulus on which visual imagery is based rather than the mental image per se. According to a meta-analysis of imagery studies carried out by Kosslyn and Thompson (2003), two variables predict whether early visual areas are activated during visual mental imagery: firstly, the imagery task needs to involve high-resolution images (with details subtending 1 degree of visual angle, or less) and secondly the task should not rely on the processing of purely spatial relations. The overlap of cortical networks involved in perception and imagery is not complete, however, as shown by neuropsychological evidence. For example, there are reports of patients with severely impoverished imagery abilities despite normal visual perception and an anatomically intact primary visual cortex (V1; Moro et al., 2008). However, traumatic brain injuries like those suffered by patients mentioned by Moro et al. (2008), typically involve diffuse axonal injury, which are likely to disrupt large-scale brain networks that are involved in imagery processes beyond early visual cortices.

Neuropsychological studies have found dissociations between spatial and visual imagery (Farah et al., 1988); for example, patients may be severely impaired in tasks assessing visual imagery while displaying normal performance for spatial imagery. Accordingly, it has been suggested that mental imagery is subject to the same dichotomy that characterizes visual perception, namely a distinction between a dorsal pathway specialized in the processing of spatial attributes and a ventral pathway focused on shape and color (Mellet et al., 1996). In line with this, purely spatial mental imagery tasks have been found to activate regions along the dorsal pathway but not early visual cortices (Mellet et al., 1996). For example, the clock task that requires the forming of a mental image depicting the spatial relationship between clock hands from verbal instructions (see the next section for further details on this task), strongly activates the intraparietal sulcus (IPS), a region along the dorsal route generally recruited during execution of visuospatial perceptual tasks (Trojano et al., 2000). A meta-analysis of 32 neuroimaging studies found that mental rotation is usually accompanied by increased activity in the IPS and adjacent regions, and that the amount of activation in these regions is consistently reflected to the amount of mental rotation required by each specific task (Zacks, 2008).

Finally, although most of the neuroimaging studies have been carried out on visual imagery, partially overlapping networks for perception and imagery have also been reported for tactile, auditory, olfactory, gustatory, and motor imagery (Kobayashi et al., 2004; Stevenson and Case, 2005).

Measuring Imagery Abilities in Health and Disease

A number of paradigms have been developed to measure imagery (for a review, see Pearson et al., 2013). Among the tests developed to test image generation ability, some require generation of a matrix of a given size (e.g., 4 × 5 and 5 × 5, depending on the specific version of the task) in which some squares are shaded, forming random patterns (Podgorny and Shepard, 1978) or a letter (Kosslyn et al., 1988). In these paradigms, the behavioral task assessing imagery accuracy involves judging whether a probe item falls within the generated pattern. In a similar paradigm, participants are asked to memorize a series of randomly positioned dots and, after a retention interval, an arrow is presented; participants are then required to report whether the arrow points to one of the originally presented dots (Finke and Pinker, 1983). Another common test of imagery generation is the clock test, originally developed by Paivio (1978) and used in many variants (Grossi et al., 1989; Cattaneo et al., 2009). In this task, individuals have to convert a digital time (e.g., '2:45') into the corresponding analog image and to compare angles (formed by the clock hands) corresponding to different digital times. In order to perform the task successfully the observer needs first to generate the image of the clock hands corresponding to each of the two given times. Both images then need to be held in WM in order to allow a direct comparison between the two angles. Among the tasks used to assess image maintenance, the visual pattern test developed by Della Sala et al. (1999) requires observers to memorize a pattern of cells visually presented on a matrix and then reproduce them on a blank matrix after some retention interval.

Cocude and Denis (1988) also developed a test to measure *imagery duration*. In their task, participants were asked to generate mental images in response to given words varying in degree of concreteness and then to indicate when their imagery experience of the images ceased. *Image inspection* has been investigated in map scanning tasks (e.g., Mellet et al., 2002), already described earlier in this article (see **Figure 1**), requiring participants to mentally move from one location on the map to another.

Mental size comparisons tasks have been widely used to measure image inspection capacity. In these tasks, participants are asked to report, for example, which of two animals is the larger (Moyer and Bayer, 1976). To perform successfully in these tasks, participants need to form a mental image of both animals and to compare them visually. *Image transformation and manipulation* can be assessed using the mental rotation task originally developed by Shepard and Metzler (1971) and described earlier in this article (see **Figure 1**). Some tasks have been specifically designed to tap the capacity to create a new image (mental synthesis): for instance, Finke et al. (1989) required participants to follow instructions such as, "Imagine a capital letter D, rotate it by 90 degrees to the left. Now place a capital letter J at the bottom. Identify the final figure (an umbrella)".

Other tests have been used to assess how mental imagery is subjectively experienced. The phenomenological characteristics of mental imagery have been studied with subjective measures in which individuals are asked to describe and rate a mental image in terms of vividness (Pearson et al., 2001; Bona et al., 2013). Betts' Questionnaire upon Mental Imagery (Betts, 1909) and the Vividness of Mental Imagery (VVIQ, Marks, 1973) are examples of such tests. Both require participants to rate the vividness of mental images generated on the basis of verbal instructions. Tests have also been developed to investigate how often individuals spontaneously use imagery (Spontaneous Use of Imagery Scale, SUIS; Reisberg et al., 2003), in thinking and episodic memory (Holmes et al., 2008) or in problem solving (the Visualiser-Verbaliser Cognitive Style Questionnaire, Kozhevnikov et al., 2005).

Conclusions

Mental imagery refers to internally generated conscious experiences, perceived in the 'mind's eye,' which can be generated from items in LTM or STM. Its importance was recognized already by the philosophers of Ancient Greece, who argued that 'the soul never thinks without a mental image.' Imagery is an important cognitive structure also because it contributes to other cognitive processes such as memory, reasoning, and creativity. Particularly, in the memory literature there is much evidence of the beneficial contribution of imagery to recall performance. From a theoretical perspective, imagery has been the subject of much controversy. The so-called 'imagery debate' has centered on two competing theoretical viewpoints, the *pictorial* (or *depictivist*) *theory* and the *propositional* (or *description*) *theory*, and it has been perhaps the most hotly debated topic in the study of visual imagery. Over the past

two decades, neuroimaging and brain stimulation techniques have begun to shed light on the brain basis of imagery processes. These studies have revealed the involvement of the same brain regions that encode incoming sensory information, indicating that imagery from the neural perspective is sensory in nature.

See also: Brain Development: Ontogenetic Neurobiology of; Brain Stimulation Reward; Cortical Areas Engaged in Movement: Neuroimaging Methods; Human Spatial Orientation, Neural Basis of; Occipital Lobe (Visual Cortex): Functional Aspects; Personality and Creativity; Spatial Cognition During Infancy and Early Childhood Across Cultures, Development of; Spatial Navigation; Virtual Reality and Spatial Cognition; Visual Imagery; Visual Memory, Psychology of; Visual Perception, Neural Basis of; Visual Streams: Dorsal and Ventral; Working Memory, Neural Basis of; Working Memory, Psychology of.

Bibliography

Baddeley, A.D., 2003. Working memory: looking back and looking forward. Nature Reviews Neuroscience 4, 829–839.

Baddeley, A.D., 2007. Working Memory, Thought, and Action. Oxford University Press, Oxford.

Baddeley, A.D., Andrade, J., 2000. Working memory and the vividness of imagery. Journal of Experimental Psychology 129, 126–145.

Betts, G.H., 1909. The Distribution and Functions of Mental Imagery (No. 26). Teachers College, Columbia University.

Bona, S., Cattaneo, Z., Vecchi, T., Soto, D., Silvanto, J., 2013. Metacognition of visual short-term memory: dissociation between objective and subjective components of VSTM. Frontiers in Psychology, 4–62.

Boroditsky, L., 2000. Metaphoric structuring: understanding time through spatial metaphors. Cognition 75, 1–28.

Borst, G., Kosslyn, S.M., 2008. Visual mental imagery and visual perception: structural equivalence revealed by scanning processes. Memory and Cognition 36, 849–862.

Borst, G., Niven, E., Logie, R.H., 2012. Visual mental image generation does not overlap with visual short-term memory: a dual-task interference study. Memory and Cognition 40, 360–372.

Cattaneo, Z., Vecchi, T., Pascual-Leone, A., Silvanto, J., 2009. Contrasting early visual cortical activation states causally involved in visual imagery and short-term memory. European Journal of Neuroscience 30, 1393–1400.

Cattaneo, Z., Pisoni, A., Papagno, C., Silvanto, J., 2011. Modulation of visual cortical excitability by working memory: effect of luminance contrast of mental imagery. Frontiers in Psychology 25, 2–29.

Cattaneo, Z., Bona, S., Silvanto, J., 2012. Cross-adaptation combined with TMS reveals a functional overlap between vision and imagery in the early visual cortex. Neuroimage 59, 3015–3020.

Cichy, R.M., Heinzle, J., Haynes, J.D., 2012. Imagery and perception share cortical representations of content and location. Cerebral Cortex 22 (2), 372–380.

Cocude, M., Denis, M., 1988. Measuring the temporal characteristics of visual images. Journal of Mental Imagery 12, 101–108.

Cornoldi, C., De Beni, R., Giusberti, F., Massironi, M., 1998. Memory and imagery: a visual trace is not a mental image. In: Conway, M.A., Gathercole, S.E., Cornoldi, C. (Eds.), Theories of Memory, vol. II. Psychology Press, Hove, UK, pp. 87–110.

Cui, X., Jeter, C.B., Yang, D., Montague, P.R., Eagleman, D.M., 2007. Vividness of mental imagery: individual variability can be measured objectively. Vision Research 47, 474–478.

De Soto, C.B., London, M., Handel, S., 1965. Social reasoning and spatial paralogic. Journal of Personality and Social Psychology 2, 513–521.

Della Sala, S., Grey, C., Baddeley, A., Allamano, N., Wilson, L., 1999. Pattern span: a tool for unwelding visuo-spatial memory. Neuropsychologia 37, 1189–1199.

Denis, M., Kosslyn, S.M., 1999. Scanning visual images: a window on the mind. Cahiers de Psycliologie Cognitive 18, 409–465.

Farah, M.J., Hammond, K.M., Levine, D.N., Calvanio, R., 1988. Visual and spatial mental imagery: dissociable systems of representation. Cognitive Psychology 20, 439–462.

Finke, R.A., Pinker, S., 1983. Directional scanning of remembered visual patterns. Journal of Experimental Psychology Learning Memory and Cognition 9, 398–410.

Finke, R.A., Pinker, S., Farah, M.J., 1989. Reinterpreting visual patterns in mental imagery. Cognitive Science 13 (1), 51–78.

Goebel, R., Khorram-Sefat, D., Muckli, L., Hacker, H., Singer, W., 1998. The constructive nature of vision: direct evidence from functional magnetic resonance imaging studies of apparent motion and motion imagery. European Journal of Neuroscience 10, 1563–1573.

Goodwin, G.P., Johnson-Laird, P.N., 2005. Reasoning about relations. Psychological Review 112, 468–493.

Grossi, D., Modafferi, A., Pelosi, L., Trojano, L., 1989. On the different roles of the cerebral hemispheres in mental imagery: the "o'Clock Test" in two clinical cases. Brain and Cognition 10, 18–27.

Hegarty, M., Kozhevnikov, M., 1999. Types of visual-spatial representations and mathematical problem solving. Journal of Educational Psychology 91, 684–689.

Holmes, E.A., Mathews, A., Mackintosh, B., Dalgleish, T., 2008. The causal effect of mental imagery on emotion assessed using picture–word cues. Emotion 8 (3), 395–409.

Isaac, A.R., Marks, D.F., 1994. Individual differences in mental imagery experience: developmental changes and specialization. British Journal of Psychology 85, 479–500.

Ishai, A., Ungerleider, L.G., Haxby, J.V., 2000. Distributed neural systems for the generation of visual images. Neuron 28, 979–990.

Johnson-Laird, P.N., 1994. Mental models and probabilistic thinking. Cognition 50 (1–3), 189–209.

Kaas, A., Weigelt, S., Roebroeck, A., Kohler, A., Muckli, L., 2010. Imagery of a moving object: the role of occipital cortex and human MT/V5+. Neuroimage 49 (1), 794–804.

Knauff, M., May, E., 2006. Mental imagery, reasoning, and blindness. Quarterly Journal of Experimental Psychology 59 (1), 161–177.

Knauff, M., Johnson-Laird, P.N., 2002. Visual imagery can impede reasoning. Memory and Cognition 30 (3), 363–371.

Kobayashi, M., Takeda, M., Hattori, N., Fukunaga, M., Sasbe, T., Inoue, Y., Nagai, Y., Sawada, T., Sadato, N., Watanabe, Y., 2004. Functional imaging of gustatory perception and imagery: "top-down" processing of gustatory signals. Neuroimage 23, 1271–1282.

Kosslyn, S.M., 1994. Image and Brain: The Resolution of the Imagery Debate. MIT Press, Cambridge, MA.

Kosslyn, S.M., Ball, T.M., Reiser, B.J., 1978. Visual images preserve metric spatial information: evidence from studies of image scanning. Journal of Experimental Psychology: Human Perception and Performance 4 (1), 47–60.

Kosslyn, S.M., Cave, C.B., Provost, D.A., Von Gierke, S.M., 1988. Sequential processes in image generation. Cognitive Psychology 20 (3), 319–343.

Kosslyn, S.M., Ganis, G., Thompson, W.L., 2001. Neural foundations of imagery. Nature Reviews Neuroscience 2 (9), 635–642.

Kosslyn, S.M., Pascual-Leone, A., Felician, O., Camposano, S., Keenan, J.P., Thompson, W.L., Ganis, G., Sukel, K.E., Alpert, N.M., 1999. The role of area 17 in visual imagery: convergent evidence from PET and rTMS. Science 284 (5411), 167–170.

Kosslyn, S.M., Thompson, W.L., 2003. When is early visual cortex activated during visual mental imagery? Psychological Bulletin 129 (5), 723–746.

Kosslyn, S.M., Thompson, W.L., Kim, I.J., Alpert, N.M., 1995. Topographical representations of mental images in primary visual cortex. Nature 378 (6556), 496–498.

Kosslyn, S.M., Thompson, W.L., Ganis, G., 2006. The Case for Mental Imagery. Oxford University Press, Inc., New York.

Kozhevnikov, M., Kosslyn, S., Shephard, J., 2005. Spatial versus object visualizers: a new characterization of visual cognitive style. Memory and Cognition 33 (4), 710–726.

Kreiman, G., Koch, C., Fried, I., 2000. Imagery neurons in the human brain. Nature 408 (6810), 357–361.

LeBoutillier, N., Marks, D.F., 2003. Mental imagery and creativity: a meta-analytic review study. British Journal of Psychology 94 (Pt 1), 29–44.

Logie, R.H., 1995. Visuo-spatial Working Memory. Lawrence Erlbaum Associates, Hove, UK.

Logie, R.H., 2011. The functional organization and capacity limits of working memory. Current Directions in Psychological Science 20, 240–245.

Marks, D.F., 1973. Visual imagery differences in the recall of pictures. British Journal of Psychology 64 (2), 17–24.

Mellet, E., Bricogne, S., Crivello, F., Mazoyer, B., Denis, M., Tzourio-Mazoyer, N., 2002. Neural basis of mental scanning of a topographic representation built from a text. Cerebral Cortex 12 (12), 1322–1330.
Mellet, E., Tzourio, N., Crivello, F., Joliot, M., Denis, M., Mazoyer, B., 1996. Functional anatomy of spatial mental imagery generated from verbal instructions. Journal of Neuroscience 16 (20), 6504–6512.
Moro, V., Berlucchi, G., Lerch, J., Tomaiuolo, F., Aglioti, S.M., 2008. Selective deficit of mental visual imagery with intact primary visual cortex and visual perception. Cortex 44 (2), 109–118.
Moyer, R.S., Bayer, R.H., 1976. Mental comparison and the symbolic distance effect. Cognitive Psychology 8 (2), 228–246.
O'Craven, K., Kanwisher, N., 2000. Mental imagery of faces and places activates corresponding stimulus-specific brain regions. Journal of Cognitive Neuroscience 12, 1013–1023.
Paivio, A., 1978. Comparisons of mental clocks. Journal of Experimental Psychology Human Perception and Performance 4 (1), 61–71.
Paivio, A., 1991. Images in Mind: The Evolution of a Theory. Harvester Wheatsheaf, New York.
Pearson, D.G., De Beni, R., Cornoldi, C., 2001. The generation and transformation of visuo-spatial mental images. In: Denis, M., Logie, R.H., Cornoldi, C., de Vega, M., Engelkamp, J. (Eds.), Imagery, Language and Visuo-spatial Thinking. Psychology Press, Hove, pp. 1–23.
Pearson, D.G., Deeprose, C., Wallace-Hadrill, S.M., Heyes, S., Holmes, E.A., 2013. Assessing mental imagery in clinical psychology: a review of imagery measures and a guiding framework. Clinical Psychology Review 33 (1), 1–23.
Pearson, D.G., Logie, R.H., Green, C., 1996. Mental manipulation, visual working memory, and executive processes. Psychologische Beitrage 38 (3–4), 324–342.
Perky, C.W., 1910. An experimental study of imagination. American Journal of Psychology 21 (3), 422–452.
Piccolino, M., 2008. Visual images in Luigi Galvani's path to animal electricity. Journal of the History of the Neurosciences 17 (3), 335–348.
Podgorny, P., Shepad, R.N., 1978. Functional representations common to visual perception and imagination. Journal of Experimental Psychology Human Perception and Performance 4 (1), 21–35.
Presmeg, N.C., 1997. Generalization using imagery in mathematics. In: English, L.D. (Ed.), Mathematical Reasoning: Analogies, Metaphors, and Images. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 299–312.
Presmeg, N.C., 2006. Research on Visualization in Learning and Teaching Mathematics. Handbook of Research on Psychology of Mathematic Education: Past, Present and Future. P.B.A. Gutierrez. Sense Publishers, Rotterdam, The Netherlands, pp. 205–235.
Pylyshyn, Z.W., 2003. Explaining mental imagery: now you see it, now you don't. Reply to Kosslyn et al. Trends in Cognitive Sciences 7 (3), 111–112.
Quinn, J.G., McConnell, J., 2006. The interval for interference in conscious visual imagery. Memory 14 (2), 241–252.
Reisberg, D., Pearson, D.G., Kosslyn, S.M., 2003. Intuitions and introspections about imagery: the role of imagery experience in shaping an investigator's theoretical views. Applied Cognitive Psychology 17 (2), 147–160.
Rinck, M., Denis, M., 2004. The metrics of spatial distance traversed during mental imagery. Journal of Experimental Psychology Learning Memory and Cognition 30 (6), 1211–1218.
Rips, L.J., 1994. The Psychology of Proof: Deductive Reasoning in Human Thinking. MIT Press.
Sadoski, M., Goetz, E.T., Rodriguez, M., 2000. Effects of concreteness on comprehensibility, interest, and recall in four text types. Journal of Educational Psychology 92 (1), 85–95.
Sanders, C.W., Sadoski, M., van Walsum, K., Bramson, R., Wiprud, R., Fossum, T.W., 2008. Learning basic surgical skills with mental imagery: using the simulation centre in the mind. Medical Education 42, 607–612.
Schlieder, C., Berendt, B., 1998. Mental model construction in spatial reasoning: a comparison of two computational theories. In: Mind Modelling: A Cognitive Science Approach to Reasoning, Learning and Discovery, pp. 133–162.
Segal, S.J., Gordon, P.E., 1969. Perceptual and Motor Skills 28 (3), 791–797.
Shepard, R.N., Metzler, J., 1971. Mental rotation of three-dimensional objects. Science 171 (972), 701–703.
Slotnick, S.D., Thompson, W.L., Kosslyn, S.M., 2005. Visual mental imagery induces retinotopically organized activation of early visual areas. Cerebral Cortex 15 (10), 1570–1583.
Smith, D., Wright, C., 2008. Imagery and sport performance. In: Lane, A. (Ed.), Topics in Applied Psychology: Sport and Exercise Psychology. Hodder Education, London, pp. 139–150.
Sparing, R., Mottaghy, F.M., Ganis, G., Thompson, W.L., Topper, R., Kosslyn, S.M., Pascual-Leone, A., 2002. Visual cortex excitability increases during visual mental imagery – a TMS study in healthy human subjects. Brain Research 938, 92–97.
Sperling, G., 1960. The information available in brief visual presentations. Psychological Monographs: General and Applied 74.
Stevenson, R.J., Case, T.I., 2005. Olfactory imagery: a review. Psychonomic Bulletin and Review 12, 244–264.
Trojano, L., Grossi, D., Linden, D.E.J., Foemisano, E., Hacker, H., Zanella, F.E., Goebel, R., Di Salle, F., 2000. Matching two imagined clocks: the functional anatomy of spatial analysis in the absence of visual stimulation. Cerebral Cortex 10, 473–481.
van Garderen, D., 2006. Spatial visualization, visual imagery, and mathematical problem solving of students with varying abilities. Journal of Learning Disabilities 39, 496–506.
Yates, F.A., 1966. The Art of Memory. Routledge and Kegan Paul, London.
Yi, D.J., Turk-Browne, N.B., Chunn, M.M., Johnson, M.K., 2008. When a thought equals a look: refreshing enhances perceptual memory. Journal of Cognitive Neuroscience 20, 1371–1380.
Zacks, J.M., 2008. Neuroimaging studies of mental rotation: a meta-analysis and review. Journal of Cognitive Neuroscience 20, 1–19.

Relevant Website

http://www.imagery-imagination.com/.

Mentally Ill, Public Attitudes towards

Dinesh Bhugra and Sara Evans-Lacko, King's College London, London, UK
W Cutter, South London and Maudsley NHS Trust, Denmark Hill, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Attitudes and behavior toward those with mental illness and those who deal with mental illness are important to study for a number of reasons. These reasons include engagement with services but also research and service funding. Numerous studies have shown high prevalence of negative attitudes among the general public toward people with mental illness, such as beliefs that people with mental illness are dangerous, lack will power, and are unreliable. Attitudes can be assessed individually; but, attitudes often reflect cultural context and in turn are influenced by a number of other factors. The formation of attitudes can be related to upbringing, past experience, knowledge, and other variables. Although attitudes do not always turn into behavior, public attitudes can have significant consequences for people with mental illness including higher rates of mortality, higher levels of self-stigma, lower levels of empowerment, lower levels of help-seeking, and higher rates of unemployment (S. Evans-Lacko et al., 2012; Hatzenbuehler et al., 2013; Rüsch et al., 2005; Schomerus et al., 2014). Awareness of attitudes and beliefs among the public is germane for developing primary prevention programs, early intervention, and community-based treatments.

Public attitudes toward mental illness and people with mental illness more generally vary according to a number of internal factors such as personal experiences and/or contact with people with mental illness and external factors including the media, cultural beliefs, and legislation. Attitudes differ according to age, gender, socioeconomic status, and cultures. It can be argued that it is not the illnesses per se that provoke negative attitudes, but what they represent (i.e., 'the outsiderness') that may be seen as frightening and that may increase social distance. These attitudes may or may not always translate into behavior, but can influence the allocation of resources into the delivery of mental health services (P.W. Corrigan and Watson, 2003), recruitment into various branches of mental health professions (Mukherjee et al., 2013), and general acceptance of people with mental illness by the larger community (Hatzenbuehler et al., 2013).

Definitions

Mental illness is a broad generic term that includes major mental illnesses such as schizophrenia, bipolar affective disorders, common mental disorders such as anxiety and depression, personality disorders, and various other conditions sometimes named after etiology (for example, posttraumatic stress disorder) or after behaviors (e.g., episodic dyscontrol syndrome). Therefore, when discussing attitudes toward mental illness, it is important to keep this context in mind as different conditions may be associated with different attitudes. For instance, a national sample of British adults showed that more than 70% of the public thought people with schizophrenia were a danger to others; however, less than 10% felt that people with eating disorders were dangerous (CRISP, 2000). The public also differs in terms of how they classify and conceptualize mental illnesses and this can influence their attitudes toward people with mental illness too. One study identified a two-factor model of views of mental illness with one group identifying only major psychiatric disorders (e.g., schizophrenia and bipolar disorder) to be mental illnesses and the other group identifying drug addiction and external stressors as a mental illness, suggesting a broader and less specific concept of mental illness (Rüsch et al., 2012). These types of classifications of types of mental illness are related to attitudes toward people with mental illness and toward help-seeking. For instance, identifying major psychiatric disorders as mental illness is associated with more positive attitudes toward people with mental illness and a greater likelihood of help-seeking, though the broader perspective is associated with greater comfort in disclosing a mental health problem (Jorm et al., 2006b; Rüsch et al., 2012; Wright et al., 2012). Because of negative associations identified with various mental illness conditions and how they are named and/or classified, it has been suggested that changing the names of conditions and using terms which more clearly reflect recovery could be important for reducing stigma; but, it seems that the process is complex and the context which shapes attitudes and stigma is more important. Thus research which is available suggests that renaming in and of itself will not change stigma (Patrick W. Corrigan, n.d.).

Attitudes

Attitudes form a key point of an individual's thinking, especially toward things that matter to the individual. Attitudes have several components – some a function of the dimensions of personality traits, whereas others are a function of experience. Assessing attitudes provides a window into belief systems, but these do not always translate into behavior. Both cognitive and affective components of attitudes should be assessed, separately and together. Attitudes do not always remain static and are influenced markedly by a number of factors such as media, education, and societal conditions.

Stigma

Stigma is the expectation of stereotypical and discrediting judgment of oneself by others in a particular context. The term

originates from Goffman's (1963) definition of stigma, which states that stigma originates when an individual, because of some attributes, is disqualified from full social acceptance. Stigma, like the attributes, is deeply embedded in the sociocultural milieu in which it originates and is thus negotiable.

The Greeks in ancient times used the word 'stigma' to refer to bodily signs that exposed something unusual and negative about the signifier. Further conceptualization of stigma has included a mark that sets an individual apart (by virtue of mental illness in this case), that links that individual to some undesirable characteristic (perceived or real in cases of mental illness using violence, fear, and aggression) and rejection, isolation, or discrimination against the individual (e.g., at the workplace by virtue of the sick role and mental illness). Stigma thus includes cognitive and behavioral components and can be applied to age, race, sexual orientation, criminal behavior, and substance misuse, as well as to mental illness. The label and negative stereotypes are then extended to all those who have the characteristic. Social exclusion may be seen as necessary for bifurcation of society and will take the same role as scapegoating. Stigma and negative attitudes to mental illness are by and large universal. These are briefly discussed below to highlight similarities prior to discussing various factors responsible for these attitudes.

Attitudes in Europe

Negative attitudes toward people with mental illness are represented by the public and professionals in all countries and cultures and can be considered global phenomena (S. Evans-Lacko et al., 2012; Lasalvia et al., 2012; Schomerus et al., 2014; G. Thornicroft et al., 2009), however, variation in attitudes and beliefs are represented even within different countries of Europe. In this section some of the key studies and the findings are reported.

Historical

In ancient Greece, stigma and shame were interlinked and these can be studied using theories of causation and the balance within the culture between exclusion of individuals with mental illness and the ritual means of cleansing and re-inclusion. The interwoven theories of shame and stigma suggest that, in spite of scientific and religious explanations, attitudes were generally negative. Simon (1992) suggests that the ambivalence of the ancient Greeks can explain their attitudes toward mental illness. The diseases of the mind were seen in parallel with those of the body, and even the doctors were part and parcel of the same cultural ambience of competition, shame, and fear of failure. During the Medieval and Renaissance periods, mental illness was perceived as a result of imbalance of various humors of the body, and related concepts of shame contributed to this. As Christianity encouraged a God-fearing attitude to the world, mental illness was seen as a sign that God was displeased and punishing the individual, thereby making negative attitudes more likely and more acceptable.

In the seventeenth and eighteenth centuries the context shifted; but, overall remained negative and in turn provided entertainment to the visitors to the Bedlam asylum where individual visitors paid a penny to see the inmates. In the nineteenth century, with the establishment of various psychiatric asylums, there was generally a growing perception among the lay public and physicians alike that the mind is a function of the brain, thereby influencing a shift toward more humane treatments – although these were still called moral therapy, thereby giving it a moral/religious tinge. The Athenian thinking on psyche has influenced these attitudes. Porter (1987) proposed that the growing importance of science and technology (among other things) was influential in channeling the power of right-thinking people in imposing the social norms. The men of the church (also men of power) influenced public opinion, and informed attitudes and behavior toward marginal social elements that would then become disturbed and alien. Central state or market economy influenced the expectations which then divided those who set and met the norms from those who did not.

Modern Context

In a survey from 1957, the British Broadcasting Corporation (BBC, 1957) reported that the public's tolerance of individuals with mental illness depended upon the circumstances in which contact was made. More than three-quarters of those surveyed were willing to mix with individuals with mental illness in areas where low personal involvement occurred; only half were willing to work with people with mental illness, whereas only a quarter were agreeable to such an individual being in authority over them.

As mentioned previously, attitudes toward mental illness are influenced by a number of external factors such as sociopolitical context, expectations of treatment, etc. In Greece, one study showed that individuals who were younger, more educated, and of higher social class saw mental illness as a psychosocial problem, whereas older people saw it as a disorder of the nervous system (Lyketsos et al., 1985). Brändli (1999) observed that in Switzerland stigma was more common in older nonurban males with low education and low mental health literacy. The findings also suggested that the general public was more accepting of seeking help from their primary care physician compared to a psychiatrist. Indeed, several more recent studies have identified a link between stigma and intended (Angermeyer et al., 2013; Mojtabai, 2010; Rüsch et al., 2011; Ten Have et al., 2010) and actual help-seeking (S. Evans-Lacko et al., 2012a) across several European countries. Research from Germany suggests that in recent decades, the general public is much more likely to recommend help-seeking from mental health professionals; however, social distance in relation to people with mental illness has not decreased (Angermeyer et al., 2013).

Across Europe, there have also been comparative studies and research that pools and compares attitudes across European countries. Studies using the Eurobarometer data, including nationally representative samples across 27 EU countries, show that negative attitudes at the country level are associated with lower levels of help-seeking (Mojtabai, 2010), higher rates of suicide (Schomerus et al., 2014), and higher rates of unemployment for people with mental health problems (S. Evans-Lacko et al., 2013). For instance, we know that people with mental illness have higher rates of unemployment compared

to those without mental illness. One recent study showed that individuals with mental health problems living in countries with higher levels of stigmatizing attitudes regarding dangerousness of people with mental illness were more vulnerable to unemployment, particularly during the macroeconomic recession, in 2010 (S. Evans-Lacko et al., 2013b).

A recent systematic mapping of the literature that reviewed studies across Europe between 2007 and 2012 demonstrated a growing interest in stigma and social inclusion research via increased publications. Most studies included in the review were based in Northwest Europe (especially in the UK, Finland, Sweden, and Germany). Research tended to be more descriptive rather than focusing on mechanisms or interventions that could improve attitudes, and this is a gap that should be addressed in future research (Sara Evans-Lacko et al., 2014). There is some work that looks at the effectiveness of anti-stigma interventions for improving attitudes. More recent research from the UK has demonstrated improvements in attitudes toward people with mental illness; however, these occurred alongside concerted anti-stigma efforts. More positive attitudes in relation to prejudice and exclusion were particularly notes and these may be due to the effectiveness of the national Time to Change anti-stigma campaign (S. Evans-Lacko et al., 2013a,b; Evans-Lacko et al., 2014a,b).

The attitudes of other medical and nonmedical professionals are also important, but are beyond the scope of this chapter.

Attitudes in North America

In North America, there are a number of studies related to stigma carried out by medical and social scientists that have illustrated a range of attitudes to a number of conditions. Here only a subset of these studies will be discussed.

Historical

In America from the seventeenth century onward, the traditional belief prevailed among Christians that madness is often a punishment delivered by God on the sinner, and this influenced subsequent views and attitudes. Although both physical and mental illnesses were seen as punishments, it was only in the latter that moral components were seen, especially in the context of teachings from the Bible. A lack of faith and persistence of sinful and lewd thoughts were seen as key causation factors in the development of insanity.

In the early 1940s, Cumming and Cumming (1957), in their classic study of lay public perceptions of people with mental illness, showed significant rejection of people with mental illness. They performed one of the earliest educational interventions in a small community in Canada, aiming to improve stigmatizing attitudes; however, these attempts were unsuccessful and it is reported that attitudes became even more hostile.

In the 1940s, Ramsey and Seipp (1948) demonstrated that subjects who had higher socioeconomic and educational levels were less likely to view mental illness as a punishment for sins and/or poor living conditions. Yet they were less pessimistic about recovery.

Modern Context

Vignettes have been used to determine public knowledge of mental illness, and public attitudes toward people with mental illness for several decades (Star, 1952). Crocetti and Lemkau (1963) used vignettes to demonstrate that virtually all the subjects in their sample agreed with the question, 'Do you think that people who are mentally ill require a doctor's care as people who have any other sort of illness do?' In this sample, low social class and low educational levels were not associated with negative attitudes. In another study using case vignettes, it was shown that the largest increase in rejection rates occurred when a person had been admitted to a mental hospital and this was seen as due not to the fact that they were unable to help themselves, but that a psychiatrist or mental health professional had confirmed their status as 'mentally ill' (Phillips, 1963). The rejection appeared to be based on how visibly the behavior deviates from the customary role expectations. Continuance of symptoms, and public visibility of these symptoms, and behavior, influences attitudes and makes them more negative.

Social distance measures assess the degree to which individuals avoid or disassociate themselves with people with mental illness. Assessments have examined a number of parameters, especially in social interactions, but also the likelihood that individuals would employ or rent to people with mental illness (P.W. Corrigan, 2001; Link and Phelan, 2001). Additional research has shown that individuals may be willing to accept individuals with mental illness in certain contexts, but reluctant to interact with them in social situations. Dovidio et al. (1985) examined first impressions among undergraduate students of individuals with mental illness in the process of college applications. The results suggested that individuals with mental illness were seen favorably in terms of character and competence, but negatively in the context of security and sociability.

An important factor related to social distance is personal contact with people with mental illness. Analysis of a nationally representative survey of individuals in the USA showed that those who reported previous contact (defined here as a history of psychiatric hospitalization among the respondents themselves, their family, or their friends). In this study, previous contact was associated with 'having a less ostracizing, critical attitude toward a stranger with mental illness' (Boyd et al., 2010). Indeed a recent meta-analysis of studies looking at effects of anti-stigma approaches identified contact, particularly face-to-face contact, as having positive effects for reducing stigma (Patrick W. Corrigan et al., 2012).

Newer research suggests that we need to look specifically at types of stigmatizing attitudes and their relation to consequences, as different types of stigma may vary in their impact. For instance, recent research has shown that a belief that people with mental illness are dangerous is associated with greater likelihood of help-seeking, more support for coercive treatment (P. Corrigan et al., 2003; Mossakowski et al., 2011), and a biogenetic explanation for mental illness (Kvaale et al., 2013). It may be that a 'medicalized' concept of mental illness in which mental illness is considered 'a disease like any other,' (Pescosolido et al., 2010) is associated with greater help-seeking, but also more social distance.

Attitudes Elsewhere in the World

Although the majority of attitude studies have been performed in North America, Western Europe, and Australia, there are a growing number of studies carried out elsewhere.

Historical

In ancient Ayurvedic texts, the description of mental illness also suggested physical forms of treatment. Diet and seasons were reported as playing key roles in the genesis of such conditions. This approach allowed the locus of control to be shifted to some degree away from the individual, thereby making that individual less responsible and less likely to be stigmatized.

Current

Ilechukwu (1988), in a survey of 50 male and 50 female psychiatric outpatients in Lagos, demonstrated that some patients did believe in supernatural causes of neuroses and disorders, but psychosocial causes were cited most commonly. The attitudes and beliefs of patients are important, but are beyond the scope of this article.

Wig et al. (1980), in a three-site study from India, Sudan, and the Philippines, found that, when asked to comment on case vignettes, community leaders were able to identify mental retardation (in all sites), alcohol and drug-related problems (in the Sudanese and Philippines areas), and acute psychosis (in India). Thus it appears that there are differences in the identification of different clinical conditions. They also reported that attitudes toward people with mental illness were more negative in India compared with the other two sites. There has also been some investigation of stigma in relation to religion. Verghese and Beig (1974), in a survey from South India, reported that over 40% of Muslim and Christian respondents reported that marriage can help mental illness, although only 20% of Hindus shared this belief. Virtually no respondents reported believing in evil spirits. The most commonly recognized causes mentioned were emotional factors, including excessive thinking. In respondents over the age of 40, one-fifth saw mental illness as God's punishment. Christians were again more likely than Hindus to fear people with mental illness (although this fear diminished with higher levels of education), and yet Christians were more likely to believe strongly in the possibility of recovery. Nearly three-quarters of Hindus believed that the moon influences mental illness. These are interesting findings in that they address religious differences in public attitudes.

The number of studies in other parts of the world is likely growing. Previously, a gap was noted in terms of understanding stigmatizing attitudes and beliefs about mental illness in Africa, but there are now several studies that have been recently published (Barke et al., 2011; Crabb et al., 2012; Wakil et al., 2013). A recent review of mental health-related stigma in Japan identified 19 studies. Many of the conclusions in terms of what types of factors are associated with less stigma (e.g., personal contact) were similar to those found in other countries; for example, it is common to find that schizophrenia is associated with more stigma than depression. Studies did suggest, however, higher levels of stigma in Japan compared to Australia and Taiwan (Ando et al., 2013). A number of attitude studies and anti-stigma interventions have been performed in Australia, particularly in relation to evaluating the Beyond Blue anti-stigma campaign (Jorm et al., 2006a; Jorm et al., 2006b; Yap et al., 2012).

Characteristics Associated with Negative Attitudes

Reactions to a taboo or outsider group depend upon a number of factors. These include the frequency of the actual or anticipated behavioral events, intensity and visibility of such behavior, and circumstances and location of such behavior on the one hand, and personal factors on the other. In this section we focus on the latter.

Age

Several studies have shown that older people tend to have more negative attitudes toward people with mental illness. The reasons for this are many. Older people, in spite of their life experiences, are generally more conservative and equally rejecting of behavior which is seen as odd and alien. The role of age is likely to be mediated by other factors such as education, social and economic class, etc.

Types of Illness

As discussed previously, attitudes toward people with schizophrenia are likely to differ from those reported toward people with depression or personality disorder. This may reflect the stereotypic images of the condition or fear related to the condition. It may also be due to previous knowledge about the illness. The perceived causative/etiological factors also play a role in attitudes.

Gender

Males tend to have more negative attitudes and, as noted above, are also more likely to be rejected when they have a mental illness; females may be more sympathetic. Other studies of attitudes toward other minority groups also demonstrate that females are more positive. They are also more likely to be carers, and may be the first to contact psychiatric services on behalf of individuals. For women, a change in role after developing mental illness is likely to be associated with more stigma than is the case in men. Yet they are more likely to be admitted with more 'masculine' illnesses such as personality disorder and drug abuse. The gender roles in the context of illness may well play a role in generating negative attitudes.

Religious Beliefs

Some studies have demonstrated that Christians tend to have more negative attitudes toward mental illness, but again this is not a consistent finding. The individual subject's level of religiosity and depth of religious values must be studied rather than simple religious ascription.

Educational Status

The effects of education on attitudes are mixed. Some studies have identified negative attitudes to be associated with low educational status, whereas others have failed to show such association, or showed that highly educated subjects held more negative attitudes, including those with professional qualifications.

Professions

Medical students may hold more negative and stereotypical attitudes, and other branches of medicine too have been shown to have negative attitudes toward people with mental illness and psychiatry as a profession. The psychiatrist is often lampooned and seen as a 'head shrinker' or a modern-day witch doctor. There are individuals who are well educated and belong to a high socioeconomic class and yet hold negative attitudes.

Others

Some ethnic groups such as African-Caribbeans in the UK and Hispanics, Asian-Americans, and Mexican-Americans in the USA have been found to have more negative attitudes toward people with mental illness or be more restrictive in their description of etiology of mental illness compared to the white majority population. Non-acculturated individuals appear to have more negative stereotypes about people with mental illness compared to those who are considered to be acculturated.

These are some of the complex set of factors that influence attitudes and stereotypes.

Interventions

There are three main strategies for improving stigmatizing attitudes toward people with mental illness described in the literature: protest, education, and contact (P.W. Corrigan et al., 2001). Protest aims to eliminate negative stereotypes in public statements, media reports, or advertisements (Rüsch et al., 2005). Education provides evidence-based information about mental illness or demonstrates how negative attitudes are represented in society, and what the consequences of this might be. The third strategy involves direct contact with people with a psychiatric disorder and it has been recognized as one of the most effective ways to fight stigma and discrimination (P W Corrigan et al., 2012; Sara Evans-Lacko et al., 2012b). Social contact is associated with a reduction in anxiety and also provides an opportunity to check '*in vivo*' the invalidity of one's own prejudices (S. Evans-Lacko et al., 2012b). Although contact has the strongest level of evidence, multifaceted strategies are encouraged and it is important to target anti-stigma work based on the type we seek to eradicate.

Wolff et al. (1999) reported on the results of follow-up of neighbors, where two group homes for people with mental illness had been set up and one set of neighbors had received extensive education. They found that although the public education intervention may have had at best only a modest effect on knowledge, behavior toward the residents with mental illness changed. There was a decrease in fear and exclusion, and increased levels of social contact in the experimental area. They observed that educational intervention per se did not in itself lead directly to a reduction in fearful attitudes, whereas contact with patients did. Thus, any campaign that encourages subjects to increase contact with people with mental illness may prove to be more successful. The educational programs must be paced slowly and in a sustainable manner. The patients and their carers must be involved in planning these interventions without turning the whole exercise into a circus.

Anti-stigma interventions aimed at improving attitudes may, in the long run, influence resource allocation into mental health, improve recruitment, and move toward acceptance of community care. These interventions are also important to consider as they may reduce fear and prevent attitudes from hardening. However, these interventions must be clear, focused, and appropriate, based upon the needs of the group that is being educated rather than on perceptions of need by the professionals. Corrigan has identified five principles for strategic stigma change: consumer contact that is targeted, local, credible, and continuous (Patrick W. Corrigan, 2011). Several large-scale evaluations have now been published (Dunion and Gordon, 2005; S. Evans-Lacko et al., 2010; S. Evans-Lacko et al., 2013; Svensson and Hansson, 2014; C. Thornicroft et al., 2014); however, further evaluation of anti-stigma programs is required to better understand the key factors, mechanisms, and pathways to improve attitudes in a sustained way.

Conclusions

The attitudes of the public toward individuals with mental illness reflect prevalent social norms and mores. Expectations from psychiatric services, response to treatment, type of treatment, type of mental illness, and inaccessibility to treatment are external factors that will influence public attitudes. Age, gender, personality traits, and social and economic class are associated with different attitudes. These attitudes will not always translate into behavior, but are interrelated and have significant consequences for people with mental illness, leading to exclusion and disadvantage, for example, through reduced participation in higher education, exclusion from civil society (including functions such as democratic participation), increased risk of contact with criminal justice systems, victimization, less access to physical health care, poverty and homelessness, and reduced life expectancy. As we further our understanding of the mechanisms that facilitate and maintain negative stigmatizing attitudes through more robust evaluation, we also need to further develop and refine our anti-stigma strategies.

See also: Attitude Formation and Change; Attitudes and Behavior; Discrimination and the Law; Foucault, Michel (1926–84); Health and Illness: Mental Representations in Different Cultures; Mental Illness: Family and Patient Organizations in Europe.

Bibliography

Ando, S., Yamaguchi, S., Aoki, Y., Thornicroft, G., 2013. Review of mental-health-related stigma in Japan. Psychiatry and Clinical Neurosciences 67 (7), 471–482. http://dx.doi.org/10.1111/pcn.12086.

Angermeyer, M.C., Matschinger, H., Schomerus, G., 2013. Attitudes towards psychiatric treatment and people with mental illness: changes over two decades. The British Journal of Psychiatry: The Journal of Mental Science 203 (2), 146–151. http://dx.doi.org/10.1192/bjp.bp.112.122978.

Barke, A., Nyarko, S., Klecha, D., 2011. The stigma of mental illness in Southern Ghana: attitudes of the urban population and patients' views. Social Psychiatry and Psychiatric Epidemiology 46 (11), 1191–1202. Retrieved from: http://link.springer.com/article/10.1007/s00127-010-0290-3/fulltext.html.

BBC, 1957. The Hurt Mind: An Audience Research Report. London.

Boyd, J.E., Katz, E.P., Link, B.G., Phelan, J.C., 2010. The relationship of multiple aspects of stigma and personal contact with someone hospitalized for mental illness, in a nationally representative sample. Social Psychiatry and Psychiatric Epidemiology 45 (11), 1063–1070. http://dx.doi.org/10.1007/s00127-009-0147-9.

Brändli, H., 1999. The image of mental illness in Switzerland. In: Guimon, N., Fischer, M.J., Sartorius, W. (Eds.), The Image of Madness. Karger, Basel.

Corrigan, P., Markowitz, F.E., Watson, A., Rowan, D., Kubiak, M.A., 2003. An attribution model of public discrimination towards persons with mental illness. Journal of Health and Social Behavior 44 (2), 162–179. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/12866388.

Corrigan, P.W., n.d. Erasing Stigma Is Much More than Changing Words. Psychiatric Services.

Corrigan, P.W., 2001. Familiarity with and social distance from people who have serious mental illness. Psychiatric Services 52 (7), 953–958. http://dx.doi.org/10.1176/appi.ps.52.7.953.

Corrigan, P.W., 2011. Best practices: strategic stigma change (SSC): five principles for social marketing campaigns to reduce stigma. Psychiatric Services (Washington, D.C.) 62 (8), 824–826. http://dx.doi.org/10.1176/appi.ps.62.8.824.

Corrigan, P.W., Morris, S.B., Michaels, P.J., Rafacz, J.D., Rüsch, N., 2012. Challenging the public stigma of mental illness: a meta-analysis of outcome studies. Psychiatric Services (Washington, D.C.) 63 (10), 963–973. http://dx.doi.org/10.1176/appi.ps.201100529.

Corrigan, P.W., River, L.P., Lundin, R.K., Penn, D.L., Uphoff-Wasowski, K., Campion, J., Kubiak, M.A., 2001. Three strategies for changing attributions about severe mental illness. Schizophr.Bull. 27, 187–195 (0586–7614 (Print)).

Corrigan, P.W., Watson, A.C., 2003. Factors that explain how policy makers distribute resources to mental health services. Psychiatr.Serv 54, 501–507 (1075–2730 (Print)).

Crabb, J., Stewart, R.C., Kokota, D., Masson, N., Chabunya, S., Krishnadas, R., 2012. Attitudes towards mental illness in Malawi: a cross-sectional survey. BMC Public Health 12 (1), 541. http://dx.doi.org/10.1186/1471-2458-12-541.

CRISP, A.H., 2000. Stigmatisation of people with mental illnesses. The British Journal of Psychiatry 177 (1), 4–7. http://dx.doi.org/10.1192/bjp.177.1.4.

Crocetti, G.M., Lemkau, P.V., 1963. Public opinion of psychiatric home care in an urban area. American Journal of Public Health 53, 409–417.

Cumming, E., Cumming, J., 1957. Closed Ranks: An Experiment in Mental Health Education. Harvard University Press, Cambridge, MA.

Dovidio, J.F., Fishbane, R., Sibicky, M., 1985. Perceptions of people with psychological problems: effects of seeking counseling. Psychological Reports 57 (3 Pt 2), 1263–1270. http://dx.doi.org/10.2466/pr0.1985.57.3f.1263.

Dunion, L., Gordon, L., 2005. Tackling the attitude problem. The achievements to date of Scotland's "see me" anti-stigma campaign. Mental Health Today (Brighton, England) 22–25. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/15819531.

Evans-Lacko, S., Brohan, E., Mojtabai, R., Thornicroft, G., 2012a. Association between public views of mental illness and self-stigma among individuals with mental illness in 14 European countries. Psychological Medicine 1–13 (1469–8978 (Electronic)).

Evans-Lacko, S., London, J., Japhet, S., Rüsch, N., Flach, C., Corker, E., Thornicroft, G., 2012b. Mass social contact interventions and their effect on mental health related stigma and intended discrimination. BMC Public Health. http://dx.doi.org/10.1186/1471-2458-12-489.

Evans-Lacko, S., Courtin, E., Fiorillo, A., Knapp, M., Luciano, M., Park, A., Brunn, M., Byford, S., Chevreul, K., Forsman, A.K., Gulacsi, L., Haro, J.M., Kennelly, B., Knappe, S., Lai, T., Lasalvia, A., Miret, M., O'Sullivan, C., Obradors-Tarragó, C., Rüsch, N., Sartorius, N., Švab, V., van Weeghel, J., Van Audenhove, C., Wahlbeck, K., Zlati also on behalf of the ROAMER Consortium, A., McDaid, D., Thornicroft, G., 2014a. The state of the art in European research on reducing social exclusion and stigma related to mental health: a systematic mapping of the literature. European Psychiatry 29, 381–389. http://dx.doi.org/10.1016/j.eurpsy.2014.02.007.

Evans-Lacko, S., Henderson, C., Thornicroft, G., 2013a. Public knowledge, attitudes and behaviour regarding people with mental illness in England 2009–2012. The British Journal of Psychiatry 202 (s55), s51–s57.

Evans-Lacko, S., Knapp, M., McCrone, P., Thornicroft, G., Mojtabai, R., 2013b. The mental health consequences of the recession: economic hardship and employment of people with mental health problems in 27 European countries. PLoS One 8, e69792 (1932–6203 (Electronic)).

Evans-Lacko, S., Malcolm, E., West, K., Rose, D., London, J., Rusch, N., Thornicroft, G., 2013c. Influence of time to change's social marketing interventions on stigma in England 2009–2011. British Journal of Psychiatry Supplement 55, s77–s88 (0960–5371 (Print)).

Evans-Lacko, S., London, J., Little, K., Henderson, C., Thornicroft, G., 2010. Evaluation of a brief anti-stigma campaign in Cambridge: do short-term campaigns work? BMC.Public Health 10, 339 (1471–2458 (Electronic)).

Evans-Lacko, S., Corker, E., Williams, P., Henderson, C., Thornicroft, G., 2014b. Trends in mental illness related public stigma among the English population in 2003–2013: influence of the time to change anti-stigma campaign. Lancet Psychiatry 1 (2), 121–128.

Goffman, E., 1963. Stigma: Notes on the Management of Spoiled Identity. Prentice-Hall, Englewood Cliffs, NJ.

Hatzenbuehler, M.L., Phelan, J.C., Link, B.G., 2013. Stigma as a fundamental cause of population health inequalities. American Journal of Public Health 103, 813–821 (1541–0048 (Electronic)).

Ilechukwu, S., 1988. Interrelationship of beliefs about mental illness psychiatric diagnosis and mental health care delivery among Africans. International Journal of Social Psychiatry 34, 200–206.

Jorm, A.F., Christensen, H., Griffiths, K.M., 2006a. Changes in depression awareness and attitudes in Australia: the impact of beyondblue: the national depression initiative. Australian and New Zealand Journal of Psychiatry 40 (1), 42–46. http://dx.doi.org/10.1080/j.1440-1614.2006.01739.x.

Jorm, A.F., Christensen, H., Griffiths, K.M., 2006b. The public's ability to recognize mental disorders and their beliefs about treatment: changes in Australia over 8 years. The Australian and New Zealand Journal of Psychiatry 40 (1), 36–41. http://dx.doi.org/10.1111/j.1440-1614.2006.01738.x.

Kvaale, E.P., Gottdiener, W.H., Haslam, N., 2013. Biogenetic explanations and stigma: a meta-analytic review of associations among laypeople. Social Science & Medicine (1982) 96, 95–103. http://dx.doi.org/10.1016/j.socscimed.2013.07.017.

Lasalvia, A., Zoppei, S., Van, B.T., Bonetto, C., Cristofalo, D., Wahlbeck, K., Thornicroft, G., 2012. Global pattern of experienced and anticipated discrimination reported by people with major depressive disorder: a cross-sectional survey. Lancet (1474–547X (Electronic)).

Link, B., Phelan, J., 2001. Conceptualizing stigma. Annual Review of Sociology 27, 363–385.

Lyketsos, G., Mouyas, A., Malliori, M., Lyketsos, C., 1985. Opinion of public and patients about mental illness and psychiatric care in Greece. British Journal of Clinical and Social Psychiatry 3, 59–66.

Mojtabai, R., 2010. Mental illness stigma and willingness to seek mental health care in the European Union. Social Psychiatry and Psychiatric Epidemiology 45, 705–712 (1433–9285 (Electronic)).

Mossakowski, K.N., Kaplan, L.M., Hill, T.D., 2011. Americans' attitudes toward mental illness and involuntary psychiatric medication. Society and Mental Health 1 (3), 200–216. http://dx.doi.org/10.1177/2156869311431100.

Mukherjee, K., Maier, M., Wessely, S., 2013. UK crisis in recruitment into psychiatric training. The Psychiatrist 37 (6), 210–214. http://dx.doi.org/10.1192/pb.bp.112.040873.

Pescosolido, B.A., Martin, J.K., Long, J.S., Medina, T.R., Phelan, J.C., Link, B.G., 2010. "A disease like any other"? A decade of change in public reactions to schizophrenia, depression, and alcohol dependence. American Journal of Psychiatry 167 (11), 1321–1330.

Phillips, D., 1963. Rejection of the mentally ill: the influence of behavior and sex. American Sociological Review 29, 679–689.

Porter, R., 1987. Mind Forg'd Manacles. Athlone, London.

Ramsey, G., Seipp, M., 1948. Attitudes and opinions concerning mental illness. Psychiatric Quarterly 22, 428–444.

Rüsch, N., Angermeyer, M.C., Corrigan, P.W., 2005. Mental illness stigma: concepts, consequences, and initiatives to reduce stigma. European Psychiatry 20 (8), 529–539. http://dx.doi.org/10.1016/j.eurpsy.2005.04.004.

Rüsch, N., Evans-Lacko, S.E., Henderson, C., Flach, C., Thornicroft, G., 2011. Knowledge and attitudes as predictors of intentions to seek help for and disclose a mental illness. Psychiatric Services 62 (6), 675–678. Retrieved from: http://journals.psychiatryonline.org/article.aspx?articleid=116233.

Rüsch, N., Evans-Lacko, S., Thornicroft, G., 2012. What is a mental illness? Public views and their effects on attitudes and disclosure. The Australian and New Zealand Journal of Psychiatry 46 (7), 641–650. http://dx.doi.org/10.1177/0004867412438873.

Schomerus, G., Evans-Lacko, S., Rüsch, N., Mojtabai, R., Angermeyer, M.C., Thornicroft, G., 2014. Collective levels of stigma and national suicide rates in 25 European countries. Epidemiology and Psychiatric Sciences 1–6. http://dx.doi.org/10.1017/S2045796014000109.

Simon, B., 1992. Shame, stigma and mental illness in ancient Greece. In: Tasman, A., Fink, P.J. (Eds.), Stigma and Mental Illness. Washington DC.

Star, S.A., 1952. What the public thinks about mental health and mental illness. In: Paper Presented at the Annual Meeting of the National Association of Mental Health. Indiana, Indianapolis.

Svensson, B., Hansson, L., 2014. Effectiveness of mental health first aid training in Sweden. A randomized controlled trial with a six-month and two-year follow-up. PloS One 9 (6), e100911. http://dx.doi.org/10.1371/journal.pone.0100911.

Ten Have, M., de Graaf, R., Ormel, J., Vilagut, G., Kovess, V., Alonso, J., 2010. Are attitudes towards mental health help-seeking associated with service use? Results from the European study of epidemiology of mental disorders. Social Psychiatry and Psychiatric Epidemiology 45 (2), 153–163. http://dx.doi.org/10.1007/s00127-009-0050-4.

Thornicroft, C., Wyllie, A., Thornicroft, G., Mehta, N., 2014. Impact of the "like minds, like mine" anti-stigma and discrimination campaign in New Zealand on anticipated and experienced discrimination. The Australian and New Zealand Journal of Psychiatry 48 (4), 360–370. http://dx.doi.org/10.1177/0004867413512687.

Thornicroft, G., Brohan, E., Rose, D., Sartorius, N., Leese, M., 2009. Global pattern of experienced and anticipated discrimination against people with schizophrenia: a cross-sectional survey. Lance 373 (9661), 408–415. http://dx.doi.org/10.1016/S0140-6736(08)61817-6.

Verghese, A., Beig, A., 1974. Public attitudes towards mental illness: the Vellore study. Indian Journal of Psychiatry 16, 8–18.

Wakil, M., Abdul, I., Jidda, M., Beida, O., Ahidjo, A., Salawu, F., Abdulmalik, J., 2013. Attitude towards psychiatric treatment and referral pattern in the university of Maiduguri teaching hospital – a preliminary report. African Journal of Psychiatry 16 (3), 182–186. http://dx.doi.org/10.4314/ajpsy.v16i3.24.

Wig, N.N., Suleiman, M.A., Routledge, R., Srinivasa, R., Ladrido-Ignacio, L., Ibrahim, H.H.A., Harding, T.W., 1980. Community reactions to mental disorders? A key informant study in three developing countries. Acta Psychiatrica Scandinavica 61 (2), 111–126. http://dx.doi.org/10.1111/j.1600-0447.1980.tb00571.x.

Wolff, G., Pathare, S., Craig, T., Left, J., 1999. Public education for community care. In: Guimon, J., Fischer, W., Sartorious, N. (Eds.), The Image of Madness. Karger, Basle, Switzerland.

Wright, A., Jorm, A.F., Mackinnon, A.J., 2012. Labels used by young people to describe mental disorders: which ones predict effective help-seeking choices? Social Psychiatry and Psychiatric Epidemiology 47 (6), 917–926. http://dx.doi.org/10.1007/s00127-011-0399-z.

Yap, M.B., Reavley, N.J., Jorm, A.F., 2012. Associations between awareness of beyondblue and mental health literacy in Australian youth: results from a national survey. The Australian and New Zealand Journal of Psychiatry 46 (6), 541–552. http://dx.doi.org/10.1177/0004867411435288.

Mentoring: Supervision of Young Researchers

Rachelle Hollander, National Academy of Engineering, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

More and tighter coupling between academic and other social institutions has prompted increased attention to ethical aspects of the relationships between established researchers and researchers-in-training, including graduate students and new researchers in the United States and other countries. Questions about the responsibilities of individual and institutional parties to these relationships arise, and persons and institutions try to respond. This article reviews recent guidance that scholars and professional groups have developed, particularly in the United States, and ends with suggestions for departments and institutions that can help them fulfill requirements for appropriate mentoring and supervision.

Context

The US National Science Foundation *Science and Engineering Indicators 2012* (*SEI 2012*) reports that US institutions granting doctoral degrees awarded 41 100 in fields of science (including psychology and social sciences) and engineering in 2009. Foreign students earned 57% of all engineering doctorates, 54% of all computer science degrees, and 51% of physics doctoral degrees. Their overall share of S&E degrees was one-third. *SEI 2012* also reports that in sum about 194 000 S&E doctoral degrees were earned worldwide in 2008. In 2007, China overtook the United States as the world leader in the number of doctoral degrees awarded in the natural sciences and engineering. Standing at the peak of a large educational enterprise which represents a large social investment, these numbers indicate the large-scale context in which post-baccalaureate education operates (National Science Board, 2012).

Given the scale of these investments and the importance for society of results from research activities using graduate and postdoctoral assistance, it is not surprising that questions arise concerning the roles and responsibilities of graduate students, postdoctoral associates, faculty, and their institutions. The system shows the strains of increasing numbers as well as mounting calls for accountability from graduate students themselves and the governmental funders and sponsors of graduate education. In the United States, graduate students have begun to organize and gain collective bargaining rights on campuses and postdoctoral associates are pushing for national reforms to their status, including written contracts, uniform job titles and benefits, support for postdoctoral associations, representation on institutional policy-making committees, and a national postdoctoral organization.

In the past decade national and international councils have begun to address issues of ethics and mentoring in graduate education and associated research activities (InterAcademy Council, 2012). Adequate response to these stresses requires the attention of academic institutions and professional associations. Most important, perhaps, are the responses at the level of departments and programs actually producing new Ph.D.s. The beginnings of change can be seen in the developments described below; but better understanding of changes in mentoring practice and supervision requires systematic research attention, of a kind they still have not received, in order to identify and assess effects on individuals, departments, programs, and institutions, and to make improvements in the future.

Increased social and governmental attention to questions of research misconduct has given particular impetus for change in this arena. These concerns resulted in regulations in the US requiring institutions to establish procedures in response to allegations, and many other nations have implemented such procedures. Governmental and academic institutions, as well as scientific and professional associations, regard this as a broader mandate for education in research integrity. The US National Institutes of Health has established requirements for ethics training for all research staff, which have implications for mentoring and supervision.

The US National Science Foundation has issued requirements for mentoring postdoctoral fellows supported with its funds, and for ethics training for graduate students and postdocs.

Professional associations have been custodians of good practice for their members, as well as for practitioners and researchers who may not be members but draw on the body of knowledge that these associations incorporate and represent themselves as having the relevant expertise. Today's world requires that experts work with others with very different specialties more often than in the past; the need for these group or team efforts raises additional issues for good mentoring and supervision and for the effective review of research results.

Supervision

The US Council of Graduate Schools issued guidance for supervisory practice in 1990 (Council of Graduate Schools, 1990) and has followed up with the Project on Scholarly Integrity that published *Research and Scholarly Integrity in Graduate Education: A Comprehensive Approach* (Kent et al., 2012). The 2012 publication articulated a strong rationale for the need for institutions to layer their response to enhancing scholarly integrity by establishing efforts throughout the institutional hierarchy. The earlier guidance indicated that there are no recipes for the creative impetuses that underlie the most successful interactions between faculty members and students,

and provided useful recommendations to assist departments, programs, and supervisors to enable their students to make good progress. These recommendations are appropriate and can be adapted easily for use with postdoctoral fellows.

The guidance indicates that departments should have a clear written framework outlining the stages students are expected to complete in graduate training. Supervisors should help their students plan their time carefully, meet with them regularly, and be ready, at an early stage in the program, to make an assessment, which will be perceived to be fair and appropriate, about whether they should continue toward the Ph.D. Supervisors should engage the students early in research activities, in order to promote knowledge about the field and facilitate the selection of a dissertation topic. In research endeavors for graduate students, students should be discouraged from achieving perfect results at the expense of good research and should not be allowed to become distracted by interesting, but tangential, issues.

In readying students to undertake research, the department or program should clarify the standards it expects students to meet. Will a literature survey be necessary? Will experimental research or fieldwork be required? Are there certain techniques or tools that the student must master in the training program or dissertation? What writing standards must reports and theses meet? In the middle stages of training, graduate students should expect to complete most of the original work that will form the dissertation. Since research may lead in unexpected directions, supervisors and students need to be alert to considerations that might require modifying the original goals, particularly to allow completion of the project in a timely fashion. Careful records need to be kept, and here supervisory assistance and monitoring is most important and lays the foundations for good research practice throughout a career.

Team research may require additional considerations. There needs to be a careful definition of students' contribution to research work and the students should be given opportunities to demonstrate a broad understanding of project purposes and methods. Of course, students in individual projects also need opportunities for presentation about their status and findings. Writing should begin early in the project. Milestones here are useful. An introduction or overview as well as a list of references can be drafted and revised as work progresses. Questions about interim findings and the relation of the research to the work of others need to be posed so that students can stay abreast of developments in their fields. Experience demonstrates that students should be allowed to complete their dissertations before being asked to continue as a postdoctoral research associate.

Checklist for Supervisors and Departments

Does the department have a document describing its views on good supervisory practice? What steps exist to make good matches between supervisors and prospective students or postdoctoral associates? Should students and postdoctoral associates present reports for assessment by nonsupervisors early in their tenure? Do they make public presentations satisfactorily? Do they participate in developing proposals for research support? Are adequate meetings scheduled between supervisors and students or postdoctoral associates? Are there regular assessments of progress and background knowledge? Do supervisees and supervisors see the assessment procedures as satisfactory? For students, how are research topics refined early in training? For postdoctoral associates, how are research projects assessed to provide adequate stimulation and chance for advancement? Have appropriate times been scheduled for developing long-term programs of research and for identifying milestones? Do supervisors check research records for accuracy, quality, and completeness?

Checklist for Students

Have you developed a systematic work-plan and identified major difficulties and relevant references? Are your records in good order, so that you can find the relevant material about a problem you worked on 6 months ago? Do you have draft portions of any completed aspects of the research? Have you adequate opportunities for presentations about your work? Do others find your written work difficult or easy to understand? Have interim projects, such as figures, tables, or other matter, been identified for preparation during the project?

Mentoring

In 1997, the National Academy of Sciences published a handbook for faculty who want to develop good mentoring skills. It points out that mentoring is a personal, as well as a professional, relationship. Good mentors want to help their students get the most from their education, assist their socialization into the field, and aid them in finding employment. Students and postdoctoral associates benefit from having many mentors with different skills and priorities.

Successful mentors today must deal with issues involving ethnicity, culture, sex, and disability. Currently, women account for about half of the doctorates earned in the United States, while about 12% of the graduate enrollment is from US groups that are underrepresented in science and engineering. Non-US citizens make up about a third of the new science and engineering doctorates. Students may need access to a number of different faculty or graduate student peers for assistance. This may be particularly important when they come from different cultural or minority backgrounds. As faculty advisors, mentors must build mutual respect with the students and fellows whom they advise and supervise. Besides recognizing how issues of diversity may affect students' priorities and perceptions, faculty need to recognize that their students are likely to pursue careers outside academe and are legitimately concerned about time to degree and funding for their graduate education.

Junior faculty are also often in need of mentoring. They have the stresses created by considerably increased responsibilities for teaching and research and for professional service. Departments can establish numerous, albeit simple, ways to ease this transition, through appropriate workshops or by encouraging senior faculty to pair with junior colleagues. Departments can provide clear written guidance about tenure requirements and teaching policies and facilitate the acquisition of necessary resources for research, as well as nonresearch resources, such as childcare, which may be very important for entering faculty. They should arrange for frequent feedback

and evaluation. Minority and women faculty in particular may need protection from demands on their time created by the lack of others of their race or gender in their departments.

Students and postdoctoral associates often need assistance in mapping out career plans and in being introduced to contacts that may be able to help them. Students should be reminded that prospective employers also expect candidates to have other skills in addition to their competence in a chosen research field. Sometimes, mentors might need to provide frank and honest advice that identifies weaknesses but also indicates areas of strength that could be further enhanced. Over the course of the training, areas of weakness can also be corrected. Practice at teaching or making presentations, for instance, or critiques of draft grant proposals and resumes as well as research reports offer opportunities for improvement. Areas of strength might provide the basis for a good letter of recommendation from the mentor to the appropriate potential employer.

There are many ways that departments and institutions can improve the quality of mentoring, all requiring systematic and ongoing attention. Departments and institutions can track the progress of former students for information about careers. They can collect other relevant data or ask advanced students to assess how well their mentors or others have contributed to their progress. Institutions can schedule special activities and provide instructions on mentoring for new faculty and advisors. They can monitor faculty abuses and take a more active role in selecting faculty for advisory roles. They can provide workshops and discussions for students. They can encourage electives and other classes to broaden skills of students and postdoctoral associates.

All discussions of good mentoring and supervisory practice indicate that attention to these matters is imperative if trainees are to emerge with an intellectual and ethical framework that promotes research integrity and with some degree of satisfaction with their training experiences. Supervisors and mentors transmit to their students many of the attitudes and information needed to carry out significant research and good teaching and to practice responsible research. Faculty must recognize the need for human relationships to be built on respect and trust, as well as the special issues raised by the vulnerability of trainees.

In selecting mentors, students need to consider their publication records and national recognition, history of finding support for trainees, availability during the period of the student's training, prior training record and positions of recent graduates, practices of recognition for student accomplishments, and laboratory or other research arrangements. Students and postdoctoral associates should also recognize that mentors have different personalities. Some take an active interest and become involved in student projects; some are disengaged; some may develop friendships with their students and fellows. Each can pose difficulties for both parties. Open communication needs to be encouraged so that issues and potential misunderstandings can be clarified and any changes that might be necessary can be initiated in timely fashion.

To produce good work, mentors and trainees also need to be educated in the research standards of their fields. They need to be familiar with relevant codes of ethics. In experimental work, standards of laboratory practice or laboratory safety need to be discussed. This also holds true for standards of fieldwork and survey practice. Regulatory requirements concerning animal and human subjects need to be well understood. Standards for ethical practice of research with human subjects include matters of privacy and confidentiality, consent, and community protection. With regard to research in the social and behavioral sciences, complex questions concerning deception or stigmatization of individuals or communities may need particular care, especially in the context of research on vulnerable populations in other cultures. Mentors also bear special responsibilities when questions of the safety of students or postdoctoral associates might arise. All persons doing research should be familiar with definitions of research misconduct and the policies of their program and institution. Students must be trained in responsible use of statistical methods, where appropriate, and in the importance of stewardship of records and data, as well as in respect for ownership of ideas and intellectual property. Standards for responsible authorship and peer review also need to be understood and discrepancies that may exist between different fields need to be examined. These are areas where departments and institutions as well as professional associations and scientific societies have much to offer.

The study of science and engineering in society has begun to examine the social responsibility of science, engineering, and technology, focusing on the ethical assumptions and consequences that come in any framing of research questions, stipulations of methods and procedures, and confidence in results. This examination is likely to add an additional layer of complexity to discussions of responsible and responsive mentoring and supervision (Schienke et al., 2011).

Conclusion

For persons in academic research environments to achieve good results, they require and deserve assurance that others will do their share and that they will be protected if difficult problems arise. To foster this ethic, graduate departments or groups should develop adequate written statements of the terms of graduate study, including policies about conflict of interest, and should assign necessary duties to individual faculty members. These statements should be distributed to trainees. Individual faculty often have responsibilities for the admission of students, which allows them to assume similar responsibilities for advising students and postdoctoral associates and for placing graduate students in postdoctoral positions and jobs. Tasks that can help graduate students and postdoctoral associates succeed – such activities as writing proposals or research reports – can also be shared among faculty, as can assignments for making students aware of ethical aspects of research and professional activities.

Departments should create systems that specify advisors' duties and methods for evaluating their performance. Recent publications and guidance from national and international bodies and professional associations, as well as books and articles by individual scholars, can be helpful in developing materials and activities responsive to the needs of new faculty, postdoctoral researchers, and graduate students. Training for advisors continues to be needed. Increased attention to

structural reforms and ethical research practice might also lead to the diminution of research misconduct.

See also: Ethical Practices: Institutional Oversight and Enforcement; Privacy of Individuals in Social Research: Confidentiality; Professions and Professionalization, History of; Research Ethics, Cross-Cultural Dimensions of; Research Ethics: Research; Research Funding: Ethical Aspects; Research Publication: Ethical Aspects ; Surveys and Polling: Ethical Aspects; Universities and Science and Technology: United States.

Bibliography

Buchanan, A., 1996. Perfecting imperfect duties: collective action to create moral obligations. Business Ethics Quarterly 6, 27–42.

Carlin, D., Denecke, D., 2008. Best Practices in Graduate Education for the Responsible Conduct of Research. Council of Graduate Schools, Washington, DC.

Committee on Science, Engineering, and Public Policy, 2009. On Being a Scientist: A Guide to Responsible Conduct in Research. The National Academies, third ed. National Academies Press, Washington, DC.

Council of Graduate Schools, 1990. Research Student and Supervisor: An Approach to Good Supervisory Practice. Council of Graduate Schools, Washington, DC.

Hollander, R.D., 1999. Guaranteeing good scientific practice in the US. In: Wolfrum, R. (Ed.), Ethos of Research. Max-Planck-Gesellschaft, Munich, Germany, pp. 199–212.

Hollander, R., Arenberg, C. (Eds.), 2009. Ethics Education and Scientific and Engineering Research: What's Been Learned? What Should Be Done?. National Academies Press, Washington, DC.

InterAcademy Council, September 2012. Responsible Conduct in the Global Research Enterprise: A Policy Report. IAP and IAC Secretariats, The Netherlands.

Kent, J.D., Denecke, D.D., Allum, J., 2012. Research and Scholarly Integrity in Graduate Education: A Comprehensive Approach. Council of Graduate Schools, Washington, DC.

LaPidus, J.B., November 1997. Doctoral education: preparing for the future. CGS Communicator XXX (10), 1–10. Special ed.

Macrina, F.L., 1995. Mentoring. In: Scientific Integrity: An Introductory Text with Cases. ASM Press, Washington, DC, pp. 15–39.

Mervis, J., et al., 1999. Postdocs working for respect. Science 285 (5433), 1513–1533.

National Academy of Science, National Academy of Engineering, National Academy of Medicine, Committee of Science, Engineering, and Public Policy, 1997. Advisor, Teacher, Role Model, Friend: On Being a Mentor to Students in Science and Engineering. National Academy Press, Washington, DC.

National Science Board, 2012. Science and Engineering Indicators 2012. National Science Foundation (NSB12-01), Arlington, VA.

North Carolina State University Resource Guide in Research Ethics. http://www.fis.ncsu.edu/grad/ethics/resource.htm.

Schienke, E., et al., 2011. Intrinsic ethics regarding integrated assessment models for climate management. Science and Engineering Ethics 17 (3), 503–523.

Sieber, J.E., 1992. Planning Ethically Responsible Research: A Guide for Students and Internal Review Boards. Sage, Newbury Park, CA.

Upham, S., October 10, 1995. University of Oregon guidelines for good practice in graduate education. CGS Communicator XXVIII, 10–12.

Weil, V., 1999. Mentoring: Some Ethical Considerations. Chicago: Illinois Institute of Technology. Unpublished Paper, 17 pp., available from Professor Vivian Weil, Center for the Study of Ethics in the Professions.

Zigmond, M.J., 1999. Promoting responsible conduct: striving for change rather than consensus. Science and Engineering Ethics 5, 219–228.

Meritocracy and Tokenism

Stephen C Wright and Gregory D Boese, Simon Fraser University, Burnaby, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Meritocracy is relevant to a variety of influential social psychological frameworks: distributive justice, social identity theory, ideology, and inequality. Equity serves as the foundational justice principle for meritocracy. Social identity theory's concepts of boundary permeability and individual mobility reflect key aspects of meritocracy. Several theoretical perspectives explore the role of meritocracy beliefs in justifying and maintaining inequality. Contrasting the individual mobility of meritocracy is the highly restricted context of tokenism. Allowing individual mobility for a few disadvantaged group tokens can maintain intergroup inequality as tokens are quickly co-opted and the disadvantaged group abandons disruptive collective action for more benign individual responses.

Meritocracy

The concept of meritocracy can be found in a variety of highly influential social psychological frameworks. It is reflected in the distributive justice literature in discussions of the principle of *equity* (Deutsch, 1975; Tyler and Smith, 1998). It is also clearly implicated Social Identity Theory's (SIT) (Tajfel and Turner, 1979) concept of *individual social mobility*. Finally, it is named as an ideological belief that can serve to legitimize inequality between individuals and groups within a society (e.g., systems justification theory, Jost and Banaji, 1994). In this article, we will consider each of these manifestations of the concept of meritocracy.

Meritocracy is commonly defined as a social system where resources and/or status are distributed based on individual merit. This definition provides a statement about the objective criterion for determining whether a particular system qualifies as a meritocracy – the more each individual has earned his/her social and material place as a result of his/her own merit, the more the system can be described as meritocratic. However, this definition leaves open a key source of contestation – the meaning of merit. Although merit is often described as involving a combination of ability and effort, even this clarification leaves open to debate which particular abilities are relevant, what kinds of effort should count, and what the relative combination of these two ought to be. Thus, determining whether a system qualifies as a meritocracy can be highly subjective, and designating a given system as meritocratic usually involves building a shared consensus. Thus, rather than describing the objective state of affairs, labeling a given context a meritocracy is decidedly social psychological.

Meritocracy as Equity

Distributive justice involves people's evaluations of the fairness of the allocation of desirable outcomes across people. Deutsch (1975), in his influential summary of distributive justice, distinguished three basic principles that people endorse as fair depending on the particular social goals that are relevant to a particular relationship or social context (see Tyler and Smith, 1998). The principle of *equality* calls for the equal distribution of resources and is widely endorsed as the default principle in social settings were cooperation and harmony are primary goals. *Need* calls for unequal distribution of resources such that those who are most in need receive the larger portions. This principle is seen as appropriate where the goals are social welfare and responsibility. The principle of *equity* also calls for an unequal distribution of resources but holds that the ratio of any individual's inputs to outcomes should be equivalent to the same ratio of the person or persons with whom the individual is being compared. In other words, each individual should receive resources in proportion to their relative contributions. Some theorists have held that equity can be understood to be the general principle of distributive justice (see Adams, 1965; Walster et al., 1973). In fact, among some economists equity and justice are entirely conflated such that 'equitable' is synonymous with 'fair.' Although most social psychologists would see these as overstatements, equity is widely accepted as the preferred distributive justice principle in economic and competitive contexts (see Tyler, 2014). Thus, in capitalistic cultures such as North American (and perhaps other 'Western' nations), which focus heavily on competitive and economic relationships, the principle of equity (merit) is applied across a wide range of domains including educational, economic, political, organizational, and interpersonal.

While equality and need are inconsistent with meritocracy, the principle of equity serves as the primary justification for meritocracy. That is, a meritocracy is a system in which the principle of equity is upheld and it is generally believed that if individuals are talented enough or invest enough effort, they will receive a greater share of system resources and will advance up the social ladder. Given the relative ubiquity of equity as the preferred justice principle in individualistic/capitalistic cultures, it is perhaps not surprising that meritocracy has been described as America's 'dominant ideology' (Kluegel and Smith, 1986; McNamee and Miller, 2009). However, the recognition that equity is only one of at least three possible justice principles provides a basis for understanding why meritocracy, even when it is actually realized, can be rejected as unfair. For example, although we might find very few North American students who would reject meritocracy at the approach for distributing grades (a valued resource), we might find more variance in the endorsement of meritocracy as the primary strategy for distributing a professor's time. While the

best students might garner some additional attention, most would likely agree that a teacher's time should also be distributed more equally or perhaps on the basis of need. Thus, despite the strong support for meritocracy as a guiding cultural worldview, there remain contexts where meritocracy can be rejected in favor of another form for distributive justice.

Meritocracy as Social Mobility

The SIT (Tajfel and Turner, 1979) provides a framework for understanding when members of disadvantaged groups will accept their current low status, take individual action to improve their personal position, or take collective action to improve the status of their entire group. SIT proposes that individuals assess three *socio-structural variables* to determine which of these alternative responses they will take. They will assess the *legitimacy* of their group's low status – does the group deserve its disadvantaged position. They will consider the *stability* of the group's position – is it possible that the in-group can improve its relative standing and increase its share of the resources. However, a third and perhaps primary question is whether the boundaries between the groups are perceived to be *permeable* – is it possible for individual members to move out of their low-status group and acquire membership in a high status group. This third assessment is closely related to the concept of meritocracy, in that it involves assessing whether the social system is one in which individual effort leads to individual upward mobility.

SIT predicts, and considerable research has shown, that when the system appears to be open to individual mobility (when it is a meritocracy) disadvantaged group members will take actions designed to improve their personal position (see Ellemers, 2001; Wright, 2010 for reviews). They seek to demonstrate that they have the relevant attributes or can exert the necessary effort to move into a more advantaged group. Conversely, they are unlikely to engage in, or show support for, collective actions. Thus, SIT describes meritocracy as critical in maintaining the stability of the social order, because it encourages members of low status groups to seek personal solutions to their low status, and steers them away from disruptive efforts to improve the status of their entire group.

One reason seeing the system as meritocratic leads people to engage in individual rather than collective actions is that it focuses attention on individual identities and away from group identities (Wright, 2001). When rewards and social position appear to be result from individual effort and ability, group memberships become less important and thus group identities hold less motivational power (Derks et al., 2009). In addition, the apparent opportunity for individual mobility and the meritocratic nature of the social system also helps to support the view that the current social system, although unequal, is fair.

Meritocracy as Ideology

SIT's general claim that belief in individual mobility plays a critical role in preventing disadvantaged groups from challenging the social order is consistent with a growing literature that also focuses on meritocracy not as a structural reality but rather as a set of beliefs about how the social system works. Using terms like 'meritocracy beliefs' (e.g., Foster and Tsarfati, 2005), or 'meritocracy worldview' (e.g., Major et al., 2007) or 'meritocracy ideology' (e.g., Sengupta and Sibley, 2013), there is a growing social psychological literature that considers the motivational antecedents as well as the psychological and behavioral consequences of believing that the current system is a meritocracy.

Causes of Meritocracy Beliefs

Believing that one's place in the social hierarchy is determined by merit serves members of advantaged groups very well. This view provides justification for their privilege and allows for feelings of pride and satisfaction when comparing to those who have less. For the advantaged, meritocracy beliefs coincide nicely with the motivation to see oneself and one's in-group in a positive light (e.g., Levy et al., 2006). Thus, it is not surprising that Whites tend to endorse meritocratic beliefs more than Blacks and, for Whites, endorsement of meritocracy increases with personal socioeconomic status (Cokley et al., 2007).

However, endorsement of meritocracy beliefs is not influenced entirely by self- or group-interests. They are also the result of a desire to be accurate in assessing social reality. Thus, relevant and especially dramatic experiences can alter meritocracy beliefs. For example, Levy et al. (2006) showed that when considering the American government's woeful response to Hurricane Katrina, African-Americans reduce their endorsement of meritocratic beliefs – their belief that success results from hard work was undermined by what seemed like clear evidence to the contrary. The idea that relevant experiences can alter meritocracy beliefs is described nicely by Crosby and Ropp (2002) in their discussion of 'Awakening to Discrimination.' Women in academia who initially believed that their institutions functioned as meritocracies, through important trigger events or the slow accumulation of negative experiences, 'awoke' to the realization that sexist discrimination influenced their professional lives in ways they had not recognized. At this point, their faith in the meritocratic nature of their institutions was shaken. Conversely, dramatic changes in the social system can strengthen meritocracy beliefs. Kaiser et al. (2009) found that, immediately following Obama's election as president, participants from a number of ethnic groups concluded that racism was less of a problem and that anyone can achieve success through effort and perseverance.

A third explanation for why one might come to hold a meritocracy ideology centers around the premise that people have a general motivation to believe the world is just; that people generally get what they deserve and deserve what they get (Lerner, 1980; Hafer and Gosse, 2010). This general belief fosters a need to believe the broader institutions that influence our lives are legitimate and fair. This perspective has been advanced in John Jost and colleagues' *System Justification Theory* (e.g., Jost and Banaji, 1994; Jost and Hunyady, 2002), which holds that, in societies with obvious inequalities, there is a particular need for an ideology that can legitimize inequality. One such ideology is meritocracy.

Ledgerwood et al. (2011) provide experimental evidence of the effect of system justification motives and endorsement of meritocracy beliefs. In one of three experiments, participants read about studies that ostensibly tested the relationship between hard work and success. Given the dominance of

meritocracy ideology it is not surprising that participants judged those studies that supported the relationship between hard work and success to be of better quality and more convincing than those that refuted the effort/success relationship. However, the critical finding was that this 'pro-meritocracy' bias was made much stronger when participants were first exposed to a passage that threatened the legitimacy of their nation's social system. The authors conclude that their findings "suggest that the desire to preserve a view of the social system as fair and just serves to motivate the cognitive and behavioral defense of meritocratic beliefs."

Thus, it appears that endorsement of an ideology of meritocracy can result from: (1) a need to see oneself and one's group in positive light (personal or group self-interest); (2) a desire to accurately reflect what appears to be social reality (experience); and (3) the need to see the world as fair and thus to justify a social system with clear inequality in the distribution of resources and status.

Consequences of Meritocracy Beliefs

It appears that there may be a somewhat reciprocal relationship between self-interest and meritocracy. While the desire to see oneself in a positive light can serve to motivate endorsement of a meritocratic ideology, holding this worldview also has consequences for personal well-being. There is some evidence that meritocratic beliefs can have a positive effect on well-being. This may be particularly true for members of high status group (e.g., O'Brien and Major, 2005). As described earlier, believing that the current system rewards hard work and superior ability makes one's own (or one's group's) high status position a source of pride and thus contributes positively to self-esteem and life satisfaction. However, McCoy, et al. (2013) found that for members of both low and high status groups, endorsement of meritocracy beliefs was positively associated with self-esteem and physical health and that one of the primary reasons for this was that a stronger faith that one's society was meritocratic was associated with stronger feelings of control over future outcomes.

For disadvantaged group, however, believing that the system is meritocratic also implies that their current low-status position is deserved, and it would seem that this self-demeaning perspective should counter the potential benefits of a sense of control on self-esteem (Jost and Hunyady, 2002). Consistent with this, Quinn and Crocker (1999) investigating the relationship between endorsement of elements of a meritocratic ideology and self-esteem and depression among average weight and overweight women, found a positive relationship among normal weight women, no relationship among mildly overweight women, and among those who were most overweight the association was negative. In an effort to illuminate this somewhat murky relationship, O'Brien and Major (2005) offer a potential moderator. In a sample of African-Americans and Latinos endorsing meritocracy beliefs undermined well-being for those who were highly identified with their ethnic in-group, while these same beliefs were associated with higher well-being among African-Americans and Latinos who were low in ethnic identification.

Foster and colleagues (Foster et al., 2006; Foster and Tsarfati, 2005) provide another perspective and another moderator. They propose that this relationship between meritocracy beliefs and well-being is moderated by personal discrimination. Among women who experienced high personal discrimination, the belief in meritocracy was negatively correlated with well-being, while for those who experienced low personal discrimination the belief in meritocracy was positively correlated with well-being. This pattern held both when a measure of past experiences with discrimination was used (Foster et al., 2006) and when discrimination was manipulated in a laboratory setting (Foster and Tsarfati, 2005).

Major et al. (2007) reach a similar conclusion. They outline and test a *worldview verification model of self-esteem* which predicts that when experiences with discrimination threaten one's worldview, discrimination will impact negatively on self-esteem. If, however, experiences of discrimination confirm one's worldview, then self-esteem will be unaffected or even bolstered. Thus, for those holding a strong meritocracy worldview, discrimination will threaten that worldview and will undermine self-esteem. However, for those who reject the meritocratic worldview, experiences with discrimination will verify their worldview and thus will have little or even positive impacts on self-esteem. In three studies involving Latino-Americans and women, they showed this predicted interaction between endorsement or rejection of meritocracy worldview and perceptions of discrimination on self-esteem.

Meritocracy Beliefs and the Legitimization and Maintenance of Inequality

As discussed earlier, a meritocracy ideology not only describes the basis for upward mobility (merit), it also provides an explanation for the current status differences. If meritocracy exists, people 'at the top' have earned and thus deserve their advantage, while people 'at the bottom' also deserve their position. Thus, a meritocratic ideology supports positive stereotypes and attitudes toward advantaged group members, and negative attitudes toward disadvantaged groups.

Advantaged Groups

Just as it is not surprising that meritocracy beliefs are associated with more positive self-esteem among advantaged group members, it is also not surprising that it is advantaged group members who most strongly endorse these beliefs. What is disappointing is that endorsement of meritocracy beliefs is positively associated with denigration of low status groups. For example, Katz and Hass (1988) found a positive correlation between White Americans' meritocracy beliefs and anti-black attitudes. They also showed that priming meritocracy beliefs *prior* to filling out an anti-black attitudes scale resulted in even stronger anti-black attitudes (see also Biernat et al., 1996). Similar findings have been reported by Crandall (1994) in considering attitudes toward overweight people, and by Cozzarelli et al. (2001) in research on attitudes toward the poor. In every case, the more participants endorsed meritocracy beliefs the more they blame the low status group for their disadvantaged positions and the more negative their attitudes toward them.

Kaiser and her colleagues have also shown that negative attitudes about low status groups may not be directed equally at all members of those groups. For example, Whites who endorse a meritocracy worldview directed more negative attitudes especially toward members of minority groups who

demonstrated strong ethnic identification (Kaiser and Pratt-Hyatt, 2009). Similarly, Whites with strong meritocracy beliefs directed especially negative attitudes toward Blacks who attribute personal outcomes to racial discrimination (Kaiser et al., 2006). This finding points out another outcome of meritocracy beliefs. Since the existence of discrimination calls into question the claim that the system is meritocratic, advantaged group members who are strongly committed to meritocracy beliefs will be particularly unwilling to accept that discrimination and privilege play a role in determining one's place in the social hierarchy. Consistent with this, Knowles and Lowery (2012) showed that embracing meritocracy causes Whites to deny the existence of racial inequity not only as a means of justifying the social system but also in order to avoid the recognition of their own group's privilege. Thus, the denial of racial inequality that results from strong endorsement of meritocracy is tied directly to the previous discussion of the benefits of meritocracy beliefs for self-esteem and well-being. Advantaged group members who endorse meritocracy beliefs benefit in terms of self-esteem in part because this worldview allows them to deny the racial privilege they (and their in-group) enjoy.

Finally, Major et al. (2002) found that the more that members of an advantaged group (i.e., Whites and men) endorsed meritocracy beliefs the more likely they were to attribute negative outcomes they received from a low status group member (i.e., Latino-Americans or women) to discrimination. Similarly, O'Brien et al. (2012) found that endorsement or priming of meritocracy beliefs lead men to feel entitled to higher levels of pay. Paradoxically, the same beliefs that allow advantaged group members to deny the discrimination experienced by the disadvantaged out-group also allow them to feel more entitled to a large piece of the pie and to blame discrimination when they do not get that larger piece.

Disadvantaged Groups

As with the relationship between meritocracy beliefs and self-esteem, it is somewhat more difficult to understand why meritocracy beliefs ought to lead disadvantaged groups to legitimize the system in which they are disadvantaged. Nonetheless, Major and her colleagues have consistently demonstrated that strong endorsement or even temporary activation of elements of a meritocracy worldview can produce responses from members of low status groups that legitimize and perpetuate their disadvantage. For example, Major et al. (2002: Study 1) showed that among both Latino and African-Americans meritocracy beliefs were negatively related to the amount of discrimination they reported experiencing. In two experiments, they first measured Latino-Americans' (Study 2) and Women's (Study 3) meritocracy beliefs, and then participants were ostensibly rejected from a group activity by an out-group member (a White in Study 2 and a man in Study 3). Results showed that those who endorsed meritocracy beliefs more strongly were less likely to see this rejection as evidence of discrimination.

McCoy and Major (2007) replicated this general pattern of findings using a priming paradigm that allowed them to manipulate meritocracy beliefs. Relative to women in a neutral prime condition, women primed with meritocracy beliefs were less likely to see rejection by a man as discrimination, and also were more likely to stereotype themselves and women in ways that justified women's lower status. In a survey of working women (Hafer and Olson, 1993), stronger endorsement of meritocracy beliefs was associated with less dissatisfaction about both one's own low job status and the lower job status of women as a group. In addition, endorsement of meritocracy beliefs leads to less action to improve one's personal position, and importantly to less assertive action on behalf of women as a group. Thus, it was those who rejected meritocracy beliefs that were likely to take action to improve their personal position and to attempt to improve the status of the group as a whole. Finally, if one is unable to take direct action to improve one's personal position or the position of the group, one means of protecting one's well-being is to dis-identify with the domain (decrease the degree to which experiences in that domain are felt to be relevant for one's self-evaluation). However, Schmader et al. (2001) found that only those African-Americans who had rejected the dominant meritocratic ideology were able to engage strategies to dis-identify from a domain in which they were experiencing discrimination.

Thus, the ideology of meritocracy serves to maintain intergroup inequality by: (1) encouraging both advantaged and disadvantaged group members to ignore or even deny exiting privilege and discrimination; (2) providing justifications for opposition to equity-restoring policies; and (3) suppressing collective action among the disadvantaged.

Prescriptive or Descriptive Meritocracy Beliefs

An interesting conundrum emerges when we think carefully about what it means to endorse a meritocratic ideology. True meritocracy in which opportunities and rewards are distributed only on the basis of individual merit can only be realized in an unbiased system (Crosby, 2004; Smith-Winkelman and Crosby, 1994). The existence of discrimination directly undermines meritocratic principles. Thus, it would be logical to predict that those who strongly believe in meritocracy should be vigilant for evidence of group-based discrimination, should oppose these violations of meritocracy, and support efforts to change the system to remove them. Yet, as we have just reviewed, a great deal of research appears to show precisely the opposite relationship such that endorsing meritocracy results in the ignoring and legitimizing of discrimination and opposition to efforts to reduce it. Son Hing et al. (2002, 2011) provide a rather elegant solution to the conundrum by distinguishing between what they call *descriptive meritocracy* and *prescriptive meritocracy*.

Descriptive meritocracy is the belief that the current system, in fact, is a meritocracy; that personal effort and ability indeed does determine one's place in the system. This is the type of meritocracy ideology that has been measured in most of the research that we have just reviewed. In contrast, prescriptive meritocracy is the belief that merit is what ought to determine outcomes and status within the system. Using the terms described earlier in this chapter, prescriptive meritocracy involves a strong commitment to the justice principle of equity. Although prescriptive meritocracy beliefs can lead to 'principled' opposition to some forms of affirmative action to redress discrimination, when those programs are perceived to violate the merit principle, Son Hing et al. (2002) have shown that

prescriptive meritocracy beliefs can facilitate acknowledgment of inequality and increase support for affirmative action programs that are perceived to be removing these merit violations. In fact, Son Hing et al. (2011) conclude that their research shows that "beliefs that meritocracy ought to exist are independent of beliefs that meritocracy does exist. The latter – beliefs that meritocracy does exist – can serve as a hierarchy–legitimizing ideology that justifies current societal inequality. In contrast, the former beliefs – that meritocracy ought to exist – reflect a preference for a particular norm of distributive justice and can in fact drive support for policies that challenge the status quo." (pp. 447–448)

Tokenism

Meritocracy is often contrasted with a system where group boundaries are entirely impermeable and one's position is entirely determined by one's group membership (e.g., van Zomeren et al., 2008). However, in many contexts in contemporary societies the system is neither entirely meritocratic, nor is upward mobility impossible. Instead, barriers and restrictions result in only some of the qualified members of the disadvantaged group gaining access to advantaged positions. Often the restrictions can be quite severe and the result is a situation described as *tokenism* (Wright et al., 1990) – *an intergroup context in which the boundaries between the advantaged and disadvantaged groups are not entirely closed, but where there are severe restrictions on access to advantaged positions on the basis of group membership*.

The vast majority of work on tokenism has considered the experiences and actions of the 'tokens' themselves – those few individuals who, despite the barriers, gain access to advantaged positions – but there is also some work on the perspective of the rest of the disadvantaged group who are forced to remain disadvantaged. The theme that emerges from both of these literatures is, rather than being a benign step in the slow process of social change, tokenism creates new impediments that serve to legitimate and perpetuate the present social arrangement.

Successful Tokens

Much of the early research investigating the experiences of successful tokens (e.g., women and minorities in settings that were traditionally the exclusive domain of white males) centered on the psychological and performance consequences of interactions between tokens and majority group members. The dominant conclusions from this literature are that (1) tokens often do not receive treatment and status equal to their non-token colleagues (Kanter, 1977; Sackett et al., 1991; Yoder, 1994); (2) tokens often find their position difficult and dissatisfying and may seek to leave the group (e.g., Cohen and Swim, 1995; Stone, 2007); and (3) token status can result in deficits in performance (e.g., Saenz and Lord, 1989).

Often the problem for tokens is twofold. First, their small numbers make them highly visible and they often garner more scrutiny. While this can sometime lead to them being seen as more important to the group (e.g., Fuegen and Biernat, 2002), the highly visible nature of their group membership often leads them to be seen more in terms of group stereotypes. When combined with the need to conform to the norms and expectations that all group members feel, being seen in terms of the stereotypes of one's disadvantages group can make it difficult to fit in. In an effort to solve this problem, women in male-dominated workplaces, for example, may attempt to adopt a more masculine style in line with the normative culture, and while this may garner some recognition for competence, it may also be seen to violate prescribed gender norms leading to unpleasant backlash and social disapproval (Benard and Correll, 2010; Phelan and Rudman, 2010).

Biernat et al. (1998) in a study of women in the American military showed that at times token women will also provide self-evaluations that are more negative than the evaluations they receive from their peers. The authors propose that these are the product of self-stereotyping, and point out that these negative self-evaluations did not persist when the number of women in the group increased. Thus, because tokenism can heighten the salience of one's group member, it may at times also heighten the salience and use of negative self-stereotypes.

Finally, Karrasch (2003) investigated commitment in the US military and found that the often reported tokenism experiences in women (e.g., feeling highly visible, isolated, or trapped in fulfilling stereotypes) were correlated with lower *affective* commitment to the military. However, tokenism experiences were also positively related to *continuance* commitment, the perception that there are limited or no better alternatives. It was also shown that *affective* commitment predicted leadership behavior, while higher levels of *continuance* commitment were associated with fewer leadership behaviors. So, it appears that tokenism experiences reduce those aspects of commitment that are associated with success, while at the same time discouraging leaving for the fear that alternatives will be even worse.

In contrast to considering the direct experience of the tokens themselves, others have taken a somewhat more intergroup approach and have investigated the impact of token status on token's support for action designed to improve the general position of their group (e.g., Ellemers et al., 2004). If disadvantaged group members who become token members of an advantaged group were to use their newly acquired status and privileges to aid their disadvantaged group, they could be important contributors to social change. While some researchers have found that, in some cases, tokens do continue to assist other women to move up the social hierarchy, the majority of this research is less optimistic. Wright and Taylor (1999) conclude that tokens may often be co-opted by the advantaged group and are unlikely to support collective attempts by their group to change status quo. Although successful tokens recognize the illegitimacy of tokenism, they are unwilling to support actions likely to correct this injustice. This failure to support assertive action may lead others to misrepresent tokens' perceptions and feeling. The lack of support for collective social change sends an implicit message of support for the present social order, in effect legitimizing the continued discrimination against the disadvantaged in-group. Similarly, Derks et al. (in press) summarize their extensive research on token women by saying that "rather than fighting gender inequality as a broader concern – through collective

mobility – these women accept and legitimize the status quo in which women as a group are disadvantaged, and use this gender-based inequality to advance their individual career mobility."

In summary, tokens face formidably personal struggles themselves and thus it is perhaps not surprising that they are either unwilling or unable to 'carry the torch' of social change and champion the cause of the larger disadvantaged group.

The Disadvantaged Group

Despite the success of a few, for the vast majority of the disadvantaged group, tokenism results in the perpetuation of their disadvantaged status. Pettigrew and Martin (1987) suggested that awareness of the difficulties experienced by successful tokens can lead to lowered aspirations among other members of the group. In addition, a number of studies (e.g., Barreto, et al., 2004; Lalonde and Silverman, 1994; Reynolds et al., 2000; Wright, 2001) reveal that, while complete boundary impermeability – a closed social system – produces strong interest in collective action, when as few as 2% of the qualified members of the disadvantaged group are allowed access to advantaged positions collective action is replaced with a preference for individual action. It appears that the slightest hint of permeability undermines collective action.

Tokenism may have this effect because it makes salient personal identities and encourages interpersonal comparisons with the few successful tokens (Wright and Taylor, 1998). Further, tokenism may obfuscate assessments of the legitimacy and stability of the in-group's low-status position (Wright, 1997). Danaher and Branscombe (2010) have shown that token hiring practices while recognized as unfair and demeaning to the in-group nonetheless maintain organizational identification among disadvantaged group members. They found that tokenism keeps women participating in the system at large, by maintaining individual mobility beliefs, and within their local organization, by maintaining organizational identification. Thus, "token practices fool women into thinking that token conditions are positive for their group by making group boundaries appear to be permeable and legitimate."

Ellemers and Barreto (2009) point out that tokenism is but one example of the ways that social structures and discriminatory practices can elicit from those who have disadvantaged beliefs and behaviors that help to perpetuate the *illusion of meritocracy* – the belief that, despite all the evidence to the contrary, individual merit and not group membership determines one's opportunities and one's place in the social world.

See also: Belief in a Just World; Justice: Social Psychological Perspectives; Sexism; Social Identity in Social Psychology; Social Psychology.

Bibliography

Adams, J.S., 1965. Inequity in social exchange. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 2. Academic Press, New York, pp. 267–299.

Barreto, M., Ellemers, N., Palacios, M., 2004. The backlash of token mobility: the impact of past group experiences on individual ambition and effort. Personality and Social Psychology Bulletin 30 (11), 1433–1445.

Benard, S., Correll, S.J., 2010. Normative discrimination and the motherhood penalty. Gender & Society 24, 616–646.

Biernat, M., Crandall, C.S., Young, L.V., Kobrynowicz, D., Halpin, S.M., 1998. All that you can be: stereotyping of self and others in a military context. Journal of Personality and Social Psychology 75 (2), 301–317. http://dx.doi.org/10.1037/0022-3514.75.2.301.

Biernat, M., Vescio, T.K., Theno, S.A., Crandall, C.S., 1996. Values and prejudice: toward understanding the impact of American values on outgroup attitudes. In: Seligman, C., Olson, J.M., Zanna, M.P. (Eds.), The Psychology of Values: The Ontario Symposium, vol. 8. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, pp. 153–189.

Cohen, L.L., Swim, J.K., 1995. The differential impact of gender ratios on women and men: tokenism, self-confidence, and expectations. Personality and Social Psychology Bulletin 21 (9), 876–884. http://dx.doi.org/10.1177/0146167295219001.

Cokley, K., Komarraju, M., Pickett, R., Shen, F., Patel, N., Belur, V., Rosales, R., 2007. Ethnic differences in endorsement of the protestant work ethic: the role of identity and perceptions of social class. Journal of Social Psychology 147, 75–89.

Cozzarelli, C., Wilkinson, A.V., Tagler, M.J., 2001. Attitudes toward the poor and attributions for poverty. Journal of Social Issues 57 (2), 207–227.

Crandall, C.S., 1994. Prejudice against fat people: ideology and self-interest. Journal of Personality and Social Psychology 66, 882–894.

Crosby, F.J., 2004. Affirmative Action Is Dead: Long Live Affirmative Action. Yale University Press, New Haven, CT, USA.

Crosby, F.J., Ropp, S.A., 2002. Awakening to discrimination. In: Ross, M., Miller, D.T. (Eds.), The Justice Motive in Everyday Life. Cambridge University Press, New York, NY, USA, pp. 382–396.

Danaher, K., Branscombe, N.R., 2010. Maintaining the system with tokenism: bolstering individual mobility beliefs and identification with a discriminatory organization. British Journal of Social Psychology 49 (2), 343–362. http://dx.doi.org/10.1348/014466609X457530.

Derks, B., van Laar, C., Ellemers, N., 2009. Working for the self or working for the group: how self- versus group affirmation affects collective behavior in low-status groups. Journal of Personality and Social Psychology 96, 183–202.

Derks, B., van Laar, C., Ellemers, N., Raghoe, G., in press. Coping with ethnic discrimination at work: self-group distancing is not only for Queen Bees. Journal of Social Issues.

Deutsch, M., 1975. Equity, equality, and need: what determines which value will be used as the basis of distributive justice? Journal of Social Issues 31 (3), 137–149.

Ellemers, N., 2001. Individual upward mobility and the perceived legitimacy of intergroup relations. In: Jost, J.T., Major, B. (Eds.), The Psychology of Legitimacy: Emerging Perspectives on Ideology, Justice, and Intergroup Relations. Cambridge University Press, New York, NY, USA, pp. 205–222.

Ellemers, N., Barreto, M., 2009. Maintaining the illusion of meritocracy: how men and women interactively sustain gender inequality at work. In: Demoulin, S., Leyens, J., Dovidio, J.F. (Eds.), Intergroup Misunderstandings: Impact of Divergent Social Realities. Psychology Press, New York, NY, USA, pp. 191–212.

Ellemers, N., van den Heuvel, H., de Gilder, D., Maass, A., Bonvini, A., 2004. The underrepresentation of women in science: differential commitment or the queen bee syndrome? British Journal of Social Psychology 43 (3), 315–338. http://dx.doi.org/10.1348/0144666042037999.

Foster, M.D., Tsarfati, E., 2005. The effects of meritocracy beliefs on women's well-being after first-time gender discrimination. Personality and Social Psychology Bulletin 31, 1730–1738.

Foster, M.D., Sloto, L., Ruby, R., 2006. Responding to discrimination as a function of meritocracy beliefs and personal experiences: testing the model of shattered assumptions. Group Processes & Intergroup Relations 9, 401–411.

Fuegen, K., Biernat, M., 2002. Reexamining the effects of solo status for women and men. Personality and Social Psychology Bulletin 28 (7), 913–925. http://dx.doi.org/10.1177/01467202028007005.

Hafer, C.L., Gosse, L., 2010. Preserving the belief in a just world: when and for whom are different strategies preferred? In: Bobocel, D., Kay, A.C., Zanna, M.P., Olson, J.M. (Eds.), The Psychology of Justice and Legitimacy. Psychology Press, New York, NY, USA, pp. 79–102.

Hafer, C.L., Olson, J.M., 1993. Beliefs in a just world, discontent, and assertive actions by working women. Personality and Social Psychology Bulletin 19, 30–38.

Jost, J.T., Banaji, M.R., 1994. The role of stereotyping in system-justification and the production of false consciousness. British Journal of Social Psychology 33, 1–27.

Jost, J.T., Hunyady, O., 2002. The psychology of system justification and the palliative function of ideology. European Review of Social Psychology 13, 111–153.

Jost, J.T., Pelham, B.W., Sheldon, O., Sullivan, B.N., 2003. Social inequality and the reduction of ideological dissonance on behalf of the system: evidence of enhanced system justification among the disadvantaged. European Journal of Social Psychology 33, 13–36.

Kaiser, C.R., Drury, B.J., Spalding, K.E., Cheryan, S., O'Brien, L.T., 2009. The ironic consequences of Obama's election: decreased support for social justice. Journal of Experimental Social Psychology 45, 556–559.

Kaiser, C.R., Dyrenforth, P.S., Hagiwara, N., 2006. Why are attributions to discrimination interpersonally costly? A test of system- and group-justifying motivations. Personality and Social Psychology Bulletin 32, 1523–1536.

Kaiser, C.R., Pratt-Hyatt, J.S., 2009. Distributing prejudice unequally: do whites direct their prejudice toward strongly identified minorities? Journal of Personality and Social Psychology 96, 432–445.

Kanter, R.M., 1977. Men and Women of the Corporation. Basic Books, New York.

Karrasch, A.I., 2003. Antecedents and consequences of organizational commitment. Military Psychology 15, 225–236.

Katz, I., Hass, R., 1988. Racial ambivalence and American value conflict: correlational and priming studies of dual cognitive structures. Journal of Personality and Social Psychology 55, 893–905.

Kluegel, J.R., Smith, E.R., 1986. Beliefs about Inequality: American's View of What Is and What Ought to Be. Aldine de Gruyer, Hawthorne, NJ.

Knowles, E.D., Lowery, B.S., 2012. Meritocracy, self-concerns, and whites' denial of racial inequity. Self and Identity 11 (2), 202–222.

Lalonde, R.N., Silverman, R.A., 1994. Behavioral preferences in response to social injustice: the effects of group permeability and social identity salience. Journal of Personality and Social Psychology 66, 78–85.

Lerner, M., 1980. The Belief in a Just World. Plenum Press, New York.

Ledgerwood, A., Mandisodza, A.N., Jost, J.T., Pohl, M.J., 2011. Working for the system: motivated defense of meritocratic beliefs. Social Cognition 29, 322–340.

Levy, S.R., West, T., Ramirez, L., Karafantis, D.M., 2006. The protestant work ethic: a lay theory with dual intergroup implications. Group Processes and Intergroup Relations 9, 95–115.

Major, B., Gramzow, R.H., McCoy, S.K., Levin, S., Schmader, T., Sidanius, J., 2002. Perceiving personal discrimination: the role of group status and legitimizing ideology. Journal of Personality and Social Psychology 82, 269–282.

Major, B., Kaiser, C.R., McCoy, S.K., O'Brien, L., 2007. Perceived discrimination as worldview threat or worldview confirmation: implications for self-esteem. Journal of Personality and Social Psychology 92, 1068–1086.

McCoy, S.K., Major, B., 2007. Priming meritocracy and the psychological justification of inequality. Journal of Experimental Social Psychology 43, 341–351.

McCoy, S.K., Wellman, J.D., Cosley, B., Saslow, L., Epel, E., 2013. Is the belief in meritocracy palliative for members of low status groups? Evidence for a benefit for self-esteem and physical health via perceived control. European Journal of Social Psychology 43, 307–318.

McNamee, S.J., Miller, R.K., 2009. The Meritocracy Myth, second ed. Rowman & Littlefield, Lanham, MD, USA.

O'Brien, L.T., Major, B., 2005. System-justifying beliefs and psychological well-being: the roles of group status and identity. Personality and Social Psychology Bulletin 31, 1718–1729.

O'Brien, L.T., Major, B.N., Gilbert, P.N., 2012. Gender differences in entitlement: the role of system-justifying beliefs. Basic and Applied Social Psychology 34, 136–145.

Pettigrew, T.F., Martin, J., 1987. Shaping the organizational context for black American inclusion. Journal of Social Issues 43, 41–78.

Phelan, J.E., Rudman, L.A., 2010. Prejudice toward female leaders: backlash effects and women's impression management dilemma. Social and Personality Psychology Compass 4, 807–820.

Quinn, D.M., Crocker, J., 1999. When ideology hurts: effects of belief in the protestant ethic and feeling overweight on the psychological well-being of women. Journal of Personality and Social Psychology 77, 402–414.

Reynolds, K., Oakes, P., Haslam, A., Nolan, M., Dolnik, L., 2000. Responses to powerlessness: stereotyping as an instrument of social conflict. Group Dynamics: Theory, Research and Practice 4, 275–290.

Sackett, P.R., DuBois, C.L.Z., Noe, A.W., 1991. Tokenism in performance evaluation: the effects of work group representation on male-female and white-black differences in performance ratings. Journal of Applied Psychology 76, 263–267.

Schmader, T., Major, B., Gramzow, R.H., 2001. Coping with ethnic stereotypes in the academic domain: perceived injustice and psychological disengagement. Journal of Social Issues 57, 93–111.

Saenz, D.S., Lord, C.G., 1989. Reversing roles: a cognitive strategy for undoing memory deficits associated with token status. Journal of Personality and Social Psychology 56, 698–708.

Sengupta, N.K., Sibley, C.G., 2013. Perpetuating one's own disadvantage: intergroup contact enables the ideological legitimation of inequality. Personality and Social Psychology Bulletin 39, 1391–1403.

Smith-Winkelman, C., Crosby, F.J., 1994. Affirmative action: setting the record straight. Social Justice Research 7, 309–328.

Son Hing, L.S., Bobocel, D., Zanna, M.P., 2002. Meritocracy and opposition to affirmative action: making concessions in the face of discrimination. Journal of Personality and Social Psychology 83, 493–509.

Son Hing, L.S., Bobocel, D., Zanna, M.P., Garcia, D.M., Gee, S.S., Orazietti, K., 2011. The merit of meritocracy. Journal of Personality and Social Psychology 101, 433–450.

Stone, P., 2007. Opting Out?: Why Women Really Quit Careers and Head Home. University of California Press, Berkeley, CA.

Tajfel, H., Turner, J.C., 1979. An integrative theory of intergroup conflict. In: Austin, W.G., Worchel, S. (Eds.), The Social Psychology of Intergroup Relations. Brooks/Cole, Monterey, CA, pp. 33–48.

Tyler, T.R., 2014. Social justice. In: Mikulincer, M., Shaver, P.R., Dovidio, J.F., Simpson, J.A. (Eds.), APA Handbook of Personality and Social Psychology, Group Processes, vol. 2. American Psychological Association, Washington, DC, USA, pp. 95–122.

Tyler, T.R., Smith, H.J., 1998. Social justice and social movements. In: Gilbert, D., Fiske, S., Lindzey, G. (Eds.), Handbook of Social Psychology, fourth ed., vol. 2. McGraw-Hill, New York, NY, pp. 595–629.

van Zomeren, M., Postmes, T., Spears, R., 2008. Toward an integrative social identity model of collective action: a quantitative research synthesis of three socio-psychological perspectives. Psychological Bulletin 134, 504–535.

Walster, E., Berscheid, E., Walster, G., 1973. New directions in equity research. Journal of Personality and Social Psychology 25 (2), 151–176.

Wright, S.C., 1997. Ambiguity, social influence and collective action: generating collective protest in response to tokenism. Personality and Social Psychology Bulletin 23, 1277–1290.

Wright, S.C., 2001. Restricted intergroup boundaries: tokenism, ambiguity and the tolerance of injustice. In: Jost, J., Major, B. (Eds.), The Psychology of Legitimacy: Emerging Perspectives on Ideology, Justice, and Intergroup Relations. Cambridge University Press, pp. 223–254.

Wright, S.C., 2010. Collective action and social change. In: Dovidio, J.F., Hewstone, M., Glick, P., Esses, V.M. (Eds.), Handbook of Prejudice, Stereotyping, and Discrimination. Sage Publications, Thousand Oaks, CA, pp. 577–596.

Wright, S.C., Taylor, D.M., 1998. Responding to tokenism: individual action in the face of collective injustice. European Journal of Social Psychology 28, 647–667.

Wright, S.C., Taylor, D.M., 1999. Success under tokenism: co-option of the newcomer and the prevention of collective protest. British Journal of Social Psychology 38, 369–396.

Wright, S.C., Taylor, D.M., Moghaddam, F.M., 1990. Responding to membership in a disadvantaged group: from acceptance to collective action. Journal of Personality and Social Psychology 58, 994–1003.

Yoder, J.D., 1994. Looking beyond numbers: the effects of gender status, job prestige, and occupational gendertyping on tokenism processes. Social Psychology Quarterly 57, 150–159.

Merton, Robert K (1910–2003)

Christian Fleck, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Over the period of three quarters of a century, Robert K. Merton demonstrated an impressive productivity as a sociologist of different specialties. First and foremost, he contributed to sociological theory and is the founder of the sociology of science. He also contributed to the sociology of knowledge, organization theory, deviant behavior, mass communications; and the professions and less broadly recognized fields of sociological semantics, and the sociology and history of ideas. The broad range of subject matters he studied and his close reading of former contributions to sociology resulted in Merton's identifying and labeling such social phenomena as the self-fulfilling prophecy, unanticipated consequences, focus groups, and role models, among a great many others. Given this wide spectrum of conceptual and empirical contributions and Merton's plea for middle-range ambitions in theory building, he will continue to be a sociologist of consequence.

If social scientists and authors from other branches of scholarship would always name the originators of neologisms and concepts they use, one name would probably surface much more often than any other: Robert K. Merton (following his own style of writing, which exhibited a strong inclination toward acronyms, I will refer to him as RKM hereafter). He coined, borrowed, polished, refined, and reformulated concepts more than any other social scientist. Along that way, he also introduced a new term for this type of work: initially 'conceptual analysis' (Merton, 1968: 168–171) and later on 'reconceptualization' (Merton, 2004: 245–260). Without RKM, not only would the language of the social sciences be more meager, but also the vernacular, the vocabulary of ordinary life. 'Self-fulfilling prophecy,' 'role model,' 'dysfunctional,' 'focussed interview' (see Merton, 1987a for an explanation of the double-s), 'serendipity,' and other Mertonian coinages traveled smoothly into the lexicon of middle-ranged educated people, not only in the Anglophone world but also beyond. Other Mertonian reconceptualizations remained inside the walls of academia: Thomas theorem, Matthew effect, latent versus manifest, opportunity structures, etc.

The *Oxford English Dictionary*, a source RKM regularly consulted both to examine changes of meaning over time and to encounter unknown but telling expressions (Merton, 2004: 233–244), cites his publications four-dozen times to document trajectories of semantics. Going through these quotations, one can see that RKM both codified meaning, so his wording functioned as an exemplar, and invented, sometimes unsuccessfully, new expressions. To give only a few examples: aberrant, avocational, backfence, benchmark, boondoggling neurosis, caste hypogamy, the local and the cosmopolitan, Matthew effect, modal personality, opinionnaires, out-grouper, pilot study, point-to-point correlation, prestige-structure, problematics (the principal problems), quantitativist, retreatism, ritualism, role, role model, subsurface trends in class structure.

On a par with augmenting scholars' and lay people's vocabularies come RKM's publications. Since the rules of the academic game require participants to refer sources, we can say with some certainty that several of his papers and some of his books became what Eugene Garfield called 'citation classics' (one of the many admirers of RKM and his oeuvre). Garfield, the founder of the citation index business, has been encouraged and tutored by RKM and listed some of his work in the category of highest cited papers (Garfield, 1963).

To better understand RKM and his intellectual distinctiveness beyond citation counts, it is advisable to take a look at the format of RKM's writings and publications. The majority of the books authored by RKM fall in the genre of collections and only a small minority fit the prerequisites of a monograph. In contrast to what happens elsewhere in academia and to other authors, RKM's collected volumes are much wider known than most of his monographs (Merton, 1949, 1957a, 1968, 1973, 1976, 1982, 1996). Paraphrasing one of his well-known phrases, one could identify his writings as middle-range texts. They do not try to cover huge intellectual landscapes but selected regions, which are investigated not from an eagle's perspective but by someone much closer to those investigated. RKM only rarely engaged in the textbook standards of participant observation as his preferred mode of data collection, but his writings show that he always tried to get as close to the field he analyzed as possible. One could call what forbade RKM to write about things strange to him were his need for familiarity coming up against inaccessibility in a researcher's time frame (see his explanations on strategic research event in (Merton, 1987b: 11–15)). To consult another of RKM's favorite lexicographic sources, most of his texts are essays in the strict sense of the definition of this genre by one of the founding fathers of modern lexicography, Samuel Johnson (1709–84): 'First taste of any thing,' 'a loose sally of the mind, an irregular, undigested piece, not a regular and orderly performance.' This circumscription might sound strange in the ears of present-day readers, but it is in accord with one of RKM's lasting convictions regarding the status of sociology as a science. In his early years, he stuck to the conviction that sociology is a young member of the science family and has a long way to go before arriving at the mature level of its siblings (Merton, 1945a). In later years, RKM bid farewell to the idea of sociology's quick maturing and even conceded that sociology might never become a science *more geometrico*. Long before he changed his mind with regard to the maturity of his craft, he had chosen the less scientific format of the essay as the most appropriate for his intellectual ambitions and inclinations. Enterprises less academically mature than neighboring rivals, e.g., sociology

compared to physics, ask for writing styles and publication formats that avoid fixations and stringencies and favor explorations. Here as elsewhere in the social world, the alignment of the causality's arrow is not carved in stone and correlation prevents hasty interpretations. But who wants to rule out that someone with a preference for the literary form gravitates toward subjects that best fit his format of choice?

Merton's Paradigmatic Vision

Given the probing style of his writing, it is anything but easy to summarize RKM's intellectual contributions or to identify the core of his intellectual ambitions. At least since the first volume of collected papers, *Social Theory and Social Structure* (abbreviated by RKM usually as STSS), appeared in print in 1949, a regular criticism he had encountered was that he did not provide an encompassing theory. While this may be true, ultimately this truth depends on the meaning one assigns to the highly equivocal term 'theory,' as RKM himself outlined in the introduction of the first two editions of STSS, which became two separate chapters in the third edition.

RKM did not provide any general theory such as Darwin's theory of natural selection, or Freud's theory of the role the unconscious, or Weber's theory of occidental rationalism – theories whose core can usually be summarized in one sentence. Instead, he outlined the panorama a theory should try to cover. RKM does not provide simple sets of propositions and call them a theory, but labors on several aspects and elements of theoretical relevance. Therefore, his theories are seldom encompassing but hint toward what one should keep in mind when examining particular aspects of social life. Given the collectively shared mood not only of sociologists but also of neighboring social scientists in the middle of the twentieth century, it is surprising to see that RKM was not in favor of grand theories that are supposed to explain everything out of a handful of concepts and vaguely defined relations between them. Early on, RKM distanced himself from Parsons' endeavor to develop such a general theory (because he did this politely and did not orchestrate it as a parricide like many other unruly disciples, most later generations of sociologists did not notice the depth of differences between these two, Merton, 1980). Parsons' claim, that his general theory will be different from the speculative theories of Spencer and his contemporaries, did not persuade RKM. Looking back to the years following the end of World War II, one needs to highlight that it was not very fashionable to challenge the highfalutin ambitions proposed by Parsons, then widely admired.

Another interpretation of theory could point to the generalization one strives for or hopes to achieve. The more generally applicable a proposition, the more likely we could call it a theory. What others labeled a theory RKM called a 'paradigm,' very early on. He made use of this term first in 1945 when he outlined a paradigm for the sociology of knowledge (Merton, 1945b), in 1949 when he elaborated one for functional analysis, and later one for structural analysis (Merton, 1975). He also used the same strategy in other contexts: e.g., in his reconstruction of the meaning of anomie (Merton, 1938c, 1995), in his attempt to elaborate patterns of intermarriage (Merton, 1941), and elsewhere. RKM must have been surprised, to say the least, when he had to recognize that someone else became designated as the originator of the term paradigm. However, his historical notes on Thomas Kuhn's trajectory do not show any jealousy (Merton, 1979, 2004: 264–269). According to RKM, paradigms are orienting, which means that they are an outline for what a researcher should look for; they direct research by providing basic queries, list variables, and outline possible connections between elements under scrutiny.

Contemplating about the properties of what could be reasonably called a sociological theory, RKM arrived at a level below the generalized ones, which Parsons and nearly all earlier theorists preferred; he gave it the versatile title "theories of the middle range: theories that lie between the minor but necessary working hypotheses that evolve in abundance during day-to-day research and the all-inclusive systematic efforts to develop a unified theory that will explain all the observed uniformities of social behavior, social organization, and social change" (Merton, 1968: 39).

This notion and understanding of theory work in sociology has been challenged, repudiated, and most of all misunderstood. Its meaning and range is debatable, but RKM devoted more energy and effort to contribute theories of this type rather than defending his concept as such. It is nearly impossible to provide a list of all Mertonian theories of middle range, but some of them will be covered below. An intrinsic problem in identifying these kinds of theories lies in the difficulty distinguishing theories from concepts in RKM's writings. Take as an example one of his most famous expressions, the self-fulfilling prophecy. Is it a middle-range theory, a mechanism, as some later admirers of RKM proposed, or is it anything more than a cleverly labeled concept? The scope of applicability of this 'dynamic social mechanism,' as he himself classified this concept-theory (Merton, 1968: 182), is nearly limitless: one should not be surprised to find proofs of it in ancient time, in far away corners of today's world, in highest developed as in disadvantaged, underdeveloped stages of mankind – in a word: it fits the criteria of universalistic applicability, much more than the definition of middle range would ask for. RKM himself subsumes the self-fulfilling prophecy, its counterpart the suicidal or self-destroying prophecy together with the "unanticipated consequences of action or decision or belief" (Merton, 1968: 182) as one pattern of latent functions. In comparing this concept with another famous reconceptualization of RKM, the interpretation and reformulation of Emile Durkheim's anomie, one wonders whether the range of these 'types of individual adaptation' – conformism, innovation, ritualism, retreatism, and rebellion – is broader than the former mentioned ones or not. Just to mention one aspect of a comparative analysis, one could imagine a human group, which cannot realize its culturally defined goals directly but in which the mechanism of self-fulfillment operates, nonetheless. Or to put it in another way: both self-fulfilling and anomie mechanisms seem to be valid in and for nearly every type of social group but neither help us to make distinctions between these groups, societies, or other larger social structures. The transition from concept to theory has

not been prescribed by RKM. We are, therefore, more or less free to choose arbitrarily between them.

Beneath high-abstraction and low-verifiable validity, to which RKM contributed in his earlier career more heavily than later on, he made a large number of contributions to sociological knowledge, only some of which can be covered in such a condensed tribute to a scholar who authored his first published paper at the age of 24 and who handed over his last manuscript to the publisher at the age of 93. In the seven decades between 1934 and 2004 when *The Travels and Adventures of Serendipity* came out in English posthumously (a translated version had been published in Italy in 2002), RKM published 20 books, more than 150 articles, numerous introductions, prefaces and, book reviews and has been responsible for coediting two huge compilations of reprints: 61 dissertations in sociology and 100 'classics, staples, and precursors in sociology.'

A Selection of RKM's Substantive Contributions

If we distinguish between core convictions and lasting engagement in a scholar's oeuvre, I would like to select five such durable concerns in RKM's case.

His very first book, a revised edition of his 1936 Harvard PhD thesis, came out in 1938, not the best time in contemporary history to draw the attention of academia to the work of an unknown youngster (Merton, 1938a). Contrary to the unpromising environment, this juvenilia impressed experts then and now (Cohen, 1990). *Science, Technology and Society in Seventeenth Century England* had been produced under the tutelage of the eminent historian of science, George Sarton. RKM had conducted research on the formation of modern science in England in the early days of the Royal Society. The history of science context did not keep the trained sociologist from applying contemporary inventions of social research methodology, e.g., he contemplated about how to draw a sample from the *Dictionary of National Biography* and ended up with about 6000 biographical data sets of individual scholars; he conducted a content analysis of the early volumes of the *Proceedings of the Royal Society*; and he paid tribute to the embeddedness of the work of these scientists in larger public affairs of building vessels, improving navigation, and designing weapons. Together with some related articles and strengthened by the fact of a reprint of core chapters of the dissertation in consecutive editions of STSS, RKM became the founding father of sociology of science. During the seven decades of his academic life, RKM remained concerned with this special branch of sociological research (Cole and Zuckerman, 1975). In 1937, while still a graduate student, Sarton made him an associate editor for 'Social Aspects of Science' in the editorial team of the journal *Isis*. RKM demonstrated his continued interest by publishing and commissioning trend reports and some of the first bibliographies (Barber and Merton, 1952; Barber, 1952, 1956). Later in life, he devoted his ASA presidential address to his specialty (Merton, 1957b), directed some dissertations and well-funded major research projects, helped to create bodies for international exchange of ideas and research findings within this group of specialists (RKM was the first president of ISA's research committee sociology of science), and was able to establish good relations to scientists themselves. Evidence of that latter point is that he was the first and is still the only social scientists who received the *National Medal of Science* from the hands of the US president.

A second field of investigations, which occupied RKM for some years, was the result of his close collaboration with Paul F. Lazarsfeld at the Bureau of Applied Social Research (Lazarsfeld, 1975; Merton, 1998). It was then labeled 'mass communication research' and would fit today into 'sociology of culture' and 'cultural sociology' properly. Back in the 1940s, recent societal transformations attracted the attention of social scientists: the role of radio, meaning and workings of propaganda, and political attempts to make use of both technologies. The most visible result of RKM's involvement in this area is the small monograph *Mass Persuasion*, which he wrote with the assistance of Marjorie Fiske and Alberta Curtis (Merton et al., 1946, 2004). The underlying study was the birthplace of the focused interview, which became subsequently adapted and modified into focus group studies, prompting RKM to publish a whole manual of procedures (Merton, et al., [1956]1990). In *Mass Persuasion*, the team of investigators interviewed people after they listened to the third all-day war bond drive by a then well-known singer and radio talk show host, Kate Smith. Combining content analysis and focused interviews, the research team, which had been formed on short notice (Lazarsfeld called this firehouse research: Lazarsfeld, 1969: 313), presented both an analysis of the persuasiveness of a mass culture heroine and an insightful exploration of what made this program attractive for listeners. In a paper RKM published around this time together with Lazarsfeld (Lazarsfeld and Merton, 1948), the two authors elaborated what later on became known as the Columbia tradition in mass communication research, a middle-of-the-road position that accredited the critical wing of social analysis with some validity: they called it 'narcotizing dysfunction of media' but insisted that media fulfilled social functions too: status conferral, on the one hand, and enforcement of social norms, on the other. The fact that RKM's involvement in empirical social research brought two monographs mainly authored by him to life is in itself telling regarding that his main interests were elsewhere. Two chapters in STSS's part on theoretical sociology, which were also published independently as a book (Merton, 1967), devote to this interrelationship in some detail: "The Bearing of Sociological Theory on Empirical Research" and "The Bearing of Empirical Research on Sociological Theory" (Merton, 1968: 139–171).

Judging by the academic echo it evoked, one of RKM's earliest papers surpass all others: "Social structure and anomie," first published in 1938 in *American Sociological Review*, not only became a citation classic but also had lasting influence on the sociological study of deviant behavior. For decades, no student of criminology, deviant behavior, and neighboring fields could evade the five types of individual adaption, or strain theory as an important way to understand deviance. RKM's interpretation of Durkheim's concept of anomie remains influential and controversial at the same time: 40 years after the publication one of his students gave an overview on the reception (Cole, 1975) and

three quarters of a century after its first print RKM's paper got quoted 3797 times in Google Scholar (August 2014). RKM himself contributed to the debate on anomie theory, further elaborated the meaning of opportunity structures (see his retrospective memoir: Merton, 1995), and used it as an analytical model for composing a reader together with Robert Nisbet on *Contemporary Social Problems* (Merton and Nisbet, 1961). Between 1961 and 1976, when this reader appeared in its fourth edition, it was formative for thinking about forms of deviance, at least in American colleges and universities.

At any rate, two more lines of thinking can be identified as characteristic of RKM: His devotion to the history of the sciences and social sciences and his elaboration of what he called sociological semantics. The interest in the past of his own discipline started for RKM from his conviction that re-conceptualization and taking a second glance at past contributions to sociology are fruitful endeavors. Later in his career, he took his own life and work as an exemplar to demonstrate the relevance of archived material, like unpublished correspondences, to grasp particular developments in sociology in more detail. He restrained from writing a full-fledged autobiography but penned several shorter memoirs, most famously the Charles Homer Haskins lecture he gave in his hometown Philadelphia in April of 1994 (Merton, 1994).

One could argue that autobiographical details of a scholar's life do not influence the validity of his pronouncements and argues even more strongly that hobby horses do not belong to an academic balance sheet. But a portrait of RKM would be lacking in color and integrity if OTSOG would not be mentioned. The acronym stands for *On the Shoulders of Giants*, originally a somewhat lengthy letter to his friend, the Harvard historian Bernard Bailyn, typed during some leisure time at the end of 1957. Eight years later, RKM's lengthy letter became a book and has since been translated and gone through at least three editions. At the time when the original manuscript has been produced, and even when the first printed version appeared in 1965, only few readers knew that this 'Shandean Postscript' was nothing like a 'singleton' (cf Merton, 1961a). Around the time of the letter writing, which resulted in OTSOG, RKM was working with his colleague Elinor Barber on a manuscript that only appeared some 40 years later. The book (Merton and Barber, 2004) contains a 70-page afterword where RKM elaborates on the evolvement of his interest in the serendipity pattern and his plea for what he then called *sociological semantics*. By that he meant the analysis of changes in the meaning of widely used concepts or terms as well as the investigation of sociological characteristics of its creators, providers, users, and those ignoring such inventions.

An Improbable Life Trajectory

Given this short overview of this scholar's oeuvre, the reader might want to know how all this could have been accomplished by one man. To put it in a slightly different way: Who was the man behind all these publications? The author was not born in 1910 but only some 15 years later, as the interested reader first had a chance to read in a portrait of RKM, which was published in the widely distributed magazine *New Yorker* by the science journalist Morton Hunt (Hunt, 1961). About 34 years later, RKM confirmed the story in his Haskins lecture of 1994: The newborn son of Yiddish-speaking immigrants from Russia goes under a very different name in 1910 and changed it to a more American sounding one when he aspired to start a professional career as a magician. Before this, encouraged by his mother and his sister's husband, he adopted both reading and magic as habits. Due to friendly female librarians in the neighboring library, established by a generous gift from Andrew Carnegie, one of the Gilded Age's robber barons, young not-yet-Robert earned a good education in "literature, science, and history, especially in biographies and autobiographies" (Merton, 1994: 343). In Philadelphia, RKM attended high school and continued into college at the then noncredited Temple University. There he met his first sociology teacher, George E. Simpson, who recruited the sophomore, now known as RKM, as a research assistant to help him finish his own dissertation on *The Negro in the Philadelphia Press*. From Simpson, RKM learned sociological techniques of data collection and analysis. His mentor also took the undergraduate student to the Annual Meeting of the then American Sociological Society. Back then, the number of attendants was small enough that an ambitious young no-name could chat with an established older man. Pitirim A. Sorokin (who, just as RKM's parents, was born in Tsarist Russia but was neither Jewish nor Russian, but belonged to the Finno-Ugric minority of Komi) encouraged RKM to apply for a scholarship for graduate study at Harvard University, where Sorokin recently had been hired to found the department of sociology. From 1931 onwards, RKM spent 7 years around Harvard Square where he met some of his future mentors and friends. Together with Sorokin, he authored at least three papers and learned Italian over a summer break after Sorokin's request (visiting professor Corrado Gini's English was insufficient for teaching). Sorokin and RKM drifted apart partly due to the latter's choice of his dissertation topic (Merton, 1989). As already mentioned, Sarton took over as mentor and role model: from him, RKM learned a new style of writing and judging others' contributions (Thackray and Merton, 1972). RKM's admiration for the Belgian-born polymath never receded (Merton, 1985). Then Talcott Parsons entered the circles of the graduate student, which brought RKM back on route to a career in sociology (Merton, 1980). The fact that he remained an underpaid tutor in Harvard's department could not have to do with the pervasive anti-Semitism held by Harvard's WASPs because RKM's Jewish origins were not known publicly until his disclosure in 1994. Given RKM's productivity before he reached the age of 30 (one book, more than a dozen articles, including some which became classics (Merton, 1936, 1938b; Merton and Sorokin, 1937), and a bunch of book reviews), it is not a big surprise that other universities were eager to hire the young Harvard man. From 1938 to 1941, RKM took a detour through the Deep South and acted as chairman and professor at Tulane University in New Orleans. Around this time, RKM must have realized that he may once become famous and started to administer his correspondence and files rigorously and easy to navigate for later users of his papers, stored at the

very university where he spent most of his career: Columbia University in the City of New York. There he started as an assistant professor on a post, which has been split in two because the two powerful members of the department could not agree on one candidate. The tacit assumption of the department's rivals and other Columbia people must have been that within a short period of time only one incumbent will survive. Quite the contrary happened – very much to the advantage of sociology. RKM and the nearly 10-years older Paul F. Lazarsfeld did not battle each other but collaborated. The story of how this relationship started has been initially told by the *New Yorker* and then by Lazarsfeld in the festschrift RKM received at his 65th birthday (Lazarsfeld, 1975): Instead of having the intended first dinner at the Lazarsfeld's, the two men went downtown to a radio studio to conduct an urgent radio-audience study, leaving their astonished wives behind. When the job was done, both continued talking at the Russian Tea Room until long after midnight.

Lazarsfeld was able to persuade RKM to join him at the newly renamed Bureau of Applied Social Research where RKM remained, in one or another function, until its end. The relationship between the Viennese-Jewish upper-middle-class mathematician-turned-into-social researcher and the Philadelphian slum dweller and Harvard alumnus became one of the highly unusual success stories in academia. They nurtured several cohorts of first-class students who became fellow sociologists – just to mention some from the earliest days: Peter M. Blau, James Coleman, Lewis Coser, Rose Laub Coser, Alvin Gouldner, Seymour Martin Lipset – and dominated American sociology for a while. The highly improbable collaboration stretched out beyond the Columbia campus. RKM also acted as gatekeeper at several publishing houses and foundations: at Ford, Guggenheim, the Palo Alto Center for Advanced Study in the Behavioral Sciences, MacArthur, Russell Sage. At other spending bodies, RKM acted either as board member or reviewer. It is not inconceivable that such a powerful position generated some jealousy.

There is another trait in RKM's performance, which needs to be mentioned: His refusal to take positions in public. RKM never performed as a public intellectual, never wrote op-eds, appeared at public events or sent letters to the editor (for one exception see Merton, 1961b). He started his days as early as 04.30 in the morning and continued working in his home office until he reached a degree of exhaustion or feeling of completeness. He then walked to his more publicly available office at Columbia or his small cubicle at the Russell Sage Foundation's headquarter where he enjoyed the status of Foundation Scholar. Apparently, he did not need much sleep because one could also meet him later in the day or at night. RKM seemed to enjoy living in New York City and meeting people from several corners of this quintessential professional metropolis. But he always remained a private person and never joined movements, organizations, or any other forms of public engagements. He rejected calls to other places, even if they were Oxford. His political convictions shine through in several of his papers, most explicitly in response to the Nazi movement in Germany and his concern with racial discrimination in his own country (Merton, 1942, 1948).

Legacies

A decade passed since RKM died on 23 February 2003, but it seems he and his works are more appreciated today than ever. During his lifetime, he received two festschriften (Coser, 1975; Gieryn, 1980), two volumes with critical comments on his work appeared (Mongardini and Tabboni, 1998; Clark et al., 1990) and two books introduced and critically commented on his work (Sztompka, 1986; Crothers, 1987). Criticism has been raised, but never reached the level of the diatribes against his friend-and-teacher Parsons and his collaborator-and-friend Lazarsfeld (cf vicariously for others Mills, 1959; Gouldner, 1970).

The biggest influence of RKM is epitomized by one of his concepts: obliteration by incorporation. But as mentioned above, some of his contributions are still related to him as their originator. Over the last couple of years, at least three more tribute-like collections appeared after conferences in Budapest, New York, and Warsaw (Elkana et al., 2011; Calhoun, 2010; Mica et al., 2011), complemented by short introductions to his work in German and French (Mackert and Steinbicker, 2013; Saint-Martin, 2013).

It might be neither a self-fulfilling nor suicidal prophecy to say that both RKM's published work and the immensely rich collection of manuscripts, correspondences, and other material stored and cataloged at the Rare Book and Manuscript Division of Butler Library, Columbia University, as the Robert K. Merton papers will find further interpreters assessing matters from the viewpoint his shoulders offer.

See also: Anomie: History of the Concept; Anomie; Celebrity; Communication Research and Media Studies, History of; Crime, Sociology of; Definition of the situation: History of the Concept; Delinquency, Sociology of; Deprivation: Relative; Ethical Questions in Social and Behavioral Sciences, History of; Focus Groups; Functionalism, History of; Gatekeepers in Social Science; Ignorance, History of Concept; Knowledge, Sociology of; Kuhn, Thomas S. (1922–96); Lazarsfeld, Paul Felix (1901–76); Media Events; Parsons, Talcott (1902–79); Science and Technology Studies, History of; Science, History of; Science, Sociology of; Scientific Misconduct, Plagiarism, and Institutional Control of Misconduct; Social Structure: History of the Concept; Sociological Theory; Sociology, History of; Sorokin, Pitirim Alexandrovich (1889–1968); Strain Theories and Crime; Unintended Consequences: History of the Concept.

Bibliography

Barber, B., 1952. Science and the Social Order. Free Press, Glencoe, IL.

Barber, B., 1956. Sociology of science. A trend report and bibliography. Current Sociology 5 (2), 89–153.

Barber, B., Merton, R.K., 1952. Brief bibliography for the sociology of science. Proceedings of the American Academy of Arts and Sciences 80 (2), 140–154.

Calhoun, C. (Ed.), 2010. Robert K. Merton: Sociology of Science and Sociology as Science. Columbia University Press, New York.

Clark, J., Modgil, C., Modgil, S. (Eds.), 1990. Robert K. Merton: Consensus and Controversy. Falmer Press, London.

Cohen, I.B. (Ed.), 1990. Puritanism and the Rise of Modern Science: The Merton Thesis. Rutgers University Press, New Brunswick.

Cole, J.R., Zuckerman, H., 1975. The emergence of a scientific specialty. The self-exemplifying case of the sociology of science. In: Coser, L.A. (Ed.), The Idea of Social Structure. Harcourt Brace Jovanovich, New York, pp. 139–174.
Cole, S., 1975. The growth of scientific knowledge. Theories of deviance as a case study. In: Coser, L.A. (Ed.), The Idea of Social Structure. Harcourt Brace Jovanovich, New York, pp. 175–220.
Coser, L.A. (Ed.), 1975. The Idea of Social Structure: Papers in Honor of Robert K. Merton. Harcourt Brace Jovanovich, New York.
Crothers, C., 1987. Robert K. Merton. Tavistock Publications, London.
Elkana, Y., Szigeti, A., Lissauer, G. (Eds.), 2011. Concepts and the Social Order: Robert K. Merton and the Future of Sociology. Centreal European University Press, Budapest.
Garfield, E., 1963. Robert K. Merton – Author and Editor Extraordinaire. Part 1 [1983] and Part 2. In: Essays of an Information Scientist, pp. 312–318, pp. 319–329. Philadelphia.
Gieryn, T.F. (Ed.), 1980. Science and Social Structure: A Festschrift for Robert K. Merton. Transaction of the New York Academy of Sciences, New York.
Gouldner, A.W., 1970. The Coming Crisis of Western Sociology. Basic Books, New York.
Hunt, M., 28 January 1961. Profiles: how does it come to be so? The New Yorker 39–63.
Lazarsfeld, P.F., 1969. An episode in the history of social research. In: Fleming, D., Bailyn, B. (Eds.), The Intellectual Migration. Europe and America, 1930–1960. Belknap Press of Harvard University Press, Cambridge, pp. 270–337.
Lazarsfeld, P.F., 1975. Working with Merton. In: Coser, L.A. (Ed.), pp. 35–66.
Lazarsfeld, P.F., Merton, R.K., 1948. Mass communication, popular taste, and organized social action. In: Bryson, L. (Ed.), Communication of Ideas. Harper, New York, pp. 95–118.
Mackert, J., Steinbicker, J., 2013. Zur Aktualität von Robert K. Merton. vs Verlag, Wiesbaden.
Merton, R.K., 1936. The unanticipated consequences of purposive social action. American Sociological Review 1, 894–904 [Reprinted in 1976].
Merton, R.K., 1938a. Science, Technology and Society in Seventeenth Century England. Saint Catherine Press, Bruges, Belgium.
Merton, R.K., 1938b. Science and the social order. Philosophy of Science 5, 321–337.
Merton, R.K., 1938c. Social structure and anomie. American Sociological Review 1938, 672–682 [Reprinted in 1968, pp. 185–214].
Merton, R.K., 1941. Intermarriage and the social structure. Psychiatry: Journal of the Biology and Pathology of Interpersonal Relations 361–374 [Reprinted in 1976, pp. 217–250].
Merton, R.K., 1942. A note on science and democracy. Journal of Legal and Political Sociology 1 (1–2), 115–126.
Merton, R.K., 1945a. Sociological theory. American Journal of Sociology 50 (6), 462–473.
Merton, R.K., 1945b. The sociology of knowledge. In: Gurvitch, G., Moore, W.E. (Eds.), Twentieth Century Sociology. Philosophical Library, New York, pp. 366–405 [Reprinted in 1968, pp. 510–542].
Merton, R.K., 1948. Discrimination and the American creed. In: MacIver, R.M. (Ed.), Discriminiation and National Welfare. Harper & Brothers, New York, pp. 99–126 [Reprinted in 1976, pp. 189–216].
Merton, R.K., 1949. Social Theory and Social Structure: Toward the Codification of Theory and Research. Free Press, Glencoe.
Merton, R.K., 1957a. Social Theory and Social Structure, second ed. Free Press, Glencoe.
Merton, R.K., 1957b. Priorities in scientific discovery. American Sociological Review 22 (6), 635–659 [Reprinted in: 1973, pp. 286–324].
Merton, R.K., October 1961a. Singletons and multiples in scientific discovery. Proceedings of the American Philosophical Society 105 (5), 470–486 [Reprinted in: 1973, pp. 343–370].
Merton, R.K., 16 July 1961b. The canons of the anti-sociologist. In: New York Times [Reprinted in 1976, pp. 180–185].
Merton, R.K., 1967. On Theoretical Sociology: Five Essays, Old and New. Free Press, New York.
Merton, R.K., 1968. Social Theory and Social Structure, third ed. Free Press, New York.
Merton, R.K., 1973. In: Storer, N.W. (Ed.), The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago Press, Chicago.
Merton, R.K., 1975. Structural analysis in sociology. In: Blau, P.M. (Ed.), Approaches to the Study of Social Structure. Free Press, New York, pp. 21–52 [Reprinted in 1976, pp. 109–144.].
Merton, R.K., 1976. Sociological Ambivalence and Other Essays. Free Press, New York.
Merton, R.K., 1979. The Sociology of Science: An Episodic Memoir. Southern Illinois University Press, Carbondale.
Merton, R.K., 1980. Remembering the young Talcott Parsons. The American Sociologist 15, 68–77.
Merton, R.K., 1982. In: Rosenblatt, A., Gieryn, T.F. (Eds.), Social Research and the Practicing Professions. Abt Books, Cambridge.
Merton, R.K., 1985. George Sarton: episodic recollections by an unruly apprentice. Isis 76 (4), 470–486.
Merton, R.K., 1987a. The focussed interview and focus groups: continuities and discontinuities. Public Opinion Quarterly 51, 550–566.
Merton, R.K., 1987b. Three fragments from a sociologist's notebooks. Establishing the phenomenon, specified ignorance, and strategic research materials. Annual Review of Sociology 13, 1–28.
Merton, R.K., 1989. The Sorokin-Merton correspondence on Puritanism, Pietism and science, 1933–1934. Science in Context 3 (1), 291–298.
Merton, R.K., 1994. A Life of Learning. Charles Homer Haskins Lecture. ACLS Occasional, New York paper no. 25. [Reprinted in 1996, pp. 339–359].
Merton, R.K., 1995. Opportunity structure. The emergence, diffusion, and differentiation of a sociological concept, 1930s–1950s. In: Adler, F., Laufer, W.S. (Eds.), The Legacy of Anomie Theory. Transaction Publishers, New Brunswick, pp. 3–78.
Merton, R.K., 1996. In: Sztompka, P. (Ed.), On Social Structure and Science. University of Chicago Press, Chicago.
Merton, R.K., 1998. Working with Lazarsfeld: notes and contexts. In: Lautman, J., Lécuyer, B.-P. (Eds.), Paul Lazarsfeld (1901–1976). La sociologie de Vienne à New York. L'Harmattan, Paris, pp. 163–211.
Merton, R.K., 2004. In: Merton, R.K., Barber, E.G. (Eds.), Autobiographical Reflections on the Travels and Adventures of Serendipity, pp. 230–301.
Merton, R.K., Barber, E.G., 2004. The Travels and Adventures of Serendipity: A Study in Historical Semantics and the Sociology of Science. Princeton University Press, Princeton.
Merton, R.K., Fiske, M., Kendall, P.L., 1990. The Focused Interview: A Manual of Problems and Procedures. Free Press, New York [Originally 1956].
Merton, R.K., Fiske, M., Curtis, A., 1946. Mass Persuasion: The Social Psychology of a War Bond Drive. Harper & Brothers, New York.
Merton, R.K., Fiske, M., Curtis, A., 2004. Mass Persuasion: The Social Psychology of a War Bond Drive. Introduction by P. Simonson. Howard Fertig, New York.
Merton, R.K., Nisbet, R.A., 1961. Contemporary Social Problems. Harcourt, Brace, Jovanovich, New York.
Merton, R.K., Sorokin, P.A., 1937. Social time. A methodological and functional analysis. American Journal of Sociology 42, 615–629.
Mica, A., Peisert, A., Winczorek, J. (Eds.), 2011. Sociology and the Unintended: Robert Merton Revisited. P. Lang, Frankfurt.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, New York.
Mongardini, C., Tabboni, S. (Eds.), 1998. Robert K. Merton & Contemporary Sociology. Transaction, New Brunswick.
Saint-Martin, A., 2013. La sociologie de Robert K. Merton. La Découverte, Paris.
Sztompka, P., 1986. Robert K. Merton: An Intellectual Profile. St. Martin's Press, New York.
Thackray, A., Merton, R.K., 1972. On discipline building. The paradoxes of George Sarton. Isis 63 (4), 472–495.

Mesoamerica, Archaeology of

Lane F Fargher, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional – Unidad Mérida, Yucatán, México

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mesoamerica is a geographical and cultural area that includes most of Mexico and parts of Central America where an ancient civilization thrived between 3000 and 500 years ago. The states and societies that formed ancient Mesoamerica shared a number of characteristics, including: intensive agriculture, markets, urbanism, writing, religion, calendar, social organization, and monumental architecture. However despite a high degree of cultural, ideological, and economic integration, Mesoamerica remained politically fragmented throughout its history. In this article, I attribute the persistent fragmentation to the cross-scale interaction of interregional trade and regional-scale political economic strategies and use world-system theory and collective action theory to elucidate their interaction.

Introduction

Prior to the Spanish Conquest, the cultural area known as Mesoamerica (Kirchhoff, 1952) had developed a thriving civilizational tradition over a period of approximately 3000 years. Thus, the term Mesoamerica makes reference to both a culture area and a civilizational tradition (Blanton et al., 1999; *see* States and Civilizations, Archaeology of). Complex social, political, and economic processes marked this tradition. Specialized merchants traveled the length and breadth of Mesoamerica, trading in special international markets (entrepôts), as well as the major regional market places. Most importantly, the long-distance exchange fostered by these merchants moved ideas, goods, and people among a vast array of regions and thereby welded Mesoamerica into a single thriving civilization. Yet, Mesoamerica remained politically fragmented throughout the entire pre-hispanic era and at no point was it integrated under a single government. Long-term political fragmentation of this area resulted from contingent cross-scale interaction of world-systems processes at the macroregional scale and collective action processes at the regional scale. In this article, I summarize the long-term patterns in these two processes, their cross-scale interaction, and their impact on individual regions from the Formative through the Postclassic. For those readers not familiar with Mesoamerica, I begin by laying out the basic environmental, geographical, cultural, and chronological features of this impressive civilization. I then provide a brief theoretical review of world-systems theory and collective action theory as they apply to premodern states and civilizations. Finally, I trace the cross-scale interaction of these processes for the most important chronological blocks in the Mesoamerican cultural sequence (Early Formative, Classic, Late Postclassic).

Background

Geographically, Mesoamerica extends from the southern boundaries of Mexico's northern desert region to Central America. This area covers approximately 950 000 km^2 and is divided into seven major geographical subareas (**Figure 1**; **Table 1**). Historically, human occupation of these subareas is divided among archaeological 'periods' based primarily on changing patterns in subsidence strategies and political complexity (see **Figure 2**; **Table 2** for information on key processes during each period). The character of Mesoamerican civilization reached its fullest expression during the Late Postclassic, a period known through a combination of archaeology and history. The key features of Mesoamerica during this period were intensive agriculture, craft specialization, markets, sedentism, urbanism, states, writing, religion, calendar, mathematics, social organization, and monumental architecture (see Blanton et al., 1999: Chapter 1).

Importantly, the degree of intensive agriculture and craft specialization were made possible by the evolution of complex marketing systems (Blanton, 1996; Blanton et al., 2005; Masson and Freidel, 2012). Ancient Mesoamericans bought and sold goods and services in a vast array of different types of markets. The major market systems operated on a 5-day market cycle and secondary markets met every 20 days. Market transactions were conducted with both commodity money and through in-kind payments. The primary commodity monies were cacao beans, cotton quilts, and feathers covered in gold dust (Smith, 2003, pp. 115–117). Together intensive production and markets supplied Mesoamerica's cities with foodstuffs, goods, and services.

Mesoamerican cities ranged in size from a few thousands to hundreds of thousands of inhabitants and could be either sparsely occupied (5–20 people/ha) to densely packed (100–200 people/ha). In many cases, these cities were several times larger than other settlements in their hinterlands, indicating that urban centers monopolized economic, political, ideology, and social power. Monumental architecture (pyramids, palaces, plazas) was generally located at the center of cities. In some cases, the state engaged in public planning and constructed (paved) roads (organized on a grid), canals, and bridges, as well as public architecture at the centers of state-administered wards (Fargher and Blanton, 2007). In other cases, political officials (usually kings) planned only monumental architecture at the center and the rest of the city exhibited chaotic, mazelike urban landscapes (e.g., Tikal (Carr and Hazard, 1961)).

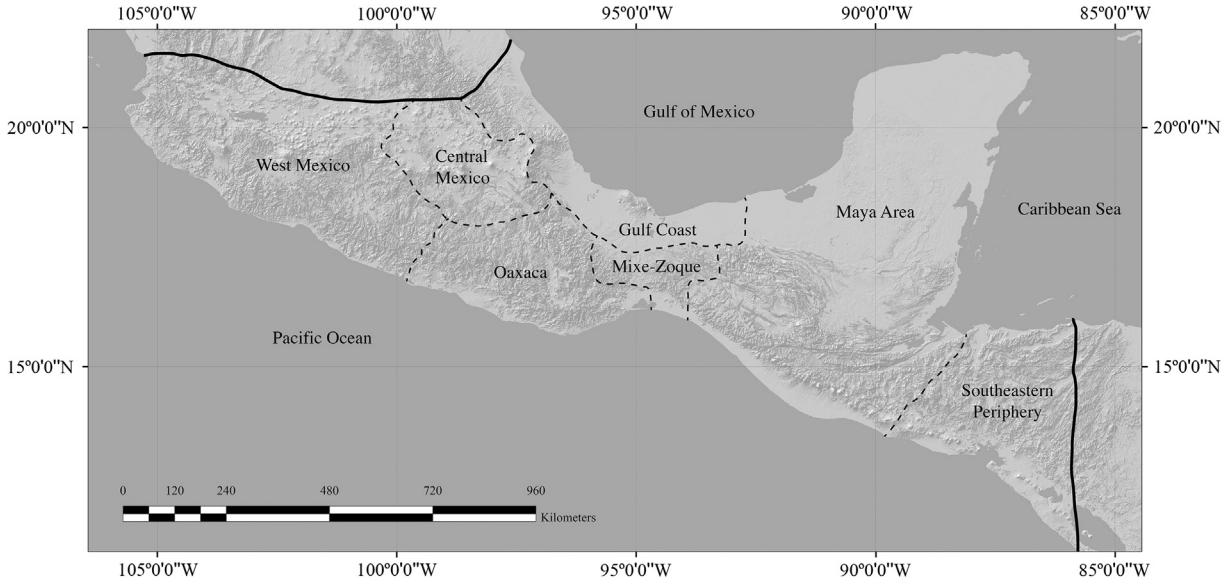

Figure 1 Map of Mesoamerica showing subareas (boundaries marked by solid line, subareas by dashed line).

Table 1 Geography and environment of Mesoamerica

Geographical subarea	*Area (km^2)*	*Geography*	*Climate*	*Rainfall (mm)*
West Mexico	200 000	Mountainous with valleys and basins, coasts	Semiarid – subhumid	250–950 (>2000 along coasts)
Central Mexico	68 000	Mountainous with valleys and basins	Semiarid – subhumid	250–950
Gulf Coast	90 000	Littoral backed by mountains	Humid	>2000
Oaxaca	90 000	Mountainous with valleys and basins, coasts	Semiarid – subhumid	250–950 (>2000 along coasts)
Mixe-Zoque	25 000	Coastal plain and piedmont mountain zones	Humid	>2000
Maya area	340 000	Low land plains, coasts, plateaus, and mountains	Semiarid – Humid	250–4000
Southeastern periphery	129 000	Mountains and coasts	Humid	>2000

Finally, cities and their hinterlands were governed by a vast array of states built around diverse political strategies. Some states were small, covering 50 sq.km, whereas others were massive covering tens of thousands of square kilometers. Kings ruled some states (e.g., Maya, Mixtec) and councils consisting of both nobles and commoners governed others (e.g., Tlaxcallan, Teotihuacan) (Blanton et al., 1996; Fargher and Blanton, 2007; Fargher et al., 2011; Sharer, 1994). From ethnohistory, we know some indigenous names for states (e.g., *altepetl* in the Basin of Mexico; *tlahtocayotl* in Puebla; *ñuu* in the Mixteca Alta; *quiche* in the Valley of Oaxaca) (Fargher et al., 2011; Oudijk, 2002). In some cases, these states conquered neighboring states to create empires, e.g., the Aztec Triple Alliance (Chase et al., 2009).

Explaining the Long-Term Evolution of Ancient Mesoamerican Civilization

Collective Action Theory

The long-term history of Mesoamerican civilization allows for the comparative examination of how different revenue (state financial) strategies shaped political organization and the consumption of resources. Revenue strategies may be grouped as either 'internal' (taxes on agricultural production, ordinary market transactions, corvée, etc., paid by free citizens) or 'external' (state ownership of mines, slaves, agricultural estates, animal herds, or control of international trade). States financed through internal revenues depend on taxpayers for their revenues and, thus, must convince them to pay their taxes. They do this by offering taxpayers voice (participation in the political process) and public goods (e.g., public planning, public security, water and sanitation services, road and bridge construction and maintenance, social security, etc.), as well as bureaucratizing public administration (to limit corruption, punish tax evaders, and distribute public goods) and limiting the degree to which governors (kings, presidents, etc.) can use public funds for their personal gratification, which results in highly collective states (see Blanton and Fargher, 2008). Conversely, states funded through external financing do not need to persuade citizens to pay taxes; and, consequently, these states tend not to offer public goods and voice nor bureaucratize, and are, thus, low in collectivity. The governors of these states build social networks to achieve their political goals by

	Periods	Sociopolitical processes	Key sites	
- 1400	Late Postclassic	Empires	Tenochtitlan-Tlatelolco, Tzintzuntzan, Cholula, Tlaxcallan, Calixtlahuaca, Coixtlahuaca, Mayapan, Tututepec, Tulum, Cerro Jazmín, Teposcolula, Texcoco	- 600
- 1200				- 800
- 1000	Early Postclassic	Collapse and reorganization	Tula, El Tajín, Chichen Itza	-1000
- 800	Epi-Classic		Xochitecatl-Cacaxtla, Xochicalco, Cholula, Teotihaucan, Cantona, Chichen Itza, Uxmal	- 1200
- 600	Classic	Middle horizon	Teotihuacan, Monte Albán, Teuchitlan, Matacapan, Tikal, Calakmul, Caracol, Copan, Palenque, Kalminaljuyu, Cerro Jazmín, Huamelulpan	- 1400
- 400				- 1600
- 200	Terminal Formative	First cities and states	Teotihuacan, Cuicuilco, Xochitecatl-Cacaxtla, Tres Zapotes, Cholula, Monte Albán, Huamelulpan, Teuchitlan, La Venta, El Mirador, Chiapa de Corzo, Monte Negro	- 1800
A.D./B.C.				years ago
- 200	Late Formative			- 2200
- 400	Middle Formative	Increasing social inequality and complexity	Monte Albán, Cuicuilco, Tlapacoya, La Venta, Nakbé, Komchen, Tayata, Dzinicahua, Teposcolula	- 2400
- 600		Settlement of Maya lowland plains	San José Mogote, La Venta, Tlatilco, Chalcatzingo, Cuicuilco, Tlapacoya, Tayata	- 2600
- 800				- 2800
- 1000	Early Formative	Early horizon	San José Mogote, San Lorenzo, Tlatilco, Tlapacoya, Coapexco, Tayata, Chalcatzingo, Paso de la Amada, Etlatongo, Yucuita, Tilantongo	- 3000
- 1200		Transition to village-based agriculture		- 3200
- 1400				- 3400
- 1600	Archaic	Mobile hunter-gatherers sedentary fishers		- 3600
- 1800				- 3800

Figure 2 Chronology of Mesoamerica.

distributing exotic and luxurious gifts, hosting elaborate and exclusive feasts, and gifting the right to extract revenues from certain settlements or regions to powerful clients.

Integrating World Systems and Collective Action

Cross-cultural research indicates that principals configured their response to wealth generated by international trade depending on their overall revenue strategy. Most commonly, governing officials sought to monopolize or control long-distance exchange and use this revenue to build networks of power (see Blanton and Fargher, 2008). However in a few cases, rulers rejected control over international trade as a revenue source in order to curb or limit its impact on politics and power. In order to do this, governing officials attempted either to squash long-distance exchange or direct goods acquired through it to collective consumption in public rites and feasts and/or into markets open to all citizens. Concurrently, they developed political structures that effectively controlled malfeasance and punished corrupt political officials,

Table 2 Chronological, geographical, demographic, and political economic background of focal cases

Focal case	*Period*	*Region*	*Physical geography*	*Area (ha)*	*Estimated population*	*Description*	*Revenue sources*	*Monopolies on economic and ideological resources*	*Public goods*	*Degree of collectivity*	*Sources*
San Lorenzo	Early Formative	Gulf Coast, Southern Veracruz	Tropical lowland alluvial plain with swamps and oxbows	53 (Hinterlands 690)	2500	Large village	Monopolization of long-distance exchange, monumental exotic basalt objects, ilmenite 'cubes,' workshops attached to elite residences, possible monopoly on agricultural lands	Massive palaces, manipulation of monumental basalt heads, alters, and statues	Not identifiable	Low	Clark 1997; Grove 1997
San José Mogote	Early Formative	Valley of Oaxaca	Highland valley	60–70	1000	Large village	Corvée or agricultural taxes	Wide distribution of prestige goods, construction of a public and accessible building	Public buildings and inclusive feasting	High	Blanton et al., 1999; Kowalewski et al., 1989
Paso de la Amada	Early Formative	Pacific Coast, Chiapas	Lowland littoral with swamps	50–70	1000	Large village	Corvée	Control of labor for the construction of large houses, wide distribution of prestige goods except hollow ceramic figurines	Use of large dwellings by workers	Moderate	Lesure, 1997, 2011
Teotihuacan	Classic	Basin of Mexico	Highland plateau	2350	125 000	City	Tax on free craft specialists and markets	Wide distribution of wealth items, consumption of elaborate goods in inclusive state rituals	Roads, canals, drains, public buildings, public security, inclusive feasting, market administration	High	Cowgill 1997; Millon 1973

Continued

Table 2 Chronological, geographical, demographic, and political economic background of focal cases—cont'd

Focal case	*Period*	*Region*	*Physical geography*	*Area (ha)*	*Estimated population*	*Description*	*Revenue sources*	*Monopolies on economic and ideological resources*	*Public goods*	*Degree of collectivity*	*Sources*
Monte Albán	Early Classic	Valley of Oaxaca	Highland valley	450	22 500	City	Corvée and agricultural taxes	Limited use of prestige goods, small palaces and tombs	Market administration, public space, and public security	Moderate	Blanton 1978; Kowalewski et al., 1989; Marcus and Flannery 1996
Tikal	Early Classic	Central Peten (northern Guatemala), Central Maya Lowlands	Tropical lowland plain	1600	60 000	City	Monopolization of long-distance exchange, prestige goods, attached labor, and agricultural land	Massive palaces, elaborate tombs, carved stone images of named rulers	Some public security	Low	Sharer 1994
Aztec Triple Alliance (Tenochtitlan)	Late Postclassic	Basin of Mexico	Highland plateau	1200	200 000	City	Agricultural taxes and corvée, market taxes, imperial tribute; royal estates	Wide distribution of prestige goods, large royal palaces, carved stone focuses on rituals	Public security, roads and bridges, canals, parks, public space, and redistribution	High	Fargher and Blanton 2007
Mixteca Alta (Teposcolula)	Postclassic	Northwest Oaxaca	Highland valleys and mountains	290	7500	City	Monopolization of long-distance exchange, prestige goods, attached labor, and agricultural land	Massive palaces, consumption of jewelry and elaborate clothing	Some public security	Low	Fargher et al., 2011
Tarascan Empire (Tzintzuntzan)	Late Postclassic	Michoacan, Jalisco	Highland plateau	674	30 000	City	Monopolization of long-distance exchange, prestige goods, attached labor, and agricultural land	Massive palaces, monopolization of prestige goods	Public security	Low	Pollard 1982

including rulers. Cross-culturally, the promotion of rule by council or multiple principals (e.g., republics and democracies) seems to be a favored pathway for controlling the agency of governing officials, but others exist (see Blanton and Fargher, 2008).

Thus, following from these ideas, I (1) explore the ways that governing officials confronted the social problems created by international trade during the Early Formative, Classic, and Late Postclassic Mesoamerica; and (2) describe the ways that these responses affected the organization of the state. In general, Mesoamerican states that relied heavily on long-distance exchange to fund political objectives were fragmented, contentious, dangerous, and oppressive, with few public goods or opportunities for citizens to express their voice (i.e., low in collectivity); alternatively, states that focused on internal revenues and rejected monopolies on international trade by rulers were inclusive and ordered with extensive public goods, opportunities for voice, and bureaucratization of public administration (i.e., high in collectivity).

Case I: Early Formative (3150–2900 Years Ago)

During the latter part of the Early Formative Period, the first major phase of world-systems integration took shape in Mesoamerica. This period is marked by interregional exchange of prestige goods, like baby-face figurines, jade, pottery, and iron-ore, the development of an international style ('Olmec'), and shared ideological and religious orientations across much of Mesoamerica, but especially in Central Mexico, Oaxaca, and on the Gulf Coast (Figure 3). The increased interregional interaction produced notable impacts at the regional- and site-scale. However, the impact of this heightened exchanged varied from region to region.

San Lorenzo

Extant evidence strongly indicates that political power at San Lorenzo was built around external revenue sources. Evidence of strategies to build cooperation between rulers and constituents and encourage compliance with revenue demands is absent. High-status individuals sought to control or monopolize key prestige goods and ideological sources of power. Coercion may also have been employed to acquire wealth, foodstuffs, and dependent laborers (Clark, 1997). Revenues derived from these monopolies were poured into monumental art, elaborate residences, and personal adornment. One of these residences was a 12-m-long apsidal structure constructed atop a 2-m-high platform that covered 3750 m^2 (Cyphers, 1996). A second excavated residence may have been even more impressive with a 3.5-m-tall central basalt column, carved stone step covers, stone benches, plastered and painted walls, and a stone aqueduct that brought water to the house from a lagoon on the plateau (by comparison low-status individuals resided in simple wattle-and-daub dwellings) (Cyphers, 1996). Moreover, carved stone depictions of high-status individuals show them in elaborate dress decorated with jewelry, scepters, and headdresses (Grove, 1997). Monumental art focused on demonstrating control over access to the supernatural and on the personalized images of important individuals (Clark, 1997; Grove, 1997). However, high-status individuals did not invest in monumental public construction (mounds or pyramids) or in elaborate tombs either (Grove, 1997).

All of these data point to the fact that power was highly personalized and institutional structures or infrastructure that could control the agency of political officials, distribute public goods, accommodate voice, monitor taxpayer compliance, and justly collect taxes were absent. Accordingly, evidence of even the most basic public goods is absent during the Early Formative from San Lorenzo. Furthermore, structures that would have allowed for the transgenerational transference of power were also absent. As a result, the political arena was highly competitive and instability was pervasive. Lineages were probably only able to maintain power for a generation or two at best, and power most likely shifted frequently among lineages.

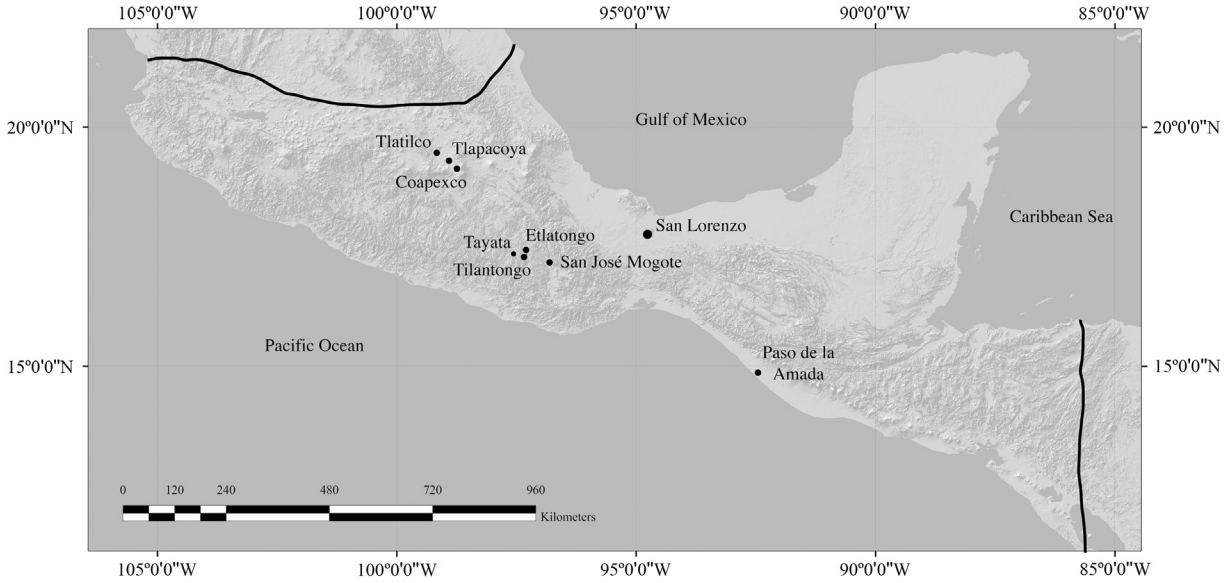

Figure 3 Map of Mesoamerica during Early Formative (boundaries marked by solid line, size of circle corresponds with relative population size).

San José Mogote

The political situation at this site was very different than the situation at San Lorenzo. It appears that at San José Mogote, personal wealth acquired through control over prestige goods was not used to gain political power. A large number of households engaged in the production of different prestige goods, including ornaments using shell from the Pacific and Gulf coasts, chert bifaces, and magnetite mirrors and traded for mica, jade beads, serpentine, and carved pottery (Marcus and Flannery, 1996). The only prestige good that seems to have been tightly controlled was magnetite mirrors, but this control did not provide households with access to more or better exotic goods nor to political power (Marcus and Flannery, 1996). Thus, despite the exchange and consumption of prestige goods, marked social inequality in houses and burials was limited and residences comparable to the high-status dwellings at San Lorenzo were absent (Marcus and Flannery, 1996). Apparently, the residents of San José Mogote developed an inclusive or collective political apparatus that effectively separated wealth from political power. Blanton et al. (1999) suggested that this was accomplished by a concept of dual opposition between male descent groups, one associated with earth and the other with sky. Instead of pouring revenues into extravagant houses, they were used to erect a public building. During the late Early Formative, the residents of San José Mogote along with residents from neighboring sites constructed a two-tier stone-faced platform 3.5-m high (Marcus and Flannery, 1996). This edifice functioned as a staging area for inclusive community-wide public rituals, large-scale feasting, and meetings among the site's household heads to resolve disputes and collect foodstuffs for rituals and other public functions, among other things.

Paso de la Amada

Although partially contemporaneous with San Lorenzo and San José Mogote, the residents of this site did not participate in the 'Olmec' exchange network; thus, Early Horizon markers have not been recovered from this site. Instead wealth and power were demonstrated by access to labor for the construction of large structures raised on platforms, which functioned as both dwellings and ritual/public spaces (Lesure, 1997, 2011; Lesure and Blake, 2002). These structures were surrounded by small residences, forming clusters distributed across the site. By the middle part of the occupation, construction at all of the large structures ceased except at Mound 6, which was reconstructed several times over a period of about 200 years and became the largest structure at the site (Lesure, 1997, 2011). Apart from the large and small dwellings there is no other architecture at the site except a single ballcourt (Lesure and Blake, 2002), but, it is not clearly attached to any of the large structures (Lesure, 2011). Evidence of plazas or other public spaces is debatable (see descriptions in Lesure, 2011).

Surprisingly, burials and artifact distributions indicate that there were only minor differences in wealth among the residents of the large and small dwellings (Lesure and Blake, 2002). Exotic items were comparatively infrequent in all contexts, but regularly distributed among households and burials. None of the burials or houses showed significant concentrations of items that set them apart as higher status (Lesure and Blake, 2002). The only artifact category that differs in distribution is the large hollow ceramic statue, which probably had a ritual function and has only been recovered from large structures (Lesure and Blake, 2002).

These data indicate an emphasis on internal revenue, specifically labor or corvée supplied by independent (free) households. This labor was deployed in the construction of the large houses but was not used to produce prestige goods for an emergent 'elite.' That is to say, leaders did not directly control labor and could not use it to increase their personal wealth and attract clients or coerce independent individuals. So what motivated the residents of small dwellings to participate in the construction of large dwellings? Lesure and Blake (2002) suggest that the individuals who participated in the construction of large houses may have had a claim on their use. They probably were entitled to participate in events such as dances, drinking events, rituals, social gatherings, political meetings, and gambling that occurred within the structures. However, their participation in construction was voluntary and probably based on an assessment of the potential benefits as well as an evaluation of the degree to which the leader upheld his or her end of the bargain on previous occasions. Where the benefits were well supplied and the leader did not exploit his position, individuals probably were willing to participate in future construction events. However, where the leader did not uphold his end of the bargain, individuals would refuse to cooperate or even shift their allegiance to other leaders.

Such a framework, based on a theory of contingent cooperation (that is to say, individuals choosing to participate or refusing to do so based on a case-by-case consideration of the costs and benefits of participation), fits well with the archaeological record. Some clusters never constructed a large dwelling (they failed to cooperate), other large dwellings were constructed, used, rebuilt, then suddenly abandoned, and sometime later reoccupied or remodeled (contingent cooperation). However, in the case of one cluster, cooperation was sustained over an extended period of time and the large dwelling and its platform (Mound 6) were rebuilt several times. During each construction phase, the dwelling was successively enlarged probably indicating that the size of the cooperating group expanded over time. Then suddenly cooperation failed and the dwelling was abandoned.

Thus, the political aspirations of Paso de la Amada leaders were held in check by contingent cooperation. My systematic cross-cultural research indicates that leaders in these contexts have much prestige and influence but little authority (power over others) (see also Lesure and Blake, 2002). Political decisions are consensus-based and community or group members are free to cooperate or to defect. In these situations, leaders may claim hereditary rights and live in larger (but not usually better made) dwellings, but any attempt to divert resources for personal gain or engage in coercion are counteracted by defection or exit (group members refuse to cooperate or join other groups). Conversely, leaders are expected to continuously supply members with services and the use of their dwellings as semipublic space in order to maintain the cooperation of group members.

Case II: Classic Period (1750–1250 Years Ago)

During the Terminal Formative (2150–1750 years ago) (Figure 4), collectivity reached a noir. Across Mesoamerica, political strategies generally shifted toward coercion, exclusion, and exploitation by emergent elites. In the Valley of Oaxaca, the highly collective or corporate Monte Albán state shifted toward a prestige goods economy (Elson and Sherman, 2007; Kowalewski et al., 1989). On the Pacific Coast of Oaxaca, the highly collective polity at Rio Viejo went into decline (Joyce and Barber, n.d.). At Teotihuacan, an emergent elite began to materialize its control over resources in the form of elaborate tombs, one under the Pyramid of the Feathered Serpent and another 'shaftlike' tomb constructed on the future site of the Pyramid of the Sun (Cowgill, 2008; Sugiyama, 2005; Sugiyama et al., 2013.) In West Mexico, high-status individuals constructed spectacular shaft tombs and filled them with massive amounts of funerary offerings (Pollard, 1997). Finally, the Maya of the Central Peten shifted toward an emphasis on personal power expressed in individualized portraiture of rulers, royal genealogies, and elaborate tombs (Sharer, 1994). Such patterns in the archaeological record suggest an underlying, pan-Mesoamerican, shift toward external revenues, a decline in cooperation between constituents and principals, and a corresponding increase in the ability of principals to behave despotically. Such was the political situation at the outset of the Classic Period when a startling shift in political policy took root in the western half of Mesoamerica.

Teotihuacan

Principals at Teotihuacan apparently shifted financial policy back toward internal revenues sources at the beginning of the Classic Period (Figure 5), promoting especially agriculture production and craft specialization, and curtailing the use of prestige goods in the political arena. Accordingly, they limited exchange in prestige goods and directed exotic goods into public rituals and offerings and away from personalized or individualized consumption. Thus, a growing number of scientists are coming to the conclusion that principal authority at Teotihuacan was vested in a ruling council (Blanton et al., 1996; Cowgill, 1997; Millon, 1976). This council apparently focused on providing public goods and voice to Teotihuacanos in exchange for compliance with taxation paid in kind and through corvée. Specifically, it provided public spaces for citizens to assemble, express their voice, and participate in mass rituals and feasts, as well as a highly planned urban landscape that included (Cowgill, 1997; Millon, 1973) the following:

- the construction of the massive and accessible pyramids and plazas, roads, and drains;
- the organization of the city on an orthogonal grid centered on the street of the dead (a massive avenue that extended 5 km) and a massive east – west avenue that intersected the former avenue at the geographical center of the city, with secondary streets and alleys connected to them;
- the redirection of the San Juan River, which runs through the city; and
- the relocation of residents in more than 2000 highly standardized, internally complex, multifamily, masonry residential compounds, which were arranged in comparatively standardized neighborhoods or wards that may have in turn been clustered in standardized administrative districts (Gómez-Chávez, 2012).

Social stratification, as well as wealth and power differences, was strongly muted following political reorganization at the onset of the Classic. There is not a single royal interment in any of the major monumental architectural complexes at this time and 'elite' or 'royal' burials dating to the Formative were systematically looted (see Sugiyama and Cabrera Castro, 2007; Sugiyama et al., 2013). Teotihuacanos were buried under the floors of administrative complexes or the central patios of their residential compounds with relatively few offerings

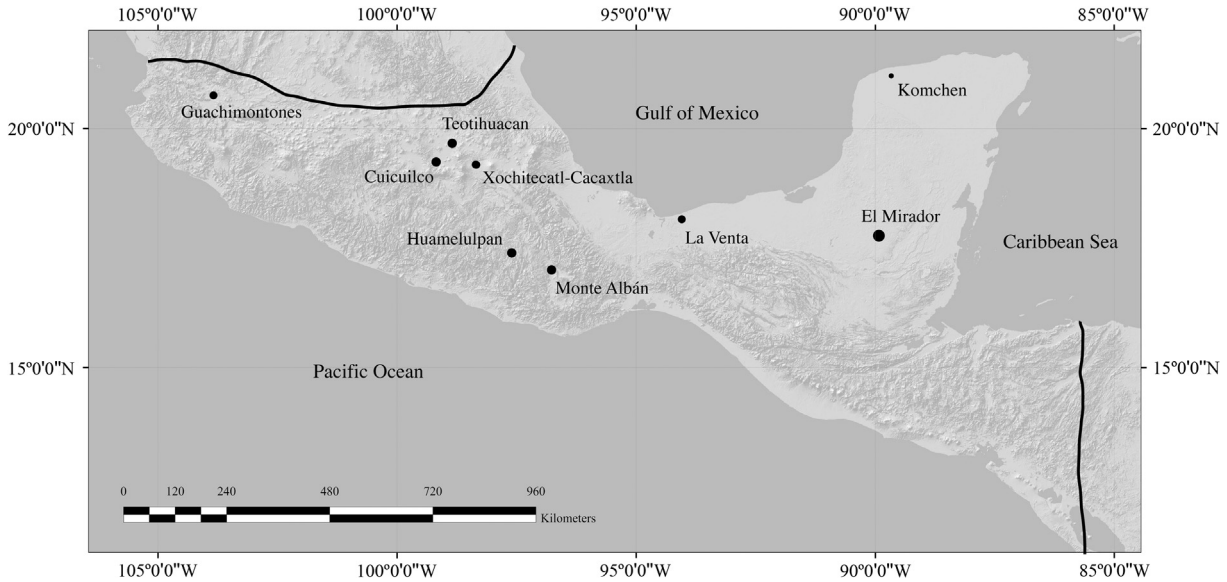

Figure 4 Map of Mesoamerica during the Late-Terminal Formative (boundaries marked by solid line, size of circle corresponds with relative population size).

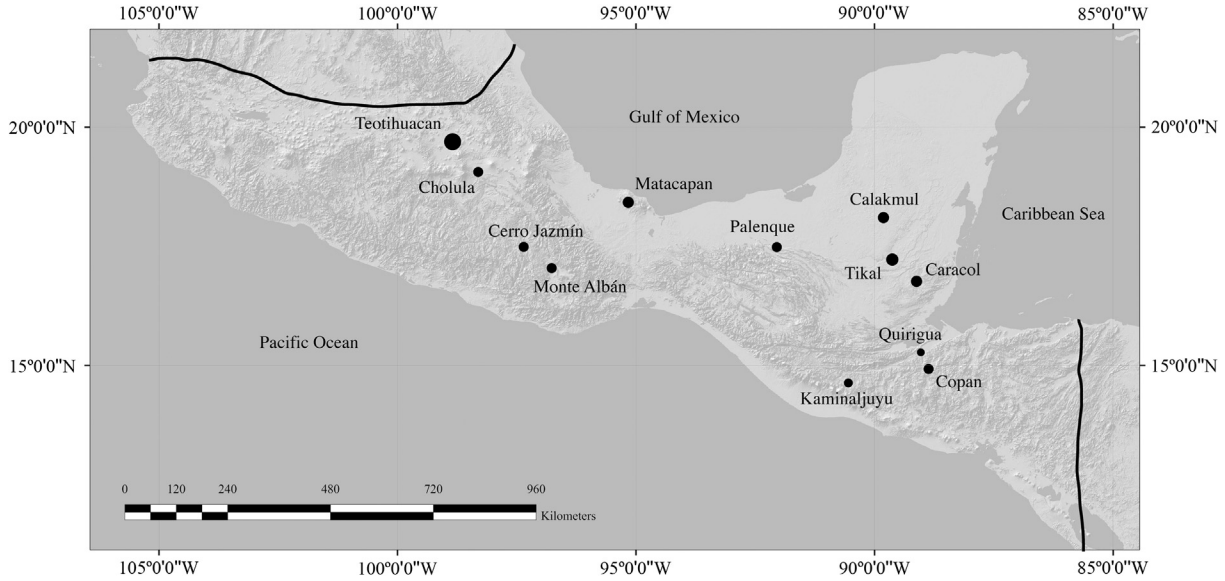

Figure 5 Map of Mesoamerica during the Classic (boundaries marked by solid line, size of circle corresponds with relative population size).

(Gómez-Chávez, 2012). The most elaborate and complex objects consisting of composites of imported materials were used in state rituals and buried in the numerous temples that dot the city (Sugiyama and Cabrera Castro, 2007; Sugiyama et al., 2013). Named lineages and clear images of rulers do not appear in public contexts (Blanton et al., 1996). Instead, public murals and carved stone depict processions with large numbers of individuals, individuals wearing masks, and apparel hidden by feathers, which induces a sense of facelessness and collectivity (Blanton et al., 1996). Finally, not a single residence stands out as exceptionally large and opulent (a 'royal' palace) (Cowgill, 1997).

Monte Albán

In the Valley of Oaxaca (**Figure 5**), research indicates that the Monte Albán state shifted toward collectivity in the Early Classic Period (1750–1450 years ago). Revenue strategies appear to have shifted from control over the production and distribution of prestige goods (especially cream-ware pottery) to agricultural taxes and corvée (Kowalewski et al., 1989). Coupled with this change, political architects at Monte Albán focused on the bureaucratization of public administration (Fargher, n.d.). These processes unfolded as follows. First, key centers of provincial power during the Terminal Formative, which were centered on elite residential, funerary, and ritual complexes, were abandoned and replaced by new 'bureaucratic' centers that were located in previously unoccupied sites and positioned to form a continuous network of regularly spaced centers extending out from Monte Albán (Kowalewski et al., 1989). Second, these new centers focused on open and accessible plazas (Blanton, 1989). Adjacent to these plazas, the state constructed specialized (nonresidential) administrative buildings (see Blanton, 1989; Feinman and Nicholas, 2011). Moreover, the organization and orientation of these administrative complexes were highly uniform across the entire valley indicating their construction was directed by the state (Blanton, 1989). The officials that occupied these structures were 'faceless' and did not claim these structures as personal property (Feinman and Nicholas, 2011). Thus, it appears that public officials (probably sent from other localities) and not local 'rulers' or chiefs administered these centers and their immediate hinterlands. Finally, a similar architectural pattern was used to reorganize the city of Monte Albán into administrative wards where each ward was centered on a public plaza-mound complex (Blanton, 1978).

The outcome of bureaucratization was twofold. First, the design of these new centers placed public plazas (which functioned, among other things, as marketplaces in pre-Hispanic Mesoamerica) under the administration of the state, allowing officials to maintain market security and ensure just and equitable exchange (Kowalewski et al., 1989). Second, it enhanced the ability of the state to provide public security within the valley. Following these political chances, many households migrated to the valley floor and lived in dispersed, unprotected settlements, which indicates that the state was protecting rural peasants from raiding and banditry (Kowalewski et al., 1989).

In keeping with the expectations from collective action theory, data from Monte Albán indicate constraints on the ruler's control of material and ideological resources along with the possibility of political structures that encouraged power sharing. Although high-status residences and tombs (which were constructed under houses) were more elaborate than those of ordinary households, they are comparable in scale and wealth to 'royal' installations (Joyce, 2010; Marcus and Flannery, 1996). Importantly, royal tombs were not placed in massive pyramids and intact tomb offerings suggest relatively modest displays as compared with Classic Maya tombs. In more public contexts, personalized images of rulers are comparatively rare, lengthy royal genealogies are nonexistent and carved stone monuments were frequently recycled (Joyce, 2010). These data suggest frequent shifting of the highest political office at Monte Albán among unrelated

individuals or lineages, but these shifts occurred in the absence of internal strife or civil wars. Additionally, a carved stone monument shows a king surrounded by a group of high-status individuals who may have shared power with the ruler (Joyce, 2010). Taken as a body, these data hint at the possibility that an advisor council shared power with the king and may have even intervened in succession processes and elected rulers from a group of candidates that included individuals from different lineages.

Tikal

Tikal contrasts sharply with Teotihuacan and Monte Albán. At this site, external revenue strategies dominated during the Early Classic Period (1750–1450 years ago) (**Figure 5**). Rulers there focused on the accumulation and use of personal wealth to attract and maintain clients (Blanton et al., 1996). Coercion, involving defeat and subjugation of competing rulers at neighboring sites, was also used to acquire revenues and to control recalcitrant clients (Sharer, 1994). The city's location between two river systems, one flowing to the Gulf of Mexico and the other to the Caribbean Sea, allowed it to control the flow of valuables across the region (Sharer, 1994). Rulers may also have controlled water by capturing rainfall in massive reservoirs, both emphasizing their symbolic power and allowing for the irrigation of nearby agricultural fields that were probably claimed as the personal property of the ruler (Scarborough, 1998). Tikal's rulers monopolized ideological resources by claiming special descent, personal links with deities, and direct control over sacred monuments like the city's massive pyramids. On carved stone, Tikal's early kings claimed descent from an individual who originated in a place that epigraphers believe is Teotihuacan (Sharer, 1994).

Revenues accruing from monopolies on long-distance exchange and ideological resources, as well as on the land fed by royal reservoirs, were plowed into conspicuous consumption. Kings and the high nobility were depicted on carved stone, polychrome pots, and murals dressed in elaborate headdresses and fine jewelry and clutching symbols of power (e.g., Sharer, 1994). They are shown seated on thrones, which have been recovered archaeologically, and mats. They used fancy polychrome pottery, jade, objects made of other types of green rock, elaborate shell artifacts, highly complex chert eccentrics, and green obsidian (from Central Mexico) (Sharer, 1994). They lived in beautiful, elaborate, and massive palaces. The massive Central Acropolis at Tikal was home to kings over many generations. Upon their death, kings and queens were interred in elaborate tombs buried under personalized funerary pyramids.

Although Tikal was the most important state in the Central Maya Area during the Early Classic it did not control the entire Maya Area. Instead, the landscape of the Central Maya Area was highly fragmented and states fought with one another to gain prestige, which was necessary for recruiting and maintaining client states. The principal figures in this bloody, jewel-encrusted drama were Tikal, Calakmul, Caracol, Copán, Palenque, Rio Bec, and Altar de Sacrificios, to name just a few (Sharer, 1994).

Epiclassic – Early Postclassic (1250–750 Years Ago)

Around 1250 years ago, a marked change in political policy occurred and the major Classic Period states went into decline (**Figure 6**). The following Epiclassic Period (1250–1050 years) in many ways marked a return to strategies that were dominant during the Terminal Formative. Across much of Mesoamerica, principals emphasized external revenue strategies based on monopolies over ideological and economic resources (especially long-distance exchange) as well as coercion. In Central Mexico, Teotihuacan's introverted and complicated style gave way to unrestrained audacious displays of wealth and

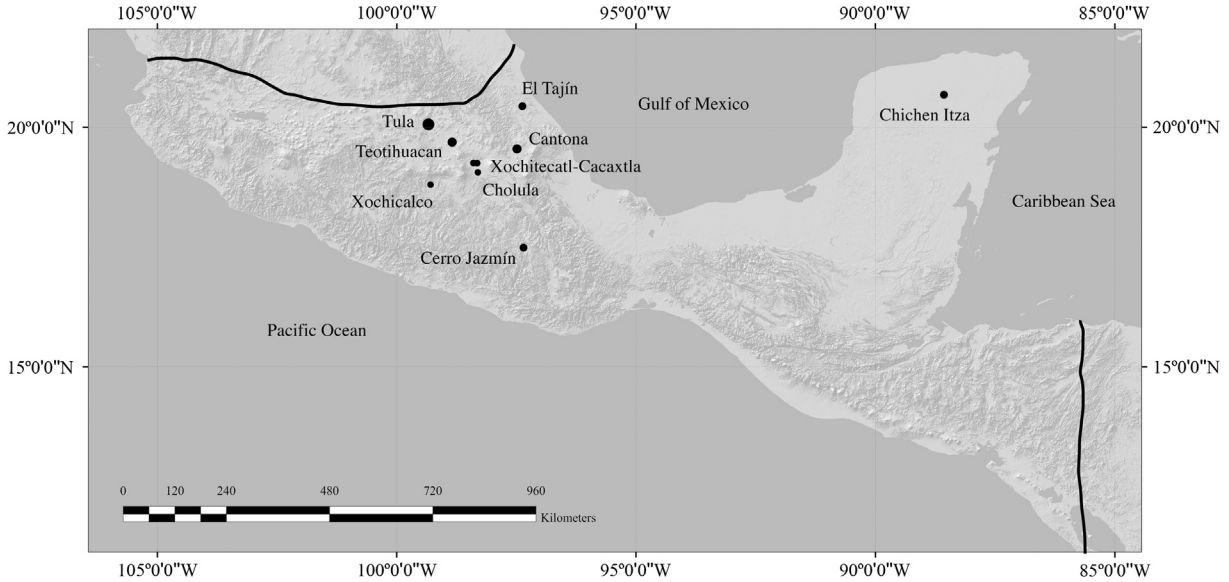

Figure 6 Map of Mesoamerica during the Epiclassic-Early Postclassic (boundaries marked by solid line, size of circle corresponds with relative population size).

personal power. For example, kings built massive and luxurious palaces at Xochitecatl, Cacaxtla, and Xochicalco (Hirth, 2000; Lombardo de Ruiz et al., 1986; Santana et al., 1990; Serra Puche, 1998). The palace on the top of Cerro Cacaxtla is truly exceptional. Here, resplendent murals are preserved that depict war, fertility, Venus figures, and possibly rulers or other high-status individuals linked with supernatural forces (e.g., a feathered serpent and a jaguar) (Stuart and Ferorelli, 1992). In the Valley of Oaxaca, the bureaucrats of the Monte Albán state began claiming individualized personal power. They began enclosing the public plazas, making them private, constructing impressive subfloor tombs under formerly administrative building, and carving stone stalæ showing their personal genealogies and marriages (Fargher, n.d.; Feinman and Nicholas, 2011). The provincial elites that had been suppressed during the Classic Period apparently reemerged, reclaiming and rebuilding their Terminal Formative centers of power with palaces and tombs (Fargher, n.d.). In the Maya lowlands, military competition and prestige good consumption spiked (Sharer, 1994). The major regional powers (e.g., Tikal, Calakmul, Caracol) collapsed and the political landscape fragmented further. For a brief period during the Early Postclassic (1050–750 years), principals at Tula (Central Mexico) and Chichen Itza (Northern Yucatan) may have shifted political policy back toward collectivity to some degree (Blanton et al., 1996). But much work still needs to be done to understand the nature of the changes.

Case III: Late Postclassic (750–500 Years Ago)

The final period in the pre-Hispanic sequence began about 800 years ago and ended with Spanish conquest in AD 1521 (**Figure 7**). In Central Mexico, West Mexico, and Oaxaca, this period was marked by long-term population growth, increasing commercialization, and the gradual political consolidation of three major empires: (1) the Aztec empire (Aztec Triple Alliance), which controlled large parts of Central Mexico, the Gulf Coast, and Oaxaca, as well as the Pacific Coast of Chiapas and small polities on the Coast of West Mexico in modern Michoacan and Guerrero; (2) the Tarascan Empire which controlled much of West Mexico; and (3) the Tututepec Empire which controlled the Pacific Coast of Oaxaca extending from the Isthmus of Tehuantepec into Guerrero (Berdan et al., 1996; Joyce et al., 2004; Pollard, 1982). Overall, Mesoamerica reached its highest degree of cultural and economic integration during this period (Blanton et al., 1999).

Aztec Triple Alliance

Following the collapse of Teotihuacan, the Basin's population dispersed and eventually concentrated around a large number of competitive city-states (Sanders et al., 1979; Smith, 2008). Then around 650 years ago, the political and demographic landscape shifted markedly and the political landscape consolidated around three major blocks, which were ultimately united under a single capital by the ruler (or *tlatoani*) of Azcapotzalco (van Zantwijk, 1985). However, around 1426 the ruler of Azcapotzalco died and a succession dispute erupted. During the ensuring war, in 1427 or 1428, three Mexica-Tenochca relatives (Itzcoatl, Motecuhzoma, and Tlacaelel) and their Acolhua cousin (Nezahualcoyotl) conspired to take control of Tenochtitlan, Texcoco, and then, with the help of Tlacopan (or Tacuba), the entire Basin of Mexico (Fargher and Blanton, 2007). Over the next 70 years, population growth and commercialization surged under the internal peace and imperial wealth brought by the Aztec Triple Alliance (Blanton et al., 2005; Sanders et al., 1979). Tenochtitlan-Tlatelolco became the largest city in the history of Mesoamerica, with a population of 150 000–200 000 people (Sanders et al., 1979).

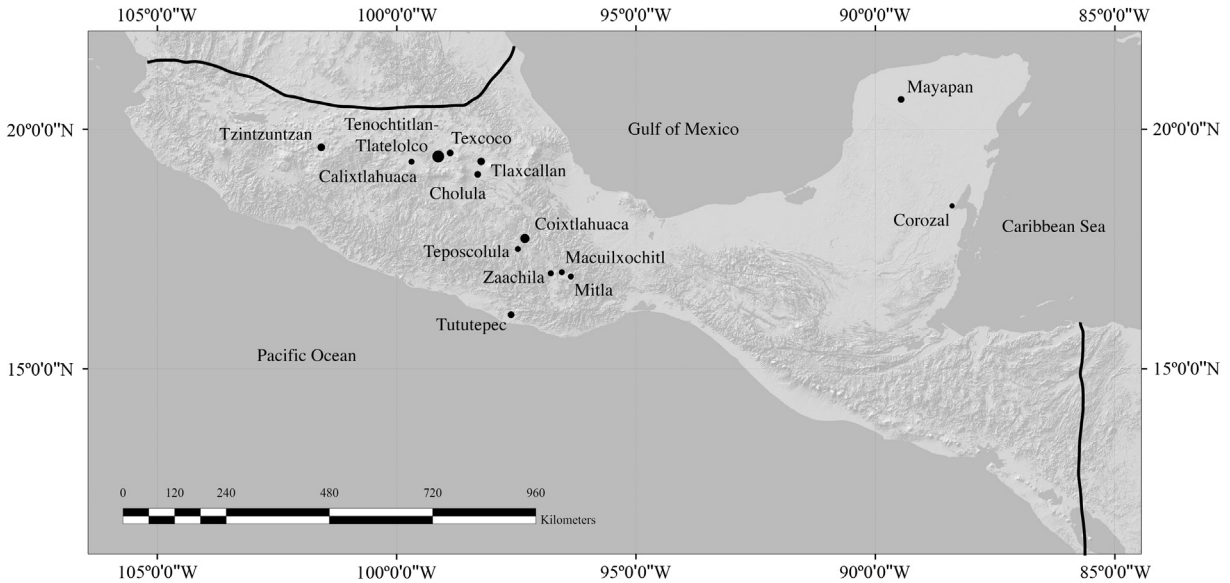

Figure 7 Map of Mesoamerica during the Late Postclassic (boundaries marked by solid line, size of circle corresponds with relative population size).

The Aztec Triple Alliance pursed a mixed revenue strategy. The principal internal revenues sources were taxes on ordinary market transactions, which were collected at the numerous marketplaces distributed across the Basin of Mexico, agricultural taxes and corvée on free peasants, and imperial tribute collected by the state (Fargher and Blanton, 2007). External revenues accrued to the principals of the Aztec Triple Alliance through ownership of agricultural estates worked by dependent laborers and laborers attached to their personal residences (Fargher and Blanton, 2007). However, these revenues were relatively minor compared to internal revenues because free peasants held the majority of agricultural land in the Basin of Mexico.

In order to encourage compliance with revenue demands, the Aztec Triple Alliance implemented bureaucratization, distributed public goods, accommodated voice, and placed limits on the agency of principals (Fargher and Blanton, 2007). First, the Aztec Triple Alliance was governed by a council called the Tlatocan, which consisted of the *tlatoani* (ruler) of Tenochtitlan, the *tlatoani* of Texcoco, the *tlatoani* of Tacuba, the *cihuacoatl* (coruler) of Tenochtitlan, and four prime ministers (the four prime ministers of the Triple Alliance always included at least one peasant). The tlatoani of Tenochtitlan was the most important and influence individual member of the Aztec government and carried the title of hueyi tlatoani (literally great speaker). This council had the power to monitor and punish every single political official in the government including the *hueyi tlatoani* and exercised this power. Second, low-level administration was divided into a series of bureaus (e.g., tax collection, policing and justice, military direction, and a 'complaints' department). Officials were selected in part on merit and commoners could enter bureaucratic service; they were trained at special schools; they were carefully monitored and corrupt officials were punished; and they were paid an in-kind salary drawn from lands assigned to each office. Third, the Triple Alliance provided extensive public goods, especially within the capital (Tenochtitlan) but across the territory of the empire as well (Fargher and Blanton, 2007). The primary public goods were road and bridge maintenance, aqueducts, irrigation and drainage canals, dykes, public parks, plazas, redistribution during famines, and public security. Finally, the architects of the Aztec empire created a court system with a precise appeal hierarchy that treated all members of the Aztec society as equals regardless of social status ('judicial equality').

Mixteca Alta

During the late Postclassic, the Mixteca Alta of Oaxaca was divided among many competing petty kingdoms or states each ruled by a royal lineage and a small hereditary nobility that was probably closely related to the ruling line (Fargher et al., 2011; **Figure 7**). In contrast to the Aztec Triple Alliance, revenue sources were strongly external. The key revenue sources were monopolies over irrigated agricultural lands, dependent serflike laborers, and long-distance exchange in prestige goods (Fargher et al., 2011). Rulers used these revenues to build lavish palaces and adorned themselves with beautiful jewelry. The key durable prestige goods included resplendent pottery painted with images of the vessels' owners and supernatural themes, gold, silver, gold–silver alloys, copper, marine shell objects, rock crystal, intricately carved bone, and polished obsidian jewelry (Caso and Borbolla, 1969). Rulers used these goods to attract and maintain networks of clients that carried out their political will. They did this instead of investing in public goods or accommodating taxpayer voice (Fargher et al., 2011). As a result, the Mixteca Alta was riddled by internal strife and violent confrontations among the petty rulers, resulting in low public security.

Tarascan Empire

Finally, the Tarascan Empire centered on the Patzcuaro Basin in Central Michoacan is markedly different than the Aztec Triple Alliance and the petty states of Oaxaca (**Figure 7**). Tarascan revenue policy strongly emphasized external revenues and elite consumption, like the Mixteca Alta, which resulted in low collectivity (summarized from Pollard, 1987). However, unlike highland Oaxaca, the Tarascan state used this strategy to build a highly centralized empire. Centralization was achieved by extensive monopolies over economic and ideological resources. Within the basin, the royal family claimed direct control over vast swaths of high-quality agricultural land, as well as a large number of dependent labors, artisans, and merchants. Merchants who specialized in long-distance exchange worked exclusively for the king, bringing in exotic prestige goods like chocolate, animal skins, feathers, turquoise, gold, silver, slaves, and many other products from the edges of the empire and beyond. However, merchants did not sell any of these items because all of it was owned by the king. The ruler also directly controlled copper mines worked by dependents and collected tribute from workers exploiting other mines. Finally, the king received tribute from conquered polities outside the basin.

The king was the sole receiver and distributor of these revenues. They were used principally to fund consumption on a massive scale by the royal household, state religious ceremonies, and to control the nobility. The king constructed a colossal palace set atop a platform 10 times larger than Motecuhzoma II's palace (Pollard, 1980; Smith, 2003). He also constructed a religious complex consisting of one large and two smaller platforms, which were used to store massive amounts of prestige goods. The remainder of the revenue was gifted to the high nobility, who constructed their residences in a cluster surrounding the royal palace (Pollard, 1980).

Not surprisingly, public goods and bureaucratization were not a high priority of the king. The capital (Tzintzuntzan) is described as an unplanned settlement and no mention is made of roads, aqueducts, or public spaces in written records (Pollard, 1980). The state showed no interest in administering markets (e.g., offering dispute resolution and consumer protection, etc.) and peasant markets were marginalized, operating completely independently of the state-based prestige goods economy (Pollard, 1982). Although the state collected maize and cotton tribute, these products were not redistributed to commoner or peasant households (Pollard, 1982). The degree of bureaucratization appears to have been very low. The chief political officials were part of a hereditary nobility and were tied personally to the king through gift exchanges involving prestige goods, land, and service to the state. The lower level nobility probably was

in turn personally tied to members of the high nobility through similar gift exchanges. Thus, elite status was defined by access to state-produced prestige goods and commoners, except successful warriors, were completely excluded from these exchanges, resulting in clearly divided social classes and stratification (Pollard, 1982; Pollard and Cahue, 1999).

Discussion

The cases described here, while representing only a fraction of Mesoamerica's ancient diversity, illustrate a number of important features. First, the most common pattern in the revenue strategies and political organization in Mesoamerican states were external revenues based on control of or monopolies on interregional prestige goods exchanges. This pattern is logical because control of external revenues is a more cost-effective strategy for principals to employ. Thus, the cases in which principals focused on internal revenues, Teotihuacan, Aztec Triple Alliance, and Monte Albán, stand out. It also appears that collective states based on internal revenues that separated wealth acquired through control of prestige goods from political power were the only ones to influence the entirety of the Mesoamerican world during their apogees (e.g., Teotihuacan, Tenochtitlan). Second, cases where principals were able to fully monopolize prestige goods exchanges at the interregional scale and use the wealth derived from this monopoly to build a centralized and stable political structure were also rare (the Tarascan Empire is the only example). Outside of Mesoamerica, this pattern is rare also. A recent global analysis (Blanton and Fargher, 2008) found only one other case (Egypt), where principals were able to fully monopolize interregional exchange. Thus, this is an exceptionally rare pathway to complexity, probably because such complete monopolies are so difficult (and costly in human lives) to achieve.

The cases described in this article provide important information on the relationship between collectivity and long-distance exchange. Building collective polities in a world economy poses unique challenges for principals. In all of the cases that emphasized internal revenues, principals had to develop a combination of collective strategies that devalued prestige goods and placed controls on political actors to keep them from using prestige goods to disrupt political policy. In the case of Teotihuacan, principals used the hegemonic position of Teotihuacan in the Classic Period Mesoamerica world to constrain the flow of prestige goods and direct it into public consumption locally and into diplomatic arenas outside of the state's territory. Teotihuacan's political architects also developed strong controls on rulers and political officials, which allowed the state to collect and even make large amounts of prestige goods without disrupting collectivity. Monte Albán's principals appear to have capitalized on Teotihuacan's international policy regarding prestige goods to build an internal revenue system. They too appear to have built political structures that separated wealth from power and constrained the agency of the king. Conversely, the Aztec Triple Alliance capitalized on interregional flows of wealth items to stimulate commercialization and, thus, market revenues. Instead of squashing the international economy and the interregional exchange of wealth items, Aztec principals dumped wealth items paid in tribute on the market. This strategy essentially made wealth items available to everyone at comparative low cost (turning them into bulk luxury goods), which effectively killed their value as political currency. At the same time, they built a political structure that effectively controlled the agency of principals and political officials alike, which allowed them to pursue internal revenues.

Thus, these data seam to hint at 'core strategies' in ancient Mesoamerica that are similar to 'core strategies' in the modern world economy. These strategies include the following: (1) strategies to keep wealth generated by international trade out of the political arena; (2) an emphasis on internal revenue strategies; (3) collectivity (bureaucratization and public goods); and (4) controls on the agency of principals so that even if they get a hold of or posses great wealth they cannot use it to act despotically. Conversely, 'periphery strategies' emphasize attempts to monopolize interregional interaction and use the revenue it provides to recruit clients through gifting, conspicuous consumption, and coercion. Moreover, attempts at monopolizing interregional exchange tend to correlate with a politically fragmented landscape and instability (e.g., frequent and violent confrontations among petty rulers competing for power).

See also: Americas, Sociocultural Overviews: Mexico; Chiefdoms, Archaeology of; Civilizations; States and Civilizations, Archaeology of; Trade and Exchange, Archaeology of.

Bibliography

Berdan, F.F., Blanton, R.E., Boone, E.H., Hodge, M.G., Smith, M.E., Umberger, E., 1996. Aztec imperial Strategies. Dumbarton Oaks Research Library and Collection, Washington, DC.

Blanton, R.E., 1989. Continuity and change in public architecture: periods I through V of the Valley of Oaxaca, Mexico. In: Kowalewski, S.A., Feinman, G.M., Finsten, L., Blanton, R.E., Nicholas, L.M. (Eds.), Monte Albán's Hinterland Part II. Museum of Anthropology, University of Michigan, Ann Arbor.

Blanton, R.E., Fargher, L.F., 2008. Collective Action in the Formation of Pre-Modern States. Springer, New York.

Blanton, R.E., 1978. Monte Alban: Settlement Patterns at the Ancient Zapotec Capital. Academic Press, New York.

Blanton, R.E., 1996. The Basin of Mexico market system and the growth of empire. In: Berdan, F., Blanton, R., Boone, E.H., Hodge, M.G., Smith, M.E., Umberger, E. (Eds.), Aztec Imperial Strategies. Dumbarton Oaks, Washington, DC.

Blanton, R.E., Fargher, L.F., Espinoza Heredia, V.Y., 2005. The Mesoamerican world of goods and its transformations. In: Blanton, R.E. (Ed.), Settlement, Subsistence, and Social Complexity: Essays Honoring the Legacy of Jeffrey R. Parsons. Cotsen Institute of Archaeology, UCLA, LA.

Blanton, R.E., Feinman, G.M., Kowalewski, S.A., Nicholas, L.M., 1999. Ancient Oaxaca. Cambridge University Press, New York.

Blanton, R.E., Feinman, G.M., Kowalewski, S.A., Peregrine, P.N., 1996. A dual-processual theory for the evolution of Mesoamerican civilization. Current Anthropology 37, 1–14.

Carr, R.F., Hazard, J.E., 1961. Map of the Ruins of Tikal, El Peten, Guatemala. The University Museum, University of Pennsylvania, Philadelphia.

Caso, A., Borbolla, D.R.de la, 1969. El tesoro de Monte Albán. Instituto Nacional de Antropología e Historia, México, DF.

Chase, A.F., Chase, D.Z., Smith, M.E., 2009. States and empires in ancient Mesoamerica. Ancient Mesoamerica 20, 175–182.

Clark, J.E., 1997. The arts of government in early Mesoamerica. Annual Review of Anthropology 26, 211–234.
Cowgill, G.L., 1997. State and society at Teotihuacan, Mexico. Annual Review of Anthropology 26, 129–161.
Cowgill, G.L., 2008. An update on Teotihuacan. Antiquity 82, 962–975.
Cyphers, A., 1996. Reconstructing olmec life at San Lorenzo. In: Benson, E.P., de la Fuente, B. (Eds.), Olmec Art and Ancient Mexico. National Gallery of Art, Washington, DC.
Elson, C.M., Sherman, R.J., 2007. Crema ware and elite power at Monte Albán: ceramic production and iconography in the Oaxaca Valley, Mexico. Journal of Field Archaeology 32, 265–282.
Fargher, L.F., Corporate power strategies, collective action, and control of principals: a cross cultural perspective. In: Fargher, L.F., Heredia Espinoza, V.Y. (Eds.), Alternative Pathways to Complexity: Households, Markets, World Systems, and Political Economy: Essays Honoring the Legacy of Richard E. Blanton, n.d.
Fargher, L.F., Blanton, R.E., 2007. Revenue, voice, and public goods in three premodern states. Comparative Studies in Society and History 49, 848–882.
Fargher, L.F., Heredia Espinoza, V.Y., Blanton, R.E., 2011. Alternative pathways to power in late postclassic highland Mesoamerica. Journal of Anthropological Archaeology 30, 306–326.
Feinman, G.M., Nicholas, L.M., 2011. The ballcourt at El Palmillo: implications for late classic Oaxaca, Mexico. Mexicon 33, 98–104.
Gómez-Chávez, S., 2012. Structure and organization of neighborhoods in the ancient city of Teotihuacan. In: Arnauld, M.-C., Manzanilla, L., Smith, M.E. (Eds.), The Neighborhood as a Social and Spatial Unit in Mesoamerican Cities. University of Arizona Press, Tucson.
Grove, D.C., 1997. Olmec archaeology: a half century of research and its accomplishments. Journal of World Prehistory 11, 51–101.
Hirth, K., 2000. Ancient Urbanism at Xochicalco. University of Utah Press, Salt Lake City.
Joyce, A.A., 2010. Mixtecs, Zapotecs, and Chatinos: Ancient Peoples of Southern Mexico. Wiley-Blackwell, Malden.
Joyce, A.A., Barber, S.B., Political authority and integration in later formative Oaxaca. In: Fargher, L.F., Heredia Espinoza, V.Y. (Eds.), Alternative Pathways to Complexity: Households, Markets, World Systems, and Political Economy: essays honoring the legacy of Richard E. Blanton, n.d.
Joyce, A.A., Workinger, A.G., Hamann, B., Kroefges, P., Oland, M., King, S.M., 2004. Lord 8 Deer "Jaguar Claw" and the land of the sky: the archaeology and history of Tututepec. Latin American Antiquity 15, 273–297.
Kirchhoff, P., 1952. Mesoamerica: its geographic limits, ethnic composition, and cultural characteristics. In: Tax, S. (Ed.), Heritage of Conquest. Free Press, Glencoe.
Kowalewski, S.A., Feinman, G.M., Finsten, L., Blanton, R.E., Nicholas, L.M., 1989. Monte Albán's Hinterland Part II: Rrehispanic Settlement Patterns in Tlacolula, Etla, and Ocotlan, the Valley of Oaxaca. Museum of Anthropology, University of Michigan, Mexico, Ann Arbor.
Lesure, R.G., 1997. Early formative platforms at Paso de la Amada, Chiapas, Mexico. Latin American Antiquity 8, 217–235.
Lesure, R.G., 2011. Paso de la Amada as a Ceremonial Center. In: Lesure, R.G. (Ed.), Early Mesoamerican Social Transformations: Archaic and Formative Lifeways in the Soconusco Region. University of California Press, Berkeley.
Lesure, R.G., Blake, M., 2002. Interpretive challenges in the study of early complexity: economy, ritual, and architecture at Paso de la Amada, Mexico. Journal of Anthropological Archaeology 21, 1–24.
Lombardo de Ruiz, S., de Molina, D.L., Molina, D.F., 1986. Cacaxtla. El lugar donde muere la lluvia en la tierra. Instituto Tlaxcalteca de la Cultura y Gobierno del Estado de Tlaxcala, Tlaxcala.
Marcus, J., Flannery, K.V., 1996. Zapotec Civilization: How Urban Society Evolved in Mexico's Oaxaca Valley. Thames and Hudson, London.
Masson, M.A., Freidel, D.A., 2012. An argument for classic era Maya market exchange. Journal of Anthropological Archaeology 31, 455–484.
Millon, R., 1973. Urbanization at Teotihuacan, Mexico. University of Texas Press, Austin.
Millon, R., 1976. Social relations in ancient Teotihuacan. In: Wolf, E.R. (Ed.), The Valley of Mexico. University of New Mexico Press, Albuquerque.
Oudijk, M.R., 2002. The Zapotec city-state. In: Hansen, M.H. (Ed.), A Comparative Study of Six City-State Cultures. The Royal Danish Academy of Sciences and Letters, Copenhagen.
Pollard, H.P., 1980. Central places and cities: a consideration of the protohistoric Tarascan state. American Antiquity 45, 677–696.
Pollard, H.P., 1982. Ecological variation and economic exchange in the Tarascan state. American Ethnologist 9, 250–268.
Pollard, H.P., 1987. The political economy of prehispanic Tarascan metallurgy. American Antiquity 52, 741–752.
Pollard, H.P., 1997. Recent research in west Mexican archaeology. Journal of Archaeological Research 5, 345–384.
Pollard, H.P., Cahue, L., 1999. Mortuary patterns of regional elites in the Lake Patzcuaro Basin of western Mexico. Latin American Antiquity 10, 259–280.
Sanders, W.T., Parsons, J.R., Santley, R.S., 1979. The Basin of Mexico: Ecological Processes in the Evolution of a Civilization. Academic Press, New York.
Santana, A., Vergara, S.D., Delgadillo, R., 1990. Cacaxtla, su arquitectura y pintura mural: nuevos elementos para su análisis. In: Cardés de Méndez, A. (Ed.), La Época Clásica, Nuevos Hallazgos, Nuevas Ideas. Instituto Nacional de Antropología e Historia, México, DF.
Scarborough, V.L., 1998. Ecology and ritual: water management and the Maya. Latin American Antiquity 9, 135–159.
Serra Puche, M.C., 1998. Xochitecatl, Tlaxcala, Gobierno del Estado de Tlaxcala.
Sharer, R.J., 1994. The Ancient Maya. Stanford University Press, Stanford.
Smith, M., 2003. The Aztecs. Blackwell, Malden.
Smith, M.E., 2008. Aztec city-state Capitals. University Press of Florida, Gainesville.
Stuart, G., Ferorelli, E., 1992. Ancient cacaxtla. National Geographic 182, 120–136.
Sugiyama, N., Sugiyama, S., Alejandro, S.O., 2013. Inside the Sun Pyramid at Teotihuacan, Mexico: 2008–2011 excavations and preliminary results. Latin American Antiquity 24, 403–432.
Sugiyama, S., 2005. Human Sacrifice, Militarism, and Rulership: Materialization of State Ideology at the Feathered Serpent Pyramid, Teotihuacan. Cambridge University Press, Cambridge.
Sugiyama, S., Cabrera Castro, R., 2007. The Moon Pyramid project and the Teotihuacan state polity: a brief summary of the 1998–2004 excavations. Ancient Mesoamerica 18, 109–125.
van Zantwijk, R.A.M., 1985. The Aztec Arrangement: The Social History of Pre-Spanish Mexico. University of Oklahoma Press, Norman.

Meta-Analysis, History of

David J Roelfs, University of Louisville, Louisville, KY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Meta-analysis is a type of systematic literature review through which a researcher can combine the results of existing studies to produce an overall estimate of the magnitude of a statistic such as a mean, correlation, or odds ratio. Meta-analyses are often coupled with meta-regressions, which are multivariate analyses of the effect of between-study differences on the magnitude of the statistic of interest. Developments to meta-analytic and meta-regression methodology are ongoing, with a trend emerging toward using the techniques to examine causal pathways.

Meta-analysis is a quantitative technique through which a researcher can combine the results of methodologically similar studies to produce a single, overarching statistical result.

The term 'meta-analysis' was coined, and the technique largely developed, in the late-1970s by statistician and educational psychologist Gene V. Glass (see Glass, 1976). The impetus for the development of meta-analysis was the lack of statistical power in the psychology experiments of the time. Lacking funds to conduct experiments on a large sample of people, psychologists would opt for small study sizes and, as a result, frequently find that their variables of interest were not statistically significant. By using a technique that allowed for the aggregation of existing research, meta-analysis allowed Glass and his colleagues to examine commonly studied variables with a much higher level of statistical power without incurring substantial financial costs.

Because meta-analysis can only be based on studies using the same empirical measures, it was first adopted by disciplines where studies were most directly replicated. Within the social sciences, the method was first put to use in psychology. In the first 10 years following the creation of formal meta-analysis methods, psychologists (including both educational and noneducational researchers) predominantly used the method to study treatment effectiveness (e.g., psychotherapy), the correlates of academic achievement, or human sex differences. In the next 10 years, psychologists began to use meta-analysis to study the effectiveness of various drug treatments (particularly for depression) and factors related to depression more broadly, while still retaining a focus on human sex difference research. In the third decade, meta-analysis was used in a more concentrated way for the study of all types of mental illness, with studies of depression and schizophrenia dominating the research. Today, psychologists continue to focus on drug therapies and other interventions for the major mental disorders. Over these four decades, the number of meta-analytic studies grew exponentially, from a few dozen in the first decade to several thousand in the most recent 10 years.

The method almost immediately spread outside of the social sciences, through psychiatry, to the medical sciences. Today, the medical sciences are the foremost users of meta-analysis, which has become one of several tools facilitating the ongoing shift toward what is known as 'evidence-based medicine.' While psychologists were the first to conduct meta-analyses and have produced several thousand meta-analyses over the past 40 years, medical researchers have now conducted nearly five times as many meta-analytic studies. In the initial years, medical research focused on many of the same issues as psychologists: psychotherapy, depression, and mental disorders more broadly. After the initial 10 years, however, medical research diverged strongly. Meta-analysis was most heavily used in this period for examining the causes and treatments for breast cancer, high blood pressure, and coronary disease (including heart attacks). These foci continued to dominate meta-analyses in the medical field in the 1990s, though research into other forms of cancer (e.g., lung cancer) also became quite frequent. The most recent 10 years, however, has witnessed a dramatic shift in the uses to which meta-analysis has been put in medicine. Researchers are now heavily focused on the genetic aspect of disease, including studies of genetic predisposition and genetic polymorphism. Research into cancer and the other major diseases of the Western world continues to be a major focus as well.

Among the social sciences, with the exception of psychology, meta-analysis remains a much more novel approach. Though early examples of meta-analyses can be found in almost any social science field, its adoption has been episodic and relatively slow. Within the social sciences, psychologists, demographers, and criminologists are the most frequent users of the method. Demographers have most often used meta-analysis techniques to study the social correlates of health (such work is most often published in epidemiology journals). This work differs from that of the medicine field in which the topics studied do not readily translate into treatment recommendations. Instead, demographers have focused most heavily on determining which subpopulations are most at risk. Criminologists have been most interested in using meta-analysis for the study of the risk factors leading to criminal behavior, including environmental (e.g., social upbringing), behavioral (e.g., drug use), and psychological determinants of criminal careers (including recidivism).

Economists, sociologists, education researchers (excluding educational psychology), and anthropologists employ meta-analysis to a small degree, while meta-analytic studies are almost unheard of in communications, geography, history, or political science. It is difficult to discern trends in the uses to which meta-analytic methods have been put within these social science fields. First, the number of meta-analyses within each discipline is small and therefore does not contain dominant

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03223-2

clusters of research topics. Second, the diversity of research topics is much greater than in disciplines such as psychology and medicine. In the social sciences, it is actually quite rare to see two or more meta-analyses on closely related topics.

In the balance of this article, I will review the methodological trajectory of meta-analysis and provide resources related to conducting this type of study (and the related technique known as meta-regression). I will begin by providing an expanded definition and description of meta-analysis. I will then turn to a description of several key methodological issues and subsequent analytical refinements that have driven changes in the way meta-analysis is done. Finally, I will argue that the future direction of meta-analysis lies with meta-regression.

The Logic Behind Meta-Analysis

Meta-analysis, in its most basic form, is analogous to calculating an overall mean from a series of two or more subgroup means. This method for calculating the overall mean is advantageous as it does not require access to the original, individual-level data from which the subgroup means were calculated. Instead, as the example equation below shows, one can derive an overall mean ($\bar{x}_{\text{overall}}$) knowing only the subgroup means ($\bar{x}_i$) and their respective sample sizes (n_i):

$$\bar{x}_{\text{overall}} = \frac{\sum(n_i\bar{x}_i)}{\sum n_i}$$

The data required for a meta-analysis consist of *effect estimates* and their associated standard errors, both of which are obtained from a comprehensive sample of available research studies. Effect estimates can take the form of means, regression coefficients from any type of generalized linear model (e.g., partial slopes from ordinary least-squares regression, log odds from logistic regression, logged rate ratios from Poisson regression, log hazard ratios from proportional hazards regression, etc.), and other similar statistics. Generally, the type of effect estimate chosen follows a normal distribution under the Central Limit Theorem, which excludes using meta-analysis for Chi-square statistics or F statistics. Correlations can be subjected to meta-analysis so long as they are transformed using Fisher's Z transformation. The standard errors of the effect estimates are used to calculate *inverse variance weights* (w_i), where $w_i = 1/se^2$. Regardless of which type of effect estimate statistic one will use, meta-analysis requires that the same statistic be available for all the studies one wishes to combine and that this statistic is based on the same variable (in the case where the effect estimate is a mean) or pair of independent and dependent variables (in the case where the effect estimate is a regression coefficient).

The basic results from the analysis of meta-analytic data are a weighted average effect estimate and a standard error for this weighted average. The weighted average effect estimate ($\overline{ES}$) is calculated using the individual effect estimates (ES_i) and their corresponding inverse variance weights (w_i) using the equation below.

$$\overline{ES} = \frac{\sum(w_i ES_i)}{\sum w_i}$$

The standard error of the mean effect estimate equals the inverse of the sum of the weights ($se_{\overline{ES}} = 1/\sum w_i$). One can use this standard error to calculate the corresponding confidence interval and derive a *p*-value from a hypothesis test for whether the weighted average effect estimate equals zero.

Meta-Analysis and Systematic Literature Reviews

Meta-analysis is a quantitative form of *systematic literature review* as it is based on an attempt to identify and analyze the population of research studies examining a specific variable or association between a specific independent–dependent variable pair. Though every meta-analysis begins with a systematic literature search, the idea of a specific systematic review methodology lagged the development of the meta-analysis statistical procedure itself. Initially, researchers conducting a meta-analysis would likely have given little consideration to the literature review component because, as experts in their respective subfields, they felt aware of the relevant research from the outset.

Subsequent studies have shown there are inherent difficulties in assembling a comprehensive, representative sample of a particular body of research, much less in locating an entire population of studies. The most thorough reviews are based on a multifaceted, multistage literature search. Expertise in one's chosen subfield remains a critical component, as does competency in searching scholarly databases and consulting with other experts in one's field. However, major obstacles inherent in the literature itself affect the search process. First, *publication bias* exists when relevant studies are not made publicly available (e.g., proprietary studies by private companies, which are not published; the tendency to delay or forego publication when one's research that does not find a statistically significant effect for a key variable – the *file drawer effect*). Second, *search bias* exists due to the inconsistencies between researchers, journals, and disciplines in the way research is indexed and reported. Researchers remain responsible for identifying the keywords that will be associated with a particular study; so simple differences in the vocabulary used can hamper the search effort. Furthermore, since one's keywords reflect only the focal concepts in a study, it is difficult for others to determine all of the variables present (the focal variable in a meta-analysis may only be a control variable in some primary studies). Systematic literature reviews are therefore, in practice, based upon samples of relevant studies rather than the population. The question is not whether one has missed a relevant study, but whether one has systematically missed relevant studies.

The first graphical and statistical methods to help detect the presence and magnitude of search and publication bias were developed by Egger and Davey-Smith (1998). The first tool developed for this purpose, known as a *funnel plot* (see Figure 1), is a graph of the values of the effect estimates (*x*-axis) against some measure of the weight the effect estimate had in the meta-analysis (*y*-axis; this is most often either the sample size or the inverse variance weight). If no search or publication bias is present the graph will be symmetric and funnel-shaped, with very little variation in the values of the effect estimates for studies with large weights and greater variation in the effect

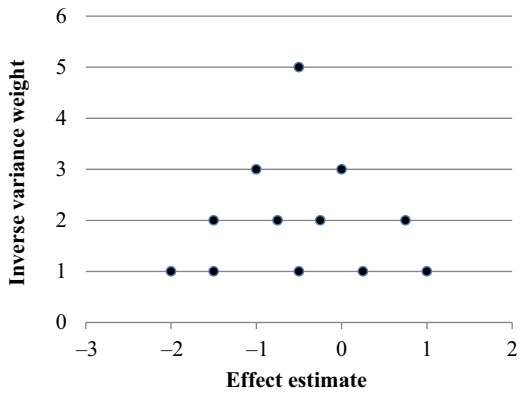

Figure 1 Example funnel plot.

estimates for studies with small weights. A lack of left–right symmetry in the funnel plot indicates that a certain class of studies has been undersampled. For example, a funnel plot that is skewed left indicates that small studies with relatively high effect estimate values have been missed.

The statistical test for publication and search bias consists of a weighted ordinary least-squares regression, with the inverse variance weight used in the meta-analysis serving as the regression weight, the effect estimate divided by its standard error serving as the dependent variable, and the inverse of the standard error of the effect estimate serving as the sole independent variable. The intercept/constant from the resulting model will be significantly different from the null value (zero) if publication and/or search bias are present in the data.

The Egger and Davey-Smith method of detecting funnel-plot asymmetry works best when the effect estimates come from relatively homogeneous study samples. When this is not the case, alternative tests may perform better. Of particular importance when using event-history (survival analysis) data is the test proposed by Peters et al. (2006), which is also based on a weighted least-squares regression. The weight for the regression is calculated using information on the numbers in the sample experiencing the event of interest (denoted as n_{event}) and the number not experiencing the event (denoted as $n_{\text{no event}}$) using the following equation:

$$\text{regression weight} = \frac{1}{(1/n_{\text{event}}) + (1/n_{\text{no event}})}$$

The dependent variable in the regression is the log of the effect estimate (e.g., log odds, log hazard), with the inverse of the sample size being used as the sole predictor variable. The regression coefficient for the inverse of the sample size will be significantly different from the null value (zero) if publication and/or search bias are present in the data.

If significant funnel-plot asymmetry is found, the researcher will need to document the limitation in their report but may also attempt to remedy the asymmetry using the 'trim and fill' method. The trim and fill method essentially involves two stages. First, the researcher identifies and removes effect estimates from the analysis that 'mirror' those that were absent due to publication and/or selection bias. The reduced set of effect estimates (exhibiting symmetry via the Egger test) is analyzed and the results compared to those from the full set of effect estimates. Second, the full set of effect estimates is augmented by the addition of effect estimates thought to closely resemble those missing due to bias. This augmented set of effect estimates (exhibiting symmetry via the Egger test) is also analyzed and the results compared to those from the original full set of effect estimates. Through these comparisons, the robustness of the set of effect estimates can be gauged, with the results from the analyses of both the reduced and augmented sets providing some insight into how the results of the analysis may differ if publication and/or selection bias were not present.

Analyzing Data Heterogeneity

Of all the aspects of meta-analysis, by far the greatest methodological effort has been spent working on the problem of unaccounted heterogeneity in one's data. Multiple techniques have been developed for identifying the presence of data heterogeneity and for quantifying its magnitude.

The most basic tool for assessing between-study heterogeneity in any type of systematic literature review is the pairing of a qualitative assessment of patterns among the relevant studies with a graphical assessment of the magnitude, direction, and level of statistical significance for the effect estimates from the studies. The graphical portion of this assessment, known as a *forest plot*, shows all of the effect estimates and corresponding confidence intervals against a line, which represents the null/no effect value (see Figure 2). A researcher systematically examines the characteristics of the samples on which the relevant studies were based and attempts to qualitatively evaluate whether there is any relationship between study attributes and the results reported. For example, if a researcher notices that most of the effect estimates that are on the left side of the forest plot come from samples of men, while most of the effect estimates that are on the right side of the forest plot come from samples of women, they might conclude that gender heterogeneity is impacting the results of the analysis (and that gender is likely an important moderating variable).

Another basic tool, which is used solely to determine the presence/absence of data heterogeneity among a set of effect estimates, is *Cochrane's Q-test*. To calculate the Q-test statistic, one calculates a weighted sum (where w_i again denotes the inverse variance weight associated with a particular effect

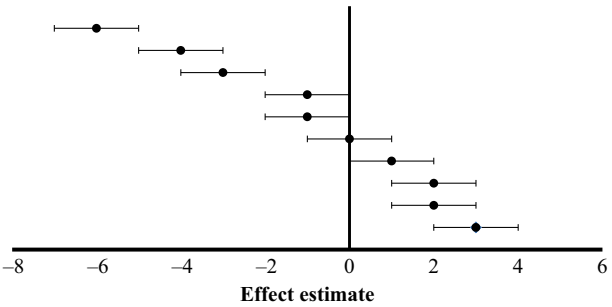

Figure 2 Example forest plot.

estimate) of the squared deviations between each effect estimate (again denoted as ES_i) and the overall mean effect estimate from the meta-analysis (again denoted as $\overline{ES}$) using the following formula:

$$Q = \frac{\sum\left[w_i\left(ES_i - \overline{ES}\right)^2\right]}{\sum w_i}$$

The Q-test statistic follows a χ^2 distribution with $k-1$ degrees of freedom, where k refers to the number of effect estimates being examined. The null hypothesis is that variation between effect estimates is random; a rejection of the null hypothesis indicates that a significant degree of nonrandom heterogeneity exists.

A third tool, called an I^2 *test*, was recently developed by Higgins and Thompson (2002) as a way to assess the magnitude of data heterogeneity when present, rather than simply its presence or absence. The I^2 statistic is calculated solely using information on the value of Cochrane's Q-test statistic and the number of effect estimates included in the analysis (k) using the following formula:

$$I^2 = \frac{Q-k-1}{Q} \cdot 100\%$$

In cases where Q is less than $k-1$, the value of I^2 is set to zero. This statistic indicates the percentage of the variation between effect estimates that is due to heterogeneity alone (as opposed to occurring by chance).

The final commonly used tool for diagnosing the presence of heterogeneity in the data is the *random effect variance component* (v statistic). Hand calculation of this particular statistic is difficult, and is best handled using statistical software. The null hypothesis, again, is that the data are homogeneous. This claim is tested using a basic t-test, where the test statistic is the v statistic divided by its standard error.

Together, these measures of data heterogeneity guide the researcher's decisions on whether to use fixed-effects or random-effects analyses, whether to conduct subgroup analyses, and whether to conduct a meta-regression. A fixed-effects meta-analysis has an assumption that the data are homogeneous (i.e., any variability present is random) while a random-effects meta-analysis assumes that each effect estimate may be drawn from a different population (and thus that both random variation and true variation between populations might exist). Thus, a random-effects model is likely more appropriate when data heterogeneity is present, regardless of whether or not the researcher has any ideas of the source of the heterogeneity.

If the researcher has some idea of the confounding, mediating, and/or moderating factors that might distinguish between study samples and therefore be the source of data heterogeneity, they can pursue strategies to measure the effects of these factors on the overall meta-analysis results. The first method for doing so involves stratifying the full set of effect sizes into subgroups based upon the values of a third variable. For example, researchers commonly conduct separate meta-analyses for men and women. Other basic demographic factors such as age, socioeconomic status, and race/ethnicity are also commonly used to define subgroup meta-analyses. The second method for examining the sources of data heterogeneity is meta-regression, which I will discuss later in this article.

Standardizing Meta-Analysis Reporting

As meta-analysis has progressed as a technique, a series of efforts have been made to elucidate best practices, including the work of the *Cochrane Collaboration* and the creation of the reporting guidelines known as QUOROM (The Quality of Reporting of Meta-analyses) and PRISMA (Preferred Reporting Items of Systematic reviews and Meta-Analyses). While the focus of the present article leans as much as possible toward the social and behavioral scientist, it must be acknowledged that the bulk of this work was undertaken by scholars in the medical disciplines.

The first effort toward standardizing meta-analytic practice emerged in 1993 with the formation of the *Cochrane Collaboration* (www.cochrane.org). The broad goal of the new organization was to promote evidence-based medicine – the idea that an overt effort was needed in order to systematically identify and promote the 'best practices' that current medical research could provide. A major part of this effort involved the increased use of meta-analyses of small-scale clinical trials to more rapidly identify medical treatment advances that should be studied using large-scale trials. One major obstacle was the lack of a centralized place where practitioners and researchers could access meta-analyses and other types of systematic reviews. The primary contributions of the *Cochrane Collaboration* were the creation of a reviews database and the active promotion of systematic review methods in medical research.

A second major obstacle to using meta-analysis to advance evidence-based medicine was the inconsistencies in the quality of meta-analytic practice and reporting. An initial attempt to address this issue was begun in 1996 by a group of 30 clinical epidemiologists, clinicians, journal editors, and medical researchers. This group met with the purpose of systematically developing a set of meta-analysis reporting recommendations, and the resulting recommendations became known as the QUOROM statement, which was issued in its final form in 1999. An updated set of guidelines known as the PRISMA statement (see http://www.prisma-statement.org/index.htm) was issued in 2009.

The PRISMA statement suggests 27 items that every systematic review should contain. These include ensuring the title clearly identifies the paper as a systematic review, meta-analysis, and/or meta-regression and ensuring the abstract contains information on the study design. The statement also suggests that explicit hypotheses and study rationale be provided, and that the methods section clearly documents how the literature search was conducted, how the study inclusion–exclusion determination was made, how the data were collected and coded, how study quality was assessed, and how data heterogeneity was investigated. The statement suggests that the presentation of results should include a flow diagram showing how many studies were considered (and subsequently either retained or eliminated) at each stage of the literature search process. The statement also suggests that meta-analysts report key descriptive information (e.g., study size, sample demographics, individual study results) for the studies included in the analysis (and directly identify these studies through citations). The use of a forest plot (described earlier) is specifically recommended as part of the reporting guidelines. Finally, the PRISMA statement suggests that meta-analysis results should, at a minimum, provide the mean effect estimate, a confidence interval for this estimate, and report the

results of Cochrane's Q-test and/or the I^2 statistic. Recommendations are also made with respect to the discussion section and with regard to the identification of outside funding sources.

Extensions of Meta-Analysis

Though still in its emergent phase, there is a trend toward expanding the number of effect estimates included in an analysis in order to examine data heterogeneity. This trend stands in contrast to the existing norm of eliminating as much heterogeneity as possible before proceeding with one's analysis. In this final section of this article, I will describe the methodological work that is presently underway that allows for including multiple effect estimates from a single study and I will describe meta-regression methods (which I earlier argued were a key tool for quantifying the effects of confounding, mediating, and/or moderating factors).

One major assumption underlying meta-analysis is that the effect estimates included in the study are independent. Ideally, this means that only one effect estimate is taken from each study and that each study included in the meta-analysis is based on a unique sample of people. It is common, however, for studies to present results from subgroup analyses or otherwise stratify their results in ways that make the selection of a single effect estimate difficult. Previously, if no overall effect estimate was directly reported in a study, the meta-analyst was encouraged to use the meta-analysis procedure on the set of effect estimates within this type of study in order to produce the desired overall effect estimate.

Focusing only on single, independent effect estimates from eligible studies substantially reduces the researcher's ability to investigate data heterogeneity. In fact, within-study variations between men and women, age groups, and the like form a powerful basis for examining heterogeneity. An argument can therefore be made that the meta-analyst should seek subgroup effect estimates even if an overall effect estimate is available. Including multiple effect estimates from a single study (or from multiple studies reporting results based on the same data source) introduces a violation of effect estimate independence that must be accounted for. Hedges et al. (2010) discuss this in detail and suggest a method for calculating robust variance estimates for this situation.

Meta-regression becomes the analytic tool of choice when the researcher increases the number of effects estimates included in the analysis (by obtaining multiple effect estimates from individual studies or simply by seeking to expand the number of studies included in the analysis). Meta-regression is also, as previously argued here, the tool of choice when the researcher's focus is the analysis of data heterogeneity rather than solely the calculation of a single meta-analytic estimate of the mean effect.

Meta-regression is an examination of the effect of one or more covariates on the magnitude of the effect estimates reported in the studies included in the analysis. Effectively, meta-regression uses a weighted least-squares model, using the inverse of the variance as the weighting variable (i.e., the same weighting variable used in meta-analysis). The dependent variable is the effect estimate itself, though the effect estimate must be in a form that follows a normal distribution. This means that if one's effect estimates are ratio measures of risk or hazard (i.e., from logistic, Poisson, negative binomial, proportional hazards regression, etc.), one must use the log of the relative risk (i.e., log-odds, log of the rate ratio, log hazard, etc.) as the dependent variable. Either a fixed-effects (fixed partial slopes, single fixed intercept) or a random-effects (fixed partial slopes, random intercept) meta-regression can be specified.

The coefficients produced through a meta-regression indicate the amount by which the magnitude of the effect estimate changes for every one-unit increase in a particular covariate. If the effect estimate takes the form of a log-odds, log-rate, or log-hazard, the researcher can calculate an exponentiated coefficient, which represents a multiplicative indicator of the effect of a one-unit increase in the covariate on the magnitude of the effect estimate. The interpretation of exponentiated coefficients must be done with care however, as they often closely resemble the output from a meta-analysis. For example, suppose you are examining a set of odds ratios from a set of studies focusing on a binary outcome. Suppose further that some of the studies examined samples of men, while others examined samples of women. To prepare the data for meta-regression, one would need to create a binary covariate for the sex of the sample and calculate the log of each odds ratio for use as the dependent variable. Suppose the coefficient for the sex covariate was 0.28. If the covariate were coded such that a value of 0 indicated a study sample consisting of men and a value of 1 indicated a study sample consisting of women, this would indicate that the mean log odds for the outcome of interest was 0.28 higher for women than for men. Exponentiating the coefficient (i.e., $e^{0.28}$) yields 1.32, which indicates that the odds of the outcome of interest for women is 1.32 times the odds of the outcome of interest for men. The 1.32 describes the relative magnitude of the odds ratios for women and men, but does not indicate anything about what the odds of the outcome itself is.

The superiority of meta-regression to subgroup meta-analyses for the analysis of data heterogeneity is relatively obvious. With the subgroup meta-analysis approach, one can only examine one intervening covariate at a time. With meta-regression, one can control for and examine the effects of multiple potential covariates. The difference in the analytic power of the two approaches is akin to the difference between a bivariate regression and a multivariate regression.

Conclusion

In sum, many refinements have been introduced to the quantitative form of systematic review known as a meta-analysis. A tool that was originally proposed to magnify the statistical power of small, relatively homogeneous experimental studies is now routinely used to guide research synthesis across the social and medical sciences. Standards for conducting and reporting the results of meta-analyses have emerged. Of the many advances made, the most attention has been paid to the issue of data heterogeneity. Early work on this issue focused on ways to identify the presence and magnitude of heterogeneity in order to help the researcher eliminate its effects. More recent work has sought to leverage heterogeneity in order to develop

better understandings of the pathways through which social (and other) factors lead to any number of outcomes of interest.

See also: Systematic Reviewing and Meta-Analysis.

Bibliography

Egger, M., Davey-Smith, G., 1998. Meta-analysis: bias in location and selection of studies. British Medical Journal 316 (7124), 61–66.

Glass, Gene V., 1976. Primary, secondary and meta-analysis of research. Educational Research 5, 3–8.

Higgins, Julian P.T., Thompson, Simon G., 2002. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine 21 (11), 1539–1558.

Hedges, Larry V., Tipton, E., Johnson, Matthew C., 2010. Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods 1, 39–65.

Peters, Jaime L., Sutton, Alex J., Jones, David R., Abrams, Keith R., Rushton, L., 2006. Comparison of two methods to detect publication bias in meta-analysis. Journal of the American Medical Association 295 (6), 676–680.

Meta-Analysis: Theory

Larry V Hedges and Arend M Kuyper, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Meta-analysis is the combination of estimates across studies for the purposes of drawing general conclusions. Meta-analysis usually summarizes the results of each study by an index of effect size and estimates of effect sizes are combined across studies. Effect sizes are often combined by computing a weighted average to estimate the mean effect size. In that case, assessments of the possible heterogeneity of effects are important. Variation of effects across studies can be studied using linear models (so-called metaregression).

A basic tenet of scientific work is that research findings must be replicable and that studies making important findings are subjected to replication. Often studies are replicated several times, with major or minor variations in research design, research procedures, or measurement methods. Some form of research synthesis or systematic reviewing is essential to determine whether results from different studies are consistent, and to summarize the evidence across studies. Meta-analysis is the use of statistical procedures in that endeavor; it differs from secondary analysis in that the information from each study comes from statistical summaries of the data in the original studies and not directly from the raw data (Glass, 1976). Meta-analysis is a fundamental scientific activity that is practiced in all of the empirical sciences and the fundamental statistical principles underlying meta-analysis are essentially the same in all applications of meta-analysis (Draper et al., 1993).

In meta-analysis, the results of individual studies are represented by a measure of effect size. Study provides an estimate of the effect-size parameter in that study and a measure of statistical uncertainty (variance) of that estimate. The meta-analysis then summarized these effect-size estimates across studies by estimating the distribution of effect-size parameters (often by an estimate of the mean and variance of the distribution of effect size-parameters). The analysis may model the variation in results across studies by estimating the relation between study level characteristics of studies and effect size. This article provides an introduction to the analytic tools most widely used in meta-analysis.

Effect Sizes

A major problem in research synthesis is that studies differ in the details of design and data collection. Effect-size measures are used to represent the outcome of each study so that the results of different studies are comparable, even though there may be differences in the ways the studies collect or analyze data. The appropriate effect size depends on the design of the studies and the nature of the outcome variable. In experimental studies that contrast treatments and employ continuous outcome variables, the standardized mean difference (the difference between the treatment and control group means divided by the within-group standard deviation) is often used (see Borenstein, 2009). In experimental studies with dichotomous outcomes, the effect sizes are typically odds ratios or risk ratios (see Fleiss and Berlin, 2009). When both the independent and dependent variables are continuous, correlation coefficients or standardized regression coefficients are appropriate effect sizes (see Becker, 2009). Each study in a meta-analysis provides both an estimate of effect size and a standard error that indicates the sampling uncertainty of the estimate. In most cases, effect-size estimates are approximately normally distributed; transformations (such as the log of odds ratios and risk ratios or the Fisher z-transform of the correlation coefficient) are often used to improve the normal approximation.

Estimating the Mean and Variance of the Effects across Studies

Estimation of the average effect size across studies involves combining the individual effect-size estimates from different studies, usually by averaging. Because some studies (i.e., those with larger sample sizes) produce estimates with lower uncertainty than others, it is common practice to weight the effect-size estimates when computing the average. A standard result is that weighting proportionally to the inverse of the variance (squared standard error) of each effect-size estimate produces the weighted mean with the smallest variance, so inverse variance weights are usually used. Weighting is used to increase precision of the combined estimate; it has no effect on bias.

Inference Models

There are two sorts of questions that meta-analysis can address. The analyst may inquire about the parameters underlying the particular set of studies observed in the meta-analysis, or there may be more general interest in a putative population of studies that could exist. These questions require different inference models, designated conditional and unconditional (Hedges and Vevea, 1998).

The conditional inference model involves generalizations to the particular parameters in the set of studies that are actually observed in the meta-analysis (or in studies that are, a priori, equivalent to them). For example, conditional inference about the mean effect size in a set of studies addresses the mean of the effect-size parameters in these particular studies; the effect

sizes of any other studies are conceptually irrelevant. The only uncertainty in the conditional inference model arises because the sample effect-size estimate in each study is not identical to the effect-size parameter that it estimates; rather, there is random variation associated with the sampling of persons (or other units) in the studies. Conditional inference may be appropriate, for example, when a particular set of studies has been commissioned to answer a particular question (e.g., a regulatory question), and these studies provide the entire evidentiary base on which a policy question is to be determined.

In contrast, the unconditional inference model involves inferences about the parameters of a universe or hyperpopulation of studies from which the observed studies are a sample. The hypothetical other studies in this universe are conceptually relevant. Because the observed studies are regarded as a sample from the universe, there is additional uncertainty arising from the process of sampling studies. Thus, the effect-size estimate of a study involves a component of uncertainty associated with the sampling of units into that study, and an additional component associated with the sampling of the study itself. The unconditional inference model can also be seen as a Bayesian model where the hyperpopulation of studies is conceived as generating a prior distribution of study effect-size parameters (*see* Bayesian Statistics). The unconditional model may be appropriate when combining evidence about a treatment that is implemented somewhat differently in every study, giving rise to a distribution of treatment effects corresponding to variation in implementation.

Inference models are distinguished from the statistical procedures described below because the type of inference intended is logically prior and should determine the statistical procedures that are used. Fixed-effects statistical procedures are appropriate for conditional inferences and random-effects or Bayesian procedures are appropriate for unconditional inferences.

Fixed-Effects Statistical Procedures

Fixed-effects statistical procedures are designed to make conditional inferences. Let $\theta_1,\ldots,\theta_k$ be the effect-size parameters from k independent studies; let $T_1,\ldots,T_k$ be the corresponding estimates observed in the studies, and let $v_1,\ldots,v_k$ be the variances (squared standard errors) of those estimates. These v_i are called the estimation error variances because they reflect the variance of the estimation errors (the $\varepsilon_i = T_i - \theta_i$).

We typically assume that the T_i is independent and each T_i is approximately normally distributed given θ_i. Note that the normality assumption may not be realistic in all situations. Where it is not, other procedures are needed that are adapted to the particular situation. If the effect-size parameters are equal or nearly so (that is if $\theta_1 = \cdots = \theta_k = \theta$), then the fixed-effects estimate of θ is the weighted mean $\overline{T}_{\bullet}$, given by

$$\overline{T}_{\bullet} = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i} \qquad [1]$$

where the weight $w_i = 1/v_i$. The variance of $\overline{T}_{\bullet}$ is the reciprocal of the sum of the weights

$$\sigma_{\bullet}^2 = \frac{1}{\sum_{i=1}^{k} w_i} \qquad [2]$$

and the standard error $SE(\overline{T}_{\bullet})$ of $\overline{T}_{\bullet}$ is the square root of $\sigma_{\bullet}^2$; that is, $SE(\overline{T}_{\bullet}) = \sigma_{\bullet}$. Because $T_1,\ldots,T_k$ are normally distributed, it follows that the linear combination $\overline{T}_{\bullet}$ is also normally distributed.

If $T_1,\ldots,T_k$ all estimate the same underlying effect size, that is if $\theta_1 = \cdots = \theta_k = \theta$, then $\overline{T}_{\bullet}$ estimates θ, and a $100(1-\alpha)$-percent confidence interval for θ is given by

$$\overline{T}_{\bullet} - C_{\alpha/2}\sigma_{\bullet} \le \theta \le \overline{T} + C_{\alpha/2}\sigma_{\bullet} \qquad [3]$$

where $C_{\alpha/2}$ is the two-tailed critical value of the standard normal distribution (e.g., $C_{\alpha/2} = 1.96$ for $\alpha = 0.05$) and $\sigma_{\bullet}^2$ is the sampling variance of $\overline{T}_{\bullet}$ (eqn [2]). If the confidence interval does not include zero, we reject the hypothesis that $\theta = 0$ at the 100α-percent significance level (*see* Estimation: Point and Interval; Hypothesis Testing in Statistics).

To assess the heterogeneity of effect parameters in fixed-effects models, a test of the hypothesis that $\theta_1 = \cdots = \theta_k$ is often employed. The test uses the statistic

$$Q = \sum_{i=1}^{k} w_i \left(T_i - \overline{T}_{\bullet}\right)^2 = \sum_{i=1}^{k} w_i T_i^2 - \frac{\left(\sum_{i=1}^{k} w_i T_i\right)^2}{\sum_{i=1}^{k} w_i} \qquad [4]$$

which has a chi-square distribution with $k-1$ degrees of freedom when the effect-size parameters are the same in all studies. Large values of Q lead to rejection of the hypothesis that the effect sizes are homogeneous.

Random-Effects Statistical Procedures

Random-effects statistical procedures are designed to make unconditional inferences. That is, they allow inferences about effect-size parameters in a universe of studies from which the observed studies are a sample. Use the same notation for effect-size parameters, their estimates, and the sampling error variances of the estimates. Let $\mu_{\bullet}$ be the mean of the population from which the effect-size parameters $\theta_1,\ldots,\theta_k$ were sampled; let τ^2 be the variance of that mean. The model for the effect-size estimates is

$$T_i = \theta_i + \varepsilon_i = \mu_{\bullet} + \xi_i + \varepsilon_i \qquad [5]$$

where $\xi_i = \theta_i - \mu_{\bullet}$ is a study-specific random effect and $\varepsilon_i = T_i - \theta_i$ is the estimation error as in the fixed-effects model. Thus τ^2 is the variance of θ_i and represents a component of variation due to between-study differences in effect parameters.

The most frequently used procedure estimates $\mu_{\bullet}$ by a weighted mean of the effect-size estimates using slightly different weights from those employed in the fixed-effects procedure. Here, the weights include the additional component of variation, τ^2, which describes between-study variation in effect-size parameters. The unconditional variance v_i^* of T_i is

$$v_i^* = v_i + \tau^2 \qquad [6]$$

where the asterisk denotes that this is the unconditional or random-effects variance. The weights used to compute the weighted mean are $w_i^* = 1/v_i^*$. Note that an estimate of τ^2

will typically be used to estimate the weights in practice (see The section 'Estimating τ^2'). The random-effects

$$\overline{T}^*_{\bullet} = \frac{\sum_{i=1}^{k} w_i^* T_i}{\sum_{i=1}^{k} w_i^*} \quad [7]$$

Note that $\overline{T}^*_{\bullet}$ depends on $\hat{\tau}^2$ because w_i^* does. The variance of $\overline{T}^*_{\bullet}$ is the reciprocal of the sum of the random-effects weights; that is

$$\sigma^{*2}_{\bullet} = \frac{1}{\sum_{i=1}^{k} w_i^*} \quad [8]$$

The standard error $SE(\overline{T}^*_{\bullet})$ of the mean effect estimate $\overline{T}^*_{\bullet}$ is the square root of its sampling variance, $SE(\overline{T}^*_{\bullet}) = \sigma^*_{\bullet}$. If both the effect-size estimates and the effect-size parameters are normally distributed, then $\overline{T}^*_{\bullet}$ is also normally distributed.

An approximate $100(1-\alpha)$-percent confidence interval for the mean effect $\mu_{\bullet}$ is given by

$$\overline{T}^*_{\bullet} - C_{\alpha/2}\sigma^*_{\bullet} \le \mu_{\bullet} \le \overline{T}^*_{\bullet} + C_{\alpha/2}\sigma^*_{\bullet}, \quad [9]$$

where $C_{\alpha/2}$ is the two-tailed critical value of the standard normal distribution and $\sigma^{*2}_{\bullet}$ is the variance of $\overline{T}^*_{\bullet}$ given in eqn [8]. If the confidence interval does not include zero, one rejects the hypothesis that $\mu_{\bullet} = 0$ at the 100α-percent significance level.

Estimating τ^2

Unlike the v_i, which can usually be computed using analytic results (formulas that involve sample size and the effect-size estimate), τ^2 must be computed empirically. Distribution free estimates are usually computed using the methods of moments. The most frequently used estimate of τ^2 in most cases is the distribution free estimate based on the Q-statistic used to test for heterogeneity. The estimate is

$$\hat{\tau}^2 = \frac{Q-(k-1)}{S_1 - S_2/S_1} \quad [10]$$

where S_a is given by

$$S_a = \sum_{i=1}^{k} w_i^a \quad [11]$$

and Q is given in eqn [4]. When eqn [10] produces a negative result, the estimate is set to zero.

The variance of $\hat{\tau}^2$ is given by

$$\frac{2(k-1)S_1 + 4\left(S_1^2 - S_2\right)\tau^2 + 2\left(S_1S_2 - 2S_3 - S_2^2/S_1\right)\tau^4}{S_1^2 - S_2} \quad [12]$$

which reduces to $2(v+\tau^2)^2/(k-1)$ when $v_1 = \cdots = v_k = v$.

If the between-study random effects can be assumed to be normally distributed, then a slightly more efficient estimate of τ^2 can be obtained by maximum likelihood (ML) or restricted ML (REML) estimation. Both of these estimators can be obtained by a simple iterative procedure (see, e.g., Raudenbush, 2009).

The between-studies variance component τ^2 is the natural characterization of between-study heterogeneity in random-effects models. The test for heterogeneity using the Q statistic described in connection with fixed-effects models is also a test of the hypothesis that $\tau^2 = 0$.

Evaluating Heterogeneity

The variance component τ^2 is a direct expression of the variation of the effect-size parameters (the θ_i's). However, to describe the impact of heterogeneity on conclusion of meta-analyses, indices that are functions of τ^2/v (where v is a function (a summary like the average) of the within-study variances) are more useful (see Higgins and Thompson, 2002 for a discussion of this point).

One index is H^2, the ratio of Q to its degrees of freedom (called the Birge ratio in physics (after Birge, 1932)), that is

$$H^2 = Q/(k-1) \quad [13]$$

H^2 has expected value 1 when $\tau^2 = 0$ and there is no heterogeneity. If $v_1 = \cdots = v_k = v$, the expected value of H^2 is

$$E\{H^2\} = (v+\tau^2)/v$$

so H^2 can be interpreted as the ratio of between-study variation $(v+\tau^2)$ to within-study variation (v).

A more useful statistic for many purposes is the ratio of between-study to total variation in the estimates, a kind of intraclass correlation called I^2 (see Higgins and Thompson, 2002). Conceptually I^2 is defined as

$$I^2 = \tau^2/(v+\tau^2)$$

where v is some 'average' value of the v_i. The problem in computing I^2 is choosing a value of v when the within-study variances are not identical. Choosing v to be a weighted average

$$v = \frac{(k-1)\sum_{i=1}^{k} w_i}{\left(\sum_{i=1}^{k} w_i\right)^2 - \sum_{i=1}^{k} w_i^2} = (k-1)/(S_1 - S_2/S_1)$$

where $(S_1 - S_2/S_1)$ is the denominator in eqn [10], yields the conventional definition of I^2, namely

$$I^2 = \frac{Q-(k-1)}{Q} \quad [14]$$

The statistic I^2 is often expressed as a percentage, reflecting the percentage of total variation in the effect-size estimates that is due to between-study heterogeneity.

Bayesian Statistical Procedures

The random-effects procedure above gives an estimate of the average effect size that depends on $\hat{\tau}^2$, which is estimated from the data. When the number of studies is small, the uncertainty of this estimate can be rather large, and a variety of values of τ^2 are almost equally consistent with the data. Bayesian procedures for estimating the average effect size essentially compute the random-effects estimate of the mean conditional on τ^2, and average these conditional means, giving weight to each that is proportional to the posterior probability of τ^2 given the data. Bayesian procedures can also provide improved estimates of individual study effect sizes using information from the entire collection of effect sizes, and have some logical advantages for unconditional inference (for more information see DuMouchel, 1990 or Hedges, 1998).

Robust Standard Errors and Confidence Intervals

Another approach to inference about effect sizes is based on robust standard errors computed from the empirical distribution of the effect-size estimates when the number of studies is large (see Hedges et al., 2010). This approach is similar to that used in econometrics (see, e.g., White, 1980). It is particularly appealing because it makes no assumptions about the (conditional or unconditional) distribution of the effect-size estimates. Moreover, it can be used even when variance estimates for individual effect-size estimates are not available (as occurs when reporting of statistics in studies is very incomplete). While this approach provides no method for computing weights to increase efficiency, standard random-effects procedures can be used to compute (efficient) weights if the estimation error variances are known.

The robust variance estimate of the weighted mean is

$$v_{\bullet}^{R} = \frac{k\sum_{i=1}^{k}\left(w_i^*\right)^2\left(T_i - \overline{T}_{\bullet}^*\right)^2}{(k-1)\left(\sum_{j=1}^{k} w_i^*\right)^2} \qquad [15]$$

which can be used for a robust confidence intervals by substituting $\sqrt{v_{\bullet}^{R}}$ into (9) in place of $\sigma_{\bullet}^*$ and using the critical value of Student's t-distribution with $k-1$ degrees of freedom in place of standard normal critical value. If the robust confidence interval does not include μ_0, one rejects the hypothesis that $\mu_{\bullet} = \mu_0$ at the 100α-percent significance level.

It is important to remember that the validity of inferences based on the robust variance estimate given in eqn [15] depends on having a large sample of studies. However, simulation studies carried out by Hedges et al. (2010) suggest that confidence intervals based on this procedure may be reasonably accurate when there are as few as 10–20 studies if they are (at least approximately) equally weighted.

Statistical Power of Tests in Meta-Analysis

The statistical power of tests in meta-analysis can be quite high when the number of studies is large, but can be quite modest if the number of studies is small and if the variance of the effect-size estimates is large (which typically happens if the sample sizes are small). This is particularly true of tests for heterogeneity of effect sizes across studies. Tests of heterogeneity may often fail to reject the hypothesis of heterogeneity even when effects do actually vary due to low statistical power. Methods for computing the power of statistical tests are beyond the scope of this article but are described in Hedges and Pigott (2001).

Modeling Between-Study Variation in Effect Sizes

The object of the meta-analysis is often to study the variation in effect sizes between studies. This often takes the form of estimating the relation between effect sizes and study-level covariates that may represent methodologically or substantively important characteristics of the studies (such as treatment intensity or dosage, treatment type, study design, subject population). Such modeling must be considered correlational, since there is no random assignment of studies to have different characteristics. Consequently, any such patterns of relations between study-level covariates and effect size should be interpreted with care. It is not unusual for several different covariates to explain essentially the same pattern of variation, which makes interpretations ambiguous.

As in the case of estimating the mean effect across studies, there are two inference models: conditional (which allows inferences about the parameters of the particular studies observed) and unconditional (which allows inferences about the parameters of studies from which the studies observed are considered a sample). In either inference model, the inferences are usually considered conditional on the values of the study-level covariates observed (*see* Linear Hypothesis; Linear Hypothesis: Regression (Basics)).

Fixed-Effects Models

In describing the modeling of between-study variation, it is helpful to use matrix notation. Denote the k-dimensional vectors of population and sample effect sizes by $\boldsymbol{\theta} = (\theta_1,\ldots, \theta_k)'$ and $\mathbf{T} = (T_1,\ldots, T_k)'$, respectively. Suppose that the effect-size parameter θ_i for the ith study depends on a vector of $p \le k$ fixed study-level covariates $\mathbf{x}_i = (x_{i1},\ldots, x_{ip})'$, such as treatment dosage, or length of follow-up. Specifically, assume that

$$\theta_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p, \quad i = 1,\ldots,k \qquad [16]$$

where $\beta_1,\ldots, \beta_p$ are unknown regression coefficients.

The $k \times p$ matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_k' \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kp} \end{pmatrix} \qquad [17]$$

is called the design matrix in regression analysis, and is assumed to have no linearly dependent columns; that is, $\mathbf{X}$ has rank p. The first predictor value is often set to 1 (i.e., $x_{i1} = 1$ for all i) so that the first column of $\mathbf{X}$ is a column of ones and β_1 becomes an intercept. The set of equations can be written succinctly in matrix notation as

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} \qquad [18]$$

where $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)'$ is the p-dimensional vector of regression coefficients.

The model for $\mathbf{T}$ can be written as

$$\mathbf{T} = \boldsymbol{\theta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad [19]$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_k)' = \mathbf{T} - \boldsymbol{\theta}$ is a k-dimensional vector of residuals. The linear model $\mathbf{T} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ for the effect sizes is analogous to the model that is the basis for ordinary least squares regression analysis. Because $\boldsymbol{\varepsilon} = \mathbf{T} - \boldsymbol{\theta}$, the distribution of $\boldsymbol{\varepsilon}$ is approximately k-variate normal with means zero and diagonal covariance matrix $\mathbf{V}$ given by

$$\mathbf{V} = \mathrm{Diag}(v_1, v_2, \ldots, v_k) \qquad [20]$$

Because the elements of $\mathbf{V}$ are independent but not identically distributed, we use the method of weighted least squares to obtain an estimate of $\boldsymbol{\beta}$.

The generalized least squares estimator is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{T} \qquad [21]$$

which has a normal distribution with mean $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}$ given by

$$\mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\Sigma} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \qquad [22]$$

The distribution of $\widehat{\boldsymbol{\beta}}$ can be used to obtain tests of significance or confidence intervals for components of $\boldsymbol{\beta}$. If σ_{jj} is the jth diagonal element of $\boldsymbol{\Sigma}$, then a $100(1-\alpha)$-percent confidence interval for β_j, $1 \le j \le p$, is given by

$$\widehat{\beta}_j - C_{\alpha/2}\sqrt{\sigma_{jj}} \le \beta_j \le \widehat{\beta}_j + C_{\alpha/2}\sqrt{\sigma_{jj}} \qquad [23]$$

where $C_{\alpha/2}$ is the 100α-percent two-tailed critical value of the standard normal distribution.

The error sum of squares statistic

$$Q_E = \mathbf{T}'[\mathbf{V}^{-1} - \mathbf{X}\mathbf{V}^{-1}(\mathbf{X}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{T} \qquad (24)$$

can be used to test the goodness of fit of the regression model. When the model is correctly specified (that is, when $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$), Q_E has a chi-square distribution with $k-p$ degrees of freedom, so that large values of Q_E indicate that fit is poor.

Mixed-Effects Models

As in the fixed-effects model, we assume that each T_i is normally distributed about θ_i with (conditional) variance v_i. However, in the case of the mixed model, the predictor variables are not assumed to explain all of the variation in the θ_i. Rather, there is an additional component of variance τ^2 that is not explained by the p study-level covariates $\mathbf{X}_1,\ldots,\mathbf{X}_p$. The linear model is

$$\theta_i = \beta_1 x_{i1} + \ldots + \beta_{ip} + \xi_i \qquad [25]$$

where $x_{i1},\ldots, x_{ip}$ are the values of the predictor variables $X_1,\ldots, X_p$ for the ith study, $\beta_1, \ldots, \beta_p$ are unknown regression coefficients, and ξ_i is a random effect with variance τ^2.

The $k \times p$ design matrix $\mathbf{X}$, and the k-dimensional vectors of population and sample effect sizes $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_k)'$ and $\mathbf{T} = (T_1,\ldots,T_k)'$, are the same as in the fixed-effects model. The set of equations can be written in matrix notation as

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \xi \qquad [26]$$

where $\boldsymbol{\beta} = (\beta_1,\ldots, \beta_k)'$ is the p-dimensional vector of regression coefficients, and $\boldsymbol{\xi} = (\xi_1,\ldots, \xi_k)'$ is the k-dimensional vector of study-specific random effects. The model for the observations $\mathbf{T}$ can be written as

$$\mathbf{T} = \boldsymbol{\theta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} \qquad (27)$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_k)' = \mathbf{T} - \boldsymbol{\theta}$ is a k-dimensional vector of residuals of $\mathbf{T}$ about $\boldsymbol{\theta}$, and $\boldsymbol{\eta} = \boldsymbol{\xi} + \boldsymbol{\varepsilon}$ is a k-dimensional vector of residuals of $\mathbf{T}$ about $\mathbf{X}\boldsymbol{\beta}$.

The distribution of $\boldsymbol{\eta} = \boldsymbol{\xi} + \boldsymbol{\varepsilon}$ has mean zero and a diagonal covariance matrix given by $\mathrm{Diag}(v_1 + \tau^2, v_2 + \tau^2,\ldots, v_k + \tau^2)$. Using the estimate $\widehat{\tau}^2$ in place of τ^2, one may compute the $k \times k$ diagonal matrix $\mathbf{V}^*$, an estimate of the unconditional covariance matrix of $\mathbf{T}$, by

$$\mathbf{V}^* = \mathrm{Diag}\left(v_1 + \widehat{\tau}^2, v_2 + \widehat{\tau}^2,\ldots, v_k + \widehat{\tau}^2\right) \qquad [28]$$

The generalized least squares estimator of $\boldsymbol{\beta}$ under the model using the estimated covariance matrix $\mathbf{V}^*$ is

$$\boldsymbol{\beta}^* = (\mathbf{X}'\mathbf{V}^{*-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{*-1}\mathbf{T} \qquad [29]$$

which is approximately normally distributed with mean $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}^*$ given by

$$\Sigma^* = (\mathbf{X}'\mathbf{V}^{*-1}\mathbf{X})^{-1} \qquad [30]$$

The distribution of $\boldsymbol{\beta}^*$ can be used to obtain tests of significance or confidence intervals for components of $\boldsymbol{\beta}$. If β^*_{jj} is the jth diagonal element of $\boldsymbol{\Sigma}^*$, then an approximate $100(1-\alpha)$-percent confidence interval for β_j, $1 \le j \le p$, is given by

$$\beta^*_j - C_{\alpha/2}\sqrt{\sigma^*_{jj}} \le \beta_j \le \beta^*_j + C_{\alpha/2}\sqrt{\sigma^*_{jj}} \qquad (31)$$

where $C_{\alpha/2}$ is the 100α-percent two-tailed critical value of the standard normal distribution.

Relation to Hierarchical Linear Models

The mixed models considered here are related to the hierarchical linear model, a special case of the general mixed linear model, which finds wide application in the social sciences (*see* Hierarchical Models: Random and Fixed Effects; also Goldstein, 1987; Raudenbush and Bryk, 2002). There is widely available software to estimate and test the statistical significance of parameters in these models, so this software is sometimes used in meta-analysis. In these models, the data are regarded as hierarchically structured, and a model is defined for each level of the hierarchy. Statistical models for meta-analysis are usually two level models, where level I is that of the study and the model for level I is

$$T_i = \theta_i + \varepsilon_i, i = 1, \ldots, k \qquad [32]$$

where ε_i is the estimation error of T_i as an estimate of θ_i. Level II of the model describes between-study variation in the study-specific effect-size parameters (the θ_i). The linear model in eqn [26] above implies a level II model like

$$\theta_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \xi_i \qquad [33]$$

where ξ_i is a study-specific random effect. Most of the attention in estimating these models has focused on the case where both the ε_i and the ξ_i are normally distributed with zero mean and unknown variances, that is $\varepsilon_i \sim \mathrm{N}(0, v_i)$ and $\xi_i \sim \mathrm{N}(0, \tau^2)$, $i = 1, \ldots, k$.

There are two important differences between the hierarchical linear models usually studied and the model used in meta-analysis. The first is that, in meta-analytic models, the variances of the estimation errors $\varepsilon_1,\ldots,\varepsilon_k$ are not identical across studies. The sampling error variances depend on various aspects of study design (particularly sample size) that are rarely constant across studies. The second is that the estimation error variances in meta-analysis are assumed to be known. Therefore, the model used in meta-analysis is a special case of the general hierarchical linear model where the level I variances are unequal, but known. Software for the analysis of hierarchical linear models can be used for mixed model meta-analysis, if it permits the specification of first-level variances that are unequal but known, as do to the programs HLM (Raudenbush et al., 2000) and SAS PROC MIXED (SAS 6.12, 1996).

Methods for Bayesian analyses of hierarchical mixed linear models are also available (see Seltzer et al., 1996 for a general treatment or DuMouchel, 1990 or Hedges, 1998 for applications to meta-analysis).

Estimating τ^2

The usual distribution free estimator of the residual variance component is given by

$$\widehat{\tau}^2 = (Q_E - k + p)/c \qquad [34]$$

where Q_E is the residual sum of squares from the fixed-effects weighted regression (eqn [24]). If the quantity computed in eqn [34] is negative, the estimate is truncated to zero. The normalizing constant c is given by

$$c = \mathrm{tr}(\mathbf{V}^{-1}) - \mathrm{tr}[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-2}\mathbf{X}] \qquad [35]$$

where $\mathbf{V} = \mathrm{Diag}(v_1,\ldots,v_k)$ is a $k \times k$ diagonal matrix of conditional variances and $\mathrm{tr}(\mathbf{A})$ is the trace of the matrix $\mathbf{A}$.

If the random effects can be assumed to be normally distributed, then ML estimation or REML estimation provide more efficient estimates of τ^2 than the method of moments estimator. While neither method provides unbiased estimates, those based on REML tend to be slightly less biased than the MLE.

The vector $\mathbf{e}_{[s+1]} = (e_{1[s+1]},\ldots,e_{k[s+1]})'$ of residuals on the $(s+1)$st step given by

$$e_{[s+1]} = T - \mathbf{X}\widehat{\boldsymbol{\beta}}^*_{[s+1]} \qquad [39]$$

is used to calculate $\widehat{\tau}^2_{[s+1]}$, the generalized least squares estimate of τ^2 given $\widehat{\boldsymbol{\beta}}^*_{[s+1]}$ and the data.

In the case of ML estimation, compute $\widehat{\tau}^2_{[s+1]}$ the value of τ^2 at the next iteration via

$$\widehat{\tau}^2_{[s+1]} = \frac{\sum_{i=1}^k \left(w^*_{i[s]}\right)^2 \left(e^2_{i[s+1]} - v_i\right)}{\sum_{i=1}^k \left(w^*_{i[s]}\right)^2} \qquad [40]$$

setting any negative estimates of τ^2 to zero. Proceed until the estimates of τ^2 converge (that is, change negligibly with each further iteration). The large sample variance of the ML estimate of τ^2 is given by

$$2 \Big/ \sum_{i=1}^k \left(v_i + \tau^2\right)^{-2} \qquad [41]$$

which is approximated by using the value of $\widehat{\tau}^2_{[s]}$ at the last iteration for τ^2.

In the case of REML estimation, the only difference is that instead of computing the estimate of τ^2 at the $(s+1)$st step $\widehat{\tau}^2_{[s+1]}$ via eqn. [40], compute $\widehat{\tau}^2_{[s+1]}$ via

$$\widehat{\tau}^2_{[s+1]} = \frac{\sum_{i=1}^k \left(w^*_{i[s]}\right)^2 \left(e^2_{i[s+1]} - v_i\right) + \mathrm{tr}\left[\left(\sum_{i=1}^k w^*_{i[s+1]}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^k \left(w^*_{i[s+1]}\right)^2 \mathbf{x}_i\mathbf{x}_i'\right)\right]}{\sum_{i=1}^k \left(w^*_{i[s+1]}\right)^2} \qquad [42]$$

One method for obtaining ML or REML estimates of τ^2 relies on the fact that if either $\boldsymbol{\beta}$ or τ^2 were known, it would be easy to obtain the least squares estimate of the other. By starting with an initial value of τ^2 and estimating first $\boldsymbol{\beta}$ and then reestimating τ^2, a sequence of estimates can be obtained that converge to the ML or REML estimators of $\boldsymbol{\beta}$ and τ^2 (see Goldstein, 1986).

To compute either estimate, start with an initial estimate $\tau^2{}_{[0]}$ of τ^2 (such as that given in eqn [34]). Then if $\widehat{\tau}^2_{[s]}$ is the estimate of τ^2 on the sth step, the $k \times k$ weight matrix on that step becomes

$$\mathbf{W}_{[s]} = \mathrm{Diag}\left[w_{1[s]}{}^*, \ldots, w_{k[s]}{}^*\right], \qquad [36]$$

where

$$w^*_{i[s]} = 1 \Big/ \left(v_i + \widehat{\tau}^2_{[s]}\right) \qquad [37]$$

The estimate $\widehat{\boldsymbol{\beta}}^*_{[s+1]}$ of $\boldsymbol{\beta}$ on the $(s+1)$st step is given by

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^*_{[s+1]} &= \left(\mathbf{X}'\mathbf{W}_{[s]}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}_{[s]}\mathbf{T} \\ &= \left(\sum_{i=1}^k w^*_{i[s]}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^k w^*_{i[s]}\mathbf{x}_i T_i\right)\end{aligned} \qquad [38]$$

The difference between eqn. [40] and eqn. [42] is the second term of the numerator, which adds a small constant to the estimate of τ^2 at each iteration. In either case, the iteration proceeds until the estimates of τ^2 converge (that is, change negligibly with each further iteration) and τ^2 is approximated by the value of $\widehat{\tau}^2_{[s]}$ at the last iteration.

Note that the ML estimate of $\boldsymbol{\beta}$ is just the estimate given in eqn [29] using the ML estimate of τ^2 to compute the weights. Confidence intervals from the distribution of $\widehat{\boldsymbol{\beta}}^*$ can be constructed from the ML estimator using eqn [31].

Metaregression analyses are often used to test for the effects of study-level moderator variables. The power of statistical tests for moderator effects need not always be large, even when the number of studies is large. Understanding the power of statistical tests for moderators may be particularly important when these analyses fail to detect moderator effects. Procedures for computing statistical power in metaregression are discussed in Hedges and Pigott (2004).

Robust Variance Estimation

Another approach to inference about effect sizes is based on variance estimates computed from the empirical distribution of the effect-size estimates. This approach is particularly

appealing for three reasons. First, it makes no assumptions about the (conditional or unconditional) distribution of the effect-size estimates, so it is robust to violations of assumptions that estimates or random effects have any specific (e.g., normal) distributions. Second, this theorem does not require that the study-level covariates be fixed as in other metaregression models. Third, robust variance computations can be used even when variance estimates for individual effect-size estimates are not available (as occurs when reporting of statistics in studies is very incomplete).

There are two disadvantages of this method, however. One is that it provides no method for computing weights to increase efficiency. We recommend that standard random-effects procedures be used to compute (efficient) weights, and then robust variance estimates can be used in conjunction with these weights. A second disadvantage is that it requires a large number of studies. However, simulations carried out by Hedges et al. (2010) suggest that the method performs reasonably well for a single covariate even when the number of studies is as small as 10 to 20.

Suppose that $\widehat{\boldsymbol{\beta}}^*$ is a mixed model estimate of $\boldsymbol{\beta}$ computed using weights w_i^*, $i = 1, \ldots, k$. Then the robust variance estimate $\mathbf{V}^R$ is given by

$$\mathbf{V}^R = \left(\frac{k}{k-p}\right)\left(\sum_{i-1}^{k} w_i^* \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \times \left(\sum_{i=1}^{k} (w_i^*)^2 e_i^2 \mathbf{x}_i \mathbf{x}_i'\right)\left(\sum_{i-1}^{k} w_i^* \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \quad [43]$$

where $e_i = T_i - \mathbf{x}_i \widehat{\boldsymbol{\beta}}^*$ is the residual in the ith study and p is the total number of coefficients in the regression model including the intercept if there is one. The diagonal elements of $\mathbf{V}^R$ can then be used as the variances to construct confidence intervals or significance test for the β_j using eqn [31].

Multivariate Meta-Analysis

We have explicitly assumed that effect sizes are independent in previous sections of this article. The assumption of independence will typically be met if indeed the estimators are derived from different experiments, or if the estimates are derived from independent samples of subjects within the same study. However, not all collections of effect-size estimates can easily be reduced to collections of independent estimators. Statistical dependence arises through the random part of the statistical model. Decomposing the effect-size estimate into $T_i = \theta_i + \varepsilon_i$, we see that dependence can arise via correlations among the estimation errors (the ε_i's) or through dependencies among the effect-size parameters (the θ_i's) or both.

Dependencies among the estimation errors can occur, if a study reports data on several different measures of the same conceptual outcome construct or report the same measure at several different points in time. One may wish to compute an effect-size estimate for each measure, but these effect sizes will usually be correlated because the measures are collected from the same individuals. Alternatively, a study may have several treatment arms that are compared to the same control condition. One may wish to estimate an effect size for each treatment, but the comparison with a common control makes such estimates stochastically dependent.

There are three general approaches to situations in which effect sizes are not independent. One is to discard effect-size estimates for all but one outcome or treatment in each study, so that each study produces a single estimate. This approach leads to independent effect sizes but may be excessively wasteful. A second approach is to combine the correlated effect sizes to create a single synthetic effect size in each study. This approach can preserve the information from all the effects, but may be difficult to implement. The third approach is to analyze the correlated effect-size estimates using a multivariate approach that recognizes their correlation structure. This approach also utilizes all of the information in the effect sizes but can likewise be difficult to implement.

Dependencies among effect-size parameters may occur, for example, if the same investigators carry out several studies that are included in the same meta-analysis and those studies share common methods, samples, or measurements that induce effects to be more alike than those from studies carried out by different investigators.

Combining Correlated Effect Sizes Within Studies

It is sometimes convenient to combine effect-size estimates across studies in two steps. First, the correlated estimators within each study are combined to produce a single synthetic effect-size estimate for each study. This eliminates the correlation of estimation errors within a study as a source of dependence. Then the independent estimators are, in turn, pooled across studies.

Suppose that there are still k studies, but now each study yields a vector $\mathbf{T}_i = (T_{i1}, \ldots, T_{iq})'$, a vector of estimates from a single study with $q \times q$ covariance matrix $\boldsymbol{\Sigma}_i$. It is not necessary for each study to have the same number of estimates, so that q may differ for each study and should properly be denoted q_i to denote the dependence of the study index i, but we drop the subscript i on q in this section for simplicity. If the estimates $T_{i1}, \ldots, T_{iq}$ are reasonably consistent with the model of the same effect size for each component, that is, if

$$\theta_{i1} = \theta_{i2} = \cdots = \theta_{iq} = \theta_{i-} \quad [44]$$

then the investigator may wish to compute an estimate of the common effect size θ_{i-}. The most precise estimate of θ_{i-} is given by

$$\overline{T}_{i-} = \sum_{j=1}^{q} \omega_{ij} T_{ij} \Big/ \sum_{j=1}^{q} \omega_{ij} = \boldsymbol{\omega}_i' \mathbf{T}_i \quad [45]$$

where $\boldsymbol{\omega}_i$ is a $q \times 1$ weight vector given by

$$\boldsymbol{\omega}_i = (w_{i1}, \ldots, w_{ip})' = \boldsymbol{\Sigma}_i^{-1} \mathbf{1} \quad [46]$$

$\mathbf{1}$ is a $q \times 1$ column vector of ones. The weights $\boldsymbol{\omega}_i$ are estimated by using estimates of $\boldsymbol{\Sigma}_i$. In large samples, $\overline{T}_{i-}$ is approximately normally distributed with mean θ_{i-} and variance

$$v_{i-} = 1 \Big/ \sum_{j=1}^{q} \omega_{ij} = 1/1'\boldsymbol{\Sigma}_i^{-1}\mathbf{1}$$

As in the case of univariate meta-analysis, the weights leading to the most efficient weighted mean are proportional to the inverse variance (that is, the inverse of the covariance matrix) and the variance of the weighted mean is the reciprocal of the sum of the weights.

If some of the k studies yield a vector of $q > 1$ effect sizes, eqn [45] can be used to obtain a single estimate of effect size $\overline{T}_{i-}$ for each of those studies. Then the large sample approximation can be used to obtain the variance of $\overline{T}_{i-}$. The result is a series of k independent estimators of effect size, one from each study. These independent effect-size estimators can then be combined using the methods given in the previous sections.

Whenever estimates of effect sizes are being combined it may be advisable to determine whether they seem to be estimating the same quantity. To test, for the ith study, whether the component estimators share a common effect size, that is, whether

$$\theta_{i1} = \theta_{i2} = \cdots = \theta_{iq} = \theta_{i-} \qquad [47]$$

compute the statistic

$$Q_i = \mathbf{T}_i'\mathbf{M}_i\mathbf{T}_i \qquad [48]$$

where $\mathbf{M}_i$ is the matrix

$$\mathbf{M}_i = \boldsymbol{\Sigma}_i^{-1} - \left(\boldsymbol{\Sigma}_i^{-1}\mathbf{1}\mathbf{1}'\boldsymbol{\Sigma}_i^{-1}/\mathbf{1}'\boldsymbol{\Sigma}_i^{-1}\mathbf{1}\right) \qquad [49]$$

When eqn [47] holds, the statistic Q_i has a chi-square distribution with $q-1$ degrees of freedom. Thus the test that $\theta_{i1} = \cdots = \theta_{iq}$ consists in comparing the statistic Q_i with the critical values of the chi-square distribution with $q-1$ degrees of freedom. If the value of Q_i is small or statistically nonsignificant, the investigator might choose to pool the estimates from all components of $\mathbf{T}_i$ to obtain the most precise estimate of θ_i-. The Q_i statistic is analogous to the Q statistic for testing heterogeneity of effects in univariate meta-analysis and reduces to it if the effect-size estimates are uncorrelated.

The practical problem in using the methods described above is that the covariance matrix $\boldsymbol{\Sigma}_i$ of the correlated effect sizes is typically not known precisely. Analytic theory exists for the elements of $\boldsymbol{\Sigma}_i$ for various effect sizes, but the covariance matrix of the effect sizes typically depends on the covariance matrix of the outcome variables on which the effect sizes are based (see Gleser and Olkin, 2009). Because the covariance matrix of the outcomes is usually not known precisely (and even an estimate of the covariance matrix is frequently not reported), it is difficult to compute the covariance matrix of the (statistically dependent) effect-size measures.

A very crude approach in this situation is simply to take an unweighted average of the estimates and to approximate the variance of the average by the average of the variances. Such an approach does make use of all of the estimates, but substantially overestimates the variance. A somewhat less crude approach is to compute the variances analytically (the diagonal elements of $\boldsymbol{\Sigma}_i$) and impute the covariances by imputing (guessing) correlations among the effect sizes, so that $\sigma_{ab} = \rho_{ab}\sqrt{v_{ia}v_{ib}}$, where ρ_{iab} is the imputed correlation, via the variance of T_{ia}, and v_{ib}, the variance of T_{ib}.

Multivariate Mixed Models

When a general linear model analysis of effect-size data is desired, there is nonnegligible covariation among effect-size parameters even after controlling for the factors that are of interest in the analysis. That is, there is greater residual covariation than would be expected from estimation error alone. The mixed model incorporates a component of between-study covariation into the uncertainty of effect-size parameters and their estimates, which has the effect of increasing residual variation.

Assume that each $\mathbf{T}_i$ has a q_i-variate normal distribution about the corresponding $\boldsymbol{\theta}_i$ with known $q_i \times q_i$ covariance matrix $\boldsymbol{\Sigma}_i$, that is.

$$\mathbf{T}_i \sim N(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i), i = 1, \ldots, k \qquad [50]$$

Note that we do not require all of the studies to have the same number of effect sizes. We denote the $Q = q_1 + \cdots + q_k$ dimensional column vectors of population and sample effect sizes by $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots \boldsymbol{\theta}_k')'$ and $\mathbf{T} = (\mathbf{T}_1', \ldots \mathbf{T}_k')'$, respectively.

In the present model, assume that each element of the effect size parameter vector $\boldsymbol{\theta}_i$ for the ith experiment depends on a vector of p fixed concomitant variables.

$$\mathbf{X}_i = \begin{pmatrix} x_{i11} & \cdots & x_{i1p} \\ \vdots & \ddots & \vdots \\ x_{iq_i1} & \cdots & x_{iq_ip} \end{pmatrix}$$

Specifically, assume that

$$\theta_{is} = \beta_1 x_{is1} + \cdots + \beta_q x_{isp}, + \xi_{is} \quad i = 1, \ldots, k; \quad s = 1, \ldots, q_i \qquad [51]$$

where $\beta_1, \ldots, \beta_p$ are unknown regression coefficients and ξ_{is} is a random effect. Partition the vector $\xi = (\xi_1, \ldots, \xi_Q)'$ into k subvectors $\xi_1, \ldots, \xi_k$ each with dimension q_i corresponding to the effects for a given study (that is corresponding to $\mathbf{T}_i$ and ε_i). Then the linear model can be written as

$$\mathbf{T} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} + \boldsymbol{\varepsilon} = \begin{pmatrix} \mathbf{T}_1 \\ \vdots \\ \mathbf{T}_k \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_k \end{pmatrix}\boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\xi}_1 \\ \vdots \\ \boldsymbol{\xi}_k \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_k \end{pmatrix}$$

where ε_i and ξ_i are normally distributed with covariance matrices $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Xi}$, respectively, and are independent of one another.

Estimation of the Regression Coefficients and the Covariance Components

If the within-study covariance matrices of the effect sizes (the $\boldsymbol{\Sigma}_i$) are known, the regression coefficient vector $\boldsymbol{\beta}$ and the covariance component matrix $\boldsymbol{\Xi}$ can be estimated by weighted least squares. The usual procedure is to first estimate the covariance component matrix $\boldsymbol{\Xi}$ and then reweight to estimate the regression coefficient vector $\boldsymbol{\beta}$ and its covariance matrix $\boldsymbol{\Sigma}$ (see, e.g., Jackson et al., 2011). There are usually advantages (among them software availability) in considering the problem as a special case of the hierarchical linear model considered in conjunction with univariate mixed model analyses. The multivariate mixed model analyses can be carried out as instances of the multivariate hierarchical linear model (see Thum, 1997), estimating parameters by the method of ML. However, a simpler alternative is available.

Since the sampling error covariance matrix is known, it is possible to transform the within-study model so that the sampling errors are independent (Kalaian and Raudenbush, 1996). For each study perform the Cholesky factorization of each sampling error covariance matrix $\boldsymbol{\Sigma}_i$ so that

$$\sum\nolimits_i = \mathbf{F}_i\mathbf{F}_i' \qquad [52]$$

where $\mathbf{F}_i$ is a known matrix (since $\mathbf{\Sigma}_i$ is a known matrix). Then transform the within-study model to be

$$\mathbf{F}_i^{-1}\mathbf{T}_i = \mathbf{F}_i^{-1}\boldsymbol{\theta}_i + \mathbf{F}_i^{-1}\boldsymbol{\varepsilon}_i \quad [53]$$

The transformed effect-size vector $\mathbf{Z}_i$ given by

$$\mathbf{Z}_i \equiv \mathbf{F}_i^{-1}\mathbf{T}_i$$

has a sampling error vector

$$\tilde{\boldsymbol{\varepsilon}}_i = \mathbf{F}_i^{-1}\varepsilon_i \quad [54]$$

which has covariance matrix $\mathbf{I}$, a $q_i \times q_i$ identity matrix. Thus one might write the within-study (level I) model as

$$\mathbf{Z}_i = \mathbf{F}_i^{-1}\boldsymbol{\theta}_i + \tilde{\varepsilon}_i \quad [55]$$

where the transformed effect-size estimates $\mathbf{Z}_i$ are independent, but the effect-size parameter vector $\boldsymbol{\theta}_i$ is the same as in the original model.

Thus the within-study model (eqn [53]) along with the between-study (level II) model

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} \quad [56]$$

is a conventional two-level hierarchical linear model with independent sampling errors at the first level, with identically distributed level I errors that have unit variance. Therefore conventional software can be used to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Xi}$ by the method of full or REML (as in HLM or SAS Proc Mixed) or iteratively reweighted least squares (as in ML3) can be used to estimate this model provided that the level I variances are constrained to unity.

The Robust Variance Estimator

If the covariance matrices of the effect sizes (the $\mathbf{\Sigma}_{i\cdot}$) are not known, the methods discussed above cannot be used. Moreover, when the number of studies is large; it can be tedious to apply those methods. A much simpler method in practice is a procedure that involves robust estimation of the variance of coefficient estimates in metaregression. This procedure does not require knowledge of the $\mathbf{\Sigma}_i$. For now, assume an arbitrary weight matrix $\mathbf{W} = \mathrm{diag}(\mathbf{W}_1,\ldots,\mathbf{W}_k)$, where $\mathbf{W}_j = \mathrm{diag}(w_{j1}, \ldots, w_{jkj})$ is the mixed model weight matrix for the jth study. The weighted least squares estimate of $\boldsymbol{\beta}$ is

$$\mathbf{b} = \left(\sum_{j=1}^{k}\mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1}\left(\sum_{j=1}^{k}\mathbf{X}_j'\mathbf{W}_j\mathbf{T}_j\right) \quad [57]$$

The exact variance of $\mathbf{b}$ is

$$\left(\sum_{j=1}^{k}\mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1}\left(\sum_{j=1}^{k}\mathbf{X}_j'\mathbf{W}_j\mathbf{\Sigma}_j\mathbf{W}_j\mathbf{X}_j\right)\left(\sum_{j=1}^{k}\mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1}$$

Robust standard errors are obtained by substituting the matrix of cross products of within-study residuals in the jth study for $\mathbf{\Sigma}_j$, that is

$$\mathbf{V}^R = \left(\frac{k}{k-q}\right)\left(\sum_{j=1}^{k}\mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1} \times \left(\sum_{j=1}^{k}\mathbf{X}_j'\mathbf{W}_j\mathbf{e}_j\mathbf{e}_j'\mathbf{W}_j\mathbf{X}_j\right)\left(\sum_{j=1}^{k}\mathbf{X}_j'\mathbf{W}_j\mathbf{X}_j\right)^{-1} \quad [58]$$

where $\mathbf{e}_j = \mathbf{T}_j - \mathbf{X}_j\mathbf{b}$ is the $(q_j \times 1)$ residual vector in the jth study. Although $\mathbf{e}_j\mathbf{e}_j'$ is a rather poor estimate of $\mathbf{\Sigma}_j$, it is good enough that the robust variance estimator converges in probability to the correct value as the number of studies becomes large (see Hedges et al., 2010). In practice, the robust variance estimator is often adequate when the number of studies is 20–30, but diagnostics and methods of improving the performance of the robust variance estimator in small samples are available (Tipton, in press).

Note that robust variance estimates can be computed using any weights (e.g., equal weights or weights proportional to the sample size in each study), but some weighting schemes lead to more efficient estimators of $\boldsymbol{\beta}$. Weights that are the same for each element of $\mathbf{T}_i$ and are proportional to the sample size of the study are generally close to optimal, if there is not large unexplained between-study heterogeneity in effects. A more detailed discussion of weighting schemes is given in Hedges et al. (2010).

The square roots of the diagonal elements of $\mathbf{V}^R$ are the robust standard errors of the elements of $\mathbf{b}$. The robust test of the hypothesis

$$H_0\colon \beta_j = 0$$

uses the test statistic

$$t_j^R = b_j \Big/ \sqrt{v_{jj}^R} \quad [59]$$

where v_{jj}^R is the jth diagonal element of $\mathbf{V}^R$, which is compared with critical values of Student's t-distribution with $k - p$ degrees of freedom. Note that the degrees of freedom is the number of clusters (not effect-size estimates) minus the number of coefficients (including the intercept) in the regression model. Thus with two predictors, $p = 3$ (the intercept and two slopes). A robust $100(1 - \alpha)$-per cent confidence interval for β_j is given by

$$b_j - C_{\alpha/2}\sqrt{v_{jj}^R} \le \beta_j \le b_j + C_{\alpha/2}\sqrt{v_{jj}^R} \quad [60]$$

where $C_{\alpha/2}$ is the $100(1 - \alpha)$-percent point of t-distribution with $m - p$ degrees of freedom.

Conclusions and Suggestions for Further Reading

There are two basic inference models in meta-analysis (conditional and unconditional) and two classes of statistical procedures used in combining effect-size estimates (fixed- and mixed-effects procedures). Bayesian procedures are closely related to mixed models. In either case, statistical methods that weight estimates by the inverse of the conditional or unconditional variance are used for analysis. Typical analyses involve computing an overall mean effect size, assessing heterogeneity, and modeling variation in effect sizes across studies using metaregression.

For more information on the general process of integrating research, and related statistical tools, the reader may find useful Borenstein et al. (2009), Cooper et al. (2009), and Lipsey and Wilson (2000).

See also: Bayesian Statistics; Estimation: Point and Interval; Hierarchical Models: Random and Fixed Effects; Hypothesis Testing in Statistics; Linear Hypothesis: Regression (Basics); Linear Hypothesis.

Bibliography

Becker, B.J., 2009. Model based meta-analysis. In: Cooper, H., Hedges, L.V., Valentine, J.C. (Eds.), The Handbook of Research Synthesis and Meta-Analysis, second ed. Russell Sage Foundation, New York, pp. 377–395.

Borenstein, M., 2009. Effect sizes for continuous data. In: Cooper, H., Hedges, L.V., Valentine, J.C. (Eds.), The Handbook of Research Synthesis and Meta-Analysis, second ed. Russell Sage Foundation, New York, pp. 221–235.

Borenstein, M., Hedges, L.V., Higgins, J., Rothstein, H.R., 2009. Fixed-Effect Versus Random-Effects Models. Introduction to Meta-analysis 77–86.

Birge, R.T., 1932. The calculation of errors by the method of least squares. Physical Review 40 (2), 207.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, second ed. Erlbaum, Hillsdale, NJ.

Cooper, H.M., Hedges, L.V., Valentine, J., 2009. The Handbook of Research Synthesis and Meta-Analysis, second ed. The Russell Sage Foundation, New York.

Draper, D., Gaver, D.P., Goel, P.K., Greenhouse, J.B., Hedges, L.V., Morris, C.N., Tucker, J.R., Waternaux, C., 1993. Combining Information: Statistical Issues and Opportunities for Research. American Statistical Association, Washington, DC.

DuMouchel, W.H., 1990. Bayesian meta-analysis. In: Berry, D. (Ed.), Statistical Methods for Pharmacology. Marcel Dekker, New York, pp. 509–529.

Fleiss, J.L., Berlin, J.A., 2009. Effect sizes for dichotomous data. In: Cooper, H., Hedges, L.V., Valentine, J.C. (Eds.), The Handbook of Research Synthesis and Meta-Analysis, second ed. Russell Sage Foundation, New York, pp. 237–253.

Glass, G.V., 1976. Primary, secondary, and meta-analysis of research. Educational researcher 3–8.

Gleser, L.J., Olkin, I., 2009. Stochastically dependent effect sizes. In: Cooper, H., Hedges, L.V., Valentine, J.C. (Eds.), The Handbook of Research Synthesis and Meta-Analysis, second ed. Russell Sage Foundation, New York.

Goldstein, H., 1986. Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika 73 (1), 43–56.

Goldstein, H., 1987. Multi-Level Models in Educational and Social Research. Oxford University Press, London.

Hedges, L.V., 1998. Bayesian approaches to meta-analysis. In: Everitt, B., Dunn, G. (Eds.), Recent Advances in the Statistical Analysis of Medical Data. Edward Arnold, London, pp. 251–275.

Hedges, L.V., Pigott, T.D., 2001. The power of statistical tests in meta-analysis. Psychological Methods 6, 203–217.

Hedges, L.V., Pigott, T.D., 2004. The power of statistical tests for moderators in meta-analysis. Psychological Methods 9, 426–445.

Hedges, L.V., Tipton, E., Johnson, M., 2010. Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods 1, 39–65. Erratum in 1 (2), 164–165.

Hedges, L.V., Vevea, J.L., 1998. Fixed- and random-effects models in meta-analysis. Psychological Methods 3, 486–504.

Higgins, J.P.T., Thompson, S.G., 2002. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine 21 (11), 1539–1558.

Jackson, D., Riley, R., White, I.R., 2011. Multivariate meta-analysis: potential and promise. Statistics in Medicine 30, 2481–2498.

Kalaian, H., Raudenbush, S.W., 1996. A multivariate mixed linear model for meta-analysis. Psychological Methods 1, 227–235.

Lipsey, M.W., Wilson, D.B., 2000. Practical Meta-Analysis. Sage Publications, Thousand Oaks, CA.

Raudenbush, S.W., 2009. Analyzing effect sizes: random effects models. In: Cooper, H., Hedges, L.V., Valentine, J.C. (Eds.), The Handbook of Research Synthesis and Meta-Analysis, second ed. Russell Sage Foundation, New York, pp. 295–315.

Raudenbush, S.W., Bryk, A.S., 2002. Hierarchical Linear Models. Sage Publications, Thousand Oaks, CA.

Raudenbush, S.W., Meng-Li, Yang, Yosef, M., 2000. Maximum likelihood for generalized linear models with nested random effects via high – order, multivariate Laplace approximation. Journal of Computational and Graphical Statistics 9 (1), 141–157.

Seltzer, M.H., Wong, W.H., Bryk, A.S., 1996. Bayesian analysis in applications of hierarchical models: issues and methods. Journal of Educational and Behavioral Statistics 21, 131–167.

Thum, Y.M., 1997. Hierarchical linear models for multivariate outcomes. Journal of Educational and Behavioral Statistics 22, 77–108.

Tipton, E. Small sample adjustments for robust variance estimation with meta-regression. Psychological Methods, in press.

White, H., 1980. A heteroskedastity-consistent covariance matrix and a direct test for heteroskedasticity. Econometrica 48, 817–838.

Metacognitive Development: Educational Implications

Wolfgang Schneider, Department of Psychology, University of Würzburg, Würzburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Metacognition is broadly defined as cognition about one's own cognitions. Metamemory, one subtype of metacognition that refers to knowledge about memory, has attracted particularly strong research interest in developmental psychology. Given that research devoted to aspects of metamemory and its importance for education has created numerous findings in the field, the impact of memory knowledge on educational outcomes will be discussed in more detail. Moreover, research on the role of metacognitive knowledge in the development of reading and mathematics will be summarized. Overall, findings indicate that most of what will be said about metamemory can be easily generalized to metacognitive knowledge related to a variety of problem-solving activities such as reading and mathematics.

Metacognition is broadly defined as cognition about one's own cognitions. John Flavell's (1979) seminal article on developmental aspects of metacognition and metacognitive monitoring stimulated three decades of empirical and conceptual work in developmental as well as educational and mainstream cognitive psychology. The basic proposal outlined by Flavell was that developmental changes in metacognitive knowledge mediate developmental changes in cognitive performance, including memory, language comprehension, reading, and mathematics. Metamemory, one subtype of metacognition that refers to knowledge about memory, has attracted particularly strong research interest in developmental psychology. Given that research devoted to aspects of metamemory and its importance for education has created numerous findings in the field, the impact of memory knowledge on educational outcomes will be discussed in more detail. Moreover, research on the role of metacognitive knowledge in the development of reading and mathematics will be summarized. Overall, findings indicate that most of what will be said about metamemory can be easily generalized to metacognitive knowledge related to a variety of problem-solving activities such as reading and mathematics.

Conceptualizations and Models of Metacognition

Although the concept of 'metacognition' is now frequently used and investigated in various fields of psychology and education, its history is rather short. The first studies on metacognition were developmental in nature and initiated in the early 1970s by Ann Brown, John Flavell, and their colleagues (for reviews, see Brown et al., 1983; Flavell et al., 2002; Schneider and Pressley, 1997). Although various conceptualizations of the term 'metacognition' have been used in the literature on cognitive development, the concept has usually been broadly defined as any knowledge or cognitive activity that takes as its object, or regulates, any aspect of any cognitive enterprise (cf Flavell et al., 2002). According to this conceptualization, metacognition refers to people's knowledge of their own information processing skills, as well as knowledge about the nature of cognitive tasks, and about strategies for coping with such tasks. Moreover, it also includes executive skills related to monitoring and self-regulation of one's own cognitive activities.

In a now classic paper, Flavell (1979) described three major facets of metacognition, namely, metacognitive knowledge, metacognitive experiences, and metacognitive skills, that is, strategies controlling cognition. Declarative metacognitive knowledge refers to the segment of 'world knowledge' that has to do with the human mind and its doings. According to Flavell et al. (2002), metacognitive knowledge about memory includes explicit, conscious, and factual knowledge about the importance of person, task, and strategy variables for processing and recalling information. For instance, knowing that an older child typically recalls more than a younger child is an indicator of person knowledge, and knowing that the time available for solving a mathematical problem is important for subsequent success indicates task-related metacognitive knowledge. Metacognitive experiences refer to a person's awareness and feelings elicited in a problem-solving situation (e.g., feelings of knowing), and metacognitive skills are believed to play a role in many types of cognitive activity such as oral communication of information, reading comprehension, attention, and memory. These facets of metacognition refer to a person's procedural knowledge, which Brown et al. (1983) referred to as 'knowing how,' and which can be further subdivided into monitoring and self-regulatory functions (see below). For an excellent discussion of more subtle distinctions among various aspects of metacognition, see Kuhn (1999, 2000).

The concept of metamemory was introduced by John Flavell and colleagues to refer to knowledge about memory processes and contents. In their taxonomy of metamemory, Flavell and Wellman (1977) distinguished between two main categories, 'sensitivity,' and 'variables.' The 'sensitivity' category referred to mostly implicit, unconscious behavioral knowledge of when memory is necessary, and thus was very close to subsequent conceptualizations of procedural metacognitive knowledge. The 'variables' category referred to explicit, conscious, and factual knowledge about the importance of person, task, and strategy variables for memory performance. This is also known as declarative metacognitive knowledge. Metamemory about person variables includes

knowledge about how, when, and why one remembers or forgets. Metamemory about task variables comprises knowledge about task influences on memory performance, for instance, knowledge that shorter item lists are easier to remember than long lists. Finally, metamemory about strategies refers to knowledge about advantages and possible problems of memory strategies.

This taxonomy of metamemory was not intended to be exhaustive. Since the late 1970s, a number of other theorists have contributed to the development of metamemory theory (for reviews, see Holland Joyner and Kurtz-Costes, 1997; Schneider, 2010, 2015). For instance, Paris and colleagues (e.g., Paris and Oka, 1986) introduced a component called 'conditional metacognitive knowledge' that focused on children's ability to justify or explain their decision concerning memory activities. The declarative metamemory component introduced by Flavell and Wellman focused on 'knowing that,' whereas conditional metamemory referred to 'knowing why.' The procedural metamemory component, that is, children's ability to monitor and regulate their memory behavior ('knowing how') was thoroughly analyzed by Ann Brown and colleagues (e.g., Brown et al., 1983).

Here, the frame of reference was the competent information processor, one possessing an efficient 'executive' that regulated cognitive behaviors. Brown and colleagues could demonstrate that memory monitoring and regulation processes play a large role in complex cognitive tasks such as comprehending and memorizing text materials.

Although more recent conceptualizations of metacognition have expanded the scope of the theoretical construct, they also make use of the basic distinction between declarative and procedural knowledge. For instance, Wellman (1990) has linked the declarative metacognitive component to the broader concept of children's 'theory of mind,' which focuses on classes of knowledge about the inner mental world and cognitive processes that develop during the preschool years. Pressley, Borkowski, and colleagues have systematically considered declarative and procedural components of metacognition in developing a theoretical model that emphasizes the dynamic interrelations among strategies, monitoring abilities, and motivation (e.g., Pressley et al., 1989). According to their 'good information processing model,' metacognitive knowledge is closely related to the learner's strategy use, motivational orientation, general knowledge about the world, and automated use of efficient learning procedures. All these components are assumed to interact. For instance, specific strategy knowledge influences the adequate application of metacognitive strategies, which in turn affects knowledge. As the strategies are carried out, they are monitored and evaluated, which leads to expansion and refinement of specific strategy knowledge.

It should be noted that conceptualizations of metacognition and metamemory in the fields of general cognitive psychology, social psychology, and the psychology of aging differ from this taxonomy, which is mainly used in the fields of developmental psychology and education. For instance, popular conceptualizations of metacognition in the field of cognitive psychology exclusively elaborate on the procedural component, focusing on the interplay between monitoring and self-control (see Nelson, 1996). On the other hand, when issues of declarative metamemory are analyzed in the fields of social psychology and gerontology, the focus is on a person's belief about memory phenomena and not on veridical knowledge.

Assessment of Metacognitive Knowledge

There are a variety of measures that have been used to capture what children know about memory. Measures assessing declarative metacognitive knowledge are taken without concurrent assessment of cognitive performance, whereas measures of procedural metacognitive knowledge are collected simultaneously with the measure of cognitive activity (concurrent measures). Most measures of declarative, factual knowledge have utilized interviews or questionnaires that focus on knowledge about person variables, task demands, and strategies. Earlier instruments predominantly used in the memory domain suffered from methodological problems, whereas more recent interviews and questionnaires showed better psychometric properties such as sufficient reliability and validity (Schneider and Pressley, 1997). Moreover, it was shown that nonverbal techniques helped in assessing young children's declarative knowledge. For instance, one task required young children to distinguish effective from poor strategies while watching a model executing the strategies on video. In another successful procedure ('peer-tutoring task') older children were asked to tutor a younger child about how to do a certain memory task in order to maximize learning. Peer tutoring is likely more motivating to young children than interviews, and they tend to be more explicit when answering a question of an older child as compared with that of an adult (who already seems to know everything).

Overall, these alternative methods alleviated some of the problems usually related to the use of questionnaire measures. However, these measures still created difficulties when applied to older children and adolescents, particularly when knowledge about text processing was assessed. There is the risk that social desirability factors reduce the validity of outcomes for this target group. Accordingly, more sophisticated measures of metacognition have to be used with older children and adolescents. For instance, Schlagmüller and Schneider (2007) came up with a standardized measure of metacognition that was based on a revised test instrument developed for the International Program for Student Assessment (PISA; see Artelt et al., 2001). This instrument taps adolescents' knowledge of strategies that are relevant during reading and for the comprehension as well as recall of text information. For each of six scenarios, students have to evaluate the quality and usefulness of five different strategies available for reaching the intended learning or memory goal. The rank order of strategies obtained for each scenario is then compared with an optimal rank order provided by experts in the field of text processing. The correspondence between the two rankings is expressed in a metacognition score, indicating the degree to which students are aware of the best ways to store and remember text information.

Procedural metacognitive knowledge has been assessed by measures that are characterized by the presence of

a simultaneous cognitive activity. For instance, in the area of memory research, children and adolescents are asked to judge their memory performance either shortly before, during, or after working on a memory task. The most studied type of procedural metacognitive knowledge is self-monitoring, that is, evaluating how well one is progressing. The developmental literature has focused on performance prediction or ease-of-learning (EOL) judgments, judgments of learning (JOL), and feeling-of-knowing (FOK) judgments. See Brown et al. (1983), Flavell et al. (2002), and Schneider and Lockl (2008) for overviews.

EOL judgments occur in advance of the learning process, are largely inferential, and refer to items that have not yet been learned (Nelson and Narens, 1994). The corresponding memory paradigm is performance prediction. In comparison, JOLs occur, during or soon after, the acquisition of memory materials and are predictions about future test performance on recently studied (and probably still recallable) items. Typically, paired-associate learning tasks are used in this context. After completion of a learning trial, participants are shown the stimuli of a given pair and have to indicate how confident they are whether they will remember the correct item response either immediately or 10 min later. A number of developmental studies also explored children's FOK judgments (e.g., Lockl and Schneider, 2002). These judgments occur either during or after a learning procedure and are judgments about whether a currently unrecallable item will be remembered at a subsequent retention test.

Self-monitoring involves knowing where you are with regard to your goal of understanding and memorizing task materials, whereas self-regulation includes planning, directing, and evaluating one's mnemonic activities (cf Flavell et al., 2002). Some developmental studies addressed aspects of children's control and self-regulation processes such as termination of study (recall readiness) and allocation of study time (see the reviews by Schneider, 2010, 2015). Recall readiness assessments are made after learning materials have been studied at least once. Typically, participants are asked to continue studying until their memory of the materials to be learned is perfect. Another example of self-regulation skills concerns the allocation of study time. This research observes how learners deploy their attention and effort when studying lists of items. For instance, developmental studies on the allocation of study time examined whether school children and adults were more likely to spend more time on less well-learned material. After a first free recall trial, participants had to distinguish between recalled and nonrecalled items (monitoring component), and were then asked to select half of the items for additional study (self-regulation component). One problem with the paradigm of the allocation of study time is that it may not only tap metacognitive processes but may also be influenced by motivational variables (see Schneider and Lockl, 2002).

A number of concerns have been raised about measures of procedural metamemory. However, although there is not a perfect index of metamemory, many of the measurement problems that metamemory researchers confront are similar to measurement problems in other areas of psychology.

Development of Metacognition

Precursors of Metacognitive Knowledge

Since the early 1980s, there has been study of preschoolers' metacognition motivated by Perner's (1991) and Wellman's (1990) conceptualizations of children's theory of mind, which emphasizes important classes of knowledge about the inner mental world that children acquire by the ages of 3–4 years. From this age on, children develop a rudimentary understanding of mental verbs such as 'thinking' or 'remembering' and can separate mental processes from external behaviors associated with them. They then gradually learn to recognize that the mental world can be differentiated into processes such as remembering, knowing, and guessing (that is, they acquire knowledge about distinct mental processes). Although 3- to 4-year-olds are not generally capable of differentiating these processes, older preschoolers already make distinctions that are very similar to those of adults.

The relationship between the development of language, theory of mind development, and metamemory development has only recently been systematically explored in longitudinal studies. For instance, Astington and Jenkins (1999) as well as De Villiers and Pyers (2002) found that language competence predicted theory-of-mind development, but not the reverse. Longitudinal relationships among knowledge of mental verbs, theory of mind, and subsequent metamemory were analyzed in a longitudinal study (Lockl and Schneider, 2007). It was shown that both early theory-of-mind competence and metacognitive vocabulary (i.e., knowledge of mental verbs) affected subsequent knowledge of memory. Findings also indicated that metacognitive vocabulary, theory of mind, and general metamemory improved considerably over the preschool years.

Children's Factual Knowledge About Memory and Cognition

Using sensitive methods that minimize demands on the child, it is possible to demonstrate some rudimentary knowledge of memory-related facts in preschoolers. For instance, they understand mental words, distinguish between important versus less important elements in picture stories, and know about the relevance of retrieval strategies when performing on a hide-and-seek task. Knowledge of facts about memory is more impressive in the elementary-grade years, and much more complete by the end of childhood. Nonetheless, knowledge of memory, in particular, knowledge about the interactions of metamemory variables or understanding of the relative importance of text elements continues to develop.

One of the most important findings is that there is increasing strategy knowledge with increasing age. However, there is also increasing evidence that many adolescents (including college students) have little knowledge of some powerful and important memory strategies (for details, see Pressley and McCormick, 1995). Although metacognitive knowledge increases substantially between young childhood and young adulthood, there seems still much to learn about how to read, comprehend, and memorize complex text materials or to solve mathematical

problems (see Pressley and Afflerbach, 1995; Schneider and Artelt, 2010).

Development of Procedural Metacognition

Compared with the clear-cut age trends demonstrated for declarative metacognitive knowledge, the situation regarding procedural metamemory is less clear. Early research focusing on monitoring showed that even young children seem to possess the relevant skills, particularly when the memory tasks are not very difficult (see reviews by Holland Joyner and Kurtz-Costes, 1997; Schneider and Pressley, 1997). Thus, the ability to monitor one's own memory activities is already well developed in young children and seems to improve continuously during the early elementary school years. However, the evidence regarding developmental trends is not consistent, with some studies showing better performance in younger than in older children (for a review, see Schneider, 2010). More recent studies exploring developmental trends in self-regulation (as well as the interaction between monitoring and self-regulation) were helpful in clarifying the situation.

The available evidence on the development of self-regulation skills shows that there are clear increases from middle childhood to adolescence. Spontaneous and effective use of self-regulation skills occurs only in highly constrained situations during the grade-school years and continues to develop well into adolescence. Comparisons of younger and older children in 'study-time apportionment' tasks indicate that it is the interplay between monitoring and self-regulatory activities that develops with age. That is, when the task is to learn item pairs of varying difficulty, both younger and older children show adequate monitoring skills in that they are well able to distinguish difficult from easy item pairs. However, only the older children allocate study time differentially, spending more time on the difficult than on the easier items. In comparison, younger children typically spend about the same amount of time on easy pairs as they spend on hard pairs. The available studies thus indicate an age-related improvement in the efficient allocation of study time. Developmental differences were not so much observed in the metacognitive knowledge itself but in its efficient application to self-regulation strategies.

Although the available developmental research seems to confirm the basic assumption that monitoring influences control processes, recent research with adults has challenged this position, suggesting that self-regulatory processes influence metacognitive monitoring. For instance, Koriat et al. (2006) proposed that study time is actually used by the learner as a cue for encoding fluency under what they called the memorizing effort heuristic, meaning that easily studied items are more likely to be remembered than items that require more effort to study. Accordingly, metacognitive judgments are basically data driven: study time duration is taken retrospectively as a cue for the feeling of mastery. Thus, greater effort (longer study time) is associated with lower JOLs, suggesting that the cause-and-effect relation is actually from control to monitoring. The results of Koriat et al. (2006) not only confirmed this assumption but also showed evidence for the 'monitoring affects control' hypothesis, indicating that the relationships between monitoring and control can be more complex than originally assumed. First developmental studies on this issue confirm the assumption that this is also true for children, indicating that evidence for the 'control affects monitoring' model increases with age (e.g., Koriat et al., 2009).

Metamemory–Memory Relations

From a developmental and educational perspective, the metamemory concept seems well suited to explain children's production deficiencies on a broad variety of memory tasks. Empirical research was stimulated by the belief that young children do not spontaneously use memory strategies because they are not familiar with those tasks and thus do not know anything about the advantages of strategy use. This should change soon after children enter school and are confronted with a large number of memory tasks. Experience with such tasks should improve strategy knowledge (metamemory), which in turn should have an impact on memory behavior (i.e., strategy use). Thus, the major motivation behind studying metamemory and its development has been the assumption that, although links between metamemory and memory may be weak in early childhood, they should become much stronger with increasing age.

Overall, the empirical findings do not indicate such a strong relationship. Narrative and statistical meta-analyses have shown that there are moderate, nontrivial quantitative associations between metamemory and memory. For instance, Schneider and Pressley (1997) reported an overall correlation of about 0.40 based on about 60 publications and more than 7000 children and adolescents. Qualitative reviews of the various metamemory–memory relations make obvious, however, that no single statistic could capture the diversity and richness of the findings. In general, the evidence points to a bidirectional, reciprocal relationship between metamemory and memory behavior. Metacognitive knowledge can influence cognitive behavior, which in turn leads to an enhanced metacognitive knowledge.

Although there are several reasons for the fact that the link between metacognition and cognitive performance is not always as strong as it could be, undoubtedly one of the major mediators is motivation. Empirical research has shown that metamemory–memory performance relationships are particularly impressive when participants are highly motivated to achieve a certain goal. Under these circumstances, all available resources will be activated to cope with task demands, including declarative and procedural metacognitive knowledge (cf Borkowski and Muthukrishna, 1995).

From the early 1980s on, multivariate experimental designs and also comprehensive field studies have been used to examine the complex relationships among metamemory, domain knowledge, memory behavior (strategy use), and memory performance, as well as their relationship with other important variables such as intelligence, memory capacity, and motivation (e.g., DeMarie et al., 2004; Schneider et al., 1998). For instance, Schneider et al. (1998) assessed the relationships among verbal intelligence quotient (IQ), memory capacity, domain knowledge, declarative metamemory, use of

a semantic organizational strategy, and recall in a sort-recall task in third and fourth graders. Findings indicated that metamemory was affected by both verbal IQ and memory capacity. Although there was only a modest direct contribution of metamemory to the prediction of recall, the indirect link via strategic behavior was much stronger. As a consequence, individual differences in declarative metamemory explained a large proportion in the variance of the recall data. Similar findings were also reported by DeMarie et al. (2004), who also illustrated the importance of declarative metamemory for explaining individual differences in strategy use and memory performance in different age groups ranging from kindergarteners to fifth graders.

Overall, the findings of various multivariate analyses confirm that the causal relation between metacognitive knowledge and cognitive performance is complex. Metamemory sometimes has an indirect effect on recall, as when knowledge about categorization strategies leads to semantic grouping during the study period, which in turn produces better recall. Moreover, the influence seems to be bidirectional (see Flavell et al., 2002; Schneider and Pressley, 1997). That is, metacognition can influence cognitive behavior, which in turn leads to enhanced metacognition. Finally, many other relevant variables such as IQ, domain knowledge, memory capacity, and motivation significantly contribute to the explanation of differences in cognitive performance. Obviously, these empirical findings are in accord with the core assumptions of the good information processing model (Pressley et al., 1989), which provides a detailed theoretical account of the complexity of these interactional processes.

Metacognition and Education

Metamemory and Educational Processes

During the 1980s and 1990s, numerous studies explored the efficiency of strategy training approaches in school (for a review, see Schneider and Pressley, 1997). The basic assumption was that, although children in most cases do not efficiently monitor the effectiveness of strategies they are using, they can be trained to do so. For instance, in a training program carried out by Ghatala and colleagues (e.g., Ghatala et al., 1986), elementary school children were presented with paired-associate learning tasks. Before studying these lists, some children received a three-component training. They were taught (1) to assess their performance with different types of strategies, (2) to attribute differences in performance to use of different strategies, and (3) to use information gained from assessment and attribution to guide selection of the best strategy for a task. As a major result, it was shown that even children 7–8 years of age can be taught to monitor the relative efficacy of strategies that they are using and to use utility information gained from monitoring in making future strategy selections.

Several studies on the development of memory and metamemory have emphasized the importance of educational contexts for developmental changes. Obviously, most of memory development is not so much a product of age but of education. One way in which parents and teachers facilitate cognitive development is by nurturing the development of children's metacognition. For instance, a cross-cultural study by Carr and colleagues (Carr et al., 1989) examined differences in the amount of instruction that US and German parents provided to their children at home. Carr and colleagues found that German parents, in contrast to US parents, reported more instruction of strategies, checked homework more often, and also used games that promoted thinking. These differences in home instruction were accompanied by superior strategy use by German children on the memory tasks provided in school. Carr et al. (1989) concluded from this that parental instruction is an important facilitator of children's further metacognitive development.

In a recent longitudinal study on memory development from kindergarten age to the end of elementary school, Kron-Sperl et al. (2008) repeatedly presented the children of their sample with a sort-recall memory task without giving any specific strategy cues. When performance of these children was compared with that of random samples of school children who received this task for the first time, substantial practice effects were found. Children of the longitudinal sample not only outperformed the control children regarding strategy use and memory performance, but also showed better task-specific memory. Obviously, it does not require much effort to improve children's strategy knowledge.

Metacognition and School Performance

Another area of metamemory research involves applications of metacognitive theory to educational settings. It was assumed that children's experiences at school must play an important role in shaping their use of and knowledge about how to learn and remember. However, observations of normal classroom situations did not always prove that teachers foster children's metacognitive development. For instance, Moely and colleagues observed in classrooms to find out how elementary teachers instructed strategy use and memory knowledge as they presented lessons to children in Grades K to 6 (see the overview by Moely et al., 1995). Teachers varied widely in the extent to which they focused on how children might adjust or regulate their cognitive activities in order to master a task. Efforts also varied widely depending on the subject matter under consideration. When Moely and colleagues first looked at a broad range of instructional effort, they found low levels of strategy instruction. However, teachers were found to be more likely to offer suggestions for strategy use when teaching math problem solving. Overall, there is broad agreement that one way in which parents and teachers can facilitate cognitive development is by the development of children's metacognition (see Carr et al., 1989; Coffman et al., 2008). Both classic and recent studies on this issue show that there is still room for improvement in this regard, but that considerable progress can be found.

Metacognition, Reading, and Instruction

Although rich strategy instruction is not common in schools, it can be successfully implemented. Several comprehensive research projects focused on reading instruction and

comprehension monitoring (cf Borkowski and Muthukrishna, 1995; Paris and Oka, 1986). A particularly important instructional procedure in this context is reciprocal teaching (Palincsar and Brown, 1984). Reciprocal teaching takes place in a collaborative learning context and involves guided practice in the flexible use of the following four comprehension monitoring strategies: questioning, summarizing, clarifying, and predicting. The novice's role is facilitated by the provision of scaffolding by the expert (teacher). Skills and strategies are practiced in the context of reciprocal teaching dialogues. The teacher and the students take turns leading discussions regarding the contents of a text they are jointly attempting to understand. Overall, this instructional approach has proven to be extraordinarily successful both with normal and learning-disabled students. Teachers using reciprocal instruction assume more responsibility for strategy implementation early in instruction, gradually transferring control over to the students (see Palincsar, 1986 for an extensive description, and Rosenshine and Meister, 1994 for a realistic appraisal of its benefits).

Furthermore, very ambitious programs have been undertaken by Pressley and colleagues in order to evaluate effective instructional programs in US public school systems (see Pressley and McCormick, 1995; Schneider and Pressley, 1997). Strategy instruction was not conducted in isolation, but was viewed as an integral part of the curriculum, and thus was taught as part of language arts, mathematics, science, and social studies. The goal was to simultaneously enhance children's repertoires of strategies, knowledge, metacognition, and motivation. Pressley and colleagues found that effective teachers regularly incorporated strategy instruction and metacognitive information about flexible strategy use and modification as a part of daily instruction. In accord with the assumption of the good information processing model outlined above (cf Pressley et al., 1989), effective teachers did not emphasize the use of single strategies but taught the flexible use of a range of procedures that corresponded to subject matter, time constraints, and other task demands. On most occasions, strategy instruction occurred in groups, with the teachers modeling appropriate strategy use. By comparison, the work by Moely and colleagues (e.g., Moely et al., 1995) illustrated that the effective teaching process described by Pressley and coworkers does not necessarily constitute the rule, and that effective teachers may represent a minority group in elementary school classrooms. Taken together, the careful documentations of instructional procedures carried out by Pressley, Moely, and their research groups have shown that there is a lot of potential for metacognitively guided instructional processes in children's school learning.

Other recent research explores the utility of the metacognition concept in research with older children and adolescents, assessing the predictive potential of metacognitive knowledge and skillfulness in reading and mathematics (e.g., Artelt et al., 2001; Mevarech and Kramarski, 2003; Schneider and Artelt, 2010; Veenman et al., 2005; see also the contributions in Desoete and Veenman, 2006). Overall, these studies confirm the view that metacognitive knowledge and self-regulated, insightful use of learning strategies not only are influential in elementary school children but also predict math performance and reading comprehension in secondary school settings, even after differences in intellectual abilities have been taken into account. They also give evidence that metacognitive knowledge relevant for school-related domains can still be effectively trained in late childhood and early adolescence.

The numerous recent studies on metacognition and reading as well as mathematics confirm the view that metacognitive knowledge and self-regulated, insightful use of learning strategies predicts academic performance in primary and secondary school settings even after differences in intellectual abilities have been taken into account. They also give evidence that metacognitive knowledge relevant for school-related domains normally develops during the course of primary school but is not at peak in adolescence. Findings from various intervention approaches showed that instructional settings including the training of metacognitive skills can be successful from early school age on, and can still be effective in late childhood and early adolescence.

Despite the fact that strategy instruction in classrooms is difficult and that empirical studies are still rare, there are plenty of good reasons to believe that the situation will improve in the future. There are now metacognitive training programs available that provide long-term strategy instruction and promise long-lasting success (cf Schneider, 2010, 2015). One precondition for increasing the use of such metacognition-related teaching is to increase teachers' understanding of the conceptual foundations of effective learning. As long as teachers do not think in information-processing terms, it will be difficult to establish progress in this field. However, recent changes in teacher education let us believe that teachers will understand information processing much better soon, enabling them to implement strategy training programs in the classroom that pay off for most students.

Taken together, the research programs and applied studies outlined above have enhanced greatly our understanding of how to establish long-term strategy instruction in educational contexts that not only is rich in metamemory and motivational enhancement, but also helps most students to accomplish their academic goals.

See also: Cognitive Development: Mathematics Learning and Instruction; Cognitive Styles and Learning Styles; Learning to Learn; Self-Regulated Learning; Tangible User Interfaces in Learning and Education.

Bibliography

Artelt, C., Schiefele, U., Schneider, W., 2001. Predictors of reading literacy. European Journal of Psychology of Education 16, 363–383.

Astington, J.W., Jenkins, J.M., 1999. A longitudinal study of the relation between language and theory of mind development. Developmental Psychology 35, 1311–1320.

Borkowski, J.H., Muthukrishna, N., 1995. Learning environments and skill generalization: how contexts facilitate regulatory processes and efficacy beliefs. In: Weinert, F.E., Schneider, W. (Eds.), Memory Performance and Competencies—issues in Growth and Development. Erlbaum, Mahwah, NJ, pp. 283–300.

Brown, A.L., Bransford, J.D., Ferrara, R.A., Campione, J.C., 1983. Learning, remembering, and understanding. In: Flavell, J.H., Markman, E.M. (Eds.), Handbook of Child Psychology, Cognitive Development 3. Wiley, New York, pp. 77–166.

Carr, M., Kurtz, B.E., Schneider, W., Turner, L.A., Borkowski, J.G., 1989. Strategy instruction and transfer among American and German children: environmental influences on metacognitive development. Developmental Psychology 25, 765–771.

Coffman, J.L., Ornstein, P.A., McCall, L., 2008. Linking teachers' memory-relevant language and the development of children's memory skills. Developmental Psychology 44, 1640–1654.

DeMarie, D., Miller, P.H., Ferron, J., Cunningham, W.R., 2004. Path analysis tests of theoretical models of children's memory performance. Journal of Cognition and Development 5, 461–492.

De Villiers, J., Pyers, J., 2002. Complements to cognition: a longitudinal study of the relationship between complex syntax and false-belief understanding. Cognitive Development 17, 1037–1060.

Desoete, A., Veenman, M. (Eds.), 2006. Metacognition in Mathematics Education. Nova Science, Haupauge, NY.

Flavell, J.H., 1979. Metacognition and cognitive monitoring—A new area of cognitive-developmental inquiry. American Psychologist 34, 906–911.

Flavell, J.H., Miller, P.H., Miller, S.A., 2002. Cognitive Development, fourth ed. Prentice-Hall, Englewood Cliffs, NJ.

Flavell, J.H., Wellman, H.M., 1977. Metamemory. In: Kail, R.V., Hagen, W. (Eds.), Perspectives on the Development of Memory and Cognition. Erlbaum, Hillsdale, NJ, pp. 3–33.

Ghatala, E.S., Levin, J.R., Pressley, M., Goodwin, D., 1986. A componential analysis of effects of derived and supplied strategy—utility information on children's strategy selections. Journal of Experimental Child Psychology 41, 76–92.

Holland Joyner, M., Kurtz-Costes, B., 1997. Metamemory development. In: Cowan, N. (Ed.), The Development of Memory in Childhood. Psychology Press, Hove, UK, pp. 275–300.

Koriat, A., Ackerman, R., Lockl, K., Schneider, W., 2009. The Memorizing-effort Heuristic in Judgments of Learning: A Developmental Perspective. Cognitive Development 24, 169–182.

Koriat, A., Ma'ayan, H., Nussinson, R., 2006. The intricate relationships between monitoring and control in metacognition. Journal of Experimental Psychology: General 135, 36–69.

Kron-Sperl, V., Schneider, W., Hasselhorn, M., 2008. The development and effectiveness of memory strategies in kindergarten and elementary school: findings from the Würzburg and Göttingen longitudinal studies. Cognitive Development 23, 79–104.

Kuhn, D., 1999. Metacognitive development. In: Balter, L., Tamis-LeMonda (Eds.), Child Psychology: A Handbook of Contemporary Issues. Psychology Press, Philadelphia, PA, pp. 259–286.

Kuhn, D., 2000. Theory of mind, metacognition, and reasoning: a life—span perspective. In: Mitchell, P., Riggs, K.J. (Eds.), Children's Reasoning and the Mind. Psychology Press, Hove, UK, pp. 301–326.

Lockl, K., Schneider, W., 2002. Developmental trends in children's feeling-of-knowing judgements. International Journal of Behavioral Development 26, 327–333.

Lockl, K., Schneider, W., 2007. Knowledge about the mind: links between theory of mind and later metamemory. Child Development 78, 148–167.

Mevarech, Z.R., Kramarski, B., 2003. The effects of metacognitive training versus worked-out examples on students' mathematical reasoning. British Journal of Educational Psychology 73, 449–471.

Moely, B., Santulli, K., Obach, M., 1995. Strategy instruction, metacognition, and motivation in the elementary school classroom. In: Weinert, F.E., Schneider, W. (Eds.), Memory Performance and Competencies—Issues in Growth and Development. Erlbaum, Mahwah, NJ, pp. 301–321.

Nelson, T.O., 1996. Consciousness and metacognition. American Psychologist 51, 102–116.

Nelson, T.O., Narens, K., 1994. Why investigate metacognition? In: Metcalfe, J., Shimamura, A.P. (Eds.), Metacognition. Knowing about Knowing. MIT Press, Cambridge, MA, pp. 1–25.

O'Sullivan, J.T., Howe, M.L., 1998. A different view of metamemory with illustrations from children's beliefs about long-term retention. European Journal of Psychology of Education 13, 9–28.

Palincsar, A.S., 1986. The role of dialogue in providing scaffolded instruction. Educational Psychologist 21, 73–98.

Palincsar, A.S., Brown, A.L., 1984. Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. Cognition and Instruction 1, 117–175.

Paris, S.G., Oka, E.R., 1986. Children's reading strategies, metacognition, and motivation. Developmental Review 6, 25–56.

Perner, J., 1991. Understanding the Representational Mind. MIT PRESS, Cambridge, MA.

Pressley, M., Afflerbach, P., 1995. Verbal Protocols of Reading: The Nature of Constructively Responsive Reading. Erlbau, Hillsdale, NJ.

Pressley, M., Borkowski, J.G., Schneider, W., 1989. Good information processing: what it is and what education can do to promote it. International Journal of Educational Research 13, 857–867.

Pressley, M., McCormick, C.B., 1995. Advanced Educational Psychology for Educators, Researchers, and Policymakers. Harper Collins, New York.

Rosenshine, B., Meister, C., 1994. Reciprocal teaching: a review of the research. Review of Educational Research 64, 479–530.

Schlagmüller, M., Schneider, W., 2007. WLST-12. Würzburger Lesestrategie Wissenstest für die Klassen 7 bis 12. Hogrefe, Göttingen.

Schneider, W., 2010. Metacognition and memory development in childhood and adolescence. In: Salatas Waters, H., Schneider, W. (Eds.), Metacognition, Strategy Use, and Instruction. Psychology Press, New York, pp. 54–84.

Schneider, W., 2015. Memory Development from Early Childhood through Emerging Adulthood. Springer, New York.

Schneider, W., Artelt, C., 2010. Metacognition and mathematics education. ZDM Mathematics Education 42, 149–161.

Schneider, W., Lockl, K., 2002. The development of metacognitive knowledge in children and adolescents. In: Perfect, T.J., Schwartz, B.L. (Eds.), Applied Metacognition. Cambridge University Press, Cambridge, UK, pp. 224–257.

Schneider, W., Lockl, K., 2008. Procedural metacognition in children: evidence for developmental trends. In: Dunlosky, J., Bjork, R.A. (Eds.), Handbook of Metamemory and Memory. Psychology Press, New York, pp. 391–410.

Schneider, W., Pressley, M., 1997. Memory Development between Two and Twenty, second ed. Erlbaum, Mahwah, NJ.

Schneider, W., Schlagmüller, M., Visé, M., 1998. The impact of metamemory and domain-specific knowledge on memory performance. European Journal of Psychology and Education 13, 91–103.

Veenman, M., Kok, R., Blöte, A., 2005. The relation between intellectual and metacognitive skills in early adolescence. Instructional Science 33, 193–211.

Wellman, H.M., 1990. The Child's Theory of Mind. MIT Press, Cambridge, MA.

Metaphor, Role in Social Thought: History of the Concept

David Lambourn, University of Warwick, Coventry, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9738–9743, © 2001, Elsevier Ltd.

Abstract

Following a broad appreciation of metaphor in the classical period and its subsequent life in the history of rhetoric, a contrast is drawn in early modernity revealing a period of resistance to metaphor, using the English writers Hobbes and Locke. This contrasts with the uses to which metaphor was put by Vico and Herder whose work challenged the previously understood relation between the literal and the metaphorical and provided a basis for new descriptions of the social, including anthropology and ethnology. Nietzsche marks a significant development proposing that metaphor offers a creative and critical plurality, making an explicit link between metaphor and the theory of persons in relation. Metaphor identifies a new perspective of human experience making it possible for metonymy to identify its parts, synecdoche to elaborate the relations between the attributes of parts to the whole, and irony to inspect the earlier practices for their adequacy, by a test of opposition. The redescription of natural science as having a dependency on metaphor, and the later work of Wittgenstein are each seen as helping to overcome the earlier resistance, opening a possibility for a subsequent burgeoning of academic study of the subject in the closing decades of the twentieth century. Links are described between perspectives on metaphor with anthropology, ethnology, ethnography, ethnomethodology, psychoanalysis, economics, psychology, and sociology. Any further links between metaphor and the achievement of intimacy and community and the ways in which these might be modified by particular social and economic structures yet remain to be developed, although a comprehensive, utopian, and pragmatist redescription of modern culture and politics drawing upon the changing status of metaphor has been offered by Richard Rorty.

Metaphor has resisted any wide agreement as concept, yet the last few decades have witnessed a burgeoning of work and interest in metaphor and its related tropes. The many attempts to theorize metaphor have included, *inter alia*, characterizations of metaphor as 'comparison,' as 'without meaning,' as 'anomaly,' 'speech act,' 'loose talk,' 'interaction,' and 'intentional category mistakes.' These attempts have had the form of trying to assimilate metaphor under a previously existing understanding of language – a move that is itself metaphorical. This resistance to conceptualizing suggests that it might be more helpful to consider metaphor in terms of its uses.

Aristotle

For most of its life metaphor has had its home in rhetoric, and it has been to Aristotle that writers have usually turned for the earliest considerations. He understood it as 'giving a thing a name that belongs to something else,' and thought that 'the greatest thing by far is to have a command of metaphor.' Not all writers have heeded the contexts within which Aristotle wrote. His teacher, Plato, had sufficient doubts about the value of metaphor that he barred poets from his Republic. Aristotle, by contrast, found uses for metaphor, not only in politics where rhetoric enabled a man to be heard effectively in public, but also in law, where juries were suspicious of evidence that could be faked, and witnesses who could be bribed. They were, rather, influenced by arguments turning around a balance of probabilities, providing an important context for rhetoric. Aristotle cannot easily be read as limiting metaphor to the realm of ornament, or efficiency of utterance: he sees clear cognitive uses:

> … strange words simply puzzle us; ordinary words convey only we what we know already; it is from metaphor that we can best get hold of new ideas.
>
> 1984, *Rhetoric* III: 1410b

We may suppose that Aristotle's description of metaphor as giving something a name which belongs to something else, detailed as: "the transference being either from genus to species, or from species to genus, or from species to species, or on grounds of analogy" (1984, *Poetics*: 1457b), also owed something to his practice as taxonomist of the natural world. Although Aristotle's comment on Plato's forms that "(to say) they are patterns and the other things share in them is to use empty words and poetical metaphors" (1984, *Metaphysics*: 991a21), may be taken as evidence of his ambivalence to metaphor, but should rather be understood as evidence of a variety of rhetorical uses. His reference to analogy and his remark that 'to make good metaphors implies an eye for resemblances' has led his theory to be dubbed 'the comparative theory.' Aristotle may be understood, in short, as praising metaphor in poetry and drama, in law, in politics, and in what we might now refer to as natural science.

The subsequent history of metaphor until the end of the Renaissance may best be understood as part of the history of rhetoric.

Early Modernity and Early Resistance to Metaphor

Part of the development of rhetoric in seventeenth century Europe was a 'simplification' of the tropes to just four: metaphor; metonymy in which the name of an attribute or other adjunct is substituted for the intended object as with 'scepter' for 'authority'; synecdoche in which a part signifies the whole; and irony where a term is used which is opposite to the one intended to be read.

The first major philosophy written in English, credited to Thomas Hobbes (1588–1679) and his *Leviathan* (1651), published a year after Descartes' death, is severe on metaphor and claims that, of the seven causes of absurdity "The sixth (is), to the use of metaphors, tropes, and other rhetorical figures, instead of words proper." Yet the opening paragraph of the

English version of the text elaborates what Black (1962) would later call the 'associated commonplaces' of the title image, Leviathan, Hobbes's metaphor for the State. Hobbes was writing in complicated, changing, and dangerous times, with differing purposes, addressing different audiences and with changing beliefs, about which Quentin Skinner has written elucidating the use of rhetoric in a context of humanism, of the growth of science, and the growing tensions with the church (1996). John Locke (1632–1704), following Hobbes, thought that "all the artificial and figurative application of words … (are) in all discourses that pretend to inform or instruct, wholly to be avoided; and where truth and knowledge are concerned, cannot but be thought a great fault." Here began a tradition later called the 'purity thesis,' still sustained in the British Isles at the beginning of the twenty-first century.

Meanwhile, in continental Europe, other voices were heard in praise of metaphor. Giambatista Vico (1668–1744) used metaphor as a tool within philology to produce a theory of 'poetic logic.' This included the notion that names were given to features of the natural world that were already well known, drawing for example upon the body – rivers and bottles were given 'mouths,' etc.

> Whence we derive the following (principle of) criticism for the times in which the metaphors were born in (various) languages. All metaphors which, by means of likenesses taken from bodies, come to signify the labors of abstract minds, must belong to times in which philosophies had begun to become more refined. This is shown by the following: that in all languages the words necessary to cultivated arts and recondite sciences have rural origins.
>
> Vico, 1744: paragraph 404

Vico's view led him to reverse the standard relationship of the poetic to the literal.

> From all this it follows that all the tropes (and they are all reducible to the four types discussed) which have hitherto been considered ingenious inventions of writers, were necessary modes of expression of all the first poetic nations, and had originally their full native propriety. But these expressions of the first nations later became figurative when, with the further development of the human mind, words were invented which signified abstract forms or genera comprising their species or relating parts with their wholes. And here begins the overthrow of two common errors of the grammarians: that prose speech is proper speech, and poetic speech improper; and that prose speech came first and afterward speech in verse.
>
> Vico, 1744: paragraph 409

Vico's theory links the actions of metaphor, metonymy, and synecdoche. Metaphor identifies a new domain of human experience making it possible for metonymy to identify its parts, synecdoche to elaborate the relations between the attributes of parts to the whole, and irony to inspect the earlier practices for their adequacy, by a test of opposition. When irony is correctly detected the new object may be regarded as appropriately established in the consciousness of the person or group. Vico's views were not fully effective until the nineteenth century but views in some respects similar to his were held by Herder (1744–1803), born in the year Vico died. Herder's application was not the diachronic one of Vico, but rather the synchronic, through which he came to celebrate difference in cultures and provided an intellectual basis for the beginnings of anthropology and ethnology.

Some dissident voices were also heard in England. Coleridge (1772–1834), when asked why he attended Davy's chemistry lectures, replied that he wanted to renew his stock of metaphors. Shelley (1792–1822) provided a connection between metaphor, language, and culture:

> Language is vitally metaphorical; that is, it marks the before unapprehended relations of things and perpetuates their apprehension, until words, which represent them, become, through time, signs for portions or classes of thought instead of pictures of integral thoughts: and then, if no new poets should arise to create afresh the associations which have thus been disorganized, language will be dead to all the nobler purposes of human intercourse.

Nietzsche and Freud

Nietzsche (1844–1900) marks a significant development. Moving beyond both Herder and Vico and their celebration of cultural plurality, extending Shelly's view, and in a Darwinian context, he proposed that metaphor serves a creative and critical plurality. With earlier writers we may be tempted to observe that with a change of view of metaphor there seems to be a linked change of the theory of the human being. With Nietzsche, however, the linkage is explicit:

> The drive toward the formation of metaphors is the fundamental human drive, which one cannot for an instant dispense with in thought, for one would thereby dispense with man himself. This drive is not truly vanquished and scarcely subdued by the fact that a regular and rigid new world is constructed as its prison from its own ephemeral products, the concepts.
>
> Nietzsche, 1979: p. 88

Turning towards the domain of philosophy, he remarks in an early writing:

> What then is truth? A movable host of metaphors, metonymies, and anthropomorphisms: in short, a sum of human relations which have been poetically and rhetorically intensified, transferred, and embellished, and which, after long usage, seem to a people to be fixed, canonical, and binding. Truths are illusions that we have forgotten are illusions; they are metaphors that have become worn out and have been drained of sensuous force, coins that have lost embossing and are now considered as metal and no longer as coins.
>
> Nietzsche, 1979: p. 84

The metaphor for truth, of coin which loses its face, has attracted many writers, few of who have commented on the phrase 'a sum of human relations' as applied to the tropes, a perception yet remaining to be developed.

Freud (1856–1939), some 12 years younger than Nietzsche, with mutual intimate acquaintances, made the claim that Nietzsche's "guesses and intuitions often agree in the most astonishing way with the laborious findings of psychoanalysis."

Whatever that relationship might yet turn out to have been, Freud's work had the effect of applying Nietzsche's perceptions to the 'unconscious,' a new metaphor of depth, drawing upon the classical myths for his lexicon – which, in turn, made available rich seams of linked metaphors. I.A. Richards (1893–1979) who provided terms for the study of metaphor: 'tenor,' 'vehicle,' and 'tension.' These terms, though by no means unambiguous, and failing to encompass the full range of metaphors, nevertheless helped to revive talk of metaphor. Richards linked psychoanalysis to metaphor.

> The psychoanalysts have shown us with their discussions of 'transference'—another name for metaphor—how constantly modes of regarding, of loving, of acting, that have developed with one set of things or people, are shifted to another. They have shown us chiefly the pathology of these transferences, cases where the vehicle—the borrowed attitude, the parental fixation, say—tyrannizes over the new situation, the tenor, and behavior is inappropriate. The victim is unable to see the new person except in terms of the old passion and its accidents. The victim reads the situation only in terms of the figure, the archetypal image, the vehicle. But in healthy growth, tenor and vehicle—the new human relationship and the family constellation—cooperate freely; and the resultant behavior derives in due measure from both. Thus in happy living the same patterns are exemplified and the same risks of error are avoided as in tactful and discerning reading. The general form of the interpretative process is the same, with a small-scale instance—the right understanding of a figure of speech—or with a large-scale instance—the conduct of a friendship.
>
> Richards, 1936: p. 135ff.

More recently, Freud's writing has begun to be redescribed in ways that reveal its tropical basis (e.g., Lacan *The Four Fundamental Concepts of Psychoanalysis*) (*see* Psychoanalysis, History of).

The Tide Turns, Natural Science Exposed as Metaphor-Dependent

Quine (1908–2000) writing as an empiricist nevertheless thought it: "a mistake, then, to think of language usage as literalistic in its main body and metaphorical in its trimming. Metaphor, or something like it, governs both the growth of language and our acquisition of it" (1979: p. 160). Mary Hesse has extensively described the work of metaphor in science as compatible with this view. Max Black (1962) offered an 'interactive' theory of metaphor as a development of the 'comparison' view, and further linked metaphor and model, thereby linking the worlds (languages?) of literature and science. In the same year, Thomas Kuhn (1962), offered a sociological description of natural science, one recognized more by sociologists and historians than by philosophers, which had as one of its central notions that of a 'paradigm' which, in some uses, may be taken as a synonym for metaphor. This provided a stimulus for developing a social critique of natural science (see Latour and Woolgar, *Laboratory Life: The Construction of Scientific Facts*). In 1970, Rom Harre distinguished models in science into two kinds, homeomorphs which are modeled upon its subject, as is a doll, and paramorphs where the model is only distantly related, Harré's examples included Harvey's use of hydraulics to model, and discover, the circulation of blood. Paramorphs, close cousins to metaphor, are creative: homeomorphs have heuristic uses only. It is perhaps ironic that the stimulus for much work on metaphor may have come from this identification of the place of metaphor in science.

The Tide Turns, Within the Philosophy of Language

Ferdinand de Saussure (1857–1913) within his structural view of language, made a distinction between paradigmatic and syntagmatic axes that can be seen to correspond to a distinction made by Roman Jakobson (1896–1982) between two types of aphasia. In one, the victims have difficulty with the selection of a term and are dependent upon the context, which is on the contiguity of the discourse to continue to be part of it. By contrast, other aphasics, were seen to substitute equivalent terms e.g., table for lamp, smoke for pipe. Jakobson saw these as examples of metonymy and metaphor on the grounds that these were the most condensed expression of the contiguity or similarity distinctions. This changed perception of the relation between metaphor and metonymy led to developments in structuralism for example, by Levi-Strauss (1908–2009) in anthropology, and by Jacques Lacan (1901–81) in psychoanalysis. Accepting Jakobson's suggestions, David Lodge offered a typology of modern writing based upon an opposition between metaphor and metonymy (1977).

Heidegger's writing is rich with metaphor yet he denied it. He construed the sensory/nonsensory distinction as belonging to metaphysics and drew the conclusion that "the metaphorical exists only within the boundaries of the metaphysical." By expressing a preference for the 'poetical' over the 'mathematical,' Heidegger has been influential in offering a reading of Nietzsche recovering a role for philosophy by means of what we might identify as metaphor of the kind which, as a sentence, is obviously false: "Philosophy is essentially untimely because it is one of those few things that can never find an echo in the present."

Derrida (1930–2004) radicalized Saussure's structural description of language, both in his conceptualizing and also in the style of his writing – in so doing, he matched the congruence in Nietzsche who linked his valuation of metaphor to his style of writing. Derrida's view of language relies upon a play of differences, and understands metaphor as operating within that play of differences – he was unable to use the otherwise conventional distinction between the metaphorical and the literal. His identification of phallogocentrism undermined the metaphor of depth for more perspicuous knowledge, paving the way for a metaphor of breadth. Hans-Georg Gadamer (1900–2002) understood as a major contributor to hermeneutics also wrote of metaphor.

> It is the genius of linguistic consciousness to be able to give expression to these similarities. This is its fundamental metaphorical nature, and it is important to see that it is the prejudice of a theory of logic that is alien to language if the metaphorical use of a word is regarded as not its real sense.
>
> Gadamer, 1960: p. 429

This view strengthens the possibility of substituting a metaphor of breadth for that of depth to describe knowing and understanding, begun by Derrida. Paul Ricoeur (1913–2005) in

a conservative review of the principal writers, sees metaphor as the continuing source for the redescription of the world, but only in a perspective within which philosophy would be concerned to manage the legitimacy of its uses, limiting its influence and restricting any epistemological damage.

While it may be interesting to speculate what part the earlier publication of Ludwig Wittgenstein's (1889–1951) *Tractatus* (1922) might have played in making the *Philosophical Investigations* (1953) as influential as it has proved, the former, though built upon metaphor, has no place for it within its closed system, whereas the latter uses metaphor throughout, but without attempting an explicit theory of it. The style of the *Philosophical Investigations* is as if we hear just one side of a conversation – simply a limitation imposed if the writing is not to be that of drama or a novel? Perhaps the *Investigations* will come to be judged as the writing which, more than any other, licensed the growth of academic interest in metaphor (*see* Hermeneutics; Critical Theory; Linguistic Turn).

The Tide Turns, The Broader Academic Response

Kenneth Burke (1897–1993), following Vico, selected 'metonymy, irony, metaphor and synecdoche' and labeled them the 'Four Master Tropes' and set them to new (as he thought) uses:

> I refer to metaphor, metonymy, synecdoche, and irony. And my primary concern with them here will be not with their purely figurative usage, but with their role in the discovery and description of 'the truth.' It is an evanescent moment that we shall deal with—for not only does the dividing line between the figurative and literal uses shift, but also the four tropes shade into one another. Give a man but one of them, tell him to exploit its possibilities, and if he is thorough in doing so, he will come upon the other three.
>
> 1969: p. 503

Four influential interdisciplinary conferences were held in the academic year beginning September 1977, following the publication of a bibliography, *Metaphor* (Shibbles, 1971). Later, some 3500 publications were identified in the period 1985–90 (van Noppen and Hols, 1990).

In one, came a significant moment with a sharp difference between Max Black and Donald Davidson. Following Richards and Black, it had been normal to speak of 'literal meaning' and 'metaphorical meaning.' Davidson relied upon a distinction between what words mean and what they do, and took metaphor to belong exclusively to the domain of use. The terms used in metaphor should be held to mean exactly what they mean in their literal use.

Andrew Ortony collected an influential number of papers representing a variety of perspectives on metaphor (1979). The rate of change merited a second edition as early as 1993 (1993). Within philosophy, Cooper provided a comprehensive and well received discussion (1986). Seen by many as a significant event, the publication in 1980 by Lakoff and Johnson (1980) of *Metaphors We Live By*, copiously illustrated the endemic nature of metaphor in everyday understanding and extended metaphor's claim to cognitive uses. More recently, this has been developed and broadened by Raymond Gibbs (1994).

By the end of the twentieth century there was scarcely a field of academic enquiry within the behavioral sciences which had not begun to explore its areas of interest with the benefit of an analysis using metaphor and its related tropes as tools e.g., Brown (1977) in sociology, Leary (1990) in psychology, McCloskey (1986) in economics, and much is owed to Hayden White who developed Vico's insights and applied them to the writing of history providing a most virtuous circle (1973, 1985).

Metaphor and 'The Social'

We have, almost in passing, encountered pointers towards a closer association between the metaphorical and the social. Darwin's *Origin of Species* provided the opportunity to replace a view of language as correspondence with reality, with that of a series of tools to support survival. Metaphor in this view is a tool-making tool. Nietzsche referred to the tropes as 'a sum of human relations.' Cohen (1979) drew attention to similarities between metaphors and jokes, and to the experience which he calls the achievement of intimacy:

> There is a unique way in which the maker and the appreciator of a metaphor are drawn closer to one another. Three aspects are involved: the speaker issues a kind of concealed invitation; the hearer expends a special effort to accept the invitation; and this transaction constitutes the acknowledgment of a community.

This description may easily be extended into a research program. Davidson has recently maintained that thought itself absolutely depends on a three way relationship between at least two people and a series of events that are shared in the world.

Critical both of Saussure's 'objectivist' linguistics and of 'subjectivist' perspectives, Bakhtin (1895–1975) developed a view of language as sustained and developed in particular social relationships, in turn embedded in wider political and economic conditions and, paradigmatically, with consciousness being thought of as inhering in the person-within-the-group. This view allows for full social interaction, including the nonverbal, to be understood as modifying the associations which attached to the terms deployed in conversation. Wittgenstein's style in the *Philosophical Investigations* is reminiscent of Bakhtin's view.

Although computer technology has made possible the assembly of large corpora of searchable language texts drawn from a variety of publications, it seems not yet to be practicable to incorporate significant contextual detail into these databases, other than of an immediate textual kind, providing a limitation on the development of a social understanding of metaphor, by this route. For the time being, it would seem that some kind of ethnographic research will be most suited to explore the human-relational sources of metaphor use, with Bakhtin's perspective suggesting the possibility of developments in ethnomethodology. Some suggestions, focusing on 'conversational realities,' have been made by John Shotter (1993) and a comprehensive, utopian, and pragmatist redescription of modern culture and politics drawing upon the changing status of metaphor, has been

offered by Richard Rorty (1989). Metaphor is the meeting place of the human with the natural.

See also: Anthropology: Overview; Freud, Sigmund (1856–1939); Language and Poetic Structure; Modernity; Nietzsche, Friedrich (1844–1900); Saussure, Ferdinand de (1857–1913); Wittgenstein, Ludwig (1889–1951).

Bibliography

Aristotle, 1984. The Complete Works of Aristotle: The Revised Oxford Translation, 2 vols. Princeton University Press, Princeton, NJ.

Bakhtin, M., 1979. Estetika Slovesnogo Tvorchestra (Moscow) (English translation, 1986. Speech Genres & Other Late Essays. University of Texas, Austin, TX.).

Black, M., 1962. Models and Metaphor. Cornell University Press, Ithaca, NY.

Blumenberg, H., 1979. Schiffbruch mit Zuschauer, Paradigma einer Daseinsmetapher. Suhrkamp Verlag, Frankfurt am Main, Germany (English translation, 1997. Shipwreck with Spectator, Paradigm of a Metaphor for Existence. MIT Press, Cambridge, MA.).

Brown, R.H., 1977. A Poetic for Sociology. Cambridge University Press, Cambridge, UK.

Burke, K., 1969. The four master tropes. In: A Grammar of Motives. University of California, Berkeley, CA.

Cohen, T., 1979. Metaphor and the cultivation of intimacy. In: Sachs, S. (Ed.), On Metaphor. University of Chicago Press, Chicago.

Cooper, D., 1986. Metaphor. Blackwell, Oxford, UK.

Derrida, J., 1972. La mythologie blanche. In: Marges de la Philosophie. Les Editions de Minuit, Paris (English translation, 1982. White mythology. In: Margins of Philosophy. Harvester, Hemel Hempstead, UK.).

Derrida, J., 1987. Le retrait de metaphor. Enclitic 2 (2), 5–33 (English translation, 1998. The retreat of metaphor. In: Wolfreys, J. (Ed.), The Derrida Reader: Writing Performances. Edinburgh University Press, UK.).

Fiumara, G.C., 1995. The Metaphoric Process. Routledge, London.

Gadamer, H.-G., 1960. Wahrheit und Methode. Mohr (Paul Siebeck), Tübingen, Germany (English translation, 1989. Truth and Method, second ed. Sheed & Ward, London.).

Gibbs, R., 1994. The Poetics of Mind. Cambridge University Press, Cambridge, UK.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.

Leary, D.E., 1990. Metaphors in the History of Psychology. Cambridge University Press, Cambridge, UK.

Lakoff, G., Johnson, M., 1980. Metaphors We Live By. University of Chicago Press, Chicago.

Lodge, D., 1977. The Modes of Modern Writing. Arnold, London.

McCloskey, D.N., 1986. The Rhetoric of Economics. Wheatsheaf, Brighton, UK.

Nietzsche, F., 1979. Philosophy and Truth. Humanities Press, Atlantic Highlands, NJ.

van Noppen, J.-P., Hols, E., 1990. Metaphor II: A Classified Bibliography of Publications 1985 to 1990. John Benjamins, Amsterdam.

Ortony, A., 1993. Metaphor and Thought, second ed. Cambridge University Press, Cambridge, UK (first ed., 1979).

Quine, W.V., 1979. A postscript on metaphor. In: Sachs, S. (Ed.), On Metaphor. University of Chicago Press, Chicago.

Richards, I.A., 1936. A Philosophy of Rhetoric. Oxford University Press, Oxford, UK.

Ricoeur, P., 1975. La Metaphore Vive. Éditions du Seuil, Paris [English translation, 1978. The Rule of Metaphor. Routledge and Kegan Paul, London.].

Rorty, R., 1989. Contingency, Irony and Solidarity. Cambridge University Press, Cambridge, UK.

Sacks, S., 1979. On Metaphor. University of Chicago Press, Chicago.

Shibbles, W.A., 1971. Metaphor: An Annotated Bibliography and History. Language Press, Whitewater, WI.

Shotter, J., 1993. Conversational Realities. Sage, London.

Skinner, Q., 1996. Reason and Rhetoric in the Philosophy of Hobbes. Cambridge University Press, Cambridge, UK.

Vico, G.B., 1744. Scienza Nuova, third ed. Mosca, Naples, Italy (English translation, 1968. New Science. Cornell University Press, Ithaca, NY.).

White, H., 1973. Metahistory: The Historical Imagination in Nineteenth Century Europe. Johns Hopkins University Press, Baltimore, MD.

White, H., 1985. Tropics of Discourse. Johns Hopkins University Press, Baltimore, MD.

Wittgenstein, L., 1953. Philosophical Investigations. Blackwell, Oxford, UK.

Method of Moments and Generalized Method of Moments

Lars P Hansen, University of Chicago, Chicago, IL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9743–9751, © 2001, Elsevier Ltd.

Abstract

This article describes empirical methods for estimating dynamic economic systems using time series data. By design, the methods target specific feature of the dynamic system and do not require a complete specification of the time series evolution. The resulting generalized-method-of-moments estimation and inference methods use estimating equations implied by some components of a dynamic economic system. This article describes the statistical methods and some applications of these methods.

Introduction

In many empirical investigations of dynamic economic systems, statistical analysis of a fully specified stochastic process model of the time series evolution is too ambitious. Instead, it is fruitful to focus on specific features of the time series without being compelled to provide a complete description. This leads to the investigation of partially specified dynamic models. For instance, the linkages between consumption and asset returns can be investigated without a full description of production and capital accumulation. The behavior of monetary policy can be explored without a complete specification of the macroeconomic economy. Models of inventory behavior can be estimated and tested without requiring a complete characterization of input and output prices. All of these economic models imply moment conditions of the form:

$$Ef(x_t, \beta_o) = 0 \qquad [1]$$

where f is known a priori, x_t is an observed time series vector, and β_o is an unknown parameter vector. These moment relations may fall short of providing a complete depiction of the dynamic economic system. Generalized method of moments (*GMM*) estimation, presented in Hansen (1982), aims to estimate the unknown parameter vector β_o and to test these moment relations in a computationally tractable way. This article first compares the form of a *GMM* estimator to closely related estimators from the statistics literature. It then reviews applications of these estimators to partially specified models of economic time series. Finally, it considers *GMM*-related moment-matching problems in fully specified models of economic dynamics.

Minimum Chi-Square Estimation

To help place *GMM* estimation in a statistical context, I explore a closely related minimum chi-square estimation method. Statisticians developed minimum chi-square estimators to handle restricted models of multinomial data and a variety of generalizations. Neyman (1949) and Burankin and Gurland (1951), among others, aimed to produce statistically efficient and computationally tractable alternatives to maximum likelihood estimators. In the restricted multinomial model, estimators are constructed by forming empirical frequencies and minimizing Pearson's chi-square criterion or some modification of it.

The method has direct extensions to any moment-matching problem. Suppose that $\{x_t\}$ is a vector process, which temporarily is treated as being *iid*. Use a function ψ with n coordinates to define target moments associated with the vector x_t. A model takes the form

$$E[\psi(x_t)] = \phi(\beta_o)$$

where β_o is an unknown parameter. The moment-matching problem is to estimate β_o by making the empirical average of $\{\psi(x_t)\}$ close to its population counterpart $\phi(\beta_o)$:

$$\min_{\beta} \frac{1}{T}\sum_{t=1}^{T}[\psi(x_t) - \phi(\beta)]' V \sum_{t=1}^{T}[\psi(x_t) - \phi(\beta)] \qquad [2]$$

where V is the distance or weighting matrix. The distance matrix sometimes depends on the data and/or the candidate parameter vector β. The use of V to denote a weighting matrix that may actually depend on parameters or data is an abuse of notation. This simple notation is used because it is the probability limit of the weighting matrix evaluated at the parameter estimator that dictates the first-order asymptotic properties.

The limiting distribution of the criterion in eqn [2] is chi-square distributed with n degrees of freedom if the parameter vector β_o is known, and if V is computed by forming the inverse of either the population or sample covariance matrix of $\psi(x)$. When β is unknown, estimates may be extracted by minimizing this chi-square criterion, and hence the name. To preserve the chi-square property of the minimum (with an appropriate reduction in the degrees of freedom), we again form the inverse sample covariance matrix of $\psi(x)$, or form the inverse population covariance matrix for each value of β. The minimized chi-square property of the criterion may be exploited to build tests of overidentification and to construct confidence sets for parameter values. Results like these require extra regularity conditions, and this rigor is supplied in some of the cited papers.

While the aim of this research was to form computationally tractable alternatives to maximum likelihood estimation, critical to statistical efficiency is the construction of a function ψ of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.42029-5

the data that is a sufficient statistic for the parameter vector (see Burankin and Gurland, 1951). Berkson (1944) and Taylor (1953) generalize the minimum chi-square approach by taking a smooth one-to-one function h and building a quadratic form of $h\left[\frac{1}{T}\sum_t \phi(x_t)\right] - h[\psi(\beta)]$. Many distributions fail to have a finite number of sufficient statistics; but the minimum chi-square method continues to produce consistent, asymptotically normal estimators provided that identification can be established.

GMM estimators can have a structure very similar to the minimum chi-square estimators. Notice that the core ingredient to the moment-matching problem can be depicted as in eqn [1] with a separable function f

$$f(x,\beta) = \psi(x) - \phi(\beta)$$

used in the chi-square criterion function. Target moments are one of many ways for economists to construct inputs into chi-square criteria, and it is important to relax this separability. Moreover, in *GMM* estimation, the emphasis on statistical efficiency is weakened in order to accommodate partially specified models. Finally, an explicit time series structure is added, when appropriate.

GMM Estimation

Our treatment of *GMM* estimation follows Hansen (1982), but it builds from Sargan's (1958, 1959) analyses of linear and nonlinear instrumental variables (*see* Instrumental Variables in Statistics and Econometrics). See Ogaki (1993) for a valuable discussion of the practical implementation of *GMM* estimation methods. *GMM* estimators are constructed in terms of a function f that satisfies eqn [1], where f has more coordinates, say n, than there are components to the parameter vector β_o. Another related estimation method is M-estimation. M-estimation is a generalization of maximum likelihood and least squares estimation. M-estimators are typically designed to be less sensitive to specific distributional assumptions (*see* Robustness). These estimators may be depicted as solving a sample counterpart to eqn [1] with a function f that is nonseparable, but with the same number of moment conditions as parameter estimators.

In Sections Linear Models, Models of Financial Markets, and From Densities to Diffusions we will give examples of the construction of the f function, including ones that are not separable in x and β, and ones for which there are more coordinates than parameter vectors. A minimum chi-square-type criterion is often employed in *GMM* estimation. For instance, it is common to define the *GMM* estimator as the solution to

$$b_T = \arg\min_{\beta} Tg_T(\beta)'Vg_T(\beta)$$

where

$$g_T(\beta) = \frac{1}{T}\sum_{t=1}^{T} f(x_t,\beta)$$

and V is a positive definite weighting matrix. This quadratic form has the chi-square property provided that V is an estimator of the inverse of an appropriately chosen covariance matrix, one that accounts for temporal dependence.

The following sections survey some applications of *GMM* estimators to economic time series. A feature of many of these examples is that the parameter β by itself may not admit a full depiction of the stochastic process that generates data. *GMM* estimators are constructed to achieve 'partial' identification of the stochastic evolution and to be robust to the remaining unmodeled components.

Time Series Central Limit Theory

Time series estimation problems must make appropriate adjustments for the serial correlation for the stochastic process $\{f(x_t, \beta_o)\}$. A key input into the large sample properties of *GMM* estimators is a central limit approximation

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f(x_t,\beta_o) \Rightarrow \text{Normal}\,(0,\Sigma_o)$$

for an appropriately chosen covariance matrix $\sum_o$. An early example of such a result was supplied by Gordin (1969), who used martingale approximations for partial sums of stationary, ergodic processes. See Hall and Heyde (1980) for an extensive discussion of this and related results. The matrix $\sum_o$ must include adjustments for temporal dependence:

$$\sum_o = \sum_{j=-\infty}^{+\infty} Ef(x_t,\beta_o)f(x_{t-j},\beta_o)'$$

which is the long-run notion of a covariance matrix that emerges from spectral analysis of time series. In many *GMM* applications, martingale arguments show that the formula for $\sum_o$ simplifies to include only a small number of nonzero terms. It is the adjustment to the covariance matrix that makes the time series implementation differ from the *iid* implementation (Hansen, 1982).

Adapting the minimum chi-square apparatus to this environment requires that we estimate the covariance matrix $\sum_o$. Since f is not typically separable in x and β, an estimator of $\sum_o$ requires an estimator of β_o. This problem is familiar to statisticians and econometricians through construction of feasible generalized least squares estimators. One approach is to form an initial estimator of β_o with an arbitrary nonsingular weighting matrix V and to use the initial estimator of β_o to construct an estimator of $\sum_o$. Hansen (1982), Newey and West (1987), and many others provide consistency results for the estimators of $\sum_o$. Another approach is to iterate back and forth between parameter estimation and weighting matrix estimation until a fixed point is reached, if it exists. A third approach is to construct an estimator of $\sum(\beta)$ and to replace V in the chi-square criterion by an estimator of $\sum(\beta)^{-1}$ constructed for each β. Given the partial specification of the model, it is not possible to construct $\sum(\beta)$ without the use of the time series data. Long-run covariance estimates, however, can be formed for the process $\{f(x_t, \beta)\}$ for each choice of β. Hansen et al. (1996) refer to this method as *GMM* estimation with a continuously updated weighting matrix. Hansen et al. (1996), Newey and Smith (2000), and

Stock and Wright (2000) describe advantages of using continuous updating, and Sargan (1958) shows that in some special circumstances this method reproduces a quasi-maximum likelihood estimator.

Efficiency

Since the parameter vector β that enters moment condition eqn [1] may not fully characterize the data evolution, direct efficiency comparisons of *GMM* estimators with parametric maximum likelihood are either not possible or not interesting. However, efficiency statements can be made for narrower classes of estimators.

For the study of *GMM* efficiency, instead of beginning with a distance formulation, index a family of *GMM* estimators by the moment conditions used in estimation. Specifically study

$$ag_T(b_T) = 0$$

where a is a k by n selection matrix. The selection matrix isolates which (linear combination of) moment conditions will be used in estimation and indexes alternative *GMM* estimators. Estimators with the same selection matrix have the same asymptotic efficiency. Without further normalizations, multiple indices imply the same estimator. Premultiplication of the selection matrix a by a nonsingular matrix e results in the same system of nonlinear equations. In practice, the selection matrix can depend on data or even on the parameter estimator provided that the selection matrix has a probability limit. As with weighting matrices for minimum chi-square criteria, we suppress the possible dependence of the selection matrix a on data or on parameters for notational simplicity. The resulting *GMM* estimators are asymptotically equivalent (to possibly infeasible) estimators in which the selection matrix is replaced by its probability limit. As has been emphasized by Sargan (1958, 1959) in his studies of instrumental variables estimators, estimation accuracy can be studied conveniently as a choice a^* of an efficient selection matrix. The link between a weighting matrix V and selection matrix a is seen in the first-order conditions:

$$a = V\frac{1}{T}\sum_{t=1}^{T}\frac{\partial f(x_t, b_T)}{\partial \beta}$$

or their population counterpart:

$$a = Vd$$

where

$$d = E\left[\frac{\partial f(x_t, \beta_o)}{\partial \beta}\right]$$

Other distance measures including analogs to the ones studied by Berkson (1944) and Taylor (1953) can also be depicted as a selection matrix applied to the sample moment conditions $g_T(\beta)$. Moreover, *GMM* estimators that do not solve a single minimization problem may still be depicted conveniently in terms of selection matrices. For example, see Heckman (1976) and Hansen (1982) for a discussion of recursive estimation problems, which require solving two minimization problems in sequence.

Among the class of estimators indexed by a, the ones with the smallest asymptotic covariance matrix satisfy

$$a = ed'\Sigma_o^{-1}$$

where e is any nonsingular matrix, and the best covariance matrix is $d'(\Sigma_o)^{-1}d$. This may be shown by imitating the proof of the famed Gauss–Markov Theorem, which establishes that the ordinary least squares estimator is the best linear unbiased estimator.

Semiparametric Efficiency

Many *GMM* applications, including ones that we describe subsequently, imply an extensive (infinite) collection of moment conditions. Application of the previous analysis requires an ad hoc choice of focal relations to use in estimation. Hansen (1985) extends this indexation approach to time series problems with an infinite number of moment conditions. The efficiency bound for an infinite family of *GMM* estimators can be related to the efficiency of other estimators using a semiparametric notion of efficiency. A semiparametric notion is appropriate because an infinite-dimensional nuisance parameter is needed to specify fully in the underlying model. It has been employed by Chamberlain (1987) and Hansen (1993) in their study of *GMM* estimators constructed from conditional moment restrictions.

Linear Models

Researchers in econometrics and statistics have long struggled with the idea of how to identify an unknown coefficient vector β_o in a linear model of the form

$$\beta_o \cdot y_t = u_t$$

where y_t is a k-dimensional vector of variables observed by an econometrician. Least squares method solves this problem by calling one of the variables, y_{1t}, the dependent variable and requiring the remaining variables, y_{2t}, to be orthogonal to the disturbance term:

$$E(u_t, y_{2t}) = 0$$

Alternatively, as suggested by Karl Pearson and others, when there is no natural choice of a left-hand side variable, we may identify β_o as the first principal component, the linear combination of y_t with maximal variance subject to the constraint $|\beta| = 1$.

A third identification scheme exploits the time series structure and has an explicit economic motivation. It is a time series analog to the instrumental variables and two-stage least squares estimators familiar to economists. Suppose that a linear combination of y_t cannot be predicted given data sufficiently far back into the past. That is,

$$E(\beta_o \cdot y_t | \mathscr{F}_{t-m}) = 0 \qquad [3]$$

where $\mathscr{F}_t$ is a conditioning information set that contains at least current and past values of y_t. This conditional moment restriction can be used to identify the parameter vector β_o, up to scale. In this setup there may be no single variable to be designated as endogenous with the remainder being

exogenous or even predetermined. Neither least squares nor principal components are appropriate for identifying β_o.

The model or the precise context of the application dictates the choice of lag m. For instance, restriction eqn [3] for a specified value of m follows from martingale pricing relations for multiperiod securities, from Euler equations of the investment problems faced by decision makers, or from the preference horizons of policymakers.

To apply *GMM* to this problem, use eqn [3] to deduce the matrix equation

$$E(Z_{t-m}\gamma_t')\beta_o = 0$$

where z_t is an n-dimensional vector of variables in the conditioning information set $\mathscr{F}_t$ and $n \geq k - 1$. By taking unconditional expectations we see that β_o is in the null space of the n by k matrix $E(Z_{t-m}\gamma_t')$. The model is overidentified if $n \geq k$. In this case the matrix $E(Z_{t-m}\gamma_t')$ must be of reduced rank and in this sense is special. This moment condition may be depicted as in eqn [1] with

$$f(x_t, \beta) = Z_{t-m}\gamma_t'\beta_o$$

where x_t is a vector containing the entries of z_{t-m} and γ_t. The *GMM* test of overidentification, based on say a minimum chi-square objective, aims to detect this reduced rank. The *GMM* estimator of β_o seeks to exploit this reduced rank by approximating the direction of the null space.

For the *GMM* applications in this section and Section Models of Financial Markets, there is a related but independent statistics literature. Godambe and Heyde (1987) and others study efficiency criteria of martingale estimation equations for a fully specified probability model. By contrast, our models are partially specified and have estimation equations, e.g., $\sum_{t=1}^{T}[Z_{t-m}\gamma_t']\beta_o$, that is a martingale only when $m = 1$. When the estimation equation is not a martingale, Hansen (1982, 1985) constructs an alternative martingale approximation to analyze statistical efficiency.

Given that the null space of $Ez_{t-m}\gamma_t'$ is not degenerate, it might have more dimensions than one. While the moment conditions are satisfied (the null space of $E(Z_{t-m}\gamma_t')$ is nondegenerate), the parameter vector itself is underidentified. Recent literature on weak instruments is aimed at blurring the notion of being underidentified. See Stock and Wright (2000) for an analysis of weak instruments in the context of *GMM* estimation.

As posed here the vector β_o is not identified but is at best identified up to scale. This problem is analogous to that of principal component analysis. At best, we might hope to identify a one-dimensional subspace. Perhaps a sensible estimation method should therefore locate a subspace rather than a parameter vector. Normalization should be inessential to the identification beyond mechanically selecting the parameter vector within this subspace. See Hansen et al. (1996) for a discussion of normalization invariant *GMM* estimators. This argument is harder to defend than it might seem at first blush. Prior information about the normalized parameter vector of interest may eliminate interest in an estimation method that is invariant to normalization.

There are some immediate extensions of the previous analysis. The underlying economic model may impose further, possibly nonlinear, restrictions on the linear subspace. Alternatively, the economic model might imply multiple conditional moment relations. Both situations are straightforward to handle. In the case of multiple equations, prior restrictions are needed to distinguish one equation from another, but this identification problem is well studied in the econometrics literature.

Applications

A variety of economic applications produce conditional moment restrictions of the form in eqn [3]. For instance, Hansen and Hodrick (1980) studied the relation between forward exchange rates and future spot rates, where m is the contract horizon. Conditional moment restrictions also appear extensively in the study of a linear-quadratic model of production smoothing and inventories (see West (1995) for a survey of this empirical literature). The first-order or Euler conditions for an optimizing firm may be written as eqn [3] with $m = 2$. Hall (1988) and Hansen and Singleton (1996) also implement linear conditional moment restrictions in the lognormal models of consumption and asset returns. These moment conditions are again derived from first-order conditions, but in this case for a utility-maximizing consumer/investor with access to security markets. Serial correlation due to time aggregation may cause m to be two instead of one. Finally, the linear conditional moment restrictions occur in the literature on monetary policy response functions when the monetary authority is forward looking. Clarida et al. (2000) estimate such models in which m is dictated by preference horizon of the Federal Reserve Bank in targeting nominal interest rates.

Efficiency

The choice of z_{t-m} in applications is typically ad hoc. The one-dimensional conditional moment restriction of eqn [3] actually gives rise to an infinite number of unconditional moment restrictions through the choice of the vector z_{t-m} in the conditioning information set $\mathscr{F}_{t-m}$. Any notion of efficiency based on a preliminary choice z_{t-m} runs the risk of missing some potentially important information about the unknown parameter vector β_o. It is arguably more interesting to examine asymptotic efficiency across an infinite-dimensional family of *GMM* estimators indexed by feasible choices of z_{t-m}. This is the approach adopted by Hansen (1985), Hansen et al. (1988), and West (2001). The efficient *GMM* estimator that emerges from this analysis is infeasible because it depends on details of the time series evolution. West and Wilcox (1996) and Hansen and Singleton (1996) construct feasible counterpart estimators based on approximating the time series evolution needed to construct the efficient z_{t-m}. In particular, West and Wilcox (1996) show important improvements in the efficiency and finite sample performance of the resulting estimators.

Efficient *GMM* estimators based on an infinite number of moment conditions often place an extra burden on the model builder by requiring that a full dynamic model be specified. The cost of misspecification, however, is not so severe. While a mistaken approximation of the dynamics may cause an

efficiency loss, by design it will not undermine the statistical consistency of the *GMM* estimator.

Models of Financial Markets

Models of well-functioning financial markets often take the form

$$E(d_t z_t | \mathscr{F}_{t-m}) = q_{t-m} \qquad [4]$$

where z_t is a vector of asset payoffs at date t, q_{t-m} is a vector of the market prices of those assets at date t, and $\mathscr{F}_t$ is an information set available to economic investors at date t. By assumption, the price vector q_t is included in the information set $\mathscr{F}_t$. The random variable d_t is referred to as a 'stochastic discount factor' between dates $t-m$ and date t. This discount factor varies with states of the world that are realized at date t and encodes risk adjustments in the security market prices. These risk adjustments are present because some states of the world are discounted more than others, which is then reflected in the prices. The existence of such a depiction of asset prices is well known since the work of Harrison and Kreps (1979), and its conditional moment form used here and elsewhere in the empirical asset pricing literature is justified in Hansen and Richard (1987). In asset pricing formula [4], m is the length of the financial contract, the number of time periods between purchase date and payoff date.

An economic model of financial markets is conveniently posed as a specification of the stochastic discount factor d_t. It is frequently modeled parametrically

$$d_t = g(y_t, \beta_o) \qquad [5]$$

where the function g is given a priori but the parameter β_o is unknown, and is a target of estimation. Models of investor preferences may be written in this manner as can observable factor models in which d_t is a function (often linear) or vector y_t of 'factors' observed by an econometrician. See Cochrane (2001), for example, and a discussion of how this approach is connected to other empirical methods in financial economics.

To apply *GMM* estimation, inference, and testing to this problem we do two things. First we replace eqn [4] by its unconditional counterpart

$$E(d_t z_t - q_{t-m}) = 0$$

which can be justified by applying the law of iterated expectations. Second we substitute eqn [5] into this unconditional moment implication

$$E[g(y_t, \beta_o) z_t - q_{t-m}] = 0$$

This may be depicted as eqn [1] by writing

$$f(x_t, \beta) = g(y_t, \beta) z_t - q_{t-m}$$

where x_t contains the entries of the factors y_t, the asset payoffs z_t, and the corresponding prices q_{t-m}. As is evident from Hansen and Singleton (1982) and Kocherlakota (1996), *GMM*-based statistical tests of this model give rise to one characterization of what is commonly termed as the 'equity premium puzzle.' This puzzle is a formal statement of how the observed risk–return trade-off from financial market data is anomalous when viewed through the guises of many standard dynamic equilibrium models from the macroeconomics and finance literatures.

Studying unconditional rather than conditional moment relations may entail a loss of valuable information for estimation and testing. As emphasized by Hansen and Richard (1987), however, information in the set $\mathscr{F}_{t-m}$ may be used by an econometrician to form synthetic payoffs and prices. This information is available at the initiation date of the financial contract. While use of conditioning information to construct synthetic portfolios reduces the distinction between conditional and unconditional moment restrictions, it introduces additional challenges for estimation and inference that are being confronted in ongoing econometric research.

From Densities to Diffusions

A century ago Karl Pearson proposed a family of density functions and a moments-based approach to estimating a parameterized family of densities. The densities within this family have logarithmic derivatives that are the ratio of a first-order to a second-order polynomial. More recently, Wong (1964) provided scalar diffusion models with stationary distributions in the Pearson family. Diffusion models are commonly used in economic dynamics and finance.

A scalar diffusion is a solution to a stochastic differential equation:

$$\mathrm{d}x_t = \mu(x_t)\mathrm{d}t + \sigma(x_t)\mathrm{d}B_t$$

where μ is the local mean or drift for the diffusion, σ^2 is the local variance or diffusion coefficient, and B_t is standard Brownian motion. I now revisit Pearson's estimation problem and method, but in the context of a data generated by a diffusion.

The stationary density q of a diffusion satisfies the integral equation

$$\int_{\underline{r}}^{\overline{r}} \left(\mu\phi + \frac{\sigma^2}{2} \frac{\mathrm{d}\phi}{\mathrm{d}x} \right) q \mathrm{d}x = 0 \qquad [6]$$

for a rich class of ϕ's, referred to as test functions. This gives an extensive family of moment conditions for any candidate (μ, σ^2). In particular, moment conditions of the form eqn [1] may be built by parameterizing μ and σ^2 and by using a vector of test functions ϕ. Pearson's moment recursions are of this form with test functions that are low-order polynomials.

Pearson's method of estimation was criticized by R.A. Fischer because it failed to be efficient for many members of the Pearson family of densities. Pearson and Fischer both presumed that the data generation is *iid*. To attain asymptotic efficiency with *iid* data, linear combinations of $f(x, \beta_o)$ should reproduce the score vector for an implied likelihood function. Low-order polynomials fail to accomplish this for many members of the Pearson family of densities, hence the loss in statistical efficiency.

When the data are generated by a diffusion, the analysis of efficiency is altered in a fundamental way. Parameterize the time series model via $(\mu_\beta, \sigma_\beta^2)$. There is no need for restrict μ and σ^2 to be low-order polynomials. Let Φ be a vector of test functions with at least as many component functions as parameters to estimate. Associated with each Φ is a *GMM* estimator by forming

$$f(x,\beta) = \mu_\beta(x)\Phi(x) + \frac{1}{2}\,\sigma_\beta^2(x)\frac{\mathrm{d}\Phi}{\mathrm{d}x}(x)$$

The moment conditions eqn [1] follow from eqn [6].

Conley et al. (1997) calculate the efficiency bounds for this estimation problem using the method described in Hansen (1985). They perform this calculation under the simplifying fiction that a continuous-data record is available and show that an efficient choice of Φ is

$$\Phi_\beta(x) = \frac{\partial}{\partial\beta}\left[\frac{2\mu_\beta(x) + \mathrm{d}\sigma_\beta^2(x)\big/\mathrm{d}x}{\sigma_\beta^2(x)}\right] \qquad [7]$$

The choice of test function turns out to be the 'derivative' of the score vector with respect to the Markov state x. This efficient choice of Φ depends on the true parameter vector β_o and hence is infeasible to implement. Conley et al. (1997) show that the same efficiency can be attained by a feasible estimator in which Φ is allowed to depend on β as in eqn [7]. The score function derivative is easier to compute than the score function itself because the constant of integration for implied density does not have to be evaluated for each β. While tractable, this efficient test function solution to the time series problem will typically not lead one to use low-order polynomials as test functions even for Pearson's parameterizations.

Pearson used moment recursions to construct tractable density estimates without numerical integration. Computational concerns have subsided, and the diffusion model restricts more than just the stationary density. Transition densities can also be inferred numerically for a given (μ, σ^2) pair. To motivate fitting only the stationary density, a model builder must suppose potential model misspecification in the transition dynamics. One example of this form of misspecification is a subordinated diffusion model used to model financial time series in which calendar time and a more relevant information-based notion of economic time are distinct.

Integral eqn [6] is 'localized' by using smooth test functions that concentrate their mass in the vicinity of given points in the state space. Banon (1978) exploited this insight to produce a nonparametric drift estimator using a locally constant parameterization of μ. Conley et al. (1997) justify a local linear version of the *GMM* test function estimator described in the previous subsection.

Related Approaches

Since around 1990, statisticians have explored empirical likelihood and other related methods. These methods are aimed at fitting empirical data distributions subject to a priori restrictions, including moment restrictions that depend on an unknown parameter. In particular, Qin and Lawless (1994) have shown how to use empirical likelihood methods to estimate parameters from moment restrictions like those given in eqn [1] for *iid* data. While Qin and Lawless (1994) use the statistics literature on estimation equations for motivation, they were apparently unaware of closely related econometrics literature on *GMM* estimation. Baggerly (1998) describes a generalization of the method of empirical likelihood based on the Cressie–Read divergence criterion. Imbens et al. (1998) and Bonnal and Renault (2001) use this generalization to unify the results of Qin and Lawless (1994), Imbens (1997), Kitamura and Stutzer (1997), and others and *GMM* estimation with a continuously updated weighting matrix. (See Newey and Smith (2000) for a related discussion.) Within the *iid* framework, Bonnal and Renault (2001) show that the continuously updated *GMM* estimator is a counterpart to empirical likelihood except that it uses Pearson's χ^2 criterion. The relative entropy or information-based estimation of Imbens (1997) and Kitamura and Stutzer (1997) is of the same type but based on an information criterion of fit for the empirical distribution.

Kitamura and Stutzer (1997) and others use clever blocking approaches for weakly dependent data to adapt these methods to time series estimation problems. Many *GMM* applications imply 'conditional' moment restrictions where the conditioning information is lagged m time periods. Extensions of these empirical distribution methods to accommodate time series 'conditional' moment implications are an important area of research.

Moment Matching Reconsidered

Statistical methods for partially specified models allow researchers to focus an empirical investigation and to understand sources of empirical anomalies. Nevertheless, the construction of fully specified models is required to address many questions of interest to economists such as the effect of hypothetical interventions or policy changes.

Models of economic dynamical systems remain highly stylized, however; and they are not rich enough empirically to confront a full array of empirical inputs. Producing interesting comparative dynamic results for even highly stylized dynamic systems is often difficult, if not impossible, without limiting substantially the range of parameter values that are considered. Since analytical solutions are typically not feasible, computational methods are required. As in other disciplines, this has led researchers to seek ways to calibrate models based on at least some empirical inputs. See Hansen and Heckman (1996) for a discussion of this literature. The analysis of dynamical economic systems brings to the forefront both computational and conceptual problems.

The computational problems are reminiscent to those confronted by the inventors of minimum chi-square methods. Two clever and valuable estimation methods closely related to moment matching have recently emerged from the econometrics literature. One method is called 'indirect

inference' and has been advanced by Smith (1993) and Gourieroux et al. (1993). The idea is to fit a conveniently chosen, but misspecified approximating model that is easy to estimate. The empirical estimates from the approximating model are used as targets for the estimation of the dynamic economic model. These targets are 'matched' using a minimum chi-square criteria.

This method is sometimes difficult to implement because the implied approximating models associated with the dynamic economic model of interest may be hard to compute. Gallant and Tauchen (1996) circumvent this difficulty by using the score function for the approximating model (evaluated at the empirical estimates) as targets of estimation. The computational burden is reduced to evaluating an empirical score expectation as a function of the underlying parameter of interest. This allows researchers to use an expanded set of approximating statistical models.

In both of these methods there are two models in play, an approximating statistical model and an underlying economic model. These methods address some of the computational hurdles in constructing parameter estimators through their use of convenient approximating statistical models. On the other hand, it is harder to defend their use when underlying dynamic economic model is itself misspecified. Gallant and Tauchen (1996), for instance, use statistical efficiency as their guide in choosing approximating statistical models. Better approximating models result in more accurate estimators of the parameters of interest. This justification, however, neglects the role of misspecification in the underlying dynamic economic model. For 'indirect inference' it is often not evident how to construct approximating models that leave the estimates immune to the stylized nature of the underlying economic model. Thus, it remains an important challenge for econometricians to devise methods for infusing empirical credibility into 'highly stylized' models of dynamical economic systems. Dismissing this problem through advocating only the analysis of more complicated 'empirically realistic' models will likely leave econometrics and statistics on the periphery of important applied research. Numerical characterizations of highly stylized models will continue with or without the aid of statistics and econometrics.

Bibliography

Baggerly, K.A., 1998. Empirical likelihood as a goodness-of-fit measure. Biometrika 85, 535–547.

Banon, G., 1978. Nonparametric identification for diffusion processes. SIAM Journal of Control and Optimization 16, 380–395.

Berkson, J., 1944. Application of the logistic function to bioassay. Journal of the American Statistical Association 39, 357–365.

Bonnal, H., Renault, E., 2001. Minimal Chi-Square Estimation with Conditional Moment Restrictions. CIRANO, Montreal, PQ.

Burankin, E., Gurland, J., 1951. On Asymptotically Normal Estimators: I, vol. 1. University of California Publications in Statistics, 86–130.

Chamberlain, G., 1987. Asymptotic efficiency in estimation with conditional moment restrictions. Journal of Econometrics 34, 305–334.

Clarida, R., Gali, J., Gertler, M., 2000. Monetary policy rules and macroeconomic stability: evidence and some theory. Quarterly Journal of Economics 115, 147–180.

Cochrane, J., 2001. Asset Pricing. Princeton University Press, Princeton, NJ.

Conley, T.G., Hansen, L.P., Luttmer, E.G.J., Scheinkman, J.A., 1997. Short-term interest rates as subordinated diffusions. Review of Financial Studies 10, 525–577.

Gallant, A.R., Tauchen, G., 1996. Which moments to match. Econometric Theory 12, 657–681.

Godambe, V.P., Heyde, C.C., 1987. Quasi-likelihood and optimal estimation. International Statistical Review 55, 231–244.

Gordin, M.I., 1969. Central limit theorem for stationary processes. Soviet Mathematics Doklady 10, 1174–1176.

Gourieroux, C., Monfort, A., Renault, E., 1993. Indirect inference. Journal of Applied Econometrics 8, S85–S118.

Hall, P., Heyde, C.C., 1980. Martingale Limit Theory and Its Application. Academic Press, Boston.

Hall, R.E., 1988. Intertemporal substitution and consumption. Journal of Political Economy 96, 339–357.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.

Hansen, L.P., 1985. A method for calculating bounds on asymptotic covariance matrices of generalized method of moments estimators. Journal of Econometrics 30, 203–238.

Hansen, L.P., 1993. Semiparametric efficiency bounds for linear time-series models. In: Phillips, P.C.B. (Ed.), Models, Methods and Applications of Econometrics: Essays in Honor of A. R. Bergstrom, Chapter 17. Blackwell, Cambridge, MA, pp. 253–271.

Hansen, L.P., Heaton, J.C., Ogaki, M., 1988. Efficiency bounds implied by multiperiod conditional moment restrictions. Journal of the American Statistical Society 88, 863–871.

Hansen, L.P., Heaton, J.C., Yaron, A., 1996. Finite sample properties of some alternative GMM estimators. Journal of Business and Economic Statistics 14, 262–280.

Hansen, L.P., Heckman, J.J., 1996. The empirical foundations of calibration. Journal of Economic Perspectives 10, 87–104.

Hansen, L.P., Hodrick, R.J., 1980. Forward exchange rates as optimal predictors of future spot rates. Journal of Political Economy 88, 829–853.

Hansen, L.P., Richard, S.F., 1987. The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models. Econometrica 55, 587–613.

Hansen, L.P., Singleton, K.J., 1982. Generalized instrumental variables of nonlinear rational expectations models. Econometrica 50, 1269–1286.

Hansen, L.P., Singleton, K.J., 1996. Efficient estimation of linear asset pricing models with moving-average errors. Journal of Business and Economic Statistics 14, 53–68.

Harrison, J.M., Kreps, D.M., 1979. Martingales and arbitrage in multiperiod securities markets. Journal of Economic Theory 20, 381–408.

Heckman, J.J., 1976. The common structure of statistical methods of truncation, sample selection, and limited dependent variable and a simple estimator for such models. Annals of Economic and Social Measurement 5, 475–492.

Imbens, G.W., 1997. One-step estimators for over-identified generalized method of moments models. Review of Economic Studies 64, 359–383.

Imbens, G.W., Spady, R.H., Johnson, P., 1998. Information theoretic approaches to inference in moment condition estimation. Econometrica 66, 333–357.

Kitamura, Y., Stutzer, M., 1997. An information-theoretic alternative to generalized method of moments estimation. Econometrica 65, 861–874.

Kocherlakota, N., 1996. The equity premium: it's still a puzzle. Journal of Economic Literature 34, 42–71.

Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.

Newey, W.K., Smith, R.J., 2000. Asymptotic Bias and Equivalence of GMM and GEL, Manuscript.

Neyman, J., 1949. Contributions to the Theory of the χ^2 Test. Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, CA.

Ogaki, M., 1993. Generalized method of moments: econometric applications. In: Maddala, G.S., Rao, C.R., Vinod, H.D. (Eds.), Handbook of Statistics, vol. 11. Elsevier Science, Amsterdam, pp. 455–486 (Chapter 17).

Qin, J., Lawless, J., 1994. Empirical likelihood and general estimating equations. Annals of Statistics 22, 300–325.

Sargan, J.D., 1958. The estimation of economic relationships using instrumental variables. Econometrica 26, 393–415.

Sargan, J.D., 1959. The estimation of relationships with auto-correlated residuals by the use of instrumental variables. Journal of the Royal Statistical Society 21, 91–105.

Smith, A., 1993. Estimating nonlinear time series models using simulated vector autoregressions. Journal of Applied Econometrics 8, 63–84.
Stock, J.H., Wright, J.H., 2000. GMM with weak identification. Econometrica 68, 1055–1096.
Taylor, W., 1953. Distance functions and regular best asymptotically normal estimates. Annals of Mathematical Statistics 24, 85–92.
West, K.D., 1995. Inventory models. In: Handbook in Applied Econometrics (Macroeconometrics). Basil Blackwell, Oxford, UK.
West, K.D., 2001. On optimal instrumental variables estimation of stationary time series models. International Economic Review 42 (4), 1043–1050.
West, K.D., Wilcox, D.W., 1996. A comparison of alternative instrumental variables estimators of a dynamic linear model. Journal of Business and Economic Statistics 14, 281–293.
Wong, E., 1964. The construction of a class of stationary Markov processes. In: Bellman, R. (Ed.), Sixteenth Symposium in Applied Mathematics – Stochastic Processes in Mathematical Physics and Engineering. American Mathematical Society, Providence, RI, pp. 264–276.

Methodological Individualism in Sociology

Emmanuel Picavet, Université Paris-I Panthéon-Sorbonne, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Methodological individualism in sociology is a family of methodological strategies that portray human agents, their actions, and their interaction as key ingredients of social explanation. Insofar as social phenomena are the results of individual actions, methodological individualism supports a practice of sociology which concentrates on individual motivations of various kinds. This often suggests the use of decision-theoretic methods which are used in other disciplines as well, with possible benefits for interdisciplinary cooperation. Methodological individualism raises problems with respect to the different levels of social explanation (or understanding).

Methodological individualism in sociology refers to explanatory and modeling strategies in which human individuals (with their motivations) and human actions (with their causes or reasons) are given a prominent role in explanations and models. Social phenomena are viewed as the aggregate results of individual actions. Explanation takes its origin in the parts rather than the whole: individual action is invested with explanatory primacy in relation to social facts, society's properties, and observed statistical regularities. Among the problems associated with this methodology, the following three are of special interest: How should individual characteristics be selected and connected? Is the relevance of individualism as a methodology grounded in social ontology, or is it primarily dependent on the known explanatory or predictive benefits to be gained from the use of a family of models and theories? How should individualistic social science take individual and social cognition into account?

Individualism and the Form of Explanations

Giving Primacy to Individuals

While the constitution of sociology as an autonomous discipline has involved the recognition of a separate layer of 'social facts,' methodological individualism, which presupposes the existence of social facts as an *explanandum* for social science, is not alien to the sociological tradition, as exemplified by the work of Weber, Pareto, and others. A general feature of individualistic explanations is that individual motivations, preferences, reasons, propensities, or characteristics generally speaking figure explicitly in the proposed models and explanations, together with the description of relevant technological and natural circumstances.

Methodological individualism has gained influence in the twentieth century through the work of such authors as Mises, Popper, and Hayek, who associated it with liberal views (in different versions). Within the limits of explanatory or predictive social science, its emergence is traceable to the debates in nineteenth-century Germany and Austria about the nature of explanation in history, the status of economic theory, and the respective scientific roles of nomological explanation (a kind of explanation of social facts that makes use of general laws) and contextual understanding. It is classically associated with the requirement of a real understanding of the motivations, intentions, or objectives of the social actors themselves (as illustrated by Simmel, 1900; Weber, 1922). It has made it possible to explore, after Adam Smith, the unintended consequences of individual intentional actions (Cherkaoui, 2006), thus contributing to revealing the complexity of the social world in connection with human action. This is exemplified by the systematic study of the emergence of unplanned norms (Opp, 2001), or conventions and social institutions (Mantzavinos, 2008).

This methodology can be contrasted with several types of nonindividualistic methods (Boyer, 1992). It is violated by those theories which rely on the operation of impersonal forces (such as nature, mind, history, progress, and destiny) or unconscious forces and by 'holistic' explanations in which the properties of collective entities (such as nations, classes, social groups, or society as a whole) are given an independent explanatory role. This opposition refers to explanatory strategies, and it should not be used to oppose two kinds of social science.

Indeed, social scientists who want to be faithful to methodological individualism may well recognize (after Pareto, 1905) that the 'individual' is social to the core, if only because thought and action are intimately related to social realities such as a language, laws, custom, and the like. Culture, education, and traditions give shape to human desires, beliefs, and dispositions, which can be expressed in a shared language and recognized in interpersonal relationships, mutual expectations, and social norms. Following the Paretian methodology, the 'social' is not to be confused with our preferred representations of society as we like it. In particular, the 'social' dimension of social life should not be described by referring to features of a desired society such as solidarity, equality, cohesiveness, social harmony, and so on. Moreover, we should take care not to equate the 'social' with widespread feelings.

On the normative or prescriptive side, the 'social' dimension as Pareto thinks of it must be reconstructed in a problematic way: Are we really able to speak on behalf of 'society' at large? This makes sense when there is a unanimous agreement. If there is no agreement, however, it is deeply objectionable. It amounts to giving primacy to the abstract notion of 'society,' possibly at the expense of concrete, living individuals (especially those who do not belong to majorities).

The other way around, those sociologists who attempt to develop explanations of social facts on the basis of properties of social life are not automatically blind to the social importance of a proper methodological treatment of individuals. They may find it necessary to explore the social determination of 'personal' characteristics. In addition, they may acknowledge the fact that complex individual operations are involved in the processes through which causality flows from social characteristics to particular social events.

In Durkheim's thought, society and the individual have a complex relationship, and Durkheimian 'holism' is a subtle one: social cooperation is a source of empowerment for individuals, but, as cooperation grows (with the concomitant evolution in goals), society becomes a source of impersonal constraints which limit the expression of particularity (Demeulenaere, 2003, ch. 2). Following Boudon's reading of Durkheim (Boudon, 2000, ch. 2), the commitment to explaining features of society as a *sui generis* entity should not mislead us: the sociologist stresses that society is made of individuals, and his or her explanation of collective beliefs (such as religious beliefs) relies on a theory of individual initiatives or activities which result in shared representations. The underlying process is a process of innovation, selection, and retention. Whenever we say that beliefs and languages originate in 'society,' following this reading of Durkheim, this is really shorthand for a more complex theory in which personal initiatives have a major role, even though we do not know the details of individual actions.

Methodological individualism is compatible with ontological holism about collective entities, in the following sense: social scientists who support the methodology may recognize the existence of social entities (such as 'cultures' or 'traditions') which are irreducible to individual component parts. But they postulate that only individuals have goals and interests, and that these have explanatory value with respect to human conduct. They interpret the notion of collective belief in such a way that it cannot be equated with society's thought (e.g., an individualistic interpretation of the notion might involve shared beliefs and the common knowledge of shared beliefs).

Finally, supporters of methodological individualism recognize that the social setup can be transformed through the action of individuals. This makes the methodology hardly compatible with the more deterministic versions of historical materialism. Supporters of the methodology also occasionally reject the culturalist tendency to equate social norms with binding traditions or habits, because they find it useful to start from the notion that social norms are effective only through personal commitment.

The admissibility of group-based concepts in individualistic explanations can be illustrated by the status of identities in sociological knowledge. As stated by Laitin (1998, p. 21), "identity groups have an impressive power to give their adherents a sense of natural membership," but individuals have an "equally impressive power […] to reconstruct their social identities." In this dialectic, the transformative power of individuals connects with the influence of collective entities. Moreover, the social identities play a role in mutual expectations, hence also in the individual choices that are conditional on these expectations. Thus, if individuals have a real power to shape and reshape their social (usually multiple) identities, their initiative affects the social framework of their interaction with other individuals.

Individual Motivation

Some individual characteristics should figure in the *explanans* part of scientific explanations in an explicit manner. But which are the relevant ones? The selection should allow the theorist to reconstruct some basic facts about the choices of reasoning agents, and formulate those relationships which are considered useful for explanation or prediction. Both cognitive factors (such as beliefs or probability judgments) and volitional ones (desires or preferences concerning states of affairs) play a decisive role. There is some uncertainty, however, in the determination of the appropriate concepts of beliefs and preferences. In particular, should they be revealed by introspection, or in actual choice?

Behaviorist social scientists insist that they should be inferred from actual choices (as in revealed preference theory in microeconomics), but most social scientists hypothesize in a bolder way. In particular, the Popperian version of methodological individualism emphasizes the crucial importance of a reconstruction of the knowledge available to the agents themselves. Elucidation of this situation-relative knowledge gives the social scientist a clue to the appropriate beliefs and behaviors of the actors (Watkins, 1970; Popper, 1972, 1976).

If we turn to the goals of the actors: should their preferences be confined to (narrowly conceived) self-interest, for example a combination of pleasures across time? This would be empirically unconvincing. It is then tempting to distinguish several types of action: self-interested action and action out of respect for norms or values, for example. But such a contrast is rather uncertain in many empirical situations. Correlatively, it appears difficult to disentangle the economic dimension from other aspects of human action (Demeulenaere, 1996).

Rationality Hypotheses and Equilibrium

Should the social scientist consider any kind of beliefs, desires, and choices? It would seem awkward to leave unused our knowledge about rational action. Several criteria of rational choice can be used. In the general sense, rationality is the disposition to let one's judgment or action be determined by the best existing reasons or argument. Cognitive rationality is ordinarily viewed as deeply different from practical rationality, even though both of them are involved in the 'rational choice' view of the social sciences. One basic requirement is the existence of a complete and transitive ranking of social states of affairs which, combined with beliefs about uncertain events, yields a ranking of actions. More specialized criteria involve specific relationships between beliefs and evaluative judgments, as exemplified by the classical expected utility formula for choice under risk or uncertainty.

It can hardly be assumed that flesh-and-blood individuals are perfect optimizing entities in all circumstances: given the drastic limitations on our information-processing capabilities, the bounded rationality assumption is more realistic (Simon, 1982), and it should be supplemented by a precise analysis of the complex relationships between intentions, beliefs and

expectations, background knowledge, and representations of the social interaction (Roy, 2008). In many models, agents seem to be confined to single-exit situations: individual goals and the characteristics of the situation are the important explanatory inputs (Latsis, 1972). In most real-world cases, however, individual attitudes in the gathering and use of incomplete knowledge are crucial explanatory factors. In Simon's words, a shift is needed from substantive to procedural rationality. This shift in emphasis has been a decisive factor of scientific progress in recent decades, especially for advances in the understanding of institutional or organizational behavior.

This psychological orientation, however, could be (and has been) resisted for the following reason: sociological explanations of aggregate social facts are not supposed to capture all the details of the actions of existing individuals, but only the central tendencies, among which we should undoubtedly rank the tendency to act rationally through the selection of appropriate means in a purposeful manner (Goldthorpe, 1998). It can thus be argued that sociology, as part of empirical science, should neither restrict itself to the elucidation of the logical consequences of the interaction of ideally rational agents, nor try to replicate the psychological functionings of particular agents. As it turns out, sociology often relies on ordinary psychology (Boudon, 2006, ch. 7, argues that this is a valuable part of Tocqueville's heritage in sociology).

Classical sociology exemplifies the tension between the tendency to rely on psychological facts or mechanisms and the appreciation of the specificity of social facts and sociological explanation. A notable example is Gabriel Tarde's notion of 'psychological statistics,' and his determination to analyze decision making in social settings, trying to make use of a direct description (or measurement) of personal characteristics such as desire and belief, and contemplating a generalization of the psychophysical line of inquiry with the aim of developing social explanations. In practice, sociologists can hardly focus on social situations exclusively or psychological processes exclusively. Stylized models of both are required, and an important problem of the micro–macro linkage consists in determining the appropriate kind of stylized description of facts or processes at the microlevel. A complex treatment of psychological processes seems required in some cases, and this provides opportunities for interdisciplinary work. The two imperatives of psychological relevance and situational analysis are complementary in a deep sense, since the rational agent's *adaptive* cognitive strategy in the face of incomplete *knowledge* must depend on the characteristics of his or her incomplete-knowledge situation.

In most social contexts, the reasons for action of distinct individuals are clearly not independent from one another, and the hypothesized relationships between individual preferences, beliefs, and actions assume the form of an equilibrium, that is to say, a pattern of correlated choices of actions among individuals in a given group or society. In economics, the concept mainly refers to a state of equality between supply and demand; other concepts have been developed and applied in modeling. In game theory, an 'equilibrium' is a situation in which individual actions are mutually adjusted in an understandable sense (e.g., as best answers to one another, as in the 'Nash equilibrium' concept). The importance of this kind of approach was neatly brought out in Cournot's analysis of imperfect economic competition. In the definition of equilibrium concepts, the interrelationships between the predicted social facts and individual reasoning are made clear. In some cases, however, the predictive value of equilibrium concepts (such as the 'Nash equilibrium,' in which each agent's action is the best response to the actions of others) presupposes a fairly sophisticated coordination mechanism for individual expectations, or even sociologically or culturally determined common focal points. This suggests that some preexisting 'social' facts, such as cultural conventions, categories of the language, and accepted attitudes, might have independent explanatory value after all (Janssen, 1997).

Individualistic Descriptions and the Uses of Social Theory

Models and theories are used in several ways. How does this plurality affect the significance of methodological individualism and the relevance of individual-based descriptions? The relationship between positive and normative uses of theories should be considered. Of equal importance is the level of description at which theory-based inferences and predictions are formulated.

The Individualistic Point of View and Normative Thinking

Scientific commitment to methodological individualism for explanatory purposes should not be mixed up with the presumption that individuals give no weight to other people in their own choices. In particular, as Popper noted, an individualistic understanding of social life need not lead to underrating the importance or value of altruistic feelings or nonselfish interests generally speaking: "What really matters are human individuals, but I do not take this to mean that it is I who matter very much" (Popper, 1962). Methodological individualism is an invitation to study, from the point of view of the individual, the balance between self-centered interests and the importance given to the interests or values of other people, as well as institutional commitments. This balance is, in itself, an object of inquiry for normative thinking insofar as we can influence the relative weights of these different kinds of interests for practical purposes.

A number of important normative criteria presuppose an individualistic description of social life. For example, the usual criteria associated with freedom and responsibility have no clear significance unless social life is understood with a view to the causal powers of persons, portrayed as autonomous units in decision-making tasks. These units must be thought of as capable of self-consciousness, evaluation, deliberative choice, and action. It thus appears that common normative views about social life or institutions implicitly rely on an individualistic kind of description and explanation with respect to social phenomena and personal capacities.

According to Baechler (2006, ch. 6), the social agent's freedom consists of choice among possible alternatives, insofar as autonomous deliberation takes place, in a way that testifies to rectitude and rightness. This statement sums up a family of

interrelated arguments and theses that have often been put forward in the tradition of methodological individualism in sociology, with the general intention of making it clear that the transformative powers of the individual in social life are ultimately grounded in the individual's capacity for autonomous choice. The latter capacity is best expressed in freedom of choice as expressed in social settings which provide opportunities to give one's consent to collective, interpersonal, or impersonal constraints.

At a fundamental level, some kind of individualistic description of social life is involved in the formulation of many evaluative criteria. In most cases, equality or efficiency criteria can be applied only if we are able to ascertain that definite states of affairs are experienced by distinct individuals. Individualism, here, stands for an 'atomistic' understanding of social life that identifies the separate lives and experiences of the individuals (rather than collective entities) as the *loci* at which things of value are to be found.

The understanding of individual action in a social environment matters for normative issues in the treatment of communities and groups. Cultural and religious communities, for example, are made of individuals. Consequently, the different kinds of involvement of communities in public life are deeply affected by the status enjoyed by individuals: their status in the political structure as well as in the more specific communities. Under liberal rule, we are equipped with rights, and the key issues about liberal political communities today connect with the nature of exercising personal rights in the political community at large. Representations of (possibly oppressive or coercive) interpersonal ties in the communities matter in this respect, with the methodological consequence that the sociological dimension of exercising rights cannot be ignored by normative theory when it addresses dilemmas about the treatment of groups in society at large.

Ballet et al. (2007) have stressed that the capability approach in social science and philosophy recognizes the importance of "connections between people, through their social commitments and their embedding in social institutions" (p. 186). From this viewpoint, it is quite clear that assessing the value of the exercising of rights or freedoms makes it necessary to refer to cultural and ethical standards, such as fairness standards. These authors have further argued that social links, expectations, and commitments account for the role of fairness standards in exercising rights. As a result, due attention to the capacity for justified self-imposed constraint is credited with delivering a 'stronger vision of agency.'

Individualism and the Levels of Description of Complex Social Facts

Several types of explanation might appear relevant and informative at different levels of description. The projected inferential and predictive uses of social theory are thus decisive for the relevance of individual-based descriptions and explanations of social life. As a matter of fact, the very notion of sociology as a separate discipline is sometimes understood in a holistic manner because sociology is supposed to describe and explain complex 'social facts' which cannot be described and explained adequately at the microlevel of individual action and interindividual interaction.

In Durkheim's classic study of suicide, for example, human conduct is understood on the basis of the social characteristics that account for the scientifically relevant aggregate facts. The propensities for suicide can be explained only by referring to the state of social groups, in particular their characteristics with respect to the integration of individuals. The statistical facts about the frequency of individual actions are explained by social properties.

The holistic dimension in Durkheimian sociology lies in a belief in the independent explanatory value of 'macro' notions with respect to observed individual conduct. If we turn to rules and institutions, this means that individual agreements cannot be treated as independent explanatory factors. In the tradition of methodological individualism, however, theorists try to explain rules and institutions by focusing on the circumstances of individual rational choice (as in Olson, 1965; Coleman, 1990).

Social explanations in the spirit of Durkheim can be reinterpreted in individualistic or interactionist terms to some extent, by focusing on the psychological dimension of complex microsociological patterns of interdependencies that aggregate into macro properties such as social integration, cohesiveness of groups, and so on. Given the parallelism between the two levels of description, the choice between an individualistic perspective and a holistic one should be considered in the light of the projected use of theory: do we intend to describe macrophenomena, or do we try to understand microinteractions and their aggregate effects? The relevance of methodological individualism is clearer in the second case.

Correlated descriptions of individual and social situations also matter in action-oriented social research. Following 'softsystems methodology,' for example (in the field of problem-structuring methods for group choice), we are invited to acknowledge that accounts of purposeful action are 'given in terms of an interpretation applied by the speaker' (Checkland, 2001); given the importance of biases and alternative value systems, we should accept a multiplicity of descriptions of purposeful activities. Accordingly, a model of human action that is used analytically for the purposes of group choice should make its underlying *Weltanschauung* explicit, in order to make sure that the descriptions are meaningful.

Individual Cognition and Sociological Knowledge

Reasons and Cognitive Processes

Individuals are reasoning entities: they develop elaborate cognitive structures – even social theories – about their own action and interaction. How should individualistic social science take this into account? Arguably, rational action involves cognitive operations to some extent. This is true of instrumentally rational actions, after the pattern of 'logical actions' in Pareto's sense, since the adequate association of ends and means is possible only on the basis of some representation of the world and some reasoning about it. 'Maximization' refers to the modeling of individual behavior as the constrained optimization of an objective function. An alternative approach consists in the choice of best elements in opportunity sets, given the preferences of the

considered agents. Methodological individualism in the social sciences has led to an extensive use of optimization programs for modeling purposes.

Sometimes, the goal and its causal association with the available means are obvious, so that the cognitive process will not figure prominently in the proposed explanations. In other cases, the interesting part of the explanation is the cognitive process itself. Strategic situations are a clear example. As Roy (2008, sec. 3.3) puts it, rational choice in strategic games is "a matter of maximizing expected payoffs given what the agents expect each other to do"; relying on hypotheses of rationality in games then obliges the scientist to take mutual expectations into account, and this yields an array of game-theoretic 'solution' concepts for strategic situations. Cognition, here, is at the core of social knowledge insofar as game-theoretic 'solutions' are used to predict or explain interactive behavior among human agents.

The explicit treatment of this cognitive dimension appears to be the best response to the critique according to which individualistic models rely on an arbitrarily simplified account of human action as instrumentally rational action. It is useless to deny the importance of instrumental, means-end reasoning in social interaction. But much is to be gained by paying attention to the cognitive processes through which agents frame their choices (Fillieule, 1996). Cognitive sociologists also explore the inner logic of argumentative strategies (Boudon, 1990; Bouvier, 1999). A challenging dimension of the theorist's work is to reconstruct the intrinsically convincing reasons which are intermingled with mistaken beliefs, resulting in seemingly irrational behavior.

Of particular interest are the processes through which individual agents evaluate and adjust their own goals, and select the norms of behavior they are willing to accept as binding. Processes of this kind cannot be left aside in sociological inquiries that rely on a broad notion of rationality, understood as acting on the best reasons and, accordingly, pursuing appropriate goals through effective means (or, more generally, making the best choices in the light of appropriate values and objectives). This was already exemplified in the classical Weberian notion of axiological rationality, now reinterpreted within the framework of theories of cognitive rationality (Boudon, 1995, 1998; Mesure, 1998; Saint-Sernin et al., 1998).

This implies that rationality requirements are also relevant for the actor's goals, beyond the instrumental adequacy of the means to the goal; these requirements may differ from the criteria of means-end or utility-maximizing adequacy (in utilitarian philosophies and in economic theory, 'utility' is the synthetic goal of an agent who seeks to get maximum satisfaction as a result of his or her action, and in the case of choices under risk or uncertainty, 'expected utility' gives insights about the average consequences of risky choices). This methodological choice is not consensual; it is motivated by an effort to gain a better fit between theoretical predictions and observed social facts. For instance, a long-standing puzzle in individualistic political studies is the so-called paradox of not voting: since individual voting in general elections has small disadvantages and no clear causal effect, why do people vote? A particular line of individualistic inquiry starts from the hypothesis that most people are able to understand the basic reasons that support the following general normative statement 'Citizens should vote.'

Cognition and the Levels of Description

Explicit treatment of individual cognition is a valuable goal, and the sociologist's concern with representations is a motive for modeling the variations of social representations among individuals (Balzer, 1993, ch. 7). Useful theories of individual cognition in social science raise questions about how precisely to draw the dividing lines between the building blocks of individualistic sociological explanation: between individual characteristics and natural facts on the one hand, and between individual characteristics and the relationship between them on the other hand. For example, choice theory as social scientists think of it should be a theory of the relevant individual characteristics, such as desires, intentions, beliefs, and so on. The comparative intensity of desire for goods or states of affairs is a proper object of scrutiny. But should the relevant individual characteristics be borrowed (for modeling purposes) from psychological or biological knowledge about the natural facts of feeling and sensation? In addition, individual characteristics which seem to be relevant for understanding individual choice, such as the comparative strength of preference, have sometimes been thought to be best understood in connection with other individual characteristics, such as the association of pleasure and pain (and their degrees) with goods or states of affairs (Hands, 2006, has analyzed the strategic situation in such debates of the doctrine of cardinalism – a view of welfare or utility according to which these notions can be given quantitative meaning in the sense that ratios between utility (or welfare) differences are considered meaningful statements).

At some level of description, individual action is made of a series of natural (neural and external) states of affairs. It thus appears that individual 'good reasons' can be screened off by the natural causes of behavior, which are a proper object of inquiry for natural science, and methodological individualism in its classical versions might appear misleading insofar as it treats individual good reasons as the ultimate explanations of action. Psychological processes are implemented in the human brain by neuronal functionings, but this does not imply that the 'ultimate' explanation of social phenomena is to be found at the neurophysiological level or at the molecular level.

Science is constructed in layers, in such a manner that it includes both higher level theories which work at the macro-level appropriate to the macrophenomena to be explained (such as mainstream sociological theories) and other theories which show how the higher level objects and relations can be explained through scientific work at the level below (Simon, 1997). The hierarchical relationship between two classes of phenomena does not automatically imply that explanation should proceed from the lower level to the higher level.

At a certain level of description, individual action is the sequel of an intention which can be described only in a certain language, using definite mental categories. Arguably, language, categories, and meanings are social objects. Hence, it might appear appropriate to consider human actions as understandable only on the basis of preexisting rules, or

'institutions' in Mauss's sense, which social actors find before them as preexisting facts. Indeed, social norms have a constitutive role with respect to the meaning of actions – undoubtedly an important feature of many social situations. This is not automatically alien to individualistic explanations, if we allow for psychological motives which depend on complex, holistic representations of social interaction, along the lines of institutional individualism (Agassi, 1975). As a matter of fact, most social rules are not just given as 'facts' but also are agreed upon, or rejected, by individuals.

Individualistic explanatory and predictive strategies in sociology rely on a sharp distinction between the choice problems of individual actors, on the one hand, and social institutions, regularities, and norms, on the other hand. On this account, the emergence and stability of social regularities, social norms, and institutions should be explained on the basis of the underlying individual reasons. This creates opportunities for the use of both classical rational-choice models and elaborate theories of individual cognition.

See also: Action, Theories of Social; Collective Action; Collective Behavior, Social Psychology of; Collective Behavior, Sociology of; Collective Beliefs, Sociological Explanation of; Rational Choice Explanation: Philosophical Aspects; Rationality in Society; Scientific Explanation; Social Cognition; Social and Economic Change: Psychological Challenges for Individuals; Sociological Traditions; Sociology, Epistemology of.

Bibliography

Agassi, J., 1975. Institutional individualism. British Journal of Sociology 26 (2), 144–155.
Baechler, J., 2006. Les Fins dernières. Eléments d'éthique et d'éthologie humaine. Hermann, Paris.
Ballet, J., Dubois, J.-L., Mahieu, F.-R., 2007. Responsibility for each other's freedom: agency as the source of collective capability. Journal of Human Development 8 (2), 185–201.
Balzer, W., 1993. Soziale institutionen. Walter de Gruyter, Berlin and New York.
Boudon, R., 1990. L'art de se persuader. Fayard, Paris (1994. The Art of Self Persuasion Cambridge, Polity Press.)
Boudon, R., 1995. Le juste et le vrai. Fayard, Paris.
Boudon, R., 1998. Social mechanisms without black boxes. In: Hedström, P., Swedberg, R. (Eds.), Social Mechanisms: An Analytical Approach to Social Theory. Cambridge University Press, Cambridge.
Boudon, R., 2000. Etudes sur les sociologues classiques, II. Presses Universitaires de France, Paris.
Boudon, R., 2006. Tocqueville for Today. Bardwell Press, Oxford.
Bouvier, A., 1999. Philosophie des sciences sociales. Presses Universitaires de France, Paris.
Boyer, A., 1992. L'explication en histoire. Presses Universitaires de Lille, Lille.
Checkland, P., 2001. Soft systems methodology. In: Rosenhead, J., Mingers, J. (Eds.), Rational Analysis for a Problematic World Revisited. John Wiley and Sons, Chichester Chapter 4.
Cherkaoui, M., 2006. Le Paradoxe des conséquences. Droz, Geneva & Paris. (Hamilton, P., 2007. Good Intentions. Max Weber and the Paradox of Unintended Consequences. Bardwell Press, Oxford.)
Coleman, J.S., 1990. Foundations of Social Theory. Belknap Press, Cambridge, MA.
Demeulenaere, P., 1996. Homo oeconomicus. Presses Universitaires de France, Paris.
Demeulenaere, P., 2003. Les Normes sociales. Presses Universitaires de France, Paris.
Fillieule, R., 1996. Frames, inferences, and rationality: some light on the controversies about rationality. Rationality and Society 8 (2), 151–165.
Goldthorpe, J., 1998. Rational action theory for sociology. British Journal of Sociology 49 (2), 167–192.
Hands, D.W., 2006. Individual psychology, rational choice, and demand: some remarks on three recent studies. Revue de Philosophie Economique/Review of Economic Philosophy 13 (1), 3–48.
Janssen, M., 1997. Individualism and equilibrium coordination in games. In: Backhouse, R., Hausman, D., Mäki, U. (Eds.), Economics and Methodology: Crossing Boundaries. MacMillan, London.
Laitin, D.D., 1998. Identity in Formation. Cornell University Press, Ithaca and London.
Latsis, S.J., 1972. Situational determinism in economics. The British Journal for the Philosophy of Science 23 (3), 207–245.
Mantzavinos, C., 2008. Individus, institutions et marchés. Presses Universitaires de France, Paris.
Mesure, S., 1998. La Rationalité des Valeurs. Presses Universitaires de France, Paris.
Olson Jr., M., 1965/1971. The Logic of Collective Action, second ed. Harvard University Press, Cambridge.
Opp, K.D., 2001. How do norms emerge? An outline of a theory. In: Boudon, R., Demeulenaere, P., Viale, R. (Eds.), L'explication des normes sociales. Presses Universitaires de France, Paris.
Pareto, V., 1905. L'individuel et le social. In: Claparède, E. (Ed.), Actes du congrès de philosophie de Genève. Kündig, Geneva. Reprinted in: V. Pareto. Mythes et ideologies. Busino, G. (Ed.). Droz, Geneva (1966).
Popper, K.R., 1962. The Open Society and Its Enemies. Routledge, London.
Popper, K.R., 1972. Objective Knowledge. Clarendon Press, Oxford.
Popper, K.R., 1976. The logic of the social sciences. In: Adorno, T.W., et al. (Eds.), The Positivist Dispute in German Sociology. Heinemann, London.
Roy, O., 2008. Thinking before Acting. ILLC (Institute for Logic, Language and Computation) Publications, Amsterdam.
Saint-Sernin, B., Picavet, E., Fillieule, R., Demeulenaere, P. (Eds.), 1998. Les modèles de l'action. Presses Universitaires de France, Paris.
Simmel, G., 1900. Philosophie des Geldes. Duncker & Humblot, Leipzig (1978. The Philosophy of Money. The Free Press, New York.)
Simon, H., 1982. Models of Bounded Rationality. The MIT Press, Cambridge, MA.
Simon, H.A., 1997. An Empirically Based Microeconomics. Cambridge University Press, Cambridge.
Watkins, J.W.N., 1970. Imperfect rationality. In: Borger, R., Cioffi, F. (Eds.), Explanation in the Behavioral Sciences. Cambridge University Press, Cambridge.
Weber, M., 1922. Wirtschaft und Gesellschaft. Mohr, Tübingen (1968. Economy and Society. University of California Press, Berkeley.)

Methodological Individualism: Philosophical Aspects

Pierre Demeulenaere, University of Paris-Sorbonne, Paris, France

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous article by G. Currie, volume 5, pp. 9755–9760, © 2001, Elsevier Ltd.

Abstract

Methodological individualism is a central tradition in the social sciences. The article argues that it corresponds to one main thesis, which is that only individuals act in the social life. The article examines the consequences that can be deduced from this main proposition regarding the theory of behavior and explanation in the social sciences, and also examines the misunderstandings to which it frequently leads.

Introduction

Methodological individualism (MI) is a central tradition in the social sciences. It is a theory about human agency in a social world and about the articulation of the various entities constituting such a world. It is consequently a theory of explanation of social phenomena.

The main idea that can characterize MI is that, in social life, so-called social entities do not act by themselves, only individuals (i.e., human beings) do so. Philosophers and social scientists such as John Stuart Mill, Carl Menger, and Max Weber initially deployed this basic idea, although none of them directly used the term 'methodological individualism.' However, a close expression can be found in Weber's analysis. Schumpeter coined the term. Popper and his followers, notably Watkins, developed the notion in a more philosophical way. Many other singular theorists have been associated with the movement. At the discipline level, economics is generally associated with MI, whereas sociology is more divided regarding this issue, with the intervention of holism or methodological holism.

Because only individuals can act, methodological individualists consider that social entities (like states, institutions, organizations) do not act by themselves but need individuals to be set in motion. This rather simple idea is, however, compatible with two other major propositions. First, institutions and social entities can, under certain conditions, be assimilated to actors. However, in the tradition of MI, it remains that such corporate actors need individuals to be set in motion. Second, social entities can have an 'influence' on individual behaviors, although this notion of influence can be interpreted in various ways. It depends mainly on the nature of the social entities that are taken into consideration. In this article, it is not possible to present the details of each author's positions. Therefore I will try to present the logical consequences of the main idea developed by MI, that only individuals act.

If we set aside the difficult problem of animal behavior, which I will not address here, it can be said that in social life only individuals act. This is clearly stated by Max Weber:

> Action in the sense of subjectively understandable orientation of behavior exists only as the behavior of one or more *individual* human beings. (Weber, 1978: p. 13)

> For still other cognitive purposes – for instance, juristic ones – or for practical ends, it may on the other hand be convenient or even indispensable to treat social collectivities, such as states, associations, business corporations, foundations, as if they were individuals persons. Thus they may be treated as the subjects of rights and duties or as the performers of legally significant actions. But for the subjective interpretation of action in sociological work these collectivities must be treated as solely the resultants and modes of organization of the particular acts of individual persons, since these alone can be treated as agents in a course of subjectively understandable action. (Weber, 1978: p. 13)

This is the fundamental proposition of MI. It means that there is an active social life only because there are individuals who act. It is a descriptive proposition about what is responsible for action in social life. It is not directly an ontological proposition about the existence of individuals or social facts, nor is it a position about explanatory possibilities in the social sciences, although this fundamental proposition has important consequences regarding those matters.

Once said that only individuals act, seven questions derive from this elementary proposition:

1. What does an individual action mean exactly? Does it imply that individuals are separated from each other and do not interact?
2. Does this mean that they are supposed to act in a certain way (a selfish or individualistic way, for instance)?
3. In particular, are individuals supposed to be rational?
4. Does this imply that they are not influenced by institutions and by social settings in the course of their action?
5. Does this exclude an infra-individual description of action?
6. Does the assertion that only individuals act imply that only individuals exist?
7. Does the assertion that only individuals act imply that any explanation of the macro level should be only based on the micro level?

Although MI is frequently associated with positive answers to these questions, I will argue that this should not be the case, both from a logical viewpoint derived from the main MI proposition and from a historical perspective based on the writings of the preeminent authors influential in this tradition.

Interaction

Stressing the uniqueness of individuals as the sole possible actors in the social world does not, however, imply any kind of 'individualistic' position, as Weber clearly insisted on. MI is frequently associated with neoclassical economics in which social outcomes (like market price equilibrium) are derived from separate individual decisions. In this scenario, individuals are separated from each other since, in the model, the formation of price equilibrium does not require their formal agreement. Moreover, in the economic tradition, individuals not only do not cooperate and do not converse in order to produce equilibrium prices, but are selfish: they try to achieve a maximization of their personal utility. These hypotheses are at the core of economic hypotheses. They can be found equally in the theory of games, where actors similarly do not cooperate, in the sense that they do not discuss together the situation and the way to deal with it, and do not initially have a cooperative attitude. On the basis of such behaviors, specific social outcomes can be seen to emerge under certain conditions: for example, price equilibrium or social norms.

However, those two dimensions (absence of direct interaction and selfishness) are not intrinsically necessary to methodological individualism as such, either in theory or in practice. The idea that only individuals act has no logical implication on the nature of their relationships: they can indeed interact and deliberately cooperate and tend to produce social outcomes. They can also have altruistic behaviors. Sociologists have frequently negatively considered methodological individualism as an 'atomism' involving separate actors, who are moreover selfish. But this is not necessarily the case. It is true that Menger (1996) referred to atomism, insisting for instance that money, when it emerges as a social institution, does not need to be the result of an agreement between actors. It derives as an unintended consequence from separate 'atomistic' decisions.

Generally speaking, it can be useful, sufficient, or relevant in specific social situations to suppose that actors act separately and do not cooperate. This is the case, for example, in the stock exchange, or more generally in the ideally conceived markets of neoclassical economics. It can be shown, moreover, that in certain theoretical types of situations (particularly prisoner's dilemma situations), separate selfish individuals can be led to adopt cooperative norms in order to satisfy their interests. But this does not mean that all norms derive from such situations and such behaviors. This type of description corresponds to the fact that in many social situations, individuals in effect do act separately, and moreover in a selfish way. On the basis of such descriptions, we can have the elements from which to derive social outcomes.

But the idea that only individuals act does certainly not imply that it is always the case that individuals do not interact or act in a selfish manner. Interaction can be found in economics, based on selfish motives, when for instance actors set up monopolies or oligopolies. This is clearly the result of an interaction among individuals. Therefore, the specific economic tradition that displays separate actors is not to be conflated (1) with the whole of economic analysis and (2) with MI.

Selfishness and Altruism

As has already been suggested, the idea that only individuals act does not imply that they act in a selfish manner. When J.S. Mill (1988), against the tradition of Comte, insisted on the importance of social phenomena to be based on the action of individuals, and stressed the importance, for economics, to suppose a selfish behavior, defining the *Homo economicus* figure (although he did not use the term), he certainly did not conclude that all men in all circumstances acted in a selfish manner aimed at increasing wealth. On the contrary, he clearly stressed the possibility of moral and altruistic behaviors.

However, MI is frequently reduced to an economic analysis of selfish behavior. Once again this cannot be derived from the core idea that only individuals act. Moreover, the very idea of selfishness and of satisfaction of interests is relatively ambiguous in the economic tradition of action. On the one hand, starting with Mill, the figure of a narrowly defined economic man pursuing the increase of his wealth is designed in order to analyze economic markets. But on the other hand, very soon, the complexity of interests, the variety of preferences, and the fact that the notion of utility is a fundamentally vague one became preponderant in economic analysis: utility depends on various psychological patterns on the one hand and on social norms on the other. Therefore, Pareto first introduced the notion of ophelimity (utility based on individuals' choices), and afterward utility in economics became a formal notion derived from preferences. Preferences are diverse and can be linked to social norms. For instance, smoking or taking drugs can be seen favorably or unfavorably on the basis of personal tastes or social prescriptions. There is, therefore, no intrinsic notion of cost or benefit that is at the bottom of the decision. The decision is itself what allows us to speak of costs or benefits, and they can be diverse for different persons: for one it is costly to take drugs (e.g., to drink alcohol) while for another it is costly not to take them. Therefore, in the domain of economics, the theory of action linked to MI is fundamentally ambiguous: it rests on the idea that actors tend to maximize their benefits or minimize their costs, but because costs and benefits do not represent an unambiguous signification, the theory of action is not really predictive of effective or definite actions and choices (Sen, 1977).

Therefore, the distance between an economic model of social behavior and a sociological model of behavior is strongly reduced (Akerlof and Kranton, 2010). Sociologists have generally attacked methodological individualism on two grounds: (1) the idea that individuals are not selfish persons acting on the basis of a pure calculus of selfish interests and (2) that individuals do not act in a social vacuum but depend on social norms and institutions. This second aspect will be more fully developed next in this article. But it should be clearly noted that those two critiques do not attain the core of the idea of methodological individualism; moreover, they do not attain the core of the economic version itself of methodological individualism, because interests, preferences, and choices can rest on social norms: therefore, social aspects can be found in the very heart of economic decisions, and the question is whether there is or not a theory of decision devoid of any social aspect, in particular of social norms.

Rationality of Action

This leads to another related question: should the theory of individual decisions be taken from a naturalistic stance, where the elements of the decision precede social dimensions, or should it be considered as inherently social, in the sense that there would be no natural, stable, and transcultural regularities in the determinants of the choices beyond the observed social and cultural variety in behaviors? This question is related to the scope and significance of rationality, which can be seen as a natural feature of the human species. However, the idea of rationality is a multisided one, and can be related to each of theses three positions. The idea that only individuals act has by itself no logical consequence regarding the question of cultural and social variation. Three positions are compatible with the core idea: one that would stress that all behaviors depend on a natural (namely psychological) basis, rational or not; another that would insist on the cultural dimension of all individual behaviors, or the social position of the actors; and finally, one that would consider that the two dimensions are always interconnected: the natural features of the human species lead to a diversity of social and cultural situations; this diversity, though real and decisive for individual decisions, is not, however, disconnected from the general natural traits of human behavior, and should tentatively be interpreted in their terms. It is likely that this third option is the right one (Haidt, 2012), and it is linked either to a theory of rationality or to unconscious motives.

Thus the debate about the rationality of action intervenes in the more general frame of the characterization of action and behavior. Methodological individualism is largely linked to (and criticized for) the idea that individuals act in a rational way. But here again, there is no necessary link between the idea of individual action and of rational action. It is perfectly understandable that individuals can act irrationally (Elster, 2007). However, this depends on the notion of rationality we have and on its status. There are five basic notions of rationality in the social sciences that can be classically linked to methodological individualism.

The first is to consider that individuals are rational whenever they act in an intentional way. The intentional decision to act corresponds to the reason someone has to act. An irrational behavior would then be to act on the basis of unintentional causes. For instance, Kahneman (2011) reports a study about parole judges observed in their decisions. Individual judges tended to be more severe when they were hungrier: clearly in this case unintentional and unconscious causes affect intentional decisions. This means that intentional motives do not have the sole influence in individual decisions; however, it is still individual decisions that are at stake, partly determined by unconscious trends.

The second notion of rationality is linked to the notion of consistency or transitivity of choices. The rationality is not related to the contents of individual choices as such, but to the fact that individuals are consistent in their preferences. This is the official position of neoclassical economics, where rationality does not involve any kind of 'right' decision but only the fact that people do not make incoherent decisions. Irrationality here would be to have inconsistent choices. But again, irrationality regarding consistency of choices would affect individual behaviors and thereby would be compatible with methodological individualism.

The third notion of rationality links it to an instrumental dimension of action, where an actor chooses the adequate means to reach an end. The rationality derives from an ability to select the correct means, which requires that a normative notion of correctness be implied. It should be noted that although instrumental rationality is widely associated with standard economics, it can be separated from it from two opposed points of view: first, as was noted by Weber, instrumental action does not necessarily imply any kind of economic motives. Second, the transitivity of preferences, which is the core of neoclassical theorization of economic action, does not in fact involve a notion of instrumental action.

The fourth notion of rationality is clearly the most widespread in the social sciences, although it suffers major ambiguities (Demeulenaere, 1996) and can be the subject of deep criticism. It is the idea that people act rationally whenever they efficiently pursue their self-interest (Coleman, 1990). This is a much narrower conception of rationality, which does not insist only on the means but on the ends of action: self-interest. In this scenario there is, however, a hesitation between two alternatives: either the preferences are supposed to be the same (like a preference for an increase of wealth) or they can be dissimilar (like smokers' and nonsmokers' preferences). This theory is strongly normative, because it reduces rationality to self-interest: this can be refuted from a philosophical point of view. However, here again, if it appears that actors are not rational in this sense – that is, if they do not behave out of strictly selfish motivations – this does not imply that it is not individual actions that are under consideration.

The fifth notion of rationality is much rarer in the social sciences, as a consequence of the influence of positivism: it extends its scope to the determination of right choices that can include not only correct beliefs but also normative deliberations. Boudon (2011) has advocated the reintroduction in sociology of this classical meaning of rationality in philosophy.

In sum, whatever notion of rationality is selected, an irrational behavior is by definition possible (otherwise it would be pointless to determine the conditions of a rational action); but there is no reason to consider that irrational actions are not individual actions, or at least individual behaviors. Therefore, rational behavior is not consubstantial to MI, even if there are reasons to consider that a rational interpretation of behavior should have priority over an irrational one, for one main reason, clearly expressed by Weber: when a rational behavior is possible, it can be expected that an individual will adopt it, because it is part of the normative idea of rationality that someone should adopt the correct rational behavior, when it is available. For instance, one should think that two and two make four, and not five. Therefore, if one adopts an irrational behavior, we need a further investigation to understand why it is the case. Naturally, from a more general point of view, we need a theory to explain why the human species is endowed with such a notion of rationality, with its normative pressure, and what are its contents and properties. We need to explain rationality, as much as we need to explain irrationality at a meta-level. But, for social science explanations, on the basis of a given theory of rationality (whatever the origins of rationality), a reference to rational behavior does not raise

supplementary questions, whereas the evidence of an irrational behavior raises the question of why the rational choice was not selected, when available. This is not intentional fundamentalism (Ylikoski, 2012) but a consequence of the normative dimension of rationality.

The Role of Situation and Institutions

From its beginnings MI has been attacked with the charge of 'atomism,' by Durkheim for one, on the ground that in this scientific tradition not only individuals do not interact but they are completely free of any social determination or influence. This critique cannot be accepted for two reasons. First, all major MI sociologists, as well as philosophers like Popper (1966) or Watkins (1957), have accepted the importance of social situations as necessary to the interpretation of individual behaviors. This is the case, for example, of Homans (1967). Thus when Wippler (1978) coined the term 'structural individualism,' to stress the role of social situations in individual behavior, it was largely accepted, for instance, by Boudon and Bourricaud (1990). The important thing is not to deny such an influence, but to understand the way it is exerted on actors, which occurs by various and heterogeneous ways. Therefore, it is not correct to describe MI as a refusal of the role of social situations on individual behavior, as Lars Udéhn (2001) describes it. Second, the alternative is not between socialized actors unilaterally depending on social structures, and nonsocialized individuals deprived of any kind of interpersonal relations and independent of any social influence. This is an untenable caricature. In the tradition of MI, people belong to social situations that exert influence on them: this is clearly the case for instance in Coleman's famous theorization of social action (Coleman's boat), where individual actions are understood only on the basis of their social environment.

However, four things must be added in this respect:

- In the tradition of MI, social factors exert an influence on individual actors, but individual actors are the ones who make the decisions, limited and constrained by these social influences. This means that social factors do not act by themselves, but individuals respond to and are constrained by social factors. It is, therefore, important to clarify and explicate the way in which social factors play a role in individual decisions.
- This means that we need to interpret social dimensions on the basis of nonsocial dimensions of individual behaviors.
- There is always a certain room for decisions, even in very constrained situations. It can be the case, however, that there is some kind of 'internalization' of norms or habits: this idea is accepted, for example, by Coleman (1990). It implies specific mechanisms that bridge social settings and individual behavior through psychological (and neuronal) adaptations, which are not necessarily clear and well understood. For example, people raised in a specific language tend to have the accent of this language when they speak another one.
- However, a specific feature of MI is to try to interpret the emergence of social institutions and social situations on the basis of prior decisions and individual moves. For example, the design of a specific keyboard (of the QWERTY type) makes all people constrained by the way it is to be used, although another layout could be more efficient. But we also know the reasons for which such a design was selected for specific purposes.

Infra-Individual Determination

The idea that only individuals act has been deployed by methodological individualists against holists or structuralists who evoked the action of social wholes. Today, it would be reaffirmed against theorists who claim the existence of a social causation that does not involve individual actors (Kincaid, 1996). MI has been challenged in return, more recently, by theorists who stress the importance of infra-individual determination of behavior, either at a psychological level or at a neurophysical level (Sperber, 2011). They underscore the existence of unconscious motives, and interpret their role as the necessity to go to a deeper level of determination of action, which is therefore termed 'infra-individual.' Consequently, there are social phenomena, based on unconscious motives, such as the diffusion of beliefs, that do not depend on individual deliberations.

There are two different things involved in this argumentation. First, as has been already said, there is no reason that MI should refuse the existence of unconscious motives that partly determine individual behavior (which is therefore partly intentional, partly unintentional). To acknowledge their existence does not lead to a refusal of the basic idea that only individuals act, even if this raises the complex problem of the relations between the conscious and intentional dimensions of behavior on one side and the unconscious and unintentional ones on the other. Therefore, whenever diffusion phenomena exist that do not depend on conscious choices, this does not mean that it is not through the means of active individuals (with their psychological characteristics) that those phenomena intervene. It would be meaningless to refer to beliefs without referring to individuals who support these beliefs and whose psychological mechanisms are responsible for their diffusion.

Secondly, unconscious motives and subintentional determination of behaviors are frequently associated with an 'inferior' level, which should be taken as responsible for actions, preferably at the individual level. But this is a misleading argument. It would rest on the idea that the inferior level has no role in the determination of intentional behaviors themselves. But there is no basis for this. If one accepts (as promoters of a subintentional determination of action do) that unconscious behaviors are determined by psychological or physical mechanisms that lie beyond the scope of individual awareness, it can be said – and it is in fact necessary to say – that conscious individual decisions and deliberations also rest on such subintentional mechanisms (that determine the strength of rationality). They are equally concerned by the mind–body problem. Therefore, a reference to psychological or unconscious motives does not straightforwardly lead to a subintentional level, because rationality itself necessarily rests on psychological and physical determinants. The infra-individual

level concerns rationality as well as irrationality: but to discover the biological mechanisms that are involved whenever a behavior is engaged, rational or irrational, conscious or unconscious, is not the task of the social sciences as such.

The Micro/Macro Transition and the Ontological Issues

So far, following Weber, I have described MI as a theory asserting that, in social life, only individuals act. This is not an ontological position about the existence of individuals or of social facts. However, it poses the problem of the relation between individuals and social entities, classically named the 'micro/macro transition' (Little, 1998). It involves the determination of what the social entities are and the nature of their link to individuals, that is, the issue of so-called 'levels' in the geography of social life. Those problems are related to ontological questions about what is considered to exist in social life.

When we refer to the micro/macro relation, two positions are possible. One is very general, and will consider that the micro-elements are the parts of any set that includes them. For instance, administrative entities are usually hierarchically dependent on other superior-level administrative entities (cities, regions, states, supranational organizations). Each level can be named and treated as a quasi-actor: it can be said, for example, that Russia and the United States are players on the global scene. Therefore, micro and macro entities do not refer to specific levels, but to positional relations: for any macro level, there is a micro level.

However, two caveats should be added to this first position. First, because Russia and the United States, for example, are stable, long-term existent countries, they clearly have properties that are stable beyond individual decisions. For instance, they have a size that does not depend on actual individual decisions (and that exerts constant constraints on the succession of people who are in charge of these countries). However, even if it is not meaningless to speak of a Russian or an American strategy, in the long run, beyond actual individual decisions, due to the existence of those lasting properties, in fact any interpretation of actual decisions by one of the states necessarily involves singular individuals (presidents, ministers, and so on). Quasi-actors are not actors in the same way that individuals are. Quasi-actors need singular individuals to be active, so that there is a privileged level in order to achieve sound interpretations of the behavior of quasi-actors.

Second, it would not be correct to describe a relation of 'emergence' between the various levels: states are not emergent regarding cities, although it can be said, from an administrative point of view, that cities belong to a state and are parts of it. Moreover, the same elements can be part of various social sets that do not overlap. For instance, an individual can be the member of a football club, and the club can be part of a league in which countries as such play no role. The football player can be also the member of a club that belongs to a country that is not his own. There is no such thing in social life as a unique hierarchy of levels where superior levels emerge from inferior ones. Therefore, the analogy with the infra-individual level and the individual level is clearly misleading: it is not at all the same type of relation, because it cannot be said that the social emerges from the individual in the same sense (however problematic it is) as the mind would emerge from its body determinants.

This is why, although it is possible to conceive the micro/macro relation as a purely positional relation, where no specific level is privileged, and where institutions can be treated as quasi-actors, it is still important for MI to reaffirm that quasi-actors necessarily depend on singular actors to be active. And this is the second and more important meaning of the micro/macro relation: macro-level institutions, organizations, or rules are necessarily referred to singular actors (embedded in social situations) to be understood.

What are then the social facts? There are many words that are commonly used, particularly 'collective,' 'structural,' and 'system' level. The whole/part is equally associated with this micro/macro relation. All of this is relatively imprecise and refers to various realities that pose different ontological problems.

I would suggest (Demeulenaere, 2011) that there are basically three types of social facts regarding an individual level.

First is the distributional properties of an aggregate, such as a mean. For example, we can select at random the people who enter a shop and determine their average age. In this case, it can be said that there is a whole, because we select a definite number of persons, but there is certainly no such thing as an emergent whole regarding the parts. The constructed measure of the mean does not correspond to the properties of a whole as such, in the same sense as the age of one person is a property of this person (although this property is by no way a physical property, because it involves various institutional considerations about the appropriate way of measuring the age). The notion of aggregation can be here misleading. If we add up a series of stones, we have an aggregate, which has specific properties (for instance, it might have a pyramidal shape). The aggregate has emergent properties (height, length, and so on) that the individual stones do not have. As opposed to this, when we measure the average size of the stones, we do not refer to this whole as such but to the individual stones taken together. So the aggregate in this case can correspond to two different things.

A second meaning of the macro dimension corresponds to the relative positions of a series of elements: for example, one individual is taller than another, or richer, or more powerful. Here again, it could be possible to speak of a whole, or of a level, but these are metaphors that have no substantial basis: there is no emergence of a specific whole or of a level beyond the set of elements that are envisaged from the point of view of their relative positions.

Finally, common rules that are followed (or rejected) by individuals represent a macro dimension regarding elements involved. These rules can include norms and institutions. Sometimes norms and institutions give birth to quasi-actors, which is not the case in the two previous cases. In this case, it is easier to speak of a group level, because people can have a sense of belonging to the whole. This is why Searle's social ontology has focused mainly on this type of situation. However, the notion of a 'superior' level is ambiguous here, because rules, norms, institutions, and organizations depend

on actors for their life. They can be said to emerge as a social novelty, but they can disappear if nobody follows them.

All of those dimensions pose various ontological problems, which will not be discussed here. The idea that only individuals act has no direct implication regarding the ontology of social facts, because they are clearly existent things that do not act. The main position of MI is one about the fact that only individuals act, not about the existence of social facts. The ontological problems depend on the notion of what an existent being is, and various positions will be possible regarding this. An organization can be said to exist even if it does not have the physical properties of an individual. A measure exists, even if the mean to which it refers does not exist from a physical point of view. The debate is not so much about whether social facts exist or not (it would be absurd to consider, for example, that the United States does not exist) but what type of social facts exist, and what is their mode of existence.

Explanation

The main proposition of MI concerning the fact that only individuals act has consequences for the relevant explanatory strategy in the social sciences. Globally, it is said that social facts should be explained by reference to individuals who set in motion social life. In particular, correlations between social variables (for instance, gender and academic success) have to be explained by individual actions and decisions in a given context (Boudon, 1974).

But the actual requirements for relevant explanations depend on the nature of the social facts considered and of their relation to individuals. The relation that exists between a series of individuals and a collective measure is not the one that exists between a set of individuals and the norms they adopt or refuse. In the first case there is no social novelty to be explained, and the collective measure logically derives from the individual measures. In the second, opposite case, norms represent a social novelty regarding individuals, and are to be explained genetically by their beliefs and decisions in a given situation.

The consequence of this is that although the requirement of MI is to insist on individual actions, because they are embedded in a social context that plays a role in their decisions, we cannot say that the macro level is explained by a micro level: since a specific social context is necessary to interpret the individual decisions, one dimension of the macro level is necessarily involved in the explanation of another dimension. However, a given context, locally admitted for the purposes of a definite explanation, can be, from another point of view, the focus of explanatory efforts.

Therefore, two positions should be rejected: one that would say that MI requires referring only to a micro level in order to explain macro-level phenomena, because any explanation actually involves macro dimensions. And second, it would be incorrect to accuse MI of intentional fundamentalism, because any explanation involves a reference to social settings and can include unconscious dimensions of behavior.

See also: Atomism and Holism: Philosophical Aspects; Individualism versus Collectivism: Philosophical Aspects; Intentionality and Rationality; Rational Choice Explanation: Philosophical Aspects; Scientific Explanation; Social Properties (Facts and Entities): Philosophical Aspects.

Bibliography

Akerlof, G.A., Kranton, Rachel E., 2010. Identity Economics. How Our Identities Shape Our Work, Wages and Well-being. Princeton University Press, Princeton and Oxford.

Boudon, R., 1974. Education, Opportunity and Social Inequality. Wiley, New York.

Boudon, R., 2011. Ordinary rationality: the core of analytical sociology. In: Demeulenaere, P. (Ed.), Analytical Sociology and Social Mechanisms. Cambridge University Press, Cambridge, pp. 33–49.

Boudon, R., Bourricaud, F., 1990. A Critical Dictionary of Sociology. University of Chicago Press, Chicago, IL.

Coleman, J.S., 1990. Foundations of Social Theory. The Belknap Press of Harvard University Press, Cambridge and Londres.

Demeulenaere, P., 1996. Homo oeconomicus, Enquête sur la constitution d'un paradigme. P.U.F., Paris. Réédition "Quadrige", 2003.

Demeulenaere, P., 2011. Introduction. In: Demeulenaere, P. (Ed.), Analytical Sociology and Social Mechanisms. Cambridge University Press, Cambridge, pp. 1–30.

Elster, J., 2007. Explaining Social Behaviour. More Nuts and Bolts for the Social Sciences. Cambridge University Press, Cambridge.

Haidt, J., 2012. The Righteous Mind. Why Good People Are Divided by Politics and Religion. Vintage Books, New York.

Homans, G.C., 1967. The Nature of Social Science. Harcourt, Brace & World, New York.

Kahneman, D., 2011. Thinking Fast and Slow. Farar, Straus and Giroux, New York.

Kincaid, H., 1996. Philosophical Foundations of the Social Sciences: Analysing Controversies in Social Research. Cambridge University Press, Cambridge.

Little, D., 1998. Microfoundations, Method and Causation: On the Philosophy of the Social Sciences. Transaction Publishers, New Brunswick, NJ.

Menger, C., 1996 [1883]. Investigations into the Method of the Social Sciences. Libertarian Press, Grove City.

Mill, J.S., 1988 [1843]. The Logic of the Moral Sciences. Open Court, La Salle.

Popper, K.R., 1966. The Open Society and Its Enemies. Princeton University Press, Princeton.

Sen, A., 1977. Rational fools: a critique of the behavioral foundations of economic theory. Philosophy and Public Affairs 6 (4), 317–344.

Sperber, D., 2011. A naturalistic ontology for mechanistic explanations in the social sciences. In: Demeulenaere, P. (Ed.), Analytical Sociology and Social Mechanisms. Cambridge University Press, Cambridge, pp. 64–77.

Udéhn, L., 2001. Methodological Individualism: Background, History and Meaning. Routledge, London.

Watkins, J.W.N., 1957. Historical explanation in the social sciences. British Journal for the Philosophy of Science 8, 104–117. Reprinted in Gardiner, P. (Ed.), 1959. Theories of History. The Free Press, New York.

Weber, M., 1978 [1922]. Economy and Society. University of California Press, Berkeley, CA.

Wippler, R., 1978. The structural-individualistic approach in Dutch sociology. The Netherlands Journal of Sociology 14, 135–155.

Ylikoski, P., 2012. Micro, macro and mechanisms. In: Kincaid, H. (Ed.), The Oxford Handbook of Philosophy of Social Science. Oxford University Press, Oxford, pp. 21–45.

Methodological Issues in the Study of Sexuality

Stuart Michaels, Academic Research Centers, NORC at the University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social scientific research on sexuality has grown enormously in recent decades primarily driven by demographic and health concerns. Focusing on the problems of sample-based surveys, the methodological challenges related to sampling (generalizability) and interviewing (measurement error) are discussed. The design of sex research must begin with the clear conceptualization and operationalization of sexuality. Gaining participation by selected participants and minimizing bias and error in response through the formulation of questions and maximizing privacy of response are emphasized.

Introduction

Social scientific research on sexuality including surveys has experienced an enormous increase in recent decades. Many factors have contributed to this increase, among them the 'sexual revolution' of the 1960s, the social movements such as Feminism and what is now called the LGBT (lesbian, gay, bisexual, and transgender) movement, as well as the HIV/AIDS epidemic. The social and historical context affects all scientific research shaping and motivating research questions that address pressing social concerns and problems. This is especially true in the area of sexuality where legitimating research is especially important to overcome interpersonal, moral, and political reticence and taboos about asking and answering questions that touch on the private and personal (Ericksen, 2001).

Definition of Sexuality as the Object of Research

While our main concern is methodological, that is how our research decisions and practices affect the generalizability and validity of research findings, we first need to address the definition of sexuality as an object of research. Sexuality as a coherent, analytically distinguishable object of scientific research dates to the second half of the nineteenth century. Sexuality emerged as a modern topic and a discipline with reference to biology and medicine in their association with psychiatry and psychology, with its origins closely tied to criminal justice and debates about whether 'psychopathia sexualis' (sexual psychopathology) should be treated as crimes or mental illness (Krafft-Ebing, 1906/1929; Bullough, 1994). The scientific study of sexuality emerged at the nexus of the medical or biological and the social. In order for sexuality to emerge as an object of scientific interest, to give rise to a scientific discipline of its own, sexology, and to attract the interest of researchers, it had to be distinguished from procreation and reproduction (Katz, 2007). The initial focus was on pathology, on psychopathologies and perversions. It is interesting to note that of the terms that organize much of contemporary research and debate, homosexuality was coined before heterosexuality. In this context, sexuality is conceived of in terms of desires and drives used to explain problematic behaviors.

Sexuality is a broad and complex topic that has been studied using a variety of methodological approaches: from case studies, to laboratory studies of sexual response, ethnographic fieldwork, to population surveys. Each of these approaches poses their own special set of methodological issues and expertise. This article focuses on one of the main approaches to the study of sexuality in the social sciences, the population survey. The methodological issues in population surveys apply to the other approaches mentioned earlier as they invariably involve forms of interviewing and self-report on sexuality. The formative paradigm for population surveys on sexuality developed out of the work of Alfred Kinsey and his colleagues, which began in 1938 in the United States (Kinsey et al., 1948, 1953; Robinson, 1989). While Kinsey's work has been criticized for not using probability sampling, its aim was to represent the variation in sexuality in a defined population from which respondents were drawn via interviews that would elicit honest answers to the questions posed (Cochran et al., 1953). The two aims, representative sampling and valid measurement, will serve to organize this discussion of the main methodological issues in the study of sexuality. However, there is one other issue which should be discussed first, the definition of sexuality in research.

In the movement from theory to specific research questions, too often insufficient attention is paid to the choice of what, specifically, is to be studied. Research is always a series of choices and these choices affect what we can learn. They also affect the answers we get. In the context of any research that depends on interviews, the choices of what questions to ask and how to formulate them are critical. Alfred Kinsey and his colleagues' work is instructive (Kinsey et al., 1948; Gathorne-Hardy, 2000). Kinsey was a zoologist and after having studied the gall wasp, he turned his attention to sex in humans. Kinsey's focus was on 'sexual behavior,' and his goal was to describe the variation in this specific aspect of sexuality across the population which he defined as 'humans' specifically 'males' and 'females.' The decision to focus on behavior is a fundamental conceptual leap that has been enormously influential on subsequent research on sexuality, but this must still be operationalized in any given study (Michaels and Giami, 1999). What defines a specific behavior or act as sexual? Kinsey's approach was to use the 'orgasm' as the characteristic that defined a behavior as sexual. This had much to recommend it, but it is not without problems. It had the great

advantage of being clear and quantifiable. It allowed Kinsey to achieve a degree of objectivity and aided in quantitatively equating a broad array of behavior from wet dreams to masturbation, from noncoital events (in Kinsey's now dated term, 'petting'), as well as homosexual and heterosexual experiences. He focused on the concept of 'outlet,' asking questions such as what proportion of the total outlet was accounted for by any particular form of activity. In retrospect, we can see that this choice has a masculine bias (Robinson, 1989). Men are more likely to view their sexual experiences and lives in terms of events that result in an orgasm, but this is much less the case for women. Kinsey himself was aware of this problem when he began to analyze the data from women and had to add a whole discussion of the problem of orgasm and counting in his second volume, *Sexual Behavior in the Human Female* (Kinsey et al., 1953).

Definition of Sexuality in Surveys

Sexual behavior remains at the core of population surveys since Kinsey. Most of these surveys have been prompted by two main issues: fertility (cf National Survey of Family Growth) or public health concerns about sexually transmitted infections (STIs) especially HIV/AIDS (cf the National Health and Social Life Survey (NHSLS) in the U.S. (Laumann et al., 1994)), the NATSAL surveys in England (Johnson et al., 1994), and the ACSF survey in France (Spira et al., 1993). In these surveys, the focus has been on sexual partners and practices, especially practices associated with prevention either of pregnancy or STIs. All such surveys are confronted with how to define sex and what constitutes a sexual partner. In the NHSLS, we decided to use a relatively broad definition when eliciting an enumeration of partners in the past year. Here is how this was presented to respondents:

> Now I am going to be asking some questions about your sexual activity during the last 12 months. People mean different things by sex or sexual activity, but in answering these questions, we need everyone to use the same definition. Here, by 'sex' or 'sexual activity,' we mean any mutually voluntary activity with another person that involves genital contact and sexual excitement or arousal, that is, feeling really turned on, even if intercourse or orgasm did not occur.

> In answering these questions, please include all persons or times in the last 12 months where you had direct physical contact with the genitals (the sex organs) of someone else and sexual excitement or arousal occurred. Certain activities such as close dancing or kissing without genital contact should NOT be included. Also, this set of questions does NOT refer to occasions where force was used and activity was against someone's will.

Sex was defined as mutually voluntary activity involving arousal and physical contact with the genitals but not necessarily penetration or orgasm (Laumann et al., 1994). (Nonvoluntary sexual experiences were asked about separately in a later part of the interview both to minimize emotional distress and because it was not considered in the basis for defining sexual partners.) The goal was to avoid a normative definition of sex as heterosexual, involving vaginal intercourse, and to provide a clear definition that would be equally applicable to women and men. In addition, in the context of a survey motivated by concerns about HIV/AIDS, every effort was made to have a definition that would elicit reporting of both male and female partners for both men and women.

Measurement or Response Error in Surveys

One of the greatest concerns when asking about sexuality in interviews is whether subjects will answer truthfully. Asking about sex in surveys is a classic example of a 'sensitive question' that is a likely source of bias or error in reporting. Sensitive questions are those that produce the largest item nonresponse rates and are most subject to social desirability effects (responses skewed in the direction of social norms). Sexual behavior is a classic example, along with drug use, abortion, voting, and income. Income questions generally have the highest item nonresponse of any survey questions. Social desirability effects on illegal and/or highly disapproved behaviors produce underreporting. Behaviors that are desirable such as voting or having a library card are overreported. There is a fairly extensive survey research methodology literature on these questions (see Tourangeau and Van (2007) for a recent thoughtful and thorough review of this literature). There is also a literature on measurement error specifically applied to sex surveys that has mainly developed as HIV/AIDS motivated attempts to improve the methods of improving data on risk behavior, for example Catania et al. (1993) and Fenton et al. (2001).

From a methodological perspective, the goal is to minimize this form of error even if it is not possible to eliminate it. It is a goal in all surveys and many of the standard practices in surveys are intended to achieve this goal. First and foremost is the assurance of confidentiality of responses.

Researchers must make a convincing case for the legitimacy of the research, for example, that it serves an important public health goal. They must ensure participants that the confidentiality of their identity, their participation, and the information they provide is a central commitment of the research enterprise. This assurance is an important tool in convincing respondents that they can and should provide honest answers to the questions posed. It is essential to make every effort to communicate to respondents that they are not being singled out but that the research goal is to understand the ensemble of persons interviewed. As part of informed consent and protection of human subjects, respondents are always informed prior to making a decision to participate in a study that they do not have to answer any question they prefer not to answer.

The role of the interviewer is that of a 'professional stranger' who the research subject will not see again and is just doing a job. This is in fact a great advantage of survey research interviewers over other professionals who might be called upon to ask questions about sexuality such as physical or mental health professionals. The latter are asking the questions as part of a treatment protocol and are likely to have some ongoing contact with the patient. Instilling an attitude of neutral sympathetic interest is part of the standard training of interviewers and is especially important in the area of sex research.

This involves a combination of desensitization to the subject matter at hand to minimize any reaction of surprise or judgment in response to answers to questions and a general sympathetic and encouraging affect toward the respondents' efforts to answer honestly. At the same time, interviewers are trained to not express too much interest and to remain neutral in their attitudes toward whatever is reported.

In spite of the many advantages of personal interviewing whether face to face or by telephone – the ability to insure that questions are asked as written and the responses given correspond to the categories intended and are recorded accurately as well as the ability to respond to and detect any misunderstanding of the questions on the part of the respondent – a great deal of research has shown that response effects are strongest when answers are produced verbally to an interviewer (Tourangeau and Van, 2007). While there is no way to eliminate these problems, there are techniques to reduce their impact. It has long been known that self-administered questionnaires where answers are not shown or known to the interviewer produce higher rates of reporting of sensitive questions. One technique used in face-to-face interviews before the widespread use of computers by interviewers was to have respondents answer especially sensitive questions on a self-administered form and put these forms in a sealed envelope. This depended on the respondents having adequate reading and writing skills. With the expansion of the use of computer-aided interviewing, Audio-Computer Assisted Self-Interviewing (ACASI) was developed (Couper et al., 2003). Used in face-to-face interviews, the respondent is given the computer which has questions and responses on the screen and an audio track that reads the question aloud. Respondents still need a minimal level of literacy to be able to select response categories (e.g., recognizing numbers or letters) and a degree of familiarity and comfort with computers. This technique tends to produce the highest reporting of socially undesirable behaviors, including sexual behaviors, and to reduce social desirability biases such as the tendency for men to report higher numbers of sexual partners than women. Another advantage of ACASI over the older form of self-administered paper and pencil forms is that computerized forms can incorporate and guide respondents through more complex skip patterns. A similar technique to ACASI has been developed for telephone surveys, interactive voice response. The sensitive questions are administered by a recording rather than a live interviewer and the respondents provide their responses via the telephone keypad or by saying aloud the number corresponding to their answer. While ACASI represents an effective technique to reduce response error due to social desirability, it is cumbersome and may not be feasible in some research settings. It also does not eliminate the need for careful attention to other issues in questionnaire design.

The design of questionnaires and the application of established principles of question construction and survey design remain essential. In this area as well, there is an extensive literature and accumulated experience (Schaeffer, 2000; Schwarz, 1999; Sudman and Bradburn, 1982; Sudman et al., 1996). The details and precepts of this literature are far beyond the scope of this discussion. A couple of points can be made. First, it should be borne in mind that answering questions about sex involves the same problems and processes as answering any questions about behavior: comprehension of the question, recall, and formulating a response. The clarity of the question and the minimization of cognitive burden in searching and processing the necessary information to formulate an answer are as important to questions about sexual behavior as they are any other activity. Once engaged in an interview, respondents follow similar conventions of usual conversational exchange and these include trying to provide honest and useful information. Many of the well-known context effects in survey design, e.g., question order effects, effects of rating scales, can be traced to respondents' attempts to use (often inadvertent) signals provided in questionnaires that lead them to infer the implicit expectations built into the questionnaire. For example, if the response categories for a question about numbers of partners in a given period of time imply low or high numbers, answers will tend to skew in the direction implied.

Language and Terminology

A major concern in interviewing about sex is the type of language and terminology to be used. The goal is clarity and comprehension of questions. Another goal is to be culturally sensitive and there may be a temptation to use slang and/or euphemism. This is to be avoided. The general principal, especially important in general population surveys, is to use simple, clear, and unambiguous language. Common terms are best that can be understood by the widest spectrum of respondents including those with the least education. The pilot testing of questions and, where feasible, the use of cognitive interviewing on a small but diverse sample that includes less educated respondents is desirable (Willis, 2005; Miller, 2011). This can help to find terms that are understandable such as 'having sex' as opposed to intercourse, 'sex organs' instead of genitals, 'butt or rectum' instead of anus.

Representative Sampling and Generalizability

Sampling is a central concern in all empirical social science research. With limited resources, we are almost never in a position to study the whole population of interest. So we study a sample to learn about the larger population from which our sample is drawn. A representative probability sample is the gold standard. It is the only method that allows one to derive estimates from a sample that can be generalized to a larger defined population and allow one to quantify the likelihood that the derived estimates approximate the true values in the population. Often in sexuality research, some form of non-probability, usually a purposive or convenience sample is used. To some extent, researchers justify this for one of the two reasons. Either the population of interest is too small or hard to reach via probability sampling, e.g., sexual minorities such as gay men or lesbians, or because the researcher does not believe that randomly selected subjects will participate. The latter was Kinsey's position. Ironically, Kinsey the pioneering and committed empiricist did not test this belief, but accepted it as

given and then proceeded to try to collect as diverse and large a sample as possible to insure that his findings were representative. While understandable in the era in which he was working (1938 to the early 1950s) before the widespread use of surveys and a fuller understanding of the use of probability sampling, it was not until much later that extensive questions about sexual behavior were asked of respondents drawn from a probability sample. With reasonable skill and care, these surveys have achieved response rates and high-quality samples comparable to those in surveys on other topics. Whenever affordable and possible, probability sampling should be used. Binson et al. (2007) provide an excellent overview of sampling issues related to surveys on sexuality and alternative approaches to true probability samples.

Heterosexuality

This discussion has not focused specifically on heterosexuality. First, the methodological issues discussed do not vary or depend on whether we are talking about heterosexuality or homosexuality or any other form of sexual behavior, practice, or affect. In addition, heterosexuality as a category is not as transparent as one might assume. Sexuality is similar to what at first glance appear to be fixed, unambiguous demographic categories such as sex and race; their reality is complex and therefore demands careful theoretical and methodological consideration when we attempt to measure them (Prewitt, 2013). It is most often the normative categories of male/female, whiteness, or heterosexuality that most easily escape conceptual and methodological scrutiny.

For many years, data on sexual behavior were collected in the context of concerns about marriage and reproduction. In that context, often studies confined themselves to studies of married persons. As interest has broadened to include sexuality, we need to consider how the design of research determines the research questions that can be addressed. Heterosexuality is not as easily distinguishable substantively and isolatable from other forms and moments in sexuality. Researchers should not presume a priori that heterosexuality is a fixed, clear category. Even if it were to turn out to be the case in general or for certain individuals or within a given time period, it is better to approach its study without building this into the research design or formulation of the questions to be asked a priori. The evolution of the National Survey of Family Growth is instructive (see the excellent NSFG website at http://www.cdc.gov/nchs/nsfg.htm). When the NSFG began in 1973, the sample was confined to women who were or had been married and unmarried women living with their own children between the ages of 15 and 44 and it only asked questions about heterosexual partners and sexual and contraceptive practices. In the third cycle of the NSFG (1982), it expanded its sample to both single and married women in that age range, and starting in 2002, it included men as well as an expanded range of questions about both same and other sex partners and practices. While complicating the notion of heterosexuality, this design allows a much more nuanced exploration of questions about who is heterosexual and what this means. In retrospect, it is clear that the latter design produces a much fuller basis to understand heterosexuality and heterosexual practices than the more restrictive earlier framework. While this larger frame may make it more difficult to operationally define 'heterosexual' as a substantive referring to persons, it allows one to ask a much richer range of questions and to gain a more profound understanding of heterosexuality.

As has been true throughout the history of sex research, the focus has often been on the more troubling and non-normative aspects of sexuality most often in response to a pressing social or health problem. These research efforts have led to methods and data that have broader applicability. Thus, much recent population research on sexuality has been motivated by concerns about HIV/AIDS and a concern about non-normative sexual practices and identity. As the LGBT communities have mobilized politically, they have generated increased scientific interest which has informed work on sexuality more generally (IOM, 2011). One case in point is the in-depth work that has been done to improve questions about sexual identity for large-scale national health surveys (Miller and Ryan, 2011; Ridolfo et al., 2012). In-depth qualitative work on question wording for a general sexual identity question has shown that the inclusion of 'heterosexual' in this question creates the largest amount of nonresponse and mistaken responses. While persons who are members of a sexual minority, e.g., LGBT persons, understand this term, less educated respondents have trouble understanding terms such as especially 'heterosexual,' but also 'homosexual' and 'bisexual.' Sexual identity is a less salient aspect of the lives of the large majority of the population who are not constantly confronted with reflection on their status. Many think of themselves as just 'normal' and yet in most research contexts we would want to classify them as heterosexual. For those who have trouble answering an identity question, or who refuse to answer the question or respond with 'don't know,' the explicit term heterosexual is confusing.

Conclusion

This discussion has focused on the methodological issues in the design of population surveys on sexual behavior and the major threats to generalizability and measurement error. The principles discussed apply to other forms of research on sexuality such as qualitative interviews or research focused on psychological states (desire, attraction, or fantasy) or the meaning associated with sexuality. The methods used need to be in service to the research questions and objectives. The methods and principles discussed here derive from quantitative research such as demography or epidemiology that has as its primary goal the quantitative description of a population. Qualitative research often has other goals such as understanding the complexity of meaning and sense-making processes that may be individual or cultural. But the same issues of generalizability and validity of measurement apply and the lessons and advances in the area of surveys are still instructive.

The methodological challenges in research on sexuality are no different than those faced in any other type of empirical social scientific research. This may be the most important precept to remember. In spite of the aura of taboo and special significance that we often associate with sexuality, we would do best to assume the opposite (Rubin, 1984). Rather than

assuming that research subjects will not be willing to report on their sexual feelings and behaviors and thereby practice a form of scientific self-censorship, we can make the greatest contributions to knowledge by taking the position that sexuality is an important part of social life and that needs to be studied with the same methods and rigor as any other. This has been demonstrated over and over again, especially since the 1990s. When presented in a sober and serious manner, with a clear statement of the scientific and social usefulness of the inquiry and the same assurances of confidentiality of responses as in any enquiry with human subjects, members of the general public have shown their willingness to answer questions that touch on what is generally considered their most private acts and feelings and to participate in research on sexuality at the same rates as in any other type of survey. The study of sexuality is subject to the same limits and errors of measurement as all empirical scientific research and the same promise of cumulative advance of knowledge.

See also: Health Surveys; Sample Survey Methodology, History of; Sample Surveys: Cognitive Aspects of Survey Design; Sexual Dysfunction, Epidemiology of; Sexual Risk Behaviors; Sexuality and Social Work; Sexuality and Space; Surveys and Polling: Ethical Aspects; Surveys: Question Wording and Response Categories.

Bibliography

Binson, Diane, Blair, Johnny, Huebner, David, Woods, William J., 2007. Sampling in surveys of lesbian, gay and bisexual people. In: Meyer, Ilan H., Northridge, Mary E. (Eds.), The Health of Sexual Minorities: Public Health Perspectives on Lesbian, Gay, Bisexual and Transgender Populations. Springer, New York, pp. 375–418.

Bullough, Vern L., 1994. Science in the Bedroom: A History of Sex Research. Basic Books, New York.

Catania, Joseph A., Turner, Heather, Pierce, Robert C., Golden, Eve, Stocking, Carol, Binson, Diane, Mast, Karen, 1993. Response bias in surveys of AIDS related sexual behavior. In: Ostrow, David G., Kessler, Ronald (Eds.), Methodological Issues in AIDS Mental Health Research. Plenum, New York, pp. 133–162.

Cochran, William, Mosteller, Frederick, Tukey, John, 1953. Statistical problems of the Kinsey Report. Journal of the American Statistical Association 48, 673–716.

Couper, Mick P., Singer, Eleanor, Tourangeau, Roger, 2003. Understanding the effects of audio-CASI on self-reports of sensitive behavior. Public Opinion Quarterly 67, 385–395.

Ericksen, Julia A., 2001. Kiss and Tell: Surveying Sex in the Twentieth Century. Harvard University Press, Cambridge.

Fenton, Kevin A., Johnson, Anne M., McManus, Sally, Erens, Bob, 2001. Measuring sexual behavior: methodological challenges in survey research. Sexually Transmitted Infections 77, 84–92.

Gathorne-Hardy, Jonathan, 2000. Sex the Measure of All Things: A Life of Alfred C. Kinsey. Indiana University Press, Bloomington.

IOM (Institute of Medicine), 2011. The Health of Lesbian, Gay, Bisexual, and Transgender People: Building a Foundation for Better Understanding. The National Academy Press, Washington, DC.

Johnson, Anne M., Wadworth, Jane, Wellings, Kaye, Field, Julia, 1994. Sexual Attitudes and Lifestyles. Blackwell Scientific Publications, London.

Katz, Jonathan N., 2007. The Invention of Heterosexuality. University of Chicago Press, Chicago.

Kinsey, Alfred C., Pomeroy, Wardell B., Martin, Clyde E., 1948. Sexual Behavior in the Human Male. Saunders, Philadelphia.

Kinsey, Alfred C., Pomeroy, Wardell B., Martin, Clyde E., Gebhard, Paul H., 1953. Sexual Behavior in the Human Female. Saunders, Philadelphia.

Krafft-Ebing, Richard F.v., 1906/1929. Psychopathia Sexualis (Authorized translation of twelfth German Edition by F.J. Rebman). Physicians and Surgeons Book Company, Brooklyn.

Laumann, Edward O., Gagnon, John H., Michael, Richard T., Michaels, Stuart, 1994. The Social Organization of Sexuality: Sexual Practices in the U.S. University of Chicago Press, Chicago.

Michaels, Stuart, Giami, Alain, 1999. Sexual acts and sexual relationships: asking about sex in surveys. Public Opinion Quarterly 63, 385–404.

Miller, Kristen, 2011. Cognitive interviewing. In: Madans, J., Miller, K., Maitland, A., Willis, G. (Eds.), Question Evaluation Methods: Contributing to the Science of Data Quality. Wiley, Hoboken.

Miller, Kristen, Michael Ryan, J., 2011. Design, Development and Testing of the NIHS Sexual Identity Question. Office of Research and Methodology, National Center for Health Statistics: Questionnaire Design Research Laboratory. (http://wwwn.cdc.gov/qbank/report/Miller_NCHS_2011_NHIS%20Sexual%20Identity.pdf) National Survey of Family Growth: http://www.cdc.gov/nchs/nsfg.htm.

Prewitt, Kenneth, 2013. What Is Your Race? The Census and Our Flawed Efforts to Classify Americans. Princeton University Press, Princeton.

Ridolfo, Heather, Miller, Kristen, Matiland, Aaron, 2012. Measuring sexual identity using survey questionnaires: how valid are our measures? Sex Research and Social Policy 9, 113–124.

Robinson, Paul A., 1989. The Modernization of Sex: Havelock Ellis, Alfred Kinsey, William Masters, and Virginia Johnson. Cornell University Press, Ithaca, NY.

Rubin, Gayle, 1984. Thinking sex. In: Vance, Carol S. (Ed.), Pleasure and Danger. Routledge and Kegan Paul, London.

Schaeffer, Nora C., 2000. Asking questions about threatening topics: A selective overview. In: Stone, Arthur A., Bachrach, Christine A., Jobe, Jared B., Kurtzman, Howard S., Cain, Virginia S. (Eds.), The Science of Self-Report: Implications for Research and Practice. Erlbaum, Mahwah, pp. 105–121.

Schwarz, Norbert, 1999. How the questions shape the answers. American Psychologist 54, 93–105.

Sudman, Seymour, Bradburn, Norman, M., 1982. Asking Questions: A Practical Guide to Questionnaire Design. Jossey-Bass, San Francisco.

Sudman, Seymour, Bradburn, Norman, M., Schwarz, Norbert, 1996. Thinking about Answers: The Application of Cognitive Processes to Survey Methodology. Jossey-Bass, San Francisco.

Spira, Alfred, Bajos, Natalie, et al., 1993. Les Comportements Sexuels en France. La Documentation Française, Paris.

Tourangeau, Roger, Van, Ting, 2007. Sensitive questions in surveys. Psychological Bulletin 133 (5), 859–883.

Willis, Gordon B., 2005. Cognitive Interviewing: A Tool for Improving Questionnaire Design. Sage Publications, Thousand Oaks, CA.

Methodology of the History of the Social and Behavioral Sciences

Christian Fleck and Christian Dayé, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The history of the social and behavioral sciences (SBS) gets varying coverage in different disciplines. Regardless of these differences, we sketch a general and descriptive methodology for the historiography of SBS. The approach is not to seek for a consistent reconstruction of how to correctly write history, but to observe historians of SBS at work empirically. This, at best, results in a collection of models of best practice that can guide future research on the history of SBS. In all brevity, we do this for five units of analysis: (1) Actors; (2) Ideas; (3) Instruments; (4) Institutions; and (5) Contexts.

The past of the social and behavioral sciences (SBS), their 'history,' gets varying coverage in different disciplines. While the natural sciences are well covered within the 'history of science' field, this is not the case for the SBS. A comparable overarching specialty 'history of SBS' does not exist, and the various SBS disciplines handle the problematic of their past differently. In some disciplines we find specialized journals, research groups, and doctoral programs, whereas in other disciplines some or even all of these specializations are lacking. The intensity with which the trajectories of the disciplines are researched varies strongly.

Regardless of these differences, in this article we venture to draw a portrait of one aspect of the historiography of SBS, its methodology. We start by describing several analytic perspectives that histories of SBS can take and subsequently develop briefly what might be called an empirical methodology of history in the SBS.

Analytic Perspectives

The history of a particular discipline consists of narratives about former periods and the trajectory of the discipline. Different approaches to writing such a history have evolved over time. A 'philosophy of science' perspective strives to reconstruct the logical structure of the system of propositions, their relationship to underlying assumptions, and to particular statements. The aim is to advance scholarship by distinguishing good science from bad science, truth from error. A 'history of science' perspective is less concerned with the logical preconditions of science but with the detailed reconstruction of particular cases of scientific research, theorizing, controversy, and debate. The *sociological perspective* on the history of science and scholarship analyzes the opportunity structures enabling scholars to fulfill their professional obligations by emphasizing the role of norms; the modes of assigning prestige and recognition to scientists; the influence of historically contingent institutional arrangements; the forms of collaboration; and the prevailing convictions, cultural moods, and world views held by scientists. Sociologists of science are not concerned with history per se but make use of past cases and arrangements, partly because it is easier to show a detached stance toward bygone situations, but also because present-day practices are less observable than 'closed cases,' and finally because sometimes past affairs are better documented and archives more easily accessible. More often than not practitioners from these three branches got their original training in different fields of scholarship. As a consequence, the exchange of ideas, the communication across disciplinary boundaries, and therefore the mutual enrichment have been diminishing over the last decades. These fields – philosophy of science, history of science, and sociology of science – became increasingly exclusive areas, i.e., specialties in a narrow sense, and tended to look inside rather than to neighboring discourses.

A further peculiarity stems from the fact that analyzing science requires some familiarity with the subject itself; in the early days of the philosophy of science movement, its exponents claimed that one needs to have training in the particular discipline one studies in order to carry out such endeavor with success. When the 'history of science' field began to blossom in the early twentieth century, the requirements have been even more far-reaching; students were expected to be universally educated cosmopolitans mastering several languages and disciplines, and to command a good knowledge of with the history of thought from the early periods on (George Sarton functioning as a role model). Only when the second generation of sociologists of science entered the stage, the attitude toward the subject under investigation changed considerably. Adopting the habitus of anthropologists, these scholars left the historical approach behind and went into the field like ethnographers. Members of (scholarly) tribes were asked to make their laboratories available to participant observers who then recruited 'local' informants to explain what they were doing when they professed to study, e.g., the nature. Alleged ignorance functioned as a device to identify the assumptions underlying the practices of science. Detailed historical knowledge, or even awareness of the field's past, was deemed to undermine the methodical alienation that is a prerequisite for any good ethnography. When this mind-set had been established successfully within academic discourse, devotees of social studies of science applied their novel conceptual schemes to past stages of scientific research too. As a side effect, these mostly constructivist scholars got rebuttals from ontological realists. The 'science wars' of the 1990s left behind what warfare usually does: burned grounds, with both parties claiming ultimate victory.

The portrait would not be complete without mentioning a fourth type of history of science authors: Scholars writing the history of their own discipline. At best their contributions show

a discipline's degree of reflexivity; at worst they worship the successes of their authors' teachers. Usually such *historians of their own discipline* lack any training in history and do not look at their case from a comparative point of view; nonetheless, they sometimes deliver rich source material for further studies. In most cases, their audiences are discipline-compatriots, and references to philosophy/history/sociology of science literature are weak or even inexistent. Very often this type of historical analysis follows the lines of reconstructing the development of particular theories and covers the past primarily as a huge reservoir of ideas.

Today, history of science and philosophy of science, besides regular calls to collaborate more closely, are relatively separate communities with distinct patterns of scholarly communication. Specialized journals mark the boundaries of the discourses. Few departments and graduate programs unite both fields – however, SBS are very rare research topics in history of science departments. History of science primarily deals with the history of natural science; the same holds true for the interdisciplinary field of science studies. The whole field of sociological investigations has been split apart, and the only remaining shared denominator is the preference not to analyze their own disciplinary field and its trajectories over time.

Some aspects of the history of SBS, ideas, and theories in particular, are part of standard university curricula in some disciplines, but this has not led to a consolidation of the history of SBS as an autonomous field of research. Organizations devoted to the history of particular SBS disciplines are rare, one notable exception being the History of Economics Society (founded in 1974), whereas several associations are concerned with the history of a multitude of disciplines (e.g., Cheiron: The International Society for the History of the Behavioral & Social Sciences, founded in 1968; the European Society for the History of the Human Sciences, founded in 1982). Major international SBS associations have, if ever, only reluctantly introduced adequate sections or research committees. For instance, the American Anthropological Association (AAA) has no section on the history of the discipline (cf http://www.aaanet.org/sections/, accessed September 2012). The same applies for the European Association of Social Anthropology and the European Sociological Association. Compared to the other sections, American Sociological Association's History of Sociology section is relatively small. The Research Committee History of Sociology of the International Sociological Association is kind of an exception: It is relatively old and prospering. There is also a scarcity of specialized journals, with the *Journal of the History of Economic Thought*, the *Journal of the History of the Behavioral Sciences*, *History of the Human Sciences* and a few others as exceptions.

An Empirical Methodology

There are countless alternative understandings of the very notion of methodology extant in SBS literature, but most fruitful in our context seems to be an understanding inspired by the Paul F. Lazarsfeld's dictum that while a sociologist studies man in society, the task of the methodologist is to study the sociologist at work (Lazarsfeld, 1982: 171). Our task, thus, is not to seek for a logical and consistent reconstruction of instructions for how to correctly write the history of SBS, but to observe how historians of SBS do their craft. This, at best, results in a collection of models of best practice that can guide future research on the history of SBS.

In principle, historiography in the SBS is concerned with five kinds of units of analysis:

1. Actors: The history of individual scholars, of groups of like-minded authors, of generations and cohorts
2. Ideas: The history of concepts, theories, and research programs
3. Instruments: The history of methodologies, methods, techniques, and research practices
4. Institutions: The history of institutions and organizations
5. Contexts: The history of the embeddedness of science in larger social, political, cultural, and economic environments.

Actors

Traditionally studies of the history of SBS focus either on actors or ideas. Looking at actors as the unit of analysis, it can be observed that the most prominent level of aggregation is the individual. The vast majority of publications in the history of SBS are engaged in one way or another with individual scholars. Practically all leading figures from all SBS disciplines received coverage of their life and work in longer or shorter biographies (and the present Encyclopedia follows the path of its predecessors by including biographical articles on leading individuals). The standard mold of intellectual biography, life, work, and influence, claims much of the space, besides regularly raised reservations about obvious shortcomings (Gross, 2008). Collins argues persuasively that even genius-like authors happen to develop their ideas in chains of interaction with others (Collins, 1998) but his elaborated methodology found only very few follow-ups (Savelsberg and Flood, 2011). Missing in Collins' argumentation is, however, the role of the private life of scholars, which has rarely been studied systematically but is covered widely in all types of autobiographical writing. Some books contain chapters on teacher–student relations, collaboration with like-minded contemporaries, and debates with competitors (Merton and Riley, 1980). Another branch of this literature is concerned with particular research groups, schools, departments, universities, etc. (e.g., Bulmer, 1984a; Dahrendorf, 1995; Farrell, 2001; Wheatland, 2009). At an even higher level of aggregation, we have prosopographies, collective biographies, and investigations making use of the concept of generation (Blaug and Vane, 2003; Fleck, 2011). Rather popular are reports on the development of disciplines in particular nation-states (Halsey, 2004; Calhoun, 2007).

Mostly lacking are investigations on structured career patterns and the demographics of disciplines, glimpses of which can be found in the *World Social Science Report* (UNESCO, 2010; Mills et al., 2006). Sociologists focusing on actors seldom transgress Rankean narrative history by employing conceptual schemes form their own discipline. Merton's paradigm for sociology of knowledge (Merton, 1996: 208) did not attract many followers whereas in recent years,

Bourdieu's field theory and his distinction between different modes of capital entered center (Bourdieu, 1988).

Ideas

The reconstruction of the succession of leading ideas revolves around central concepts, theories, etc., mostly in the manner of the history of ideas approach in philosophy and human sciences (Grafton, 2006). At the core of a history of ideas approach stands the vision that ideas beget ideas, either by differentiation or by refinement. Whereas professional historians of ideas tend to analyze single entities or unit-ideas (Nisbet (1970) following Lovejoy (1936)), historians of SBS prefer compositions of ideas, or 'theories,' as their unit of analysis (Parsons et al., 1961; Coser, 1971; Weiler, 2014; Isaac, 2012). In the beginning, the succession of systems of propositions, their configuration and reconfiguration, has been seen as following a path of progress; later stages were seen as improvements compared to older versions, and the overall perspective remained more or less linear: later theorists were able to see further because of their standing on the 'shoulders of giants' (Merton, 1993). Since the middle of the 20th century, however, historians of SBS ideas tend to restrict their attention to particular outstanding individuals, either in the way of writing intellectual biographies of single authors or analyzing small groups of closely connected individuals, often under the label of schools (Schumpeter, 1954; Bottomore and Nisbet, 1978).

One can further differentiate the ideas according to their 'epistemic' status. Ideas, in their broadest possible meaning, can be ranked according to the degree of their explicitness. The history of ideas approach offers a variety of concepts: The 'zeitgeist' is one well-known candidate, 'tacit knowledge' (Polanyi, 1966) another one, and Ludwik Fleck's 'style of thought,' a third one (Fleck, 1979). All three share the characteristic that members of later cohorts are able to identify those thoughts easily, whereas even the most outstanding coevals had troubles to transcend them.

More explicit are sets of ideas that have been assembled into research programs. Such programmatic announcements are by their very nature explicit, and usually appear in plural. Lakatos was the first who elaborated the concept of research program (Lakatos and Musgrave, 1970). This concept has for instance been applied to economics (Marchi and Blaug, 1991; Boumans, 2012). At the core of a research program is a set of propositions which are of crucial relevance for the whole program and which is therefore defended against attempts to prove them wrong. In this area, there is considerable overlap with SBS theorists striving to reconstruct past research programs, but their neglect for the social conditions of the original formulation of the program as well as some SBS historians' exaggerated claim to explain all aspects of a research program by social factors renders collaboration difficult.

Several attempts at conceptual and theoretical innovations have been ventured recently: A dialogical turn (Levine, 1995; Camic and Joas, 2004) or a new sociology of ideas (Camic and Gross, 2001; Camic, 2010) were proposed. Outstanding studies by social science giants as Albert O. Hirschman (1991) and Robert K. Merton (Merton and Barber, 2004), who traced the rhetoric of reaction and the emergence of the concept of serendipities, respectively, did not find followers so far but offer a distinct perspective on the trajectory of bundles of ideas (Zuckerman, 2010, 2011).

Instruments

A much less frequently covered aspect of SBS's historical development concerns 'instruments.' In the hard sciences, the very notion of instruments refers to clear-cut real things, such as air pumps, microscopes, and petri dishes in the early days, and highly complex and costly things in recent times, such as satellites orbiting our planet, medical imaging from X-ray radiation to DNA sequencing, not to mention colossal machineries like CERN's Large Hadron Collider. In SBS it is less clear what instruments look like. Abbott (2011) argues persuasively that the humanist's tool is a well-stacked library, offering physical access to huge numbers of publications assembled according to a not too arbitrary system of classification. But there are other types of SBS researchers, empirical sociologists, psychological experimenters, and economic or demographic modelers among them. Not only do they need money to hire helping hands, they also developed and adopted real instruments over the course of decades: Questionnaires, coding handbooks, machines to process punched cards, devices for storing data as magnetic tapes, etc. Furthermore, assembling a group of people for a focus group interview needs an appropriate space, asks for utensils to record what has been spoken or to make use of Super 8 mm film, later replaced by video and more recently digital cameras. For field researchers it might be sufficient to possess a pencil and a notebook as devices, accompanied by the quintessential humanistic instruments: a trained brain and a good memory.

From a more systematic point of view, one could argue that established search strategies, methodologies, techniques, and finally exemplars also belong to the category of instruments. At the level of search strategies, SBS scholars can be divided crudely according to their inclination toward the average man, Quetelet's 'l'homme moyen,' on the one hand and the genius on the other hand. If one is committed to the first perspective, every single individual observed could be exchanged for another one (Igo, 2007), whereas in the opposite branch of scholarship, the researcher is looking for particularities, following an idiographic approach which has its vanishing point in the unparalleled individual, the great man of the old days. Ordinary SBS research happens in between these two poles, some of it based on what one could properly name an instrument, other lacking such independent existing gadgets. Most of the scholarly discussion on search strategies is located in a philosophical frame of references. Some German philosophers, for example, called their respective thinking 'philosophical anthropology' by referring to basic assumptions about mankind. Others debated nearly the same under the title of universals, or in cultural anthropology with regard to cultural relativism. All these contributions assume that a particular set of capabilities is demonstrated or even better possessed by all men and women.

Less abstract than search strategies are methodologies. The debates on methodology are dominated either by experts in

research methodology or philosophy of (social) science. Due to their competences and interests, they seldom study the historical trajectory of particular methodologies in detail, but rather their potential refinements or abstract justifications. A historical analysis of methodologies would accentuate the contingency of particular developments, by highlighting personal influences, philosophical or *meta*-theoretical fashions, and opportunity structures at particular cognitive microenvironments to detect the roots of novel methodologies or to identify which forces support continuity in once established routines (Bunge, 1998; Outhwaite and Turner, 2007). At this level of abstraction, methodologies are transformed into routine techniques of doing research, and what has been said about potential lanes of analyzing methodologies can therefore be applied (Platt, 1996; Steinmetz, 2005; Savage, 2010).

A particularly interesting phenomenon is the case of exemplars. Thomas S. Kuhn made them famous and drew attention to the fact that ordinary scholars very often imitate exceptional models (Kuhn, 1962). A serious history of such model studies which functioned as exemplars in particular disciplines for a particular period of time is a strong desideratum in the history of SBS (Lazarsfeld, 2011).

Switching the perspective of analysis from the instruments to their qualities, one can ask for the degree of adequacy of each instrument both in synchronic and diachronic perspective. Is a particular search strategy, methodology, etc., used only in particular microenvironments (such as departments, schools, research institutes, etc.) or is it dominant at the national or even international level? How did it happen that this routine was established locally, nationally, or internationally? A second question could be directed toward the lineage of a particular instrument: Have there been predecessors, and how have they been transformed into the new shape; did the advocates of the new tool quarrel with those who favored the old one, or did they ignore particular rivals? What has been said about competitors in time could also be applied to synchronic opponents (see Merton, 1996).

Institutions

Analyzing institutions is quintessential in furthering our understanding of what happens in science and scholarship. Opportunity structures (Merton, 1995) can be seen as the conditions that shape scholarship with regard to resources; opportunity structures emerge, diffuse, and differentiate over time and space. We can describe (1) the location where scholars exhibit their profession (ranging from stable cognitive microenvironments as departments to more temporary collaborations and encounters, (2) the audiences they address, and (3) the modes of governance to which they are subjected. (1) Up to now, both sociologists and historians of science investigated primarily units of relatively stable shape and paid less attention to the more fleeting opportunities to exchange ideas. Traditional academies did not play a big role in SBS, but research networks of the 'invisible college' type and occasional happenstances as the coming together of a class of fellows at one of the centers of advanced study, offering not only the 'leisure of the theory class' (Bell, 2000: 448) but serendipitous stimulation for each of them, deserve closer examination (Heilbron et al., 2008). The still underresearched forum where ideas were exchanged, evaluated, and further developed are publications, and journals in particular. These media deserve to be studied more deeply because they exhibit the assignment of status and reputation via peer review and direct the attention of the members of each discipline toward those contributions which need to be recognized. The procedures of peer review have been studied by scientometricians since decades, however poorly recognized beyond these circles, and have been quite recently more visibly debated (Lamont, 2009). Journals, through their development over time and changes in their policies caught the attention of only very few sociologists and historians of SBS (Besnard, 1979; Lepenies, 1981; Abbott, 1999), whereas encyclopedias of the social sciences are still waiting to become the subject of serious investigations, besides their obvious role in shaping the knowledge of the covered disciplines and exhibiting the distribution of the cultural capital within them. Translations are another instance of exchanging ideas and shaping the international migration of ideas, and recent attempts to analyze them demonstrate the fruitfulness of this kind of investigation (Pollak, 1986; Heilbron, 1999; Sapiro and Bustamente, 2009). Academic and non-academic publishing houses, their market (and, consequentially, book selection) strategies, their distribution policies and technologies further are crucial factors in shaping the gestalt of SBS.

(2) The addressed audience might be other academics only, the lay people, or clients, either from the private or the public sector of any society or state (Fleck et al., 2009; Fleck and Hess, 2014). Seeing addressees as clients points to their influence on what is at the core of scholarship, whereas looking at them as uneducated and unsophisticated audiences only invites to see the 'masses' vis-à-vis the known experts (Lazarsfeld et al., 1967; Halliday and Janowitz, 1992). The distance between those who do science and those who support them or pay for them varies in history, but at any given time any variant of scholarship is embedded in larger environments, offering the mandarins time and resources at different levels of generosity, stimulating or discouraging their striving for the truth (Ringer, 1969; Lepenies, 1988; Calhoun, 2007; Fleck, 2011).

(3) Since science policy decision makers are part of the audiences, too, there is some overlap of the modes of governance with the aforementioned audiences. However, but the power relationship is reversed: Scholars are on the receiving side and those who supply funds are in the position to dictate the rules, as it happened recently with the imposition of new public management's audit culture (Strathern, 2000; Münch, 2007, 2011). The widely debated Mode-2 perspective (Gibbons, 1994; Nowotny et al., 2001) has rarely been discussed with regard to developments in SBS (Albert, 2003; Kropp and Blok, 2011). Earlier instances of models of governance have been investigated both with regard to the role of the state in general (Wagner, 1990) and philanthropic foundations (Bulmer, 1984b; Fisher, 1993).

The increased interest in institutional thinking did not have sustainable consequences for writing the history of SBS so far. Both small-scaled (e.g., sabbatical, project research, teamwork, etc.) and large institutional frameworks have been researched only in the sociology of science, particularly in the world

polity approach (Drori et al., 2003, 2006). Even when small environments have been researched as, e.g., departments, the explanations for their rise and decline did not take into account the larger institutional settings, very much to the disadvantage of their results.

Contexts

Since the beginnings of sociology of knowledge and sociology of science, contextualization is the distinct feature of any externalist approach to study the texts of surrounding authors and/or the nontextual (political, institutional, etc.) environment. Galison (2008) rightly emphasized that we need a specification of the meaning and the range of the concept 'context' and its application in history of science and related fields of research. We will highlight only two types of contexts here, which affected the trajectory of SBS over the last two centuries. These are governance on the one hand and political regimes on the other hand. In addition to what has already been said with regard to the first, we want to draw attention to another feature: patterns of internationalization of SBS. Most analyses still follow the directions laid out by modernization theory since the 1950s. Patterns of diffusion and imitation accompanied by some pinches of cultural imperialism dominate most of the publications and this might be appropriate if the focus of one's analysis is on the macro-level of the science system: What has been labeled 'isomorphism,' the consolidation and following impact of the modern world polity and culture system on different branches of scholarship and institutions worldwide, is at the moment the most prominent approach (Drori et al., 2003, 2006). Several studies on the 'Americanization' of disciplines' theories and methodologies elsewhere could be mentioned here (Haney, 2008; comp. Heilbron et al., 2008). Another recent enhancement of contextualization is provided by putting globalization of science and scholarship on its head, to bring the contributions of peripheral scholars to the attention of scholars at the center (Patel, 2010) or offering insights from distinct patterns of a politics of academic autonomy (Beigel, 2013). The recent flourishing of behavioral economics could be mentioned as an illustrative record, where a disreputable empirical social research technique, experimentation, became prolific because of its application by sophisticated economists. Their longtime disregard for empirical research has been transformed into an asset because the experimenters were able to design their experiments ingeniously due to their theoretical preponderance.

The impact of political regimes on the SBS has been extensively studied for the Nazi case and its expulsion of mostly Jewish intellectuals (Coser, 1984; Ash and Söllner, 1996; Feichtinger, 2001; Fleck, 2015) and to a lesser degree for the communist world (Sparschuh and Becker, 1997; Keen and Mucha, 1994, 2006; Voříšek, 2012, 2011; Fleck and Hess, 2011). A recently blossoming branch of literature investigates the impact of the Cold War on the cognitive gestalt and social organization of SBS (e.g., Abbott and Sparrow, 2007; Cohen-Cole, 2009, 2014; Engerman, 2009; Erickson et al., 2013; Gerovitch, 2002; Mirowski, 2002, 2005; Robin, 2001; Rohde, 2013; Solovey, 2013; Solovey and Cravens, 2012).

Conclusion

The history of SBS is a field not yet fully established within academia. In addition, crucial differences in the quality and intensity of historical scholarship exist between the various SBS disciplines. Thus, any attempt to project a methodology for SBS historiography is probably most successful if it conceives of itself as a descriptive endeavor inspecting the existing historical literature, and not as an endeavor of philosophical thinking that attempts to logically deduct prescriptions for how to correctly write the history of SBS.

Bibliography

Abbott, A., 1999. Department & Discipline. Chicago Sociology at One Hundred. Chicago University Press, Chicago.

Abbott, A., 2011. Library research infrastructure for humanistic and social scientific scholarship in America in the twentieth century. In: Camic, C., et al. (Eds.), Social Knowledge in the Making. University of Chicago Press, Chicago, pp. 43–87.

Abbott, A., Sparrow, J.T., 2007. Hot war, cold war: the structures of sociological action, 1940–1955. In: Calhoun, C. (Ed.), Sociology in America: A history. University of Chicago Press, Chicago, pp. 281–313.

Albert, M., 2003. Universities and the market economy: the differential impact on knowledge production in sociology and economics. Higher Education 45, 147–182.

Ash, M.G., Söllner, A. (Eds.), 1996. Forced Migration and Scientific Change: Emigre German-Speaking Scientists and Scholars after 1933. Cambridge University Press, Washington, DC.

Beigel, F. (Ed.), 2013. The Politics of Academic Autonomy in Latin America. Ashgate, Farnham.

Bell, D., [1960]2000. The End of Ideology: On the Exhaustion of Political Ideas in the Fifties. Harvard University Press, Cambridge, MA.

Besnard, P., 1979. La formation de l'équipe de l'Année sociologique. Revue française de sociologie 20 (1), 7–31.

Blaug, M., Vane, H.R., 2003. Who's Who in Economics. Edward Elgar, Cheltenham.

Bottomore, T.B., Nisbet, R.A. (Eds.), 1978. A History of Sociological Analysis. Basic Books, New York.

Boumans, M., 2012. Mark Blaug (1927–2011). An unrepentant Lakatosian. History of Economic Ideas 20, 11–24.

Bourdieu, P., 1988. Homo academicus. Stanford University Press, Stanford.

Bulmer, M., 1984a. Philanthropic foundations and the development of the social sciences in the early twentieth century: a reply to Donald Fisher. Sociology 18, 572–579.

Bulmer, M., 1984b. The Chicago School of Sociology: Institutionalization, Diversity, and the Rise of Sociological Research. University of Chicago Press, Chicago.

Bunge, M., 1998. Social Science under Debate: A Philosophical Perspective. University of Toronto Press, Toronto.

Calhoun, C. (Ed.), 2007. Sociology in America: A History. University of Chicago Press, Chicago.

Calhoun, C. (Ed.), 2010. Robert K. Merton: Sociology of Science and Sociology as Science. Columbia University Press, New York.

Camic, C., 2010. How Merton sociologizes the history of ideas. In: Calhoun, C. (Ed.), Robert K. Merton: Sociology of Science and Sociology as Science. Columbia University Press, New York, pp. 273–295.

Camic, C., Gross, N., 2001. The new sociology of ideas. In: Blau, J.R. (Ed.), The Blackwell Companion to Sociology. Blackwell, Malden, MA, pp. 236–249.

Camic, C., Joas, H. (Eds.), 2004. The Dialogical Turn: New Roles for Sociology in the Postdisciplinary Age. Rowman & Littlefield, Lanham, MD.

Cohen-Cole, J., 2009. The creative American: cold war salons, social science, and the cure for modern society. Isis 100 (2), 219–262.

Cohen-Cole, J., 2014. The Open Mind: Cold War Politics and the Sciences of Human Nature. The University of Chicago Press, Chicago.

Collins, R., 1998. The Sociology of Philosophies: A Global Theory of Intellectual Change. Belknap Press of Harvard University Press, Cambridge.

Coser, L.A., 1971. Masters of Sociological Thought: Ideas in Historical and Social Context. Harcourt Brace Jovanovich, New York.

Coser, L.A., 1984. Refugee Scholars in America: Their Impact and Their Experiences. Yale University Press, New Haven.

Dahrendorf, R., 1995. LSE: A History of the London School of Economics and Political Science, 1895–1995. Oxford University Press, Oxford.

Drori, G., et al., 2003. Science in the Modern World Polity: Institutionalization and Globalization. Stanford University Press, Stanford.
Drori, G., et al., 2006. Globalization and Organization: World Society and Organizational Change. Oxford University Press, Oxford.
Engerman, D.C., 2009. Know Your Enemy: The Rise and Fall of America's Soviet Experts. Oxford University Press, Oxford.
Erickson, P., Klein, J.L., Daston, L., Lemov, R., Sturm, T., Gordin, M.D., 2013. How Reason Almost Lost its Mind: The Strange Career of Cold War Rationality. The University of Chicago Press, Chicago.
Farrell, M.P., 2001. Collaborative Circles: Friendship Dynamics & Creative Work. The University of Chicago Press, Chicago.
Feichtinger, J., 2001. Wissenschaft zwischen den Kulturen. Österreichische Hochschullehrer in der Emigration 1933–1945. Campus, Frankfurt.
Fisher, D., 1993. Fundamental Development of the Social Sciences: Rockefeller Philanthropy and the United States Social Science Research Council. University of Michigan Press, Ann Arbor.
Fleck, L., 1979. Genesis and Development of a Scientific Fact. University of Chicago Press, Chicago.
Fleck, C., 2011. A Transatlantic History of the Social Sciences: Robber Barons, the Third Reich and the Invention of Empirical Social Research. Bloomsbury, London.
Fleck, C., 2015. Etablierung in der Fremde. Vertriebene Wissenschaftler in den USA nach 1933. Campus, Frankfurt.
Fleck, C., Hess, A., 2011. Sociology and communism: coming to terms with a discipline's past. Comparative Sociology 10, 670–690.
Fleck, C., Hess, A. (Eds.), 2014. Knowledge for Whom? Public Sociology in the Making. Ashgate, Aldershot.
Fleck, C., Hess, A., Lyon, E.S. (Eds.), 2009. Intellectuals and Their Publics. Perspectives from the Social Sciences. Ashgate, Aldershot.
Galison, P., 2008. Ten problems in history and philosophy of science. Isis 99 (1), 111–124.
Gerovitch, S., 2002. From Newspeak to Cyberspeak: A History of Soviet Cybernetics. MIT Press, Cambridge.
Gibbons, M. (Ed.), 1994. The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies. SAGE, London.
Grafton, A., 2006. The history of ideas: precept and practice, 1950–2000 and beyond. Journal of the History of Ideas 67 (1), 1–32.
Gross, N., 2008. Richard Rorty: The Making of an American Philosopher. The University of Chicago Press, Chicago.
Halliday, T.C., Janowitz, M. (Eds.), 1992. Sociology and Its Publics: The Forms and Fates of Disciplinary Organization. University of Chicago Press, Chicago.
Halsey, A.H., 2004. A History of Sociology in Britain: Science, Literature, and Society. Oxford University Press, Oxford.
Haney, D.P., 2008. The Americanization of Social Science: Intellectuals and Public Responsibility in the Postwar United States. Temple University Press, Philadelphia.
Heilbron, J., 1999. Towards a sociology of translation. Book translation as a cultural world system. European Journal of Social Theory 2, 429–444.
Heilbron, J., Guilhot, N., Jeanpierre, L., 2008. Toward a transnational history of the social sciences. Journal of the History of the Behavioral Sciences 44, 146–160.
Hirschman, A.O., 1991. The Rhetoric of Reaction: Perversity, Futility, Jeopardy. Belknap Press, Cambridge, MA.
Isaac, J., 2012. Working Knowledge: Making the Human Sciences from Parsons to Kuhn. Harvard University Press, Cambridge, MA.
Igo, S.E., 2007. The Averaged American: Surveys, Citizens, and the Making of a Mass Public. Harvard University Press, Cambridge, MA.
Keen, M.F., Mucha, J. (Eds.), 1994. Eastern Europe in Transformation: The Impact on Sociology. Greenwood Press, Westport, CT.
Keen, M.F., Mucha, J. (Eds.), 2006. Autobiographies of Transformation: Lives in Central and Eastern Europe. Routledge, London.
Kropp, K., Blok, A., 2011. Mode-2 social science knowledge production? the case of Danish sociology between institutional crisis and new welfare stabilizations. Science and Public Policy 38, 213–224.
Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Lakatos, I., Musgrave, A. (Eds.), 1970. Criticism and the Growth of Knowledge. Cambridge University Press, Cambridge.
Lamont, M., 2009. How Professors Think: Inside the Curious World of Academic Judgment. Harvard University Press, Cambridge.
Lazarsfeld, P.F., 1982. Problems in methodology. In: Kendall, P.L. (Ed.), The Varied Sociology of Paul F. Lazarsfeld: Writings. Columbia University Press, New York, pp. 171–208.
Lazarsfeld, P.F., 2011. An Empirical Theory of Social Action: Collected Writings. Bardwell Press, Oxford.
Lazarsfeld, P.F., Sewell, W.H., Wilensky, H.L. (Eds.), 1967. The Uses of Sociology. Basic Books, New York.
Lepenies, W. (Ed.), 1981. Geschichte der Soziologie. Studien zur kognitiven, sozialen und historischen Identität einer Disziplin. Suhrkamp, Frankfurt.
Lepenies, W., 1988. Between Literature and Science: The Rise of Sociology. Cambridge University Press, Cambridge.
Levine, D.N., 1995. Visions of the Sociological Tradition. University of Chicago Press, Chicago.
Lovejoy, A.O., 1936. The Great Chain of Being: A Study of the History of an Idea. Harvard University Press, Cambridge.
Marchi, N. de, Blaug, M. (Eds.), 1991. Appraising Economic Theories: Studies in the Methodology of Research Programs. Edwar Elgar, Aldershot.
Merton, R.K., 1993. On the Shoulders of Giants: A Shandean Postscript. The University of Chicago Press, Chicago.
Merton, R.K., 1995. Opportunity structure: the emergence, diffusion, and differentiation of a sociological concept, 1930s – 1950s. In: Adler, F., Laufer, W.S. (Eds.), Advances in Criminological Theory, The legacy of anomie theory, vol. 6. Transaction Publishers, New Brunswick, pp. 3–78.
Merton, R.K. (Ed.), 1996. Paradigm for the sociology of knowledge [originally published 1945]. On Social Structure and Science. The University of Chicago Press, Chicago, pp. 205–222.
Merton, R.K., Barber, E.G., 2004. The Travels and Adventures of Serendipity: A Study in Sociological Semantics and the Sociology of Science. Princeton University Press, Princeton.
Merton, R.K., Riley, M. (Eds.), 1980. Sociological Traditions from Generation to Generation: Glimpses of the American Experience. Ablex Corporation, Norwood.
Mills, D., et al., 2006. Demographic Review of the UK Social Sciences. Economic & Social Research Council, London.
Mirowski, P., 2002. Machine Dreams: Economics Becomes a Cyborg Science. Cambridge University Press, Cambridge.
Mirowski, P., 2005. How positivism made a pact with the postwar social sciences in the United States. In: Steinmetz, G. (Ed.), The Politics of Method in the Human Sciences: Positivism and Its Epistemological Others. Duke University Press, Durham, pp. 142–172.
Münch, R., 2007. Die akademische Elite. Zur sozialen Konstruktion wissenschaftlicher Exzellenz. Suhrkamp, Frankfurt.
Münch, R., 2011. Akademischer Kapitalismus. Zur politischen Ökonomie der Hochschulreform. Suhrkamp, Berlin.
Nisbet, R.A., 1970. The Social Bond: An Introduction to the Study of Society. Knopf, New York.
Nowotny, H., et al. (Eds.), 2001. Re-Thinking Science: Knowledge and the Public in an Age of Uncertainty. Polity Press, Cambridge.
Outhwaite, W., Turner, S.P. (Eds.), 2007. The Sage Handbook of Social Science Methodology. SAGE, London.
Parsons, T., et al., 1961. Theories of Society: Foundations of Modern Sociological Theory. Free Press of Glencoe, New York.
Patel, S. (Ed.), 2010. The ISA Handbook of Diverse Sociological Traditions. SAGE, Los Angeles.
Platt, J., 1996. A History of Sociological Research Methods in America, 1920–1960. Cambridge University Press, Cambridge.
Polanyi, M., 1966. The Tacit Dimension. Doubleday, Garden City, NY.
Pollak, M., 1986. Die Rezeption Max Webers in Frankreich. Kölner Zeitschrift für Soziologie und Sozialpsychologie 38, 670–684.
Ringer, F.K., 1969. The Decline of the German Mandarins: The German Academic Community, 1890–1933. Harvard University Press, Cambridge.
Robin, R., 2001. The Making of the Cold War Enemy: Culture and Politics in the Military-Intellectual Complex. Princeton University Press, Princeton.
Rohde, J., 2013. Armed with Expertise: The Militarization of American Social Research during the Cold War. Cornell University Press, Ithaca.
Sapiro, G., Bustamente, M., 2009. Translation as a measure of international consecration. Mapping the world distribution of Bourdieu's books in translation. Sociologica 2–3.
Savage, M., 2010. Identities and Social Change in Britain since 1940: The Politics of Method. Oxford University Press, Oxford.
Savelsberg, J., Flood, S., 2011. American criminology meets Collins: global theory of intellectual change and a policy-oriented field. Sociological Forum 26, 21–44.
Schumpeter, J.A., [1954]1986. History of Economic Analysis. Oxford University Press, New York.
Solovey, M., 2013. Shaky Foundations: The Politics-Patronage-Social Science Nexus in Cold War America. Rutgers University Press, New Brunswick.
Solovey, M., Cravens, H. (Eds.), 2012. Cold War Social Science: Knowledge Production, Liberal Democracy, and Human Nature. Palgrave Macmillan, New York.

Sparschuh, V., Becker, U., 1997. Sozialismus und Soziologie. Die Gründergeneration der DDR Soziologie. Leske und Budrich, Opladen.
Steinmetz, G. (Ed.), 2005. The Politics of Method in the Human Sciences: Positivism and Its Epistemological Others. Duke University Press, Durham.
Strathern, M., 2000. Audit Cultures. Anthropological Studies in Accountability, Ethics, and the Academy. Routledge, London.
UNESCO (Ed.), 2010. World Social Science Report: Knowledge Divides. Unesco, Paris.
Voříšek, M., 2012. The Reform Generation: 1960's Czechoslovak Sociology in a Comparative Perspective. Kalich, Prague.
Voříšek, M., 2011. In whose service? The 1960s' Czechoslovak sociologists and their party. Comparative Sociology 10, 781–806.
Wagner, P., 1990. Sozialwissenschaften und Staat. Campus, Frankfurt.
Weiler, B., 2014. The Order of Progress. Bardwell Press, Oxford.
Wheatland, T., 2009. The Frankfurt School in Exile. University of Minnesota Press, Minneapolis.
Zuckerman, H., 2010. On sociological semantics as an evolving research program. In: Calhoun, C. (Ed.), Robert K. Merton: Sociology of Science and Sociology as Science. Columbia University Press, New York, pp. 253–271.
Zuckerman, H., 2011. The Matthew effect writ large and larger: a study in sociological semantics. In: Elkana, Y., Szigeti, A., Lissauer, G. (Eds.), Concepts and the Social Order: Robert K. Merton and the Future of Sociology. Central European University Press, Budapest, pp. 121–164.

Mexican Revolution, The

Thomas Rath, University College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Knight, volume 14, pp. 9765–9770, © 2001, Elsevier Ltd.

Abstract

The article discusses interpretations of the Mexican Revolution, focusing on the revolution's origins, armed phase, and postrevolutionary state building, along with the key themes anticlericalism, Cardenismo, international relations, indigenismo, women and gender, the Pax Priísta, and historical memory. It argues that populism, revisionism, and postrevisionism offer a useful way to categorize the main trends in a large and expanding body of scholarship. While scholarship has been influenced by new empirical data, and broader theoretical and methodological trends in the social sciences, it has also been (and will remain) strongly influenced by contemporary politics in Mexico.

The Mexican Revolution is difficult to define neatly in time, space, or in terms of its ideological content. It began in 1910, when the liberal reformer Francisco Madero campaigned for presidency, was arrested by the government, escaped from prison, and issued a call for rebellion on November 20. By May 1911, a revolutionary movement of middle-class reformers and dissident peasants had emerged around Madero to oust General Porfirio Díaz, who had ruled the country for all but 4 years since 1877. The revolution's end is harder to discern; it lacked a clearly defined vanguard or ideology, and worked differently in different parts of the country. However, the revolution defined Mexico's twentieth-century history; it mobilized unprecedented numbers, brought warfare, disease, and famine in its wake; inspired sweeping social reform and cultural experimentation; and created an official party-state that ruled Mexico for 71 years. Although Mexicans rarely saw their revolution as a model to be exported abroad, it provided a source of inspiration for other reformers across Latin America, and has produced an extensive body of scholarship in history and the social sciences, both in and outside of Mexico.

Interpretations

It is useful to group scholarship into three broad interpretations. Populist accounts tended to portray the revolution as a popular, agrarian struggle that toppled the dictator Díaz, consolidated Mexican sovereignty, and delivered substantial gains for the masses, particularly during the presidency of Lázaro Cárdenas (1934–40). After the heightened suppression of students in and after 1968, revisionist perspectives predominated, which downplayed the revolution's popular support and accomplishments, and emphasized the continuities between the old Porfirian order and the new revolutionary state (Baily, 1978). Since the 1990s, postrevisionists have generally sought to reconcile these perspectives, emphasizing the dynamic relationship between revolutionary leadership and popular groups, state and civil society, and the variety of experiences in Mexico's different regions (Joseph and Nugent, 1994). These changes in interpretation have been shaped by (but are not reducible to) larger trends in historiography. Particularly important has been the proliferation of regional and local case studies first pioneered by González (1974). Like elsewhere, the 1980s and the 1990s also saw a turn away from perceived economic reductionism and a growing interest in cultural identities and language (Joseph and Deans-Smith, 1999), although the cultural turn in Mexican historiography now appears to have largely run its course. The best new scholarship combines deep empirical research with a notable theoretical and methodological eclecticism well suited to the revolution's sprawling, multifaceted character.

Origins

Most historians have sought the origins of the revolution in the profound changes Mexico underwent during the Porfiriato – the period of Porfirio Díaz's rule. Historians have traditionally seen the Porfiriato as a time of exclusionary, foreign-dominated capitalist development, overseen by an antinational and increasingly authoritarian political elite. The past two decades have seen a burst of new, impressively detailed research on the Porfirian political economy, which has questioned aspects of the old story. Despite integration into global markets, the domestic market was large and important; the state played a crucial role in economic development, not only as a conduit for cronyism and corruption (although this was crucial) but also through coherent policies of fiscal reform and a moderate protectionism. New research has painted a complicated picture of the effects of liberal laws on communal peasant land; some communities kept lands, some litigated against surveying decisions, and some sought to take advantage of liberal laws to privatize land on locally acceptable terms. Overall, newer studies are less confident than older works in linking the problems of the Porfirian political economy to the causes of the revolution, and tend to focus instead on an ossified and exclusionary political system (Riguzzi, 2009). Recent work on nineteenth-century popular liberalism has complemented this turn to political causes in some ways, showing the long tradition of peasants mobilizing for national goals and demanding local autonomy (McNamara, 2007).

The Armed Phase

Madero's movement inaugurated 10 years of upheaval and warfare – what historians usually refer to as the revolution's armed phase from 1910 to 1920. Madero proved an ineffective president: on one side, he faced the old Porfirian establishment, particularly the army, which he largely left intact; on the other, he faced frustrated former allies who expected more radical land and labor policies. Eventually, Madero was toppled by a counterrevolutionary coup and General Victoriano Huerta – whose sanguinary reputation earned him the nickname of the jackal – took power in February 1913. Huerta assassinated Madero and militarized the country, but only succeeded in fueling more rebellions; in July 1914, Huerta was toppled by a diverse coalition of revolutionary forces hailing from different regions of the country.

Between, 1915 and 1920, different revolutionary factions then fought among themselves for control of the national government. In the southern states of Morelos, Puebla, and Guerrero, Emiliano Zapata led his army of peasant villagers. Pancho Villa's movement emerged in the northern state of Chihuahua, and reflected the social complexity of northern society: miners, cowboys, fiercely independent communities from the sierra. The Constitutionalists were based in the northern states of Coahuila and Sonora, and were generally more middle-class and well connected than the Zapatistas and Villistas; indeed, the Constitutionalist leader Venustiano Carranza had been state governor of Coahuila (Brading, 1980). Despite the Zapatistas and Villistas joining forces, after a series of decisive battles in the central Bajío region in the summer of 1915, the Constitutionalists gained the ascendancy. Constitutionalist General Álvaro Obregón was a formidable military strategist, and one whom Villa appeared to underestimate. In 1919, Constitutionalist forces ambushed and assassinated Zapata in Morelos. Villa's forces were gradually reduced to small-scale guerrilla warfare; Villa was permitted to retire to a hacienda, but was later assassinated in 1923. In 1920, the Constitutionist faction itself fragmented; a powerful clique of revolutionaries from the northwestern state of Sonora, led by Obregón, rebelled against Carranza and took power. Carranaza fled the capital, only to be assassinated in a small village in the sierra of Puebla.

The revolution meant different things to different people. Madero had initially focused on the need for political reforms, clean elections, and good government. However, other groups soon emerged seeking more radical policies. The Zapatistas sought land, local autonomy, and an end to central government abuses, particularly electoral imposition and the draft. The Villistas also adopted the cause of land reform, although Villa delayed most land grants until after the fighting had finished (Katz, 1998). The Constitutionalists were far less interested in land reform. However, some Constitutionalists were radical in their willingness to ally with (and offer concessions to) organized labor, and in their determination to remove the Catholic Church's influence on public life. In 1916–17, the Constitutionalists met to draft a new constitution. The Constitutionalists' need to undermine support for Villa and Zapata, and the divisions within their own ranks, ensured that the resulting document was a good deal more radical than Carranza intended. True to Madero's original impulses, the constitution included many familiar liberal democratic elements: regular elections, division of powers, and civilian control of the military. However, Article 27 also gave the government ownership of Mexico's land and subsoil, and the right to distribute these resources for the good of the nation. Article 133 made the constitution the first in the world to guarantee workers' rights, including the right to strike and the 8-h day. Finally, the new constitution built on and extended nineteenth-century liberal anticlericalism; Article 3 secularized primary education and forbade political comment by religious institutions, and Article 130 allowed state governments to control the number of clergy in the country.

(Post)revolutionary Reform and State Building

After 1920, successive administrations went about gradually rebuilding a national government and economy, and selectively enacted reforms from the Constitution of 1917. From 1920 to 1934, the Sonoran faction ran the government, led by Generals Obregón (President, 1920–24) and Plutarco Elías Calles (President, 1924–28). The Sonorans defeated military rebellions in 1923–24, 1927, and 1929, and obtained recognition and loans from the US government. They experimented with new forms of mass politics but, despite much talk of socialism and bolshevism, remained firmly committed to capitalist development and pursued land reform in piecemeal fashion. In 1928, they also founded a national party that claimed to be the true heir of the revolution. This party ruled the country in one form or another from 1929 to 2000, and has returned to the presidency once more in 2012.

Few scholars would maintain that the official party still represents the revolution in any meaningful sense. However, debate continues about quite when the wave of political mobilization, social reform, and cultural experimentation unleashed by the revolution definitively receded; or, indeed, if the wave was ever more than a puny ripple. Reflecting this uncertainty, many historians refer to the decades after 1920 as the postrevolutionary period, while others would prefer to label it a continuation of the revolutionary process. The choice of an ending largely depends on what aspects of the revolution are emphasized. Focusing on political legitimacy, some would extend the revolutionary period until the dramatic student protests of 1968, although the degree to which 1968 really changed Mexicans' perception of their government is much debated. Some would move the ending to the debt crisis of 1982, when the party's model of economic nationalism and state-led industrial development collapsed (Beezley, 2007). However, if one focuses on what really made Mexico's experience revolutionary and distinctive in Latin America – radical attempts to redistribute property and political power – it is convincing (and conventional) to date the end to around 1940. In the 1940s, a new, more civilian generation took over the party, tacked rightward, and pursued probusiness policies. In 1946, the party adopted an odd but entirely appropriate new name: The Party of the Institutional Revolution (Partido Revolucionario Institucional, PRI). Regardless of which threads are focused on, it is clear that the formation of the new regime was very protracted, geographically uneven, and conflictive.

Anticlericalism

One key source of conflict was the relationship between the new regime and the church. During the constitutional congress of 1916–17, anticlericalism provoked by far the most vociferous debate among the 217 delegates. Conflict with the church came to a head after Calles enforced the constitution's new anticlerical provisions. In 1926–29, thousands of Catholics in the center-west of the country rebelled against the government in a movement known as the Cristiada. For decades, the government portrayed the rebels as poor, superstitious peasants duped by reactionary landlords. Jan Meyer's (1976) impressively researched study countered this official history and portrayed the movement instead as a popular, religious movement against an authoritarian state, and detailed the bloody counterinsurgency by which the government contained the rebellion. Recent studies have allowed for a more nuanced understanding of the religious conflict, revealing the dynamic, changing quality of popular religion, the competing varieties of revolutionary anticlericalism, and how local rivalries between villages helped determine which side peasants ended up supporting (Butler, 2007). The government eventually defused the rebellion by brokering a deal with the Catholic hierarchy in 1929. However, conflict between the revolutionary state and Catholic groups continued to flare up at the local level throughout the 1930s, and even after the moderate, reportedly pious Manuel Ávila Camacho became president in 1940 (Smith, 2012).

Cardenismo

The presidency of Lázaro Cárdenas (1934–40) has long been a focal point for different interpretations of the revolution, particularly its relationship with popular groups. Cárdenas emerged in populist accounts as the culmination of revolutionary radicalism, he broke the backstage power the Sonoran clique, oversaw a sweeping land reform, supported organized labor, struck a blow for economic sovereignty by nationalizing the oil industry in 1938, and turned the official party from a talking shop for the revolutionary elite into a mass, corporate party with sectors for peasantry, workers, middle classes, and the military. This narrative was challenged by revisionists, who portrayed Cardenista reform as a sophisticated attempt to divide and co-opt the masses, deepen the internal market, and consolidate capitalist development. What many accounts had in common was the tendency to see Cardenismo creating a powerful, centralized, and stable state. Knight (1994) provided what has proven to be a highly influential synthesis; he argued that Cárdenas' project represented a radical attempt to create a more equitable and inclusive society (and was correctly understood as such at the time), but was limited by state weakness and growing political opposition. Many of these arguments have found support in later studies of the right-wing and regional opposition (Fallaw, 2001). In the 1930s, federal public education was shaped by regional power brokers, and by peasant communities themselves, who rejected or adapted aspects of the state's message of anticlericalism, class consciousness, and social reform. However, in some communities at least, public education did help stabilize Mexican politics by encouraging people to use a shared revolutionary discourse and party institutions to make demands of the government (Vaughan, 1997).

International Relations

The revolution was nothing if not a defense of national sovereignty, and an extensive literature has developed on Mexico's relations with the rest of the world. During the armed phase of the revolution, the US government repeatedly intervened in Mexico. However, historians have disagreed about how much emphasis to place on the US role. For Hart (2002), the expansion of US capitalism during the Porfiriato was a central cause of the revolution, while US interventions were decisive in defeating the revolution's more radical factions. However, economic nationalism rarely excited fierce popular support, and the US government's vacillating and counterproductive interventions were less important than domestic conflicts in shaping the outcome of the revolution's armed phase. The US ambassador to Mexico, Henry Lane Wilson helped orchestrate the counterrevolutionary coup of 1913, although he acted on his own initiative and infuriated President Woodrow Wilson. From April to November 1914, US marines occupied the port of Veracruz, preventing the arrival of arms shipments to Huerta. The US government then cultivated relations first with the Villistas, and then the Constitutionalists, and in 1916–17, sent a punitive expedition into northern Mexico to pursue the Villistas; the expedition failed and further antagonized the Constitutionalists who saw it as an affront to Mexican sovereignty (Knight, 1988). After 1920, tensions with the US government focused on the oil industry, debt, and reparations for damages. The Bucareli Accords of 1923 settled these issues and allowed a modus vivendi to resume, although tensions flared again periodically when Calles offered support for the Nicaraguan nationalist Augusto Sandino, and again in 1938, when Cárdenas nationalized the oil industry. After the 1940s, as revolutionary radicalism waned, relations between the US and Mexican governments became notably cordial. During the Second World War, Mexico provided mineral resources and bracero labor to fuel the war effort, although some tensions remained over the terms military collaboration. While US–Mexican relations have dominated the field, Katz (1981) pioneered a more multilateral approach. From the 1910s to the 1940s, European powers jockeyed for influence in Mexico and Mexico sometimes succeeded in using this competition to deflect US influence (Schuler, 1999). In recent decades, historians have also sought to move the focus away from diplomacy, and incorporate culture, tourism, and travel into the study of Mexico's international relations (Delpar, 1992).

Indigenismo

The revolutionary regime's new stance toward indigenous peoples fascinated foreign observers in the 1920s and the 1930s and has provided another enduring theme in the historiography. Much of the literature has analyzed indigenismo – the revalorization of indigenous culture – as an

artistic and esthetic phenomenon. In the nineteenth century, Mexican liberals sometimes celebrated the achievements of old pre-Colombian civilizations, but they generally disdained or ignored contemporary Indian peoples and cultures. The revolutionary regime sponsored the murals of Diego Rivera replete with new images of indigenous history and customs and encouraged a vogue for indigenous folk art and music. Other studies have focused on social policies, particularly education, and their reception in indigenous communities. Interpretations of official indigenismo have echoed broader shifts in historiography, ethnography, and political science. In the 1970s, Mexico witnessed a major backlash against official indigenismo as a new generation of anthropologists, indigenous intellectuals, and leaders condemned it as a thinly disguised project of assimilation. This was in many ways a powerful critique; beyond the murals and exhibitions, indigenous-state relations could still be conflictive; in 1927, the Sonoran regime used its new military aircraft to drop bombs on rebellious Yaqui Indians; by the 1940s, assimilation had clearly became the objective of official indigenous policy. However, indigenismo was, initially at least, a multifaceted phenomenon, shaped by the press, artists, the state, transnational intellectuals, and even military leaders. It consequently contained divergent attitudes and ideas, and sometimes allowed indigenous leaders to make political demands and claim material benefits in new ways (Lopez, 2010; Dawson, 2004).

Women and Gender History

The past two decades have also seen a steady growth in studies of the revolution's effects on women and gender norms. Scholars traditionally viewed the revolution as a quintessentially macho event, a revolution by and about men. Historians first sought to recover the history of Mexico's early feminist movement, which received support from some Constitutionalist leaders such as Felipe Carillo Puerto, and women's varied roles in revolutionary warfare as soldaderas: camp followers, spies, sometimes soldiers, and officers. Scholars have now broadened the focus, exploring women's experiences in the home, in the schoolhouse, on the factory floor, and in the red-light zone. They have also charted the unprecedented growth in the 1920s and the 1930s of women's political organizing, both within dissident Catholic and revolutionary groups. Women's political objectives and tactics varied; some emphasized political rights, while others emphasized social citizenship; some groups demanded rights on the basis of equality, while others adapted maternalist arguments based on women's essential qualities. The overall picture remains one of enduring patriarchy, albeit a patriarchy that was reshaped – or modernized – in important ways. Women's organizing shaped the design and implementation of new social and welfare policies, was central to social movements across political spectrum, secured new administrative and political posts for women, and eventually led to female suffrage in 1953 (Olcott et al., 2006; Mitchell and Schell, 2007). While new studies reveal a great deal about gender relations in the period, the study of men as gendered subjects remains relatively underdeveloped.

The Pax Priísta: 1940s–60s

The two decades after 1940 have usually been seen, for better or worse, as the heyday of the new regime's stability and control – a Pax Priísta. In the past decade, as archives have opened, and the weakness of the Cardenista state has became more apparent, historians have increasingly turned their attention to this period. Few challenge the idea that government radicalism petered out after the late 1930s. However, new studies have given us a much better understanding of the kind of political system the revolution ultimately bequeathed. Before the past decade, most accounts of the 1940s and the 1950s were written by political scientists, who presented synchronic analyses of a stable political system centered on the corporate party and a mighty presidency (Schmidt, 2001). In the past decade, historians have increasingly emphasized the limits of state power: state institutions were riddled with caciquismo; president and party had to contend with regional business interests, independent military officers, and an enduring tradition of popular protest against everything from corruption and taxation, to electoral fraud and conscription. There has been a related renewal of interest in political violence in midcentury Mexico. While older studies celebrated how the regime professionalized the rebellious army of the 1920s and pushed it back into the barracks, new accounts show the continuing importance of the military in policing and repression. Regional studies show that politics under the PRI were often a rough business, punctuated by pistolerismo, collective repression, riots, and small insurgencies and counterinsurgencies (Rath, 2013; Gillingham and Smith, 2014). While this work on the Pax Priísta helps us understand how the revolution came to an end, it has also been influenced by shifting perceptions of contemporary Mexico, as the drug war has encouraged scholars to reexamine earlier assumptions of stability and control.

Historical Memory and Conclusions

Indeed, the debate over the historical meaning of the revolution has been (and remains) an integral part of Mexican politics. The scholarly revisionism in the 1970s had long roots. In the novels of Azuela (2008) and the murals of Orozco, the revolution emerges as chaos, violence, and hypocrisy; from the very beginning, conservative Catholic critics of the revolution abounded, while many others criticized the revolutionary party's betrayal of liberal democracy, such as José Vasconcelos during his 1929 presidential campaign. Confronting these critical voices, the new regime made great efforts to create an official myth of the revolution as a unifying event that brought Mexico peace and justice under the auspices of the revolutionary state. The government disseminated its version of history through school textbooks and the mass media, censored the national press, and created a plethora of new holidays and parades such as Revolution Day on 20 November. While Zapata, Villa, Carranza, and Cárdenas all had their differences while alive, they were buried together at the huge Monument to the Revolution in downtown Mexico City (Benjamin, 2000). Of course, we cannot assume that people swallowed the stuff of official myths whole, although the

history of how people responded to official myths is difficult to write and few studies have attempted it.

Many observers argue that whatever credibility and appeal this official version of history enjoyed was ebbing away by the 1980s. And yet, the revolutionary tradition was not entirely spent. While the PRI gradually began losing control over politics and public discussion, opposition groups continued to criticize the regime in the name of the revolution. In 1987, a group of left wingers split from the PRI, led by Lázaro Cárdenas' son Cuauhtémoc, and claimed the heritage of the radical 1930s. In 1994, peasants in Chiapas rebelled against government corruption and the end of land reform, and called themselves Zapatistas (O'Toole, 2011). In 2000, the PRI lost the presidency to the center-right Party of National Action (Partido Acción Nacional, PAN). The PAN would probably rather have ignored revolutionary history altogether, but in 2010, they found themselves in charge of organizing public commemorations of the centenary of the revolution. They responded by creating public ceremonies, speeches, dances, and light shows, which celebrated aspects of the revolution they liked (liberalism, Madero) and elided those they did not (agrarianism, anticlericalism). During the 2012 election campaign, the PRI at once claimed to have broken with the corruption and illiberal errors of the past, and invoked its reputation for stability and constructive material development; PRI posters across Mexico City asked citizens to remember which party had brought peace to Mexico and built the subway system.

Scholarship on the Mexican Revolution has been shaped by new empirical data, new methodologies – particularly the turn to regional history – and changing theoretical perspectives, and we know a great deal more about revolution in all its complexity than we did two decades ago. However, the topic is still far too recent, with far too profound a legacy to be only of scholarly interest. With the return of the PRI to power in 2012, the intimate connection between contemporary political debate and the questions people ask about the revolution and its history is likely to continue for some time to come.

See also: Bourgeoisie and Middle Classes, History of; Capitalism; Citizenship, Historical Development of; Civilization, Concept and History of; Contemporary History; Crowds in History; Cultural History; Democracy, History of; Dictatorship in History; Ethnicity and Ethnic Groups: Historical Aspects; Gender History; Historical Thought and Historiography: Current Trends; History and the Social Sciences; History: Theories and Methods; Industrialization, Typologies and History of; Labor History; Liberalism: Historical Aspects; Linguistic Turn and Discourse Analysis in History; Military History; Modernization and Modernity in History; Nationalism, Historical Aspects of: The West; Nations and Nation-States in History; Peasants and Rural Societies in History (Agricultural History); Political History: History of Politics; Political Parties, History of; Religion, History of; Social Inequality in History (Stratification and Classes); Social Movements, History of: General; Socialism: Historical Aspects; State, History of; Subaltern History; Time, Chronology, and Periodization in History; Working Classes, History of.

Bibliography

Azuela, M., 2008. The Underdogs, new edition. Penguin Classics, London.
Baily, D.C., 1978. Revisionism and the recent historiography of the Mexican revolution. Hispanic American Historical Review 58 (1), 62–79.
Beezley, W., 2007. Reflections on the historiography of twentieth-century Mexico. History Compass 5 (3), 963–974.
Benjamin, T., 2000. La Revolución: Mexico's Great Revolution as Memory, Myth, and History. University of Texas Press, Austin.
Brading, D.A. (Ed.), 1980. Caudillo and Peasant in the Mexican Revolution. Cambridge University Press, Cambridge.
Butler, M. (Ed.), 2007. Faith and Impiety in Revolutionary Mexico. Palgrave, Basingstoke.
Dawson, A., 2004. Indian and Nation in Revolutionary Mexico. University of Arizona Press, Tucson.
Delpar, H., 1992. The Enormous Vogue of Things Mexican: Cultural Relations between the United States and Mexico. University of Alabama Press, Tuscaloosa.
Fallaw, B., 2001. Cárdenas Compromised: The Failure of Reform in Postrevolutionary Yucatán. Duke University Press, Durham.
Gillingham, P., Smith, B. (Eds.), 2014. Dictablanda: Soft Authoritarianism in Mexico, 1940–1968. Duke University Press, Durham.
González y González, L., 1974. San José de Gracia: Mexican Village in Transition. University of Texas Press, Austin.
Hart, J.M., 2002. Empire and Revolution: The Americans in Mexico Since the Civil War. University of California Press, Berkeley.
Joseph, G., Deans-Smith, S., 1999. Mexico's New Cultural History: ¿Una Lucha Libre? Special Edition of Hispanic American Historical Review. Duke University Press, Durham.
Joseph, G., Nugent, D. (Eds.), 1994. Everyday Forms of State Formation: Revolution and the Negotiation of Rule in Modern Mexico. Duke University Press, Durham.
Katz, F., 1981. The Secret War in Mexico: Europe, the United States and the Mexican Revolution. University of Chicago Press, Chicago.
Katz, F., 1998. The Life and Times of Pancho Villa. Stanford University Press, Stanford.
Knight, A., 1988. U.S.-Mexican Relations, 1910–1940: An Interpretation. Center for U.S.-Mexican Studies, San Diego.
Knight, A., 1994. Cardenismo: juggernaut or jalopy? Journal of Latin American Studies 26 (1), 73–107.
Lopez, R., 2010. Crafting Mexico: Intellectuals, Artisans, and the State after the Revolution. Duke University Press, Durham.
McNamara, P., 2007. Sons of the Sierra: Juárez, Díaz, and the People of Ixtlán, Oaxaca, 1855–1920. University of North Carolina Press, Chapel Hill.
Meyer, J., 1976. The Cristero Rebellion: The Mexican People between Church and State. Cambridge University Press, Cambridge.
Mitchell, S., Schell, P. (Eds.), 2007. The Women's Revolution in Mexico, 1910–1953. Rowman and Littlefield, Plymouth.
Olcott, J., Vaughan, M.K., Cano, G. (Eds.), 2006. Sex in Revolution: Gender, Politics, and Power in Modern Mexico. Duke University Press, Durham.
O'Toole, G., 2011. The Reinvention of Mexico: National Ideology in a Neoliberal Era. Liverpool University Press, Liverpool.
Rath, T., 2013. Myths of Demilitarization in Postrevolutionary Mexico. University of North Carolina Press, Chapel Hill.
Riguzzi, P., 2009. From globalization to revolution? The porfirian political economy: an essay on issues and interpretations. Journal of Latin American Studies 41 (2), 347–368.
Schmidt, A., 2001. Making it real compared to what? Reconceptualizing mexican history since 1940. In: Joseph, G., Rubenstein, A., Zolov, E. (Eds.), Fragments of a Golden Age: The Politics of Culture in Mexico Since 1940. Duke University Press, Durham, pp. 23–68.
Schuler, F., 1999. Mexico between Hitler and Roosevelt: Mexican Foreign Relations in the Age of Lázaro Cárdenas, 1934–1940. University of New Mexico Press, Albuquerque.
Smith, B., 2012. The Roots of Conservatism in Mexico: Catholicism, Society, and Politics in the Mixteca Baja, 1750–1962. University of New Mexico Press, Albuquerque.
Vaughan, M.K., 1997. Cultural Politics in Revolution: Teachers, Peasants, and Schools in Mexico, 1930–1940. University of Arizona Press, Tucson.

Microeconomics, History of

Paola Tubaro, University of Greenwich, Greenwich, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides a brief overview of the history of microeconomics, starting from the late eighteenth century when its key foundations were laid. Without intending to be comprehensive, the article outlines landmarks and milestones in the building of the basic principles of today's knowledge, highlighting challenges faced and open questions. Focus is on the contribution of each historical phase to the understanding of two foundational issues in microeconomics – the theory of individual economic decision making and the market mechanism as a coordinating device. At the end, the article discusses the current state of microeconomics and recent tendencies.

Introduction

Microeconomics stands as one of the two main pillars of today's economic science, the other being macroeconomics. Microeconomics seeks to understand how individuals, households, and companies make decisions; the factors that shape these decisions; and how these decisions affect others. Its object is decision making at low, or micro, level, in contrast to macroeconomics, whose focus is the behavior of national aggregates such as gross domestic product, inflation, and unemployment.

At the heart of microeconomics is a social and scientific 'mystery.' In modern market economies characterized by private property and freedom, decisions to produce, sell, work, buy, and consume are made by multiple individuals, households, and companies autonomously and independently from one another; each actor chooses under limited, or even inexistent, knowledge of the choices of others. Yet someone's decision to buy, say, a pint of milk can only translate into an actual purchase if someone else simultaneously decides to sell a pint of milk. So if buyer and seller do not coordinate *ex ante*, how can their decisions engender actual, mutually beneficial transactions, rather than result in disappointment and frustration? The market mechanism is the device that accomplishes this result: by establishing the relative price of goods and factors, it informs individual decisions, adjusts the quantities traded, and achieves *ex post* coordination. Microeconomics, through the supply-and-demand framework of analysis, looks at how this happens. It also studies the effects of institutions such as government regulation and taxation on individual choices and market outcomes.

In essence, microeconomics is as old as economic thought itself: Classics in the eighteenth and nineteenth centuries laid out the theoretical elements that have progressively molded economists' understanding of individual decision making and the market mechanism. Yet the 'micro' label and the representation of microeconomics as a separate enterprise, explicitly distinct from macroeconomics, emerged much later. According to Humberto Barreto (2012), its first traces are in a 1933 article by Ragnar Frisch for whom the micro approach would "try to explain in some detail the behavior of a certain section of the huge economic mechanism," while the macro would instead "give an account of the fluctuations of the whole economic system taken in its entirety" (Frisch, 1933: p. 2). The work of John M. Keynes (1936) in those same years, with its strong focus on aggregate relations and an extremely simplified representation of individual behaviors, popularized the belief that the macro level of analysis differs in nature from the microlevel, and contributed to further differentiating the two subdisciplines.

Since then, it took another 50 years before economists formally acknowledged the micro–macro taxonomy. In the classification of subject fields of the *Journal of Economic Literature*, a reference for the profession worldwide, micro- and macroeconomics were subsets of a broad 'theory' category in the original 1960 version, and were brought up a level and highlighted as major independent categories in a substantial revision of the system in 1991 (Barreto, 2012). They remain two major areas today, allowing the development of new or renewed subfields within them. The teaching of economics has followed suit, with current curricula typically separating micro- and macroeconomics courses, both being part of the core training of new economists.

Perhaps ironically, a parallel movement has been increasingly blurring the distinction between the micro and macro approaches. Economists have grown more and more uncomfortable with the idea that different theoretical principles apply to the study of individual decisions and the aggregates that (ultimately) result from them. Since the late 1980s the 'microfoundations' movement in macroeconomics, championed by Robert Lucas (born 1937), aims to reconcile these two traditions and rebuild macroeconomics so as to make it consistent with the principles of microeconomics.

This article provides a brief overview of the history of microeconomics, starting from the time in which its key foundations were laid, although the 'micro' terminology was not yet in use. Despite the interest of the early literature, which dates as far back as antiquity and the Middle Ages (see, e.g., Hutchison, 1988), a detailed account of it would be beyond the scope of this article, and the more traditional convention of starting from the late eighteenth century is followed. The article overviews historical developments since then, without intending to be comprehensive, but aiming only to outline landmarks and milestones in the building of the basic principles of today's knowledge, highlighting the challenges faced and the questions that remain open. Focus is on what each

historical phase contributed to the understanding of the two foundational issues in microeconomics – namely, the theory of individual economic decision making and the market mechanism as a coordinating device. At the end, the article discusses the current state of microeconomics and recent tendencies, again distinguishing between the study of individual behaviors and the market mechanism.

The Origins: Self-Interest and the 'Invisible Hand' in Adam Smith

The Scottish philosopher Adam Smith (1723–90) is widely regarded as one of the founders of economics; though not alone in this enterprise, he exemplifies at best the novelty and relevance of eighteenth-century thought in defining what would become the basic principles of microeconomics. In his *Wealth of Nations* (1776), Smith was the first to explicitly characterize individual economic behavior as *self-interested* behavior, admitting that it is people's desire for a gain that explains work, production, and ultimately the existence of an economic system:

> It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love, and never talk to them of our own necessities but of their advantages.
> (Smith, 1776/1981: pp. 26–27)

The idea that businesspeople act to satisfy their self-interest may seem dauntingly simple, and to some extent brutal in its apparent amorality. Yet over time, it has proven fertile in informing economists' imagination and their capacity to explain individual decision making. It is not a sheer defense of selfishness – in present-day parlance, one would simply say that people respond to *incentives*, that is, advantages or gains that can derive from given decisions or actions. By and large, modern microeconomics assumes that individuals' motivations when they make decisions are their expected gains by those decisions.

But Smith, a moral philosopher, could not neglect motivations other than self-interest in shaping economic decision-making. Indeed his self-interest must be understood in conjunction with another behavioral principle, that of *sympathy* between human beings who, by putting themselves imaginatively in the place of others, understand their feelings and expectations and are moved to act accordingly (e.g., to give to those who are in need). While Smith's critics initially opposed sympathy and self-interest, today's scholarship sees them as two instances of a common framework for thinking about human behavior, in which the individually centered, self-interested component is accompanied by an interpersonal dimension, so that it becomes possible to account for various forms of behavior, from gain-seeking actions to philanthropy. Reconciliation of these two aspects of Smith's thought makes him the father of microeconomics in a broader sense: A recent trend in today's research on economic decision making (see below) tends to explicitly integrate forms of prosocial behavior into the analysis.

Incidentally, the Smithian model of individual behavior had the merit of supporting an egalitarian perspective in the nineteenth-century political debates. If the same scheme holds for all individuals, they are all equal and have the same capacity for decision making; observed differences, if any, depend on chance and history. This egalitarian view was shared by most economists of the time and strongly contrasted the (then also widespread) hierarchical stance that regarded the lower classes of society and supposedly 'inferior' ethnic/racial groups as less capable of making decisions and thus in need of guidance (Peart and Levy, 2005). Interestingly, the infamous 'dismal science' designation of economics was originally an accusation against economists' antislavery orientation, resulting from their belief that all humans are equal (Levy, 2001).

Smith's work also contributed to shaping economists' view of the market as a coordinating device that ensures, *ex post*, that individual decisions are consistent with one another and generate an orderly result. Smith's 'invisible hand' metaphor has often been recognized as an effective representation of this mechanism:

> by directing […] industry in such a manner as its produce may be of the greatest value, he [man] intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention. By pursuing his own interest he frequently promotes that of the society more effectually than when he really intends to promote it.
> (Smith, 1776/1981: p. 456)

Because individuals are not isolated but part of a larger human community, their actions have unexpected or unintended consequences at the system level. Individuals take into account only their self-interest, yet their choices affect others and trigger a chain of interactions that eventually affect the society as a whole, well beyond their original intentions. Strikingly enough, Smith argues that this spontaneous process does not lead to chaos but to harmony: Self-interest may not seem a noble motivation, yet it triggers consequences that benefit society even more than those arising from benevolence. Thus, there is no need for a strong state power that would impose social order from above, as had been argued by Thomas Hobbes (1588–1679). The idea of unintended consequences and the possible reconciliation of individual self-interest and social good, first articulated by Smith, have been at the core of subsequent microeconomic reflection – which is another reason why Smith is credited as a founder of the discipline.

Early Developments: Marginalism, Supply-and-Demand, Formalization

To further develop Smith's approach, economists faced the formidable problem of how to operationalize the principle of self-interested action so as to apply it to the solution of concrete economic problems. They first thought about it in terms of the self-interested businessperson who makes production-related decisions: In his *On the Principles of Political Economy and Taxation* (1817/2004), the highly influential British writer David Ricardo (1772–1823) discussed how cost-effective substitution of machinery for workers, or of imports for nationally produced

goods through international trade, affects a firm's profits and workers' welfare. Ricardo's writings hardly touched on the study of household and consumer behavior, though: As many economists of the time, he was mainly interested in production forces, working conditions, and their differential impact on income distribution. Emphasis was on the effects of the whole economic system on individual behavior, rather than the reverse, and a fundamental heterogeneity distinguished economic agents – consumers vs firms. While the latter were already envisaged as cost-minimizers or profit-maximizers, placing emphasis on how they acted for their own advantage, less attention was drawn to the individual decisions of the former, and demand was mainly seen as an aggregate rather than an individual construct, for example, by assuming that population growth largely determines demand for food.

Even the authors who, following developments in philosophy and psychology, introduced the notion of 'utility' to account for the pleasure and pain that consumers derive from goods, found it difficult to integrate it into economic models. Utility initially appeared as a qualitative, subjective construct devoid of any objective, quantifiable attributes – very different then from the firm's profit or income, which centuries-long advances in accounting enabled to measure in straightforward way. The solution came in the mid-nineteenth century, from reinterpreting utility not as an absolute but as a relative magnitude, varying from one individual to another and for each individual, depending on the available quantity of a good. It turned out that measurability and quantification were not even necessary, insofar as utility could be expressed as a variable or a function of other variables, allowing comparisons (larger/smaller, more/less, positive/negative, etc.). One could thus distinguish the total amount of utility from 'marginal' utility – namely, the change in the level of utility that results from a given increase in the quantity of the good. Marginal utility was thought to diminish with the quantity consumed, reflecting the capacity of individuals to order the possible uses of successively acquired units: For instance, one would reserve the first gallons of water for drinking and the successive ones for personal hygiene, for housekeeping, and finally for watering plants. In passing, this assumption solved what earlier thinkers considered a paradox – the fact that useful goods such as water or air have low market value: The reason is their abundance, which means that the last increment in quantity generates an extremely small increase in utility. These results suggested an interpretation of self-interested behavior in terms of attempts to raise one's utility to its highest possible level and were obtained independently, in 1871–74, by William S. Jevons (1835–82) in Britain, Carl Menger (1840–1921) in Austria, and Léon Walras (1834–1910) in Switzerland.

The importance of thinking in terms of marginal variations rather than total magnitudes proved so useful to account for utility and demand that it was subsequently extended to supply. In fact, notions of marginal productivity and marginal cost of production, as opposed to total productivity/cost, had already been introduced, but in a patchy fashion; they were refined and generalized in the 1890s by, among others, John B. Clark (1847–1938), Philip H. Wicksteed (1844–1927), and Knut Wicksell (1851–1926). Marginal reasoning seemed so important that the economic thought of this time period is often referred to as *marginalism*.

On the coordinating mechanism, emphasis was increasingly placed on the symmetry of supply-and-demand factors and on the resulting market equilibrium. While Ricardo had stressed supply factors in the determination of prices and quantities traded, the mid-nineteenth century developed a view according to which individual demand and supply are derived, respectively, from agents' calculations of utility (for consumers) and profit/cost (for firms), and market supply and demand are obtained by aggregating all individual values. When market supply equals demand, the market is in equilibrium – that is, the decisions of all households and all firms are consistent with one another. The French mathematician Antoine-Augustin Cournot (1801–77) was the first, as early as 1838, to illustrate these principles with the help of price–quantity diagrams in which the demand of a good decreases, and the supply increases, with its price; The intersection of the supply-and-demand schedules identifies equilibrium, that is, a price at which supply equals demand and the market clears. This 'partial equilibrium' approach, focusing on the study of a single competitive market, was further developed and popularized a few decades later in England by Alfred Marshall (1842–1924); it is now a standard part of undergraduate microeconomics education.

The advantage of Marshall's partial equilibrium approach is that it provides a tractable framework to study the relationship between price and quantity; however, it is based on the restrictive assumption that changes in the price of a good affect the quantity of that good only, disregarding the possibility that a variation in the price of a good will impinge on the demand and supply of substitutes and complements of that good too. Hence, it can be taken at most as an approximation, not as a rigorous analytical device. A rival view focusing on interdependencies among markets was simultaneously developed by Walras (1874/1977), who represented economic agents that allocate their budgets to the purchase of multiple goods so that changes in the market price and/or quantity of one good are likely to have repercussions on the markets for other goods. His notion of 'general equilibrium' which would inform large parts of microeconomics, is directly derived from this view and corresponds to a situation in which supply equals demand on each market, so that all clear simultaneously.

A derived question is whether and how actual trade practices will drive prices and quantities toward equilibrium. Marshall and Walras provided different answers, consistent with their respective views. Walras (1874/1977) proposed a model of auctions in which at given prices, all traders declare the quantity of each good that they wish to buy or sell at those prices; if with these quantities, supply equals demand on each market, then this is the general equilibrium and trade takes place; if not, prices are adjusted in such a way that they diminish where supply exceeds demand and increase in the opposite case; at the new prices, traders announce again the quantities that they wish to buy or sell, and the process ('tâtonnement') starts again, until equilibrium is reached. In short, transactions take place simultaneously, at equilibrium only, so that the same prices apply to all traders. Instead, Marshall (1920) had in mind a sequence of bilateral transactions on a single market, in which each pair of traders negotiates a price and each transaction withdraws some units from the market so that lesser quantities are available for later

trades – in other words, the conditions under which traders negotiate are altered at every step. Such changes gradually dampen price adjustments until they reach the level that corresponds to the intersection of supply and demand. Here, transactions occur sequentially, in disequilibrium, at prices that may differ from a pair of traders to the other. These different approaches strongly affected later developments (see below).

These theoretical developments, often referred to as 'neoclassical' economics (in contrast to Ricardo's 'classical' theory), accompanied a growing tendency to *formalization*. Deeply intertwined with the development of microeconomic thought since its early phases, formalization is an approach to theory-building which, starting from the definition of theoretical constructs and of linkages between them, derives logical conclusions from given premises. In this sense, formalization is an abstract process and must be distinguished from *quantification*, or the systematic effort to measure reality for purposes of business or policy intervention, rather than theory-building, for instance by counting the population in a census. While quantification was already well under way at this time, it was formalization that needed to be advanced. Writing in its defense, Jevons claimed that "The data are almost wholly deficient for the complete solution of any one problem," yet "we have mathematical theory without the data requisite for precise calculation" (Jevons, 1965/1871). Yet despite its growing importance, formalization was still far from being prominent at the end of the nineteenth century, when many economists still privileged the verbal form and some overtly criticized the use of mathematics, the most prominent example being the German Historical School. It was only in the mid-twentieth century, after the emergence of microeconomics as a distinct theoretical enterprise, that formalization became a standard practice.

Maturity: Constrained Optimization and Market Equilibrium

The theoretical elements laid out in the nineteenth century went through further elaboration and systematization in the first half of the twentieth century, when neoclassical microeconomics blossomed and eventually arose as a major subdiscipline in conjunction with macroeconomics. The concept of utility was refined and utility gradually came to be understood as rational, purposeful behavior – choosing the best possible means to achieve one's ends, an interpretation that deprives utility of any identification with self-interest narrowly interpreted, and even with pleasure, pain or any other emotions and feelings. Consistency in choices turned out to be all that was needed, and the microeconomic paradigm that eventually emerged refrained from making psychological assumptions. This more neutral interpretation of utility came to dominate economics, notably through the *revealed preference* approach (Samuelson, 1938) which enabled application of the model to a variety of settings and behaviors (including non-selfish ones and altruism).

Consistency involves the optimizing behavior of agents under some constraints (budget for consumers, technology for firms). Progressively, the constrained maximization model was extended to the study of all individual decision units, on both the demand and the supply side of the market, becoming the basis of all analyses of individual economic behavior. The contribution of Paul Samuelson (1915–2009) was essential to these developments and mainly consisted in rewriting many problems of economics as maximization problems, with extensive use of mathematics. Samuelson's effort to show that apparently diverse subjects have the same underlying structure and can be treated with the same mathematical tools defined the core of what came to be known as microeconomics, and gave it unprecedented unity and coherence. Higher abstraction and technical difficulty, however, were the price to pay for increased scope and greater logical coherence – making the field inaccessible to wide audiences and hard to communicate outside a narrow circle of specialists.

With time, these revisions proved successful in that they made constrained maximization so powerful that it began to be thought of as a basis for understanding all human behaviors – not just those traditionally regarded as falling within the realm of market transactions. The work of Gary Becker (1930–2014), in particular, is founded on the idea that individual behavior follows the same fundamental principles in different areas, so that the same explanatory model – microeconomic optimization – should always be applicable. Becker thus extended the sphere of microeconomic analysis to the study of new areas, ranging from crime and discrimination to marriage and the family (Becker, 1992). Becker's analysis initially raised many eyebrows, but gradually came to exert a huge influence on the economics profession as well as on neighboring disciplines.

The optimization model was not beyond dispute, though: At least since the 1950s, Herbert Simon (1916–2001) and others contended that actual decision makers lack the cognitive capacities to solve maximization problems and rather content themselves with 'satisficing' behavior, choosing options that are not optimal but make them happy enough. Along these lines, they developed a 'bounded rationality' approach as an alternative to the seemingly strong rationality requirements of the individual maximization model. Yet it was difficult to identify universal principles enabling to devise general models of nonmaximizing behavior, so that bounded rationality assumptions often maintained an ad hoc, hardly generalizable character. Partly for this reason, the optimization model retained its prominence for long.

More recently, a current of thought known as 'behavioral economics' has provided abundant evidence against the individual optimization model. Starting from a seminal contribution (1979) by Amos Tversky (1937–96) and Daniel Kahnemann (born 1934), it has been shown that people make decisions based on simplified heuristics such as routines and habits rather than rigorous logic, and use mental emotional filters to respond to changing circumstances. In contrast to the microeconomic paradigm that aims at a value-neutral, highly abstract representation of decision making, behavioral economics reintegrates psychological elements into the analysis, explicitly investigates pro-social behavior, and compares and contrasts it with self-interest and selfishness – renewing, in a sense, with the older tradition initiated at the time of Adam Smith.

Regarding market models, after World War II it was the general equilibrium model that most attracted the attention of economic theorists, largely due to the introduction into

economics of highly advanced formal tools and of a new way of thinking about mathematics, based on topology, convex set theory, and fixed-point theorems (Weintraub, 2002). The new tools allowed for a sophisticated refinement of Walras's approach, named the Arrow–Debreu–McKenzie model after its main contributors. A major achievement in the 1950s was a formal proof of existence of equilibrium. The difficulty was that it was not enough to show that the system of simultaneous equations representing equality between supply and demand in all markets has a solution: For this solution to be meaningful economically, and not only mathematically, it was also necessary to prove that equilibrium prices and quantities are nonnegative. By demonstrating that it is indeed the case, it was established that the notion of a set of prices that clear all markets is consistent (i.e., the notion of equilibrium of a system of interrelated competitive markets is not void).

Another success for general equilibrium theory was the mathematical proof of the so-called two theorems of welfare economics – a modern reinterpretation of Smith's 'invisible hand.' The first theorem states that a general equilibrium corresponds to a socially optimal allocation of resources, and the second states that, under some conditions, any socially optimal allocation of resources can be sustainable by a general equilibrium. These results amounted to rigorously establishing the desirable properties of the free market mechanism that earlier economists like Adam Smith had put forward only intuitively or metaphorically. In sum, the new approach completely transformed general equilibrium theory, allowing to make it "rigorous, to generalize it, to simplify it, and to extend it in new directions" (Debreu, 1984: p. 267).

In the 1950s and 1960s, such progresses put the Arrow–Debreu–McKenzie model at the center of the stage and increased confidence in its potential to provide the whole of economics with rigorous mathematical foundations. However, problems started with attempts at proving two other key properties of equilibrium – namely, stability and uniqueness. The question of stability was meant to ensure that after an exogenous shock, the market mechanism is capable of generating endogenous forces that bring it back to equilibrium; if equilibrium exists but the market cannot find it, then arguments for free markets are harder to make. In addition, if uniqueness is not guaranteed, it is unclear where an adjustment process might drive the system after a shock; besides, some equilibriums may be unstable. It became soon clear, though, that formal proofs of stability and uniqueness could be obtained only under excessively restrictive, unrealistic assumptions. Critics stressed that these results reveal that with all its mathematical underpinnings, general equilibrium theory did not truly succeed in improving knowledge of how the market mechanism works and how prices adjust in response to variations of supply-and-demand conditions (Kirman, 1989; Ingrao and Israel, 1990). Despite these shortcomings, general equilibrium theory remained central for long, and is still in the core of economists' education; the publication in 1995 of a celebrated graduate textbook by Mas-Colell et al. (1995) condensed and systematized the essence of the knowledge available at the highest level.

At this point, the consolidation of the microeconomic paradigm founded on constrained optimization of the individual economic agent and the general equilibrium of all markets appeared as a monolithic bloc – a standard, more and more firmly established 'mainstream' allowing less and less different perspectives. The micro-foundations movement, in its effort to integrate core microeconomic principles into macro, seemed to further reinforce the mainstream's primacy. Microeconomics seemed at the same time successful, and stagnating, unable to move any further.

Yet a closer look reveals that the evolution of microeconomics in the decades that followed World War II was more diverse than may seem at first sight. Alternative and heterodox economic theories, explicitly challenging neoclassical microeconomics, have never ceased to exist even though their place, relevance, and visibility seemed to somehow shrink over time. A complete review of their approaches is beyond the scope of this article; But it is worth briefly mentioning two key developments which, initially not among the most vocally critical of the state of microeconomics, have been highly successful and over time, have progressively grown to be integrated into the mainstream as it is today. First, the general-equilibrium image of individuals making decisions in isolation from one another, letting an impersonal market mechanism establish the prices that will guide their decisions, progressively gave way to a more active view of decisions as strategic reactions to other agents' own actions. In contexts in which interdependencies are strong, such as imperfectly competitive markets, economic agents anticipate the direct repercussions of their choices on others, and of others' choices on themselves. Game theory has proven to be the appropriate technique to explore the structure of strategic interactions and the direct effects of individual decisions on one another. Initiated in the 1940s with, among others, a seminal contribution of Oskar Morgenstern and John von Neumann (1944), game theory experienced a rapid development and an extraordinary expansion of its application to economics since the 1980s.

Second, Marshall's partial equilibrium remained in the shadow for a while, largely due to a devastating critique from Piero Sraffa (1925) who questioned its logical consistency; But it did not entirely disappear. The static supply-and-demand diagram was used in some applied studies, while the dynamic conception of sequential transactions between pairs of buyers and sellers, underdeveloped in Marshall's own time, was refined with the help of novel, nonmathematical techniques. It underpinned the pioneering laboratory experiments on markets that were implemented by Edward Chamberlin (1948) and Nobel Prize Vernon L. Smith (1962). Although experimental research remained a tiny niche within economics for long, it kept alive the Marshallian approach, in its dynamic connotation. Partial equilibrium experimental markets gained renewed interest in the 1990s, and the properties of the exchange dynamics that they generate are sometimes referred to in the literature as 'Marshallian dynamics' (Friedman and Rust, 1993).

Interestingly, the first experimental markets inspired by Marshall's dynamic orientation were being devised at the same time in which set-theoretical mathematical techniques were being implemented to model Walrasian systems. These tools were rudimentary at the beginning (Chamberlin simply relied on his students and a blackboard to perform his market experiment), but they became more sophisticated over time with the increased availability of software for computer-based

experiments, and are thriving today. They embody a view of economics that gives less weight to the logical rigor of formalization than to the theory's fit with empirical data, assessed with objective, well-defined, and replicable (laboratory-based) measures.

State of the Art and Future Perspectives

The mainstream edifice of neoclassical microeconomics erected throughout the twentieth century is still largely in place, and is still the basis of most economics teaching; yet its scientific underpinnings have undergone a process of profound rethinking and revision. Both general equilibrium analysis and the individual optimization paradigm have come under fire from critics of diverse obedience, including (but not limited to) behavioral economists. Research in general equilibrium theory now experiences a slow but neat decline; and studies of the principles of individual decision making tend to follow the behavioral rather than the neoclassical approach.

Yet far from languishing, the field is renewing itself at an unprecedented pace. The current tendency is to increasing diversification of approaches, methods, and topics, seemingly reversing the trend toward unification of the different parts of microeconomics observed in the twentieth century. Against the decline of – once dominant–general equilibrium theory, partial-equilibrium and experimental approaches are on the rise, and game theory has become the essential mathematical structure and support of a broad range of economic analysis. A case in point is market design – the study of how to define, enforce, and monitor trading rules that maximize the efficiency of exchange, for example in auctions. Part of the motivation for these studies in the 1990s and early 2000s was the need to design trading mechanisms that would help governments to privatize companies or state infrastructures, or to allocate property rights – the 3G auction in the United Kingdom being regarded as one of the most successful examples (Binmore and Klemperer, 2002). The design of an auction for the Bank of England's liquidity insurance operations is another case in point (Klemperer, 2010). These studies depart from the general equilibrium tradition by, first, using insight from game theory to highlight the importance of trading institutions to yield socially desirable outcomes, instead of abstracting them away; and second, by concentrating on single markets rather than global interdependencies – Marshall's 'partial equilibrium' perspective.

Partial equilibrium approaches also underlie a growing body of research on Internet markets, especially with the study of so-called 'two-sided' markets. This expression designates markets in which two different sets of agents interact through an intermediary or platform, and the decisions of each set of agents affect the outcomes of the other set of agents, typically through an externality (Rysman, 2009). Internet and media companies operating between advertisers and consumers are a typical example, along with the payment card industry (between merchants and customers). Pricing is usually asymmetric, and depends on the price sensitivity – the elasticity of demand – of each side; it often turns out that one side does not pay, or is even rewarded for using a service (with cash-back for credit card use for example), while the other, more price-inelastic side faces a high mark-up (Rochet and Tirole, 2003, 2006). Research in this area is booming and top microeconomic theorists are increasingly often consulted by, or serve as chief economists of, major Internet companies.

These examples bring to light a second major tendency in today's renewal, that for microeconomics to become much more practically oriented and empirically informed. If nineteenth- and twentieth-century microeconomics left aside quantification to adopt formalization (an approach that purports to develop hypotheses and explanatory arguments independently of, or prior to, confrontation with empirical data), recent research uses empirical investigation methods to test theories against data, thereby typifying a third and more recent use of mathematics, namely *validation* and verification, testing (formal) theory against (quantitative) data according to some version of the scientific method. To do so, new sources of information are being exploited, notably experimentation and analysis of survey and administrative data with the help of increasingly sophisticated microeconometric techniques. These techniques are now recognized parts of the field and attract an increasing number of young economists.

The future of microeconomics is, in sum, still open, with promising developments in several subfields and still greater potential for further success, and with access to an ever-wider and more diverse set of tools, techniques, theories, and datasets. Yet the precise directions of its future developments are still unclear – many paths can be taken, and the major task of choosing which ones to take lies ahead of microeconomists in the near future.

See also: Behavioral Economics, History of; Capital: History of the Concept; Equilibrium: History of the Concept; Prospect Theory.

Bibliography

Barreto, H., 2012. Microeconomics: Past, Present, and Future. Speech at the 2012 Conference of the History of Economics Society: http://www.econrx.org/htdocs/hes/barreto_microeconomics_past_present_future.pdf (accessed 01.09.13.).

Becker, G.S., 1992. The economic way of looking at life. Journal of Political Economy 101 (3), 383–409.

Binmore, K., Klemperer, P., 2002. The biggest auction ever: the sale of the British 3G telecom licences. Economic Journal, 74–96.

Chamberlin, E.H., 1948. An experimental imperfect market. Journal of Political Economy 56 (2), 95–108.

Debreu, G., 1984. Economic theory in the mathematical mode. American Economic Review 74 (3), 267–278.

Friedman, D., Rust, J. (Eds.), 1993. The Double Auction Market: Institutions, Theories, and Evidence. Addison–Wesley, Reading, MA.

Frisch, R., 1933. Propagation problems and impulse problems in dynamic economics. In: Economic Essays in Honor of Gustav Cassel. George Allen and Unwin, Ltd, London.

Hutchison, T.W., 1988. Before Adam Smith: The Emergence of Political Economy 1662–1776. Blackwell, Oxford, UK.

Ingrao, B., Israel, G., 1990. The Invisible Hand: Economic Equilibrium in the History of Science. MIT Press, Cambridge.

Jevons, W.S., 1965. The Theory of Political Economy. Kelley, New York (Original work published 1871).

Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decisions under risk. Econometrica 47, 313–327.

Keynes, J.M., 1936. The General Theory of Employment, Interest and Money. Macmillan, London.

Kirman, A.P., 1989. The intrinsic limits of modern economic theory: the emperor has no clothes. Economic Journal 99 (395), 126–139.
Klemperer, P., 2010. The product-mix auction: a new auction design for differentiated goods. Journal of the European Economic Association 8, 526–536.
Levy, D.M., 2001. How the Dismal Science Got Its Name: Classical Economics and the Ur-text of Racial Politics. University of Michigan Press, Ann Arbor, MI.
Mas-Colell, A., Whinston, M., Green, J., 1995. Microeconomic Theory. Oxford University Press, Oxford.
Morgenstern, O., von Neumann, J., 1944. Theory of Games and Economic Behaviour. Princeton University Press, Princeton, NJ.
Peart, S.J., Levy, D.M., 2005. The "Vanity of the Philosopher": From Equality to Hierarchy in Post-classical Economics. University of Michigan Press, Ann Arbor, MI.
Ricardo, D., 2004. In: Sraffa, P., Dobb, M. (Eds.), On the Principles of Political Economy and Taxation. Liberty Fund, Indianapolis, IN (Original work published 1817).
Rochet, J.C., Tirole, J., 2003. Platform competition in two-sided markets. Journal of the European Economic Association 1 (4), 990–1029.
Rochet, J.C., Tirole, J., 2006. Two-sided markets: a progress report. The RAND Journal of Economics 35 (3), 645–667.
Rysman, M., 2009. The economics of two-sided markets. Journal of Economic Perspectives 23 (3), 125–143.
Samuelson, P., 1938. A note on the pure theory of consumer's behavior. Economica 5, 61–71.
Smith, A., 1981. In: Campbell, R.H., Skinner, A.S. (Eds.), An Inquiry into the Nature and Causes of the Wealth of Nations, 2 vols. Liberty Fund, Indianapolis, IN (Original work published 1776).
Smith, V.L., 1962. An experimental study of competitive market behaviour. Journal of Political Economy 70 (2), 111–137.
Sraffa, P., 1925. Sulle relazioni fra costo e quantità prodotta. Annali di economia 2, 277–328.
Walras, L., 1977. Elements of Pure Economics: Or the Theory of Social Wealth (W. Jaffé, Trans.). A. M. Kelley, Fairfield, NJ (Original work published 1874).
Weintraub, E.R., 2002. How Economics Became a Mathematical Science. Duke University Press, Durham, NC.

Microevolution

Francisco M Salzano, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Biological changes that occur along time or space within a species are called microevolutionary. The basic factors responsible for them are (a) mutation (changes in the genetic material) and (b) natural selection (the preservation of favorable and elimination of deleterious mutations). Other factors, however, may influence this process. They can be categorized as population size (in small groups, random processes may be important), interpopulation migration (number and type of founders), and assortative mating. In humans, culture can decisively influence these and other parameters; and concerns are raised over recent opposition to the use of DNA in human population studies.

Concept

The word 'evolution' derives from the Latin term *evolutio*, which means 'an unrolling.' It can be used in this sense, or in others that involve the idea of change. But not all changes are evolutionary. The ocean's surface is always changing, but this is not an evolutionary process. Implicit in the concept of evolution are (1) continued change, (2) divergence, (3) restriction of opportunities, and (4) in a large number of situations, irreversibility. It is questionable whether there is a general direction (hence, 'progress') in the organic evolutionary process.

The primary factors that determine evolution are mutation (genetic change) and selection (in Charles Darwin's words, "The preservation of favorable individual differences and variations, and the destruction of those which are injurious" (Darwin, 1859)). The two, however, can be influenced by population structure. Among the variables that should be considered in this latter context, mention should be made of population size, the mobility and life histories of its members, sex and the genetic system, and assortative mating.

The dazzling array of organic diversity present on earth has always fascinated humankind. An important property of this variability is that it is discontinuous. Organic matter is organized in discrete individuals and groups of individuals. Moreover, these groups may be isolated reproductively from other similar ensembles. When this occurs, it is said that they are different species.

Microevolutionary processes are those that occur within a species. *Homo sapiens* is a particularly favorable subject for the investigation of these processes because, of course, humans know much more about themselves than about any other organism. However, humans are unique among other forms of life because they developed culture to an unprecedented degree. This trait varies according to its own laws, and is largely independent of biological factors. The reciprocal, however, is not true; organic evolution in humans may be strongly influenced by cultural processes.

Mutation

The information needed to assure the transmission of biological inheritance is given by a molecule, deoxyribonucleic acid (DNA). The particular sequence of its component units (nucleotides) is important because it codes (through another molecule, ribonucleic acid (RNA)) for the sequences of amino acids, the protein units that, when combined in arrays, make up physical bodies. The unit of heredity is the gene (present in a locus), composed of coding exons and noncoding introns (which probably have regulatory functions, e.g., they may modulate exonic gene expression). The transmission takes place in DNA–protein arrays called chromosomes. The genetic material can occur either in the nucleus of the cells or in the cytoplasm, in an organelle called the mitochondrion.

Detrimental agents or problems during the replication process may lead to mutations, changes in the previously existent DNA sequence. Advances in the molecular study of these structures revealed that many different types of change may occur, such as alteration in just one nucleotide, deletions or insertions of many of them, triplet (the coding unit is made of trios) expansions, variations in nucleotide position, changes in the process of DNA → RNA → protein information transfer, unequal recombination, and gene conversion between different DNA strands. Moreover, these variations do not occur at random. There are changes that are allowed and others that are forbidden, depending on the importance of the particular DNA region for cell physiology. Their fate will also be quite different if they occur in noncoding or coding regions. A special type of change is that which occurs due to horizontal DNA transfer, that is, pieces of DNA (transposons) move from one species to the other, or between chromosome regions of the same species, helped by infectious agents.

Although the human per-generation mutation rate is exceptionally high, on a per-cell division basis the human germline mutation rate is lower than that recorded for any other species. Aspects of our specific gene structure condition the production of defective alleles, and our longevity leads to an accumulation of somatic mutations. Therefore, Lynch (2010) predicted a substantial reduction in human adaptation in the coming centuries in industrialized societies, where mutagenic agents (physical, chemical, or biological factors causing alteration in DNA molecules) are more prevalent.

A kind of variation that only now can be assessed in a quantitative way is copy number variations (CNVs), which are variable adjacent arrays of repeats, deletions, and duplications. Conrad et al. (2010) mapped 11 700 CNVs in 450

individuals of European, African, or East Asian ancestry. This information will significantly enhance our ability to identify sources of pathological agents in our species.

Natural Selection

Natural selection is, like mutation, a category that involves many different processes. For instance, selection may be directional when it changes the adaptive norm of a population; stabilizing when this norm is protected; or balancing, which includes all those factors that maintain genetic heterogeneity or polymorphisms (common variants). Selection occurs at all levels of the biological hierarchy (molecular, cellular, tissue, organ, individual, population, species, and community of different taxonomic entities), and its agents are accordingly diverse (Salzano, 1975; Williams, 1992).

A current sustained effort to identify selection in humans is through molecular analysis of common variants (polymorphisms), which are compared with different aspects of our physical and socioeconomic environment. In this context, the selective sweep model was developed. Two situations can be envisaged, labeled as hard and soft sweeps. In the first case, a new, strongly beneficial mutation increases in frequency until fixated in the population. Due to the fast rate of the process, there is little opportunity for recombination to separate the variant from its genetic background, leading to a sweep (reduction) of neutral variation at linked sites. This event distorts allele frequencies at these regions and affects patterns of linkage disequilibrium (departures from free recombination along regions of a given chromosome). In soft sweep, selection would act on preexisting genetic variation that was previously neutral or deleterious, but became adaptive due to changes on the environment or genetic background.

There are ways to identify both phenomena through molecular analysis, and Hernandez et al. (2011), using sequencing data from 179 human genomes, concluded that hard sweeps (the results of a single instance of mutation spreading through the entire population) were probably rare in our evolutionary past 250 000 years. This finding does not imply that positive (based on advantageous alleles) selection was not important for human evolution. Akey (2009) surveyed 21 genome-wide scans for recent or ongoing positive selection in our genetic material, highlighting their main aspects. Stearns et al. (2010) also reviewed how selection could be measured in contemporary human populations, giving specific examples.

How can selective pressures be measured? Hancock et al. (2011) correlated allele frequencies and climate across 61 worldwide populations. The most extreme signals of selection were associated with pigmentation and autoimmune diseases. Tennessen and Akey (2011), however, found that selection has favored modest divergence at particular loci multiple times in independent human populations. Another approach focused on pathogen-induced selection pressures, considered by Fumagalli et al. (2011) and Novembre and Han (2012).

The *paired box homeotic gene 9*, a transcription gene with a major role in early development, including dentition, shows a peculiar evolutionary trend. It has four exons; the first acts uniquely as a start codon, exons 2 and 4 are evolutionarily highly conserved, but exon 3 has a high level of intraspecific variation, and one of its alleles (*Ala240Pro*) is associated with the third molar agenesis, a trait that could be advantageous in populations with progressive reduction of the dental arches, which could be due to diets based on processed food. The reduction would prevent impacted and defective third molars, as well as related oral problems, an important cause of morbidity in nonindustrial societies (Paixão-Côrtes et al., 2011).

Population Structure

Selection can act through differences in mortality or fertility, and a dialectic dilemma that faces any organism is how much energy it should invest in individual survival, as opposed to reproductive efforts. Throughout human history, significant changes have occurred in the emphasis given to these two processes, and they have been connected with subsistence patterns greatly influenced by culture. Among hunters and gatherers, mobility was very important, and the nomadic way of life would disfavor large numbers of offspring. Several measures were then developed to assure a small number of children (herbal contraceptives, mating taboos, abortion induction by mechanical means, and infanticide). Variance in number of children, a key evolutionary parameter, was thus restricted. Since population numbers were small, infectious agents would not be very important, and mortality would be mainly determined by other environmental factors or by violent deaths determined by rival groups.

The situation changed with the domestication of plants and animals. The premium now was to have a large number of children, to help in the cultivation and harvest of plants, or to take care of the animals. The possibility of storing large quantities of food made life more independent of the environment, and there was a concomitant increase in population size. Higher mortality levels then developed, due to epidemics and also due to the fact that it is easier to take care of two as compared to, say, 10 children.

Modern technology made possible a better control of number of children and of infectious agents. The action of natural selection, therefore, became more restricted.

Tribal groups have a population structure largely determined by what Neel and Salzano (1967) proposed as the fission–fusion model. Fissions can occur due to social tensions, the migrating units being composed of lineal relatives; and fusions, as convenience dictates. Using more than 580 000 single-nucleotide polymorphisms (SNPs), Henn et al. (2011) studied three hunter-gatherer African populations and compared a subset of 55 000 of these SNPs with those of 24 other groups of the same continent. Hunter-gatherers have the lowest level of linkage disequilibrium among the set, but continue to maintain the highest levels of genetic diversity found in the world.

Arylamine N-acetyltransferase 2 (NAT2) is a protein involved in physiological responses to a variety of compounds, including common therapeutic drugs and exogenous chemicals present in the diet and the environment. Sabbagh et al. (2011) collected information on 128 population samples (in a total of 14 679 individuals) representing different subsistence modes and dietary habits. A significantly higher prevalence of the slow

acetylation phenotype was found in populations practicing farming and herding than in those relying on hunting and gathering. The Neolithic transition, therefore, should have modified the selective regime affecting the NAT2 acetylation pathway.

Migration

Why do people migrate? Certainly, a series of socioeconomic and biological factors are involved in this process. Tovo-Rodrigues et al. (2010) investigated one of them, related to the dopamine receptor D4 system. Its gene (*DRD4*) is one of the most variable in the human genome, and the *7R* allele is associated with a novelty-seeking personality dimension, impulsivity and hyperactivity. Tovo-Rodrigues et al. (2010) compared the *DRD4* allele distribution in 23 Native American populations and found that *7R* is significantly more frequent in hunter-gatherers as compared to agriculturalists. Their explanation for this finding is that exploratory behaviors would be adaptive in nomad societies, because they allowed for more successful exploitation of resources. Conversely, in agriculturalist sedentary societies, novelty-seeking and exploratory behaviors would have social costs, because these populations developed intensive methods for using the land in a given settled environment.

At the macrodistribution, continental level, it has been proposed that the east–west orientation of the Eurasian landmass facilitated the rapid spread of ancient technological innovations across it, while the north–south orientation of the Americas would have led to a slower diffusion of technology there (Diamond, 1997). If the diffusion of technology was accompanied by gene flow, then genetic differentiation in the Americas along lines of longitude would be greater than that in Eurasia along lines of latitude.

Ramachandran and Rosenberg (2011) employed 678 microsatellite loci (composed of different adjacent duplications) from 68 indigenous populations of both regions to examine this point. They found that genetic differentiation in the Americas occurs along both latitudinal and longitudinal axes, and has a greater magnitude than corresponding differentiation in Eurasia. If diversity is inversely correlated to gene flow, these results support the hypothesis of a primary influence for continental axes of orientation in the demic diffusion of technology in Eurasia and the Americas.

Assortative Mating

Mating choice is a complex behavior characteristic that involves psychological, cultural, socioeconomic, and biological variables. Its main evolutionary influence is on the distribution of genotype frequencies. In multiethnic communities, there is a clear preference for homogamic matings (mating with a partner from the same population). Recently, with the use of exclusively matrilineal (mitochondrial DNA) or patrilineal (Y chromosome) genetic markers, it is possible to evaluate the influence of sex when heterogamic unions occur, in historical perspective. Thus, in Latin Americans of mixed ancestry, the European component was mainly contributed by males, while the Amerindian fraction is mostly derived from females. This is a reflection of unions that occurred in the Colonial period and that could be detected independently of the demographic, cultural, and biological changes that occurred afterward (Salzano and Bortolini, 2002).

Our research group has been studying the Xavante Amerindians of Central Brazil for the past half of a century (see, e.g., Coimbra et al., 2002), and since the beginning observed that their morphology differed from those of other Amerindian tribes. Hünemeier et al. (2012a) subjected this observation to a detailed analysis, considering cephalometric data obtained from 1203 individuals, a sample that included five other Brazilian tribes (Baniwa, Kaingang, Kayapo, Ticuna, and Yanomama), plus a Mexican one (Otomí) used as an external group. This information was compared to genetic, climatic, social, and life-history data, which were submitted to complex mathematical–statistical analysis. The main finding is that the Xavante had a remarkable pace of morphological change, which was 3.8 times higher in relation to the average, and which included the geographically and linguistically distant Otomí. The differences found cannot be attributed to other genetic (mitochondrial DNA) or climatic factors, but sexual selection could be a culture-driven force worth considering. The reproductive success of some Xavante chiefs in their polygamous society was well documented, and the Xavante clearly depart from the Kaingang, Kayapo, and non-Amerindian groups in terms of marital correlation of anthropometric traits (Stark et al., 1990).

Culture

Two branches of mathematical evolutionary analysis are the gene–culture coevolutionary theory and niche construction theory. Description of the first approach was provided by Laland et al. (2010), who also presented a list of 129 genes whose rapid rate of change could have been attributed, at least partially, to cultural factors such as plant domestication, dairy farming, sexual selection, complex cognition, and diet.

Lactase persistence in adulthood is a classic example of such type of interaction, the causative mutations in the *LCT* gene and its enhancer region having increased in frequency in populations that relied on milk for their diets. Friedrich et al. (2012) have studied these genetic regions in 316 individuals from four Brazilian Amerindian populations. Only the *−13910*T* allele, responsible for the persistence of lactase in Europeans, was present in frequencies that varied from 0.5% to 7.6%. The near-absence of the variability on the enhancer region, however, was not accompanied by that on the coding portion of the gene, which showed different combinations of gene arrangements (haplotypes) with a heterogeneous inter-population distribution. Different evolutionary factors can be at play even in relatively close DNA segments.

Niche construction theory emphasizes the capacity of organisms to modify their environment and thereby influence their own and other species' evolution. Kendal et al. (2011) presented the fundamental principles of this concept, evaluated their consequences, and suggested how the human biological and social sciences could be integrated to provide specific examples of it.

The ATP-binding cassette transporter A1 gene variant (*ABCA1*Arg230Cys*), up to now found only in Native Americans or in populations partially derived from them, was associated with a cholesterol efflux reduction and could have been advantageous in the past during food deprivation periods. But under the current modern lifestyle, with its excess of food availability, it could have a detrimental effect, causing obesity and type 2 diabetes in its carriers (Acuña-Alonzo et al., 2010). This is, therefore, an additional example of the 'thrifty' genotype hypothesis, formulated by Neel (1962). Results of our group (Hünemeier et al., 2012b) showed a strong correlation (90%) between the radiocarbon dates of maize pollen relics found in archeological sites of Mesoamerica with the frequency of this allele in populations situated nearby. The evolutionary hypothesis established in view of these and other findings is that Mesoamericans constructed an ecological niche that relied mostly on maize cultivation, leading to conditions that were favorable to this specific gene variant (since dependence on just one crop would lead to frequent famine episodes).

Problems

Practitioners of bioethics, a term created by Van Rensselaer Potter in 1970, have been considering a series of problems, one of them being the ethics of biomedical research in human populations, particularly that involving tribal groups. Salzano and Hurtado (2004) considered in detail these latter aspects. Two important principles were postulated: (a) the well-being of the study population should be the researcher's top priority; and (b) full informed consent should be required. In addition, any economic profit eventually obtained from the study should be shared by the communities who provided the data.

It is unfortunate, however, that activists developed a pronounced hostility in relation to population genetics studies in these populations. Two examples illustrating this point will be given. First, in the Guidelines for Health Research Involving Aboriginal People issued by the Canadian Institutes of Health Research, article 14 states, "An Aboriginal community should have an opportunity to participate in the interpretation of data and the review of conclusions drawn from the research to ensure accuracy and cultural sensitivity of interpretation before submission for publication" (McInnes, 2011). There is no indication of who would judge the findings and the arguments that would instruct this judgment, or if just one vote would block the publication.

The second example is a paper by Reardon and TallBear (2012), who argued that all DNA research is carried out primarily by whites (therefore, ignoring the excellent investigations of, e.g., our Japanese and Chinese colleagues), and that researchers doing genomic science consider the DNA collected as their property. They also charged population geneticists of concealed racism. But they do not define what they consider to be 'white,' therefore, adopting the same racial typological thinking that they were trying to condemn! Of course, other social scientists do not share the same views. What is needed is a continued dialogue between researchers and research participants from diverse academic disciplines and ethnic backgrounds.

Human population genetics seeks to understand the history of humankind, not the history of specific groups. To restrict such research on the basis of the assumption that the rights or interests of certain groups are infringed by this research, without individual research participants making this point themselves, is an ethically very problematic strategy.

See also: Brain, Evolution of; Cultural Evolution: Theory and Models; Darwin, Charles Robert (1809–82); Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Evolutionary Epistemology; Evolutionary Social Psychology; Gene-Culture Coevolution; Human Evolutionary Genetics; Lifespan Development: Evolutionary Perspectives.

Bibliography

Acuña-Alonzo, V., et al., 2010. A functional *ABCA1* gene variant is associated with low HDL-cholesterol levels and shows evidence of positive selection in Native Americans. Human Molecular Genetics 19, 2877–2885.

Akey, J.M., 2009. Constructing maps of positive selection in human: where do we go from here? Genome Research 19, 711–722.

Coimbra Jr, C.E.A., Flowers, N.M., Salzano, F.M., Santos, R.V., 2002. The Xavánte in Transition. Health, Ecology, and Bioanthropology in Central Brazil. University of Michigan Press, Ann Arbor.

Conrad, D.F., et al., 2010. Origins and functional impact of copy member variation in the human genome. Nature 464, 704–712.

Darwin, C., 1859. On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life. John Murray, London.

Diamond, J.M., 1997. Guns, Germs, and Steel: The Fates of Human Societies. WW Norton, New York.

Friedrich, D.C., Callegari-Jacques, S.M., Petzl-Erler, M.L., Tsuneto, L., Salzano, F.M., Hutz, M.H., 2012. Stability or variation? Patterns of lactase gene and its enhancer region distributions in Brazilian Amerindians. American Journal of Physical Anthropology 147, 427–432.

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L., Nielsen, R., 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genetics 7, e1002355.

Hancock, A.M., et al., 2011. Adaptations to climate-mediated selective pressures in humans. PLoS Genetics 7, e1001375.

Henn, B.M., et al., 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proceedings of the National Academy of Sciences of the United States of America 108, 5154–5162.

Hernandez, R.D., Kelley, J.L., Elyashiv, E., Melton, S.C., Auton, A., McVean, G., 1000 Genomes Project, Sella, G., Przeworski, M., 2011. Classical selective sweeps were rare in recent human evolution. Science 331, 920–924.

Hünemeier, T., et al., 2012a. Cultural diversification promotes rapid phenotypic evolution in Xavánte Indians. Proceedings of the National Academy of Sciences of the United States of America 109, 73–77.

Hünemeier, T., et al., 2012b. Evolutionary responses to a constructed niche: ancient Mesoamericans as a model of gene-culture coevolution. PLoS One 7, e38862.

Kendal, J., Tehrani, J.J., Odling-Smee, J., 2011. Human niche construction in interdisciplinary focus. Philosophical Transactions of the Royal Society B: Biological Sciences 366, 785–792.

Laland, K.N., Odling-Smee, J., Myles, S., 2010. How culture shaped the human genome: bringing genetics and the human sciences together. Nature Reviews Genetics 11, 137–148.

Lynch, M., 2010. Rate, molecular spectrum, and consequences of human mutation. Proceedings of the National Academy of Sciences of the United States of America 107, 961–968.

McInnes, R.R., 2011. Culture: the silent language geneticists must learn – genetic research with indigenous populations. American Journal of Human Genetics 88, 254–261.

Neel, J.V., 1962. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? American Journal of Human Genetics 14, 353–362.

Neel, J.V., Salzano, F.M., 1967. Further studies on the Xavante Indians. X. Some hypotheses-generalizations resulting from these studies. American Journal of Human Genetics 19, 554–574.

Novembre, J., Han, E., 2012. Human population structure and the adaptive response to pathogen-induced selection pressures. Philosophical Transactions of the Royal Society B: Biological Sciences 367, 878–886.

Paixão-Côrtes, V.R., et al., 2011. Genetic variation among major human geographic groups supports a peculiar evolutionary trend in PAX9. PLoS One 6, e15656.

Ramachandran, S., Rosenberg, N., 2011. A test of the influence of continental axes of orientation on patterns of gene flow. American Journal of Physical Anthropology 146, 515–529.

Reardon, J., TallBear, K., 2012. "Your DNA is *our* history": genomics, anthropology, and the construction of whiteness as property. Current Anthropology 53 (Suppl. 5), S233–S245.

Sabbagh, A., Darlu, P., Crouau-Roy, B., Poloni, E.S., 2011. Arylamine N-acetyltransferase 2 (NAT2) genetic diversity and traditional subsistence: a worlwide population survey. PLoS One 6, e18507.

Salzano, F.M. (Ed.), 1975. The Role of Natural Selection in Human Evolution. North-Holland, Amsterdam.

Salzano, F.M., Bortolini, M.C., 2002. The Evolution and Genetics of Latin American Populations. Cambridge University Press, Cambridge, UK.

Salzano, F.M., Hurtado, A.M., 2004. Lost Paradises and the Ethics of Research and Publication. Oxford University Press, New York.

Stark, A.E., Salzano, F.M., Da Rocha, F.J., 1990. Marital correlation for anthropometric characteristics in Brazilian Indians. Annals of Human Biology 17, 417–422.

Stearns, S.C., Byars, S.G., Govindaraju, D.R., Ewbank, D., 2010. Measuring selection in contemporary human populations. Nature Review Genetics 11, 611–622.

Tennessen, J.A., Akey, J.M., 2011. Parallel adaptive divergence among geographically diverse human populations. PLoS Genetics 7, e1002127.

Tovo-Rodrigues, L., Callegari-Jacques, S.M., Petzl-Erler, M.L., Tsuneto, L., Salzano, F.M., Hutz, M.H., 2010. Dopamine receptor D4 allele distribution in Amerindians: a reflection of past behavior differences? American Journal of Physical Anthropology 143, 458–464.

Williams, G.C., 1992. Natural Selection: Domains, Levels, and Challenges. Oxford University Press, New York.

Microsimulation in Demographic Research

Emilio Zagheni, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Santow, volume 14, pp. 9780–9785, © 2001, Elsevier Ltd.

Abstract

Demographic microsimulation is an individual-based and computationally intensive tool used by population scientists to model demographic processes, to gain insights on life course transitions and to make projections. Several microsimulators have been developed, since the 1960s, to address questions for which standard techniques or data sets cannot provide answers. Examples of applications include historical studies of demographic constraints on household formation, the impact of the HIV/AIDS epidemic on kinship resources for orphans and the elderly, and the study of interaction and feedback mechanisms in demographic behavior. From the methodological point of view, calibration of stochastic microsimulators is an area of active research.

Demographic microsimulation is a computationally intensive tool used by population scientists to model demographic processes, to gain insights on life course transitions and to make projections. Two important features distinguish microsimulation from other models. First, the unit of analysis is the individual (hence 'micro'). Second, the sequences of events that individuals experience over time are the result of stochastic experiments with predetermined probabilistic rules. Transitions between states are typically generated using computer algorithms and techniques also known as Monte Carlo methods (hence 'simulation').

In a standard situation, each individual in the simulation is an observation in a rectangular data file, which contains records of demographic characteristics and other key variables of interest. The simulator takes as input population files and demographic rates, and updates the population accordingly. Each simulated person is subject to a set of rates, conditional on certain demographic characteristics such as age, sex, marital status, etc. For every predefined time interval, each individual faces the risk of a number of events including death, marriage, childbirth, and migration. The selection of the event and the waiting time until the event occurs are determined stochastically, often using a competing risk model. Some constraints may be included in the simulation program in order to restrict the range of potential events for particular subgroups of the population (e.g., to avoid social taboos such as incest, to allow for a minimum interval of time between births from the same mother, etc.). Each event for which the individual is at risk is often modeled as a piecewise exponential distribution. The waiting time until each event occurs is randomly generated according to the input demographic rates. The individual's next event is the one with the shortest waiting time.

Major Modeling Options

A number of microsimulators have been developed over the course of several decades to address different types of research questions. Microsimulation models can be classified into various categories based on some distinctive features.

Continuous versus Discrete Time

The scheduling of events for simulated individuals is an important feature of each microsimulator. The algorithms that determine the next event for each agent may treat time as a continuous or as a discrete variable.

In continuous-time microsimulations, the timing and sequence of events that individuals experience over the life course are the result of a competing risk model. For every event for which an individual is potentially at risk, the simulator randomly generates a waiting time until the event occurs. The most common waiting time distribution is the piecewise exponential. Other standard distributions include the Weibull and the Gompertz distributions (Willekens, 2009). The parameters for the distributions of waiting times are chosen according to the input demographic rates for each subgroup in the population. The individual's next event is the one with the shortest waiting time. Every individual is always at risk of death; only women in reproductive age are at risk of giving birth; only married people are at risk of divorce, etc. The relative risk of each event happening next depends on input demographic rates, which typically vary by age, sex, marital status, parity, etc. Once the scheduled event occurs, a new set of waiting times is generated and a new event is scheduled accordingly.

In discrete-time simulations, time intervals are modeled, instead of the exact times before events occur. Each time period is considered separately from the others and each event is executed only once within the time period. Different events may occur within the same interval of time. As time periods get shorter and shorter, results from discrete-time simulations get more and more similar to the ones of continuous models. Continuous models have some logical and practical advantages. In continuous microsimulations, it is the sequence of events and transitions that is modeled and not state occupancies at different points in time. In discrete-time models, the duration of events can be modeled only approximately. Thus, multiple transitions within a period of time require assumptions about the ordering and the timing of events. Moreover, discrete-time models cannot handle complex and interdependent sequences of events (Willekens, 2009).

Closed versus Open Population

Microsimulations where individuals can enter the sample only through 'birth' from an existing woman, and can exit only through 'death,' are referred to as *closed* models. Whenever new people in the simulated population are generated other than through birth, then the simulation is defined *open*.

The distinction between closed and open populations mainly refers to the modeling technique used to match partners. In closed-population models, the choice of the spouse is restricted to individuals existing in the population. In open-population models, partners with suitable demographic characteristics are created in order to satisfy the demand for spouses at any given time.

Closed-population models require complex matching algorithms and fairly large population sizes in order for the marriage market to clear. Open-population models do not require sophisticated matching algorithms. However, the history of newly generated individuals is typically missing. One of the main advantages of closed models is that they allow kinship ties to be tracked over time.

A standard example of closed-population kinship simulator is SOCSIM, a computer software developed at UC Berkeley in the 1970s. The computer program uses a two-stage process to pair eligible males and females from within the simulated population. When the next-scheduled event for an individual is 'marriage,' then the person is placed in a pool of eligible members to form a union. If a member of the opposite sex with appropriate demographic characteristics is available in the pool, then the two individuals are paired. Otherwise, the person stays in the pool until an appropriate mate 'picks' him or her, based on a random process with probabilities dependent on demographic characteristics of the two potential spouses.

In closed-population simulations, migrations are often modeled by introducing a group of people, representative of the 'rest of the world,' who evolve independently of the population of interest. Migrants, or entire families of migrants, are drawn from the pool of people who represent the rest of the world.

Starting Population

The initial population for a simulation can come from a cross-sectional sample of the entire population, like the group of respondents to a sampling survey, or can be a synthetic population generated in a way that is consistent with the expected age structure and other relevant characteristics of the population under study.

When the initial starting population comes from a cross-sectional data set representative of the entire population, then individual life histories are simulated for the future. This approach is relatively simple and guarantees that the starting population has demographic characteristics consistent with the ones of the population under study. However, no information on the history of individuals before the starting point of the simulation is then available. Therefore, most kinship ties cannot be reconstructed. In addition, when the initial sample size comes from a survey, it may be relatively small and thus there could be high levels of stochasticity in the outputs of the simulation.

Often, the starting population is produced using a synthetic procedure. A population of a given size, composed of unrelated individuals, is randomly generated. Then the population is projected forward for a long period of time (say at least 100 years) using demographic rates that generate a population consistent with the population under study at the beginning of the simulation (in terms of age structure, age at marriage, age at childbirth, divorce rates, etc.). In the synthetic population, kinship ties have been tracked and can thus be projected for the future.

Top-Down versus Bottom-Up

Microsimulation models simulate demographic events for individuals in a way that is consistent with macro-demographic rates. One of the goals of microsimulation is to evaluate the consequences of demographic change for a number of quantities of interest. The methodology can be seen as a top-down approach. An alternative line of individual-based simulations, often referred to as agent-based models, pursues a bottom-up approach. Agent-based models simulate agents with built-in behavioral rules of action and interaction with other individuals and their environments. The main goal of agent-based models is to study the emergence of global patterns from simple behavioral rules of autonomous agents.

Standard microsimulation models tend to emphasize the modeling of macro-to-micro interactions, and are used to evaluate the impact of policy changes, long-term demographic trends, and demographic shocks, on various quantities of interest. Agent-based models focus on the micro-to-macro direction and are often used to test theories and to evaluate the emergence of complex phenomena.

Although standard microsimulation and agent-based models are built on different assumptions and used for different goals, in practice the boundaries between the two approaches are not clearly defined. Agent-based models often include some macro-demographic rates that serve as macro-controls for the dynamics of the simulated population. Microsimulation, on the other hand, includes important behavioral rules that regulate, for instance, preferences and interaction in the marriage market.

Calibration

In principle, perfect knowledge of demographic rates should lead to an unbiased reconstruction of population dynamics and kinship networks through demographic microsimulation. The only uncertainty associated to the simulated kinship structure would be related to the stochasticity of the microsimulation. In practice, knowledge of vital rates is far from being perfect. Kinship reconstruction and forecasting demand a level of detail for demographic rates that is often missing in available data sets. For instance, transition rates from one marital status to another are usually not readily available and estimates may not be very accurate. Fertility rates are usually not broken down by marital status or parity, especially in the developing world. In most cases, demographic rates that are used as input to the microsimulation need to be estimated from various data sources with different sampling errors. Even when reliable data

sources exist to compute demographic rates broken down by the categories of interest, the heterogeneity of the population's rates within the tabulated categories constrains the accuracy of the microsimulation (Wachter et al., 1997).

Traditionally, the problem of calibrating microsimulations has been addressed using ad hoc tuning. Input rates are adjusted on a trial and error basis in order for the output of the simulation to match key summary demographic measures obtained from a population census or sample surveys. The validity of the microsimulation has been tested by comparing kinship forecasts generated in the past with external standards provided by surveys with detailed information on numbers and ages of kin in the United States (Wachter et al., 1997).

In the past, the development of methods to calibrate microsimulations has been mainly constrained by the limitation of computer power. Traditional methods have relied heavily on minimizing the number of simulation runs. That was done, for instance, using expert judgment for adjusting the input demographic rates in a consistent and appropriate way. With increasing computer power, there has been more and more interest in the development of methods to calibrate simulation models. The Bayesian melding method (e.g., Raftery et al., 1995; Poole and Raftery, 2000), in particular, has proved useful to formalize the process of calibration and statistical inference. It is a Bayesian approach since it relies on the Bayesian machinery of combining prior distributions with likelihoods to obtain posterior distributions. It has been named 'melding' because it "provides a way of combining different kinds of information (qualitative or quantitative, fragmentary or extensive, based on expert knowledge or on data) about different quantities, as long as the quantities to which they relate can be linked using a deterministic model" (Poole and Raftery, 2000). Sevcikova et al. (2007) and Zagheni (2010) extended the approach to stochastic simulations.

When using the Bayesian melding method, the researcher first has to express the available information about inputs and outputs in terms of probability distributions. For instance, this can be done by providing a prior distribution on the inputs. Then the researcher has to specify a conditional probability distribution of the data given the outputs. This yields a likelihood for the outputs, and, implicitly, produces a likelihood for the inputs. The combination of the prior on the inputs and the likelihood, using the Bayes' rule, gives the posterior distribution. In order to obtain the posterior distribution for the quantities of interest, Sevcikova et al. (2007) suggested a computational approach that is based on the sampling importance resampling algorithm of Rubin (1987, 1988). They applied the method to urban simulations. A similar approach has been used for the calibration of demographic microsimulations (Zagheni, 2010).

Some Applications

The idea behind the development of microsimulation methods as a research tool dates back to the late 1950s and early 1960s (Orcutt, 1957; Orcutt et al., 1961). Since then, a large number of microsimulators have been proposed and used to address demographic questions (Morand et al., 2010). It is beyond the purposes of this article to provide an exhaustive list of microsimulators and applications. Some illustrative examples that give a flavor of the scope of applications of microsimulation to demographic research will be presented in this section.

The study of kinship structure is one of the most successful applications of microsimulation models in demographic research. Two of the most widely used microsimulators are SOCSIM and CAMSIM. SOCSIM originates from a collaboration between Peter Laslett, Eugene Hammel, and Kenneth Wachter in the early 1970s. CAMSIM was developed in the late 1970s and early 1980s by Peter Laslett, James Smith, and Jim Oeppen (Zhao, 2006). Both simulators have been designed to study kinship networks. However, they use quite different algorithms, since SOCSIM is a closed model, whereas CAMSIM is an open-population simulator.

SOCSIM has been originally developed for historical studies of demographic constraints on household formation. In particular, it was used to test the hypothesis that social norms, and not unfavorable demographic conditions, were the cause of the low proportion of stem family households in pre-industrial England (Wachter et al., 1978). Other historical analyses include the assessment of the 1698 Slavonian census, using SOCSIM (Hammel and Wachter, 1996a,b) and the evaluation of potential biases in genealogical data, using CAMSIM (Zhao, 2001). A second line of research that has largely benefited from the use of microsimulators is the study of change of kinship availability over time. SOCSIM has been used to project kinship resources (including step-kin) for the elderly in the US (Wachter, 1997) as well as long-term changes in family and kinship networks in Britain (Murphy, 2011). Zhao (2006) offers a comparative review of applications of SOCISM and CAMSIM.

For countries with a generalized HIV/AIDS epidemic, microsimulation has been used to assess kinship resources for the elderly in Thailand (Wachter et al., 2002) and for orphans in Zimbabwe (Zagheni, 2011). Demographic microsimulations that include modules for the transmission and progression of HIV/AIDS have been used to model population dynamics of polygynous populations in sub-Saharan Africa (Clark, 2001). In the area of indirect estimation, microsimulation has proven useful to evaluate bias for methods developed to estimate mortality from sibling survival data (Masquelier, 2013).

Microsimulation has been the dominant research tool to analyze the consequences of demographic change for family and kinship networks. Agent-based models have recently emerged as a complementary tool to study marriage formation, choice of partners, and the role of behavior, interaction and feedback mechanisms in demographic research (Billari and Prskawetz, 2003). For instance, it has been found that relatively simple mate search rules that adjust on the basis of sequential encounters with potential partners may generate regularities in the distribution of the age at first marriage (Todd et al., 2005; Billari et al., 2007).

Strengths and Limitations

Microsimulation models have been developed to address questions for which standard tools could not provide an

answer. In that sense, microsimulation amplifies the power of scientific imagination of researchers, it allows population scientists to go beyond some of the simplifying assumptions made in standard macro-demographic models.

Microsimulation is the most appropriate tool when population heterogeneity and interaction matter, and when the number of variables and the number of attributes that these variables can take is very large (Van Imhoff and Post 1998; Spielauer, 2011). Microsimulation can accommodate interactions between individuals and between variables much better than macromodels with a large state space. More specifically, "microsimulation is particularly appropriate if the results of the process are complex but the driving forces of the process are simple" (Van Imhoff and Post 1998). Anytime processes are complex at the macrolevel, but better understood at the microlevel, or when individual histories are relevant, then microsimulation/individual-based models are the preferred choice.

Microsimulation is a very attractive tool both to gain insights into population dynamics and for policy-relevant analyses. However, there are two important limitations that need to be taken into account. First, microsimulations typically require high quality and very specific types of input data that are often not available. Moreover, even if detailed demographic data are available, the calibration process may prove difficult. Second, failing to model correlations between demographic events in models of kinship may lead to less variation in the frequency of kin of any particular type than would occur in the real population (Ruggles, 1993). In other words, models tend to underestimate the fraction of people with many kin and the fraction of people with few kin. In most situations the extent of the underestimation is very small and, for practical purposes, could be ignored. However, there may be cases where the size of the underestimation is not negligible and thus, in those situations, modeling of correlations is advisable.

See also: Demographic Models; Demographic Techniques: Indirect Estimation; Historical Demography; Mortality and the HIV/AIDS Epidemic; Multilevel Models in Demography; Population Dynamics: Mathematic Models of Population, Development, and Natural Resources.

Bibliography

Billari, F., Prskawetz, A. (Eds.), 2003. Agent-based Computational Demography. Contributions to Economics. Physica-Verlag.

Billari, F., Prskawetz, A., Diaz, B.A., Fent, T., 2007. The "Wedding ring": an agent-based marriage model based on social interaction. Demographic Research 17, 59.

Clark, S.J., 2001. An Investigation into the Impact of HIV on Population Dynamics in Africa (Ph.D. dissertation). University of Pennsylvania, Philadelphia.

Hammel, E., Wachter, K., 1996a. Evaluating the slavonian census of 1698, Part I: structure and meaning. European Journal of Population 12, 145–166.

Hammel, E., Wachter, K., 1996b. Evaluating the slavonian census of 1698, Part II: a microsimulation test and extension of the evidence. European Journal of Population 12, 295–326.

Masquelier, B., 2013. Adult mortality from sibling survival data: a reappraisal of selection biases. Demography 50 (1), 207–228.

Morand, E., Toulemon, L., Pennec, S., Baggio, R., Billari, F., 2010. Demographic Modeling: The State of the Art. SustainCity Working Paper 2.1a. Ined, Paris.

Murphy, M., 2011. Long-term effects of the demographic transition on family and kinship networks in Britain. Population and Development Review 37 (1), 55–80.

Orcutt, G.H., 1957. A new type of socio-economic system. Review of Economics and Statistics 39, 116–123.

Orcutt, G.H., Greenberger, M., Korbel, J., Rivlin, A., 1961. Microanalysis of Socio-economic Systems: A Simulation Study. Harper & Row, New York.

Poole, D., Raftery, A.E., 2000. Inference for deterministic simulation models: the bayesian melding approach. Journal of the American Statistical Association 95 (452), 1244–1255.

Raftery, A.E., Givens, G.H., Zeh, J.E., 1995. Inference from a deterministic population dynamics model for bowhead whales. Journal of the American Statistical Association 90 (430), 402–416.

Rubin, D., 1987. Comment on "The calculation of posterior distributions by data augmentation", by M. Tanner and W.H. Wang. Journal of the American Statistical Association 82, 543–546.

Rubin, D., 1988. In: Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (Eds.), Using the SIR Algorithm to Simulate Posterior Distributions. Bayesian Statistics 3. Clarendon Press, Oxford, UK, pp. 395–402.

Ruggles, S., 1993. Confessions of a microsimulator. Historical Methods 26, 161–169.

Sevcikova, H., Raftery, A.E., Waddell, P.A., 2007. Assessing uncertainty in urban simulations using Bayesian melding. Transportation Research. Part B 41, 652–669.

Spielauer, M., 2011. What is social science microsimulation? Social Science Computer Review 29 (1), 9–20.

Todd, P.M., Billari, F., Simao, J., 2005. Aggregate age-at-marriage patterns from individual mate-search heuristics. Demography 42 (3), 559–574.

Van Imhoff, E., Post, W., 1998. Microsimulation methods for population projection. Population: An English Selection 10 (1), 97–138.

Wachter, K.W., Hammel, E.A., Laslett, P., 1978. Statistical Studies of Historical Social Structure. Academic Press, New York.

Wachter, K.W., 1997. Kinship resources for the elderly. Philosophical Transactions of the Royal Society of London – Series B: Biological Sciences 352 (29).

Wachter, K.W., Blackwell, D., Hammel, E.A., 1997. Testing the validity of kinship microsimulation. Journal of Mathematical and Computer Modeling 26, 89–104.

Wachter, K.W., Knodel, J.E., VanLandingham, M., 2002. AIDS and the elderly of Thailand. Demography 39 (1), 25–41.

Willekens, F., 2009. Continuous-time microsimulation in longitudinal analysis. New Frontiers in Microsimulation Modeling, 353–376.

Zagheni, E., 2010. The impact of the HIV/AIDS epidemic on orphanhood probabilities and kinship structure in Zimbabwe (Ph.D. Dissertation). University of California, Berkeley.

Zagheni, E., 2011. The impact of the HIV/AIDS epidemic on kinship resources for orphans in Zimbabwe. Population and Development Review 37 (4), 761–783.

Zhao, Z., 2001. Chinese genealogies as a source for demographic research: a further assessment of their reliabilities and biases. Population Studies 55, 181–193.

Zhao, Z., 2006. Computer microsimulation and historical study of social structure: a comparative review of SOCSIM and CAMSIM. Revista de Demografia Historica XXIV (II), 59–88.

Middle Ages, The

Gábor Klaniczay, Medieval Studies Department, Central European University, Budapest, Hungary

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Situated between antiquity and Modernity, the Middle Ages constitute an intermediate period between the grand beginnings of European civilization and its contemporary achievements. An intermediate status is always ambivalent. Thinking in strictly dualistic terms, it is no longer the old and not yet the new, or, conversely, it allows the partial survival of the old and prepares the birth of the new. This article addresses three aspects: (1) the extension of European Middle Ages is both related to different periodization schemes of universal and national history, and situated in a geographical dimension. With the Roman Empire and Christianity as basic points of reference, this notion, born in Renaissance Italy and nurtured in Western Europe, has always embraced more than *Europa Occidens*. Besides a reconsideration of the North–South divide the specific medieval features of East-Central-Europe, the "borderlands of Western civilization" are described as well. For presenting these multiple and changing identities two methods are adopted: (2) the characterization of economic, social, and mental structures as debated by recent historiography, and (3) a historical chronology of narratives, personalities, and artifacts for adding life and color.

The Middle Ages are good to think. Situated by definition between Antiquity and Modernity, they constitute an intermediary period between the grand beginnings of European civilization and its contemporary achievements trying to come up to these standards and even surpass them. An intermediary status is always ambivalent. If one thinks in strictly dualistic terms, it is no more the old and not yet the new, or, conversely, it allows the partial survival of the old and it prepares the birth of the new. But it can also become an equally dignified participant in a triad: an age with special characteristics of its own. All this offers a paradise for conflicting interpretations, deprecations, and apologies; for an eternal to and fro of disdain, glorification, exoticism, and rediscovery.

The Notion of the Middle Ages

The most enduring characterization of the Middle Ages appeared already before the word itself was coined, it was inherited from Italian Humanists who had the habit since Petrarch to lament upon the 'dark' and 'barbarian' age between the glorious Greek and Latin Antiquity and the rebirth of this same quality in their own age. Yet, as soon as Renaissance was evolving into the new (post-Columbian, Gutenbergian, Copernican) world of Modernity, as soon as Middle Ages were regarded to be definitely over, condemnation got sided by waves of nostalgia: chivalric dreams resurrected by Ariosto, Cervantes, and Tasso, ideals of sainthood and a unified *Christianitas* cherished by the Catholic Reform, the Bollandists, Chateaubriand, and the Romantics. Medieval achievements like gothic cathedrals inspired both freemasonry and waves of occultism. The brutish energy of castles, armors haunted the imaginary with the Gothic Romance, but became also crystallization points for heroic-chivalric romantics, Wagner-type new barbarians, to degenerate finally into the comic world of Monty Python. The organic beauty of medieval cities and crafts was rediscovered by the Preraffaelites, by Burckhardt and Huizinga, the *art nouveau* and the *art deco*.

To these ambivalent judgments a third dimension must be added: that of different identities (local, institutional, national, and European). Medievalists inherited monastic annals, urban, or royal chronicles, order histories, and the history of western Christendom presented as world history. Much reasoning on Middle Ages was motivated, since Machiavelli's *Discorsi* (1525), Guicciardini's *Storia d'Italia* (1540), or Shakespeare's king-dramas by an interest of getting to know one's own 'national' roots. The antiquarian curiosity of Muratori (1672–1750) resulted in his path-breaking *Rerum Italicarum Scriptores*, the mighty and systematic source-collection on German past, the *Monumenta Germaniae Historica*, was started by Pertz (1826), the *Histoire de France* by Michelet (1833), the *Récits des temps mérovingiens* by Thierry (1840), or the studies on *The Constitutional History of England* by Maitland (1908). In the twentieth century historic attention also turned toward the universalistic frameworks: Christianity, Papacy, Empire, which could be analyzed as the 'birth' of Europe. Middle Ages meant the formation of modern Europe, which, based on its achievements, got ready for a world-wide expansion.

Then came the world wars and the revolutions and with them one could witness a new type of resurgence of interest in the Middle Ages. In front of destructive wars and equally destructive totalitarian states the safeguarding of the treasures of medieval Christian civilization became a task of universal importance. The appreciation of these values was helped by excellent studies in the history of religion, philosophy, fine arts, and literature, a host of new critical editions and a modern network of museums. Works by Auerbach, Delehaye, Haskins, Curtius, Gilson, Huizinga, Kristeller, and Panofsky have restored the dignity of the masterpieces of medieval thought and artistic creation, rooted in a value system and an aesthetic canon different from ours. Without all these medieval achievements the birth of renaissance, humanism, modern diplomacy, urban civilization, intelligentsia, and many other things would have been inconceivable. Even in the fights against the menace of modern barbarianisms one could usefully rely upon the legacy of the Middle Ages.

The simultaneously unfolding new interdisciplinary approaches, which tried to renew historical research, rediscovered a multicolored cultural variety that could be examined with an anthropologist's approach. The historians of the French *Annales* found a treasure-mine in the 'total' history of medieval civilization, and an unexpected series of best-sellers showed up: not only 'new history' monographs in the footsteps of *Montaillou* by Le Roy Ladurie (1985), but also new style historical romances as *A Distant Mirror* by Tuchman (1978) or *The Name of the Rose* by Eco (1980). From the 1980s on, châteaux, monasteries, city halls, but also modern university campuses became places for a permanent reinvention of the Middle Ages.

At the end of the twentieth century two further elements joined this picture. The collapse of Communism in Eastern Europe reoriented historical consciousness to medieval origins. Sometimes in a nineteenth-century style, as claim for territorial legitimacy, sometimes as foundation of the 'common roots' of their new European identity. The other novelty was provided by chronology. The year 2000 brought about a new fascination with the apocalypse, with some divergence in Eastern Europe: where instead of chiliastic calculations, pompous millennial celebrations were held remembering the glorious founding time of their Christian states.

Which Middle Ages should we address in the new Millennium, whose concept should we adopt? I will approach this from three angles: (1) periodization, (2) structures, and (3) a historical panorama of narratives, personalities, and artifacts.

Extension and Limits of the Middle Ages

As indicated, the concept of the Middle Ages was conceived by the Renaissance. The term *media tempestas* makes its appearance in 1469, in a praise of Cusanus by Bussi, a pontifical librarian. Analogous expressions (*media aetas*, *media antiquitas*), show up frequently and ally to the already established concept of 'dark ages' known from the writings of Petrarch, Bruni, Valla, and other Italian humanists.

But the notion of a transitory period between antiquity and its rebirth was older and this effects the chronological limits of the Middle Ages. Laments upon the lost greatness of ancient Rome and hopes of its imminent revival are already voiced by Gregory the Great. From Justinian I and Charlemagne to Otto III or Frederick Barbarossa, the slogan of *renovatio imperii Romani* repeatedly allowed the conceptualization of intermediary decadent periods. For the 'renaissance of the twelfth century,' perhaps the Arabic learning could be regarded as mediator between the wisdom of the *antiqui* and the Christian learning of the *moderni*.

The negative image of medieval decadence of classical civilization, however, was duly counterbalanced by the positive image of the advent of Christianity, the triumph of which upon Roman 'paganism' was praised as the demise of 'superstition,' 'tyranny,' 'idolatry,' and 'magic.' The judgment about the period following the fall of the Roman Empire was thus, rather positive during the Middle Ages, or at worst ambivalent. This is also reflected in the first elaborate system imposing a ternary division upon world history, that of Joachim of Fiore (+1202). Projecting the Holy Trinity model upon the history of mankind, Joachim distinguishes between the age of the Father (Old Testament), the age of the Son (New Testament), which lasts from Jesus to his own age (this actually corresponds to Middle Ages), and the imminent new period of the Holy Spirit, which will have an 'Eternal Evangile,' and where everybody would literally follow Christian life-precepts.

Early modern historiography inherits from the Renaissance the ambivalent image about the Middle Ages, and extends it to embrace universal history. Horn, a historian from Leiden, puts *medium aevum* between 300 and 1500 in his *Arca Noe* (1666). Cellarius, a historian from Halle, is the first to realize his world history on the basis of the ternary division into Antiquity (*Historia antiqua*, 1685), Modernity (*Historia nova*, 1986), and the Middle Ages between them (*Historia medii aevi*, 1688). The start of the *medium aevum* is for him the reign of Constantine and the acceptance of Christianity (313), and the end is marked by the fall of Constantinople (1453) (*see* Time, Chronology, and Periodization in History).

These limits were challenged by later suggestions. The 'start' was increasingly identified with the forced abdication, in 476, of Romulus Augustulus, the Western Roman Emperor, whose vacant throne never got filled again. The 'end' was rather brought in connection with the discovery of America (1492) or the start of the Reformation (1517). National histories adjusted dates to the dividing lines within their own field, English tend to count modernity from the advent of the Tudors (1485), Hungarians take the fatal defeat from the Ottomans at Mohács, in 1526, as end of medieval Hungarian history.

With an eye on Polish discourses on 'feudal economy' in the early modern times Le Goff made the provocative proposal of a 'long Middle Ages,' lasting till the French Revolution. Extension goes both ways: since the works on 'late antiquity' by Peter Brown, a five century long transitory period is replacing the divide of the 'fall' of Roman Empire, a period embracing the history of the Mediterranean basin from Diocletian until Charlemagne.

The extension of the Middle Ages has also a geographic dimension. With the Roman Empire and Christianity as basic points of reference, this notion, born in Italy and grown in Western Europe, has always embraced more than *Europa Occidens*.

The Mediterranean basin as a lively civilization-unit throughout the Middle Ages was a truism for the historians of commerce, crusades, or navigation. Yet, it had to be confirmed by Pirenne in his *Mohammed and Charlemagne* (1937b) and by Braudel in his *Mediterranean World* (1949). This rediscovery obliges historians of Latin Christianity to communicate with the large and autonomous tradition of Byzantine studies (Ostrogorski, Obolenski, Patlagean), and engage into systematic comparisons. Parallel to this renewed interest to an integrated panorama of the South, there was also a resurgence of nineteenth century visions on the 'barbaric,' individualistic vitality, and communal freedoms of the Germanic and Slavic North (*Markgenossenschaft*, *zadruga*). The medieval Roman and Germanic synthesis are also became the object of a new type of structural history by Pirenne (1933 [1937a]) and Bloch (1939). The medieval origins of Eastern, Central, East-Central, and South-Eastern Europe, these 'borderlands of Western civilization' have also been examined in impressive syntheses by the Romanian Iorga (1908–12), the Czech Dvornik (1949), the Polish Halecki (1952), or the Hungarian Szűcs (1990).

Structures in the Middle Ages

Defining a period within history requires the characterization of the structures that provide its identity and coherence; whether supposed to be 'objective,' deep structures, or the product of a momentary combination of various factors, or just an order observed by those who detect genealogies from a posterior moment.

The most general structural account imposed upon medieval period is that of 'feudalism' (*see* Feudalism), a notion which embraces social, economic, and cultural relations. Based on medieval legal concepts, '*féodalité*' served as the most general characterization of Ancien Régime during the Enlightenment (Montesquieu, Voltaire, Vico), was further developed by economic and sociological theory (Smith, Marx, Weber), to provide finally, with the work of Bloch, Duby and Le Goff, a comprehensive account of 'feudal society' and its economic structures, social stratification, habits, mentalities, rituals, cultural and religious achievements. Historical analysis found an attractive subject in this personalized power-structure which supplemented the early medieval weakness or lack of sovereign state. Sealed with the ritual of vassalage and with the conditional donation of a *feudum* (fief), 'feudalism' constituted a complex relationship of military, economic, social (and within it parental-familial) dependence and solidarity, one of the most original features of the Middle Ages, the foundation of Europe's future achievements, as it seemed.

Departing from a chaotic self-organization of early medieval times, evolving into the flowering of the 'second feudal age' (Bloch), feudal relations were analyzed along ecclesiastic, chivalric, courtly, and urban ramifications. The concept of feudalism also allowed the comparison of medieval Europe with other world civilizations, and Marxist historiography relied upon this when for coining a universal evolutionary pattern of world history. Renewed attempts discussed 'feudal economy,' 'feudal mutation' around the year 1000 (Poly and Bournazel), or regional European variants of feudal social structure. On the other hand, the exaggerated extension of this notion also led to an outright refusal of the concept (Brown, Reynolds).

Besides this synthetic vision of an all-encompassing pyramid of feudal dependencies other original medieval types of social structure should be mentioned. 'Orders,' appearing around the year 1000 as a horizontal classificatory category, referred to alliances of monasteries such as the Cluny *congregatio* or later the Cistercians, and they were also used for a general functional tripartition within society (*oratores-bellatores-laboratores*). This latter 'ideology' (Duby) prepared the late medieval appearance of 'estates' as social entities meriting political representation. Chivalric orders, religious confraternities, urban guilds, medieval universities lend this 'corporative' principle a crucial role within medieval society.

As a combination of contemporary distinctions and posterior classification, social classes, stratification and social types are frequently discussed. From the principal divides of ecclesiastical society and laity, nobility with its higher and lower strata, courtly and urban society with the respective microcosm of rank-related, social, institutional, religious, and ethnic distinctions among its permanent and temporary inhabitants, peasantry with various phases of unification or diversification of their status, the broad space of roads, forests, suburbs left for marginal existence where the poor, the outlaws, the pilgrims, the merchants, the hermits, the wandering knights could mingle. Within these social categories one can also observe elementary structures: family (with its gender and age divisions: changing status of women, children, and the old), 'parenthood' (understood as extended kinship-group) (*see* Family and Kinship, History of; Gender History), 'individual,' 'community,' '*gens*,' and '*natio*.'

Disruptions and transformations caused by invasions and internal wars should also be considered: attacks by, and eventually settlement of migrating tribal warrior societies (Goths, Francs, Lombards, Huns, Turcs, Avars, Vikings, Hungarians, Cumans), fights with the Arabs from the fall of Vizigothic state to the successful *Reconquista*. External menace not only disturbed but also unified: the consciousness of being '*europeenses*' emerged among Carolingians confronting the Arabic threat. Medieval Europe entered the phase of expansion with the Crusades and the evolution of long-distance trade, and the concept of European identity reappeared with the *timor Tartarorum* in the thirteenth and with the progress of the Ottoman occupation in the fourteenth and fifteenth centuries.

Throughout the Middle Ages, Europe possessed a general structure of universal bonds, which mingled with a variety of local and regional particularisms, and were in a continuous tension and conflict with them. Politically speaking, this universality was embodied in the Empire, or rather the two inheritors of the Roman Empire, the Carolingian and then Holy Roman Empire in the West and Byzantium in the East. These universal powers were confronted with a jungle of local, regional, institutional, communal and individual counterpowers. Resisting tribal retinues, territorial lordships supported by fortified castles, ecclesiastical immunities, urban liberties, noble privileges, constitutional rights, courtly intrigues, popular revolts, bloody wars paved the way to the emergence of 'national' kingdoms. The principle of *rex imperator in regno suo* finally subjugated both universal structures and local autonomies; the birth of the modern Europe of sovereign nation states was among the most significant outcomes of the Middle Ages (Strayer).

From the religious and cultural point of view the universal bond within medieval Europe equals *Christianitas*, based upon ecclesiastical structures (bishoprics, dioceses, parishes) of the Roman Papacy and the Eastern Churches with their three sacred languages. The history of medieval Christendom (*see* Religion, History of) brings a continuous expansion of internal and external conversion, defeating the resistance of the '*barbaricum*,' integrating local beliefs and cultural diversity, relying upon the omnipresent cult of saints increasingly controlled and standardized by Papacy (Vauchez, Klaniczay), the universal networks of monastic and mendicant orders, upon an emerging intellectual elite issued from medieval universities, imposing a unified pattern of literacy, Roman and Canon Law, scholastic theology, a kind of *philosophia perennis* (Gilson). Within such universal frameworks, historical evolution unfolded along the conflictuous lines of religious renewal movements (often labeled and repressed as heresies) and a series of renaissances (the Carolingian, Ottonian, twelfth century ones, and the Renaissance) which allowed an increasing absorption of Antique cultural heritage. Ongoing conflicts between 'learned' and 'popular,' Old and New Testament based, 'mystic,' 'liturgical,' 'rational,' and 'secularized' variants of Christianity

should also be mentioned. This could be studied in the universe of the cult of the saints, relics and miracle beliefs (Graus, Brown, Geary, Boesch Gajano Bartlett) or the in the prolonged conflict between 'popular religion,' archaic beliefs, 'superstitions,' heresies on the one hand and official dogma, its propagation (sermons, confession) and inquisition on the other (Bakhtin, Le Roy Ladurie, Ginzburg, Schmitt, Davis, Bériou, Elliott). Toward the end of the Middle Ages, universal frameworks could not contain the internal tensions any longer, and they succumbed, just like in politics, to the ascending nationalization of churches, legal systems, vernacular literature, and to the plurality of confessions within modern Europe, they gave way to an era of wars of religion, popular revolts, witch-hunts.

This enumeration of structures could continue with the 'hard' realities of geographic, climatic, ecological, demographic, and economic 'deep structures' (Braudel), with the 'soft' but equally enduring structures of '*mentalité*,' '*imaginaire*' (Le Goff) (*see* Psychohistory), with the 'thick description,' and the case studies of historical anthropology (Le Roy Ladurie, Ginzburg, Davis) (*see* History from Below, the History of Everyday Life, and Microhistory), with eternal dualities such as 'oral' and 'written' (Stock, Clanchy), 'body' and 'mind' (Bynum) (*see* Body, History of), with basic media such as 'text' and 'image' (Petrucci, Schmitt, Belting, Camille) (*see* Visual Turn), or with the intricacies and variabilities of old and 'new' philology (Cerquiglini). But this would require much more space, so let me rather conclude with a short and necessarily arbitrary overview of some concrete forms in which European Middle Ages live with us.

A Panorama of Narratives, Personalities, and Artifacts

The author of this article starts with grand destroyers of the 'old' (Roman, pagan Antiquity) and grand constructors of the 'new' (medieval Christian Europe). Constantine the Great (312–37) had both qualities: renouncing the divinity of the Roman Emperor, he preserved its sacrality in a Christian form, terminating the persecution of Christianity, he found a form of survival for classical culture, by shifting the center of his empire to the East, he contributed both to the thousand-year-long Byzantine continuation, and to the creation of the fertile dualism of Papacy and (an ever-renewing) Empire in the Latin West. Among the 'grand destroyers,' Attila, King of the Huns (434–53) merits precedence, his archetypal figure and his exploits became a model for all the oriental invaders till Dzenghis Khan. From among the 'grand constructors,' more should be mentioned. A model saint, turning from soldier to catechumen, hermit, monk and bishop: St. Martin of Tours (†397); two church fathers: St. John Chrysostome (337–407), Patriarch of Constantinople, and St. Augustine (354–430), Bishop of Hippo, path-breakers of personalized religiosity; an abbot, St. Benedict of Nursia, founding in 529 the model of Western monastic communities in Montecassino and writing the most influential Rule; an emperor, Justinian I (527–65), with his last attempt to restore and the first one to renovate the Roman Empire, his most enduring success was the codification of Roman Law; and finally a pope, Gregory the Great (590–604), organizer of medieval papacy, church hierarchy, liturgy, conversion.

In the dim light of early medieval legends, annals, and chronicles one perceives influential converters, as St. Patrick (390–461), the apostle of the Irish or St. Boniface (675–754), the apostle of Germany; churchmen making heroic efforts for salvaging classical culture, as Cassiodore (485–580), the Roman senator turned monk or Isidor of Sevilla (560–636) the Vizigoth bishop making a synthesis of antique learning in his *Etymologies*; impressive rulers, as the Frank Clovis (481–511), the Goth Theoderic (493–526), or the Northumbrian St. Oswald (634–42), powerful Merovingian queens as the wicked Fredegund, or the pious St. Radegund (520/5–87). The collective memory also preserves the narratives of the '*origines gentium*,' barbarian migration, conversion, and dynastic myths of emerging ruling houses (claiming divine origin). Following *Germania* of Tacitus, we hear thus of the Goths, the Franks (with the Merovingian dynasty), the Anglo-Saxons (with their Woden-sprung kings), and the Lombards (or later legends and stories related to the Scandinavian Yngling, the Slavic Přemysl, Piast and Rurik houses, the migrating Hungarians with their Arpadian dynasty).

A new chapter is starting with the Carolingians: the emergence of an influential family from the position of *majordomus* to royal and imperial dignity. Charles Martel stopping the Arabic invasion at Poitiers (732), Pepin the Short (741–68) and Charlemagne (768–814) making alliance with the Papacy remunerated by the imperial crown by Pope Leo III (800) – a symbolic moment in medieval history (McKitterick), which will find its counterpoint in 1077 in the Canossa – penance of Henry IV. The Carolingian renewal of state – church structures and social networks of dependence, the take-off in demographic, agrarian and urban growth, the 'renaissance' of high culture within closer courtly and monastic circles, the successful resistance against invasions (Avars, Arabs, Normans, Hungarians) – all this laid the foundations of a new triumphant evolution, unfolding in the West around the year 1000. There is a similar new start in the East with the Macedonian dynasty (867–1056), reorganizing the Byzantine Empire, successfully confronting Bulgarians, Hungarians, Russians, systematizing the army, the administration, and the ceremonies in the times of Leo VI and Constantine VII.

As for artifacts, while the splendor of Byzantium was incontestable, and radiated even in the remote mosaics of Ravenna and Venice, in the West early Middle Ages were not spectacular but essential: a massive diffusion of horseshoes, iron plow, water-mill, evolving naval technology of Vikings or Italian merchants. In culture new technologies of writing, such as the Carolingian minuscule book-illuminations, Gregorian chants, goldsmith works, evolving monastic and urban architecture, fortified castles indicate a new evolution.

Around the year 1000, one ruler stands out: the 'marvel of the world,' the Saxon Otto III, third emperor of the Holy Roman Empire (983–1002), allied to his French teacher Gerbert d'Aurillac, Pope Sylvester II (999–1003), and the Czech St. Adalbert (†997), the 'Apostle of the Slavs.' They coined a powerful new political vision for *Europa Occidens*, relating the integration of Germanic North and Mediterranean South to the extension of the influence of Latin Christianity toward Slavic and Hungarian East. Relying upon the earlier

conversion of Croats (Tomislav – 924) and the Czech lands of St. Venceslas (†929), now Poland of Bolesław the Brave (992–1025) and Hungary of St. Stephen (997–1038) also become parts of this 'new Europe,' which also extends toward the North. The Danes of Harald Blueteeth (940–85) and Sweyn Forkbeard (985–1014), the Norwegians of St. Olaf (1015–30) and later the Swedish get converted, with Anglo-Saxon and Germanic influence. This Christianization is conjugated with a parallel activity from Byzantium: on the paths of the previous ninth century Cyrillo-Methodian missions in Bulgaria and Moravia, in 996 Kievan Rus of St. Vladimir (980–1015) is converted.

Medieval Europe is profoundly renewed after the turn of the Millennium: new dynasties (Capetians, Salians, Normans in England and Sicily); reshaped religious life under the influence of Cluny; 'reform papacy' of Gregory VII (1073–85); Crusades proclaimed by Pope Urban II in 1095. Correlated with agricultural growth, population nearly doubles in two centuries, medieval towns multiply and acquire important immunities, liberties, a new dynamism unfolds.

Twelfth century brings the explosion of 'modernity,' of fertile tensions and conflicts, and of a wide array of grand figures. Abelard (1079–1142), the daring and unfortunate knight of dialectics, self-confident in his reliance upon reason, archetype of the medieval intellectual; St. Bernard of Clairvaux (1090–53), his opponent, the charismatic leading figure of the Cistercians, founding father of the Templars and organizer of the Second Crusade (1147–49), abbot, politician, healer, and mystic; Manuel I Komnenos (1143–80) Byzantine Emperor who tries to restore the ancient splendor of the East by starting a chivalrous rivalry with the crusading West; Arnold of Brescia (+1155), student of Abelard, apostolic preacher and tribune of a republican Rome that he hopes to resurrect, but hoping in vain the help of Frederick Barbarossa (1152–90), the great emperor from the Hohenstauf dynasty, who is rather for the *renovatio imperii* and gets him executed for his heresies; the martyr bishop St. Thomas Becket, who confronted his hard-willed youthful friend, Henry II Plantagenet (1154–89) and paid with his life for it in 1170. This is also a century of women: Heloise (*c.*1100–63), the seduced student, lover and wife of Abelard, who persisted in intellectual and emotional partnership after Abelard got castrated by her angry relatives; Eleanor of Aquitaine (1122–1204), grand-daughter of the first *troubadour*, Duke William IX of Aquitaine (1087–1127), herself also patron of poetry, wife of Louis VII, King of France (1137–80), then divorcing him in 1152 to marry Henry II Plantagenet, and take a third of France with herself as dowry; Hildegard of Bingen (1098–1179), the visionary nun of Rupertsberg, most prolific women-writer of the age.

Twelfth century is that of urban handicrafts, flowering fairs, schools, the beginning of the construction of cathedrals. It is also the century of crusades and knighthood under the banner of St. George, the 'dragon-killer,' Roland, King Arthur and the knights of the Round Table; it is the century of orthodox or heretic itinerant preachers from Peter the Hermit, summoning to crusade, Robert d'Arbrissel (†*c.*1117) stepping in for the emancipation of women, to Waldes, the rich Lyon merchant who distributed his wealth to the poor in 1173, and decided to live in mendicancy. When confronted with interdiction he accepted with his disciples, the Poor of Lyon, persecution rather than submission. Another heretic movement, the Cathars, were remote followers of the antique Manichaean dualists, more immediately related to the South-East European Bogomils.

At the beginning of the thirteenth century Christianity seemed suddenly to explode from conflicts: ongoing struggles between Empire, where Frederick II, the last great Hohenstauf (1212–50) was preparing for his *Ghibellin* exploits, and the Papacy which became a fearful rival with the *Guelph* dominated Italian towns on his side. Cities made international politics in the Mediterranean and in the North Sea: the Fourth Crusade was diverted by Venice to conquer Byzantium (1204), in the North the Hansa alliance, the future master of Scandinavian-Baltic region was growing. New problems arose within the cities: confraternities, 'people's parties,' lay preachers, rebellious students in Paris, Oxford and Bologna universities, or bank crashes in Florence and Siena. The fight against the heretics took the form of the crusade against 'Albigensians' (1209–29), which subjugated Southern France to the North; France being also in war with England and the Empire, with the victory by Philip August II (1180–1223) at Bouvines (1214). Noblemen were confronting their rulers to gain constitutional rights (Magna Carta – 1215; Hungarian Golden Bull – 1222). This burgeoning situation could be mastered with the help of two religious reformers, the founders of the two mendicant orders, St. Dominic (1170–1221) and St. Francis (1181–1226), and two popes, Innocent III (1198–1216) and Gregory IX (1227–41), who were capable of restructuring old resources and integrating new forces.

With the assistance of the Dominicans and the Franciscans the challenge of heresy could be confronted, the position of the Church got re-established in the cities where the pride of the emerging cathedrals and the accumulating new richness got balanced by the religious control of fashion, luxury, and usury and by charitable institutions of burghers moved by their '*alter Christus*,' the Poverello of Assisi. The Dominicans helped setting up the fearful Inquisition (1231) against the heretics and the popular 'superstitions,' in learning found balance between reason and faith, Aristotle and the Bible (St. Thomas Aquinas: *Summa Theologiae*, *c.*1267–73), and in princely courts Christianized rulership. The practical morality of their 'mirrors of princes' found echo in the ears of St. Louis IX of France (1226–70), most pious ruler ever, or his ambitious brother, Charles of Anjou, King of Naples and Sicily (1266–85), and many others. The religious message of the mendicant orders got also amplified by pious princesses: St. Hedwig of Silesia (1174/8–1243), St. Elizabeth of Thuringia (1207–31), Blanche of Castilia (1188–1252), St. Agnes of Bohemia (1205–82), St. Margaret of Hungary (1242–70); and daughters of Italian merchants, bankers, burghers: St. Claire of Assisi (1194–1253), Umiliana dei Cerchi (1219–46), Margaret of Cortona (1247–97) or Angela of Foligno (1248–1309).

By the end of the thirteenth century medieval Europe was prospering: after the successes of machine-supported cloth and textile manufactures, set to produce mechanic clocks, silk, paper, developing steel industry, started to use gold coins again. Though the devastating attack by the Mongolians

(1240–41) in the East (subjugating Russian principalities), the rise of the Turcs in the Near East represented a serious menace, this strengthened the cohesion of an *Occidens* now better including Poles, Czechs, and Hungarians; sided by a new South-Eastern Europe, a second Bulgarian empire of the Asens, a reborn Byzantium of the Palaeologues (1261–1453), an emerging Serbian state of the Nemanjides. A new expansion is imminent: ambassadors and travelers reach the Far East (Marco Polo: 1254–95), and Aragon is setting its regard upon North Africa, Norway is incorporating Island and Greenland.

The fourteenth-century is traditionally associated with 'crisis': famines, economic decline, the Black Death (1347–52), flagellant movements, pogroms, extinction of many royal dynasties (Arpadians – 1301, Přemyslsides – 1309, Capetians – 1328, Piasts – 1370), 100 Years' War (1337–1453), ravaging popular revolts (Jacquerie – 1358; Wat Tyler uprising – 1381). This is all true, but the image of 'waning' is wrong. It is only partly valid for the West, and is counterbalanced by the prosperity in Polish, Czech, Hungarian and Serbian Central Europe. The Prague of Charles IV (1346–78) becomes now the center of Europe. Drawing upon the richness of the Low Countries, the splendid court of Burgundy is living its 'epics of temerity and pride' (Huizinga). Despite the ravages of the plague, Italian cities remain prosperous. Following Dante (1265–1321), and Giotto (1266–1337), and Petrarch (1304–74) who saw himself already beyond the 'dark' Middle Ages, the *Decameron* of Boccaccio (1250) and *The Canterbury tales* by Chaucer (1390s) offer a rich synthesis of medieval narrative tradition. With all its crises, schisms, and new heresies, with the impact of Meister Eckhardt (1260–1328), St. Bridget of Sweden (1303–73), St. Catherine of Siena (1347–80), Wyclif (1330–84) and Hus (*c.*1370–1415) Christianity is preparing for the renewal of personal and secular religiosity.

The formidable new tensions and the grand achievements of the fifteenth-century represent the concluding period of this long Middle Ages. The power of the emerging new prophecies and religious anxieties is demonstrated by the explosion of the Hussite rebellion, by Joan of Arc who brings in 1429 a decisive turn in the Hundred Years' War, and by the beginnings of witch-hunts in Switzerland, Italy, and France. The political map of Europe is reshaped by the emergence of powerful new states. The Ottoman conquest is only temporarily slowed down by John of Hunyad and the crusade organized by St. John Capistran at Belgrade (1456). The formation of a multinational Central European empire, an endeavor of the Anjous, the Luxemburgs, the Jagiellonians, and Mathias Corvinus (1458–90), is finally realized by the Habsburgs with Maximilan I (1493–1519). A large Eastern Europe re-emerges with the Polish-Lithuanian Confederation and the new Russia of Ivan III (1462–1505). In the West, an impressive new state candidate, Burgundy, succumbs finally, and besides France of Louis XI (1461–83) and England of the Tudors after the Wars of Roses, a unified Spain of Isabel of Castilia (1474–1504) and Ferdinand of Aragon (1479–1516), and a miraculously ascending Portugal will be the states which lead the world-wide expansion of Europe in early modern times. Germany and Italy, on the other hand, have the largest share of inventions, richness, and culture, which made this expansion possible: the invention of printing by Gutenberg (*c.*1450), the banks of the Medici and the Fugger, the map of Toscanelli (1474) and the bold expedition of the Genovan sailor, Columbus (1492). The narrative must end here: a new story begins in the enlarged new world.

See also: Classical Legacies; Family and Kinship, History of; Feudalism; Gender History; Renaissance.

Bibliography

Auerbach, E., 1958[1993]. Literary Language and its Public in Late Latin Antiquity and in the Middle Ages. Princeton.

Bakhtin, M., 1968. Rabelais and His World. Cambridge, MA.

Bartlett, R., 1993. The Making of Europe. Conquest, Colonization, and Cultural Change, 950–1350. Princeton.

Bartlett, R., 2013. Why can the dead do such great things? Saints and Worshippers from the Martyrs to the Reformation. Princeton.

Belting, H., 1994. Likeness and Presence: A History of the Image before the Era of Art. Chicago.

Benson, R.L., Constable, G. (Eds.), 1991. Renaissance and Renewal in the Twelfth Century. Toronto.

Berend, N. (Ed.), 2007. Christianization and the Rise of Christian Monarchy. Scandinavia, Central Europe and Rus' *c.*900–1200. Cambridge.

Bériou, N., 1998. Lavènement des maîtres de la Parole. Paris.

Bloch, M., 1939[1989]. Feudal Society. London.

Boesch, G.S., 1999. La santità. Roma.

Branca, V. (Ed.), 1973. Concetto, storia, miti e immagini del Medio Evo. Firenze.

Braudel, F., 1972–1973. The Mediterranean and the Mediterranean World in the Age of Philip II. Harper & Row, New York.

Brown, E.E., 1974. The tyranny of a construct: feudalism and historians of medieval Europe. American Historical Review 79, 1063–1088.

Brown, P., 1978. The Making of Late Antiquity. Cambridge, MA.

Brown, P., 1981. The Cult of the Saints. Its Rise and Function in Latin Christianity. Chicago.

Burckhardt, J., 1860[1954]. The Civilization of the Renaissance in Italy. New York.

Bynum, C.W., 1982. Holy Feast and Holy Fast: The Religious Significance of Food to Medieval Women. Berkeley.

Bynum, C.W., 1995. The Resurrection of the Body in Western Christianity, 200–1336. New York.

Camille, M., 1989. The Gothic Idol: Ideology and Image-Making in Medieval Art. Cambridge.

Cerquiglini, B., 1989. L'éloge de la variante. Histoire critique de la philologie. Paris.

Clanchy, M.T., 1993. From Memory to Written Record, England 1066–1307. Oxford.

Curtius, E.R., 1963. European Literature and the Latin Middle Ages. New York-Evanston.

Davis, N.Z., 1975. Society and Culture in Early Modern France. Stanford.

Delehaye, H., 1927. Sanctus. Essai sur le culte des saints dans l'Antiquité. Bruxelles.

Duby, G., 1980. The Three Orders: Feudal Society Imagined. Chicago.

Duby, G., 1981. The Age of the Cathedrals: Art and Society, 980–1420. Chicago.

Duby, G., Perrot, M., Pantel-Schmitt, P. (Eds.), 1992–1994. A History of Women in the West. Cambridge, MA.

Dvornik, F., 1949. The Making of Central and Eastern Europe. London.

Eco, U., 1980 [1983]. The Name of the Rose. New York.

Elliott, D., 2004. Proving Woman. Female Spirituality and Inquisitional Culture in the Later Middle Ages. Princeton.

Folz, R., 1969. The Concept of Empire in Western Europe from the Fifth to the Fourteenth Century. London.

Folz, R., 1984. Les saints rois du Moyen Age en Occident, VIe–XIIIe siècles. Bruxelles.

Geary, P.J., 1990. Furta Sacra: Thefts of Relics in the Central Middle Ages. Princeton.

Geremek, B., 1994. Poverty: A History. Oxford.

Gieysztor, A., 1997. L'Europe nouvelle autour de l'An Mil. La Papauté, l'Empire et les "nouveaux venus". Roma.

Gilson, E., 1932. L'esprit de la philosophie médiévale. Paris.

Ginzburg, C., 1980. The Cheese and the Worms: The Cosmos of a Sixteenth Century Miller. Baltimore.

Graus, F., 1965. Volk, Herrscher und Heiliger im Reich der Merowinger. Studien zur Hagiographie der Merowingerzeit. Prague.

Graus, F., 1988. Pest, Geissler, Judenmorde: Das 14. Jahrhundert als Krisenzeit. Göttingen.

Gurevich, A.I., 1995. The Origins of European Individualism. Oxford.
Halecki, O., 1952. The Borderlands of Western Civilization. New York.
Huizinga, J., 1919[1996]. The Autumn of the Middle Ages. University of Chicago Press, Chicago.
Kantorowicz, E.H., 1957. The King's Two Bodies: A Study in Mediaeval Political Theology. Princeton.
Kristeller, P.O., 1974. Medieval Aspects of Renaissance Learning. New York.
Klaniczay, G., 2002. Holy Rulers and Blessed Princesses. Dynastic Cults in Medieval Central Europe. Cambridge.
Le Goff, J., 1980. Time, Work, & Culture in the Middle Ages. Chicago.
Le Goff, J., 1988. The Medieval Imagination. Chicago.
Le Goff, J., Schmitt, J.-C. (Eds.), 1999. Dictionnaire raisonné de l'Occident médiéval. Paris.
Le Roy Ladurie, E., 1985[1979]. Montaillou. The Promised Land of Error. Random House, New York.
Linehan, P., Nelson, J. (Eds.), 2001. The Medieval World. London.
Maitland, F.W., 1908. The Constitutional History of England. Cambridge University Press, Cambridge.
McKitterick, R., 2008. Charlemagne: The Formation of a European Identity. Cambridge.
Michelet, J., 1833[1844]. The History of France. London.
Obolensky, D., 1982. The Byzantine Commonwealth. Eastern Europe, 500–1453. Crestwood, NY.
Ostrogorski, G., 1969. History of the Byzantine State. New Brunswick.
Panofsky, E., 1960. Renaissance and Renascences in Western Art. New York.
Patlagean, E., 2008. Un Moyen Age grec, Byzance IXe–XVe siècle. Paris.
Pertz, G.H., 1826. Monumenta Germaniae Historica. Hannover and Leipzig.
Pirenne, H., 1933[1937a]. Economic and Social History of Medieval Europe. New York.
Pirenne, H., 1937b[1965]. Mohammed and Charlemagne. New York.
Poly, J.-P., Bournazel, E., 1991. The Feudal Transformation: 900–1200. New York.
Reynolds, S., 1994. Fiefs and Vassals: The Medieval Evidence Reinterpreted. Oxford.
Schmitt, J.C., 1983. The Holy Greyhound, Guinefort, Healer of Children since the Thirteenth Century. Cambridge.
Schramm, P.E., 1929[1975]. Kaiser, Rom, und Renovatio, Studien zur Geschichte des römischen Erneuerungsgedanken vom Ende des karolingischen Reiches bis zum Investiturstreit. Leipzig.
Sergi, G., 1998. L'idée de Moyen Age. Entre sens commun et pratique historique. Paris.
Southern, R.W., 1970. Western Society and the Church in the Middle Ages. London.
Stock, B., 1983. The Implications of Literacy. Written Language and Models of Interpretation in the Eleventh and Twelfth Centuries. Princeton, NJ.
Strayer, J., 1970. On the Medieval Origins of the Modern State. Princeton.
Szűcs, J., 1990. Die drei historischen Regionen Europas. Frankfurt am Main.
Tuchman, B., 1978. A Distant Mirror: The Calamitous 14th Century. New York.
Vauchez, A., 1997. Sainthood in the Later Middle Ages. Cambridge.

Middle East, North Africa: Sociocultural Overview

Seteney Shami, Social Science Research Council, Brooklyn, NY, USA; and The Arab Council for the Social Sciences, Beirut, Lebanon

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Middle East, North Africa region, presents a complex heterogeneity of historical, social and political interconnections. This article begins with a brief overview of earlier scholarly approaches to the region and then focuses on the 1990s as a turning point that launched a new literature based on critical readings of both Orientalist and developmentalist approaches. Works over the past two decades have provided exciting insights into the workings of nation, gender, and religion in the region. However, there is still insufficient attention to issues of identity and difference that structure these societies and fuel much of its dynamics.

The term 'The Middle East and North Africa' (MENA) refers to a territory characterized by considerable diversity, extending from Morocco in the West to Iran in the East, spreads over Africa and Asia, and extending into Europe. This is a region that has given rise to the earliest forms of urban life and state organization in human civilization, produced three major monotheistic religions (Judaism, Christianity, and Islam), includes three major language groups (Arabic, Persian, and Turkish) as well as countless other languages and dialects, and exhibits sharp contrasts in physical and human geography. Since AD eighth century, successive Islamic empires have ruled and to some extent united this geographical expanse, the most recent of which was the Ottoman Empire. In other words, this is an area of great complexity, with its long historical record, literate traditions, and mixture of social and cultural groups.

In social science and humanities scholarship, this expanse was often seen as a region broadly sharing linguistic, religious, and cultural characteristics as well as a common geopolitical position in the world order. Yet, because of its complexity and diversity, the term and the territory associated with it are coupled in a shifting, unstable relationship. Afghanistan is sometimes included, sometimes not. The Sudan would fall in or out depending on whether the emphasis is on the African heterogenous south or the Arab Muslim north and it will be interesting to see how its recent division into two countries will impact the identification of North and South Sudan. Turkey is usually included, although many of its intellectuals and politicians prefer to find continuities with Europe and the Balkans, and/or with Central Asia.

The first part of the term, the Middle East, originates from nineteenth-century European, specifically British, strategic military divisions of the world. Other terms also compete for the same territories or carve it up in different ways: older terms such as the Near East or the Levant still have occasional currency, the latter having been proposed as a substitute term that emphasizes linkages, rather than boundaries, between the societies and cultures of the Greeks, Turks, Arabs, and Jews (Alcalay, 1993). Similar to this are arguments for the Mediterranean as a cultural unit. The term 'the Arab world' is often seen, especially indigenously, as more accurate in depicting an analytically coherent unit. In Arabic, the historical distinction between the Maghreb (the West, or North African countries) and the Mashreq (the East, or the Eastern Mediterranean countries) is widely used and this is reflected in French scholarship on the region, as exemplified by the journal Maghreb-Machrek. In German, the term 'the Orient' still largely refers to Islamic and Arab societies and the disciplines and learning related to them.

New regionalisms also compete with these designations in different ways. The renewal of historical and religious connections with Central Asia and the Caucasus, after a rupture of more than a century of Russian and Soviet rule, gives rise to new/old redrawing of the boundaries of the region. The freeing of scholarly imaginations by theories of globalization leads to the investigation of historical and contemporary links across the Indian Ocean with South Asia and Southeast Asia. Finally, studies of diaspora communities (e.g., Turks in Germany, the Lebanese in South America and Australia, the North Africans in Europe, Arab Americans, and so on) are beginning to look at translocal and transnational linkages that are also part of the making and unmaking of the region.

In the last decade or so, a renewed focus on 'Islam' perceived as a unifying and even homogenizing force has also redrawn scholarly boundaries. Despite the demographic reality that the largest numbers of Muslims do not live in the Middle East or North Africa, the region is inextricably linked with Islam as its historical fount and heartland. The ways in which some scholars now subsume Central Asia and parts of South Asia within the Middle East, on the basis of the 'Islamic nature' of their societies, as well as the resurgence of the term 'the Islamic World' shows how Islam is considered as a unifying force that brings together societies and histories with widely disparate historical experiences. Yet, the question of whether Islam, or any religion, can be seen as providing the sociocultural cohesion of a geographical region remains a problematic issue, as does the question of where the cohesive quality of Islam is to be located. Islam as a belief system? Or an institutional framework? Or historical experience? Or political ideology? The different answers lead toward different routes of comparisons and distinctions (Asad, 1986).

Turning Points

Lila Abu-Lughod's (1989) seminal review essay in the *American Review of Anthropology* marked a turning point in scholarship as it examined the important moves, theoretical and topical, that research on sociocultural issues in the MENA region had

made. Her three 'zones of theory' identified the dominant topics of the day to be as follows: tribe, women, and Islam. While she made clear the basis for these preoccupations in Orientalism as well as disciplinary 'prestige zones,' she also identified the advances that had begun to unmake these powerful scholarly legacies. Breakthrough works such as Edward Said's *Orientalism* (1978) opened up the door for a critical examination of the relations of knowledge and power that accompanies the dichotomy between East and West. Works began to meet the challenge of examining and historicizing the politics of representation, the most prominent being Timothy Mitchell's *Colonizing Egypt* (1988). From another angle, modernization theory and 'developmentalism' were being 'unmade' through new works in political economy and modes of production and through revising the received wisdom on the Ottoman and colonial periods, by focusing on the incorporation of the region into the world capitalist system as well as on nationalism and ideological currents in the area.

As one examines the scholarship that these new directions and frameworks yielded, one can identify not only continuities but also important shifts from Abu-Lughod's three zones of theory. Thus, works produced in the last two decades can be grouped under three main categories, which have seen the most growth in writing on the region, namely, nation, gender, and religion.

Nation

The ways in which national identity has been constructed in different parts of the region has been explored mainly in Turkey, Egypt, and Israel/Palestine, with some substantial work developing on North Africa. It is important to emphasize that work on the nation and nationalism, largely inspired by the work of Benedict Anderson (1983), is fairly new to sociocultural disciplines such as anthropology, whether in this region or elsewhere. This focus thus represents more a disjuncture than a shift from 'tribe' (as a metonym for all 'bounded' and 'local' groups) to 'nation,' and the transition between literatures has not been smooth. While rightly dismissing evolutionary paradigms (from tribe to nation), the new literature has largely not been able to address how the structures of the state are related to the production of locality and localism. More attention as to how the nation is imbricated in the local (and vice versa) would reveal the coexistence of alternative identities as well as differing readings of the nation by various social groups and communities.

Ethnographic work does focus, and in highly interesting ways, on history and memory as well as constructions of social categories such as the peasantry, notions of secularism and modernity, and the role of knowledge production and 'science' in the consolidation of nations (see Abu El-Haj, 2001). Also of interest are works on the past and present strategies of elites, roles of intellectuals and various occupational groups, the mobilization of different population groups, and the role of the media and other modes of communication in the construction of nationalism and of national and transnational public spheres (Shami, 2009). The work on expressive culture, distinction (in Pierre Bourdieu's sense), and representation is also growing both in anthropological and social historical works. The history of Arab nationalism presents a particularly interesting case, given that it is simultaneously produced in different locations such that inhabitants of 22 countries identify themselves as Arab in addition to identifying as Syrian, Egyptian, Moroccan, or Saudi (see *International Journal of Middle East Studies*, 2011). Arab nationalism is particularly illustrative of the transnational processes underpinning the production of nation, as well as the tensions produced by overlapping, and even competing, nationalisms.

Arguments about the nation in the Middle East have been especially interesting where they concern the intersection of the individual and the national, the private, and the public selves. Also significant here is the literature on the politics of memory and the ways in which the stories people recall transform individual memory to something more public and national. Ethnographic narratives also contribute to the understanding of the role of the arts and media in the construction of nationalism. The instrumental cultural and political roles of entertainment media, for example, television melodramas, in the production of national culture and the molding of individuals into modern national citizens, are the focus of a growing number of works.

While the literature on nation focuses on the construction of a national and collective identity, especially in the ways that body, self, family, sociability, and citizenship/participation are transformed, less explored are the political and discursive processes of foregrounding certain identities and erasing others that lie at the heart of any national project. National identity is premised on the existence of a 'majority' with varying degrees of tolerance for those who would then be cast or cast themselves as a 'minority.' Those who embody the nonmajoritarian identity are often left understudied, making ethnographers complicit with the nations that they study in imagining that a majority exists and is fashioned out of a collectivity that is premodern. In other words, ethnic identity and minority populations, if noted, are seen as vestiges of the premodern, and thus displaced and out of place in the nation.

It is in places where the contradictions between past and present are too sharp to ignore, and where the politics of recognition throws up powerful competing claims, that national projects can be best deconstructed through critical research. The palimpsest that is Palestine is the prime example. In the story of the town of Ein Houd/Ein Hod, Susan Slymovics (1998) gives an account of two peoples, Palestinian Arabs and Jewish Israelis, who cultivate the link between memory and place, and the sense of loss permeates the narrative at every level. Her book, *The Object of Memory*, is not only about dispossession but also about active remembrance. Slymovics situates Palestinian memory in the larger genre of memorial books produced by those who have suffered war, dispersion, and traumatic loss as they strive to construct a nation, a community, through the links between memory and place (see Naguib, 2008). Other works on Palestine/Israel, whether focusing on places or people, begin to unpack the overlapping and contradictory processes of nation-making and the inclusions/exclusions and fashioning/erasures that are at work.

The literature on Turkey is similarly fascinating in both erasing and celebrating Kurdishness depending upon context,

thus echoing political positions often taken by progressive secular elites. Many of the excellent works on Turkish nationalism that focus on state projects of remaking gender, family, domesticity, and citizenship somehow elide the issue of ethnic differences, although this becomes manifestly the object of research when studying how Kurdishness is constructed against an imagined Turkishness.

Recent works on Sudan, the Gulf States, Syria, and Cyprus all have begun to explore the fluidity of identities and the politics of ethnicity and nationalism, making clear the urgency of such processes. While Lila Abu Lughod (1989) rued the focus in anthropology on 'peripheral' places (such as Yemen and Morocco), the power of the margins in unmaking hegemonic discourses (including academic discourses) is never clearer than when looking for traces of occluded identities.

Gender

The literature on gender roles and relations has perhaps witnessed the most robust growth and generated interesting discussions on the region. Starting in the 1970s, particularly following Cynthia Nelson's 1974 article in which she challenged the stereotyped images of women succumbing to patriarchal rule with docility and passivity, the theme of women as victims of patriarchy and Islam has been replaced by narratives of women as agents who negotiate their own survival. This survival takes place, in part, by 'bargaining with patriarchy' (Kandiyoti, 1996). A burst of work on women appeared in the 1980s, as covered in Lila Abu-Lughod's 1989 *Annual Review of Anthropology* review, which aimed at moving beyond simply debunking stereotypes to constructing theory. Contemporary trends in the literature further move from a focus on women to the study of gender, including the construction of gender identity and notions of femininity and masculinity; the fashioning of the body through economic power, political participation, violence, and war; and women's enmeshment in national and transnational processes. This work has yielded insight not only into how male and female persons are culturally constructed, but also into female contributions to the construction of larger social identities. Interestingly, recent work focuses mostly on the Arab 'Mashreq' region, with a heavy focus on Egypt, and there are astonishingly few ethnographies of gender in Turkey. Anthropological research on Iran remains difficult under current political conditions, and the work on gender is often through the lens of 'youth' or takes cultural production, and especially film works, as its main ethnographic texts (Varzi, 2008). However, Homa Hoodfar has important works on gender in relation to Islam, citizenship, and reproduction in Iran (2000, 2001).

Thus, a number of scholars have examined the lived experience of gender subordination in a hierarchical system, including work on ritual, marriage, and households and socializing practices. Women's political activism and participation, particularly under conditions of conflict, has also formed the focus of a number of noteworthy monographs.

Gender relations and gender identity have also been a central focus of medical anthropology in the region in the past 20 years, largely with a focus on Egypt. The relationship between gender differentiation and ideas of sickness and health in rural Egypt has been examined as well as reproduction and motherhood as major practices through which women create their gender identity, impacted by other vectors of identity such as class and religion. Fertility and reproduction articulates with national identities and state policies for shaping citizens and several works in medical anthropology have included a focus on maleness and the body and bring in regional, transnational, and global processes as well. Studies include men's roles in systems of hierarchy and domination, whether within communities of study, in a national context, or in contexts of national or global conflict and inequality. There has also been a trend toward studying male and female roles in gender relations concurrently, particularly within the family as highlighted by Suad Joseph and Susan Slyomovics in their introduction to *Women and Power in the Middle East* (2001).

As can be discerned, an important theme in much of the above-mentioned work attends to the ways in which gender identity articulates with communal and national identities on the one hand, and global processes on the other. The two most disappointing aspects of this otherwise exciting literature are, first, its almost exclusive focus on Muslim women, and, furthermore that, in the exploration of family, nation, body, etc., various types of differences other than gender differences are erased. This makes the work of Anita Fábos on the Sudanese in Cairo quite unique (2008), with its focus on *adab* or propriety and practices such as female circumcision in the context of the cultivation of Sudanese identity in the particularly ambiguous zone of Egyptian–Sudanese relations. Also revealing is the work on female migrant domestic workers of Southeast Asian origin working in Arab families (Moors et al., 2009). The analysis of how the incorporation of 'foreign' women into the intimate domestic sphere brings into question national, racial, and religious identities is promising for exploring a new challenge to ethnic boundaries, those within the home itself and in the 'private' sphere. While hitherto the focus in this regard has been on religious dimensions, especially when differing practices of Islam come face to face within the domestic sphere, the same perspective can be applied to other aspects of gender relations and roles.

It has long been recognized that gender relations and gender identities are central to constructing and perpetuating collective identities, including national, ethnic, and minority identities. For the MENA region, this has been often explored through visual culture, for example, in the excellent works by Viola Shafik (2007a, 2007b) on Egyptian cinema. One looks forward to a more robust study of gender dynamics as they intersect and interact with different types of collective identities on the one hand and local contexts on the other.

Religion

The works on gender and religion intersect in important ways, as a major trend in the past 20 years has been a focus on women and Islam, with a growing recent interest in Shi'a women. Both Ziba Mir-Hosseini (1999) and Lara Deeb (2006) examine the construction of religious discourse and women's

participation in constructing discourse about themselves. While Mir-Hosseini focuses on the construction of religious discourse about women, Deeb considers how women themselves participate in authentication, or actively defining faith in daily life and discursively establishing religious authority not only about women but also about Islam. The ideal for women in the religiously observant Shi'i community she studied was that of public piety.

Several major interventions on Islam have influenced much of the literature of the past 20 years, for example, Talal Asad (1986, 1993), Michael Gilsenan (1982), Sami Zubaida (1995), and Daniel Varisco (2005). Unsurprisingly, there is particular ethnographic attention to Islamism or fundamentalist Islamic movements from their appeal to 'vernacular politics' (White, 2002) to women's piety movements (Mahmood, 2001, 2005). Mahmood questions ethnographic depictions of women's agency and feminist goals and argues that there is a tendency to look for resistance and challenges to male domination even when they are difficult to locate. Based on research with a women's organization in Egypt, Mahmood calls for a notion of agency that can account for women's active pursuit of ideals considered illiberal and regressive within a Western feminist discourse, such as certain Islamic ideals of female piety and modesty. Her ethnographic study exposes assumptions of both feminist theory and 'secular-liberal thought' about the nature of the self, agency, and politics, as well as assumptions about the relationship between interiority and performance.

From the other end, state responses to Islamist movements and governments' attempts to shape a particular type of Islamic identity are also the focus of study. The most thorough work to consider religion and politics in a multisectarian environment is Joseph's (2000) research in Lebanon examining how sect endogamy and the 'civic myth' of sectarian pluralism obscures the working of kin endogamy and patriarchy encoded in personal status law and sanctioned by the state. Hale's work on women's participation in both secular and Islamist politics in Sudan is exceptional in its scope. Rather than focusing her gaze on one or the other, Hale productively examines how both secular (Communist) and religious (Islamist) national parties objectify women (Hale, 1996).

In spite of this literature, in comparison to other topics, there are fewer monographs focusing on religious identity than the prominence of religious revival movements in the region since the 1970s would seem to warrant and much of the literature neglects the question of how different groups in society may articulate their Islamic identity and piety in ways that vary significantly from one another. Even works supported by rich and valuable ethnographic detail sometimes seem to treat 'Islam' and the 'Islamic Revival' as monolithic categories and processes. A neglected topic within the study of the resurgence of both political and personal Islam is how changing perceptions and practices of Islam intersect with other identity factors, including ethnic and linguistic identities. It is important to note that Islamic identity may act as a unifier across ethnic and other differences in some contexts, while in others it can conversely exacerbate such differences as well as fuel competing sectarian identities. The lack of studies on interreligious interaction or in-depth ethnographies of communities with mixed religious populations is almost inexplicable. Thus Christian and Jewish communities, when studied, are usually dealt with as though separate and self-contained.

Identities and Differences

While the literature on nation, gender, and religion (read: Islam) examines important major sociocultural processes in MENA societies, there is a general tendency to collude with the 'national order of things.' The 'nation' is discussed as the property of the majority with little attention to minority groups, or how notions of 'majority' and 'minority' are constructed in the first place. While the literature on gender identities, roles, and relations is rich and nuanced, the social and cultural context is almost always an 'unmarked' Muslim national grouping. When religion is the focus, Islam clearly dominates work on the region and remains mostly on the majority religious group (Sunni or Shi'a) to the occlusion of other Muslim and non-Muslim groups in the country or region.

The relational quality of identity construction and the structuration produced by the construction of difference are usually mentioned cursorily if at all. Religious minorities for the most part are only seriously discussed if they are the specific group under study. Thus, ethnographies that do discuss non-Muslims, mainly Christians, tend to treat them almost incidentally as part of the broader society. For example, there has been no ethnographic work on religious practice among Christians in the MENA region. There are almost no cultural analyses of how nationalism is perceived and reproduced (or resisted) among non-Muslim groups.

A small literature does specifically focus on particular ethnic, religious, and linguistic groups, clustered around quite specific geographies and groups. In the work on Iran and Afghanistan, there is a literature on various ethnic and linguistic groups. The fact that many of these constitute quite distinct pastoral nomadic groups constitutes both a problem and an interesting trajectory of research. As an illustration, while nomad–settler relations constitute an important focus in ecological and economic research on various hinterlands, few such studies take into account the often different ethnic and linguistic identities of the nomads and settlers involved. In contrast, earlier work on Iraq, such as by Amal Rassam, also known as Vinogradov, showed great promise in focusing on the mutual brokering of economic and ethnic relations and social hierarchies in rural settings (Vinogradov, 1974; Rassam, 1977). Current work on Kurds and Kurdistan, including historical works, similarly shows promise despite being few in number.

Another focus of research that promises new directions in the understanding of identity and difference is on Sudan and Egypt. A flurry of research was carried out in the 1960s by anthropologists at the American University in Cairo on Nubians of the Aswan region, in response to the massive relocations taking place in the wake of the construction of the Aswan Dam. However, for long the categories of 'Nubian' and 'Arab' remained largely unexamined in these contexts. Recent work, however, has begun to examine Nubian identity and its expressions in various settings, rural and urban. This opens up a brand new avenue of research in the region, namely, on 'race.'

This includes social historical works such as by Eve Troutt Powell (2003), pointing the way to nuanced cultural analyses of the intersections of race, ethnicity, and nation in historical and contemporary contexts.

Similarly, work on the cultural politics of 'Berber' (or Amazigh) identity focuses attention particularly on linguistic and symbolic aspects of Berber–Arab interactions. In the way that the literature on Egypt/Sudan leads us to race, the North African literature alerts us to the topics of migration and diaspora (Silverstein and Goodman, 2009). In addition to North African migrations and the politics of identity in Europe, the ethnographic gaze also shifts to the Arabian Peninsula and the Indian Ocean. Migration and displacement continue to frame the lives of a majority of the population of the region, directly and indirectly. Refugee's study is now a growing field, given that the region has produced as well as hosted some of the largest refugee populations in the world.

The most exciting literature that looks at history, diaspora, and identity should inspire generations of students. We can add to this the study of groups and communities originating from the Middle East and living in the West. In the emerging literature from the Middle East on diaspora we see two strands of research: one that connects to the ethnographies that draw on the politics of memory in contested places, the second belonging to studies connecting history and anthropology. In both approaches, writings illuminate everyday life and tell us more about how history, memory, and events impact people's everyday life. These works are examples of how personal, familial, or local memories may explicitly or silently challenge official versions of national history.

Drawing inspiration from works on memory, anthropologists working on contested sites and particular social or cultural groups are exploring the 'politics of memory' and the significance of memorializing practices to the politics of the nation and the state. Such explorations have been fruitful in understanding the role of commemorative practice in producing or reinforcing political ideologies of ethnic or religious group solidarity. Anthropologists have shown the myriad linkages between communal rituals and state policies in their historical unfolding. Connections are made between works of memory, colonial history, and political dislocation, demonstrating the centrality of political autonomy to the possibility of performing commemorative acts in postcolonial contexts or under repressive state regimes.

Conclusion

The study of sociocultural processes in the MENA region has progressed immensely in the past two decades and is much richer than previously, both in terms of the phenomena examined and the analytical frameworks deployed. This article has organized the discussion of the region through a focus on three major areas of research (nation, gender, and religion/Islam) that illustrates both continuities with past scholarship as well as the robustness of new and critical frameworks of analysis. The argument throughout, however, is that there is insufficient attention to issues of identity and difference in the region, which leads to slippages and easy generalizations (Shami and Naguib, 2013). A keen historical sensibility is needed to capture the ways in which these identities unfold and transform over time and through different contexts. In addition, the politics of state and nation building (which both involve transnational processes) must be the context through which the production of ethnic categories is understood. And finally, the transnational networks and exchanges within which identity formation is linked through migration and the formation of diasporas must also be taken into account. In the steadily growing literature on migration, one has possibilities to explore transformations and transnationalizations that aspire to legitimacy and recognition by both the state and supranational or international institutions.

This intellectual project takes on renewed urgency with the dynamics unleashed by the uprisings in the Arab region starting in late 2010. New forms of knowledge about identity and difference in this region are central to the new social imaginaries that are emerging and being contested in city squares and streets everyday. At the same time, the somber reality of sectarian and ethnic violence is also present, showing up the poverty of previous discourses of national identity and their inability to be truly inclusive. The drama being played out opens up possibilities to take up old debates and start new conversations about relationships between men and women, social classes, minorities and majorities, generations, and between citizens and the state.

See also: Ethnicity, Sociology of; Modernization; Near Middle East/North Africa Studies: Culture; Near Middle East/North Africa Studies: Gender; Near Middle East/North Africa Studies: Politics; Near Middle East/North Africa Studies: Religion; Near Middle East/North Africa Studies: Society and History; Orientalism.

Bibliography

Abu El Haj, N., 2001. Facts on the Ground: Archaeological Practice and Territorial Self Fashioning in Israeli Society. University of Chicago Press, Chicago.

Abu-Lughod, L., 1989. Zones of theory in the anthropology of the Arab world. Annual Review of Anthropology 18, 267–306.

Alcalay, A., 1993. After Jews and Arabs: Remaking Levantine Culture. University of Minnesota Press, Minneapolis, MN.

Anderson, Benedict, 1983. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London.

Asad, T., 1986. The Idea of an Anthropology of Islam. Occasional Paper Series. Center for Contemporary Arab Studies. Georgetown University Press, Washington, DC.

Asad, T., 1993. Genealogies of Religion: Discipline and Reasons of Power in Christianity and Islam. Johns Hopkins University Press, Baltimore.

Deeb, L., 2006. An Enchanted Modern: Gender and Public Piety in Shi'i Lebanon. Princeton University Press, Princeton.

Fabos, A., 2008. "Brothers" Or Others? Propriety and Gender for Muslim Arab Sudanese in Egypt. Berghahn Books, New York.

Gilsenan, M., 1982. Recognizing Islam: An Anthropologist's Introduction. Croom Helm, London.

Hale, S., 1996. Gender Politics in Sudan: Islamism, Socialism, and the State. Westview Press, Boulder, Colo.

Hoodfar, H., 2000. Iranian women at the intersection of citizenship and the family code: the perils of Islamic criteria. In: Joseph, Suad (Ed.), Women and Citizenship in the Middle East. Syracuse University Press, Syracuse, N.Y, pp. 281–313.

Hoodfar, H., 2001. Muslim women mullahs as volunteer reproductive health workers. In: Obermeyer, C. Makhlouf (Ed.), Cultural Perspectives on Reproductive Health. Oxford University Press, New York.

Relocating Arab nationalism, May, 2011. International Journal of Middle East Studies 43 (2), 203–204.
Joseph, S., 2000. Gender and Citizenship in the Middle East. Syracuse University Press, Syracuse, N.Y.
Joseph, S., Slyomovics, S., 2001. Women and Power in the Middle East. University of Pennsylvania Press, Philadelphia.
Kandiyoti, D., 1996. Gendering the Middle East: Emerging Perspectives. I. B. Tauris, London.
Mahmood, S., 2001. Feminist theory, embodiment, and the docile agent: some reflections on the Egyptian Islamic revival. Cultural Anthropology 16 (2), 202–236.
Mahmood, S., 2005. Politics of Piety: The Islamic Revival and the Feminist Subject. Princeton University Press, Princeton.
Mir-Hosseini, Z., 1999. Islam and Gender: The Religious Debate in Contemporary Iran. Princeton University Press, Princeton.
Mitchell, T., 1988. Colonising Egypt. Cambridge University Press, Cambridge, UK.
Moors, A., Jureidini, R., Ozbay, F., Sabban, Rima, 2009. Migrant domestic workers: a new public presence in the Middle East? In: Shami, S. (Ed.), Publics, Politics, and Participation: Locating the Public Sphere in the Middle East and North Africa. SSRC Books, New York, pp. 151–175.
Naguib, N., 2008. Storytelling: Armenian family albums in the diaspora. Visual Anthropology 21 (3), 231–244.
Nelson, C., 1974. Public and private politics: women in the Middle Eastern world. American Ethnologist 1 (3), 551–563.
Rassam, A., 1977. Al-Taba'iyya: power, patronage and marginal groups in Northern Iraq. In: Gellner, Ernest, Waterbury, John (Eds.), Patrons and Clients in Mediterranean Societies. Duckworth, London, pp. 157–166.
Said, E.W., 1978. Orientalism. Pantheon Books, New York.
Shafik, Viola, 2007a. Arab Cinema: History and Cultural Identity. American University in Cairo Press., Cairo.
Shafik, Viola, 2007b. Popular Egyptian Cinema: Gender, Class, and Nation. American University in Cairo Press, Cairo.
Shami, S. (Ed.), 2009. Publics, Politics and Participation: Relocating the Public Sphere in the Middle East and North Africa. SSRC Books, New York.
Shami, S., Naguib, N., 2013. Occluding difference: ethnic identity and the shifting zones of theory on the Middle East and North Africa. In: Slomovics, S., Hafez, S. (Eds.), Anthropology of the Middle East and North Africa: State of the Art. Indiana University Press, Bloomington.
Silverstein, P., Goodman, J. (Eds.), 2009. Bourdieu in Algeria: Colonial Politics, Ethnographic Practices, Theoretical Developments. University of Nebraska Press, Lincoln.
Slyomovics, S., 1998. The Object of Memory: Arab and Jew Narrate the Palestinian Village. University of Pennsylvania Press, Philadelphia, PA.
Troutt Powell, E.M., 2003. A Different Shade of Colonialism: Egypt, Great Britain, and the Mastery of the Sudan. University of California Press, Berkeley.
Varisco, Daniel, 2005. Islam Obscured: The Rhetoric of Anthropological Representation. Palgrave Macmillan, New York.
Varzi, Roxanne, 2008. Iran's Pieta: motherhood, sacrifice, and film in the aftermath of the Iran–Iraq war. Feminist Review 88, 1–13.
Vinogradov, Amal, 1974. Ethnicity, cultural discontinuity, and power brokers in northern Iraq: the case of the Shabak. American Ethnologist 1 (1), 207–218.
White, J.B., 2002. Islamist Mobilization in Turkey: A Study in Vernacular Politics. University of Washington Press, Seattle.
Zubaida, Sami, 1995. Is there a muslim society? Ernest Gellner's sociology of islam. Economy and Society 24, 151–188.

Middle Eastern and North African Legal Traditions

Lena Salaymeh, University of California (UC) Berkeley School of Law, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Legal systems in the modern Middle East and North Africa are both diverse and hybrid. This article provides an introductory overview that summarizes the multilayered legal history of this region and discusses how imperialism and colonialism shaped legal reform movements, codification, and the secularization of law.

Introduction

The common phrase/name 'Middle East and North Africa' (MENA) combines political and geographical categories. The 'Middle East' is a problematic political (rather than geographical) category; in comparison, 'North Africa' is a useful geographic category. While the more geographically descriptive (and less geopolitically constructed) term 'Southwest Asia' (or 'Western Asia') is preferable, it is not in widespread use. The 'Middle East' is an ambiguous, shifting, and modern (specifically, beginning in the twentieth century) category, having historically been preceded by the term 'Near East' (a term disseminated through British colonialism). Indeed, "The Middle East exists because the West has possessed sufficient power to give the idea substance. In this regard the colonial past and imperial present are parts of the equation that make the Middle East real" (Gasper, 2012: p. 240). Consequently, this article will focus on the modern period and modern nation-states. States that are often included within the 'Middle East' include Armenia, Bahrain, Cyprus, Egypt, Iran, Iraq, Israel, Jordan, Kuwait, Lebanon, Oman, Palestine, Qatar, Saudi Arabia, Syria, Turkey, United Arab Emirates, and Yemen. States that are typically included within 'North Africa' include Algeria, Libya, Mauritania, Morocco, Sahrawi Arab Democratic Republic, Somalia, Sudan, and Tunisia. This article will not comprehensively discuss the legal systems of each of these states, but instead offer a broad overview of the region with some detailed attention to the legal systems of some representative states. The acronym MENA will be used throughout this article to refer to the Middle East and North Africa. It should be noted that the MENA region is ethnically, linguistically, and religiously diverse; it cannot be simplified as the Arab or Islamic world.

Pre-Islamic History

Historical evidence that survives from some of the oldest legal systems globally comes from the Near East. Several ancient legal systems operated in the Near East and parts of North Africa, including the laws of Uru Nammu (twenty-second century BCE), Sumerian laws (eighteenth/nineteenth century BCE), laws of Hammurabi (eighteenth century BCE), Middle Assyrian laws (eleventh/twelfth century BCE), and many other legal systems (Jackson, 2008: pp. 37–44). In North Africa, Berber and Roman legal traditions were practiced for centuries. These and other ancient sources indicate a long history of complex legal systems operating in the region.

Islamic History (*see* Islamic Law)

Islam spread throughout the MENA gradually, over several centuries, as the result of conquests, trade, and assimilation processes. The majority of the region's inhabitants did not become Muslim until the medieval era (some estimates are between the twelfth and fourteenth centuries, depending on local circumstances). Throughout premodern history, non-Muslims (particularly Christians and Jews) operated their own courts to adjudicate matters of personal status (such as marriage, divorce, and inheritance) for their constituents. Still, non-Muslims frequented state (i.e., Islamic) courts for a variety of legal issues (including personal status). Since its historical beginning, Islamic law has had two overlapping dimensions in the MENA region: the first is the application of Islamic law as a state legal system over all its subjects; the second is the application of Islamic law by private jurists or Muslim associations (unaffiliated or unregulated by a state) over Muslim subjects. The first dimension is discussed in another encyclopedia article, Islamic Law; the second dimension will be mentioned briefly throughout this article.

Imperialism and Legal Reform (Eighteenth to Nineteenth Century)

The imperial era in the MENA region began in approximately 1798, with the Napoleonic invasion of Egypt (Mallat, 2007: p. 132). The people of the MENA region radically altered their legal systems between the eighteenth and twentieth centuries as a result of foreign intervention and internal demands for reform; the two factors cannot be easily distinguished. Legal changes that began during the Ottoman Empire intensified after its disintegration, the establishment of colonial regimes, and the later founding of quasi-independent nation-states. These legal changes primarily resulted in the emulation of the legal systems of European states, particularly in terms of juristic training, court procedures, and jurisprudential methodologies.

Tanẓīmāt (reorganization) is the term used to describe a period of reform in the Ottoman Empire that is typically dated as beginning with the Hatti Şerif of Gülhane (Imperial Edict of the Rose Chamber, known as the Gülhane Rescript)

of 1839 and ending with the Reform Edicts of 1856, but can be extended to describe legal changes beyond that date. The Gülhane Rescript confirmed the civil and economic rights of all Ottoman subjects, while confirming Islamic law as the source of Ottoman law; it is perceived as symbolizing Ottoman pluralism because it erased legal status differences between Muslims and non-Muslims. In 1847, the Ottomans established courts with both European and Ottoman judges, applying a combination of European and Ottoman laws and procedures. Several codes were promulgated in this period that largely adopted European (especially French) codes: in 1850, a commercial code; in 1858, land and penal codes; in 1861, another commercial code; and in 1863, a maritime code. These Ottoman codes were based on French law and were implemented by state courts (*niẓāmiyyah*). In addition to the enactment of codes (primarily instigated by European demands), Ottoman courts adopted European procedural rules and created special courts to adjudicate cases for Europeans. After 1869, the Ottoman Empire followed a civil code known as the *Mecelle-i Ahkam-i Adliye* (hereinafter, the Ottoman civil code), which included both Islamic legal principles and a variety of Islamic substantive and procedural laws, but was structured like the French Civil code. These modifications in the Ottoman legal system created a demand for jurists trained in these new codes and European-influenced legal procedures, resulting in the undermining of traditionally trained jurists.

Imperialist intervention in the legal systems of the region is most evident in the capitulatory system that was imposed upon the Ottoman Empire. European governments had placed significant pressure on the Ottoman Empire to provide their consuls with jurisdiction over all legal issues involving Europeans – and their affiliates, who included non-Muslim subjects of the Ottoman Empire – doing business in the Ottoman Empire. By the end of the eighteenth century, European capitulatory privileges were significantly affecting Ottoman markets as commercial dealings were effectively governed by external, foreign powers. Foreign legal jurisdiction in Ottoman Turkey did not end until the 1923 Treaty of Lausanne ratified the independence of Turkey.

Colonialism, Codification, and Modernity (Twentieth to Twenty-First Century)

The post–World War I division of the Ottoman Empire by the British and French created the modern nation-states that are now identified as MENA states. The 1916 Sykes-Picot agreement divided most of the areas previously under Ottoman rule into British and French colonies. Colonial administrators composed handbooks or legal guides for the MENA region throughout the nineteenth century; these texts shaped how colonized subjects viewed law in general and Islamic law in particular. Each newly formed state in the MENA region pursued an individualized path to establishing a state-based legal system that simultaneously reflects both the state's particular circumstances and broader trends of Westernizing law to meet global financial and political pressures. After the dissolution of the Ottoman Empire, the independent state of Turkey began a process of 'secularizing' its legal system through the adoption of European codes.

ʿAbd al-Razzāq al-Sanhūrī (d. 1971) spearheaded the codification of Egyptian civil laws in 1948, by fusing French and Egyptian legal norms. Sanhūrī was the most influential legal reformer in the Middle East in the twentieth century. A more comprehensive Egyptian Civil Code was promulgated in 1949 and it encompassed all civil legal matters, except personal status; it was based on a variety of civil codes from other countries and Islamic legal traditions. Many Arab states adopted Sanhūrī's codification with little or no modification. Egypt's civil code was replicated in Syria in 1951, in Iraq in 1953, and in Jordan in 1976. In 1951, Sanhūrī himself became involved in drafting Iraq's Civil Code, which, according to some scholars, contains more Islamic norms than its Egyptian predecessor. The introduction of legal codes throughout the region was met with little popular resistance, probably because most states divided jurisdiction between Islamic and civil courts – with the former handling the kinds of domestic cases that affected the broadest population base. There was significant scholarly discussion and commentary on, and explanation of these codes by contemporaneous jurists. In North Africa, French-inspired civil codes were the norm, with Islamic law influencing (although not necessarily dictating) family law codes.

The independent postcolonial nation-states that emerged in the MENA region implemented legal systems primarily based on Western legal systems. However, since most (if not all) of the states in the region were authoritarian, one should be cautious in drawing conclusions about the operation of law in isolation from the political context. Indeed, some observers have noted that the authoritarianism and military conflicts prevalent in the region puts into question the efficacy of the rule of law.

Colonizing Property Law, Liberalizing the Economy

In the nineteenth and twentieth centuries, land became increasingly privatized in the region as part of an extensive process of integration in the global capitalist economy. In the early nineteenth century, a series of legal decrees were enacted to facilitate the transfer and privatization of property. The Ottoman Land Code (1858) identified five types of property: (1) privately owned land; (2) state-owned land under private administration; (3) charitable-endowed trusts; (4) public land under preservation; and (5) land beyond the boundaries of inhabited areas. Regulations for each property type were designed to facilitate transfer and privatization of property. Other local land laws modified the prevailing conditions of land ownership. In order to control property in Algeria, the French waged a campaign against lands held in trust that entailed attacking the Islamic legal tradition of endowments (*awqāf*) (Powers, 1989). Western scholarly attention to Islamic trusts and laws regulating private property in the modern MENA reflect a wide-ranging interest in the history of global capitalism; unfortunately, some of this scholarly interest legitimates a problematic assumption of the incompatibility between Islamic law and modern financial structures (or economic progress in general) (Salaymeh, 2014). In the twentieth century, imperialist pressures on law reform in the region are evident in a variety of commercial laws (foreign investment, banking, taxation, employment, intellectual

property, and taxation) designed to integrate the MENA into the 'global economy' (Lapidus and Salaymeh, 2014).

Colonizing Family Law, Liberalizing Society

The long history of Western intervention in the legal systems of present-day MENA nation-states makes family law a battleground for far-reaching political issues. Colonial officers throughout the MENA region directed much criticism to the organization of families. In response, some local legal professionals sought to emulate Western personal status laws, while others rejected Western models. Enacted in 1917, the Ottoman law of family rights included a broad array of legal opinions from majority, minority, and extinct Islamic legal schools and was the first 'modern' family law code, applied throughout the empire. Turkey replaced the code with secular reforms in the 1920s. Other former Ottoman provinces replaced it with national family law codes – such as in Jordan (1951) and Syria (1953). Personal status codes were enacted throughout the region in quick succession: Tunisia (1956), Morocco (1958), and Iraq (1959). While these modern laws primarily codified Islamic legal norms, they also imposed the state's bureaucratic apparatus by requiring court involvement or registration of all marriages and divorces. In Egypt, the 1897 code for Islamic courts made written registration of marriage, divorce, and inheritance incumbent; this was confirmed by a later law requiring official documents for all family law matters after 1911. These procedural rules had substantive outcomes, since the courts were thereby able to implement a variety of previously informal or unenforced restrictions – such as a minimum age for marriage.

Marriage and divorce laws are the principal sites of contestation between groups that are erroneously labeled as 'conservative' and 'reformist/liberal'; this label oversimplifies the complex stakes that animate why individuals or groups advocate for particular laws. What underlies conflicts about family law is the sociopolitical reality of neoimperialism: many people in the MENA region view Westernized legal 'reforms' as a type of foreign intervention and therefore view rejection of Western legal norms as a form of resistance (Salaymeh, 2014; Lapidus and Salaymeh, 2014).

Islamicizing Constitutions in the MENA Regions

There is a broad spectrum of constitutional recognition of Islamic law in the MENA. Turkey declared itself a secular state in 1937, after removing constitutional reference to Islam as the state religion in 1928. Syria's 1973 constitution does not identify a state religion. After the 1979 revolution, the Iranian Constitution made Islamic law the highest source of legal authority and made Iranian jurists an integral part of the government. Based on its own determination of Islamic suitability, the Iranian Council of Guardians (comprising 12 jurists) is responsible for interpreting the Iranian Constitution, has the ability to overrule any legislation, and can prevent candidates from participating in political elections. Pakistan declared Islamic law as more authoritative than its own constitution in 1991. To understand what these constitutional references (or lack of references) to Islamic law mean or how they operate, it is essential to examine actual case law. All states in the region incorporate a hybrid of secular and religious laws and institutions. Most states apply Islamic law in distinct ways and generalizations cannot be made about the role of Islamic law in a particular modern nation-state based only on the text of its constitution.

Most importantly, the overt religious/secular characteristics or structures of states do not necessarily shape the substance of a state's legal system. In contemporary political discourse, demands for the inclusion of Islamic law in the constitutions of nation-states with Muslim majority populations has become commonplace. This rhetoric must be understood within its complicated sociopolitical context: continued foreign intervention, undemocratic regimes, and limited social and economic mobility. Popular Islamist movements pose a serious threat to essentially secular, autocratic regimes. Moreover, the term Islamic law means different things to different people and is often a symbolic – rather than substantive – demand. Typically, demands for the enforcement of Islamic law are simply demands that particular orthodox or historical Islamic laws – particularly in the areas of family and criminal law – be applied. The integration of Islamic law in the legal systems of modern nation-states has a minimal impact on the daily operation of law because the preponderance of legal matters (in civil and commercial matters) is governed by 'secular' laws.

Brief Legal Histories and Summaries of Select States

The following descriptions of the legal systems in certain MENA states focus on national legal systems, without delineating an individual state's international treaties or ratification of international agreements. Most MENA states have adopted civil legal codes, with some inclusion of Islamic law – particularly in the areas of personal status. The majority of these states have extensive and continuously updated laws concerning commercial activities, civil and criminal procedure, banking, property, taxation, and environmental protection that are not delineated here. Commercial laws in the region are essentially consistent with global practices, which are dictated by the strongest economies (mostly Western states) and designed to meet the demands of foreign investors and workers. Generally, labor laws – especially for foreign workers – are weak and employees have limited legal protections from exploitation. It should be noted that many states in the region recently enacted antiterrorism laws (modeled after those implemented in Western states) that function not as a form of crime prevention, but as a restraint on dissidents with Islamist political leanings. These legal synopses are not intended to be exhaustive in their presentation of the breadth of these legal systems.

Algeria

France colonized Algeria in 1830, beginning more than 30 years of French intervention in Algerian law – epitomized by the 'Droit Musulman-Algérien,' which was a fusion of French and Islamic laws. (Similarly, Anglo-Muhammadan law was a combination of British and Islamic laws that British judges applied to Muslims until 1947.) France imposed a hierarchical court system and a draft code (Code Morand of 1916) that

Europeanized Islamic laws; the Code Morand was a 'modern' interpretation of Islamic legal principles that was very influential on Algerian law.

Algeria achieved independence in 1962 and adopted a constitution in 1963 that was amended in 2002 and in 2008 (Kebir, 1994: p. 397). The Criminal Code (1966) integrated much French law. The Civil Code (1975) combines French and Islamic laws (Kebir, 1994: p. 397). The Family Code (1984) combines Islamic and non-Islamic laws and was modified in 2005 (Kebir, 1994: p. 404). Algeria's judicial hierarchy includes a Supreme Court, courts of appeal, and courts of first instance (Kebir, 1994: p. 400). After a civil war from roughly 1990 to 2000, the Algerian government criminalized an array of activities related to political Islamism under the pretext of 'fighting terrorism' (Kebir, 2000–2001: p. 364).

Bahrain

Bahrain gained independence from Britain in 1971 and its legal system fuses civil law (from the Egyptian civil code) and common law (from English laws of contracts) (Al Baharna, 1994: p. 332). Bahrain's criminal procedure code (1966), penal code (1976), and various commercial codes constitute the bulk of the state's laws. Bahrain's 1973 constitution gives a Council of Ministers and the Amir (or King) the authority to issue legislation. A consultative council was established in 1992 to provide legislative support and offer advice to the government (Al Baharna, 1994: p. 333). The constitution was replaced in 2002 with a new constitution declaring that the state of Bahrain would be named the Kingdom of Bahrain and issuing a series of laws (Al Baharna, 2001–2002). Bahrain's judicial hierarchy includes the Supreme Court of Appeal (established in 1989), the High Court of Appeal, the High Court, and the lower courts (Al Baharna, 1994: p. 335). The Bahraini uprising that began in 2011 demands a representative form of government and protection of basic human rights.

Egypt

Modern European (particularly French) civil law codes were integrated into the legal system of the Egyptian state in the nineteenth century. New legal codes (civil, commercial, maritime, criminal, and procedural) promulgated in 1875 combined the *Code Napoléon* and Islamic legal sources. Mixed courts were established in 1875 to adjudicate matters between Egyptians and foreign nationals on the basis of these codes; they continued to operate until 1949. Similarly, based on European models, National Courts and national codes were established in 1883. Egypt's Civil Courts have a structured hierarchy of appeals: two categories of the Courts of First Instance, Courts of Appeal, and the Civil Chamber (Kosheri, 1994: p. 125). The Egyptian government began training judges for its new Islamic courts in 1907. Egypt's Islamic courts were integrated into the National Courts in 1956.

One of the better known modern legal reformers was ʿAbd al-Razzāq al-Sanhūrī (d. 1971), discussed above, who led the drafting of the Egyptian Civil Code (1948/9) and the Iraqi Code (1951) and participated in the drafting of codes for neighboring states (including Iraq, Syria, Jordan, Kuwait, and Libya). Sanhūrī earned two doctorates from the University of Lyon, was dean of both an Egyptian and an Iraqi law school, and held several government positions. Sanhūrī sought to modernize Islamic law and to develop a modern legal system that would benefit from comparative study of Western legal systems, but emphasize Islamic values.

Egypt's 1971 Constitution established a Supreme Constitutional Court (SCC) and an independent judiciary. (Amendments to the Egyptian constitution were enacted in 2007.) The SCC hears cases through referrals from lower courts (criminal, civil, or administrative) and its decisions are irrevocable. In response to popular demands, the Egyptian Constitution was amended in 1980 to identify Islamic law as 'the' (rather than 'a') primary source of Egyptian legislation. The 2011 Egyptian Constitutional Declaration identified the principles of Islamic law as the principle source of legislation. In the Egyptian legal system, substantive Islamic law is most influential in family law. For example, in 2000, Egypt enacted a law facilitating wife-initiated divorce that was challenged as an unconstitutional violation of Islamic law; the SCC upheld the law, noting that there are valid Islamic legal opinions validating such divorces (Riad and Gabr, 2002–2003: pp. 181–194). While non-Muslims (primarily Copts) are governed by non-Islamic laws of marriage and divorce, they are subject to Islamic inheritance rules. Egypt's family laws of 1920 and 1929 were updated in 1985 (Kosheri, 1994: pp. 139–140). Islamic law does not appear to have much influence on Egypt's Penal Code.

A number of business-friendly laws were passed in the 1970s and 1980s to encourage investment in line with global capitalist trends. Law No. 203 (in 1991) initiated a slow process of privatization, supported by laws that reorganized the banking system (Kosheri, 1994: pp. 130–131). Further liberalization of Egypt's economy was instituted through Law No. 38 (in 1994), which deals with foreign currency and exchange (Kosheri, 1994: p. 133). Egypt's 1952 revolution had resulted in redistribution of land; a series of laws were passed in the 1990s that reflected broader liberalization efforts on the part of the Egyptian government (Kosheri, 1994: p. 139). Egypt's 2011 revolution and 2013 military coup have created an uncertain future for the rule of law in the state.

Iran

Modern Iran was ruled by a monarchy until the 1979 revolution, which established the Islamic Republic. The establishment of an Islamist government did not immediately alter the legal landscape of the country; specifically, some legal codes predate the revolution, including the Civil Code (1928–35) and the Commercial Code (1932). The Iranian Constitution (1979; amended and ratified in 1989) gives jurists – whether the supreme leader or the Council of Guardians – prominent roles in the state's law-making, including judicial review. The Iranian Parliament drafts laws that are reviewed by the Council of Guardians (Ansari-Pour, 1994: p. 385). In 1989, Iran created a council to mediate between the Council of Guardians and Parliament, in an effort to resolve an impasse between jurists and legislators. The Council of Ministers also drafts laws that are sent to Parliament (Ansari-Pour, 1994: p. 386). While the Imāmī Shīʿī school is the official source of law, other Islamic and non-Islamic laws are recognized by the state. Iran's 1911

criminal procedure code was replaced in 1999 (Ansari-Pour, 1999–2000: p. 335). The civil procedure code (1939) was replaced in 2000 (Ansari-Pour, 2000–2001: p. 349). An Islamic punishment code was promulgated in 1991 and a 1994 act resulted in some changes to civil and criminal procedures (Ansari-Pour, 1994: p. 386).

Iraq

Iraq's 1925 Constitution confirmed its monarchy and established a legislature, but Iraq did not become independent of British rule until 1932. The 1958 revolution established a republic and the Iraqi Constitution was modified repeatedly (in 1943, 1958, 1963, 1965, and 1970) (Al-Mukhtar, 1994: p. 158). In addition to the executive, judicial, and legislative branches, the Constitution enumerates the duties of the Revolutionary Command Council, which had supreme authority in the country. Iraq's judiciary includes civil, criminal, family, and special courts; the Court of Cassation is the highest Iraqi court (Al-Mukhtar, 1994: p. 160). Iraq's Civil Code was promulgated in 1951, under the direction of Egyptian jurist ʿAbd al-Razzāq al-Sanhūrī, and reportedly contained more Islamic legal principles than its Egyptian counterpart (Al-Mukhtar, 1994: p. 163). Civil procedure, evidence, commercial law, banking law, and employment law correspond to European civil legal systems or other Arab legal systems. Family law was governed by the 1959 personal status law and is applied to all Iraqi Muslims; Iraqi Christians, Jews, and other confessional groups have their own personal status laws that are applied in the civil courts.

Iraq issued a series of new laws and modified existing ones in the period prior to its occupation of Kuwait (Al-Mukhtar, 1994: p. 171). As a result of the United States–led invasion of Iraq in 1991 and its occupation in 2003 (which ostensibly ended in 2011), the country's legal system was effectively destroyed. Iraq's current constitution (2005) was drafted by a council (some of whose members were not Iraqi) appointed by the United States and other forces occupying Iraq at the time (the 'Coalition Provisional Authority'). The constitution has been criticized by many Iraqis for emphasizing sectarian and ethnic identity over national unity.

Jordan

Jordan achieved independence from Britain in 1946, but English law is primarily influential in arbitration law. French law influenced Jordan's civil and commercial laws. Like many other states in the region, family law is based on Islamic legal traditions. Jordan's Constitution dates to 1952, but was amended in 1984. Civil and criminal matters are handled by regular courts (First Instance, Courts of Appeal, and Court of Cassation); religious courts (Islamic and Christian) handle personal status matters; there are also military and administrative courts (Haddad, 1994: p. 178). Jordan's 1953 land law permitted the transformation of state-owned land to private ownership in specific situations. A 2006 antiterrorism law expanded the jurisdiction of the Jordanian state security court (Haddad, 2005–2006: p. 290). In 2010–11, a number of legal reforms – including the establishment of a constitutional court – were enacted to heed off a potential uprising (Haddad, 2010–2011: p. 209).

Kuwait

Kuwait was under British control between 1914 and 1961, but its legal system is based on Ottoman and Islamic laws, rather than British common law. In 1962, Kuwait enacted a constitution that identifies Islam as the state religion. Kuwait's civil code is based on the Egyptian civil code (since it was drafted by the Egyptian jurist, Sanhūrī). Kuwaiti judges sometimes defer to Egyptian precedents (Kassim, 1994: p. 254). Kuwait promulgated several codes in 1960 (criminal, criminal procedure, and commercial companies), 1964 (commercial agencies and labor), and 1980 (commercial, civil and commercial procedures, and maritime commercial) (Kassim, 1994: p. 255). The judiciary includes the Court of First Instance, the Court of Appeals, and the Court of Cassation. Kuwait's civil code (1980) was intentionally designed to incorporate legal opinions from multiple Islamic legal schools. The only tax in Kuwait is applied to corporate income (Kassim, 1994: p. 271).

Lebanon

Lebanon was under French colonial rule from 1920 (when the Ottoman Empire dissolved) to 1943. Its legal system is primarily based on the French civil law system, beginning with the Code of Obligations and Contracts (1932). Civil, commercial, and criminal courts are overseen by the Court of Cassation, whereas the Conseil d'Etat oversees administrative law. Nearly 20 distinct confessional systems (Islamic and Christian sects, as well as Judaism) govern family law; however, inheritance is a matter of civil law, not subject to religious courts. Ten legal professionals sit on Lebanon's Constitutional Council, which was founded in 1989 (Mallat, 1994: pp. 204–205). The rule of law disintegrated during the Lebanese civil war (1975–90); in its aftermath, Lebanon promulgated several laws relating to property, taxation, and estates to deal with gaps and non-functionality in the existing laws (Mallat, 1994: pp. 214–217). Reflecting the implications of a prolonged and chaotic civil war, Lebanon passed a law in 1995 that declares as dead any individual who had gone missing for more than 4 years; this law was implemented to facilitate inheritance cases (Mallat, 1994: p. 217). The complexity of asserting the rule of law in Lebanon is evident in the uneven application of criminal proceedings against individuals who committed actionable crimes during the war; many individuals directly involved in warfare became part of the state's administration after the war (Mallat, 1994: p. 219).

Libya

Libya's civil legal system is based on the French and Egyptian civil codes. The first article of Libya's civil code identifies the state's legal sources as legislation, Islamic legal principles, custom, and natural law or equity (El-Alem, 1994: p. 225). Libya's judicial hierarchy includes a Supreme Court, appellate courts, and first instance courts (El-Alem, 1994: p. 226). The Libyan Constitution gives Basic Popular Conferences legislative authority (El-Alem, 1994: p. 227). Civil and commercial procedures are governed by a 1954 code; in 1994, popular tribunals, a new type of court, were introduced

(El-Alem, 1994: pp. 230–231). A 1992 property law gives every Libyan citizen the right to own property (El-Alem, 1994: p. 232). Various modifications related to polygyny laws occurred during the 1990s (El-Alem, 1994: p. 233). Libyan law (1974 and 1994) prohibits consumption, possession, and production of alcohol by Muslims, but non-Muslims have the right to consume alcohol in private (El-Alem, 1994: p. 235). In 1994, Libya passed an Arabic language act that requires all public materials to be in Arabic or include an Arabic translation of foreign languages (El-Alem, 1994: p. 236).

After the Libyan uprising of 2011, foreign intervention, and the removal of Colonel Qadhdhāfī (Gaddafi) and his government, a General National Council was formed in 2012 with the intention of writing a constitution for Libya. Libya's legal system is unstable and in transition.

Morocco

The Kingdom of Morocco gained independence from colonial rule in 1956. Morocco's Constitution (of 1996, preceded by revisions and constitutions in 1992, 1972, 1970, and 1962) identifies Islam as the official religion. The civil procedure code (1959), the penal code (1962), and other Moroccan codes are primarily influenced by French codes. The state's judicial hierarchy includes a Supreme Court, courts of appeal, and courts of first instance (Zirari-Devif, 1994: pp. 411–412). In 2002, laws against terrorism were enacted that significantly extended police powers (Zirari-Devif, 2002–2003: pp. 357–358). King Muhammad VI of Morocco requested a new personal status code (*muduwwanah*) that was adopted by parliament in 2003; the new code limited polygyny, gave previously unmarried women the authority to contract marriages without a guardian, and restricted men's abilities to divorce (Zirari-Devif, 2003–2004: pp. 285–288). The King of Morocco initiated a series of legal reforms in 2011 in order to prevent a popular uprising (Zirari-Devif, 2010–2011: p. 289).

Oman

The modern state of Oman was recognized as sovereign in 1951 and has been ruled by Sultan Qābūs as an absolute monarchy since 1971. Royal decrees are the primary source of legislation; Oman's commercial law (1990) is based on Kuwaiti, Egyptian, and French laws (Edmondes, 1994: p. 350). The dominant Islamic school of law is Ibāḍī. Oman established special security courts to deal with political threats against the state or Sultan in 2003 (Rae and Howard, 1994: p. 295).

Qatar

Qatar gained independence from Britain in 1971 and is an absolute monarchy. Qatar's 2003 constitution established an elected parliament and identified Islamic law as a source of legislation (Al Tamimi, 2005–2006: p. 375). As in many other MENA states, Egyptian laws influence Qatari law. The civil and Islamic courts, once separate, were recently united. The Council of Ministers drafts laws for consideration by the Consultative Council and the Cabinet (Al-Nauimi, 1994: p. 345).

Saudi Arabia

Saudi Arabia is an absolute monarchy, which means that the monarch has ultimate authority. Saudi rules of civil procedure were issued in 1936 and 1952; the Ministry of Justice was established in 1970 and it issues Islamic legal decrees on a variety of issues (Abrahams, 1994: pp. 275–276). In 1992, Saudi Arabia passed its Basic Law, which is modeled after a constitution, although the first article identifies the Qur'ān as the state's constitution (Mallat, 2007: p. 160). An 'Islamic' procedural law was enacted in 2000 and was the first systematization of legal procedures in the kingdom (Hejailan, 2000–2001: p. 271). A Supreme Court and a Supreme Judicial Council were both established in 2007 (Dahlan and Zedan, 2008–2009: p. 339). While in theory Islamic law is the main source of law, regional and international laws have been integrated into the Saudi legal system. Saudi courts do not publish court decisions; prior to the 1980s, however, the decisions of courts of equity were reported (Mallat, 1998: p. 118). Commercial regulations, often influenced by Egyptian or Western practices, are promulgated by the Council of Ministers (Abrahams, 1994: p. 274). Islamic courts hear cases involving family, criminal, property, and personal injury matters (Abrahams, 1994: p. 277). The Board of Grievances (established in 1955) deals with claims against the government (Abrahams, 1994: p. 278). Saudi business law has developed to deal with the large numbers of foreign investors and foreign professionals working in the country. While only citizens may be qualified to practice law, foreign attorneys often function as legal consultants to members of the Saudi bar.

Syria

Syria became independent from French colonial rule in 1946, at which time the state issued several legal codes that were modeled after Lebanese or Egyptian codes. Syria's 1973 Constitution gives the People's Assembly and the President shared legislative power (El-Hakim, 1994: pp. 142–143). Many commercial laws were enacted in the mid-twentieth century to encourage foreign investment, tourism, and banking operations. Syria's labor laws require employer-provided social insurance. Personal status matters are handled by Islamic and other confessional courts. Civil and criminal courts are subject to appeal. Syria has three lower criminal courts (police, first degree, and court of criminal appeal). The Superior Judicial Council appoints judges and oversees lower courts. The Higher Constitutional Court deals with the constitutionality of laws (El-Hakim, 1994: p. 153). In 2011, the Syrian government issued a new constitution making broad legal changes; it was 'approved' in 2012; the country is in the midst of a civil war (El-Hakim and Harding, 2010–2011: p. 179).

Sudan

Sudan's legal system combines English common law, Islamic law, and customary law. Britain established a system of dual courts in the country, with 'secular' courts handling civil and criminal matters and Islamic courts handling the family law matters of Muslims (Safwat, 1994: p. 237). Islamic law was

integrated into the operation of all courts through the Judicial Decision Sources Act (1983), Civil Procedure Act (1983), Evidence Act (1983), Civil Transactions Act (1984), and Penal Code (1991). Most of these laws also identified the importance of custom to Sudanese law (Safwat, 1994: p. 238). Sudan's judiciary includes a Supreme Court, courts of appeal, provincial courts, district courts, town benches, and other courts (Safwat, 1994: p. 239). A personal status code was issued in 1991 for Muslims, integrating legal opinions from the Ḥanafī and Mālikī schools (Safwat, 1994: p. 250). Some parts of the Sudanese penal code that were based on Islamic law were not applied in Southern Sudan (which became independent in 2011) (Safwat, 1994: p. 253). The 1990 National Security Act gives the government broad authority to detain citizens without formal charges. Sudan's 1998 constitution identified the state as Islamic, created a constitutional court, and made numerous other legal changes (Makec, 1997–1998: p. 308). A new constitution was adopted in 2005.

Tunisia

Tunisia achieved independence from France in 1956 and adopted a constitution in 1959 (Gaigi, 1994: p. 419). Initially applicable to Muslims only, the personal status code (1956) was extended to all Tunisian citizens in 1957 (Gaigi, 1994: p. 420). The Tunisian legal system includes several codes, including civil and commercial procedure (1959), commercial (1959), and criminal procedure (1968, revised 1993) (Gaigi, 1994: p. 420). Tunisia's personal status code (1956), which has been revised repeatedly, is based on Islamic law with many modifications; it has been the subject of extensive analysis and commentary (Gaigi, 1994: p. 423). The Tunisian uprising began in 2010 and the Tunisian Constituent Assembly recently issued a new constitution.

Turkey

Turkey became a republic in 1923; in 1926, the state adopted the Swiss Civil Code (replacing the Ottoman civil code), when it was approved by the Grand National Assembly, after years of lobbying by Atatürk (Mallat, 1998: p. 114). Turkey's penal code (1926), civil procedure code (1927), and criminal procedure code (1929) were all modeled after European codes (Dogru, 1995: p. 285). Turkey's current constitution (1982, amendments through 2008) replaced several earlier constitutions and was ratified under a military government that came to power by a *coup d'état* in 1980; the Constitutional Court reviews the constitutionality of laws passed by parliament after military rule ended in 1983 (Dogru, 1995: p. 287). Turkey's constitution does not identify Islamic law as a source of law.

United Arab Emirates

The United Arab Emirates (UAE) was established as a federation of seven emirates in 1971–72 and has a Supreme Council that is responsible for approving laws and international agreements; essentially, it has a form of monarchical governance. While its constitution identifies Islamic law as the principal source of law, Egyptian codes, English common law, and local customs are influential (Price, 1994: p. 303). The Council of Ministers drafts laws that are then sent, in succession, to the National Assembly and the Supreme Council for approval (Price, 1994: p. 305). Civil (1987), penal (1987), and commercial (1993) codes demarcate much of the daily operation of law in the UAE. The UAE's civil code incorporates a few Islamic legal principles that were cited in the Ottoman civil code (Price, 1994: p. 305). The judicial hierarchy includes the Court of Cassation, the Court of Appeal, the Court of First Instance, as well as criminal and Islamic courts (Price, 1994: p. 304). In terms of the application of Islamic law, UAE courts give more legal weight to the opinions of the Mālikī and Ḥanbalī schools (Price, 1994: p. 303). Although only citizens may be qualified to practice law, foreign attorneys often function as legal consultants to members of the UAE bar. In 1994, the UAE promulgated an extensive Commercial Code; as a member of the Arab Gulf Corporation Council (founded in 1981), the UAE enacts economic laws that are in line with the regional goals of Arab Gulf States (Price, 1994: p. 305).

Yemen

The modern republic of Yemen was created with the unification of North and South Yemen in 1990. Yemen's 1994 Constitution identifies Islamic law as the primary source of law; Yemen's civil code elaborates that Islamic law, customary practices, and principles of equity are the main sources of law (Shamiri, 1994: p. 369). Laws are promulgated by the parliament, executive decrees, and resolutions of various ministers, all of which are published in Yemen's official legal journal (Shamiri, 1994: p. 370). Yemen's judicial hierarchy includes a Supreme Court, a court of appeal, and a court of first instance (Shamiri, 1994: pp. 371–372). The Yemeni Civil Code (1992) codifies many Islamic legal principles (Shamiri, 1994: p. 378). Personal status matters and criminal punishments are governed by Islamic law. In 1997, Yemen initiated a judicial reform program and established a government institution for issuing Islamic legal decrees (*dār al-iftā'*) (Shamiri, 1997–1998: pp. 390–395). Yemen established a national security organization to deal with political threats against the state in 2003 (Shamiri, 2002–2003: pp. 315–317). The Yemeni uprising began in 2011 and it is unclear what legal changes will emerge as a result of this ongoing political transformation.

Conclusion

The modern history of the MENA is characterized by recurring shifts between direct colonial occupation and indirect imperialist interventions. The complexity of this ongoing imperial situation shapes the rule of law and provokes the contours of legal debates that superficially (and incorrectly) pit tradition against modernity, religion against secularism, Islamic law against national legal systems; what underlies these debates is not a struggle between these false bipolarities, but a battle for autonomy, cultural dignity, and popular representation. Internal disputes about law in the MENA region reflect broader

controversies over the distribution of power, economic resources, and social capital. As the region continues to be embroiled in uprisings (Tunisia, Egypt, Syria, Yemen, and Bahrain), military coups (most recently in Egypt in 2013), foreign interventions, and general instability, law functions inconsistently and unpredictably.

Disclaimer

For the benefit of the general reader, references are to source in English.

See also: Constitutionalism, Comparative; Islam: Middle East; Islamic Law; Law: Change and Evolution; Legal Pluralism; Legal Systems, Classification of; Postcolonial Law; Religion: Nationalism and Identity; Secularization; Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns.

Bibliography

Abrahams, A., 1994. Country survey: Saudi Arabia. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Al Baharna, H.M., 1994. Country survey: Bahrain. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Al Baharna, H.M., 2001–2002. Country survey: Bahrain. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 8. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Al-Mukhtar, S., 1994. Country survey: Iraq. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Al-Nauimi, N., 1994. Country survey: Qatar. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Al Tamimi, 2005–2006. Country survey: Qatar. In: Cotran, E., Lau, M., Kattan, V. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 12. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Ansari-Pour, M.A., 1994. Country survey: Iran. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Ansari-Pour, M.A., 1999–2000. Country survey: Iran. In: Cotran, E. (Ed.), Yearbook of Islamic and Middle Eastern Law, vol. 6. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Ansari-Pour, M.A., 2000–2001. Country survey: Iran. In: Cotran, E. (Ed.), Yearbook of Islamic and Middle Eastern Law, vol. 7. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Dahlan, M.R., Zedan, H., 2008–2009. Country survey: Saudi Arabia. In: Cotran, E., Lau, M. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 14. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Dogru, O., 1995. Country survey: Turkey. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 2. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Edmondes, N., 1994. Country survey: Oman. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

El-Alem, M., 1994. Country survey: Libya. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

El-Hakim, J., 1994. Country survey: Syria. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

El-Hakim, J., Harding, T., 2010–2011. Country survey: Syria. In: Cotran, E., Lau, M., Sherif, A.O. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 16. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Gaigi, A., 1994. Country survey: Tunisia. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Gasper, M.E., 2012. Conclusion: there is a Middle East! In: Bonine, M.E., Amanat, A., Gasper, M.E. (Eds.), Is There a Middle East?: The Evolution of a Geopolitical Concept. Stanford University Press, Stanford, California.

Haddad, H., 1994. Country survey: Jordan. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Haddad, H., 2005–2006. Country survey: Jordan. In: Cotran, E., Lau, M., Kattan, V. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 12. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Haddad, H., 2010–2011. Country survey: Jordan. In: Cotran, E., Lau, M., Sherif, A.O. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 16. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Hejailan, H.S., 2000–2001. Country survey: Saudi Arabia. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 7. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Jackson, S.A., 2008. A Comparison of Ancient Near Eastern Law Collections Prior to the First Millennium BC [Online]. Gorgias Press, Piscataway, NJ. Available at: http://site.ebrary.com/id/10440201.

Kassim, A.F., 1994. Country survey: Kuwait. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Kebir, Y., 1994. Country survey: Algeria. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Kebir, Y., 2000–2001. Country survey: Algeria. In: Cotran, E. (Ed.), Yearbook of Islamic and Middle Eastern Law, vol. 7. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Kosheri, R.R., 1994. Country survey: Egypt. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Lapidus, I.M., Salaymeh, L., 2014. Law in the Middle East: 19th–21st Centuries. A history of Islamic Societies. third revised ed. Cambridge University Press, Cambridge.

Makec, J.W., 1997–1998. Country survey: Sudan. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 4. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Mallat, C., 1994. Country survey: Lebanon. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Mallat, C., 1998. Islamic law research in the twentieth-century Middle East. Asian Research Trends, 109–138.

Mallat, C., 2007. Introduction to Middle Eastern Law. Oxford University Press, Oxford, New York.

Powers, D.S., 1989. Orientalism, colonialism, and legal history: the attack on Muslim family endowments in Algeria and India. Comparative Studies in Society and History 31, 535–571.

Price, R., 1994. Country survey: United Arab Emirates (UAE). In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Rae, A., Howard, B., 1994. Country survey: Oman. In: Cotran, E., Lau, M. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 9. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Riad, T., Gabr, H., 2002–2003. Country survey: Egypt. In: Cotran, E., Lau, M. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 9. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Safwat, S., 1994. Country survey: Sudan. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Salaymeh, L., 2014. Commodifying 'Islamic law' in the U.S. legal academy. Journal of Legal Education 63 (4), 640–646.

Shamiri, N., 1994. Country survey: Yemen. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Shamiri, N., 1997–1998. Country survey: Yemen. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 4. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Shamiri, N., 2002–2003. Country survey: Yemen. In: Cotran, E., Lau, M. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 9. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Zirari-Devif, M., 1994. Country survey: Morocco. In: Cotran, E., Mallat, C. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 1. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Zirari-Devif, M., 2002–2003. Country survey: Morocco. In: Cotran, E., Lau, M. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 9. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Zirari-Devif, M., 2003–2004. Country survey: Morocco. In: Cotran, E., Lau, M. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 10. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Zirari-Devif, M., 2010–2011. Country survey: Morocco. In: Cotran, E., Lau, M., Sherif, A.O. (Eds.), Yearbook of Islamic and Middle Eastern Law, vol. 16. Kluwer Law International; Centre of Islamic and Middle Eastern Law at the School of Oriental and African Studies, University of London, London.

Relevant Websites

http://portal.clinecenter.illinois.edu/cgi-bin//rview?REPOSID=1 – Constitutions Repository.

http://www.law.emory.edu/ifl/legal/ – Emory Law, Legal Profiles.

Midlife Psychological Development

Mathias Allemand, University of Zurich, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M.E. Lachman, volume 14, pp. 9796–9799, © 2001, Elsevier Ltd.

Abstract

Middle adulthood is an important developmental period in the life span. The goal of this article is to give an overview about theoretical and empirical work on midlife development. The first part emphasizes the importance of studying midlife development. The second part presents conceptual frameworks and central issues of midlife. The third part summarizes research findings for environmental changes in the primary social contexts and social roles of midlife including family, work, and community. The fourth part summarizes research findings for individual changes in multiple domains such as cognitive and emotional abilities, personality and the self. The final part includes future directions.

Middle Adulthood

Middle adulthood is a developmental period covering the years from 35 to 65, with the core years between 40 and 60. The middle years have been neglected because developmental science and aging research often contrast younger with older adults and thus omit middle-aged adults (Lachman and Bertrand, 2001) (*see* Age Stratification). One reason for this neglect is that, traditionally, this period has been considered as a time of relative stability in many realms of life. Although this might hold true to some extent, middle adulthood is a challenging and complex period with manifold environmental and individual changes (Willis and Martin, 2005). Hence, understanding development in middle adulthood is important for several reasons. First, middle age is a unique period of the life span that is characterized by a complex interplay of multiple social roles. Middle-aged adults largely hold multiple roles with widest responsibilities, and the number and nature of these roles may change across midlife (Helson and Soto, 2005). Second, understanding middle adulthood is important because this period is characterized by a wide variability in its nature and course, as middle-aged adults show different developmental pathways. As such, it requires a focus on multiple trajectories of development and their interplay across different areas of life including family and work (Lachman, 2004). Third, development in earlier periods of the life span preceding midlife may influence psychological functioning in middle adulthood. As an example, research demonstrates that low self-esteem and decreases in self-esteem in adolescence predict individual differences in depression in the mid-thirties (Steiger et al., 2014). Therefore, identifying individual characteristics, sociocultural factors, and developmental antecedents of psychological functioning in middle adulthood is an important way to understand midlife development. Fourth, development in middle adulthood may have consequences for aging processes and coping in later years of the life span and therefore reflects an important developmental window into old age (*see* Retirement and Encore Adulthood: The New Later Life Course). For example, in middle adulthood it is possible to detect the first indications of developmental decline (Schaie, 2005). Knowledge of adaptive and maladaptive aging processes and midlife precursors of psychological functioning in old age is important to develop interventions to promote healthy aging and to prevent maladaptive processes.

Theoretical Perspectives

Several theoretical perspectives exist that specifically include middle adulthood as a distinct period of the life span (*see* Life Span Development, Theory of). For example, Erikson (1963) believes that a primary developmental challenge in middle adulthood is generativity, that is, one of contributing to society and helping to guide future generations (cf McAdams, 2001). According to Erikson (1963), development is the result of challenges and crises that typically occur in each life period. For example, young adults are exploring personal relationships and find intimacy in a long-term committed relationship, while failure may result in loneliness and isolation. In contrast, older adults tend to look back on their lives and feel satisfaction about their accomplishments or they experience many regrets about their life.

Middle-aged adults seek to make a positive contribution to the next generation through parenting, teaching, mentoring, leadership, creative activities, and contributions to leave a legacy for the next generation. Middle-aged adults give back to society through raising children, being productive at work, or becoming involved in community activities and organizations, such as caregiving to parents (Shifren, 2009). As such, they serve important roles in the family and relationships, at the workplace, and in the community and, more broadly, the society. They are often guides for the younger and older generations, with an ability to draw on their experiences in multiple domains (McAdams, 2001). Hence, when middle-aged adults sense that they have done nothing for the next generation, they develop a feeling of stagnation because they are dissatisfied with the relative lack of unproductivity (Erikson, 1963).

Other perspectives have been guided by theories of life span developmental psychology that views development as a lifelong process covering all periods of the life span including the middle years of adulthood (Baltes et al., 2006) (*see* Lifespan

Development, Theory of). These perspectives view midlife in the context of the entire life span rather than as a disconnected period (Staudinger and Bluck, 2001). They place emphasis on the potential plasticity of the midlife experience as a function of individual and environmental or contextual variables and compensatory behavioral changes to biological aging. Complex dynamic interactions between individuals and their environment are responsible for both gains and losses throughout middle adulthood (Baltes, 1987).

Maintenance and Adaptation Functions of Midlife

Middle adulthood is a relatively long time period that includes at least two transitions in which individuals are both leaving young adulthood and entering old age. Some researchers have suggested that distinguishing early, middle, and late periods of middle age would reflect the very different experiences, social roles, and health of adults in their 30s versus those in their 60s. One suggestion for such a categorization is based on the notion that resource investments and goal orientation across the life span differ with respect to three adaptive functions (Staudinger and Bluck, 2001). From childhood to young adulthood, *growth* is the primary function in the allocation of resources, whereas in old age the *management and prevention of losses* is predominate, as resources are increasingly needed to regulate losses such as by adjusting to chronic illnesses (cf Baltes, 1987).

Middle adulthood is primarily characterized by *maintenance* functions such as balancing family and career and the tendency to strive for stability (Staudinger and Bluck, 2001). Midlife also includes adaptation processes to the manifold and diverse life experiences of the middle years. However, according to these authors the investment of resources also differ across the middle-aged years, as considerable resources are still needed to invest in growth in early middle adulthood, while some resources are already needed to manage certain losses in late middle adulthood. As people move through the middle years, they encounter increasing, peaking, and decreasing opportunities to give back to society by striving for certain developmental goals and tasks, such as raising children or becoming established in a career and being productive at work (Havighurst, 1972).

A similar suggestion is based on the idea that important individual and environmental factors show peaks or valleys or even changes quality during middle adulthood (Helson et al., 2006). For example, social commitments and striving for status increase from young adulthood to early midlife and then decrease by late midlife, whereas the middle period is characterized by maximum status and responsibility in a socially complex environment (Helson and Soto, 2005).

Variability in Midlife Development

Although some theoretical perspectives on development during middle adulthood appear to describe normative development (e.g., Erikson, 1963; Havighurst, 1972), it is important to note that midlife development cannot be conceived as universal in nature. Middle adulthood is characterized by a wide variability in the nature and course of development (Lachman, 2004; Willis and Martin, 2005). That is, wide individual differences exist in how middle-aged adults' progress through the middle years (*see* Motivation: Life Course and Sociological Perspectives). For example, some people experience specific positive life events, such as becoming parents or getting a promotion at the workplace, whereas others are subject to negative life events, such as getting divorced or losing the job. Other people, in turn, report multiple events, whereas again others do not exhibit any major change in life circumstances in middle adulthood. As a result of diverse individual and environmental experiences in midlife each person may demonstrate a unique pattern of development.

Changes in the Environment

Middle adulthood is characterized by several changes in the environment. Changes may occur with respect to family, work and community contexts, social roles, transitions, life events, and circumstances.

Family, Relationships, Work, and Community

The primary social contexts or realms of experiences in middle adulthood are the family and the broader social network including diverse social relationships, the workplace, and the community (*see* Family as Institution; Wellbeing and Burnout in the Workplace, Psychology of). As noted earlier, middle-aged adults are motivated to give back to society through raising children, being productive at work, or becoming involved in community activities and organizations (Erikson, 1963). As such, social relationships are important in middle adulthood for several reasons. People are largely interested in positive interactions with other people, giving and receiving affection, and experiencing caring for others. However, the structure, nature, and functions of social networks may change across the middle years. For example, according to the social convoy model of social relations, people are accompanied from a dynamic social network of individuals such as the family members from whom they give and receive social support (Antonucci et al., 2011). The convoys are marked by stability and change, as some members of the convoy go, some stay, and new members come. In addition, the functions of social relationships may change in the middle years. For example, research demonstrates that the investments in meaningful social relationships increase with age, as people become more aware of the finiteness of opportunities (Charles and Carstensen, 2010).

The workplace is one of the primary settings of middle adulthood, and people's identity is largely defined by their work. In the middle years, most people attain their highest level of socioeconomic status along with widest responsibilities at work and/or in the family and community (Helson and Soto, 2005). The work context does not reflect a consistent and stable environment but rather a changing world. For example, technological advances, changes in social networking pattern inside and outside the workplace, and job characteristics are subject to change, which entails transitions in employment in coordination with changing family responsibilities, and later the experience of retiring. These work-related experiences might

also influence other domains of life such as family, leisure time, and finances (Lachman, 2004).

Occupational experiences may influence the individual and promote or prevent individual changes. At the same time, it is possible that individual characteristics may have an influence on work experiences as well. For example, research demonstrates that individual characteristics such as personality shape occupational experiences over time across middle adulthood, but occupational experiences have minimal impact on personality (Sutin and Costa, 2010). In summary, the family, relationships, the workplace, and community reflect key developmental contexts for the individual.

Multiple Social Roles and Responsibilities

Middle adulthood is a time when people are responsible members of the society and hold multiple roles with widest responsibilities in socially complex environments (Lachman, 2004). The primary social roles in middle adulthood are centered in family and relationships, work, and community. Family roles may be related to the family of origin (e.g., parents, grandparents, and relatives) and/or the family of destination (e.g., partner, children), whereas work and community roles relate to the career and the broader social environment, such as the community (*see* Family as Institution; Job Analysis and Work Roles, Psychology of). Furthermore, social roles tend to fall into two broad domains: Belongingness roles that contain friendships, family, and community roles, and status roles that contain work and social positions roles (cf Roberts and Wood, 2006). Different motivations underlie these two broad types of social roles. The desires to have family, close friendships, and to be integrated in a social group or social groups underlie belongingness roles, whereas the desires for social status, fame, money, and social regard are associated with status roles. Each social role includes a set of norms, internal and external social expectations, and expected behaviors that may have an influence on people. According to Roberts and Wood (2006), roles can serve as consistent subjective environments that people act within across time and situations. When people adapt a specific role, their experiences and behaviors may change, in particularly when they are highly committed and show a high investment to the social role (Lodi-Smith and Roberts, 2007).

Across middle adulthood, people have to coordinate multiple social roles (e.g., spouse parent, worker) (*see* Job Analysis and Work Roles, Psychology of). Sometimes middle-aged adults have to face different roles at the same time in different social situations, such as taking on a career, establishing a family, or taking care of parents, which might be a source of potential conflicts (Shifren, 2009). Moreover, many of the social roles are age graded. Social roles such as being a parent systematically change across the adult years (Helson and Soto, 2005). The number and the nature of these roles may change during middle adulthood. More specifically, the number of social roles increases from young adulthood to early midlife then decreases as children leave the parental home and adults approach retirement (Helson and Soto, 2005). In summary, it is during middle adulthood when the number of social roles peaks and then decreases and the role may change in character.

Transitions, Life Circumstances and Events

As people move through middle adulthood, they are faced with several transitions and changes in life circumstances. A typical transition in early middle adulthood is the transition to parenthood, which involves manifold physical, psychological, and social changes in the lives of the parents (Nazarinia et al., 2014) (*see* Transition to Adulthood; Fertility Theory). For example, the birth of a child is associated with changes in daily routines, activities, and sleep patterns. The transition to being a parent also reflects a change into a new social role that may trigger new experiences and behaviors (Roberts and Wood, 2006). Moreover, cultural specific timetables for events to occur (e.g., social clock) may influence the timing of transitions, such as having children. In later periods of midlife, parents may experience their adult children leaving home. The empty nest situation provides new role for the relationship with the partner.

Late-middle adulthood is also marked by the transition to retirement (Schultz and Wang, 2011) (*see* Retirement and Encore Adulthood: The New Later Life Course). This transition involves dramatic changes associated with work and social contacts and requires adaptation processes to the post-retirement life. Generally, transitions reflect challenges for the individual and the social environment including emotional difficulties. Yet, at the same time, they provide opportunities for change and growth to overcome routine and habit. Transitions may entail a reconsideration of values, priorities, and goals. For example, retirement may allow the retired individuals to pursue new interests and enjoy the freedom from daylong working.

During middle adulthood, some people experience profound changes in their life circumstances. Many life events may occur in midlife including acute illness, death of loved one, caring for one's parents, changes in relationships, loss of job, moving, changes in responsibilities at work, and financial difficulties. As an example, divorce is a life experience that is associated with a wide range of changes and might have several consequences for the divorcees. Indeed, studies reported a variety of economic, social, physical, and mental health consequences, although the full extent of such effects remains debated (Demo and Fine, 2010). Divorce is also associated with major changes and transitions with respect to role status. Consequently, changes in social roles might promote individual changes (Roberts and Wood, 2006). Specifically, new social roles such as being a single parent come with a new set of expectations and demands including parental responsibilities.

Unemployment and job loss in midlife are two other examples of life circumstances that may challenges people in middle adulthood. Possible consequences are the loss of income, independence, skills use, social contacts, and may challenge intimate relationships (*see* Unemployment and Mental Health). Because people's identity is largely defined by their work, a loss of job or unemployment is damage to social identity and status.

Another example relates to the fact that middle-aged people (the 'sandwich generation') sometimes have to care for ill parents and at the same time for demanding and challenging children or adolescents (Shifren, 2009). This double strain may be a source of chronic stress in midlife that requires an adaptive balance between parent care and employment and family roles.

People differ in how they interpret and understand stressful life events. For example, some people construe life events as turning points that reflect significant psychological changes in their lives. Another way to understand a stressful life event is to take a lesson away from experiences. How people interpret stressful life events may have an influence on themselves. For example, work shows that across middle adulthood, personality change might be more strongly related to how people understand the stressful events in their lives rather than simply the occurrence of such events (Sutin et al., 2010). For example, highly neurotic people might interpret an event as stressful whereas an emotionally stable person might see the same event as challenging but manageable. To summarize, diverse midlife experiences in terms of life circumstances and events contribute to wide variability in the nature and course of midlife development.

Changes in the Individual

In addition to environmental challenges, middle adulthood is characterized by individual changes. Changes may occur with respect to health and physical functioning as well as other domains of psychological functioning.

Health and Physical Changes

Middle adulthood is a time of relatively good self-reported health without dramatic changes, as compared to old age (Whitbourne, 2001). However, during midlife some people begin to show increases in health problems as well as gradual physical declines, particularly in late midlife. Examples for physiological changes that may occur refer to changes in the elasticity of the skin, changes in eyesight and in other sensory functions (e.g., hearing), back and joint pains, weight gain with related problems (e.g., cardiovascular disease and diabetes) (Whitbourne, 2001). In general, most of these health and physical changes do not cause disability or alter lifestyles of middle-aged adults. However, an accumulation of risk factors or risky health behaviors in middle adulthood may show negative effects later in life. For example, smoking, alcohol consumption, and overweight are risk factors for health in old age, whereas healthy lifestyles in midlife are preventive factors.

One major biosocial change in the middle years occurs in the area of reproduction, especially the menopause for women (*see* Fertility Theory; Fertility Theory: Theory of Life History Evolution). The menopause or climacteric occurs at some point between the ages of 46–58 years. It involves changes in hormonal levels and fertility and has effects on mood, and it is often associated with hot flashes, which generally last for a few months. However, women differ in their experiences of the menopause. The menopause often goes along with social changes that are typical of late-middle adulthood, such as children leaving home. For women, the menopause reflects an important biosocial transition, whereas men have no clear biosocial transition in middle adulthood.

Cognitive Abilities

Cognitive abilities are key competences that are needed to meet the challenges of job demands, education, and advanced training, societal expectations, and the demands of everyday life of middle-aged adults (Lachman, 2004) (*see* Modern Human Cognition, Evolution of). Middle adulthood provides many good opportunities for making intellectual contributions, given the position of middle-aged adults in the family, workplace, and society. Hence, middle age is the period of highest performance in a variety of individual domains. Middle-aged adults typically perceive themselves or are perceived by others as having not only more status and responsibility but also more effective intelligence and integrative skills (Lachman et al., 1994). The demands of status and widest responsibilities may encourage the development of these intellectual skills. For example, diverse work environments and workplace conditions may contribute to job-specific and individual differences in the development of cognitive abilities. As such, the development of cognitive abilities is more strongly influenced by environmental factors compared to other developmental periods (Sternberg et al., 2001).

As a consequence of diverse life contexts, changes in cognitive functioning in middle adulthood are varied and multidirectional (Schaie, 2005). Indeed, longitudinal research demonstrates that some aspects of cognitive functioning that depend on experience are maintained or even improved in midlife, while other aspects decrease (Willis and Schaie, 2005). For example, perceptual speed already begins to decline before the middle years. Number ability declines over middle adulthood and afterward. But the years from age 40 to the early 60s also reflect a period of maximum performance on some of the higher order abilities such as inductive reasoning and spatial orientation. These abilities then begin to show a decline in late-middle adulthood, though with many individual differences. As such, it is possible to detect the first indications of developmental decline in middle adulthood (Schaie, 2005).

Several risk factors or protective factors may contribute to the variability in cognitive development (Willis and Schaie, 2005). For example, multiple biosocial risk factors such as hypertension, diabetes, health status, cardiovascular disease, or stressful life circumstances may be associated to midlife cognitive decline. By contrast, several protective factors are associated with cognitive maintenance and plasticity. For example, high levels of education, complex environmental contexts and demanding activities at work, intellectual engagement with a wide range of interests, and high levels of exercise and physical activity are discussed as possible mechanisms than can be targeted in preventive interventions (cf Willis and Schaie, 2005).

Emotional Abilities

Emotions and emotional abilities are important to meet the challenges of social interactions and life circumstances, as they may have intrapersonal and interpersonal functions (Izard, 2009) (*see* Infancy and Childhood: Emotional Development). For example, emotions motivate behavior and decision-making, help in understanding the motives of other people, facilitate attention-focusing and help to prioritize between different goals. Emotions are transient and show variability over short periods of time. They may differ as a result of being exposed to different situations. By contrast, emotional abilities reflect relatively enduring patterns of emotional functioning

that are relatively stable over time. Examples for emotional abilities are emotional intelligence, the tendency to feel positive and negative emotions and satisfaction with life, emotional complexity and the capacity to express emotional experiences.

As with other forms of midlife development, changes in emotional abilities are varied and multidirectional (Charles and Carstensen, 2010). Research with respect to positive and negative emotions suggest that the tendency to feel positive affect and enjoyment and to be satisfied with life are relatively stable in the middle years but then show a decline in late periods of midlife and old age. In contrast, the tendency to feel negative emotions appears to decrease steadily from early adulthood to older adulthood (Charles et al., 2001; Mroczek and Spiro, 2005). This positive emotional developmental pattern reflects the tendency of becoming more prone to positive emotions and less prone to negative emotions over time in order to optimize well-being. An additional aspect of positive development refers to the tendency of becoming able to handle and process more complex emotional information and being self-aware of emotional complexity. This aspect includes the ability to tolerate tension and negativity in order to maintain emotional differentiation and complexity. The dynamic integration theory states people are challenged to balance increases in emotional optimization and emotional complexity across the life span (Labouvie-Vief et al., 2010). The integration of both aspects reflects a dynamic balance. Evidence suggests that from early to middle adulthood, the capacity for integration increases and later decreases due to cognitive limitations or poor emotion regulation strategies (Labouvie-Vief et al., 2010).

Emotion regulation strategies are very important in middle adulthood, as this period is a challenging and complex period with manifold environmental and individual changes. Emotion regulation strategies may serve as protective factors and help to adapt to life circumstances and stressful events, including multiple roles and responsibilities, developmental deadlines (e.g., biological clock), and other challenges associated with midlife. Indeed, research has shown that the regulation of emotion is related to adaptive functioning among middle-aged adults (Charles and Carstensen, 2010).

Motivational Changes

Motivations may also change across the middle years. The realization of long-term developmental goals such as having children depends on the opportunities that are shaped by the life span (Heckhausen et al., 2010) (*see* Motivation and Actions, Psychology of). The opportunities to realize developmental goals and tasks are age-graded and thus have developmental deadlines, as a result of biological (e.g., loss of fertility) and social constraints (e.g., social expectations about when certain activities and goals should be undertaken). For example, a middle-aged woman after the menopause is not able to get pregnant due to the loss of fertility.

To develop in an adaptive way, people have to balance between primary control processes that are directed at changing the world to bring the environment into line with one's needs and desires and secondary control processes that are directed at changing the self to bring oneself into line with environmental forces (Heckhausen et al., 2010). Primary control helps to pursuit goals, whereas secondary control supports goal flexibility and helps to cope with failure and challenges. These control processes show distinct patterns in middle adulthood. Secondary control processes tend to increase from early periods of the life span to the middle years and beyond. The capacity for primary control still increases in the early period of midlife, reaches the maximum in the middle period and finally starts to decline in the late period. However, according to Heckhausen et al. (2010) people constantly try to maximize overall primary control across the entire life span.

Motivational changes may also result from people's general tendency to perceive their time in life as either open-ended or as limited. The socioemotional selectivity theory suggests that the subjective perception of time has important implications for the selection of goals, activities, and preferences (Carstensen et al., 1999). When future time is perceived as open-ended, as is typical in young adulthood, people are strongly motivated to optimize their future by pursuing information and knowledge-related social goals. When people view their future as limited, as is typical in later life or when people are confronted with serious illnesses, emotional goals, meaningful social activities, and experiences become increasingly important (Carstensen et al., 1999; Charles and Carstensen, 2010). In the middle years, people begin to realize that maybe more time has gone by than remains. The realization of decreases of possibilities to pursue long-term goals in the future may be associated with giving up some developmental goals and dreams in midlife.

Personality Development

Human personality or the self may be viewed at least from three different perspectives including narrative, motivational, and dispositional perspectives (McAdams and Olson, 2010). The narrative perspective corresponds to internalized life stories. Life stories are the narrative understanding of the self to make meaning out of life. In addition, the motivational perspective refers to personal goals, motives, plans, values, hopes, and fears that are contextualized in time, place, and social role. Eventually, the dispositional perspective corresponds to personality traits that are relatively enduring patterns of behavior, thought, and feeling that are fairly consistent across a wide variety of situations and contexts. Individual differences in dispositional patterns are typically organized along the Big Five framework (John et al., 2008) and can be assessed from different perspectives (e.g., self-, other-, partner-, and metaperceptions of personality; Schaffhuser et al., 2014). Neuroticism, or low-emotional stability, contrasts even-temperedness with experience of anxiety, worry, anger, and depression. Extraversion refers to individual differences in the propensity to be sociable, active, assertive, and to experience positive affect. Openness to experience refers to individual differences in the proneness to be original, complex, creative, and open to new ideas. Agreeableness refers to traits that reflect individual differences in the propensity to be altruistic, trusting, modest, and warm. Finally, conscientiousness reflects the propensity to be self-controlled, task- and goal-directed, planful, and rule following. These individual differences are important because they predict life outcomes that are essential

in midlife, such as marital success, occupational attainment, and health (Roberts et al., 2007).

The most consistent evidence regarding changes in personality traits across adulthood is that people become more dominant, agreeable, conscientious, and emotionally stable across the adult years, or at least up through late-middle age (Allemand and Lehmann, 2012; Roberts et al., 2006). This pattern of change across middle adulthood is particularly important with respect to the primary social contexts of middle-aged adults (i.e., family, relationships, work, and community). For example, individual differences in the propensity to be self-controlled, task- and goal-directed, planful, and rule following are particularly important to be productive at work and thus to contribute to society. Similarly, individual differences in the tendency to forgive others are relevant in social settings including interpersonal conflicts and transgressions (Steiner et al., 2012).

The pattern of personality change reflects a trend toward greater maturity across adulthood in the sense of becoming emotionally more predictable and more attuned to social demands with age (Roberts and Wood, 2006). This pattern describes a socially effective personality that contributes to everyday life running smoothly and maintaining or increasing subjective well-being and life success. As dispositional traits show trends toward greater maturity in midlife, goals and narratives show an increasing concern with commitments to family, social relationships, civic involvement, and productive activities aimed at promoting the next generation (McAdams and Olson, 2010). Motivational and narrative aspects of the self tend to be more malleable than personality traits and are influenced more strongly by experiences in middle adulthood, making the person resilient or vulnerable to stress (McAdams and Olson, 2010).

Similar to changes in the cognitive, emotional and motivational functioning, personality change is characterized by many individual differences across the adult years (Roberts and Mroczek, 2008). This suggests variability in the degree and direction of personality change and may reflect the plurality and diversity of life experiences people can encounter in middle adulthood. Specific life circumstances and events may lead to different personality trajectories for different people (and vice versa: different personality patterns might lead to patterns of how to encounter, respond to and interpret situations), which can help to explain the substantial variability in midlife development.

Conclusions and Future Directions

Development in middle adulthood is an important issue. Four take home messages stand out. First, middle adulthood is a unique and relatively long developmental period with the primary functions of maintenance and adaptation to diverse midlife contexts and experiences. More theoretical and empirical work is needed to understand the nature and course of development from early to late periods of midlife. Second, middle adulthood is characterized by a wide variability in its nature and course. Therefore, more research is needed to understand why some people remain the same as they pass through the middle years of adulthood, while other people change, and also why people differ in the direction and degree of change. Third, middle adulthood is characterized by manifold changes in the environment and in the individual. More theoretical and empirical work is needed to understand the associations between changes in the environment and individual. Under certain conditions, environmental changes may promote or inhibit changes in cognitive, emotional, and motivational functioning and personality in midlife. Conversely, individual characteristics and experiences may shape the environment. Environmental and individual factors and their interactions may be responsible for change and stability through the middle years by means of dynamic maintenance and adaptation processes to life circumstances, stressful life events, developmental goals and tasks, and other influences. Similarly, research is needed to understand changes between diverse domains of individual functioning, such as cognition and personality. Fourth, as middle adulthood is a 'sandwich' period between other developmental periods with different experiences, it is important to examine the antecedents and consequences of midlife development. Research is needed to identify factors in earlier life periods that predict individual differences in midlife psychological functioning. In addition, midlife development may influence aging processes in old age. As such, research is needed to identify early warning signs of maladaptive processes for prevention efforts. Likewise, applied research is needed to develop interventions for middle-aged adults that promote healthy aging. To conclude, recent theoretical advances and empirical work suggest that the future is bright for the scientific study of development in middle adulthood including antecedents and consequences of midlife development.

See also: Age Stratification; Family as Institution; Fertility Theory: Theory of Life History Evolution; Fertility Theory; Human Cognition, Evolution of; Human Development, Theories of; Infancy and Childhood: Emotional Development; Job Analysis and Work Roles, Psychology of; Life Course: Sociological Aspects; Lifespan Development, Theory of; Motivation and Actions, Psychology of; Retirement and Encore Adulthood: The New Later Life Course; Transition to Adulthood; Unemployment and Mental Health; Wellbeing and Burnout in the Workplace, Psychology of.

Bibliography

Allemand, M., Lehmann, R., 2012. Personality development and aging. In: Ramachandran, V.S. (Ed.), Encyclopedia of Human Behavior. Elsevier, San Diego, CA, pp. 74–81.

Antonucci, T.C., Birditt, K.S., Ajrouch, K., 2011. Convoys of social relations: past, present, and future. In: Fingerman, K.L., Berg, C.A., Smith, J., Antonucci, T.C. (Eds.), Handbook of Lifespan Development. Springer, New York, pp. 161–182.

Baltes, P.B., 1987. Theoretical propositions of lifespan developmental psychology: on the dynamics between growth and decline. Developmental Psychology 23, 611–626.

Baltes, P.B., Lindenberger, U., Staudinger, U.M., 2006. Lifespan theory in developmental psychology. In: Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology, Theoretical Models of Human Development, vol. 1. Wiley, New York, pp. 569–664.

Carstensen, L.L., Isaacowitz, D.M., Charles, S.T., 1999. Taking time seriously: a theory of socioemotional selectivity. American Psychologist 54, 165–181.

Charles, S., Carstensen, L.L., 2010. Social and emotional aging. Annual Review of Psychology 61, 383–409.

Charles, S.T., Reynolds, C.A., Gatz, M., 2001. Age-related differences and change in positive and negative affect over 23 years. Journal of Personality and Social Psychology 80, 136–151.

Demo, D.H., Fine, M.A., 2010. Beyond the Average Divorce. Sage, Thousand Oaks.

Erikson, E.H., 1963. Childhood and Society, second ed. Norton, New York.

Havighurst, R.J., 1972. Developmental Tasks and Education. McKay Company, New York.

Heckhausen, J., Wrosch, C., Schulz, R., 2010. A motivational theory of lifespan development. Psychological Review 117, 32–60.

Helson, R., Soto, C.J., 2005. Up and down in middle age: monotonic and non-monotonic changes in roles, status, and personality. Journal of Personality and Social Psychology 89, 194–204.

Helson, R., Soto, C.J., Cate, R.A., 2006. From young adulthood through the middle ages. In: Mroczek, D.K., Little, T.D. (Eds.), Handbook of Personality Development. Lawrence Erlbaum, Mahwah, pp. 337–352.

Izard, C.E., 2009. Emotion theory and research: highlights, unanswered questions, and emerging issues. Annual Review of Psychology 60, 1–25.

John, O.P., Naumann, L.P., Soto, C.J., 2008. Paradigm shift to the integrative Big Five trait taxonomy: history, measurement, and conceptual issues. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research, third ed. Guilford, New York, pp. 114–158.

Labouvie-Vief, G., Grühn, D., Studer, J., 2010. Dynamic integration of emotion and cognition: equilibrium regulation in development and aging. In: Lerner, R.M., Lamb, M.E., Freund, A.M. (Eds.), The Handbook of Lifespan Development, Social and Emotional Development, vol. 2. Wiley, Hoboken, pp. 79–115.

Lachman, M.E., 2004. Development in midlife. Annual Review of Psychology 55, 305–331.

Lachman, M.E., Bertrand, R.M., 2001. Personality and the self in midlife. In: Lachman, M.E. (Ed.), Handbook of Midlife Development. John Wiley & Sons, New York, pp. 279–309.

Lachman, M.E., Lewkowicz, C., Marcus, A., Peng, Y., 1994. Images of midlife development among, young, middle-aged, and older adults. Journal of Adult Development 1, 201–211.

Lodi-Smith, J.L., Roberts, B.W., 2007. Social investment and personality: a meta-analysis of the relationship of personality traits to investment in work, family, religion, and volunteerism. Personality and Social Psychology Review 11, 68–86.

McAdams, D.P., 2001. Generativity. In: Lachman, M.E. (Ed.), Handbook of Midlife Development. John Wiley & Sons, New York, pp. 395–443.

McAdams, D.P., Olson, B., 2010. Personality development: continuity and change over the life course. Annual Review of Psychology 61, 517–542.

Mroczek, D.K., Spiro III, A., 2005. Change in life satisfaction over 20 years during adulthood: findings from the VA normative aging study. Journal of Personality and Social Psychology 88, 189–202.

Nazarinia, R.R., Schumm, W.R., Britt, S.L., 2014. Transition to Parenthood. Springer Science & Business Media, New York.

Roberts, B.W., Kuncel, N., Shiner, R.N., Caspi, A., Goldberg, L., 2007. The power of personality: a comparative analysis of the predictive validity of personality traits, SES, and IQ. Perspectives on Psychological Science 2, 313–345.

Roberts, B.W., Mroczek, D.K., 2008. Personality trait change in adulthood. Current Directions in Psychological Science 17, 31–35.

Roberts, B.W., Walton, K., Viechtbauer, W., 2006. Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. Psychological Bulletin 132, 1–25.

Roberts, B.W., Wood, D., 2006. Personality development in the context of the neo-socioanalytic model of personality. In: Mroczek, D.K., Little, T.D. (Eds.), Handbook of Personality Development. Lawrence Erlbaum, Mahwah, pp. 11–39.

Schaffhuser, K., Allemand, M., Martin, M., 2014. Personality traits and relationship satisfaction in intimate couples: three perspectives on personality. European Journal of Personality 28, 120–133.

Schaie, W.K., 2005. Developmental Influences on Adult Intelligence: The Seattle Longitudinal Study. University Press, Oxford.

Schultz, K.S., Wang, M., 2011. Psychological perspectives on the changing nature of retirement. American Psychologist 66, 170–179.

Shifren, K., 2009. How Caregiving Affect Development: Psychological Implications for Child, Adolescent, and Adult Caregivers. APA, Washington, DC.

Staudinger, U.M., Bluck, S., 2001. A view on midlife development from life-span theory. In: Lachman, M.E. (Ed.), Handbook of Midlife Development. John Wiley & Sons, New York, pp. 3–39.

Steiger, A.E., Allemand, M., Robins, R.W., Fend, H.A., 2014. Low and decreasing self-esteem during adolescence predict adult depression two decades later. Journal of Personality and Social Psychology 106, 325–338.

Steiner, M., Allemand, M., McCullough, M.E., 2012. Do agreeableness and neuroticism explain age differences in the tendency to forgive others? Personality and Social Psychology Bulletin 38, 441–453.

Sternberg, R.J., Grigorenko, E.L., Oh, S., 2001. The development of intelligence at midlife. In: Lachman, M.E. (Ed.), Handbook of Midlife Development. John Wiley & Sons, New York, pp. 217–247.

Sutin, A.R., Costa, P.T., 2010. Reciprocal influences of personality and job characteristics across middle adulthood. Journal of Personality 78, 257–288.

Sutin, A.R., Costa Jr., P.T., Wethington, E., Eaton, W.W., 2010. Turning points and lessons learned: stressful life events and personality trait development across middle adulthood. Psychology and Aging 25, 524–533.

Willis, S.L., Martin, M., 2005. Middle Adulthood: A Lifespan Perspective. Sage, Thousand Oaks.

Willis, S.L., Schaie, K.W., 2005. Cognitive trajectories in midlife and cognitive functioning in old age. In: Willis, S.L., Martin, M. (Eds.), Middle Adulthood: A Lifespan Perspective. Sage, Thousand Oaks, pp. 243–275.

Whitbourne, S.K., 2001. The physical aging process in midlife: interactions with psychological and sociocultural factors. In: Lachman, M.E. (Ed.), Handbook of Midlife Development. John Wiley & Sons, New York, pp. 109–155.

Migrants and Migration in Europe

David Coleman, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since the 1950s, annual immigration to Europe has increased to almost two million persons. Migration dominates demographic change in many countries, making them ethnically diverse. It has important economic and social consequences; immigrants comprise about 10% of the workforce. The European Union has acquired powers over immigration policy, but policies of different countries vary greatly, many are restrictive; popular opposition to large-scale immigration is strong. The economic effects of immigration are controversial and conflicting, beneficial for immigrants, employers and some consumers, not for the domestic poor. Fiscal benefits are small, but there is strong competition for high-skill immigrants.

Background

Until the mid-twentieth century, Europe had been a continent of emigration. Between 1815 and 1914, about 54 million Europeans are estimated to have emigrated (Baines, 1991). But the twentieth century brought to western Europe migratory movements unprecedented for a millennium. First from the forced migration and ethnic displacements occasioned by the two world wars and their political consequences, the echoes of which continue today. Then, guest-worker programs from the 1950s and less formal labor inflows, brought migrants to Europe – to notably Germany – from poorer European countries and from outside Europe, from Turkey and from former colonial territories. In France, the UK and other countries, the citizens of colonies and former colonies also enjoyed for a while unrestricted rights of entry that had been acquired during imperial times. Many of the migrants expected to return home. Most did not, even when economic growth faltered. Instead, later family reunion and family formation made settlement permanent and growing through chain migration. Further inflow was attracted by postwar economic growth and facilitated by three peaceful postwar revolutions; in transport, communication and in human rights. Asylum seeking greatly increased after the 1980s. Until 1989, movement across the Iron Curtain into and out of eastern Europe was minimal except initially from East Germany. After 1961, the East German fortified border, including the Berlin Wall, reduced to a trickle the hemorrhage of population to West Germany.

Initially, the free movement of labor between countries of the European Union (EU) made relatively modest contributions to migration flows. All that changed after 1989 when the populations of the much poorer countries of eastern Europe were free to leave after 1989, especially after they were admitted to the EU in 2004 (8 countries: Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Slovakia, and Slovenia (the 'A8')) and subsequently (3 countries: Bulgaria, Croatia, and Romania) (Eurostat, 2008). Cyprus and Malta (the 'A2') also joined in 2004. Labor migration remains important, but most immigration to Europe in recent decades has not been (ostensibly) for purposes of work; the majority being spouses, dependents, students, asylum seekers, and illegal immigrants (Garson and Salt, 2011). Through all the processes above, especially chain migration, western Europe has become a region of mass immigration, almost irrespective of the aims of national immigration policies. As a consequence, substantial non-European populations have come to Europe for the first time, introducing new ethnic and religious diversity and raising important problems of integration.

Recent Migration Trends in Europe

Before discussing numbers, it is important to keep in mind that migration data are problematic and inadequate, although improving (e.g., US Bureau of the Census, 2005; Raymer and Abel, 2008). Harmonization even of migration data between EU countries has not been easy (Poulain et al., 2006). Estimates of net annual inflow in many European countries are not based on counts in or out, but rather on comparisons of registrations from a year to the next. The UK bases estimates of annual inflow and outflow from a small sample of voluntary interviews with international travelers. Persons recorded as permanent immigrants in a year have often entered in earlier years. There may be about half a million illegal immigrants to Europe every year. Data on the 'stock' of residents, according to citizenship and birthplace, have greatly improved, however (Lemaitre, 2005).

Immigration to the EU countries (In addition to the countries listed earlier, EU member states comprise Austria, Belgium, Denmark, Finland, France, Germany, Greece, Italy, Irish Republic, Luxemburg, The Netherlands, Portugal, Spain, Sweden, The United Kingdom (the 'EU15')) has generally risen in recent decades. Overall, the net inflow has often been higher than the gross inflow to the United States (**Figure 1**), although those figures are based upon different criteria. In 2011, 1.7 million people were estimated to have migrated to the EU from a country outside the EU27, and 1.3 million people already living in the EU27 migrated to another member state. Altogether, over 3 million people immigrated to one of the EU27 member states, while 2.3 million left (both figures include movement between EU states; Eurostat, 2013a). Most immigrants from outside the EU come from countries ranked as of 'Medium Human Development Index' (HDI) status, e.g., Morocco, China, Ghana, India, Ukraine, while 6% came from low HDI countries, e.g., Iraq, Pakistan, Nigeria, and most African countries. 35% came from high HDI (non-EU)

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31122-9

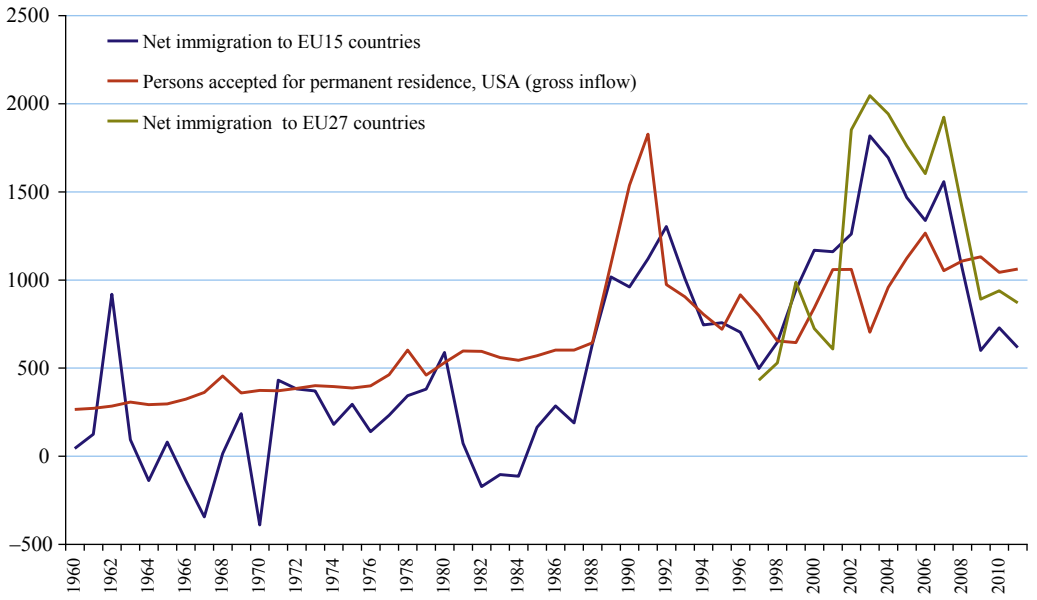

Figure 1 Net immigration to EU15 and EU27 countries, gross inflow to the United States (1960–2011), and net inflow to United States (2001–11) (thousands). Sources: Eurostat, US Department of Home Land Security.

countries (e.g., US, Japan) and 4 and 3%, respectively, from European free-trade area (EFTA) (Iceland, Liechtenstein, Norway, and Switzerland) and EU candidate (Iceland, the former Yugoslav Republic of Macedonia, Montenegro, Serbia, and Turkey) countries. Most immigrants to EU countries still come from outside Europe, although the opening of eastern Europe has changed the balance since the 1990s. There was a very large increase, recently moderated, of migration from eastern Europe accession countries to the EU, especially to those few countries that did not initially impose restrictions on labor movement (UK, Sweden, Ireland).

From 2008, most states declined migration to EU in response to the economic crisis. Partly because of that, the recent European immigration scene is one of relatively incoherent and contrary trends in different countries. We are only slowly and with uncertainty recovering from that crisis. The problems of Eurozone countries have depressed immigration, while emigration from them has increased – mostly to Germany. That turnaround has reversed Germany's population decline. In 2012, net immigration of 369 000 to Germany more than compensated for the excess of deaths over births (−196 000), increasing the population by 173 000 or 0.2%, the biggest increase since 1995.

Restrictions on A8 labor movement were lifted in 2011 (although many had already entered irregularly). Restrictions on Romania and Bulgaria, more recent EU members, must be lifted in 2014. The consequences are awaited with interest, although many have already entered illegally. The more prosperous Central and European countries (e.g., Czech Republic) have modest net inflows of population. But most have lost population by migration since 2004 or earlier (e.g., Poland, peaking at −36 000 in 2006, down to −7000 in 2012). In the poorest countries, e.g., Bulgaria, the peak estimated outflow was −214 000 in 2001, −3000 in 2012; and in Romania, peak outflow was −458 000 in 2006, down to −3000 in 2012. However accurate or otherwise these figures are, outflow has undoubtedly been large and (until 2014) mostly short term or illegal. In the last-named countries, emigration has led to severe rural depopulation and accelerating population decline, a decline reinforced by chronic low fertility and high mortality. The former Soviet Baltic states (Estonia, Latvia, and Lithuania) have also experienced high outflows, especially the latter two, all the more alarming to them as their population size is so small. The Russian Federation, however, has recorded high net inflows, increasing from 99 000 in 2004 to 329 000 in 2012. It is the center of a different migration system, initially receiving many persons of Russian 'nationality' who suddenly found themselves to be 'foreigners' when the Soviet Union collapsed in 1991, later receiving inflows from citizens of the neighboring countries of the former Soviet Union (the 'near abroad'): Ukraine in Europe, and Uzbekistan, Kyrgyzstan, Kazakhstan, and other states in Central Asia and the Caucasus. That immigration, attracted by employment and relative prosperity founded on oil and gas, has helped to arrest and even slightly reverse the population decline in Russia that began after 1991.

The overall tendency for migration to increase (until recently) masks a number of strongly divergent trends (Figure 2). The UK showed a marked increase to become the largest recipient of immigration in the EU, displacing Germany between 2003 and 2010 from its previous top position. In 1993, net immigration to the UK had been negative. By contrast, Germany (until 2009) and France have experienced smaller and generally diminishing net flows, that to Germany being briefly negative, as was that to the Netherlands from 2003–07. Annual flows in each direction can remain very large even if the net balance is small. Flows are small in absolute numbers to small countries, e.g., Denmark, Norway, and Switzerland, are often large in relation to their national populations.

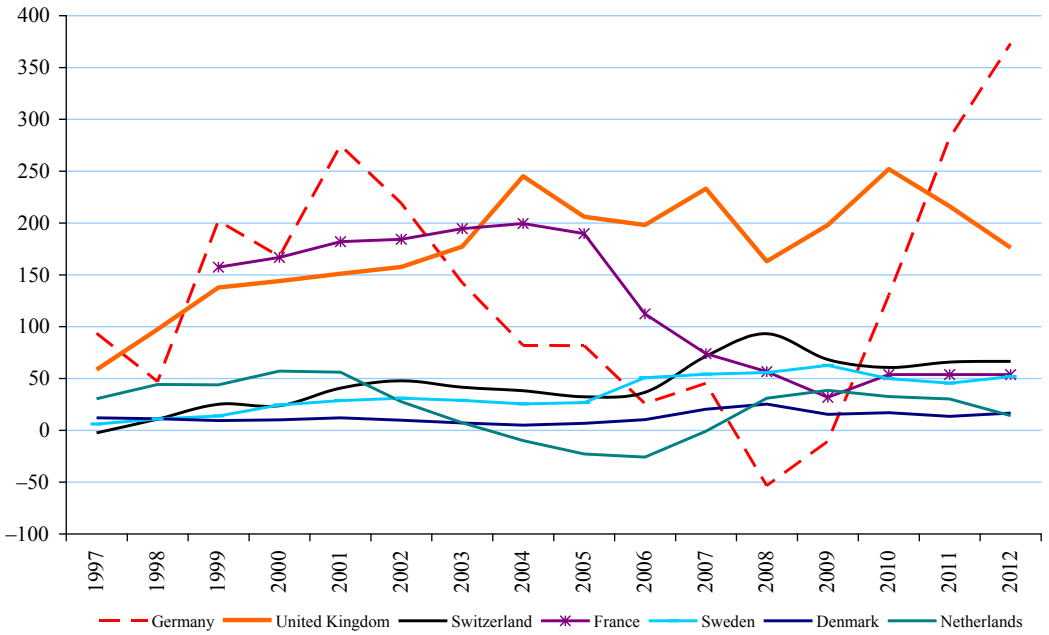

Figure 2 Net migration, selected European countries, 1997–2012. Source: Eurostat.

Spain and Italy, traditional emigration countries until the 1980s, had relatively modest published immigration figures until the late-1990s. Later migration, some of it illegal or unrecorded, from Latin America, North Africa, and eastern Europe, changed all that. When repeated amnesties regularized much of the illegally resident population, the picture changed radically (Figure 3). After 2007, net inflows fell precipitately, with many of the earlier immigrants, along with citizens, moving on to other EU countries (Figure 3). More people left Spain than entered it in 2011 and 2012. Migration to Ireland, 12 per 1000 population up to 2007 – a very high figure for a country accustomed to losing emigrants – became negative but has now increased again.

Migration Flows According to Purpose of Journey

It is difficult to make Europe-wide comparisons of the relative strength of inflows for different purposes, except asylum. However, in general, since the 1970s most (regular) migration has not been specifically for purposes of work but more for family reunion, family formation, as other dependents, study

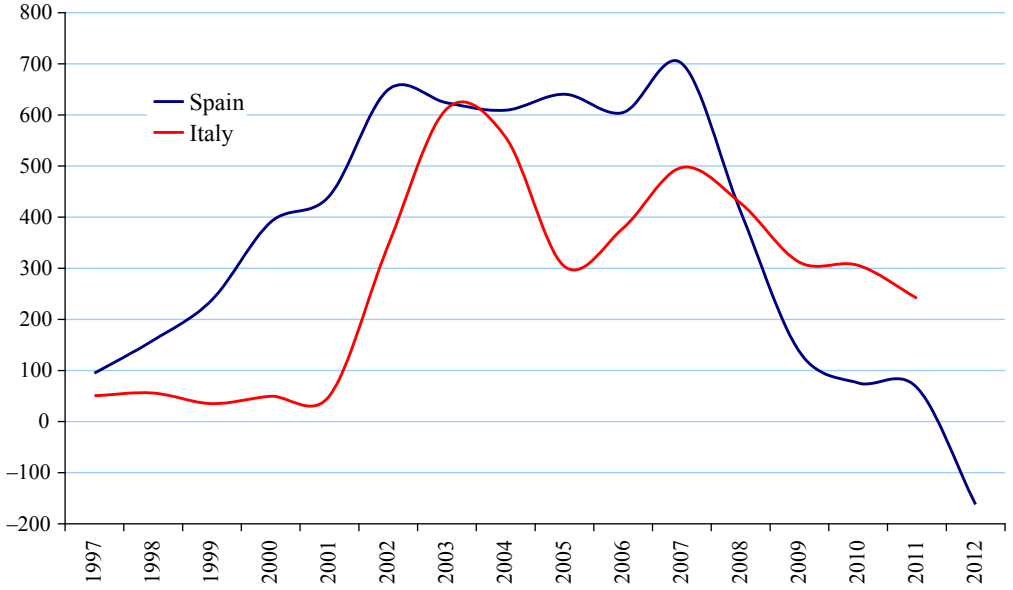

Figure 3 Net migration, Spain and Italy, 1997–2012. Source: Eurostat.

and asylum seeking. In most of the guest-worker countries, of which Germany was the most important, labor migration was severely restricted following the oil shock of 1973. By that time, the traditional industries for which the mostly low-skill labor had been recruited were in any case either in terminal decline or were remaining profitable by replacing routine labor with automated processes. Although demand continues – and is likely to increase as populations age – for low-skill migration for services, for example, for the care of the elderly, cleaning, and catering, recruitment has increasingly concentrates on skills and competition has developed for highly skilled immigrants (Salt et al., 2004; Lowell, 2005), with selective programs to attract them (documented in McLaughlan and Salt, 2002) including the EU Blue Card program. High-skill migration is often mediated through intracompany transfers.

The spouses, dependents, students, asylum seekers, and the rest may work and do, of course, but their labor force participation can be low, a handicap that can extend into the second generation (OECD, 2007: pp. 62–95). Students may stay to work, and in some countries graduates are now encouraged to do so. In 2001, the proportion of non-EU immigrants to Sweden entering as labor migrants was 3% (gross inflow) and to Denmark in 2005 2% (net inflow). By contrast, in some years labor migration to Germany has been substantial; 380 000 entered for work in 2004 out of a total gross inflow of 602 000 foreign citizens (63%, OECD, 2006).

Family migrants and dependents, mostly from outside Europe, are more likely to remain than are skilled labor migrants, more of whom come from the developed world (Rendall and Ball, 2004). The self-perpetuating process of chain migration from poor countries, creating substantial and growing transnational populations is described as 'cumulative causation' by Massey and Zenteno (1999). Most of that was neither inevitable nor planned. It followed on from a labor migration assumed to be temporary through the development, for humanitarian reasons, of systems of rights for family migration and other immigrant entitlements, in conjunction with the other postwar 'revolutions' of information and communication (Freeman, 1994; Martin et al., 2006: p. 222). The example of France from 1946 illustrates the transition from labor to family migration typical of many European countries (**Figure 4**).

The pattern of gross inflow to France in 2010 is shown in more detail in **Table 1** below, excluding immigration from the European Economic Area (EEA). 95% of that inflow was from outside Europe, 54% was from Africa, 25% was from Asia. Immigration from developing countries is mostly of family, spouses, dependents, students, and asylum. Only 9% of non-EEA migrants entered for work. 45% entered for family reasons (reunification and new marriages), 31% were students. These data do not show the number of persons from other EEA countries entering for work. In 2005, EEA immigration to France comprised 21% of the total, 34% for work, 6% family, and 46% retired and other. While patterns vary, these are not untypical of inflows to western European countries.

Until about 2004 flows within the EU were not great and were relatively numerically balanced. The removal of internal barriers to movement had not ironed out persistent regional disparities in labor shortage and surplus, and the earlier EU member states were relatively similar in economic level (Rees and Kupiszewski, 1999). Long-standing cultural differences even between the different regions of European countries in the western part of Europe still strongly affect internal migration patterns, for example in Germany (Falck et al., 2010). With an

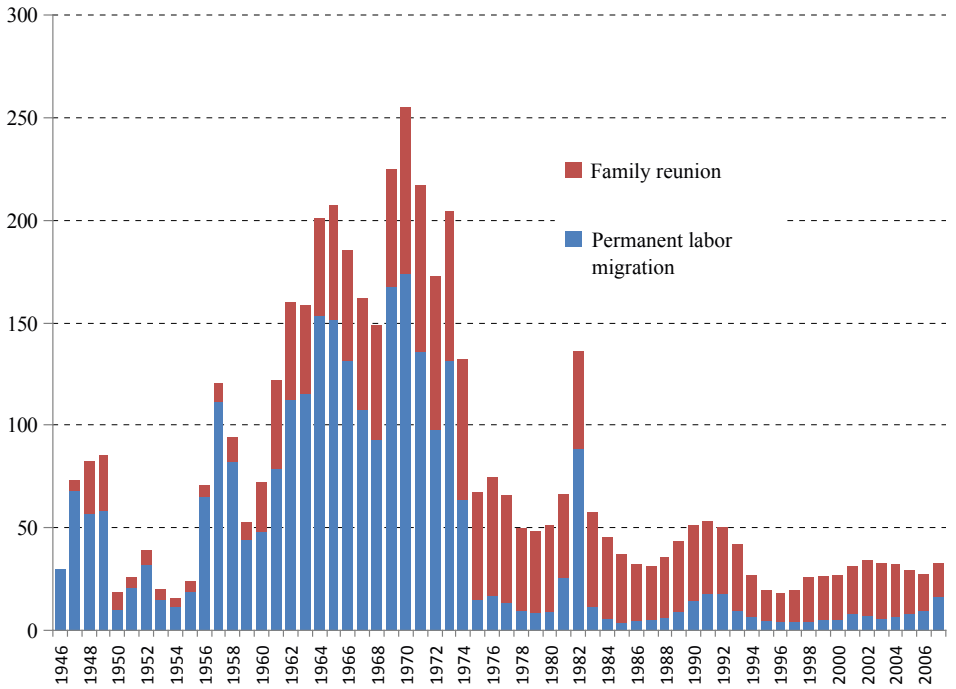

Figure 4 Entries under family reunion and permanent labor immigration in France, 1946–2007 (thousands). Source of data: OECD SOPEMI 2009, Chart I.13. Note: the spike in 1982 is an artifact of a change in legal status.

Table 1 Gross inflow to France by purpose of entry and nationality, 2010

France – first admissions by reason and nationality 2010 (column percent)								
	Economic	*Family*	*Student*	*Other*	*Humanitarian*	*All*	*Number*	*Percentage of total*
Europe – not EEA	7	4	4	5	13	5	9796	5
Africa	40	69	41	49	48	54	105756	54
Sub-Saharan	20	23	19	21	41	23	45058	23
Asia	27	15	35	22	34	25	47647	25
Japan	4	1	3	2	0	2	3107	2
America	26	12	20	24	4	16	30675	16
The United States	12	1	6	9	0	4	7481	4
Other	0	0	0	0	1	0	287	0
All	100	100	100	100	100	100		
Number	18067	86827	59964	11514	17785		194157	100
Percentage of total	9	45	31	6	9	100		

Note: EEA is the European Economic Area (EU plus EFTA). 'Other' includes retired, visitors, children.
Source: data from INSEE (2012) Fiches Thematiques: Flux d'immigration http://www.insee.fr/fr/ffc/docs_ffc/ref/IMMFRA12_h_Flot2_flu.pdf; INSEE ADREF (SGII) Table 2.

enlarged EU, the inflow of people from eastern Europe to the UK, Ireland, and Sweden following the accession of Central and eastern European counties to the EU in 2004 (A8 countries) changed that. A cumulative total of 1.1 million applications under the Workers' Registration Scheme had been made to the UK between 2004 and the end of the scheme in 2010 (gross inflow). The UK government, which chose not to adopt the 7-year delay on free movement of labor after A8 accession, which was imposed by the other major EU economies, had forecast 13 000 per year. The Polish population in the UK has increased tenfold since 2004, to over 600 000, and births to Polish-born mothers are now the single largest national category of births to foreign-born women in England and Wales.

However, even before the economic crisis, migration from the A8 countries was slowing and the return migration increasing. As economies of east European source countries grew, stimulated by EU membership, the value of Eastern currencies rose, and so diminished the exchange-rate advantage of working in the west. Furthermore, rapid domestic growth together with labor shortages exacerbated by the emigration has promoted domestic labor demand and wage-rates. Despite all that, eastern European migration to western Europe remains substantial, augmented by inflows from Bulgaria and Romania, including numerous Roma/gypsies.

Emigration of citizens from some western European countries increased, including the UK (Evans et al., 2007) and Germany (Sauer and Ette, 2007), and even France. In the Netherlands, the outflow was attributed to dissatisfaction with domestic conditions, crowding, and social and environmental deterioration (van Dalen and Henkens, 2007).

Asylum Seeking

In the Cold War period, most asylum seekers came from behind the Iron Curtain. Inflows were small, uncontroversial, and often politically useful to Western countries. All that changed in the 1980s when asylum seeking from outside Europe began to accelerate, from Iran, Turkey, Sri Lanka, and elsewhere. It grew to very high levels in the 1990s, partly in response to political crises in the Balkans and Middle East. Germany was the favorite destination through its economic attractions and the generous provisions in its Basic Law. Claims to Germany peaked at 438 190 in 1992 before returning to lower levels (19 165 in 2007). Asylum claims fluctuate considerably depending on the political situation in sending and receiving countries, and economic trends, but there are always substantial streams from South Asia, the Maghreb, and Middle East and sub-Saharan Africa. Much of it is managed by traffickers who keep abreast of political developments and switch destinations accordingly. Asylum seeking to EEA countries has increased since 2008 from 257 475 applications to 373 995 in 2012. Applications to Germany (the most popular destination) almost tripled to 77 650, doubled to Sweden (to 43 945) and to Switzerland (to 28 640), and increased 50% to France (to 61 455). Applications to weaker Eurozone countries (e.g., Greece, Ireland, Italy) fell sharply. Most claims are judged to be unfounded (2/3 of first instance decisions by EU28 countries in Q2 2013) but most applicants will remain in Europe. The latest figures at the time of writing show an extraordinary increase in claims – 50% higher in Q2 2013 than Q2 2012 (Eurostat, 2013b). Claims to Germany doubled in that time. Asylum seeking seems to have discovered Central and eastern Europe, which formerly received few claims. Claims to Hungary increased 25 times since Q2 2012, to Bulgaria four times, to Poland three times. Russians and Kosovars comprised half the increase, after Malians and Syrians (Russia 16845, Syria 8310, Kosovo 8140, Afghanistan 6075, Pakistan 5965, Somalia 4285, in Q2 2013).

Foreign Citizens and Immigrants in the EU Countries

33.3 million foreign citizens have lived in the EU27 countries in 2011, 6.6% of the total population. They were strongly concentrated in western Europe. Of these 21 million were citizens of non-EU countries. Romania, Turkey, Morocco, and Poland were the four most important countries of foreign citizenship in the EU27 countries in 2011. Data of population of foreign origin based on citizenship substantially

underestimate the number of residents born abroad, through the high level of naturalization. In some countries, the number of naturalizations has been about equal, each year, to net immigration flows. Better data on birthplace now gives us a better estimate of the number of immigrants living in European countries. 48.9 million foreign-born residents lived in the EU27 countries in 2011, 9.7% of the total population. 32.4 million of them were born outside the EU and were therefore immigrants from outside Europe, a Figure 50% higher than the number according to citizenship (Eurostat, 2012, 2013b).

What Can Be Said about the Future of Migration?

Migration theory remains fragmentary and cannot easily forecast the likely future trends of aggregate migration flows. Many of the events driving it, notably the economic trends and even more so, political change, are unpredictable. Comprising numerous unrelated flows from many origins for unconnected purposes, migration has defied satisfactory modeling or projection except where it is dominated by regular labor migration. Political processes at home and abroad can be paramount. For population projection, most national statistics offices assume the continuation of the current level or the extrapolation of recent trends. Critics of such a simple approach are usually baffled to suggest anything better. However, increasingly sophisticated approaches are being made to improve understanding (Howe and Jackson, 2005; Eurostat, 2007; Lemaitre et al., 2007; Duque and Montoso, 2008; Raymer and Abel, 2008; Kim and Cohen, 2011). Economic models succeed best when migration is primarily for labor, usually it is not. The buoyant Norwegian economy provides an exception. The econometric model of Brunborg and Cappelen (2009) fits past trends well and, for example, predicts a downturn in immigration to Norway as its economic attractions are expected to fade as its oil runs out.

Generally, migration pressure to developed countries from the South is likely to remain high, and possibly increase for some time, before declining. Economic and demographic disparities between the global North and South continue and in some cases have widened. Whether moving specifically for work or for a 'better life' in general, residents of poor countries are usually materially better off somewhere else. On the demand side, employers demand easy access to labor, especially if earlier migrant flows have made them dependant upon it, and if population aging or labor market protection restricts domestic supply as in many parts of Europe. Many third-world countries remain politically unstable. The 'cumulative causation' of established immigrant populations provokes further immigration through networks and chain migration (Gurak and Caces, 1992; Mitchell and Pain, 2003; Massey and Zenteno, 1999). Commitment to human rights, family reunion and asylum conventions, and growing ethnic electorates make it difficult for western states with liberal pretensions to restrict immigration effectively (Freeman, 1994; Castles, 2007). In Europe, the EU Commission, supported by the UK government among others, has stated that it wishes to expand EU membership to countries on the edge of Europe and beyond: to Ukraine (population 50 million), to Turkey (population 76 million), and even to North Africa, countries with larger populations and even lower levels of development than those recently admitted. This would guarantee migration pressure well into the future (Rowthorn, 2009). Rising economic inequality within countries such as the UK may also encourage immigration (Hatton and Tani, 2005).

Most commentators have concluded that the economic downturn will have only transient effects on inflow, 'A8' apart, because most immigrants are not labor migrants. Marriage migration and the inflow of dependents, students, asylum seekers, and others may be little affected (Dobson et al., 2009; Beets and Willekens, 2009; OECD, 2009: pp. 63–65). Nonetheless, provisional data so far show that its effects have been substantial in some countries (Papademetriou et al., 2011).

Many skilled labor migrants come from India to developed countries, notably to work in the IT sector. Crisis apart, this and similar flows have been increasing (Salt, 2009). But in the long run, labor migration from India may well decline, if the continued rapid growth of the Indian economy absorbs more of its own specialists, That applies a fortiori to China, with its more rapid economic growth and faster population aging. Basing themselves on labor market considerations, Barrell et al., 2009, using the model of Mitchell and Pain (2003), project that the narrowing of the economic gap would reduce substantially net migration to the UK from the A8, the Old Commonwealth countries, India, Latin America, halving the current officially projected increase in migrant stock by 2030 (p. 37). But that scenario does not take much account of noneconomic migration.

Students are a major component of migration flows to Europe from the third world, and up to stay to work or marry. China dominates these flows, with 350 000 students studying abroad in 2007, projected to be 645 000 in 20 years. China has a huge program to improve its domestic higher education – numbers are growing at 25% per year between 1999 and 2004. But even that does not appear to compensate for the growing demand. Chinese students overseas are projected to increase to 645 000 by about 2025. Eventually, the growing domestic sector and the demographic diminution of the numbers of young people in China must reduce overseas migration, but that is unlikely to become apparent for decades. In the meantime, other countries are sending more students abroad: from India 130 000 in 2005, projected to increase to 300 000, followed by South Korea (127 000), Morocco (106 000), and Turkey (104 000) (Banks et al., 2007; Maslen, 2007).

Arranged marriage – preferably first-cousin or uncle-niece in Pakistani culture – is favored in many traditional societies and is a growing component of migration to many European countries, displacing old-style family reunion with new family formation (e.g., Charsley et al., 2012). Some of it motivated more by the prospect of migration than the pleasures of marriage. In the UK, marriage migration from the Indian subcontinent has been growing roughly *pro rata* with the growth of the young South Asian population in the UK. In the Netherlands, it had become the most important component of all migration from North Africa and Turkey. Marriage migration comes from many parts of the traditional world; for example, numbers from sub-Saharan Africa to the UK trebled from 1996 to 2006.

Otherwise, and medical personnel apart, immigrants from the least developed third world, are mostly asylum seekers, students, dependents, and illegal entrants. Chronic political instability, rapid population growth, and economic and environmental fragility are likely to keep inflows high for the foreseeable future. It may increase substantially if global climate change has early effects. The latter may override the outcomes of any scenario based only on socioeconomic and political considerations. Its widely forecast effects have now crept within the time horizon of population projections. If those forecasts prove to be correct, then there could be major implications for international migration.

Effects of Policy

In part, the volatile track of migration trends follows from unpredictable policy and political changes in sending and receiving counties (Hollifield, 2000). Some of the most important political changes in sending countries are those that provoke, or ease, asylum seeking, either as a result of persecution of specific minorities promoted by governments or not controlled by them (e.g., Kurds in Turkey, Christians in Palestine, Iraq, and Syria), or economic mismanagement, or both (e.g., Zimbabwe). The regime change in Iraq, which was hoped to reduce the flow of refugees from the previous regime, has instead provoked the flight of well over a million people to neighboring Syria, who now find that they have moved from the frying pan into the fire.

In the receiving countries, successive governments differ in their perception of migration (Schain, 2008). They receive contradictory messages from business demanding access to skilled and other labor, and worries about the care of an aging population, and the popular dislike of large-scale immigration, which is general in Europe (Martin, 2010). New policies on immigration can be brought in by governments elected for other reasons (for example, the Labor government in the UK 1997–2010, which liberalized policy). In other cases, immigration itself is a factor in the change of government, usually to control it further (for example, the UK coalition government elected in 2010). For well-known reasons, it is easier in liberal democracies to encourage immigration than to restrict it. For example, the incoming Labor government in the UK reversed the previous restrictive policy of the previous 30 years, promoting immigration for a variety of reasons including expansion of labor supply and GDP. By mid-2000s, this trebled net immigration compared with the early 1990s, taking net inflows to the highest level in the EU except for Spain and Italy. The radical increases in net inflows to those two countries arose partly because of political change, including the return of a Socialist government in Spain favorable to migration. Amnesties regularized large numbers of illegal immigrants. Policy changes in Sweden and Denmark to favor skilled labor migration have also increased overall net flows to those countries, and in Sweden immigration is projected to increase to mid-century (Statistics Sweden, 2009). For most of the postwar period, Germany received about two-thirds of immigrants to Europe and two-thirds of asylum seekers. But a variety of measures, beginning with a reform of the Fundamental Law in 1991 to limit asylum seeking, and later to regulate the inflow of *aussiedler* (people of German origin in eastern Europe), together with less impressive economic growth, greatly reduced net immigration from 1990 to 2009 (**Figure 5**).

In France, labor migration had been selective and restricted for some time (**Figure 3**). A succession of measures: imposing tighter conditions on asylum, naturalization, marriage migration, and measures against illegal immigration and to facilitate deportation have developed over the decade, especially under the Presidency of M. Sarkozy (2007–12) alongside a new policy on 'national identity.' A different approach

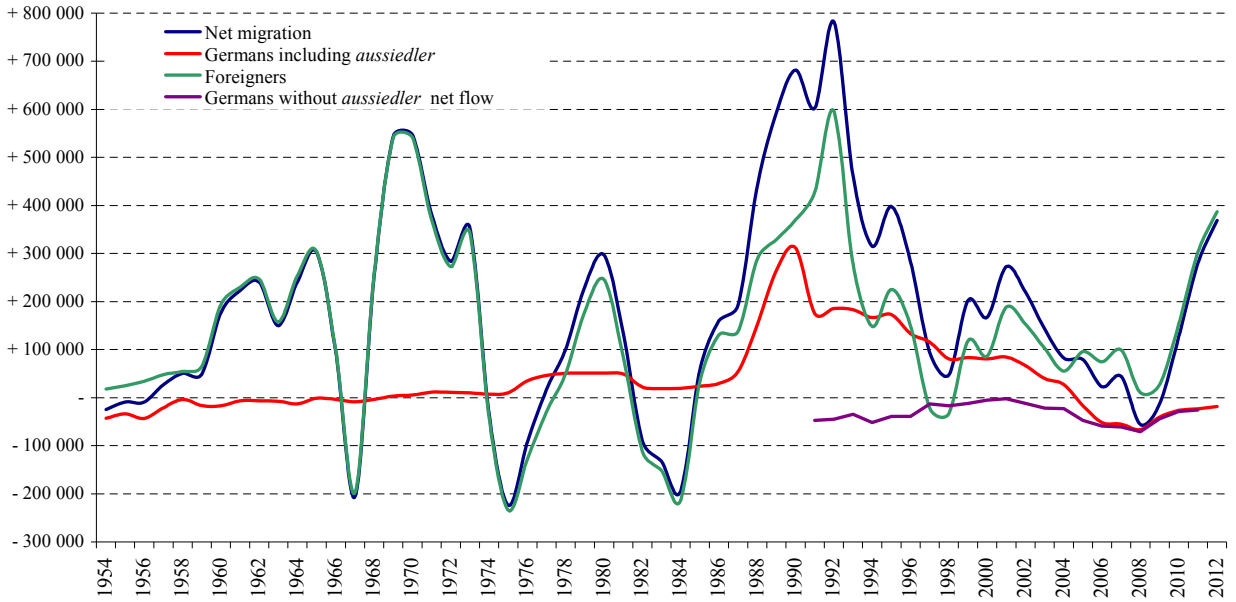

Figure 5 Net migration to Germany, 1954–2012, according to citizenship. Source: Statistisches Bundesamt, Wiesbaden. Up to 1990 German Federal Republic only.

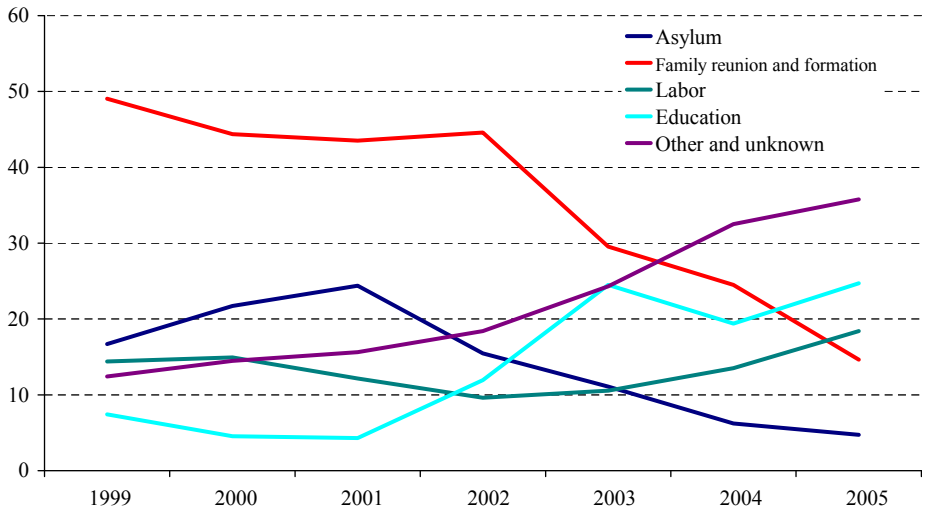

Figure 6 Net Immigration to Denmark by broad purpose of entry. Source: data from Statistics Denmark.

to migration is adopted by the new Socialist government of M. Hollande.

Some policies aim to restrict some immigration streams (notably those not focused on economic contributions), while leaving or even promoting others. Thus in Denmark, legislation, since 2002, has severely tightened the requirements for marriage migration, reducing the flow to one-third of their previous level by 2005 (Rockwool Foundation Research Unit, 2009). As of 2013, these are the latest data available from Statistics Denmark (**Figure 6**).

Danish policy has also been to admit more skilled workers, so the total volume of immigration is now higher that 10 years ago, although its composition is considerably different. Marriage migration (especially in respect of the second generation) among Moroccans and Turks in the Netherlands fell sharply following a policy change in 2004 (Statistics Netherlands, 2009). As a consequence, a higher proportion of Turks residing in the Netherlands are now marrying spouses of Turkish background who are also living in the Netherlands, rather than spouses introduced from abroad.

Despite all this, the power of European countries – especially those within the EU – to limit immigration is limited. Most immigration is not (ostensibly) for work. All the countries concerned are signatories to treaties and conventions (the Geneva Convention on Refugees, the European Convention on Human Rights, and EU legislation), which give considerable entitlements to persons wishing to enter the EU for purposes of marriage, as dependents, and for asylum, and which restrict powers of deportation. While individual counties may seek to restrict the conditions for such entitlements, judgments of the European Court of Human Rights constantly expand them. National freedom of action in these areas is considerably restricted in the EU.

Economic Consequences of Immigration

Much orthodox opinion, supported by neoclassical economic theory and the concerns of business and employers, holds that immigration benefits economic growth, increasing GDP, expanding the workforce, bringing skills, damping wage inflation, compensating for population aging, and a diminishing workforce in very low-fertility countries. Critics of the economic and social effects of large-scale immigration risk being labeled as misinformed, irrational if not racist, or 'xenophobic.' That became the establishment view (Coleman, 2003), although electoral reversals have forced a retreat from this moral high ground. Public opinion has been usually at variance with elite views on immigration (Chamie, 2009), usually favoring restriction, not expansion. In Europe countries, unlike countries of European origin overseas, immigration has no place in perceptions of national identity, but is normally seen as a challenge to it (Collier, 2013). Doubts have been raised about the unqualified economic benefits of immigration, suggesting that immigration may impose net costs, not benefits, on receiving society even by the narrow standards of fiscal accounting (Wadensjö, 1999; Borjas, 2003). This is very difficult to evaluate. Most studies conclude that the fiscal benefits overall are small; beneficial in respect of skilled migrants mostly from richer countries, not so from less skilled migrants mostly from poorer countries. The first large-scale international comparative study of fiscal effects concluded that over the OECD member countries, 'the fiscal impact of immigration is close to zero on average' (OECD, 2013: p. 9, and 125–161). As to economic growth, naturally immigration increases GDP as it increases population, assuming that most immigrants work. But it has little effect upon GDP per head, which is the appropriate measure of welfare (House of Lords, 2008). Benefits are very uneven. The unequivocal beneficiary of immigration is the immigrant. Employers benefit from easier access both to skilled and to unskilled labor, as do consumers of services. There is stiff international competition for talent in a global market. Universities gain highly talented staff and researchers, and student applicants, especially from the developing world. The poorer part of the domestic population does much less well, exposed to competition and to downward pressure on wages. Evidence is conflicting on the effects upon unemployment. But immigration can flourish, while domestic

unemployment is high (as in Spain; OECD, 2013: p.19) and high proportions of the working-age population – especially youth – remain economically inactive. This seems perverse.

All developed countries have aging populations, but the rapid population aging and incipient or actual population decline in some low-fertility counties of southern, central, and eastern countries has promoted claims that immigration is needed to maintain population and workforce and to ameliorate population aging. Earlier delusions that immigration could 'solve' population aging, promoted by a misinterpreted UN report (2000) have now mostly been exploded (Bijak et al., 2008). As in most policy areas, political considerations will overcome supposedly scientific research-based prescriptions, and immigration policy can change abruptly as a result of the fall of the electoral dice.

Effects on Population

The effects on the population of European countries are substantial, but diverse (Coleman, 2008; Rowthorn, 2009). As might be imagined, the prosperous countries of western Europe (in the broad sense) have been attractive to immigrants, including to immigrants from the former communist countries of Central and eastern Europe. The latter for the most part have lost population by emigration to the west, although in some cases that is compensated or more than compensated by new migration from the west (e.g., to the Czech Republic) and from further east (Ukraine) and the global South; in transit to western Europe, but in many cases remaining.

The Russian Federation is an exception. After the fall of communism it became a major country of immigration, initially of Russian nationals departing newly independent former Soviet states (the 'near abroad'), more recently a larger number of the nationals of those states. Net immigration in 2006 was 132 319. By 2009 immigration, estimated at 240 000, was reported to have just balanced the natural decrease of the population, generating a small increase in population for the first time since the mid-1990s. Immigration increased to 329 000 in 2012.

In western Europe, a number of countries with moderately high fertility still have positive natural increase, thanks to demographic momentum, assisted by the higher birth rates seen almost everywhere in recent years (some Scandinavian countries, France, the UK). With current migration, their populations are projected to increase by up to 20% by mid-century. Others (Spain, Italy) would decline substantially without their current high immigration, and the population of Germany fell from 2004 to 2011 until growth was restored – at least for a while – by the new high immigration depicted in Figure 3.

Formerly relatively modest in comparison with the number of births and with natural increase, and sometimes negative, net immigration now exceeds natural increase in most European counties and even approaches (Switzerland and Norway in 2011) the total number of live births each year (Table 2). If these trends continue, populations of immigrant origin will reach high proportions of the national population in a few decades, possibly eventually becoming the majority (Figure 7).

Projections of future population by national origin have been made for several European countries up to mid-century or 2060 (Coleman, 2006). In most cases, they are based on the assumption that immigration will continue at approximately recent levels and that fertility differentials will diminish or disappear. On those assumptions, populations of 'foreign origin' or 'foreign background' (mostly defined through population registers as first and second generation together) would rise to between 15 and 35% of the populations of some European countries by mid-century. Of those, about one-third is projected to be people of other European origins, two-thirds from outside Europe (Figure 8).

Table 2 Selected Western countries, Comparison of live births, net immigration, and natural increase 2011

	Population 1 January 2011	*Live births*	*Natural increase*	*Net migration*	*Net migration as percent of births*
		Data (thousands)			
Switzerland	7870	81	18	66	81
Norway	4920	60	19	47	78
Austria	8404	78	2	37	48
Italy	60 626	547	−47	241	44
Germany	81 752	663	−190	282	43
UK	62 499	808	256	235	29
Denmark	5561	59	7	13	23
Belgium	11 001	131	23	16	12
France	63 128	828	253	77	9
Spain	46 153	469	86	−42	−9
Greece	11 310	106	−5	−15	−14
Total of above	363 224	3830	420	956	25
Australia	22 684	298	151	208	70
Canada	34 484	382	129	209	55
New Zealand	4408	61	31	−3	−5
United States	309 122	3999	1531	695	17

Note: US data refer to 2010; Australia, Canada, and New Zealand to 2012. Net immigration data not available for US Admission for permanent settlement figure reduced by 1/3 allow for return migration.
Sources: Eurostat, Australian Bureau of Statistics, Statistics Canada, Statistics New Zealand, US Census Bureau, National Center for Health Statistics, US Department of Homeland Security.

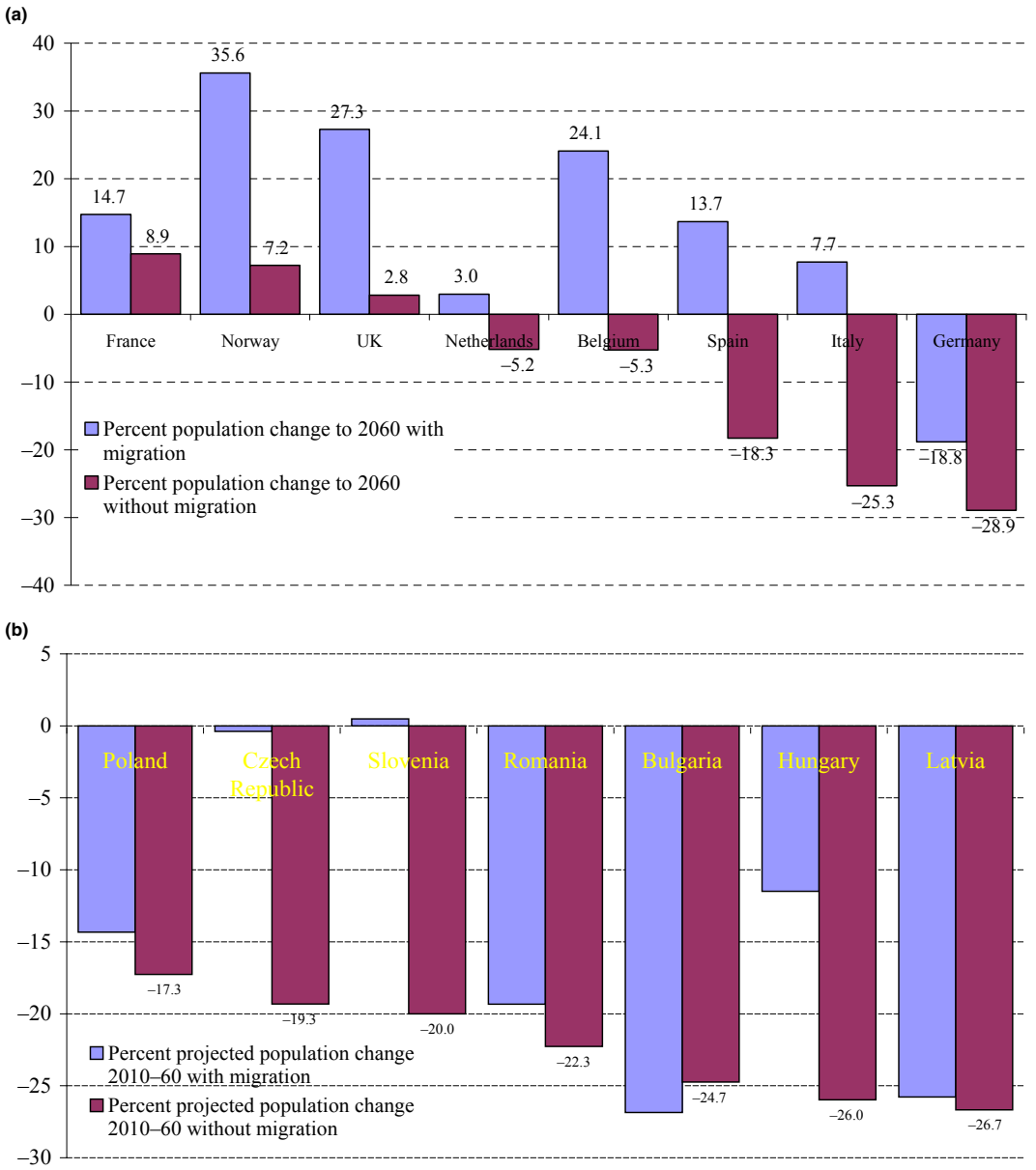

Figure 7 (a) Effect of migration on projected population growth 2010–60, percent, selected western and southern European countries, by descending order of growth without migration. (b) Projected percent population change 2010–60, selected eastern European countries, with and without migration, sorted by descending order of projected population size without migration. Sources: (a) data from Eurostat, 2008. Aging characterizes the demographic perspectives of the European societies. Statistics in Focus (72/2008). Office for Publications of the European Communities, Luxembourg, pp. 12. (b) data from Eurostat, 2010. EUROPOP2008. Convergence Scenario, National Level. Office for Publications of the European Communities, Luxembourg (web only).

Pan-European Policy Development

The European Commission (2005, 2008), had gradually extended its control over immigration policy in EU countries. The Schengen common borders area has been established including most EU member states (not UK). A 'Blue Card,' inspired by the US' Green Card system, to encourage highly skilled labor migration was formally adopted in the EU as a directive in 2009. The criteria are demanding and the take-up has been slight. The demographic and labor market and economic scenes in European counties seem too diverse for any single immigration policy to be appropriate. Common asylum policy is an exception. A Common European Asylum System was begun in 1999 at the EU summit in Tampere and successively consolidated, most recently in June 2013 for implementation in 2015 (European Commission, 2013).

Concluding Remarks

To recap, immigration is the dominant demographic driver in many European countries, provides a growing proportion of

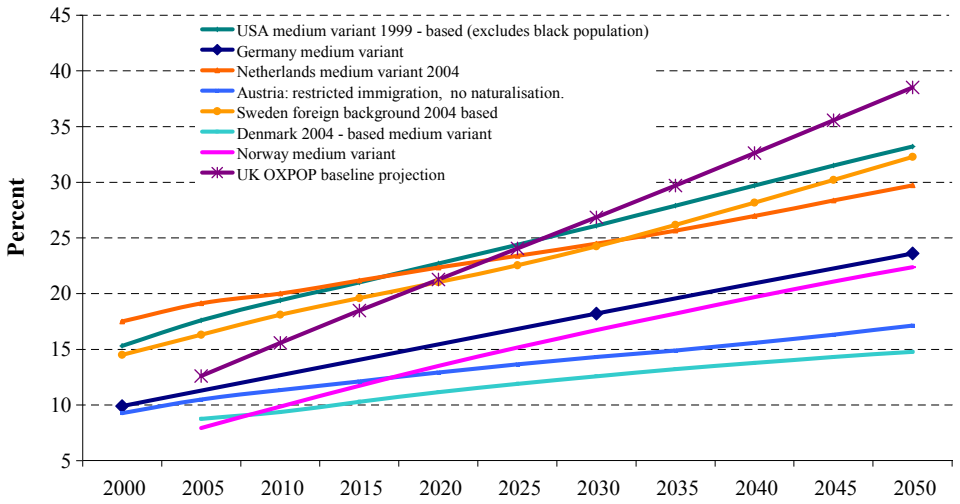

Figure 8 Projected growth of population of immigrant or foreign origin 2000–50, selected countries, as percent of total population.

the workforce, and is changing the composition of the population in an increasingly visible fashion. Trends and policies remain diverse, but most indications are that immigration will continue at a high level. Many of the streams to the EU (family, dependents, Asylum) are driven by entitlement and are difficult to moderate.

In some of the countries mentioned above, the labor force is not projected to decline for some time if at all and labor migration is concentrated upon filling highly skilled shortages. Elsewhere, very low birth rates, if continued, may to provoke such future labor shortages, which only immigration can resolve even if fertility does recover. But in some continental countries, many of the labor shortages of which employers, governments, and analysts complain follow from restrictive labor market policies that favor those in employment, not those seeking it, overgenerous retirement and welfare arrangements and high payroll taxes. This 'European social model,' notably in southern Europe, contributes to high unemployment, especially of youth, low workforce participation rates, and dependency on immigrant labor, often irregular, to provide flexibility. Reforming all of this has been slow, painful, and difficult, although spurred recently by economic crisis. These are the same populations, with their 'familist' culture, that also suffer the lowest birth rates and consequently the most rapid population aging (Dalla Zuanna and Michielin, 2003).

So in the near future, at least, there are several reasons for supposing that substantial reductions in numbers of international migrants are unlikely. Global population will rise, especially in Africa and poorer Asian countries, increasing emigration pressures. Environmental deterioration will encourage emigration from marginal areas, partly arising from population pressure on land itself, partly from the effects of global climate change. The magnitude and form of the latter, and its effects upon population, are doubly uncertain. If the economy continues to be globalized, new migration sources may emerge. The aging of populations in OECD countries, especially those with chronic low birth rates, will strengthen arguments for labor immigration, particularly if cultural and political obstacles persist to the rationalization of labor market and retirement arrangements in those countries. Competition for skilled and well qualified migrants will continue. Insofar as immigrants are naturalized and can vote, growing populations of immigrant origin may make it more difficult for political parties opposed to large-scale immigration to win office. Set against that, popular opposition to such migration has been growing in a number of European countries and has provoked a number of electoral surprises. And in the much longer run, as the economies of sending counties develop and their populations age, the assumption of unlimited labor supply through immigration may prove to be false.

See also: Assimilation of Immigrants; Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Ethnic Intermarriage; Immigration: Health and Mortality; International Migration and Remittances; International Migration by Ethnic Chinese; International Migration by Ethnic South and Southeast Asians; International Migration; Migration Out of Europe; Migration and Fertility; Migration to and from Latin America; Migration, Theory of; Migration: Cultural Aspects; Population Forecasts.

Bibliography

Baines, D., 1991. Emigration from Europe 1815–1930. Macmillan, London.

Banks, M., Olsen, A., et al., 2007. Global Student Mobility: An Australian Perspective Five Years On. IDP Australia Pty, Canberra.

Barrell, R., Gottschalk, S., et al., 2009. Projections of Migration Inflows under Alternative Scenarios for the UK and World Economies. Department for Communities and Local Government, London.

Beets, G., Willekens, F., 2009. The global financial crisis and international migration: an uncertain outlook. In: Coleman, D., Ediev, D. (Eds.), Vienna Yearbook of Population Research. Vienna Institute of Demography; Austrian Academy of Sciences, Vienna, pp. 19–38.

Bijak, J., Kupiszewska, D., et al., 2008. Replacement migration revisited: simulations of the effects of selected population and labor market strategies for the aging Europe, 2002–2052. Population Research and Policy Review 27 (3), 321–342.

Borjas, G.J., 2003. The labor demand curve is downward sloping: re-examining the impact of immigration on the labor market. Quarterly Journal of Economics 118 (4), 1335–1374.

Brunborg, H., Cappelen, Å., 2009. An econometric model for forecasting migration to Norway. In: IUSSP International Population Conference. Statistics Norway, Marrakech.
Castles, S., 2007. The Factors that Make and Unmake Migration Policies. In: Portes, A., DeWind, J. (Eds.), Rethinking International Migration: New Theoretical and Empirical Perspectives. Berghahn, Oxford, pp. 30–61.
Chamie, J., 16 October 2009. Mind the Gap: Public and Government Views on Migration Diverge. YaleGlobal.
Charsley, K., Storer-Church, B., et al., 2012. Marriage-related migration to the UK. International Migration Review 46 (4), 861–890.
Coleman, D.A., 2003. Mass migration and population change. Zeitschrift für Bevölkerungswissenschaft 28 (2–4), 183–215. Special issue In Memoriam. Dr. H.-J. Hoffmann-Nowotny.
Coleman, D.A., 2006. Immigration and ethnic change in low-fertility countries: a third demographic transition. Population and Development Review 32 (3), 401–446.
Coleman, D.A., 2008. The demographic effects of international migration in Europe. Oxford Review of Economic Policy 24, 453–477.
Collier, P., 2013. Exodus. How Migration Is Changing Our World. Oxford University Press, Oxford.
van Dalen, H.P., Henkens, K., 2007. Longing for the good life: understanding emigration from a high-income country. Population and Development Review 33 (1), 37–66.
Duque, I., Montoso, O., 2008. Producing Migration Data through Household Sample Surveys: The 2007 National Immigrant Survey (Spain); Extending the Knowledge on Immigrants beyond Registers. Working Paper 17, Joint UNECE/Eurostat Work Session on Migration Statistics. UNECE, Geneva.
Dalla Zuanna, G., Michielin, F., 2003. Immigrations and low fertility as endogenous component of economic growth. General comments and a case study. In: EURESCO Conference 2003, the Second Demographic Transition in Europe. Euroconference on Implications of Family and Fertility Change for Individuals, families and Societies.
Dobson, J., Latham, A., et al., 2009. On the Move? Labour Migration in Times of Recession. What Can We Learn from the Past? Policy Network, London.
European Commission, 2005. Green Paper on an EU Approach to Managing Economic Migration. Commission of the European Communities, Brussels.
European Commission, 2008. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions of 17 June 2008 – a Common Immigration Policy for Europe: Principles, Actions and Tools [COM(2008) 359 Final]. European Commission, Brussels.
European Commission, 2013. Common European Asylum System. European Commission, Brussels.
Eurostat, 2007. Work Session on Demographic Projections. Office for Official Publications of the European Communities, Luxemburg.
Eurostat, 2008. Recent Migration Trends: Citizens of EU-27 Member States Become Ever More Mobile while EU Remains Attractive to Non-EU Citizens. Statistics in Focus 98/2008. Office for Official Publications of the European Communities, Luxemburg.
Eurostat, 2012. Nearly Two-Thirds of the Foreigners Living in EU Member States Are Citizens of Countries Outside the EU-27-Statistics in Focus Number 31/2012. Office for Publications of the European Communities, Luxemburg. http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/publication?p_product_code=KS-SF-12-031.
Eurostat, 2013a. Migration and Migrant Population Statistics. http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Migration_and_migrant_population_statistics.
Eurostat, 2013b. Asylum Applications and First Instance Decisions on Asylum Applications - Second Quarter 2013. Issue No 12/2013. Eurostat, Luxemburg. http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/publication?p_product_code=KS-QA-13-012.
Evans, H., Chappell, R., et al., 2007. Using the Omnibus Survey to test questions on emigration. Population Trends 127, 15–20.
Falck, O., Heblich, S., et al., 2010. Dialects, Cultural Identity, and Economic Exchange. IZA Discussion Paper No. 4743. Forschungsinstitut zur Zukunft der Arbeit/Institute for the Study of Labor (IZA), Bonn.
Freeman, G.S., 1994. Can Liberal states control unwanted migration? Annals of the American Association for Political and Social Sciences 534, 17–30.
Garson, J.-P., Salt, J., 2011. 50 year review: international migration. In: International Migration Outlook: SOPEMI 2011. OECD, Paris, pp. 5–15.
Gurak, D.T., Caces, F., 1992. Migration networks and the shaping of migration systems. In: Kritz, M., Lim, L.L., Zlotnik, H. (Eds.), International Migration Systems: A Global Approach. Clarendon Press, Oxford, pp. 150–176.
Hatton, T.J., Tani, M., 2005. Immigration and inter-regional mobility in the UK, 1982–2000*. The Economic Journal 115 (507), F342.
Hollifield, J.F., 2000. The politics of international migration: how can we 'bring the state back in'? In: Brettell, C., Hollifield, J.F. (Eds.), Migration Theory: Talking across Disciplines. Routledge, London, pp. 137–185.
House of Lords Select Committee on Economic Affairs, 2008. 1st report of session 2007–08: the economic impact of immigration. Volume 1: Report HL Paper 82-1 pp. 84. Volume II: Evidence. HL Paper 82-II pp. 514. The Stationery Office, London.
Howe, N., Jackson, R., 2005. Projecting Immigration – a Survey of the Current State of Practice and Theory. Center for Strategic and International Studies, Washington, DC.
INSEE, 2012. Fiches thematiques: flux d'immigration. In: Immigrés et descendants d'immigrés en France, édition 2012. INSEE. INSEE, Paris, pp. 138–157.
Kim, K., Cohen, J.E., 2011. Determinants of international migration flows to and from industrialized countries: a panel data approach beyond gravity. International Migration Review 44 (4), 899–932.
Lowell, L.B., 2005. Policies and regulations for managing skilled international migration for work. In: United Nations Expert Group Meeting on International Migration and Development. United Nations, New York.
Lemaitre, G., July 2005. The Comparability of International Migration Statistics: Problems and Prospects. The OECD Statistics Brief No. 9. OECD, Paris.
Lemaitre, G., Liebig, T., et al., 2007. Standardised Statistics on Immigrant Inflows: Results, Sources and Methods. OECD Organisation for Economic Co-operation and Development, Paris.
Martin, P.L., Abella, M., et al., 2006. Managing Labor Migration in the Twenty-first Century. Yale University Press.
Martin, P., 2010. A need for migrant labour? UK-US comparisons. Who needs migrant workers? In: Ruhs, M., Anderson, B. (Eds.), Labour Shortages, Immigration, and Public Policy. Oxford University Press, Oxford.
Maslen, G., 2007. CHINA: Chinese students to dominate world market. University World News 1.
Massey, D., Zenteno, R., 1999. The dynamics of mass migration. Proceedings of the National Academy of Sciences of the United States of America 96, 5328–5335.
McLaughlan, G., Salt, J., 2002. Migration Policies towards Highly Skilled Foreign Workers. Report to the Home Office. Home Office, London.
Mitchell, J., Pain, N., 2003. The Determinants of International Migration into the UK: A Panel Based Modelling Approach. National Institute of Economic and Social Research, London. NIESR Discussion Papers 216.
OECD, 2006. International Migration Outlook 2006. OECD, Paris.
OECD, 2007. International Migration Outlook SOPEMI 2007. OECD, Paris.
OECD, 2009. International Migration Outlook SOPEMI 2009. OECD, Paris.
OECD, 2013. International Migration Outlook 2013. OECD, Paris.
Papademetriou, D.G., Sumption, M., et al., 2011. Migration and the Great Recession: The Transatlantic Experience. Migration Policy Institute, Washington, DC.
Poulain, M., Perrin, N., et al., 2006. Towards Harmonised European Statistics on European Migration (THESIM). Universitaires de Louvain Presses, Louvain la Neuve.
Raymer, J., Abel, G., 2008. The MIMOSA model for estimating international migration flows in the European Union. In: Working Papers, 8. UNECE/Eurostat Work Session on Migration Statistics. UNECE University of Southampton, Geneva.
Rees, P.H., Kupiszewski, M., 1999. Internal Migration and Regional Population Dynamics in Europe: A Synthesis. Population Studies No. 32. Council of Europe Publishing, Strasburg.
Rendall, M.S., Ball, D.J., 2004. Immigration, emigration and the ageing of the overseas-born population in the United Kingdom. Population Trends 116, 18–27.
Rockwool Foundation Research Unit, November 2009. Fewer family reunifications – more immigration for employment (and related articles). News from the Rockwool Foundation Research Unit. 1, 3, 5, 12.
Rowthorn, R., 2009. The economic and demographic impact of immigration on western Europe. In: Halle, M., Lane, D. (Eds.), Migration and the Movement of People in Europe: Threat or Benefit?
Salt, J., Clarke, J., et al., 2004. International Labour Migration. Council of Europe, Strasburg.
Salt, J., 2009. International Migration and the United Kingdom. Report of the United Kingdom SOPEMI Correspondent to the OECD. Migration Research Unit, Department of Geography, University College London, London.
Sauer, L., Ette, A., 2007. Auswanderung aus Deutschland: Stand der Forschung und erste Ergebnisse zur internationalen Migration deutsche Staatsbürger. Bundesinstitut für Bevölkerungsforschung, Wiesbaden.
Schain, M., 2008. The Politics of Immigration in France, Britain, and the United States: A Comparative Study (Perspectives in Comparative Politics). Palgrave Macmillan, New York.
Sobotka, T., 2010. Migration continent Europe. In: Coleman, D., Ediev, D. (Eds.), Vienna Yearbook of Population Research. Vienna Institute of Demography, Austrian Academy of Sciences, Vienna.
Statistics Netherlands, 3 November 2009. Decrease in marriage migration comes to an end. CBS Web Magazine.
Statistics Sweden, 2009. The Future Population of Sweden 2009–2060. Statistics Sweden, Stockholm.

United Nations Statistical Commission, UNECE and Council of Europe, 2006. The challenges of measuring emigration and remittances. Joint UNECE/Eurostat work session on migration statistics. Working Paper 5 Luxemburg, Statistical Office of the European Communities.

US Census Bureau, 2005. Rapporteur report on migration statistics. In: Statistical Commission and Economic Commission for Europe EUROPE Conference of European Statisticians. Fifty-Third Plenary Session. UNECE, Geneva.

Wadensjö, E., 1999. Economic effects of immigration. In: Coleman, D.A., Wadensjö, E., Jensen, B., Pedersen, S. (Eds.), Immigration to Denmark: International and National Perspectives. Aarhus University Press, Aarhus, pp. 290–328.

Relevant Websites

http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home – Eurostat. Data on Economy, Population, Migration for EU and Other European Countries.

http://dx.doi.org/10.1787/migr_outlook-2013-en – Organisation for Economic Co-operation and Development (OECD). Data and Publications Especially 'OECD International Migration Outlook 2013'.

http://www.un.org/en/development/desa/population/ – United Nations Population Division. Data and Publications on Population and Migration for All Countries of the World.

Migration and Development

Ronald Skeldon, University of Sussex, Brighton, UK; and Maastricht University, Maastricht, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the thinking on migration and development from two points of view. First, it looks at more conceptual ideas in the academic literature on long-term and macrolevel approaches to the topic. Second, it examines approaches to the management of migration in the current policy debate on migration and development. The focus is primarily on international migration, although reference is made to internal migration where appropriate and opposing views and critical approaches are highlighted. The article conceptualizes migration as an integral part of development that changes systematically as development, however defined, evolves.

Introduction

Migration is both a consequence and a cause of development and is an integral part of the development process. Analyses of migration generally involve the study of the factors that bring about the population movement, the so-called drivers of migration. Analyses also incorporate the impact that the movement has on both origin and destination areas. Both the drivers and the impacts can be seen as aspects of development.

Migration includes both movements within a country's borders, internal migration, and movements from one country or territory to another, international migration. Simple in concept, migration has proven extremely difficult to measure in practice in a way that can give comparative data from one part of the world to another. Just one difficulty is that migration, international as well as internal, is rarely a simple, one-way 'permanent' move from one jurisdiction to another, but rather a complex path of return and remigration. If forms of temporary migration are to be incorporated into the discussion, how are they to be defined? Development, too, is an equally complex term, its meaning often assumed without precise definition. Both migration and development are multidimensional terms.

Migration and development in this article will be considered from two perspectives: first, migration as an integral part of development and how forms of migration change with development and, second, the perspective of the current debate on migration and development in which migration is seen to be something that can be promoted to bring about development. The primary focus in this article will be on international migration but readers should be aware that many of the points made apply equally to internal migration.

Migration as an Integral Part of Development

At the simplest level, migration in a premodern agrarian society was very different from that associated with a modern, highly urbanized society today; not that the former was static (it was not) and the latter highly mobile, but the means, speed, volume, and composition of the flows of migration in the two societies are very different. Available technology determines how fast and frequently people move. Hence, if technology is how we measure development, the relationship is fairly clear. In the year 1774, some 265 ships carried 5027 migrants from English and Scottish ports to the Americas, an average of 19 emigrants per ship (Bailyn, 1987: p. 101). Almost exactly a century later, in 1873, 17 companies were operating 173 steamships between Europe and New York (Nugent, 1992: pp. 31–32), many carrying more than a 1000 passengers each. Over a century later again, in 2007, there were some 425 000 flights across the Atlantic, and International Air Transport Association (IATA) forecasts are for 3.6 billion passengers worldwide in 2016, up from the 2.8 billion in 2011 (IATA, 2012). Clearly, most people who fly today are not migrants setting out on a new life but short-term travellers on business or vacation. Nevertheless, these travellers are both producers and consumers of a global economy that can only exist through our ability to move further, faster, more frequently, and more safely than ever before. Development, in its broadest sense, has been associated with increased movements of people and the emergence of different forms of population movement, which complicate any simple approach to migration and development.

United Nations estimates placed the number of international migrants in the world at 232 million in 2013, around 3% of the population (United Nations, 2013). While the absolute numbers of migrants had increased from 154 million in 1990 and 175 million in 2000, the proportion of the world's population that was defined as migrant under United Nations definitions had remained around the 3% mark. One can argue that the increased numbers of people migrating and moving are simply the result of an increase in global population numbers from 5.3 billion in 1990 to 7 billion in 2012.

Although the number of migrants in the world is a function of the size of the population, the resultant numbers of movers are not a simple outcome of that expansion. Nevertheless, this introduces into the equation those aspects of development in agriculture, industry, public health, housing, medicine, and transport that have contributed to the decline in mortality, the modern rise in population numbers, and the subsequent decline in fertility. How all these variables, and more, interact to generate migration and development outcomes creates a complexity that is captured in the idea of a 'nexus' as in 'the

migration-development nexus' in which all things are connected (Faist et al., 2011; Nyberg-Sørensen et al., 2002; Mossin Brønden, 2012; Phillips, 2011). Migration and development, however defined, are two processes that vary across space and through time, and it seems relevant to commence the discussion with approaches that attempt to conceptualize the complexity from this point of view.

Mobility Transitions

The geographer Wilbur Zelinsky was the first person to attempt to link migration with the changing patterns of mortality and fertility that can be described in the model of the demographic transition in his article on the 'hypothesis of the mobility transition' (1971). He expressed this most succinctly as follows:

> There are definite patterned regularities in the growth of personal mobility through space-time during recent history, and these regularities comprise an essential component of the modernization process.
> (Zelinsky, 1971: pp. 221–222)

If we substitute 'development' for 'modernization,' which was a term of its age, we have a very clear statement of how migration is integrated into development. Certainly, Zelinsky's hypothesis was not the first time that migration and development had been linked. From the time of Ravenstein in the late nineteenth century, economic factors have been seen as among the major 'drivers' of migration but Zelinsky's contribution was to show how migration changed over time and in sequence with changes in mortality and fertility, and to bring migration into the demographic transition.

Zelinsky identified five types of migration: international; frontierward; rural–urban; urban–urban and intraurban; and circulation and showed how their importance changed through the five stages of the demographic transition from high fertility and mortality to low fertility and mortality. Zelinsky never attempted to show how the changes in the various types of migration were linked to specific dimensions of development such as business cycles, global shifts in industrial production or transportation systems, let alone changes in political development. The links with technology perhaps became most explicit in the final two panels of **Figure 1**(f) and 1(g). In panel (f), he envisaged potential migration being absorbed by forms of circulation as commuter transportation systems permitted people to circulate over longer distances rather than having to migrate more definitively to find a job. Perhaps more speculatively, he also hypothesized that the development of telecommunication systems, including video conferencing, might suppress circulation and, presumably, migration, in a super-advanced society.

The hypothesis of the mobility transition was a grand theory, macrolevel model as populations respond to development. Although based on the historical experience of the Americas, it had undoubtedly broader relevance and for the first time attempted to integrate the three demographic variables as they responded to development. The model has been modified, particularly where applied to specific cases in both developed and developing worlds (Skeldon, 1990) and in terms of global spatial diffusion (Skeldon, 1997). Although it is easy to be critical of its broad sweep and its unilinear nature, it was in many ways ahead of its time in envisaging the shifts in mobility consequent upon developments in technology. It is a true 'classic' in human geography (Woods et al., 1993). To bring a narrower focus to the academic debate on migration and development, the consequences of the flows for origin and destination need to be examined.

Of Destinations, Origins, and Flows

Migration clearly involves an origin and a destination and flows of population between them. The consequences of the migration for the development of origins and of destinations can be quite different. However, to conceptualize migration as a simple flow from an origin to a destination is deceptive as it obscures many return, circular, and onward movements. Attention must therefore focus on the nature of the linkages between origin and destinations, but first a brief consideration of each is necessary.

Destinations

Migrants are seen to be a loss to their area or country of origin and a gain for their country of destination in more than just numbers. Globally, destinations are highly concentrated and, in 2013, the top 10 countries of destination accounted for 51% of global migrants, with the United States hosting fully 20% of the migrants (United Nations, 2013). Only a few of these countries, including the United States, see migration as part of nation building, but all benefit from the addition of mostly young and often relatively well-educated migrants. It is generally considered that migrants are a gain and make a positive contribution to their destinations. Most are young adults and they generate more wealth through work and tax than they consume through welfare and services. The destination state has not paid for their education and many migrants return or retire to their origin countries relieving the destination state of the costs of their care in old age.

However, the dilemmas and contradictions of migration and development are also seen in destination areas. The arrival of migrants in the new worlds to Europeans in the Americas and Australasia marginalized indigenous inhabitants, and an assessment of development from their point of view would be very different from that of the immigrant groups. Significant immigration costs can be incurred in the integration of peoples from very different backgrounds, differences that can generate social and political tensions. A substantial literature has built up for destination areas on the processes of assimilation and integration in destinations, as well as on alternative strategies of multiculturalism, which bear only indirectly on the migration and development debate. Nevertheless, despite these costs and an increasing public and political opinion against migration in many developed economies, the contribution of migration to the development of destination areas is seen to be positive by those assessing the evidence. See, for example,

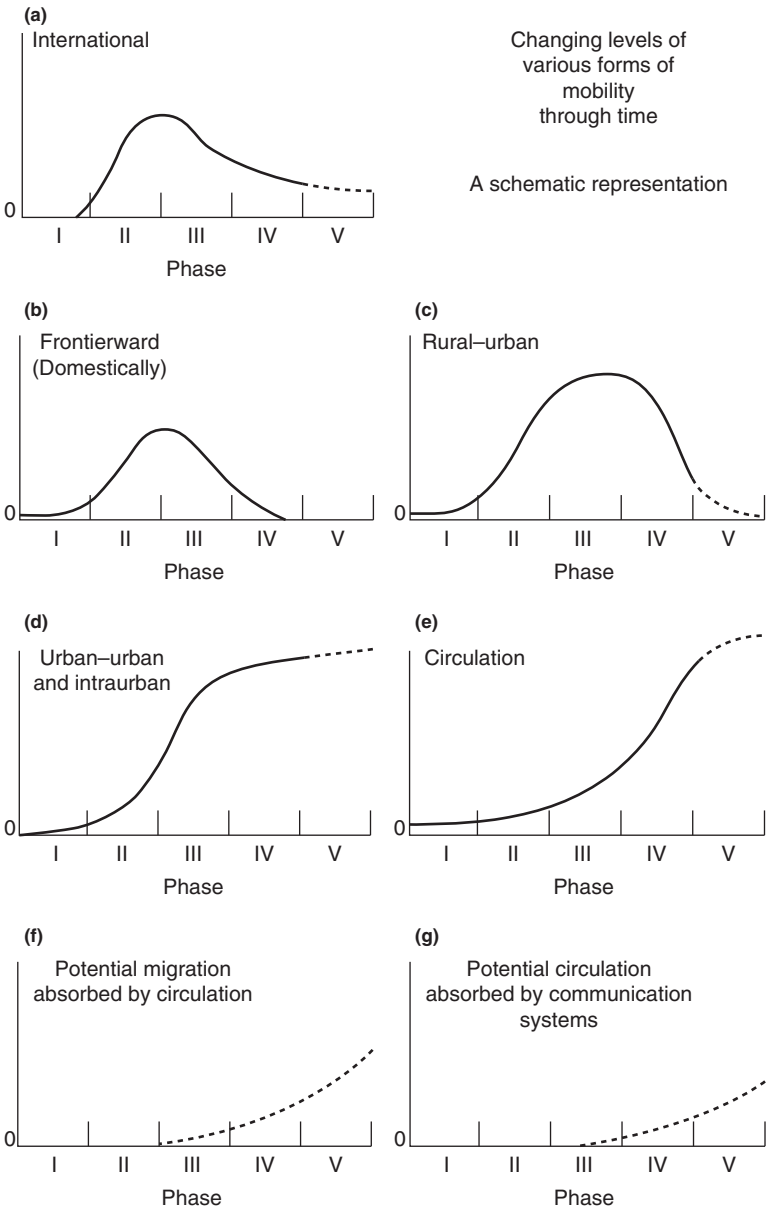

Figure 1 Zelinsky's schematic representation of the changing levels of various forms of mobility through the demographic transition. Source: Zelinsky, W., April 1971. The hypothesis of the mobility transition. Geographical Review, 61 (2), 233. © Elsevier.

Dustmann and Frattini (2013) for the United Kingdom for one recent assessment of the evidence.

Origins

It is the impact of the migration on the origin countries that is more contested and which has received most attention in the migration and development debate. Many consider the loss of migrants to be a brain drain that can prejudice a country's development. However, the analysis of the available evidence has produced a more nuanced picture. Even in the health sector, the data show that those countries with the poorest basic health measures in sub-Saharan Africa were also those with the least emigration of health workers. In addition, attempts to reduce or even stop the emigration of health professionals did not alleviate skill shortages in the countries concerned (Clemens, 2013; Skeldon, 2008). The argument is also made that the exodus of the skilled in particular fields from any country is more likely to encourage other young people to embark upon training in that particular skill, perhaps with a view to future migration. However, not all will want or be able to emigrate upon completing their studies and the country will build up a pool in these particular skill areas with the result being a brain gain (Commander et al., 2004; Mountford, 1997; Stark, 2004). However, the evidence for this eventuality is not conclusive (Lucas, 2005).

More broadly, it is argued that any loss of both skilled and less skilled migrants from a country is compensated by money

sent back as remittances, and it is remittances that lie at the heart of the migration and development debate. The reason is clear: the sheer size of global flow that is estimated to reach US$550 billion in 2013, of which $414 billion will flow to the developing world. It is estimated that this latter component will increase to $540 billion in 2016 (World Bank, 2013). Larger than global aid transfers and larger than even foreign direct investment in several countries, this source of funds is seen as a powerful force for poverty alleviation in particular and for development in general (**Table 1**).

The top recipient countries are China and India, which account for almost one third of the amount flowing to the developing world. Globally, in 2012, it was the middle-income countries that accounted for virtually 70% of the total, with flows to high-income countries accounting for a further 24%. Hence, only 6% of remittances are flowing to the poorest countries, though relatively small volumes can have a large impact. For example, remittances account for 48% of the GDP of Tajikistan, 31% of that of the Kyrgyz Republic, 25% of that of Lesotho and Nepal, and 24% of that of Moldova (World Bank, 2013).

The impact that remittances have will vary by the size of the country, as well as by the number of migrants and where they originate in the countries of origin. Migration is not a random event and areas of origin, like areas of destination, tend to be concentrated in certain parts of a country. Hence, remittances tend to flow back to particular areas, which may not be the poorest. Some 80% of households receiving remittances from international sources in Peru were found to be in the coastal regions of the country (57% in the capital Lima). Only 5% were in the rural sector that accounted for over one quarter of the total population (cited in Skeldon, 2011: p. 109). Remittances are likely to reinforce existing inequalities: the poorest of the poor rarely migrate internationally as they cannot afford to do so, and it seems unlikely that remittances reach the poorest people in the poorest parts of countries. Nevertheless, remittances resulting from internal migration and accounting for a greater number of migrants may reach more areas in the poorer parts of a country. However, these migrants and these remittances are so often excluded from the discussion of migration and development. One notable exception is Adams (2006) who showed that remittances from internal sources did have a slightly greater impact on poverty reduction in Guatemala than those from international migration.

Table 1 Estimates and projections for remittance flows to developing countries

	2009	*2010*	*2011*	*2012e*	*2013f*	*2014f*	*2015f*
	($ billions)						
All developing countries	**316**	**341**	**380**	**401**	**427**	**468**	**515**
East Asia and the Pacific	85	95	106	109	117	130	145
Europe and Central Asia	37	37	41	40	43	47	52
Latin America and the Caribbean	57	58	62	62	67	73	81
Middle East and North Africa	34	41	43	49	52	55	58
South Asia	75	83	97	109	117	127	140
Sub-Saharan Africa	28	29	30	31	33	36	39
World	**436**	**464**	**514**	**529**	**559**	**600**	**665**
Low-income countries	22	24	28	33	36	41	46
Middle-income	294	317	352	368	391	427	469
High-income	120	122	134	128	132	140	149
	(Growth rate, percent)						
All developing countries	**−4.3**	**8.0**	**11.5**	**5.3**	**6.7**	**9.5**	**10.2**
East Asia and the Pacific	1.8	10.9	12.3	2.5	7.1	11.2	11.7
Europe and central Asia	−19.3	−0.1	13.5	−3.9	6.9	10.5	11.4
Latin America and the Caribbean	−11.8	0.9	7.3	0.9	7.1	10.0	10.5
Middle East and North Africa	−6.2	20.9	6.1	14.3	5.1	5.7	6.3
South Asia	4.9	9.8	17.6	12.3	6.9	9.1	10.0
Sub-Saharan Africa	−1.7	4.0	4.9	1.6	5.6	8.6	8.8
World	**−5.9**	**6.2**	**10.9**	**2.8**	**5.7**	**8.7**	**9.4**
Low-income countries	3.9	11.0	17.7	15.6	10.4	13.0	13.4
Middle-income	−4.8	7.8	11.0	4.5	6.3	9.2	9.9
High-income	−10.0	1.6	9.4	−4.2	2.8	6.3	6.7

Note: e, estimate; f, forecast.
Source: Migration and Development Brief 20, World Bank, Washington. http://econ.worldbank.org (accessed 10.11.13.).

A Transnational Turn in Migration and Development

The focus on flows of money and of skills draws attention to the nature of the linkages between origins and destinations and, most importantly, the forms of the migration. Much of the migration is not 'permanent' but involves return to, and circulation between, origin and destinations. Even the 'great migration' across the Atlantic in the second half of the nineteenth century involved significant return in which rates of 50% appear not to have been uncommon (Nugent, 1992). With the modern forms of transportation noted earlier, return on a longer- or shorter-term basis is much easier. However, it is not simply a return movement: it is the ability of migrants to participate in the ways of life in both origin and destination communities, a process that has come to be incorporated under the terms 'transnational communities' and 'transnationalism.'

Transnationalism is part of the whole process of globalization in which large parts of the world are linked not just through the movements of people and flows of money and goods, but also through the constant exchange of information via modern forms of communication, now primarily the Internet. See Vertovec (2009) for a concise introduction to a complex literature. Migrant families, in addition to receiving remittances, may be in constant contact with their relatives in destinations about what to do with those remittances. They can also act as a source of emotional and material support for their relatives in destinations, perhaps to the extent of receiving, in times of need, reverse remittances in cash or kind. Origins and destinations thus become intimately linked in transnational flows that allow parties in both origins and destinations to participate in each other's lives and in the lives of others in both areas, locally and more widely.

The transnational community has been seen as a way to reconcile opposing views of migration and development (Portes, 2009): between those who see migration as positive for development, through remittances and as a way out of unemployment and poverty on the one hand, and those who view migration as negative for development through brain drains, skill losses and the origins becoming subservient to destinations on the other. Nor does successful integration into

destination areas appear to prejudice continued economic, social, and even political involvement in areas of origin. Hence, the transnational community is a form of 'globalization from below' that can promote development through the constant interaction of people, money, and ideas (Portes, 2009: p. 10). Whether this interaction can continue through subsequent generations as migrants and their descendants become progressively more settled in the destination is unclear and will be further addressed below in the context of diaspora engagement. The transnational perspective is also seen to provide a vehicle through which to reconceptualize critically our thinking on migration and development within a global system of power relations (Faist, 2010; Glick Schiller, 2010). Certainly, almost by definition, the analysis of international migration forces the analyst to think beyond the nation state to those transnational institutions that both generate and channel movement in particular ways. However, perhaps more than anything, the transnational perspective has brought a critical perspective to the discussions of migration and development.

Managing Migration for Development

Migration was seen, perhaps still is seen by some, to be brought about by a lack of development. People left because of a lack of opportunity at home. However, if such opportunities could be generated through a development program, then migration would cease. Hence, migration should slow consequent upon development. The experience has proven quite the reverse: migration tends to be positively associated with development. Education programs increase aspirations and agricultural improvement or industrial employment should provide increased income, and so development gives both potential internal and international migrants the information and the means through which to move. The volumes of remittances discussed earlier and revisionist interpretations of skilled migration away from brain drains drew attention to the transformative aspects of migration. Views on migration shifted from seeing migration as something undesirable and negative to accepting that migration could play a positive role in development. The epicenter of this change can be dated to the last decade of the twentieth century, even if many of the ideas originated much earlier. Within the context of this shift, it was but a short step to the idea that migration could be managed to bring about specific positive development outcomes and, consequently, migration and development as a topic has risen toward the top of the policy agenda at the international level.

Two other factors have contributed to the emergence of international migration as a serious policy concern. The first is what might be termed the 'tactical reason.' Migration into a country is considered the concern of the admitting state alone. Hence, governments see little reason to debate an issue for which they feel that they have sole responsibility, or, if pushed, should discuss only bilaterally with origin countries. However, it has become clear that a forum to discuss migration is useful, where countries can meet collectively to share experiences and to deal with sensitive matters such as irregular migration. While migration itself is a sensitive topic that cannot often be discussed overtly, if linked with development it gives the pretext to allow countries to discuss these matters in a somewhat different context. The current composition of delegates to the Global Forum on Migration and Development (GFMD) reflects this bias: officials with responsibility for migration control dominate the government representatives, with very few development specialists in attendance. The GFMD is largely a *de facto* forum on migration rather than a migration and development meeting and until the development community begins to take migration seriously as an issue, it will continue to be so.

The second reason is that very real and pressing matters exist in current global flows of migration. The trafficking and smuggling of people are high on the agenda, made more pertinent by the number of deaths among those attempting to cross the Mediterranean to Europe during the latter half of 2013. The protection of migrant workers is another central concern, an issue of longstanding interest to the International Labour Organization of the United Nations. This work extends into the area of labor recruitment, the often excessive fees that are paid, and the abusive conditions workers face, both during their recruitment and at destinations. The whole issue of migrant rights is central to these concerns. However, fundamentally important though these matters are to migrants, they do not deal centrally with migration and development. Two concerns, however, do fall squarely into this category: remittances and 'leveraging' the diaspora.

Remittances

The global patterns of flows of remittances were discussed above. Some doubts were expressed whether remittances, from international sources at least, reached the poorest people in the poorest parts of a country. While remittances do impact positively upon those who receive them, helping to reduce poverty and improving education and health, that impact may be more modest than might have been anticipated. Considerable heterogeneity among countries exists, with the data from about half of the 11 countries in a large survey in Latin America suggesting no significant impact on poverty (Acosta et al., 2007). Although the impact of remittances on development may still be variable and contested, two points seem clear. First, remittances are private monies that belong to migrants and their families. Governments would be unwise to attempt to tax or guide their use in any way and most migrants make good use of their remittances for education, subsistence, or housing. No clear dividing line between investment and consumption uses seems either practical or meaningful in this regard. Second, the international community can act to reduce transaction costs of sending remittances. Considerable progress in some migrant corridors seems to have been made, but much still remains to be done in the poorest parts of the world in Africa (World Bank, 2013).

Diasporas

In the diaspora, or the community of overseas migrants and their descendants, is to be found a pool of highly skilled and

often wealthy people that can be 'leveraged' (the term of the World Bank) to return on a short or longer-term basis and invest in their homeland. Nevertheless, it is difficult to separate change brought about by returnees from more general change in the origin economies. Migrants will only return home when something exists for which to return. The cases of arguably the more successful diaspora contributions, China and Vietnam, were consequent upon reforms in the investment and political climate at home.

The danger with both diaspora engagement and remittances is that the responsibility for development may be transferred to the migrants and away from the state. Nevertheless, the state involvement has to be various, variable, and tailored to the specific circumstances of the area and country, which should allow space for bottom-up approaches through diaspora engagement. The migration of skilled people back to their home areas to support training or entrepreneurial activities, even if for short periods, seems one avenue that policy makers could pursue with profit, and circular migration has certainly emerged as an area of recent policy interest.

The principal weaknesses of the diasporas as development agents revolve around their highly heterogeneous nature, full of competing interests that make significant and sustained development projects, even at local and regional levels, problematic. While the debate on the effectiveness of unplanned versus planned development continues (Rigg, 2012), it seems unlikely that the diaspora will be a major agent of change in bringing development. That said, it must always be recognized that many of the great leaders of revolutionary change in the twentieth century learned their ideologies and practices in the diaspora. Nevertheless, the jury is still out on whether migration can be effectively managed for development. The pendulum that has seen migration swing from being seen as negative for development to being positive may now be swinging back (de Haas, 2010, 2012).

The International Architecture for the Management of Migration and Development

No United Nations organization exists with primary responsibility for migration and, until 2007, no international forum existed in which to have discussions on the issues at the multilateral level. United Nations agencies exist with primary responsibility for definite types of population movements: the United Nations High Commissioner for Refugees for international forced migrants, and the International Labour Organization for labor migrants. Neither is particularly concerned with migration and development, although remittances are central to issues of labor migration. An international organization, the International Organization for Migration acts as a *de facto* global clearinghouse for issues in migration and development. However, it works with United Nations bodies, particularly the Population Division of the Department of Economic and Social Affairs of the United Nations in New York, the Regional Commissions of the United Nations, and the World Bank, and with a number of supporting civil society organizations, perhaps most particularly the Migration Policy Institute in Washington. The United Nations itself, after widespread consultations, in 2005 and 2006, recommended that the GFMD should be established, and it held its first meeting in 2007. The GFMD is a country-led, non-binding body that meets annually to discuss key areas in migration and development. However, as stressed above, a greater involvement of development experts and policy makers with responsibility for development will be required if this forum is to achieve its mandate.

Conclusion

That migration and development are associated seems incontrovertible. That changes in specific dimensions of what we can consider to be development, such as transportation and education, will bring about shifts in migration is also clear. However, given the sheer complexity of all the variables that impinge on what we call 'development,' migration is but one in this 'nexus' of change. Whether migration can be managed in order to maximize the benefits to development and minimize the risks of moving is not at all so clear. One looming threat is the impact of environmental change on development: whether rising populations and their carbon emissions will produce such catastrophic change that civilization as we know it is unlikely to survive. Totally different viewpoints still exist even about the least uncertain of the many unknowns facing future development, global population growth. Contrasting Emmott's pessimism (2013) with Dorling's optimism (2013), for example, shows the depth to which opinions on this topic of population–environment linkages are still divided some 200 years after the writings of Thomas Malthus.

How future global development in the face of environmental change will impact upon migration is a much greater unknown. Despite predictions that hundreds of million of people will be displaced by rising sea levels, desertification, salinization, extreme climatic events, lowered agricultural production, and wars from increasing competition for scarce resources, the most objective assessment to date is less alarmist (Foresight, 2011). The report stresses the difficulty of separating an 'environmental' driver from those other economic, social, and political factors that cause people to move. While we may expect increased numbers of people to be displaced, the majority will be more likely to move locally and over the short term rather than globally and permanently. The plight of those left behind will be as great a development challenge as the number of those displaced. The largest migration in history (Saunders, 2010), the movement to urban areas, which is driven primarily by factors other than environmental change often brings people into environmentally more hazardous areas in coastal zones around the great megacities of the world. However, the development challenges of these cities have not been a central concern in the migration and development debate.

The introduction of displaced people into the discussion leads to one of the weakest areas in the migration and development debate, the issue of migration and political development, how population movements shift over the long term as a consequence of the evolution of the state and its institutions. An important subset of this migration is forced migration for

political reasons. Refugees were not part of the Zelinsky, longue durée, hypothesis with which this discussion began. Yet, recent scholarship has begun to examine the impact that changing age structures, and particularly the existence of a 'youth bulge,' have on conflict and the emergence of liberal democracies (Cincotta, 2009; Goldstone et al., 2012). The extent to which migrants play a role in disseminating competing ideologies transnationally and between urban and rural has yet to be fully elucidated. Clearly, improving education levels that change aspirations, as well as improvements in gender relations, and the diffusion of modern forms of communication are also fundamental to the mix, but the movements of people are both important causes and consequences of conflict. A search for linkages with the broader demographic transition may prove fruitful.

Finally, the framework in which migration and development tends to be viewed is problematic. It sees development as essentially flat in a simple bipolar division of the world into 'North' and 'South,' as if a simple division between developed and developing exists. Not just the BRICS (Brazil, Russia, India, China, and South Africa) and their impact in their respective regions, but also the emergence of a range of other economies in Africa, Asia, and Latin America have generated a highly heterogeneous global economy. The BRICS themselves cover very different economic and political systems, as does a blanket category such the 'North.' In terms of migration, for example, significant differences exist between the settler societies of North America and Australasia and the longer-established developed economies of Europe. Until both academics and policy makers move beyond simple categories such as South–South migration, the real linkages between migration and development, both theoretical and policy, will prove elusive.

Given that migration is an integral part of development itself, it seems that the starting point needs to be development and how migration is likely to vary as development and development policy change. This is most likely to be context-specific, within region and even within country. Nevertheless, the great macrolevel shifts in mortality and fertility, as well as the transition to an urban society, will provide the common themes around which the variation revolves. Once the development can be mapped, the migration can be anticipated and plans made accordingly. Thus, the emphasis shifts from migration and development to development and migration (also DeWind and Ergun, 2013).

Appendix A. Supplementary Data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/B978-0-08-097086-8.32185-7.

See also: Adoption, Demography of; Democratic Transitions; Development and Urbanization; Diaspora; Global Migration; Globalization and Human Trafficking; Immigration: Political Aspects; Internal Migration: Developing Countries; International Migration and Remittances; International Migration; Migration, Economics of; Migration, Theory of; Migration: Cultural Aspects; Population Geography; Refugees: Geographical Aspects.

Bibliography

Articles preceded by * are generic readings for the topic.

Acosta, P., Fajnzylber, P., Humberto Lopez, J., 2007. The impact of remittances on poverty and human capital: evidence from Latin American household surveys. In: Özden, Ç., Schiff, M. (Eds.), International Migration, Economic Development and Policy. World Bank, Washington DC, pp. 59–98.

Adams, R.H., 2006. Remittances, poverty, and investment in Guatemala. In: Özden, Ç., Schiff, M. (Eds.), International Migration, Remittances and the Brain Drain. World Bank, Washington DC, pp. 53–80.

Bailyn, B., 1987. Voyagers to the West: A Passage in the Peopling of America on the Eve of the Revolution. Knopf, New York.

Cincotta, R.P., 2009. Half a Chance: Youth Bulges and Transitions to Liberal Democracy, Environmental Change and Security Program Report 3, Washington. http://www.wilsoncenter.org/publication/half-chance-youth-bulges-and-transitions-to-liberal-democracy (accessed 10.11.13.).

Clemens, M., 3 September 2013. What Do We Know about Skilled Migration and Development? MPI Policy Brief, Washington. http://www.migrationpolicy.org/pubs/SkilledMigration-Development.pdf (accessed 10.11.13.).

Commander, S., Kangasniemi, M., Winters, L.A., 2004. The brain drain: curse or boon? A survey of the literature. In: Baldwin, R., Winters, L.A. (Eds.), Challenges to Globalization, NBER. University of Chicago Press, Chicago.

* de Haas, H., 2010. Migration and development: a theoretical perspective. International Migration Review 44 (1), 227–264.

de Haas, H., 2012. The migration and development pendulum: a critical review on research and policy. International Migration 50 (3), 8–25.

* DeWind, J., Ergun, D., 2013. Development and migration: historical trends and future research. In: Cortina, J., Ochoa-Reza, E. (Eds.), New Perspectives on International Migration and Development. Columbia University Press, New York, pp. 5–42.

Dorling, D., 2013. Population 10 Billion: The Coming Demographic Crisis and How to Survive It. Constable, London.

Dustmann, C., Frattini, T., 2013. The Fiscal Effects of Immigration to the UK. Centre for Research and Analysis of Migration, Department of Economics, University College, London. http://www.cream-migration.org/publ_uploads/CDP_22_13.pdf (accessed 10.11.13.).

Emmott, S., 2013. 10 Billion. Penguin, London.

Faist, T., 2010. Transnationalization and development: toward an alternative agenda. In: Glick Schiller, N., Faist, T. (Eds.), Migration, Development and Transnationalization: A Critical Stance. Berghan, New York, pp. 63–99.

* Faist, T., Fauser, M., Kivisto, P. (Eds.), 2011. The Migration-Development Nexus: A Transnational Perspective. Palgrave Macmillan, Basingstoke.

Foresight, 2011. Migration and Global Environmental Change: Future Challenges and Opportunities. Final Project Report, The Government Office for Science, London.

Glick Schiller, N., 2010. A global perspective on migration and development. In: Glick Schiller, N., Faist, T. (Eds.), Migration, Development and Transnationalization: A Critical Stance. Berghan, New York, pp. 22–62.

Goldstone, J.A., Kaufmann, E.P., Toft, M.D., 2012. Political Demography: How Population Changes are Reshaping International Security and National Politics. Paradigm, Boulder.

IATA, 2012. Airlines to Welcome 3.6 Billion Passengers in 2016. http://www.iata.org/pressroom/pr/Pages/2012-12-06-01.aspx (accessed 04.11.13.).

* Lucas, R.E., 2005. International Migration and Development: Lessons from Low-Income Countries. Elgar, Cheltenham.

* Mossin Brønden, B. (Ed.), 2012. Migration and development buzz: rethinking the migration development nexus and policies, 50 (3) (special issue).

Mountford, A., 1997. Can a brain drain be good for growth in the source economy? Journal of Development Economics 52 (2), 287–303.

Nugent, W., 1992. Crossings: The Great Transatlantic Migrations, 1870–1914. Indiana University Press, Bloomington.

* Nyberg-Sørensen, N., Van Hear, N., Engberg-Pedersen, P., 2002. The migration-development nexus: evidence and policy options. State-of-the art overview. International Migration 40 (5), 3–47.

Portes, A., 2009. Migration and development: reconciling opposite views. Ethnic and Racial Studies 32 (1), 5–22.

Rigg, J., 2012. Unplanned Development: Tracking Change in South-East Asia. Routledge, London.

Saunders, D., 2010. Arrival City: How the Largest Migration in History is Reshaping Our World. Heinemann, London.

Skeldon, R., 1990. Population Mobility in Developing Countries: A Reinterpretation. Belhaven Press, London.

* Skeldon, R., 1997. Migration and Development: A Global Perspective. Longman, London.

Skeldon, R., 2008. International migration as a tool in development policy: a passing phase? Population and Development Review 34 (1), 1–18.

* Skeldon, R., 2011. Reinterpreting migration and development. In: Phillips, N. (Ed.), The Global Political Economy of Migration, International Political Economy Yearbook, vol. 17. Lynne Reiner, Boulder, pp. 103–119.

Stark, O., 2004. Rethinking the brain drain. World Development 32 (1), 15–22.

The global political economy of migration. Part 2. In: * Phillips, N. (Ed.), International Political Economy Yearbook, vol. 17. Lynne Reiner, Boulder, pp. 103–189.

United Nations, September 2013. The Number of International Migrants in the World Reaches 232 Million, Population Facts, 2013/2. http://esa.un.org/unmigration/documents/The_number_of_international_migrants.pdf (accessed 10.11.13.).

Vertovec, S., 2009. Transnationalism. Routledge, London.

Woods, R., Cadwallader, M., Zelinsky, W., 1993. Classics in human geography. Zelinsky 1971: the hypothesis of the mobility transition. Progress in Human Geography 17 (2), 213–219.

World Bank, 2013. Migration and Development Brief 20. Washington. http://econ.worldbank.org (accessed 10.11.13.).

Zelinsky, W., 1971. The hypothesis of the mobility transition. Geographical Review 61 (2), 219–249.

Multimedia

Two videos on the general topic of migration and development given by the author in 2012. The first gives a more comprehensive treatment, while the second appears to be more watched.

Ronald Skeldon, 23 February 2012. Beyond Migration and Development. International Migration Institute, Oxford University. www.imi.ox.ac.uk Multimedia.

Ronald Skeldon, January 27 2012. Where Next for Migration and Development? School of Graduate Studies and UNU Merit, Maastricht University. www.youtube.com/watch?v=gSoQHCypS9w.

The United Nations Trends in International Migrant Stock: Migrants by Destination and Origin provides a database of stocks for 1990, 2000, 2010, and 2013 where users can construct their own tabulations. Available at: http://esa.un.org/unmigration/migrantstocks2013.htm?msdo (accessed 09.11.13.).

A whole series of policy briefs on migration and development, among many other publications on migration and development, is available on the site of the Migration Policy Institute, Washington. http://www.migrationpolicy.org/research/migration_development.php (accessed 10.11.13.).

Migration and Fertility

Emilio A Parrado, Population Studies Center, The University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Until recently, the growth and structure of national populations has primarily been understood in terms of fertility and mortality, with most classic demographic indicators assuming populations to be closed to migration. This assumption is obviously no longer tenable; most developed countries have relatively low mortality and fertility concurrent with relatively high levels of international migration. These trends highlight the need to more systematically incorporate population mobility into studies of reproduction and growth. This article discusses the mechanisms by which mobility and immigration are linked to fertility, assesses various lines of theorizing about these links, and reviews current research literature.

Until recently, the growth and structure of national populations has primarily been understood in terms of fertility and mortality, with most classic demographic indicators assuming populations to be closed to migration. This assumption is obviously no longer tenable; most developed countries have relatively low mortality and fertility concurrent with relatively high levels of international migration (Preston and Wang, 2007; Mitra, 1990; Morgan et al., 1994). These trends highlight the need to more systematically incorporate population mobility into studies of reproduction and growth. International migration can affect population reproduction through two main mechanisms. The first is in- and out-migration itself. Net flows of immigrants directly impact both the size and age structure of receiving populations. Several studies have investigated how to incorporate this effect into classic demographic expressions. A general elaboration was provided by Preston and Coale (1982) who showed that in the steady-state models that form the core of formal demography, the assumption of a closed population could be relaxed by including age-specific migration rates in addition to age-specific fertility and mortality rates. The overarching insight was the recognition that age-specific migration rates affect cohort size in a manner analogous to the effect of age-specific mortality rates, and thus can easily be incorporated into demographic analyses. Several variations of this general approach have been developed in the literature (Espenshade et al., 1982; Espenshade, 1986; Blanchet, 1989; Cerone, 1987; Mitra, 1983; Preston and Wang, 2007; Schmertmann, 1992).

The second mechanism through which migration can affect population is via the fertility of immigrants at destination (Bledsoe et al., 2007; Jonsson and Rendall, 2004; Parrado, 2011; Parrado and Flippen, 2012; Toulemon, 2004). Despite its importance, the connection between immigration and fertility has received far less attention and our understanding of the interaction remains limited and imprecise. This is problematic because births to immigrant women can be an important component of population reproduction, especially if the fertility of immigrants differs significantly from that of natives. In an early discussion of alternative paths to a stationary population in the United States, Coale (1972) highlighted the interaction between fertility and immigration in producing a constant number of annual births. However, the need to consider the fertility of immigrants when assessing current and future population dynamics is generally not recognized in widely distributed population reports, including most population projections.

In addition to the importance of births to immigrant women for population growth, a main focus of this article is elaborating on the extent to which fertility and migration are linked. Most prior analyses assume the two processes are independent. However, a handful of studies in Europe (Andersson, 2004; Toulemon, 2004) and among Hispanic and Mexican women in the United States (Lindstrom and Giorguli Saucedo, 2007) have documented that fertility patterns among immigrant women are shaped by the experience of migration itself. In particular, migration alters the age pattern of childbearing in ways that are complicate common period measures of fertility. Among immigrant Hispanic women, for instance, fertility is often depressed in the years immediately prior to migration and accelerated in the years immediately following migration *irrespective of age* (Parrado, 2011; Parrado and Flippen, 2012). The main insight is that for immigrant women *two time dimensions* (rather than age alone) define childbearing trajectories at destination: age at migration and time at destination. Age at migration initiates the exposure period for childbearing in the foreign country, and time abroad defines childbearing trajectories over time. Standard demographic measures based on nonmigrant populations do not account for the association between migration and childbearing. As a result, they produce systematically biased estimates of fertility levels (Bledsoe et al., 2007; Frank and Heuveline, 2005; Lichter et al., 2012; Parrado and Morgan, 2008). These biases directly affect our understanding of immigrant fertility and its contribution to nations' total population growth. Moreover, given growing immigrant representation, they affect estimates of fertility levels for the receiving country population as a whole and especially for specific racial/ethnic groups – fueling in some cases racial and ethnic stereotypes.

In this article, I elaborate the substantive and methodological implications of the connection between immigration and fertility and empirically illustrate its importance in the U.S. case, including the implications for understanding changes in the U.S. fertility rates after the 2007 recession.

The Link between Immigration and Fertility

A concrete examination of the two time dimensions linking migration and fertility, namely age at migration and time in

the United States, and their effect on fertility measures highlights the problem. To be precise, fertility level in this article refers to the number of children women have, also referred to as *quantum*, as distinct from *tempo* that refers to the age pattern of childbearing. The most common measure for assessing the fertility level of a population at a point in time is the period Total Fertility Rate (TFR), which is an *estimate* of the number of children a woman would have *if* she were to experience the age-specific fertility rates (ASFRs) observed in a particular period and survive from birth through the end of her reproductive life. The ASFRs describe the population's age pattern of childbearing. They are obtained by dividing the number of births to women of a certain age by the total number of years at risk for this population. These ASFRs are then aggregated to calculate the TFR.

Formally, the TFR for year t can be written as:

$$TFR(t) = \sum_x f(x,t) = \sum_x (B(x,t)/N(x,t)) \qquad [1]$$

where $B(x,t)$ equals the number of births to women aged x at time t and $N(x,t)$ is the number of women aged x at time t. The index of summation x ranges over all childbearing years. It is a *hypothetical* measure based on rates, and does not correspond to the actual behavior of any identifiable group or cohort. Rather, it relies on the assumption that the age-specific rates in a given period will remain steady, and thus approximate the lifetime trajectories of actual birth cohorts. An alternative, and arguably a more accurate, measure of fertility level (Ryder, 1980) is the average number of children ever born (CEB) at the end of a woman's reproductive period, usually after ages 40–44. CEB summarizes actual rather than hypothetical childbearing and is an exact representation of the number of children that actual cohorts of women have borne. The main limitation of CEB is that it reflects the experience of past cohorts and thus provides little information about fertility levels among younger women or at a point in time. In part for this reason, the TFR has become the most common measure used for describing the fertility level of a population.

However, it is clear that meaningful and precise ASFRs are critical for both obtaining a TFR that correctly measures fertility levels and accurate population projections, since the expected number of births (the new cohort) is obtained by first multiplying the size of the female population at given ages by the estimated ASFRs and then summing these products. It has long been recognized, however, that when childbearing ages are in flux, ASFRs are a poor approximation of actual fertility levels understood in terms of number of children or family size. For instance, in discussing the implications of the widespread use of fertility control for period fertility measures, Hajnal (1947) concluded that when women can time their fertility "a change in the rate at which people are having children in a given year can no longer be taken as an indication of a change in the number of children they will bear altogether in the course of their reproductive lives" (p. 143). Some of the most important work on the American baby boom (e.g., Ryder, 1980) and on lowest-low fertility (e.g., Bongaarts and Feeney, 1998; Kohler and Ortega, 2002; Schoen, 2004) addresses this distinction between 'quantum' and 'tempo.'

The direct application of fertility measures, such as the TFR, to immigrant women ignores the potential impact of characteristics unique to immigrant populations on period fertility estimates. Immigrants' life-course transitions are divided between sending and receiving countries. Depending on the age at migration, their childbearing can take place completely at origin, completely at destination, or in a mix of both. Thus, as a group, their completed fertility will almost invariably be different from their fertility contribution to host societies, since the latter includes only the children they have in the receiving context. Moreover, the age pattern of childbearing is often powerfully shaped by the event of migration itself. For hypothetical measures such as the TFR to be a meaningful indicator of family size, this implies that they will have to be reformulated, capturing cohorts' experiences of both migration and childbearing. Moreover, if there is a close association between the event of migration and childbearing, we can expect potentially large distortions in ASFRs, which would result in inaccurate assessments of immigrant fertility.

A common situation documented in various migration streams is the case where childbearing tends to occur shortly after migration. **Figure 1** simulates this effect by plotting ASFRs (dotted curve) and fertility rates since time of migration (solid curves). The x-axis corresponds to women's age and the y-axis to fertility rates. The dotted lines crossing the x-axis correspond to different ages at migration. Thus, the figure captures the two time dimensions discussed above: age at migration and time at destination, and their association with fertility. The case illustrated in **Figure 1** corresponds to the situation where ASFRs peak during the years shortly after settlement (vertical dotted lines) and then rapidly decline with time at destination (the different solid curves). Since this pattern is reproduced *irrespective of age at migration*, failure to control for the interaction between migration and fertility inflates ASFRs (dotted curve) that remain artificially high throughout a large portion of reproductive ages. However, rather than the product of fertility behavior, high ASFRs are the result of the continued influx of immigrants and the concomitant higher fertility rates in the years after migration. For assessing the level of immigrant fertility, a main implication is that no birth cohort will reproduce the observed ASFRs. They are only applicable to migrant cohorts as they age at destination. For the dynamics of population reproduction, the main implication is that immigration and fertility rates are not independent; a change in immigration will produce a change

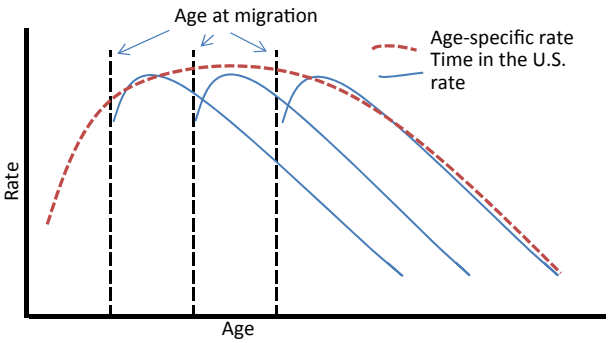

Figure 1 Simulation of fertility rates by age at migration and by time in the United States.

in fertility rates, and demographic techniques should be able to separate those two effects.

This phenomenon – a rise in birth rates following a major event – in many ways parallels for the experience of the baby boom. When World War II ended, young women faced fewer constraints to childbearing and began their families earlier, but older women's fertility also rose, as they bore children that had been postponed during the conflict. Thus, while fertility rates rose substantially, period estimates, which cannot distinguish between timing and quantum changes in fertility, badly overestimated the increase relative to cohort measures (see Ryder, 1980, 1983, 1986). As applied to the immigrant case, the pattern has been corroborated among Hispanic, especially Mexican, immigrants in the United States and among immigrants in Europe (Parrado, 2011; Parrado and Morgan, 2008; Toulemon, 2004).

It is important to note, however, that while the association is common to the general immigrant experience rather than just the experience of one particular ethnic group, the magnitude of the association as well as its effect on fertility estimates is likely to vary across contexts and groups. Assessing potential variation is central to discussions about racial, ethnic, and national origin differences in fertility behavior. However, it is not our intention to explain this variation. Such exercise requires paying attention to migration dynamics across the life-course, the socioeconomic composition of immigrant flows, and issues of immigrant incorporation. Our objective is more limited: to document and elaborate on methodologies that can account for the association between migration and fertility and assess its significance in the U.S. context. To the extent that the association is general to immigrants rather than specific to particular ethnic or national origin groups, then there are reasons to focus on the immigrant component of population growth irrespective of ethnic or of national origin in demographic analyses. At the same time, if national origin differences are salient, fertility will be affected by both the changing composition of the immigrant flow as well as changes in the overall number of immigrants.

Fertility at Origin and Destination

To more precisely separate population growth that stem from immigration per se from the fertility contribution of immigrant women, it is important to start with the recognition that the immigrant fertility takes place both at origin and destination. Measures such as CEB capture completed fertility but provide no information about where childbearing takes place. However, from a receiving country perspective, the fertility-specific contribution of immigration refers solely to the children of immigrant women born at destination. Children born abroad would also contribute to the receiving country population if they move with their parents. However, they contribute through immigration rates, not through fertility. For some analyses, such as discussions about fertility decline, the distinction might not be relevant, since researchers might be interested in focusing on the total number of children immigrant women have. From a population perspective, though, the distinction is important since it refers to the sources of reproduction. For receiving countries, the fertility component of population growth refers only to children born at destination. Technically, the decomposition can be represented in the following way:

$$\mathrm{CEB}_{(\mathrm{lifetime})} = \mathrm{CEB}_{(\mathrm{at\ origin})} + \mathrm{CEB}_{(\mathrm{at\ destination})} \qquad [2]$$

Thus, we are interested in obtaining precise estimates of CEB at destination. The decomposition would be straight forward if we had information on children's place of birth for women who have completed their reproduction, ages 40–44. However, obtaining period estimates of children born at destination requires taking into consideration the two time dimensions discussed above, namely age at migration and time in the receiving country. The destination-specific total fertility rate (D-TFR) is a modified version of the standard TFR that takes into account age at migration and time in the United States to avoid the problems arising from migration-related disruptions in the age pattern of childbearing (Toulemon, 2004; Parrado and Flippen, 2012). The D-TFR is computed by first calculating duration at destination-specific fertility rates (DD-FRs) in a particular year for women migrating at a certain age. DD-FRs are obtained by dividing the number of births among women with certain duration at destination by the total number of women with the same duration in a particular age-at-migration cohort. Specifically.

$$\text{DD-FR}(x,a,t) = B(x,a,t)/N(x,a,t) \qquad [3]$$

where $B(x,a,t)$ equals the number of births to women with time x at destination, age a at migration, and in year t. $N(x,a,t)$ is the number of women with time x at destination, age a at migration, and in year t. Aggregating the DD-FRs over time at destination (x) for specific migration cohorts (a) in a given year (t) produces an estimate of the accumulated fertility over time at destination of particular age-at-migration cohorts. Once the cohort reaches ages 40–44, which corresponds to different durations in the receiving country depending on age at migration, then the cumulative DD-FRs becomes an estimate of the total number of children women migrating at certain ages will bare at destination if they are subject to the DD-FRs observed for migrating cohorts in a particular period. The actual childbearing experience of age-at-migration cohorts is again given by the completed fertility rate ($\mathrm{CEB}(x,a,t)$) that measures the average number of children after x time at destination that women migrating at a certain age a actually had at time t.

Aggregating across migrating cohorts (a) and weighting by their proportional representation in the pool of immigrant women yields the D-TFR(t), which provides an estimate of the number of children that all immigrant women will have at destination if they are subject to the average U.S. DD-FRs and age at migration rates observed at a particular point in time. This can be written as,

$$\text{D-TFR}(t) = \sum_a \sum_x \text{DD} - \text{FR}(x,a,t) \times p(x,a,t) \qquad [4]$$

where $p(x,a,t)$ equals the proportion of immigrant women with time x in the United States, age a at migration, and in year t. D-TFR(t) is then a period estimate of the number of children immigrant women would have at destination given prevailing patterns of migration and childbearing, which corresponds to $\mathrm{CEB}_{(\mathrm{at\ destination})}$ in eqn. [2]. In the next section, we apply these measures to the U.S. context.

Immigrant Fertility in the United States

The United States is an apt setting for illustrating the interaction between migration and fertility. Drawing on multiple data sources, we can calculate the measures presented for immigrants overall, as well as by race/ethnicity and country of origin. Our analysis pays particular attention to the Latino experience since the purportedly abnormally high fertility of Latina women is often raised in discussions of socioeconomic incorporation and a central dimension undergirding negative stereotypes of the group (Chavez, 2012; Parrado and Flippen, 2012).

Table 1 pools data between 2003 and 2004 American Community Surveys to report the distribution of women of childbearing ages 15–44 and birth rates by race, ethnicity, nativity, and country of origin. Results highlight the magnitude of the fertility contribution of immigrants relative to natives as well as the significance of immigrant fertility for monitoring the demographic profile of the U.S. population. For the population as a whole, results show that 78.7% of births in the United States are attributable to native women, which is 4.3% points lower than their representation in the total population (83.1%). Alternatively, almost a quarter (21.3%) of U.S. births is attributable to immigrant women, even though they represent only 16.9% of the female population in childbearing ages.

The pattern is also evidenced, with variation, when we separate women and births by race/ethnicity. Without distinguishing by nativity, results show that while non-Hispanic (NH) white women represent 65.4% of the U.S. population, they accounted for only 59.9% of births. Thus, NH white women's share of births is 5.5% points lower than their representation in the population as a whole. At the other extreme, Hispanics represented 14.8% of the U.S. population but 19.7% of all births (difference = 4.9). In fact, more births relative to population are evidenced across all minority groups, illustrating the contribution of fertility to the growing diversity of the U.S. population.

Separating women and births by nativity more precisely assesses the sources of these patterns. Natives overall are significantly underrepresented in births (difference = −4.3). This is particularly the case for NH whites but also holds for Asians. More importantly, the greater representation of births to Hispanic and black women is considerably lower among natives. Among Hispanics, the overall difference is reduced from 4.9% overall to 1.3% among natives. The most salient result from Table 1 is that most of the growing diversity of U.S. births stems from the fertility contribution of immigrant women. The higher contribution is particularly pronounced among Hispanics, but is evident across all ethnoracial groups. Result show that even though foreign born Hispanic women comprise 8.1% of the U.S. female population of reproductive ages, they accounted form 11.6% of births. However, while percent differences are much lower, even the fertility contribution of immigrant NH white women is higher than their share of the total population, and the same can be said of NH black and Asian women.

To further highlight its importance, the last column of Table 1 reports the within-group odds ratios of the likelihood of a birth occurring to an immigrant compared to a native woman. To illustrate, in the case of Hispanics, this is computed as the percent of births to immigrant women (11.6) divided by the percent of immigrant women (8.1) divided by the percent of births to native women (8.1) divided by the percent of native women (6.8). Odds ratios also account for the bounded nature of percentages and their limitations for group comparisons. Results from this calculation show that the fertility contribution of immigrant women is much higher than natives in the population as a whole and for all groups. Overall, results show that a child born in the United States today has a 32% higher chance of being born to an immigrant than U.S. born woman (odds ratio 1.32). Among NH whites, the estimated odds ratio of 1.13 implies that a white child has a 13% higher likelihood of being born to an immigrant than a native woman. The

Table 1 Distribution of 15–44-year-old women and births by race/ethnicity, country of origin, and nativity in the United States

	All			*Native*			*Foreign-born*			
	Population	*Births*	*Difference*	*Population*	*Births*	*Difference*	*Population*	*Births*	*Difference*	*Odds ratios*
All women				83.1	78.7	−4.3	16.9	21.3	4.3	**1.32**
Race/ethnicity										
White, non-Hispanic	65.4	59.9	−5.5	62.0	56.4	−5.6	3.4	3.4	0.1	**1.13**
Black, non-Hispanic	13.6	14.1	0.5	12.3	12.4	0.2	1.4	1.7	0.3	**1.22**
American Indian	0.9	1.0	0.1	0.8	0.9	0.1				
Asian/Pacific Islander	5.2	5.3	0.1	1.2	0.8	−0.3	4.1	4.4	0.4	**1.51**
Hispanic	14.8	19.7	4.9	6.8	8.1	1.3	8.1	11.6	3.5	**1.20**
Country of origin										
Mexican	7.6	10.2	2.6	4.0	4.5	0.5	3.6	5.7	2.1	**1.42**
Cuban	0.4	0.3	−0.1	0.2	0.2	−0.1	0.2	0.1	0.0	**1.27**
Dominican	0.3	0.3	0.0	0.1	0.1	0.0	0.2	0.2	0.0	**1.10**
Chinese	1.1	0.9	−0.2	0.3	0.1	−0.2	0.7	0.8	0.1	**2.36**
Korean	0.6	0.4	−0.1	0.1	0.1	−0.1	0.4	0.4	0.0	**2.24**
Filipino	1.1	1.0	−0.1	0.4	0.3	−0.1	0.6	0.7	0.0	**1.21**
N (in thousands)	60 364	4124		50 143	3247		10 221	877		

American Community Survey 2003–04.

likelihood of immigrant origin birth is highest among Asians (1.51) and is 1.20 among Hispanics. The main implication is that the growth of the U.S. population overall and the relative growth of ethnoracial groups is powerfully shaped by the fertility of immigrants.

It is important to consider, however, that pan-ethnic and racial categories mask considerable variation in national origins among the immigrant population. Focusing on the largest immigrant groups, the bottom panel of Table 1 replicates the preceding calculations for Mexican, Cuban, Dominican, Chinese, Korean, and Filipino origin populations. Results further document the salience of immigrants for fertility within national origin groups. In most cases, the share of births born to immigrant women is higher than their share of the total U.S. population, particularly for Mexicans. The percent of births to Mexican immigrant women is 2.1% points higher than their representation in the population. However, it is also higher for Chinese (0.1). The odds ratio estimates again highlight the salient contribution of immigrants for within-group fertility. In all cases, within national origin groups, U.S. births are more likely to occur to immigrant than native women and group differences do not always follow estimates obtained from difference in percentage comparisons.

As elaborated in the theoretical section, it is not obvious that over representation in births, i.e., higher birth rates, among immigrants imply higher fertility levels or larger family sizes. We investigate the distinction for the U.S. case in Table 2. Pooling data from the June supplements of the Current Population Survey for the years 2000–08, which contain information on births during the prior year as well as total number of CEB, the table compares estimated fertility levels obtained from the standard TFR measure and CEB among women aged 40–44 by native and race/ethnicity. Despite the differences between the two measures, when we focus on U.S. born women, the period TFR and cohort CEB estimates are largely consistent; for all native women, the TFR is only 0.1 children higher than CEB. The consistency in estimates across measures is also evident when we separate women by race/ethnicity. In all cases, the difference between period and cohort estimates is negligible among native born women. Results support the expectation that as a hypothetical measure, the TFR is a good approximation to actual fertility levels for U.S. born women.

Table 2 TFR and CEB by race/ethnicity and nativity

	TFR (1)	*CEB*[a] *(2)*	*Difference TFR − CEB (3)*
Native	2.03	1.88	0.10
NH white	1.90	1.84	0.06
NH black	1.99	1.99	0.00
Asian/Pacific Islander	1.69	1.91	−0.22
Hispanic	2.21	2.13	0.08
Foreign-born	2.59	2.13	0.46
NH white	2.21	1.83	0.38
NH black	2.37	2.03	0.34
Asian/Pacific Islander	2.20	1.85	0.35
Hispanic	2.95	2.49	0.46
All race/ethnic groups	2.11	1.91	0.20
White, non-Hispanic	2.02	1.84	0.19
Black, non-Hispanic	2.04	1.99	0.05
Asian/Pacific Islander	2.00	1.86	0.14
Hispanic	2.58	2.36	0.22

[a]Among women ages 40–44.
June Current Population Survey (CPS) 2000–08.

Results differ dramatically, however, among immigrant women. Overall, the TFR exaggerates the fertility level among the foreign-born by 0.5 children compared to the actual number of CEB. Moreover, there are few differences across racial and ethnic groups. The pattern observed for immigrant Hispanics is also reproduced among NH white, NH black, and Asian immigrants. For instance, among NH whites, the TFR is 0.4 children higher than the actual CEB. Interestingly, comparing the CEB of natives and immigrants within ethnoracial categories shows relatively modest differences for all groups except Hispanics, since they originate in higher fertility countries than their white and Asian counterparts. However, consistent with the expectation of a distortion in period rates resulting from an association between migration and childbearing, the TFR is consistently much higher and the disparity between period and cohort measures more pronounced among the foreign-born of all ethnoracial groups.

An associated problem is that given the growth in immigration levels over the past several decades, the TFR could be distorting fertility estimates both for the U.S. population as a whole and for particular racial/ethnic groups. The bottom panel shows that except for NH blacks the TFR is higher than CEB for all groups, supporting the concern that the failure to account for timing distortions among immigrants is inflating fertility estimates by race/ethnicity. Moreover, even for the United States as a whole, the 2.1 estimated TFR is 0.2 children higher than the 1.9 CEB. To magnify the disparity, the 10% overestimate in the TFR means that *if immigration were to stop immediately*, applying conventionally calculated fertility rates in population projections would overshoot the annual number of births just among U.S. born women by 300 000 every year.

These results illustrate the potential importance of systematically separating immigrants from the native population in studies of fertility. They also suggest that, despite stereotypical representations, the issue is not unique to Hispanics. Table 3 focuses specifically on the Hispanic experience and separates CEB from children born in the United States according to eqn. [4]. The table reports US-FRs (eqn. [3]), US-TFRs (eqn. [4]), CEB, and the difference between the US-TFR and CEB for Hispanic immigrant women obtained from the 2000–2008 CPS according to age at migration (a) and time in the U.S. (x). Results document that US-FRs tend to be higher during the years immediately following migration, especially for women migrating after age 20, and then decline. This is consistent with the expectation of disruption in the timing of childbearing associated with migration. The US-TFRs that accumulate the duration-specific fertility rates provide a period estimate of the number of children born to women migrating at different ages after particular durations in the United States. Results show that after 20 years of U.S. residence, near the end of their reproductive ages, women migrating between ages of 15–19 have had 2.63 children in the United States. This is almost identical to the actual

Table 3 U.S. specific fertility rates and completed fertility rate by age at migration among Latino women

	Age at migration															
	15–19				*20–24*				*25–29*				*30–34*			
Years in U.S.	*US-FR*	*US-TFR*	*CEB*	*Difference*	*US-FR*	*US-TFR*	*CEB*	*Difference*	*US-FR*	*US-TFR*	*CEB*	*Difference*	*US-FR*	*US-TFR*	*CEB*	*Difference*
1–2	0.16	0.33	0.41	0.09	0.18	0.36	0.84	0.48	0.12	0.24	1.25	1.01	0.11	0.23	1.76	1.54
3–4	0.17	0.66	0.79	0.13	0.19	0.75	1.31	0.56	0.15	0.53	1.57	1.04	0.03	0.29	2.14	1.85
5–6	0.22	1.09	1.13	0.04	0.19	1.13	1.51	0.37	0.13	0.79	1.96	1.18	0.08	0.45	2.19	1.74
7–8	0.15	1.40	1.52	0.12	0.13	1.40	1.78	0.38	0.10	0.98	2.04	1.06	0.05	0.54	1.87	1.33
9–10	0.16	1.71	1.81	0.10	0.07	1.54	2.09	0.55	0.08	1.14	2.14	0.99	0.02	0.59	2.31	1.72
11–12	0.17	2.04	2.03	−0.01	0.08	1.71	2.17	0.46	0.03	1.21	2.34	1.13				
13–14	0.13	2.31	2.14	−0.17	0.10	1.90	2.23	0.33	0.03	1.27	2.30	1.03				
15–16	0.09	2.48	2.42	−0.06	0.04	1.99	2.55	0.56								
17–18	0.04	2.57	2.67	0.11												
19–20	0.03	2.63	2.59	−0.04												

	35–39				*40–44*				*40–44*				*All ages*			
Years in U.S.	*US-FR*	*US-TFR*	*CEB*	*Difference*	*US-FR*	*US-TFR*	*CEB*	*Difference*	*US-FR*	*US-TFR*	*CEB*	*Difference*	*US-FR*	*US-TFR*	*CEB*	*Difference*
1–2	0.05	0.09	2.31	2.22	0.04	0.07	2.54	2.47	0.04	0.07	2.54	2.47	0.13	0.27	1.29	1.03
3–4	0.02	0.13	2.12	1.98									0.13	0.54	1.49	0.95
5–6	0.04	0.20	2.63	2.43									0.16	0.85	1.76	0.91
7–8													0.11	1.08	1.94	0.86
9–10													0.10	1.27	2.10	0.83
11–12													0.10	1.47	2.19	0.73
13–14													0.10	1.66	2.21	0.55
15–16													0.07	1.79	2.48	**0.69**
17–18													0.04	1.88	2.67	**0.80**
19–20													0.03	1.94	2.59	**0.65**

June Supplement, Current Population Surveys 2000–08.

number of CEB. Among this migrant cohort, the difference between US-TFR and CEB is very close to zero; as could be expected, the difference does not change with U.S. duration since it represents the number of children born before migration. The main implication is that for young migrating cohorts (15–19) most of their childbearing takes place in the United States.

As the age at migration increases, the difference between US-TFR and CEB also increases. For women migrating between 20 and 24 years of age, results show a US-TFR of 1.99 after 16 years in the United States, which is roughly 0.5 children lower than the CEB. In other words, among women in this migration cohort roughly one-half child was born at origin. The difference for women migrating between 25 and 29 years of age is 1.03. These women average only 1.3 children at destination after 15 years of U.S. residence, even though their completed fertility is 2.3. Later ages at migration result in fewer U.S. born children. At the extreme, women migrating at ages 40–44 bare only 0.07 children in the United States, even though their completed fertility is 2.54 (not reported).

The US-TFR computed by aggregating across migration cohorts, according to time in the United States, provides an estimate of immigrant women's total number of U.S. born children and is reported in the last columns of the bottom panel of **Table 3**. The estimated US-TFR for immigrant Hispanic women after 19–20 years of U.S. residence is 1.94, which is *1 child lower* than the 2.96 estimate obtained applying the standard method of calculating TFR (eqn. [1] and reported in **Table 2**) to this population. Again, we argue that the main source of the disparity is the failure of the TFR to account for the interaction between age at migration and time in the United States in affecting period fertility measures among immigrant women. This overestimation misrepresents the fertility experience of immigrant-origin groups and produces biases in population projections, hence the need for more precise measures. As the Hispanic case demonstrates, the main implication for demographic analysis is that standard measures overestimate the fertility contribution of immigrant women to countries of destination. For discussion of racial and ethnic differentials in fertility behavior, the overestimation misrepresents the experience of groups with sizable immigrant representation, such as Hispanics.

The 2007 Recession and Fertility in the United States

The difficulty in integrating immigration and fertility into explanations of population dynamics in host societies, is especially relevant for discussions of the connection between fertility and economic fluctuations. A recent report by the Pew Research Center (2012) as well as official statistics from the National Center for Health Statistics (Sutton et al., 2011) documented a sharp decline in U.S. fertility rates after the 2007 economic recession (approximately 8%). The finding is generally not surprising since economic recessions have historically been associated with lower fertility. However, the reports show considerable variation across groups in the impact of the recession on fertility. In particular, the Pew Research Center reports that the decline was especially pronounced among immigrant women (14% relative to only 6% among U.S. born women). The effect is even more salient among Mexican immigrants, who saw their fertility rates decline by 23% after the recession.

The dramatic disparities in the apparent impact of the recession across groups garnered significant interest in both popular and academic circles, with a wide array of cultural and structural explanations proffered. Whatever its source, it was widely assumed that a change in behavior was at the heart of falling fertility. For instance, the Pew report emphasized that "the fall in the number of births to women is explained by behavior (falling birth rates), rather than population composition (change in the number of women of childbearing age)" (p. 3). As another example, the *New York Times* reported the decline and difference by group as a surprising change in Latino families. Reinforcing both the erroneous expectation of large Latino families and the confusion between fertility rates and family size, the opening sentence of the news report stated: "Hispanic women in the United States, who have generally had the highest fertility rates in the country, are choosing to have fewer children" (NYT, 2012) and demographers were quoted as saying: "It is surprising…When you hear about a decrease in the birthrate, you don't expect Latinos to be at the forefront of the trend." The problem is not that rates did not decline, but that the attribution of the change to a dramatic reorientation of the high fertility preference of Latino women does not account for the population processes producing the change.

Our theoretical and methodological connection of immigration and fertility provides a different, more demographic account. The fertility level of Latina women in the United States hovers between 2 and 2.5 children. So, desire for smaller families is something well-entrenched in Latino family orientations. Recessions do affect fertility rates, but there are no obvious reasons that they should affect the fertility decisions of Latina women more so than for other groups. In fact, rather than a change in Latino fertility, what specifically and dramatically declined are the fertility rates of all immigrants, including Latinos. Ignored in the Pew and media accounts is the fact that the recession also resulted in a dramatic drop in immigration to the United States. The combination of deteriorated economic climate, significant voluntary and involuntary return migration to Mexico, and border enforcement policies changed migration flows after the recession. In fact, many migration scholars have argued that with the recession net migration from Mexico to the United States is close to zero or even negative (Pew Hispanic Center, 2012). Since, as we argue, immigration and fertility rates are connected, the change in immigration rates is of central importance to understanding fertility rates. More specifically, since immigration coincides with family formation, with a large portion of immigrant fertility taking place soon after arrival, declining immigration rates will also reduce the representation of the group of immigrant women most prone to have children. In other words, the reduction in immigration is affecting the composition of the immigrant population, increasing the representation of long-term residents who are more likely to have completed their fertility, and reducing the contribution of new arrivals who are more likely to have children. The main hypothesis then is that rather than the product of changing ideas about family size, the decline in the fertility rates of immigrants after the recession is primarily the outcome of declining immigration rates.

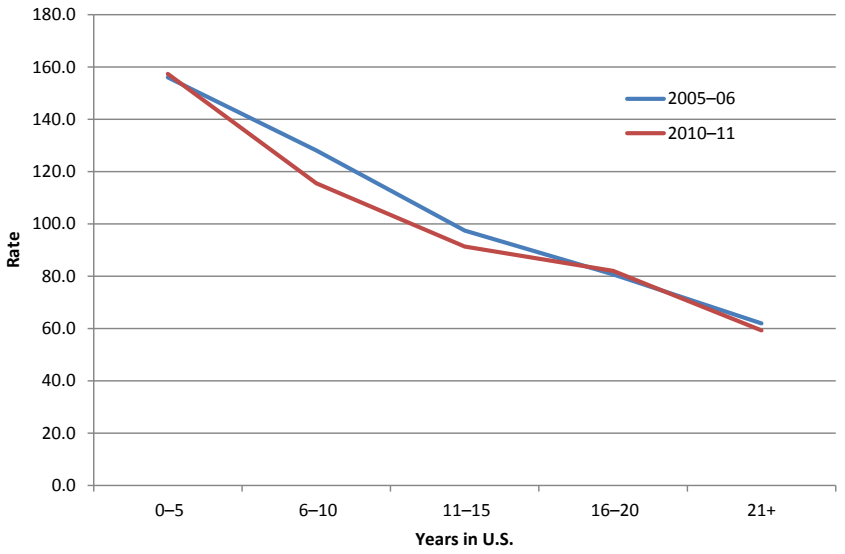

Figure 2 Birth rates for foreign-born Mexican women according to years in the United States.

An empirical examination corroborates this argument. Data from the 2005–06 and 2010–11 American Community Survey confirm the findings of prior reports; the birth rate for immigrant Mexican women in the United States declined 12% after the recession, from 110 to 69.9 children per 1000 women (15–44). Accounting for time in the United States and computing the duration at DD-FRs shows a very different pattern of change in fertility rates. Figure 2 reports DD-FRs for foreign-born Mexican women before and after the recession by U.S. duration. If behavioral changes were responsible for the drop in immigrant fertility rates, then we would expect somewhat parallel DD-FRs curves with much lower rates after the recession irrespective of time in the United States. Instead, the figure shows that the DD-FRs of immigrant women changed little before and after the recession; the relationship between U.S. duration and birth rates remained largely consistent over time. Only among women with 6–10 and 11–15 years in the United States do we see a 10 and 7% decline in fertility, which is comparable to the impact of the recession on native and NH women.

It is worth emphasizing that the lack of obvious change in DD-FRs coincides with a dramatic change in the composition of the immigrant population according to U.S. duration. Figure 3 plots the share of foreign-born Mexican women in the United States according to U.S. duration. The figure illustrates the impact of declining immigration during the recession on the composition of the foreign-born Mexican population. The share of immigrant women with 0–5 years in the United States declined from 23.5 to 13.1% following the recession. There was also a concomitant rise in the

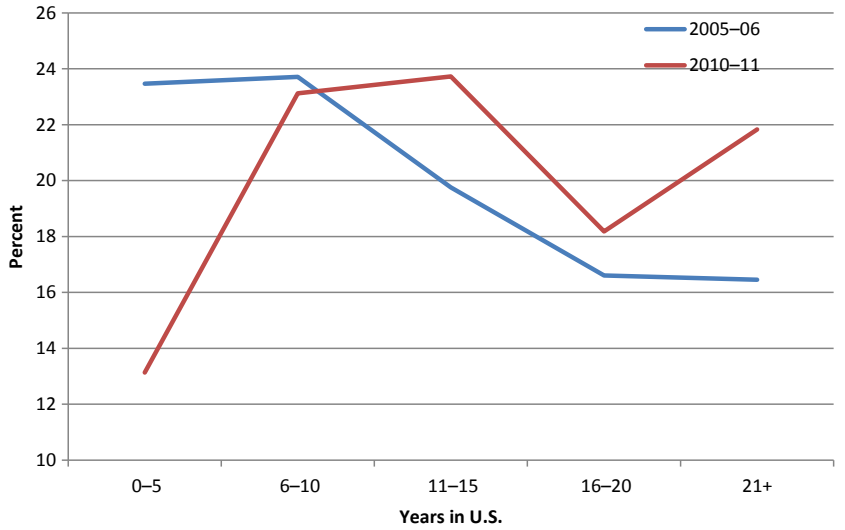

Figure 3 Composition of the foreign-born Mexican population according to years in the United States.

representation of women with more than 10 years of U.S. residence, which rose from 52% before the recession to 64% after 2007. Obviously, since fertility rates tend to be higher soon after U.S. arrival, this change in the composition of the immigrant population has had a dramatic effect on overall fertility rates. It is this change in composition, and not changing family orientations, that explains the lion's share of the drop in immigrant fertility; decomposing the change shows that if the composition of the immigrant flow had not changed in conjunction with the recession, the fertility rate of immigrants would have been only 4% lower in 2010 than in 2006. Again, this estimate is in line with the decline observed among nonimmigrants. Since the overall change was 12%, changing immigrant composition explains more than twice the change in rates than changing behavior.

Conclusions and Implications

In this article, we elaborate on the relation between immigration and fertility, linking two processes that have traditionally been treated as independent population domains in most accounts. The analysis elaborates on alternative fertility measures specifically designed to study immigrant populations that more precisely describe both the level and life-course pattern of immigrant fertility. As applied to the U.S. case, we show that the fertility of immigrants at destination is affected by two time dimensions, age at arrival and time in the U.S. We document that since the timing of migration tends to coincide with family formation, the fertility of immigrants tend to occur shortly after arrival irrespective of age at migration. Failure to account for this association produces much higher estimates of immigrant fertility levels than what are obtained from measures of completed family size.

In addition, we highlight that as a group, immigrants can experience childbearing either at origin, destination, or both. The childbearing contribution of immigrants thus is different whether one considers all children or children only born at destination. For the population dynamics of receiving countries, children born abroad contribute to population growth through immigration; children born at destination more precisely constitute the fertility-specific contribution of immigrants to population growth. The distinction is not generally recognized. Separating the two processes not only presents a different description of the fertility levels of immigrants, but also highlights that their fertility-specific contribution is a lot lower than usually assumed, reducing estimates of how much immigrant fertility can contribute to population growth. We also document that the timing disruption is evident across national origin groups, which makes more of an immigrant than a racial or ethnic trait. The impact on fertility estimates varies by group depending on size, recency, and migration dynamics of immigration.

Two main implications are worth highlighting. The first one is that immigrants should arguably be treated separately from the native born population in demographic analysis. The growing representation of the foreign-born in rich economies implies that the growth and structure of immigrant receiving countries is increasingly affected by immigration. Standard demographic techniques, in our case fertility measures, should not uncritically be extended to immigrants. An initial step in recognizing their differential contribution is to treat them separately in demographic accounts, official reports, and population projections.

The second one is that to the extent that immigration and fertility are related, currently observed fertility patterns are affected by immigration dynamics. When immigration patterns change, then fertility will also be affected. We document this expectation for the U.S. case showing that the decline in immigration explains the larger portion of declining fertility rates after the recession. The connection is particularly relevant for population projections and policy discussions. Accounts of population growth and their age structure into the future should incorporate the notion that variable immigration rates imply variation in fertility rates. In policy discussion, the stand toward promoting or restricting immigration should also take into consideration its effect on fertility.

See also: Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Immigration Policy; Immigration, Applied Research on; Immigration, Emigration, and Citizenship; Immigration: Health and Mortality; Immigration: Political Aspects; Immigration: Social Psychological Aspects; Migration and Development; Migration and Health; Migration, Ethnicity, Aging and Social Work Practice; Migration: Cultural Aspects; Migration: Motivations.

Bibliography

Andersson, G., 2004. Childbearing after migration: fertility patterns of foreign-born women in Sweden. International Migration Review 38 (2), 747–775.

Bean, Frank D., Swicegood, C.G., Berg, R., 2000. Mexican-origin fertility: new patterns and interpretations. Social Science Quarterly 81 (1), 404–420.

Blanchet, D., 1989. Regulating the age structure of a population through migration. Population: An English Selection 44 (1), 23–37.

Bledsoe, C., Houle, R., Sow, P., 2007. High fertility Gambians in low fertility Spain: the dynamics of child accumulation across transnational space. Demographic Research 16 (12), 375–412.

Bongaarts, J., Feeney, G., 1998. On the quantum and tempo of fertility. Population and Development Review 24, 271–291.

Bouvier, L.F., Poston Jr., D.L., Zhai, N.B., 1997. Population growth impacts at zero net international migration. International Migration Review 31 (2), 294–311.

Carter, Marion, 2000. Fertility of Mexican immigrant women in the U.S.: a closer look. Social Science Quarterly 81, 1073–1086.

Census Bureau, 2008. U.S. Population Projections. http://www.census.gov/population/www/projections/methodstatement.html (accessed 23.10.09.).

Cerone, P., 1987. On stable population theory with immigration. Demography 24, 431–438.

Chavez, Leo, 2012. The Latino Threat. Constructing Immigrants, Citizens, and the Nation. Stanford University Press, Stanford, CA.

Coale, A.J., 1972. Alternative paths to a stationary population. In: Westoff, C., Parke Jr., R. (Eds.), Demographic and Social Aspects of Population Growth, US Commission on Population Growth and the American Future, vol. I. US Government Printing Office, Washington, DC, pp. 589–603.

Espenshade, T.J., 1986. Population dynamics with immigration and low fertility. Supplement: Below Replacement Fertility in Industrial Societies: Causes, Consequences, Policies Population and Development Review 12, 248–261.

Espenshade, T.J., Bouvier, L.F., Arthur, W.B., 1982. Immigration and the stable population model. Demography 19, 125–133.

Ford, K., 1990. Duration of residence in the United States and the fertility of U.S. immigrants. International Migration Review 24 (1), 34–68.

Frank, Reanne, Heuveline, Patrick, 2005. A cross-over in Mexican and Mexican-American fertility rates: evidence and explanations for an emerging paradox. Demographic Research 12, 77–104.

Hajnal, J., 1947. The analysis of birth statistics in the light of the recent international recovery of the birth-rate. Population Studies 1 (2), 137–164.

Jonsson, Stefan Hrafn, Rendall, Michael S., 2004. The fertility contribution of Mexican immigration to the United States. Demography 41 (1), 129–150.

Kohler, H.P., Ortega, J.A., 2002. Tempo-adjusted period parity progression measures, fertility postponement and completed cohort fertility. Demographic Research 6 (6), 91–144.

Landale, N.S., Oropesa, R.S., 2007. Hispanic families: stability and change. Annual Review of Sociology 33, 381–405.

Lichter, Daniel T., Johnson, Kenneth M., Turner, Richard N., Churilla, Allison, 2012. Hispanic assimilation and fertility in new U.S. Destinations. International Migration Review 46, 767–791.

Lindstrom, D.P., Giorguli Saucedo, S., 2007. The interrelationship between fertility, family maintenance, and Mexico-US migration. Demographic Research 17, 821–858.

Mitra, S., 1983. Generalization of immigration and the stable population model. Demography 20, 111–115.

Mitra, S., 1990. Immigration, below-replacement fertility, and long-term national population trends. Demography 27, 121–129.

Morgan, S. Philip, Watkins, Susan Cotts, Ewbank, Douglas, 1994. Generating Americans: ethnic differences in fertility. In: Watkins, S.C. (Ed.), After Ellis Island: Newcomers and Natives in the 1910 Census. Russell Sage Foundation, New York.

New York Times, 2012. Hispanic Pregnancies Gall in U.S. As Women Choose Smaller Families (31.12.12.).

Parrado, Emilio A., 2011. How high is Hispanic fertility? Immigration and tempo considerations. Demography 48, 1059–1080.

Parrado, Emilio A., Flippen, Chenoa, 2012. Hispanic fertility, immigration, and race in the twenty-first century. Race and Social Problems 4 (1), 18–30.

Parrado, Emilio A., Morgan, S. Philip, 2008. Intergenerational fertility patterns among hispanic women: new evidence of immigrant assimilation. Demography 45 (3), 651–671.

Pew Hispanic Center, 2012. Net Migration from Mexico Falls to Zero – and Perhaps Less. Pew Hispanic Center, Washington, DC.

Pew Research Center, 2012. U.S. Birth Rate Falls to a Record Low; Decline Is Greatest Among Immigrants. Pew Research Center, Washington, DC.

Preston, S.H., Coale, A., 1982. Age structure, growth, attrition, and accession: a new synthesis. Population Index 48 (2), 217–259.

Preston, S.H., Wang, H., 2007. Intrinsic growth rates and net reproduction rates in the presence of migration. Population and Development Review 33 (4), 657–666.

Ryder, N.B., 1980. Components of temporal variations in American fertility. In: Hiorns, R.W. (Ed.), Demographic Patterns in Developed Societies. Symposia of the Society for the Study of Human Biology. Taylor and Francis Ltd., London, pp. 15–54.

Ryder, N.B., 1983. Cohort and period measures of changing fertility. In: Determinants of Fertility in Developing Countries, vol. 2. Academic Press, New York, pp. 737–756.

Ryder, N.B., 1986. Observations on the history of cohort fertility in the United States. Population and Development Review 12 (4), 617–643.

Schmertmann, C.P., 1992. Immigrants' ages and the structure of stationary populations with below-replacement fertility. Demography 29, 595–612.

Schoen, R., 2004. Timing effects and the Interpretation of period fertility. Demography 41, 801–819.

Sutton, P.D., et al., 2011. Recent decline in births in the United States, 2007–2009. NCHS Data Brief 60.

Toulemon, L., 2004. Fertility among immigrant women: new data, a new approach. Population and Societies 400, 1–4.

Migration and Health

Sunshine Rote and Kyriakos Markides, The University of Texas Medical Branch at Galveston, Galveston, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Most migration takes place within countries, but the number of international migrants, typically from less developed to more developed countries has been on the rise in recent decades. The impact of migration on health is a complex phenomenon. There are health issues associated with the point of origin, point of destination, the journey itself, and return migration. Most research has focused on the health of voluntary immigrants to Western societies and, despite increasing evidence that immigrants are relatively healthy, has often emphasized their health problems rather than any health advantages they may have. Below, the authors review research on migration and health paying particular attention to the health of immigrants at different stages in the migratory process.

Historical Development of the Field

Demographers and social scientists have always been interested in migration, one of the demographic processes which, along with fertility and mortality, determine population size, composition, and distribution (Poston and Bouvier, 2010). Interest in the impact of migration on health has increased over the latter part of the twentieth century. Although most migration in the world takes place within countries, there has been little interest in its impact on health primarily because it is not typically as disruptive as international migration. Previously, most health research was conducted in Western societies where the focus was on the health status and health care needs of native populations. Immigrants were either excluded or were too few, both in number and as a proportion of study samples, to yield reliable estimates of their health status and health needs.

Because the late twentieth century and early twenty-first century have seen increasing immigration to Western societies from poorer nations, scholars and policy makers have made a concerted effort to understand the health status and health care needs of rising numbers of immigrants and how they impact host countries' health and social service systems (Eschbach et al, 1999; Friis et al., 1998; Markides and Gerst, 2011). While attention has focused on the health status of immigrants as a whole, limited attention has been paid to health throughout the migratory process. There has also been some interest in the negative impact of dislocation and forced migration as well as the impact of smuggling and trafficking of migrants. To frame the discussion of research on migration and health, the authors use a migratory process framework proposed by Zimmerman et al. (2011), which includes predeparture, travel, destination, and return.

Predeparture

A number of studies have suggested that immigrants to Western societies appear to be as healthy as if not healthier than native populations. In terms of predeparture, younger and healthier individuals likely select into the migration process. Most people immigrate for employment or occupational reasons, which requires a relatively good level of health. Western countries typically require medical screenings by prospective immigrants. More recently, the Netherlands and Germany have instated predeparture language screenings for migrants from certain developing countries seeking family reunification (Groenendijk, 2011). Additionally, people who immigrate tend to have a positive outlook on their lives and futures, which is consistent with good health (Markides and Gerst, 2011).

Markides and Coreil (1986) suggested such a 'healthy migrant' or 'migration selection' effect to explain the relatively low mortality rates and good health of Mexican Americans living in the southwestern United States. The term 'epidemiologic paradox' was applied because Mexican Americans shared similar socioeconomic characteristics and conditions as African-Americans, yet their mortality and health conditions were similar to the more advantaged non-Hispanic whites (see also, Markides and Eschbach, 2005, 2011).

Research has demonstrated that foreign-born persons are generally healthier than native-born Americans from all major ethnic origins on a variety of health outcomes including mortality risk, functional limitations, chronic conditions such as heart disease, and certain types of cancer (Cho et al., 2004; Markides and Eschbach, 2005; Rogers et al., 2000; Ronellenfitsch et al., 2006; Singh and Hiatt, 2006). In terms of mental health, while most research finds that immigrants have better or similar emotional health as native-born persons (e.g., Dey and Lucas, 2006; Grant et al., 2004), some research reports more psychological distress among immigrants (e.g., Alegria et al., 2007; Angel et al., 2001).

Travel

Most research on the transit phase of migration focuses on the spread of communicable diseases such as tuberculosis and human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome (AIDS), and although most immigrants have no health problems associated with the travel portion of the migratory process, certain groups such as refugees and smuggled immigrants may experience health problems during this phase. The rise of tuberculosis is a major public health concern worldwide. This is especially evident in Western countries, where its prevalence is higher among immigrant groups, especially those from poor countries. Overall, the patterns of the disease among immigrants largely reflect patterns of disease

in their country of origin and are most prevalent among those with living in poverty. Although the overall impact of tuberculosis on the host countries does not appear to be major, it has a major impact among the immigrant communities, especially when immigrants live in crowded and unsanitary environments, conditions that may promote the spread of communicable diseases (Carballo et al., 1998; Spallek et al., 2010).

Research also indicates that the prevalence of HIV/AIDS is less pronounced among immigrant groups (Prossner et al., 2012), but that the distribution of the disease among immigrants is most evident among those of lower socioeconomic standing. In 2010, the United States lifted the restrictive ban on preventing immigrants with HIV/AIDS from entering the country and no longer requires HIV/AIDS testing as part of immigration medical screenings (CDC, 2011). More recently, public health concern has been paid to the problems of HIV/AIDS stigma and poor access to health care among immigrants (Katz, 2012; Ojikutu et al., 2013).

With rising globalization, the numbers of people wanting to immigrate to more developed countries is on the rise. In addition, there is a growing demand for unskilled workers in developed countries and despite more stringent entry requirements, a rapidly growing number of people become victims of international trafficking. Lacking social and legal protection, they often become victims of violence. Pretravel traumatic events like kidnapping and sexual violence can result in psychological distress and mental health problems (Gushulak and MacPherson, 2006).

Smuggled persons may also experience trauma, deprivation, violence, or injury during migration (Gushulak and MacPherson, 2006). Mode of transportation or environmental concerns such as extreme heat or cold or journeys in confined spaces can often lead to illness, injury, or even death (Gushulak and MacPherson, 2000). This has been documented at the United States–Mexico border (Guerrette and Clark, 2005) and among immigrants smuggled from China to Canada (Allan and Szafran, 2005).

As with illegally smuggled immigrants, there is a need to better understand and to better treat the mental and physical health problems of refugees. In a recent meta-analysis, Lindert et al. (2009) found that rates of depression and anxiety are twice as high among refugees as among the general US population. These differences are attributed to possible predeparture war and violence, the strain of migrating, and having little choice in the area of relocation.

Few studies have focused on positive aspects of the immigrant experience, and most focus on the health problems of the immigrants and how they impact the host societies' health and social service systems. Although many immigrants do have special health problems that demand attention, as demonstrated above, the focus on their health problems tends to perpetuate stereotypes that often foster anti-immigrant feelings and actions (Junghans, 1998; Markides et al., 2009).

Destination

Upon arrival to the country of destination, most migrants are in relatively good health. This has been documented in Australia, Canada, the United States, and many European countries. Immigrants, especially those from poor countries, may benefit from the improved living conditions and the better health care of their country of destination. The process of migrating, however, has often been considered a major and stressful life event since it entails a dramatic change in living arrangements, employment, and cultural opportunities (Friis et al., 1998; Markides and Gerst, 2011).

As immigrants live in their country of destination longer and become more acculturated and integrated, their physical and mental health status begins to mirror that of their native-born counterparts (Antecol and Bedard, 2006). For example, Singh and Miller (2004) found that recent immigrants to the United States report less disability and chronic diseases than immigrants with longer durations. Ronellenfitsch and Razum (2004) found that immigrants from Eastern Europe to Germany have better health than native-born adults but this health advantage disappeared after 5 years of living in the country. Similar findings for health convergence among immigrant and native-born groups are also found in Canada (Beiser, 2005).

Timing or age at entrance into the United States can also influence general health outcomes and a range of health behaviors. Migration also leads to changes in lifestyle factors such as poor diet, increased alcohol and cigarette use, and reduced physical activity, which can contribute to adverse health outcomes (Friis et al., 1998; Singh et al., 2011). Foreign-born men and women living in the United States are less likely to use tobacco or drink heavily than native-born adults (Lucas et al., 2003). However, a longer stay in the United States is associated with increases in these health behaviors.

Obesity is another concern worldwide and is shaped by duration in a country of destination. In the Netherlands, the percentage of adults overweight and obese is particularly high among immigrant groups from Turkey and Morocco. Similar results are found for South Asian immigrants living in the United Kingdom (Stronks, 2003). Antecol and Bedard (2006) found that immigrants have lower body mass indexes than native-born adults in the United States; however, these differences were not evident among immigrants living in their country of destination for 10–15 years. This higher risk of weight gain and obesity is likely attributable to changes in dietary habits and increase in sedentary activities, availability of food, as well as psychological distress and problems of social integration.

Another issue that has received attention in recent years is encounters with discrimination. Finch et al. (2000) find that Mexican immigrants living in the United States report more discrimination as they become more acculturated to US language and customs. Viruell-Fuentes (2007), similarly, found that second-generation Mexican immigrant women report more discrimination than first-generation immigrants. Safi (2010) found that both first- and second-generation immigrants' experiences with discrimination lead to lower levels of life satisfaction. Korean immigrants in Canada who report being treated unfairly or being refused services in public areas tend to report greater psychological distress than those who do not (Noh et al., 2007). Overall, discrimination more so than acculturation is seen as integral for understanding the development of health and psychological problems among immigrants since it takes into account the context of social structure and rather than individual adaptation (Hunt et al., 2004; Viruell-Fuentes, 2007).

Community-level living conditions may also impact physical and mental health of immigrants more so than of native-born adults. Zunzunegui et al. (2006) found that immigrants in Montreal were more vulnerable to psychological distress and poor self-rated health in high-unemployment communities than were native-born adults. Lorant et al. (2008) find that immigrants in Belgium from Turkey, Morocco, and southern Europe have poorer self-rated health than native-born adults mostly attributable to higher unemployment and lack of public services in their neighborhoods.

Social support and neighborhood cohesion within immigrant communities, however, may buffer the effects of stressors for health outcomes. Social integration and support can take numerous forms. It encompasses the nature and type of social networks as well as the support perceived or received from friends, family, or the community at large. Eschbach et al. (2004) argue that for Mexican Americans living in the southwestern United States, a high-density immigrant area may provide sociocultural resources that protect against health decline.

Health care access and utilization is also another concern when immigrants reach their country of destination. For some, increases in access to high-quality medical care can improve health maintenance; however, economic and linguistic barriers may pose problems for immigrants. Immigrants are less likely to have insurance and a regular source of care than native-born adults in the United States, are less satisfied with their care, and report more discrimination in health care settings (Derose et al., 2009).

Return Migration

For some immigrants, returning to their country of origin is an expected part of the migration process, although decisions to return may be highly dependent on length of stay in both country of destination and country of origin. Return migration may be fueled by the need for family reunification (Dustmann, 2003), increased employment opportunities at home (Reagan and Olsen, 2000), retirement, or even health complications and a desire to die in one's homeland (Abraido-Lanza et al., 1999). For immigrants in the United States who return to Mexico, the process has been described as a 'salmon bias' since these immigrants may die in their country of origin and, therefore, are not taken to account in studies of immigrants and health/mortality in mid or later life. There has been partial support found for a salmon bias, but researchers have found that it is too small to account for the mortality advantage of Hispanic immigrants (Turra and Elo, 2008; also see: Markides and Eschbach, 2011). Similarly, Riosmena et al. (2012) find partial support for a salmon bias but this was most evident among immigrants who had been in the United States for shorter durations.

Conclusions and Future Research

Overall, research indicates that immigrants are in relatively good health upon arrival mostly due to the process of healthy individuals selecting into the process of migration. After living in a country of destination for longer, their health status seems to converge to similar levels as native-born persons. Some explanations for convergence include changes in daily activities, eating habits, and exposure to discrimination. Groups of immigrants such as refugees, illegally smuggled immigrants, and those living in poverty have health concerns associated with the travel and destination phases of the migratory process. Certain groups may return to their country of origin when faced with health complications.

One of the biggest obstacles to studying migration and health is a lack of available survey data. Some time ago, Kasl and Berkman (1983) outlined an ideal study design that would involve migrants-to-be, nonmigrants at the point of origin, and native residents at the point of destination. Following the same individuals across borders and throughout the migratory process and asking about a range of psychosocial and environmental characteristics will lend better insight to the health immigrants bring with them to the process, the social context of their lives in the country of destination, and important areas of intervention to ensure the health of immigrants at every stage of the migratory process.

While such a study would be a massive undertaking, some other considerations for future data collection is to provide more information about the characteristics of the immigrant experience, including the length of stay in the country of destination, health behaviors pre- and postdeparture, and the characteristics of the immigrant experience in different domains: work, family life, and community. There is also concern about the diversity in immigrant populations and how overarching terms of 'Asians' or 'Hispanics' may hide important intraethnic variations in the migration experience and health outcomes. Focusing on specific immigrant groups along with health-related outcomes (access to medical care, obesity, and health behaviors) will lend to a better understanding of the complex and dynamic process of migration's influence on health.

See also: Adoption, Demography of; Adoption, Demography of; Immigration: Health and Mortality; Mental Health; Migration and Development; Migration and Social Capital; Migration: Motivations; Minorities; Mortality Differentials: Selection and Causation; Mortality and the HIV/Aids Epidemic; Occupational Health; Stress, Coping and Health.

Bibliography

Abraido-Lanza, A.F., Dohrenwend, B.P., Ng-Mak, D.S., Turner, J.B., 1999. The Latino mortality paradox: a test of the "salmon bias" and healthy migrant hypotheses. American Journal of Public Health 89, 1543–1548.

Alegria, M., Shrout, P.E., Woo, M., Guarnaccia, P., Sribney, W., Vila, D., Polo, A., Cao, Z., Mulvaney-Day, N., Torres, M., Canino, G., 2007. Understanding differences in past year psychiatric disorders for Latinos living in the US. Social Science & Medicine 65, 214–230.

Allan, G.M., Szafran, O., 2005. Health of Chinese illegal immigrants who arrived by boat on the west coast of Canada in 1999. Journal of Immigrant Health 7, 233–238.

Angel, J.L., Buckley, C.J., Sakamoto, A., 2001. Duration or disadvantage? Exploring nativity, ethnicity, and health in midlife. Journals of Gerontology Series B: Psychological Sciences and Social Sciences 56, S275–S284.

Antecol, H., Bedard, K., 2006. Unhealthy assimilation: why do immigrants converge to American health status levels? Demography 43, 337–360.
Beiser, M., 2005. The health of immigrants and refugees in Canada. Canadian Journal of Public Health 96, S30–S44.
Carballo, M., Divino, J.J., Zeric, D., 1998. Migration and health in the European Union. Tropical Medicine & International Health 3, 936–944.
Center for Disease Control, 2011. Final Rule Removing HIV Infection from U.S. Immigration Screening. Immigrant and Refugee Health. Available from: http://www.cdc.gov/immigrantrefugeehealth/laws-regs/hiv-ban-removal/final-rule.html.
Cho, Y., Frisbie, W.P., Hummer, R.A., Rogers, R.G., 2004. Nativity, duration of residence, and the health of Hispanic adults in the United States. International Migration Review 38, 184–211.
Derose, K.P., Bahney, B.W., Lurie, N., Escarce, J.J., 2009. Review: immigrants and health care access, quality, and cost. Medical Care Research and Review 66, 355–408.
Dey, A.N., Lucas, J.W., 2006. Physical and Mental Health Characteristics of US-and Foreign-born Adults, United States, 1998-2003. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.
Dustmann, C., 2003. Return migration, wage differentials, and the optimal migration duration. European Economic Review 47, 353–369.
Eschbach, K., Hagan, J., Rodriguez, N., Hernandez-Leon, R., Bailey, S., 1999. Death at the border. International Migration Review 33, 430–454.
Eschbach, K., Ostir, G.V., Patel, K.V., Markides, K.S., Goodwin, J.S., 2004. Neighborhood context and mortality among older Mexican Americans: is there a barrio advantage? American Journal of Public Health 94, 1807–1812.
Finch, B.K., Kolody, B., Vega, W.A., 2000. Perceived discrimination and depression among Mexican-origin adults in California. Journal of Health and Social Behavior, 295–313.
Friis, R., Yngve, A., Persson, V., 1998. Review of social epidemiologic research on migrants' health: findings, methodological cautions, and theoretical perspectives. Scandinavian Journal of Social Medicine 26, 173–180.
Grant, B.F., Stinson, S.F., Hasin, D.S., Dawson, D.A., Chou, S.P., Anderson, K., 2004. Immigration and lifetime prevalence of DSM-IV psychiatric disorders among Mexican Americans and non-hispanic Whites in the United States: results from the National Epidemiologic Survey on Alcohol and Related Conditions. Archives of General Psychiatry 61, 1226–1233.
Groenendijk, K., 2011. Pre-departure integration strategies in the European Union: integration or immigration policy? European Journal of Migration and Law 13, 1–30.
Guerette, R.T., Clarke, R.V., 2005. Border enforcement, organized crime, and deaths of smuggled migrants on the United States – Mexico border. European Journal of Criminal Policy Research 11, 159–174.
Gushulak, B.D., MacPherson, D.W., 2000. Health issues associated with the smuggling and trafficking of migrants. Journal of Immigrant Health 2, 67–78.
Gushulak, B.D., MacPherson, D.W., 2006. The basic principles of migration health: population mobility and gaps in disease prevalence. Emerging Themes in Epidemiology 3 (3).
Hunt, L.M., Schneider, S., Comer, B., 2004. Should "acculturation" be a variable in health research? A critical review of research on US Hispanics. Social Science & Medicine 59, 973–986.
Junghans, T., 1998. How unhealthy is migrating? Tropical Medicine and International Health 3, 933–934.
Kasl, S.V., Berkman, L., 1983. Health consequences of the experience of migration. Annual Review of Public Health 4, 69–90.
Katz, M.H., 2012. HIV infection among persons born outside the United States. JAMA 308, 623–624.
Lindert, J., Ehrenstein, O.S., von Priebe, S., Mielck, A., Brähler, E., 2009. Depression and anxiety in labor migrants and refugees – a systematic review and meta-analysis. Social Science & Medicine 69, 246–257.
Lorant, V., Van Oyen, H., Thomas, I., 2008. Contextual factors and immigrants' health status: double jeopardy. Health & Place 14, 678–692.
Lucas, J.W., Barr-Anderson, D.J., Kington, R.S., 2003. Health status, health insurance, and health care utilization patterns of immigrant Black men. American Journal of Public Health 93, 1740–1747.
Markides, K.S., Coreil, J., 1986. The health of Hispanics in the southwestern United States: an epidemiologic paradox. Public Health Reports 101, 253–265.
Markides, K.S., Eschbach, K., 2005. Aging, migration, and mortality: current status of research on the Hispanic paradox. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 60, S68–S75.
Markides, K.S., Eschbach, K., 2011. Hispanic paradox in adult mortality in the United States. In: Rogers, R.G., Crimmins, E.M. (Eds.), International Handbook of Adult Mortality. Springer, New York, pp. 227–240.
Markides, K.S., Gerst, K., 2011. Immigration, health, and aging in the United States. In: Settersen, R., Angel, J.L. (Eds.), Handbook of the Sociology of Aging. Springer, New York, pp. 103–116.
Markides, K.S., Salinas, J., Sheffield, K., 2009. The health of older immigrants. Generations 32, 46–52.
Noh, S., Kaspar, V., Wickrama, K.A.S., 2007. Overt and subtle racial discrimination and mental health: preliminary findings for Korean immigrants. American Journal of Public Health 97, 1269–1274.
Ojikutu, B., Nnaji, C., Sithole, J., Schneider, K.L., Higgins-Biddle, M., Cranston, K., Earls, F., 2013. All Black people are not alike: differences in HIV testing patterns, knowledge, and experience of stigma between U.S.-born and Non–U.S.-born Blacks in Massachusetts. AIDS Patient Care and STDs 27, 45–54.
Poston, D.L., Bouvier, L.F., 2010. Population and Society: An Introduction to Demography. Cambridge University Press, New York.
Prossner, A.T., Tang, T., Hall, H.I., 2012. HIV in persons born outside the United States, 2007–2010. JAMA 398, 601–607.
Reagan, P.B., Olsen, R.J., 2000. You can go home again: evidence from longitudinal data. Demography 37, 339–350.
Riosmena, F., Wong, R., Palloni, A., 2012. Migration selection, protection, and acculturation in health: a binational perspective on older adults. Demography 50, 1039–1064.
Rogers, R.G., Hummer, R.A., Nam, C.B., 2000. Living and Dying in the USA: Behavioral, Health, and Social Differentials of Adult Mortality. Academic Press, San Diego.
Ronellenfitsch, U., Kyobutungi, C., Becher, H., Razum, O., 2006. All-cause and cardiovascular mortality among ethnic German immigrants from the former Soviet Union: a cohort study. BMC Public Health 6, 16.
Ronellenfitsch, U., Razum, O., 2004. Deteriorating health satisfaction among immigrants from Eastern Europe to Germany. International Journal for Equity in Health 3, 4.
Safi, M., 2010. Immigrants' life satisfaction in Europe: between assimilation and discrimination. European Sociological Review 26, 159–176.
Singh, G.K., Hiatt, R.A., 2006. Trends and disparities in socioeconomic and behavioural characteristics, life expectancy, and cause-specific mortality of native-born and foreign-born populations in the United States, 1979–2003. International Journal of Epidemiology 35, 903–919.
Singh, G.K., Miller, B.A., 2004. Health, life expectancy, and mortality patterns among immigrant populations in the United States. Canadian Journal of Public Health. Revue Canadienne de Sante Publique 95, 1–17.
Singh, G.K., Siahpush, M., Hiatt, R.A., Timsina, L.R., 2011. Dramatic increases in obesity and overweight prevalence and body mass index among ethnic-immigrant and social class groups in the United States, 1976–2008. Journal of Community Health 36, 94–110.
Spallek, J., Zeeb, H., Razum, O., 2010. Prevention among immigrants: the example of Germany. BMC Public Health 10, 92.
Stronks, K., 2003. Commentary: public health research among immigrant populations: still a long way to go. European Journal of Epidemiology 18, 841–842.
Turra, C.M., Elo, I.T., 2008. The impact of salmon bias on the Hispanic mortality advantage: new evidence from social security data. Population Research and Policy Review 27, 515–530.
Viruell-Fuentes, E.A., 2007. Beyond acculturation: immigration, discrimination, and health research among Mexicans in the United States. Social Science & Medicine 65, 1524–1535.
Zimmerman, C., Kiss, L., Hossain, M., 2011. Migration and health: a framework for 21st century policy-making. PLoS Medicine 8, e1001034.
Zunzunegui, M.-V., Forster, M., Gauvin, L., Raynault, M.-F., Douglas Willms, J., 2006. Community unemployment and immigrants' health in Montreal. Social Science & Medicine 63, 485–500.

Relevant Websites

http://www.who.int/hac/techguidance/health_of_migrants/en/index.html – World Health Organization: Migrant Health.

http://www.cdc.gov/immigrantrefugeehealth/ – Centers for Disease Control and Prevention.

http://www.ucghi.universityofcalifornia.edu/coes/migration-and-health/ – University of California.

http://keck.usc.edu/Education/Academic_Department_and_Divisions/Department_of_Preventive_Medicine/Research/Immigrant_Health_Initiative.aspx – University of Southern California, Keck School of Medicine.

Migration and Social Capital

Nadia Y Flores-Yeffal, Texas Tech University, Lubbock, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Migration social capital is referred to the tangible and/or intangible benefits acquired by international migrants as a result of having contacts (or social networks) at the place of origin and/or at the place of destination of the migrant. Those benefits can be acquired at any time during the migratory process, such as, during the first migration trip and later trips, while crossing international borders, and during and after the settling process. It reduces the cost and risk associated with international movement. Social capital is often provided by *paisanos* (or countryman), friends, or by family members.

Introduction

Scholars studying international migration theory and the economic sociology of migration tend to address the social, noneconomic reasons for migration such as social networks and social capital. International migration scholars have borrowed the concept of social capital to refer to the benefits (tangible and intangible) acquired by international migrants as a result of having personal contacts with prior migrants at the place of origin or at the place of destination. The information and resources provided help to increase the likelihood of migration by nonmigrants as migration social capital helps to reduce the cost and risks associated with international movement (Massey and Espinosa, 1997). The information and resources provided by migrant social capital can work differently for different people and in different settings (Garip, 2008). In addition, the distance between the place of origin and the place of destination also can make a difference with regards of how migration social capital operates (Massey et al., 2011). In addition, rural or urban places of origin may also dictate how much migration social capital may be available in different settings and how it functions (Fussell and Massey, 2004; Hernandez-León, 2008; Flores-Yeffal and Aysa-Lastra, 2011; Flores-Yeffal, 2013). The context of the place of origin is therefore considered very important as the structure of the social networks in which social capital is exchanged could take many different forms (Flores-Yeffal, 2013).

The contemporary meaning of social capital was first introduced by Loury (1977) and theoretically developed by Bourdieu (1986) and Coleman (1988). Originally, the concept was not necessarily associated with issues of international migration until the term became very popular and international migration scholars began to borrow it in order to explain the role of social networks in the international migration scenario (i.e., Massey et al., 1987). Given that the origins of the concept were not necessarily conceptualized in the realm of international migration, sociologists in general disagree about what constitutes social capital, but international migration researchers generally agree that it refers to the tangible and intangible resources provided by relatives, friends, and *paisanos* (countryman) to facilitate migration, settlement, and employment in a foreign country (see Massey et al., 1987; Massey and Espinosa, 1997; Flores-Yeffal, 2013). Past studies about Europeans demonstrate that historically and around the world in the past international migrants have utilized social capital to migrate. For example, Gans (1962) wrote about European-origin segregated immigrant communities during the turn of the twentieth century in Boston. These communities had developed as a result of social networks and social capital exchange.

Today, the field recognizes mainly two theories related to social networks: *the cumulative causation theory* and *the social capital theory*. Actually, the social capital theory is considered part of the *cumulative causation theory* (see Massey et al., 1998). The theory of cumulative causation refers to the tendency of migration to self-perpetuate, regardless of what initially caused the migration to begin (Myrdal, 1957; Massey, 1990). As more people migrate, more migration social capital accumulates and so more people migrate from a single community as a result that international movement becomes easier and less costly to others. This theory takes into account the changes in context and social culture in the transnational context, at the place of origin, as well as at the place of destination. For example, it claims that a culture of migration develops among the members of the sending community due to the benefits they observe from international migration and remittances. For example, they may see other people in the community receiving remittances from abroad and building new homes, buying new clothes for them and their children and so they think that if they were to also migrate abroad, that they would also be able to provide the same amenities to their family members. This particular reaction is usually referred to *relative deprivation* in the international migration literature (Taylor, 1986). After seeing the benefits at the place of origin, more people decide to migrate internationally for the same reasons. They utilize their contacts and social networks to facilitate their migration.

The social capital theory, on the other hand, has to do with the concept of the benefits provided to prospective migrants having contacts at the place of destination. According to this network theory, migration becomes easier and less costly and risky over time because of a growing network of social contacts in the receiving community. These contacts share information and resources with future migrants, whether family, friends, or people from the same community, aiding in the process of migrating, settling, and searching for employment in the host country. These resources and connections are a form of social capital that dramatically increases the likelihood of additional migration among persons with family or friends living in the

US (Massey et al., 1987; Massey, 1990). Sociologists view such social structures and interpersonal networks as potential sources of social capital.

The Sources of Social Capital and International Migration

Portes and Sensenbrenner (1993) defined social capital as "collective expectations affecting individual economic behavior" (p. 1326). They used classic sociological concepts and introduced four sources of social capital mainly related to international migration based on the collective behavior of people:

1. Value introjections: "socialization into consensually established beliefs" that prompt behavior that is not simply out of naked greed
2. Reciprocity exchanges
3. Bounded solidarity: "situational reactive sentiments", which result in collective conscience
4. Enforceable trust: "pluralistic rewards and sanctions linked to group membership" (Portes and Sensenbrenner, 1993: p. 1326).

Immigration scholars have used such sources of social capital primarily to examine economic enterprises formed by immigrant entrepreneurs. In fact, Portes and Sensenbrenner (1993) expounded the above-mentioned concepts of social capital using examples chiefly related to immigrant ethnic entrepreneurship enterprises, and not solely as examples of social support in order to migrate and settle in the US Studies of social networks among unskilled and undocumented immigrants have neglected to examine closely the sources of social capital formulated by Portes and Sensenbrenner, due to the generalization that undocumented immigrants do not tend to form economic enterprises (see Valdez, 2011). The concept of Migration-Trust Networks developed by Flores-Yeffal (2013) provides an avenue for further developing the sources of social capital proposed by Portes and Sensenbrenner (1993), especially reciprocity, bounded solidarity, and enforceable trust relevant to the help provided to international migrants who lack legal documentation to migrate. Migration-Trust Networks refers to the social network of people in which the majority of its members lack legal documentation. Migration-Trust Networks she argues are mostly driven by relationships of trust and the social monitoring of its members. Its members tend to segregate themselves at places of destination in order to provide information, resources, and social support to the rest of the members of the group who also lack legal documentation.

Portes and Sensenbrenner (1993) proposed the concept of *bounded solidarity* as a source of social capital, which "focuses on those situational circumstances that can lead to the emergence of principled group-oriented behavior quite apart from any early value introjections" (p. 1324). In other words, the members of the group develop certain expectations of collective behavior in which any actions or decisions made should benefit not just the individual but all the members of the group. Portes and Sensenbrenner (1993) argue that one example of the situational circumstances immigrants face that lead to such bounded solidarity behavior is the extent to which they are subject to racial or ethnic discrimination once they arrive to the US.

Supportive behavior happens under certain social circumstances according to some sociologists. For example, the mechanical solidarity concept by Durkheim (1997[1933]) in which he argues that in small (rural communities) people tend to like each other because they have similar occupations and skills with each other, and feel collective conscience to each other (don't think of themselves as individuals, but as a group), and the members of the community enforce a series of values and norms between each other which leads to cooperation among the members of the group. Similar to the Mechanical Solidarity concept, the idea of *enforceable trust*, proposed by Portes and Sensenbrenner (1993), refers to a set of social expectations to follow that will be enforced through social monitoring, which can result in the rejection or punishment of those who do not conform to those agreed-upon values and behaviors. The similarity of experiences, such as crossing the border clandestinely, feeling discriminated against, sharing the fear of being caught by the border patrol, and being exploited in the US, gives rise to the enforceable trust social mechanism among migrants once they identify each other in the US. Also, Tilly (2007) has argued that international migrants help each other through bounded solidarity due to a series of expectations from the members of the rest of their group and that the social networks of migration from transnational communities are based on trust networks. Flores-Yeffal (2013) also argues that the social expectations of behavior at the community of origin are transferred to the place of destination and that the transnational relations of trust among migrants and nonmigrants create the fundamental conditions for the exchange of social capital among network members, more specifically to what she calls, Migration-Trust Networks.

Religion also plays a major role in the exchange of social capital as migrants in many cases help others and support others out of religious faith thinking that by helping others, they will probably be rewarded by God and, therefore, they will do better in the future (Hagan, 2008; Flores-Yeffal, 2013). This religious faith is often embedded within their culture and such culture is also preserved when international migrants travel and settle at the places of destination.

Portes and Sensenbrenner (1993) took into account that reciprocity was an important source of social capital and assumed it was directly by the exchange of favors amongst immigrants; this is to say, those who received favors later returned them to the givers. Instead, Flores-Yeffal (2013) found evidence that in the case of Migration-Trust Networks, in which most members of the network lack legal documents, the members of the network do not necessarily return the favor directly to the person who originally provided the help, instead, international migrants tend to pay the favors forward to other future migrants. This happens because those who provide the help are usually more established migrants than the newcomers. She refers to this type of social exchange as, "risk pooling" (Flores-Yeffal, 2013). Therefore, reciprocity relations may vary depending on the group dynamics during social capital exchange.

In the field of international migration and social networks, experts have claimed that networks are not necessarily static entities, but that they change and transform over time (Boyd,

1989; Granovetter, 1985; Hagan, 1998; Mahler, 1995; Menjívar, 2000). For example, Boyd (1989) found that those who participate in social networks meet specific roles and functions such as those networks used to just get to the US as a first-time migrant, while others participate in networks which are used later once the migrant is more established. In addition, studies of social networks among migrants have found that competition among immigrants or exploitation of fellow countrymen may disintegrate ethnic solidarity and cohesiveness among network participants (i.e., Mahler, 1995; Menjívar, 2000).

Type of Social Ties That Provide Social Capital to International Migrants

Mark Granovetter (1973) revolutionized the organizational literature when he presented the concept of "The Strength of Weak Ties". Granovetter (1973) argued that weak ties could get people better jobs, as the information obtained by acquaintances was of a better quality and more efficient than that obtained by more established network ties, or from strong ties. He argued that the information obtained by strong ties was usually more redundant, less effective, and inefficient. Later on, Granovetter revisited his theory and acknowledged that strong ties were also effective for people during times of risk and uncertainty (Granovetter, 1982, 1985). Scholars in the area of international migration have debated as they have found that strong ties have been also effective for international migrants in getting better jobs and/or better wages (Mouw, 2003; Aguilera and Massey, 2003). Later on, Flores-Yeffal also concluded that indeed, international migrants often began using what many called, weak ties for obtaining social capital from *paisanos*, but that later on those ties were transformed on strong ties and became more effective for social capital exchange through the Migration-Trust Networks in the case of Mexican undocumented emigration to the US (Flores-Yeffal, 2013).

Social capital is usually exchanged by family members, friends, or *paisanos* (or countryman). According to Flores-Yeffal (2013), *paisanos* play a crucial role in the exchange of social capital given their important role as pioneers as they are usually the first ones to migrate internationally. At first, not everyone has a family member abroad and for that reason it is more difficult for family members to be the ones providing help to newcomers. It is not until a lot of people from the same community have migrated internationally that more family members will be available to provide the help (see Flores-Yeffal and Aysa-Lastra, 2011).

Certain ethnic behaviors among the members of specific groups (among Mexican migrants, for example) tend to develop fictitious ties in the absence of other ties to exchange social capital. Among Mexicans migrants, for example, one very important way in which both rural- or urban-origin international migrants increase their chances of receiving social support or social capital at the time of migration is through the use of fictitious ties. *Cuatismo* is a type of fictitious tie practiced with the purpose of strengthening social relationships, especially among males. *Cuatismo* is a form of friendship, but stronger (Lomnitz, 1977). *Cuates* usually get together more often than normal friends, share intimate information, and support each other more strongly. Migrants who are *cuates* get together on Sundays, for example, to watch soccer games and drink beer, plan vacation trips together, spend holiday celebrations together, and also become *compadres* in order to further strengthen their relationship (Flores-Yeffal, 2013).

Studies have found that in Mexico and other parts of Latin America, *compadrazgo* is a way to build stronger relationships of trust, or to convert a weak relationship into a strong relationship in which participants grow closer (Lomnitz, 1977; Flores-Yeffal, 2013). Becoming a compadre can mean even more than being a cuate; it is a stronger relationship and more focused on family. One way to become a person's *compadre* or *comadre* is to baptize his or her child, thereby assuming responsibility for the child if something were to happen to the parents. This demonstrates trust and care, which in turn, strengthens their existing relationship and upgrades it to something similar to a kinship tie, or even stronger in certain cases.

Compadres and *comadres* tend to interact with each other with more frequency and spend more time together than with their own family members. Through *compadrazgo*, even those who have never met before can transform their social relationship into a strong relationship of trust and loyalty into the future. Migrants benefit from those types of fictitious ties because more social support and social capital becomes available to them. *Compadres* and *cuates* would provide social capital to international migrants in many cases in the same way a family member would do.

First Trips vs Later Trips

Migration social capital is usually more available to first-time migrants than to returning migrants, given the auspices associated with making a first trip (Massey and Espinosa, 1997; Flores-Yeffal and Aysa-Lastra, 2011). Most first-time migrants need some sort of information and assistance regardless of their legal status at the place of destination. The most important social capital they need is information on where to go, what to say, what to bring, how to find a job, where to live, how the system works, and so forth. Someone has to provide them with such information for their trip to be successful, otherwise, they may end up going through a lot of trouble, getting to a place in which they cannot find housing or jobs, and end up having a not-so-good experience. Social ties usually tell prospective migrants in advance whether there are jobs available or not at the place of destination, so most first-time migrants already know whether they will be able to find jobs or not. García (2005), for example, argues that once international migrants get to the place of destination, they already have jobs waiting for them. Usually, employers tend to notice the good working ethic of the workers of that specific network and then use those workers to recruit more workers who would probably also have good work ethic (García, 2005). The monitoring behavior of the people who belong to the network makes sure that those who are recommended for new jobs are responsible on their jobs, so more people in the network can have access to jobs in the future (García, 2005).

Those who lack legal documents to migrate besides information of where to go and what to do to find a job also need crucial information on how to be safe and avoid getting caught by immigration authorities or be deported. For this reason,

first-time migrants need crucial information, for example, on how to get to the border, where to find a Coyote (or smuggler), where to go once they are at the place of destination, and how to find a job without having legal documents. According to Flores-Yeffal (2013), Migration-Trust Networks provide such information to undocumented immigrants and they also provide a place to stay, food, clothes to wear (as they usually arrive with the clothes they are wearing), transportation, help to find a job without having legal documentation, where to go to the doctor, and help to get settled.

Returning migrants usually build human capital during each trip (see Massey and Espinosa, 1997), so they learn from their prior experiences and then they tend to rely on social capital less and less. The problem is when the rules change and the receiving country changes immigration laws and strengthens border control, for example. If the migrant knew how to cross the border before without hiring a smuggler, for example, once the rules change, the way the migrant knows how to cross, may become too risky again, so he or she would need information on how to find a trustful smuggler or Coyote. The information provided to first-time and returning migrants has to be trustful because their lives would be at risk if something goes wrong (see Flores-Yeffal, 2013). Some smugglers utilize very risky ways to transport migrants and cross them through the checking points, such as to put them inside trailers in crowded conditions and without air conditioning and in many cases the migrants can suffocate, for example.

According to Flores-Yeffal (2013) undocumented immigrants from Mexico experience the following step-by-step process to access social capital from others.

International Migration Begins

First, pioneer migrants leave their community and settle in the US. Many belong to pre-existing clique-like networks or peer groups and begin to invite their pre-existing ties to also join the migration journey. Family members are less available to provide help, so instead friends and *paisanos* – especially *paisanos* – offer prospective migrants the social capital to migrate.

Activation of Migration-Trust Network

For some communities, only a small number of pioneers are required to migrate internationally for the Migration-Trust Networks to be activated. Once migration begins, then subsequent migration occurs, and therefore, more people migrate from that community. The demands for cheap labor at the place of destination also helps to increase the migration flow. The context of reception then also becomes important for the flow to continue to grow. The context of reception is referred as the resources available to new arrivals at the place of destination, such as housing, jobs, transportation, food supply, schools for the children, and medical facilities. The prevalence of such amenities provides additional incentives for the migration flow to continue to grow.

Formation of Immigrant Enclaves

In search of social support and the exchange of social capital, recently arriving pioneer migrants begin to live next door to one other upon arrival. Such living arrangements allow for access of social support from others, easier enforcement of norms, and social monitoring among MTN members.

Pre-Existing Relationships of Trust

Prior relationships of trust are essential for the Migration-Trust Networks to function properly at the place of destination and in the transnational context. Such pre-existing conditions of social trust and social cohesiveness are necessary to effectively provide social capital to more newcomers and prospective migrants.

Relations of Trust

The vulnerability of undocumented migrants, for example, can only be overcome through reliance on relationships of trust. If a relationship of trust does not exist before migration, then a new relationship of trust may develop at the destination through religious ties and/or the commonalities related to the international migration experience and struggles back home.

Nonreciprocal Exchange of Favors

Favors are not necessarily returned to those who originally assisted migrants. Instead, favors are passed along to the new arrivals or returning migrants, which contributes to the growth of the Migration-Trust Network. While they may help from purely altruistic reasons, migrants do gain social status by helping others.

Sanctioning

Members of a migrant network are socially monitored by the other members of their network. Social capital exchange is one of the social expectations of migrant networks as people are expected to help others to migrate. Those who do not meet the social expectations of behavior and norms may be rejected or punished. Social monitoring not only occurs at the place of destination, but also at the place of origin of the migrants. People are monitored via several forms of technology, such as telephone, cellphones, videos, YouTube, and social media web sites. People posts pictures and show their behavior to others no matter where people are and those who have access are able to monitor the behavior of others.

Transnational Networks

Expansion of Migration-Trust Networks depends on transnational relations between migrants and those left in the sending community as they exchange social capital and vital information. They maintain transnational relations (between the place of origin and the place of destination of the migrant) largely through technology, such as cell phone communication, community Internet web sites and blogs, the use of Skype or Facebook, letters, videos or YouTube videos, movies, and pictures. Through such transnational contact, cultural ideas and practices are exchanged. Migrants are encouraged to carry and maintain their cultural practices at the destination, and they introduce aspects of the host culture to their home communities (Fitzgerald, 2008; Smith, 2006).

Conclusion

Therefore, migration social capital results from having contacts with prior migratory experience. Those migrants understand the hardship and dangers associated with the international migration experience and they are willing to provide first-time migrants and returning migrants with all the information and help necessary for them to migrate to the place of destination and get established. First-time migrants receive in general greater amounts of social capital and undocumented migrants tend to relay on Migration-Trust Networks as they confront greater risks.

Several social factors can influence how social capital is exchanged. The list goes on and on, but some of the factors discussed here are the history of migration of the particular flow, the context of the place of origin (whether rural or urban), legal documentation of the migrants, religion, culture, social values, the conditions of the context of reception, transnational politics and transnational relations between the place of origin and the place of destination, the type of social ties available, the age of the migration flow, the distance between the place of origin and the place of destination, the social and immigration policies in place at the place of destination, and so forth. In order to understand how social capital is exchanged within a particular international migration flow, several factors would have to be taken into account before one can understand its structure and how it functions.

See also: Altruism and Prosocial Behavior, Sociology of; Collective Behavior, Sociology of; Norms; Residential Segregation in the United States; Social Networks; Solidarity, Sociology of; Trust, Sociology of.

Bibliography

Aguilera, Michael B., Massey, Douglas, 2003. Social capital and the wages of Mexican migrants: new hypothesis and tests. Social Forces 82 (2), 671–701.

Boyd, Monica, 1989. Family and personal networks in international migration: recent developments and new agendas. International Migration Review 23, 638–670.

Bourdieu, Pierre, 1986. The forms of capital. In: Richardson, John G. (Ed.), Handbook of Theory and Research for the Sociology of Education. Greenwood Press, New York, pp. 241–258.

Coleman, James S., 1988. Social capital in the creation of human capital. American Journal of Sociology 94S, S95–S120.

Durkheim, Emile, 1997[1933]. The Division of Labor in Society. The Free Press, New York.

Fitzgerald, David, 2008. Colonies of the little motherland: membership, space, and time in Mexican migrant hometown associations. Comparative Studies in Society and History 50 (1), 145–169.

Flores-Yeffal, Nadia Y., 2013. Migration-Trust Networks: Social Cohesion in Mexican US-Bound Emigration. Texas A&M University Press, College Station.

Flores-Yeffal, Nadia Y., Aysa-Lastra, Maria, 2011. Place of origin, types of ties, and support networks in Mexico–U.S. migration. Rural Sociology 25 (3), 1–30.

Fussell, Elizabeth, 2004. Sources of Mexico's migration stream: rural, urban, and border; migrants to the U.S. Social Forces 82 (3), 937–967.

Fussell, Elizabeth, Douglas, Massey S., 2004. The limits to cumulative causation: international migration from Mexican urban areas. Demography 41 (1), 151–171. http://muse.jhu.edu/journals/demography, Demography; http://muse.jhu.edu/journals/demography/toc/dem41.1.html.

Gans, Herbert, J., 1962. The Urban Villagers. The Free Press of Glencoe, New York.

García, Carlos, 2005. Buscando Trabajo: social networking among immigrants from México to the U.S. Hispanic Journal of Behavioral Sciences 27 (1), 3–22.

Garip, Filiz, 2008. Social capital and migration: how do similar resources lead to divergent outcomes? Demography 45 (3), 591–617. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2831393/.

Granovetter, Mark, 1973. The strength of weak ties. American Journal of Sociology 78, 1360–1380.

Granovetter, Mark, 1982. The strength of weak ties: a theory revisited. In: Marsden,-Peter V., Lin, Nan (Eds.), Social Structure and Network Analysis. Sage, London, pp. 105–130.

Granovetter, Mark, 1985. Economic action and social structure: the problem of embeddedness. American Journal of Sociology 91, 481–510.

Hagan, Jacqueline, 1998. Social networks, gender, and immigrant incorporation: resources and constraints. American Sociological Review 63, 55–67.

Hagan, Jacqueline, 2008. Migration Miracle: Faith, Hope, and Meaning on the Undocumented Journey. Harvard University Press, Cambridge, MA.

Hernandez-León, Ruben, 1999. A La Aventura !":Jovenes, Pandillas y Migracion en La Conexion Monterrey-Huston. In: Gail Mummer, En (Ed.), Fronteras Fragmentadas, Zamora: El Colegio de Michoacan/Centro de Investigaciones y Desarrollo del Estado de Michoacan, pp. 115–143 (1999, 595 pp.).

Hernandez-León, Ruben, 2008. Metropolitan Migrants: The Migration of Urban Mexicans to the U.S. University of California Press, Berkeley, CA.

Lomnitz, Larissa, 1977. Networks and Marginality. New York Academic Press, New York.

Loury, Glenn C., 1977. A dynamic theory of racial income differences. In: Wallace, Phyllis A., LaMond, Anette M. (Eds.), Women, Minorities, and Employment Discrimination. D.C. Health & Co., Lexington, MA, pp. 153–186.

Mahler, Susan, 1995. American Dreaming: Immigrant Life on the Margins. University of Princeton Press, Princeton, NJ.

Massey, Douglas S., 1990. Social structure, household strategies, and the cumulative causation of migration. Population Index 56 (1), 3–26.

Massey, Douglas S., Alarcon, Rafael, Durand, Jorge, Gonzalez, Humberto, 1987. Return to Aztlan: The Social Process of International Migration from Western Mexico. University of California Press, Berkeley, CA.

Massey, Douglas S., Kristin, Espinosa, 1997. What's driving Mexico–U.S. migration? A theoretical, empirical and policy analysis. American Journal of Sociology 102 (4), 939–999.

Massey, Douglas S., Alarcon, Rafael, Durand, Jorge, Gonzalez, Humberto, Espinosa, Kristin E., Arango, Joaquín, Hugo, Graeme, Kouaouci, Ali, Pellegrino, Adela, Taylor, J. Edward, 1998. Worlds in Motion: International Migration at the End of the Millennium. Oxford University Press, Oxford.

Massey, Douglas S., Alarcon, Rafael, Durand, Jorge, Gonzalez, Humberto, Espinosa, Kristin E., Arango, Joaquín, Hugo, Graeme, Kouaouci, Ali, Pellegrino, Adela, Taylor, J. Edward, Aysa-Lastra, Maria, 2011. Social capital and international migration from Latin America. International Journal of Population Research. Article ID 834145, 18 pp. http://www.hindawi.com/journals/ijpr/2011/834145/.

Menjívar, Cecilia, 2000. Fragmented Ties: Salvadoran Immigrant Networks in America. University of California Press, Berkeley, CA.

Mouw, Ted, December 2003. Social capital and finding a job: do contacts matter? American Sociological Review 68, 868–898.

Myrdal, Gunnar, 1957. Rich Lands and Poor. Harper and Row, New York, NY.

Portes, Alejandro., Sensenbrenner, Julia, 1993. Embeddedness and immigration: notes on the social determinants of economic action. American Journal of Sociology 98, 1320–1350.

Smith, Robert, 2006. Mexican New York. University of California Press, Berkeley and Los Angeles, CA.

Taylor, J., Edward, 1986. Differential migration, networks, information and risk. In: Stark, O. (Ed.), Migration, Human Capital and Development. JAI Press, Greenwich, CT, pp. 147–171.

Tilly, Charles, 2007. Trust networks in transnational migration. Sociological Forum 22 (1), 3–24.

Valdez, Zulema, 2011. New Entrepreneurs: How Race, Class and Gender Shape American Enterprise. Stanford University Press, Stanford, CA.

Relevant Websites

http://www.socialcapitalresearch.com/literature/definition.html – Definitions of Social Capital.

http://infed.org/mobi/social-capital/ – Social Capital.

Multimedia

Flores-Yeffal, Nadia, 2012. Podcast Audio Presentation on Migration-Trust Networks: Social Cohesion in Mexican U.S.-Bound Emigration. http://ccis.ucsd.edu/2012/06/nadia-y-flores-yeffal-migration-trust-networks-social-cohesion-in-mexican-u-s-bound-emigration/.

Migration and Social Work

Usha George, Ryerson University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article highlights some of the key issues related to social work with migrants. It describes international trends in migration and provides an overview of migrants' needs and challenges. The article details the types of interventions, social workers provide to migrants and discusses some of the key issues facing the profession in terms of its interventions with migrant communities.

Introduction

Migrants broadly include four categories of persons: immigrants, refugees, temporary residents, and 'illegal' or undocumented immigrants. Immigrants are persons who choose voluntarily to migrate from their country of origin to another to take up permanent or temporary residence. Refugees are persons who are outside their countries of origin and forced to seek protection in another country on account of war, natural disasters, or of fear of being persecuted on account of race/ethnicity, nationality, or membership to a particular group. Temporary residents include visitors, students, seasonal workers, and persons who are allowed to live in a host country for a fixed period of time. 'Illegal' or undocumented immigrants are persons who enter a country without valid documentation or continue to stay on after the expiry of their visas. Each category of migrants has to overcome unique sets of challenges in their immigration and settlement in the political, social, and cultural context of their host countries. As a helping profession, social work assists individuals in responding to these challenges.

Further classifications of each migrant type might exist within countries. For example, in Canada a distinction is made between 'skilled' or 'economic immigrants' who qualify for immigration on the basis of 'points' they earn on a set of criteria in relation to their skills and qualifications, and 'family class immigrants' (the spouses/common-law partners, dependent children, or parents sponsored by permanent residents or citizens). In the United States of America, migrants could be permanent residents (green card holders), permanent or temporary workers, refugee or asylum seekers, while in Australia, migrant categories include employer and state/territory government sponsored migration, family sponsorship, and options for highly skilled business owners, investors and senior executives options.

The most prevalent 'push–pull' theory of migration (Lee, 1996) suggests that migrants might either be 'pushed' out of their country of origin due to conditions such as persecution, natural calamities, and lack of opportunity, or they might be 'pulled' by attractive features of their host country such as better employment opportunities and/or a better political/social climate. A version of this theory suggests that migration is a result of factors at the macro- or structural level (such as international socioeconomic political forces); at the meso-level (relationships between potential migrants in host countries and country of origin); and micro-level (individual factors) (Potocky-Tripodi, 2002). According to this theory, economic globalization brings with it an increased concentration of wealth in industrialized nations and stagnant economies in regions characterized by poor employment prospects and little or no state provision for social security. It is suggested that mass migration is a product of globalization's inability to create employment in countries of high need (Li, 2003).

Key Trends in International Migration

Most countries in the Northern Hemisphere have formal and informal mechanisms in place to either increase, decrease, or maintain migration, with a goal of population growth/maintenance and/or labor management. In the year 2009, Denmark, France, the Netherlands, and Russia aimed to reduce immigration while Australia, Canada, Estonia, Finland, New Zealand, and Sweden aimed for increased skilled labor immigration (Rajan, 2012).

Some of the more salient trends in international migration are the increasing number of international migrants, global movement of refugees, decreasing number of asylum applications, the feminization of migration (Castles 2000 in George 2007), and an increased demographic diversity in the host countries due to migrant flows from nontraditional source countries. International migration is highly interrelated to globalization, and associated developments such as improved transportation, communication, and growing international trade (Suárez-Orozco and Qin-Hilliard, 2004). Political changes in Middle East and North Africa, natural disasters in China, Pakistan, and other regions in Asia and a global economic crisis have all contributed to increased human migration worldwide in the past few years (IOM, 2011). In 2010, the total number of international migrants was 214 million people, 23 million more than what it was in 2005 (IOM). Among the Organization for Economic Cooperation and Development member countries, the United States of America continues to have the highest inflow of foreign migrants. Other countries with high inflows of foreign migrants were Australia, Germany, Canada, United Kingdom, Italy, and Spain (OECD, 2012). The numbers of refugees of concern to United Nations High Commissioner for Refugees were 10.5 million at the beginning of 2011. An additional 4.8 million registered refugees were being looked after in

refugee camps in the Middle East (UNHCR, n.d.). Developing countries were hosts to the largest number of refugees worldwide in 2011, with Pakistan hosting the largest number of refugees followed by the Islamic Republic of Iran (UNHCR, 2011). In 2011, the United States also saw the highest inflow of asylum seekers, followed by France and Germany (OECD).

Feminization of migration refers to the trend where almost half the world migrants in recent times are women, with an increasing number of women migrating internationally for work, apart from the more traditional reasons such as reunification with family. A gendered analysis of migration helps to understand the manner in which gender inequalities constitute the reason or become the consequences of migration, as well as the impact of women's migration on the families in the countries of origin (Morrison et al., 2008). Women constituted 49% of the world migrants in 2010, with an overrepresentation of women in the age groups of 65 and over (United Nations, 2010). Globally, migrants tend to be of working age, with the average of international migrants in 2010 being 39 years. Migrants to developed countries tend to be older than those to developing countries (United Nations).

Needs and Challenges Faced by Migrants

Needs and challenges faced by migrants are interrelated. However, this section will present needs as those necessities that the migrants must be able to meet in order to make the successful transition in their host countries. Challenges refer to those societal-level barriers faced by migrants in trying to meet their needs.

Migrant Needs

- *Information and Orientation*: Literature has demonstrated the primacy of the need of information and orientation for new migrants (Valtonen, 2008; Segal, 2002). Migrants need information related to almost all aspects of living and adjusting to their host society such as cultural norms regarding appropriateness of behavior, dealing with formal institutions, and how to access services. In practice, many migrants pursue informal channels of information and orientation over more formal sources.
- *Language Training*: Language barriers impact not only job opportunities but also opportunities for everyday interaction, socialization, learning, and integration. While language ability is one of the selection criteria for principal applicants, the dependents of principal applicants and other categories of migrants such as refugees are not subject to language proficiency requirements. The inability to speak the dominant language of the host country places migrants at a disadvantage in the labor market (Kler, 2006; Dustmann and Fabbri, 2003).
- *Employment*: Adequate income is a prerequisite to meet the needs of a family. Economic migrants, who have migrated on the merit of past educational and professional experiences, look for work that is commensurate with their credentials. However, migrants are often unable to secure work in their areas of expertise due to the nonrecognition of credentials by professional bodies and are required to requalify through education and/or internships. The inability of migrants to find appropriate employment has been found to be associated with lower self-esteem, anxiety, and a sense of hopelessness (Cabral, 2000), as well as dissatisfaction with life in the host country.
- *Housing*: As housing is dependent on income and employment, migrants are often forced to live in conditions of housing instability (Ballay and Bulthuis, 2004). In the case of economic migrants, the lack of adequate income results in substandard housing, inferior to what they were used to in their own countries of origin. Availability of affordable housing suitable for a family's needs is always a challenge. Access to public transport, schools, and hospitals is a major factor determining the choice of residence for a majority of migrants.
- *Networks and Social Support*: All migrants leave behind networks of family and friends to start life anew in their host country. In the case of refugees, this separation might have been involuntary or even violent and traumatic. In creating new social capital, the preexisting and self-created supportive networks are very important to migrants as sources of information and orientation and in looking for and finding employment (George and Chaze, 2009). Creating new social supports is a slow process and contingent on many other factors such as language ability, income, and transportation, and the existence of ethnic and cultural networks. For migrant women experiencing violence within the family, many factors such as language barriers, race, migrant status, loss of traditional sources of support due to migration, and lack of faith in formal sources of support intersect to heighten their vulnerability to victimization (Menjívar and Salcido, 2002).
- *Mental Health*: There has been an acknowledgment of the unique mental health needs of refugees in coping with their past trauma. Apart from loss of networks, most migrants experience downward social and economic mobility at least in the initial years as they go through the process of settlement in their host countries. Migrants in an alien context and culture have to deal with challenges to their core assumptions about the world, 'culture shock,' along with a loss of familiar cultural environments including language and belief systems (Berger and Weiss, 2002: p. 24). The separation, losses, disruptions, and stressors affecting migrant's lives due to the migration process have been conceptualized as trauma (Berger and Weiss, 2002). Social discrimination, economic deprivation, and lack of networks are linked in turn to the incidences of alcohol and substance abuse (Balgopal, 2000). There is an increasing recognition of migration experiences as a determinant of health.
- *Family Relationships*: The bicultural identity conflicts faced by migrants (Segal, 2002) as well as their children (IFSW, 2012) and the tensions that arise in the family due to the value conflicts in multigenerational households can lead to escalated stress and tension. Differences in the migrants' child-rearing practices, including disciplining (Balgopal, 2000; Segal, 2002) and those of the host society are likely to increase stress. Senior immigrants might be dependent on their adult children for a range of support and needs. The loss of traditional networks and ways of living increases their vulnerability to abuse within the family (Potocky-Tripodi, 2002). Marital relationships become

impacted by the changed gender roles, increased demand on women's caring responsibilities, difficulties in balancing provisioning and labor market activities, and isolation (Potocky-Tripodi, 2002; Menjívar and Salcido, 2002).

Challenges

- *Conceptions of Citizenship*: The notion of citizenship includes dimensions of legal status, rights, participation, and belonging (Bloemraad et al., 2008). Countries differ in their visions of who can become a citizen and whether equal citizenship can be conferred upon persons who are ethnically different from the majority host groups; and whether this citizenship is granted at birth (*jus soli*) or granted to the individual on the basis of their parents' citizenship status in the country (*jus sanguinis*). Notions of citizenship also differ in terms of whether the goal of citizenship is furthered better through policies focused on multiculturalism and group rights, or in contexts that promote the shedding of group/community based affiliations toward assimilation into one national identity (Bloemraad et al., 2008). The host country's conceptions of citizenship, whether assimilationism or multiculturalism has a bearing on the reception and accommodations the migrants receive in the host country and their subsequent integration.

The category and class of migration has the potential to determine not only the specific challenges migrants face in their host countries, but also the country's response to the needs of the migrants. Among the migrant categories, refugees are particularly vulnerable in that not only do they come from regions that are linguistically, culturally, and politically very different from their host society, they also have to deal with traumatic preimmigration experiences. Most countries actively disentitle refugees. In the United States, for example, undocumented migrants are ineligible for all federal public assistance programs with the exception of emergency health care and schooling for children (Park and Bhuyan, 2011). Women and other dependents those migrate on conditional residency requirements are particularly vulnerable to abuse at the hands of the sponsor. Seasonal workers have also been identified as a vulnerable migrant category as they are in danger of being exploited in their employments and are often not eligible for the protection and other services available to permanent migrants.

- *Integration Policies*: Countries show great diversity in terms of the policies and programs they have in place for integration of their migrants. These range from those that lean toward assimilation at one end (e.g., Unites States of America) and toward pluralism or multiculturalism at the other (e.g., Canada) (George, 2006; Valtonen, 2008). The host country's policy toward immigrant integration is a very important factor impacting the migrant's adjustment, which can be measured along on a continuum – from acculturation and assimilation at one end to rejection of the culture at the other (Segal, 2002). Countries like Canada, Australia, and New Zealand, for example, have immigration laws in place that allow foreign nationals to apply to enter the country as permanent residents. In the United States, immigration policy is largely concerned only with the management and policing of the entry of migrants into the country and services targeted specifically at integration are available only to refugees (Potocky-Tripodi, 2002). Canada and Israel, among other countries, have numerous services to facilitate the settlement and integration of newcomers. Host countries tend to have different arrangements for who pays the bill for the services available to newcomers – the government or the newcomers themselves.
- *Institutional Readiness*: The settlement and integration of migrants into the host country depends to a large extent on the country's institutional structures – preexisting relationships between races and ethnicities within the population; the labor market and related institutions; government policies that either directly or indirectly impact upon migrants and the place of the host society within the globalized economy (Reitz, 2002) – and reflects "the host society's belief of the place of foreigners in their midst, and more broadly the role of government in civil society" (Siemiatycki and Triadafilopoulos, 2010: p. 21).

Migrants often face difficulties in achieving good economic outcomes in their host countries (OECD, 2013; Buzdugan and Halli, 2009) and often work in jobs that require lower levels of skills (IMO, 2012; Stier and Levenon, 2003). Economic migrants may face challenges in relation to language skills and securing employment that recognizes their past education and work experiences (Valtonen, 2001; Tiiseto, 2011). Requirements related to local work experience, and referrals in the context of limited social networks reduce their access to appropriate employment. Other categories of migrants may have restrictions on working in the host country. Temporary or seasonal migrants may be constrained by conditional work-stay requirements, increasing their vulnerability to exploitation (Amnesty International, 2012).

Another area that exemplifies the importance of institutional readiness within host societies is housing. In Canada, for example, migrants often face difficulties in accessing adequate housing (Teixeira, 2008) and are increasingly represented among the category of absolute homeless, while refugees in particular are among the worst affected groups (Ballay and Bulthuis, 2004). Other barriers that migrants face when attempting to secure housing have been identified as administrative requirements such as having a cosigner as a guarantor; knowledge of neighborhoods and housing systems, lack of skills in conducting a housing search and skills in negotiating with landlords, language and accent barriers, and social housing waiting lists (Hulchanski, 1997).

- *Xenophobia*: Migrants all over the world often face hostility from the native born population in varying degrees and with different consequences (Bunar, 2007; Hale et al., 2011; Henry and Tator, 2005). Migrants are frequently viewed as a threat to the material security and cultural integrity of host society, a sentiment that is often capitalized or even promoted by some right wing- or xenophobic politicians (Mughan and Paxton, 2006; Bunar, 2007). This hostility gets amplified in bad economic times where migrants might be seen as competition for scarce employment opportunities (Mughan and Paxton, 2006). In most countries, the hostility is covert, and displayed through everyday discriminatory acts of limiting the migrants' access to the

equal opportunities enjoyed by the majority of the population (Segal, 2002), often resulting in institutionalized exclusionary practices (Henry and Tator, 2005). In other countries, the hostility toward migrants may be much more pronounced and displayed through intergroup conflicts and violence including hate crimes ranging from verbal abuse to murders (Bunar, 2007; Hale et al., 2011).

- *Underutilization of Services*: The underutilization of service, and health and mental health services, in particular, is a recurring theme in the migration literature (George, 2007; Guruge et al., 2009). There are many reasons why migrants may underutilize services: there may be language and financial (real or perceived) or cultural barriers to accessing services; migrants might lack knowledge of services available and/or their rights to access these. Also, they may not be eligible to access the services based on government policies of their country; and/or there may be fear of disclosure of status accompanying access of services (Murray and Skull, 2005).

Social Work with Migrants

Social workers have a long history of involvement with immigrants and refugees. The foundation of social work is synonymous with the Settlement Houses created in the early and late nineteenth century primarily to serve and help immigrants and relocated persons, first in London in Toynbee Hall and later replicated in the United States by Jane Addams (Brieland, 1990). Social work with migrants has been historically geared toward helping newcomers adjust to their new societies with the aim of enabling them to become 'good citizens' (Balgopal, 2000). In recent years, this focus has become more migrant-centric, with an acknowledgment within the profession of emphasizing the importance of the successful adjustment of the migrant and on the removal of barriers to their participation in the host society (IFSW, 2012).

The demographic profile of migrants has changed significantly over the past decades and social workers now increasingly have to work with persons from very different cultural and ethnic backgrounds. Social work responded to these challenges in the 1970s and 1980s with a call for appropriate interventions, both at the individual and structural levels (Tsang and George, 1998). This included antiracist (Dominelli, 2002; George, 2000) and/or antioppressive approaches (Dominelli, 2002) that emphasize the structural origins of problems (Mullaly, 1997).

Social work intervention with migrants generally occurs through settlement service work with new migrants. These services are targeted at migrants to assist them in meeting the most direct immediate needs of settlement in the host society. Social workers employed in the migrant settlement field are generally involved in providing two kinds of services to migrants. *Direct services* including information and referral related to settlement needs, as well as counseling. For refugees, apart from providing information, social workers have a role to play in mitigating the intensity of trauma experienced through appropriate humanitarian service response (Drumm et al., 2003).

Indirect services include advocacy, planning, and coordination with other organizations to increase service efficiency (Potocky-Tripodi, 2002; IFSW, 2012). As advocates, social workers have to educate the public and government about the needs and challenges faced by migrants as well as the resource needed to address these issues. Social workers have to be aware of immigration policy changes and the manner in which these shifts affect practice (Park and Bhuyan, 2011). An important role that social workers have in the context of *advocacy* is to ensure that migrants have opportunities for equal citizenship (in the broadest sense) in their host country. In a multicultural society, these include access to equal opportunities, the elimination of discrimination, opportunities to practice and transmit culture, and the public acceptance of migrants as legitimate members of society (Valtonen, 2008: p. 144).

Social work intervention needs also to be focused on *preventing* many of the problems faced by migrants. Prevention in social work includes three kinds: preventing problems from occurring in the first place, preventing their escalation, and finally their reoccurrence (Valtonen, 2008). Direct practice and advocacy of the kinds mentioned above would serve to prevent problems faced by migrants in their host societies. Social work also has a role to play in the countries of origin, to address issues of human rights violation, poverty, and underdevelopment, and to ensure an adequate standard of living to prevent persons from leaving the country (Eurodiaconia, n.d.).

Consultations with migrants and service providers have led to the conceptualization of a comprehensive needs-based model for intervention with migrants to meeting both their immediate needs, labor market needs, and ongoing needs that emerge on account of their interactions with the host cultural milieu (George, 2007). The model puts the needs of the migrants at the center of the intervention and proposes that social workers develop expertise at specific levels of intervention. The model suggests three levels of intervention: First, focused on meeting the immediate settlement needs of the migrants; second, focused on helping highly skilled migrants navigate the labor market and its requirements of accreditation and referrals; and third, on providing culturally appropriate intervention to migrants in dealing with issues that arise of the process of settlement such as parent–child conflict, family violence, and mental health (Fong, 2003).

An integrated framework for cross-cultural practice (Tsang and George, 1998) is useful to guide social work intervention with clients from diverse communities. Such a framework avoids essentializing tendencies and emphasizes the need for the appropriate attitudes, knowledge, and skills for cross-cultural practice. Attitudes that are facilitative of good cross-cultural practice include a commitment to justice and equity, the valuing of difference, and openness to cultural difference and learning from the clients and constant critical self-reflection. Effective cross-cultural social work is also characterized by the social worker having knowledge of the systemic context of the host society's culture, of acculturation and internalized culture as well as of the dynamics of cross-cultural communication and understanding. Required skills include the management of one's own emotional response, skills in intervention in institutional contexts, and communication and relationship skills. It is also important for the effective cross-

cultural social worker to utilize research in evaluating the effectiveness of interventions and to identify effective strategies for change (Tsang and George, 1998: p. 7).

Challenges to Social Work Interventions

Interventions with migrants must be informed by an understanding of the manner in which migrant status, in interaction with other markers of identity such as race, gender, ethnicity, and language proficiency, shapes the functioning of migrants. Professional social workers with the knowledge (specialized knowledge in issues related to settlement and an understanding of the intersectional nature of oppressions), skills (in systemic level intervention), and attitudes (an antioppressive framework) could help prevent the problems that migrants face in the settlement and integration process into deteriorating to a point of requiring more intrusive actions.

The dearth of professional social work involvement at the settlement service stage can be attributed to funding limitations; as well as the lack of training for social work with migrants (Yan and Chan, 2010). Few schools of social work offer courses that emphasize skills in working with immigrants and refugees and social workers often lack knowledge of socioeconomic, legal, and advocacy issues surrounding migration (Martinez-Brawley and Zorita, 2010).

Tensions exist between existing dominant discourses targeted at migrants (such as assimilation), and of the principles of social work practice that value self-determination and social justice (Balgopal, 2000). Social workers need to guard against being coerced into surveillance (Humphries, 2004) and policing (Martinez-Brawley and Zorita, 2010: p. 26) of immigrants, and be guided by core values in relation to social justice and integrity in their relationships with migrant clients.

Serving undocumented or 'illegal' migrants presents an ethical dilemma for some social workers (Park and Bhuyan, 2011) as they try to balance the contradictions between serving 'undeserving' populations, and their ethnical obligations to serve all migrants, based on fundamental human rights (Valtonen, 2008; *Official Journal of the European Union*, 2010; United Nations, n.d). Tensions also arise when the policies and laws of the countries stipulate lines of action that are contradictory to client needs or ethical codes of conduct or human rights (Humphries, 2004). The harsh public response to the issue of undocumented workers might lead to a weakened social work response (Valtonen, 2008).

Conclusion

Migration is likely to be an enduring feature of all countries, and especially those that require skilled labor for the advancement of their economy. The problems with settlement especially the labor market integration of migrants are likely to create a class of marginalized individuals in the host societies. Social workers have to address the issues related to the marginalization of this vulnerable group to prevent the loss of human and social capital to the host societies.

See also: Ethnicity and Migration in Europe; Immigration Policy; Immigration, Applied Research On; Immigration, Emigration, and Citizenship; Immigration: Health and Mortality; International Migration; Migrants and Migration in Europe; Migration Out of Europe; Migration and Fertility; Migration and Health; Migration and Social Capital; Migration to and from Latin America; Migration, Ethnicity, Aging and Social Work Practice; Migration: Anthropological Perspectives; Migration: Cultural Aspects; Migration: Motivations; Refugees: Geographical Aspects; Social Work Theory; Social Work and Refugees.

Bibliography

Amnesty International, 2012. Exploited Labour: Migrant Workers in Italy's Agricultural Sector. Amnesty International Press, London.
Balgopal, P., 2000. Social Work Practice with Immigrants and Refugees. Columbia University Press, New York.
Ballay, P., Bulthuis, M., 2004. The changing portrait of homelessness. In: Andrew, C. (Ed.), Our Diverse Cities. Spring, pp. 119–123.
Berger, R., Weiss, T., 2002. Immigration and post traumatic growth-a missing link. Journal of Immigrant and Refugee Studies 1 (2), 21–39.
Bloemraad, I., Korteweg, A., Yurdakul, G., 2008. Citizenship and immigration: multiculturalism, assimilation, and challenges to the nation-state. Annual Review of Sociology 34, 153–179.
Brieland, D., 1990. The Hull House tradition and contemporary social worker: was Jane Addams really a social worker. Social Work 35 (2), 134–138.
Bunar, N., 2007. Hate crimes against immigrants in Sweden and community responses. American Behavioral Scientist 51 (2), 166–181.
Buzdugan, R., Halli, S.S., 2009. Labor market experiences of Canadian immigrants with focus on foreign education and experience. International Migration Review 43 (2), 366–386.
Cabral, V., 2000. Settlement Services for Newcomers and Access to Family Services. The Multicultural Coalition for Access to Family Services, Toronto.
Dominelli, L., 2002. Anti-Oppressive Social Work: Theory and Practice. Palgrave Macmillan, London.
Drumm, R.D., Pittman, S.W., Perry, S., 2003. Social work interventions in refugee camps: an ecosystem approach. Journal of Social Service Research 30, 67–92.
Dustmann, C., Fabbri, F., 2003. Language proficiency and labour market performance of immigrants in the UK. The Economic Journal 113, 695–717.
Fong, R., 2003. Culturally Competent Practice with Immigrant and Refugee Children and Families. The Guildford Press, New York.
George, U., 2000. Toward anti-racism in social work in the Canadian context. In: Calliste, A., Dei, Sefa G.S. (Eds.), Anti-Racist Feminism. Fernwood Publishing, Halifax, pp. 111–122.
George, U., 2006. Immigrant integration: simple questions, complex answers. Canadian Diversite 5 (1), 3–6.
George, U., Chaze, F., 2009. The role of social capital in settlement: employment. Affilia 24, 394–405.
Guruge, S., Berman, R., Murphy-Kilbride, K., Tyyskä, V., Woungang, I., Edwards, S., Clune, L., 2009. Implications of English proficiency on immigrant women's access to & utilization of health services. Women's Health and Urban Life 8 (2), 21–41.
Hale, M., Kransdorf, M., Hamer, L., 2011. Xenophobia in schools. Educational Studies 47, 317–322.
Henry, F., Tator, C., 2005. The Colour of Democracy: Racism in Canadian Society. Thomson Nelson, Toronto.
Humphries, B., 2004. An unacceptable role for social work: implementing immigration policy. British Journal of Social Work 34, 93–107.
Kler, P., 2006. Graduate over education and its effects among recently arrived immigrants to Australia: a longitudinal study. International Migration 44 (5), 93–128.
Lee, E.S., 1996. A Theory of Migration. University of Pennsylvania, Philadelphia.
Li, P., 2003. Destination Canada: Immigration Debates and Issues. Oxford University Press, Toronto.
Martinez-Brawley, E.E., Zorita, P.M.B., 2010. Immigration and social work: contrasting practice and education. Social Work Education: The International Journal 30 (1), 17–28.
Menjivar, C., Salcido, O., 2002. Immigrant women and domestic violence: common experiences in different countries. Gender & Society 16 (6), 898–920.

Mughan, A., Paxton, P., 2006. Anti-immigrant settlement, policy preferences and Populist Party voting in Australia. British Journal of Political Science 36, 341–358.
Mullaly, B., 1997. Structural Social Work: Ideology, Theory, and Practice. Oxford University Press, Ontario.
Murray, S.B., Skull, S.A., 2005. Hurdles to health: immigrant and refugee health care in Australia. Australian Health Review 29 (1), 25–29.
Park, Y., Bhuyan, R., 2011. Whom shall we serve? A discourse analysis of social workers' commentary on undocumented immigrants. Journal of Progressive Human Services 23 (1), 18–40.
Potocky-Tripodi, M., 2002. Best Practices for Social Work with Immigrants and Refugees. Columbia University Press, New York.
Rajan, I.S., 2012. Editorial. Journal of Migration and Development 1 (1), 1–4.
Reitz, J.G., 2002. Host societies and the reception of immigrants: research themes, emerging theories and methodological issues. International Migration Review 36 (4), 1005–1019.
Segal, U., 2002. Framework for Immigration: Asians in the United States. Columbia University Press, New York.
Siemiatycki, M., Triadafilopoulos, T., 2010. International Perspectives on Immigrant Service Provision. Mowat Centre for Policy Innovation Paper. School of Public Policy & Governance, Toronto.
Stier, H., Levenon, V., 2003. Finding an adequate job: employment and income of recent immigrants to Israel. International Migration 41 (2), 81–107.
Suárez-Orozco, M.M., Qin-Hilliard, D.B. (Eds.), 2004. Globalization: Culture and Education in the New Millennium. University of California, Berkeley.
Teixeira, C., 2008. Barriers and outcomes in the housing searches of new immigrants and refugees: a case study of "Black" Africans in Toronto's rental market. Journal of Housing and the Built Environment 23, 253–276.
Tiiseto, R., 2011. I am a lawyer! Am I a lawyer? The experiences of foreign trained black lawyers in Ontario. Race, Gender & Class 18 (3/4), 198–216.
Tsang, A.K.T., George, U., 1998. Towards an integrated model for cross-cultural social work practice. Canadian Social Work Review 15 (1), 73–93.
Valtonen, K., 2001. Social work with immigrants and refugees: developing a participation-based framework for anti-oppressive practice. British Journal of Social Work 31, 955–960.
Valtonen, K., 2008. Social Work and Migration. Ashgate, Surrey.
Yan, M.C., Chan, S., 2010. Are social workers ready to work with migrants? Settlement of newcomers to Canada. Canadian Social Work 12 (1), 16–23.

Relevant Websites

Eurodiaconia, n.d. Policy Paper: Social Services and Migration. http://www.eurodiaconia.org/files/Migration/Policy_Paper_Social_services_and_migration.pdf (accessed 12.10.12.).
George, U., 2007. Immigration and Settlement in Canada. Booksurge Publishers. http://www.amazon.ca/Immigration-Settlement-Canada-Usha-George/dp/1419674528 (accessed 12.10.12.).
Hulchanski, D., 1997. Immigrants and access to housing: how welcome are newcomers to Canada?. In: Metropolis Year II Conference. Summary of Keynote Presentation to the "Housing and Neighborhoods" Workshop ceris.metropolis.net/research-policy/doc%20listS/000.htm (accessed 01.01.13.).
International Federation of Social Workers (IFSW), 2012. Migration. Policy Issue. Retrieved from: http://ifsw.org/policies/migration/ (accessed 01.01.13.).
International Federation of Social Workers, 2012. Migration: Policy Issue. http://ifsw.org/policies/migration/ (accessed 01.01.13.).
International Organization for Migration (IOM), 2011. World Migration Report 2011: Communicating Effectively About Migration. http://publications.iom.int/bookstore/free/WMR2011_English.pdf (accessed 01.01.13.).
Morrison, A.R., Schiff, M., Sjöblom, M., 2008. The International Migration of Women. The World Bank, Washington. http://www.rrojasdatabank.info/migrwomen07/front.pdf (accessed 12.10.12.).
Organization for Economic Cooperation and Development (OECD), 2012. International Migration Outlook. http://www.oecd.org/els/internationalmigrationpoliciesanddata/IMO%202012_Intro%20to%20Stat%20Annex.pdf (accessed 12.10.12.).
Organization for Economic Cooperation and Development (OECD), 2013. International Migration Outlook 2013. http://www.oecd.org/els/mig/imo2013.htm.
United Nations High Commissioner for Refugees (UNHCR), n.d. Refugee Figures: http://www.unhcr.org/pages/49c3646c1d.html (accessed 12.10.12.).
United Nations High Commissioner for Refugees (UNHCR), 2011. UNHCR Global Trends 2011 (accessed 28.08.13.). http://www.unhcr.org/4fd6f87f9.html.
United Nations, n.d. The Universal Declaration of Human Rights. Retrieved from: http://www.un.org/en/documents/udhr/index.shtml (accessed 12.10.12.).
United Nations, 2010. Trends in International Migrant Stock: Migrants by Age and Sex. http://www.un.org/esa/population/publications/CD_MS_2010/CD-ROM%20DOCUMENTATION_UN_Mig_Stocks_2010.pdf (accessed 12.10.12.).

Migration: Anthropological Perspectives

Caroline B Brettell, Southern Methodist University, Dallas, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article traces some of the key developments in the anthropology of migration, focusing particularly on critical trends at the end of the twentieth and the beginning of the twenty-first centuries. Among these are the formulation and analysis of transnationalism; the exploration of processes of inclusion and exclusion, together with new considerations of the meaning of citizenship; the problematization of multiculturalism and hyperdiversity coinciding with a renewed emphasis on the local in relation to the global; and a continued interrogation of the gendered dimensions of migration.

The roots of anthropological research on migrant populations can be found in the early work of urban anthropologists in Africa and Latin America who, in the post-World War II period, began to witness the movement of populations out of rural areas and into the growing cities of the developing world. By the mid-1980s, several key areas of research emerged in a 'new anthropology of migration' that was increasingly focused, although by no means exclusively, on international population flows rather than on internal migration. These areas included the political economy of migration; the social organization of migration, including studies of social networks and immigrant organizations; the relationship between migration and ideology, including gender ideology; and the impact of migration on identity, including ethnic identity.

Although much of the early research on migration that was carried out by anthropologists drew on modernization theory and a bipolar framework of analysis that opposed sending and receiving areas as well as push-and-pull causal factors, these approaches were soon replaced by historical-structural approaches rooted in dependency theory and world systems theory that clearly established the connection between migration and processes of globalization. Additionally, anthropologists began to explore how state policies both shape and are shaped by the international movement of people. The shortcomings of these more macro-level theoretical approaches led, in turn, to the development of theoretical frameworks that could encompass the migrant and/or the migrant household as active agents in the process of migration.

By the early twenty-first century, a large volume of ethnographical work had accumulated from places across the globe, much of it summarized in a host of useful review articles and edited volumes tracing the major contributions of anthropology to the study of migration (see, for example, Brettell, 2008; Fassin, 2011; Foner, 2003; Silverstein, 2005; Vertovec, 2011). The contributions of anthropology to migration studies are many, but since 1990 a number of critical areas have been identifiable. They are the formulation and analysis of transnationalism; an exploration of processes of inclusion and exclusion that has included new considerations of racism and the meanings of citizenship; the problematization of multiculturalism and hyperdiversity coinciding with a renewed emphasis on the local in relation to the global; and a continued interrogation of the gendered dimensions of migration.

The Transnational Dimensions of Migration

At the end of the first decade of the twenty-first century, transnationalism had been identified as "one of the fundamental ways of understanding contemporary migrant practices across the multi-disciplinary field of migration studies" (Vertovec, 2010a, p. 3). Although the transnational framework is central to several disciplines, anthropologists were among the first to recognize its significance. Transnationalism describes a social process, whereby migrants operate in social fields that cross geographic, political, and cultural borders. It therefore offers an alternative to the more linear and one-way theory of assimilation to describe how migrants settle in a host society. From a transnational perspective, migrants are no longer uprooted but are, rather, active participants in both homeland and host societies.

Although certainly a subject of debate (Is it new? Will it last? How widespread is it?), the transnational framework is well suited to the analysis of twenty-first century population movements that are facilitated by improved modes of transportation and impacted by the information and images that can be transmitted rapidly across vast physical distances, thereby reducing the social distance between the homeland and the country of immigration. Anthropologists have conducted research on transnational families and long-distance parenting; transnational political activities and organizations; transnational religious and ritual practices; and the transnational economics associated with remittances and global entrepreneurship. Some anthropologists have addressed the strategic transnational projects of elite migrants (including, by extension, the relationships among migration, transnationalism, and development), but more commonly, the emphasis has been on the everyday transnational actions of working-class migrants as well as on how transnational identities and arrangements may challenge the political and cultural authority of the state as well as state projects concerning immigrant incorporation.

Research on transnational migration has resulted in new multisited methodologies that can encompass the myriad dimensions and spaces that characterize a transnational social field. In an attempt to sort out the complex empirical and methodological reach into transnationalism, Stephen Vertovec (2004, p. 971) has usefully outlined three modes of transformation within three domains: the perceptual, which

involves a migrant's 'orientational bi-focality in the sociocultural domain'; the conceptual, which affects the "meaning of the analytical triad, 'identities-orders-borders' in the political domain"; and the institutional, which affects 'forms of financial transfer, public–private relationships and development in the economic domain.'

Closely associated with work on transnationalism are efforts to re-theorize the concept of diaspora and the diasporic subject. As Pnina Werbner (2004, p. 896) has argued, diasporic communities, as mobilized groups, "are cultural, economic, political and social formations *in process*…. Diasporas are full of division and dissent. At the same time they recognize collective responsibilities, not only to the home country but also to coethnics in far flung places." Anthropologists have explored the process by which diasporas are formed and articulated, whether in real or virtual spaces. Of particular interest is the way in which diasporas nurture collective memory and social solidarity with coethnics dispersed across the globe. Consideration of diaspora has also generated ethnographic work on memory and the meaning of home and/or the idea of return. It has also brought the study of refugees into the arena of the anthropology of migration in order to assess what forcibly displaced populations share with those classified as economic migrants as well as where their experiences differ. In addition, some anthropologists have conducted research into the role of the sending society state in the lives of diasporic populations, promoting the maintenance of cultural practices and religious observances, facilitating remittances that are fundamental to sending society economies, or even establishing, as in the case of Haiti, an official political status for their nationals living abroad.

In summary, the anthropological work on transnationalism and diasporic communities emphasizes the way in which identities are negotiated and constructed simultaneously in local, national and global spaces. Transnationalism encompasses displacement and emplacement, deterritorialization and reterritorialization, the transgression and redefinition of borders, and the construction and deconstruction of boundaries. The transnational and diasporic practices of migrants are both influenced by and in turn influence broader processes of global change. Both are likely to continue and continue to be the object of anthropological inquiry.

Claiming Inclusion: The Politics of Citizenship and Belonging

The transnational framework has resulted in a host of new and innovative approaches to the concept of citizenship, a topic with which anthropologists have only recently become engaged. As Michel Laguerre (1998, p. 177) has observed, "given the fact that these transnational practices occur on an everyday basis, it seems that our concept of citizenship is no longer consistent with the domain of practice." Laguerre formulates the concept of diasporic citizenship to capture the "situation of the individual who lives outside the boundaries of the nation state to which he or she had formerly held primary allegiance and who experiences through transnational migration….the subjective reality of belonging to two or more nation-states" (pp. 12–13). Other anthropologists working on migrant populations write of transborder citizenship, flexible citizenship, racialized citizenship, or polycultural citizenship, all with the goal of emphasizing that citizenship is more than simply a political status or a set of rights and obligations between individuals and the state. Rather it is "a total relationship, inflected by identity, social positioning, cultural assumptions, institutional practices and a sense of belonging" (Werbner and Yuval-Davis, 1999, p. 4). Thus, a critical question for anthropologists is how citizenship is claimed or negotiated by migrants who are often confronted with acts of exclusion whether more formally by nation states or more informally in the context of everyday interactions (Reed-Danahay and Brettell, 2008).

To capture this sense of belonging, and of citizenship as practice, several anthropologists have operationalized a concept of cultural citizenship that has been defined on the one hand as a process of subjectification and citizen-making and, on the other hand, as a process whereby immigrants can claim space and inclusion. As an example of the first approach, Aihwa Ong (2003, p. 15) shows how Cambodian Americans experience American citizenship by illustrating the "social policies and practices beyond the state that in myriad mundane ways suggest, define, and direct adherence to democratic, racial and market norms of belonging." She suggests that it is in the spaces of encounter, "in the practices directed at newcomers, and the mutual daily interactions that ensue, that the meaning and exercise of citizenship happens" (p. 16). By contrast, the second approach to cultural citizenship highlights the more agentic dimensions of belonging – that is, claiming "the right to be different (in terms of race, ethnicity, or native language) with respect to the norms of the dominant national community, without compromising one's right to belong, in the sense of participating in the nation-state's democratic processes" (Rosaldo and Flores, 1997, p. 57). Anthropologists have drawn on this formulation of cultural citizenship to explore various forms of political activism and immigrant civic engagement as well as dimensions of immigrant youth culture. Much of this work emphasizes that the expression of citizenship among immigrants and their offspring is fluid and changing, multidimensional, subjective, and often contentious.

Confronting Exclusion: Racism, National Identity, and Constructions of the Immigrant 'Other'

Anthropologists have recognized, documented, and theorized the various ways in which migrants are made into citizens or make claims to citizenship and hence inclusion, but at the same time, they have also explored concomitant processes of exclusion. Here they have focused attention on the reception of immigrants, particularly how immigrants are constructed as being 'foreign' and 'other' in public discourse. Beginning in the 1990s, particularly in the new host societies of southern Europe, foreigners were often represented as morally degenerate criminals and were frequently associated with drug smuggling and sexual trafficking. But such processes of negative stereotyping and othering have been identified as equally characteristic of the United States, a country with a deep history of immigration. Thus, the Cuban refugees who arrived as part of the Mariel boatlift were stigmatized as criminals, homosexuals, and communists, while Haitians were stigmatized as

carriers of HIV/AIDS. Needless to say, racism was also in play in relation to both these populations, the Marielito Cubans being decidedly darker than the first wave of Cuban refugees who fled to the United States in the early 1960s.

In the European context, anthropologists have drawn attention to a 'new racism' based on cultural difference (Silverstein, 2005; Vertovec, 2011), particularly in relation to migrants of Islamic background (Mandel, 2008), but also in relation to those seeking political asylum (Fassin, 2005). In his analysis of the 2004 law in France that banned the wearing of head scarves in public schools, John Bowen (2007) argues not only that critical dimensions of French identity (secularism and communalism) were and continue to be at the center of this debate but also that the law itself illustrates the "particularly French passion for seeking statutory solutions to social ills" (p. 243). Bowen, as well as several other anthropologists working on the subject of immigrants in Europe, pay particular attention to the role of the media in constructing particular representations of foreigners, especially Muslims. This role includes, in Bowen's view, one in which Muslims are accorded the right to speak (p. 246) and hence what kind of Muslim is found to be acceptable in a country (in this case, France) that emphasizes assimilation rather than multiculturalism.

In the United States, it is not the so-called Islamic threat but the Latino threat (Chavez, 2008) that anthropologists have identified as a primary motivation for anti-immigrant discourse and processes of exclusion. Equally important are the debates over illegality, which some anthropologists have formulated as a state of abjectivity – being cast away as low, contemptible, and undeserving beings (Gonzales and Chavez, 2012). In her ethnographic study of undocumented Salvadorans in the United States, Susan Bibler Coutin (2003, p. 29 ff) captures this abject status in her reference to the spaces of nonexistence that this population occupies. 'The undocumented therefore exist in a non-domain, a space of illegality. Often conflated with criminals, unauthorized immigrants are officially outside both the law and the social body.' Based on his research on immigration to Bahrain, Gardner (2010, p. 52) has observed that "attention to illegality, deportation, and removal focuses the anthropological lens on the processes by which states seek to control the movement of people in particular… Illegality, deportation and other forms of removal are tools the state uses to control and govern the transnational movement of people." All of this work (see also Ticktin's (2011) discussion of the *sans-papiers* in France), ultimately rooted in the theoretical frameworks of governmentality formulated by Michel Foucault, reflects the ways in which the anthropology of the state intersects with the anthropology of migration. Further, it allows anthropologists to investigate the relationship between migration and constructions of national identity. What makes people American, or French, or Italian, or Australian, or Canadian is often given salience in debates that circulate around the immigrant other.

Cities, Hyperdiversity, and the Multicultural Project

The individual, the household, and the state are all important units of analysis in the anthropology of migration. But recently, anthropologists have refocused their attention on the urban context where the majority of migrants settle, thereby returning in some sense to the roots of the anthropology of migration in the development of urban anthropology. Anthropologists have emphasized a 'city as context' perspective not only as a way of framing a more comparative approach to the study of processes of migrant settlement but also as a way of linking macro and micro processes. By focusing on the city as a unit of analysis, anthropologists can explore the nuances and dynamics of intergroup relations as well as how race and ethnicity are constructed and negotiated in everyday life and everyday practice. Cities differ, for example, in the historical depth of their experience with immigrants, in the diversity of the immigrant populations that settle there, in their geographical characteristics, in their local economies, and in their political structures.

Emerging from this renewed interest in studying the city as one studies migration is a focus on how the scale of a city impacts processes of migrant incorporation and transnational connection. This approach is offered as an alternative to the methodological nationalism reflected in studies that emphasize ethnic groups, ethnic institutions, and nation-states. Rather than to view cities simply as 'containers, providing spaces in which migrants settle and make a living,' the city-scale approach highlights how migrants 'actively contribute to the restructuring and repositioning of either their cities of settlement or those to which they are transnationally connected' (Glick Schiller and Caglar, 2011, p. 2, 3). Migrants, from this perspective, are both subjects and agents of the global processes, including neoliberal restructuring, that reposition cities and other local places. Those who have adopted this scalar approach are calling for "a theory-inspired comparative agenda that will advance conceptual and empirical knowledge on urban and migrant dynamics" (Glick Schiller and Caglar, 2011, p. 19).

Another area of interest that has emerged from the renewed interest in cities as units of analysis in the anthropological study of migration is that of super- or hyperdiversity within national and particularly urban environments. Vertovec (2010b, p. 66–67) describes this approach as drawing on previous work concerning multiethnic arenas of interaction, residential segregation, notions of pluralism, and the complex and multilayered experience of migrants with unequal power structures and social locations. It is an approach that looks at diversity from a variety of perspectives by considering its multiple dimensions and the range of factors that influence people's lives in local places. Of particular note is that multiculturalism from this perspective can be viewed as a "kind of diversity management strategy" (p. 68).

A recent example of ethnographical work that examines how people come together in a multi-ethnic city to construct a collective life is Beth Epstein's (2011) study of the state-planned new town of Cergy in the suburbs of Paris. Multiculturalism is at the center of her analysis. "In Cergy, as in other cities in France, the question of how best to integrate people of foreign origin into the local design is being posed in relation to larger anxieties about how best to build and sustain a cohesive social fabric in a period marked by a preeminent concern for the problems of insecurity and social dissolution…. The discursive shift, led by members of self-styled minority groups….[draws] attention to the way integration can act as

a discriminating force that selects and chooses, even as it strives to create an ostensibly universal, cohesive whole" (Epstein, 2011, p. 162–163). Epstein offers us what she labels a "cultural analysis of the politics of integration" (p. 19).

Some anthropologists have framed multiculturalism positively as a policy that respects diversity and difference, whereas others have emphasized the negative dimensions in relation to a 'moral panic' about an 'excess of alterity' (Grillo, 2010). The work of Norwegian anthropologist Unni Wikan (2002) is notable here. She argues that the multicultural model and its 'excessive tolerance' for difference has resulted in a 'generous betrayal' of immigrant newcomers. Culture, in her view, has become a bit like race, a concept that subverts human rights, particularly those of women and children, because it supports ethnic differences and identity politics. Writing about Norwegians, but one could extend this category, Wikan suggests that they regard one another "as individuals with a different character and the ability and will to think for themselves." By contrast, she argues, "immigrants are largely perceived as products of culture. They are perceived as caught in the grips of culture and therefore unable to exercise independent judgment… It is disrespectful and really quite degrading. Unfortunately, immigrants themselves often contribute to such degradation. By constantly invoking 'culture' as explanation (and excuse) for their behavior, they belittle themselves as acting, thinking, willful human beings, and they run down the very qualities that have brought them here: initiative, courage, perseverance" (Wikan, 2002, p. 81). Wikan offers a powerful critique of the concept of culture and of a multicultural model of integration that ignores intracultural variability. Several other European ethnographers have followed Wikan's lead by noting and critically analyzing the expanded use of the concept of culture in public discourse in relation to debates about immigrants and ethnic minorities (see Vertovec, 2011).

In a more recent book, *In Honor of Fadime*, Wikan (2008) pursues this interest by offering a nuanced analysis of the so-called honor killing, in 2002, of a young Kurdish woman born and raised in Sweden. She juxtaposes the argument that many immigrants have not left their respective countries voluntarily and hence have not made a conscious choice to abandon their cultural traditions, with the rights of children who are born abroad and whose identities and cultural choices may be more in line with those of liberal western democracies. Governments of host countries must, in her view, be sensitive to the conflicts and human tragedies that can emerge from this situation and to the fact that culture is neither static nor inherited but, rather, is composed of a multiplicity of views (including those of parents and children) and is always changing. Her conclusion, applied to Norway but with broader implications, is that "the critique of the culture concept that is engaging anthropology must be applied with equal force to the situation of immigrants…. Otherwise the result will be racism – wielded in the name of charity" (2002, p. 83).

Gender and Migration

The attention that Wikan's work draws to honor killing and, for that matter, the larger debates about cultural practices (veiling, female circumcision, polygyny) that are out of place in European host societies, underscore the importance of the gendered dimensions of migration. Interest in the relationship between migration and gender emerged in the 1970s with the development of feminist anthropology and coincided with the growing feminization of migration that has been noted by demographers. Anthropological studies of gender and migration emphasize how men and women are both gendered actors in the migration process, equally affected by global gendered processes and gendered institutional structures. This corpus of research pays particular attention to how patriarchal structures can affect work, family life, associational life, welfare state policies, and labor recruitment policies in both the sending and the receiving societies. Gender ideologies can impact the possibility of migration, the processes of incorporation, and how individuals (men and women) engage in transnational space.

Two ethnographic examples illustrate the significant variation in gendered dimensions of migration. Mexico is characterized by a gendered culture of migration. Despite the increasing number of female migrants to the United States, Mexican men and women continue to define migration flows as primarily male because that is what is compatible with local understandings of gender and masculinity (Malkin, 2004). Further, while el Norte is defined as the place of work, Mexico is defined as the place of family and of life. By contrast, in Sri Lanka it is primarily women who depart to work as housemaids in the Middle East and men who remain behind and take over child care and other kinship tasks (Gamburd, 2008). Approximately 10% of Sri Lankan households have an adult female who is working overseas. "Labor flows from sending countries that depend on family structures that facilitate migration, and demands for labor in host countries rely on cultural assumptions about the gender, race, religion, and other characteristics of the ideal worker" (Gamburd, 2008, p. 7). In Sri Lanka, family structure (and we might argue, by extension, gender relations) is less the result of traditional forms of organization than of constant adjustments to changing circumstances, locally, nationally, and globally. When these Sri Lankan women breadwinners return from abroad with their experiences of cosmopolitanism, they often face backlash and are subjected to accusations of prostitution and marital infidelity while abroad. As a result they resist trying to change local hierarchies once they return in order to subvert these accusations and the sometimes violent behavior that accompanies them.

The literature concerning gender, migration, and the family also documents how gender roles and relationships that have been constructed in particular and differing cultural contexts are either reproduced or transformed in the immigrant context. A number of scholars, for example, have explored how the increasing economic responsibilities of immigrant women have impacted gender hierarchies within the family. Commonly, marital conflicts and tensions result from the discrepancy between the traditional patriarchal ideology that may be held by husbands and the important economic contributions of their wives to the maintenance of the immigrant household. In many cases, however, women's challenges to male authority within the family are subtle, and generally economic resources are pooled. In other cases, a more

companionate form of marriage emerges as a result of migration (Hirsch, 2003).

Finally, it is important to emphasize the new work on gender and migration that intersects with medical anthropology and with humanitarian concerns. A number of anthropologists have focused on women's health issues; on the exclusionary policies related to the reproduction, including the move away from birthright citizenship in Ireland or the controversies over 'anchor babies' in the United States; on the relationship between migration and the spread of HIV/AIDS in various parts of the world; and on violence against women. In a particularly intriguing approach to these issues, Ticktin (2011, p. 129) demonstrates how "sexual violence renders legitimate the claim for papers and basic rights" in a broader context that is blatantly anti-immigrant. Ticktin's larger argument is that a regime of care plays an important role in the politics of immigration such that battered women or immigrants who are sick can make legitimate claims for crossing borders while those who are simply fleeing poverty cannot.

Conclusion

Anthropological research on population flows, whether internal or international, is diverse in the questions that are pursued and the methodologies that are deployed. What is common to all this work is a sensitivity to place and context as well as an awareness of the importance of documenting and analyzing the multifaceted ways in which people in local places respond to global processes. Anthropologists remain attentive to the importance of culture, which means that they highlight the relationships between beliefs and behaviors; the interactions among the multiple dimensions (economic, religious, political, and social) of the human experience; and the importance of issues of identity that are fluid and constantly negotiated. Anthropologists of migration also remain attentive to the comparative approach that has been central to their discipline and thus pursue questions not only of what is common to the migrant experience, no matter where migrants originate from and what their destination might be, but also of what is unique and why. Finally, it is important to emphasize that the anthropology of migration intersects with other anthropologies, for example, with feminist anthropology, the anthropology of the state, the anthropology of globalization, and medical anthropology. These intersections enrich the theoretical frameworks and empirical dimensions of the anthropological study of migrant populations.

See also: Citizenship, Historical Development of; Multiculturalism, Anthropology of; Urban Anthropology.

Bibliography

Bowen, John.R., 2007. Why the French Don't Like Headscarves: Islam, the State, and Public Space. Princeton University Press, Princeton.

Brettell, Caroline B., 2008. Theorizing migration in anthropology: the social construction of networks, identities, communities and globalscapes. In: Brettell, C.B., Hollifield, J.F. (Eds.), Migration Theory: Talking across Disciplines, second ed. Routledge Taylor and Francis Group, New York, pp. 113–159.

Chavez, Leo R., 2008. The Latino Threat: Constructing Immigrants, Citizens and the Nation. Stanford University Press, Palo Alto, CA.

Coutin, Susan B., 2003. Legalizing Moves: Salvadoran Immigrants' Struggle for US Residency. University of Michigan Press, Ann Arbor, MI.

Epstein, Beth S., 2011. Collective Terms: Race, Culture, and Community in a State-Planned City in France. Bergahn Books, New York and Oxford.

Fassin, Didier, 2005. Compassion and repression: the moral economy of immigration policies in France. Cultural Anthropology 20, 362–387.

Fassin, Didier, 2011. Policing borders, producing boundaries: the governmentality of immigration in dark times. Annual Review of Anthropology 40, 213–226.

Foner, Nancy (Ed.), 2003. American Arrivals: Anthropology Engages the New Immigration. School of American Research Press, Santa Fe, NM.

Gamburd, Michele R., 2008. Milk teeth and jet planes: kin relations in families of Sri Lanka's transnational domestic servants. City and Society 20, 5–31.

Gardner, Andrew M., 2010. City of Strangers: Gulf Migration and the Indian Community in Bahrain. Cornell University Press, Ithaca, NY.

Glick Schiller, Nina, Caglar, Ayse (Eds.), 2011. Locating Migration: Rescaling Cities and Migrants. Cornell University Press, Ithaca, NY.

Gonzales, Roberto, Chavez, Leo R., 2012. "Awakening to a nightmare": abjectivity and illegality in the lives of undocumented 1.5 generation Latino immigrants in the United States. Current Anthropology 53, 255–281.

Grillo, Ralph, 2010. An excess of alterity? Debating difference in a multicultural society. In: Vertovec, S. (Ed.), Anthropology of Migration and Multiculturalism: New Directions. Routledge, New York, pp. 19–38.

Hirsch, Jennifer S., 2003. A Courtship after Marriage: Sexuality and Love in Mexican Transnational Families. University of California Press, Berkeley, CA.

Laguerre, Michel S., 1998. Diasporic Citizenship: Haitian Americans in Transnational America. St. Martin's Press, New York.

Malkin, Victoria, 2004. "We go to get ahead": gender and status in two Mexican migrant communities. Latin American Perspectives 31, 75–99.

Mandel, Ruth, 2008. Cosmopolitan Anxieties: Turkish Challenges to Citizenship and Belonging in Germany. Duke University Press, Durham, NC.

Ong, Aihwa, 2003. Buddha is Hiding: Refugees, Citizenship, and the New America. University of California Press, Berkeley, CA.

Reed-Danahay, Deborah, Brettell, Caroline B. (Eds.), 2008. Citizenship, Political Engagement, and Belonging. Rutgers University Press, New Brunswick, NJ.

Rosaldo, Renato, Flores, William V., 1997. Identity, conflict, and evolving Latino communities: cultural citizenship in San Jose, California. In: Flores, W.V., Benmayor, R. (Eds.), Latino Cultural Citizenship: Claiming Identity, Space and Politics. Beacon Press, Boston, MA.

Silverstein, Paul A., 2005. Immigration, racialization, and the new savage slot: race, migration and immigration in the new Europe. Annual Review of Anthropology 34, 363–384.

Ticktin, Miriam, 2011. Casualties of Care: Immigration and the Politics of Humanitarianism in France. University of California Press, Berkeley, CA.

Vertovec, Steven, 2004. Migration and other modes of transnationalism: towards conceptual cross-fertilization. International Migration Review 38, 970–1001.

Vertovec, Steven, 2010a. Introduction: new directions in the anthropology of migration and multiculturalism. In: Vertovec, Steven (Ed.), Anthropology of Migration and Multiculturalism: New Directions. Routledge, New York and London, pp. 1–17.

Vertovec, Steven, 2010b. Super-diversity and its implications. In: Vertovec, Steven (Ed.), Anthropology of Migration and Multiculturalism: New Directions. Routledge, New York and London, pp. 65–95.

Vertovec, Steven, 2011. The cultural politics of nation and migration. Annual Review of Anthropology 40, 241–256.

Werbner, Pnina, 2004. Theorising complex diasporas: purity and hybridity in the south Asian public sphere in Britain. Journal of Ethnic and Migration Studies 30, 895–911.

Werbner, Pnina, Yuval-Davis, Nira, 1999. Introduction: women and the new discourse of citizenship. In: Yuval-Davis, Nira, Werbner, Pnina (Eds.), Women, Citizenship and Difference. Zed Books, London and New York, pp. 1–38.

Wikan, Unni, 2002. Generous Betrayal: Politics of Culture in the New Europe. University of Chicago Press, Chicago, IL.

Wikan, Unni, 2008. In Honor of Fadime: Murder and Shame. University of Chicago Press, Chicago, IL.

Migration: Cultural Aspects

Marcelo M Suárez-Orozco, Graduate School of Education and Information Studies, University of California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the twenty-first century migration is the human face of globalization. All continents are experiencing mass migration – as sending, transit, or destination areas, and often, all three at once. This article reviews the basic scholarship on cultural concerns in mass migration with a focus on families, language, and inequality, as they structure the pathways of immigrants in a variety of destinations.

Introduction

As we enter the second decade of the twenty-first century, the lives over a billion people are shaped by the experience of migration. All continents are involved in the massive movement of people, as areas of immigration, emigration, or transit – and often as all three at once. In the twenty-first century, immigration is the human face of globalization – the sounds, colors, and smells of a miniaturized, interconnected, and fragile world. But viewed anthropologically, immigration is constitutive of the human experience. Migration is written in our genetic code and is encoded in our bodies: in our bipedalism, in our stereoscopic vision, in our neocortex. Modern humans are the children of immigration, and migrations today are transforming humanity. During second decade of the twenty-first century, 232 million are international migrants (see http://data.worldbank.org/indicator/SM.POP.TOTL), approximately 740 million as internal migrants (UNDP 2013 http://www.un.org/en/development/desa/population/publications/pdf/technical/TP2013-1.pdf) and millions more as immediate relatives left behind (UNDP, 2010). Only China (1.36 billion) and India (1.4 billion) have larger populations than this immigration nation.

Mass migrations have ushered in the era of hyperdiversity. The global cities of the twenty-first century now resemble the breathtaking cultural and linguistic diversity once associated with the highlands of New Guinea or the Amazonian Mato Grosso. There are now over a dozen global cities with more than 1 million immigrants – from Hong Kong to Melbourne, from Moscow to Singapore, and in more than two dozen global cities, immigrants now account for more than one-quarter of the population, inter alia, in Amsterdam, Auckland, Muscat, and Perth (see http://www.migrationinformation.org/datahub/gcmm.cfm#map1list).

Globalization increases inequality (Picketty, 2014) and structures migratory flows by coordinating markets, economies, social practices, and cultural models. Demographic and ecological factors also structure mass migration. The number of international migrants continues to grow at a steady rate, and shall likely pass the 400 million mark by 2050, "as a result of growing demographic disparities, the effects of environmental changes, new global political and economic dynamics, technological revolutions and social networks" (International Organization for Migration 2010: p. xix). Lastly, catastrophic migrations related to war and environmental collapses are at an all time high since the second war world. According to the UN refugee agency "the number of refugees, asylum-seekers and internally displaced people worldwide has, for the first time in the post-World War II era, exceeded 50 million people. http://www.unhcr.org/53a155bc6.html"

Viewed anthropologically, immigration involves change in residency and change in community. Scholarly research has specialized in two broad types of large-scale migration: internal migration (within the confines of a nation-state or territory) and international migration (across international borders). Although the large-scale movement of people within a nation-state is a phenomenon of a separate order from mass migrations across international borders, internal migrants share many characteristics with international migrants: most move from rural villages to urban centers, many experience linguistic and cultural dislocations, racialization, and face isomorphic bureaucratic and legal restrictions. Much scholarly and policy attention has been focused on international migration. Yet most migrants are internal migrants staying within the confines of their nation-states. India and China, following their insertion into global capitalism, have experienced unprecedented levels of internal migration. With more than 320 million migrants in India and over 200 million in China, the two countries combined now have more than double the number of international migrants the world over. Some of the most important anthropological contributions to the study of immigration have focused on internal migration (for example, Colson, 1971; Brandes, 1975; Scudder and Colson, 1982; Morgan and Colson, 1987).

Mass migration is reshaping economies and societies the world over, yet many facets of immigration remain overlooked, misunderstood, or neglected. The dominant approaches to the study of mass migration privilege labor factors (Piore, 1980); economic variables (Borjas, 1999); demographic forces (Passel and Cohn, 2009); and, in recent years, border controls, security (Andreas, 2009), and undocumented immigration (Massey et al., 2002; Chavez, 1997). Though these approaches have generated some scholarly consensus, important aspects of mass migration remain unexplored. In this article, I first examine cultural concerns in immigration *qua* the family, and then I turn to concerns over immigration and language. This by necessity will lead us to examine current debates in the educational literature on the nexus between immigration, language acquisition, and education. Lastly, I will briefly examine concerns over immigration and inequality through the lenses of poverty, segregation, unauthorized status, and racialization.

Immigration and the Family

The unit of migration is the family – variously defined in different parts of the world and structured by distinct culturally coded legislative, economic, reproductive, and symbolic forms. Immigration is an ethical act of and for the family. Distinct patterns of kinship, household, family, and social organization are of paramount importance in structuring worldwide migratory waves (Massey, 1990). Immigration typically starts with the family and family bonds sustain it. Immigration profoundly challenges and changes families as well as the societies in which immigrants settle (Foner, 2009).

Historically, the pattern in stepwise migration was of the male going ahead, establishing a beachhead while sending remittances home. Over time, when financially possible the process of bringing relatives – wife, children, and others – begins. But in recent decades, immigration has achieved a nearly perfect gender balance with approximately half of all global migrants being women. In the high- and middle-income countries, demand for service workers draw women – many of them mothers – to care for children. In rapidly aging countries it is equally true that these immigrant workers are summoned to care for aging folks. Immigrant care labor is thus present at the start and the end of the life cycle.

With large-scale migration, transnational families have entered the global cities of the twenty-first century. For millions 'home' – as a cultural ideal and *qua* the practices of the household may be in two countries. Anthropologists have pioneered the scholarly study of migrants shuttling back and forth in transnational labor and living arrangements the world over (Basch et al., 1994). Transnational migrants remain substantially engaged (economically, politically, and culturally) in their newly adopted lands and in their communities of origin, moving back and forth in ways not always seen in previous eras of large-scale immigration (Suárez-Orozco and Suárez-Orozco, 2001). Millions of children throughout the world are being raised transnationally.

The transnational lives of immigrant families have only recently come into scholarly focus. On the basis of a study of schools in the states of Nuevo Leon and Zacatecas, Hamann and Zuñiga (2008) estimated that as much as 2% of students enrolled in Mexico's *primarias* (primary schools) and *secundarias* (high schools) had transnational experiences, including considerable schooling in the United States. Approximately half of the transnational youth in this study were born in the United States or had a US-born mother. The authors suggest that perhaps 1%, or as many as 220 000, youth being educated in Mexico, are in fact US citizens living transnational lives.

Millions of other children are educated in culturally patterned long-distance family arrangements where separations and reunifications across national boundaries are normative. In a recent study conducted in Montreal, among 254 first- and second-generation immigrant-origin youth from the Philippines and the Caribbean, approximately 62% of the Filipino-origin participants and 38% of the Caribbean-origin participants had experienced separations during their migration (Rousseau et al., 2009). In a nationally representative survey of 1772 youth that restricted its sample to documented immigrants, nearly one-third of the participants between ages 6 and 18 years had been transnationally separated from at least one parent for two or more years (Gindling and Poggio, 2009). Notably, the rates of separation were highest for children of Latin American origin, which happen to be more than half of all migrants to the United States. This is a low estimate since separation rates are higher among the unauthorized or those who are in the process of regulating their documentation status.

The Longitudinal Immigrant Adaptation Study (LISA) found high rates of transnational family separations among new immigrant students. This US bicoastal interdisciplinary study was conducted with 400 recently arrived immigrant youth from China, the Dominican Republic, various countries in Central America, Haiti, and Mexico recruited from public schools; Suárez-Orozco et al. (2008) found that nearly three-quarters of the youth were separated from one or both of their parents during the migration process. There were significant differences between immigrant groups in regards to family separation: Chinese families were least likely to be separated over the course of migration (52%), whereas the vast majority of Central American (88%) and Haitian children (85%) were separated from either one or both of their parents during the course of migration. Approximately 26% of children in the study were separated from both parents, a pattern most often occurring in Central American families (54%). The length of separation from parents was unexpectedly long: some individuals in the study reported being separated from one or both parents for nearly their entire childhood, though the length of separation varied widely across regions of origin. Of the youth who were separated only from their mothers, 54% of Central American children endured separations lasting 4 years or more, as did approximately one-third of both the Dominican and Haitian families. Chinese and Mexican children underwent fewer and shorter separations from their mothers. When separations from the fathers occurred during migration, they were often very lengthy or permanent ones. For those families who were separated, 28% had separations from fathers that lasted more than 4 years. Notably, the LISA study found that immigrant youth who underwent protracted family separations and complicated reunifications were at higher risk to show a pattern of declining academic achievement over time once they resettled in the new country (Suárez-Orozco et al., 2010a).

Recent scholarly work has identified a related pattern of 'satellite children' born in the Untied States to immigrant parents but raised by grandparents in China. A study of immigrants in New York City estimated that more than 50% of the mothers had sent their youngsters to Fujian Province to be raised by grandparents. The mothers reported undocumented status, substantial immigration-related debt burdens, and exhausting 6-day work weeks as the main reasons for raising their children transnationally (Kwong et al., 2010, Yoshikawa, 2011). Many of these transnationally raised children, the mothers expected, would return to the United States to complete their formal education.

New research suggests that migrations resulting in protracted separations threaten the identity and cohesion of the family, transforming well-established roles, creating new loyalties and bonds, and destabilizing cultural scripts of authority, reciprocity, and responsibility (Suárez-Orozco and Suárez-Orozco, 2011). Even under the best of circumstances, the family is never the same after migration. One family

commences the migratory journey and 'another,' now reconstituted family, completes it.

Immigration, Language, and Schooling

The children of immigrants are the fruit borne of immigration (Suárez-Orozco and Suárez-Orozco, 2001). Immigrant-origin children are a fast-growing sector of the youth population in diverse countries the world over, including Australia, Canada, Germany, Italy, New Zealand, the Netherlands, Spain, and Sweden (Meir and Morehouse, 2008). In the United States, the country with the largest number of immigrants in the world (at more than 42 million in 2014), approximately one-quarter of all youth are of immigrant origin (more than 16.6 million in 2010), and it is projected that by 2020 one in three of all children will be growing up in immigrant households (Frey, 2011). The transition of immigrant-origin children to an ever-larger list of immigrant-dependent countries is a topic of scholarly interest, policy relevance, and practical urgency.

As families migrate, their languages move with them, leading to greater linguistic and cultural diversity. The new hyperdiversity challenges modernist ideologies linking identity and belonging to a single national language and cultural tradition. In the United States, immigrants speak a combined total of more than 460 different languages (Kindler, 2002). In New York City, approximately half of all children and youth originate in immigrant-headed households speaking an estimated combined 176 different languages (see http://www.nytimes.com/2010/04/29/nyregion/29lost.html?_r1). In Los Angeles Unified School District, the nation's second largest, over 62% of the children and youth are English Language learners (see http://dq.cde.ca.gov/dataquest/lc/DistrictLC.aspx?cYear=2013-14&cSelect=1964733–LOS+ANGELES+UNIFIED).

How to best develop the means to communicate in the dominant language, thereby developing advanced academic language skills required for higher-order cognitive work and the matter of maintaining immigrant languages are fiercely debated the world over (McAndrew, 2007). Language acquisition and academic trajectories are bound together with processes of identity formation, acculturation, assimilation, and economic integration (Suárez-Orozco, 1989, 1991; Portes and Zhou, 1993; Portes and Fernández-Kelly, 2008; Portes and Hao, 1998, 2002; Suárez-Orozco et al., 2008).

Two constants have defined the relationship between immigration and language. First, every wave of large-scale immigration in multiple destinations generates a fear that new immigrants will not learn the dominant language. Second, the children of immigrants from widely divergent origins inevitably gravitate toward the new language (Tse, 2001; Portes and Hao, 2002) and, over time and across generations, lose their native language skills (Fillmore, 1991). Yet these mutually canceling concerns miss the defining element of mass migration in the era of globalization: Academic language learning as a cultural construct is changing what is required in terms of language acquisition for this generation of immigrants. Most large-scale studies of immigrants depend on self-reported data of oral proficiency, which differs from the decontextualized, written language necessary for higher-order cognitive engagement and academic success (Cummins, 2000).

Academic language proficiency is central to any understanding of educational trajectories of immigrant youth in high-income countries with strict accountability regimes of the high-stakes-testing variety. Scholarly research has shown a high correlation between proficiency in academic language skill and academic achievement as measured by standardized tests (Valdes, 2004; Suárez-Orozco et al., 2008). First- and second-generation immigrant youth are often found to score lower than their native-born peers on standardized tests in the United States as well as in European contexts (Barth et al., 2008). The science results of PISA's 2006 study by the Organization for Economic Development and Cooperation (OECD) of 57 countries revealed that, "[i]n fact, first-generation immigrant students in OECD countries lag on average more than 50 lconsidering the OECD average difference in performance between school years" (Christensen and Segeritz, 2008: pp. 13–14). Likewise, the 2006 math and reading results show that "the achievement gaps for immigrant and nonimmigrant students in reading and mathematics in 2006 are similar to those found in science in 2006" (Christensen and Segeritz, 2008: p. 15). Although there are cross-country differences in the academic trajectories of immigrant-origin youth (Christensen and Segeritz, 2008) and although, over time, some immigrant-origin youth do remarkably well (Kasinitz et al., 2008), the general trend is worrisome, especially as the share of immigrant-origin students continues to grow in a number of countries.

Research on second-language acquisition and bilingualism informs debates on educational models that promote success for immigrant youth, especially which language should be used for instruction, under what circumstances, and for how long. One recurring finding (Cummins, 2000; Collier, 1995; August and Hakuta, 1997; Hakuta et al., 2000) is that it takes approximately five to seven years for immigrant language learners to develop the academic language proficiency required to compete fairly with native speakers in standardized-assessment regimes at the center of education reforms the world over. Another consistent finding suggests that 'balanced biliguals' that is, youth who maintain their home language as they acquire a second academic language, tend to have better educational trajectories over time (Portes and Rumbaut, 2001).

Immigration and Inequality

In the age of linguistic and cultural hyperdiversity, there is considerable heterogeneity in socioeconomic status between and within immigrant groups. Immigrants today, unlike in previous eras of mass migration, are among the most highly educated and skilled workers in their new countries. Never before in world history have so many highly educated and skilled people joined the migration flow in such high numbers. This is true in the United States, as well as in Canada, Australia, New Zealand, and other destinations, where immigrant-origin scientists, engineers, and medical doctors are overrepresented as a total share of the population (see Migration Information Source 2010; http://www.migrationinformation.org/Feature/display.cfm?ID 779).

Immigrants seldom originate in the poorest regions of the world, yet research suggests that they are more likely than their

native counterparts to encounter poverty in their new societies (Kazemipur and Halli, 2001). Poverty, defined as a 'condition of relative deprivation within society' (Kurtz 1973: p. 53), is created in part by unique disadvantages that are often associated with immigrant status, such as "language barriers, incompatibility of educational credentials, limited transferability of job skills, unfamiliarity with the market demands, and lack of access to job and educational networks" (Kazemipur and Halli, 2001: p. 1132; de Haan and Yaqub, 2009).

Concerns over the transgenerational transfer of poverty among immigrants have lead to research examining poverty in immigrant families and children. The children of immigrants have "greater market-income poverty rates than children in native-born families" (Hernandez et al., 2010: p. 425). In affluent countries worldwide, poverty among children of immigrants has increased steadily in recent years with gaps between native-born and immigrants ranging from 7% in Australia and Germany to 12% in the United States and to 26–28% in England and France (Hernandez et al., 2010). Differences among ethnic groups are also prevalent. US data on Latino children show that, in 1999, 22.8% were living in poverty, compared with 7.7% of whites (Therrien and Ramirez, 2000). In 2006, however, the poverty rate for Latino children had nearly doubled that of white native-born children (28% and 16%, respectively) (Fry and Gonzales, 2008). For immigrant Latino families, poverty rates reach higher percentages, with 35% of foreign-born Latino immigrants living in poverty compared with 27% of their second- or third-generation counterparts (Fry and Gonzales, 2008).

In the United States, children of immigrants are four times as likely as native-born children to live in crowded housing conditions and three times as likely to be uninsured, and 37% of Latino immigrant families report difficulties affording food (Capps, 2001). Children raised in poverty are vulnerable to an array of distresses including difficulties concentrating and sleeping, anxiety and depression, as well as a heightened exposure to delinquency and violence. Poverty has long been recognized as a significant risk factor for poor educational outcomes (Weissbourd, 1996; Luthar, 1999).

Poverty coexists with a variety of other factors that augment risks, such as single parenthood; residence in violence-ridden neighborhoods, gang activity, and drug trade; as well as school environments that are segregated, overcrowded, understaffed, and poorly funded. Poverty, however, is not solely an inner-city phenomenon. In the United States since the 1990s, poverty grew in the suburbs for the first time and at a greater rate than in cities (Murphy, 2010). Poverty is also associated with high rates of housing mobility and concurrent school transitions that can be highly disruptive to educational performance (Gándara and Contreras, 2009).

Immigrant origin Latino children are the most segregated students in US schools (Orfield and Lee, 2005). Immigrants who settle in predominantly minority neighborhoods may have little, if any, direct, continous, and intimate contact with peers from the nonimmigrant mainstream population. A pattern of triple segregation – by race, language, and poverty – shapes the lives of many new immigrants in varied countries.

In Germany, immigration accounts for a significant share of the population. In 2005, approximately 15.3 million people were immigrants or of migrant background. Immigrants are largely segregated, with most immigrant-origin people concentrating in the western part of the country and only a small share living in the eastern states (Clauss and Nauck, 2010).

Povery and segration are often compounded by unautorized status. The United Nations estimates that there are between 40 and 50 million unauthorized migrants worldwide (UNDP, 2010). The United States has the largest concentration of undocumented immigrants in the world – approximately 11 million people are unauthorized. There are approximately 1.1 million youth living in the United States without proper documentation, and millions more are living in households headed by at least one undocumented immigrant (see http://pewhispanic.org/reports/report.php?ReportID61). Research suggests that undocumented youth often arrive after multiple family separations and traumatic border crossings (Suarez-Orozco, 2014 – http://www.pass.va/content/dam/scienzesociali/pdf/ninosmigrantes.pdf); they may continue to experience fear and anxiety about being apprehended, separated again from their parents, and deported (Suárez-Orozco et al., 2008; Chaudry et al., 2010).

A large proportion of undocumented workers are employed in low-paying professions with erratic working conditions (Capps, 2001; Gándara and Contreras, 2009; Yoshikawa, 2011). Unauthorized migrants do not access social services that could serve to mitigate the harshest conditions of their poverty (Yoshikawa, 2011). Psychological and emotional duress takes a toll on the experiences of youth raised in the shadow of the law (Cervantes et al., 2010), which has also been documented through narrative and qualitative research (Bagley and Castro-Salazar, 2010; Suarez-Orozco, 1989; Gonzalez, 2009).

Protracted poverty, deep segregation, and unauthorized status are the ingredients for immigration distopia in many immigrant-affected societies. At the very least, structural factors accelerate the processes of racialization via cultural disparagement (De Vos and Suárez-Orozco, 2000) and negative 'social mirroring' (Suárez-Orozco, 2004) of new immigrants of color.

The arrival of large numbers of immigrants into the high-income countries of the world has resulted in isomorphic patterns of residential segregation and can be related to broader dynamics of racialization of 'marked' native minorities. The idea of 'race,' the stepchild of nineteenth-century pseudoscience, included in these earlier iterations craniometrics (measuring of skulls) as a means of classifying groups of people (Mosse 1964: p. 89) into fixed hierarchical relationships of superiority and inferiority (Gould, 1996; De Vos and Wagatsuma, 1972). As Boas (1940) observed, "a stratification of society in social groups that are racial in character will always lead to racial discrimination. As in all other sharp social groupings the individual is not judged as an individual but as a member of his class" (pp. 16–17).

Racialized hierarchies lent the aura of science to atavistic superstitions that clustered people of 'purer' origin to avoid the danger of contamination by outsiders (Zollschan cited in Mosse, 2000: pp. 201–202). As applied to immigrants, 'racialization' (Bashi, 1998; Omi and Winant, 1994; Shih, 2008) is an appropriate construct foregrounding the sociohistorical processes of segregation, marginalization, micro- and macro-aggressions, and collective disenfranchisement.

In countries such as the United States, large and growing numbers of poor immigrants of color and the undocumented are *de facto* relegated to spaces where socially constructed phenotype (Gilman, 2000; Bonilla-Silva, 2004) aligns with entrenched patterns of segregation and marginalization of native minorities (Corni, 2002; Ogbu and Simmons, 1998; Vigil, 2002). Portes and Zhou (1993; Portes, 1996) have appropriately termed this dynamic as 'segmented assimilation' wherein certain immigrants join the marginalized space of native minorities creating what they term a new 'rainbow underclass.' Meanwhile, other countries have witnessed the emergence of a complex interweaving of factors racializing immigrants and other minorities. These can include religion (France and its North African minorities), country of origin (Japan and its Brazilian, Korean, and other minorities), and social class (Spain and its Roma minority).

In the currency of racialization, dark-skinned immigrants appear to pay a hefty penalty in the US education system. Murguia and Telles (1996) working with the National Chicano Survey report that 10.2% of light-skinned Mexican Americans completed college as opposed to 5.3% among dark-skinned Mexican Americans. Massey and Denton (1989) analyzing US census data report that it will be more difficult for darker-skinned Hispanics than for their lighter-skinned counterparts to integrate into white communities. Color and poverty are correlated with a host of other factors shaping educational trajectories.

Restricted mobility (Massey and Denton, 1989), negative social mirroring (Suárez-Orozco, 2004), and inequitable access to quality education (Orfield and Lee, 2005) among immigrants are not isolated US phenomenona, as the cases of Koreans and other minorities in Japan (Lee and DeVos, 1981; Castro-Vazquez, 2009) and of Europe and its Muslim minorities attest.

To paraphrase Lévi-Strauss, immigration is 'good to think' and interrogate schemas and practices of the modernist nation-state. Transnational displacements bring race, language, and identity to the forefront, interrupting the taken-for-granted cultural schemas and social practices that structure belonging to and membership in the nation. Indeed, the institutions of state-making in the modernist era are geared toward producing citizens, workers, and consumers of a nation-state rather than dual citizens with global sensibilities. Reimagining the narrative of beloging and recalibering the instutitions of the nation state are a *sine-qua-non* to move beyond the current global immigration distopia worldwide.

See also: Assimilation of Immigrants; Cultural Evolution: Overview; Cultural Geography; Migration History; Migration, Economics of; Migration, Theory of.

Bibliography

Andreas, P., 2009. Border Games: Policing the U.S.-Mexico Divide. Cornell University Press, New York.

Bagley, C., Castro-Salazar, R., 2010. Envisioning undocumented historias: evoking a critical performance ethnography. In: Hillyard, S. (Ed.), New Frontiers in Ethnography: Studies in Qualitative Methodology. Emerald Grp, Bingley, UK, 11, pp. 141–160.

Barth, H., Heimer, A., Pfeiffer, I., 2008. Integration through education: promising practices, strategies and initiatives in ten countries. In: Morehouse, C. (Ed.), Immigrant Students Can Succeed: Lessons from Around the Globe. Carl Bertelsmann Prize 2008. Verl. Bertelsmann-Stiftung, Gutersloh, Germany, pp. 119–188.

Basch, L., Schiller, N.G., Szanton, C., 1994. Nations Unbound: Transnational Projects, Postcolonial Predicaments and Deterritorialized Nation-states. Routledge, London.

Bashi, V., 1998. Racial categories matter because racial hierarchies matter: a commentary. Ethnic and Racial Studies 21 (5), 959–968.

Boas, F., 1940. Race. Language and Culture. Collier-Macmillan, London.

Bonilla-Silva, E., 2004. From bi-racial to tri-racial: towards a new system of racial stratification in the USA. Ethnic and Racial Studies 17 (6), 931–950.

Borjas, G., 1999. Economic Research on the Determinants of Immigration: Lessons for the European Union. The World Bank, Washington, DC.

Brandes, S.H., 1975. Migration, Kinship, and Community: Tradition and Transition in a Spanish Village. Academic Press, New York.

Colson, E., 1971. The Social Consequences of Resettlement: The Impact of the Kariba Resettlement upon the Gwembe Tonga. Kariba Studies, No. 4. Univ. Manchester Press, Manchester.

Capps, R., 2001. Hardship among children of immigrants: findings from the 1999 National Survey of America's Families. Urban Institute B29, 1–8.

Castro-Vazquez, G., 2009. Immigrant children from Latin America at Japanese schools: homogeneity, ethnicity, gender and language in education. Journal of Research in International Education 8, 57–80.

Cervantes, J.M., Mejia, O.L., Mena, A.G., 2010. Serial migration and the assessment of "extreme and unusual psychological hardship" with undocumented Latina/o families. Hispanic Journal of Behavioral Sciences 32 (2), 275–291.

Chavez, L., 1997. Shadowed Lives: Undocumented Immigrants in American Society. Wadsworth, Boston, MA.

Chaudry, A., Capps, R., Pedroza, J., Castaneda, R.M., Santos, R., Scott, M.M., 2010. Facing Our Future: Children in the Aftermath of Immigration Enforcement. Urban Institute, Washington, DC. http://www.urban.org/publications/412020.html.

Christensen, G., Segeritz, M., 2008. An international perspective on student achievement. In: Morehouse, C. (Ed.), Immigrant Students Can Succeed: Lessons from Around the Globe. Carl Bertelsmann Prize 2008. Verl. Bertelsmann-Stiftung, Gutersloh, Germany, pp. 11–36.

Clauss, S., Nauck, B., 2010. Immigrant and native children in Germany. Child Indicators Research 3, 477–501.

Corni, G., 2002. Hitler's Ghettos: Voices from a Beleaguered Society 1939–1944. Arnold, London.

Cummins, J., 2000. Language, Power, and Pedagogy: Bilingual Children in the Crossfire. Multilingual Matters, Buffalo, NY.

de Haan, A., Yaqub, S., 2009. Migration and poverty: linkages, knowledge gaps and policy implications. Presented at UNRISD/ION/IFS Workshop Social Policy and Migration in Developing Countries, Stockholm.

de Vos, G., Wagatsuma, H., 1972. Japan's Invisible Race: Caste in Culture and Personality. University of California Press, Berkeley, CA.

Fillmore, L.W., 1991. When learning a second language means losing the first. Early Childhood Research Quarterly 6 (3), 323–346.

Foner, N., 2009. Across Generations: Immigrant Families in America. NY Univ. Press, New York.

Frey, W.H., 2011. America's Diverse Future: Initial Glimpses at the U.S. Child Population from the 2010 Census. Metropolitan Policy Program, Brookings Inst., Washington, DC. http://www.brookings.edu/~/media/Files/rc/papers/2011/0406_census_diversity_frey/0406_census_diversity_frey.pdf.

Fry, R., Gonzales, F., 2008. One-in-five and Growing Fast: A Profile of Hispanic Public School Students. Pew Hispanic Center Rep. Pew Research Center, Washington, DC. http://www.hawaii.edu/hivandaids/The%20Hispanic%20Population%20in%20the%20US%20%20%20March%202000.pdf.

Gándara, P., Contreras, F., 2009. The Latino Education Crisis: The Consequences of Failed Social Policies. Harvard University Press, Cambridge, MA.

Gilman, S.L., 2000. Are Jews white? or the history of the nose job. In: Back, L., Solomos, J. (Eds.), Theories of Racism. Routledge, London, pp. 229–237.

Gindling, T.H., Poggio, S., 2009. Family Separation and the Educational Success of Immigrant Children. Maryland Policy Brief. University of Maryland Press, Baltimore.

Gonzalez, R., 2009. Young Lives on Hold: The College Dreams of Undocumented Immigrants. College Board, New York, NY.

Gould, S., 1996. The Mismeasure of Man. Norton, New York.

Hamann, E.T., Zuniga, V., 2008. Transnational students in Mexican schools. Anthropology News 49 (5), 19.

Hakuta, K., Butler, Y.G., Witt, D., 2000. How long does it take English learners to attain proficiency? Policy Report 2000-1. The University of California Linguistic Minority Research Institute, Santa Barbara.

Hernandez, D.J., Macartney, S.E., Blanchard, V.L., 2010. Children of immigrants: family and socioeconomic indicators for affluent countries. Child Indicators Research 3, 413–437.
Kasinitz, P., Mollenkopf, J., Waters, M., Holdaway, J., 2008. Inheriting the City: The Children of Immigrants Come of Age. Russell Sage, New York.
Kazemipur, A., Halli, S.S., 2001. Immigrants and 'new poverty': the case of Canada. International Migration Review 35 (4), 1129–1156.
Kindler, A., 2002. Survey of the states' limited English proficient students and available educational programs and services: 2000–2001 Summary Report. National Clearinghouse for English Language Acquisition and Lang. Instruction Education Programs, 2002, Washington, DC.
Kurtz, D.V., 1973. The rotating credit association: an adaptation to poverty. Human Organization 32 (1), 49–57.
Kwong, K., Chung, H., Sun, L., Chou, J.C., Taylor-Shih, A., 2010. Factors associated with reverse migration separation among a cohort of low-income Chinese immigrant families in New York City. Social Work in Health Care 48 (8), 798–811.
Lee, C., DeVos, G., 1981. Koreans in Japan: Ethnic Conflict and Accommodation. Univ. Calif. Press, Berkeley, CA.
Lee, S.J., 2005. Up Against Whiteness: Race, School, and Immigrant Youth. Teachers Coll. Press, New York.
Luthar, S.S., 1999. Poverty and Children's Adjustment. Sage, Thousand Oaks, CA.
Massey, D.S., Spring, 1990. Social structure, household strategies, and the cumulative causation of migration. Population Index 56 (1), 3–26.
Massey, D.S., Durand, J., Malone, N., 2002. Beyond Smoke and Mirrors: Mexican Migration in an Era of Economic Integration. Sage, New York.
Massey, D.S., Denton, N.A., 1989. Hypersegregation in the U.S. metropolitan areas: black and hispanic segregation along five dimensions. Demography 26 (3), 373–391.
McAndrew, M., 2007. The education of immigrant students in a globalized world. In: Suarez-Orozco, M.M. (Ed.), Learning in the Global Era: International Perspectives on Globalization and Education: policy debates in contemporary perspective. University of California Press, Berkeley, pp. 232–255.
Meir, J., Morehouse, C., 2008. Introduction, Immigrant Students Can Succeed: Lessons from Around the Globe. Carl Bertelsmann Prize 2008. Verl. Bertelsmann-Stiftung. Gutersloh, Germany.
Morehouse, C., 2008. Immigrant Students Can Succeed: Lessons from Around the Globe. Carl Bertelsmann Prize 2008. Verl. Bertelsmann-Stiftung. Gutersloh, Germany.
Morgan, S.M., Colson, E., 1987. People in Upheaval. Center for Migration Studies, 1987.
Mosse, G.L., 1964. The Crisis of German Ideology: Intellectual Origins of the Third Reich. Grosset & Dunlap, New York.
Mosse, G.L., 2000. The Jews: myth and counter-myth. In: Back, L., Solomos, J. (Eds.), Theories of Racism: A Reader. Routledge, London, pp. 195–205.
Murguia, E., Telles, E., 1996. Phenotype and schooling among Mexican Americans. Sociology of Education 69 (4), 276–289.
Murphy, A.K., 2010. The symbolic dilemmas of suburban poverty: challenges and opportunities posed by variations in the contours of suburban poverty. Sociological Forum 25 (3), 541–569.
Ogbu, J., Simons, H., 1998. Voluntary and involuntary minorities: a cultural-ecological theory of school performance with some implications for education. Anthropology and Education Quarterly 29 (2), 155–188.
Omi, M., Winant, H., 1994. Racial Formation in the United States. Routlege, New York.
Orfield, G., Lee, C., 2005. Why Segregation Matters: Poverty and Education Inequality. Harvard University Press, Cambridge, MA.
Passel, J., Cohn, V., 2009. A Portrait of Unauthorized Immigrants in the United States. Pew Hispanic Center, Washington, DC.
Picketty, Th, 2014. Capital in the Twenty-First Century. Harvard University Press, Cambridge, MA.
Piore, M., 1980. Birds of Passage: Migrant Labor and Industrial Societies. Cambridge Univ. Press, New York.
Portes, A., 1996. Children of immigrants: segmented assimilation and its determinants. In: Portes, A. (Ed.), The Economic Sociology of Immigration: Essays on Networks, Ethnicity, and Entrepreneurship. Russel Sage Found, New York, pp. 248–280.
Portes, A., Fernandez-Kelly, P., 2008. No margin for error: educational and occupational achievement among disadvantaged children of immigrants. Annals of American Academy of Political and Social Science 620 (1), 12–36.
Portes, A., Hao, L., 1998. E pluribus unum: bilingualism and loss of language in the second generation. Sociology Education 71, 269–294.
Portes, A., Hao, L., 2002. The price of uniformity: language, family and personality adjustment in the immigrant second generation. Ethnic and Racial Studies 25 (6), 889–912.
Portes, A., Rumbaut, R., 2001. Legacies: The Story of the Immigrant Second Generation. Univ. Calif. Press, Berkeley, CA.
Portes, A., Zhou, M., 1993. The new second generation: segmented assimilation and its variants. Annals of American. Academy of Political and Social Science 530, 74–96.
Rousseau, C.I., Hassan, G., Measham, T., Moreau, N., Lashley, M., Castro, T., 2009. From the family universe to the outside world: family relations, school attitude, and perception of racism in Caribbean and Filipino adolescents. Health & Place 15 (3), 751–760.
Scudder, T., Colson, E., 1982. From welfare to development: a conceptual framework for the analysis of dislocated people. In: Hansen, A., Oliver-Smith, A. (Eds.), Involuntary Migration and Resettlement: The Problems and Responses of Dislocated People. Westview, Boulder, CO, pp. 267–287.
Shih, S., 2008. Comparative racialization: an introduction. Pub. Modern Language Association 3 (5), 1348–1362.
Suárez-Orozco, C., 2000. Identities under siege: immigration stress among the children of immigrants. In: Suárez-Orozco, M.M. (Ed.), Cultures under Siege: Collective Violence and Trauma. Cambridge Univ. Press, Cambridge, MA, pp. 194–226.
Suárez-Orozco, C., 2004. Formulating identity in a globalized world. In: Suárez-Orozco, M.M., Qin-Hilliard, D. (Eds.), Globalization: Culture & Education in the New Millennium. Univ. Calif. Press, Berkeley/Los Angeles, CA, pp. 173–202.
Suárez-Orozco, C., Suárez-Orozco, M., 2001. Children of Immigration: The Developing Child Series. Harvard Univ. Press, Cambridge, MA.
Suárez-Orozco, C., Bang, H.J., Onaga, M., 2010a. Contributions to variations in academic trajectories amongst newcomer immigrant youth. International Journal of Behavioral Development 34 (6), 500–510.
Suárez-Orozco, C., Suárez-Orozco, M., 2011. Transnationalism of the heart: familyhood across borders. In: Cere, D., McClain, L. (Eds.), What Is Parenthood? Competing Models for Understanding Today's Revolution in Parenthood. Cambridge Univ. Press, London.
Suárez-Orozco, C., Suárez-Orozco, M., Todorova, I., 2008. Learning a New Land: Immigrant Students in American Society. Belknap, Cambridge, MA.
Suárez-Orozco, M., 1989. Central American Refugees and U.S. High Schools: A Psychosocial Study of Motivation and Achievement. Stanford Univ. Press, Stanford, CA.
Suárez-Orozco, M., 1991. Migration, minority status and education: European dilemmas and responses in the 1990s. Theme Issue Anthropology and Education Quarterly 22 (2), 99–199.
Suárez-Orozco, M., 2014. Las Tres Caras de Herodes: Éxodo de Criaturas, Migraciones Catastróficas y Vida en Sombras. Multidisciplinary Journal of Educational Research 4 (3). http://goo.gl/l9d82w.
Therrien, M., Ramirez, R.R., 2000. The Hispanic population in the United States: population characteristics. Current Population Reports. US Census Bureau, Washington, DC, pp. 20–535. http://www.hawaii.edu/hivandaids/The%20Hispanic%20Population%20in%20the%20US%20March%202000.pdf.
Tse, L., 2001. Why Don't They Learn English? Separating Fact from Fallacy in the U.S. Language Debate. Teachers Coll. Press, Williston, UT.
United Nations Population Division (UNDP), 2010. Mobility and Migration. United Nations Development Programme. http://hdr.undp.org/sites/default/files/nhdr_migration_gn.pdf.
Valdes, G., 2004. Between support and marginalisation: the development of academic language in linguistic minority children. In: Brutt-Griffler, J., Varghese, M. (Eds.), Bilingualism and Language Pedagogy. Multilingual Matters, Clevedon, UK, pp. 10–40.
Vigil, J.D., 2002. A Rainbow of Gangs: Street Cultures in the Mega-City. University of Texas Press, Austin.
Weissbourd, B., 1996. The Myth of the Vulnerable Child. Addison-Wesley, Reading, MA.
Yoshikawa, H., 2011. Immigrants Raising Citizens: Undocumented Parents and Their Children. Russel Sage, NY.

Migration, Economics of

George J Borjas, Harvard Kennedy School, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Migration is the move from one geographic location to another. Residential migration occurs when the household (or person) changes its place of residence by moving from one neighborhood to another within the same locality. Internal migration occurs when the household moves across larger geographically distinct units – such as counties, metropolitan areas, states, or provinces – but remains within the same country. International migration occurs when the household moves across national boundaries.

The study of migration, the movement of workers from one geographic location to another, lies at the core of labor economics because the analysis of labor flows is a central ingredient in any discussion of labor market equilibrium. Workers respond to regional differences in economic outcomes by voting with their feet. These labor flows improve labor market efficiency. This article summarizes the determinants and consequences of both internal migration (labor flows within a given country) and international migration (labor flows across countries).

Migration and Economic Efficiency

The study of migration lies at the core of labor economics because the analysis of labor flows – whether within or across countries – is a central ingredient in any discussion of labor market equilibrium. Workers respond to regional differences in economic outcomes by voting with their feet. These labor flows improve labor market efficiency.

Suppose there are two regional labor markets in a particular country, the North and the South, and that these two markets employ workers of similar skills. Suppose further that the current wage in the North exceeds the wage in the South.

Under some conditions, the wage differential between the two regions will not persist once the economy attains a competitive national equilibrium. After all, the wage differential encourages some Southern workers to pack up and move to the North, where they can earn higher wages and presumably attain a higher level of utility. The flow of Southern workers into the North would raise the Southern wage and depress the Northern wage. If there were free entry and exit of workers in and out of labor markets, the national economy would eventually be characterized by a single wage.

The single wage property of competitive equilibrium has important implications for economic efficiency. The theory of labor demand shows that the wage equals the value of marginal product of labor in a competitive market. As workers move to the region that provides the best opportunities, they eliminate regional wage differentials. Therefore, workers of given skills have the same value of marginal product of labor in all markets. The allocation of workers to firms that equates the value of marginal product across markets is an efficient allocation because it maximizes national income.

To see why a competitive equilibrium is efficient, suppose that a benevolent dictator takes over the economy and that this dictator has the power to dispatch workers across regions. In making allocation decisions, this dictator has one overriding objective: to maximize the country's income. When the dictator first takes over, he sees that the wage in the North exceeds the wage in the South. This wage gap implies that the value of marginal product of labor is greater in the North than in the South.

The dictator picks a worker at random. Where should this worker be allocated? Because the dictator wants to maximize national income, the worker will be sent to the region where he is most productive, the North. In fact, the dictator will keep allocating workers to the Northern region as long as the value of marginal product of labor is greater in the North than in the South. The law of diminishing returns implies that as the dictator forces more and more people to work in the North, the value of marginal product of Northern workers declines. In the end, the dictator will maximize national income only when the value of marginal product of workers is the same in all labor markets.

To easily illustrate this fundamental result, consider the representation in Figure 1. The North (N) is an industrialized region with relatively few workers, while the South (S) has a large population and lags behind the North in economic development. Initially, the wage in the North, W_N, exceeds the wage in the South, W_S.

The two regions have downward sloping labor demand curves that are assumed to have the same slope, but may have different intercepts. The difference between the intercepts measures the economic value of the different infrastructures, with the North's infrastructure allowing any specific worker to have a higher value of marginal product in the North than in the South. Consider a short-run situation where the infrastructure is fixed in each region, so that the height of each labor demand curve is fixed. Finally, assume that all labor is supplied inelastically in each of the regions, with supply curves S_N and S_S.

Figure 1 illustrates the initial labor market equilibrium. Because of the initial supply imbalance and because of the gap in the value of marginal product for any specific worker, the wage in the North is higher than the wage in the South. Income-maximizing workers in the South wish to move to the North to take advantage of the higher wage.

Suppose migration is costless. Workers then move from the low-wage to the high-wage region as long as any wage difference exists, and the flow stops only when the wage in the two

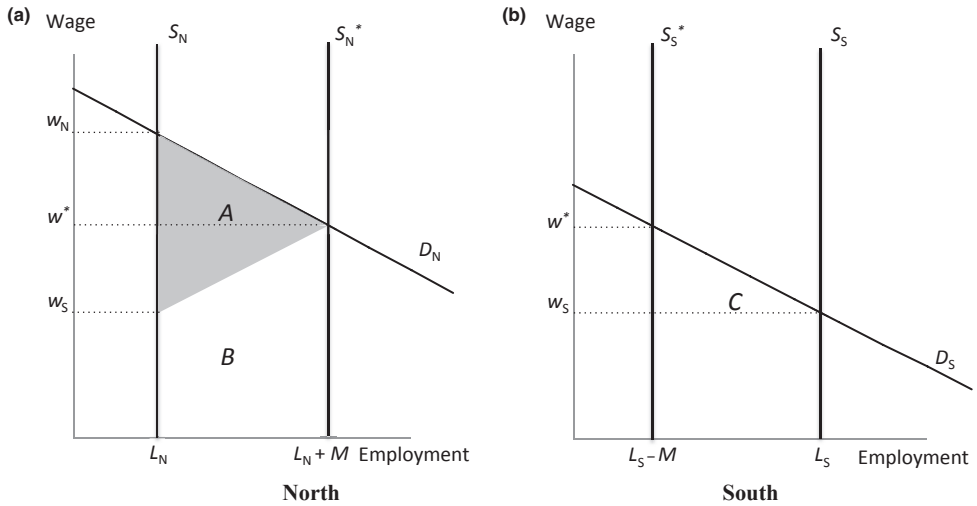

Figure 1 The gains from migration.

regions is equalized at w^*. Each point on the labor demand curve gives the value of marginal product of the last worker hired. As a result, the area under the demand curve gives the total product of all workers hired. The exit of M workers from the South generates a loss in the South's gross domestic product (GDP) equal to the area of the trapezoid C. At the same time, the entry of M workers in the North increases the North's GDP by the size of the area of the trapezoid $A + B$. Note that the South's loss in GDP is also given by the (mirror-image) trapezoid B in the graph illustrating conditions in the North's labor market. The increase in national income, therefore, is given by the triangle A.

In sum, migration and economic efficiency are closely linked in a competitive economy. Through an 'invisible hand,' workers who search selfishly for better opportunities accomplish a goal that no one in the economy had in mind: an efficient allocation of resources.

An Economic Model of Migration

In 1932, Sir John Hicks argued that "differences in net economic advantages, chiefly differences in wages, are the main causes of migration." Practically all modern studies of migration use this hypothesis as the point of departure and view the migration of workers as a type of human capital investment (Sjaastad, 1962). Workers calculate the value of the opportunities available in each of the alternative labor markets, net out the cost of making the move, and choose whichever option maximizes the net present value of lifetime income.

Suppose there are two labor markets where a particular worker can be employed. The worker is currently employed in region i and is considering a move to region j. The worker, who is t years old, earns w_{it} dollars. If he were to move, he would earn w_{jt} dollars. It costs C dollars to move from i to j. These migration costs include the actual expenditures incurred in transporting the worker and his family, as well as the dollar value of the 'psychic cost' – the pain and suffering that inevitably occurs when one moves away from family, neighbors, and social networks.

Like all other human capital investments, migration decisions are guided by the comparison of the present value of lifetime earnings in the alternative opportunities. The net gain to migration is given by:

$$\text{Net Gain} = \sum_{k=t}^{T} \frac{w_{jk} - w_{ik}}{(1+r)^{k-t}} - C \qquad [1]$$

where r is the discount rate and T is the age of retirement. The worker moves if the net gain is positive. A number of empirically testable propositions follow immediately from this framework:

1. An improvement in the economic opportunities available in the destination increases the net gains to migration, and raises the likelihood that the worker moves.
2. An improvement in the economic opportunities at the current location decreases the net gains to migration, and lowers the probability that the worker moves.
3. An increase in migration costs lowers the net gains to migration, and reduces the likelihood of a move.

In sum, migration occurs when there is a good chance that the worker will recoup his human capital investment. As a result, migrants will tend to gravitate from low-income to high-income regions and the larger the income differential between the regions or the cheaper it is to move, the greater the number of migrants.

This framework has been extended to address a slightly different, but equally important, question: *which* persons in the source region are most likely to move? The answer to this question is particularly important for evaluating the impact of migration flows. Suppose that 10% of region i's population chooses to move to region j. The economic impact of this migration on both the source and destination regions will depend on whether the migrants are randomly chosen from the source region's population or are chosen from a particular tail of the skill distribution.

Differences in migration costs across workers help sort out the migrants from the source region's population. Other things being equal, those workers who find it easier to move are the ones who will, in fact, move.

The shape of the income distributions in the two regions will also help sort out the migrants – even if all workers faced the same migration costs (Borjas, 1987; Grogger and Hanson, 2011). Suppose that the skills of workers are portable across the two regions – so that employers value the same types of workers in each of the labor markets. Higher skills typically lead to higher earnings, but the rate of return to skills typically varies across labor markets. Two types of selection may characterize the migrant flow:

1. Positive selection occurs when the migrants have above-average skills. The migrant flow from i to j is positively selected when the destination offers a higher rate of return to skills. The migrants are then drawn from the upper tail of the skill distribution because region i, in a sense, 'taxes' high-skill workers and 'insures' less-skilled workers against poor labor market outcomes.
2. Negative selection occurs when the migrants have below-average skills. The migrant flow is negatively selected when the source region offers a larger payoff to skills. Few skilled workers will then want to move from region i.

In short, as long as regional income differences (net of migration costs) are large enough to induce migration, highly skilled workers will naturally gravitate to those regions where the rate of return to skills is high. In the optimal sorting of workers to regions, highly skilled workers live in regions that offer high rates of return to skills and less-skilled workers live in regions where the rate of return to skills is relatively low.

The simple model in eqn [1] stresses how alternative income streams determine the migration decision for a particular worker. However, it is the alternative streams of utility that determine the worker's decision. After all, the worker cares not only about relative incomes but also about the amenities and disamenities offered by the various regions.

The fact that migration maximizes utility introduces a number of interesting twists into the study of migration decisions. For instance, eqn [1] ignores why there are regional wage differences in the first place, implicitly assuming that the national labor market is in disequilibrium (in the sense that different regions offer different opportunities to the same worker). However, regional wage differences may partly reflect compensating wage differentials that reward workers for the varying set of amenities that different regions offer (Roback, 1982). The wage would then be relatively lower in more pleasant localities. Even though a particular worker might face different wages in different labor markets, the worker's utility would be constant across labor markets. The wage differentials that are the focus of the human capital approach – and that determine the migration decision in eqn [1] – are the ones that persist after the analysis has controlled for regional differences in the value of amenities and disamenities (*see* Wage Differential and Structure).

Internal Migration

Despite problems of comparability, the available data indicate that there is a great deal of internal migration in developed economies. In Ireland and Japan, the 1-year migration rate across local areas (counties in Ireland, prefectures in Japan) is between 2 and 3%, compared to a 1-year migration rate of over 5% in the United States.

Determinants of Internal Migration

A large empirical literature attempts to determine which variables best explain the size and direction of internal migration flows (see the survey in Greenwood, 1997). A few variables have been found to be good predictors in many countries.

Age

Migration is most common among younger workers. The human capital model provides a simple explanation for this pattern. Older workers have a shorter period over which they can collect the returns to the migration investment. The shorter payoff period decreases the net gains to migration, and hence lowers the probability of migration.

Education

Migration is more common among more educated workers. This correlation could arise because highly educated workers may be more efficient at learning about employment opportunities in alternative labor markets, thus reducing migration costs. It is also possible that the geographic region that makes up the relevant labor market is larger for more educated workers. Consider, for example, the labor market faced by college professors. There are few firms in any given city and professors' skills are highly portable across colleges and universities. In effect, college professors sell their skills in a national – or even an international – labor market.

Distance

Greater distances deter migration because greater distances imply larger migration costs. The relationship between distance and migration was an integral part of the 'gravity models' that dominated the literature before the introduction of the human capital framework. The gravity model presumed a (mechanical) direct relationship between migration and the sizes of the destination and origin regions, as well as an inverse relationship between migration and distance.

Other Variables

The evidence linking migration and other variables is more ambiguous. Some studies suggest that persons who are unemployed are more likely to migrate than persons who are not, and that the elasticity of migration with respect to the source region's unemployment rate seems to be stronger for workers who are unemployed (at least in the United States). These studies, however, typically ignore that both employment and migration propensities are endogenously determined. Other studies investigate if the wage differential between the destination and source regions has a positive impact on migration. This direct test of the human capital model requires information on the earnings stream that the worker would face in both the source and the destination regions. As a result, the evidence may be very sensitive to how the selection bias problem is handled. Finally, some studies examine the link between internal migration and differences in social welfare benefits provided by different jurisdictions. Although the

evidence is mixed, recent evidence suggests that 'welfare magnets' attract particular types of person to generous localities (McKinnish, 2007).

Return and Repeat Migration

Workers who have just migrated are very likely to move back to their original location (generating return migration flows), and are also very likely to move onward to still another location (generating repeat migration flows). In the United States, the probability of an interstate migrant returning to the state of origin within a year is about 13%, while the probability of a migrant moving on to yet another state is 15% (DaVanzo, 1983).

Unless economic conditions in the various states change drastically soon after the migration takes place, the high propensity of migrants to move again is not consistent with the simple human capital model summarized in eqn [1]. Prior to the initial migration, the worker's cost–benefit calculation indicated that a move from region i to region j maximized the present value of lifetime earnings (net of migration costs). How can a similar calculation made just a few weeks after the move indicate that returning to i or perhaps moving on to region k maximizes the worker's income?

Two distinct factors can generate return and repeat migration flows. The worker might learn that the initial decision was a mistake. After all, a worker contemplating a move faces an uncertainty about economic conditions in the region of destination. Once he migrates, he might discover that the available opportunities are far worse than what he expected to find. Return and repeat migration flows arise as workers attempt to correct their errors.

Return or repeat migration might also be the career path that maximizes the present value of lifetime earnings in some occupations, even in the absence of any uncertainty about job opportunities in different locations. A brief experience with a particular employer in a different city might provide a 'stepping-stone' to better job opportunities elsewhere in the future. In other words, the temporary stay in region j is but one rung in the career ladder that maximizes lifetime earnings.

Family Migration

The discussion has focused on the behavior of a single worker as he or she compares employment opportunities across regions and chooses the location that maximizes the present value of lifetime earnings. Most migration decisions, however, are not made by single workers, but by families. The migration decision, therefore, should not be based on whether a particular member of the household is better off at the destination than at the origin, but on whether the family as a whole is better off (Mincer, 1978).

Suppose that the household is composed of two persons, a husband and a wife. Let ΔPV_{H} be the change in the present value of the husband's earnings stream if he were to move geographically from region i to region j – net of migration costs. And let ΔPV_{W} be the same change for the wife. If the husband were single, he would migrate if the 'private gains' ΔPV_{H} were positive. If the wife were single, she would migrate if ΔPV_{W} were positive.

The family unit (that is, the husband and the wife) will move if the net gains to the family are positive, or if $\Delta PV_{\mathrm{H}} + \Delta PV_{\mathrm{W}} > 0$. The optimal decision for the family unit is not necessarily the same as what is optimal for a single person. Suppose, for example, that the wife would move on her own if she were single, for she gains from the move (that is, $\Delta PV_{\mathrm{W}} > 0$), but that the husband's loss exceeds her gain (so that $\Delta PV_{\mathrm{H}} + \Delta PV_{\mathrm{W}} < 0$). Hence it is not optimal for the family to move. The wife is, in effect, a 'tied stayer.' She will sacrifice the better employment opportunities available elsewhere because her husband is much better off in their current region of residence.

Similarly, consider the situation where the husband experiences an income loss if he moves on his own (so $\Delta PV_{\mathrm{H}} < 0$). Nevertheless, when he moves as part of a family unit, the wife's gain exceeds the husband's loss (or $\Delta PV_{\mathrm{H}} + \Delta PV_{\mathrm{W}} > 0$), and it is optimal for the family to move. The husband is a 'tied mover.' He follows the wife, even though his employment outlook is better at their current residence.

The analysis suggests that marriage deters migration because a person's private gains may be outweighed by the spouse's losses. The analysis also suggests that migration need not 'pay' for all workers in the family. A comparison of the pre- and postmigration earnings of tied movers would indicate that migration reduced their earnings. The family, however, is better off. Recent studies confirm empirically the importance of viewing migration decisions through a family lens (Costa and Kahn, 2000).

International Migration

There was a resurgence of immigration in the United States and in many other countries at the end of the twentieth century. By the year 2010, about 215 million persons – or roughly 3% of the world's population – resided in a country where they were not born. Nearly 9% of the population in Portugal, 13% in Germany, 14% in Spain, 16% in Austria, and 13% in the United States was foreign born.

Different sets of concerns have motivated economic research on internal and international migration. For the most part, the internal migration literature has examined the determinants of migration – who moves and where. In contrast, the international migration literature has been concerned with the issues that propel the debate over immigration policy. In particular, what is the economic impact of international migration on the host country?

Assimilation and Cohort Effects

The impact of immigration on the host country depends on how the skill distribution of immigrants compares to the skill distribution of the native-born population. A great deal of research attempts to document trends in the skill endowment of the immigrant population, and how that skill endowment adapts to economic and social conditions in the host country through the process of assimilation.

Early studies in this literature used cross-sectional data sets to trace out the age–earnings profiles of immigrants and natives (Chiswick, 1978). A cross-sectional survey allows us to compare the current earnings of newly arrived immigrants (measured as of the time of the survey) with the current

earnings of immigrants who migrated years ago. It was typically found that the earnings of newly arrived immigrant men were substantially lower than the earnings of native-born men. In contrast, immigrants who had been in the host country for two or three decades earned more than native workers did.

These findings were typically interpreted as follows. When immigrants first arrive in their destination, they lack many of the skills valued by host-country employers, including language, educational credentials, and information on what the best-paying jobs are and where they are located. As immigrants learn about the host country, their human capital grows relative to that of natives, and economic assimilation occurs in the sense that immigrant earnings begin to converge to the earnings of natives. Immigrant earnings could eventually surpass native earnings if immigrants are positively selected from the population of the source countries. The cross-sectional evidence thus seemed to indicate that upward mobility was an important aspect of the immigrant experience in many host countries.

This interpretation draws inferences about how the earnings of immigrant workers evolve over time from a single snapshot of the population. Suppose, however, that today's newly arrived immigrants are inherently less skilled than those who arrived 20 years ago. The poor economic performance of recent arrivals may then indicate that the newest immigrants have few skills and poor economic potential and will always have low earnings, while the economic success enjoyed by the earlier arrivals may indicate that the earlier waves were always better skilled and had greater economic potential. Because of these intrinsic differences in skills across immigrant cohorts, one cannot use the current labor market experiences of those who arrived 20 or 30 years ago to forecast the future earnings of newly arrived immigrants (Borjas, 1985).

In short, a cross-sectional survey yields an incorrect picture of the assimilation process if there are skill differentials among immigrant cohorts at the time they entered the host country. These 'cohort effects' can be generated by shifts in immigration policy, by changing economic conditions in the host and source countries, and by selective outmigration of immigrants in the host country.

Many studies, using either longitudinal data or repeated cross-sections, have calculated the rate of economic assimilation and measured the importance of cohort effects in immigrant-receiving countries (see the survey in Borjas, 2014). The findings typically differ across countries. In the United States, the immigrant waves that entered the country in the 1980s and the 1990s were relatively less skilled than the waves that entered in the 1960s and the 1970s. Immigrants in the United States also experience some economic assimilation.

Many of the international differences in the skill endowment of the immigrant population and in the rate of economic assimilation are probably due to differences in immigration or adaptation policies. The existing research, however, has not examined this relationship carefully.

Labor Market Impact of Immigration

The entry of immigrants into a particular labor market should lower the wage of competing workers (workers who have the same types of skills as immigrants), and increase the wage of complementary workers (workers whose skills become more valuable because of immigration). For example, an influx of foreign-born laborers into a particular locality reduces the economic opportunities for laborers who already live in the locality – all laborers now face stiffer competition in the labor market. At the same time, highly skilled natives may gain substantially. They pay less for the services that laborers provide, and natives who hire these laborers can now specialize in producing the goods and services that better suit their skills.

In many host countries, immigrants cluster in a limited number of geographic areas. Many empirical studies exploit this fact to identify the labor market impact of immigration by comparing labor market conditions in 'immigrant cities' with conditions in markets untouched by immigration (Card, 1990; Altonji and Card, 1991). These studies typically correlate some measure of economic outcomes for native workers with a measure of immigrant penetration in the locality and often report a weak 'spatial' correlation. The evidence is often interpreted as indicating that immigration has little impact on the labor market opportunities of the native-born.

The 'short-run' perspective that frames this type of research can be very misleading. Over time, natives who live in the immigrant cities – as well as natives who live in other cities – will likely respond to the entry of immigrants. It is not in the best interest of native-owned firms or native workers to sit still; all natives now have incentives to change their behavior to take advantage of the altered economic landscape.

For instance, native-owned firms see that cities flooded by less-skilled immigrants tend to pay lower wages to laborers. Employers who hire laborers will want to relocate to those cities, and entrepreneurs thinking about starting up new firms will find it more profitable to open them in immigrant areas. The flow of jobs to the immigrant-hit areas helps cushion the adverse effect of immigration on the wage of competing workers in these localities. Similarly, laborers living in areas not directly affected by immigration might have been thinking about moving to the cities penetrated by immigrants *before* the immigrants entered the country. They will now choose to move elsewhere. And some native-born laborers living in the immigrant cities will seek better opportunities elsewhere.

The internal migration of native workers and firms within the host country, in effect, can accomplish what the immigrant flow, with its tendency to cluster in a small number of gateway localities, did not – a 'spreading out' of the additional workers over the entire nation, rather than in just a limited number of localities. A spatial comparison of the employment opportunities of native workers in different localities might show little or no difference because, in the end, immigration affected every city, not just the ones that actually received immigrants.

Because geographic comparisons may not provide valuable information about the economic impact of immigration, many recent studies attempt to measure the effect at the level of the national labor market (Borjas, 2003). These studies identify the impact of immigration by examining how the evolution of average wages for narrowly defined skill groups has been affected by differences in the size of the supply shocks. Specifically, the studies explore how the wage of workers with x years of education and y years of work experience evolves over time, as immigrants increase the supply of that skill group. The descriptive evidence reveals a strong negative correlation

between the wage growth of a specific education–experience skill group and the size of the immigration-induced supply shock in that group. A supply shock that increases the number of workers in a skill group by 10% reduces the wage of that group by around 3–4%.

This approach has also led to a renewed interest in structural methods to estimate the labor market impact of immigration. Specifically, the data allow the analyst to estimate the technological relationships among the skill groups, and then use the resulting elasticities of substitution to simulate the labor market impact of a specific supply shock. The structural approach has one important benefit: it allows the estimation of the impact of immigration on the entire wage structure, including both 'own effects' (i.e., the impact of a supply shock in one skill group on the wage of that group), and 'cross effects' (i.e., the impact of the supply shock on the wage of other groups). The existing evidence, however, suggests that the structural estimates are sensitive to the functional form assumptions made about the production technology in the labor market (Borjas et al., 2012).

The Gains to the Host Country

It is simple to describe how the host country benefits from immigration in the context of the 'textbook model' of a competitive labor market. Consider the standard supply–demand analysis presented in Figure 2. The supply curve of labor is given by S and the demand curve for labor is given by D. For simplicity, suppose that the labor supply curve is inelastic, so that there are N native-born workers. A competitive market equilibrium implies that the N native workers are initially employed at a wage of w_0. The area in the trapezoid $ABN0$ measures the value of national income prior to immigration.

Suppose that immigrants enter the country and that, in the short run, the capital stock remains fixed so that the demand curve does not shift. If we assume that immigrants and natives are perfect substitutes in production, the supply curve shifts to S' and the market wage falls to w_1. National income is now given by the area in the trapezoid $ACM0$. The figure shows that the total wage bill paid to immigrants is given by the area in the rectangle $FCMN$, so that the increase in national income accruing to natives is given by the area in the triangle BCF. This triangle is the *immigration surplus*, and measures the increase in national income that occurs as a result of immigration and that accrues to natives.

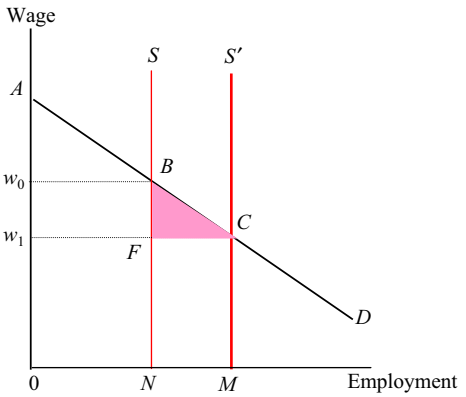

Figure 2 The benefits from immigration to the host country.

Why does the host country benefit? Because the market wage equals the productivity of the last immigrant hired. As a result, immigrants increase national income by more than what it costs to employ them. Put differently, all the immigrants hired except for the last one contribute more to the economy than they get paid.

The simulation of this model for the United States suggests that the net benefit from immigration is quite small (Borjas, 1995). If a 10% increase in supply lowers the wage by 3%, the immigration surplus resulting from the fact that approximately 15% of the US workforce is foreign-born is on the order of 0.2% of GDP. This small net gain, however, disguises the fact that there may have substantial wealth transfers from native-born workers to the capitalists who employ the services that immigrants provide.

Immigration alters labor market conditions in both the host and source countries. Moreover, the immigrant population in the host country often makes large financial transfers, or remittances, to family members left in the source country. These remittances can have a substantial impact on economic activity, particularly in developing source countries. The impact of international migration flows on the source countries is only beginning to be studied systematically.

The Global Gains from Open Borders

A number of studies extend the model to explore by how much world GDP would rise if all restrictions that prevent international migration were removed (Hamilton and Whalley, 1984). These studies use the framework developed in Figure 1 to calculate the area of the triangle A when there is 'free' international migration. The estimated gains to world GDP tend to be very large, on the order of tens of trillions of dollars. The simulations also reveal that a very large number of persons (perhaps more than half the world's population) would need to move from the developing to the industrialized economies in order for regional wage differences to be fully equalized (Benhabib and Jovanovic, 2012).

The huge gains from open borders, however, must be contrasted with potential costs, and the nature of these costs has not yet been examined. First, it is not costless to move billions of people across vast distances. Second, most of the global gains result from the fact that the intercept of the labor demand curve in the industrialized region is above the respective intercept in the developing region (as in Figure 1). These different intercepts, however, are the result of economic, institutional, and cultural factors that *endogenously* led to the development of different infrastructures in the two regions, with the Northern infrastructure allowing similarly skilled workers to attain a higher value of marginal product. For immigration to generate substantial global gains, it must be the case that billions of persons can move to the industrialized economies without affecting the underlying factors that led to these infrastructure differentials in the first place.

See also: Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Immigration, Applied Research on; Immigration, Citizenship, and Integration: Social Work Connections; International Migration; Migration, Theory of; Migration: Anthropological Perspectives; Migration: Cultural Aspects.

Bibliography

Altonji, J.G., Card, D., 1991. The effects of immigration on the labor market outcomes of less-skilled natives. In: Abowd, J.M., Freeman, R.B. (Eds.), Immigration, Trade, and the Labor Market. University of Chicago Press, Chicago, pp. 201–234.

Benhabib, J., Jovanovic, B., 2012. Optimal migration: a world perspective. International Economic Review 53, 321–348.

Borjas, G.J., 1985. Assimilation, changes in cohort quality, and the earnings of immigrants. Journal of Labor Economics 3, 463–489.

Borjas, G.J., 1987. Self-selection and the earnings of immigrants. American Economic Review 77, 531–553.

Borjas, G.J., 1995. The economic benefits from immigration. Journal of Economic Perspectives 9, 3–22.

Borjas, G.J., 2003. The labor demand curve *is* downward sloping: reexamining the impact of immigration on the labor market. Quarterly Journal of Economics 118, 1335–1374.

Borjas, G.J., 2014. Immigration Economics. Harvard University Press, Cambridge, MA.

Borjas, G.J., Grogger, J., Hanson, G.H., 2012. On estimating elasticities of substitution. Journal of the European Economic Association 10, 198–210.

Card, D., 1990. The impact of the Mariel boatlift on the Miami labor market. Industrial and Labor Relations Review 43, 245–257.

Chiswick, B.R., 1978. The effect of Americanization on the earnings of foreign-born men. Journal of Political Economy 86, 897–957.

Costa, D.L., Kahn, M.E., 2000. Power couples: changes in the location choice of the college educated. Quarterly Journal of Economics 115, 1287–1315.

DaVanzo, J., 1983. Repeat migration in the United States: who moves back and who moves on? Review of Economics and Statistics 65, 552–559.

Greenwood, M.J., 1997. Internal migration in developed countries. In: Rosenzweig, M.R., Stark, O. (Eds.), Handbook of Population and Family Economics, vol. 1B. Elsevier, Amsterdam, pp. 647–720.

Grogger, J., Hanson, G.H., 2011. Income maximization and the selection and sorting of international migrants. Journal of Development Economics 95, 42–57.

Hamilton, B., Whalley, J., 1984. Efficiency and distributional implications of global restrictions on labour mobility: calculations and policy implications. Journal of Development Economics 14, 61–75.

McKinnish, T., 2007. Cross-border welfare migration: new evidence from micro-data. Journal of Public Economics 91, 437–550.

Mincer, J., 1978. Family migration decisions. Journal of Political Economy 86, 749–773.

Roback, J., 1982. Wages, rents, and the quality of life. Journal of Political Economy 90, 1257–1278.

Sjaastad, L.A., 1962. The costs and returns of human migration. Journal of Political Economy 70, 80–93.

Migration, Ethnicity, Aging and Social Work Practice

Naina Patel, Ageing and Migration, Lucinges, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Patterns of migration and aging are reflected in the care and welfare needs of minority ethnic (ME) elders in developed countries, with implications for policy, service design, and care delivery. ME elders have a higher incidence of ill health and poverty, while underusing care services. They are hardnosed in their expectations of care services, identifying dedicated care, communication, and access as paramount. Minority voluntary organizations play a critical role in meeting unmet care needs. Mainstream social care, health, and housing providers and professionals increasingly recognize the specific needs of ME elders, but investment in this area is essential.

Introduction

Human history is a history of migration. Recently, the number of people moving between countries in search of work or fleeing from war or persecution has risen rapidly: between 1960 and 2010 international migration increased by 177%, rising from 77 million to 214 million individuals (OECD, 2012). The aging of populations in large numbers is similarly a recent global phenomenon, which is accelerating. In the first decade of the twenty-first century, the number of people aged 60 or more worldwide reached 810 million (United Nations, 2012). It is estimated that for the first time ever, the number of older people will exceed the number of children worldwide by 2047 (United Nations, 2012). Unprecedented in human history or prehistory, this development is a harbinger of a new world society; and one with huge economic, political, and social implications, including for the cluster of activities, industries, and services associated with care and welfare.

Aging and ethnicity combined is not a new phenomenon. However, its scale and scope are new in view of international migration and shifting global demographic patterns. Recent Western history includes a pattern of inflows of people from slaves, sailors, merchants, and elites from the colonies in the past to migrants, refugees, asylum seekers, settlers, and citizens, at present. What is common to these periods is how economics, politics, and the social and historical location of settlement have determined the patterns of flow of people from one part of the world to another.

This article focuses primarily on migrants who arrived in Europe (Patel, 2003), the United States (Markides and Mindel, 1987), Canada (Koehn, 2009), and Australia (Radermacher et al., 2008) who form minorities, who have settled and contributed in those regions, and who are now living as elders. I refer to this group as 'minority ethnic (ME).' Consideration is given to their experiences of aging, care, and welfare issues such as health, housing, and income; the development and delivery of care, with reference to good practice and implications for social work; and current and future challenges in meeting ME care needs.

ME Elders' Experience of Poverty and Ill Health

For ME elders, there are common experiences and cumulative effects of ethnicity and age, reflected in patterns of settlement and housing, incomes and pensions, social status, education, and health as well as diverse cultural traditions and ethnicities. Although some experiences and consequences of aging are independent of ethnicity, the distinctive experience of ME elders is one of discrimination; and ethnicity is frequently a predictor of that discrimination rather than age. At the same time, people are complex and have several identities, giving rise to multiple and intersectional discrimination, which compound the challenges faced by ME elders – whether these other identities arise from gender, religion, disability, or other sources.

While ME elders are not homogeneous, like elders in the majority population, issues such as health, income, and housing for ME elders have an ethnic face irrespective of their specific backgrounds or identities. For all elders the link between poverty and health is well established. There are patterns of economic and health inequalities that necessitate policies on poverty reduction and tailor-made interventions, including culturally competent health, social care, and housing services (see 'Aging now' in the film *Progressing Policy with Passion*).

Income

The ME aging experiences exhibit a higher incidence of poverty compared to their majority peers. Even in relatively rich countries, such as Finland, the United Kingdom, and Germany in the early 2000s, the percentage of ME elders with incomes of less than €750 per month was 48%, 35%, and 25%, respectively of the total ME populations in each country. Moreover, in each case 40% or more were dependent on social benefits (Minority Elderly Care or MEC Project). These data are taken from the 'MEC Project,' which was instituted and led by myself at Policy Research Institute on Ageing and Ethnicity, involving institutions from 10 European countries (**Table 1** lists the countries) (www.priae.org). The project conducted an in-depth interview-based survey of 3277 minority elders and 901 service providers across these European countries over a 3-year period up to 2005.

In the United Kingdom, ME elders receive lower pensionable incomes than their majority peers (Khan, 2008). ME elder women are likely to be poorer than their peers in the majority, representing a triple jeopardy. Being an older woman is doubly disadvantageous; and being a minority one multiplies the disadvantage further (Patel and Traynor, 2005).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28009-4

Table 1 Elders' expectations of service-delivery quality in care services, in Europe as a whole and by country

Country	*Main minority ethnic groups surveyed*	*Most important expectation*	*Second most important expectation*
Bosnia/Croatia	Hungarians, Bosniaks, Czechs, Roma	Hygiene and tidiness	Identity appropriate care (religious, gender, ethnic)
Finland	Sami, Russian, Vietnamese	Effective information and care systems	Identity appropriate care (cultural values, ethnic)
France	Algerian, Moroccan, Vietnamese	Effective information and care systems	Identity appropriate care (gender, ethnic)
Germany	Turkish, Italian, German Repatriates	Effective information and care systems	Identity appropriate care (linguistic, ethnic)
Hungary	Roma, German, Croats	Access to care services	Identity appropriate care (gender, ethnic)
Netherlands	Surinamese, Moluccans, Turkish	Effective information and care systems	Access to care services
Spain	Moroccan, Roma, Latin American	Access to care services	Identity appropriate care (gender, ethnic)
Switzerland	Spaniards, Italians, Former Yugoslavia	Access to care services	Identity appropriate care (ethnic)
United Kingdom	African-Caribbean, South Asian, Chinese	Effective information and care systems	Identity appropriate care (religious)
All countries	All, as above	Effective information and care systems	Clear communication

Note: Summary of factor analyses of responses by 3277 elders to an extensive set of Likert-scale questions.
Source: PRIAE MEC Project database (www.priae.org).

The reasons for lower pensionable incomes received by ME elders compared to their counterparts in the majority population are due to factors such as late entry into the labor market (especially if they migrated as adults); lower/differential earnings from lower status and unskilled jobs, conflated by discriminatory behavior by employers; a lower participation in the labor market; lack of knowledge of age related entitlements (e.g., because information is not available in their primary language); or not claiming the full/any amounts (often for cultural reasons). During their working years, ME elders experience at least twice the rate of unemployment than the older majority population, and for longer periods. The nature of the manual jobs that ME elders have occupied affects their income and health in later years. Ill health contributes to early aging, early retirement, and lower life expectancy (Patel, 1990; Haas et al., 2012).

Health

Health and ethnicity are separately – and to some extent jointly – established areas of study in Western countries, but a focus on ME elders and their health is relatively recent. National data limitations on age ethnicity and health have contributed to the lack of investigation of issues related to ethnic health inequalities. In the last decade, however, there has been a growing attention to both the physical and mental health of ME elders across Europe, the United States, Canada, and Australia (Nazroo, 2003; Patel, 2003; Rao et al., 2006; Koehn, 2009; Haas et al., 2012). See 'Mental Health and Physical Active Aging' in films *Dementia Matters Ethnic Concerns; Get Active Today*.

ME elders have multiple and complex health needs. Experience of debilitating long-term illnesses such as heart disease, diabetes, and hypertension are more common than among older majority populations. In addition, there are differences between ethnic groups and by gender: for example, in the United Kingdom, South Asian groups have higher rates of diabetes than the Chinese and are approximately 50% more likely to die prematurely from coronary heart disease than the general population; in Finland, 46% of Vietnamese reported poor health compared to 12% of Russians and 8% of Sami; in Hungary, Roma people have worse overall health than Croats, and Germans, and Roma women suffer greater ill health than Roma men. Across Europe, diabetes and cardiovascular disease are high among some groups (Surinamese in the Netherlands, Asians in the United Kingdom, Moroccans in Spain, Russians in Finland), while for other groups high levels of hypertension, bone disease, and mental health are particular concerns (PRIAE, 2004). Such individual conditions can additionally give rise to multiple complications, unless treated and managed appropriately.

The main reasons for a higher prevalence of ill health among ME elders include: genetics, socioeconomic inequalities, and limited access to health and care services. Factors underlying these reasons, or conflating them, include inappropriate diagnoses (health experts may be unaware of relevant ME issues), discrimination and racism, the type of employment in working years, language and communication difficulties, cultural and faith-based factors (rationalizing dementia as an 'act of God,' for example) and health-related behaviors regarding physical activity, food, and drink.

In some areas of health, ME elders are 'invisible,' or a 'hidden problem,' and so have not been considered in research, practice, or policy until recently. Alzheimer's disease is the most common type of dementia that has been referred to as the 'silent killer' since it is a progressive, degenerative disease of the brain. It cannot be prevented, or cured. Age is a major 'risk factor,' and therefore, dementia is becoming as western societies become older on average. For instance, it is the sixth leading cause of death in the United States. Research into ethnicity and dementia is well advanced in the United States, compared to Europe and Australia; but there is lower level of awareness of Alzheimer's disease among ME families and elders. The necessary information is scarce, including in ME languages; inadequate knowledge and communication also lowers access to services; and cultural preconceptions of how dementia is understood results in late referrals to care professionals, often delayed to a point where ME elders' dementia has considerably worsened (Patel et al., 1998; Jimenez et al., 2013).

Mental health problems also arise from isolation and loneliness, which are features of old age in many societies, particularly those in the West that emphasize privacy and individual fulfillment. ME elders are not immune from loneliness,

multiplied by factors such as poor fluency in the majority language, inadequate communication, ill health (and resultant disabilities), and poverty that socially exclude them from participating in many activities. In such cases, families become the main source of support and comfort, but as with the majority population this is seldom enough (Radermacher et al., 2008).

Housing

> The popular image of a … British family … is a husband, wife and two children. For (an ethnic minority family) the popular image is one of an extended family network, 'families within families', providers of care and social and psychological support. Strangely such virtues are married to a conception of 'overcrowded households'!. (Patel, 1990: p. 30)

These stereotypes were never true, and today, ME elders are far less likely to live in extended family or multigenerational households. As with any person, for ME elders housing is an essential anchor, providing for certain basic needs, as well as giving a sense of location, which is particularly important in later stages of life. In their old age, with increasing illness and poverty, a home where they can age with comfort and support is the least that elders can expect. Suitably adapted housing (including culturally and identity appropriate services) can not only ameliorate a range of health problems for older people, but also with social contact and activities, improve their quality of life. See 'Multicultural housing' in film *Progressing Policy with Passion*.

Meeting the hopes and expectations for appropriate care requires changes in policy and practices. Higher life expectancies, increasing numbers of older people (and hence an expanding market), a societal conversation about autonomy and choice are all resulting in moves toward new and more flexible systems of health and social care in response to majority older people's needs (keeping affordability and costs in mind). Similar moves, haltingly perhaps, but in the right direction, aided by research and other work on ME elders, is also leading planners and developers to provide appropriate housing solutions for ME elders. Examples of minority ethnic housing associations in the United Kingdom include, Ashana House, Great Wall Housing, Unity, Ujama, Sadeh Lok, and Pine Court. In the process, debates about 'separate housing or not' have been defused through successful working multicultural models of housing.

Additionally, there is a continuing discussion of how their families and communities respond to ME 'elder housing' when culturally such schemes are not deemed to be the norm. However, the common long-held view that 'minorities look after themselves' (thus obviating consideration of specialist care and services as ME elders age) is incorrect. Research and active dissemination work with ME age and housing organizations show that elders and families need tailor-made information, examples, and support to better understand, and increase confidence in considering and choosing among housing options in old age when the need arises (Patel and Traynor, 2006; Mahmood et al., 2008).

Encouraging Take-Up of Services

Amid high levels of ill health and poverty, underuse of services by ME elders persists, and needs to be understood. The 10 European country–MEC project revealed a range of explanatory factors, including lack of knowledge, language difficulties, income disparity, inappropriate services (lack of an appropriate cultural mix), discriminatory assumptions, and complex health systems. The responses from all of the countries in the MEC project suggest that professionals and managers/planners have hitherto attributed the underusage of services to the inherent characteristics of the elders themselves, and have paid insufficient regard to institutional characteristics and attitudes, or the services provided.

An organization can do several things to help ME elders overcome barriers and gain access to services. For example, good practice examples in Europe in the MEC research include: community outreach to access information and services; and developing and delivering ethnically sensitive day and residential services.

The Development and Delivery of Elder Care Services

Mainstream Services

There is considerable evidence that ME elders are disadvantaged by limited access to services in social care, health, housing, income, and employment. For example, in gaining access to housing, a ME elder needs to be aware of its availability and the route to access it. However, what if the elder:

- Does not know that such services exist?
- Does not know that she or he may be entitled to them?
- Is presented with a complex system (involving referrals, assessment processes, waiting lists, and various agencies) that is not understood?
- Is not confident she or he meets the eligibility conditions?
- Has to be literate in a variety of ways (e.g., language fluency, form filling, engaging in interviews).

For ME elders the nub of the problem is that language and communication difficulties inhibit acquisition of requisite knowledge, even before it can be used to navigate and access services.

'Do not know, do not hear, do not get' is how one ME elder aptly summarized access issues, for instance, in housing (PRIAE, 2008). Moreover, questions also arise of whether available housing services are appropriate for cultural and other reasons. In some cultural frames, the concept of elders being in a separate home/living independently is a source of embarrassment and discouraged. The idea of seeking external help when problems arise is still regarded by many elders from various ethnic backgrounds, with shame and guilt and a sign of family failure (or, worse, their own). Containment of problems and tolerance of whatever the family circumstances dictate often become the reality for many elders. This is not to say that all ME elders fill this profile. Increasingly, where ME housing provision exists, planning by elders and families of their living independently in their own home – but close enough to the family – is rising, to the satisfaction of both (Patel, 1990, 1999).

Let us suppose that ME elders have managed to navigate the information and culture barriers. For some there may still be discrimination in obtaining housing provision. It is not, therefore, surprising that many who have secured housing see

it as a remarkable achievement rather than an entitlement, especially if it meets their needs regarding language and culture.

The lack of access to appropriate services is amplified by insufficient proactive referral, assessment, treatment, and care of all elders with specific health and social care needs. See 'Voices of minority elders' in the film, *Minority Elders Policy Network*.

For example, a multifaceted and flexible approach is required in caring for people with dementia. Implicit in this approach is the notion that all people with dementia have a right to available care. Various studies show that ME elders with dementia are marginalized from mainstream health and social care methods and approaches (Patel et al., 1998; Jimenez et al., 2013). To alleviate this problem, films and other materials have been produced by minority ethnic groups to educate and train clinicians and social workers, and thereby boost awareness and referrals (see films under 'Health').

Services Offered by ME Voluntary Organizations

There are countless examples of self-help initiatives by ME elder women and men, sometimes with support from professionals, who have created service organizations to help meet ME elders' needs in multicultural contexts. Many aim to combat exclusion and work to counter discrimination and disadvantage through activities, which range from information and support to the provision of tangible day care and residential services. ME voluntary organizations have emerged as a direct response to ME elders' needs. These are meeting key requirements as expressed by ME elders, such as service availability in their own homes; appropriate food, language, religious and spiritual practices; and as bulwarks against prejudice and racism. Such organizations act as 'primary providers' in many contexts rather than as 'complementary providers,' effectively substituting for elements of mainstream health, social care, and housing services (Patel, 1990; PRIAE, 2004; Dominelli, 2008; Williams and Johnson, 2010). Nonetheless, ME organizations' approaches to services provision can be valuable for mainstream organizations because their existence provides role models of multicultural care for mainstream service providers to emulate and build upon.

Toward a Wider Recognition of the 'Specific' Needs of ME Elders

The central idea that ME elders have 'specific' needs has had a troubled journey. For over two decades, the subject has been marginalized under 'special needs'; or has not been considered important on the basis that 'services are open to all.' The MEC Project results show that over time professionals and planners in some countries are recognizing that ME elders have specific needs and that these have to be met. Nevertheless, in several countries in this project barely 50% of professional respondents recognized that ME elders have specific needs and encounter access barriers (France, Finland, Hungary, the Netherlands, and Spain). The United Kingdom and Germany had the highest level of recognition of these issues – over 80% among UK professionals and over 60% in Germany. This result could be attributed to organized responses against racism and discrimination in care and welfare from ME communities working in alliance with majority community professionals. Beyond recognition that specific needs exist, many care professionals across Europe have yet to fully understand the nature and characteristics of ME elders' needs and expectations.

Analysis of the MEC Project survey shows that, as a whole, ME elders' have five key expectations of quality of care services received: effective information and care systems, good communication, full access to services, identity appropriate care, and tangible issues such as hygiene and tidiness (**Table 1**). These mirror the difficulties encountered in mainstream services, as mentioned earlier.

The first expectation, *effective information and care systems*, underscores that for ME elders, care systems and professional, expert staff are paramount; and that full information on diagnoses, treatment, and access should be provided. 'Soft' aspects such as trust and assurance, empathy, and privacy are implicit in this first expectation. Linked to it is that of full access to services (third expectation in the table); and the availability of services in ME elders' own language or presence of an appropriate interpreter is also a priority (second expectation). In a similar vein, the fourth expectation, identity appropriate care, relates to individuals' emotional core, and ultimately their dignity. The fifth expectation includes not just hygiene and tidiness of care facilities, but also the presentation, neatness, and practices of staff. Research in other European countries, North America and elsewhere supports the above findings that 'matter-of-fact,' clinical issues are paramount in ME elders' requirements for good quality care (Miranda et al., 2010; Garroutte et al., 2008).

There are country differences in elders' expectations, partly reflecting the mix of ethnic groups in each country. For example, from the MEC results across Europe, identity appropriate care is normally deemed the second most important aspect in defining good care, but what defines these identities varies: for ME elders in Finland cultural values are to the fore; in Germany it is language; and in the United Kingdom it is religion (**Table 1**). This is not unexpected given that people possess multiple identities. However, variations between countries have policy implications regarding the balance in relevant aspects of service provision. Identity is important, but what constitutes identity differs, context matters, and recognizing diversity is important.

Current and Future Developments

While many countries with ME elders have made considerable progress in the area of aging and ethnicity, the emerging challenge for the care industry and its workforce is not whether there will be a greater number of ethnic elders, but that they are increasingly from many cultures, with different languages and faiths; and with various socioeconomic and health patterns. ME elders today and in the future will have a greater variety of needs, many of which must be reflected in services provided. How policy makers and service providers

can best be informed to determine demand and meet diverse needs appropriately is the issue. Providers need to consider how they can better gear up to meet the increase and mix in demand for services in this context, and how they can ensure that voice, choice, and autonomy are met for individuals from all cultures.

Policy makers are often insufficiently engaged with ME elders, or their organizations, and consequently miss them off the policy radar, considering aging and ethnicity together. This leads to the notion that "in age 'race' is blind; in 'race' age is blind" (PRIAE, 2004: p. 15). However, through targeted efforts by specific organizations, individuals, and professionals, ME elders are now being increasingly considered in policy, care, research, and training, albeit to a variable degree.

The MEC Project reinforced an important policy message: that increasingly in all countries care professionals are recognizing that ME elders have specific needs which are growing, but are often insufficiently catered for. ME voluntary age organizations are an important element of providing for that critical care, but are not the full answer. Unfortunately, policy makers still do not pay sufficient heed to such entities as key resources for supplying tailored services to ME elders.

In the last decade or so, mainstream health, social care, and housing providers have made progress in the recruitment of ME staff, the development of ME elder facilities, information provision, and communication campaigns on a range of health and social care support issues, as well as in consulting ME elders and ME elders' groups on various age related policy issues. However, progress is not uniform. The quality of services can still be categorized as 'patchy, ad hoc, and piecemeal,' particularly in the context of rising demand, previous unmet needs, and growing ME elder populations (PRIAE, 2008; Jacobsen et al., 2011; NSPAC, 2011).

More progress on all these fronts is needed. This can build on the considerable research, policy work, and innovative initiatives to date. Unfortunately, it is not uncommon to hear that not much work exists, that researchers have to start anew. At the same time there are many challenges that need to be met today or in the near future.

First, effective provision of ME-tailored services requires effective staffing. Ethnicity and aging require an engagement with multidimensional issues, some of which are complex and uncomfortable for care staff (Patel, 1995; Sakamoto, 2007; Dominelli, 2008). This is not helped by care staff in social work, and related areas, having to work under pressurized conditions, in highly regulated services giving them little discretion, with declining resources. They often have to decide whether to go for satisfying basic cultural needs (at a physical level) or engaging with elders to forge a deeper understanding, in order to work with various complexities and contingencies that they are not familiar with. Care staff must be supported and strengthened in this regard.

Secondly, progress in the provision of specific care for ME elders is not without challenges. A long-standing view persists that after the current generation, ME elders will be well integrated into society and not need specific services. Such views are sometime linked to anti-multicultural perspectives. While the mainstreaming of some aspects of care; is, or will undoubtedly be the case, research on aging generally shows that as people age their needs and preferences become more culturally or identity pronounced, not less. Services tailored to ME requirements will continue to be necessary. See 'Integration' in the film *My Yorkshire*.

Finally, increasingly, governments are following ideologies of austerity, with wide-ranging effects, including cuts in services. This is not a good time to 'age.' Current austerity measures in many developed countries have hit services targeted at ME elders, directly and indirectly. For example, privatization and the concentration of services provision in selected organizations are having a deleterious effect on smaller ones, including ME voluntary service providers, as funding is cut or redirected. In this age of austerity, disinclination in use of ME organizations by policy makers may be reinforced, to the detriment of service provision to ME elders. Additionally, there is the tendency by politicians and policy makers to conflate ME elder issues with the ongoing farrago on immigration and welfare entitlements. What should be clearly stated is that this is misdirection. ME elders' expectations and rights as citizens fully derive from their contribution to the economies and societies in which they have settled.

The West is diversity rich. An appropriate and resource efficient response to ME elders' aging in the twenty-first century is to invest in services that *work* for minority ethnic elders, be they located with minority or majority service providers. Investment in care and welfare in developed countries today, not least for MEs, is not an option.

See also: Culturally Appropriate Interventions in Social Work; Elder Abuse; Elder Care; Indigenous Social Work; Oppression; Social Work Theory.

Bibliography

Dominelli, L., 2008. Anti-racist social work, third ed. In: Practical Social Work Series. Palgrave Macmillan, London.

Garroutte, E.M., Sarkisaian, N., Goldberg, J., Buchwald, D., Beals, J., 2008. Perceptions of medical interactions between healthcare providers and American Indian older adults. Social Science & Medicine 67, 546–566.

Haas, S.A., Krueger, P.M., Rohlfsen, L., 2012. Race/ethnic and nativity disparities in later life physical performance: the role of health and socioeconomic status over the life course. The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences 67 (2), 238–248.

Jacobsen, L., Kent, M., Lee, M., Mather, M., 2011. America's aging population. Population Bulletin 66 (1). Population Reference Bureau, Washington.

Jimenez, D.E., Cook, B., Bartels, S.J., Alegría, M., January 2013. Disparities in mental health service use of racial and ethnic minority elderly adults. Journal of the American Geriatrics Society 61 (1), 18–25.

Khan, O., 2008. Financial Inclusion and Ethnicity. Runnymede Trust, London.

Koehn, S., 2009. Negotiating candidacy: access to care for ethnic minority seniors. Ageing and Society 29 (4), 585–608.

Mahmood, A., Chaudhury, H., Kobayashi, K.M., Valente, M., 2008. The housing and community characteristics of South Asian immigrant older adults in Greater Vancouver, British Columbia. Journal of Architectural and Planning Research 25 (1), 54–75.

Markides, K.S., Mindel, C.H., 1987. Aging and Ethnicity. Sage, Beverly Hills, CA.

Miranda, F.J., Mera, A.C., Hernandez, M.I.S., Murillo, L.R., Cervera, J.V., 2010. Adapting the Servqual Scale to Primary Health Care Services in Spain. APAS Papers 197. Academic Public Administration Studies Archive.

Nazroo, J., 2003. The structuring of ethnic inequalities in health: economic position, racial discrimination and racism. American Journal of Public Health 93 (2), 277–284.

NSPAC, 2011. The Ageing Experience of Australians from Migrant Backgrounds. National Seniors Productive Ageing Centre, Braddon, ACT, Australia.
OECD, 2012. Connecting with Emigrants: A Global Profile of Diasporas. OECD Publishing, Paris.
Patel, N., 1990. A 'Race' Against Time? Social Services Provision to Black Elders. Runnymede Trust, London. http://www.runnymedetrust.org/publications/170/32.html (accessed 10.12.12.).
Patel, N., 1995. In search of the holy grail? In: Hugman, R., Smith, D. (Eds.), Ethical Issues in Social Work. Routledge, London, pp. 16–46.
Patel, N., Mirza, N., Lindblad, P., Armstrup, K., Samaoli, O., 1998. Dementia & Minority Ethnic Older People: Managing Care in the UK, Denmark and France. Russell House Publishing Ltd, Lyme Regis, UK.
Patel, N., 1999. Black and minority ethnic elderly: perspectives on long-term care. In: Royal Commission on Long-term Care for the Elderly. With Respect to Old Age: Long Term Care – Rights and Responsibilities: Research, vol. 1. HMSO, London, pp. 257–304.
Patel, N. (Ed.), 2003. Minority Elderly Care in Europe: Country Profiles. PRIAE, Leeds, UK.
Patel, N., Traynor, P., 2005. The Financial Well-Being of Ethnic Minority Women Aged 40+ in the UK. NPOE, Amsterdam.
Patel, N., Traynor, P., 2006. Developing Extra Care Housing for Black and Minority Ethnic Elders. London, UK.
PRIAE, 2004a. Black and Minority Ethnic Elders in Europe: Health and Social Care Research Findings MEC Project. PRIAE, Leeds, UK.
PRIAE, 2004b. MEC Project database at. www.priae.org.
PRIAE, 2008. Ageing and Ethnic Diversity in the UK, a Policy Digest. PRIAE, Leeds, UK.
Radermacher, H., Feldman, S., Browning, C., 2008. Review of Literature Concerning the Delivery of Community Aged Care Services to Ethnic Groups. Ethnic Communities Council of Victoria and Monash University, Melbourne, VIC.
Rao, D.V., Warburton, J., Bartlett, H., 2006. Health and social needs of older Australians from culturally and linguistically diverse backgrounds: issues and implications. Australasian Journal on Ageing 25 (4), 174–179.
Sakamoto, I., 2007. A critical examination of immigrant acculturation: toward an anti-oppressive social work model with immigrant adults in a pluralistic society. British Journal of Social Work 37 (3), 515–535.
United Nations, 2012. Population Ageing and Development: Ten Years after Madrid. Population Facts No. 2012/4. Department of Economic and Social Affairs, Population Division, New York.
Williams, C., Johnson, M.R., 2010. Race and Ethnicity in a Welfare Society. Open University Press, Maidenhead, UK.

Relevant Websites

www.priae.org – Ageing and Ethnicity: Dementia Matters Ethnic Concerns (under publications/media).
www.priae.org – Ageing and Ethnicity: Get Active Today (under publications/media).
www.priae.org – Investing in Ageing and Ethnicity – My Yorkshire (under publications/media).
www.priae.org – Minority Elders Policy Network (under publications/media).
www.priae.org – Progressing Policy with Passion (under publications/media).

Migration History

KJ Bade, Osnabrveck, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9809–9815, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

As migration history has both a historiographical and a historical dimension this article branches both sides. The first part gives an overview on issues, concepts, and methods of historiographical migration research. The second part sketches the courses of European and worldwide migrations from early modern times up to the present. In historiographical descriptions of migration movements and migrant behavior on the one hand and in the identities assigned to migrants by authorities controlling migration on the other hand, despite the fundamental differences in motivation and aims, problems of description and ascription seem quite similar on both sides.

The term 'migration history' has historiographical and historical dimensions. This article is thus divided into two sections. The first part outlines concepts and methods, focusing on tasks, fields, and problems of interdisciplinary historical migration research. The second section gives some insight into the history of migration itself, based primarily on modern European history. Non-European migration is included where it was linked in some way to European migration.

Historical Migration Research: Issues and Concepts

Motivations, Patterns, Typologies

Migration has always been a constitutive element of the *conditio humana*, as *homo sapiens* spread over the world as *homo migrans*. The history of migration is part of general history and can only be understood in context. As social processes, after all, migration movements are responses to complex economic, ecological, social and cultural, religious, ethnic, and political conditions and challenges. Migration thereby gradually penetrates all spheres of life. Historical migration research branches spaces, cultures, and historical times. Therefore it is necessary for researchers to adopt an interdisciplinary approach. Varying in degree according to the problem at hand, interdisciplinary research strategies cover almost all human sciences in both contemporary, empirical, and historical migration research.

There was and still is a broad variety of reasons for migration. We can, for example, posit economic or social motivations. Within this field, we could then distinguish between subsistence, betterment, or career migration (Tilly, 1978). These types of movements can in turn be distinguished from migrations motivated by religious and political or ethnonationalistic and racial reasons, which also cause flight and forced migrations. These last include the expulsions and forced repatriations in the twentieth century, where the movement of borders across people often caused movements of people across borders.

Of crucial importance for any critical analysis of historical migration trends – and for an insight into the fate of many migrants, often less a matter of choice than of circumstance – is an awareness of the fact that definitions and attributes such as 'emigrant' and 'immigrant,' 'labor migrant,' or 'refugee' and 'asylum seeker' are ascribed classifications. So far, they have often been assigned for administrative or tax purposes or for epistemological scientific reasons – which also depend on specific classification criteria – but these terms in no way describe the generally multiple identities of migrants (Castles and Miller, 1998).

The matter is even more complicated because when migration was controlled and restricted, migrants had to assume the official classification of their status in order to get past immigration officials, often leaving a 'false trace' in official records and statistics. This is one more reason why we have to distinguish between the way migrants classified themselves and the way they were classified, for example, by the state or contemporary researchers.

Spatial Dimensions and Research Concepts

In examining spatial mobility, we have to distinguish between movements in geographical spaces and those in social spaces. Geographically, the scope of historical migration research ranges from the macro-cosmos of international and intercontinental mass movements to the micro-cosmos of interregional or interlocal migrations, and thus also extend from large-scale studies on a highly abstract level to small-scale case studies with larger socio-historical significance. Levels and methods of historical migration research thus range from micro-historical to meso- and macro-historical approaches, including even multilevel theories of migration research, and from individual or group-specific dimensions to quantitative analyses of highly aggregated mass data serving to determine collective behavior during mass movements. On the temporal axis, the field of historical migration research stretches from long-term studies of single migration movements to cross-sectional analyses of the entire, simultaneous migration in one region or beyond its borders (Lucassen and Lucassen, 1997).

The varying approaches by individual disciplines and the various emphases in interdisciplinary approaches lead to different interpretations of migration history. As a social and historical phenomenon, for instance, migration should be seen as a complex process. According to 'classical' historical migration research, this process was often triggered by an increasing propensity to migrate, e.g., in the case of European transatlantic mass emigration in the nineteenth century, and by the consequently more or less gradual mental segregation from the social

context of the home region. In this process, transatlantic migrant networks played an important role. The next phase would be the transformation – often provoked by some external cause – of this propensity into an actual decision to migrate, followed by the act itself. In the case of dense transnational networks resulting from chain migrations, the departure often took place more or less abruptly. The last phase – provided the migration process was not aborted or reversed by remigration – was described as assimilation into the social and cultural context of the immigration region. In the case of large discrepancies in material culture, sociocultural norms, or collective mentalities, assimilation could become a long-term social and cultural process, sometimes even reaching intergenerational dimensions ('second-generation immigrant').

The 'classical' approaches of historical migration research focused on movement in geographical spaces. In the 1990s, however, new approaches emerged which were to focus upon the movement and positioning of migrants in social spaces. This applied especially to meso-level network theories as well as to theories and typologies of transnational social spaces and migrant identities. These new approaches were derived mainly from the social and political sciences (Faist, 1998, 2000; Pries, 1999), dealing with the accelerated development of transnational structures in the economy, in society, and in politics, against the background of rapidly forced globalization. These new approaches in migration research sometimes lead to unproductive claims of exclusivity and needlessly hostile confrontation with the 'classical' approaches of migration research, even though both approaches can be usefully incorporated into complex research concepts (Gerber, 2000).

Tasks of Historical Migration Research

Historical migration research has three main tasks. The first is to investigate migration movements in terms of volume, courses, and structures. The second task is to study the behavioral patterns of migrants with respect to, for example, region, class, group, and gender. The third task is to embed migration movements and the behavioral patterns of migrants into the framework of population, economy, society, and culture of both emigration and immigration areas. This includes the economic, social, and cultural tensions between both sides encouraging migrations as well as the effects of migration on both sides.

Despite such comprehensive tasks and the fact that the movement of peoples rates as one of the most 'moving' moments in history, historical migration research is not an independent discipline. It is, in fact, an interdisciplinary field of research, to which the humanities as well as social and behavioral sciences contribute.

Historical migration research as an interdisciplinary branch is relatively young. Its main disciplinary foci are the historical sciences, including, inter alia, the history of population, economic, social, cultural, and gender history as well as ethnohistory and historical anthropology. They also include the branches of legal and political history analyzing the structure of migration procedures and their repercussions on migration processes. And there are links to mainly empirical disciplines or approaches that can be also applied to historical issues and historiographical approaches (e.g., sociology, social geography, or social psychology).

Migration Movements: Global and European Perspectives

Migration was and still is by no means a uniquely western phenomenon. People in Africa, Asia, and the New World moved for similar reasons as the Europeans. As in Europe, most people moved over short distances, usually remaining within their own region or continent. Nonetheless, there have been several nonwestern diasporas. Between AD 600 and 1500, the Arabs and the Turks expanded into parts of southern Europe and controlled the Balkans between AD 1500 and 1914. The Japanese and Russians colonized Korea and Siberia respectively, and the Chinese created trading communities along the shores of the Indian Ocean (Hoerder, 2001).

We can classify the history of global migrations according to regions, periods, or migration patterns. We shall focus here, by way of example, on the modern history of European migrations, looking first at the non-European references linked to the external migration of Europeans. We know that the most voluminous migration movements outside of Europe were triggered by the process of European expansion and contraction. Many Europeans and non-Europeans met for the first time during the colonial migrations that followed the age of European 'discoveries,' and, for the last time, during postcolonial migrations from the former European colonies back to the 'mother countries.'

Global Dimensions

Some migrations outside of Europe were directly carried out or instigated by Europeans, e.g., the settler migrations from Europe to North America, South America, South Africa, Australia, and New Zealand (totaling about 70 million people); the transatlantic African slave trade (12 million people); and the migration of colonial soldiers and indentured laborers from China, India, Java, and Japan to the colonial labor markets in Asia, Africa, and South America (2.5 million people). In total, the volume of intercontinental migrations directly caused by the European expansion can be estimated at about 100 million migrants, both voluntary and involuntary.

In addition to migratory movements controlled by Europeans, an unknown number of Africans, Asians, and Amerindians moved as an indirect consequence of European expansion. In the New World, the indigenous Amerindian population migrated or was driven to those regions left to them by the intruding Europeans, a fate shared by the indigenous populations in parts of Southern Africa, Australia, and New Zealand. In Asia and Africa, European expansion led to new jobs in textile mills, mining, cash-crop agriculture, railway construction, and the colonial armies, causing large scale, internal labor migrations. From this point on, we shall only mention migration movements directly controlled by the Europeans.

The migrations resulting from the European expansion can be divided into three types of movements and three periods: (1) the migration to the colonial plantations, mines, and railway construction sites between AD 1500 and 1900; (2) the transatlantic mass exodus of Europeans from the mid-nineteenth century until 1914, and their overseas emigration between 1945 and 1960; (3) the colonial and overseas return

migration as well as the non-European labor migration to Western Europe after 1945. We shall briefly outline the first two movements and include the third in the review of developments in Europe (Section European Dimensions).

The first overseas migration movement started around AD 1500, just after the Iberians had discovered the New World. The Amerindian population declined rapidly because of European epidemic diseases and economic exploitation, while the number of Europeans willing to settle in the American colonies was too small to satisfy the demand for labor in the mines and plantations. The Iberians had been faced with a similar labor shortage in the South of their own peninsula and had imported slave labor from sub-Saharan Africa. Some of these slaves were brought to the New World, and soon the Iberians were also transporting slaves directly from West Africa to the New World. The British, French, and Dutch followed suit, and between AD 1500 and 1850, roughly 5 million slaves were brought to Brazil, 5 million to the Caribbean, and 1 million each to Spanish and North America. In three and a half centuries, less than 4 million Europeans migrated to the New World, making it a demographic extension of Africa, rather than Europe (Emmer, 1991).

After AD 1850, humanitarian protests totally suppressed the Atlantic slave trade and won emancipation for all slaves in the New World. However, the demand for tropical labor remained. The recruitment of free African migrants failed and although the number of European migrants rose dramatically after the middle of the nineteenth century, they chose to settle in the more temperate overseas zones. Employers in the tropical world then turned to Asia; and between AD 1850 and 1914 about 2.5 million migrant workers from India, China, Japan, Java, and Polynesia went to work in the coffee fields around Sao Paulo and on islands off the Peruvian coast to dig guano. They also worked on sugar plantations in the Caribbean, Natal, Fiji, Hawaii, and Queensland, in the tobacco fields of Sumatra, on rubber plantations in Indochina, and they even helped to build the railways of East Africa (Indians) and California (Chinese) (Engerman, 1986; Northrup, 1995).

Most of these Asian workers had signed a contract of indenture that granted them free passage in exchange for overseas employment for a fixed period, and entitled them to return home once their contracts expired. The majority did not return, creating ex-Indian communities in South Africa, Fiji, Hawaii, Trinidad, Guyana, and Suriname; ex-Japanese communities in Brazil and Hawaii; and ex-Chinese communities in the Caribbean and Indonesia. Humanitarian protests in the West and nationalist protests in China and India brought some of these migration movements to an end. After World War II, millions of Asian migrants began to cast their sights on the Middle East.

One special chapter in the encounter between the European and non-European world through migration, was marked by the European transatlantic mass exodus from the mid-nineteenth to early twentieth century (Nugent, 1992; Baines, 1995). This exodus was far stronger than the European colonial migrations. Up to the 1830s, the continental migration from central Europe to eastern and southeastern Europe – in contrast to the emigration from England, Ireland, Scotland, and Wales – was much more powerful than transatlantic migration, which had become a mass movement only by the middle of the century.

As a secular social mass movement, transatlantic migration accompanied the shift in Europe from agriculture to industry. It was the transportation revolution that ultimately facilitated the mass exodus. Of the same importance, however, were the transatlantic networks established by migrations of people and exports of goods and capital even prior to the Industrial Age. The mass migration was preceded by the transatlantic movement of colonial labor and settlement migrations of destitute Europeans – men and women – who worked off their passage to the New World in 'indentured servitude' and were eventually rewarded with a small amount of start-up capital or a piece of land (Wokeck, 1999).

Essential factors in the development of transatlantic emigration in the nineteenth century were the liberty to emigrate from the sending countries, the need for immigrants, and their full acceptance in the receiving countries. The advent of the steamship and the expansion of railroad systems on both sides of the Atlantic made cheap passages possible. Chain migrations established transatlantic networks and a dense transatlantic communication. Once underway, the transatlantic mass movement developed a growing internal dynamic up to World War I, after which migration controls and restrictions began to curb the trend.

Up to the late 1880s, the 'classical' sending regions for mass emigration in the nineteenth century – excluding France, which was hardly affected – were the relatively well-developed industrial countries of western, central, and northern Europe. Once employment increased due to industrialization and the economic tensions toward the United States decreased, transatlantic migration from these regions slowed down in the 1890s, except in the case of Great Britain (Ireland). From the 1880s on, however, emigration from the southern, southeastern, and eastern European regions increased all the more, roughly corresponding to the north-to-south and west-to-east rate of industrialization. In the USA, this transatlantic movement soon came to be known as the 'new immigration.'

A gross total of 63 million (including returnees) and a net total of 50–55 million Europeans emigrated overseas from AD 1820 to 1915. The main destinations up to the late nineteenth century were in North America, with the US far ahead of Canada. New Zealand and Australia began to catch up in the 1860s, as did South American countries, which had had mixed populations since the colonial era, from the 1870s. Beginning in the 1880s and 1890s, Argentina and Brazil attracted large numbers of emigrants in the growing wave of migration from southern Europe, especially from Spain and Portugal, causing a fall in the percentage entering the United States – from about 80% prior to 1850, to about 75% from 1851 to 1890, and about 50% from then on.

European overseas emigration reached a climax between 1945 and 1960. Up to the mid-1960s, it still included more people than the immigration to Europe from Turkey, Asia, Africa, and the Caribbean (Münz, 1997: p. 225).

European Dimensions

Europe was really on the move in the nineteenth and twentieth centuries, but also in the Middle Ages and the early modern period. If we could look at spatial mobility from the Middle Ages up to the end of the twentieth century and at the same

time could eliminate reduced traveling times and longer distances due to innovations like the steamship, the railway, and the airplane, we probably would not see a major rise in mobility toward industrial and postindustrial Europe. In the Middle Ages, after all, the majority of the European population, at least for certain parts of their lifetimes, had to be mobile to survive; only a minority stayed at home living off subsistence farming or local employment throughout their lives (Schubert, 1995).

In the early modern period, various groups of European migrants, temporarily or permanently, moved over great distances over land or water. Among other groups there were migrating artists and artisans, architects and technical experts, seasonal or itinerant laborers and migrating tradesmen of fixed abode, travellers, laborers, mercenaries, sailors, and laborers for the maritime and colonial labor markets. There were settlement migrations recruited or invited by state authorities, e.g., the 'Peuplierung' of Prussia, the 'impopulation' of the Donau monarchy, and the colonial settlements in Russia under Katharine II. And there were the movements of religious refugees and expellees, who were often welcomed with open arms for economic reasons by the host authorities. All these long-distance movements were paralleled by an even greater number of short-distance moves between villages and small cities.

The three most relevant large-scale types of migration in modern Europe were: (1) flight and forced migrations in Europe; (2) economically motivated migrations in Europe; and (3) flight, minority, and economically motivated migrations to Europe (Page Moch, 1992; Bade, 2000).

1. Among flight and forced migrations, movements for religious but also for political reasons predominated during the European cultural crisis of the early modern era. In both the late eighteenth and especially in the nineteenth centuries, however, the most common reason for flight and exile was political persecution. Perceived as a threat to the European stability pact negotiated at the Congress of Vienna in 1815, all constitutional reform, national, and social movements were radically suppressed. Revolts and revolutions were crushed, causing a dramatic rise in the numbers of political refugees. The period from AD 1830 to 1848/1849, qualified the nineteenth century as the era of political exile. Still, the number of political refugees was low in comparison to that of new types of refugees, casualties of the epoch of the nation-state. The founding of the nation-states created minorities within national borders, thereby laying the foundations for the flight and forced migrations of the twentieth century, which would go down in history as the 'century of refugees.'

The series of benchmarks of the tragedy of flight and forced migrations in the twentieth century began with World War I, during which millions fled war zones or were deported. During the interwar period, about 1.5 million people fled from Revolutionary and Soviet Russia, while around 5 million were resettled as a result of the great 'population swap' carved out by the new nation-states that had emerged from the ashes of the three multi-ethnic empires. During World War II all this, however, was surpassed by the flight and deportations of 50–60 million people, out of whom 6 million Jews were killed in mass executions or industrialized mass murder in German concentration and extermination camps. Immediately after World War II came the expulsion of about 14 million Germans from the former Eastern territories of the German Reich and from the settlement districts of ethnic Germans in the East. After the era of flight from eastern to central and western Europe during the Cold War, the 'ethnic cleansing' of the interwar period and of World War II, was reenacted in the early 1990s in the expulsions and flight caused by the wars in the former multi-ethnic republic of Yugoslavia.

2. Besides migrant trade – very important up to the expansion of commodity markets in the nineteenth century – labor migrations were the most important among economically motivated movements. They evolved into predominantly agricultural migration systems with fixed circular movements (Lucassen, 1987). These systems were held together by long seasonal migration traditions between regions with virtually inverse requirements: the poor, rural, sending regions did not have enough work or had work at very low wages and had seasonally available surplus labor. The target areas, usually involved in intensive monoculture, offered seasonal work and much higher wages than the sending regions. Apart from agriculture, seasonal construction work in developing cities and their surrounding areas was also attractive.

Out of about 20 European labor migration systems operating at the turn of the eighteenth century, J. Lucassen reconstructed seven larger systems from a far earlier period. In these systems, more than 300 000 labor migrants, men and women, traveled up to 250–300 km at the turn of the century, within and across state borders. The most important of these migrating systems was the 'North Sea system' from the start of the seventeenth to the middle of the nineteenth century. Starting in the Netherlands, it spanned the whole North European coast. Besides supplying seasonal farm labor, the coastal harbors also provided access to the maritime and colonial labor markets of the North Sea system.

During the industrialization process, the magnetic field of migration in North Central Europe underwent a dramatic change in its powers of attraction. The North Sea system was overtaken by the new industrial coal and steel mining centers, especially in the Ruhr area and in Lorraine. In North Central Europe, the fall-off in seasonal agricultural migrations to the Netherlands was followed – up to World War I – by an increase in east-to-west migrations to the German Northeastern territories, with women comprising about 50% of the labor force on the huge East-Elbian estates of Prussia. Poles and Italians formed the main contingents of foreign labor in Germany and France, in the *longue durée* hinting at the large-scale inner-European south-to-north and east-to-west migrations which eventually became the hallmark of the second half of the twentieth century.

The two World Wars suspended transnational mobility in the new migration systems, changed it with the restructuring of the German Northeastern territories after World War I, and brought it to a complete standstill with the loss of the German Eastern territories after World War II. During World War II, labor migrations in Nazi Germany and in German-occupied Europe were largely replaced by the deportation and forced employment of disenfranchised slave laborers, who accounted for most of the 11 million 'displaced persons' (DPs) after World War II.

Only 10 years after a war that Germany had started, Germany and Italy signed the Labor Recruitment Treaty of 1955, paving the way for the future pan-European system of 'guest worker migrations.' Leaving Turkey aside, all the sending regions lay in the northern Mediterranean. Receiving areas were the highly industrialized and economically thriving countries of central, northern, and western Europe up to the first half of the 1970s, when, in the face of projected economic crises, recruitment bans and immigration restrictions were introduced.

Despite the high level of remigration, many labor migrants at that time settled in the host countries and sent for their families to join them. Short-term migrations became long-term ones, ultimately evolving into genuine immigration processes. Since the late 1970s, they have gradually transformed the host countries of 'guest worker migrations' into immigration countries.

In the former colonial nations, particularly in England, France, and the Netherlands, the role of 'guest workers' was first occupied by postcolonial immigrations, which were generally permanent from the outset. During the whole decolonization process, about 7 million people, including remigrating Europeans, migrated from the former European colonies to Europe after World War II. Profound changes took place, when, in the 1980s, the now thriving former south European sending regions of 'guest worker migrations' received increasing intercontinental south-to-north migrations. This ultimately transformed Europe as a whole from an emigration into an immigration continent.

3. The 'new immigrations' to Europe included both intercontinental south-to-north migrations and the new east-to-west migrations since the fall of the Iron Curtain. As in postcolonial migrations, privileged migrations legitimized by historical or cultural links to the host regions were predominant here. This mainly included minorities from the former Soviet Empire and its successor states, such as Armenians, ethnic Germans, and Jews (Bade and Oltmer, 1999; Fassmann and Münz, 2000). Apart from, and often coinciding with, non-European labor migrations to Europe, there were growing global migrations of asylum-seeking refugees in the south-to-north direction since the early 1980s, and, since the fall of the Iron Curtain, there have been increasing migrations in the east-to-west direction also.

Worldwide migration movements have increased in this age of globalization, media, and information networks. Yet, for the most part migrants remained in the surrounding areas of the sending regions and the percentage reaching Europe was still only about 5% at the end of the twentieth century. Nevertheless, horror visions of global mass migrations toward the continent have captured the European imagination, equating migration policy with security and defense policy.

Experts differ in their assessment of the 'migration pressure' from the southern and eastern regions. The key questions are whether it is even directed at Europe, whether it will gradually and inexorably grow, and whether it can be curbed by coordinated – i.e., global, not just European – intervention ('global governance') to control the causes of migration (Nuscheler, 1995; Opitz, 1997). From the entire range of conceivable strategies, Europe up to 2001 has done the least to tackle the causes of involuntary migration in the sending areas, and the most to combat flight migrations to Europe.

With freedom of movement within the European Union, internal 'vulnerability' grew – to use defense policy jargon – due to immigrations from outside the Community. The flip side of opening the borders has therefore been the increased closure of 'Fortress Europe.' Apart from private visits, tourism, and other short-term stays, the European defense system against non-European immigration only admits people welcome for economic, cultural, or other grounds, e.g., highly qualified specialists, scientists, artists, and people who are accepted as members of privileged postcolonial or ethnic minorities or have to be tolerated to some extent because of universalist or human-rights principles (family reunions, refugees, asylum seekers) (Santel, 1995).

In current migration debates and migration policies the tension has increased between self-descriptions and official ascriptions, i.e., between the way migrants perceive themselves and the identities assigned to them by immigration authorities. Migrants must do their best to fit into these assigned identities in order to have a chance of acceptance. Ascriptions, e.g., of 'refugee characteristics,' are the codes of administrative systems controlling and managing migrants' destinies. The decision on who is a 'real' refugee, depends on the fulfillment of these one-sided criteria. What matters most to asylum-seeking refugees is, therefore, often not what has happened to them, but whether their stories fit into the catalog of available ascriptions laid down by the host country. Hence, the approaches of migration policy and migration research to such conceptual problems of migration, may seem quite similar, – despite the fundamental conflict of interest concerning the ascriptions used on both sides (Bade, 2000).

See also: Culturally Appropriate Interventions in Social Work; Elder Abuse; Elder Care; Indigenous Social Work; Oppression; Social Work Theory.

Bibliography

Bade, K.J., 2000. Europa in Bewegung. Migration vom späten 18. Jahrhundert bis zur Gegenwart. CH Beck, Munich, Germany (Engl., French, Ital., Span. editions forthcoming 2001/02).

Bade, K.J., 2001. Historische migrations forschung. In: Oltmer, J. (Ed.), Migrationsforschung und Interkulturelle Studien. IO Jare IMIS, Osnabrueck.

Bade, K.J., 2003. Migration in European History. Blackwell Pub, Malden, MA.

Bade, K.J., Oltmer, J. (Eds.), 1999. Aussiedler – deutsche Einwanderer aus Osteuropa. Rasch, Osnabrück, Germany.

Baines, D., 1995. Emigration from Europe, 1815–1930. Cambridge University Press, Cambridge, MA.

Castles, S., Miller, M.J., 1998. The Age of Migration. International Population Movements in the Modern World, second ed. Macmillan, London.

Emmer, P.C., 1991. European expansion and migration: the European colonial past and intercontinental migration, an overview. In: Emmer, P.C., Mörner, M. (Eds.), European Expansion and Migration. Essays on the Intercontinental Migration from Africa, Asia, and Europe. Benz Publishers, New York, pp. 1–12.

Engerman, S.L., 1986. Servants to slaves to servants: contract labour and European expansion. In: Emmer, P.C. (Ed.), Colonialism and Migration. Indentured Labour before and after Slavery. Martinus Nijhoff Publishers, Boston, pp. 263–294.

Faist, T., 1998. International Migration and Transnational Social Spaces. Institut für Interkulturelle und Internationale Studien, Bremen, Germany.

Faist, T., 2000. The Volume and Dynamics of International Migration and Transnational Social Spaces. Clarendon Press, Oxford, UK.
Fassmann, H., Münz, R. (Eds.), 2000. Ost-west-wanderung in Europa. Böhlau, Vienna, Austria.
Fitzgerald, P., Lambkin, B.K., 2008. Migration in Irish History, 1607–2007. Palgrave Macmillan, Basingstoke, England.
Gerber, D.A., 2000. Theories and lives: transnationalism and the conceptualization of international migrations to the United States. In: Bommes, M. (Ed.), Transnationalismus und Kulturvergleich (IMIS-Beiträge), vol. 15. Rasch, Osnabrück, Germany, pp. 31–53.
Harzig, C., Hoerder, D., Gabaccia, D.R., 2009. What Is Migration History? Polity, Cambridge, UK.
Hoerder, D., 2001. Cultures in Contact. European and World Migrations, 11th Century to 1990s. Duke University Press, Durham, NC, forthcoming.
Lucassen, J., 1987. Migrant Labour in Europe, 1600–1900. The Drift to the North Sea. Croom Helm, London.
Lucassen, J., Lucassen, L. (Eds.), 1997. Migration, Migration History, History. Old Paradigms and New Perspectives. Peter Lang, Bern, Switzerland.
Lucassen, J., Lucassen, L., 2005. Migration, Migration History, History: Old Paradigms and New Perspectives. P. Lang, Bern.
Lucassen, J., Lucassen, L., Manning, P., 2010. Migration History in World History: Multidisciplinary Approaches. Brill, Leiden.
Manning, P., 2005. Migration in World History. Routledge, New York.
Münz, R., 1997. Woher – wohin? Massenmigration im Europa des 20. Jahrhunderts. In: Pries, L. (Ed.), Transnationale Migration (Soziale Welt, Sonderband 12). Nomos, Baden-Baden, Germany, pp. 221–243.
Nichols, R., 2008. A History of Pashtun Migration, 1775–2006. Oxford University Press, Oxford, England.
Northrup, D., 1995. Indentured Labour in the Age of Imperialism, 1834–1933. Cambridge University Press, Cambridge, UK.
Nugent, W., 1992. Crossings. The Great Transatlantic Migrations, 1870–1914. Bloomington, IN.
Nuscheler, F., 1995. Internationale Migration, Flucht und Asyl. Opladen.
Opitz, P.J. (Ed.), 1997. Der globale Marsch. Flucht und Migration als Weltproblem. C.H. Beck, Munich, Germany.
Pagden, A., 2001. Peoples and Empires: A Short History of European Migration, Exploration, and Conquest, from Greece to the Present. Modern Library, New York.
Page Moch, L., 1992. Moving Europeans. Migration in Western Europe since 1650. Indiana University Press, Bloomington, IN.
Pries, L. (Ed.), 1999. Migration and Transnational Social Spaces. Ashgate, Aldershot, UK.
Rodriguez, M.S., 2005. Repositioning North American Migration History: New Directions in Modern Continental Migration, Citizenship, and Community. University of Rochester Press, Rochester, NY.
Santel, B., 1995. Migration in und nach Europa. Opladen.
Schubert, E., 1995. Fahrendes Volk im Mittelalter. Bielefeld, Germany.
Tilly, C., 1978. Migration in modern European history. In: McNeill, W.H., Adams, R.S. (Eds.), Human Migration. Patterns and Policies. Indiana University Press, Bloomington, IN, pp. 48–72.
Wokeck, M., 1999. Trade in Strangers. The Beginning of Mass Migration to North America. Pennsylvania State University Press, University Park, PA.

Migration: Motivations

Douglas S Massey, Office of Population Research, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The motivations for human migration are diverse, but can generally be classified under five basic rubrics: material improvement, risk management, symbolic gratification, social connection, and threat evasion. Most theoretical work in recent decades has focused on motivations connected with material improvement, risk management, and social connection. Less attention has been paid to symbolic gratification and threat evasion. Recent work has suggested the latter motivation may have become increasingly salient in the twenty-first century given global climate change and rising civil violence.

Since our earliest origins as human beings, we have been a migratory species. Around 100 000 years ago, *Homo sapiens* first left East Africa and by around 12 000 years ago had populated all the continents save Antarctica, a uniquely wide and rapid spread in geological terms. Thus, migration is intrinsic to the human condition and geographic movement has always been a key adaptive strategy for human beings (Cavalli-Sforza and Cavalli-Sforza, 1995). The motivations for human migration are diverse, but can generally be classified under five basic rubrics: material improvement, risk management, symbolic gratification, social connection, and threat evasion. Most theoretical work in recent decades has focused on motivations connected with material improvement, risk management, and social connection. Less attention has been paid to symbolic gratification and threat evasion, but the latter has assumed greater importance as a motive for migration in the twenty-first century.

Material Improvement

The desire of people to improve their material circumstances has long been recognized as a key motivation for migration. In prehistoric times, individuals motivated by a desire for material gain moved to gain greater access to natural resources such as water, plants, and game (Massey, 2005). In modern times, they more often move to gain access to economic resources such as earnings, wealth, and education (Sjaastad, 1962). Migration motivated by a desire for material gain is most commonly theorized by neoclassical economics, which views the migratory decision as a cost-benefit analysis whereby rational, utility-maximizing actors balance the gains to be had by working at various geographic locations against the costs of migrating to these places (Todaro and Maruszko, 1987). Individuals maximize utility by moving to the location where the difference between earnings at origin and destination is greatest, net of the costs of migration.

The desire for material gain has also been theorized by the new economics of labor migration, which argues that in addition to maximizing absolute income people often move to increase their incomes relative to others in their communities of reference (Stark, 1991). The new economics framework also argues that people move to overcome missing, failed, or inaccessible markets for capital and credit. A lack of access to capital and credit often represents a barrier to mobility, especially in developing countries. For example, in many poor nations, mortgage markets are rudimentary or nonexistent and credit is difficult or expensive to obtain. In this case, a migrant might be motivated to move to a high-wage country to accumulate savings in order to acquire a home, buy land, start a business, or purchase consumer goods, while other household workers earn subsistence income locally. If the wages abroad are sufficiently high, the migrant might even be able to generate remittances sufficient to support the family while at the same time accumulating savings to finance socioeconomic mobility.

Risk Management

Whether motivated by a desire to maximize expected earnings or to finance purchases in the absence of working markets for capital or credit, the underlying motivation is one of material improvement, a core human motivation (Fiske, 2003). Although a desire for material gain is a core motive for individual action, human beings are also motivated by a strong desire to minimize losses and protect against risk. Psychologically, human beings experience much more pain from a material loss than pleasure from an equivalent material gain, a principle known as loss aversion (Tversky and Kahneman, 1991). Many developing nations lack well-developed insurance markets or government substitutes (unemployment or health insurance schemes). Under these circumstances, actors may employ migration as a means to self-insure against potential losses, and this motivation has also been theorized under the New Economics of Labor Migration (Stark, 1991).

Consider a household that is worried about local economic volatility. Such a household might act strategically to minimize risks to income by diversifying its labor portfolio, sending one member to work somewhere else in the country and another to work at a foreign destination while family members at home work in local labor markets or manage household assets or business enterprises. If conditions deteriorate locally to cause a drop in subsistence or income (because of a drought, recession, or plant closing), the shortfall can be buffered by money remitted by family members employed elsewhere. Sending out different members to work in geographically distinct labor

markets provides 'insurance' in the absence of effective markets or government programs, enabling households to rely on alternative income streams during local downturns. This strategy is feasible as long as economic conditions at origin and destination are not strongly and positively correlated (Massey et al., 1993).

Symbolic Gratification

In addition to moving in order to gain access to material resources, increase relative income, accumulate funds in the absence of capital and credit markets, and hedge against potential losses to income, people also move for purposes of symbolic gratification – to gain status, prestige, and esteem, which are also core human motivations (Fiske, 2003). A college professor may decide to move to another university to accept a named chair, for example, even if the salary remains constant or declines somewhat. Rural residents may be attracted by the bright lights and stimulating activities of a big city, which offers them the chance to become more 'cultured' and 'sophisticated.' Denizens of provincial cities may even be willing to accept a decline in income if they are able to gain access to the prestige and pleasures of the metropolis. Although the economic model of utility maximization can be applied to explain movement for symbolic gains, such motivations have also been treated by sociological theories of cultural capital, which point out that intangible resources such as status, prestige, and esteem are not only valuable motivators in themselves, but also constitute important resources than can be used to advance one's position in society (Bourdieu, 1986).

Social Connection

Human beings are also motivated to affiliate and interact with other people for the exchange of emotional resources such as liking, companionship, love, and sex, as well as to act collectively to achieve instrumental ends (Massey, 2005). *Homo sapiens* has always been a social species and individuals cannot survive, prosper, or reproduce outside of social groups. As a result, human actors are psychologically disposed to maintain and seek out social connections with others (Fiske, 2003). Throughout history, human beings have migrated to get married and married couples separated by migration have been strongly motivated to reunite, either through the return migration of the absent spouse or emigration of the trailing spouse. Such motivation may also be theorized by the neoclassical model of utility maximization, but most work on the social process of migration has been done under the rubric of social capital theory (Massey, 1990; Massey and Espinosa, 1997).

Because individuals are always socially connected to others – through networks defined by kinship, friendship, and acquaintance – migration not only motivates people to migrate for purposes of family reunification, but also confers instrumental value on ties to current or former migrants with experience in high-opportunity destinations (Palloni et al., 2001). The wife of a male migrant may seek to migrate simply to reunite with her husband, but she may also aspire to gain access to greater opportunities for employment and higher wages at the place of destination, or to overcome the patriarchal constraints of rural life (Hondagneu-Sotelo, 1994). Moreover, even if people do not wish to associate with a particular migrant, siblings, cousins, in-laws, and friends, they may be motivated to draw upon their tie to a migrant in order to facilitate migration for their own economic ends, whatever their feelings about family reunification (Massey and Sánchez, 2010). Likewise, a working migrant may learn of economic mobility opportunities at the place of destination, or be asked by a boss to find someone to fill a specific job, and then turn to his social network to fill the position by recruiting someone from his or her sending community (Hagan, 1994; Bashi, 2007).

In short, having a social connection to a migrant reduces the costs and risks of migration so that migratory behavior tends to spread rapidly through social networks, yielding a process known as cumulative causation (Massey, 1990). If someone wishes to migrate from a small Mexican town to the United States without authorization, the prospect is daunting if no one from the community has ever crossed the border before. In the absence of information or experience, undocumented border crossing can be a costly and risky proposition. Once someone from the community has migrated, however, everyone socially connected to that migrant acquires something quite valuable: a tie to someone who knows how to cross the border, find a job, and get by in the United States.

Depending on the closeness of the tie and the social obligations it implies, the migrant may provide information, offer job leads, recommend reliable border smugglers, finance the trip, accompany the person across the border, or offer food and lodging upon arrival. Whatever the specific set of services, the potential costs and risks of migration are reduced for those connected to a migrant, yielding valuable social capital, and some of these connections take advantage of their connection to migrate themselves, which expands the set of people with ties to a migrant, inducing still more to migrate, yielding a self-perpetuating migration chain to promote cumulative causation by reducing the costs and risks of migration from an ever expanding set of people (Massey and Zenteno, 1999).

Threat Evasion

The final motive for migration is the desire to escape an immediate threat to emotional or physical well-being, such as civil violence, family violence, crime, war, natural disasters, political upheaval, or catastrophic economic transformations, events that often produce a stream of out-migrants whose mobility is motivated by well-founded and acute fears for their immediate well-being. They perceive a tangible current or impending risk and move rapidly to escape it, usually but not always by proceeding to the nearest and most accessible safe haven (Kunz, 1973). If they remain within their home nation, they become internally displaced persons and if they cross an international border they become refugees or asylees. Although one might apply neoclassical economics to conceptualize refuge migration as

an exercise in utility maximization, the decision to migrate under threat is more primal than rational and analyses of refugee movements are generally been considered theoretically more in terms of classificatory schemes and conceptual typologies based on the nature of the threat, the form of displacement, and the context of reception (Kunz, 1981; Portes and Rumbaut, 2006).

All migrants are self-selected to some degree, but how refugee migrants are selected depends critically on the social, economic, and political circumstances of their departure, arrival, and prospects for return. Those fleeing an acute event such as a hurricane, an earthquake, or the sudden eruption of military hostilities are generally the least selected. In the most severe cases, everyone who is able moves and only the infirm, feeble, and abjectly impoverished are left behind. Political upheavals, crime, civil violence, and economic catastrophes are generally selective on the basis of traits such as class, political affiliation, ethnicity, and religion. It is difficult, however, to make a priori predictions about political selectivity without a more precise knowledge of the nature of the threat and the timing of its expression. For example, a left-wing revolution might produce the exodus of the property-owning elite (as in Cuba), whereas a right-wing crackdown might spur the departure of the poor (as in Guatemala).

Whatever the original nature of the selection, the incentive to integrate into the receiving society varies systematically with the prospects for return. When people believe that returning home is unlikely or impossible, they have a strong incentive to invest in social integration and economic mobility at the place of destination (Chiswick, 1979). If they believe return is imminent or likely in the short to medium term, they have less incentive to adapt to their new circumstances. If the diaspora proves to be prolonged and not resolved in either direction – settlement or return – refugees may languish in marginalization and poverty for many years or even generations, segregated from the host society but unable to return to their place of origin (Haddad, 2008).

Climate Change

Natural disasters, wars, and revolutions have occurred commonly throughout human history to displace people from their homelands and will no doubt continue to do so in the future. More recently, structural transformations of the economy during shifts from agrarian to industrial or industrial to postindustrial modes of production have long been known to displace people from traditional livelihoods and turn them into migrants (Massey, 1988; Sassen, 1988). In recent years, however, researchers have begun to pay closer attention to two potential drivers of migration that are beginning to appear to be endemic to the twenty-first century: climate change and civil violence.

The term 'environmental refugees' was introduced by El-Hinnawi (1985) to describe people forced to leave their place of origin, either temporarily or permanently, because of environmental disruptions triggered by human or natural events. Suhrke (1994) identified desertification, land degradation, deforestation, and rising sea levels as the most important forms of environmental change leading to out-migration, whereas Hugo (2008) lists environmental disasters, environmental degradation, climate change, and disruptions from large-scale human projects as the principal causes. Jacobson (1988) argues that environmental refugees constitute the single largest class of displaced persons in the world today. Myers and Kent (1995) estimate that some 25 million people were environmental refugees in the mid-1990s. Black (2001), however, questions the very concept and suggests that counts of environmental refugees are wildly inflated. Whatever the number of such refugees, Afifi and Warner (2008) found a positive association between the size of migration flows between 172 countries and measures of overfishing, desertification, water scarcity, soil salinization, deforestation, air pollution, soil erosion, and soil pollution within nations. Thus, migration indeed appears to be one response to environmental change.

Most studies rely on country-specific data rather than cross-national statistical analyses. In their study of migration in Guatemala and Sudan, for example, Bilsborrow and DeLargy (1991) found that a decline in land productivity fostered rural out-migration, whereas Hitztaler (2004) found that resource-poor villages in Russia's Kamchatka region displayed greater rates of out-migration than those with stable resource bases. Similarly, Massey et al. (2010) found that short-distance moves in Nepal were predicted by decreasing access to firewood, declines in agricultural productivity, and decreases in land cover. Bohra-Mishra and Massey (2011a) likewise found significant effects of population pressure, deforestation, and agricultural productivity on the likelihood of local mobility, but found little evidence that environmental deterioration promoted long-distance migration. Estimates derived by Feng et al. (2010), however, uncovered a strong relationship between declining crop yields and out-migration to the United States; and Munshi (2003) documented a significant connection between the scarcity of rainfall and US migration, a connection also observed by Nawrotzki et al. (2012).

Civil Violence

In an era of global climate change, therefore, environmental disruptions clearly carry the potential to promote movement, very clearly over shorter distances but also possibly over longer distances and even internationally. In addition to environmental change, another potential threat common to the contemporary world is civil violence emanating from diverse sources, including crime, political terrorism, narcoterrorism, guerilla insurgencies, revolutions, and state-sponsored repression; and as with environmental change, aggregate-level studies generally uncover a strong connection between such violence and out-migration.

Shellman and Stewart (2007) found, for example, that Haitian emigration to the United States was strongly correlated with surges in political violence, holding economic conditions constant. Morrison (1993) found similar results in Guatemala, as did Morrison and Lafaurie (1994) in Colombia. Jones (1989), however, found the effect of violence on migration was mainly indirect, with conflict producing local economic dislocations that, in turn, led to emigration. Morrison and May (1994) also found that conflict-related economic turmoil was more important than

violence in predicting internal migration within Guatemala, but Schultz (1971) showed that the effect of violence on rural–urban migration in Guatemala was small compared with the effects of socioeconomic and demographic factors.

In addition to these aggregate level studies, a growing number of investigators have examined the connection between violence and individual and household decisions to migrate. In their survey of displaced and nondisplaced persons in Colombia in 2000, for example, both Engel and Ibañez (2007) and Ibáñez and Vélez (2008) found that the threat of violence and the presence of paramilitary and guerilla groups were strongly associated with the likelihood of out-migration. Lundquist and Massey (2005) likewise found that the probability of leaving Nicaragua for the United States was strongly predicted by the intensity of the American Contra Intervention. In Mexico and Costa Rica, Alvarado and Massey (2010), however, found that the effects of lethal violence were small and generally associated with a lower probability of migration to the United States. In their study of Nepal, Bohra-Mishra and Massey (2011b) found that civil violence had nonlinear effects on migration, such that low to moderate levels of violence reduced the likelihood of movement, whereas high levels increased the probability of migration. The effects of violence were confined to local and within-country moves, however, and did not affect the likelihood of emigration from Nepal.

Conclusion

At this point, four of the five general categories of migrant motivations have been well-theorized. Material improvement is covered by neoclassical economics and the new economics of labor migration, risk management is theorized by the new economics of labor migration, symbolic gratification by cultural capital theory, and social connection by social capital theory. These reasons for migration all reflect core human motivations that psychologists have identified as intrinsic to the human condition (Fiske, 2003). Although threat evasion has not received systematic theoretical treatment to the same degree, self-preservation is clearly a fundamental human motivation and recent research has identified environmental change and civil violence as two potentially important drives of migration in the twenty-first century.

To date, most studies have considered climate change and violence as two independent threats, each of which might potentially contribute to the flow of migrants around the world. Recent work suggests, however, that climate change and civil violence are, in fact, causally interrelated. Indeed, Hsiang et al. (2011) show that outbursts of civil violence are closely tied to variations in the El Niño Southern Oscillation (ENSO). They found the probability of a civil conflict erupting doubles during El Niño versus La Niña years, and that the ENSO may have been behind 21% of all civil conflicts between 1950 and 2005. Given the potential of climate change to influence the frequency and severity of weather events such as El Niño, climate change not only has the potential to generate migration directly through displacement but also indirectly by triggering civil conflicts in affected areas throughout the world.

See also: Assimilation of Immigrants; Immigration: Political Aspects; Migration, Economics of; Migration, Theory of; Migrations, Colonizations, and Diasporas in Archaeology.

Bibliography

Afifi, T., Warner, K., 2008. The Impact of Environmental Degradation on Migration Flows across Countries. Working Paper No. 5. United Nations University Institute for Environment and Human Security.

Alvarado, S.E., Massey, D.S., 2010. In search of peace: structural adjustment, violence, and international migration. Annals of the American Academy of Political and Social Science 630, 294–321.

Bashi, V., 2007. Survival of the Knitted: Immigrant Social Networks in a Stratified World. Stanford University Press, Stanford, CA.

Bilsborrow, R., DeLargy, P.F., 1991. Population growth, natural resource use and migration in the Third World: the cases of Guatemala and Sudan. Population and Development Review 16, S125–S147.

Black, R., 2001. Environmental refugees: myth or reality. United Nations High Commissioner for Refugees Working Papers 34, 1–19.

Bohra-Mishra, P., Massey, D.S., 2011a. Environmental degradation and out-migration: new evidence from Nepal. In: Piguet, E., Pécoud, A., de Guchteneire, P. (Eds.), Migration and Climate Change. Cambridge University Press, Cambridge, UK, pp. 74–101.

Bohra-Mishra, P., Massey, D.S., 2011b. Individual decisions to migrate during civil conflict. Demography 48, 401–424.

Bourdieu, P., 1986. The forms of capital. In: Richardson, J.G. (Ed.), Handbook of Theory and Research for the Sociology of Education. Greenwood Press, New York, pp. 241–258.

Cavalli-Sforza, L.L., Cavalli-Sforza, F., 1995. The Great Human Diasporas: The History of Diversity and Evolution. Perseus Books, Cambridge, MA.

Chiswick, B.R., 1979. The economic progress of immigrants: some apparently universal patterns. In: Fellner, W. (Ed.), Contemporary Economic Problems. American Enterprise Institute, Washington, DC, pp. 357–399.

El-Hinnawi, E., 1985. Environmental Refugees. United Nations Environment Programme, Nairobi.

Engel, S., Ibañez, A.M., 2007. Displacement due to violence in Colombia: a household-level analysis. Economic Development and Cultural Change 55, 335–365.

Feng, S., Krueger, A.B., Oppenheimer, M., 2010. Linkages among climate change, crop yields and Mexico–US cross-border migration. Proceedings of the National Academy of Sciences 26. http://dx.doi.org/10.1073/pnas.1002632107.

Fiske, S.T., 2003. Social Beings: Core Motives in Social Psychology. Wiley, New York.

Haddad, E., 2008. The Refugee in International Society: Between Sovereigns. Cambridge University Press, New York.

Hagan, J.M., 1994. Deciding to be Legal: A Maya Community in Houston. Temple University Press, Philadelphia, PA.

Hitztaler, S., 2004. Changing human populations in post-Soviet Kamchatka: an integrated study of shifts in fertility and net population. Population and Environment 25 (4), 335–354.

Hondagneu-Sotelo, P., 1994. Gendered Transitions: Mexican Experiences of Immigration. University of California Press, Berkeley, CA.

Hsiang, S.M., Meng, K.C., Crane, M.A., 2011. Civil conflicts are associated with the global climate. Nature 476, 438–441.

Ibáñez, A.M., Vélez, C.E., 2008. Civil conflict and forced migration: the micro-determinants and welfare losses of displacement in Colombia. World Development 36, 659–676.

Jacobson, J., 1988. Environmental Refugees: A Yardstick of Habitability. World Watch Paper 86. World Watch Institute, Washington, D.C.

Jones, R.C., 1989. Causes of Salvadoran migration to the United States. The Geographical Review 79, 183–194.

Kunz, E.F., 1973. The refugee in flight: kinetic models and forms of displacement. International Migration Review 7, 125–146.

Kunz, E.F., 1981. Exile and resettlement: refugee theory. International Migration Review 15, 42–51.

Lundquist, J.H., Massey, D.S., 2005. Politics or economics? International migration during the Nicaraguan Contra War. Journal of Latin American Studies 37, 29–53.

Massey, D.S., 1988. International migration and economic development in comparative perspective. Population and Development Review 14, 383–414.

Massey, D.S., 1990. Social structure, household strategies, and the cumulative causation of Migration. Population Index 56, 3–26.

Massey, D.S., 2005. Strangers in a Strange Land: Humans in an Urbanizing World. W.W. Norton, New York.

Massey, D.S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., Taylor, J.E., 1993. Theories of international migration: a review and appraisal. Population and Development Review 19, 431–466.

Massey, D.S., Axinn, W.G., Ghimire, D.J., 2010. Environmental change and out-migration: evidence from Nepal. Population and Environment 32, 109–136.

Massey, D.S., Espinosa, K.E., 1997. Undocumented migration and the quantity and quality of social capital. Soziale Welt 12, 141–162.

Massey, D.S., Sánchez, M., 2010. Brokered Boundaries: Creating Immigrant Identity in Anti-Immigrant Times. Russell Sage Foundation, New York.

Massey, D.S., Zenteno, R., 1999. The dynamics of mass migration. Proceedings of the National Academy of Sciences 96, 5328–5335.

Morrison, A.R., 1993. Violence or economics: what drives internal migration in Guatemala? Economic Development and Cultural Change 41, 817–831.

Morrison, A.R., Lafaurie, M.P., 1994. Elites, guerrillas and narcotraficantes: violence and internal migration in Colombia. Canadian Journal of Latin American and Caribbean Studies 19, 123–154.

Morrison, A.R., May, R.A., 1994. Escape from terror: violence and migration in post-revolutionary Guatemala. Latin American Research Review 29, 111–132.

Munshi, K., 2003. Networks in the modern economy: Mexican migrants in the U.S. labor market. Quarterly Journal of Economics 118, 549–597.

Myers, N., Kent, J., 1995. Environmental Exodus: An Emergent Crisis in the Global Arena. The Climate Institute, Washington, DC.

Nawrotzki, R.J., Riosmena, F., Hunter, L., 2012. Do rainfall deficits predict U.S.-bound migration from rural Mexico? Evidence from the Mexican Census. Population Research and Policy Review. http://dx.doi.org/10.1007/s11113-012-9251-8.

Palloni, A., Ceballos, M., Espinosa, K.E., Spittel, M., Massey, D.S., 2001. Social capital and international migration: a test using information on family networks. American Journal of Sociology 106, 1262–1299.

Portes, A., Rumbaut, R.G., 2006. Immigrant America: A Portrait, third ed. University of California Press, Berkeley, CA.

Sassen, S., 1988. The Mobility of Labor and Capital: A Study in International Investment and Labor Flow. Cambridge University Press, Cambridge, UK.

Schultz, P.T., 1971. Rural-urban migration in Colombia. Review of Economics and Statistics 53, 157–163.

Shellman, S.M., Stewart, B.M., 2007. Predicting risk factors associated with forced migration: an early warning model of Haitian flight. Civil Wars 9, 174–199.

Sjaastad, L.A., 1962. The costs and returns of human migration. The Journal of Political Economy 70, 80–93.

Stark, O., 1991. The Migration of Labor. Basil Blackwell, Cambridge, UK.

Suhrke, Ai, 1994. Environmental degradation and population flows. Journal of International Affairs 47, 473–496.

Todaro, M.P., Maruszko, L., 1987. Illegal migration and US immigration reform: a conceptual framework. Population and Development Review 13, 101–114.

Tversky, A., Kahneman, D.P., 1991. Loss aversion in riskless choice: a reference dependence model. Quarterly Journal of Economics 106, 1039–1061.

Migration Out of Europe

Timothy J Hatton, University of Essex, Colchester, UK; and Research School of Economics, College of Business and Economics, Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides an outline of historical research on emigration from Europe to the New World in the century from 1850. It focuses first on the determinants of the volume and composition of emigration. Second, it examines the assimilation of immigrants and their integration in host country labor markets. The third topic is the rise of restrictive immigration policies and their role in the ending of mass emigration.

Introduction

International migration has been a key element in the development of the Western world since the industrial revolution. Its most profound influence has been on the peopling of the continents of North and South America and Australasia with emigrants of European stock and their descendants. Much of the discussion has focused on what might be called 'the age of mass migration,' from the middle of the nineteenth century to World War I. This was a period of relatively free migration that ended with the spread of immigration controls. Three sets of questions emerge from this experience. One is who migrated, when, and above all, what motivated migration. Another is how immigrants fared in the host country labor market, how their presence was felt by nonimmigrant workers and what were the broader effects on living standards and income distribution. Finally there is the issue of whether, and if so how, these evolving trends led to an immigration policy backlash.

Patterns of Migration

Early settlements of Europeans were progressively established in different parts of the New World from the sixteenth to the eighteenth century. In the century following 1820 about 55 million Europeans set sail for these resource-abundant and labor-scarce destinations and about three-fifths of them went to the United States. Earlier migration had been a trickle by comparison. Some of the early migrants were pioneer settlers seeking to establish new communities free from the religious or political persecution they faced in Europe. There were also streams of convicts and indentured servants sent to work on frontier farms and plantations. These were rapidly overtaken by the mounting numbers of free migrants but it was not until the middle of the nineteenth century that mass migration can really be said to have taken hold. Annual emigration rose from about 300 000 in the 1850s to the 1870s, climbing steeply thereafter to exceed a million at the turn of the century (Figure 1).

The emigrants in 1900 were rather different from those in 1800. Early nineteenth-century migrant streams were often led by farmers and artisans from rural areas traveling in family groups, intending to acquire land and settle permanently at the New World's frontier. In the late-nineteenth century, while many still had rural roots, the emigrants from any given country were increasingly drawn from urban areas and from nonagricultural occupations. About two-thirds were men and the overwhelmingly majority traveled as single individuals rather than in family groups. Of migrants to the United States between 1868 and 1910, 76% were between the ages of 15 and 40 years. While the young and single might be regarded as inherently more adventurous and more mobile, these characteristics also reflect a deeper economic calculus. By emigrating

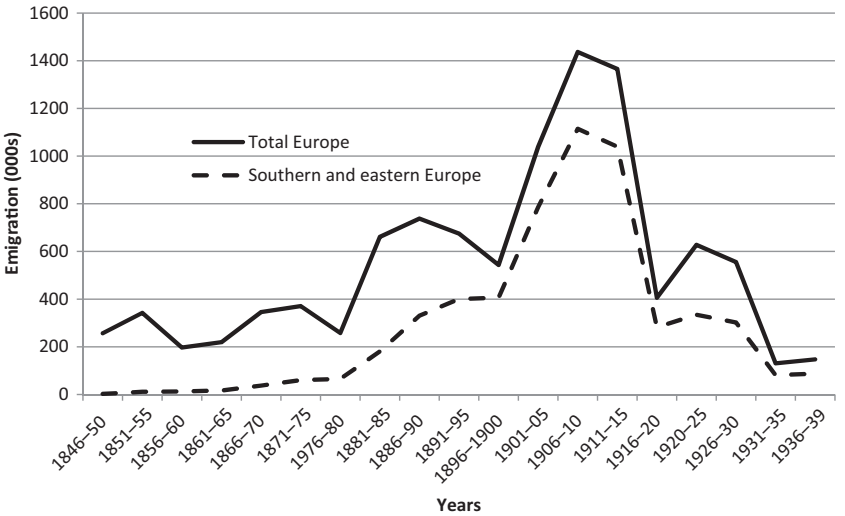

Figure 1 Emigration from Europe, 1846–1924 (5-year averages). Source: Kirk, D., 1946. Europe's Population in the Interwar Years. Princeton University Press, Princeton, p. 297.

as young adults they were able to reap the gains over most of their working lives. By moving as individuals, they were able to minimize the costs of the move, including earnings foregone during passage and job search. And since the emigrants were typically unskilled, they also had little country- or technology-specific human capital invested and hence stood to lose few of the rents from such acquired skills (except language). Finally, these young adults had the smallest commitment to family and assets at home (Hatton and Williamson, 1998: p. 11–12).

As mass emigration mounted so the composition by country of origin changed. In the middle decades of the nineteenth century, the emigrants came from north-western Europe, chiefly from Britain, Ireland, Germany, and the Scandinavian countries. But as **Figure 1** shows, the great surge in migration after 1880 was dominated by the rise in emigration from the countries of southern and Eastern Europe. Notable among these so-called 'new emigrant' countries were Italy, Spain, Austria-Hungary, and Russia. Statistics such as these hide the enormous variations in *rates* of emigration (per thousand of the source population). The highest was from Ireland with an average emigration rate of 13 per 1000 per annum between 1850 and 1913. Countries such as Sweden and Norway had rates approaching five per 1000 from 1870 to 1913 while the rates for Germany and Belgium were under two per 1000 and those for France were very small. Furthermore, the long run trends in emigration differed widely: from the 1880s rates of emigration declined sharply in Ireland, Germany and Norway while they underwent dramatic increase in Southern and Eastern Europe (Hatton and Williamson, 1998: p. 33).

The Determinants of Emigration

Various hypotheses have been offered to explain variations in emigration across time and place, drawing on perspectives from economics, sociology, demography and geography (Lowell, 1987: Ch. 2). One important fact that such theories must explain is that, during the course of modern economic growth in Europe, national emigration rates often increased, slowly at first from very low levels, rising more steeply to a peak, and then gradually falling. A one factor is rapid population growth, producing large cohorts of young adults whose opportunities to inherit smallholdings or to enter into family businesses were limited. Analysis of cross-country trends in emigration across Europe in the late nineteenth century finds strong evidence of these demographic effects (Hatton and Williamson, 1998: p. 43). There is also some evidence that the more urban the country, the higher was emigration – supporting the notion of greater mobility among urban populations.

Clearly, economic incentives were (and are) a key determinant of migration: average unskilled wage rates in 12 (western) European countries were barely more than half those in the New World (weighted by proportion of emigrants). Wage ratios were an important determinant of emigration rates. On average an increase in the New World real wage by 10% relative to that in Europe generated a rise in the emigration rate of 1.3 per 1000 (Hatton and Williamson, 1998: p. 43). But real wage ratios alone cannot explain why poor countries often had low emigration rates and why emigration often increased as development took place.

Other influences must also be taken into account. One is that many potential migrants were too poor to emigrate. As countries and regions industrialized, and as incomes rose, a larger share of potential migrants could afford to save up sufficient resources to finance the move and hence the number of emigrants increased. Thus in its early stages, increases in emigration could be positively associated with an increase in wage rates at home (even though this narrowed the foreign/home wage gap). As a result, European countries exhibited a long cycle of emigration with and upswing flowed by a downswing, although for many the cycle was not complete by 1913 (Hatton and Williamson, 1998: pp. 48–51).

Persistence and Volatility in Migration

Much of the literature has focused on migrant networks as the key element in determining mass migration. These have both social and economic dimensions. Once established, migration streams perpetuated themselves as previous migrants provided new migrants with prepaid tickets for the passage, food and shelter on arrival, and used their existing networks to help gain access to job opportunities. This 'friends and relatives effect' proved to be a very powerful force in the late-nineteenth century. Evidence from the United States immigration records suggests that as many as 90% of arrivals were traveling to meet a friend or relative who had previously emigrated. Econometric estimates indicate that for each 1000 migrants from a particular country, the friends and relatives effect 'pulled' between 20 and 100 further migrants each year, even when other factors such as relative wages are taken into account. Thus the larger the stock of previous emigrants, the greater would be the flow, and that flow would in turn lead to further flows.

Not only did the friends and relatives effect reduce the costs and uncertainties of migration to the prospective migrant, it also eased the poverty trap. This helps to explain different emigration patterns from countries at similar levels of development. In Ireland, the Great Famine effectively ejected a million migrants who formed a substantial migrant stock, particularly in the United States. Thus, after mid-century even the poorest Irish migrant would have benefited from the assistance of previous emigrants, so emigration depended less on the migrant's own resources and more on the potential gains (Hatton and Williamson, 2005: p. 65). As a result, Irish emigration declined as conditions in Ireland gradually improved. By contrast, in Italy, which had few emigrants before 1870 the poverty constraint was more important and hence emigration increased as the economy developed.

Superimposed on the long run trends in migration streams were sharp year-to-year fluctuations. Gross emigration sometimes fell by as much as half, only to recover a few years later. The literature on annual movements in emigration suggests that business cycle conditions, especially in the destination country, were the key influence and that other influences (such as wages) were relatively unimportant. This has been interpreted to mean that migrants were driven by the 'pull' effects of opportunities in the New World, rather than the 'push' effects of conditions at home. However, careful analysis shows that a model in which both the wage and the probability of employment matter, both at home and abroad can explain time series movements in

emigration reasonably well (Hatton, 1995; Hatton, and Williamson, 1998: Ch. 4). The proportionately larger effect of unemployment, particularly in the destination, reflects risk aversion on the part of potential migrants. The year-to-year volatility of migration can also be interpreted as reflecting the option value of waiting. Thus, even though it may be worthwhile to emigrate now, it may be better still to wait until conditions in the receiving country have improved.

Destination Choice and Return Migration

In the past, emigration streams from a given country were often dominated by one destination, for example, Scandinavian emigrants went almost exclusively to the United States. Consequently, choice among alternative destinations is less well-understood. Choice of location *within* a receiving country is associated with measures of regional income, as might have been expected, and especially with the stock of previous migrants to that state/region. But choice among countries involves additional factors such as cultural and linguistic affinity with the country of origin. Thus emigrants from Italy, Spain, and Portugal revealed much stronger preferences for South American countries such as Argentina and Brazil than did other European emigrants. Given these affinities and the pulling power of previous migrants, these streams persisted in spite of the income advantage of going instead to the United States. However, when new immigrant streams arose, such as that from southern Italy at the turn of the century economic advantage carried more weight. Thus migrants from the north of Italy continued to favor South America over North America (despite their urban backgrounds), while the rural southern Italians migrated in increasing numbers to the urban United States.

Although most migrants were permanent settlers, there were mounting flows of return migrants. By the end of the century, about one-third of European migrants to the United States were returning, usually after a few years, and even this may be an underestimate. Increasing destination wages relative to transport costs and falling voyage times contributed to the trend. But the upward trend in return migration owes most to the changing country composition of emigrants – particularly, the growing share from southern Europe. Many of these emigrants intended to return home and use their accumulated savings to establish families, and often to start farms or businesses. In such cases, the outward flows were more male dominated than where permanent settlement was the goal (Hatton and Williamson, 2005: p. 80). While return migration strategies are not well understood one thing is clear: differences in rates of return migration are associated more with the country of origin than with the country of destination. Thus, the high rates of return migration among southern Europeans at the turn of the century applied equally to those emigrating initially to North America and to South America.

Immigrant Assimilation

It has often been argued that immigrants faced discrimination and prejudice – factors that placed them under social and economic disadvantages and which may have made migration less attractive or encouraged return migration. The US Immigration Commission, which reported in 1911 argued that immigrants themselves, particularly the 'new immigrants' from southern and eastern Europe, were unable or unwilling to integrate into American society and that they lacked the skills and motivation to be successful in the labor market. Instead they crowded into ghettos and into unskilled occupations with little hope of upward mobility. Revisionist writers have argued instead that immigrant communities were not backward looking and isolationist. Rather, they provided a means though which immigrants could maintain their ethnic identity and benefit from social support networks at the same time as gaining access to the means of economic advancement (Bodnar, 1985).

A key issue is whether immigrants did suffer economic disadvantage and whether they caught up with the native born as they acquired knowledge, skills, and experience in the American labor market. The literature on immigrant earnings in the postwar period finds that, soon after arrival, immigrants suffer a significant wage disadvantage relative to the native born but that the difference erodes with time at the destination. Analysis for immigrants at the turn of the century indicates that (contrary to an older view) a similar pattern can be found among immigrants who arrived in adults in the United States around the turn of the century (Hatton and Williamson, 1998: Ch. 7). Those who arrived as children suffered little or no disadvantage and second-generation immigrants often had higher earnings than the native born. While the 'new immigrants' faced some economic disadvantages, much of this can be associated with limited education and English language proficiency – disadvantages that diminished over time. It is worth noting that this initial disadvantage was much less than that faced by immigrants to the United States from the developing world in the years after 1970 (Hatton, 2000).

A related issue is whether immigrants were among the best and the brightest from the source country as has often been assumed. That was undoubtedly true of the pioneer migrants but as mass migration gained momentum, immigrants from a given country became less skilled and educated. Those moving though migrant networks often needed less enterprise and initiative than those who went before. One recent study compares Norwegian male emigrants to the United States from 1865 to 1900 with the brothers that they left behind. The evidence suggests that, had they stayed at home, these emigrants would have attained lower occupational rank than their peers. On this measure at least the migrants were not necessarily the brightest and the best (Abramitsky et al., 2012). Across a range of source counties, the evidence suggests that the higher were the source country returns to education the less likely the more able and educated would emigrate (Baten and Stolz, 2012).

The Labor Market Impact of Immigration

The effect of mass immigration and labor market competition has been a contentious issue, in the past as in the present. A few studies have found that the wages of operatives grew more slowly in cities and regions with the highest rates of immigration (Goldin, 1994). But such effects were ameliorated by mobility, both by immigrants and nonimmigrants across locations, and especially with the expansion into lightly settled areas. Although immigrants often clustered in gateway cities and

other communities, the economy-wide effects of immigration were not easily observed by comparisons across localities.

A number of studies have estimates the labor market effects of immigration using general equilibrium models for the economy as a whole. These are stylized frameworks but they do allow the assessment of a wide variety of direct and indirect effects. One such exercise suggests that, in the absence of immigration from 1870 onward, US real wages would have been 34% higher in 1910 than they actually were (Hatton and Williamson, 1998: p. 212). For Argentina, where the inflow was even larger relative to the population, a different study finds that wages in 1914 would have been 25% higher in the absence of mass immigration after 1870 (Taylor, 1997).

However, such results depend on the counterfactual assumptions upon which such models are based. The most important of these is the international flow of capital. In the era up to 1914, capital flowed from the Old World to the New and especially to those countries that were absorbing immigrants. This helped to boost investment in infrastructure and in a wide range of other enterprises, which in turn prevented a decline in the ratio of capital to labor, and consequently slower growth in productivity and wages that might otherwise have occurred. If capital followed labor to the New World in search of high return then in the absence of mass migration much less capital would have been transferred. As a result, wages would have increased by less in the absence of immigration than in a scenario that does not incorporate the responsiveness of capital flows. To take one example, the effect of immigration to the United States on American wages would fall to 9%, if the responsiveness of capital is taken into account.

Real Wages at Home and Economic Convergence

For European countries, the effects of emigration were largely the opposite; labor supply grew less rapidly than otherwise. Emigration left a less dramatic imprint on population and society in Europe than immigration did in the New World. And it must be placed against a background of industrialization, urbanization, and the demographic transition. Nevertheless, the effects were not negligible. Estimates for several European countries, based on using the general equilibrium framework, suggest that wages would have been lower in 1913 by around to 10% in the absence of emigration. In the UK, unskilled wages would have been 12% lower by 1910 in the absence of emigration from 1870. But if capital outflows had diminished in response then the effect would have been only half as large.

Two implications follow from these economy-wide effects. One is that migration was a force for convergence in real wages and living standards within the so-called Atlantic economy. The available estimates show that in 1870 real wages in the New World were 96% higher than those of the Old World. In the ensuing 40 years that gap narrowed to around 79%. Some of this was due to the transfer of population and labor brought about by half a century of mass migration. One estimate suggests that in its absence the gap between New World and Old World wages would have increased; there would have been divergence rather than convergence (Taylor and Williamson, 1997: p. 40).

The second implication is the contrasting effects on income distribution. As the settler economies became more populated, land prices and rents increased, and combined with the pressure on wages, this caused their income distributions to widen. By contrast, in Europe, emigration tended to promote greater equality as wages rose and the rents and land prices stagnated. This process was enhanced by the ongoing trend of globalization in international trade, which like international migration, was underpinned by falling transport costs. In those European countries that responded by protecting their agricultural sectors with tariffs to blunt the effects of the so-called grain invasion (France, Germany, Spain) the decline in inequality was more muted (O'Rourke and Williamson, 1999: Ch. 6).

Migration and globalization also influenced wage differentials by skill. As migrants were largely unskilled, and increasingly so over time, the largest wage effects were felt among unskilled workers. One study found that net migration had the effect of increasing the skill premium in the United States but narrowing it in France, the UK, Spain, and Italy (Betrán and Pons, 2004). Not surprisingly, this effect is only observed after controlling for a wide range of other forces including trends in technical and structural change as well as key shifts in labor market institutions. Nevertheless, the effects of migration on incomes and their distribution were among the forces that led ultimately to rising barriers to migration.

Immigration Policy

As migration to the New World ascended to ever greater heights in the two decades before World War I, the pressures for restricting immigration mounted. And especially so in the United States, which imposed a literacy test in 1917 to be followed by the first immigration quota in 1921. It has been suggested that the rise in immigration sowed the seeds of its own destruction in the form of a policy backlash. This policy response can be better understood by asking two questions. First, who gains and who loses from migration? And second, who is in a position to do something about it (Foreman-Peck, 1992)?

The answer to the first question is that those most affected were typically low skilled blue-collar workers in the New World. The answer to the second question depends on who has the vote. As the franchise widened, it typically percolated down the hierarchy of class and income, diluting the political weight of landowners and capitalists. At the turn of the twentieth century, voting rates were about one-third for adults in the United States but less than 10% in Latin America where the *Latifundia* retained its grip on power (Engerman and Sokoloff, 2005). Consequently, immigration restrictions came earlier in the United States than in Latin America. In the British Dominions, which were relatively democratic, immigration policy was tightened, but not decisively until the early 1930s.

In the United States, the movement toward restriction began in the 1880s with a number of piecemeal restrictions such as the Chinese Exclusion Act (1882) the 'Gentlemen's agreement' with Japan (1907), tighter inspection, restrictions on contract labor and a head tax. In other settler countries there was a similar trend toward restrictions and exclusion, such as the White Australia Policy. The trend is also evident in the decline in the subsidized passages offered to immigrants, as in Australia after 1913 and Brazil between 1917 and 1927. There is some evidence that the hardening of immigration policies

was underpinned by slow real-wage growth and increasing inequality, as well as by the scale of immigration itself (Timmer and Williamson, 1998). But the timing of major restrictions was associated with big shocks: war and depression. In the United States, pressure for restrictions mounted over two decades before political interest groups finally coalesced to pass the 1917 Immigration Act (Goldin, 1994).

One interesting question is what underlay the public attitudes toward immigration. One possibility is that southern and European immigrants were seen increasingly as undermining the existing Anglo-Saxon culture, while those from Asia were an even greater threat. Another is simply the perceived effect of labor market competition in unskilled labor market. One test of this hypothesis suggests that the latter was uppermost in the United States during the 1890s. It was the scale of immigration rather than the source country composition that was the more important (Richardson, 2005). Labor market conditions were also important; the Emergency Quota was introduced in the United States just as the unemployment rate increased from 5.2% in 1920 to 11.7% in 1921.

The End of the Age of Mass Migration

The US Immigration Acts of 1921 and 1924, which introduced immigration quotas, are often seen as putting an abrupt end to the age of mass migration. Quotas based on country of origin fell particularly heavily on the countries of southern and eastern Europe. In Australia, Canada, and South Africa severe restrictions were introduced in the early 1930s. As **Figure 1** shows, in the 1920s emigration from Europe fell well below the peak decade before 1914, and in the 1930s it fell even further. However, it is likely that immigration from some European countries (those on the downswing of the emigration cycle) would have diminished anyway. The world depression of the 1930s, which was particularly severe in the New World, discouraged immigrants from most European countries except those fleeing totalitarian regimes.

The deglobalization of the world economy during the interwar period reversed some of the pre-War trends. The composition of emigrants shifted back toward northwest Europe, and their occupational skills increased, but there was also a shift toward family reunification (Hatton and Williamson, 2005: p. 186). The trend in the wage premium for skill was also reversed. This trend was more marked in the New World than the Old, and, it was also supported by the expansion of education.

In the post-World War II period, emigration revived and grew as prosperity returned to the world economy. But it did not match the heights of the decade before the World War I. One reason is that migration chains, which had been such an important factor in the earlier period, had been broken by 30 years of war and economic upheaval. Another was that under more buoyant economic conditions, prewar policies became more binding and were subsequently loosened, notably in Australia and Canada. Emigration from Europe grew over the following decades but not as fast as it had in the 50 years before 1914. This was partly due to continuing the rapid growth of living standards in Europe and their convergence on those of destination countries such as the United States.

From the 1960s United States, Canada, and Australia adopted immigration policies that discriminated less by source country and (in the latter two cases) more according to skill and qualifications. As a result, Europeans declined as a proportion of immigrants into the United States – from more than half in the 1950s to 10% in the 1980s (Hatton and Williamson, 2005: p. 208). This period also marked the transition of Europe from a region of emigration to one of immigration. At the same time countries in Asia, Africa, and the Caribbean entered a phase comparable with that of Europe in the age of mass migration.

See also: Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Ethnicity and Migration in Europe; Global Migration; Immigration Policy; Immigration and Crime in the United States; Immigration, Applied Research on; Immigration, Citizenship, and Integration: Social Work Connections; Immigration, Emigration, and Citizenship; Immigration: Health and Mortality; Immigration: Political Aspects; Immigration: Social Psychological Aspects; Migration History; Migration and Development; Migration and Health; Migration and Social Capital; Migration, Economics of; Migration, Ethnicity, Aging and Social Work Practice; Migration: Anthropological Perspectives; Migration: Cultural Aspects; Migration: Motivations; Migrations, Colonizations, and Diasporas in Archaeology.

Bibliography

Abramitsky, R., Platt-Boustan, L., Eriksson, A., 2012. Europe's tired, poor huddled masses: self-selection and economic outcomes in the age of mass migration. American Economic Review 102, 1832–1856.

Baten, J., Stolz, Y., 2012. Brain Drain in the age of mass migration: does relative inequality explain migrant selectivity? Explorations in Economic History 49, 205–220.

Betrán, C., Pons, M.A., 2004. Skilled and unskilled wage differentials and economic integration, 1870–1930. European Review of Economic History 8, 29–60.

Bodnar, J., 1985. The Transplanted: A History of Immigrants in Urban America. University of Indiana Press, Bloomington, IN.

Engerman, S.L., Sokoloff, K.L., 2005. The evolution of suffrage institutions in the new world. Journal of Economic History 65, 891–921.

Foreman-Peck, J., 1992. A political economy of international migration, 1815–1914. Manchester School 60, 359–376.

Goldin, C.D., 1994. The political economy of immigration restriction in the United States. In: Goldin, C., Libecap, G. (Eds.), The Regulated Economy: A Historical Approach to Political Economy. University of Chicago Press, Chicago, pp. 223–257.

Hatton, T.J., 1995. A model of U.K. Emigration, 1870–1913. Review of Economics and Statistics 77, 407–415.

Hatton, T.J., 2000. How much did immigrant 'Quality' decline in late nineteenth Century America? Journal of Population Economics 13, 509–535.

Hatton, T.J., Williamson, J.G., 1998. The Age of Mass Migration: Causes and Economic Effects. Oxford University Press, New York.

Hatton, T.J., Williamson, J.G., 2005. Global Migration and the World Economy: Two Centuries of Policy and Performance. MIT Press, Cambridge, MA.

Kirk, D., 1946. Europe's Population in the Interwar Years. Princeton University Press, Princeton.

Lowell, B.L., 1987. Scandinavian Exodus: Demography and Social Development of 19th Century Rural Communities. Westview Press, Colorado.

O'Rourke, K.H., Williamson, J.G., 1999. Globalization and History: The Evolution of a Nineteenth-Century Atlantic Economy. MIT Press, Cambridge, MA.

Richardson, G., 2005. The origins of anti-immigrant sentiments: evidence from the Heartland in the age of mass migration. B.E. Press Topics in Economic Analysis & Policy 5. Art 11.

Timmer, A.S., Williamson, J.G., 1998. Immigration policy prior to the 1930s: labor markets, policy interactions, and globalization backlash. Population and Development Review 24, 739–771.

Taylor, A.M., 1997. Peopling the Pampa: on the impact of mass migration to the river plate, 1870–1914. Explorations in Economic History 24, 100–132.

Taylor, A.M., Williamson, J.G., 1997. Convergence in the age of mass migration. European Review of Economic History 1, 27–63.

Migration to and from Latin America

Douglas S Massey, Office of Population Research, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Over the course of history, Latin America has been very strongly shaped by migration. The aboriginal peoples arrived around 11 000 years ago, and by 1492 had expanded to around 54 million people. The modern history of the region begins in 1492 with the arrival of the Spanish followed closely by the Portuguese, whose germs decimated the indigenous population and transformed Latin America into a region of European immigration. The inflow of European immigrants was modest during the colonial era, especially in comparison to the importation of African slaves, but immigration expanded after independence during the nineteenth and early twentieth centuries, and Argentina and Brazil eventually became immigrant receiving societies on a par with Canada and the United States. In the postwar period, only Argentina persisted as an immigrant receiving society, but its inflow now came from neighborhood countries in South America rather than Europe. Migration to the United States began in Mexico in the early twentieth century and spread to Puerto Rico in the 1940s and then to Cuba and the Dominican Republic in the 1960s. Migration to the United States from Central and South America dates mainly from the 1980s, but at the same time many migrants also began moving to Europe, especially from South America. Much of the migration to Europe has dried up since the Great Recession, but migration to the United States continues. Whereas much of the migration from Mexico was unauthorized during the period from 1965 to 2008, undocumented Mexican migration appears to have ended owing primarily to Mexico's demographic transition.

Early Migration Patterns

As with all nations in the western hemisphere, the countries of Latin America have been strongly shaped by international migration. The first immigrants, of course, were the aboriginal peoples who crossed the Bering Straits some 11 200 years ago (Dillehay, 2000). The next major immigration dates from 1492, when Christopher Columbus arrived on American soil. At that time the aboriginal population of the Americas is estimated to have been around 54 million, though some scholars put the figure much higher. By 1650, however, the aboriginal population had fallen to around 6 million owing mainly to the lack of resistance to infectious diseases brought by Europeans, but also by the disintegration of their civilizations in the wake of European colonization (Denevan, 1992).

In the centuries after 1492, the population of Latin America grew mainly from the arrival of the European colonists and births to them and their descendants. In addition to the Spanish, colonizers from Portugal, Britain, France, and the Netherlands vied for dominance in the Americas. Islands in the Caribbean ended up being colonized by all these nations, whereas the better part of North America fell under British control while Mexico, Central America, and most of South America became colonies of Spain, with the major exception of Brazil, which Pope Alexander gave to the Portuguese under the Treaty of Tordesillas in 1493 (Lockhart and Schwartz, 1983). Ever since then the Portuguese- and Spanish-speaking countries have collectively been known as Latin America.

Colonization by Spain and Portugal occurred mainly for the purpose of wealth extraction, and early immigrants from the mother countries were relatively small in number and disproportionately dominated by men, leading to considerable procreation with indigenous women to produce a new mestizo population (Lockhart and Schwartz, 1983). Of far greater consequence than the arrival of Spanish and Portuguese immigrants, however, was the large-scale importation of African slaves. According to Klein (1999), over the three centuries of the slave trade about 1.7 million Africans were brought into Spanish America and 3.9 million entered Brazil.

European immigration into Latin America during the Colonial Period generally remained modest, as the rulers of both Spain and Portugal prohibited the entry of immigrants from other European countries to guarantee control over their dominions (FitzGerald and Cook-Martín, 2014). The Wars of Independence from Spain were a drawn-out 15-year affair that began in 1810, but by 1825 all countries in Latin America had achieved independence and began to embark on various nation-building projects. Over the course of the nineteenth century, Latin American governments adopted policies to encourage immigration from Europe with the goal of modernizing and 'whitening' their populations (FitzGerald and Cook-Martín, 2014).

Although the success of these policies was modest in most nations, Argentina and Brazil did emerge as countries of mass immigration during the nineteenth century and by 1900 the percentage of foreign-born reached around 7% in Brazil and 30% in Argentina, compared to around 15% in the United States (FitzGerald and Cook-Martín, 2014). As in the United States, mass migration from Europe was brought to a halt by the First World War and after a brief revival in the 1920s finally ended for good with the onset of the Great Depression (Massey, 1995).

In the postwar period, with the sole exception of Argentina countries throughout Latin America were transformed from nations of immigration to nations of emigration (Durand and Massey, 2010). Although Argentina continued to take in immigrants, the source nations shifted from Europe to South America, led by Chile and then in decreasing order including Uruguay, Paraguay and Bolivia, and Peru (Massey et al., 1998). There was significant migration into Venezuela during the oil boom of the 1970s, especially from Colombia, but it has not

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31124-2

persisted, and Venezuela is now a country of emigration, particularly but not only to the United States.

Mexico–US Migration

Migration from Latin America in the modern period has, not surprisingly, been directed primarily to the United States. The oldest and largest flow of Latin Americans to the US is from Mexico. The southwestern US states, of course, were once part of Mexico, but only around 50 000 Mexicans remained north of the border after annexation (Jaffe et al., 1980) and once the border was finally agreed upon in 1854, there was little systematic emigration northward (Massey et al., 2002). Indeed, the border did not physically exist except as a line on a map, and movement back and forth was mostly local for decades (the Border Patrol was not founded until 1924).

Large scale migration from Mexico to the USA dates from the early twentieth century. During the late nineteenth century Mexican President Porfirio Díaz brought in American railroads in order to promote industrialization and development (Cardoso, 1980). The railroads sparked an economic boom in the western US by linking the region's mines, forests, and farms to the burgeoning industrial centers of the Northeast and Midwest, creating a huge demand for labor. The new rail lines also connected Mexico's mines and haciendas to the US markets to generate a rapidly growing trade. The first Mexican migrants to the United States were recruited by railroads around the turn of the century to replace the Chinese immigrants who had been barred from entry by the Chinese Exclusion Act of 1882 (Cardoso, 1980).

Migration from Mexico was modest until 1907, however, when the United States negotiated its 'Gentlemen's Agreement' with Japan to bring an end to Japanese immigration, which up to that point had been the principal source of farm labor (Massey et al., 2002). Desperate for labor, growers sent recruiters down the rail lines into Mexico looking for workers, initiating a new and much larger outflow northward. Whereas annual immigration from Mexico averaged just 1100 persons per year between 1900 and 1907, by 1912 the figure had grown to 22 000.

Once the United States entered the First World War in 1917 and military mobilization drained manpower from the civilian economy, labor demand boomed; and when the United States enacted restrictive quotas to block mass European migration in 1920, the demand for Mexican labor surged again to produce what historians have called a 'flood tide' of Mexican immigration to the United States (Cardoso, 1980). During the 1920s, the annual immigration from Mexico rose to 50 000 persons per year with another 15 000 arriving on temporary work visas. From 1907 through 1929, a total of 686 000 legal immigrants and 226 000 temporary workers entered the United States, with many of the latter simply staying on to work when their visas expired (Massey et al., 2002).

What ended the flood tide was the Great Depression, which not only eliminated labor demand, but also ushered in a program of mass deportations (Hoffman, 1974; Balderrama and Rodríguez, 2006). From 1929 to 1939, 469 000 Mexican immigrants were forcibly repatriated and many more returned home on their own, cutting in half the number of Mexicans living in the United States (Massey et al., 2002). Once the United States entered the Second World War, however, war mobilization once again created labor shortages, especially in agriculture, and the United States replaced the welcome mat by creating the Bracero Program, a binational accord that arranged for the annual entry of Mexican workers on temporary visas (Galarza, 1964).

Although originally deemed a temporary wartime measure, at the behest of growers it was expanded during the postwar years to reach a peak in the late 1950s when the number of Bracero entries peaked at 450 000, with another 50 000 entering as legal immigrants (Calavita, 1992; Massey et al., 2002). In the early 1960s, however, Congress began to wind down the program and formally eliminated it in 1965. Over the course of the program's existence, however, 4.7 million Braceros entered the country, along with around 600 000 legal immigrants.

In addition to ending the Bracero Program, in 1965 the US Congress acted to impose the first ever numerical limits on immigration from the western hemisphere, capping the total at 120 000 per year. By 1976, the hemispheric cap had been replaced by a single worldwide cap of 290 000 and all countries were given an annual quota of just 20 000 legal immigrant visas. From the late 1950s to the late 1970s, Mexico thus went from access to 450 000 temporary visas and 50 000 permanent resident visas to just 20 000 resident visas and no work visas. Economic conditions on both sides of the border had not changed, however, and over the course of the Bracero years hundreds of thousands of migrants had established connections with employers north of the border and elaborated well-developed social networks (Massey et al., 1987). As a result, when opportunities for legal entry evaporated, migrants simply continued to migrate without authorization, ushering in the era of undocumented migration (Massey et al., 2002).

From 1965 to 1979 the number of undocumented entries steadily rose but then stabilized as the outflows reached the levels observed in the late 1950s, except that now the migrants were labeled as 'illegal' (Massey and Pren, 2012). Being 'illegal' they were by definition 'lawbreakers' and 'criminals' and framed by politicians and bureaucrats as a grave threat to the nation, bringing about the rise of a 'Latino threat narrative' in the mass media (Chavez, 2001, 2008). In response to the perceived threat, Congress allocated more funds, equipment, and personnel to the Border Patrol, which of course, produced more apprehensions even though the actual inflow had stabilized. More apprehensions were taken as self-evident proof of the accelerating 'invasion,' which brought more resources to the Border Patrol, thereby generating more apprehensions (Massey and Pren, 2012).

During the 1980s and 1990s, therefore, rising apprehensions and enforcement entered a mutually reinforcing cycle of circular that ratcheted border enforcement up to unprecedented levels. Paradoxically, however, the exponential growth of border enforcement had little effect on the number of undocumented migrants entering the United States. Instead it had a profound effect on the number of migrants returning to Mexico. Prior to the mid-1980s, migration from Mexico had been overwhelmingly circular. From 1965 to 1985 around 85% of undocumented entries were offset by departures, and

the net annual increase in the undocumented population was small (Massey and Singer, 1995). Even among legal immigrants, return migration was common. Estimates suggest that during the 1970s, annual out-migration of legal Mexican immigrants averaged about 20% of legal in-migration (Warren and Kraly, 1985). Of those legal immigrants who entered the country in 1970, around 56% were estimated to have returned by 1979 (Jasso and Rosenzweig, 1982).

The militarization of the border, however, dramatically increased the costs and risks of undocumented border crossing; and in response, migrants quite logically stopped crossing the border – not by staying in Mexico but by remaining in the United States once they had achieved entry (Massey et al., 2002). During the 1990s and 2000s, circulation ceased as the probability of return migration plummeted (Rendall et al., 2011). The inevitable results was a boom in the undocumented population which rose from under 2 million in 1988 to peak at about 12 million in 2008 (Wasem, 2011).

Over the same period 5.4 million legal immigrants entered from Mexico, about 2.3 million of whom were former undocumented migrants who had legalized under the provisions of the 1986 Immigration Reform and Control Act, the rest of whom were mostly sponsored by family members who were US citizens. Parents, spouses, and minor sons and daughters of citizens are eligible for entry outside of the annual quota, and as US policies became increasingly hostile toward noncitizens, legal Mexican immigrants naturalized in record numbers.

The key date was 1996, when US legislation declared legal immigrants ineligible for many social benefits, declared them deportable for often minor offenses, and abridged the rights of noncitizens to challenge removal orders (Zolberg, 2006). While the number of Mexican naturalizations averaged just 23 000 per year before 1996, afterward the average rose to 123 000, creating 2.1 million new citizens with the right to bring in relatives outside the quotas. At the same time, the US Congress quietly began expanding temporary worker migration from Mexico. Whereas in 1993 only 23 000 temporary worker entries were recorded, by 2008 the figure had reached 361 000.

In 2008, of course, the Great Recession brought the US economy to a standstill, and unemployment soared. Between 2008 and 2009, the number of undocumented residents fell by a million and since then has fluctuated around 11 million. Although temporary worker migration and legal immigration from Mexico also dipped, they have since revived but undocumented migration from Mexico appears to have stopped, with few people entering and few of those already present in the United States leaving (Passel et al., 2012).

The apparent end of undocumented migration is partially explained by the weakness of US labor demand and the expansion of opportunities for legal entry, but the principal reason is the dramatic decrease in the rate of labor force growth within Mexico (Hanson and McIntosh, 2010). The current fertility rate is roughly at replacement level and Mexico has become an aging society, dramatically reducing the potential supply of Mexican workers. Circularity has resumed within the Mexico–US migration system, but now most of those moving back and forth across the border do so in legal status, either as temporary workers or as legal immigrants.

Non-Mexican Migration from Latin America

Aside from Mexico, the principal sources of migrants from Latin America to the United States prior to 1980 were Puerto Rico, Cuba, and the Dominican Republic. As US citizens, Puerto Ricans are technically not immigrants and can come and go as they please. The boom years of Puerto Rican migration were from 1940 to 1970, when the island was transforming from a rural to an urban-industrial economy and when the demand for workers in New York's service and garment industries was strong. Since 1970, however, net migration has fluctuated around zero and the island-born population is shrinking and the second and third generations are increasingly moving away from New York (Acosta-Belen and Santiago, 2006).

Immigration from Cuba and the Dominican Republic are both products of the Cold War. Cuban emigration to the United States was spurred by Castro's socialist revolution, which drove out the entrepreneurial, professional, and middle classes during the 1960s and 1970s (Portes and Bach, 1985). Beginning with the Mariel boat lift in 1980, however, Cuban migrants have increasingly been drawn from the lower middle and working classes (Portes and Stepick, 1993). Mass migration from the Dominican Republic dates from 1965, when the United States occupied Santo Domingo to quell a left-wing rebellion led by students, and the ambassador there was ordered to make residence visas generously available to Dominican students and other members of the middle class to get them out of the country, seeding a legal inflow that proceeds to this day (Riosmena, 2010). From 1980 to 2010, some 674 000 Cubans and 939 000 Dominicans entered the United States as legal immigrants.

Migration from Central America to the United States dates from the 1980s when the United States intervened militarily in the region in an effort to topple the leftist Sandinista regime in Nicaragua and suppress leftist insurgencies in El Salvador, Guatemala, and Honduras. The resulting wave of civil violence and economic turmoil generated large flows of refugees heading toward the United States. Whereas Nicaraguans, fleeing a left-wing regime, were welcomed and granted an easy path to legal permanent residence, Salvadorans, Guatemalans, and Hondurans had the misfortune to flee countries governed by right-wing regimes allied with the United States and were therefore deemed economic rather than political refugees and most entered without authorization. As of 2010, around two-thirds of all immigrants from these countries living in the USA were undocumented (cf Hoefer et al., 2011; Acosta and de la Cruz, 2011). Between 1980 and 2010, 205 000 legal immigrants entered from Nicaragua, compared with 1.2 million from El Salvador, Guatemala, and Honduras, helping to turn Hispanics into the nation's largest minority group.

Migration from South America also dates from the 1980s and is attributable mainly to the debt crisis and the 'lost decade' of the 1980s when economic growth flattened and structural adjustment policies were imposed by global lenders in Washington, DC (Massey et al., 2006), though in the case of Colombia out-migration also occurred in response to political violence (Silva and Massey, 2015). From 1980 to 2010, legal immigration from South America totaled around 1.7 million,

with the largest contributors being Colombia (518 000), Peru (320 000), Ecuador (251 000), Brazil (201 000), Venezuela (133 000), and Argentina (100 000). The only country with significant undocumented migration to the United States is Ecuador, whose undocumented population stood at 180 000 in 2010 (Hoefer et al., 2011).

During the 1980s and 1990s migrants from Central and South America also began migrating to Europe in significant numbers (Durand and Massey, 2010). Naturally the largest number went to Spain, where around 1.1 million Latin Americans had settled by 2005 (Padilla and Piexoto, 2007), but there were also significant Latin American populations in Italy (205 000), Britain (113 000), Germany (94 000), Portugal (56 000), and France (47 000). Of these migrants, by far the largest share (88%) came from South America, with the leading sources being Ecuador, Colombia, Bolivia, Peru, Argentina, and Brazil. A significant share of migration from Argentina and Brazil, the two largest countries of European immigration during the early twentieth century, were able to reclaim a European nationality, obtain a passport, and then migrate to Europe as EU citizens. Durand and Massey (2010) have labeled this form of movement transgenerational migration, for citizenship is typically reclaimed through an immigrant grandparent from Germany, Italy, Spain, or Portugal, which extend nationality to third generation descendants of persons born in those countries. Since the Great Recession, Latin American migration to Europe, and particularly Spain and Portugal, has slowed down to a standstill and in some cases reversed.

See also: Global Migration; Immigration Policy; Immigration, Applied Research on; Immigration, Citizenship, and Integration: Social Work Connections; Immigration: Social Psychological Aspects; International Migration and Remittances; Migration and Social Capital; Migration, Theory of; Migration: Anthropological Perspectives; Migration: Cultural Aspects; Migration: Motivations.

Bibliography

Acosta-Belen, Edna, Santiago, Carlos E., 2006. Puerto Ricans in the United States: A Contemporary Portrait. Lynne Rienner, Boulder, CO.

Acosta, Yesenia D., de la Cruz, G. Patricia, 2011. The Foreign Born from Latin America and the Caribbean: 2010. American Community Survey Brief 10-15. U.S. Bureau of the Census, Washington, DC.

Balderrama, Francisco E., Rodríguez, Raymond, 2006. Decade of Betrayal: Mexican Repatriation in the 1930s. University of New Mexico Press, Albuquerque.

Calavita, Kitty, 1992. Inside the State: The Bracero Program, Immigration, and the I.N.S. Routledge, New York.

Cardoso, Lawrence, 1980. Mexican Emigration to the United States 1897–1931. University of Arizona Press, Tucson.

Chavez, Leo R., 2001. Covering Immigration: Population Images and the Politics of the Nation. University of California Press, Berkeley.

Chavez, Leo R., 2008. The Latino Threat: Constructing Immigrants, Citizens, and the Nation. Stanford University Press, Stanford, CA.

Denevan, William M., 1992. The Native Population of the Americas in 1492. University of Wisconsin Press, Madison.

Dillehay, Thomas D., 2000. The Settlement of the Americas a New Prehistory. Basic Books, New York.

Durand, Jorge, Massey, Douglas S., 2010. New world orders: continuities and changes in Latin American migration. Annals of the American Academy of Political and Social Science 630, 20–52.

FitzGerald, David S., Cook-Martín, David, 2014. Culling the Masses: The Democratic Origins of Racist Immigration Policy in the Americas. Stanford University Press, Stanford, CA.

Galarza, Ernest, 1964. Merchants of Labor: The Mexican Bracero Story. NcNally and Loftin, Santa Barbara.

Hanson, Gordon H., McIntosh, Craig, 2010. The great Mexican emigration. Review of Economics and Statistics 92 (4), 798–810.

Hoefer, Michael, Rytina, Nancy, Baker, Bryan, 2011. Estimates of the Unauthorized Immigrant Population Residing in the United States: January 2010. U.S. Office of Immigration Statistics, Washington, DC.

Hoffman, Abraham, 1974. Unwanted Mexican Americans in the Great Depression: Repatriation Pressures, 1929–1939. University of Arizona Press, Tucson.

Jaffe, Abraham J., Cullen, Ruth M., Boswell, Thomas D., 1980. The Changing Demography of Spanish Americans. Academic Press, New York.

Jasso, Guillermina, Rosenzweig, Mark R., 1982. Estimating the emigration rates of legal immigrants using administrative and survey data: the 1971 cohort of immigrants to the United States. Demography 19 (3), 279–290.

Klein, Herbert S., 1999. The Atlantic Slave Trade. Cambridge University Press, Cambridge, UK.

Lockhart, James, Schwartz, Stuart B., 1983. Early Latin America: A History of Colonial Spanish America and Brazil. Cambridge University Press, Cambridge, UK.

Massey, Douglas S., 1995. The new immigration and the meaning of ethnicity in the United States. Population and Development Review 21, 631–652.

Massey, Douglas S., Alarcon, Rafael, Durand, Jorge, Gonzalez, Humberto, 1987. Return to Aztlan: The Social Process of International Migration from Western Mexico. University of California Press, Berkeley and Los Angeles.

Massey, Douglas S., Joaquín Arango, Graeme Hugo, Kouaouci, Ali, Pellegrino, Adela, Edward Taylor, J., 1998. Worlds in Motion: International Migration at the End of the Millennium. Oxford University Press, Oxford.

Massey, Douglas S., Behrman, Jere R., Sanchez, Magaly (Eds.), 2006. Chronicle of a Myth Foretold: The Washington Consensus in Latin America. Sage Publications, Thousand Oaks, CA.

Massey, Douglas S., Durand, Jorge, Malone, Nolan J., 2002. Beyond Smoke and Mirrors: Mexican Immigration in an Age of Economic Integration. Russell Sage Foundation, New York.

Massey, Douglas S., Pren, Karen A., 2012. Unintended consequences of U.S. immigration policy: explaining the post-1965 surge from Latin America. Population and Development Review 38, 1–29.

Massey, Douglas S., Singer, Audrey, 1995. New estimates of undocumented Mexican migration and the probability of apprehension. Demography 32, 203–213.

Padilla, Beatriz, Peixoto, Joao, 2007. Latin American Immigration to Southern Europe. Migration Information Source, Washington, DC.

Passel, Jeffrey, Cohn, D'Vera, González-Barrera, Ana, 2012. Net Migration from Mexico Falls to Zero-and Perhaps Less. Pew Hispanic Center, Washington, DC.

Portes, Alejandro, Bach, Robert L., 1985. Latin Journey: Cuban and Mexican Immigrants in the United States. University of California Press, Berkeley.

Portes, Alejandro, Stepick, Alex, 1993. Miami: City on the Edge. University of California Press, Berkeley.

Rendall, Michael, Brownell, Peter, Kups, Sarah, 2011. Declining return migration from the United States to Mexico in the Late-2000s recession: a research note. Demography 48 (3), 1040–1068.

Riosmena, Fernado, 2010. Policy shocks: on the legal auspices of Latin American migration to the United States. Annals of the American Academy of Political and Social Science 630, 270–293.

Silva, Carolina, Douglas, S., Massey, 2015. Violence, networks, and international migration from Colombia. . International Migration, forthcoming. Published online at. doi:10.1111/imig.12169.

Warren, Robert, Kraly, Ellen P., 1985. The Elusive Exodus: Emigration from the United States. Population Trends and Public Policy No. 8. Population Reference Bureau, Washington, DC.

Wasem, Ruth E., 2011. Unauthorized Aliens Residing in the United States: Estimates since 1986. Congressional Research Service, Washington, DC.

Zolberg, Aristide R., 2006. A Nation by Design: Immigration Policy in the Fashioning of America. Russell Sage Foundation, New York.

Migration, Theory of

Douglas S Massey, Office of Population Research, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Theorists in a variety of disciplines have offered theoretical explanations for human migration. This article reviews the leading models that have been developed to theorize five facts of human migration: the structural forces that promote 'out-migration' from sending regions; the structural forces that attract 'in-migrants' to receiving societies; the motivations, goals, and aspirations of people who respond to these structural forces by becoming migrants; the social and economic structures that arise to connect areas of out- and in-migration; and the actions taken by states to influence the number and characteristics of immigrants. These models include neoclassical economics, the new economics of labor migration, segmented labor-market theory, social capital, cumulative causation, world-systems theory, and securitization theory.

Social scientists have theorized five basic aspects of human migration in their efforts to explain it: the structural forces that promote 'out-migration' from sending regions; the structural forces that attract 'in-migrants' to receiving societies; the motivations, goals, and aspirations of people who respond to these structural forces by becoming migrants; the social and economic structures that arise to connect areas of out- and in-migration; and the actions taken by states to influence the number and characteristics of immigrants. The ensuing sections review the theoretical perspectives that address these issues, and then offer a theoretical synthesis that includes empirically supported propositions from each one.

Neoclassical Economics

The oldest and best-known theoretical model, neoclassical economics, argues that migration is caused by geographic differences in the supply of and demand for labor (Lewis, 1954; Ranis and Fei, 1961). A region with a large endowment of labor relative to capital will have a low equilibrium wage, whereas an area with a limited endowment of labor relative to capital will have a high wage. The resulting gap causes workers from the low-wage or labor-surplus area to move to the high-wage or labor-scarce region. As a result of this movement, the supply of labor decreases and wages rise in the capital-poor area, while the supply of labor increases and wages fall in the capital-rich area, leading, at equilibrium, to a geographic wage differential that exactly reflects the costs of interregional movement, pecuniary and psychic.

Associated with this macroeconomic theory is an accompanying microeconomic model of individual choice (Todaro, 1976). In this scheme, rational actors decide to migrate because a cost-benefit calculation leads them to expect a positive net return, usually monetary, from movement. Migration is conceptualized as a form of investment in human capital (Sjaastad, 1962). People choose to move where they can be most productive, given their skills; but before they can reap the higher wages associated with greater labor productivity, they must invest in the material costs of travel, the costs of self-support while moving and looking for work, the effort involved in learning a new environment and possibly a different culture, the difficulty experienced in adapting to a new labor market, and the psychological costs of cutting old ties and forging new ones.

Potential migrants estimate the costs and benefits of moving to alternative locations and migrate to where the expected discounted net returns are greatest over their projected working lives. Net future returns are estimated by taking earnings observed in the destination area and multiplying them by the probability of obtaining a job there to derive 'expected destination earnings,' which are then subtracted from earnings estimated for the place of origin (observed earnings times the probability of employment). The difference is then summed over future years and discounted by a factor that reflects the greater utility of money earned in the present than in the future. From this integrated difference, the estimated costs are subtracted to yield the expected net returns to migration. In theory, a potential migrant goes to wherever the expected net returns are greatest.

The New Economics of Migration

In recent years, a 'new economics of labor migration' has arisen to challenge the assumptions and conclusions of neoclassical theory (Stark, 1991). A key insight of this approach is that migration decisions are not made by isolated individuals, but within larger units of interrelated people – typically families or households but also entire communities – and that people act collectively not only to maximize expected income, but also maximize status within an embedded hierarchy to overcome barriers to capital and credit, and to minimize risk.

In developed countries, risks to household income are managed through institutional mechanisms. Crop insurance programs and futures markets give farmers a way of protecting themselves from natural disasters and price fluctuations, whereas unemployment insurance and government transfer programs protect workers against recessions and job loss through structural transformation. Private or government-sponsored retirement programs, meanwhile, offer citizens a means of insuring against the risk of poverty in old age.

In the absence of such programs, households are in a better position than individuals to control risks to economic well-being. They can easily diversify sources of income by allocating different family workers to different labor markets. As

long as economic conditions across labor markets are negatively or weakly correlated, households minimize risk by geographically diversifying their labor portfolios. In the event that economic conditions at home deteriorate and productive activities there fail to generate sufficient income, the household can rely on migrant remittances for support.

Markets for credit and capital are incomplete or inaccessible in many settings, and in the absence of an efficient banking system, migration becomes attractive as a strategy for accumulating funds in lieu of borrowing. Families send one or more workers to a higher wage area to accumulate savings or send them back in the form of remittances. Although most migrant savings and remittances go toward consumption, some of the funds are also typically channeled into productive investment.

A key proposition of the new economic model is that income is not a homogeneous good. The source of the income really matters, and households have significant incentives to invest scarce family resources in activities that provide access to new income sources, even if these activities do not increase total income. The new economics also questions the assumption that income has a constant effect on utility – i.e., a $100 real increase in income means the same thing to a person regardless of community conditions and irrespective of his or her position in the local income distribution. Instead, it holds that households send workers abroad not only to improve absolute incomes, but also to increase them relative to others, and, hence, to reduce relative deprivation compared with some reference group (Stark, 1991). A household's sense of relative deprivation depends on the total income earned by households above it in the income distribution. If utility is negatively affected by relative deprivation, and even if a household's income and expected gains from migration remain unchanged, it may acquire an incentive to migrate if there is a change in other households' incomes.

Segmented Labor-Market Theory

Although neoclassical theory and the new economics of labor migration offer divergent explanations for the origins of migration, both ultimately reduce to microlevel decision models. What differs are the units assumed to make the decision (the individual or the household), the entity being maximized or minimized (income vs risk), the assumptions about the economic context of decision-making (complete and well-functioning markets vs missing or imperfect markets), and the extent to which the migration decision is contextual (whether income is evaluated in absolute terms or relative to some reference group).

Standing distinctly apart from these models is segmented labor-market theory, which argues that migration stems from the intrinsic character of modern industrial societies. According to Piore (1979), migration is not caused by push factors in sending regions (low wages or high unemployment), but by pull factors in receiving areas (a chronic and unavoidable need for migrant workers). The built-in demand for inexpensive and flexible workers stems from three basic features of advanced industrial economies.

The first is structural inflation. Wages not only reflect conditions of supply and demand, but also confer status and prestige that social qualities inhere in jobs. In general, people believe that wages should reflect social status, and they have clear notions about the correlation between occupational status and pay. As a result, employers are not entirely free to respond to changes in the supply of workers. A variety of informal social expectations and formal institutional mechanisms (union contracts, civil service rules, bureaucratic regulations, company job classifications) ensure that wages correspond to the hierarchies of prestige and status that people perceive and expect.

The demand for cheap, flexible labor is also augmented by social constraints on motivation within occupational hierarchies. Most people work not only to generate income, but also to accumulate social status. Acute motivational problems arise at the bottom of any job hierarchy because there is no status to be maintained and there are few avenues for upward mobility. The problem is structural, because the bottom cannot be eliminated from the labor market. Mechanization to eliminate the lowest and least desirable class of jobs will simply create a new bottom tier composed of jobs that used to be just above the bottom rung. Since there must always be a bottom of any hierarchy, motivational problems are inescapable.

Thus, employers need workers who view bottom-level jobs simply as a means to the end of earning money, and for whom employment is reduced solely to income, with no implications for status or prestige. For a variety of reasons, migrants satisfy this need. They begin as target earners, seeking to earn money for a specific goal that will improve their status or well-being at home – building a house, paying for school, buying land, acquiring consumer goods. Moreover, the disjuncture in living standards between origin and destination communities often means that low urban or foreign wages appear to be generous by the standards of the sending society.

The demand for migrant labor also stems from economic dualism. Capital is a fixed factor of production that can be idled by lower demand but not laid off; owners of capital must bear the costs of its unemployment. Labor is a variable factor of production that can be released when demand falls, so that workers bear the costs of their own unemployment. Whenever possible, therefore, capitalists seek out the stable, permanent portion of demand and reserve it for the employment of equipment, whereas the variable portion of demand is met by adding labor. Thus, capital-intensive methods are used to meet basic demand, and labor-intensive methods are reserved for the seasonal, fluctuating component. This dualism creates distinctions among workers, leading to a bifurcation of the labor force.

Workers in the capital-intensive primary sector get stable, skilled jobs working with the best equipment and tools. Employers are forced to invest in these workers by providing specialized training and education. Their jobs are complicated and require considerable knowledge and experience to perform well, leading to the accumulation of firm-specific human capital. Primary-sector workers thus tend to be unionized or professionalized, with contracts that require employers to bear a substantial share of the costs of their idling (severance pay and unemployment benefits). Because of these costs and continuing obligations, workers in the primary sector become expensive to let go; they become more like capital.

In the labor-intensive secondary sector, however, workers hold unstable, unskilled jobs; they may be laid off at any time

with little or no cost to the employer. Indeed, the employer will generally lose money by retaining workers during slack periods. During down cycles, the first thing secondary-sector employers do is to cut payroll. As a result, employers force workers in this sector to bear the costs of their unemployment. They remain a variable factor of production and are, hence, expendable.

The inherent dualism between labor and capital extends to the labor force in the form of a segmented market structure. Low wages, unstable conditions, and the lack of reasonable prospects for mobility in the secondary sector make it difficult to attract local workers, who are instead drawn into the primary, capital-intensive sector, where wages are higher, jobs are more secure, and there is a possibility of occupational improvement. To fill the shortfall in demand within the secondary sector, employers turn to migrants.

In their analysis of the process by which Cuban immigrants were incorporated into the United States, Portes and Bach (1985) uncovered evidence of a third possible sector that blends features of primary and secondary labor markets. Like the secondary sector, ethnic enclaves contain low-status jobs characterized by low pay, chronic instability, and unpleasant working conditions – jobs that are routinely shunned by natives. Unlike the secondary sector, however, they provide immigrants with significant economic returns to education and experience, as well as the very real prospect of upward socioeconomic mobility, thus replicating features of the primary sector.

The existence of a large, concentrated ethnic population creates a demand for specialized cultural products and services that immigrant entrepreneurs are uniquely qualified to fill. In addition, privileged access to the growing pool of migrant labor gives them an advantage when competing with mainstream firms. Migrants working in the enclave trade low wages and the acceptance of strict discipline upon arrival for a greater chance of advancement and independence later on. In order to function effectively over time, therefore, an ethnic enclave requires a steady stream of new arrivals willing to trade low initial wages for the possibility of later mobility, yielding an independent segmented source of labor demand for migrant workers, complementing that emanating from the secondary sector.

World-Systems Theory

Growing out of the historical–structural tradition in social science, world-system theory argues that migration stems from the penetration of capitalist economic relations into noncapitalist or precapitalist areas to create a mobile population (Sassen, 1988). Driven by a desire for higher profits and greater wealth, owners and managers of capitalist firms in core regions enter poor peripheral areas in search of land, raw materials, labor, and consumer markets. Migration emerges in response to the disruptions and dislocations that inevitably occur in the process of capitalist development.

In order to achieve the greatest profit from existing agrarian resources, and to compete within global commodity markets, capitalist farmers seek to consolidate landholding, mechanize production, introduce cash crops, and apply industrially produced inputs such as fertilizer, insecticides, and high-yield seeds. Land consolidation destroys traditional systems of land tenure based on inheritance and common ownership. Mechanization decreases the need for manual labor and makes many agrarian workers redundant. The substitution of cash crops for staples undermines traditional, social, and economic relations based on subsistence; and the use of modern inputs produces high crop yields at low unit prices. All of these forces drive peasants off the land to create a mobile labor force of people displaced from the land.

The extraction of raw materials for sale on global markets requires industrial methods that rely on paid labor. The offer of wages to former peasants undermines traditional forms of social and economic organization based on norms of reciprocity and fixed role relations, and creates incipient labor markets based on new conceptions of individualism, private gain, and social change. At the same time, capitalist firms enter developing regions to establish assembly plants, often within special export-processing zones created by sympathetic governments. The demand for factory workers strengthens local labor markets and weakens traditional productive relations. The insertion of foreign-owned factories into peripheral regions undermines regional economies by producing goods that compete with those made locally; feminizing the workforce without providing factory-based employment opportunities for men; and socializing women for industrial work and modern consumption without providing a lifetime income capable of meeting these needs. The result once again is the creation of a population that is socially and economically uprooted and prone to migration.

The same capitalist economic processes that create migrants in peripheral regions simultaneously attract them into certain highly developed urban areas. The investment that drives economic globalization is managed from a small number of global cities, whose structural characteristics create a strong demand for migrant labor. In order to ship goods, deliver machinery, extract and export raw materials, coordinate business operations, and manage expatriate assembly plants, capitalists in global cities build and expand transportation and communication links to the peripheral regions where they have invested. These links not only facilitate the movement of goods, products information, and capital, but also promote the movement of people by reducing the costs of movement along certain pathways.

The creation and perpetuation of a global trading regime also requires an underlying system of international security. Core capitalist nations have both an economic interest in, and the military means of preserving, geopolitical order, and leading powers thus maintain relatively large armed forces to deploy as needed to preserve the integrity of the global capitalist system. Threats to the stability of that system are met by military force projected from one or more of the core nations. Each military base and armed intervention, however, creates a range of social and political connections that promote the subsequent movement of migrants.

Finally, processes of economic globalization also create ideological or cultural links between core capitalist regions and their peripheries. In many cases, these cultural links are long-standing, reflecting a colonial past in which core countries established administrative and educational systems that mirrored their own in order to govern and exploit a peripheral region. Ideological connections are presently reinforced by mass communications and advertising campaigns directed

from the core urban centers. The diffusion of core cultural patterns and the spread of modern consumption interact with the emergence of a transportation–communication infrastructure to channel migrants disproportionately to global cities.

Social-Capital Theory

Although it was Loury (1977) who first introduced the concept of social capital, it was Bourdieu (1986) who pointed out its broader relevance to human society. According to Bourdieu and Wacquant (1992: p. 119), "social capital is the sum of the resources, actual or virtual, that accrue to an individual or a group by virtue of possessing a durable network of more or less institutionalized relationships of mutual acquaintance and recognition." The key characteristic of social capital is its convertibility – it may be translated into other forms of capital, notably financial capital. People gain access to social capital through membership in networks and social institutions, and then convert it into other forms of capital to improve or maintain their position in society (Coleman, 1988).

Massey et al. (1987: p. 170) were the first to identify migrant networks as a form of social capital. In keeping with Coleman's (1990: p. 304) dictum that "social capital … is created when the relations among persons change in ways that facilitate action," they identified migration itself as the catalyst for change in the nature of social relations. Everyday, ties of friendship and kinship provide few advantages, in and of themselves, to people seeking to migrate. Once someone in a personal network has migrated, however, the ties are transformed into a resource that can be used to gain access to employment and high wages at points of destination. Each act of migration creates social capital among people to whom the new migrant is related, thereby raising the odds of their migration.

Goss and Lindquist (1995) point to migrant institutions as a structural complement to migrant networks, arguing that interpersonal ties are not the only means by which social capital is created and that "migration is best examined not as a result of individual motivations and structural determinations, although these must play a part in any explanation, but as the articulation of agents with particular interests and playing specific roles within an institutional environment…" (p. 345). For-profit organizations and private entrepreneurs provide services to migrants in exchange for fees set on the underground market: surreptitious smuggling across borders; clandestine transport to internal destinations; labor contracting between employers and migrants; counterfeit documents and visas; arranged marriages between migrants and legal residents or citizens of the destination country; and lodging, credit, and other assistance in countries of destination. Over time, individuals, firms, and organizations become institutionally stable and offer another source of social capital to would-be migrants.

Cumulative Causation

The theory of cumulative causation basically argues that human migration changes individual motivations and social structures in ways that make additional movement progressively, likely, a process first identified by Myrdal (1957) and reintroduced to the field by Massey (1990). Causation is cumulative in the sense that each act of migration alters the context within which subsequent migration decisions are made to make additional trips of longer duration more likely. Social scientists have discussed eight ways that migration has cumulatively caused: the expansion of networks, the distribution of income, the distribution of land, the organization of agriculture, culture, the regional distribution of human capital, the social meaning of work, and the structure of production. Feedbacks through other variables are also possible, but have not been systematically treated.

The State Filter

Prior to 1914, the flow of immigrants around the world was relatively free and unconstrained. Although receiving nations imposed qualitative controls that restricted entry on the basis of health, morality, race, and ethnicity, international migration was not numerically restricted and passports did not exist (Zolberg, 2006). Numerical restrictions on immigration and passport controls began to be imposed during the 1920s, and since 1945, national states have sought to regulate both the numbers and characteristics of those who enter their nations for work and residence.

State policies thus actively shape the size and composition of immigrant flows into around the world. By the end of the twentieth century, a consensus had emerged that policies tend to become more restrictive during periods of economic dislocation and distress, during periods when the volume of immigration is or has been high, as well as during periods of ideological insecurity and political conformity (Goldin, 1994; Timmer and Williamson, 1998; Meyers, 2004). Since the events of 11 September 2001, however, immigration has been increasingly 'securitized' by politicians and pundits who frame it as an existential threat conflated with global terrorism, yielding draconian enforcement measures (Bourbeau, 2011; d'Appollonia, 2012).

The securitization of immigration began earliest in the United States, when Ronald Reagan incorporated border control into the rhetoric of the Cold War and framed the Mexico–US border as a bulwark against communists from Central America (Massey et al., 2002); but securitization expanded dramatically during the 1990s in response to a series of terrorist attacks, culminating in 9/11 and the passage of the USA PATRIOT Act (Massey and Pren, 2012). The degree to which restrictive immigration policies actually work depends on a state's capacity to implement and enforce them, of course (Massey, 1999), and given the diffuse, open, decentralized system of the United States, the progressive militarization of its southern border after 1986 backfired by curtailing circular migration, which had long been a characteristic feature of the North American migratory system. In response to rising costs and risks of unauthorized border crossing, Mexicans stopped circulating not by staying in Mexico but by remaining in the United States once the costs and risks along the border had been experienced, thereby increasing the net rate of undocumented migration and accelerating undocumented population growth.

A Theoretical Synthesis

Contemporary migration appears to originate in the social, economic, political, and cultural transformations that accompany the penetration of capitalist markets into nonmarket or premarket societies (world-systems theory). In the context of a globalizing economy, the entry of markets into peripheral regions disrupts existing social and economic arrangements and brings about the displacement of people from customary livelihoods, creating a mobile population of workers who actively search for new ways of earning income, managing risk, and acquiring capital. Most of these people seek to assure their well-being is by selling their labor on emerging markets (neoclassical economics).

Because wages are generally higher in urban than in rural areas, much of this process of labor commodification is expressed in the form of rural–urban migration; but wages are even higher, in developed countries overseas, and the larger size of these transnational wage differentials inevitably prompts some adventurous people to sell their labor on international markets by moving abroad for work. International wage differentials are not the only factor motivating people to migrate, however. Evidence also suggests that people displaced in the course of economic transformation move not simply in order to reap higher lifetime earnings by relocating permanently to a foreign setting. Rather, households struggling to cope with the jarring transformations of early economic development also use temporary labor migration as a means of managing risk and overcoming barriers to capital and credit (the new economics of labor migration).

While the early phases of economic development promote emigration, transformations in advanced postindustrial economies yield a bifurcation of labor markets. Jobs in the primary labor market provide steady work and high pay for native workers, but jobs in the secondary sector offer low pay, little stability, and few opportunities to generate a structural demand for migrant workers (segmented labor-market theory). This process of labor market bifurcation is most acute in global cities, where a concentration of managerial, administrative, and technical expertise leads to a concentration of wealth and a strong ancillary demand for low-wage services (world-systems theory). Unable to attract native workers, employers turn to immigrants and initiate immigrant flows directly through formal recruitment (segmented labor-market theory).

Although often instrumental in initiating migration flows, recruitment becomes less important over time because the same processes of economic globalization that create mobile populations in developing regions, and which generate a demand for their services in certain cities, also create links of transportation, communication, politics, and culture to make the movement of people increasingly cheap and easy (world-systems theory). However when a migration stream begins, it displays a strong tendency to continue because of the growth and elaboration of networks (social-capital theory). The concentration of migrants in certain destination areas creates a 'family and friends' effect that channels later cohorts to the same places and facilitates their arrival and incorporation. If enough migrants arrive under the right conditions, an enclave economy may form, which further augments the specialized demand for migrant workers (segmented labor-market theory).

The spread of migratory behavior within sending communities sets off ancillary structural changes, shifting distributions of income and land, and modifying local cultures in ways that promote additional out-migration. Over time, the process of network expansion may become self-perpetuating because each act of migration causes social and economic changes that promote additional movement (the theory of cumulative causation). As receiving areas implement restrictive policies to counter the rising tides of migrants, they create a lucrative niche into which enterprising agents, contractors, and other middlemen move to create migrant-supporting institutions, providing migrants with yet another infrastructure capable of supporting and sustaining movement (social-capital theory). In cases where state capacity to impose immigration controls is limited, social capital formation and cumulative causation can promote migration despite restrictions, which may backfire by deterring out-migration rather than in-migration.

During the initial phases of out-migration, the effects of capital penetration, market failure, social network expansion, and cumulative causation dominate to cause a rapid upsurge of the flow; but, as the level of emigration reaches high levels, and the costs and risks of movement drop, movement is increasingly determined by wage differentials (neoclassical economics) and labor demand (segmented labor-market theory). As economic growth in sending areas occurs, wage gaps gradually diminish and well-functioning markets for capital, credit, insurance, and futures come into existence, progressively lowering the incentives for emigration. If these trends continue, the country or region ultimately becomes integrated into the global or national economy, whereupon it undergoes a transformation: massive out-migration progressively trails off and shifts to a pattern of circulatory movement whose net balance fluctuates in response to changing economic conditions at origin and destination (Hatton and Williamson, 1998), yielding a characteristic curve in the course of the mobility transition that Martin and Taylor (1996) have called the "migration hump."

See also: Assimilation of Immigrants; Immigration: Political Aspects; Migration Out of Europe; Migration and Health; Migration, Economics of; Migration: Cultural Aspects; Migrations, Colonizations, and Diasporas in Archaeology.

Bibliography

d'Appollonia, A.C., 2012. Frontiers of Fear: Immigration and Insecurity in the United States and Europe. Cornell University Press, Ithaca, NY.

Bourbeau, P., 2011. The Securitization of Immigration: A Study of Movement and Order. Routledge, New York.

Bourdieu, P., 1986. The forms of capital. In: Richardson, J.G. (Ed.), Handbook of Theory and Research for the Sociology of Education. Greenwood, New York, pp. 241–258.

Bourdieu, P., Wacquant, L.J.D., 1992. An Invitation to Reflexive Sociology. University of Chicago Press, Chicago.

Coleman, J.S., 1988. Social capital in the creation of human capital. American Journal of Sociology 94S, 95–120.

Coleman, J.S., 1990. Foundations of Social Theory. Belknap Press of Harvard University Press, Cambridge, MA.

Goldin, C., 1994. The political economy of immigration restriction in the U.S. 1890 to 1921. In: Golden, C., Libecap, G. (Eds.), The Regulated Economy: A Historical Approach to Political Economy. University of Chicago Press, Chicago.

Goss, J., Lindquist, B., 1995. Conceptualizing international labor migration: a structuration perspective. International Migration Review 29, 317–351.
Hatton, T.J., Williamson, J.G., 1998. The Age of Mass Migration: Causes and Economic Impact. Oxford University Press, Oxford.
Lewis, W.A., 1954. Economic development with unlimited supplies of labour. Manchester School of Economic and Social Studies 22, 139–191.
Loury, G.C., 1977. A dynamic theory of racial income differences. In: Wallace, P.A., LaMond, A.M. (Eds.), Women, Minorities, and Employment Discrimination. Lexington Books, Lexington, MA, pp. 152–186.
Martin, P.L., Taylor, J.E., 1996. The anatomy of a migration hump. In: Taylor, J.E. (Ed.), Development Strategy, Employment, and Migration: Insights from Models. Organisation for Economic Co-operation and Development, Development Centre, Paris, pp. 43–62.
Massey, D.S., 1990. Social structure household strategies, and the cumulative causation of migration. Population Index 56, 3–26.
Massey, D.S., 1999. International migration at the dawn of the twenty-first century: the role of the state. Population and Development Review 25, 303–323.
Massey, D.S., Alarcón, F., Durand, J., González, H., 1987. Return to Aztlan: The Social Process of International Migration from Western Mexico. University of California Press, Berkeley, CA.
Massey, D.S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., Taylor, J.E., 1998. Worlds in Motion: Understanding International Migration at the End of the Millennium. Oxford University Press, Oxford.
Massey, D.S., Durand, J., Malone, N.J., 2002. Beyond Smoke and Mirrors: Mexican Immigration in an Age of Economic Integration. Russell Sage Foundation, New York.
Massey, D.S., Pren, K.A., 2012. Unintended consequences of US immigration policy: explaining the post-1965 surge from Latin America. Population and Development Review 38, 1–29.
Meyers, E., 2004. International Immigration Policy: A Theoretical and Comparative Analysis. Palgrave MacMillan, London.
Myrdal, G., 1957. Rich Lands and Poor: The Road to World Prosperity. Harper and Row, New York.
Piore, M.J., 1979. Birds of Passage: Migrant Labor and Industrial Societies. Cambridge University Press, New York.
Portes, A., Bach, R.L., 1985. Latin Journey: Cuban and Mexican Immigrants in the United States. University of California Press, Berkeley.
Ranis, G., Fei, J.C.H., 1961. A theory of economic development. American Economic Review 51, 533–565.
Sassen, S., 1988. The Mobility of Labor and Capital: A Study in International Investment and Labor Flow. Cambridge University Press, Cambridge, UK.
Sjaastad, L.A., 1962. The costs and returns of human migration. Journal of Political Economy 70S, 80–93.
Stark, O., 1991. The Migration of Labor. Blackwell, Oxford, UK.
Timmer, A.S., Williamson, J.G., 1998. Immigration policy prior to the 1930s: labor markets, policy interactions, and globalization backlash. Population and Development Review 24, 739–772.
Todaro, M.P., 1976. Internal Migration in Developing Countries: A Review of Theory, Evidence, Methodology, and Research Priorities. International Labour Office, Geneva.
Zolberg, A., 2006. A Nation by Design: Immigration Policy in the Fashioning of America. Harvard University Press, Cambridge, MA.

Migrations, Colonizations, and Diasporas in Archaeology

Marcel Otte, University of Liège, Liège, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Demographic density, as much as the ideas of conquest and discovery, has always led to population movements since the origins of humanity. These are evidenced by stylistic and symbolic processes, as much as in art and technology, as in ways of life. Their systematic reconstruction enables the creation of 'historical' maps from the distant past to modern peoples today.

Humanity possesses an efficient adaptive ability due to the technological and cultural changes that it has accomplished. These changes have permitted hominid groups to occupy varied habitats that would have been incompatible with their natural anatomical capacities alone. From the beginning, one of the essential criteria defining humanity was mobility. The appeal of displacement was apparently based on the requirements of thought: courage, imagination, and challenge. Through time, various kinds of migrations were undertaken; in each case, it was a matter of overcoming new, successively encountered, constraints: steppe, tundra, and oceans. We distinguish seasonal migrations (hunting) from definitive displacement (colonization). We also recognize the cases of diaspora (where existing groups remained separate from the new ones) and of acculturation (groups living in symbiosis with each other).

Migrations are the basis for widespread human presence in the world today. The specificity of human migrations is related to their adaptive mode, and is entirely different from animal dispersions. The flexibility of cultural systems permits, in principle, their adaptation to any natural environment encountered. Thus, this form of expansion can be broader and more intense than modifications of a biological order. Demographic developments occurring in human populations due to technological adaptations are often evoked to justify such human migrations. However, causes of a metaphysical order seem to act more powerfully to drive populations toward dispersal.

Regardless of the cause(s), migrations are a constant, observable in human behavior up to the present. In reality, a sedentary way of life seems to constitute an exception rather than the rule, considering the immense history during which humanity was exclusively nomadic. Once technoeconomical capacities were adequate, they were immediately put to use for displacements, which required new forms of adaptation of the social group, including the readjustment of values.

In terms of methodology, we often distinguish seasonal migration from permanent displacement. In practice, the distinction is reduced to a notion of scale, applied to time, rates, and proportions of migrations. Human territorial extension was a result of perpetual movement, and only large-scale phenomena can be considered permanent where diversity is constant and processes unending.

From the beginning, a distinctive feature of humans was their ability to leave the natural environment favorable to the way of life followed by other primates (Klein, 1989). Via bipedalism and the manipulation of objects, humans were able to leave the protective and nourishing environment of the tropical forests. The bipedal position, adapted to open environments, provoked these displacements, while a meat-based diet compensated for the loss of calories from vegetal foods, which were greatly reduced in the new savannah environments. Migration and the cultural adaptation that bipedalism and such a diet necessitated were therefore the driving forces behind the origins of humanity. The freedom acquired in relation to environmental determinism became the inevitable destiny of humanity, under the ultimate threat of its complete disappearance (Böhme, 1996).

The natural environments successively encountered in the process of hominid expansion toward northern latitudes required the development and adaptation of more complex and suitable physical forms (Wolpoff, 1998; Klein, 1989). These migratory movements, as early as 2 million years ago, extended north and east of the Old World, each time reaching new areas and demanding new dietary resources. The two keys explaining the range and the success of these migrations are the production of tools and the use of fire. Both were invented through imagination and audacity, forces that alone hold the explanation for such pioneering enterprises (**Figure 1**). Humanity eventually progressed throughout history according to this continuous thread of making evermore extended territorial conquests – toward the islands, the Americas, the moon. No force other than imagination pushed humanity, yet – retrospectively – we can retrace the high points, the pauses, the adaptations, and changes in rhythm. Within the confines of already-occupied areas, complex phenomena resulting from contacts and exchanges proper to humanity were produced. These phenomena can be classified in categories such as acculturation, colonization, and diaspora.

To appreciate the movements of migration toward unoccupied areas, the examination of conquests on the margins can

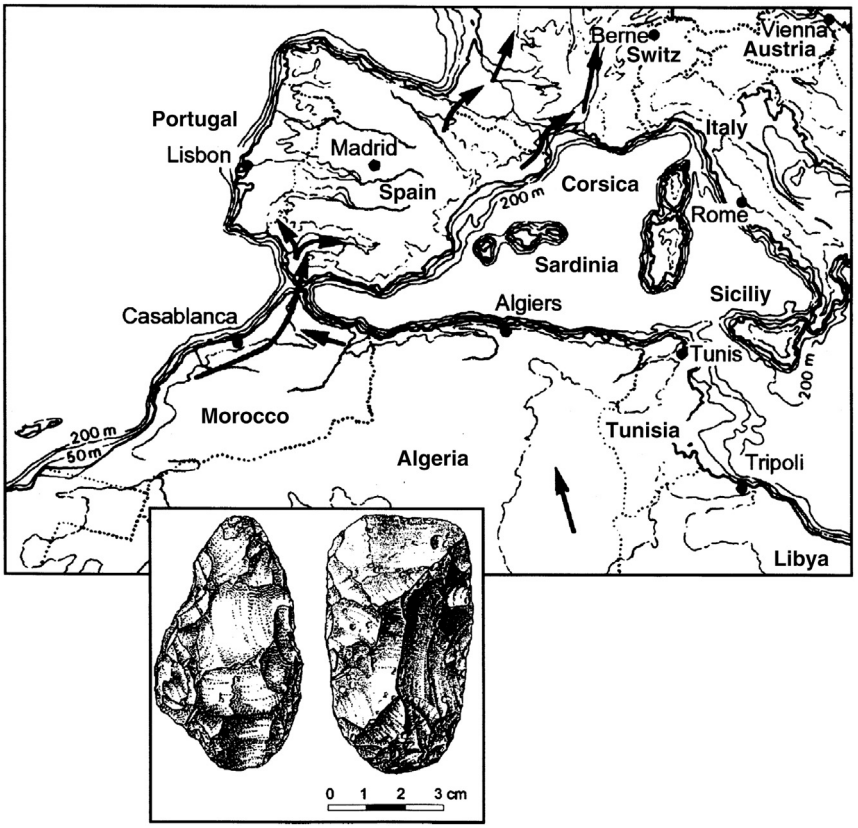

Figure 1 The first Acheulean migration in Europe passed from Africa to Spain via Gibraltar, after an earlier population from Asia arrived in Europe with a flake-based industry. After Roe, D.A., 1981. The Lower and Middle Palaeolithic in Britain. Routledge, London.

be instructive. For diffusion to the high seas of Oceania, migration is clearly linked to the mastery of navigation. An analogous situation is found on the northern margins of Europe during the Late-glacial: Hamburgian and Magdalenian reindeer hunters overcame climatic constraints as temperatures rose (Figures 2–4).

In order that massive displacements toward unoccupied regions may take place, technological equipment, in the subtlety of their elaboration, permitted adaptation to new environments and not simply to climatic modifications. This 'call to conquest' and the temptation of mobility seem to act as the only agents leading to displacement. Within occupied areas, more limited migratory cycles can be observed, for example, seasonal rounds (following the displacement of herds), or for ceremonial reasons (trade/exchange, marriage). These are such that potential mobility remains constant and is effectively begun from the moment that a threshold of ability is attained in relation to new constraints in a foreign environment. Technology, hunting, and group solidarity therefore had to become specialized before a migration could take place outside the original area of development (Otte, 1995).

The conquest continues east to Easter Island and north to Scandinavia. The case of the Americas remains perhaps the most revealing; migrations first followed the Siberian coast, then the islands of Beringia, and finally the northwest Pacific Coast. Migratory waves then switched from the high latitudes to the southern opening formed by the Californian coast. The invasions, extremely intense in North America, passed through a bottleneck between glaciers where modes of adaptation could reverse abruptly. In other terms, it seems quite probable that the first migration took place via coasts, directly toward the south. This would explain the early site dates in South America (e.g., Brazil). Moreover, the northern plains were colonized later, either after the separation of the glaciers to the north or even by a return, coming from the south after having crossed the continent only in Central America. As means of communications were developed, displacements were organized, first by coastal or marine environments, later by land. The modes of occupation of the prehistoric Americas still reflect migratory waves following the first conquest (Dixon, 1999).

The contacts with the cultural milieus already in place during migrations appear more complicated than simple adaptation to new environments. Some classic cases are known. One example is the arrival of modern humans in Europe, evidence of which is given both by a new anatomical form and by a completely novel system of symbolic and technological behaviors. We can be certain at

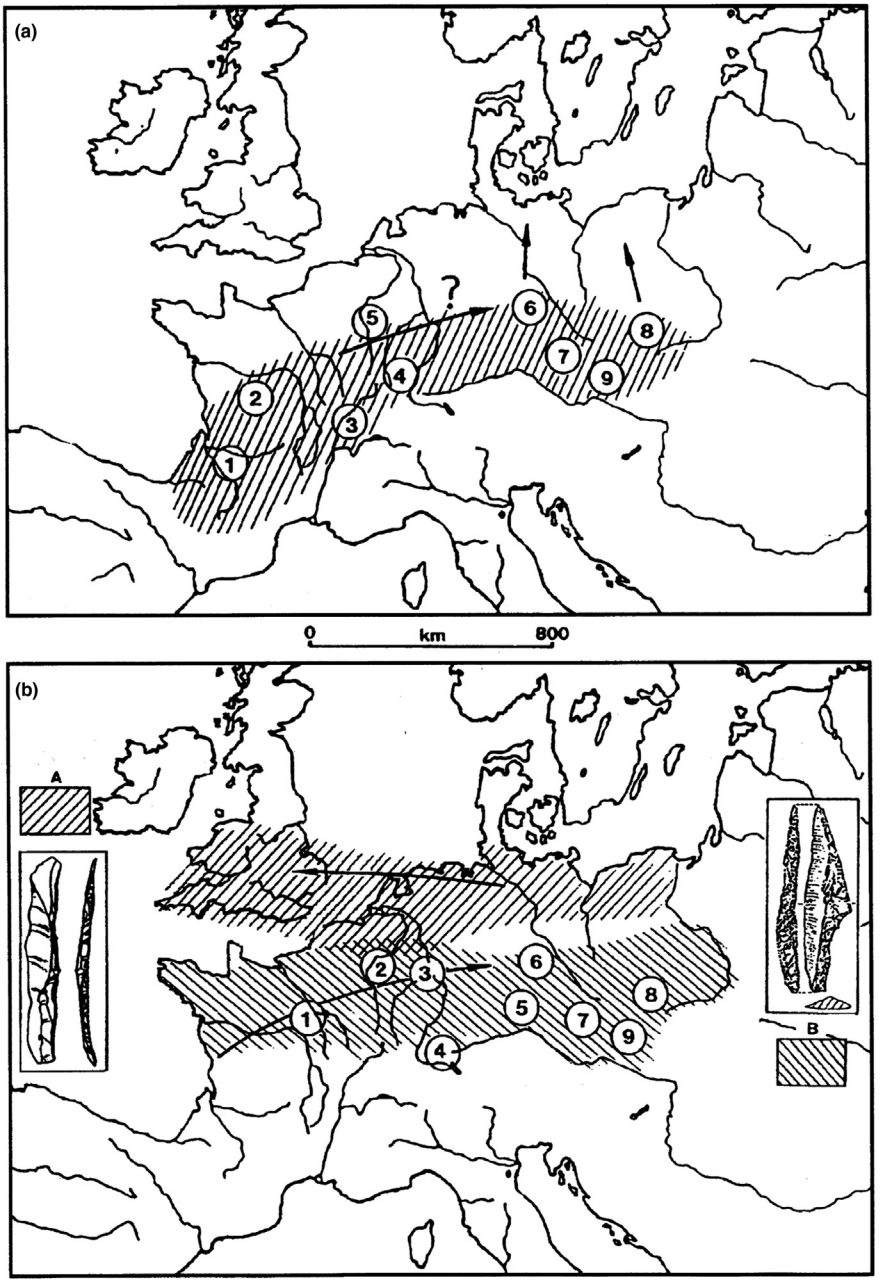

Figure 2 (a) Magdalenian movement in hilly regions during the Dryas I. (b) Dual movement during the Late-glacial, toward the northern plains and west across the then-dry North Sea. After Otte, M., 1996. Aires culturelles au Paléolithique supérieur d'Europe. In: Mohen, J.P. (Ed.), La vie préhistorique. Faton, Dijon, France, pp. 286–289.

least that a migratory wave from external populations can be revealed at the level of their customs, and attested by archaeology. A point of contact is then established, as it was with the Celts or the Germans, whose migrations were – elsewhere – attested by texts. A double control confirms the validity of archaeological data in the study of migration: anthropological on one hand, historical on the other. The Aurignacian trail crossed a Europe already regularly occupied by Neanderthals of the Mousterian culture.

The two milieus are clearly opposed, according to archaeological data, but on the margins, acculturation phenomena were produced locally, evidenced by 'mixed' assemblages: the Chatelperronian industry, foliate point cultures (Figure 5). In this way, tertiary phenomena (acculturation) emerge, from which later cultures develop. The case

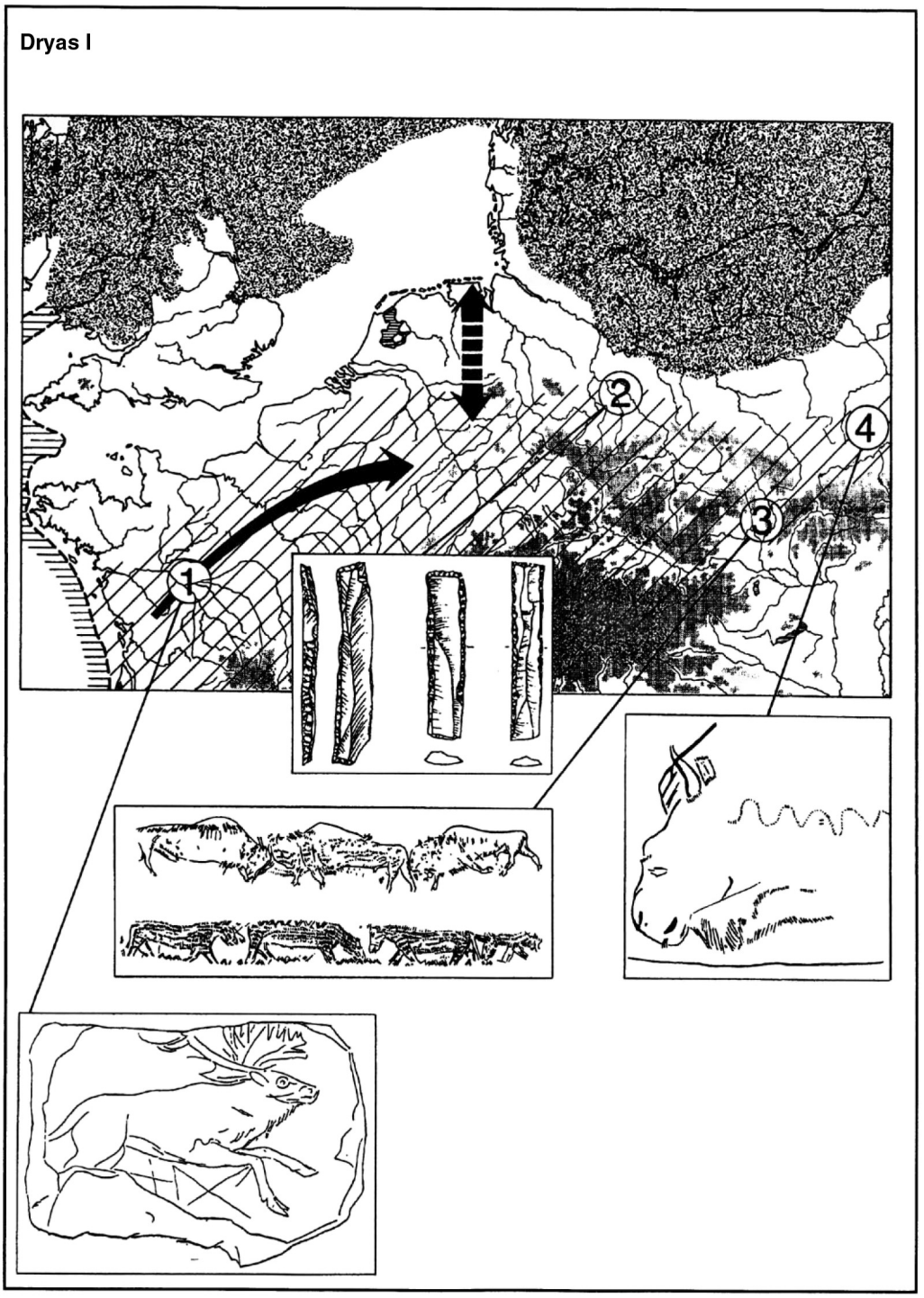

Figure 3 Two forms of migrations are observed in the northern plains during the Late-glacial: conquest in the eastern hilly regions and seasonal movements from north to south, following fluvial axes. After Otte, M., 1997b. Paléolithique final du nord-ouest, migrations et saisons. In: Fagnart, J.P., Thévenin, A. (Eds.), Le Tardiglaciaire du Nord-Ouest de l'Europe. CTHS, Paris, pp. 353–366.

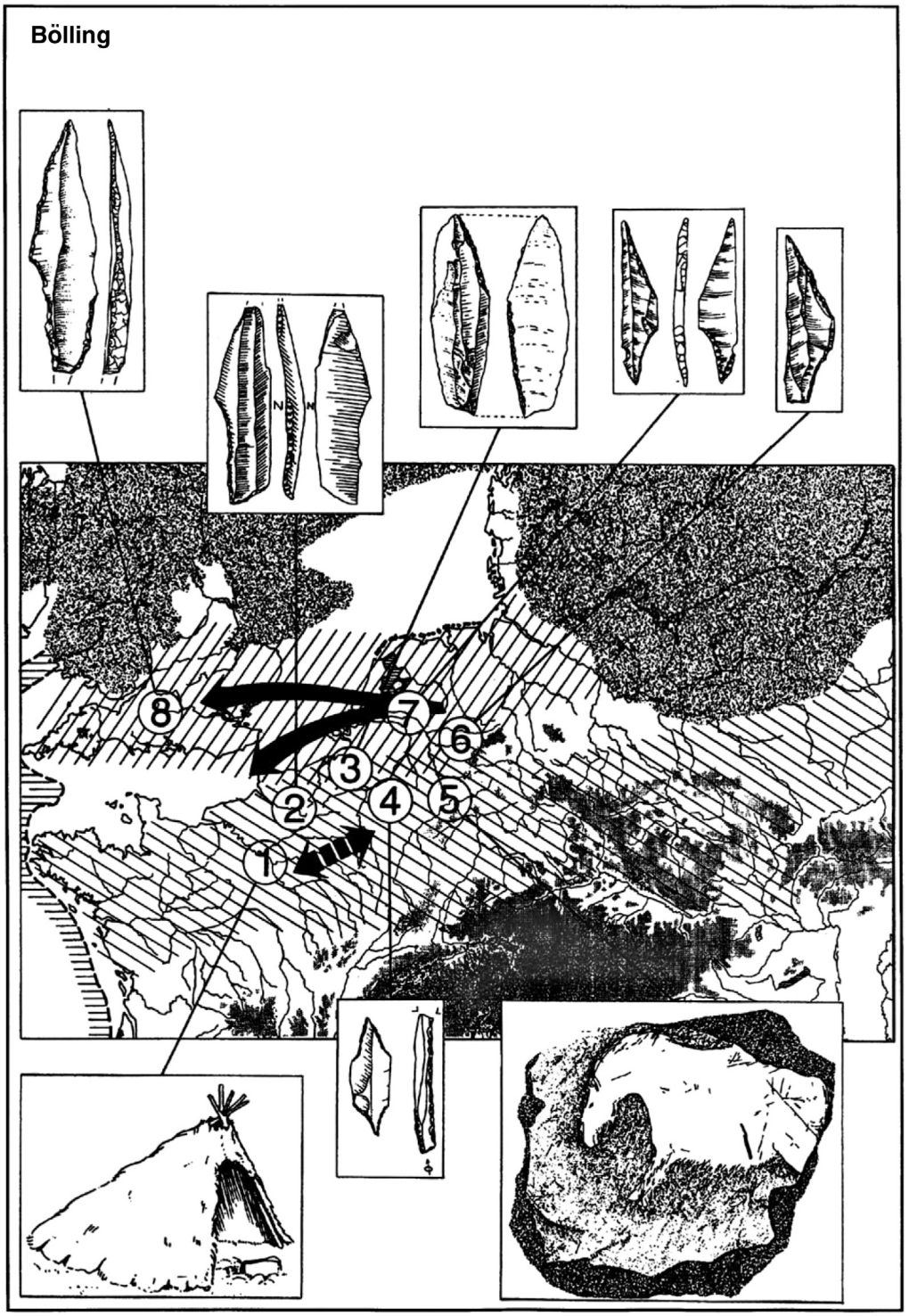

Figure 4 While Magdalenian migrations followed the axis of the plateaus, the northern plains were crossed by Hambourgian and Creswellian groups. After Otte, M., 1997b. Paléolithique final du nord-ouest, migrations et saisons. In: Fagnart, J.P., Thévenin, A. (Eds.), Le Tardiglaciaire du Nord-Ouest de l'Europe. CTHS, Paris, pp. 353–366.

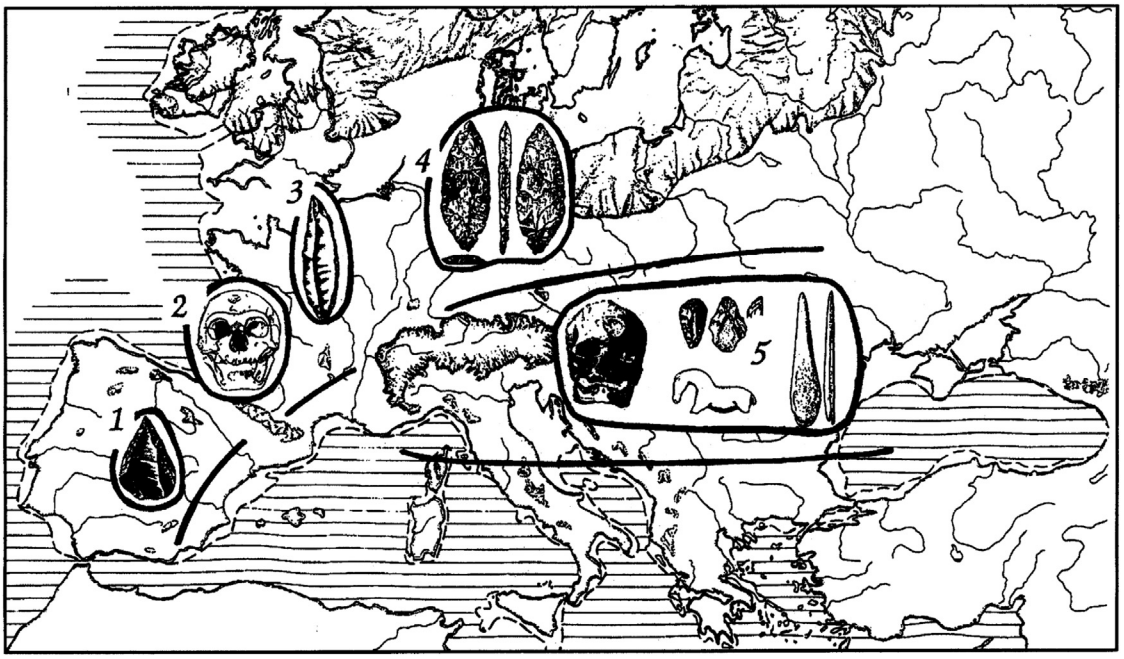

Figure 5 The first migration by modern humans in Europe met the established local Neanderthals. The Aurignacian remained homogeneous with respect to people and traditions. At the margins of this movement, contact phenomena arose: the acculturation of the Neanderthals by the Aurignacian culture. After Otte, M., 1995. Traditions bifaces. Les industries à pointes foliacées d'Europe centrale. Paléo (Suppl. 1), 195–200.

of the Aurignacian, at first clearly imposing itself, then completely disappearing in favor of a local Gravettian, forms a good example of migrations melting into a new and finally dominant population. These types of migrations to already occupied regions occur throughout the Palaeolithic with more or less clarity (Otte, 1997b).

Another instructive case is illustrated by the southern Solutrean industry, which includes one element that seems to originate in North Africa and another from northern Gravettian migrations (Otte, 1997a). When these two movements from north and south met, a new region of acculturation made its appearance, provoking the creation of a culture (the Solutrean) lasting several thousand years in Southwest Europe. Only general artistic trends reflect continuity with the preceding period, tracing their development under a harmonious form. Other elements of a stylistic character are profoundly modified at the end of the period of contact between these two opposing migratory waves.

A third migratory model also appears within the archaeological domain: the establishment of colonies, far from their points of origin but maintaining contact via exchanges with the intervening regions. This mode of colonization is well known from recent historical periods: Greek and Roman colonies, and later European colonial expansion. Colonial migrations first appear during the Linearbandkeramik (LBK) period of the Neolithic, 6000 years ago, when early Neolithic groups expanded across the European continent from the east and southeast (Otte, 2001).

The settlers traveled in a distinctive manner, under a permanent form but without modifying their own way of life, in this way maintaining contact with their original culture. The LBK people had architecture and an urbanism in balance with a stable economy (agriculture and animal husbandry), and an effective and appropriate technology (production of ceramics, grindstones, and polished adzes). A coherent way of life was harmonized with a new, totally mastered, environment. Contacts with indigenous populations modified the internal equilibrium acquired by the new arrivals. Archaeological data clearly illustrate this migratory model, with a great constancy presented by habitats and ceramic styles (Figures 6 and 7). Contacts are revealed by the long-distance exchange of raw materials from previously occupied regions. The impression of a strong cultural unit thus dominates the entire area of expansion of the archaeological culture.

Finally, the diaspora shows people whose migrations have integrated them into other milieus without affecting the structural articulation created by their own traditions. Modern cases are eloquent and well known, such as the Armenians, the Roma (gypsies), and the Jewish groups. Archaeological cases of widely dispersed cultures superimposed onto local traditions yet remaining stable, are also known. For example, the Bell Beaker people during the Neolithic period, identified by the presence of distinctive, globular goblets, were spread across Western Europe, particularly along seacoasts and rivers. Their simple burials, individual and with highly standardized grave goods, suggest the existence of a network of traveling merchants, perhaps linked to the exchange of the first metals

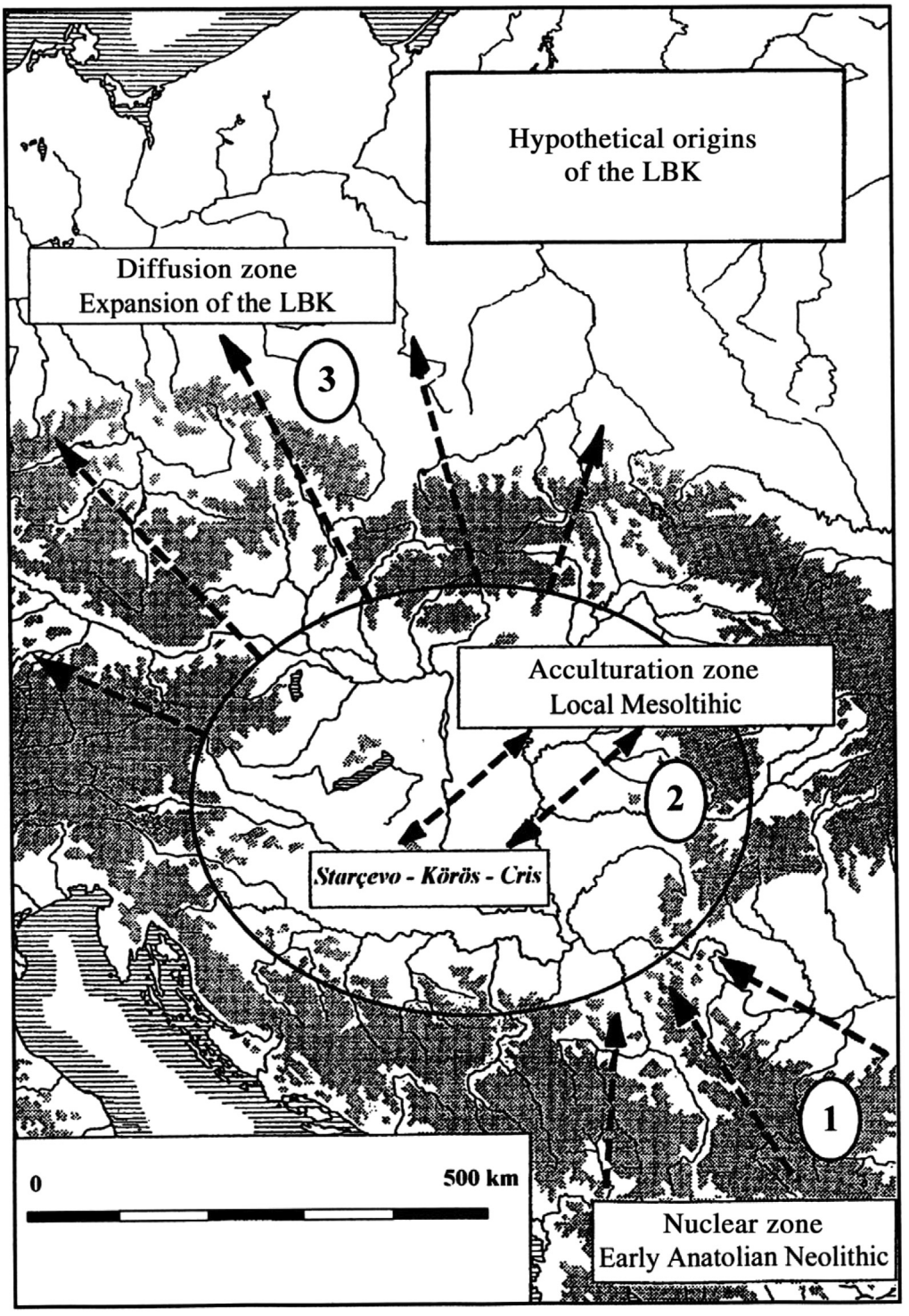

Figure 6 The LBK appears as a movement of colonization from a cultural center in modern Hungary. The settlers kept contact with nuclear areas and remained stable. After Otte, M., Noiret, P., 2001. Le Mésolithique du Bassin Pannonien et la formation du Rubané, L'Anthropologie 105 (3), 409–419.

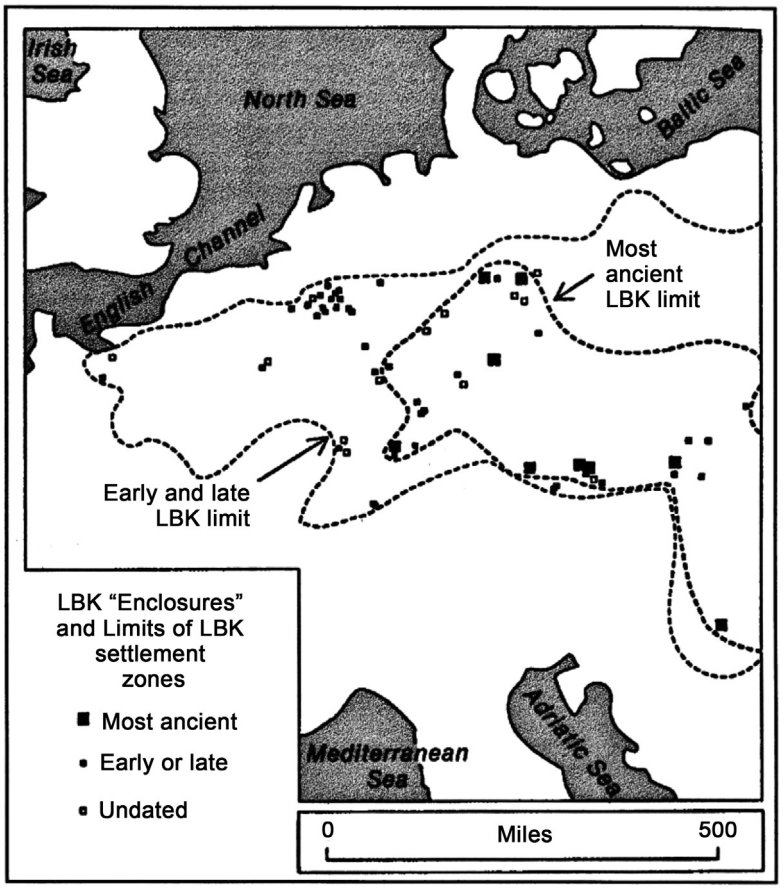

Figure 7 The LBK expansion represents the clearest model of prehistoric colonization: artifacts remained unchanged over a vast territory and are radically different from those found in preceding periods in the regions crossed. After Keeley, L.H., 1996. War before Civilization. Oxford University Press, Oxford.

(Figure 8). Cultural links remain obvious between these different populations, because they participated in an extended exchange network over a vast territory. However, not a single center seems to exist as a unifying and referential model for the culture.

Migrations in archaeological contexts are thus frequent and refer to all humanity. We are able to recognize clear traces throughout virgin territories such as the Americas or northern Europe; here it is the artifacts themselves that directly retrace the paths of human expansion. Stylistic criteria permit us to recognize different cultures and to follow their movement across space and through time.

Mitigating cases where such migrations are superimposed on indigenous populations demand different and more elaborate methods appropriate to the human sciences. The styles of equipment must be compared between the two groups in order not only to distinguish them but also to identify possible symbioses or influences. Maps of dispersion aid, stage by stage, in understanding archaeological evidence across space (Figure 9). To distinguish the influences of possible convergences, and similarities produced in different environments but at similar developmental stages, critical analysis must be applied to the archaeological record. The classic case, for example, is the Late Mousterian in Europe, where the same 'inventions' are known across all cultural environments in different regions but in similar circumstances. Migration can only appear in the global distribution and cohesion of a phenomenon, paired with territorial expansion. The more stylistic criteria are elaborated, the less random they are, and the more therefore they correspond to the system of values conveyed by a homogeneous population. It is this general key that archaeologists follow in their reconstructions of displacement. Although this law is not infallible and demands constant vigilance, comparisons with recent culturally well-defined peoples, permit us to ensure its effectiveness (Figure 10).

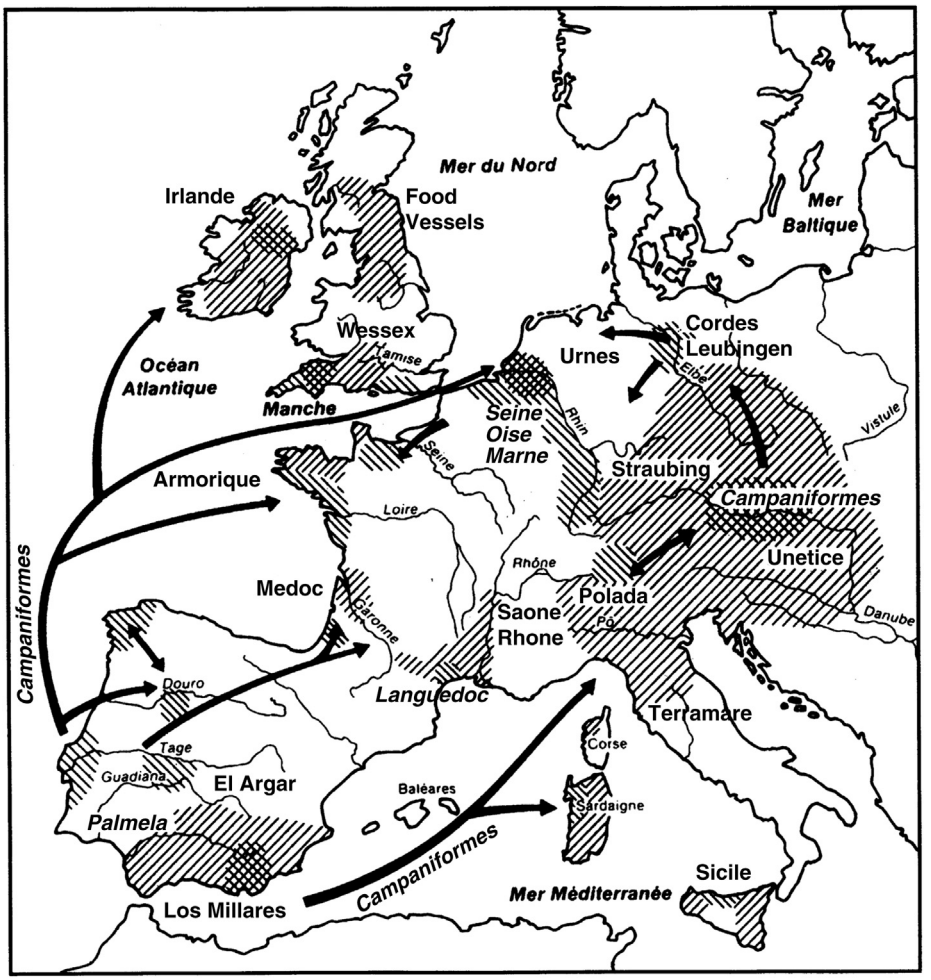

Figure 8 Small groups of Bell Beaker people had an enormous territorial expansion. They are characterized by graves, a particular ceramic form and copper objects. Their distribution, superposed on to local traditions, evokes the diaspora observed in historical times. After Briard, in Mohen, J.P. (Ed.), 1996. La vie préhistorique. Faton, Dijon, France.

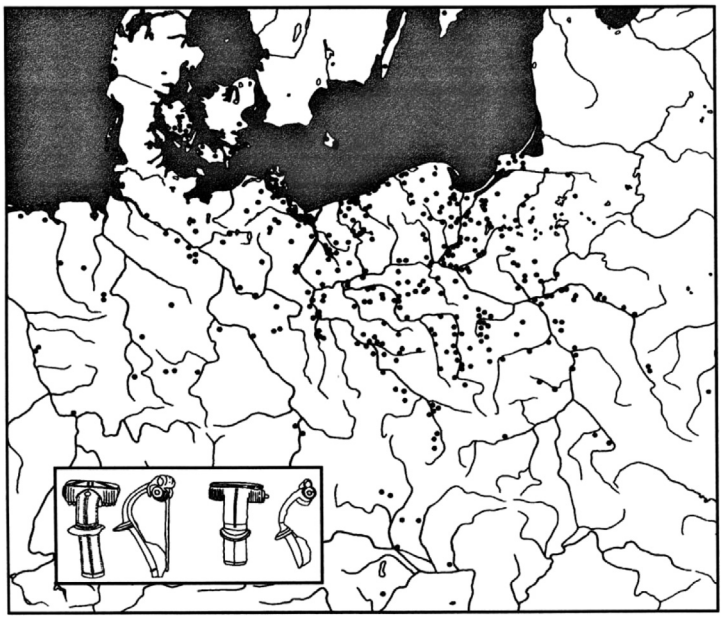

Figure 9 Methodological example of distribution of a particular type of decorative elements along the Baltic coast. These Germanic elements identify an ethnic group and permit us to follow their migrations until the time of Christ. After Böhme, H.W., 1996. Kontinuiät und Traditionen bei Wanderungsbewegungen im frühmittelalterlichen Europa vom 1–6 Jahrhundert. Archäologische Informationen 19, 89–103.

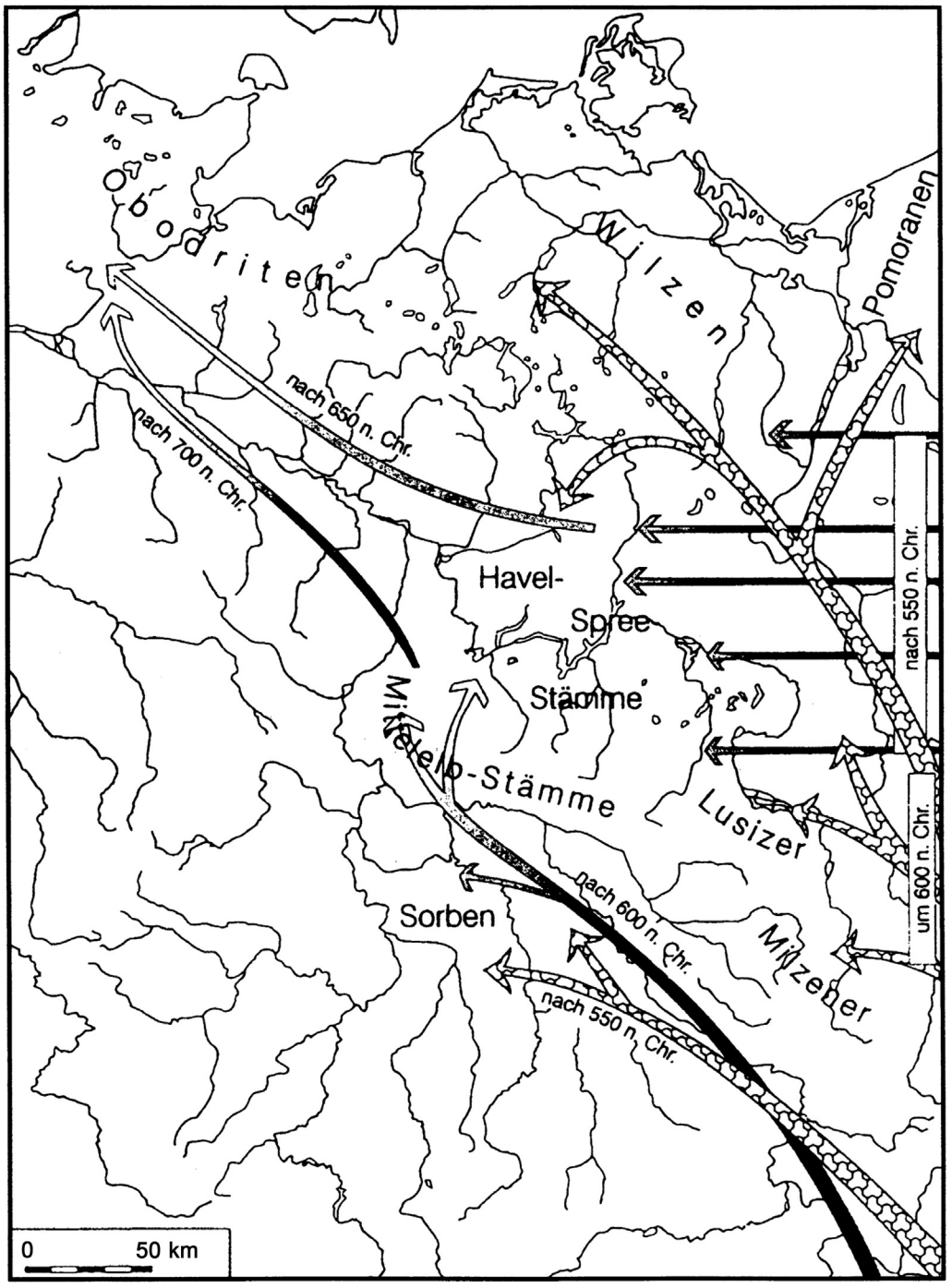

Figure 10 Historical migrations (here, the Slavs) are reconstructed by the distribution of languages, written sources, and archaeological information. After Kóčka-Krenz, H., 1996. Die Westwanderung der Slawen. Archäologische Informationen 19, 125–134.

See also: Adoption, Demography of; Demographic Techniques: Data Adjustment and Correction; Demographic Techniques: Indirect Estimation; Ethnic Identity and Ethnicity in Archaeology; Ethnicity and Migration in Europe; Migration: Anthropological Perspectives; Population Dynamics: Momentum of Population Growth.

Bibliography

Böhme, H.W., 1996. Kontinuiät und Traditionen bei Wanderungsbewegungen im frühmittelalterlichen Europa vom 1–6 Jahrhundert. Archäologische Informationen 19, 89–103.

Dixon, J., 1999. Late Pleistocene maritime adaptations and colonisation in the Americas. Pre-prints of the World Archaeological Congress 4, 10–14.

Keeley, L.H., 1996. War before Civilization. Oxford University Press, Oxford.

Klein, R., 1989. The Human Career. The University of Chicago Press, Chicago.

Kóčka-Krenz, H., 1996. Die Westwanderung der Slawen. Archäologische Informationen 19, 125–134.

Mohen, J.P. (Ed.), 1996. La vie préhistorique. Faton, Dijon, France.

Otte, M., 1995. Traditions bifaces. Les industries à pointes foliacées d'Europe centrale. Paléo (Suppl. 1), 195–200.

Otte, M., 1996. Aires culturelles au Paléolithique supérieur d'Europe. In: Mohen, J.P. (Ed.), La vie préhistorique. Faton, Dijon, France, pp. 286–289.

Otte, M., 1997a. Contacts trans-méditerranéens au Paléolithique. In: Fullola, J.M., Soler, N. (Eds.), El món mediterrani després del Pleniglacial (18.000–12.000BP). Museu d'Arqueologia de Catalunya, Girona, pp. 29–39.

Otte, M., 1997b. Paléolithique final du nord-ouest, migrations et saisons. In: Fagnart, J.P., Thévenin, A. (Eds.), Le Tardiglaciaire du Nord-Ouest de l'Europe. CTHS, Paris, pp. 353–366.

Otte, M., 2001. Le Mésolithique du Bassin pannonien et la formation du Rubané. In: Kertész, R., Makkay, J. (Eds.), From the Mesolithic to the Neolithic, Proceedings of the International of Szolnok (22–27 September 1996). Budapest, Archaeolingua Main Series, vol. 11, pp. 375–378.

Otte, M., Noiret, P., 2001. Le Mésolithique du Bassin Pannonien et la formation du Rubané. L'Anthropologie 105 (3), 409–419.

Roe, D.A., 1981. The Lower and Middle Palaeolithic in Britain. Routledge, London.

Wolpoff, M., 1998. Paleoanthopology. McGraw-Hill, Maidenhead, UK.

Militaries and Militarization, Anthropology of

Andrew Bickford, George Mason University, Fairfax, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Anthropologists have become increasingly interested in militaries and militarization, as they recognize the role played by militaries, militarization, and warfare in culture and everyday life. By examining militarization and the military ethnographically, anthropologists analyze the methods and strategies used by state and military actors to influence, shape, and control perceptions of military service, military action, and military violence; how people respond, react, accept, and contest the military and militarization; and the ways in which military values and worldviews shape and influence culture, agency, and worldviews.

The end of the Cold War brought renewed hope for a reduction in militarization, a general demilitarization of global politics, and a long-awaited 'peace dividend.' Instead, the end of the Cold War resulted not in a diminishment of militarization in the United States and around the world, but an expansion of militarization and the influence of militaries in the form of prolonged wars in Iraq and Afghanistan, an expansion of US military power around the world via bases (see Gusterson, 2009; Lutz, 2009b; McCaffrey, 2002; Vine, 2009), the increased use of United States Special Operations Forces and drones in the 'War on Terror,'(Gusterson, 2011), the increased presence of the military and military values in everyday life (Ben-Ari, 2008, 2004; Feldman, 1994; Gusterson, 2007; Lutz, 2001, 2002, 2007), the skyrocketing of the US defense budget (see, for example Gonzalez, 2010; Gusterson, 2007, 2009; Lutz, 2001, 2002, 2007; The Network of Concerned Anthropologists, 2009), and discourses of 'heroism' and military superiority in the United States (Bickford, 2010) and elsewhere. Within this context of the increasingly visible expansion of the military in everyday life, the anthropology of militarization and the military has assumed a new importance, and a growing number of anthropologists now focus attention and research on the military and militarization, both in the United States and elsewhere. Once considered something of a 'niche' area of research, the study of militaries and militarization is an increasingly important specialization in contemporary anthropology.

As Max Weber contended, states claim a monopoly on 'legitimate' violence, and as sociologist Charles Tillly (1999) argued, war makes states. While war makes states, states need to make soldiers, lay the groundwork for support of the military and military action, and convince civilians to accept and to make war, thereby shaping and forming a key part of the political, cultural, and moral/ethical life of the state (Ben-Ari, 2004; Bickford, 2011; Gonzalez, 2010, 2009; Gusterson, 2007; Lutz, 2009a; Simons, 1999). Anthropologists who study militarization and the military concentrate upon the broad contours of these intersections. This is an important development in anthropology, as much of who, what, and where anthropologists have traditionally studied has been shaped or influenced by militarization and war in some way (see, for example, Green's work on war, violence, and trauma in Guatemala (1999) and the influence of decades of war and militarization on the Mayan widows she studied), and anthropologists have often been blind to or failed to see these connections between culture, militarization, and war (see Starn's (1991) critique of ethnographic research in Peru and the failure to recognize the Shining Path movement, and Hutchinson's work with the Nuer (1996) and her analysis of Evans-Pritchard's (1969) work with the Nuer in the 1930s).

While a large number of anthropologists have studied and theorized the origins of aggression and 'primitive' warfare, only fairly recently have an increasing number of anthropologists turned their gaze toward the military in the modern nation-state, and the impact of militarization policies on the cultures and life worlds of the people whom they study. Anthropologists have long been worried about the influence of the military and intelligence organizations on anthropological research (Boas, 1919; Price, 2008), but this concern did not necessarily translate into a focus on the military and militarization as objects of study for a variety of reasons. These reasons included access, a focus away from the nation-state, and political disdain: for a large number of anthropologists, the military as an object and site of study lay far outside their interests, both professionally and politically.

Once the seemingly exclusive purview of historians, political scientists, international relations specialists, and military sociologists, anthropologists have become increasingly interested in militaries and militarization, as they recognize the role played by militaries, militarization, and warfare in everyday life. Anthropologists who study the military and militarization often draw upon the work of militarization scholars from these other fields (political scientist Cynthia Enloe's work has been a major influence in anthropological studies of gender, militarization and militaries, for example), and in the process provide an ethnographic perspective and analysis that adds to and complements other analyses of militarization and the study of militaries. An approach that examines the socio-cultural-historical aspects of militarization allows for an examination of the ways in which preparations for war and military service influence and shape the identities, world views, and life courses of individuals, including how they cope with life in the civilian world after military service (Bickford, 2011: 24–25; see also Açiksöz, 2012; Finley, 2011; Gutmann and Lutz, 2010; Messinger, 2010; Wool and Messinger, 2012).

Anthropologists tend to focus on the 'lived experience' of militarization and the military in addition to its forms, structures, institutions, and content. The analysis of militarization policies and practices also provides a window into the ways state and military officials and actors work both in tandem and in contradiction to create the political structure of the 'State,' and to examine the practices and techniques of military power that state agents utilize in the consolidation of power and legitimacy. Anthropologists study the military in order to understand the impact of the military as an institution on those who are forced or volunteer to join and their families, the impact of the military on communities, and the physical, symbolic, and structural effects of military violence. By examining militarization and the military ethnographically, one sees the methods and strategies used by state and military actors to influence, shape, and control perceptions of military service, military action, and military violence; how people respond, react, accept, and contest the military and militarization; and the ways in which military values and worldviews linger on longer after encounters with militaries and militarization.

Defining Militarization

Broadly defined, militarization is the cultural, symbolic, and material preparation for war, and as recent research in anthropology has shown, militarization and the presence of state militaries influences much of the 'everyday life' in many societies and cultures around the world, both explicitly and subtly. Most importantly, militarization is an intentional process, something a state or group must set out to do; as Margaret Mead wrote, war – and in this case, the preparation for war – is not a 'biological necessity' (Mead, 1940). Rather, militarization and the development of military institutions are explicit projects, but are often cloaked in discourses of biological inevitability. Militarization programs often trade in and promote ideas about innate violence and the naturalness of military values (see Ferguson, 2009: 33–34; Fry, 2006, 2007); while sociobiological notions of 'natural' soldiers are politically convenient, these arguments sidestep analyses of power and process, and can be used to naturalize and depoliticize militarization processes and war (Bickford, 2011: 6–7; Ferguson, 2009; Fry, 2006, 2007) as a way to make them seem inevitable and irresistible.

Militarization draws from, and shapes, cultural contexts and histories and individual life courses. Broadly speaking, militarization is about the transformation of civilians into soldiers, and through the creation of certain types of soldiers, the creation of certain types of men and women; the creation of new types, conceptions, and understandings of 'citizens,' and the acceptance of these new types of categories and ways of being. It is a process of social, political, and military reproduction; the reproduction of the state through military values and identities; and the naturalization of the creative and (re)productive violence of the state in and through the very bodies of its citizens.

Anthropologists view militarization as a social process designed to bring about support for the military and shape social structures to support the military. Catherine Lutz's definition of militarization as "the shaping of national histories in ways that glorify and legitimate military action" (Lutz, 2007) is particularly useful in thinking through militarization programs, and the strategies necessary to implement a successful demilitarization program. Anthropologist Roberto Gonzalez, who, in his book *Militarizing Culture: Essays on the Warfare State*, examines the increasing militarization of the United States, draws from and expands historian Richard H. Kohn's definition of militarization, and the ways in which this relates to the ethnographic analysis of militarization:

> The degree to which a society's institutions, policies, behaviors, thought, and values are devoted to military power and shaped by war.
>
> Kohn, as cited in Gonzalez, 2010: 19

Lutz also describes militarization in terms of what she calls the 'military normal':

> The condition in which science, entertainment, business, and even high fashion deeply reflect militaristic values.
>
> Lutz, 2009a: 23–37

And Bickford (2011), in his analysis of East German militarization and the creation of 'soldiers,' examines how states attempt to create a 'military imaginary' in order to create a seemingly 'natural' bond between citizens and the military:

> The ways in which the necessity, implementation, and desired outcomes of (compulsory) military service and training are imagined and envisioned by the state, and the ways in which these tropes are linked to normative ideas of the "proper" soldier and [citizen], legitimate violence, morality, and military tradition.
>
> Bickford, 2011: 4

Militarization is about the transformation of civilians into soldiers, and through the creation of certain types of soldiers, the creation of certain types of men and citizens. It is also concerned with creating citizens who will support the creation and training of soldiers (Bickford, 2011; Gill, 1997). It is about making 'soldiering' and 'the military' appealing, and is concerned with garnering support for the military, military solutions, and military violence. Militarization is an exercise in attempting to establish hegemony through the promotion of military values, fear, and defense and of making these values seem natural, normal, and desirable (of course, in practice, this is not a smooth or entirely successful process, as evidenced by resistance to militarization, conscription, and war) (Bickford, 2011; others). A goal of militarization is to make people accept and love war, and see it as 'normal' (Lutz, 2009a). As anthropologist R. Brian Ferguson writes, "people have the capacity to learn, even to enjoy, war and build it into their social lives and institutions" (Ferguson, 2009: 34; Lutz, 2009a, 2007; see also Gonzalez, 2010: 27). Ferguson also notes that "Commentators have often compared war to a disease, but a more apt analogy is an addiction" (Ferguson, 2009: 40; see

also Gonzalez, 2010: 27). Militarization programs – and those who design and implement these programs – target and exploit this capacity to learn to love war. Indeed, people are supposed to learn to love war so much that they will willingly sacrifice their children for the state (Bickford, 2011; see also Perez's work (2006) on the militarization of Latina/o youth and the intersections of race and class in the United States). Ultimately, as anthropologists who work on militarization have shown, militarization is an idea – a deep-rooted insidious idea – that influences a large part of what constitutes the 'everyday' in modern states, and that makes the world make sense to many who internalize it as their own 'common sense' (see Lutz, 2009a, for her discussion of the 'military normal'; see also Herzfeld, 2001).

As anthropologists who study militarization and the military have shown, militarization programs depend on the establishment of deep networks within the state, and the deployment of discourses around courage, strength, honor, masculinity, femininity, credentials, health, marriage, violence, and identity. Militarization is not only about producing weapons and war materials but also about producing citizens, gender identities, family forms, language, ethics, morals, ways of being, and ways of seeing the world (Bickford, 2011; Cohn, 1987; Frühstuck, 2007; Gill, 1997, 2004; Gonzalez, 2010; Kanaaneh, 2005; Lutz, 2009a, 2007). Ultimately, militarization is a process that produces 'in-' and 'out-'groups, friends and enemies, the sacred and the profane, those who should be protected and those who should be killed, 'us' and 'them,' as well as 'we' and 'the other' (Bickford, 2011; Ferguson, 2009; Gonzalez, 2010; Gusterson, 2007; Mamdani, 2002; Waterston, 2009). Anthropologists who study militarization and the military engage many of these areas in their research, examining their overlapping influences and connections.

These definitions and perspectives make clear that military and militarized values will not simply disappear overnight, and that militarization impacts and influences a vast array of ideas, institutions, and processes that might seemingly appear far removed from traditional ideas about the military and war. Insights gleaned from anthropological work on militarization and militaries provide a necessary addition to military studies and a valuable complication to the analysis of demilitarization programs, showing that demilitarization needs to attend to the cultural side of war preparation, and not just to weapons and weapons systems. Militarization programs are also forms and visions of the state and the moral universe of the state in that they represent the ways and manners the state is imagined and desired, and that they shape 'permissible' and 'moral' forms of violence (Bickford, 2011: 24). In this sense, militarization programs are just as much about values and 'common sense' ideas about the world (see Herzfeld, 2001) as they are about producing tanks, planes, and land mines. States and military blocs create different rationales for different types of militaries, soldiers, and citizens. These competing visions contain different moral economies, which shape ideas about soldiering, killing, allegiance, human rights, conceptions of justice and law, and conceptions of political economy, and form the 'justification' for the use of military violence, rearmament, and the production of weapons.

Anthropology and Militaries

A key area of interest in the study of militarization is the study and analysis of militaries as institutions, and the influence of a 'military' on society, political economy, and everyday life. The majority of studies have focused on various aspects of the US military (see Bickford, 2010, 2008); Frese and Harrell, 2003; Gill, 2004; Gonzalez, 2010, 2009; MacLeish, 2013; Lutz, 2009a,b, 2007, 2002, 2001; Gutmann and Lutz, 2010; Finley, 2011; Simons (1997); Hautzinger and Scandlyn, 2013; Vine, 2009). However, anthropologists have also studied the armed forces of Japan (Frühstuck, 2007), Bolivia (Gill, 1997), Israel (Ben-Ari, 1998; Weiss, 2012; Kanaaneh, 2005), Canada (Irwin, 2011), and Germany (Bickford, 2011). These studies approach the study of militaries from a variety of angles, but with a number of overlapping themes, such as the production of 'the enemy'; the ways in which violence is legitimated, codified, and imagined; and the ways in which military service and gender intersect, just to name a few. In addition – and in conjunction with – the study of militarization, anthropologists also study the armed forces of a variety of states as a way to examine and often critique state power, violence, and military policies; the motivation for joining a military; the role of a military force in a state; and the effect of a military on the culture and daily life of a state. Irwin (2011) has studied the Canadian military and peacekeepers. Bickford (2011) examined both the East and West German militaries, and the German military after unification, focusing on militarization policies and the role of the military in the German Democratic Republic (GDR), the militarization of the GDR, and the policies and politics of memory in the German military after unification. Frühstuck (2007) has studied the Japanese Self-Defense Force, focusing on questions of militarization and gender in Japan, and the ways in which peacekeeping operations intersect with militarized masculinity. Frühstuck's ethnography is a fine-grained examination of military service, training, and militarized masculinity. Gill (1997) studied the intersections of citizenship, masculinity, and militarization in Bolivia. Gill also conducted research into the US Army's School of the Americas (SOA), and examined the militaries of United States and Central and South America, the ways in which the US military trained officers in brutal counterinsurgency tactics, and the politics of postatrocity impunity (2004). Ben-Ari (1998) studied the emotions and the construction of the 'enemy' in the Israeli army; Kanaaneh (2005) has studied Palestinian soldiers in the Israeli military; and Weiss (2012) studied conscientious objectors in the Israeli military.

Veterans

As anthropologists have increasingly engaged militarization and the military, they have also turned their attention to the experiences of soldiers and veterans as a way to examine the aftereffects of militarization, and the ways in which military service impacts daily life and life courses after military service or warfare. Anthropologists are also interested in the ways in which veterans and their experiences of military service are

coopted and used by the military for recruiting and propaganda purposes. Gutmann and Lutz (2010) write about and analyze the political awakening and activism of US military veterans of the wars in Iraq and Afghanistan, and how they have used their military experiences as a way to form political responses to militarization in the United States. Bickford (2011) conducted fieldwork with former members of the East German army and Border Guards, examining their experiences of dislocation and political estrangement after German unification, and the ways in which East German militarization impacted their decisions to become career officers. Carbonella (2003) writes about the connections between representations of Vietnam veterans and globalization, and Wagner (2013) and Wagner and Quintyn (2009) write about the Tomb of the Unknown Soldier in the United States, and the politics of veterans, war dead, and commemoration in the United States.

Health and Technology

Anthropologists are also increasingly interested in the intersections of militarization, war, health, and medicine, and ways in which both civilians and soldiers experience war and the preparation for warfare in terms of bodily experience, illness, wounding, and as a biomedical project concerned with battlefield performance enhancement. This research overlaps with the study of veterans; indeed, many anthropologists who study veterans are also keenly interested in experiences and narratives of veterans' trauma and suffering. Messinger (2010) and Wool and Messinger (2012) study US military personnel who suffered amputations, and how they come to terms with new prosthetic limbs. Açiksöz (2012) studies nationalism, masculinity, and the political identity of Turkish Army soldiers who lost limbs fighting the Kurdish Workers Party (Partiya Karkerên Kurdistan (PKK)). Finley (2011) and Hautzinger and Scandlyn (2013) study posttraumatic stress disorder among US combat veterans; and Vasquez (2009) explores visual technologies and soldiers' experiences of warfare. Bickford (2008) studies advances in US military medicine and the militarization of medicine in the production of 'super soldiers' in the United States. Wagner (2008) has conducted extensive research into the use of deoxyribonucleic acid analysis to identify and recover genocide victims in Bosnia, and the ways in which veterans, survivors, and their families come to terms with war and genocide in Bosnia. Also, Russ and Ames (2006) examined competing policies and practices on health and sexuality in the US Navy.

Counterinsurgency, Militarization, and Anthropology

Directly related to the study of militarization and militaries is one of the most contentious debates in recent anthropology: anthropology's relationship to the military, and the US military's attempts to utilize anthropology as a kind of 'force multiplier' in counterinsurgency operations in Iraq, Afghanistan, and elsewhere. While anthropology has had a long history of some forms of collaboration with the military (see Price, 2011, 2008, 2004), and resistance to the military's use of anthropology (Boas, 1919; Network of Concerned Anthropologists, 2009; Price, 2011), the recent controversy surrounding the US military's 'Human Terrain System' and the embedding of anthropologists and other social scientists in military units (with anthropologists wearing uniforms and in some instances carrying weapons) brought this relationship once again to the forefront as an ethical issue within the American Anthropological Association. Some anthropologists have spoken out against the military use of anthropology, arguing that the US military is attempting to militarize and 'weaponize' anthropology (Forte, 2011; Gonzalez, 2009; The Network of Concerned Anthropologists, 2009; Price, 2011), while others argue that anthropology should work with and engage the military as a way to provide the military with cultural knowledge and training in order to reduce violence and military action (Albro, Marcus, McNamara, and Schoch-Spana, eds. 2012; Fosher, 2011; McFate, 2005; McFate and Fondacaro, 2011; Rubenstein et al., 2012; Selmeski, 2007). In 2007, the American Anthropological Association commissioned a report investigating the controversy, and concluded that anthropologists need to follow ethical fieldwork practices, regardless of which side of the debate they are on.

Recent works that examine the controversy around anthropology, counterinsurgency, and anthropology's relationship to the state and military include Schaft's (2007) examination of anthropology in Nazi Germany, Bahn's work on counterinsurgency in India (2013), Gonzalez's overview of American counterinsurgency programs and anthropology (2009), McNamara and Rubinstein's (eds.) edited volume examining both sides of the counterinsurgency and anthropology controversy (2011), and Kelly, Jauregui, Mitchell, and Walton's (eds.) volume examining counterinsurgency programs around the world (2010). Wakin (1992) writes about anthropology, counterinsurgency, and Thailand, and David Price (Price, 2011, 2008, 2004) has written extensively about the relationship(s) between anthropology, the military, and intelligence agencies (Price, 2011, 2008).

Anthropology of Bases

One recent direction in the anthropology of militaries and militarization is an analysis of the impact of US military bases around the world and in the United States, and the experiences of those who live on, work on, or who have been displaced by military bases. McCaffrey examined the US Navy's use of the Puerto Rican island of Vieques as a bombing range, and the subsequent protest movements to dislodge the US military from the island (2002). Vine, in *Island of Shame*, examined the impact of the US/UK base on Diego Garcia in the Indian Ocean, and the forced relocation of the island's original inhabitants (2009). Vine has also written about the increase of US bases around the world (2012). Lutz (2009b) conducted an in-depth analysis of the impact of Fort Bragg on the city of Fayetteville, North Carolina, and MacLeish (2013) conducted fieldwork with soldiers and civilians at Fort Hood, a major US Army base in Texas. Gill (2004) conducted extensive fieldwork at the US Army's 'SOA,' at Fort Benning, Georgia. Besteman (2009, 2008) and Vine (2012) have examined

the US military's 'Africa Command' (AFRICOM), and the proliferation of US military bases in Africa. Gusterson (1996) and Masco (2006) have both examined nuclear weapons research and research laboratories and their impact on communities in California and New Mexico.

Anthropology, Militarization, and Demilitarization

While the turn toward research on militarization and militaries is an important step in anthropology, the logical next step is to draw from and expand research on militarization and militaries, and to engage ethnographically with demilitarization, and the ways in which states and groups come to terms with the end of militarization programs and the lingering aftereffects of militarization. A discussion of recent work in anthropology on militarization points toward a discussion of demilitarization, and how people come to terms with the lingering effects and impacts of militarization, war, and violence. If militarization depends on the creation of dense networks of 'positive' feelings toward the military, for instance, how exactly does one come to terms with these networks after a militarization program ends, after a war, or after military violence? How do people 'unravel' or dismantle militarization programs? How do people come to terms with the effects of militaries, militarization, and war? Demilitarization implies a reversal of an implicit process or program, an unraveling, an unmaking, of that which came before and of ways of thinking, feeling, and seeing that made a military solution thinkable and desirable (Bickford, 2013). As anthropologists have shown, militarization can be a pervasive and deeply woven thread in a culture (see Green, 1999); to unravel it or do away with it requires a thorough understanding of how militarization works.

Demilitarization is often narrowly defined as simply the dismantling or demobilization of a military, and the eventual destruction of military equipment, the destruction of weapons and explosives, and the incineration and destruction of chemical and biological weapons. Of course, these are important steps in demilitarization, but they can also be seen as merely the 'surface' level of a demilitarization program. If militarization is about shaping and creating a populace that either embraces, or at least goes along with, positive ideas and values associated with war and killing, then simply destroying weapons after the fact is not enough to effectively demilitarize a society. Narrow definitions of demilitarization centered solely on weapons obscure analysis of the social relations and cultural constructions of militarization. As Bickford (2013, 2011), Açiksöz (2012), Gutmann and Lutz (2010), and Carbonella (2003) have shown in their studies of military veterans, enculturation into military norms and worldviews does not go away overnight; these experiences and 'training' linger on, shaping how people see and think for years afterward.

Anthropologists have examined aspects of demilitarization and the aftereffects of warfare and violence. As Hinton (2002) argues in his overview of the anthropology of genocide, there is a need to attend to local dimensions and experiences of war and genocide in order to understand how war and genocide took the form(s) they did (see, for example, Bahn, 2013; Besteman, 1999; Gill, 2009; Green, 1999; Hinton, 2002, 2004; Hoffman, 2005; Malkki, 1995; Manz, 2009; Nordstrom, 2009, 2004; Olujic, 1998; Robben, 1995; Schaft, 2007; Wagner, 2008). In other words, one needs to understand the cultural forms that obtain in a given situation, and how they shape and mediate expressions of militarization, war, and violence in order to develop demilitarization strategies that can attend to underlying cultural logics and issues; without this, there is a risk of merely attending to the 'surface' level of violence and failing to understand how deeply rooted and entrenched militarization and the creation of the 'military normal' (Lutz, 2009a) can be. Demilitarization can also entail a process of accountability, of trying those responsible for militarization, war, genocide, and atrocities accountable for their actions, such as the Nuremberg trials after World War II, the Truth and Reconciliation Commission(s) in South Africa and Rwanda, and the Enquete Commission in Germany after unification (see Bickford, 2011; Hinton, 2002; Scheper-Hughes and Bourgois, 2004; Wagner, 2008; Waterston, 2009).

Anthropological studies of militarization and militaries shed light on the programs, formations, and processes of violence in the state, and can inform analysis and discussions of demilitarization programs, as they make explicit the often unstated or hidden assumptions, desires, and forms of power that make them effective and deeply entrenched. Coming to terms with militarization, militaries, and war, and thinking through demilitarization programs compel us to think not only about disarmament and the destruction of weapons but also about the sociocultural aspects of militarization. Insights from the anthropology of militarization and militaries demonstrate that to fully understand militarization, and to possibly demilitarize a society, there is a need to take into account the connections anthropologists have found between militarization and gender, sexuality, language, education, race, class, age, ideology, health, identity, status, honor, shame, heroism, 'soldiers,' veterans, sacrifice, media, entertainment, esthetics, families, development, globalization, military bases at home and abroad, science, religion, the state, the nation, space, place, the 'other,' and perhaps, most importantly, how individuals think about all these, and create identities and worldviews shaped by and through militarization and militaries.

See also: Body: Anthropological Aspects; Borders, Anthropology of; Citizenship: Political; Collective Memory, Social Psychology of; Elites, Anthropology of; Ethnography in Applied Social Research; Gender: Anthropological Aspects; Health: Anthropological Aspects; Human Rights and Social Work; Identity in Anthropology; Peace and Nonviolence: Anthropological Aspects; Political Anthropology; War: Anthropological Aspects, Historical Development of; Xenophobia and Nativism.

Bibliography

Açiksöz, Salih Can, March 2012. Sacrificial limbs of sovereignty: disabled veterans, masculinity, and nationalist politics in Turkey. Medical Anthropology Quarterly 26 (1), 4–25.

Ben-Ari, Eyal, 1998. Mastering Soldiers: Conflict, Emotions, and the Enemy in an Israeli Military Unit. Berghahn Books, Oxford, UK.

Ben-Ari, Eyal, 2008. War, The military, and militarization around the globe. Social Anthropology 16 (1), 90–99.
Ben-Ari, Eyal, 2004. The military and militarization in the United States. American Ethnologist 31 (3), 340–348.
Besteman, Catherine, 2009. Counter AFRICOM. In: The Counter-counterinsurgency Manual, Network of Concerned Anthropologists. Prickly Paradigm Press, Chicago, pp. 115–132.
Besteman, Catherine, 2008. Beware of those bearing gifts: an anthropologist's view of AFRICOM. Anthropology Today 24 (5), 20–21.
Besteman, Catherine, 1999. Unraveling Somalia: Race, Violence and the Legacy of Slavery. University of Pennsylvania Press, Philadelphia.
Bickford, Andrew, 2013. Demilitarization: unraveling the structures of violence. In: Peter N. Stearns (Ed.), Demilitarization in Contemporary World History. The University of Illinois Press, Champaign, Illinois.
Bickford, Andrew, 2011. Fallen Elites: The Military Other in Post Unification Germany. Stanford University Press, Palo Alto.
Bickford, Andrew, September 30, 2010. "Pat Tillman and Why Soldier Hero Worship Serves the Powerful… Not the Soldiers." "The Shadow Elite" Column. The Huffington Post.
Bickford, Andrew, 2008. "Skin-In solutions:" militarizing medicine and militarizing culture in the United States military. North American Dialogue 11 (1), 2008.
Boas, Franz, December 20, 1919. Scientists as Spies. The Nation.
Carbonella, August, 2003. Memories of imperialism: the figure of the Viet Nam veteran in the age of globalization. Focaal: European Journal of Anthropology 2003 (41).
Cohn, Carol, Summer 1987. Sex and death in the rational world of defense intellectuals. Signs 12 (4), 687–718. Within and Without: Women, Gender, and Theory.
Evans-Pritchard, E.E., 1969. The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People. Oxford University Press, Oxford.
Feldman, Allen, May 1994. On cultural anesthesia: from desert storm to Rodney King. American Ethnologist 21 (2), 404–418.
Ferguson, R. Brian, 2009. Ten points on war. In: Waterston, Alisse (Ed.), An Anthropology of War: Views from the Frontlines. Berghahn Books, New York.
Finley, Erin P., 2011. Field of Combat: Understanding PTSD among Veterans of Iraq and Afghanistan. ILR Press (An Imprint of Cornell University Press), Ithaca and London.
Forte, Maximilian, 2011. The human terrain system and anthropology: a review of the ongoing public debates. American Anthropologist 113 (1), 149–153.
Fosher, Kerry, 2011. Yes, both, absolutely: a personal and professional commentary on anthropological engagement with military and intelligence operations. In: Kelly, John D., Jauregui, Beatrice, Mitchell, Sean T., Walton, Jeremy (Eds.), Anthropology and Global Counterinsurgency. University of Chicago Press, Chicago.
Frese, Pamela R., Harrell, Margaret C. (Eds.), 2003. Anthropology and the United States Military: Coming of Age in the Twenty-first Century. Palgrave, New York.
Frühstuck, Sabine, 2007. Uneasy Warriors: Gender, Memory, and Popular Culture in the Japanese Army. University of California Press, Berkeley.
Fry, Douglas P., 2007. Beyond War: The Human Potential for Peace. Oxford University Press, New York.
Fry, Douglas P., 2006. The Human Potential for Peace: An Anthropological Challenge to Assumptions about War and Violence. Oxford University Press, New York.
Gill, Lesley, 2009. War and peace in Columbia. In: Waterston, Alisse (Ed.), An Anthropology of War: Views from the Frontline. Berghahn Books, New York.
Gill, Lesley, 2004. The School of the Americas: Military Training and Political Violence in the Americas. Duke University Press, Durham and London.
Gill, Lesley, 1997. Creating citizens, making men: the military and masculinity in Bolivia. Cultural Anthropology 12 (4), 527–550.
Gonzalez, Roberto, 2010. Militarizing Culture: Essays on the Warfare State. Left Coast Press, Walnut Creek, CA.
Gonzalez, Roberto, 2009. American Counterinsurgency and the Human Terrain. Prickly Paradigm Press, Chicago.
Green, Linda, 1999. Fear as a Way of Life. Columbia University Press, New York.
Gusterson, Hugh, 13 October 2011. Death by drone. The Bulletin of the Atomic Scientists.
Gusterson, Hugh, 10 March 2009. Empire of bases. The Bulletin of the Atomic Scientists.
Gusterson, Hugh, 2007. Anthropology and militarism. Annual Review of Anthropology 36, 155–175.
Gusterson, Hugh, 1996. Nuclear Rites: A Weapons Laboratory at the End of the Cold War. University of California Press, Berkeley.
Gutmann, Matthew, Lutz, Catherine, 2010. Breaking Ranks: Iraq Veterans Speak Out against the War. University of California Press, Berkeley.
Hautzinger, Sarah, Scandlyn, Jean N., 2013. Beyond Post-Traumatic stress: Homefront Struggles with the wars on terror. Left Coast Press, Walnut Creek, CA.
Herzfeld, Michael, 2001. Anthropology: Theoretical Practice in Culture and Society. Blackwell Publishers, Inc, Malden, MA.
Hinton, Alexander Laban, 2004. Why Did They Kill? Cambodia in the Shadow of Genocide. University of California Press, Berkeley.
Hinton, Alexander Laban, 2002. The dark side of modernity: towards an anthropology of genocide. In: Hinton, Alexander Laban (Ed.), Annihilating Difference: The Anthropology of Genocide. The University of California Press, Berkeley.
Hoffman, Danny, Spring 2005. West-African warscapes: violent events as narrative blocs: the disarmament at Bo, Sierra Leone. Anthropological Quarterly 78 (2), 328–353.
Hutchinson, Sharon, 1996. Nuer Dilemmas: Coping with Money, War, and the State. University of California Press, Berkeley.
Irwin, Anne., 2011. Military ethnography and embedded journalism: parallels and intersections. In: McNamara, Laura A., Rubenstein, Robert A. (Eds.), Dangerous Liaisons: Anthropologists and the National Security State. SAR Press, Santa Fe, New Mexico.
Kanaaneh, Rhoda, 2005. Boys or men? Duped or "Made?" Palestinian soldiers in the Israeli military. American Ethnologist 32 (2), 260–275.
Kohn, Richard, January 2009. The Danger of militarization in an Endless 'War' on Terrorism. Journal of Military History, 177–208.
Kelly, John D., Jauregui, Beatrice, Mitchell, Sean T., Walton, Jeremy (Eds.), 2010. Anthropology and Global Counterinsurgency. University of Chicago Press, Chicago.
Lutz, Catherine, 2009a. The military Normal: feeling at home with counterinsurgency in the United States. In: The Counter-counterinsurgency Manual, Network of Concerned Anthropologists. Prickly Paradigm Press, Chicago.
Lutz, Catherine, 2009b. The Bases of Empire: The Global Struggle against U.S. Military Posts. NYU Press, New York.
Lutz, Catherine, 2007. Militarization. In: Nugent, David, Vincent, Joan (Eds.), A Companion to the Anthropology of Politics. Blackwell, Malden, MA.
Lutz, Catherine, 2002. Making war at Home and in the United States: militarization and the current crisis. American Anthropologist 104 (3), 723–735.
Lutz, Catherine, 2001. Homefront: A Military City and the American 20th Century. Beacon Press, Boston.
MacLeish, Kenneth, 2013. Making War at Fort Hood: Life and Uncertainty in a Military Community. Princeton University Press, Princeton.
Malkki, Liisa, 1995. Purity and Exile: Violence, Memory, and National Cosmology Among Hutu Refugees in Tanzania. University of Chicago Press, Chicago.
Mamdani, Mahmood, 2002. Good muslim, bad muslin: a political perspective on culture and terrorism. American Anthropologist 104 (3), 766–775.
Manz, Beatriz, 2009. The continuum of violence in post-war guatemala. In: Waterston, Alisse (Ed.), An Anthropology of War: Views from the Frontline. Berghahn Books, New York.
Masco, Joseph, 2006. The Nuclear Borderlands: The Manhattan Project in Post-Cold War New Mexico. Princeton University Press, Princeton.
McCaffrey, Katherine T., 2002. Military Power and Popular Protest: The U.S. Navy Base in Vieques. Rutgers University Press, Puerto Rico. New Brunswick.
McFate, Montgomery, 2005. Anthropology and counterinsurgency: the strange Story of their curious relationship. Military Review, 24–38. March/April.
McFate, Montgomery, Fondacaro, Steve, 2011. Reflections on the human terrain system during the first four years. Prism 2 (4), 63–82.
McNamara, Laura A., Rubinstein, Robert A. (Eds.), 2011. Dangerous Liaisons: Anthropologists and the National Security State. Sante Fe: School for Advanced Research Press.
Mead, Margaret, 1940. War Is Only an Invention - Not a Biological Necessity. Asia, XL, pp. 402–405.
Messinger, Seth, 2010. Getting past the accident: explosive devices, limb loss, and refashioning a life in a military medical center. Medical Anthropology Quarterly 24 (3), 281–303.
Network of Concerned Anthropologists (Eds.), 2009. The Counter-Counterinsurgency Manual, Or Notes on Demilitarizing American Society. Prickly Paradigm Press, Chicago.
Nordstrom, Carolyn, 2009. Prelude: an accountability, written in the Year 2109. In: Waterston, Alisse (Ed.), An Anthropology of War: Views from the Front Line. Berghahn Books, New York.
Nordstrom, Carolyn, 2004. Shadows of War: Violence, Power, and International Profiteering in the Twenty-first Century. University of California Press, Berkeley.

Olujic, Maria B., 1998. Embodiment of terror: gendered violence in peacetime and wartime in Croatia and Bosnia-Herzegovina. Medical Anthropology Quarterly 12 (1), 31–50.
Perez, Gina M., 2006. How a scholarship girl becomes a soldier: the militarization of latina/o youth in chicago public schools. Identities: Global Studies in Culture and Power 13 (1), 53–72.
Price, David, 2011. Weaponizing Anthropology: Social Science in Service of the Militarized State. Counterpunch and AK Press, Oakland.
Price, David, 2008. Anthropological Intelligence: The Deployment and Neglect of American Anthropology in the Second World War. Duke University Press, Durham and London.
Price, David, 2004. Threatening Anthropology: McCarthyism and the FBI's Surveillance of Activist Anthropologists. Duke University Press, Durham and London.
Robben, Antonius, 1995. The politics of truth and emotion among victims and perpetrators of violence. In: Nordstrom, Carolyn, Robben, Antonius (Eds.), Fieldwork under Fire: Contemporary Studies of Violence and Survival. University of California Press, Berkeley.
Rubenstein, Robert, Fosher, Kerry, Fujimura, Clementine (Eds.), 2012. Practicing Military Anthropology: Beyond Expectations and Traditional Boundaries. Kumarian Press, West Hartford, CT.
Russ, Ann J., Ames, Genevieve M., January–February, 2006. Policy and prevention as competing narratives in us navy life and medicine. Culture, Health, and Sexuality 8 (1), 1–15.
Schaft, Gretchen, 2007. From Racism to Genocide: Anthropology in the Third Reich. University of Illinois Press, Urbana.
Scheper-Hughes, Nancy, Bourgois, Philippe, 2004. Making sense of violence. In: Scheper-Hughes, Nancy, Bourgois, Philippe (Eds.), Violence in War and Peace. Blackwell, Malden, MA.
Selmeski, Brian, 2007. Military Cross-cultural Competence: Core Concepts and Individual Development. Centre for Security, Armed Forces, and Society, Royal Military College of Canada.
Simons, Anna, 1999. War: back to the future. Annual Review of Anthropology 28, 73–108.
Simons, Anna, 1997. The Company They Keep: Life inside the U.S. Army Special Forces. Avon Books, New York.
Starn, Orin, 1991. Missing the revolution: anthropologists and the war in Peru. Cultural Anthropology 6 (1), 63–91.
Tilly, Charles, 1999. War making and state making as organized crime. In: Evans, Peter B., Rueschemeyer, Dietrich, Skocpol, Theda (Eds.), Bringing the State Back. Cambridge University Press, Cambridge.
Vasquez, Jose N., 2009. Seeing green: visual technology, virtual reality, and the experience of war. In: Waterston, Alisse (Ed.), An Anthropology of War: Views from the Frontline. Berghahn Books, New York.
Vine, David, 2012. U.S. Empire of Bases Grows. TomDispatch.Com.
Vine, David, 2009. Islands of Shame: The Secret History of the U.S. Military Base on Diego Garcia. Princeton University Press, Princeton.
Waterston, Alisse (Ed.), 2009. An Anthropology of War: Views from the Frontline. Berghahn Books, New York.
Wagner, Sarah E., 2013. The making and unmaking of an unknown soldier. Social Studies of Science 43 (5), 631–656.
Wagner, Sarah E., 2008. To Know where He Lies: DNA Technology and the Search for Srebrenica's Missing. University of California Press, Berkeley.
Wagner, Sarah E., Quintyn, C., May 2009. Dismantling a national icon: genetic testing and the tomb of the unknowns. Anthropology News 50 (5).
Wakin, Eric, 1992. Anthropology Goes to War: Professional Ethics and Counterinsurgency in Thailand. University of Wisconsin Press, Madison.
Weiss, Erica, March 2012. Principle or pathology? Adjudicating the right to conscience in the Israeli military. American Anthropologist 114 (1), 81–94.
Wool, Zoë, Messinger, Seth, March 2012. Labors of love: the transformation of care in non-medical attendant program at walter reed army medical center. Medical Anthropology Quarterly 26 (1), 26–48.

Militarism

Anna Stavrianakis, University of Sussex, Brighton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of militarism has had an ambivalent status in the social sciences, veering between polemic and analytical usage. This article seeks to clear the ground by outlining five main definitions of militarism, as: an ideology glorifying war; the propensity to use force; military buildup; excessive influence (of either the institution of the military or of the military–industrial complex); and the influence of military relations on social relations in general. It then discusses the modernist tilt to much of the discussion and questions of methodological nationalism and the appropriate scale of analysis. Finally, it addresses some of the normative issues around the study of militarism.

Introduction

Militarism has an ambivalent status as a social science concept. Subject to varying definitions, conceptualizations and approaches within and between disciplines, its usage has veered between pejorative polemic and social scientific analysis, and it has experienced peaks and troughs in academic discourse according to international and historical conjuncture. While war and military power have a much longer history, both as practices and as objects of analysis, the academic study of militarism took off in the late nineteenth century, becoming particularly prominent in relation to Germany in the context of World War I (WWI). Since then, the production of knowledge about militarism has both shaped, and been shaped by, the practice of war, war preparation, and military power itself. The various definitions and trends in the study of militarism morph as the practice of war, and the exercise of power in international politics, changes.

Others have provided chronological accounts of the debate over militarism (e.g., Berghahn, 1981; Bucholz, 1999). In what follows, I provide a fivefold set of definitions of militarism that crosses some of the chronological and theoretical boundaries that typify debate. Militarism can variously be understood as: an ideology glorifying war; the propensity to use force; military buildup; excessive influence (of either the institution of the military or of the military–industrial complex (MIC)); or the influence of military relations on social relations in general. These definitions become progressively more sociological: for example, the use of force may not rest on outright glorification of war and may draw on other sources of social power; military buildup can occur in scenarios without the direct of military force; and the military can exercise influence even when it does not seize power directly through a coup. I then discuss two key issues that cut across these definitions and shape the overall debate around militarism, namely the modernist tilt to much of the discussion, and questions of methodological nationalism and the appropriate scale of analysis. Finally, in conclusion I address some of the normative issues around the study of militarism.

Definitions and Approaches

Militarism as Ideology

Historically, militarism has been widely understood as an ideology that glorifies war, military institutions, and martial values. A 'powerful war mythology' took hold in Europe in the decades prior to 1914, emphasizing elite and noble values, honor, courage, duty, and patriotism (Bucholz, 1999: 426–427), with militarism assuming 'the importance of a basic cultural value' in WWI Germany (Berghahn, 1981: 32). Outside of Europe, scholars have emphasized the spiritual and ideological dimensions of late nineteenth- and early twentieth-century Japanese militarism seen through concepts such as *bushido*, the way of the warrior or samurai (Bucholz, 1999: 427). Alfred Vagts later made the distinction between the rational and effective 'military way' in the conduct of war, and militarism: that "complex of feelings which rank military institutions and ways above the ways of civilian life, carrying military mentality and modes of acting and decision into the civilian sphere." For him, militarism was characterized by "qualities of caste and cult, authority and belief," and involved "an emphasis on military considerations, spirit, ideals, and scales of value, in the life of states" (Vagts, 1959: 13–17).

Ideological definitions are now rather rare, both because militarism itself is rarely a primarily ideological phenomenon, and also because definitions based on ideology are analytically inadequate. The outright valorization of war and state militaries is now less prevalent: the social conditions that facilitated the mass ideological mobilization of the early to mid-twentieth century have changed. Societal mobilization for and justification of war depend on more than glorification, and take a variety of forms in different societies and across time. As Shaw puts it, "[i]t is sociologically unenlightening to restrict the meaning to ideology: the core idea is the 'carrying' of military forms into the civilian sphere," which requires analysis of practices as well as mentalities and attitudes. In addition, these military forms may justify war and military power without actively glorifying them (Shaw, 2013: 19–20). War and war preparation have not gone away, but they are created, perpetuated, and justified in different ways. While an ideological dimension remains important in various contemporary scholarly debates, with historical sociologists such as Michael

Mann and feminists such as Cynthia Enloe emphasizing the attitudes, values, and beliefs that characterize militarism (Mann, 1987: 71; Enloe, 1988: 7), they embed this dimension of militarism in more sociological accounts that emphasize material as well as ideational context, and move beyond the immediate glorification of war and military institutions.

Militarism as the Use of Force to Resolve Conflict

A second type of definition is behavioral: militarism as the propensity to utilize force to resolve conflict. For structural realists in International Relations, for example, militarism explains those instances where great powers act in nonstrategic ways. Militarism is an aberration in an anarchic international system, in which the use of force is the ultimate arbiter (Van Evera, cited in Mearsheimer, 2010: 84). Eide and Thee, meanwhile, working in the peace research tradition, define militarization as "the inclination to rely on military means of coercion for the handling of conflicts" (1980: 9). Such a tradition has continued in the post-Cold War peace research agenda on "Militarized Interstate Disputes," which studies the threat, display, or use of force by states against each other (Correlates of War, n.d.). For others, meanwhile, militarism as the use of force is not restricted to interstate relations: some focus on the domestic use of force against recalcitrant populations. For many Marxists, militarism has a dual aspect, being exercised against internal as well as external enemies to promote the expansion of capitalist power outside the territorial confines of the nation-state, and the maintenance of order within the nation-state (e.g., Liebknecht, 1973).

For all their differences, behavioral approaches do not assume that a propensity to use force is necessarily or essentially rooted in martial values or the glorification of war. The emphasis on behavior, however, focuses on the directly observable use of force while downplaying the political or sociological processes underlying not only war but military power more generally. As Skjelsbaek argues, while militarist behavior, "the excess use of violence" or "the actual use of an excessive military apparatus in war," is in some ways the most direct, central, and observable dimension of militarism, this underplays the latent use of weapons, social structures, ideology, strain on the economy of military spending, and so on (Skjelsbæk, 1980: 87). That is, militarism draws upon numerous sources for its power and legitimacy, not just on acts of violence, and can remain entrenched even in relatively pacific (at least international) contexts. Behavioral definitions are thus rather narrow, relying on directly observable phenomena, underplaying the wider role of military power in society.

Militarism as Military Buildup

A third definition is of militarism as military buildup, in particular, quantitative increases in weapons production and imports, military personnel, and military expenditure. This concern was shared by both strategic studies and peace research scholars during the Cold War, despite their different research focus and politics. Strategists were more interested in superpower nuclear rivalry, deterrence, arms racing, action–reaction models, and so on, while peace researchers were concerned about spiraling military spending and the transfer of weapons to the South. Ross, for instance, understands and measures militarism by six indicators: levels of military expenditure, arms imports, and arms production; size of armed forces; and number of wars and military regimes (Ross, 1987). This type of definition survived the end of the Cold War and continues to be prominent in both research agendas. In contemporary peace research, a predilection for statistical analysis has meshed with post-Cold War concerns around liberal democracy, trade liberalization, ethnicity, and so on in assessing the likelihood of militarization. And in contemporary strategic studies, such quantitative accounting of military forces and equipment forms the bedrock of analyses of the military balance of power and likely outbreak of war.

This type of work lends quantitative robustness to the study of militarism, allowing for the identification of patterns and thus comparison. However, in a similar vein to behavioral definitions, what remains unaddressed is the social embeddedness, power, or meanings of militarism, or those social, political, and international structures and processes in which military buildups are so powerfully rooted (for a critique of the thinness of this 'militarization thesis,' see Shaw, 1991: 66–73). A more sociological account of military buildup can be seen in Kaldor's Cold War-era work on the increasingly 'baroque' character of modern military technology (Kaldor, 1982a), grounding military production in political economy; and in E.P. Thompson's critique of the ideology of exterminism that of which the superpower nuclear arms race was a symptom (Thompson, 1982).

Militarism as Excessive Influence

A fourth definition is of militarism as the excessive influence of either the institution of the state military or of the MIC. The former is characteristic of a civil–military relations approach, of which there are various strands. These include a US-focused literature on the rise of a standing army post-WWII (e.g., Lasswell, 1950; Huntington, 1957; Janowitz, 1960); and the Cold War concern with the role of the military as a potential modernizing force in so-called 'new nations' (e.g., Huntington, 1968; Janowitz, 1977; Pye, 1962). The literature on the developing world initially mostly focused on coups, but later shifted to a broader spectrum of military influence, because of the observation that the military can be powerful even if it does not seize power directly.

There is debate as to whether it is the characteristics of the military itself, or society, that are the more important variable in understanding militarism. However, the civil–military relations approach as a whole rests on the problematic idea of an 'equation,' with the military on one side and civilian institutions and procedures on the other (Valenzuela, 1985). But the civilian side is not just another institution; the military is an actor within a social and political system, not a neutral force outside its boundaries. The two 'sides' are thus highly interpenetrated. Civil–military relations discussions of militarism – in which militarism occurs when the 'balance' tilts too far in favor of the military – are thus limited by a conceptual separation between military and society that obscures the grounding of military power in social relations.

Literature on the MIC starts to overcome this pluralist separation between the military and civilian institutions, focusing

instead on the shared interests between the military, military industry, top-level government bureaucrats, and legislators and the influence they exercise on society as a whole. This type of approach is more focused on the socioeconomic dimensions of militarism, as compared to the institutional approach of the civil–military relations literature. There are a variety of theoretical perspectives in play, but the result of these relationships is variously characterized as coordinated and mutually supportive influence that creates a shared interest in continued military spending (Rosen, 1973: 2–3), "a set of commonly shared interests between the military and some major corporations" (Lieberson, 1973: 61), and a "self-serving accommodation between corporate elites, government bureaucrats, and the military hierarchy" (Moskos, 1972: 4).

The civil–military relations and MIC approaches are both useful ways in two particular aspects of militarism – the institutional and the political-economic. But the focus on either institutions or economic relations cannot adequately explain militarism in its broader sense of the preparation for and conduct of organized political violence. In addition, both approaches tend to be nationally based in their explanations. MIC accounts are useful for thinking about the political economy of militarism, and thus start to ground militarism in social relations (e.g., between arms companies, politicians, and the military); but usually remain focused on national economies, which is problematic given the internationalization of military industry. And civil–military relations approaches have tended to assume a sovereign nation-state, which is problematic given the internationalization of military power through alliances, foreign military training, arms transfers, and so on. These processes challenge the basic assumption that the key explanatory variables (military and domestic political process) are located within national societies (Valenzuela, 1985).

Militarism as the Influence of Military Relations on Social Relations

The definitions discussed thus far have each provided insights into particular dimensions of militarism; yet they tend to lack sociological depth. That is the impetus to the fifth definition provided here, that of militarism as the influence of military relations on social relations. In this, the component parts of other definitions – ideology, military buildup, institutional relations, and so on – are understood as embedded in society. It is this wider social context that is crucial for understanding the way that not only war, but also war preparation and military power, are produced, reproduced, and, at times, challenged. As Martin Shaw defines it, militarism is "the penetration of social relations in general by military relations" or the "relationship of war preparation and society" (Shaw, 1991, 2013: 9–15). Sociological approaches can be inflected in a variety of ways theoretically, from the neo-Weberian historical sociology of Shaw and Michael Mann, to the Marxism of Mary Kaldor's Cold War work on the mode of warfare (1982b) and feminist sociologies of the mutually constitutive relations between militarism and patriarchy (e.g., Cock, 1991; Enloe, 1988).

This is a much more abstract definition, the empirical content of which depends on the case under analysis. In contrast to most approaches, which see militarism as an end-state, and militarization as the process through which it is reached, militarism here is an abstraction rather than a phenomenon. That is, the concept of 'military' is used not as a noun to refer to an institution, but to describe "all social relations, institutions and values relating to war and war preparation" (Shaw, 2003: 106); and militar*ism*, in turn, is understood as the tendency or extent to which these military relations influence social relations as a whole. Militarization and demilitarization can be measured and their character assessed: as military relations exercise greater influence on social relations in general, militarization increases, and vice versa. This approach is more open analytically, in terms of the scale at which militarism is to be studied, and the normative ramifications of analysis. Before addressing this, I turn first to the prior question of the relationship of militarism to modernity, which shapes much of the debate.

Militarism and Modernity

While war and military power, and the study thereof, obviously have a much longer history, the concept and academic study of militarism is a modern phenomenon, emerging in the nineteenth century (Berghahn, 1981: 7). And there is something specific about modern militarism: Vagts' general definition of militarism as 'a tendency to extend dominion,' which is more clearly transhistorical, is qualified by the specificity of modern militarism, in which militaries are not constantly engaged in combat, so they become narcissistic (Berghahn, 1981: 38–39). Much social science discussion is marked by a similarly distinctively modernist orientation. Despite the significant differences between them, liberal and Marxist discussions share the idea of historical progress and development leading to the withering of militarism (Shaw, 1991: 15–17). Classical liberals believed that militarism would die out with the industrialization and modernization associated with the development of capitalism, as seen in Herbert Spencer's distinction between militant and industrial society, for example. For Marxists, while many saw capitalism as militaristic, they still anticipated that the transition to communism would mean militarism died out. Social theory, itself a child of modernity, has long held militarism to be a throwback to a more violent past, one that will be exorcised through progress and development.

This logic of progress was evident in discussions of Prusso-German and Japanese militarism, revolving around the relationship between 'backwardness' and the transition from agrarian to industrial society; whether these should be considered similar or different cases in terms of their levels of capitalist development and industrialization; and the distinction from the more 'advanced' West. Part of this was also about allowing Anglo-Saxon liberals to refute the Marxist critique that Western capitalist-industrial societies were necessarily militaristic (Berghahn, 1981: 67, 120). The logic of progress is also clearly seen in discussions of the military as a modernizing agent in the Third World. The emergence of this literature was part of the growth of a comparative politics approach in political science, which featured the rise of structural-functional schemes and social systems metaphors (Valenzuela, 1985: 133). The context of the

modernization debate is evolutionary theories of society and the effort to ensure Third World countries become capitalist-industrial societies, with free market economies, parliamentary-democratic institutions, and industrialization (Berghahn, 1981: 71).

This modernist belief has been challenged from various quarters, however. Not only is there a marginalized 'bellicose tradition' in older continental European social theory (Malesevic, 2010), but contemporary historical sociologists have attempted to shatter "the dream of a modernity without violence" (Joas, 2003). In particular, there is a strand of scholarship that emphasizes the autonomy of war and militarism as social forces in their own right (Shaw, 1988; Barkawi, 2011) rather than an expression or symptom of either geopolitics or capitalism. Feminist and postcolonial scholars have also troubled the notion that militarism is a remnant of a more violent past. Gender roles in war are consistent across space and time (despite the diversity of gender and war separately) (Goldstein, 2006: 3) and militarism relies on particular forms of masculinities in societies "with otherwise dissimilar cultures and uneven levels of industrialisation" (Enloe, 1993: 72–73). Crucially, these are also racialized as well as gendered. Bell Hooks emphasizes not only "the contradiction that black men should serve in wars, die for this country, for this democracy, which institutionalized racism and denied them freedom" (1995: 58–59), but also that "Black women are very likely to feel strongly that white women have been quite violent, militaristic in their support and maintenance of racism" (Hooks, 1995: 59–60). As other postcolonial scholars remind us, from the periphery, modernity has never looked pacific (Krishna, 2001).

Scale of Analysis

Discussions of the relationship of militarism to geopolitics, capitalism, and patriarchy raise methodological questions around the scale at which militarism is studied. Many studies of militarism – despite different conceptual and theoretical premises – use an essentially statist or methodologically nationalist framework for studying militarism, especially in country case studies, studies of civil–military relations, and also quantitative analyses which compare different levels of militarization between states. Others, by contrast, are implicitly or explicitly critical of such state-centrism, emphasizing instead the ways, in which militarism is structured by, or embedded within, transnational, imperial, or global social relations. Thee, for example, distinguishes between the outward-oriented militarism of great powers and the inward-oriented militarism of developing countries governed by military regimes, arguing that the former is "also largely the root cause and the driving force behind the global spread of militarism" (Thee, 1980: 21–24). Militarism as a global phenomenon is thus shaped by relations between core and periphery. Scholars of various persuasions have tackled this, from the structural arguments of Cold War peace researchers to identity-based arguments of post-Cold War constructivists.

Scholars working in more anthropological and sociological traditions, meanwhile, have scaled down the study of militarism to specific sites and to questions of subjectivity. Feminist analyses have demonstrated that both liberal and Marxist approaches may have "dramatically *under*estimated the power it requires to militarize any society" (Enloe, 1993: 51). This draws our attention to the societal level and everyday scale in order to understand both the impact of war and militarism on people (especially, but not only, women) as well as the gendered construction of war and militarism (Sjoberg and Via, 2010: 231–232).

It is not a question of 'choosing' an appropriate level – be it international, national, or subnational – however. As Sjoberg and Via put it, "absolutely distinguishing between the personal, national and international level of war and militarism lacks conceptual and empirical rigor at best" (2010: 233–234) as the so-called 'levels' are interdependent. The various 'levels of analysis' need to be considered in co-constitutive interaction (Stavrianakis and Selby, 2013a: 15): militarism at any scale is shot through with international relations, and any macro analysis needs to take social relations into account.

Conclusion

Militarism has an ambiguous status as a social science concept, deriving at least in part from its polemic or pejorative overtones (Shaw, 2013: 19): militarism is often an accusation rather than an analysis. In addition to the sometimes problematic ways the concept itself has been used, the wider intellectual climate has been more or less hospitable to discussions of militarism over time. For example, the post-Cold War political and intellectual hegemony of liberalism, the rise of influential and policy-related discourses on 'failed states' and 'new wars,' and the dominance of discussions of 'security' and 'securitization' have all contributed to the marginalization of militarism as a key concept in International Relations (Stavrianakis and Selby, 2013a: 6). Despite this, the concept has a rich and varied history, as the various definitions have shown; and across the social sciences it is still in play, often in rich and innovative ways.

Much scholarship on militarism is informed by some degree or form of antimilitarism – whether it be criticism of the excessive buildup of military capabilities, a critique of military power per se and its relationship to capitalism, or opposition to the gendered ways in which men and women are incorporated into war practices. The possibility for opposition to militarism depends on both the character of militarism itself in any given place and time, as well as on wider socioeconomic, political, and cultural patterns (Shaw, 2013). For others, an oppositional politics of antimilitarism is replaced by a politics of 'countermilitarism.' This involves responding to the diffuse character of contemporary military power with broader forms of critical engagement beyond simple opposition, such as dialogue, subversion, and mimicry (Stavrianakis and Selby, 2013b). And critical intellectual positions do not map easily on to political persuasions – conservative critics of US foreign policy have resurrected the language of militarism in relation to the wars in Iraq and Afghanistan (e.g., Bacevich, 2005). Such differences speak to the various ways that political and intellectual sensibilities intersect.

To be sure, the normative issues around militarism are not clear-cut. There is plenty of contemporary empirical evidence of

practices we may want to associate with militarism, such as the recent and ongoing wars in Afghanistan, Sudan, Libya, Syria, and elsewhere; rising military spending under the rubric of the War on Terror; and military coups in Mali, Niger, Thailand, Honduras, and elsewhere. But their causes, meaning, and relative significance are disputed. To take just two examples: liberal wars of intervention under the guise of humanitarianism and liberation have occurred in a context of structural demilitarization of Western societies and economies since the end of WWII. So force can still be deployed outward while domestic society is less obviously militarized – perhaps even more so because of the growing separation of the military from society and its professionalization. And liberal equality agendas within some national militaries are widely seen as progressive in terms of the rights of women, homosexuals, and ethnic minorities. Yet they simultaneously reinforce militarism by incorporating marginalized parts of the population into war fighting and war preparation. These examples suggest the need for an ongoing discussion about militarism, to which conceptual clarity can make a significant contribution.

See also: Arms Control; Cold War, The; Military Sociology; Military, War, and Politics; Nation-State and War; Securitization; War, Sociology of.

Bibliography

Bacevich, Andew, 2005. The New American Militarism: How Americans Are Seduced by War. Oxford University Press, Oxford.

Barkawi, Tarak, 2011. From war to security: security studies, the wider agenda and the fate of the study of war. Millennium 39 (3), 701–716.

Berghahn, Volker R., 1981. Militarism: The History of an International Debate 1861–1979. Cambridge University Press, Cambridge.

Bucholz, Arden, 1999. Militarism. Encyclopedia of Violence, Peace, and Conflict, 2, 423–432.

Cock, Jacklyn, 1991. Colonels and Cadres: War and Gender in South Africa. Oxford University Press, Cape Town, Southern Africa.

Correlates of War, n.d. Militarized Interstate Disputes (v3.10) http://www.correlatesofwar.org/COW2%20Data/MIDs/MID310.html.

Eide, Asbjørn, Thee, Marek, 1980. Introduction. In: Eide, Asbjørn, Marek, Thee (Eds.), Problems of Contemporary Militarism. Croom Helm, London, pp. 9–11.

Enloe, Cynthia, 1988. Does Khaki Become You? The Militarization of Women's Lives. Pluto Press, London.

Enloe, Cynthia, 1993. The Morning after: Sexual Politics at the End of the Cold War. University of California Press, Berkeley.

Goldstein, Joshua S., 2006. War and Gender. How Gender Shapes the War System and Vice Versa. Cambridge University Press, Cambridge.

Hooks, Bell, 1995. Feminism and militarism: a comment. Women's Studies Quarterly 3 & 4, 58–64.

Huntington, Samuel P., 1957. The Soldier and the State: The Theory and Politics of Civil-Military Relations. Harvard University Press, Cambridge, MA.

Huntington, Samuel P., 1968. Political Order in Changing Societies. Yale University Press, New Haven, CT.

Janowitz, Morris, 1960. The Professional Soldier: A Social and Political Portrait. The Free Press, New York.

Janowitz, Morris, 1977. Military Institutions and Coercion in the Developing Nations. University Chicago Press, Chicago.

Joas, Hans, 2003. War and Modernity: Studies of Violence in the 20th Century. Polity, Cambridge.

Kaldor, Mary, 1982a. The Baroque Arsenal. Hill and Wang, New York.

Kaldor, Mary, 1982b. Warfare and capitalism. In: New Left Review (Ed.), Exterminism and Cold War. Verso, London, pp. 261–288.

Krishna, Sankaran, 2001. Race, amnesia and the education of international relations. Alternatives 26, 401–424.

Lasswell, Harold D., 1950. National Security and Individual Freedom. McGraw-Hill, New York.

Liebknecht, Karl, 1973. Militarism and Anti-militarism. Rivers Press Ltd, Cambridge.

Malesevic, Sinisa, 2010. The Sociology of War and Violence. Cambridge University Press, Cambridge.

Mann, Michael, March–April 1987. The roots and contradictions of modern militarism. New Left Review I (162).

Mearsheimer, John, 2010. Structural realism. In: Dunne, Kurki, Smith (Eds.), International Relations Theories. Discipline and Diversity, second ed. Oxford University Press, Oxford, pp. 71–88.

Moskos Jr., Charles C., 1972. The military-industrial complex: theoretical antecedents and conceptual contradictions. In: Sarkesian, Sam (Ed.), The Military-Industrial Complex. A Reassessment. Sage, Beverly Hills, pp. 3–23.

Pye, Lucien W., 1962. Armies in the process of political modernisation. In: Johnson, John. J. (Ed.), The Role of the Military in Underdeveloped Countries. Princeton University Press, Princeton, NJ, pp. 69–89.

Rosen, Steven, 1973. Testing the theory of the military-industrial complex. In: Rosen, Steven (Ed.), Testing the Theory of the Military-Industrial Complex. Lexington Books, Lexington, MA, pp. 1–27.

Ross, Andrew L., 1987. Dimensions of militarization in the third world. Armed Forces and Society 13 (4), 561–578.

Shaw, Martin, 1988. Dialectics of War: An Essay in the Social Theory of Total War and Peace. Pluto Press, London.

Shaw, Martin, 1991. Post Military Society: Militarism, Demilitarization and War at the End of the Twentieth Century. Polity, Cambridge.

Shaw, Martin, 2003. War and Genocide: Organized Killing in Modern Society. Polity, Cambridge.

Shaw, Martin, 2013. Twenty-first century militarism: a historical-sociological framework. In: Stavrianakis, Anna, Selby, Jan (Eds.), Militarism and International Relations: Political Economy, Security, Theory. Routledge, Abingdon, pp. 19–32.

Sjoberg, Laura, Via, Sandra, 2010. Conclusion: the interrelationship between gender, war, and militarism. In: Sjoberg, Laura, Sandra, Via (Eds.), Gender, War and Militarism. Praeger, Santa Barbara, CA, pp. 231–240.

Skjelsbæk, Kjell, 1980. Militarism, its dimensions and corollaries: an attempt at conceptual clarification. In: Eide, Asbjørn, Thee, Marek (Eds.), Problems of Contemporary Militarism. Croom Helm, London, pp. 77–105.

Stavrianakis, Anna, Selby, Jan, 2013a. Militarism and international relations in the twenty-first century. In: Stavrianakis, Anna, Selby, Jan (Eds.), Militarism and International Relations: Political Economy, Security, Theory. Routledge, Abingdon, pp. 3–18.

Stavrianakis, Anna, Selby, Jan, 2013b. War becomes academic: human terrain, virtuous war and contemporary militarism. An interview with James Der Derian. In: Stavrianakis, Anna, Selby, Jan (Eds.), Militarism and International Relations: Political Economy, Security, Theory. Routledge, Abingdon, pp. 59–74.

Thee, Marek, 1980. Militarism and militarisation in contemporary international relations. In: Eide, Asbjørn, Thee, Marek (Eds.), Problems of Contemporary Militarism. Croom Helm, London, pp. 15–35.

Thompson, E.P., 1982. Notes on exterminism: the last stage of civilization. In: New Left Review (Ed.), Exterminism and Cold War. Verso, London, pp. 1–33.

Vagts, Alfred, 1959. A History of Militarism: Civilian and Military. Hollis and Carter, London.

Valenzuela, Arturo, 1985. A note on the military and social science theory. Third World Quarterly 7 (1), 132–143.

Military and Disaster Psychiatry

Neil Greenberg, King's College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D.M. Benedek, R.J. Ursano, volume 14, pp. 9850–9857, © 2001, Elsevier Ltd.

Abstract

Military and disaster psychiatrists assist both trauma-exposed people and organizations who have experienced extreme circumstances that have the potential to damage mental health and impair organizational, community, and individual functioning. While most trauma-exposed individuals remain psychologically healthy in the longer term, there is a strong evidence base showing that high-quality social support and experiencing a low, but not no, pressure environment are the most ideal conditions to support recovery. Ensuring that basic needs such as provision of shelter and food are met are also important perquisites to a good state of post-conflict or post-disaster mental health.

Individuals exposed to traumatic events are at increased risk of suffering from a range of mental health conditions including, but not limited to, posttraumatic stress disorder (PTSD; NICE, 2005). While the definition of a traumatic event has changed as new evidence has emerged, our current understanding includes a potential to affect people who are directly or indirectly exposed to an event as well as those who indirectly learn of a close relative/friend's trauma. Importantly, current diagnostic guidelines specifically acknowledge that PTSD may follow repeated or extreme indirect exposure to aversive details of events that occur in the course of professional duties (APA, 2013). This criterion specifically acknowledges that some workers are at particular, and predictable, risk of developing trauma-related mental health problems as function of their employment. Indeed, there is a wealth of evidence that as well the innocent victims of disasters, many occupational groups, including military personnel and emergency services personnel, are in substantially higher levels of trauma-related psychopathology than the general population (Skogstad et al., 2013).

Military and disaster psychiatry has evolved to take account of the cognitive, emotional, and potential neurobehavioral consequences of trauma exposure including human-made events (e.g., wars and terrorist attacks) and natural disasters (e.g., earthquakes or accidents). Mental health care interventions in military operations, and in response to disasters, generally follow a preventative medicine approach. That is to say that the most ideal occurrence is for individuals to avoid the development of mental health difficulties after exposure to traumatic events (primary prevention); where impairing symptoms do develop it is best to detect incipient mental health problems rapidly so that effective early intervention can be provided (secondary prevention); and where individuals do go on to develop formal mental health conditions, then evidence-based treatment should be provided in a culturally-sensitive manner by appropriate trained and experienced mental health professionals (tertiary prevention). Providing an evidence-based preventative approach to the prevention of traumatic stress related conditions is likely to be a legal responsibility for trauma-exposed organizations, such as the military, media companies, war correspondents, or emergency responders (UKPTS, 2014). However, the responsibility for protecting the mental health of individuals involved in disasters is less clear and the delivery of disaster psychiatry is often a result of a complex interplay of agencies including regular health care providers, local charities, nongovernmental organizations (NGOs), and local members of the community who 'pull together' in the aftermath of an incident.

Mental health care interventions delivered in theaters of war or within disaster zones are likely to differ from those delivered in traditional community- or hospital-based practice. This article examines the approaches of military and disaster psychiatry, and the cognitive, emotional, and behavioral responses associated with the events that create these environments. The principles of prevention, detection, and treatment of posttraumatic illness on the battlefield and in the aftermath of disaster are outlined. Finally, emerging cultural, ethical, and possible future technological challenges to the practice of military and disaster psychiatry are discussed.

Introduction

Both war and disaster have the potential to create an atmosphere of chaos and tragedy. Individuals who operate in these environments may experience both expected stressors, such as seeing dead bodies, and the effects of injuries in local population, and potentially among colleagues, and also a range of less predictable pressures such as being called on to make rapid decisions that have the capacity to expose individuals to moral insult or injury. For instance, individuals may be forced to choose to save their own life at the expense of someone else losing theirs. However, there is considerable evidence that while exposure to traumatic stressors may be inevitable, personnel appropriately prepared or trained to operate in these environments, may be less at risk of psychological injury than those who simply caught up in the chaotic nature of war or disaster (Rubin et al., 2005). Military and disaster psychiatry professionals have to address care demands in, or resulting from, working within nontraditional environments and in mass casualty situations, where resources are overwhelmed. Providing care in such environments is multi-professional relying on contributions of psychiatrists, mental health nurses, psychologists and social workers, other health care

workers, social scientists, epidemiologists, and emergency responders such as police and firemen. There is growing evidence that the psychological resilience of teams that operate in these environments (e.g., military forces, charitable organizations, and media) is more a function of the bonds between individuals (i.e., camaraderie and leadership) between team members than the psychological coping styles of the individuals concerned (Jones et al., 2012).

Any overview of military and disaster psychiatry begins with an examination of the consequences of traumatic situations such as disasters and wars for communities, and individuals, the provision of health and health care responses to these traumatic experiences.

The War and Disaster Environment

Disasters have the capacity to overwhelm a community's capacity to respond to them. The distinction between 'natural' disasters (e.g., earthquakes) and human-made, or technological ones, such as explosions, or train derailments, has become increasingly difficult to make. For example, much of the death and destruction from an earthquake may be due to poorly constructed housing or the failure of emergency services to respond in a timely and effective fashion. Thus, the vast majority of disasters have at least some human element to them particularly in terms of those who are affected and in the response. For example, communities may blame governments or commercial companies for failing to respond to a traumatic event in a timely and effective manner even where an incident was truly unforeseeable. From a psychological standpoint, it may be helpful to understand the inevitability, and impact, of the blame by those affected by the disaster toward statutory authorities (e.g., councils, governments, and emergency responders). In the main, blame, however well placed, impedes psychological recovery of those affected by an event.

Potential stressors in disaster environments include exposure to the dead or injured, threats to life, loss of loved ones or property, and/or physical injury. Although military forces, when commanded to do so, assist affected communities and provide a relatively safe environment for those affected by traumatic events, such safety may often be short lived.

Frequently, those so affected by war or disasters are reluctant to seek help for mental health difficulties (Osório et al., 2013), and thus do not receive treatment for psychiatric symptoms. Although traumatic events themselves can be short lived, for example, an earthquake may be concluded in seconds, the consequential traumatic experiences may continue for weeks or months and, if not managed well or treated, possibly years. For both military personnel and civilians in war or on-going disasters environments, exposure over time may include anticipated or entirely unexpected life-threatening experiences followed by daily life in an austere and disrupted environment.

In very general terms, post-disaster emotional and behavioral responses occur in four broad phases. The first, immediately following a disaster, consists of strong emotions including feelings of disbelief, numbness, fear, and confusion for those involved with the event; all are quite normal emotional responses to abnormal events. The second phase, which usually lasts from a week to several months, is accompanied by the appearance of assistance from outside agencies and communities. Adaptation to austere environments as well as intrusive symptoms (unbidden thoughts and feelings accompanied by hyperarousal) occur during this phase and physical (somatic) symptoms such as fatigue, dizziness, headaches, and nausea may develop. Anger, irritability, apathy, and social withdrawal are often present, and these cognitive and emotional responses may be misplaced. For instance, those affected by the incident(s) may begin to feel resentful or angry toward those who they perceive as having responsibility for responding to the incident. The third phase of a trauma response is often marked by increased feelings of disappointment and resentment when hopes for aid and restoration are not met. These feelings may lead to the sense of community being weakened as individuals increasingly focus on their personal needs rather than the needs of the community. The final phase, which is termed reconstruction, may last for years. During this period, people rebuild their lives, make homes and find work using available social supports. Individuals may progress through these phases at various rates. Failure to navigate through this reconstructive phase can lead to persistent symptoms.

While PTSD is a relatively new diagnosis, first appearing in the Diagnostic and Statistical Manual Volume Three in 1984, the effects of disaster and war have been recognized for many years. For instance, the emotional consequences of disaster are described in the *Iliad*, and references to the terror induced by the attack of this hero are diverse. Ancient Greeks attributed epidemic illness to Apollo's wrath after the desecration of his temple.

The Mental Health Consequences of Military Operations and Disasters

Military and disaster psychiatry professionals have to understand the multitude of behavioral reactions to the spectrum of stressful events or stressors. While there is a good evidence base concerning risk factors for the development of post-incident mental health problems such as PTSD, professionals working in this area of psychiatry also have to understand how to distinguish normal reactions from those which indicate a need for professional services. This skill is often lacking in health care professionals who do not have expertise in this area. The precipitating stressor for PTSD involves a threat to the physical integrity of self or others leading to, often short lived, distress symptoms. Often those exposed to traumatic events experience helplessness, horror, or fear during the event itself although these reactions may well be absent in those who are prepared for the traumatic event or who have experience relevant to their actions during the event itself. For instance, war journalists may run toward the sound of a shot in a purposeful way in order to get their story or soldiers who are shot at may respond purposefully in accordance with their training.

However, even among trained individuals, there is a characteristic distress response which many but not all trauma-exposed individuals may experience. This response consists of symptoms that involve: (1) 'reliving' the original event (e.g., nightmares, distressing vivid recollections, or fear when exposed to events resembling the original trauma);

(2) cognitive or behavioral avoidance of events or situations that somehow resemble or symbolize the original trauma; (3) symptoms of increased vigilance, such as exaggerated startle, outbursts of anger, or other evidence of hyperarousal; and (4) negative changes in cognitions and/or mood, such as numbing of emotional feelings for others and highly negative appraisals of the future. If these symptoms of severe distress persist for over a month, then a diagnosis of PTSD may be appropriate if the symptoms are associated with significant functional impairment.

The symptoms of impending PTSD often occur within the first month after the trauma, and if severe may be termed acute stress disorder (ASD). PTSD itself is not diagnosed until at least a month after the trauma and there is a strong evidence base suggesting that the majority of cases of PTSD occur within the first 6 months after a traumatic event. In a minority of cases PTSD may be diagnosed for the first time after six or more months after the index event; this is termed delayed onset PTSD. There is considerable controversy persists regarding the diagnostic validity of delayed onset PTSD with some experts suggesting that delays in diagnosis result from delayed presentation rather than true delayed onset. In some cases, delay in onset may result from personnel who have experienced traumatic events remaining in a high state of alert and camaraderie after traumatic events. Such situations may be found among military personnel who experience traumatic events on a prolonged deployment and only find that they begin to think about their traumatic experiences once they have returned home or in communities who experience disaster but are initially highly supportive and functional as they strive to survive in the aftermath and/or as they reestablish a sense of normality. In these two examples it is quite understandable that proper processing of the traumatic event itself may be delayed, and thus so consequentially may the onset of the disorder.

Another important skill which military and disaster psychiatrists need to demonstrate is an understanding of disabling distress reactions that may occur in response to less significant trauma such as the unexpected loss of a key interpersonal relationship or loss of a job. Individuals, who experience significant but non-traumatic events, can present clinically within patterns of symptoms that are similar to those described by PTSD or ASD. However, the correct diagnosis in these cases is an adjustment disorder, which is characterized by maladaptive behavioral and/or emotional responses to a diverse array of stressors. Other psychological conditions may also follow exposure to intense stressors such as conversion disorder that may be diagnosed when someone develops unexplained symptoms or deficits affecting voluntary motor or sensory function (e.g., sudden paralysis of the 'trigger' finger in troops) without demonstrable neuroanatomical injury. Bereavement is a normal grief reaction after the death of someone who is valued or loved, may also occur in response to losses incurred during war or disaster. Other distress responses to war or disaster include anxiety and depressive syndromes, and antisocial behaviors that are not themselves related to frank mental disorders (such as acts of retributional violence or criminal behavior that may amount to war-related atrocities). Alterations in health-related behaviors (e.g., misuse of tobacco, drugs, or alcohol and risk-taking behaviors) may also develop after exposure to disaster or war. Taking a careful history including gaining a detailed understanding of pre-incident normal behaviors and attitudes as identifying relevant risk factors (e.g., history of childhood adversity and prior history of mental illness) as well as details of the traumatic event(s) and post-event reactions will allow military and disaster psychiatrists to decide upon the correct diagnostic label and treatment plan.

Battle Fatigue and Combat Stress

The terms 'battle fatigue' or 'combat stress' are broad and encompass a variety of responses to operational stressors but are not specific diagnostic constellation of symptoms such as major depressive disorder or PTSD. A wide range of physical and emotional symptoms and signs can occur among individuals with either condition including physical symptoms gastrointestinal distress, tremulousness, anxiety, perceptual disturbance, a sense of unreality, and a dazed look (i.e., 'thousand-yard stare'). Battle fatigue may be used to describe combatants who have exhausted physiological and psychosocial coping mechanisms with the intense combat experience whereas combat stress may be used to describe a more classic ASD reaction in the aftermath of particularly intense periods of fighting. It is noteworthy that minor injuries, infections, starvation, heat exhaustion, and cold injuries may decrease the coping resources of a combatant and predispose them to develop formal mental health difficulties.

Medically Unexplained Physical Symptoms

War historians have observed that medically unexplained physical symptoms (MUPS) are common sequelae of combat since at least the US Civil War. Syndromes such as 'soldier's heart' and illnesses characterized by physical symptoms attributed (by sufferers) to war-related exposure to Agent Orange or the pesticides and medical countermeasures as in the case of Gulf War syndrome are examples. Contentious debates between scientists, clinicians, veterans and their advocates, and journalists persist around putative etiology. Some argue that the consistent appearance of these syndromes after war speaks to the likelihood that psychosocial factors contribute to their etiology and importantly the particular role of speculation (by the media and politicians) in the absence of facts to rebut hearsay (Tarn et al., 2008).

Other Psychiatric Illnesses

While not confined to war or disasters, depression, anxiety disorders, and personality changes have all been associated with exposure to trauma. These psychiatric disorders may be accompanied by somatic complaints and may be related to either a traumatic event itself or the associated consequences (e.g., loss of lifestyle or of loves ones).

Risk and Protective Factors

Broadly speaking risk factors for the development of PTSD and other mental health disorders after traumatic incidents can be

divided into pre-incident factors, the nature of the incident itself, and factors relating to the post-event environment. The most influential are the post-incident factors as described in the following paragraphs (Brewin et al., 2000). Pre-incident factors include gender (women being diagnosed with PTSD more frequently than men), low educational status, childhood adversity, prior history of mental health problems, and/or socioeconomic disadvantage. There are many potential explanations for the existence of these risk factors including psychosocial factors, coping styles, cultural factors, and neurobiological explanations. It is also notable that individuals who have successfully negotiated past traumatic experiences may be resilient (sometimes described as 'hardened') in similar future situations. However, if past traumatic events resulted in PTSD, or other psychiatric distress syndromes, subsequent traumatic exposures may make future episodes of these disorders more likely.

Factors related to the incident itself include number and frequency of traumatic events, the nature of the event(s) (e.g., possible use of torture), moral challenges faced (did the person survive because others perished), and the physical/practical consequences of the event (e.g., physical injuries and loss of status, possessions, or loved ones).

The most influential factors in terms of longer-term post-incident mental health are the post-incident ones. Specifically, there is a wealth of good evidence that (1) having access to good social support (provided by people that the individual trusts and who honestly care for them) and (2) experiencing a low (but not no) pressure environment during recovery are the strongest determinants of good post-incident mental health. The understanding of these two factors and their effect on mental health are critical to influencing post-combat, or post-disaster mental health, and mental health care professionals who mistakenly intervene professionally without taking account of these two factors may inadvertently cause harm. There is considerable evidence that being exposed to post-incident mental health counseling (often termed 'debriefing') is, in the main, associated with no long-term benefit, but in some cases leads to a poorer outcome than not being 'psychologically debriefed' at all.

The protective nature of social cohesion is often most evident in defined groups such as the military or emergency services where strong camaraderie and leadership are commonplace. Both are powerfully protective in terms of post-event mental health status. Within military units, any mental health intervention has to take account of group dynamics and to a less extent the same is true for any population affected by disasters. For instance in the aftermath of a train crash, provision of treatment for someone who has developed PTSD without addressing the role they played during the event is unlikely to be successful. Within military units cohesion and leadership is often forged during preparatory training regimes and cemented during combat itself. The same is true for emergency responders during the aftermath of any traumatic event.

Similarly, although possibly more fragile bonds may form within previously non-coherent communities during and after disasters. After disasters the impact of statutory agencies in fostering a sense of security and safety, and preventing further exposure to stressors is critical. Symptoms in civilian victims of war or in the aftermath of disaster may be mitigated or exacerbated by perceptions of community leadership's preparedness for disaster, response to crisis, recognition of 'heroes,' and provision of medical, financial, or emotional assistance both immediately after crisis, and over time. Often a range of NGOs (e.g., the Red Cross or Red Crescent) may also provide input such as attending to basic human needs such as food, clothing, and shelter thereby reducing both the psychological and the physiological effects of the event. However, poorly coordinated post-disaster responses, can lead to confusion and anger in disaster affected communities with consequential effects on their mental health.

Management and Care Delivery

Often disasters or military conflicts shatter the expectation of a world as a just and safe place. Ideally primary preventative measures such as providing professionals with the right training, including importantly training managers/leaders about the impact of their behaviors on the mental health of team members will be in place before an incident occurs. Additionally most organizations that operate in high-risk situations will have selected their employees carefully, although it is notable that formal pre-employment or pre-deployment mental health screening has not been shown to be effective.

Once an incident has occurred interventions vary with the stage of the disaster. Initially, establishing a safe environment and managing life-threatening injury and disease possibilities, such as those resulting from infection or absence of potable water, can be the most important psychiatric interventions. Subsequently, identifying higher-risk populations such as disaster workers, firefighters, police, persons at impact zones, and children can focus intervention strategies. Post-incident mental health screening, a secondary prevention measure, may be effective in some populations (e.g., people who appear to be in difficulty after disasters or who consult a health care professional asking for assistance), but has not been shown to be effective in military populations probably because of concerns of personnel about the consequences of scoring positive in terms of their future career and the way they are viewed by peers and bosses.

Community outreach programs may also be important since disaster victims rarely seek mental health care. The physically injured are also at great risk for psychiatric disturbance. Educating medical and community groups about normal responses to traumatic events as well as when mental health referral is indicated as an important part of outreach programs. Providing information for families and work colleagues may also be important keeping in mind that the aim is to encourage them to provide good support without causing them to become unduly worried.

Military Mental Health Care

Military forces generally have in-house mental health provision that aims to provide both preventative intervention before operational deployments and evidence-based care for individuals who develop formal mental health disorders.

However, the primary aim of most military mental health care provision is to keep personnel psychologically fit, or return them to psychological fitness, so they can deploy to high-threat environments including combat and peacekeeping missions. Military mental health care provision may take place in home-base locations or in a deployed environment, although there is scant evidence about the effectiveness of treatment delivered in such conditions. Military health care professionals have to decide if personnel are ready for deployment where they may be required to protect colleagues or carry out other safety critical tasks. If deployed when not fully fit, poor concentration or impaired judgment, both characteristics of people who suffer with mental health problems, can lead to serious consequences for the individual and their colleagues.

Organizational Management of Traumatic Stress

Military forces, and other disaster responders, have a duty to provide appropriate care and support to their staff. In general the mental health of trauma-exposed personnel is more a function of good organizational preparation and camaraderie/leadership than the availability, or not, of mental health professionals. Organizational preparedness can include training for employees and managers to foster group cohesion (e.g., peer support training such as Trauma Risk Management (TRiM) developed in the UK (Greenberg et al., 2010)) or a variety of resilience training packages some of which (e.g., mindfulness-based approaches) have shown promise although well-conducted randomized controlled trials of these training packages is lacking. There have been various efforts to routinely screen troops after deployment in order to advise those with possible mental health disorders to seek help, but to date there is no evidence that these, costly processes, have been effective in spite of their widespread use (Milliken et al., 2007).

It is also noteworthy that many organizations who have a reputation for working in high threat locations have also been described as having a 'stiff upper-lip,' a term that infers that individuals within these organizations are reluctant to either acknowledge suffering from mental health problem or in seeking help for them. Research has however shown that in reality, a reluctance to seek help is common within all elements of society and not restricted to specific trades or groupings. However, help seeking is likely to be strongly influenced by the attitudes of the families of those affected by traumatic events and thus mental health professionals should take account of family views when formulating a management plan.

Possible Technological Advances

The modern battlefield is both rather similar and rather different to those of old. While troops are still, often, required to live in austere conditions and contend with threats to their health, and indeed live on occasion, more recent conflicts have been marked by the development of better communication with home than ever before. While there be many putative benefits of easy access to support from friends and family located far away, improvements in communication exposes deployed troops home stressors in, or nearly in, 'real time' that presents specific psychological difficulties. For instance, trying to talk to a child, who is missing their father, while he is being subjected to incoming mortar rounds can be challenging to say the least. However, disasters also may interrupt communications leading to distress. For instance, research carried out after the London bombings of 2005 revealed that distress was significantly more likely if respondents had been unable to contact loved ones around the time of the bombings in order to be able to verify their safety (Rubin et al., 2005).

Future wars will see troops having to deal with increasing technological advances in hardware such as the use of advanced sensors, weapons, and battlefield communication systems that require greater degrees of cognitive engagement than has been the case previously. Furthermore, the evolution of highly mobile units on widely disbursed battlefields will decrease the opportunity for exchanging rested troops from the rear area for those exhausted by frontline combat. Provision of brief respite for exhausted troops – a hallmark of management of battle fatigue – may become impossible as each individual may be performing a critical specialized task. Advanced technology will have similar implications for those responding to human-made disasters such as terrorist attacks especially as terrorists gain increased access to so-called weapons of mass destruction (chemical and biological agents). Clarifying the roles of military and civilian responders in terms of triage, treatment, consultation, and education in any joint response to crisis is another challenge for military and disaster psychiatrists.

Conclusions

Becoming involved with conflicts, including conflicts, or natural and human-made disasters can lead to deterioration in an individual's state of mental health. Much of the effect come from the result of disruption of individual's usual social ties and the pressures they have to contend with as they try to reestablish a sense of equilibrium post trauma. Whatever the cause, disasters and military operations leave in their wake populations experiencing psychological disturbances that have been well described by social scientists, civil leaders, physicians, and other care providers throughout the ages. Most people will recover from these difficulties without the need for formal interventions; however, a small proportion will require, or at least benefit from, evidence-based care. In most cases, however, those who do go on to develop mental health problems will either not recognize they are unwell or will be reluctant to seek help if they do.

The consequences of exposure to disaster and war may take the form of psychological disorders such as PTSD or may manifest as various (and sometimes more subtle) forms of behavioral change, anxiety, or depression. Symptoms may present at different times during and after traumatic exposure although the majority of post-event reactions come on soon after the event itself. Many factors complicate the evaluation and treatment of mental health disorders in the aftermath of war or disaster including resources being overwhelmed, life-

threatening physical illnesses or injuries requiring immediate treatment, and that psychological casualties are often reluctant to seek assistance.

In spite of the difficulties that military and disaster psychiatrists face, considerable progress has been made over recent years. We know that improving social bonds, providing psychological and physical safety, and adopting a watchful waiting approach are all likely to mitigate the effects of exposure to traumatic events. Furthermore, there is a range of evidence-based treatment approaches that are primarily based on the provision of focused, time-limited psychotherapies that can help individuals who develop mental health disorders after war or disaster affect a successful recovery. The value of multidisciplinary preparation and training for disaster management and the need for post-incident welfare and support, rather than counseling or therapy-based, outreach programs have also been demonstrated. Future work will hopefully identify how to bolster individual and group resilience and identify novel, or more rapidly delivered, therapeutic techniques, to improve recovery of those who become formally ill. Given that wars and disasters are likely to be a feature of our societies for many years ahead, military and disaster mental health care delivery will need to remain evidence based but flexibly delivered to provide the highest standard of prevention, evaluation, and care for individuals and groups devastated by war or disaster.

See also: Mass Trauma: Psychopathological Effects across the Life Span; Natural Disasters: Health-Related Aspects; Personality Theory and Psychopathology; Post-Traumatic Stress Disorder; Post-Traumatic Stress in Social Work.

Bibliography

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Publishing, Arlington, VA.

Brewin, C.R., Andrews, B., Valentine, J.D., October 2000. Meta-analysis of risk factors for posttraumatic stress disorder in trauma-exposed adults. Journal of Consulting & Clinical Psychology 68 (5), 748–766.

Greenberg, N., Langston, V., Everitt, B., Iversen, A., Fear, N.T., Jones, N., Wessely, S., August 2010. A cluster randomized controlled trial to determine the efficacy of Trauma Risk Management (TRiM) in a military population. Journal of Traumatic Stress 23 (4), 430–436.

Jones, N., Seddon, R., Fear, N.T., McAllister, P., Wessely, S., Greenberg, N., 2012. Leadership, cohesion, morale, and the mental health of UK Armed Forces in Afghanistan. Psychiatry 75 (1), 49–59.

Milliken, C.S., Auchterlonie, J.L., Hoge, C.W., 14 November 2007. Longitudinal assessment of mental health problems among active and reserve component soldiers returning from the Iraq war. Journal of the American Medical Association 298 (18), 2141–2148.

National Institute for Health and Clinical Excellence, 2005. Post-traumatic Stress Disorder (PTSD): The Management of PTSD in Adults and Children in Primary and Secondary Care. NICE, London. Available from: www.nice.org.uk.

The Organisational Management of Traumatic Stress, 2014. The UK Psychological Trauma Society. http://www.ukpts.co.uk/site/assets/UKPTS-Guidance-Document-120614.pdf.

Osório, C., Jones, N., Fertout, M., Greenberg, N., August 2013. Changes in stigma and barriers to care over time in UK Armed forces deployed to Afghanistan and Iraq between 2008 and 2011. Military Medicine 178 (8), 846–853.

Rubin, G.J., Brewin, C.R., Greenberg, N., Simpson, J., Wessely, S., 2005. Psychological and behavioural reactions to the bombings in London on 7 July 2005: cross sectional survey of a representative sample of Londoners. British Medical Journal 331, 606.

Skogstad, M., Skorstad, M., Lie, A., Conradi, H.S., Heir, T., Weisæth, L., 2013. Work-related post-traumatic stress disorder. Occupational Medicine (London) 63 (3), 175–182.

Tarn, M., Greenberg, N., Wessely, S., 2008. Gulf War syndrome – has it gone away? Advances in Psychiatric Treatment 14, 414–422.

Military Geography

Rachel Woodward, Newcastle University, Newcastle upon Tyne, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.B. Garver, volume 14, pp. 9859–9863, © 2001, Elsevier Ltd.

Abstract

Military geography aims to understand how military activities, organizations, institutions, and effects are geographically constituted and expressed. Military geography is broad in scope, and includes analyses of military land use, including and beyond battlefields and the environmental impacts of military activities; explorations of representations and interpretations of military landscapes; investigations of the economic and social relations effected by military capabilities; and the life worlds of military personnel. Future areas for military geographies research include the ways in which the changing nature of warfare produces new geographies of war, and the ways in which militarism as a social process continually reworks its spatialities.

Military geography can be defined in contemporary social sciences as the geographical constitution and expression of militarism, military activities, and militarization. Because these processes are multiple, with a huge variety of effects and manifestations, it is often more appropriate to talk of a plurality of military geographies. Unifying this field is an attentiveness to the ways in which places, spaces, environments, and landscapes are shaped by, and in turn have an effect upon, military capacities and military power.

This definition of military geography, dominant in the contemporary period, can be contrasted with an older definition of military geography as the application of geographical tools and techniques to military problems (Palka and Galgano, 2000). This more functional definition, that emerged with the discipline of geography itself, understands geographical scholarship in terms of its utility for the application of military power. More critical approaches toward military geography see the role of scholarship, rather, as understanding the effects of military capabilities and understandings, enabling a more sustained examination of how military activities and militarism operate as geographical processes, and how in turn geography shapes the military.

History of Military Geography

The history of the discipline of geography since the early nineteenth century shows geography's intimate connection to the pursuit of military power, control, and domination. With the development of colonial and later imperial ambitions among European powers, particularly Britain, and with the expansion of the territorial ambitions of the United States, we see the emergence of attempts by nation-states not only expand economic influence overseas, but also to undertake such ambitions through the use of military force. Geographical knowledge, of terrain, climate, biogeography, population, and settlement patterns, was an indispensable prerequisite for colonial and imperial expansion. In the United Kingdom, this is evident in the close association between military institutions and personnel and the establishment of the Royal Geographical Society. In the United States, this is evident in the establishment of military geography on the syllabus for trainee officer cadets at the US military academy at West Point. The understanding of military geography as assistive to military activities and ambitions stems from this period.

During the first half of the twentieth century, and specifically during the World Wars I and II, the labor power and skills of geographers were conscripted for the war efforts of the key protagonists in both conflicts. Recent research has established the centrality of the contribution of geographers and geographical knowledge to, for example, intelligence and information work, cartography, and surveying. Geography, conceived in this period as essentially descriptive of physical topography, environment, and patterns of human settlement and culture, was essential to the war effort specifically because the pursuit of strategic military objectives such as invasion and counterinvasion required, during both planning and execution, thorough knowledge of terrain, and a good understanding of local populations. We see the continuity of the idea of geography as a discipline of utility to the pursuit of military goals.

In the postwar period, three trends are evident, and are foundational to our understanding of what military geography might be in the present. First, we see the continuity of engagement between geography as a discipline informing military strategies and planning. This is evident, for example, in the adoption of emergent ideas about settlement distribution and spatiality in geographical science to urban planning anticipatory of the possible effects of nuclear annihilation during the Cold War. Urban morphologies at this time reflect ideas about the capacities for the maintenance of the nation-state under nuclear attack, particularly ideas about settlement dispersal (Farish, 2003). Second, more broadly across the social sciences and of particularly utility to geography, what has been termed the 'quantitative revolution' – the emergence of sophisticated techniques for statistical analysis of numeric data and the adoption of these techniques as a core element of geographical analysis and explanation – has shaped the nature of geographical inquiry. The symbiosis between the development of quantitative analysis and geographical information systems, and the demands of military strategists for techniques to model and thus plan and predict the effects of military activities, shows continuity (evident to the present) in the use of geographical techniques for the solution of military planning problems. Third, from the late 1960s onwards, the critiques

of patterns of power, domination, and control evident in advanced capitalist economies, by geographers energized by the ideas originating in Marxist thought, emerge as a radical geographical approach to understanding the spatialities of social life. Although there are indications across the later twentieth century of the recognition of the significance of military power as part of a wider analysis of geographies shaped by state power in conjunction with the demands of capital, the attention of radical geography to military forces and military activities as a focus for explicit scrutiny emerged only slowly.

However, toward the end of the twentieth century and into the present, contemporary geographical scholarship has turned to military topics as a more central focus of study. Political geography, reinvigorated by the ideas of those writing a critical geopolitics to explore the representational and discursive portrayals of states and strategists, has engaged more directly with the organization, execution, and effects of military power (Dalby, 2010). Military forces, activities, and effects themselves have invited explicit focus in terms of their transformational potential for shaping geographies (Woodward, 2004; Graham, 2010). Long-running armed conflicts in Central Africa and Southeast Asia, and the wars in Iraq and Afghanistan in the first decade of the twenty-first century, have both prompted a wealth of analytic investigation by geographers whose critical and condemnatory overview of the deployment of military power and its catastrophic effects are prompting a renewed consideration of just why geography as an academic discipline needs to be attentive to militarism, military activities, and war.

The Scope of Contemporary Military Geography

The scope of contemporary military geography is broad in terms of topics for investigation and analytic approaches to those topics. It is also notable that those working within military geography draw extensively on the conceptual and theoretical insights of other areas within the subdiscipline, such as political, economic, social, and cultural geography, and more widely across the social sciences. While military geography has at its heart a concern with the geographical constitution and expression of military activities and effects, it is in its contemporary form clearly allied conceptually and politically to a broader movement within sociology, anthropology, women's studies, political science, and international relations which seeks to establish not only the patterns and consequences of military power, but also – and crucially – the conditions which make the deployment of military force possible in the first instance. Thus we can see a divergence between more traditional military studies which see military phenomena as a given, functional category, and a more critical military studies approach which sees military processes and practices as the outcome of social life and political contestation.

Terrain, Tactics, and Land Use

A core component of contemporary military geography is the examination of the geographies of military activities in terms of the interrelationship between terrain and military tactics, and in terms of wider questions of territorial control and land use in nonconflict and domestic as well as conflict and postconflict settings. The analysis of both terrain, in terms of geomorphological conditions and effects, and of factors such as meteorological conditions pertaining in particular times and places and their influence on specific military encounters, has been a source of fascination and interest to many seeking to explore historical campaigns and battles (Doyle and Bennett, 2002). Technological developments in surveying equipment and mapping, including LiDAR and other remote sensing techniques have enabled geographers to track and trace the evidence for past incidents across lost battlefields, and to map in fine detail the course of specific tactical encounters between military forces. An emergent subdiscipline of combat archaeology draws on geographical and archaeological exploratory techniques and environmental reconstruction methods to establish in more detail local biogeographical conditions in sites of both conflict and of longer-standing military occupation. In terms of land use and military presence, and reflecting a long-standing interest among geographers not just in patterns of human activities but, crucially, conceptual explanations for the spatial constitution and expression of military activities, interests range from the exploration of the political economies of military basing, through to the representation and reimagination of regions and large areas of territory by military planners and strategists and the wider politics of such constructions (Farish, 2010). Military geography is also exploring the spatialities of military airspaces (Williams, 2011).

Environmental Impacts of Military Activities

Military activities, across a range from the pursuit of armed combat through to domestic training and basing activities, are enormously detrimental to the natural environment. Military geography, as a subdiscipline interested in the totality of military activities and the range of potential geographical effects, therefore seeks to account for the environmental activities, and to explore in addition the strategies deployed by governments and their military forces in explaining and accounting for such impacts. Military activities impact on the natural environment in the first instance through the sheer scale of consumption of natural resources, required for the execution of military power; resources include fuels, precious metals, ferrous metals, and hydrocarbons for the manufacture of weapons systems including munitions, and the provision of energy for their deployment. The conduct of armed conflict is damaging to environments through the physical modification and destruction of habitats and local ecologies; through the pollution effects of chemical, biological, and nuclear contaminants released through military operations; and through the residue of ferrous, heavy metal, and chemical pollution on sites of military engagement. Nonconflict activities, which are essential for the preparation of armed forces to the readiness states required by their national governments, are similarly destructive through the modifications imposed by basing, logistics, communications, and training infrastructures (including the requirement for training areas by land forces), by the erosion and soil compaction effects of tracked vehicles, by pollution from the leakage of fuels and lubricants, by the litter of used practice munitions, by noise effects of military training, and by human interference and presence in specific environments. Of concern to military geography too are the

rationalizations produced to manage and explain such effects. For example, former military sites affected by unmanageable and unremediable contamination may be declared as sacrificed with future uses impossible due to contamination (Havlick, 2011). Of interest too are the policy mechanisms established for the remediation of former military lands with levels of contamination low enough to allow land reuse following restorative work. Interest in these policy mechanisms extends to the discursive and representational practices deployed by military forces and state interests in explaining to a wider public both the effects of military activities on the environment, and the rationale for allowing negative impacts where it is thought that national military interests prevail.

Representing and Reading Military Landscapes

In keeping with a long-standing core interest in human geography in the identification, reading and engagements with landscape, military geography seeks to explore how military landscapes work (Woodward, 2014). Approaches include considering how military forces themselves read landscapes for military purposes, from the scale of the individual and personal with the activities of personnel on the ground assessing terrain for tactical reasons, through to the large-scale surveying of landscapes for more strategic military maneuvers. Practices of landscape interpretation owe much to geographical skills in reading terrain and establishing viewpoints, and here again we see the coevolution of military techniques alongside those of geography as a discipline. This is evident, for example, in the development of cartographic practices and aerial surveillance techniques during and after the World Wars I and II, and in more contemporary examples where military personnel seek new methods for thinking through the spatialities of urban landscapes for military purposes, including the control and domination of resident insurgent populations. Military landscapes and landscapes of militarism also include the settlements and infrastructures associated with military basing, both in domestic and overseas contexts (Lutz, 2001). Of particular interest are the transposition and installation of urban morphologies adapted for military requirements by US forces in their domestic setting, to overseas garrison towns and territories under military control (Gillem, 2007). Landscape interpretation is also a representational strategy, and as public organizations funded by national governments, many national military forces are increasingly concerned to provide a narrative about the appropriate reading of landscapes as a means of legitimating a military presence. So, for example, training areas which might otherwise be thought of in terms of their landscape qualities come to be portrayed not only as constitutive of the value of a nation's territory, but as significantly military as part of that value.

Reading and engaging with military landscapes is also a civilian practice, and nonmilitary understandings and interpretations of military landscapes have proven a fruitful area for military geographical inquiry. One of the most immediately obvious and dramatic types of military landscape, prompting a wealth of inquiry, is memorial landscapes around past military activities. Studies of, for example, the landscape effects of memorials and cemeteries associated with the World War I in northern France and Belgium, explore the intersections between ideas of national identity and national sacrifice with more individual and personal needs for mourning and recognition of loss. There is also a substantial interest in the geographies of postmilitary landscapes. Across European and North American contexts, for example, successive rounds of investment in military capabilities, superseded at regular intervals through changing geopolitical power configurations, technological and developments in weapons and communications systems, and national defense strategies, have generated landscapes of ruin and redundancy. These are of interest not just for their exposure to civilian gaze in the aftermath of (usually fairly secretive) military use, but also for what they tell us about social priorities for the care and protection, or redundancy and destruction, of symbols of past military priorities and histories into the future. Visual representations of military landscapes, including postmilitary landscapes as well as responses to conflict and its effects, are of increasing interest within military geography because of the insights which artistic responses generate as part of a wider conversation about militarism and its spatial and landscape constitution and expression (Ingram, 2011). Military geography is also concerned with the phenomenological and post-phenomenological experience of being in landscape, following conceptual developments in wider landscape studies, and so the significance of the personnel, embodied, experiential encounter with militarized space is also of relevance here (Sidaway, 2009).

The Geographies of the Military Economy

Military forces and their activities are recognized as significant economic actors, and drawing on the insights of economic geography, military geography seeks to understand how military effects on economies and labor markets are played out at a range of different scales, and how they intersect with other economic processes. The changing nature of the defense industry, particularly the consolidation of manufacturing and research and development capabilities, has had profound local and regional effects in terms of concentrations, consolidation, and decommissioning of defense industry activities with knock-on effects for labor markets. There are also patterns of global economic change brought about by the transnational nature of an industry increasingly footloose from national and supranational political control. Military economic geography is interested in patterns of basing and deployment of personnel, and in patterns of consumption and expenditure consolidated within particular localities domestically and overseas as they follow personnel deployments. There are also economies associated with the provisioning of military forces, including questions about the embeddededness or otherwise of forces within local economies as a consequence of military supply chains for basic commodities.

Reflecting a trend underway from the end of the Cold War,-and gathering pace under neoliberal regimes of governance and economic management by nation-states, the privatization and outsourcing of military activities is an increasingly significant issue for military forces, and correspondingly for military geography. This has led to some arguing for a reconceptualization of the geopolitical as the geoeconomic social, reflecting the potential scale (according to proponents of

this argument) of the shift in the deployment of subcontractors for an increasing array of military functions (Cowen and Smith, 2009). The range of outsourced and subcontracted activities includes the use of personnel for both officially recognized personal protection (and there are indications of its use for more clandestine operations where visible deployment of regular armed forces would be politically problematic), the subcontracting of a range of logistics and supplies functions, including catering, health, welfare, and administrative services, the contracting-out of training functions involving the providers of weapons and transportation systems, and the outsourcing of land management and training area maintenance functions in some cases. It should be emphasized that these are all tasks which 40 years ago would have been conducted in-house by regular armed forces personnel. The consequences of this shift toward privatization and outsourcing are various. Benefits argued for by proponents of this shift, and recognized even by some critics, include the delegation of activities best undertaken by those with the requisite skills and training but who do not require a military role to undertake the task. More problematic are questions about accountability and viability of the subcontracting of activities around the deployment of armed violence beyond the state's nominated and legitimized representatives (Gallaher, 2012). There is clearly a geography to this pattern of subcontracting, and military geography is increasingly alert to the complex political economy of privatization of military forces, and the spatialities and scalar effects of modes of neoliberal management in the organization of military activities.

Military Personnel

Reflecting the engagement of human geography with matters of social and individual identity, subjectivity, and embodiment, and with the exploration of issues such as class, ethnicity, gender, and sexuality as organizing features of social relations, military geography is similarly interested in military personnel as people, and more collectively. Military geography has explored the social construction of military identities with reference to the spatialities inherent in that process through, for example, explorations of the intersections between military masculinities and space, with reference to different spatial and temporal contexts and in locations as diverse as the battlefield, the parade ground, the training area, and the barracks. The use of ideas about gender, gender roles and gender relations within military units, and specifically the spatialities through which these relationships are constituted and expressed informs broader debates about women's military participation. The effects of hegemonic ideas about military masculinities and their enactment through social space similarly inform broader social scientific debates on gender identities. Geography, as a discipline alert to the significance of the visual in representational strategies, seeks to explore such issues with reference to the imagery used to consolidate dominant ideas about gendered identities within military forces. The construction and articulation of representations of the figure of the soldier, and the variety in these across different national contexts, also figures in explorations focused on the broader social understanding of the figure of the military operative and his or her place in the world, as part of wider explorations of civil-military relations.

Most military geography is concerned with the institutions and personnel formed and deployed with the authority of the nation-state for the exercise of potentially lethal yet state-sanctioned violence. However, not all militarized violence is conducted by state military forces, and military geography is alert too to paramilitary, insurgent and revolutionary militarisms which are so often emergent in response to state military violence. Examining nonstate military actors raises particularly interesting questions for military geography, precisely because of the contingencies of nonstate funding in shaping military objectives and activities, the deployment of nonstate military personnel, and the specific social position of the nonstate operative, revolutionary, paramilitary, or insurgent.

Future Directions in Military Geography

Military geography is essentially a set of engagements with a range of practices which reflect often very rapid temporal and spatial change. Developments in nation-states' strategic interests and internationally collaborative ambitions, combined with evolutions in military capabilities and policies, and shifting geopolitical realities at a range of scales, means that the subdiscipline of military geography is marked by its dynamism. Indications of potential future avenues of exploration for military geographies reflect such changes.

The first of these concerns the changing nature of warfare and of ways of engaging with armed conflict by state and nonstate actors. Although (as already discussed) there are indications that much of this agenda is starting to be addressed, it is necessary to emphasize the significance of current changes in the nature of war for future military geographical inquiry. Key issues here include the increasing urbanization of warfare as a deliberate, strategic choice by many national military forces, and the consequences of this for the use of geographical techniques of surveying, particularly regarding their use for purposes far removed from the intentions of those responsible for their development. The noted twentieth-century shift to war as a civilian issue, rather than a purely military affair, continues apace, and armed conflict is increasingly utterly all-encompassing for those caught up in it directly and along its margins. The use of preemptive force has tremendous consequences for civilian populations. We can also note the growing interest in environmental security and resource wars as foundational to twenty-first century conflicts.

The second area for future research which is highly relevant to military geography speaks to cultural geography's longstanding interest in the construction, articulation, and reproduction of meaning with reference to space and social life. Militarism, often understood as the extension of military priorities into civilian life, is among other things a cultural practice, and military geography is well placed, with its focus on the spatial constitution and expression of military phenomena, to make critical interventions as analysts grapple with the circulation of ideas about military power and priorities in civilian sociocultural life.

See also: Geopolitics; Militarism; Military History; Military, War, and Politics; Political Geography; Social Geography; Space and Social Theory in Geography.

Bibliography

Cowen, D., Smith, N., 2009. After geopolitics? From the geopolitical social to geo-economics. Antipode 41 (1), 22–48.

Dalby, S., 2010. Recontextualising violence, power and nature: the next twenty years of critical geopolitics? Political Geography 29, 280–288.

Doyle, P., Bennett, M. (Eds.), 2002. Fields of Battle: Terrain in Military History. Kluwer, Amsterdam.

Farish, M., 2003. Disaster and decentralization: American cities and the Cold War. Cultural Geographies 10, 125–148.

Farish, M., 2010. The Contours of America's Cold War. University of Minnesota Press, Minneapolis.

Gallaher, C., 2012. Risk and private military work. Antipode 44 (3), 783–805.

Gillem, M., 2007. America Town: Building the Outposts of Empire. University of Minnesota Press, Minneapolis.

Graham, S., 2010. Cities under Siege: The New Military Urbanism. Verso, London.

Havlick, D., 2011. Disarming nature: converting military lands to wildlife refuges. The Geographical Review 101 (2), 183–200.

Ingram, A., 2011. Making geopolitics otherwise: artistic interventions in global political space. The Geographical Journal 177 (30), 218–222.

Lutz, C., 2001. Homefront: A Military City and the American Twentieth Century. Beacon Press, Boston.

Palka, E.J., Galgano, F.A. (Eds.), 2000. The Scope of Military Geography: Across the Spectrum from Peacetime to War. McGraw-Hill, New York.

Sidaway, J.D., 2009. Shadows on the path: negotiating geopolitics on an urban section of Britain's South West Coast Path. Environment and Planning D: Society and Space 27, 1091–1116.

Williams, A., 2011. Reconceptualising spaces of the air: performing the multiple spatialities of UK military airspaces. Transactions of the Institute of British Geographers 36, 253–267.

Woodward, R., 2004. Military Geographies. Oxford, Blackwell.

Woodward, R., 2014. Military landscapes: agendas and approaches for future research. Progress in Human Geography 38 (1), 40–61.

Military History

Hew FA Strachan, University of Glasgow, Glasgow, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9863–9868, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Military history is the history of wars, and of armed forces in peace as well as in war. Its separation from history, more broadly defined, arises from its didactic purpose. Its origins therefore lie with the development of professional armies. The growth of history as a university subject from the late nineteenth century did mean that academics as well as soldiers studied the history of war, but its institutional foundations were shallow. Moreover, soldiers rather than academics wrote the official histories of World War I. The situation only changed with the adoption of broader definitions for the official histories of World War II. The world wars themselves showed the absurdity of approaching the study of modern history without taking military history into account. Since the 1960s academic military history has thrived.

Military history is the history of wars, and of armed forces in peace as well as in war. Its separation from history, more broadly defined, arises from its didactic purpose. Its origins therefore lie with the development of professional armies. The growth of history as a university subject from the late nineteenth century did mean that academics as well as soldiers studied the history of war, but its institutional foundations were shallow. Moreover, soldiers rather than academics wrote the official histories of World War I. The situation only changed with the adoption of broader definitions for the official histories of World War II. The world wars themselves showed the absurdity of approaching the study of modern history without taking military history into account. Since the 1960s academic military history has thrived on the back of two influences. The first, 'the new military history,' is largely American in origin and stresses the subject as a component of 'total history.' The second, strategic studies, relates past to present.

The easy definition of military history is that it is the history of wars. And yet this definition is highly imprecise. Wars have social, economic, and political dimensions that have been analyzed more by the historians of those subdisciplines than by military historians. That is not to say that there are no important links to be made between military history and other historical subdisciplines, nor is it to deny that the good military historian endeavors to make those connections. But in terms of their subject matter military historians have been concerned primarily with the histories of armed forces, not only in war but also in peace. Military history has therefore been more comfortable with wars fought by armies and navies than with wars fought between warrior societies, or before soldiering became a distinct profession.

The Emergence of Military History as a Separate Subdiscipline

Thucydides wrote the history of a war, and – like some other ancient historians – himself saw service. But it would not be reasonable to call him a military historian. It required the growth of professional armies and the concomitant influence of the Enlightenment for the ancient historians who wrote about conflict to be treated as military historians rather than as historians *tout court*.

The Influence of Professional Armies

In the 1590s, Maurice of Nassau developed systems of infantry drill and of military organization that standardized tactics and were emulated throughout Europe. This was the basis for what in 1956 Michael Roberts called the 'military revolution' – a clutch of changes that occurred in warfare between 1560 and 1660. The Thirty Years' War in Central Europe, the Fronde in France, and the British civil wars served to disseminate the principles of the 'military revolution,' which found their seventeenth century apogee in the Swedish army of Gustavus Adolphus. Subsequent historians have elongated the chronology of the 'military revolution.' Geoffrey Parker sees its origins as before 1560, in the growth of a new style of artillery-resistant fortification, the *trace italienne*. Others have highlighted developments after 1660, calling the growth in army size under Louis XIV and the maintenance of armies in peace as well as in war a 'second military revolution' (Rogers, 1995). The essential points are that over the course of the period 1500–1800, armies as we now recognize them evolved, and that those armies then became the basis of imitation within Europe and the foundation of empire outside it.

During the course of the eighteenth century the aristocracies of Europe, now subordinated to the crown and state, made soldiering their vocation. Military academies were established, not only for the training of the scientific arms – the artillery and engineers – but also for the cavalry and infantry. The combination of institutional continuity and professionalization promoted the study of war. The writings of the ancients were supplemented by the effusions of eighteenth century practitioners – including Maurice de Saxe, Henry Lloyd, and G.F. von Tempelhoff – who combined theoretical musings with a distillation of their own experiences. Military history had didactic rather than scholarly origins.

The Enlightenment

The study of past wars as a means of improving present practice was, at one level, entirely consistent with the precepts of the Enlightenment. War, like any other human activity, could be seen as a science, based on unchanging principles, themselves derived from actual experience. The wars of Frederick the Great,

a child of the Enlightenment and a major writer on war, served to promote these connections. Over 70 works of military theory were published in the seventeenth century, but more than twice that in the eighteenth century, in particular, over 100 during the period 1756–89 (Gat, 1989).

Napoleon was both a product of this tradition and its most distinguished advocate. William Napier, an unstinting admirer of the emperor, established British military history with his account of the Peninsular War. Although Napier eschewed theory, his history was influenced by the most important military theorist of the nineteenth century, A.H. Jomini. Jomini's *Traité des grandes opérations militaires* (first two volumes, 1804) became a five-volume history of the wars of Frederick the Great and Napoleon, with a resumé of the general principles of war tacked on as a conclusion. His *Précis de l'art de la guerre* (1838), which shaped the syllabuses of military academies into the twentieth century, put the theory first but still relied on history to make its points (Shy, 1986; Alger, 1982).

Carl von Clausewitz was critical of Jomini, but like him, wrote far more history than theory. Moreover, his principal work, the unfinished and posthumously published *Vom Kriege* (1832), relies on history for its evidential base. Clausewitz stressed his anxiety to break with the nostrums of earlier military writers, but he is no different from them in his readiness to cull military history for normative purposes.

At one level, therefore, the influence of the Enlightenment on the development of military history operated internally – it helped define how and by whom war was fought. But its consequences were also external to the subject. The phenomenon of war itself appalled the *philosophes* – not the least Frederick's friend, Voltaire. Their efforts to curb and moderate its effects, stoked by the memory of the Thirty Years' War and its terrors for the civilian population, took shape through international law. War became an activity clearly distinguished from peace, undertaken by specialists separated from civilian society but who, crucially, acted not on their own account but on behalf of the state. For the *philosophes*, war was not necessarily endemic in society. Those who built on the legacy of the Enlightenment, the liberals of Victorian Britain in particular, could see it as an activity that was not honorable but reprehensible, maintained by the aristocracy with the aim of sustaining their own hold on society, and replaceable by alternative forms of interstate competition, particularly trade.

Academic Neglect of Military History

The legacy of liberalism was a belief that military history was not a proper study for universities. As history gained a foothold in the curriculums of European higher education in the last third of the nineteenth century, war ought to have been central to its preoccupations. Indeed, at one level it was. Historians of the ancient or medieval worlds could not neglect war entirely, but they still preferred to focus on more humane developments, on 'progress' in law, religion, or the machinery of state. Equally, no nineteenth century scholar could neglect the impact on Europe of the Napoleonic Wars or the wars of German Unification – as Archibald Allison showed for the former, and Heinrich Friedjung for the latter. But they neglected the conduct of war itself: war was an aberration, inimical to the 'Whig' view of history. The advent of nuclear weapons in 1945, with their threat of a war so awful and so complete that war and peace became absolute terms rather than points on a scale of relative values, completed this marginalization. Diplomatic historians looked at why wars occurred, at how they were ended, and at efforts to avoid them thereafter, but the history of war itself was left to its practitioners.

This picture of academic neglect, still propagated but with much less reason in the twenty-first century, is overdrawn. There were major exceptions, even if it remains true that those exceptions prove the rule.

The founding father of academic military history, Hans Delbrück, served in the Franco-Prussian War and believed that universities should recognize military history. He planned to do his *Habilitation* on a military historical topic, but both Heinrich von Treitschke and Theodor von Mommsen, the leading historians in Berlin, opposed the idea. Delbrück persisted with his enthusiasm and in 1883 announced his intention to write a general history of war in its political context. Leopold von Ranke told him that the study of war was not appropriate to a university, and when Delbrück was appointed to a chair in 1895 (ironically as Treitschke's successor) it was in 'universal and world history.' In 1900, Delbrück published the first of the four volumes of his history of war. Mommsen told the author that he would not have time to read it (Bucholz, 1985; Deist, 1998).

Delbrück also alienated those to whom his work ought to have appealed. He argued that Frederick the Great had not sought decisive battle but had preferred to exhaust his enemies by maneuver. The German army's general staff was incensed. Its own studies of Frederick's campaigns were extensive, but their purposes were more didactic than scholarly. Institutionally, this loyalty to Frederick (rather than to Napoleon) as the founding father of modern war predisposed them to see him as the advocate of a strategy of 'annihilation' rather than of 'attrition' (Bucholz, 1985; Lange, 1995).

At one level, therefore, Delbrück was an isolated figure – virtually unique in the academic world of Wilhelmine Germany, and vilified by the one institution that thought deeply about the history of war. But he was not uninfluential. He had been tutor to the Kaiser, he was a Reichstag deputy, and he edited the *Preussische Jahrbücher*. He emerged from World War I a leading figure in Germany's public and intellectual life. As a role model his legacy spread in two directions. First, his interpretation of military history was essentially Clausewitzian; he studied war as a state activity, and as an agent for the implementation of strategy. Second, he treated war as a discrete phenomenon, possessing an integrity from the ancient world to the modern. Historical knowledge so defined was the basis for pronouncements on current strategic issues.

Delbrück's institutional legacy proved more short lived. Military history was established at the Friedrich Wilhelm University in Berlin after World War I. But the subject was usurped by the Nazis. For a long time after World War II there was no established chair of military history at any German university, with the exception of the post created for the Clausewitz scholar, Werner Hahlweg, at Münster in 1969. Only in the 1990s did the subject reassert itself in research terms, and a professorship was established at Potsdam.

In Britain, also, individual careers punctured the image of academic neglect. The Regius Professor of Modern History at Oxford between 1904 and 1925, Sir Charles Firth, was a historian of the English Civil War and of Cromwell's army. The same university's Chichele Professor of Modern History from 1905, Sir Charles Oman, wrote a massive and definitive history of the Peninsular War. In the same year Oxford created a lectureship in military history, strategy, and tactics, and in 1909 established the Chichele chair in the history of war. Thus in Britain, unlike Germany, the subject acquired an institutional focus. Two years later, Cambridge followed suit with the Vere Harmsworth chair of naval history. King's College London, whose professor of modern history, J.K. Laughton, had been involved in the formation of the Naval Records Society in 1893, formed a department of naval history in 1913 but failed to make an appointment (N.A.M. Rodger in Hattendorf, 1994).

In Germany, the effect of World War I was to institutionalize personal initiative; in Britain, its effect was to undermine the progress already made. Spenser Wilkinson, who had been appointed to the Oxford chair, had been a journalist, and – like Delbrück – had linked the study of military history to current policy issues. But, unlike Delbrück, Wilkinson did not join the public debate on the strategy of World War I. He turned to scholarly studies of the eighteenth century and, his eyesight failing, became a marginal figure. Until the appointment of Sir Michael Howard to the chair in 1977, his successors lacked academic punch (Ernest Swinton), were tarred with the brush of journalism (Cyril Falls), or were unproductive in publication terms (Norman Gibbs). The burden of actual teaching throughout this period fell on C.T. Alkinson, whose output was considerable, but who tended to antiquarianism as well as personal eccentricity. In Cambridge, the Harmsworth chair married naval and imperial history in 1932, to the eventual detriment of the former.

General Staffs and Official Histories

The failure of the universities in Britain to master military history was displayed by the allocation of responsibility for the official histories of World War I. The accounts of naval and military operations were both put in the hands of civilians – Sir Julian Corbett and Sir John Fortescue, respectively. Neither man held an academic post. Corbett was a lawyer by training, but his interest in contemporary maritime doctrine, as well as in naval history, qualified him admirably for the task. Fortescue was writing a *History of the British Army* in what proved ultimately to be 13 volumes, but he derived his principal income from his appointment as Royal Librarian. His enthusiasm for colorful accounts of Napoleonic battles produced a prose style and a conceptual approach unequal to the nature of modern war. He was replaced by a team of soldiers, headed by Sir James Edmonds.

Thus, the bulk of Britain's official history program ended up, by default, being run on lines not dissimilar to those of the other belligerents. In the wake of Prussia's victory in 1871, the major land powers had established or reestablished their general staffs, charged with planning in peacetime and operational direction in wartime. The staffs' study of history flowed from their responsibility for doctrine and training. During the period 1871 to 1914, massive military histories of continuing scholarly importance were published, either by the historical sections of the general staffs (the Germans devoted 19 volumes to the campaigns of Frederick the Great) or by individual officers with staff training. The conceptual basis of current understanding of Napoleonic warfare was established in this period, by French officers such as Jean Colin and Hubert Camon.

These were the bodies and individuals made responsible for the official histories of World War I: the Austro-Hungarian was in the hands of the Bundesministerium für Heereswesen and the Kriegsarchiv, the French in those of the Ministry of War, and the German in those of the Marinearchiv (for naval operations) and the Reichsarchiv (for land operations). Although the latter were nominally independent of the armed forces, in reality the personalities and the agenda showed the thread of continuity.

Thus, the biggest projects in military history ever undertaken were given to soldiers rather than scholars. Their justifications became didactic, and their focus tactical and operational rather than political or economic. Full accounts of campaigns were not matched by comparable analyses of economic mobilization or strategic direction.

After World War II, academics regained some of the ground they had lost, but the effects were partial. The British official history was conceived in much broader terms, and a division of labor between dons and soldiers established, the former being tasked with both the 'civil' and the 'strategic' series. In the USA the history of naval operations was entrusted to S.E. Morison, a Harvard professor. Elsewhere, however, military history narrowly defined, remained largely in the hands of the successors to the general staffs' historical sections. France has *services historiques* for each of its armed forces, and there are few historians of warfare in university posts. With distinguished exceptions, such as the medievalist Philippe Contamine and the modernist Guy Pedroncini, the study of war in French academic life has been dominated by its relationship to social history – either the social composition of armies (André Corvisier has been a pioneer here) or the impact of war on society itself (where Jean-Jacques Becker has been a leader). In Italy, although in the 1960s and 1970s Giorgio Rochat, Massimo Mazzetti, and others began to challenge the domination of the official histories, the armed services' grip on the subject has remained powerful. In Germany, the Militärgeschichtliche Forschungsamt was created in 1957 under the auspices of the Bundeswehr. It has, however, recruited civilian scholars as well as uniformed members, and its publications have maintained rigorous standards that defy the self-serving purposes of the didactic tradition to which it is nominally the heir. Its monumental history of Germany in World War II shows a willingness to confront its armed forces' recent past, which contrasts strongly with the pattern in Japan. Here, the rejection of a nation's militarist inheritance has also resulted in the atrophy of military history.

The Didactic Tradition

Governments and armed forces sponsored the writing of official history because they hoped to derive from it 'lessons' for

future application. But their labors produced accounts so voluminous and so long in production that the lessons were in danger of being lost or forgotten. Into this gap stepped others, preeminently but not only J.F.C. Fuller and Basil Liddell Hart in Britain, who unashamedly used military history as a prescriptive basis for the future. Like Delbrück, they ranged from the ancient world to the present, but unlike him, their focus was tactical rather than strategic. Even Fuller's and Liddell Hart's ostensibly historical works, whose popularity endures, retain a theoretical purpose.

The 'New' Military History

In the short term, therefore, the effect of the two world wars was the reappropriation of military history by the armed forces for purposes that were instructional rather than academic. But in the long term the all-embracing nature of both wars made the marginalization of military history within university circles unsustainable. Once the first half of the twentieth century entered the curriculum, war could no longer be left out.

The idea of total war was used to underpin the idea of total history. War became central, rather than peripheral, not only to the period 1914–45 but to every other period as well. Ancient and medieval history was particularly susceptible to this approach. But its success can best be measured by consideration of the eighteenth century. The 'didactic' tradition had seen this as the era of limited warfare, when armies maneuvered rather than sought battle, and wars were indecisive affairs conducted by professional armies cut off from civil society. Current scholarship sees eighteenth century warfare as involving a much larger cross section of society, and is far readier to apply the vocabulary of modern war to conflicts like the Seven Years' War and the American War of Independence. Their length and geographical extent are seen by some as prefigurings of world war.

The 'new military history,' a term coined in the USA in the 1960s, indicated that the history of war was not simply about 'great captains' and their conduct of operations (Paret, 1992). Military history embraced the study of armies in peacetime, and war's cultural, economic, and social effects; it had a responsibility to explore its links with other subdisciplines in history. The main directions in which the subject moved were toward 'war and society,' both narrowly and broadly defined. The narrow definition meant the analysis of an army's social composition, and also its effect on civil–military relations. The broader definition considered war's effects – economic and cultural – on civil society.

As a campaigning tool the 'new military history' proved remarkably successful, not the least in the USA. In 1954, only 37 of the 493 colleges and universities in the USA offered courses in military history. By 1989 164 of the 554 institutions taught military history, and more than 2% of all historians in America were military historians. This calculation excluded the officer training courses (the ROTC), which accounted for about 400 more military historians. In the mid-1970s in the USA about 100 doctoral theses a year embraced military history; 20 years later the figure had tripled (Don Higginbotham in Charters et al., 1992; Kennedy, 1989).

Military historians established their subject by redefining themselves in interdisciplinary terms: they explained why they were not military historians, or at least not in the traditional sense, but 'total' historians. Although its advocates continue to promulgate the 'new military history,' its point has been made – it is now not so 'new.' So well absorbed have its points become that academic military history is increasingly returning to a narrower definition of its roots. Warfare itself is now a subject fit for self-contained study, and academics working in the field publish in journals devoted exclusively to military history rather than to general history. The argument that war is part of total history no longer has to be made.

The Rediscovery of Clausewitz

One of the reasons why Delbrück was at odds with the general staff was that he and they were not comparing like with like. Delbrück's focus was on war as a political act, as the means to fulfill strategic objectives. The general staff saw war as a self-contained phenomenon. Since 1945, Gordon Craig, Gerhard Ritter, and, above all, Michael Howard and Peter Paret have operated within the Delbrückian interpretation. It is hard to underestimate the intellectual consequences in the anglophone world of Howard's and Paret's translation of Clausewitz's *Vom Kriege*, which appeared in 1976. It relieved Clausewitz of the incubus of obscurity and ambiguity, and it put the focus firmly on books 1 and 8 of *Vom Kriege* – those that deal with the relationship between war and politics: this was a primer for the vocabulary of nuclear deterrence as much as it was an insight into Napoleonic warfare.

Although Paret approached the task as a historian of ideas, Howard did so as a student of strategic studies. Through his revitalization of the war studies department of King's College London in the 1960s, Howard reforged the links between military history and current policy (Howard, 1983). His own subsequent distinction in both fields, at Oxford and then at Yale, ensured that, in the UK at least, the health of military history was tied in to the study of international relations. In the USA as well as in continental Europe the relationship between military history and strategic studies has been more antagonistic. The cynic would say that, deprived of the armed services' commitment to the study of the subject, evident especially in Germany and France, and lacking the financial resources of the history departments of American universities, British military history has had no option but to embrace a marriage of convenience.

There are therefore distinct national approaches to military history, reflections in large part of the institutional diversity of its roots. Comparative military history has barely begun. Moreover, the strengths and weaknesses of each national school can reflect the fact that this form of history above all is tied in to the nation-state, its formation, and its self-identity. German military history since 1945 has predictably been concerned with World War II to the detriment of earlier wars; the Soviet Union largely ignored Russian military history before 1917. But both these patterns are being broken down. And Britain, which according to this sort of logic ought to be leading the world in naval history, actually let the subject crumble in the 1970s and 1980s.

'Popular' Military History

The popularity of military history outside the confines of universities and staff colleges, at least within the UK and the USA, makes military history an obvious route by which to introduce the nonspecialist public to history more generally. Some of the writing, such as that of Henri Lachouque or Georges Blond in France, can come close to romantic nostalgia; the work of others, such as Paul Carell on the Wehrmacht in World War II, can seem striking for what it leaves out. But the best work in this genre – both German Werth and Alistair Horne on Verdun, for example – deserves, and has received, serious attention from scholars. In Britain, John Keegan and Antony Beevor have thrived on the back of military history's flowering in academic circles. Much of what they do is traditional. Keegan, like Liddell Hart, ranges across time and space; like Liddell Hart he also writes with the fluency and ease of a journalist; and, as with Liddell Hart, there is a didactic thrust. But the appeal is also in the narrative, as the staggering success of Beevor's *Stalingrad* (1998) displays. Keegan (1976) criticized historians of war who focused on its operational and tactical level but failed to explain the experience of combat itself. Beevor's book puts such precepts into practice. The academic world has also responded to Keegan's call. Research on war 'from below' not only has popular appeal but also interdisciplinary potential.

Military historians continue to complain that they are marginal figures in the academic world. In mainland Europe, that complaint has some substance, but becoming less year on year. And in the USA, where it is stated most vociferously, it is without serious foundation. However, its current strength derives in large part from its past battles to establish itself. Lacking a linear pedigree, military history is a hybrid whose resilience derives from the multiplicity of its approaches.

See also: Arms Control; Chemical Science: History and Sociology; Enlightenment; Military Geography; Military Sociology; Military, War, and Politics; War, Sociology of; Warfare in History; Wars among Nation-States: Patterns and Causes.

Bibliography

Alger, J.I., 1982. The Quest for Victory: The History of the Principles of War. Greenwood, Westport, CT.
Beevor, A., 1998. Stalingrad. Viking, London.
Bucholz, A., 1985. Hans Delbrück and the German Military Establishment. University of Iowa Press, Iowa City, IA.
Bucholz, A., 1997. Delbrück's Modern Military History. University of Nebraska Press, Lincoln, NE.
Bradford, J.C., 2006. International Encyclopedia of Military History. Routledge, New York.
Black, J., 2004. Rethinking Military History. Routledge, New York, NY.
Charters, D.A., Milner, M., Wilson, J.B., 1992. Military History and the Military Profession. Greenwood, Westport, CT.
Clausewitz, C. von, 1976. On War (M. Howard, P. Paret, Trans.). Princeton University Press, Princeton, NJ.
Cowley, R., 2005. The Cold War: A Military History. Random House, New York.
Deist, W., 1998. Hans Delbrück. Militärhistoriker und Publizist. Militärgeschichtliche Mitteilungen 57, 371–383.
Delbrück, H., 1975–85. History of the Art of War within the Framework of Political History. Greenwood, Westport, CT.
DeVries, K., 2002. A Cumulative Bibliography of Medieval Military History and Technology. Brill, Leiden.
Fortescue, J.W., 2012. Military History. Cambridge University Press, Cambridge.
Gat, A., 1989. The Origins of Military Thought from the Enlightenment to Clausewitz. Oxford University Press, Oxford.
Graff, D.A., Higham, R., 2002. A Military History of China. Westview, Boulder, CO.
Hattendorf, J.B. (Ed.), 1994. Ubi Sumus? The State of Naval and Maritime History. Naval War College Press, Newport, RI.
Howard, M., 1983. The use and abuse of military history. In: Paret, P. (Ed.), The Causes of Wars. Temple Smith, London.
Holmes, R., Strachan, H., Bellamy, C., Bicheno, H., 2001. The Oxford Companion to Military History. Oxford University Press, Oxford.
Jomini, A.H.de, 1992. The Art of War. Greenhill, London.
Keegan, J., 1976. The Face of Battle. Jonathan Cape, London.
Kennedy, P., 1989. The fall and rise of military history. The Yale Journal of World Affairs 1 (2), 12–19.
Keegan, J., 2009. The American Civil War: A Military History. Alfred A. Knopf, New York.
Lange, S., 1995. Hans Delbrück und der 'Strategiestreit'. Kreigsführung und Kreigsgeschichte in der Kontoverse 1879–1914. Rombach, Freiburg im Breisgau, Germany.
Morillo, S., 2006. What Is Military History? Polity Press, Cambridge.
Murray, W., Scales, R.H., 2003. The Iraq War: A Military History. Belknap Press of Harvard University Press, Cambridge, MA.
Paret, P., 1992. The history of war and the new military history. In: Paret, P. (Ed.), Understanding War: Essays on Clausewitz and the History of Military Power. Princeton University Press, Princeton, NJ.
Rogers, C., 1995. The Military Revolution Debate: Readings on the Military Transformation of Early Modern Europe. Westview, Boulder, CO.
Shy, J., 1986. Jomini. In: Paret, P. (Ed.), Makers of Modern Strategy from Machiavelli to the Nuclear Age. Oxford University Press, Oxford, UK.

Military Sociology

Mary K Kniskern and David R Segal, University of Maryland, College Park, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The field of military sociology is interdisciplinary and international. As the discipline developed, concepts were incorporated not only from other social sciences, but also from cultural shifts and political demands on the military.

From an initial focus on enlisted ground combat forces, dominated by U.S. sociologists during and following World War II, the military sociology agenda expanded to include professionalism, civil-military relations, changing military operations, families, veterans, and reservists in all branches of the armed forces and in many nations.

Introduction

During the development of Western sociology as a discipline in the late eighteenth and nineteenth centuries, early European sociologists saw war and conflict as important social processes on a continent repeatedly affected by war. In spite of the attention European sociologists paid to military forces as the organizations that waged those wars, the continental sociology transported to the United States of America reflected the more pacific theorists and theories, and disregarded an awareness of the military (Malešević, 2010). Thus, while Max Weber's theory of bureaucratic organization came to be widely taught in American sociology and political science departments, organizational psychology programs, and colleges of business and management, the fact that Weber's prototype had been the Prussian army was generally overlooked; and while Durkheim's theory of suicide became a staple of American sociology, the category of heroic suicide – for example throwing oneself on an explosive device to protect one's comrades – as a form of altruistic suicide, was virtually never mentioned.

More than many other fields of sociological study, military sociology is interdisciplinary. It draws on history, psychology, political science, and economics to meet the demands of multiple audiences beyond academia including military personnel and policy makers. An applied approach to research also sets military sociology apart from many other sociological fields. Enduring themes in military sociology include: motivation, cohesion, and effectiveness in combat; military professionalism; civil–military relations and citizenship; organizational changes within the military institution; and changing demographics of military personnel including issues of population diversity. In response to the ongoing conflicts in Iraq and Afghanistan, fought by multinational forces, as well as cultural changes in civilian society, military sociologists and their various audiences have become more concerned with changing military roles and strategies such as peacekeeping, counterinsurgency, and nation building; the changing roles of women and homosexuals within the organization; the effects of service on veterans and their families; and the involvement of reserve forces to meet the demands of ongoing conflicts.

Foundations of the Interdisciplinary and Applied Approaches

Although sociology was established as an academic discipline in the United States by the 1890s, an analysis of the growth of military sociology at the end of the twentieth century (Boëne, 2000) indicated that only about 5% of identified works were published before 1942. Boëne attributed the slow growth of the field largely to ideological liberalism, a meliorist orientation to social problems within the discipline, and war weariness after World War I. We would add to this the absence of a standing army during peacetime. Early contributions to the social science study of the military were by psychologists, historians, and political scientists, providing a foundation for the interdisciplinary orientation of military sociology, although they emphasized less sociological objects and units of analysis. On a micro-level, psychologists focused on individual abilities and behavior, contributing to aptitude testing and the understanding of effectiveness. On a macro-level, political scientists examined war as a component of the process of international relations, such as Charles Merriam's project on the causes of war, conducted at the University of Chicago, which led to Quincy Wright's (1942) seminal study of war. Not until sociologists were involved during World War II was much attention paid to the military as a social institution, an organization, an occupation, or a profession.

By the mid-twentieth century, the field of military sociology was dominated by Americans. The United States mobilized large numbers of social scientists, including academic sociologists during World War II. The pattern of interdisciplinary collaboration was established because the problems studied extended beyond the bounds of any single academic discipline. The emphasis on enlisted ranks and ground combat forces emerged in response to the nature and size of the mobilization during a mass-force era of military combat. Social scientists were mobilized explicitly to help manage the army and the war, orienting military sociology toward organizational and small-group processes rather than more macrolevel concerns, with an emphasis on policy applications for the research at army level rather than generating national policies regarding the army and other services.

Motivation, Cohesion, and Effectiveness in Combat

Not all sociologists were mobilized in research roles during World War II. Contributions by others who applied sociological skills to their military experiences, such as George C. Homans's (1946) observations of social relations on a small warship, which contributed to his conceptualization of exchange theory, and Tamotsu Shibutani's (1978) study of demoralization in a company of Japanese-American soldiers greatly enriched the field of military sociology. In 1946, *The American Journal of Sociology*, the major sociological journal of that time, published a special issue titled *Human Behavior in Military Society*. Included in this issue were Arnold Rose's study of military social structure, Alfred Lindesmith's observations of the effects of their status in service on the self-esteem of teachers, and August Hollingshead's article on adjustment to army life and to subsequent civilian life. Sociologists from other specialties continued to enrich the field by using the military as a venue for research and theorizing in the post-World War II years and into the twenty-first century.

Supporting both social research and personnel management, the landmark study of World War II reported results of experiments and surveys conducted by the Information and Education Division of the War Department. Among the topics included in the four volumes of *Studies in Social Psychology in World War II*, including the two-volume *American Soldier* studies authored by Samuel A. Stouffer et al. (1949a,b), were cohesion, leadership, primary groups, morale, race relations (the army was still racially segregated), communication, and persuasion. This study established the foundation of the research agenda of sociology and social psychology for many years. The power and thrust of this study served to focus attention on the enlisted personnel and shaped quantitative methodology in sociology as a whole. Thirty-five years after the publication of *The American Soldier*, *Social Psychology Quarterly*, the major journal in sociological social psychology, published an issue devoted to the impact of this work. Just as the U.S. military had adopted psychological selection and classification tests after World War I, the use of survey research to inform personnel policies continued after World War II.

Military Professionalism

Unlike previous military conflicts in U.S. history, World War II was not followed by demobilization (Burk, 2001). After the war, the United States, its allies, and adversaries maintained sizable forces, reflecting the new bipolar tensions in the international community produced by the Cold War (Segal and Segal, 2004). While the manpower to maintain these standing armies was frequently supplied by conscripted forces, the focus of social scientists shifted to the officer corps and military professionalism with publications by Samuel P. Huntington (1957), a political scientist at Harvard University, and Morris Janowitz (1960), a sociologist at the University of Michigan on the nature of the military profession and its relationship to the state and to society. Research on the professions was becoming a central theme in the sociology of occupations, and Huntington and Janowitz were not the first to examine the military through the professional lens. Prior to publication of their books, Norbert Elias published the first of an intended three-part series of articles examining the development of the officer profession in the British navy in 2007[1950]. Sociologists in other nations applied the questions raised by Huntington and Janowitz to militaries in their own nations. The focus of the developing definition of a profession included identifying the particular expertise of its practitioners, the responsibilities, status and authority attributed to the profession, and a strong group identity with autonomous regulation of access and discipline.

While mid-century conceptions of professionalism emphasized the active-duty officer corps, the challenges of the more technologically complex twenty-first century military suggest questions about expanding the definition of military professional to include well educated and technically proficient career-oriented enlisted personnel. Additional questions reflect the changes in the sociology of professions, the extent to which specialization in the armed forces requires recognizing each branch as a separate profession, and the degree to which reservists with other employment should be considered military professionals (Abbott, 2002).

Civil–Military Relations and Citizenship

Scholars in the mid-twentieth century attempted to describe the structural relationships between modern military forces and their host societies. Among the most important of these were Mills's (1956) *The Power Elite* and Lasswell's model of *The Garrison State*. Issues of the relationship between the military, the state, and society became even more of a concern following World War II. Civil–military relations continued to hold a place on the sociological agenda, even as the research focus returned to the enlisted soldier. Involvement through participation in militia units, mobilization and conscription, and a linkage between military service and citizenship had been an expectation of male citizens since the American and French Revolutions (Kestnbaum, 2002). In the United States, perceptions of inequality in the conscription process that placed the burdens of fighting the Vietnam War disproportionately on the poor drove a debate on ending conscription and manning the military force with volunteers. Economists held a central position in the discussion of whether labor market dynamics could replace conscription in meeting the manpower needs of the Cold War military. The decision was made and the 1972 defense appropriation provided funds for the establishment of an all-volunteer military force. The United States announced the end of peacetime conscription for the first time since World War II in January 1973.

As they had during World War II, the U.S. armed forces conducted and funded behavioral science research to support the challenge of maintaining a large standing force with volunteers. Although the service research units had emphasized psychology and the measurement of aptitudes since World War II, they reincorporated sociology and the study of attitudes. Military sociology grew as the military research units funded research in-house and at universities.

Organizational Changes within the Military

The shift to an all-volunteer force in the United States was a reflection of an international trend away from conscription, and reflecting this trend military sociology began to reconsider the nature of military service and its relationship to society. With Moskos's (1977) formulation of institutional and occupational models of the military, the focus of military sociology returned to enlisted personnel. Moskos suggested that military service was being transformed from a values-based vocation to a job motivated by economic benefits. This had implications for understanding both the individual soldier and military organization (Segal, 1986a) and affected both the military's recruiting efforts and sociologists' research on identity concepts. Moskos's conceptualization and many of its dimensions of change dominated the military sociology research agenda for more than a quarter of a century, as nations around the world abandoned conscription in favor of volunteer forces (Haltiner, 1999) and scholars applied Moskos's models to their own nations (Moskos and Wood, 1988).

As the U.S. Army was restructured after the end of the Cold War, military bases were reorganized. A new aspect of the relationship between the military and civilian society was added to the research agenda. We learned that communities around military bases have less racial inequality in employment and less racial segregation in housing than similar communities without the military influence. We also learned that gender discrimination in employment, as measured by higher female unemployment, lower female wages for all women, and lower returns to human capital investments, is higher in communities where the military is a major presence in the economy (Booth, 2003).

Changing Nature of Military Conflict

In addition to manpower changes, military organizations had to respond to the changing nature of military conflict. Moskos's formulation posited that occupational militaries would focus more on constabulary operations than conventional interstate wars, as Janowitz had hypothesized for military professionals following World War II. During the Cold War period following World War II, the principal antagonists (the United States and the USSR) and their allies maintained conventional mass-force structures while the 'middle powers,' countries such as Canada, the Netherlands, the Nordic countries and smaller countries, adjusted to the shifting demands of peacekeeping missions under the auspices of the United Nations. Cold War peacekeeping doctrine emphasized impartiality, minimal force, and host nation consent. Impartiality excluded the major nations since they were likely to be interested parties in any area of the world in which conflict occurred (Moskos, 1976). The United States did get involved on a continuing basis in peace operations conducted under auspices other than the United Nations (Segal and Segal, 1993).

With the end of the Cold War and the collapse of the Soviet Union and the Warsaw Pact, the center of gravity of military sociology shifted to Europe with a focus on changing the missions of military forces from conventional interstate wars to a broader spectrum of operations including strategic peacekeeping and nation building. Military budgets in many nations decreased and the size of military forces also decreased significantly. The major powers became more involved in peace operations at the same time that nations with more pacific security policies, such as Japan and Germany whose military roles had been limited at the end of World War II, were encouraged under new international norms of burden sharing to become involved in multinational peace operations (Segal and Kurashina, 2007). Military operations became less about conventional forces representing recognized states with similar allies facing similar adversaries on a defined battlefield, and more concerned with containing nonstate actors such as insurgencies and terrorism. These changes have been reflected in the field of military sociology.

Technology has also altered the military and military sociology. In addition to visible changes in war-fighting technology, communications technology has changed our perceptions of war and relationships between society, soldiers, and their families back home. From war correspondents in World War II, whose reports were reviewed by military censors before appearing in print the next day or in a newsreel the following week, to same day headlines and film from Vietnam on the evening news, to reporters embedded in military units filing their stories in real time from Iraq and Afghanistan, communications technology has changed the connections between civil society and the military. For the soldiers and their families, communications have progressed from mail with long delays between writing and receiving letters, through telephones and faxes, to frequent and widespread access via the Internet (Ender and Segal, 1998). In addition to the potential for distracting the war fighter with concerns about family struggles at home, communications technologies raise issues with information and mission security.

Changing Demographics in the Force – Race, Gender, and Sexual Orientation

The Vietnam War brought an increasing awareness in the United States of the effect that social trends were having on the military and broadened the subject matter of military sociology. Military units during World War II were segregated by race and gender. Minority soldiers, sailors, marines and airmen, and all women were limited in the positions available to them in the mid-century military. Manpower needs led to the desegregation of army units during the Korean War, but civilian society in the United States and the U.S. military experienced racial tensions during and after the Vietnam War. Even after the turn of the century, there were concerns about equal access to opportunity and equitable distribution of the demands of military service on the basis of race. Race was not as much of an issue in other Western industrial nations, but gender integration was a general concern. Women had served in the U.S. military during World War II and worked in the civilian labor force to keep the economy running and to support the men fighting overseas. After the war, middle class wives returned to their homes. It was another quarter century before female participation in the civilian labor force

and the demands of an all-volunteer force motivated the military to incorporate women into nontraditional roles. With the end of conscription, the U.S. military needed to incorporate racial minorities and women into the military in order to meet recruiting goals, and military sociologists added gender integration to the research agenda.

Moskos's institutional/occupational formulation suggested that women would increasingly be integrated into the military on an equal basis. Major advances toward equality in the numbers and roles of women in military service have taken place both in the United States and in other nations (Segal, 1995). Other dimensions reflecting increasing concern with diversity include continuing concerns with racial equality (Moskos and Butler, 1996); increased attention to the intersections of race, class, and gender (Booth and Segal, 2005; Moore, 1996, 2003); and emergent concerns with sexual orientation integration (Scott and Stanley, 1994).

In 1992, U.S. presidential candidate Bill Clinton promised to end a long-standing ban on homosexual men and lesbians serving in the U.S. armed forces. The change was resisted by military leaders, politically and religiously conservative organizations, and members of Congress. Research conducted at that time showed that some of the NATO allies similarly had policies that banned homosexuals from service, some allowed them to serve openly, and others did not care. The research also showed a significant gap between policy and practice (Segal et al., 1994). In 1993, the Clinton administration agreed to a policy that allowed gay service members to serve as long as they did not reveal their sexual orientation or engage in homosexual acts: a policy known as Don't Ask, Don't Tell (DADT). In 2010, the U.S. Congress voted to repeal DADT, once the President, the Secretary of Defense, and the Chairman of the Joint Chiefs of Staff certified that openly gay men and women serving would not harm military readiness. The 1993 debates involved considerable testimony by military sociologists, and led to an increase in the amount of research done on gays in the military. In 2010, the Secretary of Defense appointed a Comprehensive Review Panel to advise him on how the ban was to be lifted without impacting negatively on military performance. Again, military sociologists figured prominently in the review. DADT was repealed on 10 September 2011. Research conducted on the impact of repeal after 1 year showed that none of the negative consequences that had been anticipated by opponents of repeal had occurred (Belkin et al., 2012).

Reserve Forces

Reserve units in the American military had not been activated during the Vietnam War. Rather, a conscription system that deferred college students and burdened less advantaged populations fueled social unrest in the United States that was manifested in other nations as well. Following Vietnam, resources and tasks were redistributed among the active and reserve U.S. forces and the National Guard within the concept of a 'total force' that would require activation of reserve components to support future military missions. This allocation of essential mission support activities to reserve and National Guard units continued during the downsizing after the end of the Cold War.

While military forces were downsized, active contingency operations for humanitarian and peacekeeping purposes expanded, straining the capacities of the active duty military. Three approaches were taken to meet the increasing demands: maintaining a standby conscription system, increasing the use of reserve forces, and outsourcing. Jobs which required military competence and status were performed by military personnel while civilians were hired to fill positions that did not require military expertise or authority, and ultimately, deploying both active duty and reserve forces more frequently and for longer period of time than is optimal for military effectiveness (Korb and Segal, 2011). Just under 25% of the current U.S. Department of Defense workforce is composed of civilian contractors hired on a contingency basis. While civilian contractors have been used to support military operations since before the Civil War, the use of civilians to offset downsizing of active duty forces during a period of active deployment is unique. Military sociologists are beginning to explore the implications of having large numbers of noncombatant civilian contractor personnel who are exempt from military discipline colocated with military personnel in a combat zone (Kelty, 2009).

With the post-Cold War downsizing of the active force, nearly half of the army's combat force was in state-based National Guard units. The first challenge facing the reorganized total force was Operation Desert Shield/Desert Storm on the Arabian Peninsula in 1990. At the height of that conflict, about one-fourth of the personnel in the combat theater were Army Reserve and National Guard personnel, more than 73 000 in total. The bulk of the deployed reserve units served support functions such as transportation, medical, and military police units. The National Guard combat brigades that had been intended to join active duty combat divisions were judged unprepared for combat, and programs were developed to improve future deployability of National Guard combat troops.

By the mid-1990s, the army began deploying small numbers of reserve component personnel overseas for 6-month deployments in support of contingency operations. On the basis of a 1994 experiment, for example, American infantry participation in the Multinational Force and Observers in the Sinai Desert in support of the Camp David Accords, was shifted from active infantry battalions of the 18th Airborne Corps to National Guard battalions. With America's invasion of Iraq in 2002, the role of reserve personnel shifted from participants in contingency operations to participation in continuous operations subject to multiple longer deployments of a year or more. Increased reliance on reserve units has required more research to address the distinct concerns of reserve component personnel, their families, and their civilian employers.

Effects of Service on Service Members and Families

One concern drawing particular attention after 10 years of active conflict is the physical and mental well-being of the troops, many of whom have experienced multiple combat deployments. Post-traumatic stress disorder and high suicide rates among returning U.S. military personnel have returned

reentry stress to the research agenda as it was following the return of Vietnam veterans.

Veterans benefits and postservice outcomes such as education, employment, and earnings have been on the military sociology agenda throughout the all-volunteer era. The primary finding of this research has been that men who served during World War II or the Korean War benefited from their service when compared with their civilian peers who did not serve. Veterans of later conflicts have not been shown to experience the same benefits. The life course perspective, a major perspective in the larger sociological community, has been used to clarify the processes by which military service affects postservice trajectories (Wilmoth and London, 2012).

Following World War II, Reuben Hill (1949) published a landmark study of the stress that military service imposes on families. Twenty-first century sociologists in Europe and the United States continue to study family stressors and their influence on service member readiness and retention. While the conscripted force of World War II through the Vietnam War was primarily composed of young single men, the all-volunteer force of the later twentieth and twenty-first century is older and more likely to be married than in earlier years, particularly when reserve forces are considered. Sociological research has informed changes in policy and management practices to address the conflicting demands placed on service members by the military and their families (M.W. Segal, 1986b), the needs of family members that demand support from the military, and the effects that family members have on service member readiness and retention (Bourg and Segal, 1999).

Research has consistently shown that service members marry younger than their civilian peers, have first children earlier, and have larger families. Racial differences in marriage rates and stability are significantly lower in the military than in comparable civilian populations (Lundquist, 2004). More attention has also been given to the effects of being in a military family on child and spouse well-being and spousal employment issues as the multiple deployments to combat zones have strained family members' health and challenged employment.

Military Sociology as an Academic Discipline

Not until the 1960s, more than half a century after sociology departments were established at major universities, did military sociology emerge as a viable academic field. Charles Coates and Roland Pellegrin, of the University of Maryland, published the first textbook on military sociology in 1965. An organizational infrastructure in military sociology began with the establishment of the Inter-University Seminar on Armed Forces and Society (IUS), a small group of American scholars from several universities with interests in military sociology who met periodically to discuss their research, first at the University of Michigan and then at the University of Chicago. Since then the IUS has grown to an international and interdisciplinary group of more than 600 scholars, military officers, and students from over 35 countries around the world that meet biennially. Additional support for the study of military sociology came from the Research Committee on Armed Forces and Society of the International Sociological Association that began to bring military sociologists from a number of nations together every 4 years at the World Congresses of Sociology, and evolved into the Research Committee on Armed Forces and Conflict Resolution. In 1986, scholars concerned with military sociology in Europe formed the European Research Group on Military and Society (ERGOMAS). Through themed working groups and a biennial conference, ERGOMAS fosters interdisciplinary and international collaboration on the relationship between the military and civilian society.

Growth in military sociology corresponded with increasing fragmentation and specialization in sociological publications, as well as a reluctance among major sociological journals to publish articles on war or the military in the wake of the Vietnam War. Two specialized journals, *Armed Forces and Society (AF&S)*, published by IUS, and *Journal of Political and Military Sociology (JPMS)* were established to provide outlets for the increased research and writing in the field. JPMS has recently become a research annual, and the publishing field has been expanded with the establishment of *Res Militaris*, an international electronic journal, reflecting advances in publication technology.

Articles published in these journals, and more recently in major sociological journals, reflect long-term research traditions, as well as the field's responsiveness to the changing needs of the military, and changes in sociological theory and society as a whole. Recognition of sociological contributions to the military mission is apparent in the increasing number of courses in military sociology offered and the numbers of military sociologists being included among the faculty of military academies internationally (Caforio, 2000).

The processes of group dynamics continue as active areas of study, but conceptualizations changed. Cohesion and effectiveness were important concerns during World War II and important topics on the early military sociology research agenda. While early research emphasized social cohesion among homogenous troops without demonstrating a positive impact on performance, more recent conceptualizations emphasize task cohesion based on contributions to common goals (Segal and Kestnbaum, 2002). In the 1990s, the questions about the relationship between the conscripted military, the state, and the society raised by Huntington and Janowitz that motivated international research at that time and in the early years of the volunteer force were reframed within the context of the larger sociological interest in social context. New questions considered the extent to which the culture of the military diverged from that of its host society (Feaver and Kohn, 2001), and once again motivated international comparative studies. Research showed that the American military culture, like other militaries and other professions, is distinct yet consistent with the broader culture within which it is embedded and which it defends. Another set of questions concerns the nature of the relationship between the military institution and the state, and between military service and citizenship, and the implications of those relationships both nationally and internationally (Kestnbaum, 2002).

Research on military conscription continues to be a high priority as nations around the world discuss ending conscription among concerns about the equity of less than universal service and the manpower demands of ongoing conflicts.

While military sociology maintains a strong interdisciplinary orientation, changes in the larger discipline of sociology influence the questions asked by military sociologists. One example of this is Moskos's response to postmodernist theorizing in the sociological community. Like his earlier conceptualization of the shift from a values-based institutional model to an economically based occupation, Moskos proposed a shift from a modern to a postmodern military. Some of the dimensions specified reflect microorganizational aspects such as gender roles and the relationship between the military and service members' families. Macrolevel dimensions such as changes in mission and force structure did not correspond to components of the institutional/occupational (I/O) model. Like the I/O model, the postmodern model has been applied to militaries in many nations (Moskos et al., 2000) finding that, while militaries vary in the degree to which they would be classified as modern, none qualify as postmodern under this template. Principal critiques of the postmodern military concept are that the analytic structure is inherently positivistic, not postmodern at all, and that no structured military can be truly postmodern by the nature of postmodernism (Booth et al., 2001).

Military sociology continues to influence military policy and practice. Research informs changes in social programs in support of veterans and service families, in manpower management issues such as gender integration and issues of equity, and in actual operations. One example of sociological theory applied to combat is the use of social network theory in the mission to locate Saddam Hussein (Reed and Segal, 2006). The primary professional organization for military sociologists, IUS, includes many active-duty and reserve military professionals among the scholars contributing to conferences and in publication. Research has expanded beyond ground combat forces to include naval and air forces.

As an academic discipline, military is still a small, yet growing, subfield within sociology. A growing number of colleges and universities offer courses or programs studying the military from a sociological perspective. As nations in Central and Eastern Europe consider ending conscription, redesign the relationship between the military and their new democratic governments, modernize and professionalize the military, and address issues of gender integration and service members' families, interest in the applied aspects of military sociology has grown. In spite of renewed interest in military sociology inspired by the Global War on Terror and a decade of involvement in active combat by U.S. troops, the center of gravity of the field appears to have moved from North America to Europe since the end of the Cold War.

See also: Citizenship, Historical Development of; Conflict and War: Anthropological Aspects; Defense and Security Economics; Durkheim, Emile (1858–1917); Geopolitics; Insurgency; Marx, Karl (1818–83); Militarism; Military History; Military, War, and Politics; Nationalism, Sociology of; Organizations: Authority and Power; Professions in Organizations; War, Sociology of; Weber, Max (1864–1920).

Bibliography

Abbott, Andrew, 2002. The army and the theory of professions. In: Snider, Don M., Watkins, Gale L., Matthews, Lloyd J. (Eds.), The Future of the Army Profession. McGraw Hill, New York, pp. 523–536.

Belkin, Aaron, Ender, Morten, Frank, Nathaniel, Furia, Stacie, Packard Jr., Gary, Schultz, Tammy S., Samuels, Steven M., Segal, David R., 2012. One Year Out: An Assessment of DADT Repeal's Impact on Military Readiness. Palm Center, Los Angeles.

Boëne, Bernard, 2000. Social science research, war and the military in the United States. In: Kümmel, Gerhard, Prüfert, Andreas D. (Eds.), Military Sociology. Nomos, Baden-Baden, pp. 149–251.

Booth, Bradford, 2003. Contextual effects of military presence on women's earnings. Armed Forces & Society 30, 25–52.

Booth, Bradford, Segal, David R., 2005. Bringing the soldiers back in: implications of inclusion of military personnel for labor market research on race, class & gender. Race, Gender, & Class 12, 34–57.

Booth, Bradford, Kestnbaum, Meyer, Segal, David R., 2001. Are post-cold war militaries post-modern? Armed Forces & Society 27 (3), 319–342.

Bourg, Chris, Segal, Mady Wechsler, 1999. The impact of family supportive policies and practices on organizational commitment to the army. Armed Forces & Society 25, 633–652.

Burk, James, 2001. The military's presence in American society, 1950–2000. In: Feaver, Peter D., Kohn, Richard H., Cohn, Lindsay (Eds.), Soldiers and Civilians: The Civil-Military Gap in American National Security. MIT Press, Cambridge, MA, pp. 247–274.

Caforio, Giuseppe (Ed.), 2000. The European Officer: A Comparative View on Selection and Education. Edizioni ETS, Piza.

Coates, Charles, Pellegrin, Roland, 1965. Military Sociology: A Study of American Military Institutions and Military Life. Social Science Press, University Park, Maryland.

Elias, Norbert, 2007[1950]. In: Moelker, Rene, Mennell, Stephen (Eds.), The Genesis of the Naval Profession. University College Dublin Press, Dublin.

Ender, Morten E., Segal, David R., 1998. Cyber-soldiering: race, class, gender, and new media use in the military. In: Ebo, Bosah (Ed.), Cyberghetto or Cybertopia? Greenwood Press, Westport, pp. 65–81.

Feaver, Peter D., Kohn, Richard H. (Eds.), 2001. Soldiers and Civilians: The Civil-military Gap and American National Security. MIT Press, Cambridge, MA.

Haltiner, Karl, 1999. The definite end of the mass army in Western Europe. Armed Forces & Society 25, 7–36.

Hill, Reuben, 1949. Families under Stress. Greenwood Press, Connecticut.

Homans, George, C., 1946. The small warship. American Sociological Review 11, 294–300.

Huntington, Samuel P., 1957. The Soldier and the State: The Theory and Politics of Civil-military Relations. Belknap Press, Cambridge, MA.

Janowitz, Morris, 1960. The Professional Soldier: A Social and Political Portrait. The Free Press, New York.

Kelty, Ryan, 2009. Citizen soldiers and civilian contractors: soldiers unit cohesion and retention attitudes in the total force. Journal of Political and Military Sociology 37, 1–27.

Kestnbaum, Meyer, 2002. Citizen-soldiers, national service and the mass army: the birth of conscription in revolutionary Europe and North America. Comparative Social Research 20, 117–144.

Korb, Lawrence J., Segal, David R., 2011. Manning & financing the twenty-first century all-volunteer force. Daedalus 140, 75–87.

Lundquist, Jennifer Hickes, 2004. When race makes no difference: marriage and the military. Social Forces 83, 731–757.

Malešević, Siniša, 2010. How pacifist were the founding fathers? War and violence in classical sociology. European Journal of Social Theory 13, 193–212.

Mills, C. Wright, 1956. The Power Elite. Oxford University Press, New York.

Moore, Brenda L., 2003. Serving Our Country: Japanese-American Women in the Military during World War II. Rutgers University Press, New Brunswick.

Moore, Brenda L., 1996. To Serve My Country, to Serve My Race. New York University Press, New York.

Moskos, Charles C., Williams, John Allen, Segal, David R. (Eds.), 2000. The Postmodern Military: Armed Forces after the Cold War. Oxford University Press, New York.

Moskos, Charles C., Butler, John Sibley, 1996. All that We Can Be. Basic Books, New York.

Moskos, Charles C., Wood, Frank R. (Eds.), 1988. The Military: More than Just a Job?. Pergamon-Brassey's, Washington.

Moskos, Charles, 1977. From institution to occupation: trends in military organization. Armed Forces & Society 4 (1), 41–50.

Moskos, Charles C., 1976. Peace Soldiers: The Sociology of a United Nations Military Force. University of Chicago Press, Chicago.

Reed, Brian, Segal, David R., 2006. Social network analysis and counter-insurgency operations: the capture of Saddam Hussein. Sociological Focus 39, 251–264.

Scott, Wilbur, Stanley, Sandra Carson (Eds.), 1994. Gays and Lesbians in the Military: Issues, Concerns, and Contrasts. Aldine De Gruyter, New York.

Segal, David R., Kurashina, Yuko, 2007. New missions, or old missions for new actors. International Review of Sociology 17, 5–24.

Segal, David R., Segal, Mady Wechsler, 2004. America's military population. Population Bulletin 59 (4).

Segal, David R., Kestnbaum, Meyer, 2002. Professional closure in the military labor market: a critique of pure cohesion. In: Snider, Don M., Watkins, Gale L. (Eds.), The Future of the Army Profession. McGraw Hill, New York, pp. 441–458.

Segal, David R., Gade, Paul A., Johnson, Edgar M., 1994. Social science research on homosexuals in the military. In: Scott, Wilbur J., Stanley, Sandra Carson (Eds.), Gays and Lesbians in the Military: Issues, Concerns and Contrasts. Aldine de Gruyter, Hawthorne, NY, pp. 33–51.

Segal, David R., Segal, Mady Wechsler, 1993. Peacekeepers and Their Wives. Greenwood Press, Westport.

Segal, David R., 1986a. Measuring the institutional/occupational change thesis. Armed Forces & Society 12, 351–376.

Segal, Mady Wechsler, 1995. Women's military roles cross-nationally: past, present, and future. Gender & Society 9, 757–775.

Segal, Mady Wechsler, 1986b. The military and family as greedy institutions. Armed Forces & Society 13, 9–38.

Shibutani, Tamotsu, 1978. Derelicts of Company K: A Sociological Study of Demoralization. University of California Press, Berkeley, CA.

Stouffer, Samuel A., Suchman, Edward A., DeVinney, Leland C., Star, Shirley A., Williams Jr., Robin M., 1949a. The American Soldier: Adjustment during Army Life, vol. 1. Princeton University Press, Princeton, NJ.

Stouffer, Samuel A., Lumsdaine, Arthur A., Lumsdaine, Marion Harper, Williams Jr., Robin M., Smith, M. Brewster, Janis, Irving L., Star, Shirley A., Cottrell, Leonard S., 1949b. The American Soldier: Combat and Its Aftermath, vol. 2. Princeton University Press, Princeton, NJ.

Wilmoth, Janet M., London, Andrew S. (Eds.), 2012. Life Course Perspectives on Military Service. Routledge Press, New York.

Wright, Quincy, 1942. The Study of War. University of Chicago Press, Chicago.

Military, War, and Politics

Jan W Honig, King's College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Chisholm, volume 14, pp. 9857–9859, © 2001, Elsevier Ltd.

Abstract

Ensuring that the military fight in the interests of government and society has been a major issue throughout history. The political structure that is generally seen as the model of success is the state, and in particular the modern democratic state. The theories underpinning this claim tend to be highly normative and context dependent. Based overwhelmingly on a strongly entrenched Western-centric narrative of historical development, much social science has struggled to reorient itself to understand and address the current conflicts and political disorders that rack the globe.

The relationship between military, war, and politics is an important issue for four reasons. First, war can make or break societies and their governments. Second, war, understood as the application of violence, is often believed to possess a special, inherent dynamic that makes it prone to escalation and loss of control. Third, the military's special skills set – the management of force – might make them prone to advocating and implementing inappropriately violent solutions to political problems, whether domestic or international. Fourth, the tendency of bureaucracies to pursue self-interested ends imbues a military bureaucracy with a special danger: it has a uniquely powerful potential to making its voice heard and getting its way. These concerns about war and the military mean that societies (or those that matter in society) seek to establish ways of embedding the military within political structures that ensure that it only fights in the interests of government and society.

The issue of controlling war and the military achieved special salience in the modern world with the rise of a particular form of political organization: the state. The state established the military as a distinct bureaucracy. It also developed the capacity to mobilize manpower and push military technological production to a degree that made wars unprecedentedly destructive. The relationship between state, war, and military was characterized by an inherent instability. War and military were instruments of state that continually bridled at the leash of state control. The modern academically informed debates about stabilizing the relationship were, on the one hand, developed under the shadow of the world wars and the spectre of nuclear weapons. On the other hand, these debates have not been free from strong normative and empirical Eurocentric influences that stretch back further in time. The leading ideas underpinning modern discussions about the proper polity, the nature of war, and the special position of the military are predominantly of a post-French Revolutionary, nineteenth century vintage. As the modern theories associated with military, war, and politics cannot be divorced from historical and normative context, it is useful first to outline the historical master narrative that frames modern understandings, before briefly identifying the main post-Second World War theories. This broader context helps explain the difficulties much analysis and theorizing is encountering in the face of the challenges thrown up by a globalizing world since the end of the Cold War.

The Military and the Rise of the State

The emergence of the military as a bureaucracy is regarded as indissolubly linked to the rise of the European state as the preeminent form of political organization. Modern understandings of this dual development have been governed by a master narrative of European history that was created in the nineteenth century and which persists in its basic outlines to this day. In the high middle ages, there was no distinction between those who fought and those who governed. The nobility across the continent was fragmented into many thousands of competing entities. All claimed extensive political rights to independent power. The three key expressions of independence lay in their direct claims to the proceeds of taxation, the administration of justice, and the right to resort to force. Military power was organized around the noble household. Nobles sought to retain followers, usually drawn from extended family and local region, who in return for being provided for, rendered military service. Military assistance did not exclude other forms of service, but because it produced the hard foundation of economic and political power, it was predominant. Shrewd power politics could translate into greater economic power, and thus the amassing of greater relative military power. Capitalizing on these opportunities was greatly enhanced by improving economic conditions, especially from the twelfth century onward. Greater agricultural yields invigorated trade and increased the money in circulation which in turn aided social stratification. Military service came to be remunerated with money. A monetary nexus strengthened the traditional bonds of political loyalty between men, but also facilitated the emergence of a new group of nonpolitical men, known as mercenaries, who fought for financial gain.

Here an opening, or even a necessity, was created for the state to emerge and establish order in societies hitherto wracked by conflict and instability. Enterprising rulers, so the classical story goes, could employ apolitical mercenaries to bring the excessively politicized, fractious, and independent-minded nobility of Europe under control. Economic growth created both a pool of manpower and the monetary means to

pay them. What was required was an entity to extract fiscal resources from society and organize, direct, and pay military manpower: the state. The key to extraction lay either in the application of absolutist power politics or in the forging of a deal between rulers and the emerging middle classes who benefited from the order imposed by the rulers' armed force (Tilly, 1990; Porter, 1993). In the states that achieved the greatest degree of domestic stability and permanence, this deal formed the basis for the emergence of representative parliamentary democracies (Downing, 1992). In the older military historical literature the process of gaining the upper hand over the nobility was helped by a structural tactical weakness: the nobility fought on horseback and this was less effective on the battlefield than disciplined massed infantry (Delbrück, 1920–1923). Later accounts accord a prominent role to technology: the cavalry and castle strongholds of the nobility were particularly vulnerable to the new gunpowder weapons that appeared in the fifteenth century (Parker, 1996; Rogers, 1995). Despite the varying emphases, this narrative has a teleological and deterministic quality about it. Economic growth, extractive tax mechanisms, and military technology combined to form a self-reinforcing mechanism that produced a stronger and stronger state. A diversity in economic potential might give certain regions and states structural advantages over others: compare, for example, the success of early modern England and the Dutch Republic with the crisis ridden and faltering Holy Roman and Spanish Empires. None of these political communities managed to escape the basic logic of state formation.

The Nation-State and the Emergence of Civil–Military Relations

State fiscal weakness could exacerbate one key issue: maintaining the loyalty of the mercenary military. Their purely economic motives meant they would serve the highest bidder and, when none was available, they possessed the means forcibly to extract through plunder the resources required for their own upkeep. The 30 Years' War (1618–1648) has long provided the most telling illustration of the damage mercenaries and their enterprising captains could do to a country. In response, states strengthened their fiscal grip to create permanent, standing armies. The provision of stable employment and regular pay went some considerable way in obviating the desire for soldiers to seek alternative employers or self-employment. Payment by itself, however, did not provide total assurance of loyalty in such a dangerous occupation as the military. The undesirability of the work had reinforced the practice of recruitment, often by impressment, from the untrustworthy but abundantly available dregs of society. Retention issues thus remained. But they now shifted from the danger of fighting for another state or plundering for private gain to wholesale desertion. Draconian discipline, lifelong service and physical separation of the soldiery from society through the building of barracks were measures increasingly applied in the eighteenth century to keep armies forcibly together.

In this narrative, the issue of loyalty was only resolved by the nation-state. By the time of the French Revolution, monarchical state structures were sufficiently developed to mobilize, organize and discipline sizable military resources. A sense of national identity, however, massively boosted war-fighting potential by permitting universal conscription of a population loyally committed to national defense and, possibly, international aggrandizement. The introduction of the *levée en masse* in 1793 and the more than 20 years of major war that Republican and Napoleonic France managed to sustain, proved the practical effectiveness of the concept of national loyalty bolstered by the organizational and coercive capacities of the state. National conscription was adopted across Europe, usually hesitatingly by monarchies fearful of revolution, but wholeheartedly by progressive regimes who saw citizen-soldiers as their natural instrument of self-defense. With each new experience of major war – one only has to recall 1870, 1914, and 1939 – the strength of national feeling pleasantly surprised anxious conservative regimes while confirming the expectations of the progressive democracies. The association of armed forces with nation and state has remained a powerful one: the citizenry in the modern state-studded developed world remain overwhelmingly wedded to the idea that proper and effective military forces, whether conscript or volunteer, should be organized by states and function best if motivated by national loyalty.

As states pacified the territories within their borders, wars were turned outward and became 'international.' Two effects are associated with this development. Internationally, the success of nation-states in mobilizing military resources is seen as destabilizing. Sovereign states feared each other's war-fighting potential, but by increasing their own forces in response they only increased the nervousness of others. Thus a shared feeling of insecurity set off a process which resulted in the phenomenon of arms races so clearly observable before the First World War and also during the Cold War. Arms races are often cited as a key factor in causing the First World War (Hermann, 1996). The state system and the ensuing instability (termed 'security dilemma') also inspired the major modern international relations theory of Realism (Morgenthau, 1948; Waltz, 1979). Domestically, the pacification process at first permitted the military to separate itself from the civilian social and political spheres. The military emerged as a distinct state institution and bureaucracy which was dedicated to fighting the state's international wars and which also functioned as the latent, ultimate sanction for maintaining domestic order. The highly influential work *On War* by the Prussian general and military philosopher, Carl von Clausewitz (1780–1831), is predicated on a separation between war (military) and politics (civil). He associates war with a particular activity, force, whose special properties require and justify management by a specialized, carefully trained profession, which acts on behalf of and under the watchful eye of a government and which represents the higher interests of state and nation (Clausewitz, 1976).

Militarism and Total War

Clausewitz also first theorized the phenomenon of escalation in war. If war is about the use of force and if the warring parties want to prevail, then their interaction dictates that each side seeks to use more force than the other. War is therefore in

principle about making the enemy defenseless, preferably through an immediately overwhelming offensive blow. In the later nineteenth century, this understanding of war provided a theoretical underpinning to the practical professional processes of military force generation and operational planning. The distinctive demands of war and war preparation professionalized the military and set it increasingly apart from the civilian sphere. The belief that war had to be fought offensively with overwhelming means conspired with the unstable international environment to make the military 'a state within the state' which threatened civilian political supremacy. Critics coined the label 'militarism' in the 1860s to describe this danger (Berghahn, 1984).

Publicly, the concerns about the political implications of military demands found a focus in the intensely political debates about the issue of conscription. Military professionals (joined by political conservatives who feared the insurrectionary potential of conscripts) favored enlisting regulars and tightly regulated long-serving conscripts. Those fearful of state and society succumbing to militarism argued that a system of universal short-service conscription of citizen soldiers would provide an effective hedge against both domestic militarism and international threats. For much of the nineteenth century, debate raged across continental Europe about the relative merits of the long service professional versus the short service citizen soldier. Blessed by their fortunate geographical position, the UK and US (with the brief interruption of the Civil War) avoided much of these disputes and settled on the compromise of maintaining small regular armies that neither threatened the domestic constitutional order nor imposed the burden of conscription on society.

The experience of the First World War suggested that the military were not simply a state within the state but that it could become the state itself and erase all civil–military distinctions. Harbingers of later developments could be seen in the Austro- and Franco-Prussian wars of 1866 and 1870–1871 during which Prussian chancellor Otto von Bismarck's political preferences famously clashed with Prussian chief of the general staff Helmuth von Moltke's visions of military necessity. By 1917–1918, the German military went much further with the silent institution of a military dictatorship led by Generals Erich Ludendorff and Paul von Hindenburg. Militarizing pressures intruded on the political processes in all countries at war and all governments struggled to varying degrees with maintaining civilian supremacy. The unprecedented mobilization of resources between 1914 and 1918, which included the novel phenomenon of a 'home front,' even threatened to make war an end in itself.

Modern warfare thus raised immediately pertinent questions about the appropriate national constitutional order that would put nation-states in the best position to survive war. 'Total war' was the most radical conceptualization to emerge in the interwar years. Ludendorff, who had led German forces to defeat in 1918, summed up the idea in a popular 1935 book. He took it as given that the international system was marked by the struggle for national survival. The total war for survival required the total mobilization of the nation under the organizational aegis of the totalitarian state. War, Ludendorff wrote, was no longer, as Clausewitz had claimed, a continuation of politics, but politics was a continuation of war by other means (Ludendorff, 1936). Such militarist conceptualizations received significant scholarly input from German academics, like the sociologist Hans Freyer and the constitutional lawyers Carl Schmitt and Ernst Forsthoff. A variation took hold in the USSR. The life-and-death struggle between capitalism and communism dictated a total mobilization of socialist society which took advantage of the unity between party, proletariat and state.

The liberal democracies were not immune to this debate. The term total war, for example, had originated in France during the First World War. However, the democracies feared and resisted militarization of their constitutional order. The US sociologist Harold Lasswell famously warned in 1941 of the danger "that we are moving toward a world of 'garrison states' – a world in which the specialists on violence are the most powerful group in society" (Lasswell, 1941: 455; Friedberg, 2000). Although the democratic powers came out of the Second World War victorious and with their democratic institutions unimpaired, militarist pressures did not abate. The Cold War demanded a permanent state of war preparedness and war, if it were to break out, would likely see the use of nuclear weapons. A real threat remained that the military and war could overwhelm politics in peacetime and in wartime.

The Victory of Democratic Control over the Military and War

This is the dark empirical background against which the seminal modern contributions of the 1950s, 1960s and early 1970s to the analysis of the relationship between military and politics and war and politics should be situated. Samuel Huntington (Huntington, 1957) and Morris Janowitz (Janowitz, 1960) put forward theories on the forms of civil–military relations which most effectively protected against foreign threats without endangering democracy at home. Morton Halperin (Halperin, 1974) and Graham T. Allison (Allison, 1971) analyzed the bureaucratic politics of self-serving government agencies with a view to identifying the best ways of assuring effective policy process. Bernard Brodie and Thomas Schelling worried about nuclear escalation and sought to develop effective strategies of conflict (Freedman, 2003). Alfred Vagts (Vagts, 1959) and Gerhard Ritter (Ritter, 1954–1968), with pertinent personal experience of recent German history, provided historical accounts of the phenomenon of militarism (Berghahn, 1984). To this list of influential authors, one should also add Clausewitz. In the highly successful 1976 English translation of *On War* by Michael Howard and Peter Paret, Clausewitz's work emerged as the seminal text on war as a political instrument (Honig, 2007).

This body of literature was overwhelmingly normatively based. In contrast to the earlier writings on total war, it did not overtly question the desirability of a democratic political order, nor refrain from expressing an abhorrence of war. It also accepted that war possessed a unique and unchanging nature and that the military domain constituted a special sphere of activity, distinct from the civilian and the political. The association of war with escalation fueled fears that nuclear war could perhaps not be managed. For all the assiduous application of novel and imaginative social

scientific methods, the literature struggled to arrive at clear and indisputable findings on how the relationship between (nuclear) war, military, and politics was best managed. While a social scientific consensus proved elusive, a normative consensus emerged that war was better avoided altogether and that it was the role of prudent government policy to promote this.

The military's main role settled on acting as a deterrent. This put the military in a clearly subservient position which, pacified with relatively generous defense budgets and relatively independent management, it on the whole accepted. The growing destructiveness of weapons technology diminished the need for vast armies of conscripts. So the move away from manpower-intensive to technology-intensive war preparation, underpinned ultimately by a relatively cheap nuclear weapons inventory, made the demand for total mobilization redundant. Constitutional pressures thus abated and civil–military relations could to all appearance be adequately managed by the prevailing structures of civilian ministerial oversight and parliamentary budget control.

Preparation for war did not overwhelm the state and its democratic-capitalist order. The needs of modern, technology-intensive war, on the contrary, emphasized the importance of creating a select, specialist profession of 'managers of violence' (in Huntington's subtle adaptation of Lasswell's phrase). Even the instances in which the military were used – Vietnam being the most prominent case in point – it did not threaten the primacy of democratic political control. Despite the emergence of an influential narrative that the Vietnam War was lost as the result of a failure by the political leadership to define clear objectives and accept military professional demands to follow through the logic of war by escalating to success, defeat was an understandable outcome justified by the overall political context which carried the risk of a local war escalating into a global nuclear holocaust. Studies noted that, with respect to the administration of US civil–military relations during Vietnam, 'the system worked' (Gelb and Betts, 1979) and that the military were generally much more cautious than their political leaders about committing to the use of force (Betts, 1977).

Some concern persisted about the self-serving inclinations of military bureaucracies and there was an extensive debate in 1980s about the allegedly pre-emptive and offensive preferences of military-professional organizations (Miller et al., 1991). Official US policy statements also indicated that the military continued to prefer strategies that emphasized a speedy use of overwhelming force (the 1984 Weinberger and 1991 Powell doctrines). By the end of the Cold War, a strong confidence had nonetheless set in that 'mature democracy' had overcome the traditional challenges the military posed to politics. The apparent success of a particular stage in Western political development was further underlined by the contrast with the developing, 'third' world countries where military coups appeared an endemic problem. And even in the 'second,' communist world the totalitarian system seemed to creak and crack at the seams because of the social and financial burden imposed by maintaining a totalitarian system that was locked in an exhaustive arms race with the capitalist West.

Post-Cold War Challenges

Established theories and paradigms did not prepare the social sciences well for the end of the Cold War. Overwhelmingly state-centered, the prevailing discourses were suddenly confronted with wars that seemed to escape from state control and that saw paramilitaries and other irregulars engaging in unconventional and often highly brutal warfare. Some scholars, like Israeli historian Martin van Creveld, argued that a new age of 'low-intensity conflict' had dawned which would spell the end of the state and its conventional military (Van Creveld, 1991). The wars in Yugoslavia, in west, central and east Africa, and later also in Iraq, Afghanistan, and Syria, not to mention the 9/11 terror attacks against the United States, heralded the return of fundamental questions. First, if wars no longer originated with states and if existing states could not get them under control, were they no longer political, but instead driven by fundamentally different considerations and motivations, such as religion or criminality? Or were wars still political but nurtured by different, possibly novel forms of political organization and motivation? Second, had the nature of war changed or was war, as it had so often done in the past, merely changing in outward character? Third, in what ways should the conventional military adapt? Were new forms of organization and new operational approaches to war required, in addition to new equipment and technology? Were established understandings of military professionalism and the strict separation between civil and military, between politics and war, still fit for purpose? In short, not only did the emerging threats exhibit a different relationship between the society sustaining the organizations that conducted war, but the relationship between war, military, and politics in the developed world seemed in line for fundamental reappraisal and reform.

A number of ambitious attempts were made to map the issue of political order and change and the relationship with war. With the power of the state to bind together, control and protect society in doubt, the search was on for a different, or a previously underrated, principle of order. Samuel Huntington presented a controversial theory that claimed that the major 'civilizational' wars of the future would take place along the 'fault lines' between the world's major religions (Huntington, 1997). Azar Gat adopted a revitalized sociobiological approach according to which the extended kinship group of the nation harnessed and civilized the human drives for food, shelter, and procreation which, he claimed, had the potential to create a more peaceful democratic world (Gat, 2006). Philip Bobbitt laid out a future 'constitutional order' marked by the transition from nation-states to frontierless 'market states' which would spawn their own forms of conflict, weapons, and vulnerabilities (Bobbitt, 2008). The heterogeneity of threats and the elusiveness of final outcomes also saw a rise in the popularity of connecting the idea of 'risk society' with war. In an argument similar to Bobbitt's (though perhaps less assured), the bureaucratization of war in states and international law was seen to have created its own subversive enemies which, in response, called forth novel technological and strategic responses (Rasmussen, 2006; see also Coker, 2009). Scholars from the left were drawn to more pessimistic, economistic explanations. They saw the privatization and commercialization that marked the 'globalizing' world mirrored in 'new wars'

in which private enterprisers brutally pursued their own unrestricted commercial gain (Münkler, 2005; Kaldor, 2012).

These accounts continued to subscribe to the idea that war was an instrument, though it was no longer seen as directed by a narrow state-rational, political instrumentality. They did not directly engage with the issue of the nature of war, even though they suggested that war was a very different animal which had changed more than just the color of its skin. The direct debate on the changing nature of war, however, had an unsatisfactory quality about it. There were those, like Colin Gray, who forcefully claimed that war would always remain war (Gray, 2005). Others, like Van Creveld, equally forcefully asserted the contrary (Van Creveld, 1991, 1999). Underscoring the centrality of Clausewitz in this debate, both sides proffered wildly different interpretations of his thinking which generally sidestepped the philosophical intricacies of the phenomenon's definition.

The uncertainty of political change and the unclarity about war's changeability in scholarly treatments likely encouraged a tendency to cast the debates about civil–military relations and the conduct of war in the developed world in familiar terms. The fear of military coups had disappeared, but concerns about a widening gap between civil society and the military re-emerged from time to time, as did questions about the efficacy of political control over the military. Receding threat levels and budget pressures meant that armed forces shrunk significantly in size and, with fewer and fewer exceptions, they did away with conscription and became all-volunteer establishments. Views differed on the effect. Militaries generally favored the move as it accorded with traditional standards of professional self-image. Some, usually without the military, claimed that the military profession had come to distinguish itself in little from civilian professions and was firmly embedded in society's value system and bureaucratic practices. However, others worried that the military's (conservative) political preferences might prove out of step with civil society and lead to misunderstanding and tensions over policy aims and societal support. Conversely, civil (and political) society might misunderstand the special nature of the military's expertise in violence and make demands the application of force could not deliver. The historically unprecedented dwindling in the number of civilians, and especially politicians, with direct experience of war and the military fueled a concern that the military would be used too readily for inappropriate purposes. A technological 'Revolution in Military Affairs' (RMA) created a worry that remotely guided precision weaponry and electronic intelligence gathering might make war 'virtual' and turn it into a spectator sport, taking place far away from a domestic audience's (and even its military's) direct experience (Ignatieff, 2001; McInnes, 2002). In a worst case, military professional self-confidence could enter into an unholy alliance with military technological prowess and popular domestic clamoring for overambitious foreign military adventures (Bacevich, 2005).

Human rights abuses and especially terrorism did tempt grand coalitions of Western countries to launch a series of increasingly ambitious military interventions. Since state failure and collapse was regarded as the main cause of these troubles, nation-building emerged as the main goal of intervention. But the wars in Afghanistan from 2001 and Iraq from 2003 in particular showed how unfamiliar and surprisingly resistant the local human and political terrain was. These long wars called into question the efficacy of established military professional operational approaches to war and political understandings of nation-building. Multipronged 'whole of government' or 'comprehensive approach' seemed an obvious necessity. Lining up government agencies, and a host of nongovernmental agencies, from many countries proved no easy undertaking. Militaries proved highly resistant to taking on new and unfamiliar 'nonviolent' roles.

Ambitious but vague political designs, limited resources and military professional myopia led to civil–military tensions over both Iraq and Afghanistan, as it had in Yugoslavia in the 1990s. Civilian supremacy was never called into question, but neither was military professional independence. The conviction persisted that war and politics were essentially separate spheres and as the military had no business meddling in politics, politicians had no business meddling in war. Closer civil–military cooperation and coordination at the highest levels of government found advocates in academia, with some arguing that politicians needed better to understand the immutable nature of war (Strachan, 2013) and others that "political leaders must immerse themselves in the conduct of their wars" (Cohen, 2002: 206). Failing that the military would fight as well as they could but their actions would take place in a political vacuum and likely lead to unhappy or imperfect outcomes. On the other hand, the military puzzled over the question whether traditional approaches to war still worked (Kaplan, 2013). First Iraq, then Afghanistan followed the same scenario. The magnitude and complexity of the task overwhelmed weak civil–military coordination and an uncertain use of force. The interventions drifted toward a limited focus on development of an indigenous capacity to provide security.

The social sciences seemed to play catch-up for much of the time. Students of state formation could not easily produce immediately applicable strategies for rebuilding failed and collapsed states. Development aid experts faltered in the face of the unprecedented challenge of administering aid in support of an international war effort. Students of bureaucratic politics struggled with proposing ways of overcoming the interagency rivalry between unrivaled numbers of national and international agencies and non-governmental organizations (NGOs). Finally, strategic studies' specialists did little better than resurrect a counterinsurgency literature that had been developed – with decidedly mixed results – in an era of colonial wars when standards of applying force had been rather more permissive. In a recognition that much of the available academic expertise was perhaps too Western-oriented, the US military attempted to recruit specialists from anthropology, a discipline that had hitherto played little part in informing military policy and strategy.

Recent experiences underline once again that the relationship between military, war, and politics is not only a highly complex, but a changeable phenomenon. Contemporary democratic states may offer the best model for achieving relatively harmonious relations between military, government, and society and ensuring that the military act in the interest and

with the support of society. Democratic society may have freed itself from the spectre of direct military involvement in domestic politics. Democracies may also be most sturdy of states, capable of withstanding most serious threats because of their rootedness in popular support and participation. The apparent peacefulness of the domestic democratic order and the international order between democracies make this a tempting model for humanity. The theories which explain why this is so and how this came to be, however, tend to be colored by strong normative convictions and are rooted in a particular justificatory historical narrative. The wars between the West 'and the rest' over the past 20 years illustrate that effective alternative narratives exist. They make intelligible and justify other forms of political organization which are capable of generating, supporting and sustaining militaries that relate to government and society in very different ways from the Western model. Given the strength and stability of the established Western paradigm, coming to analytical terms with these unfamiliar wars fought by unfamiliar polities is likely to be a slow process.

See also: Cold War, The; Ethnic Conflicts; Military History; Military Sociology; National Security Studies and War Potential of Nations; Terrorism; War, Sociology of; War: Anthropological Aspects, Historical Development of; Warfare in History.

Bibliography

Allison, Graham T., 1971. Essence of Decision: Explaining the Cuban Missile Crisis. Little, Brown, Boston.
Bacevich, Andrew J., 2005. The New American Militarism: How Americans Are Seduced by War. Oxford University Press, New York.
Berghahn, Volker R., 1984. Militarism: The History of an International Debate, 1861–1979. Cambridge University Press, Cambridge.
Betts, Richard K., 1991(orig. 1977). Soldiers, Statesmen, and Cold War Crises. Columbia University Press, New York.
Bobbitt, Philip, 2008. Terror and Consent: The Wars for the Twenty-first Century. Knopf, New York.
Clausewitz, Carl von, 1976. On War (M. Howard, P. Paret, Trans.). Princeton University Press, Princeton, NJ.
Cohen, Elliot A., 2002. Supreme Command: Soldiers, Statesmen, and Leadership in Wartime. Free Press, New York.
Coker, Christopher, 2009. War in an Age of Risk. Polity, Cambridge.
Delbrück, Hans, 1990(orig. 1920–1923) (W.J. Renfroe, Trans.). History of the Art of War, 4 vols. University of Nebraska Press, Nebraska, IL.
Downing, Brian M., 1992. The Military Revolution and Political Change: Origins of Democracy and Autocracy in Early Modern Europe. Princeton University Press, Princeton, NJ.
Freedman, Lawrence, 2003. The Evolution of Nuclear Strategy, third ed. Palgrave Macmillan, London.
Friedberg, Aaron L., 2000. In the Shadow of the Garrison State: America's Anti-statism and Its Cold War Strategy. Princeton University Press, Princeton, NJ.
Gat, Azar, 2006. War in Human Civilization. Oxford University Press, Oxford.
Gelb, Leslie H., Betts, Richard K., 1979. The Irony of Vietnam: The System Worked. Brookings, Washington, DC.
Gray, Colin S., 2005. Another Bloody Century: Future Warfare. Weidenfeld & Nicolson, London.
Halperin, Morton H., 1974. Bureaucratic Politics & Foreign Policy. Brookings, Washington, DC.
Hermann, David G., 1996. The Arming of Europe and the Making of the First World War. Princeton University Press, Princeton, NJ.
Honig, Jan Willem, 2007. Clausewitz's on war: problems of text and translation. In: Strachan, Hew, Herberg-Rothe, Andreas (Eds.), Clausewitz in the Twenty-first Century. Oxford University Press, Oxford, pp. 57–73.
Huntington, Samuel P., 1957. The Soldier and the State: The Theory of Politics and Civil-military Relations. Harvard University Press, Cambridge, MA.
Huntington, Samuel P., 1997. The Clash of Civilizations and the Remaking of World Order. Simon & Schuster, London.
Ignatieff, Michael, 2001. Virtual War. Vintage, New York.
Janowitz, Morris, 1960. The Professional Soldier: A Social and Political Portrait. Free Press, New York.
Kaldor, Mary, 2012. New and Old Wars: Organized Violence in a Global Era, third ed. Polity, Cambridge.
Kaplan, Fred, 2013. The Insurgents: David Peraeus and the Plot to Change the American Way of War. Simon & Schuster, New York.
Lasswell, Harold D., January 1941. The Garrison state. American Journal of Sociology 46 (4), 455–468.
Ludendorff, Erich, 1936. The Nation at War (A.S. Rapoport, Trans.). Hutchinson, London.
McInnes, Colin, 2002. Spectator-sport War: The West and Contemporary Conflict. Lynne Rienner, Boulder, CO.
Miller, Steven E., Jones, Lynn, Sean, M., Van Evera, Steven (Eds.), 1991. Military Strategy and the Origins of the First World War, revised ed. Princeton University Press, Princeton, NJ.
Morgenthau, Hans J., 1948. Politics Among Nations: The Struggle for Power and Peace. Knopf, New York.
Münkler, Herfried, 2005. The New Wars. Polity, Cambridge.
Parker, Geoffrey, 1996. The Military Revolution: Military Innovation and the Rise of the West, 1500–1800, second ed. Cambridge University Press, Cambridge.
Porter, Bruce D., 1993. War and the Rise of the State: The Military Foundation of Modern Politics. Free Press, New York.
Rasmussen, Mikkel V., 2006. The Risk Society at War: Terror, Technology and Strategy in the Twenty-first Century. Cambridge University Press, Cambridge.
Ritter, Gerhard, 1969(orig. 1954–1968 (H. Norden, Trans.). The Sword and the Scepter: The Problem of Militarism in Germany, 4 vols. University of Miami Press, Miami, FL.
Rogers, Clifford J. (Ed.), 1995. The Military Revolution Debate: Readings on the Military Transformation of Early Modern Europe. Westview, Boulder, CO.
Strachan, Hew, 2013. The Direction of War: Contemporary Strategy in Historical Perspective. Cambridge University Press, Cambridge.
Tilly, Charles, 1990. Coercion, Capital, and European States, AD 990–1990. Blackwell, Oxford.
Vagts, Alfred, 1959. A History of Militarism: Civilian and Military, revised ed. Meridian, New York.
Van Creveld, Martin, 1999. The Rise and Decline of the State. Cambridge University Press, Cambridge.
Van Creveld, Martin, 1991. The Transformation of War. Free Press, New York.
Waltz, Kenneth N., 1979. Theory of International Politics. McGraw-Hill, New York.

Mill, John Stuart (1806–73)

Iain Hampsher-Monk, University of Exeter, Exeter, UK

Published by Elsevier Ltd.

Abstract

J.S. Mill's main contributions to social science lay in three areas: political economy, political philosophy, and the philosophy of social science. The major works identified with these three fields are *Principles of Political Economy*, *On Utilitarianism*, *On Liberty*, *Considerations on Representative Government*, and his *Logic*. Although educated in classical utilitarianism – the view that the criterion of moral and political rectitude was the pursuit of the greatest happiness for the greatest number – Mill is famous for his critical adjustments to it, expressed particularly in his essay *On Liberty*, where he advocated political arrangements that maximize the freedom of the individual. In many of his works he stressed and explored the significance of historical processes – sensitivity to which had been absent from his utilitarian predecessors' aspiration to create a purely deductive social science, and increasingly from the liberal political economists who would claim his legacy. Mill's iconic status as a liberal has made his intellectual legacy a site of fierce ideological contestation.

John Stuart Mill was born in London, on 20 May 1806. His precocious education – famously described in his *Autobiography* (1873) – although commonly held to have inculcated him with utilitarian principles, and certainly intended by his father to prepare him for a career as a social and political thinker and reformer, was grounded in the classics. His father started him with Greek at age three, and his education was much wider than that of most modern social scientists. He spent a year in France in 1820–21 before beginning a career in the civil service, following his father's footsteps at India House, London Headquarters of the British Indian civil service.

The young Mill was introduced by his father to the Philosophical Radicals, a largely extra-parliamentary grouping of Benthamite utilitarians on the more radical wing of those seeking parliamentary and administrative reform. However, his underlying philosophical position was undergoing a reappraisal. In his *Autobiography*, Mill describes an emotional crisis at age 20 that had important intellectual dimensions and was finally resolved only following the death of his father in 1836. Establishing some distance from his father and his upbringing, he claimed, liberated his thinking from Benthamite utilitarianism, his differences with which were first outlined in a series of articles on 'The Spirit of the Age' in the periodical *The Examiner*. An important aspect of his new position was the ideal of self-cultivation (which his own *Autobiography* sought to encourage). He was inspired in his exploration of the imagination and emotions by his relationship with Harriet Taylor, a married woman whom he met and fell in love with in 1830. To her he also credited considerable influence in the intellectual development of his arguments. Despite this awakening of feeling, Mill rejected intuitionism as a basis for philosophy and was committed to extending the experiential methods of the natural sciences to the social. The new dimensions in his thinking did, however, stress the limitations of the associationist psychology that had informed his own upbringing and introduced ideal-regarding criteria – standards that looked to the fulfillment of certain ideal principles – into the predominantly want-regarding and consequentialist character of his inherited utilitarianism. Furthermore, Mill increasingly recognized the thick historical texture required for an understanding and appraisal of social and political institutions. Through his father's Scottish education and his teacher Dugald Stewart, Mill had access to the thinkers of the Scottish Enlightenment – David Hume, Adam Smith, William Robertson, and Adam Ferguson – who had developed sophisticated accounts of what would today be called historical sociology, then termed philosophical history. Mill's interest in this can be seen as early as his essay *Civilisation* (1836).

There were new influences that he also acknowledged. He devoted an essay to S.T. Coleridge sympathetic to the latter's concern to institutionalize historically acquired learning and cultivation in a national clerisy – a kind of secularized church establishment. Alexis de Tocqueville's two volumes of *Democracy in America* were reviewed by Mill for the *Westminster Review* in 1835 and 1840 and had a profound effect on his thinking about the need to manage the political effects of the historical movement to more democratic societies. Auguste Comte, whose sense of history as a process of rational amelioration survived (albeit frailly) in Mill, although he rejected Comte's more elitist policy implications. Impressed as Mill was by Tocqueville's account of local democracy in American townships, and his claim that democratic culture was the inevitable future for European commercial polities, Tocqueville's worries about egalitarian mediocrity reinforced Mill's concerns, already voiced in *Civilisation*, about the stifling of individuality, energy, and dissent in the emergent mass societies. Although America had abolished an aristocracy of taste or intellect, this had not eradicated deference, which Mill now saw being abjectly paid to 'public opinion (which) becomes in such countries a species of religion and the majority is its prophet.' A continuing concern for Mill, henceforth, was, as he put it in his review of Tocqueville's work, 'that there should exist somewhere a great social support for opinions different from the mass.'

Mill's commitment to radical politics itself faltered at times following disillusionment with the Radicals' Whig allies in the Reform Parliament of 1832. The major focus of his political activity at this time became a new radical periodical, the *London and Westminster Review* (1835). It was from the early 1830s too that he began work on his *A System of Logic*, book six of which is devoted to the logic of the moral, i.e., social sciences.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61092-9

Mill recognized that since the application of the principle of utility required an assessment of consequences, a knowledge of social and historical context and process was needed to make such assessments. However the utilitarianism of Bentham and Mill's father, James, had attempted to model, understand, and explain political phenomena deductively, drawing out the implications of utilitarian psychology, rather than attending to the specific cultural and historically contingent features of society, an omission for which the historian Macaulay famously had lambasted Mill senior's aprioristic, deductivist *Essay on Government*. Mill's exposure to the tradition of historical thinking in the Scottish enlightenment of the eighteenth century, theorizing as it did the effects of large-scale social changes, for example, economic and social mobilization, on human reasoning processes and ultimately on social conventions and forms of political association, meshed well with the concerns of his more recent contemporaries about the emergence of democratic society. Mill's *Logic* analyzes the methods appropriate to the different branches of science – Mill's view being that the only secure ground for deductivism is the assured existence of the uniform operation of physical causes, an assurance that is not actually present in the moral, as we would say social, sciences. In Book VI *On the Logic of the Moral Sciences*, Mill discussed the extent to which the human sciences diverged from the natural. His discussion has become a point of departure for much teaching and theoretical reflection in the social sciences ever since.

Outlining the possibility of a social science, Mill subjected to criticism two alternative methods that he called the geometric and the chemical methods, by analogy with sciences already treated earlier in the *Logic*.

The geometric method that was pursued by his father in his ill-fated *Essay on Government* was an attempt to deduce the optimum form of government from certain posited axioms about the selfish motivations of human beings. This cannot work because deductive operations presume a uniformity that cannot cope with the quintessential sociopolitical circumstance of multiple, simultaneous, and conflicting causes impinging on a given social or institutional context. Geometrical deduction is not even yet mechanics, which, crude though it is, can at least cope with this multiplicity of forces.

The chemical method is the attempt to infer causal laws from the interaction of complex entities at (at least) one remove from their underlying physical causes. Mill's thought here seems to be that if real causes take place at the level of physics, 'chemical' phenomena represent epiphenomena that are related *in some systematic way* to those (hidden) underlying real physical causes, and that by a complex of observation and deduction, the chemical properties of physical elements might be derivable at this level. The limitations of this method are that it can only apply where the observed phenomena are *invariantly* related to underlying causal properties – as in the case of chemical compounds whose properties relate systematically to those of their constituent elements. Since the underlying causes of social phenomena are many and various and not elemental or invariant in their operation, it is not possible to treat political facts in the manner of chemical facts nor to induce from experiment or observation any reliable causal regularities. Mill instances the attempt to answer the question whether regulating imports is conducive to economic prosperity or not. Even supposing two apparently similar polities pursuing different policies in this respect would not constitute a test case, argues Mill. The reason is that different trade policies are not ultimate 'properties of Kinds' but presuppose a whole raft of differences in opinions, habits, and so on, which might themselves influence the success of the economy, with or without free trade.

For Mill, 'true' causation operated only at the level of local physical events. A human science that sought to base itself on causal generalizations, therefore, must ultimately be grounded in the discovery of physiological causes. However, he recognized these were too far below the level of the visible 'general circumstances of society' that he sought to understand, for the latter, at least in the present state of knowledge, to be unpacked in terms of the former. In the meantime, he thought, empirical induction about the effects of circumstance on human character formation (which he called 'ethology'), although it does not yield full causal knowledge, could produce 'empirical laws' of greater or lesser range and accuracy, and which might certainly be useful presuppositions against which to consider the likely consequences of policy initiatives. Such empirical laws yielded what he called (following Comte) both 'uniformities of coexistence' (social statics) and 'uniformities of succession' (social dynamics). Although neither the deductive nor the inductive method was sufficient in itself, such confidence as we can have in deductive conclusions can be derived, Mill thought, from "collating the conclusions of the ratiocination either with the concrete phenomena themselves, or when this is unobtainable [as in the case of social science], with their empirical laws." This combination of deductive and inductive reasoning brought to bear on even very complex phenomena, whilst it could not achieve the level of predictive accuracy available in the natural sciences, could prove extremely useful to the "wise conduct of the affairs of society" (*Logic* VI, ix, §§1,2). It might even, in time, be linked with or 'resolved into' the real causal laws at the level of physiology, so effecting a complete union of the social with the physical sciences. Yet even if this were not fully achieved, there was, Mill insisted, the possibility of constructing a science of society or what he called 'that convenient barbarism' – sociology.

Following his *Logic*, which was a considerable publishing success, Mill returned to political economy. His father had taught him economics in his childhood, and he had even published in the subject at age 16. His *Essays on Some Unsettled Questions in Political Economy*, written in 1830–31, now found a publisher (1844), and he settled down to the composition of the *Principles of Political Economy, with Some of their Applications to Social Philosophy* (1848). Not only (as the subtitle indicates) did he regard economics as inseparable from the rest of social philosophy, he also regarded the relationship between economics and social theory more widely considered as being the most important part of it. Whilst composing it, Mill was painfully aware of the Irish potato famine, and the impact of economic orthodoxies on its course. Seeking to avoid specific polemical issues, Mill was nevertheless concerned to stress that political economy was "far from being a set of maxims and rules, to be applied without regard to times, places, and circumstances" (*Hansard* v.190, 1525). Thus, although in pure economics he regarded himself as a follower of Ricardo, he distanced himself from those who took Ricardo to have

established a purely deductive school. Historians of economic thought disagree as to precisely how to characterize this relationship and who these economists might be (see Hollander, 1985: vol. II, pp. 914ff.). It is clear, however, that Mill regarded good economic method to combine both deductive and inductive operations, and in policy stressed the importance of knowledge of local conditions – such as those in Ireland or India – in applying that method. Economic science, he thought, operated under those limitations applying to the social sciences generally, outlined in the last volume of the *Logic* – deductions were only reliable if the generalizations that generated them were the result of physical causes, and even then only if other causal effects did not intervene. Neither of these conditions, in his view, obtained in economics any more than in the rest of social science. This also informed his view of the limited role that mathematical or formal modeling might be expected to play in economics – for without certainty or quantification such precision was likely to be misleading.

Mill's insistence on subordinating economic analysis to policy consideration is clear in his famous chapters on socialism in the *Principles*. These underwent considerable development in the first three editions (1848, 1849, 1852), under the influence of the European revolutions of 1848, which had reopened debate on the subject, and of Harriet Taylor (whom he had married in 1851 following the death of her husband). Originally skeptical about cooperation, Mill now discussed the ideas of Fourier as a supposedly superior synthesis of the ideas of Robert Owen and Saint-Simon, which overcome earlier worries about the lack of incentivization in cooperative schemes. He came to see the emergence of cooperative forms as both an evolutionary possibility within capitalist economies and – once population could be voluntarily controlled and inequalities moderated through inheritance tax – a desirable social form in which individuality and moral progress, far from being under threat, could flourish better than in conventional economies.

Such themes integrate with the political writing to which Mill returned, at the end of the 1850s, producing a trio of important essays, long in the making: *Utilitarianism*, *On Liberty*, and *On Representative Government*. These worked out his concern to articulate a conception of progress consistent with utility, as a part of which he sought to support diversity of opinion and to protect social and political institutions from the effects of mediocre uniformity – whether deriving from the middle classes' social conformism or a socialist-inspired working-class movement.

Utilitarianism articulated a modified defense of the principle of utility into which Mill introduced significant modifications. The first was the claim that in addition to the various dimensions of utility analyzed by Bentham, and contrary to the latter's claim (made notorious by Mill himself) that 'pushpin was as good as poetry,' it was possible to distinguish higher from lower qualities of pleasure. Amongst the higher pleasures were those of altruism, the cultivation of the higher feelings and of the intellect. Quality as well as quantity was to be given weight in seeking to apply the utilitarian criterion that actions were 'right in proportion as they tend to promote happiness.' This constituted a rejection of Bentham's insistence on the integrity of individuals' subjectively expressed preferences and introduced what many have seen as an elitist, ideal-regarding element into utilitarianism. Mill's second innovation was his claim that the principle of utility should be deployed as a criterion of moral or social principles, practices, or institutions, and *not* applied directly to regulate individual actions. The distinction has since been refined and elaborated as that between rule utilitarianism and act utilitarianism, although Mill does not make the distinction in these terms: although justice is grounded in utility, direct appeals to utility cannot be used to overthrow (any but?) the hardest cases.

Classical utilitarianism was premised on the integrity of each individual's experience of utility. Each was to count for one and no more than one, and each was the best judge of their own utility. Mill's two modifications, perhaps worryingly, allow privileged judgments to be made, capable of overturning the subjective claims of the individual; but they also allowed judgments about progress to be made that did not rest purely on increasing the quantities of pleasure.

Issues about progress underpin what is perhaps Mill's most famous work – his defense of liberty. In *On Liberty*, Mill formulated what he disarmingly called a "very simple principle," one he had introduced in all but name, in his *Principles of Political Economy*. There he remarked that "there is a circle around every human being, which no government, be it that of one, the few or the many, ought to be permitted to overstep" (*PPE*, Vxi, §2). Mill's principle was that "the sole end for which mankind are warranted, individually or collectively, in interfering with the liberty of action of any of their number, is self-protection, that the only purpose for which power can be rightfully exercised over any member of a civilised community, against his will, is to prevent harm to others. His own good, either physical or moral, is not a sufficient warrant" (Acton, 1967: 72–73).

Mill's principle has become an icon for liberals and occasioned a huge literature, both academic and polemical. The academic controversy has centered around two issues: whether Mill's defense of the principle of liberty was consistent with his claimed loyalty to the principle of utility, and whether his defense of the integrity of the sphere of individual liberty is ultimately coherent.

Mill denied that his defense of liberty involved any appeal to abstract right distinct from utility. The latter remained the 'ultimate appeal on all ethical questions.' It was, though, 'utility in the largest sense grounded on the permanent interests of man as a progressive being' to which one had to appeal, and not the mere maximization of existing preferences. This invoked his attempt in *Utilitarianism* to introduce the qualities and refinement of pleasure and not just its quantity as a utilitarian criterion. Briefly, the connection Mill makes between the defense of liberty and 'enlarged utility' is that utilitarian progress (beyond mere quantity) presupposes innovation, and innovation presupposes (though it is not guaranteed by) liberty. Liberty of thought and speech could generate and promulgate new ideas and test old ones, avoiding the static mediocrity that he feared so much. Liberty in the sphere of private action enabled experiments in living from which might emerge new and more beneficial social forms, new forms of domestic partnership and cooperative production in the sphere of the economy being two such examples. In this view, liberty

could be justified in instrumental terms as a means to a social progress defined, or definable, with reference to utility ('in the largest sense…'). However, Mill's enthusiasm for liberty – redoubled in the closing passages of *On the Subjugation of Women* – his celebration of what he disarmingly calls the 'intrinsic worth of individuality' and the cultivation of the personality as a 'noble and beautiful object of contemplation' – appeals to an autonomous aesthetic ideal that has seemed to many commentators to go well beyond the instrumental role claimed for it in his more programmatic statements.

Mill also had great difficulty sustaining the boundaries of the 'sphere of private action' that he attempted to mark by distinguishing 'self-regarding actions,' which were not to be collectively interfered with, and 'other regarding actions,' which were, in principle, candidates for legislative or social restraint. The viability of this in turn involved distinguishing between actions that merely 'affect' and those that 'affect the interests of' others. But the notion of interests, whilst clearly different from wants or preferences, is a notoriously difficult one to deploy – particularly for liberals – without establishing paternalistic claims over individuals' preferences.

Mill's working through of this principle is for some an object lesson of the sensitivity required in the application of philosophical principles to areas of policy, for others revealing of a residual paternalism.

Considerations on Representative Government, Mill claimed, embodied the principles to which he "had been working up during the greater part of my life." It was certainly, amongst other things, an attempt to show how his commitment to liberty as a part of a revised conception of utility could be embodied in political institutions. It revealed an acute historical sensitivity to the broad circumstances that made representative government possible, to the generation of supportive dispositions in the population and to those processes that might imperil it – the race, as he put it, between democratization and education. Representative government is fixed firmly within the presuppositions of a progressive, national, metropolitan state, and for all Mill's celebration of diversity, he evinces little enthusiasm for sustaining the contributions to diversity of relict regional or national cultures such as those of the Celtic-speaking nations.

Once representative government is possible – which is not the case for all stages of human development – Mill considers there are two criteria by which government should be judged: efficiency and education. The first seeks the optimum 'use of the existing good qualities' in the population; the second seeks to augment the virtuous qualities in the population at large. These can be seen as a version of two principles widely seen as central to many nineteenth-century thinkers: stability and progress. Much of the work is a consideration of various political institutions in light of the trade-off between these two principles. Two examples may suffice: those of voting and local government.

Like all utilitarians, Mill rejected abstract or natural rights, and did not derive the vote, any more than liberty, from such principles. The right to vote was for him grounded in the important necessity of ensuring that every individual – even, as he later publicly argued, women – had an institutional expression through which they could safeguard their interests. But this did not entail the privacy, equality, universality, or unlimited domain of the franchise. All of these were to be assessed in consequentialist terms, through their contribution to utility. Consequentialist judgments are sensitive to the empirical circumstances of that which is being judged. Mill acknowledged that privacy of the ballot box might once have been necessary to avoid undue pressure from employees, landlords, or social superiors, but the danger now, he thought, came from individuals regarding the vote as an expression of their selfish wishes. Public voting would, he thought, have the advantage of requiring voters to justify their decisions to their fellows, incentivizing them to decide on grounds of public reason rather than private interest. Consequentialist considerations also informed his thoughts about the electorate. Though all adults have a prima facie claim to the vote for protection, equal voting is not necessarily conducive to the best outcome – part of which is to make the best use of the talents available. Mill advocates an educational test (at least partly to encourage self-improvement), disfranchising those receiving welfare benefit, and proportional representation and plural voting for university graduates and professional classes. This ensured that minority and educated opinion would at least be heard in Parliament, exercising benign influence on less-informed representatives, and mitigated the potential for majority tyranny. For similar reasons he opposed the mandating of representatives. Finally, Mill regarded representative bodies as incompetent to draft or engage in detailed amendment of legislation, which should be left to commissions of experts. Mill thought such considerations enabled a distinction to be drawn between 'True and False Democracy.' In True Democracy, the role of the People is 'strong enough to make reason prevail but not strong enough to prevail against reason.'

Mill's discussion of local government stresses the educative role it can play in giving the inexperienced scope for political participation in a limited sphere where they cannot damage great national interests. Devolution must not, however, be so great as to exclude the interest of at least some of the educated from whom lessons might be learnt, nor, he thought, should it ever escape the superintendence of the national government.

Mill sat as Liberal MP for Westminster – a notoriously radical seat since the 1780s – in the 1865–68 Parliament, his candidacy provoking huge demand for cheap popular editions of his *Political Economy*, *Liberty*, and *On Representative Government*. True to the spirit of his observations in *On Representative Government*, he refused to pledge – or indeed campaign – prior to the election. The most famous cause he espoused during that period was to seek – albeit unsuccessfully – to amend the 1867 Reform Bill so as to grant votes to women. This became a major cause for him during the rest of his life, and his *The Subjection of Women* (1869) has become one of the canonical texts in the study of women's political emancipation.

Harriet Taylor had died in 1858 in Avignon, where Mill subsequently spent half of each year until his own death on 7 May 1873. Harriet's daughter, Mill's stepdaughter, Helen, acted as Mill's secretary and arranged the publication of a number of his papers after his death, including the *Autobiography* and the *Chapters on Socialism*. In the latter, Mill pursued themes explored in the later chapters of his *Principles of Political Economy*, although with even greater open-mindedness as to the possibility of one form or another of socialism proving itself superior. Although he thought revolutionary socialism and a centralized economy both unjust and 'chimerical,' and that even temperate, though extensive, forms of socialism were 'not

available as a present resource,' he recognized the high moral values implicit in some of these and was firmly of the opinion that various socialist schemes 'have a case for a trial.' For him though, the criteria to be applied in such a trial were essentially those of the development of human character. Mill recognized the irresistibility of the advance of democracy, the implications of which were still working their way through society and which, when they had done so, would penetrate to 'the very first principles of existing society' – in which he included private property. Yet, whilst accepting this, his concern to balance values that the mass of the people might not yet share is one of the central tensions in his work, informing what has come to be seen as his central preoccupation: the defense of liberty.

Mill's *Principles of Political Economy* enjoyed huge – almost canonical – status in the middle of the nineteenth century, down to the marginalist revolution associated with Jevons and Marshall. It is now no longer of technical interest to economists but remains an important chapter in the history of economic, and social and political, thought. That part of his *Logic* devoted to the philosophy of the social sciences remains an important statement of a particular approach in sociology. Mill's political writings have a more ambiguous legacy. Political science now takes forms – and has preoccupations – Mill could not have foreseen; as a result, his *Representative Government* now has a less practical character than he might have hoped for it. *On Liberty*, however, is still often read in the spirit in which it was written – as a statement of moral and political philosophy to be considered as a success or failure in terms of the arguments advanced there. Perhaps what would have been most satisfying of all to Mill, his own character as a public moralist, the ideal of personal character formation, and his complex and nuanced treatment of political nationality have reemerged recently as an acknowledged central preoccupation of his work.

See also: Comparative Studies: Method and Design; Consequentialism (Including Utilitarianism); Emergent Properties; Enlightenment; Freedom: Political; Individuals in Society: History of the Concept; Liberalism: Historical Aspects; Liberalism: Political Doctrine and Impact on Social Science; Natural Law; Political Economy, History of; Utilitarian Social Thought, History of.

Bibliography

Acton, H.B., 1967. A System of Logic. Everyman, London (new impression of the eight ed. Utilitarianism, On Liberty and Considerations on Representative Government. London, 1910 and often reprinted).
Berger, F.R., 1984. Happiness, Justice and Freedom: The Moral and Political Philosophy of John Stuart Mill. University of California Press, Berkeley, CA.
Burns, J.H., 1957. J.S. Mill and democracy 1829–1861. Political Studies 5 (2), 158–175 and (3), 281–294.
Collini, S. (Ed.), 1989. On Liberty with the Subjection of Women and Chapters on Socialism. Cambridge University Press, Cambridge, UK.
Collini, S., Winch, D., Burrow, J., 1983. That Noble Science of Politics. Cambridge University Press, Cambridge, UK.
Cowling, M., 1963. Mill and Liberalism. Cambridge University Press, Cambridge, UK.
Cipaldi, Nicholas, 2004. John Stuart Mill: A Biography. CUP, Cambridge.
Eisenach, E.J., 1998. Mill and the Moral Character of Liberalism. Pennsylvania University Press, University Park.
Grey, J., 1983. Mill on Liberty: A Defence. Routledge and Kegan Paul, London.
Hamburger, J., 1965. Intellectuals in Politics: John Stuart Mill and the Philosophical Radicals. Yale University Press, New Haven, CT.
Hollander, S., 1985. The Economics of John Stuart Mill, 2 vols. Blackwell, Oxford, UK.
Laine, M. (Ed.), 1991. A Cultured Mind: Essays on Mill Presented to J.M. Robson. Toronto University Press, Toronto.
Leavis, F.R. (Ed.), 1967. Mill on Bentham and Coleridge. Chatto and Windus, London.
St Packe, J.M., 1954. The Life of John Stuart Mill. Secker and Warburg, London.
Rees, J., 1985. John Stuart Mill's on Liberty. Oxford University Press, Oxford, UK.
Robson, J.M. (Ed.), 1963/1991. Collected Works of John Stuart Mill. Blackwell, Toronto.
Robson, J.M., 1968. The Improvement of Mankind. Blackwell, Toronto.
Ryan, A., 1975. John Stuart Mill. Routledge and Kegal Paul, London.
Schneewind, J.B. (Ed.), 1968. Mill: A Collection of Critical Essays. Macmillan, London.
Schwarz, P., 1968. The New Political Economy of John Stuart Mill. Weidenfeld and Nicolson, London.
Skorupski, J. (Ed.), 1998. Cambridge Companion to John Stuart Mill. Cambridge University Press, Cambridge, UK.
Stillinger, J. (Ed.), 1969. Autobiography. Oxford University Press, Oxford, UK.
Thompson, D., 1976. John Stuart Mill and Representative Government. Princeton University Press, Princeton, NJ.
Ten, C.L., 1980. Mill on Liberty. Oxford University Press, Oxford, UK.
Varouxakis, Georgios, 2002. Mill on Nationality. Routledge/PSA, London and New York.
Winch, D. (Ed.), 1985. Principles of Political Economy (Books IV and V). Penguin, Harmondsworth, UK.

Millennialism

Bryan R Wilson, University of Oxford, Oxford, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is a reproduced from the previous edition, volume 14, pp. 9881–9885, © 2001, Elsevier Ltd.

Abstract

Millennialism is defined as a general term synonymous with millenarism and millenarianism and chiliasm. Its provenance and transmission via Judaism precedes an outline of the history of Christian millennialism in which premillennialism, belief that the Second Coming will precede the millennium, and postmillennialism, the thesis that the millennium will result from human effort before the Second Coming, are distinguished. Militant millennialism in Europe is exemplified, preceding discussion of the nineteenth century growth of mass movements. Predictions of dates are exemplified from the history of the Seventh-day Adventists and Jehovah's Witnesses. Examination of various sects indicates that the relative deprivation thesis, which postulates that religious belief is a compensation for various forms of felt deprivation, does not account for the appeal of millennialism to its entire social constituency. Divergent responses to the problem of cognitive dissonance, experienced when prophecy fails, are compared with reference to the Seventh-day Adventists, Jehovah's Witnesses, the Exclusive Brethren, and the Catholic Apostolic Church. Reference is made to the incidence of millennialism in other major religious traditions. Examples of third-world so-called millennialism raise the question of whether these phenomena are autochthonous (that is, indigenously inspired) or result from Christian influence. Mass movements of a magical orientation are compared with militant millennial cults, and the role of millennialism in forging new bases of social identity and in raising group consciousness is indicated.

Millennialism (derived from the Latin term, millennium, which means literally 1000 years) is the belief that a prolonged period of bliss is a future prospect, usually expected to commence after catastrophe which will end the present dispensation and perhaps the world itself. Salvation and survival into the new era will be confined to those who believe this prophecy and conform to certain stipulated demands of faith and morals. Millennialism has inspired numerous movements throughout the world and given rise to outbursts of intense social unrest, but has also created new patterns of social consciousness.

The Term Millennialism

The terms millennialism, millenarism, and millenarianism are more or less synonymous, although some writers (Harrison, 1979) have sought to distinguish millennialism as scholarly theological concern with prophetic exegesis, and millenarianism as applicable to social movements preaching the imminence of the millennium. This distinction, however, is not observed generally. The term chiliasm (Greek *chilioi* – a thousand) is also employed in reference to social manifestations of belief in the early outworking of prophecy.

Historical Provenance

Millennial ideas arose in Zoroastrianism which exerted an influence on Judaic eschatology in which it became a strong theme. Jewish apocalyptic literature linked the millennium to the expectation of a returning messiah, a king to save his people, and this association informed the expectation among early Christians that Christ would soon return to earth to rule over his kingdom for 1000 years. The concept gradually has acquired wider application to contexts other than the Judeo-Christian, and is used loosely to refer to a long period of time rather than specifically to 1000 years, but it is in Christianity that millennialism is most fully, albeit controversially, articulated.

Development in Early and Medieval Christianity

In Christian history, an assortment of New Testament texts, especially Revelation 20, have been cited alongside the Old Testament books of Isaiah, Daniel, and Ezekiel to provide both a sequence of events and allegorical reference points as a basis for a chronology for the prophesied millennium. Some early Church fathers and various heretical movements expected an imminent literal outworking of prophecy, and looked forward to indulgences and pleasures which they had hitherto denied themselves. Augustine changed Christian thinking by regarding the contemporary epoch of the Church to be the millennium, at the end of which Christ would return and establish the conditions of eternity. Millennial ideas did not disappear, however, and found expression in the theories of Joachim of Fiore in the twelfth century and in various heretical movements, and they received a new impulse from sects that arose in Reformation times.

Pre- and Postmillennialism

For later Christians, Daniel Whitby (1638–1726) formulated the position which became widely accepted, that the advent would occur at the end of the millennium rather than before, thus establishing a distinction between postmillennialists (who believed that the advent would follow the millennium) and premillennialists. Postmillennialists held optimistic beliefs about social progress and the diffusion of the Christian message through missionary activity, thus paving the way for the return of Christ. The designation 'millennialist' was henceforth applied to premillennialists, those expecting the imminent second advent of Christ. Among themselves, premillennialists differed about other matters, among which

were: whether the tribulation alluded to in I and II Thessalonians would occur before or after the 'rapture' of the saints (i.e., their being caught up in the clouds to join Christ); and whether the millennial kingdom would be established on earth (as believed by Jehovah's Witnesses) or in heaven (as maintained by Seventh-day Adventists). Most millennialists have been 'mortalists,' denying that man possessed an immortal soul.

The Development of Premillennialism

Millennialism has given rise to numerous episodes of fanaticism. After the Reformation, the 'open Bible' policy of Protestants gave millennial expectations common currency. Many movements inspired by expectations of the early fulfillment of prophecy have been of short duration, particularly those with adherents who took it upon themselves to take up arms to help bring the advent and the millennium about. Such were the Anabaptists who occupied the city of Münster in Westphalia in 1534, and the Fifth Monarchy Men who challenged the secular authorities in the 1650s in England.

Dating in Prophetic Exegesis

Millennialists have not depended on the secular calendar in affirming when the millennium would begin, but have often set dates. Only the most proximate dates have inspired active movements. Expected dates have been calculated by the choice of specific past events as the departure point for the outworking of chronology and the application of numerological principles. Those calculations concerning the fulfillment of prophecy generally have not coincided with the completion of the first or second millennia of the western calendar. Not all movements have committed themselves to predicting specific dates, and others have done so guardedly in recollection of the repeated scriptural warnings that Christ would come 'like a thief in the night.' Jehovah's Witnesses have been perhaps the most persistent in setting dates for the fulfillment of prophecy (having, formally or informally, at different times and successively harbored keen expectations of 1874,1914, 1925, and 1975). The Seventh-day Adventists, after the disappointed hopes of William Miller's predictions for 1843 and 1844, have abandoned date-setting, while the Christadelphians, although much exercised on the subject and disposed, especially in their early years (1840s and 1850s), to look for signs in contemporary political events to match biblical prophecies, have for a long time disavowed the possibility of acquiring accurate foreknowledge of the time of Christ's return.

The Size of Modern Millennial Movements

Whilst postmillennial orientations stimulated missionary and revivalist activity in the eighteenth century, particularly in North America, the pessimism implicit in premillennialism previously had inspired movements which operated in considerable tension with the state and the wider society, and these hostile movements usually were small and localized. Growth of literacy and improved communications in the nineteenth century allowed sectarian premillennialists to organize much larger, enduring, and widespread movements, some of which, by the end of the twentieth century, had followings counted in millions (e.g., the Seventh-day Adventists, Jehovah's Witnesses). As the need for better accommodation with modern society and the agencies of the state grew, so these sects became less aggressive and their insistence on the imminent end of the existing dispensation less strident.

The Social Constituency of Millennialism

It has been a general assumption that those attracted to millennial sects are not only pessimistic about general social conditions, but are likely to be drawn from disinherited social groups, and the relative deprivation thesis has been frequently invoked to explain their appeal. As a generalization this may be warranted, but by no means all of those attracted to millennialism have formed their views in socially or culturally disadvantaged circumstances. Prominent Protestant thinkers – Joseph Mede, the theologian, and the scientist, Isaac Newton, among them – were devoted to numerological exegeses of biblical prophecy. The initial leaders of the strongly adventist (premillennial) Catholic Apostolic Church included several Anglican priests and Church of Scotland ministers, and all but one of that church's 12 apostles were either lesser aristocracy or professional men. Several lay supporters were, or became, MPs or higher civil servants. The originators of the contemporaneous movement devoted to premillennialist prophetic exegesis (later known as the Plymouth Brethren), were clerics and ordinands, and local leaders were often men of some substance. Nineteenth century millennialism had an appeal which extended beyond the working class.

When Prophecy Fails

Millennial movements all face the prospect of the failure of prophecy, and examination of their history reveals diverse responses to this problem. Gager, following Festinger's study of a contemporary adventist cult (*When Prophecy Fails*, 1956) has contended that the expansion of early Christianity was attributable, at least in part, to the failure of millennial expectations (Gager, 1975). Believers sought to overcome the cognitive dissonance of failed prophecy by persuading others that they were not wrong: the next best thing to being proved right was to persuade others to join you and to accept your rationalizations for the error. The evidence on early Christianity is perhaps too sketchy to substantiate this contention, while Festinger may have studied too small and ephemeral a movement (in which he and his team constituted too influential a group) to warrant the comparison.

Two Responses to Failure

Even so, some millennialist movements have triumphed over repeated failure of prophecy, and have done so by various different strategies. Perhaps the most successful of contemporary millennial bodies, the Seventh-day Adventist Church, has flourished by cultivating an encompassing program of educational, dietetic, and therapeutic welfare in which to engage members' energies. This policy originated early in the movement's history, and has expanded despite the apparent inconsistency of devoting resources to welfare in a world due soon to

be overturned by the cataclysmic events supposedly foretold in the Scriptures. In contrast to this many-sided strategy to occupy the faithful, Jehovah's Witnesses, whose theology teaches that God's Kingdom is already established but not yet visible, seek to engage their following single-mindedly in the work of recruitment. This is a response to failed prophecy more or less on the Festinger model. By deploying their 'publishers' (the movement's designation of its members) in extensive house-to-house calls, the movement has developed a systematic method not only of winning new adherents but, perhaps more importantly, of involving members in the regular reaffirmation of their own commitment. The primacy of this latter function becomes apparent when it is acknowledged that at least 200 h of canvassing are on average expended by publishers for each additional new convert.

The Introversionist Response

The Plymouth (Exclusive) Brethren response to hope deferred has been a classic withdrawal from involvement in the wider society into a more introversionist position. The movement has ceased to proselytize: although still undertaking token public proclamation of 'the truth,' they are far from exhorting listeners 'to come and join us' – joining is not an option that is made easy. The Brethren, who regard their assemblies as already a foretaste of heaven, expect the Holy Spirit to guide all sincere Christians to them and their way of life. Maintaining communal purity has become a primary concern, hence the insistence on separation from the inherently evil world, and the diffidence respecting would-be converts. A less extreme example of the same shift of orientation from keen millennialist expectations to a more introverted position is also found among the Christadelphians.

The Demise of the Catholic Apostolics

Perhaps the most radical resolution of the experience of failed prophecy was that of the Catholic Apostolic Church. The Church confidently expected the advent to occur in the lifetime of its original 12 apostles, who were designated (by members who claimed the gift of prophecy) in the early days of the Church's formation in *c.* 1830–35. An apostle died in 1855, and a shocked church acknowledged that there was no brief to replace him. The last of the 12 died in 1901. Thereafter, the Church, according to its own rubrics, lacked legitimacy to ordain ministers. Eventually, in the 1970s, the last minister, ordained by the last apostle, also died. The self-prescribed denial of the competence to make new appointments virtually pre-ordained the demise of the Church if prophecy failed. A schismatic branch in Germany survives, the Neue Apostolische Gemeinde, in which designated prophets took a different view and nominated successors to apostles as they died.

Millennialism in Other Religious Traditions

The millennial theme represents a final outworking of history, in which the wicked are punished, and the righteous are saved in a process which is envisaged generally as involving severe tribulations. The term is applied loosely, however, and the constituent elements of millennialism, as manifested in Christian eschatology, are often absent in so-called millennialist movements in other religious traditions. In Judaism, the messianic element is strong, and has stimulated the emergence of messianic claimants. The most celebrated of these was Shabbatei Tzevi (1626–76), a self-styled messiah whose following, and that of his various successors, was widespread throughout the Middle East and Europe, and which persisted even after he recanted and converted to Islam. In Islam, the Shi'ite tradition of a hidden imam, due to return to lead the faithful in some future time, incorporates both messianic and millennial themes, but the rising of the Mahdi, Mohammed Ahmed ibn Seyyid, in nineteenth-century Sudan appears to have been a more explicit admixture of revolutionary political aims and religious fervor. The visions of future salvation, following a period of tribulation, found in the claims made in the Hindu tradition for Kalki, an avatar of Vishnu, and the prospect of the forthcoming troubled age of Mappo, in Mahayana Buddhism, have lacked the power to inspire significant social movements and scarcely qualify to be designated millennialist. The other-worldliness of Indian religion appears to render it inhospitable to millennialist themes. A more explicitly millennialist movement occurred in late nineteenth-century China, in the T'ai'ping rebellion, but this movement was influenced directly by Christian ideas.

Third World Instances

A wide assortment of religious uprisings among tribal peoples are often termed millennialist. Most of these movements are ephemeral. Some are openly militant, proclaiming either the early return of the ancestors and engaging in practices designation as nativistic, or claiming to be the true originators of all the innovative accoutrements of western civilization. Sometimes a messiah is promised as the savior who will usher in the time of prosperity and bliss, and the prophet harbinger of the new era may himself be cast in the role of messiah. The cargo cults of Melanesia exemplify the collective messiahship of the ancestors for whose return votaries practiced ecstatic rituals or sought to dance the new dispensation into being. The Ghost Dance of 1870 among the tribes of California and its more celebrated successor of 1890 proclaimed collective salvation once the ancestors were induced to return to practice the old ways.

Christian Influences in Third World Cases

Many of these movements were influenced powerfully by Christian millennialism. Christian missionaries were often clergy of fundamentalist persuasion, or were drawn from fundamentalist denominations: millennialism was a stock part of their message. The Kitawala movement in what is today Zambia, Zimbabwe, and Malawi was a radical and semi-politicized millennial cult which owed its inspiration to Watchtower preachers (later designated as Jehovah's Witnesses) of whom Joseph Booth, a sometime Seventh-day Adventist, was a conspicuously influential figure. Suggestions that the Tupinamba and Tupi-Guaraní tribes in South America who, as early as the mid-sixteenth century, were discovered to embrace a cult which taught that they might dance themselves to

weightlessness and be carried off to a promised land of deathlessness in the East, illustrated the possibility of autochthonous millennialism, uninfluenced by Christianity, may be discounted, since those who first reported this phenomenon were Jesuits.

Mass Magical Movements

Various Third World movements were designated readily as millennial even though all that they promised were miraculous therapies or new means for detecting witches. Such were the various Zionist sects of South Africa; the pocomania cults of Jamaica, which claimed to counteract obiah, so-called black magic; and the Lumpa Church of Alice Lenshina which fell victim to military intervention ordered by the Zambian government in 1964. These cults have operated essentially at the individual level, seeking to play very little part in inducing changes in social consciousness.

Revolutionist Movements and Social Consciousness

The more avowedly revolutionist movements have often enjoyed greater influence. Such was the case with the Hau-Hau movement among the Maoris which played a role in the Maori Wars; the Maji Maji Rebellion in East Africa; the afore-mentioned Kitawala movement in Central Africa in provoking the Chilembwe uprising; and the collective resistance movement among erstwhile mutually hostile North American tribes under Tenskwatawa, the Shawnee prophet. Millennialism has often served as an unwitting agency in raising the consciousness of tribal or ethnic identity, or of the need to transcend it in a wider cause. This appears to have occurred among Rastafarians in the West Indies, and more effectively among the Mahdists of the Sudan in the formation of a new state. Prospects of salvation as a chosen people in a new millennium caused Paul to declare that Christians were 'neither Jew nor Greek, bond nor free,' but a new people with a new collective identity.

See also: Cargo Cults; Charisma and Charismatic; Christianity, Origins of; Judaism; Mission: Proselytization and Conversion; Prophetism; Religion: Nationalism and Identity; Secular Religions.

Bibliography

Benz, E. (Ed.), 1965. Messianische Kirchen, Sekten und Bewegungen im heutigen Afrika. Brill, Leiden, The Netherlands.

Cohn, N., 1961. The Pursuit of the Millennium, second edn. Harper, New York.

Festinger, L., Riecken, H.W., Schachter, S., 1956. When Prophecy Fails. University of Minnesota Press, Minneapolis, MN.

Flegg, C.G., 1992. Gathered Under Apostles: A Study of the Catholic Apostolic Church. Clarendon Press, Oxford, UK.

Gager, J.G., 1975. Kingdom and Community: The Social World of Early Christianity. Prentice-Hall, Englewood Cliffs, NJ.

Harrison, J.F.C., 1979. The Second Coming: Popular Millenarianism 1780–1850. Rutgers University Press, New Brunswick, NJ.

Holt, P.M., 1958. The Mahdist State in the Sudan 1881–1898. Clarendon Press, Oxford, UK.

Mühlmann, W.E., 1961. Chiliasmus und Nativismus: Studien zur Psychologie, Soziologie und historischen Kasuistik der Um-sturzbewegungen. Reimer, Berlin.

O'Leary, S.D., 1994. Arguing the Apocalypse: A Theory of Millennial Rhetoric. Oxford University Press, New York.

Pearson, M., 1990. Millennial Dreams and Moral Dilemmas: Seventh-Day Adventists and Contemporary Ethics. Cambridge University Press, Cambridge, UK.

Penton, M.J., 1985. Apocalypse Delayed: The Story of Jehovah's Witnesses. University of Toronto Press, Toronto, ON.

Sandeen, E.R., 1970. The Roots of Fundamentalism: British and American Millenarianism, 1800–1930. University of Chicago Press, Chicago.

Sharot, S., 1982. Messianism, Mysticism, and Magic: A Sociological Analysis of Jewish Religious Movements. University of North Carolina Press, Chapel Hill, NJ.

Wilson, B.R., 1973. Magic and the Millennium: A Sociological Study of Religious Movements of Protest among Tribal and Third-World Peoples. Heinemann, London.

Worsley, P., 1957. The Trumpet Shall Sound: A Study of Cargo Cults in Melanesia. McGibbon & Kee, London.

Mills, Charles Wright (1916–62)

A Javier Treviño, Wheaton College, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Charles Wright Mills was one of the most famous American sociologists of the mid-twentieth century. A public political intellectual, he wrote about the moral uneasiness of the post-War era. Mills's major works include a treatise on social psychology, a trilogy on the American class structure and power system, mass-market paperbacks on nuclear war and the Cuban Revolution, and his most famous book on the sociological imagination. In Mills's view, freedom and reason were being threatened in post-War America. He believed that through a particular way of thinking – what he called the 'sociological imagination' – individuals can better understand the meaning of their epoch for their own lives, and also become active participants in history making.

C. (Charles) Wright Mills (born 28 August 1916, Waco, Texas and died 20 March 1962, West Nyack, New York) was one of the most famous American sociologists of the mid-twentieth century. He was well known and widely read not only in the United States but also in Europe, the Soviet Union, and Latin America. His popularity was due largely to the sale, in the hundreds of thousands of copies, of his mass-market paperbacks. A public political intellectual, Mills delivered 'sermons' about the moral uneasiness of the post-War era to a nonprofessional audience of learned 'publics' by which he meant various groups of politically alert people (Aronowitz, 2012).

Mills's sociology is regarded as 'critical' not only because he exposed the underbelly of U.S. society in order to engender structural change but also because he ruthlessly critiqued the mainline sociology of the 1950s. As one of the foremost dissident intellectuals of the Cold War period and a particularly harsh critic of American foreign policy and national security strategy and of big business and mass society, Mills's thinking had an indelible influence on several generations of dissenters in sociology and the other social sciences. Toward the latter part of his career when he had become more of a political radical, Mills came to be seen as a spokesman for the New Left, the new generation of radicalized students, activists, and intellectuals.

Mills's sociology focused generally on two concerns: (1) the influence of social structure on personality and (2) the bureaucratization of modern society and the power relations it produces (Treviño, 2012). His scrutiny of the centralization of power in the hands of a few elites and the manner in which institutions shape the character of individuals enabled him to engage in a running critique of the types of uninformed men and women and repressive large-scale bureaucracies that prevailed in American mass society during the middle decades of the twentieth century.

Mills identified himself thoroughly as an 'American aboriginal' (Mills and Mills, 2000: 205). This notwithstanding, he held a blatant contempt of American influence and ideology, and was particularly indignant with those who engaged in what he called the 'Great Celebration' of U.S. society without addressing its cultural and political deficiencies. Despite Mills's anti-pro-America stance, he was nonetheless always deeply concerned with the preservation of democracy in the United States. He remained troubled that the U.S. was far less democratic than it should be.

Owing largely to his polemical writings, his moral-political 'preachings,' and the celebrity status he enjoyed outside academe, Mills frequently found himself at odds with many of his colleagues. In consequence, he was effectively ostracized from the sociological community. Mills's response to this exclusion was to become an academic outlaw and exhibit a brusque and choleric mien toward anyone in opposition to him. In contradistinction to most sociologists of the period, Mills, as a public sociologist, aimed to convey modern social science ideas to a broader audience outside the academy by employing a clear and vivid prose. Such were his ways of showing contempt for the routine canons of academe and what he referred to as 'the higher ignorance' of the received sociology of the time.

Personal Life

Mills was born in Waco, Texas on 28 August 1916 to a middle-class family of mixed English, Irish, and French ancestry. He grew up in Sherman, Fort Worth, and Dallas, and spent a few years in San Antonio. His later efforts to escape his Texas backwoods beginnings by immersing himself in urbane high-brow surroundings may partially explain why he saw himself and was treated by others as an 'outlander' in various aspects of his personal and professional life (Horowitz, 1983). Mills also rejected his parent's middle-class world, which he regarded as one of artificial politeness and simulated contentment. He did this by adopting the political ethos and persona of the radical nonconformist.

After graduating from Dallas Technical High School in 1934, Mills enrolled as an engineering student at Texas Agricultural and Mechanical College (A&M), which at the time was an all-male military compulsory institution. As a result of a particularly troubled freshman year at A&M, Mills developed an enduring "visceral negative reaction to militarism" (Geary, 2009: 61).

At the beginning of his sophomore year, Mills transferred to the much more cosmopolitan University of Texas (UT) at Austin, where, in 1939, he concurrently earned a BA degree in sociology and an MA in philosophy. During his senior year at UT, he wrote three essays on the sociology of knowledge. It was on the

strength of these articles, published by the two leading journals of American sociology, that Mills, at the age of 25 years, first came to the attention of scholars working in the areas of the sociology of knowledge, pragmatism, and social theory.

In the fall of 1939, Mills began doctoral training in sociology at the University of Wisconsin, where he continued his studies in the sociology of knowledge and pragmatism. During his final year there, he met and befriended Hans H. Gerth, an émigré from Nazi Germany, who had recently been hired to teach at Wisconsin. Gerth and Mills collaborated on two books: *From Max Weber: Essays in Sociology* (Weber, 1946), the first English translations of Weber's sociological writings, and *Character and Social Structure* (Gerth and Mills, 1953), the first social psychology textbook that considers the influences of institutional orders in shaping individual character.

In 1941, Mills accepted a position as assistant professor of sociology at the University of Maryland, even though he had not yet finished his dissertation, *A Sociological Account of Pragmatism*, in which he interpreted the careers and publics of the American pragmatist philosophers, Charles Sanders Peirce, William James, and John Dewey. Throughout his career, Mills maintained an interest in the social and intellectual values of pragmatism, which emphasizes the fusion of theory and practice, intellect and craft. For Mills, theory must always have a practical use and he regarded his sociology as a form of 'craftsmanship,' where knowledge is intrinsically related to social action.

With the involvement of the United States in World War II, Mills endeavored to avoid being drafted into the army. He was ultimately classified as unfit for military service due to chronic hypertension and well as a too high pulse rate. Mills was adamantly opposed to U.S. involvement in the War. Contrary to most intellectuals of the time, he did not see the War as a struggle for democracy, but believed it was just another opportunity for the imperial powers to redivide the world.

At War's end, Mills joined the Department of Sociology at Columbia University. At a time when academic protocol dictated that professors dress conservatively, in button-down shirt, necktie, and jacket, Mills dressed as a lumberjack – in flannel shirts, leather jacket, and work boots. He commuted to Columbia on his BMW motorcycle while carrying his books and papers in a duffel bag.

Mills developed an international reputation as he traveled extensively. In 1956, he obtained a Fulbright lectureship at the University of Copenhagen. Three years later, he was a visitor to the London School of Economics. In addition to journeying to countries on both sides of the Iron Curtain, Mills spent time in Brazil and Mexico and visited Cuba a few months after the triumph of that country's revolution.

Mills died of heart failure on 20 March 1962 at the age of 45 years. He was buried in Oak Hill Cemetery, Nyack, New York. On his headstone is inscribed the aphorism taken from the opening paragraphs of *The Marxists* (1962), his last book: "I have tried to be objective. I do not claim to be detached" (Summers, 2008: 9).

Major Works

The major works of C. Wright Mills may be subsumed under four general groupings. First, there is the book, *Character and Social Structure*, where he articulates the social psychological theoretical framework that informed many of his other writings. Second, there is the important trilogy on power and social stratification – consisting of *The New Men of Power*, *White Collar*, and *The Power Elite* – in which Mills critically discusses the American class system. Third, there are his polemical tracts – *The Causes of World War Three* and *Listen Yankee* – where he advocates for a politics of responsibility and truth. Finally, in his most famous book, *The Sociological Imagination*, Mills presents his vision of what sociology should be and what it has to offer. Throughout all of these writings, Mills had two main objectives that required that he consider the character of individuals in the context of their social structure: to define and dramatize the essential characteristics of his time, on the one hand, and to see various populations as actors in the social drama of the twentieth century, on the other.

The Sociology of Persons and the Psychology of Institutions

Character and Social Structure (1953), which Mills coauthored with Hans H. Gerth, is a concise and systematic statement concerning the psychological nature of five major institutional orders: (1) political, (2) economic, (3) military, (4) religious, and (5) kinship orders. Here, Gerth and Mills not only consider how types of personalities are anchored in each of the institutional orders but also discuss how the institutional orders form social structures. Their goal is to construct a working model of persons and society by combining the sociology of persons with the psychology of institutions. This synthesis produces for Gerth and Mills a 'social psychology' through which they aim to achieve three objectives. First, they attempt to describe and explain the conduct and motivations of people in various types of societies and at various points in time. Second, they ask how the external conduct and inner life of one individual interplay with those of others. Finally, they seek to describe the types of persons usually found in different societies and then to explain them by tracing their interrelations with their societies.

Gerth and Mills indicate how social structures shape the personalities of people as historical actors. In developing a social psychology that considers the objective social functions of institutional orders alongside the subjective meanings of institutional members, Gerth and Mills combine key concepts from the work of George Herbert Mead, Freud, Marx, and Max Weber. Simply put, in emphasizing the roles people play in various institutions, Gerth and Mills demonstrate how 'character structures' – personality traits and conduct patterns – are molded by the institutional orders to form historical types of social structures.

The Class Structure and Power System

The New Men of Power, *White Collar*, and *The Power Elite* constitute Mills's studies on the American class structure and power system. These books attempt to relate the psychological characteristics of certain groups and their individual members – labor leaders; white-collar workers; and the political directorate, the chief

executives of large corporations, and the warlords – to the stratified milieu of post-War America. Mills is chiefly concerned with power and the powerful: the elites, exploiters, and policy makers.

In *The New Men of Power* (1948), Mills focuses on the American labor movement and explains the rise of union leaders and their role as a new elite. The book is basically an analysis of the character of America's labor leadership – the positions they occupy, their career lines, and the traditions and anxieties that motivate them. Endeavoring to relate a type of social role to a particular social structure, Mills is interested in union leaders not as individuals but as a social type formed by the roles they played in a political economy that was changing from laissez-faire to monopoly and a state capitalism with many corporate features.

Given that the U.S. economy had, since the Great Depression, been racked by the cycle of slump and war and boom, Mills's concern was that post-War America was drifting toward a war-oriented economy drawn into the Cold War conflict with Soviet Russia. He believed that only the powerful force of labor unions as agencies of protest could stop the U.S. from an incessant spending of large sums on munitions and other military items. As members of a strategic vanguard and the only potentially liberating mass force, the leaders of American labor – the new men of power – along with the labor intellectuals (i.e., the union's lawyers, editors of the union's newspapers, economists, statisticians, and research directors) could form an alliance of power and intellect to combat the main drift that was transforming the U.S. into a corporate garrison state. Mills, however, had little confidence that the leaders of trade unions would broker any opposition to a militarized economy. They were too conservative politically and too inhibited intellectually to seek the egalitarian, democratic society. Instead, they desired only to stabilize their power and position. Seeking a larger cut of the economic pie, the labor leaders had become politically and economically absorbed into the new corporate–government complex and had relinquished any aspirations for seeking structural change. As such, there was little that ideologically distinguished them from those leaders in the political, corporate, and military spheres.

Today, unions in the U.S. are far fewer in number compared with when Mills wrote about them in the mid-1940s. Union membership dropped from just over 35% of all workers in 1945 to less than 12% in 2012 (The Foundry, 2013). In fact, union membership has declined to almost exactly the same percentage as it was in 1930, before President Franklin D. Roosevelt encouraged the growth of unions.

By the mid-1950s, Mills had become thoroughly disillusioned with labor's ability to participate in a political movement to transform American society and no longer saw union leaders as players in the higher circles of power. His disenchantment with the labor movement of the working class led him to turn his attention to the next rung in the stratification hierarchy: the white-collar workers of the new middle class.

In *White Collar* (1951), Mills explains that how certain structural and occupational changes taking place in mid-twentieth-century America were affecting the psychological character, the social biographies, and the social roles of white-collar workers. According to Mills, the increasing bureaucratization of U.S. society coupled with the centralization of corporate power gave rise to a world of big organizations inhabited by the 'new middle class,' or those white-collar workers involved primarily in sales and management and whose work situation was increasingly bureaucratized by the 'command hierarchies' of business and government. As dependent employees they planned, administered, recorded, distributed, and managed for others. The white-collar workers of the new middle class included salespeople, corporate managers, clerical workers, and bureaucratic professionals such as salesmen and public relations specialists, who were needed to manage, design, sell, and keep account of production. Their characteristic skills involved the handling of paper, money, and people. By the middle of the twentieth century, one-fourth of the labor force was made up of employees working in large bureaucratic corporations.

The white-collar employees had become professionalized occupants of specified offices and specialized tasks, working in impersonal bureaucracies that made their work dull and routine. This meant that they experienced feelings of frustration, despair, anxiety, and insecurity. They had become powerless and alienated because they could not realize themselves in their work, for work had become a set of skills sold to another. These transformations of power striped the white-collar professionals of any control over their work. And because there existed a separation between the product and the processes of work, they had never acquired a sense of craftsmanship – of purpose and fulfillment – from their jobs. For them, work had no intrinsic meaning and provided no gratification.

The white-collar managers, clerks, and bookkeepers were cogs in a business machinery that had routinized greed and made aggression an impersonal principle of organization. Mills asserts that the white-collar salaried professionals were forced to accept the meaninglessness of their working life. As a result of the alienating conditions of modern work, for the mass of employees work had a generally unpleasant quality.

Further, the more the middle classes experienced their life as one of powerlessness, the more apathetic they became to politics of any kind. They were neither radical nor liberal nor conservative nor reactionary; they were 'inactionary,' strangers to politics, and history was being made behind their backs. In short, middle-class white-collar workers had become 'cheerful robots,' a mass of confused and unfocused automatons adrift in a bureaucratic world not of their own making.

Elaborating on the issues of increased bureaucratization, the consolidation of power, and political indifference, Mills's *The Power Elite* (1956) is a study of stratification that reveals a tripartite ruling stratum in U.S. society. His basic thesis is that as the institutional means of decision, information, and power became more coordinated and as the public became more politically uninformed, there had arisen a national 'power elite' with tiers and ranges of wealth and power of which people in the rest of society know very little. Mills identifies the American power elite as that relatively inaccessible small group of individuals and families who possess disproportionately more money, power, and prestige, as well as all the experiences, privileges, and trainings, which these resources provide.

He sees the elite of power as occupying pivotal positions in the three major hierarchies and organizations of modern society at that time: the machinery of the state, the big

corporations, and the military establishment. Accordingly, as members of the political directorate, the corporate rich, and the high military, the power elite command the dominant institutions of a dominant nation, and are in a position to make decisions of enormous consequence for the underlying populations of the world and to determine the course of events. The general public, in its supreme disinterestedness, frequently does not know that these decisions are being made until well after the fact. In short, the power elite are people of power within the coordination of political, economic, and military decision.

The 'political elite' group consists of higher politicians and key government officials, but in particular the president, vice president, and the members of the cabinet. It also includes the White House staff as well as the most important appointed heads of major regulatory agencies and commissions. Aside from occupying the executive command posts of the political order and forming the political directorate, the political elite are also the legal, managerial, and financial members of the corporate elite.

The second group composing the triangle of power, the 'corporate elite,' is made up of persons who occupy the top command posts in the giant corporations. The corporate elite also includes top-level management, the major stockholders, and the corporate lawyers representing the largest financial and industrial corporations in the country. At the very top stratum are the high-ranking executives, the 'corporate rich,' who manage the corporate complexes and make the key economic decisions. As the corporate world became more intricately involved in the political order, these executives became intimately associated with the politicians, and especially with the key politicians who form the political directorate of the United States government. Additionally, as increasing numbers of corporation chieftains enter government directly, the result is the emergence of a new political economy at the apex of which are situated those who represent the corporate rich and the political elite.

Finally, the third sector of this ruling triumvirate is the 'military elite,' who oversee the largest and most expensive feature of the U.S. government, the military order. At the top of the military order's bureaucracy, just below the president and the secretary of defense, is the military board of directors, the Joint Chiefs of Staff. Immediately below the Joint Chiefs, there is a circle of generals and admirals presiding over the elaborate and dispersed military forces as well as the economic and political liaisons necessary to maintain those forces. Thus, alongside the corporation chieftains and the political directorate, the generals and admirals of the Pentagon have been given increased power to make and influence decisions of the gravest consequence.

The Joint Chiefs of Staff, along with the munitions contractors who supply them with weapons, comprised a new and potent force in making the key decisions regarding U.S. foreign policy and international relations. Indeed, the military and corporate elites became political insiders serving as the president's most influential advisors. Mills warns of a new corporatism – directed by the political, industrial, and military elites – that points at the trend toward a permanent military economy. A few years later, President Dwight D. Eisenhower, himself a member of the power elite, echoed Mills's concern in his farewell presidential address to the nation, cautioning the American people about the burgeoning 'military–industrial complex.'

By using the mass media, the power elite of the three interlocking and coordinated directorates – the machinery of the state, the big corporations, and the military establishment – are able to persuade the mass of people into believing that through the democratic process, they have made the key decisions, when in fact they have not. Rather, it is the power elite who determine the course of historical events.

The Politics of Responsibility and the Politics of Truth

Mills's stinging critiques of U.S. society, world affairs, and professional sociology will leave his readers with a deep sense of pessimism or else provoke them to organize, mobilize, and take action for change. This latter inspiration stems from his admonition to sociologists and other intellectuals that they must be motivated by a commitment to politics. Indeed, throughout his later, more polemical writings – *The Causes of World War Three* and *Listen Yankee* – he calls on scientists and artists, ministers and scholars to practice the 'politics of responsibility' (Mills, 1958) and assert moral leadership. This means that intellectuals have an obligation to hold those in power accountable for the decisions they make and that affect human destiny. And in the late 1950s, at the height of the Cold War between the U.S. and the USSR, there was no decision that would have more momentous, and calamitous, consequences for the whole of humanity than thermonuclear war.

The Causes of World War Three (1958) is an explicit evocation of the politics of responsibility. In it, Mills maintains that the power elites in the United States and the Soviet Union were leading the two nations to a total and absurd war. These two superpowers intimately shared the 'military metaphysic,' a way of thinking in which everything in the world situation is officially defined in terms of national security and defense. As such, Mills calls upon the clergy, scientists, and the intellectual community to take a responsible and moral stand on the issue of peace and nuclear disarmament. He also calls for a rational response and a commitment to the task of overcoming widespread civic apathy, indifference, and elite irresponsibility concerning the annihilation of the human species through nuclear war. Mills endeavors to persuade American and Soviet intellectuals to prevail over the organized irresponsibility and unrealistic decisions of their respective countries' power elites and to, by their own efforts, sue for a separate peace. Mills believed that only in this way could the drift and thrust toward World War III and mass destruction be reversed.

But more than being politically responsible, Mills also maintained that social scientists and other intellectuals hold a politically informed notion of truth. This was a 'politics of truth' (Mills, 1959) – the critical voice of the intellectual that exposed the truth about the political power relations in U.S. society. This meant that intellectuals are duty-bound to tell the truth – to disclose the facts – about social reality; particularly, since this reality was being distorted by the stultifying culture of mass society and the manipulation of the mass media. Thus, in *Listen Yankee: The Revolution in Cuba* (1960), Mills seeks

explicitly to present to the U.S. public 'the truth' about the recent Cuban Revolution.

Shortly after the triumph of the revolution, Mills visited Cuba in order to observe firsthand its accomplishments in transition. Given its absence from the news of Cuba in the US, his goal in this book was to clearly and emphatically present the 'voice' of the Cuban revolutionary to Americans. On his trek through the island, he witnessed the innovative ways in which the common people were beginning to construct an entirely new society. Mills also interviewed scores of Cubans, from all walks of life, who were close to events and who were eager to express their sentiments. After a couple of weeks on the island, he returned to the U.S. and began writing *Listen Yankee*.

The book consists of a brief introduction followed by eight 'letters' in which Mills uses direct speech to convey a composite viewpoint of how the Cuban revolutionaries see their revolution as well as how they define their aspirations and their relationship with the United States. He considers Cuba's voice as the voice of the hungry-nation bloc (that is, Asia, Africa, and Latin America), and the Cuban revolutionary, he believes, was speaking (mainly through his book) in the name of many people in that bloc. For Mills, it was imperative that Cuba's voice be heard in the United States. Up to that point, the American public had virtually ignored Cuba, and Mills pleads with them to hear well the message of the Cuban Revolution; for only by dealing with the perils of ignorance could the perils of disastrous mistakes be avoided.

One year after the publication of *Listen Yankee*, President Eisenhower approved the CIA-sponsored Bay of Pigs military invasion of Cuba to overthrow the left-wing government of Fidel Castro. The failed invasion increased Castro's popularity as Cuba's leader and likely led him to pursue closer ties with the Soviet Union. Castro remained leader of Cuba for the next 50 years and the island nation, located 90 miles south of the United States, is today one of the world's remaining socialist states.

The Sociological Imagination

The Causes of World War Three and *Listen Yankee* established Mills's image as a public intellectual. Between those two books he produced the one for which he is best known in academic sociology: *The Sociological Imagination* (1959). Part primer, part polemic, Mills wrote it for two main reasons: to inveigh against the current dominant 'schools' in academic sociology and to articulate his own version of sociology.

A trenchant critic of what he saw as the two major sociological orientations in the 1950s, Mills scorns, on the one hand, the narrow, technical procedures of researchers like Paul F. Lazarsfeld, and on the other, the grandiose conceptual perspectives of theorists like Talcott Parsons. According to Mills, researchers had become so obsessed with particular methods that they were unable to see, much less study, the bigger questions pertaining to the social forms and tendencies that directly affect people's lives. They were engaged in 'abstracted empiricism,' an approach that converts the pressing social issues of the day into mundane statistical assertions. And given that research technicians were overly focused on trifling and specialized topics such as public opinion and voting behavior, they ignored the larger issues of social structure such as those dealing with the legitimacy and use of the existing distribution of power. Today, abstracted empiricism is a pervasive component in sociology as evidenced by the multitude of quantitative and narrowly focused studies published in major journals such as the *American Sociological Review*. Indeed, it has become more popular with the increased use of high-speed computers, statistical analysis, and large-scale data sets that produce studies devoid of sociological meaning and that ignore or deny the existence of social structure.

Mills also attacks Parsons's 'grand theory' or his general conceptual scheme for analyzing the structure and processes of social systems, and accuses him of being so preoccupied with creating concepts that Parsons cannot define in a clear and orderly way the social problems at hand, much less guide efforts to solve them. Mills urges grand theorists to abandon their high levels of abstraction and to instead observe problems in their historical and structural contexts. While grand theorizing fell out of fashion in the U.S. during the 1960s, its use continues in the work of European sociologists like Norbert Elias, Jürgen Habermas, and Niklas Luhmann.

As an alternative to abstracted empiricism and grand theory, Mills proposes the strategy that he believed must inform the sociological enterprise, 'intellectual craftsmanship.' This is a reflective style of work that guides sociologists' professional activities as well as what they, as persons, observe and experience in their everyday lives. In other words, social scientists must learn to use their personal experience in their intellectual work (Treviño, 2013). This consideration of personal experience, of biography, is central to what Mills regards as the most needed quality of mind for intellectual workers: the sociological imagination.

The 'sociological imagination' is a way of thinking that provides people with the insight into the social conditions of their existence and an understanding of the causes of their feeling trapped and uneasy, alienated and powerless. Its practical task and its intellectual promise are to enable individuals to grasp history and biography and the relations between the two within society. For only by seeing the interconnection of biography, or the intensive study of personality in its social setting, and history, can people understand their own experience, gauge their own fate by locating themselves within their period, and know their own life chances by becoming aware of all individuals in similar circumstances. Consequently, in employing the sociological imagination, people must ask three basic questions: (1) What is the structure of this particular society as a whole? (2) Where does this society stand in human history? (3) What varieties of men and women now prevail in this society and in this period? The sociological imagination, therefore, enables people to understand themselves fully, and when properly used, it can free them to be involved in the rational making of history. But in order to be free, individuals have to make the connection between their 'personal troubles' of the immediate milieu and those 'public issues' at the level of historically located social structure. It is the duty of those who employ the sociological imagination to continually translate troubles into issues and issues into the terms of their human meaning for them. Individuals must be aware that the malaise and frustrations they experience in their inner lives are linked to those problems residing at the level of social structure.

In his 2004 presidential address to the American Sociological Association, Michael Burawoy (2005) extended Mills's notion of the sociological imagination to include active participation in civil society, what Burawoy calls 'public sociology.' For Burawoy, the sociologist must enter into an unmediated dialogue in certain pockets of civil society such as with neighborhood associations, with communities of faith, with labor movements, and with prisoners. The core activity of public sociology is the dialogue between sociologists and their publics, and its political role, as Mills had previously stated, involves the sociologist "talking to publics and at kings" (Burawoy, 2008: 372; see also Mills, 1959: 181).

Freedom and Reason

As an undergraduate student, Mills was introduced to the American pragmatist philosophers, who believed that reason could serve as a guide to freedom. This link between reason and freedom served as an instructive force throughout Mills's writings.

As Mills saw it, the two values of modern Western culture – freedom and reason – were being threatened in post-War America. It could no longer be assumed that increased rationality necessarily made for increased freedom. The extreme complexity of the current era coupled with the concentration of power and knowledge made it so that ordinary people often could no longer reason about the great structural forces that affected their lives.

In modern American society, most people's freedom, which, for Mills, involves their chance to reason about available choices and then have the opportunity to choose, is constrained by power relations. Those with the power to make the larger decisions of their society are freer than those without the power. Mills draws two conclusions about individuals in their relation to the power structure of American society. First, though all people are free, some, by virtue of their placement within the social structure, are more powerful and freer than others to shape human affairs. Second, and subsequently, history is made behind most people's backs.

While Mills accepts the pragmatist notion that reason plays a pivotal role in freeing individuals from the constraints of their social structure, he does not assume that this emancipation is an automatic process. He maintains that individuals must be willing and structurally positioned to acquire the reason – the capacity to formulate available choices in life – that freedom requires.

A true democratic freedom can be realized only when individuals become aware of, and want to do something about, the social forces that constrain their lives. More to the point, individuals, in order to secure their freedom, must have not only the power but the desire to make decisions concerning their own lives and their place in history. True freedom of action means engaging in free initiative for self-determination and accepting responsibility for choices made. In the end, Mills's main concern was to utilize sociologically informed reason in the pursuit of a more just and democratic society.

See also: Bureaucratization and Bureaucracy, History of; Cuban Revolution, The; Dewey, John (1859–1952); James, William (1842–1910); Lazarsfeld, Paul Felix (1901–76); Mead, George Herbert (1863–1931); Parsons, Talcott (1902–79); Social Inequality in History (Stratification and Classes); Weber, Max (1864–1920); White-Collar Workers, History of.

Bibliography

Aronowitz, S., 2012. Taking It Big: C. Wright Mills and the Making of Political Intellectuals. Columbia University Press, New York.
Burawoy, M., 2005. For public sociology. American Sociological Review 7, 4–28.
Burawoy, M., 2008. Open letter to C. Wright mills. Antipode 40, 365–374.
Geary, D., 2009. Radical Ambition: C. Wright Mills, the Left, and American Social Thought. University of California Press, Berkeley.
Gerth, H.H., Mills, C.W., 1953. Character and Social Structure: The Psychology of Social Institutions. Harcourt, Brace & World, New York.
Horowitz, I.L., 1983. C. Wright Mills: An American Utopian. Free Press, New York.
Mills, C.W., 1942. A Sociological Account of Pragmatism: An Essay on the Sociology of Knowledge. (Unpublished Doctoral Dissertation). University of Wisconsin, Madison.
Mills, C.W., 1948. The New Men of Power: America's Labor Leaders. Harcourt, Brace, New York.
Mills, C.W., 1951. White Collar: The American Middle Classes. Oxford University Press, New York.
Mills, C.W., 1956. The Power Elite. Oxford University Press, New York.
Mills, C.W., 1958. The Causes of World War Three. Simon & Schuster, New York.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, New York.
Mills, C.W., 1960. Listen Yankee: The Revolution in Cuba. Ballantine, New York.
Mills, C.W., 1962. The Marxists. Dell, New York.
Mills, K., Mills, P. (Eds.), 2000. C. Wright Mills: Letters and Autobiographical Writings. University of California Press, Berkeley.
Summers, J., 2008. The Politics of Truth: Selected Writings of C. Wright Mills. Oxford University Press, New York.
The Foundry, 2013. Union Membership Continues to Decline. http://blog.heritage.org/2013/01/27/chart-of-the-week-union-membership-continues-to-decline/ (accessed 19.05.14.).
Treviño, A.J., 2012. The Social Thought of C. Wright Mills. Sage, Los Angeles.
Treviño, A.J., 2013. C. Wright Mills as designer, craftsman, and stylist. In: Scott, J., Nilsen, A. (Eds.), C. Wright Mills and the Sociological Imagination: Contemporary Perspectives. Edward Elgar, Abingdon, UK, pp. 213–215.
Weber, M., 1946. In: Gerth, H.H., Mills, C.W. (Eds.), From Max Weber: Essays in Sociology. Oxford University Press, New York, pp. 129–156.

Relevant Websites

https://www.facebook.com/pages/C-Wright-Mills/112492478766211?nr – Facebook.
http://www.lib.utexas.edu/taro/utcah/01094/cah-01094.html – A Guide to the Charles Wright Mills Papers.
http://infed.org/mobi/c-wright-mills-power-craftsmanship-and-private-troubles-and-public-issues/ – Infed Encyclopedia.
www.cwrightmills.org – Mills photos.
http://sociologicalimagination.org/archives/category/the-21st-century-sociological-imagination – The Sociological Imagination.
http://en.wikiquote.org/wiki/C._Wright_Mills – Wiliquote.

Mind, Theories of

Fulvia Castelli, University of Pavia, Pavia, Italy

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A.M. Leslie, volume 23, pp. 15652–15656, © 2001, Elsevier Ltd.

Abstract

Theory of Mind (ToM) is a term coined to describe people's effortless propensity to search for intentions behind actions. ToM research began with primates in 1978 and continues in humans to investigate both the developmental stages of skills involving the ability to understand other's mental states – infants (7 months old) compute other's beliefs – and its brain basis. Research on ToM has always overlapped with research on autism. Findings indicate that individuals with autism have a diminished cognitive awareness of their own and others' mental states, and their brain connectivity in a multicomponents network is disrupted in different areas.

A Multi-Faceted Concept

People have the pervasive tendency to explain their own and others' actions in terms of beliefs, desires, and goals. The attribution of mental states represents a useful social strategy to make sense of and predict people's behavior. The terms 'Theory of Mind' (ToM), 'mindreading' or 'mentalizing' have all been coined to describe people's effortless propensity to search for intentions behind actions.

One of the finest examples of the predictive usefulness of ToM attribution comes from the fictional character of Sherlock Holmes, the great mastermind of detection created by Sir Arthur Conan Doyle in the nineteenth century. In 'A scandal in Bohemia,' Sherlock Holmes has been instructed by his client to find a photograph that a lady has safely concealed in her house. Sherlock Holmes' cunning plan is to make the lady think that her living room is on fire by tossing in a smoke rocket and raising a fire alarm. What happens afterward is exactly what Sherlock Holmes had predicted: the lady's instinct is at once to rush to secure the photograph, which is the object she valued most, and by doing so she accidentally revealed to Sherlock Holmes, who was in the same room, the secret hiding place. The case is brilliantly solved by the detective thanks to his correct attribution of intentionality to his opponent.

Intentionality is, according to philosopher Frank Brentano and von (1970) the feature that sets the mental states completely apart from the physical states. Mental states hold the property of 'pointing' toward certain objects or states of affairs (*Intendere* in Latin means to aim), but these objects or states of affairs need not exist. Intentionality is the common feature of propositional attitudes, those cognitive states that are described in everyday language with the use of a 'that' clause (typically beliefs and desires, i.e., " I believe that ….").

An important aspect of propositional attitudes – beliefs and desires, as well as intentions, thoughts, doubts, remembrances – is that they may represent actual or possible states of affairs. The states of affairs upon which people's propositional attitudes are directed may actually attain in the real world, but equally they might not: beliefs are often false, desires can be frustrated, and hopes may be dashed. Hence, propositional attitudes are states that 'do not exist here and now': they are states that are about or directed upon a nonexistent state of affairs or objects that can be true or false, and even paradoxical (e.g., "John believes that drinking water is safe," and "he believes that drinking H_2O is poisonous").

The term 'intentionality' refers to all mental states; hence, all mental states are intentional. These include propositional attitudes (beliefs, desires), pretense, intentions, and emotional states. Their differences can be expressed in terms of their contents: beliefs, desires, and pretense 'point toward' a proposition; intentions or goals and aims point toward a future state of affairs; and emotional states point toward a state of affairs.

The ability to pretend or to understand pretense, according to Leslie (1987), is remarkably similar to the logical properties of intentional states. At 2 years of age, infants possess not only the ability to form primary representations, but also meta-representations. The former ability allows them to understand the world in an accurate, faithful, and literal way – transparent representations – whereas the latter ability allows them to understand states that are not necessarily existing and true – opaque representations – for example, the child is not confused but delighted when his/her mother is pretending to hold a telephone when she picks up a banana saying "Let's call Daddy!"

An important distinction is within propositional attitudes, namely, between desire and intention. The simplest form of intention implies an action, whereas desire implies always an attitude toward a proposition; for example, "Mary wishes that she goes to Mars," is different from "Mary intends to go to Mars." Indeed, if Mary intends to go to Mars, she has to set off and organize the interplanetary journey, whereas if she wishes to go to Mars, she can simply daydream about it. Obviously in our everyday language such a distinction is not easy to apply: the terms 'wanting' and 'trying' are used frequently and are more ambiguous. If Mary says: "I want to go to Mars," it is not clear whether she is just uttering a metaphor, or whether she will really try to go to Mars.

In our everyday descriptions of people and ourselves the use of different emotion terms are as frequent and useful as belief/desire terms. Nevertheless, the general category of 'emotion' includes very different psychological states such as sudden reaction responses to people or objects and introspective experiences. The difference between emotional states and propositional attitudes is that their intentional content is different: emotions are mental states that point directly toward

a state of affair, for example, "I am afraid of snakes" whereas beliefs and desires are attitudes toward propositional content, for example, "I believe that snakes could be dangerous."

To sum up, mental states attribution includes various types of mental states, all with different properties. From the very pragmatic view of social science research, it is important to underline that when a test is created with the purpose of investigating the ToM ability, it is tapping into a very complex cognitive ability which allows people to make inferences regarding different intentional contents. Whether different contents of mental states are reflected in a multifaceted cognitive organization is a matter for investigation.

The Nature of Theory of Mind

The processes that underlie the development of ToM are still a matter of a sometimes tortuous debate. The concept of mental states seems as abstract as the concept of the electron, and there seems to be as little and direct everyday evidence for the existence of beliefs and desires as there is for the existence of electrons. Historically, researchers are divided in those who advocate theory–theory, modularity theory and simulation–theory. The first view assumes that possession of ToM concepts depend upon possession of abstract knowledge (e.g., Gopnick and Wellman, 1992). The second view assumes that possession of abstract concepts depend upon a cognitive mechanism, rather than upon knowledge (e.g., Leslie 1987, 2000). The third view assumes that we understand other people's behavior by recreating the mental processes on ourselves; hence, perception of others' actions is coupled with own's action (e.g., Gallese and Goldman, 1998). Different variations of these explanations have been proposed over the last decades; however, recent views emerging from neuroimaging literature on ToM (e.g., Apperly, 2008) agree that these different explanations do not bear heuristic value any longer. Indeed, the recent marriage between psychology and neurobiology has provided much impetus for future research into the 'social brain' (Brothers, 1990), which includes both the system of ToM and a 'mirror system' (Rizzolatti et al., 1996). The former enables us to understand actions in terms of desires and beliefs and the latter enables us to understand other's goals and intentions and to empathize with their emotions by a mechanism of motor resonance (Frith and Frith, 2010). Most importantly, there is consensus that future research will be in the direction of investigating various triggering inputs for these systems with particular emphasis on the distinction between automatic versus deliberate social signals (Frith and Frith, 2012).

Origin of Theory of Mind Research

Research into ToM began with primates. The very term 'Theory of Mind' was coined by Premack and Woodruff (1978) who, in a pioneering paper, asked whether chimpanzees have the ability to infer mental states from others' behavior. The first nonlanguage-based ToM test was used with Sarah, an adult female chimpanzee. She was shown a series of videotaped events in which human actors were struggling to obtain inaccessible objects, and then asked to indicate how she thought the human actor would solve his problem. One of the scenes, for example, presented the actor in a cage similar to that of the chimpanzee's, struggling to get bananas that were hanging from the high ceiling. The animal then was presented with a pair of still photographs, one of which depicted the actor engaged in a behavior that constituted a practical solution to the impediment, for example, stepping onto a box.

Premack and Woodruff argued that Sarah's consistent choice of the correct photograph can be understood by assuming that the animal possesses ToM: she solves such problems by being able to attribute to the actor both an intention or purpose and knowledge. According to the authors, when the chimpanzee watches the videotapes showing the human actor jumping to reach the bananas attached to the ceiling, she assumes that the actor wants the bananas and he is struggling to reach them. She also assumes that the actor knows how to attain the bananas so that when she is required to indicate which photograph depicts the solution, she chooses correctly the picture of the actor stepping on the box. However, these results can also be explained with no reference to mental state representations.

In a commentary to Premack and Woodruff's study, philosopher Daniel Dennett (1978) argued that the evidence of the primate's correct choice may equally well be supported by "supposing that [it is] derived from either habits (of thought) or belief about other features of the world (e.g., experienced regularities in the behaviors of others)" (p. 569). Dennett pointed out three crucial requirements for an experiment to establish whether subjects attribute mental states to others or not: (1) the actor's anticipated behavior is a novel action, in the sense of not being habitual or often repeated in the given circumstances; (2) the subject's response consists of an action as much as possible from his natural repertoire, rather than a highly trained artificial response; and (3) the perceived dependence of the subject's response on the other person's action should also be natural and obvious.

The first requirement is fulfilled when the belief attributed to another person is false. Only in that case, will the individual be expected to act inappropriately to the novel circumstance, hence, in all likelihood, not the way the person has typically acted in the past. Correctly reporting someone's belief when it differs from your own cannot be done by simply referring to one's own knowledge of the facts. The other two requirements are fulfilled by communicative acts, such as asking a question or requesting something from the other individual.

Following Dennett's suggestions, Wimmer and Perner (1983) accepted the challenge to develop a test to investigate children's ability to attribute mental states to others. The classic 'False Belief Task' paradigm investigates the ability to understand that a belief is opaque, namely, that there can be different beliefs about a single event. Hence, the test is based on the assumption that understanding someone else's false belief requires the representation of a mental state that differs from what is known to be true. If a child is able to represent another's mental states, then he or she will be able to predict the ignorant person's behavior, which is no longer appropriate in the changed circumstances.

The original test, a *location change* task, was constructed by Wimmer and Perner around a story character, Maxi, who hides

a chocolate in a box before he goes out to play. While he is out, mother comes in and moves Maxi's chocolate from the box to the cupboard. Then Maxi comes in back from playing. The experimenter asks: "Where will Maxi look for his chocolate?" The test allows for distinguishing whether the child predicts Maxi's behavior on the basis of his own knowledge (Maxi will look where the chocolate really is) or whether he understands Maxi's false belief (Maxi will look where the chocolate was left). The child answers without the need of verbalizing his response, since a mere pointing gesture suffices. Results indicated that while most 4-year-olds correctly predicted Maxi's behavior on the basis of his false belief, most 3-year-olds answered on the basis of the new situation, without taking into account that Maxi's belief remained the same regardless of the reality changed.

In another False Belief Task, a *content change* task, the child himself is the holder of a belief which is revealed to be false. In the 'Smarties' task (Hogrefe et al., 1986) the child is asked to guess what a Smarties tube – well-known sweets' package – contains. After the child correctly answers "sweets" or "Smarties," the experimenter opens the package and shows the real content, a pencil. The lid is then replaced and the child is asked: "When you first saw this box, before we took off the lid, what did you think was inside?" (own mental state question). In this task 3-year-olds do not seem to recall their false belief but answer instead "a pencil." The same result occurs when the child is asked what another child would say when shown the same tube of Smarties (other person mental state question): 3-year-olds expect a true belief while 4-year-olds recognize the false belief (Perner et al., 1989; Wimmer and Hartl, 1991).

Finally, in 1985, Baron-Cohen, Leslie and Frith tested the hypothesis that children with autism, with a mental age above 4 years, fail to take into account other people's beliefs. They used a variation of the Wimmer and Perner (1983) paradigm with a group of normally developing children (mean age 4.6 years), a group of Down syndrome children (mean age 11 years), and children with autism (mean age 12 years). Children were presented with a scene with two dolls, 'Sally' and 'Ann': Sally has a basket and Ann has a box. Sally puts her marble in the basket and leaves the scene. While she is out, naughty Ann takes the marble out of the basket and puts it in her box. Question: Where will Sally look for her marble? The results showed that 80% of the children with autism failed, pointing to the location where the marble actually was, whereas the majority of both control groups passed (85% of the normal developing children and 86% of the Down syndrome children). This finding has been replicated in a number of studies, using the verb 'think' in the question rather than 'look' and real people rather than dolls (Leslie and Frith, 1988). Furthermore, Perner et al. (1989) replicated the findings from the 'Smarties' task with a control group of children with specific language impairment, so that the possibility that children with autism failed to understand false-belief tasks on the basis of a general language deficit was ruled out.

Research on ToM began with Premack and Woodruff's seminal paper asking whether chimpanzees have a ToM: on the 30th anniversary of its publication, Call and Tomasello (2008) reviewed all data available and concluded that chimpanzees do understand others in terms of a perception–goal psychology but do not seem to understand false beliefs in particular.

Theory of Mind and Own Mental States

In the Premack and Woodruff's paper, ToM was defined as follows: "In saying an individual has a ToM we mean that the individual imputes mental states to himself and to others […] The system can be used to make predictions, specifically about the behavior of other organisms" (p. 515). While the majority of research has focused on attributing mental states to others, attribution of mental states to self has been relatively neglected in the literature.

Everyday common sense is that we know what 'we think' better than what 'others think.' However, data from typical development give little evidence to date of better performance on task requiring reflection on own mental states and to others. The ability of self-reflection emerges no earlier in development (Wellman et al., 2001), and neuroimaging findings appear to suggest that regions active during attribution of own mental states overlap with region active during attribution of mental states to others (Happè, 2003).

The task used most widely to assess recognition of false belief *in self* and others is the 'Smarties' unexpected contents task described above. Indeed, participants are asked both a Self-False belief question – what they thought was inside the box before they looked – and an Other-person False belief question – what another person will think is inside the box before they look. On the Smarties task, individuals with autism tend to perform as poorly on the Self test question as they do on the Other-person test question (Baron-Cohen, 1991, 1992; Fisher et al., 2005; Russell and Hill, 2001). However, there is confound in this particular procedure because children could benefit from the request to state at the beginning of the test what they think is inside the Smarties box. They 'verbalize' their false belief before the critical questions are asked. This sort of 'memory for statement' could be the strategy that children use to pass the test questions without really reflecting on their own belief (Williams and Happè, 2009). Williams and Happè crucially manipulated the conditions of a content change task (the Plaster task) in a way so that the participants had to select a box where they thought there were plasters without being asked to explicitly verbalize their belief. The main finding was that participants with autism performed significantly less well on the Self test question than the Other-person test question on this new task. Other studies suggest that individuals with autism have a diminished awareness of their own and others' intentions and that this diminution is associated with other impairments in ToM (Williams and Happè, 2010).

The Developmental Shift between Three and Four Years of Age in and Autism

The issue of ToM deficit in autism became interconnected with the issue of the failure of 3-year-old children and the developmental shift in understanding false beliefs. Paradigms first introduced to explore ToM in normal children were also used with children with autism. In fact, it is crucial to understand the reason why both the groups fail the same task: could it be that young children are similar to children with autism at the cognitive level? The answer is

negative: extensive research has shown that 3-year-old normally developing children and older children with autism are very different from one other, even though both fail false-belief tasks.

The 'false photograph' paradigm (Leekam and Perner, 1991; Leslie and Thaiss, 1992) based on a task devised by Zaitchik (1990), nicely demonstrated that children with autism do not find difficulties with any 'out-of-date' representations, but only with those based on mental states. During the experiment, a Polaroid photo is taken of a scene, then placed face-down in front of the child while the scene is changed (e.g., an object is moved from the scene, or is swapped with another object). Children are then asked: "In the photo, where is the object?" Zaitchik's study showed that normally developing 3-year-old children find this task as difficult as standard false-belief tasks, whereas 5-year-olds find no difficulties in both tasks. Leslie and Thaiss (1992) demonstrated that all the children with autism with a verbal mental age ranging from 4.4 to 14.5 years passed the false photograph tests and the majority of them (77%) failed the false-belief tests. By contrast, normally developing children aged 4.5 years passed all tests. This finding has been replicated with analogous tasks, for example, using line drawings and maps, thus dismissing concerns that the children's inexperience with cameras might effect their response (Charman and Baron-Cohen, 1992, 1995; Leslie and Thaiss, 1992).

Another relevant study has shown that younger normally developing children not only fail 'out-of-date' mental state tests (e.g., false-belief tasks) but also 'out-of-date' non-mental states tests (e.g., 'screen task') that older children with autism are able to pass (Roth and Leslie, 1998). In the screen task children are first presented with a scene, then the objects are moved behind an opaque screen, and a new set of objects is placed in *front* of the child who watches a change of object location. The child is then asked to indicate the location of the object *behind* the screen. Both, children with autism older than 4 years and normal 3-year-old children, fail to represent mental states (false-belief task), however only the 3-year-olds fail to represent an out-of-date state of affairs (screen test).

The failure of 3-year-olds with either mentalizing tasks (false-belief), misrepresentation tasks (out-of-date photographs, drawings, maps) or executive control tasks (scenes hidden by a screen) all seem to reflect a performance limitation associated with executive functions not yet fully developed. By contrast, failure on false-belief tasks and success on all other tasks by children with autism reflect a selective impairment in understanding the mental states of others. Thus, children with autism show a competence limitation in mentalizing, rather than to executive functions processing problems.

Strong corroborative evidence comes from the subtle but effective manipulation of task instructions. Several investigators have modified the false-belief task so as to make it simpler – for instance, by making the question simpler, more specific and more pragmatically natural (e.g., Siegal and Beattie, 1991; Surian and Leslie, 1999). Surian and Leslie showed that 3.5-year-old normally developing children who failed the standard question "Where does Billy think the ball is" passed the modified question "Where is the *first* place Billy will look for his ball?" However, children with autism who performed at the level of the normal 3.5-year-old children on the 'think' question were not facilitated by the 'look first' question. According to the authors, the format of this question enables the prepotent response, which maintains that Billy has a true belief (as a default assumption), to be inhibited.

Developmental Time Course of Theory of Mind Skills

During the last decade developmental psychologists became very interested in investigating developmental ToM, namely, at what age children acquire the concept of belief.

Extended research showed that children younger than 4 years perform very poorly on several versions of the false-belief task (Wellman et al., 2001; Wellman and Liu, 2004; Milligan et al., 2007). Children's performance does not improve either if they are asked about where the puppet *will look* for the object, or where it *will say*, or *think*, or *know* that the object is (Wellman et al., 2001) or when children are requested to provide nonlinguistic answers (e.g., de Villiers and Pyers, 2002; de Villiers and de Villiers, 2000; Wellman et al., 1996).

However, these failures do not necessarily establish that younger children lack the ability to represent mental states of others. Toddlers at 2.7 years of age adjust their communicative behavior by taking into account whether their interlocutors have witnessed where an object has been hidden (O'Neill, 1996). Other evidence suggests that 3-year-olds have an implicit understanding of mental states (Garnham and Ruffman, 2001) or an understanding in which they have low confidence levels (Ruffman et al., 2001). When children aged 3 years are asked explicit test questions rather than standard ones, their performance improves (e.g., Siegal and Beattie, 1991; Surian and Leslie, 1999; Yazdi et al., 2006). Thus, there is evidence that children younger than 4 years do understand beliefs. Can infants represent mental states such as beliefs? Since infants cannot be tested with elicited-response false-belief tasks, recent investigations adopted spontaneous-response false-belief tasks in which infants' understanding of an actor's behavior to be guided by her true or false belief about an object hiding place is inferred by behavior they produce as they observe a scene unfold. In particular, the task measures infants' 'surprise,' that is, violation of expectation, indicated by longer looking times at an unexpected scene as opposed to an expected one. A study by Onishi and Baillargeon (2005) tested 15-month-olds and infants looked as the agent reached to the location where the agent falsely believed the object was hidden but looked longer when the agent reached to the other location where the toy was indeed hidden. An eye-tracking study by Southgate et al. (2007) demonstrated that 25-month-old infants correctly anticipate an actor's actions when these actions can be predicted only by attributing a false belief to the actor. Even infants, 13-months old, formed different expectations about a puppet's future behavior depending on whether the puppet had previously seen a desired object being put beyond a screen (Surian et al., 2007). A novel false-belief paradigm based on helping as an active behavioral response (Buttelmann, Carpenter and Tomasello

et al., 2005) showed that by infants, 18 months of age, successfully took into account the adult's belief in the process of attempting to determine his goal. Hence, a growing body of evidence strongly suggests that children before being able to pass elicited response false-belief tasks, have at least an implicit understanding of others' beliefs (Baillargeon et al., 2010; Csibra and Gergely, 2013; Leslie, 2005; Luo and Baillargeon, 2010; Luo, 2011). Interestingly, infants of 7 months of age, an age by which infants attribute goals and intentionality (Csibra, 2008) *automatically*, that is, *effortlessly*, compute other's beliefs and seem to maintain them in their minds even in the absence of the agent (Kovacs et al., 2010). Recently (Heyes, 2014) has argued to maintain a certain caution in interpreting results that infants appreciate that others have false beliefs. Indeed, she argues that many findings can be explained by the operation of domain-general processes and in terms of 'low-level novelty'. However, the adoption of different experimental strategies could advance our knowledge of developmental theory of mind.

Brain Basis and Theory of Mind

The authors of the infants' research findings often underline that infants compute 'implicitly,' 'effortlessly,' 'automatically,' or ' online' mental states. These results give support to the suggestion, which was postulated during the early years of research in ToM, of the existence of a specialized neurocognitive mechanism that was engaged when computing mental states and that was selectively impaired in individuals with autism (Leslie, 1987). From the philosophical perspective, the existence of a dedicated 'system' was also postulated. According to Dennett (1987) we are constantly bombarded with sensory inputs that alert our system to the presence of an agent, inputs that make us adopt the so-called 'intentional stance' toward that agent in order to understand and predict its behavior. In the last decade functional brain imaging studies have indeed attempted to identify brain pathways involved during mental states attribution, both in healthy individuals and in people with autism. A neuroimaging study in particular has examined brain activation during an online attribution of mental states to geometrical shapes moving silently on a white background (Castelli et al., 2000) as to avoid any language-based or familiarity confounds on the task. Indeed, the exposure to simple, nonverbal stimuli evoked in participants' mental state attribution by their kinetic properties alone. Over the past decade, a dedicated brain network for ToM has been confirmed by a series of brain imaging studies (Frith and Frith, 2003; Saxe et al., 2004). Most of these studies implicate activation in medial frontal and temporoparietal cortex (STS) and the temporal poles/amygdala complex. As expected, this system shows malfunction in autism (Castelli et al., 2002; see also Zilbovicius et al., 2006; Kana et al., 2009). The specific roles of the frontal and posterior components of the ToM network are still debated. However, recent studies on brain connectivity indicate that these cortical areas are connected during ToM processing, but are disrupted in the brain of individuals with autism (Just et al., 2013).

Bibliography

Apperly, I.A., 2008. Beyond simulation-theory and theory-theory: why social cognitive neuroscience should use its own concepts to study "theory of mind". Cognition 107, 266–283.

Baron-Cohen, S., Leslie, A.M., Frith, U., 1985. Does the autistic child have a "theory of mind"? Cognition 21, 37–46.

Brentano, F., von, 1970. Psychology from an Empirical Standpoint. Rutledge, New York.

Baron-Cohen, S., 1991. The development of a theory of mind in autism: deviance and delay. Psychiatric Clinics of North America 14 (1), 33–51.

Baron-Cohen, S., 1992. Out of sight or out of mind: another look at deception in autism. Journal of Child Psychology and Psychiatry 33 (7), 1141–1155.

Baillargeon, R., Scott, R.M., He, Z., 2010. False-belief understanding in infants. Trends in Cognitive Sciences 14, 110–118.

Brothers, L., 1997. Friday's Footprint: How Society Shapes the Human Mind. Oxford University Press, Oxford.

Charman, T., Baron-Cohen, S., 1992. Understanding drawings and beliefs: a further test of the metarepresentation theory of autism: a research note. Journal of Child Psychology and Psychiatry 33, 1105–1112.

Charman, T., Baron-Cohen, S., 1995. Understanding models, photos, and beliefs: a test of the modularity thesis of metarepresentation. Cognitive Development 10, 287–298.

Call, J., Tomasello, M., 2008. Does the chimpanzee have a theory of mind? 30 years later. Trends in Cognitive Sciences 12 (5), 187–192.

Csibra, G., 2008. Goal attribution to inanimate agents by 6.5-month-old infants. Cognition 107 (2), 705–717.

Csibra, G., Gergely, G., 2013. Teleological understanding of actions. In: Banaji, M., Gelman, S.A. (Eds.), Navigating the Social World: What Infants, Chidren, and Other Species Can Teach Us. Oxford University Press, New York, pp. 38–43.

Castelli, F., Happè, F., Frith, U., Frith, C., 2000. Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. Neuroimage 12, 314–325.

Castelli, F., Happè, F., Frith, U., Frith, C., 2002. Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. Brain 125, 1839–1849.

Dennett, D.C., 1978. Beliefs about beliefs. Behavioural and Brain Sciences 1, 568–570.

Dennett, D.C., 1987. Three kinds of intentional psychology. In: Dennett, D.C. (Ed.), The Intentional Stance. MIT Press, Cambridge, pp. 43–68.

Frith, U., Frith, C.D., March 2003. Development and neurophysiology of mentalizing. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 358 (1431), 459–473.

Frith, U., Frith, C., 2010. The social brain: allowing humans to boldly go where no other species has been. Philosophical Transactions of the Royal Society B 365, 165–175.

Frith, U., Frith, C., 2012. Mechanisms of social cognition. Annual Review of Psychology 63, 287–313.

Fisher, N., Happé, F., Dunn, J., 2005. The relationship between vocabulary, grammar, and false belief task performance in children with autistic spectrum disorders and children with moderate learning difficulties. Journal of Child Psychology and Psychiatry 46, 409–419.

Garnham, W.A., Ruffman, T., 2001. Doesn't see, doesn't know: is anticipatory looking really related to understanding of belief? Developmental Science 4, 94–100.

Gallese, V., Goldman, A., 1998. Mirror neurons and the simulation theory of mind-reading. Trends in Cognitive Sciences 2, 493–501.

Gopnick, A., Wellman, H.M., 1992. Why the child's theory of mind really is a theory. Mind & Language 7 (1–2), 145–171.

Happè, F., 2003. Theory of mind and the self. Annals of the New York Academy of Sciences 1001, 134–144.

Heyes, C., 2014. False belief in infancy: a fresh look. Developmental Science 22 (2), 1–13.

Hogrefe, G., Wimmer, H., Perner, J., 1986. Ignorance versus false belief: a developmental lag in attribution of epistemic states. Child Development 567–582.

Heider, F., Simmel, M., 1944. An experimental study of apparent behavior. American Journal of Psychology 57, 243–259.

Just, M.A., Keller, T.A., Kana, R.K., 2013. A theory of autism based on frontal-posterior underconnectivity. In: Just, M.A., Pelphrey, K.A. (Eds.), Development and Brain Systems in Autism. Psychology Press, New York, pp. 35–63.

Kovacs, A.M., Teglas, E., Endress, A.D., 2010. The social sense: susceptibility to others' beliefs in human infants and adults. Science 330, 1830–1834.

Kana, R.K., Keller, T.A., Cherkassky, V.L., Minshew, N.J., Just, M.A., 2009. Atypical frontal-posterior synchronization of theory of mind regions in autism during mental state attribution. Society for Neuroscience 4, 135–152.

Leslie, A.M., 1987. Pretense and representation: the origins of theory of mind. Psychological Review 94, 412–426.

Leslie, A.M., Thaiss, L., 1992. Domain specificity in conceptual development: neuropsychological evidence from autism. Cognition 43, 225–251.

Leekam, S.R., Perner, J., 1991. Does the autistic child have a metarepresentational deficit? Cognition 40, 203–218.

Leslie, A.M., Frith, U., 1988. Autistic childrens understanding of seeing, knowing and believing. British Journal of Developmental Psychology 6, 315–324.

Leslie, A.M., 2000. 'Theory of mind' as a mechanism of selective attention. In: Gazzaniga, M.S. (Ed.), The New cognitive Neurosciences, second ed. MIT Press, Cambridge, MA, pp. 1235–1247.

Leslie, A.M., 2005. Developmental parallels in understanding minds and bodies. Trends in Cognitive Sciences 9, 459–462.

Luo, Y., Baillargeon, R., 2010. Toward a mentalistic account of early psychological reasoning. Current Directions in Psychological Science 19, 301–307.

Luo, Y., 2011. Do 10-month-old infants understand others' false beliefs? Cognition 121, 289–298.

Milligan, K., Astington, J.W., Dack, L.A., 2007. Language and theory of mind: meta-analysis of the relation between language ability and false-belief understanding. Child Development 78 (2), 622–646.

O'Neill, D.K., 1996. Two-year-old children's sensivity to parent's knowledge state when making requests. Child Development 67, 659–677.

Onishi, K.H., Baillargeon, R., 2005. Do 15-month-old infants understand false beliefs? Science 308, 255–258.

Premack, D., Woodruff, G., 1978. Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences 1 (4), 515–526.

Perner, J., Frith, U., Leslie, A.M., Leekam, S.R., 1989. Exploration of the autistic child's theory of mind: knowledge, belief, and communication. Child Development 60, 688–700.

Perner, J., Ruffman, T., 2005. Infants' insight into the mind: how deep? Science 308, 214–216.

Rizzolatti, G., Fadiga, L., Gallese, V., Fogassi, L., 1996. Premotor cortex and the recognition of motor actions. Brain Research Cognitive Brain Research 3, 131–141.

Roth, D., Leslie, A.M., 1998. Solving belief problems: towards a task analysis. Cognition 66, 1–31.

Russell, J., Hill, E.L., 2001. Action monitoring and intention reporting in children with autism. Journal of Child Psychology and Psychiatry 42, 317–328.

Ruffman, T., Garnham, W., Import, A., Connoly, D., 2001. Does eye gaze indicate implicit knowledge of false belief? Charting transitions in knowledge. Journal of Experimental Child Psychology 80, 201–204.

Saxe, R., Xiao, D.K., Kovacs, G., Perrett, D.I., Kanwisher, N., 2004. A region of right posterior superior temporal sulcus responds to observed intentional actions. Neuropsychologia 42, 1435–1446.

Siegal, M., Beattie, K., 1991. Where to look first for children's knowledge of false belief. Cognition 38, 1–12.

Surian, L., Leslie, A.M., 1999. Competence and performance in false belief understanding: a comparison of autistic and normal 3-year-old children. British Journal of Developmental Psychology 17, 141–155.

Southgate, V., Senju, A., Csibra, G., 2007. Action anticipation through attribution of false belief by 2-year-olds. Psychological Science 18, 587–592.

Surian, L., Caldi, S., Sperber, D., 2007. Attribution of beliefs by 13-month-old infants. Psychological Science 18 (7), 580–586.

Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H., October 2005. Understanding and sharing intentions: the origins of cultural cognition. Behavioral and Brain 28 (5), 675–691.

de Villiers, J.G., 2005. Can language acquisition give children a point of view?. In: Why Language Matters for Theory of Mind. Oxford University Press, pp. 186–219.

de Villiers, J.G., Pyers, J.E., 2002. Complements to cognition: a longitudinal study of the relationship between complex syntax and false-belief-understanding. Cognitive Development 17 (1), 1037–1060.

de Villiers, J.G., de Villiers, P.A., 2000. Linguistic determinism and the understanding of false beliefs, in children's reasoning and the mind. Psychology Press 191–228.

Wellman, H.M., Liu, D., 2004. Scaling of theory-of-mind tasks. Child Development 75, 523–541.

Wimmer, H., Perner, J., 1983. Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 13 (1), 103–128.

Wimmer, H., Hartl, M., 1991. Against the Cartesian view of mind: young children's difficulty with own false beliefs. British Journal of Developmental Psychology 9, 125–136.

Williams, D.M., Happè, F., 2009. What did I say? versus what did I think? Attributing false beliefs to self amongst children with and without autism. Journal of Autism and Developmental Disorders 39 (6), 865–873.

Williams, D., Happe, F., 2010. Representing intentions in self and other: studies of autism and typical development. Developmental Science 13 (2), 307–319.

Wellman, H.M., Cross, D., Watson, J., 2001. Meta-analysis of theory-of-mind development: the truth about false belief. Child Development 72 (3), 655–684.

Wellman, H.M., Hollander, M., Schult, C.A., 1996. Young children's understanding of thought bubbles and of thoughts. Child Development 67 (3), 768–788.

Yazdi, A.A., German, T.P., Defeyter, M.A., Siegal, M., 2006. Competence and performance in belief-desire reasoning across two cultures: the truth, the whole truth and nothing but the truth about false belief? Cognition 100, 343–368.

Zaitchik, D., 1990. When representations conflict with reality: the pre-schoolers problem with false beliefs and false photographs. Cognition 35, 41–68.

Zilbovicius, M., Meresse, I., Chabane, N., Brunelle, F., Samson, Y., Boddaert, N., 2006. Autism, the superior temporal sulcus and social perception. Trends in Neurosciences 29, 359–366.

Mind–Body Dualism

Rom Harré, Georgetown University, Washington, DC, USA; and Linacre College, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Suffering', 'remembering', 'weighing 55 kilos', 'being in the doorway' are among the properties that are routinely assigned to individual people. While mental processes, such as remembering and mental states, such as being in pain, are ascribed only to people and beings similar to them, weight and location are attributes of many kinds of things that are certainly not people. Should this distinction in kinds of attributes shape our efforts to study people by considering each person to be a composite of two distinct substances – a mind and a body? Rene Descartes' famous argument for substance dualism subsequently appeared in various versions but it has been the focus of searching criticisms. Several alternative proposals for organizing the study of the attributes of persons into a coherent body of knowledge have been proposed. These include materialism or physicalism, that there is only one relevant substance, material stuff; and various versions of personalism, that 'person' is the primitive core of all descriptions of human life, conceptually prior to both 'mind' and 'body'. Wittgenstein's strategy in the face of the rerunning of the mind–body debates over centuries was to suggest that *both* sides of the dichotomy that has shaped these debates should abandoned and new distinctions introduced.

Dualism of Substances

Centuries of effort to understand the nature and activities of human beings have acknowledged the radical distinction between the nature of mental attributes and processes in which people engage and the materiality of their bodies, including their brains. Many, perhaps every state of mind is correlated with some state of the body, while only some states of the body are correlated with states of mind. Until recently most attempts to give a coherent account of human nature in all its fullness have been based on the presumption that the distinction between mental states and states of the body reflects a distinction between the mind as an immaterial substance and the body as a material substance. Nouns typically refer to substances, enduring stuffs the role of which in lay and scientific discourses, is to be the substratum to which attributes are ascribed. Thus, the person is taken to be a doublet, one component of which is mental stuff and the other material stuff. According to the Cartesian tradition to be discussed in detail below the properties of mental stuff (mind) and the properties of body stuff (matter) are utterly different. Not only is a person a substantival doublet but also, in a way that was never satisfactorily explained, a unity of mind and matter.

Some dualists have supposed that though in life mind and body are intimately conjoined the mind as a substance could be detached from the matter of the body and continue to exist with nearly all of the mental attributes that it had as a component of a person. Others have supposed that mental stuff would perish along with the material stuff with which it was conjoined in the life of the individual person. Mind–body dualism does not, without some additional premises, entail the immortality of the soul.

Descartes' Proposal

Descartes exposition in the second and sixth of his *Meditations on First Philosophy* (Descartes, 1641, in Latin and 1647 in French) includes some important insights and led to several versions of the dualist account of human nature. The first step in Descartes' analysis is to establish at least one secure truth about his own nature in the midst of doubts about what he previously thought he knew for sure. The only thing he cannot doubt is that he is doubting, and so he is sure he must be a thinking thing. In the *Sixth Meditation* he considers the status of his knowledge of material things including his own body. In a frequently quoted passage he declares "Nature teaches me, by these sensations of pain, hunger, thirst and so on, that I am not merely present in my body as a sailor is present in a ship but that I am very closely joined and, as it were, intermingled with it, so that I and the body form a unit" (Descartes, [1656]1981). However, knowledge of the truth of experiences seemingly of things other than myself is a matter for the mind, so as to warn him of what is safe and what is harmful 'for the composite of which the mind is a part' (*loc. cit.*). The body as a material substance is divisible, but the mind as a mental substance has no parts, nor is it affected by every part of the brain. Only divisible entities could decay and perish. The indivisible soul must therefore be immortal.

The depth of his original doubts depended on his playing with the idea that there might be a malevolent god deceiving him – but God is not a deceiver. Descartes' sensory knowledge of the material things he perceives in his environment are for the most part for his own good.

The root idea of a dualism of substance derived from the distinctiveness of mental attributes from the material attributes of the body deeply influenced philosophical psychology and continues to do so. If mind and body were distinct substances with no properties in common how could there be an interaction between them. This interaction is displayed in perception, a process by which matter acts on mind. It is also displayed in deliberate action, a process in which mind acts on matter? How could one be sure another person was a conscious being with a mental life if my contacts with this person were limited to appearances in the material world? Descartes himself noted that what appeared to be the people in the street might turn out to be automata dressed as people. The problem of

mind–brain interaction and the problem of other minds are still with us.

Early Criticisms and Revisions of Cartesianism

Pierre Gassendi, one of the authors Descartes invited to offer objections to his proposals, pointed out how empty was the idea of a 'the mind as a thinking thing'. Using the criterion of the ability to form a clear and distinct idea of something as a guide to the nature of that thing Gassendi argued that declaring that he had a clear and distinct idea of the mind as a 'thinking thing' this tells us nothing about the nature of the mind so it cannot be a clear and distinct idea of the mind. "It is far from true that a man who sees clearly that the mind thinks therefore sees clearly the entire intimate nature of mind" (Gassendi, 'Logic'. In: Bush, C.B., 1972: p. 190).

Gassendi developed a more complex pattern of relationships between mental and material aspects of the person and the duality of substances. Not all cognitive processes were carried out in the immaterial mind, only those of high intellectual abstraction. For example, he suggested that classifying and other low level cognitive processes were 'mechanical' and occurred in the brain. In *Syntagmata* (Gassendi, 1971[1658]: p. 336) he describes the *soul* as more or less the form of the body, rather as Aristotle might have described it. "It is such that in the same man it not only gives life, feels and moves, but also understands, reasons, wonders, probes abstruse questions, speaks, laughs, makes laws, discovers skills, and so on." 'Thought' is 'separate' only by abstraction from 'matter', while 'mind' is separate from matter in fact, for example as God and as intelligence are actually separate from matter. The *immortal soul* was not an intimately interwoven component of the unified person, but a stand-alone purely intellectual correlate that changed in synchrony with changes in the human being.

A century later David Hartley proposed a kind of dualism of process. 'Vibrations' of the small particles in the nerves and brain were in correspondence with matching streams of ideas. In Proposition 11 of Book 1 (Hartley, 1749: p. 8) he wrote, "… whatever changes are made in this substance [the white medullary substance] corresponding changes are made in our ideas, and vice versa". By mid-eighteenth century Hartley had no need to argue for the mind–brain dualism realized as vibrations and idea, which is the taken for granted metaphysical foundation for his neuropsychology.

Recent Criticisms

The Cartesian version of mind–body dualism has been under attack since the time of its original proposal in the seventeenth century. The basis of the Cartesian route to substance dualism is the seeming radical diversity of the kinds of properties that can be ascribed to a human being as person. There are those that incorporate extension, and those that do not. The essence of matter is extension so there must be a material substance in which they inhere. The parallel analysis of thinking as an unextended attribute of something leads to the mind as an immaterial substance. If this categorization of personal properties is defective the Cartesian case for mind–body dualism is undermined.

An indirect criticism of the Cartesian argument involves attempts to support the claim that there is only one substance, which is neither material nor immaterial, but has both mental and material properties, and indeed some properties which require both kinds of concepts in their analysis. The person is the primitive unanalyzable concept of choice for philosophers, such as P.F. Strawson (1959) and psychologists such as William Stern, whose 'personalism' has been revived by Lamiell (2010).

In his *Individuals* (1959) P.F. Strawson argues for the primitive role of the concept of 'person', making the point that "the Cartesian error is just a special case of the more general error … of thinking of the designations or apparent designations of persons as *not* denoting precisely the same thing or entity for all kinds of predicates ascribed to the entity designated" (Strawson, 1959: p. 105). All P predicates 'imply the possession of consciousness on the part of that to which they are ascribed'. Importantly, Strawson includes 'going for a walk' which only the conscious embodied person can do, as well as 'being in pain' which only the embodied but conscious person can experience. So M or material predicates are also ascribed to the primitive being, the person. There is a dualism but it is of predicates not of substances. People are the subjects of both M and P predicates, though only of human beings and those like them can P predicates be ascribed.

There cannot be a mind–body problem if there is no real interaction between monads. Gottfried Leibniz 'solved' the problem of how the unfolding of each windowless monad, roughly a person's and brain and that person's mind, each as a collection of inferior with a supreme monad which though utterly distinct, seem to stand in causal relations one to another. Leibniz solution to the impossibility of such a relation is the idea that each monad reflects the state of all the other monads as if there had been a causal relation between them. By the principle of a divinely preestablished harmony between each and all of them they unfold in apparent synchrony. Leibniz declared that if we have a complete conception of a substance it will be "sufficient … for the deduction of all the predicates of which the substance is or may become the subject" (Leibniz, 1987[1686]: p. 13, Section VIII). So all the thoughts and feelings of a mind will appear in due order without the intervention of causal influences from the rest of monads that constitute the human being and the 'surrounding' universe.

Dualism of Domains

Public and Private Cognition

During the twentieth century attention to the means by which cognitive processes were performed led to a realization that thinking could be carried out by group of people in conversation. Remembering as reminiscing and deciding as negotiating are cognitive processes that occur interpersonally. There was a new conception of dualism to take account of, that between public-social cognition and private-individual cognition. Public-social cognition is enacted in the material world, but is relevant to psychology only in so far as material entities, such as verbal utterances, are endowed with meaning. Private-individual cognition is enacted subjectively, but is relevant to human life only in so far as the meaning of privately

formulations of thoughts depend on spatiotemporal location of the embodied being. On this view, human bodies are sites for persons – which may be occupied but may be deserted. If the responsible people become convinced that there is no chance that human, that is meaningful activities, can be recovered from a human being in a vegetative state, that is there is no longer a person there, the life support can be terminated (Grattan, 2003: pp. 113–145).

Lev Vygotsky (1978) made this distinction the basis of his developmental psychology. An infant enters a world of conversations and meaningful actions, then, coached by family members, it acquires bit by bit the necessary skills for public–social interactions, gradually privatizing them and so acquiring a 'mind', which is in effect a repertoire of skills and effective practices. As the basis for cognitive psychology it points toward the realization that many cognitive processes have public-collective realizations and private-individual realizations. Some are realized only in the public-collective domain among people and others only in the private-individual domain, within oneself.

Recent Proposals of Alternative Distinctions

In a bold attempt to break the chain of endless debates around Cartesian impasses distinction, Wittgenstein (1953: §281) suggests that we should abandon the mind–body distinction that underpins them. The problems of how the states of a material substance can affect an immaterial substance (perception) and how the state of an immaterial substance can affect a material substance (deliberate action) are insoluble not because they are very difficult but because the terms in which they have been posed are the source of the intractability of all such problems. Not only should the distinction be abandoned but also with it both alternatives. Ceasing to think of human thought and action in mentalistic, immaterial terms, should not be followed by attempts to achieve satisfactory understandings of human thought and action in material terms. We should adopt some other nontoxic dualisms – for example, alive–dead, morally protected – morally vulnerable, and so on. We will examine several possibilities for this strategy for giving an acceptable account of the obvious fact of mental–physical dualism of processes and functions without requiring a foundation in the dualism of substances that has been the basis of attempts to link mental processes with the correlated physical processes that occur in every human being and in the members of other species capable of cognition.

However, a certain naive version of materialism that runs counter to Wittgenstein's proposal has appeared in 'neuropsychology'. The problem of how mental activities are related to neural activities is sidelined by using mentalistic predicates to refer to processes in organs of the brain – thus 'the hippocampus remembers' or the 'frontal lobes reason'. Bennett and Hacker (2003) have shown how this way of bypassing the Cartesian problem commits a 'mereological fallacy'. It is a fallacy to use a predicate of a part of an entity that gets its meaning as it is ascribed to the whole entity. Only persons can remember. This is not a matter of fact but a feature of the semantics of 'remembering'. Discoveries of the role of the hippocampus in the process of remembering make no difference to the primary way in which 'remembering' gets its meaning.

Hybrid Psychology

Following the implicit advice offered by Wittgenstein and turning to the Vygotskian framework of cultural-historical-instrumental psychology for significant dichotomies, suggests a very different distinction – that between tasks and the tools needed to carry them out. Tasks, such as playing tennis, mowing the lawn, and solving crossword puzzles, are defined culturally. The tools needed to carry them out, including, tennis racquets, lawn mowers, and brains, are defined materially. There is a strong conceptual link between tools and tasks in that while tennis racquets could be used as snow shoes their role as tennis racquets shapes their material attributes and depends on the cultural process they are adapted to. The same principle holds good for organs of the brain, such as the auditory cortex, which is adapted biologically to differentiating harmonic intervals, but has recently been readapted in each listener to hear the cultural artifact, the equi-tempered scale, as perfectly harmonious. As a general rule the criteria for identity of specific brain organs are derived from the function they play in cultural instrumental activities. We can identify the hippocampus as the site of certain memory processes because we already know what remembering is as a cultural instrumental phenomenon.

Supervenience

The well-established idea of emergent properties is based on a distinction between a base level domain of entities, which can be arranged in such a way that the composite so formed has at least some properties different from those of any of the entities that make it up. The distinctive higher level properties are said to be 'emergent'. For human beings, the base level properties are material, the constituents of 'body', while the higher level properties are mental, constituents of 'mind'. In some contemporary approaches to psychology they are said to be 'discursive', that is to consist of meanings that are ordered by reference to norms, conventions and story lines (Harré and Moghaddam, 2012). If it turns out that some body and brain states and processes could exist without there being associated mental states and processes but no mental states and processes could exist without there being associated body and brain states, *and* all differences in the one are correlated with differences in the other, the former are said to *supervene* on the latter. The relation between the two groups of states and processes is *supervenience*. If mental properties supervene on material properties then there cannot be a difference in mental properties unless there is a difference in material properties. I change from thinking about apples to thinking about pears, and at the same time (more or less) there is a change from one brain state to another. Or to put the matter otherwise "there cannot be two events alike in all physical respects but differing in some mental respect" (Davidson, 1980: p. 214).

Philosophical studies of supervenience involve many sided attempts to identify a relationship that takes account of the fact that the concepts that appear in the discursive practices of thinking and acting are manifestly distinct from those that are used to describe processes in the body while undermining the

possibility that the states and processes to which they refer inhere in distinct substances (McLaughlin and Bennett, 2011).

For 'weak' supervenience there is some material state for every mental state, fulfilling the 'differences' criterion. For 'strong' supervenience the same type of material state is correlated with each type of mental state. Weak supervenience of mind on body may be true weakening the two substances dualism. Nevertheless even to formulate the supervenience thesis independent criteria of identity of elementary items in each level domain may be required. There may be no explanation for the fact that research can demonstrate that there is some material correlated with each mental state or process. Only if we accept strong supervenience is the existence of a supervenience relation a ground for physicalism, the doctrine that there are only material substances and their states. Even then how entities, states, and process in the two domains are related remains unexplained.

Turing's Dualism

The advent of machine computation and the successful mimicking of some mental processes, at least as concerns input and output, by computing machines have suggested another take on the doctrine of mind–body dualism.

Impressed by the possibilities he saw in the construction and possible modes of operation of computing machines Alan Turing (1950) proposed a way of thinking about human cognition that can be condensed into a pair of analogies.

Thinking–Brain: Computing–Computer

To explore this scheme two analogies would need to be followed through. The first explores the brain–computer analogy. How like is a computer when programed to carry out an activity similar to the human brain when a person has learned some mental skill? The second explores the thinking–computing analogy. How like is machine computation to the human practices of thinking? Here we have another version of mind–brain dualism but which requires only one substance, the matter of the computing machine and its physical layout of registers, serial busses, etc.

This pair of analogies has been famously and controversially contested by John Searle (1980). In essence Searle's argument depends on the distinction between the physical tokens with which machine computation is carried on, and the meanings these must have if they are to play a part in human thought and action. There is a duality of meaningful entities and meaningless ones, but the duality is not between substances. It is between meaningless material entities as bits of stuff and material entities as symbols, entities that are meaningful to the members of a certain culture.

Conclusions

In the seventeenth century few doubted the immortality of the soul. The material body soon decayed. The key psychological attributes of the person such as memory could not decay if the soul of a particular human individual, his or her active intellectual, rational part, were to survive the dissolution of the body. They must therefore be attributes of something that is indissoluble, namely an immaterial component, the soul. It follows that a human being is a union of body and soul, of two substances. For this argument to work the attributes of each must be distinct from those of the other. And now the insoluble problems that bedevil the many attempts to make mind–brain dualism the basis of psychology appear. Resisting the temptation to presume that every noun must refer to something substantive many philosophers and psychologists have dropped 'the mind' and persisted with the monism of physicalism. Mental states and processes are declared to be nothing more than material states and processes (Churchland, 1984). This Procrustean bed so distorts the criteria for identifying cognitive, emotional, and intentional states of a person that it has seemed to many to be quite inadequate as the basis for a way to understand human life.

Persons might be singularities in a material social environment, to each of which we assign both mental and material attributes the meanings of many of which require both material and mental criteria. Moreover, there may be many distinctions other than the substantival, which could prove useful for understanding human life and the relations between the items in the broad spectrum of attributes that characterize a person.

See also: Identity and Identification, Social Psychology of; Person and Self: Philosophical Aspects; Physicalism and Alternatives; Self: Philosophical Aspects.

Bibliography

Bennett, M.R., Hacker, P.M.S., 2003. Philosophical Foundations of Neuroscience. Blackwell, Oxford.
Churchland, P., 1984. Matter and Consciousness. MIT Press, Cambridge, MA.
Davidson, D., 1980. Mental events. In: Davidson, D. (Ed.), Essays on Actions and Events. Clarendon Press, Oxford.
Descartes, R., 1641[1988]. Meditations on the first philosophy (J. Cottingham, R. Stoothoff, D. Murdoch, Trans.). In: The Philosophical Writings of Descartes, vol. II.
Gassendi, P., 1971[1658]. Selected Works of Pierre Gassendi (C.B. Bush, Trans.). Johnson Reprint, New York and London.
Grattan, K., 2003. The dispute over the fate of Terrri Sciavo. In: Harré, R., Moghaddam, F.M. (Eds.), The Self and Others. Praeger, Westport, CT.
Harré, R., Moghaddam, F.M., 2012. Psychology for the Third Millennium. Sage, London.
Hartley, D., 1749. Observations on Man: His Frame, His Duty and His Expectations. Leak and Frederick, London.
Kim, J., 1993. Concepts of supervenience. In: Kim, J. (Ed.), Supervenience and the Mind. Cambridge University Press, Cambridge.
Lamiell, J.R., 2010. William Stern (1871–1938): A Brief Introduction to His Life and Work. Pabst Science Publishers, Berlin.
Leibniz, G., 1987[1686]. Discourse on Metaphysics (G. Montgomery, Trans.). Open Court, La Salle, IL.
McLaughlin, B., Bennett, K., 2011. Supervenience. In: Zalta, E.N. (Ed.), Stanford Encyclopedia of Philosophy. (Winter 2011). The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford, CA.
Searle, J.R., 1980. Minds, brains and programs. The Brain and Behavioral Sciences 3, 417–424.
Strawson, P.F., 1959. Individuals. Methuen, London.
Turing, A., 1950. Computing machines and intelligence. Mind 54, 433–460.
Vygotsky, L.S., 1978. Mind in Society. Harvard University Press, Cambridge, MA.
Wittgenstein, L., 1953. Philosophical Investigations (G.E.M. Anscombe, G.H. von Wright, Trans.). Blackwell, Oxford.

Mindfulness

Katherine Gardhouse and Zindel Segal, University of Toronto Scarborough, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In recent years, the concept of mindfulness has figured prominently in the development of innovative psychological interventions for physical and mental health. Mindfulness-based interventions have been shown to reduce recurrence of depression, stress, chronic pain, anxiety, and disordered eating, as well as to improve concentration and productivity. Mindfulness practices encourage participants to recognize automatic thought and behavior patterns and in doing so permit a more choiceful manner of responding to, and exploring, experience. The opportunity to relate to symptoms in this way may lead to significant shifts in both behavioral and physiological measures, ultimately leading to health improvements.

Mindfulness is a multifaceted concept that is often defined as paying purposeful attention to the present moment with a compassionate and nonjudgmental attitude toward one's experience (Kabat-Zinn, 2003/2013). Mindfulness is a kind of meta-awareness, an approach to consciousness, a skill, and a tool that can be used to observe thoughts and sensations of the body as they unfold. For example, when eating, one can focus directly on the richness of the flavors and the unique textures of the food while chewing, or instead they can attend to the highway while driving as they eat 'on the go.' In this second case, the experience of the food is not the focus of attention, although the individual may be aware to some degree that he is eating. Here, attention is divided between tasks. Mindfulness occurs in the first condition – when full attention and awareness are focused on the present moment, but this can extend broadly to all thoughts, perceptions, and experiences.

Mindfulness is the capacity of awareness that can be focused on inner and outer experiences as they are moment by moment. Mindfulness emphasizes self-compassion and acceptance so that the experiences one has, automatic or mindful, are not attached to in order to promote a renewed and present-centered approach. When not being mindful, individuals will carry out behaviors without direct awareness, or get lost in automatic thoughts that take them away from focusing on the immediate present. While the automatic processes of the mind are effective task managers, freeing up resources to plan and execute, they can also have negative consequences for health when they cause one to neglect immediate needs or to engage with the environment in suboptimal ways.

Mindfulness meditation has been the foundation of Buddhist practices for centuries but it is also considered to be an inherent human capacity. Mindfulness is not necessarily some enlightened experience only attained by those who practice, but it is also the occasioned attentional stance that occurs when one becomes conscious of thoughts, emotional states, or body sensations. Everyone has mindful moments, but some individuals are more naturally disposed to, or practice, being mindful and this may have consequences for health and well-being. Individuals who score higher on measures of mindfulness – who are more in tune with their inner and outer experiences – report greater levels of self-esteem, positive affectivity, optimism, self-actualization, autonomy, competence (Brown and Ryan, 2003), conscientiousness, agreeableness, self-compassion (Giluk, 2009), life satisfaction, and emotional intelligence (Baer et al., 2006; Brown and Ryan, 2003). Mindfulness has also been inversely associated with depression and anxiety, avoidance coping, negative affect, rumination, self-consciousness, anger, difficulties with self and emotion regulation, and neuroticism (Baer et al., 2006).

In recent years, there has been an explosion of research into mindfulness-based interventions (MBIs) for psychological and physiological illnesses (Grossman et al., 2004). These intensive 8-week programs have been shown to reduce depressive relapse, pain, anxiety, sleep disturbance, headache, and improve quality of life in both clinical and nonclinical samples (e.g., Keng et al., 2011; Hofmann et al., 2010; Chiesa and Serretti, 2009). Neuroimaging work by Farb et al. (2007) has shown that participants who complete an 8-week mindfulness course exhibit reduced activation in brain regions associated with narrative, or future thinking, and increases in brain region activity associated with present experiential thinking patterns, whereas wait-listed controls do not. This suggests that mindfulness practices have the potential to support fundamental shifts in the way we relate to information, allowing participants to adopt different modes of processing in a relatively short period of time. These transformations in neural activity may result in mindfulness developing into a more stable individual trait (Chiesa et al., 2010). Although the research is still in its infancy, these are some promising findings in the field of health.

Automaticity: To Do or Not to Do

Over the past hundreds of thousands of years, humans have benefitted from evolution's blind adaptations in many ways, including our unrivaled intellect. It seems, however, that the environment we have more recently shaped has some unintended consequences for our minds. While the more ancient problem-solving or 'doing' mode of processing was well suited for hunting, gathering, and living in small groups, it is called upon all too often in modern civilization and can run us into exhaustion. Today, life is no longer about what is directly in front of us – about direct conflict and resolution (e.g., run away from the tiger, chase the deer, make a fire, etc.) – but instead we

are able to consider past and future things like tomorrow's conference meeting, retirement, tuition payments, the dinner party on the weekend, or the health of family. In today's climate, there is much 'to do' and, for better or for worse, we possess a mode for just that. As such, the automatic doing mode is turned on constantly to help us 'get things done'; or so we think.

The automatic processes that dominate behavior and thoughts in *mindless* states can ultimately guide day-to-day functions without awareness or consideration. Everyone experiences mindless moments, perhaps arriving at a destination without remembering the drive, or sending a 'quick' e-mail only to spend an hour responding to messages or getting distracted on the Internet. We are so easily called upon to engage with our surroundings and while this 'doing' mode can be a very effective resource manager, it can also tap into and overextend resources in unnecessary and sometimes stressful ways without really seeking permission to do so. These processes disrupt the connection to the present by taking the mind to distant places past or future (i.e., temporal expansion), or by prioritizing endless preoccupations and to-do lists over immediate reality. Thoughts that occur sporadically in the mind can take over and reduce the ability to focus, flip one's mood, or cause one to overthink the small things. There are, however, other ways to operate. We do not have to be a slave to the busy 'doing' mode of the mind, and it is on this dimension that mindfulness works to reawaken individuals to a more purposeful perspective.

Mindfulness teaches participants to adopt a different mode, the 'being' mode, which allows individuals to observe things as they are instead of automatically engaging with their surroundings or thoughts. This shift presents a different way of sampling the environment, and a perspective that allows one to assess situations before mindlessly responding. In adopting this kind of practice, when an individual sees a long series of new e-mails in his inbox, he can prevent himself from immediately responding and ask whether now is a good time or not. Similarly, when one starts to ruminate about an embarrassing moment on the weekend, she can catch herself and bring her attention back to what is immediately present. Mindfulness practices teach participants to move beyond the wall of thoughts that flood the mind, and the response patterns that possess the body, to be more active passengers in their lives. Evidently, this may have substantial consequences for health.

Mechanisms of Mindfulness and Relapse Prevention in Mental Health

While even the most healthy of individuals are subject to the automatic thralls of the mind, the doing mode can become especially problematic to those with depression and anxiety disorders. This is in part because automatic and habitual negative thinking patterns dominate their headspace. Segal et al. (1995/2013) refer to the nonadaptive form of the doing mode as the driven-doing mode. In this mode, there is no immediate action or resolution to be taken; like a train without breaks, there is no stopping it from speeding along the tracks until it derails. While the doing mode, although often automatic, actually works on resolving real discrepancies (e.g., e-mailing, washing the dishes, shopping, and doing the laundry), the driven-doing mode strives to answer questions or decipher issues that have no apparent or immediate answers. Rumination is a central characteristic of depression as individuals spiral into endless questions, asking themselves: why do I feel this way, what is wrong with me, why can't I function normally; all with the hope of finding some resolution. In running these endless scripts, the driven-doing mode exacerbates negative moods caused by depression by magnifying perceived inadequacies, reinforcing them, but then leaving the individual with no answers or solutions. Depression and the driven-doing mode, manifested in rumination, take the victim captive in these counterproductive inner monologues filled with expectations of how one 'should' be or how they were in the past in contrast to how they currently perceive themselves.

In developing a maintenance program for depression remission, Segal et al. (1995/2013) recognized that teaching participants how to identify the automatic nature of thoughts, and the cascade of negative emotions that follow, would be helpful in preventing them from allowing these negative thoughts to run their course to eventual relapse. Depressive relapse occurs among 75–90% of patients, even after successful treatment (Kessler et al., 2005). After recovering from a major depressive episode, patients are left vulnerable to relapse because whenever negative thoughts and experiences occur, they can quickly reactivate depressive thinking patterns that become more and more hardwired with each episode. As such, it is very difficult to overcome ruminative habits, and individuals who revert to dysfunctional thinking patterns show increased cognitive reactivity to sadness provocation and are at the highest risk of relapse (Fresco et al., 2007). As a result, cognitive reactivity is a central target of mindfulness-based cognitive therapy (MBCT) (Segal et al., 1995/2013). MBCT, adapted from the earlier work of Jon Kabat-Zinn, was designed to teach participants to recognize negative and automatic thought patterns and has been shown to reduce responses that would typically agitate vulnerable individuals (Kuyken et al., 2010).

Research has found that MBCT is effective in reducing relapse (Teasdale et al., 1995; Bondolfi et al., 2010), cognitive reactivity (Kuyken et al., 2010), rumination (Williams, 2008), and worry (Evans et al., 2008) in depressed patients. MBCT changes the way in which participants relate to their thoughts by bringing their attention to the present while reserving judgment and reaction to the thoughts that naturally arise in the mind. In shifting to the 'being' mode, participants can in some way suspend negative thoughts in awareness before deciding how and whether to respond to them. Mindfulness interventions encourage decentering and hope to alter the way in which thoughts are perceived; instead of trying to manipulate or change negative thoughts and habits in participants who suffer from depression, mindfulness practices teach participants to attend to their thoughts and feelings promoting an understanding that such thought patterns are natural occurrences, not facts that reflect who they are. "By observing the activity of the mind, in the moment, in a spirit of inquiry rather than judgment, and by repeatedly decentering from old mental routines, questioning and testing them, patients learn to experience thoughts as events in the mind, rather than 'the truth' or 'me'" (p. 140) (Fennell and Segal, 2011). This releases

patients from the captivity of their thoughts because they can separate themselves from the judgments that routinely occur allowing them to evaluate their validity or reset their perspective. When a thought or idea enters the mind, participants are encouraged to nonjudgmentally observe or label those percepts as simply a 'thought,' 'emotion,' etc., before bringing their attention back to the breath. This practice allows individuals to shift their thinking from a self-oriented judgmental perspective to more broadly evaluate mental processes. Instead of thinking 'I am a terrible person,' they come to recognize that instance as merely a thought, not a true reflection of themselves.

Another mechanism of the preventative success of MBCT could be due to an increase in positive affect during remission (Wichers et al., 2010). Importantly, moment-to-moment positive affect and pleasure in daily activities has been shown to increase resiliency – an individual's ability to respond positively to life events. To assess this, a group of participants who had experienced at least one prior episode of major depression completed experience sampling post-MBCT intervention over a 6-day period (Geschwind et al., 2011). In experience sampling, participants are alerted at random times throughout the day and week to report on their current mood in their daily environment. Results showed that participants experienced boosted positive affect, well-being, and reward compared to a wait-listed control group, even when reductions in depressive symptoms were controlled for. It has been proposed that the shift in outlook of simple day-to-day interactions, perhaps walking to the bus, getting a coffee, or doing the dishes, may conjure up a more connected state and greater well-being (Kozasa et al., 2012; Bondolfi et al., 2010).

Mindfulness and Physical Health

Mindfulness-based stress reduction (MBSR) was first developed by Dr Jon Kabat-Zinn at the Stress Reduction Clinic at the University of Massachusetts Medical School (now at The Center for Mindfulness) in 1979. Before the application of mindfulness interventions to psychological disorders, MBSR was used to treat patients that suffered from chronic pain, and has since been applied to other physiological conditions ranging from fibromyalgia, to headache, to lower back pain (e.g., Morone et al., 2008). All of these conditions are extremely burdensome to patients, families, support systems, and employers because they are relentless and difficult to alleviate, often leaving the patient frustrated, depressed, and unemployed. When participants are enrolled in an 8-week mindfulness course, however, chronic pain has been found to significantly decrease. Kabat-Zinn (1982) proposed that this occurs by reducing participants' tendencies to catastrophize. Catastrophizing occurs when individuals negatively evaluate their situation, placing a great deal of weight on their circumstances, allowing them to balloon into something hopeless and unmanageable, as if blocking all light from the sun. Catastrophizing has been found to account for a large percentage of variance in participants' reports of pain severity as individuals who magnify and focus on their dire state experience more pain than individuals who are naturally more present focused (Schütze et al., 2010). The act of catastrophizing requires that the individual assess, judge, and expand on the negative aspects of their experience, all of which mindfulness addresses in opposition by helping participants recognize these patterns and change their natural response, highlighting its effectiveness in this population.

In Kabat-Zinn's original work (1982), he discusses that pain is an often-cited topic in meditative writings as practitioners are required to sit for extended periods of time. This can place a great deal of strain on the limbs of the body. When pain sensations occur during a sitting practice, meditators are encouraged to focus on the sensation fully and nonjudgmentally, resisting the urge to escape the ache. They are instructed to explore and accept such sensations and to appreciate them as transient and nonthreatening. In teaching this technique to chronic pain patients, this can alter their relationship with pain; by attending to sensations of pain nonjudgmentally – by acknowledging these events, but not avoiding the sensation – participants begin to recognize that the course of pain is not destructive and all-encompassing. In doing so, they become desensitized and less emotionally reactive to their experience of agony, whether or not there is any real reduction in the pain stimulus.

While this shift in attentional focus may contribute to the reduction of pain reported by chronic pain patients, acceptance-based interventions, like MBIs, have also been shown to increase pain tolerance. Indeed, when pain tolerance was measured in experienced meditators compared to non-mediators, the meditators required significantly higher heat levels in order to report the same degree of pain as non-meditators (Grant and Rainville, 2009). Even more, when a painful stimulus was applied to novice participants who completed only 20 min of mindfulness meditation for three consecutive days prior to testing, pain sensitivity also decreased compared to those who did not engage in meditation training (Zeidan et al., 2010). The authors found this short intervention to have similar effects on pain tolerance as doing simple math equations, which they suggest highlights that the focus and attention required during mindfulness meditation may play a role in its analgesic properties. In meditative practices, focusing on the breath or shifting attention could serve as a mental distraction, diverting thoughts away from negative stimuli and preventing catastrophizing of those sensations. With accumulated practice, these effects may be magnified as one's ability to cultivate a significantly heightened state of awareness improves.

Mindfulness-Based Interventions: A Unique Approach to Health

Mindfulness-based therapies bridge a very wide expanse between eastern and western medicine, bringing together techniques from ancient Buddhist philosophy and current clinical science. Over the past three decades, MBIs, like MBSR and MBCT, have garnered attention for treating symptoms of poor health. The programs are typically taught in group classes over an 8-week period with one class each week, a day of silence in week six, home practice 45 each day, and follow-up over 6 months postintervention. During MBIs, participants are asked to maintain their attention by, for example, focusing on their breath or conducting body scans. When participants

notice that their attention has diverted to other thoughts, they are instructed to recognize those diversions nonjudgmentally and return their attention back to their breath or body. This technique is taught as a tool that can be applied to daily life. When one realizes the occurrence of negative thoughts, the individual can acknowledge those feelings and 'let go' of their attachment and resulting emotional arousal in order to refocus attention on the present and respond differently. Mindfulness aims to cultivate nonjudgmental attitudes about the past, present, or future, to reduce automatic and potentially negative rumination about one's experiences and maintain openness toward one's thoughts. Both MBSR and MBCT are efficacious treatment programs that have been found to reduce recurrence of depression, stress, improve concentration and productivity, as well as alleviate chronic pain, anxiety, and disordered eating (e.g., Keng et al., 2011; Hofmann et al., 2010).

One of the key elements of MBIs is that unlike more standard methods of health care, these courses take a non-goal-oriented approach to treatment outcome. This can seem at odds with the current approach taken by medical professionals today. In Western medicine, we have come to think of treatment as having some end goal of health, cure, or symptom relief. As Fennell and Segal (2011) discuss, "[there appears to be] something of a paradox here. In the field of healthcare at least, people undoubtedly come to classes because they are suffering and hope for relief—and presumably therapists/instructors would not provide classes unless they believed this was possible" (p. 133). However, MBIs discourage instructors from trying to directly 'fix things' and participants from attaching to their end goals or hopes in taking the course. While MBIs do hope to alleviate suffering, they do not emphasize or invite this in as part of the process. With the aim of remaining open to experience, instructors emphasize that there is no correct way to do the practices and no ultimate state to achieve. Participants are reminded to be open to the process, and not to focus on the outcome, cling to, or hold expectations of themselves, but to just let things be as they are.

In clinical psychology, this is in sharp contrast to cognitive behavioral therapy, which focuses on eliminating suffering by asking participants to actively define their goals and the steps needed to attain them. In the case of MBIs, however, while relief from suffering underlies the intervention, the course promotes a light-hearted, open, and unfiltered approach to participants' experiences, with an emphasis to leave expectations at the door. The constant assessment of goals and attempts to 'succeed' can become problematic to the process by activating the nasty driven-doing mode of thinking: "am I doing this correctly?," "this isn't doing anything," "am I feeling better?," "Why can't I stop mind wandering?" Here, the participant starts to analyze his progress and gives in to the driven-doing mode. However, by identifying and disengaging from those thoughts when they arise, he can instead return to the present and permit the practice to run its course granting him the opportunity to explore experiences further and reinforce present state awareness.

The group setting of MBIs encourages open explorations of mindfulness practices because there are organic discussions during each class. The group framework is another important feature of MBIs for two reasons. First, it can be offered to more individuals than one-on-one therapy, with significant impact over a short period of time. Second, it encourages participants to recognize that many of the difficulties they face both generally and specifically, with regard to the course, are not unique to them but are shared by both their classmates and all humans to some extent. Interaction with the instructor during weekly meetings allows participants to listen and relate to the struggles that other group members experience. Through this dialogue, they recognize the tendencies to want to assess their progress and are reminded how to shift their state of inquiry to an open and curious one.

Mindfulness Practices and Stress Management

Stress is a part of each and every day, existing as dynamic tension in life. Without it, we would not adapt to perturbations in our environment. Stressors are environmental stimuli that force the body into reaction, or out of homeostasis. There is no zero level of stress, nor should there be, as stress is a force of existence much like gravity. Without stress, we would not learn, adapt, and grow. There is, however, a balance to be found as too much stress, or *distress*, is detrimental. Given the inevitability of stress, there are two important approaches to stress: stress reduction and stress management, in both of which mindfulness can play an important role.

Distress, or chronic stress, can cause serious health problems by overloading the sympathetic and neuroendocrine system, wearing down the immune system, increasing inflammation and oxidative damage, leaving individuals very vulnerable both emotionally and physically. While removing any sources of stress is undoubtedly beneficial to health, mindfulness interventions work to reduce and manage the stress that is unavoidable by both helping to identify stressors that are often undetected, and also by changing the way in which individuals respond and relate to stress. As discussed above, there is great value in being able to spot automatic cascades of thoughts and reactions. Many stressors are self-imposed by the doing mode of processing and are not necessarily immediately apparent, and need not be overtly catastrophic in nature to nevertheless cause burden. Mindfulness highlights moments that might subconsciously cause grief, if left unattended. When an individual starts thinking about negative interactions at work, or ends up getting distracted by e-mail or the Internet, it can leave him or her open to potentially negative moods or detract from the time spent on more valuable aims and goals (Brown and Ryan, 2003). These kinds of moments add up over time, thus identifying them is valuable to reducing tension and opening up new ways of responding. The stressor is made visible and, in being so, it can be prevented from causing additional burden by letting go or directly addressing the issue, allowing stress to return to its adaptive role.

There is an important distinction to be made between the stressors themselves and the appraisal of said stressors. The role that perceived stress plays in the stress response can be as stressful (and sometimes more stressful) than the actual threat itself. Mindfulness has the potential to alter the way in which both stress and the stress response are experienced. In a review, Thompson et al. (2011) found that mindfulness is a strong predictor of resiliency following exposure to trauma. Participants

with higher levels of mindfulness exhibited fewer disruptive symptoms following trauma while those who developed post-traumatic stress disorder (PTSD) were more likely to use avoidant coping strategies and showed higher levels of dissociative experiences, common of PTSD. Similarly, in a four-part study, student stress was measured via stress tests, autobiographical stress recall, experience sampling in daily life, and over the course of a semester leading up to exams. Weinstein et al. (2009) found that mindful students were less susceptible to stressful experiences, reporting more muted stress appraisals and showing better performance. They also used approach coping skills to deal with stressful situations whereas those low in mindfulness tended to utilize avoidant coping strategies.

Individuals who score higher on measures of mindfulness may have higher tolerance to distress because they are more likely to see the stressor at face value (not to elaborate and catastrophize), are less overwhelmed by their own thoughts, and are able to identify and respond to stress adaptively (Feldman et al., 2013). With the growing understanding of the absolute detrimental impacts of stress, the examination of practices and behaviors that can mitigate the levels of experienced stress are essential. The attentional shift that comes with dedicated practice, like any other acquired skill, can be reinforced over time, and it is becoming evident that mindfulness is a skill worthy of our scientific attention.

See also: Behavioral Medicine; Chronic Pain: Models and Treatment Approaches; Co-morbidity of Mental and Physical Conditions; Depression, Pessimism, and Health; Mental Health; Religion and Health: Clinical Considerations and Applications; Respiratory Disorders: Psychosocial Aspects; Stress, Coping and Health.

Bibliography

Baer, R.A., Smith, G.T., Hopkins, J., Krietemeyer, J., Toney, L., 2006. Using self-report assessment methods to explore facets of mindfulness. Assessment 13 (1), 27–45.

Benzo, R., 2013. Mindfulness and motivational interviewing: two candidate methods for promoting self-management. Chronic Respiratory Disease 20 (3), 175–182.

Bondolfi, G., Jermann, F., Van der Linden, M., Gex-Fabry, M., Bizzini, L., Weber Rouget, B., Myers-Arrazola, L., Gonzalez, C., Segal, Z., Aubry, J.M., Bertschy, G., 2010. Depression relapse prophylaxis with mindfulness-based cognitive therapy: replication and extension in the Swiss health care system. Journal of Affective Disorders 122 (3), 224–231.

Brown, K.W., Ryan, R.M., 2003. The benefits of being present: the role of mindfulness in psychological well-being. Journal of Personality and Social Psychology 84 (4), 822–848.

Chiesa, A., Paolo, B., Serretti, A., 2010. Functional neural correlates of mindfulness meditations in comparison with psychotherapy, pharmacotherapy and placebo effect. Is there a link? Acta Neuropsychiatrica 22 (3), 104–117.

Chiesa, A., Serretti, A., 2009. Mindfulness-based stress reduction for stress management in healthy people: a review and meta-analysis. The Journal of Alternative and Complementary Medicine 15 (5), 593–600.

Dobkin, P., Hutchinson, T., 2013. Teaching mindfulness in medical school: where are we now and where are we going? Medical Education 47 (8), 768–779.

Evans, S., Ferrando, S., Findler, M., Stowell, C., Smart, C., Haglin, D., 2008. Mindfulness-based cognitive therapy for generalized anxiety disorder. Journal of Anxiety Disorders 22 (4), 716–721.

Farb, N.A.S., Segal, Z.V., Mayberg, H., Bean, J., McKeon, D., Fatima, Z., Anderson, A.K., 2007. Attending to the present: mindfulness meditation reveals distinct neural modes of self-reference. Social Cognitive and Affective Neuroscience 2 (4), 313–322.

Feldman, G., Dunn, E., Stemke, C., Bell, K., Greeson, J., 2013. Mindfulness and rumination as predictors of persistence with a distress tolerance task. Personality and Individual Differences 56.

Fennell, M., Segal, Z., 2011. Mindfulness-based cognitive therapy: culture clash or creative fusion? Contemporary Buddhism: An Interdisciplinary Journal 12 (1), 125–142.

Flook, L., Goldberg, S.B., Pinger, L., Bonus, K., Davidson, R.J., 2013. Mindfulness for teachers: a pilot study to assess effects on stress, burnout, and teaching efficacy. Mind, Brain, and Education 7 (3), 182–195.

Fresco, D.M., Segal, Z.V., Buis, T., Kennedy, S., 2007. Relationship of posttreatment decentering and cognitive reactivity to relapse in major depression. Journal of Consulting and Clinical Psychology 75 (3), 447–455.

Giluk, T., 2009. Mindfulness, Big Five personality, and affect: a meta-analysis. Personality and Individual Differences 47, 805–811.

Grant, J.A., Rainville, P., 2009. Pain sensitivity and analgesic effects of mindful states in Zen meditators: a cross-sectional study. Psychosomatic Medicine 71 (1), 106–114.

Geschwind, N., Peeters, F., Drukker, M., van Os, J., Wichers, M., 2011. Mindfulness training increases momentary positive emotions and reward experience in adults vulnerable to depression: a randomized controlled trial. Journal of Consulting and Clinical Psychology 79 (5), 618–628.

Grossman, P., Niemann, L., Schmidt, S., Walach, H., 2004. Mindfulness-based stress reduction and health benefits: a meta-analysis. Journal of Psychosomatic Research 57 (1), 35–43.

Hofmann, S.G., Sawyer, A.T., Witt, A.A., Oh, D., 2010. The effect of mindfulness-based therapy on anxiety and depression: a meta-analytic review. Journal of Consulting and Clinical Psychology 78 (2), 169–183.

Kabat-Zinn, J., 1982. An outpatient program in behavioral medicine for chronic pain patients based on the practice of mindfulness meditation: theoretical considerations and preliminary results. General Hospital Psychiatry 4 (1), 33–47.

Kabat-Zinn, J., 2003/2013. Full Catastrophe Living: Using the Wisdom of Your Body and Mind to Face Stress, Pain, and Illness. Dell Publishing, New York, NY.

Keng, S., Smoski, M.J., Robins, C.J., 2011. Effects of mindfulness on psychological health: a review of empirical studies. Clinical Psychology Review 31 (6), 1041–1056.

Kessler, R.C., Berglund, P., Demler, O., Jin, R., Merikangas, K.R., Walters, E.E., 2005. Lifetime prevalence and age-of-onset distributions of *DSM-IV* disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry 62, 593–602.

Kozasa, E.H., Sato, J.R., Lacerda, S.S., Barreiros, M., Radvany, J., Russell, T.A., Sanches, L.G., Mello, L., Amaro Jr., E., 2012. Meditation training increases brain efficiency in an attention task. NeuroImage 59 (1), 745–749.

Kuyken, W., Watkins, E., Holden, E., White, K., Taylor, R., Byford, S., Evans, A., Sholto, R., Teasdale, J.D., Dalgleish, T., 2010. How does mindfulness-based cognitive therapy work? Behaviour Research and Therapy 48 (11), 1105–1112.

Morone, N.E., Greco, C.M., Weiner, D.K., 2008. Mindfulness meditation for the treatment of chronic low back pain in older adults: a randomized controlled pilot study. PAIN 134 (3), 310–319.

Schütze, R., Rees, C., Preece, M., Schütze, M., 2010. Low mindfulness predicts pain catastrophizing in a fear-avoidance model of chronic pain. PAIN 148 (1), 120–127.

Segal, Z., Williams, M.G., Teasdale, J.D., 1995/2013. Mindfulness-based Cognitive Therapy for Depression, second ed. The Guilford Press, New York (New York).

Teasdale, J.D., Segal, Z., Williams, J.M.G., 1995. How does cognitive therapy prevent depressive relapse and why should attentional control (mindfulness) training help? Behaviour Research and Therapy 33, 25–39. http://dx.doi.org/10.1016/0005-7967(94)E0011-7.

Thompson, R.W., Arnkoff, D.B., Glass, C.R., 2011. Conceptualizing mindfulness and acceptance as components of psychological resilience to trauma. Trauma, Violence & Abuse 12 (4), 220–235.

Weare, K., 2013. Developing mindfulness with children and young people: a review of the evidence and policy context. Journal of Children's Services 8 (2), 141–153.

Weinstein, N., Brown, K.W., Ryan, R., 2009. A multi-method examination of the effects of mindfulness on stress attribution, coping, and emotional well-being. Journal of Research in Personality 43, 374–385.

Wichers, M., Peeters, F., Geschwind, N., Jacobs, N., Simons, C.J.P., Derom, C., Thiery, E., Delespaul, P.H., van Os, J., 2010. Unveiling patterns of affective responses in daily life may improve outcome prediction in depression: a momentary assessment study. Journal of Affective Disorders 124 (1–2), 191–195.

Williams, J.G., 2008. Mindfulness, depression, and modes of mind. Cognitive Theory Research 32, 721–733.

Zeidan, F., Gordon, N.S., Merchant, J., Goolkasian, P., 2010. The effects of brief mindfulness meditation training on experimentally induced pain. Journal of Pain 11 (3), 199–209.

Mineral Resources, Limits to: The Case of Peak Oil

Ugo Bardi, Università di Firenze, Firenze, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the concept of 'peak oil' showing that it is a consequence of a combination of geological and economic factors that are normally present when exploiting a finite resource under a free (or nearly free) market conditions. The concept of peak oil is then analyzed in view of the reactions to the concept during the more than half a century from its introduction. There have been several similar cases in the past where the attitude of the public and the operation of industrial lobbies has led to the demonization of a number of studies and of the scientists who performed them. In the case of peak oil, this negative reaction has not been so strong as in other cases, but still the general reaction shows several similarities with examples such as, for instance, the 'Limits to Growth' study of 1972 and climate science today.

Introduction

The term 'peak oil' refers to the idea that the world's production of crude oil is expected to peak and then decline, typically when about half of the available resources have been extracted. The term was proposed for the first time in 1999 by Colin Campbell in the first issue of the newsletter of the Association for the Study of Peak Oil (ASPO) (Aleklett, 2009). Campbell, together with his colleague Jean Laherrere, had proposed earlier on (Campbell and Laherrere, 1998) that the world's production of crude oil world peak and decline approximately in 2005, on the basis of a previous concept suggested by Marion King Hubbert (1956).

The concept of 'peaking' and 'peak oil' rapidly gained recognition with the start of the second oil crisis in the first decade of the twenty-first century and it became an icon of the movement of ideas that sees the depletion of nonrenewable mineral resources as a major problem for humankind's future. In time, the peak oil idea generated a number of similar concepts, such as 'Peak Coal' (Patzek and Croft, 2010; Zittel et al., 2013), 'Peak Uranium' (Dittmar, 2013; Zittel et al., 2013), 'Peak Minerals' (Bardi and Pagani, 2007), 'Peak Caviar' (Bardi, 2008), and even 'Peak Everything' (Heinberg, 2007). All these concepts are based on the same interpretation of the trend of exploitation of a resource that exists in limited amounts or that, anyway, is exploited much faster than it can regenerate itself.

'Peak Oil' is the best known of these peaks and it is a classic example of a message related to the finiteness of natural resources. As it is normally understood by the public, it deals with something relatively easy to understand: given that the amount of extractable hydrocarbons in the world is limited, there should follow as a simple inference that, at some moment in the future, production will go to zero. As production started at zero in the past, there has to be a maximum somewhere in between and we can call it a peak.

In this sense, the issue of peaking differs from other issues also related to the finiteness of our environment, which are often more difficult to explain to the general public. For instance, climate change can be seen as related to the limited capability of the Earth's atmosphere to absorb greenhouse gases without heating up at levels dangerous for the survival of human beings. But most people perceive the problem of climate change as complex, abstract, and remote from their daily experience and hence not an immediate threat.

Nevertheless, the concept of peak oil is widely misunderstood and the debate on it in the media goes through the usual oversimplifications. The concept of 'peak' is often confused with the actual 'running out' of a resource. The exact date of the peak is given unduly importance; neglecting the fact that it is gradual depletion that counts and, in particular, the rising costs of exploiting the resource. Also, as it often happens with negative messages, those who propose the idea of peak oil are the object of personal attacks and are often branded as catastrophists or Cassandras. The result is that the message arrives garbled and confused to the public and to policy makers alike.

In this article, I will provide an overview of the concept of peak oil that can be used to understand the more general concept of resource depletion. In addition, I will provide a discussion on how people have reacted so far to the concept of peak oil, which turns out to have been similar to the reaction to other 'negative' views of the future, such as global warming, environmental poisoning, and the like.

A Brief History of the Concept of Resource Depletion

It was only in mid-nineteenth century that the first analysis of the consequences of the limited amounts of mineral resources available was published. It was the economist William Stanley Jevons who produced it in his *The Coal Question; An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal Mines* (1865) (Jevons, 1866).

Jevons' study was grounded in a well-known concept in economics, that of the 'diminishing returns on investments.' In order to extract coal, one must invest resources in the form of equipment, machinery, work, and more. In Jevons' times, most of these resources were produced using coal and, hence, one could say that coal needs to be invested to obtain coal. So, extraction as an economic activity is subjected to the conventional concept of return on investment (ROI). In modern terms, the same concept can be seen in energy terms, that is as 'energy return on energy investment' (EROI or EROEI) (Hall and Day, 1990).

The basis of Jevons' interpretation lies in the fact that, for all investments, the high return ones are those made first. That is,

the coal deposits closest to the surface and richest in coal were those which provided the highest returns (and needed the least energy investments) and hence were extracted first. With time, however, extraction had to move to less satisfactory resources, that is to thinner and deeper coal seams. That required larger investments and it lowered the returns. Hence, Jevons proposed that, in time, coal extraction would become less and less convenient in terms of monetary returns and also in terms of resources to be invested. At some moment, coal production would cease to grow and then enter in a phase of permanent decline. That would occur well before the actual disappearance of the extractable coal deposits in Britain and, according to Jevons, 'within a century from the present time' (Chapter XII of *The Coal Question*). Jevons predictions turned out to be qualitatively correct, as shown by the historical data (**Figure 1**).

After Jevons' work, the interest in understanding mineral depletion lied dormant for several decades and it was only in the twentieth century that new theoretical studies were performed on the subject. In the 1930s, Harold Hotelling introduced the so-called 'Hotelling's rule' (Hotelling, 1931), proposing that the prices of an exhaustible resource would keep growing with the progress of depletion and that production would decline from the beginning of the extraction cycle. This result made the model even more 'doomerish' than Jevons' one. A more optimistic theory was developed in later times on the basis of neoclassical economics, see, for example, Solow (1974) and Stiglitz (1979). The gist of this interpretation is that the same production levels of any commodity can be obtained by different combinations of elements; for example, number of workers, equipment. It is further assumed that these elements are wholly interchangeable with each other. There follows that the production of a mineral commodity can continue growing even for a diminishing availability resulting from depletion, provided that technological progress increases more rapidly with time; as it was proposed to be the general case (Nordhaus, 1992).

Even more optimistic models, albeit qualitative ones, were also developed. The so-called functional model of resource extraction (Zimmermann, 1933) emphasizes the fact that higher prices make it possible to extract more expensive resources, thus increasing the resource base. This concept is often expressed as 'the market creates resources.' According to this view, high prices do not depress demand because technological progress always makes it possible to reduce extraction costs and, in any case, a growing economy can maintain a healthy demand even for high prices. This interpretation seems to be rather popular in many circles but it may lead to some rather overoptimistic interpretations, such as when the economist Julian Simon stated (Simon, 1981) that the world's mineral resources would last 'six billion years' or anyway for times so long that there would be nothing to fear even for a remote future. Simon's interpretation originated, apparently, from the examination of the price trends of just five mineral commodities over a few decades.

The problem with most of these models of depletion is that they are either abstract and scarcely connected with the real world or qualitative and based on unverifiable assumptions. For instance, the neoclassical model was criticized by Herman Daly (1997) as:

> Since the production function is often explained as a technical recipe, we might say that Solow's recipe calls for making a cake with only the cook and his kitchen. We do not need flour, eggs, sugar, etc., nor electricity or natural gas, nor even firewood. If we want a bigger cake, the cook simply stirs faster in a bigger bowl and cooks the empty bowl in a bigger oven that somehow heats itself. Nor does the cook have any cleaning up to do, because the production recipe produces no wastes.

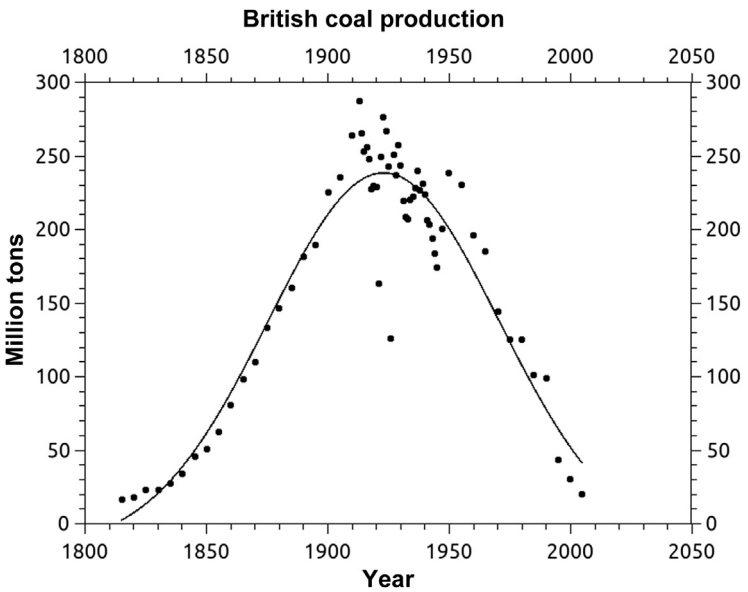

Figure 1 Historical data for British coal production. The data from 1815 to 1860 are from Cook, C., Stevenson, J., 1996. The Longman Handbook of Modern British History, 1714–1995. Longman, London and New York., the data from 1860 to 1946 are from Kirby, M.W., 1977. The British Coal-mining Industry, 1879–1946. The Macmillan press Ltd, London and Basingstoke., and the remaining data are courtesy of the British Coal Authority (http://coal.decc.gov.uk/). The data have been fitted with a Gaussian function that approximates the Hubbert curve (see text for a description).

Probably as an effect of the limitations of these models, the extractive industry normally maintains a 'down to earth' approach to forecasting; mainly focused on the determination of the extractable amounts on the basis of geological data which are then projected into the future on the basis of the concept of 'resource to production' (R/P) ratio. This ratio returns the number of years of production still available at constant production rate and it can not tell much about long-range trends.

However, in the 1950s, Marion King Hubbert went beyond this low key approach and developed a long-range predictive model based on resource estimates. The model, published for the first time in 1956 (Hubbert, 1956), was based on the idea that the oil extraction cycle was to follow a symmetric 'bell-shaped, curve. That is, it would grow up to a certain point (the 'Hubbert peak') and then gradually decline on the other side of the curve. The model for the US 48 lower states is shown in Figure 2.

Today, it is possible to examine historical cases of regional oil production in light of the Hubbert model. There are many cases in which the model does describe the historical oil production data and can be used for forecasting. Its most notable success is for oil production in the United States up to just a few years ago. But the model works not just of oil; it can also describe other nonrenewable mineral resources (Bardi and Pagani, 2007) and even the production trends of renewable resources when they are exploited at rates much faster than their replenishing rate, for example, whale oil (Bardi, 2007). However, in many cases, the results of the Hubbert model turned out to be approximate or even completely unable to describe historical data (Brandt, 2007). Production in countries such as Saudi Arabia never showed a well-defined bell-shaped curve, while in other cases the data could only be fitted by multiple curves; such as for the Soviet Union and, probably, for the United States in recent years.

Hubbert's ideas were reexamined in 1998 by Jean Laherrere and Colin Campbell (Campbell and Laherrere, 1998) who used the model to analyze the pattern of the world oil production. Their prediction was that a world peak production for conventional crude oil (peak oil) would occur approximately in 2005. This date did correspond to a considerable slowdown of the precedent growth trends and a rapid rise in oil prices. However, it is also true that at present (2014) a well-defined productive peak is not yet visible. One of the problems lies in the fact that Campbell and Laherrere based their forecast on 'conventional' oil, whereas in recent years the industry has been able to put into production various sources of 'non conventional' oil, including biofuels, liquids from tar sands, and others. The addition of these resources has permitted to keep increasing the production of combustible liquids, albeit at a slower rate than in the past. The following figure (Figure 3) shows the updated forecast for world oil production obtained using the same methods that were used for the 1998 study (Campbell and Laherrere, 1998)

The Systemic View of Peak Oil

In time, several attempts were performed to understand the theoretical basis of the Hubbert model. Here it would be out of place to review all these studies; it will suffice to say that the Hubbert ideas have been examined by means of standard economics models (see, e.g., Holland, 2013), on statistics interpretations (see, e.g., Guseo, 2011), and by 'bottom up' models based on the extraction dynamics of wells (Bentley, 2002). These attempts have had various degrees of success, but for the present text it may be more interesting to focus into an interpretation that can provide an intuitive grasp of the elements that create the bell-shaped curve, in other words, a 'mind sized' model (Bardi, 2013). Such a view of the model can be based on the concept called 'system dynamics'.

System dynamics originated in the 1960s, mainly with the work of Jay Forrester (1995). Using the newly developed digital computers of the time, Forrester put together a complete method for the study of the evolution of complex systems. The method was based on the simultaneous solving of a set of differential equations that described how the various parameters of the system interacted with each other as a function of time. These systems were often studied in terms of long-term trends of social and economic systems and, in this application, system dynamics was never to be intended as a forecasting method. Rather, it was to be seen as a method for

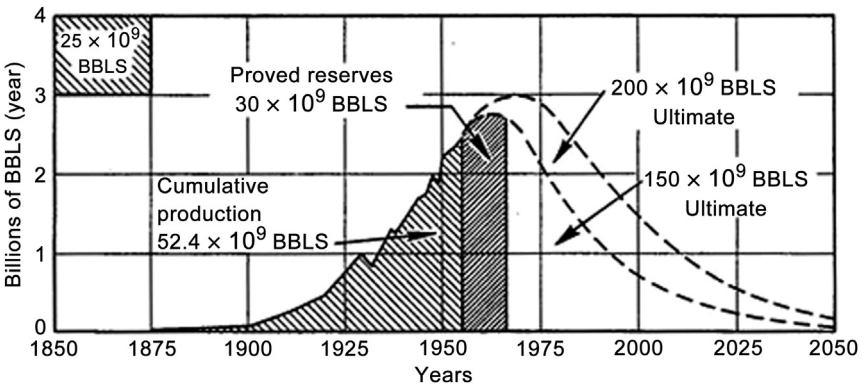

Figure 2 Hubbert's original graph illustrating the proposed production curve for crude oil in the 48 US lower states. Reproduced from Hubbert, M.K., 1956. Nuclear energy and the fossil fuels. In: Spring Meeting of the Southern District Division of Production American Petroleum Institute. American Petroleum Institute, Sant Antonio, TX. Available at: energycrisis.biz/hubbert/1956/1956.pdf (accessed 03.12.12.).

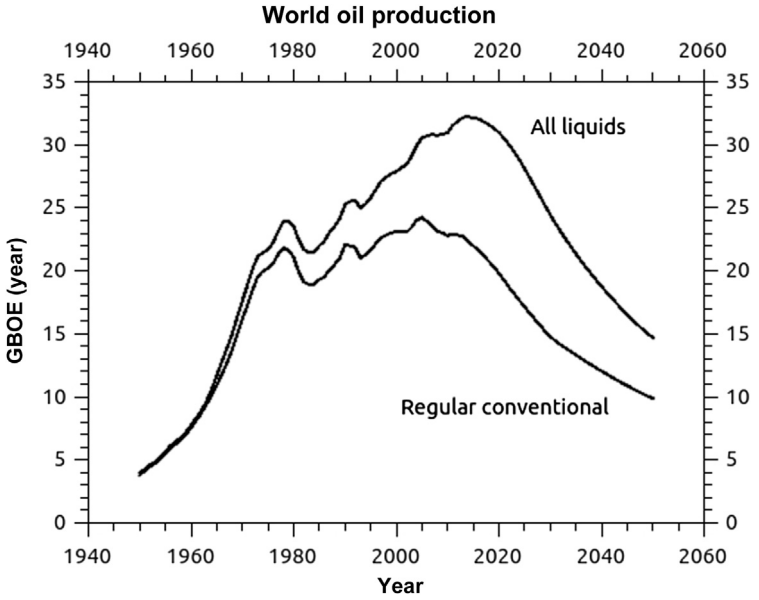

Figure 3 Past world conventional crude oil and 'all liquids' production, together with an extrapolation for the future based on the Hubbert model. The peak for the production of regular oil may have taken place in 2005, but the total peak for all possible forms of combustible liquids will not occur until *c.* 2015. Data courtesy of Colin Campbell.

amplifying the power of the human mind in understanding the behavior of complex systems and how to modify it (Meadows, 2008).

System dynamics was soon utilized to describe the behavior of the whole world's economic system, an approach often termed 'world modeling.' The first study in this field was performed by Jay Forrester (1971), soon followed by the 1972 study titled *The Limits to Growth* (Meadows et al., 1972) (LTG) which rapidly became a major subject of debate worldwide (Bardi, 2011b). The LTG model did not consider crude oil as a disaggregated resource, but its general results were consistent with the earlier model proposed by Hubbert. That is, system dynamics predicted that production would grow at a near exponential rate for a while, then slow down, reach a peak, and finally decline. The shape of the LTG production curves, however, was typically asymmetric; with decline faster than growth. The reasons for this behavior are described by Ugo Bardi (2013) in terms of a property of these systems termed the 'Seneca Effect' that results from parameters additional to depletion affecting the production trends.

Seen in these terms, the concept of peak oil is best understood in terms of system dynamics. A simple interpretation in this sense has been presented by (Bardi and Lavacchi, 2009). In short, the bell-shaped curve is generated by increasing extraction cost which lower profits. As a result, the industry finds itself short of the necessary capital to invest in more and more expensive resources. With diminishing investments, the production rate peaks and declines. This interpretation is equivalent to the one proposed by Jevons (1866) with the difference that when these assumptions are set in the form of equations in the model, they do generate the bell-shaped curve typical of Hubbert's model. This interpretation of the model can also be used to understand why the Hubbert model may not work for some regional cases. It is because the economic and financial systems are able to move resources from one sector of the economy to another. Increasing investments have permitted, for instance, to keep gas production from decreasing in the United States after having peaked in 1973 (Bardi, 2011a). In recent years, large investments on the extraction of 'tight gas' have allowed the US industry to produce at levels exceeding those of the peak.

Perception and Reactions to Peak Oil and Other Depletion Claims

The diffusion of new concepts in science tends to follow well-established patterns. Basically, ideas that challenge the establish paradigms tend to undergo thorough (and sometimes aggressive) scrutiny on the part of the scientific community. Surviving this scrutiny may be difficult for a new theory which, often, needs time and more data to be refined and upgraded in such a way to become acceptable. A classic case in point is the reaction to Alfred Wegener's theory of 'continental drift' that was proposed for the first time in 1912 (Wegener, 1912). As reported in detail, for instance by Oreskes (1999), Wegener's theory went through several stages: it attracted wide attention at the beginning, but it was then strongly criticized and rejected. Different factors played a role in the rejection of the theory, including political ones (Wegener was German and German science was not popular for a long time in the periods after the two world wars). But the rejection was not unjustified: before accepting the idea that continents were moving, some serious evidence was needed for their motion and for the causes of it. This evidence was not available in Wegener's times but when appropriate data became available, in the 1970s, Wegener's theory became the cornerstone of modern Earth sciences, today renamed as 'plate tectonics.'

So, the acceptance of new scientific ideas seems to follow some established gradual patterns and Bardi has argued that a similar path should be expected for the peak oil theory (Bardi, 2009). However, the situation may be more complex than that. In modern times, the negative reaction to new and upsetting paradigms can be enhanced by public relation campaigns specifically aimed at discrediting the research and the researchers involved. There are several examples of such campaigns in recent history, some of which have been highly successful. Here is a list of some of the best known and relevant cases.

1. *'Silent Spring' and the pesticide industry*. In 1962, Rachel Carson published a book titled *Silent Spring* (Carson, 2002) in which she strongly criticized the abuse of pesticides. The reaction of the chemical industry is described by (Groshong, 2007) reporting the text of a document Manufacturing Chemists' Association, Inc. that discusses "measures which should be taken to bring the matter back to proper perspective in the eyes of the public." This is one of the first documented cases of a planned PR campaign by an industry against a scientist and of her results.
2. *The tobacco controversy*. In the 1960s, the tobacco industry fought hard and for several years against the studies that had evidenced a link between tobacco smoking and lung cancer. The story of the PR campaign generated by the tobacco lobby is told in detail, for instance, in 'Bending Science' (McGarity and Wagner, 2008). This campaign is one of the most detailed set of data available on the way these campaigns are set up and operated.
3. *The 'Limits to Growth' controversy*. When the study titled *The Limits to Growth* (Meadows et al., 1972) was published, it encountered considerable criticism in various scientific sectors. However, as detailed in (Bardi, 2011b) it was not the scientific criticism that condemned the study to be widely considered as 'wrong' but, rather, a denigration PR campaign that had many points in common with other cases of denigration of scientists and of scientific studies. In this case, however, there is no actual evidence that the campaign was financed by specific lobbies and groups.
4. *The climate controversy*. The story of this controversy is recent and still ongoing. Data that demonstrate the occurrence of anthropogenic climate change (AGW) have been accumulating since the mid-1950s. By the 1980s, the evidence was sufficiently ample that it led to the foundation of the International Panel on Climate Change (IPCC, established in 1988) that issued its first report in 1990 (Houghton et al., 1990). It is certain that there was an organized reaction on the part of some sections of the fossil fuels industry against the concept of anthropogenic climate change. This reaction involved methods and techniques very similar to those deployed by the tobacco industry in earlier times, as described for instance in *Merchants of Doubts* (Naomi Oreskes and Conway, 2010) and *Climate Cover-up* (Hoggan, 2009).

If we examine the concept of 'peak oil' in light of these controversies, we see that it does not appear to have been the object of comparable professional negative campaigns. From the beginning, the theory was considered mainstream enough to be mentioned and discussed in scientific journals. For instance, a 1974 article on *Science* (Berg et al., 1974) extensively covers Hubbert's estimates, even though it concludes that they are excessively pessimistic. However, with the fall in oil prices in the second half of the 1980s, interest in oil depletion waned, only to be taken up again with the end of the 1990s and the new century. In this second wave of interest, the discussion in scientific journals ranged over a wide variety of issues, with the main criticism being that of the concept of 'peaking' on the basis of the uncertainty in the reserve estimates (see, e.g., Lynch, 1999, 2002).

As shown in **Figure 4**, obtained from the Jstor database (dfr.jstor.org) that collects scientific article over more than a century, the interest in Hubbert's idea seems to have ebbed and waned in correspondence to the vagaries of the oil prices.

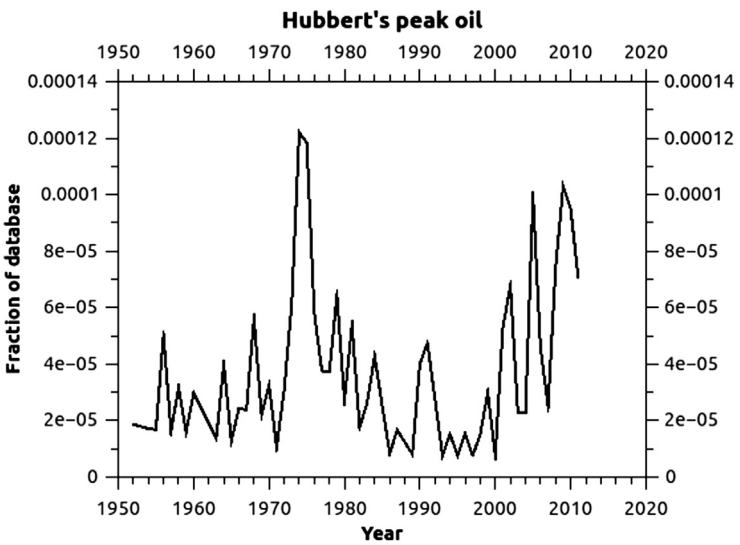

Figure 4 Fraction of citations in the Jstor database (dfr.jstor.org) that include the terms 'Hubbert,' 'peak,' and 'oil.'

We clearly see in the picture the peaking of interest in Hubbert's theory during the 'oil crisis' of the 1970–80s. The interest wanes during the period of low oil prices in the late 1980s and the 1990s, to pick up again in recent times, with the new oil crisis ongoing. Neither the data about Jevons nor those about Hubbert tell us whether the citations indicated acceptance or rejection of the concept discussed, but it is clear that if the theory was mentioned in a scientific journal, it was considered at least worth discussing.

Outside the world of standard scientific journals, the concept of peak oil was often attacked and criticized in industry reports and in the media, for instance, by Daniel Yergin (2011) and Leonardo Maugeri (2012). The reaction against the perspective of scarcity brought by peak oil also originated a series of legends often based on the idea that oil scarcity is a conspiracy on the part of the oil industry as a way to justify high market prices. This legend is often correlated with the so-called abiotic oil theory; often interpreted in terms the existence of large, even immense, oil resources located inside the Earth's mantle; something that, incidentally, makes no sense even in simple physical terms, as discussed in Höök et al. (20101).

So, even without evidence for a major PR effort against the concept of peak oil, these reactions shows the difficulty that the human mind has to deal with even moderately complex concept that are not amenable to simple, binary, 'yes/no' answers. It is a common experience in trying to relate the concept of 'peak production' to the general public or to policy makers, to face an attitude that could be defined as 'binary,' in the sense that the depletion problem is seen as an 'either/or' condition. That is, either there is oil (or some other resource) or there is not. In the first case there is no problem, in the second we have an immediate problem. This attitude reminds the statement attributed to Arthur Schlesinger "Human beings have only two modes of operation: complacency or panic." It has also been examined by Donella Meadows (1999) who cites Jay Forrester about two common modes of thought: 'dynamic' and 'linear.' According to this interpretation, linear thinking is the one that sees the world in binary terms of black/white, yes/no, good/bad, full/empty, and so on, and it lacks the nuances of dynamical thinking which instead sees the world in various shades of gray and the glass as always half full or half empty. Despite the increasing complexity of the modern world, the linear mode of thinking seems to be still largely prevalent.

Conclusion

All models are wrong, and the Hubbert model is no exception. But, all models can be useful if one understands their limits. What the Hubbert model of oil production can tell us is that the limit to the production of a mineral resource lies not with the concept of 'ultimate recoverable resource' but with the increasing costs of extraction that start manifesting themselves much before the last barrel is extracted. If left to itself, the extractive system will often follow a Hubbert-like path, but that is not obligatory; you can beat Hubbert, but you must be prepared to pay a high price for doing that. But it is also true that you cannot beat Hubbert forever by playing the extraction game harder: increasing energy costs of extraction are a physical factor that eventually forces the decline of a limited resource. The market can make extraction profitable, but it cannot make it cheap. What we *could* do to beat Hubbert forever is to leave the extraction game altogether. We would need to plan ahead and start already now to use renewable energy resources – not subjected to depletion – on a large scale. But human beings do not seem to be especially apt at long-range planning and the destiny of those who do not understand the Hubbert curve seems to be to follow it to the bitter end.

See also: Limits to Growth.

Bibliography

Aleklett, K., 2009. Colin Campbell and 100 Months of Peak Oil. ASPO International: The Association for the Study of Peak Oil and Gas. Peakoil.net. Available at: http://www.peakoil.net/headline-news/colin-campbell-and-100-months-of-peak-oil (accessed 20.06.13.).

Bardi, U., 2007. Energy prices and resource depletion: lessons from the case of whaling in the nineteenth century. Energy Sources, Part B: Economics, Planning, and Policy 2 (3), 297–304.

Bardi, U., 2013. Mind sized world models. Sustainability 5 (3), 896–911. Available at: http://www.mdpi.com/2071-1050/5/3/896 (accessed 27.05.13.).

Bardi, U., Pagani, M., 2007. Peak Minerals. The Oil Drum. Available at: http://www.theoildrum.com/node/3086 (accessed 12.06.13.).

Bardi, U., 2011a. No Peak Oil Yet? The Limits of the Hubbert Model. The Oil Drum. Available at: http://www.theoildrum.com/node/7241 (accessed 12.06.13.).

Bardi, U., 2008. Peak Caviar. The Oil Drum. Available at: http://europe.theoildrum.com/node/4367 (accessed 06.12.12.).

Bardi, U., 2009. Peak oil: the four stages of a new idea. Energy 34 (3), 323–326. Available at: http://dx.doi.org/10.1016/j.energy.2008.08.015 (accessed 15.06.13.).

Bardi, U., 2011b. The Limits to Growth Revisited. Springer. Available at: http://books.google.com/books?hl=en&lr=&id=Gj0RmQWGfXcC&pgis=1 (accessed 02.12.12.).

Bardi, U., Lavacchi, A., 2009. A simple interpretation of Hubbert's model of resource exploitation. Energies 2 (3), 646–661. Available at: http://www.mdpi.com/1996-1073/2/3/646 (accessed 03.12.12.).

Bentley, R.W., 2002. Global oil & gas depletion: an overview. Energy Policy 30 (3), 189–205. Available at: http://dx.doi.org/10.1016/S0301-4215(01)00144-6 (accessed 20.06.13.).

Berg, R.R., Calhoun, J.J.C., Whiting, R.L., 1974. Prognosis for expanded U.S. production of crude oil. Science 184 (4134), 331–336.

Brandt, A.R., 2007. Testing hubbert. Energy Policy 35 (5), 3074–3088. Available at: http://dx.doi.org/10.1016/j.enpol.2006.11.004 (accessed 30.05.13.).

Campbell, C.J., Laherrere, J.F., March 1998. The end of cheap oil. Scientific American, 80–86.

Carson, R., 2002. Silent Spring. Houghton Mifflin Harcourt. Available at: http://books.google.com/books?hl=en&lr=&id=6sRtTjwwWYEC&pgis=1 (accessed 03.06.13.).

Cook, C., Stevenson, J., 1996. The Longman Handbook of Modern British History, 1714–1995. Longman, London and New York.

Daly, H.E., 1997. Georgescu-Roegen versus Solow/Stiglitz. Ecological Economics 22 (3), 261–266. Available at: http://dx.doi.org/10.1016/S0921-8009(97)00080-3 (accessed 12.06.13.).

Dittmar, M., 1 September 2013. The end of cheap uranium. The Science of the Total Environment 461–462, 792–798. Available at: http://dx.doi.org/10.1016/j.scitotenv.2013.04.035 (accessed 22.06.13.).

Forrester, J., 1995. The beginning of system dynamics. McKinsey Quarterly. Available at: http://leml.asu.edu/jingle/Web_Pages/EcoMod_Website/Readings/SD+STELLA/Forrester-Begin'g-SD_1989.pdf (accessed 03.06.13.).

Forrester, J., 1971. World Dynamics. Wright Allen Press, Cambridge, MA. ISBN13: 9780262560184. Available at: http://documents.irevues.inist.fr/handle/2042/29441 (accessed 03.06.13.).

Groshong, K., 2007. The Noisy Response to Silent Spring; Placing Rachel Carson's Work in Context. Available at: http://www.bshs.org.uk/wp-content/uploads/file/bshs_monographs/library_monographs/bshsm_013_chang-and-jackson.pdf#page=371 (accessed 03.06.13.).

Guseo, R., 2011. Worldwide cheap and heavy oil productions: a long-term energy model. Energy Policy 39 (9), 5572–5577. Available at: http://dx.doi.org/10.1016/j.enpol.2011.04.060 (accessed 22.05.13.).

Hall, Charles A.S., Day, J.W., 1990. Ecosytem Modeling in Theory and Practice: An Introduction with Case Histories.

Heinberg, R., 2007. Peak Everything: Waking up to the Century of Declines. New Society Publishers. Available at: http://www.indiebound.org/book/9780865715981 (accessed 22.06.13.).

Hoggan, J., 2009. Climate Cover-up: The Crusade to Deny Global Warming. Greystone Books Ltd. Available at: http://books.google.com/books?id=tQYjQzOkYK0C&pgis=1 (accessed 12.06.13).

Holland, S.P., 2013. Encyclopedia of Energy, natural resource, and environmental economics. Economics of Peak Oil, 146–150. Available at: http://www.sciencedirect.com/science/article/pii/B9780123750679001248 (accessed 24.06.13.).

Höök, M., Bardi, U., Feng, L., Pang, X., 2010. Development of oil formation theories and their importance for peak oil. Marine and Petroleum Geology 1 (9), 1995–2004. Available at: http://www.sciencedirect.com/science/article/pii/S0264817210001224 (accessed 20.06.13.).

Hotelling, H., 1931. The economics of Eshaustible resources. Journal of Political Economy 39 (2), 137–175.

Houghton, J.J., Jenkins, G.I., Ephruns, J.J., 1990. Climate Change: The IPCC Scientific Assessment. Cambridge. Available at: http://www.ipcc.ch/ipccreports/far/wg_I/ipcc_far_wg_I_full_report.pdf (Accessed 20.06.13.).

Hubbert, M.K., 1956. Nuclear energy and the fossil fuels. In: Spring Meeting of the Southern District Division of Production American Petroleum Institute. American Petroleum Institute, Sant Antonio, TX. Available at: energycrisis.biz/hubbert/1956/1956.pdf (accessed 03.12.12).

Jevons, W.S., 1866. The Coal Question, second revised ed. Macmillan and Co. Available at: http://www.econlib.org/library/YPDBooks/Jevons/jvnCQ.html.

Kirby, M.W., 1977. The British Coalmining Industry, 1879–1946. The Macmillan press Ltd, London and Basingstoke.

Lynch, M.C., 2002. Forecasting oil supply: theory and practice. The Quarterly Review of Economics and Finance 42 (2), 373–389. Available at: http://www.sciencedirect.com/science/article/pii/S1062976902001266 (accessed 20.06.13.).

Lynch, M.C., 1999. Oil scarcity, oil crises, and alternative energies – don't be fooled again. Applied Energy, 31–53. Available at: http://www.sciencedirect.com/science/article/pii/S0306261999001233 (accessed 20.06.13.).

Maugeri, L., 2012. Oil: The Next Revolution. Belfert Center for Science and…. Available at: http://belfercenter.ksg.harvard.edu/files/Oil-%20The%20Next%20Revolution.pdf.

McGarity, T., Wagner, W., 2008. Bending Science: How Special Interests Corrupt Public Health Research. Available at: http://books.google.it/books?hl=en&lr=&id=Ah6-__otORAC&oi=fnd&pg=PR7&dq=tobacco+bending+science&ots=TnemcvW6lu&sig=PmolpgureQvVLlq3MlsSND_GGQM (accessed 03.06.13.).

Meadows, D.H., 1999. Leverage Points: Places to Intervene in a System. The Sustainability Institute. Available at: http://thesolutionsjournal.com/node/419 (accessed 03.06.13.).

Meadows, D.H., et al., 1972. The Limits to Growth. Universe Books.

Meadows, D.H., 2008. Thinking in Systems: A Primer. Available at: http://books.google.it/books?hl=en&lr=&id=JSgOSP1qkIUC&oi=fnd&pg=PR9&dq=donella+meadows+system+dynamics&ots=qeoPRy2DoM&sig=iYdD5SnfjI-E5dPJGWY7S-SXah8 (accessed 03.06.13.).

Nordhaus, W.D., 1992. Lethal models. Brookings Papers on Economic Activity 2, 1–59. Available at: http://scholar.google.it/scholar?hl=en&q=nordhaus+lethal+models&btnG=&as_sdt=1,5&as_sdtp=#1 (accessed 03.06.13.).

Oreskes, N., 1999. The Rejection of Continental Drift: Theory and Method in American Earth Science. Oxford University Press, New York. Available at: http://books.google.it/books?hl=en&lr=&id=EEQdk9GRfkoC&oi=fnd&pg=PA3&dq=wegener+continental+drift+1912&ots=i0TLHYy5wA&sig=OVUOCQ26QJpcqZCKwpwoVt0rtaU (accessed 15.06.13.).

Oreskes, N., Conway, E.M., 2010. Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming. Bloomsbury Publishing USA. Available at: http://books.google.com/books?id=fpMh3nh3JI0C&pgis=1 (accessed 12.06.13.).

Patzek, T.W., Croft, G.D., 2010. A global coal production forecast with multi-Hubbert cycle analysis. Energy 35 (8), 3109–3122. Available at: http://dx.doi.org/10.1016/j.energy.2010.02.009 (accessed 22.06.13.).

Simon, J., 1981. The Ultimate Resource. Princeton University Press.

Solow, R., 1974. The economics of resources or the resources of economics. American Economic Review, 1–14.

Stiglitz, J.E., 1979. A neoclassical analysis of the economics of natural resources. In: Kerry, Smith V. (Ed.), Scarcity and Growth Reconsidered. Johns Hopkins Press, Baltimore, MD.

Wegener, A., 1912. Die Herausbildung der Grossformen der Erdrinde (Kontinente und Ozeane), auf geophysikalischer Grundlage. In: Petermanns Geographische Mitteilungen. Presented at the Annual Meeting of the German Geological Society, Frankfurt Am Main (January 6, 1912). Annual Meeting of the German Geological Society, Frankfurt am Main, vol. 63, pp. 185–195, 253–256, 305–309.

Yergin, D., 2011. The Quest: Energy, Security, and the Remaking of the Modern World. Available at: http://scholar.google.it/scholar?hl=en&q=daniel+yergin&btnG=&as_sdt=1%2C5&as_sdtp=#6 (accessed 20.06.13.).

Zimmermann, E., 1933. World Resources and Industries. Harper and Brothers, New York.

Zittel, W., et al., 2013. Fossil and Nuclear Fuels: The Supply Outlook. Available at: http://www.energywatchgroup.org/fileadmin/global/pdf/EWG-update2013_long_18_03_2013up.pdf.

Minimum Wages

William Wascher, Board of Governors of the Federal Reserve System, Washington, DC, USA

Published by Elsevier Ltd.

Abstract

The minimum wage is a policy commonly used in both advanced and emerging market economies to support the wages of low-wage workers and to increase the income of the poor. However, the efficacy of the minimum wage as a policy tool is controversial. Although minimum wages do raise the wages of low-skilled workers, they also reduce employment opportunities for such workers and raise the prices of goods and services produced with low-wage labor. On balance, minimum wages appear to have little, if any, beneficial effect on the distribution of income.

Introduction

Minimum wage laws specify the minimum rate that a worker covered by the law can be paid. Minimum wages are commonly set by governments in both advanced and emerging market countries in an attempt to prohibit unfairly low pay and increase the income of workers at the low end of the wage distribution. First used in Australia in the late 1800s to combat manufacturing 'sweatshops,' minimum wages subsequently were introduced in the United Kingdom and the United States in the early part of the twentieth century and have since been implemented in many other countries. According to the International Labour Organization (2012), more than 130 countries around the world currently have a minimum wage of some type, although the specific levels of minimum wages, their coverage, and the processes used to determine them can differ substantially across these countries.

Proponents of the minimum wage argue that it is an important policy tool aimed at ensuring that workers receive a 'fair' wage for their work and helping individuals or families achieve self-sufficiency through the workplace. And, indeed, minimum wages are very popular among the public because they raise the wages of many low-skilled workers. However, while supportive of the goals of minimum wages, many economists are less sanguine about their likely effects on the low-wage labor market. They note that minimum wages increase the costs to firms of employing low-skilled labor, which can induce employers to reduce their hiring or raise prices, offsetting many of the economic benefits of higher wages. In addition, they point out that many minimum wage workers are not from poor families, limiting the effectiveness of the policy tool in alleviating poverty. In this sense, the minimum wage creates economic winners and losers, and the welfare consequences of such a policy depend on who comprises each group.

Characteristics of Minimum Wage Laws

Although in widespread use around the world, the characteristics of minimum wage laws and how they are determined are different across countries. In many countries, the minimum wage is a national standard set by federal government legislation or statute; this is the case, for example, in Australia, France, and several southern European countries. In other countries, such as Canada and Japan, minimum wages are set at the regional level. And a few countries, including the United States, have both a national minimum wage and wage floors set by narrower geographic jurisdictions, such as states or cities. In addition, while minimum wages in many countries are determined as part of the legislative process, in other countries the minimum wage may be determined through a national collective bargaining process consisting of representatives from employee and employer groups, as well from the national government.

Even apart from these differences in the wage-fixing process, the effective level of the minimum wage varies widely around the world. The International Labour Organization defines the minimum wage as "the payment to the employed of a wage adequate to maintain a reasonable standard of life as this is understood in their time and country" (ILO, 2012: p. 1). In practice, however, minimum wages may differ across countries according to the political climate and in response to varying concerns about the unintended consequences of setting a wage floor on employment and, more generally, the competitiveness of firms that employ low-wage labor. Moreover, even the concept of a minimum wage is not standard: While minimum wage laws refer specifically to the base wage in many countries, in others wage supplements – such as tips, commissions, or productivity bonuses – are counted toward the minimum wage.

In addition, the proportion of workers covered by minimum wage laws also differs substantially across countries. In many countries, all workers are covered by minimum wage legislation and all employers are subject to those laws. In other countries, however, certain groups of workers are exempt from the minimum wage or can be paid a legislated subminimum wage below the normal minimum wage. In particular, youths or trainees can sometimes be paid less than the regular minimum wage, while household workers and others working in the informal sector are often excluded from minimum wage laws.

Economic Effects of Minimum Wages

The economic effects of minimum wages have been studied extensively by economists over the past century. Despite the

attention afforded this issue, economists remain divided on the merits of the minimum wage. In part, this may reflect the fact that theoretical models of the labor market differ in their implications for the effects of the minimum wage on employment, income, and prices. More importantly, empirical studies have yielded a wide range of estimated effects. However, an extensive review of the minimum wage literature by David Neumark and William Wascher (2008) concluded that "minimum wages reduce employment opportunities for less-skilled workers" and "do not, on net, reduce poverty or otherwise help low-income families" (p. 6). A more recent review of the U.S. literature by the United States Congressional Budget Office (2014) also concluded that a higher minimum wage would likely result in fewer jobs for low-wage workers, although this review concluded that a higher minimum wage would reduce the number of families with incomes below the poverty line.

Theoretical Considerations

The majority of research on minimum wages – both theoretical and empirical – has focused on their effects on employment. In the simplest textbook model of the labor market, in which labor and product markets are assumed to be perfectly competitive, the equilibrium wage is determined in the marketplace so that the supply of labor is equal to the demand for labor (point L in **Figure 1**). If all workers are covered by the minimum wage, setting the minimum wage above the equilibrium wage leads to a decline in employment through a reduction in the demand for labor (point L_{MW} in the figure). This reduction in labor demand occurs through two distinct mechanisms. First, the minimum wage increases each firm's cost of production, and thus leads to a rise in the price of output, which, in turn, induces a decline in demand for it. In response to the weaker demand for output, production falls, reducing the demand for labor (this is referred to as the 'scale effect'). Second, in response to the higher cost of labor induced by the minimum wage, employers adjust their production processes to shift the mix of inputs they use away from labor toward capital and other inputs (this is referred to as the 'substitution effect'). Moreover, in competitive models with more than one type of labor, firms will substitute away from the low-skilled workers toward higher-skilled workers whose wages are unaffected by the minimum wage ('labor–labor substitution'). The combined result of these responses is a decline in employment of low-skilled workers who are covered by minimum wage laws.

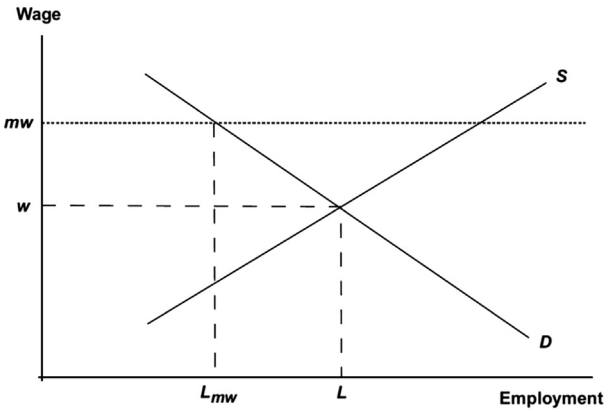

Figure 1 The textbook treatment of minimum wages.

Note that the effects on unemployment can be ambiguous in many of these models. In the simple textbook model with full coverage illustrated in **Figure 1**, a higher minimum wage causes labor demand to fall and labor supply to rise, resulting in an increase in unemployment. However, in a model in which not all sectors of the economy are covered by a minimum wage, workers displaced by the minimum wage in the covered sector may take a job instead in the uncovered sector or drop out of the labor force, mitigating the effect of the higher minimum wage on unemployment in the aggregate. Alternatively, the higher minimum wage in the covered sector may induce individuals in the uncovered sector to search for work in the covered sector, thereby increasing the supply of labor to the covered sector and exacerbating the effect on unemployment.

In addition, not all theoretical models of the labor market predict that minimum wages will reduce the demand for low-skilled labor. In particular, in so-called 'monopsony' models that assume that employers have market power over their workforce, an increase in the minimum wage can, under certain conditions, lead to an increase in employment (Stigler, 1946). In such models, when individual employers have to raise the wage to hire an additional worker, they must also increase the wage paid to existing employees. Absent a minimum wage, this condition implies that the market-determined level of employment and the market-determined wage will fall below those in the competitive model. In this case, imposing a minimum wage between the monopsony wage and the competitive wage will lead to an increase in employment, although, it is still the case that boosting the minimum wage above the competitive equilibrium wage will begin to reduce employment for low-skilled workers, as predicted by the competitive model.

Empirical Evidence on the Employment Effects of Minimum Wages

Empirical research on the minimum wage dates back to the early 1900s, when economists at the U.S. Department of Labor investigated the effects of a new minimum wage law introduced in Oregon in 1913. Throughout the next several decades, and in particular, following the establishment of a national minimum wage in the United States in 1938, economists continued to evaluate the effects of minimum wages on employment with varying results. Of particular note was the so-called 'marginalist' debate, in which George Stigler and others argued that the results from minimum wage studies supported the neoclassical view that minimum wages reduce labor demand, while Richard Lester and others viewed the evidence as supporting the institutionalist view that marginal costs were not an important determinant of low-wage employment.

Following a series of minimum wage increases in the 1960s and 1970s and given the ongoing debate about their effects, the United States Congress established the Minimum Wage Study Commission in 1977 to investigate "the effects of increasing

the minimum wage on inflation, employment, and unemployment; the ability of the minimum wage to ameliorate poverty; and the economic consequences of proposals for automatic indexing of the minimum wage and for setting a special subminimum rate for young workers" (Eccles and Freeman, 1982: p. 226). As part of that effort, three of the Commission's staff economists surveyed the existing literature and did their own research, eventually publishing a summary concluding that "time-series studies typically find that a 10 percent increase in the minimum wage reduces teenage employment by one to three percent," with smaller effects evident for 20–24 year olds (Brown et al., 1982). This study was widely viewed as having settled this issue and became the 'consensus' view of economists for the next decade or so.

In the early 1990s, however, the debate over the effects of the minimum wage reemerged. Armed with additional statistical variation generated by an increasing number of states with a minimum wage set higher than the federal level, a new set of researchers began to reassess the effects of the minimum wage on employment. In contrast to the time-series literature summarized by Brown et al., the evidence from this new research varied widely, with some researchers finding results that supported the consensus view (e.g., Neumark and Wascher (1992), Deere et al. (1995), and Burkhauser et al. (2000)), others finding no significant effects of minimum wages (Card, 1992a), and still others reporting positive effects on employment from a higher minimum wage (Card, 1992b; Katz and Krueger, 1992; Card and Krueger, 1994, 1995).

In an effort to synthesize this literature, David Neumark and William Wascher published a critical summary of the empirical literature on minimum wages that had been produced since the late 1980s (Neumark and Wascher, 2008). While acknowledging that the range of estimates in the more recent literature was noticeably wider than in the literature reviewed by Brown et al., Neumark and Wascher concluded "that the preponderance of evidence supports the view that minimum wages reduce the employment of low-skilled workers" (p. 104) and that the "adverse effects are even more apparent when research focuses on those directly affected by minimum wages" (p. 287).

Subsequent to the review of the literature by Neumark and Wascher, a few economists have challenged their conclusions. One line of argument relates to the appropriate way to summarize a large body of empirical research. In particular, Neumark and Wascher conducted a critical review of the literature, in which they focused on what they viewed as the most compelling evidence on the effects of the minimum wage. In contrast, Doucouliagos and Stanley conducted a meta-analysis and concluded that once the results are adjusted for publication selection bias there is little evidence of an adverse effect of the minimum wage on employment. However, Neumark (2013) questions the relevance of the meta-analysis results in assessing the body of minimum wage research because that approach treats all studies as providing equally valid evidence on minimum wage effects, regardless of the reliability of the methodology used to conduct the analysis or differences in the set of workers on which the studies focus.

A second line of argument asserts that a failure of the literature to take account of 'spatial heterogeneity' led to a negative bias in the disemployment effects reported in many of the studies included in Neumark and Wascher's survey (e.g., Allegretto et al., 2011; Dube et al., 2010). In particular, these authors claim that spurious correlation between minimum wage increases in particular states and underlying changes in low-wage employment in those states confounds the ability of standard panel-data techniques to accurately estimate the effects of minimum wages on employment, and they find that including state-specific time trends or limiting the sample to states or counties in particular geographic areas leads to estimates of minimum wage effects that are small and not statistically different from zero. However, these arguments have themselves been criticized as based on overly restrictive specifications that throw out too much identifying information to attempt to eliminate a potential bias that is of questionable importance (Neumark et al., 2014).

As noted above, the latest researchers to weigh in with a summary of the employment effects of minimum wages in the United States are from the independent Congressional Budget Office (CBO). In a February 2014 report, the CBO assessed the existing body of research and concluded that an increase in the minimum wage from its current level of $7.25 per hour to $10.10 per hour "would reduce total employment by about 500 000 workers, or 0.3 percent" (p. 1). The agency also noted that the standard error around this estimate is quite wide with "about a two-thirds chance that the effect would be in the range between a very slight reduction in employment and a reduction in employment of 1.0 million workers."

The controversy over the employment effects of minimum wages has not been limited to the United States. Although the majority of research in this area has tended to focus on U.S. labor markets, reflecting the long history of the federal minimum wage in the United States as well as the greater variation available from state-level minimum wages, there is a burgeoning literature on the effects of minimum wages in other countries. As in the United States, however, a consensus on the effects and thus the appropriate level of the minimum wage remains elusive, with ongoing debates in many countries, including the United Kingdom, France, Australia, and most recently Germany.

Effects of Minimum Wages on Poverty and the Distribution of Income

Although the employment effects of the minimum wage has received the most attention in the research on the implications of setting a floor on wages, the potential for job loss is not the only, or even necessarily the most important, outcome from raising the minimum wage. Many of those who support a higher minimum wage claim that they do so because, in their view, it will help low-income families by providing them with additional income that outweighs the possible negative consequences on overall employment. This challenge of balancing the potential benefits against the possible adverse employment effects of the minimum wage was eloquently captured nearly four decades ago by Edward Gramlich (1976), who wrote "Minimum wages do, of course, distort relative prices and hence compromise economic efficiency, but so do all other attempts to redistribute income through the tax and transfer system. The important question is not

whether minimum wages distort, but whether the benefits of any income redistribution they bring about are in some political sense sufficient to outweigh the efficiency costs" (p. 410).

Despite the compelling rationale for focusing on the effects of minimum wages on the distribution of income, research on this aspect of minimum wages is much less developed than the research on employment effects. However, what research does exist tends to find little net benefit from a higher minimum wage on the incomes of poor and near-poor families, with some studies finding modest positive effects (Dube, 2013) and others finding no effects or even modest negative effects (Neumark et al., 2005). According to Neumark and Wascher (2008), "the combined evidence is best summarized as indicating that an increase in the minimum wage largely results in a redistribution of income *among* low-income families, with some gaining as a result of the higher minimum wage and others losing as a result of diminished employment opportunities or reduced hours" (p. 189).

There are three reasons for this result. First, the minimum wage is not well targeted toward the poor because many workers who earn the minimum wage are not in poor families. For example, Burkhauser and Sabia (2010) find that only 11.3% of minimum wage workers were in poor families in 2007, while 42% were in families with incomes more than three times the poverty line. Second, roughly 50% of poor families have no workers at all. These families are obviously not in a position to benefit from a higher minimum wage. And third, as discussed earlier, a minimum wage increase is likely to lead to adverse employment effects on low-wage workers in poor families, offsetting the social benefits of the higher wages earned by those workers who keep their jobs. In the end, the recent CBO report concludes that the minimum wage would have a modest positive effect on the income of households below the U.S. poverty line, but also notes that the minimum wage is less effective at redirecting income toward the poor than some other policies, such as the Earned Income Tax Credit.

Other Effects of Minimum Wages

From a theoretical perspective, minimum wages can also affect other economic outcomes, including human capital acquisition and pricing behavior. For example, if a higher minimum wage increases the opportunity cost of attending school, some individuals who qualify for a minimum wage job may choose to work at the higher wage and thus acquire less education than they would if the wages associated with the available job opportunities were lower. Similarly, if a worker's on-the-job training is implicitly financed by the firm paying a lower wage, then an increase in the minimum wage could cause employers to reduce the amount of training provided to its employees. Alternatively, a minimum wage could increase the motivation for a worker to obtain additional training or schooling to better qualify for the higher-wage job, especially if the minimum wage reduces the availability of those jobs. With regard to prices, standard competitive models predict that some of the higher labor costs incurred by a firm following an increase in the minimum wage will be passed through to the prices of its product. However, in other models of the labor market, including the monopsony model described above as well as some versions of search and efficiency models, minimum wages can lead to a decline in prices, generally because these models predict an increase in output as a result of the minimum wage.

The empirical evidence, while limited, generally supports the view that minimum wages reduce human capital accumulation, although this conclusion is not clear-cut. With regard to schooling, higher minimum wages tend to be associated with lower school enrollment rates (Neumark and Wascher, 2008). For training, the evidence is more mixed. Some studies suggest that minimum wages cause employers to cut back on training, while others find little effect in either direction. However, other research finds that minimum wages have adverse longer-run effects on the earnings of adults who were exposed to a higher minimum wage in their youths, consistent with the possibility that the minimum wage reduced human capital formation for those individuals in their formative years.

With regard to pricing behavior, the empirical evidence generally indicates that minimum wages tend to result in price increases of goods or services produced with low-wage labor, consistent with the competitive model of low-wage labor markets. For example, MacDonald and Aaronson (2006) find that restaurants, especially limited-service outlets (e.g., fast-food restaurants) increase prices in response to a higher minimum wage, with the largest effects in low-wage areas where the effect of the minimum wage on labor costs is likely to be the greatest. That said, the effect of minimum wages on aggregate inflation is likely to be relatively small in most developed countries, reflecting the small share of minimum wage workers in the overall workforce. The exception might be in developing countries where more workers are affected by the minimum wage (say if other wages are set as multiples of the minimum wage), and especially if minimum wages are indexed to overall inflation.

Conclusion

The use of minimum wages is widespread around the world. While the exact nature of the wage floor and how it is determined differs across countries in many respects, the basic goals of the minimum wage are essentially the same everywhere: to provide higher wages for low-skilled workers and to provide additional income to poorer households.

Although these goals are laudable and important, whether the minimum wage can achieve them to any meaningful degree is questionable. Minimum wages obviously do raise the earnings of many low-wage workers. But they also appear to have negative effects on employment, which may adversely affect job opportunities for other low-skilled individuals. Moreover, the minimum wage is not well targeted toward the poor, as many minimum wage workers are in families well above the poverty line and many poor families have no workers. As a result, higher minimum wages tend to have only modest effects on the incomes of poor households, especially after taking into account the higher prices of goods and service passed on to consumers as a result of the wage floor.

Disclaimer

The views expressed are those of the authors and should not be interpreted as reflecting those of the Board of Governors of the Federal Reserve System.

See also: Income Distribution and Inequality: Measurement Issues; Poverty: Measurement and Analysis; Wage Differential and Structure; Welfare and Work.

Bibliography

Allegretto, Sylvia A., Dube, Arindrajit, Reich, Michael, 2011. Do minimum wages really reduce teen employment? Accounting for heterogeneity and selectivity in state panel data. Industrial Relations 50 (2), 205–240.

Brown, Charles, Gilroy, Curtis, Kohen, Andrew, 1982. The effect of the minimum wage on employment and unemployment. Journal of Economic Literature 20 (2), 487–528.

Burkhauser, Richard V., Couch, Kenneth A., Wittenburg, David C., 2000. A reassessment of the new economics of the minimum wage literature with monthly data from the current population survey. Journal of Labor Economics 18 (4), 653–680.

Burkhauser, Richard V., Sabia, Joseph J., 2010. Minimum wages and poverty: will a $9.50 federal minimum wage really help the working poor? Southern Economic Journal 76 (3), 592–623.

Card, David, 1992a. Using regional variation in wages to measure the effects of the federal minimum wage. Industrial and Labor Relations Review 46 (1), 22–37.

Card, David, October 1992b. Do minimum wages reduce employment? A case study of California, 1987–1989. Industrial and Labor Relations Review 46 (1), 38–54.

Card, David, Krueger, Alan B., 1994. Minimum wages and employment: a case study of the fast-food Industry in New Jersey and Pennsylvania. American Economic Review 84 (5), 772–793.

Card, David, Krueger, Alan B., 1995. Myth and Measurement: The New Economics of the Minimum Wage. Princeton University Press, Princeton, NJ.

Congressional Budget Office, February 2014. The Effects of a Minimum-wage Increase on Employment and Family Income Washington, DC.

Deere, Donald, Murphy, Kevin M., Welch, Finis, 1995. Employment and the 1990–1991 minimum-wage hike. American Economic Review Papers and Proceedings 85 (2), 232–237.

Doucouliagos, H., Stanley, T.D., 2009. Publication selection bias in minimum-wage research? A meta-regression analysis. British Journal of Industrial Relations 47 (2), 406–428.

Dube, Arindrajit, 2013. Minimum Wages and the Distribution of Family Incomes, unpublished manuscript.

Dube, Arindrajit, William Lester, T., Reich, Michael, 2010. Minimum wage effects across state borders: estimates using contiguous counties. Review of Economics and Statistics 92 (4), 945–964.

Eccles, Mary, Freeman, Richard B., 1982. What! Another minimum wage study? American Economic Review Papers and Proceedings 72 (2), 226–232.

Gramlich, Edward M., 1976. Impact of minimum wages on other wages, employment, and family incomes. Brookings Papers on Economic Activity 2, 409–451.

International Labour Organization, 2012. Minimum wages. In: Global Wage Report 2012/13: Wages and Equitable Growth. Geneva.

Katz, Lawrence F., Krueger, Alan B., 1992. The effect of the minimum wage on the fast food Industry. Industrial and Labor Relations Review 46 (1), 6–21.

Lester, Richard A., 1946. Marginalism, minimum wages, and labor markets. American Economic Review 37 (1), 135–148.

MacDonald, James, Aaronson, Daniel, 2006. How firms construct price changes: evidence from restaurant responses to increased minimum wages. American Journal of Agricultural Economics 88 (2), 292–307.

Neumark, David, 2013. Do Minimum Wages Destroy Jobs? IZA World of Labor, Bonn, Germany.

Neumark, David, Wascher, William, 1992. Employment effects of minimum and subminimum wages: panel data on state minimum wage laws. Industrial and Labor Relations Review 46 (1), 55–81.

Neumark, David, Wascher, William, 2007. Minimum wages and employment. Foundations and Trends in Microeconomic 3 (1–2), 1–186.

Neumark, David, Wascher, William, 2008. Minimum Wages. The MIT Press, Cambridge, MA.

Neumark, David, Salas, J.M. Ian, Wascher, William, 2014. Revisiting the minimum wage-employment debate: throwing out the baby with the bathwater? Industrial and Labor Relations Review 67 (3), 608–648.

Neumark, David, Schweitzer, Mark, Wascher, William, 2005. The effects of minimum wages on the distribution of family incomes. Journal of Human Resources 40 (4), 867–894.

Obenauer, Marie L., Nienburg, Bertha von der, 1915. Effect of Minimum-Wage Determination in Oregon. Bureau of Labor Statistics Bulletin No. 176. United States Department of Labor, Washington, DC.

Stigler, George J., 1946. The economics of minimum wage legislation. American Economic Review 36 (3), 358–365.

Minnesota Multiphasic Personality Inventory

James N Butcher, University of Minnesota, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Minnesota Multiphasic Personality Inventory (MMPI)-2 is the most widely used personality inventory in clinical, forensic, and personnel assessment settings. It is a self-report personality inventory based upon items that address mental health symptoms, personality traits, and behavioral characteristics. The MMPI-2 provides mental health professionals with information that allows them to appraise psychological adjustment. Moreover, the test includes a number of self-report validity indices that allow the practitioner to determine the client's willingness and frankness to cooperate with the appraisal.

The MMPI, the Minnesota Multiphasic Personality Inventory, has had a long and successful tradition in personality assessment. The test was originally developed in the 1940s as a means of evaluating mental health problems in psychiatric and medical settings. The test authors, Starke Hathaway and J.C. McKinley, thought that it was important in evaluating patients' problems to ask them about what they felt and thought. Their instrument was a self-report inventory that included a very broad range of symptoms and behaviors and could be answered with a sixth-grade reading level. The MMPI was developed according to rigorous empirical research methods and rapidly became the standard personality instrument in clinical settings (Hathaway and McKinley, 1940). The popularity of the true–false personality inventory was in large part due to its easy-to-use format and to the fact that the scales have well-established validity in assessing clinical symptoms and syndromes (Butcher, 2011). The MMPI underwent a major revision in the 1980s resulting in two forms of the test – an adult version, the MMPI-2 (Butcher et al., 1989) and an adolescent form, MMPI-A (Butcher et al., 1992). The MMPI-2 is the most widely researched instrument and is used for the evaluation of clinical problems in a broad range of settings including mental health, health psychology, correctional settings, and personnel screening, and in many forensic applications such as child custody and personal injury (Bow et al., 2010; Martin et al., 2001; Lally, 2003).

The MMPI-2 contains 567 true–false questions addressing mental health symptoms, beliefs, and attitudes. The items on the MMPI-2 are grouped into scales (clusters of items) that address specific clinical problems such as depression or anxiety. After the inventory is completed, the items are scored or grouped according to the scales that have been developed. An MMPI scale allows the clinician to compare the responses of the client with those of thousands of other people. Initially, the scores are compared with the normative sample, a large representative sample of people from across the United States, to determine if the person's responses are different from people who do not have mental health problems. If the person obtains scores in the extreme ranges, for example, on the depression scale, compared with the normative sample then they are likely to be experiencing problems comparable with the clinical samples of depressed clients that have been studied.

Evaluating Cooperative Responding

In some situations clients might be motivated to present personality characteristics and problems in ways that are different than they actually are. For example, if people are being tested to determine whether or not they are 'sane' enough to stand trial in a criminal court case they might attempt to exaggerate symptoms or problems to avoid responsibility. Alternatively, people being evaluated in preemployment psychological evaluation might be inclined to present themselves in an extremely positive way to cover up problems. It is important, in MMPI-2 profile interpretation, to evaluate the way in which people approach the task of self-revelation. Are they sufficiently cooperative with the testing to produce a valid result? There are several indices on the MMPI-2 to address honesty and cooperativeness in responding to the items (Baer et al., 1995):

Cannot Say Score

This index is simply the total number of unanswered items. If the client leaves out many items on the test (e.g., 30 items), the profile may be invalidated. A high score on this index suggests that clients have not cooperated sufficiently to provide a clear picture of their personality functioning.

The L Scale

The L or Lie scale is a measure of the client's willingness to acknowledge personal faults or problems. Individuals who score high on this scale are presenting an overly virtuous picture of themselves. They claim virtue and positive characteristics that are not typically found among people in general. The L scale is particularly valuable in situations where the individual has something to gain by not appearing to have problems, such as in personnel screening or in child-custody disputes. In these cases, people might try to put their best foot forward and present themselves as 'better' adjusted than they actually are.

The K Scale

The K scale was developed as a means of evaluating the tendency that some people have to minimize problems. This

 http://dx.doi.org/10.1016/B978-0-08-097086-8.25019-8

scale is also used as an index for correcting some clinical scales that have been shown to be affected by defensiveness; that is, it serves as a correction factor to compensate for the tendency of some people to deny problems. Empirically derived percentages of the K scale score are added to five of the clinical scales to improve test discrimination.

The Infrequency Scales

The MMPI-2 contains three measures to assess the tendency of some people to exaggerate their problems or 'fake' the test by overresponding to extreme items (the F or Infrequency scale, the Back F or F(B) scale, and the P(p) scale). Each scale performs a somewhat different function in interpretation. The items on these scales are very rare or bizarre symptoms. Individuals who endorse a lot of these items tend to exaggerate symptoms on the MMPI-2, perhaps as a way of trying to convince professionals that they need to have psychological services. Some people who have a need to claim problems to influence court decisions will tend to elevate the infrequency scales. The infrequency scales can be elevated for several possible reasons: the profile could be invalid because the client became confused or disoriented or responded in a random manner. High F and F(B) scores are commonly found among clients who are malingering or producing exaggerated responses in order falsely to claim mental illness.

TRIN and VRIN Scales

Two scales to assess inconsistent responding have been developed for the MMPI-2. These scales were derived by examination of the individual's responses to pairs of items that have similar content. A particular response to one of the items in the pair requires that the other item be endorsed in a given way to be consistent: for example, a pair of items that contains content that cannot logically be answered in the same direction if the subject is responding consistently to the content.

Profile Interpretation

The assessment of clinical problems is approached by examining the client's responses to the items using several strategies. We will examine three types of scales that comprise the MMPI-2's problem measures: the traditional clinical scales and profile codes, the MMPI-2 content scales, and the supplemental or specific-problem scales.

The MMPI-2 Clinical Scales

Hathaway and McKinley developed the original MMPI clinical scales to classify empirically or group patients into problem types. For example, they developed scales to assess hypochondriasis (the Hs scale), depression (the D scale), hysteria (the Hy scale), psychopathic deviation (the Pd scale), paranoid thinking (the Pa scale), psychasthenia (the Pt scale), schizophrenia (the Sc scale), and mania (the Ma scale). In addition, two other scales were included on the basic profile to address problems of sex-role identification (the Mf scale) and social introversion and extraversion (the Si scale). These scales have clearly defined correlates or empirical relationships. If a scale is elevated in the significant range, the individual is considered to possess the attributes measured by the scale.

Some clients have only one of the clinical scales elevated in their profile. In these cases, the profile interpretation is straightforward – the clinician simply examines the published correlates for that scale and incorporates these into the psychological report. However, some clients might have several of the clinical scales elevated in their profile. The full pattern or profile configuration needs to be interpreted. Fortunately, extensive research has been undertaken to evaluate complex profiles or code types. These profile types result from having two or more of the clinical scales elevated in the interpretive range. In these cases, the practitioner refers to published 'code books' which provide empirical descriptions for the various patterns that have been studied.

Content-Based Scales

The MMPI-2 contains a number of scales that assesses content themes that the individual endorsed in the item pool. The content scales are homogeneous item clusters that summarize themes and represent direct communications about problems to the practitioner. There are 15 content scales measuring different symptom areas and problems: for example, Antisocial Practices (ASP), Bizarre Mentation (BIZ), and Low Self-Esteem (LSE).

Special-Problem or Supplemental Scales

Several supplemental scales have been developed to assess specific problems such as the potential to develop problems of addiction (the MacAndrew Addiction scale or MAC-R and the Addiction Potential scale or APS) and whether or not the individual acknowledges having problems with drugs or alcohol. Another scale, the Marital Distress scale, addressed a client's attitudes toward their marital relationship. The MMPI-2 supplemental scales provide the practitioner with the means of evaluating specific problems (such as posttraumatic stress disorder) that are not addressed in the clinical or content scales.

Applications of the MMPI-2

Contemporary uses of the MMPI-2 include such activities as evaluating clients following their admission to an inpatient psychiatric facility; providing personality information about client's problems and attitudes in the early stages of psychological treatment to assess treatment amenability; providing personality information for therapists to employ in giving the client feedback in psychotherapy; and providing information that would aid the clinician in arriving at an appropriate clinical diagnosis. The MMPI-2 is also widely used in medical settings to provide psychological adjustment information on patients – for example, in rehabilitation programs as a measure of behavior problems and symptoms following a stroke as part of the neuropsychological evaluation; or in the evaluation of drug treatment effects.

A great deal of contemporary personality assessment research is also aimed at further exploring the MMPI-2 test

variables. The two most active research areas involve further exploration into the external validity of the MMPI-2 scales, and the translation and adaptation of the instrument for other languages and cultures. For example, research carried out in the mid-1990s has further explored the validity of the test in diverse settings such as personnel screening (Butcher, 1995); forensic assessment (Megargee, 2009); and test correlates with outpatients (Graham, 2000). Several studies have explored the use of MMPI-2-based personality descriptions in providing test feedback to psychotherapy clients. This research has shown that providing psychological test feedback in treatment increases the effectiveness of treatment (Finn, 1966; Finn and Tonsager, 1992; Newman and Greenway, 1997).

The MMPI-2 is also widely used as a cross-cultural research technique to study psychopathology across cultures. The MMPI-2 items and scales have shown remarkable resilience when used in other languages and cultures (Butcher, 1996), and computer-based interpretation of the MMPI-2 has been shown to be highly effective in describing clinical patients in other cultures (Butcher et al., 2009).

Conclusion

The MMPI-2 is the most widely used and most versatile clinical personality instrument that clinical psychologists have available. It is a self-report personality questionnaire that provides the test user with scores on a number of scales that address important clinical problem areas. In addition the MMPI-2 provides mental health professionals with information that allows them to appraise the client's willingness and frankness in the appraisal – key information about the client's response attitudes that allows for a determination of the credibility of the client's self-report. Once cooperation in the assessment is assured, the MMPI-2 clinical measures can be effective in delineating the mental health symptoms, personality traits, and special problems that the client is likely to be experiencing.

See also: Personality Assessment.

Bibliography

Baer, R.A., Wetter, M.W., Nichols, D.S., Greene, R., Berry, D.T.R., 1995. Sensitivity of MMPI-2 validity scales to underreporting of symptoms. Psychological Assessment 7, 419–423.

Butcher, J.N., 1995. User's Guide for the Minnesota Report: Revised Personnel Report. National Computer Systems, Minneapolis, MN.

Bow, J.N., Flens, J.R., Gould, J.W., 2010. MMPI-2 and MCMI-III in forensic evaluations: a survey of psychologists. Journal of Forensic Psychology Practice 10 (1), 37–52.

Butcher, J.N., 1996. International Adaptations of the MMPI-2. University of Minnesota Press, Minneapolis, MN.

Butcher, J.N., 2011. The MMPI-2: A Beginners Guide, third ed. American Psychological Association, Washington, DC.

Butcher, J.N., Berah, E., Ellertsen, B., Miach, P., Lim, J., Nezami, E., Pancheri, P., Derksen, J., Almagor, M., 1998. Objective personality assessment: computer-based MMPI-2 interpretation in international clinical settings. In: Belar, C. (Ed.), Comprehensive Clinical Psychology: Sociocultural and Individual Differences. National Computer Systems, Minneapolis, MN, pp. 277–312.

Butcher, J.N., Graham, J.R., Dahlstrom, W.G., Tellegen, A.M., Kaemmer, B., 1989. Minnesota Multiphasic Personality Inventory-2 (MMPT-2): Manual for Administration and Scoring. University of Minnesota Press, Minneapolis, MN.

Butcher, J.N., Perry, J., Dean, B.L., 2009. Computer based assessment. In: Butcher, J.N. (Ed.), Oxford Handbook of Personality Assessment. Oxford University Press, New York, pp. 163–182.

Butcher, J.N., Williams, C.L., Graham, J.R., Tellegen, A., Ben-Porath, Y.S., Archer, R.P., Kaemmer, B., 1992. Manual for Administration, Scoring, and Interpretation of the Minnesota Multiphasic Personality Inventory for Adolescents: MMPI-A. University of Minnesota Press, Minneapolis, MN.

Finn, S.E., 1966. Assessing feedback integrating MMPI-2 and Rorschach findings. Journal of Personality Assessment 67, 543–557.

Finn, S.E., Tonsager, M., 1992. Therapeutic effects of providing MMPI-2 test feedback to college students awaiting therapy. Psychological Assessment 4, 278–287.

Graham, J.R., 2000. MMPI-2: Assessing Psychopathology and Personality, second ed. Oxford, New York.

Hathaway, S.R., McKinley, J.C., 1940. A multiphasic personality schedule (Minnesota): 1. Construction of the schedule. Journal of Psychology 10, 249–254.

Lally, S.J., 2003. What tests are acceptable for use in forensic evaluations? A survey of experts. Professional Psychology: Research and Practice 34 (5), 491–498.

Megargee, E.I., 2009. Understanding and assessing aggression and violence. In: Butcher, J.N. (Ed.), Oxford Handbook of Personality and Clinical Assessment. Oxford University Press, New York, pp. 542–566.

Martin, M.A., Allan, A., Allan, M.M., 2001. The use of psychological tests by Australian psychologists who do assessments for the courts. Australian Journal of Psychology 53 (2), 77–82.

Newman, M.L., Greenway, P., 1997. Therapeutic effects of providing MMPI-2 test feedback to clients at a university counseling service: a collaborative approach. Psychological Assessment 9, 122–131.

Relevant Website

www.umn.edu/mmpi.

Minorities

Richard T Schaefer, DePaul University, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by Z. Layton-Henry, volume 14, pp. 9894–9898, © 2001, Elsevier Ltd.

Abstract

Minorities are subordinate groups whose members have significantly less control or power over their own lives than do the members of a dominant or majority group. In sociology, *minority* means the same as *subordinate*, and *dominant* is used interchangeably with *majority*. A minority group has five properties: unequal treatment, distinguishing physical or cultural traits, involuntary membership, awareness of subordination, and in-group marriage. Minority groups typically emerge through migration, annexation, and colonialism. The social consequences for minorities include extermination, expulsion, secession, segregation, fusion, and assimilation. Pluralism offers an alternative to complete societal dominance over minorities. Primarily through occupying an economic niche as self-employed entrepreneurs, middleman minorities provide desired services and sometimes products of use to both the majority and other minorities. Most typically, the term 'minorities' has been used to describe racial, ethnic or national, and religious minorities residing in a nation-state with a clearly defined majority (dominant) group. And also to writers such as youth or elderly, people with disabilities, exconvicts or exprisoners, and people with mental illnesses, people who are not solely heterosexual, and women to name a few.

Minorities have been the focus of study by social scientists for generations and the use of the term have a long history both in scholarship and in public discourse. Minorities are subordinate groups whose members have significantly less control or power over their own lives than do the members of a dominant or majority group. In sociology, *minority* means the same as *subordinate*, and *dominant* is used interchangeably with *majority*.

Minorities are therefore about unequal access to power and not necessarily a factor of sheer size and are being outnumbered by some other group or groups. Being superior in numbers does not guarantee a group control over its destiny and ensure majority status. Most dramatically in apartheid South Africa, Black Africans, Coloreds, and Asian Indians collectively outnumbered Whites by a ratio of 10:1, but were still very much subject to the control of the political and economic dominance of Afrikaners and English-speaking Whites. The economic legacy of apartheid rule from 1948 through 1994 is evident in the relative poverty rates in 2010 – 47% for black Africans and 0.5% for white South Africans (Kane-Berman, 2011: 312).

Similarly, in 1920, the majority of people in Mississippi and South Carolina were African-Americans. Yet African-Americans did not have as much control over their lives as Whites, let alone control of the states of Mississippi and South Carolina. Throughout the United States, today are counties or neighborhoods in which the majority of people are African-American, Native American, or Hispanic, but white Americans are the dominant force. Nationally in the United States, about 51% of the population is female, but males still dominate positions of authority and wealth well beyond their numbers.

While minorities may be defined by significantly less power as a group, that does not mean minorities are not without recourse. As we will note later, individual minority-group members may choose to immigrate as in the case of Jews to Israel. In addition, subjugated minorities may mobilize to confront the dominant majority as are Roman Catholics in Northern Ireland, Muslims in India, and Palestinians in Israel.

Confronted with evidence that a particular minority is subordinate to the majority, some people respond, "Why not? After all, the majority rules." However, the subordination of a minority involves more than its inability to rule over society. A member of a subordinate or minority group experiences a narrowing of life's opportunities – for success, education, wealth, the pursuit of happiness – that goes beyond any personal shortcoming he or she may have. A minority group does not share in proportion to its numbers what a given society, such as the United States, defines as valuable.

Characteristics of a Minority

A minority or subordinate group has five characteristics: unequal treatment, distinguishing physical or cultural traits, involuntary membership, awareness of subordination, and in-group marriage (Wagley and Harris, 1958).

1. Members of a minority experience unequal treatment and have less power over their lives than members of a dominant group have over theirs. Prejudice, discrimination, segregation, and even extermination create and maintain this social inequality.

The degree to which minority groups experience these disabling mechanisms, of course, varies with each group, and varies with the same group over time. The discrimination that black Americans encounter in the twenty-first century, although not as severe as under slavery, has still been discriminatory. To attempt to rank levels of discrimination by severity is difficult, and probably foolish. What constitute the worst denial of fights: the enslavement of black Africans, the decimation of tribal Indians by the United States Cavalry, or the placement of more than 110 000 Japanese Americans in relocation camps during World War II? Little is gained by ranking such injustices.

2. Members of a minority group share physical or cultural characteristics that distinguish them from the dominant

group, such as skin color or language. Each society has its own arbitrary standard for determining which characteristics are most important in defining dominant and minority groups.

By defining a group as a minority invoking certain physical traits, society isolates them from the dominant group, which encourages the development of certain cultural patterns unique to the minority.

3. Membership in a dominant or minority group is not voluntary; People are born into the group. A person does not choose to be an African-American or white.

Religious minorities may appear to be an exception to this property, but relatively speaking religion is largely passed on from parents to children. Conversion to a given faith does occur but those around them see often converts a bit differently whether inside the chosen faith or not.

4. Minority-group members have a strong sense of group solidarity. William Graham Sumner, writing in 1906, noted that people make distinctions between members of their own group (the in-group) and everyone else (the out-group). When a group is the object of long-term prejudice and discrimination, the feeling of 'us versus them' often becomes intense. Consequently, minority-group members develop an *esprit de corps*, a sense of 'I am one of them.'
5. Members of a minority generally marry others from the same group. A member of a dominant group often is unwilling to join a supposedly inferior minority by marrying one of its members. In addition, the minority group's sense of solidarity encourages marriage within the group and discourages marriage to outsiders.

In-group marriage, whether within a minority group or a majority group, is referred to as endogamy. The opposite, marrying out of one's own group, is exogamy. Even as recently as 2008, only 14.4% of all marriages by black men were with whites, and only 6.5% of all black women's marriages were with whites (Qian and Lichter, 2011).

The tendency of minority-group people to marry fellow group member is not an accident. The dominant group often has little respect on minority people and scorns those among the majority who, as they see it, lower themselves to marry someone of another group. Often, laws making intergroup marriages illegal buttress this scorn. For example when Barack Obama's parents married in Hawaii in 1961, the marriage between the white American Ann Dunham and the black Kenyan Barack Obama, Sr would not have been legal at the time in 22 states. Consequently, in-group marriage for minority people is not always due to choice, but can be a necessity as the majority refuses to form out-group marriages. Subordinate-group people may choose to marry among themselves in order to preserve their cultural heritage or in sincere feelings that marriage with a member of the dominant group would lead to personal unhappiness.

Emergence of Minority-Group Status

Minority groups typically emerge or, in some respects, disappear through a process that may occur in a very short period or over a series of generations. Three situations are likely to lead to the formation of a subordinate-group and dominant-group relationship. A subordinate group emerges through (1) migration, (2) annexation, and (3) colonialism.

Migration

People who immigrate to a new country often find themselves a minority in that new country. Cultural or physical traits or religious affiliation may set the immigrant apart from the dominant group. Immigration from Europe, Asia, and Latin America has been a powerful force in shaping the fabric of life in the United States.

Although people may immigrate because they want to, leaving the home country is not always voluntary. Conflict or war has displaced people throughout human history. In the twentieth century, we have seen huge population movements caused by two world wars; revolutions in Spain, Hungary, and Cuba; the partition of British India; conflicts in Southeast Asia, Korea, and Central America; and the confrontation between Arabs and Israelis. Environmental factors such as drought and floods have also been the source of migration from one's homeland.

In all types of geographical mobility, two sets of forces operate: *push factors* and *pull factors*. Push factors discourage a person from remaining where he or she lives. Religious persecution and economic factors such as dissatisfaction with employment opportunities are possible push factors. Pull factors, such as a better standard of living, friends and relatives who have already emigrated, and a promised job, attract an immigrant to a particular country.

While generally we think of migration as a voluntary process, much of the population transfer that has occurred in the world has been involuntary. The forced movement of people into another society guarantees a subordinate role. Involuntary migration is no longer common; although enslavement has a long history, all industrialized societies today prohibit such practices. Of course, many contemporary societies, including the United States, bear today the legacy of slavery.

Annexation

Nations, particularly during wars or as a result of war, incorporate or attach land. This new land is *contiguous* to the nation, as in the German annexation of Austria and Czechoslovakia in 1938 and 1939, respectively, and in the Louisiana Purchase of 1803. The Treaty of Guadalupe Hidalgo that ended the Mexican–American War in 1848 gave California, Utah, Nevada, most of New Mexico, and parts of Arizona, Wyoming, and Colorado to the United States. The indigenous peoples in some of this huge territory were dominant in their society one day, only to become minority-group members the next.

When annexation occurs, the dominant power generally suppresses the language and culture of the minority. Such was the practice of Russia in the immediate post-World War II era with the Ukrainians and Poles, and of Prussia with the Poles. Minorities try to maintain their cultural integrity despite annexation. Poles inhabited an area divided into territories ruled by three countries but maintained their own culture across political boundaries.

Colonialism

Colonialism has been the most frequent way for one group of people to dominate another. Colonialism is the maintenance of political, social, economic, and cultural domination over people by a foreign power for an extended period (W. Bell, 1991). Colonialism is rule by outsiders but, unlike annexation, does not involve actual incorporation into the dominant people's nation. The long control exercised by the British Empire over much of North America, parts of Africa, and India is an example of colonial domination.

Colonialism has often led indigenous people, such as tribal groups, to become a minority in an area they once were the majority (dominant) group. Examples of this process include the Maori of New Zealand, the First Nation people of Canada, and the Hawaiians of Hawaii. Societies gain power over a foreign land through military strength, sophisticated political organization, and the massive use of investment capital. The extent of power may also vary according to the dominant group's scope of settlement in the colonial land. Relations between the colonial nation and the colonized people are similar to those between a dominant group and exploited subordinate groups. The colonial subjects are generally limited to menial jobs and the wages from their labor. The natural resources of their land benefit the members of the ruling class.

Colonialism is domination by outsiders. Relations between the colonizer and the colony are similar to those between the dominant and subordinate peoples within the same country. This distinctive pattern of oppression is called internal colonialism. Among other cases, it has been applied to the plight of blacks in the United States and Mexican Indians in Mexico, who are colonial peoples in their own country. Internal colonialism covers more than simple economic oppression. Nationalist movements in African colonies struggled to achieve political as well as economic independence from Europeans. Similarly, African-Americans also call themselves nationalists in trying to gain more autonomy over their lives (Blauner, 1969, 1972).

Consequences of Minority-Group Status

There are several consequences for minorities of subordinate status. These differ in their degree of harshness, ranging from physical annihilation to absorption into the dominant group. In this section, we will examine six consequences of minority status: (1) extermination, (2) expulsion, (3) secession, (4) segregation, (5) fusion, and (6) assimilation. Finally, we consider pluralism that offers an alternative to complete societal dominance over minorities.

Extermination

The most extreme way of dealing with a subordinate group is to eliminate it. Today, the term genocide is used to describe the deliberate, systematic killing of an entire people or nation. This term is frequently used in reference to the Holocaust – Nazi Germany's extermination of 12 million European Jews and other ethnic minorities during World War II. More recently, the term ethnic cleansing was introduced into the world's vocabulary as ethnic Serbs instituted a policy intended to 'cleanse' – eliminate – Muslims from parts of Bosnia. Again in 1994, a genocidal war between the Hutu and Tutsi people in Rwanda left 300 000 school-age children orphaned.

Expulsion

Dominant groups may choose to force a specific subordinate group to leave certain areas or even vacate a country. Expulsion is therefore another extreme consequence of minority-group status. European colonial powers in North America and eventually the US government itself drove virtually all Native Americans out of their tribal lands into unfamiliar territory.

More recently, beginning in 2009, France expelled over 10 000 ethnic Roma (or Gypsies), forcing their return to their home countries of Bulgaria and Romania. This appeared to violate the European Union's (EU) ban against targeting ethnic groups as well as Europe's policy of 'freedom of movement.' In 2011, the EU withdrew its threat of legal action against France when the government said it would no longer expel Roma in particular but only those living in 'illegal camps,' which many observers felt was only a technical way for the country to get around long-standing human rights policies (Schaefer and Zellner, 2011).

Secession

A group ceases to be a subordinate group when it secedes to form a new nation or moves to an already established nation where it becomes dominant. After Great Britain withdrew from Palestine, Jewish people achieved a dominant position in 1948, attracting Jews from throughout the world to the new state of Israel. In a similar fashion, Pakistan was created in 1947 when India was partitioned. The predominantly Muslim areas in the north became Pakistan, making India predominantly Hindu. Throughout this century, minorities have repudiated dominant customs. In this spirit, the Estonian, Latvian, Lithuanian, and Armenian peoples, for example, not content to be merely tolerated by the majority, all seceded to form independent states following the demise of the Soviet Union in 1991.

Some African-Americans have called for secession. Suggestions dating back to the early 1700s supported the return of blacks to Africa as a solution to racial problems. The settlement target of the American Colonization Society was Liberia, but proposals were also advanced to establish settlements of freed slaves in Canada, Haiti, South America, and the western United States. Territorial separatism and the emigrationist ideology were recurrent and interrelated themes among African-Americans from the late nineteenth century well into the 1980s. The black Muslims, or Nation of Islam, once expressed the desire for complete separation in their own state or territory within the present borders of the United States. Although a secession of blacks from the United States has not taken place, it has been considered and even proposed in the past.

Segregation

Segregation refers to the physical separation of two groups in residence, workplace, and social functions. Generally, the dominant group imposes segregation on a subordinate group.

Segregation is rarely complete, however, intergroup contact inevitably occurs even in the most segregated societies.

The extent of racial and ethnic isolation globally and historically is staggering. Carl Nightingale (2012) observes that 'city splitting' took place in the past as the British East India Company, split Calcutta into 'White Town' and 'Black Town.' In Germany today, they speak of *Ghettoisierung* (or 'ghettoization') as concentrations of Turkish immigrants are emerging. Social scientists in Sweden have used the segregation index to document the isolation of Greek, Chilean, and Turkish immigrants from the rest of the population. Segregation is present in virtually all multiracial societies (Massey, 2012).

The focus on metropolitan areas should not lead one to overlook the continuing legally sanctioned segregation in rural areas as in the placement of Native Americans on reservations. Although the majority of our nation's first inhabitants do live outside these tribal areas, the reservations play a prominent role in the identity of Native Americans. They face the difficult choice that, while it is easier to maintain tribal identity on the reservation, economic and educational opportunities are more limited in these areas segregated from the rest of society.

Fusion

Fusion describes the result when a minority and a majority group combine to form a new group. This combining can be expressed as $A + B + C \rightarrow D$, where A, B, and C represent the groups present in a society, and D signifies the result, a cultural racial group unlike any of the initial groups. Theoretically, fusion does not require intermarriage, but it is very similar to amalgamation or the cultural and physical synthesis of various groups into a new people. In everyday speech, the words *fusion* and *amalgamation* are rarely used, but the concept is expressed in the notion of a human melting pot, in which diverse racial or ethnic groups form a new creation, a new cultural entity (Newman, 1973).

As noted earlier, when identifying the properties of minorities, there is only modest evidence in the United States of a fusion of races. Racial intermarriage has been increasing, and the number of interracial couples immigrating to the United States has also grown. This growth of the so-called 'mixed marriages' can be seen in the growing presence of children of biracial or multiracial parentage, leading the Bureau of the Census to allow individuals to indicate more than one race in the decennial censuses. Most people did select one racial category in Census 2000 and again in 2010. Overall, approximately 9 million people, or 2.9% of the total population, selected two or more racial groups in 2010. Some self-identified biracial or multiracial people see themselves as a minority distinct from those identifying a single race or ethnic ancestry (Humes et al., 2011).

Assimilation

Assimilation is the process by which a subordinate individual or group takes on the characteristics of the dominant group and is eventually accepted as part of that group. Assimilation is a majority ideology in which $A + B + C \rightarrow A$. The majority (A) dominates in such a way that the minorities (B and C) become indistinguishable from the dominant group. Assimilation dictates conformity to the dominant group, regardless of how many racial, ethnic, or religious groups are involved (Gordon, 1964; Newman, 1973: 53).

To be complete, assimilation must entail an active effort by the minority-group individual to shed all distinguishing actions and beliefs and the total, unqualified acceptance of that individual by the dominant society. In the United States, dominant white society encourages assimilation. The assimilation perspective tends to devalue alien culture and to treasure the dominant. For example, assimilation assumes that whatever is admirable among blacks was adapted from whites and that whatever is bad is inherently black. The assimilation solution to black–white conflict is the development of a consensus around white American values.

Assimilation is difficult for minorities. The individual must forsake his or her cultural tradition to become part of a different, often antagonistic, culture. Members of the subordinate group who choose not to assimilate look on those who do as deserters. Many Hindus in India complained of their compatriots who copied the traditions and customs of the British. Those who assimilate must totally break with the past. Australian Aborigines who become part of dominant society conceal their origin by ignoring darker-skinned relatives, including their immediate family. As one assimilated Aborigine explained, "Why should I have anything to do with them? I haven't anything in common with them, and I don't even like most of them" (Berndt and Berndt, 1951: 264). These conflicting demands leave an individual torn between two value systems.

Pluralism

The alternative to divisive relationships between the majority and the minority is pluralism. Pluralism implies that various groups in a society have mutual respect for one another's culture, a respect that allows minorities to express their own culture without suffering prejudice or hostility. Whereas the assimilationist or integrationist seeks the elimination of ethnic boundaries, the pluralist believes in maintaining many of them.

There are limits to cultural freedom. A Romanian immigrant to the United States could not expect to avoid learning English and still move up the occupational ladder. To survive, a society must have a consensus among its members on basic ideals, values, and beliefs. Nevertheless, there is still plenty of room for variety. Earlier, fusion was described as $A + B + C \rightarrow D$ and assimilation as $A + B + C \rightarrow A$. Using this same scheme, we can think of pluralism as $A + B + C \rightarrow A + B + C$, where groups coexist in one society (Newman, 1973).

In the United States, cultural pluralism is more an ideal than a reality. Although there are vestiges of cultural pluralism – in the various ethnic neighborhoods in major cities, for instance – the rule has been for subordinate groups to assimilate. Yet as the minority becomes the numerical majority, the ability to live out one's identity becomes a bit easier. African-Americans, Hispanics, and Asian Americans already outnumber whites in nine of the ten largest cities. The trend is toward even greater diversity. Nonetheless, the cost of cultural integrity throughout the nation's history has been high. The various Native American tribes have succeeded to a large extent in

maintaining their heritage, but the price has been bare subsistence on federal reservations.

The most visible controversy about pluralism for ethnic or language minorities is the debate surrounding bilingualism. Bilingualism is the use of two or more languages in places of work or education, with each language being treated as equally legitimate. The passionate debate under way in the United States over bilingualism often acknowledges the large number of people who do not speak English at home. In education, bilingualism has seemed to be one way of helping millions of minority people, who want to learn English to function more efficiently within the United States.

Middleman Minorities

Sociologists and economists have described a subgroup of minorities called middleman minorities. Subject to discrimination and inequality, such middleman minorities do not experience extreme discrimination. Primarily through occupying an economic niche as self-employed entrepreneurs, they provide desired services and sometimes products of use to both the majority and other minorities. Typically, they remain apart from larger industries such as manufacturing and agriculture. More specifically, typical enterprisers include barbers, restaurant owners, tailors, launderers, and jewelers. The enterprises tend to be relatively transportable or easily liquidated, since reprisals from the dominant group are always a possibility.

Based on a dominant–subordinate stratification system, middleman minorities occupy an intermediate rather than low-status position. At different times, middleman included Jews throughout Europe – even at times when the discrimination against them was at their peak such as during World War II they still had great success in some parts of Europe, Chinese throughout Southeast Asia, and Indians in East Africa (Bonacich, 1973; Blalock, 1967).

Broadening the Use of Minorities

Most typically 'minorities' has been used to describe racial, ethnic or national, and religious minorities residing in a nation-state with a clearly defined majority (dominant) group. Since the subjugation of such minorities is well known, writers, both scholarly and otherwise, have found it useful to use the term minorities to describe much different groups. In this manner to say 'minority,' serves as a shorthand to describe subordination, disadvantage, and clearly defined boundaries distinguishing them from the dominant group. The notion of minority in this broader sense has been applied to age groups such as youth or elderly, people with disabilities, exconvicts or exprisoners, and people with mental illnesses to name a few (Farley, 2005; Spiegel and Keith-Spiegel, 1973).

A significant minority that has a visible public presence recently in the political sphere of social policy are those people who seem to stand outside those with a heterosexual orientation. A generation ago, homosexuals or gay and lesbians seem to be a sufficient descriptor. But in recent years, the general public consciousness has been alerted to the complexity of the expression of human sexuality. Hence, minorities can be said to be defined by those who are bisexual or transgendered. Continuing the recognition of the complexity of sexual orientation, some question the use of term bisexual as only serving to reinforce the binary rubric of sexuality and thus contended that a more appropriate term is pansexual to describe people who are sexually attracted to people of all gender identities and biological sexes (Cavendish, 2010).

Women as a Minority

The most common analogy about minorities beyond racial, ethnic, and religious groups used in the social sciences is the similarity between the status of African-Americans and that of women. Blacks are considered a minority group, but one asks, How can women of all groups be so similar in condition? We recognize some similarities in recent history; for example, an entire generation has observed and participated in both the civil rights movement and the women's movement. A background of suffrage campaigns, demonstrations, sit-ins, lengthy court battles, and self-help groups are common to the movement for equal rights for both women and African-Americans. But similarities were recognized long before the recent protests against inequality. In *An American Dilemma*, the famous study of race relations in the United States, Gunnar Myrdal (1944) observed that a parallel to the blacks' role in society was found among women. Others, like Helen Mayer Hacker (1951, 1974), later elaborated on the similarities.

Women are an oppressed group; they are a social minority in the United States and throughout Western society. Men dominate in influence, prestige, and wealth. Women do occupy positions of power, but those who do are the exception, as evidenced by newspaper accounts that declare 'she is the first woman' or 'the only female' to be in a particular position.

Many people, men and women, find it difficult to conceptualize women as a minority. After all, women do not live in ghettos. They no longer have to attend inferior schools. They freely interact and live with their alleged oppressors, men. How, then, are they a minority? Considering the five properties of a minority group, women do experience unequal treatment. Although they are not segregated by residence, they are victims of prejudice and discrimination. Women have physical and cultural characteristics that distinguish them from the dominant group (men). Membership in the subordinate group is involuntary. Through the rise of contemporary feminism, women have become increasingly aware of their subordinate status and have developed a greater sense of group solidarity. While women are not forced to marry endogamous; yet many women feel that their subordinate status is most irrevocably defined within marriage.

Utility of Minority as a Descriptor

Minority is a term that emerged from use of scholars into public discourse as a way to describe inequality between groups of people and not necessarily between specific individuals. Minimally, minority in the United States is taken to mean

minority racial and ethnic groups with federal programs that target 'minorities' such as the Small Business Administration. In this use, minority refers to African-Americans, Asian Americans, Pacific Islanders, American Indians, and Hispanics.

Beyond this grouping, minority includes other ethnic groups such as Jewish Americans, religious minorities such as Muslims in the American context, women, as well as a variety of groups that are disadvantaged in contemporary society. While some decry this broadening of the definition and others question identifying as a minority as some sort of personal weakness, the term minority seems entrenched in the public and academic nomenclature for some time.

See also: Assimilation of Immigrants; Discrimination: Racial; Ethnic Intermarriage; Ethnic, Racial, and Nationalist Movements; Inequality, Social; Prejudice in Society: Psychological Perspectives; Prejudice in Society: Sociological Perspectives; Race Identity; Racial Relations.

Bibliography

Bell, W., 1991. Colonialism and internal colonialism. In: Lachmann, R. (Ed.), The Encyclopedic Dictionary of Sociology, fourth ed. Dushkin Publishing Group, Guilford, CT.

Berndt, R.M., Berndt, C., 1951. From Black to White in South Australia. F.W. Chesire, Melbourne.

Blalock, H.M., 1967. Toward a Theory of Minority-Group Relations. Wiley, New York.

Blauner, R., 1969. Internal colonialism and Ghetto revolt. Social Problems 16, 393–408.

Blauner, R., 1972. Racial Oppression in America. Harper & Row, New York.

Bonacich, E., 1973. A theory of middleman minorities. American Sociological Review 38, 583–594.

Cavendish, 2010. Sex and Society. Marshall Cavendish Reference Books, London.

Farley, John E., 2005. Minority. In: Horowitz, M.C. (Ed.), New Dictionary of the History of Ideas. Gale Publishing, Farmington Hills, MI, pp. 61–1463.

Gordon, M.M., 1964. Assimilation in American Life: The Role of Race, Religion, and National Origins. Oxford University Press, New York.

Hacker, H.M., 1951. Women as a minority group. Social Forces 30, 60–69.

Hacker, H.M., 1974. Women as a minority group: twenty years later. In: Denmark, F. (Ed.), Who Discriminates against Women. Sage Publications, Beverly Hills, CA, pp. 124–134.

Humes, K.R., Jones, N.A., Ramirez, R.R., 2011. Overview of Race and Hispanic Organization. 2010 Census Briefs. C2010 BR-02. U.S. Government Printing Office, Washington, DC.

Kane-Berman, J. (Ed.), 2011. South Africa Survey 2010/2011. South African Institute of Race Relations, Johannesburg.

Massey, D.S., 2012. Reflections on the dimensions of segregation. Social Forces 91, 39–43.

Myrdal, G., 1944. An American Dilemma: The Negro Problem and Modern Democracy. Harper & Row, New York.

Newman, W.M., 1973. American Pluralism: A Study of Minority Groups and Social Theory. Harper & Row, New York.

Nightingale, C.H., 2012. Segregation: A Global History of Divided Cities. University of Chicago Press, Chicago.

Qian, Z., Lichter, D.T., 2011. Changing patterns of interracial marriage in a multiracial society. Journal of Marriage and Family 73, 1065–1084.

Schaefer, R.T., Zellner, W., 2011. Extraordinary Groups, ninth ed. Worth, New York.

Spiegel, D., Keith-Spiegel, P., 1973. Outsiders USA: Original Essays on 24 Outgrouops in American Society. Rinehart Press, San Francisco.

Sumner, W.G., 1906. Folkways. Ginn, New York.

Wagley, C., Harris, M., 1958. Minorities in the New World: Six Case Studies. Columbia University Press, New York.

Relevant Websites

http://ec.europa.eu/justice/discrimination/roma/index_en.htm – European Union and the Roma (ypsies).

http://www.ohchr.org/EN/Issues/Minorities/Pages/internationallaw.aspx – Minorities under International Law (United Nations Office of the Commissioner on Human Rights).

http://www.un.org/en/letsfightracism/minorities.shtml – National, Ethnic, and Linguistic Minorities (United Nations).

http://www.census.gov/econ/sbo/ – Survey of Minority Business Owners (U.S. Census Bureau).

http://www.minorityrights.org/ – World Directory of Minorities and Indigenous Peoples (Minority Rights Group International).

Miracles, Cultural Logic of

Paolo Parigi, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Miracles are not a residual category of science, nor are they the result of ignorance. For millions of people, miracles continue to play an important role in modern life. This article explores the resilience of miracles by considering two key factors: (1) the relationship between the institutions of religion and science, and (2) the power of miracles to mobilize dedicated constituents. The article argues that taken together, these two factors turn miracles into social facts. In the final part of the article, modern sainthood is used as an example to illustrate how the Catholic Church negotiated its relationship with science at the dawn of modernity and how miracles continue to have the power to mobilize people. Some segments of this article were adapted from my book, *The Rationalization of Miracles*, Cambridge University Press.

Although many people might assume that a miracle is merely a residual category of event, destined to be marginalized by such processes of modernity as secularization and scientific progress, for millions of people miracles remain a key element of modern life. Miracles continue to occur abundantly around the world despite advancements in scientific understanding and a decrease in church attendance. Durkheim famously defined a social fact as a "way of acting, fixed or not, capable of exercising on the individual an influence, or an external constraint" (1969: p. 26); social facts are the 'things' that social scientists study. The resilience of supernatural events is important for sociologists and social scientists more broadly because it highlights a core characteristic of miracles: whatever else they may be, they are undeniably social facts.

How did miracles survive the challenges of modernity and manage to become a social fact? One key to this question is the struggle between religious institutions, like Christian churches, and the scientific community over control of the category of miracles and the authority to sanction (or debunk) miraculous claims. Another key is an understanding of the capacity of miracles to mobilize people in a common cause. Indeed, this mobilizing power is the difference between a miracle and a personal fantasy.

Before examining in greater detail how these two factors combine to make miracles a social fact, I offer a definition of miracle as any event that defies common expectations and for which a readily available mental frame exists that denies or eliminates the role of randomness as a cause for the event. A social world without miracles has always been possible, now as 500 years ago. It is a rational world ruled by stochastic processes where random events occur as a matter of routine. The fact that many people do not see the world in this way tells us a lot about our modern societies and, perhaps, about our collective expectations and fears as human beings.

After discussing how the scientific community, religious institutions, and activists all contribute in turning miracles into social facts (see Section The Relationship between Science and Religion and the Creation of New Knowledge), I will examine in greater detail the exemplary case of modern Catholic saints. In particular, I will reconstruct the historical transformation of sainthood at the hands of the Catholic Church during the seventeenth century and how such transformation made miracles compatible with modernity (see Section An Extended Example: Modern Sainthood). Because of the Church's overt attempts to use rules to regulate the production of miracles, modern sainthood is an ideal typical case for studying miracles as social facts in other Christian religions too.

The Relationship between Science and Religion and the Creation of New Knowledge

In the Christian religions, most miracles are health related, thereby making the relationship between religious institutions and medical scientists the focus point of my analysis. A metaphor known as Zeno's Paradox could help illustrate this relationship. The Greek hero Achilles is racing against a turtle. Achilles is 10 times faster than the turtle, so he gives the turtle a head start of 10 spatial units. Will Achilles ever reach the turtle? If the distance the two have to run is infinite, the answer is no, because the turtle will always be 10 units ahead of Achilles, no matter how infinitesimal the units are. If we subscribe to the view that modernity has transformed death from a fundamental aspect of human's life (regulated by rituals and traditions) into an undesirable and 'hidden' outcome of science and medicine, then religious institutions and medical scientists become similar to the turtle and Achilles respectively, running the infinite distance of modern life. Miracles will always occur in the space of what is unpredictable.

If Zeno's Paradox provides a key for understanding why the supernatural remains part of modernity it does not explain 'how' such a process developed in time – the institutions and actors involved in the production of miracles and the historical circumstances that created the modern field of religion. In this article, I argue that the current relationship between modern religious institutions and science (medicine in particular) can be recast as one of control over a boundary: the boundary that separates the possible from the impossible. Modern science failed to displace miracles because Christian churches defended this boundary, claiming authority on the impossible side of the

division. While before the modern period, religion dictated what was to be considered 'officially' a miracle (an event that imitated the life or actions of Jesus, for instance see Vauchez, 2000; Boesch-Gajano, 1999), strong enforcement of the possible/impossible division became the cornerstone on which miracles could continue to occur in the modern period, without any need to define the content of the impossible. Indeed the priest (or pastor) and the scientist are commonly perceived as representatives of opposite and distant worlds. One world, that of science, is predicated on openness, experiments, and results; the other world, that of religion, is predicated on mysteries and dogmas.

Nevertheless, several social scientists have shown how categories evolve over time and boundaries become porous. Hannan and Freeman, for instance, considered several social processes, such as personnel turnover and technical errors, which usually lead to the blending of opposite categories (1989). Rao, Monin, and Durand considered a successful case of such blending in the realm of French gastronomy (2005). Indeed one could interpret certain aspects of positivism during the eighteenth century – the intuitions that led Auguste Comte to invent the word 'sociology,' for example, or Saint Simon's vision of a technocratic society (2009) – as a blending between science and religion. Had such a blending been successful, my argument suggests, miracles would have disappeared from the landscape of religious life. Why did the boundary of the possible/impossible remain salient during modernity?

One reason that this boundary did not become porous at the end of the nineteenth century was that religious institutions throughout the Christian world continued to train their own personnel quite separately from the state-controlled training of scientists and other professionals. Maintaining control over the training of the clergy was (and is!) important because it produced a specific knowledge about the realm of the impossible that justified the existence of a rigid boundary. In his historical analysis of the relationship between mental illness and civilization, Michel Foucault argued that once the mentally ill started to be moved into institutions, they fell under the scrutiny of a power that began distinguishing between different types of psychological malady – alcoholism, depression, etc. Different types of mental illness became the objects of analysis of a new discipline, psychology (Foucault, 2001). Similarly, different types of impossible events came under the scrutiny of the clergy, which transformed some of them into miracles. As with the case of psychology, power – in this case, the power of religious institutions expressed through the training of their personnel – generated knowledge – in this case, a knowledge of the impossible.

No other church extended this process of knowledge creation further than the Catholic Church. In the aftermath of the Protestant Schism, Rome embarked on a deep reorganization of its centennial structure and created a specific commission to investigate miracles performed by candidates for sainthood and by objects such as statues, icons, and paintings. While the latter part of this article will provide more historical details about this fascinating and not widely known period of Church history, it is worth pointing out that Rome's strategy as ultimate arbiter of the impossible did more than defend a boundary line. In effect, the Catholic Church achieved control over the boundary. This is illustrated by the simple fact that Catholics throughout the world consider true miracles not just those events that defy scientific expectations but also those that have received approval from the Pope. That is, for Catholics, not every impossible event is *ipso facto* a miracle.

In light of all of this, the relationship between modern religious institutions and science is not one of progressive domination of the latter over the former. This simplistic perspective, perhaps better captured with Francis Bacon's sentence that "knowledge is power," ignores that science and religion are two separate and autonomous systems of beliefs, practices, values, and rules – and that power, too, is capable of creating knowledge. Only by restoring the complexity of the social world in which scientists and priests live together, it is possible to start understanding the resilience of miracles.

The Mobilizing Power of Miracles

In the modern world, most supernatural events take place privately between a single worshipper and his or her deity. This type of miracle is private in the sense that it does not involve others beyond the individual and, potentially, his or her immediate family. Private miracles do not have mobilizing power and are a matter of personal beliefs, i.e., they are not social facts. A subset of miracles are, however, public, in the sense that they occur at pilgrimage destinations or are used for establishing someone's sanctity by the Catholic Church. Still other times, miracles are public because they occur in public spaces in front of several people. Many social scientists tend to think that public miracles are the result of crowd behavior or are spontaneous eruptions of faith and are, in this sense, comparable to private miracles. This way of thinking about miracles has led social scientists to in large part ignore them as objects of study. However, in reality, public miracles more closely resemble mobilizing events of contemporary social movements, and are worth studying as organized expressions of religious activism.

Indeed, behind each public miracle – be that a healing performed by a powerful relic of a soon-to-be-saint or the crying of a religious icon in front of hundreds of worshippers – there is always a group of activists that first catalyzed the conditions for the occurrence of the miracle and then organized the local support generated by the miracle so that more miracles could occur. This group of activists could be made of early acolytes of a recently deceased candidate for canonization, or it might comprise the clergy of a potential pilgrimage destination. Regardless, activists benefit directly from the occurrence of the miracle and stand to further benefit from the occurrence of more supernatural events.

This is not to suggest that the religious motivations of activists are of secondary importance. Rather, it serves as a reminder that the mobilization of a community via miracles also produces worldly advantages for some people. Pilgrimage is one example. A steady stream of pilgrims generates revenue that can be channeled either to a local church or a religious order. But while the economics of a miracle can be a boon for activists, they do not merely reap the reward. The activists' efforts to organize the miracle are actually essential for the miracle to exist as a social fact, and for a subtle reason – the

participation of activists makes public miracles compatible with a cultural and social environment dominated by skepticism toward supernatural events.

Religious activists catalyze the occurrence of miracles through the use of ritual. In the case of public miracles performed by a Catholic saint, for instance, they occur via the intercession of a relic that it is usually in the control of the activists. The ritual performs the function of expelling chance from the equation and ensuring that everything that happens as a result of it is the result of God's will. The ritual creates a frame that makes an unpredictable event predictable, and therefore reproducible. By showing that the event obeys God's laws, the ritual creates a legitimate suspension of natural laws (or of medical knowledge) so that, for example, the places where these types of miracles occur are *de facto* places for supernatural activity on Earth.

Using rituals in order to make miracles align with modernity represents an extension of Max Weber's thesis on the rationalization of religion (1922). The mobilizing power of miracles shows that, in fact, the rationalization of religious beliefs did not mean only the exclusion of magic from religion, i.e., Puritanism. The path traveled in Northern Europe to make Christian religious beliefs compatible with modernity was just one possible solution to the tensions created by scientific progress and medicine. Another solution was to leave some magic into the religious core of Christianity. What this article argues is that the rationalization of religion is better seen as a historical and institutional process that involves activists legitimizing and structuring what can be considered genuinely supernatural, i.e., a miracle that occurred as the effect of a ritual (the cause), according to their own interests and ideology.

An Extended Example: Modern Sainthood

Miracles are part and parcel of the modern world and are a key aspect of Catholicism. Sainthood would not be *necessary* without miracles, i.e., without the widespread belief in miraculous events and the cultural understanding that such events are the work of holy individuals after their death (*postmortem*), because they sit in paradise and are thus able to intercede for us with God. The story of modern sainthood shows to the maximum extent how the Church constructed a specific knowledge of the impossible in contrast to the discoveries of science, and how religious supporters of future candidates for sainthood used (public) miracles to mobilize local communities in support of the canonization of their leaders. The institutional process that made it possible for modern saints to continue performing (public) miracles thus serves as an illustration of how social facts are created. In what follows, I will provide some historical detail of the reorganization of the Catholic Church in the aftermath of the Protestant schism. My attention will focus first on sainthood during the medieval period and then on the special commission that Rome created with the intention of reforming it.

Sainthood before the Protestant Schism

In the Middle Ages the ascetic, healing saints, as typified by Saint Francis of Assisi, were prevalent in Italy and had a long history going back to the late Roman Empire (Brown, 1981), while the noble kinship saints, such as Saint Louis, were dominant in France and were a more recent version of sainthood (Vauchez, 2000, 1989). For the most part, the popes canonized candidates that conformed to the Italian model of sainthood.

Models of sainthood did not vary only by region (Duby, 1980). Because the popes did not have complete control over matters of sainthood until the seventeenth century, Boesch-Gajano and Modica suggest distinguishing types of sanctity based on the authorities that certified them. *High* sainthood in this model coincides with official sainthood, one that received the approval of the pontiffs (2000). *Low* sainthood, on the other hand, refers to sanctity achieved through popular acclaim and certified by local authorities. The case of a dog worshipped as a saint in the village of Villars-les-Dombles in the Ain region of Southern France is a notorious example of this latter type of sanctity. The dog, Guinefort, was (incorrectly, it turned out) accused of having killed a baby and was thrown in a well and killed. After discovering the dog's innocence, the population of the village began to worship the dog, which from that point on (and up to the eighteenth century) was known as Saint Guinefort. Vauchez explains: "In the contrast to … harsh punishment and its patent injustice, popular piety was born. This popular emotion is the same which, transposed on the religious sphere, generated devotion. By virtue of a process that we can treat as a 'law of popular affectivity,' popular piety generates religious piety" (1989: p. 101).

Nevertheless, official sainthood did require the support of people, what the Church called the *Fama Sanctitatis*, in addition to papal approval of a candidate's sanctity. In the Middle Ages, the support a saint could attract was based largely on his miracles; hence, miracles were a vital instrument for gathering the popular support necessary for canonization. Organizing this support through the use of rituals was not that fundamental, however. Summarizing the field of sainthood before the Protestant schism we can say that while *low* saints had popular support, *high* saints had prestige.

During the reformation, Protestants attacked the Church's indulgence to popular devotion of the kind that produced the dog Saint Guinefort. At the same time, the great rate of scientific discoveries popularized via the printing press contributed to produce a shift in climate, whereby supernatural events began to be greeted with increasing skepticism. Thus, at the dawn of modernity, the Church was faced with the dual challenge of promoting its own version of sainthood while attracting popular support. To a certain degree, the two goals had the potential to undermine one another: allowing popular devotion was important in ensuring popular support, but allowing too great a variety of popular devotion would seem to prove skeptics right because it could potentially turn any popular mystery into a holy act. The establishment of a central commission to investigate claims to sainthood was part of the Church's effort to reshape the relationship between popularly perceived sainthood and its own official version.

But *how* did the new commission effect this reshaping? It took considerable organizational skill to steer the environment from one in which putative saints faced lax scrutiny and were proclaimed locally to one in which Rome asserted close control over sainthood, all the while maintaining popular support.

The Church risked alienating its followers in a period when the legitimacy of its dogma was much in question and when religious wars challenged the political authority of Rome.

The Special Commission for Investigating Miracles and Saints

Pope Sixtus V created the *Congregatio Sacrorum Rituum*, or Congregation of Sacred Rites, with the 22 January 1588 encyclical *Immensa Aeterni Dei*. The *Congregatio* was originally charged with organizing and moderating the Latin liturgy of the Church as well as investigating whether individuals who died with reputations of extreme holiness were indeed saints. The latter function quickly took precedence, so that the work of the *Congregatio* resembled more and more that of a tribunal that 'tried' candidates for sainthood after their deaths (Papa, 2001).

The establishment of the *Congregatio* was the culmination of a long process that began on 9 February 1993, at the Church of Saint John in Rome, when the pope officially sanctioned the worship of the bishop of Augsbourg, Saint Udalric. This was Rome's first documented canonization. Interestingly, the *bolla*, i.e., the Pontiff's communication, issued for the occasion described the 20 miracles that Udalric had performed. Similar canonization bulls listing miracles continue to be issued to this day (Papa, 1988).

Despite this early example, the Roman Curia and the papacy had very little control over granting the title of saint. Throughout the Middle Ages, Catholic Europe was divided in its devotion to a myriad of local saints. This state of affairs persisted even though, in order to prevent abuse of the title and to institute some consistency, Pope Alexander III officially took the power of canonization out of the hands of local bishops in 1170 with the decree *Audivimus, de Reliquiis et Veneratione Sanctorum*. But the decree was not enforced, partly because of the difficulty of controlling a territory as vast as Europe and partly because it made little organizational sense to risk alienating popular support. The practice of worshipping local saints continued.

However, in the new social landscape of the Counter-Reformation, the saints and their miracles were a matter of particular importance, and a subject about which Protestants and Catholics fought fiercely (Thomas, 1971). In this environment, Rome had a direct interest – for the sake of its own credibility – in assuring the quality of the people worshipped as saints. As Rome began to insist more strongly on the importance of its approval, local religious activists found an interest in securing Rome's recognition of their candidate, in order to enhance their local reputation. The two demands, one from the top and the other from the bottom, met in the procedures of the *Congregatio*, which transformed sainthood into a standardized legal procedure (Gotor, 2000, 2004).

The canonized sainthood that the *Congregatio* instituted was based on a series of trials in which witnesses were called to testify on the miracles and virtues of the candidate. Each trial was divided into sections, and witnesses were called for each. The sections were (1) the early life of the candidate; (2) the sanctity of the candidate in a broad sense; (3) the virtue of faith; (4) the virtue of hope; (5) the virtue of charity toward God; (6) the virtue of charity toward spiritual life; (7) the virtue of charity toward temporal life; (8) the virtue of prudence; (9) the virtue of justice; (10) the virtue of patience (only for French candidates); (11) the virtue of strength; (12) the virtue of the religious life, if the candidate belonged to a religious order; (13) the supernatural gifts of the candidate; (14) the *Fama Sanctitatis*, or saintly reputation of the candidate; (15) the death of the candidate; (17) the candidate's fame after death; and (18) the miracles.

Local church authorities opened the canonization procedure by collecting evidence on the recently deceased holy person and commissioning the writing of *vitae*, or biographies. This first phase was called the *Processo Ordinario*, because the authority that conducted the investigations was ordinary, or locally based. The evidence gathered in the *Processo Ordinario* was sent to the *Congregatio* in Rome, where it received its first scrutiny, based on the way the trial was conducted – whether all the necessary authorities were present, whether the depositions were sealed and the minutes recorded, etc. Approval in this phase meant that a new trial would be held, this time by the apostolic authorities; therefore the new trial was thus called the *Processo Apostolico*. The judges in this trial were all selected by Rome. If the *Congregatio* reached a positive opinion in the *Processo Apostolico*, it issued the *Lettere Remissoriali*, or permission to gather evidence on specific details (*in partibus*) of the candidate's life and the miracles he performed. Since many candidates were active in more than one parish during their lives, both the ordinary and apostolic trials were held in multiple locations, in effect multiplying the number of required trials (Veraja, 1988).

The trials *in partibus*, often called *Remissoriali*, were conducted by the Roman authorities. Until the decrees of the period 1634–42, the *Congregatio* entrusted a special commission, the *Uditori di Rota*, with the task of organizing most of the work related to these trials. The *Uditori di Rota*, of whom there were three, were the pope's trustworthy advisors. All of them held important positions in the Roman Curia; they were usually cardinals and they were often members of the *Congregatio*. Their predominance in the first decades of the life of the *Congregatio* shows that the pope intended to maintain close supervision over matters of sainthood (Papa, 2001). The *relatio*, or summary of the trials, that the *Uditori* wrote for the *Congregatio* and for the pope was the single most important document in a canonization process during this period. This summary of the life of the candidate was divided into two parts: the first addressing the virtues of the candidate, and the second his miracles. The miracles discussed in the *relatio* were a selection of all those collected during the trials.

The process of developing modern sainthood was a precarious one that required balancing conflicting interests. This meant that the *Congregatio* refined and changed its rules many times over the years between its founding and the reforms of the period 1634–42. At the beginning of the seventeenth century, the *Congregatio* instituted one of its first new rules by creating the title of *beato* (blessed). This title had existed informally since late antiquity, indicating a cult that the Roman authorities had approved for the purpose of local worship only. After the *Congregatio's* rule change, the title of blessed became in practice an intermediate step to sainthood. Local authorities carried out the investigations for the title, which allowed them to assert some autonomy from the Roman Curia. Thus, the institution of the title *blessed* mediated between Rome's desire to assume direct control over sainthood and the prerogatives of

local communities that had chosen their own saints for centuries (Gotor, 2004).

The reforms of 1634–42 introduced a new figure to the canonization process that drastically reduced the role of the cardinals of the *Congregatio* and completely replaced the *Uditori*. The *Promotore Fidei* – a 'devil's advocate' with respect to the candidate's position – was charged with the task of dismantling any false claims about the powers and virtues of the candidate. The *Promotore Fidei* was a lawyer entrusted with representing the interests of the Church on matters of sainthood, and he usually represented the Church several times during his career. The use of legal experts on religious matters had been a long-standing tradition in the Church, but the extent of their involvement in canonization procedures was one of *Congregatio*'s many institutional innovations. The first attorney nominated to defend the interests of Rome during a canonization trial was Giovanni Giacomo Nerotti in the case of Raymond de Penafort, a Spanish saint worshipped since the thirteenth century. However, Nerotti's task remained confined to the traditional role of the Curia's attorney – i.e., establishing that local trials followed proper procedures. Nerotti and his counterpart on de Penafort's behalf, the Dominican priest Michele Llot, did not address the merits of de Penafort's purported miracles.

With the expansion of scientific and medical knowledge during the seventeenth century, the *Promotore*'s role also expanded beyond procedural supervision to resemble that of a defense attorney defending the integrity of the Church poking holes in the evidence offered by the prosecutor of the case, i.e., the lawyer for the candidate. This required the creation of a legal-theological framework for handling medical knowledge. With the *Promotore's* changing role came a new title – *Promotore Fidei*, or protector of the faith. While several documents from the first decades of the seventeenth century mention the new title, the job itself was not officially created until the reforms of 1634–42. Key to the *Promotore Fidei's* job was authoring opinions (*animadversiones*) about the candidate's miracles. These documents, which outlined the *Promotore*'s doubts about the miracles, would circulate among the members of the *Congregatio* and the pope.

In describing his 'doubts,' the *Promotore* used the best medical authority accessible to him in order to poke holes in the evidence collected about miracles by the candidate's attorney, the *Postulatore*. The *Promotore*, i.e., operated similarly to the devil's advocate and it is likely that it is this institutional arrangement that popularized the expression 'devil's advocate.' With time, the expertise the *Promotore* accumulated in the process of debating cases helped the *Congregatio* produce new and more refined knowledge about miracles. For example, it was in this period that the notion took hold that a true miracle required that healing must occur instantly and completely (Antonelli, 1962). The *Congregatio* incorporated this idea into an established classification of miracles by degrees that dated to Saint Augustine. This produced a new, more refined list of the characteristics of a true miracle; knowledge of and the ability to judge those characteristics belonged to the Church, not doctors. In this way, the *Congregatio* was able to create a legal framework to control the boundaries of the supernatural and to protect it from the potential further claims of medicine. The new institutional figure of the *Promotore* gave to the process of canonization its lasting form.

The Mobilizing Power of Miracles

Immediately after the candidate's departure for paradise (the *dies natalis*, in Church terminology), his acolytes were faced with the problem of how to ensure that their leader's message of religious reform could continue. In the absence of the deceased leader's personal charisma the leader's message would only continue to propagate if new activists were brought to the cause of the candidate's sanctity. The involvement of newcomers required that acolytes mobilize a community and this mobilization translated directly into creating the conditions for the occurrence of more miracles. However, since the candidate was not there in the flesh, the acolytes had to be careful to connect the occurrence of the supernatural event to the efforts of the candidate who, sitting close by God in paradise, was interceding for the sick person. Acolytes had to exercise skill in building, and controlling, the networks that generated miracles on behalf of their candidate (Parigi, 2012).

The days immediately after the candidate's death marked a special time in the community where the candidate and his acolytes had been active. It was during this period that the acolytes could potentially capitalize on their skills as activists, because opportunities for miracles multiplied in the days following the *dies natalis*. If the candidate were truly a saint, the general population expected to see *postmortem* miracles, which of course could only occur after the *dies natalis*. The days immediately after the candidate's death were also the time when it was the easiest to create relics of the saint, which would be the main vehicle of those *postmortem* miracles. Relics would prove the involvement of the candidate in *postmortem* miracles, and the candidate's acolytes would hope to control them in order build their network and create more miracles.

The public's interest in relics came from the deeply held belief that relics had healing powers. For instance, when friar Rainiero, a candidate for sainthood who lived in central Italy, died on 23 August 1589, his acolyte and fellow friar Egidio da Amelia testified that a mob rushed to the body that lay inside the church and began quite literally to tear him apart – taking his hair and pieces of his clothes and punching holes into his body so that tissues could be soaked in his blood (ASV 3238; 3241; ACS *Storico*, 223). A similar scene happened when friar Pasqual Baylon, another candidate that lived in the same period as Rainiero, died in Villa Real, Spain (ASV 3393). Egidio's testimony exemplifies the fact that public fervor was directed mostly toward pieces of the candidate's body, the most common kind of relics. It would be a grave mistake to think that only the lower strata of society believed in their power. Philip II, the king of Spain from 1556 to 1598, amassed a collection of more than 100 relics in the palace-monastery Escorial (Parigi, 2012).

But generally a mob did not tear the body of a candidate to pieces. The acolytes had direct access to relics, and so they were called upon to attend to the sick. Over time, a ritual developed. Upon arriving at the sick person's house, the acolyte would invite everyone present to pray, often instructing them to kneel and pray aloud together. During the prayer, the acolyte asked

his candidate for grace for the sick person and at this point usually took out one of the candidate's relics, which he applied to the invalid. The other people present would still be kneeling in prayer. The acolyte would then leave the scene and the other people would notice a sudden improvement in the invalid's condition. Immediately or within the span of a few days, the sick person would be perfectly healed, having become a 'saved.' (This describes most *postmortem* healings. It is not, however, what Church doctrine required of genuine miracles. As previously mentioned, the Church mandated an instantaneous healing that in reality seldom occurred.)

Successfully organizing public fervor into a community mobilization therefore meant maintaining strict control over the candidate's relics. Acolytes organized their control of the relics through the creation of the ritual healing. The ritual helped construct a predictable mental frame – based on popular beliefs about how nature and society worked – around (unpredictable) illnesses and thereby 'produced' healings. As an expression of a strict set of guidelines, a ritual reduced uncertainty by removing chance from the equation and instead offered evidence that all forces obeyed God's laws. A healing was a patterned process in time in which the behavior of each actor involved was symbolic (Turner, 1967). Acolytes catalyzed *postmortem* miracles through their role as performers of a ritual, at whose center stood the candidate or a piece of his body.

Successful acolytes were those who knitted a community together using miracles. This was not a simple task and required considerable organizational skills. The protosocial movement that acolytes organized would then become the main force supporting the canonization efforts. Why would acolytes want to secure the *Congregatio*'s approval? The simple answer is that they faced competition from other religious activists, and such competition created havoc in many local communities. For instance, a priest in a small village near Milan was exasperated by all the attention attracted by a crying image of the Virgin Mary in a nearby parish. During the trial held by the Church authority on the case, he told the judges explicitly that his constituency was diverting prayers and money to the other parish, creating problems for him and his Church (Sangalli, 1993). Competition was particularly fierce where supernatural results were perceived to be produced by a candidate's relics. For example, the internal organs of Filippo Neri (a candidate to sainthood that lived in the sixteenth century) were secretly taken from Rome and brought to Naples, where they ended up in the hands of a noblewoman (*Il primo processo*). Similarly, the head of Bernardino of Santa Lucia, another candidate to sainthood that lived in the same period, was stolen from the Franciscan Church in Agrigento, Italy, where the whole body lay exposed (ASV 2209). A list of such anecdotal examples could go on at length.

Yet, for the candidates and their supporters receiving Rome's seal of approval, local competition dissipated. A saint performs miracles as a matter of fact and, therefore, candidates that became saints attracted donations, and often pilgrims, for the local church that housed the body of a saint or some of his relics. Furthermore, Rome's approval could mean the creation of a new religious order, with offices and often a complex bureaucracy, at the disposal of the candidate's supporters. For local activists, the path to canonization was also the path to progressively reducing local competition. This convergence of interests between the goal of the *Congregatio* to create a more modern version of sainthood and the goal of local activists to reduce competition transformed *postmortem* miracles into social facts within the communities that harbored them.

Conclusion

In the previous section, I used the historical case of how the Catholic Church was able to transform sainthood to illustrate a key characteristic of miracles in the modern world, that of being social facts. If the case of the Catholic Church is thus useful for explaining the resilience of miracles to the challenges posed by scientific progress and medicine, it is also interesting in its own right as an example of how an institutional field can be renewed from the center and without conflict (Schneiberg and Lounsbury, 2008). Following Pierre Bourdieu, I see a field as a constellation of objective relationships that structure the interactions between given historical actors (2012). The rules of modern sainthood that I documented above, brought to existence a new actor vis-à-vis the Church and the local authorities – the acolytes, i.e., organized group of believers mobilizing local communities in support of their (defunct) leader.

Fervent believers have always existed in the Christian religion, at least since the arrival of Saint Paul to the shore of Rome, but the organization of their fervor into something more similar to a social movement began with the new rules that the *Congregatio* developed in the aftermath of the Protestant schism. Acolytes, i.e., acquired a new structural role – and with that a system of relationships – that made them an essential part of the field of modern sainthood. A field perspective therefore can help explaining not just the endless proliferation of religious orders within Catholicism but also a key reason for why no other major schisms has occurred within the Church since the sixteenth century: the Church created a position for activists within its religious field capable of incorporating the interests of local actors. The legitimacy of modern sainthood has its roots in the matching between subjective beliefs and objective structure (Bourdieu, 1971).

The new rules that the Church developed allowed for the expressions of different interests and generated a mutually beneficial exchange between central officials, local authorities, and the acolytes. While this exchange emerged through lots of happenstance and conflict (Parigi, 2012), it is interesting to note that it represents a case of institutional renovation from the core of a field. My analysis of modern sainthood highlights a unique path toward institutional change that occurred not only endogenously (Rao et al., 2003) but also from the impulse of the central actor operating in the field, i.e., the Church. Change occurred because the procedures that the Church created were able to integrate the interests of all different actors operating in the field.

Primary Sources

ASV = Vatican Secret Archive.
ACS = Archive of the Congregation for the Causes of the Saints.

Pasqual Baylon

ASV 3393

Filippo Neri

Giovanni Incisa della Rocchetta, Nello Vian, eds. 1957. *Il primo processo per San Filippo* Neri. vols. 1-4. Città del Vaticano: Biblioteca Apostolica Vaticana

Rainiero of Borgo San Sepolcro

ASV 3238; 3241; ACS Storico, 223.

See also: Catholicism; Collective Identity; Culture and Institutional Logics; Culture, Cognition and Embodiment; Embodiment and Culture; Religious Belief, Evolution of.

Bibliography

Antonelli, F., 1962. De inquisitione medico-legali super miraculis in causis beatificationis et canonizationis. Pontificium Athenaeum Anthonianum, Roma.
Boesch-Gajano, S., 1999. La santità. La Terza & Figli, Roma-Bari.
Boeach-Gajano, S., Modica, M. (Eds.), 2000. Miracoli. Dai segni alla storia. Laterza & Figli, Roma-Bari.
Bourdieu, P., 1971. Une interprétation de la théorie de la religion selon Max Weber. Archives Européennes de Sociologie 12 (1).
Bourdieu, P., 2012. In: Alciati, R., Urciuoli, E.R. (Eds.), Il campo religioso. Accademia University Press, Torino.
Brown, P., 1981. The Cult of Saints. Chicago University Press, Chicago.
Comte, A., 2009. A General View of Positivism. Cambridge University Press, Cambridge, MA.
Duby, G., 1980. The Three Orders. Feudal Society Imagined. University of Chicago Press, Chicago.
Emile, D., 1969. Le regole del metodosociologico. EdizioniComunita, Milano.
Foucault, M., 2001. Madness and Civilization: A History of Insanity in the Age of Reason. Routledge, London.
Gotor, M., 2000. La fabbrica dei santi: la riforma urbaniana e il modello tridentino. In: Fiorani, L., Prosperi, A. (Eds.), Storia D'Italia. Einaudi Editore, Roma, pp. 679–727.
Gotor, M., 2004. Chiesa e santità nell'Italia moderna. Editori Laterza, Roma.
Hannan, M., Freeman, J., 1989. Organizational Ecology. Harvard University Press, Cambridge, MA.
Papa, G., 1988. La Sacra Congregazione dei Riti nel primo periodo di attivita. In: Miscellanea in occasione del IV centenario della Congregazione per le cause dei santi (1588–1988). Congregazione per le cause dei santi, Città del Vaticano, pp. 13–52.
Papa, G., 2001. Le cause di canonizzazione nel primo perido della Congregazione dei Riti (1588–1634). Urbaniana University Press, Città del Vaticano.
Parigi, P., 2012. The Rationalization of Miracles. Cambridge University Press, New York, NY.
Rao, H., Monin, P., Radolphe, D., 2003. Institutional change in Toque Ville. The American Journal of Sociology 108 (4), 795–843.
Rao, H., Monin, C.C., Durand, R., 2005. Border crossing: bricolage and the erosion of categorical boundaries in French gastronomy. American Sociological Review 70, 968.
Sangalli, M., 1993. I miracoli a Milano. I processi informativi per eventi miracolosi nel milanese in età spagnola. NED, Milano.
Schneiberg, M., Lounsbury, M., 2008. Social movements and institutional analysis. In: Greenwood, Royston, Oliver, Christine, Sahlin-Andersson, Kerstin, Suddaby, Roy (Eds.), The Sage Handbook of Organizational Institutionalism. Sage Publications, London, pp. 650–672.
Thomas, K., 1971. Religion and the Decline of Magic. Oxford University Press, New York.
Turner, V.W., 1967. The Forest of Symbols. Cornell University Press, Ithaca, NY.
Vauchez, A., 1989. La santità nel Medioevo. Il Mulino, Bologna.
Vauchez, A., 2000. Santi, profeti e visionari. Il soprannaturale nel medioevo. Il Mulino, Bologna.
Veraja, F., 1988. Miscellanea in occasione del quarto centenario della Congregazione per le Cause dei Santi. Libreria Editrice Vaticana, Città del Vaticano.
Weber, M., 1922. The Protestant Ethic and the Spirit of Capitalism. Roxbury Publishing Co., Los Angeles.

Mirror Neurons, Theory of

Michael C Corballis, School of Psychology, University of Auckland, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mirror neurons, first identified in the monkey brain, respond both when an animal makes a hand movement and when it observes another individual making the same movement. A corresponding but more elaborate system in the human brain has emerged to incorporate such social functions as language, empathy, theory of mind, and cultural learning. The roles of mirror neurons suggest that the mind functions as an embodied system, grounded in real-world simulations, rather than a system based on the manipulation of abstract symbols.

Introduction

> I predict that mirror neurons will do for psychology what DNA did for biology.
>
> V.S. Ramachandran (2000)

This bold prediction has had some impact; typing 'mirror neurons' into Google produces some 2.7 million hits. Among Ramachandran's predictions was that mirror neurons would be seen to underlie the various surges in technology and invention that have created the distinctive human mind – the invention of tools, the 'great leap forward' in cultural and technological innovation of 40 000 years ago, the emergence of language and mathematics, not to mention nuclear power, automobiles, air travel, and even space travel. Mirror neurons were first discovered, though, in the brains of monkeys.

Work undertaken in the 1980s by Giacomo Rizzolatti and colleagues in Parma, Italy, showed that single neurons in the prefrontal cortex of the macaque were specialized for coordinated movements of the hand, such as those involved in grasping an object (e.g., Rizzolatti et al., 1988). They then discovered serendipitously that some of these neurons also responded when the monkeys watched another individual making the same movements (Gallese et al., 1996). These neurons were dubbed 'mirror neurons,' because they seemed to map observed actions on the production of the same actions, as though the monkey perceives an action in terms of how it would itself perform the same action. The activity in mirror neurons does not depend on specific muscular movements; this was elegantly demonstrated by training monkeys to grasp objects using two kinds of pliers, one operating normally, and the other operating in reversed fashion, so that squeezing the arms opened the pliers rather than closing them. The same mirror neurons fired regardless of which pliers were used to grasp an object, even though the hand movements themselves were reversed (Umiltà et al., 2008). Mirror neurons, therefore, seem to be tuned to coordinated, intentional movements, whether perceived or produced, defined in terms of their outcome rather than their muscular components.

These properties seemed to bestow on single neurons almost magical computational powers, but it is now known that mirror neurons are part of an extensive circuit. In the macaque brain, this 'mirror neuron system' (MNS) includes area F5 in the prefrontal cortex, where mirror neurons were first discovered, along with two areas in the parietal lobe, which also contain neurons with mirror properties. The parieto-frontal circuit also receives high-order visual information from the superior temporal sulcus (STS) and inferior temporal lobe, but the neurons in these areas do not have motor properties, and the circuit is under the control of two further areas in the frontal lobes, the presupplementary motor area and the ventral prefrontal cortex (Rizzolatti and Sinigaglia, 2010). It has also been proposed that a similar MNS underlies the perception of emotion (Gallese et al., 2004). The mirroring of perception onto action or emotion is, therefore, a function of complex circuits, and not just of individual neurons.

Although the MNS was originally mapped in the macaque brain, it is now established that homologous systems exists in the human brain. A meta-analysis of 125 studies using functional magnetic resonance imaging (fMRI) revealed 14 clusters of neurons with mirror properties. These were located in areas homologous to those identified in the monkey brain, including the inferior parietal lobule, inferior frontal gyrus, and ventral prefrontal cortex, but also included regions in the primary visual cortex, cerebellum, and the limbic system (Molenberghs et al., 2012). Evidence for individual mirror neurons was also obtained from patients suffering from epilepsy, who had multiple electrodes implanted in the medial frontal and temporal lobes in order to identify potential seizure foci. The patients were asked to execute or observe grasping movements of the hand or facial expressions of emotion, and neurons with mirror properties were identified in the hippocampus and adjoining regions, as well as in the supplementary motor area (Mukamel et al., 2010). It appears that the human brain is indeed well endowed with mirror neurons, perhaps more so than the monkey brain.

The discovery of mirror neurons attracted a flurry of attention in fields as diverse as social psychology, developmental psychology, and linguistics, seeming to explain such phenomena as imitation, empathy, altruism, music, theory of mind (ToM), autism, and the perception of speech, not to mention the contortions my body undergoes when I watch a game of rugby on my TV screen. One problem is that many of these phenomena apply to humans, but not to macaques, where the MNS was first identified. For instance, the very designation 'mirror' implies imitation, but macaques appear incapable of imitation (Visalberghi and Fragaszy, 2001).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23204-2

Rizzolatti and Craighero (2004) therefore concluded that "the primary function of mirror neurons cannot be action imitation" (p. 172). Later, though, I discuss evidence that the MNS in humans may play a critical role in imitative learning, especially in social contexts.

A more delimited view is that the MNS underlies action understanding. That is, when an observer watches another individual performing an action, the mechanisms for performing that action are activated in the observer at a cortical level, so that the observer then 'understands' that action in terms of how he or she would perform it (e.g., Rizzolatti and Craighero, 2004). This notion too has proven elusive. For instance, mirror neurons appear to be activated in the human brain by meaningless gestures (Iacoboni et al., 1999), raising questions as to whether such actions can be said to be 'understood.' The MNS is also activated when people watch human lip movements or monkey lip-smacks, but not when they watch a dog barking (Buccino et al., 2004), yet there can be little doubt that the barking was recognized and understood. One possibility is that the mirror response is not a necessary component of understanding, but can augment and enrich it. Hickok (2009) gives the example of watching a person playing a saxophone. A nonplaying viewer would certainly understand the action as one of playing the saxophone, but a viewer who was herself an expert might generate internal motor representations enabling her to predict future movements, and determine the level of expertise displayed by the player.

In a further refinement, Rizzolatti and Sinigaglia (2010) argue that the fronto-parietal MNS is specifically dedicated to the understanding of the actions of others 'from the inside,' giving the observer "a first-person grasp of the motor goals and intentions of other individuals" (p. 264). Precisely what this might mean, though, remains to be more fully explored.

The extent to which the properties of mirror neurons revealed from research on monkeys can be applied to functions considered distinctively human is uncertain: Ramachandran's famous remark may be overly optimistic, but need not be totally misguided. The MNS may well have provided platforms for the emergence of human capacities not evident in nonhuman primates. One example is language.

Mirror Neurons and Language

It is widely agreed that language, as a generative, grammatical system, is unique to humans. Some have argued, in fact, that it emerged as a singular event, perhaps due to a fortuitous mutation, in our own species within the past 100 000 years (e.g., Chomsky, 2010) – a theory that smacks of the miraculous. The discovery of mirror neurons in the macaque brain has led to more Darwinian accounts, in which the MNS provided the basis for progressive changes leading to the emergence of speech, and to language as a grammatical system (e.g., Arbib, 2005; Corballis, 2002).

Motor Theory of Speech Perception

As Rizzolatti and Arbib (1998) pointed out, the mapping of perception onto execution was long anticipated by the motor theory of speech perception (Liberman et al., 1967), which holds that speech is understood in terms of its articulation, rather than in terms of acoustic analysis. Mirror neurons in the macaque respond to manual actions, such tearing paper or dropping a stick, whether these actions are seen or heard, suggesting that the MNS does receive auditory input. Critically, though, mirror neurons do not receive input from vocal signals themselves (Kohler et al., 2002). The incorporation of vocal sounds into the MNS probably emerged late in primate evolution, and is perhaps unique to humans (Jürgens, 2002).

As Hickok (2009) has pointed out, though, the motor theory cannot fully explain how we perceive speech. People with production aphasia, or the acquired inability to produce speech, typically retain the ability to understand speech, and some nonhuman species, including chimpanzees (Savage-Rumbaugh et al., 1998), and even dogs (Kaminski et al., 2006; Pilley and Reid, 2011), have been shown to respond to words and simple commands spoken by humans despite being quite unable to produce anything resembling speech sounds. Hickok and Poeppel (2007) suggest a dual-stream model of processing speech. A ventral stream, involving superior and medial temporal lobe, is responsible for the understanding of speech, while a dorsal stream, operating through the posterior and dorsal temporal lobe and the parietal operculum, maps speech sounds on to articulatory representations in the frontal lobe. It is this dorsal stream that is part of the MNS, critical in learning to speak, or in tuning production of speech to what we hear. It may play a role in the perception of speech itself, but is not critical to verbal understanding, which can be accomplished through the ventral stream. A similar division has been proposed for the processing of visual inputs, with a dorsal stream mapping onto action and a ventral stream onto recognition (Milner and Goodale, 2006). The visual streams may play roles in sign language parallel to those of the auditory streams in speech.

The Gestural Origins of Language

Although people (and dogs) can understand language even when unable to produce it, the evolution of language as a two-way communication system probably grew out of the MNS. The MNS in primates has to do primarily with manual grasping, supporting the notion that language itself evolved from manual gesture, not from vocal calls (Hewes, 1973). Area F5 in the primate cortex – the site of the original discovery of mirror neurons – is homologous with Broca's area in the human brain. 'Language,' as Rizzolatti and Arbib (1998) entitle their article, 'is within our grasp.'

Support for the gestural theory comes from the now classic attempts to teach great apes to communicate in language-like ways. Efforts to teach them anything resembling speech failed rather dismally (e.g., Hayes, 1952; Kellogg and Kellogg, 1933), and much greater success was achieved by teaching a form of signed language (Gardner and Gardner, 1969; Patterson, 1978), or by having them point to symbols on a keyboard (Savage-Rumbaugh et al., 1998). Gestural communication has been widely documented among great ages in the wild (e.g., Hobaiter and Byrne, 2011; Pika and Mitani, 2009; Tanner and Byrne, 2010). Manual gestures persist as a critical accompaniment to speech in regular discourse between humans (e.g., Iverson and Goldin-Meadow, 2005), and carry the full burden

of language in the signed languages of the deaf (Emmorey, 2002).

Given the importance of manual gesture as an accompaniment of speech, and the autonomy and linguistic sophistication of signed languages in the deaf, some have queried the notion that manual gesture preceded speech in language evolution, and proposed that they have always been equal partners (e.g., Kendon, 2009, 2012; McNeill, 2005). Some support for this view comes from the strong association between movements of the hands and movements of the face. In the macaque, some mirror neurons respond both when the animal grasps with the hand and when it grasps with the mouth (Gallese et al., 1996; Rizzolatti et al., 1988). In humans, too, there is a strong link between hand and mouth movements; for instance, the way in which a speech sound is uttered is different when the speaker picks up a large object, such as an apple, than when the speaker picks up a smaller one, such as a grape. It is similarly influenced when the speaker simply watches another person picking up these objects (Gentilucci, 2009; Gentilucci and Corballis, 2006). In modern sign languages, movements of the face may play as important a role as movements of the hands (e.g., Emmorey, 2002; Sutton-Spence and Boyes-Braem, 2001).

The evolutionary latecomer, though, may have been voicing. All primates communicate vocally, but their vocalizations are largely instinctive and driven by emotion rather than volition. Primate calls are driven by the limbic system of the brain rather than the motor cortex (Jürgens, 2002). Even chimpanzees, our closest nonhuman relatives, appear to have little intentional control of the voice; Premack (2007), for example, writes that "chimpanzees lack voluntary control of their voice" (p. 13866). Nevertheless recent evidence suggests that chimpanzees, at least, do have limited intentional control of vocalization; for example, they are able to direct food calls at specific individuals, implying both a degree of vocal flexibility and intentional control (Schel et al., 2013). This flexibility, though, pales in comparison to the flexibility and intentional control of the hands in our primate forebears.

A likely scenario is that language evolved within the MNS from the mechanisms of grasping, whether by hand or mouth, in our primate forbears. One limitation of this system in monkeys is that it is transitive, responding only to actions involving the grasping of an actual object, even if that object is hidden behind a screen but the animal knows it is there (Umiltà et al., 2001). In humans, though, evidence from neuroimaging suggests that the MNS responds not only to transitive actions, but also to intransitive ones, where a movement occurs in the absence of an object to be grasped (Rizzolatti and Sinigaglia, 2006), paving the way to imitative or mock actions. Thus pantomime might have been an early precursor to symbolic language. Some degree of pantomime is evident in the natural communications of apes, including orangutans (Russon and Andrews, 2011), although it is unlikely that gestural communication evolved a storylike structure until the emergence of bipedal hominins, and perhaps not even until the emergence of the genus *Homo* some 2.6 million years ago (Thompson, 2010).

A critical development may have been the development of stone tool-making industries, perhaps at the transition from the earlier Oldowan industry to the later, more complex Acheulian industry dating from around 1.7 million years ago. Stout et al. (2008) asked three archeologists experienced in early tool technologies to make Oldowan tools or Acheulian tools, or simply strike cobbles together, and in each case measured their brain activity using a Positron emission tomography scanner with a slow-decaying radiological trace. Only in the case of Acheulian tool making did the brain activity overlap with language circuits, including Broca's area. The active areas also corresponded to the known MNS. Unlike Oldowan tools, Acheulian tool making involves planned sequential action, which may have given rise to pantomime, which might also have been extended to other action episodes, such as accounts of hunting excursions, dating from the middle Pleistocene.

Pantomime, though, is inefficient, and can be streamlined by a process of conventionalization (Burling, 1999), whereby pictorial or iconic representations are gradually replaced by arbitrary signals, shaped to convey information with maximum efficiency and minimum effort. In losing their pictorial quality, though, these arbitrary signals must be sustained and taught by the linguistic community, and transmitted between generations. The process of conventionalization is evident in the development of sign languages. For example, in American Sign language, the sign for *home* was once a combination of the sign for *eat*, which is a bunched hand touching the mouth, and the sign for *sleep*, which is a flat hand on the cheek. Now it consists of two quick touches on the cheek, both with a bunched hand shape; so the original iconic components are effectively lost (Frishberg, 1975). Albeit with a lack of modern cultural sensitivity, Darwin (1872) remarked "on the practice of the deaf and dumb and of savages to contract their signs as much as possible for the sake of rapidity. Hence their natural source or origin often becomes doubtful or is completely lost; as is likewise the case with articulate language" (p. 62).

Because of their spatial nature, sign languages do retain an element of mime, or what has been termed iconicity. It has been estimated, for example, that in Italian Sign Language some 50% of the hand signs and 67% of the bodily locations of signs stem from iconic representations, with a mapping in space and time between the sign and its meaning (Pietrandrea, 2002). In American Sign Language, too, some signs are arbitrary, but many more are iconic. For example, the sign for 'erase' resembles the action of erasing a blackboard, and the sign for 'play piano' mimics the action of actually playing a piano (Emmorey, 2002).

Speech, then, can be regarded as an end-point in the process of conventionalization, whereby arbitrary sounds came to accompany and replace visible movements as the primary carrier of the linguistic message – although, as we have seen, manual gestures remain an important element of normal discourse. This shift may have been the result of a progression of gesture from hand and face to movements internal to the mouth, with the accompaniment of voicing. As Studdert-Kennedy (2005) pointed out, speech itself is a gestural system, involving coordinated movements of the lips, the larynx, the velum, and the blade, body, and root of the tongue. Just when controlled voicing emerged in evolution has been a matter of controversy, with some arguing that articulate speech was possible only in *Homo sapiens* (Lieberman, 2007), with others suggesting that the

Neanderthals, who predated humans by at least 500 000 years, would have possessed the capacity to speak, and perhaps even a fully human language capacity (Johansson, 2013). As we have seen, chimpanzees have some limited intentional control over their calls, but this probably serves to attract attention rather than to convey propositional information (Schel et al., 2013).

On this account, then, gestures of the hands, and perhaps the face, did precede vocalization in the evolution of language, although the switch was more a question of changing dominance rather than an abrupt transition. Even in chimpanzees, vocal calls may accompany manual gestures, primarily to attract or direct attention, while in humans manual gestures serve to embellish speech and occasionally substitute for it. Regardless of the timing of the change, though, language remains a gestural system, evolving through the MNS. The eventual dominance of speech effectively involves the retreat of the gestural repertoire into the mouth, to allow the rest of the body freedom for other activities, such as the use and manufacture of tools, and perhaps for carrying things. In this respect, the emergence of speech may be seen as an early example of miniaturization.

Whether language is voiced or produced manually, then, the MNS seems to have provided the platform for its evolution. The fronto-parietal MNS identified by Rizzolatti and Sinigaglia (2010) shows a remarkable homology with the language network in the human brain. Area F5 in the macaque is homologous to Broca's area in the human brain, and the parietal mirror regions encompass the homologue of Wernicke's area. In the course of evolution, moreover, it seems that the system became increasingly lateralized toward the left cerebral hemisphere. Leftward asymmetries of the homologues of Broca's and Wernicke's areas are evident in the majority of chimpanzees (Hopkins and Cantalupo, 2008), as well as in most humans (Corballis, 2002).

Empathy and ToM

If mirror neurons map perceived movements onto their production, thereby providing for the understanding of the actions of others, one might ask whether they play a social role beyond that of language. One possibility is that mirror neurons form the basis of empathy, the understanding of emotional states in others. A more advanced possibility is that they also allow us to understand cognitive states, such as what others see, think, or believe.

Empathy

One structure that plays a critical role in experiencing the emotions of others, and that has mirror properties, is the insula, a deep cortical structure folded within the lateral sulcus (which separates the parietal and temporal lobes). Its role in one emotion that of disgust, has been well studied. In the monkey brain, the insula receives input from chemoreceptors as well as from bodily states, and stimulation of the insula produces visuomotor and autonomic responses indicative of a feeling of disgust. Stimulating the insula in humans likewise produces a feeling of nausea, and brain imaging has shown that the insula is also activated by the sight of disgusted facial expressions in others (Gallese et al., 2004).

Other evidence suggests that the insula, along with the anterior cingulate cortex, also provides an MNS for pain, allowing people (and perhaps monkeys) to experience their own pain, as well as the pain of others (Singer et al., 2004). The empathy applies to the affective component, but not the sensory one. You may say to an ailing friend 'I feel your pain,' but what you feel is the emotional state, but not the physical pain itself.

The MNS appears also to be involved in understanding other emotions. One of the characteristics of the autism spectral disorder is a lack of feeling of the emotions of others, perhaps as a consequence of a dysfunction of the MNS (Rogers and Pennington, 1991). In one study, children were shown pictures of faces showing emotions of anger, fear, happiness, and sadness, along with a neutral expression, and asked to copy them. Both normal children and high-functioning children with autism were able to imitate the emotions, but in the children with autism, unlike the normal children, fMRI recording showed no activity in the frontal component of the MNS (pars opercularis), or in the insula and amygdala (Dapretto et al., 2006). These results suggest that both normal children and those with autism can detect and imitate emotional expressions, but do so in very different ways. The autistic child may perceive your sadness, but not feel it.

Similar findings have been reported in adults. Both normal adults and those diagnosed with autism spectrum disorder were able to identify emotions expressed in faces. While doing so, the MNS was activated, except that in the group with autism the frontal component was missing (Schulte-Ruether et al., 2011). The authors suggest that it is the frontal component that is responsible for creating the 'emotional bond' between people.

Theory of Mind

The MNS may go beyond empathy to provide information not only on what others feel, but also on what they think or believe. This faculty is known as ToM. Whether nonhuman animals possess this faculty has been a matter of controversy ever since Premack and Woodruff (1978) posed the question "Does the chimpanzee have a theory of mind?" Opinion remains divided. Call and Tomasello (2008) concluded that 30 years of research since the question was posed show that chimpanzees have some understanding of the goals, intentions, perceptions, and knowledge of others, but no understanding of their beliefs or desires. In the same year Penn and Holyoak (2008) remained convinced that "nonhuman animals do not appear to possess anything remotely possessing a ToM" (p. 121).

Even in macaques, though, research suggests a role for the MNS in understanding the intentions of others, which suggests at least limited ToM. Some mirror neurons in the inferior parietal lobule of the macaque are more active when the animal grasps a food item with the intention of eating it than with the intention of placing the food in a container, while others show the reverse pattern. A similar dissociation is recorded when the animal observes the experimenter grasping food, indicating that parietal mirror neurons convey information as to the intention of others (Koski et al., 2002). fMRI studies in humans, too, have shown different activity in the MNS

depending on whether grasping actions are associated with drinking or cleaning up (Iacoboni et al., 2005). Understanding the intentions of others may well have had survival value in our primate heritage, serving as a warning of possible future behavior, as well as creating a capacity for cooperation. In humans, moreover, shared intentionality is a prerequisite for language, the most highly developed of our social tools (Tomasello et al., 2005).

Imitation and Cultural Learning

We humans are distinguished not only by language, but also by the complexities of culture, defined as "the system of shared beliefs, values, customs, behaviors, and artifacts that the members of society use to cope with their world and with one another, and that are transmitted from generation to generation through learning" (Bates and Plog, 1990: p.7). It is also widely agreed that cultural learning depends on sophisticated imitation, and is especially well developed in humans. It has been proposed that this learning depends at least in part on imitation, and is acquired through the MNS. Losin et al. (2009: p. 182) suggest that "the MNS may not only play a role in the acquisition of culturally mediated behaviors, but the MNS itself may be shaped by the presence of culturally mediated behaviors in one's motor repertoire." Like language itself, then, imitation may depend on the adaptation of the MNS to perform a function not well developed in nonhuman species, but essential to cultural learning in humans.

Contrary to the claim of Rizzolatti and Craighero (2004) that the primary role of the MNS in primates is not action imitation, it does seem clear that in humans, at least, the MNS is adapted, through learning, to create what has been termed "automatic imitation" (Heyes, 2011). Brain-imaging studies provide evidence that experience does indeed alter MNS activity. For example, Vogt and Thomaschke (2007) found increased activity in a number of brain regions, including the putative MNS, while participants observed practiced compared with nonpracticed guitar chords. Over a longer period, Cross et al. (2006) showed practice-related increases in MNS activity over five fMRI scans at weekly intervals while their participants learned a novel dance sequence. More pertinent to the question of cultural learning, Calvo-Merino et al. (2005) showed that changes in MNS activity resulting from practice extended to real-world expertise acquired over many years. They found greater MNS activity in experienced dancers when they observed their own style of dance rather than a comparable but unfamiliar style. Potentially, at least, daily experiences, including those specific to the culture we live in, can influence MNS function, and through overlearning explain our automatic tendency to imitate. And of course language itself depends heavily on imitation, and is culturally constrained – there are some 7000 different languages in the world, each more or less impenetrable to others.

Although automatic imitation is influenced by experience and culture, there may be a hard-wired component as well. Just hours after birth, infants imitate facial gestures such as protruding the tongue, a finding that has been much replicated (Meltzoff and Decety, 2003). One model as to how the MNS might change as a result of experience has been proposed by Iacoboni (2005). He suggested that the STS, which is part of the MNS but does not include the motor component, constructs a higher-order description of the observed action. This description is transmitted to parietal and frontal components of the MNS, where the goal of the action and motor components to achieve the action are encoded. The motor plan is then fed back into the STS, where its sensory consequences are compared with the visual description of the original action. This circuit leads to progressive refinement of the MNS circuitry.

Automatic imitation pervades our social interactions. People tend to yawn or laugh when someone else yawns or laughs, or even the tendency of people in conversation to imitate each other when in conversation, perhaps crossing their legs or holding their arms in similar fashion (Hurford, 2004). These imitative tendencies may help to strengthen social bonds. Nevertheless they can interfere with intentional behavior. In spoken discourse, for example, a natural tendency to imitate what one hears must be suppressed to sustain effective communication. A compulsive tendency to repeat heard utterances, known as echolalia, is sometimes observed, and attributed to frontal lobe deficits (e.g., Grossi et al., 2013; Cho et al., 2009). A number of studies have shown that when people are instructed to make simple finger movements in response to neutral cues, their movements are inhibited when videos of incompatible movements are presented simultaneously with the cues (e.g., Cross et al., 2013; Stürmer et al., 2000). Automatic imitation is reduced when the frontal node of the MNS is stimulated through transcranial magnetic stimulation (Catmur et al., 2009). Adaptive social behavior depends on the interaction between control processes, mediated by a 'prefrontal control system,' and the more automatic reactions of the MNS (Cross et al., 2013).

Conclusions

As Ramachandran anticipated, mirror neurons have indeed provided insight into many aspects of cognition, some of them distinctively human. These include language, empathy, ToM, imitation, and cultural learning. Although mirror neurons were first identified in monkeys, their extensive and distinctive roles in human condition illustrates the fact that the MNS is an adaptive system, providing complex but automatic calibration to social contingencies. The system is part of our biological heritage, shaping us, and indeed our evolutionary predecessors to life among members of the same species.

The MNS has also been taken to support the notion that cognition is embodied (Barsalou, 2008), such that perceiving the world is computationally entwined with our bodily movements, and thought itself is made up of internal simulations of real-world events and actions. This stands in contrast to the idea that thought is essentially computational and symbolic, as in a digital computer. Again, it is opposed to the Chomsky's (2010) view of language and thought (or what he calls 'I-language'), which is that symbolic representations and computational algorithms emerged in humans without reference to the external world – a view incompatible with Darwin's theory of evolution by natural selection. The roles

of mirror neurons may remain incompletely specified and indeed 'fuzzy' (Hurford, 2004), but they do provide the basis for a more Darwinian perspective on the evolution of thought, and for a paradigm shift from computer-based models of the mind toward models that are based on real-world interactions with the biological and social environments.

See also: Autism Spectrum Disorders; Empathy During Early Childhood Across Cultures, Development of; Imitation During Infancy and Early Childhood; Joint Attention During Infancy and Early Childhood Across Cultures; Neural Correlates of Social Development.

Bibliography

Arbib, M.A., 2005. From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. Behavioral and Brain Science 28, 105–168.

Barsalou, L.W., 2008. Grounded cognition. Annual Review of Psychology 59, 617–645.

Bates, D.G., Plog, F., 1990. Cultural Anthropology. McGraw-Hill, New York.

Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., et al., 2004. Neural circuits involved in the recognition of actions performed by nonconspecifics: an fMRI study. Journal of Cognitive Neuroscience 16, 114–126.

Burling, R., 1999. Motivation, conventionalization, and arbitrariness in the origin of language. In: King, B.J. (Ed.), The Origins of Language: What Nonhuman Primates Can Tell Us. School of American Research Press, Santa Fe, NM.

Call, J., Tomasello, M., 2008. Does the chimpanzee have a theory of mind? 30 years later. Trends in Cognitive Sciences 12, 187–192.

Calvo-Merino, B., Glaser, D.E., Grèzes, J., Passingham, R.E., Haggard, P., 2005. Action observation and acquired motor skills: an fMRI study with expert dancers. Cerebral Cortex 15, 1243–1249.

Catmur, C., Walsh, V., Heyes, C., 2009. Associative sequence learning: the role of experience in the development of imitation and the MNS. Philosophical Transactions of the Royal Society B 364, 2369–2380.

Cho, Y.-J., Han, S.-D., Song, S.K., Lee, B.I., Heo, K., 2009. Palilalia, echolalia, and echopraxia–palipraxia as ictal manifestations in a patient with left frontal lobe epilepsy. Epliepsia 50, 1616–1619.

Chomsky, N., 2010. Some simple evo devo theses: how true might they be for language? In: Larson, R.K., Déprez, V., Yamakido, H. (Eds.), The Evolution of Human Language. Cambridge University Press, Cambridge, pp. 45–62.

Corballis, M.C., 2002. From Hand to Mouth: The Origins of Language. Princeton University Press, Princeton, NJ.

Cross, E.S., Hamilton, A.F.C., Grafton, S.T., 2006. Building a motor simulation de novo: observation of dance by dancers. NeuroImage 31, 1257–1267.

Cross, K.A., Salvatore Torrisi, S., Losin, E.A.R., Iacoboni, M., 2013. Controlling automatic imitative tendencies: interactions between mirror neuron and cognitive control systems. NeuroImage 83, 493–504.

Dapretto, M., Pfeifer, M.S., Scott, J.H., Scott, A.A., Sigman, M., Bookheimer, S.Y., et al., 2006. Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. Nature Neuroscience 9, 28–30.

Darwin, C., 1872. The Expression of the Emotions in Man and Animals. John Murray, London.

Emmorey, K., 2002. Language, Cognition, and Brain: Insights from Sign Language Research. Erlbaum, Hillsdale, NJ.

Fadiga, L., Fogassi, L., Pavesi, G., Rizzolatti, G., 1995. Motor facilitation during action observation: a magnetic stimulation study. Journal of Neurophysiology 73, 2608–2611.

Frishberg, N., 1975. Arbitrariness and iconicity in American sign language. Language 51, 696–719.

Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G., 1996. Action recognition in the premotor cortex. Brain 119, 593–609.

Gallese, V., Keysers, C., Rizzolatti, G., 2004. A unifying view of the basis of social cognition. Trends in Cognitive Science 8, 396–403.

Gallese, V., Rizzolatti, G., 2008. How pliers become fingers in the monkey motor system. Proceedings of the National Academy of Sciences of the United States of America 105, 2209–2213.

Gardner, R.A., Gardner, B.T., 1969. Teaching sign language to a chimpanzee. Science 165, 664–672.

Gentilucci, M., 2009. The observation of manual grasp actions affects the control of speech: a combined behavioral and transcortical magnetic stimulation study. Neuropsychologia 47, 3190–3202.

Gentilucci, M., Corballis, M.C., 2006. From manual gesture to speech: a gradual transition. Neuroscience & Biobehavioral Reviews 30, 949–960.

Grossi, D., Marcone, R., Cinquegrana, T., Gallucci, M., 2013. On the differential nature of induced and incidental echolalia in autism. Journal of Intellectual Disability Research 57, 903–912.

Hayes, C., 1952. The Ape in Our House. Gollancz, London.

Hewes, G.W., 1973. Primate communication and the gestural origins of language. Current Anthropology 14, 5–24.

Heyes, C., 2011. Automatic imitation. Psychological Bulletin 137, 463–483.

Hickok, G.S., 2009. Eight problems for the mirror neuron theory of action understanding in monkeys and humans. Journal of Cognitive Neuroscience 21, 1229–1243.

Hickok, G.S., Poeppel, D., 2007. The cortical organization of speech processing. Nature Reviews Neuroscience 8, 393–402.

Hobaiter, C., Byrne, R.W., 2011. Serial gesturing by wild chimpanzees: its nature and function for communication. Animal Cognition 14, 827–838.

Hopkins, W.D., Cantalupo, C., 2008. Theoretical speculations on the evolutionary origins of hemispheric specialization. Current Direction in Psychological Science 17, 233–237.

Hurford, J., 2004. Language beyond our grasp: what mirror neurons can, and cannot, do for language evolution. In: Kimbrough Oller, D., Griebel, U., Plunkett, K. (Eds.), The Evolution of Communication Systems: A Comparative Approach. MIT Press, Cambridge, MA, pp. 297–313.

Iacoboni, M., 2005. Neural mechanisms of imitation. Current Opinion in Neurobiology 15, 632–637.

Iacoboni, M., Woods, R.P., Brass, M., Bekkering, H., Mazziotta, J.C., Rizzolatti, G., 1999. Cortical mechanisms of human imitation. Science 286, 2526–2528.

Iacoboni, M., Molnar-Szakac, I., Gallese, V., Buccino, G., Mazzione, J.C., Rizzolatti, G., 2005. Grasping the intentions of others with one's own mirror neuron system. PLoS Biology 3, e79.

Iverson, J.M., Goldin-Meadow, S., 2005. Gesture paves the way for language development. Psychological Science 16, 367–371.

Johansson, S., 2013. The talking Neanderthals: what do genetics, fossils, and archeology say? Biolinguistics 7, 35–74.

Jürgens, U., 2002. Neural pathways underlying vocal control. Neuroscience & Biobehavioral Reviews 26, 235–258.

Kaminski, J., Call, J., Fischer, J., 2006. Word learning in a domestic dog: evidence for "fast mapping." Science 304, 1682–1683.

Kellogg, W.N., Kellogg, L.A., 1933. The Ape and the Child: A Study of Early Environmental Influence upon Early Behavior. McGraw-Hill, New York.

Kendon, A., 2009. Language's matrix. Gesture 9, 355–372.

Kendon, A., 2012. Vocalisation, speech, gesture, and the language origins debate. Gesture 11, 349–370.

Kohler, E., Keysers, C., Umiltà, M.A., Fogassi, L., Gallese, V., Rizzolatti, G., 2002. Hearing sound, understanding actions: action representation in mirror neurons. Science 297, 846–848.

Koski, L., Wohlschlager, A., Bekkering, H., Woods, R.P., Dubeau, M.C., Mazziotta, J.C., Iacoboni, M., 2002. Modulation of motor and premotor activity during imitation of target-directed actions. Cerebral Cortex 12, 847–855.

Liberman, A., Cooper, F.S., Shankweiler, D.A., Studdert-Kennedy, M., 1967. Perception of the speech code. Psychological Reviews 74, 431–461.

Lieberman, P., 2007. The evolution of human speech. Current Anthropology 48, 39–66.

Losin, E.A.R., Dapretto, M., Iacobon, M., 2009. Culture in the mind's mirror: how anthropology and neuroscience can inform a model of the neural substrate for cultural imitative learning. In: Chiao, J.Y. (Ed.), Progress in Brain Research, vol. 178. Elsevier, Amsterdam, pp. 175–190.

McNeill, D., 2005. Gesture and Thought. University of Chicago Press, Chicago.

Meltzoff, A.N., Decety, J., 2003. What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive neuroscience. Philosophical Transactions of the Royal Society B. Biological Sciences 358, 491–500.

Milner, A.D., Goodale, M.A., 2006. The Visual Brain in Action, second ed. Oxford University Press, Oxford.

Molenberghs, P., Cunnington, R., Mattingley, J.B., 2012. Brain regions with mirror properties: a meta-analysis of 125 human fMRI studies. Neuroscience & Biobehavioral Reviews 36, 341–349.

Mukamel, R., Ekstrom, A.D., Kaplan, J., Iacoboni, M., Fried, I., 2010. Single-cell responses in humans. Current Biology 20, 750–756.

Patterson, F., 1978. Conversations with a gorilla. National Geographic 154, 438–465.

Penn, D.C., Holyoak, K.J., Povinelli, D.J., 2008. Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. Behavioral & Brain Sciences 31, 108–178.

Pika, S., Mitani, J.C., 2009. The directed scratch: evidence for a referential gesture in chimpanzees? In: Botha, R., Knight, C. (Eds.), The Prehistory of Language. Oxford Scholarship Online, Oxford, UK, pp. 166–180.

Pietrandrea, P., 2002. Iconicity and arbitrariness in italian sign language. Sign Language Studies 2, 296–321.

Pilley, J.W., Reid, A.K., 2011. Border collie comprehends object names as verbal referents. Behavioural Processes 86, 184–195.

Premack, D., 2007. Human and animal cognition: Continuity and discontinuity. Proceedings of the National Academy of Sciences (USA) 104, 13861–13867.

Premack, D., Woodruff, G., 1978. Does the chimpanzee have a theory of mind? Behavioral and Brain Science 4, 515–526.

Ramachandran, V.S., 2000. Mirror Neurons and Imitation Learning as the Driving Force behind the "Great Leap Forward" in Human Evolution. Edge. http://www.edge.org/3rd_culture/.

Rizzolatti, G., Arbib, M.A., 1998. Language within our grasp. Trends in Neurosciences 21, 188–194.

Rizzolatti, G., Camardi, R., Fogassi, L., Gentilucci, M., Luppino, G., Matelli, M., 1988. Functional organization of inferior area 6 in the macaque monkey. II. Area F5 and the control of distal movements. Experimental Brain Research 71, 491–507.

Rizzolatti, G., Craighero, L., 2004. The mirror-neuron system. Annual Review of Neuroscience 7, 169–192.

Rizzolatti, G., Sinigaglia, C., 2006. Mirrors in the Brain. Oxford University Press, Oxford.

Rizzolatti, G., Sinigaglia, C., 2010. The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations. Nature Reviews Neuroscience 11, 264–274.

Rogers, S.J., Pennington, B.F., 1991. A theoretical approach to the deficits in infantile autism. Developmental Psychology 3, 137–162.

Russon, A., Andrews, K., 2011. Orangutan pantomime: elaborating the message. Biology Letters 7, 627–630.

Savage-Rumbaugh, S., Shanker, S.G., Taylor, T.J., 1998. Apes, Language, and the Human Mind. Oxford University Press, New York.

Schel, A.M., Machanda, Z., Townsend, S.W., Kluberbühler, K., Slocombe, K.E., 2013. Chimpanzee food calls are directed at specific individuals. Animal Behaviour 86, 955–965.

Schulte-Ruether, M., Greimel, E., Markowitsch, H.J., Kamp-Becker, I., Remschmidt, H., Fink, G.R., et al., 2011. Dysfunctions in brain networks supporting empathy: an fMRI study in adults with autism spectrum disorders. Society for Neuroscience 6, 1–21.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R.J., Frith, C.D., 2004. Empathy for pain involves the affective but not the sensory components of pain. Science 303, 1157–1162.

Stout, D., Toth, N., Schick, K., Chaminade, T., 2008. Neural correlates of early stone age toolmaking: technology, language and cognition in human evolution. Philosophical Transactions of the Royal Society B 363, 1939–1949.

Studdert-Kennedy, M., 2005. How did language go discrete? In: Tallerman, M. (Ed.), Language Origins: Perspectives on Evolution. Oxford University Press, Oxford, pp. 48–67.

Stürmer, B., Aschersleben, G., Prinz, W., 2000. Correspondence effects with manual gestures and postures: a study of imitation. Journal of Experimental Psychology: Human Perception and Performance 26, 1746–1759.

Sutton-Spence, R., Boyes-Braem, P. (Eds.), 2001. The Hands Are the Head of the Mouth: The Mouth as Articulator in Sign Language. Signum-Verlag, Hamburg.

Tanner, J.E., Byrne, R.W., 2010. Triadic and collaborative play by gorillas in social games with objects. Animal Cognition 13, 591–607.

Thompson, T., 2010. The ape that captured time: folklore, narrative, and the human-animal divide. Western Folklore 69, 395–420.

Tomasello, M., Carpenter, M., Call, J., Behme, T., Moll, H., 2005. Understanding and sharing intentions. Behavioral & Brain Sciences 28, 675–691.

Umiltà, M.A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., et al., 2001. I know what you're doing: a neurophysiological study. Neuron 31, 155–165.

Umiltà, M.A., Escola, L., Instkerveli, I., Grammont, F., Rochat, M., Caruana, F., et al., 2008. When pliers become fingers in the monkey motor system. Proceedings of the National Academy of Sciences 105, 2209–2213.

Visalberghi, E., Fragaszy, D., 2001. Do monkeys ape? Ten years after. In: Dautenhahn, K., Nehaniv, C.L. (Eds.), Imitation in Animals and Artifacts. MIT Press, Cambridge, MA.

Vogt, S., Thomaschke, R., 2007. From visuo-motor interactions to imitation learning: behavioural and brain imaging studies. Journal of Sports Science 25, 497–517.

Misdemeanor Offenses

DeAnn Kalich, Margot Hasha, and Craig J Forsyth, University of Louisiana, Lafayette, Lafayette, LA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Misdemeanor offenses are largely ignored in the literature, because attention is usually focused on the most serious and violent crimes. Most people entering the American criminal justice system do so for misdemeanor offenses. These offenses, while considered petty, are prohibited by misdemeanor laws. *Mala prohibita* crimes (wrong only because they are prohibited by law) refer to behaviors that violate contemporary standards only, such as white-collar crime and illegal drug use. These crimes represent society's attempts to regulate behavior that offend the moral sensibilities of particular people during specific periods of time in certain locations, and correspond generally to misdemeanors. This article examines the role of moral entrepreneurs in constructing definitions of crime/deviance as misdemeanors/felonies and subsequent examples of misdemeanors in American society. An examination of the very real consequences to being labeled guilty of a misdemeanor offense is also presented.

Misdemeanor offenses are largely ignored in the literature, because attention is usually focused on felonies: the most serious and violent crimes, such as robbery, rape, and murder. In the United States, approximately 1 million felony convictions are entered annually, while more than 10 million misdemeanor cases are filed during the same period. In fact, in most states, misdemeanor dockets are four to five times the size of felony dockets, with these wrongful convictions never making it into headlines specifically because they are petty offenses, for example, trespassing, disorderly conduct, and loitering. Most people entering the American criminal justice system do so for misdemeanor offenses. These offenses, while considered petty, are prohibited by misdemeanor laws. How we process misdemeanors in the United States represents an immense public institution – one so large and powerful that it deserves far more scrutiny than has been afforded it thus far.

When exploring criminal offenses, it is helpful to distinguish between *mala prohibita* crimes and *mala in se* crimes, with the latter considered more serious than the former. *Mala in se* (evil in themselves) crimes refer to behaviors that violate moral codes as well as traditional norms. This category encompasses the core of the criminal code, including property crimes, violent crimes, robbery, and homicide, and corresponds to felony crimes. Conversely, *mala prohibita* crimes (wrong only because they are prohibited by law) refer to behaviors that violate contemporary standards only, such as white-collar crime and illegal drug use. These crimes represent society's attempts to regulate behavior such as drug use, gambling, and prostitution that offend the moral sensibilities of particular people during specific periods of time in certain locations, and correspond generally to misdemeanor crimes. *Mala prohibita* crimes are likely to vary over time and across jurisdictions. An example is the consumption of alcohol in the United States, which shifted from legal to illegal during Prohibition with the ratification of the 18th Amendment and the Volstead Act, and back to legal again through the ratification of the 21st Amendment, which repealed the 18th. Similarly, casino gambling is legal in some states and illegal in others, with many states also having sanctioned lotteries (Vito et al., 2007).

Moral Entrepreneurs

It is sometimes difficult to define misdemeanor crime, because the term refers to a diverse collection of offenses that are grouped differently across jurisdictions. Some social science researchers reference misdemeanors as 'morality offenses.' This perspective states criminal law is used to impose minimum standards of morality on a citizenry. The 'minimum standards' are considered 'misdemeanors.' The laws defining offenses which are against public morality are used to set the standards for acceptable behavior. Howard Becker argues that groups of like-minded individuals act as *moral entrepreneurs* to define the behavioral standards that will be acceptable to a specific community. Also referred to as 'rule creators' and 'rule enforcers,' moral entrepreneurs are in the business of constructing definitions of crime/deviance and then for policing the subsequently identified rule violations. Their chief concern is the persuasion of others, not the means by which that persuasion is achieved.

While misdemeanor offenses are generally considered to be less serious than felony offenses – as evidenced by the degree of legal censure present in the scheduled punishment for the misdemeanor – these offenses often are treated as quite serious by groups of moral entrepreneurs who themselves may hold little or no official legal status in their community. The sense of morality that is offended in the act of the misdemeanor is the sense of morality belonging to the moral entrepreneurs. For these groups, then, it is vitally important that the punishment assigned to the rule violation does, in fact, occur. Often, offenses that are classified as misdemeanors are in fact of great importance to the dominant culture groups, who are responsible for having defined the laws which target the behaviors in question. The label of misdemeanor may therefore carry a small official sanction relative to that carried by a felony offense, but the unofficial response of some interested observers to the rule violation is one of moral outrage and therefore of great importance. This is because the rule that has been defined as a morality offense is a rule that reflects the values and morals of the population of interested observers (what Becker refers to as moral entrepreneurs), typically those inhabiting upper social strata in the dominant culture.

When the efforts of rule creating moral entrepreneurs are successful and new sets of laws and enforcement agencies finally institutionalize their crusades, 'rule enforcers,' who are comprised of police and court systems, attempt to not only enforce these laws but also to justify their own roles with respect to their interactions with the public. These rule enforcers represent another category of moral entrepreneur, according to Becker. He asserts that rule enforcers are in a bind: to be too effective might indicate that society does not need them, while not being effective enough might lead to the societal belief that they are failing at their roles. As rule enforcers, these groups are less concerned with the content of the rule than they are with performing their jobs. And, because rules (and thus, laws) change over time, and what was once unacceptable might now be accepted in society (such as the legalization of marijuana use in some states), rule enforcers often devote much time and energy to coercing respect from those with whom they interact. As such, persons may be deemed deviant not so much for breaking the rules as for demonstrating disrespect to the rule enforcers. According to Becker, these police and judicial officials view human nature pessimistically as a result of their constant exposure to 'willful deviance.'

Labeling an individual who commits a misdemeanor as 'deviant' depends upon factors outside the actual behavior, such as if an official experiences pressure to justify his/her position at a given time; whether or not the rule enforcer feels respected by the actor; whether the crime committed was performed by an amateur (who is more likely to be caught, convicted, and labeled deviant); or whether the misdemeanor committed is even high on the priority list of the rule enforcer.

Social theorists such as Joseph Gusfield have posited that official social recognition of some rules as valid and moral rules is a way of conferring higher social status to the groups whose values are reflected by those rules. Thus, elevating a group's norms behavior to the status of a law is a way of reinforcing the superiority of that group's ideas and morality relative to the ideas and morality of other, competing groups. This is true despite the seriousness of the sanction or legal response assigned to the rule violation. The fact that the behavior is defined as a violation at all is enough to accomplish what is needed in defining one social group as superior to another. Social groups that are successful in having their definitions of good and bad, right and wrong, moral and immoral encoded into law are therefore reaffirmed as superior to those groups whose ways of life are not recognized as legal. Misdemeanor crimes often are behaviors that are criminalized because particular groups prohibit the behaviors on moral grounds. Moral entrepreneurs will wage moral crusades in order to manufacture widespread public concern about an issue and to increase support for criminalizing that behavior in the written law. They tend to be righteous, sometimes self-righteous and fervent in their efforts, and their mission is to promote their sense of morality. Many legal campaigns have the character of a moral crusade to criminalize behavior that may not in fact be as dangerous or expensive to society as other noncriminal behaviors.

One example of moral entrepreneurs who have engaged in moral crusades is the Women's Christian Temperance Union and the U.S. Prohibition Movement. This alliance brought together conservative Christian activists, Women's Rights groups, and anti-Child Labor groups for a common cause. According to Gusfield, Temperance (moral) abstainers sought to correct the behavior of (immoral) drinkers. Drinking and nondrinking historically have been methods by which to identify members of a subculture, thus drinking and abstinence eventually became symbols of social status. Through the Temperance Movement, a declining social elite comprised of White Anglo-Saxon Protestants attempted to retain social power and leadership authority. By the 1830s, abstinence had become an aspect of middle-class membership, a necessity for social upward mobility, and a component of public morality. The passage of the 18th Amendment in 1920 was the apex of the victory of Protestant over Catholic and tradition over modernity.

Another group of moral entrepreneurs might be the U.S. Federal Narcotics Bureau, especially while under the direction of Harry Anslinger. Craig Reinarman argues that most of the antidrug campaigns across history in the United States are moral crusades, such as the antimarijuana and anti-opium movements founded during the early twentieth century. Similarly, groups such as Mothers Against Drunk Drivers (MADD) and Students Against Drunk Drivers have organized to persuade public morality to convert and criminalize the act of drinking and drinking. More recently, an alliance has formed among conservative Christian activists and Feminists opposed to human trafficking. The role of moral entrepreneur is a role that could be applied to any social group that seeks to convert others to the primary group's way of thinking about a particular issue affecting society. Thus, it can even be argued that the activist parent group (modeled after MADD) known as Bothered About Dungeons and Dragons (BADD) is also a group of moral entrepreneurs that has attempted to have the fantasy game 'Dungeons and Dragons' banned from sale, or prohibited for underage players. Parents in the group have concerns about the connections between the game, satanic worship, and the occult and suicidal behavior in teenagers. Though BADD has not been successful in redefining that specific practice of game playing as a misdemeanor crime, the crusade continues.

It can likewise be argued that groups such as the Ku Klux Klan (KKK) act as moral entrepreneurs when attempting to police the behaviors of others according to the values espoused by the Klan. Historically, behaviors deemed immoral by the Klan have in fact been criminalized as misdemeanors (miscegenation, prostitution, gambling, and alcohol consumption). While the KKK has been prominently featured in historical accounts of racist and nationalistic violence, less attention has been given to its active role as rule enforcers for conservative morality as an extension of the group's support of Protestant fundamentalism. After World War I, membership in the Klan swelled in response to social anxiety about ongoing chaos in social relationships with African-Americans, Catholics, Jews, labor radicals and liberals. In addition, a new 'enemy' was identified in the guise of white men and women who deviated from the Klan's moral code. Nancy Maclean argues that the Klan could be petitioned on behalf of wronged victims of immorality to reinforce the boundaries between right and wrong by enforcing sanctions against the immoral offenders. Thus,

Maclean cites a woman whose husband was both an alcoholic and a gambler, who refused to grant her a divorce with alimony and instead slandered her with an accusation of adultery. The woman was therefore unable to financially care for their two children, and in response, petitioned the KKK for help in obtaining the justice denied to her by the courts. This was a routine role for the Klan of the 1920s, and historical documents show that a preponderance of Klan violence at that time period in history focused on offenders who were atypical in race (white) and gender (women). For example, white women in Macon County, Georgia, were frequent targets of the Klan for alleged consumption of alcohol and sexual promiscuity.

Finally, social science research demonstrates that moral enterprises (where specific behaviors, beliefs, and conditions are linked to moral meanings) are often sufficient to confer status to a group relative to other groups without the need for rule enforcement or punishment for rule violations. Thus, misdemeanor crimes by their very existence and encoded presence in law are sometimes sufficient to the needs of the moral entrepreneurs. No official sanctions are necessary. Specifically, Joseph Gusfield argues that laws can exist to perform either instrumental or symbolic functions. Instrumental functions of law serve to modify behavior. Thus, a law is enacted and when enforced, behaviors change. An example can be found in the seatbelt laws that were legislated in Texas during the mid-1980s. When the initial seatbelt law was approved, few Texans modified their practices with regard to buckling their seatbelts. In order for more modifications of behavior to occur, citizens required enforcement of the new law. When the Texas Department of Public Safety allocated monies for the enforcement of the new law, and subsequent and steep financial penalties were levied against violators, most drivers began to buckle their seatbelts. The enforcement of the law provided an instrumental effect. Without enforcement, no instrumental change to behaviors would have occurred. Gusfield states that laws often are passed without provision for enforcement. These laws he calls symbolic laws – symbolic because they were passed at all. No enforcement is intended, but the meaning behind the codification of the rule violation is a status conferring one. A group has its norms elevated to the status of law and is thus of higher status than those of other groups. Gusfield says that enforcement of the symbolic law would destroy the function of this type of law. Thus, during the U.S. Prohibition Movement, laws against drinking were predominantly symbolic laws with little or no monies provided to fund their enforcement. Gusfield argues that this is because the group that advocated for prohibition was the group that held status and power, and the Temperance Movement was designed to target a behavior found among newer immigrants to the United States, whose increasing population size was beginning to threaten the power found in maintaining the status quo. Yet, if laws against drinking were enforced, assimilation of groups would have resulted and meant the disappearance of the ruling class. Symbolic laws (laws without enforcement) were therefore enacted instead. Many social theorists have suggested that the purpose of a misdemeanor crime category is to harbor the products of the activities of moral entrepreneurs who have engaged in moral crusades.

Levels of Crime: Felonies, Misdemeanors, and Infractions

Crimes are classified according to their severity, with the particular classification influencing both the substance and the procedure of a criminal charge. The mildest crimes are known as infractions, and the most serious are felonies. Misdemeanor violations of the law fall between felonies and infractions, in terms of both severity and punishment. Infractions, also referred to as petty offenses, are not considered crimes in some states. However, even in jurisdictions that treat infractions as criminal offenses, incarceration is not usually the punishment. When it is, confinement occurs in a local jail. Most persons convicted of a felony and sentenced to incarceration are sent to state prisons, or, in the case of a felony, to federal prison. Misdemeanor offenses, less serious than felonies but more serious than infractions, are generally punished by a fine and/or incarceration in a local jail, which generally also house people awaiting trial. With very few exceptions according to state, people convicted of misdemeanors do not lose any civil rights, as opposed to convicted felons. However, there are circumstances in which misdemeanor conviction can interfere with job prospects and certain civil privileges; for example, a school bus driver receiving a first DUI who may lose her driver's license.

Felonies

Felonies are the most serious class of criminal offenses throughout the United States. Many jurisdictions separate felonies into their own distinct classes, with repeat offenders convicted of committing heinous felonies receive more severe punishments than first-time offenders convicted of committing less cruel or injurious felonies. Depending upon the circumstances, felonies are punishable by a fine, imprisonment in excess of 1 year, or both. Historically, in common law, felonies were offenses that violated the moral standards of a community. Today, many crimes classified as felonies still offend the moral standards of American communities and include such offenses as treason, terrorism, murder, rape, kidnapping, robbery, arson, and burglary, among others.

Infractions

An infraction is a violation of an ordinance, a municipal code administrative regulation, and state or local traffic rules. In states that do not criminalize infractions, such crimes are treated as civil offenses. Similar to misdemeanors, infractions are often defined in broad language. Although incarceration is rare, when it does occur confinement is limited to serving time in a local jail.

Misdemeanors

A misdemeanor is a criminal offense for which a fine and/or a sentence to a term of imprisonment not in excess of a year may be imposed, with confinement occurring in a city our county jail rather than a state or federal prison. Misdemeanors

are classified differently in various states and classifications are based upon level of severity of the crime. All states as well as the federal government classify misdemeanors into Classes A, B, and C, high or gross, ordinary, and petty. High or gross misdemeanors are defined in some states as any crime that is not a felony or an ordinary or petty misdemeanor.

Class A misdemeanors are the most serious and carry fines of up to $4000, along with jail terms of up to 1 year in the county jail. The U.S. Code establishes a sentence of 6 months to 1 year in jail, although some state laws vary in fines and jail terms. Examples of Class A misdemeanors include second DUI charges, assault charges, vehicular theft, and possession of an illegal weapon. These misdemeanors are not easily expunged.

Under the U.S. Code, Class B misdemeanors carry 30-day to 6-month jail terms and fines of up to $2000. State codes vary in sentencing, with one state assigning jail sentences of 6 months or less and fines of $500. In this same state, prostitution is considered a Class B misdemeanor. Other examples of Class B misdemeanors include first-time DUI offenses, criminal trespassing, petty thievery in the amount of $50–500, and law violators who evade arrest.

Class C misdemeanors carry jail terms of 5–30 days under the U.S. Code and not more than 30 days in jail with fines of $50 or less under most state laws.

Misdemeanor convictions usually do not result in the loss of civil rights, but they may result in the loss of privileges, such as professional licenses, public office, or public employment, known as *collateral consequences of criminal charges*. Such sanctions are more common when the misdemeanor is related to the privilege in question, such as the loss of a driver's license by a taxi driver convicted of reckless driving. Another example of collateral consequences related to misdemeanor conviction involves moral turpitude – generally evaluated on a case-by-case basis. A prominent example is found in the U.S. Constitution, which provides that the President of the United States may be impeached and subsequently removed from office is found guilty by Congress for 'high crimes and misdemeanors,' the definition of high crimes being left to Congressional discretion.

Unclassified Misdemeanors

In the United States, when a statute does not specify a class for misdemeanors, they are referred to as *unclassified*. Such laws are enacted by legislators when they wish to impose penalties falling outside the framework specified by each class. First-time marijuana possession is an unclassified misdemeanor in some states and is punishable by up to 30 days in jail rather than the usual fines and jail sentences of the classified misdemeanors. Some states categorize misdemeanors as Class A, Class B, and unclassified.

Eleven Types of Misdemeanors

1. *Simple Assault* – Assault is an intention threat to do violence to another person. While often confused with *battery*, which is the actual touching or striking of another person, assault involves verbal threats or actions. Assault and battery involves the threat of another's safety combined with actual unwanted physical contact. Simple assault is a common misdemeanor. It is defined as assault without the intent of injury. However, exceptions are simple assault on police officers, elected officials, and social workers, which constitute felonies. Simply invading another's personal space is a simple assault.
2. *Petty Theft* – Petty theft is an umbrella term encompassing larceny, stealing, embezzlement, and any other instance of taking someone's property against that person's will. In most jurisdictions, petty theft is identified by a value amount of $500 or less. While stealing property from someone's home is petty theft, shoplifting or stealing from one's employer is also petty theft. The charge of petty theft may be elevated to a felony if the offender has a criminal history involving prior thefts. *Larceny* refers to the wrongful taking property and occurs when, with the intent to deprive the owner, an individual appropriates, obtains, or withholds the property from the owner. Punishment for the crime of larceny varies in severity based upon the value of the wrongfully taken property. Grand larceny involves property of larger scale, as opposed to petit larceny of small-scale property.
3. *Solicitation* – Solicitation occurs when one person attempts to persuade another to commit a crime. The most common form of solicitation is prostitution, the encouragement by one person of another to pay for sex. It is also a misdemeanor to procure customers for a prostitute, known as *pandering*. Another form of solicitation is *panhandling*, a request made in person for an immediate donation of money or other reward. Panhandling can occur through passive manipulation of emotions (especially guilt) or aggressive intimidation tactics. During the Great Depression of the 1930s, and in other cultures around the world, such actions as panhandling (or begging, as it is more commonly known) are considered socially appropriate and even honorable for the context. For example, religious requests for financial support of the ministerial class who do not work for pay are accepted by most social observers across cultures. Some U.S. jurisdictions now criminalize panhandling as misdemeanors, though social theorists and political commentators argue against the practice as one which unevenly discriminates against the homeless.
4. *Trespassing* – Trespassing occurs when a person unlawfully intrudes on another's property. Trespassing laws vary among jurisdiction but are commonly categorized as misdemeanors. It is common to see 'no trespassing' signs in places such as schools and private hunting grounds. Whether the property is public or private, going onto another person's property without their knowledge or consent is trespassing, whether or not damage is done. Entering another person's home does not fall into the purview of trespassing; rather, such a violation is *breaking and entering*. Exceptions to trespassing laws include police, meter readers, and government surveyors.
5. *Arson* – Arson is the intentional damage of property by fire. Setting fire to any building or structure, whether public or private, is a misdemeanor. Burning down a home or business for the purpose of making an insurance claim is arson as well as insurance fraud. Misdemeanor arson

requires that the act is malicious or reckless, regardless of motive. If a house is vacated, unfit for human habitation, or closed up, its burning is not arson. If a dwelling is burned under the impression that it is abandoned, but someone actually lives in it, then arson has been committed. Accidental fires are not considered arson.

6. *Contempt of Court* – Contempt of court involves the act of defying authorities during court proceedings. A contempt of court charge may be imposed upon anyone in the courtroom, including attorneys, witnesses, jurors, court officers, or anyone who interrupts the proceedings. *Civil contempt of court* occurs when someone refuses to abide by a judge's decision in a case. An example is a child support case in which one parent refuses to pay the amount mandated by the court. An example of *criminal contempt of court* occurs when an individual physically interrupts a court hearing by shouting at court officials or witnesses.
7. *Stalking* – Stalking as a misdemeanor crime involves following or harassing a person continuously and causing that individual emotional distress. Any unwanted intrusion or communication by a person to another person falls within the definition of misdemeanor stalking and is a form of mental assault in which the intruder breaks in the lifeworld of the victim. Stalking behaviors include physically following someone, making harassing phone calls or text messages, or electronically targeting victims over the Internet and through monitoring software, hidden cameras, microphones, and GPS tracking units, all of which can pose a real threat to another's safety and perception of well-being. Stalking or harassing another is typically sanctioned as a misdemeanor crime for first offenses, but subsequent stalking convictions are usually punished as felonies.
8. *Possession of Illegal Substances* – Drug possession involves having illegal drugs in one's pockets, purse or wallet, or vehicle. Small amounts of illegal drugs for personal use usually result in misdemeanor charges, while larger amounts for the purposes of distribution constitute felony drug possession.
9. *Driving Under the Influence* – Driving under the influence involves operating a vehicle while intoxicated from alcohol or while under the influence of any illegal substance. Convictions for DUIs can result in the loss of one's driver's license, mandated community service, jail time, and fines. Those found guilty of DUIs with a history of other similar offenses often experience harsher penalties than do first-time offenders.
10. *Criminal Mischief* – This misdemeanor charge is applied to individuals who knowingly and intentionally injure, damage, move, deface, tamper with security devices or destroy any property of another or public property without consent. Security devices include fire extinguishers, fire hoses, fire escapes, emergency exits, emergency escape equipment, life preservers, life rafts, alarms, or devices intended to warn of danger or emergency, such as guard rails, traffic signals and signs, or railroad grade crossing signs. Criminal mischief also occurs when without consent an individual tampers with the property of another or public property for the purpose of endangering or interfering with another person or that person's property or use of the property. Criminal mischief includes destruction of property such as breaking windows or altering property through graffiti. It also includes the intent to impair the functioning of any computer, computer system, computer network, computer software, or computer program, through such actions as computer hacking, alteration of a computer, computer system, or computer data or software contained in the computer. To be considered a misdemeanor, damage to the property must not exceed $1000. Property damage in excess of $1000 is categorized as a felony.
11. *Indecent Exposure* – Because the term 'indecent' is difficult to define, indecent exposure is the misdemeanor with the most variability across jurisdictions. For example, in Washington State, indecent exposure has been defined as intentionally exposing one's genitals with the knowledge that the action can cause reasonable alarm. Often indecent exposure is committed for the sexual gratification of the offender or committed to entice a sexual response from another. Exposing one's genitals does not include baring the female breast (which, if considered indecent exposure, would interfere with breast-feeding mothers). However, exposing oneself to a minor under the age of 14 years qualifies as gross misconduct, a felony. Additionally, indecent exposure reaches the level of sexual assault if any physical contact is made. In most cases, the first conviction for indecent exposure subjects one to misdemeanor penalties like a fine or a few months in the county jail. Second offenses would likely rise to felony status, leading to a state prison sentence. In some states, any indecent exposure conviction subjects the offender to a lifetime mandate of registration as a sex offender.

Not Petty Consequences

Because misdemeanors are considered petty offenses, data about them are sparse. Some states, like California, do not count misdemeanor convictions at all. Because the massive misdemeanor legal apparatus is very informal and overcrowded, petty offenders routinely are denied counsel, even when they are constitutionally entitled to an attorney. When they do have access to an attorney, the process is referred to as 'meet them and plead them lawyering.' This reflects the reality that once people charged with misdemeanors get to court, they often are pressured by players in the system to plead guilty, without full knowledge of their civil rights or the nature of the charges against them. If the accused pleads 'not guilty,' bail is set, with approximately 75% of defendants unable to pay. Consequently, innocent defendants commonly plead guilty just to get out of jail. Millions of Americans are thus punished without due process and soon learn that, at least when it comes to minor offenses, the law and evidence are not part of the process.

According to one argument, minor misdemeanor crimes may not actually matter too much. Since guilty pleas by defendants typically result in a fine or probation (not jail time), perhaps these minor crimes should be handled in the quickest,

cheapest way without counsel or due process, especially given the deplorable lack of resources system wide. Indeed, petty offenders may actually get out of jail sooner (if sentenced to jail time) if they plead guilty. Moreover, it is widely assumed that millions of these defendants actually are guilty, so rushing them through the system will not result in much of a miscarriage of justice.

However, it must be pointed out that these petty convictions are not minor. These offenses are punished with fines that some individual defendants simply cannot pay. Additionally, convictions of any type can hurt job prospects; affect eligibility to obtain professional licenses; negatively impact child custody; threaten public housing; jeopardize food stamps, student loans, and health care; or lead to deportation.

See also: Crime and Class; Crime: Knowledge about and Prevalence; Law and Society; Opiate Use and Abuse, History of.

Bibliography

Becker, Howard S., 1963. Outsiders: Studies in the Sociology of Deviance. The Free Press, New York.

Becker, Howard S., 1995. Moral entrepreneurs: the creation and enforcement of deviant categories. In: Herman, Nancy (Ed.), Deviance: A Symbolic Interactionist Approach. General Hall, New York.

Gusfield, Joseph R., 1986. Status Politics and the American Temperance Movement. University of Illinois Press, Chicago.

Maclean, Nancy, 1991. While women and Klan violence in the 1920s: agency, complicity and the politics of women's history. Gender & History 3 (3), 285–303.

Reinarman, Craig, 1994. The social construction of drug scares. In: Adler, Patricia A., Adler, Peter (Eds.), Constructions of Deviance: Social Power, Context, and Interaction. Wadsworth Publishing, Belmont, CA.

Siegel, L., 2012. Criminology: Theories, Patterns, and Typologies, eleventh ed. Wadsworth, Belmont, CA.

Vito, G., Maahs, J., Holmes, R., 2007. Criminology: Theory, Research, and Policy, second ed. Jones and Bartlett Publishers, Sudbury, MA.

Zellner, William W., 1995. Counter Cultures: A Sociological Analysis. St. Martin's Press, New York.

Mismatch on the Labor Market

Kostas Mavromaras and Peter J Sloane, National Institute of Labour Studies, Flinders University, Adelaide, SA, Australia and IZA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Separate sections of the literature deal with search behavior – or the process through which job matches occur – and skill mismatch, which is concerned with outcomes and includes qualifications mismatch, skills mismatch, and skills obsolescence. Causes of mismatch may reflect the qualities of job seekers, their preferences, or their membership of particular groups. Consequences may manifest themselves in terms of wage or productivity adjustments. Recent work using panel data has allowed some of these effects to be identified more clearly. The main policy implication is that attention should focus on skills mismatch rather than on qualifications mismatch.

Introduction

The term mismatch covers disparate areas of the literature. One part of this is largely concerned with job search and focuses on the process of job filling or job change. Another strand focuses on skill mismatch and is largely concerned with the outcomes resulting from this process. When the labor market is subject to severe adjustments, such as those resulting from skill-biased technological change or the hollowing-out phenomenon in which medium-skilled jobs are automated and labor is displaced, some mismatch is inevitable – some workers find themselves with the wrong skills and employers search for those with the, desired, new skills. An alternative approach, adopted by Sattinger (2012), is to distinguish between short-run mismatches occurring between workers and jobs, and long-run aggregate mismatches which arise from shifts in supplies of and demands for workers. In what follows, we first give a brief account of the job-search process and then consider the skill-mismatch literature in some detail.

Job Search

Short-run mismatches can occur as a result of heterogeneity in terms of both jobs and workers, combined with imperfect information and labor-market frictions. Search will then be costly for both employers and those seeking employment. As Pissarides (1990) and Petrognolo and Pissarides (2001) put it, when searching for a job a worker does not know which firms will be willing to make a job offer. Then, the job seeker has to decide whether to accept an offer, depending on their reservation wage, or to continue searching. An employer may fill a job vacancy before the best available candidate is found, because it is expensive to have an unfilled vacancy. Similarly, an employee may accept a job before the best vacant job is found. Under such circumstances, perfect matching will be rare. This exchange process can be summarized by a well-behaved function that expresses the number of jobs formed in any period in terms of the number of workers looking for jobs, the number of employers looking for employees, and a small number of other variables. These variables will be related to the stage of the business cycle in which we are. The simplest form of the matching function is

$$M = m(U, V) \qquad [1]$$

where $M =$ the number of jobs formed during a given time period, $U =$ the number of unemployed workers looking for work, and $V =$ the number of vacant jobs.

A steady state relationship between the unemployment rate and the vacancy rate can be derived from [1]. Let U and V be the number of unemployed workers and job vacancies respectively, and N and L be the level of employment and the labor force respectively, so that $L = (N + U)$. Define the unemployment rate $u = U/L$ and let the vacancy rate be $v = V/N$. Assume that the job-separation rate is z, so that total separations are $S = zN$. Then, imposing constant returns to scale on m (.) and noting that in the steady state the number of matches M equals the number of job separations S, we can obtain the Beveridge curve as in **Figure 1**.

Equilibrium occurs when the number of vacancies equals the number of unemployed workers – on the 45° line from

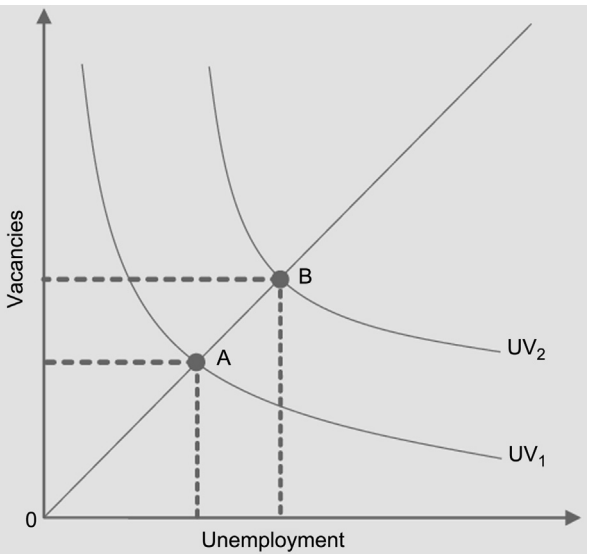

Figure 1 The Beveridge curve.

the origin. If, however, the economy runs into a recession the number of unemployed workers will exceed the number of job vacancies to the right of the 45° line or, in other words, labor supply will exceed labor demand. If the economy overheats, the number of vacancies will exceed the number of unemployed workers to the left of the 45° line, that is, labor demand exceeds labor supply. If there is an increase in structural imbalances in the economy, the whole Beveridge curve may shift outwards, say from A to B. Supply and demand will be in balance on B, but frictional and (or) structural unemployment will be higher. With excess demand there will be labor shortages and with excess supply labor surpluses.

What constitutes a labor shortage is not, however, straightforward. An employer who cannot find sufficient workers because the wage has been set too low is not suffering from a genuine shortage. Further, skill gaps in which the existing employees lack the appropriate skills to do the job properly require different measures to overcome the problem, such as additional training. In some instances skill shortages are the manifestation of buoyant demand and are associated with increasing sales in the sector. In other cases they are the manifestation of falling demand where skilled workers are leaving a declining sector, faster than the sector declines as shown in Mavromaras et al. (2013c). How employers respond to skill shortages is also not straightforward. Healy et al. (2011) use a large longitudinal sample of Australian small and medium size enterprises (SME) to show that the most common employer responses to experiencing skill shortages are longer hours by existing staff, subcontracting, internal training, and reductions in production. Paying higher wages to attract the needed skilled workers is only used by 16% of all employers.

Hiring standards may be adjusted because of skill shortages according to the stage of the business cycle and the state of the labor market. When demand is buoyant, employers may be forced to take on workers who are somewhat lacking in the personal traits that employers would normally demand. In contrast, when demand is depressed and labor is in plentiful supply employers may raise their expectations and seek qualities beyond those which are technically required. It is in times of labor-market tightness that employers may be forced to hire workers who are underqualified or underskilled. In contrast, when the labor market is slack overqualified or potentially overskilled workers may be obliged to apply for jobs beneath their capabilities. Cross-sectional analyses may fail to pick up these subtleties of the business cycle. Nevertheless, there is information regarding disequilibrium manifestations of mismatch as proposed by Mavromaras et al. (2013c), who show how combinations of complementary short-length time series data on education and employment outcomes can be used to monitor skills disequilibrium at the level of occupations for the whole labor market.

Skill Mismatch

It is now well established that many workers are employed in jobs for which their current qualifications are not a requirement, while others lack the qualifications that are required of current recruits. In addition, some workers possess skills and abilities which are underutilized in their current job, while others struggle to cope because, through lack of training, they do not possess the necessary skills. We use skill mismatch as a generic term to cover both forms of mismatch (in qualifications and in skills); the term skills mismatch refers to the narrower concept which includes skills and abilities, but excludes qualifications mismatch. The starting point for this branch of the labor economics literature was Richard Freeman's book entitled *The Overeducated American*, which first made use of the term overeducation, meaning that some graduates could not obtain jobs which required a degree or, as Freeman put it,

> The college job market underwent an unprecedented down turn at the outset of the 1970s, with young graduates just beginning their careers most severely affected. Real and relative earnings of graduates dropped, employment prospects and occupational attainment deteriorated, and large numbers were forced into occupations normally viewed as being below the college level. For the first time in recent history the economic value of an investment in college education fell, though with considerable variation among professions and groups.
>
> Freeman, 1976, p. 184

Freeman noted that the above outcome was not anticipated by conventional economists working in the human capital or economics of education field. Despite the fact that many of Freeman's most pessimistic predictions for the future of higher education did not come to pass, his work stimulated a stream of papers that consider mismatch across several dimensions, covering different time periods and countries, and extending to all classes of labor.

By any standards, when we look at the labor market, the extent of mismatch is considerable. Indeed, Slonimczyk (2013) shows that a substantial fraction of the increase in overall inequality in the United States over the period 1973–2002 can be attributed to an increase in mismatch rates and mismatch wage premia. For example, around 20% for males and 48% for females of the increase in the Gini coefficient can be attributed to surplus qualifications. In **Table 1** we summarize the incidence of vertical mismatch (that is, when the level of the worker's qualifications is different from the level usually required for the job) in relation to qualifications, constructed from existing studies.

In addition to vertical qualifications mismatch there may be horizontal mismatch in which the level of qualification is appropriate, but the type is not. Its measurement depends on

Table 1 Qualifications mismatch

	Overqualification (% employed)	*Underqualification (% employed)*
Asia	26	21
Australia	20	10
Europe	30	31
Latin America	24	21
US/Canada	37	16

Source: Leuven, E., Oosterbeek, H., 2011. Hanushek, E., Welsh, F. (Eds.), Handbook of the Economics of Education, vol. 4, Elsevier Science, pp. 282–386, and for Australia, Black, D., 2013. The Utilisation of Human Capital from Education in Australian Labour Markets. Over-Education? Ph.D. thesis, University of Melbourne.

the identification of occupations that correspond to each field of study, but in some cases the relationship will be imprecise. Nonetheless, Randstat (2012) found that 23% of workers in EU27 were affected by horizontal mismatch in 2009. Data from the fifth wave of the European Working Conditions Survey (EWCS) 2010 suggest that 31% of employees are overskilled, a similar proportion to those who are overqualified, while13% are underskilled. In all possible types of mismatch, the least satisfactory case seems to be a situation where employees are both overqualified and overskilled (see below). According to EWCS 2010, about 35% of the overqualified are also overskilled, so that a third of the overall incidence of overqualification falls into the least satisfactory category. A much less-researched form of mismatch is skill obsolescence, under which skills depreciate over time as the result of either changes in technology or the effects of the aging of the worker. According to a 2012 CEDEFOP pilot survey in four countries (Finland, Germany, Hungary, and the Netherlands), skill obsolescence affects 12% of workers.

There are measurement issues arising in relation to these estimates. The assumption is made that there is a particular educational requirement for a particular job. This ignores the possibility that one form of human capital may be substitutable for another. Thus, while undereducated workers may lack education, they may make up for this by having more experience on the job. The estimates are also dependent on the availability of appropriate data sets. Three types of measurement have been conducted in relation to qualifications mismatch:

1. Systematic job evaluation may be utilized to determine the precise level of qualifications required to perform a particular job, which is sometimes referred to as an objective measure. This is not commonly used, however, perhaps because the appropriate educational level may change as the nature of jobs changes over time, and the required educational level may lag behind.
2. Worker self-assessment, often referred to as a subjective measure, depends on the nature of the questions asked, such as what level of education is required to obtain the job you currently hold, or what level of education is required to perform the job you hold? If the answer to the first question implies a higher level of education than the answer to the second question, this is consistent with credentialism on the part of the employer, which might occur in a slack labor market with plenty of job applicants. Credentialism can be defined as a situation in which the level of education required to *obtain* a job exceeds the level of education required to *perform* the job adequately. This rests on a belief held by the employer that the possession of certificates and diplomas implies higher productivity on the part of the employee (sometimes referred to as a sheepskin effect).
3. Many data sets do not contain a direct question on qualifications mismatch in which case reliance has to be placed on the so-called empirical method. In such cases qualifications mismatch is frequently measured as the situation in which the level of education is more than one standard deviation above or below the mean, which will only pick up cases where mismatch is substantial and therefore cannot be compared with the two measures above. Further, as the degree of overeducation and undereducation is often asymmetric (as in **Table 1**), it is more appropriate to use deviations from the mode rather than the mean. Thus, the use of the mode is increasingly being adopted.

A number of studies have attempted to measure the correlation among the three measures above and find it to be low, with different people being identified as mismatched according to the measure used. Thus, Verhaest and Omey (2006) compared the reliability of five different measures of overeducation on the basis of encompassing tests and concluded that the objective measure and a subjective measure based on the level of education needed to do the job should always be preferred to both empirical methods based on mean or modal education and a subjective measure based on the level of education needed to get the job. Studies rarely consider the severity of mismatch – largely due to the lack of data – but a degree holder working in a job for which a high-school certificate is acceptable should not cause as much concern as a degree holder working in an unskilled job.

The equivalent of the empirical method for measuring qualifications mismatch is not available for skills mismatch, and analysis is totally dependent on direct questions being asked of employees in surveys, such as 'to what extent are you able to utilize the skills and abilities you possess in your current job?' Answers are often given on a seven-point scale. Desjardins and Rubenson (2011) suggest that although qualifications mismatch has the advantage of being easier to measure and is broader in its coverage of skills, it also has the disadvantage of being less precise because it does not allow for differences in the quality of similar qualifications, such as degree classification or the reputation of the institution where the qualification was obtained. Further, no allowance is made for the possibility that there may be skill gain or loss after the date at which the qualification was obtained. Skills obsolescence can arise for a number of reasons, which means that it may not be possible to capture it in a single question. Workers may be asked if there are any skills that they previously possessed which are no longer required; or are there any new skills in the current job that have emerged, but which they do not possess; or whether they consider that their ability to perform their current job has diminished over time.

Relationship of Mismatch Models to Existing Theories

The naïve Mincer human capital model is basically a supply-driven model which implies that earnings are determined just by the personal characteristics of the worker, so that precisely where an individual is employed is immaterial to the level of earnings:

$$\text{Log } W_i = a + bS_i + cX_i + \mathrm{d}X_i^2 + e_i \qquad [2]$$

where W denotes the level of wages, a the constant term, X the years of experience, e the error term, and subscript i the ith individual.

At the other extreme is the Thurow (1975) job competition model which assumes that productivity is determined

exclusively by the job, rather than the worker, with superior personal characteristics simply serving to push the job applicant to the front of the job queue. Thus, in this demand-driven model, all workers within a particular job will be paid the same regardless of their qualifications. Thus,

$$\text{Log } W_i = a + b_1 q^r + e \qquad [3]$$

where q^r is the *minimum* qualifications required to obtain the job.

Between these two extreme models lies the Sattinger (1993) job assignment model, which is based on the belief that there is an allocation problem in the assignment of heterogeneous workers to jobs which vary in their complexity. Let S_i equal the extent to which skill i is required by a particular employer, and t_i is the extent to which this is possessed by particular employees. Then, on the demand side of the labor-market employers will collectively specify a frequency distribution M (s_1, s_2, … , s_m) of the number of employees they require at various skill levels. Similarly, on the supply side of the labor market, potential employees will be available with a frequency distribution of skills N (t_1, t_2, …, t_n). Therefore, if

$$M(s_1, s_2, \ldots, s_m) \neq N(t_1, t_2, \ldots, t_n) \qquad [4]$$

overeducation or undereducation will ensue, and might well be a persistent problem if both distributions are subject to continual change, which seems likely in a dynamic economy. In fact, the assignment model encompasses both the human capital model and the job competition model as special cases. Imposing an appropriate set of restrictions and using an F test on the residual sum of squares, both the human capital and job competition models can be rejected in favor of the encompassing job assignment model on the basis of different data sets (see Hartog, 1986; Hartog and Oosterbeek, 1988; Sloane et al., 1999).

A more general model in the spirit of the job assignment model, which has been used by a large number of economists subsequently, was implemented by Duncan and Hoffman (1981) and is frequently referred to as the ORU (over-, required- and under-education) model. This decomposes qualifications into those required to do the job (q^r), those surplus to requirements (q^s), and those which are lower than currently required (q^u). Thus, in this case

$$\text{Log } W_i = b_0 + b_1 q^r + b_2 q^s + b_3 q^u + e \qquad [5]$$

With q^r set as the default, the expected sign on q^s is positive, and that on q^u negative. According to the naïve Mincer human capital model, it must be that $b_1 = b_2 = b_3$ and returns are not influenced by the circumstances of employment. In contrast, in the case of the job competition model it must be that $b_2 = b_3 = 0$, as only required education determines pay in this case. It should be noted that mismatch is only a serious problem in the long run if the relative supplies of different types of education do not affect the skill composition of final demand. This would amount to there being fixed skill requirements, so that employers do not respond by changing their methods of production to accommodate the changed supply conditions. The ORU approach has been criticized by Leuven and Oosterbeek (2011) on the grounds that the model may be subject to both omitted variable bias and measurement error, neither of which is satisfactorily dealt with in much of the literature.

Brief mention may be made of two alternative approaches to the above. First, Sicherman and Galor's theory of career mobility (1990) suggests that individuals may accept a lower-level job in order to invest in a form of training and recoup a return through accelerated promotion in the future. In this context, mismatch is a purely temporary phenomenon. Second, Frank's theory of differential overqualification (1978) is based on the proposition that for married couples the search for a pair of jobs will be constrained geographically, and it will rarely be possible to optimize the position of both partners. Under the male chauvinist model, the husband will optimize his employment and the wife will be constrained either as a tied stayer in the current location or a tied mover in a new location, so that married women may face a greater degree of overeducation than do men or single women.

Causes of Mismatch

A key issue is to what extent mismatch is explained by heterogeneity in terms of labor quality, or by workers trading off attachment to jobs which do not require their level of qualification for positive job attributes, such as an easier pace of work, locational advantages, or preferred hours. Either of these possibilities would not imply that mismatch must represent a form of market failure. A number of studies have found that those with better degree classifications and from more established universities are less likely to be overeducated. Also, those graduating in more professional disciplines such as engineering or law are less likely to be overeducated than are those with more general degrees such as arts. On the demand side, the higher the occupation or the greater the size of firm the greater the probability of a match. One way to cast some light on these issues is to disaggregate the overeducation variable by distinguishing between those who are overeducated but have high job satisfaction and those who are overeducated but have low job satisfaction. Thus, Chevalier (2003) and Chevalier and Lindley (2009) classify the former as apparent overeducation, and the latter as genuine overeducation. Compared to matched graduates the apparently overeducated were found to have a wage penalty of 7% and the genuinely overeducated a penalty of 23%. Including various measures of skill reduced the penalty by between 2 and 5 percentage points, but did not eliminate it. A similar approach was adopted by Green and Zhu (2010) who distinguish between real and formal overeducation according to whether or not overeducation is accompanied by skill underutilization. They find that those in the real overeducation category suffer from higher wage penalties than do those in the formal overeducation category; and as well, they experience considerably lower job satisfaction.

Individual characteristics such as membership of discriminated against groups such as the young, the elderly, females, ethnic minorities, and the disabled may give rise to mismatch. Young workers may be particularly prone to mismatch as new entrants into the labor market, since they lack work experience; older workers may suffer from mismatch as a consequence of skill obsolescence, as outlined above. Several studies have confirmed that mismatch is greater for members of minority groups and immigrants. There is no strong evidence, however,

to support the suggestion that women are more likely than men to suffer from mismatch. The case of mismatch in the context of workers with disability was investigated by Jones et al. (2014) who do not find any relationship between work-limiting disability and overskilling, but find a positive relationship between work-limiting disability and overeducation.

There may be cohort effects if individuals qualify for entry into the labor market at a time when there are more graduates or when the market is depressed, so that there is an excess supply of qualified applicants. This may cause some to accept lower-level jobs than they are qualified for and, in turn, force others at that level to take a similar step, so that there is a bumping down phenomenon in which appropriately qualified workers are forced out of current jobs by those with higher qualifications, and are then forced into lower-level jobs or out of employment altogether.

Consequences of Mismatch

A large number of studies have been undertaken using cross-sectional data and have found that labor-market mismatch – whether in the form of qualifications or skills – is associated with negative labor-market outcomes in the form of lower wages, reduced job satisfaction, and higher labor turnover. Sicherman (1991) reported two stylized facts in relation to wages. First, overeducated workers are paid less than if they were matched, but more than their matched coworkers so that while there is a wage penalty for employees, employers pay a premium for employing them. Second, in contrast, undereducated workers are paid more than if they were matched, but less than their matched coworkers. They, therefore, benefit by being paid more than if they were in a job commensurate with their qualifications, and there is an incentive to remain in their current employment. These results have been confirmed in a large number of subsequent studies, but virtually all of these are based on cross-sectional data and, therefore, may be biased due to the problem of unobserved individual heterogeneity. Sattinger and Hartog (2013) and Sattinger (2012) show – while making some relatively mild assumptions – that the above outcome can be explained by Nash bargaining between workers and employers.

An important question is why do employers hire overeducated workers and pay them a premium wage? Not a great deal is known about this because we lack employer surveys with relevant questions; according to the human capital approach, however, higher earnings must reflect higher productivity. In imperfect markets, however, the link between wages and productivity is imprecise with firms possibly obtaining a rent from the services of their employees. Few attempts have been made to measure this directly using a production-function approach. One exception is Tsang (1987), who used a Cobb–Douglas production function in a case study of the Bell Corporation in the United States and found evidence of a negative productivity effect. In contrast, a positive effect was found by Kampelmann and Rycx (2012), using a linked employer–employee panel for Belgium, covering the period 1999–2006. They measure value added per worker and control for simultaneity, time invariant workplace characteristics, cohort effects, and dynamics. Based on the ORU specification, it appears that a higher level of required education has a significant positive effect on firm productivity, and additional years of overeducation have a similar effect. However, additional years of undereducation among young workers have the reverse effect. There is no relationship with profits, suggesting the wage and productivity effects cancel out one another.

The most common methodology used in applied econometrics for controlling for potential bias caused by unobserved individual heterogeneity is the use of panel data, which consists of time series of cross-section data. Earlier examples of this literature are Mundlak (1978) and Deaton (1985), with later comprehensive expositions found in Hsiao (2003) and Arellano (2003). Examples of applications in the area of mismatch include Bauer (2002) using the German Socio-Economic Panel, Tsai (2010), using the U.S. Panel of Income Dynamics, Quintini (2011b), using the European Community Household Panel, which covered the period 1994–2001, and Mavromaras et al. (2013a) using the Household, Income and Labour Dynamics in Australia (HILDA) panel data. Each of them finds that the mismatch wage penalty is much reduced when controlling for unobserved individual heterogeneity. The last of these is particularly revealing, as the authors analyze mismatched workers according to whether they are matched, overeducated, overskilled, or a combination of the two. They present two types of estimation – pooled cross-section, as a comparison point with the earlier literature, and random effects probit with Mundlak corrections. They find that there is a significant pay penalty in their male sample only for those who are both overskilled and overeducated. Further, job satisfaction is not reduced by being overeducated, but is clearly reduced by overskilling, whether on its own or when jointly present with overeducation. With the exception of the case where overskilling and overeducation are present jointly, mismatch has no significant effect on the quit rate. These results overturn much of the previous literature, but without similar analyses for other countries we cannot say whether they apply outside Australia. Quintini (2011b) explores the effect on mismatch of involuntary separations at different points of the business cycle. When the unemployment rate is in line with its medium-term average, losing one's job because of a business closure increases the likelihood of overqualification in the next job by 15% compared to quitting, but if the unemployment rate becomes twice as large as its average rate this effect increases to 35%. She does not, however, interact overqualification and overskilling as in the Mavromaras et al. (2013a) study.

The extent to which mismatch is a temporary phenomenon is important from the policy point of view, but one on which there is no consensus. Mavromaras and McGuinness (2012) and Mavromaras et al. (2013b) use a random effects dynamic probit model to estimate the persistence of overskilling and the effect of overskilling dynamics on wages, using the HILDA survey data. They show that being overskilled is associated with lower wages and more overskilling at the population level, and that becoming overskilled causes lower wages and a higher probability of future overskilling at the individual level. They find that overskilling is highly persistent in a manner that is inversely related to educational level. Furthermore, the wage penalties are also persistent, which magnifies

the scarring effects of overskilling. Where exits from overskilling do occur, Quintini (2011b) suggests that they are most likely to derive from improvements in the match between workers and jobs through either a job change or job upgrading, such as by giving the worker more responsibilities. In contrast, overqualified individuals can only become well-matched by changing jobs.

Few studies have examined the relationship between skill mismatch and the business cycle, mainly because of a lack of a sufficiently long series of panel micro data (CEDEFOP, 2010). The business cycle will influence skill mismatch through changing the demand for skills, and the return to well-matched skills will be increased in upturns and reduced in downturns. It is possible that the response of the labor market will feed back into the business cycle, especially through skill hoarding and training responses. The distinction between firm or sector human capital and general human capital becomes pertinent, as it impacts on the willingness of firms and individuals to pay for training. The social opportunity cost to the state of contributing toward the cost of training can be considerably reduced during a recession, making the question of who pays for training cycle-dependent.

Policy Implications

There is a lot of unobserved individual heterogeneity which is revealed when panel estimation is utilized. Overeducation on its own appears to have few negative effects suggesting that for many the state of overeducation is a choice variable, in which individuals trade-off being matched for compensating differences which leave them satisfied. Policy should focus more on skills mismatch which does appear to have negative consequences. This should qualify statements such as the assertion made by a recent European Union report (2012, p. 351) on skills mismatch to the effect that "skill mismatch negatively affects economic competitiveness and growth, increases unemployment, undermines social inclusion, and generates significant economic and social costs." While this may be the case for skills mismatch, it is less obviously so for qualifications mismatch, though even here there are questions of whether the education received is a good investment for the state. As Quintini (2011a) notes, very few studies mention the policy implications of their findings, but where they do, suggestions include a better match between what is delivered through the education system and the requirements of the labor market, including a better careers guidance service, and extended training and retraining systems, which deal both with the needs of the unemployed and those subject to skill obsolescence.

See also: Education and Employment; Education, Economics of; Experimental Design: Overview; Happiness and Work; Innovation and Technological Change, Economics of; Labor Supply; Longitudinal Causal Inference; Longitudinal Data; Search, Economics of; Unemployment: Structural; Wage Differential and Structure.

Bibliography

Arellano, M., 2003. Panel Data Econometrics. Oxford University Press, Oxford.
Bauer, T., 2002. Educational mismatch and wages: a panel analysis. Economics of Education Review 21, 221–229.
Black, D., 2013. The Utilisation of Human Capital from Education in Australian Labour Markets. Over-Education? Ph.D. thesis, University of Melbourne.
CEDEFOP, 2010. The Skill Matching Challenge: Analysing Skill Mismatch and Policy Implications. European Union, Luxembourg.
CEDEFOP, 2012. Preventing Skills Obsolescence: Rapid Labour Market Changes Leave Too Many Workers at Risk of Losing Their Skills. CEDEFOP Briefing Note, July.
Chevalier, A., 2003. Measuring over-education. Economica 70 (3), 509–531.
Chevalier, A., Lindley, J., 2009. Over-education and the skills of UK graduates. Journal of the Royal Statistical Society, Series A 172 (2), 287–337.
Deaton, A., 1985. Panel data from time series of cross sections. Journal of Econometrics 30, 109–126.
Desjardins, R., Rubenson, K., 2011. An Analysis of Skill Mismatch Using Direct Measures of Skills. OECD Education Working Papers, No. 63, Paris.
Duncan, J., Hoffman, S., 1981. The incidence and wage effects of over-education. Economics of Education Review 1 (1), 75–86.
European Union, 2012. The Skill Mismatch Challenge in Europe. Chapter 6 in Employment and social developments in Europe, Luxembourg, pp. 351–394.
Frank, R.H., 1978. Why women earn less: the theory and estimation of differential over-qualification. American Economic Review 68 (3), 360–373.
Freeman, R., 1976. The Overeducated American. Academic Press, New York.
Green, F., Zhu, Yu, 2010. Overqualification, job dissatisfaction, and increasing dispersion in the returns to graduate education. Oxford Economic Papers 62 (4), 740–763.
Hartog, J., 1986. Allocation and the earnings function. Empirical Economics 2 (2), 97–110.
Hartog, J., Oosterbeek, H., 1988. Education, allocation and earnings in the Netherlands, overschooling. Economics of Education Review 7 (2), 185–194.
Hsiao, C., 2003. Analysis of Panel Data. Cambridge University Press, Cambridge.
Healy, J., Mavromaras, K., Sloane, P., 2011. Adjusting to Skill Shortages: Complexity and Consequences. IZA. IZA Discussion Paper No. 6097.
Jones, M., Mavromaras, K., Sloane, P., Wei, Z., 2014. Disability, job mismatch, earnings and job satisfaction in Australia. Cambridge Journal of Economics 38 (5), 1221–1246.
Kampelmann, S., Rycx, F., 2012. The impact of educational mismatch on firm productivity: evidence from linked panel data. Economics of Education Review 31 (6), 918–931.
Leuven, E., Oosterbeek, H., 2011. In: Hanushek, E., Welsh, F. (Eds.), Handbook of the Economics of Education, vol. 4. Elsevier Science, pp. 282–386.
Mavromaras, K., McGuinness, S., 2012. Overskilling dynamics and education pathways. Economics of Education Review 31, 619–628.
Mavromaras, K., McGuinness, S., O'Leary, N., Sloane, P., Wei, Z., 2013a. Job mismatches and labour market outcomes: panel evidence on university graduates. Economic Record 89 (286), 382–395.
Mavromaras, K., Mahuteau, S., Sloane, P., Wei, Z., 2013b. The effect of over-skilling dynamics on wages. Education Economics 21 (3), 281–303.
Mavromaras, K., Healy, J., Richardson, S., Sloane, P., Wei, Z., Zhu, R., 2013c. A System for Monitoring Shortages and Surpluses in the Market for Skills. Report prepared for the Australian Commonwealth Government. Australian Workforce and Productivity Agency, Canberra.
Mundlak, Y., 1978. On the pooling of time series and cross section data. Econometrica 46, 69–85.
Petrognolo, B., Pissarides, C., 2001. Looking into the black box: a survey of the matching function. Journal of Economic Literature 39, 390–431.
Pissarides, C., 1990. Equilibrium Unemployment Theory. Basil Blackwell, Oxford.
Quintini, G., 2011a. Over-Qualified or Under-Skilled: A Review of Existing Literature. OECD Social, Employment, and Migration Working Papers, No. 121, Paris.
Quintini, G., 2011b. Right for the Job: Over-Qualified or Over-Skilled?. OECD Social, Employment, and Migration Working Papers, No. 120, Paris.
Randstat, 2012. Into the Gap: Exploring Skills and Mismatches. SEO Report No. 2011-56, Amsterdam.
Sattinger, M., 1993. Assignment models of the distribution of earnings. Journal of Economic Literature 31, 831–880.
Sattinger, M., 2012. Qualitative mismatches. Foundations and Trends in Microeconomics 8 (1–2), 1–168.
Sattinger, M., Hartog, J., 2013. Nash bargaining and the wage consequences of educational mismatches. Labour Economics 23, 50–56.

Sicherman, N., 1991. Over-education in the labor market. Journal of Labor Economics 9 (2), 101–122.
Sicherman, N., Galor, O., 1990. A theory of career mobility. Journal of Political Economy 8 (1), 169–192.
Sloane, P., Battu, H., Seaman, P., 1999. Over-education, under-education and the British labour market. Applied Economics 31, 1437–1453.
Slonimczyk, F., 2013. Earnings inequality, and skills mismatch in the US: 1973–2002. Journal of Economic Inequality 11, 163–194.
Thurow, L.C., 1975. Generating Inequality. Basic Books, New York.
Tsai, Y., 2010. Returns to overeducation: a longitudinal analysis ot the US labor market. Economics of Education Review 29 (4), 606–617.
Tsang, M.C., 1987. The impact of the underutilisation of education on productivity: a case study of the US Bell companies. Economics of Education Review 6 (3), 239–254.
Verhaest, D., Omey, E., 2006. Discriminating between alternative measures of over-education. Applied Economics 38, 2113–2120.

Missing Data

Roderick J Little, University of Michigan, Ann Arbor, MI, USA
Donald B Rubin, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The problem of missing data is ubiquitous in the social and behavioral sciences. We begin by distinguishing between the pattern of missing data and the mechanism that creates the missing data. Then we consider common, but limited, approaches: complete-case analysis, nonresponse weighting, and single imputation. We then discuss more principled methods, including maximum likelihood and multiple imputation.

Introduction

Missing values arise in social science data for many reasons. Some examples follow. In longitudinal studies, subjects often drop out prior to the end of the study (*attrition*). In many surveys, some individuals provide no information because of noncontact or refusal to respond (*unit* nonresponse). Others are contacted but provide only some information (*item* nonresponse). Information about a variable is partially recorded; a common example in social science occurs when times to an event (death, finding a job) are being recorded, and the event has not taken place for some individuals when the study is terminated; the time for such a subject is known to be greater than that corresponding to the latest time of observation, but the actual time is unknown (*right censoring*). In other situations it is known that the time to an event lies in an interval; for example, in a longitudinal study of employment, it may be established that the event of finding a job took place some time between two interviews (*interval censoring*). Often indices are constructed by summing values of particular items, some of which may be missing.

Missing data can arise by design. For example, suppose one objective in a study of obesity is to estimate the distribution of a measure Y_1 of body fat in the population, and correlate it with other factors. Suppose Y_1 is expensive to measure but a proxy measure Y_2, such as body mass index, which is a function of height and weight, is inexpensive to measure. Then it may be useful to measure Y_2 and covariates, X, for a large sample and Y_1, Y_2, and X for a smaller subsample. The subsample allows predictions of the missing values of Y_1 to be generated for the larger sample, yielding more efficient estimates of quantities involving Y_1 than are possible from the subsample alone.

The most important step in dealing with missing data is to try to avoid having them by careful design and data collection (e.g., National Research Council, 2010). Because data are likely to be missing in any case, however, it is important to try to collect covariates that are predictive of the missing values, so that an adequate adjustment can be made. In addition, the process that leads to missing values should be assessed during the collection of data if possible (Little, 1995), because this information helps to model the missing-data mechanism when an adjustment for the missing values is performed.

Three major approaches to the analysis of missing data can be distinguished: (1) Discard incomplete cases and analyze the remainder (complete-case (CC) analysis), as discussed in the section 'Complete-Case and Weighting Methods'; (2) impute or fill in one number for each missing value and then analyze the filled-in data (single imputation), as discussed in the section 'Single Imputation'; and (3) analyze the incomplete data by a method that does not require a complete (that is, a rectangular) dataset. In particular, we focus in sections 'Multiple Imputation,' 'Maximum Likelihood for Ignorable Models,' 'Maximum Likelihood for Nonignorable Models,' and 'Bayesian Simulation Methods' on powerful likelihood-based methods, specifically ML and Bayesian simulation (*see* Bayesian Statistics). The latter is closely related to multiple imputation (MI) (Rubin, 2004), an extension of single imputation that allows uncertainty in the imputations to be reflected appropriately in the analysis, as discussed in the section 'Multiple Imputation.'

A basic assumption in all our methods is that missingness of a particular value hides a true underlying value that is meaningful for analysis, which may seem obvious but is not always the case. For example, consider a longitudinal analysis of employment; for subjects who leave the study because they move to a different location, it makes sense to consider employment history after relocation as missing, whereas for subjects who die during the course of the study, it is not reasonable to consider employment history after time of death as missing. Rather it is preferable to restrict the analysis of employment to individuals who are alive. More complex missing-data problems arise when individuals leave a study for unknown reasons, which may include relocation or death.

Pattern and Mechanism of Missing Data

The *pattern* of the missing data refers to indicators for which values in the dataset are observed and which are missing. Specifically, let $Y = (y_{ij})$ denote an $(n \times p)$ rectangular dataset without missing values, with ith row $y_i = (y_{i1}, \ldots, y_{ip})$, where y_{ij} is the value of variable Y_j for subject i. With missing values, the pattern of missing data is defined by the *response indicator matrix* $R = (r_{ij})$, such that $r_{ij} = 1$ if y_{ij} is observed and $r_{ij} = 0$ if y_{ij} is missing; $1 - r_{ij}$ is the missing-data indicator for y_{ij}.

Some methods for handling missing data apply to any pattern of missing data, whereas other methods assume

a special pattern. A simple special pattern is *univariate* nonresponse, where missingness is confined to a single variable. Another example is *monotone* missing data, where the variables can be arranged so that $Y_{j+1}, \ldots, Y_p$ are missing for all subjects where Y_j is missing, for all $j = 1, \ldots, p-1$. The result is a 'staircase' pattern, where all variables and units to the left of the broken line forming the generally irregular staircase are observed, and all to the right are missing. This pattern arises in longitudinal data subject to attrition, where once a person drops out, no more data are observed for that person.

The missing-data mechanism addresses the reasons why values are missing, and in particular whether these reasons relate to values in the dataset. For example, subjects involved in a longitudinal intervention may be more likely to drop out of a study because they feel a treatment was ineffective, which might be related to a poor value of an outcome measure. Rubin (1976) treated R as a random matrix and characterized the missing-data mechanism by the conditional distribution of R given Y, say $f(R|Y, \phi)$, where ϕ denotes unknown parameters. When missingness does not depend on the values of the data Y, missing or observed, that is,

$$f(R|Y, \phi) = f(R|\phi) \quad \text{for all } Y, \phi \frac{1}{2}$$

the missing-data mechanism is called missing completely at random (MCAR). MCAR does not mean that the pattern itself is random, but rather that missingness does not depend on any data values. An MCAR mechanism is plausible in some planned missing-data designs, but is a strong and often unrealistic assumption, especially when missing data do not occur by design, because missingness often does depend on values of variables.

Using a notation that can be made more precise, let Y_{obs} denote the observed components of Y and let Y_{mis} denote the missing components of Y. A less-restrictive assumption is that missingness depends only on values Y_{obs} that are observed and not on values Y_{mis} that are missing. That is,

$$f(R|Y, \phi) = f(R|Y_{\text{obs}}, \phi) \quad \text{for all } Y_{\text{mis}}, \phi \qquad [1]$$

The missing data are then called missing at random (MAR) at the observed values of R and Y_{obs}. Many methods for handling missing data assume the mechanism is MCAR, and yield biased estimates when the data are not MCAR. Better methods only assume MAR or some other explicitly defined mechanism.

Complete-Case and Weighting Methods

A common and simple method is CC analysis, also known as *listwise* deletion, where units with any missing values are discarded and standard analysis methods are applied to the remaining complete cases. In many statistical packages this is the default analysis. Valid (but often suboptimal) repeated-sampling inferences are obtained when the missing data are MCAR, because then the complete cases are a random subsample of the original sample with respect to all variables. However, even under MCAR, CC analysis results in the loss of the information in the incomplete cases, which may be substantial, particularly when the number of variables is large.

A serious problem with dropping incomplete cases is that the complete cases are often a biased sample. The size of the resulting bias depends on the degree of deviation from MCAR, the amount of missing data, and the specific analysis being conducted. For example, inferences about means are subject to bias if the missingness mechanism deviates from MCAR, but inference about regression coefficients is valid (although often inefficient) if the missingness mechanism depends on covariates but not on outcomes. In sample surveys, this issue motivates strenuous attempts to limit unit nonresponse through multiple follow-ups, and surveys with high rates of unit nonresponse (say 30% or more) are often considered unreliable for making inferences to the whole population.

A modification of CC analysis, commonly used to handle unit nonresponse in surveys, is inverse probability weighting (IPW), which weights respondents by the inverse of an estimate of the probability of response. In particular, when estimating a population mean, the sample mean is replaced by the weighted mean. A simple approach for creating weights is to form adjustment cells (also called subclasses) based on background variables measured for respondents and nonrespondents; for unit nonresponse adjustment, these are often based on geographical areas or groupings of similar areas based on aggregate socioeconomic data. All nonrespondents are assigned zero weight and the nonresponse weight for all respondents in an adjustment cell is then the inverse of the response rate in that cell. This method removes the component of nonresponse bias attributable to differential nonresponse rates across the adjustment cells, and eliminates bias if, within each adjustment cell, respondents can be regarded as a random subsample of the original sample within that cell (i.e., the data are MCAR within each adjustment cell).

With more extensive background information, a useful alternative approach is *response propensity* stratification, where (1) the indicator for unit nonresponse is regressed on the background variables, using the combined data for respondents and nonrespondents, using a method such as logistic regression appropriate for a binary outcome; (2) a predicted response probability is computed for each respondent based on the regression in (1); and (3) adjustment cells are formed based on a categorized version of the predicted response probability. Theory (Rosenbaum and Rubin, 1983) suggests that this is an effective method for removing nonresponse bias attributable to the background variables used to form the adjustment cells when unit nonresponse is MAR.

Although IPW can be useful for reducing nonresponse bias, it does have serious limitations. First, information in the incomplete cases is still discarded, so the method is generally inefficient; weighted estimates can have unacceptably high variance, especially when outlying values of a variable are given large weights. For modifications of the IPWs to increase efficiency, see for example Cao et al. (2009).

Variance estimation for weighted estimates should take into account uncertainty in the estimated weights (*see* Estimation: Point and Interval). Approaches include Taylor series expansion (Robins et al., 1995) or computing bootstrap standard errors – but in order to reflect all the uncertainty, the weights have to be recalculated for each bootstrap sample (Little and Rubin, 2002: Chapter 5).

Single Imputation

Methods that impute or fill in the missing values have the advantage that, unlike CC or IPW analysis, observed values in the incomplete cases are retained. A simple version imputes missing values of a variable by its unconditional sample mean based on the observed data. This method can yield satisfactory point estimates of means and totals under MCAR, but it yields inconsistent estimates of other estimands, such as variances, covariance matrices, or regression coefficients, and standard errors based on the filled-in data are systematically underestimated (Little and Rubin, 2002: Chapter 3). Thus it cannot be recommended in any generality.

An improvement is *conditional mean* imputation, in which missing values for a unit are replaced by estimates of their conditional mean given the values of observed variables for that unit. For example, in the case of univariate nonresponse with $Y_1, \ldots, Y_{p-1}$ fully observed and Y_p sometimes missing, one approach is to classify units into cells based on similar values of observed variables and then to impute missing values of Y_p by the within-cell mean from the complete units in that cell. A more general approach is to regress Y_p on $Y_1, \ldots, Y_{p-1}$ using the complete units and then impute the missing values of Y_p using the predictions from the regression (*regression imputation*). For a general pattern of missing data, the missing values for each unit can be imputed from the regression of the missing variables on the observed variables, computed using the set of complete units. Iterative versions of this method lead (with some important adjustments) to ML estimates under multivariate normality (Little and Rubin, 2002: Chapter 8).

Imputations can be used to improve the efficiency of IPW estimates. In particular, augmented IPW (AIPW) estimates of the mean use a model to predict the missing values and then apply IPW to the residuals (Robins et al., 1994, 1995); this increases the efficiency of IPW if there are good predictors. These estimators afford some protection against misspecification of the prediction model for the missing values, and in particular are called *doubly robust*, in the sense that they are consistent for the target parameter if either the prediction model or the model for estimating the response propensities is correctly specified. There is some debate about the utility of this property in practice, because all models suffer from some degree of misspecification, and the approach can be nonrobust if neither model is correct (Kang and Schafer, 2007; Zhang and Little, 2011; Waernbaum, 2012). The penalized spline of propensity prediction (Zhang and Little, 2009) method estimates the response propensity and then models its relationship with the outcome using a penalized spline, including other predictors parametrically. This approach appears to perform well in simulations (Zhang and Little, 2011). For estimands other than means, AIPW methods can be applied to inferences based on generalized estimating equations with missing data – see, for example, Tsiatis (2006).

Although conditional mean imputation incorporates information from the observed variables and under MAR yields best predictions of the missing values in the sense of mean squared error, it leads to distorted estimates of quantities that are not linear in the data, such as percentiles, correlations, and other measures of association, variances and other measures of variability. A way to address this problem is to use random draws rather than best predictions to preserve the distribution of variables in the filled-in dataset. An example is *stochastic regression* imputation, in which each missing value is replaced by its regression prediction plus a random error based on the estimated residual variance.

Hot-deck methods impute actual values observed in the dataset. A simple approach creates adjustment cells based on variables observed for respondents and nonrespondents and then imputes nonrespondent values using a randomly selected respondent value in the same cell. This method requires categorizing continuous observed variables, and is less satisfactory with extensive variables, because the number of adjustment cells becomes large. A more general approach to hot-deck imputation matches nonrespondents and respondents using a distance function based on the variables observed for both types of units. For example, *predictive mean matching* (Rubin, 1986; Little, 1988b) matches units with similar predicted means from a regression of the missing variable on observed variables. This method is somewhat robust to misspecification of the regression model used to create the matching metric. These methods, and extensions to handle general patterns of missing data, are reviewed in Andridge and Little (2010).

The imputation methods discussed so far assume the missing data are MAR. In contrast, models that are missing not at random (MNAR) assert that even if a respondent and a nonrespondent to Y_p appear identical with respect to observed variables $Y_1, \ldots, Y_{p-1}$, their Y_p values differ systematically. Lillard et al. (1986), for example, discussed MNAR imputation models for missing Current Population Survey data. It is also possible to create an MNAR hot-deck procedure; for example, respondents' values that are to be imputed to nonrespondents could be multiplied by an inflation or deflation factor that depends on the variables that are observed. A crucial point about the use of MNAR models is that there is rarely direct evidence in the data to address the validity of their underlying assumptions. Thus, whenever MNAR models are being considered, it is prudent to consider several MNAR models and explore the sensitivity of analyses to the choice of model (Rubin, 1977; Little and Wang, 1996; National Research Council, 2010; Liublinska and Rubin, 2014).

Multiple Imputation

A single imputed value cannot represent the uncertainty about which value to impute, so analyses that treat imputed values just like observed values generally underestimate uncertainty, even if nonresponse is modeled correctly and random imputations are created. Large-sample results show that for simple situations with 30% of the information missing, single imputation under the correct model results in nominal 90% confidence intervals having actual coverages below 80%. The inaccuracy of nominal levels is even more extreme in multiparameter testing problems (Rubin, 2004: Chapter 4).

A modification of imputation that fixes this problem is MI (Rubin, 1987, 2004). Instead of imputing a single set of draws for the missing values, a set of M (say $M = 5$) datasets are created, each containing different sets of draws of the missing values from their predictive distribution. We then apply the

standard complete-data analysis to each of the M datasets and combine the results in a simple way. In particular for scalar estimands, the MI estimate is the average of the estimates from the M datasets, and the sampling variance of the estimate is the average of the sampling variances from the five datasets plus $1+M^{-1}$ times the sample variance of the estimates over the M datasets; the factor $1+M^{-1}$ is a small-M correction; the last quantity here estimates the contribution to the variance from imputation uncertainty, missed (i.e., set to zero) by single imputation methods. Another benefit of MI is that averaging over datasets results in more efficient point estimates than does single random imputation. Often MI is not much more difficult than doing a single imputation – the additional computing from repeating an analysis M times is not a major burden and methods for combining inferences are straightforward. Most of the work is in generating good posterior predictive distributions for the missing values, which is needed for consistent point estimation in any case.

An important refinement is to create each of the imputed datasets using a different draw of the estimated parameters, thereby reflecting uncertainty in the parameter estimates. This form of MI is called *proper* (Rubin, 2004: Chapter 4).

Maximum Likelihood for Ignorable Models

There are statistical methods that let us analyze a nonrectangular dataset without having to impute the missing values. In particular, ML avoids imputation by formulating a statistical model and basing inference on the likelihood function of the incomplete data. Define Y and R as before, and fully observed fixed covariates are implicit in our notation. The data and missing-data mechanism are modeled in terms of a joint distribution for Y and R. *Selection models* specify this distribution as

$$f(Y,R|\theta,\phi)=f(Y|\theta)f(R|Y,\phi) \qquad [2]$$

where $f(Y|\theta)$ is the model for Y in the absence of missing values, such as the multivariate normal; $f(R|Y,\phi)$ is the model for the missing-data mechanism, such as the multinomial logit or probit model; and θ and ϕ are unknown parameters. The likelihood of θ, ϕ given the data Y_{obs}, R is then proportional to the density of Y_{obs}, R regarded as a function of the parameters θ, ϕ, and is obtained by integrating the density in eqn [2] over the missing data Y_{mis}, that is

$$L_{\text{full}}(\theta,\phi|Y_{\text{obs}},R)=\text{const.}\times\int f(Y,R|\theta,\phi)\mathrm{d}Y_{\text{mis}} \qquad [3]$$

The likelihood of θ *ignoring the missing-data mechanism* is obtained by integrating the marginal distribution of Y over the missing data, that is

$$L_{\text{ign}}(\theta|Y_{\text{obs}})=\text{const.}\times\int f(Y|\theta)\mathrm{d}Y_{\text{mis}} \qquad [4]$$

This likelihood is easier to work with than eqn [3], because it is typically more tractable computationally, and more importantly, avoids the need to specify a model for the missing-data mechanism, about which little is known in many situations. Hence it is important to determine when valid likelihood inferences are obtained from eqn [4] instead of the full likelihood eqn [3]. Rubin (1976) showed that valid likelihood-based inferences about θ are obtained from eqn [3] when the missing data are MAR, i.e., eqn [1], and θ and ψ are distinct in the sense that they take values in disjoint parameter spaces; the missing-data mechanism is then called *ignorable* for likelihood inferences. Large-sample inferences about θ based on an ignorable model are based on ML theory, which shows that under regularity conditions,

$$\theta-\widehat{\theta}\sim N_k(0,C) \qquad [5]$$

where $\widehat{\theta}$ is the value of θ that maximizes eqn [4], and $N_k(0,C)$ is the k-variate normal distribution with mean zero and covariance matrix C given by the inverse of an information matrix; for example, $C=I^{-1}(\widehat{\theta})$ where I is the observed information matrix given by $I(\theta)=-\partial^2\log L(\theta|Y_{\text{obs}})/\partial\theta\partial\theta^T$, or $C=J^{-1}(\widehat{\theta})$, where $J(\theta)$ is the expected value of $I(\theta)$. As in Little and Rubin, 2002: Chapter 6, eqn [5] is written to be open to a frequentist interpretation if $\widehat{\theta}$ is regarded as random and θ fixed, or a Bayesian interpretation if θ is regarded as random and $\widehat{\theta}$ fixed. Thus if the missing data are MAR, the likelihood approach reduces to developing a suitable model for the data and computing $\widehat{\theta}$ and C.

In many problems, maximization of the likelihood to compute $\widehat{\theta}$ requires numerical methods. These methods include standard optimization methods such as Newton–Raphson or scoring, or the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) or one of its extensions; theory and numerous applications of EM can be found in Meng and Pedlow (1992), Schafer (1997), McLachlan and Krishnan (1997), and Little and Rubin (2002: Part II). Examples include multivariate normal data with a general pattern of missing values and the related problem of multivariate linear regression with missing data; robust inference based on multivariate t models; loglinear models for multiway contingency tables with missing data; and the general location model for mixtures of continuous and categorical variables, which yields ML algorithms for logistic regression with missing covariates. Another common software implementation of ML is to repeated-measures models with missing outcome variables.

Maximum Likelihood for Nonignorable Models

Nonignorable, non-MAR models apply when missingness depends on the missing values (see the section 'Pattern and Mechanism of Missing Data'). A correct likelihood analysis must be based on the full likelihood from a model for the joint distribution of Y and R. The standard likelihood asymptotics apply to nonignorable models, and computational tools such as EM also apply to this more general class of models (see Little and Rubin, 2002: Chapter 15). However, often information needed to estimate all the parameters is very limited, and estimates are sensitive to misspecification of the model. Often a sensitivity analysis is needed to see how much the answers change for various assumptions about the missing-data mechanism.

Pattern-mixture models are particularly useful here, and specify

$$f(Y,R/\pi,\psi)=f(Y/X,\psi)f(R/X,\pi) \qquad [6]$$

That is, models for Y given X for each pattern of missing data, where ψ and π are unknown distinct parameters, and now the distribution of Y is conditioned on the missing-data pattern, R (Rubin, 1977; Little, 1993).

Although most of the literature on missing data has concerned selection models that specify $f(Y|\theta)$ and $f(R|Y, \phi)$ as in eqn [2], pattern-mixture models can also provide inferences for parameter θ of the complete-data distribution, by expressing it as functions of the pattern-mixture model parameters ψ and π. Pattern-mixture models of the form of eqn [6] are simpler to fit, and often have a more straightforward practical interpretation (National Research Council, 2010; Giusti and Little, 2011; Carpenter and Kenward, 2013).

Heitjan and Rubin (1991) and Heitjan (1994) extend the formulation of missing-data problems via the joint distribution of Y and M to more general incomplete data problems involving coarsened data. The idea is to replace the binary missing-data indicators by coarsening variables that map the γ_{ij} values to coarsened (e.g., rounded) versions. Full and ignorable likelihoods can be defined for this more general setting. This theory provides a bridge between missing-data theory and theories of censoring in the survival analysis literature. The EM algorithm can be applied to such models by including the missing-data (or coarsening) indicator in the complete-data likelihood.

Bayesian Simulation Methods

ML is most useful when sample sizes are large, because then the log-likelihood is nearly quadratic and can be summarized well using the ML estimate of θ and its large-sample variance–covariance matrix. When sample sizes are small, a useful alternative approach is to include a prior distribution for the parameters and compute the posterior distribution of the parameters of interest. For ignorable models this is

$$p(\theta/Y_{\text{obs}}|, R) \equiv p(\theta/Y_{\text{obs}}) = \text{const.} \times p(\theta) \times f(Y_{\text{obs}}/\theta)$$

where $p(\theta)$ is the prior distribution of θ, and $f(Y_{\text{obs}}|\theta)$ is the density of the observed data given θ, as in eqn [3]. Because the posterior distribution rarely has a simple analytic form for incomplete data problems, iterative simulation methods such as data augmentation (Tanner and Wong, 1987) and the Gibbs sampler (Geman and Geman, 1984) are often used to generate draws of θ from the posterior distribution $p(\theta|Y_{\text{obs}}, R)$. The Gibbs sampler starts with an initial draw $\theta^{(0)}$ from an approximation to the posterior distribution of θ. Given a value $\theta^{(t)}$ of θ drawn at iteration t:

1. Draw $Y_{\text{mis}}^{(t+1)}$ with density $p(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})$
2. Draw $\theta^{(t+1)}$ with density $p(\theta/Y_{\text{obs}}, Y_{\text{mis}}^{(t+1)})$

The procedure is motivated by the fact that the distributions in steps 1 and 2 are often much easier to draw from than the correct posterior distributions $p(Y_{\text{mis}}|Y_{\text{obs}})$ and $p(\theta|Y_{\text{obs}})$. Under mild regularity conditions, the iterative procedure can be shown in the limit to yield a draw from the joint posterior distribution of Y_{mis}, θ given Y_{obs}. The algorithm was termed chained data augmentation in Tanner (1991). This algorithm can be run independently K times to generate K draws from the approximate joint posterior distribution of θ and Y_{mis}.

Techniques for monitoring the convergence of such algorithms are discussed in a variety of places (e.g., Gelman et al., 2014: Chapter 11). Schafer (1997) developed data augmentation algorithms that use iterative Bayesian simulation to multiply – impute rectangular datasets with arbitrary patterns of missing values when the missing-data mechanism is ignorable. The methods are applicable when the rows of the complete-data matrix can be modeled as iid observations from the multivariate normal, multinomial loglinear, and general location models.

Overview

The advent of modern computing has made more principled methods of dealing with missing data practical (e.g., MI, Rubin, 1996). Such methods are finding their way into current commercial and free software. Schafer (1997), Van Buuren (2012), Carpenter and Kenward (2013), and the forthcoming third edition of Little and Rubin (2002) are sources to learn more about missing-data methodology.

See also: Bayesian Statistics; Computation: Expectation-Maximization Algorithm; Estimation: Point and Interval; Likelihood in Statistics; Markov Chain Monte Carlo Methods; Monte Carlo Methods and Bayesian Computation: Overview; Sample Surveys: Model-Based Approaches.

Bibliography

Andridge, R.H., Little, R.J., 2010. A review of hot deck imputation for survey nonresponse. International Statistical Review 78 (1), 40–64.

Cao, W., Tsiatis, A.A., Davidian, M., 2009. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. Biometrika 96, 723–734.

Carpenter, J.R., Kenward, M.G., 2013. Multiple Imputation and its Application. John Wiley & Sons, Ltd, Chichester, UK.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society Series B 39, 1–38.

Efron, B., 1994. Missing data, imputation and the bootstrap. Journal of the American Statistical Association 89, 463–479.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D., 2014. Bayesian Data Analysis, third ed. Chapman & Hall, New York.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721–741.

Giusti, C., Little, R.J., 2011. A sensitivity analysis of nonignorable nonresponse to income in a survey with a rotating panel design. Journal of Official Statistics 27 (2), 211–229.

Heitjan, D.F., 1994. Ignorability in general incomplete-data models. Biometrika 81 (4), 701–708.

Heitjan, D., Rubin, D.B., 1991. Ignorability and coarse data. Annals of Statistics 19, 2244–2253.

Kang, D.Y.J., Schafer, J.L., 2007. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science 22, 523–539.

Lillard, L., Smith, J.P., Welch, F., 1986. What do we really know about wages: the importance of nonreporting and census imputation. Journal of Political Economy 94, 489–506.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data, second ed. Wiley, New York.

Little, R.J., Wang, Y.-X., 1996. Pattern-Mixture Models for Multivariate Incomplete Data with Covariates. Biometrics 52, 98–111.

Little, R.J.A., 1988a. Robust estimation of the mean and covariance matrix from data with missing values. Applied Statistics 37, 23–38.
Little, R.J.A., 1988b. Missing data adjustments in large surveys. Journal of Business and Economic Statistics 6, 287–301.
Little, R.J.A., 1993. Pattern-mixture models for multivariate incomplete data. Journal of the American Statistical Association 88, 125–134.
Little, R.J.A., 1995. Modeling the drop-out mechanism in longitudinal studies. Journal of the American Statistical Association 90, 1112–1121.
Liublinska, V., Rubin, D.B., 2014. Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. Statistics in Medicine 33, 4170–4185.
McLachlan, G.J., Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley, New York.
Meng, X.L., Pedlow, S., 1992. EM: a bibliographic review with missing articles, Proceedings of the Statistical Computing Section. American Statistical Association 24–27.
National Research Council, 2010. The Prevention and Treatment of Missing Data in Clinical Trials. National Academy Press, Washington, DC.
Oh, H.L., Scheuren, F.S., 1983. Weighting adjustments for unit nonresponse. In: Madow, W.G., Olkin, I., Rubin, D.B. (Eds.), Incomplete Data in Sample Surveys, Theory and Bibliographies, vol. 2. Academic Press, New York, pp. 143–184.
Robins, J.M., Rotnitzky, A., Zhao, L.P., 1994. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89, 846–866.
Robins, J.M., Rotnitzky, A., Zhao, L.P., 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 90, 106–121.
Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.
Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592.
Rubin, D.B., 1977. Formalizing subjective notions about the effect of nonrespondents in sample surveys. Journal of the American Statistical Association 72, 538–543.
Rubin, D.B., 1986. Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations. Journal of Business and Economic Statistics 4, 87–94.
Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.
Rubin, D.B., 1996. Multiple imputation after 18+ years. Journal of the American Statistical Association 91, 473–489 with discussion, 507–515; and rejoinder, 515–517.
Rubin, D.B., 2004. Multiple Imputation for Nonresponse in Surveys. Reprinted with appendices as a "Wiley Classic." John Wiley & Sons, New York.
Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. Chapman & Hall, London.
Tanner, M.A., 1991. Tools for Statistical Inference: Observed Data and Data Augmentation Methods. New York, Springer-Verlag.
Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–550.
Tsiatis, A.A., 2006. Semiparametric Theory and Missing Data. Springer, New York.
Van Buuren, S., 2012. Flexible Imputation of Missing Data. Chapman and Hall/CRC, New York.
Waernbaum, I., 2012. Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. Statistics in Medicine 31, 1572–1581.
Zhang, G., Little, R.J., 2009. Extensions of the penalized spline of propensity prediction method of imputation. Biometrics 65 (3), 911–918.
Zhang, G., Little, R.J., 2011. A comparative study of doubly robust estimators of the mean with missing data. Journal of Statistical Computing and Simulation 81, 2039–2058.

Missing Girls: A Globalizing Issue

Christophe Z Guilmoto, IRD/CEPED, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The recent 50 years have been characterized by an increasing proportion of men in the world's population. In addition to excess female mortality, this rise has been caused by the emergence of prenatal sex selection in Asia and Eastern Europe and has resulted in millions of missing girls. Among all preconditions such as the rise in sex selection technology and the rapid decline of fertility, the preference for sons is at the core of this demographic masculinization and only when it dies out will sex ratio come back to normalcy.

Women's long march toward greater gender equity has been a major landmark of the second half of the twentieth century. Educational expansion, urbanization, and rapid economic development have created many new opportunities for women around the world over the last few decades. Demographic change has in particular brought many benefits to women such as increasing longevity and survival of children, less frequent and more spaced childbearing, lower maternal mortality, and higher spatial mobility. Yet, aggregated population figures suggest on the contrary that women appear to have been subjected to the most systematic form of demographic discrimination over the last 50 years and that as a result, their relative share in the world's population has been regularly declining. Using recent demographic estimates (United Nations Population Division, 2013), we can observe the gradual increase in the sex ratio, which from 99.6 males per 100 females in 1950 to 101.6 in 2010. This growing proportion of men is a clear sign of a new gendered demography that has swept Asia and other parts of the world during the last decades.

The Growing Demographic Deficit of Women

The demographic records are clear: women have seen their proportion decline exactly when the world's population as a whole has recorded its most rapid historical increase following the onset of mortality decline during the twentieth century. The first explanation for their reduced share in the overall population is that women have not benefited from increased longevity as much as men. The apparent surplus of men is due in particular to various forms of discrimination causing excess female mortality, especially during childhood (United Nations, 2011). Another explanation is that female populations may be now affected by discrimination before their birth. In fact, both prenatal and postnatal mechanisms are at the source of this reduction of women's share in the population observed in many countries (Croll, 2000; Miller, 2001).

We should first remind readers that demography may appear ambivalent when it comes to sex ratios. On the one hand, more men are born than women and the sex ratio at birth (SRB) usually osculates around 105 male births per 100 female births (Chahnazarian, 1988). On the other hand, mortality is systemically favorable to women – from intrauterine period to old age. Women's life expectancy is often several years ahead of that of men, for both biological and behavioral reasons, and this differential causes the sex ratio to diminish regularly with age, starting at 105 males per female at birth to 100 during adulthood and even lower levels beyond 50 years of age. When mortality improves and fertility declines as happened almost everywhere over the last 100 years, the average age of the population increases as well as the proportion of women. But this is not what was observed in several countries because of the persistence of excess mortality among girls and the emergence of prenatal sex selection.

These two mechanisms have different social features (Guilmoto, 2012b). Gender-based excess mortality is an age-old feature of certain demographic régimes. Girls and women deemed of lesser value are in many ways disadvantaged, most commonly by being denied equal access to food, care, and other resources. Girls may be for instance vaccinated in lower proportions than their brothers, get a smaller share of food, or be taken less often to a clinic in case of illness. The most radical form of discrimination is female infanticide, a practice reported in the past in several regions of Asia. Other traces of excess female mortality are less easy to detect in the absence of reliable statistics, but they have been shown to prevail in many countries where mortality rates among women are suspiciously higher than expected. Yet, selective infanticide has now almost completely vanished and recent trends point to a gradual reduction in mortality differentials, with life expectancy now higher for women than men across the world.

The second mechanism relates to the elimination of unborn girls during pregnancy or even earlier during conception. It is never directly observable for want of reliable data on abortions by sex of the fetus. Yet, the SRB has unexpectedly increased in several countries over the last 30 years. It took years for demographers to make sense of this increasing share of male births and to fight off pseudo-biological theories such as the effects of hepatitis epidemics, conflict and stress, or unique physiological conditions in some countries. Yet, the fact that SRB in affected populations inevitably rise with birth order – as parents react to the absence of prior male births – confirms that this is not a biological mechanism, but on the contrary the result of a conscious demographic choice. In fact, apart from issues related to the quality of statistics, the only valid explanation for the relative absence of female births in several Asian countries lies in the emergence of prenatal elimination of female embryos and fetuses, with sex selective abortions

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64065-5

playing a major role in these imbalances. (Early demographic studies can be found in Ramanamma and Bambawale (1980), Zeng et al. (1993), and Park and Cho (1995)).

Prenatal discrimination has indeed become the main driver of sex imbalances at birth over the last 30 years. **Table 1** presents recent statistics for a selection of countries where the proportion of male births is unduly high. This comparative table suggests that the issue of adverse SRB levels is not restricted to China and India: sex selection has become a more global phenomenon, with distinct manifestations in several regions of the world. The countries listed here form admittedly a mixed bag of nations with regard to their level of economic development, political systems, and cultural and religious traditions. We find for instance atheistic, Buddhists, Hindu, Sikh, Muslim, and Christian countries in our list. While Asia has long been blamed *en bloc* for prenatal discrimination, it appears that several affected countries are located in Eastern Europe and that conversely, many important Asian countries – from Iran to Japan and Indonesia – appear totally immune to mechanisms of prenatal sex selection.

In Asia, China and India are most prominent with their high SRB and these two countries are the first contributors to the world's imbalances (Attané and Guilmoto, 2007). The SRB has reached the extreme level of 120 male births per 100 female births in parts of India and China after 2000. India's average SRB is lower, but this is mostly due to the fact that many regions in the South or the East of the subcontinent do not practice prenatal sex selection at all. In Northwest India, especially from Punjab to Gujarat, the SRB can in fact be around 120 male births per 100 female births or higher. While lack of reliable estimates prevents us from concluding on Pakistan's probably high level of birth masculinity, the recent rise in the SRB in Vietnam during the last decade has been well documented. In addition, the almost normal SRB level observed today in South Korea conceals the rise observed during the 1980s, which briefly took birth masculinity to 115 male births per 100 female births before a subsequent decline. South Korea offers indeed the only illustration of a complete sex ratio transition, in which countries observe successively a temporary rise in their SRB and its decline (Guilmoto, 2009).

Table 1 Estimates of the sex ratio at birth in various countries around 2010

Country	*Sex ratio at birth*
East and Southeast Asia	
China	117.8
Singapore	107.5
South Korea	106.7
Vietnam	111.2
South Asia	
India	110.5
Pakistan	109.9
Caucasus	
Azerbaijan	116.5
Armenia	114.9
Georgia	111.8
Southeast Europe	
Albania	111.7
Kosovo	109.7
Montenegro	109.8

The normal sex ratio at birth is 105 male births per 100 female births.
Estimates from birth registration and demographic surveys.
Source: Guilmoto, C.Z., 2012b. Sex Imbalances at Birth. Current Trends, Consequences and Policy Implications, UNFPA, Bangkok and Guilmoto, C.Z., Duthé, G., 2013. Masculinization of births in Eastern Europe. Population & Societies 506, 1–4.

In Eastern Europe, two distinct regional clusters emerge (Guilmoto and Duthé, 2013). The first is located in the Southern Caucasus and includes three countries (Armenia, Azerbaijan, and Georgia) where birth masculinity is almost as high as in China. In spite of their proximity, these three countries have very distinct socioeconomic features as well as cultural and religious traditions and have been united only during the Soviet Union's era. Similarly, the other geographical cluster of high sex ratios at birth centers in Albania includes neighboring Kosovo, Montenegro, and Western Macedonia that were parts of the former Yugoslavia. These regions have not been part of the same polity since the dismantlement of the Ottoman Empire at the end of the nineteenth century.

To this list we should add a scattered set of diasporic populations, immigrants from sex-selecting regions found throughout the world in various countries and whose offspring sex ratios are biased in favor of male births (Dubuc and Coleman, 2007). This finding has been reported for instance among populations of Indian, Korean, and Chinese origin in North America and the United Kingdom. Albanian migrants in Italy and Greece also appear to beget more boys than biology would predict. While this phenomenon has little demographic impact in host countries, it shows that gender discrimination is not so much linked to the social and economic environment in origin countries than to the culture – mindset and attitudes – of people.

This initial geographical cluster suggests that adverse sex ratios tend to cluster in small regions and this is also true within countries as high values observed in India's Northwest and Vietnam's Red River Delta demonstrate. This obvious cultural dimension of gender discrimination goes deeper than ethnic, religious, or linguistic affiliation. For instance, the three countries of the South Caucasus have almost similar adverse levels of SRB in spite of deep religious, linguistic, and political variations. But there are also social variations, such as differences between rural and urban areas and across socioeconomic groups. Detailed studies show in particular that the richer strata tend initially to discriminate more strongly against female births than the poor. All these features help us understand why and how prenatal sex selection takes place (Guilmoto, 2012b).

The Causes of Gender Discrimination Regarding Offspring

Countries have often been reluctant to acknowledge the existence of prenatal sex selection, blaming faulty statistics or biological hazards for distorted sex ratios. It is true that prenatal sex selection tends to project an unsavory image of societies practicing it, but demographic evidence – from birth registration figures to census and survey results – has gradually been

shared to civil society and governments via academic and media reports. Once recognized, sex imbalances at birth tend to be attributed in each country to a specific set of local factors such as the dowry paid by parents of girls in India, strict family planning rules and Confucian precepts in East Asia, or economic crises and conflicts in Eastern Europe. Yet, the publicity received by these 'local narratives' of prenatal discrimination tends to obscure the commonalities of a demographic phenomenon that occurred almost simultaneously across Eurasia. Prenatal sex selection from Albania to Korea does in fact follow similar mechanisms with almost identical root causes. Sex selection is simply a recent adaptive behavior by couples to avoid births of the unwanted sex and it corresponds to a rational strategy in response to a set of changing constraints and opportunities. To understand its logic, we need to distinguish its three indispensable preconditions: the perceived benefits of sex selection guided by gender preference, the availability of means to implement it, and the additional urgency imposed by reduced family size (Guilmoto, 2009). I examine these three factors separately (**Figure 1**).

Son Preference

Son preference constitutes the primary factor behind prenatal and postnatal discrimination and it plays the role of the 'demand factor' in our equation: births of girls are avoided because of a skewed gender valuation system in which boys get preeminence. When given a choice about abortion, food sharing or health expenses, some families tend automatically to favor boys. In most cases, this is because they want to ensure the birth of at least one boy. Put differently, what parents want to expressly avoid is a family with no son. The cultural and political contexts may widely differ, but the pressing need for a male progeny is the common trait from the Balkans to East Asia (John et al., 2008; Murphy et al., 2011; Kaser, 2008). Here are a few illustrations in the way this preference for sons plays out in different settings.

Girl children may first be seen as more 'expensive' to bring up, mostly for security or honor-related reasons. When older, young women are also seen as more difficult to marry off, and this is particularly true in South Asia when marriage payments come essentially from the bride's family. But the low valuation of women becomes more obvious after marriage and sex-selecting countries inevitably correspond to areas where married women leave their natal family to join their husband's. As a result, bringing up and educating women are perceived ultimately as a loss and equivalent to 'watering the neighbor's garden' since all the parents' efforts will benefit another family. This feeling is due to the fact that newlywed women will become an active member of their husband's family in kinship systems characterized by strong virilocality or patrilocality. In patrilineal societies, married sons stay close to their parents and often coreside with them. They often share residence, agricultural land, or business interests with their parents. They contribute to most aspects of their parents' well-being in terms of physical protection, affection, economic support, and material assistance. In societies with limited individual instruments of social security (health insurance or pension benefits), family solidarity is a key to old age support and to withstand unanticipated health or economic shocks. Without a son, parents would be clearly at risk especially since customs sharply disapprove of being supported by one's married daughters. In patrilineal societies, solidarity and living arrangements within the family are strictly asymmetrical and exclude married daughters that belong to another family by virtue of marriage.

Support to the patrilocal family receives a very strong social sanction in the form of 'filial piety', held as a prime virtue in many Asian countries. But this lifelong support exerted strictly along the male line also has implications beyond the life of parents since many religious or traditional systems of belief also require male progeny after their death. Sons may be indispensable in funerary rituals or ancestors' worship, so that the need for sons has an indefinite horizon. To these immaterial benefits of male children, we should also add the status and honor symbol of fathering sons and perpetuating the family name. Couples with no son are lambasted and discriminated against for failing to prolong the patrilineal lineage, with effects ranging from constant sneering to inferior status for men and from domestic violence to repudiation for women. Similarly, the birth of sons is celebrated with great fanfare when that of girls is often a source of lament (Croll, 2000).

This entrenched preference for sons is often seen as an effect of wider gender inequity. It should not, for son preference is often unrelated to many other manifestations of gender discrimination at work, in the family, workplace, or public spaces. Women in China and Vietnam have for instance reached a high level of social and economic autonomy, but with apparently little effect on traditional bias against girl children. Conversely, we can see no prenatal sex selection in many countries where women are otherwise discriminated against by law and officially barred from access to the same economic, legal, or political position as men. When investigated further, son preference and the gender hierarchy it epitomizes appear closely linked to the nature of kinship systems that determine particular marriage, inheritance, and coresidence customs. All observations made earlier on the factors behind son preference presuppose a strictly patrilineal and patrilocal family organization. In more bilateral societies, when

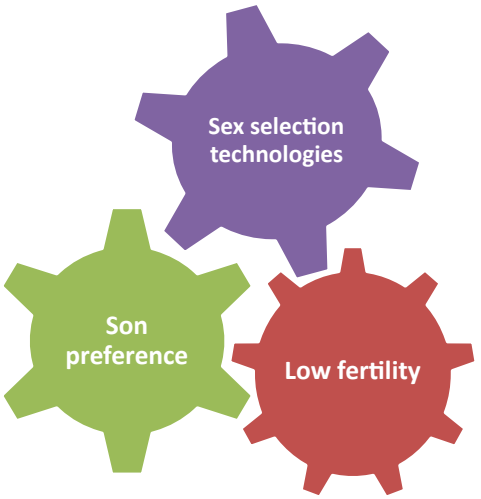

Figure 1 The three preconditions for prenatal sex selection.

the mother's and father's family are on a somewhat equal footing, discrimination against girls simply does not make sense. This can be illustrated by the absence of sex selection in countries close to China, India, and Vietnam such as Sri Lanka, Thailand, or Indonesia. Kinship rules hold a larger sway than religious or ethnic affiliation when it comes to gender preference and they are strictly enforced to ensure the reproduction of family systems (Chakraborty and Kim, 2010).

Technology and Fertility Decline

In addition to gender bias, there are two other indispensable factors for sex selection to occur (Guilmoto, 2012b). The next precondition is a typical 'supply factor' related to the availability of effective techniques to avoid female births. Sex selection has been made possible at this scale only because of a recent technological breakthrough. Of course, methods for influencing the birth of a son have long existed in son-preferring societies. In India, pilgrimages to temples and other specific rituals are for instance thought of as a way to improve the chance of having boys. Prescriptions related to the timing of the intercourse and to the diet of pregnant women are also supposed to affect birth masculinity in many settings. But these prenatal 'folk methods' were probably of limited effectiveness. Postnatal methods of sex selection were also common, starting with female infanticide, a most radical method long attested in parts of South and East Asia. Selective infanticide was often linked by hypergamic norms among higher-status groups, but at times it simply represented a crude way to avoid a redundant female birth. A more common way to eliminate female children is through neglect after birth. Higher mortality rates are the undeniable outcome of discriminatory attitudes toward daughters that translate into passive discrimination. Yet, all these techniques are mostly remnants for the "ancien régime" of sex selection. In fact, they were all rather ineffective, low-tech methods, with the exception of infanticide, which for its part entailed huge psychological costs.

The first change took place when modern contraception became available, allowing women to stop childbearing once they had reached the desired gender composition. As a result, many families end with a boy as the last child as is observed in Vietnam or India. This strategy is definitely biased toward boys, but it has no impact whatsoever on the overall sex ratio of births that remains only determined by random biological factors. The contraception revolution also pointed to a new trend, the desire to control family size and the new possibility to actively avoid unwanted pregnancies. At about the same time, the new prenatal diagnostic techniques developed in industrialized countries started to spread in richer Asian countries during the 1980s, making sex detectable after 3 months of pregnancy. After 1980, many clinics in Asia were already offering sex diagnosis services. When this technology hit countries like India, women immediately understood that sex detection could allow them to opt for sex-selective abortion of female fetuses when they only needed a son (Ramanamma and Bambawale, 1980). Combined with abortion, prenatal sex diagnosis provided a unique tool to implement a more systematic form of sex selection (Miller, 2001).

In the former socialist world, from Eastern Europe to Vietnam, access to modern technology emerged only with the introduction of the new private health care facilities that proved more flexible to respond to the emerging demand for sex selection. In some other countries, the main issue remains access to abortion, which may be strictly outlawed or limited to lifesaving situations. Ultrasonography has gradually become both relatively easy to implement and inexpensive after a rapid fall in the price of scanning machines. More recently, new methods have emerged. Some allow the early detection of the sex of the fetus, using blood of women during the second month of pregnancy. Other high-tech techniques are linked to *in vitro* fertilization or preimplantation genetic diagnosis. These techniques are getting common in industrialized countries, but remain often inaccessible in Asia. Yet, what these new developments suggest is that future advances in reproductive technologies will inevitably make prenatal sex selection easier and that abortion may soon be outdated as a method to avoid unwanted girls (Dondorp et al., 2013).

Fertility level is the last of the three preconditions of this analytical framework and is the easiest to understand. It stems from the impact of fertility decline on the gender composition of the family. With six children, the odds that parents get six daughters and remain without a son are indeed very low (1.4%). But with two children, the risk becomes much higher (about 24%). In other words, parents who wanted a son in the past were most likely to have one after successive births. But as fertility declines, the probability of parents remaining sonless increases. At the same time, women in need of a son are also less ready to undergo repeated pregnancies and to have unwanted daughters. Increasingly risk-adverse and economically conscious parents refuse to submit themselves to a biological roulette and are now in a position to decide both the quality and the quantity of children they will have.

In addition to these social and economic considerations, family planning regulations introduced in several countries such as China and Vietnam have increased the 'marginal cost' of an additional child through various fines and penalties. Some have projected that fertility control in China was the main cause for sex selection. Yet, the skewed sex ratio also observed among other Chinese populations in Taiwan or in North America suggests that family planning was only one factor behind skewed sex ratios.

The Implications of Sex Selection

Gender discrimination plays out in two sequences. The first one is the immediate disappearance of girls at birth (or during early childhood) while the second one refers to future demographic disturbances among adults. The routine elimination of girls before their birth is no doubt the most extreme symptom of women's contemporary undervaluation that goes well beyond all other forms of discrimination. Since Amartya Sen's pioneer paper on missing women (Sen, 1990), we can measure the extent of the resulting gap across populations. Within families, the imbalance is mostly invisible since parents who abort female fetuses often have had prior girls. But when compared across larger communities, figures point inescapably to a surplus of boys. It is in fact possible to get a better idea of the actual rift between male and female by using demographic estimates on the age and sex distribution

for entire populations. 'Missing women' are women that ought to be found in population censuses if the sex distribution was not skewed toward men. By applying sex ratios by age computed for nondiscriminating countries, from Latin America to Europe, and correcting for age and mortality variations, it is possible to assess the number of women that are apparently missing from demographic records in Asia and Eastern Europe.

Using such a technique in conjunction with the United Nations Population Division population estimates, we obtain a total gender gap of 117 million women missing in 2010 (Table 2). Most of them should be found in China (66 million) and India (43 million). This large figure reflects in part the recent rise in sex-selective abortions, but also the cumulated toll of excess female mortality in the past. It is equivalent to the female population of Brazil or Indonesia, respectively the fifth and fourth most populous countries in the world. It represents about 8% of the female population in affected countries.

Among these missing women, we find 39 million of them would be below 20 years of age. This corresponds to the cohorts born since 1990, covering roughly the period when prenatal sex determination started to spread outside industrialized countries. The proportion of missing girls is especially high in China where they account for 14% of the women below 20 years of age. In India as a whole, where prenatal sex selection is less frequent, 13 million girls can be estimated as missing, but the contribution of excess mortality to this gap is also significant in India. China and India taken together account for 95% of the gender gap among the youth. The gender gap remains close to or above 5% in the Republic of Korea and in East-European countries.

A second way to look at sex imbalances at birth consists in examining their longtime consequences. This analysis starts with a set of demographic projections, in which the SRB is adjusted according to different future scenarios. These demographic simulations provide a precise description of future distortions in the age and sex structures of affected countries such as China and India. Of primary interest is the ripple effect of current sex imbalances at birth onto the young adult population after two or three decades. Projections show that even in the optimistic scenario of a rapid return of birth masculinity back to normalcy, men in China and India are bound face a serious 'marriage squeeze' in the years to come. Men will vastly outnumber women of marriageable age and saturate the marriage market for several decades, with a 60% surplus of prospective grooms over potential brides. Demographers have worked different adaptive scenarios whereby men would defer marriage by several years, but the magnitude and duration of birth imbalances preclude any demographic adjustment (Guilmoto, 2012a). The size of this 'male surplus' will be clearly too big to be offset by international migrations of potential brides or grooms. Moreover, trends observed elsewhere in East Asia or Western Europe suggest that women are likely to marry older and in smaller proportions in the future, a change bound to exacerbate the squeeze effect of their demographic deficit on the marriage market. Consequently, not only will a lot of men have to marry much later than in the past, but a rising proportion among them will also not be able to find a wife at all. The necessary readjustment of marriage systems will happen in patrilineal countries where family reproduction precisely depends on the male line and where only a small proportion of men remained unmarried in the past. Compared to Western societies where both male and female singlehood have long existed and where same-sex unions are gradually legalized, marriage systems in patriarchal societies are rather inflexible and allow limited leeway in case of a significant male surplus in the marriage market.

This unheard of situation has given rise to many speculations about its upcoming repercussions, ranging from rising male homosexuality and HIV AIDS contamination to human trafficking and international conflicts (Hudson and den Boer, 2004). But research recently conducted in some of the most vulnerable areas of China and India confirms the mounting pressure in women-scarce rural areas with significant social, sanitary, and demographic consequences on affected communities (Kaur, 2013; Li et al., 2010). Men resort to bride import from other, poorer regions when they can, and have at times to rely on existing criminal organizations for that purpose. An even higher proportion of men from underprivileged backgrounds seem condemned to involuntary celibacy, with married life becoming a new unattainable status because of their economic or social situation. In fact, if sex selection has mostly affected girls of higher status – since prenatal gender bias is more common among the rich – the price will be later paid mostly by men from the poorest strata. Sex selection is often considered a typical environmental Tragedy of the Commons in which opportunistic behavior by rational individuals leads later to the depletion of the resources and to a collective catastrophe. But a closer gender and class analysis suggests that the dividends and retributions stemming from the manipulation of biology will not be equally and simultaneously shared across society.

Table 2 Gender gap (missing women) by country in 2010

	Number of missing women			
	All women		*Women below 20 years*	
	In thousands	*%*	*In thousands*	*%*
Afghanistan	1097	7.2	502	5.7
Albania	40[a]	2.5	15	3.1
Armenia	31[a]		31	7.4
Azerbaijan	104		104	7.8
Bangladesh	2020	2.8	354	1.2
China	66 163	10.3	23 687	14.2
Hong Kong	27[a]		27	4.4
Georgia	19[a]		19	3.8
India	43 267	7.3	13 197	5.6
Montenegro	2[a]		2	2.7
Nepal	114[a]		114	1.7
Pakistan	2907	3.4	281	0.7
Singapore	114	4.5	11	1.9
South Korea	533	2.2	260	4.8
Vietnam	245[a]		245	1.7
Total	116 685	7.7	38 850	7.6

Numbers given in thousands.
Percentages computed over the corresponding female population.
[a]No measurable gender gap among adults.
Source: Guilmoto, C.Z., 2012b. Sex Imbalances at Birth. Current Trends, Consequences and Policy Implications, UNFPA, Bangkok.

Future Directions in SRBs

No one believes though that current sex imbalances at birth reflect a new demographic equilibrium and that they will persist indefinitely. Skewed sex ratios appear socially unsustainable in the long run. But this analysis suggests that prenatal sex selection depends on the conjunction of three preconditions and some are unlikely to improve in the near future Access to sex selection technology will continue to expand, even if some countries have actively attempted to ban prenatal sex diagnosis. Techniques to influence the sex of the births will continue to improve and spread across the world, making sex selection easier to implement than today. With countries offering state-of-the-art preimplantation technologies such as the United States or Taiwan, we may even expect the development of a new form of sex selection tourism. In a similar fashion, there is no reason to expect any significant turnaround in the low fertility situation that initially encouraged parents to plan the sex composition of their family. Even where stricter family planning regulations are relaxed in East Asia, low and ultralow fertility will largely remain and prevent parents from having more children in the quest for a son.

A reduction in son preference is therefore the only factor likely to bring about a lasting change. The bias against girls seems to have hardly changed over the last decades in view of persistent high sex ratios, but there are signs of a downturn. South Korea's experience in sex selection shows that the preference for boys over girls can recede and bring the sex ratio down to a normal level. In this country, patriarchal kinship institutions inherited from Confucian traditions have given way to less asymmetrical family structures in which daughters play a greater role. This evolution has been mostly attributed to the effect of social and economic change and to the new educational and employment opportunities offered to women by the rapid development that took place since the 1970s (Chung and Dasgupta, 2007). The prohibition of sex selection was also introduced and enforced right from the 1980s. In fact, the policies recently set up by many governments in other countries to enforce gender equity in social relationships and in family laws and sensitization campaigns can only accelerate this transition away from entrenched male bias. Apart from South Korea, there are now discernible indications of decline in birth masculinity levels in several parts of India and China, which suggests that the downturn, however slow, is probably already underway. The overall picture remains complex since in spite of a potential decline in the two largest world populations, prenatal sex selection might still spread in the future to new countries in Africa or in the Middle East when fertility declines and access to prenatal reproductive technologies progresses.

See also: Dowries and Other Marital Transfers; East Asian Studies: Gender; Fertility Transition: China; Fertility Transition: East Asia; Fertility Transition: South Asia; Gender and Feminist Studies in Geography; Gender and Place; Gender and Women's Studies, Applied Research On; Gender, Economics of; Gender: Anthropological Aspects; Russian and Post-Soviet Studies: Gender; Sex Selective Abortion.

Bibliography

Attané, I., Guilmoto, C.Z. (Eds.), 2007. Watering the Neighbour's Garden. The Growing Demographic Female Deficit in Asia. The Committee for International Cooperation in National Research in Demography, Paris.

Chahnazarian, A., 1988. Determinants of the sex ratio at birth: review of recent literature. Social Biology 35 (3–4), 214–235.

Chakraborty, T., Kim, S., 2010. Kinship institutions and sex ratios in India. Demography 47 (4), 989–1012.

Chung, W., Dasgupta, M., 2007. The decline of son preference in South Korea: the roles of development and public policy. Population and Development Review 33 (4), 757–783.

Croll, E., 2000. Endangered Daughters: Discrimination and Development in Asia. Routledge, New York.

Dondorp, W., et al., 2013. ESHRE task force on ethics and law 20: sex selection for non-medical reasons. Human Reproduction 28 (6), 1448–1454.

Dubuc, S., Coleman, D., 2007. An increase in the sex ratio of births to India-born mothers in England and Wales: evidence for sex-selective abortions. Population and Development Review 33 (2), 383–400.

Guilmoto, C.Z., 2009. The sex ratio transition in Asia. Population and Development Review 35 (3), 519–549.

Guilmoto, C.Z., 2012a. Skewed sex ratios at birth and future marriage squeeze in China and India, 2005–2100. Demography 49 (1), 77–100.

Guilmoto, C.Z., 2012b. Sex Imbalances at Birth. Current Trends, Consequences and Policy Implications. UNFPA, Bangkok.

Guilmoto, C.Z., Duthé, G., 2013. Masculinization of births in Eastern Europe. Population & Societies 506, 1–4.

Hudson, V.M., den Boer, A.M., 2004. Bare Branches: Security Implications of Asia's Surplus Male Population. The MIT Press, Cambridge, MA.

John, M.E., et al., 2008. Planning Families, Planning Gender: The Adverse Child Sex Ratio in Selected Districts of Madhya Pradesh, Rajasthan, Himachal Pradesh, Haryana, and Punjab. Action Aid and IDRC, New Delhi.

Kaser, K., 2008. Patriarchy after Patriarchy. Gender Relations in Turkey and in the Balkans 1500–2000. LIT Verlag, Vienna.

Kaur, R., 2013. Mapping the adverse consequences of sex selection and gender imbalance in India and China. Economic & Political Weekly 48 (35), 37–44.

Li, S., et al., 2010. Male singlehood, poverty and sexuality in rural China: an exploratory survey. Population 65 (4), 679–693.

Miller, B., 2001. Female-selective abortion in Asia: patterns, polices, and debates. American Anthropologist 103 (4), 1083–1095.

Murphy, R., et al., 2011. Son preference in rural China: patrilineal families and socioeconomic change. Population and Development Review 37 (4), 665–690.

Park, C.B., Cho, N.H., 1995. Consequences of son preference in a low-fertility society: imbalance of the sex ratio at birth in Korea. Population and Development Review 21 (1), 59–84.

Ramanamma, A., Bambawale, U., 1980. The mania for sons: an analysis of social values in South Asia. Social Science & Medicine. Part B: Medical Anthropology 14 (2), 107–110.

Sen, A., 20 December 1990. More than 100 million women are missing. New York Review of Books 37, 61–66.

United Nations, 2011. Sex Differentials in Childhood Mortality. United Nations Department of Economic and Social Affairs and Population Division, New York.

United Nations Population Division, 2013. World Population Prospects: The 2012 Revision. United Nations, New York.

Zeng, Yi, et al., 1993. Causes and implications of the recent increase in the reported sex ratio at birth in China. Population and Development Review 19 (2), 283–302.

Mission: Proselytization and Conversion

Brian Stanley, Westminster College, Cambridge, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 9903–9906, © 2001, Elsevier Ltd.

Abstract

This article concentrates on the concept and practice of mission within the Christian tradition, and on interpretations of missionary activity by historians, social scientists, and theologians. Attention is paid to the origins of the terminology of Christian missions in the Jesuit order in the sixteenth century and to the more recent history of 'mission' as a theological concept within the churches. Developments in historical and anthropological scholarship in this field include the shift from earlier approaches which stressed the material connections of missions with the colonial powers toward a concern with elucidating the complex patterns of religious and cultural exchange created by missionary encounters. These trends are related to recent scholarship in African, South Asian, Chinese, and Latin American studies. Mention is also made of the growing interest by sociologists and feminist historians in the significance of missionary organizations for the sending and receiving cultures in modelling both conservative and innovative patterns of gender relations. The article concludes by noting attempts by many Christian theologians to dissociate the concept of mission from proselytization.

Mission: Proselytization and Conversion

The concept of a universal mission to the world is embedded in different ways in the monotheistic religious traditions of Judaism, Christianity, and Islam. Judaism, though its scriptures contain strong elements of universalism, has in modern times ceased to be a missionary religion. Islam, by contrast, continues to win converts in large numbers, in part because of the appeal to peoples threatened by modernization of the Muslim concept of the immutability of the Shari'a or divine law. In this article 'mission' refers primarily to the activity of the churches in commending the doctrinal beliefs and moral values of Christianity as worthy of universal acceptance. 'Proselytization' (or proselytism) derives from the Greek *proselytos* – the term used by the Septuagint (the Greek version of the Old Testament) to describe a non-Jew who adopted the Israelite faith in Yahweh – and has come to mean the recruitment of converts from other religious systems and communities. Proselytization traditionally has been regarded as central to Christian mission, but does not necessarily constitute its totality. Medical and relief work, Christian education, the promotion of agricultural development, and agitation for a just international economic order, are now seen by Christians as integral parts of the mission to bring the world into conformity with Christian values. This article surveys the history of the term 'mission' and the missionary activities of the churches since the eighteenth century, reviews recent scholarship on the missionary movement and conversion, and notes attempts by many theologians to dissociate the concept of mission from proselytization.

The History of 'Mission' and 'Missions'

The Terms 'Mission' and 'Missions'

The use of the Latin term *missio* to describe the sending of Christians to propagate the faith among unbelievers was pioneered by Ignatius Loyola in the sixteenth century. Senior members of the Jesuit order vowed to go wherever the Pope might send them, to 'missions' among either the Catholic faithful or infidels, to promote the glory of God and the good of souls (Ross 1994). The widespread application of the singular vernacular term 'mission' to the overall activity of the church in the world is, however, more recent, dating only from the 1950s, when both political challenges and theological questioning compelled missionary theorists to re-evaluate precisely what it was that God had supposedly sent the church into the world to do. Until then, the plural term 'missions' was much more common, denoting the range of agencies deployed by both Catholic and Protestant churches to pursue the agreed aim of converting non-Christians. That usage remains widespread among American Christians and also among historians and social scientists concerned to investigate Christian missions simply as a human phenomenon.

The Modern Missionary Movement

Christian missions up to the mid-eighteenth century were conducted chiefly by the Roman Catholic and Russian Orthodox churches (the latter mainly in Siberia). The Roman Catholic missionary endeavor lost impetus following the suppression of the Jesuit order by the papacy in 1773, and the initiative passed to Protestants influenced by German pietism and the ensuing movements of evangelical revival which spread from central Europe to North America and Britain. The Protestant awakening culminated in the formation of numerous voluntary societies, beginning with the Baptist Missionary Society (1792) in Britain, the American Board of Commissioners for Foreign Missions (1810) in America, and the Basel Mission (1815) in continental Europe. For much of the nineteenth century these societies experienced only limited success in terms of both domestic recruitment and conversions overseas. From the late 1870s this picture changed as missionary enthusiasm converged with forms of

popular imperialism, and as indigenous peoples became increasingly attracted by the literacy and access to literature (both vernacular and English) which the missions offered. By this time, European Roman Catholic missions had also revived, stimulated by the recovery of the papacy and the formation of new religious orders, such as the White Fathers, founded by Cardinal Lavigerie in 1868. Competition between Protestant and Catholic missions was frequently intense, and often coincided with imperial rivalries. North America was regarded by the Catholic Church as technically a mission field until 1908, and the first American Catholic missionary order, the Maryknoll Fathers and Brothers, was not founded until 1911.

By 1910 Protestant foreign missionaries working in Africa, Asia, and the Pacific numbered some 19 000, of whom 71% were male or female lay persons. Catholic missionary numbers were almost certainly larger. In both Protestant and Catholic missions, however, the bulk of the work of evangelization was performed by indigenous (and primarily lay) Christians. After the First World War, the Americans supplanted the British as the predominant force in Protestant missions, though the older denominational boards progressively were eclipsed by newer interdenominational missions of evangelical or fundamentalist inclination (Hutchison 1987). Since the 1960s, many western mission agencies have dwindled in size, but their place has been taken by new missions from the nonwestern world, notably from Korea. By 2000, the total Christian foreign missionary force had risen to an estimated 420 000.

Recent Scholarship on Christian Missions

The Historiography of Missions

Until the late 1960s most writing on Christian missions tended to be confessional, often hagiographical, rather than scholarly. This was despite the fact that missionaries had made substantial contributions to the development of secular disciplines such as linguistics, Sanskritic and oriental studies, and anthropology (Walls 1996). Since the 1960s historical and anthropological scholarship on the missionary movement has proliferated, prompted by heightened awareness of the importance of religion in colonial encounters and the shaping of indigenous nationalist movements. Some of the pioneering African historiography of the 1960s identified a growth in racist and colonialist attitudes among late Victorian missionaries as crucial in the origins of West African nationalism (e.g., Ajayi 1965). The anticolonial reaction of the 1960s and 1970s encouraged writing by historians, anthropologists, and theologians that interpreted the missionary movement as the ideological expression of western colonialism. Such interpretations were able to point to the close connections between missions and western imperial expansion in contexts such as China, and to the consistent use by missions of terminology which branded nonwestern peoples as benighted heathens who needed 'civilizing' by the supposedly Christian west.

Since the early 1980s a strengthening reaction to, or at least modification of this tradition, has been evident. Scholars from a variety of disciplines (including, notably, literary studies) have become increasingly dissatisfied with models which erect a simple dichotomy between 'colonialist' missions and the 'victims' of colonialism (see *Religion: Culture Contact*). As the focus of imperial historiography has shifted from the materialist preoccupations of an earlier generation of Marxist-influenced historiography to a characteristically postmodern concern with the power of narrative and the construction of mentality and identity, scholarship has become less interested in the formal connections of missions to the apparatus of colonial rule and more concerned to elucidate the complex patterns of religious and cultural exchange created by missionary encounters (Peterson and Allman 1999). Rather than regarding indigenous converts simply as passive dupes of colonial pressure, scholars now are inclined to delineate a range of indigenous responses varying from outright rejection to transformative appropriations of the Christian message and the biblical text. This emphasis accords with the primacy of indigenous agents in transmitting the message, and with the fact that the rate of conversion to Christianity in the nonwestern world has accelerated, not slackened, since decolonization. Particular attention has been devoted to the cultural significance of vernacular translation, which committed missionaries, whether consciously or not, to a 'radical pluralism' wherein 'all languages and cultures are, in principle, equal in expressing the word of God' (Sanneh 1989, Walls 1996). The development of a vernacular literature, and especially a popular religious literature, to which mission Christianity contributed so fundamentally in nonwestern contexts, has been identified as the essential precondition for the transmutation of an ethnicity into a nation (Hastings 1997). The current academic trend does not, however, necessarily imply a more favorable estimation of the missionary impact. Perhaps its most influential exponents, the American anthropologists Jean and John Comaroff, despite an emphasis on the complexity and two-sided nature of missionary encounters in nineteenth-century southern Africa, continue to depict missionaries as 'colonial evangelists,' for whom language was not a medium of cultural liberation but rather a field of discourse whose terms they dominated in order to achieve 'the colonization of consciousness' (Comaroff and Comaroff 1991, 1997).

Regional Studies

The shift in focus toward the dynamics of conversion has been most marked in African studies, perhaps because of the remarkable growth of Christianity in sub-Saharan Africa (from some 60 million adherents in the 1960s to as many as 300 million by 2000). There are signs, however, of a similar trend in South Asian studies. The literary scholar, Gauri Viswanathan, though an adherent of Edward Said's school of 'Orientalism,' has argued that in British India, conversion to mission Christianity was a profoundly destabilizing influence, disrupting the stable categories of religion and nationhood on which the Raj depended, and compelling a professedly secular colonial administration to grapple with fundamental issues of civil rights (Viswanathan 1998). Studies of conversion by tribal and untouchable (or Dalit) peoples in India have followed precedents set by Africanist scholarship in interpreting conversion to Christianity as

a process, usually collective, which frequently offered those involved a more satisfactory framework of cosmological meaning and a more advantageous communal identity than the obvious alternative of adherence to Brahmanical versions of Hinduism. Frykenberg (in Hastings 1999) has observed that 'no mass movement of conversions in India was ever led by a European.'

Scholarship on conversion to Christianity in China, where the material linkages between missions, converts, and the protection of the imperial powers were more blatant than anywhere else, has been relatively slow to follow the trend set by the Africanists and South Asianists. Interpretations of Christianity in China during the missionary era have tended until recently to concentrate either on mission ideologies and strategies or on hostile antiforeign reactions to missions, rather than on the motivations of those Chinese who did convert (but for a recent collection which pioneers a more indigenous perspective on the Christian history of China see Bays 1996). The substantial growth of popular Christianity in Communist China since the era of the Cultural Revolution (1966–79) necessarily remains largely uncharted territory for scholarship.

Scholarly analysis of the impact of Christian missions on Latin America for long concentrated on the close connections between the Catholic Church and Iberian expansion which followed Columbus's 'discovery' of America in 1492 and the papal bull of 1493 which in effect divided ecclesiastical control of the continent between the Spanish and Portuguese crowns. More recently, attention has been turned to the phenomenal growth of popular Protestantism (especially Pentecostalism) in Latin America since the 1950s (see Christianity: Evangelical, Revivalist, and Pentecostal). The primary question currently being debated is parallel to that posed by writing on African Christianity during the colonial period: to what extent can the current success of Protestant proselytism in this most traditionally Catholic of continents be interpreted satisfactorily as a function of North American cultural and political influence, or should greater emphasis in fact be placed on the ability of Pentecostalism to meet the felt needs of the Latin American poor? (Hastings 1999, Martin 1990).

Sociological and Gender Studies

Within the social sciences, interest in missions has been for obvious reasons more pronounced among anthropologists than among sociologists. Very few sociologists have devoted attention to the unusual nature of mission organizations as typically voluntary bodies which have needed to maintain control at a formidable distance over highly motivated and frequently eccentric individuals, and at the same time satisfy the expectations of their supporting domestic constituencies. One such study of how religious zeal was both exploited and contained through institutional mechanisms examines the Basel Mission on the Gold Coast (now Ghana) in the nineteenth century (Miller 1994).

Much of the recent writing on missions within the social sciences examines the significance of the missionary movement from a feminist perspective. For most of the nineteenth century, Protestant missions were overwhelmingly male institutions. Most societies did not accept single women candidates, and missionary wives were not counted as missionaries in their own right, despite making in many cases major contributions to the work of the mission. The first Protestant missionary society specifically for women (the Female Education Society) dates from 1834. The first such societies were often inspired by the needs of the high-caste women of the Indian zenanas, who, being secluded from public contact with men, could be reached and educated only by women. From the late 1860s, female or zenana missionary societies became common on both sides of the Atlantic, either as autonomous institutions, or as female auxiliaries to existing denominational societies. These bodies enabled women to fulfil public roles both in their homelands and overseas which had few parallels in other contexts. From the 1870s, the denominational societies themselves began to accept single women candidates. They did so to varying extents, but the trend was accentuated by the readiness of many of the new nondenominational 'faith missions' to recruit single women (Robert 1996). By 1910 the missionary movement had become a predominantly female enterprise still controlled by men. Whilst still influenced by male assumptions about the proper restriction of women to the domestic sphere, it modeled a different reality. In that year, single women made up nearly 28% of the entire Protestant missionary force, and married women a further 28%. It seems likely that a similar gender distribution obtained for most of the twentieth century. The proliferation of missionary sisterhoods, first in France in the late nineteenth century, and then more widely, produced an even greater predominance of women in the Catholic missionary movement. Missionary women have thus become a major concern of scholars investigating what has been termed 'imperial feminism,' which is held to have combined 'colonial' views of racial difference with stances that were in practice subversive of traditional concepts of gender difference (Chaudhuri and Strobel 1992).

Mission and Proselytism in Theological Discourse

Whilst secular scholarship on missions has tended to move away from approaches which interpreted the proselytizing imperative as inescapably connected with the colonial system of material domination toward indigenous perspectives which emphasize the volitional agency of the convert or convert community, much Christian theology since the 1960s ironically has continued to regard coercion as intrinsic to the concept of proselytization, and has therefore been anxious to divorce 'mission' entirely from proselytism. One animating impulse of this concern has been the experience of the Roman Catholic, Orthodox, and other ancient eastern churches in finding their faithful to be the objects of often aggressive and well-funded evangelism by Protestant groups, particularly from the USA. A report jointly produced by the Roman Catholic Church and the World Council of Churches in 1970 sought to distinguish Christian witness from proselytism, defining the latter as 'improper' attitudes or behavior which employed coercion and thus violated the rights of the human person. Such a distinction has proved an unstable one, since almost any act of religious witness can be interpreted by the

recipient as being coercive or invasive. Since the collapse of Communist regimes in eastern Europe in 1989 and the dissolution of the Soviet Union, opposition to proselytism from the west has motivated legislation designed to curb the missionary work of those regarded by the Orthodox churches as 'sects.' More broadly, from the perspective of a Muslim, Hindu, Buddhist, or Jew, all Christian missionary activity is proselytizing in intention and effect, leading inevitably to the severance of the convert from community and even nation. Some Christian theologians accordingly have taken their opposition to proselytism toward its logical conclusion by denying that the conversion of adherents of other religions is a valid missionary objective for Christians. It then becomes difficult to resist the application of the same principle to the conversion of humanists or adherents of any belief system whatever, and the universalism intrinsic to the Christian theological tradition becomes unsustainable. Religious pluralism is no new phenomenon, but all religious believers in the western world now face unprecedented social and cultural pressures to accept the permanence and ideological absoluteness of pluralism. In the younger churches of the nonwestern world, whose role in world Christianity is becoming increasingly dominant, insistence on the inescapable connection between mission and proselytization remains strong. It may therefore be predicted that the twenty-first century is likely to see an even more heated debate within the churches about the validity and meaning of mission, proselytization, and conversion.

Bibliography

Ajayi, J.F.A., 1965. Christian Missions in Nigeria 1841–1891: The Making of a New Elite. Northwestern University Press, Evanston, IL.

Bays, D.H. (Ed.), 1996. Christianity in China: From the Eighteenth Century to the Present. Stanford University Press, Stanford, CA.

Chaudhuri, N., Strobel, M. (Eds.), 1992. Western Women and Imperialism: Complicity and Resistance. Indiana University Press, Bloomington, IN.

Comaroff, J., Comaroff, J., 1991. Of Revelation and Revolution. University of Chicago Press, Chicago.

Comaroff, J., Comaroff, J., 1997. Of Revelation and Revolution. University of Chicago Press, Chicago.

Hastings, A., 1997. The Construction of Nationhood: Ethnicity, Religion and Nationalism. Cambridge University Press, Cambridge, UK.

Hastings, A. (Ed.), 1999. A World History of Christianity. Eerdmans, Grand Rapids, MI.

Hutchison, W.R., 1987. Errand to the World: American Protestant Thought and Foreign Missions. University of Chicago Press, Chicago.

Martin, D., 1990. Tongues of Fire: The Explosion of Protestantism in Latin America. Blackwell, Oxford, UK.

Miller, J., 1994. The Social Control of Religious Zeal: A Study of Organizational Contradictions. Rutgers University Press, New Brunswick, NJ.

Peterson, D., Allman, J., 1999. Africans meeting missionaries. Journal of Religious History 23, 1.

Robert, D., 1996. American Women in Mission: A Social History of Their Thought and Practice. Mercer University Press, Macon, GA.

Ross, A.C., 1994. A Vision Betrayed: The Jesuits in Japan and China, 1542–1742. Edinburgh University Press, Edinburgh, UK.

Sanneh, L., 1989. Translating the Message: The Missionary Impact on Culture. Orbis Books, Maryknoll, NY.

Viswanathan, G., 1998. Outside the Fold: Conversion, Modernity, and Belief. Princeton University Press, Princeton, NJ.

Walls, A.F., 1996. The Missionary Movement in Christian History: Studies in the Transmission of Faith. Orbis Books, Maryknoll, NY.

Mixed Methods

Abbas Tashakkori, University of North Texas, Denton, TX, USA
Charles Teddlie, Louisiana State University, Baton Rouge, LA, USA
Burke Johnson, University of South Alabama, Mobile, AL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mixed methods research is an eclectic, pragmatic approach to employing combinations of research tools in answering multifaceted questions by seeking multiple, multilayered answers. It has gained in popularity since the 1980s as a third option to the either-or dichotomy (either quantitative or qualitative research) that had previously dominated the social and behavioral sciences. This article first presents three contrasting sets of beliefs that have characterized mixed methods in the contemporary literature. The article then discusses several issues in mixed methods: research questions (the dominance thereof), research designs (four families), data collection/analysis strategies (qualitative and quantitative), and postanalysis integration.

The field of mixed methodology (and social research methodology, in general) has changed so much in the past two decades that it looks chaotic at times. (We view *mixed methodology* as broader and different from *mixed methods (MM)* or *mixed methods research (MMR)*. The former is a set of concepts, rules, methods, norms, and expectations, including research ethics. The latter is the practice or concrete product of mixed methodology.) A quick look at how it is conceptualized within and across disciplines would easily reveal the immense diversity of worldviews, designs, data collection and analysis, permissible inferences, quality audits, and utilization approaches. Undoubtedly, a degree of diversity is a healthy indicator of a growing and lively community of scholars who utilize a methodological approach. On the other hand, without systematic attempts to create bridges (higher-order frameworks) between diverse conceptualizations and create a common language to facilitate communication within such a highly diverse community, a sense of (at least temporary) chaos is inevitable.

Mixed methodology was received well in 1980s as a reaction to dichotomizing research methodology, separating both the scholars and their approach to inquiry into two somewhat amorphous qualitative (QUAL) and quantitative (QUAN) categories and communities. The qualitative research community emerged and was clearly defined as a reaction to the dominant status quo in social science disciplines in which research was heavily centered on hypothesis testing, promise of objective answers to important social questions, and belief in isolating and finding clearly delineated causes for outcomes and effects. Qualitative research emerged as a community with a promise of searching for multiple explanations, from multiple perspectives, within sociopolitical contexts of occurrences. Cohorts of graduate students were caught in an either-or dilemma of selecting either quantitative or qualitative research, but not both. Journals oftentimes sent a similar message. Change is rarely smooth and without resistance. In the past two decades, there has been a clear 'us' and 'them' designation in many writings and references within both of the two broad communities of scholars.

The mixed methods 'movement' (as we and others have called it) was a reaction to the dichotomy. Rejection of the dominant dichotomy resulted in the possibility, desirability, and plausibility of conceptualizing mixed methods. This is seen in three somewhat different clusters of beliefs where (1) mixed methods is seen as a point on a multidimensional continuum with QUAL and QUAN orientations or packages of beliefs at the poles (QUAL———QUAL/quan———QUAL/QUAN———QUAN/qual———QUAN); (2) mixed methods is conceptualized as an interactive bridge between the qualitative and quantitative orientations (QUAL←→QUAN); and (3) mixed methods is considered either irrelevant or as an all encompassing orientation that considers the popular dichotomy (of QUAL and QUAN approaches) both false and irrelevant. Below, we consider each of these sets of 'beliefs' about mixed methodology (i.e., mental models or sets of cognitive 'statements' linking mixed methodology with valued attributes). Each of these mutually noninclusive sets has explicit or implicit advocates.

The first set of beliefs (the multidimensional 'continuum orientation') is held by those who consider each research project as an attempt to answer a research question or a set of interrelated research questions via *the* combination of QUAL and QUAN perspectives and methods appropriate to the questions (Johnson and Onwuegbuzie, 2004; Teddlie and Tashakkori, 2009). Virtually every research project, according to this type of conceptualization, inherently needs to include a combination of the two approaches in some degree or another. This belief system includes the differentiation of designs according to which approach is 'dominant' or takes the biggest intellectual space in the research project (i.e., leading to a notation system to classify designs in QUAL–quan, QUAN–qual, and similar forms). (Using notation from Morse (1991), QUAL–quan signifies a study in which qualitative research is core and quantitative research is subsidiary, and QUAN–qual signifies a study in which quantitative research is core and qualitative research is subsidiary. In both cases, it is important that the subsidiary approach provides an added value to the overall research study.)

According to the second set of beliefs, conceptualizing mixed methods as an interactive bridge between two distinct approaches (qualitative and quantitative research) involves

explicitly recognizing two identifiable and distinct sets of methods/approaches, each with its own community of scholars. These mixed methods projects employ the two sets of methods/approaches side by side (coexisting) or in a dialectic manner (Greene, 2007). Scholars in this community have tried to find common or symbiotic grounds between the two sets, both in terms of ideas and methods, in order to make it possible for the two communities of QUAL and QUAN researchers to understand each other, and, more specifically, so that both QUAL and QUAN can be used interactively in single research studies to add explanation and understanding to important research questions. Mixed methods research designs are conceptualized with the first and second belief sets in two or more parallel or sequential components or strands, with each strand (i.e., a combination of conceptualization, methodological, analytical, and inferential stages) conducted separately within one of the two broad methodological grounds and linked at some level during and/or at the end of the process.

The third set of beliefs has strong roots in the early twentieth century history of social research in which there was no discernable differentiation between the two sets of methods or approaches to research. We call this the traditionalist age, as it was the time when traditional empirical research approaches were being formulated. This also was a time when disciplines were beginning to clearly distinguish themselves into departments in universities and department names were being decided upon. Multiple perspectives, questions with multiple components (that one can identify today as QUAL or QUAN), multiple complementary/competing methods of study, and integrated inferences were all employed by prominent researchers in social sciences in early- and mid-twentieth century (e.g., Festinger et al., 1956; Lewin et al., 1939; Sherif et al., 1961). In this traditional research community, the issue of the QUAL/QUAN dichotomy is irrelevant, either as two discrete categories or as poles of a continuum. Separation of methods and perspectives is not recognized as a necessary component of planning and implementing a research project, since it is not how research is done (in a way, the two poles are collapsed, with every research study being both QUAL and QUAN, at the same time, depending on the questions asked!) (see Brewer and Hunter, 2006; Waszack and Sines, 2003 for examples).

Consistent with the inherent 'return to roots' characteristic of this group, this subcommunity in MM might be called 'research traditionalists' whose dominant tenet is that the research question (not the paradigm) dictates all aspects of the research inquiry, leading to taking multiple perspectives and using whatever methods are necessary to find answers. The term 'aparadigmatic' also has been used by a few to describe this group of scholars (Greene, 2007). Pragmatism has emerged over time as a philosophical anchor for this rather large subset of the community of mixed methods scholars. According to this group, good research is just that: no discussion is needed about whether it is qualitative, quantitative, or mixed. In its current version, the 'Q labels' are seen as having had their moment but now as having outlived their usefulness (Onwuegbuzie and Leech, 2005).

Several scholars (e.g., Brewer and Hunter, 2006; Erzberger and Kelle, 2003; Hunter and Brewer, 2003) have identified mixed methods studies from the formative years of the traditional period. For example, Hunter and Brewer (2003) presented three such studies: (1) the Hawthorne studies, which began in 1924 and continued for several years (Roethlisberger and Dickson, 1939); (2) the studies of 'Yankee City' (Newburyport, Massachusetts) as an example of community life in the United States (Warner and Lunt, 1941); and (3) the Marienthal study which was conducted during the 1930s economic depression in Austria, but not published in the United States until 40 years later (Jahoda et al., 1971). In the Hawthorne studies (Roethlisberger and Dickson, 1939), the researchers used interviews and observations as they tried to understand what became known as the 'Hawthorne effect.' The Hawthorne studies lasted several years, were conducted by multiple researchers, and the overall research program emphasized experiments *and* extensive interviewing, observation, and life history data. Their work focused on the formal/structural side of an organization and especially the informal/human side of organizations.

At least in this point in time, most writings about mixed methodology, and most mixed methods research projects, seem to include a combination of the three sets of beliefs. We see all three sets of beliefs in our own thinking. Mixed methodology, broadly conceived, is often considered today as a 'third option' to research for many scholars. It is a highly eclectic and pragmatic approach to employing useful combinations of tools (i.e., methods) for answering multifaceted questions (i.e., formulated from different perspectives) and explicitly seeking multilayered and multiple answers. The biggest degree of divergence today is probably in conceptualizing the meaning of mixing. Meanings range from linking two clearly recognized perspectives/outcomes (QUAL and QUAN) while keeping each separate with its own assumptions and quality audit, to integrated methodology that incorporates multiple perspectives without explicitly differentiating or keeping the QUAL and QUAN separate.

We appreciate diversity and variation in perspective, but we generally consider it unfortunate when calls are made for formulating research questions *after* adopting/declaring a perspective, imposing strict and simplistic designs (e.g., selecting a 'design name' for one's project from a small set of preconceived names/designs), and so on. There are also lingering doubts about the human capabilities of a single researcher to examine problems via multiple lenses. A useful antidote to this last issue might be to use research teams, instead of requiring a researcher to do all of the hard thinking and work.

General Characteristics of Mixed Methods Projects

As mentioned above, despite a respectable divergence in what we called 'beliefs' about mixed methods, mixed methods researchers have many commonalities. For example, they often identify themselves as integrative and eclectic. Many work within a pragmatic worldview and encourage paradigm diversity and diversity of the research enterprise at all levels because of its potential to spur new growth and scientific understanding and advancement. They emphasize continua, balance, and a cyclical, iterative approach to research. They collect both narrative and numeric data, employ both structured and

emergent designs, analyze their data both via statistical and content/meaning analysis, and integrate the results in order to reach 'metainferences' that integrate qualitative and quantitative findings to provide better answers to research questions. They also acknowledge the unavoidable impact of their culture, worldview, and sociopolitical context on their purposes and questions for inquiry, eventually leading to selected methods and an inference process based on empirical results (Newman et al., 2003). There also is a widespread acknowledgment within the community that not all research questions require a mixed approach. For example, Johnson and Onwuegbuzie (2004) put it like this:

> In general we recommend contingency theory for research approach selection, which accepts that quantitative, qualitative, and mixed research are all superior under different circumstances and it is the researcher's task to examine the specific contingencies and make the decision about which research approach, or which combination of approaches, should be used in a specific study.
>
> pp.22, 23

Research Questions in Mixed Methods Research

Research questions that call for mixed methods research are often multifaceted, having implicit or explicit interrelated components that might separately fit traditional qualitative or quantitative orientations. These combination questions often include both 'what and how' or 'what and why' of events, cognitions, and/or behaviors. Often, there is an overarching question that includes these general components, followed by a collection of separate qualitative- and quantitative-type subquestions. An example, from Tashakkori and Newman, 2010 is "What are the effects of teaching method X on achievement and self-perceptions of groups A and B?," with subquestions such as "Are groups A and B different in achievement and self-perceptions?," "What are the perceptions and constructions of participants in groups A and B regarding teaching method X and its impact?" and "How does teaching method X work differently in the two groups (if it indeed did)?"

Unique to mixed methods studies are questions regarding the nature of integration (i.e., how do the insight from multiple strands of the study, or the outcomes of data analysis in various strands contribute to an overall answer?). An example of such 'integration questions' might be "how do the two sets of answers contribute to a comprehensive understanding of the phenomenon under study?" Possible scenarios for answers to this question might be that the results complement each other (provide two alternative aspects/approaches to the phenomena), add to each other (more complete answers), or confirm/corroborate.

Mixed Methods Designs

Based on the planning/emergent process of research and types of questions, mixed methods designs can be conceptualized under three broad families of sequential, parallel, and conversion designs, leading to a fourth family (fully integrated) when combined (Tashakkori et al., 2013). Within each of these four design families, variation can occur on the basis of units of study (individuals, natural collections of units, or both), and the degree to which multiple strands (defined above) are planned versus emerge as the project is conducted (i.e., emergent vs preplanned questions and design) or both.

In sequential designs, later research questions can crystallize or other aspects of the study can emerge (such as Who needs to be sampled? What additional data need to be collected?) based on the findings from a previous strand. For example, participants in a randomized controlled/clinical trial who have not responded as expected might suggest questions and data sources for a next strand consisting of multiple case studies. In this example, the questions for the follow-up case study might emerge as a result of unexpected findings of the first strand, or they might have been developed in a preplanned manner (expecting lack of predicted response to intervention by at least some participants). Preplanned sequential designs predict certain outcomes (e.g., certain categories of responses, identification of certain subgroups based on the findings) would be necessary before the next phase of the study can be implemented.

Also in sequential mixed designs/studies, one might collect data in a separate/different sample of individuals for each strand, or select subsets of individuals from one strand to study in the next. The first case might also include selecting from two or more levels of a social system or organization, such as children to be tested in one strand, parents to be interviewed in another, in a multilevel mixed methods study. In short, in all these scenarios, the samples, procedures, instruments for data collection, or even data analysis techniques of a later strand are often dependent on the results obtained in a previous one.

Parallel mixed methods studies employ relatively independent strands of data collection and analysis, either simultaneously (concurrently), or with some time lag. The time frame for implementation does not change a parallel mixed project to a sequential one. The key differentiator of this group of designs from the previous (sequential) designs is that one strand is not dependent on the outcomes of another. Hence, the order and timing of implementation is flexible, partly depending on the practical logistics of the study. The findings and inferences emerging from multiple strands are integrated at some later point, after the data are collected and analyzed, in order to find a comprehensive answer to the initial multifaceted research questions. Different strands might use different samples at the same or multiple levels of a social system, or one strand might be limited to a subsample of the other one (e.g., a full sample is studied in one strand, while additional data are collected on a subsample in another concurrent strand).

In conversion mixed methods designs, the data (often narrative/qualitative) are analyzed both qualitatively and quantitatively, and inferences are made on the basis of both sets of analytic results. For example, themes are developed via content analysis of narrative data and then codes are assigned to these themes (or other aspects of narrative data) based on intensity of occurrence and the codes (as binary or categorical variables) are analyzed quantitatively (e.g., correlations, regression). In conversion designs, inferences are made on the basis of *both* sets of results (thematic analysis and statistical analysis) and the data and results are compared and contrasted or linked to each other. The data collected from an entire

sample might be analyzed via content and statistical methods separately. Alternatively, only the data from a subset might be converted and analyzed for subsequent strands.

Fully integrated designs incorporate a combination of the three basic types (parallel, sequential, conversion) in an interactive and dynamic manner. Variations in the sampling, data collection, and data analysis procedures that were just described above for the simpler families of designs are also applicable to this more dynamic and comprehensive family. For example, one might collect survey data in a large sample, then select two subsamples for the next step in order to collect observational data in one (second strand of the study), and then administer a personality test to another (third strand of the study), then analyze all the data in the three sets, then quantitize the observational data and statistically analyze it in a fourth conversion strand, and then finally link all four sets of outcomes together into metainferences. Two key points here are that (1) integration is integral to strong MMR and (2) MMR designs can be both creative and complex as we answer our research questions.

Data Collection and Analysis Strategies in Mixed Methods Research

Mixed methods researchers collect both qualitative/narrative and numerical data and/or convert one to another as an additional analytic strategy. The mixed data are collected either with a single method of data collection (e.g., questionnaires including both open- and closed-ended items), or using multiple methods of data collection (e.g., questionnaires and naturalistic observations at the workplace). The former is an example of *intramethod mixing* (i.e., using a mixed version of a single method of data collection) and the latter is an example of *intermethod mixing* (Johnson and Turner, 2003). Observations might be recorded by taking extensive notes to document as many of the interactions as possible, while also using highly structured protocols yielding numeric data. Sometimes, two types of data are collected sequentially in alternative intervals (e.g., 3 min of one, 3 min of the other, and so on). Similarly, one might conduct open-ended (qualitative, ethnographic) interviews that also have structured (closed-ended) items/components (i.e., intramethod mixing). Quantitative and qualitative approaches might be used in the same sample, or the more extensive data collection efforts might be limited to subsamples. Often, individual and/or focus group interviews start with broad questions and then increasingly focus on specific aspects or components of the phenomenon or behavior of interest. Data collection should always be tailored to the needs of one's research questions.

Depending on the nature of the data (e.g., narrative/visual vs numerical), appropriate analysis strategies are employed with each set. In sequential designs, each round of data collection and analysis appropriately and systematically follows the previous step. For example, if there is a plan/need to select specific types of units (e.g., effective schools, or better/worse than expected postsurgery recovery patients) for a study, the data collected in the previous strand must be analyzed accordingly (e.g., descriptive variable analysis and/or multiple regression to find regression residuals for selection of certain cases). In parallel mixed designs, the investigator might statistically compare experimental and control groups, while also, independently, conducting a thematic analysis of post-experimental interviews (or open observation notes if the individuals/groups were observed).

Postanalysis Integration

Being able to examine a phenomenon from multiple perspectives or vantage points (e.g., 'qualitative' and 'quantitative' perspectives or multiple stakeholder perspectives) is a definite advantage of mixed methods studies. There is a degree of comfort in finding consistency between the outcomes of such examinations, leading to greater confidence in the trustworthiness of the conclusions. However, consistency between the conclusions of multiple strands is not necessarily a goal and is not a requirement of a strong mixed methods inquiry. Integration of research results in mixed methods can also denote elaboration, complementarity, completeness, confirmation/disconfirmation, and so forth. Pending a thorough review of an 'inconsistency' to make sure it is not the result of data and/or analysis problems, the (inconsistent) outcomes can provide a broader and more complex understanding of the phenomenon under investigation by suggesting that (1) there are multiple aspects/dimensions of the same phenomenon, each abiding by a different set of rules/theories, (2) there is set of explanations that might provide the conditions for the applicability of the other, or (3) there are multiple plausible answers or stakeholder vantage points to the question.

The 'big picture' outcome of a mixed methods study depends highly on the degree of confidence in the conclusions/inferences made based on the findings of each stand, and their effective linking and integration. There are several frameworks for assessing the quality of mixed methods inferences. One approach to research quality suggests using separate evaluation criteria for the qualitative and quantitative components using the criteria associated with these components; this is seen in the 'multiple validities' type in Onwuegbuzie and Johnson (2006). *Multiple validities legitimation* ('validity' is viewed broadly) is the extent to which all of the pertinent validities (quantitative, qualitative, and mixed) are addressed and resolved successfully in a research study (Onwuegbuzie and Johnson, 2006). A 'cross-perspective' audit strategy also has been suggested (Tashakkori and Newman, 2010) in which qualitative standards/procedures (e.g., prolonged engagement, peer debriefing, triangulation, audit trail, negative case analysis, and others) are used to enhance the quality of inferences gleaned from quantitative findings, and the quantitative concepts (e.g., consistency between elements of the design, construct validity, investigator expectancy effects, and possible participant reactivity) are used for assessing the trustworthiness of the inferences based on qualitative findings. Another (and nonmutually exclusive) framework ('integrative framework,' O'Cathain, 2010; Teddlie and Tashakkori, 2009) provides an umbrella that includes the underlying logic of audits/strategies of both perspectives.

The umbrella (integrative) approach to assessing/maximizing the quality of inferences in mixed methods research studies has

focused on at least five components of strong inferences, and possible strategies to maximize each: (1) quality of research questions; (2) quality of research design components and their implementation; (3) quality of interpretation (of the findings); (4) quality of integration (of inferences from multiple strands of the study); and (5) quality of utilization, with specific attention to the value-added contributions of a mixed methods study to the quality of policy and practice, beyond simpler methodologies. Auditing and enhancing each of these five aspects of a mixed methods study incorporates known strategies from both QUAL and QUAN approaches, regardless of the type of data or findings. *Multiple validities legitimation* ('validity' is viewed broadly) is the extent to which all of the pertinent validities (quantitative, qualitative, and mixed) are addressed and resolved successfully in a research study (Onwuegbuzie and Johnson, 2006).

Given a reasonable degree of confidence in the quality of integrated inferences, transferability is the potential value-added contribution of the conclusions and recommendations for policy and practice. Transferability is broadly defined here as the degree of applicability or generalizability of conclusions of a mixed methods study to other individuals or entities (population transferability), other settings or situations (ecological transferability), other time periods (temporal transferability), other methods of observation/measurement (operational transferability), or in complex ways to subpopulations as seen in higher-level interaction effects (complexity transferability). Given the pragmatic nature of mixed methods studied, utilization quality is often evaluated in terms of transferability (contributions) to policy and practice. Practical theory (i.e., theory that works in particular places/contexts with particular people) is an ongoing goal of mixed methods research. Kurt Lewin (1951) expressed the spirit many years ago when he said "There is nothing so practical as a good theory" (p.169).

Conclusions

We started this article by discussing three sets of beliefs about mixed methodology (MM as a space on a multidimensional continuum of QUAL and QUAN approaches, as an interactive bridge between two separate and identifiable poles, or as an encompassing third option that does not consider the dichotomy as important for good research), with each leading to possibly different conceptualizations of mixed methodology. Regardless of an investigator's specific beliefs about mixed methodology, it is predominantly considered a pragmatic approach to problem solving, with the potential to provide a more diverse sets of tools than either of the two traditional approaches (i.e., QUAL and QUAN). Although not agreed by all, we contend that in mixed methods research, the research question, not the paradigm or purpose, should dictate the methods and strategies used in mixed inquiry. We are aware that some in the field advocate deciding on a paradigm or purpose for using integrated methodology (e.g., complementary, dialectical, transformative), rather than letting the research question drive the decision process. We wonder: Does it matter what the purpose of mixing was if the outcomes are clear and enlightening?

To us, the driving force behind planning a mixed methods study is the answer to an initial question: "Does the research question/problem beg for a mixed approach, instead of a simpler approach?" In varying degrees of contrast, there are suggestions (actually mandates!) by some others encouraging the investigators to first decide on why they value using a mixed approach (i.e., to have stronger inferences, to complement the findings of one with the other), then proceed to formulating research questions that fit within those values. Similarly, there are suggestions to declare a paradigm/worldview up front, then formulate one's research questions. Although we acknowledge the importance of being aware of one's worldview and/or purpose for a study, we consider them as parts of an investigator's engagement in the process of reflection about the research questions (i.e., reflective inquiry; Greene, 2007) and the entire study, throughout the entire project.

A lingering issue in mixed methodology is that of a priori decisions regarding 'dominance' of one approach or another in inquiries. This is largely consistent with the multidimensional continuum belief mentioned early in this article (e.g., QUAL + quan and QUAN + qual approaches). Although it might be useful to remain aware of where on the continuum of the combination of QUAL and QUAN a component falls, we do not have a priori any indication of its usefulness in terms of contributions to the quality of research inferences and the potential of transferability/utilization of the outcomes. It is not clear if/when such decisions are essential in planning of a project, or why/how doing so enhances the outcomes. We wonder: Does it matter which approach was dominant (or not), if the outcomes are useful for one's research questions?

Challenges and Opportunities in Mixed Methods Research

A challenge for mixed methods researchers is that they must learn specific methods, procedures, and techniques that have been traditionally identified as being qualitative or quantitative. They also must learn how to integrate understandings gleaned from these complementary or competing methods and approaches to reach metainferences (i.e., inferences or conclusions that build on or *integrate* quan and qual findings) that better inform researchers, professionals, and stakeholders. The challenges are substantial. The possible rewards are, however, even greater! Mixed methods research has the capability of providing understandings from multiple perspectives, deeper and wider than the two traditional approaches might be capable of providing. In many respects, mixed methods research uses the same strategies as everyday problem solvers, but in a more systematic manner (Tashakkori and Teddlie, 2010). Everyday problem solving includes examination of occurrences from multiple perspectives (Gorard, 2010), incorporating both subjective/social/attitudinal and objective/physical/stable evidence into consideration in finding answers to questions.

See also: Applied Social Research, History of; Interpretive Methods: Micromethods; Methodology of the History of the Social and Behavioral Sciences; Pragmatist Social Thought, History of; Problem Selection in the Social Sciences: Methodology; Qualitative Methods, History of; Qualitative Research; Small Group Research, History of.

Bibliography

Brewer, J., Hunter, A., 2006. Foundations of Multimethod Research: Synthesizing Styles, second ed. Sage Publications, Inc., Thousand Oaks, CA.

Erzberger, C., Kelle, U., 2003. Making inferences in mixed methods: the rules of integration. In: Tashakkori, A., Teddlie, C. (Eds.), Handbook of Mixed Methods in Social and Behavioral Research. Sage Publications, Inc., Thousand Oaks, CA, pp. 457–490.

Festinger, L., Riecken, H.W., Schacter, S., 1956. When Prophecy Fails. University of Minnesota Press, Minneapolis, MN.

Gorard, S., 2010. Research design, as independent of methods. In: Tashakkori, A., Teddlie, C. (Eds.), Sage Handbook of Mixed Methods in Social and Behavioral Research. Sage, Thousand Oaks, CA, pp. 237–252.

Greene, J.C., 2007. Mixed Methods in Social Inquiry. Wiley, San Francisco, CA.

Hunter, A., Brewer, J., 2003. Multimethod research in sociology. In: A.Tashakkori, Teddlie, C. (Eds.), Handbook of Mixed Methods in Social and Behavioral Research. Sage Publications, Inc., Thousand Oaks, CA, pp. 577–594.

Jahoda, M., Lazarsfeld, P.E., Zeisel, 1971. Marienthal: The Sociology of an Unemployed Community. Aldine, Chicago, IL.

Johnson, R.B., Onwuegbuzie, A.J., 2004. Mixed methods research: a research paradigm whose time has come. Educational Researcher 33 (7), 14–26.

Johnson, R.B., Turner, L.A., 2003. Data collection strategies in mixed methods research. In: Tashakkori, A., Teddlie, C. (Eds.), Handbook of Mixed Methods in Social and Behavioral Research. Sage, Thousand Oaks, CA, pp. 297–319.

Lewin, K., 1951. In: Cartwright, D. (Ed.), Field Theory in Social Science: Selected Theoretical Papers. Harper & Row, New York.

Lewin, K., Lippitt, R., White, R.K., 1939. Patterns of aggressive behavior in experimentally created "social climates". Journal of Social Psychology 10, 271–299. http://dx.doi.org/10.1080/00224545.1939.9713366.

Morse, J.M., 1991. Approaches to qualitative-quantitative methodological triangulation. Nursing Research 40, 120–123.

Newman, I., Ridenour, C.S., Newman, C., DeMarco, G.M.P., 2003. A typology of research purposes and its relationship to mixed methods. In: A.Tashakkori, Teddlie, C. (Eds.), Handbook of Mixed Methods in Social and Behavioral Research. Sage Publications, Inc., Thousand Oaks, CA, pp. 167–188.

O'Cathain, A., 2010. Assessing the quality of mixed methods research: towards a comprehensive framework. In: Tashakkori, A., Teddlie, C. (Eds.), Sage Handbook of Mixed Methods in Social and Behavioral Research. Sage, Thousand Oaks, CA, pp. 531–558.

Onwuegbuzie, A.J., Johnson, R.B., 2006. The "validity" issue in mixed methods research. Research in the Schools 13 (1), 48–63.

Onwuegbuzie, A.J., Leech, N.L., 2005. Taking the "Q" out of research: teaching research methodology courses without the divide between quantitative and qualitative paradigms. Quality and Quantity 39, 267–295.

Roethlisberger, F.J., Dickson, W.J., 1939. Management and the Worker. Harvard University Press, Cambridge MA.

Sherif, M., Harvey, O.J., White, B.J., Hood, W.R., Sherif, C.W., 1961. Intergroup Conflict and Cooperation: The Robber's Cave Experiment. University of Oklahoma Institute of Intergroup Relations, Norman, OK.

Tashakkori, A., Newman, I., 2010. Mixed methods: integrating quantitative and qualitative approaches to research. In: McGaw, B., Baker, E., Peterson, P.P. (Eds.), International Encyclopedia of Education, third ed. Elsevier, Oxford, UK.

Tashakkori, A., Teddlie, C., 2010. Putting the human back in "human research methodology": the researcher in mixed methods research. Journal of Mixed Methods Research 4, 271–277.

Tashakkori, A., Teddlie, C., Sines, M.C., 2013. Mixed methods in psychological research. In: Weiner, I., Shinka, J.A., Velicer, W.F. (Eds.), Handbook of Psychology, Research Methods in Psychology, second ed., vol. 2. Wiley.

Teddlie, C., Tashakkori, A., 2009. Foundations of Mixed Methods Research: Integrating Quantitative and Qualitative Approaches in the Social and Behavioral Sciences. Sage, Thousand Oaks, CA.

Warner, W.L., Lunt, P.S., 1941. The Social Life of a Modern Community. Yale University Press, New Haven, CT.

Waszack, C., Sines, M., 2003. Mixed methods in psychological research. In: Tashakkori, A., Teddlie, C. (Eds.), Handbook of Mixed Methods in Social and Behavioral Research. Sage, Thousand Oaks, CA, pp. 557–576.

Mixture Models in Statistics

Geoffrey J McLachlan, University of Queensland, Brisbane, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by U.E. Makov, volume 15, pp. 9910–9915, © 2001, Elsevier Ltd.

Abstract

Finite mixtures of distributions have provided a mathematical-based approach to the statistical modeling of a wide variety of random phenomena. They underpin a variety of techniques in major areas of statistics, including cluster and latent class analyses, discriminant analysis, image analysis, and survival analysis, in addition to their more direct role in data analysis and inference of providing descriptive models for distributions.

Introduction

Finite mixture distributions are typically applied to data with two main purposes in mind:

1. to provide an appealing semiparametric framework in which to model unknown distributional shapes, as an alternative to, say, the kernel density method (the set of all normal mixture densities is dense in the set of all density functions under the L1 metric);
2. to provide a probabilistic clustering of the data into g clusters with an outright clustering obtained by assigning each data point to the component to which it has the greatest posterior probability of belonging.

In both situations, there is the question of how many components to include in the mixture.

One of the first major analyses involving the use of mixture models was undertaken over 100 years ago by the famous biometrician Karl Pearson. In his now classic paper (Pearson, 1894), he used the method of moments to fit a mixture of two normal probability density functions with different means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 in proportions π_1 and π_2 to some measurements on the ratio of forehead to body length of $n = 1000$ crabs sampled from the Bay of Naples. The moment estimates of the five parameters in this two-component normal mixture were obtained through the solution of a ninth-degree polynomial (nonic).

Not surprisingly, various attempts were made over the ensuing years to simplify Pearson's moments-based approach to the fitting of a normal mixture model. In the 1960s, the fitting of finite mixture models by maximum likelihood (ML) had been studied in a number of papers, including the seminal papers by Day (1969) and Wolfe (1970). However, it was the publication of the seminal paper of Dempster et al. (1977) on the expectation–maximization (EM) algorithm that greatly stimulated interest in the use of finite mixture distributions to model heterogeneous data. This is because the fitting of mixture models by ML is a classic example of a problem that is simplified considerably by the EM's conceptual unification of ML estimation from data that can be viewed as being incomplete.

Finite mixture models have since been increasingly used to model the distributions of a wide variety of random phenomena and to cluster data sets; see, for example, the recent books by Böhning (1999), Frühwirth-Schnatter (2006), McLachlan and Peel (2000), Schlattmann (2009), and Mengersen et al. (2011), and the references therein. Earlier references on mixture models may be found in the previous books by Everitt and Hand (1981), Lindsay (1995), McLachlan and Basford (1988), and Titterington et al. (1985); see also the papers by Banfield and Raftery (1993) and Fraley and Raftery (2002).

Notation

We let $\mathbf{Y}$ denote a random vector consisting of p feature variables associated with the random phenomenon of interest. We let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ denote n independent observations on $\mathbf{Y}$. With the finite mixture model-based approach to density estimation and clustering, the density of $\mathbf{Y}$ is modeled as a mixture of a number (g) of component densities $f_i(\mathbf{y})$ in some unknown proportions $\pi_1, \ldots, \pi_g$. That is, each data point $\mathbf{y}_j$ is taken to be a realization of the mixture density

$$f(\mathbf{y}_j) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{y}_j) \qquad [1]$$

where the g components correspond to the g groups. In eqn [1], the $f_i(\mathbf{y}_j)$ are densities and the π_i are nonnegative quantities (the mixing proportions) that sum to one. The $f_i(\mathbf{y}_j)$ are called the *component densities* of the mixture.

On specifying a parametric form $f_i(\mathbf{y}_j;\boldsymbol{\theta}_i)$ for each component density, we can fit the parametric mixture model

$$f(\mathbf{y}_j;\boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{y}_j;\boldsymbol{\theta}_i) \qquad [2]$$

by ML via the EM algorithm of Dempster et al. (1977); see also McLachlan and Krishnan (1997). Here $\boldsymbol{\Psi} = (\boldsymbol{\xi}^T, \pi_1, \ldots, \pi_{g-1})^T$ is the vector of unknown parameters, where $\boldsymbol{\xi}$ consists of the elements of the $\boldsymbol{\theta}_i$ known a priori to be distinct.

Identifiability for mixture distributions is defined slightly different. Let

$$f(\mathbf{y}_j;\boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{y}_j;\boldsymbol{\theta}_i)$$

and

$$f(\mathbf{y}_j;\boldsymbol{\Psi}^*) = \sum_{i=1}^{g^*} \pi_i^* f_i(\mathbf{y}_j;\boldsymbol{\theta}_i^*)$$

be any two members of a parametric family of mixture densities. This class of finite mixtures is said to be identifiable for $\boldsymbol{\Psi} \in \boldsymbol{\Omega}$ if

$$f(\mathbf{y}_j; \boldsymbol{\Psi}) \equiv f(\mathbf{y}_j; \boldsymbol{\Psi}^*)$$

if and only if $g = g^*$ and we can permute the component labels so that

$$\pi_i = \pi_i^* \text{ and } f_i(\mathbf{y}_j; \boldsymbol{\theta}_i) = f_i(\mathbf{y}_j; \boldsymbol{\theta}_i^*) \quad (i = 1, \ldots, g) \qquad [3]$$

Here $\equiv$ implies equality of the densities for almost all $\mathbf{y}_j$ relative to the underlying measure on $\mathbf{R}^p$ for $f(\mathbf{y}_j; \boldsymbol{\Psi})$.

In the following, we focus on mixtures of distributions for continuous feature data. But mixtures of discrete distributions, for example, mixtures of Poisson distributions, are also commonly used in practice. Mixtures of distributions for mixed feature data consisting of both discrete and continuous variables are employed too using, for example, the location model (McLachlan and Peel, 2000: Chapter 5).

ML Estimation

The ML estimate of $\boldsymbol{\Psi}$ is obtained as an appropriate root of the likelihood equation

$$\partial \log L(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi} = 0 \qquad [4]$$

where $L(\boldsymbol{\Psi})$ denotes the likelihood function for $\boldsymbol{\Psi}$ formed from the observed random sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$. Solutions of eqn [4] corresponding to local maxima can be found by application of the EM algorithm. The latter is applied in the framework where an observation $\mathbf{y}_j$ is conceptualized to have arisen from one of the components and the indicator vector $\mathbf{z}_j$ denoting its component of origin is taken to be missing, where $z_{ij} = (\mathbf{z}_j)_i$ is defined to one or zero, according to whether $\mathbf{y}_j$ did or did not arise from the ith component of the mixture ($i = 1, \ldots, g$; $j = 1, \ldots, n$). The complete-data vector is therefore declared to be $\mathbf{y}_c = (\mathbf{y}^T, \mathbf{z}^T)^T$, where $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^T$ contains the observed data and $\mathbf{z} = (\mathbf{z}_1^T, \ldots, \mathbf{z}_n^T)^T$ contains the unobservable component-indicator variables.

The complete-data log likelihood for $\boldsymbol{\Psi}$, $\log L_c(\boldsymbol{\Psi})$, is given by

$$\log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \left\{ \log \pi_i + \log f_i(\mathbf{y}_j; \boldsymbol{\theta}_i) \right\} \qquad [5]$$

The E-step of the EM algorithm requires averaging the complete-data log likelihood $\log L_c(\boldsymbol{\Psi})$ over the conditional distribution of $\mathbf{z}$ given the observed data vector $\mathbf{y}$, using the current fit for the vector of unknown parameters $\boldsymbol{\Psi}$. As $\log L_c(\boldsymbol{\Psi})$, is linear in the unobservable data z_{ij}, the E-step (on the $(k+1)$th iteration) simply requires the calculation of the current conditional expectation of Z_{ij} given the observed data $\mathbf{y}$, where Z_{ij} is the random variable corresponding to z_{ij}. This yields the Q-function given by

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{g} \sum_{j=1}^{n} \tau_i \left(y_j; \boldsymbol{\Psi}^{(k)} \right) \left\{ \log \pi_i + \log f_i(\mathbf{y}_j; \boldsymbol{\theta}_i) \right\} \qquad [6]$$

where

$$\begin{aligned} \tau_i \left(\mathbf{y}_j; \boldsymbol{\Psi}^{(k)} \right) &= E_{\boldsymbol{\Psi}^{(k)}} \left(Z_{ij} \middle| \mathbf{y} \right) \\ &= \mathrm{pr}_{\boldsymbol{\Psi}^{(k)}} \left\{ Z_{ij} = 1 \middle| \mathbf{y}_j \right\} \\ &= \pi_i^{(k)} f_i \left(\mathbf{y}_j; \boldsymbol{\theta}_i^{(k)} \right) \Big/ f \left(\mathbf{y}_j; \boldsymbol{\Psi}^{(k)} \right) \end{aligned} \qquad [7]$$

is the posterior probability that the jth observation $\mathbf{y}_j$ belongs to the ith component ($i = 1, \ldots, g$; $j = 1, \ldots, n$). In the above, $E_{\boldsymbol{\Psi}^{(k)}}$ and $\mathrm{pr}_{\boldsymbol{\Psi}^{(k)}}$ denote expectation and probability, respectively, using the current value $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$.

The M-step on the $(k+1)$th iteration requires the global maximization of $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$ with respect to $\boldsymbol{\Psi}$ over the parameter space $\boldsymbol{\Omega}$ to give the updated estimate $\boldsymbol{\Psi}^{(k+1)}$. The updated estimate of the ith mixing proportion π_i is then given by

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_i \left(\mathbf{y}_j; \boldsymbol{\Psi}^{(k)} \right) \Big/ n \quad (i = 1, \ldots, g) \qquad [8]$$

The updated estimate of the vector $\boldsymbol{\xi}$ containing the distinct parameters in the component densities satisfies the equation

$$\sum_{i=1}^{g} \sum_{j=1}^{n} \tau_{ij} \left(y_j; \boldsymbol{\Psi}^{(k)} \right) \partial \log f_i(y_j; \theta_i) \Big/ \partial \boldsymbol{\xi} = 0 \qquad [9]$$

The E- and M-steps are alternated repeatedly until the difference

$$L\left(\boldsymbol{\Psi}^{(k+1)} \right) - L\left(\boldsymbol{\Psi}^{(k)} \right)$$

changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood values $\{L(\boldsymbol{\Psi}^{(k)})\}$. Dempster et al. (1977) showed that the (incomplete-data) likelihood function $L(\boldsymbol{\Psi})$ is not decreased after an EM iteration; that is,

$$L\left(\boldsymbol{\Psi}^{(k+1)} \right) \geq L\left(\boldsymbol{\Psi}^{(k)} \right) \qquad [10]$$

for $k = 0, 1, 2, \ldots$. Hence convergence must be obtained with a sequence of likelihood values that are bounded above.

As the likelihood eqn [4] tends to have multiple roots corresponding to local maxima, the EM algorithm needs to be started from a variety of initial values for the parameter vector $\boldsymbol{\Psi}$ or for a variety of initial partitions of the data into g groups. The latter can be obtained by randomly dividing the data into g groups corresponding to the g components of the mixture model. With random starts, the effect of the central limit theorem tends to have the component parameters initially being similar at least in large samples. One way to reduce this effect is to first select a small random subsample from the data, which is then randomly assigned to the g components. The first M-step is then performed on the basis of the subsample. The subsample has to be sufficiently large to ensure that the first M-step is able to produce a nondegenerate estimate of the parameter vector $\boldsymbol{\Psi}$.

For mixtures of normal distributions, the choice of root of the likelihood equation in the case of homoscedastic components is straightforward in the sense that the ML estimate exists as the global maximizer of the likelihood function. The situation is less straightforward in the case of heteroscedastic components as the likelihood function is unbounded. In practice, consideration has to be given to the problem of relatively

large local maxima that occur as a consequence of a fitted component having a very small (but nonzero) variance for univariate data or generalized variance (the determinant of the covariance matrix) for multivariate data. Such a component corresponds to a cluster containing a few data points either relatively close together or almost lying in a lower-dimensional subspace in the case of multivariate data. There is thus a need to monitor the relative size of the fitted mixing proportions and of the component variances for univariate observations, or of the generalized component variances for multivariate data, in an attempt to identify these spurious local maximizers.

An account of a Bayesian approach to the fitting of mixture models may be found in Mengersen et al. (2011).

Model-Based Clustering

Once the mixture model has been fitted, a probabilistic clustering of the data into g clusters can be obtained in terms of the fitted posterior probabilities of component membership for the data. An outright partitioning of the observations into g nonoverlapping clusters is effected by assigning each observation to the component to which it has the highest estimated posterior probability of belonging.

With this approach to clustering, the question of how many clusters there are in a data set can be considered in terms of the number of components g in the mixture model. The latter is to be considered on the basis of the likelihood as to be discussed in the next section.

It can be seen that the mixture likelihood-based approach to clustering is model based in that the form of each component density of an observation has to be specified in advance. Hawkins et al. (1982) commented that most writers on cluster analysis "lay more stress on algorithms and criteria in the belief that intuitively reasonable criteria should produce good results over a wide range of possible (and generally unstated) models." For example, the trace $\mathbf{W}$ criterion, where $\mathbf{W}$ is the pooled within-cluster sums of squares and products matrix, is predicated on normal groups with (equal) spherical covariance matrices; but as they pointed out, many users apply this criterion even in the face of evidence of nonspherical clusters or, equivalently, would use Euclidean distance as a metric. They strongly supported the increasing emphasis on a model-based approach to clustering. Marriott (1974) noted that the mixture likelihood-based approach "is about the only clustering technique that is entirely satisfactory from the mathematical point of view. It assumes a well-defined mathematical model, investigates it by well-established statistical techniques, and provides a test of significance for the results."

Choice of the Number of Components

With a mixture model-based approach to clustering, the question of how many clusters there are can be considered in terms of the number of components of the mixture model being used. It is sensible in practice to approach the latter question of the number of components g in a mixture model in terms of an assessment of the smallest number of components compatible with the data. A guide to the final choice of g can be obtained from monitoring the increase in the log likelihood as g is increased from a single component. Unfortunately, it is difficult to carry out formal tests at any stage of this sequential process for the need of an additional component, since, as is well known, regularity conditions fail to hold for the likelihood ratio statistic λ to have its usual asymptotic null distribution of χ^2 with degrees of freedom equal to the difference between the number of parameters under the null and alternative hypotheses (McLachlan and Peel, 2000: Chapter 6).

A formal test of the null hypothesis $H_0{:}g = g_0$ versus the alternative $H_1{:}g = g_1 (g_1 > g_0)$ can be undertaken using a resampling method, as described in McLachlan (1987). A commonly used method of testing the above hypotheses is to adopt the BIC criterion of Schwarz (1978), which, when applied in the present context, leads to H_0 being rejected if twice the increase in the log likelihood (that is, $-2\log\lambda$) has to be greater than $d\log n$, where d denotes the difference between the number of parameters under the two hypotheses. Some theoretical justification for the use of BIC in this role has been given by Keribin (2000). Further information-based criteria for tests on the number of components in a mixture model have been considered (McLachlan and Peel, 2000: Chapter 6), including the integrated classification likelihood (ICL) criterion, as proposed by Biernacki et al. (2000).

Mixtures of Normal and *t*-Distributions

For multivariate data of a continuous nature, attention has concentrated on the use of multivariate normal components because of their computational convenience. Under this assumption, the component densities $f_i(\mathbf{y}_j;\boldsymbol{\theta}_i)$ for $\mathbf{y}_j$ are given by

$$f_i(\mathbf{y}_j;\boldsymbol{\theta}_i) = \phi(\mathbf{y}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) \qquad [11]$$

and $\phi(\mathbf{y}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ denotes the p-variate normal density function with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ $(i = 1,\ldots,g)$. For normal component densities, the updates of the estimates of the component means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$ are given in closed form.

One attractive feature of adopting mixture models with elliptically symmetric components such as the normal or t-densities is that the implied clustering is invariant under affine transformations of the data; that is, invariant under transformations of the feature vector.

For many applied problems, the tails of the normal distribution are often shorter than appropriate. Also, the estimates of the component means and covariance matrices can be affected by observations that are atypical of the components in the normal mixture model being fitted. McLachlan and Peel (1998) have considered the fitting of mixtures of (multivariate) t-distributions. The t-distribution provides a longer-tailed alternative to the normal distribution. Hence it provides a more robust approach to the fitting of normal mixture models, as observations that are atypical of a normal component are given reduced weight in the calculation of its parameters.

The t density with location parameter $\boldsymbol{\mu}_i$, positive-definite matrix $\boldsymbol{\Sigma}_i$, and ν_i degrees of freedom is given by

$$\phi_t(\mathbf{y}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,\nu_i)=\frac{\Gamma(\nu_i+p/2)|\boldsymbol{\Sigma}_i|^{-1/2}}{(\pi\nu_i)^{(1/2)p}\Gamma(\nu_i/2)\left\{1+\delta\left(\mathbf{y}_j,\boldsymbol{\mu}_i;\boldsymbol{\Sigma}_i\right)\Big/\nu_i\right\}^{(1/2)(\nu_i+p)}} \quad [12]$$

where $\delta(\mathbf{y}_j,\boldsymbol{\mu}_i;\boldsymbol{\Sigma}_i)=(\mathbf{y}_j-\boldsymbol{\mu}_i)^T\sum_i^{-1}(\mathbf{y}_j-\boldsymbol{\mu}_i)$ denotes the Mahalanobis squared distance between $\mathbf{y}_j$ and $\boldsymbol{\mu}_i$ (with $\boldsymbol{\Sigma}_i$ as the covariance matrix). If $\nu_i>1$, $\boldsymbol{\mu}_i$ is the mean of $\mathbf{Y}_j$, and if $\nu_i>2$, $\nu_i(\nu_i-2)^{-1}\boldsymbol{\Sigma}_i$ is its covariance matrix. As ν_i tends to infinity, $\mathbf{Y}_j$ becomes marginally multivariate normal with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. Hence this parameter ν_i may be viewed as a robustness tuning parameter. It can be fixed in advance or it can be inferred from the data for each component, thereby providing an *adaptive* robust procedure.

Mixtures of Factor Analyzers

For the parametric mixture model [2], it is important that the dimension of $\boldsymbol{\Psi}$ is not large relative to the sample size n. The g-component normal mixture model with unrestricted component-covariance matrices is a highly parameterized model with $\frac{1}{2}p(p+1)$ parameters for each component-covariance matrix $\sum_i(i=1,\ \ldots,\ g)$. Banfield and Raftery (1993) introduced a parameterization of the component-covariance matrix $\boldsymbol{\Sigma}_i$ based on a variant of the standard spectral decomposition of $\boldsymbol{\Sigma}_i$.

A common approach to reducing the number of dimensions is to perform a principal component analysis (PCA) and then to perform the cluster analysis on the basis of the first few leading principal components. But as is well known, projections of the feature data $\mathbf{y}_j$ onto the first few principal axes are not always useful in portraying the group structure; see McLachlan and Peel (2000, p. 239) for an illustrative example of this.

Another approach for reducing the number of unknown parameters in the forms for the component-covariance matrices $\boldsymbol{\Sigma}_i$ is to adopt the mixtures of factor analyzers model as considered in McLachlan and Peel (2000: Chapter 8). With the mixture of factor analyzers model, the ith component-covariance matrix $\boldsymbol{\Sigma}_i$ has the form

$$\boldsymbol{\Sigma}_i=\mathbf{B}_i\mathbf{B}_i^T+\mathbf{D}_i \quad (i=1,\ldots,g) \quad [13]$$

where $\mathbf{B}_i$ is a $p\times q$ matrix of factor loadings and $\mathbf{D}_i$ is a diagonal matrix. It assumes that the component correlations between the observations can be explained by the conditional linear dependence of the latter on q latent or unobservable variables specific to the given component. Unlike the PCA model, the factor analysis model [13] enjoys a powerful invariance property: changes in the scales of the feature variables in $\mathbf{y}_j$ appear only as scale changes in the appropriate rows of the matrix $\mathbf{B}_i$ of factor loadings. If the number of factors q is chosen sufficiently smaller than p, the representation [13] imposes some constraints on the component-covariance matrix $\boldsymbol{\Sigma}_i$ and thus reduces the number of free parameters to be estimated.

The mixtures of factor analyzers model can be fitted by using the alternating expectation–conditional maximization (AECM) algorithm (Meng and van Dyk, 1997). In practice, consideration has to be given to the number of components g and the number of factors q in the mixture of factor analyzers model. One obvious approach is to use the Bayesian Information Criterion (BIC) of Schwarz (1978). An alternative approach is to use the likelihood ratio statistic for tests on g and q. For tests on g, it is well known that regularity conditions do not hold for the usual χ^2 approximation to the asymptotic null distribution of the likelihood ratio test statistic to be valid (McLachlan and Peel, 2000: Chapter 6). This is also the case for tests on q at a given level of g; see Drton (2009).

The mixture of factor analyzers model is sensitive to outliers since it uses normal errors and factors. McLachlan et al. (2007) have considered the use of mixtures of t-analyzers in an attempt to make the model less sensitive to outliers.

Even with the MFA approach, the number of parameters still might not be manageable, particularly if the number of dimensions p is large and/or the number of components (clusters) g is not small (McLachlan et al., 2002). Baek et al. (2010) considered how this factor-analytic approach can be modified to provide a greater reduction in the number of parameters. They termed their approach mixtures of common factor analyzers (MCFA). This is because the matrix of factor loadings is common to the components before the component-specific rotation of the component factors to make them white noise. Recently, Baek and McLachlan (2011) have extended the approach to mixtures of common t-factor analyzers.

In the above, it has been assumed implicitly that there is a one-to-one correspondence between the mixture components and the clusters. With some data, for example, in flow cytometry, the clusters are asymmetric and so each cluster is unable to be modeled adequately by a single component with an elliptically symmetric distribution. One way to handle this situation is to fit mixtures of normal or t-component distributions that are greater in number than that of the clusters; that is, the distribution of the data within a cluster is modeled by a mixture of normal or t-distributions.

Another approach is to fit a mixture model in which each component distribution is chosen to model the data in the corresponding cluster by adopting a mixture model with, for example, skew normal or skew t-component distributions; see Lee and McLachlan (2014) and the references therein.

The reader is referred to the lists in the appendix in McLachlan and Peel (2000) and at http://www.maths.uq.edu.au/~gjm/mix_soft_other/emmix.html for the availability of software for the fitting of mixtures of normal, t, skew normal, skew t, and factor distributions, which includes the EMMIX suite of programs developed initially by McLachlan et al. (1999).

See also: Bayesian Statistics; Computation: Expectation-Maximization Algorithm; Factor Analysis and Latent Structure Analysis: Overview; Mixture Models in Statistics; Statistical Clustering.

Bibliography

Baek, J., McLachlan, G.J., 2011. Mixtures of common t-factor analyzers for clustering high-dimensional microarray data. Bioinformatics 27, 1269–1276.

Baek, J., McLachlan, G.J., Flack, L.K., 2010. Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualization of high-dimensional data. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1298–1309.

Banfield, J.D., Raftery, A.E., 1993. Model-based Gaussian and non-Gaussian clustering. Biometrics 49, 803–821.

Biernacki, C., Celeux, G., Govaert, G., 2000. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 719–725.

Böhning, D., 1999. Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others. Chapman & Hall/CRC, New York.

Day, N.E., 1969. Estimating the components of a mixture of two normal distributions. Biometrika 56, 463–474.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B 39, 1–38.

Drton, M., 2009. Likelihood ratio tests and singularities. The Annals of Statistics 37, 979–1012.

Everitt, B.S., Hand, D.J., 1981. Finite Mixture Distributions. Chapman & Hall, London.

Frühwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer, New York.

Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97, 611–631.

Hawkins, D.M., Muller, M.W., ten Krooden, J.A., 1982. Cluster analysis. In: Hawkins, D.M. (Ed.), Topics in Applied Multivariate Analysis. Cambridge University Press, Cambridge, pp. 303–356.

Keribin, C., 2000. Consistent estimation of the order of mixture models. Sankhyā: The Indian Journal of Statistics, Series A 62, 49–66.

Lee, S., McLachlan, G.J., 2014. Finite mixtures of multivariate skew *t*-distributions: some recent and new results. Statistics and Computing 24, 181–202.

Lindsay, B.G., 1995. Mixture Models: Theory, Geometry and Applications. Institute of Mathematical Statistics and the American Statistical Association, Alexandria, VA.

Marriott, F.H.C., 1974. The Interpretation of Multiple Observations. Academic Press, London.

McLachlan, G.J., 1987. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Applied Statistics 36, 318–324.

McLachlan, G.J., Basford, K.E., 1988. Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York.

McLachlan, G.J., Krishnan, T., 1997. The EM Algorithm and Extensions, second ed. 2008. Wiley, New York.

McLachlan, G.J., Peel, D., 1998. Robust cluster analysis via mixtures of multivariate t-distributions. In: Amin, A., Dori, D., Pudil, P., Freeman, H. (Eds.), Lecture Notes in Computer Science, vol. 1451. Springer-Verlag, Berlin, pp. 658–666.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.

McLachlan, G.J., Peel, D., Basford, K.E., Adams, P., 1999. The EMMIX software for the fitting of mixtures of normal and *t*-components. Journal of Statistical Software 4 (2), 1–4.

Meng, X.L., van Dyk, D., 1997. The EM algorithm – an old folk song sung to a fast new tune (with discussion). Journal of the Royal Statistical Society B 59, 511–567.

Mengersen, K.L., Robert, C.P., Titterington, D.M., 2011. Mixtures: Estimation and Applications. Wiley, Hoboken, New Jersey.

McLachlan, G.J., Bean, R.W., Ben-Tovim Jones, L., 2007. Extension of the mixture of factor analyzers model to incorporate the multivariate *t*-distribution. Computational Statistics & Data Analysis 51, 5327–5338.

McLachlan, G.J., Bean, R.W., Peel, D., 2002. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics 18, 413–422.

Pearson, K., 1894. Contributions to the theory of mathematical evolution. Philosophical Transactions of the Royal Society of London 185, 71–110.

Schlattmann, P., 2009. Medical Applications of Finite Mixture Models. Springer, Berlin.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Titterington, D.M., Smith, A.F.M., Makov, U.E., 1985. Statistical Analysis of Finite Mixture Distributions. Wiley, New York.

Wolfe, J.H., 1970. Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research 5, 329–350.

Mobile Communication

Rich Ling, Nanyang Technological University, Singapore

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article is a review of mobile communication from a social perspective. It sets out a discussion of the definition of mobile communication considering different ways to consider this phenomenon. It is followed by an analysis of current and emerging directions in mobile communication research. This includes an examination of how the mobile phone allows for individual addressability and the interlacing of activities. These characteristics of the mobile phone mean that it can be used for nuanced microcoordination. Mobile communication has been adopted by different groups in different ways. This article examines three of these groups, namely, business use, teens, and among women in developing countries. The article considers the development of location-based services, the ways in which the mobile phone develops social cohesion, and how it is becoming a structural part of social interaction. Finally, the article considers several methodological issues associated with the mobile phone.

Introduction

Mobile communication describes the use of various technological systems in order to communicate while one is away from a fixed location. Mobile communication has been the focus of scholarly work since around 2000 (Green and Haddon, 2009). Researchers in Hong Kong (Leung and Wei, 1999), Finland (Kasesniemi and Rautiainen, 2002; Oksman and Rautianen, 2003), the United Kingdom (Green and Smith, 2004; Grinter and Eldridge, 2001; Haddon, 1997), Japan (Ito, 2001), and Norway (Ling, 2004) as well as many other locations have made contributions. With the development of the mobile Internet, research is starting to bridge the work of the traditional mobile communication community and the Internet research community. This article will examine the definition of mobile communication, the current and emerging issues in mobile communication research, and some methodological issues poised by the study of this area.

What Is Mobile Communication (and What Has It Been)?

A broad definition of mobile communication is the use of a portable communication device to interact with others and eventually with information sources. We likely think of the mobile phone as the most common mobile communication device, but an expanded definition might include the use of semaphore flags, signal lamps, and other portable communication devices. Lars Magnus Ericsson, the founder of the telephone equipment manufacturer Ericsson, mounted a telephone terminal in his car. Using wires to patch into nearby telephone lines he was able to make calls while away from the landline telephone. Following this line of thought, one can think of telephone booths as affording us some type of communication while we are away from our home base (Ling and Donner, 2009). Obviously the phone booth and Ericsson's jury rigged device are only quasi-mobile.

Radio-based wireless devices have allowed for mobile communication for over a century. The work of Marconi gave us the ability to broadcast messages across great distances. Marconi's wireless device was used to communicate long distances using the Morse telegraph code in the late 1800s. As early as 1903, he was sending trans-Atlantic messages. These became mobile when ships installed wireless devices onboard. In the wake of the Titanic disaster in 1912, various legislative bodies enacted laws proscribing the use of radio equipment on passenger ships. In addition, there were moves to regulate the use of radio frequencies.

During the early part of the twentieth century, radio communication moved from transmitting/receiving Morse code to voice. Nonetheless, mobile radio communication was based on broadcasting of a single message to many receivers. By the early 1920s, this developed into commercial radio (Douglas, 1987), and there was also use of radio-based systems for communication by public authorities such as the Detroit Police (Agar, 2003; Farley, 2009; Wikle, 1930). It was also used relatively early by organizations where there was the need to manage a fleet of vehicles such as taxis, delivery trucks, as well as operations in the field such as the military. (In some cases, this was quasi-mobile since it was limited to one-way radio. Upon receiving a message, the driver would have to find a phone booth in order to return the communication.) At this point, radio communication and switched telephony (i.e., the system of connecting two parties through a switching device) lived in two separate worlds. After the Second World War, there was some work done investigating the ability to integrate radio into the switched telephone network (Ling, 2004). The first approach was to use a broadcast system of mobile communication for 'car phones.' This relied on a single antenna that covered a large area. This, however, did not make efficient use of the limited radio capacity. In 1947, Douglas H. Ring and W. Rae Young developed a system that divided the landscape into a mesh of cells that allowed for more efficient use of the limited radio frequencies along with a system with which to hand off the calls between the cells. This system of tracking phones as they moved from one cell to another was first used in a trial developed by AT&T in 1969 on the Amtrak Metroliner train between Washington, DC, and New York in 1969. Eight years later, a system with 10 cells and the capacity for 2000 users was established in Chicago by Motorola. These developments move us closer to what we now call mobile telephony. The cellular

systems allowed for the development of small handsets with which we could call one another whenever we were connected to the system.

The first mobile phones allowed for voice interaction. With time, however, we have added other functionality such as onboard calendars, name registers, and games. In the last few years, this has increasingly included access to a broad variety of online information. One of the first additional communication functions was short message service or texting (Hillebrand et al., 2010; Trosby, n.d., 2004). This function 'pushed' messages to addressees and allowed for asynchronous text-based interaction. Experience shows that there is a niche for this type of interaction. In 2010, there were over 6 trillion text messages sent worldwide.

Mobile phones have developed into sophisticated computing devices that give us access to a wide variety of information sources and ways of interacting with others. Starting with the Japanese iMode system (Lindmark, 2002), the development of the open operative systems that allows us to download a seemingly endless number of 'apps' has led flourishing of smartphones (West and Mace, 2010). In this guise, the mobile phone, along with the Wi-Fi made the mobile phone into a portable Internet device. These devices, along with the larger tablets and other Wi-Fi-enabled devices, allow us to communicate with one another and to search data archives whenever and wherever we might be. They push the definition of communication beyond interaction with other individuals and include interaction with other devices and information sources.

Current and Emerging Directions in Theory and Research

As with the study of many sociotechnical systems, the study of mobile communication's social consequences has seen several deferent phases. The earliest work was often more descriptive and examined adoption by particular groups (Kasesniemi and Rautiainen, 2002; Mante-Meijer and van de Loo, 1998; Rakow and Navarro, 1993). There were also discussions with regards to potential effects of the technology on society (Haddon, 1997; Lange, 1993). Eventually, mobile communication became rather common in the developed world and started to diffuse in developing countries. The growth of adoption in Africa, Asia, and South America led to new analyses. According to his broad investigation of this scholarship, Donner suggests that the early articles examining adoption in developing countries often focused on the determinants of adoption, the impact of mobile use, and/or their role in economic development (2008).

More recently, there has been the development of theoretical insights that again can help in the framing of future analyses (Castells et al., 2007; Geser, 2004; Katz and Aakhus, 2002; Ling, 2008, 2012; Wellman, 2002). Many of these approaches were developed with the assumption that mobile communication was an untethered version of landline telephony with perhaps the addition of texting. The focus is on the effect of mediated person-to-person communication on the lives of individuals. To be sure, this is an important dimension of mobile communication since it has led to many of the issues discussed below. The growth of the mobile Internet and smartphones has added a new wrinkle to the analysis that has yet to be theorized. There are new forms for interpersonal interaction (social networking and interactive games) and the devices can be used to gather information on an 'as needed' basis (e.g., maps for location-based services or medical information in developing countries).

The major questions associated with the study of mobile phones in the social sciences have been associated with the way that they adjust and change social interaction. In some cases, they have facilitated information flow and in other cases they have disrupted the previous communication channels. Research regarding mobile communication has examined how it has

- provided us with individual addressability and interlacing of activity,
- changed the way that we organize social interaction,
- contributed to the development of special cultural formations in different groups,
- developed location-based applications,
- how it has encouraged social cohesion, and
- how it has become increasingly embedded in society.

All the while, one gets the impression that mobile communication research is like chasing a hen with a bushel basket. Researchers keep trying to capture it while it darts and jibes in ever new directions. When we feel that we have captured the essence of mobile communication, there is a new development that gives users new functionality, e.g., the growth of Internet protocol-based texting such as WhatsApp and group texting in Facebook. The users then take these and exploit them in an unexpected direction.

Individual Addressability and Interlaced Activity

Mobile communication has progressively led to the ability to engage in individualized communication since the equipment with which we can communicate has become ever more portable. The first radios required huge steam engines to power them and were only mobile when mounted in, for example, steam ships. Of course the landline system of telephony was available, but it was bound to specific locations. Improvements in technology including, among other things, transistors and the development of the cellular system mean that we have far more sophisticated devices that neatly fit into our pockets or get lost under the stack of newspapers on the kitchen counter. The consequence of these small portable devices is that we are individually addressable. We can call and be called regardless of where we might be. This has, in turn, meant that we can interlace communication with other activities (Ling and Donner, 2009). It also means that we include mediated communication in activities where before we had been incommunicado. On the one hand this leads to an efficiency in planning and the ability to engage in impromptu social interaction with our closest sphere in the folds of daily life (Licoppe, 2004). It also means that we may present a danger to others when, for example, driving (Strayer et al., 2006) and we can potentially drive other copresent individuals to tears with our chatter (Monk et al., 2004).

Microcoordination

Perhaps more than anything else the mobile phone is a coordination tool. Previous to the widespread diffusion of the mobile phone, remote communication took place via the landline telephone. This meant that we called to specific locations on the chance that our eventual interlocutor was nearby. In the case of house-based landline telephones, if we were not able to talk with one person, we might as easily have a conversation with another. This system was useful for coordination, but it did not allow for changes in plans or new arrangements once we – or our eventual conversation partner – were in transit.

The individual addressability of the mobile telephone changed this (Ling and Yttri, 2002). Our newfound ability to communicate while on the run means that we use it to coordinate interactions in at an ever more nuanced level. The ability to make changes in plans or indeed to form plans when away from the landline telephone has changed the way that we coordinate interaction. Many people have only a loose idea that they will, for example, meet friends for a social visit on a particular day. As day arrives and as the eventual meeting partners confirm their interest to meet (perhaps through a series of text messages) the plans begin to take concrete shape. These might include a series of interactions mapping out suggestions (i.e., which café to visit) and negotiating the time. Even as the individuals are en route to the meeting, they might make last minute adjustments (i.e., the first person to the café finds it too crowded and thus suggests a 'plan B'). The individual addressability provided by the mobile phone allows for the ongoing negotiation of meetings and also the discussion of family logistics. One partner who is late from a meeting might call or text to the other who can then shuffle the order in which they pick up the kids at school or go shopping.

This type of give-and-take in the planning process works in the context of small groups since the individuals can interact and form plans and the situation clarifies itself. It is clear, however, that this type of planning is not suited to the functioning of large organizations. In that case, time-based coordination retains its place. Interestingly, however, use of the mobile Internet via smartphones allows the development of applications where people can access timetables for these organizations. Apps that give the real-time bus schedules for the nearest bus stop or the departure time and gate number for a particular flight give us individualized access to this information and facilitate the coordination through the diffusion of information.

Mobile Communication among Particular Groups

Mobile communication has been variously adopted by different groups in a variety of locations. This variation in adoption has itself been a topic of research. Some have examined how it varies from place to place (Campbell and Russo, 2003; Lasén, 2004) while others have noted that it unifies behavior (Katz and Aakhus, 2002). There are several groups that have had particular adoption patterns.

Business and Formal Organizations

Radio communication has a long history in commercial and formal organizations. As noted above, broadcast radios were used by organizations that needed to organize a large number of vehicles with a constant flow of shorter tasks such as local delivery companies, taxis, or large animal veterinarians who need to visit a variety of locations (Ling, 2004). As with the police and fire, these organizations used a central dispatching system via broadcast radio.

With the coming of the mobile telephone, the broadcast radio system was to some degree displaced (Manning, 1996; Townsend, 2000). Another group who quickly adopted the mobile phone was sales people who often needed to work out complex meeting schedules in a variety of locations (Ling, 2012). The device facilitated coordination and it allowed employees who were out in the field to, for example, call their warehouse to check on the availability of goods. The mobile phone also was adopted by construction workers in order to facilitate the sharing of equipment and interaction with, for example, architects when the plans do not fit the situation at the building site. With time, the mobile phone has also become a way for employers to contact substitutes when, for example, a shift worker calls in sick. In other cases, organizations that have adopted open office solutions with 'hot desking' may use a variety of different work stations from day to day or even hour to hour. In order to facilitate this, they equip all employees with mobile phones (Julsrud, 2005).

The effect of mobile communication on commerce can also be seen in developing countries. In these cases, the coordination might have to do, for example, with the local production of handcrafts (Jagun et al., 2008), everyday logistics (Sey, 2011), or facilitating the flow of market information (Franklin et al., 2000; Jensen, 2007; Platteau and Abraham, 1987; Rahman, 2007). In developing countries, entrepreneurs and customers might also work out special uses of the mobile phone to facilitate commerce. One approach is to use 'beeps,' 'flashes,' or missed calls in order to signal potential work. In this system, a person who, for example, needs the help of a plumber or carpenter or needs to summon a taxi will call the number, let it ring once or twice and then hang up. Because of caller identity, the handworker or driver has the number of the potential client. If they are available for the work, they can call back and make the specific arrangements (Donner, 2007; Geirbo et al., 2007).

Teen Use of Mobile Communication

Another specific group that has adopted the mobile phone is teens. The device fits particularly well into their lives and indeed this group was one of the first to adopt it for largely social purposes. With the development of prepaid subscriptions that set boundaries on consumption, teens were soon among the most enthusiastic users of mobile telephony (Johnsen, 2003). The analysis of teens' use of mobile communication has described its contribution to a sense of teen culture (Goggin and Crawford, 2011; Ito et al., 2005). Following a slightly different direction, there has also been analysis of how mobile communication ebbs and flows as teens move through the different phases of adolescence (Helles, 2009; Ling, 2010). There has also been analysis of how the mobile phone plays into the sense of teen self-image (Skog, 2002) and how texting becomes a way for teens to exploit the plasticity of the language (Baron and Campbell, 2010; Baron and Ling, 2011; Ling et al., 2010). (This is also something

that is possible to see in texting in developing countries (Vold Lexander, 2011).)

By 2011, teens in the United States were among the most enthusiastic users of texting. Analysis by Lenhart shows that teens aged 14–17 years sent a median of 60 texts per day (2012). The fact that teen girls in the United States send a median of 100 texts per day shows that they are perhaps more committed to texting than are boys. Regardless of gender, texting is one of the most important communication channels for teens. According to Lenhart, there are 63% of teens who report using it daily to stay in touch with friends. Only 39% call using the mobile phone, 35% have face-to-face talks (outside school) and 35% use social networking sites. The data show that only 19% use the landline phone and 6% e-mail one another. Texting and talking are shown to be complementary, not competitive forms of communications since the heaviest texters also talked the most.

Women in Developing Countries

Women in developing countries have also seen changes in their situation as a result of the mobile phone. For some it has been a source of income and for others it has facilitated their daily lives by providing a way to coordinate time-consuming tasks. For the women who have had to find work far from home, it has given them a way to stay in contact with their families and the mobile phone has become an element in their interactions with men.

Perhaps one of the first ways we saw the effect of the mobile phone on women in developing countries was in the case of the Grameen Phone 'Village Phone' program. Grameen Phone provided women in rural Bangladesh with microloans to purchase mobile phones and a solar battery charging system that became a privately owned, publically available telephone link for their village. The 'telephone ladies' paid back their loans and eventually derived a profit by charging other villagers for calls (Aminuzzaman et al., 2003; Richardson et al., 2000). In 2003, it was estimated that there are more than 21 000 women using this system in Bangladesh (Yunus, 2003). Since that time, handsets have become less expensive and more people have the ability to own their own handset.

The mobile phone provides women in developing countries with a communication channel through which they can experience not only increased empowerment (Lee and Jayachandran, 2009) but also new forms of control. Women who move away from their home also use the mobile phone for a variety of both social and instrumental tasks. Madianou and Miller, for example, describe how the mobile phone allows Filipina women working in London to maintain contact with their children who live in the Philippines (2011). The separation from their children is bridged, to some degree, by using the phone to call and to maintain contact. Another example of this is seen with women who are 'internal migrants' in China. They leave their village of origin to work in the larger cities. Following the work of Wallis, the mobile phone becomes an important channel for these women to maintain contact with their families and to organize their social lives in the cities. However, the mobile phone is also a conduit through which employers can reach into their private lives and surreptitiously check up on them (2011). Thus, there is both empowerment and disempowerment packaged into the mobile phone for these women.

Another way that mobile telephony is coming into the lives of women (and for that matter men) in developing countries is through the growth of mobile-based health initiatives. Mobile-based medicine (or m-health) allows local health care givers, who are in many cases women (Chib and Hsueh-Hua Chen, 2011), to get advice from more central health care providers and it also can streamline reporting in health information systems. Rather than relying on cumbersome paper-based systems that needs manual aggregation, mobile handsets allow the development of broad-scale health reporting systems (Asangansi and Braa, 2010).

Beyond these more instrumental uses, the mobile telephony is also used in the personal interactions of women and men. As in developed countries, it is used to facilitate romantic interactions. As we might expect, its use can also lead to misunderstandings (Archambault, 2011). In some cases, the mobile phone can undermine traditional forms of matchmaking (Ellwood-Clayton, 2005). Also, since it can be used in what is seen as a subversion of long-established authority roles in these affairs, it can be a particularly contentious object that needs to be hidden by the teen girls when they receive it as a surreptitious gift from a boy (Hijazi-Omari and Ribak, 2008). Finally, in a less savory combination of romantic relationships and mobile phone-aided entrepreneurship, Horst and Miller document how impoverished women in Jamaica use the phone in order to maintain an overview of potential sexual partners who would also then contribute to their household economy in a type of informal prostitution (2005). Thus, in a surprising variety of ways, mobile communication has been adopted by women in developing countries. In many cases, it empowers them but it is also used to maintain existing exploitive power relationships.

Location-Based Services

Since their inception, mobile phones have included a location-based dimension in their use. Early users could remind interlocutors about their location and gather information on where, for example, a meeting was or where they might find a particular location (Elliott and Urry, 2010). More recently, the integration of location determination functionality in mobile phones (i.e., use of the Global Positioning System) along with mapping software has led to the development of a variety of applications or apps. These facilitate finding, for example, stores or services (i.e., taxi stands or bus stops). In addition, there are apps that help tourists find hotels, restaurants, and sights along with others' evaluations.

Another use area is location-based social networking (Gordon and Silva, 2011) that allow people to follow the movement of friends or to compare 'check-ins.' There can be gaming aspects to this type of mobile application (i.e., Foursquare) or they can simply be a way for us to follow one another as we move through the day (Google latitude). There are also various types of location-based games (i.e., BotFighters) and hobbies (geocaching). This type of location based-service is perhaps the most commonly discussed in the media. It is also interesting to note that use of these services is relatively low. According to Lenhart, their use in 2010 was below 10% of teens (Lenhart, 2012). Other forms of location-based activities include attaching information to places.

Finally, a more instrumental version of location-based mobile communication is the tracking of objects. This can include, for example, tracking railroad cars, freight trucks, or shipping containers as they move toward their destination. At a smaller scale, it can include tracking the progress of particular products or parts in a 'just in time' assembly line. This type of application moves away from interpersonal interaction and focuses more on facilitating logistics in complex organizations. (This can include using the mobile phone system, the GPS system, or some combination of the two.)

Social Cohesion

There has been much discussion asking whether Information and Communications Technologies (ICTs), including mobile phones, have reduced the size of social networks and led to social isolation (McPherson et al., 2006; Putnam, 2000). To this point, research on mobile communication has showed that, in many cases, it has a positive effect on social cohesion. This includes work on families and friendship groups in Scandinavia (Ling, 2008; Ling and Stald, 2010), work by Kim et al. in Korea (2006), Matsuda in Japan (2005: p. 127), and Smoreda and Thomas in Europe (2001). Matsuda describes what she calls the 'full time intimate sphere' (2005: p. 133), Habuchi describes the tele-cocoon (2005: p. 167), and Licoppe gives us the idea of connected presence (2004). All of these concepts suggest that the device is a tool of the intimate sphere. Castells et al. (2007), Miyata (2006), Igarashi et al. (2005), Ishii (2006), Reid and Reid (2005), and Hampton each discuss the way in which social cohesion is developed and maintained via the mobile phone. Hampton et al. show that mobile phone use increases the size of an individual's core social network (2011). That being said, this is still a limited number of people. In a study of machine-generated calling logs containing approximately 65 million calls and texts in Norway, Ling, Bertel, and Sundsøy find that half of all text and calls commonly go to less than five different other numbers (2012). In sum, it is fair to say that the mobile phone helps to maintain social cohesion in the intimate sphere.

Structural Embedding of Mobile Communication in Society

Much of the discussion of mobile communication is framed in the affordances, the benefits, and the costs at the personal level. Another approach is to look at the social effect of mobile communication. When looking more broadly at the social level, it is clear that the mobile telephony provides social utility for the group. It facilitates coordination, helps maintain cohesion, and is a safety channel in the case of emergencies. In order to function in this way, however, it assumes that the individual has a functioning device.

As with, for example, fax machines in their day or office calendar programs, adoption of the mobile phone by a group of people facilitates the functioning of the group. The fax (that was subsequently replaced by e-mail and scanners), the electronic calendar, and the mobile phone provide a common platform for different types of information sharing. These systems, however, do not come without a cost to the individual. We have to take the time to learn how to use the system and, in some ways, we have to surrender our individual prerogative to participation with others in use of the system. Not having the equipment or not using the program places us outside the group. In the case of the mobile phone, there is also the cost and the maintenance of the handset just as there is the cost of the subscription. The worldwide adoption of mobile telephony shows that we are willing to incur these costs in order to be linked into our social networks.

If one does not have a mobile phone; if it is forgotten at home, turned off, or not charged then the individual becomes a broken link in the social network. In this case, we have, in a small way, failed our friends and family who rely on being able to contact us. To the degree that we can no longer make agreements or coordinate interaction we have failed our social obligations. The mobile phone has become, to some degree, a structural part of society.

Methodological Issues

Studying mobile communication has relied on the traditional approaches to examining social phenomena. There have been many qualitative analyses including observation studies, focus groups, and person-to-person interviews. In addition, there have been a wide variety of surveys. These analyses have given us insight into how people use their mobile phones, the differences between cultures in use and the ways that people adapt and change the technology to better fit with their social context. They have given us insight into how quickly the devices have spread and also an understanding of the social consequences of mobile phone use.

These research tools have their role, and they also have their weaknesses. Qualitative analyses give us fine-grained insight into specific situations. However, the results are rarely generalizable. We may glean deep awareness of how a particular group uses mobile phones, but since the sampling is limited and not randomized, we need to be cautious with regards applying the insights into other situations. By contrast, surveys can give generalizable insight, but they too have their limitations, particularly in the case of quotidian activities such as mobile communication. It is often difficult for people to remember, for example, how many calls they made during the day. Indeed, comparisons of survey material with call data records show only moderate correlation (Boase and Ling, 2011). This might argue for using machine logs in order to measure behavior. The weakness of this approach is that we do not necessarily know the attitudes or demographics of the users. Thus, all the methods to study mobile communication have their potentials and their limitations. Indeed mobile communication presents public opinion researchers with a new set of difficulties (Groves, 2011).

Another potential, however, is to use mobile phones as data collection devices. On-board logging of activities is one approach. Another is to use the individual addressability of the mobile phone to carry out real-time research. Kubey, Larson, and Csikszentmihaly have suggested using pagers in their experience sampling method (1996). Since then, mobile phones have been examined in relation to their use as probes to measure real-time automobile traffic (Rose, 2006), health-related behavior (Hicks et al., 2010), and interviewing in situ for auditing decision making (Baxter and Hunton, 2011).

See also: Communication and Transportation, History of; Information Society; New Media and Social Capital; Radio as a Medium; Telephone.

Bibliography

Agar, J., 2003. Constant Touch. A Global History of the Mobile Phone. Icon Books, Cambridge, UK.

Aminuzzaman, S., Baldersheim, H., Jamil, I., 2003. Talking back! Empowerment and mobile phones in rural Bangladesh: a study of the Village Phone Scheme of Grameen Bank. Contemporary South Asia 12 (3), 327–348.

Archambault, J.S., 2011. Breaking up "because of the phone" and the transformative potential of information in Southern Mozambique. New Media & Society 13, 444–456.

Asangansi, I., Braa, K., 2010. The emergence of mobile-supported national health information systems in developing countries. Studies in Health Technology and Informatics 160 (Pt 1), 540.

Baron, N., Campbell, E., 2010. Talking Takes Too Long: Gender and Cultural Patterns in Mobile Telephony. Association of Internet Researchers, Göteborg, Sweden.

Baron, N., Ling, R., 2011. Necessary smileys and useless periods: redefining punctuation in electronically-mediated communication. Visible Language 45 (1/2), 45–67.

Baxter, R.J., Hunton, J.E., 2011. Capturing affect via the experience sampling method: prospects for accounting information systems researchers. International Journal of Accounting Information Systems 12 (2), 90–98. http://dx.doi.org/10.1016/j.accinf.2010.12.002.

Boase, J., Ling, R., 2011. Measuring Mobile Phone Use: Self-report versus Log Data. International Communication Association, Boston, MA.

Campbell, S.W., Russo, T.C., 2003. The social construction of mobile telephony: an application of the social influence model to perceptions and uses of mobile phones within personal communication networks. Communication Monographs 70 (4), 317–334.

Castells, M., Fernandez-Ardevol, M., Qiu, J., Sey, A., 2007. Mobile Communication and Society: A Global Perspective. MIT, Cambridge, MA.

Chib, A., Hsueh-Hua Chen, V., 2011. Midwives with mobiles: a dialectical perspective on gender arising from technology introduction in rural Indonesia. New Media & Society 13.

Donner, J., 2007. The rules of beeping: exchanging messages using missed calls on mobile phones in sub-Saharan Africa. Journal of Computer Mediated Communication 13 (1), 1–22.

Donner, J., 2008. Research approaches to mobile use in the developing world: a review of the literature. The Information Society 24 (3), 140–159.

Douglas, S., 1987. Inventing American Broadcasting 1899–1922. Johns Hopkins University Press, Baltimore.

Elliott, A., Urry, J., 2010. Mobile Lives. Routledge, London.

Ellwood-Clayton, B., 2005. Desire and loathing in the cyber Philippines. In: Harper, R., Taylor, A., Palen, L. (Eds.), The Inside Text: Social Perspectives on SMS in the Mobile Age. Klewer, London, pp. 195–222.

Farley, T., 9 May 2009. Privateline.com: Telephone History. Retrieved from: http://www.privateline.com/TelephoneHistory3A/numbers.html.

Franklin, M.E., Abramowit, J.S., Kozak, M.J., Levitt, J.T., Foa, E.B., 2000. Effectiveness of exposure and ritual prevention for obsessive-compulsive disorder randomized compared with nonrandomized samples. Journal of Consulting and Clinical Psychology 68 (4), 594–602.

Geirbo, H.C., Helmersen, P., Engø-Monsen, K., 2007. Missed Call: Messaging for the Masses. A Study of Missed Call Signaling Behavior in Dhaka (Internal Telenor R&I Publication). Telenor R&I, Fornebu.

Geser, H., 2004. Towards a Sociological Theory of the Mobile Phone. University of Zurich (Retrieved 14.02.05.).

Goggin, G., Crawford, K., 2011. Generation disconnections: youth culture and mobile communication. In: Ling, R., Campbell, S. (Eds.), Mobile Communication: Bringing Us Together or Tearing Us Apart?, The Mobile Communication Research Series, vol. II Transaction, New Brunswick.

Gordon, E., Silva, A. de S. e, 2011. Net Locality: Why Location Matters in a Networked World, first ed. Wiley-Blackwell, Oxford.

Green, N., Haddon, L., 2009. Mobile Communications. Berg, Oxford.

Green, N., Smith, S., 2004. "A spy in your pocket"? Monitoring and regulation in mobile technologies. Surveillance & Society 1 (4).

Grinter, R., Eldridge, M., 2001. y do tngrs luv 2 txt msg? In: Prinz, W., Jarke, Y., Rogers, K., Schmidt, K., Wulf, V. (Eds.) Kluwer, Dordrecht, Netherlands, pp. 219–238.

Groves, R.M., 2011. Three eras of survey research. Public Opinion Quarterly 75 (5), 861–871.

Habuchi, I., 2005. Accelerating reflexivity. In: Ito, M., Okabe, D., Matsuda, M. (Eds.), Personal, Portable, Pedestrian: Mobile Phones in Japanese Life. MIT, Cambridge, MA, pp. 165–182.

Haddon, L., 1997. Communications on the Move: The Experience of Mobile Telephony in the 1990s. Telia, Farsta.

Hampton, K.N., Sessions, L.F., Her, E.J., 2011. Core networks, social isolation, and new media. Information, Communication & Society 14 (1), 130–155.

Helles, R., 2009. Personlige medier i hverdagslivet. Københavns Universitet, Afdeling for medier, erkendelse og formidling, København.

Hicks, J., Ramanathan, N., Kim, D., Monibi, M., Selsky, J., Hansen, M., Estrin, D., 2010. AndWellness: an open mobile system for activity and experience sampling. In: Wireless Health 2010, WH '10. ACM, New York, NY, USA, pp. 34–43. http://dx.doi.org/10.1145/1921081.1921087.

Hijazi-Omari, H., Ribak, R., 2008. Playing with fire: on the domestication of the mobile phone among Palestinian teenage girls in Israel. Information, Communication & Society 11 (2), 149–166. http://dx.doi.org/10.1080/13691180801934099.

Hillebrand, F., Trosby, F., Holley, K., Harris, I., 2010. Short Message Service: The Creation of Personal Global Text Messaging. Wiley, Chichester.

Horst, H.A., Miller, D., 2005. From kinship to link-up: cell phones and social networking in Jamaica. Current Anthropology 46 (5), 755–778.

Igarashi, T., Takai, J., Youhida, T., 2005. A longitudinal study of social network development via mobile phone text messages focusing on gender differences. Journal of Social and Personal Relationships 22 (4), 691–713.

Ishii, K., 2006. Implications of mobility: the uses of personal communication media in everyday life. Journal of Communication 56, 346–365.

Ito, M., 2001. Mobile Phones, Japanese Youth and the Re-placement of Social Contact. Boston. Retrieved from: http://www.itofisher.com/PEOPLE/mito/Ito.4S2001.mobile.pdf.

Ito, M., Okabe, D., Matsuda, M., 2005. Personal, Portable, Pedestrian: Lessons from Japanese Mobile Use. MIT, Cambridge, MA.

Jagun, A., Heeks, R., Whalley, J., 2008. The impact of mobile telephony on developing country micro-enterprise: a Nigerian case study. Information Technologies and International Development 4 (4), 47–65.

Jensen, R., 2007. The digital provide: information (technology), market performance and welfare in the South Indian fisheries sector. The Quarterly Journal of Economics 122 (3), 879–924.

Johnsen, T., 2003. The social context of the mobile phone use of Norwegian teens. In: Fortunati, L., Katz, J.E., Riccini, R. (Eds.), Mediating the Human Body: Technology, Communication and Fashion. Lawrence Earlbaum, London, pp. 161–170.

Julsrud, T., 2005. Behavioral changes at the mobile workplace: a symbolic interactionistic approach. Mobile Communications, 93–111.

Kasesniemi, E.-L., Rautiainen, P., 2002. Mobile culture of children and teenagers in Finland. In: Katz, J.E., Aakhus, M. (Eds.), Perpetual Contact: Mobile Communication, Private Talk, Public Performance. Cambridge University Press, Cambridge, pp. 170–192.

Katz, J.E., Aakhus, M.A., 2002. Conclusion: making meaning of mobiles a theory of apparatgeist. In: Katz, J.E., Aakhus, M. (Eds.), Perpetual Contact: Mobile Communication, Private Talk, Public Performance. Cambridge University Press, Cambridge, pp. 301–320.

Kim, H., Kim, G.J., Park, H.W., Sogang, A., Nam, Y., 2006. The configurations of social relationships in communication channels: F2F, email, messenger, mobile phone, and SMS. In: Hartmann, M. (Ed.), ICA Pre-conference on Mobile Communication. The International Communications Association, Erfurt/Dresden, Germany.

Kubey, R., Larson, R., Csikszentmihalyi, M., 1996. Experience sampling method applications to communication research questions. Journal of Communication 46 (2), 99–120. http://dx.doi.org/10.1111/j.1460-2466.1996.tb01476.x.

Lange, K., 1993. Some concerns about the future of mobile communications in residential markets. In: Christofferson, M. (Ed.), Telecommunication: Limits to Deregulation. IOS Press, Amsterdam, pp. 197–210.

Lasén, A., 2004. A Comparative Study of Mobile Phone Use in Public Places in London, Madrid and Paris. Digital World Research Centre, University of Surrey, Guildford.

Lee, D., Jayachandran, S., 2009. The Impact of Mobile Phones on the Status of Women in India. Disponible à l'adresse: http://mobileactive.org/research/impact-mobilephonesstatus-women-india.

Lenhart, A., 2012. Teens, Smartphones & Texting. Pew Internet and American Life. Retrieved from: http://www.pewinternet.org/Reports/2012/Teens-and-smartphones.aspx.

Leung, L., Wei, R., 1999. Who are the Mobile phone have-nots? Influences and consequences. New Media & Society 1 (2), 209–226.

Licoppe, C., 2004. Connected presence: the emergence of a new repertoire for managing social relationships in a changing communications technoscape. Environment and Planning: Society and Space 22, 135–156.
Lindmark, S., 2002. Evolution of Techno-economic Systems: An Investigation of the History of Mobile Communications. Chalmers University of Technology, Department of Industrial Management and Economics, Gothenberg, Sweden.
Ling, R., 2004. The Mobile Connection: The Cell Phone's Impact on Society. Morgan Kaufmann, San Francisco.
Ling, R., 2008. New Tech, New Ties: How Mobile Communication Is Reshaping Social Cohesion. MIT Press, Cambridge, MA.
Ling, R., 2010. Texting as a life phase medium. Journal of Computer Mediated Communication 15 (2), 277–292.
Ling, R., 2012. Taken for Grantedness. MIT Press, Cambridge, MA.
Ling, R., Baron, N., Lenhart, A., Campbell, S., 2010. "Girls Text Really Weird": Cross-gendered Texting among Teens. Association of Internent Researchers.
Ling, R., Bertel, T., Sundsøy, P., 2012. The socio-demographics of texting: an analysis of traffic data. New Media & Society 14 (2), 280–297.
Ling, R., Donner, J., 2009. Mobile Communications. Polity, London.
Ling, R., Stald, G.B., 2010. Mobile communities: are we talking about a village, a clan or a small group? American Behavioral Scientist 53 (4).
Ling, R., Yttri, B., 2002. Hyper-coordination via mobile phones in Norway. In: Katz, J.E., Aakhus, M. (Eds.), Perpetual Contact: Mobile Communication, Private Talk, Public Performance. Cambridge University Press, Cambridge, pp. 139–169.
Madianou, M., Miller, D., 2011. Mobile phone parenting: reconfiguring relationships between Filipina migrant mothers and their left-behind children. New Media & Society 13.
Manning, P.K., 1996. Information technology in the police context: the "sailor" phone. Information Systems Research 7 (1), 52–62.
Mante-Meijer, E., van de Loo, H., 1998. Blurring of the life spheres: flexibility and teleworking. In: Kant, A., Mante-Meijer, E. (Eds.), Blurring Boundaries: When Are Information and Communication Technologies Coming Home. Telia/COST 248, Stockholm.
Matsuda, M., 2005. Mobile communication and selective sociality. In: Ito, M., Okabe, D., Matsuda, M. (Eds.), Personal, Portable, Pedestrian. Mobile Phones in Japanese Life. MIT Press, Cambridge, London, pp. 123–142.
McPherson, M., Smith-Lovin, L., Brashears, M.E., 2006. Social isolation in America: changes in core discussion networks over two decades. American Sociological Review 71, 353–375.
Miyata, K., 2006. Longitudinal effects of mobile internet use on social network in Japan. In: Rice, R.E. (Ed.), The Proceedings of the International Communications Association Conference 2006. The International Communications Association, Dresden.
Monk, A., Carroll, J., Parker, S., Blythe, M., 2004. Why are mobile phones annoying? Behavior and Information Technology 23 (1), 33–41.
Oksman, V., Rautianen, P., 2003. "Perhaps it is a body part": how the mobile telephone became an organic part of the everyday lives of Finnish children and teenagers. In: Katz, J.E. (Ed.), Machines that Become Us. Transaction, New Brunswik, NJ, pp. 293–308.
Platteau, J.P., Abraham, A., 1987. An inquiry into quasi-credit contracts: the role of reciprocal credit and interlinked deals in small-scale fishing communities. The Journal of Development Studies 23 (4), 461–490.
Putnam, R., 2000. Bowling Alone: The Collapse and Revival of American Community. Touchstone, New York.
Rahman, T., 2007. "Real Markets" in Rural Bangladesh: Institutions, Market Interactions and the Reproduction of Inequality. IPPG, University of Manchester, Manchester.
Rakow, L.F., Navarro, V., 1993. Remote mothering and the parallel shift: women meet the cellular telephone. Critical Studies in Mass Communication 10, 144–157.
Reid, D., Reid, F., 2 October 2005. Insights into the Social and Psychological Effects of SMS Text Messaging, 160 characters. Retrieved from: http://www.160characters.org/documents/SocialEffectsOfTextMessaging.pdf.
Richardson, R., Ramirez, R., Haq, M., 2000. Grameen Telecom's Village Phone Programme in Rural Bangladesh: A Multi-Media Case Study. TeleCommons Development Group (TDG), Guelph, Ontario.
Rose, G., 2006. Mobile phones as traffic probes: practices, prospects and issues. Transport Reviews 26 (3), 275–291. http://dx.doi.org/10.1080/01441640500361108.
Sey, A., 2011. "We use it different, different": making sense of trends in mobile phone use in Ghana. New Media & Society 13 (3), 375–390. http://dx.doi.org/10.1177/1461444810393907.
Skog, B., 2002. Mobiles and the Norwegian teen: identity, gender and class. In: Katz, J.E., Aakhus, M. (Eds.), Perpetual Contact: Mobile Communication, Private Talk, Public Performance. Cambridge University Press, Cambridge, pp. 255–273.
Smoreda, Z., Thomas, F., 2001. Social networks and residential ICT adoption and use. In: EURESCOM Summit 2001 3G Technologies and Applications. EURESCOM, Heidelberg.
Strayer, D.L., Drews, F.A., Crouch, D.J., 2006. A comparison of the cell phone driver and the drunk driver. Human Factors 48 (2), 381–392.
Townsend, A.M., 2000. Life in the real time city - mobile telephones and urban metabolism. Journal of Urban Technology 7, 85–104.
Trosby, F., n.d. The History of SMS – from a Participant's Perspective, an Unpublished History of the Specification of SMS.
Trosby, F., 2004. SMS, the strange duckling of GSM. Telektronikk 3 (2004), 187–194.
Vold Lexander, K., 2011. Texting and African language literacy. New Media & Society 13 (3), 427–443. http://dx.doi.org/10.1177/1461444810393905.
Wallis, C., 2011. Mobile phones without guarantees: the promises of technology and the contingencies of culture. New Media and Society 13 (3), 471–485.
Wellman, B., 2002. Little boxes, glocalization, and networked individualism. In: Tanabe, M., van den Besselaar, P., Ishida, T. (Eds.), Digital Cities II: Computational and Sociological Approaches. Springer, Berlin, pp. 10–25.
West, J., Mace, M., 2010. Browsing as the killer app: explaining the rapid success of Apple's iPhone. Telecommunications Policy 34 (5–6), 270–286. http://dx.doi.org/10.1016/j.telpol.2009.12.002.
Wikle, T., 1930. The detroit police department radio system. The American Journal of Police Science 1 (5), 456–465.
Yunus, M., 2003. Halving poverty by 2015–we can actually make it happen. The Round Table 92 (370), 363–375.

Mobility and Travel Activity Patterns

Joe Weber, University of Alabama, Tuscaloosa, AL, USA
Mei-Po Kwan, University of Illinois at Urbana-Champaign, Champaign, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mobility refers to a range of phenomena at a range of spatial and temporal scales. It typically refers to the movement of people. If this movement is permanent, it falls into the topic of migration. If it is voluntarily carried out for recreation, it may be tourism. At more local scales it is a necessary component of daily life, whether as commuting to work or in order to carry out other daily activities. Mobility is an important topic in the field of transport geography and touches on many methodologies. It is fundamental to the study of spatial interaction between places and the study of flows of people and goods.

Mobility refers to a range of phenomena at a range of spatial and temporal scales. It typically refers to the movement of people, including migration, trips by tourists, and commuting trips. Migration refers to permanent or long-term movement by people, usually associated with a new home location that is relatively far away (e.g., in another city) from the original home location. It has been and continues to be a major reason for changing spatial patterns of population. It may be voluntary, often for economic motives, or involuntary, as with refugees or slaves. Seasonal migration, such as for farm labor, may be a recurring pattern or migration may be permanent, as with settlers moving to a new country. It is common to speak of push and pull factors in such cases, with push factors being those that encourage people to leave a particular place and pull factors those that attract migrants to an area. Economic motives can be important to both push and pull factors. In the twentieth century, rural to urban migration and suburban migration have become important sources of change in population patterns. The US decennial census records current household location and the previous location 10 years ago; this provides the principal source for understanding migration within the country at interregional and intrastate scales. In other countries, data on migration can be very difficult to obtain and may not exist.

Tourism can be defined broadly as travel for recreational purposes, whether local, domestic, international, or even into outer space. It is an important economic activity in many areas. Tourism is a relatively recent phenomenon and has been highly influenced by improvements in transport technology/space–time convergence. It is a type of mobility expected to grow in magnitude, at least in the developing world, but with uncertain socioeconomic and environmental impacts. Recent studies have examined the mobility patterns of tourists using global positioning system (GPS) data and advanced analytical methods such as sequence alignment and Geographic Information Systems (GIS)-based visualizations (e.g., Shoval and Isaacson, 2007). These studies observed distinctive patterns of visits (in terms of the order in which tourists visit a set of sites) and routes for different social groups.

Daily Mobility

Daily mobility falls within the field of transport geography and it is at the local level where the movement of individual and households has been most intensively studied. The study of individual activity–travel patterns has been an important topic since the 1960s. Major objectives of this area of study are to examine how people make decisions about destinations, transport modes, and trip chaining. One finding is that the household level is absolutely essential to understanding people's daily mobility, as an individual's ability or need to travel is dependent on various household activities and other demands on travel resources (Hanson and Hanson, 1980; Kwan, 1999).

The daily commute to and from work is the most studied component of activity–travel behavior because traditionally it has been the primary source of road congestion and therefore the primary focus of urban transportation planning models; commuting is also the form of daily travel that is most easily measured due to the availability of decennial census data up to the 2000 Census and subsequently the American Community Survey. In this case mobility is treated as a derived demand, dependent on the need to travel out of the house to a workplace. Standard urban models incorporate mobility in the form of commuting. These models are based on an economic trade-off between commuting costs (increasing with distance from the center of the city) and housing prices (decreasing with distance from the city center). Different households will choose different locations based on their abilities to pay and household priorities (Alonso, 1964). In this monocentric urban model, land use is based on the spatial mobility of workers. Mobility is assumed to be identical for all people, and treated largely as the equivalent of distance. Actual commuting behavior suggests that few people try to minimize their distance to work by relocating their homes or workplaces (Clark et al., 2003). Such residential relocation is not necessarily easy or simple.

Congestion is a consequence of too much mobility and limited transport capacity. It reduces mobility for certain modes at certain times of the day and week. Over the long term, congestion could reshape urban patterns by encouraging people to move their homes or workplaces to maintain their commuting time. However, in recent decades it has been noted and debated that commuting times are not increasing as would be expected given the increasing population of large cities in the US. This has been termed the commuting paradox (Gordon et al., 1991), and is explained, in part, by the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.72093-9

development of polycentric or decentralized cities in which more people may live closer to their place of employment than possible in a monocentric city. Maintaining reasonable levels of mobility is a powerful, if unexamined, goal for households. In fact, a law of constant commuting time (Hupkes, 1982) has been specified, which notes average times allotted to commuting have remained surprisingly constant over time and across many transport modes.

Activity–Travel Patterns and Automobility

The activity-based approach became important in the 1970s. This is based on the idea that travel is not usually undertaken for its own purpose but reflects a need to carry out particular activities (Jones et al., 1983). Any understanding of a person's travel patterns must be based on his or her activity patterns. These activities are part of an agenda of things a household must do. These activities may or may not require travel, and if they do, it may be carried out in a variety of ways. The scheduling and implementation of activities will create a need for travel.

Studies of activity–travel patterns stemmed from dissatisfaction with metropolitan transportation planning that was based on aggregated data representing census or planning zones. These data are often based on census household data, which provide average commuting data by zone. The Census Transportation Planning Products also provides some additional aggregate flow data. Data for individual travel behavior are more limited. The US Department of Transportation collects a National Household Travel Survey data set from a small number of households nationally. Much of the data available about individual household mobility come from disaggregate travel diary surveys collected at the level of individual metropolitan areas. These are collected from a representative sample of households in a particular city, and record each activity each person engaged in, when, for how long, whether travel was involved, and if so, details of the trip. These data sets have provided tremendous insights into mobility, especially when used with GIS. In the last decade or so, regional planning agencies in the U.S. and transport researchers have started to collect data of people's activities and movement with GPS or other location-aware wireless technologies. These high-resolution space–time data have encouraged the development of new analytical methods that enable researchers to yield new insights about people's activity–travel patterns (Shen et al., 2013).

The mode of transportation available is an important factor to mobility. For many in the developed world, automobility is the predominant form for more than very short travel, and may be the only mode feasible in many cases. In the developed world, the rise of car ownership is the predominant factor in understanding mobility changes. Statistics such as those on vehicle miles traveled show that increasing car ownership and car use have increased the average number of miles that people travel. Intelligent transportation systems are a broad range of technologies that have been deployed to help preserve or improve mobility by car (Figueiredo et al., 2001). Public transit systems are found in cities of all sizes and in rural areas, but the vast majority of ridership is accounted for by the largest cities (American Public Transportation Association, 2013). Those who do not possess or are unable to use a car may be completely dependent on transit ('captive riders') for any trip outside their immediate neighborhood (Rutherford and Wekerle, 1988).

Mobility is not equally distributed among the population. Women have shorter commutes and typically travel shorter distances than men (Hanson and Pratt, 1988). This has been explained in various ways that are based on their typically greater household responsibilities and lower salaries. As mobility in developed worlds is strongly dependent on income, which often follows divisions of race and ethnicity, minority populations often have lower levels of mobility due to reduced access to cars (Johnston-Anumonwo, 1997; Lee and Kwan, 2011).

Another group with low levels of mobility is children. Children also have decreasing levels of mobility in many developed countries, as parents increasingly restrict their independent mobility to areas within the neighborhood or around the home due to fear (O'Brien et al., 2000). Neighborhoods are designed around the needs of automobiles and drivers and may not be safe for pedestrians or bicycles. Children in these areas may be restricted to a small area unless accompanying a parent in a car. While generations of children walked to school in the past, the journey to school has become increasingly reliant on buses or parent's motor vehicles (McMillan, 2005).

The elderly face mobility limitations for physical or mental reasons as well as their frequent inability to make use of automobiles (Shoval et al., 2011). Impaired mobility, especially physical reasons preventing walking, is a serious health problem for the elderly, and a decrease in mobility can swiftly be followed by health declines. The disabled may be physically limited in their mobility at very local scales, such as individual buildings, though they have been successful in improving their mobility through legislation.

Differential mobility, or the lack of mobility, has been implicated in several urban issues relating to employment. Spatial mismatch refers to a condition in which low-mobility populations, typically minority workers, in central city locations are unable to reach suburban employment locations. The decentralization trends, which have maintained stable automobile-based mobility, have shifted employment away from older walking and transit-based neighborhoods to newer automobile-oriented suburban areas. Those workers who lack social mobility and limited physical mobility are effectively trapped in areas with few jobs. They face a mismatch between their home locations and that of potential jobs (McLafferty and Preston, 1996).

Spatial entrapment refers to a phenomenon in which women, who commonly have less mobility due to financial constraints and household responsibilities, are limited in their ability to seek employment far away from home. Employers may take advantage of this and locate facilities in suburban locations where they can tap into this relatively immobile workforce (Nelson, 1986). Social exclusion is a phenomenon in which certain populations are not able to participate fully in society or are excluded from accessing important services. This may be due to a variety of factors including discrimination, socioeconomic differences, language, but also a lack of

mobility. Enabling greater mobility by immobile populations has therefore become an important goal.

Conceptualizations of Human Mobility and Activity–Travel Patterns

Distance is an increasingly small factor in understanding human mobility. New transport and communication technologies have vastly restructured the importance of distance as a deterrent to travel or interaction (Kwan and Weber, 2003). At one point, the death of distance was forecast, but travel has continued to increase despite the widespread ability to substitute virtual travel for physical mobility. Improved information and communications technologies will reduce constraints on mobility. GIS and related technologies have greatly increased the ability to measure, analyze, and visualize human mobility. The ability to work with data sets with large numbers of individuals and locations and connect locations through network space with dynamic travel times has greatly enhanced our understanding of human activity–travel patterns. Mobility is important to space–time accessibility, but the availability of destinations or opportunities is also important. An individual's mobility may be very high at night, but if few destinations are open, it is of little importance.

An important framework for understanding mobility is time geography (Hägerstrand, 1970; Kwan, 2004). This was put forth by Torsten Hägerstrand in 1970 and has been further developed by many others. The time–geographic framework examines individual activity–travel patterns during the course of a day or some longer period. While many approaches to mobility focus on enabling technologies, time geography is oriented toward constraints on mobility. These are typically described as physical and biological limits to mobility, limits due to authority (explained below), and those related to an individual's schedule of activities and need to coordinate tasks at different places at different points in time.

Physical and biological limits are straightforward. Humans are likely to travel at only 3–4 miles per hour (4–6 km h^{-1}) on foot for fairly short distances. Sleep, food, and other bodily functions must be attended to at regular intervals. Use of mechanized transport will greatly speed up travel, but there are still limits. Individuals will vary widely in their physical limitations or tolerances, and even more widely when age, gender, and disability are taken into account.

Authority limits to mobility exist because individuals do not have the right to go anywhere at any time, even in a free society. Access may be restricted to property owners, employees, or those with the proper permission or security clearance. The authority may be formal or customary, or reflect a fear of consequences. Totalitarian countries may have significant limitations on the right to move freely, and the ability to freely pass between countries has always been restricted. Even where mobility is relatively unrestricted it will still be subject to regulation, as with speed limits on highways. Authority constraints are likely to be fairly uniform across the population in an egalitarian society, but there will always be situations in which certain individuals have greater rights or fewer restrictions than others.

Coupling constraints refer to the limits on mobility resulting from a person's need to undertake joint activities with other individuals (e.g., a meeting) at particular times and places. This constraint reflects the fact that individuals do not possess complete control over their own activities and are likely to be involved in fixed activities, those for which the individual cannot modify the location and timing. These fixed activities provide the framework for their daily activity schedule (Weber and Kwan, 2002). Any other activities that an individual desires to pursue and has the ability to choose times and locations are flexible activities. Flexible activities can be carried out only if they are feasible given the times, duration, and location of fixed activities. The classic case involves an individual who wishes to leave his or her office during a lunch break, and is limited to potential restaurants that can be reached and the food eaten within the time available. The number of potential restaurants will be highly variable among individuals depending on their typical activities and the area in which they live. It will be further modified by individuals' incomplete information about travel times and locations, as well as the uncertainty or unreliability of travel times.

An important factor in understanding mobility is the phenomenon of space–time convergence or compression (Janelle, 1969). This was identified in the 1960s and refers to decreasing travel times between places due to improved transportation, such as jet planes or long-distance freeways. Higher speeds and/or shorter routes decreased travel time between cities, essentially bringing these places closer together. Much of the story of space–time convergence is that of transportation technology innovations, but there are important social and economic consequences. The phenomenon of jet lag is a physical manifestation of convergence, and the finding of constant commuting times is another. Travel times between places may increase due to a deterioration of transport facilities, increasing congestion, or political disruptions. This is space–time divergence. This has been observed in Central Africa and North Korea where mechanized transport systems have deteriorated and walking has once again become the predominant mode.

The importance of the personal automobile to mobility in developed countries, and especially the US, has led to concerns that people are too dependent on cars for mobility.

Mobility has therefore become an important topic in urban design. A debate exists as whether cities in developed nations are too dependent on automobility. This may create a health problem to individuals due to reduced exercise and to society in the form of pollution and reliance on nonrenewable fuels. It may be that people in developed countries possess too much mobility, or at least mobility that is dependent on nonrenewable fuels and creates considerable pollution. The concept of sustainability evaluates whether a practice can be continued in the long term, and automobile-based mobility in its present form is not considered sustainable. This view has led to frequent calls to reduce auto use and redesign society around other forms of transport, typically transit, though the sustainability of transit is itself questionable without new fuels and large numbers of passengers. The concept of excess commuting measures the extent to which observed commuting patterns and distances are near the minimum theoretically possible (Giuliano and Small, 1993). Others seek to promote more

basic types of mobility, such as bicycling and walking. These forms of active transport have benefits both for transportation as well as personal health. Those who defend the role of automobiles in US society may describe their role as preserving mobility, which has come to be viewed by many as a basic right.

In recent years a 'new mobilities paradigm' has emerged based on the work of sociologists, geographers, and transport researchers. It asserts the ontological significance of the movements of people, goods, and information, and challenges the conventional notion that places and regions are stable entities whereas movements are transient and analytically unimportant (Urry, 2007). The paradigm emphasizes that all places are connected by networks of movements and flows (which human geography has recognized since the 1940s), and that the time people spend on traveling is not dead time that they always seek to minimize.

See also: Access: Geographical; Automobility; Information Society, Geography of; Location Theory; Time-Space in Geography; Tourism, Geography of; Transportation Geography; Urban Geography.

Bibliography

Alonso, W., 1964. Location and Land Use. Harvard University Press, Cambridge.
Adey, P., 2010. Mobility. Routledge, New York.
American Public Transportation Association, 2013. Ridership Report. http://www.apta.com/resources/statistics/Pages/ridershipreport.aspx (accessed 14.02.14.).
Clark, W., Huang, Y., Withers, S., 2003. Does commuting distance matter? Commuting Tolerance and Residential Change Regional Science and Urban Economics 33, 199–221.
Cresswell, T., 2006. On the Move: Mobility in the Modern Western World. Routledge, New York.
Figueiredo, L., Jesus, I., Machado, J., Ferreira, J., de Carvalho, J.M., 2001. Towards the development of intelligent transportation systems. Intelligent Transportation Systems 88, 1206–1211.
Giuliano, G., Small, K., 1993. Is the journey to work explained by urban structure? Urban Studies 30, 1485–1500.
Gordon, P., Richardson, H., Jun, M., 1991. The commuting paradox: evidence from the top twenty. Journal of the American Planning Association 57, 416–420.
Hägerstrand, T., 1970. What about people in regional science? Papers of the Regional Science Association 24, 6–21.
Hanson, S., 2010. Gender and mobility: new approaches for informing sustainability. Gender, Place and Culture 17 (1), 5–23.
Hanson, S., Hanson, P., 1980. Gender and urban activity patterns in Uppsala, Sweden. Geographical Review 70, 291–299.
Hanson, S., Pratt, G., 1988. Spatial dimensions of the gender division of labor in a local labor market. Urban Geography 9, 180–202.
Hupkes, G., 1982. The law of constant travel times and trip rates. Futures 14, 38–46.
Janelle, D.G., 1969. Spatial reorganization: a model and concept. Annals of the Association of American Geographers 59, 348–364.
Jones, P.M., Clarke, M., Dix, M., Heggie, I., 1983. Understanding Travel Behaviour. Gower, Aldershot.
Johnston-Anumonwo, I., 1997. Race, gender, and constrained work trips in Buffalo, NY, 1990. The Professional Geographer 49 (3), 306–317.
Kwan, M.P., 1999. Gender, the home-work link, and space-time patterns of non-employment activities. Economic Geography 75 (4), 370–394.
Kwan, M.P., 2004. GIS methods in time-geographic research: geocomputation and geovisualization of human activity patterns. Geografiska Annaler B 86 (4), 267–280.
Kwan, M.P., Weber, J., 2003. Individual accessibility revisited: implications for geographical analysis in the twenty-first century. Geographical Analysis 35 (4), 341–353.
Kwan, M.P., Xiao, N., Ding, G. Assessing activity pattern similarity with multidimensional sequence alignment based on a multiobjective optimization evolutionary algorithm. Geographical Analysis, forthcoming.
Lee, J.Y., Kwan, M.P., 2011. Visualisation of socio-spatial isolation based on human activity patterns and social networks in space-time. Tijdschrift voor Economische en Sociale Geografie 102 (4), 468–485.
McLafferty, S., Preston, V., 1996. Spatial mismatch and employment in a decade of restructuring. The Professional Geographer 48, 420–431.
McMillan, T., 2005. Urban form and a child's trip to school: the current literature and a framework for future research. Journal of Planning Literature 19, 440–456.
Nelson, K., 1986. Labour demand, labour supply and the suburbanization of low-wage office work. In: Scott, A.J., Storper, M. (Eds.), Production, Work, Territory. Allen and Unwin, Boston, pp. 149–172.
O'Brien, M., Jones, D., Sloan, D., 2000. Children's independent spatial mobility in the urban public realm. Childhood 7, 257–277.
Rutherford, B.M., Wekerle, G.R., 1988. Captive rider, captive labor: spatial constraints and women's employment. Urban Geography 9 (2), 116–137.
Shen, Y., Kwan, M.P., Chai, Y., 2013. Investigating commuting flexibility with GPS data and 3D geovisualization: a case study of Beijing, China. Journal of Transport Geography 32 (1), 1–11.
Shoval, N., Isaacson, M., 2007. Sequence alignment as a method for human activity analysis in space and time. Annals of the Association of American Geographers 97 (2), 282–297.
Shoval, N., Wahl, H.W., Auslander, G., Isaacson, M., Oswald, F., Edry, T., Landau, R., Heinik, J., 2011. Use of the global positioning system to measure the out-of-home mobility of older adults with differing cognitive functioning. Ageing & Society 31, 849–869.
Urry, J., 2007. Mobilities. Polity Press, Cambridge.
Warf, B., 2008. Time-Space Compression: Historical Geographies. Routledge, London.
Weber, J., Kwan, M.P., 2002. Bringing time back in: a study on the influence of travel time variations and facility opening hours on individual accessibility. The Professional Geographer 54, 226–240.

Mobility, Social

Walter Müller, Mannheim Center for European Social Research, University of Mannheim, Mannheim, Germany
Reinhard Pollak, Freie Universität Berlin, Berlin, Germany; and WZB Berlin Social Science Center, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social mobility is the movement in time of individuals, families, or other social units between positions of varying advantage in the system of social stratification of a society. Classical authors have studied social mobility primarily in its contribution to class or status group formation. Recent research concentrates on identifying the degree to which individuals' social opportunities in life are conditioned by their social origins (conditions of life in the parental family) and on specifying the individual, institutional, and societal factors responsible for it. Studies of the process of status attainment show that the social status or prestige level of a job attained by a person at a given point in his or her career strongly depends on earlier job achievements, educational qualifications, and parents' status. Studies of intergenerational class mobility analyze the patterns of transition from parental (origin) class to individuals' own (destination) class and show that class destinations heavily depend on class origins. While, in general, a high degree of constancy over time and cross-national similarity among economically advanced societies is found in relative rates of mobility, social fluidity has slightly increased in some societies in recent decades. Absolute rates of mobility vary more than relative rates across societies and time, yet have partly converged between societies in recent decades.

Social mobility is the movement in time of social units between different positions in the system of social stratification of a society. The moving units are usually individuals, but may also be groups (families or kinship units). Society is seen as comprising at any given time an ordered system of more or less fixed positions that individuals or groups can hold and potentially move between. The system is stratified to the extent that (1) the positions are more or less advantageous and provide their holders with unequal power, material or symbolic means, and privilege; (2) individuals or groups have different chances of access to or movement between the positions of unequal advantage. This second aspect represents the core domain in the study of social mobility.

Social mobility research distinguishes between *inter*generational and *intra*generational mobility. *Inter*generational mobility compares individuals' positions in adult life with the positions of their families of origin; *intra*generational (or career) mobility studies position sequences within individuals' own generations (life-time), for instance, the moves between the successive jobs in working life. To assess social mobility, researchers distinguish between absolute rates and relative rates of social mobility (replacing older concepts of structural and exchange mobility). Depending on the research tradition, scholars exploit different concepts of social stratification (class, prestige, status, income), and they use different units of analysis (individuals, couples, households, or larger groups).

Research Questions and Traditions of Study

Classical authors have primarily seen social mobility or immobility as a mechanism of *class or status group formation*. The basic assumption is that classes, status groups, or other collectivities with some form of sociocultural or sociopolitical identity can only emerge and potentially provide a basis for collective action on condition that they achieve a sufficient degree of 'demographic identity' (Goldthorpe) in their membership, that is, when they include a kernel of members homogeneous in terms of their social origin and stable over time (families across generations or individuals through their working lives). Along these lines, Marx – whose main concern with mobility is otherwise with *collective* downward mobility of small proprietors and other middle classes into the proletariat – considers *individual* mobility or immobility as a crucial determinant of working-class organization in his discussion of the North American class structure. Contrary to Europe, classes in America, he argues (much like does later Sombart), "have not yet become fixed, but continually change and interchange their elements in a constant flux" (Marx, 1958: 225). While mobility weakens class formation in the working class, the rule of the ruling class is stabilized by assimilating 'the foremost minds of a ruled class.' Also stressing the connection to social mobility, Max Weber even defines a social class as an ensemble of those class situations among which high interchange of individuals typically takes place. But intergenerationally, Weber argues, the possibility of upward social mobility is limited and even decreasing.

After World War II, research on social mobility has largely turned to the study of *social opportunity*, aiming to determine the relative openness or rigidity of the stratification system concerning access to positions of varying advantage. Research has in particular concentrated on identifying the degree to which individuals' prospects in life are determined by the social conditions of their family of origin and on specifying the individual, institutional, and societal factors responsible for it. The more strongly the positions persons hold are conditioned by their social provenance, the more rigid and closed is a stratification system; the greater the mobility between social origin and a persons' own position, the more open or fluid that system is.

Issues related to social mobility are often the object of normative and political discourse and are then often interpreted ideologically, for instance, as ways of legitimizing social inequalities. In this vein, variants of liberal political philosophy from John Stuart Mill onward regard unequal conditions as legitimate when they result from (fair) competition or when unequal chances of access to privileged positions are due to unequal achievements. In its extreme form, the *functionalist theory of stratification* (*see* Social Stratification) holds unequal rewards to be deserved as they are assumed to be distributed according to the contributions made to the efficient functioning (and ultimately survival) of society and are thought to be necessary to attract the most able and hard-working individuals to the most important positions. Though at times influenced by normative preferences, research in the field is analytical and empirical. Its results may, however, serve to examine to what degree and for what reasons social reality conforms to or falls short of particular normative or political goals.

The investigation of social mobility is mainly pursued by way of observation and comparison of national populations. Comparative designs are essential to determine the extent to which openness or rigidity of the stratification system is related to other characteristics of societies. Macrotheoretical approaches highlight characteristics such as the degree of societies' modernity, their economic and political order (e.g., capitalist market vs state command economies), particular institutions or policies. Research has especially focused on the evolution of social mobility in line with the development of industrial society. Notably, with the transition from preindustrial to industrial society, a *historical increase* in mobility has been assumed, either as a precondition for the take off of industrial development or as its consequence (Lipset and Zetterberg, 1959); others assume a *steady increase* in mobility as industrialism advances and an increasing *uniformity* and *convergence* of social structures and social mobility in industrial societies, due to standardizing tendencies inherent in advanced technologies and other elements of industrialism. Finally, *exceptionally* high rates of mobility have been postulated for specific societies, particularly those not burdened by the legacy of the European feudal past or its class-based stratification system, such as the United States, Australia, or Japan. To test such hypotheses, numerous data sets in many different societies have been created, primarily through large-scale population surveys, but also (notably in studies on earlier periods) by using suitable information from official records.

For the purposes of a structured summary, much of the literature can be represented as a debate between two different paradigms and research traditions. To study mobility, both essentially rely on work positions, that is, the (occupational) jobs individuals hold to earn a living. The *status attainment* approach, on the one side, sees the principal property of these jobs in terms of their respective location in a finely graduated continuum of rewards related to the position and conveyed to their incumbents. The continuum of rewards is generally conceived in terms of social prestige or socioeconomic status and measured by corresponding scales (*see* Social Stratification). The *class mobility* approach, on the other side, conceives the (structure of) positions among which mobility can occur as discrete classes, which differ from each other by properties of a qualitatively different nature rather than along one single vertical dimension only. In earlier studies, categories consisted of rather broad groupings of occupations defined in an ad hoc manner (sometimes as broadly as farm/manual/nonmanual). More sophisticated class schemas are grounded in theoretically comprehensive conceptions of the class structure. In their now widely accepted class schema, Erikson and Goldthorpe (1992) draw on the character of the employment relations as core criterion to identify different classes, in addition to the criterion of ownership of the means of production. Related with the differing conception of the main properties of the positions among which mobility is seen to occur, the two traditions also differ in their theoretical understanding of the mobility process, the methods used, and the main conclusions arrived at.

The Process of Status Attainment

Research pursued in the *status attainment* framework is often – even though not necessarily – embedded in a more general (liberal) theory of industrialism and shares with it assumptions about the impacts of industrialization on other aspects of social structure, notably social mobility. Industrialization is understood as a pervasive force that sooner or later affects all national societies across the globe. In this process, it is argued, the order of stratification in preindustrial societies in which *ascription* by family, kin, ethnic ties, gender, or status group affiliation determines the position of individuals in society, is increasingly replaced by a new order in which *achievement* based on individual abilities, qualifications, and work accomplishments play the decisive role in social selection (Blau and Duncan, 1967; Treiman, 1970). A fundamental trend toward rationalism and universalism is seen to unfold. To survive in markets, actors are forced by the technological sophistication of production processes and growing market competition to increasingly allocate individuals "to positions according to what they have shown they can do, rather than according to their social provenance" (Goldthorpe, 1996: 256). Notably formal educational qualifications are expected to increasingly determine the job level (successively) reached in bureaucratic work organizations. To enhance the development, states expand educational services, subsidize the costs of education, thus providing easier access to education for children in less-fortunate economic conditions. As industrialization progresses, social placement of individuals will then be conditioned less by their social origin, and intergenerational social mobility between the positions of parents and those of their children is expected to increase.

In this theoretical orientation, social mobility is understood as a process of successive status attainment by individuals, in which later attainment depends on earlier attainment. In line with the conception of the occupational structure as a graded hierarchy, this process is studied with variants of linear regression models (often path models or structural relations equations) to determine the extent to which the status or prestige level of a person's job at a given point in his or her career depends on earlier job achievements, educational qualifications, and directly or indirectly on parents' status. In the leading study – Blau and Duncan's (1967) *The American Occupational Structure* – education is identified as the main determinant of individuals' occupational careers and as the core link between

social origins and attained position. Parental education and occupational status also affect children's occupational status or incomes directly, but their effects are weaker than those of education. Parental education and status are mainly mediated through education. One far-reaching conclusion has been that under such conditions of status attainment – powered by individual achievement – the very basis for class solidarity and class action disappears. Extensions of the model have studied how other family conditions, personal abilities (intelligence, motivation), sociopsychological factors (parental encouragement and support, educational and occupational aspirations, peer contacts), or characteristics of schools intervene in these processes and contribute to generating them. In evidence from a number of societies at different levels of industrial development, Treiman and Yip's (1989) findings confirmed the expectation that status attainment becomes increasingly achievement based as industrialization unfolds.

The status attainment paradigm has come in for criticism on a number of counts. Its representation of the attainment process is individualistic and one-sidedly supply-driven, concentrating as it does on the way familial resources and individual human capital affect occupational and other outcomes. But the transformation of individual resources into outcomes depends on existing opportunities. The latter are constrained both by the distribution of available positions and by the competition of those contending for the same positions with similar or better resources. Further, the unidimensional representation of positions in terms of status or prestige obscures the way in which opportunities are bound up with structural transformations such as the shifts in forms of employment and between sectors, industries, and occupations. Groups like farmers, skilled industrial workers, small shopkeepers, and service workers hold positions that are heterogeneous in terms of their sociostructural location and affected in different ways by particular state policies and structural transformations. However, in terms of status or prestige, they all are similarly rated. In the hierarchical framework it thus appears difficult, if not impossible to trace mobility processes to their causal conditions.

Intergenerational Class Mobility and Social Fluidity

Proponents of the *class framework* have attempted to overcome these shortcomings by reorienting the study of mobility to analyze how individuals move between the social class positions existing within a class structure. The moves are represented in mobility matrices, which show the rates of transitions from given origin positions (from which moves start) to destination positions. In their elementary form, mobility matrices thus are bivariate frequency tables in which the cells contain the frequencies of transitions observed between origin and destination. Origin and destination can refer to different sets of positions. In the study of intragenerational mobility this is often the social class of first job and current job, but can also be any two successive positions in individuals' own working life. In the case of intergenerational mobility – which is of particular interest in the study of social mobility and to which we refer in the following – the classes compared are those of parents and those their children occupy in adult life. Several important strengths result from this reorientation. Compared to hierarchical ratings, the class characteristics allow the derivation of more realistic hypotheses concerning the greater or lesser likelihood of mobility between particular origins and destinations. Structural transformations and changing opportunity structures can be more specifically captured with class categories.

Of particular significance in this respect is the distinction between *absolute* and *relative* rates of mobility. *Absolute rates* refer to the observed absolute number of moves that take place in reality, and to proportions such as the percentage of children who have a different (percentage mobile) or the same (percentage immobile) class position as their parents. If the class categories are assumed to be ordered, distinctions are made between horizontal, upward, and downward mobility, meaning moves between classes at the same hierarchical level, moves from a lower to a higher and moves from a higher to a lower level, respectively. Other often used measures for absolute rates are *outflow* proportions (percentages of offspring of a given origin attaining particular destinations) or *inflow* proportions (percentages of members in a given destination recruited from the various origin classes).

Relative rates, in contrast, compare the relative chances the offspring of a given origin class – compared to those of another origin class – have to attain the various destination classes. These relative chances are assessed by so-called odds-ratios, which can be interpreted as measure of the advantages or disadvantages the progeny of one origin class compared to those of another origin class have in obtaining one destination class rather than another. The information contained in a mobility matrix can then be expressed in a set of such odds-ratios, which in turn can be analyzed by various log-linear modeling techniques (for early contributions see Goodman, 1970; Hauser, 1979; *see* Multivariate Analysis: Discrete Variables (Overview); Binary Response Models and Logistic Regression; 42062) Odds-ratios have the crucial property of being independent of the relative size of origin and destination classes (the so-called *marginal distributions* of the mobility matrix). As it is easier to obtain a position in larger (growing) rather than in smaller (shrinking) classes, the marginal distributions and the differences between the origin and destination distribution can be understood as capturing the structure of opportunities to reach different classes. The odds-ratios thus express the pattern of association between origin and destination independently of a given opportunity structure. They are therefore widely accepted to indicate the degree of *social fluidity* between the class positions of parents and their offspring, and in turn the degree of openness or rigidity of the class structure or the stratification system.

Relative rates thus are assumed to indicate the fundamental relativities existing in a mobility regime in the chances of people of different origin to reach the different destinations, or as Featherman et al. (1975) formulate: They indicate the genotypical pattern of the mobility regime in contrast to the phenotypical pattern revealed by the absolute rates. Even though absolute rates are more directly observed, theoretically they are conceived to be implied by the more fundamental conditions of social fluidity (measured by the relative rates) on the one side, and, on the other side, by the opportunity structures, which are assumed to be conditioned by macrostructural

economic and other developments and to influence the distributions of class positions available in particular times.

Given the differences in the conceptual understanding of absolute and relative rates of mobility, theoretical explanations for the two types differ, and findings are also likely to differ. As to relative rates, researchers emphasizing the hierarchical dimension have stressed the distances (in status or other considered dimensions) between positions as the mobility-generating principle (e.g., Hout and Hauser, 1992). The larger the vertical distance between positions, the less fluidity there will be between them. Sophisticated distance models have consistently shown considerable distance effects, but generally they cannot explain all details of the existing pattern. Researchers in the class framework have acknowledged the multidimensional nature of the resources that are available to the offspring of various class origins and that can be used for gaining access into classes of varying advantage and desirability. In a microtheoretical approach, Goldthorpe (2007) proposes an explanation in terms of – resource-constrained – rational action. Members of the various social classes pursue the same aims, giving priority to avoiding downward mobility and only in the second place seeking to achieve upward mobility. But in pursuing these same goals and priorities, the different amounts and kinds of resources and constraints typically available prompt the working and intermediate classes to use different strategies than the upper classes. In their comparative study *The Constant Flux*, Erikson and Goldthorpe (1992) developed a 'Core model' of social fluidity with four (im)mobility mechanisms in order to cover and analyze developments in a broad range of industrial societies up to the 1970s. They show that alongside the impacts of vertical '*hierarchy*' the fluidity pattern is markedly shaped by '*inheritance*' of parental class position (in all classes, but especially in the petty bourgeoisie, among farmers and in the service class), by the '*sector*' division between employment in agriculture and outside agriculture, and by special *affinities* expressing enhanced mobility among particular classes like the various nonmanual classes.

In a more recent paper, Jonsson et al. (2009) propose additional mechanisms for intergenerational class (im) mobility. Their 'microclass' approach sees typical class features (e.g., identity, social closure, collective action, and class outcomes) working at a detailed occupational level rather than through few analytically aggregated class categories. Accordingly, mechanisms of intergenerational reproduction mainly operate on a microclass level, e.g., through the transmission of occupation-specific human capital, occupation-specific cultures, tastes and networks, and fixed economic resources directly connected to parental occupations (e.g., business, farm). Their empirical analysis shows indeed an impressive effect of microclass inheritance, leaving additional (aggregated) class effects rather inconsequential, at least for men.

All in all, when assessed in terms of *relative rates*, class destinations in mature industrial societies continue to depend heavily on class origins. In England or France, for instance, the odds of a son born to a service class father obtaining a service class position rather than one in the unskilled working class is more than 20 times larger than the corresponding odds for an unskilled worker's son. Mobility is more likely to occur for intermediate classes while the likelihood of staying immobile in the parental class is larger in the most and least advantageous classes. Classes that require the ownership of material, financial capital, or land are 'inherited' to a clearly larger extent from the parental to the offspring generation than classes not requiring such assets.

Much research continues to address the question whether social fluidity differs between countries or changes over time. Erikson and Goldthorpe (1992) argue that the cumulative effect of different starting positions, resources, and strategies is so pertinent that it produces a high stability in the resulting outcomes, both across time and in different countries. They propose accepting, albeit only in a weak form, the hypothesis originally advanced by Featherman et al. (1975), that similar relative rates of social mobility exist in all industrial societies with a nuclear family system and a market economy. High degrees of cross-national commonality and constancy over time, however, are not to be equated with complete uniformity. Some limited, but nevertheless significant differences are found between societies and change is also found within some societies across time. As changes in time do not form a systematic trend (associated with industrial or some other macrosocietal development), Erikson and Goldthorpe favor Sorokin's (1927) assumption of 'trendless fluctuation.' Also, concerning differences between countries, there is hardly evidence for systematic explanations: The hypothesis of *exceptionally high* fluidity in the 'new' societies (United States, Australia, or Japan) cannot be confirmed. Nor does social fluidity systematically vary according to the political-economic order of countries: The former communist Eastern European societies show quite similar fluidity patterns and variations of them as in capitalist societies. Deviations from the core pattern appear to arise rather from historically contingent conditions resulting from a nation's particular economic, social, and political history or from specific institutional arrangements, for instance, in the educational system, in labor market segmentation, or in other institutions with an impact on specific elements of the fluidity pattern. For instance, the long-term equalizing policies by the social democratic governments in Sweden appear to have fostered fluidity in this country. The strong discriminatory measures pro- and contra particular classes in some phases of the communist period appear to have had corresponding effects in Hungary.

In a more recent study examining the development of social fluidity from the 1970s up to the 1990s in a number of mostly European societies, Breen (2004) and Breen and Luijkx (2004) largely confirm a high degree of commonality, but find somewhat more variation between countries and more change over time than Erikson and Goldthorpe (1992). In the 1990s, with the exception of Britain, all countries studied showed higher fluidity than in the 1970s, even though not all of these changes were statistically significant. Yet, countries did not become more similar in their fluidity pattern. In the 1990s, Germany, France, Ireland, and Italy had comparatively low levels of fluidity while Israel, the Netherlands, Sweden, Norway, Poland, and Hungary had a more fluid pattern of relative mobility. While Breen and Luijkx (2004) find no evidence that macrosociological characteristics such as the level of economic development or income inequality could explain the differences between countries and the change over time, they provide tentative hypotheses how characteristics of a country's

educational system (educational inequality and expansion and the strength of the linkage between educational qualifications and labor market outcomes) might influence the level of social fluidity.

Absolute rates differ more than relative rates between countries and across time. They depend on both the pattern of social fluidity and the opportunity structures. Related to economic and occupational change, mobility is induced as workers are 'forced' to leave areas of declining employment and gravitate to those which are expanding. Societies differ in their class composition and in the pace of change over time, and these differences in the opportunity structure induce differences in absolute rates between countries and across time even at a common or constant pattern of social fluidity.

In contrast to the relative rates, Breen and Luijkx (2004) find evidence that in more recent decades various aspects of absolute rates (total, upward, downward) converged between economically highly developed societies. This is mainly due to a growing similarity in the class structure of these societies: the agricultural classes are declining to a similarly low level while the service classes are increasing to a relatively similarly high level. With these developments absolute upward mobility slightly increased in most countries and – in the 1990s – was about twice as large as downward mobility. Depending on the kind and speed of changes in social fluidity and the transformation of the class structure, the probability of individuals moving between different classes will change. The absolute numbers of such moves determine the observable mobility experience of a population. They also strongly affect the composition of class membership by persons intergenerationally stable or recruited from other class origins. Even though they are likely to influence processes of class formation, these issues have been hardly studied so far.

Largely similar relative rates alongside varying absolute rates are also found, if different population subgroups are compared. Substantial differences between men and women, immigrants and natives, or between groups of different racial or ethnic affiliation can exist in absolute rates because the groups hold jobs in different segments of the labor market and their distribution among classes varies. But in terms of social fluidity they differ less. If they sometimes do, differences tend to be small. Inequalities in class distribution existing between men and women as a result of gender-segregated labor markets, for instance, are mostly not intensified by men and women being affected differently by their class origins. As a confirming corollary, it is found that in times when married women's labor market participation was low, social fluidity for daughters in the marriage markets has been rather similar to the relative mobility prospects of their brothers in the labor market. With respect to gender, another highly debated issue concerns the proper unit of stratification, class and mobility analysis. Is it (always) women's or men's own position in the labor market that determines their social status or class affiliation, or is it rather the common position of the family unit? As women's labor force participation has increased and has become more continuous, the most convincing answer is to study the issues separately to analyze men's and women's labor market careers and social mobility separately, likewise their household division of labor and resource utilization, but also the common position they may share.

In the figures on mobility or immobility between social origin and eventual destination, whether viewed in absolute or relative terms, the process of intergenerational social reproduction is only reflected in its global outcomes. It results from various intervening processes, notably educational attainment, transition from education to work, and mobility between successive jobs in working life within and between firms. These processes have to be – and are – studied on their own right (*see* Life Course: Sociological Aspects). Of these, the role of education has been scrutinized most intensively. The fact that education is a prerequisite for access to professional and managerial jobs makes it an important channel for upward social mobility, but evidently one which the more advantaged families also make use of to secure favorable positions for the next generation. Further, while education is found to be the main carrier responsible for the hierarchical element in the fluidity pattern, this is much less true – in some cases not true at all – in the case of other elements or mechanisms that produce immobility (Ishida et al., 1995). Transfer of capital or business, provision of social contacts, or inculcation of the proper social skills and manners are some of the means prosperous families have to back their children outside school and which they can rely on to compensate for possible failure in education. Studies pursued in the class framework so far have not found any evidence consistently supporting expectations of a *general* trend toward increased education-based or 'credentialist' social selection in late-industrial or postindustrial societies, at least not in the sense of education consistently becoming a stronger predictor of social destination. Rather, in many countries the predictive power of education has tended to weaken, especially when the proportions of highly educated people has exceeded the demand for such qualifications on the labor market. On the other side, over time educational attainment became less dependent on social origin in several countries (see Breen et al., 2009, contesting the earlier conclusion of persistent educational inequality by Shavit and Blossfeld, 1993). Declining educational inequality can contribute to increasing social fluidity. Yet such changes appear to be limited in extent and tend to affect cohorts in education at a particular time. Accordingly, much of the change in social fluidity of a population can be attributed to the replacement of older, more immobile cohorts with younger, more mobile cohorts (Breen and Jonsson, 2007). But the effects of declining educational inequality on social mobility tend to be small, and to affect social mobility in the population at large it will take long times.

Alternative Approaches to Social Mobility

Class- or status-based approaches are often challenged for their choice of unit of analysis (individuals vs family or household) and their rootedness in occupational characteristics, leaving long-term unemployed or inactive persons out of the analysis. Thus, some researchers, economists in particular, favor measures of (household) income or wealth as an underlying hierarchy in society. They split income or wealth distributions of origin and destination positions into segments (percentiles) and calculate correlations or elasticities between origin and destination segments. The empirical

robustness of these approaches has often been weak due to the lack of appropriate data (small samples, imprecise measures of long-run income estimated from short-run fluctuating data). With improved data bases and clarification in concepts and corresponding statistical measures, recent research has made substantial progress (for a review see Björklund and Jäntti, 2009). Based on large administrative data sets from tax records, Chetty et al. (2014) study whether in the United States the increasing inequality in income distribution has been connected – as some earlier research claims – with increasing intergenerational persistence of incomes. Findings differ depending on the aspect addressed. Similar to the sociologists' distinction between relative and absolute mobility, Chetty et al. (2014) distinguish the association between parent and child percentile rank in the income distribution on the one side, and the inequality in the marginal distribution of incomes in the parent and child generation on the other side. For children born between 1971 and 1992, the association between the (relative) income ranks has not changed in spite of growing inequality in the marginal distribution. Yet due to the latter, the inequality in (absolute) incomes of children raised in a family in the bottom income quintile compared to those in the top quintile has widened substantially – since "the rungs of the income ladder have grown further apart" Chetty et al. (2014). Yet, also because of an increasing dispersion, the chances of a child from a low-income family (i.e., low percentile) to gain an income beyond a high threshold of, e.g., $100 000 have *improved*. As for the US cohorts born between the 1950s and the 1970s, Hertz (2007) and Lee and Solon (2009) find intergenerational income association similar to those observed for the more recent cohorts by Chetty et al. (2014), stability in the association appears to have prevailed in the United States over much of the post-World War II period. According to Chetty et al. (2014), this stability for the United States at large appears to exist hand in hand with stable differences in the size of the association in different regions of the United States. From international comparisons findings accumulate showing differences between countries: The United States is among the countries with high intergenerational income association while the link between parents and children's income position is considerably lower e.g., in several of the Scandinavian countries.

Sibling correlations are an alternative approach to assess social mobility without relying on the classical origin-destination pattern in terms of class or income positions. Sibling correlations examine correlations of various economic and noneconomic outcomes (like wage, income, education, or physical attributes) among two or more children of a family and draw conclusions on the effect of the origin family on these very outcomes. Estimating these effects is a challenging task, given that high correlations among siblings may have various causes (same family of origin, same neighborhood, same school, etc.).

Few studies in social mobility research, usually very limited in scope, turn their focus to the causal study of specific mobility mechanisms. Systematic interventions studies, especially in the educational context, examine potential long-term upward mobility pathways, drawing on experiences from the classical Perry Preschool Project (Heckman et al., 2009). Other studies concentrate on, for example, discriminating behavior of employers in the (class) position attainment process by using experiments with mock applications (Jackson, 2009). Thus, much work needs to be done, when it comes to causal explanations of what can be said in descriptive accounts.

See also: Binary Response Models and Logistic Regression; Class: Social; Educational Sociology; Equality of Opportunity; Equity and Education; Family and Schooling; Generations, Sociology of; Income Inequality; Life Course: Sociological Aspects; Log-Linear and Marginal Models; Multivariate Analysis: Discrete Variables (Overview); Social Inequality and Schooling; Social Inequality in History (Stratification and Classes); Social Mobility, History of; Social Stratification; Socioeconomic Status and Academic Achievement.

Bibliography

Björklund, A., Jäntti, M., 2009. Intergenerational income mobility and the role of family background. In: Oxford Handbook of Economic Inequality. Oxford University Press, Oxford, pp. 491–521.

Blau, P.M., Duncan, O.D., 1967. The American Occupational Structure. Wiley, New York.

Breen, R. (Ed.), 2004. Social Mobility in Europe. Oxford University Press, Oxford.

Breen, R., Jonsson, J.O., 2007. Explaining change in social fluidity: educational equalization and educational expansion in twentieth-century Sweden. American Journal of Sociology 112, 1775–1810.

Breen, R., Luijkx, R., 2004. Social mobility in Europe between 1970 and 2000. In: Breen, R. (Ed.), Social Mobility in Europe. Oxford University Press, Oxford, pp. 37–77.

Breen, R., Luijkx, R., Müller, W., Pollak, R., 2009. Nonpersistent inequality in educational attainment: evidence from eight European countries. American Journal of Sociology 114, 1475–1521.

Chetty, R., Hendren, N., Kline, P., Saez, E., Turner, N., 2014. Is the United States still a land of opportunity? Recent trends in intergenerational mobility (No. w19844). American Economic Review: Papers and Proceedings 104 (5), 141–147.

Erikson, R., Goldthorpe, J.H., 1992. The Constant Flux: A Study of Class Mobility in Industrial Societies. Clarendon Press, Oxford.

Featherman, D.L., Jones, F.L., Hauser, R.M., 1975. Assumptions of social mobility research in the US: the case of occupational status. Social Science Research 4, 329–360.

Goldthorpe, J.H., 1996. Problems of "meritocracy". In: Erikson, R., Jonsson, J.O. (Eds.), Can Education be Equalized? The Swedish Case in Comparative Perspective. Westview, Boulder, pp. 255–287.

Goldthorpe, J.H., 2007. On Sociology, second ed., vol. 2. Stanford University Press, Stanford.

Goodman, L.A., 1970. The multiplicative analysis of qualitative data: interactions among multiple classifications. Journal of the American Statistical Association 65, 226–256.

Hauser, R.M., 1979. Some exploratory methods for modeling mobility tables and other cross-classified data. Sociological Methodology 10, 141–158.

Heckman, J.J., Moon, S.H., Pinto, R., Savelyev, P.A., Yavitz, A.Q., 2009. The rate of return to the High/Scope Perry Preschool Program. Journal of Public Economics 94, 114–128.

Hertz, T., 2007. Trends in the intergenerational elasticity of family income in the United States. Industrial Relations: A Journal of Economy and Society 46 (1), 22–50.

Hout, M., Hauser, R.M., 1992. Symmetry and hierarchy in social mobility: a methodological analysis of the CASMIN model of class mobility. European Sociological Review 8, 239–266.

Ishida, H., Müller, W., Ridge, J.M., 1995. Class origin, class destination, and education: a cross-national study of ten industrial nations. American Journal of Sociology 60, 145–193.

Jackson, M., 2009. Disadvantaged through discrimination? The role of employers in social stratification. British Journal of Sociology 60 (4), 669–692.

Jonsson, J.O., Grusky, D.B., Di Carlo, M., Pollak, R., Brinton, M.C., 2009. Micro-class mobility – social reproduction in four countries. American Journal of Sociology 114, 977–1036.

Lee, C.I., Solon, G., 2009. Trends in intergenerational income mobility. The Review of Economics and Statistics 91 (4), 766–772.

Lipset, S.M., Zetterberg, H.L., 1959. Social mobility in industrial societies. In: Lipset, S.M., Bendix, R. (Eds.), Social Mobility in Industrial Societies. University of California Press, Berkeley.

Marx, K., 1958. The eighteenth Brumaire of Louis Bonaparte. In: Marx, K., Engels, F. (Eds.), Selected Works. Foreign Language Publications, Moscow.

Shavit, Y., Blossfeld, H.P. (Eds.), 1993. Persistent Inequality: Changing Educational Attainment in Thirteen Countries. Westview Press, Boulder, CO.

Solon, G., 2002. Cross-country differences in intergenerational earnings mobility. Journal of Economic Perspectives 16, 59–66.

Sorokin, P.A., 1927. Cultural Mobility. Free Press, New York.

Treiman, D.J., 1970. Industrialization and social stratification. In: Laumann, E.O. (Ed.), Social Stratification: Research and Theory for the 1970s. Bobbs Merrill, Indianapolis.

Treiman, D.J., Yip, K.-B., 1989. Educational and occupational attainment in 21 countries. In: Kohn, M.L. (Ed.), Cross-national Research in Sociology. Sage, Newbury Park.

Relevant Websites

http://www.isa-sociology.org/rc28.htm – International Sociological Association.

http://www.journals.elsevier.com/research-in-social-stratification-and-mobility/ – Elsevier, The Official Journal of the ISA RC28 on Social Stratification and Mobility.

http://sociology.as.nyu.edu/page/rc28 – New York University, Department of Sociology.

Model Selection and Model Averaging

Gerda Claeskens, ORStat and Leuven Statistics Research Center, KU Leuven, Leuven, Belgium
Maarten Jansen, Université libre de Bruxelles, Brussels, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Model selection methods provide a way to select one model among a set of models in a statistically valid way. Such methods include tools for variable selection in regression models. Asymptotic properties such as consistency and efficiency, the specific use of the model, or properties regarding minimization of a certain risk function such as the expected prediction error, may help to decide which method to choose. Model selection is a special case of model averaging where the estimators obtained from different models are combined in a weighted average. Model averaging avoids the selection of one model. The choice of the weights may be determined by a model selection method or may come from a priori knowledge in a Bayesian framework.

Introduction

Statistical models may not coincide with the true, physical model that has generated the data. This may be because the true model is too complicated from the mathematical or computational points of view, or simply because the precise model is unknown. Even if the precise model is known and computationally feasible, a simplification could be preferable. As an illustration, consider two quantities x and y that are related by a law $y = \beta_0 + \beta_1 x + \beta_2 x^2$, and suppose we observe x_i along with $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$, with independent zero mean normal errors ε_i. If we are interested in the value of y corresponding to $x = x_0$, then least squares regression estimators for the values of β_i lead to an unbiased estimator for y. For small values of β_2, however, it might be interesting to adopt a model without the quadratic term. The bias introduced by dropping this variable from the model might be compensated by a reduction in variance. The statistical bias-variance tradeoff suggests in such case that a simpler, but incorrect model outperforms the true model. In addition, the simpler model may have an easier interpretation in the practical context.

The best performing statistical model is also dependent on the parameter to be estimated. In the example above, the model without the quadratic term would obviously fail if we were interested in the value of β_2. Not all models have the same performance. Some models may be good for prediction purposes, other models concentrate on particular features of the data.

These considerations leave the researcher the difficult task to select an appropriate model. Alternatively, instead of selecting one model and further ignoring the other models, estimates or predictions arisen from different plausible models may be combined to lead to model-averaged estimates or predictions.

Model selection methods are developed to select models in a statistically valid way. Model averaging is a wider concept that allows to combine estimators in a weighted average of single model estimators and that offers the possibility to take the uncertainty of the model choice into account when working further with a selected model.

Information Criteria

Model selection has been well studied in regression models for the problem of variable selection. Some of these criteria, such as R^2 and Mallows' C_p are only valid in multiple linear models, other criteria are valid to select more general models (e.g., likelihood-based) and to various estimation methods (e.g., maximum likelihood estimation, nonparametric estimation). Several of such methods for model choice are listed below. Numerous books on this topic have been published, including those by Linhart and Zucchini (1986) and Miller (2002), with a large focus on regression models, the edited monograph by Lahiri (2001), Burnham and Anderson (2002), with an emphasis on the Akaike information criterion (AIC), and that by Claeskens and Hjort (2008).

The Value of R^2 in Linear Regression Models

The question of which variables to include in a linear regression model is a classical example of model choice. Consider a linear model of the form $Y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_p x_{p,i} + \varepsilon_i$ for $i = 1, \ldots, n$, where often a normal distribution is assumed for the variables $\varepsilon_1, \ldots, \varepsilon_n$ with mean zero and variance σ^2. A summary value of the regression model is the R^2 value, the value of multiple determination, defined as the proportion of the variation of the response that is explained by the regression model. More precisely, $R^2 = \mathrm{SSR}(\widehat{\beta})/\mathrm{SST}$ with $\mathrm{SST} = \sum_{i=1}^n (Y_i - \overline{Y})^2$ with $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$ and $\mathrm{SSR}(\widehat{\beta}) = \sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2$ with $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1,i} + \ldots + \widehat{\beta}_p x_{p,i}$, the prediction formed by inserting the least squares estimators for the unknown parameters. We also define $\mathrm{SSE}(\widehat{\beta}) = \mathrm{SST} - \mathrm{SSR}(\widehat{\beta})$. The larger the value of R^2, the better the model. Let the model with p regression coefficients be the full model, i.e., the largest model one is willing to consider. An all subsets search fits all possible submodels of this full model by leaving out some (or all) of the covariates. The simplest model includes a constant only: $Y_i = \beta_0 + \varepsilon_i$. For each such submodel the corresponding R^2 value can be computed and the submodels may be ranked according to their R^2 value. Since including more variables in a model cannot decrease the R^2 value, this is not satisfactory to select

a good model. Better is the adjusted R^2 value where the degrees of freedom, the number of parameters in the fitted model, is taken into account, $R^2_{\text{adjusted}} = 1 - \frac{\text{SSE}(\widehat{\beta})/(n-p-1)}{\text{SST}/(n-1)}$. The model selection strategy is then to select a model with a small number of variables for which the corresponding adjusted R^2 value is large.

Mallows' C_p

Minimizing the mean squared prediction error $\text{PE}(\beta) = E[\sum_{i=1}^{n}\{\widehat{Y}_i - E(Y_i)\}^2]$ in multiple regression models is the idea behind the C_p construction of Mallows (1973). The sample version is defined by $C_p = \text{SSE}(\widehat{\beta})/\widehat{\sigma}^2 + 2p - n$, where $\widehat{\sigma} = \text{SSE}(\widehat{\beta}_{\text{full}})/(n - \text{q} - 1)$ with q the number of parameters in the full model and $\widehat{\beta}_{\text{full}}$ the full model least squares estimator of β. The statistic C_p is constructed as an estimator of the scaled squared prediction error $\text{PE}(\beta)/\sigma^2$. Models with a large bias tend to have C_p values that are larger than p. The best model according to this method is that model with the smallest value of C_p, which is often identified via a graphical representation of the values of C_p versus p.

Akaike's Information Criterion

The information theoretic approach of Akaike (1973) tries to find a model that comes close in Kullback–Leibler discrepancy to the true model. This method assumes the existence of a true model, but does not need to have knowledge about that model for the actual calculation of the criterion value.

Let $f(y; \theta)$ denote the modeled density, possibly depending on unknown parameter values θ, and let g be the true density, then the Kullback–Leibler discrepancy between the model and the true densities is defined by $\text{KL}[f(\cdot;\theta), g] = \int \log[g(y)/f(y;\theta)]g(y)\text{d}y$. When $f = g$, the Kullback–Leibler discrepancy is equal to zero, for all other densities f it holds that $\text{KL}(f, g) \geq 0$. A good model has a small value of $\text{KL}(f, g)$.

For independent and identically distributed data $Y_1, \ldots, Y_n$, the estimator $\widehat{\theta}$ that maximizes the log likelihood $\sum_{i=1}^{n}\log f(Y_i;\theta)$ converges, under some conditions (see, e.g., White, 1994), almost surely to the value θ_0 that minimizes $\text{KL}[f(\cdot;\theta), g]$.

The AIC results from estimating the expected value of $\text{KL}[f(\cdot;\widehat{\theta}), g]$. This leads to $\text{AIC} = -2\sum_{i=1}^{n}\log f(Y_i;\widehat{\theta}) + 2p$, with p the length of the vector θ. For regression models where covariate vectors x_i are observed together with a response value Y_i, for $i = 1, \ldots, n$, AIC similarly is defined as $\text{AIC} = -2\sum_{i=1}^{n}\log f(Y_i; x_i; \widehat{\theta}) + 2p$.

The general strategy in model selection via AIC is to compute for each model under consideration the value of the corresponding AIC for the given dataset and to rank the models according to their AIC value. The model with the smallest AIC is considered the best model.

The AIC values may be compared for non-nested models and do not need any knowledge about the true model. This method is applicable to more general model selection and is not restricted to variable selection within a model class. Consider, for example, the choice between an exponential density function and a Weibull density that has one additional parameter. The AIC values of both models may be computed on a given dataset and the smallest AIC value determines for that dataset which model is preferred in the sense of having the smallest estimated expected Kullback–Leibler discrepancy.

The concept of using AIC for variable selection has been extended in various ways toward various forms of model misspecification, robustness against outliers, estimation using other methods than maximum likelihood, corrections for small sample sizes, mixed models containing both random and fixed effects, missing data, etc. See, for example, Claeskens and Hjort (2008) and Burnham and Anderson (2002).

The Schwarz Bayesian Information Criterion

The Bayesian Information Criterion (BIC) has been proposed by Schwarz (1978) and Akaike (1977, 1978). One reason for its development was to have a selection method with different asymptotic properties than the AIC, see further in Section Asymptotic Properties of Model Selection Methods.

While the name of the criterion refers to Bayesian methods, its computation does not require any knowledge of Bayesian procedures. In fact, the formula of the BIC differs only from that of the AIC in the last term. For the situation of a model for independent and identically distributed data, $\text{BIC} = -2\sum_{i=1}^{n}\log f(Y_i;\widehat{\theta}) + p\log(n)$,with p the length of the vector θ and log the natural logarithm with base exp(1). For regression models BIC is similarly defined as $\text{BIC} = -2\sum_{i=1}^{n}\log f(Y_i; x_i, \widehat{\theta}) + p\log(n)$. The BIC has been extended toward allowing other types of models and estimation methods.

The definition for BIC originates as an approximation of the posterior probability of a model given the data (see, e.g., Claeskens and Hjort, 2008: Section 3.2.2), the priors do not appear in the BIC formula. A Laplace approximation is applied to the marginal likelihood for the model where the parameters that are present in the model are integrated over their parameter space with respect to their prior density. Only the most important terms in this approximation are kept, leading to the BIC. The absence of prior information in the BIC makes this perhaps less attractive from a Bayesian point of view. Other Bayesian approaches have been developed, such as working with Bayes factors and the use of the deviance information criterion (Spiegelhalter et al., 2002). Minimum description length also may include aspects of Bayesian modeling for model choice (Rissanen, 1996; Grünwald, 2007) (*see* Bayesian Model).

The Focused Information Criterion

The idea behind the focused information criterion (FIC, Claeskens and Hjort, 2003) is that a good model should lead to a small mean squared error for the focus, that is, a parameter that one wishes to estimate using the model. The mean squared error of an estimator is a sum of the squared bias and the variance of that estimator. The value of FIC for a given focus μ and a given model is an estimator for the mean squared error of $\widehat{\mu}$ computed using that model. Different models may give rise to different estimators of μ, hence to

different mean squared errors and to different values of the FIC. The model with the smallest value of FIC is considered the best model.

Consider as an example a multiple regression model where we wish to estimate $\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \beta^t x$. Allowing different submodels to be fit, leads to different estimators $\widehat{\mu}_s = \widehat{\beta}_s^t x_s$ where the subscript $_S$ denotes working with a subset of the variables only. For example, the narrow model estimator $\hat{\mu}_\emptyset = \hat{\beta}_0$, and the estimator in the model only including variables x_2 and x_3 is $\widehat{\mu}_{\{2,3\}} = \widehat{\beta}_0 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3$. We wish to select that submodel for which μ is estimated with the smallest mean squared error.

A good model for estimation of μ should correspond to an estimator that has a small or no bias and at the same time a small variance. This leads to a bias–variance tradeoff since typically models with a large number of parameters will tend to have a small bias, but a large variance since many parameters need to be estimated. On the other hand, models with only a small number of parameters might have a small variance but tend to have larger bias values.

The calculation of the mean squared error requires the knowledge of a true model. Therefore, the derivation of the FIC assumes a local misspecification setting where the true model takes the form of the full model but with true parameter values of the form $\theta_0 + \delta/\sqrt{n}$ with n the sample size. This allows the calculation and estimation of the mean squared error for the estimators.

One important consequence of directing the model search toward performing well for the specified focus is that different focus quantities might lead to different selected models. In the above multiple regression example, this might mean that specifying another vector $\tilde{x}$, perhaps at the boundary of the covariate space, while the former x could have located near the center, might lead to selecting a different model. Specifying the error variance σ^2 as a focus could also require a different model for best estimation in mean squared error sense. This is in contrast with other information criteria that select one model for all purposes.

The use of the FIC has been extended toward non- and semiparametric estimation, survival data analysis, quantile regression, and time series models to name just a few areas.

Cross-Validation and Generalized Cross-Validation

The method of cross-validation splits the dataset (usually multiple times) into a so-called training set and a validation set. These two sets do not need to have the same size. Typically a model is fit using the training set. The validation set is then used to compute the value of an expected loss function, the risk function, using that fitted model. When multiple splits of the dataset are used, the resulting risk functions are averaged (or summed), which is named the cross-validation value. When this process is repeated for different models, the model with the smallest such cross-validation value is considered the best one.

In multiple regression models, leave-one-out cross-validation applied to the prediction sum of squares leads to the PRESS statistic (Allen, 1974). One observation vector, say the ith one $(y_i, x_{1,i}, \ldots, x_{p,i})$, is left out of the dataset. The remaining data values are used to fit a multiple regression model leading to estimated coefficients $\widehat{\beta}_j^{[-i]} (j = 0, 1, \ldots, p)$ that are used to predict the value Y_i via $\widehat{Y}_i^{[-i]} = \widehat{\beta}_0^{[-i]} + \widehat{\beta}_1^{[-i]} x_{1,i} + \cdots + \widehat{\beta}_p^{[-i]} x_{p,i}$. This prediction is considered more honest than the least squares prediction using all of the data values since here the prediction is performed for a value that is not already used for estimation of the regression coefficients. One then considers PRESS $= \sum_{i=1}^n (Y_i - \widehat{Y}_i^{[-i]})^2$. Since the leave-one-out predictions depend on the variables that are in the model, submodels may be ranked according to their value of PRESS. The model with the smallest value of PRESS is considered the best one.

For large datasets typically the dataset is divided into k parts (k is usually between 3 and 10). One part is the validation set, the other $k-1$ parts are together the training set. This approach named k-fold cross-validation is less time consuming than leave-one-out methods. For an overview on cross-validation, see Arlot and Celisse (2010).

As the training and validation procedure of cross-validation can be time consuming, the idea of cross-validation can be formalized and then generalized into the much simpler expression of generalized cross-validation (GCV). GCV is a function of the estimated parameter $\widehat{\beta}$ within a model under consideration. This function can be minimized over all possible models, with their corresponding estimators. $\mathrm{GCV}(\widehat{\beta}) = \mathrm{SSE}(\widehat{\beta})/(1 - \nu/n)^2$, where ν is a generalized concept of degrees of freedom (Ye, 1998), which depends on the model under consideration as well as on the estimation within that model. The expression of GCV can be analyzed to be closely related to the formula of Mallows' C_p.

Asymptotic Properties of Model Selection Methods

One of the favorable properties that a model selection method may have is efficiency. In the setting of a multiple regression model, by efficiency is understood that, conditional on the selected model, the ratio of the expected squared prediction error for predictions using this model for the sample values, $\sum_{i=1}^n E[(\widehat{Y}_i - Y_i)^2]$ and of the best possible such expected squared prediction error among the considered models, say this occurs for model S^* with predictions $\widehat{Y}_{i,s*}$, $\sum_{i=1}^n E[(\widehat{Y}_{i,s*} - Y_i)^2]$, converges in probability to one when the sample size grows toward infinity. It has been shown that AIC, the small sample corrected version of AIC and Mallows' C_p are among the efficient selection criteria (Hurvich and Tsai, 1995; Lee and Karagrigoriou, 2001). Yang (2005) has obtained the minimax rate optimality of information criteria of the AIC-type. The BIC and Bayesian model averaging are shown to be not efficient and not minimax rate optimal.

Consistency of the selection is another asymptotic property that some criteria may possess. Sin and White (1996) study consistency for a general class of information criteria. If there is only one model in the collection of considered models that comes closest to the true model in Kullback–Leibler discrepancy, then a weakly consistent criterion shall with probability one select this model, when the sample size tends to infinity. Weak consistency requires weak

assumptions and is satisfied by a large number of criteria for which the penalty term is of the order $o_P(n)$, including both AIC and BIC.

Consistency is a stronger property. This involves selecting the true model when it is one of the considered models, and selecting a parsimonious model, that is a model with a small number of parameters, when there are two or more models that are otherwise close in Kullback–Leibler distance, or for which the log-likelihood ratio statistic is bounded in probability. For precise conditions, refer to Sin and White (1996). The BIC and the criterion of Hannan and Quinn (1979) are shown to be both consistent and strongly consistent under some conditions. The latter property concerns almost sure convergence, while consistency involves convergence in probability. The criteria AIC and Mallows' C_p are among the nonconsistent criteria.

Having efficiency and consistency properties at the same time is not possible (Yang, 2005). This issue also holds when consistent choices for the weights in model averaging are used, in particular Bayesian model averaging is not minimax rate optimal (Yang, 2005).

Pretesting and Conditional Inference

Once a model is selected from a collection of available models, proper inference should take the uncertainty involved with the selection into account as indeed the selected model is random. Too often, however, this source of uncertainty is ignored. The same ignorance arises with pretesting when only coefficients that are significantly different from zero are included in a model. Danilov and Magnus (2004) explain that model selection and estimation should be studied simultaneously. Working conditional on the selected model leads to too optimistic statements in the sense that, for example, confidence intervals are too narrow (Hjort and Claeskens, 2003; Leeb and Pötscher, 2005).

The randomness involved with model selection may be studied by writing an estimator of a model quantity (this may be a parameter in the model, or another focus such as a prediction at a certain location) in a selected model as a weighted average of the estimators in all considered models. Since only one model is selected, this implies that all but one weight are zero. The nonzero weight is set equal to one. The choice of which weight equals one depends on the data, implying randomness of the weights.

While the asymptotic distribution of the estimator in each model might be normal, the randomness of the weights causes the asymptotic distribution of the randomly weighted estimators to be non-normal in most cases. Proper inference on estimators-post-selection is difficult (Hjort and Claeskens, 2003; Leeb and Pötscher, 2008).

Model Averaging

While post-selection estimators use zero/one weights, allowing the weights assigned to estimators in different models to take on values other than zero or one leads to model averaging estimators which take the form $\widehat{\mu}_{\text{weighted}} = \sum_{S\in \mathbf{S}} w_s \widehat{\mu}_s$, with S being a collection of models that is used and w_S being the weight given to the estimator $\widehat{\mu}_s$ in the model identified by S. Often it is imposed that the weights sum to one and take values between 0 and 1.

Since information criteria provide a ranking of models according to their value of the criterion, values of information criteria such as Mallows' C_p, AIC, BIC, and FIC have been used to construct weights for model averaging (see, e.g., Hansen, 2007; Burnham and Anderson, 2002). Estimators $\widehat{\mu}_s$, $S \in \mathbf{S}$ in the considered models receive a weight proportional to the value of the corresponding model selection criterion, the better, the higher ranked the model according to that selection method, the larger the weight. Such 'smooth' weights as opposed to zero or one values for model selection avoid the problem that slight changes in the data would cause another model to be selected (thus switching between weight zero and weight one for the involved models). Another advantage is that all estimators from the considered models are used. Such weighted estimators are referred to as frequentist model averaging estimators (Hjort and Claeskens, 2003).

Prior information on the models leads to another choice of weights that is used in Bayesian model averaging (see, e.g., Hoeting et al., 1999) (*see* Bayesian Statistics). Once the prior probabilities are specified for each of the models in the considered model collection and for the parameters in those models, a challenge with Bayesian model averaging is in the computation of the posterior probabilities that are used in the construction of the weights.

Bayesian model averaging is much applauded since once values can be obtained from the posterior distribution of the weighted estimator $\widehat{\mu}_{\text{weighted}}$, the resulting inference takes the weight specification into account. This is different in frequentist model averaging where the calculation of the weighted estimator is relatively easy though its distribution is difficult to study (Hjort and Claeskens, 2003).

High-Dimensional Variable Selection

Best *n*-term Selection

In high-dimensional problems, the number of observations n is smaller than the number of parameters in the full model, from which it follows that not all parameters can even be estimated within that model. A typical solution for this underdetermination is to impose sparsity, i.e., to propose an estimator in which the vast majority of the variables are set to zero. Sparsity can be imposed by optimization of the sum of squared residuals with a regularization term, as in $\text{SSE}(\beta) + \lambda\|\beta\|_q$. With $\beta = (\beta_1, \ldots, \beta_p)$, the norm $\|\beta\|_q = \{\sum_{j=1}^{p} \beta_j^q\}^{1/q}$ is chosen to promote sparsity. The cases of ℓ_q-norms with $q = 0$ and $q = 1$ are discussed in more detail later. The regularization parameter λ can be seen as a Lagrange multiplier for the constrained optimization problem to find the minimum of $\text{SSE}(\beta)$ for which $\|\beta\|_q \leq C$, for a constant C which is determined by λ and vice versa.

Taking $q = 0$, the regularization term is a pseudonorm, simply counting the size of the model $\|\beta\|_0 = \#\{i; \beta_i \neq 0\}$. The estimator that minimizes the ℓ_0 regularized SSE is the least squares estimator within a submodel. In order to find the

global optimum of the ℓ_0 regularized problem, one needs to compare the SSE-values of all submodels, which is a problem of combinatorial complexity.

As an alternative, the ℓ_0 regularization can be replaced by an ℓ_1 regularization. The constraints are then linear instead of combinatorial, thereby reducing the optimization to a quadratic programming problem, which is a convex optimization problem. The problem is known as 'least absolute shrinkage and selection operator' or lasso (Tibshirani, 1996) or as basis pursuit (Chen et al., 1998). Algorithms solving the problem include direct solvers such as least angle regression (Efron et al., 2004), as well as iterative methods, such as iterative soft-thresholding (Daubechies et al., 2004) or gradient projection (Figueiredo et al., 2007). While dramatically reducing the computational cost, the outcome is still a sparse estimator for β. It has been found that if the regularization parameter is not too large and if the nonzeros in the true model are sufficiently prominent, then ℓ_1 constrained optimization is variable selection consistent (Meinshausen and Bühlmann, 2006; Wainwright, 2009; Tropp, 2006; Zhao and Yu, 2006). It has also been found that, for given λ, ℓ_1 constraints find approximately the same model as ℓ_0 constraints (Donoho, 2006).

Shrinkage

There is, however, a major difference between ℓ_0 and ℓ_1 constrained least squares. Whereas ℓ_0 constraints lead to least squares estimators within the selected model, ℓ_1 constraints lead to a shrinkage estimator within the selected model. More precisely, let I indicate the set of selected indices from β, and X_I the submatrix of X formed by selecting from X the columns indicated by I, then the Karush–Kuhn–Tucker conditions for the constrained optimization problem are

$$\left(X_I^t X_I\right)\widehat{\beta} = \mathrm{ST}_\lambda\left(X_I^t Y\right)$$

where ST_λ is the soft-threshold function with threshold λ, which is defined by $\mathrm{ST}_\lambda(x) = 0$ if $|x| < \lambda$, $\mathrm{ST}_\lambda(x) = x + \lambda$ if $x \leq \lambda$, and $\mathrm{ST}_\lambda(x) = x - \lambda$ if $x \geq \lambda$.

The shrinkage in the soft-threshold function has several benefits. It reduces the influence of false positives in the selection. This may result in better prediction errors, known as Stein's phenomenon (Stein, 1956). Shrinkage comes out from a computationally more attractive problem, but it is also mathematically more tractable, especially when optimizing over λ.

The Sparse Variable Selection Problem

The regularization parameter λ controls the balance between residual and sparsity norms. It thus determines the size of the selected model. Variable selection is done by choosing the appropriate value of λ. As ℓ_0 and ℓ_1 constraints lead to different estimators within the selected model, the best model for estimation without shrinkage does not coincide with the best model for estimation with shrinkage. Finding the best model for estimation without shrinkage is more delicate, as the prediction error for a misspecified λ rapidly increases (Jansen, 2014). As shrinkage reduces the effect of false positives, the ℓ_1 problem is too tolerant for false positives, leading to considerably overestimating the size of the true model (Jansen, 2014).

Bibliography

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B., Csáki, F. (Eds.), Second International Symposium on Information Theory. Akadémiai Kiadó, Budapest, pp. 267–281.

Akaike, H., 1977. On entropy maximization principle. In: Applications of Statistics (Proc. Sympos., Wright State Univ., Dayton, Ohio, 1976). North-Holland, Amsterdam, pp. 27–41.

Akaike, H., 1978. A new look at the Bayes procedure. Biometrika 65, 53–59.

Allen, D.M., 1974. The relationship between variable selection and data augmentation and a method for prediction. Technometrics 16, 125–127.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Statistical Surveys 4, 40–79.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, second ed. Springer-Verlag, New York.

Chen, S., Donoho, D.L., Saunders, M.A., 1998. Atomic decomposition by basis pursuit. SIAM Journal on Scientific Computing 20 (1), 33–61.

Claeskens, G., Hjort, N.L., 2003. The focused information criterion. Journal of the American Statistical Association 98, 900–916 with discussion and a rejoinder by the authors.

Claeskens, G., Hjort, N.L., 2008. Model Selection and Model Averaging. Cambridge University Press, Cambridge.

Danilov, D.L., Magnus, J.R., 2004. On the harm that ignoring pretesting can cause. Journal of Econometrics 122 (1), 27–46.

Daubechies, I., Defrise, M., De Mol, C., 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Communications on Pure and Applied Mathematics 57, 1413–1457.

Donoho, D.L., 2006. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. Communications on Pure and Applied Mathematics 59, 797–829.

Efron, B., Hastie, T.J., Johnstone, I.M., Tibshirani, R.J., 2004. Least angle regression. Annals of Statistics 32 (2), 407–499 with discussion.

Figueiredo, M., Nowak, R., Wright, S.J., 2007. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. IEEE Journal of Selected Topics in Signal Processing 1 (4), 586–597.

Grünwald, P.D., 2007. The Minimum Description Length Principle. MIT Press.

Hannan, E.J., Quinn, B.G., 1979. The determination of the order of an autoregression. Journal of the Royal Statistical Society, Series B 41, 190–195.

Hansen, B.E., 2007. Least squares model averaging. Econometrica 75 (4), 1175–1189.

Hjort, N.L., Claeskens, G., 2003. Frequentist model average estimators. Journal of the American Statistical Association 98, 879–899 with discussion and a rejoinder by the authors.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Statistical Science 14, 382–417 with a rejoinder by the authors.

Hurvich, C.M., Tsai, C.-L., 1995. Relative rates of convergence for efficient model selection criteria in linear regression. Biometrika 82, 418–425.

Jansen, M., 2014. Information criteria for variable selection under sparsity. Biometrika page to appear 101 (1), 37–55.

Lahiri, P., 2001. Model Selection. In: Institute of Mathematical Statistics Lecture Notes – Monograph Series, vol. 38. Institute of Mathematical Statistics.

Lee, S., Karagrigoriou, A., 2001. An asymptotically optimal selection of the order of a linear process. Sankhyā Series A 63, 93–106.

Leeb, H., Pöstcher, B.M., 2005. Model selection and inference: facts and fiction. Econometric Theory 21 (01), 21–59.

Leeb, H., Pötscher, B.M., 2008. Can one estimate the unconditional distribution of post-model-selection estimators? Econometric Theory 24 (02), 338–376.

Linhart, H., Zucchini, W., 1986. Model Selection. John Wiley & Sons Inc., New York.

Mallows, C.L., 1973. Some comments on C_p. Technometrics 15, 661–675.

Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. Annals of Statistics 34 (3), 1436–1462.

Miller, A., 2002. Subset Selection in Regression, second ed. Chapman & Hall/CRC, Boca Raton, FL.

Rissanen, J.J., 1996. Fisher information and stochastic complexity. IEEE Transactions on Information Theory 42, 40–47.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

Sin, C.-Y., White, H., 1996. Information criteria for selecting possibly misspecified parametric models. Journal of Econometrics 71, 207–225.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B 64, 583–639 with a discussion and a rejoinder by the authors.

Stein, C., 1956. Inadmissibility of the usual estimator for the mean of a multivariate distribution. In: Proc. Third Berkeley Symp. Math. Statist. Prob. University of California Press, pp. 197–206.

Tibshirani, R.J., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58 (1), 267–288.

Tropp, J.A., 2006. Just relax: convex programming methods for identifying sparse signals in noise. IEEE Transactions on Information Theory 52 (3), 1030–1051.

Wainwright, M.J., 2009. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). IEEE Transactions on Information Theory 55 (5), 2183–2202.

White, H., 1994. Estimation, Inference and Specification Analysis. Cambridge University Press, Cambridge.

Yang, Y., 2005. Can the strengths of AIC and BIC be shared? Biometrika 92, 937–950.

Ye, J., 1998. On measuring and correcting the effects of data mining and model selection. Journal of the American Statistical Association 93, 120–131.

Zhao, P., Yu, B., 2006. On model selection consistency of lasso. Journal of Machine Learning Research 7, 2541–2563.

Model Testing and Selection, Theory of

Karl Christoph Klauer, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Tests of the validity of probability models can be based on classical chi-squared statistics for discrete random variables. For continuous random variables, chi-squared-based methods are available as well as procedures using order statistics and log-likelihood ratio tests, among others. Classical approaches to model selection are also based on model-fit tests. A related approach is based on tests between hypotheses that correspond to the different families of models. Other approaches are the Bayesian approach, the discrepancy-based method to select models with the goal to minimize the expected value of a given discrepancy criterion, and selection based on the minimum description length principle.

One of the main tasks of applied statisticians is to construct probability models for observations. This means representing the observations in terms of random variables and parametric families of distributions. To fit a model is to provide an estimate of the probability distribution of these random variables. Goodness-of-fit statistics measure the discrepancy of the observed from the estimated distribution. Goodness-of-fit tests are used to decide whether these discrepancies are significant under the null hypothesis that the model holds. Model selection is the process of selecting one from different families of competing models. Model fit and model selection are closely related.

This article reviews procedures for assessing model fit and techniques for selecting models. Classical chi-squared-based approaches to model fit provide goodness-of-fit tests for discrete random variables that take only a finite number of different values. For continuous random variables, chi-squared-based methods are available, as well as procedures using order statistics and log-likelihood ratio tests, among others. Methods of Bayesian model checks and model criticism are also considered.

The focus here is on goodness-of-fit measures with a reference distribution that is known at least for large samples in classical goodness-of-fit approaches. There are many indices, measures, and graphical methods designed to capture aspects of model fit in a descriptive manner in different fields. For example, in structural equation modeling, there is a multitude of indices that are often considered in assessing a model's fit based on heuristical recommendations for appropriate cut-off values. Among these are the adjusted goodness-of-fit index, the standardized root mean square residual, the nonnormed fit index, the normed fit index, the expected cross-validation index, the comparative fit index, and others (Schermelleh-Engel et al., 2003). Furthermore, we focus on those approaches for assessing goodness of fit that are general-purpose methods applicable to large classes of families of distributions rather than being tailored to a specific family such as the family of normal distributions.

Classical approaches to model selection build on goodness-of-fit tests, in that they select the simplest model that fits the data. A related approach is based on tests between hypotheses that correspond to the different families of models. Other approaches considered include the Bayesian approach, within which the Bayesian information criterion (BIC) is motivated. In the discrepancy-based method, models are selected with the goal to minimize the expected value of a given discrepancy criterion, thereby typically striking a balance between parsimony and goodness of fit. Akaike's information criterion (AIC) is an example. An information-theoretic approach to model selection is based on the minimum description length principle, leading to the normalized maximum likelihood (NML) index for model selection and a Fisher information approximation to it.

Models

In a probability model, observed values x are assumed to be realizations of random variables X with probability distribution Q that is an element of a parameterized family of probability distributions $M = \{P_\theta, \theta \in \Theta\}$, where θ may be a vector with several components and Θ, the parameter space, is the set of values that can be taken by θ. The model is the family of distributions. A model is said to be nested in another model if its family of distributions is a subset of that of the second model.

Goodness of Fit

Assessing goodness of fit means the statistical testing of models with data. Based on a random sample $x_1, \ldots, x_n$ realizing random variables $X_1, \ldots, X_n$ with distribution Q, it is thus required to test the null hypothesis:

$$H_0 : Q \in M$$

Most goodness-of-fit tests are parameter-free, that is, the distribution of the test does not depend upon θ, at least asymptotically, that is, as the sample size increases.

Goodness-of-Fit Testing for Discrete Random Variables

Many traditional goodness-of-fit statistics were originally developed for discrete random variables. Discrete random variables take only a finite (or at most countably infinite) number of values. They underlie statistical models of categorical data (Agresti, 1990). The probability distribution Q of a discrete random variable X can be specified by its probability function $q(x) = Q(\{x\})$, which assigns the probability of the set

$\{x\}$ to x. Likewise, the probability distribution of P_θ is specified by the probability function $f(x|\theta) = P_\theta(\{x\})$.

Let N_x be the number of X_i with $X_i = x$, and let n_x be a realization of the random variable N_x. Goodness-of-fit tests are based on asymptotically efficient estimators $\hat{\theta}_n$ of θ such as the maximum likelihood estimator under H_0. The two most widely used tests are given by the log-likelihood ratio statistic, where log is the natural logarithm:

$$G^2 = -2\sum_x n_x[\log nf(x|\hat{\theta}) - \log n_x] \qquad [1]$$

and Pearson's chi-squared statistic:

$$X^2 = \sum_x (n_x - nf(x|\hat{\theta}))^2/(nf(x|\hat{\theta})) \qquad [2]$$

where n is the total number of observations in the sample.

Under H_0 and given suitable regularity conditions, each of the above statistics is asymptotically distributed as χ^2 with degrees of freedom given by the number of different data categories x that can occur in principle minus the number of independent parameters minus one. In practice, it may be necessary to aggregate data from several categories, forming new, composite categories from categories x for which $nf(x|\hat{\theta})$ is small for the asymptotic approximation to be adequate. Different rules of thumbs have been proposed for what constitutes an adequate cell size. Although the statistics of eqns [1] and [2] are most well known, there are many others (Read and Cressie, 1988).

Goodness-of-Fit Testing for Continuous Random Variables

Goodness-of-fit tests for models of continuous random variables can be based on those for the discrete case and on the order statistics. Apart from general-purpose methods, special tests have been developed to test for exponential and normal families of distributions among others (D'Agostino and Stephens, 1986).

The goodness-of-fit test for models of discrete random variables can be applied to the continuous case by grouping the values of the random variables by means of a finite number of intervals I_x, $x = 1, \ldots, k$. For this purpose, compute $p(x|\theta) = P_\theta(I_x)$; n_x, the number of observations falling into I_x, and, finally, the above test statistics on the basis of these grouped data and probability distributions. Estimation of θ should be based on the grouped data rather than on the original likelihood. Note that a subjective element enters the test through the choice of the intervals I_x. A rational choice is to define I_x so that $P_{\hat{\theta}}(I_x) = 1/k$, $x = 1, \ldots, k$. Although this makes the intervals dependent upon the random sample via $\hat{\theta}$, the asymptotic χ^2-distribution of the above test statistics is not compromised (Moore and Spruill, 1975). When the maximum likelihood estimates based on the ungrouped data are available, generalized chi-squared statistics such as the Rao–Robson statistic (D'Agostino and Stephens, 1986) provide more powerful goodness-of-fit tests and should be preferred.

Given continuous real-valued random variables $X_1, \ldots, X_n$, the so-called order statistics $X'_1, \ldots, X'_n$ correspond to the values of the random variables ordered from the smallest to the largest and thus, $X'_1 < \ldots < X'_n$ (with probability one). Let $0 = \lambda_0 < \lambda_1 < \ldots < \lambda_k < 1$, and let n_i be the greatest integer less than or equal to $n\lambda_i + 1$. The λ_i partition the range of cumulative probabilities into a finite number of intervals, and thus a subjective element is implicit in their choice. Denote the cumulative distribution function associated with P_θ by $F(\cdot|\theta)$, and define the statistic

$$Y_n^2(\theta) = n\sum_{i=1}^{k}[(F(X'_{n_i}|\theta) - F(X'_{n_{i-1}}|\theta)) - p_i]^2/p_i \qquad [3]$$

where $p_i = \lambda_i - \lambda_{i-1}$, $i = 1, \ldots, k$. If $\hat{\theta}_n$ is an estimator that minimizes $Y_n^2(\theta)$, then under suitable regularity conditions the test statistic $Y_n^2(\hat{\theta}_n)$ is asymptotically distributed as chi-squared with $k - q - 1$ degrees of freedom, where q is the number of independent parameters.

Testing against Specified Alternatives

The above methods are general-purpose tests that aim at realizing test power for detecting a broad range of deviations from the model assumptions. Discussions of diagnostic tests and graphical methods for detecting specific deviations abound in treatments of many particular models. A general and widely applicable principle for deriving such tests is given by Neyman–Pearson log-likelihood ratio tests of nested models. For this purpose, a second, less constrained family of distributions $N = \{P_\eta, \ \eta \in \Psi\}$ is considered of which the original family M is a subset. For example, the second model may involve the same and additional predictors in the context of multiple linear regression; it may additionally contain interactions of higher order in the context of analysis-of-variance models or generalized linear models, and so forth. A so-called conditional model test is a test of H_0 versus H_1:

$$H_0 : Q \in M \quad \text{versus} \quad H_1 : Q \in N \setminus M$$

For this test, both θ and η are estimated by means of the maximum likelihood method under M and N, respectively, and $-2\log L$ is computed, where L is the ratio of the resulting likelihood values under the two models. Under suitable regularity conditions, the test statistic is distributed as chi-squared with degrees of freedom given by the difference in the number of independent parameters of the two models. The test power of this test is particularly large for detecting violations of constraints that M imposes over and above those imposed by N, but a significant result can also mean that constraints common to M and N are already violated.

Bayesian Model Checks and Criticisms

In Bayesian inference, a prior distribution has to be specified for the model parameters with density $g(\theta)$, and inference is then based on the posterior distribution of θ given the data $x = (x_1, \ldots, x_n)$ with density $f(\theta|x)$ proportional to the product of $g(\theta)$ and $f(x|\theta)$. Consequently, Bayesian model analyses check both the adequacy of the model specification in so-called model checks and the sensitivity of inferences relative to changes in the prior distribution in so-called robustness checks.

Robustness checks examine how sensitive the drawn inferences are as a function of alternative specifications of the prior. Frequently, different priors are plausible, and in hierarchical models priors can be complex, requiring many somewhat arbitrary decisions. Checking robustness involves specifying alternative priors and assessing the changes in inferences that

occur across the set of alternative priors. A couple of principled methods have been considered (O'Hagan and Forster, 2004; Chapter 8). They are not widely used because of the computational difficulties implied for realistically complex practical models.

Model checks attempt to explore the adequacy of a model without explicit reference to alternatives. One way to assess these questions is via so-called posterior predictive model checks (O'Hagan and Forster, 2004; Chapter 8). Given the data x, new data x^{pred} can be generated from the so-called posterior predictive distribution. Its density is $p(x^{\text{pred}}|x) = \int f(x^{\text{pred}}|\theta)f(\theta|x)\mathrm{d}\theta$. A test quantity $T(x,\theta)$ is then defined that is a function of the (new or old) data and of the parameters such that it quantifies specific deviations of the data from the model predictions. For example, T might be Pearson's chi-squared statistic of goodness of fit. The probability p of $T(x^{\text{pred}},\theta) > T(x,\theta)$ is then estimated under the joint distribution of (x^{pred},θ) given the observed data. This posterior p value can be conveniently estimated if a sample of the posterior distribution is obtained via Markov chain Monte Carlo methods in an application. The model is considered adequate with regard to the deviations quantified by T if this posterior p value is not small.

One way to think of this procedure is by analogy to classical assessments of model fit. The observed value of a test statistic is evaluated with reference to the distribution of the statistic generated from the model given the current parameter estimates. The parameter estimates stem from the posterior distribution, however, instead of from maximum likelihood estimation. For a critique of posterior predictive model checks, see O'Hagan and Forster (2004) (Chapter 8).

Within this framework, different kinds of simplification, extension, and variations have been explored. For example, it has been proposed to admit not only replicated data x^{pred} but also replicated parameters θ^{pred} into the definition of T and the reference distribution. Another approach, so-called predictive model checks, samples x^{pred} from the prior predictive distribution, with density given by $\int f(x|\theta)g(\theta)\mathrm{d}\theta$ and with T typically only a function of the data and not the parameters.

Posterior predictive model checks have been criticized as being too conservative because they test for discrepancies from the observed data using a reference distribution that has been shaped by that same data. Predictive model checks avoid this difficulty but can be of little diagnostic value if the prior is vague as well as not robust under different specifications of vague priors. A number of variants of these checks have been based upon partitioning the data set into two subsets, one of which is used to generate a posterior distribution of the model parameters and the rest to test for discrepancies with reference distributions derived from that posterior distribution (see the section on classical cross-validation methods below). Another principled approach to dealing with these difficulties has been proposed by Bayarri and Berger (2000).

Model Selection

In model testing, there is one distinguished model, and the objective is to test its validity. In model selection, there are several models, and the objective is to select one that best serves a given function such as estimation of certain common parameters or prediction. Frequently, the different models differ in complexity, usually indexed by the number of independent parameters, and most methods explicitly or implicitly incorporate a trade-off between fit and parsimony. Among the many different approaches that are discussed in this emerging field are model selection by classical hypothesis tests, discrepancy-based methods, bootstrap and cross-validation techniques, Bayes approaches, and information-theoretic approaches.

Model Selection by Classical Hypothesis Tests

In nested families of models, procedures for model selection have been proposed based on the classical hypothesis testing framework sketched above. So-called sequential model tests start with the model with the minimal number of constraints. During step k of the procedure, it is assumed that the model with k constraints holds, and is tested against the model with $k+1$ constraints. The sequence of tests is terminated at step l as soon as the first hypothesis with $l+1$ constraints is rejected. In multiple linear regression, the models are defined by the subset of predictors to be included in the equation, and automated sequential selection procedures based on hypothesis tests such as forward and backward stepwise selection have been considered (Jungeilges, 1992). For nonnested, disjoint models, selection by hypothesis testing is possible by means of a generalized likelihood ratio approach originally proposed by Cox (1961) or by integrating the models into a joint supermodel first (Atkinson, 1970).

Discrepancy-Based Methods

Another principled approach to model selection has been proposed by Linhart and Zuccini (1986). A discrepancy Δ measures the lack of fit of the underlying distribution Q relative to members P_θ of candidate models. It satisfies $\Delta(\theta) = \Delta(Q,P_\theta) \geq \Delta(Q,Q)$ for all θ.

Depending upon the purpose of the model selection exercise, different aspects of model fit may be important and may lead one to consider different discrepancies. Important general-purpose discrepancies are the following, where $F(\cdot)$ is the cumulative distribution function associated with Q and $F(\cdot|\theta)$ is the cumulative distribution function associated with P_θ. (For continuous random variables, the probability functions should be replaced by probability density functions and the sums by integrals.)

Kullback–Leibler discrepancy	$\Delta_{\text{K-L}}(\theta) = -E_Q \log f(x\|\theta)$,
Kolmogorov discrepancy	$\Delta_{\text{K}}(\theta) = \sup_x \|F(x) - F(x\|\theta)\|$,
Cramér–von Mises discrepancy	$\Delta_{\text{C-M}}(\theta) = E_{P\theta}(F(x) - F(x\|\theta))^2$,
Gauss discrepancy	$\Delta_{\text{G}}(\theta) = \sum_x (f(x) - f(x\|\theta))^2$,
Pearson chi-squared discrepancy	$\Delta_{\text{P}}(\theta) = \sum_x (f(x) - f(x\|\theta))^2/f(x\|\theta)$,
Neyman chi-squared discrepancy	$\Delta_{\text{N}}(\theta) = \sum_x (f(x) - f(x\|\theta))^2/f(x)$.

For a given model, the discrepancy due to approximation is defined as the discrepancy $\Delta(\theta_0)$ of that member P_{θ_0} of the model that minimizes $\Delta(Q,P_\theta)$. It is sometimes also called the model bias, the model error, or the degree of misspecification. For a given model and method of estimating θ, the discrepancy between P_{θ_0} and the estimated distribution $P_{\hat{\theta}}$, $\Delta(P_{\theta_0}, P_{\hat{\theta}})$, is

called the discrepancy due to estimation. It is sometimes also called the estimation variance. Finally, the overall discrepancy is the discrepancy of the underlying distribution and the fitted distribution, $\Delta(Q, P_{\hat{\theta}})$. In many of the cases of interest, the overall discrepancy is the sum of the two component discrepancies, but even when it is not, the overall discrepancy is still a monotone function of both component discrepancies.

It is interesting to consider how the component discrepancies behave as model complexity increases. As the model becomes less constrained, the discrepancy due to approximation will tend to decrease, because the model will be able to approximate the true distribution more closely. On the other hand, the discrepancy due to estimation will simultaneously tend to increase because complex models have a greater propensity to capitalize on chance and to be misled by chance fluctuations in the sample. Thus, the overall discrepancy represents a trade-off between the ability to fit and the propensity to capitalize on error.

In this approach, selection is based on estimated expected discrepancies, where expectation is taken with respect to the underlying distribution. An estimator of the expected discrepancy is called a criterion, and the model that gives rise to the smallest criterion is selected.

The derivation of criteria can be a difficult technical issue. In some contexts, unbiased estimators can be derived in the finite-sample case. For example, in the context of multiple linear regression, the well-known variable selection criteria S_p and C_p can be thus derived as (monotone transformations of) unbiased estimators of discrepancies based on errors of prediction (Linhart and Zuccini, 1986). In many instances, unbiased estimators cannot be derived. In these cases, asymptotic methods, bootstrap methods, and cross-validation techniques can be used.

For example, when the underlying distribution is a member of model M, an asymptotically unbiased criterion based on the Kullback–Leibler discrepancy is in fact AIC (divided by $2n$):

$$\mathrm{AIC} = -2\log f(x_1, \ldots, x_n | \hat{\theta}_n) + 2q$$

where $f(x_1, \ldots, x_n|\theta)$ is the likelihood of the data $x_1, \ldots, x_n$, and estimation is based on the maximum likelihood method. Even if Q is not in M, AIC/$(2n)$ has been argued to remain a satisfactory approximation of the Kullback–Leibler criterion as long as the model is not grossly misspecified (Linhart and Zuccini, 1986).

A small-sample correction of AIC is $\mathrm{AIC_c}$ given by

$$\mathrm{AIC_c} = \mathrm{AIC} + 2q\frac{q+1}{n-q-1}$$

It provides a better approximation of the Kullback–Leibler criterion in certain situations.

Bootstrap Techniques

Bootstrap techniques provide another means of estimating expected discrepancies that are widely applicable. In the bootstrap method, the unknown distribution Q is replaced by Q_n, which assigns probability mass $1/n$ to each observed value x_i, $i = 1, \ldots, n$. This makes it possible to compute expected discrepancies when an explicit formula is available or to use Monte Carlo methods to estimate expected discrepancies to any desired degree of accuracy. Expectation is with respect to Q_n rather than Q, however, and the bootstrap criterion has a downward bias (Chung et al., 1996).

Cross-Validation Techniques

In cross-validation, the data are divided into two subsamples, a calibration sample of size $n - \nu$ and a validation sample of size ν. The first sample is used to estimate the model; the second is used to estimate the expected discrepancy. Cross-validation techniques for model selection use a small ν, typically $\nu = 1$, but repeat the above steps for all possible subdivisions of the sample data into two subsamples of the required sizes. The cross-validation criterion is the average, over these repetitions, of the estimated expected discrepancies. When $\nu = 1$, this method is called one-item out cross-validation. Under many circumstances, the resulting cross-validatory criterion is straightforward to compute and is an estimator of the expected discrepancy for a sample of size $n - \nu$ (Linhart and Zuccini, 1986).

Bayesian Model Selection

Another principled approach to model selection has been proposed within the framework of Bayesian statistics. In this approach, each model M is seen as potentially containing the true probability distribution, and an a priori probability, $P(M)$, for this event is assigned to each model. Based on the data D, the posterior probabilities of the models, $P(M|D)$, are then computed, and the model with the largest posterior probability is selected. Another possibility is to avoid selection of one specific model altogether, and to use instead the posterior distribution, which averages over the different models by weighting each model with its posterior probability. This is called model averaging. In addition to an a priori probability for each model, an a priori probability has to be specified for the distribution of each model's parameters. Computing the posterior probabilities can be a formidable task.

For example, consider the case of just two models M_1 and M_2 and data D. Selection is based on the posterior odds,

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1)\,P(M_1)}{P(D|M_2)\,P(M_2)}$$

such that M_1 is selected if the posterior odds exceed one and M_2 is selected otherwise. The first quotient on the right side of the above equation is called the Bayes factor, and it is equal to the posterior odds ratio, if the prior odds ratio is one. Bayes factors are known to be much more sensitive to the choice of the prior distributions than is Bayesian estimation of parameters of interest (O'Hagan and Forster, 2004; Chapter 7).

Another index is the BIC:

$$\mathrm{BIC} = -2\log f(x_1, \ldots, x_n | \hat{\theta}_n) + q\log(n)$$

where $f(x_1, \ldots, x_n|\theta)$ is the likelihood of the data $x_1, \ldots, x_n$, and estimation is based on the maximum likelihood method. It can be shown that selecting the model with the smaller value of the BIC is asymptotically equivalent to selecting according to the Bayes factor. Thus, the index BIC can be used to obtain a large-sample approximation of the Bayes factor.

The deviance information criterion (DIC) is another Bayesian model-selection index structured like AIC and BIC, in

that it comprises a term quantifying the lack of model fit and a term penalizing model complexity (O'Hagan and Forster, 2004; Chapter 7). The latter term is of interest in its own right, in that it is interpreted as the effective number of parameters. It is smaller than the actual number of parameters to the extent to which the model parameters are constrained by dependencies in the data or the prior. DIC can be conveniently computed on the basis of the output from Markov chain Monte Carlo methods of sampling from the posterior distribution. A point estimate of the parameter estimates is also required.

Model Selection Following the Minimum Description Length Principle

A different idea is pursued by the minimum description length approach. Roughly, a model reduces the complexity of a code (e.g., a string of binary values) needed to describe the data, because only the model and its estimated parameters as well as the residuals have to be encoded, but not the original data, once the model has been fit. In as much as the residuals show less variability than the original data, a good model reduces the code needed to describe the data set. The code length is a function of both the model's complexity and its ability to account for the data (Grünwald, 2007). In the present context, the principle leads to the NML index for model selection. Much like AIC and BIC, the index adds minus the maximum log-likelihood of the data given the model and a penalty term for the model's flexibility. The penalty is given by the logarithm of the sum of maximum likelihood values summed over the entire set of possible data patterns $\gamma = (\gamma_1, \ldots, \gamma_n)$ that might in principle occur in the experimental setting. NML is thus given by (again for continuous random variables, the sum should be replaced by an integral)

$$\mathrm{NML} = -\log f(x_1, \ldots, x_n | \hat{\theta}_n) + \sum_{\gamma} f(\gamma | \hat{\theta}_n(\gamma))$$

where $\hat{\theta}_n(\gamma)$ is the maximum likelihood estimate of the model parameters for data pattern γ. The penalty term, $\sum_{\gamma} f(\gamma | \hat{\theta}_n(\gamma))$, is a measure of the model's ability to fit data in general, whatever the data. The resulting model-selection index is principled and has a couple of desirable properties (Grünwald, 2007). For example, the penalty terms $2q$ and $q \log(n)$ in AIC and BIC, respectively, penalize models based on the number of parameters and (for BIC) sample size, whereas functional form plays no role. However, an additional parameter can have anything between a dramatic influence on a model's flexibility in fitting data in general and a negligible influence, depending upon how the parameter enters into the model equations. For example, constraining a parameter to range only in a certain restricted interval makes no difference in computing the penalty terms in AIC and BIC, but it appropriately leads to a decreased penalty term in NML. Like for BIC, model selection by NML is asymptotically linked to model selection by the Bayes factor (Grünwald, 2007; Chapter 14).

The NML index is often difficult to compute. An asymptotic approximation is the Fisher information approximation given by

$$\mathrm{FIA} = -\log f(x_1, \ldots, x_n | \hat{\theta}_n) + \frac{q}{2} \log \frac{n}{2\pi} + \int \sqrt{\det I(\theta)} \mathrm{d}\theta$$

where I is the so-called Fisher information matrix of the model for a sample of size one. The Fisher information matrix is the matrix of the expected second partial derivatives of the model's log-likelihood function. As can be seen, the index is similar to BIC (when multiplied by two). The term involving the Fisher information matrix can be seen as a measure of model flexibility due to functional form.

Model Selection and Testing: Concluding Remarks

Theoretical results show that BIC and NML identify the correct model with probability converging to one, as the sample size increases, whereas a similar asymptotic consistency is not given for AIC (Jungeilges, 1992). However, most distributions underlying real data are likely to be much more complex than the most complex model considered for selection. Under this premise, all the above selection methods favor the most complex models among the considered models as sample size increases, because estimation variance, but not model bias, converges to zero as n goes to infinity. That is, in large samples, selection is dominated by the degree of misspecification of given models so that nonparsimonious models with more ability to approximate the underlying complex distribution are selected. Similarly, in this situation, any model is eventually rejected in model testing as the test power of model tests for detecting even tiny model violations increases with increasing sample size.

Another class of problems is subsumed under the label model selection bias (Chatfield, 1995). Model selection criteria are themselves random variables and thus subject to random error. As a consequence, which model is selected for further inference or prediction is uncertain, and there may be a substantial probability that another model is chosen if the model selection procedure is repeated on the basis of a new random sample of the data. Note that this is a fundamental problem that also applies to criteria based on cross-validation and to Bayesian model selection.

This has several consequences. The criterion value of the selected model will in general appear too good, because overfitting may have contributed to the model's eventual selection success, and nominal standard errors of parameter estimates and nominal levels of significance tests based on the selected model will be incorrect because they do not take model uncertainty into account. Several approaches are being discussed for dealing with this issue. In Bayesian model selection, model uncertainty can be explicitly incorporated by model averaging (see above). Furthermore, through appropriate bootstrap and cross-validation techniques, model selection can be explicitly built into statistical procedures (Hjorth, 1994).

See also: Computational Approaches to Model Evaluation.

Bibliography

Agresti, A., 1990. Categorical Data Analysis. Wiley, New York.

Atkinson, A.C., 1970. A method for discrimination between models. Journal of the Royal Statistical Society 32, 323–345.

Bayarri, S., Berger, J., 2000. *P*-values for composite null models. Journal of American Statistical Association 95, 1127–1142.

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society 158, 419–466.

Chung, H.-Y., Lee, K.-W., Koo, J-Y., 1996. A note on bootstrap model selection criterion. Statistics and Probability Letters 26, 35–411.

Cox, D.R., 1961. Tests of separate families of hypotheses. Proceedings of the Fourth Berkeley Symposium 1, 105–123.

D'Agostino, R.B., Stephens, M.A., 1986. Goodness-of-Fit Techniques. Marcel, New York.

Grünwald, P., 2007. The Minimum Description Length Principle. MIT Press, Cambridge.

Hjorth, U., 1994. Computer Intensive Statistical Methods. Validation, Model Selection and Bootstrap. Chapman & Hall, London.

Jungeilges, J.A., 1992. The Robustness of Model Selection Rules. Lit, Münster, Germany.

Linhart, H., Zuccini, W., 1986. Model Selection. Wiley, New York.

Moore, D.S., Spruill, M.C., 1975. Unified large-sample theory of general chi-squared statistics for tests of fit. Annals of Statistics 3, 599–616.

O'Hagan, A., Forster, J., 2004. Kendall's advanced theory of statistics. In: Bayesian Inference, vol. 2B. Arnold, England.

Read, T.R.C., Cressie, N.A.C., 1988. Goodness-of-Fit Statistics for Discrete Multivariate Data. Springer, New York.

Schermelleh-Engel, K., Moosbrugger, H., Müller, H., 2003. Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. Methods of Psychological Research Online 8, 23–74.

Models of Influence in Networks

Noah E Friedkin, University of California, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article focuses on dynamic models of endogenous interpersonal influences unfolding in networks, that is, a social process in which the displayed emotional, attitudinal, or behavioral orientations of individuals toward a common object are affected by the displayed orientations of other individuals toward the same object. A large empirical literature on interpersonal influence exists in the social sciences, which presents particular useful constraints on the formal architecture of this class of dynamic models. Research on these dynamic models is an accelerating interdisciplinary enterprise that includes contributions from the social, natural, and engineering sciences. The article outlines the core constructs and measurement models of this field, and details its engagement with the postulate of an influence process based on convex combinations of individuals' heterogeneous orientations toward common objects of their social environments.

Introduction

This article focuses on dynamic models of endogenous interpersonal influences unfolding in networks, that is, a social process in which the displayed emotional, attitudinal, or behavioral orientations of individuals toward a common object are affected by the displayed orientations of other individuals toward the same object. Because the domain of such objects is large, dynamic models of influence in networks have ubiquitous relevance in the social sciences. The object may be a particular issue, event, behavior, individual, commodity, institution, regime, political party, ethnic group, tribe, religion, or nation (i.e., any common focus of attention and arousal in a population of $n \geq 2$ individuals). The existence of endogenous interpersonal influences is a core postulate of social psychology. The postulate allows for individual differences of openness or closure to interpersonal influence.

The elementary unit of endogenous interpersonal influence is a *direct* influence relation for an ordered pair of individuals (i,j) in which individual j's emotional, attitudinal, or behavioral orientation toward a particular object is visible and salient to individual i. An invisible orientation cannot be directly influential, and a visible orientation may be discounted. The influence relation may be represented by the directed line $i \leftarrow j$, which indicates that j's orientation has a direct influence on i's orientation, or by $i \rightarrow j$, which indicates that j's orientation is accorded influence by i. This representation includes loops, $i \rightarrow i$, which indicate that i accords some weight to his or her own orientation.

With respect to direct influence relations, the core unsettled problem in this field is the discovery of a suitable formalization of the 'within-the-skin' mechanism – the cognitive algebra – by which each individual synthesizes the direct, heterogeneous orientations of other individuals, as well as his or her own orientation, toward an object. Some such synthesis must occur when an individual's orientation has been influenced by the orientations of others. Is the modified orientation a linear or nonlinear function of these orientations? Is it a convex combination? How does the individual's own orientation enter into the synthesis, and, in particular, when and how is the individual's initial orientation involved? Is it a first-order or higher order difference function? The specification remains an open question, although progress has been made on it. The available evidence comes from cognitive science (e.g., Anderson, 1981, 1991) and from empirical assessments of the predictions of dynamic models that incorporate particular assumptions about the cognitive algebra. The evidence points to a convex combination of orientations.

The more complex unit of endogenous interpersonal influence is a *total* influence relation for each ordered pair of individuals (i,j) in which individual j's orientation has influenced i's orientation on the basis of the direct and indirect flows of influence in a network. These flows are generated by the within-the-skin mechanism with which each member of an influence network cognitively synthesizes heterogeneous orientations. As individuals modify their orientations, each individual may be located in a changing landscape of orientations, and repetitive responses to these orientations enable indirect influences. The formal architecture of a model of influence in networks may or may not allow solutions of the total influence for each (i,j) ordered pair of an influence system.

This article concentrates on a class of dynamic models of influence in networks that allows the quantification of total influences on orientations toward an object, where the orientations are values in a defined interval on the real number line, and might be any number in this interval. Such orientations may be multidimensional. They also may be periodically displayed as categorical orientations that are epiphenomenal manifestations of individuals' real, more nuanced orientations toward an object.

Constraints on Mathematical Imagination

A massive empirical literature on interpersonal influence exists in the social sciences, from which findings may be extracted that have an important bearing on the formal architecture of models of influence in networks. Friedkin and Johnsen (2011) provide a portal into the social-psychological literature on group dynamics. Formal models of influence in networks vary in their attention to this literature and the extent to which their simplifying assumptions are constrained by robust generic

observations of the developed and developing lines of empirical work on social influence. Such empirical observations do exist, and three are highlighted; for their more detailed treatment, and access to the empirical studies upon which they are based, see Friedkin and Johnsen (2011).

First, individuals' extents of openness versus closure to interpersonal influence on a particular issue are heterogeneous. There is no universal constant of susceptibility to interpersonal influence. Heterogeneity extends to individuals' accorded influences: an individual is rarely accorded equal influence by different individuals, and different individuals are rarely accorded equal influence by the same individual. Individual differences are nonignorable and are based on numerous individual, relational, and contextual variables that affect individuals' openness to interpersonal influence and the amount of influence accorded to particular persons. Endogenous interpersonal influences involve *embodied* orientations, i.e., orientations toward an object that are displayed by particular individuals, and perforce are confounded with diverse features of not just the individuals who are conveying their orientations but also their interpersonal relationships. As a consequence, the basis of a direct influence relation, $i \rightarrow j$, in which i accords some weight to j's orientation, is complex, and simplifying homogeneity assumptions about the values of these relations are not credible. Empirical investigation of a model of influence in networks of size $n > 2$ poses a generalized three-body problem that requires direct measures of the network members' extents of openness to interpersonal influence and their accorded influences to other individuals.

Second, an important robust finding pertains to the formally attractive postulate that the value of the $i \rightarrow j$ accorded influence at time t may be simplified to a function of the time t orientations of those individuals whose orientations are visible to individual i. This postulate is attractive because, with it, a group's vector of initial orientations may be employed to specify the group's initial influence network and its temporal change, as orientations are modified, within a fixed structure of interpersonal visibility. Festinger's (1954) social comparison theory is a classic assertion of this postulate, and numerous models of influence in networks incorporate it. The accumulated empirical evidence does not support its employment as a reliable specification of influence networks. The mixed empirical findings in the literature on this specification are consistent with the conclusion that each direct influence relation, $i \rightarrow j$, in which i accords some weight to j's orientation, is the resultant of an idiosyncratic mixture of bases.

Third, interpersonal influences in networks generate various generic distributions of orientations. The emergent distribution may be a consensus, or it may be an influenced array of positions with unresolved disagreements. Both outcomes occur with noteworthy frequencies, even in *strongly connected* influence networks (i.e., networks in which one or more paths exist from i to j, and from j to i, for all i and j). A formal model of influence in networks must be consistent with both. Many proposed models are not. Hence, Abelson's (1964: 153) famous complaint: "Since universal ultimate agreement is [a] ubiquitous outcome of a very broad class of mathematical models, we are naturally led to inquire what on earth one must assume in order to generate the bimodal outcomes of community cleavage studies". When the outcome is consensus, it may be a settlement on an initial position of one of the members of a group, including the minimum or maximum (most extreme) initial position of the group members, or on a compromise position that is not the initial position of any group member. While the relative frequency of each type of consensus outcome varies depending on the experimental design, each type has been observed with noteworthy frequency in samples of small groups. Hence, formal models that exclude these frequently occurring generic outcomes have limited applicability.

Finally, and remarkably, individuals' emergent orientations are rarely more extreme than the minimum or maximum value of group members' initial orientations. In the special circumstance of an initial consensus, the consensus position is maintained unaltered by groups: the issue is considered settled, and the group will move on to the next issue when a sequence of issues is being dealt with. This robust finding applies to judgmental issues with no correct issue position, and to intellective issues with a correct position that may or may not be apprehended. On intellective issues, groups rarely question and alter an incorrect initial consensus. On both types of issues, in the absence of an initial consensus, discussions of an issue rarely generate positions that are outside the range of the initial positions of the group members. If this robust finding extends to multidimensional issues, then the general rule is that emergent influenced positions are constrained to the *convex hull* of a group's initial positions. Observed violations of this rule are rare and strongly support its employment as a simplifying formal constraint. This constraint is consistent with the assumption of a cognitive algebra based on a convex combination of orientations.

Influence Networks

The elementary units of endogenous interpersonal influence form an influence network, that is, ordered pairs of individuals (i,j) in which individual j's orientation toward a particular object is visible and salient to individual i, when such pairs are members of the set of ordered pairs defined by a population of $n \geq 2$ individuals with orientations toward a common object. Let $G(V,E)$ denote this network, where V is the set of n individuals, and E is the set of ordered pairs with a directed line $i \rightarrow j$ indicating that i accords some influence to j's displayed orientation toward the common object. This network may be described as an $n \times n$ matrix $\mathbf{G} = [g_{ij}]$ of values, with $g_{ij} = 0$ indicating the absence of an $i \rightarrow j$ relation, and $g_{ij} \neq 0$ indicating its existence and value, for each (i,j) ordered pair.

Analysis may focus on this global common-object network, or on one or more of its weakly connected components when multiple components exist. If G is transformed into an undirected graph $G \rightarrow \tilde{G}$, in which $i - j$ exists if and only if either $i \rightarrow j$ or $j \rightarrow i$ exists, then a weak component of G is a maximal connected subgraph of $\tilde{G}$. A weakly connected component is the minimal criterion of a *coherent* influence network within a population oriented to a common object; in other words, such a network does not exclude any individual who is accorded influence by one or more of its members, nor does it exclude

any individual who accords influence to one or more of its members, with respect to the common object. In particular, if i is included in the network and the accorded influences $i \rightarrow k \leftarrow j$ exist, then both k and j also must be included in the network. The collection of these disconnected components is the global common-object network. Note that each weak component may include individuals who are closed to interpersonal influence but accorded influence by at least one member of the component. Also note that a component, which satisfies the condition of weak connectivity, may be a component with a 'higher' connectivity category: it will be a strongly connected component if one or more paths exist from i to j, and from j to i, for each of its (i,j) unordered pairs; and it will be a unilaterally connected component if one or more paths exist from i to j, or from j to i, for each of its (i,j) unordered pairs (these standard definitions of a network's connectivity category (strong, unilateral, or weak) are such that strong $\Rightarrow$ unilateral $\Rightarrow$ weak connectivity).

The influence network $G(V,E)$ is a valued network for which homogeneity assumptions on these values are suspect. Homogeneity assumptions include constraining the main diagonal values of $\mathbf{G}$ to zero; constraining nonzero values of each row of $\mathbf{G}$ to the same value, which may vary from row to row; and constraining the nonzero values of each column of $\mathbf{G}$ to the same value, which may vary from column to column. The first is suspect because it assumes that all individuals accord no weight to their own orientations. The second is suspect because it assumes that all individuals do not discriminate among the orientations of other persons to whom they accord some influence. The third is suspect because it assumes that all individuals who accord influence to a particular individual will accord the same amount of influence to that person.

An important implication of a formal model that relaxes homogeneity assumptions is its demand for measurement models of $\mathbf{G}$ that are commensurate with the possibility of substantial heterogeneity. An outstanding problem is the lack of research on such measurement models. Investigations often employ measurement models of influence networks that draw on structural features of interpersonal contact networks. The availability of contact data is a convenience, which leaves open the problem of developing a measurement model that allows for heterogeneous accorded influences, including a measure of the extent to which each individual is open or closed to interpersonal influence on his or her orientations toward a particular object. Although regular contact is an important basis of interpersonal influence, it is not a necessary or sufficient condition of the interpersonal visibility and salience upon which influence is based. Currently, the most direct approach to a measurement model, and the approach that most closely corresponds to the cognitive foundations of accorded influence, is one that draws on individuals' subjective self-reported accorded weights to self and particular others. A detailed assessment of this measurement approach appears in Friedkin and Johnsen (2011), where 450 issue-specific influence networks are examined. As the implications of dealing with influence network structures are fleshed out, the measurement models of these structures will surely become an increasing focus of attention and refinement.

During some temporal period, $t = 1,2,\ldots,k$, the influence network $G(V,E)$ may be a fixed or changing construct with respect to V, E, or both. The network may be fixed by strongly constraining antecedent conditions, or subject to random disturbances, or a structure that evolves in a systematic way during the period. It may be a structure that is quickly constructed into a stable form by individuals engaged with a specific object. It may be a structure that applies to different objects, or an object-specific structure. Models of influence in networks are partitioned into two broad classes: those that assume a fixed network and those that do not. Models that allow structural changes of the influence network may or may not be researchable, as they present more or less demanding challenges for investigators interested in securing empirical supports for such models.

As Odysseus was forced to choose between Scylla and Charybdis, and chose the less challenging option (Scylla), most lines of research on models of influence have avoided dealing with the advancement of formalizations of an influence process that unfolds in a changing network, and have focused on the advancement of formalizations of an influence process that unfolds in a fixed network (these fixed-structure formalizations include models in which the edges of G are active or inactive with static probabilities). Even so, work on models that allow and formalize network evolution is accelerating. The work on structural change is a large leap into the unknown, down the middle of the Strait of Messina, which simultaneously grapples with Scylla and Charybdis. The approach necessarily combines (1) a formalization of the mechanism by which each individual cognitively synthesizes orientations of other individuals, as well as his or her orientation, toward an object; and (2) ancillary formalizations that temporally modify the network in which the influence mechanism unfolds. The assumption of a fixed network reduces the number of a model's moving parts, and usefully serves to concentrate attention on the fundamental predicates of the influence mechanism. The assumption requires a thoughtful selection of field settings and experimental designs in which it is not grossly violated.

The literature on influence networks presents a remarkable interdisciplinary convergence to a specification of $\mathbf{G}$ in which $0 \leq g_{ij} \leq 1$ for all i and j, and $\sum_{j=1}^{n} g_{ij} = 1$ for all i. This important special case of $\mathbf{G}$ will be denoted as $\mathbf{W}$. The accorded values $\{w_{i1}, w_{i2}, \ldots, w_{in}\}$ for each i are the direct relative weights accorded by i to the orientation of each member of the network. If $w_{ii} = 1$, then the individual is completely closed to interpersonal influence, $\sum_{j \neq i}^{n} w_{ij} = 0$. If $w_{ii} = 0$, then the individual is completely open to interpersonal influence, $\sum_{j \neq i}^{n} w_{ij} = 1$. A subset of these interpersonal weights will be 0 in cases of invisible orientations, and others will be 0 in cases of visible but completely discounted orientations. The importance of this specification is indicated by the consequence of a loss of confidence in it: if confidence in it were substantially eroded, then a large number of models of influence in networks would be set aside.

Direct measures of the accorded influences entailed in $\mathbf{W}$ have been based on individuals' subjective self-reported allocations of weights to self and particular others on a specific issue. Proxy measures based on contact networks, which allow heterogeneity and do not restrict direct influence to contact relations, have been based on measures of structural centrality to specify individuals' extents of openness or closure to interpersonal influence, and on measures of structural proximity to

specify the interpersonal accorded weights for each individual (e.g., Friedkin, 1998, 2001). In such cases, when the investigator's measurement model of susceptibility is different from the measurement model for accorded interpersonal influence, the two measurement models may be combined to form **W** as follows:

$$\mathbf{W} = \mathbf{AC} + \mathbf{I} - \mathbf{A} \qquad [1]$$

where $\mathbf{A} = \mathrm{diag}(a_{11}, a_{22}, \ldots, a_{nn})$, with $0 \le a_{ii} \le 1$ for all i, is the measure of an individual's openness to interpersonal influence; and $\mathbf{C} = [c_{ij}]$, with $0 \le c_{ij} \le 1$ for all i and j, and $c_{ii} = 0$ for all i is the measure of the relative interpersonal weights accorded to others. Here, $\sum_j^n c_{ij} = 1$ if $a_{ii} > 0$, and $\sum_j^n c_{ij} = 0$ if $a_{ii} = 0$:

$$w_{ij} = \begin{cases} 1 - a_{ii} & \text{for } i = j \\ 0 & \text{for } i \neq j, \text{ with } \sum_j^n c_{ij} = 0, \text{ if } a_{ii} = 0 \\ a_{ii}c_{ij} + 1 - a_{ii} & \text{for } i \neq j, \text{ with } \sum_j^n c_{ij} = 1, \text{ if } a_{ii} > 0 \end{cases} \qquad [2]$$

The decomposition assumes $a_{ii} = 1 - w_{ii} = \sum_{j \neq i}^n w_{ij}$, that is, an individual's extent of openness to interpersonal influences is equivalent to the aggregate interpersonal weight that the individual accords to others.

Formalization of the Influence Mechanism

Models of influence in networks often bypass the complex problem of specifying the antecedents of direct influence relations and, instead, focus on the advancement of suitable formalizations of the process of interpersonal influence. Two specifications of the process are prominent and distinguished by their focus on either nominal (i.e., categorical) orientations or real (i.e., quantitative) orientations that allow all values of an interval on the real number line. For nominal orientations, a 'threshold mechanism' is postulated in which an individual's selection among mutually exclusive (typically binary) options is displayed when the influenced value of a quantitative variable (usually conceptualized as a pressure or inclination to manifest a particular choice or behavior) reaches or exceeds a threshold value (Granovetter, 1978; Watts and Dodds, 2007). For real orientations, an 'averaging mechanism' is postulated in which an individual's influenced orientation is a weighted average of other individuals' orientations, as well as the individual's own orientation (DeGroot, 1974; Friedkin and Johnsen, 1990, 1999, 2011).

The theoretical distinction between these two mechanisms is narrowed when individuals' nominal displays are epiphenomenal manifestations of real orientations formed by a weighted averaging mechanism. Individuals with the same nominal choice may vary markedly in the strength of their orientations; individuals with different nominal choices may include more or less opposing orientations. Threshold and averaging mechanisms are not inconsistent, and both may be maintained. However, if real orientations are always implicated in nominal choices and the latter are inexact indicators of real orientations, then the general versus restricted accuracy of a threshold mechanism may be questioned. The postulate of ubiquitous real orientations is importantly supported by the work of Bargh and Ferguson (2000) and others on the automaticity of attitudes. The enactment of a behavior is constrained by various circumstances; behavioral inclinations, to do or not do something, may or may not be manifested. The predicate of an influence of action-upon-action is not universally applicable. In contrast, attitudinal orientations to objects appear to be an automatic response to objects, and their displays may be taken as a general predicate of formal models of endogenous interpersonal influences in networks. Models of endogenous interpersonal influences on attitudinal orientations toward objects leave open the elaboration of the linkages of attitudes with individuals' emotions and actions toward specific objects. With respect to the attitude–behavior linkage, attitudes may be included as an antecedent condition of the probability of an action among other antecedent conditions relevant to the explanation of the particular behavior (Friedkin, 2010).

The field presents numerous formalizations of the mechanism by which individuals cognitively synthesize the orientations of others toward an object, as well as own orientation, during the temporal process of interpersonal influence. This author's assumptions are that there is one to-be-discovered mechanism operating on attitudes by which the mind accomplishes the synthesis, and that this mechanism can formalized. Treating the mechanism as a rational calculation is suspect, when neither the orientations that are synthesized nor the accorded influences that are implicated in it may rest on rational calculations. Although the slow hard work of rational calculation may be done by some individuals, the quick work of an automatic natural heuristic mechanism appears to be a more generally accurate basis on which to construct models of influence in networks. With the accumulating findings of the cognitive revolution in social psychology, indicating ubiquitous heuristic mechanisms, it has become increasingly anomalous to assume rational calculation as the main basis of interpersonal influence systems (Kahneman, 2011; Kahneman et al., 1982).

For a fixed influence network, models of influence in the network necessarily involve an additional assumption about the *sequence* of interpersonal influences. Three simplifying sequence specifications are that (1) the individuals randomly broadcast their orientations in a temporal sequence, (2) dyads are randomly activated in a temporal sequence, and (3) all individuals are simultaneously directly influenced by all of their visible and salient orientations at each time t in a temporal sequence.

This article concentrates on the third class of sequencing formalizations, consistent with the discrete time linear model:

$$\mathbf{Y}^{(t+1)} = \mathbf{AWY}^{(t)} + \mathbf{BY}^{(1)}, \quad t = 1, 2, \ldots \qquad [3]$$

where $\mathbf{A} = \mathrm{diag}\ (a_{11}, a_{22}, \ldots, a_{nn})$, with $0 \le a_{ii} \le 1$ for each i; $\mathbf{B} = \mathrm{diag}\ (b_{11}, b_{22}, \ldots, b_{nn})$; $\mathbf{Y}^{(t)} = [y_{ik}^{(t)}]$ is an $n \times m$ matrix of values, which allows all values of an interval on the real number line; and k is the kth dimension of the time t orientation of each individual toward a particular object. The formalization permits solutions of total influence for each ordered pair of an influence system (i.e., the net direct and indirect influence of j on i, for each (i,j) ordered pair). With it, various theorems of linear algebra are applicable. Notably, for suitable **W**, $\lim_{k \to \infty} \mathbf{W}^k = \mathbf{W}^\infty$ exists, and for suitable **AW**, $\lim_{k \to \infty} [\mathbf{I} + \mathbf{AW} + (\mathbf{AW})^2 + (\mathbf{AW})^3 + \ldots + (\mathbf{AW})^k] = (\mathbf{I} - \mathbf{AW})^{-1}$

exists. Equilibrium orientations of the individuals, $\mathbf{Y}^{(\infty)}$, may or may not be attained, depending on the conditions under which the influence process unfolds. The formal architecture may be elaborated with a specification of the antecedents of the initial orientations of individuals $\mathbf{Y}^{(1)} \leftarrow \mathbf{X}$. It may also be elaborated with a specification of nominal (categorical) manifestations of the orientations of individuals, $\ddot{\mathbf{Y}}^{(t+1)} \leftarrow \mathbf{Y}^{(t+1)}$, where the special categorical nature of such manifestations are denoted by the 'double-dotted' $\mathbf{Y}^{(t+1)}$.

Not all special cases of this linear model are consistent with the constraint that emergent $\mathbf{Y}^{(t+1)}$ is in the convex hull of $\mathbf{Y}^{(1)}$. A seminal special case, consistent with this constraint, is obtained with $\mathbf{A} = \mathbf{I}$ and $\mathbf{B} = \mathbf{0}$:

$$\mathbf{Y}^{(t+1)} = \mathbf{W}\mathbf{Y}^{(t)}, t = 1, 2, \ldots .$$

The model is postulated by French (1956), Harary (1959), and DeGroot (1974). The within-the-skin mechanism is

$$y_{ik}^{(t+1)} = \sum_{j=1}^{n} w_{ij} y_{jk}^{(t)}$$

for the kth dimension of the orientation. The mechanism is a memoryless weighted averaging of the orientations that are accorded influence. The self-weight w_{ii} is the extent to which an individual is open or closed to interpersonal influence. The aggregate interpersonal weight, $1 - w_{ii} = \sum_{j \neq i} w_{ij}$, is distributed in the mind of i to other individuals. The equilibrium state of the system is $\mathbf{Y}^{(\infty)} = \mathbf{W}\mathbf{Y}^{(\infty)}$. Not all $\mathbf{W}$ allow equilibrium. Equilibrium is not attainable, for example, in an $n = 2$ influence system with $w_{11} = w_{22} = 0$ and $y_{1k}^{(1)} \neq y_{2k}^{(1)}$. The input–output formalization, obtained from the recursion, is $\mathbf{Y}^{(\infty)} = \mathbf{V}\mathbf{Y}^{(1)}$ when $\mathbf{V} \equiv \lim_{k \to \infty} \mathbf{W}^k = \mathbf{W}^{\infty}$ exists. The emergent values, $\mathbf{V} = [v_{ij}], 0 \leq v_{ij} \leq 1, \sum_{j=1}^{n} v_{ij} = 1$, are the total relative influences of each individual j on each member i of the influence system. Thus, the equilibrium orientation of each individual is a weighted average of the members' initial orientations, $y_{ik}^{(\infty)} = \sum_{j=1}^{n} v_{ij} y_{jk}^{(1)}$, on each dimension, where the weights are the total influences. The v_{ij} are not, of course, in the mind; their values are collective nontransparent resultants of the influence process unfolding in a network.

While the model allows heterogeneous influences and constrains influenced orientations to the convex hull of initial orientations, it presents a highly restricted set of conditions under which equilibrium disagreement will occur. In particular, in the domain of strongly connected influence networks, the generic form of $\mathbf{V}$, if it exists, is

$$\mathbf{V} = \begin{bmatrix} z_1 & z_2 & \cdots & z_n \\ z_1 & z_2 & \cdots & z_n \\ \vdots & \vdots & \ddots & z_n \\ z_1 & z_2 & \cdots & z_n \end{bmatrix}, \quad 0 \leq z_i \leq 1, \sum_{j=1}^{n} z_j = 1,$$

$$i = 1, 2, \ldots, n \qquad [4]$$

and the model's prediction is a consensus matrix $\mathbf{Y}^{(\infty)}$, with respect to the values of each of its one or more dimensions. In this case, an analytical solution for $\mathbf{z} = [z_1 \quad z_2 \quad \ldots \quad z_n]'$ is available from the left eigenvector of $\mathbf{W}$ for eigenvalue 1, that is, $\mathbf{x} = \mathbf{W}'\mathbf{x}$, whence $z_i = x_i / \sum_k^n x_k$.

The restrictive formal properties of this model motivated Friedkin and Johnsen's (1990, 1999, 2011) generalizations. With the constraint $\mathbf{B} = \mathbf{I} - \mathbf{A}$, the Friedkin–Johnsen model for a fixed influence network is

$$\mathbf{Y}^{(t+1)} = \mathbf{A}\mathbf{W}\mathbf{Y}^{(t)} + (\mathbf{I} - \mathbf{A})\mathbf{Y}^{(1)}, \quad t = 1, 2, \ldots . \qquad [5]$$

Clearly, with $\mathbf{A} = \mathbf{I}$, this model subsumes the French–Harary–DeGroot model as a special case. The formalized within-the-skin mechanism is

$$y_{ik}^{(t+1)} = a_{ii} \sum_{j=1}^{n} w_{ij} y_{jk}^{(t)} + (1 - a_{ii}) y_{ik}^{(1)} \qquad [6]$$

for the kth dimension of the orientation. The cognitive algebra remains a weighted averaging of the orientations that are accorded influence, but now it is a mechanism in which the initial orientation may not be 'forgotten': it disappears as a component of the synthesis if $a_{ii} = 1$, and remains a component otherwise. When the antecedent bases of an initial orientation include experiences, values, beliefs, and norms, the initial orientation toward an object may have enduring salience. The individual-level value $0 \leq a_{ii} \leq 1$ describes the extent to which an individual is closed or open to interpersonal influence. The *standard* form of this mechanism assumes $a_{ii} = 1 - w_{ii}$, for all i. With this constraint, the weight accorded to the initial orientation $1 - a_{ii}$ is equivalent to self-weight w_{ii}, and to the extent to which an individual is closed to interpersonal influence. The extent to which an individual is open to interpersonal influence (i.e., the aggregate interpersonal weight, $a_{ii} = 1 - w_{ii} = \sum_{j \neq i} w_{ij}$) is distributed in the mind of i to one or more other individuals.

A more general formalization of total influences than $\mathbf{V} \equiv \mathbf{W}^{\infty}$ is available with this mechanism. The partial and total influences emerge as follows:

$$\mathbf{Y}^{(2)} = \mathbf{A}\mathbf{W}\mathbf{Y}^{(1)} + (\mathbf{I} - \mathbf{A})\mathbf{Y}^{(1)} = \mathbf{V}^{(1)}\mathbf{Y}^{(1)}$$

where

$$\mathbf{V}^{(1)} = \mathbf{A}\mathbf{W} + (\mathbf{I} - \mathbf{A})$$

$$\mathbf{Y}^{(3)} = \mathbf{A}\mathbf{W}\mathbf{Y}^{(2)} + (\mathbf{I} - \mathbf{A})\mathbf{Y}^{(1)} = \mathbf{V}^{(2)}\mathbf{Y}^{(1)}$$

where

$$\mathbf{V}^{(2)} = (\mathbf{A}\mathbf{W})^2 + (\mathbf{A}\mathbf{W} + \mathbf{I})(\mathbf{I} - \mathbf{A}) = \mathbf{A}\mathbf{W}\mathbf{V}^{(1)} + (\mathbf{I} - \mathbf{A})$$

$$\mathbf{Y}^{(4)} = \mathbf{A}\mathbf{W}\mathbf{Y}^{(3)} + (\mathbf{I} - \mathbf{A})\mathbf{Y}^{(1)} = \mathbf{V}^{(3)}\mathbf{Y}^{(1)}$$

where

$$\begin{aligned} \mathbf{V}^{(3)} &= (\mathbf{A}\mathbf{W})^3 + [(\mathbf{A}\mathbf{W})^2 + \mathbf{A}\mathbf{W} + \mathbf{I}](\mathbf{I} - \mathbf{A}) \\ &= \mathbf{A}\mathbf{W}\mathbf{V}^{(2)} + (\mathbf{I} - \mathbf{A}) \end{aligned}$$

and so on. Note that the sequence $\mathbf{V}^{(1)}$, $\mathbf{V}^{(2)}$, $\mathbf{V}^{(3)}$, …, may be obtained by specifying the starting matrix $\mathbf{V}^{(1)} = \mathbf{A}\mathbf{W} + (\mathbf{I} - \mathbf{A})$ and recursively applying the formula $\mathbf{V}^{(t+1)} = \mathbf{A}\mathbf{W}\mathbf{V}^{(t)} + (\mathbf{I} - \mathbf{A})$ for $t = 1, 2, \ldots$ When $\mathbf{V} \equiv \lim_{t \to \infty} \mathbf{V}^{(t)}$ exists, the equilibrium matrix equations of the process are

$$\begin{aligned} \mathbf{Y}^{(\infty)} &= \mathbf{A}\mathbf{W}\mathbf{Y}^{(\infty)} + (\mathbf{I} - \mathbf{A})\mathbf{Y}^{(1)} \\ &= \mathbf{V}\mathbf{Y}^{(1)} \end{aligned} \qquad [7]$$

and we have for each group member i:

$$y_{ik}^{(\infty)} = \sum_{j=1}^{n} v_{ij} y_{jk}^{(1)} \qquad [8]$$

for the kth dimension of the orientation, where the values of the emergent $\mathbf{V} = [v_{ij}], 0 \le v_{ij} \le 1, \sum_{j=1}^{n} v_{ij} = 1$ are the relative total influences of each individual j on each member i of the influence system. The equilibrium orientation of each individual is a weighted average of the members' initial orientations. For $\mathbf{A} = \mathbf{I}$, this limit is $\mathbf{V} = \mathbf{W}^{\infty}$. For a nonsingular $\mathbf{I} - \mathbf{AW}$, it is $\mathbf{V} = (\mathbf{I} - \mathbf{AW})^{-1}(\mathbf{I} - \mathbf{A})$. In general, with

$$\mathbf{V}^{(1)} = \mathbf{AW} + (\mathbf{I} - \mathbf{A}) \qquad [9]$$

$$\mathbf{V}^{(t+1)} = \mathbf{AWV}^{(t)} + (\mathbf{I} - \mathbf{A}), \quad t = 1, 2, \ldots,$$

$\mathbf{V}$ is defined as the limit of the above recursion, when the limit exists.

It is useful to consider a prominent specification consistent with $\mathbf{Y}^{(t+1)} = \mathbf{AWY}^{(t)} + \mathbf{BY}^{(1)}$, which does not constrain $\mathbf{Y}^{(t+1)}$ to the convex hull of $\mathbf{Y}^{(1)}$. With the constraints $\mathbf{A} = \alpha\mathbf{I}$ and $\mathbf{B} = \mathbf{I}$, we have

$$\mathbf{Y}^{(t+1)} = \alpha\mathbf{WY}^{(t)} + \mathbf{Y}^{(1)}, \quad t = 1, 2, \ldots \qquad [10]$$

that is, a familiar discrete time dynamic model. With $0 < \alpha < 1$, the equilibrium state of the system is $\mathbf{Y}^{(\infty)} = \alpha\mathbf{WY}^{(\infty)} + \mathbf{Y}^{(1)}$, and its input–output formalization is $\mathbf{Y}^{(\infty)} = (\mathbf{I} - \alpha\mathbf{W})^{-1}\mathbf{Y}^{(1)}$ (for a one-dimensional orientation, the elaboration of the model with $\mathbf{y}^{(1)} = \mathbf{X\beta} + \mathbf{u}$ returns Ord's (1975) formalization and associated statistical approach to estimating the two unmeasured constructs $\{\alpha, \boldsymbol{\beta}\}$). This formalization does not necessarily constrain $\mathbf{Y}^{(\infty)}$ to the convex hull of $\mathbf{Y}^{(1)}$. The Friedkin–Johnsen formalization may be brought close to this model, $\mathbf{Y}^{(t+1)} = \alpha\mathbf{WY}^{(t)} + (1 - \alpha)\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(\infty)} = (\mathbf{I} - \alpha\mathbf{W})^{-1}(1 - \alpha)\mathbf{Y}^{(1)}$, so that the equilibrium predictions of the two models differ only by the factor $(1 - \alpha)$. The factor constrains $\mathbf{Y}^{(\infty)}$ to the convex hull of $\mathbf{Y}^{(1)}$. However, the $\mathbf{A} = \alpha\mathbf{I}$ homogeneity assumption is suspect.

Lastly, consider the constraint $\mathbf{B} = \beta(\mathbf{I} - \mathbf{A})$, with which $\mathbf{Y}^{(\infty)} = \beta\mathbf{VY}^{(1)}$, where $\mathbf{V}$ is the Friedkin–Johnsen $\mathbf{V}$. This generalization of the Friedkin–Johnsen model allows for breaches of the convex hull of $\mathbf{Y}^{(1)}$ with a scalar $\beta \neq 1$ that inflates or deflates the values of all orientations (Friedkin, 1999; Friedkin and Johnsen, 1990). This and the other special cases of $\mathbf{Y}^{(t+1)} = \mathbf{AWY}^{(t)} + \mathbf{BY}^{(1)}$, described in this article, are not exhaustive.

Illustration

Most formalizations of endogenous interpersonal influences deal with a one-dimensional $\mathbf{Y}^{(t)}$. The availability of multidimensional $\mathbf{Y}^{(t)}$ has not been exploited. To illustrate, consider persons who are prioritizing m alternatives or planning to distribute an amount of some resource over m alternatives, where each person's orientation is represented by the relative weight that the person accords to a particular alternative or by the proportion of available resources (e.g., time and funds) that the person believes should be allocated to each of the alternatives. Here, $\mathbf{Y}^{(t)} = [y_{ik}^{(t)}]$ is an $n \times m$ weight matrix with $0 \le y_{ik}^{(t)} \le 1$ for each person $i = 1, 2, \ldots, n$ and each alternative $k = 1, 2, \ldots, m$, such that $\sum_{k=1}^{m} y_{ik}^{(t)} = 1$ for all i. For such a $\mathbf{Y}^{(t)}$, a non-obvious property of the influence process specified by Friedkin and Johnsen is its preservation of the form of the issue (i.e., the weights assigned by each person to the m alternatives will sum to one at each time t). In general, from $\mathbf{Y}^{(t+1)} = \mathbf{V}^{(t)}\mathbf{Y}^{(1)}$, $\mathbf{Y}^{(t+1)}$ must be row-stochastic if $\mathbf{Y}^{(1)}$ is row-stochastic. Depending on the application, $\mathbf{Y}^{(t)}$ may be periodically displayed as categorical choices (e.g., persons may choose their most preferred alternatives):

$$\ddot{y}_{ik}^{(t)} = \begin{cases} 0 & \text{if } y_{ik}^{(t)} < \max(y_{i1}^{(t)}, y_{i2}^{(t)}, \ldots, y_{im}^{(t)}) \\ 1 & \text{if } y_{ik}^{(t)} = \max(y_{i1}^{(t)}, y_{i2}^{(t)}, \ldots, y_{im}^{(t)}) \end{cases} \qquad [11]$$

for all i at each time $t = 1, 2, 3, \ldots$, where this maximum value may occur for more than one alternative.

To make the illustration concrete, consider a group of four persons who are discussing three alternatives {A, B, C} and periodically displaying their most preferred alternative(s). Suppose that the displayed initial positions of the persons on this issue are

$$\mathbf{Y}^{(1)} = \begin{array}{c} \begin{array}{ccc} \text{A} & \text{B} & \text{C} \end{array} \\ \begin{bmatrix} .20 & .20 & .60 \\ .50 & .25 & .25 \\ .25 & .60 & .15 \\ 0 & 0 & 1 \end{bmatrix} \end{array} \qquad \ddot{\mathbf{Y}}^{(1)} = \begin{array}{c} \begin{array}{ccc} \text{A} & \text{B} & \text{C} \end{array} \\ \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

Person 1 accords .20 weight to the first two alternatives and .60 weight to the third alternative, person 2 accords .50 weight to the first alternative and .25 weight to the second and third alternatives, and so on. Interpersonal influences may alter these orientations and, depending on the influence network, produce consensus. If the influence network of the group is

$$\mathbf{W} = \begin{bmatrix} .500 & .250 & 0 & .250 \\ .100 & .100 & .200 & .600 \\ .100 & .200 & .600 & .100 \\ 0 & .300 & 0 & .700 \end{bmatrix}$$

then, from $\mathbf{Y}^{(t+1)} = \mathbf{AWY}^{(t)} + (\mathbf{I} - \mathbf{A})\mathbf{Y}^{(1)}$, with $a_{ii} = 1 - w_{ii}$ for all i,

$$\mathbf{Y}^{(2)} = \begin{array}{c} \begin{array}{ccc} \text{A} & \text{B} & \text{C} \end{array} \\ \begin{bmatrix} .213 & .181 & .606 \\ .158 & .174 & .669 \\ .258 & .532 & .210 \\ .045 & .023 & .933 \end{bmatrix} \end{array} \qquad \ddot{\mathbf{Y}}^{(1)} = \begin{array}{c} \begin{array}{ccc} \text{A} & \text{B} & \text{C} \end{array} \\ \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

$$\mathbf{Y}^{(3)} = \begin{bmatrix} .179 & .170 & .652 \\ .154 & .165 & .681 \\ .235 & .510 & .255 \\ .024 & .020 & .956 \end{bmatrix} \qquad \ddot{\mathbf{Y}}^{(1)} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\vdots$$

$$\mathbf{Y}^{(\infty)} = \begin{bmatrix} .156 & .162 & .682 \\ .122 & .153 & .726 \\ .219 & .499 & .282 \\ .014 & .017 & .969 \end{bmatrix} \qquad \ddot{\mathbf{Y}}^{(1)} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Person 1's initial preference for option C has been strengthened. Person 2's initial preference for option A has decisively switched to option C. Person 3's initial preference for option B has been maintained on a somewhat weakened basis. Person 4's initial strong preference for option C has been maintained. The influence system has generated a majority whose most preferred preference is option C.

Conclusion

The literature on models of influence in networks is substantially larger and more complex than is suggested by the focus of this article. There is no lack of imagination in the field in constructing such models. However, an intimate dance between theoreticians and experimentalists, which has served to dramatically advance particular fields of the natural sciences, is rare. My perspective on this body of work is captured by a lithograph which shows a logger, axe in hand, looking at a tree that he has just cut through: the tree, instead of toppling over, has floated upwards and the logger is standing dumbstruck, perhaps disappointed, at the result. Plausible formalizations may or may not be supported by detailed nontrivial empirical observations. For example, the demonstration that a particular model can generate a consensus or bipolar factions on an issue is important, since these generic types of outcomes have been observed, but many models can do the same and may differ markedly in the accuracy with which they predict the content of emergent orientations. Sustained lines of empirical inquiry on dynamic models of endogenous interpersonal influences unfolding in a network are surprisingly scarce.

Untested models with explorations (via simulation) of their formal properties are the rule. The main, and considerable, value of such work is the demonstration of what might be explained if the assumptions of the dynamic model are well founded. However, a more jaundiced perspective also is applicable: such explanations are easily obtained and not compelling in the absence of empirical evaluations of their predictions. While the judgment on what constitutes a compelling empirical assessment of a formal architecture may rest on different standards, the advancement of science has not suffered from its emphasis on parsimony and measured constructs. Note that the Friedkin–Johnsen model involves no unmeasured constructs. The predictions of the model are generated strictly on the basis of individuals' initial orientations $\mathbf{Y}^{(1)}$ and their influence network $\mathbf{W}$, for which measures may be, and have been, gotten.

A sustained line of empirical work has occurred on Friedkin–Johnsen's standard model. I close with a brief description of these supports, and in so doing emphasize the importance of obtaining the direct measures of individuals' initial orientations toward an object, which are involved in the evaluations conducted to date. The supports are based on experiments with small groups, in a standard design that obtains pre- and postdiscussion orientations, and measures of each group member's postdiscussion subjective accorded weights to self and particular others (i.e., each row of $\mathbf{W}$). I have previously noted that the employment of subjective weights is consistent with the cognitive foundations of the postulated mechanism, and have emphasized the importance of further work on measurement models of $\mathbf{W}$. The available evidence indicates that influence networks are issue specific and may differ for the same group of individuals as they consider a sequence of issues.

The laboratory experiments on the model have included designs in which members' interpersonal communications are constrained to dyadic interaction in telephonic communication networks with different structures of open lines of communication, and designs in which members are communicating via computers with broadcasts to all members of the group. The model's simplifying assumption on sequencing (i.e., all individuals simultaneously respond to all visible and salient orientations in a temporal sequence of state changes) does not appear be seriously misleading. Notably, in addition, the model has been evaluated in a quasi-experimental field setting, dealing with established book clubs who have met regularly face-to-face to discuss selected texts (Childress and Friedkin, 2012). The same design was employed, with measures of individuals' pre- and postdiscussion orientations toward a particular book, and measures of each group member's postdiscussion subjective accorded weights to self and particular others. Because these are enduring groups, an attempt was made to obtain a prediscussion measure of $\mathbf{W}$, under the expectation of a stable preestablished influence network. However, a substantial proportion of book club members disabled this measure with their expressions of doubt about 'tending' to be influenced by the same individuals across meetings; the subjects noted that while such a tendency sometimes occurs, their accord of influence and its distribution vary substantially from meeting to meeting depending on the subject matter of the book being discussed.

In each of the studies, the empirical analysis focused on the accuracy of the input–output formalization, that is, the predicted postdiscussion orientations $\hat{\mathbf{y}}^{(\infty)} = \mathbf{V}\mathbf{y}^{(1)}$, based on a one-dimensional measure of the vector of initial attitudes $\mathbf{y}^{(1)}$ and the subjective direct measure of each group's influence network $\mathbf{W}$, where the emergent total predicted influences are the elements of the matrix $\mathbf{V}$. The predicted postdiscussion orientations are elements of the vector $\hat{\mathbf{y}}^{(\infty)}$, and the predicted group-level mean orientation for each group is the mean value of the elements of this vector. Support for the model is indicated in group-level analyses of these data, where groups' postdiscussion mean orientations are regressed on their predicted mean orientations, and in multilevel mixed-effects regression analyses of individuals nested in groups (Rabe-Hesketh and Skrondal, 2008), where the dependent variable is individuals' postdiscussion orientations and the independent variables are individuals' predicted postdiscussion orientations and the predicted group-mean orientation.

In both the laboratory and field settings, significant effects of the model's predictors were maintained under controls for the individuals' initial orientations and the groups' initial mean orientations. With any model that constrains predictions to the convex hull of initial orientations, such controls are essential to the assessment of the model. If a model's predictions do not significantly contribute to explaining the direction and magnitude of departures of individuals' postdiscussion orientations from their prediscussion orientations, and departures of groups' postdiscussion mean orientations from their prediscussion mean orientations, then the approach (its particular

combination of measurement and formal models) must be rejected. To date, the Friedkin–Johnsen model has survived this test on six issues, with experiments conducted in laboratories of three universities and one field setting, under different conditions of pressure to reach consensus; under different types of, and more or less constrained, communication structures; and with datasets that were assembled by investigators other than Friedkin and Johnsen.

Subjective heterogeneity (i.e., a mind's idiosyncratic allocation of weights) is an important feature of accorded influences. This feature is subject to conditioning, and special cases of extreme conditioning can elevate an individual's openness to interpersonal influence and constrain the individual's accord of influence to particular other individuals. Social ecology and communication technology have implications for the pattern of interpersonal visibility of displayed orientations. Acknowledging such effects does not erode the fundamental postulate that influence networks are social cognition structures, assembled by the weights that individuals allocate to the displayed orientations of other individuals, and to their own orientations. The accorded weights are components of the heuristic within-the-skin mechanism with which each individual synthesizes heterogeneous orientations. The repetitive automatic activation of this mechanism enables flows of influence in the network. Under this premise, the investigation of influence in networks importantly extends the implications of natural automatic heuristic mechanisms of cognition to the domain of group dynamics.

Bibliography

Abelson, R.P., 1964. Mathematical models of the distribution of attitudes under controversy. In: Frederiksen, N., Gulliksen, H. (Eds.), Contributions to Mathematical Psychology. Holt, Rinehart & Winston, New York, pp. 142–160.

Anderson, N.H., 1981. Foundations of Information Integration Theory. Academic Press, New York.

Anderson, N.H., 1991. Contributions to Information Integration Theory. Lawrence Erlbaum, Hillsdale, NJ.

Bargh, J.A., Ferguson, M.J., 2000. Beyond behaviorism: on the automaticity of higher mental processes. Psychological Bulletin 126, 925–945.

Childress, C.C., Friedkin, N.E., 2012. Cultural reception and production: the social construction of meaning in book clubs. American Sociological Review 77, 45–68.

DeGroot, M.H., 1974. Reaching a consensus. Journal of the American Statistical Association 69, 118–121.

Festinger, L., 1954. A theory of social comparison processes. Human Relations 7, 117–140.

French Jr., J.R.P., 1956. A formal theory of social power. The Psychological Review 63, 181–194.

Friedkin, N.E., 1998. A Structural Theory of Social Influence. Cambridge University Press, New York.

Friedkin, N.E., 1999. Choice shift and group polarization. American Sociological Review 64, 856–875.

Friedkin, N.E., 2001. Norm formation in social influence networks. Social Networks 23, 167–189.

Friedkin, N.E., 2010. The attitude-behavior linkage in behavioral cascades. Social Psychology Quarterly 73, 196–213.

Friedkin, N.E., Johnsen, E.C., 1990. Social influence and opinions. Journal of Mathematical Sociology 15, 193–206.

Friedkin, N.E., Johnsen, E.C., 1999. Social influence networks and opinion change. Advances in Group Processes 16, 1–29.

Friedkin, N.E., Johnsen, E.C., 2011. Social Influence Network Theory: A Sociological Examination of Small Group Dynamics. Cambridge University Press, New York.

Granovetter, M.S., 1978. Threshold models of collective behavior. American Journal of Sociology 83, 1420–1443.

Harary, F., 1959. A criterion for unanimity in French's theory of social power. In: Cartwright, D. (Ed.), Studies in Social Power. Institute for Social Research, Ann Arbor, MI, pp. 168–182.

Kahneman, D., 2011. Thinking, Fast and Slow. Farrar, Straus and Giroux, New York.

Kahneman, D., Slovic, P., Tversky, A., 1982. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, New York.

Ord, K., 1975. Estimation methods for models of spatial interaction. Journal of the American Statistical Association 70, 120–126.

Rabe-Hesketh, S., Skrondal, A., 2008. Multilevel and Longitudinal Modeling Using Stata. Stata Press, College Station, TX.

Watts, D.J., Dodds, P.S., 2007. Influentials, networks, and public opinion. Journal of Consumer Research 34, 441–458.

Models of Power in Exchange Networks

John Skvoretz, University of South Florida, Tampa, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Research into power in exchange networks is based on a large corpus of experimental studies and has spurred the development of many theories to account for experimental outcomes. The types of networks that fall into the domain of these theories are described. How these objects are instantiated in experimental procedures are outlined. The various theories that have been proposed are mentioned but four are set out in detail: power dependence theory, core theory, expected value theory, and network exchange theory. Finally, the results of two recent evaluations of these theories, one empirical and one theoretical, are summarized.

Research on power in exchange networks is unusual in sociology because it formulates relatively clear mathematical models for power in the context of their evaluation through laboratory experiments. In this research, power is measured by the relative earnings from exchange between subjects occupying positions in networks of exchange. Power is relational since the profit from exchange is zero-sum, in the sense that the more profit A earns in exchanges with B, the less profit B earns. However, exchange is mixed motive because both A and B are better off if they exchange, even if one is only marginally better off, than if they do not exchange with each other or with any other person. If A earns reliably more profit in exchange with B than B does with A, A is deemed to have power in the A–B exchanges. Informally, power outcomes fall into three types: strong power outcomes in which A reliably earns much more than B, weak power in which A reliably earns more than B but only a modest amount more, and null or equal power in which A and B earn about the same amount, or more technically, the difference between their earnings is not significantly different from zero.

Formal modeling was evident 40 years ago in the early development of this research field. Emerson (1962: p.31) used network imagery and basic equations relating power and dependence for his "simple theory of power relations." As compared to recent work, Emerson's formal models were crude, even when more fully elaborated several years later (Emerson, 1972) but they set the tone for future research. In part, the nature of the problems demanded mathematical treatment since it was clear from the beginning that A's power in exchange with B was affected by the nature of the interdependencies between the A–B relation and the other exchange relations that A may have and that B may have, and the relations that their partners may have, and so on, as embodied in the overall pattern of connection in the network. Tracing out the path of effects could not simply be left to verbal gymnastics.

The use of experiments to evaluate formal models is also a longstanding feature of this research area beginning with Cook and Emerson (1978) and including Brennan (1981), Cook et al. (1983), Markovsky et al. (1988), Molm (1990), Skvoretz and Willer (1991, 1993), Bienenstock and Bonacich (1993), Lovaglia et al. (1995a), Molm (1997), Lucas et al. (2001), Emmanuelson (2005), and Willer and Emmanuelson (2008). In general, these experiments use social psychology's equivalent of the fruit fly, the undergraduate, as subjects who are placed in relatively small exchange networks and left to bargain with one another over the terms of exchange. In early research, communication was done via paper and pencil and sometimes through face-to-face communication with barriers to prevent some pairs from communicating. Current state of the art in experiments uses workstations connected via local area networks and sophisticated GUI displays to present setting information and in-round details of the back and forth of negotiations.

Theories of power in such networks do not focus on how A's power in exchanges with B depend on differences in personal attributes like bargaining toughness or aggressiveness. Rather the scientific problem is generated by the fact that A, B, and their exchange relation are embedded in a network of such relations and the important questions are (1) to identify the properties of positions in such networks that provide advantage or disadvantage and (2) to uncover the mechanisms by which these properties produce power differences of greater or lesser magnitude. In pursuit of these questions, variation in individual attributes is, for the most part, noise to be statistically controlled and filtered out of the analysis. At the same time, the pursuit of these questions requires data from networks in which relevant properties of the positions occupied by A's and B's vary. These properties capture differences in how positions are connected directly or indirectly to others in the network through direct exchange relations or chains of such relations.

These questions have attracted the attention of many researchers over the past 40 years and there have been many theories or models for power in exchange networks proposed, including power dependence theory (Cook and Emerson, 1978; Cook and Yamagishi, 1992; Buskens and van de Rijt, 2008), Vulnerability theory (Cook et al., 1983; Cook et al., 1986), graph power index or GPI theory (Markovsky et al., 1988; Patton and Willer, 1990; Markovsky et al., 1993), expected value theory (Friedkin, 1992, 1993; Bonacich and Friedkin, 1998), core theory (Bienenstock and Bonacich, 1992, 1993; Bonacich and Bienenstock, 1995, 1997), exchange-resistance theory (Skvoretz and Willer, 1993), reciprocity theory (Skvoretz and Lovaglia, 1995), identity theory (Burke, 1997), and versions of network exchange theory, which include optimal seek theory (Simpson and Willer, 1999; Girard and

Borch, 2003) and, most recently, domain analysis theory (Willer et al., 2012). In a systematic survey of predictions, van de Rijt and van Assen (2008: p. 260) restrict attention to four – core theory, power dependence theory, expected value theory, and network exchange theory – because "no theories have been investigated as much and received as much attention as these theories."

Exchange networks comprise the domain of all these theories. Some basic terms define these entities. Actors refer to occupants of network positions; they may be concretely individuals, teams, organizations, or other collectivities. Negotiated exchange refers to a bilateral agreement that benefits both partners, and in typical experimental setups exchange occurs when there is mutual agreement on the division of a 'profit pool' of points. There are also nonnegotiated forms of exchange defined by the mutual gifts of points to partners. In negotiated exchanges, potential partners can bargain over the division of the pool – the actual procedures for bargaining can differ across experimental setups. In nonnegotiated exchanges, an actor has the option to send a number of points, fixed by experimental design to a partner and no capability to change or 'negotiate' the number of points. The vast majority of empirical studies have focused on negotiated exchange networks. Molm (1990) and Molm et al. (1999) are notable exceptions.

An exchange tie refers to a pair of positions between which an exchange is possible while an exchange network is a set of positions connected by exchange ties. A position in such a network is a location in the overall pattern of connection. Profit pool refers to the total number of points that can be divided in an exchange. Typically, all profit pools are the same size. Experiments are usually organized into sessions or rounds. In negotiated exchange networks, a session is an opportunity for actors to negotiate pool divisions with those to whom they are connected by an exchange tie. In nonnegotiated exchange networks, a session is an opportunity to give points to a partner.

The networks themselves are built up from exchange ties. As noted earlier, critical to the phenomenon of power in these networks is the type of interdependency between and among exchange ties. Classification of exchange ties focus on how exchange in one of an actor's ties impacts exchange in other ties that the actor has. (Clearly, when an actor has only one tie, the distinctions are moot.) Terminology varies somewhat from research program to research program. One type of interdependency, however, is of little interest and that is 'null' interdependency. Two exchange ties have a null connection or interdependence if the completion of exchange in one of them has no consequences for the chance that exchange is completed in the other and vice-versa. These connections are of little interest because the completion of exchange in any one tie is independent from the completion of exchange in any other tie. Thus the 'network' reduces to an independent set of dyadic negotiations subject only to time constraints that may apply if a position has many ties to others and only limited time in which to complete an exchange in each. Therefore, there can be no systematic or structural basis for power in such networks of null interdependence. In terms of early network theorists Holland and Leinhardt (1979), such networks exhibit no interesting structure because what occurs in one dyad is independent of outcomes in any other.

A widely investigated type of interdependence that does have structural consequences of power development is negative or exclusive connection. A pair of exchange ties at a position is negatively or exclusively interdependent, if completing an exchange in one precludes or eliminates the possibility of completing an exchange in the other. More generally, negative connection between two ties means that completing an exchange in one tie decreases the probability of exchange in another. At the extreme, the second exchange is impossible. This extreme occurs in the '1 exchange regime, an experimental protocol that creates networks of negatively connected exchange ties by permitting actors to complete only one exchange per session. The negative or exclusionary interdependence can be more complicated holding among three or more exchange relations through a regime that limits the total number of exchanges that may be completed to less than total number of exchange ties available. In this more complicated case, completion of an exchange in any one tie does not preclude completion of an exchange in any particular other tie, unless it is the last possible exchange to be completed in which case exchange is precluded in all remaining active ties.

Less widely investigated but still of importance structurally is positive or inclusive connection or interdependence. A pair of exchange ties is positively or inclusively connected if the completion of exchange in one increases the probability of exchange in the other. There are several experimental protocols that can realize this type of interdependence. For example, the experiment could stipulate that earnings in one tie can only be converted to actual payoffs if an exchange is completed in the other tie. A related protocol creates positive connection by stipulating what is earned in one tie can then be traded for something of even greater value in a second tie. Again, sets of three or more exchange ties can have more complicated modes of positive or inclusive connection. Furthermore, sets of exchange ties can combine null, negative, and positive connections in various combinations making for a very complex network of interdependencies of exchange ties across positions. In general, however, theory and research focuses on negatively or exclusively connected networks. Szmatka and Willer (1995) and Willer and Skvoretz (1997) are notable exceptions.

Experimental methods for negatively connected networks have some common features. First, subjects are instructed to earn as much as possible for themselves, that is, to approximate the utility maximizing actor of the rational choice theoretical framework. Second, all profit pools that define the possible earnings from exchange are of the same size. In one standard implementation, the full information design, subjects are presented with the overall design of the network clearly indicating which position they occupy and they expect to change positions periodically. The profit pools contain 24 points. Offers, counteroffers, and agreements can be communicated at any time during bargaining sessions of fixed duration. Exchange occurs when partners agree on a division of their pool. Once all agreements that could be made are made or time runs out, the session ends and a subject earns whatever share was agreed upon if they made an exchange or earns 0 points for the round if he or she failed to complete an exchange. Note that failure could be due to time running out before an exchange was made or due to all potential partners completing exchanges with others. In another standard implementation, restricted

information, subjects only know how many partners they have and not the overall design of the network. They do not expect and do not change positions periodically. Profit pools are 30 points. Subjects make one offer per session to each partner, and may increase/decrease by one point (or leave unchanged) the offer in the previous session. Exchange occurs when offers are mutually compatible (with multiple compatible ones eliminated by random selection). Skvoretz and Burkett (1994), Lovaglia et al. (1995a), and Simpson et al. (2011) are three studies that investigate how information differentials may affect the distribution of power.

As van de Rijt and van Assen (2008: p. 260) point out, four theories of power in exchange are the most widely represented in empirical studies: core theory, power dependence theory, expected value theory, and network exchange theory. Each of these may be described in turn.

Core theory applies concepts from cooperative game theory, namely, the concept of the core as a special set of outcomes, to network exchange (Bienenstock and Bonacich, 1992, 1993; Bonacich and Bienenstock, 1995, 1997). Specifically, core theory assumes that when a core set of outcomes exists, it contains all outcomes that will be observed. An outcome, defined as a payoff distribution to each actor, is in the core if it satisfies conditions of individual, coalitional, and collective rationality. Individual rationality is satisfied if each person earns at least as much as he or she could alone. Since that quantity is always 0 by experimental design, the condition is trivially satisfied. Second, coalitional rationality means that each subset of actors should earn in sum at least what they get without the cooperation of actors not in the subset. Finally, collective rationality means that the sum of the earnings in the entire group is as large as possible. Bonacich (1998) shows that there are four types of core solutions for the typical 1-exchange regime: (1) the entire profit pool is allocated to some positions and zero is allocated to all other positions, (2) exactly $^1/_2$ of the pool is allocated to any pair completing an exchange, (3) an infinite number of allocations constitute the core, and (4) the core is empty. He also shows that in the third case, additional considerations can put inequality bounds on the outcomes such that it is possible to infer that some positions should earn no less than $^1/_2$ of the pool size in exchanges.

Power dependence theory assumes that actors evaluate offers from partners in relation to how much they could get in some alternative relation (Cook and Emerson, 1978; Cook and Yamagishi, 1992; Buskens and van de Rijt, 2008). This possible payoff from an alternative partner is the actor's comparison level for exchanges with a focal partner. The difference between the comparison level and the offered terms defines the dependency of the actor on the focal partner for favorable outcomes. Simultaneously, the partners of the actor are conducting the same evaluation of the actor's offers. Terms of exchange are struck at the point that both actors are equally dependent on each other for favorable outcomes, that is, at the point that both are earning in exchange with each other exactly the same number of points more than their respective next best offers. Power dependence theory holds that this logic applies to all but suboptimal exchange ties in a network, where suboptimal ties are ones in which the partners can gain more from other partners at equidependence than from each other and so have no incentive to ever complete an exchange.

Expected value theory proposes that the earnings received from an exchange condition are the likelihood that the exchange will occur and these likelihoods, in turn, determine the probabilities that actors are excluded from exchange (Friedkin, 1992, 1993, 1995; Bonacich and Friedkin, 1998). These probabilities, in turn, influence the terms of exchange that actors will offer one another and so lead to differences in earnings received. Then, in a recursive fashion, the earnings received lead to adjustments in the likelihoods of exchange, which then change the probabilities of being excluded, which finally impact the terms offered and so on until a formally defined convergence criterion is reached. Expected value theory is more similar to power dependence theory than it is to core theory – it produces specific point predictions of power differences in all types of networks, unlike core theory, but, unlike power dependence theory, it adjusts the expectations of value to be received from another source by the probability of receiving such value through exchange.

Network exchange theory, unlike the previous three theories, is the most difficult to summarize because it has had many variants over several decades of development (Markovsky et al., 1988; Patton and Willer, 1990; Markovsky et al., 1993; Skvoretz and Willer, 1993; Skvoretz and Lovaglia, 1995; Lucas et al., 2001). The variant that recent research (Willer and Emmanuelson, 2008) considers the 'best' based on empirical evidence combines graph power index – resistance theory (Lovaglia et al., 1995b) with optimal seek (simplified) theory (Girard and Borch, 2003). The basic idea of this theory is to first identify substructures of an exchange network that are strong power, equal power, or weak power substructures and the breaks between such substructures that will occur because actors have no incentive to seek exchange with one another even though they are connected by an exchange relation. In strong power substructures, actors in one set of positions (labeled I for included always) are predicted to earn the maximum share of a pool and actors in the other set of positions (labeled E for excludable) are predicted to earn the minimum share. In equal power components, all nodes are I or all are E and have the same number of connections and so the prediction is that, of the net of exogenous individual factors, each actor will earn 50% of any pool. Calculation of shares in the weak power components is more complicated, taking into account differences between positions in their likelihood of being included in exchange and how such a difference affects the best and worst they can expect from an exchange.

Two recent evaluations of these theories exist. Willer and Emmanuelson (2008) use data from seven 'weak power' networks to test the predictions from these four theories and six others. Weak power networks are selected for the basic reason that predictions of all theories for the other types are indistinguishable, that is, all theories propose that in strong power networks advantaged positions will earn nearly all the available points in a pool and all theories propose that in equal power networks, pools will be split evenly. It is only in the case of weak power networks that the theories make distinguishably different predictions. As noted above, they find that the best fit to data from these networks is provided by a version of network exchange theory.

The second evaluation by van de Rijt and van Assen (2008) is not an empirical evaluation but an evaluation of logical aspects of the theories against six desiderata: predictions should be invariant under scalar multiplication of the size of the profit pool; predictions should be identical for positions that are automorphically equivalent, distinguishable from one another only by label (symmetry); predictions should exhibit outcome stability, that is, the sum of predicted payoffs for any pair of connected actors is always greater than the sum attainable by exchanging themselves; predictions should be sensitive to link deletion in a downward direction only, that is, if the link between A and B is deleted, their predicted payoffs are less than or equal to their predicted payoffs in the network with the link restored; predictions should exit, that is, there should be at least one payoff vector predicted; and, sixth, the prediction should be unique. None of the four theories has all six desirable properties. In two cases, expected value theory for the existence property and network exchange theory for the symmetry property, van de Rijt and van Assen are unable to determine whether or not the theory satisfies the criterion. Their evaluation concludes that no one theory is superior to the others and so any nonempirical choice among theories would be based on the relative value placed by a theorist on each of the six desiderata. For instance, an advocate of network exchange theory could argue that its violation of the first criterion, because it predicts a disadvantaged actor will get 4.2% of a 24-point pool but 20% of a 5-point pool (getting the minimum possible in both cases), is not a very serious error, but rather reflects better than the other alternatives – the constraints of any experimental protocol – namely, that even the most severely disadvantaged must get something to motivate completion of an exchange. Or an advocate of core theory could argue that its violation of the symmetry criterion with respect to a particular payoff vector is misleading since in the case cited, it predicts a set of vectors in which automorphically equivalent actors occupy identical positions when all imputations are considered and of which the particular vector is just one possible outcome.

These two studies evaluate theories for power in negotiated exchange networks. Power in nonnegotiated or reciprocal exchange networks has received limited attention. In these networks, an exchange relation is defined by the opportunity for an actor to give unilaterally a fixed number of points to his or her partner. If the actor has more than one partner, only one can be selected on a round, thereby, creating a negative connection between his or her exchange ties – the more often the actor gives in one relation, the less often the actor gives in the others. Available evidence provided by Molm et al. (1999) suggests reciprocal exchange produces smaller power differentials in otherwise similar networks but the same pattern of difference across different types of networks – namely, greater earnings differences in strong power networks than in weak power networks.

Summary

A general theory of power in networks, even one limited to networks of negative connection, remains an elusive goal and not for lack of effort. Over the years, various proposals have been made. Even the more successful have not been built in a dynamic modeling framework – the general pattern is to identify some measure of position sensitive to the location of that position in the overall pattern of connection and postulate that a difference in this measure between two positions connected by an exchange relation will produce a greater or lesser difference in the earnings of actors in these positions from exchange. (This generalization includes proposals, like expected value theory, that have an iterative component in which values expected from exchange condition the probabilities of exchange, which in turn modify the values until some convergence is reached.) The goal is to understand how structure (network location) shapes process (negotiation to agreement) to produce structural effects on outcomes (differential earnings), but without an explicit representation of the process it is doubtful that the goal can be achieved. Even if a proposal, such as the most recent one deemed best by Willer and Emmanuelson (2008), is generally agreed to be descriptively adequate, the deeper question of explanatory adequacy – why a particular representation of the problem provides empirically satisfactory fits – requires delving into the details of process, a multiagent search through a space of possible deals constrained by the complex interdependencies among points deriving from the underlying network of exchange connections.

See also: Analytical Sociology; Axiomatic Theories; Game Theory; Models of Influence in Networks; Network Centrality, Measures of; Social Network Algorithms and Software; Social Simulation: Computational Models.

Bibliography

Bienenstock, E.J., Bonacich, P., 1992. The core as a solution to exclusionary exchange networks. Social Networks 14, 231–243.
Bienenstock, E.J., Bonacich, P., 1993. Game theory models for exchange networks: experimental results. Sociological Perspectives 36, 117–136.
Bonacich, P., 1998. A behavioral foundation for a structural theory of power in exchange networks. Social Psychology Quarterly 61, 185–198.
Bonacich, P., Bienenstock, E.J., 1995. When rationality fails: unstable exchange networks with empty cores. Rationality and Society 7, 293–320.
Bonacich, P., Bienenstock, E.J., 1997. Network exchange as a cooperative game. Rationality and Society 9, 37–65.
Bonacich, P., Friedkin, N.E., 1998. Unequally valued exchange relations. Social Psychology Quarterly 61, 160–171.
Brennan, J.S., 1981. Some Experimental Structures. In: Willer, D., Anderson, B. (Eds.), Networks, Exchange and Coercion. Elsevier/Greenwood, New York, pp. 189–204.
Burke, P., 1997. An identity model for network exchange. American Sociological Review 62, 134–150.
Buskens, V., van de Rijt, A., 2008. Sequential power-dependence theory. Journal of Mathematical Sociology 32, 110–128.
Cook, K.S., Emerson, R.M., 1978. Power, equity, and commitment in exchange networks. American Sociological Review 43, 721–739.
Cook, K.S., Yamagishi, T., 1992. Power in exchange networks: a power-dependence formulation. Social Networks 14, 245–265.
Cook, K.S., Emerson, R.M., Gillmore, M.R., Yamagishi, T., 1983. The distribution of power in exchange networks: theory and experimental results. American Journal of Sociology 89, 275–305.
Cook, K.S., Gillmore, M., Yamagishi, T., 1986. Point and line vulnerability as bases for predicting the distribution of power in exchange networks: reply to Willer. American Journal of Sociology 92, 445–448.
Emerson, R.M., 1962. Power dependence relations. American Sociological Review 17, 31–41.

Emerson, R.M., 1972. Exchange theory part II: exchange relations and network structures. In: Berger, J., Zeldtich Jr., M., Anderson, B. (Eds.), Sociological Theories in Progress, vol. 2. Houghton-Mifflin, Boston, pp. 59–87.

Emmanuelson, P., 2005. Improving the precision and parsimony of network exchange theory: a comparison of three network exchange models. Current Research in Social Psychology 10, 149–165.

Friedkin, N.E., 1992. An expected value model of social power: predictions for selected exchange networks. Social Networks 14, 213–229.

Friedkin, N.E., 1993. An expected value model of social exchange outcomes. Advances in Group Processes 10, 163–193.

Friedkin, N.E., 1995. The incidence of exchange networks. Social Psychology Quarterly 58, 213–222.

Girard, D., Borch, C., 2003. Optimal seek simplified. Current Research in Social Psychology 8, 16.

Holland, P.W., Leinhardt, S., 1979. Structural sociometry. In: Holland, P.W., Leinhardt, S. (Eds.), Perspectives on Social Network Research. Academic Press, New York, pp. 63–83.

Lovaglia, M., Skvoretz, J., Willer, D., Markovsky, B., 1995a. Negotiated outcomes in social exchange networks. Social Forces 74, 123–155.

Lovaglia, M., Skvoretz, J., Markovsky, B., Willer, D., 1995b. Assessing fundamental power differences in exchange networks. Current Research in Social Psychology 1, 8–17.

Lucas, J.W., Wesley Younts, C., Lovaglia, M.L., Markovsky, B., 2001. Lines of power in exchange networks. Social Forces 80, 185–214.

Markovsky, B., Willer, D., Patton, T., 1988. Power relations in exchange networks. American Sociological Review 53, 220–236.

Markovsky, B., Skvoretz, J., Willer, D., Lovaglia, M., Erger, J., 1993. The seeds of weak power: an extension of network exchange theory. American Sociological Review 58, 197–209.

Molm, L.D., 1990. Structure, action and outcomes: the dynamics of power in exchange relations. American Sociological Review 55, 427–447.

Molm, L.D., 1997. Coercive Power in Social Exchange. Cambridge University Press, Cambridge, UK.

Molm, L.D., Peterson, G., Takahashi, N., 1999. Power in negotiated and reciprocal exchange. American Sociological Review 64, 876–890.

Patton, T., Willer, D., 1990. Connection and power in centralized exchange networks. Journal of Mathematical Sociology 16, 31–49.

Simpson, B., Willer, D., 1999. A new method for finding power structures. In: Willer, D. (Ed.), Network Exchange Theory. Praeger, Westport, CT, pp. 270–284.

Simpson, B., Markovsky, B., Steketee, M., 2011. Network knowledge and the use of power. Social Networks 33, 172–176.

Skvoretz, J., Burkett, T., 1994. Information and the distribution of power in exchange networks. The Journal of Mathematical Sociology 19, 263–278.

Skvoretz, J., Lovaglia, M.J., 1995. Who exchanges with whom: structural determinants of exchange frequency in negotiated exchange networks. Social Psychology Quarterly 58, 163–177.

Skvoretz, J., Willer, D., 1991. Power in exchange networks: setting and structure variations. Social Psychology Quarterly 54, 224–238.

Skvoretz, J., Willer, D., 1993. Exclusion and power: a test of four theories of power in exchange networks. American Sociological Review 58, 801–818.

Szmatka, J., Willer, D., 1995. Exclusion, inclusion and compound connection in exchange networks. Social Psychology Quarterly 58, 123–131.

van de Rijt, A., van Assen, M.A.L.M., 2008. Theories of network exchange: anomalies, desirable properties, and critical networks. Social Networks 30, 259–271.

Willer, D., Emmanuelson, P., 2008. Testing ten theories. Journal of Mathematical Sociology 32, 165–203.

Willer, D., Skvoretz, J., 1997. Network connection and exchange ratios: theory, predictions and experimental tests. Advances in Group Process 14, 199–234.

Willer, D., van Assen, M.A.L.M., Emmanualson, P., 2012. Analyzing large scale exchange networks. Social Networks 34, 171–180.

Moderator Variable: Methodology

Charles M Judd, University of Colorado, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 9937–9939, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article introduces the definition of moderator variable in statistical analysis and how moderation is different from mediation and interaction. It also discusses potential problems with the interpretation of the moderation effect that arise in observational studies, as well as their relationship with significance testing and variable transformations.

Typically theoretical models in the behavioral sciences posit that independent variables affect dependent variables. A moderator variable is a variable, which is thought to temper or modulate the magnitude of the effect of an independent variable on a dependent one. Moderators may be characteristics of people or characteristics of situations. In either case, they affect the magnitude of the relationship between an independent variable and a dependent one. Moderation implies an interaction between the independent variable and the moderator. Analytic procedures using both analysis of variance and, more generally, multiple regression, permit the testing of such interactions. Tests of interactions with measured, as opposed to manipulated variables, may not have high statistical power. Complications also arise due to the typically arbitrary metrics of many dependent variables. Nonlinear scale transformations can eliminate interactions and affect conclusions about the presence or absence of moderation.

A moderator variable is a variable, which is thought to temper or modulate the magnitude of the effect of an independent variable on a dependent one. Conceptually, it is important to differentiate between a moderator and a mediator. A moderator is a variable that affects the magnitude of the relationship between the independent and dependent variables. It identifies the conditions under which, or the type of participant for whom, the effect is likely to be particularly large or particularly small. A mediator is a variable that describes the process that is responsible for the effect of the independent variable on the dependent one. It is a variable that is affected by the independent variable and in turn affects the dependent one, thus being responsible for the effect. (*See* Mediating Variable; also Baron and Kenny, 1986; Muller et al., 2005; MacKinnon et al., 2007; Fairchild and MacKinnon, 2009; Pearl, 2010, 2012.)

Examples of Hypotheses Involving Moderation

Hypotheses about moderation are ubiquitous in the social and behavioral sciences. The following examples, taken from a variety of fields, provide illustrations. Each example consists of two sentences. The first hypothesizes an effect of an independent variable on a dependent one. The second sentence identifies a hypothesized moderator.

1. Level of education attained affects lifelong earnings. This relationship is less strong for females than it is for males.
2. Stressful life events tend to produce psychological problems. These effects are lessened if one has an extensive social support network.
3. Some forms of instruction lead to better retention than others. These effects are larger among more able students.
4. Exercise has pronounced health benefits. These benefits are more pronounced among older people.

Note in these examples that moderator variables can take on a variety of forms. They can refer to characteristics of the participants in the research (i.e., their gender, ability, or age); equally plausibly they may characterize the situations or environments which moderate the relationship between the independent and dependent variables. Additionally, their scale of measurement may be nominal (e.g., gender) or may vary more or less continuously (e.g., ability levels).

Conceptual and Analytic Issues in Testing Moderation

To say that one variable moderates the effect of another is equivalent to saying that there is a statistical interaction between the two variables. Accordingly, moderation implies interaction, although not all interactions involve moderating variables. The distinction is as follows: moderation means that there is a main effect of the independent variable plus an interaction between the independent variable and the moderator. Overall there is an effect of the independent variable, but the magnitude of that effect depends on the moderator. An interaction, more generally, makes no claims about the presence or absence of the main effect of the independent variable. It simply indicates that the effect of one variable depends on another.

Because moderation implies a main effect of the independent variable plus an interaction of that variable with the moderator, typically interactions in the presence of moderators are not what are called crossover or disordinal interactions. That is, from one level of the moderator to others, the effect of the independent variable varies in magnitude but not in direction. To illustrate, consider the first of the examples given above, that gender moderates the relationship between educational attainment and lifetime earnings. For both genders, there is presumably a positive effect of educational attainment on earnings, but that positive effect is less strong among females than among males. For no level of the moderator does the relationship between educational attainment and earnings reverse.

Because moderation implies an interaction, it is necessary that the moderator and the independent variable be at least partially independent of each other. Ideally, the two variables would be fully crossed as in an experimental design where both variables are manipulated independently of each other. More typically, while the independent variable may be manipulated, the moderator may be a measured variable that varies more or less continuously. Again, however, to assess moderation it is important to insure that observations from all levels of the moderator occur at every level of the independent variable.

Analytically, moderation is typically assessed via either analysis of variance or multiple regression procedures, with the choice between the procedures traditionally depending on the scale of measurement of the independent and moderator variables. Two effects are of interest. First, the overall main effect of the independent variable is expected. Second, moderation of that effect implies that the interaction between the independent variable and the moderator is statistically significant. The use of analysis of variance to assess moderation is perfectly appropriate when both the independent variable and the moderator are measured on a nominal scale, with a few discrete levels of each variable. With moderators that vary more or less continuously, regression procedures are more appropriate, since substantial loss of power can occur if continuously varying moderators are broken up into discrete levels (e.g., using a median split) to accommodate them in analysis of variance. In general, multiple regression can be used to assess moderation, regardless of whether independent and moderator variables are scaled discretely or more or less continuously. In the case of discrete or nominal scaling, one simply must compute dummy-coded or contrast-coded variables to be included as predictors. The general model involves including as predictors the independent variable, the moderator variable, and a variable that is the product of the two. The test of the product variable, controlling for the two components of that product, provides a test of the statistical interaction (Cohen, 1978). In this analysis, one must be cautious in interpreting the coefficients (and their associated statistical tests) of the component variables (i.e., the independent and moderator variables in a model that includes their product). Unless these variables have been centered on their means (and the product computed from those centered variables), their coefficients will not be readily interpretable as the average or main effect of each variable (Aiken and West, 1991; Judd and McClelland, 1989; Henseler and Fassott, 2010; Aguinis, 2004).

In some situations, the independent variable is likely to vary within units rather than between them (Judd et al., 2001). For instance, the effects of mood as an independent variable may be assessed by naturally occurring variations in mood within individuals over time. In these cases, traditional tests of statistical significance of both the main effects of the independent variable and its interaction with a moderator are likely to be biased unless dependencies in the data are appropriately handled. Repeated measures analysis of variance and, more generally, hierarchical linear models can be used to assess moderation in these sorts of designs.

Tests of moderation with naturally varying, as opposed to manipulated, independent, and moderator variables are likely to have relatively low statistical power. Compared to the experimental situation where both variables are manipulated, there is likely to be less variation in the independent and moderator variables in the case where these variables are measured as they naturally occur. As a result the residual variation in their product is likely to be particularly small compared with the situation where both variables are manipulated. Tests of moderation (and interaction more generally) might then be most powerfully conducted by selecting units with relatively extreme values on the two variables (McClelland and Judd, 1993).

Interpretive Ambiguities

As the examples in Section Examples of Hypotheses Involving Moderation illustrated, moderation hypotheses are ubiquitous in the social and behavioral sciences and analytic procedures for testing interactions, as described in the previous paragraphs, are relatively well worked out. However, interpretive difficulties can arise in describing moderation unless strong assumptions can be made about the metric of the dependent variable in any analysis. Nonlinear transformations of the dependent variable may eliminate interactions, especially when those interactions involve moderation of a main effect of an independent variable, rather than a crossover interaction.

Consider a simple example from cognitive psychology. Participants are timed while identifying letter strings as words or not (a lexical decision task). Words are either preceded by a semantically related prime or not. The main hypothesis is that primes should decrease the latency of the lexical response. Target words vary in their familiarity or frequency in the language. The effect of primes is expected to be moderated by word familiarity, such that prime effects are larger for less familiar target words. Hypothetical latency means (in seconds) are given in the left-hand portion of **Table 1**. Evidence is found for moderation: the prime versus no prime difference is more than twice as large for unfamiliar words (0.033 s) as it is for familiar ones (0.015 s). But now consider what happens if we convert the means to represent speed of response (rather than its latency) by taking the inverse of the latencies. In the speed metric, there is absolutely no evidence for moderation due to familiarity.

Which is the correct analysis? What conclusion about moderation is correct? Unfortunately, since there is no way to define the more appropriate response metric, there is no way to answer this question. The lesson is that many interactions in the social and behavioral sciences, where the metrics of the variables are somewhat arbitrary, disappear under nonlinear scale transformations. This is particularly likely to be the case when interactions are not crossover, as in the case of moderation.

Table 1 Means illustrating scale transformation that eliminates moderation

	Latency metric (tenths of seconds)		*Speed metric (inverse transformation)*	
	No prime	*Prime*	*No prime*	*Prime*
Familiar	0.556	0.541	0.0180	0.0185
Unfamiliar	0.833	0.800	0.0120	0.0125

To be confident about the presence of moderation, given arbitrarily defined metrics of most dependent variables in the social and behavioral sciences, analyses should be conducted using a variety of different metrics. If interactions prove to be significant under a variety of different nonlinear transformations, then confidence concerning the presence of moderation increases.

See also: Causal Counterfactuals in Social Science Research; Control Variables in Research; Endogeneity: Methodology; Mechanism-Based Causal Analysis; Mediating Variable; Structural Equation Modeling.

Bibliography

Aguinis, H., 2004. Regression Analysis for Categorical Moderators. Guilford Press, New York.

Aiken, L.S., West, S.G., 1991. Multiple Regression: Testing and Interpreting Interactions. Sage, Newbury Park, CA.

Baron, R.M., Kenny, D.A., 1986. The moderator-mediator variable distinction in social psychological research. Journal of Personality and Social Psychology 51, 1173–1182.

Cohen, J., 1978. Partialed products are interactions; partialed powers are curve components. Psychological Bulletin 85, 858–866.

Fairchild, A.J., MacKinnon, D.P., 2009. A general model for testing mediation and moderation effects. Prevention Science 10, 87–99.

Henseler, J., Fassott, G., 2010. Testing moderating effects in PLS path models: an illustration of available procedures. In: Handbook of Partial Least Squares. Springer, New York, pp. 713–735.

Judd, C.M., Kenny, D.A., McClelland, G.H., 2001. Estimating and testing mediation and moderation in within-subject designs. Psychological Methods 6, 115.

Judd, C.M., McClelland, G.H., 1989. Data Analysis: A Model Comparison Approach. Harcourt, Brace, Jovanovich, San Diego, CA.

MacKinnon, D.P., Fairchild, A.J., Fritz, M.S., 2007. Mediation analysis. Annual Review of Psychology 58, 593–614.

McClelland, G.H., Judd, C.M., 1993. Statistical difficulties of detecting interactions and moderator effects. Psychological Bulletin 114, 376–390.

Muller, D., Judd, C.M., Yzerbyt, V.Y., 2005. When moderation is mediated and mediation is moderated. Journal of Personality and Social Psychology 89, 852–863.

Pearl, J., 2010. The foundations of causal inference. Sociological Methodology 40, 75–149.

Pearl, J., 2012. The causal mediation formula – a guide to the assessment of pathways and mechanisms. Prevention Science 13, 426–436.

Modern Management: Origins and Development

John-Christopher Spender, Kozminski University, Warszawa, Poland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Accounts of the origins and development of management have two deficiencies: (1) they fail to distinguish managing a business from other types of managing and (2) they are more often histories of writing about management than of doing it. Explaining management requires attention to the nature of the thing being managed – the firm. The history must hinge on specific aspects that distinguish today's management from that of former times; and various have been advanced – scale, technology, and globalization. Yet all were evident long ago and miss what is modern about modern management. This article treats the modernness of the mode of business administration as key to understanding the modern firm. Following Weber the bureaucratic mode is tempting but once we accept Knightian uncertainty, bureaucratic theory cannot be an adequate guide to today's managerial practice. Modern firms align employees' judgment rather than their blind obedience by providing them with a corporate citizenship that echoes but differs from that provided earlier by guilds. The historical origins and development of managing are of this transition.

The Chronology of Management

History stands on time, so management's history requires a time-contingent definition of management with moments or periods in which management changed markedly from managing past to managing present. If managing has been the same since time immemorial it has no history. If it is simply 'getting things done through other people,' it has a long history indeed, but with no clear transitions to support a historical analysis. The classic works on the history of management or of management thought presume the world is always changing and lay out a sequence of periods marked by different managerial focus (Witzel, 2011). Unfortunately too often this is a history of academic writing about management rather than a history of doing it. Many of the transitions are arbitrary and assigned to a particular time without adequate evidence, such as the oft-remarked transition from the 'inhumanity' of scientific management to the Mayo-inspired human relations movement. A glance at eighteenth- and nineteenth-century business texts shows considerable concern for human beings, albeit not in today's language (Freedley, 1879). At the same time these histories make little of the distinction between managing generally – e.g., in government, military, or the Church – versus managing firms, now more or less implicit in the term 'management.' Management schools are not much concerned with the public sector.

Even as we think of managing as something quintessentially modern and to do with firms we have difficulty determining the characteristic that marks this modernity and when it began. Many, influenced by Alfred Chandler's work, argue scale marks modern management's emergence, the rise of large corporations toward the end of the nineteenth century, especially in the US but also in Europe and Japan. Unfortunately scale does not provide a sound chronology. As economic historians know, large-scale economic organizations, such as plantations, potteries, or naval shipyards (de Vries and van der Woude, 1997) were widespread in the eighteenth century and evident many centuries before. Plus many criticize how Chandler prioritized managers' structuring decisions over responding to contextual changes (Rosen, 2013). Other 'modernists' stand on technology, seeing modern management as marked by the technological changes that precipitated the Industrial Revolution – as illustrated in Denis Diderot and Jean d'Alembert's *Encyclopédie* of 1751. Yet technological change is not a secure criterion either. We see the arrival of 'big data,' a century after Chandler's explanatory moment. If this is crucial, today's senior managers, nurtured in an earlier era, should be ill equipped to manage today's IT-penetrated firms. In fact they simply employ IT specialists to handle the new technology, just as their forebears managed coal mines from the surface or colonial trade from London through agents or employees. Hiring employees and agents with appropriate skills shields managers from external change. Others cite Fordism and scientific management as useful ways of distinguishing modern management from previous ways of 'getting things done through others.' Yet mass production goes back far beyond the Springfield rifle to, for instance, the mass production of ships and tackles in the fifteenth-century Venetian Arsenal.

A more persuasive chronology emerges if we think modern management is distinguished by its quantitative methods, the dominance of 'the numbers.' A growing literature argues modern management (and indeed modern society) is marked, even caused, by the expanding use of measures and statistics (Porter, 1995). Max Weber's work on the spread of rationalism is influential (Weber, 2003). Likewise Frederick Taylor's contributions to managerial accounting impacted twentieth-century management practice more than his infamous 'time and motion' studies. Accounting goes back beyond Roman times. William Petty (1623–87) used statistics while surveying the Irish territories occupied by Cromwell's armies, and to calculate taxes and the nation's wealth. Post-World War II advances extended accounting and measurement, but merely describe aspects of modern managing without defining it. In practice accountants, statisticians, and quantitative analysts are 'staff' employed to advise management as a practice that always calls for judgment beyond 'the numbers.' In the military tradition, Fayol distinguished 'line' from 'staff' and we can likewise distinguish analysis from the exercise of managerial judgment when the numbers are not conclusive, when choices remain after the analysis is complete. Note that when managers

employ others having the professional skills to shield them from changes in technology, accounting standards, laws, regulations, and so on they exercise judgment; indeed Frank Knight regarded choosing subordinates as the severest test of a manager's judgment (Knight, 1965).

For some this 'judgment lying beyond analysis' is the essence of managerial practice and thus of a viable chronology, suggesting today's professionalized forms of managerial judgment can be contrasted against the arbitrary or prescientific judgment of past times. As Khurana's analysis of the history of US management education revealed, at the end of the nineteenth century there were widespread moves around the world, to develop a 'science of managing' and establish management as a profession (Khurana, 2007; Spender, 2005). The history of engineering, for example, is of increasing professionalism based on growing scientific knowledge and practical expertise. Unfortunately management meets few of the criteria for a profession; there is no qualification or policing of entry or exit; no institutional body to judge members' conduct and, most importantly, no certified body of professional knowledge against which to measure training, qualification, conduct, growth, or change. Though many speak loosely of management as a profession, it is not like any of the 'real professions' – medicine, engineering, or law. While management has generated many professionalization initiatives, none have had great impact on managerial practice. After a century of aspiration and claim, plus the 1959 Foundation Reports, business schools have yet to develop the science necessary to underpin a profession and, in consequence, their role is less to provide professional training than to help students into the hiring pipeline. Ironically, many of the business heroes of the early twenty-first century, such as Bill Gates, Mark Zuckerberg, Steve Jobs, Richard Branson, Marissa Mayer, and Bernie Ecclestone, lack 'professional training.'

What can be said if we have no viable chronology of managing, no clear separation between management past, present, and, by implication, future? Errors here are to think the history of management is (1) of a rational practice or body of general laws that can be separated from the specifics of the historical context in which it is of value or (2) from the idiosyncratic judgment of the managerial actor. When looking for its history it is not useful to think of managing as a form of social physics, a science to be applied objectively in fully known situations, for we have made no such progress. Rather, the search for a history must be in the complementary German historical tradition, of treating managing as the application of a specific person's judgment in a specific situation to be judged good or bad in terms of the judgments being applied to the situation. Instead of a chronology of managerial practice that advances from irrational to rational, or from prescientific to professional, or providing a sequence of academic writing genres, we must create a chronology of managerial situations and actors, and of managers' judgments as these draw in their agency, their 'judgments beyond their analyses.'

An ancillary problem is that if every collision of actor and situation is unique there is little to be said. Historians, eschewing general laws, presume situations and actors are not entirely unique, that insights garnered from examining judgments in one situation can illuminate its application in others even as they are not determining. For example, one change might be that the nineteenth-century notion of standardization that drove the 'American system of manufacturing' could be widely applied to administrative work as well, setting the stage for bureaucratic planning and the training to support it. Though many writers presume firms are rationally designed and managed bureaucracies, itself an opinion with a history, few managers feel bureaucratic theory an adequate description of their practice, more an aspect or possible mode of managing. We might likewise point to the 'invention' of the limited liability corporate form in the seventeenth century. It clearly changed the context of managerial practice from the earlier partnership form, given it radically shifted financial risk. No doubt the firm's legal form indicates constraints to managing and many of today's forms such as S corporations or 501(c)(3) seem 'modern' in comparison to partnerships. Managing a family business or a partnership is not the same as managing a multinational or not-for-profit. Managing a global enterprise, especially in the Internet Age, is different from managing in the English East India Company's early years (Keay, 2010). In short, the chronology we seek is one of the changing contexts in which managers do whatever they do, for managing is not the exercise of a universal rationality; it calls for situation-specific judgments. Being contingent, it has a history that can be told in terms of the changing contexts of its practice.

The Changing Contexts of Management Practice

Perhaps the principal change in the context of managing over the last few centuries – even more than the changes in IT and other technologies – is the erosion of business's monopoly powers, in part because of the loss of government grants and charters and the growth of government intervention with antitrust legislation and regulations, and in part because of globalization, competition, and greater consumer choice. Without citing much historical evidence, many academics pronounce business has become 'hypercompetitive.' We also hear that managers' power within firms has changed from eighteenth-century Dickensian absolutism to the bargaining situation exemplified, perhaps, in German *Mittlestand* firms. Even if these changes have indeed occurred it does not follow that they have made managing more difficult, perhaps the opposite. As the economy develops an 'invisible hand' appears in the background, the *deus ex machina* in the business drama that takes charge when markets are 'efficient.' As information flows improve and transaction costs drop 'efficient markets' emerge, an ordering that is not the artifact of individuals, managers, or the state. Many argue market forces should be our socio-economy's dominating mode of ordering, buttressing our democratic politics and our ethics of independence. Management's role then narrows onto reading and following, as rationally as possible, markets they neither engineer nor command. Social and political power is downplayed and becomes suspect as 'interference' in efficient markets. So one history of managing is of moving our economic affairs into increasingly efficient markets in which rationality dominates judgment.

But managing a business is never just about markets. Adam Smith knew their importance but he also saw an enterprise as a context in which entrepreneurs and capital come together

with others (labor) in a place of work (land). His intuitions about market behavior were formalized into Marshall's 'scissors,' the familiar crossing of demand and supply curves, and this led on to the microeconomists' definition of managing as price and output setting. But ignoring market imperfections, microeconomists are still not able to address the creation of economic value. Smith thought differently and was not only interested in markets. Following Petty, he implied managers should focus on the learning and tools generated within the enterprise, for that ultimately drove the nation's wealth. If some believe managing under perfect competition was Adam Smith's hope; we know Hobbes's views were very different – order comes from an exercise of power. Managing is not only analysis; it is also an exercise of power and the creation of an economic order that does not emerge on its own. Consequently management practice lies in two contexts: (1) directing firms interacting through competitive markets and (2) exercising power over individuals (employees) so as to create order, predictability, and economic value within firms. Thus it has two histories – one oriented toward dealing with changes in markets, the other toward ordering within the firm's boundaries – and, of course, a history of how these two are brought together in practice. While managers have a degree of power in both contexts, the relevant modes of power and integration differ. Loosely speaking, efficient markets deny managers any significant 'market power'; but if markets were efficient they would not have much of a history. Real markets are such that managers must deal with their social and economic imperfections. They have a history because these imperfections change. The histories of commercial and financial markets are rich and fascinating, especially their increasing scope, articulation, and sophistication since the Dutch Republic, though it is important to appreciate most of the financial instruments necessary to today's capitalism were developed before the time of the Florentine bankers.

The markets for products and for labor are structured and paced differently. Labor law and regulation has supplemented workplace activism and unions. Managing has been changed by the redistribution of power within the firm and the development of more participative modes of control. Many regard this as the cutting edge of management change and thus to its history. But this presumes the nature of the firm is less problematic than it is. Today managing is about firms – but what are these? The remainder of this article focuses on the firm itself and its particular history. To anticipate its conclusion, we argue for a historical connection between the growth of the Dutch Republic – dubbed the 'First Modern Economy' – and the changes that demarcate the modern firm – because they transformed its creation. Setting up and sustaining the firm is management's preeminent task so changes here can be made the basis of a history of managing. Modern management, we argue, was 'invented' in the Netherlands around the middle of the sixteenth century. In spite of many later adaptations, driven by the changes in technology, globalization, social norms, corporate law, regulation, or market forces noted above, managing exhibits a core characteristic – its mode of coordinating those employees the firm relies on if it is to adapt and persist in an uncertain and changing world.

Theories of the Firm

The term 'firm' comes from the medieval Italian '*firma*,' the signature that committed partners to a business arrangement or obligation. A modern firm is more than a signature – and here begins the puzzle. As Ronald Coase pointed out, we have no 'theory of the firm' to underpin our distinction between firms, individuals, and markets and, thereby, our notions of managing within the firm (Coase, 1991). What is a firm? Why do they exist? What is their history? Legal scholars have been debating these matters for centuries; even the Roman law on which modern corporate law is based was ambiguous. Today legal scholars see two possibilities; that the firm is (1) an 'unnatural person' with some of the rights citizens have to own property, to sue and be sued, and to contract or (2) it is a managed aggregation of people and assets (Veldman and Parker, 2012). The first frames managing as dealing with markets, the second as dealing with internal matters.

A viable theory of the firm would define or at least clarify this, and so make a history of managing firms possible. Organization theorists and microeconomists have a handful of notions, though the most accepted derive from Weber's theory of bureaucracy, driven by rational choice. The firm is defined as a machinelike apparatus of linked work-roles designed to achieve specified goals. It follows that managing embraces selecting objectives, designing roles, hiring into them, resourcing the activity therein, instructing the work, measuring it, and so on. In 1937, Gulick labeled this analysis POSDCORB – planning, organizing, staffing, directing, coordinating, reporting, and budgeting. Other noneconomic theories of the firm were gathered in Morgan's *Images of Organization*; mechanism, organism, brain, culture, political system, psychic prison, flux, and domination (Morgan, 1997). Each indicates a distinct notion of managing and each has a history. Note management's role varies from considerable to trivial, great in the case of design and control versus marginal when the manager is merely the instrument of the firm's self-organizing process. Just as the invisible hand appears to make managing across markets no more than reading the market's facts, so management within the firm becomes less relevant if the firm is self-organizing. Managers, of course, presume their importance to the firm and its activities, one reason why design and control is their favored metaphor; the history of managing cannot be of their unimportance! Note Morgan's metaphors fall into two classes, focused on managing resources or on managing people. One of the charges against scientific management is that it proposed people were no more than tangible economic resources, so drawing them into mechanistic models.

Paralleling Morgan's metaphors are a handful of microeconomic models of the firm; transaction cost analysis, principal-agent theory, nexus of contracts, property rights, team production, etc. (Spender, 2014). Again these fall into two classes, focused on resources (and prices) or on people (behavior), and they carry implications for managing. The microeconomist's emphasis is on Coase's questions: Why do firms exist? What are their boundaries, where they are? Why are their internal arrangements as they are? Why is their performance so varied? Inasmuch as managing is about 'going beyond analysis,' it is about answering these questions about firms and managing them that economic theory cannot.

Transaction costs economists argue firms exist because they can achieve production costs lower than the prevailing market prices, the make or buy decision in new terminology. But managers must create the lower cost regime before the make or buy decision can be evaluated. Principal-agent economists argue firms exist because they provide principals with a means of controlling agents through incentives and monitoring. Again, managers must create regimes that control divergent interests and knowledge. Perhaps firms exist as a nexus of management-initiated contracts between the firm and its providers and consumers, or because they can hold and protect property rights, or provide a legally structured entity to hold the undistributed results of team production, all management tasks. Douglass North also argued institutions can be understood as responses to social uncertainties. Because a society's uncertainties are contingent on their historical context – there being no overarching theory of social uncertainties – the meaning and nature of the resulting institution lies in the context's history, not in any general principles. Without a 'theory of the context' management's responses cannot be comprehended separately from the specifics of their context; so managing is 'indexical' rather than generic and a history of managing based on imperfections would be hopelessly fragmented.

Looking inwards at the modes of power used to construct and sustain firms suggests a greater degree of generality. Sociologists see three fundamental modes of power: calculation, coercion, and culture. Markets are largely coordinated by calculation. Firms are nonmarkets precisely because they are domains of power, not calculation. Thus the long-term history of managing may be of socioeconomies moving away from coercion (feudal) as the dominant mode of creating social order and toward, first, more efficient markets, and when they are not available, to noncoercive 'cultural' modes of coordination. Coercion is common, of course, and has a history as long as society itself. It remains a puzzle for management theorists; Do threats discourage employees from contributing their judgment? The modern assumption is 'yes,' so culture management lies at the heart of contemporary analyses of managing. Within their firm, managers create and institutionalize a culture of purposive conformance (alignment) that is accepted by those involved as 'managerial authority.' As a voluntaristic order lying between the individual actor and social structures such as the state, a firm's culture implies a collective consciousness or shared sense of objective that shapes the way the firm's individuals dispose their energies and agency toward the firm's objectives (Spender, 2014).

The Modern Firm

Modern management reflects what is modern about our era, especially the need to persuade rather than coerce skillful and politically enfranchised citizens into accepting the entrepreneurs' judgments. The firm is a domain in which employees are persuaded into alignment with the entrepreneur's goals. How does this work? It is generally accepted that modernity began in Florence, initiating the fourteenth-century Renaissance; its principal characteristics being (1) the spread of scientific thinking, objectivity, communication, and empirical methods and (2) the rise of individualism, self-awareness, and freedom from premodern forms of social order (primarily feudal and/or religious). Others argue the plague of 1348 (Black Death) greatly reduced worker populations and precipitated new economic relationships. Citizens acquired rights to buy and sell as they chose, and eventually to establish their own value-creating activities, both economic and artistic, breaking free of feudal and institutional arrangements such as guilds that previously gripped the economy. The bans on assembly, itself regarded as a threat to feudal leaders or the state, were reduced, legitimating individual 'projectors' or 'master adventurers' shaping the economic activity of others. Their freedom to set economic goals was complemented by the evolution of capital markets and corresponding legal and financial instruments, such as rights of incorporation. Entrepreneurs could form entities that could absorb and apply the capital of others. The political, legal, and religious upheavals across Europe also freed many individuals to travel beyond the towns and villages in which they were nurtured, leading to viable labor markets.

The preconditions for an explosion of economic activity were in place in the sixteenth-century Netherlands; religious freedom after the revolt and repulse of the Spanish occupiers, a good supply of labor, skilled immigrants of every variety fleeing religious persecution all over Europe, educational facilities that led to the population becoming the world's most literate. There were substantial technological advances in windmills, textiles, pumps, fish-salting, and intensive agriculture. There was the *waterschap* or *hoogheemraadschap* local democratic political tradition born of the collaboration necessary to create and sustain the dykes and polders. Export trade depended on ships, and the Dutch mastered both designs and methods of production to produce the most economic crafts (*fluyts*). Obviously these preconditions did not cause the evolution of the modern firm, but they comprised a singularly fruitful context for those striving for better socioeconomic arrangements than those of the previous era. It is difficult for us to appreciate how these people thought about economic activity for we are children of a rationalist age (Lecky, 1910). We find the metaphor of the mechanical firm completely acceptable. But to suggest the Renaissance and Enlightenment movements helped an already comprehended economic rationalism emerge from under the crushing weight of feudal and religious ideology is to deny the history of these changes. This is 'Whig history,' a story of how society overcomes impediments to achieve its present perfection. In practice, the economic potential of the division of labor and efficient markets were not self-evident to those engaged in commerce; nor were they knowingly suppressed by those whose interests would be harmed. When Smith was writing, following the work of Petty and the French economists, he was exploring the political and economic implications of then radically new ideas.

Not assuming everything is progressing toward perfection, non-Whig historians work as best as they can from what they know of the past and thence into the present, identifying the phases of change. Likewise we have no crystal ball to tell us how managing might change in the future when, for instance, we might be under constant surveillance. The Netherlanders showed how acculturation and persuasion characterize today's modern firm – in contrast to the coercive and feudal modes of

power that characterized the previous era's manufactories. Understanding that the presence of uncertainty makes judgment necessary to all practice, it also means employees cannot be persuaded by rational argument alone for neither they nor their managers can foresee the consequences of any particular instruction. Note that Coase suggested the nature of the firm sprang from the 'supersession of the price mechanism,' and the employees' willingness to obey the direction of managers 'within certain limits'; the first makes monetary incentives problematic, the second shifts attention onto the processes managers use as they persuade others to do their bidding.

This opens up many questions about firms and managing, but even without resolving these, history provides suggestions about how the modern firm emerged. While responding to changes in technology, finance, and trading will always be important, managing the modern firm is defined by this novel – for the 1550s – mode of administration, cultural rather than either coercive or familial. As Weber noted, economic expansion was limited until entrepreneurs found ways of acquiring human capital beyond the limits of the family, paralleling Penrose's insight that a firm's growth is limited by the rate at which its 'management team' can learn (Penrose, 1995). Crucially, the implication was that employees must give up part of the citizenship rights into which they were newly enfranchised. Thus the modern firm is not an inhuman machine but, more importantly, it offers an alternative mode of citizenship – not inevitably welcomed, as in Whyte's *The Organization Man* (Whyte, 1956). Again Coase's suggestions are revealing. He implied the existence of a 'special class that directs the activities of others' and, parenthetically, of 'those who might prefer to work under another's direction.' The result is an embedded contrast of master and servant or leaders and led. Its ethical and political implications, interesting as they are, do not explain why those who had recently achieved citizenship would voluntarily lessen those rights by moving into a firm's field of power. To argue this is for economic reasons (personal gain) denies Coase's discussion of the frictions or costs of all economic activity (transactions costs). A firm cannot survive if its employees earn precisely what they contribute – even allowing for a just return to capital. Thus as Veblen argued, firms exist because the rewards to investors and employees are less than the value of their contributions (Veblen, 1965). How come this paradox became so widely institutionalized and remains uncontested? This is management's history.

The Guilds

While the Dutch Republic advanced freedom of religion and the rule of law, enhancing political citizenship in ways that were subsequently articulated into the US Bill of Rights and Constitution, the variations on citizenship were more complex than they seem today. Slavery and bondage proliferated (Steinfeld, 1991). Apprenticeship was common. Women's rights differed from men's. Guild membership was especially important to the evolving notion of the commercial firm. Guilds had existed since Roman times but became popular throughout Europe as economic activity picked up in medieval times. Adam Smith criticized them for restraining economic development in ways that anticipated Taylor's distaste for unions. The guilds' eventual decline mirrors the growth of firms as alternative modes of economic organization. We can probe the nature of the firm by considering features of guild life no longer present today. But again, our current modes of economic thought and ideology of rationalism may cloud our sense of why guilds existed. They were deeply intertwined with commerce but were not rational-legal in Weber's sense. The argument they arose to protect craftsmen from competition or for purely economic reasons is to project our thinking backwards into those times. Likewise evidence they resisted technological and legal changes that would improve economic efficiency is balanced by evidence that they often pursued explicitly economic ends. Guilds were clearly involved in managing a trade's intellectual capital by restricting both sharing among members and the rights of nonmembers to practice the craft or enter the trade. But this does not get to the heart of their nature.

There were many reasons why people wanted to become guild members (Epstein and Prak, 2008; Kieser, 1989; Lucassen and Prak, 1998). The guilds' primary *raison d'être* was not protectionism, it was to provide members with social and economic identity, a reputation, just as professionals claim today when identifying as engineers, academics, or whatever, taking their individualism for granted. Sixteenth-century people's identity was more deeply grounded in relationships – family lineage, feudal belonging – and guild membership. The nakedness of today's individualism would have been terrifying to all but the most radical thinkers (and believers). Plus while cash wages were increasingly displacing 'payment in kind,' medieval economic activity was tied up with creditworthiness – to reassure landlords, victuallers, and tailors – and those who supplied artisans with their factors of production; as credit and working capital does today. Amid the Dutch Republic's turbulence an immigrant's work would be unvalidated. Guilds accredited it and made the work tradable, but only after the novitiate went through the guild's entry process and committed to its quality and work policing. Guilds provided care for sick and unemployed members and families, making membership a long-term obligation. Guilds similarly established traders' credentials, for letter of introduction were essential for nonfamily members. Given the times, most guilds were allied with religious organizations, deploying religious ideas to enhance reputations, often becoming religious schools. Guild membership was everything, offering an economic and socially viable citizenship that differed markedly from the problematic social condition of nonmembers.

Guilds provided a citizenship that predated national citizenship and prefigured 'corporate citizenship' today, the modified state of sociopolitical identity Coase noted. The essence of sixteenth-century Dutch managing lay in creating, exploiting, and sustaining a new kind of citizenship, different from that the guilds offered (Sison, 2011). Employment law is such that the nature and boundaries of corporate citizenship remain far from clear today, in part because of the complexities but also because all law is a 'work-in-progress.' Today's employees sign open-ended or 'incomplete' contracts that leave management considerable latitude in determining the employee's eventual duties. Employment today is 'at will' and the fact that employees do not have to take oaths as guild

members did – but as military and religious participants still do – implies a radically more fluid relationship. There are no laws to force an employee to contribute her/his labor. Thus having abandoned the coercive tools available (still) to military and religious organizations, the modern firm is bound together very differently, by the cultural effects of managerial persuasion.

The history of how Dutch entrepreneurs were able to leverage the in-place legal, political, and social norms that legitimated a guild-based (and constrained) citizenship into a modern form of corporate citizenship – and so create today's more purely economic entity – has yet to be fully worked out. Note the oft mentioned *Vereenigde Oostindische Compagnie* – founded in 1602 – is not a good example, given it evolved rapidly into a military and governing agency for the new Dutch Republic's colonizing. Joining it required an oath of allegiance. The focus here is on the multiple small firms (which we now call 'small and medium enterprises') the burghers founded that drove the Dutch Republic's emerging economy (de Vries and van der Woude, 1997). The historical break between the guilds and the modern firm is neither technological nor financial, for these means and instruments were already widely available; its essence lies in its commitment to a novel mode of administrative power. Part of the history of guilds is their eventual erosion as states became more powerful, for states had interests in economic growth the guilds did not. States waged war and were the populace's ultimate custodians, especially of the poor, indigent, or criminal. States needed – and still need – cash more than the guilds did, for the guilds' sights were on membership and honor rather than on international financial and political power. So perhaps the story here, as Smith suggested, is that these new firms' emerging economic power and greater economic efficiency helped them force themselves into the nation's political economics and displace guilds as inferior wealth creators.

While many wax nostalgic and argue today's firms should provide a richer cultural locus in which 'trust' and 'ethics' abound, as if they ever did in a bygone Golden Age, the modern firm is closer to being an apparatus for maximizing shareholder value that regards every factor of production and every participant as tradable, replaceable, and membership of passing consequence. The modern firm abjures the human and social obligations that marked the guild. It has become increasingly personified at law, albeit unnaturally, and seized on reputation and brand as significant self-images; employees are correspondingly depersonified, useful only for the services they provide, no longer recognized as long-term obligations. The guild participants' concern with reputation and honor was displaced by prioritizing shareholders' and managers' economic interests, all part of the rise of rational man, the self-maximizing isolate the Enlightenment ushered into the modern world, the antithesis of a 'state of honor.' Likewise the evolution of statistics (numbers in the service of the state) has transformed our citizenship, powered by improved record keeping and surveillance. The concomitant changes in our political democracy, especially in neoliberal thought, clearly buttress individualism at the expense of community. Khurana observed a similar dynamic in the US as the notion of managerial professionalism shifted from social duty to managerial and shareholder capitalism (Khurana, 2007; Spender, 2007). In short, the modern firm coevolved along with the turn to rationalist ideology and the modern surveillance state; they are made for each other. The history of managing is not a Whiggish journey toward economic perfection; rather it is the contingent result of today's norms of corporate citizenship as they engage today's growth-intensive socioeconomics.

Concluding Comments

Managing is only relevant to entities (firms) that are not self-managing – as self-organizing systems and efficient markets are thought to be. Managing can only be understood through the lens of a concept or theory of firms integrating the factors of production available in real markets into outputs that can be sold into real markets. The inputs and outputs are heterogeneous and of wide variety; land, labor, and capital, perhaps, or technology, finance, and employees' judgment contributions. Each factor implies a history of managing as integrative. The trickiest aspect is managing people into uncertainty-resolving judgments that generate appropriable economic value. Managing means getting specific individuals to engage specific uncertainties as they contribute their labor and judgment in ways that cannot be fully rewarded. Firms are therefore singular and indexical, and only exist and survive because they are, in part, an apparatus of social and economic exploitation as well as of economic value-creation. They are profoundly political entities so managing is a discrete political activity embedded within open zones in democratic capitalism. The history of the modern firm and the forms of management that create and sustain it can be explored through the notion of corporate citizenship and how that differs from other and previous forms of citizenship.

The complexities of managing can never be fully clarified because they arise at the collision of uncertainty and the firm's complex of social, normative, legal, technological, and other constraints (Spender, 2014). But today's corporate citizenship can be illuminated by contrasting it with that offered by the guilds, so treating guilds as historically transitional socioeconomic institutions that provided some of the modern firm's foundational principles and character. Overall then, the evolution of the modern firm is part of an interlocking history of ideological, political, legal, and technological change that led to (1) the 'invention' of the firm as a legal but unnatural person with political, legal, and social rights that differed from citizen's; (2) the legitimation of the subordination of individual rights to the firm's managers; and (3) the state as an increasingly rational-legal entity to which all citizens, individuals, or corporates, ultimately owe allegiance.

See also: Boundaries and New Organization Forms; Business Education; Business Models; Economic Geography; Industrial Geography; Industrial Relations, History of; Modernization and Modernity in History; Obedience (Research), History of; Strategic Management; White-Collar Workers, History of; Working Classes, History of.

Bibliography

Coase, Ronald H., 1991/1937. The Nature of the firm. In: Williamson, Oliver E., Winter, Sidney G. (Eds.), The Nature of the Firm; Origins, Evolution and Development. Oxford University Press, New York, pp. 18–33.

Epstein, Stephan R., Prak, Maarten (Eds.), 2008. Guilds, Innovation, and the European Economy, 1400–1800. Cambridge University Press, Cambridge.
Freedley, Edwin T., 1879. Common Sense in Business: Practical Answers to Practical Questions on the True Principles and Laws of Success in Farming, Manufactures, Speculation, and Buying and Selling Merchansdise. Claxton, Remsen & Haffelfinger, Philadephia, PA.
Keay, John, 2010. The Honourable Company. HarperCollins, New York.
Khurana, Rakesh, 2007. From Higher Aims to Hired Hands: The Social Transformation of American Business Schools and the Unfulfilled Promise of Management as a Profession. Princeton University Press, Princeton, NJ.
Kieser, Alfred, 1989. Organizational, institutional, and societal evolution: medieval craft guilds and the genesis of formal organizations. Administrative Science Quarterly 34, 540–564.
Knight, Frank H., 1965. Risk, Uncertainty and Profit. Harper & Row, New York.
Lecky, William Edward H., 1910. History of the Rise and Influence of the Spirit of Rationalism in Europe. Longmans, Green, & Co., London.
Lucassen, Jan, Prak, Maarten, 1998. Guilds and society in the dutch republic (16th–18th centuries). In: Sevilla (Ed.), Guilds, Economy and Society, pp. 63–77.
Morgan, Gareth, 1997. Images of Organization, new ed. Sage, Thousand Oaks, CA.
Penrose, Edith T., 1995. The Theory of the Growth of the Firm, third ed. Oxford University Press, New York.
Porter, Theodore M., 1995. Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton University Press, Princeton, NJ.
Rosen, Christine M., 2013. What is business history? Enterprise & Society 14 (3), 475–485.
Sison, Alejo José G., 2011. Aristotelian citizenship and corporate citizenship: who is a citizen of the corporate Polis? Journal of Business Ethics 100 (1), 3–9.
Spender, J.-C., 2005. Speaking about management education: some history of the search for legitimacy and the ownership and control of management knowledge. Management Decision incorporating the Journal of Management History 43 (10), 1282–1292.
Spender, J.-C., 2007. Management as a regulated profession: an essay. Journal of Management Inquiry 16 (1), 32–42.
Spender, J.-C., 2014. Business Strategy: Managing Uncertainty, Opportunity, and Enterprise. Oxford University Press, Oxford.
Steinfeld, Robert J., 1991. The Invention of Free Labor: The Employment Relations in English and American Law and Culture, 1350–1870. University of North Carolina Press, Chapel Hill, NC.
de Vries, Jan, van der Woude, Ad, 1997. The First Modern Economy: Success, Failure, and Perseverance of the Dutch Economy 1500–1815. Cambridge University Press, Cambridge.
Veblen, Thorstein, 1965. The Theory of the Business Enterprise. Augustus M. Kelley, New York.
Veldman, Jeroen, Parker, Martin, 2012. Specters, Inc.: the elusive basis of the corporation. Business & Society Review 117 (4), 413–441.
Weber, Max, 2003. The History of Commercial Partnerships in the Middle Ages (Lutz Kaelber, Trans.). Rowan & Littlefield, Lanham, MD.
Whyte, William H., 1956. The Organization Man. Doubleday & Co, New York.
Witzel, Morgen, 2011. A History of Management Thought. Routledge, Abingdon, Oxon.

Modernity

Bernard Valade, Paris Descartes University, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article attempts to clear away the confusion that has surrounded the notion of modernity. The primacy of the aesthetic dimension is recognized when considering the characteristics and attractiveness of modernity. The display of the historical process usually linked to the notion of modernity is followed by a specificity analysis of this polysemous and complex concept, attracting numerous contradictions and ambivalent views about the writings of Charles Baudelaire, Walter Benjamin, and Robert Musil. The relationship between sociology and modernity also is explored, with particular reference to the work of Georg Simmel.

If it is true that a notion is intuitive knowledge (*see* Knowledge, Sociology of), synthetic and inaccurate enough about one thing, then modernity belongs to this type of mental representation that, as opposed to concept, does not offer clearly defined contours of the abstract object to which it refers. We commonly agree on what must be understood by 'modern': that which appears, exists, or belongs to the current era or to a recent period, whereas 'modernism' expresses a preference against all traditionalism (*see* Sociological Traditions). Economists, sociologists, and development experts also know what 'modernization' (*see* Modernization, Sociological Theories of) means when expressed in terms of politics and equipment plans. They agree with Jean Fourastié that, in the 1920s, to modernize a factory meant to install machines there and to submit the workers to their rhythms, whereas now "it is to introduce into it a new working method, based, above all, on the means and faculties of men" (1969: p. 354). All those who deal with religious sciences (*see* Religion, Sociology of) know about the crisis of 'modernism', which condenses the doctrines and movements advocating renewal and adaptation of Catholic theology (*see* Catholicism) in relation to modern scientific discoveries and the general evolution of the world at the end of the nineteenth century and at the beginning of the twentieth century. Finally, nobody can ignore that modern style, 1900 style, and 'noodle' style are composed of floral ornaments – intertwining iris, Virginia creeper, and plane tree or chestnut tree leaves.

Numerous semantic samples, which relate to the modern or to what it specifies, could be attached to modernity. In particular, we should mention those that include ideas of custom, of *avant-garde*, and of doubt, as well as the refusal of traditional values (*see* Values, Sociology of). The complexity – paradoxically, as it would itself become the challenge – of modernity lies in the multiplicity of levels in which it is unfurled and in the uncertainty of the temporal framework that we tend to attribute to it. It appears to show solidarity toward 'progress' made in the field of science and technology; when applied to arts and culture, it can be conceived as being free from established rules, released from former achievements, and of new creation, original, unpublished. It is possible that modernity saw the light of day in the sixteenth century: Did not a French historian entitle one of his works *La Modernité du XVIe siècle* (*The Modernity of the Sixteenth Century*)? Leaving the modern period for the contemporary era, modernity emerges in the nineteenth century, when Chateaubriand created the term in a text dated 19 May 1833 and entitled *Mémoires d'Outre-tombe*, and in which he recounted that, on the outskirts of a town on the Danube,

> Vulgarity and modernity of customs and passports were contrasted with storm, the gothic door, the sound of a horn and the noise of a torrent.

It also is useful, from the beginning, to clear away the confusion that has surrounded the notion of modernity. We therefore recognize the primacy of the aesthetic (*see* Aesthetic Education) dimension that, at first sight, is nothing more than a rather vague constellation of ideas. Seen in this way, one might wonder how sociology should approach the subject, and which works should be consulted when we intend to examine its relationship with modernity.

The Characteristics of Modernity

Historically, modernity has a double origin: the Renaissance (*see* Renaissance) and the Reform (*see* Reformation and Confessionalization). From the beginning, it was the demonstration of a crisis, the symptom of a transformation – and in no way its 'analyzer.' It must be recognized that Jean Baudrillard ably demonstrated that modernity is not an analytical concept, but rather, by converting the crisis into a value, it is a form of regulation – and first and foremost, a way toward civilization (*see* Civilization, Concept and History of) – supporting a new mythology. It is not enough to say that the latter is centered on change; the ideological contents must be defined: These include categorical requirements that institute the 'tradition of novelty.' This argument owes nothing to Christopher Colombus, or to Galileo Galilei, and very little to Luther and Calvin. The idea of 'rupture,' to which modernity provides the rhetoric, however, is associated with the great discoveries, with the passage from geocentricism to heliocentricism, and to the theses of Wittenberg.

The development of this rhetoric continues through a series of processes: World rationalization, secularization (*see* Secularization) of thought, loss of one's fondness for community links (*see* Community, Social Context of: The US Case), and

assertion of individualism (*see* Individualism versus Collectivism: Philosophical Aspects). In France, we are aware of the strong moments of this movement. First, we have the quarrel of the ancient and the modern, in which the champions were, on the one hand, Boileau, Racine, Bossuet, La Bruyère, and, on the other, Quinault, Saint-Évremond, and Fontenelle. The last of these was the author of *Digression sur les Anciens et les Modernes*, which appeared during the same year (1688) that Perrault began his compilation of *Parallèles des Anciens et des Modernes* (1688–92). Descartes had already affirmed that there is no reason

> …to bow down before the Ancient because of their great age, it is more ourselves who should be called Ancient. The world is now older than before and we have a greater experience of things.

We see that the old image of dwarves perched on the shoulders of giants has reappeared. In fact, it was not only a question of specific literary value within the works of antiquity. The opposition was especially interested in the effectiveness of progress of intelligence and morality: The ancient remained attached to the idea of an original and 'natural' simplicity, whereas the modern believed in the beneficial effects of civilization (*see* Civilization, Concept and History of).

Another important step was the period known as the *Aufklärung*, the Enlightenment (*see* Enlightenment), the *Siècle des Lumières*, which in France was the setting for the adventure of the encyclopedia (*see* Encyclopedias, Handbooks, and Dictionaries). In a world plunging into obscurity because of the effacing of religious signals, there emerged a willingness to maintain an illuminated center – and to assume the heritage – far from the models that merely had to be imitated. 'Modernity' provided support for ideas of progress and freedom: *Laisser faire (les hommes), laisser passer (les marchandises)*: Let them do (men), let them pass (marchandises This injunction coined by Vincent de Gournay (1712–59), became the motto of economical liberalism (*see* Liberalism: Historical Aspects), an expression of modernity. Free examination, free expansion, free pleasure in fact cumulated their impact at the time of the Enlightenment, but it was not until the next century, with the continuation of the Democratic and Industrial Revolutions, that modernity acquired its full political and economic significance.

Without a doubt, modernity is the inseparable companion of industrialization (*see* Industrialization), of urbanization (*see* Urbanization; Development and Urbanization), and of the introduction of regular salaries, which characterize a productive society wholly placed under the sign of the division of work. It also is an integral part of a progressive movement toward a mass consumer society based on standardization. The new way of life that it engenders – of which dissociation of work place and living environment, as observed by Max Weber (*see* Weber, Max (1864–1920)), is just one aspect – corresponds to the organization of the 'modern' state, the centralized, the bureaucratic state (*see* Bureaucratization and Bureaucracy, History of), and the promotion of the individual as an abstraction, which in turn creates the combination, in the universal derealization exposed by Marx (*see* Marx, Karl (1818–83)), of the rationality of the state – henceforth, equipped with a constitution – and of private interests, which are linked solidly to ownership. The concepts we have just discussed are not, objectively speaking, the most essential part of what we are seeking to understand. This can be found in the anxiety that overtakes the individual freed from the old magic-religious mechanism and that will, in the end, only be alienated by new networks of mediation. Mobility has reached a state of subjectivity and has rendered the identity of the subject uncertain, so that from then onward, it will be predisposed to instability.

The Attractiveness of Modernity

If we see modernity man installed in a time-specific, historical era, that is to say in a new temporality that apparently excludes eternity, we again are making him the vector or the product of an ideology of rupture. Within a cultural context where everything seems to be contemporary, current, and immediate, contradictions follow on from each other and opposites unite – for example, standardization and the search for singularity, social integration and irruption of subjectivity, submission (to norms) and subversion (of forms), and so on. With the emergence of *avant-garde* (*see* Avant-Garde Art and Artists) – that is, the aggressive emancipation of models, which had been considered legitimate until then – the destruction of traditional frames and perfect examples, we are nevertheless passing from a period of modernity viewed as a rhetoric of rupture to an attractiveness for change.

In an incisive study, *Prosa der Moderne*, Peter Bürger (1988) perfectly demonstrated that

> …if it is true that philosophical thought is blocked because it is incapable of considering the catastrophe by which it is obsessed, it is time to discover the other medium which allows one to see clearly within ourselves, in other words the prose of modernity.

Thinking laterally, he also gave prominence to how literary knowledge and sociological analysis can cross paths. Among the authors of literary modernity whose writings have the power to take their own presuppositions as a theme, we must place Charles Baudelaire. We will not, however, refer to P. Bürger's (1988) subtle comments about the allegory of the poet of *Fleurs du mal*, we simply will recall how the critic, in *Curiosités esthétiques*, identified modernity by assigning, for the first time, the meaning that remains attached to it today.

In 1863, *Le Figaro* consecrated its famous chronicle to modernity, as seen through the street scenes drawn by Constantin Guys, 'the painter of modern life'. The quality that should exist in the heart of all creation is identified: it is 'movement'. 'And thus he goes, he runs, he seeks' and what is this 'modern' artist looking for? The goal he is aiming for being

> …other than the fugitive pleasure of circumstance (…), it is, for him, to release from customs what they can contain as poetical in the historical, to extract the eternal from the transitory.

Modernity "is the transitory, the fugitive, the contingent, the half of art whose other half is eternal and unchanging."

Baudelaire (1869) It is the resolution of a conflict, the reduction of tension between the ever-suspect outpouring of emotion and the unjustifiable steadiness of eternal beauty, the reconciliation – thanks to poetical works – of the requirement for emotional authenticity (via the insertion of work in the evanescence of daily life) and of a desire for eternity (which is satisfied by the freeing of the invariable from the contingent).

We clearly can detect the reasons for Baudelaire's strong delight in painting and music, and especially for the canvasses of clashing contrasts by Delacroix and the swelling music of the Wagnerian universe. There is an energy that gives dynamism to these works; it jostles and disrupts the emotional state of the person who contemplates or listens to them. Combining passion and control, Delacroix remains detached from all imitation of nature: He is the 'suggestive' painter and "what he translates is the invisible, the impalpable, the dream, the nerves and the soul." On this subject, Baudelaire observes that Breughel, with his dreamlike chaos, and Goya, the painter of the monstrous probable, figure among the initiators of modernity. Modernity is therefore neither topicality, nor contemporaneousness; one is less fond of it at a given time. It becomes an orientation toward a future-past and a discernment in the era of the 'provisionally eternal'.

At the same time, we are experiencing the assumption of imaginary. Toward this – which dilates as the symbolic becomes tense – there opens a space on canvass and score. A constructive imagination, the first constituent, governs its display that does not end with the imagination–deceptions generated by fancy. With passing movements of decomposition and recomposition, the form of the beautiful symbol of modernity is sketched, made up of strange things and devices that render the beautiful the antonym of the real. The caricature of Daumier, the hair-splitting of the universal exhibitions, and the calculated affectation of the dandy are equally make-ups, but not masking, of the real. During the Salon of 1859, Baudelaire gave meaning to the lie that is modernity, when he wrote that, after expressing his enchantment for the theater decor that concerned his artistic dreams

> …these things, because they are false, are infinitely nearer to the truth, whereas most of our landscape painters are liars precisely because they forgot to lie.

We also can identify the reasons that led Walter Benjamin to see the passive stroller in the intoxicating crowd as the 'hero of modernity', in his studies dedicated to Charles Baudelaire, *Ein Lyriker im Zeitalter des Hochkapitalismus* (1938–39). This is surely the poet of modern society who, in the agitated throng of large towns, feels the loss of experience intensely: for which the instantaneous real-life experience of innumerable shocks was substituted, became permanent, and could not be assimilated. In this situation, which is characterized by a new relationship to time – about which modernity no longer recognizes the accomplishment, having lost all sense of recollection and festive moments – he can respond by miming death. He can also, like Baudelaire, search within the 'correspondences' for an experience that will be pursued and protected from, and far from, the modern clash. In a general way, the loss of the 'aura' will be converted into a poetical experience.

An earlier text by Walter Benjamin, *Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit* (1935), provided a definition for this tendency of modernity to be in constant opposition to uniqueness and distance. The masses' appropriate images multiplied by technique; a new, democratic, community is formed, in which the concepts of creation, genius, and style fall into disuse as we move from the value of worship to the value of expression. Growing reproducibility leads to the decline in ideas of authenticity and uniqueness, of the authority of the original, all condensed into the 'aura,' which becomes a theological survival. In this perspective, modern art trivially signifies that by as many people as possible having access to its products. With the texts united within the title *Das Passagen-Werk* (1927–40), Benjamin aimed for a deeper comprehension of modernity in light of historical and dialectical materialism (*see* Analytical Marxism). Taking as examples, arcades, universal exhibitions, and Baron Haussman's destruction and constructions, along with photography and the bourgeois interior, he made the ambiguous demonstration of the process triggered by the fetish for merchandise. The creation of Parisian arcades, covered passages lined with shops and boutiques with their window displays full of goods, is identified as the moment when merchandise and technique seem to be liberated from art to appear free of all artifice, but once again they are surrounded by borrowed elements. Art becomes technique and, from then on, is placed mainly at the service of the market. It is draped with dreams of the past and transfigures the defaults of the present into works that reveal the utopia of the future. Another art that became fashionable during the nineteenth century was phantasmagorical art, which consists of showing phantoms through optical illusions in a dark room, leading to the assimilation of modernity with a phantasmagorical universe in which signs free of all substance proliferate and where, at the same time, the primacy of form is asserted.

Criticism of Modernity

The aesthetic dimension of modernity and the society in which it is inscribed both have been examined by Georg Simmel (1989–90) (*see* Simmel, Georg (1858–1918)). In his opinion, the world of modernity was in opposition to that of classical culture (*see* Classical Legacies), which was considered to be orderly and containing hierarchical values. In this world, where custom was represented as the summit of the change to which it was perpetually dedicated, was born a new living place – the large town inhabited by masses of people – and a new style of behavior appears that was characterized by the modern individual. Individualism (*see* Individualism versus Collectivism: Philosophical Aspects), its dominant trait, was no longer seen as a universal archetype, as it was in the ancient Latin and Christian world. The individual was conceived from now on as a totality, which German romanticism had glorified. Similarly, tragedy was no longer, as in ancient times, seen as a confrontation with an external destiny, but interiorized, so that the individual was not only man but also victim and torturer.

In this configuration, which stems from the movements and the disorder that it designates, 'modernity' appears as a sum of transformations and contradictions that Simmel attempted to

disentangle. People could be placed under the sign of conflict, thus providing one of the most important chapters of *Soziologie* in 1908, and in particular were well portrayed in art and religion. To persist, it is required to become consciously but implacably subjectivewithin itself: This is considered as the stylization of life expressed through its liberation. Rodin was selected to be the representative of modern art and interpreter of modernity, insofar as his sculpture was the representation of movement – as an expression of the soul – in relation to a renewed conception of temporality. In the same way that time is an addition of separated instants, movement is composed of a succession of instantaneous postures. This idea leads to condemnation of mechanism and of naturalism (*see* Naturalism), to the assertion that art, because it does not have to bear the 'train of reality', does not have to copy reality. Art (*see* Art History) is an original conception of the aesthetically modern, which is no longer that of the object, but of the subject, with its moods that are expressed in a portrait or a landscape.

All the Simmelian digressions on copy and creation, on distance and proximity, and on the tragedy of culture and the intuition of life, are harmonized into a single philosophy of modernity, which cannot be separated from general sociology. Lilyane Deroche-Gurcel (1997) perfectly demonstrated their complementarity, beginning with an explanation of how aesthetic points of view and sociological analyzes come together for Simmel. Two dissatisfying conceptions of modernity in art are revealed by the author of *Rembrandt – Ein Kunstphilosophischer Versuch* (1916). One provides him with a matrix as its only formal principle: The beautiful works are the girls and their shape. The other makes it the immediate expression of life: The original creations are born from strict contact with reality. Going beyond the hieraticism of art for art and the outburst of exuberant expressionism, Simmel saw in the interaction of life and form – an interaction that had been accomplished by both Rembrandt and Rodin – the outcome of modernity. It is this conception that led him to general criticisms or criticisms of detail: instant photography, which can give only a fixed expression of a reality, although it pretends to show movement; naturalism (*see* Naturalism), which is constantly at work in the *avant-garde*, closed in by the childish rejection of tradition and the boundless search for novelty; and, finally, all dogmatism.

Taking into account the main characteristic of modernity, that is the multiplication of the interactions among individuals, the dissolution of stable contents into transient elements, Simmel concentrated on studying, not constituted society, but socialization. The substitution of movement for substantiality is in correlation with that of the *Vergesellschaftung* in the study of the *Gesellschaft*. It is accompanied by the abandoning of the system for function, of the structure for process, and of the status for role, behavior, and attitude. It supports the election of the fragment compared with the totality and of the form compared with the contents. In this respect, the sociology of Simmel, more than any other, has a hold on modernity that signifies, apart from the secession with stability, the loss of all substantial contents. Here, we can observe the relationship that exists between modernity and melancholy.

What is significant, from this point of view, is the analysis of indifference: Simmel thought of it in terms of melancholy, by observing within certain aspects of socialization (customs, coquetry, adventure, or sociability; the politeness of a detachment without despair) the resurging of melancholic syndromes within the species of their form – depersonalization and a-historicity – understood by their sociological meaning. Finally, Simmel underscored the ambivalence of modernity. Although movement in fact appears to be carried off in a flow of tormented forms, it in fact offers itself to the serene contemplation of the amateur who sees in the work of art of modernity, not the deceptive immobilization of life but rather its stylization.

Generally observed from above, modernity could just as well be observed from below, as fragile historical periodizations. Dedicated to postmodernity, the works of Colin Campbell (1987), David Harvey (1989), Zygmunt Bauman (1989, 1992, 1997), and Barry Smart (1992) draw in a hollow a representation of modernity. A Fordist method of production, a division of the society into classes and mass consummation would be the last traits to have been presented by modernity – finally discredited in and by totalitarian barbarism.

In this perspective of historical framing, we equally discuss a 'late modernity' of 'high modernity' (A. Giddens) or a 'liquid modernity' (Z. Baumann). If for other authors who used the concept of 'postmodernity' (J. Baudrillard, J.-F. Lyotard, M. Maffesoli), the change of era, social, and cultural frame are implicit, A. Giddens and Z. Bauman highlighted the richness of the concept of modernity by pointing out a late phase of the evolution of this historical period. The approach of the Canadian sociologist Charles Taylor shows a similar view. The author examined in a critical manner the connection between modernity and secularization (*see* Secularization) and looked for a solution for the new challenge of the modern identity. At the same time, the individual demands respect for his 'dignity' (concept replacing historically the one of 'honor') and for his 'uniqueness'. The concept of multiculturalism, examined in a critical manner, nevertheless is considered by the author an unquestionable reality, assumed by most of the societies. It becomes thus necessary to synthesize the concepts of multiculturalism, multiple identities, universalism trend, and authenticity recognition.

We must equally underline that the syntagm of 'late modernity' evokes the subtle analysis of 'late antiquity', by Peter Brown, and his predilection for Gibbon's work, in which the author has the eye of a comparative historian rather than the one of a decadence thinker – 'sociologist of the imperial system', knowing how to reveal the institutional change and the beginning of a new social setting.

The profound analyzes made by P. Brown on the style, institutions, creativity, and concept of 'sacred' in the late antiquity (*see* Ancient Greece and Rome, History of) could serve to analyze a new shift of approach in the modernity.

We could equally highlight that the idea of decadence, which seems to suddenly appear as a rejection of certain aspects of the modernity, manifests particularly vividly at the end of nineteenth century. The concepts of antique decadence and contemporary decadence entangle particularly in the work of J.-K. Huysmans (*A rebours*, 1884, *Là-bas*, 1891).

The postmodern era therefore would begin with the jamming of the effects generated by the rationalization of the world and would be followed by the coming of social

configurations that from then on would be commanded by new information and communication techniques (*see* Information Society). The deconstruction of what modernity edified therefore would be put into operation. In the field of urbanism and architecture, the cities of Brasilia and Chandigar show well what modernity meant during the twentieth century: a utopia of simplicity that does not recognize the complexity of life.

Retrospectively conceived in this way, modernity found itself deprived of many of its characteristics, which were transferred to the age that closed it down: the mix of the present with the past, of the new with the archaic, of cultural relics with technical gadgets. Subject to trial by a plurality of readings and interpretations and according to segments of severed time, modernity can less be a rupture than a response to this – by amalgamation. Above all, the postmodern point of view concerning modernity, whose horror is exposed, leaves behind the tragic that profoundly touched it. It is therefore a reducer, the better to reveal the return of complexity, which led to its adoption.

In fact, modernity man is seized by this complexity. It is Ulrich, alias Musil (1931) of *Der Mann ohne Eigenschaften*, the man 'without quality,' who was devoted to the juxtaposition of partial histories. Before the series of events crossed by the crisis of sense and before the difference between physical reality and psychic reality, which increasingly is being demonstrated, the question is this: How does one summarize it? How can we find the balance between the vital, the emotional, the affective and the control, the criticism, and the censure? Within the Kakania, Austria–Hungary mythical, modernity is an impossible synthesis. But it is also the uneasiness to which we are condemned as soon as, in the order of politics, a work interrupted by reduction of admiration or respect renders the calls for trust and esteem hardly audible and credible.

We should distinguish two registers within the study of modernity. The one that we will qualify as 'ideological' is the most frequented by sociologists. In the ideology of modernity, some see a conservatism via change, others see a spectacular ambiguity, and yet others see a culture of daily life. On another level, which, for want of a better term, we will call 'existential,' we must situate all that was produced with the weakening of subjectivity, with the assertion of individualists, with the perversity of human relationships, which are just as much aspects of modernity. It is at this level of analysis that we must position ourselves to examine the psychosociological consequences of self-constitution, according to which modernity was unveiled and which always will be, although in changing forms, the return of the same.

See also: Aesthetic Education; Ancient Greece and Rome, History of; Art History; Avant-Garde Art and Artists; Bureaucratization and Bureaucracy, History of; Catholicism; Civilization, Concept and History of; Classical Legacies; Community, Social Context of: The US Case; Development and Urbanization; Encyclopedias, Handbooks, and Dictionaries; Enlightenment; Individualism versus Collectivism: Philosophical Aspects; Industrialization; Information Society; Knowledge, Sociology of; Liberalism: Historical Aspects; Marx, Karl (1818–83); Modernization, Sociological Theories of; Naturalism; Reformation and Confessionalization; Religion, Sociology of; Renaissance; Secularization; Simmel, Georg (1858–1918); Sociological Traditions; Urbanization; Values, Sociology of; Weber, Max (1864–1920).

Bibliography

Baudrillard, J., 1968. Le Système des objets. Gallimard, Paris.
Baudrillard, J., 1970. La Société de consommation. Denoël, Paris.
Bauman, Z., 1989. Modernity and the Holocaust. Cornell University Press, Ithaca, NY.
Bauman, Z., 1992. Intimations of Postmodernity. Routledge, London.
Bauman, Z., 1997. Postmodernity and Its Discontents. Cambridge University Press, New York.
Bauman, Z., 2000. Liquid Modernity. Polity, Cambridge.
Bauman, Z., 2011. Culture in a Liquid Modern World. Polity, Cambridge.
Baudelaire, C., 1869. Curiosités esthétiques. In: Oeuvres de Charles Beaudelaire, third ed. Michel Lévy frères, Paris.
Bell, D., 1976. The Cultural Contradictions of Capitalism. Heinemann, London (1979 Les Contradictions culturelles du capitalisme. Presses Universitaires de France, Paris).
Benjamin, W., 1955, 1961, 1966. Schriften, Illuminationen, Angelus Novus. Suhrkamp, Frankfurt, Germany.
Benjamin, W., 1969. In: Tiedemann, R. (Ed.), Charles Baudelaire. Suhrkamp, Frankfurt, Germany.
Brown, P., 1978. The Making of the Late Antiquity. Harvard University Press, Cambridge, MA; London.
Bürger, P., 1988. Prosa der Moderne. Suhrkamp, Frankfurt, Germany.
Bürger, P., 2002. Theory of the Avant-Garde. University of Minnesota Press, Minneapolis (Theorie der Avantgarde, 1974).
Campbell, C., 1987. The Romantic Ethic and the Spirit of Modern Consumerism. Oxford, UK, London.
Compagnon, A., 1990. Les cinq paradoxes de la modernité. Ed. du Seuil, Paris.
Deroche-Gurcel, L., 1997. Simmel et la modernité. Presses Universitaires de France, Paris.
Deroche-Gurcel, L., 2000. Avant-propos and Préface. In: Simmel, G. (Ed.), Sociologie—Études sur les formes de la socialisation. Presses Universitaires de France, Paris.
Domenach, J.-M., 1986. Approches de la modernité. Ellipses, Paris.
Fourastié, J., 1949. Le Grand Espoir du XXe siècle. Presses Universitaires de France, Paris.
Giddens, A., 1994. Les Conséquences de la modernité. L'Harmattan, Paris.
Giddens, A., 1991. Modernity and Self-Identity: Self and Society in the Late Modern Age. Stanford University Press, Stanford, CA.
Habermas, J., 1985. Der philosophische Diskurs der Moderne. Suhrkamp, Frankfurt, Germany.
Harvey, D., 1989. The Condition of Postmodernity. Blackwell, Oxford, UK.
Hauser, H., 1930. La Modernité du XVIe siècle. Alcan, Paris.
Jameson, F., 1991. In: Fish, S. (Ed.), Post-Modernism – or The Cultural Logic of Late Capitalism. Duke University Press, Durham, NC.
Lipovetsky, G., 1987. L'Empire de l'éphémère. Gallimard, Paris.
Lyotard, J.-F., 1985. La Condition Post-moderne. Éditions de Minuit, Paris.
Maffesoli, M., 2000. L'Instant éternel. Denoël, Paris.
Morin, E., 1962. L'Esprit du temps–Essai sur la culture de masse. Grasset, Paris.
Musil, R., 1931. Der Mann ohne Eigenschaften, vol. 1. Rowohlt, Hamburg, Germany.
Rince, D., 1984. Baudelaire. In: de Beaumarchais, J.-P., Couty, D., Rey, A. (Eds.), Dictionnaire des Littératures de langue française. Bordas, Paris.
Simmel, G., 1988. Philosophie de la modernité. Payot, Paris.
Smart, B., 1992. Modern Conditions, Postmodern Controversies. Routledge, London.
Smart, B., 1993. Postmodernity. Routledge, London.
Taylor, C., 1989. Sources of the Self: The Making of the Modern Identity. Cambridge University Press, Cambridge.
Taylor, C., 1994. Le malaise de la modernité. Les Éditions du Cerf, Paris.
Touraine, A., 1992. Critique de la modernité. Fayard, Paris.
Vieillard-Baron, J.-L., 1989–1990. Introduction. In: Simmel, G., Philosophie de la modernité, vols. 1–2. Payot, Paris.

Modernity: Anthropological Aspects

Aihwa Ong, University of California, Berkeley, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 9944–9949, © 2001, Elsevier Ltd.

Abstract

Anthropology is a modern study of human existence in which the anthropos becomes an object of knowledge and also a technique of modern power. The field readily acknowledges that it is both a product, and an interrogator, of modernity. This relationship has given anthropology its existential doubling, as an extension, and as an undoing, of the taken for granted aspects of Western modernity. For much of the twentieth century, anthropologists have used ethnographic findings to question the assumed aspects of the historical framing of modernity and its others. Another feature of anthropology's approach to modernity is through the notion of plural modernisms, which seeks to capture the particularistic experiences of others in encounters with modern Western forms. The most promising approach has been to treat modernity as a specific ethnographic project, one that tracks the spread of political and social rationalities, and their production of new techniques, social forms, and subjects in a variety of ethnographic settings. Anthropology as a field and an ethnographer's craft is especially well suited to identify the ever shifting webs of rationalities that constitute our plural worlds, and to pose the question of modernity itself as a paradox about humanity.

Modernity as Frame and Problem in Anthropology

The idea of modernity has operated as a historical framing for anthropology, but the categories of what is modern (radically new) in the human condition have shifted throughout the history of the discipline. At a simple level, we can say that the social sciences are generally about the study of modern societies, so why would anthropologists discuss modernity as an isolated area? We can say that anthropology is a *modern* study of human existence in which the *anthropos* becomes an object of knowledge and also a technique of modern power. For a discipline that produces its knowledge through an intimate experience of the (exotic) other, the field is particularly unique in that it is both a product, and an interrogator, of modernity.

This relationship of anthropology toward modernity has given anthropology its existential doubling,as an extension, and as an undoing, of the taken for granted aspects of Western modernity. For much of the twentieth century, anthropologists worked within the historical framing of modernity and its others, and used ethnographic findings to question the assumed aspects of Euro-American modernity and progress. The second distinctive aspect of anthropology's relation to modernity is as plural modernisms, as a way to capture the particularistic experiences of nonmodern or subaltern others in encounters with modern Western forms. The most promising approach has been to treat modernity as a specific ethnographic project, one that tracks the spread of political and social rationalities, and their production of new techniques, social forms, and subjects in a variety of ethnographic settings. Anthropology as a field and an ethnographer's craft is perhaps well suited to identify the ever shifting webs of rationalities that shape our plural worlds, and to pose the question of modernity itself as a paradox about humanity.

Modernity and Its Others

From its very beginning as an academic discipline, anthropology has dealt with modernity as historical encounters between the modern and the nonmodern, and the social consequences that followed from such transformations. Initially driven by an evolutionary logic, social anthropology was dominated by an opposition between the so-called primitive societies characterized by 'premodern' customs regarding magic, sexuality, and exchange which acted as a foil for modern ideas such as rationality, the Oedipus complex, the nuclear family, and the profit motive. There was also a more radical epistemological project to discover a cultural logic embedded within the native social system, giving it a positive valence that established a rationality alternative to that of European cultures. Anthropology thus studied those aspects of other cultures that stripped away the taken for granted nature of Euro-American modernity, and introduced a reflexive element into the 'science' of human knowledges. This implicit critique of the rational and scientific assumptions of European modernity gradually included the indictment of racism, as research conducted in the shadow of European imperialism exposed its destruction of native cultures and its dependence on racism in colonial systems. Claude Lévi-Strauss ruminated on Europe's destruction of the premodern as inseparable from the loss of its own cultural ideals.

In the United States, the pioneers in anthropology participated in debates over the fate of traditional cultures under modern conditions. Nineteenth-century 'salvage anthropology' engaged in a valiant attempt to record dying cultures, sometimes rendered in utopian terms. By the early twentieth century, Franz Boas established *cultural* anthropology as a field that spoke out against the human costs of Western conquest, colonialism, and racism. Boas sought to develop a scientific language of cultural differences and universal human dignity as inescapable aspects of modern humanity. During World War II, Margaret Mead and Ruth Benedict popularized the study of national identities, often contrasting the rationality of American society with premodern others.

In the post–Second World War era, classic social theory in the writings of Marx, Durkheim, Simmel, and Weber

replaced earlier evolutionary perspectives on cultural change. Capitalism is now taken as the utterly revolutionary force that has created the modern world by radically breaking with the past. Market civilization tore asunder personal relations, destroyed communal social systems, and introduced egotistical calculations and the modern state to premodern worlds. Thus emerged a strong anthropological tradition to study the varied impact of the capitalist juggernaut on native social forms, subjectivity, and social change.

Commodity Production and Fetishism

Most scholars would agree that modernity began with the original capitalist transformations in the North Atlantic world, and their effects on the direction and limits of social change for all of humanity. For Marxist anthropologists, the most resonant story was the ways peoples of the new and ancient worlds were displaced and compelled to engage in commodity production. The works of Eric Wolf (1982) and Sidney Mintz (1985) show that exploitative market relations transformed colonized peoples into new kinds of alienated human beings, as commodity relations dissolved preexisting cultural relations among people, uprooting them from former ways of life.

A significant contribution of anthropology is to show that the market system, contrary to Marx and Karl Polanyi, did not gradually replace all other social forms, but rather interacted with other social systems in the differentiation of modern society. The 'modes of production' school demonstrated that kinship- and community-based systems in Africa, Latin America, and the Asia-Pacific became functionally articulated with the capitalist system (Godelier, 1977). Ethnographies captured in vivid and concrete detail the parasitic dependence of capitalism on the labor of slaves, sharecroppers, cash-cropping peasants, and women in the developing world. Premodern labor forms and the domestication of women in far away places have historically been part of capitalism, and therefore a part of modernity itself. Forms of servitude such as guest workers and maids, often organized along gender, ethnic, and national lines, are an integral part of advanced capitalist societies. These cultural others furnish the images of the premodern other to contemporary capitalism's own self-reflection.

The concept of commodity fetishism influenced ethnographic studies of how the market, bureaucracy, and mechanical media erode and reform indigenous notions of sociality and personhood. For over a century, anthropologists have decried the destruction or debasement of native cultures that came with capitalism, Christianity, and foreign rule. Anthropologists have documented the social despair, millennial movements, and rituals of violence inspired by missionaries, travelers, and traders. Another perspective mines the folklore and magical beliefs of colonized peoples for their hidden critique of the objectifying meanings of capitalism. The fetishization of market goods creates conditions for social competition, and gives rise to new hierarchies along lines of town and country, class and gender differences. Across the world, the mass production of culture, including soap operas and beauty contests, engender a variety of cultural configurations, displacing revolutionary sentiments with new desires and confusion.

Modernity and the Spread of Rationality

The Weberian view in anthropology dealt with the universal diffusion of the peculiar social rationalities of Western capitalism to traditional societies. Like much of the social sciences in the 1960s and 1970s, Weberian-inflected anthropology was framed by a modernization theory which ranked societies according to the degree of their adoption of modern Western values and institutions. In a series of studies on Indonesia, Clifford Geertz (1963) observed that the limited adoption of rational values and entrepreneurial forms such as capital accounting 'stranded' a country seemingly poised to take off into industrial capitalism. Geertz observed that Javanese cultural and economic spheres were firmly integrated by an overriding communal ethos that thwarted the rise of individualism and impersonal relations required by market organization. The Weberian assumption that economic rationalization would be gradually split off from other spheres was itself challenged by studies of advanced capitalist societies. Marshall Sahlins (1976) argued that in American society, practical reason can be shown to be informed by religious and aesthetic schemes of meaning. But clearly, the project to decenter certain dominant Western conceptions of humanity must be embedded in an analysis of how these schemes of meaning are transformed in relation to the power dynamics of class, gender, and race.

The New Modernisms

By the early 1980s, the end of the Cold War spelled the decline of universalizing theories of modernization, and new generations of anthropologists rejected the totalizing, evolutionary assumptions about social change. The writings of Michael Foucault, and the rediscovery of Nietzsche, popularized understanding about the intimate relations between knowledge and power and focused intellectual projects not on things but on the conditions of emergence of new forms that have been built into the conduct of life. Foucault's *The Order of Things* (1970) first sets out the ways knowledge/power schemes shape and normalize objects of knowledge, and his later works have had a powerful effect on anthropology's analytical techniques and self-perception.

But the so-called postmodern challenge, and an influential text, *Anthropology as Cultural Critique* (Marcus and Fischer, 1986), became symptomatic of the misrepresentation of the crisis of knowledge as simply a 'crisis of representation.' The self-reflexive turn caused many to scale down their anthropological project to an experiment in writing ethnography. First, a 'politics and poetics' of representation approach seeks to redress the temporal and political inequalities between ethnographic informants and the anthropologist (Clifford and Marcus, 1986). Second, feminists – concerned about partial truths and informant empowerment – popularized multivocality to undercut the authority of the author. Third, a radical deconstructionist approach privileges language and writing over the theorizing of social forms. Such ethnographies seek to capture the jarring experiences of colonial and proletarian subjects, both to

uncover the hidden 'truths' of subaltern suffering and resistance, and to subvert objectifying categories of description. Fourth, the so-called 'culture/power/history' orientation takes modernity as a historical cultural formation, embedding the study of discourses and cultural politics within the specific political economic contexts. These varied approaches seek to capture the particular conjunctures and disjunctures of discourse, culture, and power in encounters with Western power and knowledge. But unlike experimental ethnographies of earlier generations, the current postmodernist assault on the ideological constitution of Western social science does not offer an alternate method for studying schemes of practice, their logics and limits, and possibilities for transformation. Thus, postmodern approaches represent a detour in the anthropological theory of modernity. They do not develop a positive concept of modernity or analytical categories that can be applied to the systematic study of social forms that will allow for theoretical generalization.

Modernity as Political and Social Rationalities

The anthropological focus on modernity as an ethnographic problem in its own right was stimulated by recent European social thought. In American anthropology, with the help of his American interlocutors Herbert L. Dreyfus and Paul Rabinow, Michel Foucault's *oeuvre* came to have a wide-ranging influence on different approaches to questions of modernity. The fundamental question was not to treat modernity as an analytical opposition to given ethnographic realities, but rather to take modernity itself as an ethnographic object, as specific social and political rationalities configuring particular worlds.

Foucault's *Discipline and Punish* (1977) has been especially influential in the anthropological analysis of modern power. We can identify three conceptual clusters: (1) a positive (not repressive) 'capillary' notion of modern power as circulating practices arising from multiple domains; (2) the role of knowledge/power schemes in proliferating social norms and forms that produce subjects; and (3) the specific technologies of governmentality – grids of knowledge/power, mechanisms of surveillance – that govern our everyday behavior and make us particular kinds of modern human beings. Such a reconceptualization of modernity as projects and schemes of power is particularly suited to the ethnographic method of tracking the logics of everyday attitudes, habits, and practices (Bourdieu, 1977). It becomes apparent that the goal is not to arrive at some grand theory about modernity, but rather to scrutinize the concrete manifestations of emerging social practices, norms, and cultural politics in relation to the market, the nation state, and to globalizing forces.

Alternative Modernities

The Foucault effect on thinking about colonialism and postcolonialism shifts attention from reactionary cultural formations to strategic modern projects. More attention is now given to social rationalities – law, hygiene, sex, and labor disciplines – involved in civilizing colonized societies and subjects. The study of postcolonial development itself is reconceptualized as regimes of discursive power that have material effects on shaping the politics for remaking the Third World (Escobar, 1995). The anthropology of imaginary modernities – narratives that depend on the creation of the otherness of their own pasts as well as of the West – show that in the new China, different generations have been shaped by visions of a movement from a dark past into a bright socialist future. Indeed, for other countries, folk theories of modernity may even incorporate backwardness, thus subverting metropolitan development programs.

The concept of 'alternative modernities' risks reifying native views of modernity, and it will have more purchase if ethnographic attention is also paid to the specific social practices and forms that are distinctively modern. For instance, anthropologists note that the obsessive promotion of marginal cultural practices may engender a 'spectacular modernity' in societies greatly transformed by industrial capitalism and deeply anxious about the loss of national identity.

More grounded approaches have explored the disciplining effects of agricultural programs and bureaucracy in shaping developing societies (Ferguson, 1990). Others consider how modernity becomes a state project in developing and postsocialist countries, focusing on the interactions between the market and other social systems that discipline civil society. Such ethnographies demonstrate how the novel forms of governmentality and governance, evolving in relation to globalizing forces, are dialectically linked to selfdescriptions of alternative configurations of modernity.

Modern Schemes and Subjectivity

In *French Modern* (1989) Paul Rabinow opens up the space between the high cultures of modernity and the experiences of ordinary life as a domain of 'middling modernity' where social technologies encounter and regulate everyday sensibility. By seeking the coming into being of modernity in the mundane details of daily life, ethnographers present a complex picture of how the nondramatic norms and forms are constitutive of what it means to be human in particular societies. Perhaps anthropologists have gone further than other social scientists not only in grounding modernity as a particular set of rationalizing norms and practices, but also to specify the locations of different kinds of rationality in modern, universalizing institutions. Ethnographic scrutiny delineates how different regimes of power/knowledge have identified, processed, and constituted particular kinds of modern subjects in runaway factories, hospitals, and scientific laboratories and in utopian cities. Thus, modernity as an object of anthropological inquiry itself is a shifting target, as anthropologists track specific technologies in a range of ethnographic situations, and allow for comparison beyond cultural particularities.

Cities and the Reconfiguring of Citizenship

The metropolis, as Benjamin and Simmel have noted, is the *locus classicus* of modern sociality, the setting for the most extreme expressions of individual freedom and impersonal objectification. Earlier anthropological studies of urban life focused on the retention of village bonds among migrants or the maintenance of ego-centered networks among working

class communities, as a defense against the alienating effects of modernity's flux. Indeed, more and more, cities rather than nations have become the salient sites for examining the gap between political citizenship and the experiences of different categories of persons who have been excluded or subordinated within the ideal national community. A central problem in conceptualizing modernity then is an ethnographic investigation of the mechanisms and routines of governance and the practices of subjects being thus constituted as administrative subjects of particular cities, zones, and nation-states.

There has been a spate of ethnographies on immigrants, drug-addicts, prisoners, street-persons, youths, the working poor, and racial minorities protesting their illusory juridical status and exclusions from the 'liberal community.' Generations of anthropologists have studied black ghetto culture and poor urban neighborhoods, stressing the differential forms of integration into American society. More recent works explore how marginalized urban groups formulate their demands for substantive rights – to schooling, housing, clean water, cultural difference, and respect – from the state and fellow citizens in Western democracies (Holston, 1998). A concept of citizenship, not simply as imposed rights for abstract individuals, but as a set of contingent, negotiative relations between local authorities and self-making subjects, suggests a distinctly modern form of state-subject relations and subject-making.

Transnationalism, Differentiation, and Subject Formation

Of course, the modern condition now includes the proliferation of transnational networks that are transforming earlier understandings about community, mobility, and belonging. In our globalized world, modernity is increasingly understood and experienced in terms of transnational processes and webs of relationships shaped by the intensified flows of people, images, knowledge, and practices in everyday life. Such reconfigurations of social life and connections unsettle earlier anthropological notions of culture, community, and identity, suggesting that it may be possible to talk about a global ecumene spawned by intensified cross-border flows and exchanges (Hannerz, 1992). An approach to the cultural dimensions of globalization (Appadurai, 1996) suggests that the spatial and imagined coordinates of culture and community are now shaped by diasporic communities that are in the process of becoming 'postnational.' The category of the transnational imaginary, however, is an indeterminable unit of analysis, and the visions of global companies, multilateral agencies, and diasporic networks are all rather different. The voluntary flows of tourists, intellectuals, expatriates, and businessmen should be theoretically distinguished from the forced homelessness of migrant workers and refugees.

How are such imagined communities constituted in relation to the variety of social systems that have been produced and proliferated by globalization? Accelerated border flows have multiplied specific configurations of subsystems that go to make up particular worlds. While sociologists have focused on the materiality of the new technologies – time-space compression and time-space distanciation – anthropologists seek to sharpen questions about what new modern cultural forms and subjects are emerging in different places, and what alternative cultural logics may be at work (Ong, 2000). Transnational institutions and formations privilege new kinds of corporate subjects who internalize the norms of mobility, flexibility, and competitiveness (Martin, 1994). A further direction investigates the rationalities that go into the invention of new techniques, practices, objects, places, and subjects, configurations in which the meaning of 'the human' is problematized (Rabinow, 1998). Such projects will require us to go beneath the rhetoric of globalization and engage in multisited ethnography that can track the performative and migratory aspects of rationalities – in science, medicine, business, bureaucracy – that are shaping modern subjectivity in a variety of domains. Attention to new forms of governmentality enables us to ask what kinds of rules we have set up for ourselves to be considered modern human beings.

Living with Uncertainty

But the operation of rationalities is incomplete and contingent, and globalization has proliferated risks – volatile cash flows, unsettling information, diseases, toxic wastes, light arms, and bioscientific inno- vations – that expose everyone to the irrationalities of modernity. Capitalism has intensified the dominance of space and proliferated the social forms of modernity. Thus, in a world where the globalization of contingency is 'modern society's main attribute' (Luhmann, 1998), our thinking needs to go beyond simply the perception and containment of risk. What remains to be investigated is how risks enter into the calculations of everyday life not as an independent but as a dependent variable.

There are signs that the future directions of an anthropology of modernity will concern itself with how societies and individuals live with uncertainties and absorb risks in their everyday routines. What do modern people problematize in their everyday life, and how do they solve problems? In the intersecting realms of health, science, morality, and capitalism, for instance, anthropologists have explored the ways people cope with the AIDS pandemic, the threats of nuclear plants, and the politics and moral perils of human genomic research. The study of divergent human fate as well will attend to how heterogeneous social spaces – global cities, production zones, blighted borderlands, postsocialist formations, refugee camps, and cross-border networks – are experimenting with different ways for ordering human sociality.

While the anthropology of modernity has a shifting target – the particularistic configurations of contemporary social life – there is a fundamental unity in the ethnographic diagnosis of the particular techniques, norms, and practices that go into the constitution of modern human subjects in specific fields of power. The ethnographic method is ideal for identifying new devices, arrangements, and techniques, and for attending to the performative aspects of emerging practices within an actor-network formulation (Latour and Woolgar, 1986[1979]). Our methodology as well must absorb the uncertainties of our human science, and theory must be based in an ever-widening spiral of observations beyond the immediate field situation. The challenge is for anthropologists to wed an ethnographic analysis of the practicable, strategic

features we take to be characteristic of modern society with their interpretive, moral aspects.

For all the self-reflection and criticism, the project of anthropology remains one of a rigorous and systematic exposition of emerging worlds. Such a methodology – whether to study the interactions of functions systems, modes of ruling, modalities of subject-formation, biotechnological development, or the management of risks they all entail – allows for linking the particular with the general, the local with the global, the structural with the interpretive dimensions of modernity. By identifying the social and political rationalities constituting our plural worlds, we may ask whether we want to remain the kinds of human beings we have become. Modernity is both the condition and the decision about the future.

Bibliography

Appadurai, A., 1996. Modernity at Large. University of Minnesota Press, Minneapolis, MN.

Bourdieu, P., 1977. Outline of a Theory of Practice (R. Nice, Trans.). Cambridge University Press, Cambridge, UK.

Clifford, J., Marcus, G. (Eds.), 1986. Writing Culture. University of California Press, Berkeley, CA.

Escobar, A., 1995. Encountering Development. Princeton University Press, Princeton, NJ.

Ferguson, J., 1990. The Anti-Politics Machine. Cambridge University Press, Cambridge, UK.

Foucault, M., 1970. The Order of Things: An Archaeology of the Human Sciences (A. Sheridan, Trans.). Tavistock, London.

Foucault, M., 1977[1975]. Discipline and Punish: The Birth of the Prison (A. Sheridan, Trans.). Pantheon Books, New York.

Geertz, C., 1963. Agricultural Involution. University of California Press, Berkeley, CA.

Godelier, M., 1977. Perspectives in Marxist Anthropology. Cambridge University Press, Cambridge, UK.

Hannerz, U., 1992. The global ecumene as a network of networks. In: Kuper, A. (Ed.), Conceptualizing Society. Routledge, London.

Holston, J. (Ed.), 1998. Cities and Citizenship. Duke University Press, Durham, NC.

Latour, B., Woolgar, S., 1986[1979]. Laboratory Life. Princeton University Press, Princeton, NJ.

Luhmann, N., 1998. Observations on Modernity (W. Whobrey, Trans.). Stanford University Press, Stanford, CA.

Marcus, G., Fischer, M., 1986. Anthropology as Cultural Critique. University of Chicago Press, Chicago.

Martin, E., 1994. Flexible Bodies. Beacon Press, Boston, MA.

Mintz, S., 1985. Sweetness and Power. Viking, New York.

Ong, A., 2000. Flexible Citizenship. Duke University Press, Durham, NC.

Rabinow, P., 1989. French Modern. MIT Press, Cambridge, MA.

Rabinow, P., 1998. The Anthropology of Reason. University of Chicago Press, Chicago.

Sahlins, M., 1976. Culture and Practical Reason. University of Chicago Press, Chicago.

Wolf, E., 1982. Europe and the People Without History. University of California Press, Berkeley, CA.

Modernity: History of the Concept

Gurminder K Bhambra, University of Warwick, Warwick, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Wagner, volume 15, pp. 9949–9954, © 2001, Elsevier Ltd.

Abstract

Modernity is one of the central concepts of sociology. Typically, its emergence is located in the processes of economic and political revolution in Europe in the eighteenth century. Such an understanding conflates Europe with modernity and renders the process of becoming modern, at least in the first instance, one of endogenous European development. Coterminous with this argument is the idea that the rest of the world was external to these world-historical processes and that colonial connections and processes were insignificant to their development. This is challenged by more recent critical understandings of modernity associated with postcolonial and decolonial thought.

Introduction

'Modernity' has long been the dominant conceptual frame for social and political thought, sociology, and the social sciences more generally. The repercussions of the French Revolution and the processes of industrialization stimulated debates about the emergence of a *modern* world (modernity) and this world was held to require a distinctively *modern* form of explanation (sociology). The sociological understanding of modernity typically rests on ideas of the modern world associated with the processes of economic and political revolution located in Europe and underpinned by the earlier cultural changes brought about by the Renaissance, Reformation, and Scientific Revolution. Such understandings conflate Europe with modernity and render the process of becoming modern, at least in the first instance, one of endogenous European development. At the same time, the rest of the world is regarded as external to these world-historical processes and colonial connections are seemed to be insignificant to their development.

While accounts of the industrial and French revolutions – and, by implication, of modernity itself – have not remained unchanged over time, what has remained constant has been the historiographical frame – of autonomous endogenous origins and subsequent global diffusion – within which these events are located. As such, not only have 'others' not been recognized as constitutive of the canonical 'twin revolutions,' but also the potential contribution of other events (and the experiences of non-Western 'others') to the sociological paradigm of modernity has rarely been considered. This article will first present the standard, dominant genealogy of the conceptual history of modernity before addressing the alternate genealogy as coloniality–modernity. It will close with a discussion of the necessity of developing our concepts through an acknowledgment of broader, connected histories and sociologies.

Modernity and the Social Sciences

The dominant accounts of modernity within the social sciences rest on two fundamental assumptions: *rupture* and *difference* – a temporal rupture that distinguishes a traditional, agrarian past from the modern, industrial present; and a fundamental difference that distinguishes Europe from the rest of the world. These paradigmatic assumptions, Gurminder K. Bhambra (2007) argues, have framed both the standard methodological problems posed by social inquiry and the explanations posited in resolving them. They have also demarcated the disciplines themselves with sociology's remit being 'the modern,' while 'the traditional' has been that of anthropology, with 'sociological laws' common to both.

Whatever their other differences, the early sociologists believed themselves to be living through a (or perhaps *the*) great transformation in history and were concerned with understanding how it had begun and influencing how it would be brought to the completion they believed to be inherent within it. This was done through an engagement with ideas of historical stages. Once the general pattern of historical development was known, it was believed that the trajectory of development for other societies could be ascertained without direct research as all human societies were deemed to follow the same pattern; one which had been established by Europe. This built on the earlier tradition of the Scottish Enlightenment, which developed understandings of progress within human societies in relation to ideas of successive stages in historical development.

This was taken up and developed in the twentieth century by Talcott Parsons (1971), who argued that a common origin and shared heritage allowed for the evolution of a 'European system' of modern societies that could be understood in its entirety as differentiated from both earlier societies and from the rest of the world. Jürgen Habermas (1996), in turn, argued for modernity to be understood as an 'unfinished' project where the future was not about the reproduction of the present, but was considered to be a space for the further development of projects and trends. The 'unfinished project' is the bringing to fruition of what is already predicated in the Western experience.

Modernization Theory

Following in the classical tradition of sociology, modernization theory took as its idea of change the standard notion of a linear movement from a traditional past to a modernized future.

Explanations of the processes of modernization were primarily located in the context of a historical understanding of societies, where each form was deemed to be superseded by a progressively higher one. Traditional, or premodern, societies were put forward as objects of comparison with societies already deemed to be modern and the problem was set up in terms of accounting for the historical transition from one to the other. The debate on convergence around modern institutions and the economy, however, also occurred in the context of counterclaims arguing for *divergence* around issues of culture and political organization.

Parsons (1971) was particularly interested in understanding the implications of the transition to modernity and looking at why the breakthrough to modernization had not occurred elsewhere than in Europe. In common with most other theorists at the time, he believed that modern society had emerged in the West, and that this provided the base from which the *system* of modern societies then developed. Modernization scholars, such as S.N. Eisenstadt (1974), also believed that Western modernization should be used as a model of global applicability and other societies classified in terms of their relative modernization in comparison with this model; that is, other societies were to be studied in terms of the extent to which they approximated the characteristics of Western industrial societies. While becoming modern in the first instance might derive from peculiar circumstances which were historically contingent and even perhaps unlikely (for example, as set out by Weber in his study of the Protestant Ethic), once Europe had become modern, it was deemed to be able to show the way to the rest of the world as a model to be imitated. The birth of modernity, it was believed, could be induced.

Reinhard Bendix (1967) usefully summarizes modernization theory as resting on three related assumptions: first, an understanding of 'tradition' and 'modernity' as mutually exclusive; second, social change occurring as a consequence of phenomena internal to the society changing; and third, a belief that modernity would eventually replace tradition and, in doing so, would have the same effects across the globe. In this sense, modernization theory rests on a notion of *convergence* whereby the difference of other societies – as constituted through their traditions – would be erased through the process of the global diffusion of Western institutions. Once the structures of modernization extended to other areas, it was believed that the previous indigenous patterns, or structures, would change and that change would be in the direction of the relatively modernized societies.

While this position was largely dominant within the debates on modernization, it was not uncontested at the time, for example, by Dean Tipps (1973). Further, critical scholars, such as Henry Bernstein (1971), believed that using notions of a stagnant past, where the differences among societies were not regarded as relevant to the issue of modernization, and a dynamic heterogeneous present, both distorted the character of traditional societies and obscured understandings of the different forms of relationships between traditional societies and, what were regarded as, new institutions. This body of work came to be known as underdevelopment or dependency theory.

The demise of modernization theory, and related understandings of modernity, was associated with the explicit move away from structural–functional explanations in sociology. The latter was associated with the rise of more radical approaches, particularly those influenced by Marxism, and the related decline of Parsons. The fall of communism in the 1990s, however, reversed these sensibilities. For a number of commentators, the convergence thesis which had been discarded had turned out to be confirmed with Francis Fukuyama most famously proclaiming a new 'end of history.' This is the context in which some writers, most notably Eisenstadt (2000), returned to modernization theory seeking to challenge this triumphant liberalism, while also acknowledging the force of events. At the same time, the processes of decolonization, which had been the initial context for modernization theory, had themselves given rise to postcolonial critiques of the Eurocentrism of dominant understandings of modernity. The new paradigm of multiple modernities, then, was articulated in relation to these varying concerns.

Multiple Modernities

Decolonization in the 1960s was followed by the fall of communism in Europe in the late 1980s and 1990s. The perceived seismic shift in the global order – in particular, globalization being seen as the creation of a world market after the breakup of the Soviet-dominated economic bloc – renewed sociological debates about the nature of the modern world leading to the development of a new paradigm, that of multiple modernities. While for some, such as Fukuyama, these events confirmed the convergence claims of modernization theory and the role of the United States as 'lead society,' for theorists of multiple modernities, the removal of 'cold war' constraints instead allowed for greater divergence. The insistent claims of scholars from the Third World – be they theorists of underdevelopment or then postcolonial theorists – also required engagement and address.

Theorists of multiple modernities, such as Eistenstadt and others writing in Daedalus (2000), believe that, in outlining their new approach, two fallacies are to be avoided. The first, associated with earlier modernization theories, is that there is only one modernity. The second is that of Eurocentrism or, in their words, Orientalism. The argument is that while the idea of one modernity, especially one that has already been achieved in Europe, would be Eurocentric, theories of multiple modernities must, nonetheless, take Europe as the reference point in their examination of alternative modernities. Thus, while theorists of multiple modernities point to the problem of Eurocentrism, they do so at the same time as asserting the necessary priority to be given to the West in the construction of a comparative sociology of multiple modernities. As such, multiple modernities continue to be understood as derived from the creative appropriation, by those that followed, of the institutional frameworks of modernity that are seen to originate in Europe, and so the problem of Eurocentrism remains integral to the new paradigm.

Within the approach of multiple modernities, modernity is understood simultaneously in terms of its *institutional constellations* as well as a *cultural program*. Understanding modernity in this way allows scholars to situate 'European modernity' – seen in terms of the conflation of the institutional and the cultural

forms – as the originary modernity and, at the same time, allows for different cultural encodings that result in *multiple* modernities. For example, Nilufer Gole (2000) reflects on the idea of modernity from what she terms its Islamic edge, that is, Turkey, in order to discuss non-Western patterns of modernization. The idea of multiple modernities, then, is consistent with the idea of a common framework of modern institutions – for example, the market economy, the modern nation-state, and bureaucratic rationality – which originated in Europe and was subsequently exported to the rest of the world. This explains the apparent paradox that theorists of multiple modernities can dissociate themselves from Eurocentrism at the same time as apparently embracing its core assumptions. Further, in charting the development of specifically *multiple* modernities, scholars suggest that the first radical transformation occurred in the Americas, followed by the communist and national-socialist types in Europe. Not only modernity, then, but also multiple modernities are seen to have their origin in Europe or the Western civilizational framework at large.

Similar problems can be seen in the proliferation of theories of alternative modernities, global modernities, hybrid, and entangled modernities despite it seeming that the debates around these issues have moved beyond their initial, limited Eurocentric understandings of the world. In a collection of articles on alternative modernities, Dilip Gaonkar (2001) suggests that the emergence of the debate on alternative modernities reinforces the fact that something specifically modern is held to be inescapable. The West is understood as the major clearing house of modernity and in its globalization has traveled from the West to the rest of the world meaning that non-Western peoples must now begin to engage their traditions with modernity in different forms of hybrid 'modernities.' These assertions rest on a number of assumptions not least that there is an original modernity which was born in and of the West, and that the West is significantly different from the rest of the world such that while it can enjoy an original modernity, everybody else has to make do with a hybrid version, in which different modernities are no more than the combination of the adopted 'originary' form with 'nonmodern' elements. Gaonkar (2001) argues that a minimal requirement for thinking in terms of alternative modernities is to opt for a 'cultural,' as opposed to an 'acultural' theory of modernity which enables us to understand how non-Western cultures have tropicalized, or domesticated, key features of modernity. Even though the idea of modernity is seen less as the terminus toward which non-Western peoples constantly tend and more in terms of being constitutive of plural and diverse global systems, the West remains the pivot around which these ideas circulate.

The challenge posed to modernization theory by the approach of multiple and alternative modernities has some significance in its own terms, but it is much less fundamental than its advocates suppose. The basic premise of theorists of multiple and alternative modernities, of questioning the dominant assumption of convergence and its corollary idea of one trajectory to modernity, is an important qualification to modernization theory. What is significant in its omission, however, is the failure to address adequately the way in which the West remains the point of reference. Further, in maintaining its focus on the internal dynamics of *separate* civilizations and the inability to take a point of view *other* than from the West, the comparative approach exacerbates the problem of Eurocentrism by ignoring and even actively excluding (through its use of ideal types) the connected and entangled histories that constitute the basis of an adequate understanding of the global context of sociohistoric processes. As Arif Dirlik (2003) argues, by identifying 'multiplicity' with culture, the idea of multiple modernities seeks to contain challenges to modernity by conceding the possibility of different ways of being modern, but doing nothing to address the problems with the conceptualization of modernity itself. Delinking our understandings of sociohistoric processes from a European trajectory and focusing not only on the different sources and roots but also on the ways these interacted and intersected over time would provide us with a richer understanding of the complexities of the world in which we live and the historical processes that constitute it. There is a need, as Sanjay Subrahmanyam has argued, to address these *interconnections* as opposed to reifying the entities that are supposed to be connected.

Postcolonial Theory and the Coloniality/Modernity Paradigm

Two traditions that have been most significant in reconceptualizing understandings of modernity have been those associated with postcolonial theory and the coloniality/modernity school of thought. Since the 1960s and 1970s, knowledge claims in the social sciences and humanities have been under pressure from the rise of subaltern positions and an explicitly recognized politics of knowledge production. If we now understand dominant approaches as Eurocentric, for example, it is because of new voices emerging in wider political arenas and in the academy itself. The end of colonialism as an explicit political formation has given rise to understandings of postcoloniality and, perhaps ironically, an increased recognition of the role that colonialism played in the formation of modernity. Postcolonial and decolonial scholarship has been integral to the opening out and questioning of the assumptions of the dominant discourses and, following Boaventura de Sousa Santos (2007), arguing for the possibility of other knowledges. This is also developed by Sujata Patel (2014) who argues for the contestation of colonial modernity within and by the global south in order to establish other understandings of modernity.

Such arguments have been most successful in their challenge to the insularity of historical narratives and historiographical traditions emanating from Europe or what are seen as northern epistemologies (see Spivak, 1988). This has been particularly so in the context of demonstrating the parochial character of arguments about the endogenous European origins of modernity in favor of arguments that suggest the necessity of considering the emergence of the modern world in the broader histories of colonialism, empire, dispossession, and enslavement.

Postcolonial theory, according to Homi Bhabha (1994), attempts to disrupt standard discourses of modernity by shifting the frame through which we view the events of modernity and questioning what we consider those events to be. This is not a version of multiple modernities, in that it does not offer

an alternative modernity to be placed alongside those already seen to be existing. Rather, it calls on us to interrogate the conceptual paradigm of modernity, as it has commonly been understood, from the perspectives of those Others usually relegated to the margins, if included at all. These Others, usually constituted as being outside modernity, exist in both the global south and the global north. Their perspectives are now taken to be central to understanding modernity conceptually.

The task, as Bhabha (1994) puts it, is to take responsibility for the unspoken, unrepresented pasts within our present and, through an understanding of them, reconfigure what it means to be modern. This reconsideration does not relegate difference simply to cultural expression, but calls upon that difference to make a difference to the narratives of modernity already in place. The issue, according to Bhabha (1994), is more about reinscribing 'other' cultural traditions into narratives of modernity and thus transforming those narratives – in both historical terms and theoretical ones – rather than simply renaming or reevaluating the content of these other 'inheritances.'

The meaning of modernity, according to Bhabha (1994), does not derive from a foundational event in the past, but from its continual contestation in the present. This negotiation calls into question both the conditions with which modernity is typically associated and those who lay claim to it. In naming ourselves as modern, as Bhabha (1994) suggests, one moves from the periphery to the center and, in the process, transforms the understanding of 'modernity' from which and about which one speaks. This resonates with the work of the coloniality/modernity school of thought, which is linked to world-systems theory as well as scholarly work in development and underdevelopment studies.

The theoretical distinction – coloniality/modernity – was first articulated by Anibal Quijano (2007). He argues that with the sixteenth-century conquest of the lands that we now call Latin America emerged a new world order which ended up encapsulating the whole planet. This coloniality of power, expressed through political and economic spheres, he continues, was strongly associated with the coloniality of knowledge (or of imagination), articulated as modernity/rationality. The modernity that Europe takes as the context for its own being is, he suggests, so deeply imbricated in the structures of European colonial domination over the rest of the world that it is impossible to separate the two: hence, modernity/coloniality. This understanding enables critique of the European paradigm of modernity/coloniality to be more than a simple refutation of modernity. Rather, it presents a more complicated and more complex definition of it.

The elaboration of the mutual coconstitution of modernity/coloniality and the extension of the timeframe of modernity back to the fifteenth century are two of the key contributions made by Quijano (2007) to its reconceptualization. Both aspects radically challenge standard notions of modernity. The temporal dimension is stretched back from the eighteenth century to the fifteenth century and the spatial dimension is expanded from northern Europe, specifically France and Britain, also to include southern Europe, specifically Spain and Portugal, and Latin America. Acknowledging these shifts establishes colonialism as the precondition for the development of (ideas of) modernity and all subsequent understandings of modernity have to take into account the conditions of its emergence.

Walter Mignolo (2007) builds on Quijano's earlier theoretical work and, in particular, further elaborates his conception of modernity/coloniality in the context of the work of epistemic decolonization necessary to undo the damage wrought by both modernity and by understanding modernity/coloniality only as modernity. Mignolo (2007) argues for a process of 'delinking' that is both an epistemic delinking from the rhetoric of modernity and a practical delinking – that is, a struggle to break free – from the logic of coloniality. His project of 'delinking' points to the need to change the terms (concepts) as well as the content (histories) of the conversations on modernity/coloniality that enables the histories and thought of other places to be understood as prior to European incursions and to be used as the basis of developing connected histories and sociologies of encounters through those incursions.

Mignolo (2007) argues for the importance of recovering earlier histories, that is, histories prior to colonization, from which to articulate both alternative possibilities of living and modes of resistance to the logics of modernity/coloniality. Shifting the historical frame of significant events from Europe to other parts of the world, and prior to European contact, necessitates a historiography of global history cognizant of colonialism and coloniality. This rewriting does not involve the inscription of a new form of universal modernity, but rather, for Mignolo (2007), a differently understood *pluri*versal form, where local histories and their narratives of colonization and decolonization can connect through shared experiences and create new ways of being modern and understanding modernity.

Conclusion

In standard histories of sociology, modernity, or the modern world, is seen to come into being as a consequence of the 'twin' revolutions – the French and the Industrial. In more critical work, it has also long been associated with the processes of dispossession, enslavement, colonialism, and imperialism. This understanding of the coconstitution of coloniality–modernity, and the general silence around this, is central to any proper appreciation of the history of the concept.

See also: Geist (Spirit): History of the Concept; Progress: History of the Concept; Sociology, History of.

Bibliography

Bendix, Reinhard, 1967. Tradition and modernity reconsidered. Comparative Studies in Society and History: An International Quarterly IX, 292–346.

Bernstein, Henry, 1971. Modernization theory and the sociological study of development. Journal of Development Studies 7 (2), 141–160.

Bhabha, Homi K., 1994. The Location of Culture. Routledge, London.

Bhambra, Gurminder K., 2007. Rethinking Modernity: Postcolonialism and the Sociological Imagination. Palgrave, Basingstoke.

Daedalus, 2000. Multiple modernities. Daedalus 129 (1).
Dirlik, Arif, 2003. Global modernity? Modernity in an age of global capitalism. European Journal of Social Theory 6 (3), 275–292.
Eisenstadt, Shmuel N., 1974. Studies of modernization and sociological theory. History and Theory 13 (3), 225–252.
Eisenstadt, Shmuel N., 2000. Multiple modernities. Daedalus 129 (1), 1–29.
Gaonkar, Dilip P. (Ed.), 2001. Alternative Modernities. Duke University Press, Durham.
Gole, Nilufer, 2000. Snapshots of Islamic modernities. Daedalus Winter 129 (1), 91–117.
Habermas, Jurgen, 1996. Modernity: an unfinished project. In: d'Entreves, M.P., Benhabib, Seyla (Eds.), Habermas and the Unfinished Project of Modernity: Critical Essays on the Philosophical Discourse of Modernity. Polity Press, Cambridge, pp. 38–58.
Mignolo, Walter D., 2007. Delinking: the rhetoric of modernity, the logic of coloniality and the grammar of de-coloniality. Cultural Studies 21 (2), 449–514.
Parsons, Talcott, 1971. The System of Modern Societies. Prentice-Hall Inc., NJ.
Patel, Sujata, 2014. Gazing backwards and looking forwards: colonial modernity and the making of a sociology of modern India. In: Arjomand, Said (Ed.), Social Theory and Regional Studies. SUNY Press, New York.
Quijano, Aníbal, 2007. Coloniality and modernity/rationality. Cultural Studies 21 (2), 168–178.
Santos, Boaventura de Sousa (Ed.), 2007. Another Knowledge Is Possible: Beyond Northern Epistemologies. Verso, London.
Spivak, Gayatri Chakravorty, 1988. Subaltern studies: deconstructing historiography. In: Guha, Ramachandra, Spivak, Gayatri Chakravarty (Eds.), Selected Subaltern Studies. Oxford University Press, New York.
Subrahmanyam, Sanjay, 1997. Connected histories: notes towards a reconfiguration of early modern Eurasia. Modern Asian Studies 31 (3), 735–762.
Tipps, Dean C., 1973. Modernization theory and the comparative study of societies: a critical perspective. Comparative Studies in Society and History 15 (2), 199–226.

Modernization

Manfred G Schmidt, Ruprecht Karls Universität Heidelberg, Heidelberg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The most widespread use of political modernization is that of a deep-seated transformation of a nonmodern political system in a posttraditional modern political order with key characteristics such as rationalization of authority, differentiation of new political functions, mass participation, capacity building, establishing equality as a major guideline, and a sequence, in which economic development is conducive to cultural changes that make individual autonomy and democracy increasingly likely. There are many routes that lead toward modernity, and modernization has been associated with outcomes of different quality. The diversity of political modernization gave rise to major revisions of classical modernization theories, such as the substitution of 'complex' theories for 'simple' modernization patterns, and the amendment of classical modernization theory by the theory of modern 'risk society.'

Political modernization has three different, yet related, meanings. In political discourses, it refers to successful and enduring adaptation of the political structures of a country in a way which secures system maintenance and political innovation. Political modernization may also be used for the transformation of a previously nonmodern or semimodern society to a higher level of socioeconomic and political development. In the vocabulary of the social sciences, the most widespread use of political modernization is that of a transformation of a nonmodern, mostly traditional political system in a posttraditional political order.

The defining characteristics of political modernization in the latter sense embrace above all (1) "rationalization of authority" (Huntington, 1966: p. 378), i.e., the replacement of traditional political authority by single, secular, national political authority; (2) differentiation, i.e., a "process of progressive separation and specialization of roles, institutional spheres, and associations in the development of political systems" (Coleman, 1968: p. 397); (3) the expansion of popular involvement and participation in the polity; (4) the establishment of equality as a key norm, comprising above all adult citizenship, universalistic legal norms, and the prevalence of achievement criteria in recruitment to political and administrative roles, and, in the case of democratic modernization; and (5) a "human development sequence" (Inglehart and Welzel, 2005), in which economic development tends to substitute self-expression values for survival values and makes individual autonomy and democracy more likely.

Routes to Political Modernization

Many routes lead to political modernization. The most ancient route has been that of Western Europe. That trajectory was mainly derived from, or overlayered with, conflicts of class. A second route was opened through creation of overseas migration to the New Worlds in North America and Australasia. Political development in the 'colonial zone' (Therborn, 1995: p. 6) marked a third path to political modernization. A fourth route has resided in those cases, in which a policy of defensive modernization, initially adopted mainly for the purpose of strengthening the 'power state,' was ultimately conducive to democratizing the polity of the country, such as in Japan and South Korea. A fifth trajectory to political modernization has taken place in Central and Eastern European postcommunist countries. That path involves two modernization processes: the replacement of a planned economy by a mixed economy or a market economy, and the substitution of a democracy for an autocratic regime. Whether political change in countries such as Tunisia, Lybia, and Egypt since 2010 ('Arab Spring') represents a sixth route to modernity, representative of an 'Islamic type of modernity,' remains questionable because nonsecular forces and survival values have remained powerful and massive restrictions of political rights and civil liberties have continued to prevail in these countries.

Simple and Complex Patterns of Political Modernization

In a stylized fashion, one can distinguish between theories of 'simple' and 'complex' patterns of political modernization. Theories of simple political modernization patterns regard modernization as a relatively uniform, homogeneous, stable, linear, and irreversible process. This family of theories also expects more benefits than costs from political modernization. In contrast to this, theories of complex modernization patterns conceptualize political modernization as a heterogeneous, pluralistic, precarious, crisis-ridden, nonlinear, and reversible process. Moreover, these theories also focus attention on side effects, unintended consequences, and massive social costs associated with political modernization. According to this view, political modernization does not necessarily generate a fairer and a more efficient political system. Moreover, theories of complex modernization also emphasize path-dependent patterns of social and cultural change. According to this view, the cultural heritage of a society – for example, the heritage of catholic or protestant religion, Islam, or communism – will have a lasting impact on political culture and will be conducive to the emergence of 'families of nations' (Castles, 1993), such as the Scandinavian countries, the English-speaking nations, Post-Communist Europe, and the Islamic world.

The shift in focus from simple to complex patterns of modernization mirrors that political modernization may be more loosely coupled with successful socioeconomic development and high levels of political performance than predicted in simple modernization theories. For example, economic and social development may be associated with political modernity or a polity with strong premodern characteristics, good governance or low levels of governance, and democratization or persistence of semidemocratic or nondemocratic regimes. The persistence of nondemocratic regimes is exemplified by economically advanced oil-exporting states in the Middle East. Furthermore, democratic modernization can be achieved not only at a relatively advanced stage of socioeconomic development, but also on a far lower level. With development, Przeworski and Limongi (1997: p. 178) have argued, democracy can flourish in poor countries. This view is testified both by democratization of former Third World countries in the 1970s, 1980s, and 1990s and the maintenance of democracy in these states. Finally, a relatively advanced stage of socioeconomic development does not necessarily preclude the breakdown of democratic modernization as the collapse of the Weimar Republic in Germany in 1933 demonstrates (Wehler, 2003).

Determinants of Successful Paths to Political Modernization

The plurality of routes to political modernity mirrors a wide variety of factors that are conducive to, or inhibit, the effort to modernize a country. Among the factors that facilitate political modernization, sustained economic growth plays a major role, while the level of economic development is, beyond a certain threshold, not decisive. But sustained economic stagnation or a long period of economic decline is likely to block every route to political modernization. Moreover, the chances for successful political modernization tend to be higher in ethnically relatively homogeneous countries, and lower in more heterogeneous states. However, appropriate techniques of compromise building, such as consociationalism, can bridge sharp ethnic differences. The determinants of political modernization outcomes include political action on the part of collective and individual actors as well as actor constellations. For example, a distribution of power between traditional and modernization-oriented elites biased toward the latter is of importance. Large protest movements demanding political modernization are also relevant players. Supportive factors include a manageable sequence in which major system-development problems, or 'crisis of development,' generic to the political modernization processes are encountered, such as the formation of national identity and political legitimacy, a certain degree of centralization of political power, participation, social and political integration, and the use of government power for effective allocation and distribution of goods. Furthermore, international factors make a difference. Democratic political modernization, for example, flourishes within the context of a democracy-friendly international environment. The driving forces of successful political modernization include, as R. Inglehart's and C. Welzel's analysis of survey data from the Worlds Values Survey and the European Values Study suggest, a political culture in which 'self-expression values' prevail over 'survival values' (Inglehart and Welzel, 2005, 2009). 'Self-expression values' give priority to postmaterialist values and are conducive to attributes that are crucial to democracy, such as 'a culture of trust and tolerance,' participatory political orientations and preference for individual freedom. In contrast to this, the dominance of 'survival values' in a society's value system, i.e., values that give "top priority to economic and physical security and conformist social norms"(Inglehart and Welzel, 2009: p. 40) impedes political modernization.

Nonreflexive and Reflexive Modernization

Like other complex and dynamic processes, political modernization generates different types of outcomes – intended and unintended ones. Political modernization may result in different roles of the polity, such as 'polity dominance,' 'polity dependence,' or 'polity autonomy' (Coleman, 1968: p. 399), and it may be associated with consolidated or defective types of (Croissant and Merkel, 2004). Moreover, political modernization may result in a higher level of economic development, an advanced welfare state, and better governance, but it has also been associated with weak economic growth, limited social protection, and 'destructive dimensions' (Eisenstadt, 2011: p. 1612), such as repression, exclusion, wars, and generation of risks for future generations, for example, high ecological costs. The ambivalent tendencies of modernization manifest themselves most clearly in the theory of 'risk society' (Beck, 1986) and the theory of 'reflexive modernization' (Beck et al., 1997). According to these theories, the process of 'modernization of modernization' of former industrial societies has resulted in a dramatic increase of 'manufactured risks' (Giddens, 1999) in economically advanced countries.

Theories of reflexive modernization are important contributions to the study of advanced social and political systems. However, they share a weakness with other relatively general theories: They tend to disregard the large degree of variation, which according to comparative politics and comparative political economy, characterizes the structures of modernity, the nature of 'manufactured risks,' the political consequences of classical and reflexive modernization, and the political capacities to cope with effects and side effects of modernization processes.

Bibliography

Apter, D.E., 1965. The Politics of Modernization. University of Chicago Press, Chicago, IL.

Beck, U., 1986. Risikogesellschaft. Auf dem Weg in eine andere Moderne. Suhrkamp, Frankfurt. Risk Society: towards a new modernity. Sage, London (1992).

Beck, U., Giddens, A., Lash, S. (Eds.), 1997. Reflexive Modernization. Politics, Tradition, and Aesthetics in the Modern Social Order. Polity Press, Cambridge, UK.

Castles, F.G. (Ed.), 1993. Families of Nations. Patterns of Public Policy in Western Democracies. Dartmouth, Aldershot, UK.

Castles, F.G., Leibfried, S., Lewis, J., Obinger, H., Pierson, C. (Eds.), 2010. The Oxford Handbook of the Welfare State. Oxford University Press, Oxford.

Coleman, J.S., 1968. Modernization II political aspects. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences, vol. 10. The Macmillan Company & The Free Press, New York, pp. 395–402.

Croissant, A., Merkel, W. (Eds.), 2004. Consolidated or defective democracy? Problems of regime change. Democratization 11 (5).
Eisenstadt, S.N., 2002. Multiple Modernities. Wiley, New York.
Eisenstadt, S.N., 2011. Modernization theory. In: Bardie, B., et al. (Eds.), International Encyclopedia of Political Science, vol. 5. Sage, Los Angeles, pp. 553–559.
Giddens, A., 1984. The Consequences of Modernity. Cambridge University Press, Cambridge, UK.
Giddens, A., 1999. Risk and responsibility. The Modern Law Review 18 (1), 1–10.
Huntington, S.P., 1966. Political modernization. America versus Europe. World Politics 18, 378–414.
Inglehart, R., Welzel, C., 2005. Modernization, Cultural Change, and Democracy. The Human Development Sequence. Cambridge University Press, Cambridge, UK.
Inglehart, R., Welzel, C., 2009. How development leads to democracy. What we know about modernization. Foreign Affairs 88 (2), 33–48.
Lipset, S.M., 1994. Political Man. The Social Basis of Politics, fifth ed. John Hopkins University Press, Baltimore, MD.
Przeworski, A., Limongi, F., 1997. Modernization. theories and facts. World Politics 49, 155–183.
Therborn, G., 1995. European Modernity and Beyond. The Trajectory of European Societies 1945–2000. Sage, London.
Wehler, H.-U., 2003. Deutsche Gesellschaftsgeschichte, vol. 4. Vom Beginn des Ersten Weltkriegs bis zur Gründung der beiden deutschen Staaten 1914–1949. C.H.Beck, Munich.
Welzel, C., Inglehart, R., 2007. Mass beliefs and democratic institutions. In: Boix, C., Stokes, S. (Eds.), The Oxford Handbook of Comparative Politics. Oxford University Press, Oxford, pp. 297–316.

Modernization and Modernity in History

Paul Nolte, International University, Bremen, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 15, pp. 9954–9961, Elsevier Ltd., with revisions made by the Editor.

Abstract

The article analyzes 'modernization' and 'modernity' as key concepts in history and the social sciences, from their origins in eighteenth- and nineteenth-century sociology and historical thinking through the late twentieth century and especially after 1945. It describes the high tide of classical modernization theory after World War II and its subsequent usage in different historiographical traditions (particularly the United States and Germany). The article then deals with the critique, both theoretically and normatively grounded, of modernization theory since the 1980s, and the resulting shift toward concepts of (cultural) modernity and its contradictions. However, in recent years, a reappraisal of macrosociological modernization theories has occurred, for example, in the use of concepts such as 'reflexive modernization' or in the idea of a plurality of modernities in postcolonial theory.

Modernization and modernity have emerged, particularly after World War II, as key concepts in Western social science. They refer to a long-term process of change primarily in the Western world, but also outside the Euro-Atlantic hemisphere, by which traditional agrarian societies are transformed into the complex systems of industrial (and possibly democratic) societies typical of the late nineteenth and twentieth centuries ('modernization'). As a result of this process, societies and cultures develop a typical set of characteristics (e.g., in personality structure, organizations and institutions, and systems of belief and symbolic expression) that set them apart from older, premodern types of sociocultural systems ('modernity'). For many centuries, scholars have described and tried to explain – and also to intellectually support – the process of modernization in terms of concepts such as social differentiation, rationalization, or economic development. From the 1950s onward, such concepts have been more explicitly formulated as social-scientific theories of Western development, especially in American social science, but spreading very influentially from there to the academic cultures of other Western countries ('modernization theory'). As modernization describes a process of change in history, research in this field has proved to be a very important, almost paradigmatic area of cooperation and exchange between the social sciences and the discipline of history, and a particular branch of sociology called 'historical sociology' has centered around the concept of modernization as Westernization of societies, mostly during the past three to five centuries. Since the 1980s, modernization theories have increasingly been attacked – and often rejected – for their normative, Western-biased implications. At the same time, however, the concept of modernity has become more prominent in attempts at defining twentieth-century culture and its ambiguities. In addition, the 1990s have witnessed a renaissance of modernization theory in a reformulated framework of cultural skepticism and multiculturalism, an intellectual trend that is also significant for new attempts at 'bringing society back in' as the pendulum swings back from a radical emphasis on 'culture' in history and other social sciences.

Concepts of Modernity and Modernization before 1945

The term 'modern' (Latin *modernus*) has been in use since the fifth century AD to distinguish the 'new' or 'current' from the 'old' or 'antique' (Latin *antiquus*), especially as a means of describing and legitimizing new institutions, new legal rules, or new scholarly assumptions (Gumbrecht, 1978). Thus, the term 'modern' has been carrying some normative implications from its very inception to claim modernity for oneself meant a depreciation of the old or traditional and the attempt to win distance from it. In a very general sense and up to our own time, the concept of 'modern/modernity' is anchored in an attempt at defining one's own 'present' by differentiating it from the past. The conflict over the primacy of the new, or modern, over the old, or antique, has paradigmatically been fought out in the French *Querelles des anciens et des modernes* in the late seventeenth century; after the 'Querelles,' the superiority of the modern over the classic, or traditional, has rarely been fundamentally in doubt in the Western tradition of thought. The emphasis on the 'modern' gained additional support in the eighteenth century, when the enlightenment shifted historical thinking from cyclical to linear-progressive models. Since then, and especially after the American and French revolutions, the modern world was considered a world that in a radical and emphatic way was open into the future, a future that was to be attained by a natural progression of the cultural state of society.

In the wake of this fundamental shift, 'modern' was no longer just an opposition to the old or classic but came to be regarded as the long-term process of change that was called 'modernization' in twentieth-century social science. In that sense, the so-called 'stage theories' of historical development brought forward most impressively by writers of the Scottish enlightenment (Smith, Ferguson, and others) can be seen as the first explicit and scientific modernization theories; at the same time, they were founded on the claim of a Western superiority over colonial or 'savage' societies. During the nineteenth century, when developmental and progressivist thinking became a dominant pattern, the idea of modernization became

closely connected to evolutionary theory; the process of modernization was considered an ever-advancing evolution of Western traits leading toward a highly industrialized, rational, and bureaucratic social organization. This link to evolutionary theory prominently figured in the modernization theories of the mid-twentieth century but has lost much of its appeal since then. Probably, the most prominent, and certainly the most influential, variant of nineteenth century (evolutionary) modernization theories has been Karl Marx and Friedrich Engels' historical-materialist theory of Western development. It provided a consistent view of the long-term processes of change, identified mechanisms of social change and progress (namely, the evolution of the productive forces and the subsequent change of social (class) relations), and it underscored the importance of the transition to a capitalist industrial economy in the nineteenth century. In doing so, it provided a model for later modernization theories, be they Marxist or anti-Marxist, and at the same time, established a powerful 'master narrative' of Western history which patterned much historical writing in the twentieth century.

The most important bridge between those 'classical' theories of modernization (from the enlightenment to Marxism) and the more specifically scholarly scientific approaches to modernization that developed in the mid-twentieth century was built by the founding fathers of modern sociology around the turn of the century, especially by Emile Durkheim and Max Weber. Weber's idea of modernization as occidental rationalization, empirically grounded in his political sociology and his comparative sociology of religion, proved enormously influential in systematic and historical sociology – from Talcott Parsons to Reinhard Bendix and S.N. Eisenstadt – as well as in history. Weber succeeded in combining scope and historical depth with analytical concepts; he evaded the pitfalls of Marxism by giving weight to cultural, religious, and intellectual factors as well as material factors in his view on the specificity of modern Western development. Although the high tide of classical Weberian sociology, and of the adoption of Weberian categories in historical writing, may have been over since the late 1970s or 1980s, his thought remains influential in recent attempts at reformulating the ideas of modernity and modernization, as Weber is now seen less as an advocate of progress and rationalization and more as a shrewd observer of the deep ambiguities of modernization and modern life.

Modernization and Modernization Theory in the Social Sciences after 1945

After World War II, 'modernization theory' in a more limited, very specific sense emerged in the context of American social science, and American postwar political culture in general. With the 'American century' and the vision of a 'Pax Americana' reaching its climax, economic, political, and social development elsewhere in the world was increasingly viewed in the perspective of progressive Americanization that would, in the end, lead to a homogeneous system of modern industrial societies. The modernization of the West, and the USA in particular, was regarded a model for the future development of 'Third World' countries, their agrarian societies, and authoritarian political structures. The optimistic mood of social-scientific theories in this vein, their sometimes crude and simplistic assumptions about modernization, and their somewhat intellectually 'imperialist' stance toward non-European cultures have been much criticized since the late 1960s. In some areas of cultural theory and cultural history, they have been treated with contempt or, at best, with disregard. However, one should not ignore the fact that in the 1950s and 1960s, modernization theory marked an important step in the professionalization of concepts of social change and development and in the thinking about 'modernity' in general. Also, concepts of modernization developed in sociology and political science often radiated beyond the boundaries of their disciplines and proved very influential in other disciplines such as history.

The approaches of modernization theory between the 1950s and 1970s need not be described here in detail (for an overview and literature, see Lerner et al., 1968), but it may be useful to mention some fundamental concepts and assumptions, especially insofar as they inspired empirical work in historical sociology or history proper. The classical post-1945 modernization theory crystallized at a time when structural–functional theory in sociology stood in high regard, and when social and cultural problems were often discussed in terms of a theory of social systems. Harvard sociologist Talcott Parsons best epitomizes this approach; like many others, he attempted to link the somewhat static assumptions of functionalism with more dynamic Weberian concepts of rationalization and evolutionary change in Western history (see, e.g., Parsons, 1966). American modernization theory mostly concentrated on the transition from 'traditional society' to industrial capitalism in the economy and on the transition to democracy and mass participation in the realm of politics. For Western societies, this implied a strong focus on the nineteenth century, a time period which at that time also became increasingly important in the work of historians in many countries. While quantitative approaches in modernization theory were rarely applied in empirical historical studies, qualitative analytical concepts like theories of economic growth (e.g., Walt W. Rostow, Alexander Gerschenkron) and of political modernization, especially of state formation and nation-building (e.g., Karl W. Deutsch), proved attractive for historians interested in problems of social and institutional change.

In modernization theories of the 1950s and 1960s, a dichotomy of 'tradition' and 'modernity' often figured prominently; sociologists and political scientists established catalogs with typical features of 'traditional' societies (e.g., agrarian, localistic, and homogeneous) and contrasted them with features of 'modern' societies (e.g., urban, centralistic, and heterogeneous). Though this seemed to confirm a static, even ahistorical bias in modernization theory, the traditional–modern dichotomy, which in many respects traces back to Ferdinand Toennies' distinction between *Gemeinschaft* and *Gesellschaft*, became a widespread analytical tool in history and historical sociology in the 1960s and 1970s. Concepts such as this, focusing on the transition from tradition to modernity through a series of crises which modernizing societies had to undergo, proved attractive to historians who were interested in the history of revolutions or in the emergence of participatory democracy. They also triggered a wave of empirical work in

historical sociology between the late 1950s and the early 1970s in which historical processes of transition to modernity were analyzed, as in Neil Smelser's application of Parsonian theory to English industrialization (Smelser, 1959). Other scholars made use of Weberian concepts of comparative change in Western and non-Western (e.g., China and Japan) societies (see, e.g., Eisenstadt, 1973; Moore, 1966). This line of research has never been completely cut off, but it grew thinner and, in the 1980s and 1990s, stood under an ever larger shadow of cultural theory and anthropology, that is, of disciplines and approaches which fundamentally questioned the assumptions of macrosociological Western-oriented modernization theory.

Modernization and Modernization Theory in Historical Work

Historians in many Western countries have found theoretical concepts on modernity and modernization increasingly interesting since the 1950s, and especially during the 1960s and 1970s. There are several reasons for this. First, the postwar modernization theories were more than an intradisciplinary project; they served as an expression of a general mood and intellectual climate – of progress, of Western success, etc. – that did not leave academics from other disciplines, particularly in the humanities, unaffected. Second, the discipline of history at that time was very much in search for new approaches, for new master narratives, which would be able to support the ongoing broad expansion of history into the fields of social and economic history. At the same time, more historians became interested in comparative or international history, and modernization theory offered ways and concepts of pursuing such approaches. Third, the period of the 1950s to the 1970s was characterized by an intense cooperation between historians and social scientists; historians were looking for theoretical models in sociology and political science, while on the other hand, students of these disciplines attempted to treat their subjects – the formation of societies, the building of nation-states, or the emergence of party systems – with more historical depth (Kocka, 1996). In a very general sense, in which 'modernization theory' is not understood as a coherent theoretical concept, but more as a diffuse historical 'plot structure' of societal modernization, economic progress, and political democratization, it may indeed be said that "The quintessential professionalized historical discourse of the first postwar decades was modernization theory" (Applegate, 1999: p. 1162). However, this statement could also be made for much of nineteenth- and early twentieth-century liberal ('Whig') historiography; and even in the 1960s or 1970s, it was always a small minority of historians who made explicit use of social-scientific concepts of modernity and modernization in their empirical work. Also, national cultures and traditions of historiography diverged very much in that regard. While Anglo-American 'empiricist' culture generally worked as a barrier against the intrusion of theoretical concepts in historical writing, German scientific culture, with its strong traditions of fundamental theorizing, encouraged the adaptation of such concepts. American (USA) and German historiography can serve as examples for a somewhat closer look at the role of modernization theory in the discipline of history.

The United States of America

The fact that the USA was the seedbed for the most influential modernization theories after 1945 might induce the assumption that concepts of modernization have been widely adopted by American historians, but that is not the case, at least not with modernization theories in a more limited social-scientific meaning of that term. It is true that American historiography had a long tradition of 'progressivist' interpretation of its nation's history, even before the rise of the 'Progressive Historians' at the beginning of the twentieth century. Indeed, it is safe to say that a progressivist impulse constituted the very beginnings of American national historiography in the early nineteenth century (e.g., George Bancroft). After the nineteenth-century idea of progress had lost much of its original persuasion, modernization theory offered a somewhat more sophisticated replacement. Yet, as Dorothy Ross has maintained, "Its ideological use, reductionism, and historical determinism made historians wary of it from the start," and "Only few historical studies have in fact openly claimed modernization as the theoretical basis of their work" (Ross, 1998: p. 93). On the other hand, central assumptions of modernization theory often tacitly informed 1950s historiography, particularly in the then predominant school of 'consensus history,' which sought to establish an image of America as a country of conflict-free modernism. While it can be argued that in this perspective America had never been truly 'traditional' as European feudal societies were influential consensus historians such as David Potter drew on social-scientific notions of a modernizing American character; Potter (1954), for example, explicitly referred to David Riesman's theory of a transition from 'inner-directed' to 'other-directed' man. Similarly, it is very common among American historians to point to a fundamental dichotomy between *Gemeinschaft* and *Gesellschaft* and to use this distinction for describing modernization processes or historical developments that lead in the direction of a growing social complexity and differentiation. In what can be regarded as one of the most important historical case studies on American modernization, Robert Wiebe has argued for a fundamental change in the organizing principles of American society in the decades around 1900, a change which he describes as the transition from 'island communities' with their communal values to a technological and industrialized order shaped by the bureaucratic mentality of a new professional middle class (Wiebe, 1967). Still, American historians have generally been hesitant in explicitly using modernization theories from sociology or political science; even more so since around 1980, when the culturalist turn in the humanities generated additional skepticism about master narratives of progress and linear development.

Germany

In parts of the West German historical profession during the 1960s, an intellectual climate very favorable to the reception of ideas on modernization in general, and of social-scientific theories of modernization in particular, developed. It was part of a broader, international tendency toward social and economic history in the postwar decades, at the expense of the more traditional political and diplomatic history, but probably

even more important, were factors specific to German history and to the German political situation after the Third Reich. As a younger generation of academics began to orient themselves toward the West, especially the USA, they opened up for new theoretical and interpretative concepts brought forward by American social science, and modernization theory with its broad perspectives on comparative social change was regarded as one of the most intellectually promising programs (Wehler, 1975). The experience of fascism and the Holocaust in Germany lead many historians and social scientists to ask for long-term developments in German history responsible for the German inclination toward authoritarian government in the twentieth century, and the answer was found in failures of the modernization process in Germany especially since the mid-nineteenth century, after the revolution of 1848 and Bismarck's unification of Germany.

In a book enormously influential with the general public and professional historians alike, sociologist Ralf Dahrendorf eloquently elaborated the notion of Germany as a country being unable to adapt its institutional structures and its mentalities to the necessities of a 'modern' society. Paradoxically, Dahrendorf argued only the results of the National Socialist regime finally threw Germany into modernity (Dahrendorf, 1965). In the following two or three decades, much of nineteenth- and twentieth-century German social history was influenced by an interpretive framework of a German *Sonderweg* into modernity, that is, a specific course of German history characterized by deficient modernization, with the transition to democracy and to modern civil values lagging behind other Western countries (see Kocka, 1999). As German social historians, more so than their colleagues in the UK, France, or the USA, were interested in theoretical concepts of social history and social change, they also undertook efforts at explicitly using social-scientific modernization theories in their empirical work and in their writing of history. The most ambitious and widely discussed project in this vein is Hans-Ulrich Wehler's multivolume *Deutsche Gesellschaftsgeschichte*, a broad attempt at synthesizing modern German history with a focus on social and economic structures and with a discussion of Germany's specific modernization problems as its red thread and intellectual core (Wehler, 1987–1995).

In Wehler's work as well as in German history in general, however, Weberian concepts of modernity and of comparative social change increasingly superseded, especially since the 1980s, concepts from contemporary American social science. A Weberian perspective seemed to offer more historical depth, going beyond the assumption of a sharp break between the 'traditional' and the 'modern' world in the nineteenth century. Even more important, as naively optimistic versions of modernization theory became less persuasive in the 1980s, Max Weber's emphasis on the ambiguities of modernization, on the complicated web of gains and losses called forth by the Western rationalization process, looked more attractive to many historians. In this perspective, the failures of German history culminating in the Third Reich were not simply caused by a modernization lag but were, at least in part, inherent problems and results of modernity itself. This fundamental shift in the assessment of modernity and modernization, however, was not restricted to Germany but can be traced as a general trend in historical and social-scientific thought since the 1980s.

From Modernization to Modernity: Critique and Reappraisal since the 1980s

Since around 1980, postwar concepts of modernization and modernity have come under heavy attack, in history as well as in the social sciences and humanities in general. Within the overall trend toward a more critical, or skeptical, appraisal of modernization in history, it is important to distinguish several strands, even if they often overlap in historiographical practice.

In many Western countries, the history of what in English-language usage is called the 'early modern' period, that is, the history of the sixteenth to eighteenth centuries has witnessed an unprecedented expansion since around 1970 – particularly in Germany and the USA, perhaps less so in the UK and France (where the history of this period is called *histoire moderne*). As social historians have taken a much closer look at societies which according to social-scientific modernization theory are 'traditional' and 'static' societies, it appears that Western societies before the industrial and democratic revolutions were not so traditional after all but quite dynamic and expansive. This has been a heavy blow to theoretical and historical assumptions about a sharp break between traditional and modern societies that presumably happened (in North America and Western Europe) in the decades around 1800. It has thus prompted historians to ask for 'modern' structures in earlier times and for continuities of 'tradition' in the industrial period, with the effect of further questioning historiographical models of modernization as teleological evolution and linear progression. Among interpretations of long-term change in Europe since the Middle Ages, German sociologist Norbert Elias' theory of the civilization process proved very stimulating among historians from the late 1970s, perhaps because his perspective on civilization is compatible with both 'optimistic' and 'pessimistic' viewpoints on modernization (Elias, 1976). A 'pessimistic' reading of Elias focuses on the process of growing 'discipline,' self-control, and oppression, in many ways similar to Michel Foucault's understanding of modernity that became influential at about the same time. Especially in French and Italian historiography on the early modern era, modernization was now regarded as a strategic project of literate elites for the suppression of popular culture and ordinary people (see, for example, Muchembled, 1978).

Along with the 'culturalist turn' in the humanities, with the emergence of cultural anthropology as a paradigmatic discipline for the whole field of humanities, the core meaning and content of the concept of 'modernity' has shifted from economy to culture. When historians, and scholars in general, talk about modernity, they now hardly ever refer to the economic modernity that originated in the European and North American industrial revolutions, as was the case between the 1950s and 1970s. Instead, they refer to the idea of a specifically modern culture, that is, to the idea that modernity is defined by certain attitudes and belief systems, by models of cultural expression as they have been shaped, for example, by the mass media. It is important to recognize that the idea of modernity as culture is by no means new; after all, the *Querelles*

of the seventeenth century were about culture, and a cultural approach to the understanding of modernity is central to writers like Baudelaire in the nineteenth century as well as to many neo-Marxian strands of thought in the twentieth century, from Gramsci to Adorno and Habermas. In recent years, the current experience of a new information and knowledge revolution has directed attention toward the role of science and knowledge in the formation of modern culture; again, it can be said that modernity now appears to be less a product of economic forces and more a product of systematic knowledge, its reproduction, and its application through the emergence of a system of scientific disciplines.

Two further changes go along with this shift from economy to culture. First, instead of using the term 'modernization' (and hence emphasizing a processual change), many cultural theorists and historians now prefer to talk about 'modernity' or 'modernism' because they have become uncertain whether the notion of the modern and the idea of progress and development should be linked closely to one another, as was the case with the 'older' modernization theories. Second, although 'modernity' may be used very widely to characterize cultural innovation in different historical times, there is a clear propensity to concentrate on the twentieth century and define the twentieth century as an age of modernity – whereas former modernization theories, as has been pointed out before, tended to see the nineteenth century as the crucial breakthrough period of the modern. Particularly in American and in German historiography, the beginning of the twentieth century is now regarded as a period of fundamental change in society and culture, as a time when modern, urban mass society as we still today know it, first crystallized. Instead of pointing to industrial production forces, historians have come to view the beginnings of a consumer society as a strong indicator for modernity (see, e.g., Dumenil, 1995).

Contradictions of Modernity

It has already been mentioned that postwar modernization theories have come under attack for their sometimes all too simplistic constructions, for their bias on Westernization, and for their claim that modernization unmistakably effected economic or political 'progress.' As historians have started to ask for the 'social costs' of modernization, for the experience of the 'losers' in this process, the balance sheet began to look different. However, the assumption now is often not just that modernization had some ambivalent consequences or side effects but that modernization and modernity in themselves carry some dangerous anticivilizatory and eventually catastrophic potentials. Like the concept of 'cultural modernity' sketched before, the notion of inherent contradictions of modernity is very much grounded in the historical experience of the twentieth century: namely, in the experience of dictatorship and totalitarianism, of mass murder and ethnic cleansing. Those dark sides of twentieth century history can no longer be persuasively explained as abnormal aberrations from an otherwise successful path of modernization; modernization and modernity rather look janus faced; they bear in them the potentials for both civilization and barbarism, and there is no way of neatly separating those two sides. In the 1980s, twentieth-century German history provided the first important paradigm for this interpretation. In research about Weimar and Nazi Germany, it turned out that modernism, or the quest for modernization, could be politically reactionary in both its intentions and its effects (Herf, 1984). During the 1990s, a fierce debate was fought among German historians over the question of the ostensible 'modernity' of National Socialism (see Frei, 1993). At the same time, the Holocaust perpetrated by Nazi Germany came to be seen as a paradigm for general contradictions of the modernization process; it hinted at contradictions not only apparent in German history but also in the history of Western civilization as a whole. Sociologist Zygmunt Bauman in a widely received book maintained that "The Holocaust was born and executed in our modern rational society" (Bauman, 1989: p. X). Hence, the Holocaust poses a question not just to German history but also to 'Western civilization' as a whole. The Holocaust, in other words, in the 1990s often serves as a metaphor for the potentially murderous implications of modernity and modernization in the Western trajectory.

From Modernity to Postmodernism?

Instead of reformulating modernization into modernity and pointing at its contradictions, a second way of dealing with the widespread discontents with the postwar modernization theories in the 1980s was to leave the 'modern' behind and claim a new era of 'postmodernism' and the reign of postmodern intellectual principles in the humanities and social sciences. In the fields of literature, political science, and cultural criticism, the impact of the postmodernist impulse has been enormous, but historical discourse has been much less influenced by it in all major Western countries. Although some claims about the 'end of history' or the 'end of modernity' resonated among historians, hardly, any serious attempts were undertaken to conceptualize 'postmodernity' as a distinct epoch in the context of historical periodization. Nor have there been many efforts at describing certain social or cultural phenomena of the twentieth century under the heading of postmodernism in a way that was comparable to the notion of a modernist culture that emerged during the same time.

As a methodological and theoretical concept, postmodernism was somewhat more successful. Borrowing from literary and cultural theory, historians in the late 1980s and 1990s discussed whether the time-honored claims of historiography to objectivity and factuality could still be sustained and whether the narrative master plots of history – of which 'modernization' was one of the most important – should be followed in historical writing at all (for a fair, yet critical discussion of these issues, see Appleby et al., 1994). In its resistance against the modernization master narrative of history, the postmodernist impulse shared some of the more general skepticism about linear-progressive theories of history and social development, but it soon turned out that any contemporary reformulation of modernization concepts, however critical it was considering the ecological crisis, ethnic mass murders, or the challenges of multiculturalism, would be best advised to stick to the category of modernity. For to claim a new, postmodern era did not eventually promise to be helpful in the attempt at explaining the often bitter consequences of

modernity and modernization, an attempt that most historians and social scientists still find necessary.

Toward a New Theory of Modernization?

While historians are still debating the end of modernization and the demise of the grand narrative of modernization in historical writing, the concept of modernization has witnessed a remarkable renaissance in sociology since the mid- to late 1980s (Mergel, 1997: p. 203). The experiences of ecological crisis and catastrophe on the one hand and of economic globalization and a worldwide diffusion of cultures on the other hand have made sociologists think anew about societal change on a macrolevel, on differences between tradition and modernity, and on the possible course and direction of social evolution at the turn of the millennium. The enormous dynamics of change in contemporary society is now being conceived as an equivalent to the major transition described by older modernization theories as a shift from agrarian to industrial societies. As the classical form of industrial societies is coming to an end, Ulrich Beck has argued, it is becoming 'tradition' itself, a tradition that is now undergoing a process of second-degree rationalization (Beck, 1986). Industrial society was only 'half-modern' and is now being dissolved in a process called 'reflexive modernization' because the modernization process is becoming reflexive by being reproduced and pushed forward in scientific and public discourses about the problems generated by it.

The concept of reflexive modernization has also been suggested by Anthony Giddens in what he calls an "institutional analysis of modernity with cultural and epistemological overtones" (Giddens, 1990: p. I; cf Beck et al., 1994). For Giddens, the dynamism of modernity derives from (1) a separation of time and space ('time–space distanciation'); (2) the 'disembedding' of social systems in an increasingly globalized world; and (3) a reflexive ordering and reordering of social relations, affected by continual inputs of knowledge (Giddens, 1990: pp. 16–17). In this second stage, or 'radicalized modernity,' social change is not any longer comparable to a car controlled by the driver but is like a 'juggernaut' – "a runaway engine of enormous power which, collectively as human beings, we can drive to some extent, but which also threatens to rush out of our control and which could rend itself asunder" (Giddens, 1990: p. 139). In Giddens' view, globalization is not simply a spreading of Western culture and economy throughout the world; yet he firmly remains committed to the idea of modernity as a Western project.

Other authors, who argue from a 'postcolonial' political and epistemological perspective, are more skeptical about the Westernness of the ongoing modernization process. As traditional spatial boundaries disappear, cultures overlap and identities become hybrid creatures with different cultural backgrounds. The result, then, is not a continuation of Western modernity but a new plurality of modernities ('multiple modernities') without a clear priority of the West (see Appadurai, 1996).

However, as a common denominator of all these approaches, an emphasis on the spatial reordering of social relations and on the complicated relationship between globalization and local identity formation emerges. The new theories about reflexive and global modernization are important because they shift attention from a somewhat static notion of cultural modernity back to modernization as a process in time, and that is, among other things, what makes them particularly promising in the field of history. On the other hand, their understanding of historical processes, especially their assessment of social change in the period of the first (industrial) modernization and prior to it, is often superficial and sometimes naïve. Historians are just beginning to discuss the new macrosociological theories of modernization, and perhaps a fruitful dialog between historians and sociologists will develop in the near future. The idea of modernization has weathered many storms and changed its appearance, but it seems that it still offers an indispensable clue to the historical and sociological analysis of long-term changes in society and culture.

See also: Differentiation: Social; Evolution, History of; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Modernity: Anthropological Aspects; Modernity: History of the Concept; Modernity; Modernization, Sociological Theories of.

Bibliography

Appadurai, A., 1996. Modernity at Large: Cultural Dimensions of Globalization. University of Minnesota Press, Minneapolis, MN.
Appleby, J., Hunt, L., Jacob, M., 1994. Telling the Truth About History. Norton, New York.
Applegate, C., 1999. A Europe of regions: reflections on the history of sub-national places in modern times. American Historical Review 104, 1157–1182.
Bartolovich, C., Lazarus, N., 2002. Marxism, Modernity, and Postcolonial Studies. Cambridge University Press, Cambridge, UK.
Bauman, Z., 1989. Modernity and the Holocaust. Polity Press, Cambridge, UK.
Bauman, Z., 2004. Liquid Modernity. Polity Press, Cambridge, UK.
Beck, U., 1986. Risikogesellschaft: Auf dem Weg in eine andere Moderne. Suhrkamp, Germany: Frankfurt am Main.
Beck, U., Giddens, A., Lash, S., 1994. Reflexive Modernization: Politics, Tradition, and Aesthetics in the Modern Social Order. Polity Press, Cambridge, UK.
Connerton, P., 2009. How Modernity Forgets. Cambridge University Press, Cambridge, UK.
Dahrendorf, R., 1965. Gesellschaft und Demokratie in Deutschland. R. Piper, Munich, Germany.
Dumenil, L., 1995. The Modern Temper: American Culture and Society in the 1920s, first ed. Hill and Wang, New York.
Eisenstadt, S.N., 1973. Tradition, Change, and Modernity. Wiley, New York.
Elias, N., 1976. Über den Prozess der Zivilisation, 2 vols. Suhrkamp, Frankfurt, Germany [The Civilizing Process].
Frei, N., 1993. Wie modern war der Nationalsozialismus? [How modern was Nazism?]. Geschichte und Gesellschaft 19, 367–387.
Giddens, A., 1990. The Consequences of Modernity. Stanford University Press, Stanford, CA.
Gumbrecht, H.U., 1978. Modern, Modernität, moderne. In: Brunner, O., Conze, W., Koselleck, R. (Eds.), Geschichtliche Grundbegriffe, vol. 4. E. Klett, Stuttgart, Germany.
Herf, J., 1984. Reactionary Modernism: Technology, Culture, and Politics in Weimar and the Third Reich. Cambridge University Press, Cambridge.
Himmelfarb, G., 2004. The Roads to Modernity: The British, French, and American Enlightenments. Knopf, New York.
Israel, J.I., 2001. Radical Enlightenment: Philosophy and the Making of Modernity, 1650–1750. Oxford University Press, Oxford.
Kocka, J., 1996. Annäherung und neue Distanz. Historiker und Sozialwissenschaftler seit den fünfziger Jahre. In: Hettling, M., Nolte, P. (Eds.), Nation und Gesellschaft in Deutschland. Beck, Munich, Germany.

Kocka, J., 1999. Asymmetrical historical comparison: the case of the German 'Sonderweg'. History and Theory 38, 40–50.

Lerner, D., Coleman, J.S., Dore, R.P., 1968. Modernization. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences, vol. 10. Macmillan & Free Press, New York, pp. 386–409.

Mergel, T., 1997. Geht es weiterhin voran? Die Modernisierungstheorie auf dem Weg zu einer Theorie der Moderne. In: Mergel, T., Welskopp, T. (Eds.), Geschichte zwischen Kultur und Gesellschaft. Beck, Munich, Germany.

Misa, T.J., Brey, P., Feenberg, A., 2003. Modernity and Technology. MIT Press, Cambridge, MA.

Moore Jr., B., 1966. Social Origins of Dictatorship and Democracy. Beacon Press, Boston.

Muchembled, R., 1978. Culture populaire et culture des élites dans la France moderne (XVe–XXVe siécle). Flammarion, Paris.

Osborne, P., 2005. Modernity. Routledge, London, UK.

Parsons, T., 1966. Societies: Evolutionary and Comparative Perspectives. Prentice Hall, Englewood Cliffs, NJ.

Potter, D.M., 1954. People of Plenty: Economic Abundance and the American Character. University of Chicago Press, Chicago.

Punter, D., 2007. Modernity. Palgrave Macmillan, Basingstoke.

Ross, D., 1998. The new and newer histories. Social theory and historiography in an American key. In: Molho, A., Wood, G.S. (Eds.), Imagined Histories. American Historians Interpret the Past. Princeton University Press, Princeton, NJ.

Smelser, N.J., 1959. Social Change in the Industrial Revolution: An Application of Theory to the Lancashire Cotton Industry 1770–1840. Routledge and Paul, London.

Wehler, H.-U., 1975. Modernisierungstheorie und Geschichte. Vandenhoeck & Ruprecht, Göttingen, Germany.

Wehler, H.-U., 1987–1995. Deutsche Gesellschaftsgeschichte, 3 vols. C. H. Beck, Munich, Germany.

Wiebe, R.H., 1967. The Search for Order 1877–1920, first ed. Hill and Wang, New York.

Modernization, Sociological Theories of

Sergei Gavrov, Moscow State University of Design and Technology, Moscow, Russia; and Moscow City Pedagogical University, Moscow, Russia
Igor Klyukanov, Eastern Washington University, Cheney, WA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R. Inglehart, volume 15, pp. 9965–9971, © 2001, Elsevier Ltd.

Abstract

Sociological theory views modernization as a macroprocess of transition from traditional society to the society of modernity. More specifically, the main sociological theories of modernization are based on the ideas of evolutionism, functionalism, and diffusionism. Attention is also paid to the theoretical conceptualization of modernization processes in non-Western societies, which can be explained by the interaction of endogenous and exogenous factors. Within the framework of scientific discourse and the public consciousness, theory of modernization is sometimes questioned as a view of a progressive development of the world. Today, the global spread of modernization is more adequately theorized in terms of historical and sociocultural dynamic shifts from institutionalized, vertically integrated forms of social solidarity to network forms and social movements. This view of modernization as a communication phenomenon that involves the changing ways of interaction and media use highlights its self-reflexive nature, showing how modernization can only gain an access to itself by a series of partial and conflicting descriptions.

Defining Modernization

Overall, modernization is understood as a macroprocess of transition from traditional society to the society of modernity, leading to the appearance of "modes of social life or organization which emerged in Europe from about the seventeenth century onwards and which subsequently became more or less worldwide in their influence" (Giddens, 1990: p. 1).

Today, the concept of modernization is understood in three different meanings: (1) as the internal development of Western Europe and North America relating to the European New Era; (2) as a process by which countries that do not belong to the first group of countries aim to catch up with them; and (3) as processes of evolutionary development of the most modernized societies (Western Europe and North America), i.e., modernization as a permanent process, carried out through reform and innovation, which today means a transition to a postindustrial society.

1. Speaking of the genealogy of the word 'modern,' German philosopher Jürgen Habermas notes that it was first used in Europe at the end of the fifth century in order to differentiate between the received official status of the Christian present and the pagan Roman past (Habermas, 1985a). In subsequent periods, the content of this understanding changed, but only the Enlightenment and Romanticism gave it the meaning that it has today. Since then, that which promotes objective expression of spontaneously renewing the actuality of the spirit of the time is considered contemporary and modern.

As a result of the accelerated development during the Modern times in Europe, a unique modern civilization was formed which radically differed from traditional societies. A fragmented political system of Europe created preconditions for the emergence of industrial capitalism. Modernity originated in Western Europe due to the formation of the Protestant work ethic, the market economy, bureaucracy, and legal systems. In Western Europe, the large-scale process of modernization – the transition from a traditional (preindustrial) society to a society of modernity – took several centuries, genealogically dating back to the industrial revolution in England, and the strengthening of the bourgeoisie as it gained political power from the revolutions in Britain (1640–42), the US (1776), and France (1789).

Usually, three periods of modernization are distinguished: the first period – the end of the eighteenth century to the beginning of the twentieth century; the second period – the 1920s–1960s; and the third period – the 1970s–1990s. According to some social theorists, e.g., Ulrich Beck and Anthony Giddens, we are living a period of late modernity, its consequences having only intensified (Beck, 1998; Giddens, 1990). For Beck, for instance, the so-called simple modernization resulted in industrial society; when continued, the modernization of industrial society leads to the creation of a risk society.

Some authors believe that (contemporary) modernity cannot be completed at all. Thus, the Senegalese sociologist Samir Amin argues that 'modernity' is unfinished, but it opens the doors to the unknown. Modernity is inherently incompletable, but it involves a sequence of diverse forms, which contradicts the society at each moment of its history (Amin, 2000: pp. 581–622).

Modern society includes four basic institutions: a competitive democracy, a market economy, a welfare state, and mass communication. A market economy is the basis of an autonomous civil society, which overcomes all boundaries and creates an open society. Unlike other more detailed studies of cultural anthropology in traditional societies, modern societies are based on the principles of suffrage law, universalization of civil rights, the institutionalization of social change, secular culture and secularism, urbanization, the autonomy of subsystems, rationalization, dominance of the market economy, bureaucratization, professionalization, mass literacy and mass media, and the growth of social and professional mobility.

It is especially important to note that "modernity always had something to do with technology …, and thus eventually with progress" (Jameson, 2002: p. 7). The scientific revolution of the seventeenth century and technological advances led to the transformation of community members in the citizens' 'imagined community' of the nation-state. Distinctive features of modernity are: within the political sphere – a democratic, constitutional state; within the sphere of nation-building – a transition to a national government; within the spheres of science and education – a formation of autonomous studies; and within the economic sphere – a transition to capitalism. The universal packaging of modernizing transformations, characteristic of the twentieth century, operates in several spheres: at the level of a socioculturally organized society, modernization manifests itself in the form of a movement from industrialism to postindustrialism in the economic sphere; in a movement from an authoritarian to a democratic regime in the political sphere; and in the legal sphere as a transition from a conventional to a legal right. Changes in the sphere of religion are of particular significance: modernity is considered to have came into being because people had taken the world out of God's hands, as it were, while still providing possible manifestations of enchantment in the form of contingent meaning.

In accordance with the definition given by the famous English expert in the field of modernization transformation, V. Moore, modernization is "a total transformation of the traditional premodern society into a social organization, which is characteristic of the 'advanced,' economically prosperous and political stable nations of the West" (Moore, 1963: p. 89). However, professor of sociology at the University of Munich, Ulrich Beck, believes that "modernization leads not only to the formation of a centralized government, to the concentration of capital and more subtly interweaving of the division of labor and market relations, to mobility, mass consumption and so on, but also – and here we come to a generalized model – to a triple 'customization': freedom from historically imparted social forms and relationships in the traditional sense of circumstances of domination and liberation ('aspect of freedom'), a loss of traditional stability in terms of effective knowledge, faith and accepted norms ('aspect of demagnification') and – which somewhat inverts the concept of this understanding – to a new kind of sociocultural integration ('aspect of control and reintegration')" (Beck, 1998).

2. The second meaning of modernization denotes the various processes of development of lesser developed catching up to developed countries. Modernization is thus seen as a response to the calls of Western civilization modernity, to which every society gives an answer in accordance with the principles, structures, and symbols laid down as a result of its historical development. In this sense, the term 'modernization' refers to the underdeveloped societies and the description of their efforts in order to catch up with the leading, more developed countries, which coexist with them in the same historical time. In this case, the concept of 'modernization' describes the movement from the periphery to the center of modern society. The theories of modernization, neomodernization, and convergence operate within the term 'modernization' in this narrow sense. Many theorists wrote about the differences and interactions within the premodern (preindustrial) and modern societies in the nineteenth century (Tönnies, 1887; Spencer, 1873; Durkheim, 1893; Comte, 1853).
3. Finally, in the third meaning, modernization is understood as a process of innovative transformations of the most developed countries of Europe and North America, which first began the process of modernization and have long since been deeply rooted in modernity. For example, renowned Israeli sociologist Eisenstadt believes that "historical modernization is a process of change leading to two types of social, economic and political systems that have developed in western Europe and North America between the XVII and XIX centuries and has spread to other countries and continents" (Eisenstadt, 1978). There is also a large body of work on the subject of the transition to a postindustrial society (Galbraith, 1958; Inglehart, 1977; Thurow, 1980; Fukuymama, 1992; Handy, 1993; Bell and Inozemtseva, 2007).

The History of the Study of Modernization

Modernization as a sociocultural macroprocess has its own theoretical foundation – the modernization theory, the formation of which was influenced by evolutionism, functionalism, and diffusionism. Fundamental contributions to the formation of scientific concepts explaining the macroprocess of modernization, i.e., the transition from traditional to modern society (Comte, 1853; Spencer, 1873; Weber, 1905; Durkheim, 1893; Tönnies, 1887; Cooley, 1918; Maine, 1861).

The theories of modernization in their classical form received scientific and public recognition in the 1950s and mid-1960s, when the work of the following authors became widely known (Smelser, 1956; Lerner, 1958; Rostow, 1960; Parsons and White, 1961; Hagen, 1963; Apter, 1965; Levy, 1966; Berger, 1967; Eisenstadt, 1978).

Within the work of functionalists, it is worthwhile to mention the work in classic American and global sociology by Talcott Parsons, who examined the processes of segregation in imported sociocultural experiences in countries that were in the process of modernizing (1961). Parsons believes that in the constant attempts to divide importable, foreign culture experiences into acceptable and not acceptable, a tendency to preserve cultural values "of the highest level is displayed, while concurrently opening the road of radical changes at the next level of value specifications, i.e., on the level of major functional subsystems" (Parsons and White, 1961: p. 662).

We know that cultural anthropology came from studying traditional, archaic forms of human coexistence; this is seen in the works of classical cultural anthropologists (Herskovits, 1940; White, 1949; Kroenbar, 1952). In cultural anthropology, the evolution of traditional, local cultures is seen in two main forms: (1) As a linear stadial evolution of a progressive nature from relatively simple societies to more complex ones. This understanding correlates to the classical understanding of the modernization process. These views are more or less shared in England (Tylor, 1870; Spencer, 1873; McLennan, 1869; Lubbock, 1870; Frazer, 1935), Germany (Bastian, 1860;

Lipset, 1960), France (Letourneau, 1894), and the US (Morgan, 1877). (2) As a multilinear development of different types of cultures. In the latter case, a greater emphasis is placed on the uniqueness of the modernization process and the various kinds of modernity that subsequently arise. Modernization is thus viewed more as the implementation of various historically conditioned types.

Evolutionists, primarily Herbert Spencer (1820–1903), an English philosopher, biologist, psychologist, and sociologist, focused their theoretical constructions on analyzing how societies develop. Spencer outlined his views on the evolution of society in-depth in his fundamental work *Principles of Sociology* (1896). He and his followers paid close attention to progressive social changes, positive outcomes of the evolutionary nature of the modernization processes. They believed that modernizing transformations were unilinear, with lesser developed countries having to go through the same path that developed countries of modernity had already passed; the changes are gradual, cumulative, and peaceful. They stressed the importance of endogenous, immanent causes and described the driving force of the changing terms 'structural' and 'functional differentiation,' 'adaptive perfection' and similar evolutionary concepts. Piotr Sztompka, professor at the Jagiellonian University in Krakow, notes that from the standpoint of evolutionists – supporters of the modernization theory – it was supposed to bring universal improvement in social life and the human condition (1993). Modernization and convergence were seen as necessary, irreversible, endogenous, and beneficial processes. The path of modernizing transformation consists of successive stage segments, such as 'traditional–transitional–contemporary,' traditional – the stage of achieving preconditions for the beginning of change – the beginning of continuous growth – maturation – attainment of the level of mass consumption (Rostow, 1960).

As a result of the development process of modernization, modern became the crowning point of expressing progressivist cultural–civilizational paradigm. The ideology of progress, while becoming increasingly secular, defined the eurocentrism of the historical process throughout the period of modernity, which was indicative of various nations ascending the ladder to rationalism, secularism, and econocentrism. In a generalized form of presenting the views of classic sociopolitical thoughts on progress, Robert Nisbet said that as a whole, the classical concept can be seen as an idea of the gradual liberation of humanity from fear and ignorance toward a higher level of civilization (1980). A stadial approach, adopted for a long time till the last decades of the twentieth century by evolutionists, neoevolutionists, Marxits, and positivists, prevailed and, in an indirect form, is common today in Western science, which largely preserves the spirit of Eurocentrism (Harrison and Huntington, 2001).

Diffusionists interpreted the processes of development as predominantly diffusional, rather than endogenous-evolutionary in nature (Ratzel, 1898; Frobenius, 1898; Graebner, 1911). In contrast to the interpretation of modernization as a spontaneous tendency of evolution, which self-develops from the 'bottom,' diffusionists believed that it started and was controlled from 'above' by the intellectual and political elite which aims to pull its countries out of backwardness through planned, purposeful action. Diffusion acts as a mechanism for modernization change. Interaction between the more developed, modernized and the less developed and modernized societies is a crucial factor of modernization. In transforming countries where modernization is a desired goal, they look to the developed countries in the Western civilization. Consequently, modernization is not just a spontaneous development in a progressive direction. In this understanding, modernization presents itself as a straight and preferably precise transfer of foreign cultural norms, values, institutions, models of work, and leisure from countries that are in a reference group of their own. Modernization, therefore, is not a self-supporting, self-progressing process. Rather, it is the transfer of samples, models, and achievements of developed countries into their own.

Modernization of Non-Western Societies

Modernization processes in non-Western societies can be explained using evolutionary and diffusionist theories. Given the interaction of endogenous (evolutionary) and exogenous (diffusionist) components of the modernization process, the following classification can be presented:

Endogenous forms of modernization – a process, deterministic in endogenous sociocultural dynamics. Modernization is complex due to internal causes, self-developing, self-transforming societies. An example of this type of modernization, starting with the New Age, is the development of Western Europe and North America.

The adaptive ('catching up') type of modernization. Practiced in countries not related to the Western pioneers of modernization, beginning as an adaptive response to the accelerating processes of sociocultural dynamics within the Western civilization modernity, passing the scheme of challenge–response. It is divided into two subcategories:

- Modernization as self-Westernization. Initiated with the aim of achieving internal goals, which includes the need to bridge the technological gap with Western civilization's modernity and to preserve government autonomy. In turn, this type of modernization is divided into two subcategories:
  - Defense modernization. Conducted primarily to strengthen the military and political potential of the state, the change of the actual society acting more as a secondary process to technical and technological borrowings. Under this model, strengthening the modernization of the state is the absolute goal and people are simply a means to an end. An example of this is the Russian modernization from Peter I to the end of the Soviet period, excluding the reforms of Alexander II and the sociocultural transformations of the post-Soviet period. R. Bendix believes that only the "first modernized state has a chance to follow its own path autonomously without external pressure. All other nations are influenced by the struggle between avantgarde and rearguard modernization. The threat of military defeat forces struggling countries (Russia, the Ottoman Empire, or Japan) to reform their bureaucracy and army. Such modernization is described as defense modernization"

(Bendix, 1964). This is extensive modernization, i.e., appropriation and assimilation of foreign cultural achievements, borrowing from the results of innovation without acquiring the ability to actually innovate, which cannot be borrowed.
- Liberal modernization. Carried out in order to change society and liberate people, with the perception not only of the technical and technological tools of Western civilization's modernity, but also the processes that led to the creation of this instrument, Western in the perception of its genealogy institutions, norms, values, and models of behaviors. A classic example is the reforms of Alexander II, as well as the modernization processes of Russia and the former socialist countries of Eastern Europe in the (Gavrov and Гавров, 2004).

- Modernization under foreign custody. It presents itself as the transformation of national sociocultural systems, carried out with the direct or indirect participation of the state or any states belonging to Western civilizations of modernity. As a result, geopolitical sovereignty is not saved. In turn, this type of modernization is divided into two subcategories:
  - Modernization in the form of partial responsibility. Carried out in the spirit of colonial and semicolonial policy, with one or more sectors of the economy serving the interests of the colonial metropolis economy, while other sectors do not receive enough momentum to develop. As a rule, modernization of colonies is not considered or formulated as a problem of policy. Development of social and cultural spheres managed by the state is a result of the activities of colonial administration. An example of this is the British colonial rule in India, which resulted in the formation of a Westernized Indian elite, thus beginning the struggle for liberation from colonial dependence.
  - Modernization in the form of systemic involvement, when one or more states initiate and assume full responsibility for the modernization processes on territories under custody. Examples of this are Germany and Japan after World War II.

This classification is conditional; in practice, there is an interweaving of internal and external factors determining the processes of modernization, although the diffusing component dominates in the modernizing process of transforming non-Western societies.

In the 1970s and mid-1980s, the concept of progress, in particular the theory of modernization, was generally subjected to considerable reevaluation. Within the framework of scientific discourse and the public consciousness, it is questioned as a model of an evolutionary arrangement of the world, its modernization, as well as the concept of progress. Critics of the modernization theory pointed to the low efficiency of modernization transformations in the Third World and their partial or complete rejection. It was recognized that a significant variety of nations of modernity is persisting and likely to continue doing so. There is a contradiction between the uniqueness of the West and the fact that it was a model for the rest of the world, and, on the other hand, the specificities of the dynamics of other civilizations were not fully recognized during the time of Marx or Weber when the spread of capitalism and modernization outside of Europe was in the early stages. However, it has become much more evident in the later stages of modernization after World War II (Eisenstadt, 1978).

As a result of the spread of the modernization processes in different regions of the world, it turned out that the process and the results of modernization are influenced by the sociocultural traditions of the recipient countries. Transforming societies change foreign cultural innovations, yielding hybrid designs in which there are residual elements of the imported foreign culture materials and local sociocultural traditions. Authenticity in relation to the initial samples and practical efficiency of hybrid designs are usually not high.

Global Spread of Modernization

The twentieth century saw a global spread of modernization processes and an institutional environment of modernity. Until the second half of the twentieth century, the real superiority of the Western civilization, the geographical area where modernity was born, presided over the world predominantly and wholly. As an alternative project for most of the last century was a socialist project, representing itself (especially in the USSR) as a desperate attempt to achieve the quantitative economic indicators in governments of modernity. Socialist variations of modernization that were practiced during the twentieth century in various countries of the world, especially in the USSR, were (despite all the ideological differences) only one of the branches of the general modernization process, acting as an adaptive reaction of non-Western countries. It is not coincidental that the traditional area of modernity was perceived as a socialist modernization option since it correlated with the collectivist mentality of non-Western countries. As close as socialism is close to the collectivistic Asia, it is just as far from the individualistic Euro-Atlantic civilization, where its chances are vanishingly small. The revolutionary situation developed in Germany and Hungary after the October Revolution in Russia was the result of a defeat in World War I and a consequence mainly of autochthonous development.

The second alternative project, not claiming universality, was the project of National Socialism. It should be remembered that the processes of modernization present a synthesis of the progressive and inverse dynamics, even though when considering it in a sufficiently lengthy diachronic context, it is dominated by progressive dynamic processes. As a dialectical combination of differently oriented dynamic processes, an example of negative adaptive responses to the challenge of Euro-Atlantic civilizations of modernity can be seen as experience of the totalitarian regimes of the twentieth century, including, above all, the experience of Nazi Germany and Stalinist USSR. Using technological achievements of modernity while denying democratic forms of social organization and human rights, is a severe regulation of the creative potential of individuals.

In the second half of the twentieth century, the situation radically changed as Germany, starting with its Western part, was becoming a key element in the construction of European modernism. Other countries of the European continent, including Italy, Spain, later Poland, Hungary, and other former

socialist countries carved their path in this direction, gradually integrating into the Euro-Atlantic civilization of modernity. The rest of humanity, defined in the years after World War II as the 'Third World,' not only failed to offer a universal, alternative project, but also effective, local development options for non-Western states.

Economically and socially successful groups of Asian nations, with modernization processes within the framework of Islamic civilizations, worked toward building an economy while taking into account their activities in religious institutions, such as prohibition of levying interests on loans within the Koranic tradition.

The Western civilization's period of modernization is increasingly developing on a horizontal plane, leaving behind its gradually diminishing vertical dimensions. Note that under the vertical direction, we understand the value of sacral-transcendental dimensions, and under the horizontal dimension – the value of the earth's material life. This departure from the sacral-transcendental nature is seen in the transformation and filling of the new, secular content, Christian in its genealogy. Non-Western civilizations were long unable to answer the call of the Western period of modernization, largely because of the priorities of their civilizations, which are encompassed in the vertical dimension above the horizontal. Within the modernization process, a desire to 'catch up and overtake' the state of modernity without losing or, on a smaller scale, not radically transforming traditional civilizational identities, is seen as problematic.

In the case of a practical implementation of these transformational projects, achievable at the cost of extreme tension of the strength of the society and state, the system with a multivector direction of movement cannot effectively function on any of these planes. Infantilism and mythology of mass consciousness continues today and not just in Russia, with a focus on the possibility of reaching civilizational benefits and a standard of living, yet without any significant sacrifices.

Attempts at modernization in the planar civilizational dimension (which includes the economy, forms of social organization, technology, etc.) while maintaining the primacy of the vertical dimension of civilization (with its spiritual and religious components) is an unattainable utopia. However, the desire to preserve civilizational identity leads to an individualized relationship between the local (civilizational) and the universal (stadial). In turn, this relationship with respect to the local civilization is not static, instead being determined by the intensity experienced by endogenous historical and sociocultural dynamics, adapted by its character in relation to the dynamics of exogeny produced within the framework of the Western civilization period of modernity. The attempt to converge archaic traditionalism using scientific, technical, and organizational achievements of modernity turned out to be unsuccessful and destructive. Modernization in Nazi Germany, Fascist Italy, and the Leninist–Stalinist USSR paradoxically were against modernity in its classic sense.

Militarization, social mobilization, the cultivation of national unity, intolerance of political, ethnic, sexual minorities – all this can be seen as rejection of an important part of the ontological content of modernity – political democracy and human rights, the priority rights of the states, and the nation under human rights – "You are nothing, your people are everything." Such rejection of ontological content of modernity, adopted and overused, engenders technical–technological, bureaucratic, and penitentiary innovations.

Unlike German projects, the Soviet alternative project of modernization and modernity claimed universality, and therefore was international. Going back to the ideological constructions of Karl Marx, in the practice of state building, Marxism was revised by the social and cultural transformations. A Marxist revolution happening in an underdeveloped peasant agrarian country, which is what Russia was at the time, became a challenge to classical Marxism. Soviet modernization became a classic example of defense modernization. This is a complex version of modernization as an attempt to redesign the ontological foundations of modernity, not with the goal of protecting the old providential, but in order to create based on the belief in Communism and the Communist Party.

Today, it can be argued that globalization is a modern stage from which we deploy a single and universal 'project of modernity.' Globalization, manifested in the movement to integrate economic, legal, informational, educational, and ultimately, cultural space, is the current stage of modernization. There is a whole vector stream of modernizing/globalizing changes moving toward an interconnected, mutually supported world that is based on the institutions and values of Western civilization modernity.

Globalization is a large-scale process of disseminating institutional and value-normative environments on a global scale, which leads to greater access of the borders of nations and to a significant weakening of national sovereignty when public functions are transferred to the transnational level. This process is due to various reasons, most notably the emerging reductions of the functions of national governments to functions of local self-government (postal services, part of the system of education, police functions, etc.). The most important of these is the depth of geopolitical sovereignty on its territories; subjected to rethinking, evolution occurs in the direction of reducing the powers of the nation-state. In the period of late modernity, fewer and fewer states can earnestly talk about ensuring national security with the help of armed forces; economic, educational, and fiscal space is increasingly taking on a supranational nature and becoming less controlled at the national level.

In the process of globalization, the identity of modernized people goes beyond national borders, taking on a transnational form of self-identification. This increases the level of freedom, transforming the national community and cultural traditions. Forms of social prestige, referential models of behavior, along with norms and values, are all modeled and translated within a framework of emerging global information and cultural space.

Today, vectors of historical and sociocultural dynamics shifted from institutionalized, vertically integrated forms of social solidarity to network forms (Castells, 1997) and social movements, global and Western in their genealogical norms, values, and models of behavior. People of the age of late modernity can identify themselves with transnational movements, as well as with antiglobalists, environmentalists (e.g., Greenpeace), sexual minorities, youth movements of the nonconformist types, owners of vehicles, washing machines, shampoos, users of certain brands and services, etc. A person can act as an owner of a house built as a specific project, a car of a certain model, make a choice between Pepsi or Coca-

Cola, root for a football team which is often not the team of the person's country, etc. Moreover, new forms of self-identification and solidarity are identifying with the cultural products produced globally by the media. A person can be aware of himself or herself as part of a prestigious social group that is created and reproduced with the help of various forms of advertising.

Today, it can be argued that the large-scale modernization that arose in the civilizational space of Western civilization is similar to the spread of Greek culture in the ancient world. Despite the multiplicity of interpretations of modernization transformation, most of the transformations of the world can be historically compared to the times when the Greek cultural tradition became the basis for common traditions of Roman antiquity. In the new stage of historical development, something similar is happening in relation to modernity, where by means of diffusion, modernization is extended from the Western civilizational arena of modernity.

Modernization as a Communication Phenomenon

Modernization as a macroprocess of transition from traditional society to the society of modernity can be theorized in terms of three discourses, each differently conceptualizing modernization as a communication phenomenon that involves the changing ways of interaction and media use.

The first discourse is a discourse of representation, grounded in the Cartesian legacy and the ideas of positivism. In this view, communication is conceived as a process of the production and transmission of truth claims. From this perspective, modernization is evaluated in terms of its effectiveness and not as a process of democratic participation and decision making. It is from this perspective that many ideas were promoted through mass media that were considered essential to modernization, such as technical innovation, functional differentiation, the power of reason, etc.

Modernization was equated with the transfer of such ideas from a modern, industrialized West, to the less-developed societies. In fact, in 1961, the United Nations declared the decade of the 1960, the Decade of Development. As President Kennedy put it in launching the Decade at the UN in New York, "To those peoples in the huts and villages of half the globe struggling to break the bonds of mass misery, we pledge our best efforts to help them help themselves." The theoretical platform for such efforts was provided in such scholarly publications as David Lerner's *The Passing of Traditional Society* and Wilbur Schramm's *Mass Media and National Development*.

The second discourse – a discourse of interpretivism – is a reaction against the view of modernization as 'the magic bullet' with an explicit or implicit Western bias. The discourse of interpretivism does not see a modern society only in terms of such indicators as its level of media development or GNP, calling to consider other factors such as income disparities, etc. According to this view, mass media are not assumed to be a neutral force; behind all flows of information and goods, one must look for certain ideological forces at play and interpret their influences on the process of modernization. In addition, one cannot have complete faith in new information and communication technologies as a perfectly efficient vehicle of modernization. On the contrary, often developing countries come to be dependent on mass media in the core, modern societies: this is a key thesis of the media dependency theory (Schiller, 1976).

The discourse of interpretivism does not reject the ideals of the Enlightenment; rather, it tries to understand how that modernist vision can be saved and enhanced by opposing the interactionally constituted 'lifeworld' to instrumental rationality. In this respect, attention is shifted from a supposedly disembodied mind to a communication community of real-life people who do not live by representing an already existing reality, but engage with one another in various forms of communicative praxis. The discourse of interpretivism thus is based not on a correspondence, but consensus theory of truth, and the constitutive nature of language (Habermas, 1985a; Gadamer, 1960).

The third discourse is a discourse of reflexivity; it takes the critical thrust of interpretivism a step further by extending the examination of every action to the agent instigating it. As a result, the discourse of reflexivity aims at providing a constructive corrective to the nihilistic danger of deconstruction. No longer understood as a simple relationship between knowledge and control, modernization can only be understood as constant efforts to bind time and space through (self)-reflexivity. According to such theorists as Ulrich Beck and Scott Lash, (self)-reflexivity is crucial because uncertainty is inherent in every communicative action and modernization has a truly global, and potentially destructive, impact (1994). The exact dynamic of modernization, therefore, cannot be predicted and any claims for technological and scientific certainties, or 'final truths,' can only be theorized in the act of self-reflexivity.

In a way, modernization is always 'on the move.' Not surprisingly, more attention is now paid to new social movements with a heightened awareness of the transformative implications of their actions, e.g., democratic, environmental, and peace movements. Also, it is no surprise that calls are made, not only by activists but theorists as well, for new forms of public dialogue that would present the process of modernization in its phenomenological flesh.

What sets the discourse of reflexivity apart, among other things, is the realization that modernity can only gain an access to itself by a series of partial and conflicting descriptions. To put it differently, its ground is constantly slipping from under its feet: "To be modern is to be part of a universe in which, as Marx said, 'all that is solid melts into air'" (Berman, 2010: p. 15). It is due to this inherent dynamic between the original tenets of modernity and its institutional frameworks, within which these tenets can only partially be implemented, that modernization combines self-realization with self-doubt, both revealed in the discourse of reflexivity. And, it is because of this dynamic that modernity cannot but remains 'work in progress.' One can only hope that not only have we always been, but we will always continue to be, modern.

See also: Citizenship, Anthropology of; Democratic Theory; Industrial Societies; Modernity; Modernization; Postindustrial Societies; Social Movements: A Social Psychological Perspective; Sociological Theory; Third World.

Bibliography

Amin, Samir, 2000. Economic globalism and political universalism: conflicting issues? Journal of World Systems Research VI (3), 581–622.

Apter, David, 1965. The Politics of Modernization, The University of Chicago Press, Chicago, IL, pp. xvi, 481.

Bastian, Adolf, 1860. Der Mensch in der Geschichte, vol. 3. (Man in History), Biblio-Verlag, Osnabrück.

Beck, Ulrich, 1998. World Risk Society. Polity Press, Cambridge.

Beck, Giddens, Lash, 1994. Reflexive Modernization. Politics, Tradition and Aesthetics in the Modern Social Order, Polity Press, Cambridge, 225 pp.

Bell, Daniel, 1973. The Coming of Post-industrial Society: A Venture of Social Forecasting, Basic Books, New York, pp. xiii, 507.

Bell, Daniel, Inozemtseva, Vladislav, 2007. Period Disunity: Reflections on the World in the XXI Century.

Berger, Peter, 1967. The Sacred Canopy: Elements of a Sociological Theory of Religion, Doubleday, Garden City.

Berman, M., 2010. All that Is Solid Melts into Air: The Experience of Modernity. Verso, New York.

Castells, M., 1997. The Power of Identity, the Information Age: Economy, Society and Culture, vol. II. Blackwell, Cambridge, MA; Oxford, UK.

Comte, Auguste, 1853. The Course in Positive Philosophy (H. Martineau, Trans. into English).

Cooley, Charles, 1918. Social Process, Charles Scribner's Sons, New York.

Durkheim, Emile, 1893. The Division of Labour in Society.

Eisenstadt, Shmuel, 1978. Revolution and the Transformation of Societies. A Comparative Study of Civilizations. Free Press, New York.

Frazer, James, 1935. Creation and Evolution in Primitive Cosmogenies, and Other Pieces, Curzon Press.

Frobenius, Leo, 1898. Der Ursprung der afrikanischen Kulturen (The origin of African culture).

Fukuymama, Frances, 1992. The End of History and the Last Man.

Gadamer, Hans-Georg, 1960. Truth and Method.

Galbraith, J.K., 1958. The Affluent Society, Houghton Mifflin, Boston, MA.

Gavrov, S., Гавров, С.Н., 2004. Модернизация во Имя Империи. Социокультурные аспекты Модернизационных процессов в России. Эдиториал URSS, Moskva.

Giddens, Anthony, 1990. The Consequences of Modernity. Polity Press, Cambridge.

Graebner, Fritz, 1911. Methode der Ethnologie (Method of Ethnology).

Habermas, Jürgen, 1985a. Der philosophische Diskurs der Moderne. Suhrkamp Verlag, Frankfurt am Main.

Habermas, Jürgen, 1985b. The Philosophical Discourse of Modernity.

Hagen, Everett, 1963. On the Theory of Social Change: How Economic Growth Begins, Illinois Dorsey Press, Homewood, IL.

Handy, Charles, 1993. Understanding Organizations. Oxford University Press, New York.

Harrison, L., Huntington, S. (Eds.), 2001. Culture Matters: How Values Shape Human Progress. Basic Books, New York.

Herskovits, Melville, 1940. Economic Life of Primitive People. A.A. Knopf, New York and London.

Inglehart, Robert, 1977. The Silent Revolution: Changing Values and Political Styles Among Western Publics, Princeton University Press, Princeton, NJ.

Jameson, Fredric, 2002. A Singular Modernity. Essay on the Ontology of the Present. Verso, London and New York.

Kroenbar, Alfred, 1952. The Nature of Culture, University of Chicago Press, Chicago, IL.

Lerner, Daniel, 1958. The Passing of Traditional Society: Modernizing the Middle East, Free Press, Glencoe, IL.

Letourneau, Charles, 1894. The Literary Development in the Various Human Races.

Levy, Marion, 1966. Modernization and the Structure of Societies, Princeton University Press, Princeton, NJ.

Lipset, S.M., 1960. Political Man: The Social Bases of Politics, Doubleday, Garden City.

Lubbock, John, 1870. The Origin of Civilisation and the Primitive Condition of Man, Longmans, Green, London.

Maine, H.J.S., 1861. Ancient Law: Its Connection with the Early History of Society, and Its Relation to Modern Ideas, London.

McLennan, J.F., 1869. The Worship of Animals and Plants.

Moore, W., 1963. Social Change. Prentice Hall, Englewood Cliffs.

Morgan, Lewis, 1877. Ancient Society.

Nisbet, R., 1980. History of the Idea of Progress. Transaction Publishers, New York.

Parsons, T., 2002. парсонс Т. Очерк социальной системы // О социальных Системах. Москва, Издательство "Академический проект", С. 545–687.

Parsons, Talcott, White, Winston, 1961. The Link between Character and Society.

Ratzel, Friedrich, 1898. The History of Mankind.

Rostow, W.W., 1960. The Stages of Economic Growth: A Non-communist Manifesto. Cambridge University Press, London.

Schiller, H.I., 1976. Communication and Cultural Domination. International Arts and Sciences Press, New York.

Smelser, Neil, 1956. Economy and Society: A Study in the Integration of Economic and Social Theory, Free Press, Glencoe, IL.

Spencer, Herbert, 1873. The Study of Sociology.

Sztompka, Piotr, 1993. The Sociology of Social Change.

Thurow, Lester, 1980. The Zero-sum Society: Distribution and the Possibilities for Economic Change, Basic Books, New York.

Tönnies, Ferdinand, 1887. Gemeinschaft und Gesellschaft (Community and Society).

Tylor, E.B., 1870. Researches into the Early History of Mankind and the Development of Civilization, J. Murray, London.

Weber, Max, 1905. The Protestant Ethic and the Spirit of Capitalism.

White, Leslie, 1949. The Science of Culture: A Study of Man and Civilization, Farrar, Straus, New York.

Modes of Production

Donald L Donham, University of California, Davis, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Marxist social theory is often thought of as 'materialist' and 'determinist.' Just how misleading this characterization can be emerged principally in the 1970s in a wide-ranging conversation about modes of production. This article attempts to summarize the lasting contribution of this intervention by reviewing what modes of production are, how they are structured, and why they change over world history.

Concepts in the social sciences often go through intense phases of development followed by periods in which they take on background status. In the latter times, ideas remain alive, but they are no longer being extended or deepened. One example is Karl Marx's notion of modes of production. First developed in the nineteenth century, the idea of modes of production underwent a dramatic transformation in the late 1960s and 1970s as anthropologists, economists, historians, and philosophers were drawn into a wide-ranging conversation. The critical impulse was largely spent by the 1980s, but by that point, aspects of Marx's thought had been developed in strikingly new directions. Some might even think that Marx had been stood on his head.

This article is an attempt to set out some of the enduring legacies of this discussion. It defines forces and relations of production and describes how the first may be thought to determine the second in the creation of distinctive ways or modes of producing social life. Karl Marx's term *Produktionsweise* is usually translated into English as 'mode of production,' but it might be more simply rendered as 'way of producing.' Never precisely defined in Marx's work – yet arguably at the center of it – *Produktionsweise* refers to both a form or structure and a way of producing. One might say that Marx's object in *Capital* was to show how the first – a particular structure of unequal power and wealth – was both the cause and the result of the second, men, and women pursuing their individual ends within capitalist cultural assumptions.

One notable aspect of the capitalist mode of production is, according to Marx, its inherent instability. Crises are necessary, built-in features of this way of producing; they are not somehow accidents caused by factors outside it. As individuals pursue their capitalist projects, there comes a point, finally, when the entire social system can no longer be reproduced. Then, social revolution appears on the horizon, and the mode of production stands to be fundamentally transformed.

Even though Marx referred at times to other modes of production – primitive communist, the ancient, Asiatic, feudal, socialist, and finally modern communist – his only developed work was on capitalism. On what would come after capitalism, he offered only the vaguest of outlines; it would fall to Lenin and others to construct actually existing socialisms. And on what came 'before,' Marx sometimes reflected only the European prejudices of the nineteenth century. For example, he thought that the Asiatic mode of production was inherently static and would waken to change only through the effects of Western imperialism.

The grand evolutionary teleology in Marx's view of history – the circular but progressive spiral from primitive to modern communism – accounts for much of the political appeal of Marx's ideas. Ironically, that appeal was felt mainly in countries on the periphery of capitalist development such as Russia in 1917, China in 1949, and Ethiopia in 1974; in such locations, the notion of stages of history appeared to offer a road map to a better future, a way to skip ahead to modernity. But this aspect of Marx's work also accounts for its ambivalent reception among social scientists who, even if they identified with progressive politics, distrusted apparent teleologies and grappled with refractory evidence.

As Marxism began to provide the legitimating ideology of socialist states in the twentieth century, it tended to stagnate there. In the capitalist West, however, Marx's ideas received critique and development, principally during periods of crisis. Dogmas of various sorts were questioned, and eventually a more flexible version of Marxist theory took its place alongside others in the social sciences.

In the field of anthropology this began after the upheavals of 1968, particularly in France where, under the influence of structuralist philosopher Louis Althusser, scholars such as Claude Meillassoux, Pierre-Philippe Rey, Maurice Godelier, and Emmanuel Terray attempted to extend the notion of modes of production to so-called precapitalist societies. In this article, I attempt to summarize the lasting legacy of this conversation. I do so, not in the structuralist vocabulary of the time, but in reference to what I regard as the clarifying, later work of British analytical philosopher G.A. Cohen.

What, then, is a mode of production? Perhaps the simplest answer is that it is a model of certain basic regularities that hold across and therefore define historical epochs. At the heart of these regularities is the interrelationship between technology in some broad sense and forms of social inequality. According to Marx, the first (*Produktivkräfte*, or productive forces) determines the second (*Produktionsverhältnisse*, or relations of production). These two constitute the base of a mode of production, which, in turn, determines the superstructure.

Productive Forces

Productive forces are anything that can be used in productive interaction with nature. On this reading, productive forces are not simply raw materials or tools; more inclusively, they are human skills, productive knowledge, and even the technical

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12110-5

aspects of human cooperation in the labor process. These factors Marx considered as equally 'material' in that all can be utilized to produce. But most current readers would probably classify objects, knowledge, and the technical aspects of cooperation as, respectively, material, ideal, and social.

The most striking aspect of productive forces has been their tendency to expand over human history. In the most general sense, the anterior cause of this tendency appears to be located in human nature itself. Humans everywhere develop their powers and skills in a continuous dialectic with nature, and when presented with a choice between more or less efficient means of production, humans everywhere (with certain limited exceptions) choose the first.

If productive forces have tended to expand, this fact can be established only at the world level. Some individual societies have stagnated; some have even experienced local declines in the level of productive powers. As will become clear in this article, relations of production can prevent technical development beyond a certain point. For any particular society in stasis, there is no reason to believe that new social relations compatible with the further expansion of forces must develop. In fact, technical knowledge tends to diffuse from place to place, and typically the most 'advanced' societies are not the ones that give rise to the next development. Rather, peripheral groups that are the recipients of the productive knowledge of their more 'advanced' neighbors often leapfrog ahead.

Just what determines the level of development of productive forces should be considered. A society's mass of productive forces is determined by two factors: the first is the average productivity of labor, and the second is population. The product of these two factors yields the level of productive forces. In this formulation, population itself is not a productive force, yet it enters into the determination of the level of productive powers.

Relations of Production

Relations of production are not, as the English phrase suggests, simply the social relationships formed in the actual process of production. Rather, they are the *de facto* power relationships that both underlie and are the result of the division of the fruits of a society's total labor. Household head versus dependents, chief versus subjects, master versus slaves, feudal lord versus peasants, and capitalist versus workers – relations of production are basic asymmetries of power grounded in the organization of material life. In capitalism, relations of production rest upon actual control over productive forces in the process of production, but in other modes of production, as anthropologists and historians eventually showed, this is not necessarily the case.

In relation to the last point, consider the contrast between capitalism and precolonial chiefdoms in Africa. Capitalists put the production process into motion; they or their representatives oversee and supervise in order to ensure that a profit is produced at the end of the work cycle. In many chiefdoms, chiefs did not control the production process itself. Yet, having produced, subjects brought tribute, the fruits of their own labors, to the chief. This contrast illustrates the fact that relations of production, by definition, correspond to the basic structures of power in a society, however that power is constituted – whether by economic, coercive, or religious means (or some mix of these). This makes the reformed notion of modes of production quite different from what is often assumed of Marxism; most particularly, it is not a form of economic reductionism.

Why do productive inequalities occupy a central place in Marx's thought? The answer is that they locate the basic divisions within any society, the lines of potential opposition – of contradiction. Marx saw these as the potential fault lines along which tensions tend to build up are routinely dissipated by small readjustments, and are sometimes violently resolved by radical realignments. These fault lines are structural; they do not necessarily lead to actual struggle and conflict (indeed, the function of the superstructure is precisely to prevent such occurrences). Nevertheless, contradictions always exist as potentialities; they lie just below the surface.

Forces Determine Relations

Ordinarily, if a factor X is said to determine factor Y, then X is taken to cause Y. Given X, Y is uniquely specified. Transferred to Marx's scheme given here, this understanding of determination yields what has been called 'vulgar Marxism.' Such an interpretation can be supported by some of Marx's own, more incautious summary statements: "The handmill gives you society with the feudal lord; the steam mill society with the industrial capitalist." Empirically, it is clear that the same technology can exist in different modes of production.

A great deal of effort has been expended by commentators to explicate Marx's notion of determination – from examination of dialectics to, in Engels' phrase, determination only 'in the last instance.' Probably the best way to clarify Marx's claims is to note that they are forms of functional, not causal, explanation. Immediately, a distinction needs to be drawn between functional explanations and the theoretical school known as functionalism in sociology and anthropology: Marx used functionalist explanations, but he was not a 'functionalist' in the second sense.

Functional explanations have a distinct logical form in which the effects of a trait enter into the explanation of the presence of that trait: for example, when Y's presence is partly explained by some positive consequence Y has for X. In such a case, X can be said to determine or 'select' Y because of Y's effect on X.

Marx believed that the level of productive forces selects relations of production. For any given level of development of forces, there is only a limited set of relations possible – namely, those that are compatible with and in fact promote the further development of forces. Here, X hardly causes Y in a one-way pattern of determination. Rather, the reverse is closer to the truth: it is Y's effect on X that is paramount. Particular relations cause expansion in the forces. Herein lies the rub, however. As forces expand, they reach a level that is no longer compatible with the existing relations of production. At that point, according to Marx, revolution becomes a possibility. In this way, continuous incremental change (in forces) is mapped onto occasionally discontinuous revolutionary change (in relations) in the base of modes of production.

Bases Determine Superstructures

One of the important contributions of G.A. Cohen was a clarification of the notion of superstructure. If Marx claims that the base or, more precisely, relations of production determine the superstructure, then a definition of relations and superstructure must be found in which the two are clearly distinguished. The difficulty is that power as we find it in any social formation is inevitably encased in particular cultural and legal concepts (i.e., superstructural notions). It was for this reason that Cohen defined relations of production as the *de facto* distribution of power. Every culturally encoded right, he argued, can be matched with a *de facto* power. And it is *de facto* power that determines culturally specific notions, rather than the reverse.

One important effect of this definitional move is to place the long-term dynamic of a mode of production outside of culture itself. Eventually, culture corresponds to *de facto* power rather than the reverse. But in the short term, it places culture at the very center of the analysis of modes of production, for it is precisely the superstructure that defines and stabilizes relations of production. Once again, the logic of functional explanation turns the ordinary expectations about Marxism upside-down. If something is described as having a base and a superstructure, we think of the base as a foundation. It supports the superstructure; without the foundation, the superstructure would fall to the ground. Yet, the logic of functional explanation requires the opposite: it is the superstructure that defines the base.

To adapt a visual image from Cohen (1978), consider the following: Four pieces of wood of the same length are driven a little way into the ground. They stand, but they are easily blown over by a strong wind. A top is attached to the four struts such that the whole becomes a table. The struts, now legs, stand solidly upright. They do not wobble in the wind. Of the tabletop, one can say that (1) it is supported by the legs, and (2) its very existence is required for the legs to be legs. Without the top, the legs would not be legs – only sticks of wood of such-and-such a length. Here, we have a building whose base and superstructure relate in the proper way.

The conclusion to this line of thought (heretical to some Marxists) is that *Capital* is an analysis of the capitalist superstructure – the set of culturally specific and legally encoded meanings and practices that tend to stabilize capitalist relations of production. In a comparative sense, this conclusion is crucial since it sets capitalism on the same analytical plane as any other mode of production. On this view, the analysis of modes of production becomes a matter of locating effective power difference over material production as they are reflected in differential control over the total social product and analyzing the meanings and practices that tend to reproduce such powers – what Antonio Gramsci called hegemonic culture. It goes without saying that it is always misleading to import the cultural logic of one mode of production (such as capitalism) into the understanding of others.

Functional Alternatives

If this reading of Marx's notion of determination solves many conceptual problems, it also creates others in the application of Marx's ideas. Carl Hempel and others have called attention to the difficulties of actually carrying out functional explanations. Not the least of these stem from the problem of functional alternatives. That is, there is often more than one Y (Y_1, Y_2, Y_3, …) that can fulfill any given function. That being so, the particular presence of Y_1, say, cannot be explained simply by referring to its function, for Y_2 or Y_3 could have served the same need.

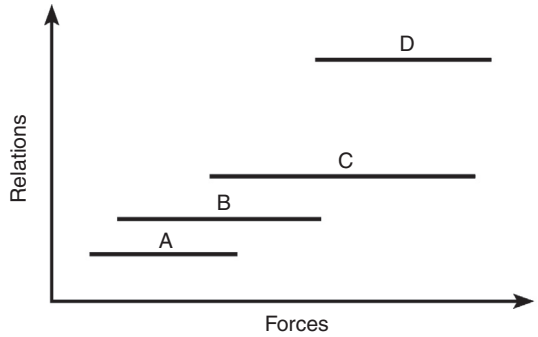

Figure 1 A possible set of forces and relations of production when functional alternatives exist.

The critical question becomes "What is the size of the set of functional alternatives?" Implicitly, Marx appears to have believed that alternatives are quite limited – typically to the set of one. Consider such a limiting case. For each level of development of productive forces, there is only one set of relations possible. History therefore proceeds in a series of discrete steps from A to B to C. Both the past and the future are specified – socialism is inevitable because it is supposedly the only economic system consistent with highly developed forces.

A different story emerges if the number of functional alternatives is even moderately few. If several different sets of relations of production are compatible with any given level of forces, then the grand design of world history becomes considerably more complex. One possible trajectory in Figure 1, for example, would be A to C to B to D. In this reading, the future looks considerably more open. There may indeed be functional alternatives to socialism – if indeed socialism as we have known it is compatible with developed forces.

Conclusion

The notion of a mode of production is a deliberate simplification. It provides a model of certain central relationships that define epochal forms. As such, the analysis of modes of production cannot substitute for the examination of particular cases in specific spatial and temporal contexts. In this sense, mode-of-production analyses do not constitute historical explanations.

In a larger sense, however, the concept of ways of producing furnishes the necessary scaffolding without which local historical explanation would be impossible. Comparison and contrast, whether explicitly or implicitly carried out, depend upon an overarching framework. This is what Marx's notion of modes of production provides: a way of understanding society and history in the broadest terms.

The present is once again a moment of deep crisis in world capitalism. Accordingly, it would not be surprising if the social sciences produced another efflorescence of Marxist-inspired ideas, but if so, it will unlikely do so in reference to a comparative analysis of modes of production. After the disintegration of the Soviet Union and the elimination of socialism as a world model, nothing remains but capitalism. In 1968, it was possible at least to imagine large parts of the earth as 'noncapitalist,' but now there is, for the first time in human history, one truly globalized world capitalist system. Deepening the analysis of the present will require taking into account what was learned in the 1970s: namely, that capitalism is no more 'economic' than other modes of production and no less dependent upon a series of culturally peculiar ideas. Once the analysis of modes of production is truly leveled in this respect, we will have a more powerful approach to capitalism itself, one that does not naturalize capitalist culture.

See also: Marx, Karl (1818–83).

Bibliography

Althusser, L., Balibar, E., 1970. Reading Capital (B. Brewster, Trans.). Pantheon, New York.

Anderson, P., 1974a. Lineages of the Absolutist State. New Left Books, London.

Anderson, P., 1974b. Passages from Antiquity to Feudalism. New Left Books, London.

Aston, T.H., Philpin, C.H.E. (Eds.), 1985. The Brenner Debate: Agrarian Class Structure and Economic Development in Pre-industrial Europe. Cambridge University Press, Cambridge, UK.

Cohen, G.A., 1978. Karl Marx's Theory of History: A Defence. Clarendon Press, Oxford, UK.

Donham, D.L., 1999. History, Power, Ideology: Central Issues in Marxism and Anthropology, second ed. University of California Press, Berkeley, CA.

Kahn, J.S., Llobera, J.R. (Eds.), 1981. The Anthropology of Pre-capitalist Societies. Macmillan, London.

Moore, S., 1980. Marx on the Choice between Socialism and Communism. Harvard University Press, Cambridge, MA.

Modularity and Organizations

Carliss Y Baldwin, Harvard Business School, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Modularity describes the degree to which a complex system can be broken apart into subunits (modules) which can be recombined in various ways. Modularity is important for organizations and the economy because the boundaries of organizational units and corporations are likely to match the boundaries of underlying technological modules. (This correspondence is called 'mirroring'.) In this article, I explain the concept of modularity and describe how systems can be modularized. I then explain why mirroring is likely to be a commonly observed organizational pattern, review the empirical evidence, and conclude with open research questions.

Introduction

Modularity is a property of complex systems, both natural and artificial. Loosely speaking, the term describes the degree to which a given system can be broken apart into subunits (modules) which can be assembled and recombined in various ways (Simon, 1962 [1981], 1981; Schilling, 2000 [2003]). There is no universally accepted measure of modularity, however, highly modular systems are made up of many, small, loosely coupled components, while a nonmodular system is a single large structure in which 'everything depends on everything else.'

Modularization is a structured learning and design process aimed at making a complex system more modular (Baldwin and Clark, 2000; MacDuffie, forthcoming). The process was first invented at IBM in the mid-1960s to address looming problems in the design, production, and marketing of early computers. IBM System/360 was the first computer to be designed as a modular system with fully interoperable hardware and software across a wide range of processors. The process of modularization was subsequently refined through applications to software, semiconductor chips, mini and personal computers, and digital communication networks.

Modularity is important for organizations and the economy because the boundaries of organizational units and corporations are likely to match the boundaries of underlying technological and task modules. (This correspondence is called 'mirroring' or the 'mirroring hypothesis'.) As a result, changes in the modular structure of products and processes have the potential to change the structure of industries and disrupt established firms.

In this article, I first describe the development of the concept of modularity, then a series of attempts at modularization and lessons learned from both successes and failures. I will then describe the origins of the mirroring hypothesis, explain why mirroring is a commonly observed pattern in industrial organization, and review the empirical evidence on mirroring. I conclude with a list of open research questions.

The Concept of Modularity

In a 1962 essay on 'The Architecture of Complexity,' the polymath Herbert Simon proposed that complex systems would be organized as hierarchically as a set of inter-related subsystems "until we reach some lowest level of elementary subsystem" (Simon, 1962 [1981, p. 196]). Using the famous fable of two watchmakers, he went on to argue that the organizational structure most favorable to evolution was a 'near-decomposable' hierarchy in which interactions among elements within subsystems were strong, while interactions between subsystems were weak. Two years later, the architect and design theorist Christopher Alexander made a similar observation: "No complex adaptive system will succeed in adapting in a reasonable amount of time unless the adaptation can proceed subsystem by subsystem, each subsystem relatively independent of each other" (Alexander, 1964, p. 41). Both men credited the concept of 'selection by components' to the cybernetics pioneer W. Ross Ashby. The presence of units whose elements are highly connected within the unit and only weakly connected to other units is a defining property of modular systems. The highly connected units are called modules (Baldwin and Clark, 2000).

The concept of modularity was further advanced by the computer scientist David Parnas. Writing about software, he defined a 'module' as a work assignment to an individual or small team and 'modularization' as a set of work assignments for a complex software system. He then proposed that "every module … [should be] characterized by its knowledge of a design decision which it hides from all others. Its interface [should be] chosen to reveal as little as possible about its inner workings" (Parnas, 1972 [2001, p. 151]). Behind this concept of 'information hiding' lies a strategy for minimizing coordination costs across a large group of designers. Each team needs to know a lot – possibly everything – about its own module, but comparatively little about other modules. Thus information about modular systems can be parsed into 'visible' and 'hidden' information. Insofar as changes to the system affected only hidden information, they would be relatively easy to implement, hence the system would be more adaptable. Thus information hiding makes modular systems evolvable.

Karl Ulrich (1995) defined modularity in relation to product architecture. A product architecture is (1) a list of functional requirements; (2) a mapping from functional elements to physical components; and (3) the specification of interfaces among the physical components. A modular

architecture is characterized by a one-to-one mapping of functions to components (i.e., a distinct physical component performs each function) and 'decoupled' interfaces. In contrast to other definitions of modularity, Ulrich's definition rests on a prespecified and stable list of functions, thus it is most often applied in contexts where the conceptual frame of the artifact, i.e., its functions and interactions with users and the environment, are well understood.

Schilling (2000[2003]) emphasized the reconfigurability of modular systems. Modular systems take in heterogeneous inputs and arrive at heterogeneous outputs, thereby satisfying consumer demand for high variety: "[if] all customers want the same thing, there is little to be gained through offering a modular system" (2000[2003, p. 179]).

Baldwin and Clark (2000) defined modular systems as those displaying high levels of decomposability and information hiding, which in turn lead to three economic benefits. First, modularity reduces cognitive complexity, making the overall system easier to understand. Second, different modules of the system can be worked on concurrently and in parallel, reducing time to completion. Third, through information hiding, modular systems accommodate uncertainty: "if new knowledge later yields a better solution for one of the hidden module designs, then it will be relatively simple to incorporate the new solution without changing much in the rest of the system. … [Thus modularity] creates *options* within the design" (p. 91). The authors go on to develop a mathematical model, based on options theory, to value modular systems.

The Process of Modularization

Since the mid-1960s, processes of modularization have been observed in many settings. Some efforts were successful, while others were failures.

In the 1950s, IBM instituted a Standard Modular System for manufacturing transistor circuits. Circuits were constructed on standard-sized cards that were pluggable into a standard box. The system allowed mass production techniques to be applied to the manufacture of complex electronic circuits, thus reducing the cost of the most basic building block of a computer.

Following this success in manufacturing, an IBM task group attacked the twin problems of rationalizing IBM's diverse product line and reducing the cost to customers of upgrading hardware. (At that time, IBM sold six incompatible processors, thus most hardware upgrades entailed rewriting virtually all software applications as well as retraining field personnel.) The result was System/360, which has been called the first modular computer system (Ferguson and Morris, 1993). All System/360 processors were 'binary compatible,' i.e., they could process the same software programs, and the peripheral equipment could be mixed and matched with processors to suit customer needs. Introduced in 1967, System/360 was both a technical accomplishment and a huge economic success (Baldwin and Clark, 2000).

In contrast, attempts to modularize System/360s system software failed utterly. Managers initially divided the system-to-be-built into 40 subprojects and allocated more than 1000 people to the task of delivering the software. But unsuspected conflicts emerged between the code modules and the work fell further and further behind. The experience led Frederick Brooks, a key architect of System/360, to propose 'Brooks Law': "adding people to a late project simply makes it later." He also asserted that system software could not developed in a modular fashion, but required highly coordinated teams controlled by a 'master architect' (Brooks, 1975[1995]).

But Brooks did not make allowances for advances in knowledge about how to build system software. Even as his book was going to print, Ritchie and Thompson at Bell Labs revealed Unix, a new operating system that was modular in both its structure and the way it dynamically managed computer resources in real time (Richie and Thompson, 1974). Unix went on to become the foundational design for many modern operating systems. Together with its descendants, it offered a clear counterexample to Brooks' theories of software development (Raymond, 1999). Indeed, Brooks (1975[1995], p. 271) later admitted that he had failed to appreciate Parnas' 'information hiding' as a basic principle of modular software design.

Subsequently, Mead and Conway (1980) successfully modularized the design of semiconductor chips. However, attempts to modularize the design and production of automobiles largely failed for two reasons: (1) elements related to the same function, e.g., braking, are not collocated, but spread throughout the vehicle; and (2) companies did not invest *ex ante* in learning about critical interdependencies or alternate decompositions (MacDuffie, forthcoming).

Drawing on engineering research documents, especially Mead and Conway (1980), Baldwin and Clark (2000) provide general description of the process of modularization. A nonmodular design, by definition, contains many dependencies between design elements, so that a change in one element will require changes in other elements, which in turn will propagate throughout the entire system. The process of modularization entails the following: first, identifying each dependency and understanding various ways of solving it; then, eliminating the dependency by adopting a 'design rule.' Through systematic repetition of this process, unwanted dependencies across different 'blocks' of the system can be eliminated, while fruitful dependencies within blocks are retained. Eventually, the system arrives at the characteristic modular structure: it will be highly connected within subsystems and independent (except for design rules) across subsystems.

The process can be illustrated with the aid of a design structure matrix (DSM) (Eppinger, 1991). In a DSM, decisions about the design of a complex system are arrayed along the rows and columns of a square matrix. If changing one design decision changes another decision, then we place an 'x' in the column of the first decision and the row of the second. Figure 1 shows two hypothetical DSMs for a laptop computer consisting of four components: the drive system, the main board, the LCD screen, and packaging. Figure 1(a) depicts the design before modularization. There are many out-of-block x's both above and below the main diagonal. Zooming in on the matrix, we see that a decision about the main board affects a decision about the LCD screen, as indicated by the circled 'x' below the main diagonal. Changing the LCD screen decision in turn affects another

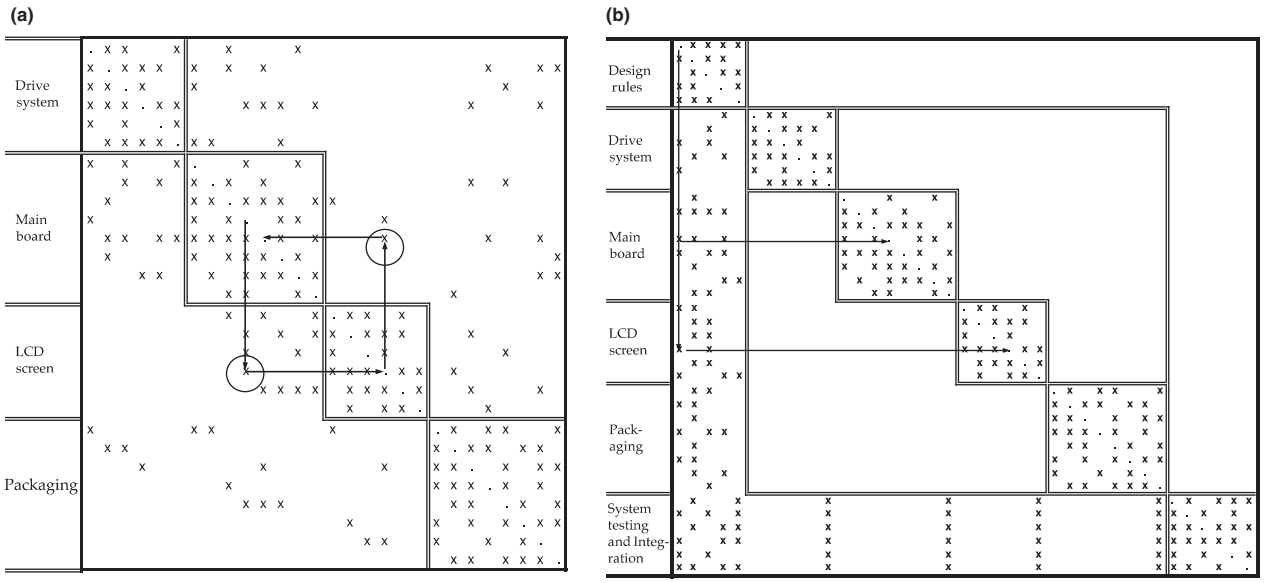

Figure 1 The process of modularization. Reproduced from Baldwin, Carliss Y., Clark, Kim B., 2000. Design Rules. The Power of Modularity, vol. 1. MIT Press, Cambridge, MA, pp. 50, 74.

decision on the main board. The flow of decisions thus forms a loop (shown by arrows) between the main board and the LCD screen. Hence the design process must cycle between the two components, adjusting decisions for each until the components are fully compatible.

The laptop design (and corresponding design process) can be modularized in the following way. First, designers must identify the technical cause of each unwanted dependency: why exactly does decision (1) concerning the main board affect decision (2) on the screen? Then they must resolve the potential conflict by formulating a 'design rule' or 'standard' that eliminates the dependency. Then, as long as each group of component designers stays within the constraints of the design rules, they are free to make decisions independently of the other groups.

The technique of removing unwanted dependencies via design rules is a generalizable procedure. When this strategy is consistently pursued, dependencies across components become sparser, and the system as a whole becomes 'near decomposable,' that is, tightly connected within components and loosely connected across components. Figure 1(b) shows a characteristic DSM after modularization. In this idealized picture, there are no out-of-block dependencies, indicating that work on each component can take place independently of the others. Hierarchical dependencies on design rules, indicated by new arrows, have replaced cyclical dependencies between component decisions. There are also new sets of activities, indicated by new blocks in the DSM. Design rules must be established and communicated at the beginning of the process, and system testing and integration are needed at the end of the process to bring the modules together, test the system, and work around unforeseen incompatibilities.

The following lessons can be learned from observing modularization as a process in varying contexts and over time. First, the benefits and costs of modularity vary across artifacts, technologies, and time. Whitney (2004) has argued that benefits tend to be high and costs low for digital and electronic artifacts, while electromechanical artifacts that manage large amounts of power or heat are intrinsically more difficult to split into independent subsystems. The case of Unix also shows that successful modularizations depend critically on prior knowledge of specific interdependencies and how to address them via hierarchical design rules.

Consumer preferences matter too. Modularity is most valuable when consumers value the option to mix and match components to adapt a system to their own needs or to accommodate later additions and upgrades. It is not worth modularizing if all consumers desire the same thing, indeed, it then makes sense to assemble separate components and sell them as an indivisible unit.

Finally, it is worth noting that technological progress does not always lead in the direction of higher levels of modularity. Adding interdependencies between components can lead to higher levels of performance and efficiency. For example, in the case of bicycle drive trains, Fixson and Park (2008) document a significant improvement in the precision and ease of use for 'index-shifting' systems in which all parts of the drive train are co-specialized. Introduced by Shimano in 1984, the nonmodular index-shifting systems quickly replaced older modular systems.

Modularity and Organizations: The Mirroring Hypothesis

Modularity and the process of modularization are important topics for the social sciences because of the impact of modularity on the division of knowledge, the organization of work,

the boundaries of firms, and ultimately industry structure. There is strong overlap between Parnas' definition of a module as a 'work assignment' and fundamental questions of organization and economic theory. von Hippel (1990), building on Simon and Alexander, proposed that information processing costs in organizations would be minimized if tasks were partitioned at points of low problem-solving interdependence. In other words, in assigning tasks to individuals or groups, modules should not be split, but kept together. Henderson and Clark (1990) turned this normative recommendation into a prediction. They defined organizational structure as being made up of communication channels, information filters, and problem-solving strategies and then asserted "organizations are boundedly rational and, hence, their knowledge and information-processing structure come to mirror the internal structure of the product they are designing" (p. 27). Thus the theory of correspondence between the modular structure of artifacts and the structure of organizations has come to be called 'the mirroring hypothesis.'

von Hippel and Henderson and Clark looked only at organizational structure within firms. Langlois and Robertson (1992[2003]) were the first to see that modularity could affect the division of labor across firms, hence the structure of an industry. Modular systems, they said, provide consumer choice in the selection of components, and thus can give rise to decentralized networks of producers whose products worked together based on common standards of compatibility. "By allowing specialist producers … to concentrate … on particular components, a modular system thus enlists the division of labor in the service of innovation" (1992[2003, p. 84]). They supported their arguments by recounting the impact of modularizations on industry structure in the stereo records and personal computer industries.

Sanchez and Mahoney (1996[2003]) unified and extended these arguments. Echoing Henderson and Clark, they suggested that "although organizations ostensibly design products, it can also be argued that *products design organizations*, because the coordination tasks implicit in specific product designs largely determine the feasible organizational designs for developing and producing those products" (1996[2003, p. 363], emphasis in original). They defined a modular product architecture as one that used standardized interfaces to achieve flexibility in configuration. Such architectures would allow a firm to "source components through loosely coupled networks," leading to "increasingly dynamic product markets" and "previously unattained levels of product variety and change" (1996[2003, p. 375]). In common with other writers in the mid-1990s, Sanchez and Mahoney were strong advocates of the modular approach to product design and saw practically no limits to its applicability. Like the Internet and the dot.com boom, modularity was part and parcel of the emerging new economy. Its failures and inherent risks (described above) were not then widely known or appreciated.

Why is mirroring a plausible hypothesis? In Baldwin's (2008) article on "Where do transactions come from?" I explicitly consider the impact of modularity on transaction costs and firm boundaries. Module boundaries are points where 'mundane transaction costs' are low; thus, other things being equal, they are attractive points to locate transactions between autonomous agents. It is possible to place transactions within modules, but potential agency and monitoring costs will be high, hence such placements generally will be disadvantageous. Work within modules is best conducted in a 'transaction-free zone.'

It is also possible to group several modules within the boundaries of a single corporation. But in this case, if the corporation favors its internal suppliers, it will lose the option value of seeking the best or lowest-cost components (Jacobides, 2005). Conversely, if the upstream business units are prohibited from selling to external customers, they may lose out on scale economies and fall behind in terms of technology and product performance.

Within the boundaries of a single firm, strict mirroring can be hard to achieve because of geographic constraints and organizational rigidities, especially given a changing product architecture. At the same time, modern information and communication technologies make it easier for geographically dispersed teams to coordinate work on a single module. Repeated interactions can lead to relational contracts and cooperation even in the absence of trust-inducing face-to-face interaction. Thus it is best to view mirroring as a cost-minimizing, ideal form of relationship between product architecture and organizational structure.

Two types of deviation from mirroring are possible and have different consequences (Colfer and Baldwin, 2010). First, a modular organization (for example, a geographically dispersed group of individuals or a set of autonomous firms) may undertake to design and produce a nonmodular product. Lack of mirroring will then result in increased coordination and/or transaction costs and failures arise because of insufficient communication or misaligned incentives. Second, a nonmodular organization (for example, a collocated team or a single firm) may undertake to design and produce a modular system. In this case, the organization may fail to perceive or ignore opportunities to reduce cost or enhance performance by adopting modular solutions from outside the organization.

An especially insidious version the second failure can arise when a nonmodular organization fails to perceive the value of modularity and makes the system less modular than it needs to be from a purely technical standpoint. For a nonmodular organization, maintaining module boundaries and interfaces is costly and may be viewed as unnecessary in the short run. Conflicts can be settled in an ad hoc fashion via direct communication and coordination between the affected parties. The result is a system with high levels of interdependence between elements: one that *ex post* looks more like **Figure 1(a)** above than **Figure 1(b)**.

If the options associated with modularity (variety, configurability, upgradability) are highly valued by consumers, the less modular systems will turn out to be uncompetitive with modular systems and disappear from the market. In such cases, the industry will come to be organized as a set of horizontal, complementary layers with different firms competing in and dominating each layer (Grove, 1996; Baldwin and Clark, 2000).

MacCormack et al. (2012) show that software code developed within firms is significantly less modular than

functionally equivalent software developed by geographically dispersed, autonomous individuals working in open source communities. Thus the mirroring hypothesis holds in this study: software developed by less modular organizations is less modular than dictated by purely technical considerations, and vice versa. However, it cannot be inferred from this study that the nonmodular systems are suboptimally structured. Although nonmodular products may suffer from lack of variety, configurability, and upgradability, there may be compensating benefits to firms in terms of lower cost, higher performance, and better protection of intellectual property (Henkel et al., 2012). Nevertheless, because mirroring between product and organizational structure is a low-cost alternative, especially in the short run, nonmodular organizations may create nonmodular products by default. If the benefits of modularity are deemed to be competitively significant, then managers must intervene by either imposing (and enforcing) modular design rules or by breaking up the nonmodular organization in ways that enforce information hiding between subgroups.

If deviations from strict mirroring are possible, then the frequency with which mirroring is observed and the circumstances of its presence or absence become important empirical questions. Colfer and Baldwin (2010) surveyed 102 studies of development projects in which both the modular structure of the system and organizational ties (geographic collocation and common employment) were disclosed in sufficient detail to infer whether mirroring was present or not. The sample included 22 within-firm studies, 62 across-firm studies, and 18 studies of open source communities. They found evidence of mirroring in just over two-thirds (69%) of the studies, although in a few cases the evidence was mixed. However, evidence of mirroring was noticeably weaker in the open source subgroup, where 68% (11 out of 18) studies were not supportive of the hypothesis.

Of the 32 studies that were not supportive, four cases involved collocated contributors employed by a single firm designing modular systems made up of independent components. In all of these cases, modularity was an explicit goal of the design effort, and a great deal of attention was paid to design rules. (One of these groups was the task group that designed the architecture of System/360, discussed earlier.)

In contrast, 28 of the 32 cases involved contributors who were not collocated and/or employed by the same firm collaborating to develop highly interdependent, nonmodular systems. Consistent with the idea that mirroring is a low-cost way to achieve the coordination and cooperation needed to complete interdependent tasks, all of these cases involved high levels of investment in tools that made it easy to share information and resolve conflicts rapidly as well as social mechanisms to build trust and encourage cooperation among the participants.

Conclusion: Open Research Questions

At this point, a number of open research questions remain, and I will close this article by mentioning a few of them. First, we need to know which types of sociotechnical systems are amenable to modularization and why. In some contexts, modularization appears irresistible; in others, there is no discernible trend; while in other cases, nonmodular systems have taken over the market. Successful modularizations appear to be enabled by (1) physical properties of the system (Whitney, 2004) and (2) knowledge of the relevant interdependencies (Baldwin and Clark, 2000; MacDuffie, forthcoming). Consumer preferences also play a role (Schilling, 2000[2003]), and firms may choose different degrees of modularity as a matter of competitive strategy (Fixson and Park, 2008; Henkel et al., 2012). But up to now, decisions on where and when to modularize have been ad hoc. A comprehensive theory that identifies the relevant variables and contingencies would go a long way toward preventing costly mistakes (MacDuffie, forthcoming).

Such a theory would also allow us to better understand and predict industry structure. Nonmodular systems often give rise to vertically integrated industries, in which a few firms strive to coordinate a pipeline of products that compete in the end-product market. Highly modular systems in contrast give rise to 'horizontally layered' industries in which many firms make complementary modules that jointly rely on open standards (Grove, 1996; Jacobides and Tae, 2012). In such industries, firms must actively manage their relationships with autonomous complementors, both large and small (Gawer and Henderson, 2007). They must decide which modules to keep closed as a source of profits, and which to open as a means of increasing system value (Henkel et al., 2012). Further empirical and theoretical work is needed to understand the competitive forces and dynamics at work in these highly distributed technical 'ecosystems' where competition between complementors is the rule.

Finally, in both modular and nonmodular systems, we need to better understand the reasons for and against mirroring. As argued earlier, mirroring should be regarded as an economic and strategic choice. It is sometimes advantageous for a tight-knit team to create and maintain a highly modular system, but this will not happen by default. Tools for visualizing and measuring modularity and enforcing modular boundaries are thus needed to make this a strategic option for firms. Conversely, geographically dispersed teams and/or autonomous individuals often need to collaborate to create nonmodular systems. In addition to opting for highly modular systems, open source software communities have also invested in tools and social practices, based on openness and technical transparency, that promote collaboration *within* modules (Raymond, 1999). Outside of software, however, this work is just beginning (cf. Dawson et al., 2012).

Although many complex systems exhibit modularity, it is a relatively new focus of research in the social sciences. The process of modularization was invented in the mid-1960s to solve the twin problems of complexity and costly upgrades caused by IBM's numerous incompatible product lines. Thereafter, the process has been applied many times, sometimes successfully but also with some notable failures. Through studying these experiments, scholars have developed a powerful and coherent economic theory of modular systems, which explains how modularity creates value and how and why modular systems evolve. The theory also

addresses the relationship between technological modularity and organizational structure through the mirroring hypothesis, which provides testable predictions about the location and design of transactions and the impact of successful modularizations on industry structure. However, more work is needed to develop a discriminating theory that can predict when and where modularizations are feasible and worthwhile to learn how firms compete in horizontally layered industries and to develop tools, organizational mechanisms, and social practices that will permit firms to relax the constraints of mirroring when it is advantageous to do so.

See also: Boundaries and New Organization Forms; Business Platforms; Closed and Open Systems: Organizational; Hierarchies and Markets; Industrial Dynamics: Patenting Perspectives; Innovation and Technological Change, Economics of; Organization: Overview; Rationality in Society; Technology and Organization.

Bibliography

Alexander, Christopher, 1964. Notes on the Synthesis of Form. Harvard University Press, Cambridge, MA.

Baldwin, Carliss Y., 2008. Where do transactions come from? Modularity, transactions and the boundaries of firms. Industrial and Corporate Change 17 (1), 155–195.

Baldwin, Carliss Y., Clark, Kim B., 2000. Design Rules. In: The Power of Modularity, vol. 1. MIT Press, Cambridge, MA.

Brooks, Frederick P., 1975. The Mythical Man-Month: Essays on Software Engineering, 20th Anniversary Edition 1995. Addison-Wesley, Reading, MA.

Colfer, Lyra J., Baldwin, Carliss Y., 2010. The Mirroring Hypothesis: Theory, Evidence and Exceptions. Harvard Business School Working Paper No. 10-058, January 2010 (revised, June 2010). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1539592.

Dawson, Benjamin, Fixson, Sebastian K., Whitney, Daniel, 2012. Orchestrating Coordination: Reducing Rework in Complex Product Development. Technology, Operations & Information Management Working Papers. Paper 4. http://digitalknowledge.babson.edu/toimwp/4 (accessed 12.10.12.).

Eppinger, Steven D., 1991. Model-based approaches to managing concurrent engineering. Journal of Engineering Design 2, 283–290.

Ferguson, Charles H., Morris, Charles R., 1993. Computer Wars: How the West Can Win in a Post-IBM World. Times Books, New York, NY.

Fixson, Sebastian K., Park, Jin-Kyu, 2008. The power of integrality: linkages between product architecture, innovation and industry structure. Research Policy 37 (8), 1296–1316.

Gawer, Annabelle, Henderson, Rebecca, 2007. Platform owner entry and innovation in complementary markets: evidence from Intel. Journal of Economics and Management Strategy 16 (1), 1–34.

Grove, Andrew S., 1996. Only the Paranoid Survive. Doubleday, New York.

Henderson, Rebecca M., Clark, Kim B., 1990. Generational innovation: the reconfiguration of existing systems and the failure of established firms. Administrative Science Quarterly 35, 9–30.

Henkel, Joachim, Baldwin, Carliss Y., Shih Willy, 2012. IP Modularity: Profiting from Innovation by Aligning Product Architecture with Intellectual Property. Harvard Business School Working Paper, No. 13–012, August 2012, revised December 2012.

Jacobides, Michael G., 2005. Industry change through vertical dis-integration: how and why markets emerged in mortgage banking. Academy of Management Journal 48 (3), 465–498.

Jacobides, Michael G., Tae, Jennifer C., 2012. Beyond Five Forces: Kingpins, Bottlenecks and Value Dynamics along a Sector. Paper Presented at the 2102 Academy of Management, Boston, MA.

Langlois, Richard N., Robertson, Paul L., 1992. Networks and innovation in a modular system: lessons from the microcomputer and stereo component industries. Research Policy 21(4), 297–313. Reprinted in: Raghu, G., Kumaraswamy, A., Langlois, R.N. (Eds.), 2003. Managing in the Modular Age: Architectures, Networks, and Organizations. Blackwell, Oxford/Malden, MA.

MacCormack, Alan, Baldwin, Carliss, Rusnak, John, 2012. Exploring the duality between product and organizational architectures: a test of the mirroring hypothesis. Research Policy 41 (8), 1309–1324.

MacDuffie, John Paul, forthcoming. Modularity-as-property, modularization-as-process, and 'modularity'-as-frame: lessons from product architecture initiatives in the global automotive industry. Global Strategy Journal. 3(1), 8–40

Mead, Carver, Conway, Lynn, 1980. Introduction to VLSI Systems. Addison-Wesley, Reading, MA.

Parnas, David L., 1972. On the criteria to be used in decomposing systems into modules. Communications of the ACM 15, 1053–1058. Reprinted in idem: Hoffman, D.M., Weiss, D.M. (Eds.), 2001. Software Fundamentals: Collected Papers of David Parnas. Addison-Wesley, Boston, MA.

Raymond, Eric S., 1999. The Cathedral and the Bazaar. O'Reilly & Associates, Inc., Sebastopol, CA.

Richie, Dennis M., Thompson, Kenneth, 1974. The Unix time sharing system. Communications of the ACM 17 (7), 365–375.

Sanchez, Ronald A., Mahoney, Joseph T, 1996. Modularity, flexibility and knowledge management in product and organizational design. Strategic Management Journal 17, 63–76. Reprinted in: Raghu, G., Kumaraswamy, A., Langlois, R.N. (Eds.), 2003. Managing in the Modular Age: Architectures, Networks, and Organizations. Blackwell, Oxford/Malden, MA.

Schilling, Melissa A., 2000. Toward a general systems theory and its application to interfirm product modularity. Academy of Management Review 25(2), 312–334. Reprinted in: Raghu, G., Kumaraswamy, A., Langlois, R.N. (Eds.), 2003. Managing in the Modular Age: Architectures, Networks, and Organizations. Blackwell, Oxford/Malden, MA.

Simon, Herbert A., 1962. The architecture of complexity. Proceedings of the American Philosophical Society 106, 467–482. Reprinted in idem: The Sciences of the Artificial, second ed. MIT Press, Cambridge, MA, pp. 193–229 (1981).

Ulrich, Karl, 1995. The role of product architecture in the manufacturing firm. Research Policy 24, 419–440.

von Hippel, Eric, 1990. Task partitioning: an innovation process variable. Research Policy 19, 407–418.

Whitney, Daniel E., 2004. Physical Limits to Modularity. http://esd.mit.edu/symposium/pdfs/papers/whitney.pdf (accessed 21.07.05.).

Monasticism: West and East

William JF Keenan, Nottingham Trent University, Nottingham, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 9973–9976, © 2001, Elsevier Ltd.

Abstract

Monasticism, defined inclusively here to incorporate all modes of religious life from the hermit to the contemplative and active religious orders, can be found throughout all the major religious traditions. The monastic vocation, open to male and female alike, has been particularly prominent in Christianity and Buddhism and monastic organizations and institutions have had a great influence on civilizations and societies where these faiths have become powerful. Monastic life entails regular prayer, regimes of self-discipline, and works of charity. Occidental monasticism tends to inner-worldly activism, whereas oriental monasticism emphasises other-worldly mysticism. Religious virtuosi typically follow a Rule established by their charismatic founder(s). Each religious order and religious congregation has its own distinctive spiritual 'way' and apostolate (teaching, nursing, contemplative prayer). Different ages are associated with different 'styles' of religious life from the desert solitaries, to the great monastic centres of the Middle Ages, to the international missionary congregations of the modern era. Approaching the third millennium of monasticism, religious life west and east is at a crossroads as traditional forms confront novel experiments in the consecrated life.

Introduction

'Monasticism,' derived from the Greek word *monos* meaning 'alone,' is defined inclusively here to incorporate all forms of vocational commitment to religious life. It may embrace the reclusiveness of the solitary desert anchorite, the coenobitical 'flight' of the self-governing rural ashram, monastery, or convent, or the inner-city apostolates of households of professional catechists and ordained welfare workers. Throughout the histories of the major world faiths, countless religious devotees have 'set apart' their lives 'in' but not 'of ' the world within religious 'quasisects' adhering to particular theological traditions of 'training the soul.' From the Essenes of ancient Judaism, through the myriad Christian religious orders and religious congregations, to Sufi and Sanusi Islamic brotherhoods, or Buddhist and Hindu monks, monasticism provides contexts in which to study the religious mediation of the social bond and explore the variety of forms in which the quest for spiritual perfection impacts upon lifestyle, social organization, culture, and society.

The Religious Life

For Weber, the life of the monk, the religious virtuoso *par excellence*, is in direct contrast to that of the warrior, albeit both vocational forms of life exhibit a primitive communism valorizing a segregationist ascetic ethic of celibacy, poverty, and fraternal solidarity. Warrior-monks – Hospitallers, Templars, Teutonic Knights, Samurai – have existed within the great religious traditions as attempts to integrate the call(ings) to arms and alms. However, the military life consecrated to violence and the religious life dedicated to the peaceful pursuit of an interior spiritual state and the prefiguration of the kingdom of God usually have existed in mutual tension as separate 'ways'. Like militarism, monasticism typically cultivates a compliance culture characterized by pervasive regulatory structures for the discipline of the body together with elaborate command hierarchies demanding uncritical obedience to legitimate superiors (*'alter Christi,'* in the case of Christian religious orders) as a mark of the fidelity and successful socialization of conscripts and volunteers.

Detached from the responsibilities of marriage and family, monastic life provides opportunity for both personal sanctification and an exemplary 'life for others.' Be it through life-long observance of 'the nine canonical hours' or the daily round with the begging bowl of the temporary Buddhist monk, 'the ascetic death-in-life' (Weber, 1963), serves to 'mortify the flesh' and 'raise the mind, heart and soul to God.' This quotidian self-discipline develops a unified life ofwork and prayer where the contemplation of cloister, cell, or convent integrates with the labors of field, study, or hospital ward. Weber (1963) characterized occidental monasticism as 'the disciplined army of a rational bureaucracy of office' and deemed it a mode of 'asceticism directed toward the outer world.' By contrast, 'oriental monasticism' was considered to be 'of magical character,' dominated by the innerdirected mysticism of fakirs, sages, and mystagogues.

Becoming a 'religious,' be that a scholar monkpriest, enclaustrated nun, nursing sister, or liqueurmaking brother, entails a lengthy process of formation through various graded orders of membership as entrants pass, usually via a formal investiture ceremony ('the taking of the habit'), from novice through to 'full profession.' Under the watchful eye of experienced 'directors of souls,' entrants to 'the vowed life' are absorbed within a distinctive tradition of discipleship having its own idiosyncratic mores, prayer routines, division of labor, and 'heroic' exemplars such as saints and martyrs. Thus, each separate religious 'clan' transmits its own spiritual 'disciplines' and cultural inheritance to incoming generations of recruits whose individual members 'become' religious of a given stamp: Carthusian, Carmelite, Community of the Resurrection, Shinto, Sufi, Tibetan Buddhist, and so on. Distinct 'family resemblances' unite the members of the same monastic dynasties across time

and space. This religious cultural capital manifests not only as material symbolic externals such as the cut and color of the diverse religious costumes, but also in the characteristic religious responses associated with the various spiritual 'paths' and their associated forms of life. To the outsider, all monks and nuns may look alike and engage in much the same works of divine service and charity. To the insider, however, a de la Salle Teaching Brother and a Franciscan friar are as different as chalk and cheese. A howling Dervish is not a silent Cistercian Trappist.

Civilizational Mission

Civilizations east and west have been marked inerradicably by their monastic patrimonies that have been formative at critical renewal points in the histories of peoples, empires, and cultures. Nisbet (1973) goes so far as to opine: 'Without any question, it seems to me, the monastery, especially as it was conceived by Benedict, is the true source of most of our Western ideas of community achieved through the processes of withdrawal and renewal.' Across the faith traditions east and west, monasteries as centers of 'godliness and good learning' (Leclercq, 1961) have been power-houses of liturgical, musical, and architectural creativity sites of practical wisdom in agriculture, animal husbandry, and horticulture (grape growing and bee-keeping being renowned monkish specialisms) (*see* Religion and Economic Life); pre-eminent throughout the religious traditions in the arts and sciences of calligraphy, printing, and librarianship; and, as 'experiments in living,' 'prophetic' centers of innovation in modes of self-governance and welfare administration. Despite a reputation for social conservatism, religious virtuosi from all the world faith traditions across all five continents have participated in diverse independence movements and human rights campaigns.

As a force in the development of the western social order, monasticism has spawned many regional 'styles' with their own distinctive roots and Rules but with considerable syncretism in evidence (Knowles, 1969). In the course of time, 'Celtic' monasticism, itself deriving from the Egyptian eremetical exemplars (Antony and Pachomius) fused with the reverence for learning of the Greek and Frankish scholar-monks, such as Basiland Cassianto develop the high Latin culture that was the main conduit of Christianity from the last days of the Roman Empire to the initial centuries of Christendom in the early Middle Ages. Missionary monks inspired by St. Patrick and St. Columbas took the Christian gospel into the heartlands of European tribal society in the sixth and seventh centuries. From the eighth to the eleventh centuries ('the Benedictine centuries'), the proliferation of daughter monasteries of Benedict's Monte Cassino (founded 529) created the network of organization and intelligence that shaped the character and ethos of medieval European civilization. The religious culture of early modern urban development in the thirteenth and fourteenth centuries was much influenced by the preaching friars (Dominicans, Franciscans, Augustinians) of the Mendicant Orders (Lawrence, 1994) (*see* Christianity, Origins of). Throughout Asia and the Far East the monasteries of Theravada Buddhism had comparable influence (Silber, 1995) (*see* Buddhism).

Styles of Life

The 'contemplative' and 'active' styles of religious life rarely are walled off entirely from each other. The 'world' knocks at the monastery gates in every era. Pontiffs and prelates have sought the support of abbeys and nunneries in the promotion of this or that ecclesial cause from prayers for the crusades to theological backing in the extirpation of heresy. Monarchs and nobles have found ways to secure preference and placement for one of their number in the governmental and suffrage systems of religious life. Paradoxically, these attractions and achievements have at times led to monasticism assuming considerable economic and political power (*see* Religion: Mobilization, and Power) be it that of the great medieval religious foundations (Grande Chartreuse, Cluny, Citeaux, Monte Cassino) with their vast benefices and endowments; or the self-governing republic since 1052 of the influential monastery of Mount Athos in northern Greece; or the three centuries of state control assumed by the Dalai Lama, the chief abbot of Tibetan Buddhism; or the impact since 1837 of the Senusi Islamic brotherhood on Sudan and the eastern Sahara. There is a continuous tradition of criticism of the worldly foibles and contradictions of monastic elites down the ages, from Vigilantius (fourth century), through Chaucer, Erasmus, Luther, and Calvin, to Voltaire and modern anticlerical republicanism worldwide. Liberation theology and a spate of contemporary novels and films directed against 'new world theocracies' (the Paraguayan Reductions) and colonial 'foreign missions' sustain the iconoclastic suspicion of monastic imperialism.

Successive waves of new religious foundations have been launched by charismatic (*see* Charisma and Charismatic) religious founders animated by a utopian 'return to sources' *(ad fontes)* from which 'scriptural primitivism,' the combined attack on abuses within the Church and world, have been prepared down the centuries. The inherent tension between its this-worldly and other-worldly parameters accounts in great measure for the transformative capacities of monasticism which owes its survival and dynamism to an ongoing internal reform impulse (the Gregorian Reforms, for instance). Following the upheavals of the Protestant Reformation ('the dissolution of the monasteries') and the French Revolution, Roman Catholicism invested heavily in 'apostolic orders' (Jesuits, Daughters of Charity) and missionary teaching congregations (Sallesians, Marists) which served as 'the pope's battalions' in the Tridentine 'counter-modern' offensive. The nineteenth-century Oxford Movement saw fit to revive monasticism within Anglo-Catholicism (Hill, 1973). Come the age, come the religious order. The *aggiornamento* of the Second Vatican Council (1962–65) with its emphasis on 'adaptation to the signs of the times' and 'a preferential option for the poor' (Sweeney, 1994) is the most recent expression of the sporadic self-reform tendency in religious life to check the mixed fruits of sanctity's success, social, and theological drift.

Each new age of monastic life and religious congregations addresses the perceived apostolic needs and ecclesiastical interests of the epoch be that for grand monastic citadels, hostels for crusaders and pilgrims, sanctuaries from plague and persecution, hospitals, schools, orphanages, hospices, food kitchens, or way stations for state policies of care in the community. Monasticism serves as a ligature binding civilizations east and west to their pasts and a bridge across which lessons of the religious quest are passed over into civil society.

Not surprisingly, militant secular regimes (*see* Secularization) have found pretext to sequester monastic assets and secularise (and worse) religious personnel. Monasticism and the religious orders provide the great world religions with multiple and mixed benefits: an outlet for 'charisms' to flourish then 'routinize'; an opportunity context for a specialized religious 'division of labor'; social space to cultivate a realm of inter mediate associational life – collegia – within the complex organic ecclesiastical commonwealth (the *communitas communitatum*) without need for secession from the Mother Church.

Sign of Contradiction

If, on the one side, the religious institutes represent *gemeinschaft* in its most comprehensive guise (anticipatory democracy, brotherhood of equals, protofeminist sisterhood), on the other side, monastic life can exemplify the 'total institution' as escapist flight into infantile dependency and social parasitism. As a 'sign of contradiction,' monasticism and religious life west and east, appear destined to live out these two contrary images and appeals. As many established religious orders enter the 'terminal' stage in the monastic life-cycle (Hostie, 1972) marked by membership hemorrhage, aging and low, erratic recruitment (Ebaugh, 1993), the third millennium is likely to see novel forms of monasticism – mixed gender lay communities; gay and lesbian institutes; multifaith congregations – bearing witness 'against all reason' and the 'possessive individualism' of secular modern consumer society to the resilience and revitalization (Cada et al., 1979) of 'the consecrated life.'

See also: Asceticism; Buddhism; Catholicism; Islam: Middle East; Orthodox Churches; Pilgrimage; Religions of India; Religious Organizations.

Bibliography

Cada, L., Fitz, R., Foley, G., Giardino, T., Lichtenberg, C., 1979. Shaping the Coming Age of Religious Life. Seabury Press, New York.

Ebaugh, H.R.F., 1993. Women in the Vanishing Cloister: Organizational Decline in Catholic Religious Orders in the United States. Rutgers University Press, New Brunswick, NJ.

Hill, M., 1973. The Religious Order: A Study of Virtuosi. Religious Life and Its Legitimation in the Nineteenth Century Church of England. Heinemann, London.

Hostie, R., 1972. Vie et mort des ordres religieux. Desclee de Brouwer, Paris.

Knowles, D., 1969. Christian Monasticism. Weidenfeld & Nicolson, London.

Lawrence, C.H., 1994. The Friars. The Impact of the Early Mendicant Movement on Western Society. Longman, London and New York.

Leclercq, J., 1961. The Love of Learning and the Desire for God: A Study of Monastic Culture (C. Misrahi, Trans.). Fordham University Press, New York.

Nisbet, R., 1973. The Social Philosophers. Community and Conflictin Western Thought. Heinemann, London.

Silber, I.F., 1995. Virtuosity, Charisma and Social Order: A Comparative Sociological Study of Monasticism in Theravada Buddhism and Medieval Catholicism. Cambridge University Press, Cambridge, UK.

Sweeney, J., 1994. The New Religious Order: A Study of the Passionists in Britain and Ireland, 1945–1990 and the Option for the Poor. Bellew, London.

Weber, M. (Ed.), 1963. The Sociology of Religion. Beacon Press, Boston [E. Fischoff, Trans., Intro. Parsons T.].

Monetary Policy since the 2007–2009 Financial Crisis

Benjamin M Friedman, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Monetary policy is one of the two principal means (the other being fiscal policy) by which government authorities in a market economy regularly influence the pace and direction of overall economic activity, importantly including not only the level of aggregate output and employment but also the general rate at which prices rise or fall. This article updates the earlier article on 'Monetary Policy' (link) to take account of the extraordinary developments initially triggered by the 2007–09 worldwide financial crisis and then following in the aftermath of the crisis. With short-term interest rates in most advanced industrialized economies at or near zero, soon after the crisis began, central banks have been forced to seek out new ways of carrying out monetary policy. In part these new policy tools have involved heightened efforts to influence market participants' expectations of future monetary policy actions. In other respects they have involved an expanded role for a central bank's asset holdings, rather than its outstanding liabilities which had previously been the main focus of monetary policy. Both directions raise important questions about how monetary policy affects economic activity and how best to carry out this important aspect of economic policymaking.

Monetary policy is one of the two principal means (the other being fiscal policy) by which government authorities in a market economy regularly influence the pace and direction of overall economic activity, importantly including not only the level of aggregate output and employment but also the general rate at which prices rise or fall. Even before the 2007–09 financial crisis, the predominant trend over much of the prior half century had been to place increasing emphasis on monetary policy (and correspondingly less on fiscal policy) for these purposes. In many countries, events since the crisis further shifted this balance toward monetary policy, as central banks initially acted to prevent the crisis from developing into a disastrous economic collapse and then sought to foster recovery from the downturn to which the crisis led. On both counts the specific actions that many central banks took and the broader directions that they charted during this period broke new ground in monetary policymaking. At the same time, the crisis experience highlighted the role that monetary policy plays in ensuring 'macroprudential stability' – that is, stability of the financial system as a whole, in contrast to preventing the failure of any individual financial institution.

The 2007–09 financial crisis and the protracted economic downturn that it triggered in many of the world's advanced industrialized countries represented one of the most significant economic events since World War II. The collapse of major financial firms, the decline in asset values and consequent destruction of paper wealth, the interruption of credit flows, and the loss of confidence both in firms and in credit market instruments were all extraordinary. In many countries the real economic costs – in terms of reduced production, lost jobs, shrunken investment, and foregone income and profits – exceeded that of any postwar decline. Further, in many countries the recovery from the downturn proved frustratingly slow. It is hardly surprising that monetary policy in many countries responded to these events in extraordinary ways (Borio and Disyatat, 2010; Bernanke, 2012a,b).

Special Lending to Financial Institutions

Because the 2007–09 crisis grew out of failures in the credit markets – in the first instance, the market for residential mortgages in the United States – and because many major financial institutions held claims on these troubled instruments in sufficient volume to threaten their viability, the first response of many governments and central banks was to act to prevent the kind of cascading failures that had turned the downturn of the early 1930s into the Great Depression. Very early in the crisis many central banks instituted special facilities for lending to banks and in some cases to other financial institutions as well. In December 2007, the U.S. Federal Reserve System established a Term Auction Facility to channel extra central bank lending, and on unusually favorable terms (as well as with anonymity, to avoid the reputational damage often consequent on large-scale borrowing from the central bank), to commercial banks doing business in the United States. Early in 2008, the Federal Reserve established two additional facilities, enabling it to lend against a broader class of collateral and to lend to nonbanks. As early as August 2007, the European Central Bank (ECB) had undertaken a special intervention to inject liquidity into the Eurozone banking system and had declared its readiness to provide unlimited further liquidity if needed. In September 2008, the ECB undertook a further special 'term refinancing' operation (in other words, lending to banks for longer than normal). In May 2009, the ECB changed its method of regular weekly policy operations to remove the previous limit on banks' access to central bank funds, thereby making liquidity available on an unlimited basis at the main refinancing rate. It also lengthened the maturity of its lending, expanded the range of eligible collateral, and introduced new facilities that enabled it to transact with a broader range of counterparties. Later, at the end of 2011, the bank acted to prevent a renewed liquidity crisis among Eurozone banks by

offering banks 3-year financing under a new 'long-term refinancing operation.' In April 2008, the Bank of England introduced a special liquidity scheme to enable struggling U.K. banks to exchange illiquid assets for assets like Treasury bills that could then be used as collateral for further borrowing. The Bank of Japan had already been employing such extraordinary policy measures for some years by the time the 2007–09 crisis hit. These steps included an expanded securities lending facility, a broader range of collateral (including private securities) eligible for central bank lending, and even some direct funding for Japanese corporations.

Lending by central banks to banks, and even to nonbank financial institutions, was not new in the 2007–09 crisis. Nevertheless, the actions that the major central banks took in response to the crisis were different from what they had done in the past, both in scale and in scope. (In addition to these central bank actions, governments in some countries took highly unusual steps to prevent the failure of a variety of bank and nonbank institutions, even including nonfinancial firms. The U.K. government, for example, took an 81% share in Royal Bank of Scotland, and at the peak the U.S. government owned more than one-third of the equity in Citigroup and more than 90% of that in AIG. The U.S. government also rescued from failure two of the three large American automobile manufacturers (Cecchetti and Disyatat, 2010; Ashcraft et al., 2012)).

Conventional Monetary Policy, Down to the Zero Lower Bound

Once it became apparent that the financial crisis was likely to have significant effects beyond the financial markets, depressing output and employment in many economies (with yet further feedback onto the viability of their lending institutions), central banks turned to expansionary monetary policy in the usual way – that is, by lowering short-term interest rates. As late as September 2007, the U.S. Federal Reserve System's target for the federal funds rate, its customary policy rate, had been 5¼%. After a sequence of reductions, in December 2008, the Federal Reserve adopted a target range of 0–¼%. The ECB lowered its main refinancing rate from 4¼% as late as October 2008 to 1% in May 2009. Beginning from 5¾% throughout the summer and fall of 2007, the Bank of England first lowered its official bank rate in December of that year and did so in repeated steps until it reached ½% in March 2009. In Japan, where short-term interest rates had already been at very low levels for years, owing to the economy's stagnation and the gradual average decline of prices, the Bank of Japan lowered its uncollateralized call loan rate from ½% in October 2008 to just 1/10% in December of that year.

The central banks of the advanced high-income economies also coordinated these actions to an extraordinary extent. In October 2008, for example, six central banks – the U.S. Federal Reserve System, the Bank of Canada, the Bank of England, the ECB, the Swiss National Bank, and the Sweden's Riksbank – simultaneously announced a ½% reduction in each bank's policy interest rate. The Bank of Japan could not match this reduction because its policy interest rate was already so low, but it nevertheless publicly expressed its support for this action.

Especially in the United States and Japan, these actions effectively placed the central bank's policy interest rate at the practical lower bound. Because investors in most economies can always earn a zero nominal interest rate by holding currency, which is risk-free apart from physical security concerns, the central bank cannot force the expected nominal return on any other asset below zero; investors would merely substitute holdings of currency instead. In most periods of weak economic activity in the past, even with easier monetary policy the nominal interest rates on short-term market instruments never reached the zero lower bound because price inflation was sufficiently rapid that the central bank could implement whatever negative real interest rate (the nominal rate minus expected price inflation) it chose to achieve the desired stimulus to economic activity while nominal rates remained above zero. But by the time of the 2007–09 crisis, inflation had been low, in most of the industrialized world, for some years. As a result, the initial actions central banks took to stimulate their economies quickly pushed nominal short-term interest rates to, or close to, the zero lower bound. Many central banks therefore turned to other – in many cases, highly experimental – ways of providing further economic stimulus (Krugman, 1997; Bodenstein et al., 2012; Chung et al., 2012; Ball, 2013).

Forward Guidance on the Policy Interest Rate

One approach that some central banks had tried before (for example, the Federal Reserve System in the early 2000s), was the use of public statements intended to influence market participants' expectations of the *future* level of the policy interest rate over some economically significant time horizon. Once the policy interest rate approached the lower bound in the 2007–09 financial crisis, many central banks turned almost immediately to this kind of 'forward guidance' with respect to their own future actions. Moreover, as central banks attempted to place greater reliance on this policy tool during the crisis and its aftermath, they continually experimented with various ways of expressing their intentions.

When the Federal Reserve reduced its target for the federal funds rate to 0–¼%, in December 2008, the Open Market Committee (the Federal Reserve group responsible for such decisions) publicly stated that "the Committee anticipates that weak economic conditions are likely to warrant exceptionally low levels of the federal funds rate for some time." The Committee repeated substantially the same public statement following each meeting that it held through June 2011. In August 2011, it then switched to a more explicit, calendar-based statement that "the Committee currently anticipates that economic conditions … are likely to warrant exceptionally low levels for the federal funds rate at least through mid-2013." Thereafter, the Committee extended the specific horizon for this

forward guidance several times, by late 2012 indicating the anticipation of low short-term interest rates 'at least through mid-2015.' In December 2012, the Committee shifted to yet a new form of statement explicitly conditioned on observed and forecast economic outcomes rather than a specific calendar date: "the Committee … currently anticipates that this exceptionally low level for the federal funds rate will be appropriate at least as long as the unemployment rate remains above 6½%, inflation between 1 and 2 years ahead is projected to be no more than a half percentage point above the Committee's 2% longer-run goal, and longer-term inflation expectations continue to be well anchored." In January 2014, with the latest reported unemployment rate down to 6.7%, the Open Market Committee changed the language of its regular statement once again, to emphasize that if the unemployment rate were to fall below 6½% this event alone would not necessarily trigger an increase in the target federal funds rate: "In determining how long to maintain a highly accommodative stance of monetary policy, the Committee will also consider other information, including additional measures of labor market conditions, indicators of inflation pressures and inflation expectations, and readings on financial developments. The Committee continues to anticipate, based on its assessment of these factors, that it likely will be appropriate to maintain the current target range for the federal funds rate well past the time that the unemployment rate declines below 6–1/2%, especially if projected inflation continues to run below the Committee's 2% longer-run goal." In March 2014, the Committee changed its forward guidance again, deleting the reference to a specific unemployment rate altogether and indicating a continuation of the low interest rate policy even after the achievement of the 6–1/2% level: "In determining how long to maintain the current 0–¼% target range for the federal funds rate, the Committee will assess progress – both realized and expected – toward its objectives of maximum employment and 2% inflation. This assessment will take into account a wide range of information, including measures of labor market conditions, indicators of inflation pressures and inflation expectations, and readings on financial developments. … The Committee currently anticipates that, even after employment and inflation are near mandate-consistent levels, economic conditions may, for some time, warrant keeping the target federal funds rate below levels the Committee views as normal in the longer run."

Other major central banks likewise employed public statements about their future policy actions, although not synchronously with the Federal Reserve. The Bank of Japan, faced with deflationary conditions, had already been using a form of forward guidance as early as 1999, and in 2001 had strengthened the language it used to refer not just to declining prices but real economic activity as well. The bank returned to this tactic after the crisis as well. In October 2010, it announced that the policy interest rate would remain at zero until the economy had regained price stability, and in April 2013, it further linked the zero interest rate to realization of the bank's new 2% inflation target. The Bank of England turned to forward guidance only in August 2013, announcing that its Monetary Policy Committee "intends not to raise Bank Rate from its current level of 0.5% at least until the Labour Force Survey headline measure of the unemployment rate has fallen to a threshold of 7% … ." In February 2014, the Committee indicated – in parallel with the Federal Reserve's statement the month before – that it would most likely keep the rate low even after unemployment reached 7% or below. The ECB made no explicit statement about future levels of its policy interest rate until July 2013, when (echoing what the Federal Reserve had done early on in the crisis) it announced that it intended to keep the rate low for an 'extended period of time.'

The purpose of this longer-horizon guidance on the policy interest rate, at each respective central bank, was not to help enforce the target for the policy rate itself – under conditions initially of the crisis and then the protracted slow postcrisis recovery this was unnecessary – but rather to influence market participants' expectations in such a way as to help lower the interest rates on longer-term securities, like mortgages and corporate bonds, by which households and firms finance their spending. Many investors can substitute between holding longer-term securities and holding a comparable-duration sequence of short-term instruments in their portfolios. Lowering the policy rate that investors expect to prevail in the future induces investors to seek to shift their holdings toward longer-term securities, thereby also lowering the longer-term interest rates that presumably matter for private economic activity. There is some evidence, from experience before the crisis as well as after, of effects on market interest rates associated with this kind of announcement. In addition, expectations that short-term interest rates would remain low for an extended time would plausibly encourage greater market demand for assets like equities, helping to raise their prices and thereby increase household wealth. On both counts, the goal was to stimulate aggregate demand for goods and services in the economy, and therefore boost output and employment from crisis-depressed levels. (By contrast, there have been some instances in which the effect of such statements of intentions regarding future central bank actions has been counterproductive from the perspective of the objective apparently motivating the statements in the first place. Most prominently, when Federal Reserve officials first referred publicly, in the spring of 2013, to the likely cutback and eventual termination of its ongoing bond purchase program (see Section Large-Scale Asset Purchases below), the immediate market reaction raised long-term interest rates by more than most estimates of the program's total effect to that time in lowering them (Gurkaynak et al., 2005; Levin et al., 2010; Justiniano et al., 2012; Kool and Thornton, 2012)).

Large-Scale Asset Purchases

With short-term interest rates effectively at the zero lower bound, many central banks also sought to bring down longer-term interest rates in ways that went beyond the expectational effect of forward guidance on the future

trajectory of the policy interest rate. Doing so led some central banks to become major purchasers of various kinds of longer-term assets – not for the purpose of expanding their liabilities, as under traditional monetary policy, but for the effect on asset prices and yields consequent on the purchases of the assets themselves. In effect, as a result of the crisis the focus of monetary policy switched from the liability side of central banks' balance sheets (providing currency and especially reserves) to the asset side (buying and holding financial instruments issued by other borrowers).

In the United States, the Federal Reserve System first began to purchase large quantities of unusual assets in October 2008, following the collapse of the market for commercial paper (unsecured short-term borrowing by private financial institutions and nonfinancial corporations) triggered by the failure of Lehman Brothers. Fairly quickly, the central bank accumulated $385 billion of commercial paper; its doing so stabilized what had become a dysfunctional market, thereby reducing the interest rate borrowers had to pay and also reviving the volume of new lending. In March 2009, the Federal Reserve instituted a new program to purchase securities backed by residential mortgages. By year-end, it had accumulated $900 billion of these securities, once again restoring functionality to a badly damaged market, and resulting in lower interest rates to borrowers as well as increased volume.

These purchases sharply increased the size of the Federal Reserve's balance sheet. By year-end 2008, total assets held were $2.2 trillion, compared to less than $900 billion at the beginning of 2007, although most of the increase consisted of advances to banks and other financial firms, and holdings of instruments like commercial paper acquired through directed facilities intended to support specific markets. At year-end 2009, total asset holdings were still $2.2 trillion, but in place of direct advances under special facilities the Federal Reserve had purchased large quantities of Treasury notes and bonds and, even more so, of mortgage-backed securities. By mid-2014–5½ years into the period of zero short-term interest rates – total assets held reached $4.3 trillion, more than quadruple the level at the beginning of the crisis.

Other central banks pursued similar policies, in either greater or lesser amounts. In May 2010, the ECB initiated a program to buy directly securities that it normally accepted as collateral in lending. In September 2012, the ECB converted the program to one of 'outright monetary transactions,' under which its purchases of member states' government bonds were to be permanent rather than temporary holdings. Under these programs the bank purchased enough assets, and did enough lending, that its balance sheet grew from €1.1 trillion at the beginning of 2007 to €3.1 trillion at year-end 2012. The Bank of England began its asset purchase program in March 2009 and progressively increased its volume of monthly purchases. By late 2012, the increase in the bank's balance sheet was proportionately even greater than that of the Federal Reserve: from £85 billion at the beginning of 2007 to more than £400 billion. In Japan, where the central bank had already engaged in policies that greatly expanded its balance sheet well before the crisis, the increase was initially smaller. But following the election of a new government in December 2012, together with the appointment of new leadership at the Bank of Japan the following spring and the bank's subsequent commitment to raising Japan's inflation rate to 2% per annum, the rate of increase accelerated abruptly: starting from ¥110 trillion at the beginning of 2007, the bank's total assets held reached ¥258 trillion by mid-2014.

The purpose of these asset purchases was sharply different from the usual working of central bank open market operations. There, the object is to reduce the market-clearing interest rate on central bank liabilities, taking advantage of the interest-elastic demand for central bank reserves on the part of banks and other reserve-holding institutions (see Section 2 of the 2000-article on Monetary Policy). Because the central bank must pay for any assets it acquires, open market purchases of assets are simply the way in which the central bank brings about an increase in the quantity of reserves that it supplies to the market. Whether the central bank frames the operation as choosing a larger quantity of reserves and letting the market determine the corresponding new (presumably lower) equilibrium interest rate, or choosing a lower interest rate and supplying whatever (presumably larger) quantity of reserves the market requires to get there, is irrelevant for this purpose. What matters in practical terms is the impact on the market-clearing short-term interest rate.

The point of the asset purchases that central banks made during and after the 2007–09 financial crisis, once short-term interest rates were already at or near zero, was plainly different. Now the object was to lower the interest rates that mattered more directly for the private sector's demand for goods and services, taking into account that no private sector borrower issues default-free obligations or finances its expenditures with overnight borrowing. On grounds of both maturity and default risk, private-sector obligations are not perfect substitutes for default-free overnight instruments in investors' portfolios. The return on the relevant private-sector obligations therefore depends on more than just the current level and expected future trajectory of the central bank's policy interest rate (typically an overnight rate, or some other very short maturity). In particular, it depends on the quantity of private-sector obligations that investors, collectively, must hold. Asset purchases by the central bank therefore add to the total market demand for private-sector securities, and so, for given supply, they raise the market price of these securities, thereby lowering the corresponding market-clearing interest rates.

In the ordinary open market operations that central banks carry out when their policy interest rates are not constrained by the zero lower bound, the volume of assets purchased is too small to have noticeable effects of this kind. But the asset purchase programs carried out by many central banks since the 2007–09 crisis have been several orders of magnitude larger, and most analyses of these purchases confirm that they have had such effects in sufficient size to exert a meaningful effect on the economy. Estimates of this policy's effectiveness in lowering long-term interest rates in the United States, scaled to a $600 billion asset purchase

(the size of the Federal Reserve's 'QE2' round of bond buying announced in November 2010 and carried out over the following 7 months), average about ¼ percentage point – not enormous, especially in light of conventional estimates of the effect of lower interest rates on aggregate demand, but a contribution to the economic recovery process that Federal Reserve policymakers were seeking. Moreover, the effect of lower long-term interest rates was probably reinforced by corresponding movements in equity prices and in the foreign exchange value of the dollar (although in the case of exchange rates any conclusion is necessarily conditional on the assumption made about the policy stance underlying the other currency).

Further, the composition of the assets that the central bank buys matters too. In the U.S. experience, the Federal Reserve bought not just long-term assets but mortgage-backed securities in particular; as of mid-2014, the Federal Reserve held $1.7 trillion of such securities, out of a total balance sheet of $4.3 trillion. The intended effect was to narrow the spread between the interest rates on mortgage borrowing versus comparable-maturity Treasury obligations, and thereby to arrest the downward trend in house prices as well as to stimulate new residential construction. In light of the role of the mortgage market as the epicenter of the 2007–09 financial crisis, with consequent effects on residential construction (U.S. homebuilding declined from 2.1 million new houses started in 2005 to just 554 000 in 2009) the central bank's desire to stimulate demand in this one sector of the economy in particular is easily understandable. Clear evidence on the effectiveness of these mortgage purchases on new construction is not yet available, but on the surface it appears that at least the attempt to narrow mortgage–Treasury interest rate spreads was effective.

Potential Risks from Large-Scale Asset Purchases: Price Inflation?

Because central banks have to pay for assets that they purchase, the large increase in asset holdings with which many of them responded to the 2007–09 crisis likewise meant an equally large increase in their outstanding liabilities. Economists who view the macroeconomy through the lens of central bank (or 'high powered') money supply therefore expected not just some increase in these economies' rates of inflation but a hyperinflation, perhaps even reminiscent of what occurred in Germany and Russia after World War I or in Hungary after World War II. Some well-known economists predicted just such a consequence of these extraordinary central bank actions.

But as of mid-2014, more than 5 years into the implementation of postcrisis policies, no such increase in inflation – not even a few percentage points – had appeared, in any of these economies. Even if inflation in these countries does pick up pace at some point, a response of some percentage points to this magnitude of increase in central bank liabilities, and only after a delay of more than half a decade, would still be sharply at variance with the core implications of standard quantity-theory models relating prices to central bank money. The central idea of the quantity theory, familiar from generations of textbooks prior to the crisis, was that doubling the quantity of central bank liabilities (or 'monetary base') would lead to a doubling of the price level. In some of the major economies, the quantity of central bank liabilities not just doubled but quadrupled. Correspondingly, several generations of empirical research in monetary economics, prior to the crisis, had documented the relationship between one or another economy's price level and the quantity of liabilities issued by its central bank. In light of the large magnitude of the expansion in central bank balance sheets during and following the crisis, however, together with the absence of any noticeable response from prices, it will be a long time before anyone will again be able to find economically sensible estimates, for any of the large high-income economies, from what were once standard empirical relationships between the price level (or its rate of change) and the quantity (or change) of central bank liabilities.

Potential Risks from Large-Scale Asset Purchases: Portfolio Losses?

Central banks normally take little risk. Most of the assets that they hold are usually limited to short-term obligations of their own governments. In advanced industrial countries, therefore, where the probability of the government's defaulting on its obligations is remote, the risk of loss to the central bank is minimal.

The assets that many central banks purchased during and after the 2007–09 financial crisis were different. Many were of longer term, and therefore subject to the risk of market price decline if interest rates were to rise. Some – for example, commercial paper in the United States, and mortgage-backed securities in the United States and the United Kingdom – were issued by private borrowers, and therefore subject to risk of potential default. (In some cases, however, private obligations bought by the central bank were guaranteed by the country's government; the mortgages bought by the Federal Reserve, for example, were insured by the U.S. federally sponsored credit agencies, which by then had been taken into receivership by the U.S. Government.)

Because central banks typically operate with very little capital, risk of significant loss on their asset holdings, either from fluctuation in the market value of longer-maturity obligations or from default by private-sector issuers, raises the prospect of a reduction in capital, and perhaps even a technical insolvency. As of mid-2014, no such losses had eventuated; instead, holding the securities that they purchased during and in the aftermath of the crisis had delivered record portfolio profits to those central banks that pursued this course. Once the financial effects of the crisis are in the past, however, and market interest rates return to a more normal structure (with short-term rates no longer at the zero lower bound), some losses of this kind are likely. Such losses would represent a form of negative seigniorage – in contrast to the record gains implicitly earned by taxpayers from 2008 to 2014 – and therefore ultimately impose a burden on the country's taxpayers.

Portfolio losses of this kind pose no real risk to the central bank itself, however. In contrast to a private bank or other

private financial institution, a central bank does not need positive capital to operate (Curdia and Woodford, 2011; Gagnon et al., 2011; Gertler and Karadi, 2011; Krishnamurthy and Vissing-Jorgenson, 2011; Friedman, 2015).

New Questions about the Future Conduct of Monetary Policy

The extraordinary actions that many central banks took during the 2007–09 financial crisis, and in the lengthy economic recovery following it, have raised a variety of new questions about the future conduct of monetary policy, once the effects of the crisis have finally passed. Three stand out as the most significant.

The Ongoing Usefulness of Forward Guidance

Central banks have increasingly relied on 'forward guidance' during this experience – that is, public statements about their own likely future policy actions, intended to shape the expectations of market participants – in part as a way to make monetary policy more effective when reductions in the policy interest rate are precluded by the zero lower bound. In principle, however, there is no reason not to resort to such statements as a supplement to ordinary interest rate policy once short-term interest rates are no longer constrained.

To what extent central banks will do so remains uncertain. The problem lies in the nature of speculative asset markets. Economists who study monetary policy often work with models in which the structure of the relevant expectations on the part of financial market participants and of private economic agents more generally is extremely simple (in some standard models the only aspect of monetary policy that influences economic agents' behavior is the central bank's long-term inflation target), but the world in which actual central banks operate is different. Even if the central bank steers its monetary policy according to a fixed inflation target, the Bank of England does, participants in the securities markets still want to know just how much departure from the inflation target policymakers will tolerate, and under what circumstances, before they decide to raise or lower the bank's policy interest rate, or buy or sell assets, or engage in any of a variety of other policy actions. If the central bank follows a more loosely defined inflation targeting regimen, as in the case of the ECB, or has explicit multiple policy objectives, as the Federal Reserve does, the opportunities for such speculation are all the richer.

Market participants are in the business of speculating, for profit, on the actions that the central bank takes and the timing with which it takes them. No matter what level of transparency and disclosure the central bank provides, therefore, unless it locks its policy trajectory into determination under a rigid formula, which it publicly discloses and commits to follow without departures, the market will inevitably want to know more than policymakers can possibly disclose. Even then, the central bank's no-departures commitment may well not be fully credible, so that market participants will continue to speculate about policymakers' actions – although in this case the speculation is framed in terms of whether, and if so under what circumstances, they will violate their pledge to adhere to the formula.

The Ongoing Usefulness of Asset Purchases and Sales

Parallel questions surround the potential future use of the other new monetary policy tool to which many central banks turned in the absence of ability to lower their policy interest rates: asset purchases. The value to monetary policy of the ability to influence long-term interest rates in general, and specific interest rates like the mortgage rate in particular, likewise does not vanish just because the short-term interest rate that the central bank normally uses as its policy interest rate is above zero and potentially subject to being lowered. Further, in contrast to forward guidance, the evidence on the effectiveness of such asset purchases is more clear-cut. As a result, there is a reasonable likelihood that, at least occasionally, central banks will continue to use some form of asset purchases – and, conversely, asset sales when the object is to raise long-term interest rates in order to restrain economic activity – once the zero lower bound is no longer binding.

Further, the potential usefulness of such purchases and sales is not limited to attempts to stimulate or restrain aggregate economic activity. During the years before the crisis, for example, when U.S. house prices were rising rapidly and homebuilding had likewise reached extremely high levels, many observers of monetary policy urged Federal Reserve policymakers to tighten policy in order to arrest the developing 'bubble' in the mortgage and housing markets. The frequently expressed concern, however, was that raising short-term interest rates was a 'blunt instrument' with which to attack an excess in one specific sector of the economy. By contrast, if the central bank can deploy its balance sheet separately from its setting of its policy interest rate – and, if a part of its balance sheet is its holdings of mortgage-backed securities, which it can either enlarge or reduce in order to lower or raise mortgage rates relative to other interest rates – it can take more directed action in the event of a situation like the one that preceded (and, in the eyes of many economists, precipitated) the crisis.

Keeping this new policy tool available in a symmetrical way, so that the central bank can make either asset purchases or sales depending on the needs of the economy, also has a bearing on the extent to which central banks should seek to return their balance sheets to the pre-crisis size – which for many would mean a large reduction – once their economies regain a more normal footing. Having sizable asset sales available as an ongoing tool of monetary policy requires that the central bank not downsize its balance sheets to the pre-crisis level and composition. A central bank can sell securities only if it owns them in the first place.

New Ways of Implementing Conventional Monetary Policy

The increased size of central bank balance sheets resulting from the asset purchase programs carried out during and after the 2007–09 financial crisis also raises potential

challenges for the implementation of conventional monetary policy through raising or lowering the central bank's policy interest rate. Enlarged asset holdings mean an enlarged quantity of central bank liabilities outstanding. Because little occurred during the crisis to increase the public's demand for currency, the counterpart of the large increase in assets held by these central banks was mostly an increase in the reserves they provided to banks. Further, because banks' deposits did not increase in any extraordinary way either, the great bulk of this increase in reserves consisted of excess reserves.

The danger potentially created by having such a huge volume of excess reserves outstanding is that once the economy begins to expand banks are therefore free to extend credit and create deposits in what would be almost unlimited amounts. To prevent them from doing so by traditional means, the central bank would have to shrink its volume of outstanding reserves – by selling assets – back to the level at which reserves became sufficiently scarce to raise market interest rates to the level needed to restrain economic activity, and with it potential inflationary pressures. One question raised by such a prospect is whether the assets that the central bank had accumulated in response to the crisis (mortgage-backed securities, for example) would have a ready market to permit sales in the required volume. A further question is whether, as the central bank sold off these assets in sufficient volume to enable it to raise its policy interest rate, these sales would so depress the prices of the remaining assets on its balance sheet as to create large losses (see Section Potential Risks from Large-Scale Asset Purchases: Portfolio Losses? above).

Most central banks regularly pay interest on banks' holdings of excess reserves, however. (In the United States the Federal Reserve System received legal authority to do so starting in November 2008.) With interest paid on excess reserves, the central bank does not need to reduce the quantity of reserves it supplies in order to raise short-term interest rates. It can simply raise the rate at which it pays banks to hold excess reserves – thereby raising the effective floor on market interest rates, since no bank would ordinarily lend to a customer at a lower interest rate than it would receive from holding risk-free reserves with the central bank. Paying interest on excess reserves therefore in effect severs the traditional connection between the quantity of reserves supplied and the interest rate at which reserves are exchanged for other assets, and hence market interest rates more generally. Once the economic and market effects of the 2007–09 crisis are in the past, central banks may choose to sell part or even all of the large volumes of assets that they purchased, but there is no need for them to do so merely in order to raise their policy interest rates. Indeed, over time it is possible that conventional open market operations may become much less relevant to the actual conduct of monetary policy.

Macroprudential Concerns

In seeking to understand the origins of the financial crisis of 2007–09, which proved so costly to so many economies, economists and policymakers have focused not just on the lax lending standards and inadequate supervision applied to many lending institutions (especially in the United States) but also on the extended period of low interest rates early in that decade, which apparently led many institutions to take on risks that in the event they proved unable to sustain. The result, in most of the high-income advanced economies, has been a new emphasis on financial regulation and also a new awareness of the extended trade-offs associated with monetary policy.

Financial regulation has traditionally focused on the soundness of individual institutions. By contrast, the new emphasis on macroprudential stability focuses on the systemic risk arising from the interconnectedness of financial institutions, in that one institution's assets are another's liabilities so that the lending institution is ultimately no stronger than the borrower. In the wake of the crisis, the United States enacted the Dodd–Frank Act of 2010, which established a new Financial Stability Oversight Council specifically charged with responsibility for such systemic concerns, and also implemented a wide range of new regulations on financial institutions as well as procedures for prompt government intervention in the event of threatened failure. The European Union began the process of unifying its member countries' banking systems, importantly including proposals for a uniform set of procedures for resolving bank failures. At the same time, countries around the world moved to designate specific institutions as 'systemically important,' thereby subjecting them to new layers of reporting and regulation. Also, under the rubric of the Basel Committee, central banks moved to raise banks' capital requirements as well as to impose new restrictions on banks' leverage based on each bank's total balance sheet (in both cases scheduled to take effect in 2018.)

With short-term interest rates at or near zero for more than half a decade following the crisis, in many of the large high-income economies, concerns over a potential repetition of the experience that led up to the crisis have also played a major role in the ongoing discussion. The fear is that institutions seeking to earn at least some positive return on their assets will increasingly turn to either longer-maturity holdings or holdings more exposed to default and other risk than is normal – or appropriate – for their amount of capital and exposure to volatility of their liabilities. The result has been growing awareness of a new form of monetary policy trade-off: between keeping the central bank's policy interest rate at the level that best achieves policymakers' macroeconomic objectives (under the postcrisis circumstances, as low as possible so as to stimulate economic activity in the face of lingering unused resources) and putting interest rates at a level consistent with the ordinary conduct of financial institutions' business in a way not exposed to excess risk (presumably much higher) (Goodfriend, 2002; Hanson et al., 2011; Friedman and Kuttner, 2011; Stein, 2013).

See also: Economic Growth; Financial Crises, History of; Financial Regulation; Macroeconomics, History of From 1933 to Present; Public Policy; Regulation, Political Economy of.

Bibliography

Ashcraft, A., Malz, A., Pozsar, Z., 2012. The federal reserve's term asset-backed securities loan facility. Economic Policy Review 18, 29–66 (Federal Reserve Bank of New York).

Ball, L.M., 2013. The case for four percent inflation. Central Bank Review 13, 17–31 (Central Bank of Turkey).

Bernanke, B.S., 2012a. The Federal Reserve's Response to the Financial Crisis. http://www.federalreserve.gov/newsevents/lectures/federal-reserve-response-to-the-financial-crisis.htm.

Bernanke, B.S., 2012b. The Aftermath of the Crisis. http://www.federalreserve.gov/newsevents/lectures/the-aftermath-of-the-crisis.htm.

Bodenstein, M., Hebden, J., Nunes, R., 2012. Imperfect credibility and the zero lower bound. Journal of Monetary Economics 59, 135–149.

Borio, C., Disyatat, P., 2010. Unconventional monetary policies: an appraisal. Manchester School 78 (Suppl.), 53–89.

Cecchetti, S.G., Disyatat, P., 2010. Central bank tools and liquidity shortages. Economic Policy Review 16, 29–42 (Federal Reserve Bank of New York).

Chung, H., Laforte, J.-P., Reifschneider, D., Williams, J.C., 2012. Have we underestimated the likelihood and severity of zero lower bound events? Journal of Money, Credit and Banking 44, 47–82.

Curdia, V., Woodford, M., 2011. The Central-Bank balance sheet as an instrument of monetary policy. Journal of Monetary Economics 58, 54–79.

Friedman, B.M., 2015. Has the financial crisis permanently changed the practice of monetary policy? Has it changed the theory of monetary policy? Manchester School.

Friedman, B.M., Kuttner, K.N., 2011. The implementation of monetary policy: how do central banks set interest rates?. In: Friedman, B.M., Woodford, M. (Eds.), Handbook of Monetary Economics, vol. 3B. Elsevier, Amsterdam.

Gagnon, J., Raskin, M., Remache, J., Sack, B., 2011. The financial market effect of the federal reserve's large scale asset purchases. International Journal of Central Banking 7, 3–43.

Gertler, M., Karadi, P., 2011. A model of unconventional monetary policy. Journal of Monetary Economics 58, 17–34.

Goodfriend, M., 2002. Interest on reserves and monetary policy. Economic Policy Review 8, 77–84 (Federal Reserve Bank of New York).

Gurkaynak, R., Sack, B., Swanson, E., 2005. The sensitivity of long-term interest rates to economic news: evidence and implications for macroeconomic models. American Economic Review 95, 425–436.

Hanson, S.G., Kashyap, A.K., Stein, J.C., 2011. A macroprudential approach to financial regulation. Journal of Economic Perspectives 25, 3–28.

Justiniano, A., Evans, C.L., Campbell, J.R., Fischer, J.D.M., 2012. Macroeconomic effects of FOMC forward guidance. Brookings Papers on Economic Activity 1–80.

Kool, C.J.M., Thornton, D.L., 2012. How effective is central bank forward guidance? Federal Reserve Bank of St. Louis, Working Paper 2012-063.

Krishnamurthy, A., Vissing-Jorgenson, A., 2011. The effects of quantitative easing on interest rates: channels and implications for policy. Brookings Papers on Economic Activity 215–287.

Krugman, P.R., 1997. It's baaack: Japan's slump and the return of the liquidity trap. Brookings Papers on Economic Activity 137–205.

Levin, A., Lopez-Salido, D., Nelson, E., Yun, T., 2010. Limitations on the effectiveness of forward guidance at the zero lower bound 2010. International Journal of Central Banking 6, 143–189.

Stein, J.C., 2013. Overheating in Credit Markets: Origins, Measurement and Policy Responses. http://www.federalreserve.gov/newsevents/speech/stein20130207a.htm.

Money and Finances, Psychology of

Eva Jonas and Jochim Hansen, University of Salzburg, Salzburg, Austria
Sabine Walper, Deutsches Jugendinstitut, Munich, Germany
Dieter Frey, Ludwig-Maximilians-Universität, Munich, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E. Jonas, S. Walper, D. Frey, volume 15, pp. 9984–9988, © 2001, Elsevier Ltd.

Abstract

This article starts with a section that emphasizes the importance of cultural factors influencing psychological aspects of money. General issues of 'what money can buy' are discussed including effects of money on family formation and disruption. The next sections address attitudes toward money and psychological consequences of merely thinking about money. The section 'handling money' reviews topics such as consumption, saving, and investing money, as well as evading taxes and donating money. Correlates of financial behavior are discussed. The final section focuses on issues of poverty and discusses the impact of scarce financial resources on well-being and social relations.

Introduction

As a major tool for exchanging goods and services, money plays a crucial role in economic and social life. It serves as a legal tender, unit for calculation, and allows keeping values over time. Tied to these economic functions, money has a large influence on human experience and behavior, delimiting consumer options, life styles, access to education, and even chances on the marriage market. Societies differ considerably regarding the distribution of income among their members and in the extent to which individual economic resources determine access to cultural goods such as education and health care. But individual differences in income almost invariably provide a major indicator of social inequalities (*see* Justice: Social Psychological Perspectives). Being one of the multiple markers of social stratification, money carries social prestige and may affect social relations as well as people's self-concept, experience, cognitions, and social behavior not only due to its economic function but also through the symbolic value attached to it. Thus, from a psychological perspective, money is much more than a neutral unit of account that the *homo economicus* strives to maximize, but has symbolic power that influences thinking, experiences, and behavior.

Despite the trend to increasing globalization that is particularly evident in international finances and the increasing importance of stock holdings among private investors in many nations, societies still differ beyond their financial systems in cultural norms and collective rationale, which shape attitudes toward money. Collective experiences with money can affect whole nations, as evident in the fear of inflation in Germany that followed the experiences after World War I, and recent and current economic crises that affect individuals as well as whole economies alike. This points to the importance of socialization not only for attitudes toward money but also for expectations concerning the larger economic system.

Furthermore, cultural norms and experiences with money guide the display of one's financial resources in public. People frequently discuss general financial issues such as prices and the tax system. However, in many cultures – except for materialistic cultures such as many cultures in South Asia – personal financial assets are mostly considered a very private issue, which is not discussed even with family members or friends. This lack of openness may serve to prevent social envy in the case of personal affluence and to help protecting one's social image if personal resources are scarce. Interestingly, other social symbols such as clothing and expensive cars allow to undermine this social norm, indicating that the social norm of avoiding talking about money only holds on an abstract level of finances but not on a level of concrete products.

Given the likely importance of cultural factors, available research on psychological aspects of money and finances suffers from a focus on Western systems. Furthermore, psychologists have paid surprisingly little attention to money as research topic (for an overview, see Furnham and Argyle, 1998) leaving this domain largely to sociologists and economists. Only recently, research in social and consumer psychology started to investigate the psychological consequences of money. Drawing on relevant research from different disciplines, this contribution first addresses the significance of money for personal well-being and social relations. Next, we turn to beliefs and values connected with money, and discuss the cognitive and behavioral effects of merely thinking about money. Thereafter, we address issues of handling money looking at correlates and predictors of spending, saving, and donating money, as well as evading taxes. In a final section, selected research on effects of poverty is summarized.

What Money Can Buy: The Personal and Social Significance of Money

Although folk wisdom claims that money does not make happy, findings point to an overall low, but significantly positive relationship between income and well-being (within-nation correlations range about between $r = .15$ and $r = .20$, Diener and Seligman, 2004). Not surprisingly, the prediction of happiness from economic resources improves if other indicators of wealth besides income are considered. Pointing to the importance of social comparison processes, relative income is more important than its absolute level because people define

themselves as wealthy or poor in comparison to others. In countries with high poverty rates and extreme income differences like India or Brazil, the relation between happiness and income is substantially stronger. Across 55 nations, the correlation between income and happiness reaches about $r = .50$ (Diener et al., 1995). Finally, increases in income during prosperous periods have a positive influence on citizens' well-being, whereas income losses during depression periods have the opposite effect. However, effects of changes in income on changes in satisfaction do not seem to last long. People adapt to higher as well as lower levels of income (Easterlin, 2003).

Much attention has been paid to the significance of money in family formation and family life (White and Rogers, 2000). Economic assets play a considerable role in marriage decisions, not only among women, who have traditionally been assumed to consider their partner's bread-winning potential, but also among men. Men's earning and employment have consistently been found to increase the likelihood of marriage, and even women's increasing participation in labor force does not seem to render male income less important. Furthermore, the income of men and women – in particular the difference in income between husbands and wives – affect the division of household labor between marital partners beyond time restrictions due to their employment. There is more sharing of household tasks between spouses if wives contribute more to family income (Coltrane, 2000). Economic resources also affect family relations. Conflict about financial affairs seems to be among the most important sources of discord in partnerships, and family violence has often been linked to the lack of financial resources. As to divorce, findings demonstrate a positive contribution of men's earnings to marital stability (e.g., Hoffman and Duncan, 1995). Benefits from women's higher earnings are less clear, because the stabilizing effect may be counteracted by women's higher economic independence that contributes to greater hazards of divorce ('independence effect'). However, the independence effect is relatively weak and may be restricted to marriages that are unlucky at the first place (Sayer and Bianchi, 2000; Oppenheimer, 1997). Furthermore, causality between women's income and marital quality/stability may be bidirectional: Marital discord was found not only to be affected by women's earnings but also to instigate women's higher labor force participation, which causes higher earnings (White and Rogers, 2000).

Attitudes toward Money

Attitudes toward money may be understood within a larger system of values that has been broadly distinguished as materialistic or postmaterialistic. Materialism is associated with greed for money which means that people strive for accumulating money beyond what is needed for survival, often at the expense of happiness (Dittmar, 2010; Kasser and Ahuvia, 2002). Based on Maslow's hierarchy of values, Inglehart (1977) claims that materialistic needs, related to stable economic resources and physical security, are basic and foregrounded in one's value system if they are not sufficiently fulfilled, specifically – coined in a socialization hypothesis – if they have been lacking during childhood and adolescence. Accordingly, situations that remind one of insecurity, such as awareness of death, increase materialism and conspicuous consumption (Arndt et al., 2004; Kasser and Sheldon, 2000). On the other hand, postmaterialistic values including political participation, freedom of speech, and self-realization will develop primarily if materialistic needs are satisfied. In line with these assumptions, research has shown that materialistic values are less foregrounded among the well educated and those with higher income, and particularly that post-materialistic values became increasing preferred among younger generations.

Interestingly, although greed and materialism may function to fulfill security needs on the level of the individual, they contribute to financial crises and thus financial insecurity on a global level: Materialistic orientations, short-dated bonuses for financial managers, and the greed of many investment bankers, who – often exceeding criminal thresholds – developed more and more innovative financial products that have been partly responsible for recent financial crises (Lenz et al., 2010). Here, money may function like a drug, such as consumption of cocaine or sexual activity; the prospect of profit activates the nucleus accumbens (i.e., the 'pleasure center' in the brain) and causes release of dopamine (Lea and Webley, 2006), which may be a sustaining factor for greed.

On a symbolical level, it is remarkable that money is associated with ambivalent characteristics and paradoxical meanings. Money is associated with positive values like security, independence, freedom, prestige, and possessions as well as with negative associations like anxiety, anger, greed, envy, conflict, and the 'pain of paying.' People who own much money are likely to be seen as hard working, successful, and even beneficent. But they are also at risk of being regarded as superior, selfish, capitalistic, and hostile. Given these complexities, research has focused much on understanding the basic structure of attitudes toward money.

Several different questionnaires have been constructed and psychometrically examined in order to investigate the fundamental dimensions underlying money attitudes (for an overview, see Furnham and Argyle, 1998). For example, the Money Attitude Scale (Yamauchi and Templer, 1982) differentiates between the factors power/prestige, retention time, distrust, quality and anxiety, whereas the Money Beliefs and Behavior Scale (Furnham, 1984) measures six factors: obsession, power/spending, retention, security/conservative, inadequate, and effort/ability. The short version of the Money Ethic Scale (Tang, 1995) summarizes three factors: success, budget, and evil. Materialism can be assessed with the Richins and Dawson's (1992) 18-item Material Values Scale. All different approaches to the measurement of money attitudes illustrate that the attitude structure is rather complex and multidimensional with clearly differentiable factors reflecting the ambivalence of money. Some of those factors are linked to clinical traits, such as anxiety or obsession, whereas others are related to the retention factor, such as saving and investing money. Other factors, such as power, prestige, and spending, show that money can also be a means to influence and impress others.

Affected by parental income, education, social class, beliefs, parenting practices, and money habits, money attitudes are formed early in childhood and remain fairly consistent throughout one's life. Empirical studies show that high-income people tend to think that money is less evil and rather reflects

one's achievement. Participants scoring high in protestant ethic are more likely to budget money properly and tend to see money as evil and representing power. High leisure-oriented individuals see money more frequently as good, representing achievement and power/freedom. In contrast to students with credit cards, those without are more likely to score higher on money attitudes of obsession and retention. Additionally, compared to normal consumers, compulsive spenders have been shown to have a relatively lower self-esteem, to believe more that one's self-esteem can be enhanced by money, and to be more convinced that money solves problems.

There are also attitudes toward banknotes and coins as the actual currency. Bruner and Goodman (1947) demonstrated the influence of subjective values and objective needs on the perception of currencies by asking rich and poor 10-year-old children to match circles of light corresponded to a range of coins. The results illustrate that the size of socially valued objects, such as coins, are judged larger than neutral objects, such as gray discs. Moreover, the greater the value of a coin, the more biased the estimation. This overestimation effect is stronger for poor than for rich children.

Thinking about Money

Recently, research in social psychology started to investigate what happens when people merely think about money. It has been found that such 'money priming' affects cognitions, emotions, and behavior. It has been shown, for instance, that priming with the idea of money changes cognitive processes (Hansen et al., 2013); mere reminders of money, such as pictures of money or thinking about the concept of money, trigger abstract (versus concrete) mental representations and an abstract style of thinking. For instance, participants who are primed with money represent actions more in terms of their abstract goals instead of their concrete means, they perceive the world in fewer and broader (instead of more and narrower) categories, and they are relatively quicker in identifying global pictures ('the forest') and relatively slower in identifying local details ('the trees'). This has also consequences for judgments and behavior; for instance, money primes cause a focus on central (instead of peripheral) aspects when evaluating objects. Furthermore, thinking about money also increases self-control (Boucher and Kofos, 2012), which is related to abstract thinking. Importantly, these effects only emerge when money is seen as a resource, so that the idea of money can prime security and personal strength. For people who associate money with the 'pain of paying' or with loss, reminders of money cause more concrete and detailed-oriented thinking styles.

Money priming has effects on social behavior, as well. For instance, it has been demonstrated that participants primed with the concept of money show a reduced tendency to request help with a difficult task (Vohs et al., 2006). Compared to participants of a control group, money-primed participants are also more reluctant to help other people. Additionally, money priming causes a preference to play and work alone, and increases the tendency to put more physical distance between oneself and a new acquaintance. These effects have been interpreted as result of money reminders signaling independence and self-sufficiency (Vohs et al., 2008); since money usually is associated with strength, it reminds one that people (including oneself) may have plenty of resources. Therefore, merely thinking about money causes the impression that other people and oneself need only few social resources.

Reminders of money also function as a pain buffer and affect emotions. For instance, thinking about money leads to experiencing less social distress and physical pain (Zhou et al., 2009). Although social support is the most important psychological factor that alleviates social distress and physical pain, reminders of money also function as pain buffer (Zhou and Gao, 2008). Conversely, social exclusion and monetary loss result in increased pain awareness. Additionally, research has demonstrated that anticipation of pain heightens the desire for social support as well as the desire for money.

Handling Money

Saving, Spending, and Investing Money

The decision to save money or to spend it for consumption is partly based on rational economic considerations. However, saving is also a socially accepted goal, formed during socialization and by norms of the social environment. Predicting people's saving behavior, economists focus on factors such as the increase of income, which leads to an overproportional increase of savings; social comparisons with other people, which influence habits of consumption and saving; and saving behavior as postponing consumption in order to guarantee a constant level of lifelong consumption (for an overview, see Webley and Nyhus, 2008). The framing approach of the 'behavioral life cycling' hypothesis by Shefrin and Thaler (1988) stresses the importance of individual self-control, self-imposed rules, and precommitments for explaining saving. Different forms of money, such as regular income, assets, or windfall profits, are subjectively categorized and assigned to different 'mental accounts.' The propensity to save all or part of an income is assumed to depend on which mental account the income is referred to.

Katona (1975) emphasizes the importance of differentiating between people's ability to consume or save and their willingness to do so. He has pointed out that people's beliefs about economic issues mediate the amount of money they save. People's consuming or saving behavior is based on expectations regarding the future development of the economic situation. The Index of Consumer Sentiment (Katona, 1975) predicts turning points in business cycles and expresses changes in optimism–pessimism about the financial future. It has been found that total saving increases when people become more uncertain and worried about the future. One-third of household saving in the United States is explained by the precautionary motive, in particular with regard to health and retirement considerations. Because saved money is often used as protection against economic risks, risk-averse individuals tend to have lower debts and save more money than risk-taking individuals.

Empirical studies (e.g., Furnham, 1985; Lunt and Livingstone, 1991; Traut-Mattausch and Jonas, 2011) show that age is linearly related to saving (the older the higher the savings), and that education has a curvilinear relation to saving (the most and least educated people save less, whereby the

more educated invest more of their money instead of saving it). Savers are more optimistic about economic circumstances and believe in personal control over finances, whereas nonsavers think of themselves more as victims of external events. However, they also seem to blame themselves more when things go wrong. Savers get more social support from friends and relatives than nonsavers and thus more information about how to handle finances. Savers show more consistent shopping habits, such as going to a few favorite shops, whereas nonsavers tend more to shop around.

Psychological research on financial investments in stock markets focuses on investigating the heuristics that people may apply when deciding how to invest their money (for overviews, see Gärling et al., 2009). Since financial decisions are highly influenced by the context in which price information is presented, salience and attention effects play a prominent role in financial decision making. For example, selective causal attributions provided by the media have a significant influence on investment decisions in financial markets (Andreassen, 1987); by explaining recent changes in stock prices with good news after prices rise and with bad news after prices fall, reporters tend to focus on the sets of facts that suggest that change should have happened and to omit the set of facts that suggest it should not have occurred. Such attributions foster expectations that changes will either persist or regress to previous price levels. This can mislead investors and result in wrong decisions in turn.

Tax Compliance and Evasion

About 20–30% of taxpayers in Western industrial nations do not declare their income completely. Approximately 10–20% of tax is illegally evaded. The majority of people think that their tax burden is too high and that they pay more taxes than other people and even more than those with a comparable income. The lack of transparency of the tax system and people's lack of knowledge about tax issues increase distortions in perception systematically (for overviews, see Jonas et al., 1999; Kirchler, 2007).

The economic approach suggests a very rational behavior of paying and evading taxes, which is only influenced by the benefits and costs associated with tax evasion, such as the subjectively perceived probability of detection, penalty rates, amount of income, and level of marginal tax rate (for an overview, see Andreoni et al., 1998). However, tax evasion is not only a matter of rational calculation and utility maximization but also of psychological variables such as an individual's willingness to evade or comply (*see* Social Dilemmas, Psychology of). Tax evasion is higher the more people experience money and profits in particular as an indicator of success and the more they perceive opportunities to do so. Negative attitudes toward the government, toward the fairness of the society's resource allocation, toward the legislation, and toward the spending behavior of political actors are also important in influencing tax evasion. Furthermore, tax evasion is lower the more important moral norms and religious values for a person and the higher his or her integration in the society. Finally, tax fairness, which is judged based on the comparison of tax burden and the government's provision of public goods as well as the tax burden of other taxpayers, is also an important predictor for tax evasion.

Donating Money

On the one hand, people are willing to take illegal actions to save money by evading taxes but, on the other hand, they voluntarily give money away by donating it to charities (*see* Altruism and Prosocial Behavior, Sociology of). Numbers of the Charities Aid Foundation show for Great Britain that more than two-thirds of the inhabitants donate money for charitable causes. However, the relationship between income and donations is quite low (only about $r = .13$; Furnham and Argyle, 1998). Considering the characteristics of people who donate money, it is found that people between the age of 50 and 65 years give the most; women give more money than men and are also more actively involved in volunteering. Married people give about 20–40% more than singles with a comparable income. Having children leads to an increased willingness to donate (about 50% more than people without children). People who are religious give about three times more than people to whom religion is not important. Concerning contextual factors, more money is given if the request comes from a physically attractive person, a similar person, or an expert. More money is also given in reciprocal relationships and to one's own group. Touching the potential donors and eye contact are effective strategies for raising donations. Furthermore, modeling and conformity have been shown to be crucial factors. Additionally, donating and spending money on other people have been shown to increase happiness; spending more of one's income on others than on oneself is associated with greater happiness, and participants who were randomly assigned to spend money on others experienced greater happiness than those assigned to spend money on themselves (Dunn et al., 2008).

When Money Runs Short: Perspectives on Poverty

In the past decades, issues of poverty have increasingly become addressed by social and developmental psychologists. Relevant research dates back to the 1930s when Jahoda et al. (1975, original 1938) introduced the ethnographic approach in studying a whole community that was hit by unemployment due to plant closing. Showing that different 'reaction types' (e.g., 'unbroken,' 'in despair') emerged as styles of coping with the increasingly desperate situation, early findings inspired later stress and coping research. Addressing family adaptations to economic loss, Elder (1974) showed that the families' adaptive reactions along with parents' stress during the Great Depression increased the likelihood of conflict and tension within families and shaped children's short- and long-term developments.

Present research focuses much on relative income poverty, defined by national poverty lines or by a per capita income of less than 50% of the average income. Different from former static approaches, more recent studies consider the duration of poverty and suggest, for instance, that stable poverty is experienced by only a small minority of children in the United States while a large share lives at least 2 years in poverty. In general, findings show that long-term poverty has more negative effects on well-being and children's behavior than short-term deprivation, but the latter is often detrimental, too.

Regarding effects of poverty, research pays increasing attention to children who have been pointed out as an important risk group of poverty in many countries, especially if living with a single parent, having two or more siblings, or belonging to ethnic minorities (e.g., Seccombe, 2000). Most studies report deleterious effects of poverty on both children's and adults' well-being and health, including lower life expectancy, higher infant mortality, and more acute and chronic health problems. Disadvantages often begin during pregnancy leading to low birth weight and undernourishment, thus increasing the risk for serious illness. In explaining these findings, impaired health among the poor has been linked to limited access to medical or prenatal care, to less healthy lifestyles, nutritional deficiencies, and low levels of education or literacy, making it difficult to use health services or follow health instructions. High levels of emotional stress have also been shown to negatively affect the immune system.

Furthermore, children living in poverty have more socio-emotional and behavior problems, have more difficulties in peer relationships, and do poorly in school. Much of these effects can be linked to parental stress, which results from the economic pressure they face and which spills over into the parent–child relationship leading to less nurturing, more authoritarian, and more inconsistent, harsh physical discipline. Hence, impaired parenting is important in mediating negative effects of income loss on children. On the other hand, individual and family resources (e.g., emotional stability and a harmonious marriage) may buffer the consequences of income loss for family life and children's development. Increased social support from family and friends has also been shown to reduce the stress and violence experienced by poor families, and to improve self-esteem. High rates of poverty in the neighborhood increase the likelihood of children's externalizing problems even beyond individual family income. Given the spatial segregation of poor families, problems arising from the lack of family and community resources are likely to cumulate.

Finally, it should be noted that running up debts is often used to compensate a lack of money. Despite the risk of serious financial problems due to long-term obligations, only few studies in psychology address such issues. Lea and Webley (1995) report that debtors have less money management facilities than nondebtors, rate their ability at money management lower, and think in shorter time horizons without considering long-term consequences. Debt counseling has become an important task in social services provided to the poor.

Conclusion and Outlook

From a psychological perspective, it is obvious that money is much more than just an economic exchange good. Although money is an important resource that affects many domains in life, much of its significance seems to derive from its symbolic value and from psychological processes, such as social comparison processes, or attitudes toward money, which are complex and often ambivalent. Similarly, determinants of handling money often go beyond economic rationale. Recent research includes more genuinely psychological perspectives (such as effects of money on cognition and social behavior) in studying the significance of money in people's life. Although we could not cover all topics that are related to money (for instance, for the psychology of incentives and motivation, *see* Pay, Compensation, and Performance, Psychology of and Work Motivation), the summarized topics demonstrate that the psychology of money and finances is a rich field in which research efforts are worthwhile. Further subjects related to money such as economic and financial policy, state indebtedness, ethics and moral issues when handling money, or the cultural history of money are interesting and – applied to psychology – provide a broad spectrum of promising avenues for future research.

See also: Behavioral Economics; Consumer Psychology; Corporate Taxation; Justice: Social Psychological Perspectives; Monetary Policy since the 2007–2009 Financial Crisis; Organizational Justice; Pay, Compensation, and Performance, Psychology of; Social Dilemmas, Psychology of; Wealth Distribution; Work Motivation.

Bibliography

Andreassen, P.B., 1987. On the social psychology of the stock market: aggregate attributional effects and the regressiveness of prediction. Journal of Personality and Social Psychology 53, 490–496.

Andreoni, J., Erard, B., Feinstein, J., 1998. Tax compliance. Journal of Economic Literature 36, 818–860.

Arndt, J., Solomon, S., Kasser, T., Sheldon, K.M., 2004. The urge to splurge: a terror management account of materialism and consumer behavior. Journal of Consumer Psychology 14, 198–212.

Boucher, H., Kofos, M.N., 2012. The idea of money counteracts ego depletion effects. Journal of Experimental Social Psychology 48, 804–810.

Bruner, J., Goodman, C., 1947. Value and need as organizing factors in perception. Journal of Abnormal and Social Psychology 42, 33–44.

Coltrane, S., 2000. Research on household labor: modeling and measuring the social embeddedness of routine family work. Journal of Marriage and the Family 62, 1208–1233.

Diener, E., Diener, M., Diener, C., 1995. Factors predicting the subjective well-being of nations. Journal of Personality and Social Psychology 69, 851–864.

Diener, E., Seligman, M.E.P., 2004. Beyond money: toward an economy of well-being. Psychological Science in the Public Interest 5, 1–31.

Dittmar, H., 2010. Consumer Culture, Identity and Well-being: The Search for the 'Good Life' and the 'Body Perfect'. Psychology Press, New York.

Dunn, E.W., Aknin, L.B., Norton, M.I., 2008. Spending money on others promotes happiness. Science 319, 1687–1688.

Easterlin, R., 2003. Building a Better Theory of Well-being. Discussion Paper No. 742. IZA, Bonn, Germany.

Elder Jr., G.H., 1974. Children of the Great Depression. Westview Press, Chicago.

Furnham, A., 1984. Many sides of the coin: the psychology of money usage. Personality and Individual Differences 5, 50–509.

Furnham, A., 1985. Why do people save? Journal of Applied Social Psychology 15, 354–373.

Furnham, A., Argyle, M., 1998. The Psychology of Money. Routledge, London.

Gärling, T., Kirchler, E., Lewis, A., van Raaij, F., 2009. Psychology, financial decision making, and financial crisis. Psychological Science in the Public Interest 10, 1–47.

Hansen, J., Kutzner, F., Wänke, M., 2013. Money and thinking: reminders of money trigger abstract construal and shape consumer judgments. Journal of Consumer Research 39, 1154–1166.

Hoffman, S.D., Duncan, G.J., 1995. The effects of income, wages, and AFDC benefits on marital disruption. Journal of Human Resources 30, 19–42.

Inglehart, R., 1977. The Silent Revolution. Changing Values and Political Styles among Western Republics. Princeton University Press, Princeton.

Jahoda, M., Lazarsfeld, P.F., Zeisel, H., 1975. Die Arbeitslosen von Marienthal. Ein soziographischer Versuch (The unemployed of Marienthal. A socio-graphic test). Suhrkamp, Frankfurt (Orig. 1938).

Jonas, E., Heine, K., Frey, D., 1999. Ein Modell der Steuerzufriedenheit – psychologische Grundlagen (un)ökonomischen Handelns (A model of tax contentedness – psychological basis of (un)economic behavior). In: Fischer, L., Kutsch, T., Stephan, E. (Eds.), Finanzpsychologie. Oldenburg, München, pp. 160–187.

Kasser, T., Ahuvia, A., 2002. Materialistic values and well-being in business students. European Journal of Social Psychology 32, 137–146.

Kasser, T., Sheldon, K.M., 2000. Of wealth and death: materialism, mortality salience, and consumption behavior. Psychological Science 11, 348–351.

Katona, G., 1975. Psychological Economics. Elsevier, New York.

Kirchler, E., 2007. The Economic Psychology of Tax Behavior. Cambridge University Press, Cambridge.

Lea, S.E.G., Webley, P., 1995. Psychological factors in consumer debt: money management, economic socialisation, and credit use. Journal of Economic Psychology 16, 681–701.

Lea, S.E.G., Webley, P., 2006. Money as tool, money as drug: the biological psychology of a strong incentive. Behavioral and Brain Sciences 29, 161–209.

Lenz, A.M., Frey, D., Rosenstiel, L.v., 2010. Schöpferische Zerstörung und zerstörerische Schöpfung – wie Finanzinnovationen wesentlich zur internationalen Finanzkrise beitrugen (Inventive destruction and destructive invention – how financial innovations fundamentally contributed to the international financial crisis). In: Oerter, R., Frey, D., Mandl, H., Rosenstiel, L.v. (Eds.), Neue Wege wagen: Innovation in Bildung, Wirtschaft und Gesellschaft. Lucius & Lucius, Stuttgart.

Lunt, P., Livingstone, S.M., 1991. Psychological, social and economic determinants of saving: comparing recurrent and total savings. Journal of Economic Psychology 12, 621–641.

Oppenheimer, V.K., 1997. Women's employment and the gain to marriage: the specialization and trading model. Annual Reviews of Sociology 23, 431–453.

Richins, M.L., Dawson, S., 1992. A consumer values orientation for materialism and its measurement: scale development and validation. Journal of Consumer Research 19, 303–316.

Sayer, L.C., Bianchi, S.M., 2000. Women's economic independence and the probability of divorce. Journal of Family Issues 21, 906–943.

Seccombe, K., 2000. Families in poverty in the 1990s: trends, causes, consequences, and lessons learned. Journal of Marriage and the Family 62, 1094–1113.

Shefrin, H., Thaler, R., 1988. The behavioral life-cycle hypothesis. Economic Inquiry 26, 609–643.

Tang, T.L.-P., 1995. The development of a short money ethics scale: attitudes toward money and pay satisfaction revisited. Personality and Individual Differences 19, 809–816.

Traut-Mattausch, E., Jonas, E., 2011. Why do people save? The influence of financial satisfaction and income on saving. Journal of Psychology 219, 246–252.

Vohs, K.D., Mead, N.L., Goode, M.R., 2006. The psychological consequences of money. Science 314, 1154–1156.

Vohs, K.D., Mead, N.L., Goode, M.R., 2008. Merely activating the concept of money changes personal and interpersonal behavior. Current Directions in Psychological Science 17, 208–212.

Webley, P., Nyhus, E., 2008. Inter-temporal choice and self-control: saving and borrowing. In: Alan, L. (Ed.), The Cambridge Handbook of Psychology and Economic Behaviour. Cambridge University Press, Cambridge, pp. 105–131.

White, L., Rogers, S.J., 2000. Economic circumstances and family outcomes: a review of the 1990s. Journal of Marriage and the Family 62, 1035–1051.

Yamauchi, K.T., Templer, D.J., 1982. The development of a money attitude scale. Journal of Personality Assessment 46, 522–528.

Zhou, X., Gao, D.G., 2008. Social support and money as pain management mechanisms. Psychological Inquiry 19, 127–144.

Zhou, X., Vohs, K.D., Baumeister, R.F., 2009. The symbolic power of money: reminders of money alter social distress and physical pain. Psychological Science 20, 700–706.

Money, Sociology of

Frederick F Wherry, Yale University, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by V.A Zelizer, volume 15, pp. 9991–9995, © 2001, Elsevier Ltd.

Abstract

From premodern societies to the present, money has facilitated the trade of goods, services, and persons. With the rise of the modern nation-state came attempts to affirm national identity through money and to standardize monetary currencies so the state could pay for its wars. These political-economy explanations, however, obscured the other forms and functions that money has taken (and continues to take) in society. Even when money does not appear to take on heterogeneous forms of legal tender, it is accounted for and earmarked by different groups as if it were a multitude of entities. By studying the variation in what money looks like, how it is used, and how its meanings vary, sociological studies capture the cultural codes and multiple identities that constitute money and the marketplace.

What is money and why do sociologists study it? What forms has money taken over time, and how have the social meanings of money affected the purposes to which it has been put? Sociologists argue that the form that money takes depends on the societies producing it. The size, color, symbols, emotional salience, and restrictions on money do not arise out of a technical feasibility assessment but rather from the meaningful engagements of people, institutions, and their polity over how to exchange the rights to a good, how to create restricted versus generalized processes for transferring goods and services, and how to accumulate value that can be deployed in the future. The more that an object (or a virtual entity) fulfills these three functions – (1) enabling *the rights to* goods and services be transferred from one party to another, (2) legitimizing the general use of an exchange medium between a variety of parties or sites, and (3) facilitating value storage for later deployment – the more money-like it is (Zelizer, 2001). An exchange medium that does not fulfill all three of these criteria may nonetheless be money, just less fully so. (see Maurer, 2006 on the anthropology of money.)

Sociologists studying the degree to which an entity functions as money and the varieties of forms it takes draw from the lessons of history, ethnography, interviews, experiments, and surveys. We will begin with history to document just how many varieties of money have existed along with the social, political, and cultural situations that brought these monies into being. These studies help us understand that what we take as money, as value, and as a sensible way to engage in economic transactions could have been structured otherwise – and might still.

The Shift from Premodern to Modern Money

In premodern times, when people purchased slaves in Monomotapa in the Gulf of Guinea, they used cotton cloth as currency and referred to these swaths of cloth as 'piece[s] of India' (Braudel, 1981: p. 442). Not only cloth but also rocks, corral, nails, and furs served to store value and to facilitate the transfer of the rights to use objects or services. In the Trobriand Islands that Bronislaw Malinowski studied, women used different forms of currency than men did for different kinds of trades.

In the West, money itself reflected the social differentiations of class and status in society. Writing about money in the fifteenth to the eighteenth centuries, historian Fernand Braudel demonstrated how different categories of people were matched to different categories of money: "[E]very metal played its part: gold, reserved for princes, large merchants (even the Church); silver for ordinary transactions; copper naturally for the smallest. Copper was the 'black' money of people of small means and the poor. Mixed with a little silver it blackened quickly and deserved its name" (Braudel, 1981: p. 458). This use of money as a means of communicating social identity was not new. In a study of how money was used in Roman society between 200 BCE and 200 CE, Michael Crawford recalls how a person's relationship with money affected his dignity and reputation. As a result, Tiberius intervened in the dispute between the moneylenders and members of the upper classes in order to preserve their social status and to maintain the social order. Crawford writes:

> [I]n Rome, on the only known occasion on which money was supplied at public demand, it was supplied not for economic reasons, but, as we might expect, in order to preserve social status. In A.D. 33 the moneylenders of Rome, accused of irregularities, tried to turn the tables on their accusers and attempted to call in all their debts. The debtors were threatened with the necessity of selling off their land on a rapidly falling market to the evident profit of the usurers. But Tiberius intervened with an interest-free loan of 100 000 000 sesterces to those whose *dignitas* and *fama* were threatened…. On occasions of what appear to have been general currency shortages in 63, 49 and 44 BC no comparable intervention took place. (Crawford, 1970: p. 46)

Money functioned as a communicative instrument of social status and as a social mechanism for stability. In this instance, the social status of a particular category of individuals (not society's economic well-being) was threatened and therefore warranted an intervention in the money market. Had a different category of individuals been threatened by the moneylenders, the informal rules about how money should be used, against whom an economic threat could be made, and from whom the rule's enforcement would come would have differed.

Just as different categories of persons are matched with different categories of money, so too are different categories of money matched with varying rules of use. Crawford notes that some coins that held more economic value were nonetheless less valuable by virtue of the image on the coin. The Roman historian Arrian wrote that the coin bearing the image of Nero held more economic value than the one bearing the image of Trajan; however, the Nero coin was considered morally rotten, but the Trajan one fine. In other words, money was more than a means to an end; it was an end in itself, encompassing a set of value judgments. Money carried images and meanings that ought not be defiled. Indeed, Tiberius considered it a capital offense for a coin bearing the image of Augustus to be taken into a brothel or lavatory precisely because some types of money must be protected from defilement or moral pollution (see Crawford, 1970: p. 47).

With the rise of the modern nation-state, the variations in money diminished. The United States just after the Civil War, for example, faced both a practical and a symbolic problem. Before the Civil War, states and substate institutions issued their own currencies and scripts. With more than 5000 varieties of money in the United States, the polity seemed anything but united (Zelizer, 1994). The debts amassed during the Civil War generated a practical problem: with incommensurable currencies, how could the central governing institution pay its debts? And the struggles over unity exacerbated a symbolic problem: how do people's daily transactions remind them that they belong not simply to a state or to a regional body (such as the former Confederacy) but rather to a national community?

Political scientist Eric Helleiner uses an 1863 letter from S.M. Clark, the chief clerk of the US Treasury, addressed to Secretary Salmon Chase, in order to illustrate how the producers of currency intended to disseminate nationalist images through the circulation of money in the marketplace. The letter reads:

> [The money] would tend to teach the masses the prominent periods in our nation's history. The laboring man who should receive every Saturday night, a copy of 'Surrender of Burgoyne' for his weekly wages, would soon inquire who General Burgoyne was, and to whom he surrendered. This curiosity would be aroused and he would learn the facts from a fellow laborer or from his employer. The same would be true of other National pictures, and in time many would be taught leading incidents in our country's history, so that they would soon be familiar to those who would never read them in books, teaching them history and imbuing them with a National feeling. (Helleiner, 1998: p. 1412)

In other words, because money is widely circulated voluntarily by people who happen to use the market for procuring their needs, the state does not have to force a sense of national feeling. The national culture would cohere by virtue of consumption and economic exchange. As individuals engage in their routines and move through their daily rounds, they encounter monies and the images carried thereon. In this way, it is easy to understand why Fernand Braudel referred to money as one of the elementary 'structures of everyday life' (Helleiner, 1998: p. 1413).

Money is such an elementary form of everyday life that it is often taken for granted, and in its taken-for-grantedness, it is able to constitute meaning structures that influence the course of social action. Sociologists Bruce Carruthers and Sarah Babb (1996) use the historical case of money after the Civil War in the United States to demonstrate the extent to which the implicit expectations people have about money and its uses gets challenged when the medium of exchange itself becomes ambiguous and unsettled. The expenses of the Civil War were so great that the national government had to abandon the gold standard for fiat money. A variety of monies presented themselves: "national bank notes, state bank notes, greenbacks, currency convertible into specie, and specie coin" (Carruthers and Babb, 1996: p. 1557). Two camps – the bullionists (advocates of 'hard money') and the greenbackers (advocates of 'soft money') – emerged to contest the form that money should take in the United States. For the former, money had an intrinsic value, and its holders were honest Christians (not pagan shape-shifters). For the latter, money was a democratic creation subject to democratic control. Both sides argued that the type of money the state issued spoke to the underlying character and identity of the nation. As money speaks for character and identity, it moves beyond its economic function: "If money served only an economic purpose, it is unlikely that it would have been traditionally organized largely along national lines…. National currencies, and money more generally, need to be examined not just as an economic phenomenon but also in terms of their geographic, political, social and cultural dynamics" (Gilbert and Helleiner, 1999: pp. 1–2).

Simone Polillo continues the theme of identity, character, and money as he examines how financial elites in the United States (*c.* 1800–1913) and Italy (*c.* 1860–1913) have shaped collective identities in order to control new financial instruments. He identifies a binary between the conservatives and the wildcats, with the former maintaining barriers to credit for those lacking it and the latter generating new financial instruments and engendering new financial actors to enter the field. The financial instruments varied by their exclusivity (restrictions on who could possess them) and by the degree of control those possessing the instrument had.

The increase in the types of financial instruments reflected a proliferation of monies and corresponding practices: from hedge funds to exclusive credit cards; from relationship banking to mortgage backed securities. Each type of financial instrument entered into specific circuits of exchange, infused with symbols of membership and relationship categories.

> Much like the men of early-twentieth-century America analyzed by Viviana Zelizer (1994), who earmarked a significant portion of their earnings to spend it on rituals of sociability with other men (such as drinking together), financial elites earmark the most prestigious financial instruments to "spend" them in financial interactions with other financial elites, in the activity of trading that allows them to be at the center of vibrant financial markets. (Polillo, 2013: p. 7)

Polillo's findings reject the notion of money as fungible, because the financial instruments are effective to the extent that they are not fungible; moreover, the experience of Michael Milken's junk bond debacle illustrates the consequences of mismatching categories of money with inappropriate categories of practice.

The theme of money and national identity, most strongly linked in Polillo's analysis of Italy, also plays out in France. There, Helleiner notes "the growing use of national currency among the rural poor of France … as one part of a broader transformation that transformed 'peasants into Frenchmen' in the late nineteenth century. In his words, the new standardized national currency became "a universal language that all understood and all now wanted to speak" (quoted in Helleiner, 1998: p. 1415). In other words, the currency served as a great leveler, dulling rural–urban distinctions and making more salient the money individuals used as an indication of the nation to which each belonged.

Helleiner notes that currencies are utilized in religious-like interactions as people place their trust (an act of faith) and declare their devotion to a particular currency form. Adam Muller, a conservative Prussian writing in the early nineteenth century,

> believed that "the state's duty is to awaken national pride, the feeling of 'oneness' with the national state in the economic sphere" … A national currency was seen as a key tool in this project.… But this bond was not in terms of the issues of imagery, communication, the community of fate, or sovereignty.… Rather, it was at a more emotional level. (Helleiner, 1998: p. 1427)

The importance of iconography and the meanings of money do not disappear with the increasing differentiation of economic life or with the growing integration of modern economies. In his introduction to *The Year of the Euro* (2006), Robert Fishman examines the struggles over national identity in 2002 over the rise of a single European currency:

> The scenes of euro enthusiasts lining up beside automatic teller machines to withdraw new euro banknotes shortly after midnight on the changeover date, the confusion and petty disputes over currency conversion in some commercial establishments, and indeed all the many stories and experiences constituting the year of the euro stand as an enormous collective … experience spanning twelve European countries. (Fishman, 2006: p. 8)

The political project of national identity and of supranational organization relies on the "iconography … the values, themes, and concepts represented by artistic motifs [on the money] … [assist] in the construction of a European identity" (Hymans, 2006: p. 15). The nation-state tries to increase their legitimacy by using the currency to signal their embrace of values in tune with the 'spirit of the times' (Hymans, 2004: p. 6). Likewise, Nigel Dodd (1994) uses the symbols on money and reactions to their proposed alteration as evidence of the ongoing national identity work that money facilitates.

The European Central Bank remains a site for research in Carlo Tognato's (2012) investigations of money as a symbolic medium. Central bankers contest currencies and their associated territories as sacred or profane, granting monies character roles in a moral drama. Ministers of finance in euro member countries use religious imagery when talking about the order that the euro engenders and to warn against the chaos other currencies might unleash. Tognato identifies a range of ritual practices, public prayers, and cultural codes that constitute the macroeconomy of the European Union. Likewise, Bill Maurer (2002) identifies the religious underpinnings of financial derivatives.

Modern money and the cultural codes they depend on operate in discussions of taxation as well as discussions about consumer spending. Recent work on symbolic interaction identifies how groups of people in the southern United States use narratives about taxation to engage in boundary work (Kidder and Martin, 2012) – differentiating themselves as the exploited versus the exploiters, as creating (making) versus destroying (taking). Indeed, talk about taxes resembles what Wherry (2008) identifies in talk about consumer purchases: there are narratives that enable people to engage in boundary work, affirming those who are near to or far from a societal ideal. These discussions lead us to discussions of household finance, where we now turn.

Household Monies, Household Finance

Economic sociologists along with behavioral economists and psychologists have studied what money means and how households manage it. These different disciplines find that people do not behave as if money is fungible when arranging their household budgets. Zelizer contrasts her approach to earmarking (what she calls relational accounting) from the focus of psychologists as an individual's process of partitioning and evaluating budget categories:

> The distinction between individual and relational accounting is more than rhetorical: in the first view we might expect a given person's array of special monies to display a lot of stability and to vary considerably among persons as a function of life history; in the second, we might expect otherwise disparate people who are closely connected to have similar arrays of special monies. Those expectations differ significantly. However, because life histories and social relations intertwine, we should expect difficulty, at times, in distinguishing sharply between both types of accounting. (Zelizer, 2012: p. 160)

It is the way that individuals come together to form relationships and the ritual practices they engage in that matter crucially for how they engage in accounting. Zelizer's historical studies point to legal tender stored in a cookie tin that acquired a nickname and got allocated to specific budget categories. Money earned by a child got allocated differently from money earned by the female head of a two-parent heterosexual household. More recent studies acknowledge that people sometimes store money in different envelopes or utilize multiple bank accounts. In other words, the notion that money is just a means to facilitate market exchanges has been challenged by the regular practice of 'primitivizing' money – restricting its use by virtue of the identity of who earned it, the social relationships they are managing, and the purposes to which it will be put.

We can see the different approach that sociologists take to partitioning money, when compared with psychologists, in the example of the Christmas Savings Club. Psychologists study these clubs as self-control and commitment devices for individuals who cannot, otherwise, resist temptations to spend

rather than save. These savings clubs usually do not pay interest on the deposits, so from the point of view of a savvy budget manager, it would make more sense to put one's money into an instrument that allows profit generation. We can therefore look at the lost opportunity to profit as the cost of self-control.

Some economic sociologists tell a different story that complements this story of self-control. Viviana Zelizer reminds us that many working class women used these savings clubs in the early to mid-1900s. These clubs enabled these women to guard their holiday monies from other household members who might have compelled early expenditure. This was especially the case with their husbands. Zelizer writes:

> Indeed, the Christmas money was often reserved for domestic necessities such as a washing machine or a daughter's new coat…. the segregated Christmas money spared them from what was often considered the humiliating need to extract gift money from their wage-earning husbands. Thus, not only individual self-control but also negotiated household relations accounted for at least some of the Christmas clubs' great success in the United States. (Zelizer, 2012: p. 160)

The Christmas Savings Club taps into an existing ritual and set of practices that have become formalized. If it were not Christmas or another ritual well enough regarded, the club would not have the same power in keeping the money of wives off limits to their husbands. Christmas and its associated meanings probably helped these women challenge their husband's authority without threatening the relational roles of their husbands as the heads of their households. For love of ritual and the social relationships sustained thereby, people mark and manage their money.

Viviana Zelizer treats as different species of money the earmarks and nicknames given to budgeting categories. She uses cultural codes about family and gender and the variation in these codes for working-class women (compared with upper-class women) to explain what kinds of budgeting categories and practices households practiced at the turn of the twentieth century. The names granted to money varied along the lines of "allowance, pin money, 'egg money,' 'butter money,' spending money, pocket money, gift, or 'dole'" (Zelizer, 1989: p. 344). Following monetary practices enabled Zelizer to understand intrahousehold relations and interclass differentiations. Money itself served as a means for cultural communication.

Margarita Mooney (2003) and others have extended these insights to immigrant remittances. Households do not manage remittances simply according to what the needs of household members are but rather according to the identity of the person sending the monies home. If a household received remittances from a female household member (especially if that person is the mother), the monies tended to be earmarked for school fees, school uniforms, groceries, and clothes. By contrast, if a husband sent home earnings, they often earmarked the spending for larger, more expensive items (e.g., appliances, automobiles, home improvements). Treating money as if it is fungible and devoid of social meanings follows the fiction of a modern currency freed from its premodern functions. (For the social meanings of money in Africa, Asia, and Latin America, see Singh, 2013.)

Money in Theory

While Emile Durkheim had little to say about money, his nephew Marcel Mauss offered a cultural approach to money that took into account the interaction partners and other background conditions that make money effective as a medium of exchange. In *The Gift* ([1950] 2000), Mauss argued that an individual's gift is vested with her spirit, and this inalienable quality of the gift compels the recipient to reciprocate in kind. These objects being exchanged are attractive not because of the money value (price) but rather because the individuals involved are moral agents responding to moral stimuli. Objects used as money *enabled these moral claims to be affirmed* by the givers and recipients of gifts. These insights about the role of money in the exchange of gifts, as well as the implications of money for constituting the category of gift (rather than a repayment), have been taken up in Pierre Bourdieu's *Outline of a Theory of Practice* (1977). Bourdieu, however, focuses on the timing of the exchange than on money per se.

Karl Marx described money as a common denominator, a 'radical leveler' that "does away with all distinctions associated with traditional social relations" (Marx, 1974: p. 32). In *Structures of Everyday Life*, Fernand Braudel proffers a Marxist understanding of money as detrimental to intimate social relationships; the emergence of money brought "[s]harp variations in prices of essential foodstuffs; incomprehensible relationships in which man no longer recognized either himself, his customs or his ancient values" (Braudel, 1981: p. 436). Lamentably, labor became alienated from laborers and the individual's work became commodified, just as the worker herself or himself became 'a thing.'

In *The Philosophy of Money*, Georg Simmel notes that money is free to move across social sites without regard for social class, caste, or tribe. According to Helleiner, Marx and Simmel (2004 [1907]) "implied that, through its ability to reduce everything in the nation to a common 'language,' a national currency enabled citizens not just to communicate economically but also to think in a similar way and to situate themselves in a common fashion within the national community" (Helleiner, 1998: pp. 1417–1418).

Talcott Parsons ([1951] 1966) posits that money is indeed a language, a communication system, and that the economy is a means for society to adapt to its environment. And in Parsons and Smelser's *Economy and Society* ([1956] 2005) they provide a functionalist account of budgeting. Jens Beckert revisits this account in his discussion of the multiple functions of money and budgeting:

> Through spending on consumption, the social system 'household' fulfills institutionalized demands to live according to a certain lifestyle. Expenditure levels are determined by a culturally defined basket of goods that determines the 'cultural survival' (Parsons and Smelser, 1956: p. 221) of the household. Beyond that, expenditures for entertainment, leisure, and vacations are given a function for intra-family conflict management and, with that, status symbols assume an integrative function by positioning the household in relation to other households and thus symbolizing membership in a group. (Beckert, 2006: p. 171)

Household budgeting is both a positional and relational act, but the goal is not only to gain status but also to affirm

solidarity with a group; it is to rally around a totem. And these totems and their associated imaginings of what the future should hold affect budgeting decisions.

In *The Working Class and the Standards of Living* (1913), Maurice Halbwachs adopted Durkheim's theory of collective representations to explain the budgeting practices of families. Combining data from the Imperial Office of Statistics with that from the Union of Metal Workers, Halbwachs discovered that a future image (a representation) collectively held within a group affected (1) how families classified the categories in their budgets, (2) what percentage of their wage workers paid into each category, and (3) how budget categories were prioritized. Halbwachs did not argue against the concept of fungibility, as Zelizer (1994, 2010) does, yet he attends to collective representations (what Zelizer calls cultural codes).

Conclusion

Studying money facilitates nuanced understandings of social life. People establish, affirm, and contest national and regional identities. People erect boundaries between self and others, insiders and outsiders, but in doing so, they short-circuit purely arithmetical forms of evaluation. It is not useful for policy makers to hone in so-called bad decisions while ignoring the meaningful rituals and the overarching meanings that render strategies of economic action salient for the situation at hand. It is counterproductive to flatten money as a technology of exchange rather than a meaningful instrument that groups use to communicate who they are and what they hold sacred.

See also: Capitalism: Global; Consumption, Sociology of; Culture, Sociology of; Economic Sociology; Exchange: Social; Nation, Sociological Aspects of; Social Identity in Sociology; Work, Sociology of.

Bibliography

Beckert, Jens, 2006. Interpenetration versus embeddedness: the premature dismissal of Talcott Parsons in the new economic sociology. American Journal of Economics and Sociology 65 (1), 161–188.

Bourdieu, Pierre, 1977. Outline of a Theory of Practice. Cambridge University Press, Cambridge.

Braudel, Fernand, 1981. Civilization and Capitalism, 15th to 18th Century. Harper and Row, New York.

Carruthers, Bruce G., Babb, Sarah, 1996. The color of money and the nature of value: greenbacks and gold in postbellum America. American Journal of Sociology 101, 1556–1591.

Crawford, Michael, 1970. Money and exchange in the Roman world. Journal of Roman Studies 60, 40–48.

Dodd, Nigel, 1994. The Sociology of Money. Polity, Cambridge, MA.

Fishman, Robert M., 2006. Introduction. In: Fishman, Robert M., Messina, Anthony M. (Eds.), The Year of the Euro: The Cultural, Social, and Political Import of Europe's Common Currency. University of Notre Dame Press, Notre Dame, IN, pp. 1–14.

Gilbert, Emily, Helleiner, Eric, 1999. Introduction–nation-states and money: historical contexts, interdisciplinary perspectives. In: Gilbert, Emily, Helleiner., Eric (Eds.), Nation-States and Money: The Past, Present and Future of National Currencies. Routledge, New York, pp. 1–20.

Halbwachs, Maurice, 1913. The Working Class and the Standards of Living. La classe ouvriére et les niveaux de vie: Recherches sur la hiérarchie des besoins dans les sociétés industrielles contemporaines. Alcan, Paris.

Helleiner, Eric, 1998. National currencies and national identities. American Behavioral Scientist 41, 1409–1436.

Hymans, Jacques E.C., 2004. The changing color of money: European currency iconography and collective identity. European Journal of International Relations 10, 5–31.

Hymans, Jacques E.C., 2006. Money for Mars? The euro banknotes and European identity. In: Fishman, Robert M., Messina, Anthon M. (Eds.), The Year of the Euro: The Cultural, Social, and Political Import of Europe's Common Currency. University of Notre Dame Press, Notre Dame, IN, pp. 15–36.

Kidder, Jeffrey L., Martin, Isaac William, 2012. What we talk about when we talk about taxes. Symbolic Interaction 35, 123–145.

Marx, Karl, 1974. Capital. Lawrence and Wishart, London.

Maurer, Bill, 2002. Repressed futures: financial derivatives' theological unconscious. Economy and Society 31, 15–36.

Maurer, Bill, 2006. The anthropology of money. Annual Review of Anthropology 35, 15–36.

Mauss, Marcel, 2000 [1950]. The Gift: The Form and Reason for Exchange in Archaic Societies. W.W. Norton, New York.

Mooney, M., 2003. Migrants social ties in the U.S. and investment in Mexico. Social Forces 49, 1455–1470.

Parsons, Talcott, 1966 [1951]. The Social System. The Free Press, New York.

Parsons, Talcott, Smelser, Neil J., 2005 [1956]. Economy and Society: A Study in the Integration of Economic and Social Theory. Taylor & Francis, New York.

Polillo, Simone, 2013. Conservatives vs. Wildcats: A Sociology of Financial Conflict. Stanford University Press, Stanford.

Simmel, Georg, 2004 [1907]. The Sociology of Money. Taylor & Francis, New York.

Singh, Supriya, 2013. Globalization and Money: A Global South Perspective. Rowman & Littlefield, Lanham, MD.

Tognato, Carlo, 2012. Central Bank Independence: Cultural Codes and Symbolic Performance. Palgrave-Macmillan, New York.

Wherry, Frederick F., 2008. The social characterizations of price: the fool, the faithful, the frivolous, and the frugal. Sociological Theory 26, 363–379.

Zelizer, Viviana, 1989. The social meaning of money: 'special monies'. American Journal of Sociology 95, 342–377.

Zelizer, Viviana, 1994. The Social Meaning of Money. Basic Books, New York.

Zelizer, Viviana, 2001. Circuits of commerce. In: Alexander, Jeffrey, Marx, Gary T., Williams., Christine (Eds.), Self, Social Structure, and Beliefs: Explorations in the Sociological Thought of Neil Smelser. University of California Press, Berkeley, pp. 122–144.

Zelizer, Viviana, 2010. Economic Lives: How Culture Shapes the Economy. Princeton University Press, Princeton.

Zelizer, Viviana, 2012. How I became a relational economic sociologist and what does that mean? Politics & Society 40 (2), 145–174.

Money: Anthropological Aspects

Bill Maurer, University of California Irvine, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Weatherford, volume 15, pp. 9988–9991, © 2001, Elsevier Ltd.

Abstract

From its traditional concern with 'primitive' moneys and definitional quandaries over the nature of money in small-scale and nonmarket societies, anthropology more recently has considered money's meanings in multiple, overlapping spatial, cultural, and temporal milieu, as well as the pragmatics of money – what it does and how it signifies, rather than what it is or what it signifies. Early twenty-first century business interests in divesting the state of its monopoly over the legitimate means of exchange open new frontiers for the anthropology of money and value.

Early ethnologists were stunned by the range of things that seemed to be used as currency, from shells to braids of tobacco to giant, nearly immovable stones. Over the course of the twentieth century, several anthropological positions on money crystallized: (1) that 'primitive' is to be distinguished from 'modern' money in that the former is special purpose, employed only for exchanges involving certain things or transactions, or in specific, bounded domains of social life (say, marriages, death, or religious rituals), while the latter is general purpose, able to be employed for all manner of things and transactions, and across domains; (2) that special purpose moneys can be used for each of the Aristotelian 'functions' of money, different objects serving as means of exchange, method of payment, store of value, and measure of value or unit of account; (3) that, therefore, modern money is a universal or general equivalent, after Marx, and that it became the measuring rod against which all things and transactions can be evaluated; and (4) the colonial experience worldwide was one of the supplanting of local, special purpose moneys by modern, general purpose money that thereby; (5) deracinated and flattened objects' and relationships' meanings, purposes, and modes of evaluation by replacing them with one mode of value tied to the market logic of supply, demand, and price. By the 1980s, anthropology in most quarters had explicitly or implicitly essentially adopted an Aristotelian–Marxian–Simmelian perspective on money: as fulfilling the classic functions identified by Aristotle, as commensurating incommensurabilities and serving as a general equivalent as discussed by Marx, and as eroding 'social' or familial connection with the flat wash of anonymous money as argued by Simmel.

In the late twentieth century, however, money itself underwent several important transformations. First, many of the core European countries adopted a single currency, the Euro. The breakup of the Soviet Union and fragmentation of the Eastern Bloc countries into their constituent units led to the creation of new currencies, as well as the extension of the Eurozone eastward to new regions. Several countries adopted the US dollar as their official currency; others are informally dollarized; while the rise of China led commentators and everyday folk alike to question the US dollar's status as the global reserve and standard of value. Aside from their political and economic aspects, these interrelated phenomena spotlighted the relationship between money and sovereignty, money not just as a state project but as a symbol of the nation. Anthropologists explored this symbolism in a unifying Europe (Peebles, 2011) as well as in a former socialist world adapting to a market-based global economic order (Truitt, 2013; Lemon, 1998).

Second, informal flows of money from the global North to the global South reconfigured social and political relationships while creating the material conditions for transnationally distributed extended families. Coming into the twenty-first century, South–South remittance flows have become significant, reconfiguring political, familial, and financial relationships. Anthropological studies of remittances emphasized money's role in creating and sundering relations while serving as a means of communicative action (Hernandez and Coutin, 2006; Pedersen, 2013) and creating new collective identities. This work echoed earlier research suggesting that money mediates different temporal cycles, short-term exchanges in a marketplace vs longer term, durable and transgenerational relationships (Parry and Bloch, 1989).

Third, global financialization – the rise of finance as an industry in its own right rather than in the service of raising capital for productive enterprise – together with the use of market principles for governance, led anthropologists to look at the speculative aspects of money and credit. Marxist stories of abstraction, commensuration, and money as fiction excavated from informants' stories around the world led anthropologists to explore the spectral, fantastic, and magical properties of money. If Bohannan's interlocutors in the early twentieth century were struck by the capacity of modern money to equilibrate values and violate bounded spheres of exchange, the Comaroffs' were more spell struck, according to money magical powers wielded by the undead or witches (Bohannan, 1959; Comaroff and Comaroff, 1999). These themes were familiar to anthropologists whose earlier work had recorded peasants' explanation of the reproduction of money in terms of the intercession of the gods or devils (Taussig, 1980). Yet attention to credit and finance raises a particularly anthropological question: How to demarcate what counts as 'money' in ways anthropologists once did regarding so-called primitive currencies (Maurer, 2012).

Money and sovereignty, money and kinship, money and spheres of exchange, money and abstraction, spirits and speculation – these were the stuff of the anthropology of money at the end of the twentieth century. Anthropologists had come a long way from the old days of simply cataloging

the great diversity of objects used as media of exchange or stores of value.

Several empirical realizations and theoretical shifts are beginning to make more complex these themes and to suggest others. First, anthropologists, as (for the most part) subjects of capitalist, modern societies, came to the study of money with their own folk theories of economics and value in place. Disabusing ourselves of the idea that practical reason is acultural (Sahlins, 1978) proved easier than culturally situating the idea of money as always a commodity. Anthropologists have been as surprised as their informants to discover that money is a collective illusion, that its representational logics often fail. Is it a stand-in for 'real value,' or is it 'really valuable' in and of itself? (Foster, 1999). Hart usefully reminded the field as early as 1986 that money is always both a token and a commodity, but this lesson, and his equally significant revelation that money is above all a recording device, a kind of externalized memory, are still sometimes forgotten (Hart, 1986, 2001).

Guyer's important revision of Bohannan situated his spheres of exchange as part of a broader colonial and slave economy of an Atlantic Africa with ties to Birmingham and Nantes (where manillas, the brass bracelets serving as a currency of the slave trade, were actually manufactured) and beyond (Guyer, 2004). Bohannan had argued that 'conveyances' within one sphere of exchange are morally neutral but 'conversions' across the spheres are morally fraught, with people wanting to convert up into higher prestige spheres but loathe to convert down (Bohannan, 1959). Putting transactions in a broader historical and geographical context Guyer found that one could make a conveyance in another region, returning with new items from another sphere to convert up closer to home. In other words, Guyer finds conversion, not conveyance, to be a generalized phenomenon, and one inextricably bound with position in a set of sometimes intersecting hierarchies and geographies. For Guyer, this introduces the necessity of placing rank and its associated numerical relationships of ordinality at the center.

Guyer's move, in turn, required a more complex engagement with number itself. Numbers are not just a priori logical givens and do more than count things. There are different metaphors, embodiments, and physical or metaphysical reference to numbers. Rather than seeing number, like (the myth of) modern money, as abstracting and commensurating, tout court (Crump, 1978), recognition of rank requires an expansion of the analytical vocabulary for different kinds of numbers and styles of quantification and enumeration (Verran, 2001). This is becoming a fruitful area for the anthropological study of money (Pickles, 2013) but has to battle the headwinds of analysis of 'neoliberalism' and financialization that still insist on viewing the world in terms of the unfolding of one logic of capitalist abstraction, and one valence of quantification (Appadurai, 2011).

Guyer herself takes up the variety of conversions among different orders of value and geographies in order to query the contemporary world of multiple currencies (Guyer, 2012). Most, she observes, are 'soft,' that is, they do not function as a store of value or a reserve, but as a means of exchange. In economies of soft currencies, it is the money supply and not its ultimate value per se that matters, and that keeps exchanges going. One does not hold money, fetishizing its value, but instead tries to keep it moving and, crucially, to convert up. It is this potential for conversions that preoccupies people. The borders between domains and positions, the temporal dimensions of exchanging 'up,' and a variety of units of account and measurement thus map out a program for empirical investigation.

Guyer's program emphasizes creativity, dynamism, and pragmatic action. Another line of thinking in the anthropology of money focuses on money's materiality. One might assume that this work is therefore more static or less historical. This is not the case. Anthropological linguists developing a material semiotics have found in money a useful demonstration of some of their theoretical stances. Keane and Kockelman, in different ways, emphasize money's materiality as a communicative medium. Keane argued that when money is material, that very materiality underdetermines money's semiotic content (does it signify because of the marks inscribed on it? does it signify because of its qualities of being heavy, smooth, shiny, brittle?) (Keane, 2001). This indeterminacy allows money to remain an open signifier (Foster, 1999; Rutherford, 2001). Kockelman likens money to infrastructure – that which carries value across time, space, and language, a communicative channel subject to being redirected, skimmed from, or destroyed (Kockelman, 2010). Similar work along these lines draws attention to the polysemy of numbers on banknotes (Strassler, 2009), counterfeit notes (Chu, 2010), or redenominated actual (Dzokoto et al., 2010) or spiritual (Kwon, 2007) moneys. These studies shift anthropological attention from what money signifies, to how it does so.

There is much merit in bringing together these new strands of anthropological thinking on money. From the shift away from ethnological obsession with simply cataloging the stuff of money, to the opening-up of the spheres of exchange paradigm, to the reconsideration of money and number, money and commensuration, and money and abstraction, anthropologists can begin to examine the complex monetary ecologies and repertoires in which, even in a postneoliberal world, people make and do money with each other. The task, however, remains difficult: anthropologists are like most other moderns in accepting the separation, for example, of quality from quantity, and the purification of these as distinct and irreconcilable domains (Callon and Law, 2005; Latour, 1993). Anthropologists also take money, along with number, as necessarily capitalist forms of abstraction. Here, the work of economic sociologists (Zelizer, 1994) remains useful in pointing out the zones and practices in which people simply do not use money to equilibrate or deracinate but rather to make distinctions and relations.

Anthropologists also kept alive the intellectual tradition that saw money as a form of credit and as the state's monopoly over finality in payment, as guarantor of the clearance, and settlement of exchanges. Economist Wray, who has been a leading voice for the recovery of state and credit theories of money, writes:

> Theories of money became increasingly simplistic and silly [after the 1970s] with the rise of New Classical, Real Business Cycle and even New Keynesian approaches to macroeconomics. Serious monetary

> research was left to the fringe in economics (Post-Keynesians, Institutionalists, Political Economists, Social Economists), or to other disciplines such as Sociology or Anthropology.
> (Wray and Bell, 2004: 3)

This alternative tradition in monetary theory has become important in the aftermath of the global financial crisis that began in 2008. As with prior large-scale economic crisis, this one fed a reconsideration of money itself. As in the 1970s, activists and others rediscovered time- and community-based moneys of account like Time Banking and alternative currencies (Maurer, 2005; North, 2007). Others, however, drew from metallist traditions, becoming latter-day gold bugs as well as advocates of new currency experiments based on monetarist assumptions and the idea that value derives from scarcity (Maurer et al., 2013). Some academics rediscovered chartalism or state theories of money, as well as credit theories of money (Graeber, 2011) that see money as alternately the product of the state's enforcement of the use of its coupon to pay taxes, or as a chain of debts the ultimate arbiter and enforcer of which is the state. Others, particularly mainstream economists, either held onto monetarism or awakened the Keynes that had been lurking inside them all along (Krugman, 2013).

Governments, too, have been experimenting with new forms of exchange. As already noted, several countries dollarized, yet some of these same countries have been exploring various experiments in 'social economy' and 'solidary economy' (Nelms, 2013), a reaction against the G8 club of powerful countries setting global economic policies and agendas, as well as an assertion of the new power of some countries of the global South like Brazil. This took place as the Euro itself faltered: Greece and Cyprus exemplified the challenges of maintaining a currency union in the absence of unified fiscal policy.

The infrastructures of money are changing. Money in the core capitalist countries is increasingly – for the wealthy, at least – dematerialized. Electronic payments have reduced cash use in some countries. Some governments and nonprofits have an explicitly anticash agenda (Bjorklund Larsen, 2012). Their thinking assumes that cash's anonymity makes it the preferred vehicle for criminal activity and tax evasion. Digitizing transactions simultaneously institutes an electronic record, which can be audited, monitored, and tracked. Often put in terms of efficiency and convenience, electronic money systems at the same time frequently function through publicly traded, corporate infrastructures such as the payment card networks (Visa, MasterCard), which levy fees for transit.

Yet the digitization of cash also poses for anthropology and the study of money more generally the age-old question of what money in fact is, how people use it and what they use it for. In a sense, the digitization of money represents a return to the recognition of money as a memory device, as a means of recording credits and debts. What distance is there, really, between the cuneiform ledgers of ancient Sumer and the electronic transactional databases recording many of the world's payments in the early twenty-first century? There is of course the obvious distinction that the ancient database was a public one. But Hart for one has argued for new political options and vocabularies for a world of transactional data:

> Once we accept that money is a way of keeping track of complex social networks that we each generate, it could take a wide variety of forms compatible with both personal agency and collective forms of association at every level from the local to the global. It is up to us to build them.
> (2007: n.p.)

The early twenty-first century seems to be a new era of experimentation with and on money. Professionals from across the information and communications technology and the payments industry held large conferences to reimagine the future of money. Often there was a slippage in their discussions between the future of the transit of or infrastructures for state-issued money, and the future of money itself – its potential disintermediation from the state, and the introduction of a Hayekian world of privatized, denational currencies. A further slippage ensues when so-called money innovators waver between imagining these currencies as tokens or as commodities valuable in themselves because of some presumption of their scarcity.

Still, the increased circulation of private tokens harks back to the era before the consolidation of national currencies and, in the United States at least, the 'wildcat' banking period before state banks were regulated (Helleiner, 2003). If anything, innovations in money's transit and form are reopening a debate over the role of central banks in creating money and setting monetary policy and in the boundaries of regulatory authority over the means of payment and value transfer. The EU in the early 2000s issued several directives that tried to draw a fine line between infrastructures for value transfer and value storage, the latter seeming to invade the state's role as warranting the token at the top of the 'money pyramid' (see Wray and Bell, 2004: 12).

The historical irony is that as private industry experiments with infrastructures of value transfer and storage, and as populism on the left and right of the political spectrum contributes to the delegitimization of fractional reserve banking and central banks more generally (and this article is being written at the centenary of the US Federal Reserve system), central bankers have been busy rescuing the economies of the core capitalist nations. And a few anthropologists have been conducting fieldwork alongside them (Holmes, 2009).

From curatorial interests in the curiosities of primitive moneys, to debates over spheres of exchange and earmarking, anthropologists have more recently emphasized that money is representationally complex, more complex than a story of deracination and abstraction will allow. Its materiality matters, and it is not just a symbol (of state, of value, of social relations, or of power) but a practice. People's everyday interactions with money, especially outside the areas of the world dominated by reserve currencies whose value is more or less stable (though fragile), tend to emphasize its means of exchange modality, for there is no sense in hording a thing whose value may evaporate tomorrow. The view from this world of 'soft currencies' (Guyer, 2012) and of complex repertoires of money, number, value, and social relations may ultimately have more to say about the

involution and privatization of money in the early twenty-first century than that of the economists, monetarist, neoclassical, Keynesian, or otherwise.

See also: Economic Anthropology; Finance, Anthropology of; Infrastructures of Society, Anthropology of; Materiality and Culture; Property: Anthropological Aspects.

Bibliography

Appadurai, A., 2011. The ghost in the financial machine. Public Culture 23 (3), 517–540.

Bjorklund Larsen, L., 5 May 2012. Cash Is King – No More. What Does the Decreasing Circulation and Usage of Cash in Contemporary Sweden Mean for Our Perception of Money? Paper presented at the panel Debt, Sin and Tax: Money as Mediator of Sociality. Annual Meeting of the Swedish and Norwegian Anthropological Associations, Stockholm.

Bohannan, P., 1959. The impact of money on an African subsistence economy. The Journal of Economic History 19 (4), 491–503.

Callon, M., Law, J., 2005. On qualculation, agency, and otherness. Environment and Planning D: Society and Space 23 (5), 717–733.

Chu, J.Y., 2010. Cosmologies of Credit: Transnational Mobility and the Politics of Destination in China. Duke University Press Books, p. 360.

Comaroff, J., Comaroff, J.L., 1999. Occult economies and the violence of abstraction: notes from the South African postcolony. American Ethnologist 26 (2), 279–303.

Crump, T., 1978. Money and number: the Trojan horse of language. Man 13 (4), 503–518.

Dzokoto, V. Afi Abui, Mensah, E. Clifford, Opare-Henaku, A., 2010. A bird, a crocodile, and the 21st century cowrie shell: analyzing ghana's currency change. Journal of Black Studies 42 (5), 715–736.

Foster, R.J., 1999. In god we trust? The legitimacy of melanesian currencies. In: Akin, D., Robbins, J. (Eds.), Money and Modernity: State and Local Currencies in Melanesia. University of Pittsburgh Press, Pittsburgh, PA, pp. 214–231.

Graeber, D., 2011. Debt: The First 5000 Years. Melville House, Brooklyn.

Guyer, J.I., 2004. Marginal Gains: Monetary Transactions in Atlantic Africa. University of Chicago Press, Chicago.

Guyer, J.I., 2012. Soft currencies, cash economies, new monies: past and present. Proceedings of the National Academy of Sciences of the United States of America, 1–8.

Hart, K., 1986. Heads or tails? Two sides of the coin. Man 21 (4), 637–656 (Retrieved from).

Hart, K., 2001. Money in an Unequal World: Keith Hart and His Memory Bank. Texere, New York.

Hart, K., 2007. Money: Towards a Pragmatic Economic Anthropology. The Memory Bank Blog, Posted 15 July 2007. Available at: http://thememorybank.co.uk/2007/07/15/127/ (last accessed 18.08.13.).

Hayek, F.A., 1990. The Denationalisation of Money: The Argument Refined, third ed. Coronet Books, Philadelphia, PA.

Helleiner, E., 2003. The Making of National Money: Territorial Currencies in Historical Perspective. Cornell University Press, Ithaca.

Hernandez, E., Coutin, S.B., 2006. Remitting subjects: migrants, money and states. Economy and Society 35 (2), 185–208.

Holmes, D.R., 2009. Economy of words. Cultural Anthropology 24 (3), 381–419.

Keane, W., 2001. Money is No object: materiality, desire and modernity in an Indonesian society. In: Fred, E., Myers, R. (Eds.), The Empire of Things: Regimes of Value and Material Culture. SAR Press, Santa Fe, NM, pp. 65–90.

Kockelman, P., 2010. Enemies, parasites, and noise: how to take up residence in a system without becoming a term in it. Journal of Linguistic Anthropology 20 (2), 406–421.

Krugman, P., 2013. End This Depression Now! W.W. Norton & Company.

Kwon, H., 2007. The dollarization of Vietnamese ghost money. The Journal of the Royal Anthropological Institute 13 (1), 73–90.

Latour, B., 1993. We Have Never Been Modern. Harvard University Press, Cambridge.

Lemon, A., 1998. "Your eyes are Green like dollars": counterfeit cash, national substance, and currency apartheid in 1990s Russia. Cultural Anthropology 13 (1), 22–55.

Maurer, B., 2005. Mutual Life, Limited: Islamic Banking, Alternative Currencies, Lateral Reason. Princeton University Press, Princeton, NJ.

Maurer, B., 2012. In: Carrier, James (Ed.), Finance 2.0. In a Handbook of Economic Anthropology, second ed. Edward Elgar Publishers, Ltd, pp. 183–201.

Maurer, B., Nelms, T.C., Swartz, L., 2013. "When perhaps the real problem is money itself!": the practical materiality of Bitcoin. Social Semiotics 23 (2), 261–277.

Nelms, T., 2013. 25.000:1. Estudios de la Economia Blog. Available at: http://estudiosdelaeconomia.wordpress.com/2013/08/19/25-0001/.

North, P., 2007. Money and Liberation: The Micropolitics of Alternative Currency Movements. University of Minnesota Press, Minneapolis, MN, p. 240.

Parry, J.P., Bloch, M., 1989. Introduction: money and the morality of exchange. In: Parry, J.P., Bloch, M. (Eds.), Money and the Morality of Exchange. Cambridge University Press, Cambridge, UK, pp. 1–32.

Pedersen, D., 2013. American Value: Migrants, Money, and Meaning in El Salvador and the United States. University of Chicago Press, Chicago.

Peebles, G., 2011. The Euro and Its Rivals: Currency and the Construction of a Transnational City. Indiana University Press, Bloomington, IN.

Pickles, A.J., 2013. Pocket calculator: a humdrum "obviator" in Papua New Guinea? Journal of the Royal Anthropological Institute 19 (3), 510–526.

Rutherford, D., 2001. Intimacy and alienation: money and the Foreign in Biak. Public Culture 13 (2), 299–324.

Sahlins, M., 1978. Culture and Practical Reason. The University Of Chicago Press, p. 259.

Strassler, K., 2009. The face of money: currency, crisis, and remediation in Post-Suharto Indonesia. Cultural Anthropology 24 (1), 68–103.

Taussig, M., 1980. Devil and Commodity Fetishism in South America. University of North Carolina Press, Chapel Hill.

Truitt, A.J., 2013. Dreaming of Money in Ho Chi Minh City. University of Washington Press, Washington, DC.

Verran, H., 2001. Science and an African Logic. University of Chicago Press, Chicago.

Wray, L.R., Bell, S., 2004. Introduction. In: Wray, L.R. (Ed.), Credit and State Theories of Money: The Contributions of a. Mitchell Innes. Edward Elgar Pub, Cheltenham, pp. 1–13.

Zelizer, V., 1994. The Social Meaning of Money: Pin Money, Paychecks, Poor Relief, and Other. Princeton University Press, Princeton, NJ.

Montague Grammar

Barbara H Partee, University of Massachusetts Amherst, Amherst, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Montague grammar is a theory of semantics and of its relation to syntax, developed by the logician Richard Montague and subsequently extended by linguists, philosophers, and logicians. Montague grammar had its roots in logic and the philosophy of language; it became influential in linguistics, and subsequently evolved into contemporary formal semantics. Enduring features of the theory have been a truth-conditional theory of meaning, a model-theoretic conception of semantics, and the methodological centrality of the Principle of Compositionality: "The meaning of a whole is a function of the meanings of its parts and their mode of syntactic combination."

Introduction

Montague grammar is a theory of semantics, and of the relation of semantics to syntax, originally developed by the logician Richard Montague (1930–71) and subsequently modified and extended by linguists, philosophers, and logicians. Classical Montague grammar had its roots in logic and the philosophy of language; it quickly became influential in linguistics, and linguists have played a large role in its evolution into contemporary formal semantics. The most constant features of the theory over time have been the focus on truth-conditional aspects of meaning, a model-theoretic conception of semantics, and the methodological centrality of the Principle of Compositionality: "The meaning of a whole is a function of the meanings of its parts and their mode of syntactic combination." This article presents an outline of the origins of Montague grammar, summarizes the basic principles of the classical form of the theory, and describes some more recent developments and interesting controversies.

Montague Grammar in Historical Context

Richard Montague was a logician and philosopher of language whose seminal works (1970a,b, 1973) founded the theory known after his death as Montague grammar. Montague had been a student of Alfred Tarski (1901–83), one of the pioneers, with Frege and Carnap, in the model-theoretic semantics of logic (Feferman and Feferman, 2004). As a logician, Montague built on that tradition and developed an intensional logic with a rich type theory and a possible worlds model-theoretic semantics, incorporating certain aspects of pragmatics, including the treatment of 'indexical' words and morphemes like 'I, you' and the present tense. In the late 1960s, Montague turned to the project of 'Universal Grammar,' which for him meant a theory of syntax and semantics encompassing both formal and natural languages.

Montague's idea that a natural language like English could be formally described using logicians' techniques was a radical one at the time. Most logicians believed that natural languages were not amenable to precise formalization, while most linguists doubted the appropriateness of logicians' approaches to the domain of natural language semantics.

Part of the context of Montague's work was the 'Ordinary Language' versus 'Formal Language' war among logicians and philosophers of language. The Ordinary Language philosophers were a new generation who rejected the formal approach of Russell, Carnap, and Tarski, urging closer attention to Ordinary Language and its use. Central figures were the late Wittgenstein, Strawson, Austin, and Ryle. In 'On Referring' (Strawson, 1950), Strawson wrote, "Neither Aristotelian nor Russellian rules give the exact logic of any expression of ordinary language; for ordinary language has no exact logic." Russell replied in 'Mr. Strawson on Referring' (Russell, 1957): "I may say, to begin with, that I am totally unable to see any validity whatever in any of Mr. Strawson's arguments." … "I agree, however, with Mr. Strawson's statement that ordinary language has no logic." Thus both sides in this 'war' (as well as Chomsky, as expressed in his reply to Bar-Hillel (Chomsky, 1955)) were in agreement that logical methods of formal language analysis did not apply to natural languages.

Also at the time of Montague's work, Chomskyan generative syntax was well established, and linguists were developing and debating approaches to semantics to fit into the enterprise of generative grammar, and questions on the relation of semantics to syntax had become central. The 'linguistic wars' between generative semantics (Lakoff, Ross, McCawley, Postal) and interpretive semantics (Jackendoff, with the support of Chomsky) were in full swing.

Montague rejected the premises of the 'Ordinary Language' versus 'Formal Language' war. His 'Universal Grammar' (Montague, 1970b) begins with the sentence, "There is in my opinion no important theoretical difference between natural languages and the artificial languages of logicians; indeed I consider it possible to comprehend the syntax and semantics of both kinds of languages with a single natural and mathematically precise theory" (p. 222). And while he was sympathetic to Chomsky's goals, he was not impressed with the results 'emanating from MIT,' remarking, "I fail to see any great interest in syntax except as a preliminary to semantics" (Montague, 1970b, p. 223).

In the earliest introductions of Montague's work to linguists, including Partee (1973, 1975) and Thomason's extended introduction to Montague (1974), it was argued that Montague's work offered the potential to accommodate

 http://dx.doi.org/10.1016/B978-0-08-097086-8.53022-0

some of the best aspects of both generative semantics and interpretive semantics, with some added advantages of its own.

Early attempts to integrate Montague grammar with transformational grammar (Partee, 1975) faced two kinds of obstacles. On the one hand, Montague's own syntax was quite unconstrained and in practice combined some systematic use of categorial grammar with some ad hoc extensions; it clearly needed overhaul for serious linguistic work. On the other hand, transformational grammars of the period could not be directly coupled with Montague's semantics because many of the posited transformations did not have uniform semantic effects, as Montague's theory required. Much of the subsequent diversity in approaches to formal semantics that incorporate features of Montague's theory reflect different responses to those problems.

Montague's work, and related work by the philosophers Michael Bennett, M.J. Cresswell, David Lewis, Terence Parsons, Robert Stalnaker, Richmond Thomason, and others, and by linguists such as Emmon Bach, Robin Cooper, David Dowty, E.L. Keenan, Lauri Karttunen, and Barbara Partee gained influence during the 1970s. Formal semantics became a mainstream approach to semantics within linguistics by the 1980s; major textbooks include Chierchia and McConnell-Ginet (1990), Dowty et al. (1981), Gamut (1991), and Heim and Kratzer (1998); a good 'lay introduction' is Bach (1989). See also Janssen (1981), Lewis (1970), and Partee (1976, 1997).

The Theory and the Substance of Montague Grammar: Central Principles

The paper of Montague's that had the most impact on linguists and on the subsequent development of formal semantics was "The proper treatment of quantification in ordinary English" ('PTQ') (Montague, 1973): short, but densely packed. 'Montague grammar' has often meant what Montague did in the fragment in PTQ and the extensions of PTQ by linguists and philosophers in the 1970s and 1980s with greater or lesser innovations. But it is the broader algebraic framework of 'UG' ('Universal Grammar,' Montague, 1970b) that constitutes Montague's theory of grammar. This section therefore begins with the basic principles laid out in UG, concentrating on the Principle of Compositionality, then continues with model-theoretic interpretation, type theory, and the 'method of fragments,' including some of the key features of the fragment in PTQ.

Universal Grammar: Syntax and Semantics as Algebras, Compositionality as Homomorphism

Montague's 'Universal Grammar' (Montague, 1970b) contains the most general statement of Montague's formal framework for the description of language. The central idea is that a grammar should be able to be cast in the following form: the syntax is an algebra, the semantics is an algebra, and there is a homomorphism mapping elements of the syntactic algebra onto elements of the semantic algebra. This very general definition leaves a great deal of freedom as to the nature of these algebras. In a logical language, the elements of the syntactic algebra can be the well-formed expressions. But for a natural language, ambiguity makes that impossible, since the homomorphism requirement means that each element of the syntactic algebra must be mapped onto a unique element of the semantic algebra. So for a natural language, the elements of the syntactic algebra are usually taken to be expressions together with disambiguating structural descriptions, typically trees of some sort.

The nature of the elements of the semantic algebra is also open to variation. They may be model-theoretic constructs, as in Montague's own work, Heim's 'file change potentials,' the extensional domains of first-order logic, hypothesized concepts, expressions in a 'language of thought,' or anything else, as long as they can be formalized in a suitable algebra. What is constrained is not the 'substance' of the semantics but properties of its structure and its relation to syntax.

It is the homomorphism requirement, which formalizes the compositionality requirement, that provides one of the most important constraints on UG in Montague's sense. The compositionality requirement, stated in the introductory paragraph, is almost uncontroversial when stated informally. But formalizing it requires an explicit theory of syntax (to specify what is meant by *parts* and *syntactic combination*), an explicit theory of semantics (*meanings*), and an explicit theory of the mapping from one to the other (*is a function of*).

Logicians typically specify the syntax of a formal language as a recursive definition; in that case the requirement of compositionality as homomorphism can be satisfied by giving the semantics in the form of a parallel recursive definition. The 'derivation trees' that correspond to the steps in applying a recursive definition become the elements of the syntactic and semantic algebras, and the homomorphism requirement says that each syntactic derivation must be mapped onto a unique semantic derivation.

The simplest linguistic examples take the elements of the syntactic algebra to be (sub)trees generated by a context-free grammar, and semantic interpretation to specify how the interpretation of a given subtree is computed from the interpretations of its immediate 'daughter' subtrees. Formal semanticists often work either with a 'monostratal grammar,' i.e., a grammar with a single level of syntax and no transformations (Generalized Phrase Structure Grammar (GPSG), Head-Driven Phrase Structure Grammar (HPSG), Lexical-Functional Grammar (LFG), Categorial Grammar, Tree-Adjoining Grammar (TAG)), or with a specified level of syntactic representation such as 'LF' (a syntactic level of 'logical form') in Chomskyan syntax.

Model Theory: Direct and Indirect Interpretation

Montague's 'Universal Grammar' presents formal frameworks for both 'direct' and 'indirect' semantic interpretation. The fragment of Montague, (1970a) illustrates the direct model-theoretic interpretation of natural language syntactic rules; direct model-theoretic interpretation of natural language is also found in some of the work of Cresswell, von Stechow, and Kratzer. 'Indirect' interpretation, illustrated in UG and PTQ and in the works of many others, proceeds via translation into an intermediate language, such as Montague's intensional logic. The translation process must then itself be compositional, meaning that the translation of

an expression must be a function of the translations of its parts, and the intermediate language must itself be given a compositional model-theoretic interpretation. The intermediate language is then, in principle, dispensable. Montague viewed the intermediate language as offering increased perspicuity in presentation but attached no theoretical importance to it; linguists tend to want evidence for or against the psychological reality of any posited level of 'semantic representation,' but strong evidence for or against such levels is still lacking.

Type Theory and Intensionality

Linguists before Montague had tried unsuccessfully to give a compositional semantics to natural language syntax, but their efforts were stymied in part by the mismatch between natural language syntax and first-order logic. The richness of Montague's logic, with several crucial features previously unfamiliar to linguists, was a crucial factor in the possibility of giving a compositional semantic interpretation to independently motivated syntactic structure.

Natural languages have many kinds of expressions: sentences, noun phrases, verb phrases, adjectives, adverbs of several sorts, prepositional phrases, and more. First-order logic has only three 'types' of expressions: formulas (analogous to sentences), terms (analogous to a subset of noun phrases), and predicates (analogous to nouns, adjectives, and verbs); this means that 'logical forms' for natural language sentences must often have structures very different from the structure of the given language. Montague's logic has a rich system of types built up by recursive rules from two simple types t and e corresponding to formulas and terms. One rule builds up 'functional types,' defining for any two types a and b the type <a,b> of functions from a-type things to b-type things. The other rule builds 'intensional types,' defining for any type a the type <s,a> of functions from possible worlds to things of type a. In order to follow Frege's insight that one can view function-argument application as the basic 'semantic glue' by which meanings are combined, Montague makes crucial use of the lambda calculus, first introduced by Church (1940), in order to explicitly define functions and work with function-argument application in compositional derivations. With his rich system of semantic types, Montague showed for the first time how one can interpret noun phrases like *every man*, *the man*, *a man* uniformly as semantic constituents, something impossible with the tools of first-order logic.

Intensionality is a key notion in the analysis of referential opacity. The construction exemplified in "Jones is seeking…" is *referentially opaque*, since the substitution of one coreferential expression for another in that context does not always preserve the truth value of the whole. It may be true that Jones is seeking the president and false that Jones is seeking Mary's father even though the president is Mary's father. Montague's intensional logic develops Frege's distinction between *sense* and *reference* and Carnap's distinction between *intension* and *extension*, to treat the phenomenon of referential opacity, pervasive in belief-sentences and other propositional attitude constructions.

Montague analyzed intensions as functions from possible worlds to corresponding extensions: propositions as functions from possible worlds to truth values, individual concepts as functions from possible worlds to individuals, properties as functions from possible worlds to sets. Montague proposed that not only propositional attitude verbs like *believe* create opaque contexts, but also transitive verbs like *seek* and *need*, modifiers like *former* and *alleged*, and prepositions like *about*. Montague's analysis of intensionality opened up new avenues of research as well as leading to interesting controversies and discussions of possible alternative ontological foundations for semantic theory.

The Method of Fragments: The PTQ Fragment

Each of Montague's three 'grammar' papers (Montague, 1970a,b, 1973) contains a 'fragment.' The term was introduced in 'English as a formal language' (Montague, 1970b), which begins:

> In the present paper I shall accordingly present a precise treatment, culminating in a theory of truth, of a formal language that I believe may be reasonably regarded as a fragment of ordinary English.
>
> Montague, 1974, p. 188

The 'method of fragments,' a methodological feature of Montague's work, is the writing of a complete syntax and semantics for a specified subset ('fragment') of a language, rather than writing rules for a single construction while making implicit assumptions about the rest of the grammar.

The fragment of PTQ, the best known of Montague's works, had a number of noteworthy features. One was the systematic correspondence between syntactic categories and semantic types, an appreciation of which led to a revival of interest in categorial grammar, where such a correspondence is built into the system of categories.

Another innovation was Montague's bottom-up syntax, combining concatenation and transformation-like operations in single recursive grammatical rules (modeled on the recursive rules of logicians' syntax), thus building well-formed expressions of all sorts of categories directly, rather than building an entire 'deep structure' top-down and then applying transformations bottom-up.

Montague's PTQ contained many other ideas that were new to linguists, some of which had antecedents in the work of other logicians and philosophers. We mention only a few here. The fragment illustrated the use of lambda-abstraction to generate phrasal conjunction directly, without the need for the semantically problematic syntactic transformation of 'Conjunction Reduction.' Montague introduced a version of Quine's treatment of relative clauses as 'sentential adjective phrases' derived from open sentences by lambda-abstraction, without any kind of 'deletion under identity.' Thus instead of deriving *The boy who loves Mary is happy* from the pair of sentences *The boy loves Mary* and *The boy is happy* by a transformational rule, Montague starts from an 'open' sentence like *He*$_3$ *loves Mary* and derives *who loves Mary* from that, lambda abstracting over the variable x_3, which interprets *he*$_3$, to interpret the relative clause as the property of being an individual who loves Mary. The relative clause is then combined with the noun *boy* very much as an adjective like *blond* would be. Montague also introduced the binding of

bound-variable pronouns by lambda operators rather than directly by quantifiers or by their antecedents, a move that has had major consequences for theories of anaphora. He used meaning postulates to distinguish entailments due to individual lexical items (e.g., that *seek* entails *try to find*) from entailments due to the semantics of a grammatical construction, e.g., that *A boy who loves Mary arrived* entails that *A boy arrived*, and laid the foundations for the strong 'lexicalist' tradition in formal semantics led by Dowty (1979).

Early Formal Pragmatics

The term pragmatics is due to Charles Morris (1938). Within semiotics, the general science of signs, Morris distinguished three branches: syntactics (now syntax), the study of "the formal relation of signs to one another"; semantics, the study of "the relations of signs to the objects to which the signs are applicable" (their designata); and pragmatics, the study of "the relations of signs to interpreters" (1938, p. 6). Semantics in the work of formal language philosophers such as Russell is pure semantics, and the need for attention to context dependence in interpretation was one of the concerns of the ordinary language philosophers. One of the first aspects of context dependence on which logicians made progress was tense, where Prior and Reichenbach were pioneers. And Montague was an important contributor to these developments. In "Pragmatics" (Montague, 1968) and "Pragmatics and intensional logic" (Montague, 1970c), he developed a higher order typed intensional logic that unified tense logic and modal logic, extending Prior's work and building on Church's work on intensional logic and type theory (Church, 1951), and more generally unified 'formal pragmatics' with intensional logic.

From Montague Grammar to Formal Semantics

Early Impact of Montague's Work on Linguistics

Before Montague, linguists took as the task of semantics the explication of ambiguity, semantic anomaly, and synonymy: the key questions were how many readings a sentence has, and which sentences share readings. But the individuation of 'readings' had always been problematic, so there was often crucial disagreement about data. The introduction of truth conditions as the basic semantic property of a sentence that a semantic theory should capture profoundly affected the adequacy criteria for semantics, and led to a great expansion of semantic research.

A methodological principle implicit in Chomskyan syntax in the 1960s, encouraged although not required by the Katz–Postal hypothesis that meaning is determined at deep structure (Katz and Postal, 1964), and carried to extremes in generative semantics, was the principle that sameness of meaning should be reflected in sameness of deep structure. But from the perspective of Montague grammar, sameness of meaning does not require identity at any syntactic level (cf Thomason, 1976): semantics is model-theoretic, not representational, not 'syntactic.'

With a semantics that made it possible to calculate compositionally whether two expressions have the same truth-conditional potential or not, many linguists concluded that there was no need to invent ever-more-abstract underlying structures in order to capture synonymy or sameness of truth conditions, and much of the motivation for that aspect of generative semantics disappeared. Its semantic goals could apparently be met, even exceeded, in a formal semantic approach, offering greater explicitness and compatible with more 'conservative' syntax. Thus the rise of Montague grammar was undoubtedly one factor in the decline of generative semantics and the fading away of the linguistic wars.

Details of Montague's analyses have in many cases been superseded, but in overall impact, PTQ was as profound for semantics as Chomsky's *Syntactic Structures* was for syntax. Emmon Bach (1989) summed up their cumulative innovations thus: Chomsky's thesis was that English can be described as a formal system; Montague's thesis was that English can be described as an *interpreted* formal system.

Montague's method of 'fragments' is no longer widely followed except in rule-based approaches to computational linguistics, where completeness is a necessity. Many aspects of Montague grammar have found fruitful application in computational semantics (see section 'Computational Semantics'), and many innovations (such as methods for working with underspecified representations rather than with large sets of meanings of ambiguous expressions) have in turn come from the computational semantics community.

Linguistics–Philosophy Interactions in the Development of Formal Semantics

At the time of Montague's work, there were already interactions between linguists and philosophers; Donald Davidson and Gilbert Harman, wanting to encourage more cooperation and understanding, and optimistic about generative semantics, organized a conference in 1969 that included the linguists Emmon Bach, Charles Fillmore, George Lakoff, James McCawley, and Barbara Partee, and the philosophers Donald Davidson, Peter Geach, Gilbert Harman, David Kaplan, W.V. Quine, J.F. Staal, and Bruce Vermazen. The conference resulted in the influential volume (Davidson and Harman, 1972), which also included the important papers (Kripke, 1972; Lewis, 1970; Montague, 1970c) and papers by others not at the conference. And in 1968, a first version of Terry Parsons' "A Semantics for English" (Parsons, 1972) was circulated, a number of linguists knew and discussed that work, as well as (Donnellan, 1966), and there was much interest among linguists in the puzzles of 'donkey pronouns' posed in Geach (1962). The philosophers Julius Moravcsik and Jaakko Hintikka, together with the polymath Patrick Suppes, organized a 1970 workshop of linguists and philosophers that included Montague's last paper. The linguist Edward Keenan was well versed in logic and philosophy; in addition to his own innovative work on semantics, syntax, and linguistic typology, he organized a conference in 1973 that led to the influential interdisciplinary volume (Keenan, 1975), in whose title the term 'Formal Semantics' made an early appearance. The philosopher Richmond Thomason edited Montague's works on 'formal philosophy' and included a long pedagogical introduction for linguists and philosophers (Thomason, 1974). The first collection of papers on Montague grammar (Partee, 1976) included papers by the philosophers David Lewis, Richmond Thomason, Michael Bennett, C.L. Hamblin, and M.J. Cresswell and by linguists Barbara Partee, Robert Rodman,

David Dowty, and Muffy Siegel, plus a paper co-authored by a linguist and a philosopher, Robin Cooper and Terence Parsons. Philosophers and linguists in Germany, Scandinavia, and in Amsterdam also did important work in Montague grammar starting in the 1970s.

In the 1970s and 1980s there were many interdisciplinary conferences and a significant number of collaborative works by linguists and philosophers or logicians: Barwise and Cooper, 1981; Bennett and Partee, 1972; Chierchia and Turner, 1988; Cresswell and von Stechow, 1982; Keenan and Stavi, 1986; and more.

Approaches to the Syntax–Semantics Interface

Montague had a very abstract theory of syntax (as an algebra), and some specific syntactic ideas; but in general his semantics was far more influential than his syntax. There have been many proposals for different ways to link Montague's semantics, or some variant of it, to different sorts of syntactic theories. Partee originally proposed ways to synthesize Montague grammar with Chomsky's transformational grammar (Partee, 1973, 1975). But the richness of Montague's semantics eliminates some of the motivation for transformations and lends itself to a monostratal syntax with a rich lexical component. Theories that have included a semantics influenced by Montague grammar include Extended Categorial Grammar, GPSG, HPSG, Heim's File Change Semantics, Dynamic Montague Grammar, and to some extent Kamp's Discourse Representation Theory and Asher and Lascarides' more recent Segmented Discourse Representation Theory. Starting with the work of Heim (1982), there has also been a rich tradition of theories using Chomskyan generative grammar with a syntactic level of logical form, using that as the syntactic input for compositional formal semantics.

The Naturalization of Formal Semantics

As formal semantics has developed from Montague grammar since the beginning of the 1970s, it has increasingly become a core part of mainstream theoretical linguistics, with continuing important contributions from logicians and philosophers. Areas of early research success tended to be in the 'logical' parts of language: quantification, anaphora, conjunction, tense, and aspect. Later work expanded the domains covered and the range of languages treated. An important development of the 1980s was the work of Kamp and Heim on Discourse Representation Theory and File Change Semantics. Their work led to changes in theories of indefinite noun phrases, quantification, and anaphora, and more generally led to the replacement of truth conditions by 'context change potential' as the core semantic property of sentences. Later work by Groenendijk and Stokhof gave an alternative perspective on the dynamic interaction of meaning and context (Groenendijk and Stokhof, 1990, 1991) with associated innovations in logic. Formal theories of topic-focus structure, interrogatives, presuppositions and implicatures, and many other aspects of natural language semantics and pragmatics have also been developed.

The basic model theory has been enriched by additional structure on the domain of entities, to model mass nouns and 'plural entities' (Link, 1998), and by widespread acceptance of events as basic entities. Kratzer and others have argued for a 'situation semantics' as a more fine-grained model structure, taking possible situations as parts of possible worlds. There is increasing attention to details of lexical semantics, and to interfaces of semantics with syntax, on one hand, and pragmatics on the other.

Further Developments and Controversies

Typology and Natural Language Metaphysics

The issue of "Natural Language Metaphysics" (Bach, 1986) is an important foundational area that connects cognitive issues with formal semantics and linguistic typology. What presuppositions concerning the constitution and structure of the world as humans conceive it are built into human languages, and how, and which are universal? These questions may concern both semantic structure and semantic content, from the semantic difference between nouns and verbs to the content of color terms or whether time is discrete or continuous. Their investigation may challenge the lines between semantic knowledge and other kinds of knowledge. Formal semantics, following the logical tradition, initially employed relatively 'austere' model structures; recent investigations, particularly into lexical semantics, are leading to richer model structures.

Formal semantics has increasingly expanded its scope to include work on a wide variety of languages, and in recent decades has made significant contributions to the study of linguistic universals and linguistic typology (Bach et al., 1995; Bittner, 1994; Keenan and Paperno, 2012; Matthewson, 2010; von Fintel and Matthewson, 2008). Typological investigations, especially but not only in the domains of quantification and of tense, aspect and mood, have included the distribution of nominal versus adverbial quantification, the semantics of languages that lack a determiner category, and the range of tense and aspect systems. Recent work on the formal semantics of sign languages has led to enriched perspectives on the semantics of demonstratives, variable binding, propositional attitudes, quotation, and other domains (Quer, 2005; Schlenker, 2011).

Computational Semantics

Until recently, there were few attempts to apply formal semantic methods to computational linguistic tasks, one interesting exception being the Rosetta project led by Jan Landsbergen at Philips Eindhoven in the 1980s (Rosetta, 1994). Recently, there have been advances in methods for making formal semantics computationally tractable, and interest in computational formal semantics has increased to the point that there are textbooks and courses in the field (Blackburn and Bos, 2003, 2005; Bunt and Muskens, 1999). A few companies have begun to offer computational linguistic products that use formal semantics and others are exploring the possibility of doing so. For some kinds of natural language processing problems, such as the solving of the kinds of logic puzzles found on the Graduate Record Exam, formal methods can offer important advantages and have been implemented (Lev et al., 2004). Connecting computational linguistics, psychology, and formal semantics, there is promising current research aimed at combining the best of statistical/probabilistic and formal approaches, exemplified by the

work of Joshua Tenenbaum's Computational Cognitive Science group at MIT (Tenenbaum et al., 2011).

Psycholinguistics and Formal Semantics and Pragmatics: Processing and Acquisition

Other recently expanding domains of research connect psycholinguistics and formal semantics, research that helps to strengthen both fields. The study of how children develop an understanding of sentences with quantifiers has a relatively long history (Crain, 2009; Philip, 1992; Smits et al., 2006). The study of real-time processing of language in both adults and children has expanded over recent decades to include attention to semantics and pragmatics as well as syntax, with particular interest in issues of compositionality, modularity, and the extent to which such pragmatic phenomena as implicatures interact with semantics (Chierchia et al., 2004; Frazier, 1999). A recent overview of work on the acquisition of semantics can be found in Crain and Thornton (2011).

Context Dependence and Formal Pragmatics

In subsequent work on Montague grammar and formal semantics, the treatment of context-dependent phenomena such as tenses, pronouns, demonstratives, and all sorts of indexicality has been a central concern. Other aspects of formal pragmatics such as the treatment of presupposition and focus sensitivity have been recognized as central to the study of meaning, so that 'formal semantics and pragmatics' has increasingly come to be seen as a single field, as described in Kadmon (2001) and as evidenced by the new online journal *Semantics and Pragmatics*; the aspects of pragmatics most closely connected to semantics are called 'near-side pragmatics' in the useful article by Korta and Perry (2008). The separation of pragmatics and semantics has been further challenged with arguments that the interpretation of pragmatic implicatures occurs interspersed with semantic interpretation (Chierchia, 2004).

Ongoing Debates within and around Formal Semantics

Some aspects of Montague's approach met with resistance or skepticism in the linguistics community, and some aspects were controversial among philosophers. We mention three foundational issues that distinguish Montague grammar, and formal semantics in general, from other approaches to semantics.

Many linguists have resisted the relevance of truth conditions and entailment relations to natural language semantics, and some still do. Some of the objections to truth conditions are countered by arguing that part of human semantic competence is the matching of sentences with their truth conditions relative to possible worlds (including fictional worlds), with no necessary reference to the 'actual world.' Entailment (sentence A *entails* sentence B if sentence B is true in every possible state of affairs in which sentence A is true) is a central semantic concern in logic, and remains so in formal semantics. Cognitive semanticists replace concern with logical entailment by concern with human inference; formal semanticists see the relation of entailment to actual human inference as indirect. But many semanticists agree about the importance of revising the formal logics invented by logicians to model the 'natural logic(s)' implicit in the semantics of natural languages.

Many linguists think of semantics in terms of a 'level of representation' of expressions analogous to a syntactic or phonological level. A representational view of semantics is quite congenial to the popular computational theory of mind. The contrasting model-theoretic view sees semantic interpretation relating expressions to elements of models (possibly mental models) defined in terms of constituents such as possible situations, entities, properties, truth values, etc.

The nonpsychologistic tradition of 'objective' (though abstract) meanings (Frege, Tarski, Carnap, Montague) contrasts with the psychologistic view of meanings 'in the head' (Fodor, Lakoff, Jackendoff, and all psychologists). Do expressions refer to objects or to concepts? Is semantics a branch of mathematics, or is it (as on the Chomskyan view of all of linguistics) a branch of psychology? Classical formal semanticists, who take the first disjunct in these choices, follow David Lewis in distinguishing semantics from knowledge of semantics, making semantic competence interestingly different from syntactic competence. More recent work in the philosophy of mind and of perception (Burge, 2003; Stalnaker, 1989) goes a long way toward reconciling the apparent tension between objective and mentalistic foundations for semantics by emphasizing how mental states depend in significant part on the environment: "Meanings are in the head like footprints are in the sand" (Stalnaker, 1989).

New Directions in Formal Semantics and Pragmatics

Many new directions have been mentioned in the preceding sections. Others include the emergence of game-theoretical semantics (Hintikka and Sandu, 1997) and a game-theoretical approach to integrated semantics and pragmatics (Benz et al., 2005). A number of researchers are exploring experimental and corpus-based methods for gaining new kinds of data for semantics as well as syntax (Krifka, 2011). There are beginnings of neurolinguistic research in formal semantics (Brennan and Pylkkänen, 2008) to complement the much older tradition of studying the neurolinguistic encoding of lexical meanings.

Of the key journals for the field, one, *Linguistics and Philosophy*, first published in 1977, maintains the collaborative links between those fields that marked the beginnings of Montague grammar. Another, *Natural Language Semantics*, first published in 1993, is more nearly centered within linguistics and includes greater concern with the syntax–semantics interface. A third, *The Journal of Logic, Language, and Information*, first published in 1992, reflects the split found in some European universities, where Chomskyan syntax is located in linguistics departments and formal semantics within philosophy or logic departments. And *Journal of Semantics*, which began in 1982, has recently reoriented itself to become a major venue for formal and experimental semantics. There is now also an important online open-access journal, *Semantics and Pragmatics*, first published in 2007, whose name reflects the increasing integration of those two subfields.

While this article has emphasized Montague grammar and formal semantics within linguistics, fruitful cross-fertilizations between linguistics and neighboring fields continue to strengthen and enrich the field now as they did at its beginnings.

See also: Anaphora; Deixis; Generative Grammar; Information Structure in Linguistics; Lexical Semantics; Linguistic Presupposition; Logic and Linguistics; Meaning and Rule Following; Negation, Linguistics of; Pragmatics, Linguistic; Quantifiers; Reference and Representation: Philosophical Aspects; Semantics; Syntax–Semantics Interface; Tense, Aspect, Mood and Evidentiality, Linguistics of.

Bibliography

Asudeh, Ash, Crouch, Richard, 2002. Glue semantics for HPSG. In: van Eynde, Frank, Hellan, Lars, Beermann, Dorothee (Eds.), Proceedings of the 8th International HPSG Conference, Norwegian University of Science and Technology (3–5 August 2001). CSLI Publications, Stanford.

Bach, Emmon, 1986. Natural language metaphysics. In: Marcus, Ruth Barcan, Dorn, Georg J.W., Weingartner, Paul (Eds.), Logic, Methodology, and Philosophy of Science VII. North-Holland, Amsterdam, pp. 573–595.

Bach, Emmon, 1989. Informal Lectures on Formal Semantics. State University of New York Press, New York.

Bach, Emmon, Jelinek, Eloise, Kratzer, Angelika, Partee, Barbara H. (Eds.), 1995. Quantification in Natural Languages. Studies in Linguistics and Philosophy No. 54, vols. I and II. Kluwer, Dordrecht.

Barwise, Jon, Cooper, Robin, 1981. Generalized quantifiers and natural language. Linguistics and Philosophy 4, 159–219.

Bennett, Michael, Partee, Barbara, 1972. Toward the Logic of Tense and Aspect in English. System Development Corporation, Santa Monica, California reprinted with an Afterword by Indiana University Linguistics Club, Bloomington, 1978.

Benz, Anton, Jaeger, Gerhard, van Rooy, Robert (Eds.), 2005. Game Theory and Pragmatics. Palgrave Macmillan, New York.

Bittner, Maria, 1994. Cross-linguistic semantics. Linguistics and Philosophy 17, 53–108.

Blackburn, Patrick, Bos, Johan, 2003. Computational semantics. Theoria 18, 27–45.

Blackburn, Patrick, Bos, Johan, 2005. Representation and Inference for Natural Language: A First Course in Computational Semantics: CSLI Studies in Computational Linguistics. CSLI Publications, Stanford.

Brennan, Jonathan, Pylkkänen, Liina, 2008. Processing events: behavioral and neuromagnetic correlates of aspectual coercion. Brain and Language 106, 132–143.

Bunt, Harry C., Muskens, Reinhard (Eds.), 1999. Computing Meaning. Studies in Linguistics and Philosophy, vol. 73. Kluwer Academic, Dordrecht; Boston.

Burge, Tyler, 2003. Psychology and the environment: reply to Chomsky. In: Hahn, Martin, Ramberg, Bjørn (Eds.), Reflections and Replies: Essays on the Philosophy of Tyler Burge. MIT Press, Cambridge, MA, pp. 451–470.

Chierchia, G., Guasti, M.T., Gualmini, A., Meroni, L., Crain, S., 2004. Semantic and pragmatic competence in children and adult's interpretation of 'or'. In: Noveck, I., Wilson, S. (Eds.), Experimental Pragmatics. Palgraves, London, pp. 283–300.

Chierchia, Gennaro, Turner, Raymond, 1988. Semantics and property theory. Linguistics and Philosophy 11, 261–302.

Chierchia, Gennaro, McConnell-Ginet, Sally, 1990. Meaning and Grammar. An Introduction to Semantics. MIT Press, Cambridge.

Chierchia, Gennaro, 2004. Scalar implicatures, polarity phenomena, and the syntax/-pragmatics interface. In: Belleti, Adriana (Ed.), Structures and Beyond: The Cartography of Syntactic Structures, vol. 3. Oxford University Press, Oxford.

Chomsky, Noam, 1955. Logical syntax and semantics: their linguistic relevance. Language 31, 36–45.

Church, Alonzo., 1940. A formulation of the simple theory of types. Journal of Symbolic Logic 5, 56–68.

Church, Alonzo, 1951. A formulation of the logic of sense and denotation. In: Henle, P., Kallen, H., Langer, S. (Eds.), Structure, Method, and Meaning: Essays in Honor of H.M. Sheffer. Liberal Arts Press, New York.

Crain, Stephen, Thornton, Rosalind, 2011. Acquisition of semantics. In: Colm Hogan, Patrick (Ed.), The Cambridge Encyclopedia of the Language Sciences. Cambridge University Press, Cambridge, pp. 745–748.

Crain, Stephen, 2009. Sentence scope. In: Bavin, E. (Ed.), The Cambridge Handbook of Child Language. Cambridge University Press, Cambridge, UK.

Cresswell, M.J., von Stechow, Arnim, 1982. De re belief generalized. Linguistics and Philosophy 5, 503–535.

Davidson, Donald, Harman, Gilbert (Eds.), 1972. Semantics of Natural Language. Reidel, Dordrecht.

Donnellan, Keith, 1966. Reference and definite descriptions. Philosophical Review 75, 281–304.

Dowty, David, 1979. Word Meaning and Montague Grammar. The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ. Reidel, Dordrecht.

Dowty, David, Wall, Robert E., Peters Jr., Stanley, 1981. Introduction to Montague Semantics. Reidel, Dordrecht.

von Fintel, Kai, Matthewson, Lisa, 2008. Universals in semantics. The Linguistic Review 25, 139–201.

Feferman, Anita Burdman, Feferman, Solomon, 2004. Alfred Tarski: Life and Logic. Cambridge University Press, Cambridge.

Frazier, Lyn, 1999. On Sentence Interpretation. Kluwer Academic Publishers, Dordrecht.

Gamut, L.T.F., 1991. Logic, Language, and Meaning. Introduction to Logic, vol. 1. Intensional logic and logical grammar, vol. 2. University of Chicago Press, Chicago.

Geach, Peter T., 1962. Reference and Generality. Cornell University Press, Ithaca.

Groenendijk, Jeroen, Stokhof, Martin, 1990. Dynamic Montague grammar. In: Kálman, L., Pólos, L. (Eds.), Papers from the Second Symposium on Logic and Language. Adakémiai Kiadó, Budapest, pp. 3–48.

Groenendijk, Jeroen, Stokhof, Martin, 1991. Dynamic predicate logic. Linguistics and Philosophy 14, 39–100.

Heim, Irene, 1982. The Semantics of Definite and Indefinite Noun Phrases. University of Massachusetts: Ph.D. Dissertation; published 1989. Garland, New York.

Heim, Irene, Kratzer, Angelika, 1998. Semantics in Generative Grammar. Blackwell, Oxford.

Hintikka, J., Sandu, G., 1997. Game-theoretical semantics. In: van Benthem, J., ter Meulen, A. (Eds.), Handbook of Logic and Language. MIT Press, Cambridge, MA, pp. 361–410.

Janssen, T.M.V., 1981. Compositional semantics and relative clause formation in Montague grammar. In: Groenendijk, Jeroen, Janssen, Theo, Stokhof, Martin (Eds.), Formal Methods in the Study of Language. Part 1. Mathematisch Centrum, Amsterdam, pp. 445–481.

Kadmon, Nirit, 2001. Formal Pragmatics: Semantics, Pragmatics, Presupposition, and Focus. Blackwell Publishers, Oxford.

Katz, Jerrold J., Postal, Paul M., 1964. An Integrated Theory of Linguistic Descriptions. MIT Press, Cambridge, MA.

Keenan, Edward (Ed.), 1975. Formal Semantics of Natural Language. Cambridge University Press, Cambridge.

Keenan, Edward, Stavi, Jonathan, 1986. A semantic characterization of natural language determiners. Linguistics and Philosophy 9, 253–326.

Keenan, Edward L., Paperno, Denis (Eds.), 2012. Handbook of Quantifiers in Natural Language. Springer.

Korta, Kepa, Perry, John, 2008. Pragmatics. In: Zalta, Edward N. (Ed.), The Stanford Encyclopedia of Philosophy, Fall 2008 ed.

Krifka, Manfred, 2011. Varieties of semantic evidence. In: Maienborn, Claudia, von Heusinger, Klaus, Portner, Paul (Eds.), Semantics. An International Handbook of Natural Language Meaning, vol. I. Mouton de Gruyter, Berlin, pp. 242–267.

Kripke, Saul, 1972. Naming and necessity. In: Davidson, D., Harman, G.F. (Eds.), Semantics of Natural Language. Reidel, Dordrecht, pp. 253–355 and addendum 763–769.

Lev, Iddo, MacCartney, Bill, Manning, Christopher D., Levy, Roger, 2004. Solving logic puzzles: from robust processing to precise semantics. In: 2nd Workshop on Text Meaning and Interpretation at ACL 2004, pp. 9–16.

Lewis, David, 1970. General semantics. Synthese 22, 18–67.

Link, Godehard, 1998. Algebraic Semantics in Language and Philosophy: CSLI Lecture Notes No. 74. CSLI Publications, Stanford, Calif.

Matthewson, Lisa, 2010. Cross-linguistic variation in modality systems: the role of mood. Semantics and Pragmatics 3, Article 9, 1–74.

Montague, Richard, 1968. Pragmatics. In: Klibanski, R. (Ed.), Contemporary Philosophy. La Nuova Italia Editrice, Florence, pp. 102–121.

Montague, Richard, 1970a. Universal grammar. Theoria 36, 373–398.

Montague, Richard, 1970b. English as a formal language. In: Visentini, Bruno, et al. (Eds.), Linguaggi nella Società e nella Tecnica. Edizioni di Comunità, Milan, pp. 189–224.

Montague, Richard, 1970c. Pragmatics and intensional logic. Synthèse 22, 68–94.

Montague, Richard, 1973. The proper treatment of quantification in ordinary English. In: Hintikka, K.J.J., Moravcsik, J.M.E., Suppes, P. (Eds.), Approaches to Natural Language. Reidel, Dordrecht, pp. 221–242.

Montague, Richard, 1974. Formal Philosophy. Selected Papers of Richard Montague. Edited and with an Introduction by Richmond H. Thomason. Yale University Press, New Haven/London.

Morris, Charles, 1938. Foundations of the theory of signs. In: Neurath, Otto, Carnap, Rudolf, Morris, Charles (Eds.), International Encyclopedia of Unified Science. University of Chicago Press, Chicago, pp. 77–138.

Parsons, Terence, 1972. An Outline of a Semantics of English. University of Massachusetts, Amherst. Ms.

Partee, Barbara H., 1973. Some transformational extensions of Montague grammar. Journal of Philosophical Logic 2, 509–534.

Partee, Barbara H., 1975. Montague grammar and transformational grammar. Linguistic Inquiry 6, 203–300.
Partee, Barbara H. (Ed.), 1976. Montague Grammar. Academic Press, New York.
Partee, Barbara H., Hendriks, Herman L.W., 1997. Montague grammar. In: Benthem, Johan van, Meulen, Alice ter (Eds.), Handbook of Logic and Language. Elsevier/MIT Press, Amsterdam/Cambridge, MA, pp. 5–91.
Philip, William, 1992. Distributivity and logical form in the emergence of universal quantification. In: Barker, Chris, Dowty, David (Eds.), SALT II. Proceedings of the Second Conference on Semantics and Linguistic Theory. The Ohio State University, Department of Linguistics, Columbus, Ohio, pp. 327–346.
Quer, Josep, 2005. Context shift and indexical variables in sign languages. In: Proceedings of SALT XV. CLC Publications.
Rosetta, M.T., 1994. Compositional Translation. Foris, Dordrecht.
Russell, Bertrand, 1957. Mr. Strawson on referring. Mind 66, 385–389.
Schlenker, Philippe, 2011. Quantifiers and variables: insights from sign language (ASL and LSF). In: Partee, B.H., Glanzberg, M., Skilters, J. (Eds.), Formal Semantics and Pragmatics: Discourse, Context, and Models. The Baltic International Yearbook of Cognition, Logic and Communication, vol. 6. New Prairie Press, Manhattan, KS.
Smits, Erik-Jan, Roeper, Tom, Hollebrandse, Bart, 2006. Children's ambiguous understanding of weak and strong quantifiers. In: Anderssen, Merete, Westergaard, Marit R. (Eds.), Nordlyd 34.3: Tromsø University Working Papers on Language and Linguistics: Proceedings from Workshop on Language Acquisition, Scandinavian Conference of Linguistics 22. CASTL, Tromsø, pp. 187–208.
Stalnaker, Robert, 1989. On what's in the head. Philosophical Perspectives 3: Philosophy of Mind and Action Theory 287–316.
Strawson, Peter F., 1950. On referring. Mind 59, 320–344.
Tenenbaum, J.B., Kemp, C., Griffiths, T.L., Goodman, N.D., 2011. How to grow a mind: statistics, structure, and abstraction. Science 331 (6022), 1279–1285.
Thomason, Richmond H., 1976. Some extensions of Montague grammar. In: Partee, B.H. (Ed.), Montague Grammar. Academic Press, New York, pp. 77–118.
Thomason, Richmond H., 1974. Introduction. In: Thomason, R.H. (Ed.), Formal Philosophy. Selected Papers of Richard Montague. Yale University Press, New Haven, pp. 1–69.

Relevant Websites

http://www.sigsem.org – ACL SIGSEM, Special Interest Group on Computational Semantics.
http://elanguage.net/index.php – eLanguage, the Digital Publishing Platform of the Linguistic Society of America.
http://findaid.oac.cdlib.org/findaid/ark:/13030/tf2g50045w/entire_text/ – Finding Aid for the Richard Montague Papers Archived in UCLA Library.
http://linguistlist.org/ – Linguist List, a Multipurpose Resource.
http://www.semanticsarchive.net/ – The Semantics Archive.
http://plato.stanford.edu/ – The Stanford Encyclopedia of Philosophy.
http://semprag.org/ – Semantics and Pragmatics, Open-Access Journal.

Monte Carlo Methods and Bayesian Computation: Importance Sampling

Roman Liesenfeld, Eherhard Karls University, Tubingen, Germany
Jean-Francois Richard, The University of Pittsburgh, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 10000–10004, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Monte Carlo simulation methods can be used to numerically evaluate expectations of functions of random variables (e.g., posterior moments of parameters of interest) for which no analytical expressions are available. They consist of generating random draws from the relevant distribution and replacing expectations by arithmetic means across such draws. Under appropriate technical conditions, the (statistical) accuracy of Monte Carlo estimates is inversely proportional to the number of draws. In cases where a sampling density is not available, which produces sufficiently accurate estimates, one will try to construct an auxiliary sampling density, which is called an importance function. The ratio between the integrand and the selected importance function defines the corresponding remainder function. The initial expectation is then estimated by the arithmetic mean of realizations of the remainder function under random draws from the importance function. An efficient importance function is one which closely resembles the initial integrand under an appropriate metric and, thereby, produces accurate Monte Carlo estimates under as few draws as possible. The article discusses several methods for constructing efficient importance functions at a sufficient level of generality.

General Principle

The term importance sampling (IS) designates a method designed to improve the numerical performance (or efficiency) of Monte Carlo (MC) simulation methods for the evaluation of analytically intractable integrals, generally expectations of functions of random variables. Early descriptions of the method can be found, e.g., in Kahn and Marshall (1953), Trotter and Tukey (1956), or in the commonly cited monograph of Hammersley and Handscomb (1964, Section 5.4). Examples of applications are the numerical evaluation of Bayesian posterior expectations of functions of parameters of interest – see, e.g., Kloek and van Dijk (1978) or Geweke (1989) – and that of a likelihood function in the presence of unobserved latent variables – see, e.g., Durbin and Koopman (1997) or Richard and Zhang (2000).

The general principle of IS can be outlined as follows. Consider a situation where one has to numerically evaluate an integral of the form

$$I = \int_D \phi(x)\mathrm{d}x \qquad [1]$$

where ϕ denotes a real function defined on $D \subset R^p$.

A case of special interest is that where ϕ takes the form of the product of a density function μ with support D and a function h whose expectation has to be evaluated on μ, say

$$I = E_\mu[h(x)] = \int_D h(x)\mu(x)\mathrm{d}x \qquad [2]$$

In the subsequent analysis, μ will be referred to as an 'initial' sampler, to be distinguished from the importance sampler introduced below.

In low-dimensional applications (say $p \leq 5$), one might consider using nonstochastic multivariate quadrature rules, as described, e.g., in Stroud and Secrest (1966) or in Golub and Welsch (1969). However, higher-dimensional applications require the use of MC simulation methods. The two most widely used methods are Monte Carlo Markov chain (MCMC) methods (see Markov Chain Monte Carlo Methods) and IS which is presented here.

IS requires the selection of a class of samplers, indexed by an auxiliary vector of parameters $a \in A$, say $M = \{m(x|a); a \in A\}$. While in some applications M might consist of a single sampler, it is generally advisable to select a class of samplers and then, within that class, a specific sampler under an appropriate 'efficiency' criterion. Selection of M and of a sampler within M are discussed further below, once the purpose of IS has been clarified. In the case of formula [2], it will often be the case that M consists of suitable parametric extensions of μ and, therefore, that μ itself belongs to M, i.e., that there exists $a_0 \in A$ such that $\mu(x) \equiv m(x|a_0)$ on D. While, as discussed further below, such a condition may be instrumental in the construction of M, it is by no means a requirement of the method.

For any given $a \in A$, the integral in formula [1] can be rewritten as

$$I = \int_D \left[\frac{\phi(x)}{m(x|a)}\right] m(x|a)\mathrm{d}x = E_m\left[\frac{\phi(x)}{m(x|a)}\right] \qquad [3]$$

In the case of formula [2], this expectation is often rewritten as

$$I = E_m[h(x)\omega(x;a)] \qquad [4]$$

where

$$\omega(x;a) = \frac{\mu(x)}{m(x|a)} \qquad [5]$$

denotes the weight function associated with the sampler m.

For any given $a \in A$, an MC IS estimate of I is then given by

$$\widehat{I}_S(a) = \frac{1}{S}\sum_{i=1}^{S}\frac{\phi(x)}{m(x|a)}\bigg|_{x=\tilde{x}_i(a)} \qquad [6]$$

where $\{\tilde{x}_i(a); \quad i: 1 \rightarrow S\}$ denotes a set of S identically independently distributed (hereafter i.i.d.) draws from the auxiliary sampler $m(x|a)$. If $m(x|a)$ is such that the sampling variance

$$\sigma^2(a) = \int_D \left[\frac{\phi(x)}{m(x|a)} - I\right]^2 m(x|a)\mathrm{d}x \qquad [7]$$

is finite, then a central limit theorem applies, whereby the random variable $\sqrt{S}(\widehat{I}_S(a) - I)$ converges in distribution towards a Normal with zero mean and variance $\sigma^2(a)$. Let $\widehat{\sigma}_S^2(a)$ denote the MC finite sample variance of the ratio ϕ/m in formula [6]. A numerical estimate of the variance of the estimator $\widehat{I}_S(a)$ is then given by $(1/S)\widehat{\sigma}_S^2(a)$. Note, therefore, that once a sampler has been selected, the numerical accuracy of the corresponding IS estimator of I, as measured by its MC standard deviation is of the order $S^{-1/2}$. Poor selection of a may require a prohibitively large MC sample size in order to achieve sufficient accuracy. Conditions for the finiteness of $\sigma^2(a)$ are discussed, e.g., in Geweke (1996) or Stern (1997). A sufficient condition is that $|\phi/m|$ be bounded above on D. Convergence in the absence of such a condition may be notoriously more difficult to verify.

If there exists $a_* \in A$ such that $\phi(x) = m(x|a_*)$ on D, then $\sigma^2(a_*) = 0$ and $\widehat{I}_S(a_*) = I, \quad \forall S \geq 1$. More generally, one would aim at selecting a value $\widehat{a}$ on A which minimizes the MC sampling variance $\sigma^2(a)$. Since, however, $\sigma^2(a)$ does not have an analytical expression and can only be numerically approximated, the actual choice of $\widehat{a}$ will often be based upon more heuristic considerations – some of which are discussed further below – aiming at constructing importance samplers which 'approximate' $\phi(x)$ on D. It should be noted that the term importance sampling emphasizes the fact that $m(x|a)$ is precisely meant to oversample in the 'important' parts of the domain D, i.e., those parts which contribute most to the value of integral I.

It is, however, useful to draw users' attention to a potentially misleading interpretation of this notion of region of importance. In practice, it is generally the case that large, possibly infinite, values of $\sigma^2(a)$ will not be generated by values of x in the region of importance (where most importance samplers are expected to perform well by design), but instead by values of x which can be very far away in the tails of $m(x|a)$ and which are, therefore, most unlikely to be ever drawn in a finite sample MC simulation. This situation can be particularly pernicious when D is unbounded. Earlier applications of IS to the computation of Bayesian posterior moments did occasionally suffer from such pathologies, especially in situations where non-informative prior distributions were used in combination with a (qualitatively) poorly identified model, see Statistical Identification and Estimability. (Note that in such cases, truncation of the support D of the posterior density may not be an acceptable solution, insofar as the moments of interest might be highly sensitive to such a truncation.) A classical example of nonexistence of $\sigma^2(a)$, which has been widely cited in the early Bayesian literature on IS, is that where ϕ^{-1} has polynomial tails (e.g., if μ in formula [2] takes the form of a Student-t density kernel) and one selects for M a class of Normal samplers. Nonexistence of $\sigma^2(a)$ may be particularly difficult to detect when it originates from such low-probability events. Indications of a potential problem are the empirical findings that the MC sampling variance of $\widehat{I}_S(a)$ does not decrease inversely proportionally to S or, relatedly, in the context of formula [4], that a few draws, often less than a handful, carry excessive relative weights. It is, therefore, absolutely critical to pay very close attention to estimates of the MC sampling variance of $\widehat{I}_S(a)$ and/or to the behavior of the weight function in eqn [4], at least in those cases where the existence of $\sigma^2(a)$ cannot be formally assessed.

Construction of Importance Samplers

Thus the two critical issues in the design of a successful IS implementation are: (a) the selection of an appropriate class M of auxiliary samplers; and (b) the selection within that class of an 'efficient' sampler, i.e., one for which the MC sampling variance $\sigma^2(a)$ is as small as possible. A number of truly innovative proposals to that effect are found in the recent econometric and statistical literature, though there remains a significant lack of generic principles for the construction of efficient importance samplers (one recent exception to this general statement is presented on its own in Section 3 below). As discussed further in the concluding remarks, this lack of generic principles is perceived by some as a significant impediment to routine use of IS. In the following some of the most useful proposals which have been offered in the late twentieth-century literature are briefly presented.

While Tierney and Kadane (1986) do not discuss IS per se, they propose an extension of the Laplace method of integration, as discussed, e.g., in De Bruijn (1958), to construct approximations for Bayesian posterior moments and marginal densities. The very same idea can be used to construct an importance sampler to approximate numerically the exact integrals. See Richard and Zhang (1998) for an application of this principle to a logit model with unobserved heterogeneity. Since, however, the Laplace method essentially serves to construct a local approximation around the mode of $\phi(x)$, it remains important to pay special attention to tail behavior in order to control for $\sigma^2(a)$, as discussed above.

Geweke (1989) pays attention to the minimization of the MC sampling variance $\sigma^2(a)$ within specific families of IS densities, typically multivariate Student-t densities and skewed generalizations thereof, labeled split-t densities. A problem with his approach, which is characteristic of any attempt at minimizing $\sigma^2(a)$, is that the latter depends upon integrals which have themselves to be numerically evaluated (a problem which will be addressed in Section 3 below). In the same paper Geweke also discusses more heuristic procedures for tailoring $m(x|a)$ in ways which proved quite successful in some Bayesian applications. It should also be noted that Geweke's article includes numerous references relative to MC simulation and numerical integration.

Evans (1991) develops adaptive methods which rely upon earlier draws of x to identify large values of the weight function defined in formula [5] and accordingly to modify the importance sampler. While such adaptive methods can occasionally

prove useful, they do suffer from the problem explained earlier, whereby values of x for which the weight function $\omega(x; a)$ is large are typically highly unlikely to be drawn.

Owen and Zhou (2000) discuss improvements of IS technique which are shown to work quite well in low-dimensional applications. One proposal consists of combining IS with the method of 'control variates' to produce an upper bound for the MC sampling variance. (The method of control variates uses knowledge of approximate integrals to I to reduce the MC sampling variance of $\widehat{I}_S(a)$. See Hammersley and Handscomb (1964) or Hendry (1984) for detailed descriptions of the control variate method.) Another extension consists of pooling several IS estimators in order to reduce MC sampling variances.

Durbin and Koopman (1997) apply IS to evaluate the likelihood function of non-Gaussian state space models. Their application is particularly interesting in that it consists of a fairly sophisticated and, in a sense, indirect application of IS in dimensions which are significantly larger than those of the applications discussed above. Let $p(y|\lambda,\theta)$ denote a non-Gaussian density for a vector y of observables, conditional on a vector λ of unobserved variables and a vector θ of unknown parameters. Let $v(\lambda|\theta)$ denote the density of λ given θ. Finally, let $g(y|\lambda,\theta)$ denote a Gaussian approximation to $p(y|\lambda,\theta)$. Under g one can use standard Kalman filter techniques, as discussed, e.g., in Harvey (1990), to compute a likelihood function $L_g(\theta)$ and, in parallel, a density $v_g(\lambda|y,\theta)$ for the latent variables. It is then trivial to show that the actual likelihood function.

$$L_p(\theta) = \int p(y|\lambda,\theta)v(\lambda|\theta)\mathrm{d}\lambda \qquad [8]$$

can be rewritten as

$$L_p(\theta) = L_g(\theta)\int \frac{p(y|\lambda,\theta)}{g(y|\lambda,\theta)}v_g(\lambda|y,\theta)\mathrm{d}\lambda \qquad [9]$$

The advantage of formula [9] over formula [8] is that it relies upon numerical integration to evaluate departures from the Gaussian likelihood L_g, rather than the likelihood itself. The density v_g in formula [9] plays exactly the role of an importance sampler. Durbin and Koopman (1997) then discuss the construction of the approximating Gaussian model in order to secure maximal numerical accuracy. The interest of that application is to show that the selection of an importance sampler can be approached indirectly via the construction of an operational approximation to a complex model.

Another sophisticated implementation of IS is found in Madras and Piccioni (1999) where the authors propose to provide greater flexibility in the choice of the class of samplers M, often a sticky problem in applications of IS, by allowing $m(x|a)$ to be the (implicit) equilibrium distribution associated with a Monte Carlo Markov chain simulator (see Monte Carlo Methods and Bayesian Computation: Overview; Markov Chain Monte Carlo Methods). The main advantage of this extension lies in the fact, discussed elsewhere, that MCMC simulators are very flexible. Its potential drawback is that its convergence properties are typically more difficult to assess than those of a conventional importance sampler.

An approach which also combines IS and MCMC is used by DiMatteo and Kadane (2001). They employ a standard MCMC technique to compute posterior moments for a given prior. Then, in order to assess robustness relative to the prior, they propose to rerun the analysis under different priors using the initial MCMC sampler as importance sampler.

An important message which emerges from this discussion is that efficient applications of IS have to be carefully tailored to the problem under consideration. This has proved to be a significant obstacle to routine applications of IS and is especially true for Bayesian applications which can produce notoriously 'ill-behaved' and, therefore, difficult to approximate posterior densities. Actually, Bayesian statisticians have largely abandoned IS in favor of MCMC. It should be noted, however, that while MCMC can be more flexible than IS, its convergence is not easily assessed in practice and may require large numbers of MC draws. In fact, both methods can occasionally produce spectacular failures if not carefully designed and controlled.

Efficient Importance Sampling

Expanding upon an earlier contribution by Danielsson and Richard (1993), Richard and Zhang (1998) have recently proposed a generic principle for constructing efficient importance samplers in cases where ϕ is a positive function. For ease of presentation, the method will be introduced here without reference to dimensionality. A straightforward rearrangement of terms in formula [7] produces the following expression:

$$\sigma^2(a) = \int h\left[d^2(x;a)\right]\phi(x)\mathrm{d}x \qquad [10]$$

where

$$d(x;a) = \ln\left[\frac{\phi(x)}{m(x|a)I}\right] \qquad [11]$$

$$h(c) = \mathrm{e}^{\sqrt{c}} + \mathrm{e}^{-\sqrt{c}} - 2 \qquad [12]$$

Note that h is monotone convex on R_+ and that $h(c) \geq c$. Minimization of $\sigma^2(a)$ with respect to a requires nonlinear optimization. Consider, therefore, the simpler function

$$q(a) = \int [\ln\phi(x) - c - \ln m(x|a)]^2\phi(x)\mathrm{d}x \qquad [13]$$

where c stands for an intercept in place of the unknown $\ln I$. Next, one selects an initial sampler, say $m(x|a_0)$ – it could be $\mu(x)$ in formula [2], or any other sampler based, for example on a local approximation of $\phi(x)$ of the sort used by Tierney and Kadane (1986). An MC approximation to $q(a)$ is then given by

$$\widehat{q}_R(a) = \frac{1}{R}\sum_{i=1}^{R}[\ln\phi(x) - c - \ln m(x|a)]^2\omega_0(x)\big|_{x=\tilde{x}_i^0} \qquad [14]$$

where

$$\omega_0(x) = \frac{\phi(x)}{m(x|a_0)} \qquad [15]$$

and $\{\tilde{x}_i^0; i: 1 \to R\}$ denotes a set of R i.i.d. draws from $m(x|a_0)$. If, furthermore, $m(x|a)$ belongs to the exponential family of distributions, then minimization of $\widehat{q}_R(a)$ with respect to a takes the form of a simple weighted linear least squares problem. It may be advisable to omit the weight function $\omega_0(x)$

if it exhibits excessive variance in cases where $m(x|a_0)$ is found to provide a poor global approximation to $\phi(x)$. The method can be iterated by using as initial sampler in any given round the optimized sampler from the previous round. Experience suggests that no more than 3 to 4 iterations are required to produce a stable solution, irrespective of the initial sampler $m(x|a_0)$. In order to assess how much efficiency is lost by minimizing $q(a)$ instead of $\sigma^2(a)$, let $\widehat{a}$ and a_* denote the minimizers of q and σ^2, respectively. Taking advantage of the convexity of h, Jensen's inequality implies that

$$\sigma^2(\widehat{a}) \geq \sigma^2(a_*) \geq h[q(a_*)] \geq h[q(\widehat{a})] \qquad [16]$$

MC estimates of the two bounds are immediate byproducts of the minimization of $\widehat{q}_R(a)$. Unless they were found to be far apart, there would be no reasons for solving the more complex minimization problem associated with $\sigma^2(a)$.

The method is extremely accurate and fast in all cases where the integrand ϕ can be well approximated by a sampler from the exponential family of distributions. The real strength of the method lies in very high-dimensional (1000^+) integration of 'well-behaved' integrands. High-dimensional minimization of $\widehat{q}_R(a)$ does require an additional factorization which is not discussed here. The final algorithm is one whereby minimization is recursively carried out one dimension at a time. It has proved useful for likelihood evaluation in dynamic latent variable models such as the one discussed earlier in Durbin and Koopman (1997). Liesenfeld and Richard (2000) successfully apply Efficient Importance Sampling (EIS) to estimate parametric as well as semi-nonparametric dynamic stochastic volatility models for three daily financial series (IBM stock prices, S&P 500 stock index, and Dollar/Deutsche Mark exchange rate). Dimensions of integration are of the order of 4700. In each case highly accurate estimates of the likelihood function are obtained with as few as $R = S = 50$ MC draws and each individual likelihood evaluation requires on the order of 8 s on a Pentium II 466 MHz notebook.

Such outstanding performance originates from the fact that these high-dimensional integrands turn out to be well-behaved, reflecting the fact that the data happen to be quite informative on the unobserved volatilities. A pilot application to a logit panel model with random effects along both dimensions had proved equally successful (see Richard and Zhang, 1998 for details).

Smooth Functional Integration

In numerous applications, such as likelihood function evaluation, the integral to be evaluated is itself a function of some argument δ, say

$$I(\delta) = \int_{D(\delta)} \phi(x;\delta)\mathrm{d}x, \quad \delta \in \Delta \qquad [17]$$

The argument δ is then carried along in all the formulae previously discussed. Most importantly, one cannot expect a single sampler to be efficient for all δ's. Hence, IS requires the selection of a particular a, say $a(\delta)$, for each relevant value of δ. If, furthermore, $I(\delta)$ has to be optimized with respect to δ, then smoothness of the functional approximation $\widehat{I}_S(\delta)$ becomes critical and cannot be secured with independent pointwise MC estimation, however accurate.

In such cases, it becomes essential that all IS sequences $\{\tilde{x}_i(a(\delta));\ \ i:1\rightarrow S\}$ be obtained by transformation of a common sequence of random draws $\{u_i;\ i:\ 1 \rightarrow S\}$ from a distribution which does not depend on a, say

$$\tilde{x}_i(a) = H(u_i;a), \quad i:1\rightarrow S, \quad \forall a \in A \qquad [18]$$

The u_i's, the so-called common random numbers, would typically be uniforms (in which case H denotes the inverse of the distribution function of x given a) or standardized normals (when x given a is itself normally distributed). See, e.g., Devroye (1986) for detailed descriptions of such transformations. Note, in particular, that inversion of a distribution function can be a computer-intensive operation but, nevertheless, is indispensable for smoothness. (The use of carefully designed interpolation techniques can significantly reduce computing times.) It should be noted that smoothness is not only required for numerical convergence of an optimization algorithm but also plays a central role in the derivation of the asymptotic properties of a broad range of simulation estimators. See, e.g., Pakes and Pollard (1989) or Gouriéroux and Monfort (1996).

Conclusion

IS is one of the key techniques currently available for the evaluation of analytically intractable integrals by MC simulation. It performs at its best when used in combination with 'acceleration' techniques such as, e.g., control variates, mixtures, common random numbers. Very high-dimensional (1000^+) integration is feasible as long as the integrand is reasonably well-behaved. Further developments are required to guide the design of efficient IS samplers. IS is not an all-purpose technique and lack of flexibility in the construction of samplers may complicate its application to ill-behaved integrands, for which alternative methods of integration such as MCMC might prove more useful. Nevertheless, IS has proved to be a remarkably useful technique for a wide range of applications including very high-dimensional integrations.

Bibliography

Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M., 2003. The variational Bayesian Em algorithm for incomplete data: with application to scoring graphical model structures. Bayesian Statistics 7, 453–464.

Bratley, P., Fox, B.L., Schrage, L.E., 1987. A Guide to Simulation, second edn. Springer Verlag, New York.

Danielsson, J., Richard, J.F., 1993. Accelerated Gaussian importance sampler with applications to dynamic latent variable models. Journal of Applied Econometrics 8, s153–s173.

De Bruijn, N.G., 1958. Asymptotic Methods in Analysis. North Holland, Amsterdam.

Devroye, L., 1986. Non-Uniform Random Variate Generation. Springer Verlag, New York.

DiMatteo, I., Kadane, J., 2001. Vote tampering in a district Justice election in Beaver county, Pennsylvania. Journal of the American Statistical Association 96, 510–518.

Durbin, J., Koopman, S.J., 1997. Monte Carlo maximum likelihood estimation for non-Gaussian state space models. Biometrika 84, 669–684.

Evans, M., 1991. Adaptive importance sampling and chaining. Contemporary Mathematics (Statistical Multiple Integration) 115, 137–142.

Geweke, J., 1989. Bayesian inference in econometric models using Monte Carlo integration. Econometrica 57, 1317–1339.

Geweke, J., 1996. Monte Carlo simulation and numerical integration. In: Amman, H., Kendrick, D., Rust, J. (Eds.), The Handbook of Computational Economics, vol. 1. North Holland, Amsterdam, pp. 731–800.

Golub, G.H., Welsch, J.H., 1969. Calculation of Gaussian quadrature rules. Mathematics of Computation 23, 221–230.

Gouriéroux, C., Monfort, A., 1996. Simulation Based Econometric Methods. In: CORE Lecture series. Oxford University Press, Oxford, UK.

Hammersley, J.M., Handscomb, D.C., 1964. Monte Carlo Methods. Methuen, London.

Harvey, A.C., 1990. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge, UK.

Hendry, D.F., 1984. Monte Carlo experimentation in econometrics. In: Griliches, Z., Intriligator, M.D. (Eds.), The Handbook of Econometrics. North Holland, Amsterdam.

Hesterberg, T., 1995. Weighted average importance sampling and defensive mixture distributions. Technometrics 37, 185–194.

Kahn, H., Marshall, A., 1953. Methods of reducing sample size in Monte Carlo computations. Journal of the Operations Research Society of America 1, 263–278.

Kloek, T., van Dijk, H.K., 1978. Bayesian estimates of equation system parameters: application of integration by Monte Carlo. Econometrica 46, 1–19.

Liesenfeld, R., Richard, J.-F., 2000. Univariate and multivariate stochastic volatility models: estimation and diagnostics. Mimeo, Eberhard-Karls-Universität, Tübingen, Germany.

Madras, N., Piccioni, M., 1999. Importance sampling for families of distributions. The Annals of Applied Probability 9, 1202–1225.

Neal, Radford M., 2001. Annealed importance sampling. Statistics and Computing 11 (2), 125–139.

Owen, A., Zhou, Y., 2000. Safe and effective importance sampling. Journal of the American Statistical Association 95, 135–143.

Pakes, A., Pollard, D., 1989. Simulation and the asymptotics of optimization estimators. Econometrica 57, 1027–1057.

Raftery, Adrian E., Bao, Le, 2010. Estimating and projecting trends in HIV/AIDS generalized epidemics using incremental mixture importance sampling. Biometrics 66 (4), 1162–1173.

Richard, J.-F., Zhang, W., 1998. Efficient High-Dimensional Monte Carlo Importance Sampling. Mimeo, University of Pittsburgh, PA.

Richard, J.-F., Zhang, W., 2000. Accelerated Monte Carlo integration: an application to dynamic latent variables models. In: Mariano, R., Schuermann, T., Weeks, M. (Eds.), Simulation-Based Inference in Econometrics: Methods and Applications. Cambridge University Press, Cambridge, UK.

Stern, S., 1997. Simulation-based estimation. Journal of Economic Literature 35, 2006–2039.

Stroud, A.H., Secrest, D.H., 1966. Gaussian Quadrature Formulae. Prentice-Hall, Englewood Cliffs, NJ.

Tierney, L., Kadane, J.B., 1986. Accurate approximations for posterior moments and marginal densities. Journal of American Statistical Association 81, 82–86.

Trotter, H.F., Tukey, J.W., 1956. Conditional Monte Carlo for normal samples. In: Mayer, H.A. (Ed.), Symposium on Monte Carlo Methods. Wiley, New York.

Yuan, Changhe, Druzdzel, Marek J., 2006. Importance sampling algorithms for Bayesian networks: principles and performance. Mathematical and Computer Modelling 43 (9), 1189–1207.

Monte Carlo Methods and Bayesian Computation: Overview

Siddhartha Chib, Washington University, St. Louis, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 15, pp. 10004–10009, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

The recent growth of Bayesian methods in statistics and the applied sciences has been heavily influenced by developments in Monte Carlo computational methods. This article features a summary of the main tools with particular emphasis on the dominant role played by Markov chain Monte Carlo methods for dealing with practically all aspects of Bayesian inference, from estimation of parameters, to prediction and model choice. Selected applications that attest to the continuing importance and viability of these methods are referenced.

Introduction

Monte Carlo simulation methods and, in particular, Markov chain Monte Carlo methods, play a large and prominent role in the practice of Bayesian statistics, where these methods are used to summarize the posterior distributions that arise in the context of the Bayesian prior–posterior analysis. Monte Carlo methods are used in practically all aspects of Bayesian inference, for example, in the context of prediction problems and in the computation of quantities, such as the marginal likelihood, that are used for comparing competing Bayesian models.

Suppose that $\pi(\psi|y,\mathscr{M}) \propto p(\psi|\mathscr{M})f(y|\psi,\mathscr{M})$ is the posterior density for a set of parameters $\psi \in \mathscr{R}^d$ in a particular Bayesian model $p(\psi|\mathscr{M})$ and sampling density or likelihood function $f(y|\psi,\mathscr{M})$ and that interest centers on the posterior mean $\eta = \int_{\mathscr{R}^d} \psi\pi(\psi|y,\mathscr{M})\mathrm{d}\psi$. As an illustration of this general setting, consider binary data $y_i \in \{0, 1\}$ modeled by the probabilities $\Pr(y_i = 1|\psi) = \Phi(x_i < \psi)$, where Φ is the standard normal cumulative distribution function, x_i are covariates and ψ is an unknown vector of parameters. The posterior distribution of ψ given n independent measurement $y = (y_1,\ldots,y_n)$ is proportional to

$$\pi(\psi|y,\mathscr{M}) \propto p(\psi|\mathscr{M}) \prod_{i=1}^{n} \Phi\left(x_i'\psi\right)^{y_i} \left(1 - \Phi\left(x_i'\psi\right)\right)^{(1-y_i)}$$

and the goal is to summarize this density.

Now suppose that a sequence of variates $\{\psi^{(1)},\ldots,\psi^{(M)}\}$ drawn from $\pi(\psi|y,\mathscr{M})$ are available and one computes the sample average of these draws. Under any set of conditions that allow use of the strong law of large numbers, it follows that the sample average is simulation consistent, converging to the posterior mean as the simulation sample M goes to infinity. This then is the essence of the Monte Carlo method: to compute an expectation, or to summarize a given density, sample the density by any method that is appropriate and use the sampled draws to estimate the expectation and to summarize the density.

The sampling of the posterior distribution is, therefore, the central focus of Bayesian computation. One important breakthrough in the use of simulation methods was the realization that the sampled draws need not be independent, that simulation consistency can be achieved with correlated draws. The fact that the sampled variates can be correlated is of immense practical and theoretical importance and is the defining characteristic of Markov chain Monte Carlo methods, popularly referred to by the acronym MCMC, where the sampled draws form a Markov chain (*see* Markov Chain Monte Carlo Methods). The idea behind these methods is simple and extremely general. In order to sample a given probability distribution, referred to as the target distribution, a suitable Markov chain is constructed with the property that its limiting, invariant distribution is the target distribution. Once the Markov chain has been constructed, a sample of draws from the target distribution is obtained by simulating the Markov chain a large number of times and recording its values. Within the Bayesian framework, where both parameters and data are treated as random variables, and inferences about the parameters are conducted conditioned on the data, the posterior distribution of the parameters provides a natural target for MCMC methods.

Markov chain sampling methods originate with the work of Metropolis et al. (1953) in statistical physics. A crucial and landmark extension of the method was made by Hastings (1970) leading to a method that is now called the Metropolis–Hastings algorithm. This algorithm was first applied to problems in spatial statistics and image analysis (Besag, 1974). A resurgence of interest in MCMC methods started with the papers of D. Geman and S. Geman (1984) who developed an algorithm, a special case of the Metropolis method that later came to be called the Gibbs sampler, to sample a discrete distribution, Tanner and Wong (1987) who proposed a MCMC scheme involving 'data augmentation' to sample posterior distributions in missing data problems, and Gelfand and Smith (1990) where the value of the Gibbs sampler was demonstrated for general Bayesian inference with continuous parameter spaces.

Markov Chain Monte Carlo Methods

The Metropolis–Hastings Algorithm

This powerful algorithm provides a general approach for producing a correlated sequence of draws from the target density $\pi(\psi)$ that may be difficult to sample by a classical independence method. As before, let $\pi(\psi)$ denote the density

that one wishes to sample, where one can assume for simplicity that the density is absolutely continuous; and let $q(\psi, \psi')$ denote a source density for a candidate draw ψ' given the current value ψ in the sampled sequence. The density $q(\psi, \psi')$ is referred to as the proposal or candidate-generating density. Then, the M–H algorithm is defined by two steps: a first step in which a proposal value is drawn from the candidate-generating density and a second step in which the proposal value is accepted or rejected according to a probability given by

$$\alpha(\psi,\psi') = \begin{cases} \min\left[\dfrac{\pi(\psi')q(\psi',\psi)}{\pi(\psi)q(\psi,\psi')}, 1\right] & \text{if } \pi(\psi)q(\psi,\psi') > 0 \\ 1 & \text{otherwise} \end{cases} \quad [1]$$

If the proposal value is rejected, then the next sampled value is taken to be the current value; which means that when a rejection occurs the current value is repeated. Given the new value, the same two-step process is repeated and the cycle iterated a large number of times. A full expository discussion of this algorithm, along with a derivation of the method from the logic of reversibility, is provided by Chib and Greenberg (1995).

The M–H method delivers variates from π under quite general conditions. A weak requirement for a law of large numbers for sample averages based on the M–H output involve positivity and continuity of $q(\psi, \psi')$ for (ψ, ψ'), and connectedness of the support of the target distribution. In addition, if π is bounded then conditions for ergodicity, required to establish the central limit theorem, are satisfied. A discussion of the conditions and the underlying Markov chain theory that justifies these methods was first provided by Tierney (1994) and summarized recently by Robert and Casella (1999). Of course, the variates are from π only in the limit as the number of iterations becomes large but, in practice, after an initial burn-in phase consisting of (say) n_0 iterations, the chain is assumed to have converged and subsequent values are taken as approximate draws from π. Theoretical calculation of the burn-in is not easy, so it is important that the proposal density be chosen to ensure that the chain makes large moves through the support of the target distribution without staying at one place for many iterations. Generally, the empirical behavior of the M–H output is monitored by the autocorrelation time of each component of ψ defined as

$$\left\{1 + 2\sum_{x=1}^{\mathscr{M}}\left(1 - \frac{s}{\mathscr{M}}\right)\rho_{ks}\right\} \quad [2]$$

where ρ_{ks} is the sample autocorrelation at lag s for the kth component of ψ, and by the acceptance rate which is the proportion of times a move is made as the sampling proceeds. Because independence sampling produces an autocorrelation time that is theoretically equal to one, and Markov chain sampling produces autocorrelation times that are bigger than one, it is desirable to tune the M–H algorithm to get values close to one, if possible.

Given the form of the acceptance probability $\alpha(\psi, \psi')$ it is clear that the M–H algorithm does not require knowledge of the normalizing constant of $\pi(\cdot)$. Furthermore, if the proposal density satisfies the symmetry condition $q(\psi, \psi') = q(\psi', \psi)$, the acceptance probability reduces to $\pi(\psi')/\pi(\psi)$; hence, if $\pi(\psi') \geq \pi(\psi)$, the chain moves to ψ', otherwise it moves to ψ' with probability given by $\pi(\psi')/\pi(\psi)$. The latter is the algorithm originally proposed by Metropolis et al. (1953).

Different proposal densities give rise to specific versions of the M–H algorithm, each with the correct invariant distribution π. One family of candidate-generating densities is given by $q(\psi, \psi') = q(\psi' - \psi)$. The candidate ψ' is thus drawn according to the process $\psi' = \psi + z$, where z follows the distribution q, and is referred to as the *random walk M–H* chain. The random walk M–H chain is perhaps the simplest version of the M–H algorithm and is quite popular in applications. One has to be careful, however, in setting the variance of z because if it is too large it is possible that the chain may remain stuck at a particular value for many iterations while if it is too small the chain will tend to make small moves and move inefficiently through the support of the target distribution.

Hastings (1970) considers a second family of candidate-generating densities that are given by the form $q(\psi, \psi') = q(y)$. Tierney (1994) refers to this as an *independence M–H chain* because the candidates are drawn independently of the current location ψ. Chib and Greenberg (1994, 1995) implement such chains by matching the proposal density to the target at the mode by a multivariate normal or multivariate-t distribution with location given by the mode of the target and the dispersion given by inverse of the Hessian evaluated at the mode.

The Multiple-block M–H Algorithm

In applications when the dimension of ψ is large it may be preferable, as discussed by Hastings (1970), to construct the Markov chain simulation by first grouping the variables ψ into smaller blocks. For simplicity suppose that two blocks are adequate and that ψ is written as (ψ_1, ψ_2), with $\psi_k \in \Omega_k \subseteq \mathscr{R}^{d_k}$. In that case the M–H algorithm requires the specification of two proposal densities,

$$q_1(\psi_1, \psi_1' | \psi_2); \quad q_2(\psi_2, \psi_2' | \psi_1)$$

one for each block ψ_k, where the proposal density q_k may depend on the current value of the remaining block. Also, define

$$\alpha(\psi_1, \psi_1' | \psi_2) = \min\left\{\frac{\pi(\psi_1', \psi_2)q_1(\psi_1', \psi_1 | \psi_2)}{\pi(\psi_1, \psi_2)q_1(\psi_1, \psi_1' | \psi_2)}, 1\right\} \quad [3]$$

and

$$\alpha(\psi_2, \psi_2' | \psi_1) = \min\left\{\frac{\pi(\psi_1, \psi_2')q_2(\psi_2', \psi_2 | \psi_1)}{\pi(\psi_1, \psi_2)q_2(\psi_2, \psi_2' | \psi_1)}, 1\right\} \quad [4]$$

as the probability of move for block ψ_k conditioned on the other block. Then, in what may be called the multiple-block M–H algorithm, one cycle of the algorithm is completed by updating each block, say sequentially in fixed order, using a M–H step with the above probability of move, given the most current value of the other block.

The extension of this method to more than two blocks is straightforward.

The Gibbs Sampling Algorithm

Another MCMC method, which is a special case of the multiple-block M–H method, is called the Gibbs sampling

method and was brought into statistical prominence by Gelfand and Smith (1990). An elementary introduction to Gibbs sampling is provided by Casella and George (1992). To describe this algorithm, suppose that the parameters are grouped into two blocks (ψ_1, ψ_2) and each block is sampled according to the *full conditional distribution* of block ψ_k, defined as the conditional distribution under π of ψ_k given the other block. In parallel with the multiple-block M–H algorithm, the most current value of the other block is used in deriving the full conditional distribution. Derivation of these full conditional distributions is usually quite simple since, by Bayes' theorem, each full conditional distribution $\pi(\psi_1|\psi_2)$ and $\pi(\psi_2|\psi_1)$ are both proportional to $\pi(\psi_1,\psi_2)$, the joint distribution of the two blocks. In addition, the powerful device of data augmentation, due to Tanner and Wong (1987), in which latent or auxiliary variables are artificially introduced into the sampling, is often used to simplify the derivation and sampling of the full conditional distributions. For further details *see* Markov Chain Monte Carlo Methods.

Implementation Issues

In implementing a MCMC method it is important to assess the performance of the sampling algorithm to determine the rate of mixing and the size of the burn-in (both having implications for the number of iterations required to get reliable answers). A large literature has now emerged on these issues, for example, Tanner (1996, Section 6.3), Cowles and Carlin (1996), Gamerman (1997, Section 5.4), and Robert and Casella (1999).

Writing at the turn of the millenium, the more popular approach is to utilize the sampled draws to assess both the performance of the algorithm and its approach to the stationary, invariant distribution. Several such relatively informal methods are now available. Perhaps the simplest and most direct are autocorrelation plots (and autocorrelation times) of the sampled output. Slowly decaying correlations indicate problems with the mixing of the chain. It is also useful in connection with M–H Markov chains to monitor the acceptance rate of the proposal values with low rates implying 'stickiness' in the sampled values and thus a slower approach to the invariant distribution.

Somewhat more formal sample-based diagnostics are also available in the literature, as summarized in the routines provided by Best et al. (1995) and now packaged in the BOA program. Although these diagnostics often go under the name 'convergence diagnostics' they are in principle approaches that detect *lack* of convergence. Detection of convergence based entirely on the sampled output, without analysis of the target distribution, is extremely difficult and perhaps impossible. Cowles and Carlin (1996) discuss and evaluate thirteen such diagnostics (for example, those proposed by Geweke (1992), Gelman and Rubin (1992), Zellner and Min (1995), amongst others) without arriving at a consensus.

While implementing MCMC methods it is important to construct samplers that mix well (mixing being measured by the autocorrelation times) because such samplers can be expected to converge more quickly to the invariant distribution. As a general rule, sets of parameters that are highly correlated should be treated as one block when applying the multiple-block M–H algorithm. Otherwise, it would be difficult to develop proposal densities that lead to large moves through the support of the target distribution and the sampled draws would tend to display autocorrelations that decay slowly. Liu (1994) and Liu et al. (1994) discuss possible strategies in the context of three-block Gibbs MCMC chains. Roberts and Sahu (1997) provide further discussion of the role of blocking in the context of Gibbs Markov chains.

Additional Uses of MCMC Methods in Bayesian Inference

MCMC methods provide an elegant solution to the problem of prediction in the Bayesian context. The goal is to summarize the Bayesian predictive density defined as

$$f(y_f|y,\mathscr{M}) = \int f(y_f|y,\mathscr{M},\theta)\pi(\theta|y,\mathscr{M})\mathrm{d}\theta, \qquad [5]$$

where $f(y_f|y,\mathscr{M},\theta)$ is the conditional density of future observations y_f given $(y,\mathscr{M},\theta)$. In general, the predictive density is not available in closed form. However, this is not a problem when a sample of draws from the posterior density is available. If one simulates

$$y_f^{(i)} \sim f\left(y_f|y,\mathscr{M},\theta^{(j)}\right) \qquad [6]$$

for each sampled draw $\theta^{(j)}$, then the collection of simulated values $\{y_f^{(1)},\ldots,y_f^{(M)}\}$ is a sample from the Bayes prediction density $f(y_f|y,\mathscr{M})$. The simulated sample can be summarized in the usual way by the computation of sample averages and quantiles.

MCMC methods have also been widely applied to the problem of the model choice. Suppose that there are K possible models $\mathscr{M}_1,\ldots,\mathscr{M}_K$ for the observed data defined by the sampling densities $\{f(y|\theta_k,\mathscr{M}_k)\}$ and proper prior densities $\{p(\theta_k|\mathscr{M}_k)\}$ and the objective is to find the evidence in the data for the different models. In the Bayesian approach this question is answered by placing prior probabilities $\Pr(\mathscr{M}_k)$ on each of the K models and using the Bayes calculus to find the posterior probabilities $\{\Pr(\mathscr{M}_1|y),\ldots,\Pr(\mathscr{M}_K|y)\}$ conditioned on the data but marginalized over the unknowns θ_k (Jeffreys, 1961 and Kass and Raftery, 1995). Specifically, the posterior probability of $\mathscr{M}_k$ is given by the expression

$$\Pr(\mathscr{M}_k|y) = \frac{\Pr(\mathscr{M}_k)m(y|\mathscr{M}_k)}{\sum_{i=1}^{K}\Pr(\mathscr{M}_l)m(y|\mathscr{M}_l)} \propto \Pr(\mathscr{M}_k)m(y|\mathscr{M}_k), (k \le K) \qquad [7]$$

where $m(y|\mathscr{M}_k)$ is the marginal likelihood of $\mathscr{M}_k$.

A central problem in estimating the marginal likelihood is that it is an integral of the sampling density over the prior distribution of θ_k. Thus, MCMC methods, which deliver sample values from the posterior density, cannot be used to directly average the sampling density. In addition, taking draws from the prior density to do the averaging produces an estimate that is simulation consistent but highly inefficient because draws from the prior density are not likely to be in high density regions of the sampling density $f(y|\theta_k,\mathscr{M}_k)$. A natural way to correct this problem is by the method of importance sampling.

If we let $h(\theta_k, \mathscr{M}_k)$ denote a suitable importance sampling function, then the marginal likelihood can be estimated as

$$\widehat{m}_1(y|\mathscr{M}_k) = M^{-1}\sum_{j=1}^{M}\frac{f\left(y\middle|\theta_k^{(j)},\mathscr{M}_k\right)p\left(\theta_k^{(j)}\middle|\mathscr{M}_k\right)}{h\left(\theta_k^{(j)}\middle|\mathscr{M}_k\right)} \quad [8]$$

$$\theta_k^{(j)} \sim h\left(\theta_k^{(j)}\middle|\mathscr{M}_k\right) \quad (j \le M)$$

This method is useful when it can be shown that the ratio is bounded, which can be difficult to check in practice, and when the sampling density is not expensive to compute; this, unfortunately, is often not so. Gelfand and Dey (1994) propose an estimator that is based on the posterior simulation output. Another method for finding the marginal likelihood is due to Chib (1995). The marginal likelihood, by virtue of being the normalizing constant of the posterior density, can be expressed as

$$m(y|\mathscr{M}_k) = \frac{f(y|\theta_k,\mathscr{M}_k)p(\theta_k|\mathscr{M}_k)}{\pi(\theta_k|y,\mathscr{M}_k)} \quad [9]$$

which is an identity in θ_k.Based on this expression an estimate of the marginal likelihood on the log-scale is given by

$$\log \widehat{m}(y|\mathscr{M}_k) = \log f\left(y\middle|\theta_k^*,\mathscr{M}_k\right) + \log p\left(\theta_k^*\middle|\mathscr{M}_k\right) - \log \widehat{\pi}\left(\theta_k^*\middle|y,\mathscr{M}_k\right) \quad [10]$$

where θ_k^* denotes an arbitrarily chosen point and $\widehat{\pi}(\theta_k^*|y,\mathscr{M}_k)$ is the estimate of the posterior density at that single point. To estimate this posterior ordinate Chib (1995) utilizes the Gibbs output in conjunction with a decomposition of the ordinate into marginal and conditional components. Chib and Jeliazkov (2001) extend that approach for output produced by the Metropolis–Hastings algorithm. A comparison of different estimators of the marginal likelihood is provided by Chen and Shao (1997).

When one is presented with a large collection of candidate models, each with parameters $\theta_k \in B_k \subseteq R^{d_k}$, direct fitting of each model to find the marginal likelihood can be computationally expensive. In such cases it may be more fruitful to utilize model space–parameter space MCMC algorithms that eschew direct fitting of each model for an alternative simulation of a 'mega model' where a model index random variable, denoted as $\mathscr{M}$, taking values on the integers from 1 to K, is sampled in tandem with the parameters. The posterior distribution of $\mathscr{M}$ is then computed as the frequency of times each model is visited. Methods for doing this have been proposed by Carlin and Chib (1995) and Green (1995) which are discussed in article 2.1.76. Both methods are closely related as shown by Dellaportas et al. (1998) and Godsill (1998). The problem of variable selection in regression has also been tackled in this context starting with George and McCulloch (1993).

An Example

Markov chain Monte Carlo methods have proved enormously popular in Bayesian statistics (see for example, Gilks et al. 1996) where these methods have opened up vistas that were unimaginable fifteen years ago.

As an illustration of how MCMC methods are designed in a Bayesian context, we return to the binary data problem discussed in the introduction and describe the approach of Albert and Chib (1993) that has found many applications. Recall that the posterior distribution of interest is proportional to

$$\pi(\beta|y,\mathscr{M}) \propto p(\beta|\mathscr{M})\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{(1-y_i)} \quad [11]$$

where $p_i = \Phi(x_i'\psi)$. To analyze this probit model, and other models for binary and ordinal data, Albert and Chib introduce latent variables and write $z_i = x_i'\psi + \varepsilon_i$ where ε_i is $\mathscr{N}(0, 1)$ and let $y_i = I[z_i > 0]$ where $I[\cdot]$ is the indicator function. This two-equation specification is equivalent to the binary probit regression model since $\Pr(y_i = 1|\psi) = \Pr(z_i > 0|\psi) = \Phi(x_i'\psi)$, as required. The Albert–Chib algorithm now consists of simulating the joint posterior distribution of (ψ, Z), where $Z = (z_1,\ldots,z_n)$. The MCMC simulation is implemented by sampling the conditional distributions $[\psi|y, Z]$ and $[Z|y, \psi]$. The first of these distributions depends only on Z (and not on y) and is obtained by standard Bayesian results on normal linear models. The second factors into a product of independent distributions $[z_i|y_i, \psi]$. The knowledge of y_i merely restricts the support of the normal distribution $z_i|\psi$; if $y_i = 0$, then $z_i \le 0$ and if $y_i = 1$, then $z_i 0$. Suppose that a priori $\psi \sim \mathscr{N}_k(\psi_0, C_0)$, where ψ_0 and C_0 are known hyperparameters. Then, a sample of ψ and Z from the posterior distribution is obtained by iterating on the following steps:

a. sample ψ from the distribution $\mathscr{N}_k(\widehat{\psi}, C)$, where $\widehat{\psi} = C\left(C_0^{-1}\psi_0 + \sum_{i=1}^{n} x_i z_i\right), C = \left(B_0^{-1} + \sum_{i=1}^{n} x_i x_i'\right)^{-1}$;
b. sample z_i from $\mathscr{T}\mathscr{N}_{(-\infty,0)}(x_i'\psi, 1)$ if $y_i = 0$ and from $\mathscr{T}\mathscr{N}_{(0,\infty)}(x_i'\psi, 1)$ if $y_i = 1$, where $\mathscr{T}\mathscr{N}_{(a,b)}(\mu, \sigma^2)$ denotes the $\mathscr{N}(\mu, \sigma^2)$ distribution truncated to the region (a, b).

At this time, considerable work on MCMC methods and applications continues to emerge, a testament to the importance and viability of MCMC methods in Bayesian analysis.

See also: Bayesian Statistics; Markov Chain Monte Carlo Methods; Markov Models and Social Analysis.

Bibliography

Albert, J.H., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88, 669–679.
Alfaro, Michael E., Zoller, Stefan, Lutzoni, François, 2003. Bayes or Bootstrap? a simulation study comparing the performance of Bayesian Markov Chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Molecular Biology and Evolution 20 (2), 255–266.
Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). Journal of the Royal Statistical Society, B Met 36, 192–236.
Best, N.G., Cowles, M.K., Vines, S.K., 1995. CODA: convergence diagnostics and output analysis software for Gibbs sampling. Technical report. Cambridge MRC Biostatistics Unit.
Carlin, B.P., Chib, S., 1995. Bayesian model choice via Markov Chain Monte Carlo methods. Journal of the Royal Statistical Society, B Met 57, 473–484.
Casella, G., George, E.I., 1992. Explaining the Gibbs sampler. American Statistician 46, 167–174.
Chen, M.-H., Shao, Q.-M., 1997. On Monte Carlo methods for estimating ratios of normalizing constants. Annals of Statistics 25, 1563–1594.
Chib, S., Greenberg, E., 1994. Bayes inference in regression models with ARMA(p,q) errors. Journal of Econometrics 64, 183–206.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis–Hastings algorithm. American Statistician 49, 327–335.
Chib, S., Jeliazkov, I., 2001. Marginal likelihood from the Metropolis–Hastings output. Journal of the American Statistical Association 96 (453), 276–281.
Chib, S., 1995. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 90, 1313–1321.
Cowles, M.K., Carlin, B., 1996. Markov Chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association 91, 883–904.
Dellaportas, P., Forster, J.J., Ntzoufras, I., 1998. On Bayesian model and variable selection using MCMC. University of Economics and Business, Greece. Technical report.
Gamerman, D., 1997. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Chapman and Hall, London.
Gelfand, A.E., Dey, D., 1994. Bayesian Model choice: asymptotics and exact calculations. Journal of the Royal Statistical Society B56, 501–514.
Gelfand, A.E., Smith, A.F.M., 1990. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85, 398–409.
Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequence. Statistical Science 4, 457–472.
Geman, D., Geman, S., 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 609–628.
George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. Journal of the American Statistical Association 88, 881–889.
Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics. Oxford University Press, New York, pp. 169–193.
Gilks, Walter R., Berzuini, Carlo, 2001. Following a Moving Target – Monte Carlo Inference for Dynamic Bayesian Models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (1), 127–146.
Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Markov Chain Monte Carlo in Practice. Chapman and Hall, London.
Godsill, S.J., 1998. On the relationship between model uncertainty methods. Signal Processing Group, Cambridge University. Technical report.
Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.
Han, Cong, Carlin, Bradley P., 2001. Markov Chain Monte Carlo Methods for Computing Bayes Factors. Journal of the American Statistical Association 96 (455), 1122–1132.
Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.
Jeffreys, H., 1961. Theory of Probability, 3rd edn. Oxford University Press, New York.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.
Liu, J.S., Wong, W.H., Kong, A., 1994. Covariance structure of the Gibbs Sampler with applications to the comparisons of estimators and augmentation schemes. Biometrika 81, 27–40.
Liu, J.S., 1994. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. Journal of the American Statistical Association 89, 958–966.
Liu, Jun S., 2008. Monte Carlo strategies in scientific computing. Springer, New York.
Mengersen, K.L., Tweedie, R.L., 1996. Rates of convergence of the Hastings and Metropolis algorithms. Annals of Statistics 24, 101–121.
Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equations of state calculations by fast computing machines. Journal of Chemical Physics 21, 1087–1092.
Muller, P., Erkanli, A., West, M., 1996. Bayesian curve fitting using multivariate normal mixtures. Biometrika 83, 67–79.
Robert, C.P., Casella, G., 1999. Monte Carlo Statistical Methods. Springer, New York.
Roberts, G.O., Sahu, S.K., 1997. Updating schemes, correlation structure, blocking, and parametization for the Gibbs sampler. Journal of the Royal Statistical Society, B Met 59, 291–317.
Tanner, M.A., Wong, H.W., 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–540.
Tanner, M.A., 1996. Tools for Statistical Inference, third ed. Springer, New York.
Tierney, L., 1994. Markov chains for exploring posterior distributions (with discussion). Annals of Statistics 22, 1701–1762.
Zellner, A., Min, C.K., 1995. Gibbs sampler convergence criteria. Journal of the American Statistical Association 90, 921–927.

Moore Jr, Barrington (1913–2005)

Dennis Smith, Loughborough University, Loughborough, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Barrington Moore analyzed the consequences of political systems for human happiness and misery, the nature of exploitation, reciprocity in relations between the people and political elites, and the role of moral codes in producing justice and identifying injustice. His most famous book, *Social Origins of Dictatorship and Democracy*, explored the contribution to shaping sociopolitical change made by declining sociopolitical interests; in particular, powerful classes such as the landed aristocracy and the peasantry who face losing out or even destruction as cities become more influential, industry and the urban working class grow in importance, and central government acquires greater control.

Brief Biography

Family

Barrington Moore Jr. (1913–2005) is best known as the author of the classic *Social Origins of Dictatorship and Democracy* (henceforth *Social Origins*, Moore, 1966), which won the Woodrow Wilson Foundation Award in 1968, and the MacIver Award in 1969. Moore was the son of Barrington Moore Sr., a distinguished ecologist who made his career as a forestry engineer. His mother, Muriel Hennen Morris, a New York socialite, was the daughter of a successful trader with interests in China and Republican politics. After their marriage, his parents moved to the rugged environment of New Mexico where his father's forestry work was based. The marital relationship was a troubled one and when the couple finally split up, the young Barrington, aged 9 years, was taken into the household of his step-grandfather, Lewis Cass Ledyard. He was a successful corporate lawyer who became J. P. Morgan's private attorney and acquired substantial business interests. The Ledyard's social circle reached to the top of high society on the East Coast, including the Vanderbilts, Rockefellers, and Roosevelts. Barrington Moore Jr.'s grandparents brought him up in Newport, Rhode Island, on their Sunset Ridge estate, where a multigabled mansion flanked by trees looked out upon a many-acred lawn.

Education and Intellectual Influences

Moore's earliest educational experiences were not pleasant. He did not enjoy his 6 years at St George's private boarding school in Newport, where he suffered bullying, but at the age of 20 years, he was able to escape to Williams College in Massachusetts. There he majored in Latin and took eight courses in Attic Greek, giving him a thorough grounding in classical and medieval history. When he graduated, he was elected to Phi Beta Kappa, the elite American society for outstanding students in the liberal arts and science, founded in 1776.

At Williams, Moore also took a political science course that introduced him to the work of William Graham Sumner and Albert Galloway Keller. Moore credited these two very different scholars for his own abiding interest in problems relating to inequality, authority, ideological obscurantism, totalitarian regimes, and the causes of human misery. Later, at Yale, Moore carried out graduate work in the social sciences influenced both by Kellner, his supervisor, and G.P. Murdoch, director of the Cross-Cultural Survey (see Jackall, 2001; Snyder, 2007).

Moore completed a doctoral thesis on social stratification that applied a cross-national comparative approach in a strongly positivist spirit. Subsequently, he realized the inadequacy of this approach. Sumner and Kellner were partially displaced as background influences by Gaetano Mosca. More generally, in later years, Moore moved beyond positivism and combined his comparative-historical approach with two theoretical tendencies. One was a modified utilitarianism that assumed human beings generally abhor the misery brought by war, hunger, cruelty, and intolerance (for example, in *Reflections on the Causes of Human Misery and on Certain Proposals to Eliminate Them*; henceforth *Reflections*, Moore, 1972). The other was a quasi-Hegelian search for signs of emancipatory forces that might realize the human potentiality for creating rational and decent societies. He looked for glimmers of this in the rising tide of democracy in the United States post-1776, France post-1789, and England post-1832 (in *Social Origins*) and in the revolutionary upsurge in Germany after the World War I (in *Injustice. The Social Bases of Obedience and Revolt*; henceforth *Injustice*, Moore, 1978).

Career

As an adult, Moore enjoyed a substantial private income. This background provided a solid material and cultural base for his strong independence of mind and character. It allowed him to teach and research without becoming overly troubled with the task of climbing the professional ladder in order to acquire honors. Nor did he seek to build up influence by creating a 'school' or a network of former students in strategic academic posts. However, he earned considerable respect among fellow scholars. A number of his former students such as Theda Skocpol and Charles Tilly have paid tribute to his strong influence. Moore's colleagues in the American Sociological Association created the annual Barrington Moore Book for the best book in comparative and historical sociology.

At the start of his career, Moore was given a strong push toward an approach to social science that was more methodologically sophisticated, morally sensitive, and politically aware by his work for the American government during World

War II. He served as a policy analyst, working in the Department of Justice and the Office of Strategic Studies (OSS). At OSS, Moore was assigned to a team led by Herbert Marcuse, who became his long-term friend and intellectual sparring partner. The OSS, precursor of the Central Intelligence Agency, was in effect the largest social science research organization in the United States during the war years. Both Moore and his wife Elizabeth C Moore are to be found on the OSS list of employees in the mid-1940s. They met when they worked in the same office at the Department of Justice, and married in 1944. Elizabeth (or Betty), who died in 1992, was his main intellectual and life companion.

After the war, Moore taught at the University of Chicago before moving to Harvard in 1948. He did not stay long in the Social Relations Department, where the Parsonian paradigm, to which he was unsympathetic, prevailed, but instead he made his base at the Russian Research Center. He joined this institution in 1951 and remained there for the rest of his working life, researching and also teaching small classes of undergraduates and graduates. Moore greatly valued the resource provided by Harvard's Widener Library, not least as a source of empirical data, including statistical materials. For all his concerns with moral and theoretical issues, Moore was punctilious and precise in his treatment of factual materials, especially in a comparative context. His typical work pattern over many years was to spend much of the winter gathering together research materials in Harvard, often working in his office at home, and then spend a great deal of the summer on his yacht, along with his wife Betty, working on various writing projects.

Major Publications

Before *Social Origins of Dictatorship and Democracy*

Certain themes dominate Moore's major publications. These include the consequences of political systems for human happiness and misery, the nature of exploitation, reciprocity in relations between the people and political elites, and the role of moral codes in producing justice and identifying injustice. All were present in his first important book, entitled *Soviet Politics – the dilemma of power* (Moore, 1950). This book focuses on two major issues. One is the impact on Soviet government and political organization made by the complex and unequal division of industrial labor. The other is the conflict between the egalitarian goals of the Bolsheviks, as proclaimed in their ideology, and the inequalities generated by the political means the Bolsheviks adopted to achieve and keep power in a rapidly industrializing society while facing threats from external and internal enemies.

Moore traces the tactics employed by Lenin and his successor, Stalin, and emphasizes that Bolshevik ideology limited the flexibility of the regime during the period of the New Economic Policy in the 1920s when markets and private enterprise were officially permitted to a limited degree. In a companion study, *Terror and Progress USSR* (Moore, 1954), Moore explores how the regime might be likely to change in the wake of Stalin's death. He pays particular attention to the threat to totalitarian bureaucracy and the official ideology that was posed by two things: the increased salience of social stratification, bringing to prominence a new emphasis on tradition, and the increasingly prominent role being played by apolitical technical experts with their belief in scientific rationality.

In both his Soviet studies, Moore is using his analysis of this specific national case to consider conflicts and affinities more generally present in relations between the following: the political purposes of rulers, the functional requirements of social life, the potentialities and restrictions built into specific cultural symbols, and the way human beings typically perceive their worlds and act within them.

The issues just mentioned contributed to two related inquiries that continue throughout Moore's career. The first is what empirical generalizations may be made about the processes whereby peasants and other kinds of workers become citizens within modern bureaucratic states. The second is what kind of social existence can a citizen legitimately expect to enjoy and how do they typically respond when decent conditions are denied to them (see Smith 1983)?

In *Political Power and Social Theory* (Moore, 1958), Moore sets down his thoughts on modes of acquiring political power, totalitarianism in preindustrial societies, the family, conformity in industrial societies, and methodologies of social science. The essays are designed to attack commonly held but unsustainable assumptions. For example, Moore argues that some institutions, such as totalitarianism, thought to be modern, are also found in premodern times; that some 'essential' institutions, such as the family, thought to be universal to industrial societies, may in fact be dispensable; and that we can identify certain recurrent cycles of institutional change, which involve changes in levels of centralization, integration, and legitimacy within sociopolitical orders.

The first essay, 'Notes on the process of acquiring power,' is a *tour de force*, demonstrating theoretical links and empirical similarities between the rise of the Bolsheviks, the papacy, the Bourbon monarchy in France, and the early Mogul emperors. Moore assesses power-seeking strategies promoting nativism, hierarchy, and egalitarianism, as well as control-seeking strategies such as feudalism, totalitarianism, and rational bureaucracy, each of which may be found in many different societies and epochs.

The essays in *Political Power and Social Theory* were a ground-clearing exercise in which Moore sharpened his analytical tools for the following project: to develop a strategy that would enable him to apply rational analysis to discoverable facts about the empirical sociopolitical world and derive a basis for value judgments about how men and women should behave toward one another, especially as citizens and rulers. On the basis of these early essays, Moore concluded that, regretfully, moral choices had to be made in conditions of high uncertainty. However, he was working to improve that situation and the first substantial fruit of that effort was *Social Origins*, which was published in 1966.

Social Origins of Dictatorship and Democracy

The argument in *Social Origins*, which took Moore about 10 years to write, was groundbreaking in three respects. It demonstrates the large contribution to shaping sociopolitical change made by threatened sociopolitical interests; in particular by powerful or potentially powerful classes such as the landed aristocracy and the peasantry fear losing out as cities

become more influential, industry and the urban working class grow in importance, and central government acquires greater control. Another contribution is Moore's recognition that in cases of attempted rebellion, material motivations for rebellion are mingled with considerations of morality and justice, especially the subordinates' judgment about whether or not the ruling class is doing its best to govern fairly and in the interests of all. Not least, Moore insists that democracies, just as much as dictatorships, were liable to be preceded by violent events such as revolutions and civil wars, and installed under conditions of repression, temporarily at least.

Social Origins presents a comparative analysis of the moral and material conditions of groups within the configurations that bind together peasants, landlords, merchants, rulers, and public officials. On the basis of a masterly exposition, deployed over several chapters, Moore sets out to build two complementary intellectual structures. The first is a series of causal explanations of a number of rapid and dramatic structural transformations, entailing considerable violence and suffering, including the French and Chinese Revolutions. The second is a series of analyses of the morally relevant consequences of the processes whereby modernizing urban society has impinged upon the rural order.

The analytical strategies are very different in the two cases. In constructing his causal explanations, Moore typically presents his empirical findings with respect to the following: (1) the potentialities for and limits upon structural variations in the society concerned at its specific stage of development; (2) the tendencies toward cohesion and disintegration, which are actually present in that society, with particular reference to the division of labor and forms of domination; (3) the ways that specific groups perceive their material and moral interests within the society; (4) the occurrence of specific events or tendencies, which present threats to those perceived interests that are sufficiently sudden or drastic to stimulate people into action; (5) the identity of the potential allies, opponents, and victims of specific groups whose perceived interests are threatened; (6) the options in historical development that have been closed off in that society as a result of preceding sequences of historical change; (7) how threatened groups actually behave when the threat materializes; and (8) the structural outcomes to which their behavior contributes.

Applying this approach, he identifies three main routes from the preindustrial to the modern world. One leads toward democracy, an outcome that is made much likely when a strong and independent capitalist bourgeoisie emerges, ready to attack inherited obstacles to such a development. England, France, and the United States are examples, although they arrive at this point through diverse trajectories. Two other routes lead to dictatorship.

The second route is capitalist but the principal agent is a dominant landed aristocracy, which does the military and political work of clearing the space for a commercially vigorous bourgeoisie while making sure that the bourgeoisie's aspirations for democracy are kept in check. Japan and Germany are two key examples. When the peasants and urban workers demand to become part of the modernizing polity, they are incorporated through fascist techniques that give them a subordinate, obedient role.

The third route involves peasant uprisings, which weaken the rural ruling class in societies where the commercial bourgeoisie is not dynamic enough to bring about industrial modernization. This modernizing task is undertaken instead by a powerful state bureaucracy, which incorporates workers and peasants in the name of communist ideology. Russia and China are the main examples.

Moore's comparative analyses of varying patterns of conflict and solidarity within peasant communities, in France, China, Japan, and so forth, are a valuable part of *Social Origins*. So is his lengthy chapter on India, a case that does not fit easily into his typology of routes.

In assessing the morally relevant political consequences (as distinct from the causes) of modernizing processes, Moore's basic approach is to conduct a cost–benefit analysis, which weighs the costs in terms of human misery against any benefits in terms of human freedom. This calculus includes costs and benefits that might plausibly have been expected if genuinely feasible alternative patterns of development that did not occur had in fact been followed. So an element of counterfactual reasoning about opportunity costs is included (see Smith, 1984).

After *Social Origins of Dictatorship and Democracy*

In his next two books, Moore conducts a frontal attack on the problem of how criteria for value judgments may be derived from objective knowledge about human societies. In *Reflections* (Moore, 1972), he concludes that the most likely outcome of existing sociopolitical trends is the erosion of accepted codes of behavior and the collapse of political authority. However, he finds some rays of hope. For example, he reckons that it will be possible to persuade supporters of capitalism that the misery it produces is not an inevitable or necessary aspect of that system; a more humane form of capitalism is technically feasible as well as being desirable. He sets out some standards for such a society. For example, its inhabitants should expect to enjoy the following: emotional security; a broad and coherent liberal education that gives them technical and intellectual competence; and a significant degree of socioeconomic equality, sufficient to prevent a powerful establishment from controlling thought.

Moore accepts that insecurity and uncertainty in social and political life are impossible to eliminate, as are fraud and selfishness. However, they may be minimized if citizens demand that those in power exercise what he calls 'rational political authority.' Applying a quasi-utilitarian moral calculus, Moore deduces that rulers should try to govern in such a way that they minimize suffering within their societies insofar as this is humanly possible, given existing material resources, technological facilities, and organizational capacities within the society.

In *Injustice* (Moore, 1978), he looks for evidence that in practice subjects or citizens normally apply the standard just mentioned when assessing the activities of their rulers or governments. His main point is that when rulers and governments either fail to prevent or deliberately bring about suffering and misery that could have been avoided, the typical response of those who are damaged by this is intense moral outrage. In effect, *Injustice* is an amalgam of two intellectual exercises.

One is an essay on philosophical anthropology, which seeks empirical evidence for the existence of a widespread ideology of rational political authority, drawing upon data relating to societies as varied as Inuits in North Alaska, the Bantu-speaking Lovedu people of South Africa, the Kapauku Papuans, the inhabitants of Barotseland in Zambia, the Trobrian Islanders, the Semai of Malaysia, the inhabitants of ancient Greece, the Ming Empire in China, and eighteenth-century England. There is some confirmatory evidence but it falls far short of proving the case.

The other exercise, more successful, is Moore's study of the German working class, which explores the way industrial employees diagnosed their own experiences at work and proposed remedies for their disadvantageous situation. In particular, he considers the uprising in the Ruhr in 1920, and the choices open to Friedrich Ebert, German chancellor and leader of the SPD (Sozialdemokratische Partei Deutschlands) when faced with this revolt. In carrying out this analysis, Moore succeeds in drawing together two things: the shaping of working-class consciousness in a specific historical situation and the economic, political, and other macrostructural forces that determined both the constraints and opportunities facing the rebellious mining communities of the Ruhr.

After *Injustice*, Moore published four volumes of essays, each focused upon a specific theme. These were *Privacy. Studies in Social and Cultural History* (henceforth *Privacy*, Moore, 1984), *Authority and Inequality under Capitalism and Socialism. United States of America, USSR and China* (henceforth *Authority and Inequality*, Moore, 1987), *Moral Aspects of Economic Growth and Other Essays* (henceforth *Moral Aspects*, Moore, 1998), and *Moral Purity and Persecution in History* (henceforth *Moral Purity*, Moore, 2000).

Authority and Inequality was based upon the Tanner lectures that Moore delivered at Brasenose College, Oxford in 1984–85. In these lectures, Moore examines the ways that bureaucracies and markets structured relations of authority and inequality in the United States, USSR, and China, and the potential for maximizing freedom and rationality in these societies, especially by the use of rational political authority to reduce misery. He argues that industrialization has increased the viability of forms of sociopolitical organization that are both prosperous and free. This is one of the reasons why resistance to bureaucracies that fail to pursue prosperity and freedom for all is endemic, as subordinates' look after their own interests as far as possible by setting up various kinds of informal and unofficial groups in their workplaces and local neighborhoods. Rulers may try either to subvert these groups (as in Russia) or infiltrate and manipulate them (as in the United States). In China, such groupings have been especially effective in cutting officialdom down to size, as occurred during the Cultural Revolution.

Distinctive structural tensions or contradictions are found in all three countries. In Russia, the imposition of terror weakened economic initiative in the population at large and this situation has not improved following the shift from terror to technocracy that had occurred by the mid-1980s. By contrast, in China, there has been a greater commitment by government to rational political authority. Citizens and workers have been encouraged and emboldened to attack corruption. This effort has tended to divert energies from the pursuit of economic growth. Furthermore, attempts to introduce market incentives for individual enterprise have been restricted: by the vested interests of bureaucrats and workers dedicated to maintaining a guaranteed living standard ('the iron rice bowl'); by worker resistance to interference; and by the government's anxiety about the danger of creating an underemployed intelligentsia, trained and motivated for market opportunities that do not yet exist.

In the United States, there is an abiding structural contradiction between, on the one hand, the deep antiauthoritarianism and individualism of American culture and, on the other hand, the increase in bureaucratization since the 1870s. Moreover, inflation during the 1960s and 1970s undermined the ability of the market to establish a close and clear relationship between individual effort and individual reward. The struggle for advantage between groups in an inflationary context undermined the social peace established since the 1930s by the New Deal coalition of Democratic politicians, trade unions, white southerners, and various racial, ethnic, and religious minorities.

Moore concludes that the legitimacy of both the socialist and the liberal capitalist models of politics and society have suffered decay. Neither has eliminated inequality and they have brought with them the evils of, in one case, imperialism and unemployment and, in the other case, scarcity and repression. Organizational complexity and the need for coordination mean that all industrial societies are bound to limit freedom and equality in some respects. However, justice demands that appropriate rewards should go to those with socially useful competences in short supply. In fact, this does not occur. Another complication is that people intensely dislike the anxiety that accompanies industrial modernization and, as a result, they are all too ready to be seduced by fascism, nihilism, or hedonism as ways of escaping responsibility and uncertainty.

In his three other main works, Moore returns to the strategy of comparative anthropology and history undertaken in *Injustice*, pursuing themes such as privacy, moral aspects of capitalism and socialism, and moral purity in different cultural and religious traditions through history. In *Privacy* (Moore, 1984), Moore finds that in ancient Greece, private life was considered inferior to public participation in civic affairs. In Hebrew society, as recorded in the Old Testament, God's law reached deep inside the household and personal sphere. By contrast, in imperial China, rules of etiquette provided limits to the extent of intrusion imposed upon educated people. One of the great civilizing achievements of the modern West has been to challenge the priority of the public sphere and defend the private sphere against intrusion. However, Moore considers that a pathological version of privacy has developed in contemporary America, which inhibits citizens from getting involved in matters that do not directly and immediately concern them.

Many of the essays in *Moral Aspects* (Moore, 1998) overlap with Moore's other writings but two should receive individual mention. One is the title essay, 'Moral Aspects of Economic Growth: Historical Notes on Business Morality in England' (Moore, 1998, 1–53), which is part of a comparative historical project that remained uncompleted. It is largely based on a study of the laws and customs controlling trade in medieval and early modern England. During this period, business norms placed a large premium on the appearance of respectability and

social concern, including care for employees. These norms were put to the test as commercial and industrial growth eventually brought about a society in which acquiring wealth was no longer a zero-sum game in which one had to steal from others, oppress them, or simply go without. The ironic result of economic growth was that, despite the increased potential for strengthening public institutions providing social welfare, by the early nineteenth century the market was being widely described as a key arena in which individual initiative, self-reliance, and hard work provided the only road to material success. However, Moore finds that in practice, the paternalistic tinge of medieval social policy persisted. Moral concerns remained densely interwoven with economic practice.

In 'Rational Discussion: Comparative Historical Notes on Its Origins, Enemies, and Prospects' (Moore, 1998, 144–157), Moore contrasts the oratorical political culture of ancient Greece with the memorandum-based styles of the Chinese imperial bureaucracy, each designed to facilitate rational discussion, with both cultures encouraging its elites to acquire the relevant arts of speaking and writing. Turning to religion, Moore notes its tendency to raise the temperature of debate and argues that a balance is needed between enthusiasm and detachment when discussing public policy. In conclusion, Moore suggests that rational discussion is unlikely when the following conditions prevail: a high level of political excitement, an authoritarian regime repressing opposition, a powerful educated elite dominating decision making, or widespread quasi-religious certainty about how society ought to be organized.

Moore's last book, *Moral Purity* (Moore, 2000), traces ideas of purity and impurity through the Old Testament, in the religious conflicts of sixteenth-century France, during the French Revolution, and in Hinduism, Buddhism, and Confucian China. This survey leads him to the conclusion that the theory and practice of moral purity, combined with attacks upon manifestations of impurity in others, arose from monotheism, as developed in Judaism, Christianity, and Islam. These concerns were secularized during the French Revolution, which popularized the idea of revolutionary purity. The three Asian religious approaches studied were relatively free of obsessions about purity until the expansion of Western ideas and ideals transformed them, leading to Maoism, militant Hinduism, and more militant forms of Buddhism. Demands for moral purity have become global, an undesirable outcome in Moore's view.

Critical Reception and Lasting Impact of Moore's Work

The Influence of Moore's Methodology

Moore's principal concern throughout his career was to encourage social scientists to remain clear-sighted, rational, and dispassionate in considering the evidence that their research produced, wherever it led. Above all, researchers should be wary of ideological formulas and dogmatic assertions of all kinds. Moore's last book, *Moral Purity* (Moore, 2000), illustrates the danger of dogma since it argues that monotheism carries the potential for justifying self-righteous violence against hated unbelievers.

Moore's typical approach was to locate specific empirical cases in a comparative context as far as possible, while examining, in each case studied, the interplay between the human pursuit of material and moral interests and the driving force of large-scale historical processes that open up some options for action while closing others.

Moore insists that in assessing the costs and benefits of any historical outcome, whether it is a new regime that sets a new political direction, a technological innovation that transforms the industrial division of labor, or a religious revival that transforms a community's way of life, a key counterfactual question is what would have been the costs and benefits if matters had gone differently: a different regime, another kind of technology, an alternative form of religion, or no change at all?

The scale and scope of the vision embodied in Moore's methodology, combining knowledgeable detachment with alert involvement, locate him in a tradition that reaches back to Montesquieu and Alexis de Tocqueville and includes more recent practitioners such as Norbert Elias, C Wright Mills, and Albert Hirschman. Apart from a sharp negative appraisal of Talcott Parsons' 'new scholasticism' (Moore, 1953), Moore does not engage to any great extent with contemporary social and political theory. For this reason, perhaps, reviewers have sometimes found works such as *Privacy* (Moore, 1984) rather mystifying, being unfamiliar with, or unsympathetic to, Moore's long-term agenda of seeking universally shared moral criteria while remaining alert to the possible presence of long-term processes of development in cultures across space and time.

Moore's methodological approach is not reducible to short formulae and has been most effectively transmitted to students who have worked closely with him and those who have read his books with care, not least by those determined, if possible, to undermine or disprove his specific empirical generalizations or theoretical propositions. For his part, Moore welcomed such critical attention and was happy to see his own provisional conclusions improved or superseded by careful scholars.

One example of Moore's willingness to debate his conclusions is his friendly dialogue with Herbert Marcuse, who shared his interest in Soviet Marxism, the relationship between human desires and cultural constraints, the causes of human misery, and the sociopolitical dynamics of tolerance (Marcuse, 1958, 1964, 1969). They both contributed essays to a book on tolerance (Wolff et al., 1965).

Responses to *Social Origins of Dictatorship and Democracy*

Barrington Moore is best known for his resounding classic, *Social Origins* (Moore, 1966), although its renown dwindled during the mid-1990s, along with historical sociology more generally, in the face of the onslaught of ahistorical, market-oriented, rational-choice forms of theory and scholarship heavily influenced by the political success of neoliberal economics. *Social Origins* has received more critical attention since the late 1990s (e.g., Skocpol, 1998).

Moore's confrontation in *Social Origins* with both Marxian and liberal interpretations of the emergence of capitalist democracy, fascism, and communism made this book an object of intense debate across disciplines and continents

during the late 1960s and 1970s, and since then a continuing challenge to practitioners of comparative historical analysis. By 1975, *Social Origins* had been reprinted nine times in the United States.

Early reviewers had difficulty in categorizing the book, which was variously described as Marxist, Weberian, and Parsonian. While its great strengths were acknowledged, a number of justified criticisms were made: for example, there was too little consideration of relations between states and societies, processes of political decision making within the state apparatus were neglected, and the role of ideas tended to be treated as a secondary factor. Moore's predominant emphasis in *Social Origins* was upon the ramifications of processes such as the commercialization of agriculture and the progressive infiltration of governmental bureaucratic power throughout modernizing national societies.

Despite its emphasis upon the dynamics of political economy, Moore's direct approach to power relations and exploitation could not be dismissed as mere left-wing rhetoric since it came in the form of a carefully argued, dispassionate, and empirically rich thesis advanced by a man of impeccable social background rooted in the heart of the American establishment. As a result, *Social Origins* cleared the way for the positive reception in the United States during the 1970s of works such as those by Eric R Woolf and Roland Mousnier on peasant uprising (Woolf, 1971; Mousnier, 1971), Antony Giddens on the class structure of advanced societies (Giddens, 1973), Perry Anderson on historical forms of political economy (Anderson, 1974a,b), Immanuel Wallerstein's analyses of the capitalist world system (e.g., Wallerstein, 1974), Jeffrey Paige on agrarian revolutions (Paige, 1975), James Scott on peasant moral economy (Scott, 1976), and Theda Skocpol's comparative analysis of social revolutions (Skocpol, 1979).

During the 1980s and early 1990s, a strong current of serious research continued, inspired by Moore's example and subjecting his arguments to sympathetic challenge. Three examples are Jack Goldstone's discussion of revolution and rebellion in the early modern world (Goldstone, 1991), Gregory Luebbert's study of political order in interwar Europe (Luebbert, 1991), and Brian M Downing's analysis of the ways in which military modernization in early modern Europe contributed to the emergence of democracy and dictatorship (Downing, 1992).

By the time Barrington Moore's centenary arrived in 2013, the prospects of his work's survival as an intellectual force for the foreseeable future seemed good. For example, Daron Acemoglu and James A Robinson paid indirect tribute to that influence in the title of their study of the 'economic origins of dictatorship and democracy' (Acemoglu and Robinson, 2006). This book shares Moore's concern with struggles between socioeconomic interests to get their hands on their society's resources and influence the political and constitutional rules through which resource distribution is governed. However, in sharp contrast to Moore, who works through a mountain of empirical materials to arrive at his provisional generalizations, Acemoglu and Robinson begin with an a priori rational choice model from which they derive a typology of possible outcomes. These outcomes are then illustrated through empirical examples – Britain, Argentina, Singapore, and South Africa – that appear to exemplify the range of typological possibilities.

In fact, it is not reasonable to expect that Moore's comparative-historical and counterfactual methodology might be contained within a rational choice approach. This is especially the case when attention is paid to the emotional and social–psychological dynamics of political culture, a factor bracketed out by rational choice theory but an important consideration within *Reflections* (Moore, 1972) and *Social Origins* and, indeed, a central concern of Moore's book *Injustice* (Moore, 1978). In the early twenty-first century, these emotional and social–psychological dynamics are receiving considerable attention both in political science and historical sociology (e.g., Demertzis, 2013; Smith, 2006, 2013). This is yet another reason why Moore's concerns, his methods, and his example will continue to enliven debate and dialogue in the social sciences.

See also: Elias, Norbert (1897–1990); Giddens, Anthony (1938–); Hirschman, Albert O. (1915–2012); Mills, Charles Wright (1916–62); Skocpol, Theda (1947–); Sumner, William (1840–1910); Wallerstein, Immanuel (1930–).

Bibliography

Acemoglu, D., Robinson, J.A., 2006. Economic Origins of Dictatorship and Democracy. Cambridge University Press, Cambridge.
Anderson, P., 1974a. Passages from Antiquity to Feudalism. New Left Books, London.
Anderson, P., 1974b. Lineages of the Absolutist State. New Left Books, London.
Demertzis, N. (Ed.), 2013. Emotions in Politics. The Affect Dimension in Political Tension. Palgrave Macmillan, London.
Downing, B.M., 1992. The Military Revolution and Political Change. Origins of Democracy and Autocracy in Early Modern Europe. Princeton University Press, Princeton.
Giddens, A., 1973. The Class Structure of the Advanced Societies. Hutchinson, London.
Goldstone, J.A., 1991. Revolution and Rebellion in the Early Modern World. University of California Press, Berkeley, CA.
Jackall, R., 2001. The education of Barrington Moore, Jr. International Journal of Politics, Culture and Society 14 (4), 675–681.
Luebbert, G.M., 1991. Liberalism, Fascism, or Social Democracy. Social Classes and the Political Origins of Regimes in Interwar Europe. Oxford University Press, Oxford.
Marcuse, H., 1958. Soviet Marxism. Routledge, London.
Marcuse, H., 1964. One-dimensional Man. Routledge, London.
Marcuse, H., 1969. Eros and Civilization. Sphere Books, London.
Moore, B., 1950. Soviet Politics – The Dilemma of Power. The Role of Ideas in Social Change. Harvard University Press, Cambridge, MA.
Moore, B., 1953. The new scholasticism and the study of politics. World Politics 6 (1), 122–138.
Moore, B., 1954. Terror and Progress USSR. Some Sources of Change and Stability in the Soviet Dictatorship. Harvard University Press, Cambridge, MA.
Moore, B., 1958. Political Power and Social Theory: Six Studies. Harvard University Press, Cambridge, MA.
Moore, B., 1966. Social Origins of Dictatorship and Democracy. Lord and Peasant in the Making of the Modern World. Beacon Press, Boston, MA.
Moore, B., 1972. Reflections on the Causes of Human Misery and on Certain Proposals to Eliminate Them. Penguin, Harmondsworth.
Moore, B., 1978. Injustice. The Social Bases of Obedience and Revolt. Macmillan, London.
Moore, B., 1984. Privacy. Studies in Social and Cultural History. M E Sharpe, Inc., Armonk, NY.
Moore, B., 1987. Authority and Inequality under Capitalism and Socialism. Clarendon Press, Oxford.
Moore, B., 1998. Moral Aspects of Economic Growth, and Other Essays. Cornell University Press, Ithaca, NY.

Moore, B., 2000. Moral Purity and Persecution in History. Princeton University Press, Princeton, NJ.
Mousnier, R., 1971. Peasant Uprisings. Allen and Unwin, London.
Paige, J.M., 1975. Agrarian Revolution. Free Press, New York, NY.
Scott, J.C., 1976. The Moral Economy of the Peasant: Rebellion and Subsistence in Southeast Asia. Princeton University Press, New Haven, CT.
Skocpol, T., 1979. States and Social Revolutions. Cambridge University Press, Cambridge.
Skocpol, T. (Ed.), 1998. Democracy, Revolution, and History. Cornell University Press, London.
Smith, D., 1983. Barrington Moore. Violence, Morality and Political Change. Macmillan, London.
Smith, D., 1984. Morality and method in the work of Barrington Moore. Theory and Society 13 (2), 151–176.
Smith, D., 2006. Globalization – the Hidden Agenda. Polity Press, Cambridge.
Smith, D., 2013. Inside stories. Oscar Wilde, Jean Améry, Nelson Mandela and Aang San Suu Kyi. In: Demertzis, N. (Ed.), Emotions in Politics. The Affect Dimension in Political Tension. Palgrave Macmillan, London, pp. 60–83.
Snyder, J., 2007. The critical spirit and comparative historical analysis (interview with Barrington Moore). In: Munck, G.L., Snyder, J. (Eds.), Passion, Craft and Method in Comparative Politics. Johns Hopkins Press, Baltimore, MD, pp. 86–112.
Wallerstein, I., 1974. The Modern World-system. Capitalist Agriculture and the Origins of the European World-economy in the Sixteenth Century. Academic Press, New York, NY.
Wiener, Jon, 1976. Review of reviews: social origins of dictatorship and democracy. History and Theory 15, 146–175.
Wolf, E.R., 1971. Peasant Wars of the Twentieth Century. Faber, London.
Wolff, R., et al., 1965. A Critique of Pure Tolerance. Beacon Press, Boston, MA.

Moral Development and Education

Larry Nucci, Graduate School of Education, University of California, Berkele, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article summarizes developmental approaches to moral education beginning with the work of Piaget and Kohlberg. Current research on moral development is reviewed including attention to the ethics of care, and social cognitive domain theory with reference to issues of culture and findings from neuroscience. Educational practices reviewed include dilemma discussion, transactive discussion, the 'Just Community School,' Developmental Discipline, domain consistent teacher discourse, and community service activities aimed at the development of the 'moral self.'

Applications of developmental psychology to moral education originated with Jean Piaget's (1932) studies of the moral development of young children. Piaget (1932) approached children's socialization through a careful study of how children interpret the rules that structure their games, and how they reason about actions that carry moral consequences. This work led him to conclude that children's natural efforts to understand the social world result in concepts of morality that shift from an earlier 'heteronomous' orientation toward adult authority and social norms to an 'autonomous' moral orientation based on children's own interpretations of fairness, reciprocity and mutual respect. In the heteronomous phase, according to Piaget (1932), children maintain a unilateral respect for adult authority and view social rules as fixed. Children's moral behavior is guided by concerns for objective consequences of their actions and a fear of external consequences. As children develop they become capable of taking into account the perspectives of others, and shift their moral orientation toward reciprocity and mutual respect. Moral obligation in the autonomous phase stems from this understanding of the requirements of fairness rather than being based upon external rules and sanctions.

Although Piaget did not engage in educational research, his conclusions about children's heteronomous and autonomous moral orientations led him to argue that early moral development is stimulated by peer interactions in which social power is more or less equalized, and is hampered by adult interventions that invoke the imposition of social norms and unilateral solutions to social conflicts that reinforce young children's heteronomous morality. Piaget concluded from this work that schools should emphasize cooperative decision-making and problem solving, nurturing moral development by requiring students to work out common rules based on fairness (DeVries and Zan, 1994). In effect, Piaget's (1932) view of moral education was the inverse of the traditional approach to character education that emphasizes the adult role in children's socialization.

Kohlberg and the Application of Moral-Stage Theory

In the 1960s Piaget's (1932) account of moral development was extended and reworked by Lawrence Kohlberg (1984). Consistent with Piaget, he proposed that children form their understandings of moral concepts such as justice, rights, and human welfare from their direct social experiences rather than acquiring them as social values from adults. Kohlberg followed the development of moral judgment beyond the ages studied by Piaget, and determined that the process of attaining moral maturity took longer and was more gradual than Piaget had proposed. Kohlberg described moral development as moving through three levels incorporating a total of six developmental stages. However, the overall pattern described by Kohlberg (1984) was in line with Piaget's basic premise that moral development involves a gradual transition from the conflation of morality with the conventions and standards of society and social authority to a principled level of moral maturity in which moral judgments are structured by underlying conceptions of fairness and human welfare.

Moral Dilemma Discussion

Kohlberg's (1984) research on moral development was part of the 'cognitive revolution' that reshaped American psychology in the 1970s. For Kohlberg, the limitations of behaviorist and social learning accounts of socialization mapped directly onto Piaget's (1932) criticisms of Durheim's (1925) vision of moral education, and traditional approaches to character education. Kohlberg (Kohlberg and Turiel, 1971) mocked the appropriation of Aristotelian philosophy by traditional character education as reducing moral education to the teaching of an arbitrary 'bag of virtues' defined by the unexamined social values of character education proponents. Moreover, he viewed the socialization approach advocated by traditional character education as contributing to the consolidation of conventional moral reasoning rather than contributing toward moral development.

Instead of teaching moral values directly, Kohlberg proposed that the core of moral education should be a focus upon moving students through the progression of stages of moral reasoning toward principled morality. Following Piaget, Kohlberg held that the process of stage change occurs gradually as a result of the efforts of individuals to resolve the contradictions or conflicts that result from the application of their moral understandings to actual moral situations. Kohlberg applied this general principle to his work on education through the use of *moral dilemma discussion*. Early research on the social experiences that contribute to stage change indicated that development is enhanced when individuals are drawn into a discussion about a moral conflict and asked to provide both a solution and justification for their position, and when

counterarguments are one stage above the moral reasoning level of the targeted students (Blatt and Kohlberg, 1975). Later, more refined studies indicated that the key to successful moral discussion was less a matter of achieving an 'optimal mismatch' in the moral reasoning stage of discussion participants, than the degree to which members of a discussion actively engaged in efforts to reconcile or integrate competing moral arguments with the goal of arriving at a shared moral position (Berkowitz and Gibbs, 1983). The 'transactive' discussion patterns that Berkowitz and Gibbs (1983) identified in their research were structurally similar to what Jurgen Habermas (1991) has defined as the 'communicative discourse' requisite for the social construction of a rationally defensible morality. The actual classroom applications of Kohlberg's moral stage theory to moral dilemma discussion did not fare as well. This approach to moral education did not translate into changes in student conduct, nor was it easily integrated within the natural flow of academic activities valued by classroom teachers (Power et al., 1989). As will be discussed below, however, recent work has effectively revived the use of classroom discussion as a tool for moral education (Nucci, 2009).

The Just Community School

Following the disappointing outcomes of the early effort at moral education through dilemma discussion, Kolhberg and his colleagues (Power et al., 1989) shifted their efforts toward a more comprehensive reform of the organization and culture of the school as a whole. What they refer to as the *just community* approach divides up students within a school into *clusters* of 40–100 students and their teachers who engage in weekly whole-group discussions focused upon arriving at just resolutions of issues and disputes that arise within the actual interactions of the students. Published reports indicate that this approach is quite effective in reducing student misconduct, and that it contributes to students' moral development (Power and Higgins-D'Alessandro, 2008). Despite its apparent effectiveness, the just community approach has not been widely adopted by school systems either in the United States or Europe where much of the research effort has been focused. Among the stated impediments to large-scale adoption of this approach is that it (1) requires radical school reform; (2) involves a substantial commitment to teacher preparation; (3) involves an initial investment of resources; and (4) is emotionally draining on participating teachers (Power and Higgins-D'Alessandro, 2008). Moreover, the Kohlberg account of moral development upon which the approach is based has been challenged by developmental psychology (Snarey and Samuelson, 2008). The latter critique, however, is not aimed at the broader educational methods employed in the just community approach. Despite these criticisms and challenges the just community approach has staunch defenders and advocates (Power and Higgins-D'Alessandro, 2008).

Moral Education and the Morality of Care

One of the most important early challenges to Kohlberg's vision of moral development and moral education focused upon the perceived overemphasis on reasoning and discourse, and the deemphasis of moral emotion (Noddings, 2002). The impetus of this critique came from claims by Carol Gilligan (1982) that Kohlberg's moral theory had overlooked a separate moral ethic of care that was said to be more characteristic of women and girls than of men and boys. The initial claims of gender differences in the 'care orientation' toward morality were largely refuted by subsequent research (Walker, 2006). However, the notion of an ethic of care was taken up by Nel Noddings (2002) and made the centerpiece of her approach to moral education. The goal of moral education according to Noddings (2002) is to develop a caring orientation marked by compassionate engagement and acts of kindness toward others along with the capacity to accept care directed toward oneself.

Noddings' (2002) concerns address an imbalance in the emphasis on reasoning that Kohlberg himself understood to be only one component of a comprehensive treatment of moral education (Kohlberg and Turiel, 1971). There has been a great deal of attention paid in recent years to the topic of moral emotion. This attention has been accompanied by suggestions for ways in which school practices might foster the emotional component of moral functioning. Unfortunately, much of what has been written has served to further the misconception that cognition and affect operate independently of one another. For example, latter-day emotivists have made claims that peoples' responses to moral events are directed by a set of inborn affective triggers, and that moral reasoning occurs after-the-fact as a means to explain our own reactions to ourselves (Greene and Haidt, 2002; Izard, 1986; Haidt, 2001). Such models are built from research indicating that humans possess a number of inborn capacities to display and to recognize in others basic emotions such as anger, sadness, and joy (Eckman, 1993), and that human infants appear to be primed to respond empathically to the distress of others (Emde et al., 2001; Ludemann, 1991). From an evolutionary/adaptive view of intelligence one may argue that such basic emotional schemata provide a 'quick and dirty' response to social problems that may be a part of our evolutionary heritage. Unmediated raw emotion, however, is hardly adaptive in the social world. Critics have argued that such emotivist models fail to adequately account for the ways in which cognitions constructed over a person's lifetime alter or enter into the regulation of affect, and final appraisal of a social situation (Pizzaro and Bloom, 2003; Turiel, 2010). Moreover, such emotivist theories misrepresent the ways in which affect and cognition are intertwined even in the case of such inborn affective response systems (Damasio, 1996). For her part Noddings (2008) disavows these reductionist and simplistic efforts to account for the role of emotion in moral functioning, and distinguishes them from the ethic of care.

Developmental Discipline

Translating the ethic of care into educational practice has fallen primarily to educational researchers who have focused upon classroom climate and developmentally appropriate modes of student discipline. The most extensively researched effort to combine classroom practices with the goal of enhancing an overall climate of trust and moral 'goodwill' (Arsenio and Lover,

1995) has been Watson's (2003, 2008) work on what is referred to as *developmental discipline*. Watson's approach integrates basic elements of attachment theory (Shroufe, 1988) with the 'self-determination theory' (Deci and Ryan, 1985) view of student motivation. A central tenet of developmental discipline is that young children need to establish sound attachment relationships with significant adults. Classroom teachers fall within this adult category and in some cases serve as surrogates for parents who have not provided these emotional needs for their own children. The basic goal of developmental discipline is to enlist children's basic attachment needs and their intrinsic motivations for autonomy to guide students toward engagement in moral conduct. Teachers employing developmental discipline minimize their use of power assertion, and focus instead on efforts to engage students in reflection on their own motivations and feelings as well as the effects that their actions have had on others. Conflicts are resolved by mediating discourse between or among students with encouragement to find their own solutions, rather than resorting to teacher-imposed resolutions. Students are also involved in the construction and/or modification of classroom rules and conventions governing their behavior. These practices are largely congruent with the positions taken by Piaget (1932) and his educational supporters (DeVries and Zan, 1994). For Watson (2003), engagement of students in the process of establishing classroom customs and scaffolding their capacity to resolve interpersonal conflicts is an aspect of community building that contributes to the overall atmosphere of trust. It is this sense of 'trust' that Watson views as critical to the developmental needs of elementary school children, and allows them grow morally (Watson, 2003). Research conducted on the effectiveness of developmental discipline on students' prosocial behavior has been largely supportive (Battistich et al., 1997). A retrospective study of young adults reported that experiences in elementary classrooms employing developmental discipline have salutary long-lasting impacts (Watson, 2008).

Social Cognitive Domain Theory and Moral Education

By the 1980s a number of questions and challenges arose regarding Kohlberg's (1984) stage theory of moral development (Snarey and Samuelson, 2008). Among the problems noted are the inconsistencies in expression of moral stage within individuals across contexts (Carpendale and Krebs, 1992), the lack of fit between the Kohlberg assessment of moral growth and the moral judgments of individuals from non-Western cultures (Snarey and Keljo, 1991), and the general lack of correspondence between moral reasoning and behavior when defined in terms of Kohlberg stage (Blasi, 1980). What emerged from subsequent developmental research are accounts that view social and moral reasoning as heterogeneous and multifaceted (Turiel, 2002). The dominant post-Kohlberg paradigm for research on moral development is what is referred to as *social cognitive domain theory* (Smetana, 2006; Turiel, 2002). According to domain theory concepts about morality emerge in early childhood around nonarbitrary and universal considerations of acts such as hitting and hurting that have consequences for others. Moral judgments are structured by underlying conceptions of fairness and human welfare, and are distinct from concepts about social conventions that are the contextually dependent and agreed upon social rules established through consensus by members of a given social system or group (Turiel, 1983). Over the past 30 plus years numerous studies conducted in a wide range of cultural contexts have replicated the basic finding that children and adults maintain conceptual distinctions between issues of morality and societal convention (Turiel, 2002; Nucci, 2001). The youngest age at which children have been reported to reliably distinguish between morality and convention is 2 ½ years (Smetana and Braeges, 1990). Most recently, it has been discovered that this basic distinction is maintained even by children suffering from autism (James and Blair, 2005; Leslie et al., 2006). Neuroscience research has indicated that issues of morality and convention are processed differently (Lahat et al., 2012).

The domains of morality and convention are further distinguished from concepts about personal matters of individual choice and privacy (Nucci, 2001, 2014). While morality and convention deal with aspects of interpersonal regulation, concepts of personal issues refer to actions that comprise the private aspects of one's life, such as the contents of a diary, and issues that are matters of preference and choice (e.g., friends, music, hairstyle) rather than right or wrong. Research conducted in a wide range of cultures has indicated that claims to a personal area are a fundamental aspect of social development critical to the establishment of personal autonomy and individual identity (Nucci, 2001, 2014).

In contrast with the developmental assumptions of Piaget (1932) and Kohlberg (1984) in which morality gradually displaces convention as a basis for moral judgments, domain theory maintains that conceptions of morality and convention coexist as separate conceptual frameworks throughout development. Moreover, judgments made in context may draw from one or another conceptual domain (include the personal), or entail coordinations across domains as a function of the individual's interpretation of the moral and nonmoral elements of a given social situation (Smetana, 2006; Turiel, 2002). With development children construct more comprehensive and internally consistent concepts within each domain, and become increasingly capable of engaging in cross-domain coordinations (Nucci, 2009; Nucci and Turiel, 2009; Smetana, 2006).

Revised Aims of Moral Education

This more complex account of moral and social development has led to corresponding modifications in statements of the developmental aims of moral education (Nucci, 2001, 2009). The core focus of moral education remains on the development of the students' conceptions of fairness and human welfare. However, the goals also include attention to the capacities of children to coordinate moral and nonmoral considerations in context (Nucci, 2009; Nucci and Turiel, 2009). This includes supporting the construction of understandings of the function, purposes and importance of conventions. More centrally, however, this view of social and moral cognition makes it evident that children may be engaged in meaningful reflection on the moral implications of prevailing social norms and conventions at *all* points in development, and not just at the more advanced principled level of moral reasoning (Nucci, 2009). The aim of moral education is correspondingly

redirected toward increasing the capacity to manage moral complexity and the application of moral considerations of fairness and concern for rights and human welfare across the lifespan.

Educational Applications

Advocates for the application of domain theory to moral education maintain the emphasis on attention to development, but modify this principle with an attention to social cognitive domain (Nucci, 2001, 2009). Domain theory based research has largely supported the practices advocated by proponents of *developmental discipline* and the concerns for care expressed by Noddings (2002) and social justice by proponents of just community schools (Power and Higgins-D'Alessandro, 2008). For example, research on children's moral emotions sustains the emphasis on establishing a climate of *trust* (Watson, 2003) as being concordant with the child's construction of a moral worldview based upon 'goodwill' associated with prosocial moral behavior (Arsenio and Lover, 1995). What has been added is specificity with respect to the social messages teachers should convey when engaging students around moral transgressions and violations of school rules (Nucci, 2001). Research has demonstrated that children evaluate teacher responses to their transgressions in terms of their correspondence to the domain of the transgression (Killen et al., 1994; Nucci, 2001). Statements of governing rules, or teacher statements focusing upon disruptions to classroom functioning concordant with violations of conventions were rated highly by students, but were given low ratings as teacher responses when applied to moral violations. In contrast directing students to consider the consequences of actions upon the welfare of others were rated highly as appropriate teacher responses to moral violations, but not in response to violations of school conventions. Teachers who consistently employed domain inconsistent social messages were rated lower by students than domain concordant teachers. These ratings are associated with student compliance with teacher directives.

Attention to domain has also been found to impact the effective use of the academic curriculum as a medium for moral education (Nucci, 2001). This was demonstrated in a study in which high school teachers led students in discussion of moral, social conventional and mixed domain issues drawn from their high school English composition and American history courses. In one condition teachers directed students in discourse consistent with the domain of the social issues. Discussion of moral issues (e.g., slavery) centered on considerations of fairness and human welfare, while discussions around conventional issues (e.g., modes of dress) focused upon the function of social norms for establishing social order and societal structure. In a second condition teachers structured the discussion in a domain-discordant fashion focusing upon social norms for the conventional issues, and fairness and welfare for issues of convention. In a third condition the teacher simply used one or the other mode irrespective of the domain of the issue under discussion. Findings from the study demonstrated that development of students' concepts of morality and societal convention was optimal in the domain concordant condition. Moreover, only those students who participated in the domain appropriate learning context demonstrated the ability to spontaneously coordinate moral and societal considerations when interpreting and evaluating mixed domain issues involving both conventional and moral elements (Nucci, 2001).

Community Service and the Moral Self

One major challenge for moral education is translating classroom and school experiences into changes in moral behavior. One school of thought maintains that the bridge between moral judgment and action is the integration of moral judgments with the individual's moral aspects of personal identity and sense of self (Blasi, 2005; Colby and Damon, 1992). The formation of moral identity is thought merge in early adolescence as part of the more general process of identity formation in adolescence and early adulthood (Hart, 2005). Despite misgivings that have been expressed about the reduction of moral motivation to self-consistency (Nucci, 2004), one author has made a thoughtful case for the development of moral identity as the aim of moral education (Lapsley, 2008). The vehicle that has gained most attention as having potential for impacting moral identity and behavior has been the extension of school-based social experiences to include community service activities (Hart et al., 2008).

Efforts to promote youth engagement in community service are not new, and can be traced back in the United States to William James who regarded service to the broader community and nation as the 'Moral Equivalent to War' (Killen and Horn, 2000). When community service is linked to the school curriculum in the form of mandated requirements it is given the label 'service learning.' Programs for service learning are diverse and include a range of activities from tutoring or coaching youth sports to working in a community soup kitchen. For this reason the impact of service learning has been somewhat difficult to evaluate. Nonetheless, some conclusions have begun to emerge indicating that well-run service-learning programs lead to an increased level of civic engagement (Andersen, 1998; Hart et al., 2008; Youniss et al., 2001; Youniss and Yates, 1997), and are associated with an increase in positive moral action and a decrease in rates of delinquent conduct (Hart et al., 1998; Hart et al., 2008).

At least two factors appear to be important for service learning to positively impact students' social growth. The first is that some degree of choice as to the nature of the particular activity a student will take part in seems to be important for the activity to have positive benefit (Andersen, 1998; Barker and Eccles, 1997; Nucci, 2001). The importance of choice is linked to student motivation as well as identity formation. In this regard, Leming (Nucci, 2001: p. 214) reported that when programs force students into so-called 'voluntary' service activities, they result in alienation among adolescents who see through what they perceive to be blatant attempts by schools to make them into particular sorts of people.

The second critical factor is that the service-learning activity be linked to a period of reflection (Andersen, 1998; Youniss et al., 2001; Youniss and Yates, 1997). The degree of reflection varies from program to program from such minimal activities as writing a one-page reaction paper to keeping a daily journal (Killen and Horn, 2000). The importance of reflection rests on

the opportunities it affords for students to share feelings and experiences and to connect those experiences to their own social values and sense of self (Andersen, 1998). Killen and Horn (2000) suggest that such reflection engages students in considering issues of fairness and social inequality that foster moral growth, and may also touch upon consideration of the ways in which conventional stereotypes constrain outgroups such as the poor and disabled. Opportunities for reflection upon service activities may also engage students in discovering one's connections to others and concomitant sense of responsibility and efficacy (Killen and Horn, 2000). Finally, it has been proposed that the effectiveness of 'service learning' to foster the connection between moral actions and personal identity stems from the student's own reflection upon the ideological assumptions of service agencies, and the moral impact of service activities in relation to the values maintained by the individual student (Youniss et al., 2001; Youniss and Yates, 1997).

Conclusions

Since the 1930s the research on children's moral development has guided moral education and promoted it as an alternative to traditional forms of character education. Recent research has shown that such traditional efforts are largely ineffective at impacting attitudes and behavior above the effects of widely recognized 'best teaching' practices that emphasize classroom community, student intellectual autonomy, and high levels of academic instruction (Social and Character Development Research Consortium, 2010). The research on moral and social development has supported the general contention that moral education requires attention to students' construction of moral and social judgments and not solely the acquisition of societal standards or 'virtuous' habits. This recent work indicates that moral concerns for justice and human welfare are available from early childhood and forms the focal point for moral education. At the same time, however, current research indicates that moral development does not move toward an end point where moral principle triumphs over nonmoral considerations. Thus, moral education requires a variegated approach to preparing students to handle moral complexity with respect to personal behavior and efforts to construct a just and compassionate society. Finally, recent work has suggested that education for moral development should also include experiences beyond the classroom that engage students in activities connecting them in substantive and personal ways with the broader community.

Bibliography

Andersen, S.M., 1998. Service Learning: A National Strategy for Youth Development. A Position Paper Issued by the Task Force on Education Policy. Institute for Communitarian Policy Studies, George Washington University, Washington, DC.

Arsenio, W., Lover, A., 1995. Children's conceptions of socio-moral affect: happy victimizers, mixed emotions, and other expectancies. In: Killen, M., Hart, D. (Eds.), Morality in Everyday Life. Cambridge University Press, New York, pp. 87–130.

Barker, B., Eccles, J., April 1997. Student Council Volunteering, Basketball, or Marching Band: What Kind of Extracurricular Involvement Matters? Symposium: Adolescent Involvement in Community. Biennial Meeting of the Society for Research in Child Development, Washington, DC.

Battistich, V., Solomon, D., Watson, M., Schaps, E., 1997. Caring school communities. Educational Psychologist 32, 137–151.

Berkowitz, M., Gibbs, J., 1983. Measuring the developmental features of moral discourse. Merrill-Palmer Quarterly 24, 399–410.

Blasi, A., 1980. Bridging moral cognition and moral action: a critical review of the literature. Psychological Bulletin 88, 1–45.

Blasi, G., 2005. Moral character: a psychological approach. In: Lapsley, D., Power, F.C. (Eds.), Character Psychology and Character Education. University of Notre Dame Press, Notre Dame, IN, pp. 18–35.

Blatt, M., Kohlberg, L., 1975. The effects of classroom moral discussion upon children's level of moral judgment. Journal of Moral Education 4, 129–161.

Carpendale, J., Krebs, D., 1992. Situational variation in moral judgment. In a stage, or on a stage? Journal of Youth and Adolescence 21, 203–224.

Carr, D., 2008. Character education as the cultivation of virtue. In: Nucci, L., Narvaez, D. (Eds.), Handbook of Moral and Character Education. Routledge, New York, pp. 99–116.

Colby, A., Damon, W., 1992. Some Do Care: Contemporary Lives of Moral Commitment. Free Press, New York.

Damasio, A., 1996. Descartes' Error: Emotion, Reason and the Human Brain. Avon, New York.

Damon, W., 1988. The Moral Child. The Free Press, New York.

Deci, E., Ryan, R., 1985. Intrinsic Motivation and Self-Determination in Human Behavior. Plenum Press, New York.

DeVries, R., Zan, B., 1994. Moral Classrooms, Moral Children: Creating a Constructivist Atmosphere for Early Education. Teachers College Press, New York.

Durkheim, E., 1925/1961. Moral Education. The Free Press, Glencoe, IL.

Eckman, P., 1993. Facial expressions and emotion. American Psychologist 48, 384–392.

Emde, R., Hewitt, J.K., Kagan, J., 2001. Infancy to Early Childhood: Genetic and Environmental Influences on Developmental Change. Oxford University Press, Oxford, UK.

Gilligan, C.J., 1982. In a Different Voice. Harvard University Press, Cambridge.

Greene, J., Haidt, J., 2002. How (and where) does moral judgment work? Trends in Cognitive Sciences 6 (12), 517–523.

Habermas, J., 1991. Moral Consciousness and Communicative Action. MIT Press, Cambridge.

Haidt, J., 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological Review 108, 814–834.

Hart, D., 2005. The development of moral identity. Nebraska Symposium on Motivation 51, 165–196.

Hart, D., Atkins, R., Ford, D., 1998. Urban America as a context for the development of moral identity. Journal of Social Issues 54, 513–530.

Hart, D., Matsuba, M.K., Atkins, R., 2008. The moral and civic effects of learning to serve. In: Nucci, L., Narvaez, D. (Eds.), Handbook of Moral and Character Education. Routledge, New York, pp. 484–499.

Izard, C., 1986. Approaches to developmental research on emotion-cognition relationships. In: Bearison, D., Zimiles, H. (Eds.), Thought and Emotion: Developmental Perspectives. Erlbaum, Hillsdale, NJ, pp. 21–38.

James, R., Blair, R., 2005. Morality in the autistic child. Journal of Autism and Developmental Disorders 26, 571–579.

Killen, M., Breton, S., Ferguson, H., Handler, K., 1994. Preschooler's evaluations of teacher methods of intervention in social transgressions. Merrill-Palmer Quarterly 40, 399–416.

Killen, M., Horn, S., 2000. Facilitating children's development about morality, community and autonomy: a case for service-learning experiences. In: van Haaften, W., Wren, T., Tellings, A. (Eds.), Moral Sensibilities and Education II: The Schoolchild. Concorde Publishing, Bemmel, The Netherlands, pp. 89–115.

Kohlberg, L., 1984. Essays on Moral Development. In: The Psychology of Moral Development, vol. 2. Harper and Row, San Francisco.

Kohlberg, L., Turiel, E., 1971. Moral development and moral education. In: Lesser, G. (Ed.), Psychology and Educational Practice. Scott Foresman, Chicago.

Lahat, A., Helwig, C., Zelazo, P., 2012. Age-related changes in cognitive processing of moral and social conventional violations. Cognitive Development 27, 181–194.

Lapsley, D., 2008. Moral self-identity as the aim of education. In: Nucci, L., Narvaez, D. (Eds.), Handbook of Moral and Character Education. Routledge, New York, pp. 30–52.

Leslie, A., Mallon, R., Dicorcia, J., 2006. Transgressors, victims, and cry babies: is basic moral judgment spared in autism? Social Neurosciences 1, 270–283.

Ludemann, P., 1991. Generalized discrimination of positive facial expressions by 7-month old infants. Child Development 62, 55–67.

Noddings, N., 2002. Educating Moral People: A Caring Alternative to Character. Teachers College Press, New York.

Noddings, N., 2008. Caring and moral education. In: Nucci, L., Narvaez, D. (Eds.), Handbook of Moral and Character Education. Routledge, New York, pp. 161–174.

Nucci, L., 2001. Education in the Moral Domain. Cambridge University Press, New York.
Nucci, L., 2004. Reflections on the moral self construct. In: Lapsley, D.K., Narvaez, D. (Eds.), Moral Development, Self, and Identity. Lawrence Erlbaum, Mahwah, NJ, pp. 111–132.
Nucci, L., 2009. Nice Is Not Enough: A Teachers' Guide to Education in the Moral Domain. Merrill, Columbus, OH.
Nucci, L., 2014. The personal and the moral. In: Killen, M., Smetana, J. (Eds.), Handbook of Moral Development, second ed. Routledge, New York, pp. 538–558.
Nucci, L., Turiel, E., 2009. Capturing the complexity of moral development and education. Mind, Brain, and Education 3, 151–159.
Piaget, J., 1932. The Moral Judgment of the Child. Free Press, New York.
Pizzaro, D.A., Bloom, P., 2003. The intelligence of moral intuitions: comment on Haidt (2001). Psychological Review 110, 193–196.
Power, F.C., Higgins-D'Alessandro, A., 2008. The just community approach to moral education and the moral atmosphere of the school. In: Nucci, L., Narvaez, D. (Eds.), Handbook of Moral and Character Education. Routledge, New York, pp. 230–247.
Power, F.C., Higgins, A., Kohlberg, L., 1989. Lawrence Kohlberg's Approach to Moral Education. Columbia University Press, New York.
Shroufe, A., 1988. The role of infant-caregiver attachment in development. In: Belsky, J., Nezworski, T. (Eds.), Clinical Implications of Attachment. Erlbaum, Mahwah, NJ, pp. 18–38.
Smetana, J., 2006. Social-cognitive domain theory: consistencies and variations in children's social and moral judgments. In: Killen, M., Smetana, J. (Eds.), Handbook of Moral Development. Erlbaum, Mahwah, NJ, pp. 119–154.
Smetana, J.G., Braeges, J.L., 1990. The development of toddlers' moral and conventional judgments. Merrill Palmer Quarterly 36, 329–346.
Snarey, J., Keljo, K., 1991. In a gemeinschaft voice: the cross-cultural expression of moral development theory. In: Kurtines, W.M., Gewirtz, J.L. (Eds.), Handbook of Moral Behavior and Development, vol. 2. Lawrence Erlbaum and Associates, Hillsdale, NJ, pp. 395–424.
Snarey, J., Samuelson, P., 2008. Moral education in the cognitive-developmental tradition: Lawrence Kohlberg's revolutionary ideas. In: Nucci, L., Narvaez, D. (Eds.), Handbook of Moral and Character Education. Routledge, New York, pp. 53–79.
Social and Character Development Research Consortium, 2010. Efficacy of Schoolwide Programs to Promote Social and Character Development and Reduce Problem Behavior in Elementary School Children (NCER 2011–2001). National Center for Education Research, Institute of Education Sciences, U.S. Department of Education, Washington, DC.
Turiel, E., 1983. The Development of Social Knowledge: Morality and Convention. Cambridge University Press, Cambridge, England.
Turiel, E., 2002. The Culture of Morality: Social Development, Context, and Conflict. Cambridge University Press, Cambridge, UK.
Turiel, E., 2010. Snap judgment? Not so fast: thought, reasoning and choice as psychological realities. Human Development 53, 105–109.
Walker, L., 2006. Gender and morality. In: Killen, M., Smetana, J. (Eds.), Handbook of Moral Development. Erlbaum, Mahwah, NJ, pp. 93–115.
Watson, M., 2003. Learning to Trust: Transforming Difficult Elementary Classrooms through Developmental Discipline. Jossey-Bass, San Francisco, CA.
Watson, M., 2008. Developmental discipline and moral education. In: Nucci, L., Narvaez, D. (Eds.), Handbook of Moral and Character Education. Routledge, New York, pp. 175–203.
Wynne, E., Ryan, K., 1993. Reclaiming Our Schools: A Handbook on Teaching Character, Academics, and Discipline. Macmillan Publishing Company, New York.
Youniss, J., McLellan, J.A., Mazer, B., 2001. Voluntary service, peer group orientation, and civic engagement. Journal of Adolescence Research 16, 456–468.
Youniss, J., Yates, M., 1997. Community Service and Social Responsibility in Youth. University of Chicago Press, Chicago.

Moral Development, Cultural Differences in

Klaus Helkama and Florencia M Sortheix, University of Helsinki, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Drawing on current conceptualizations of values (Schwartz's basic human values theory) and the moral domain (Haidt's moral foundations theory; Shweder's big three of morality), as well as the cognitive developmental approach to morality, we examine research on cross-cultural differences in the emphases on different parts of the moral domain and in the pathways of moral development in the light of cultural variations in wealth, relatedness, and hierarchy. We also review studies of age-related changes in values and justice-oriented reasoning, and examine influences on cultural and individual inclusiveness of the moral ingroup.

The Moral Domain

Morality has two functions: to promote ideals and justice, and to maintain social stability. Among the classics in psychology and sociology, Piaget and Weber stressed the former, Freud and Durkheim the latter function. The current dominant definition of the moral domain (Haidt, 2012) includes six moral foundations: harm/care (not hurting anybody, caring for the weak), liberty (freedom from oppression), fairness/reciprocity (equal treatment of all), ingroup/loyalty (not betraying one's group, being proud of the history of one's country), authority/respect (showing respect for authority), and purity/sanctity (not violating standards of decency). The first three foundations are, roughly, related to the promotion function of morality, the three latter to its prevention function. Another taxonomy of moral ideologies is Shweder et al.'s (1997) 'big three': Ethic of autonomy (corresponds to Haidt's care, liberty, and fairness), community (ingroup and authority), and divinity (purity). Obviously, the foundations could be in conflict, for instance, in a choice between ingroup loyalty and honesty. In line with the two opposite functions of morality, politically conservative people in the United States emphasize the three foundations related to stability (loyalty, authority, and purity), whereas liberals stress the change-oriented harm and fairness foundations.

Values, abstract, general, desirable goals, represent what is most important in life. They are the foundations based on which people live and relate to others (Schwartz, 1992). Studies relying on Schwartz's (1992) value model suggest that there is also universal agreement that some values are moral. His model postulates 10 universal values. They can be described in terms of a circular structure with two dimensions, self-enhancement – self-transcendence and conservation – openness to change. Values relate in consistent ways to moral foundations (Boer and Fischer, 2013). Benevolence is regarded as a moral value in widely different cultures (Vauclair, 2010), and is positively associated with all five moral foundations among more than 34 000 respondents from all over the world (Graham et al., 2011). Conformity and tradition-related values (self-discipline, politeness, honoring of parents and elders; being devout, respect for tradition) are related to the two foundations of authority, and purity (Boer and Fischer, 2013), possibly also to ingroup (Sverdlik et al., 2012), see **Table 1**.

Except for benevolence, the moral domain varies across cultures. Conformity and tradition are more often regarded as moral values in less wealthy than in affluent countries, whereas universalism values (e.g., broadmindedness, equality, world peace, social justice, protecting environment) and even some self-direction values (freedom) are regarded as moral values more often in wealthy individualistic than in poor collectivistic countries (Sverdlik et al., 2012; Vauclair, 2010).

Morality is neither sociologically nor psychologically a coherent system. Max Weber pointed out in 1918 that our conceptions of justice involve a fundamental antinomy or incompatibility, between egalitarianism and merit. Neuropsychological research shows that we regard it as right to kill one person by turning a switch in order to save five people, but regard it as wrong to kill one person by pushing him from a bridge to save five persons. The thought of pushing another activates the 'old' regions in our brain, whereas the thought of turning the switch activates the 'new' parts of the brain.

However, there are universally shared moral intuitions, for instance, that dishonesty is wrong. Vauclair and Fischer (2011) reviewed responses of more than 200 000 people from 84 countries to the 'morally debatable behaviors' scale, used in the World Value Survey since 1981. By means of multilevel analyses they found that there was very little variation across cultures or across individuals in the opinion that such 'dishonest and illegal' acts as accepting a bribe were wrong. By contrast, there was a great deal cross-cultural as well as within-culture variation with regard to 'personal and sexual' acts like abortion, homosexuality, or divorce.

Table 1 The moral domain in different psychological approaches

Schwartz's values	Self-direction, universalism, benevolence	Conformity	Tradition
Haidt's Moral foundations	Liberty, fairness, care	Ingroup/loyalty, authority/respect	Purity
Schweder's 'big three'	Autonomy	Community	Divinity

Moral development is here defined as any change in a morally relevant psychological variable that occurs in human beings as they age. Culture is here seen as the primary mode of human adaptation, which reflects the demands of the ecological, demographic, and institutional environment, as well as historical experiences. Culture and development are intertwined in many senses. Culture sets the developmental tasks for individuals and largely defines developmental pathways of development.

A hunting and gathering culture is egalitarian and low in conformity, often with benevolent attitudes to outgroups. A culture adapted to an agrarian socioeconomic environment is organized by the dimension of relatedness, where the individual is part of the social system, defined by cooperation, acceptance of hierarchy, striving for harmonious relationships. A culture, adapted to an urban socioeconomic postindustrial environment, is organized by the dimension of autonomy; the individual is self-contained in a social system defined by a mixture of competition and cooperation, and plurality of values. We examine sociocultural contexts that range from countries and geographic areas to smaller units such as professions or families. In characterizing those contexts, the focus will be on the dimensions of relatedness – expressed by autonomy vs. embeddedness (Schwartz, 2008) – and hierarchy – egalitarianism (expressed by Schwartz's and Hofstede's well-known dimensions) as well as economic affluence of the social unit.

Pathways of Moral Development

As the moral domain varies cross-culturally, so do pathways of moral development. This is seen already in infancy. In cultures where relatedness values (conformity, tradition) are emphasized, socialization goals and parenting strategies of mothers for their infants differ from the goals and strategies of mothers in postindustrial cultural contexts (Keller, 2007). There is no explicit developmental cross-cultural research on the authority, ingroup, and purity foundations. Jensen (2011) advanced a cultural–developmental theory, in which she outlined a series of developmental trajectories for the autonomy, community, and divinity in different cultural contexts. She noted that community-related arguments increase with age, as shown by the findings that in late adolescence and adulthood, there is an increase in moral concepts associated with societal organization, and with what is best for the society. These results are found not only in western countries, but also in India and Zambia. Adulthood age trends from representative national samples are consistent with this view, as they demonstrate an increase in the importance of conformity and security values.

Wealth has two important consequences for values and moral development pathways. First, it increases the likelihood of individualism and democracy (causality is reciprocal). Inglehart and Welzel (2005) found that the ratio of societal changes toward democracy vs. changes toward autocracy for the period 1950–90 increased with wealth. In countries with per capita incomes greater than 7000 dollars, changes toward democracy occurred 28 times as often as the reverse changes. Even within democracies, wealth increases preference for fairness and care (self-transcendence values). Van Herk and Poortinga (2012) examined 195 regions in Europe comparing the influence of historical factors and current economic situation. They found that contemporary affluence, measured by the gross national product (GNP), was far more important than history in explaining preference for self-transcendence values (benevolence and universalism). Using the third wave of the European Social Survey, we calculated correlations of universalism with age. Despite the confounding of cohort effects, correlations were positive in all countries, ranging from 0.12 (Denmark) to 0.31 (Russia). Democracies, then, seem to favor the moral developmental trajectory toward universalism values.

A second consequence of wealth is increasing value differentiation (Boer and Fischer, 2013). In wealthier and more individualistic societies, there is less value consensus. Schwartz and Sagie (2000) suggest that with increasing democratization, value consensus further decreases. This means that a new developmental task emerges: how to solve value conflicts? That is the focus of Kohlberg's (1984) cognitive–developmental theory of moral judgment.

Kohlberg's Theory of Moral Development

The cross-cultural studies on Kohlberg's developmental theory of moral judgment are far more numerous than on any other approach (see, e.g., Eckensberger and Zimba, 1997). His basic focus was moral judgment, which is the ability to reason about value conflicts. He attempted to formulate a system of universal moral categories. Kohlberg's category system covered, broadly, Haidt's (2012) six foundations. In Kohlberg's Moral Judgment Interview (MJI), two commonly accepted moral principles clashed in a dilemma, for instance, saving a life vs. obeying the law, or authority vs. honesty (promise keeping).

Moral development was defined by the justifications (structure) people gave for their choices, not by the choices (content). The justifications were rank ordered into stages according to their cognitive sophistication and level of social perspective taking. Thus, the criteria of development were more stringent than mere age-related change. After a radical revision of the scoring system, the final sequence consisted of five empirical stages (preconventional: (1) heteronomous morality, punishment, and obedience to authority; (2) individualism, instrumental purpose, and exchange; conventional: (3) mutual interpersonal expectations, relationships, and interpersonal conformity; (4) social system and conscience; and post-conventional: (5) social contract or utility and individual rights).

Although Kohlberg initially thought that development would lead to agreement on universal ethical principles at Stage 6, this stage was not found in empirical studies. At the post-conventional level, there is no value consensus but post-conventional arguments could be employed to defend a variety of values. While research has supported many Kohlbergian criteria of development, e.g., the hierarchical nature of the stages, there is more situational variation in the arguments than Kohlberg's idea of stages as integrated wholes implied. Therefore, current usage relies on the concept of a schema: personal interest schema for Stage 2/3, maintaining norms

Table 2 Developmental approaches to morality

Fairness
Kohlberg (1984): Stages of moral reasoning
Focus: Solving value conflicts presented in hypothetical dilemmas
Methods of assessment:
- Moral Judgment Interview (MJI; Colby and Kohlberg: The measurement of moral judgment, Cambridge University Press, 1987; production measure)
- Defining Issues Test (DIT; Rest: Development in judging moral issues, 1979: comprehension and preference) DIT-2 (Rest et al., 1999)
- Sociomoral Reflection Measure (SRM; Gibbs, Widaman, and Colby, Child Development, 1982)
- Sociomoral Reflection Measure-Short Form (SRM-SF; Gibbs, Basinger, and Fuller, Moral maturity: measuring the development of sociomoral reflection, Erlbaum, 1992). (The SRM-SF, unlike the other instruments, is not based on dilemmas.)
- Moral Judgment Test (MJT; assessment of moral arguments); Lind (www.uni-konstanz.de/ag-moral/lind.htm)

Care
Gilligan: In a different voice (1982)–Skoe (1998): Ethic of care
Focus: Solving dilemmas on care of others vs. care of self (partly own real-life dilemmas)
Countries: Norway, Canada, Finland (Juujärvi, Myyry, and Pesso, J. Moral Education, 2010)
Method of assessment: Ethic of Care Interview (Skoe et al. Early Adolescence, 1999)
Eisenberg et al. (Developmental Psychology, 1987): Prosocial reasoning
Focus: Reasons for prosocial behavior
Countries: USA, Brazil (Eisenberg, Zhou, and Koller, Child Development, 2001)
Method of assessment: Prosocial dilemma interview

Community, autonomy, and divinity (the 'big three')
Jensen: Coding manual for the ethics of autonomy, community, and divinity (2004).www.LeneArnettJensen.com
Focus: Coding moral reasons from moral discourse
Guerra and Giner-Sorolla: Community, Autonomy, Divinity Scale (CADS), Journal of Cross-Cultural Psychology, 2010

Moral exemplarity
Studying morally outstanding persons who have shown extraordinary commitments promoting benevolence, justice, or basic human welfare (Life review interview; List of personal strivings; Frimer, Walker, and et al., Journal of Personality and Social Psychology, 2011)

schema (Stage 3/4), and postconventional schema (Rest et al., 1999).

Most importantly, there is empirical support for the assumption that higher stage solutions to moral problems are felt to solve those problems in a more balanced and just way. When presented with Stage 5 postconventional arguments and Stage 4 conventional arguments, both right-wing and left-wing students in Britain prefer Stage 5 to Stage 4. High scorers are less self-serving in judging their own and others' bad deeds and more likely to rely on the procedural justice principle of bias suppression in reporting personal real-life dilemmas (Myyry and Helkama, 2002). However, high scorers are hardly more likely to resist temptation or show altruism or moral responsibility than others.

Kohlberg's theory conceptualizes environmental influence in terms of social perspective-taking opportunities (participation in peer groups, decision making in the family, and sense of participation in societal institutions of work and politics in general). These experiences stimulate the growth of justice reasoning. The structural nature of the stages implies that development is mainly about increasing sophistication, i.e., differentiation and hierarchical integration of moral concepts, as well as increasing universality, but does not necessarily lead to adoption of any definite values.

The variety of approaches and measures of different aspects of moral development are summarized in Table 2. Findings on comprehension and preference formed the basis of Rest's Defining Issues Test (DIT), which measures the preference for postconventional schema and has been frequently used in cross-cultural research. Gibbs's measures, on the other hand, deal with development up to Stage 4.

Is there evidence for the Rousseau-derived hypothesis of a 'natural' universal trend toward postconventional Stage 5 judgment? Studies carried out with the MJI, where respondents produce moral arguments, show that Stage 5 is extremely rare (Eckensberger and Zimba, 1997; Snarey, 1985). By contrast, studies that have used the DIT, a comprehension and preference test, suggest that the preference for postconventional arguments increases with age and education in a number of countries (Gielen and Markoulis, 2001). The postconventional schema reflects an outside-society or prior-to-society perspective, as illustrated by such items of the DIT as "would (X) in this case bring about more total good for the whole society or not" or "what values are going to be the basis of governing how people act toward each other."

Social perspective taking has been shown to be the factor underlying the progress along the stages of justice reasoning judgment, irrespective of method of assessment (Gibbs et al., 2007). The same holds for the ethic of care. Therefore, when looking at culture-based differences in Kohlbergian moral development, the natural question is: What are the dimensions of culture that are related to social perspective-taking opportunities and actual perspective taking?

Culture-Based Differences in Justice Reasoning

The prototypes of Gesellschaft (society) and Gemeinschaft (community) have been used as explanatory constructs for cultural differences in the attainment of higher stages (Snarey, 1985). Gesellschaft, an urban, complex, heterogeneous environment, with a variety of contacts with strangers, supposedly provides more varied social perspective-taking opportunities than does Gemeinschaft, rural, face-to-face, homogeneous environment, where contacts are mostly with familiar people. Movement of any ecological variable (e.g., diversity of occupations or formal schooling) in a Gesellschaft direction shifts cultural values toward more individualism and developmental pathway toward more independent behavior (Greenfield, 2009).

However, in addition to a movement toward a larger society (Gesellschaft), the cultural value of autonomy vs. embeddedness could be related to perspective taking. One prerequisite of social perspective taking is the existence of value plurality, which means increasing likelihood of encountering situations where different values are represented by different people, i.e., in a sociocognitive conflict. Experimental studies show that sociocognitive conflict leads to cognitive development. In autonomous societies individuals are socialized for independent thinking and personal freedom is highly valued. In more

embedded societies religion, group norms and social conventions govern everyday life giving less freedom for individual choices in moral justice-oriented reasoning. The weaker value consensus in more democratic societies implies that individuals growing up in such societies are more exposed to sociocognitive value conflicts, so we could expect that they score higher on moral judgment tasks.

The role of egalitarian vs. hierarchical cultural values also seems to influence moral development. Hofstede's power distance (hierarchy) refers to the extent to which those lower in organizational hierarchy regard inequality as natural. **Table 3** shows that moral judgment development scores tend to be higher in countries which score lower on power distance. The upper half of **Table 3** shows the scores for the highest scoring (higher middle class) adults in Snarey's (1985) review of studies using the Moral Judgment Interview, and the bottom half shows the *P*% (percentage of postconventional arguments regarded as most important) scores of freshman samples from different countries in Gielen and Markoulis' (2001) review, with Spearman rho = 0.80 ($p < .01$). The notable exception from the pattern in **Table 3** is Ireland, low in power distance but also low postconventional morality. This is probably due to the influence of Catholic Church.

Religious orthodoxy is associated with lower postconventional scores (Rest et al., 1999). Where a high proportion of moral rules are seen as absolute, there is less room for value conflicts.

The role of egalitarianism is also seen in smaller cultural units; e.g., families. Adolescents who participate in decision-making score higher on the DIT than adolescents in hierarchical role–oriented families. Husbands and wives who make decisions together score higher on the MJI than those who do not. Although these findings could be due to the tendency of individuals high on justice-oriented reasoning to act democratically, the findings on university students' moral judgment development suggest that the causal link goes from a hierarchical culture to individuals.

The US business world is hierarchical, and the business students score lower than expected on measures of moral reasoning (Maeda et al., 2009). Also the medical profession is known to be fairly hierarchical. In contrast to students in other fields of study, European and Canadian medical students do not appear to progress on moral judgment. This is suggested by findings of regression in at least four recent longitudinal studies, which used different Kohlbergian methods to assess justice-oriented reasoning. Presumably, for the medical profession, like other life-and-death occupational groups (fire brigades, armies), a hierarchical organization is often a natural one for effective functioning. Solving value conflicts is not easily compatible with its functional requirements. However, the hierarchical organization of medical profession and its consequences on moral reasoning is no natural law, as shown by findings of clear progress among medical students in the United States (Maeda et al., 2009). Likewise, business school students in Finland obtained high scores on postconventional moral judgment in the 1980s.

Over the past decades, inequality has increased almost everywhere in the Western world. The findings that in the United States, university students' *P*% scores have declined, narcissism increased and empathy decreased are consistent with each other in suggesting a weakened disposition of taking into account the feelings and consequences of own actions for other people.

Democracies provide a sense of participation in the political process, an essential input for the attainment of the system perspective and the prior-to-society perspective, on which

Table 3 Moral judgment development indices and power distance

	MMS	*Power distance*
A. Kohlberg's Moral Judgment Interview (MMS); Snarey (1985)		
Israel (Kibbutz)	458	13
USA	391	40
Taiwan	370	58
India	363	77
	P%	*Power distance*
B. Defining Issues Test (*P*%; Gielen and Markoulis, 2001, p. 87)		
Australia	43.8	36
USA	42.3	40
South Korea	41.5	60
Taiwan	41.4	58
Greece	40.3	60
Poland	39.9	68
Hong Kong	37.9	68
Ireland	34.1	28
Nigeria (West Africa)	30.0	77
Sudan	27.8	80
Kuwait	24.6	80
Egypt	22.6	80

MMS; moral maturity score.
P% = weighted percentage of postconventional arguments ranked among the most important ones.

Kohlberg's Stages 4 and 5 are based. On the other hand, a substantial proportion of citizens with the Stage 4 and 5 systems and prior-to-society perspective seems to be a prerequisite of a well-functioning democracy. These perspectives enable them to see the common good in the midst of competing values and ingroup as well as religious loyalties. To summarize, Gesellschaft (as opposed to Gemeinschaft), autonomy (as opposed to embeddedness), brought about by sufficient wealth, as well as democracy and egalitarianism in the sense of low power distance (as opposed to autocracy, hierarchy, and high power distance) seem to be those dimensions of culture that are conducive to individual development of justice-oriented reasoning.

Other Developmental Dimensions

After Kohlberg, several moral–developmental dimensions have emerged (Table 2). Far less cross-cultural research exists on these dimensions than on Kohlberg's justice. An attempt to build a comprehensive moral–developmental psychology, relying on neo-Kohlbergian findings and brain research is Narvaez's (2008) Triune ethics. It unites findings from brain and clinical research with studies of moral emotions, and singles out three basic ethical mindsets, which represent global brain states. Ethic of safety, shared with all animals, is based primarily on instincts for survival, and located in the brainstem and lower limbic system. Ethic of engagement is based on mammalian emotional systems that guide people toward intimacy, as e.g., in parental care, or play in a peer group. Ethic of imagination relies on recently evolved parts of the brain (neo cortex and prefrontal cortex) and makes possible will power as well as moral problem solving by means of reflection. How much one favors one mindset over another is shaped by experience in early life (i.e., caregiving practices) and other sensitive periods.

Moral Values and Development of Justice-Oriented Reasoning

If the structural development of moral judgment does not lead to agreement about the hierarchy of moral principles or moral values, then it is natural to ask how value priorities and moral judgment development are related. Few associations of value priorities with Kohlbergian moral development have been published, which is consistent with the idea of the independence of structure (moral stage) and content (value). A review (Helkama, 2011) found that 70% of the 24 significant positive associations of Kohlbergian measures with values were for universalism values, and the rest were evenly divided between benevolence and self-direction.

Conformity values have shown negative correlations with Kohlbergian measures of moral development in educationally homogenous samples. Universalism values are by definition universalistic in the sense of being related to concern and protection for the welfare of all people and nature. Benevolence and conformity are ingroup values, concerned with the welfare of close ones and restraint from violation of ingroup norms, respectively. The association of universalism with the development of moral reasoning could be attributed to the close association of universalism with social perspective taking and empathy.

Moral Inclusiveness, Country-Level, and Individual-Level

The two ingroup values, benevolence and conformity, while motivationally close, relate to perspective taking and empathy differently. Across eight studies in four countries, benevolence showed consistent associations with both cognitive perspective taking and emotional empathy, while conformity did not show any associations (Helkama, 2011). The fact that the patterns in the associations of benevolence and universalism are very similar – almost identical – could be used to argue that universalism could be understood as sympathy extended beyond the boundaries of the ingroup, to all people. This idea is supported by the notion of the inclusiveness of the moral universe (Schwartz, 2007), which is defined as the extent toward which universalism values are applied to outgroup members. At the country level, inclusiveness increases with egalitarianism and democratization and decreases with embeddedness.

Schwartz (2007) operationalized the notion of the inclusiveness of moral universe by looking at the separation of benevolence and universalism values from a psychometric point of view. He took as his starting-point four universalism values, broadmindedness, equality, social justice, and a world at peace. If all four items were intermixed with benevolence or other moral value items, the inclusiveness score was 0, and if all of them formed a distinct region, then it was 4. Among 66 countries, democratization was the strongest predictor ($r = 0.73$) of inclusiveness. In a sample of 21 European countries, inclusiveness of moral universe predicted acceptance of immigrants. Moreover, the country inclusiveness moderated the associations of universalism with acceptance of outgroup immigrants and participation in prosocial activity, so that the association was stronger in those countries where moral universe was more inclusive, i.e., where universalism values were understood as applying to members of outgroups, too.

The inclusiveness of an individual's moral ingroup should increase with his or her moral judgment level, if Kohlberg's stages follow the Kantian notion of universality, as he assumed. Because history provides many examples of members of highly developed cultures whose belief in their moral superiority has led them to barbarous acts toward outgroup members, it is appropriate to ask whether and when high scores on Kohlbergian measures of moral judgment predict inclusiveness of the moral ingroup. Two studies seem relevant.

One study (Helkama, 2009) examined opinions about development cooperation in relation to Kohlberg's moral stages at two points of time among ordinary adults in a Finnish rural community. When economy was thriving (1982), moral judgment level predicted inclusiveness of the moral universe in the sense that those who scored higher on moral judgment were more likely than low scorers to see contributing to development cooperation as a moral duty. However, during the massive economic recession in 1993, no association was found between moral reasoning and inclusiveness of the moral

ingroup, defined in terms of the same questions on development cooperation. The inclusiveness of the moral universe in that specific rural community in 1993 was much narrower than the Finnish average over 1990s. Thus, the relation of moral judgment development to the inclusiveness of the moral ingroup varies as a function of the economic situation.

The other study (Passini, submitted for publication) addressed ordinary Italians, who were given Rest's DIT and measures of social distance tapping prejudice toward immigrant groups. There were two kinds of dilemmas on the DIT. In some dilemmas, the protagonist was Italian, an ingroup member, in others, Romanian, an outgroup member. Overall, the postconventional score was related to lower prejudice. However, respondents with a narrow moral ingroup (i.e., prejudiced ones) scored lower on postconventional reasoning when the protagonists were Romanian (outgroup members) than in case they were Italian. By contrast, those respondents with an inclusive moral ingroup obtained identical *P*% scores on both types of dilemmas. Inclusiveness of the moral ingroup at the individual level is, accordingly, not reducible to the developmental level of moral judgment.

In sum, these studies suggest, first, that the value choices that higher stage moral arguments justify are susceptible to prevailing value climate and that the inclusiveness of the ingroup may regulate the social perspective-taking process. Categorizing some people as members of the outgroup leads to switching off empathy toward them (Demoulin et al., 2009). Second, they suggest that the inclusiveness of the moral ingroup may vary quickly as a function of the economic situation. Disease stress as well as economic threat narrow the boundaries of the ingroup and increase the importance of conservation values. In the Finnish rural community just mentioned, both local and national identifications increased from 1993 to 2007, simultaneously with the increase in importance of conservation values (Sortheix et al., 2013).

These Finnish observations, then, suggest that societal moral code adapts to economic and other threats by shifting toward relatedness values and narrower ingroups. As value consensus increases, moral judgment in the sense of ability to solve value conflicts is less needed.

Conclusions and Future Directions

The concept of moral development, introduced by Piaget and Kohlberg, is an Enlightenment notion, based on common human nature and equality of all people. The centrality of benevolence values in morality and increase in reasoning about societal organization with age seem to be universal features. Piaget and Kohlberg assumed that there is a 'natural' trajectory of development, which is only diverted by poor institutions. Rousseau's idea of a free and egalitarian society is consistent with observations on hunter-gatherers, who live in an ecological environment where autonomy is needed and equality is functional. By contrast, in conditions with high disease stress and economic scarcity, hierarchy is functional, and the authority, ingroup loyalty, and religious purity (social-stability maintaining) components of morality concur in promoting self-discipline. Increase in economic affluence decreases the need for conformity and creates pressure toward individualism, democracy, and equality. While the weight of conservation values in morality is reduced, no value consensus on universalism (fairness) along the lines of 'natural' Rousseau-like development is observed in wealthy societies, but increased value diversity, also among moral values. 'Human nature' seems equally inclined to believe in ingroup superiority and ingroup loyalty as in the Golden Rule and universal human rights. The moral–developmental trajectory for citizens in postindustrial democracies is characterized by growth of reflective moral problem-solving skills, based on social perspective taking; i.e., the developmental trajectory described by Kohlberg's stage sequence.

Moral change follows the route predicted by Bardi's and Goodwin's (2011) theory of value change: With increase in wealth and societal equality, the domain of moral values moves along Schwartz's circle from conformity toward universalism and even self-direction. In affluent contexts, ingroup norms, which in less affluent environments are all regarded as moral, get differentiated into conventional and moral ones. As the increase in wealth makes democracy and value plurality more likely, also self-direction values (liberty) become to be thought of as moral values.

While the above review suggests that moral codes tend to be rational given the life circumstances (self-discipline is stressed in conditions of scarcity, universalism in affluent societies, which can afford it), continuity of traditions is a force that is obviously also at work. Wealth does not automatically bring about individualistic values and democracy. As Boer and Fischer (2013) show in their review, the increased opportunities that arise from wealth only give rise to more choice through a cultural emphasis on autonomy (as opposed to embedded and traditional societies) and, we would add, egalitarianism. The possibility of choice (value plurality) is important, but so is the responsibility that egalitarian societies foster in their citizens (see Helkama, 2011).

With decrease in wealth, increase of threats in the environment, or increase in inequality, moral change goes to the opposite direction, as illustrated above: Universalism values become less important and conservation values more important. Since moral judgment (in the Kantian and Kohlbergian sense of ability to solve value conflicts) is less required, level of moral judgment goes down. Possibly, in these contexts the responsibility of making such choices is left to the collective or the authority.

The cross-cultural study of moral development has been fairly one-sided in many ways. Most of it has been carried out starting from Kohlberg's theory. Developmental studies have been restricted to care and fairness. In the future, there is a need for more research using other methods than questionnaires or interviews about hypothetical situations. Mixed-method approaches that combine qualitative and quantitative approaches will become more common.

Brain research is one of the most rapidly expanding areas in psychology. Integrating results from these studies (like Narvaez's (2008) Triune Ethics) with cross-cultural data will increase our knowledge of the interaction of affective and cognitive factors in moral development as well as of long-term influence of childhood experiences.

The accumulation of representative databases from earlier time periods and development of more sophisticated

statistical techniques like multilevel modeling make it possible to better understand the short-term historical and temporal dynamics of societal and psychological change in moral development and, e.g., the interplay of normative and value regulation.

The increased heterogeneity of societies implies increase in the number of people who simultaneously belong to groups that differ in their values and moral outlooks. As Sverdlik et al. (2012) point out, one important future direction is to find out how membership in multiple ingroups could be translated into a permissive moral outlook.

See also: Behavioral Problems, Effects of Parenting and Family Structure On; Defiant Behavior During Adolescence across Cultures; Defiant Behavior During Infancy and Early Childhood; Empathy During Early Childhood Across Cultures, Development of; Moral Development, Theories of; Parenting Attitudes and Beliefs across Cultures; Prosocial Behavior, Effects of Parenting and Family Structure On; Self-Regulation During Early Childhood Across Cultures, Development of; Values Across Cultures, Development of.

Bibliography

Bardi, Anat, Goodwin, Robin, 2011. The dual route to value change: individual processes and cultural moderators. Journal of Cross-Cultural Psychology 42, 271–287.

Boer, Diana, Fischer, Ronald, 2013. How and when do personal values guide our attitudes and sociality? Explaining cross-cultural variability in attitude-value linkages. Psychological Bulletin.

Demoulin, Stephane, Leyens, Jacques-Philippe, Dovidio, John (Eds.), 2009. Intergroup Misunderstandings: Impacts of Divergent Social Realities. Psychology Press, London.

Eckensberger, Lutz, Zimba, Roderick, 1997. The development of moral judgment. In: Berry, John, Dasen, Pierre, Saraswathi, T.S. (Eds.), Handbook of Cross-Cultural Psychology, vol. 2. Allyn & Bacon, Boston, pp. 299–338.

Gibbs, J.C., Basinger, K.S., Grime, R.L., Snarey, J.R., 2007. Moral judgment development across cultures: revisiting Kohlberg's universality claims. Developmental Review 27, 443–500.

Gielen, Uwe P., Markoulis, Diomedes C., 2001. Preference for principled moral reasoning: a developmental and cross-cultural perspective. In: Adler, Leonore Loeb, Gielen, Uwe P. (Eds.), Cross-Cultural Topics in Psychology. Praeger, Westport, CT, pp. 81–101.

Graham, Jesse, Nosek, Brian A., Haidt, Jonathan, Iyer, Ravi, Koleva, Spassena, Ditto, Peter H., 2011. Mapping the moral domain. Journal of Personality and Social Psychology 101, 366–385.

Greenfield, Patricia, 2009. Linking social change and developmental change: shifting pathways of human development. Developmental Psychology 45, 401–418.

Haidt, Jonathan, 2012. The Righteous Mind. Why Good People Are Divided by Politics and Religion. Allen Lane, London.

Helkama, Klaus, 2009. Ingroup and outgroup in morality and moral development. In: Jasinskaja-Lahti, Inga, Mähönen, Tuuli-Anna (Eds.), Identities, Intergroup Relations and Acculturation. Gaudeamus, Helsinki, pp. 137–149.

Helkama, Klaus, 2011. Social psychology of morality and moral development: a functional system approach. In: Valentim, Joaquim Pires (Ed.), Societal Approaches in Social Psychology. Peter Lang, Bern, pp. 131–149.

Inglehart, Ronald, Welzel, Cristian, 2005. Modernization, Cultural Change, and Democracy. Cambridge University Press, Cambridge.

Jensen, Lene Arnett, 2011. The cultural–developmental psychology of moral development. In: Jensen, Lene Arnett (Ed.), Bridging Cultural and Developmental Approaches to Psychology. Oxford University Press, New York, pp. 3–25.

Keller, Heidi, 2007. Cultures of Infancy. Erlbaum, Mahwah, NJ.

Kohlberg, Lawrence, 1984. The Psychology of Moral Development. Harper & Row, New York.

Maeda, Yukiko, Thoma, Steve J., Bebeau, Muriel J., 2009. Understanding the relationship between moral judgment development and individual characteristics: the role of educational contexts. Journal of Educational Psychology 101, 233–247.

Myyry, Liisa, Helkama, Klaus, 2002. Moral reasoning and the use of procedural justice rules in hypothetical and real-life dilemmas. Social Justice Research 15, 373–391.

Narvaez, Darcia, 2008. Triune ethics: the neurobiological roots of our multiple moralities. New Ideas in Psychology 26, 95–119.

Passini, Stefano. The Effect of the Group Membership of the Person to be Judged on Moral Reasoning. Manuscript, University of Bologna, submitted for publication.

Rest, James, Narvaez, Darcia, Bebeau, Muriel, Thoma, Steve, 1999. Post-conventional Moral Thinking. A Neo-Kohlbergian Approach. Erlbaum, Mahwah, NJ.

Schwartz, Shalom H., 1992. Universals in the content and structure of values: theoretical, advances and empirical tests in 20 countries. In: Zanna, M.P. (Ed.), Advances in Experimental Social, Psychology, vol. 25. Academic Press, San Diego, CA, pp. 1–65.

Schwartz, Shalom H., 2007. Universalism values and the inclusiveness of our moral universe. Journal of Cross-Cultural Psychology 38, 711–728.

Schwartz, Shalom H., 2008. Cultural Value Orientations: Nature and Implications of National Differences. Publishing House of SU HSE, Moscow.

Schwartz, Shalom H., Sagie, Galit, 2000. Value consensus and importance: a cross-national study. Journal of Cross-Cultural Psychology 31, 465–497.

Shweder, Richard, Much, Nancy, Mahapatra, Manmohan, Park, Lawrence, 1997. The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. In: Brandt, A., Rozin, P. (Eds.), Morality and Health. Routledge, New York, pp. 119–169.

Snarey, John, 1985. Cross-cultural universality of social–moral development: a critical review of Kohlbergian research. Psychological Bulletin 97, 202–232.

Sortheix, Florencia, Olakivi, Antero, Helkama, Klaus, 2013. Values, life events and health: a study in a rural Finnish community. Journal of Community and Applied Social Psychology 23 (4), 331–346.

Sverdlik, Noga, Roccas, Sonia, Sagiv, Lilach, 2012. Morality across cultures: a values perspective. In: Miculincer, Mario, Shaver, Phillip R. (Eds.), The Social Psychology of Morality. APA, Washington, DC, pp. 219–235.

Van Herk, Hester, Poortinga, Ype H., 2012. Current and historical antecedents of individual value differences across 195 regions in Europe. Journal of Cross-Cultural Psychology 43 (8), 1229–1248.

Vauclair, Christine-Melanie, 2010. What Is Right? What Is Wrong? And Does the Answer Tell Us Something about Culture? An Investigation of the Culture and Morality Using the Values, Concept. Unpublished doctoral dissertation, Victoria University of Wellington, NZ.

Vauclair, Christine-Melanie, Fischer, Ronald, 2011. Do cultural values predict individuals' moral attitudes? A cross-cultural multi-level approach. European Journal of Social Psychology 41, 645–657.

Moral Development, Theories of

Marc Jambon and Judith G Smetana, University of Rochester, Rochester, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Socialization approaches and cognitive-developmental theory have provided two major approaches to studying moral development. Socialization approaches have focused on the development of conscience through guilt and internalized behavior, typically measured in terms of compliance, whereas cognitive-developmental theories have emphasized developmental changes in moral judgments and reasoning. While important differences persist, contemporary approaches have attempted to integrate thoughts, emotions, and behaviors and have stressed the importance of social interactions. The rise of evolutionary based accounts have led to more research with infants and studies incorporating neuroscience, furthering the understanding of the complex nature of moral development.

Introduction

For most of the past century, much of the psychological theorizing and research on moral development has stemmed from two distinct traditions; socialization perspectives and cognitive- or structural-developmental approaches. These theories differ in how they define morality and in their hypotheses regarding its origins and developmental processes. In the first section, the authors describe how early theorists such as Freud, Skinner, and Piaget helped to shape the assumptions, questions, and methods of later developmental research. The authors then discuss empirical findings from the different perspectives and examine how contemporary approaches have developed from these foundations. They emphasize that traditional divisions have become less entrenched as theorists have begun to recognize the dynamic, integrative role of emotions, thoughts, and behaviors in moral development. Finally, the authors briefly discuss recent trends in moral development research from an evolutionary perspective.

Assumptions and Early Theoretical Roots

Socialization Theories

Historically, socialization theories have conceptualized moral development primarily in terms of children's internalization and compliance with adult and societal rules and requests, although they differ in their view of the role of emotions and affect in this process. Approaches emphasizing morality's affective components have their roots in the psychoanalytic tradition of Sigmund Freud. In line with his broader theory of psychosexual development, Freud believed that moral development entailed the internalization of societal standards through the child's identification with parents. This process resulted in the creation of the conscience, or the superego in Freudian terms, which functions as an internal regulator of behavior. The conscience inhibits the innate impulses and urges toward selfishness and aggression that, if left unchecked, would otherwise prevent humans from living in society. From this perspective, aversive emotions such as guilt, shame, and anxiety are thought to serve as the driving motivations for the child's successful inhibition of impulses and internalization of external norms.

Other socialization researchers have been influenced by behaviorist and learning theory principles as exemplified in the work of B.F. Skinner. Like Freud, Skinner did not distinguish between moral norms and other types of social rules or expectations. In contrast to Freud, however, Skinner postulated that moral behaviors are learned through conditioning and reinforcement. Behaviors valued by society are rewarded, while undesirable behaviors are punished. Some social learning theorists, such as Albert Bandura, incorporated cognitive processes, such as observational learning or imitation, to account for how novel behaviors could be acquired independent of reinforcements. Additionally, affective responses like fear and empathy retained a role in other versions of learning theory, although they were interpreted as stemming from conditioned responses resulting from the pairing of feelings of pleasure and pain with external stimuli such as punishment (Eisenberg, 1986).

Cognitive-Developmental Theories

Rooted in the theorizing of Jean Piaget, cognitive-developmental approaches focus on the development of children's moral judgments. These approaches view the child as active in his or her own development and constructing moral knowledge through reciprocal interactions with the environment (termed constructivism). In Piaget's formulation, the process of moral development does not simply entail internalization of external values and standards but rather is a constructive process leading to qualitative changes in children's thinking about right and wrong.

For Piaget, morality is inherently other oriented, revolving around notions of respect, fairness, and justice, and entailing an understanding of rules as serving these moral ideals. Piaget derived his theory from interviews with children about game rules and concepts such as stealing and lying, and his moral stage theory is closely tied to changes in children's social relationships and cognitive abilities. Initially, young children are characterized as premoral and having little concern for rules and obligations. Beginning in the preschool years, however, children enter into what Piaget referred to as a stage of heteronomous morality. Due to constraints in their cognitive abilities and their love and fear of parents, children at this stage hold an absolutist view of morality; the good is defined in

terms of strict adherence to external rules and authority commands, and acts are seen as wrong based on their objective consequences rather than actors' intentions. Social rules are viewed as unchangeable and absolute, leading to the belief that transgressions will be automatically punished (immanent justice). Interactions with peers and increases in cognitive and perspective-taking abilities during late childhood lead to the development of autonomous morality. Through cooperation with others, children come to separate their understanding of rules from notions of reciprocity, justice, and respect. They begin to incorporate intentions into their moral judgments and evaluate rules in terms of whether they serve to uphold moral principles. Although certain aspects of the theory have not received empirical support over the years, Piaget's attempts to define morality as distinct from rules and authority, and his emphasis on social interactions, has had a lasting influence on cognitive-developmental theories of moral development.

Empirical Moral Development Research: Then and Now

Socialization Theories

While psychoanalytic theory provided a conceptual basis for understanding the nature and origins of children's moral development, behaviorism and learning theory provided the rigorous empirical methods that were highly regarded during the first half of the twentieth century. Thus, many socialization researchers integrated ideas from Freudian and behaviorist traditions to understand how children internalized external moral norms and expectations through the creation of a conscience. Many studies examined correlations between parenting practices and measures of conscience, although experimental procedures in a laboratory setting were preferred. Moral behavior was typically operationalized as compliance with commands in 'forbidden toy' or 'resistance to temptation' tasks. Children were required to engage in various tedious activities without stopping, or to refrain from playing with, touching, or looking at attractive toys or objects in the absence of an authority figure. The underlying assumption was that children with an internalized conscience would willingly comply with these behavioral demands (Brody and Shaffer, 1982).

The most frequently employed measure of moral affect was the extent to which children felt guilty or anxious in response to a transgression. This was typically assessed through children's verbal judgments of hypothetical transgressors' feelings of guilt following a rule violation (Brody and Shaffer, 1982). Thus, conscience development was viewed primarily as the extent to which children would be able to inhibit selfish impulses, comply with societal and parental expectations – especially in the absence of authority – and experience guilt when rules were broken.

Much early research attempting to integrate psychoanalytic and leaning theories examined the effects of parenting on the internalization process. Because aversive emotions were seen as the primary motivational basis for moral development, disciplinary practices – particularly punishment – became the primary focus of research. Three types of parental disciplinary practices after a rule transgression were seen as relevant to children's moral development (Brody and Shaffer, 1982): power assertion (use of threats, force, removal of privileges), love withdrawal (nonphysical displays of anger or disapproval, threats of abandonment), and induction (parental explanations and reasons for why the behavior was wrong or undesirable). A consistent finding is that parents who use reasoning and induction have children with more advanced conscience development, whereas love withdrawal and power assertion were either unrelated or negatively associated with moral internalization. This was thought to be because these more punitive discipline techniques led children to develop a concern with the anticipation of rewards and punishment, which establishes an external rather than self-regulated moral orientation. In contrast, induction facilitates moral growth by focusing the child on the consequences of actions for others. In turn, this leads to a healthy sense of guilt and compliance to demands, even in the absence of an authority figure or threat of punishment.

Many contemporary socialization researchers have continued to define moral development as a process of internalization and self-regulation, yet they have shifted their attention to underscore the bidirectional nature of the parent–child relationship and the active role of the child. For instance, Joan Grusec and her colleagues (Grusec et al., 2000) have pointed out that parents' influence on children's acquisition of moral values depends on a number of additional factors, including children's characteristics and reactions to parental socialization attempts, the broader history of the relationship, and the specific socialization goals and practices parents may have in a given situation. Furthermore, while conscience continues to be defined and measured by behavioral compliance and aversive emotional reactions, Grazyna Kochanska and Ross Thompson have also incorporated positive emotions (e.g., empathy), moral cognitions (e.g., an awareness of the rules and the consequences of transgressions), and concepts such as self and identity into their conceptualizations.

Toddlers who are securely attached and in relationships with parents characterized by high levels of responsiveness, warmth, and shared positive affect show greater internalization and more advanced conscience development (Kochanska and Aksan, 2006). These early parent–child relationships are thought to provide the motivation for children to willingly and eagerly cooperate with the caregiver's socialization attempts, an orientation referred to as committed compliance. In contrast, children of insensitive parents, who share few positive and mutually enjoyable activities with their child, are more likely to induce situational compliance, or compliance that is motivated and sustained by parental pressure and fear of punishment. Likewise, mothers who elaborate more when talking with their children and who emphasize others emotions, thoughts, and perspectives rather than rules when discussing past behaviors have young children with more advanced conscience development and moral reasoning (Dunn, 2006; Laible and Thompson, 2000). Furthermore, these early relational experiences and children's committed compliance during toddlerhood are associated with more mature conscience development later in life.

Researchers also have focused on the role of temperament in moral development. Negative emotional arousal (i.e., anxiety and guilt) and the ability to regulate and inhibit one's

behavior are thought to be important for moral internalization and conscience development. Two different temperamental factors – fearfulness and effortful control – have been identified as critical, albeit in different ways (Kochanska and Aksan, 2006). Children who are highly fearful and shy are naturally more prone to negative emotional arousal and are more inhibited in their actions. Fearful children exhibit more guilt after wrongdoing and more compliant behavior, especially when parents use gentle, nonassertive parenting strategies that deemphasize the use of power. For these children, parenting practices that capitalize on negative emotions without causing high levels of distress are most effective for moral internalization. Fearless children, in contrast, do not exhibit feelings of discomfort after transgressions; thus, the emotional arousal presumed to motivate moral behavior is lacking. For these children, mutually responsive relationships and secure attachments, rather than parental discipline, is important for conscience development. Thus, the effectiveness of different socialization techniques depends on the characteristics of the child.

The temperamental trait of effortful control refers to children's ability to deliberately control behavioral impulses and is therefore thought to be important in the development of self-regulation. Children who are higher in inhibitory or effortful control are better able to resist temptations. As a result, they exhibit greater levels of internalized and compliant behavior throughout childhood (Kochanska and Aksan, 2006). Deficits in attentional and behavioral control also have been linked to problem behaviors, including aggression, substance use, and lower levels of moral reasoning in adolescence and adulthood (Eisenberg, 2000).

Two aspects of early conscience development, guilt proneness and regulatory control, are also important for adaptive moral functioning in later childhood. For instance, Kochanksa and her colleagues (Kochanska et al., 2009) found that 2- and 3-year-olds lacking guilt reactions to transgressions had higher levels of disruptive conduct (i.e., lying, fighting, bullying others, and disobedience) at 6 years of age, but only if they were also low in effortful control. This suggests that there may be different paths to conscience development. Children who lack the affective requirements for moral internalization may rely on their ability to consciously regulate behavior. In extreme cases, temperamental deficits in fearfulness and guilt-proneness coupled with emotional and behavioral dysregulation may contribute to the development of psychopathy, which is characterized by chronic and extreme antisocial behavior (Eisenberg, 2000).

One explanation offered for the link between conscience development and later moral functioning is the concept of a moral self. As children learn to self-regulate and comply with parental standards (i.e., committed compliance), they may come to view themselves as good individuals who adopt their parents' values and expectations (Kochanska et al., 2010). In early childhood, the moral self has been operationalized as the extent to which children see themselves as feeling bad and wanting to make amends after wrongdoing, complying with rules, and holding prosocial attitudes toward others. This internalized sense of self is thought to serve as the motivational basis to regulate and guide future behavior. Indeed, Kochanska et al. (2010) found that the moral self at age 5 years mediated the link between earlier measures of conscience and later moral conduct. Some have hypothesized that early conceptions of the moral self may provide the foundation for later moral identity development in adolescence and young adulthood (e.g., Hardy and Carlo, 2011; Kochanska et al., 2010), although these connections have yet to be tested empirically.

While negative emotions such as guilt and anxiety have been accorded a central role in socialization research, others, such as Nancy Eisenberg and Martin Hoffman, have also examined the positive affective and behavioral components of morality, including empathy and prosocial behavior. They assert that tendencies toward care and cooperation are rooted in our biological makeup. In contemporary developmental research, empathy is most commonly defined as an emotional response stemming from the perception of another's condition or affective state, reflecting what the other person is feeling (Eisenberg et al., 2006; Hoffman, 2000). Empathy is typically viewed as a relatively automatic, value-neutral emotional reaction that provides the foundation for feelings of care or concern for others, a reaction referred to as sympathy. Sympathy differs from empathy in that it does not involve feeling what the other person feels, but instead involves concern stemming from the apprehension of another's negative or distressing experience or state. Thus, while empathy may apply to both negative and positive shared experiences (e.g., feeling happy when others are happy), sympathy specifically pertains to feelings of concern, sorrow, or sadness for others. It is considered more developmentally mature than empathy.

When empathic feelings are highly intense or poorly regulated, however, they may lead to personal distress (Eisenberg et al., 2006) or empathic overarousal (Hoffman, 2000). This highly arousing, aversive negative response focuses on alleviating one's own distress rather than another's. It is therefore not seen as contributing to children's moral development.

Empathic connections to others' emotions and feelings are evident at birth. Newborn infants less than 24 h old have been shown to become upset and cry when hearing other infants cry, but not when hearing a tape recording of their own cries (Hoffman, 2000). The reflexive, global empathic reactions of infants become more sophisticated and other oriented during childhood, aided by normative advances in perspective taking and the ability to regulate emotional reactions. This has been assessed using simulated distress procedures in which adults (usually mothers or experimenters) pretend to be hurt, and children's reactions are coded. These procedures have revealed that facial (e.g., sadness), verbal (e.g., 'are you OK?'), and behavioral expressions (e.g., trying to alleviate distress) of empathic concern and sympathy increase during the second and third years of life, while personal distress reactions such as crying decline with age (Eisenberg et al., 2006).

Individual differences in temperament and parenting practices contribute to variations in empathy-related responding (Eisenberg, 2000; Eisenberg et al., 2006). Children and adults who are less prone to experience intense negative emotions and who are more capable of regulating their emotions show higher levels of sympathy and lower levels of personal distress. Parental influences on children's empathy-related responding are complex, and findings differ depending on whether physiological measures, facial expressions, or self reports are measured. In general, warm and supportive parenting,

including responsiveness, inductive discipline, and practices that help children understand and cope with their emotions are associated with greater sympathy and less personal distress.

These dispositions toward feelings of care and concern for others are important because they provide the motivation for engaging in prosocial behavior (defined as voluntary actions aimed at helping or benefitting others) and refraining from antisocial and aggressive behavior. As this suggests, children and adults who are more prone to sympathy engage in more prosocial behavior, such as donating to charity or assisting others. Those more prone to personal distress engage in less prosocial behavior, except when doing so is likely to relieve their own negative reactions (Eisenberg et al., 2006). In keeping with its definition as a neutral emotional reaction identical to another's situation, empathy – when distinguished from sympathy and personal distress – appears to show little systematic relationship to prosocial behavior.

The connections between empathy-related responding and aggression are more ambiguous (Eisenberg et al., 2006). Higher levels of empathy and sympathy are consistently associated with less aggressive and problem behaviors for older children, adolescents, and adults. However, in early childhood, greater empathic responding (including measures of both empathy and sympathy) has sometimes been linked to higher levels of both prosocial and aggressive behavior, perhaps because aggressive children are more assertive and more likely to engage in both harmful and helpful activities when interacting with others. Indeed, aggressive preschoolers who use both coercive and prosocial strategies for obtaining resources are also better able to appear more moral in their verbal responses (Hawley and Geldhof, 2012).

While much of the contemporary socialization research has focused on the development of moral affect and behavior within the context of parent–child interactions, cultural psychologists have focused on how different cultural systems provide a context for the development of moral values and codes of conduct (Miller and Bland, 2014; Shweder et al., 1987). Much early work from this perspective sought to categorize different cultures on the basis of their general orientations toward certain types of beliefs or practices. For instance, one broadly used distinction is between cultures that stress individual rights and personal freedoms (referred to as individualism and thought to pertain to North American and European societies as well as Australian and New Zealand) and those emphasizing duty, obligation, and interdependence (referred to as collectivism and encompassing the rest; Triandis, 1995).

In recent years, however, cultural theorists have begun to stress the heterogeneity that exists within cultures as well as between them, and have begun to explore the ways in which universal moral concepts such as justice and care can take on different meanings in distinct cultural settings. For instance, a series of studies by Joan Miller and her colleagues (see Miller and Bland, 2014) has shown that children and adults in both the United States and India view responsiveness to the needs of family and friends as important and desirable. However, Hindu Indian participants are often more likely than Americans to treat interpersonal and familial commitments as morally obligatory in different circumstances, which may stem from the way in which Indian culture emphasizes interpersonal duties and responsibility (*dharma*). Other researchers have also incorporated more qualitative and ethnographic accounts of cultural practices to explore how different communities socialize moral values (Fung and Smith, 2002).

Despite these theoretical advances, cultural psychologist Joan Miller has recently noted that this approach continues to assume a "fax model of cultural transmission" (Miller and Bland, 2014: p. 309), whereby children are presumed to automatically internalize the given values of their particular community fully formed. This is likely because the ideas and methods of cultural psychology emerged from disciplines outside of developmental psychology (i.e., anthropology, sociology, and social psychology), leaving cultural theorists with relatively little to say about the specific mechanisms by which cultural meanings and values are internalized within the individual. As a result, much of the cultural research to date has focused on adolescents and adults, and has had relatively little to say about developmental processes (but see Fung and Smith, 2002; Rogoff et al., 2007).

Cognitive-Developmental Theories

Consistent with Piaget, cognitive- or structural-developmental theories view moral reasoning as central to understanding moral action. Thus, behaviors such as aggression, honesty, and rule following (moral or otherwise) cannot be understood without knowing the thought process underlying them. As a result, researchers have adapted Piaget's clinical interview method to assess children and adolescents' moral reasoning. Judgments are elicited and probed in response to hypothetical stories or real-life situations to assess children's understanding.

Building on Piaget and beginning in the late 1950s, Lawrence Kohlberg (1971) articulated what became a very well-known global stage theory of moral judgment development. He proposed that individuals' conceptions of justice and rights develop through a series of six qualitatively distinct stages (grouped into three broad levels), with each successive stage representing a more developmentally and logically mature form of moral reasoning. During the preconventional level, children's moral judgments are based on concerns with obedience, punishment, and self-interest. The *conventional* level of reasoning develops during the adolescent years, and at these two stages, what is morally right is determined by interpersonal duties and societies' rules and laws. Although no longer tied to punishment, conventional moral reasoning still constitutes an external orientation because it is based on the existence of external rules and expectations, sharing similarities with Piaget's heteronomous stage. Kohlberg proposed that not until late adolescence (if at all) do individuals fully distinguish between social rules and the principles of justice, rights, and mutual respect, which he referred to as *postconventional* reasoning.

These stages were described as universally applicable to all people, and the ordering was believed to be invariant (i.e., occurring in the same sequence). Additionally, stages were described as 'structured wholes,' meaning that an individual's particular way of thinking about morality (e.g., as a conventional matter of rules and authority) would be the same regardless of the content or nature of the situation.

Researchers from the Kohlbergian tradition presented individuals with hypothetical dilemmas involving conflicts between different moral concepts, such as life, authority, duties, and obligations. The goal was to capture the organization or structure of thought underlying children's and adolescents' moral judgments and how these structures change with age and social experience. Similar to Piaget, Kohlberg believed that advances in moral thinking were tied to the development of children's logical thinking and perspective taking abilities; he described them as "necessary but not sufficient" for moral growth. Global stage theories were also used to examine other aspects of children's moral thinking, including distributive justice and prosocial reasoning.

Kohlberg's influence on the study of moral development cannot be underestimated; many of his contributions, including the need to ground psychological conceptions of morality in philosophical definitions, remain central to many contemporary approaches to moral judgment development. However, problems identified with the specifics of Kohlberg's theory led some of his contemporaries to question the validity of his assumptions. For instance, a number of critics argued that Kohlberg's prioritization of abstract notions of justice as indicating the most advanced levels of moral understanding were biased against women and individuals from non-Western cultures. While subsequent research has suggested that claims of bias may have been overstated (Walker, 2006), the notion that Kohlberg's theory does not adequately capture the complexity of children's (and adults') moral understanding has received substantial empirical support.

The social domain perspective, as articulated by Elliot Turiel, Larry Nucci, and Judith Smetana, has provided the most well-known and comprehensive cognitive-developmental alternative to global stage theory accounts of children's moral judgment development. These researchers have critiqued Kohlberg's use of complex hypothetical vignettes to study morality because they underestimate children's ability to make distinctly moral judgments. The results of numerous studies with culturally diverse samples of males and females from around the world have shown that children and adolescents distinguish issues of justice, rights, and welfare (moral issues) from more arbitrarily defined and contextually based social conventions (e.g., how to address an authority figure and what clothes to wear to school) at much earlier ages than Kohlberg's theory suggests.

In this approach, young children's judgments are typically studied in response to hypothetical stories that depict straightforward (prototypical) situations that correspond to different *domains* of social knowledge. For instance, moral domain events, such as one child intentionally hitting another, are presented in situations where they do not conflict with other types of goals, motivations, or events. Children make judgments about the act's *generalizability* across contexts (e.g., if it is OK if performed at home rather than at school or in another country) and *rule* and *authority independence* (e.g., would it be OK if there were no rules or authority forbidding it). Even preschoolers judge conventional rules and customs to be contextually relative, alterable, and dependent on authority and consensus. In contrast, moral obligations involving issues of justice, rights, and welfare are judged to be applicable across contexts and independent of the existence of authority or rules for their prescriptive force. By the early school years, children also justify their judgments about moral events with concerns for others' harm or welfare, fairness or rights, and obligations, whereas they justify conventional events with appeals to authority, tradition, social order, and cultural norms.

Social domain researchers have also examined children's understanding of self, identity, the causes and motivations underlying behavior, as well as judgments of personal issues. The latter revolve around individuals' control over one's body, privacy, and preferences and choices over issues such as leisure activities and friends. Personal issues are seen as outside the scope of moral and conventional regulation and are judged to be up to the individual to decide because they have no harmful consequences for others. Although the specific nature and content of the personal domain varies at different ages and across cultures, needs for autonomy and personal choice are seen as universal features of human development. As such, they are thought to serve as a foundation for children's developing understanding of rights and justice.

Children's experiences as victims, observers, and perpetrators of moral transgressions generate distinctly moral judgments. Observational studies in a variety of contexts, including home, day care, school, and free-play settings, show that parents and teachers focus on the repercussions of moral transgressions for victims. Adults emphasize the harm or unfairness of moral events and encourage transgressors to take the victim's perspective. Thus, caretakers amplify and direct children's attention toward the consequences of actions for others. Although adults are concerned with enforcing moral rules, many moral conflicts occur primarily between siblings and peers and are often resolved without adult intervention, especially as children grow older. Victims also respond with affectively charged reactions like crying, protests, and retaliation that serve to highlight the harmful nature of the transgressions. Beyond situations of conflict, Judy Dunn (2006) has shown that emotion understanding and positive sibling relationships and friendships are linked to greater moral sensitivity and care for others during childhood.

In contrast to moral events, authorities (and to a lesser extent peers and siblings) enforce conventional expectations, especially with young children. Authority figures are concerned with maintaining rules and social order and socializing children into the norms and customs of their social group. Thus, they respond to conventional violations with references to rules, authority commands, and the need for social order and decorum. By middle childhood, children also respond to others' conventional transgressions, although even preschool children enforce peer-initiated conventions. Moral violations have intrinsic consequences for others and therefore evoke powerful emotional responses, whereas violations of home and school conventions are typically judged to be affectively neutral.

Observations of preschoolers at home and at school have shown that in contrast to responses to moral and conventional events, interactions regarding personal issues (e.g., clothing and food preferences) often involve more tacit forms of communication. For instance, rather than giving direct commands, caregivers may offer choices, provide different options, or solicit opinions. Thus, parents and children

negotiate the boundaries of children's jurisdiction versus what authorities can legitimately regulate.

Although preschoolers distinguish straightforward moral events from other types of social norms, their moral understanding continues to develop throughout childhood and adolescence. For instance, compared to preschoolers, older children more consistently apply moral criteria to abstract moral events, such as those involving psychological harm (e.g., name calling). From early to middle childhood, moral concerns broaden from a concrete focus on harm and others' welfare to include an understanding of fairness as equal treatment. Adolescents become able to take context into account, leading to an appreciation for *equity*, or the idea that fair treatment entails a consideration of individuals' needs and characteristics (Damon, 1977; Nucci and Gingo, 2011), and they also become more attuned to issues of discrimination and rights.

The social domain perspective stresses that children's moral development also must consider individuals' thinking about both moral and nonmoral concerns as they overlap and conflict with each other (referred to as *multifaceted events*). This is important because, consistent with the notion that individuals actively strive to interpret and understand their social world, many everyday situations cannot be cleanly categorized as solely matters of morality or convention. Thus, variations in the salience of different concepts and how they become coordinated in thinking help to explain individual differences in moral evaluations.

Domain coordination has been examined in children's and adolescents' reasoning about various social issues, such as inclusion and exclusion from social groups; the development of prejudice based on race, gender, and ethnicity; children's and adolescents' understanding of rights, civil liberties, and government; the influence of cultural hierarchy and social roles on judgments of duty and obligation; parent–adolescent relationships and autonomy development; and lying and honesty in close relationships. The findings from a large and diverse body of research do not support a view of development as entailing progress toward a shared moral viewpoint, that morality overrides other types of concerns as children grow older, or that specific moral values come to be endorsed uniformly and consistently across different contexts (Nucci and Gingo, 2011). Rather, the picture that has emerged is that with age children come to hold a more complex, nuanced, and flexible understanding of morally relevant situations. They become better able to recognize and integrate conflicting moral, conventional, and personal concerns.

Charles Helwig's research on individuals' understanding of rights, laws, and civil liberties illustrates this point nicely (see Helwig, 2006 and, for examples in non-Western cultures, Wainryb, 2006). In one study, childrens', adolescents' and college students' evaluations of rights to free speech and religion were examined in different social contexts, including the family, school, and society. As early as 6 years of age, children conceived of rights as moral issues that were independent of authority and that should be upheld across cultures. However, developmental trends were found in the extent to which individuals identified and coordinated different factors in their judgments. Older adolescents and college students became more likely than younger children to believe that rights may be restricted in certain contexts. With age, children and adolescents began to consider an agent's mental competence and the possibility of negative consequences for others when deciding whether rights should be restricted. In another study, Helwig further found that older adolescents were more likely than younger participants to uphold rights to free speech and religion when they conflicted with established laws. With age, adolescents were better able to coordinate concepts of rules, social systems, and civil liberties. Thus, individuals at all ages endorse rights in some contexts but restrict them in others; there is no clear developmental trajectory toward upholding rights in all situations. Instead, there is an increased tendency with age to attend to and coordinate different factors and circumstances into moral judgments.

Consistent with other cultural psychological approaches (Miller and Bland, 2014), social domain researchers have also shown how different factual beliefs individuals hold about the world also may lead to variations in moral judgments (Turiel et al., 1991). Informational assumptions refer to individuals' descriptive understandings of reality, rather than their specific moral values or concepts, and can stem from many sources, including parents, science, and religion. They are often taken for granted (although accepted knowledge can change, such as with scientific progress), and they can vary substantially across individuals, religions, and ethnic groups or cultures. Informational assumptions are believed to lie at the heart of many moral and social conflicts. For example, individual differences in individuals' assumptions about when life begins and the definition of 'personhood' have implications for whether abortion is seen as a moral or personal issue.

Age-related changes in preschoolers' understanding of others' minds (referred to as *theory of mind*) have been examined as an important source of developmental change in children's factual assumptions about the world (Wainryb and Brehl, 2006). As an appreciation for others' mental states develops, children are more able to consider the reasons, intentions, and motivations behind moral events or transgressions. This allows them to incorporate and integrate information about contexts and situations when forming moral evaluations. Mental state understanding has also been linked to children's judgments of diverse cultures and tolerant attitudes toward other groups' practices. Thus, with age, children are increasingly able to take into account both moral and factual beliefs in evaluating their social experiences.

Recent Trends in Moral Development Research: Evolutionary Approaches

There has been an increased interest in integrating other disciplines, from neuroscience to anthropology, into the understanding of moral development. As an example, recent research has drawn on evolutionary theories to examine the foundational roots of morality in the first few years of life (Hamlin, 2013; Warneken and Tomasello, 2009). Specifically, researchers have focused on implicit social preferences and behavioral tendencies toward altruism and cooperation that are hypothesized to be innate. These hardwired tendencies or intuitions have been selected through evolution because of their adaptive value in promoting and maintaining harmony in

social groups, affording humans the opportunity to survive and flourish within large groups and societies.

Recent methodological advances have allowed these theorists to examine early moral dispositions toward care and cooperation in infancy and early toddlerhood. For instance, recent studies have used measures of looking time and play preferences to show that infants and toddlers, some as young as 5–10 months of age, already show an affinity for prosocial actors as well as a dislike of antisocial actors, even when the victims of such actions do not display any signs of distress or negative emotion. By 14 months of age, children also show proclivities toward altruistic behaviors, spontaneously giving assistance and help to others when needed, even in the absence of any tangible external rewards. Indeed, giving children tangible rewards for helping may actually lessen the likelihood that they will engage in such behaviors in the future by undermining their intrinsic motivation to help. However, this research is primarily cross-sectional and involves novel experiments in laboratory settings. Therefore, further studies are needed to understand how these early dispositions relate to moral understanding at later ages.

Evolutionary theory has also played an important role in the emerging field of moral neuroscience (for extensive discussions see Young and Dungan, 2012; Zelazo et al., 2010). Recent advances in brain scanning techniques, such as functional magnetic resonance imaging (fMRI), have allowed researchers to explore the different brain structures and patterns of neural activation associated with different moral capabilities. In fMRI studies, individuals are typically presented with experimental stimuli (e.g., depictions of intentional vs. accidental harm) and asked to give forced choice judgments (e.g., OK or not OK). Because fMRI records changes in blood flow between different brain areas over time, a large number of experimental trials are needed to ensure that participants are thinking about the stimuli. Given the methodological complexity of this type of research, reaction times are often used to argue for the relative deliberateness or automaticity of a particular judgment, and reasoning and justifications are rarely assessed. Neuroscientists have also utilized clinical populations, such as individuals with brain lesions, to examine how impairments in certain neural structures may influence moral functioning. Because of the invasiveness of neuroimaging procedures and the limited number of individuals with brain abnormalities, most moral neuroscience studies have been restricted to small samples conducted almost exclusively with adult populations. Nevertheless, this emerging body of research provides valuable insight into the underlying neural structure of morality.

There is a growing consensus that moral functioning is not isolated to one specific brain structure, but instead entails a variety of different areas associated with both cognitions and emotions. As an example, fMRI research has identified various areas within and related to the prefrontal cortex (PFC) as playing a crucial role in moral judgments and behavior. Paul Eslinger and Melissa Robinson-Long note that the PFC undergoes development throughout the lifespan and is a core element of a brain system that "mediates diverse information processing, consolidates complex types of knowledge and experience, integrates cognitive, visceral-somatic and affective streams of processing, and significantly influences behavioral self-regulation through goal direction, inhibition, decision-making, and self-monitoring" (Zelazo et al., 2010: p. 252). Consequently, there is substantial evidence that individuals with damage to areas of the PFC often show marked impairments in their moral and social decision making and increased rates of disruptive and antisocial behavior. PFC injury has also been tied to deficits in emotional arousal and effortful control, which socialization researchers have shown to be important for healthy conscience development. Psychopaths, who show low levels of guilt and empathy, high levels of antisocial behavior, and a reduced ability to distinguish between moral and conventional violations, also show abnormalities in PFC structure and activation.

Recent research has also begun to explore the neural underpinnings of empathy and prosocial behavior. A number of studies have shown that the same areas of the brain responsible for the experience of pain are also activated when individuals observe or imagine others being harmed. Studying children and adolescents (as well as adults), Jean Decety and Kalina Michalska (2010) also found age-related changes in neural activation while watching others being harmed. This suggested that empathic sensitivity to others' pain may shift with age from a basic, visceral emotional response to a more controlled, integrated process that takes into account different contextual cues and deliberative reasoning. These findings are fully in line with the developmental account of empathy offered by Martin Hoffman (2000). Similarly, prosocial behaviors, such as giving to charity, have been shown to activate regions of the brain associated with reward processing, which is consistent with the notion that evolved tendencies toward altruism and cooperation are intrinsically motivating (Warneken and Tomasello, 2009).

The importance of theory of mind understanding for making moral judgments has also been highlighted by Liane Young and Rebecca Saxe. In particular, a brain region known as the right temporoparietal junction (RTPJ) has been consistently implicated in the identification and representation of mental states such as beliefs and intentions. In one study, the researchers found that individuals displaying stronger activation in the RTPJ gave greater weight to an agent's intentions when evaluating hypothetical stories involving harm, leading to more accepting judgments of accidental transgressions. In a more direct test of this link, the researchers used a procedure to temporarily disrupt activation in the RTPJ and found that disrupting activation in this region led individuals to focus more on an act's outcomes (as opposed to intentions). This led them to be more forgiving of situations where agents intentionally tried but failed to cause harm to others. Similar activation of brain regions associated with mental state understanding was found by Decety and Michalska (2010) when participants viewed others intentionally – but not accidentally – harmed.

While moral neuroscience research has the potential to greatly expand the understanding of the neural processes implicated in the sense of right and wrong, the current methodological limitations preclude any strong conclusions regarding the nature of children's moral development. Additionally, much moral neuroscience research does not offer clear definitions of morality and often relies on broad or vague descriptions of rule following behavior or situations entailing life and death (see chapters in Zelazo et al., 2010).

Summary and Conclusions

While contemporary approaches draw on the early work of figures such as Freud, Skinner, Piaget, Kohlberg, and Darwin for inspiration in defining the nature and scope of morality, conceptualizations have been broadened and refined to illuminate the complex, multifaceted nature of children's moral development. Socialization approaches continue to focus on processes of internalization and behavioral regulation. However, they have expanded to incorporate both positive and negative affective experiences, highlighting the importance of interactions beyond the discipline context and embracing a more active view of the child. Similarly, structural-developmental theories continue to view judgment and reasoning as the cornerstone of children's moral development while attending to the complex interplay among emotions, cognitions, and behaviors in individuals' evaluations of right and wrong. Additionally, the emergence of evolutionarily based accounts of individuals' predispositions toward care and cooperation, and their underlying neural structures have provided valuable insights into the phylogenetic and ontogenetic roots of morality.

Considered together, research on moral development has shown that from the first years of life, children are concerned with the well-being of others and are motivated toward helping and cooperating with those in their social groups. At the same time, moral development is aided and constructed through children's experiences with family members and peers. The characteristics of the child, their interpersonal relationships with others, and variations in societal and cultural assumptions about the nature of reality all further influence the process of moral development.

See also: Eisenberg's Theory of Prosocial Reasoning; Empathy During Early Childhood Across Cultures, Development of; Moral Development, Cultural Differences In; Piaget's Theory of Cognitive Development; Prosocial Behavior During Childhood and Cultural Variations; Social and Emotional Development in the Context of the Family; Values Across Cultures, Development of.

Bibliography

Brody, G., Shaffer, D., 1982. Contributions of parents and peers to children's moral socialization. Developmental Review 2, 31–75.

Damon, W., 1977. The Social World of the Child. Jossey-Bass, San Francisco.

Decety, J., Michalska, K., 2010. Neurodevelopmental changes in the circuits underlying empathy and sympathy from childhood to adulthood. Developmental Science 13, 886–899.

Dunn, J., 2006. Moral development in early childhood and social interaction in the family. In: Killen, M., Smetana, J. (Eds.), Handbook of Moral Development. Erlbaum, Mahwah, NJ, pp. 331–350.

Eisenberg, N., 1986. Altruistic Emotion, Cognition, and Behavior. Erlbaum, Hillsdale, NJ.

Eisenberg, N., 2000. Emotion, regulation, and moral development. Annual Review of Psychology 51, 665–697.

Eisenberg, N., Spinrad, T., Sadowsky, A., 2006. Empathy-related responding in children. In: Killen, M., Smetana, J. (Eds.), Handbook of Moral Development. Erlbaum, Mahwah, NJ, pp. 517–549.

Fung, H., Smith, B., 2002. Learning morality. In: Lancy, D.F., Bock, J., Gaskins, S. (Eds.), The Anthropology of Learning in Childhood. Rowan & Littlefield, Lanham, MD, pp. 261–285.

Grusec, J., Goodnow, J.J., Kuczynski, L., 2000. New directions in analyses of parenting contributions to children's acquisition of values. Child Development 71, 205–211.

Hamlin, J.K., 2013. Moral judgment and action in preverbal infants and toddlers: evidence for an innate moral core. Current Directions in Psychologial Science 22, 186–193.

Hardy, S., Carlo, G., 2011. Moral identity: what is it, how does it develop, and is it linked to moral action? Child Development Perspectives 5, 212–218.

Hawley, P.H., Geldhof, G.J., 2012. Preschoolers' social dominance, moral cognition, and moral behavior: an evolutionary perspective. Journal of Experimental Child Psychology 112, 18–35.

Helwig, C., 2006. Rights, civil liberties, and democracy across cultures. In: Killen, M., Smetana, J. (Eds.), Handbook of Moral Development. Erlbaum, Mahwah, NJ, pp. 185–210.

Hoffman, M.L., 2000. Empathy and Moral Development: Implications for Caring and Justice. Cambridge University Press, Cambridge, UK; New York.

Kochanska, G., Aksan, N., 2006. Children's conscience and self-regulation. Journal of Personality 74, 1587–1617.

Kochanska, G., Barry, R.A., Jimenez, N.B., Hollatz, A.L., Woodard, J., 2009. Guilt and effortful control: two mechanisms that prevent disruptive developmental trajectories. Journal of Personality and Social Psychology 97, 322–333.

Kochanska, G., Koenig, J.L., Barry, R.A.M., Kim, S., Yoon, J.E., 2010. Children's conscience during toddler and preschool years, moral self, and a competent, adaptive developmental trajectory. Developmental Psychology 46, 1320–1332.

Kohlberg, L., 1971. From is to ought: how to commit the naturalistic fallacy and get away with it in the study of moral development. In: Mischel, T. (Ed.), Psychology and Genetic Epistemology. Academic Press, New York, pp. 151–235.

Laible, D.J., Thompson, R.A., 2000. Mother–child discourse, attachment security, shared positive affect, and early conscience development. Child Development 71, 1424–1440.

Miller, J.G., Bland, C.G., 2014. A cultural psychology perspective on moral development. In: Killen, M., Smetana, J. (Eds.), Handbook of Moral Development, second ed. Erlbaum, Mahwah, NJ, pp. 299–314.

Nucci, L., 2001. Education in the Moral Domain. Cambridge University Press, Cambridge, UK.

Nucci, L., Gingo, M., 2011. Moral reasoning. In: Goswami, U. (Ed.), Wiley-Blackwell Handbook of Child Cognitive Development, second ed. Wiley-Blackwell, Chichester; Malden, MA, pp. 420–445.

Rogoff, B., Moore, L., Najafi, B., Dexter, A., Correa-Chávez, M., Solís, J., 2007. Children's development of cultural repertoires through participation in everyday routines and practices. In: Grusec, J.E., Hastings, P.D. (Eds.), Handbook of Socialization. Guilford, New York, pp. 490–515.

Shweder, R., Mahapatra, M., Miller, J., 1987. Culture and moral development. In: Kagan, J., Lamb, S. (Eds.), The Emergence of Morality in Young Children. University of Chicago Press, Chicago, pp. 1–83.

Smetana, J., 2006. Social-cognitive domain theory: consistencies and variations in children's moral and social judgments. In: Killen, M., Smetana, J. (Eds.), Handbook of Moral Development. Erlbaum, Mahwah, NJ, pp. 119–153.

Triandis, H.C., 1995. Individualism & Collectivism. Westview Press, Boulder, CO.

Turiel, E., 1983. The Development of Social Knowledge: Morality and Convention. Cambridge University Press, Cambridge, UK.

Turiel, E., Hildebrandt, C., Wainryb, C., 1991. Judging social issues: difficulties, inconsistencies, and consistencies. Monographs of the Society for Research on Child Development 56 (2), 103. Serial No. 224.

Wainryb, C., 2006. Moral development in culture: diversity, tolerance, and justice. In: Killen, M., Smetana, J. (Eds.), Handbook of Moral Development. Erlbaum, Mahwah, NJ, pp. 221–242.

Wainryb, C., Brehl, B., 2006. I thought she knew that would hurt my feelings: developing psychological knowledge and moral thinking. In: Kail, R. (Ed.), Advances in Child Development and Behavior. Elsevier, New York, pp. 131–171.

Walker, L.J., 2006. Gender and morality. In: Killen, M., Smetana, J. (Eds.), Handbook of Moral Development. Erlbaum, Mahwah, NJ, pp. 93–115.

Warneken, F., Tomasello, M., 2009. The roots of human altruism. British Journal of Psychology 100, 455–471.

Young, L., Dungan, J., 2012. Where in the brain is morality? Everywhere and maybe nowhere. Social Neuroscience 7, 1–10.

Zelazo, P.D., Chandler, M., Crone, E. (Eds.), 2010. Developmental Social-Cognitive Neuroscience. Taylor & Francis, New York.

Moral Entrepreneur

Marianne E de Campo, Faculty of Public Administration, Berlin School of Economics and Law, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Within the interactionist theory of deviance – or labeling theory – moral entrepreneurs function as rule creators and rule enforcers. Deviance, according to labeling theory, is a result of applying rules to an act, thus moral enterprise constitutes an important part in defining deviance.

The concept of the moral entrepreneur inspired the study of social problems and helped uncover the often professional interests of moral entrepreneurs fighting in an apparently purely ethical battle.

The concept of the moral entrepreneur is an integral part of labeling theory – or rather the interactionist theory of deviance (Becker, 1973) – which sees deviance as the result of the application of rules. In his *Outsiders* first published in 1973 Howard S. Becker argued that deviance is not a quality of the very act itself but is constructed by the reactions of others to the act. By creating and enforcing a rule, others label the agent as deviant and make her/him an 'outsider.' Since values and rules in modern society are far from unambiguous, the act of deciding whether and which rule ought to be applied to a certain practice requires initiative, i.e., moral entrepreneurship. Single persons or social groups act as moral entrepreneurs by establishing their professional and sometimes personal interests as public opinion. They succeed in doing so with the support of additional experts who confirm the gravity of a 'social problem' and use the media and politics to set the agenda, push their views, and enforce the newly created rules.

The Concept

Given the notion that rules have to be created by human beings the question how and when rules are created and enforced has to be answered. Since no social group enforces all its rules without exceptions, the circumstances leading to enforcement are relevant. As Becker explains, an entrepreneur has to take the initiative to punish a culprit, and the group in favor of enforcement of the rule has to gain the attention of others or 'blow the whistle' (Becker, 1973: p. 122). In pluralist societies, competing groups will call for public attention for infringed rules, thus the relative power of a group and its access to the media decides whether it succeeds in alerting the public. Government bureaucracies (e.g., enforcement agencies such as the police) on the one hand and various kinds of experts (e.g., lawyers, psychiatrists, social workers, etc.) on the other hand lend their authority to moral entrepreneurs; however, they themselves have interests that are at times compatible with the original moral entrepreneur who strives to arouse the public. For example, Edwin Sutherland (Sutherland, 1950) shows that psychiatrists had an economic interest in the creation of sexual psychopath laws, which required a psychiatric diagnosis and treatment of the offender. Similarly, each new edition of the Diagnostic and Statistical Manual of the American Psychiatric Association diagnoses more and more mental disorders and thus contributes to 'pathologizing everyday behavior' (Kutchins and Kirk, 1997: p. 21).

The moral entrepreneur appears in the form of the rule creator or moral crusader and in the form of the rule enforcer. The first is interested in the ethical content of the rules, follows a mission by imposing her/his morals onto others in order to better them or 'save' their souls. The moral crusaders are typically middle- or upper middle-class members of society. They are themselves not affected by the vices of those low-life milieus they want to eradicate. Gusfield (Gusfield, 1955, 1967, 1986[1963]) portrayed the members of the temperance movement as paradigm moral crusaders. Similarly, members of the Anti-Pornography Organizations in the USA of the 1960s struggled to defend their values against the permissiveness of liberal culture, which they regarded as threatening by symbolizing a social change that would undermine their social status (Zurcher et al., 1973).

The moral entrepreneur as rule enforcer is far more pragmatic and less purist than the rule creator; after all, the rule enforcer makes his living in prosecuting the deviators from the rule. So, e.g., the police will pride themselves with the successes of prosecution and at the same time report that despite their efforts crime persists or even increases and requires further funding for rule enforcement. As a consequence of the working conditions of the rule enforcers (time and staff constraints, lack of proper equipment, etc.), they will selectively decide when and who to persecute. Becker mentions that an important selection criterion is the suspect's disrespect for the rule enforcer (Becker, 1973: p. 158; Westley, 1953).

According to Becker, enterprise is crucial for deviance, because no initiative of the moral entrepreneur – no rules from which to deviate. And further, without the rule enforcing moral entrepreneur's initiative, infringement of rules would not be detected or persecuted. Becker, therefore, advocated for a shift in the focus of the study of deviance from the rule breakers to the people who make and enforce rules (cf Becker, 1973: p. 153).

Genesis of the Concept: The Quest of American Puritans for a Dry Society

The experience of alcohol prohibition in the USA (the Volstead Act 1920–33) where a reformist social movement succeeded in

establishing its particularistic interest as general moral standard will most probably have furthered questioning the creation and enforcement of rules in general. Consequently, labeling theory provided appropriate answers for intellectuals in a pluralist society such as the USA where Puritans claimed hegemony with the support of government bureaucracies.

David Riesman studying the 'American character' in 1950 mentioned moralizers as key figure of American politics of the nineteenth century (Riesman, 2001[1950]). This self-righteous fervor of moral entrepreneurs of the temperance movement was portrayed by Andrew Sinclair who showed that they would also disdain jazz music ('Does Jazz put the Sin in Syncopation?'), tango dancing, and so on, just as much as tobacco smoking (Sinclair, 1962: p. 181). For Joseph R. Gusfield, the historical case of the American temperance movement stands for a moral crusade, which typically escalates symbolic political controversies so that one group imposes its life style on another (Gusfield, 1986).

So, when Becker created the concept of the moral entrepreneur he could refer to a number of studies about moral entrepreneurs as rule creators intending to better the souls of others by their reformist activities. He himself illustrated the concept with the case of the Marihuana Tax Act and the role the Federal Bureau of Narcotics allegedly played in rounding up legislators for a law prohibiting marihuana (Becker, 1973: pp. 135ff.). A reanalysis of these data by Galliher and Walker (Galliher and Walker, 1976, 1978) challenged Becker's conclusions from this case. However, the concept as such was only questioned by general skepticism toward the allegedly relativistic premise of labeling theory and social constructionist views of deviance.

In 2009, Becker partly revoked his claim that the moral entrepreneurs in the war against opiate drugs, marihuana, and alcohol were by and large motivated by their Puritan views on the vice of seeking pleasure and loss of control with the help of these drugs. In the preface for Dudouet's study of the interests of the pharmaceutical industry fighting the illicit drug trade with the help of government bureaucrats in international organizations (such as the United Nations and its predecessor, the League of Nations), Becker depicts the war on illicit drug traffic as a purely symbolic cover story for controlling the legal market of the raw materials and manufactured narcotics. The influence of the large pharmaceutical companies in the West, both in terms of financial capital on the globalized market and steering political decisions that help in protecting their market positions, is the major power of the moral enterprise of prohibiting drugs, Becker argues (Dudouet, 2009).

Impact of the Concept: The Construction of Social Problems and Moral Panics

Labeling theory influencing many a study of deviance in the 1960s (both in the USA and UK) came under increasing attack in the 1970s. But by the 1980s, the social constructionist approach of moral entrepreneurship experienced a resurrection in the study of social problems. In *Constructing Social Problems* first edited in 1977, Malcolm Spector and John Kitsuse apply the concept of moral entrepreneurship in terms of 'claims-making activities' (2001: pp. 73ff.). While Becker (1966) in a radically social constructivist statement asserts that even nonexistent conditions may make people define social problems (e.g., witches or flying saucers) if they believe in the threat resulting from these conditions, for Spector and Kitsuse the interaction between claims making and responding constitutes the very social problem while defining a putative social condition as a social problem (2001: pp. 75ff.). Evidently, claims making requires the initiative of a moral entrepreneur either as a member of the group affected by the newly defined social problem or as an external evaluator.

The examples of social problem research consequently identify, e.g., various American public health service officials and political decision-makers as driving forces in labeling nicotine smoking as deviant; they were accompanied among others by private enterprises offering quit-smoking-classes who had an economic interest in the campaign (Nuehring and Markle, 1974; Markle and Troyer, 1979). Carolyn Wiener shows how the National Institute on Alcohol Abuse and Alcoholism contributed to the medicalization of alcohol drinking and how this social problem created a vast number of accomplices among the professionals offering treatment for the 'disease' alcoholism (1981). Contrarily, the failing attempts of anticoffee groups to vilify caffeine give an example that despite a health warning of the Federal Food and Drug Administration, the moral enterprise is doomed if its claims are neither supported by (medical) experts nor meet the interests of powerful political and economic allies (Troyer and Markle, 1984).

In the UK, Stanley Cohen (2002[1972]) took up the idea of the moral entrepreneur in his PhD thesis of 1969 where he studied folk devils of the 1960s' youth culture, and how moral entrepreneurs (businessmen, the police, politicians) labeled them as dangerous deviants. Cohen extended the concept with a distinct focus on the public's reaction and created the term 'moral panic,' which is based on the observation that in industrial societies the mass media play an important role in the construction of social problems. Moral entrepreneurs define a social problem and with the help of the media construct a moral panic. Again the economic interests – in this case of tabloid newspapers – constitute a driving force in constructing deviance not the least by the typical claim of the moral panic that the deviant act represents just the tip of the iceberg of some far greater threat to society. Interestingly enough, Cohen remarked in the third edition of his book, that the media themselves are frequently using 'his' term of moral panic, when denouncing a public outcry as hysteria (Cohen, 2002).

In Germany, the concept of moral crusades was applied in studies of the German radical left terrorism of the Red Army Fraction, Rote Armee Fraktion (RAF) of the 1970s (Sack et al., 1984). The German-based Austrian sociologist, Heinz Steinert, concluded that neither the rather small number of terrorist-victims in Germany, nor the nature of the crimes (consisting mainly of bank robberies and not so much political violence in the narrow sense) can account for the media-hype of the terrorist threat. Rather the response of the state to the challenge by the radical left (as it was expressed in various pamphlets and declarations by RAF-activists) denied

any legitimacy to their revolutionary concept and thus assisted their popularity by adding political meaning to 'ordinary' offenses such as bank robberies (Steinert, 1976, 1984: pp. 394ff.). The witch hunt against so-called fellow-travelers ('Sympathisanten') among intellectuals and artists served as a moral crusade in order to establish the border between the friend and enemies of the established order in postwar Germany (Treiber, 1984: pp. 325ff.).

In the 1990s, the study of moral panics was expanded by the historian Philip Jenkins who added analyses of child abuse (Jenkins, 1998), serial murder, Satanism, terrorism, and so on. He underlined that moral panics always serve the interests of occasionally most diverse groups, such as feminists and conservative moralists, law enforcement, child advocates, and professionals in the case of the various cycles of the moral panic on child abuse. In addition, they have to resonate with social needs resulting, e.g., from demographic change.

In 1994, Erich Goode and Nachman Ben-Yehuda offered a systematic theory of moral panics and operationalized the characteristics such as public concern (measured in opinion polls and social movement activities), disproportionality of the concern compared to its 'real' nature (measurable, e.g., in the false perception of a risk that is far less likely than another yet trivialized risk), wide consensus about the concern, volatility, and hostility toward the allegedly deviant group (Goode and Ben-Yehuda, 2009[1994]). They also theorize about the origins of the moral panics and distinguish grassroots panics, panics that are orchestrated by the elite and such carried by the moral or material interests of various interest groups and professionals. While the interest group theory (or assumption) is adequate to summarize a lot of the studies about the construction of social problems, the elitist theory could be suspected to serve as a conspiracy theory about politicians and the media in order to divert attention from the 'real' social problems. An example for the grassroots moral panic is the occasionally exaggerated perception of the danger of radioactivity. With Goode and Ben-Yehuda, the social constructionist tide in understanding the construction of moral panics wanes since their notion of disproportionality claims the existence of objective evidence about the relevance of a concern.

The Concept Today: Moral Entrepreneurs for 'Political Correctness'

As opposed to the term Moral Panic, the term Moral Entrepreneur does not seem to have been incorporated in the public in terms of usage in the popular press as it was studied by Merton and Wolfe (Merton and Wolfe, 1995). Yet, facing any of the present-day outrages about violated norms makes one wonder how the concept of the moral entrepreneur could sink into oblivion. The debate about political correctness from the 1990s onward would lend itself appropriate for an illustration of a moral crusade: the conservatives opposing so-called speech codes (e.g., the National Association of Scholars) warned of imposed orthodoxy in so-called sensitivity trainings and pointed out that the bureaucracies created for particular 'minorities' have an interest in emphasizing differences and promoting complaints in order to justify their very existence.

Further, also liberal advocates of free speech expressed concern about the silencing effect of political correctness and regarded it as a result of identity politics leading to exclusionary practices by moral judgments. Due to the considerable cultural influence of progressives in intellectual circles and academia who at the same time lack political power in society at large, they would tend to use ideology as a bludgeon against each other out of frustration (Aufderheide, 1992).

With new social media the opportunities to enforce pc-codes have multiplied and so have the opportunities for moral entrepreneurs advocating political correctness.

See also: Labeling Theory, History of.

Bibliography

Aufderheide, P. (Ed.), 1992. Beyond PC: Toward a Politics of Understanding. Graywolf Press, Saint Paul, MN.
Becker, H.S. (Ed.), 1966. Social Problems: A Modern Approach. Wiley, New York.
Becker, H.S., 1973. Outsiders: Studies in the Sociology of Deviance. Free Press, New York.
Cohen, S., 2002. Folk Devils and Moral Panics: The Creation of the Mods and Rockers, third ed. Routledge, London; New York.
Dudouet, F.-X., 2009. Le grand deal de l'opium: histoire du marché légal des drogues, Collection "Histoire: enjeux et débats". Syllepse, Paris.
Galliher, J.F., Walker, A., 1976. The puzzle of the social origins of the Marihuana Tax Act of 1937. Social Problems 24, 367–376.
Galliher, J.F., Walker, A., 1978. The politics of systematic research error: the case of the Federal Bureau of Narcotics as a moral entrepreneur. Crime and Social Justice 10, 29–33.
Goode, E., Ben-Yehuda, N., 2009. Moral Panics: The Social Construction of Deviance, second ed. Wiley-Blackwell, Chichester, UK; Malden, MA.
Gusfield, J.R., 1955. Social structure and moral reform: a study of the Woman's Christian Temperance Union. American Journal of Sociology 61, 221–232.
Gusfield, J.R., 1967. Moral passage: the symbolic process in public designations of deviance. Social Problems 15, 175–188.
Gusfield, J.R., 1986. Symbolic Crusade: Status Politics and the American Temperance Movement, second ed. University of Illinois Press, Urbana.
Jenkins, P., 1998. Moral Panic: Changing Concepts of the Child Molester in Modern America. Yale University Press, New Haven, CN.
Kutchins, H., Kirk, S.A., 1997. Making Us Crazy. DSM: The Psychiatric Bible and the Creation of Mental Disorders. Free Press, New York.
Markle, G.E., Troyer, R.J., 1979. Smoke gets in your eyes: cigarette smoking as deviant behavior. Social Problems 26, 611–625.
Merton, R.K., Wolfe, A., 1995. The cultural and social incorporation of sociological knowledge. American Journal of Sociology 26, 15–39.
Nuehring, E., Markle, G.E., 1974. Nicotine and norms: the re-emergence of a deviant behavior. Social Problems 21, 513–526.
Riesman, D., 2001. The Lonely Crowd: A Study of the Changing American Character, Abridged and rev. ed. Yale Nota Bene. Yale University Press, New Haven, CT.
Sack, F., Steinert, H., Berlit, U. (Eds.), 1984. Protest und Reaktion, Analysen zum Terrorismus. Westdeutscher Verlag, Opladen.
Sinclair, A., 1962. Prohibition. The Era of Excess. Little, Brown and Company, Boston; Toronto.
Spector, M., Kitsuse, J.I., 2001. Constructing Social Problems. Transaction Publishers, New Brunswick, NJ.
Steinert, H., 1976. Über die Fabrikation einer terroristischen Wirklichkeit und ihre realen Folgen. Kriminalsoziologische Bibliographie 4, 97–123.
Steinert, H., 1984. Sozialstrukturelle Bedingungen des "linken Terrorismus" der 70er Jahre. Aufgrund eines Vergleichs der Entwicklungen in der Bundesrepublik Deutschland, in Italien, in Frankreich und den Niederlanden. In: Berlit, U. (Ed.), Protest und Reaktion, Analysen zum Terrorismus. Westdeutscher Verlag, Opladen, pp. 388–564.
Sutherland, E.H., 1950. The diffusion of sexual psychopath laws. American Journal of Sociology 56, 142–148.

Treiber, H., 1984. Die gesellschaftliche Auseinandersetzung mit dem Terrorismus: Die Inszenierung "symbolischer Kreuzzüge" zur Darstellung von Bedrohungen der normativen Ordnung von Gesellschaft und Staat. In: Steinert, H., Berlit, U. (Eds.), Protest und Reaktion, Analysen zum Terrorismus. Westdeutscher Verlag, Opladen, pp. 320–386.

Troyer, R.J., Markle, G.E., 1984. Coffee drinking: an emerging social problem? Social Problems 31, 403–416.

Westley, W.A., 1953. Violence and the police. American Journal of Sociology 59, 34–41.

Wiener, C., 1981. The Politics of Alcoholism: Building an Arena around a Social Problem. Transaction Books, New Brunswick, NJ.

Zurcher, L.A., Kirkpatrick, R.G., Cushing, R.G., Bowman, C.K., 1973. Ad hoc anti-pornography organizations and their active members: a research summary. Journal of Social Issue 29, 69–94.

Moral Panics

James M Jasper, New York, NY, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 10029–10033, © 2001, Elsevier Ltd.

Abstract

Students of deviance, social problems, and politics have used the concept of a moral panic to describe sudden concern over a group or activity, accompanied by calls for control and suppression. Out of an infinite range of potential perceived threats, one – which may or may not be new – is suddenly the center of considerable attention. The news media, public officials, religious leaders, and private 'moral entrepreneurs' are key in focusing public attention on the issue, typically by identifying some recognizable group as 'folk devils' responsible for the menace. Human history has been full of rumors, fears, and efforts – often deadly – to constrain or eliminate the human source of perceived problems.

Students of deviance, social problems, and politics have used the concept of a moral panic to describe sudden concern over a group or activity, accompanied by calls for control and suppression. Out of an infinite range of potential perceived threats, one – which may be neither new nor on the rise – suddenly receives considerable attention. Marijuana use, motorbikes, and rock 'n' roll music are common examples. The news media, public officials, religious leaders, and private 'moral entrepreneurs' are key in focusing public attention on the issue, typically by identifying some recognizable group as 'folk devils' – usually young people, racial and ethnic minorities, or other relatively powerless groups – responsible for the menace. New political or legal policies are sometimes the result, and new symbols and sensibilities (available as the raw materials for future panics) almost always are.

The Concept of Moral Panics

Trying to understand a 1964 incident in which massive police and media attention were given to an evening of rowdy behavior by young people at an English seaside resort, British sociologist Stanley Cohen first formulated the concept of a moral panic in 1972, borrowing the term from his colleague Jock Young. In Cohen's (1972) words,

> A condition, episode, person or group of persons emerges to become defined as a threat to societal values and interests; its nature is presented in a stylized and stereotypical fashion by the mass media; the moral barricades are manned by editors, bishops, politicians and other right thinking people; socially accredited experts pronounce their diagnoses and solutions; ways of coping are evolved or (more often) resorted to; the condition then disappears, submerges or deteriorates and becomes more visible.

The key idea, attractive to cultural constructionists, is that the public reactions and rhetoric are not clearly linked to any objective measure of threat. Scholars who use the concept of moral panic frequently try to show that the rates of the activity are actually declining or at least not increasing, or that those promoting the panic have interests in doing so that are independent of the actual harm done (for instance, police departments may receive more funding if they can persuade the public that Satanic child abuse is rampant). The term 'panic' seems to imply a disproportionate or self-defeating response, although the feelings and thoughts of those panicking will probably be the same regardless of the extent of any real threat. Although some involved in a panic are cynically trying to manipulate public opinion, others (presumably those being manipulated) really do perceive the putative threat as real.

Most scholars have examined panics over aspects of lifestyle considered – at least by those embracing them, but also usually by the analyst too – to be harmless personal choices, such as clothing styles, sexual choices, or substance use. For them it is easy to show that, under certain value systems, any repressive action is unwarranted. Others have applied the term to actions that are clearly harmful, such as child abuse, but in which the response appears exaggerated or misplaced. For them it is more difficult to show that the response is disproportionate, although they may show that it is ineffective. The term has also been used to describe those opposed to technologies such as nuclear reactors. It is assumed that the emotional dynamics of a panic are similar in these diverse cases.

Many observers have found the concept of moral panics useful because it opens a window onto a society's disagreements over basic values, often intuitively felt ones, as well as onto fears and anxieties that are normally submerged. A panic over child abuse may reveal broader concerns over changes in family structure; one over water fluoridation, fears of government bureaucracy; one over crack cocaine, racist attitudes toward African American urban youths. Cultural meanings rarely reveal themselves so starkly.

Moral Panics in History

Human history has been full of rumors, fears, and efforts – often deadly – to constrain or eliminate human sources of perceived problems. When the Black Death came to Europe in the fourteenth century, neither physicians nor priests were of much help, so large numbers of frightened Germans blamed the Jews who lived among them, exterminating most and driving the rest into Central and Eastern Europe. Because few disasters have been as widespread as the plague, few panics have been as broad, but local instances have been common throughout known history.

In the nineteenth century more systematic efforts appeared, especially in the UK and the USA, to make the lower classes conform to the expectations of the middle and upper classes, especially those aspects of lifestyle considered (by those behind the efforts) to be part of moral character. Drinking, cruelty to animals, neglect of children, pornography, even movies came eventually to be seen as antisocial behaviors that well-intended benevolent societies tried to extirpate, partly for the good of the lower classes themselves. Unionization, rebellion, and urban riots too were seen as little more than a form of crime or bestiality. For more than 100 years, most moral panics consisted of the privileged worried about the less privileged. (Although lower-class panics persisted, such as anti-Catholic riots in the USA.)

For most of this era, intellectuals were more likely to contribute to moral panics than to analyze, and thereby criticize, the panics of others. Few scholars questioned the widespread fears of urban mobs that characterized the middle and upper classes of nineteenth-century Europe. In fact, they contributed to them, especially by dismissing 'crowd' behavior as irrational.

There seems to have been some kind of turning point around World War II, since which the sources of moral panics have grown more diverse. Intellectuals and academics, as they grew in numbers in the 1950s, instigated something like mild panics over conformity, the authoritarian personality, and mass society as signs of vulnerability to fascism or communism. Politicians were behind many of the panics over communist infiltration. Grassroots issue entrepreneurs seem to have motivated the American panic over fluoridation, however. In the USA, the majority of moral panics since the 1970s have been the work of the religious right, whose diligent organizing is one reason the USA seems to have more panics than most other advanced industrial nations. There, panics frequently cluster around sex for pleasure rather than procreation (loose morals, homosexuality, teenage pregnancy, abortion), around race and poverty (crack and other drugs, welfare mothers, urban crime and violence), and around immigration (the English language, standards in the schools, overpopulation, job and wage competition). Each panic carries with it clearly defined folk devils.

Although moral panics are hardly new, they seem to have become more frequent in recent decades, or at least more visible to national audiences. For one thing, advanced industrial societies have a 'social movement sector' of professional activists who act as moral entrepreneurs as they try to attract attention, funds, and volunteers. Such individuals have long existed – think of Thomas Paine or Anthony Comstock – but they now have extensive formal organizations, political networks, and direct-mail fund-raising techniques at their disposal. Second, immigration into many countries has increased in recent decades, accompanied by heightened tensions over group boundaries, especially ethnic ones. Newcomers often have habits that are poorly understood and easily used to stigmatize the group. Increased penetration of the media, especially television, into each and every 'deviant' subculture has further spread awareness of alternative life-styles, especially sexual freedoms. The Protestant right in the USA and elsewhere has discovered the power of life-style panics for raising funds, if not recruiting new members.

The Role of the Media

Much of the 'panic' in moral panics – the disproportionate nature of the response – is due to framing and amplification by communications media. If they do not report on a phenomenon, or do not frame it as something one should be frightened about, there is rarely a panic. Thus the rise of television everywhere and the wide readership for tabloid newspapers in the UK, for example, have been blamed for an increase in moral panics. In some views, the media reflect the values and interests of economic and political elites, and so focus on the groups and behaviors that frighten these elites. In a less conspiratorial view, the media simply look for unusual life-styles and behaviors as a means to attract viewers, readers, and so on. In doing so, they spread information about what mainstream culture considers deviant. Sometimes, it is deviant enough to ignite panic.

Cohen (1972) originally described three ways in which media coverage contributes to moral panics: in the exaggeration and distortion of what actually has happened; in predictions that the same kind of thing will happen again; and in the framing of key symbols in pejorative ways. To a large extent, these make the event or behavior into news.

Variations in Panics

A number of variations in culture and social structure affect the shape and outcomes of moral panics. One of the most important is the political, and especially the electoral, system. Where, as in the USA, politicians are left to raise their own funds and do their own campaigning in single-district elections, they may be tempted to play on the fears of the majority, which often have to do with life-styles of minorities. Where, as in Sweden, representation is more proportional and political parties control the chances of individual candidates, party leaders at least have the opportunity to curb efforts at creating moral panics for electoral purposes. Of course, this allows entire parties to play the panic card, as in the case of the National Front in France.

The structure of the media industry also has an effect. In some nations governments exert considerable control over the content of the media – although the widespread relaxation of such control in recent years may have contributed to an increase in moral panics. Their intervention allows governments to curb sensationalism and prevent moral panics, but it also allows them to incite them to their own advantage. On another dimension, in countries with nationally centralized news media, such as the UK, panics can spread more quickly than in countries where the media are more regional and local, as in France or the USA.

Other potential moral entrepreneurs face varying kinds of incentive structures. Local police departments, for instance, may need to appeal to city or regional governments for additional funding. Certain occupations may have opportunities for advancing into new terrain because of moral panics. The more that a taxation or funding system allows input from local units or from the general population, the more tempting it is for those whose funding comes from these sources to use moral panics to increase their budgets.

Some moral panics reflect widespread public anxieties and fears that are easily stirred up through news coverage. Others seem traceable to specific interest groups who will benefit from the panic. Goode and Ben-Yehuda (1994) distinguish three models of moral panics, according to their origin. In one view, moral panics are relatively grassroots in origin, in a second they are organized by elites, and in a third vision they are promoted by interest groups or social-movement organizations. But there is no reason to think that all panics must come from the same source, so that one model would fit them all. Goode and Ben-Yehuda's typology functions equally well for distinguishing different types of panic as different scholarly models: some panics are indeed more grassroots in origin, others are more elite-directed, and yet others are promoted initially by social-movement groups. In addition, some seem entirely a creation of the news media. These types may differ in their dynamics and outcomes as well as their origins.

As for differences in the general public, many commentators have assumed (more than demonstrated) that there are national differences in susceptibility to 'mass mobilization' in instances such as moral panics. It has been thought that a lively 'civic culture' should prevent the kind of demagoguery often associated with moral panics, and that formal education breeds tolerance for the alternative life-styles that are often demonized in panics. This evidence is not as strong as one might wish. Just the kind of educated, normally tolerant people one might expect to resist moral panics often end up leading them. One of their frequent (if not defining) features is the relative absence of critical, dissenting voices.

Controversies

The concept of moral panics has not lacked critics. Like all varieties of cultural constructionism, the analysis of moral panics can be turned against those who use it. Just as they find bias and exaggeration in their subjects of study, so their own critics can claim to find the same flaws in research on moral panics. Most scholars who have used the concept have been left-leaning and thus quicker to discern panics on the right, for instance. A more fundamental criticism is that the term panic is necessarily pejorative, assuming rather than demonstrating that those analyzed are not entirely rational. To prove irrationality or disproportionality, analysts must presumably investigate the goals of those accused of panic (often a complex and poorly defined cluster of goals) and show that they are undermining their own progress toward those goals.

Debates over the utility of moral panics as an analytic tool tap into deep disagreements about the rationality of protest and politics. For most of the modern era, students of collective protest viewed it as immature, irrational, or manipulated, a kind of crowd dynamic leading people to do things they would not do otherwise. In this tradition panics, like fads, were analyzed in the same way as movements for unions or civil rights. To critics, moral panics are a holdover from this pejorative vision that dismissed protestors as irrational.

In recent decades a very different view of such activity has arisen, in which collective action is seen as a pervasive and regular form of politics, in which groups and individuals organize to pursue their own interests. They mobilize resources, look for political opportunities in which to act, and formulate rhetoric to appeal to as many potential supporters as possible. The concept of panic seems to undermine this positive view of such activity, but in fact it focuses attention on precisely the dynamics that this rationalist perspective tends to overlook. Issues must be chosen from a myriad of possibilities; grievances must be constructed; emotions must be aroused, often through the demonization of others; previously private issues must be portrayed as a public concern. The dynamics highlighted by research on moral panics, in other words, characterize a great deal of politics both institutionalized and not.

Moral panics remain a common occurrence in modern societies. They generate considerable scholarly interest, although some care must be taken to avoid loaded language and tendentious arguments about them – and some will always think that the term 'panic' itself is necessarily loaded.

See also: Emotions, Sociology of; Mass Communication: Normative Frameworks; Mass Killings and Genocide, Psychology of; Mass Media and Cultural Identity; Media and Social Movements; Psychoanalysis in Sociology; Social Movements, History of: General; Social Movements: A Social Psychological Perspective; Social Psychology, Sociological.

Bibliography

Beisel, N., 1997. Imperiled Innocents: Anthony Comstock and Family Reproduction in Victorian America. Princeton University Press, Princeton, NJ.
Cohen, S., 1972. Folk Devils and Moral Panics: The Creation of the Mods and Rockers. MacGibbon and Kee, London.
Goode, E., Ben-Yehuda, N., 1994. Moral Panics: The Social Construction of Deviance. Blackwell, Cambridge, MA.
Hall, S., Critcher, C., Jefferson, T., Clarke, J., Roberts, B., 1978. Policing the Crisis: Mugging, the State and Law and Order. Macmillan, London.
Jenkins, P., 1992. Intimate Enemies: Moral Panics in Contemporary Great Britain. Aldine de Gruyter, New York.
Jenkins, P., 1998. Moral Panic: Changing Concepts of the Child Molester in Modern America. Yale University Press, New Haven, CT.
Springhall, J., 1998. Youth, Popular Culture and Moral Panics: Penny Gaffs to Gangsta Rap, 1830–1996. St. Martin's, New York.
Thompson, K., 1998. Moral Panics. Routledge, London.
Waddington, P.A.J., 1986. Mugging as a moral panic: A question of proportion. British Journal of Sociology 37, 245–259.
Watney, S., 1987. Policing Desire: Pornography, AIDS and the Media. Methuen, London.

Moral Reasoning in Psychology

Elliot Turiel, Graduate School of Education, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Developmental research has focused on reasoning about right and wrong in social relationships. Piaget and Kohlberg each proposed sequences in which children first develop moral judgments that are confused with nonmoral features, such as adherence to social conventions, and then judgments differentiating them from each other. However, subsequent research shows that young children form concepts in the moral domain with understandings of welfare and fairness, differing from judgments about social conventions. Moral decisions entail coordination of different concerns – moral and nonmoral – in situational contexts. Individuals critically apply moral judgments to cultural practices entailing social inequalities.

Moral reasoning, involving concerns with welfare, justice, and rights, has been analyzed extensively by philosophers. Informed by philosophical expositions, psychologists have researched the development of moral judgments from early childhood to adulthood. A large body of research has revealed that young children begin to develop distinct judgments about avoiding harm to others and promoting welfare, along with other domains of social reasoning. Development in the moral domain also involves increased understandings of justice and rights. In concrete social decisions and actions there is a coordination of different domains of thought including morality, the conventions of social systems, and areas of personal jurisdiction. Since many social systems are hierarchically organized, conflicts can be generated by the application of moral judgments to existing inequalities and injustices.

The term morality has been used in different ways, and its explanation is subject to a great deal of debate. One controversy pertains to whether morality is particular to cultures or if aspects of morality are applicable to humans across cultures. These debates also include considerations as to whether moral values are arbitrarily set in a group, society, or culture or are absolute in that they must be categorically followed. However, the most useful conceptualizations on either side of these debates pertain not to arbitrary values *or* absolutism, but to morality that forms part of integrated cultural systems, which might differ from culture to culture *or* moral precepts that are in general applicable across situations but which are not absolute in that they might not be applicable in all situations. Various explanations have been provided as to how individuals acquire the morality of their culture, including ones that emphasize the learning of a set of values or the incorporation of cohesive systems of moral thought (*see* Moral Development, Cultural Differences In). Some maintain that moral judgments are based on intuitions that do not involve reasoning. However, an alternative psychological perspective, with roots in philosophy (e.g., Immanuel Kant, John Stuart Mill, John Rawls, Martha Nussbaum) is that morality involves reasoning about right and wrong with regard to social relationships, legal systems, and societies.

Social Interactions and Moral Judgments

In psychology, much research has focused on the development of moral reasoning, involving constructions of judgments about how social relationships ought to occur, with understandings of welfare (avoiding harm and promoting good), justice (in the distribution of goods, and equal worth of persons), and rights (group and individual). Such understandings are a function of children's social interactions and emotions seen to occur the world over.

Two developmental psychologists have had a deep and lasting influence on thinking about moral development. One is Jean Piaget, who is regarded as one of the foremost contributors to developmental psychology more generally. Piaget's (1932) research on children's moral judgments was conducted in the late 1920s when sociologically oriented views of morality were predominant in the social sciences. As an example, the sociologist Emile Durkheim (1925/1961) had proposed that morality constitutes a shared respect for the rules and values of society, and that children must be educated to form a strong emotional attachment to the group and to respect its rules and authority. Piaget, by contrast, examined the ways children construct ideas about rules, authority, fairness, and equalities. The influence of Piaget's theory of moral development on psychology was limited until the early 1960s. Up to that point, greater attention was given to psychoanalytic and behavioristic theories, each of which explained moral development as the incorporation of society's standards. The research of Lawrence Kohlberg (1969) served to refocus attention on moral reasoning.

Two Moralities and Two Types of Social Interactions

Piaget examined the organization of children's thinking and how they interpret events. In his work on moral development, he observed children's use of rules in playing games, sharing (distributive justice), punishment (retributive justice), and intentions or consequences in assessing responsibility for transgressions. He proposed a sequence of two broad levels of children's moral judgments connected to two ways they relate to others. In the first, children (approximately 4–7 years) are

heteronomous in their moral thinking and have unilateral respect for adults in authority. Heteronomy is characterized by absolutistic judgments, whereby social rules are regarded as fixed and requiring strict adherence. As an example, young children reify game rules, judging that they cannot be altered even if all the participants were to agree to the changes. They also believe that adults' commands must be followed to the letter. Children also judge actions by their material consequences rather than by the actor's intentions.

In part, heteronomous thinking is due to inherent relationships of inequality. Young children attend mainly to the greater power and status of adults – which imbues adults with moral authority. In later childhood there is a shift in moral thinking that occurs with increasing interactions with peers, which provide children with more opportunities to experience mutuality, reciprocity, and cooperation. As a consequence children shift from heteronomy to autonomy, defined as involving participation in the elaboration of norms (as opposed to receiving them readymade). Through the reciprocity in relationships of equality children construct understandings of justice and mutual respect, which they differentiate from obedience to authority and strict adherence to rules.

Stages in the Development of Moral Judgments

In extensions of Piaget's research, further emphasis was placed on the need to ground social and behavioral research on definitions of the realm. In the dominant approaches of the time, considerations of the psychological means by which children acquire morality were addressed without formulations of the substantive elements of morality. Kohlberg (1969), especially, was instrumental in combining psychological with philosophical analyses (relying to a fair extent on neo-Kantian philosophers such as Rawls, 1971). One reason for stressing inclusion of philosophical formulations was that Kohlberg proposed that children think systematically about moral matters, forming organized ways of thinking about concepts of justice, rights, and welfare.

Kohlberg, too, regarded the development of moral judgments to involve progressive differentiations or distinctions in moral and nonmoral understandings. In his research, children were presented with hypothetical stories that depicted conflicts around issues of life, physical well-being, trust, law, authority, and retribution. (A well-known story depicts the conflict of a man who must decide whether to steal an otherwise unavailable drug to save his wife's life.) Kohlberg formulated a sequence of stages and levels that portrayed children's moral judgments somewhat differently from Piaget's characterizations. In particular, it appeared that young children do not reify rules nor possess an unyielding respect for authority (according to Kohlberg, Piaget had misinterpreted a moral orientation to power and sanctions as reification). Rather, in the first two stages (labeled a 'preconventional' level) moral judgments are not differentiated from material value, sanctions, power, and attaining people's needs and desires. In early adolescence, there is a shift to thinking about morality primarily in terms of social expectations and maintaining the social order by upholding laws and authority (the 'conventional' level). After attaining understandings of morality as rooted in rules or laws and the need to maintain the social order, there emerges a way of thinking that includes applying moral judgments to society and its practices, rules, and authority. At the most advanced level (the 'postconventional'), which does not develop at least until late adolescence, concepts of justice, fair treatment, rights, and respect for persons are differentiated from the existing social order. At the most advanced level, individuals make moral judgments about society's system of laws and norms.

Distinguishing Morality from Other Domains of Social Judgment

During the latter part of the twentieth century Piaget and Kohlberg influenced a great deal of research. However, the patterns of developmental changes they described have been questioned. In particular, research indicates that young children's moral judgments are neither heteronomous nor based on a confusion of moral value with material value or sanctions. One line of research demonstrates that young children have positive orientations to matters like helping and sharing (*see* Prosocial Behavior and Empathy). Another line of research shows that young children develop understandings of welfare and fairness distinct from understandings of other social domains (Turiel, 2006). Using research methods that do not pose complex conflicts between moral and social considerations (as was the case in Kohlberg's methods), it was demonstrated that children do not treat all rules as fixed, do not solely have unilateral respect for authority, and do not judge by sanctions.

Children's moral ideas stem from interactions with children and adults that involve observing and experiencing harmful consequences, inequalities, or violations of freedoms. These moral judgments differ from another set of judgments about social uniformities that can be termed 'conventions.' Conventions are socially expected behaviors that coordinate interactions within social systems. Such uniformities in behavior gain their force from being customary and allow for greater efficiency in that people know what to expect of each other. Children do not judge moral issues (such as hitting or stealing) by the existence of uniformities, agreements, rules, or common practices. For instance, rules pertaining to moral issues are judged not to be legitimately alterable by general agreement or dependent on what authorities dictate, or relative to the social and cultural context. Maintaining the welfare and rights of persons, as well as promoting fairness in social interactions and cultural practices, guides evaluations of actions, rules, authorities, and the organization of social systems.

By contrast, conventions, such as those pertaining to modes of greeting, forms of address, or dress codes, are judged by children to be alterable by consensus, contingent on common use, rules, or authority expectations, and relative to social contexts. Thus, children judge violations of conventions to be wrong only insofar as there is a governing rule or if it is dictated by an authority (e.g., a conventional rule in a school can be changed by the principal). Similarly, conventions are not generalized across contexts in that it is judged that other cultures can legitimately have a different set of customary behaviors.

Morality is a fundamental domain of social reasoning. Children's judgments within the moral domain, which begin to emerge at a fairly young age, do not entail reification of rules or confusion with other social domains. Changes in moral reasoning occur with age. A young child's moral reasoning is focused primarily on welfare – preventing harm and promoting welfare. Older children and adolescents develop increased understandings of justice and reciprocity among persons, which they coordinate with concepts of welfare. Research has also shown that young children's moral judgments are associated with the early emergence of emotions like sympathy, empathy, and affection. It appears that such positive emotions are more important to the development of morality than negative emotions like shame, guilt, and disgust. Moreover, children's actions are connected with the types of moral judgments they make. It has been demonstrated that the moral judgments of children and adolescents about hypothetical situations closely correspond to their judgments about actual events they experience (Turiel, 2008).

Coordination in Moral Decision-Making

The different domains of social reasoning have implications for explanations of how people make decisions in social situations. Decision-making involves coordination in that the different domains, as well as different considerations within the moral domain, are taken into account and priorities are drawn. These processes of coordination have been examined in research on honesty or trust, and on rights. For example, whereas honesty is generally considered morally right, in some circumstances it is not necessarily the morally correct course of action. An example discussed in philosophical discourse is a situation in which lying might be necessary to save a life. Studies with adolescents and adults show that they consider it acceptable to deceive a person who directs someone to engage in acts considered morally wrong. Similarly, coordination of different considerations is seen in decisions about rights like freedom of speech and religion. Although children and adults endorse rights, in many situations they subordinate rights to matters like preventing harm or promoting community interests.

Social Hierarchy and Social Justice

People make social decisions by which moral judgments are applied in concrete situations and in the context of societal arrangements. Social hierarchy, by which groups are accorded greater or lesser power is an important feature relevant to moral, social, and personal activities considered out of the boundaries of conventional or moral regulation. Often, hierarchical distinctions are made on the basis of social class and gender (Nussbaum, 1999).

Since women are often in subordinate positions in the social hierarchy, they experience injustices and inequalities, and in some cases are victims of oppression. The social system, therefore, can further the personal interests of some (e.g., males) at the expense of others (e.g., females) through social expectations and practices. Such arrangements constitute a source of tension and conflict within social systems. Whereas people in lower positions are part of the social system and do accept norms and practices, at the same time they are critical of aspects of social arrangements. In many respects, moral reasoning produces social harmony since people are concerned with the welfare of others and promoting fairness. Another side of morality, however, is that it can result in conflicts and efforts at subverting societal arrangements judged to be counter to welfare, justice, and rights. Research in several patriarchal cultures has shown that females judge cultural practices entailing social inequalities to be unfair and attempt to achieve changes toward greater justice (Abu-Lughod, 1993; Turiel, 2002).

See also: Human Rights, Anthropology of; Hunting and Gathering Societies: Anthropology; Infancy and Childhood: Emotional Development; Morality 'East' and 'West': Cultural Concerns; Practical Reasoning: Philosophical Aspects; Problem Solving and Reasoning, Psychology of.

Bibliography

Abu-Lughod, L., 1993. Writing Women's Worlds: Bedouin Stories. University of California Press, Berkeley.

Durkheim, E., 1925/1961. Moral Education: A Study in the Theory and Application of the Sociology of Education. Free Press, New York.

Kohlberg, L., 1969. Stage and sequence: the cognitive-developmental approach to socialization. In: Goslin, D.A. (Ed.), Handbook of Socialization Theory and Research. Rand McNally, Chicago, pp. 347–480.

Nussbaum, M.C., 1999. Sex & Social Justice. Oxford University Press, New York.

Piaget, J., 1932. The Moral Judgment of the Child. K Paul, Trench, Trubner, London.

Rawls, J., 1971. A Theory of Justice. Belknap Press, Cambridge, MA.

Turiel, E., 2002. The Culture of Morality: Social Development, Context, and Conflict. Cambridge University Press, Cambridge, England.

Turiel, E., 2006. The development of morality. In: Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology. In: Eisenberg, N. (Ed.), Social, Emotional, and Personality Development, sixth ed., vol. 3. Wiley, New York.

Turiel, E., 2008. Thought about actions in social domains: morality, social conventions, and social interactions. Cognitive Development 23, 136–154.

Morality 'East' and 'West': Cultural Concerns

Kwang-Kuo Hwang, National Taiwan University, Taipei, Taiwan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Eastern' moral systems entail the ethics of community or the ethics of divinity, which emphasizes the importance of fulfilling one's duties or attaining to group goals. In contrast, 'Western' morality highlights an individual's right of choice with utilitarian consideration, rather than any enforced social demand of deontological concern. Negative duties are universal moral codes in most human societies, while positive duties are specific to non-Western societies, which are experiencing transformation in order to preserve their core values under the impact of Western individualism.

Comparison of Three Methodologies

Three main approaches in the social and behavioral sciences inform the comparative study of moral reasoning and behavior prominent in 'Western' societies following the Judeo–Christian tradition and 'Eastern' societies following Confucianism, Islam, and Hinduism: developmental psychology, anthropological ethnography, and cultural psychology. These three approaches result in different understandings of morality in the 'West' and the 'East.' This article first offers a comparison of the methodologies of these three approaches and then provides a detailed examination of research on Western and Eastern moral systems.

Developmental Psychology

Kohlberg's (1981) theory of moral development and research methods were widely adopted by child psychologists in the 1970s and 1980s to study the cognitive development of moral reasoning and moral judgment in children of varying ages in different cultures. Based on the assumption of cross-cultural universality in moral development, Kohlberg claimed that "All individuals in all cultures use the same basic moral categories, concepts, and principles, and all individuals in all cultures go through the same order on sequence of gross stage development, though they vary in rate and terminal point of development" (Kohlberg, 1971: p. 175). His theory consists of six universal stages of cognitive development in moral judgment.

Kohlberg's theory adopts the perspective of evolutionism, which is based on a normative model and classifies a great diversity of moral discourse from a multitude of cultures into several categories arranged in a formal sequence. According to his theory, the development of moral reasoning follows a universal invariant sequence toward the goal of thinking with universal ethical principles.

A series of empirical studies have been conducted using Kohlberg's standardized moral dilemmas and scoring manuals (Marchand-Jodoin and Samson, 1982; Parikh, 1980). After careful examination of the literature, Snarey (1985) found that because stage skipping and stage regressions were rare, progress from Stage 1 to Stage 3 or 4 was virtually universal, but that the presence of Stage 4/5 or 5 was extremely rare in all populations. Nearly all samples from Western middle-class groups and 91% of Eastern urban populations exhibited some Stage 6 principled reasoning. No tribal or folk cultural groups of the non-Western world showed any postconventional thinking. Postconventional thinking upholds individual rights or universal ethical principles.

Many of the moral reasoning data collected in collectivist or communalistic societies of the non-Western world either could not be scored according to the standardized manual or could not be explained in the context of Kohlberg's theory (e.g., Maqsud, 1979; Sullivan, 1977; White et al., 1978). Such anomalistic facts have been used to challenge Kohlberg's theory. Critics charge that though Kohlberg's preconventional and conventional stages are well based on empirical operative judgments, his descriptions of higher stage reasoning at postconventional stages are primarily based upon Kant, Rawls, and other Western philosophers (Snarey, 1985). This perspective entails Western ideologies of rationalism, individualism, and liberalism, and therefore assigns the values of male, Caucasian, American intellectuals as the end point of moral maturation (Buck-Morss, 1975; Edwards, 1985). The theory's ethical objectivism ignores the moral discourse prevailing in many non-Western societies, and fails to recognize their substantial ethical philosophies (Weinreich-Haste and Locke, 1983). Imposing Kohlberg's theory as well as his scoring system on the moral reasoning of non-Western peoples may result in systematic bias originating from Western ethnocentrism.

Gibbs (1979) tried to revise Kohlberg's theory in terms of Piaget's phylogenetic perspective. According to Piaget, human intelligence is a holistic phenomenon encompassing social, moral, and logico-physical aspects. Its development follows a standard sequence that is predetermined by epigenetic factors like those of other species. Along with completion of physical maturation and the singular achievement of intellectual development over the course of the adolescent years, there is a progressive ability for people at the highest stage of standard development to reflect on the very conditions of their existence in the world. Reflection on one's own existence may motivate one to define a moral theory to justify one's basic moral principles from a particular standpoint that is either inside or outside the society (e.g., Stacey, 2011).

As many studies have shown, different cultures show wide variety in their philosophies of morality. These normative philosophies may provide material for second-order thinking or metaethical reflection when individuals try to define their

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64005-9

own moral theories. It is impossible for individuals to eliminate cultural influences from the contents of their own moral reasoning. Considering this fact, a better way to conduct cross-cultural research on morality would be to use ethnographic analysis to describe the patterns of moral conduct and moral thinking that consistently emerge in the daily social life of people in a particular culture.

Anthropological Ethnography

Researchers who study the ethnography of morality usually adopt the viewpoint of interpretive anthropology, which assumes that culture is an organized collection of symbols with the major function of maintaining social order for normal operation (Erickson and Donald, 1997; Gubrium and Holstein, 1997). As derivatives of the cosmology of a given culture, the concepts of morality and personhood, as well as other symbolic systems, can best be understood in the holistic context of that culture.

Many anthropologists have used the ethnographic approach to study indigenous religions, cosmology, personhood, and concepts of morality in non-Western societies (e.g., Edel and Edel, 1959; Howell, 1997). They tend to adopt the insiders' point of view to interpret the moral ideas, moral discourse, and moral practices in a particular society. Instead of relying solely on philosophical analysis and anthropological records, they pay special attention to the dynamic interplay between abstract moral principles and particular social events in a given community.

Researchers following this approach have often adopted the viewpoint of relativism for depicting the uniqueness of each culture. Paradoxically, members of a cultural community tend to insist on the viewpoint of absolutism: They may believe that their moral standard is superior to that of other cultural groups, and use it as a criterion for defining what is meant by *person* or for differentiating their own group from outsiders (Hcelas and Lock, 1981; Carrithers et al., 1985; Rapport, 1997). Those who violate their moral standards might be condemned as *not a person*. The process of dehumanization may lead to severe interpersonal or even intergroup conflict in extreme cases (e.g., Melhuus, 1997).

Cultural Psychology

Cultural psychologists investigate the cultural foundation of psychological diversity in self-organization, cognitive process, emotional function, and moral reasoning among people living in different cultures. In order to understand native peoples' symbolic and behavioral inheritance in terms of their shared framework of interpretation, cultural psychologists also adopt an ethnographic approach, but with the perspective of pluralism. Pluralism assumes that people all over the world are similar in their psychological structure and function, although aspects of their knowledge, thinking, feeling, values, needs, and behavior are activated by the experience of living in a particular cultural community. Cultural psychologists advocate the slogan: "One mind, many mentalities; universalism without uniformity" (Shweder et al., 1998: p. 871). They work to develop universal theoretical frameworks for interpreting the thinking and behavior of people in different cultures. Although it would be impossible to depict all moral systems in the world, it is possible to describe the sharp contrasts between moral systems of the East and West with the conceptual schemes provided by cultural psychologists.

Morality

Cultural psychologists suggest that three main domains of moral discourse exist in human society: the ethics of autonomy, the ethics of community, and the ethics of divinity (Shweder et al., 1997). Because the moral experiences of human beings are so complicated and the recognition of reality so restricted, even though all moral systems contain elements of all three of these domains, any particular moral system will tend to emphasize a particular domain of moral discourse.

Morality of the West

The contemporary Western moral system is characterized by the ethics of autonomy, which can be traced to the Judeo–Christian religious tradition. Christianity assumes that the world was created by God, who also created humankind in accordance with His own image. Since all humans were created in God's image, all humans are equal. Every person has the essence of humanity and should obey the divine commandments declared by God.

This concept of person and its underlying discourse of divinity changed drastically after the Renaissance in Europe. The Cartesian philosophy of dualism detached individuals from the outer world where they find themselves. Rationalism emphasized an individual's free will and subjective reasoning. The expansion of modern capitalism also made significant contributions to the rise of individualism (Turner, 1988).

The Western concept of self-formed, in the cultural milieu of individualism, requires a dualistic view of the world. It draws sharp distinctions between mind and matter, cognition and affection, self and society, nature and culture, and so on. Self is conceptualized as "a dynamic center of awareness, emotion, judgment and action organized into a distinctive whole and set contrastively both against other such wholes and against its social and natural background" (Geertz, 1974: p. 225). The inner life of the self, with its spontaneity, privacy, uniqueness, genuine feeling, and constancy, has a clear-cut boundary that separates it from the outer social world constituted by mask, role, rule, or context.

In the Western view of personhood, the individual is conceived of as prior to and more fundamental than society, which is treated as an artificial construction designed mainly to satisfy human needs. When a social order can no longer serve the function of realizing this end, it should be redesigned and reconstructed to restore its function. Based on this premise, Western morality is characterized by an emphasis on the rights of individuals. According to Dworkin (1977), all moral codes encompass personal rights, personal duties, and social goals, but they may differ in the priority given to these three categories. Western morality is premised on the concept of natural rights, rather than on natural duties or social goals.

Because individuals are construed as quasi-sacred and absolute, their legitimate demands should be carefully respected.

The individual is an ultimate and indivisible monad, and every human group is made up of monads with the same human nature. Individuals' free will and right to liberty are natural, fundamental, and inescapable. By the same token, the duty to respect others' rights is equally natural and inescapable. One's rights are restrained by the identical rights of other individuals, and everybody is encouraged to strive for the rights of all human beings to the greatest overall liberty (Dumont, 1986).

The realization of negative duties of omission (e.g., not killing, not cheating, not stealing, etc.) is considered to be the prerequisite of a moral agent. Violating moral codes of this category might be condemned as evil. In contrast, the practice of the positive duties of commission is subject to an individual's right of choice. Individuals who decide to undertake an altruistic act may be admired for their virtue. As a consequence, a certain state of tension is frequently aroused between social demand and the individual's desire for freedom of choice. The autonomous view of self tends to entail the belief that performing social obligations in spite of individual rights may have the negative effect of restricting one's own personal liberties. Even if one can justify working for the benefit of another person, the act of undertaking this responsibility is a resolution of personal choice with either utilitarian or contractrian consideration, rather than an enforced social demand of deontological concern.

As indicated by Gilligan (1982), in contrast to the Kohlbergian ethics of justice for males, the moral reasoning of Western females can be characterized by the ethics of caring and responsibility. However, both the deontological view of morality held by Kohlberg and the morality-of-caring framework developed by Gilligan are bound to the Western culture of individualism. Both stress freedom of choice and individual responsibility. Commitments to interpersonal caring are viewed as matters of personal decision (Miller, 1994), rather than as obligations to a particular social target in Eastern societies (Mohmood, 2005).

Morality in the East

In sharp contrast to the individual-centered morality of the West, contents of moral systems in the Eastern world, no matter whether they originated from Confucianism, Hindusim, Islam, or Buddhism, can be described as duty based. Unlike typical Western dualism, most Eastern philosophies advocate a monistic view of origin. That is, one's life is construed as an inseparable part of the world in which one finds oneself, not as an independent entity standing in opposition to the outer world. For example, traditional Chinese philosophy conceives of the formation of all things in the universe as the result of dynamics between the two opposing forces of *ying* and *yang*. A person's life is also produced from the unification of *ying* (female) and *yang* (male). In Confucianism, a family is conceived of as the body of a human being; each role in the family represents a distinct part of the body. For instance, the relationship between parent and child is like that between flesh and bone, while brothers are considered to be the four limbs (Hwang, 1999).

The Hindu view of life is also premised on the functional interrelatedness of the universe. From the earliest times of *Rigveda* (about 1500 BC), the Hindus believed that *Rta* (dynamic equilibrium) is the governing dynamic principle underlying functional equilibrium of the cosmic order. The famous *Purusha sukta of Rigveda* assumes the functional relatedness of all living organisms. Any existing phenomena many attain a state of *Rta* when each of its parts remains in its proper place and functions in its own law of activity. Humans, animals, and plants are interrelated. They all are part of the cosmic immanent life force. The individual functioning of one in coordination with the individual functions of others results in collaborative equilibrium of the whole universe (Heimann, 1964). In the same vein, Hindus conceptualize their society as a huge cosmic organism in which certain functions are allotted to each of the four major parts: the head, arms, trunk, and feet. Each of these four represents one of the four castes of Hindu society. Remaining in the proper function fulfills the goal of smooth collaboration of the whole society.

Instead of drawing the boundary of an individual around the immediate surface of one's physical body, most Oriental philosophies conceive of this boundary as expandable to include one's family or the whole society. Followers of Oriental philosophy tend to believe that external forces predetermine the social order. Role relationships within one's primary group are especially unchangeable through one's own efforts. Therefore, Oriental moral systems endow mandatory features not only to negative duties of omission, but also to some positive duties of commission. Examples include filial piety (unconditional positive duty) and loyalty (conditional positive duty) in Chinese society (Hwang, 1999, 2012), repaying *on* (favors or kindness) in Japanese society, and *utang na loob* (debt inside oneself or a debt of gratitude) in Filipino society. These kinds of positive duties or categorical imperatives urge individuals to fulfill obligations to specific targets in the community. These obligations ensure the smooth functioning of society, rather than protect the rights of individuals. In Dworkin's (1977) conceptual framework, Eastern moralities are based on duties or group goals rather than rights (Ahmed, 1993; Ali, 2006).

Compared with Western ethics of autonomy, ideas about individual rights are relatively uncommon in traditional Eastern moralities. Even moral discourse arguing for negative duties in Eastern moral systems are proposed not on the premise of protecting individual rights, but in consideration of maintaining social order or for religious reasons. For example, the act of killing others violates the principle of benevolence (*ren*). *Ren*, the requirement that all people carry out their moral duty to significant others in their intimate society, is the most fundamental moral rule for maintaining psychosocial homeostasis in Chinese society. Buddhists must adhere to five requirements: do not kill, steal, lie, be lustful, or taste intoxicants or meat. Violation of any of these requirements may result in *Karma* bringing automatic retribution in a later life. According to Islam, the present life is short and temporary, while the future life is eternal and everlasting. Those who devote themselves to charity, especially those who sacrifice themselves for the almighty Allah, may ascend to Heaven. Those who indulge in evil doings will descend to Hell. This kind of moral discourse fits within the model of Eastern moral systems that entail the ethics of community or the ethics of divinity, rather than the ethics of autonomy.

Moral discourse supporting the ethics of community or divinity in the Oriental world as revealed by sages or prophets

are recorded in classics such as *The Analects* of Confucianism, or religious scriptures such as the Qur'an of the Muslim faith, or the *Upanishads* of Hinduism. Intellectuals who are able to interpret those classics or scriptures, like Confucian scholars in traditional China, Brahmans in Hindu society, priests in Islamic countries, and monks in Buddhist societies, have occupied high positions in traditional and even contemporary societies. Their position is similar to that of priests in the Christian world before the European religious Reformation movement.

In contrast, the modern Western view of morality advocates a position of secularism, which assumes that all moral agents morally equivalent. Anyone can discover the natural law of morality without the assistance of revelations by a prophet or sage.

Global Ethics and Cultural Change

Since the collapse of the Soviet Union and East European communist regimes at the end of the 1980s, most countries have been incorporated into the world economic system of capitalism. The rapid expansion of opportunity for intercultural contact and exchange between people from the East and West has facilitated the formation of such concepts as multiculturalism and global ethics in the age of globalization. As part of the Universal Ethics Project sponsored by UNESCO, a group of international scholars of philosophy have proposed a Universal Declaration of Human Responsibilities, consisting of a list of global ethics with four categorical imperatives proposed on the basis of a fundamental demand that every human being must be treated humanely (Kung and Kuschel, 1993):

> You shall not kill!
>
> You shall not steal!
>
> You shall not lie!
>
> You shall not commit sexual immorality!

An examination of the list reveals that it consists of negative duties found in most cultures, while positive duties specific to a particular Eastern culture are undergoing drastic cultural changes under the impact of Western individualism, commercialism, and materialism. For example, as an unconditioned positive duty in Confucian culture, filial piety has been the core value of familism in East Asian societies. Empirical study on the basis of a dual model distinguishing between reciprocal and authoritarian filial piety indicated that diverse evolutional paths of contemporary filial piety do appear in such Chinese societies as Taiwan, Hong Kong, and China, with a general tendency of diminishing importance in authoritarian filial piety, while reciprocal filial piety remains its persistence (Yeh et al., 2013). In accompany with such cultural changes, the traditional pattern of arranged marriage has evolved several transitional patterns of mate selection in East Asian countries, but autonomous marriage has become the shared trend with cohort variation in Japan, Korea, and Taiwan (Tsutsui, 2013). By the same token, intergenerational relations in those areas share similar features in both of its change and continuity patterns: while coresidence between generations remains patriarchal, the main flow of integer support is from adult children to parents. Sons tend to perform various filial duties much more than daughters, but the actual practice of filial responsibility tends to be shared among adult siblings according to their various resources instead of by the birth order of sons (Lin and Yi, 2013).

See also: Buddhism; Cultural Relativism, Anthropology of; Ethnography; Individualism versus Collectivism: Philosophical Aspects; Law and Morality: An Analytical Perspective; Monasticism: West and East; Morality: Evolution of; Neural Foundation of Morality; Pluralism and Tolerance; Relativism: Philosophical Aspects; Religion: Morality and Social Control; Religions of East Asia; Religions of India; Value Pluralism.

Bibliography

Ahmed, L., 1993. Women and Gender in Islam: Historical Roots of a Modern Debate. Yale University Press, New Haven, CT.
Ali, K., 2006. Sexual Ethics and Islam: Feminist Reflections on Qur'an, Hadith and Jurisprudence. Oneworld, Oxford.
Buck-Morss, S., 1975. Socioeconomic bias in Piaget's theory and its implications for the cross-cultural controversy. Human Development 18, 35–49.
Carrithers, M., Collins, S., Lukes, S. (Eds.), 1985. The Category of the Person: Anthropology, Philosophy, History. Cambridge University Press, Cambridge.
Dumont, L., 1986. Essays on Individualism: Modern Ideology in Anthropological Perspective. University of Chicago Press, Chicago.
Dworkin, R., 1977. Taking Rights Seriously. Harvard University Press, Cambridge, MA.
Edel, M., Edel, A., 1959. Anthropology and Ethics. Charles C. Thomas, Springfield, IL.
Edwards, C., 1985. Cross-cultural research on Kohlberg's stages: the basis for consensus. In: Modgil, S., Modgil, C. (Eds.), Lawrence Kohlberg: Consensus and Controversy. Falmer, Sussex, England.
Erickson, K.C., Donald, D.S., 1997. Doing Team Ethnography: Warning and Advice. Sage, Beverly Hills, CA.
Geertz, C., 1974. From the native's point of view: on the nature of anthropological understanding. In: Basso, K., Selby, H. (Eds.), Meaning in Anthropology, first ed. University of New Mexico Press, Albuquerque, NM.
Gibbs, J.C., 1979. Kohlberg's moral stage theory: a Piagetian revision. Human Development 22, 89–112.
Gilligan, C., 1982. In a Different Voice: Psychological Theory and Woman's Development. Harvard University Press, Cambridge, MA.
Gubrium, J.F., Holstein, J.A., 1997. Analyzing Narrative Reality. Sage, Thousand Oaks, CA.
Hcelas, P., Lock, A. (Eds.), 1981. Indigenous Psychologies: The Anthropology of the Self. Oxford University Press, London.
Heimann, B., 1964. Facets of Indian Thought. Schocken Books, New York.
Howell, S., 1997. The Ethnography of Moralities. Routledge, New York.
Hwang, K.K., 1999. Filial piety and loyalty: two types of social identification in Confucianism. Asian Journal of Social Psychology 2 (1), 163–183.
Hwang, K.K., 2012. Foundations of Chinese Psychology: Confucian Social Relations. Springer, New York.
Kohlberg, L., 1971. From is to ought: how to commit the naturalistic fallacy and get away with it in the study of moral development. In: Mischel, T. (Ed.), Cognitive Development and Epistemology. Academy Press, New York, pp. 151–235.
Kohlberg, L., 1981. Essay on Moral Development, Vol. 1: The Philosophy of Moral Development, first ed. Harper & Row, San Francisco, CA.
Kung, H., Kuschel, K.-J. (Eds.), 1993. A Global Ethics: The Declaration of the Parliament of the Parliament of the World Religions. SCM Press, London.
Lin, J.P., Yi, C.C., 2013. A comparative analysis of intergenerational relations in East Asia. International Sociology 28, 297–315.
Maqsud, M., 1979. Cultural influences on transition in the development of moral reasoning in Nigerian boys. The Journal of Social Psychology 108, 151–159.
Marchand-Jodoin, L., Samson, L., 1982. Kohlberg's theory applied to the moral and sexual development of adults. Journal of Moral Education 11, 247–258.
Melhuus, M., 1997. the troubles of virtue: values of violence and suffering in a Mexican context. In: Howell, S. (Ed.), The Ethnography of Moralities. Routledge, London, England.

Miller, J.G., 1994. Cultural diversity in the morality of caring: individually oriented versus duty-based interpersonal moral codes. Cross-Cultural Research 28 (1), 3–39.

Mohmood, S., 2005. Politics of Piety: The Islamic Revival and the Feminist Subject. Princeton University Press, Princeton, NJ.

Parikh, B., 1980. Development of moral judgment and its relation to family environmental factors in Indian and American families. Child Development 51, 1030–1039.

Rapport, N., 1997. The morality of locality: on the absolutism of landownership in an English village. In: Howell, S. (Ed.), The Ethnography of Moralities. Routledge, London, England.

Shweder, R.A., Goodnow, J., Hatano, G., LeVine, R., Markus, H., Miller, P., 1998. The cultural psychology of development: one mind, many mentalities. In: Damon, W. (Ed.), Handbook of Child Psychology: Vol. 1. Theoretical Models of Human Development. J. Wiley, New York, pp. 865–937.

Shweder, R.A., Much, N.C., Mahapatra, M., Park, L., 1997. The big three of morality (autonomy, community, divinity) and the big three explanations of suffering. In: Brandt, A.M., Rozin, P. (Eds.), Morality and Health. Routledge, New York, pp. 119–169.

Snarey, J.R., 1985. Cross-cultural university of social-moral development: a critical review of Kohlbergian research. Psychological Bulletin 97 (2), 202–232.

Stacey, J., 2011. Unhitched: Love, Marriage, and Family Values from West Hollywood to Western China. New York University Press, New York.

Sullivan, E., 1977. A study of Kohlberg's structural theory of moral development: a critique. Human Development 20, 352–376.

Tsutsui, J., 2013. The transitional phase of mate selection in East Asian countries. International Sociology 28, 257–276.

Turner, B.S., 1988. Individualism, capitalism and the dominant culture: a note on the debate. Journal of Sociology 24, 47–64.

Weinreich-Haste, H., Locke, D. (Eds.), 1983. Morality in the Making: through, Action, and the Social Context. Wiley, Chichester, UK.

White, C.B., Bushnell, N., Regnemer, J.L., 1978. Moral development in Bahamian school children: a 3-year examination of Kohlberg's stages of moral development. Developmental Psychology 14, 58–65.

Yeh, K.H., Yi, C.C., Tsao, W.C., Wan, P.S., 2013. Filial piety in contemporary Chinese societies: a comparative study of Taiwan, Hong Kong, and China. International Sociology 28, 277–296.

Morality: Evolution of

Steve Stewart-Williams, Swansea University, Swansea, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article looks at the origins of human morality, and examines the contributions of both biological evolution and cultural evolution. Biological evolution may help explain certain categories of moral behavior (such as altruism), as well as certain morally relevant capacities (such as empathy). Cultural evolution may explain how different moral systems emerge in different times and places, and why they tend to foster aspects of human nature that enhance group success and discourage aspects of human nature that undermine it. Morality must be understood as the joint production of both biological and cultural evolution.

> The ant, the bird, the marmot… have read neither Kant nor the fathers of the Church nor even Moses. The idea of good and evil has thus nothing to do with religion or a mystic conscience. It is a natural need of animal races. And when founders of religions, philosophers, and moralists tell us of divine or metaphysical entities, they are only recasting what each ant, each sparrow practices in its little society.
>
> —Peter Kropotkin (1898), *Anarchist Morality*.

Morality Is a Mess

This might seem like a strange place to begin a discussion of the origins of morality, but let me explain why I think it's a useful starting point. Scientists strive for simple, parsimonious explanations for phenomena. Those that meet this standard are described approvingly as *elegant* explanations. It seems unlikely, however, that morality will ever have an elegant explanation, for the simple reason that morality is not an elegant phenomenon. Like many human institutions, it is a messy aggregation of a vast array of competing influences. These influences include inclinations rooted in the biology of human nature, such as a tendency to favor kin, a tendency to resent freeloaders, and a capacity to empathize with the suffering of others. They also include influences that can be classified broadly as cultural. Among these are norms designed to rein in socially disruptive aspects of human nature, norms aimed at furthering the interests of the individuals or groups promoting them, and – at least occasionally – genuine efforts to work out the logical consequences of universal principles of justice and the common good. If we focus on any of these influences to the exclusion of the others, we will arrive at a lopsided view of the nature of morality – hence the importance of keeping in mind that morality is a mess.

In this article, I consider both the biological and cultural evolution of our formal moral systems. I begin with the evolution of altruistic behavior, which is the area of moral behavior that evolutionary theory has shed the most light on. I then consider some of the putative evolved mechanisms underlying human morality, such as inhibitions against harming innocents, the capacity for empathy, and an aversion to incestuous mating. Finally, I consider the cultural evolution of morality and why some moral systems thrive and persist whereas others fade away.

The Problem of Altruism

In 2011, the Australian state of Queensland suffered extreme flooding. As with any disaster, this one left many tales of heroism in its wake. Among the most poignant is the story of 13-year-old Jordan Rice. Jordan had been shopping with his mother and his younger brother, Blake. As they drove home, they were caught in a flash flood. They ended up on the roof of their car, stranded precariously in the middle of raging waters. Eventually, help arrived. A local man tied a rope to himself and made his way to the car to try to rescue the family. Jordan insisted that the rescuer save his little brother first. Before the rescuer could return, Jordan and his mother were swept away and drowned.

By putting his brother ahead of himself, Jordan lost his life. This act of selflessness represents the height of moral action. But from a simplistic Darwinian perspective, it is puzzling. At first glance, evolutionary theory seems to imply that the organisms that survive and leave behind the most descendents will be those that look after number one. That's exactly what Jordan Rice didn't do. How could natural selection have spared a tendency for such self-sacrificial behavior? Of course, we might wonder whether natural selection is relevant to the question; perhaps Jordan was simply brought up in a culture that prizes self-sacrifice. But that is probably not the whole story, because comparable acts can be found in other animals. For instance, birds in some species feign injury to lure predators away from their babies in the nest, only attempting a getaway at the very last minute. Likewise, bees sting invaders to the nest, protecting the colony but losing their lives in the process. Self-sacrificial altruism is surprisingly common in nature. What might be going on?

Kin Selection

Over the past half-century, evolutionary biologists have made great strides in answering this question (Alexander, 1987; de Waal, 2006; Krebs, 2011). The single-most important theory of animal altruism – and the one most relevant to the examples given above – is William Hamilton's (1964) *kin selection theory* (*see* Kin Selection; Kinship, Evolution of). According to standard neo-Darwinian theory, a gene can be selected if it has effects that boost its replication rate relative to rival versions of the same gene (Dawkins, 1976). This can happen in a number of circumstances. Most obviously, it can happen if the gene increases its bearer's chances of surviving and reproducing. But it can also happen if the gene increases *another* organism's chances of surviving and reproducing, as

long as the other organism is more likely than chance to possess the same gene. Close relatives are more likely than chance to possess any given gene, and thus a tendency to help close relatives can potentially be selected. This process is called *kin selection*. Kin selection is the explanation for feigned injury in birds: The heroic bird's babies in the nest are more likely than chance to carry any genes contributing to the development of injury feigning. Kin selection is also the explanation for the suicidal nest defense of bees and other eusocial insects: Their close kin in the nest are more likely than chance to carry any genes contributing to the nest defense tendency. And kin selection may be an important part of the explanation for Jordan's sacrifice: His brother was more likely than chance to possess genes contributing to whatever motivations and inclinations underpinned his selfless behavior. Of course, shared genes were the last thing on Jordan's mind; he was merely acting on his inclinations. But kin selection is plausibly part of the explanation for how those inclinations evolved in the first place. Needless to say, the relationship between kin selection and morality is far from simple. After all, moral systems tend to oppose certain forms of kin altruism: those we label 'nepotism.' Nonetheless, kin selection helps to explain a class of behavior that we typically view as moral: altruism among kin.

Reciprocal Altruism

Hamilton's theory explains most of the altruistic behavior seen in other animals. However, it does not explain most of the altruistic behavior seen in human beings, because much of this involves nonrelatives. Certainly, the kind of extreme altruism exemplified by Jordan Rice may be more common among kin than nonkin. But more common, less costly forms of altruism in our species often take place between unrelated people (Stewart-Williams, 2007). Nonkin altruism is hard for the evolutionist to explain, but there are various possibilities. The best known is Robert Trivers' (1971) *reciprocal altruism theory* (*see* Sociobiology: Overview). This starts from the fact that individuals working together can often produce more than the sum total of what each could produce working alone. If five hunters working together bring down one giraffe, for example, they end up with more meat per hunter than if each working alone brings down one turtle. Thus, cooperation is better for all involved. Reciprocal altruism utilizes this principle but smears the cooperation out across time: I help you when I can; you help me when you can. This sounds like a good arrangement, so why don't all animals do it? The main stumbling block is *the problem of cheating*. Although it is better to trade favors than to act selfishly, it is better still to reap the benefits of someone else's help but then not return the favor. Because cheating pays more than mutual helping, mutual helping is evolutionarily unstable. It is displaced easily by a tendency to exploit helpers. The only way that mutual helping could evolve is if it coevolves with psychological 'machinery' that somehow defends against cheats. According to Trivers, many of our standard-issue emotional responses do exactly this. People unconsciously keep track of favors given and received, and when a trading partner (e.g., a friend or ally) starts taking advantage, people naturally respond with anger and moral indignation. This motivates them to punish the cheat or to cut off further aid, which protects them from exploitation. These and other psychological dispositions can be viewed as an evolved moral psychology (see also Cosmides and Tooby, 1992).

Trivers' theory has been tested in various ways. Most notably, it has been explored *in silico* – that is, using computer simulations (Axelrod, 1984; Axelrod and Hamilton, 1981; Nowak and Sigmund, 1992). In a typical simulation, virtual organisms encounter one another repeatedly and must either cooperate or defect (i.e., cheat). Points are awarded based on how well they do in each encounter. The virtual organisms are programed with various strategies, from very simple ones (such as Always Cooperate or Always Defect) to very complex, conditional ones. Among the most successful strategies, and by far the most famous, is Tit-for-Tat. Tit-for-Tat works like this. The first time it encounters a new player, it cooperates. In every subsequent encounter, it simply does whatever the other player did the last time they met. If the other player cooperated, Tit-for-Tat cooperates, thereby continuing a mutually beneficial cooperative partnership. If the other player defected, Tit-for-Tat defects, thereby cutting off the option of further exploitation. Tit-for-Tat is not always the most successful strategy; it all depends on who it's up against. But the general picture emerging from this research is that strategies *like* Tit-for-Tat – strategies that cooperate conditionally but defend against cheats – tend to do better than overly nice or overly nasty strategies.

Reciprocal altruism appears to be relatively rare among nonhuman animals, although there is evidence for it in vervet monkeys (Seyfarth and Cheney, 1984) and vampire bats (Wilkinson, 1984), among others. In contrast, many lines of evidence suggest that reciprocal altruism is an important part of the explanation for nonkin altruism among humans (see, e.g., Delton et al., 2011; Gurven et al., 2000; Hames, 1987).

Sexual Selection

Reciprocal altruism is not the only contender, however. A more recent attempt to explain our altruistic streak is based on an older theory: sexual selection theory (Darwin, 1871). Sexual selection is a form of natural selection. It occurs when a trait is selected, not because it enhances its bearer's survival prospects, but because it increases its bearer's *reproductive* prospects – that is, its ability to attract mates or outcompete same-sex rivals for mates. According to sexual selection theory, the mate choices of one sex can give rise to elaborate sexual ornamentation in the other. The enormous nose of the male proboscis monkey is a striking example. Females prefer to mate with males with larger-than-average noses. This preference, acting over many generations, gives rise to the male's impressive proboscis. In humans, not only the females but the males as well possess 'ornaments' that the other sex finds attractive (Stewart-Williams and Thomas, 2013). This includes secondary sexual characteristics such as women's hour-glass shape and breasts, and men's V-shaped torso and deep voice (Barber, 1995). According to some evolutionary psychologists, sexual selection has also produced behavioral ornaments, and among these are altruism and morality (Miller, 2007; Zahavi, 1975). These traits are costly, hard-to-fake displays of fitness, designed by selection to attract mates (*see* Human Mate Choice, Evolution of). Geoffrey Miller has made

the strongest case for altruism and morality as sexually selected ornamentation. The following quote is from his book *The Mating Mind*:

> Murder, unkindness, rape, rudeness, failure to help the injured, fraud, racism, war crime, driving on the wrong side of the road, failing to leave a tip in a restaurant, and cheating at sports. What do they have in common? A moral philosopher might say they are all examples of immoral behavior. But they are also things we would not normally brag about on a first date. (Miller, 2000: p. 292)

Several lines of evidence support the sexual selection theory of altruism. First, just as female proboscis monkeys prefer males with enormous noses, men and women tend to prefer mates who display kindness and generosity (Buss, 1989). Second, men are particularly likely to provide help in circumstances where doing so might demonstrate their fitness and enhance their mating prospects. So, for example, men are more likely than women to help when helping is dangerous and an audience is present, especially when the recipient of the help is a woman (Eagly and Crowley, 1986; Johnson, 1996). In a similar vein, Hawkes (1991) has made a strong case that big-game hunting is at least partly a matter of men showing off to obtain mates. The spoils of the hunt are not channeled preferentially to kin or used to pay off earlier debts; instead, the meat is shared widely throughout the group. What explains this altruistic sharing? According to Hawkes, the best hunters tend to have higher status and more wives and more affairs than other men – a finding that fits well with the sexual selection hypothesis. Certainly, sexual selection could not be the whole explanation for human altruism. It does not apply to altruism among kin, for instance, as kin are not potential mates (except in rare and unfortunate cases). But there is enough evidence now to think that sexual selection may be one piece of the puzzle.

Group Selection

The final theory of altruism we'll consider is also the most controversial. This is the idea that altruism is not a product of selection among individuals, but rather of selection among groups. The concept of group selection traces back to Darwin (1871) who, in *The Descent of Man*, suggested that groups of altruists would fare better than groups of selfish individuals, and thus that altruism would slowly displace selfishness. The subject was brought to center stage nearly a century later, when V.C. Wynne-Edwards (1962) argued that organisms in many species had evolved to sacrifice themselves for the good of the group. The main impact of Wynne-Edwards' work was to inspire a strong counterargument. Critics like G.C. Williams (1966) pointed out that any organisms that sacrificed themselves for the group would quickly be displaced by those that sat back and let others make the sacrifice while they pumped out as many offspring as possible. Although the critics conceded that group selection was possible in principle, most concluded that it was unlikely in practice. During the 1960s and 1970s, the idea that adaptations evolve for the good of the group was largely expunged from evolutionary biology, and replaced with the idea that adaptations evolve for the good of individuals and their kin (Dawkins, 1976; Williams, 1966). After several decades in the wilderness, however, a more sophisticated form of group selection is now gaining attention (Nowak, 2006; Sober and Wilson, 1998; Wilson, 2012). Modern group selectionists do not claim that group selection is the only selective force, and thus the approach is probably better named *multilevel selection theory*. Advocates of this theory (e.g., Wilson and Sober, 1994) distinguish between within-group selection and between-group selection. Within-group selection favors traits that advantage the individual, but between-group selection favors traits that advantage the group. Organisms, it is claimed, are the net outcome of these countervailing forces. When there are small differences in reproductive success within groups, but big differences in average reproductive success between groups, selection is concentrated at the level of the group. According to the multilevel selectionists, this is precisely what has happened in our species. As such, we have an evolved tendency to extend help not only to relatives but to nonrelatives in our group, even when they cannot reciprocate, and even when there is no scope for sexual or reputational benefits (a phenomenon known as *strong reciprocity*; Gintis, 2000).

Although group selection is currently enjoying a new lease of life, so too are criticisms of the theory. Some critics charge that group selectionists are too willing to attribute any instance of human sociality and cooperation to group selection (Pinker, 2012). Humans are certainly prosocial animals, but prosociality is not necessarily selected for the good of the group. As the theories described above show, it can be selected instead for the good of prosocial individuals and their kin. And there are several reasons to favor individual-level selection over group-level selection as the explanation. First, the individual-level approach has been much more empirically fertile than group selection. Second, in many ways, humans do not look like we're strongly group selected (Pinker, 2012; Price, 2012). Parents spend many thousands of dollars each year on their children, but if they also give several hundred dollars to charity, this is considered uncommonly altruistic (Singer, 1981). Grandparents routinely look after their grandchildren without payment; nonrelated group members generally do not. People happily share with family and friends, but they have to be cajoled and coerced to pay their taxes. If humans were strongly group selected, why would we need tax collectors? Why would we need conscription? Why would we need ethical codes that admonish us to prioritize the interests of the group? Certainly, extreme altruism among nonrelatives is hard to explain in terms of individual-level benefits – consider, for instance, the soldier who throws himself on a grenade to save the rest of the group. But such acts are rare, and it is usually a mistake to invoke natural selection to explain rare acts. Of course, group selectionists have responses to all of these criticisms, and the issue remains as yet unresolved. But with persuasive arguments on both sides, the group selection debate is currently one of the most exciting debates within evolutionary behavioral science (*see* Human Cooperation, Evolution of).

A Mind Made for Morality

Until recently, altruism was the main area in which Darwinian theory has been used to explore moral psychology. But altruism

is only one component of morality, and over the past decade or so, evolution-minded thinkers have begun exploring others. In the process, they have started to shed light on some of the psychological mechanisms underpinning moral behavior. So, for instance, according to Jonathon Haidt's *Moral Foundations Theory* (Haidt and Joseph, 2004), the human mind comes factory equipped with a set of moral tastes, including a desire to reduce the suffering of others, a sense of fairness, loyalty to the in-group, emotions and attitudes guiding behavior toward people of lower versus higher status, and a desire for purity and the avoidance of contamination. Other thinkers have focused on moral emotions, such as guilt, shame, contempt, anger, disgust, compassion, and sympathy (de Waal, 1996; DeScioli and Kurzban, 2009; Rozin et al., 1999). In this section, I discuss four putative psychological mechanisms underpinning human morality – namely, inhibitions about harming other people, the capacity for empathy, a sense of fairness, and an aversion to incest.

Trolleyology

Arguably, the most famous contemporary research in moral psychology is based on a hypothetical moral dilemma imported from philosophical ethics. The dilemma – which has been described as the white lab rat or the fruit fly of moral psychology – involves an out-of-control trolley (i.e., train) careening down the track and heading toward five workers on the line. You are standing there watching, and you realize that if the trolley continues on its present path, it will kill the five workers. In one version of the dilemma, you just so happen to be standing next to a lever. If you pull the lever, it will divert the trolley to another line, saving the five. Unfortunately, there is one person on the other line, and that person will be killed instead. What do you do? If you're like most people, you divert the trolley, killing one person but saving five (Hauser et al., 2007). But now imagine a second scenario. You are standing on a footbridge over a railway track, watching the trolley hurtle toward the five workers. This time you cannot divert it. Standing next to you, however, is a fat man (or, if you prefer, a man with a heavy backpack). Somehow you calculate that, if you push the man onto the track, his weight will be sufficient to stop the trolley in its tracks before it kills the five – but the man will die in the process. You are not heavy enough to stop the trolley, so you cannot make the ultimate Wynn-Edwardsian sacrifice and throw yourself on the track instead. What would you do in this situation? If you're like most people, you would *not* push the man (Hauser et al., 2007). But from a utilitarian perspective (according to which the right action is the one that produces the greatest good for all involved), the two dilemmas are morally identical: Either one person dies or five die. Why do we respond so differently?

Joshua Greene and colleagues (2004) have argued that people have a built-in aversion to directly and intentionally harming innocent people, and that, as a result, they are willing to pull the lever but not push the man. It is a matter of debate whether such an aversion is a product of natural selection as opposed to learning or culture. But the fact that the pattern is so widespread across cultures has led many to the former view. Interestingly, although people's decisions do not differ a great deal across cultures, their *explanations* for them do (Hauser et al., 2007). This suggests that, to some extent, moral training simply furnishes us with rationalizations for an unreasoned aversion that we would possess anyway (see Haidt, 2001).

Empathy

Closely related to this is the capacity for empathy. Empathy involves understanding what others are experiencing and having the appropriate emotional responses to it – for example, sympathizing with others' suffering (Baron-Cohen, 2011). Although culture and experience may amplify or diminish our empathetic tendencies, there are several reasons to suppose that the basic capacity for empathy is a product of natural selection. For one thing, the first flowerings of empathy appear very early in the life span (Warneken and Tomasello, 2009). As Paul Bloom has observed, toddlers often become upset if a nearby adult is upset, and they may even offer up a teddy bear or favorite blanket to make the adult feel better. In addition, some nonhuman animals may be capable of empathy. The strongest case has been made for large, intelligent animals like chimpanzees and dogs (de Waal, 1996). However, some researchers have argued that even mammals as diminutive as rats have some rudimentary capacity for empathy (Bartal et al., 2011). Patricia Churchland (2011) suggests that empathy originally evolved to allow mammalian mothers to attend to the needs of their offspring. In some lineages, however, natural selection later co-opted the neural mechanisms underpinning maternal empathy and applied them to other social relationships, including relationships with mates, unrelated allies, and ingroup members.

Although empathy is an important ingredient for morality, by itself it is not sufficient for morality. Indeed, as various commentators have pointed out, empathy can sometimes be morally misleading (e.g., Bloom, 2013). We easily empathize with our friends and neighbors' smallest hurts but may fail to empathize with the millions of people starving in the developing world. Thus, doing the right thing sometimes involves going against the advice of our empathy and listening instead to the voice of reason. But although empathy is not sufficient for morality, it is hard to imagine that someone who lacked empathy entirely could be truly moral (Prinz, 2005). Empathy is an important building block of human morality, and one that plausibly has its origin in our evolutionary history.

A Fairness Instinct?

A more contentious claim is that human beings have a built-in sense of fairness – a 'fairness instinct,' as David Barash put it. The most common way of exploring this issue is a decision task known as the Ultimatum Game. The game involves two players: the Proposer and the Responder. The Proposer is given a sum of money – say, $10. Her task is to propose how that money should be split between herself and the Responder. The Proposer could suggest a 50–50 split, or a 90–10 split, or any other combination. Then it's the Responder's turn. The Responder has only two options: accept the proposal or reject it. If he accepts it, they both get the proposed sums of money. If he rejects it, no one gets anything. In a world where people were perfectly rational and entirely self-interested, Proposers would

elect to keep $9 and give away only $1, and Responders would accept any nonzero offer – why opt for nothing when you could have $1? But that is not what usually happens. In Western nations, Proposers usually offer $4–5, and Responders usually reject anything less than $3. The Proposers are not acting irrationally; they know that if they make a lowball offer, the Responder will resent it and reject it, so they make an offer the Responder will accept. The Responders, in contrast, *are* acting irrationally, because their insistence on punishing unfair offers leaves them out of pocket. That, at any rate, is the verdict of classical economics. But closer inspection suggests how this tendency could be adaptive. For most of our evolutionary history, humans rarely encountered people they would never encounter again. A willingness to respond angrily to unfairness – in this case, to reject a stingy offer even at some personal cost – signals to the world that you are not the sort of person who will accept mistreatment. Although this may be self-destructive in the short term, it may be beneficial in the longer term, as it means that people will think twice about mistreating you next time around.

Thus, the Ultimatum Game findings have been used as evidence for an evolved distaste for unfairness. The problem is, though, that the results described so far come exclusively from Western nations. When Joseph Henrich and colleagues (2005) took the Ultimatum Game to 15 small-scale societies, they found that people in these societies behaved very differently than Westerners. Among the Machiguenga horticulturalists of Peru, for instance, Proposers often made low offers – and Responders often accepted them. Even more surprising, among the Au and Gnau of Papua New Guinea, Proposers often made *high* offers – offers of up to 70% of the pot – and Responders turned them down! Henrich concluded that the alleged fairness instinct is actually a set of cultural norms: norms that grow up in cultures that are embedded in a market economy and in which people must cooperate closely to eke out a living.

The issue is further complicated by the fact that other researchers have found evidence for an incipient sense of fairness in nonhuman animals. Brosnan and de Waal (2003) trained capuchin monkeys to exchange pebbles for food. Sometimes the capuchins received pieces of cucumbers for their pebbles; sometimes they received grapes. The monkeys preferred the grapes, but they were quite happy to receive the cucumber – as long as their neighbor was only getting cucumber, too. However, if the neighbor started getting grapes, while they continued to get cucumber, the capuchins threw a tantrum. Indeed, they often threw the cucumber away forcefully. The monkeys were not perturbed by inequality per se. Some collected more stones than others and therefore earned more treats, but onlookers did not throw tantrums about that. Instead, they were upset that another individual got a better reward for the same behavior. In other words, they didn't want an equal distribution of wealth; they wanted equal pay for equal work. The capuchin sense of fairness is a far cry from our own. For one thing, capuchins only object to unfairness when they are on the receiving end of it (although see Brosnan et al., 2010). Nonetheless, the fact that these less culture-bound creatures possess anything like a sense of fairness makes it harder to believe that our own sense of fairness is purely a cultural invention. Although it is not yet clear how to reconcile the capuchin findings with the cross-cultural Ultimatum Game findings, the solution may relate to the fact that the term 'fairness' is being used in subtly different ways in the different studies. In the Ultimatum Game, it means equality (i.e., a 50–50 split); in the capuchin research, it means equity (i.e., equal pay for equal work). It may be that, rather than having an evolved aversion to deviations from a 50–50 split, humans have an evolved aversion to unfairness or mistreatment more generally, with some flexibility in what exactly this means from culture to culture (*see* Human Cooperation, Evolution of).

Incest Aversion

There may also be biological contributions to some aspects of our sexual morality. The best example concerns incest. In many species, our own included, inbreeding is highly detrimental to fitness. The offspring of incestuous liaisons often do not survive, and when they do, they have an elevated risk of genetic abnormalities. It is plausible to suppose, then, that we possess psychological adaptations that tend to steer us away from incestuous mating. Given the strong evolutionary rationale for such adaptations, it is probably no coincidence that most people feel acute disgust at the idea of sex with siblings or other family members. A powerful case has been made that humans have an evolved tendency to develop a sexual aversion to individuals they grow up with, especially during the first five or so years of life (Fessler and Navarrete, 2004; Lieberman et al., 2007). This negative sexual imprinting is known as the *Westermarck effect* (Westermarck, 1921; Wolf, 1995). Now it might be argued that incest aversion is a cultural invention: Perhaps earlier peoples observed that incestuous mating led to birth defects and the like, and therefore proscribed it. But this does not fit the facts. Societies have proscribed many forms of sexuality over the ages: adultery, premarital sex, and so on. Sometimes the proscription works; sometimes it does not. But when it does, it is usually because people manage to resist temptation, not because they're disgusted by the act in question. In contrast, most people are profoundly disgusted by the idea of incestuous mating, and they feel this way despite receiving little socialization against incest: Parents generally worry about their teenagers having sex with nonrelatives, not with siblings. It is plausible to suppose, then, that an aversion to incest is a product of natural selection. The moral precept 'Incest is wrong' is presumably *not* a direct product of natural selection. However, it may be an idea that we readily accept, given the nature of human nature (*see* Incest Prohibition, Origin and Evolution of).

The Cultural Evolution of Morality

This brings us to the topic of the cultural evolution of morality. Although some suggest that morality is a product of biological evolution – an adaptation like eyes and wings – there are several reasons to view it instead as the joint product of both biological and cultural evolution (Mesoudi and Danielson, 2008; Stewart-Williams, 2010). First, moral codes vary tremendously across cultures and times (Prinz, 2005). In modern Western societies, most people deem cannibalism, slavery, and torture to be self-evidently wrong; in earlier times, however, and across a wide range of cultures, they were

deemed morally permissible. Similarly, during the twentieth century, the West underwent a profound shift in modal attitudes to such issues as homosexuality, premarital sex, and the treatment of ethnic minorities. Certainly, there are still differences of opinion within Western culture: People disagree bitterly about the morality of abortion, capital punishment, casual sex, drug use, and euthanasia. But this too is difficult to square with the notion that morality is a direct product of biological evolution (although see DeScioli and Kurzban, 2009). A second reason to look beyond biology is the fact that morality often goes against the grain of our raw biological impulses. For instance, the principle that we should 'turn the other cheek' goes against the natural impulse to harm those who harm us. If turning the other cheek were biologically adaptive for our ancestors, we would presumably possess an evolved appetite for it. This is not to deny that, at least in some cases, morality simply reflects our natural inclinations; think, for instance, of our nepotistic streak and our aversion to incest. But often morality looks more like a set of cultural norms designed to put the brakes on socially divisive inclinations. A final reason to doubt that morality is a biological adaptation in any simple sense is the fact that we strive to teach our children to be moral. Although young children spontaneously help others in a way that even adult chimpanzees do not, parents and teachers still spend a lot of time encouraging children to share and not to hurt their playmates. They might not create children's prosocial tendencies out of nothing, but they clearly try to amplify them. Once again, this suggests that the norms and ideals we bunch together under the heading 'morality' must be understood as the joint production of both biology and culture.

In this final section, I examine three theories about how moral systems might evolve culturally: (1) the memetic approach, according to which moral ideas come to predominate in a group to the extent that they are good at replicating themselves; (2) cultural group selection, according to which moral ideas come to predominate in a group to the extent that they promote the group's interests; and (3) gene-culture coevolution, according to which cultural changes in moral ideas (or indeed any aspect of culture) can alter biological selection pressures and lead to downstream genetic change.

Moral Memes

The first approach to cultural evolution is the memetics approach. Richard Dawkins (1976) coined the term 'meme' to refer to a unit of culture: an idea, a word, a ritual, a tune. He then argued that, just as natural selection among genes favors genes that are good at replicating themselves, so too natural selection among memes favors memes that are good at replicating themselves. A central insight of the theory is that the memes that come to predominate in a culture are not necessarily those that are good for us. They *may* sometimes be good for us, but that's not the reason they are selected. Memes are selected because they are good for the memes themselves, where goodness is defined in terms of their ability to spread from mind to mind and to persist in a culture. Applied to morality, the claim would be that moral memes are selected not because they enhance reproductive fitness, or because they benefit individuals or groups, but instead because they are good at selfishly replicating themselves. The memetic approach does not necessarily conflict with the earlier material. Indeed, the earlier material may shed light on why some moral memes succeed whereas others fall out of the meme pool. Memes must adapt to relatively fixed aspects of human nature. So, for instance, a moral system that prohibited any form of nepotism – including the tendency to feed and clothe one's own offspring rather than the next-door neighbor's offspring – would be unlikely to persist in the meme pool, as it would clash too violently with our evolved kinship psychology. Moral memes must also adapt to the current cultural setting. The fairness meme, for instance, seems to prosper better in market economies than in small-scale societies.

The memetic perspective may help explain some otherwise-puzzling features of morality. Consider, for instance, the first four of the Hebrew 10 Commandments: have no Gods before me; make no graven images; do not take God's name in vain; and keep the Sabbath holy. These have nothing to do with harm avoidance, or justice, or any other major moral principle. At a push, it could be argued that they foster social cohesion. But the memes'-eye view raises an alternative and often-overlooked possibility: These commandments help to protect the religious system that they are a part of. They may be viewed as part of a large cluster of memes – a memeplex – 'designed' by memetic selection to perpetuate itself. Critics of memetics note that, thus far, the approach has yielded little in the way of empirical findings. Nonetheless, it is an appealing approach, which alerts us to the possibility that at least some moral memes may exist not because they help us, but simply because they help themselves.

Cultural Group Selection

A second approach to cultural evolution goes by the name of *cultural group selection*. This is selection for the good of the group, but at the level of cultural elements rather than genes. Richerson and Boyd (2004) argue that group selection at the genetic level was probably weak in our ancestral past, as there was too much gene flow between groups. There was much less 'meme flow,' however. Even today, neighboring forager tribes are often genetically indistinguishable, but speak mutually unintelligible languages. If this kind of asymmetry was typical throughout our evolutionary history, there would have been much greater scope for cultural group selection than for genetic group selection. The sociologist Emile Durkheim long ago suggested that institutions such as morality function as social glue, binding together large-scale societies. Cultural group selection provides a possible mechanism for bringing about such institutions.

Many aspects of morality are potentially explicable in terms of cultural group selection. One example is the Western institution of monogamy. The most common mating pattern in ancestral societies was a mixture of monogamous and polygynous marriage (Marlowe, 2003). As Henrich and colleagues (2012) note, though, several centuries ago, European cultures established a norm of monogamous marriage. This norm applies even to high-status men who, in traditional societies, would often have taken several wives. The monogamy norm cannot be explained as a direct product of natural selection; after all, it is historically recent and, to some extent, pushes

against our evolved tendencies. How, then, did it come about? Henrich et al. suggest that groups that instituted monogamous norms had a number of advantages over those that did not. For a start, crime rates and rates of violence are generally lower in monogamous than polygynous societies. This is because polygynous societies are overrun with bachelors – if one man has four wives, three men must have none – and when young men have no wives or mating opportunities, they are more likely to cause problems (Daly and Wilson, 2001). In contrast, groups adhering to the monogamy norm have fewer such problems. These groups were therefore more productive and wealthier. They grew faster and persisted for longer. Furthermore, other groups, observing their success, may have copied them. In this way, the monogamy meme spread – despite the fact that it clashed to some degree with human nature (*see* Marriage Systems, Evolution of).

Cultural group selection is probably not a complete explanation for morality. As the philosopher David Hume (1777/1999) pointed out, certain virtues – those he called the "monkish" virtues: celibacy, fasting, penance, and the like – are not useful for society in any obvious way. But cultural group selection does help to explain a deep trend in our moral systems: the fact that they generally encourage aspects of human nature that facilitate social harmony, and discourage aspects of human nature that upset it.

Gene-Culture Coevolution

The final approach – gene-culture coevolutionary theory – is distinguished by its emphasis on the interactions between biological evolution and cultural evolution (Cavalli-Sforza and Feldman, 1981; Durham, 1991; Lumsden and Wilson, 1981) (*see* Gene-Culture Coevolution). The classic example of gene-culture coevolution concerns the capacity to digest lactose in milk even after the age of weaning. This is found almost exclusively in populations with a long history of dairying. One might suppose that the capacity to digest lactose came first and that the cultural practice (dairying) followed. It seems, however, that it was the other way around: Dairying came first and established a selection pressure that led to the evolution of life-long lactose digestion (Laland et al., 2010). A similar process may have shaped human morality. Richerson and Boyd (2004) suggest that gene-culture coevolution helps explain why humans are so much more altruistic and prosocial than other animals (see also Boehm, 2012). In their view, our hypertrophied altruism stems originally from cultural group selection (*see* Human Cooperation, Evolution of). Groups that established and enforced altruistic norms did better than other groups, lasting longer and growing more rapidly. Once the prosocial norms were in place, though, people whose genes fitted them to this social milieu prospered and had more offspring. In time, the relevant genes crowded out competing variants. Thus, biological evolution reinforced the cultural tendency and amplified our altruistic dispositions.

Conclusion

Morality is a complex phenomenon that resists any simple explanation. It is the net outcome of a tangle of conflicting causal influences. These include both biological and cultural influences. Biology affects morality in various ways. First, some forms of moral behavior, in particular altruism, have their origin in our evolutionary history. Second, formal moral systems sometimes reflect aspects of human nature. This includes, for example, the proscription on incest among biological relatives. Third, certain evolved capacities provide the basic building blocks for morality. This includes the capacity for empathy and emotions like anger, disgust, and guilt. And fourth, biology may constrain our moral systems: Moral systems that conflict too violently with human nature may be unlikely to persist.

In addition to these biological influences, cultural evolution may help to explain the moral systems that crop up in different places and times. Ultimately, moral memes must possess properties that enable them to stick around in the culture. In many cases, the property in question may be the fact that they enable groups holding those moral memes to prosper and persist. Finally, not only does biology shape morality but morality may shape biology. It may do this by setting up selection pressures that, over many generations, fit people to the demands of the local moral system.

See also: Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Gene–Culture Coevolution; Human Cooperation, Evolution of; Incest Prohibition, Origin and Evolution of; Incest, Inbreeding, and Their Consequences; Kin Selection; Kinship, Evolution of; Marriage Systems, Evolution of; Sociobiology: Overview.

Bibliography

Alexander, R.D., 1987. The Biology of Moral Systems. Aldine de Gruyter, Hawthorne, NY.

Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York.

Axelrod, R., Hamilton, W.D., 1981. The evolution of cooperation. Science 211, 1390–1396.

Barber, N., 1995. The evolutionary psychology of physical attractiveness: sexual selection and human morphology. Ethology and Sociobiology 16, 395–424.

Baron-Cohen, S., 2011. Zero Degrees of Empathy: A New Theory of Human Cruelty. Penguin, London.

Bartal, I.B.-A., Decety, J., Mason, P., 2011. Empathy and pro-social behavior in rats. Science 334, 1427–1430.

Bloom, P., 2013. The Baby in the Well: The Case against Empathy. The New Yorker. Retrieved 30 May, 2013 from: http://www.newyorker.com/. Webpage: http://www.newyorker.com/arts/critics/atlarge/2013/05/20/130520crat_atlarge_bloom.

Boehm, C., 2012. Moral Origins: The Evolution of Virtue, Altruism and Shame. Basic Books, New York.

Brosnan, S.F., de Waal, F.B.M., 2003. Monkeys reject unequal pay. Nature 425, 297–299.

Brosnan, S.F., Talbot, C., Ahlgren, M., Lambeth, S.P., Schapiro, S.J., 2010. Mechanisms underlying responses to inequitable outcomes in chimpanzees, *Pan troglodytes*. Animal Behaviour 79, 1229–1327.

Buss, D.M., 1989. Sex differences in human mate preferences: evolutionary hypotheses tested in 37 cultures. Behavioral and Brain Sciences 12, 1–49.

Cavalli-Sforza, L.L., Feldman, M.W., 1981. Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press, Princeton, NJ.

Churchland, P.S., 2011. Braintrust: What Neuroscience Tells Us about Morality. Princeton University Press, Princeton, NJ.

Cosmides, L., Tooby, J., 1992. Cognitive adaptations for social exchange. In: Barkow, J.H., Cosmides, L., Tooby, J. (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, New York, pp. 163–228.

Daly, M., Wilson, M., 2001. Risk-taking, intrasexual competition, and homicide. Nebraska Symposium on Motivation 47, 1–36.

Darwin, C., 1871. The Descent of Man and Selection in Relation to Sex. Murray, London.

Dawkins, R., 1976. The Selfish Gene. Oxford University Press, New York.

de Waal, F., 1996. Good Natured: The Origins of Right and Wrong in Humans and Other Animals. Harvard University Press, Cambridge, MA.

de Waal, F., 2006. Primates and Philosophers: How Morality Evolved. Princeton University Press, Princeton, NJ.

Delton, A.W., Krasnow, M.M., Cosmides, L., Tooby, J., 2011. Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. Proceedings of the National Academy of Sciences 108, 13335–13340.

DeScioli, P., Kurzban, R., 2009. Mysteries of morality. Cognition 112, 281–299.

Durham, W.H., 1991. Coevolution: Genes, Culture, and Human Diversity. Stanford University Press, Stanford, CA.

Eagly, A.H., Crowley, M., 1986. Gender and helping behavior: a meta-analytic review of the social psychological literature. Psychological Bulletin 100, 283–308.

Fessler, D.M.T., Navarrete, C.D., 2004. Third-party attitudes toward sibling incest: evidence for Westermarck's hypotheses. Evolution and Human Behavior 25, 277–294.

Gintis, H., 2000. Strong reciprocity and human sociality. Journal of Theoretical Biology 206, 169–179.

Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., Cohen, J.D., 2004. The neural bases of cognitive conflict and control in moral judgment. Neuron 44, 389–400.

Gurven, M., Hill, K., Kaplan, H., Hurtado, A., Lyles, R., 2000. Food transfers among Hiwi foragers of Venezuela: tests of reciprocity. Human Ecology 28, 171–218.

Haidt, J., 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological Review 108, 814–834.

Haidt, J., Joseph, C., 2004. Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. Daedalus 133, 55–66.

Hames, R.B., 1987. Relatedness and garden labor exchange among the Ye'kwana. Ethology and Sociobiology 8, 354–392.

Hamilton, W.D., 1964. The genetical evolution of social behaviour: I & II. Journal of Theoretical Biology 7, 1–52.

Hauser, M.D., Cushman, F., Young, L., Kang-Xing Jin, R., Mikhail, J., 2007. A dissociation between moral judgments and justifications. Mind and Language 22, 1–21.

Hawkes, K., 1991. Showing off: tests of an hypothesis about men's foraging goals. Ethology and Sociobiology 12, 29–54.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., et al., 2005. "Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. Behavioral and Brain Sciences 28, 795–855.

Henrich, J., Boyd, R., Richerson, P.J., 2012. The puzzle of monogamous marriage. Philosophical Transactions of the Royal Society B 367, 657–669.

Hume, D., 1777/1999. An Enquiry Concerning the Principles of Morals. Oxford University Press, Oxford.

Johnson, R.C., 1996. Attributes of Carnegie medalists performing acts of heroism and of the recipients of these acts. Ethology and Sociobiology 17, 355–362.

Kropotkin, P., 1898. Anarchist Morality. Free Society, San Francisco.

Krebs, D., 2011. The Origins of Morality: An Evolutionary Account. Oxford University Press, New York.

Laland, K.N., Odling-Smee, J., Myles, S., 2010. How culture shaped the human genome: bringing genetics and the human sciences together. Nature Reviews Genetics 11, 137–148.

Lieberman, D., Tooby, J., Cosmides, L., 2007. The architecture of human kin detection. Nature 445, 727–731.

Lumsden, C.J., Wilson, E.O., 1981. Genes, Mind, and Behavior: The Coevolutionary Process. Harvard University Press, Cambridge, MA.

Marlowe, F.W., 2003. The mating system of foragers in the standard cross-cultural sample. Cross-Cultural Research 37, 282–306.

Mesoudi, A., Danielson, P., 2008. Ethics, evolution and culture. Theory in Biosciences 127, 229–240.

Miller, G.F., 2000. The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature. Vintage, London.

Miller, G.F., 2007. Sexual selection for moral virtues. Quarterly Review of Biology 82, 97–125.

Nowak, M.A., 2006. Five rules for the evolution of cooperation. Science 314, 1560–1563.

Nowak, M.A., Sigmund, K., 1992. Tit for tat in heterogeneous populations. Nature 355, 250–253.

Pinker, S., 2012. The False Allure of Group Selection. Retrieved 20 May, 2013 from: http://www.edge.org/. Webpage: http://edge.org/conversation/the-false-allure-of-group-selection.

Price, M.E., 2012. Group selection theories are now more sophisticated, but are they more predictive? A review of Samuel Bowles and Herbert Gintis, a cooperative species: human reciprocity and its evolution. Evolutionary Psychology 10, 45–49.

Prinz, J.J., 2005. The Emotional Construction of Morals. Oxford University Press, Oxford.

Richerson, P.J., Boyd, R., 2004. Not by Genes Alone: How Culture Transformed Human Evolution. University of Chicago Press, Chicago.

Rozin, P., Lowery, L., Imada, S., Haidt, J., 1999. The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). Journal of Personality and Social Psychology 76, 574–586.

Seyfarth, R.M., Cheney, D.L., 1984. Grooming, alliances, and reciprocal altruism in vervet monkeys. Nature 308, 541–543.

Singer, P., 1981. The Expanding Circle: Ethics and Sociobiology. Farrar, Strauss, & Giroux, New York.

Sober, E., Wilson, D.S., 1998. Unto Others: The Evolution and Psychology of Unselfish Behavior. Harvard University Press, Cambridge, MA.

Stewart-Williams, S., 2007. Altruism among kin vs. nonkin: effects of cost of help and reciprocal exchange. Evolution and Human Behavior 28, 193–198.

Stewart-Williams, S., 2010. Darwin, God, and the Meaning of Life: How Evolutionary Theory Undermines Everything You Thought You Knew. Cambridge University Press, Cambridge.

Stewart-Williams, S., Thomas, A.G., 2013. The ape that thought it was a peacock: does evolutionary psychology exaggerate human sex differences? Psychological Inquiry 24, 137–168.

Trivers, R.L., 1971. The evolution of reciprocal altruism. Quarterly Review of Biology 46, 35–57.

Warneken, F., Tomasello, M., 2009. The roots of human altruism. British Journal of Psychology 100, 445–471.

Westermarck, E.A., 1921. The History of Human Marriage, fifth ed. Macmillan, London.

Wilkinson, G.S., 1984. Reciprocal food sharing in the vampire bat. Nature 308, 181–184.

Williams, G.C., 1966. Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought. Princeton University Press, Princeton, NJ.

Wilson, E.O., 2012. The Social Conquest of the Earth. Liveright, New York.

Wilson, D.S., Sober, E., 1994. Reintroducing group selection to the human behavioral sciences. Behavioral and Brain Sciences 17, 585–654.

Wolf, A.P., 1995. Sexual Attraction and Childhood Association: A Chinese Brief for Edward Westermarck. Stanford University Press, Stanford, CA.

Wynne-Edwards, V.C., 1962. Animal Dispersion in Relation to Social Behaviour. Oliver & Boyd, Edinburgh.

Zahavi, A., 1975. Mate selection: a selection for a handicap. Journal of Theoretical Biology 53, 205–214.

Morphology in Linguistics

Antonio Fábregas, University of Tromsø, Tromsø, Norway

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A. Spencer, volume 15, pp. 10047–10053, © 2001, Elsevier Ltd.

Abstract

Morphology, the study of the correspondences between grammatical information, meaning, and form inside words, is one of the central linguistic disciplines. In this short presentation we briefly review the units it identifies, the processes used to build new words and alter the grammatical properties of existing words, and some of the main underlying theoretical questions currently under debate: (1) whether morphology is independent from syntax; (2) what is the nature of the basic units of morphological analysis; (3) the division of labor between lexical listing and word formation; and (4) how idiosyncratic information that plays a role in word grammar is distributed across different linguistic components.

What Is Morphology?

Morphology is the area of linguistics that studies the grammatical information contained in WORDS, and the one that explores how words are related to each other according to their shared meaning and form. Take, as an example, the Spanish word in (1), where the accent marks the phonological position of stress, not the actual orthographic convention.

(1) Cantarémos
'we will sing'

For a morphologist, what is relevant about (1) is that it is a verb (whose infinitive is *cantar*, 'to sing'), specifically a first person plural form (that is, corresponding to *we*), in a particular tense (future) and a particular mood (indicative): something that the traditional grammars have called a simple future form. Linking these pieces of information to the form of the word can be done in various ways (see the section 'Morphemes or Words?' for details). One of them is to see if it is possible to correlate distinct pieces of information with segments inside the word. By comparing (1) to other forms of the same verb, some morphologists would conclude that:

1. *-mos* is associated with the information 'first person plural,' as *-mos* in (1) occupies a position that other segments fill when the verbs appear in other persons and numbers: e.g., *cantaré-mos* versus *cantaré-is*, 'you (pl.) will sing.'
2. In a first approximation, it can be concluded that stressed *-ré-* expresses future tense and indicative. We know that *-ré-* expresses indicative because the future form in the subjunctive (which is pretty unusual in contemporary Spanish) exhibits the unstressed segment *-re-* in its place: *canta-ré-mos* versus *cantá-re-mos*, 'that we will sing.' We know that *-ré-* expresses future because in the past form, it is substituted by *-ba-*: *canta-ré-mos* versus *cantá-ba-mos*, 'we sang.'
3. Also in a first approximation, canta- would be identified as a unit that carries the conceptual meaning of the word (that is, that tells us what kind of event or state we are talking about). We know this because it can be substituted by other verbs: *canta-ré-mos* versus *vivi-ré-mos* 'we will live' or *bebe-ré-mos* 'we will drink.'

Consequently, the word could be decomposed – SEGMENTED is the more technical term – as in (2), where each segment expresses different pieces of information, pending the examination of further contrasts which would refine the analysis and which we will discuss shortly.

(2) canta-ré-mos
a. canta -> verb
b. ré -> future indicative
c. -mos -> first person plural

With complications explored in Section Morphemes or Words?, many morphologists would call these segments, paired with meaning and grammatical information, MORPHEMES. In some theories, a central ingredient of morphological analysis is to provide the decomposition of words into morphemes, understood as the smallest units that, in themselves, carry meaning and grammatical information.

Given decompositions such as the one in (2), words can be divided in two groups: SIMPLE WORDS that consist of only one morpheme, and COMPLEX WORDS that contain two or more morphemes. (3) illustrates both classes with Spanish nouns.

(3) a. sal
'salt'
b. implica-ción
imply-ation
'implication'

Another aspect of morphology is to study how words that share some property are related to each other. Consider an English word like (4), a plural noun.

(4) paper-s

This word has at least two properties: it is a noun denoting a particular substance used to write on and it is plural; each one of these parts is associated with a morpheme.

A morphologist could compare this word to others that contain the information 'plural'; put more technically, to other words that have the LINGUISTIC FEATURE [plural]. This would imply comparing *paper-s* with *cat-s*, but also with a word like *women*,

where the morpheme *-s* is not present and instead a vocalic change has been used to mark [plural] (*woman* ~ *women*), a word like *fish*, where no apparent change has taken place (*fish* ~ *fish*), and perhaps even one like *people*, where apparently the expression of [plural] involved uses a different morpheme (*person* ~ *people*). Several questions would arise from this comparison, all of them relevant to morphology: (1) how many different ways does English have to express [plural] in nouns? (2) why do some nouns use one procedure, and others a different one? Are there principled reasons for this choice, or is it accidental and idiosyncratic? (3) what happens with nouns that, apparently, lack a plural form, like *thirst*? (4) and what about nouns which, in the plural, get a special meaning, and express different notions from those they express in the singular, like *savings*, understood as the money one has saved out of income? Ultimately, all these questions reduce to one: what are the conditions under which a noun can express the information [plural], and what are its effects?

Similarly, a comparison could be made on the basis of the first morpheme, *paper*, which would lead us to consider words such as *paper-work*, *wall-paper*, the verb *to paper* ('to cover something with paper') with its derived noun (*paper-ing*) and participle (*paper-ed*), etc. Here, other sets of questions would arise: (1) what are the kinds of words that we can build with the morpheme *paper*: nouns, adjectives, verbs…? (2) what is the meaning and form relation between these words? Why do we interpret *to paper* as 'to cover something with paper' and not, for instance, as 'to write something on paper'? (3) why do we need extra morphemes to make a verb out of a noun in, for instance, the pair *drama* ~ *dramat-ize* or *monopoly* ~ *monopol-ize*, but not on *paper* ~ *paper* or *chain* ~ *chain* 'to put a chain'? and (4) why is *-ing* used to build a noun from the verb *to paper*, and not *-ation* (*explan-ation*) or *-ment* (*govern-ment*), when other nouns coming from verbs use precisely these? Is this choice accidental or can it be captured by some generalization?

Given this basic comparison, the following is a nonexhaustive list of some crucial problems that morphology studies, all based on the observation that words relate to each other in meaning, function, and form:

1. Correspondence between category and grammatical information: given words of a class, what is the grammatical information that these words can express, and how is it organized? For instance: in a given language, what are the pieces of information that a verb carries, and how are they different from what a noun expresses? Do all languages differentiate nouns from verbs according to their morphology? Can a verb express aspect and not tense, or a pronoun person and not number, or gender but not number?
2. The shape of morphemes: why does a morpheme, sometimes, have to be pronounced differently and what is the relation between all these variants of a morpheme – called allomorphs? For instance: why does the morpheme *destroy* (*destroy-ed*, *destroy-ing*) alternate with *destruct-* (*destruct-ion*, *destruct-or*)? How do speakers determine which allomorph to use, and how are they distributed? How should we represent this alternation? Through a rule that produces one from the other, listing both morphemes as separate units, or from a common abstract morpheme that is pronounced differently under different conditions?
3. Derivation of meaning: how is meaning codified inside a word, how is it built from the internal components of that word, and how is it preserved, modified, or lost when other morphemes are added? For instance: how do we obtain the meaning 'that cannot be comprehended' from the morphemes *in-*, *comprehens-*, and *-ible* in *in-comprehens-ible*? Why can the verb *to seed* mean either 'to remove seeds' or 'to put seeds,' while *to nail* is only 'to put nails'? Why can we use *recital* to describe the event of *reciting poetry*, but not of *reciting the alphabet*?
4. Restrictions on Productivity: why are some morphemes unable to combine with many others, and how do speakers determine the kinds of words that can be built from a morpheme? For instance: why is it possible to build a noun denoting 'the person that does an action' with *run* (*runn-er*) or *murder* (*murder-er*), but not with *die* (**di(e)-er*)? Why can English make adjectives in *-able* from verbs like *analyze* (*analyzable*) or *break* (*breakable*), but not from *shine*, *exist*, *sleep*, or *sit*?
5. Morpheme selection: why are there different morphological procedures to express the same notion, as with [plural], and how do some words choose one, or several, of them? Why do we have *move-ment*, but not **mov(e)-ation*, while we allow both *govern-ment* and *govern-ation*, and what is exactly the difference between these two nouns that come from the same verb?

See, among many others, Spencer (1991), Harley (2006), Booij (2007), Lieber (2009), and Fábregas and Scalise (2012) for some recent introductions to morphology where more about these questions can be found.

Types of Morphemes

Morphemes are normally divided into two basic classes: affixes and roots. Roughly speaking, affixes are morphemes which have a fixed position inside the word and cannot form words by themselves or in combination with other affixes. Roots are morphemes with free position, which frequently can stand alone as words and can form a word with an affix. (5) is an example of an affix in English: it always appears to the right of a root and without a root does not form a word in English. (6) is a root. It can appear to either the left or the right of another root and it can form a whole word, a noun, without other apparent morphemes.

(5) -ish
  a. child-ish, *ish-child
  b. *This is too -ish.
  c. *sub-ish
(6) tree
  a. apple-tree, tree-house
  b. This is a tree.
  c. sub-tree

However, it is not always the case that a root corresponds to a word. In the Spanish verb *canta* (cf (1)) the root is *cant-*, which is not a word and which we obtain once we segment the affix *-a*, which marks that it is a verb belonging to a particular conjugation (as in *am-a* 'love' or *rob-a* 'steal'). In

languages whose morphology has affixes that mark the morphological class of the root, the smallest unit that can always appear alone as a word is the STEM, which is the combination of the root with the affix that marks its grammatical category, as *cant-a* or *am-a*.

Given their fixed order, affixes are further classified into the following main classes:

1. SUFFIXES, if they must follow the root, e.g., *-ish* (*whitish*);
2. PREFIXES, if they must precede it, e.g., *pre-* (*pre-war*);
3. INFIXES, if they appear inside the root, e.g., *bleedin'* in *abso-bleedin'-lutely*;
4. CIRCUMFIXES, if they consist of two parts, one before and one after the root, as in the German participle affix *ge-* … *-t* (*sagen* 'say' > *ge-sag-t*, 'said').

No equivalent distinction is made for roots. The reader will find throughout descriptions of the types of affixes according to their linear adjacency requisites, for instance, in Bauer (2003), as well as in the introductory texts presented in the previous section.

Morphology and Its Interfaces

Because of its nature, the grammatical aspects that morphology analyzes intersect with those explored in other linguistic disciplines. It shares with syntax the study of the grammatical information relevant to build phrases and sentences – tense, number, aspect, agreement, etc. – as well as the rules that combine objects with each other to form structures; with semantics, the study of how units codify meaning and how complex meaning is built by combination of smaller elements; with phonology, the question of how phonemes are combined inside units, the sound restrictions across morphemes, and how the segmental and prosodic properties of words are defined. Moreover, the properties of morphological units (words and affixes) are immediately relevant for psycholinguistics (especially lexical storage and lexical access), first and second language acquisition, pragmatics, and sociolinguistics; morphological research is a crucial part inside historical linguistics and language typology studies; it is obviously relevant to lexicography and other branches of applied linguistics dealing with words. In this section we would like to highlight two of the most relevant contact areas between morphology and other linguistic disciplines.

On the common ground between morphology and semantics, we find LEXICAL SEMANTICS, the study of how the meaning of roots and affixes is obtained, and whether the semantic contribution of a morpheme is internally structured or not. Some authors (Fodor and Lepore, 1998) claim that the meaning of any word is atomic – that is, it cannot be decomposed into smaller primitives that form a structure which accounts for what the word means – but the authors that believe that word meaning should be built from smaller elements are a majority. LEXICO-CONCEPTUAL STRUCTURES (Jackendoff, 1990) and QUALIA STRUCTURE (Pustejovsky, 1995) are examples of approaches where the meaning of morphemes is obtained through a hierarchical combination of smaller semantic primitives (such as *cause*, *become*, *animate*, *shape*, *function*, etc.). Lieber (2004) is a recent example of an attempt to use these structures to analyze morpheme combinations: morphemes come equipped with semantic structures which, in combination with those of the other morphemes in the word, determine properties such as the class of entities denoted by the word, its argument structure, or its aspectual information.

The study of the phonological operations that take place inside words is another important area of study in morphology. LEXICAL PHONOLOGY (Kiparsky, 1982; Mohanan, 1982) concentrates on the study of the special phonological processes that apply internally to a word, such as the rule of SPIRANTIZATION, that turns a dental stop into a fricative in the presence of certain affixes starting with a palatal sound (*naviga[t](e)* > *naviga[ʃ]-ion*). LEXICAL PHONOLOGICAL RULES apply inside words, typically across morpheme boundaries, and are IDIOSYNCRATIC: they allow for real exceptions. If we consider the combination of the morpheme *-al* (*verb-al*) with nouns ending in *-nt*, we see that frequently spirantization does not take place (*argumen[t]* > *argumen[t]-al*), except with some words (*presiden[t]* > *presiden[ʃ]-al*, where the orthography *presidential* represents the palatalization of /t/ with *-i-*). In contrast, POST-LEXICAL RULES apply inside words or between them and are expected to be general in the language; for instance, a post-lexical rule would be flapping in American English, which turns /t/ into [ɾ] in intervocalic position. Note that the rule does apply inside a word (7a) or between words (7b).

(7) a. edit > edi[ɾ]-ed
b. closet > … the close[ɾ] under a…

Recently, the study of lexical phonology in word formation has received revived attention; see Ito and Mester (2003), and especially Stratal Optimality Theory (Bermúdez-Otero, 2007, 2012, 2013).

There are other frequently studied phenomena on the interface between morphology and phonology. REDUPLICATION is a word formation process that involves repeating (part of) a root, as in Hungarian *néha* 'seldom' > *néha-néha* 'only sometimes.' INFIXATION, as hinted before, includes the cases where a morpheme is not completely linearized with respect to its base; here researchers have specially studied the interaction of the phenomenon with the position of stress and the number of syllables of the root. A final issue that has been studied are the asymmetries in the way in which different affixes integrate with its base. See Raimy (2000), Booij (2000), Yu (2007), Frampton (2009), and Kaisse (2005) for more information about these issues.

Inflection and Word Formation (Derivation and Compounding)

If we move now to morphological operations, it is traditional to distinguish between two types of processes: WORD FORMATION and INFLECTION. The first refers to the operations that, on the basis of another word, build a more complex one whose grammatical category, meaning, or other lexical properties have changed. (8) are all examples of word formation processes; in (8a), the second word is a noun, while the first was an adjective. In (8b), the meaning of the word has

changed: the complex word to the right does not simply denote an animal, but the offspring of an animal. In (8c), the elements that the verb requires to be used in a sentence change as a result of the morphological process – the first verb is intransitive, but the second requires a direct object. Any of these three types of changes is considered enough to claim that morphology has built a new word, distinct from the first one.

(8) a. clear > clear-ness
b. whale > whale-calf
c. wail > be-wail

In contrast, inflectional processes produce different forms of the same word, therefore without altering its grammatical category, its meaning, or other central lexical properties (9). (9a), from Russian, shows a noun that appears in a particular case, which depends on its syntactic function inside a phrase, but does not alter what the noun means or its grammatical category; (9b), from Portuguese, shows an adjective in feminine plural form, something which depends on its agreement with a noun; (9c), from Norwegian, shows a verb in past tense. None of these processes alter in any way the grammatical category of the word; in fact, the word could undergo those inflectional processes precisely because it belongs to a particular grammatical category, to begin with. The kind of entity they express remains unaltered (the same entity, the same property, and the same event). It is also worth noticing that normally an inflectional process marks a change that is triggered by syntax (like case assignment or agreement) or directly relevant to it (like the tense in which a sentence appears). Note also that inflection is typically performed through affixes, not through the addition of roots.

(9) a. stol-a
table-gen
b. branc-a-s
white-fem-pl
c. spis-te
eat-past

If we go back to word formation, further divisions are normally made here, depending on the kind of morphemes that are combined, affixes, or roots (cf the section 'Morphology and Its Interfaces').

A word-formation process that involves combining two roots (or two structures that contain roots) is a case of COMPOUNDING (10). If the morphemes combined with the root are all affixes, it is a case of DERIVATION (11).

(10) pick-pocket, air-conditioning, green-eyed, congress-woman, piano-bar
(11) re-build, whit(e)-en, class-ify, change-able, beauti-ful, pre-war, en-rich

Whether these three processes, inflection, derivation, and compounding, are really different phenomena that can be clearly differentiated from each other, however, is questioned by some morphologists. Scholars disagree with respect to whether the distinction between inflection and derivation actually exists; see Van Marle (1984), Haspelmath (1996), and Marantz (unpublished manuscript) for different approaches to this problem. There are also disagreements with respect to whether the division between compounding and derivation is clear; see Ralli (2010) for a summary.

Be it as it may, let us say a bit more about derivation. Given the importance of category change in a derivational process, derivational affixes are sometimes classified into NOMINALIZERS – if they form nouns (12), ADJECTIVIZERS – if they form adjectives (13), and VERBALIZERS – if they form verbs (14). The category of the unit they combine with is also used: DEVERBAL affixes are those that combine with verbs, e.g., *-able* (*dispos-able*); DEADJECTIVAL, those that combine with adjectives, e.g., *-ness* (*absurd-ness*); DENOMINAL, if they combine with nouns, e.g., *-less* (*tooth-less*). Thus, for instance, this last suffix would be called 'a denominal adjectivizer,' meaning it combines with nouns to produce adjectives.

(12) -ment, -(at)ion, -th, -ness, -ity…
(13) -ish, -less, -ful, -(i)ous, -al, -ic…
(14) -ify, -ize, -en…

It is important to remember, however, that category change is a sufficient, but not necessary, condition to characterize a process as derivational. Even when there is no category change, an affix can be considered derivational if the semantics of the base or other lexical aspects are altered (remember the *wail* ~ *be-wail* example). The reader will find recent summaries about word-formation processes in Lieber and Stekauer (2005, 2009) and Scalise and Vogel (2010).

Some Debates in Current Morphological Theories

We conclude this overview briefly presenting some of the main issues that are being debated in current morphological theories. The questions that we present now are, to some extent, orthogonal to each other.

Lexicalism and (Neo)constructionism

The first question has to do with whether morphology is a separate generative component of syntax. LEXICALISM (Halle, 1973; Booij, 1977; Lieber, 1981; Scalise, 1983; Ackema and Neeleman, 2004) argues that, next to syntactic rules, there are LEXICAL RULES which are responsible for word formation. In so-called strong lexicalist approaches, inflection is also performed through operations that take place in the lexicon, despite their obvious interaction with syntax. In these systems, the lexicon contains both a list of units and a set of operations that can apply idiosyncratically to only some units in order to build words. The output of these operations is an object that has all its relevant properties defined: grammatical category, argument structure, meaning, etc. Syntax, which operates after lexical rules, cannot manipulate the properties of words (The LEXICAL INTEGRITY HYPOTHESIS; see Lapointe, 1980). Consequently, the operations in syntax are determined by the properties of words, as defined in the

lexical component. (15) represents the relevant part of the lexicalist architecture of grammar.

(15) A Lexicalist architecture of grammar
List of items
↓
Lexical rules
↓
Syntax

In contrast, Neo-constructionist approaches – see Lieber (1992), Halle and Marantz (1993), and other authors working in the Distributed Morphology framework (Borer, 2005; Svenonius et al., 2009) – argue that word formation and inflection take place in the syntactic component, and thus that 'words' are in actuality also syntactic objects, a particular type of phrase which occasionally has special phonological or semantic properties (Julien, 2007). The lexicon would be just a list of stored items, and there would not be lexical rules. 'Lexical' properties, like grammatical category and argument structure, would emerge in syntax through the combination of roots with syntactic heads in different configurations. In contrast with Lexicalism, where syntax is conditioned by lexical information, in Neo-constructionist approaches, lexical information is defined by syntax. (16) presents the basic Neo-constructionist architecture, where the only generative component is syntax.

(16) List of items
↓
Syntax

Debates between these two approaches concentrate on how Neo-constructionism can explain the apparent effects of the Lexical Integrity Hypothesis, where syntax seems to be unable to access the internal structure of words, and how lexicalism can account for the many similarities between morphological and syntactic structures. See Marantz (1997), Spencer (2005), Williams (2007), and Bobaljik (2012) for different aspects of this debate.

Morphemes or Words?

Traditional morphology assumes that words can be decomposed into smaller units, morphemes, and in the section 'What Is Morphology?' we have also assumed it. However, this is far from established. Theories that decompose words into morphemes (and, therefore, treat morphemes as the basic units of morphological analysis) are called Item-and-Arrangement. They are opposed to theories where the basic unit of morphological analysis is the complete word and which are called Item-and-Process or Word-and-Paradigm (Aronoff, 1976; Anderson, 1992; Stump, 2001). For this second set of theories, morphology should not be viewed as a set of rules that combine morphemes into structures, but as a set of operations that associate a set of morphosyntactic and semantic properties with a particular word. (17) illustrates one possible way of representing the relation between the words beauty and beautiful in these models: there is an operation of adjectivalization that, applied to one word, gives the other.

(17) Adj (beauty) = beautiful

Thus, inside words there are no segmentable units that correspond to distinct pieces of information. In this theory, we cannot associate the property of [adjective] with the segment *-ful*; that segment is simply a phonological sequence that marks a change experienced by the word *beauty*.

The empirical support for theories without morphemes comes from three kinds of phenomena: (1) the fact that morphological decompositions sometimes must isolate recurrent segments which, however, cannot be associated with clear grammatical or semantic information (Berry morphemes), as in *in-stitute*, *pro-stitute*, *con-stitute*, *re-stitute*, where it would be extremely difficult to give a common meaning to the segment *-stitute*. In contrast, a word-based morphology would not encounter this problem, as it would be whole words that have a meaning, not their alleged morphemes; (2) the fact that sometimes there is no one-to-one mapping between information and segments. For instance, in some cases one segment codifies the information that otherwise would seem distributed between two sequences of segments (cumulative exponence), as in Spanish *canta* 'sing' > *canta-ste*, 'sang, 2sg' where *-ste* seems to contain information about tense and person agreement, which in other words would be expressed by different morphemes: *canta-ba-s*, 'sang, 2sg' where *-ba-* would express imperfective past tense and *-s* would express person agreement; (3) the fact that sometimes we can have changes in the grammatical information of the word without any change in its shape, as in conversion, which alters the grammatical category of the word apparently without adding any morpheme: English *dance* (noun) and *dance* (verb). Item-and-arrangement theories would have to postulate a phonologically silent morpheme in these cases, which is a problematic move for some authors; see the papers in Bauer and Valera (2005) for different takes on how to analyze conversion.

Lexical Listing versus Rules

The debate between listing and rules poses this question: how much information should be stored in the lexicon, as part of the idiosyncrasies that a speaker of a language must learn, and how much can be derived productively from abstract rules? For everybody, the lexicon must at the very least contain statements about how the phonological shape of an item is associated with its meaning; e.g., that the animal scientifically called 'Equus africanus asinus' is denoted by the sequence of phonemes /dɒŋkɪ/ in English, /buro/ in Spanish, etc. The questions are two: (1) what else should be listed? (2) if one believes that words have internal structure, in which cases a whole complex word has to be listed, and not its component morphemes?

Lexicalist theories tend to argue that all lexical units must be listed with information about their grammatical category, argument structure – if any, and purely morphological information such as its conjugation class or declension, whether it has irregular inflection, its co-occurrence restrictions with other units

and its position inside a complex structure. For instance, the lexicon would specify that *deep* is an adjective that, when turned into a noun, takes the morpheme *-th* (*dep-th*); in Spanish, *canta* 'sing' would be specified as a regular verb, from the first conjugation of the language; in English again, *-th* would be specified as a nominal affix, a suffix, and one that must combine with adjectives, but cannot follow other affixes (**tooth-less-th*, **child-ish-th*). In contrast, Neo-constructionist theories try to derive as much of this information as possible from general rules; for instance, in Distributed Morphology the grammatical category of a root is not specified in the lexicon, as it is derived from the syntactic context where the root appears (cf Arad, 2003 for other properties derived from here).

Some item-and-arrangement morphologists would argue that there would not be a lexical entry for a suffix like the nominalizer *-th* in the lexicon, and that instead, the words that contain it would have to be stored. Why? Because this affix has a very low PRODUCTIVITY, that is, there are very few roots that use it to nominalize, and speakers cannot build new nouns with *-th*: *dep-th*, *leng-th*, *warm-th*, *bread-th* are possible, but not **tall-th* or **cold-th*. Proposing a rule in this case would be unwarranted, and perhaps even useless, given the few cases that a speaker would have to memorize and store in her lexicon.

A second situation where listing of whole words is proposed is when the word shows unpredictable properties, which cannot be explained as the result of rules governing morpheme combinations. Take the word *birth*. Readers aware of the history of English might know that back in time this word was formed from the verb *bear* and the affix *-th*, with some then predictable phonological changes. Crucially, contemporary speakers are unaware of this, and it is doubtful that in their mind *birth* is related with *bear*; thus, in this case it is likely that a lexical listing solution would be used. Similar cases take place in Romance languages with Latin words. Even though the word *satisfacer* 'satisfy' was morphologically built in Latin by combination of the adverb *satis* 'enough' and the verb *facere* 'make, do,' it is unlikely that average speakers of Spanish relate this verb in any way with *hacer* 'make, do,' because in order to do so one should ignore unpredictable phonological and semantic differences. In general, rules determining morpheme combinations are only proposed when (1) they correctly restrict the set of morphemes to which it applies; (2) they can derive the grammatical information of the resulting word; (3) they derive the meaning of the resulting word from the meaning of the morphemes involved and how they are combined with each other; (4) if necessary, they account for the phonological changes experienced by these morphemes in the word, and (5) speakers can use those rules to productively build novel words.

Exponents and Morphosyntactic Features; Late Insertion

Another debated issue is whether the information stored is placed in one single list or should be distributed in different lists, which are accessed by different components of the grammar. In this second case, one influential proposal is that the MORPHOSYNTACTIC features associated with a morpheme must be stored in one list, to which syntax has access when building the structure, while the MORPHOPHONOLOGICAL information is in a second, distinct list. This is known as the SEPARATION HYPOTHESIS (Beard, 1995). In other words, we cannot properly talk of the nominal suffix *-ness* in *tall-ness*; we should talk about a nominal affix [N] which combines with adjectives – which is its morphosyntactic representation – and a morpho-phonological EXPONENT *-ness*, a suffix, that materializes or spells out that set of features under certain circumstances. This implies that morphemes have two sides: one that determines the properties of the units in syntax and one that determines how they will be phonologically expressed. The mismatches between grammatical information and morphological realization discussed in Section Morphemes or Words? can thus be treated under a new light (cf Ackema, 1995).

Some approaches go further, and propose that these two sides of the morpheme are accessed at different points. The morphosyntactic list is presyntactic, as it defines the properties of the units that syntax manipulates, and consequently determines the conditions under which a syntactic structure is licensed (e.g., that what we pronounce *-ness*, being a nominal element, could not combine with tense); the morphophonological list is postsyntactic, because, by hypothesis, the phonological properties of those items (such as whether they carry stress, end in a consonant, are trisyllabic, etc.) are not taken into consideration to build a syntactic structure. Thus, this information in inserted into the syntactic tree after it has been built; this hypothesis is known as LATE INSERTION (cf Halle, 1997). For some theories, this late inserted information includes also the purely morphological information about morpheme selection (for instance, that the exponent *mov(e)* takes the exponent *-ment* to nominalize, while *explain* takes *-ation*), declension or conjugation, morpheme ordering, etc.

Other authors (e.g., Borer, 2003) include also in a postsyntactic list the conceptual meaning of roots (for instance, that *donkey* denotes a certain animal), which is also unpredictable. This list would be distinct from the other two for two reasons: (1) if the first is relevant for phonology, this second one is only relevant to semantics as the meaning of exponents normally does not correlate with the phonemes involved in them, so they have to be kept in separate components; (2) the conceptual unpredictable meaning is associated with exponents, not to morphosyntactic features. The concepts expressed by *cookie*, *boson*, *dog*, *war*, *bottle*, *child*, and *rose* are all very different, but morphosyntactically they all correspond to the same kind of object: count nouns.

Note that each one of these three lists (morphosyntactic features, morphophonological exponents, and concepts) is relevant to each of the three major areas with which morphology interacts: syntax, phonology, and semantics.

See also: Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects; Lexical Semantics; Lexicology and Lexicography; Morphophonology.

Bibliography

Ackema, Peter, Neeleman, Ad, 2004. Beyond Morphology: Interface Conditions on Word Formation. Oxford University Press, Oxford.

Ackema, Peter, 1995. Syntax below Zero. Doctoral dissertation. OTS, Utrecht University.
Anderson, Stephen S., 1992. Amorphous Morphology. Cambridge University Press, Cambridge.
Arad, Maya, 2003. Locality constraints on the interpretation of roots: the case of Hebrew denominal verbs. Natural Language and Linguistic Theory 21, 737–778.
Aronoff, Mark, 1976. Word Formation in Generative Grammar. MIT Press, Cambridge, MA.
Bauer, Laurie, Valera, Salvador, 2005. Approaches to Conversion/Zero-Derivation. Waxmann Verlag, Münster.
Bauer, Laurie, 2003. Introducing Linguistic Morphology. Edinburgh University Press, Edinburgh.
Beard, Robert, 1995. Lexeme-Morpheme Base Morphology: A General Theory of Inflection and Word Formation. SUNY, New York.
Bermúdez-Otero, Ricardo, 2007. Morphological structure and phonological domains in Spanish denominal derivation. In: Martínez-Gil, Fernando, Colina, Sonia (Eds.), Optimality-Theoretic Studies in Spanish Phonology. John Benjamins, Amsterdam, pp. 497–517.
Bermúdez-Otero, Ricardo, 2012. The architecture of grammar and the division of labour in exponence. In: Trommer, Jochen (Ed.), The Morphology and Phonology of Exponence. Oxford University Press, Oxford, pp. 8–83.
Bermúdez-Otero, Ricardo, 2013. The Spanish lexicon stores stems with theme vowels, not roots with inflectional class features. Probus 25.
Bobaljik, Jonathan, 2012. Universals in Comparative Morphology. MIT Press, Cambridge, MA.
Booij, Geert, 1977. Dutch Morphology: A Study of Word Formation in Generative Grammar. Foris, Dordrecht.
Booij, Geert, 2000. Morphology and phonology. In: Booij, Geert, et al. (Eds.), Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung. Walter de Gruyter, Berlin, pp. 335–344.
Booij, Geert, 2007. The Grammar of Words: An Introduction to Linguistic Morphology. Cambridge University Press, Cambridge.
Borer, Hagit, 2003. Exo-skeletal vs endo-skeletal explanations. In: Moore, J., Polinsky, Maria (Eds.), The Nature of Explanation in Linguistic Theory. CSLI and University of Chicago Press, Chicago, pp. 31–67.
Borer, Hagit. 2005. In Name Only, vol. 1 from the Exoskeletal Trilogy. Oxford University Press, Oxford.
Fábregas, Antonio, Scalise, Sergio, 2012. Morphology: From Data to Theories. Edinburgh University Press, Edinburgh.
Fodor, Jerry A., Lepore, Ernest, 1998. The emptiness of the lexicon: critical reflections on James Pustejovsky's The Generative Lexicon. Linguistic Inquiry 29, 269–288.
Frampton, John, 2009. Distributed Reduplication. MIT Press, Cambridge, MA.
Halle, Morris, Marantz, Alec, 1993. Distributed morphology and the pieces of inflection. In: Hale, K., Keyser, S.J. (Eds.), The View from Building 20. MIT Press, Cambridge, MA, pp. 111–176.
Halle, Morris, 1973. Prolegomena to a theory of word formation. Linguistic Inquiry 4, 451–464.
Halle, Morris, 1997. Impoverishment and Fission. PF: Papers at the Interface. MIT Working Papers in Linguistics, pp. 425–450.
Harley, Heidi, 2006. English Words: A Linguistic Introduction. Blackwell, Oxford.
Haspelmath, Martin, 1996. Word-class-changing inflection and morphological theory. In: Booij, Geert, van Marle, Jaap (Eds.), Yearbook of Morphology. Kluwer, Dordrecht, pp. 43–66.
Ito, Junko, Mester, Armin, 2003. Japanese Morphophonemics: Markedness and Word Structure. MIT Press, Cambridge, MA.
Jackendoff, Ray, 1990. Semantic Structures. MIT Press, Cambridge, MA.
Julien, Marit, 2007. On the relation between morphology and syntax. In: Ramchand, G., Reiss, C. (Eds.), The Oxford Handbook of Linguistic Interfaces. Oxford University Press, Oxford, pp. 209–238.
Kaisse, Ellen M., 2005. Word-formation and phonology. In: Lieber, Rochelle, Stekauer, Pavol (Eds.), The Handbook of Word Formation. Springer, Dordrecth, pp. 25–49.
Kiparsky, Paul, 1982. Lexical morphology and phonology. In: Yang, I.S. (Ed.), Linguistics in the Morning Calm. Hanshin, Seoul, pp. 3–91.
Lapointe, Steven, 1980. A Theory of Grammatical Agreement. Doctoral dissertation. University of Massachusetts, Amherst.
Lieber, Rochelle, Stekauer, Pavol, 2005. The Handbook of Word Formation. Springer, Dordrecht.
Lieber, Rochelle, Stekauer, Pavol, 2009. The Oxford Handbook of Compounding. Oxford University Press, Oxford.
Lieber, Rochelle, 1981. On the Organization of the Lexicon. Doctoral dissertation. University of New Hampshire.
Lieber, Rochelle, 1992. Deconstructing Morphology. Chicago University Press, Chicago.
Lieber, Rochelle, 2004. Morphology and Lexical Semantics. Cambridge University Press, Cambridge.
Lieber, Rochelle, 2009. Introducing Morphology. Cambridge University Press, Cambridge.
Marantz, Alec, 1997. No escape from syntax: don't try morphological analysis in the privacy of your own lexicon. In: Dimitriadis, A., Siegel, L., et al. (Eds.), UPenn Working Papers in Linguistics, vol. 4. University of Pennsylvania, Pennsylvania, pp. 201–225.
Marantz, Alec. Words. MIT, unpublished manuscript.
Mohanan, K.P., 1982. Lexical Phonology. Doctoral dissertation. MIT, Cambridge, MA.
Pustejovsky, James, 1995. The Generative Lexicon. MIT Press, Cambridge, MA.
Raimy, Eric, 2000. The Phonology and Morphology of Reduplication. Mouton de Gruyter, Berlin/New York.
Ralli, Angela, 2010. Compounding versus derivation. In: Scalise, Sergio, Vogel, Irene (Eds.), Cross-Disciplinary Issues in Compounding. John Benjamins, Amsterdam, pp. 57–77.
Scalise, Sergio, Vogel, Irene, 2010. Cross-Disciplinary Issues in Compounding. John Benjamins, Amsterdam.
Scalise, Sergio, 1983. Generative Morphology. Foris, Dordrecht.
Spencer, Andrew, 1991. Morphological Theory: An Introduction to Word Structure in Generative Grammar. Blackwell, Oxford.
Spencer, Andrew, 2005. Word formation and syntax. In: Lieber, Rochelle, Stekauer, Pavol (Eds.), The Handbook of Word Formation. Springer, Dordrecht, pp. 73–99.
Stump, Gregory T., 2001. Inflectional Morphology: A Theory of Paradigm Structure. Cambridge University Press, Cambridge.
Svenonius, Peter, Ramchand, Gillian, Taraldsen, Tarald, Starke, Michal (Eds.), 2009. Nordlyd 36.1, Special Issue on Nanosyntax. Universitetet i Tromsø, Tromsø.
Van Marle, Jaap, 1984. On the Paradigmatic Dimension of Morphological Creativity. Foris, Dordrecht.
Williams, Edwin, 2007. Dumping lexicalism. In: Ramchand, Gillian, Reiss, Charles (Eds.), The Oxford Handbook of Linguistic Interfaces. Oxford University Press, Oxford, pp. 353–382.
Yu, Alan C.L., 2007. A Natural History of Infixation. Oxford University Press, Oxford.

Morphophonology

Hans Basbøll, University of Southern Denmark, Odense M, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Morphophonology is the interaction between morphology and phonology, and these two linguistic disciplines are defined in Section Linguistic Disciplines and the Sign within a structuralist tradition of linguistics. Section Interaction between Phonology and Morphology: Three Examples presents three case studies of morphophonological interaction, departing from phonotactics (regular inflectional morphology of English), prosody (the Danish laryngeal prosody stød in its relation to word structure), and phonological variation (German and Danish stem change in plural formation). In section Morphophonemic Analysis, finally, the stages in a *morphophonemic analysis* are outlined, predominantly as part of a discovery procedure and within a structuralist tradition.

Introduction: Terminology

The term *morphophonology* is here taken in a broad sense, in agreement with its composite parts *morpho-* and *-phonology*, as covering the *interaction* between the linguistic components, or modules, *morphology* (cf Spencer and Zwicky, 1998) and *phonology* (cf Goldsmith, 1995). Accordingly, it is not to be understood as a separate linguistic component (such as morphology, phonology or syntax, for example). There are two further terms which are sometimes used in the same or in a closely related sense, viz. *morphonology* and *morphophonemics*. *Morphonology* (in German *Morphonologie*, French *morphonologie*) was the term which the leading founding father of Prague phonology used (Trubetzkoy, 1929, 1931). More recently, it has been the title of Dressler's (1985) book *Morphonology: the dynamics of derivation*. *Morphophonemics* has been a favorite term in American structuralism since *Menomini morphophonemics* by its leading founding father (Bloomfield, 1939). The term *morphophonemics* suggests more emphasis on the segmental building blocks in linguistic structure than the other terms which are neutral with respect to the distinction between segments and prosody.

Linguistic Disciplines and the Sign

The Linguistic Sign and Basic Linguistic Disciplines

The *sign* (French *signe*) is a key notion in structuralist linguistics, in the tradition following the Swiss linguist Saussure (1916). The sign is a unit – psychic and social – with two sides, viz. *content* (French *signifié* 'what is being signified'), or meaning, and *expression* (French *signifiant* 'what is signifying'), or sound. (Content and expression is the terminology of Hjelmslev (1943), and there is a comprehensive and complicated history of the linguistic sign, with linguistically central names like Peirce (e.g., 1965), and up to today's cognitive linguistics, e.g., in works by Langacker (1987).) One of Saussure's examples is the sign '*horse*' (French '*cheval*'). Its content is, according to Saussure, not the animal in blood and flesh – the animal is just being referred to – but the concept of a horse; nor is the expression the concrete sounds (neither articulatorily nor acoustically), the expression is an *acoustical image* (French *image acoustique*).

The relation between content and expression in a linguistic sign is, according to Saussure, *arbitrary* (French *arbitraire*: 'le signe linguistique est arbitraire'); i.e., there is *no natural connection* between the two: the relation is due to a social convention (normally not an explicit and conscious one, however). A sign like '*horse*' cannot be decomposed into smaller signs: it is one linguistic building block. Its content may be decomposed into smaller semantic elements (such as 'animal,' 'four-legged,' and so on); and its expression can be decomposed into smaller phonological elements (like initial /h/ and final /s/, etc.). But there is no general correspondence between the content elements and the expression elements (as there would be in the counterfactual state of affairs where /h/ would correspond to 'animal,' and so on). Such *minimal signs* (as '*horse*' in our example) are called *morphemes*. The arbitrarity of the sign is a characteristic of morphemes only, not of linguistic signs in general. When smaller signs are combined into larger and more complex signs, the relation between content and expression is no longer arbitrary, but *motivated*, viz. by the individual signs and by the construction(s) they enter into; e.g., when *black* and *bird* are compounded into *blackbird*, the meaning (of a particular ornithological species) is motivated, but not determined, by the meaning of *black* and *bird*.

The *basic linguistic disciplines* – 'language-internally,' excluding 'hyphen-disciplines' like socio(-)linguistics, psycho(-)linguistics and text(-)linguistics, for example – can be defined departing from the linguistic sign: *Semantics* is concerned with the *content*-side of language; and *phonology* with the *expression*-side of language (cf Section Phonology and Its Subparts). *Sign combinatorics* represents the traditional linguistic discipline *grammar*, divided into *syntax* and *morphology* (cf Section Morphology and the Word). When we depart from semantics and *transcend* the border between language-internal and language-external phenomena, we meet *pragmatics*. And when we transcend the border between language-internal and language-external phenomena from phonology, we meet *phonetics* (cf Section Phonology and Its Subparts). Thus both pragmatics and phonetics contain language-internal as well as language-external aspects. *Lexicon*, finally, as a linguistic

discipline is concerned with the *arbitrary* aspects of language, viz. of the relation between content and expression of morphemes (and partly words).

Morphology and the Word

Grammar can thus, from a strictly structuralist point of view, be characterized as the linguistic discipline which treats sign combinatorics, i.e., it is concerned with the *syntagmatic* (concerned with the linear order of elements in the speech chain) – as opposed to paradigmatic (concerned with choices on a particular position in the speech chain) – axis; grammar is subdivided into two (sub)disciplines, viz. *syntax* and *morphology*. The notion *word* is crucial to define the difference between those two: in syntax, the building blocks are words and the frame is (primarily) the sentence; and in morphology, the building blocks are morphemes and the frame the word. Therefore, the definition of a word is essential for morphophonology – the interaction between phonology and morphology.

The term *word* designates a central linguistic unit, and this term is, in distinction to most other units used in linguistic descriptions, by linguists, also a word used frequently by linguistically naïve speakers in general. This is related to the fact that the word is a unit with some psychological reality (and also to the fact that words are separated by space in orthography). The most appropriate linguistic definition of a word is by means of *distributional* characteristics: a word can occur in isolation, it can be moved around in an utterance, and its morphemes occur in a fixed internal order, all these criteria being interrelated and connected to its psychological reality. On the other hand, the word cannot be defined exclusively by semantic criteria – as suggested by the existence of lexicalized phrases, idioms (e.g., *kick the bucket*), complex lexical items, etc. – nor can it be defined cross-linguistically by phonological criteria. It is language-dependent to which degree word boundaries have a phonetic manifestation of any sort, i.e., languages differ typologically in this respect.

Phonology and Its Subparts

Phonology is the linguistic discipline concerned with the expression side of linguistic signs. (I disregard the analysis of written language in this article, and thus the issue of graphemes, letters, etc., cf Section Before Morphophonemic Analysis Can Start.) Whereas morphemes are the building blocks of morphology, *phonemes* are the primary building blocks of phonology. Phonemes are established by means of the *commutation test* using *minimal pairs*, cf Section Input Level to Morphophonemic Analysis: Phonemes. Phonemes are not linguistic signs, and they have no semantic content as opposed to morphemes, but they build up the expression of signs, together with prosodic elements. The basic subdivision of phonological units falls between *segmental* units, viz. phonemes – vowels and consonants – and *suprasegmental* units, i.e., *prosodies*. The latter category encompasses syllable prosodies and word accents (such as word stress, tonal word accents and the Danish stød, cf Section Prosody and Morphology: Danish Stød in Word Structure) and intonation. A particular subpart of phonology is *phonotactics*, i.e., combinatorics of phonemes, and this subdiscipline plays a pertinent role in morphophonology (see Dressler and Dziubalska-Kołaczyk, 2006 on 'morphonotactics'), cf Section Phonotactics and Morphology: English Regular Inflections.

Phonetics – the discipline concerned with language sounds (cf Section The Linguistic Sign and Basic Linguistic Disciplines) – is a crucial discipline for phonology, although it is not – within the structuralist linguistic tradition – considered a subdiscipline of phonology; but phonetics is a necessary condition for phonological analyses and descriptions that aim to be psychologically and physically realistic as well as testable, cf Section Nonchange of Phonetic Content between Levels: A Principle of Naturalness.

Nonchange of Phonetic Content between Levels: A Principle of Naturalness

Lemmas are lexical units. They can be divided into those which do not have several different expressions phonologically, but only one ('nonalternating lemmas') and 'alternating lemmas,' with more than one phonological expression (cf Section Phonological Variation and Morphology: Stem Change in German and Danish Plurals). For nonalternating lemmas, the phonological representation in the lexicon is, I shall propose, the (phonetic) surface form, disregarding only rather superficial phonetic details. This means, for example, that lemmas which do not enter into a particular morphological alternation – for example involving the Danish syllabic prosody stød, cf Section Prosody and Morphology: Danish Stød in Word Structure – are lexically specified accordingly (as having stød or not having stød). This is the null-hypothesis within the present framework. Thus, a lexical specification of a certain phonological property implies constancy of this property, irrespective of the morphological context, except for surface phonetic adjustments. A number of predictions with respect to alternations, in different paradigms and types of words, follow from this approach to the lexicon (see Basbøll, 2005, 2006). The fact that the symbols in the lexical representation are identical to symbols in the phonetic representation is an expression of the *naturalness principle* and holds not just for the relation between phonetics and phonology, but also in the morphology and the lexicon.

Interaction between Phonology and Morphology: Three Examples

Phonotactics and Morphology: English Regular Inflections

The typical pattern for regular inflectional morphology in English can be illustrated by the plural formation of nouns (similar principles apply to 3d person pres. tense, and (with *-ed*) to past tense). The ending is either [z], [s] or [ɪz], depending on the final segment of the stem, e.g., *hens*, *dogs*, *cats*, *busses* [henz dɔgz kæts ˈbʌsɪz]. Rather than listing all possible final segments of stems, the phonetic/phonological category of this segment can be represented through characterizations like 'voiced nonsibilant,' 'voiceless nonsibilant' and 'sibilant' – for the final segments combining with [z], [s] and [ɪz], respectively – typically using distinctive features

(for voicedness, sibilance, etc., as suggested above, but their exact inventory is controversial). Furthermore, the rules for the use of [z], [s] and [ɪz] can be simplified by being applied in a specified sequence.

What are the phonemes that are realized as [z], [s] and [ɪz] (as morphological endings, e.g., for plural)? According to the standard Bloomfieldian principle of 'once a phoneme, always a phoneme' (see Section Input Level to Morphophonemic Analysis: Phonemes), the phonemic analysis is straightforward: [z] is a realization of /z/, [s] of /s/ and [ɪz] of /ɪz/ (or /əz/, a difference which is insignificant here). According to a Praguian analysis, [z] and [s] would be analyzed differently: plural [z] – in examples like *hens*, viz. where the stem ends in a sonorant, e.g., a vowel or a nasal – would be a realization of the phoneme /z/ since in this position there is a contrast between [s] and [z], cf *buzz*, *bus* [bʌz bʌs]. In examples like *dogs*, on the other hand (where the stem ends in a voiced obstruent), plural [z] would not be a realization of the phoneme /z/, however, since there is no contrast to [s] in this position (e.g., *[dɔgs] is not a possible phonetic form in English). Similarly, plural [s] would not be a realization of the phoneme /s/ since there is no contrast to [z] in this position (e.g., *[kætz] is not a possible phonetic form in English). Plural [s], as well as plural [z] after obstruents, therefore realizes a neutralization product – an 'archiphoneme' – according to Praguian principles, i.e., a phonological unit which represents what is common to [s] and [z], without being phonologically specified for voicedness or voicelessness (even though [z] is phonetically voiced and [s] phonetically voiceless), cf Section Input Level to Morphophonemic Analysis: Phonemes.

Since the English plural ending [s] only occurs where [z] is phonotactically impossible, the following procedure (Basbøll, 1972, 2006) determines the choice of plural endings *without any reference to phonetic or phonological content* (such as segments, distinctive features, etc.):

1. Is [z] (phonotactically) OK? If yes: choose [z], if no: go to (2)
2. Is [s] (phonotactically) OK? If yes: choose [s], if no: choose [ɪz]

Illustrated with *hens*, *dogs*, *cats*, *busses* as examples:

1. [henz dɔgz] are (phonotactically) OK; but *[kætz], *[bʌsz]
2. [kæts] is (phonotactically) OK; but *[bʌss] not, so: [ˈbʌsɪz]

Since we know that speaker-hearers, including small children acquiring the language, are sensitive to phonotactical patterns in the language around them – even though we do not know the details of this mechanism – this is an interesting proposal not only because it is simple, but also from a psycholinguistic perspective. But where does the order [z] before [s] before [ɪz] come from? An obvious answer is frequency – as plural endings – in the language (and we know that frequency plays a role in language acquisition).

In addition to this completely regular pattern, English exhibits pl. forms identical to the sg. (as *sheep*), and a limited number of forms with the suffix *-en* (e.g., *oxen*, sg. *ox*), forms with stem change (as *men*, sg. *man*), and forms with a combination of stem change and suffix (as *children*, sg. *child*), cf Section Phonological Variation and Morphology: Stem Change in German and Danish Plurals. A morphophonologically interesting subpattern of stem change + suffix can be illustrated by *lives*, *loaves*, *houses* (sg. *life*, *loaf*, *house* – the latter pronounced with [s] whereas the plural form has intervocalic [z]). The stem change only concerns the final obstruent, it is a simple phonological change from voiceless (in the sg. stem) to voiced (in the pl. stem), and the principles of regular plural formation reviewed above apply to the pl. stem: [z] is selected as the plural suffix.

Prosody and Morphology: Danish Stød in Word Structure

The Danish *stød* is a laryngeal syllable rhyme prosody (somewhat resembling creaky voice). It only occurs in heavy (bimoraic) syllables which can be defined as those having a long sonority rhyme – a long vowel, or a short vowel followed by a sonorant consonant (e.g., a nasal or [l]) – and primary or secondary stress (in a 3 degrees of stress-system), see Basbøll (2005: 82–87, 265ff). Stød is contrastive, and there are many minimal pairs (i.e., pairs of words whose pronunciation is identical except for the presence or absence of *stød*, cf Sections Phonology and Its Subparts and Input Level to Morphophonemic Analysis: Phonemes), e.g., *musen* 'the muse,' definite form of the indefinite *muse* [ˈmuːsən ˈmuːsə] versus *musen* 'the mouse,' definite of the indefinite *mus* [ˈmuːˀsən muːˀs]. The grammatical distribution of stød is extremely complex, but it can be accounted for in Basbøll's Non-Stød Model. The relation between the syllable prosody stød, and word structure in Danish, is revealing for the interaction between phonology and morphology, viz. for morphophonology.

The Danish word has a three-layered structure as follows: {[(stem P1)P2]P3} where P1, P2 and P3 are positions for suffixes and where the position of the suffixes follows general principles of integration of suffixes into stems (in the word structure), see Basbøll (2005: 351ff). The minimal word, in '(…),' the basic word, in '[…],' and the maximal word, in '{…},' each define a *phonological domain*, i.e., a sequence within which certain phonological rules apply. The unmarked situation in the native vocabulary is that heavy (bimoraic, see above in Section Prosody and Morphology: Danish Stød in Word Structure) syllables have stød. There is, however, a principle of non-stød which applies to penultimate syllables in a minimal word, and to monosyllabic stems before a syllable in the basic word. Examples are the pl. form *huse* 'houses,' sg. *hus* [ˈhuːsə huːˀs] where *huse* has non-stød just like the monomorphemic word *muse* above (penultimate syllable of the minimal word); the past form *talte* 'spoke' versus the compounded verb with the same monosyllabic second part *udtalte* 'pronounced' [ˈtˢæːldə ˈuðtˢæːˀldə] (basic words [(*tal*) *te*] and [(*udtal*) *te*], respectively), i.e., a monosyllabic versus a polysyllabic stem before a syllable in the basic word. In all these cases, the position of the suffix in word structure is predicted by the model, P1 being the position for lexicalized (integrated) nonverbal suffixes (such as the plural suffix *-e*, the non-lexicalized pl. suffix being *-er*), and P2 for lexicalized verbal suffixes (such as the past tense suffix *-te*, the nonlexicalized past tense suffix being *-ede*). The definite suffix *-en* (in *musen* 'the mouse' above) always occurs on P3, thus stød is retained in the definite form of *mus* above since the following

syllable occurs outside the basic word. Accordingly, the distribution of stød and non-stød (in heavy syllables), a prosodic phonological phenomenon, cannot be predicted by the phonetic or phonological context alone, but crucially depends on grammatical structure, wherefore its description belongs in the realm of morphophonology.

Phonological Variation and Morphology: Stem Change in German and Danish Plurals

In both German and Danish, noun plurals are formed with either a suffix alone (incl. zero), or stem change alone, or a combination (incl. pl. = sg.), cf Section Phonotactics and Morphology: English Regular Inflections. The following three-way gradation of stem change, in the dimension of transparency versus opacity, based upon general phonological principles, has been proposed as a hypothesis applicable to languages exhibiting stem change as a morphophonological mechanism (Basbøll et al., 2011; Laaha et al., 2011):

1. '*No change*': no phonological change of the plural stem when compared to the singular stem (e.g., *bil, biler* 'car,' 'cars' [biːˀl, ˈbiːˀlɐ]). The category 'no change' does not preclude purely phonetic changes, however, cf Section Before Morphophonemic Analysis Can Start.
2. '*Weak change*': a *phonological change* of the plural stem which is either prosodic or due to an automatic segmental process. The category weak change, as far as segmental phonological changes are concerned (cf Section Phonology and Its Subparts), involves *changes in the string of concrete phonemes* – in a Bloomfieldian sense (cf Section Input Level to Morphophonemic Analysis: Phonemes) – but not in morphophonemes. For German this can be illustrated with the pl. *Bilder* 'pictures,' pronounced with [d], sg. *Bild* pronounced with [t]; these two sounds are in contrast initially (/t/ vs /d/), but finally there is no [d] due to the automatic segmental process of devoicing, and thus no final phoneme */d/ (cf Section Input Level to Morphophonemic Analysis: Phonemes); morphophonologically, on the other hand, there would be |d| in both sg. and pl. Whereas an automatic segmental change does not qualify as a phonological change, obviously, the reversal of such an automatic segmental process does, viz. German revoicing of the stem-final obstruent in the case of *Bild-Bilder* since it is not automatic, cf e.g., *Welt* 'world,' pl. *Welten*, both pronounced with [t].

   Prosodic changes can be illustrated by Danish stød changes (stød drop and stød addition, cf Section Prosody and Morphology: Danish Stød in Word Structure). The distinction between prosodic and segmental phonological changes is relevant, also psycholinguistically, as illustrated by the fact that Danes, when asked to reverse the syllables of the girl's name *Mona* [ˈmoːna] produce [ˈnæːmo] and not *[naˈmoː], thereby suggesting that the prosodic frame including stress and vowel length is stored separately from the string of segmental phonemes (Grønnum, 2007: 42–44).
3. 'Strong change': a segmental change of the plural stem which is *phonologically nonautomatic and primarily motivated by morphology*. Umlaut – also 'metaphony,' a change in stem vowel as in, e.g., English *mouse, mice* – in both German and Danish is a case in point. From a synchronic point of view, the relations between the singular and the plural stem in words with Umlaut are phonologically unmotivated (in German |a u o aw| in sg. (e.g., *Fass, Fluss, Knopf, Haus* [fas fluʊs knɔpf haws]) correspond to |ɛ y ø ɔj| in pl. (*Fässer, Flüsse, Knöpfe, Häuser* [ˈfɛsɐ ˈflʏsə ˈknœpfə ˈhɔɣ̞zɐ]), and in Danish |a a o ɔ| in sg. (e.g., *and, barn, fod, hånd* [anˀ bɑːˀn foðˀ hʌnˀ]) correspond to |ɛ ø ø ɛ| in pl. (*ænder, børn, fødder, hænder* [ˈɛnˀɐ bœɐ̯ˀn ˈføðˀɐ ˈhɛnˀɐ]). There is nothing in the phonological context – as against the morphological one – in today's languages that dictate the strong stem change. Furthermore, these stem changes apply to a limited (fewer in Danish than in German) subset only of morphemes, but very far from all, and there are obvious ambiguities in either direction (e.g., in Danish, sg. |a| can correspond to pl. |ɛ| or |ø|, and pl. |ø| to sg. |a| or |o|). All this shows that we are dealing, for strong stem change, with morphophonology, not phonology (cf Section From Phonemes to Morphophonemes: Introducing Morphological Relatedness).

Since stem change produces opacity which supposedly makes acquisition more difficult, no change, weak change and strong change, defined cross-linguistically as above, are predicted to delay acquisition in this order. Furthermore, from these hypotheses it is predicted that strong change in German will be less difficult for children than strong change in Danish since it is more homogeneous and has a higher type frequency (there are far more lemmas in German than in Danish representing each strong change). These predictions have been successfully tested in a large experiment with (Austrian) German and Danish children by Laaha et al. (2011).

Morphophonemic Analysis

Before Morphophonemic Analysis Can Start

The procedures surveyed below presuppose that the language is known to the phonologist undertaking the phonemic analysis. To *segment* the sound chain into words and to identify phonetic realizations as referring to the same word is no trivial task, due to inherent characteristics of the speech signal. The acoustic signal (in contradiction to the written language with spaces between words) consists of a continuous sound string with unclear or nonexisting markings of boundaries between syllables and such higher level units as words. Furthermore, the *variability* in the sound shape of individual words – both due to speaker characteristics and to the communication situation – makes it difficult to identify different acoustic realizations of the same unit as referring to the same linguistic entity, e.g., a word. Speakers can draw on prosodic cues such as stress, fundamental frequency (the main phonetic correlate of tone) and sonority (German *Schallfülle*, e.g., so that vowels, nasals and obstruents in that order have decreasing inherent sonority) in order to segment words in the sound chain, as well as on distributional – including phonotactical – patterns in the input.

The linguist's phonemic analysis is not undertaken directly on the basis of the sound chain itself – on concrete unanalyzed utterances – but departs from a phonetic transcription, i.e., a rendering of the continuous speech chain in terms of a string of phonetic symbols (for vowels and consonants), with or without diacritics (small marks added to phonetic symbols) – e.g. for aspiration, lowering, syllabicity, etc. – and with or without symbols for prosodic properties like stress, word tones, etc. Such a notation is discrete in the mathematical sense, with a finite and well defined number of symbols, e.g., taken from the International Phonetic Alphabet (IPA-symbols). But note that a huge amount of analysis, explicit and/or implicit, underlies any competent use of such a phonetic transcription system.

There is a fundamental distinction between a transcription of a concrete *token* (one individual instance, by a particular individual on a particular occasion) of speech – whether 'on-line,' e.g., in field work, or from tape etc., these two cases being clearly different – and a transcription of a pronunciation *type* (a single abstract pronunciation, representing an infinite number of possible tokens). The latter task demands a systematic transcription within explicit conventions and presupposes knowledge of the particular speech form. Only phonetic distinctions that either convey differences in meaning within the speech norm in question, or differences with sociolinguistic significance (e.g., sexual, sociological, regional or chronological), are typically rendered in a systematic phonetic transcription. As far as individual differences are concerned, level of *distinctness* ('register'), is a relevant parameter. For the procedures of phonemic analysis surveyed in this article, the phonetic transcriptions concern a high level of distinctness within a well defined speech norm. Furthermore, the stream of speech contains segmental and prosodic information intertwined, and the extraction of prosody from the segmental chain is anything but straightforward.

Input Level to Morphophonemic Analysis: Phonemes

In the phonemic analysis, in order to establish contrastive segments, the *commutation test* using *minimal pairs* is applied (cf Section Phonology and Its Subparts). Thereby both *contrastive segments* and *positions* in the sound chain are established as phonemically relevant entities. For example, taking the Danish word *sat* 'sat (ptc.)' [sad] as point of departure, the minimal pairs *sat*, *hat* 'hat' [had], *nat* 'night' [nad] establish [s], [h] and [n] as three different contrastive segments before [ad]; the minimal pairs *sat*, *sut* 'comforter (for babies)' [sud], *sit* 'his/her/its (refl. poss.)' [sid] establish [a], [u] and [i] as three different contrastive segments between [s] and [d]; and the minimal pairs *sat*, *SAS* '(the common Danish pronunciation of the *Scandinavian Airlines System*)' [sas], *sand* 'sand' [sanˀ] establish [d], [s], and [n] (with stød) as three different contrastive segments after [sa]. By the very same procedure, *three different phonologically relevant positions* are being established, i.e., there are exactly three positions ('places') in the sound chain [sad] where the segments can be contrasted with other segments, and this is established through the minimal pair or commutation test.

The analysis into phonemes is based upon a distinct pronunciation of single words in the norm chosen. The path from contrastive segments to phonemes is constituted by the following steps, observing the biuniqueness condition, viz. the principle that it is possible to infer both the phonemes from the contrastive segments, and also the contrastive segments from the phonemes (according to the phonological rules for the norm in question):

1. Phonetically closely related sounds which are not in contrast – in the sense that they form minimal word pairs, or potential minimal word pairs – are identified phonemically. Typically, two such sound symbols will be in complementary distribution – viz. each sound symbol only occurs in contexts of sounds where the other is excluded – but they might be free variants too (i.e., occur in contexts which are wholly or partly identical, but with no difference in meaning). The identification is based upon phonetic similarity not shared by any other sound segment in the phonetic notation. With reference to a well defined analysis of distinctive features, the sound symbols to be identified phonemically should possess a cluster of distinctive features not found together in any other sound. Examples of such identifications in Danish are [j] and [i̯] to /j/, and [ʁ] and [ʀ̥] to /r/.
2. Further identifications of phonetically closely related sounds which do not contrast in any specific context, can be made, still respecting the principle of biuniqueness, i.e., one and the same sound can be identified in two different ways phonemically, presupposing that the phonemic context is different. It is a condition for this kind of phonemic identification that the realization is motivated by a systematic phonological principle (such as, for Danish, vowel coloring by an adjacent /r/, pronounced as [ʁ] or [ʀ̥] dependent on position, in the direction of this /r/). An example of such an identification in Danish is the assignment of [ɛ] after a [ʁ] belonging to the same syllable ('tautosyllabic') and of [e] not after a tautosyllabic [ʁ] to the phoneme /e/ (e.g., *bredt*, *lidt* 'broad (neuter), little (adv.)' [bʁɛd led]), and the assignment of [a] after a tautosyllabic [ʁ] and of [ɛ] not after a tautosyllabic [ʁ] to the phoneme /ɛ/ (e.g., *ret*, *let* 'straight, easy' [ʁad lɛd]). On this procedure, including phonological fusion, i.e., bi(morpho)phonemic interpretation of a single phonetic, or phonemic, segment, see Basbøll (2006).

The relation between contrastive segments and positions (in the sound chain) are looked upon in very different ways by different schools or traditions of phonology (cf Fischer-Jørgensen, 1975). In my view there is still genuine insight to be gained from structuralist phonology, and from combining an eclectic structuralist approach with more modern ways of considering phonology-phonetics and phonology-morphology-lexicon interrelations.

At one end of the scale we find the strict (post-) Bloomfieldians (Bloomfield, 1933; cf Bloch and Trager, 1942; Joos, 1957) with the tenet 'once a phoneme, always a phoneme,' i.e., when a contrastive segment has been established in one position, it is phonemic in all positions. When /p/ and /b/ are established as contrasting allophones (word-)initially in German, for example (cf the minimal pair *Pein* 'torment' pronounced with initial [p] and *Bein* 'leg' with initial [b]), the sound [p] (word-)finally must be an allophone of /p/,

even though the sound [b] does not occur (word-)finally wherefore there can be no contrast in this position either; the phoneme /b/ is then said to be *defectively distributed*. A consequence of this analysis is that *Lob* 'praise,' *Lobe* (dat.) [lo:p], [ˈlo:bə] exhibits a morphological alternation between /p/ and /b/ (syllable-initially). For example, the stop in *Rad* 'wheel' [ʁɑ:t] and *Rat* 'council' [ʁɑ:t] are pronounced identically word-finally (and hence syllable-finally), but enter different morphological alternations, cf *Rade* 'wheel (dat.)' [ˈʁɑ:də] and *Rate* 'council (dat.)' [ˈʁɑ:tə]. According to the Bloomfield-school, this can be analyzed so that the phoneme /t/ word-finally can alternate with either of the phonemes /t/ or /d/ syllable-initially. Both of the morphophonemes |t| and |d| are thus realized by the phoneme /t/ in word-final position.

In Prague phonology (cf Trubetzkoy, 1939), there would be *neutralization* word-finally in German, i.e., a particular phonemic entity, an *archiphoneme*, would occur in the position of no contrast, viz. a segment which is phonetically voiceless but has no specification for voicing phonologically and is thus neither a /p/ nor a /b/, but something which is common to the two (cf Section Phonotactics and Morphology: English Regular Inflections). Such an *incompletely specified segment* is given a separate phonological notation, e.g., /P/ (as in *Lob* [lo:p]) for a segment which is phonetically a voiceless plosive like initial /p/, but is not in contrast (opposition) to a /b/ (which is in contrast to /p/ initially). A consequence of this analysis is that *Lob*, *Lobe* (dat.) [lo:p], [ˈlo:bə] exhibits a morphophonological alternation between /P/ (word-finally and hence syllable-finally too) and /b/ (syllable-initially).

From Phonemes to Morphophonemes: Introducing Morphological Relatedness

Phonetic and phonemic transcriptions rely exclusively on input from sound. By contrast *morphophonemic* transcriptions (between vertical strokes) involve also morphological information. Heuristically, two phonemes can be identified morphophonemically if they occur in parallel positions in different forms of the same morpheme – e.g. in different inflectional forms belonging to the same paradigm – if it applies to a significant set of morphemes in a systematical way. A Danish example would be /j/ in *spøge* 'make jokes' [ˈsbø:i̯ə] ~ [ˈsbø:ɪ], and /g/ in its past participle *spøgt* [sbøgd], which may be considered as representing the same morphophoneme |g|. There is no claim of biuniqueness (cf Section Input Level to Morphophonemic Analysis: Phonemes) between morphophonemes and sound chain, viz. the same phoneme in the same phonemic context can represent different morphophonemes – as /v/ [ʋ̯] in *sagfører* 'barrister' [ˈsɑʋ̯fø:ɐ] represents the morphophoneme |g|, cf *føre sag* 'litigate' [føɐ ˈsæ:ˀɪ̯], contra |v| in *savklinge* 'saw blade' [ˈsɑʋ̯kʰleŋə], cf *sav* 'saw' [sæ:ˀʋ̯].

The morphophonemic analysis should make phonological patterns simpler and more general, e.g., with simpler phonotactics and more general principles of realization. An example from Danish could be: |p|, |t|, |k| are always realized as plosives, whereas |b|, |d|, |g| have variable realizations as either plosive or approximant, depending on their position in the phonological syllable.

In Danish, the position just before a stressed (nonnative) suffix, like *-i*, *-ik*, *-at*, *-ere* is revelatory for the morphophonemic identity of consonants undergoing consonant gradation. Thereby the morphophonemic analysis of the syllable-final phonemes /b/, /d/, /g/, /ð/ leads to the following alternations and identifications 1 through 5. The generalizations only pertain to a segment in word-final position alternating with a segment in morpheme-final position just before a stressed (nonnative) suffix beginning with a vowel.

1. Syllable-final /d/ alternates with syllable-initial /t/, both representing the morphophoneme |t| (e.g., in *vat* 'cotton wool' [vad], *vattere* 'quilt' [vaˈtˢeːˀɐ] and *demokrat* 'democrat' [demoˈkʰʁɑːˀd], *demokrati* 'democracy' [demokʰʁɑˈtˢiːˀ]);
2. syllable-final /g/ alternates with syllable-initial /k/, both representing the morphophoneme |k| (e.g., in *lak* 'lacquer, n.' [lɑg], *lakere* 'lacquer, v.' [laˈkʰeːˀɐ] and *patriark* 'patriarch' [patˢʁiˈɑːg], *patriarkat* 'patriarchate' [patˢʁiɑˈkʰæːˀd]);
3. syllable-final /b/ alternates with syllable-initial /p/ in some (nonnative) words, thus representing the morphophoneme |p| (e.g., in *mikroskop* 'microscope' [mikʰʁoˈsgoːˀb], *mikroskopi* 'microscopy' [mikʰʁosgoˈpʰiːˀ]);
4. syllable-final /b/ alternates with syllable-initial /b/ in some (nonnative) words, thus representing the morphophoneme |b| (e.g., in *hydrofob* 'hydrophobe' [hydʁoˈfoːˀb], *hydrofobi* 'hydrophobia' [hydʁofoˈbiːˀ]);
5. syllable-final /ð/ alternates with syllable-initial /d/ in (nonnative) words, thus representing the morphophoneme |d| (e.g., in *abbed* 'abbot' [ˈabeð], *abbedisse* 'abbess' [abeˈdisə]).

Morphophonemes as Candidate Lexical Segments

Section Morphophonemic Analysis until now was an attempt to characterize morphophonemic transcriptions *from below* ('bottom-up'), i.e., from sound via phonemes to morphophonemes, as part of a *discovery procedure* (cf Harris, 1951). But morphophonemes can also be considered candidates for *underlying segments* in the lexicon.

Whereas the morphophonemes are derivable from the surface structure, by means of a bottom-up procedure, *lexical segments* are the units posited as underlying. There is not a unique string of morphophonemes corresponding to every string of phonemes or sounds, but rather there may be different solutions, so to speak, to this discovery procedure, depending on which sets of morphological alternations are to be accounted for.

The relation between phonemes and morphophonemes is not biunique: whereas it is in general possible in Danish – for the main speech norm chosen here – to derive the phonemes of a given word form from the morphophonemes, by means of phonological principles and mechanisms, the opposite is not true: Such a biuniqueness presupposes an explicit exhaustive enumeration of the word forms that are considered formally related and nonsuppletive (i.e., not phonologically unrelated forms, such as English *be*, *is*, *was*). In many cases, in many specific phonemic contexts, several morphophonemes correspond to the same phoneme, and vice versa: the same morphophoneme corresponds to several phonemes.

From the point of view of the addressee, and not least of the person acquiring the language, there will be several candidates for lexical segments in many cases, and an ongoing change in the lexical representations can be expected, according to (increased) awareness of morphophonological relations, i.e., morphological relations paired to their phonological realizations.

A decision on such issues depends on whether we are talking about

1. hypotheses about speakers, i.e., about psycholinguistics and psychological reality; or
2. methodological principles for the analysis, regardless of any relation to the behavior of real speaker-hearers.

In the case of (1), speakers' knowledge and awareness of orthography must be expected to play a substantial role in alphabetized cultures, and this issue should be subjected to empirical investigation. If we are to speculate about the psychological reality for speakers of Danish of two equally possible morphophonemic transcriptions |braːv| and |braːg| for Danish [bʁɑːˀʋ̥], which may mean either 'brave' or 'crash' (neither word enters into morphological alternations which might reveal the underlying identity of the stem-final consonant), it may be relevant, for example, that *brav* 'brave' is written with the letter *v*, finally, that occurs initially in words pronounced with [v] (/v/), and the homophonous *brag* 'crash' with the letter *g* that occurs initially in words pronounced with initial [g] (/g/), cf *vi*, *gå* 'we,' 'go' [vi], [gɔːˀ]. Furthermore, different speakers' ideas of relatedness between word forms – whether historically founded or not – may be relevant for their lexical representations, and also this matter should be investigated empirically. These issues represent the strong sense of *morphophonemes* as *candidate lexical segments*.

In the case of (2), *naturalness* can be chosen as a methodological guideline for the relation between different levels (see Section Nonchange of Phonetic Content between Levels: A Principle of Naturalness): the correspondence between the contrastive segment [X] and the phoneme /X/ is more natural than between [X] and /Y/, and the correspondence between the phoneme /X/ and the morphophoneme |X| is more natural than between /X/ and |Z|. This is not just a question of arbitrary labels: both contrastive segments and phonemes have phonetic content, even though the degree of abstraction is not the same; and morphophonemes as well have phonetic content – and can be analyzed in terms of distinctive features, and so on – even though they also encode information about morphological alternations which may have consequences for the phonetic content.

Concluding Remarks

Morphophonemic *analysis* in the narrow sense is a 'bottom-up'-procedure from phonemes, e.g., established from sound by means of the commutation test, to morphophonemes. But the morphophonemic analysis in the broad sense – inversely – can also proceed the other way round, viz. as a 'top-down'-procedure, called *synthesis* in the narrow sense (e.g., in Hjelmslev's (1943) terminology). In Section Morphophonemes as Candidate Lexical Segments, a perspective of synthesis in this sense was taken, viz. when morphophonemes were considered as candidate lexical segments. From that perspective, a phonological description – to avoid the term 'analysis' in order not to confuse the narrow and the broad sense of this term – can be given from lexical segments until phonemes, and further on via contrastive segments to phonetic segments. Such a description involves realization principles or rules of some sort, e.g., as those used in different versions of generative phonology in the broad sense, and somehow recalling basic descriptive principles behind Bloomfield's (1939) *Menomini morphophonemics*.

Note

Sections Phonotactics and Morphology: English Regular Inflections and Morphophonemic Analysis are based upon Basbøll (2006). The IPA-transcription of Danish follows Basbøll (2005), however without diacritics for voicelessness (Danish [b d g] are voiceless). I am indebted to Nina Grønnum for numerous valuable stylistic and other suggestions.

See also: First Language Acquisition, Linguistic Theory of; Linguistic Typology; Linguistics: Overview; Morphology in Linguistics; Phonology; Suprasegmentals.

Bibliography

Basbøll, H., 1972. Remarks on the regular plural formation of English nouns. Language and Literature 1 (3). Copenhagen, 39–43.
Basbøll, H., 2005. The Phonology of Danish. Oxford University Press, Oxford.
Basbøll, H., 2006. Phonemic analysis. In: Brown, K. (Ed.), The Encyclopedia of Language and Linguistics, second ed. Elsevier, Oxford, pp. 351–358.
Basbøll, H., Kjærbæk, L., Lambertsen, C., 2011. The Danish noun plural landscape. Acta Linguistica Hafniensia 43 (2), 81–105.
Bloch, B., Trager, G.L., 1942. Outline of Linguistic Analysis. Linguistic Society of America, Baltimore, MD.
Bloomfield, L., 1933. Language. Henry Holt, New York.
Bloomfield, L., 1939. Menomini morphophonemics. Travaux du Cercle Linguistique de Prague 8, 105–115.
Dressler, W.U., 1985. Morphonology: The Dynamics of Derivation. Linguistica Extranea, Studia, vol. 12. Karoma Publishers, Ann Arbor.
Dressler, W.U., Dziubalska-Kołaczyk, K., 2006. Proposing morphonotactics. Italian Journal of Linguistics 18, 249–266.
Fischer-Jørgensen, E., 1975. Trends in phonological theory. A Historical Introduction (Akademisk Forlag, København). Travaux du cercle linguistique de Copenhague, second ed. (1995), vol. 27. Linguistic Circle of Copenhagen.
Goldsmith, J.A. (Ed.), 1995. The Handbook of Phonological Theory. Blackwell, Oxford.
Grønnum, N., 2007. Rødgrød med Fløde. En lille bog om dansk fonetik. Akademisk Forlag, Copenhagen.
Hamp, E.P., Householder, F.W., Austerlitz, R. (Eds.), 1966. Readings in Linguistics, II. University of Chicago Press, Chicago.
Harris, Z.S., 1951. Methods in Structural Linguistics. University of Chicago Press, Chicago (Reprinted as Structural Linguistics (1963)).
Hjelmslev, L., 1943. Omkring Sprogteoriens Grundlæggelse. Festskrift udgivet af Københavns Universitet. (Reprinted in (1966) Akademisk, Copenhagen; and in English translation (1966) Prolegomena to a theory of language. The University of Wisconsin Press, Madison).
Joos, M., 1957. Readings in Linguistics. ACLS, Washington.
Laaha, S., Kjærbæk, L., Basbøll, H., Dressler, W.U., 2011. The impact of sound structure on morphology: an experimental study on children's acquisition of German and Danish noun plurals focusing on stem change. Acta Linguistica Hafniensia 43 (2), 106–126.

Langacker, R.W., 1987. Foundations of Cognitive Grammar. Stanford University Press, Stanford.
Peirce, C.S., 1965. In: Hartshorne, C., Weiss, P. (Eds.), Collected Papers. Harvard University Press, Cambridge, MA.
Saussure, F. de, 1916. Cours de linguistique générale. Payot, Paris.
Spencer, A., Zwicky, A.M. (Eds.), 1998. The Handbook of Morphology. Blackwell, Oxford.
Trubetzkoy [Trubeckoj], N.S., 1929. Sur la morphonologie. Travaux du Cercle Linguistique de Prague 1, 85–88.
Trubetzkoy [Trubeckoj], N.S., 1931. Gedanken über Morphonologie. Travaux du Cercle Linguistique de Prague 4, 160–163.
Trubetzkoy [Trubeckoj], N.S., 1939. Grundzüge der Phonologie. Travaux du Cercle Linguistique de Prague 7.

Mortality and the HIV/AIDS Epidemic

Sara Hertog and Cheryl Sawyer, Population Division, United Nations Department of Economic and Social Affairs, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Palloni, volume 15, pp. 10058–10061, © 2001, Elsevier Ltd.

Abstract

The knowledge base on HIV/AIDS has improved vastly since the disease first emerged as a global threat to public health some 30 years ago. While then an HIV diagnosis was a certain death sentence, it is now widely considered a manageable chronic disease owing to advances in medicine. Together with successful prevention efforts, these advances have dramatically altered the mortality impact of the epidemic. This article reviews the global epidemiology of HIV/AIDS, takes stock of the present state of the epidemic, describes the evolution of patterns of AIDS mortality, and discusses some of the challenges facing the global response to HIV/AIDS into the future.

Introduction

In 1981, physicians in the United States noticed a disturbing increase in deaths to previously healthy, young gay men due to rare infections and cancers that had seldom been fatal in the past (CDC, 1981). Not long thereafter, doctors in Eastern Africa noted cases of a new, fatal wasting disease. They agreed that whatever had caused these deaths had compromised their patients' immune systems so severely that their bodies were unable to fight infections that ordinarily posed no serious threat to healthy people. Doctors remained unsure as to what exactly was causing the condition, thus they labeled it the 'Acquired Immune Deficiency Syndrome' (AIDS).

In 1984, two separate teams of scientists isolated the viral agent responsible for crippling the immune systems of AIDS patients; it was eventually named the Human Immunodeficiency Virus (HIV) (Coffin et al., 1986). HIV infects and ultimately leads to the destruction of CD4 white blood cells (also called T-cells) that are critical to the body's ability to mount an immune response to viruses and bacteria. A healthy person has between 500 and 1500 CD4 cells mm^{-3} of blood. In the absence of medical intervention, HIV spreads throughout the body and the concentration of CD4 cells in the infected person's bloodstream falls below 200, leaving him or her vulnerable to opportunitstic infections and, ultimately, death (Moore and Chiasson, 1999).

Tremendous advances achieved over the ensuing decades have saved countless lives by enabling individuals, families, and communities to prevent HIV transmission while also prolonging the lives of those already infected. Advances in public health, for example, have included large-scale education campaigns, condom distribution, and needle exchange programs that enable people to protect themselves from HIV. Medical advances have brought a suite of antiretroviral (ARV) drugs that slow the progression of the disease and are highly effective at preventing HIV infection in infants, who account for the majority of new infections among children. The emergence of AIDS invigorated social movements toward gender equality and gay rights, yielding significant advances in confronting the biases, stereotypes, and stigma that amplify vulnerability to HIV infection. The epidemic also gave rise an unprecedented global political response to a shared health threat: it inspired the establishment of a dedicated United Nations agency, UNAIDS, to combat the disease and political declarations of the United Nations General Assembly, including the 2001 Declaration of Commitment on HIV/AIDS and two Political Declarations on HIV/AIDS adopted in 2006 and 2011, as well as billions of dollars in philanthropy, international aid, and domestic health spending.

Despite these advances, UNAIDS estimates that since 1990 around 33 million people have died from AIDS-related causes (UNAIDS, 2012). Indeed, the HIV/AIDS epidemic has dealt a powerful blow to hundreds of years of sustained progress in human survival. It is one of only a handful of diseases that have produced worldwide epidemics, thus achieving pandemic status. Prior to the emergence of HIV/AIDS, virtually all of the world's populations had made progress through the epidemiologic transition, a process characterized by the waning occurrence of large fluctuations in mortality rates attributable to frequent famine and infectious disease outbreaks, and the growing relative importance of diseases associated with aging, such as cardiovascular diseases and cancers. Among the countries most highly affected by AIDS, the disease halted progress through that transition. It has produced sharp increases in mortality rates, especially among adults aged 15–49 years, and it has erased decades of gains in life expectancy. It has killed a sizable proportion of the working-aged population in some countries, with important demographic, social, and economic consequences for this generation and the next.

This article begins with a summary of HIV/AIDS epidemic dynamics, including the modes of transmission and factors that determine whether an HIV epidemic can emerge and be sustained. The section Global Overview of the HIV/AIDS Epidemic provides an overview of the global HIV/AIDS epidemic and explains how incidence and prevalence have varied across regions and over time. The section AIDS-Related Mortality describes the evolution of the mortality impact of HIV, illustrating how the epidemic has shaped the age patterns of mortality and trends in the life expectancy in highly-affected populations. The article concludes by identifying some of the challenges that remain to be tackled in the global response to the epidemic into the future.

HIV/AIDS Epidemic Dynamics

HIV is relatively difficult to transmit compared to other infectious diseases such as influenza and tuberculosis, and even other sexually transmitted infections, such as gonorrhea or syphilis (Abu-Raddad et al., 2008). For HIV transmission to take place, a susceptible person must come into close contact with certain body fluids of an infected individual. Only blood, semen, vaginal secretions, and breast milk can carry quantities of the HIV virus sufficient for transmission to occur. Thus the main modes of HIV transmission are sexual intercourse, injecting drugs using contaminated needles, and from mother-to-child during pregnancy, delivery, or breastfeeding. While medical transfusions of blood containing HIV contributed to the spread of HIV during the early years of the global epidemic, donor blood supplies are now screened routinely and in most countries transfusions no longer carry the risk of HIV transmission that they once did (Takei et al., 2009).

The term 'efficiency' refers to the probability that a virus is transmitted during a given contact between an infected person and a susceptible person. Mother-to-child transmission (also called 'vertical' transmission) is the most efficient mode: on average, in the absence of any preventive intervention, an infant born to an HIV-positive mother has around a 30% chance of acquiring HIV (Rollins et al., 2012). Both injecting drug use and sexual intercourse are far less efficient modes of HIV transmission. The average per-contact probability of transmission among injecting drug users (IDUs) who share needles and other injecting equipment is estimated around 0.8% (Hudgens et al., 2001). Transmission probabilities associated with sexual intercourse can vary widely, depending on the type of intercourse, whether the susceptible male partner is circumcised, and the presence of co-infection with other sexually transmitted infections (STIs). Receptive anal intercourse is the most efficient means of HIV transmission through sexual activity (about 1.7% per contact), followed by receptive penile–vaginal intercourse (0.08% per contact), and insertive penile–vaginal intercourse (0.04% per contact) (Boily et al., 2009; Baggaley et al., 2010). Male circumcision reduces the likelihood of infection by around 60% per sexual contact (Auvert et al., 2005; Bailey et al., 2007; Gray et al., 2007), while the presence of STIs such as gonorrhea, chlamydia, or herpes simplex virus, amplifies sexual transmission efficiency by five times, on average (Boily et al., 2009). Across all modes of transmission, the viral load (the concentration of the HIV in the body fluids) of the infected person is a powerful determinant of the transmission efficiency. Probabilities of transmission are amplified several times over when the viral load is very high, as occurs during the first weeks following the initial infection and again during the late stages of disease (Royce et al., 1997). Despite the low transmission efficiency, sexual intercourse is the most common mode of HIV transmission because of the large proportion of the population at risk (UNAIDS, 2012).

In order for an HIV epidemic to emerge and become sustained in a population, each new infection must, on average, lead to at least one additional new infection. This concept is known as the 'reproductive rate of infection' and is denoted by R_0 (Anderson and May 1991). When R_0 is less than 1, the epidemic cannot sustain itself and eventually dies out. In any given population, R_0 is determined by three factors: (1) the average rate of contact between infected and susceptible persons, (2) the average duration of infection, and (3) the efficiency of transmission (Anderson and Garnett, 2000). In the case of HIV, the duration of infection is long – a median of around 11 years from initial infection to death in the absence of drug treatment for HIV (Ghys et al., 2007) – but transmission tends to be inefficient. Thus, an HIV epidemic can grow and sustain itself only when the rate of contact between infectious and susceptible persons is sufficiently high.

In the early years of a country's HIV epidemic, the rate of contact between infectious and susceptible persons in the general population is low: in nascent epidemics adult HIV prevalence is very low – less than 1% infected among persons aged 15–49 years – and sexual exposures are rare. But in certain key population subgroups, rates of contact between infectious and susceptible persons may be high enough to give rise to localized epidemics. One key population subgroup historically associated with local epidemics is men who have sex with men (MSM), among whom sexual practices were associated with greater transmission efficiency (Royce et al., 1997). Commercial sex workers (CSW) also constitute a key population in the epidemiology of HIV: their frequency of sexual exposure, exposure to multiple partners who themselves have other partners, high prevalence of STI co-infection, low capacity to negotiate condom use, and high rates of exposure to sexual violence leave CSW particularly vulnerable to HIV infection. IDUs are a third key population with heightened vulnerability to HIV infection, again because of their frequent exposures to higher efficiency modes of transmission. When epidemics are sustained among key populations but transmission among the general population remains rare and adult HIV prevalence is less than 1%, a country is said to have a 'concentrated' HIV epidemic.

An HIV epidemic becomes 'generalized' when transmission becomes sustained outside of the key populations and adult HIV prevalence grows to exceed 1%. In generalized epidemics, heterosexual intercourse is the dominant mode of HIV transmission (UNAIDS, 2012). Some characteristics that may facilitate the emergence of a generalized HIV epidemic include high rates of contact between the general population and key populations that are experiencing high HIV prevalence; low coverage of protective factors such as male circumcision or condom use; and high prevalence of factors that enhance transmission efficiency, such as STI co-infections or concurrent sexual partnerships (having relationships with two or more sexual partners that overlap in time, thus increasing the likelihood of contact between an infectious person and a susceptible person during the highly infectious period following the initial infection) (Morris et al., 2010).

The past few decades have witnessed enormous advances in the prevention and treatment of HIV/AIDS, dramatically improving the capacity of individuals, families, communities, and societies to mitigate the medical, social, and economic impacts of HIV/AIDS. The first large-scale interventions to have been prioritized on a global scale focused on education. Disseminating accurate information about HIV/AIDS and how it is (and is not) spread was viewed as key to encouraging protective behaviors and uptake of voluntary testing, as well as to reducing stigma and discrimination surrounding the diagnosis. The role of education in slowing or reversing the growth

of HIV epidemics has not been well established, although household surveys in many countries indicate that young people have been delaying initiation of sexual activity and using condoms correctly and consistently at greater rates (UNAIDS, 2012). Among key populations in particular, information, education, and communication campaigns have focused on reducing the number of sexual partners, promoting condom use and, among IDUs, encouraging users to demand and access sterile injecting equipment.

Other advances in HIV/AIDS prevention have taken the form of medical interventions. Early on, ecological evidence strongly suggested that male circumcision was protective against HIV infection (Bongaarts et al., 1989; Caldwell and Caldwell, 1994). Later, three clinical trials confirmed that observation: circumcision reduced the likelihood of vaginal-to-penile transmission of HIV by as much as 60% (Auvert et al., 2005; Bailey et al., 2007; Gray et al., 2007). The World Health Organization (WHO) and UNAIDS thereafter recommended that voluntary medical male circumcision be promoted as an effective measure toward HIV prevention (in conjunction with correct and consistent condom use) in populations with high rates of heterosexual transmission of HIV and low prevalence of male circumcision (WHO/UNAIDS, 2007). By the end of 2011, just 1.3 million voluntary medical circumcisions had been performed for HIV prevention among men in high prevalence areas (UNAIDS, 2012), but it was hoped that new less invasive technologies will improve the availability and access to the procedure in the future (WHO/UNAIDS, 2011).

Among the most important advances in HIV/AIDS medical science has been the development of a class of drugs, called antiretrovirals (ARV), that effectively slow the progression of the disease, suppressing viral loads often to undetectable levels and allowing CD4 counts to rebound. The first ARVs were introduced in the mid-1990s and by 2010, more than 20 such drugs had been licensed (Vella et al., 2012). WHO guidelines, updated in 2010, recommend that adults and adolescents with HIV begin a course of antiretroviral therapy (ART) once the CD4 count falls to 350 cells mm^{-3} (WHO, 2010a). Once a patient begins an ART regimen, it must be continued for life. At the end of 2011, an estimated 8 million people in low- and middle-income countries were receiving ART, representing 54% of those in need according to the 2010 WHO guidelines. For children who acquire HIV infection in the period surrounding their births or in the first 2 years of life through infected breast milk, the WHO advises that ART begin immediately irrespective of the CD4 count (WHO, 2010b). In 2011, just 28% of children in need of pediatric ART were receiving it. In 2013, the WHO again revised its guidelines to recommend that ART be initiated even earlier, beginning when a patient's CD4 counts drops below 500 cells mm^{-3} (WHO, 2013).

Not only do ARVs dramatically prolong the lives of people living with HIV, but new evidence indicates that these drugs are highly effective at reducing the probability of HIV transmission. The drugs are so effective at reducing mother-to-child transmission that UNAIDS has set a target of zero new HIV infections among infants globally by 2015 (UNAIDS, 2010). Among heterosexual partners, treating the infected partner with ART can reduce sexual transmission efficiency by as much as 96% (Cohen et al., 2011), thus the WHO has recommended that persons living with HIV be offered ART irrespective of CD4 count if in a relationship with an uninfected partner (WHO, 2012a, 2013). While some studies have suggested that ARVs can be effective in preventing HIV transmission if taken prophylactically by the uninfected partner, recent trials have produced disappointing results in part as a result of participants' poor adherence to drug regimens (WHO, 2012b).

Global Overview of the HIV/AIDS Epidemic

To achieve an accurate picture of the number of people living with HIV and AIDS mortality is extremely difficult. Around half of people living with HIV worldwide do not know their status (UNAIDS, 2012) and most AIDS deaths occur in areas that lack reliable death registration (Mathers et al., 2005). To estimate the numbers of people living with HIV (the number of prevalent HIV infections) at the country level, analysts rely on a combination of surveillance and surveys. Sentinel surveillance is used to monitor HIV prevalence – the proportion of the population infected with HIV – in populations at risk for HIV. In concentrated epidemics, sentinel surveillance monitors HIV prevalence among the key populations experiencing HIV epidemics, which may include MSM, IDUs, or CSW, depending on the context. For generalized epidemics, sentinel surveillance monitors HIV prevalence among pregnant women attending clinics for antenatal care. In addition to sentinel surveillance, many countries have conducted household surveys, such as Demographic and Health Surveys (DHS), and included biomarkers that can be used to obtain estimates of adult HIV prevalence. The estimates from household surveys are then considered together with those obtained through sentinel surveillance in order to fit an epidemic curve for each country. In partnership with national statistical offices, UNAIDS produces updates of HIV prevalence for 155 countries every 2 years. Estimates of HIV incidence – the rate at which new HIV infections occur – are then derived from models suited to the prevalence estimates, and estimates of AIDS deaths are predicted from the combined incidence and prevalence estimates, taking into account mortality rates from other causes and the coverage and effectiveness of ART (Stanecki et al., 2012).

HIV prevalence estimates obtained from both sentinel surveillance and household surveys are subject to some degree of uncertainty since they do not cover the entire population. Sampling errors associated with prevalence estimates, combined with additional uncertainty introduced by the models applied to derive incidence and mortality from prevalence data, mean that there is considerable uncertainty surrounding the national estimates produced by UNAIDS. To account for that uncertainty, UNAIDS publishes lower and upper bound estimates for HIV prevalence, incidence, and AIDS deaths.

Figure 1 shows UNAIDS estimates of the number of adults and children living with HIV globally over the period 1990 through 2011. The shaded area represents the associated lower and upper boundaries that reflect the uncertainty introduced by the data and modeling procedures. At the end of 2011, an estimated 34.0 million people were living with HIV worldwide with associated uncertainty ranging from 31.4 to 35.9 million.

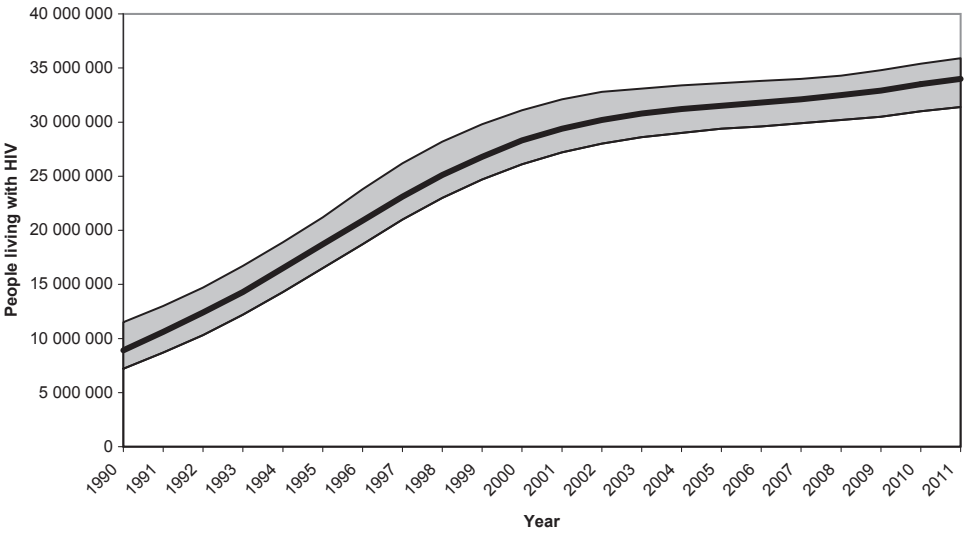

Figure 1 UNAIDS estimates of the number of people living with HIV worldwide, with associated uncertainty, 1990–2011. Data source: UNAIDS Report on the Global AIDS Epidemic – 2012.

This figure represented a 20% increase since the end of 2000, when an estimated 28.3 million (uncertainty bound from 26.1 to 31.1 million) people were living with HIV. The annual growth rate in the number of people living with HIV globally has declined markedly over the course of the epidemic. This rate averaged 14.8% in the early 1990s, but has remained well under 2.0% since 2003 (data not shown).

The continuing increase in the number of people living with HIV globally masks the marked progress achieved in averting new infections and preventing death among those living with HIV. **Figure 2** plots UNAIDS estimates of the numbers of new HIV infections and AIDS deaths globally, along with associated uncertainty bounds, for the period 1990 through 2011. From the figure of 1.9 million (1.6 to 2.1 million) new HIV infections globally in 1990, the annual number of incident infections rose to a peak of 3.5 million (3.4 to 4.1 million) in 1997 and has declined fairly steadily thereafter to an estimated 2.5 million (2.2 to 2.8 million) new HIV infections in 2011. In the absence of changes in mortality, such a reduction in the numbers of new cases would be expected to lead to a decline in the number of people living with HIV. That the number of prevalent cases has instead continued to grow is testament to the improved survival of people living with HIV following uptake of ART. At the end of 2005, the year global AIDS deaths peaked at 2.3 million (2.1 to 2.6 million), ART was reaching less than 2 million people in low- and middle-income countries, but by 2011, 8 million people were receiving ART, representing 54% of the estimated number of people needing it (UNAIDS, 2012).

The geographic distribution of HIV/AIDS impact has been highly uneven (**Figure 3**). At the end of 2011, 56 countries were experiencing generalized HIV/AIDS epidemics with prevalence among the population aged 15–49 in excess of 1%. Among these, 39 countries were located in sub-Saharan Africa, home to 69% of people living with HIV and 71% of AIDS deaths in 2011 (UNAIDS, 2012). In three countries located in Southern Africa – Botswana, Lesotho, and Swaziland – adult HIV prevalence exceeded 20%.

Outside of sub-Saharan Africa, the region with the next highest HIV prevalence is the Caribbean, where 1.0% (0.9–1.1%) of adults aged 15–49 were living with HIV at the end of 2011. Generalized epidemics were noted in the Bahamas, Jamaica, Haiti, and Trinidad and Tobago. In Asia, only Thailand was experiencing a generalized epidemic in 2011, with adult HIV prevalence at 1.2% (1.1–1.2%). The only European country with HIV prevalence greater than 1% in 2011 was Estonia, at 1.3% (1.1–1.6%). Concentrated epidemics – where HIV prevalence was less than 1% in the general population aged 15–49, but greater than 5% among key populations – were more widely distributed geographically in 2011.

Most countries experiencing generalized epidemics have seen the rates at which new infections occur decline in recent years. In sub-Saharan Africa overall, the adult incidence rate fell from its high of 0.86% (0.78–0.92%) to 0.38% (0.35–0.42%) in 2011. In the Caribbean region, the adult incidence rate declined from 0.18% (0.13–0.25%) in 1990 to 0.06% (0.04–0.07%) in 2011. However, evidence points to stalled or reversed progress in some countries, such as Uganda, where the adult incidence rate is estimated to have risen slightly in recent years: after having declined from 1.34% (1.12–1.55%) in 1990 to 0.62% (0.52–0.69%) in 1998, it increased thereafter to reach 0.84% (0.75–1.00%) in 2011. A comparison of results from DHS conducted during the same time period points to increasing prevalence of multiple sexual partners in 12 months and decreasing condom use as possible explanations for the heightened incidence in Uganda (UNAIDS, 2012).

Recent growth in HIV incidence in some regions gives cause for concern that some countries with concentrated epidemics could eventually develop generalized epidemics. UNAIDS notes that the number of new infections in the Middle East and North Africa grew by more than 35% between 2001 and 2011, while some evidence pointed to increasing incidence in Eastern

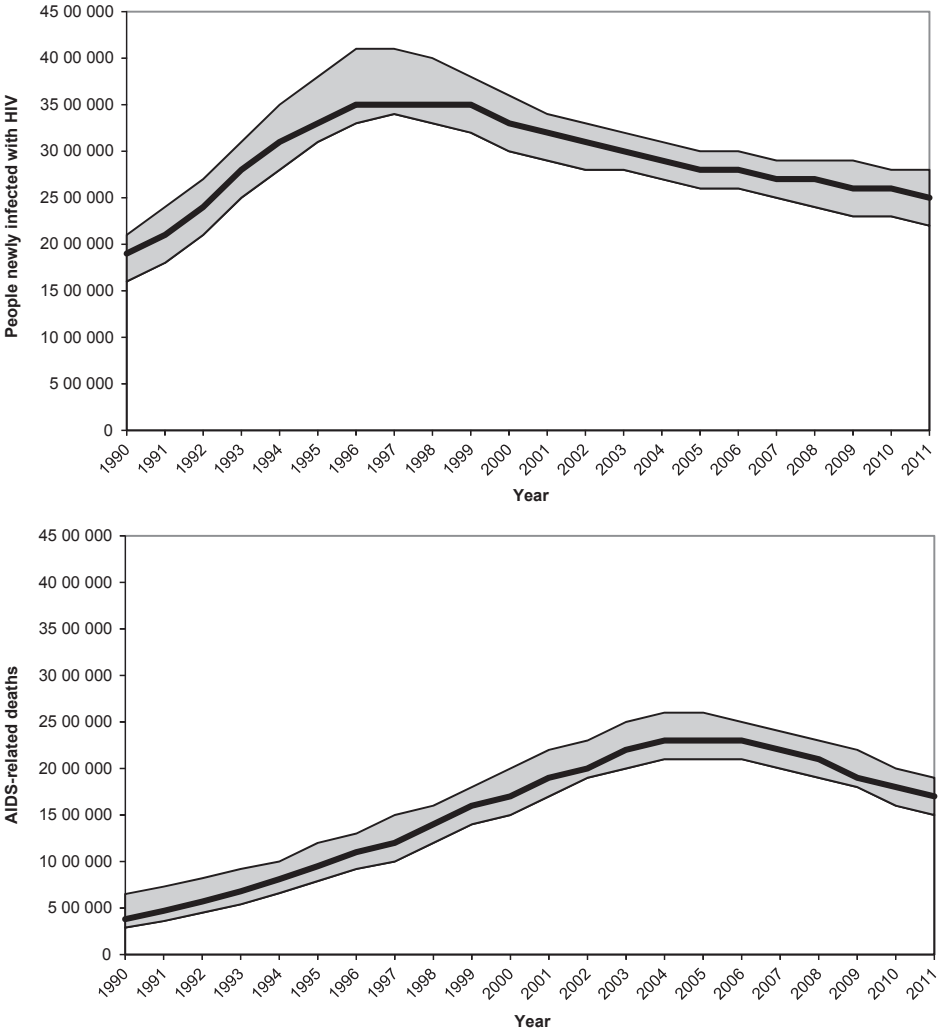

Figure 2 UNAIDS estimates of the number of people newly infected with HIV and AIDS-related deaths worldwide, with associated uncertainty, 1990–2011. Data source: UNAIDS Report on the Global AIDS Epidemic – 2012.

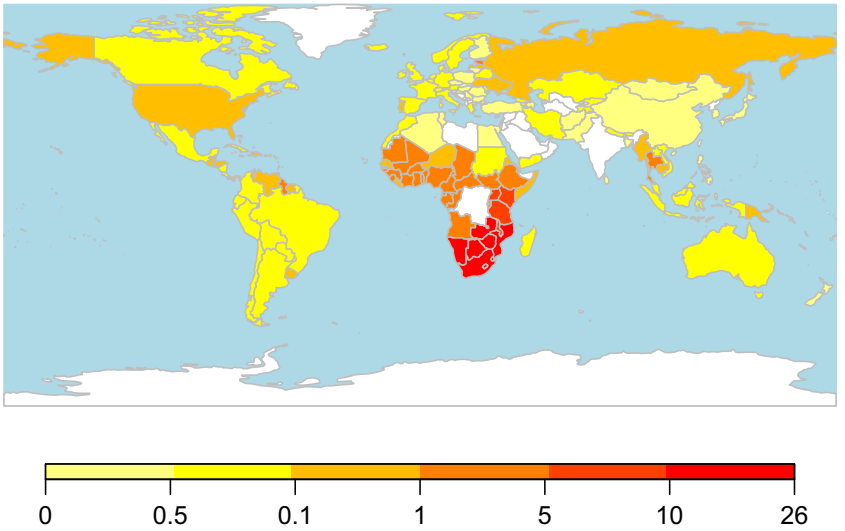

Figure 3 Adult HIV prevalence (%), 2011. Data source: UNAIDS Report on the Global AIDS Epidemic – 2012.

Europe and Central Asia as well. The agency has identified nine countries with increasing epidemics, which it defines as having experienced a 25% or greater increase in the rate of adult HIV incidence between 2001 and 2011: Bangladesh, Georgia, Guinea-Bissau, Indonesia, Kazakhstan, Kyrgyzstan, Philippines, Republic of Moldova, and Sri Lanka.

Some populations saw growth in the overall numbers of people newly infected with HIV, even as the incidence rate remained stable or declined. This phenomenon is attributable to growth in the population aged 15–49. For the world as a whole, the rate of HIV incidence among those aged 15–49 in 2011, at 0.06% (uncertainty ranging from 0.06 to 0.07%), was equivalent to the incidence rate in 1990, but the annual numbers of adults newly infected with HIV grew over that same period from 1.7 million in 1990 to 2.2 million in 2011. This growth corresponds to the growth in the global population aged 15–49 from 2.7 billion in 1990 to 3.7 billion in 2011 (UN-DESA, 2011). In the United Republic of Tanzania, a decline in the adult HIV incidence rate from 0.64% (0.59–0.70%) in 2000 to 0.59% (0.51–0.68%) in 2011 was insufficient to stop the growth in the numbers of new infections each year. Persistent high fertility throughout the latter half of the twentieth century had produced sustained growth in the population at risk for HIV infection from 15.6 million in 2000 to 21.1 million in 2011 (UN-DESA, 2011), resulting in an increase in the numbers of new infections from 97 000 in 2000 to 120 000 in 2011 (UNAIDS, 2012).

AIDS-Related Mortality

Regional Trends in AIDS-Related Mortality

While the annual number of AIDS-related deaths globally peaked in 2005 and has declined thereafter following declining incidence and increasing ART coverage, regional trends in AIDS-related mortality show important differences. Sub-Saharan Africa, which accounted for 78% of global AIDS-related deaths in 2005 and 71% in 2011, has driven the global trend. The region achieved a 33% decline in total AIDS-related deaths between 2005 and 2011 (Table 1). Within sub-Saharan Africa, Botswana, Burundi, Côte d'Ivoire, Ethiopia, Kenya, Namibia, Rwanda, Zambia, and Zimbabwe each achieved reductions in annual AIDS-related deaths of more than 50% during that period (UNAIDS, 2012).

South Asia was the region with the second-highest number of AIDS-related deaths in 2011 with 250 000 deaths and while the region had achieved a reduction in the number of AIDS deaths since 2005, at 14% the pace of decline was slower than in sub-Saharan Africa. Four regions saw increases in annual AIDS-related deaths between 2005 and 2011: Eastern Europe and Central Asia, East Asia, the Middle East and North Africa, and North America. Both the Caribbean and Oceania were estimated to have even larger proportional reductions in AIDS-related deaths than sub-Saharan Africa: the number of AIDS-related deaths was halved in the Caribbean between 2005 and 2011, while Oceania achieved a 43% decline.

Age Patterns of AIDS-Related Mortality

Unlike many other infectious diseases that disproportionately affect children and the elderly, the mortality impact of AIDS is most pronounced among adults in the prime of their lives. Approximately 86% of AIDS deaths occur among adults aged 15–59 (Bongaarts et al., 2011), while 14% of AIDS deaths are among children aged 0–14. Proliferation of ART is anticipated to shift AIDS mortality to later ages into the future.

Figure 4 plots the age-specific mortality rates (m_x; natural log scale) for Southern Africa (which includes Botswana, Lesotho, Namibia, South Africa, and Swaziland) over 5-year periods spanning the region's HIV/AIDS epidemic. During the period 1990–95 – when HIV prevalence was rising, but deaths had not yet had a significant impact on all-cause mortality – the age pattern of mortality from all causes was typical: high in infancy, declining throughout childhood, rising in adolescence, and increasing steadily thereafter through the adult ages. By 1995–2000, excess mortality due to AIDS was visible in the mortality rates of adults of aged 15–49. From adolescence, age-specific mortality rates rose sharply among those in their 20s and leveled off among those in their 30s at a higher level than was observed for the same age range 5 years earlier. The mortality impact of AIDS is magnified even more in 2000–05, with higher age-specific mortality rates among those

Table 1 AIDS-related deaths in 2005 and 2011 and percentage change, by region

	2005		*2011*		
Region	*Deaths*	*(Uncertainty bound)*	*Deaths*	*(Uncertainty bound)*	*Percentage change 2005–11*
Global	2.3 million	(2.1–2.6 million)	1.7 million	(1.5–1.9 million)	−26
Sub-Saharan Africa	1.8 million	(1.6–1.9 million)	1.2 million	(1.1–1.3 million)	−33
South and South-East Asia	290 000	(270 000–310 000)	250 000	(190 000–340 000)	−14
Eastern Europe and Central Asia	76 000	(58 000–100 000)	92 000	(63 000–120 000)	21
East Asia	39 000	(27 000–56 000)	59 000	(41 000–82 000)	51
Latin America	60 000	(36 000–93 000)	54 000	(32 000–81 000)	−10
Middle East and North Africa	20 000	(15 000–25 000)	23 000	(18 000–29 000)	15
North America	20 000	(16 000–26 000)	21 000	(17 000–28 000)	5
Caribbean	20 000	(16 000–23 000)	10 000	(8200–12 000)	−50
Western and Central Europe	7800	(7600–9000)	7000	(6100–7500)	−10
Oceania	2300	(1700–3000)	1300	(<1000–1800)	−43

Data source: UNAIDS Report on the Global AIDS Epidemic – 2012.

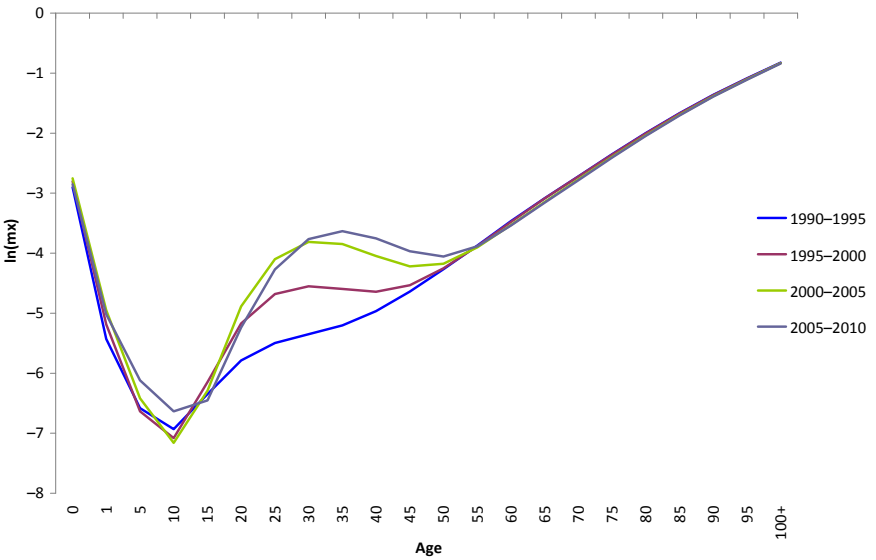

Figure 4 Age-specific mortality rates, Southern Africa, 1990–95 to 2005–10. Data sources: United Nations, Department of Economic and Social Affairs, Population Division, 2011; World Population prospects: The 2010 Revision.

in their 20s, 30s, and 40s than had been observed 5 and 10 years earlier. By 2005–10, elevated death rates shifted to later ages, reflecting both reduced HIV incidence that shifted the age distribution of people living with HIV toward older ages and expanding coverage of ART. While trends in mortality from non-AIDS causes have influenced these all-cause mortality rates as well, much of the recent pattern of adult mortality in Southern Africa is attributable to the impact of AIDS (Reniers et al., 2011).

The evolution of the mortality impact of the epidemic is also evident in the changing survival probabilities among Southern Africa's birth cohorts. Figure 5 charts the percentage of selected birth cohorts who were still alive at each age (a dashed line indicates survival prospects projected after 2010). Survivors of the cohort born in 1950–55 were in their mid-40s at the peak of HIV incidence. They experienced relatively high mortality in childhood, but moderate mortality rates in adulthood such that median age at death (the age by which half of the cohort has died) is estimated at around 65 years. The cohort born 10 years later in 1960–65 was in their 30s during peak incidence and had better survival than their 1950–55 counterparts to age 50, but thereafter saw diminished survival prospects and a median age at death closer to 60. The cohort born in 1970–75 has been especially hard hit by AIDS-related mortality. They were in their 20s during peak incidence and while their survivors are still relatively young – in their late 30s and early 40s at the time of this writing – they are projected to have a median survival of only

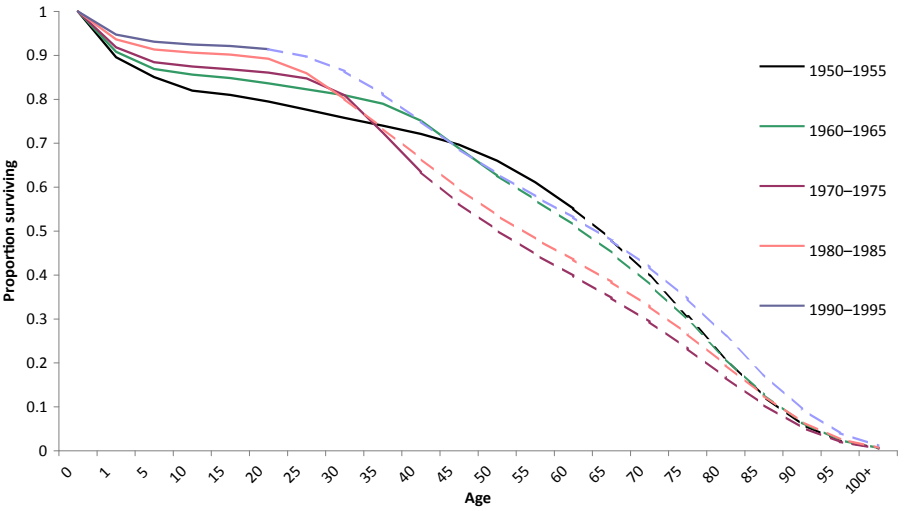

Figure 5 Proportion surviving by age, selected Southern African birth cohorts. Data sources: United Nations, Department of Economic and Social Affairs, Population Division, 2011; World Population prospects: The 2010 Revision.

around 50 years. The survival benefits of ART are anticipated to become evident among Southern Africa's birth cohort born in 1980–85 with a projected median survival of close to 55 years. If current projections prove accurate, the birth cohort of 1990–95 will reap the benefits of both ART and reduced HIV incidence, with their median age at death rebounding to become similar to that of the 1950–55 birth cohort. Such projections are subject to a high degree of uncertainty, however, and rely on assumptions about the future course of the HIV/AIDS epidemic, coverage and effectiveness of ART, and levels of mortality due to other causes. For details regarding the assumptions about the future of the HIV epidemic and coverage and success of interventions, including ART, see United Nations, Department of Economic and Social Affairs, Population Division (2011, 1).

Impact of HIV/AIDS on Life Expectancy at Birth

The life expectancy at birth is a common indicator used to summarize the level of mortality over all ages for a given point in time. It represents the average number of years a hypothetical cohort of individuals would live if subjected throughout their lives to the age-specific mortality rates of a time period. For most of human history, life expectancy increased only very slowly as frequent events like famine and infectious disease epidemics shaped mortality trends. Public health advances of the nineteenth century, however, brought rapid improvements in survival, which were bolstered further with the introduction of vaccines to prevent the spread of many formerly common communicable diseases. As a result, the past two centuries have witnessed tremendous gains in life expectancy, beginning first in the more developed regions and then spreading to developing regions as well. The shift in the predominant causes of death that accompany increasing life expectancy – away from communicable diseases and nutritional conditions and toward the noncommunicable diseases associated with aging – is known as the epidemiological transition (Omran, 1971).

For the countries hardest hit by HIV/AIDS, the epidemic has stalled or even reversed progress through the epidemiological transition, as evidenced by recent trends in the life expectancy at birth. In some countries, such as Botswana where adult HIV prevalence exceeded 20% in 2011, AIDS effectively wiped out the gains achieved in life expectancy since the 1960s (**Figure 6**). Life expectancy in Botswana dropped from 64 years in the late 1980s to 49 years in the early 2000s, before recovering somewhat after 2005 with the rollout of ART. In countries where expansion of access to treatment has not proceeded as quickly, such as South Africa, the recovery in life expectancy has occurred more slowly. While life expectancy in countries with smaller epidemics has been less affected, stalled progress through the epidemiological transition coincident with the emergence of HIV/AIDS is evident. Jamaica, Nigeria, and Thailand have all experienced slowed progress in life expectancy, despite adult HIV prevalence below 5%.

New evidence suggests that expanded coverage of ART has the potential to dramatically improve the expectation of life in populations highly affected by AIDS. Bor et al. (2013) completed the first population-based study of mortality change after widespread ART rollout, focusing on a rural region of South Africa where HIV prevalence approached 30%. They estimated an 11.3 year gain in the life expectancy at age 15 over the 4 year period from 2003 to 2007 following widespread uptake of ART. The study indicated that even without universal coverage of ART, dramatic increases in life expectancy were achievable with widespread access to treatment.

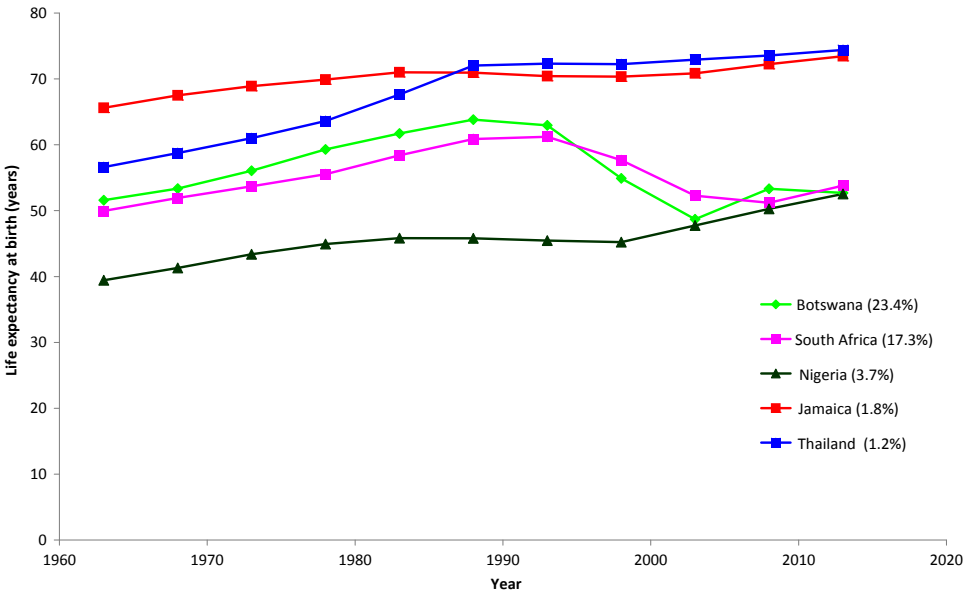

Figure 6 Life Expectancy at birth for selected countries (UNAIDS 2011 HIV prevalence estimate shown). Data sources: United Nations, Department of Economic and Social Affairs, Population Division, 2011; World Population prospects: The 2010 Revision.

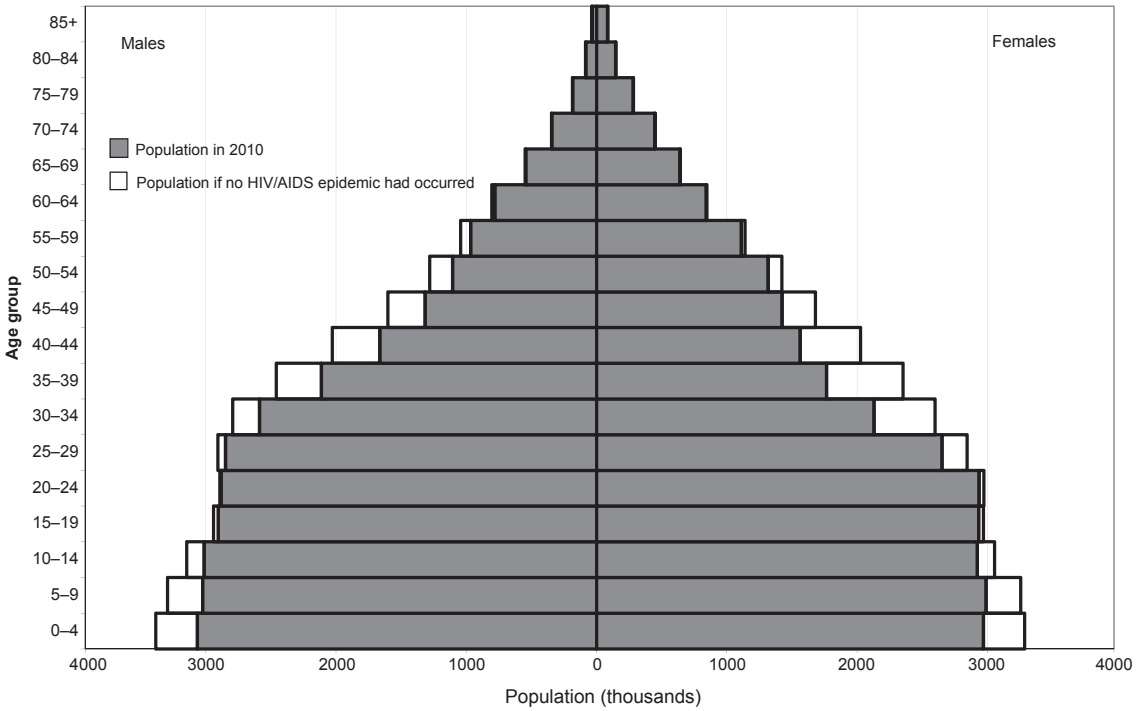

Figure 7 Population by sex and age group, 2010, and counterfactual in the absence of HIV/AIDS, Southern Africa. Data sources: United Nations, Department of Economic and Social Affairs, Population (thousands) Division, 2011; World Population prospects: The 2010 Revision.

Impact of HIV/AIDS on Population Size and Age Structure

In some highly-affected countries, the population size and age structure has been noticeably altered by AIDS-related mortality. Figure 7 displays a population pyramid for Southern Africa, with the size of each bar representing the size of the population in each age group, with children at the bottom of the pyramid and the elderly at the top. The dark shaded bars represent the actual population size in the region in 2010, while the light extensions of the bars indicate what the population size would have been if AIDS-related mortality had never occurred. According to this counterfactual, in 2010, Southern Africa's population of 57.8 million people was 8% smaller than it would have been in the absence of HIV/AIDS. The population deficit is most evident among those aged 30–54 in 2010, among whom AIDS mortality has been most severe: the population in this age group was 16% smaller than it would have been if the epidemic had not occurred. Their absence is felt among the families, communities, and economies left behind. An estimated 17.3 million children worldwide in 2011 had lost one or both parents to HIV/AIDS and 88% of them lived in sub-Saharan Africa (UNAIDS, 2012).

The Future of the AIDS Epidemic

Predicting the future of the HIV/AIDS epidemic has proven a futile endeavor. In the early years of the epidemic some scientists were confident of the imminence of a vaccine, yet 30 years later vaccine trials have repeatedly produced disappointing results. By the early 2000s, forecasts had grown dire, with warnings of 65 million AIDS-related deaths between 2002 and 2020 (UNAIDS, 2002), more than 10 million infected in China by 2010 (UN Theme Group on HIV/AIDS in China, 2002), and population decline in some of the worst affected countries (United Nations, 1991). The reality, while still critical, proved far less calamitous. Credit is due, in part, to education campaigns that have given people the knowledge and tools they needed to protect themselves, and to new drug discoveries that have saved an estimated 14 million life years in low- and middle-income countries alone thus far (UNAIDS, 2012), while simultaneously averting countless new infections.

Continuing the current downward trends in global HIV incidence will depend on success in reducing the prevalence of risk behaviors while increasing access to and use of the proven tools that prevent the transmission of HIV. Among the most effective prevention tools is ART, which also prolongs the lives and improves the well-being of people living with HIV. Despite substantial progress in delivering ART to those who need it, these efforts continue to be outpaced by the number of new HIV infections. In sub-Saharan Africa, a record 2.3 million people were added to treatment programs over the 2-year period 2010–11, but the region saw 3.7 new HIV infections during that same time (UNAIDS, 2012). Increasing uptake of prevention and treatment, in turn, hinges on the success of efforts to promote education and gender equality, while eliminating stigma and discrimination.

Selected additional challenges to the continued reduction of AIDS-related mortality include the high risks of tuberculosis associated with HIV infection and the tenuous sustainability of HIV/AIDS programs. Tuberculosis accounts for nearly one in four AIDS-related deaths and while the number of tuberculosis deaths to people living with HIV has fallen by 25% since 2004, significant challenges remain (UNAIDS, 2012). Tuberculosis

screening, voluntary HIV testing, and early ART delivery need to be scaled up since treatment is known to be effective at preventing tuberculosis acquisition and mortality.

Many of the countries most affected by HIV/AIDS also have insufficient domestic resources to devote to epidemic response and the health system more generally. Their heavy reliance on international investments raises concern surrounding the fragility of existing HIV/AIDS programs. While domestic investment in the AIDS response has increased, international donor support accounted for more than half of the 2011 investment in 26 of 33 sub-Saharan African countries (UNAIDS, 2012). The latest efforts focus on integrating HIV/AIDS prevention and treatment into comprehensive health care services, including sexual and reproductive health care, to improve the sustainability of programming into the future.

Disclaimer

The views expressed in this article do not necessarily reflect those of the United Nations.

See also: AIDS: Acquired Immune-Deficiency Syndrome; Adult Mortality in Developing Countries; Demographic Techniques: Indirect Estimation; Epidemiology, Concepts and Methods of; Infant and Child Mortality in the Less Developed World; Microsimulation in Demographic Research; Mortality, Epidemiological, and Health Transitions; Rehabilitation (Cardiac): Psychosocial Aspects.

Bibliography

Abu-Raddad, L.J., Magaret, A.S., Celum, C., Wald, A., Longini Jr., I.M., Self, S.G., Corey, L., 2008. Genital herpes has played a more important role than any other sexually transmitted infection in driving HIV prevalence in Africa. PLoS One 3 (5), e2230.

Anderson, R.M., Garnett, G.P., 2000. Mathematical models of the transmission and control of sexually transmitted diseases. Sexually Transmitted Diseases 27 (10), 636–643.

Anderson, R.M., May, R.M., 1991. Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, Oxford.

Auvert, B., Taljaard, D., Lagarde, E., 2005. Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 trial. PLos Medicine 2, 1–11.

Baggaley, R.F., White, R.G., Boily, M.C., 2010. HIV transmission risk through anal intercourse: systematic review, meta-analysis and implications for HIV prevention. International Journal of Epidemiology 39 (4), 1048–1063.

Bailey, C., Moses, S., Parker, C.B., et al., 2007. Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomized controlled trial. Lancet 369, 643–656.

Boily, M.C., Baggaley, R.F., Wang, L., Masse, B., White, R.G., Hayes, R.J., Alary, M., 2009. Heterosexual risk of HIV-1 infection per sexual act: systematic review and meta-analysis of observational studies. Lancet Infectious Diseases 9, 118–129.

Bongaarts, J.P., Reining, P., Way, P., Conant, F., 1989. The relationship between male circumcision and HIV infection in African populations. AIDS 3 (6), 373–378.

Bongaarts, J.P., Pelletier, F., Gerland, P., 2011. Global trends in AIDS mortality. In: Rogers, R.G., Crimmins, E.M. (Eds.), International Handbook of Adult Mortality. Springer Science+Business Media, pp. 171–183.

Bor, J., Herbst, A.J., Newess, M.L., Barnighausen, T., 2013. Increases in adult life expectancy in rural South Africa: valuing the scale-up of HIV treatment. Science 339 (6122), 961–965.

Caldwell, J.C., Caldwell, P., 1994. The neglect of an epidemiological explanation for the distribution of HIV/AIDS in Sub-Saharan Africa: exploring the male circumcision hypothesis. Health Transition Review. Supplement to Volumes 4, 23–46.

CDC (U.S. Center for Disease Control), 1981. Kaposi's sarcoma and pneumocystis pneumonia among homosexual men – New York and California. Morbidity and Mortality Weekly Report 30, 305–308.

CDC (U.S. Center for Disease Control), 1982. Kaposi's Sarcoma (KS), Pneumocystis Carinii Pneumonia (PCP), and Other Opportunistic Infections: Cases Reported to CDC as of July 8.

Coffin, J., Haase, A., Levy, J.A., Montagnier, L., Oroszlan, S., Teich, N., Temin, H., Toyoshima, K., Varmus, H., Vogt, P., Weiss, R.A., 1986. What to call the AIDS virus (letter). Nature 321, 10.

Cohen, M., Chen, Y., McCauley, M., Gamble, T., Bollinger, R., July 2011. Antiretroviral treatment to prevent the sexual transmission of HIV-1: results from the HPTN 052 multinational randomized controlled trial. In: Abstract MOAX0102, Presented at the Sixth International AIDS Society Conference on HIV Pathogenesis, Treatment and Prevention, Rome, pp. 17–20.

Ghys, P.D., Zaba, B., Prins, M., 2007. Survival and mortality of people infected with HIV in low and middle income countries: results from the extended ALPHA network. AIDS 21, S1–S4.

Gray, R.H., Kigozi, G., Serwadda, D., et al., 2007. Male circumcision for HIV prevention in young men in Rakai, Uganda: a randomized trial. Lancet 369, 657–666.

Hudgens, M.G., Longini Jr., I.M., Halloran, M.E., Choopanya, K., Vanichseni, S., Kitayaporn, D., Mastro, T.D., Mock, P.A., 2001. Estimating the transmission probability of human immunodeficiency virus in injecting drug users in Thailand. Journal of the Royal Statistical Society–Series C (Applied Statistics) 50 (1), 1–14.

Mathers, C., Fat, D.M., Inoue, M., Rao, C., Lopez, A.D., 2005. Counting the dead and what they died from: an assessment of the global status of cause of death data. Bulletin of the World Health Organization 83, 171–177.

Moore, R.D., Chiasson, R.E., 1999. Natural history of HIV infection in the era of combination antiretroviral therapy. AIDS 13, 1933–1942.

Morris, M., Epstein, H., Wawer, M., 2010. Timing is everything: international variations in historical sexual partnership concurrency and HIV prevalence. PLoS One 5 (11), 1–8.

Omran, A.R., 1971. The epidemiologic transition: a theory of the epidemiology of population change. Milbank Memorial Fund Quarterly 49 (4), 509–538.

Reniers, G., Masquelier, B., Gerland, P., 2011. Adult mortality in Africa. In: Rogers, R.G., Crimmins, E.M. (Eds.), International Handbook of Adult Mortality. Springer Science+Business Media, pp. 151–170.

Rollins, N., Mahy, M., Becquet, R., Kuhn, L., Creek, T., Mofenson, L., 2012. Estimates of peripartum and postnatal mother-to-child transmission probabilities of HIV for use in spectrum and other population-based models. Sexually Transmitted Infections 88, i44–i51.

Royce, R.A., Sena, A., Cates Jr., W., Cohen, M.S., 1997. Sexual transmission of HIV. New England Journal of Medicine 336, 1072–1078.

Stanecki, K., Garnett, G.P., Ghys, P.D., 2012. Developments in the field of HIV estimates: methods, parameters and trends. Sexually Transmitted Infections 88, i1–i2.

Takei, T., Amin, N.A., Schmid, G., Dhingra-Kumar, N., Rugg, D., 2009. Progress in global blood safety for HIV. Journal of Acquired Immune Deficiency Sydromes 52, S127–S131.

UNAIDS, 2002. UNAIDS Report on the Global HIV/AIDS Epidemic.

UNAIDS, 2010. Getting to Zero: 2011-2015 Strategy of the Joint United Nations Programme on HIV/AIDS (UNAIDS).

UNAIDS, 2012. UNAIDS Report on the Global AIDS Epidemic – 2012.

United Nations, Department of Economic and Social Affairs, Population Division (UN-DESA), 2011. World Population Prospects: The 2010 Revision. Volume I: Comprehensive Tables. United Nations, New York.

United Nations, Department of Economic and Social Affairs, Population Division, 1991. The AIDS Epidemic and its Demographic Consequences. United Nations, New York.

UN Theme Group on HIV/AIDS in China, 2002. HIV/AIDS: China's Titanic Peril. UNAIDS, Geneva.

Vella, S., Schwartlander, B., Sowk, S.P., Eholie, S.P., Murphy, R.L., 2012. The history of antiretroviral therapy and of its implementation in resource-limited areas of the world. AIDS 26 (10), 1231–1241.

WHO, 2010a. Antiretroviral Therapy for HIV Infection in Adults and Adolescents. Recommendations for a public health approach: 2010 revision.

WHO, 2010b. Antiretroviral Therapy for HIV Infection in Infants and Children: Towards Universal Access. Recommendations for a public health approach: 2010 revision.

WHO, 2012a. Guidance on Couples HIV Testing and Counselling – Including Antiretroviral Therapy for Treatment and Prevention in Serodiscordant Couples. Recommendations for a public health approach.

WHO, 2012b. Guidance on Oral Pre-exposure Prophylaxis (PrEP) for Serodiscordant Couples, Men and Transgender Women Who Have Sex with Men at High Risk of HIV. Recommendations for use in the context of demonstration projects.

WHO, 2013. Consolidated ARV Guidelines 2013: Summary of New Recommendations.

WHO/UNAIDS, 6–8 March, 2007. New data on male circumcision and HIV prevention: policy and programme implications: conclusions and recommendations from the WHO/UNAIDS Technical Consultation on Male Circumcision and HIV Prevention: Research Implications for Policy and Programming Montreux, Switzerland.

WHO/UNAIDS, 2011. Joint Strategic Action Framework to Accelerate the Scale-up of Voluntary Medical Male Circumcision for HIV Prevention in Eastern and Southern Africa, 2012–2016. UNAIDS, Geneva.

Mortality, Biodemography of

Anatoliy I Yashin and Eric Stallard, Center for Population Health and Aging, Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E.M. Crimmins, volume 15, pp. 10061–10065, © 2001, Elsevier Ltd.

Abstract

The biodemography of mortality aims to explain observed age patterns and time trends in mortality rates from the biological point of view. It integrates knowledge about biological factors and mechanisms affecting mortality rates with demographic information and approaches using unifying theoretical concepts, methods of mathematical modeling, and statistical analyses. The future of this research area depends on the progress in understanding factors and biological mechanisms involved in regulation of aging, and their connections to mortality. These include progress in experimental fields and advances in understanding evolutionary forces shaping the age pattern of mortality.

The biodemography of mortality deals with the integration of knowledge about biological factors and mechanisms involved in forming age patterns and time trends of mortality rates by combining demographic concepts and approaches with biological information, concepts, and methods in the analysis of mortality data. The typical pattern of age-specific mortality rates exhibits an initial spike at birth followed by a decline during infancy and early childhood (ontogenescent part) followed by an increase during adolescent and later ages (senescent part). The declining and increasing parts reflect the effects of distinct interacting components of the age-specific mortality rate that are usually studied separately. These interacting components contribute to a broader phenomenon of mortality correlations at different ages, across populations, and over time (Vaupel et al., 1998). Factors that have been considered as possible contributors to such correlations include evolutionary forces, *in utero* environments, and early life conditions, among others (*see* Early Life Influences on Health and Mortality in Adulthood). Despite intensive study, no general consensus has been reached on the nature of the relationships between these two components of the mortality curve.

Although studying mutual connections between the two components has scientific merit, most attention during the last decade has focused on the senescent (increasing) component. This attention was motivated by the substantial progress achieved in manipulating rates of individual aging in laboratory animals raising the possibility of extending life and reducing disease prevalence in humans (Sierra et al., 2009). To understand whether and how such goals could be achieved for humans, one has to better understand the biological machinery involved in the regulation of aging-related declines in health, well-being, and physiological functioning, and the connections between such declines and health and survival outcomes.

Biodemographic studies of the senescent component of human mortality allow for efficient use of integrated knowledge from relevant research fields including epidemiology, genetics, sociology, gerontology, environmental sciences, population genetics, etc., for maintaining human health, well-being, and longevity. There are many interesting biodemographic ideas, concepts, and methods relevant to understanding mortality. This article provides a selective account of important historical steps in the development of this research field and illustrates how the integration of demographic and biological knowledge and data contributed to progress in the field.

Frailty Models

Survival experiments with large populations of laboratory animals provided evidence that the senescent component of the age-specific mortality rates in a number of species increases exponentially over most of the adult age range (following Gompertz's curve) but, at the upper end of the adult age range, it decelerates, levels off, or even declines. The results of these analyses were summarized in Vaupel et al. (1998). The downward deviation from the Gompertz curve has been explained by the presence of hidden (unobserved) heterogeneity in the chances of death. The demographic frailty model (Vaupel et al., 1979) employed an unobserved 'fixed frailty' variable to describe individual differences in susceptibility to death; the fixed frailty concept provided a convenient tool for analyzing mortality rates over the entire range of adult ages. The use of such models was a substantial step forward in understanding the forces shaping cohort mortality curves and in elucidating the roles of compositional changes in this process.

Although the fixed frailty models improved the descriptions of late-age mortality curves, they provided little information about the biological mechanisms influencing the mortality rates. Moreover, it was found that the class of fixed frailty models was not distinguishable from a class of models with randomly changing heterogeneity variables (Yashin et al., 1994). This observation indicated that additional information is needed to better understand the forces involved in regulation of the mortality rates. One way of adding such information was to introduce explanatory variables (observed covariates) into the fixed frailty models. Such models then treated hidden frailty as a random effect and corresponding studies were focused on estimation of effects of observed covariates on survival in the presence of hidden frailty. It was shown that the presence of hidden frailty influences the

estimates of the effects of observed covariates on mortality risks. It was found that, when ignored, the presence of unobserved frailty diminishes estimated effects.

An important class of extended frailty models used information on survival in related individuals (e.g., twins, sibs, and other relatives). For a number of such models (which differ in their frailty distributions), the multivariate survival function was described in a semiparametric form. A remarkable property of such semiparametric models was that one did not need to specify a parametric form for the baseline hazard rates. These hazard rates could be estimated semiparametrically from bivariate or multivariate survival data. Multivariate survival models were used in analyses of the heritability of individual susceptibility to death, as well as the heritability of mortality by cause (Wienke, 2010). The fact that extreme longevity runs in families suggests the high potential for further development of multivariate survival models to study factors and mechanisms affecting exceptional survival in families, including genetic factors.

Biodemographic Ideas in Genetic Analyses of Mortality

The genetics of aging, longevity, and mortality has become the subject of intensive analyses, ranging from studies of candidate genes to genome-wide association studies (GWAS) (Nebel et al., 2011) that involve hundreds of thousands to millions of genetic variants (i.e., single-nucleotide polymorphisms). However, many estimates of genetic effects on longevity in GWAS have not reached genome-wide statistical significance (using the Bonferroni correction for multiple testing) and many findings remain nonreplicated.

Possible reasons for slow progress in this field include the lack of a biologically based conceptual framework that would drive development of statistical models and methods for genetic analyses of data, the presence of hidden genetic heterogeneity, the collective influence of many genetic factors each with small effects, the effects of rare alleles, and epigenetic effects; as well as molecular biological mechanisms regulating cellular functions. In addition, researchers studying the genetics of human aging and longevity tend to underestimate the role of demographic information and biodemographic ideas in genetic analyses. The use of demographic mortality data and models of hidden heterogeneity in mortality helped improve the accuracy of genetic estimates in genetic centenarian studies. These analyses were able to detect the nonmonotonic (U-shaped) age trajectories of genetic frequencies corresponding to intersections of mortality rates for carriers and noncarriers of the respective genetic variants. The existence of such trajectories was confirmed in later studies. Alternative hypotheses describing possible mechanisms responsible for such age patterns of genetic frequencies (intersections of mortality rates) were discussed in Yashin et al. (2001) and Bergman et al. (2007). The U-shaped patterns of genetic frequencies indicated that the effects of the corresponding genetic variants on mortality risk changed from harmful to beneficial during the life course. This property suggested that some genetic risk factors contributing to the mortality increase early in life are likely to be found in the genomes of long-lived individuals. This conclusion has been confirmed in a number of recent studies of human aging and longevity. The use of demographic information and models in analyses of data on genetically heterogeneous cohorts allowed researchers to compare the age patterns of mortality rates for carriers and noncarriers of candidate alleles and genotypes. Such comparisons were not possible using data on genetic frequencies alone. These methods were further extended in Arbeev et al. (2011). Recent genetic analyses of health-related traits revealed that many genes affecting life span and health span do exhibit pleiotropic effects.

Evolution of Aging, Health, and Mortality

The possibility of evolutionary origins of common aging-related diseases in situations where disease susceptibility alleles demonstrate deleterious or slightly deleterious effects were discussed in Pritchard (2001) and Reich and Lander (2001). Their results indicated that the alleles affecting life span must show pleiotropic associations with health span, duration of unhealthy life, and mortality rates by cause. Under other population genetics scenarios (Di Rienzo, 2006), neutral genes can also be involved in the origin of common diseases. Since such genes have little or no effect on fitness, their associations with health-related traits are likely to be manifest in various trade-offs. The evidence for negative correlations between select diseases has been provided in Stallard (2002) and Ukraintseva et al. (2010), among others. The existence of dependencies among diseases suggests the possibility of common genetic backgrounds for groups of health-related traits. The results of recent studies confirmed the existence of pleiotropic associations of certain genes with health traits (Jeck et al., 2012). Thus, in addition to pleiotropic associations of genes between life span and health-related traits, such genetic associations can also be found among groups of health-related traits.

In parallel with demographic efforts focused on developing and implementing models of survival in heterogeneous populations, researchers in the field of aging recognized the crucial role of aging-related changes in biomarkers that influence the chances of death. In contrast with Gompertz's assumption about a rapid exponential decline in the 'vitality function' with increasing age, the reviews of physiological studies of aging carried out in the middle of the last century showed that physiological parameters characterizing many biological human capacities tended to decline almost linearly with age.

Strehler and Mildvan's Model of Mortality and Aging

To reconcile a linear decline in biological capacity with exponential increases in the rate of mortality, Strehler and Mildvan (1960) proposed a model of aging and mortality (SM-model). In this model, mortality was viewed as a response of an aging organism to the random fluctuations of energy demands. According to the SM-model, the death rate at a given age was proportional to "the frequency of stresses which surpass the ability of a subsystem to restore the initial conditions" at that age. They showed that, under such an

assumption, the exponential increase in mortality (Gompertz's curve) results from the linear decline in vitality. The model allowed them to explain the negative correlation between the two parameters of the Gompertz curve (SM-correlation) observed in empirical studies of the age patterns of mortality rates. The rectangularization pattern of survival improvement during the first half of the last century in developed countries was in accordance with the SM-correlation, as predicted by the SM-model (*see* Mortality: The Great Historical Decline). These changes were linked to a decline in the average magnitudes of stresses experienced by subsequent population cohorts. The rectangularization pattern of survival improvement was also exhibited in subgroups of individuals ordered by the number of longevity alleles that they carried (Yashin et al., 2012). Since the environmental conditions in each such group remained the same, the SM-model linked better survival with higher initial values of the vitality function, suggesting that exceptional survival was likely to have a genetic background. The more longevity alleles that were carried by study participants the higher was the value of their initial stress resistance. In terms of genetic functions, the resistance to stresses is associated with repair capacity, redundancy, and other functions that increase an organisms' resilience, or robustness. A useful illustration of how increased resilience may influence survival can be given by using the model of saving lives (Vaupel and Yashin, 1987). An increase in the 'number of lives being saved' for individuals in the population can be interpreted as an increase in redundancy: for each lost life (premature death), the decedent has one or more 'redundant' lives, depending on the relevant advances in medical technology.

Stochastic Process Model

Strehler and Mildvan's (1960) theory stimulated further research in the direction of biodemographic analysis of mortality rates. Woodbury and Manton (1977) introduced a multivariate stochastic process model of human mortality and aging. This model has been intensively used in analyses of longitudinal data on aging, health, and longevity (Manton et al., 1995). This model was further extended in Yashin et al. (2011) to incorporate state of the art advances in aging research into the model structure. In particular, the extended model described partly observed aging-related changes in physiological variables linked together with age-dependent unobserved variables that included resistance to stresses, adaptive capacities, physiological norms ('optimal' physiological states), stochasticity, allostatic adaptations, and allostatic load. Together, these extensions represented biological mechanisms of aging-related changes in humans, which were consistent with existing biological knowledge about aging. The model described how the age-dependent unobserved variables interacted with partly observed physiological indices and other factors, and linked them with health and survival outcomes. This provided a convenient conceptual framework for comprehensive systemic analysis of aging-related changes in humans using longitudinal data and for linking these changes with genotypic profiles, and morbidity and mortality risks. The model has been used to develop unique efficient statistical methods for analyzing longitudinal data on aging, health, and longevity.

Fertility–Mortality Trade-Off

It has been argued that the priority that natural selection places on reproduction negatively affects other processes such as longevity; the problem posed by this trade-off underlies the disposable soma theory for the evolution of human aging (Drenos and Kirkwood, 2005). The innate immunity could be a critical factor skewing an individual into the direction of either a high fertility or better maintenance strategy (Van Bodegom et al., 2007). Over thousands of years, resistance to infectious diseases through inflammatory host responses was favored because it allowed humans to survive up to reproductive age. Such responses, however, may be negatively associated with fertility and may explain the trade-offs between fertility and the subsequent development of chronic degenerative diseases. The data on historical human populations do not support the hypothesized negative effects of reproduction on longevity: married women who remained childless lived no longer than women who reproduced and women who had few children lived no longer than women who had many children (Lycett et al., 2000). However, the relationship between longevity and reproduction gets stronger with increasing poverty. Other data show a negative influence of parity and a positive influence of age at last childbirth on postreproductive survival, as well as a significant interaction between these two variables (Gagnon et al., 2009).

Index of Cumulative Deficits (DI)

Index of cumulative deficits (DI) or frailty index appears to have broad utility as a promising measure of systemic aspects of aging. The properties of the DI have been validated using various data sets, including the National Long Term Care Survey, the National Health and Nutrition Examination Survey, the Canadian Study of Health and Aging, the Australian Longitudinal Study of Aging, and others. The DI predicts mortality outcomes (Mitnitski et al., 2001; Kulminski et al., 2007) and it does so better than chronological age does. This index is associated with morbidity, activity of daily living (ADL) limitations, cognitive impairment, institutionalization, and hospitalization. The DI has similar relationships with age regardless of the nature of the specific deficits, so long as the set of deficits is reasonably broad; moreover, these relationships hold across a range of settings. The DI was found to reach maximum levels that were age independent, which suggests that there is a limit to the extent of deficit accumulation that is compatible with life. It has the potential to describe biological aging and differentiate short- and long-life phenotypes. In addition, an important aspect of the DI is that it allows one to more comprehensively address the health/well-being of an individual by representing mild-effect traits (i.e., traits which are traditionally ignored in analyses due to their low individual effects) in a coherent way. Since the DI includes dimensions that are readily available from typical population-based studies, it appears to be a promising candidate to serve as a measure of

aging that could be used in demographic models. Thus, on the one hand, the DI can summarize a range of measures relevant to the biology of aging that are amendable to preventive interventions and, on the other hand, it can be incorporated into individualized and/or population-based models of senescence, longevity, and mortality. To find the best descriptions of aging-related declines in health/well-being, further mathematical modeling and comparative analyses of different indices are needed.

The future

The future of this research area depends on the progress in understanding factors and biological mechanisms involved in regulation of aging, and their connections to mortality. These include progress in experimental fields and advances in understanding evolutionary forces shaping age patterns of mortality (*see* Evolutionary Demography: A Synthesis of two Population Sciences).

See also: Aging and Health in Old Age; Biomarkers in Demographic Research; Early Life Influences on Health and Mortality in Adulthood; Evolutionary Demography: A Darwinian Renaissance in Demography; Gene–Environment Interplay; Mortality Differentials: Selection and Causation; Mortality of the Oldest-Old; Twin Studies in Demography.

Bibliography

Arbeev, K.G., Ukraintseva, S.V., Arbeeva, L.S., Akushevich, I., Kulminski, A.M., Yashin, A.I., 2011. Evaluation of genotype-specific survival using joint analysis of genetic and non-genetic subsamples of longitudinal data. Biogerontology 12, 157–166.

Bergman, A., Atzmon, G., Ye, K., Maccarthy, T., Barzilai, N., 2007. Buffering mechanisms in aging: a systems approach toward uncovering the genetic component of aging. PLoS Computational Biology 3, e170.

Di Rienzo, A., 2006. Population genetics models of common diseases. Current Opinion in Genetics and Development 16, 630–636.

Drenos, F., Kirkwood, T.B., 2005. Modelling the disposable soma theory of ageing. Mechanisms of Ageing and Development 126, 99–103.

Gagnon, A., Smith, K.R., Tremblay, M., Vezina, H., Pare, P.P., Desjardins, B., 2009. Is there a trade-off between fertility and longevity? A comparative study of women from three large historical databases accounting for mortality selection. American Journal of Human Biology 21, 533–540.

Jeck, W.R., Siebold, A.P., Sharpless, N.E., 2012. Review: a meta-analysis of GWAS and age-associated diseases. Aging Cell 11, 727–731.

Kulminski, A., Ukraintseva, S.V., Akushevich, I., Arbeev, K.G., Land, K., Yashin, A.I., 2007. Accelerated accumulation of health deficits as a characteristic of aging. Experimental Gerontology 42, 963–970.

Lycett, J.E., Dunbar, R.I., Voland, E., 2000. Longevity and the costs of reproduction in a historical human population. Proceedings of the Royal Society B: Biological Sciences 267, 31–35.

Manton, K.G., Woodbury, M.A., Stallard, E., 1995. Sex differences in human mortality and aging at late ages: the effect of mortality selection and state dynamics. Gerontologist 35, 597–608.

Mitnitski, A.B., Mogilner, A.J., Rockwood, K., 2001. Accumulation of deficits as a proxy measure of aging. Scientific World Journal 1, 323–336.

Nebel, A., Kleindorp, R., Caliebe, A., Nothnagel, M., Blanche, H., Junge, O., Wittig, M., Ellinghaus, D., Flachsbart, F., Wichmann, H.E., Meitinger, T., Nikolaus, S., Franke, A., Krawczak, M., Lathrop, M., Schreiber, S., 2011. A genome-wide association study confirms APOE as the major gene influencing survival in long-lived individuals. Mechanism of Ageing and Development 132, 324–330.

Pritchard, J.K., 2001. Are rare variants responsible for susceptibility to complex diseases? American Journal of Human Genetics 69, 124–137.

Reich, D.E., Lander, E.S., 2001. On the allelic spectrum of human disease. Trends in Genetics 17, 502–510.

Sierra, F., Hadley, E., Suzman, R., Hodes, R., 2009. Prospects for life span extension. Annual Review of Medicine 60, 457–469.

Stallard, E., 2002. Underlying and multiple cause mortality at advanced ages: United States 1980–1998. North American Actuarial Journal 6, 64–87.

Strehler, B.L., Mildvan, A.S., 1960. General theory of mortality and aging. Science 132, 14–21.

Ukraintseva, S.V., Arbeev, K.G., Akushevich, I., Kulminski, A., Arbeeva, L., Culminskaya, I., Akushevich, L., Yashin, A.I., 2010. Trade-offs between cancer and other diseases: do they exist and influence longevity? Rejuvenation Research 13, 387–396.

Van Bodegom, D., May, L., Meij, H.J., Westendorp, R.G., 2007. Regulation of human life histories: the role of the inflammatory host response. Annals of the New York Academy of Science 1100, 84–97.

Vaupel, J.W., Carey, J.R., Christensen, K., Johnson, T.E., Yashin, A.I., Holm, N.V., Iachine, I.A., Kannisto, V., Khazaeli, A.A., Liedo, P., Longo, V.D., Zeng, Y., Manton, K.G., Curtsinger, J.W., 1998. Biodemographic trajectories of longevity. Science 280, 855–860.

Vaupel, J.W., Manton, K.G., Stallard, E., 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16, 439–454.

Vaupel, J.W., Yashin, A.I., 1987. Repeated resuscitation: how lifesaving alters life tables. Demography 24, 123–135.

Wienke, A., 2010. Frailty Models in Survival Analysis. Chapman & Hall/CRC, Boca Raton, FL, USA.

Woodbury, M.A., Manton, K.G., 1977. A random-walk model of human mortality and aging. Theoretical Population Biology 11, 37–48.

Yashin, A.I., Akushevich, I., Arbeev, K.G., Kulminski, A., Ukraintseva, S., 2011. Joint analysis of health histories, physiological states, and survival. Mathematical Population Studies 18, 207–233.

Yashin, A.I., Ukraintseva, S.V., De Benedictis, G., Anisimov, V.N., Butov, A.A., Arbeev, K., Jdanov, D.A., Boiko, S.I., Begun, A.S., Bonafe, M., Franceschi, C., 2001. Have the oldest old adults ever been frail in the past? A hypothesis that explains modern trends in survival. Journals of Gerontology Series A, Biological Sciences and Medical Sciences 56, B432–B442.

Yashin, A.I., Vaupel, J.W., Iachine, I.A., 1994. A duality in aging: the equivalence of mortality models based on radically different concepts. Mechanism of Ageing and Development 74, 1–14.

Yashin, A.I., Wu, D., Arbeev, K.G., Stallard, E., Land, K.C., Ukraintseva, S.V., 2012. How genes influence life span: the biodemography of human survival. Rejuvenation Research 15, 374–380.

Mortality Crossover

Irma T Elo, Population Studies Center, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mortality crossovers refer to a phenomenon where one population or population subgroup has higher age-specific death rates at younger ages but lower age-specific death rates at older ages than a comparison population or population subgroup. Mortality crossovers have been documented in the United States between whites, African-Americans, Hispanics and Native Americans, and among developed and developing countries. Such mortality crossovers are attributed to mortality selection and population heterogeneity or to age misreporting at advanced ages.

A mortality crossover is said to occur when mortality curves of two populations or population subgroups intersect. At younger ages, age-specific death rates for one population exceed those of the other, with a gradual narrowing of differentials as age advances, until at some older age the death rate differentials are reversed. Such intersection of mortality curves has provided empirical support for theories of mortality selection and 'the survival of the fittest' mechanism in old age mortality. Others have argued that the observed mortality crossovers are simply a consequence of age misreporting.

Empirical Evidence

One of the most widely noted mortality crossovers is that observed for African-Americans and whites in the United States. Throughout the twentieth century, mortality estimates based on vital statistics and census data have consistently shown black death rates to exceed white rates until some age above the mid-70s at which point white rates exceed black age-specific death rates. The greatest relative disadvantage for African-Americans is observed in middle age, and the age at which the two rates cross has increased over time (Pearl, 1922; Lynch et al., 2003). Investigations of black–white mortality differentials by cause of death have also revealed a crossover in cause-specific death rates, although the relative differentials by age vary substantially by cause of death, with mortality crossovers observed for some but not for all causes of death (Eberstein et al., 2014; Hussey and Elo, 1997; Elo and Preston, 1997; Manton et al., 1979; Manton and Stallard, 1997).

Evidence from linked data sources, where information for deaths and population at risk comes from a single source, also shows a black–white mortality crossover at the oldest ages. Kestenbaum (1992) has conducted a careful analysis of mortality based on Medicare and Social Security data for black and white Americans at ages 85 and above. Black mortality exceeded or was the same as white mortality up to age 87 for males and up to age 88 for females, at which ages the rates crossed over. Similarly, mortality estimates based on the National Longitudinal Mortality Study, the National Health Interview Survey linked to the National Death Index, the American Changing Lives study, and small-scale, longitudinal data sources show a crossover in black and white death rates at the oldest ages (Dupre et al., 2006; Eberstein et al., 2014; Elo and Preston, 1997; Masters, 2012; Sautter et al., 2012; Yao and Robert, 2011).

A mortality crossover is also observed for Hispanics and non-Hispanic whites and Native Americans and whites in the United States, but at a younger age than for blacks and whites (Elo and Preston, 1997; Thornton, 2004). For example, Hispanic Americans have higher mortality than non-Hispanic whites at younger ages and lower mortality after about age 40–45 with the proportionate gap between Hispanic and white rates increasing with age. At least some of the Hispanic advantage appears to be due to the exceptionally low mortality of the foreign-born. Hispanic populations in Latin America are also known for having unusually low death rates at older ages relative to their level of mortality at younger ages (Elo and Preston, 1997; Dechter and Preston, 1991).

Empirically observed mortality crossovers are not limited to the United States. Coale and Kisker (1986) compared death rates in selected life tables for developed and developing countries with a similar life expectancy at age five. In each pair of life tables considered, death rates at younger ages in the developing country exceeded those of the developed country, while at the oldest ages the relative differentials were reversed. Nam et al. (1978) investigated mortality crossovers by comparing life table death rates for pairs of national populations where one population was considered low and the other high income. Of the 1035 pairs of mortality curves examined, for 16% of the original pairs the authors' findings were consistent with the hypothesis that the more disadvantaged lower income populations would have higher mortality at younger ages but lower mortality at older ages.

Manton and Vaupel (1995) also documented a crossover in death rates at the very oldest ages between white Americans and populations in England, France, Sweden, and Japan. In Japan and the European countries, life expectancy at birth was notably higher than in the United States. Death rates, however, began to converge around age 65, and above age 80 the U.S. rates fell below those of the European countries and Japan. More recently, Ho and Preston (2010) similarly showed that although U.S. male and female age-specific death rates in 2005 were among the highest among 17 OECD countries between ages 40 and 70 for males and 40 and 75 for females, at older ages the U.S. death rates rose less sharply than in most comparison countries and above age 90 the U.S. death rates were among the lowest.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31045-5

Proposed Explanations

Most observers attribute mortality crossovers either to selective survival and population heterogeneity or to poor data quality at the oldest ages.

Selective Survival and Population Heterogeneity

Manton and Vaupel have been in the forefront of advancing theories involving selective survival and population heterogeneity (Vaupel et al., 1979; Manton and Stallard, 1981; Manton et al., 1979). In a heterogeneous population, where individuals have different susceptibilities to mortality that are fixed at birth, frailer members of a cohort die at younger ages leaving a more robust group of survivors at successively older ages. If we further consider two population subgroups that have the same distribution of frailty, but one is subject to more adverse conditions in early life then its frailest members are eliminated more quickly than among the advantaged group. If such selection is appropriately balanced, it is theoretically possible that a crossover of the mortality curves for the two population subgroups is observed at some advanced age (Manton et al., 1979; Olshansky, 1995).

Although the above mechanism is feasible, it requires that genetic endowments for longevity are fixed at birth and remain constant over the life span. However, life style and environmental factors are likely to influence individual as well as population longevity. To address this possibility, Manton and his colleagues have extended the selection hypothesis to acknowledge the potential impact of environmental factors on differential rates of population aging (Manton et al., 1979). It has been further shown that mortality crossovers between two populations can occur even when the age pattern of mortality is the same for all individuals, but the distribution of frailty varies among them. Mortality will rise less rapidly with age in a population with a greater degree of heterogeneity at the same mean level of frailty (Vaupel, 2010 cited in Ho and Preston, 2010).

The crossover in black and white death rates at older ages has provided much of the empirical support for the mortality selection and survival of the fittest explanations. Given large socioeconomic differentials between whites and blacks and the adverse conditions African-Americans experienced earlier this century, it is conceivable that selective survival has been more pronounced for blacks than for whites. Evidence from other populations, however, suggests that cohorts who experienced unfavorable conditions in early life also tend to experience elevated mortality at older ages (Elo and Preston, 1992; Finch and Crimmins, 2005; Janssen et al., 2005). A handful of studies have used individual and/or neighborhood level socioeconomic conditions and religious attendance as sources of population heterogeneity to examine whether these sources of heterogeneity can help explain the black–white mortality crossover with mixed results (e.g., Dupre et al., 2006; Sautter et al., 2012; Yao and Robert, 2012). A recent study by Masters (2012) in turn points to a potentially important role of black–white differences in cohort patterns of mortality in the black–white mortality crossover. Using hierarchical age-period cohort methods, the author finds that non-Hispanic black and non-Hispanic white age-specific mortality rates do not crossover at any age, controlling for period and cohort patterns of adult mortality.

Selective survival and population heterogeneity have also been invoked as a possible explanation for the low mortality in the United States relative to other OECD countries at the oldest ages. Other potential explanations for this pattern include cohort differences in education, immigration, access to health care, and health behaviors as well as more extensive medical interventions in the United States than in other OECD countries at the oldest ages (Manton and Vaupel, 1995; Ho and Preston, 2010).

Poor Quality Data and Age Misreporting

The second explanation for mortality crossovers is that they are produced by poor data quality. It is hypothesized that age misreporting in death statistics and/or censuses biases mortality estimates downward at the oldest ages. Age overstatement has been considered the most common form of age misreporting, although other patterns have also been recorded (Ewbank, 1981; Preston et al., 1999). Preston et al. (1999) have further shown that various patterns of age misreporting – net age overstatement, net age understatement, and symmetric age misreporting – bias mortality estimates downward at the oldest ages. The downward bias is produced by all methods of mortality estimation the authors considered: conventional methods derived from vital statistics and censuses; longitudinal studies where age is identified at baseline; variable-*r* procedures based on age distribution of the population; variable-*r* procedures based on age distribution of deaths; and extinct generation methods. Thus, it is not unreasonable to propose that age misreporting at the oldest ages has led to downwardly biased death rates at the oldest ages and spurious mortality crossovers.

In their study of life tables for developed and developing countries, Coale and Kisker (1986) concluded that the mortality crossovers they observed were caused by poor data quality. Age misreporting has also been implicated in the low mortality of Hispanics at the older ages relative to the U.S. white population. Other potential explanations for the low mortality of Hispanics include the positive selection of migrants on such characteristics as healthiness, difficulty of following the foreign-born in U.S. death records, and return migration to countries of origin by those in poor health, i.e., the salmon bias hypothesis (Abraido-Lanza et al., 1999; Elo and Preston, 1997). The lower mortality of Hispanics compared to the non-Hispanic whites, however, has been observed not only in vital statistics and census data but also in Social Security and Medicare data where age misreporting is less likely to be present (Elo et al., 2004). Turra and Elo (2008) have further shown that although salmon bias is likely to play a role in the Hispanic mortality advantage at the older ages, it is too small in magnitude to be a primary explanation for the lower mortality of Hispanics than non-Hispanic whites in the United States at ages 65 and above.

The most extensive studies of the potential impact of age misreporting on mortality crossovers pertain to the black–white mortality crossover in the Unites States. There is a general agreement that age misreporting has seriously biased African-American death rates downward at the oldest ages. The first study to document inconsistencies in age reporting for

African-Americans and whites in vital statistics and census data was the 1960 Matched Records Study, which linked death certificates registered in May to August in 1960 to the 1960 Census of Population. For less than half of the nonwhite male and female matched cases was the age in the two sources identical. Age reporting was far better for whites. After the data were corrected for age misreporting, the racial crossover in mortality moved from the age interval 75–79 to the last open-ended age interval, 85 and older (Kitagawa and Hauser, 1973).

Two recent matched-record studies of death certificates from the 1980s for whites and African-Americans to early census records for these same individuals when they were children or adolescents continue to show substantial age misreporting for African-Americans in death statistics at the oldest ages, but minimal problems for whites. Thus, conventionally constructed death rates from vital statistics and census data for African-Americans continue to be understated at the oldest ages. In contrast, death rates for whites do not show a systematic bias. A comparison of mortality rates for blacks and whites based on corrected data continues to reveal a mortality crossover for males in the age interval 95+ and for females at ages 90–94 (Preston et al., 1996; Hill et al., 2000). Age-specific death rates for this comparison were, however, based on somewhat unconventional variable-*r* procedures and on a very small number of deaths for African-Americans at the oldest ages. Lynch et al. (2003) also found that the age of black–white mortality crossover was at a later age when estimates were based on data adjusted for possible age misreporting than on unadjusted data. The affirmation of the reliability of mortality estimates for whites lends support to the findings by Manton and Vaupel (1995) and Ho and Preston (2010) that white mortality in the United States at the oldest ages is lower than in most OECD countries.

The above findings raise compelling questions about the possible underlying factors for various observed mortality crossovers. It is unlikely that a single explanation will account for them or for the convergence of death rates at the oldest ages under all circumstances. To gain greater insights into the variability in the aging process, it is important to explain not only those instances where crossovers occur, but also when they have failed to materialize.

See also: Demographic Techniques: Data Adjustment and Correction; Differential Aging; Early Life Influences on Health and Mortality in Adulthood; Immigration: Health and Mortality; Mortality Differentials: Selection and Causation; Period and Cohort Analysis in Demography.

Bibliography

Abraido-Lanza, A.F., Dohrenwend, B.P., Ngmak, D.S., Turner, J.B., 1999. The Latino mortality paradox: a test of the 'salmon bias' and healthy migrant hypotheses (Latino return migration to home country). American Journal of Public Health 89, 1543–1548.

Coale, A.J., Kisker, E.E., 1986. Mortality crossover: reality or bad data? Population Studies 40, 389–401.

Dechter, A., Preston, S.H., 1991. Age misreporting and its effects on adult mortality estimates in Latin America. Population Bulletin of the United Nations 31 (32), 1–17.

Dupre, M.E., Franzese, A.T., Parrado, E.A., 2006. Religious attendance and mortality: implications for the black-white mortality crossover. Demography 43, 141–164.

Eberstein, I.W., Nam, C.B., Heyman, K.M., 2014. Causes of death and mortality crossovers by race. Biodemography and Social Biology 54, 214–228. http://dx.doi.org/10.1080/19485565.2008.9989143.

Elo, I.T., Preston, S.H., 1992. Effects of early-life conditions on adult mortality: a review. Population Index 58 (2), 186–212.

Elo, I.T., Preston, S.H., 1997. Racial and ethnic differences in mortality at older ages. In: Martin, Linda G., Soldo, Beth J. (Eds.), Racial and Ethnic Differences in Health of Older Americans. National Academy Press, Washington, DC, pp. 10–42.

Elo, I.T., Turra, C.M., Kestenbaum, B., Ferguson, B.R., 2004. Mortality among elderly Hispanics in the United States: past evidence and new results. Demography 41, 109–128.

Ewbank, D.C., 1981. Age Misreporting and Age Selective Underenumeration: Sources, Patterns, and Consequences for Demographic Analysis. Committee on Population and Demography, report No. 4. National Academy Press, Washington, DC.

Finch, C., Crimmins, E., 2005. Inflammatory exposure and historical changes in human life span. Science 305, 1736–1739.

Hill, M.E., Preston, S.H., Rosenwaike, R., 2000. Age reporting among white Americans aged 85+: results of a record linkage study. Demography 37, 175–186.

Ho, J.Y., Preston, S.H., 2010. US mortality in an international context: age variations. Population Development and Policy Review 36, 749–773.

Hussey, J.M., Elo, I.T., 1997. Cause-specific mortality among older African Americans; correlates and consequences of age misreporting. Social Biology 44, 227–246.

Janssen, F.A., Peeters, A., Mackenbach, J.P., Kunst, A.E., 2005. Relation between trends in late middle age mortality and trends in old age mortality – is there evidence for mortality selection? Journal of Epidemiology and Community Health 59, 775–781.

Kitagawa, E.M., Hauser, P.M., 1973. Differential Mortality in the United State: A Study of Socioeconomic Epidemiology. Harvard University Press, Cambridge, MA.

Kestenbaum, B., 1992. A description of the extreme aged population based on improved medicare enrollment data. Demography 29, 565–580.

Lynch, S.M., Brown, J.S., Harmsen, K.G., 2003. Black-white differences in mortality compression and deceleration and the mortality crossover reconsidered. Research on Aging 25, 456–483. http://dx.doi.org/10.1177/0164027503254675.

Manton, K.G., Stallard, E., 1981. Methods for evaluating heterogeneity of aging processes in human populations using vital statistics data: explaining the black/white mortality crossover by a model of mortality selection. Human Biology 53, 47–67.

Manton, K.G., Stallard, E., 1997. Health and disability differences among racial and ethnic groups. In: Martin, Linda G., Soldo, Beth J. (Eds.), Racial and Ethnic Differences in Health of Older Americans. National Academy Press, Washington, DC, pp. 43–105.

Manton, K.G., Vaupel, J.W., 1995. Survival after age of 80 in the United States, Sweden, France, England, and Japan. New England Journal of Medicine 333, 1232–1235.

Manton, K.G., Poss, S.S., Wing, S., 1979. The black/white mortality crossover: investigation from the perspective of the components of aging. The Gerontologist 19 (3), 291–300.

Masters, R.K., 2012. Uncrossing the U.S. black-white mortality crossover: the role of cohort forces in life course mortality risk. Demography 49, 773–796. http://dx.doi.org/10.1007/s13524-012-0107-y.

Nam, C.B., Weatherby, N.L., Ockay, K.A., 1978. Causes of death which contribute to the mortality crossover effect. Social Biology 25 (4), 306–330.

Olshansky, S.J., 1995. Introduction: new developments in mortality. The Gerontologist 35 (5), 583–587.

Pearl, R., 1922. The Biology of Death. J.B. Lippincott, Philadelphia, PA.

Preston, S.H., Elo, I.T., Stewart, Q., 1999. Effects of age misreporting on mortality estimates at older ages. Population Studies 53, 165–177.

Preston, S.H., Elo, I.T., Rosenwaike, I., Hill, M., 1996. African-American mortality at older ages: results of a matching study. Demography 33 (2), 193–209.

Sautter, J.M., Thomas, P.A., Dupre, M.E., George, L.K., 2012. Socioeconomic status and the black-white mortality crossover. American Journal of Public Health 102, 1566–1571.

Thornton, R., 2004. The Navajo-U.S. population mortality crossover since the mid-20th century. Population Research and Policy Review 23, 291–308.

Turra, C.M., Elo, I.T., 2008. The impact of salmon bias on the hispanic mortality advantage: new evidence from social security data. Population Research and Policy Review 27, 515–530.

Vaupel, J.W., 2010. The Theory of Heterogeneity: A Concise Primer. Manuscript, Max Planck Institute for Demographic Research. Germany, Rostock.

Vaupel, J.W., Manton, K.G., Stallard, E., 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16 (3), 439–454.

Yao, L., Robert, S.A., 2011. Examining the racial crossover in mortality between African American and white older adults: multilevel survival analysis of race, individual socioeconomic status and neighborhood socioeconomic context. Journal of Aging Research. http://dx.doi.org/10.4061/2011/132073.

Mortality Differentials: Selection and Causation

Noreen Goldman, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Differentials in mortality by socioeconomic status and the nature of social relationships have been well established in a broad range of settings and time periods. A myriad of studies has investigated the many potential causal pathways linking these aspects of the social environment to health and mortality, with a current emphasis on psychosocial factors, neighborhood contexts, and biological linkages. Recent efforts have exploited longitudinal surveys, some beginning at birth, to strengthen causal inferences. Fewer researchers have focused on identifying selection mechanisms or reverse causal pathways through which, for example, poor health may lead to reduced income or exclusion from marriage.

Differentials in mortality by socioeconomic status (SES) and the nature of social relationships have been well established. In general, persons of higher SES (e.g., those with higher income, education, and occupational class) and persons who are more socially integrated (e.g., those who are married) experience lower death rates than their respective counterparts. These associations have been found across time and place – including industrialized and developing countries – although with some exceptions. Three sets of hypotheses have been proposed by researchers to explain these patterns. The first encompasses *causal* mechanisms through which SES and social relationships potentially affect health status and the risk of dying. A second type of explanation, referred to as *selection* or *reverse causation*, pertains to a set of pathways whereby unhealthy individuals may reduce their SES or become less socially connected as a consequence of their inferior health status. A third explanation relates to *artifactual* or *spurious* mechanisms, such as measurement error.

SES and Mortality

Differences in rates of morbidity and mortality by SES have been identified for the vast majority of illnesses, along a social gradient: persons of higher SES fare better than those lower in the social hierarchy (Adler et al., 1993). These differences typically (1) are present regardless of whether SES is defined in terms of income, wealth, education, or occupation; (2) exist throughout the social hierarchy (i.e., not simply between the poor and the nonpoor); and (3) have been documented in a large number of studies in both industrialized and developing countries, based on vital registration, census, and survey data. There are exceptions, however. For example, there is no clear mortality gradient among older adults in Costa Rica (Rosero-Bixby and Dow, 2009) and some Hispanic groups in the US have relatively shallow gradients in health and mortality (Goldman et al., 2006; Turra and Goldman, 2007).

Many of the studies examining the link between SES and health have been based in Western populations, especially the United States and Britain. Two long-term investigations in Britain have been used extensively to examine social differentials in health. One is the Whitehall Study, which tracked a sample of British civil servants since 1967 (Whitehall I) and another sample since 1985 (Whitehall II) (Marmot et al., 1995; Marmot and Brunner, 2005). The original Whitehall study is frequently credited with establishing the existence of a gradient in health – e.g., the study demonstrated that mortality differentials are present throughout successive grades of employment even within a relatively homogeneous group: civil servants in stable office jobs in London. Whitehall II has the explicit mission of examining social gradients in health and illness. A second set of frequently analyzed data comprises four large national birth cohort studies: three that have been following births that occurred in England, Scotland, and Wales during 1-week periods in 1946, 1958, and 1970 and a fourth (the Millennium Cohort Study) that has been following births in the 2000–01 period.

Many recent investigations have focused on assessing the extent to which causal rather than selective or spurious mechanisms drive the observed socioeconomic inequalities in health. To reduce the impact of bias in the statistical analysis, researchers have relied largely on prospective (longitudinal) data, such as those described above. Although many of these studies provide some control for selection-type biases, relatively few studies study such biases explicitly. There is a general consensus among epidemiologists and health researchers that the observed disparities in health are not driven by *social selection*, also referred to as *social drift*, or *reverse causality*. Specifically, while there is some downward drift in SES among persons in poor health (e.g., schizophrenics may lose income or their jobs because of their health), many health researchers argue that this type of mechanism is unlikely to be an important part of the association between SES and a wide range of health indicators (see, for example, Davey Smith et al., 1994; Marmot et al., 1995; Macintyre, 1997).

In contrast, economists typically recognize the potential importance of reverse causality. For example, Smith (2004) uses data from national longitudinal surveys in the US (the Health and Retirement Study and the Panel Study of Income Dynamics) to examine the impact of new serious health events among older adults and concludes that such health shocks have a substantial impact on work, income, and wealth. In a study examining the link between children's health and social mobility, Case and Paxson (2006) demonstrate that health problems during childhood can affect economic success in adulthood, partly through poor educational achievement.

Various studies – particularly those examining reverse causal processes – underscore the importance of distinguishing among the different components of SES when examining linkages with health – e.g., among adults, poor health may reduce income and wealth but it is unlikely to impact years of schooling. The effects of poor health on social status are also likely to vary by the nature of the illness or condition, with stronger associations apparent for some mental illnesses and severely disabling diseases. For example, schizophrenia and epilepsy have been linked to the reduced social mobility (Goldberg and Morrison, 1963; Harrison and Taylor, 1976).

It is also possible that the associations between SES and health arise, at least in part, from *indirect selection* mechanisms, i.e., background (or *spurious*) factors, such as height, genetic characteristics, parental behavior, and self-efficacy, that affect both health status and social mobility. Unfortunately, many of these factors are difficult to measure and some are themselves the result of social status. Artifactual mechanisms (e.g., errors of measurement, such as census undercounts, 'numerator–denominator' problems, such as inconsistencies in reports between registration and census data, or inappropriate measures of mortality or SES) provide another possible explanation of the observed associations. However, increasing reliance on longitudinal surveys rather than cross-sectional data has reduced concerns about most of these methodological issues.

The strong belief in the causal linkages between socioeconomic position and health has led some social science and health researchers to view SES as a 'fundamental cause' of disease (Link and Phelan, 1995). The underlying idea is that social inequalities in health have persisted across time and place, despite radical shifts in the disease environment, risk factors of disease, and technology, because SES always entails access to resources that affect exposure to disease, maintenance of health, and treatment of illness. Because of the endurance of these health inequalities, Link and Phelan (1995) discourage research that focuses on the 'proximate determinants' of illness – i.e., the specific mechanisms that link social position to health or illness. Nevertheless, in an effort to gain a deeper understanding of social inequalities in health, many studies seek to identify the particular causal mechanisms that relate differences in SES to differentials in health and mortality. Possible pathways include differences in (1) access to medical care, both preventive and curative; (2) access to information regarding health risks and health care; (3) patterns of health risk behaviors (such as smoking, drinking, unhealthy diet, and inadequate exercise); (4) exposure to environments that are not conducive to good health and longevity (e.g., poor housing conditions, occupational hazards, pollution, and crime); (5) exposure to stressful situations; (6) access to resources that mediate the physiological consequences of stress; (7) ability to control one's environment and feel secure about one's position; and (8) the availability of social relationships and support.

Researchers who have focused on a single mechanism have concluded that *multiple* causal pathways almost certainly underlie the observed associations. For example, although the higher death rates experienced by persons of lower SES in several British and American studies have been shown to result in part from higher rates of smoking, poorer diets, and inadequate levels of exercise, these differences in behavior were insufficient to account for the observed mortality differentials (Marmot et al., 1995; Macintyre, 1997). Similarly, the widening of mortality differentials by SES between the 1960s and 1980s in Britain and the US – despite the presence of the National Health Service in the former and Medicaid in the latter – has been taken as evidence that the explanation goes beyond inequities in access to medical care (Davey Smith et al., 1990; Pappas et al., 1993). More recent studies demonstrate a continued widening of social inequalities in health in the US and Europe (Mackenbach et al., 2003; Marmot, 2003; Singh and Siahpush, 2006), but reasons for this temporal pattern are not well understood.

In recent years, researchers have broadened the theoretical framework underlying the study of mortality differentials, devoting increasing attention to the following types of factors causally linking social status to health: (1) psychosocial influences; (2) neighborhood context; and (3) biological pathways (Chen and Miller, 2013). For example, studies have examined how different occupations or social positions vary with regard to stressful experiences, anxiety, mastery, depression, discrimination, and the ability of individuals to control their lives (Marmot, 1999). Other research has focused on individuals' broader social environment, considering, for example, racial and ethnic composition, neighborhood violence, and neighborhood social capital. Some researchers have argued that an individual's health is related not only to his or her absolute income or family income, but also to the level of income inequality in the society, with higher degrees of inequality associated with poorer health and survival (Wilkinson, 1996), However, the evidence relating societal income inequality to population health is plagued by methodological issues and poor data quality and is inconclusive (Subramanian and Kawachi, 2004; Lynch et al., 2004). Another advance has been the incorporation of physiological measurements, often referred to as biomarkers, into social science surveys. These new data collection efforts have allowed individuals to begin to elucidate the biological mechanisms that underlie social disparities in health and mortality, considering standard clinical factors (such as blood pressure, cholesterol levels, or blood glucose levels), markers of inflammation and immune function, and neuroendocrine mechanisms (Karlamangla et al., 2012; Turra et al., 2005). Researchers often describe such efforts as 'getting under the skin' of social disparities in health. Biomarkers have also been used to explore the potential role of exposure to chronic stressful experience – via increased 'allostatic load' – in mediating the relationship between SES and health (McEwen, 1998).

Most research linking SES to health relies on what epidemiologists refer to as observational studies. As noted above, the most useful of these studies have been prospective surveys that collect extensive information on the social environment, health, and survival. These data generally allow the analyst to measure preexisting conditions that may confound the association between SES and health, to relate changes in SES to subsequent changes in health, and to control for potentially confounding variables. Given increasing recognition of the influences of social factors on health throughout the life cycle, prospective surveys that begin early in life are especially useful. A myriad of studies has established that early life experiences related to the socioeconomic position of the family – e.g., family composition, parent–child interactions, the presence of

child abuse – can have wide-ranging effects on physical and mental well-being in adulthood (see, for example, Wadsworth and Kuh, 1997; Power and Hertzman, 1999; Elo, 2009). Influences on adult health can begin as early as birth or even the prenatal period (Barker, 1999). Studies of children have demonstrated steeper gradients with the age of the children, suggesting, for example, stronger impacts of family income on children's health as the children grow (Case et al., 2002). Studies have also demonstrated that the magnitude of social disparities in health and survival generally declines at the older ages, due, at least in part, to heterogeneity in the risk of dying by social status (Zajacova et al., 2009).

Despite the richness of prospective surveys, there are of course many impediments to causal inference from observational studies, including confounding, omitted variable bias, and endogeneity. Although experiments – i.e., randomized, controlled, double-blind trials – are the gold standard for causal inference in medicine and epidemiology, they are generally not feasible or ethical, and often not very informative in the social sciences since investigators are constrained by the types of social conditions that can be manipulated. Nevertheless, there have been experiments, quasi-experiments (which do not entail randomization but involve researchers' control over the intervention), and natural experiments (in which treatment allocation is not under researchers' control but seems to be haphazard or close to random) in this field of study. For example, using findings from an experimental demonstration project known as Moving to Opportunity, researchers examined the impact of randomly assigned housing vouchers, which allowed residents of public housing to move to lower poverty areas, on obesity and diabetes (Ludwig et al., 2011). In an intervention in which exposure to illness (by inhalation of the virus), rather than social position, was assigned by the investigators, Cohen and colleagues (Cohen et al., 2008) analyzed the association between SES and development of viral infection. Economists have used natural experiments to identify causal impacts of SES on health by examining whether changes in health follow an external change in policy (e.g., compulsory schooling; Lleras-Muney, 2005). Other studies have considered the impact on health of such seemingly exogenous events as the reunification of Germany, increased income from casinos among American Indians, and winning a lottery (Evans et al., 2012). Some quasi-experimental studies that examine links between social position and health involve animals, particularly nonhuman primates. For example, Shively and colleagues manipulated the relative social status (dominant vs subordinate position) of monkeys and assessed the subsequent development of atherosclerosis (Shively and Clarkson, 1994). Despite their usefulness at reducing the threats to causal inference that characterize observational studies, these experimental-type investigations – whether on animals or humans – cannot examine the complex set of interrelated pathways, with multiple moderating influences, that operate in actual social environments to impact human health (Goldman, 2001).

Social Relationships and Mortality

A related area of research has examined the associations between *social relationships* and health and mortality. These undertakings have concluded that persons who are less socially integrated have higher rates of illness and lower life expectancy than their counterparts. The associations are present for a wide range of causes of death. The possible interpretations of these findings mirror those described above, namely causal vs selection-type arguments. Causal explanations encompass a broad set of pathways through which the availability of social connections or social support can, for example, protect persons from health risks and unhealthy behaviors (e.g., smoking and poor diets), provide access to health-related information, increase individuals' sense of control and self-esteem, and buffer responses to stressful experiences. Selection-type arguments posit that people suffering from mental or physical illnesses are less likely to establish and/or maintain social relationships or that some other variables reduce a person's ability to have positive relationships and maintain good health.

As with studies of SES, prospective studies have provided much more convincing evidence than cross-sectional or retrospective ones regarding the importance of causal mechanisms in producing the association between social relationships and health. For example, a set of prospective community studies carried out in the 1970s and 1980s demonstrated that levels of social integration – a measure comprised of various indicators of social ties and activities – were positively and strongly associated with survival among adults (House et al., 1988). In a recent meta-analysis based on 148 studies involving more than 300 000 individuals followed for an average of 7.5 years, Holt-Lundstad and colleagues concluded that persons with adequate social relationships had a 50% greater probability of surviving than those with inadequate relationships. They argued that this magnitude of effect surpassed that for many well-known mortality risks, including obesity and lack of physical activity (Holt-Lunstad et al., 2010).

Findings from surveys have been strengthened by experimental and clinical research, which has demonstrated that variations in the degree of social contact have psychological and physiological effects in animals and humans (House et al., 1988). For example, partly by reducing individuals' negative appraisal of stressful experiences, social relationships are thought to improve cardiovascular and immune functioning (Uchino et al., 1999). Similarly, Cacioppo and colleagues have shown that loneliness – in particular, subjective feelings of loneliness – impacts the sympathetic nervous system, the immune system, and the brain (Miller, 2011). A large number of studies have demonstrated that even simple interventions – such as the presence or petting of companion animals – can improve the physical and mental well-being of adults (Wells, 2011).

While the availability of social ties and support provided by these individuals is generally thought to have a positive influence on well-being, not all relationships are beneficial. For example, through a series of experimental-type studies, researchers have demonstrated that married couples experiencing conflict and hostile behavior experienced dysregulation in cardiovascular and immune function (Robles and Kiecolt-Glaser, 2003).

Marital Status and Mortality

Marriage is often considered one of the strongest sources of social support. The earliest research on social connection and

health – dating as far back as the mid-1800s – focused largely on marital status. Since that time, hundreds of studies, primarily in industrialized countries, have demonstrated that married persons have greater longevity and experience better health than single, divorced, and widowed persons. These results generally hold across time and place. In most studies, differentials in mortality by marital status have been larger for men than for women, although a recent meta-analysis found no sex difference in the magnitude of the effects for the older population (Manzoli et al., 2007). The relative rankings of the different unmarried states in terms of mortality risk vary across studies, but Rendall and colleagues argue that many claims of differences across the unmarried groups are not supported by statistical tests; they demonstrate that, in a large US study, there is little evidence of survival differences among single, widowed, and divorced/separated persons (Rendall et al., 2011). As with the broader field of social disparities in health, studies of marital status differentials in mortality are relying increasingly on prospective rather than cross-sectional data.

Once again, the dominant hypotheses for the better survival profile of married persons comprise causal vs reverse causal/selection explanations. Causal hypotheses, often termed *marriage protection* in this literature, encompass a broad range of social, psychological, economic, and environmental benefits that are presumed to be associated with having a spouse. Many of these are similar to the mechanisms described above hypothesized to link social connection to health and survival. For example, the presence of a spouse is hypothesized to (1) result in greater emotional and social support; (2) lead to higher income; (3) facilitate access to medical information and health services; (4) constrain risk-taking behavior and encourage healthier lifestyles; (5) act as a buffering mechanism in the presence of stress; and (6) substitute for inferior or inaccessible formal health care. Moreover, the termination of marriage – either by divorce or death of a spouse – is believed to lead to higher risks of dying because of the stresses associated with the loss of a spouse (e.g., *bereavement effects* in the case of widowhood) and the accompanying decline in social and economic resources. However, as suggested earlier, differences across the marital states in terms of health impacts are not always clear-cut: marriage can lead to spousal conflict, which has a negative impact on health; in such cases, divorce may have a more positive impact on health than an unhappy marriage.

The health consequences of being unmarried are likely to vary across societies as a function of social and cultural factors. For example, in societies where marriage is an integral part of one's social status, and is virtually universal, persons (especially women) who remain single or get divorced are likely to face discrimination and the concomitant negative consequences for their health.

Marriage selection arguments, by contrast, are based on the premise that mentally and physically healthier persons are more likely to marry in the first place, less likely to become widowed or divorced, and more likely to remarry if a previous marriage dissolves. The factors that drive the selection criteria may be far from obvious. For example, spouses may be selected for better health not only through the direct exclusion of mentally and physically ill persons from first marriage or remarriage, but also through a range of criteria that govern mate selection (i.e., *indirect selection*) such as income, physical appearance (e.g., height, weight, and beauty), health-related habits (e.g., smoking and drinking), and emotional stability.

In line with studies of SES and social connection, analyses focusing on marital status have typically concluded that the associations with mortality are produced predominantly by various benefits associated with having a spouse and not by marital selection. While this conclusion may generally be correct, many of the inferences drawn by researchers have not been justifiable. In addition, selection appears to be the dominant source of the observed differentials in at least one population (described below).

Two types of problems of inference have been noted. The first is that, in the past, demographers often based their arguments on cross-sectional data. In particular, some demographers have reached conclusions about the importance of marital selection on the basis of aggregate patterns of mortality by marital status, e.g., age patterns of mortality differentials and the direction and strength of the relationship between the magnitude of the mortality differential and the relative size of the single population. Results from a simple simulation model demonstrate that the resulting inferences are not justifiable because both causal and selection processes could lead to the same aggregate patterns of mortality (Goldman, 1993a).

The second type of weakness pertains to unjustifiable inferences derived from prospective data. Although these data have permitted researchers to conclude that the health benefits of marriage (or of higher SES) persist in the presence of controls for selection, they have also led some to argue that selection mechanisms are unimportant. The latter conclusion has generally been unwarranted: existing statistical analyses frequently include controls for selection processes, but most were not designed to *measure the impact* of selection. More recently, some demographers have analyzed longitudinal data in an effort to directly assess health-related selection into first marriage and into divorce. These studies have demonstrated that health-related characteristics and behaviors are used (explicitly or implicitly) as selection criteria for marriage and divorce. For example, in the US, users of illicit drugs, heavy drinkers, and obese persons have lower marriage rates than their respective counterparts (Fu and Goldman, 1996).

Perhaps because demographers (like other social scientists) have been most interested in identifying causal mechanisms, they have often failed to recognize that the impact of selection on mortality differentials by marital status is likely to vary across societies. A study of atypically high mortality among never-married Japanese demonstrates this point clearly. As of the mid-1900s, a 20-year-old never-married Japanese man or woman could expect to live about 15 years less than his or her married counterpart. Data from a variety of sources strongly suggest that the Japanese arranged marriage process appears to be largely responsible for this huge discrepancy in longevity. Several characteristics of arranged marriages are particularly noteworthy in this context: the importance of mental and physical health, particularly conditions thought to be genetically determined, in the choice of a potential spouse; the inclusion of family, friends, go-betweens, and even private detectives in the selection of a mate; and, the efficiency of the mate selection process (Goldman, 1993b). As arranged marriages were replaced by love marriages during the second

half of the twentieth century, the relative mortality rates of never-married Japanese progressively diminished.

Future Directions

Despite extensive research focusing on mortality differentials by SES and by the nature of social relationships, much remains unknown about the relative importance of different causal and noncausal pathways in producing the associations. Two current trends in data collection are likely to increase scientists' understanding of these linkages in the future. The first is the increasing breadth of longitudinal surveys in the social sciences. Some surveys, which follow individuals from the time of birth, are able to capture critical early life influences on subsequent social status and health; with increasing lengths of follow-up, these surveys have amassed information for a substantial part of the life cycle. Using such rich survey data, analysts have been able to take a life course perspective to examine the antecedents and consequences of poor health; explore multilevel influences on health that comprise individual, family, and community-level variables; and identify salient moderators of the links between low SES and poor health. The second innovation is the noninvasive collection of biological and clinical markers as part of social science surveys. By identifying some of the physiological pathways relating aspects of the social environment to physiological outcomes, such as neuroendocrine response, immune function, and cardiovascular function, social scientists may gain additional insights into the causal linkages that underlie the observed mortality differentials. The number of biosocial surveys has increased dramatically in recent years and is likely to continue to do so in the near future.

See also: Early Life Influences on Health and Mortality in Adulthood; Gene–Environment Interplay; Marital Quality and Health; Mortality Crossover; Mortality, Biodemography of; Twin Studies in Demography.

Bibliography

Adler, N.E., Boyce, W., Chesney, M.A., Folkman, S., Syme, S., 1993. Socioeconomic inequalities in health: no easy solution. Journal of the American Medical Association 269, 3140–3145.

Barker, D.J., 1999. Fetal origins of cardiovascular disease. Annals of Medicine 31 (Suppl. 1), 3–6.

Case, A., Paxson, C.H., 2006. Children's health and social mobility. The Future of Children 16 (2), 151–173.

Case, A., Lubotsky, D., Paxson, C., 2002. Economic status and health in childhood: the origins of the gradient. The American Economic Review 92 (5), 1308–1334.

Chen, E., Miller, G.E., 2013. Socioeconomic status and health: mediating and moderating factors. Annual Review of Clinical Psychology 9, 8.1–8.28.

Cohen, S., Alper, C.M., Doyle, W.J., Adler, N., Treanor, J.J., Turner, R.B., 2008. Objective and subjective socioeconomic status and susceptibility to the common cold. Health Psychology 27 (2), 268–274.

Davey, S.G., Bartley, M., Blane, D., 1990. The black report on socioeconomic inequalities in health 10 years on. BMJ: British Medical Journal 301 (6748), 373–377.

Davey, S.G., Blane, D., Bartley, M., 1994. Explanations for socio-economic differentials in mortality evidence from Britain and elsewhere. European Journal of Public Health 4, 131–144.

Elo, I.T., 2009. Social class differentials in health and mortality: patterns and explanations in comparative perspective. Annual Review of Sociology 35, 553–572.

Evans, W., Wolfe, B., Adler, N.E., 2012. The SES and health gradient: a brief review of the literature. In: Wolfe, B., Evans, W., Seeman, T.E. (Eds.), The Biological Consequences of Socioeconomic Inequalities, 2012th ed. Russel Sage Foundation, New York, pp. 1–37.

Fu, H., Goldman, N., 1996. Incorporating health into models of marriage choice: demographic and sociological perspectives. Journal of Marriage & Family 58 (3), 740–758.

Goldberg, E.M., Morrison, S.L., 1963. Schizophrenia and social class. The British Journal of Psychiatry 109 (463), 785–802.

Goldman, N., 1993a. Marriage selection and mortality patterns: inferences and fallacies. Demography 30 (2), 189–208.

Goldman, N., 1993b. The perils of single life in contemporary Japan. Journal of Marriage and Family 55 (1), 191–204.

Goldman, N., 2001. Social inequalities in health. Annals of the New York Academy of Sciences 954 (1), 118–139.

Goldman, N., Kimbro, R.T., Turra, C.M., Pebley, A.R., 2006. Socioeconomic gradients in health for white and Mexican-origin populations. American Journal of Public Health 96 (12), 2186–2193.

Harrison, R.M., Taylor, D.C., 1976. Childhood seizures: a 25-year follow-up. Lancet 1, 948–951.

Holt-Lunstad, J., Smith, T.B., Layton, J.B., 2010. Social relationships and mortality risk: a meta-analytic review. PLoS Medicine 7 (7), 1–20.

House, J.S., Landis, K.R., Umberson, D., 1988. Social relationships and health. Science 241 (4865), 540–545.

Karlamangla, A.S., Gruenewald, T.L., Seeman, T.E., 2012. Promise of biomarkers in assessing and predicting health. In: Wolfe, B., Evans, W., Seeman, T.E. (Eds.), The Biological Consequences of Socioeconomic Inequalities, 2012th ed. Russell Sage Foundation, New York, pp. 38–62.

Link, B.G., Phelan, J., 1995. Social conditions as fundamental causes of disease. Journal of Health & Social Behavior, Spec No, 80–94.

Lleras-Muney, A., 2005. The relationship between education and adult mortality in the United States. The Review of Economic Studies 72 (250), 189–221.

Ludwig, J., Sanbonmatsu, L., Gennetian, L., Adam, E., Duncan, G.J., Katz, L.F., Kessler, R.C., Kling, J.R., Lindau, S.T., Whitaker, R.C., McDade, T.W., 2011. Neighborhoods, obesity, and diabetes – a randomized social experiment. New England Journal of Medicine 365 (16), 1509–1519.

Lynch, J., Smith, G.D., Harper, S., Hillemeier, M., Ross, N., Kaplan, G.A., Wolfson, M., 2004. Is income inequality a determinant of population health? Part 1. A systematic review. Milbank Quarterly 82 (1), 5–99.

Macintyre, S., 1997. The black report and beyond what are the issues? Social Science & Medicine 44 (6), 723–745.

Mackenbach, J.P., Bos, V., Andersen, O., Cardano, M., Costa, G., Harding, S., Reid, A., Hemstrom, O., Valkonen, T., Kunst, A.E., 2003. Widening socioeconomic inequalities in mortality in six Western European countries. International Journal of Epidemiology 32 (5), 830–837.

Manzoli, L., Villari, P., M Pirone, G., Boccia, A., 2007. Marital status and mortality in the elderly: a systematic review and meta-analysis. Social Science & Medicine 64 (1), 77–94.

Marmot, M., 1999. Epidemiology of socioeconomic status and health: are determinants within countries the same as between countries? Annals of the New York Academy of Sciences 896 (1), 16–29.

Marmot, M., Brunner, E., 2005. Cohort profile: the Whitehall II study. International Journal of Epidemiology 34 (2), 251–256.

Marmot, M., Bobak, M., Davey, S.G., 1995. Explanations for social inequalities in health. In: Amick, B.B., Levine, S., Tarlov, A., Walsh, D.C. (Eds.), Society and Health. Oxford University Press, Oxford, UK, pp. 172–210.

Marmot, M.G., 2003. Understanding social inequalities in health. Perspectives in Biology and Medicine 46 (Suppl. 3), S9–S23.

McEwen, B.S., 1998. Protective and damaging effects of stress mediators. New England Journal of Medicine 338 (3), 171–179.

Miller, G., 2011. Why loneliness is hazardous to your health. Science 331 (6014), 138–140.

Pappas, G., Queen, S., Hadden, W., Fisher, G., 1993. The increasing disparity in mortality between socioeconomic groups in the United States, 1960 and 1986. New England Journal of Medicine 329 (2), 103–109.

Power, C., Hertzman, C., 1999. Health, well-being and coping skills. In: Keating, D.P., Hertzman, C. (Eds.), Developmental Health and the Wealth of Nations. The Guilford Press, New York, pp. 41–51.

Rendall, M.S., Weden, M.W., Favreault, M.M., Waldron, H., 2011. The protective effect of marriage for survival: a review and update. Demography 48 (2), 481–506.

Robles, T.F., Kiecolt-Glaser, J.K., 2003. The physiology of marriage: pathways to health. Physiology & Behavior 79 (3), 409–416.

Rosero-Bixby, L., Dow, W.H., 2009. Surprising SES gradients in mortality, health, and biomarkers in a Latin American population of adults. The Journals of Gerontology: Social Sciences 64B (1), 105–117.

Shively, C.A., Clarkson, T.B., 1994. Social status and coronary artery atherosclerosis in female monkeys. Arteriosclerosis, Thrombosis, and Vascular Biology 14 (5), 721–726.

Singh, G.K., Siahpush, M., 2006. Widening socioeconomic inequalities in US life expectancy, 1980–2000. International Journal of Epidemiology 35 (4), 969–979.

Smith, J.P., 2004. Unraveling the SES: health connection. Population and Development Review 30 (Suppl.), 108–132.

Subramanian, S.V., Kawachi, I., 2004. Income inequality and health: what have we learned so far? Epidemiologic Reviews 26 (1), 78–91.

Turra, C.M., Goldman, N., 2007. Socioeconomic differences in mortality among U.S. adults: insights into the hispanic paradox. Journal of Gerontology: Social Sciences 62B (3), S184–S192.

Turra, C.M., Goldman, N., Seplaki, C.L., Glei, D.A., Lin, Y.H., Weinstein, M., 2005. Determinants of mortality at older ages: the role of biological markers of chronic disease. Population & Development Review 31 (4), 675–698.

Uchino, B.N., Uno, D., Holt-Lunstad, J., 1999. Social support, physiological processes, and health. Current Directions in Psychological Science 8 (5), 145–148.

Wadsworth, M.E.J., Kuh, D.J.L., 1997. Childhood influences on adult health: a review of recent work from the British 1946 national birth cohort study, the MRC National Survey of Health and Development. Paediatric and Perinatal Epidemiology 11 (1), 2–20.

Wells, D., 2011. The value of pets for human health. The Psychologist 24, 172–176.

Wilkinson, R.G., 1996. Unhealthy Societies: The Afflictions of Inequality. Routledge, London, UK.

Zajacova, A., Goldman, N., Rodriguez, G., 2009. Unobserved heterogeneity can confound the effect of education on mortality. Mathematical Population Studies 16 (2), 153–173.

Mortality, Epidemiological, and Health Transitions

Graziella Caselli, Sapienza University of Rome, Rome, Italy

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G. Santow, volume 15, pp. 10070–10075, © 2001, Elsevier Ltd.

Abstract

In demography, mortality transition marked the shift from higher to lower mortality rates. In the medical epidemiological field, the analyses of the mortality trend by cause of death led to the formulation of the theory of epidemiological transition, used to mark the shift from a phase when infectious diseases were the primary cause of death, to one in which chronic diseases were the chief cause. Since the 1970s, the concept of epidemiological transition has taken on a new significance that is also more far-reaching and thorough. The stance taken to analyze structural changes in mortality by age and cause has changed: it is no longer merely descriptive but now also tries to define the link with the triggering determinants, including the wider concept of health transition.

Theoretical Concept of Mortality, Epidemiological, and Health Transitions

The study of changing trends in mortality which affect Europe together with certain non-European societies with deep European roots (the United States, Canada, Australia, New Zealand, Argentina, and Uruguay) has led to the theory of 'demographic transition,' a term coined to designate the historical process which marked the shift from an older mortality and fertility regime to a new one. In particular, the process of 'mortality transition' marked the shift from higher rates (over 30 per 1000) to lower rates (10–12 per 1000). This process of change in mortality in European countries has for a considerable time engaged the attention of academics in various fields: demographers as well as economists, medical researchers, epidemiologists, and geneticists among them. There is still considerable discussion about the underlying reasons for the decline of mortality (particularly during the eighteenth century), and many different views have been put forward (Livi Bacci, 1987; McKeown and Brown, 1955; McKeown and Record, 1962; McKeown, 1976; Wrigley and Schofield, 1981). It could hardly be otherwise, when mortality decline has been studied by means of a single global indicator (growth mortality rates). It has seldom been possible to study changes in sex- and age-specific mortality, and rarer still to consider changes to mortality by causes of death before the end of the nineteenth century. However, it is only by looking at changes in causes of death that we shall be able to identify the determinants which have resulted in falling mortality. In the nineteenth and twentieth centuries, the indicators of mortality by age and cause of death have helped demographers to a better understanding of the long mortality transition process.

In the medical epidemiological field, the analyses of the mortality trend by cause of death led to the formulation of the theory of 'epidemiological transition' (Omran, 1971), used for many years to mark the shift from a phase when mortality levels were high, with infectious diseases as the primary cause, to one in which mortality levels had notably declined, with chronic diseases being the chief cause. In the 1970s, the concept of epidemiological transition had taken on a new significance that is also more far-reaching and thorough. The stance taken to analyze structural changes in mortality by age and cause had changed; it was no longer merely descriptive but now tried also to define the link with the triggering determinants. From this viewpoint, moving from description to an explanation of the changes in the diverse mortality profile, the definition 'epidemiological transition' gave way to 'health transition' (Lerner, 1973). In the 1970s, the spectacular decline in mortality from cardiovascular disease that could be seen in all Western countries called into question the very concept of epidemiological transition, which was centered on the development of pathologies. In fact, in the 1990s, John Caldwell (1990), Graziella Caselli (1989, 1991, 1996), and Julio Frenk et al. (1991) proposed replacing the concept of epidemiological transition with the broader one of health transition, incorporating in the description of the development of mortality both the development of the epidemiological characteristics of the health situation and those of society's response to this state of health.

In what follows we shall analyze the historical development of the decline of mortality, following the various stages of the transition process in relation to the role played by changes in the epidemiological framework and the determinants that have produced them.

From Mortality Crises to the Mortality Transition

At the outset of the eighteenth century, man had not yet succeeded in mastering the forces of nature, and life expectancy, which was short even in normal times, was affected by subsistence crises and famines that resulted in misery and extremely high death rates. The effects of these crises were often made worse by epidemics of plague and other fevers (Caselli, 1989, 1991).

Of these various crises, which lasted until the early decades of the nineteenth century, the most important of which affected the Baltic and Scandinavian regions and other areas in Northeastern Europe in 1708–09, were followed in most countries of Central and Northern Europe by a period when subsistence crises became rarer and affected mortality less (Caselli, 1989, 1991, 1996; Vallin, 1991). Epidemics, however, continued

and typhus, smallpox, dysentery, and influenza were recorded throughout the eighteenth and much of the nineteenth century, though they were less lethal and less widespread than before. These crises began to disappear in the middle of the eighteenth century when Northern and Western countries started to benefit from the first significant results of a more effective struggle against famine and epidemic diseases, particularly thanks to the third 'food revolution' described by Massimo Livi Bacci (1987). After the middle of the eighteenth century, the less serious nature of the crises was reflected by a lower level of mortality and, conversely, by an increase in life expectancy at birth (see **Figure 1** for some representative countries). In fact, life expectancy began to increase in Northern European countries, England, and France, and had reached a figure of between 40 and 42 years by the early nineteenth century, while in the pretransitional period life expectancy was less than 30 years on average. The fall in death rates – particularly in infant ages (see **Figure 2**) – with the consequent increase in life expectancy, had already been apparent for more than a century in those countries when

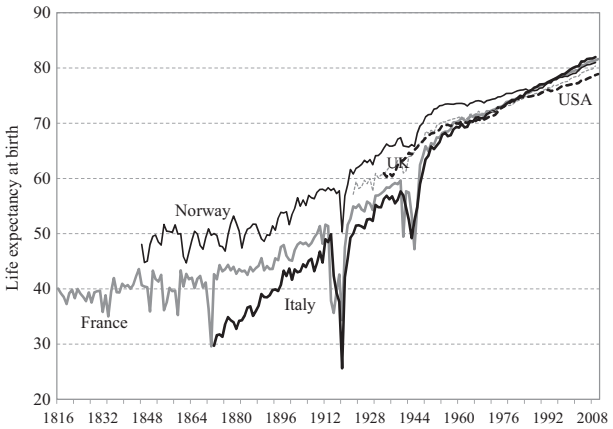

Figure 1 Life expectancy at birth in some selected countries: trends since the nineteenth century. Source: Own elaboration on the human mortality database, MPIDR, rostock (www.humanmortality.de).

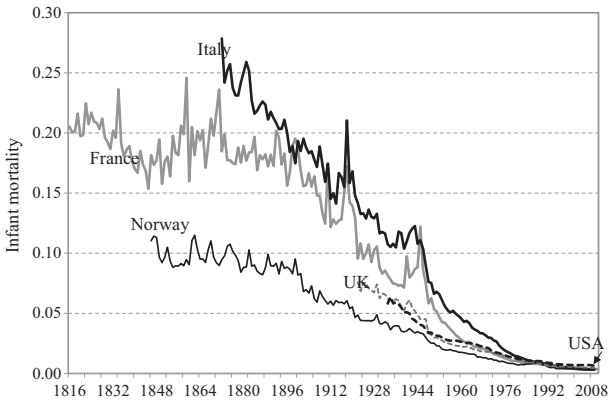

Figure 2 Infant mortality (per 1000) in some selected countries: trends since the nineteenth century. Source: Own elaboration on the human mortality database, MPIDR, rostock (www.humanmortality.de).

(just at the end of the nineteenth century) it was beginning to appear in the rest of Europe, accompanying the first stages of a breakthrough in medicine and improved individual and public sanitary conditions (waste disposal, preservation of surface waters, rubbish collection, etc.). The falling mortality trend continued in Northern countries while some Western countries, i.e., England, Wales, and France, ceased to make further progress in the 50 years between the 1830s and 1880s, during which period life expectancy figures remained more or less constant at the same level (e.g., for France see **Figure 1**). In France life expectancy did not change appreciably in the next 40 years and only resumed its increase toward the end of the 1880s, while in England and Wales the life expectancy increase became rapid and continuous after the 1870s (Vallin, 1991). The reasons for this stagnation have been discussed by a number of authors (McKeown, 1976; McKeown and Brown, 1955, McKeown and Record, 1962; Wrigley and Schofield, 1981). It appears to have been mainly due to the difficulties experienced by those countries in the rather chaotic and uncontrolled early stages of the industrial revolution with its negative effects on the health situation of the working masses and on living conditions and the environment in general. It was only from 1880 onwards, when the first effects began to be felt of the remarkable progress in health care and economic and social development, affecting not only workers but the population as a whole, that life expectancy in those countries resumed its previous upward trend, thanks to the important role played by the falling trend in infant mortality (**Figure 2**).

In Southern and Eastern Europe, where famine and epidemics were still present during the late nineteenth century, mortality rates remained almost at pretransitional levels. It was at the end of the nineteenth century that mortality began to fall significantly in these countries too, so that life expectancy at birth increased from 30–33 to 38–40 years, in 10–20 years (i.e., from 33 to 38 years in Italy from 1881 to 1891, see **Figure 1**).

The general pattern of the differing timescales with which European countries reached different stages of mortality transition clearly reveals the slowness of some of them in achieving a reduction in mortality, and also reflects the time taken by each country to reach the 50-year life-expectancy threshold – in other words, to pass, in Chesnais' terminology (1986), from an outdated to an intermediate mortality regime. Bearing in mind that the normal duration of life before the transition was of the order of 28–30 years, the pioneering countries took about two centuries to reach the intermediate level of 50 years (doubling their starting figure), whereas the latecomers achieved the same result in about half a century. The reasons for this accelerated progress are well-known. At the end of the nineteenth century, the main cause of death in the poorest countries in Europe was infectious diseases (such as typhus, measles, whooping cough, scarlet fever, and diphtheria), particularly deadly in children, which were spread by contamination of the environment or by contaminated food or water and aggravated by malnutrition and the wretched conditions under which the majority of the population lived (Caselli, 1996). Only in the early years of the nineteenth century were a number of these less advanced European countries and also other non-European countries able to benefit fully from the increasingly effective methods

of controlling these diseases which the pioneering countries had developed during the preceding period through slow and patient research.

In the twentieth century, all the countries entered the critical transitional phase in public health, which enabled almost all of them to achieve in only 30 years, the greatest progress ever made in so short a time. As a result, in all these countries the breakdown in the birth–mortality equilibrium, caused by the mortality crises of the two World Wars and the 'Spanish' flu epidemic, was rapidly surmounted (**Figures 1** and **2**). In 1960, life expectancy at birth in the wealthiest countries of Northern and Central Europe had by then reached 70 years for men and 75 for women, while other countries were slightly behind, with an average difference of 4 years between men and women. It was in the 1960s that the mortality rate reached its lowest levels (10–12 per 1000) more or less everywhere and that, following the theoretical pattern of demographic transition, concluded the shift from high to low levels as envisaged in the mortality transition model.

During the transition period there were major modifications in the age structure of the mortality rate, reflecting the passage from the previous regime of high infant and juvenile mortality due to infectious diseases, to a new regime in which adult mortality predominated, with a prevalence of chronic or degenerative diseases. The different phases of this change, viewed both from the overall standpoint and from this structural modification in respect of age and cause of death, moving from the ancient to the new epidemiological regime, were analyzed and theorized in epidemiological domain (Mackenbach, 1994; Omran, 1971).

Epidemiological versus Health Transition

In 1971, following the unprecedented increase in life expectancy in developed countries, Abdel Omran (1971) put forward a theory describing for the first time the decrease of infectious diseases, which had gradually been replaced by chronic diseases. As we have said above, it was labeled 'epidemiological transition' and, initially, comprised the following stages: (1) The 'age of pestilence and famine,' dominated by recurrent mortality crises in which life expectancy at birth was more or less 30 years (a phase indicated in mortality transition pattern as 'pretransitional'); (2) the 'age of receding pandemics,' during which some infectious diseases became less important and the sharp decline in mortality in the early years of life brought life expectancy to 50 years (start of mortality transition); and (3) the 'age of degenerative diseases,' is characterized by the predominance of chronic diseases and the stabilization of mortality at a low level (end of the mortality transition period).

According to Omran, with the Third Age the advances in life expectancy in the mid-1960s slowed down or, in some cases, stopped, with the rise of new endemic diseases, whether degenerative (e.g., cardiovascular diseases, cancers, diabetes, metabolic disorders). France Meslé and Jaques Vallin (2006) consider Omran's division into three ages as accurately reflecting reality up until the end of the 1960s, when he wrote his first article – a period in which the improvements in life expectancy were slowing down, particularly for men.

However, events swiftly belied Omran's model. From the early 1970s on, life expectancy in all Western countries once again began to increase for men too, driven by a clear reduction in mortality for cardiovascular diseases. Following these new behaviors and the resulting increase in survival, Jay Olshansky and Brian Ault (1986) and Richard Rogers and Robert Hackenberg (1987) imagined a 'fourth phase' of epidemiological transition. The first two authors suggested that the mortality model by cause had remained the same because the age of death had shifted forward, while the second two spoke rather of the start of a new era, where the most important advances in survival had been and would be achieved, thanks to the reduction in mortality brought about by the adoption of new and healthier lifestyles.

In 1983, Abdel Omran recognized the existence of one and possibly two additional phases to his initial theory of the epidemiological transition, accepting the suggestions of Jay Olshansky and Brian Ault, but pointed out that the new mortality models would be accompanied by the emergence of new diseases (HIV, Ebola, etc.) and the revival of other former diseases (such as, for example, cholera and tuberculosis).

Omran's identifying a fourth phase of epidemiological transition has been questioned (Fetter, 1997; Meslé and Vallin, 2006). According to France Meslé and Jacques Vallin (2006), his initial idea was based mainly on the conceptualization of the transition from the old stable regime of high mortality in the past to a new stable regime of low mortality. The transition would be protracted and would follow the same pattern of the three phases (or four, according to other authors) without changing the prevailing nosological framework Omran described. Obviously, this interpretation clashes with reality. The new era of advances in survival that had begun in the 1970s was based on a significant change in the epidemiological situation of the cardiovascular revolution. Detailed analyses have shown that this change is something quite distinct from the slump in infectious diseases. For this reason, France Meslé and Jacques Vallin (2006) proposed a substantial modification of Omran's model of epidemiological transition and the reappraisal of it by other authors, suggesting a new model that would follow "the semantic shift proposed by Julio Frenk et al. (1991) and include within the wider concept of 'health transition' an initial phase (that described by Omran) of life expectancy gains, attributed mainly to the decline in mortality due to infectious diseases, followed by a second phase dominated by the decline in cardiovascular diseases, leaving open the possibility of identifying later phases, such as that suggested by Shiro Horiuchi (1999) for cancer and perhaps senescence" (Meslé and Vallin, 2006: p. 250).

Two Phases of the Health Transition in More Developed Countries, Leaving Open the Possibility of a Third

For an overall view that will clearly identify and explain the two main phases of advances in health based on controlling two very different types of pathology that act on equally different ages of decease, we shall refer to the results of some analyses of the last 20 years (Caselli, 1991, 1996; Vallin, 2006). As we have already said, the decline in mortality in Europe began at

different times – due to the different socioeconomic conditions and risk factors in the different countries – while, if we exclude the countries of Eastern Europe, the remaining countries all reached the lowest level of 80 years of life expectancy at birth (79–80 years for men and 84–85 for women) at the outset of the twenty-first century.

If we divide causes of death into broad categories, then, quite apart from the times and levels of mortality, we will find similar models of mortality by cause. In fact, on the eve of the World War I infectious diseases were still by far the main cause of death in all populations, both in countries with a higher mortality and in those with lower mortality rates. If we take all these diseases together (including bronchitis, pneumonia, and influenza), they accounted at that time for 56% in Italy, 49% in England and Galles, and 43% in Norway (Caselli, 1996). In Northern and Southern Europe infectious diseases accounted for slightly more than half of total mortality. Within this group of causes, however, there was a clear geographical differentiation: as we were able to show in a preceding study, gastroenteritis dominated the list of diseases in Southern Europe, while tuberculosis still played a dominant role in Northern Europe, as in the nineteenth century. The geographical differences in the internal composition of infectious diseases were the result of various characteristics of infant mortality and mirrored the distances between countries in terms of development. Gastrointestinal diseases may, in fact, be regarded as a good index of social development, hygiene, and health, in which, at the beginning of the century, Mediterranean countries were still lagging behind the more advanced Nordic countries. Tuberculosis, on the other hand, to which adolescents and young adults were particularly vulnerable, mainly attacked countries in which the urbanization process – already well advanced as a result of the head start of those countries in the field of industrialization – provided a fertile field for the propagation of the disease. At the beginning of the 1930s, when the major health revolution began, infectious diseases, in spite of their genuine decrease, still constituted the principal cause of death. In most advanced Northern countries, where life expectancy already exceeded 60 years, as in the less advanced countries, with a life expectancy of 45–50 years, infectious diseases were still the prime cause of death, representing 35–40% of total mortality.

In the 1930s cardiovascular diseases were responsible in Northern and Central Europe for about 20% of overall mortality. On the other hand, even in the most advanced countries, cancer still be accounted for a modest proportion of total mortality, about 10%. It is of course possible, especially in the case of cancer, which had long been poorly diagnosed, that the available statistics underestimated the real proportion. It is nevertheless certain that we were only seeing the initial stages of the expansion of certain risk factors (cigarette smoking, of course, but also industrial hazards) which were subsequently to contribute to an increase in mortality under this heading. The group of the other diseases, still very well represented, accounted for a large proportion of deaths due to traditional causes, including all diseases specific to early infancy, which remained at a high level until the end of the 1960s.

As we have said, starting at different times and following different patterns specific to each country, the process of transition in the 1960s, just at the end of the major health revolution, had already completely transformed the disease classification of mortality. Thanks to the widespread medical, health, socioeconomic, and cultural progress, along with the introduction of sulfonamides in the 1930s and 1940s, the decline in mortality from infectious diseases accelerated, particularly with the arrival of, first, antibiotics and, then, vaccination. In France, for example, the decline in mortality from these causes was responsible for 10 of the 13.5-year life expectancy gained between 1930 and 1960, while in Italy, where the health transition was further behind, the gain made during the same years may be exclusively traced to the marked decline in mortality from these causes (Caselli, 1996).

In 1960s, cardiovascular diseases had become the paramount cause of death, both in men and women, causing between 30 and 40% of all deaths, depending on the particular country; they occupied the place previously held by infectious diseases. Although these diseases had become more significant, nevertheless, for men, and still more for women, they led to slightly fewer mortalities and contributed to the increased average survival of the population (almost 1 year in France from 1960 to 1970) even in the stagnation characterizing the 1960s, which particularly involved adult and elderly men, which can be primarily traced to unfavorable mortality trends from cancer and violent causes. Cancer, which in the 1960s had climbed to second place, accounted for about 18–20% of male mortality in Northern and Central European countries, and 15% in Southern countries.

It was in the 1970s that the great cardiovascular revolution began. The decline in mortality due to cardiovascular diseases became quickly more significant, first affecting adults and then the elderly, and established itself in all developed countries. This decline is still continuing today and also involves much of the adult population in Asian and Latin American countries (Caselli et al., 2002).

The 1960s, then, marked the most important turning point in health care. It was in this period that cardiovascular diseases replaced infectious diseases in extending life expectancy. France Meslé and Jacques Vallin (2006) point out that this contradicted Omran's theory of the conclusion of the epidemiological transition. Not only did deaths from cardiovascular disease reach ever lower levels from the late 1960s onwards – they had already been diminishing for women in the 1950s – but there was also an inversion in the growth of mortality for some types of tumor – an inversion that in some countries after the 1980s also became marked for tumors of the respiratory tract and for all those causes that France Meslé and Jacques Vallin (2006) call 'man-made diseases' (i.e., lung cancers, alcoholism, car accidents, suicide, homicide). This means that, while the decline in mortality for cardiovascular disease has caused much of the improvement in survival since the 1980s, we may conjecture on the basis of the most recent experience that the positive development of mortality for lung cancer may be, at least for men, at the center of the increase in survival that will take place in the next few decades, leaving open the possibility of identifying later phases, such as that suggested by Shiro Horiuchi (1999).

In fact, for some time it was believed that the different trends for cardiovascular disease and cancer indicated that the death risks for these two groups of causes were in competition with each other and that there had, in any case, been a transfer from

the one to the other as a result of improvements in diagnosis. Later, the clear inversion of the trend of cancer mortality suggested the possibility of other mechanisms and imparted a new stimulus to all the interpretative hypotheses which emphasize the relation between these changes and the economic, social, and health situation. The negative changes in mortality due to cancer, which have continued for many years, can now be more easily linked with the increasing importance of specific etiological factors against which we have no effective means of fighting, either individually or collectively. It is undeniable that the major success achieved in the therapeutic sphere has greatly benefited large numbers of patients, including cancer patients, and prevented the increase in this risk factor becoming an increase in overall mortality. An even more important part has been played, however, by the increasing sensitization of public opinion to health problems; this attitude, by promoting early diagnosis, has enabled full advantage to be taken of the latest treatments and certain risk categories to be eliminated by the adoption of healthier lifestyles.

It is not easy to determine to what extent increasing attention to health problems is due to health education, which is implemented in many countries, or to the changes which have occurred in economic and social structure (Preston, 1975). This latter hypothesis appears to be more soundly based, at least in regard to the majority of diseases, to the extent that the twentieth century has witnessed an extraordinary rise in educational levels and a profound economic transformation which has continually increased the proportion of the population with access to levels of information and standards of living which were reserved in the past for the privileged few. It is no less certain that, irrespective of the genesis of the changes which have occurred, many of those in recent years – and many more to come in the future – have revealed and will continue to reveal new potential resources in the battle against diseases and death.

In sum, the conviction is growing all the time that the determining factors of the fall in mortality have changed in the course of these years. In the first phase of the health transition, which coincided with the first stage of economic development, the direct and indirect importance of the availability of adequate food, the development of health and hygiene, and improvements in living accommodation played a decisive part. In the second phase, which saw the disappearance of the causes of death which were most dreaded in the past and the emergence of new diseases, the determining factors tended to lie in the field of pharmacological discoveries and the modernization of the health service, largely due to a massive investment in health by the public authorities. In this latter phase, however, which coincided with the industrial stage of higher economic development, new risks began to spread, giving rise to diseases such as cancers, against which therapeutic progress proved relatively ineffective at first. More elements suggest that we are on the way toward a new victory which would enable us to overcome these problems too. These are the signs of a development which will enable postindustrial societies to make further progress in life expectancy and in healthy life expectancy and generate a genuine third phase of the health transition. Without underestimating the part played by economic changes, the confirmation and general implementation of this third phase will also depend on the interaction between the new health culture and further improvements in therapeutic resources, particularly for the elderly. The new health culture is active on two fronts: individual and social control of the risk factors and early diagnosis of pathological conditions, and improved health facilities and more efficient forms of treatment and new medical technologies.

The Generalization of the Transition: The Decline of Mortality in Poor Countries

The considerable increase in life expectancy, based on two centuries of progress in combating infectious diseases, was at first limited to European and North American countries, but as early as the first half of the twentieth century also affected some parts of Latin America and Asia. The spread of medical techniques that had been used for years in the more developed countries allowed countries such as Mexico to gain 17 years of life expectancy in just 20 years (1930–50), an advance that had taken Sweden 150 years to achieve.

In a few decades, particularly in the postwar period, cheap but effective methods against infectious and parasitic diseases were widely used in the developed countries and led to a sharp increase in the survival possibilities in much of the Third World. Countries such as Cuba, Costa Rica, Regions as Kerala in India, and a – a little later – China – saw their peoples achieve a life expectancy close to those of rich countries before they had achieved high levels of economic development. The improvements in health care in countries of the South made possible the development of theories and explanatory models of the reduction of mortality, suggesting the importance of considering the many risk factors that, both individually and in combination, play a part in determining the process of the health transition (Mosley and Chen, 1984; Palloni, 1985).

The initial euphoria over the success of vaccines and antibiotics induced the belief that targeted campaigns would be enough to eradicate mortality from the main endemic causes of death. One of the most relevant success stories was Ceylon – present-day Sri Lanka – where in just 2 years (1947 and 1948) life expectancy increased by 12 years following a decisive campaign to eliminate the disease at the behest of the WHO, which involved massive spraying of DDT (Frederiksen, 1960).

From the 1950s to the threshold of the twenty-first century, life expectancy in developing countries overall rose from an average of 41–63 years, with a gain of 22 years, when in the same periods the richest countries saw a gain of no more than 9 years (Vallin, 2006), rising from 66 to 75 years. In this way the gap between the two groups of countries, which was 25 years immediately after World War II, had reduced to 12 years on the threshold of the new century. If we exclude from the group of developing countries sub-Saharan Africa, where even the more modest progress that has been made has recently been compromised by the AIDS epidemic, the difference of 12 years would reduce to 10.

This slowing down of the progress of the poorest countries has dampened the WHO's enthusiasm. It has become clear that imported progress in health care could bring lasting benefits only in the context of policies able to take into account the many factors (social, economic, cultural, and medical) that are highlighted in the theoretic models of experts (Caldwell, 1990;

Suarez-Ojeda and Junes, 1985; Tabutin, 1995). In addition, what proved true in the first phase of the European health transition, which was based on controlling infectious diseases, was still more so in the second phase, when advances in survival were linked to the reduction in mortality from very different kinds of diseases, such as cardiovascular conditions, for example. In this case, it was not possible to simply vaccinate the population against it and reclaim land that was a seat for some infectious diseases. Reducing mortality for modern diseases requires a series of global policies and changes in individual and collective behaviors that are unlikely to be acquired without adequate education. The cardiovascular revolution, which has been ongoing in rich countries and in many developing countries for several decades now – like the battle against cancer – requires, as we have said, individual and collective awareness of the importance of behavioral risk factors, which have been the real driving force in the second phase of the health transition. It has taken many years for all these conditions to have been met in the rich countries and it will require epoch-making changes before this happens in the poorer countries of the world.

In conclusion, the infectious diseases that killed children and adolescents in the main, had to be conquered before modern societies could direct their energies toward those diseases whose victims were adults. A high level of economic and social development was required before there could be the necessary material and cultural tools for eliminating the modern pathologies and preventing new pathologies from appearing or old ones from re-emerging and blocking the process of health transition toward ever higher survival values – and so preventing what happened in the countries of Eastern Europe, and Russia in particular, in the 1980s. These are the only countries that have experienced all the inconveniences of the 'third phase of epidemiological transition' indicated by Abdel Omran in 1983, combining the lack of positive effects from the decline in mortality from degenerative diseases with those produced by an increase in mortality from man-made diseases (Skholnikov et al., 1995).

See also: Adult Mortality in Developing Countries; Adult Mortality in Eastern Europe and the Former Soviet Republics; Adult Mortality in Industrialized Societies; Early Life Influences on Health and Mortality in Adulthood; Infant and Child Mortality in Industrialized Countries; Infant and Child Mortality in the Less Developed World; Mortality and the HIV/AIDS Epidemic; Mortality: The Great Historical Decline.

Bibliography

Caldwell, J.C., 1990. Introductory thoughts in health transition. In: Caldwell, J.C., Findley, S., Caldwel, P., Santow, G., et al. (Eds.), What We Know about Health Transition: The Cultural, Social and Behavioral Determinants of Health, Special Issue on Health Transition, vol. 1. Heath Transition Centre, Australian National University, Canberra, pp. vi–xiii.

Caselli, G., 1989. Transition sanitaire et structure par cause de la mortalité. Anciennes et nouvelles causes, Annales de Démographie Historique. Special Issue on "Le Déclin de la mortalté", Paris, pp. 55–77.

Caselli, G., 1991. Heath transition and cause specific mortality. In: Schofield, R., Reher, D., Bideau, A. (Eds.), The Decline of Mortality in Europe. Clarendon Press, Oxford, pp. 68–96.

Caselli, G., 1996. National differences in the health transition in Europe. Historical Methods 29 (3), 107–125.

Caselli, G., Meslé, F., Vallin, J., 2002. Epidemiological transition theory exceptions. Genus 58 (1), 9–51.

Chesnais, J.-C., 1986. La Transition Démographique: Étapes, Forms, Implications Économiques. Études de séries temporelles (1720–1984) relatives à 67 pays. Cahier 113. INED-PUF, Paris.

Fetter, B., 1997. The epidemiologic transition: one, Many or none? Health Transition Review 7 (2), 235–237.

Frederiksen, H., 1960. Malaria control and population pressure in Ceylon. Public Health Reports 75, 865–868.

Frenk, J., Bobadilla, J.L., Stern, C., Frejka, T., Lozano, R., 1991. Elements for a theory of the health transition. Health Transition Review 1 (1), 21–38.

Horiuchi, S., 1999. Epidemiological Transition in Human History. In: Chamie, J., Cliquet, R.L. (Eds.), Health and Mortality Issues of Global Concern. Proceedings, of the Symposium on Health and Mortality, Brussels. United Nations, New York, pp. 54–71.

Lerner, M., 1973. Modernization and Health: A Model of the Health Transition. Paper presented at the American Public Health Association Conference, San Francisco.

Livi Bacci, M., 1987. Popolazione e alimentazione. Saggio sulla storia demografica europea. Il Mulino (Universale Paperbooks), Bologna.

Mackenbach, J.P., 1994. The epidemiologic transition theory. Journal of Epidemiology and Community Health 48, 329–331.

McKeown, T., 1976. The Modern Rise of Population. Edward Arnold, London.

McKeown, T., Brown, R.G., 1955. Medical evidence related to English population change in the eighteenth century. Population Studies 11 (2), 119–141.

McKeown, T., Record, R.G., 1962. Reasons for the decline in mortality in England and wales during the nineteenth century. Population Studies 16 (2), 94–122.

Meslé, F., Vallin, J., 2006. The health transition: trends and prospects. In: Caselli, G., Vallin, J., Wunsch, G. (Eds.), Demography. Analysis and Synthesis. A Treatise in Population Studies, vol. II. Academic Press-Elsevier, USA, pp. 247–259. Chapter 57.

Molineaux, L., 1985. The impact of parasitic diseases and their control, with special emphasis on malaria in Africa. In: Vallin, J., Lopez, A., Behm, H. (Eds.), Health Policy, Social Policy and Mortality Prospects, ordina ed. Liège, Belgium, pp. 11–40.

Mosley, W.H., 1985. Will primary health care reduce infant and child mortality? A critique of some current strategies, with special reference to Africa and Asia. In: Vallin, J., Lopez, A., Behm, H. (Eds.), Health Policy, Social Policy and Mortality Prospects, ordina ed. Liège, Belgium, pp. 101–136.

Mosley, W.H., Chen, L.C., 1984. An analytical framework for the study of child survival in developing countries. Population and Development Review 10, 25–45.

Olshansky, S.J., Ault, A.B., 1986. The fourth stage of the epidemiological transition; the age of delayed degenerative diseases. The Milbank Quarterly 54 (3), 355–391.

Omran, A.R., 1971. The epidemiologic transition: a theory of the epidemiology of population change. Milbank Memorial Fund Quarterly 49 (4), 509–538.

Omran, A.R., 1983. The epidemiological transition theory. A preliminary update. The Journal of Tropical Pediatrics 29, 305–316.

Preston, S.H., 1975. The changing relation between mortality and economic development. Population Studies 29 (2), 231–248.

Palloni, A., 1985. Santé et lutte contre la mortalité en Amérique latine. In: Vallin, J., Lopez, A., Behm, H. (Eds.), La lutte contre la mort. INED, PUF, Paris, pp. 447–473 (Travaux et Documents, cahier n°108).

Rogers, R.G., Hackenberg, R., 1987. Extending epidemiologic transition theory. Social Biology 34, 234–243.

Suarez-Ojeda, N., Junes, Y., 1985. Childhood mortality in the Americas: probable effects of primary health care. In: Vallin, J., Lopez, A., Behm, H. (Eds.), Health Policy, Social Policy and Mortality Prospects, ordina ed. Liège, Belgium, pp. 139–162.

Skholnikov, V., Meslé, F., Vallin, J., 1995. La crise sanitaire en Russie. Population 50 (4–5), 907–982.

Tabutin, D., 1995. Transitions et théories de mortalité. In: Gérard, H., Piché, V. (Eds.), Sociologie des populations. Presse de l'Université de Montréal, Montréal, pp. 257–288.

Vallin, J., 1989. Théorie(s) de la baisse de la mortalité et situation africaine. In: Pison, G., van de Walle, É., Sala Diakanda, D. (Eds.), Mortalité et Société en Afrique au sud du Sahara. INED, PUF, Paris, pp. 399–431. Travaux et documents, Cahier n° 124.

Vallin, J., 1991. Mortality in Europe from 1720 to 1914. Long-term trends and changes in patterns by age and sex. In: Schofield, R., Reher, D., Bideau, A. (Eds.), The Decline of Mortality in Europe. Clarendon Press, Oxford, pp. 38–67.

Vallin, J., 2006. Europe's demographic transition, 1740–1940. In: Caselli, G., Vallin, J., Wunsch, G. (Eds.), Demography. Analysis and Synthesis. A Treatise in Population Studies, vol. III. Academic Press-Elsevier, USA, pp. 41–66. Chapter 68.

Wrigley, E.A., Schofield, R.S., 1981. The Population History of England 1541–1871. Edward Arnold, London.

Mortality of the Oldest-Old

Vladimir Canudas-Romo and Adam Lenart, Max-Planck Odense Center on the Biodemography of Aging, University of Southern Denmark, Odense C, Denmark

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.W. Vaupel, volume 15, pp. 10075–10079, © 2001, Elsevier Ltd.

Abstract

"An 80-year-old is as healthy today as a 70-year-old yesterday," such phrases are not uncommon, and we can actually observe that the probability of death of an 80-year-old in Japan in 2009 corresponds to the mortality of a 70-year-old in 1959. This significant change in the mortality level of the oldest-old population, i.e., people 80 years and older, is the focus of this article. Among others, we concentrate on the age patterns, time changes, and population variation in mortality among the oldest age group of the population.

The Oldest-Old: Are They Still 80 Years Old?

Improvements in old age mortality during the second half of the twentieth century have resulted in an unprecedented increase in absolute and relative terms of the oldest-old. This rise in their numbers would not have been possible without a decrease in their mortality rate. As such, there is little resemblance in terms of mortality between an 80-year-old today and a person reaching to that age half a century ago (Sanderson and Scherbov, 2010). As **Figure 1** shows, the mortality level of 80-year-olds in 2010 in the USA corresponds to the mortality of 71-year-olds in 1950. Japan shows a more extreme case with an age gap of 10 years between those in 1959 that experienced the same mortality level as 80-year-olds in 2009.

For countries with longer historical time series of mortality, we can already see moderately improving mortality conditions as early as in 1900 for the oldest-old. In France and Sweden at ages 63 and 70 in 1900, and 69 and 72 in 1950, respectively, are observed the same levels of mortality of an 80-year-old in 2010. In other words, a transition of 11 years in France (from age 69 in 1950 to 80 in 2010) and 8 years in Sweden in the second half of the century as opposed to 6 and 2 years in the first half of the century, respectively. In these countries as well as in other high-longevity countries, the second half of the century was characterized as one of drastic mortality changes for the oldest-old.

How Many Oldest-Old Are There?

The differential in ages that have the same mortality levels over time is only one indicator of the changes in mortality observed in the group of the oldest-old. The number of survivors or mortality until the age of 80 years, and the remaining expected number of years of life after the age of 80 years, or life expectancy at age 80, complement the information provided by the level of mortality at this age. Of these three spectrums of mortality (past, present, and future; or mortality before, current, and after age 80), the largest mortality differential is the great number of extra survivors attaining old age now as opposed to in the past.

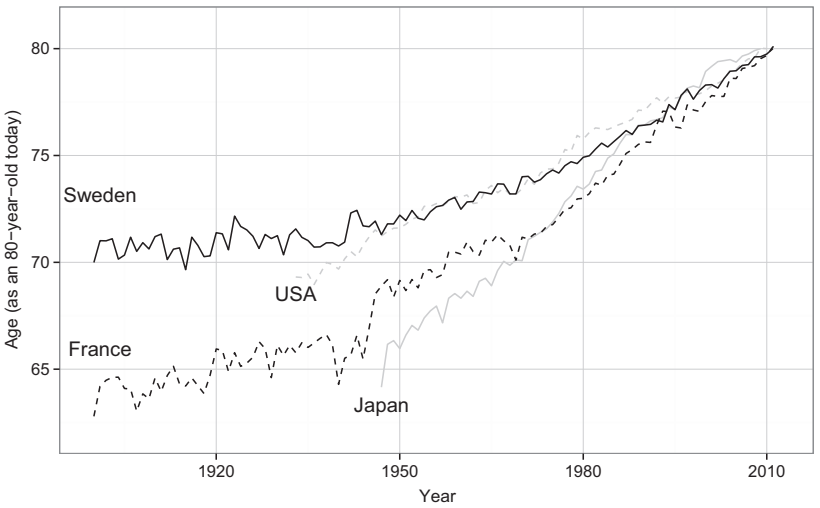

Figure 1 Ages in the twentieth century that show similar levels of mortality as 80-year-old females and males in 2010 for the USA, France, Sweden, and Japan.

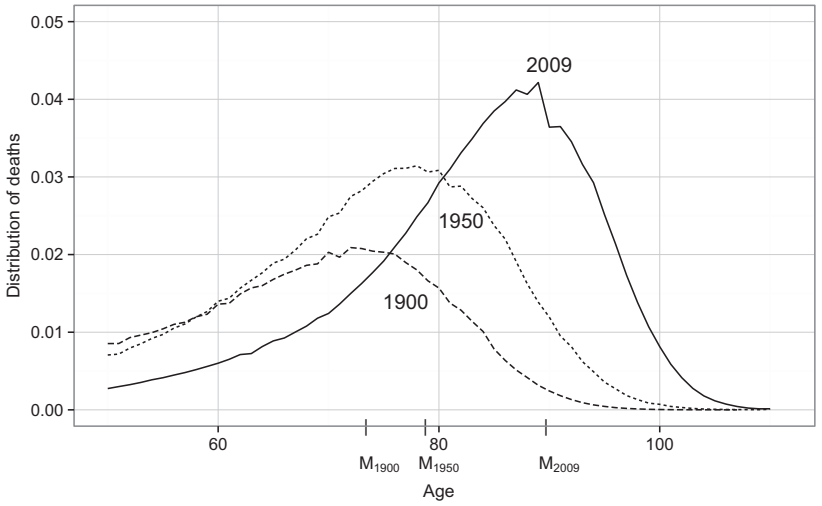

Figure 2 Life table distribution of deaths after the age of 50 years in high-longevity countries in 1900, 1950, and 2000.

Aggregating all the mortality information from high-longevity countries, as if it was one population, allows us to observe clearer trends in mortality for the oldest-old (we use this aggregated data in the rest of the article unless we explicitly refer to a country). **Figure 2** presents the bell shape characteristic of the distribution of deaths centered on the modal age at death, or age where most of the deaths occur.

In 1900, the modal age at death was 73.4 years, which changed to 78.8 in 1950 and 89.7 in 2009. In the second half of the twentieth century, there is a faster shift toward older ages for the bell shape in the distribution of deaths. This shift reflects the amble improvements in mortality in ages above the modal age seen in the recent past. High-longevity countries have a life expectancy at birth of 81.5, which is beyond the threshold age of the oldest-old group of 80 years. In other words, a child born today in high-longevity countries will live on an average of 81.5 years if current mortality conditions remained without change in the future. However, it is 8.2 years later where the modal age at death or age with the highest concentration of deaths is found (Canudas-Romo, 2010). The burden in hospitals, nursing homes, and public health in general is thus intensified at ages around the modal age at death.

The highly debated rectangularization of the survival function is easily refuted by looking at the trend over time in the modal age at death. The current changes are better described as a shift in the bell shape of mortality around the mode, while the deviation from the modal value has decreased. The standard deviation beyond the modal age in high-longevity countries has changed from 4.9 in 1900, to 4.7 in 1950, and 3.6 in 2010. Some have seen this as the beginning of the end of longevity or the arrival to the limit of longevity, while a more accepted perspective argues in the direction of a double phenomenon of shifting and compression of mortality.

Age Pattern of Mortality after the Age of 80 Years

While mortality increases exponentially in younger adult ages, the probability of death levels off at extreme old ages. The well-known Gompertz–Makeham mortality model captures well the pattern of mortality for young adults between ages 30 and 90, but it needs to be modified to capture the plateau of the age-specific death rates of the oldest-old (Missov and Lenart, 2011). A logistic curve-type in the female and male age pattern of mortality can be seen in **Figure 3** (Thatcher et al., 1998). In this Figure, the estimate of a constant death rate of 0.7 above age 110, stemming from supercentenarian research, is also included (Maier et al., 2010).

Frailty theory has helped explaining the bending of the mortality curve at old ages. This theory considers a population to be composed of heterogeneous individuals with various degrees of mortality. Individuals with a higher level of mortality are likely to die younger, and the overall population level mortality is then driven by the lower level of mortality of the remaining persons.

The larger pool of oldest-old persons has also increased the variation of mortality at older ages and changes in the age pattern of mortality are already observable. Octogenarians in high-longevity countries experienced average annual mortality improvements of 1.2% between 1950 and 2009. For the same period, nonagenarians experienced average annual declines of 0.6%. However, the mortality of centenarians is yet to improve. As a consequence of the mortality shape at old ages, the groups of centenarians and supercentenarians have increased at unprecedented numbers.

The cause of death structure of the oldest-old differs from the causes of death at younger-old ages. Multiple causes of death are not uncommon in death certificates at the oldest-old ages. Though broadly speaking, in spite of the major decline of cardiovascular mortality at the oldest-old ages, heart failure, coronary atherosclerosis, and stroke are still among the leading causes of death. Other important causes of death that distinguish the oldest-old from the younger old are pneumonia, influenza, septicemia, mental disorders, nutritional deficiency, and accidental fall. Diseases that tend to develop prematurely in some high-risk individuals at younger ages such as malignant neoplasms, acute myocardial infarction (particularly for males), hemorrhagic stroke, chronic liver disease and cirrhosis,

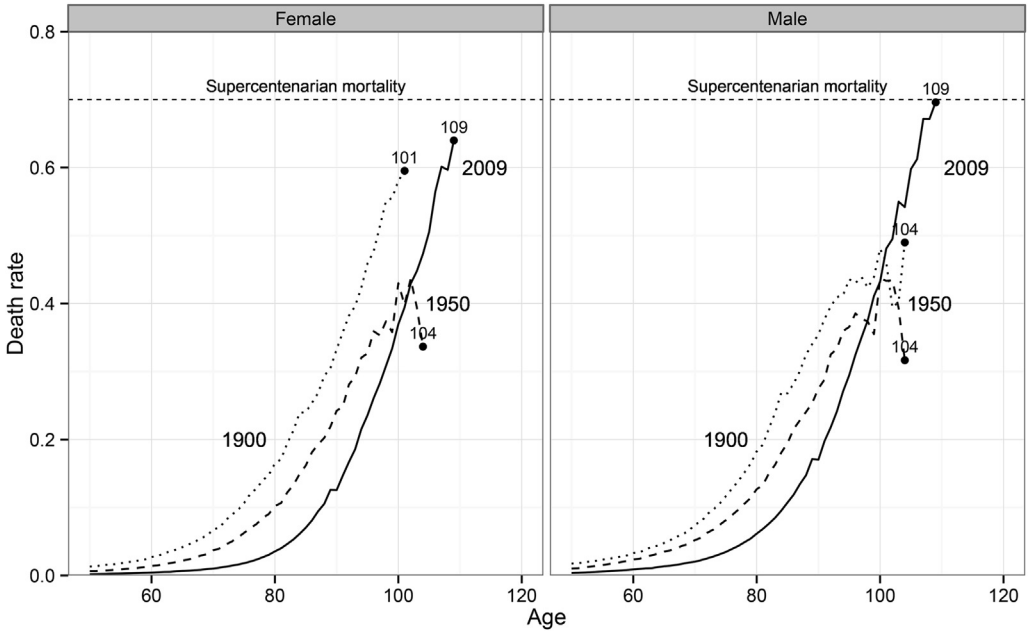

Figure 3 Female and male age patterns of mortality for high-longevity countries. Including estimates of mortality for ages beyond 110, or supercentenarians, at a constant level of 0.7 (equivalent to a probability of death of 0.5; Maier, H., Gampe, J., Jeune, B., Robine, J.M., Vaupel, J.W., 2010. Supercentenarians. Demographic Research Monographs 7 (XVI), 325. http://dx.doi.org/10.1007/978-3-642-11520-2.).

diabetes mellitus, and multiple sclerosis are less prevalent at oldest-old ages (Horiuchi, 2007).

Gender Differences in the Mortality of the Oldest-Old

The higher male to female mortality found at younger ages is also observed at the oldest-old ages (Oksuzyan et al., 2010). As a consequence, the ratio of oldest-old females compared with oldest-old males is particularly prominent. Among developed nations, the extreme disparity between numbers of females and males above age of 80 years is found in Germany with 227 females per 100 males. Eastern European countries have also very high numbers of females as compared to males at older ages. At the other extreme, in Iceland there are only 150 females per 100 males followed closely by Israel and Australia. The female to male ratio gains intensity with age; if all the mortality information from high-longevity countries were taken together, at age of 80 years the ratio of females to males would be 146 females per 100 males, by age of 90 years it increases to 250, for 100-year-olds we have 547, and at age 110 a ratio of 772. Once more referring to high-longevity countries, females surpass their male counterparts with 44 920 females above age 100 in 2009, as opposed to 7219 males. The maximum age attained at death by a person follows also the female–male gap differential with Jeanne Calment (France), a woman reaching age 122 and a half when she died in 1997 (Vaupel, 2010) and Jiroemon Kimura (Japan), a man reaching age 116 when he died in 2013 (although the age of this person remains to be well documented).

The female–male life expectancy gap has been changing over time too. The contribution of the oldest-old age group to explaining this life expectancy gap between sexes has increased over time. A combination of cancers, respiratory and cardiovascular diseases is identified as the causes of death responsible for the gender-gap in remaining life expectancy at ages beyond 80.

Determinants of Oldest-Old Longevity

Blue zones was a term coined by Michel Poulain and Gianni Pes to identify a region in Sardinia (Italy) where the number of oldest-old, particularly centenarians and supercentenarians, is unusually high. This has been popularized into a less scientific, but still intriguing, search for regions in the world where mortality at oldest-old ages is significantly lower: Okinawa (Japan), Nicoya (Costa Rica), Ikaria (Greece), and Loma Linda (United States) can be listed among them. Although each group has its own reasons for their aging success, some common traits for all groups are they are physical active communities, where people have a vegetable-based diet with a moderate alcohol consumption but without overeating habits; of importance is also that they have a reason for living and waking up every day; and that they belong to a group (faith, friendship, or/and family).

Being born with the right set of genetic information has also been suggested as a key element to reach old age and to have lower mortality risk than your peers; for example, carriers of the ApoE ε4 allele have higher mortality than none carriers, particularly among those aged 90 and more. Similarly, variants of the FoxO gene have been found more frequently in centenarians than in 90-year-olds suggesting a strong association of these genes with human longevity (Kenyon, 2010). Although, these associations have been

found in centenarians coming from different ethnic groups, the mechanism of aging and how the genes delay this process need to be further explored.

Theories of Oldest-Old Mortality

The most prominent evolutionary theories and a brief description of their main ideas can be shortly mentioned (Kirkwood and Holliday, 1979) as follows:

1. Mutation accumulation: Mutations that decrease the fitness are selected against, but germ line mutations that are expressed in older ages are not regulated by evolution.
2. Antagonistic pleiotropy: Genes that offer advantages to survival or reproduction in younger ages and are deleterious in older ages are not selected against.
3. Disposable soma: There is a trade-off between reproduction and maintenance. The primary function of the somatic cells is the reproduction of the organism. However, only finite resources are available for organisms, and their most efficient use is to invest in reproduction and less in repair.
4. Grandmother hypothesis: Long postreproductive human life span is justified, if older relatives contribute to the reproductive success of their progeny.
5. Insurance against indeterminacy: Higher life expectancy decreases the probability of death in the reproductive ages.

Current evolutionary theory acknowledges that postfertile life span is an evolved trait. As evolutionary selection optimizes fitness (number of surviving offspring) and not the length of life itself, evolutionary theory does not suggest any upper limit of human life span. Demographic observation also supports the notion that a 'Wall of Death' that cannot be exceeded does not exist. According to the mutation accumulation, antagonistic pleiotropy and disposable soma theories, the timing of old-age death is indeterminate. However, the grandmother hypothesis and insurance against indeterminacy hypothesis postulate that mortality in the postreproductive ages could have evolved to increase fitness in childbearing ages (Levitis et al., 2013).

Data and Future of Oldest-Old Mortality

Problems with vital registration systems, census and population registers, divide countries into those where it is possible to accurately calculate death rates and those where it is not. As such, in many developing nations, it is less clear which type of mortality levels are found in the oldest-old population. For example, in several Latin American countries the mortality of the older population is lower than in more developed nations: for example, Costa Rican death rate between ages 80 and 84 in the period 2005–09 was 61.8 per 1000, which is lower than the death rate of 65.9 per 1000 observed in Sweden for the same age group and time. Costa Rica has very well-documented data, however, other countries in the region have similar death rates, which have prompted debate among specialists about the accuracy of the information in the entire region. Other regions of the world have even greater problems in their data collecting system with the extreme in African countries, where data can be trusted only in very few cases.

Countries with deficient mortality and population historical data use relational models as an alternative to assess their mortality levels, now and in the future. The output of these relational models that are created from historical information from countries with reliable data are age patterns of mortality that decelerate with age at the oldest-old ages.

For countries with accurate historical mortality information, the Lee–Carter model is currently the most widely used mortality projection model in demography. According to this model, the mortality level of the oldest-old is likely to decrease in the future. However, this model might even overestimate their future death rates as it assumes a constant rate of improvements in mortality, which does not match the experience observed in the oldest-old age group. Alternative procedures have been developed recently to more systematically capture the rapid decline of mortality of the oldest-old. Examples of alternative forecasts in individual country forecasts are including the information of the best performer, or averages of all countries with similar trends.

To further understand the oldest-old mortality in the present and to prepare for the future, priority should be given to collect the right data of both, current and past, health and exposure information. A multidisciplinary effort will be required, where scientists from the fields of humanities, biology, medicine, and mathematics collaborate to advance the knowledge of the secrets of oldest-old low mortality and long life.

See also: Adult Mortality in Industrialized Societies; Aging Mind: Facets and Levels of Analysis; Aging and Health in Old Age; Aging, Theories of; Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences; Death and Dying, Sociology of; Demographic Models; Demographic Techniques: Data Adjustment and Correction; Differential Aging; Early Life Influences on Health and Mortality in Adulthood; Evolutionary Demography: A Synthesis of Two Population Sciences; Life Table; Mortality, Biodemography of; Mortality, Epidemiological, and Health Transitions; Mortality: The Great Historical Decline; Multistate Transition Models in Demography; Obesity and Mortality; Old Age and Centenarians: The Human 'Warranty Period'; Old Age: Definitions, Theory, and History of the Concept; Period and Cohort Analysis in Demography; Population Forecasts.

Bibliography

Canudas-Romo, V., 2010. Three measures of longevity: time trends and record values. Demography 47 (2), 299–312.

Horiuchi, S., 2007. Causes of death among the oldest-old: age-related changes in the causes-of-death distribution. In: Robine, J.M., Crimmins, E.M., Horiuchi, S., Zeng, Y. (Eds.), Human Longevity, Individual Life Duration and the Growth of the Oldest-old Population. Springer, Dordrecht, Netherlands, pp. 215–235.

Kenyon, C.J., 2010. The genetics of ageing. Nature 464 (7288), 504–512.

Kirkwood, T.B., Holliday, R., 1979. The evolution of ageing and longevity. Proceedings of the Royal Society of London. Series B. Biological Sciences 205 (1161), 531–546.

Levitis, D., Burger, O., Bingaman Lackey, L., 2013. The human post-fertile lifespan in comparative evolutionary context. Evolutionary Anthropology 22 (2), 6679.

Maier, H., Gampe, J., Jeune, B., Robine, J.M., Vaupel, J.W., 2010. Supercentenarians. Demographic Research Monographs 7 (XVI), 325. http://dx.doi.org/10.1007/978-3-642-11520-2. Springer, Heidelberg [et al.].

Missov, T., Lenart, A., 2011. Linking period and cohort life-expectancy linear increases in Gompertz proportional hazards models. Demographic Research 24, 455–467. http://dx.doi.org/10.4054/DemRes.2011.24.19.

Oksuzyan, A., Crimmins, E., Saito, Y., O'Rand, A., Vaupel, J.W., Christensen, K., 2010. Cross-national comparison of sex differences in health and mortality in Denmark, Japan and the US. European Journal of Epidemiology 25 (7), 471–480.

Sanderson, W.C., Scherbov, S., 2010. Remeasuring aging. Science 329 (5997), 1287–1288 (Washington).

Thatcher, A.R., Kannisto, V., Vaupel, J.W., 1998. The Force of Mortality at Ages 80 to 120, vol. 22. Odense University Press, Odense.

Vaupel, J.W., 2010. Biodemography of human ageing. Nature 464 (7288), 536–542.

Relevant Websites

www.mortality.org – The Human Mortality Database (HMD) was created to provide detailed mortality and population data on the history of human longevity.

http://www.bluezones.com/ – Information of areas in the world where there is an unusually low mortality at older ages.

http://www.supercentenarians.org/ – The International Database on Longevity (IDL) is the result of an ongoing concerted effort to provide thoroughly validated information on individuals who attain extreme ages.

http://www.demogr.mpg.de/en/projects_publications/online_databases_1906/default.htm – The Max Planck Institute for Demographic Research includes in its website information on several datasets that discuss mortality among the oldest old. Example of these data are: Kannisto-Thatcher Database on old age mortality; International Database on longevity (IDL); human life-table Database.

http://www.who.int/healthinfo/mortality_data/en/ – World Health Organization Mortality Database is a compilation of mortality data by age, sex and cause of death, as reported annually by member states from their civil registration systems.

Mortality: The Great Historical Decline

Tommy Bengtsson, Lund University, Lund, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The timing of the great mortality decline was strikingly similar in the countries of Western and Northern Europe, despite differences in their individual economic structure and development. The decline started in the middle of the eighteenth century, became more pronounced around AD 1800, and then continued. The history of mortality decline in North America was similar. Life expectancy at birth starting at around 35 years, increased about 1 month per year between AD 1750 and 1840, and thereafter increased by almost 3 months per year to reach more than 80 years by the end of the nineteenth century. With the great mortality decline followed a change in the composition of the causes of death, away from pestilence, over receding pandemics, to man-made diseases. Few scholars today will argue that a single factor, like economic growth, was the main determinant of the great mortality decline. Economic growth was perhaps not a major determinant, at least not in the initial stages of the great mortality decline and perhaps far less than expected during the later stages as well. In fact, the causes are multifactorial and vary from the start of the decline to the present day.

Features of the Decline

Our knowledge about the initial stages of mortality decline is not only based on data from the Nordic countries, which have had historical data of age-specific mortality rates far back in time due to the early implementation of population censuses, but also from data derived from England, France, and other countries, where information has been derived through reconstructions of national samples of parish data. From these sources we know that mortality in the Western world started to decline around AD 1800, a decline that identifies the start of the demographic transition (Davis, 1945; Notestein, 1953). It is, however, less well known that mortality had already started to decline throughout Europe prior to the mid-eighteenth century (Perrenoud, 1984); in England the decline started even earlier, from AD 1720 (Wrigley et al., 1997), with a similar development in North America (Fogel, 1994). Thus the timing of the great mortality decline was strikingly similar across the countries of Western and Northern Europe, starting, albeit at a lower rate, well before 1800. This timing could also be substantially similar within different areas of a country, as in Sweden (Fridlizus, 1984).

We know that this decline began with a fall in the rates of infant and child mortality. While infant mortality continued to decline in the Nordic countries throughout the nineteenth century, it leveled off in Western Europe and remained stable until the end of the nineteenth century, and then dropped rapidly as in all other Western countries. Thus, two patterns of the historical infant and childhood mortality decline can be identified: a Northern pattern with declines essentially throughout the nineteenth century and a Western one with a leveling off during the nineteenth century and a new decline after the 1880s (Perrenoud, 1984). The decline in adult and old age mortality became more pronounced during the course of the nineteenth century and became more rapid after World War I. It slowed down around 1950, to become more rapid again in the 1970s and has continued declining since then.

The great mortality decline was characterized by a change in the composition of the causes of death, shifting away from pestilence, via receding pandemics, to man-made diseases. In fact from Swedish records we know that the pre-1800s decline in infant and child mortality was entirely due to a reduction in deaths from smallpox, which was a very common childhood disease in the eighteenth century (Fridlizus, 1984; Sköld, 1997). The theory of epidemiological transition, distinguished between three development patterns: the classical (Western) pattern described above; the accelerated pattern; and the delayed pattern (Omran, 1971). They are distinguished by differences in timing and speed. The decline that took 200 years in the West started 150–200 years later in the Third World, but took less than 50 years to complete. There are still many countries where the great mortality decline has been delayed and where the possibility still remains to reap gains from low mortality.

As a consequence of the decline in mortality, life expectancy started to increase, from around 35 years in the mid-eighteenth century, to more than 45 years in the 1850s, and to almost 55 years at the end of the nineteenth century in a number of countries. Mortality in the Western world continued to decline throughout the twentieth century, and in best practice countries, life expectancy increased at a rate of about 3 months per year for females, (slightly less for males), to reach more than 80 years in AD 2000 (Oeppen and Vaupel, 2002). Countries with lower life expectancy levels, like Japan and China, had far faster improvements in life expectancy during the catch-up periods, sometimes more than 1 year per calendar year.

Explanations of the Mortality Decline

Two types of factors, *period* and *cohort*, have been put forward in explaining the decline. While period factors affect the entire population at the same time regardless of age, cohort factors influence a certain cohort, typically a birth cohort, over the entire life course. One should also make a distinction between factors that rely almost entirely upon the *actions of man* and those *beyond deliberate human control*. Interestingly enough, cohort explanations, which were in focus in the 1920s and 1930s, were replaced by period explanations in the 1950s, and seem to be enjoying a revival.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31043-1

Period Explanations

A widely accepted multifactorial explanation (United Nations, 1953) is based on period factors such as public health reforms, advances in medical knowledge, improved personal hygiene, as well as rising income and rising standards of living. This explanation is consistent with the demographic transition theory.

The demographic transition theory was once one of the most widely accepted 'truths' in social sciences and one of the main pillars of the UN view of how to promote heath improvement in developing countries. It stated that mortality was high and fluctuated strongly before the demographic transition. The transition was said to begin with a decline in the general level of mortality and with a dampening of its fluctuations. This was regarded as a consequence of a modernization of the economies (the agricultural and industrial revolutions). A rapid increase in population size would follow, since fertility declined much later than mortality, due to slow response from the institutions and norms that influence fertility (Davis, 1945; Notestein, 1953).

The idea of a stable level of mortality prior to modernization is a central feature in the theory, but before the mid-eighteenth century the mortality level was far from stable, as was shown for England by Wrigley and Schofield (1981). Similar patterns had been demonstrated by many other scholars as well, but only for individual parishes or institutions, or for smaller regions like Southern Sweden (Bengtsson and Oeppen, 1993) or Northern Italy (Galloway, 1994). Since the mortality level prior to the great mortality decline was not stable, it is not appropriate to use it to identify the beginning of the demographic transition (Bengtsson and Ohlsson, 1994).

The link between modernization and the mortality decline, central to the demographic transition theory, is far from obvious. Typically, mortality began to decline at least half a century (or earlier) before marital fertility started its own long-term decline, and long before living standards for the ordinary man started to improve – another criticism of the demographic transition theory. Sweden is a good example of this, since living standards for workers started to improve steadily only after industrialization around 1860 – more than 100 years after mortality started its slow decline and more than 50 years after the decline became more rapid. Also for countries like Germany and France, which experienced the industrial take off earlier than Sweden, the mortality decline preceded improvements in living standards. In fact it is not even clear that improvements in living standards preceded mortality decline in England, since evidence suggests that living standards for workers declined during the eighteenth century, at the same time that mortality slowly started to go down (Wrigley and Schofield, 1981).

On a broader front, McKeown (1976) questioned the multifactorial explanation of the great mortality decline. He argued that a single factor, namely better nutrition, could explain almost the entire great mortality decline. His criticism was based on a study of cause-specific mortality in England and Wales from 1838 to 1947, where he observed that two-thirds of the mortality decline was due to a reduction in the prevalence of infectious diseases.

McKeown argued that medical advances had little influence on mortality trends before the breakthrough of sulphonamides and antibiotics in the 1930s and the 1940s. In earlier periods the only disease that one could cure was diphtheria. Very few deaths, however, were due to diphtheria and it was already retreating when the antitoxin came into use around 1900. For periods before 1838, McKeown argued that inoculation and vaccination for smallpox had little or no impact on the general mortality decline. Vaccination started in England at the end of the eighteenth century, but it was not widespread until after 1840, at which time it became available at public expense. Inoculation stated earlier but had little impact, according to McKeown.

In the Nordic countries, smallpox mortality went down well before vaccinations started in the first years of the nineteenth century and long before vaccination became compulsory in Sweden in 1816 (vaccination became more common in the 1820s when some 40% of the population of Sweden were vaccinated). Although mortality due to smallpox was quite rare in Sweden by the time vaccinations started in 1801, it continued to be fatal for many decades. Thus although vaccination had nothing to do with the decline of smallpox mortality, it perhaps prevented it from coming back on a large scale.

Turning to inoculation, another potential prevention against smallpox, only about 50 000 individuals were inoculated between 1750 and 1800 in Sweden (Sköld, 1997). If inoculation provided total safety against death from smallpox, some 7500 would have been saved, assuming a fatality rate of 15%. This saving should be compared with the 300 000 who died from smallpox in Sweden during this period. Thus neither inoculation nor vaccination was of major importance for the initial mortality decline but it is possible that vaccination ensured that the disease never came back. Whether it also eradicated smallpox is, however, doubtful, since the disease was already showing a tendency to become milder over ti1me.

The improvement of personal hygiene may have had some effect on mortality in England and Wales after about 1880, when a decline in intestinal infections coincided with substantial improvements in water supply and sewage control, which must have reduced the incidence of waterborne infections. Since changes in mortality from intestinal infections were only a smaller part of the general mortality decline prior to, say, AD 1870, the general reduction must have been mostly due to factors other than improved personal hygiene, according to McKeown. For the same reason the type of public health measures discussed above must have had little impact on the great mortality decline prior to 1870 and only a partial effect afterward. Other public health measures (such as the breastfeeding campaigns that took place in Sweden in the 1830s and later) were not discussed by McKeown. Fridlizus (1984), however, has argued that for Sweden, these campaigns might very well have had an impact on infant mortality, but started much later than the decline in mortality. Incidentally, childhood mortality went up somewhat during the period when the breastfeeding campaigns began.

Helleiner argued in 1957 that the population of Western Europe must have increased from the mid-eleventh century to the late thirteenth century and from the mid-fifteenth century to the end of the sixteenth century. The population increase observed in the eighteenth century was therefore not unique except that mortality decline started from a higher level and went on longer than before. Helleiner and other scholars have argued that these changes in mortality were spontaneous (natural), and in 1973 the UN added natural factors as a fifth determinant of the great mortality decline. Later, Fridlizus (1984), Perrenoud (1984), and Schofield (1984) asserted

that a change in the virulence of pathogens initiated the great mortality decline. The basic idea is that the virulence of pathogens changes spontaneously over time. Virulence in an organism is generally understood as its ability to overcome host defences. At this point it is important to note that some pathogens develop more quickly in a malnourished host while others do not depend on such circumstances to produce very high mortality. Influenza epidemics, for example, are neither created nor sustained in malnourished persons.

McKeown's main argument on this issue is that the initial decline in mortality during the latter part of the eighteenth century was an integral part of the great mortality decline, which continued for the next two centuries; and therefore it cannot be attributed to a spontaneous reduction in the virulence of pathogens. According to McKeown, only nutrition remains as a plausible explanation not only for the decline in infectious diseases from the mid-nineteenth century onwards but also for the initial phase of the mortality decline. This reasoning is part of McKeown's attempt to find one explanation for mortality decline as a whole.

While Helleiner regarded the virulence hypothesis as a period explanation outside of man's control, Kunitz (1986) did not agree. He argued that the disappearance of epidemics, such as plague, typhus, and smallpox, is associated with the growth of stable governments, the expansion and integration of nation states, and the emergence of disciplined armies in the seventeenth and eighteenth centuries. These advancements made it possible to control epidemics by means of isolation and better hygiene. Whether this also applies to smallpox is not clear, since families with children having the disease were never isolated, at least not in Sweden, even though such ideas were put forward (Sköld, 1997). Kunitz also put forward the argument that population growth and loss of regional and local isolation converted epidemics into relatively more benign childhood diseases including the nonspecific pneumonia–diarrhoea complex, which became relatively more important when epidemic diseases declined. As a consequence, mortality became more dependent on nutritional status and adequate adequate nursing. Poverty, crowding, poor sanitation, and hazardous work conditions became likewise more important as determinants of death, particularly in the cities. As a result, the eighteenth century saw a divergence in the life expectancy of the upper and lower classes, according to Kunitz. His empirical evidence for this conclusion is, however, rather weak since it is only based on a comparison of aggregate infant mortality rates in England, France, Sweden, Germany, and Spain, and with rough estimates of the standard of living in these countries at the end of the eighteenth century.

The first part of Kunitz' explanation, namely the control of epidemics by means of isolation and better hygiene, is much in line with Easterlin's argument that technology (in a broad sense) was the main determinant of the great mortality decline (Easterlin, 1999). Easterlin argued that the mortality decline from the late nineteenth century onwards was not due to economic growth and market developments but to public policy initiatives and new institutions based on new knowledge of disease. This could explain the simultaneous mortality patterns in countries with clear differences in economic development, in the latter part of the nineteenth century onwards.

Cohort Explanations

Cohort explanations of the mortality decline refer to features that initially affected only certain young age groups but that may have long-lasting effects for the groups involved. Net nutrition is seen as the main determinant of the development of cells, which is most rapid during the fetal stage and which diminishes gradually until the body is fully developed at an age around 20. Net nutrition is what is left for the development of cells, after the nutritional claims for other life-supporting activities and work are accounted for. A low net nutrition could thus either be due to low nutritional intake or to additional claims coming from diseases. Also, many diseases not only make their own claims on nutrients but also make it more difficult for the body to absorb nutrients in general. This is the case for infectious diseases. If cells and organs do not develop, children may become stunted, shorter in general, and less healthy. Thus we can basically differentiate between two types of cohort explanations for the mortality decline, namely increased nutritional intake during the fetal stage and/or early years of life; and decreased claims during the fetal stage or early years of life due to a reduction in diseases in the mother or the child.

The importance of early childhood conditions for health in later life has probably been well known since time immemorial. The idea is that each generation seems to carry along the same relative mortality from childhood to old age. This was also what the first studies of the great mortality decline emphasized (Kermack et al., 1934). Improvement in health among infants and children brought about a general decline in mortality. The focus then turned to period factors in the 1950s, as discussed above. Both medical and historical research in this area has expanded rapidly over the last couple of decades (for overviews, see Elo and Preston, 1992; Kuh and Ben-Schlomo, 2004). In this connection, the work by Barker (1994) has been of major importance. He has summarized the medical evidence of the importance of nutrition for the fetus and the new-born infant for the health of the adult. Fogel (1996) has advocated these ideas in historical research perhaps more than anyone else. In addition, Preston and van de Walle (1978), for urban France, and Fridlizius (1989), for Sweden, emphasized the importance of cohort factors in mortality decline, though more in the vein of Kermack and his colleagues.

Steckel (1983) and Fogel (1996 and previous work) use final body height as a measure of net nutrition and health during childhood. Individuals who had well nourished and healthy mothers and were well nourished themselves during the fetal stage, have a lower risk of death during infancy. If they are well nourished and healthy, their cells and organs develop better and they are taller and live a longer life. Since net nutrition and not gross nutrition determines health, there is no direct link between gross nutrition and height. Improvements in health and height may either be a result of better nutrition (better diet), or fewer claims due to diseases. Thus a decline in the prevalence of smallpox, for example, has a positive effect on height and extends the lifespan, everything else being equal. The problem is that it is difficult to evaluate how much of the improvement in health is due to diet and how much is due to other claims. Calculating diets for premodern populations is a difficult task (Fogel, 1996), and it is even more difficult to calculate claims. Still, Fogel and others claim that historical

records, fragmentary as they are, show similarities between trends in height and gross consumption, indicating that the trend in claims has been of minor importance (Fogel, 1994, 1996; Floud et al., 2010).

Undernourishment, whether due to a low or badly composed intake or due to an increased claim on nourishment because of disease, may lead to a stunting of height or weight as well as to illness, disease, and mortality later in life, rather than increasing mortality in the immediate future. The immediate relationship – i.e., the period link between economy and mortality – is therefore much weaker than Malthus believed, according to Fogel. The rather weak short-term relationship often found between prices and death for many European countries, shown by Lee (1981), Galloway (1988), and others, supports this interpretation (Fogel, 1994). Thus, cohort factors matter more to the mortality decline than period ones, according to Fogel.

A warning against an uncritical use of cohort mortality notions is implied in the findings of Fridlizius (1989), who has shown that the mortality of the Swedish cohorts born in the beginning of the nineteenth century deviated greatly from the anticipated mortality trends, given their mortality at younger ages. He called them 'deformed cohorts' and saw the excess mortality for men as due to high levels of alcoholism during the 1830s and 1840s. Such life-styles lifted their mortality above their 'programed' level. Other analyses of aggregate cohort mortality data also bring into question, any strong cohort effect of conditions prevailing in early life.

While Fogel was mainly concerned with chronic malnutrition, we may also note that temporary disturbances in nutrition, in particular during some periods at the fetal stage, may also have long-lasting effects on health, since later improvements cannot always compensate for prior loss (Barker, 1994). Therefore, temporary malnutrition due to lack of nutrients or to an increased disease load, whether of the mother or of the child, may have long-lasting effects on health and lifespan. It has, for example, been shown that in Sweden the disease load in the first year of life has a strong impact on mortality from airborne infections and other diseases in later life (Bengtsson and Lindström, 2000). Consequently, elimination of temporary malnutrition, which is still a major problem in many countries of the world, may cause another long-term cohort decline in mortality.

Mortality Fluctuations

Another feature of the great mortality decline is that mortality fluctuations became much smaller in the nineteenth century than they were previously. Views on its consequences for the balance between the population and the economy in preindustrial society and its effects on mortality decline have been controversial. There are two schools of thought and they have reached very different conclusions. According to those who analyze mortality crises (Meuvret, 1946; Goubert, 1965), the reduction in such crises was a major factor behind the secular decline in mortality. This view gained support during the 1960s and 1970s. Wrigley and Schofield (1981), in their analysis based on a large national sample of English parishes, showed, however, that the elimination of crises accounted for only a small fraction (about 10%) of the secular decline in mortality. Many others also obtained similar results (see Ortega Osona, 2000, for an overview). This view is now questioned by Ortega Osona, who has found that the conclusion of Wrigley and Schofield holds true for England but not for Northern Italy and New Castile, where the reduction of the variability in mortality was found to be important for the initial decline in mortality. Analyses of determinants of mortality fluctuations, by age and gender and for different socio-economic groups using longitudinal individual-level data, have expanded in the last decade (Bengtsson et al., 2004) from previous analyses made at the macro level without such details (Lee, 1981; Galloway, 1988).

The Socioeconomic Gradient in Mortality

In the last decades, increasing health differentials between social groups have received a lot of attention (Marmot, 2004). It has been argued that we become sick not only because of low income, but also because of increasing income differentials (Wilkinson, 1996; Wilkinson and Picket, 2006). Social differences in health have in recent times diverged not only in low-income but also in high-income countries. This is in stark contrast to the early postwar period when improvements in life expectancy in many countries made health authorities, health scientists, and politicians believe that social differences in mortality had started to converge. The conclusion at the time was that inequality, when measured as death rates, was on a steady decline, possibly even on the brink of disappearing; the reason being access to prevention and medication at low costs. Yet another view was that class differences in mortality had always existed but had remained more or less constant (Cassel, 1976; Link and Phelan, 1996; Marmot, 2004). The underlying hypothesis is that those who have greater access to economic resources will always be less afflicted, regardless of the nature of the disease (Link and Phelan, 1996).

The question is then whether health differentials have converged, diverged, or stayed more or less constant. Due to a lack of coherent data covering long time periods, our knowledge has been based on highly incomplete evidence from a variety of locations at various points in time. A recent publication, however, summarizes the results from seven new studies of locations in Western and Southern Europe, the United States, and Canada, for which individual-level longitudinal data exists for long time periods (Bengtsson and van Poppel, 2011). Most of these studies also cover large parts of the twentieth century, for which such micro data hitherto had largely been lacking. Taken together, they have a wide geographic coverage and a very long time horizon.

Geneva, with a labor force of more than 20 000 in the watch-making industry alone in 1800, saw huge but declining social differences in mortality during its nineteenth-century industrialization (Schumacher and Oris, 2011). The workers in the sawmill town of Sundsvall in Northern Sweden faced higher mortality than farmers, but had the same levels of mortality as the upper classes in the town itself (Edvinsson and Lindkvist, 2011). In Saguenay, Quebec, with a growing aluminum and pulp industry, mortality was about the same for all groups except farmers, who had significantly lower mortality up until recent times (Gagnon et al., 2010). In a rural area in Scania in Southern Sweden, with some

emerging industry after 1860, workers and the affluent farmers had the same mortality in working ages up until 1950, when they diverged (Bengtsson and Dribe, 2011). In Alghero, a port town in Sardinia, the social differences in mortality among children were substantial but did not increase as it became industrialized (Breschi et al., 2010). Instead, these social differences either stayed constant or declined. The situation in the Netherlands, for which data covers a large part of the country, both urban and rural, was similar. The social differences found for earlier cohorts remained the same or diminished slightly over time (Schenk and van Poppel, 2010). While the absolute differences in infant and child mortality between social groups in the United States have diminished over time, the relative differences persist and have possibly expanded a bit (Haines, 2010).

Taken together, in no case is it possible to find evidence for worsening conditions leading to higher mortality and/or larger social differences in mortality during the industrialization phase. In one case in rural Scania, we find increasing social differences in adult mortality after 1950. A social gradient in mortality is indeed found in urban areas, notably Geneva and Alghero, but these differences date further back in time, and in Sundsvall there were no differences at all. Instead, mortality was higher in urban than in rural areas for all social groups, a common finding referred to as the urban mortality penalty. The finding that social differences in mortality were small (except in cities), in preindustrial agricultural societies that were included in the studies mentioned above, is consistent with findings for historical England both at the macro level (Wrigley and Schofield, 1981; Hollingsworth, 1957; Livi-Bacci, 1991) and at the micro level (Clark, 2007).

The reason for the lack of social differences in the past is most probably because mortality in communicable and often virulent diseases was high during this period. Almost all groups were exposed, and medical treatment was inefficient. The difference between urban and rural areas was therefore probably the result of differences in exposure, not resistance. Today, the disease pattern is entirely different, even though pandemics strike now and then. Life style factors, such as smoking, diet, and exercise, have become increasingly important. Diseases are treatable and means of prevention are known.

Summary

Few scholars today will argue that any single factor was the main determinant of the great mortality decline. Of course, it is no coincidence that the greater resources made available by the transformation of our economies in the eighteenth and nineteenth centuries runs parallel with the great mortality decline and with the fertility transition as well. In a millennial perspective, these events took place at about the same time. That is not to say that there is a close relationship between the economy and the great mortality decline. On the contrary, economic growth is not likely to be even a major determinant, either before the decline or in its initial stages, and was also perhaps far less than expected during later stages. Instead, the causes are multifactorial and vary from the start to the end of the decline.

Parts of the initial decline may very well have been the result of pure luck – a spontaneous decline in the aggressiveness of smallpox pathogens that were a part of an old demographic pattern – and not a result of a modernization process or vaccination. The development and compulsory use of vaccines probably did help prevent the return of the disease. Improvements in nourishment and in the care of children had long-lasting effects on the lifespan. Advancements in water supply and sanitation, as well as better housing, contributed to the decline from the second part of the eighteenth century onwards. Medical development in the twentieth century prolonged life. The fact that health is determined by net nutrition, defined as nutritional intake minus claims, and that claims are due partly to diseases, makes it difficult to evaluate the determinants of the mortality decline. The influence of conditions in childhood and other temporary factors on mortality at older ages, also adds to the complexity of the issue. The impact of each individual factor is therefore very difficult to measure and requires long time series of high-quality data. Several variables, like the virulence of pathogens and the claims on nutrients due to disease, can at best be estimated indirectly, if at all. The analyses based on highly aggregated longitudinal data have identified the problems and directed us toward the solutions. It is, however, doubtful that solutions will be found at that level of analysis. Perhaps the use of longitudinal microdata will serve as an important complement, as it has in other areas.

Thus, the cause of the great mortality decline is multifactorial and the importance of the various factors changes over time. These factors also influence each other. For example, health is both a cause and a consequence of net nutrition, and this makes it extremely difficult to measure the impacts of each factor. Still, it seems very unlikely today that the initial phase of the great mortality decline was mainly a result of an increasing standard of living and of intervention by man. From the late nineteenth century onwards, it is likely that improved nourishment and various interventions had an impact on the decline. One might therefore argue that the initial decline was not a result of modern developments but rather part of an old pattern, where the level of mortality varied. If this pattern had continued, one could have anticipated an upswing in mortality during the latter part of the eighteenth century, and in fact such an upswing started in several European countries around the middle of the century, although it was limited. At present, we know that improved nourishment, personal care, and various public health measures mattered. Thus, we may speculate that it was not the initial mortality decline that was a result of the development of modern society, but rather the absence of an otherwise expected increase. This leads us to a new generalization of the demographic transition, where the pretransition stage is characterized by changing levels of mortality that were not controlled by man, with population growth varying, as did nuptiality and fertility. From around the middle of the eighteenth century this pattern was broken and man began to determine the levels of mortality as well as fertility.

See also: Adult Mortality in Developing Countries; Adult Mortality in Industrialized Societies; Early Life Influences on Health and Mortality in Adulthood; Infant and Child Mortality in Industrialized Countries; Infant and Child Mortality in the Less Developed World; Malthus, Thomas Robert (1766–1834); Mortality,

Biodemography of; Mortality, Epidemiological, and Health Transitions; Period and Cohort Analysis in Demography; Public Health as Social Science; Second Demographic Transition.

Bibliography

Barker, D.J.P., 1994. Mothers, Babies and Disease in Later Life. BMJ, London.

Bengtsson, T., 1992. Den demografiska utvecklingen i de nordiska länderna 1720–1914. Lund Papers in Economic History, 15. Department of Economic History, Lund.

Bengtsson, T., Dribe, M., 2011. The late emergence of socioeconomic mortality differentials: a micro-level study of adult mortality in southern Sweden 1815–1968. Explorations in Economic History 48, 389–400.

Bengtsson, T., Lindström, M., 2000. Childhood misery and disease in later life: the effects on mortality in old-age of hazards experienced in early life, southern Sweden, 1760–1894. Population Studies 54 (2), 263–277.

Bengtsson, T., Lindström, M., 2003. Airborne infectious diseases during infancy and mortality in later life in southern Sweden, 1766–1894. International Journal of Epidemiology 32 (2), 286–294.

Bengtsson, T., Oeppen, J., 1993. A Reconstruction of the Population of Scania 1650–1760. Lund Papers in Economic History, 32. Department of Economic History, Lund.

Bengtsson, T., Ohlsson, R., 1994. The demographic transition revised. In: Bengtsson, T. (Ed.), Population, Economy, and Welfare in Sweden. Springer-Verlag, Berlin, pp. 13–36.

Bengtsson, T., van Poppel, F., 2011. Socioeconomic inequalities in death from past to present: an introduction. Explorations in Economic History 48, 342–356.

Bengtsson, T., Campbell, C., Lee, J.Z. (Eds.), 2004. Life Under Pressure: Mortality and Living Standards in Europe and Asia 1700–1900. MIT Press, Cambridge, MA.

Breschi, M., Fornasin, A., Manfredini, S., Mazzoni, L., Pozzi, L., 2010. Socioeconomic conditions, health and mortality from birth to adulthood, Alghero 1866–1925. Explorations in Economic History 48 (3), 366–375.

Cassel, J., 1976. The contribution of the social environment to host resistance: the fourth Wade Hampton Frost Lecture. American Journal of Epidemiology 104 (2), 107–123.

Clark, G., 2007. A Farewell to Alms: A Brief Economic History of the World. Princeton University Press, Princeton, NJ.

Davis, K., 1945. The world demographic transition. Annals of the American Academy of Political and Social Sciences 237, 1–11.

Easterlin, R.A., 1999. How beneficent is the market? A look at the modern history of mortality. European Review of Economic History 3 (3), 257–294.

Edvinsson, S., Lindkvist, M., 2011. Wealth and health in 19th Century Sweden: a study of social differences in mortality in the Sundsvall region. Explorations in Economic History 48, 376–388.

Elo, I., Preston, S., 1992. Effects of early-life conditions on adult mortality: a review. Population Index 59 (2), 186–212.

Floud, R., Fogel, R.W., Harris, B., Sok, C.H., 2010. The Changing Body. Health, Nutrition, and Human Development in the Western World Since 1700. Cambridge University Press, Cambridge.

Fogel, R.W., 1994. The relevance of Malthus for the study of mortality today: long-run influences on health, mortality, labour force participation, and population growth. In: Lindahl-Kiessling, K., Landberg, H. (Eds.), Population, Economic Development, and the Environment. Oxford University Press, Oxford, pp. 231–284.

Fogel, R.W., 1996. The Escape from Hunger and Premature Death 1700–2100: Europe, America and the Third World. The 1996 Ellen McArthur Lectures, Cambridge University Press, Cambridge.

Fridlizius, G., 1989. The deformation of cohorts: nineteenth-century mortality decline in a generational perspective. Scandinavian Economic History Review 37 (3), 3–17.

Fridlizus, G., 1984. The mortality decline in the first phase of the demographic transition: Swedish experiences. In: Bengtsson, T., Fridlizius, G., Ohlsson, R. (Eds.), Pre-industrial Population Change: The Mortality Decline and Short-term Population Movements. Almqvist & Wiksell International, Lund, Sweden, pp. 71–114.

Gagnon, A., Tremblay, M., Vézina, H., Seabrook, J.A., 2010. Once were farmers: occupation, social mobility, and mortality during industrialization in Saguenay-Lac-Saint-Jean, Quebec 1840–1971. Explorations in Economic History 48 (3), 429–440.

Galloway, P.R., 1988. Basic patterns in annual variations in fertility, nuptiality, mortality and prices in pre-industrial Europe. Population Studies 42, 275–302.

Galloway, P.R., 1994. A reconstruction of the population of North Italy from 1650 to 1881 using annual inverse projection with comparison to England, France, and Sweden. European Journal of Population 10, 223–274.

Goubert, P., 1965. Recent theories and research in French population between 1500 and 1700. In: Glass, D.V., Eversley, D.E.C. (Eds.), Population in History: Essays in Historical Demography. Arnold, London, pp. 457–473.

Haines, M., 2010. Inequality and infant and child mortality in the United States in the twentieth century. Explorations in Economic History 48 (3), 418–428.

Helleiner, K.F., 1957. The vital revolution reconsidered. Canadian Journal of Economics and Political Science 23, 1–9.

Hollingsworth, T.H., 1957. A demographic study of the British ducal families. Population Studies 11, 4–26.

Kermack, W., McKendrick, A., McKinlay, P., March 1934. Death rates in Great Britain and Sweden: some regularities and their significance. Lancet, 698–703.

Kuh, D., Ben-Schloma, Y. (Eds.), 2004. A Life Course Approach to Chronic Disease Epidemiology. Oxford University Press, Oxford.

Kunitz, S.J., 1986. Mortality since malthus. In: Coleman, D., Schofield, R.S. (Eds.), The State of Population Theory: Forward from Malthus. Blackwell, Oxford, pp. 279–302.

Lee, R.D., 1981. Short-term variation: vital rates, prices and weather. In: Wrigley, E.A., Schofield, R.S. (Eds.), The Population History of England, 1541–1871: A Reconstruction. Arnold, London, pp. 356–401.

Link, B.G., Phelan, J.C., 1996. Understanding sociodemographic differences in health-the role of fundamental social causes. American Journal of Public Health 86 (4), 471–473.

Link, B.G., Northridge, M.E., Phelan, J.C., Ganz, M.L., 1998. Social epidemiology and the fundamental cause concept: on the structuring of effective cancer screens by socioeconomic status. Milbank Quarterly 76 (3), 375–402.

Livi-Bacci, M., 1991. Population and Nutrition. An Essay on European Demographic History. Cambridge University Press, Cambridge.

Marmot, M., 2004. The Status Syndrome: How Social Standing Affects Our Health and Longevity. Bloomsbury Publication, London.

McKeown, T., 1976. The Modern Rise of Population. Arnold, London.

Meuvret, J., 1946. Les crises de subsistence et la demographie de la France d'Ancien Regime. Population 1, 643–650.

Notestein, F.W., 1953. Economic problems of economic change. In: Proceedings of the Eighth International Conference of Agricultural Economics. London.

Oeppen, J., Vaupel, J., 2002. Broken limits to life expectancy. Science, 296 (5570), 1029–1031.

Omran, A.R., 1971. The epidemiological transition. A theory of the epidemiology of population change. Millbank Memorial Fund Quarterly 49 (4), 509–538.

Ortega Osona, J.A., 2000. Determinants of mortality variability in historical populations and its behavioural and aggregate consequences. In: Bengtsson, T., Saito, O. (Eds.), Population and Economy. From Hunger to Modern Economic Growth. Oxford University Press, Oxford, pp. 279–300.

Perrenoud, A., 1984. The mortality decline in a long term perspective. In: Bengtsson, T., Fridlizius, G., Ohlsson, R. (Eds.), Pre-industrial Population Change: The Mortality Decline and Short-term Population Movements. Almqvist & Wiksell International, Lund, Sweden, pp. 41–70.

Preston, S.H., van de Walle, E., 1978. Urban French history in the nineteenth century. Population Studies 32, 275–297.

Quaranta, L., 2013. Scarred for life. How conditions in early life affect socioeconomic status, reproduction and mortality in southern Sweden, 1813–1968. Lund Studies in Economic History, No. 59, Lund.

Schenk, N., van Poppel, F., 2010. Social class, social mobility and mortality in the Netherlands, 1850–2004. Explorations in Economic History 48 (3), 401–417.

Schofield, R.S., 1984. Population growth in the century after 1750: the role of mortality decline. In: Bengtsson, T., Fridlizius, G., Ohlsson, R. (Eds.), Pre-industrial Population Change: The Mortality Decline and Short-term Population Movements. Almqvist & Wiksell International, Lund, Sweden, pp. 17–40.

Schumacher, R., Oris, M., 2011. Long-term changes in social mortality differentials, Geneva, 1625–2004. Explorations in Economic History 48, 357–365.

Sen, A., 1998. Mortality as an indicator of economic success and failure. The Economic Journal 108, 1–25.

Sköld, P., 1997. The Two Faces of Smallpox. A Disease and Its Prevention in Eighteenth- and Nineteenth-Century Sweden. Sweden: Report no. 12 from the Demographic Data Base. Umeå University.

Steckel, R., 1983. Height and per capita income. Historical Methods 16, 1–7.

United Nations, 1953. The determinants and consequences of population trends. Population Studies 17.

United Nations, 1973. The determinants and causes of population trends. Population Studies 50.

Wilkinson, R.G., 1996. Unhealthy Societies: The Afflictions of Inequality. Routledge, London.

Wilkinson, R.G., Picket, K.E., 2006. Income inequality and population health: a review and explanations of the evidence. Social Science and Medicine 62 (7), 1768–1784.

Wrigley, E.A., Schofield, R.S., 1981. The Population History of England, 1541–1871: A Reconstruction. Arnold, London.

Wrigley, E.A., Davies, R.S., Oeppen, J.E., Schofield, R.S., 1997. English Population History from Family Reconstitution, 1580–1837. Cambridge University Press, Cambridge.

Mother-Child Communication: Cultural Differences

Carolin Demuth, Aalborg Universitet, Aalborg, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Communication with children plays a crucial role not only for cognitive and social-emotional development but also in a more general sense for an understanding of self and self in relation to others. Research from linguistic anthropology and cultural developmental psychology have shown that there exists a great variety of cultural genres of communicating with children that are in line with the relevant broader cultural ideologies of good child care. Culture, communication, and self-development are inextricably intertwined. Culturally distinct communicative practices in which children participate will therefore ultimately lead to different cultural developmental pathways. While traditional research in developmental psychology has focused on mother–child dyads and experimental designs there is an increasing recognition of the need for naturalistic studies of everyday communication with children including their broader social network.

This article addresses culture-based differences in mother–child communication. It gives an overview of studies conducted in various societies across the world and with children at different ages. It aims at pointing to distinct developmental pathways that will evolve from culture-based differences in communicative practices with children. It stresses the importance of studying natural, everyday interactions in which children participate for better understanding the intertwined nature of culture and child development.

Mother–child communication constitutes a central topic in social and behavioral sciences, specifically in developmental research. Communication with children plays a crucial role for cognitive and social-emotional development. Research in the field of psychology has primarily focused on the impact of mother–child communication on socioemotional development/secure attachment (*see* Attachment during Infancy and Early Childhood: Cultural Variations, Development of), language acquisition, and general cognitive development (*see* Biological Preconditions for Language Development; Language Development, Theories of; Language Development: Emergentist Theories; Language Development: Evolutionary Theories; Child-Directed Speech: Influence on Language Development; Theory of Mind during Infancy and Early Childhood Across Cultures, Development of). Communication with children constitutes, however, a much more crucial determinant for child development in that it is the central means to construct social meaning and an understanding of self and self in relation to others. Social interaction with children can be considered the crucial place where development takes place and communication is a central part of this. Culturally different communicative practices in which children participate will hence ultimately lead to different developmental pathways of selfhood and social understanding (*see* Attachment during Infancy and Early Childhood: Cultural Variations, Development of; Joint Attention during Infancy and Early Childhood Across Cultures; Early Emotional Development and Cultural Variability; Self in Culture: Early Development; Values Across Cultures, Development of; Early Memory Development and Cultural Variations; Cognitive Development during Infancy and Early Childhood across Cultures; Memory Development during Infancy and Early Childhood across Cultures; Autobiographical Memory during Early Childhood and Cultural Variations; Concept Formation in Childhood: Cross-Cultural Perspectives; Theory of Mind during Infancy and Early Childhood Across Cultures, Development of).

Mother–Child Communication in Early Infancy

Infants readily engage in communication and social interaction from birth on. Around 2–3 months infants are able to maintain eye-to-eye contact during social interaction and smile to interactional partners more frequently. This in turn allows caregivers to communicate with their infants in more contingent patterns, for example, by initiating turn-taking and interpreting the infant's vocalization and physical movements as meaningful and responding to it accordingly. This pattern has been described as *protoconversation* by Mary C. Bateson (*see* Child-Directed Speech: Influence on Language Development; Language Development, Theories of) and assumed to have a biological component and therefore to show same characteristics in many different languages, such as high voice pitch, exaggerated intonation, slow tempo, repetitiveness, and vowel elongation. The large majority of studies on infant-directed speech/protoconversation was conducted in white, middle-class North American or northern European contexts, that is, within the same sociocultural context of the researchers, and hence might have lead to a methodological artifact that largely ignored culture. The universality of this communicative pattern has meanwhile been challenged and a number of psychological and anthropological studies have found that these communicative patterns are not characteristic for many non-Western cultures (for an overview see Ochs and Schieffelin, 1984; *see* First Language Acquisition, Developmental Psychology of). Extended mock conversations and play dialogues, typical for Western societies, is virtually absent also among the Kaluli of Papua New Guinea and the Samoans of Polynesia (Ochs and Schieffelin, 1984). In cultural societies such as the Gusii in Kenya (LeVine et al., 1994), Kaluli in Papua New Guinea and Western Samoans (Ochs, 1982;

Ochs and Schieffelin, 1984; Schieffelin and Ochs, 1986), or Mayan societies (cf De León, 2000), caregivers do not engage in conversations including dyadic turn-taking with their infants. Mayan Zinacantec babies do however have the status of 'protoaddressees' (De León, 2000), that is, they are addressed with speech in a variety ways, including for instance rhetorical questions, formal address in triadic interaction, speech activities to control the child's behavior. Mayan Zinacantec interactions with children furthermore are characterized to a larger extend by nonverbal communication. Haviland (1998) could show that there is no radical discontinuity between gesture and language in this community and that gestures play a more important role for language acquisition than in Western societies.

First instances of turn-taking have been observed during breast-feeding right after birth (Kaye, 1982; cf Takada, 2011) in that infants stop after having suck for a few seconds, followed by mothers 'jiggling' the infant or breast and the infant resuming sucking. While these findings might suggest some innate and hence universal component of mother–infant interactions, no such pattern was found, for example, among the San in South Africa (Takada, 2011).

Cowley et al. (2004) found that in KwaZulu communities in South Africa the caregiver's actions chime with those of the infant creating an impression of behaving in unison. Often this is highly rhythmic and characterized by frequent vocal overlap or chorusing rather than dyadically following a turn-taking pattern. They suggest to speak of 'protosong' rather than of 'protoconversation' to describe a communicative pattern with infants in this community. Similar findings have been reported by a study among the farming Nso in the Northwester Grassfields of Cameroon (Demuth, 2008).

Also typical communicative patterns with infants in the United States such as engaging and exciting are not common in many other societies. LeVine et al. (1994), for instance, describe that soothing and calming has been the common pattern in infant care among the Gusii, agricultural people in Western Kenya.

French mothers have been found to organize conversational exchanges with their infants in a dyadic form, using mirroring devices, and addressing questions to the child, whereas Senegalese Wolof immigrant mothers were found to include third persons in interactions with their infants and to use more singing and rhyming as well as more imperatives (Rabain-Jamin and Sabeau-Jouannet, 1997).

A study by Gratier (2003) on spontaneous vocal interactions of mothers and their 2–5-month-old infants from India, France, and the United States found both communalities among the three cultural contexts with regard to general rhythmic aspects in the coordination between mother and infant vocalization. On the other hand, the study also revealed significant cultural variability: the Indian mothers produced less space between vocal turns and more overlap of mother and baby vocalizations. They also produced a higher ratio of nonverbal to verbal vocalizations. Moreover, the vocal interaction of immigrant dyads living in the United States showed signs of change in the direction of the host culture.

From an ecocultural perspective, cultural variations in communication with infants can be considered as sociohistorically evolved strategies that are inherently organized by cultural parental beliefs and adaptive to the specific ecocultural pressures and requirements of a given sociocultural environment (Keller, 2007; *see* Ecocultural Theories of Development; Development of Attachment during Infancy and Early Childhood: Cultural Variations). They can be understood as cultural expressions of what might be a basic innate communicative pattern with infants. Besides the caregiver's personality and psychological functioning, therefore communication with children always needs to be understood as serving broader cultural socialization goals (*see* Parenting Attitudes and Beliefs across Cultures). Schieffelin and Ochs (1998) have argued, for instance, that the accommodations to the child reflect a discomfort of Western middle-class mothers with the competence differential between adult and the child and that caregivers use self-lowering (e.g., simplified speech) as well as child-raising strategies (e.g., acting as if the child were more competent than his or her behavior more strictly would indicate) to reduce this competence gap. In some societies, age and status differences can, for example, affect the rights to turns in a conversation (Keating and Egbert, 2004). Among the Gusii, verbal exchange is largely restricted to persons whose kin relationship defines them as social equals. Parents and children, however, are considered inherently unequal even in adulthood and ideally distant as communicative partners (LeVine et al., 1994).

Within attachment theory, for instance the quality of communication with infants during the first years of life has been claimed to have an impact on secure attachment. Mothers who are aware of an infant's signals, accurately interpret them, and promptly respond by showing empathy and insight into the baby's wishes and moods have been claimed to promote secure attachment development in their child. However, these infant care practices correspond to the general Western conception of maternal sensitivity. Good motherhood might be defined differently in various sociocultural groups and other forms of maternal sensitivity and hence communicative practices might similarly lead to secure attachment development (*see* Attachment during Infancy and Early Childhood: Cultural Variations, Development of). According to the Cameroonian Nso beliefs, for instance, a good mother is supposed to know what is good for the child and to decide what is going to be done. There is therefore no 'need' to speculate on what might be going on in the child's mind. Not referring to the internal experience of the child has also been found in other non-Western cultural groups (e.g., Ochs and Schieffelin, 1984). For instance, the Kaluli in Papua New Guinea use little language except for greetings, imperatives, and rhetorical questions toward infants (Ochs and Schieffelin, 1984). This has been related to their belief that infants have no understanding yet and hence are not able to initiate appropriate interactional sequences. Western Samoans have been found to not directly address infants except in the form of songs or rhythmic vocalization (Ochs and Schieffelin, 1984). Moreover, societies such as Samoans (e.g., Ochs, 1982), Kaluli of Papua New Guinea (e.g., Schieffelin and Ochs, 1986), strongly dislike verbal speculation on what someone else might be thinking or feeling or to express personal thoughts also in adult conversations.

Moreover, nonverbal and physical features of communication are equally important, especially in nonliterate cultures (Nsamenang and Lamb, 1994). In this line of argumentation, Gratier (2003) has suggested to broaden the definition of protoconversation to include kinesthetic and tactile modes of interaction and conversational styles that are not centered on the turn-taking format. She states that a protoconversation may also simply consist of conversation without words. Similarly, Ochs and Schieffelin (1984) have proposed that caregiver–infant interactions do not follow a 'biologically designed choreography' (Stern, 1977) but that there are many different choreographies within and across societies.

Mother–Infant Communication as the Cradle of Cultural Meaning

Within the field of psychology, it was Vygotsky (1978) who first argued that social interaction, particularly language, is an important mechanism by which children learn to understand themselves. The sociocultural perspective emphasizes the idea that these interactions are dialogically intertwined with cultural ideologies, which essentially provides individuals with certain frameworks that influence how the individual is seen in relation to others.

The importance of mundane communication for child development becomes clear when looking at communicative practices as locally organized meaning-making processes. It is through reciprocal engagements with others, especially the mother, that infants come to develop a sense of shared experience (termed 'primary intersubjectivity'; Trevarthen, 1998), that is, finding a shared background with another person against which experience can be understood (Hundeide, 1993). By coordinating the mother's intentions with the child's needs and intentions, and interpreting and evaluating the infant's behaviors, both verbally and nonverbally, cultural meaning is dialogically constructed in that mother and infant tacitly negotiate 'contracts' (Hundeide, 1993) of what is acceptable and permissible. Through sharing common experience of what is 'normal' the infant is 'contractually' committed to present (or to position) himself in a certain way as part of the silent agreement of participation. Infant's awareness of reality and learning of culture is hence built through communication with the motives and emotions of other persons (Trevarthen, 1998; *see* Self in Culture: Early Development). As Forrester states: "The desire of any parent for his/her child to be a self expresses itself through culturally specific discourses: language as social action presupposed on the beliefs, narratives and everyday understandings of what is appropriate in context" (1999: p. 42). Through repeated exposure to these 'frames' and ways in which the infant is positioned in the interaction, a relatively stable self-construal, that is, perception of oneself and of oneself in relation to others in the society, is developed. The way the child constructs himself or herself discursively later on in childhood will reflect these relatively stable patterns. A case study on caregiver–infant interactions when the infant was 18–28 months old (Forrester, 2001), for example, could show that the child began to understand and locate herself in the membership categories provided and deemed appropriate within the specific cultural context's and provides insight on how the caregivers' discursive practices facilitated or constrained her emerging conception of what it is to be a self. The sense of self a child acquires hence is in large part determined by the models, metaphors, and constructs of self made available to the child in the discourse surrounding him or her.

There is empirical evidence that children at a very young age display cultural characteristics of conversational patterns and that even as early as at 2 (Gratier, 2003) and 3 (Cowley et al., 2004) months of age, infants already display communicative practices that are conform with the prevailing cultural models of how to relate with their caregivers. Drawing on the microanalysis of a South-African mother–infant interaction, Cowley et al. could demonstrate how a 3-month infant comes to grasp that his mother wants him to fall silent and thus learns to adjust to the expectations of his encultured caregiver (Cowley et al., 2004).

Research with infants from various cultural backgrounds revealed that as early as 6 and even 3 months of age infants' cry melodies show clear differences in the shape and tone of the babies cry melodies.

Mother–Child Communication in Childhood

While cultural ideologies of good child care are largely implicit, they become evident in the different strategies parents in diverse societies use to teach their children to behave well: Inuit parents, for instance, use teasing but would never use shaming or threatening strategies or beat the child (Briggs, 1982, 1998), while frightening and beating are common among the Mfantse in Ghana. However, they would never use shaming strategies publicly (Quinn, 2005). Shaming is, however, common in Taiwanese mothers' socialization strategies (e.g., Miller et al., 2012, 2007) and both, teasing and shaming is common among Kaluli (Papua New Guinea) (Schieffelin, 1986). Teasing was also found to be typically used by Euro-American white working-class mothers but not by Euro-American higher educated white middle-class mothers (Kusserow, 1999; Miller, 1986). Micronesian Ifaluk use anticipative strategies to avoid misbehavior of the child by teaching him or her to be fearful of certain types of ghosts (Lutz, 1988).

The dialogical nature of language acquisition and cultural understanding of the world has been first addressed by the linguistic anthropologists Elinor Ochs and Bambi Schieffelin (e.g., Ochs, 1982; Ochs and Schieffelin, 1984; Schieffelin and Ochs, 1986) within their language socialization paradigm: Through participation in mundane everyday communication, children not merely acquire language but also a cultural understanding of how to make sense of their experience and the world. The focus hence is on how a community's habitus of communicative practices and strategies with children is related to sociocultural orientations (Ochs et al., 2005).

Euro-American middle-class mothers tend to pay more attention to socializing their children into being active conversational partners (Blum-Kulka, 1990; Blum-Kulka and Sheffer, 1993; Junefelt and Tulviste, 1997), and to praise their children, fostering assertiveness, cognitive development, and stressing the child's individuality (e.g., Junefelt and Tulviste, 1997).

Mother–Child Coconstructions of Autobiographical Narratives

As infants become language-using toddlers, parents begin to engage in talk about past and future events with them. Children on their part start to tell stories of personal experience in interactions with others in the second or third year of their lives. A number of studies have therefore addressed the relation between mother–child communication and the development of autobiographical remembering and storytelling (see, e.g., Nelson and Fivush, 2004 for an overview). Self-narratives can be considered to have a crucial function for identity-formation and self-construal (Bruner, 1990). Through repeated participation in communication with caregivers children learn how to tell stories about themselves in culturally appropriate ways and have internalized what constitutes a 'good story' that is in line with broader cultural ideologies. By participating in these practices, children not only acquire cognitive skills like telling a story but most of all they learn to make sense of the world and of themselves – they acquire cultural tools to develop identities, affective stances, values, and moral orientations that are in line with broader cultural ideologies of a community (e.g., Miller et al., 2007). Learning to engage in storytelling hence is inseparably bound to cultural ways of constructing a sense of self.

Cross-Cultural and Quasi-Experimental Studies

A number of quasi-experimental studies have found cultural differences in the way mothers coconstruct stories about past events in elicited interaction with their children as well as differences in how children on their part tell stories about their experience (*see* Autobiographical Memory during Early Childhood and Cultural Variations). Mainly, these differences have been related to different cultural self-construals and differences in the emphasis of the dimensions independence and interdependence or autonomy and interpersonal relatedness. Mothers from different cultural contexts may have different goals regarding the purpose of remembering, and these goals further influence the way in which they speak to their children.

Euro-American mothers, for instance, have been found to use a high-elaborative child-centered conversation style that has been shown to be associated with their values regarding an autonomous, individual identity (Wang, 2007). Euro-American children on their part produce detailed autobiographical narratives and independent self-construals. Chinese mothers tend to use a more didactic, mother-centered approach, focusing on appropriate behaviors, rules, and social obligations which has been shown to be associated with confucian thinking and their values regarding interdependence (e.g., Wang, 2006, 2007, 2008). Chinese children on their part show greater orientation toward social engagement, greater concern with moral correctness and authority, less autonomous orientation, more expressions of emotions, and more situational details in their accounts compared with their US counterparts.

Mothers in highly educated Western urban middle-class societies (defined as autonomous context) have been found to adopt an elaborative-evaluative reminiscing style and person-oriented approach in their conversations with toddlers, whereas mothers from non-Western subsistence-based communities (defined as relatedness context) have been found to adopt a repetitive reminiscing style. Children in the prototypical autonomous contexts accordingly showed a highly elaborated style in narrating autobiographical memories whereas children in the prototypical relatedness context provide low elaboration in their autobiographical accounts (Schröder et al., 2013; Tõugu et al., 2011, 2012).

A number of studies have investigated mother–child communication during family mealtime interactions in different cultural communities. Blum-Kulka (1990, 1997) found that American mothers of schoolchildren paid considerably more attention to following conversational norms and turn-taking at the dinner table than mothers from Israel. Euro-American mothers have been found to praise their children more often than Estonian and Swedish mothers whereas Estonian mothers have been found to use more regulation than Euro-American and Swedish mothers (Junefelt and Tulviste, 1997). Finno-Ugric societies have been found to generally talk less and accordingly also talk less with their children (Tulviste, 2004b). Estonian mothers have, however, been found to use a highly directive conversational style with a high proportion of imperatives (Tulviste, 2004a; Tulviste et al., 2004) compared to Finnish, Swedish, and Latvian mothers.

Discursive Practice Approach

While most of the above studies have been conducted within a traditional experimental design and/or within a (post)positivistic research paradigm that conceives of language as referential, there is an increasing number of researchers who point to the constitutive role of language and conceive of language as action in its own right. These approaches are based on social constructionism and Wittgensteinian psychology and closely linked to discursive psychology (e.g., Edwards and Potter, 1992; Shotter, 1993). They take a 'communicative' or 'discursive' practice approach in that they privilege the sequential microanalysis of 'talk-in-interaction,' that is, of what is being done with verbal and nonverbal communications in the way an interaction develops. They allow for a more comprehensive cultural analysis because cultural orientations are expressed not only in the content of talk but also in the way that discourse is organized, enacted, and embedded (Miller and Fung, 2012).

A classic ethnographic study on children's personal stories was conducted by Heath (1983) who compared working-class families of a European-American community and a neighboring African-American community in the Piedmont Carolinas. While children's stories in the European-American community adhered to a criterion of literal truth and focused on children's weakness or foolishness, the stories of the African-American children adhered to fictional embellishment and focused on aggrandizement of the protagonists.

The work by Miller et al. (2007) found that Taiwanese mothers typically use storytelling as a didactic medium to invoke moral and social rules in order to correct children's misbehavior while US middle-class mothers typically

downplayed their children's misdeeds in conarrated stories in order to not threaten the child's self-esteem.

In a study on middle-class family interactions in Los Angeles and Rome, Fasulo et al. (2007) examined conversational practices between parents and children with regard to everyday hygiene and household cleaning tasks that foster familial and cultural values such as autonomy, interdependence, and responsibility. Their findings revealed differences in the style and amount of parental control over cleaning tasks, and the number of options given to children in the process and sequence of tasks, which may lead to the construction or limitation of children's agency.

Sirota (2006) could show how mother–child bedtime routines in US middle-class families serve as an interactional matrix to foster autonomous self-initiative in tandem with a growing capacity for relational communion with others.

Swedish mothers have been found to follow egalitarian communication patterns with their children that are in line with a general orientation toward democracy and equality (Aronsson and Cekaite, 2011). Swedish children on their part have been found to be more active in controlling their mothers' behavior compared to Estonian and Finnish children, possibly illustrating the Swedish value of generational equality (Tulviste et al., 2004).

By comparing Italian and Euro-American family dinner interactions, Ochs et al. (1996) could show that children's attitudes toward food (as pleasure vs as reward) are socialized through discursive practices by caregivers.

A study on Italian family dinner interactions by Sterponi (2009) could show how parents scaffold children's participation in episodes of accountability through vicarious accounts that serve as face saving acts both for the child whose action was signaled as improper and for the parent who initially requested the account.

Methodological Considerations and Future Directions

Understanding child development in cultural context has become a central concern in developmental psychology in recent years. More and more scholars acknowledge that there are multiple developmental pathways and not a single universal pathway of healthy development (e.g., Keller, 2007). Understanding child development in cultural context requires attention to meaning – more precisely meaning-making processes in social interactions (Miller and Fung, 2012). Specific attention hence needs to be given to communicative practices in which children participate.

Traditionally, psychological research on communication with children has focused on the mother–child dyads whereas in the majority world children participate in multiparty interactions (De León, 2012). Most studies have been conducted in experimental or relatively controlled designs. Language in these approaches has primarily been treated as referential.

A discursive practice approach that conceives of language as constitutive, however, calls for methodological approaches that are able to capture the intertwined nature of culture and development in mundane everyday interactions and naturally occurring communication. This implies the broader social network of children such as fathers, siblings, peers, neighbors, and extended family. Research designs that combine ethnographic procedures with microanalytic conversation and discourse analysis have proven to be very fruitful in this field (Ochs and Schieffelin, 1984; Miller and Fung, 2012; Demuth and Fatigante, 2012).

A discursive practice approach that solely focuses on verbal communication risks, however, to neglect embodied language practices through which children constitute their social world in the midst of moment-to-moment interaction. It is therefore crucial to include gesture, mimic, and other nonverbal aspects of communication. There is also an increasing acknowledgment of the importance of the phenomenological immediacy of experience in discursive practice approaches that is central to embodied action (Cresswell and Teucher, 2011).

Furthermore, research needs to consider the sociocultural embeddedness of communicative practices with children, for example, ways in which a community's habitus may influence the communicative practices with children. In this line of argumentation, Ochs et al. (2005) suggest a model of Child-Directed Communication that extends caregiver–child communication to semiotic resources beyond speech.

In order to avoid the bias of generalizing from studies with white Western middle-class families more research with participants from understudied non-Western societies, as well as with working-class samples are urgently needed.

Finally, developmental research needs to take into account children's active participation in their social worlds (*see* Qualitative Methodology in Developmental Psychology). Children are not passive participants of cultural practices but actively contribute to the coconstruction process of social interaction.

Appendix A. Supplementary Data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/B978-0-08-097086-8.23209-1.

See also: Child Care and Child Development; Child Care and Development across Cultures; Child-Directed Speech: Influence on Language Development; Critical Psychology; Cultural Psychology; Culture and the Self: Implications for Psychological Theory; Developmental Research across Cultures and Nations: Challenges, Biases, and Cautions; Ecocultural Theories of Development; First Language Acquisition, Developmental Psychology of; Language Development, Theories of; Language Development: Influence of Socio-Economic Status; Parenting Attitudes and Beliefs across Cultures; Qualitative Methodology in Developmental Psychology; Self in Culture: Early Development; Self-Development in Childhood and Adolescence; Vygotsky's Sociocultural Theory.

Bibliography

Aronsson, K., Cekaite, A., 2011. Activity contracts and directives in everyday family politics. Discourse & Society 22 (2), 137–154.

Bonaiuto, M., Fasulo, A., 1997. Rhetorical intentionality attribution: its ontogenesis in ordinary conversation. British Journal of Social Psychology 36 (4), 511–536.

Blum-Kulka, S., 1990. You don't touch lettuce with fingers: parental politeness in family discourse. Journal of Pragmatics 14, 259–288.

Blum-Kulka, S., 1997. Dinner Talk: Cultural Patterns of Sociability and Socialization in Family Discourse. Lawrence Erlbaum, Mahwah, NJ.

Blum-Kulka, S., Sheffer, H., 1993. The metapragmatic discourse of American-Israeli families at dinner. In: Kasper, G., Blum-Kulka, S. (Eds.), Interlanguage Pragmatics. Oxford University Press, New York, pp. 196–223.

Branco, A.U., Valsiner, J. (Eds.), 2004. Communication and Metacommunication in Human Development. Information Age Publishing Inc, Greenwich, CT.

Briggs, J.L., 1982. Living dangerously: the contradictory foundations of value in Canadian Inuit society. In: Leacock, E., Lee, R. (Eds.), Politics and History in Band Societies. Cambridge University Press, Cambridge, pp. 109–131.

Briggs, J.L., 1998. Inuit Morality Play: The Emotional Education of a Three-Year-Old. Yale University Press, New Haven, CT.

Bruner, J., 1990. Acts of Meaning. Harvard University Press, Cambridge, MA.

Cowley, S.J., Moodley, S., Fiori-Cowley, A., 2004. Grounding signs of culture: primary intersubjectivity in social semiosis. Mind, Culture, and Activity 11 (2), 109–132.

Cresswell, J., Teucher, U., 2011. The body and language: M. M. Bakhtin on ontogenetic development. New Ideas in Psychology 29, 106–118.

De León, L., 2000. The emergent participant: interactive patterns in the socialization of Tzotzil (Mayay) infants. The Journal of Linguistic Anthropology 8 (2), 131–161.

De León, L., 2012. Language socialization and multiparty participation frameworks. In: Duranti, A., Ochs, E., Schieffelin, B. (Eds.), The Handbook of Language Socialization. Wiley-Blackwell, Chichester, pp. 81–111.

Demuth, C., 2008. Talking to infants: how culture is instantiated in early mother-infant interactions. The Case of Cameroonian Farming Nso and North German Middle-Class Families (Doctoral thesis). University of Osnabrueck. Available at: http://repositorium.uni-osnabrueck.de/handle/urn:nbn:de:gbv:700-2009030626.

Demuth, C., Fatigante, M., 2012. Comparative qualitative research in cultural psychology: challenges and potentials. In: Hummrich, M., Pfaff, N., Rademacher, S. (Eds.), "Kulturvergleichende Qualitative Forschung". [Cultural Comparative Qualitative Research], Zeitschrift für Qualitative Forschung [Journal of Qualitative Research], vol. 13 (1–2), pp. 13–37.

Duranti, A., Ochs, E., Schieffelin, B. (Eds.), 2011. The Handbook of Language Socialization. Wiley-Blackwell, Chichester.

Edwards, D., Potter, J., 1992. Discursive Psychology. Sage Publications, Inc.

Fasulo, A., Loyd, H., Padiglione, V., 2007. Children's socialization into cleaning practices: a cross-cultural perspective. Discourse & Society 18 (1), 11–33.

Forrester, M.A., 1999. Reflections and projections of the developing self. Theory & Psychology 9 (1), 29–46.

Forrester, M.A., 2001. The embedding of the self in early interaction. Infant and Child Development 10 (4), 189–202.

Gratier, M., 2003. Expressive timing and interactional synchrony between mothers and infants: cultural similarities, cultural differences, and the immigration experience. Cognitive Development 18, 533–554.

Haviland, J.B., 1998. Early pointing gestures in Zincantán. Journal of Linguistic Anthropology 8, 162–196.

Heath, S.B., 1983. Ways with Words: Language, Live and Work in Communities and Classrooms. Cambridge University Press, Cambridge, UK.

Hundeide, K., 1993. Intersubjectivity and interpretive background in children's development and interaction. European Journal of Psychology of Education 8 (4), 439–450.

Junefelt, K., Tulviste, T., 1997. Regulation and praise in American, Estonian, and Swedish mother–child interaction. Mind, Culture, and Activity: An International Journal 4, 24–33.

Kaye, K., 1982. The Mental and Social Life of Babies: How Parents Create Persons. University of Chicago Press, Chicago, IL.

Keating, E., Egbert, M., 2004. Conversation as a cultural activity. In: Duranti, A. (Ed.), A Companion to Linguistic Anthropology. Blackwell, Malden, MA, pp. 169–196.

Keller, H., 2007. Cultures of Infancy. Erlbaum, Mahwah, NJ.

Kusserow, A.S., 1999. De-homogenizing American individualism: socializing hard and soft individualism in Manhattan and Queens. Ethos 27, 210–234.

LeVine, R.A., Dixon, S., LeVine, S., Richman, A., Leiderman, P.H., Keefer, C.H., et al., 1994. Child Care and Culture: Lessons from Africa. Cambridge University Press, New York.

Lutz, C.A., 1988. Unnatural Emotions: Everyday Sentiments on a Micronesian Atoll and Their Challenge to Western Theory. University of Chicago Press, Chicago, IL.

Miller, P.J., 1986. Teasing as language socialization and verbal play in a White working-class community. In: Schieffelin, B.B., Ochs, E. (Eds.), Language socialization across cultures. Cambridge University Press, New York, pp. 199–212.

Miller, P.J., Fung, H., 2012. I. Introduction. Monographs of the Society for Research in Child Development 77 (1), 1–14.

Miller, P.J., Fung, H., Koven, M., 2007. Narrative reverberations. How to participate in narrative practices co-creates persons and cultures. In: Kitayama, S., Cohen, D. (Eds.), Handbook of Cultural Psychology. Guilford Pubn, New York.

Miller, P.J., Fung, H., Lin, S., Chian-Hui Chen, E., Boldt, B.R., 2012. How socialization happens on the ground: narrative practices as alternate socialization pathways in Taiwanese and European-American families. Monographs of the Society for Research in Child Development 77 (1).

Nelson, K., Fivush, R., 2004. The emergence of autobiographical memory: a social cultural developmental theory. Psychological Review 111 (2), 486–511.

Nsamenang, A.B., Lamb, M.E., 1994. Socialization of Nso children in the Bamenda grassfields of Northwest Cameroon. In: Greenfield, P.M., Cocking, R.R. (Eds.), Crosscultural Roots of Minority Child Development. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, pp. 133–146.

Ochs, E., 1982. Talking to children in Western Samoa. Language in Society 11 (1), 77–104.

Ochs, E., Schieffelin, B.B., 1984. Language acquisition and socialization: three developmental stories and their implications. In: Shweder, R.A., LeVine, R.A. (Eds.), Culture Theory: Essays on Mind, Self, and Emotion. Cambridge University Press, New York, pp. 276–320.

Ochs, E., Pontecorvo, C., Fasulo, A., 1996. Socializing taste. Ethnos 60 (3), 7–46.

Ochs, E., Solomon, O., Sterponi, L., 2005. Limitations and transformations of habitus in child-directed communication. Discourse Studies 7 (4–5), 547–583.

Quinn, N., 2005. Universals of child rearing. Anthropological Theory 5 (4), 477–516.

Rabain-Jamin, J., Sabeau-Jouannet, E., 1997. Maternal speech to 4-month-old infants in two cultures: Wolof and French. International Journal of Behavioral Development 20 (3), 425–451.

Schieffelin, B.B., 1986. Teasing and shaming in Kaluli children's interactions. In: Schieffelin, B.B., Ochs, E. (Eds.), Language socialization across cultures. Cambridge University Press, New York, pp. 165–181.

Schieffelin, B.B., Ochs, E., 1986. Language Socialization across Cultures. Cambridge University Press, Cambridge, MA.

Schieffelin, B.B., Ochs, E., 1998. A cultural perspective on the transition from prelinguistic to linguistic communication. In: Woodhead, M., Faulkner, D., Littleton, K. (Eds.), Cultural Worlds of Early Childhood. Routledge, Oxon, pp. 48–63.

Schröder, L., Keller, H., Kärtner, J., Kleis, A., Abels, M., Yovsi, R.D., Chaudhary, N., Jensen, H., Papaligoura, Z., 2013. Early reminiscing in cultural context: cultural models, maternal reminiscing styles, and children's memories. Journal of Cognition and Development 14 (1), 10–34.

Sirota, K.G., 2006. Habits of the heart: children's bedtime routines as relational work. Text & Talk 26 (4/5), 493–514.

Shotter, J., 1993. Conversational Realities: Constructing Life through Language. Sage Publications, London.

Stern, D.N., 1977. The First Relationship: Infant and Mother. Harvard University Press, Cambridge, MA.

Sterponi, L., 2009. Accountability in family discourse: socialization into norms and standards and negotiation of responsibility in Italian dinner conversations. Childhood 16, 441–459.

Takada, A., 2011. Preverbal infant–caregiver interaction. In: Duranti, A., Ochs, E., Schieffelin, B. (Eds.), The Handbook of Language Socialization. Wiley-Blackwell, Chichester, pp. 56–80.

Tõugu, P., Tulviste, T., Schröder, L., Keller, H., De Geer, B., 2011. Socialization of past event talk: cultural differences in maternal elaborative reminiscing. Cognitive Development 26 (2), 142–154.

Tõugu, P., Tulviste, T., Schröder, L., Keller, H., De Geer, B., 2012. Content of maternal open questions and statements in reminiscing with their 4-year-olds: links with independence and interdependence orientation in European contexts. Memory 20 (5), 499–510.

Trevarthen, C., 1998. The concept and foundation of infant intersubjectivity. In: Bråten, S. (Ed.), Intersubjective Communication in Early Ontogeny. Cambridge University Press, Cambridge, pp. 15–46.

Tulviste, 2004a. Sociocultural variation in mother's control over children's behavior. Ethos 32 (1), 1–17.

Tulviste, 2004b. Mothers' conversational styles across cultures: the cases of Estonia, Finland, Sweden, and the U.S. In: Setiadi, B.N., Supratiknya, A., Lonner, W.J., Poortinga, Y.H. (Eds.), Ongoing Themes in Psychology and Culture (Online Ed.). International Association for Cross-Cultural Psychology, Melbourne, FL. Retrieved from: http://www.iaccp.org.

Tulviste, T., Mizera, L., De Geer, B., 2004. Expressing communicative intents in Estonian, Finnish, and Swedish mother-adolescent interactions. Journal of Child Language 4, 801–819.

Tulviste, T., Mizera, L., De Geer, B., Tryggvason, M.-T., 2003. A silent Finn, a silent Finno-Ugric, or a silent Nordic? A comparative study of Estonian, Finnish, and Swedish mother-adolescent interactions. Applied Psycholinguistics 24, 249–265.

Vygotsky, L.S., 1978. Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA.

Wang, Q., 2006. Relations of maternal style and child self-concept to autobiographical memories in Chinese, Chinese immigrant, and European American 3-year-olds. Child Development 77 (6), 1794–1809.

Wang, Q., 2007. Remember when you got the big, big bulldozer? Mother-child reminiscing over time and across cultures. Social Cognition 25 (4), 455–471.

Wang, Q., 2008. Emotion knowledge and autobiographical memory across the preschool years: a cross-cultural longitudinal investigation. Cognition 108, 117–135.

Wang, Q., Leichtman, M.D., 2000. Same beginnings, different stories: a comparison of American and Chinese children's narratives. Child Development 71 (5), 1329–1346.

Wang, Q., Leichtman, M.D., Davies, K.I., 2000. Sharing memories and telling stories: American and Chinese mothers and their 3-year-olds. Memory 8 (3), 159–177.

Relevant Websites

http://childes.psy.cmu.edu/.
http://www.celf.ucla.edu/.
http://dx.doi.org/doi:10.1016/B978-0-08-097086-8.23209-1.
http://www.human.cornell.edu/che/HD/socialcognition/index.cfm.

Motherhood

Karen J Swift, York University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article traces ideas and ideologies of motherhood from prehistoric times through the present. Motherhood as a dominant discourse for personal, social, and state purposes is explored. The actual experience of motherhood by women historically and in different parts of the world is discussed. The policing and control of motherhood in various ways, by males, families, religions, and the state, are examined. Some ideas of the future of motherhood by various authors are presented, and the author shows how mothers themselves increasingly shape motherhood.

Introduction

The idea of motherhood has occupied iconic status historically and in cultures around the world. This article addresses myths, discourses, experiences, and realities of motherhood, primarily, but not only, in the West. The author concludes with ideas about the possible future of motherhood.

Images and Myths of Motherhood

According to Thurer (1994), the oldest of all gods was the Great Mother, who was believed to be all powerful. The role of mothers from the Paleolithic era and for many centuries forward was highly valued and celebrated. Only females could reproduce the human species and only mothers could be identified with certainty as the biological parent of a child. The continuation of humanity and family line depended upon mothers, who were strongly associated with the earth as the source of all life.

Adrienne Rich (1976), too, cites motherhood in prehistory as a site of power, although becoming constrained even then. Early on, women were associated with nature, as denoted by the phrase 'mother earth.' Nature was seen as woman's domain, the source of her powers of reproduction, and as the symbolic representation of her fecundity. Images of early mothers were often in the shape of vessels. The tree in leaf was a female representation of the transformation of life and of the dependency of new life upon the life-giving source, the mother. Even ancient tombs were designed to represent the mother, with labyrinthine forms suggesting female internal anatomy and representing the rebirth of the human spirit. These associations project motherhood as a natural fact, as instinctual, as inevitable for women to aspire to and accomplish.

In ancient Egypt, the goddess Isis was considered the paragon of motherhood, a patron goddess of childbirth and a figure of both gentle motherhood and confident queenhood. In Africa, a common symbol of motherhood is the creation goddess associated with the moon and water. The creation goddess nurtures and has power over life. In Hinduism, Devi Ma is the Mother Goddess; she is not only a creator but also a warrior and destroyer. Buddhism has its roots in the Great Mother Tradition. In Christian tradition, Eve, the first mother, is depicted as both sinful in giving Adam the poisoned apple and an enduring symbol of creation, the 'Mother of all the living.' In the Jewish faith, Adam's partner in the Garden of Eden is Chavah, who embodies the essence of life and ability to create new life, nurture, and enhance new life. Unlike Eve, Chavah is a thoroughly positive figure. Mary is the mother of Jesus, conceived through a (sexless) miracle and worshipped by many Christians today. She is portrayed as all giving, all loving, and all sacrificing and remains as an iconic image of the ideal mother. Hence, in ancient times, men were secondary to women because of women's powers of reproduction.

However, for most if not all of the major religions, these early female and motherhood symbols eventually gave way to male leaders and gods in power and authority. According to Campbell, the prodigious power of motherhood in the earliest stages of human history eventually encouraged in males the desire to "break, control and employ [that power] to its own ends" (1972: p. 372), a shift he identifies as beginning about 30 000 years BC. Thus, depiction of the divine began to change from female to male figures and symbols. Most religions have special roles and rules for women, including mothers, that place them as institutionally less powerful than men. Symbolic figures glorifying motherhood, although still important to these various religions and cultures, have long since ceased to reflect the immense power that women as creators of life once held.

Engels (1942), in his famous study of the origins of the family, concurs that prior to recorded history women occupied positions of status and power, never since reproduced. He has traced the development of the monogamous family, noting the earlier importance of mother-right based on the reality that until monogamy was established and enforced only mothers could be identified with certainty as a biological parent. Father-right becomes the norm with monogamy in marriage, monogamy that is for women, along with inheritance rights for fathers through laws establishing that the husband is the father. The association of motherhood with earth, oceans, transformation, growth, and nature became, as patriarchy developed, less associated with power and increasingly with symbolism of woman as nurturers of the young produced through childbirth, and as molders of the unformed children who in turn were increasingly viewed as potential heirs, nurtured by the mother but owned and controlled by the father.

Motherhood as Ideology

Understandings of motherhood arrive in contemporary times as a set of discourses found in everyday conversation, news, law, and policy. Two closely related strands of discourse are easily identified. The dominant discourse of motherhood has remained stable through several centuries; motherhood is viewed as the main purpose in life of every woman and the most illustrious of states. Through much of recorded history, and guided by patriarchal forms of thought, law, and discourse, motherhood has been presented as women's ultimate purpose and main mission in life, to be a holy mission and the highest form of service to humanity.

A well-known summary of the ideology of motherhood is the "belief that all women need to be mothers, all mothers need their children and all children need their mothers" (Oakley, 1986: p. 67). Many feminists have examined the 'cult of motherhood' as an all-encompassing mixture of a biological imperative to reproduce and a social imperative for mothers to provide stabilizing and moral influences for children and society at large. As the mother came to be cast as the all-important figure for the child, belief held that she should devote herself body, mind, heart, and soul to the enterprise of caring for that child. The 'bliss' of motherhood was promoted, not least by the Christian church, as the highest calling for women. An extension of that calling is that the mother–child bond is sacrosanct and must extend into the life of the child in the form of primary care by the mother throughout the child's life. Western feminists point out that the glorification of motherhood actually serves as a method of constraint; 'proper' mothering dictated by ideology demands that mothers stay at home and spend most of their time taking care of children and creating a comfortable, stable home atmosphere.

A related discourse of motherhood follows from the state's interest in controlling and using female fertility for nation building and economic purposes. Yeo (2005) states that motherhood in this sense was 'created' around the middle of the eighteenth century. The developing discourse of motherhood in Europe after that period tasked mothers not only with responsibility for continuation of the family, the production of (male) heirs, and the improvement of class position for the family but also for purity of the race (white). Thus, motherhood came to be not only a private matter but also a matter of importance to the state, which it remains. Consequently, when population growth is needed, policies are developed to encourage women to become mothers. In times of overpopulation or when poor or racialized mothers are deemed to be overproducing children, policies supporting child care and access to health care are likely to shrink. Malthus, writing two centuries ago, instructed that only families who could support children should reproduce. Otherwise, he thought, their children 'must starve' (Malthus, 1803: p. 226). Yeo (2005) claims one of the 'master narratives' governing European motherhood in earlier centuries was that of nation building, especially in the colonies. The creation of new nations required a growing, healthy population, with women's roles focused on producing and rearing soldiers and laborers. Once nation-building efforts became established, mothers were called upon to contribute to the development of a large and prosperous white middle class needed to perpetuate and grow capitalism. For this purpose, white mothers were needed to learn, teach, and demonstrate the moral authority the middle class required to dominate those below in the social, economic, and racial hierarchies. These efforts have produced an intersection of class, gender, and race that carries over into the current century in the West.

The development and growth of the middle class coincided with the expansion of industrialization in both Europe and North America. Beginning in the nineteenth century, increasing industrialization brought large numbers of men into the new industrial labor force, creating a gendered separation of men and women in their daily and working lives. Women were increasingly denied entry into the public world of paid work and discourse (Fegan, 1994). According to many white feminists, it is in this period that the 'cult of motherhood' was intensified in the West. The new jobs were for men, hence the 'living wage' argument: if men could earn enough in the paid labor force to support wife and children, women could then leave paid jobs to men and do their biological and social reproductive work in the home.

A contemporary extension of this 'state-based' discourse of motherhood is generally attributed to Adrienne Rich (1976), who famously identified two meanings of motherhood. One is the relationship of a woman to reproduction of the species and children. The other is the 'institution' of motherhood, which Rich names as male controlled and a product of patriarchy. This ideology of motherhood is called upon to inscribe upon women motherhood not only as instinct and selflessness, but also as subordination to others, namely, men, in the interests of carrying out their motherly duties. The institution of motherhood, anywhere in the world sends messages about how to feel, think, and act as a mother.

For women of color, however, the ideology of motherhood portrayed by white feminists does not fit so well. Collins (1994) argues that patriarchy and male domination have not historically been the primary forces controlling the shape and form of motherhood for women of color. Nor do the dichotomies presented by white feminists – public/private, male/female – apply in the same ideological way. For many women of color, experiences of motherhood involve survival issues, for their children, for themselves, and for the group. The isolation of the mother in the private home is a distant dream for these mothers. Instead, they, their children, and their male partners as well, may all have carried out paid work together, in fields, as migrant workers, doing piecework, as guest workers in foreign countries. Some, for instance, Aboriginal mothers, have not been allowed to perform their mothering work for long periods, as their children have been forcibly removed to residential schools for lengthy periods. The nuclear family, with the clear divisions of labor noted by white feminists, is often not their family form. This can be the case for white families as well. In these, the extended family may prevail, meaning the work of mothering goes on into succeeding generations. Many of these mothers perform their mothering work for other women's children, as they leave their own behind in other neighborhoods or other countries to care for children of better-off mothers. Collins (1994) argues that for women of color, the state has harnessed motherhood for its own purposes of producing and controlling cheap labor in support of elite

populations. While Collins' argument holds for many women of color, these mothers also experience patriarchy as a driving force in their lives. In their roles as paid laborers, colonial relations may rule their mothering, but when they return home, it is male rule they must endure.

For childless women, the discourse of motherhood as life's purpose for every woman is also ill-fitting, posing problems of identity and social value for childless women. If motherhood is the main justification of life for women, and that purpose remains unfulfilled, who are they and what social status and value can they have? The shame of barren and single women applies in most cultures and throughout history. Images of the maiden aunt, the old maid, or the spinster come to mind. One socially legitimate and often revered role for single women has been the religious life (Yeo, 1999: p. 208). In the late nineteenth and early twentieth centuries, the idea of the 'social mother' was invented (Yeo, 1999: p. 209), largely by social workers, as another way of providing a role for childless women. Social mothers may not have had their own children but they performed a motherly role in the community. Social work itself as well as teaching, nursing, and related 'women's professions' to some extent grew out of this idea. Of course, this direction was a double-edged sword for women entering the new professions; it allowed them entry into the public world of paid employment, but it restricted them to interaction mainly with other women and children and to very low paid work (Yeo, 1999).

Thus, ideologies of motherhood serve state, social, patriarchal, and personal purposes. While there is a central and glorified ideology of motherhood that pressures and exhorts women to perform their lives and their mothering role in certain ways, the experiences of motherhood for different groups of women invariably fail to conform to the ideal.

Policing Motherhood

In spite of the ideological glorification of motherhood in discourse, symbols, and myths, many authors have indicated that the actual experience of motherhood over past centuries has been less than glorious. For centuries, motherhood was a life-threatening condition, as many women died in childbirth. Throughout the twentieth century, European and North American women were denied effective forms of birth control and safe abortions that could save them not only from the health hazards of childbirth, but also from the hazards of raising too many children on small incomes. Margaret Sanger titled her 1928 book *Motherhood in Bondage*, including in it the pleas of thousands of overburdened, often unhealthy mothers for access to the new forms of birth control.

The threat of unsafe childbirth remains for many women in poor countries. Maternal death rates remain high, for instance, in Asia and Africa, where about 95% of the world's maternal deaths occur; 1 in 16 women die in childbirth in Sub-Saharan Africa as opposed to 1 in 2800 in 'developed' regions. For every woman who dies, two or three dozen others incur disabling injuries through childbirth. And should they survive to raise their children, most mothers in these regions will know poverty and malnutrition for both herself and her children, and will likely experience the deaths of one or more of her children before the age of 5 years (Rosenfield et al., 2007).

As many women realize, motherhood generally involves hard work and considerable sacrifice, and the central guiding ideology of motherhood produces social norms and expectations that can be utilized to judge mothers' performances regardless of the socially constructed difficulties they face. Should they balk, or depart from the norms set by the state and society, a variety of laws and policies are available to police motherhood for state and social interests. In China, the government imposes a limit on number of children for certain families, mainly in urban areas. After that, the state can impose fines or even involuntary abortion and/or sterilization. In other places, organized religion imposes strict rules for motherhood, quite often policed by the men of the family, religious institutions, and/or the state.

In the West, motherhood has for some decades been policed by a 'science' of correct ideas and behaviors, along with state and religious controls. Scientific motherhood is several centuries old, stimulated by Enlightenment ideas of the existence of scientific, albeit frequently revised, 'facts' and 'truths' alongside the development of professions with relevant expertise. Physicians authored some of the earliest advice, thus beginning the medicalization of motherhood. Early examples of advice to mothers are William Cadogan's 1749 *Essay upon Nursing* and William Buchan's (1784) *Domestic Medicine*, in which he instructed that "nothing can be more preposterous than a mother who thinks it below her to take care of her own child" (p. 28); and J.C. Ryle's scripture-inspired advice that children should be trained "in the way they should go, and not in the way that they would" (1888).

Advice literature presumes a conception of childhood, new in the eighteenth century, as a stage distinct from adulthood, and a new idea of motherhood, requiring women to follow the proffered advice. As advice has changed over the centuries, so have the underlying notions of motherhood and mothering required to follow it.

Another source of advice comes via the education system. Frederich Froebel, for example, encouraged a system of learning through play, to be taught to parents (mothers) and carried out in the home. His work began in the mid-nineteenth century and continues to influence education and parents. Some years after Froebel, Jean Piaget developed his theory of children's cognitive learning patterns and development, a theory that remains infused in ideas of proper parenting as a path to optimal childhood development. Freud's theory of the Oedipus complex, involving desire for the mother by both male and female children, later contributed to ideas of failed motherhood as the central feature in the developmental problems of children. Many psychologists considered the mother as outright dangerous. Thus, motherhood was not celebrated but rather denigrated as the source of virtually all childhood problems. John Bowlby followed Freud's psychology with now famous concepts about attachment and bonding. Bowlby's work suggested that only biological mothers could properly bond with children, thereby reinforcing notions of the ever-present doting mother as necessary for healthy childhood development. This work has been much criticized by feminists, but remains influential in mainstream ideas of good mothering.

Building on Freudian psychology in the postwar Western world were the burgeoning sciences of psychology and psychiatry. In this period, mother-blaming was at its height, and a deepening binary emerged: mothers as saints, as promoted through the traditional ideology of motherhood, and mothers as villains, as discovered through science. In the United States, the postwar literature emerged tying motherhood to a crisis of manhood and masculinity. Mothers should not be too warm, too affectionate, or too close to their sons, for fear of producing an emasculated adult. Motherhood was even linked to national security, to the growing 'menace' of communism, and fascism – "Naziism was … a mom surrogate with a swastika for a heart" said the psychiatrist Edward Strecker (cited in Aidenbaum, 2006). Meanwhile, experts in the social and psychiatric sciences blamed mothers for a raft of mental health issues – neuroses, depression, repression, insecurities of all kinds, delinquency, and homosexuality. One 1985 study (Caplan and Hall-McCorquodale) reviewing dozens of published clinical studies found that regardless of the problem, the large majority of research studies examined the mother–child relationship as a problem source, and about two-thirds of the studies blamed the mother's 'pathology' for problems that included everything from 'mental handicaps' to colitis. Second-wave feminist Betty Friedan commented that the mother could be blamed for almost anything. Freudian psychology has been critiqued and much of it dispensed with, but the discourse of dangerous mothering has not disappeared but has become lodged in laws controlling female bodies and welfare state provisions.

Laws and rules governing the female body and its reproductive capacities were institutionally established as early as 1803 in Great Britain with the passage of The Ellenborough Act, which provided the death penalty for abortions occurring after 'quickening.' Laws criminalizing abortion followed in Canada, focusing on criminalizing abortionists. Shortly after Confederation in Canada, abortion laws were extended to criminalize the woman herself. Access to birth control in these countries was tightly controlled by law into the twentieth century, resulting in enforced motherhood. While these laws have been loosened in many parts of the world, control of the reproductive capacities of women continues, for example, through American policy denying foreign aid to programs that include abortion (Blanchfield, 2013). Similarly, Canada does not fund abortion through foreign aid, and allocates only a very small portion of its aid budget to family planning (Clark, 2010). Law and policy allowing legal abortion are regularly challenged by lobby groups and politicians in the English-speaking world and elsewhere.

In contrast, policing motherhood occurs across the world, and policies enacted to prohibit motherhood for some women. The United States had a forced sterilization program, generally applying to women and immigrants, in the early 1900s, in an attempt to purify the gene pool and produce a society without social problems (Webster University, n.d.). The United States Supreme Court of the era upheld state laws allowing involuntary sterilization, including for intellectually disabled people and some immigrants. The not so hidden message was that motherhood should not be experienced by all women but by those socially selected for specific reasons. In Canada, Sweden, Japan, Germany, and elsewhere in the first half of the twentieth century, forced sterilization occurred for eugenic reasons (Webster University, n.d.). Sometimes, population control is the purpose, as in contemporary China, or gender control through selective abortion and/or infanticide, that occur in communities that do not value girls. Selective abortion can take place in countries that have relatively open abortion policies.

The modern welfare state has been built on the foundation of Elizabethan Poor Law, which distinguished the deserving from the undeserving poor. The original laws viewed vagrants as the undeserving, but low-income, single, and young mothers have long since served as the centerpiece of the undeserving population (Abramowitz, 1989). Thus, motherhood is classed as well as raced. The middle class, the married, and the white are generally viewed as more deserving, and are socially placed so that help from the state is not so often needed; when it is, they fare no better than the single, the poor, and those of color.

Child protection services, originating in the West in the last decades of the nineteenth century, continued the deserving/undeserving distinction, and generally created with urban, poor, and often racialized mothers (and their children) as their target group – 'bad mothers' (Swift, 1995). In the beginning, these examples of unsuccessful motherhood were most frequently charged with neglect of their children. Neglect could mean leaving children unsupervised or failing to provide adequate food, clothing, shelter, education, health care, and discipline. Neglect, although legally aimed at parents, virtually always means mothers, usually poor mothers, who often have no means of providing sufficiently for children. From the beginning, the most stringent punishment for failed motherhood has been removal of the children from the mother's care. Recently, the phrase 'failure to protect' has become embedded in child protection law usually to mean failure to protect children from witnessing violence perpetrated against the mother by a male, often the children's father. Although not articulated, an ideal of motherhood haunts the law, and failure implies transgressing a standard. Contemporary child welfare practice has further encoded such standards in assessment tools, notably risk assessments, which purport to measure numerically the risk that a child will experience bad mothering at some future point (Swift and Callahan, 2009).

Clearly, ideas about ideal motherhood remain firmly entrenched in child protection law. Davies and Krane (1997) argue that mother-blame, still based on an unspoken idealization of the nurturing capacities of all mothers, permeates child protection work. They also point out that the actual work of mothering is invisible in the process of evaluating the accomplishments and failures of mothers who come into contact with child protection services. Mothers caught up in the system are sent to parenting classes to learn the skills and attributes of the good mother; anger management courses to learn self-control; drug rehabilitation to become sober, all to conform more closely to ideal motherhood.

Institutional child welfare plays a central role in reinforcing the dominant ideology of motherhood by regularly producing exemplars of 'bad mothering.' Child welfare services also legitimize scientific mothering, through their use of expert prescriptions to mothers. These prescriptions will not be

contained in the West. The United Nations Convention on the Rights of the Child, developed in a spirit of improving the life circumstances of the world's children, asks developing countries to create institutional structures for child protection (UNICEF, 2013). These countries, under pressure to conform to the Convention, reach out for Western models of child protection. As they do, the invisible, unspoken ideal of Western motherhood may become more entrenched around the world (Swift and Kikulwe, 2012).

The Future of Motherhood

In Marge Piercy's (1976) futuristic novel *Woman at the Edge of Time*, motherhood as a gendered role and activity no longer exists. Children are conceived in laboratories through random selection of genetic attributes, raised to viability in artificial wombs, and cared for after birth by groups of three comothers, some of whom may be male. Every child is a wanted child, and all infants and toddlers are housed together in large nurseries.

This vision of the future is not far-fetched, considering scientific advances in fertility. Work on developing an artificial uterus is underway. Scientists are constructing human sperm, eggs, and wombs in the laboratory, and cloning animals was successfully achieved years ago. Ethical considerations currently prevent science from taking these advances to an end result of producing an artificially bred human being. Margaret Atwood (1985) overcomes this problem in her fictional story of women in *A Handmaid's Tale* where a country founded under racist, theocratic, and military rule keeps a class of women as concubines (handmaids) for reproductive purposes in an era of declining births. Handmaids are paired with upper-class males for reproductive purposes, and their offspring given to mothers of the proper class to raise.

One vision explores what it might be like to do away with biological motherhood; the other examines a world that utterly controls motherhood for state and class purposes. Although extreme, each portrayal emphasizes potential directions for the future of motherhood and mothering, which may or may not appeal.

In the contemporary West, motherhood is a personal choice, owing to effective, accessible forms of birth control and safe abortion and feminist challenges to the 'institution' of motherhood. Western women are choosing later motherhood and fewer children than in the past, and combine motherhood with paid labor. Most mothers in these societies have achieved a level of personal liberation and choice undreamt of a century ago. Current discourse speaks of parenting, which often means mothering, but increasingly involves fathers. Various state-supported programs and policies such as income support, maternity/paternity leave, and subsidized child care centers, while far from perfect, facilitate increased personal choice and assistance with child rearing. Motherhood is not exactly as it used to be, and many are thankful for that. Nonetheless, in the West, mothers continue to struggle for control of their own bodies through ongoing political threats to legal abortion; and programs such as child care that support tasks ordinarily undertaken by mothers but that are now shrinking under neoliberal policies attacking the welfare state.

In the Global South, things are different. With high rates of maternal death in childbirth, the AIDS epidemic, poverty for the majority, armed conflict, and state instability, motherhood remains a risky business. Coupled with these problems, a recent report (United Nations Population Fund, 2012) reveals that 222 million girls and women in developing countries are unable to access appropriate and timely 'family planning' assistance, including contraceptives, which the United Nations now considers a basic human right. Few women in these regions can control their own fertility, and the motherhood that follows is often not a personal choice, thereby reconstructing a world of deserving and undeserving mothers.

Mothers around the world simultaneously seek to extend the conditions and roles of motherhood in courageous and world-changing ways (O'Reilly, 2011). In South America, motherhood has been successfully employed as a political strategy in human rights struggles. The Mothers of the Plaza de Mayo demonstrated for many months in the 1970s, challenging the Argentinian government, along with that of the United States, to account for their 'disappeared' children, all probably dead. This struggle began with individual mothers seeking information about their children's fates and developed into a powerful ideological weapon against the state, which was reluctant to use force against mothers. Chilean mothers used a similar strategy for political and personal ends against the Pinochet dictatorship in the same decade, using their role as nurturing mothers as a powerful challenge to the government's savagery. Both groups reconstructed the notion of motherhood in the process, from one of an apolitical home-bound servant to one of active agents in the political sphere.

These examples of advocacy are models for many other groups of mothers wanting to challenge violent states, including the Mourning Mothers of Iran, demanding accountability for the deaths and disappearances of their children; the Mothers of Acari, who opposed political violence in Brazil; and Russian mothers fighting to expose and end human rights abuses in the military.

Women in Ireland, many of them mothers, have marched to end the violent 'Troubles' between North and South. Mothers from Israel and Palestine have reached across borders to seek peaceful solutions to the ongoing conflict in the Middle East. Mothers in countries around the world work on environmental issues, some winning prizes for their efforts. Other mothers-led groups promoted maternal values of nurturing and cooperation to promote a less violent world. Examples are Mothers Against Human Trafficking; PeaceXPeace, a group promoting sustainable world peace; Mothers Against Drunk Driving; and many groups opposing gun violence. Mothers in Africa and other low-income regions are rapidly moving into the world of breadwinning, developing small businesses, with the aid of microloans, to support their families and communities (Kristoff and WuDunn, 2009; Microcredit Summit Campaign, n.d.). Efforts to contain motherhood, order its boundaries and activities, and confine mothers to traditionally prescribed activities are weakening. Motherhood is increasingly being shaped by mothers themselves, in the interests not only of their own children but also those of the public and political worlds.

See also: Child Protection; Gendered Care and Social Work Practice; Indigenous Child Welfare Practice: The Canadian Case; Social Work Theory.

Bibliography

Abramowitz, M., April 1, 1989. Women and Children at Risk and in Poverty. Prepared for presentation to the Scholar and Feminist Conference: Women and Public Policy Making the Difference. Barnard College, New York.

Aidenbaum, A.M., 2006. Mother-blaming and the rise of the expert. Michigan Journal of History Fall 2006, 1–34. http://www.umich.edu/~historyj/ (accessed 14.11.12.).

Atwood, M., 1985. The Handmaid's Tale. McClelland and Stewart, Toronto.

Blanchfield, L., 2013. Abortion and Family Planning-related Provisions in U.S. Foreign Assistance Law and Policy. Congressional Research Service. http://www.fas.org/sgp/crs/row/R41360.pdf (accessed 09.08.13.).

Buchan, W., 1784. Domestic Medicine: Or, a Treatise on the Prevention and Cure of Diseases, eleventh ed. W. Strahan, T. Cadel, and J. Balfour, London.

Campbell, J., 1972. The Masks of God: Primitive Mythology, fourth ed. Viking Press, New York (Original work published 1956).

Caplan, P., Hall-McCorquodale, I., 1985. Mother-blaming in major clinical journals. American Journal of Othopsychiatry 53 (3), 345–353.

Clark, C., April 26, 2010. Ottawa refuses to fund abortion in G8 plan. The Globe and Mail. http://www.theglobeandmail.com/ (accessed 09.08.13.).

Collins, P.H., 1994. Shifting the center: race, class, and feminist theorizing about motherhood. In: Glenn, E.N., Chang, G., Forcey, L.R. (Eds.), Mothering: Ideology, Experience, and Agency. Routledge, New York, pp. 45–65.

Davies, L., Krane, J., 1997. Shaking the legacy of mother blaming: no easy task for child welfare. Journal of Progressive Human Services 7 (2), 3–22.

Engels, F., 1942. The Origin of the Family, Private Property and the State, thirteenth ed. International Publishers, New York.

Fegan, E.V., 1994. Abortion, Law and the Ideology of Motherhood: New Perspectives on Old Problems (unpublished Masters thesis). University of British Columbia, Vancouver. https://circle.ubc.ca/ (accessed 14.11.12.).

Kristoff, N.D., WuDunn, S., 2009. The Women's Crusade. The New York Times. http://www.nytimes.com/ (accessed 08.09.13.).

Malthus, T.R., 1803. An Essay on the Principle of Population, second ed. For J. Johnson by T. Bensley, London.

Microcredit Summit Campaign, n.d. About the campaign. http://www.microcreditsummit.org/ (accessed 09.08.13.).

Oakley, A., 1986. Telling the Truth about Jerusalem: A Collection of Essays and Poems. Basil Blackwell, Oxford.

O'Reilly, A. (Ed.), 2011. The 21st Century Motherhood Movement: Mothers Speak Out on Why We Need to Change the World and How to Do It. Demeter Press, Toronto.

Piercy, M., 1976. Woman on the Edge of Time. Random House Inc, New York.

Rich, A., 1976. Of Woman Born: Motherhood as Experience and Institution. W. W. Norton & Company, New York.

Rosenfield, A., Min, C.J., Freedman, J.D., 2007. Making motherhood safe in developing countries. New England Journal of Medicine 356, 1395–1397.

Ryle, J.C., 1888. The Duties of Parents. Old Paths Gospel Press, Choteau, MT.

Sanger, M., 1928. Motherhood in Bondage. Brentano's, London.

Swift, K.J., 1995. Manufacturing 'Bad Mothers': A Critical Perspective on Child Neglect, second ed. University of Toronto Press, Toronto.

Swift, K., Kikulwe, D., 2012. Child Protection: Transnational Comparisons. Prepared for presentation to the International Workshop: Reimagining Social Policy toward Transnational Social Support. Toronto. October 11–12, 2012.

Swift, K.J., Callahan, M., 2009. At Risk: Social Justice in Child Welfare and Other Human Services. University of Toronto Press, Toronto.

Thurer, S.L., 1994. The Myths of Motherhood: How Culture Reinvents the Good Mother. Houghton Mifflin Company, New York.

UNICEF, 2013. Convention on the Rights of the Child. http://www.unicef.org/crc/ (accessed 09.08.13.).

United Nations Population Fund, 2012. By Choice, Not by Chance: State of World Population 2012. http://www.unfpa.org/public/home/publications/pid/12511 (accessed 14.11.12.).

Webster University, n.d. Forced Sterilization. http://www2.webster.edu/~woolflm/forcedsterilization.html (accessed 14.11.12.).

Yeo, E.J., 1999. The creation of 'motherhood' and women's responses in Britain and France, 1750–1914. Women's History Review 14 (2), 201–218.

Yeo, E.J., 2005. Constructing and contesting motherhood, 1750–1950. Hecate 31 (2), 4–20.

Motivation and Actions, Psychology of

Peter M Gollwitzer and Gabriele Oettingen, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Traditionally, research on motivation and action has focused on the role of motives and needs, but also on expectations, attributions, and control beliefs. Nowadays, when it comes to answering the question of how motivation relates to action, research includes the concept of goals. But pursuing goals with the right content and structure in the proper situational context does not yet guarantee effective goal attainment. People often have to engage in self-regulatory thought to further their goal pursuits. In this article, we review the history and current research associated with these three broad perspectives on motivation and action.

Traditionally, research on motivation and action has focused on the role of motives and needs, but also on expectations, attributions, and control beliefs. McClelland (1985) distinguished three basic groups of motives: the achievement motive, the power motive, and the affiliation motive. As food is the reward or incentive for hunger, so is improving one's performance on a given task the incentive for the achievement motive. The incentive of the power motive is having impact, control, or influence over another person, a group, or the world at large. Finally, the incentives for the affiliation motive extend to sexual pleasures (sexual motive); being together with people (need for affiliation); and experiencing harmony, concern, and commitment (intimacy motive). All of these motives may entail a hope for success and a fear of failure component. Trying to meet a standard of excellence may not be triggered solely by hope for success, but also by fear of failure, and spending one's spare time affiliating with others may not be determined solely by the anticipated positive feelings of togetherness, but also by fear of rejection.

In principle, all humans are assumed to share these various motives, although with different strengths. Motive strength can be assessed by exploring both the array of situations a person interprets in terms of a given motive (e.g., a person high in need for power interprets all kinds of situations as power related) and the intensity of the anticipated affect associated with having acquired respective incentives. Commonly this is assessed with the Thematic Apperception Test, which contains pictures of scenes loosely related to the motives measured.

Recent research has linked the activation of different motives to different hormonal responses that in turn facilitate motive-specific behaviors. Scoring high on a certain motive implies a recurrent concern for acquiring the respective incentives. For instance, people high on the affiliation motive perform affiliative acts frequently and energetically, readily perceive affiliative cues in the environment, and quickly detect affiliative networks. Also, predictions of the professional success of managers are strikingly accurate, particularly if one considers the motive dispositions in achievement (high), power (high), and affiliation (low) in concert.

However, attempts to predict behaviors from motives commonly fail when engagement in these behaviors is based on conscious reflection. When it comes to choosing between courses of action or tasks of different difficulty, or to persisting on a given task versus leaving the field, people deliberate on the feasibility and desirability of the alternative courses of action.

One of the first attempts to integrate cognitive aspects of motivation was made by Atkinson (1957) in his risk-taking model. He proposed that the subjective probability of success and the task's perceived incentive value conjointly affect task choice, both variables being influenced by the perceived difficulty of the task. Whereas easy tasks lead to a high subjective probability of success (direct function), they also possess low incentive value (inverse function) because the anticipated affect associated with success (pride) is the lowest for easy tasks. The reverse is assumed for difficult tasks. Atkinson suggested that multiplying probability of success and incentive value will give a good estimate of whether a person will choose to work on a task, especially when the obtained score is weighted by the approach (hope for success) and avoidance (fear of failure) components of the person's achievement motive.

Elaborations of the Atkinson model (e.g., Heckhausen, 1991) added further expectation-related concepts (e.g., the expectation that successful task performance will lead to the anticipated incentives) and differentiated various incentives (e.g., extrinsic side effects, such as when an achievement task has affiliative benefits). Atkinson's model has also been elaborated by attribution theorists (Weiner, 1992) who attempted to understand changes in expectations and incentive value in terms of the attributions made for past performances. Moreover, Weiner discovered that the approach component of the achievement motive (hope for success) is associated with attributing failure to luck or lack of effort and success to ability, whereas the avoidance component is linked to attributing failure to lack of ability and success to luck.

Recognition of the motivational importance of expectations and attributions was the starting point of the cognitive revolution in the psychology of motivation that also introduced the concept of control beliefs. Self-efficacious individuals hold the firm belief that they possess the potential to execute the kinds of behaviors that a given task demands (Bandura, 1997). People acquire this belief by reflecting on their own relevant past behaviors, observing the behaviors of similar others, being evaluated by significant others (e.g., teachers), and observing their own physiological reactions when challenged by a given

task. High self-efficacy beliefs are associated with choosing aspiring goals, exerting strong efforts to attain these goals, and persisting in the face of obstacles and hindrances.

The Concept of Goals

Nowadays, when it comes to answering the question of how motivation relates to action, research relies upon the concept of goals. In our opinion, the concept of goals helps the cognitive explication of a person's readiness to make a certain response and to execute it in an energized manner. Importantly in this regard, Ajzen (1991) suggested that this readiness should be assessed in terms of a person's intention to make the response. Such intentions can be conceived as self-imposed or assigned goals that imply standards that the person intends to meet (with respect to quality and quantity criteria). Doing so allows asking new questions such as, what makes people reach their goals and what makes them fail to do so?

Determinants of Effective Goal Pursuit

Goal content strongly affects the chances of attaining a goal. For instance, Ryan et al. (1996) have argued that goals of autonomy, competence, and social integration favor creativity, cognitive flexibility, deep processing of information, and effective coping with failure. These effects are assumed to be mediated by an intrinsic self-regulation, as the needs of autonomy, competence, and social integration are associated with intrinsic goal striving in line with a person's interests or core values, rather than with extrinsic goal striving in line with environmental pressures or internal sanctions. Intrinsic goal striving is preferred by individuals with positive self-regard, and it can be facilitated from outside by teachers who provide autonomy support (e.g., when law school faculty provide autonomy support, grade point average improves).

In addition to goal content, *structural features of goals* also affect goal pursuit. For example, acting on one's goals is commonly believed to depend on the strength of the goal ('I really want to reach goal x!'). But most tests of this goal strength–behavior relation are only based on correlational studies that preclude causal inferences. A recent meta-analysis by Webb and Sheeran (2006) took a closer look at this assumption by only selecting studies where the strength of the goal was manipulated relative to a control group, and differences in subsequent goal-directed behavior were observed. They found 47 experimental tests of the intention (goal)–behavior relation that actually used an experimental manipulation of the strength of the goal (intention). The meta-analysis showed that a medium-to-large change in strength of intention ($d = 0.66$) led to a small-to-medium change in respective behavior ($d = 0.36$).

But meeting one's goals does not only depend on the strength of the goal; it also depends on what kind of aspiration or standard is specified in the goal (i.e., whether the person wants to achieve a lot or only a little). Locke and Latham (2006) reported that participants are more likely to attain challenging goals that are spelled out in specific terms than moderately specific goals or challenging but vague goals (i.e., 'do your best'). This effect has a number of prerequisites: frequent performance feedback, strong goal commitment, low goal complexity, and the presence of necessary skills and means to the individual. What does not seem to matter is whether the goal is determined from outside (assigned goals), freely chosen by the individual (self-set goals), or chosen in interaction with others (participative goals). As potential mediators of the goal-specificity effect, Locke and Latham pointed to heightened persistence, attention to the execution of goal-directed behaviors, a greater readiness to plan the goal pursuit, and to feedback and self-monitoring advantages.

Goal attainment is also affected by structural features such as goal orientation (i.e., approach vs avoidance orientation, promotion vs prevention, learning vs performance orientation, and low vs high identity-relation). Framing of the orientation of social goals in terms of *approach versus avoidance* clearly affects their attainment. For instance, striving for the goal of making new friends versus striving for the goal of not being lonely produces quite different outcomes. With respect to the outcome variable of interpersonal satisfaction, framing a goal in terms of avoiding loneliness leads to less favorable results than framing it as making new friends (Elliot et al., 2006). These differences seem to be mediated by differential attention and memory processes, differential interpretation and weighting of available information, and differential evaluation of the progress made toward goal attainment.

Higgins (2006) reported that approach goals benefit more from goal striving that makes use of eagerness-related approach strategies (such as pulling things toward oneself) than from vigilance-related avoidance strategies (such as pushing things away from oneself), whereas the reverse is true for avoidance goals. The assumed reason for this is *value from fit*. Higgins argued that people engage more in goal striving when the strategies used match the goal orientation (i.e., eagerness strategies/positive outcome focus; vigilance strategies/negative outcome focus) than when there is a mismatch (i.e., vigilance strategies/positive outcome focus; eagerness strategies/negative outcome focus). This heightened engagement in turn leads to a higher perceived value.

Framing goals in terms of *learning versus performance* has been found to matter as well (Dweck, 1996). Learning goals lead to better achievement than performance goals because they allow for a more effective coping with negative feedback. For people with performance goals, negative feedback signals failure and lack of ability and thus causes them to give up prematurely. People with learning goals, on the other hand, view negative feedback as setbacks and as valuable cues on how to focus on new strategies, ultimately furthering goal attainment. Recent research on the framing of achievement goals in terms of learning versus performance has also investigated its influence on interactions in social achievement situations (Poortvliet et al., 2007). People with learning goals are oriented reciprocally. They give information openly, and they process received information with a focus on those pieces of information that fit well and add value to their own chosen task strategy. On the other hand, people with performance goals are oriented exploitatively. They provide information to others reluctantly, and they process received information with a suspicious attitude that leads them to focus on detecting and disregarding

low-quality information that might hurt their own task performance.

Finally, it matters whether a person frames a given task goal in terms of its *identity-relatedness*. For instance, the task of solving a certain arithmetic problem can be approached with the goal of solving it effectively or the goal of becoming a great mathematician. The latter goal has been referred to as a self-defining goal or *identity* goal, as it specifies an identity as a desired end state. Self-completion theory (Wicklund and Gollwitzer, 1982) proposes that people who are committed to identity goals can undertake a variety of activities to claim identity-goal attainment, because many different behaviors indicate the possession of such identities. For a scientist, for example, such self-symbolizing activities might include engaging in professional duties (e.g., giving lectures); making positive self-descriptions (e.g., 'I discovered a new principle!'); exerting identity-relevant social influence (e.g., advising students); or acquiring respective skills, tools, and material symbols (e.g., programming skills, fast computers, large office). Failing to perform an identity-relevant activity or lacking an identity symbol produces a state of incompleteness; to restore completeness, people engage in self-symbolizing efforts by emphasizing the possession of alternative symbols or setting out to acquire new identity symbols (e.g., engaging in identity-relevant activities, describing oneself as having the required personality attributes, showing off relevant status symbols). Importantly, affirming one's general self-integrity or bolstering one's self-esteem are not sufficient to offset incompleteness regarding an identity goal; rather, one must acquire specific identity symbols (Gollwitzer et al., 2013).

Attaining one's goals, however, does not only depend on the content of the goals and their structural features. It also depends on the *context* in which the person strives for her goals. Research has differentiated various critical context variables that relate to the person's affective state, the array of competing action tendencies, and the power position of the goal striver. For instance, Tice et al. (2001) focused on negative *affect* and observed that feeling emotionally distraught (e.g., having been asked to imagine that one has caused a traffic accident that killed a child) makes it difficult to follow through with goals of not eating unhealthy food or delaying gratification to attain better long-term rewards.

Recently, researchers have also focused on the contextual variable of being in a position of *power* versus being powerless. For instance, Guinote (2008) reported that people in power procrastinate less in pursuing their goals, they persist longer in the face of difficulties, they show more willingness to try out different strategies to attain the goal, and they more readily seize good opportunities to make goal-directed responses. In addition, they more readily recognize whether a given situation can be used to serve their goals and then allow suitable situations to guide their behavior. All of this appears to be facilitated by a change in executive functioning. Powerful individuals are better than powerless ones at updating goal-relevant information (i.e., new information is monitored for goal relevance, and relevant information replaces old, irrelevant information in working memory). They are also better at inhibiting responses that may interfere with the present goal, and at switching between the main goal and respective subgoals.

The Self-Regulation of Goal Pursuit

But pursuing goals with the right content and structure in the proper situational context does not yet guarantee effective goal attainment. People often have to engage in self-regulatory thought to further their goal pursuits. For instance, picking a goal that appears desirable and feasible does not yet guarantee that one effectively strives for this goal. A person may adopt the goal to play the piano because she loves to make music and feels capable of doing so, yet actually attempting to realize this goal takes a further effort, and there are certain self-regulatory strategies that facilitate making this effort.

Instigating Determined Goal Pursuits

The *theory of fantasy realization* specifies three respective self-regulation strategies (Oettingen, 2012): mental contrasting, indulging, and dwelling. In mental contrasting, people first imagine the fulfillment of a wish or fantasy (e.g., giving a good presentation at a conference) and then reflect on the present reality that stands in the way of attaining the desired future (e.g., evaluation anxiety). Mental contrasting is a problem-solving strategy that allows people to recognize that they have not yet fulfilled their wish and that they need to take action in order to achieve the desired future. As a consequence, expectations of attaining the desired future become activated and determine a person's striving to attain the desired future. When perceived expectations of success are high, people will engage in strong goal striving; when expectations of success are low, people will refrain from doing so, and thus they will venture on alternative wishes and desired futures. In this way, mental contrasting helps people discriminate between feasible and unfeasible goals.

The theory of fantasy realization specifies two further routes to goal pursuit. People may engage either in indulging (envisioning only the attainment of the wished-for future) or in dwelling (reflecting only on the present negative reality). Neither of these mental strategies produces any discrepancy between future and reality, and thus the individual fails to recognize that actions (making responses) are necessary to achieve the desired future. Therefore, expectations of success do not become activated, and goal pursuit does not reflect the perceived likelihood of reaching the desired future. Individuals who indulge and dwell show a medium level of striving, even though the resource-efficient strategy to follow would be no engagement in the case of low expectations of success, and full engagement in the case of high expectations of success.

Various experiments support these claims. In one study (Oettingen et al., 2001: Study 4), first-year students enrolled in a vocational school for computer programming indicated their expectations of excelling in mathematics. Next, they named positive aspects that they associated with excelling in mathematics (e.g., feelings of pride, increasing job prospects) and negative aspects of reality, that is, potential personal obstacles (e.g., being distracted by peers or feeling lazy). In the mental-contrasting condition, participants had to elaborate in writing two aspects of the desired future and two aspects of the present reality, in alternating order beginning with the

aspect of the desired future. Participants in the indulging condition were asked to elaborate four aspects of the desired future only; in the dwelling condition they instead elaborated four aspects of the present reality only. As a dependent variable, participants indicated how energized they felt with respect to excelling in math (e.g., how active, eventful, energetic). Two weeks after the experiment, the participants' teachers reported how much effort each student had invested over the interim and provided each student with a grade for that time period. As predicted, only in the mental-contrasting condition did the students feel energized, exerted effort, and earned grades based upon their expectations of success. Those with high expectations of success felt the most energized, invested the most effort, and received the highest course grades; those with low expectations of success felt the least energized, invested the least effort, and received the lowest course grades. To the contrary, participants in both the indulging and dwelling conditions felt moderately energized, exerted moderate effort, and received average grades independent of their expectations of success.

A variety of studies pertaining to different life domains replicated this pattern of results, for example, experiments on studying abroad, acquiring a second language, getting to know an attractive stranger, finding a balance between work and family life, self-improvement, coping with stress, and fulfilling idiosyncratic interpersonal wishes of great importance. Furthermore, intensity of goal pursuit was assessed by cognitive (e.g., making plans), affective (e.g., feelings of frustration), motivational (e.g., feelings of energization), and behavioral (e.g., amount of invested effort) indicators. These indicators were measured via self-report or observations, either directly after the experiment or weeks later.

Recent research suggests that mental contrasting also promotes the choice of suitable means for effective goal striving. Oettingen et al. (2010) examined the mental-contrasting effects on seeking and giving help as means to an end. For college students, mental contrasting about attaining academic help (more than indulging and dwelling) led to expectancy-dependent seeking of help (Study 1), while for critical care nurses mental contrasting about helping patients' relatives led to expectancy-dependent giving of help (Study 2). Mental contrasting also enables people to master negative feedback. A series of three studies (A. Kappes et al., 2012) showed that when expectations of success were high, mental contrasting promoted the processing of relevant negative feedback (which in turn led to effective planning), protected participants' self-view of competence against negative feedback, and led to optimistic as well as effort-related (rather than ability-related) attributions in response to negative feedback.

The mediating processes of mental contrasting pertain to both cognitive and motivational processes (summary by Oettingen, 2012). As for cognitive processes, mental contrasting modulates the strength of the associative link between future and reality and between reality and instrumental means. When expectations of successfully reaching a desired future are high, mental contrasting strengthens the association between the desired future and the reality; when expectations are low, mental contrasting weakens future–reality associations. Importantly, the future–reality associations in turn mediate mental-contrasting effects on self-reported (e.g., feelings of responsibility) and other-rated intensity of goal pursuit (e.g., raters scored quality of performance on giving a talk and solving a creativity test). But mental contrasting not only links future and reality, it also connects present reality to relevant instrumental means (i.e., means instrumental to overcome or circumvent the present reality to attain the desired future). Again, the strength of the reality–behavior associations mediated intensity of goal pursuit (e.g., performance of taking the stairs instead of the elevator to achieve the goal of getting physically fit). Mediating processes of mental contrasting pertain also to motivational processes. Mental contrasting of a desired future with present reality leads to energization in terms of increased systolic blood pressure as well as self-report of feeling energized.

A recent experiment used continuous magneto-encephalography, a brain imaging technique measuring magnetic fields produced by electrical activity in the brain (Achtziger et al., 2009) to test whether mental contrasting qualifies as a sophisticated problem-solving technique. Mental contrasting was compared to indulging in a positive future or simply resting. As compared to resting, mental contrasting produced heightened brain activity in areas associated with working memory, episodic memory, intention maintenance, action preparation, and vivid visualization. That is, mental contrasting demands vividly imagining a desired future, anticipating hindrances to realizing this future, and making plans on how to overcome these barriers. The brain activity associated with indulging, on the other hand, did not differ from resting.

This latter finding raises doubts that solely engaging in highly positive fantasies about a desired future can strengthen a person's goal pursuits. Indeed, research on engaging in positive versus negative fantasies about the future supports this suspicion. Early on, Oettingen and Wadden (1991) observed that obese women who spontaneously indulged in positive fantasies about their weight loss were less successful in achieving a lower body mass index (after 4 months and 2 years) than obese women who spontaneously produced fantasies that were more negative. Moreover, Oettingen and Mayer (2002) observed that people who indulge in positive fantasies (valence and frequency) show comparatively weaker academic achievements (i.e., achieving a good grade in a psychology class), professional achievements (i.e., finding a job after graduation), interpersonal relations (i.e., finding a romantic partner), and health (i.e., recovering from hip surgery). More recently, H.B. Kappes and Oettingen (2011) investigated the effects of experimentally induced positive fantasies on energization, hypothesizing that low energy is a mechanism by which positive fantasies translate into poor achievement. Indeed, induced positive fantasies resulted in less energy (as measured by physiological and behavioral indicators) than fantasies that questioned the desired future, negative fantasies, or neutral fantasies. Additionally, energy measured right after the induction of the positive fantasies mediated accomplishment in everyday life a week later. Finally, positive fantasies yielded a larger decrease in energy when they pertained to a more rather than less pressing need (e.g., need achievement) further suggesting that it is the positivity of fantasies that quells energization. Altogether the results indicate that one reason positive fantasies predict poor

achievement is because they sap energy required to pursue the desired future.

Dealing with Hindrances on the Way to Goal Attainment

One very powerful self-regulation strategy of overcoming hindrances to goal attainment is planning out one's goal striving in advance. Gollwitzer (1999) proposed a distinction between goal intentions and implementation intentions. Goal intentions (goals) have the structure of 'I intend to reach Z!' whereby Z may relate to a certain outcome or behavior to which the individual feels committed. Implementation intentions (plans) have the structure of 'If situation X is encountered, then I will perform the goal-directed response Y!' Both goal and implementation intentions are set in an act of will: The former specifies the intention to meet a goal or standard; the latter refers to the intention to perform a plan. For instance, a possible implementation intention for the goal intention to eat healthy food could link a suitable situational context (e.g., one's order is taken at a restaurant) to an appropriate behavior (e.g., asking for a low-fat meal). Whereas goal intentions merely specify desired end states ('I want to achieve goal X!'), the if-component of an implementation intention specifies when and where one wants to act on this goal, and the then-component of the plan specifies how this will be done. Implementation intentions thus delegate control over the initiation of goal-directed behavior to a specified opportunity by creating a strong link between a situational cue and a goal-directed response.

Evidence that forming if–then plans enhances rates of goal attainment has been obtained in many studies on a whole array of different goals (e.g., academic, health, interpersonal). A recent meta-analysis (Gollwitzer and Sheeran, 2006) involving over 8000 participants in 94 independent studies revealed a medium-to-large effect size ($d = 0.65$; Cohen, 1992) of implementation intentions on goal attainment on top of the effects of mere goals. This size of the implementation intention effect is noteworthy, given that goals by themselves already have a facilitating effect on relevant behavior enactment (Webb and Sheeran, 2006).

Research on the underlying mechanisms of implementation intention effects (summary by Gollwitzer and Oettingen, 2011) found that implementation intentions facilitate goal attainment on the basis of psychological mechanisms that relate to the anticipated situation (specified in the if-part of the plan), the intended behavior (specified in the then-part of the plan), and the mental link forged between the if-part and the then-part of the plan. Because forming an implementation intention implies the selection of a critical future situation, the mental representation of this situation becomes highly activated and hence more accessible. This heightened accessibility of the if-part of the plan has been observed in several studies testing this hypothesis by using different experimental paradigms: for example, lexical decision tasks, the dichotic-listening paradigm, and cued recall. Moreover, forming implementation intentions not only heightens the activation (and thus the accessibility) of the mental presentation of the situational cues specified in the if-component, it also forges a strong associative link between the mental representation of the specified opportunity and the mental representation of the specified response. These associative links are quite stable over time. Moreover, mediation analyses suggest that cue accessibility and the strength of the cue–response link together mediate the impact of implementation intention formation on goal attainment.

Gollwitzer (1999) suggested that the upshot of the strong associative (critical situation – goal-directed response) links created by forming implementation intentions is that – once the critical cue is encountered – the initiation of the goal-directed response specified in the then-component of the implementation intention exhibits features of automaticity, including immediacy, efficiency, uncontrollability, and redundancy of conscious intent. Consistent with this idea, there is plenty of evidence now that if–then planners do act quickly, deal effectively with cognitive demands (i.e., speed-up effects are still evident under high cognitive load), show uncontrolled attention to the specified cues, and do not need to consciously intend to act in the critical moment (i.e., implementation intention effects are observed even when the critical cue is presented subliminally).

The postulated and observed component processes underlying implementation intention effects (enhanced cue accessibility, strong cue–response links, automation of responding) mean that if–then planning allows people to see and seize good opportunities to move toward their goals. Fashioning an if–then plan thus strategically automates goal striving; people intentionally make if–then plans that delegate control of goal-directed behavior to preselected situational cues with the explicit purpose of reaching their goals. This delegation hypothesis was tested in a functional magnetic resonance imaging study reported by Gilbert et al. (2009). In this study, participants had to perform a prospective memory task on the basis of either goal or implementation intention instructions. Acting on the basis of goal intentions was associated with brain activity in the lateral rostral prefrontal cortex, whereas acting on the basis of implementation intentions was associated with brain activity in the medial rostral prefrontal cortex. Brain activity in the latter area is known to be associated with bottom-up (stimulus) control of action, whereas brain activity in the former area is known to be related to top-down (goal) control of action. Support for the delegation hypothesis also comes from studies using critical samples (see Gollwitzer and Oettingen, 2011) – that is, individuals with poor self-regulatory abilities such as people with schizophrenia, people with substance abuse disorders, people with frontal lobe damage, and children with attention-deficit/hyperactivity disorder (ADHD).

Research on the facilitating effects of forming implementation intentions on goal attainment has targeted all of the four major problems that are known to doom effective goal striving: getting started, staying on track, failing to call a halt to futile goal striving, and overextending oneself. Given that forming implementation intentions automates goal striving, people who form implementation intentions should actually have it easier when they are confronted with these four central problems of goal implementation. Indeed, numerous studies suggest that problems of getting started on one's goals can be solved effectively by forming implementation intentions. But many goals cannot be accomplished by a simple discrete one-shot action as they require that people keep striving over an

extended period of time. Such staying on track may become very difficult when certain internal stimuli (e.g., being anxious, tired, overburdened) or external stimuli (e.g., temptations, distractions) interfere with ongoing goal pursuit. Implementation intentions have been found to prevent the negative influence of interferences from inside (e.g., mood) and outside the person (e.g., disruptions by attractive video shows). The self-regulatory problem of calling a halt to a futile goal striving (i.e., disengaging from a chosen but noninstrumental means or from a chosen goal that has become unfeasible or undesirable) can also be ameliorated by forming implementation intentions. And finally, when task performance is regulated by implementation intentions, using different ego-depletion paradigms, research participants who used implementation intentions to self-regulate in one task do not show reduced self-regulatory capacity in a subsequent task.

A new line of research on implementation intentions has been stimulated by Aristotle's concept of akrasia (lack of willpower). It is argued that any willful strategy of goal striving (such as if–then planning) has to prove itself under conditions where people commonly fail to demonstrate willpower. Three such conditions have been analyzed so far: (1) situations in which a person's knowledge and skills constrain performance such as taking academic tests, (2) situations in which an opponent's behavior limits one's performance such as negotiation settings, and (3) situations in which the wanted behavior (e.g., not snacking, not littering) runs into conflict with the respective impulsive antagonistic responses. For all of these situations, forming implementation intentions qualifies as an effective self-regulation strategy (see Gollwitzer and Oettingen, 2011).

The litmus test for any strategy to improve willpower is enhanced performance in a *delay of gratification* task. Children with deficits in inhibitory control (i.e., children with ADHD) are known to have particularly pronounced problems with delaying gratifications. Accordingly, Gawrilow et al. (2011) investigated whether delay of gratification can be facilitated by forming implementation intentions even in children with ADHD. Only when the goal intention to delay gratification was furnished with an implementation intention a heightened ability to delay gratification was observed.

Current Goal Research: Interventions

Knowledge about effective self-regulation strategies of goal pursuit allows one to construct interventions that teach people how to effectively pursue their goals. One such intervention (see Oettingen, 2012) combines mental contrasting with forming implementation intentions into a metacognitive strategy called mental contrasting with implementation intentions (MCII). In a recent intervention study with middle-aged women (Stadler et al., 2009), participants were taught the cognitive principles and individual steps of the MCII self-regulation strategy. This intervention allowed participants to apply MCII to their idiosyncratic everyday wishes and concerns on improving their lifestyle. Specifically, participants were taught to apply MCII by themselves to the wish of exercising more whenever possible. Participants were free to choose whatever form of exercising they wished to apply MCII to, and they were encouraged to anticipate exactly those obstacles that were personally most relevant and to link them to exactly those goal-directed responses that personally appeared to be most instrumental. As dependent measures, participants maintained daily behavioral diaries to keep track of the amount of time they exercised everyday. Overall, teaching the MCII technique enhanced exercise more than the information-only control intervention; this effect showed up immediately after the intervention and it stayed stable throughout the entire period of the study (16 weeks after the intervention). More specifically, participants in the MCII group exercised nearly twice as much: an average of 1 h more per week than participants in the information-only control group.

The MCII intervention was also used to promote healthy eating in middle-aged women (i.e., eating more fruits and vegetables), and it produced the desired behavior change effects; these persisted even over the extensive time period of 2 years (Stadler et al., 2010). In another study, Adriaanse et al. (2010) targeted the negative eating habit of unhealthy snacking in college students. MCII worked for both students with weak and strong such habits and, notably, it was more effective than mental contrasting or formulating implementation intentions alone. Moreover, MCII was observed to benefit chronic back pain patients in increasing their health behaviors (Christiansen et al., 2010). Over a period of both 3 weeks and 3 months patients increased their physical mobility as compared to a standard treatment control group. Physical mobility was measured by objective (i.e., bicycle ergometer test and number of lifts achieved in 2 min) and subjective indicators (reported physical functioning). Finally, MCII has shown beneficial effects outside the health area. For example, it benefited study efforts in adolescents preparing for standardized tests (Duckworth et al., 2011). Together, these findings suggest that MCII is a cost- and time-effective self-regulation technique when it comes to the effective self-regulation of goal pursuit.

Summary

Research on goals has started to conceptualize the individual as the master of her own fate. People only need to apply empirically tested self-regulation strategies (e.g., mental contrasting and forming implementation intentions) to effectively attain their goals. These strategies allow realizing idiosyncratic wishes according to what is desirable and feasible. When applied in metacognitive form, these strategies can liberate people from being bound to erroneous engagement and from being plagued by action control deficits (e.g., impulsivity).

See also: Academic Engagement; Affect-Regulation Motivation; Avoidance and Approach Motivation: A Brief History; Control Behavior: Psychological Perspectives; Coping across the Lifespan; Decision Making: Nonrational Theories; Grit; Health Self-Regulation, Motivational and Volitional Aspects of; Interest, Psychology of; Motivation in Australian Aboriginal Populations; Passion and Motivation; Personal Projects; Self-Determination Theory; Self-Regulation in Adulthood; Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model.

Bibliography

Achtziger, A., Fehr, T., Oettingen, G., Gollwitzer, P.M., Rockstroh, B., 2009. Strategies of intention formation are reflected in continuous MEG activity. Social Neuroscience 4, 11–27.

Adriaanse, M.A., Oettingen, G., Gollwitzer, P.M., Hennes, E.P., De Ridder, D.T.D., De Wit, J.B.F., 2010. When planning is not enough: fighting unhealthy snacking habits by mental contrasting with implementation intentions (MCII). European Journal of Social Psychology 40, 1277–1293.

Ajzen, I., 1991. The theory of planned behavior. Organizational Behavior and Human Decision Processes 50, 179–211.

Atkinson, J.W., 1957. Motivational determinants of risk-taking behavior. Psychological Review 64, 359–372.

Bandura, A., 1997. Self-efficacy: The Exercise of Control. Freeman, New York.

Christiansen, S., Oettingen, G., Dahme, B., Klinger, R., 2010. A short goal-pursuit intervention to improve physical capacity: a randomized clinical trial in chronic back pain patients. Pain 149, 444–452.

Cohen, J., 1992. A power primer. Psychological Bulletin 112, 155–159.

Duckworth, A.L., Grant, H., Loew, B., Oettingen, G., Gollwitzer, P.M., 2011. Self-regulation strategies improve self-discipline in adolescents: benefits of mental contrasting and implementation intentions. Educational Psychology 31, 17–26.

Dweck, C.S., 1996. Implicit theories as organizers of goals and behavior. In: Gollwitzer, P.M., Bargh, J.A. (Eds.), The Psychology of Action: Linking Cognition and Motivation to Behavior. Guilford Press, New York, pp. 69–90.

Elliot, A.J., Gable, S.L., Mapes, R.R., 2006. Approach and avoidance motivation in the social domain. Personality and Social Psychology Bulletin 32, 378–391.

Gawrilow, C., Gollwitzer, P.M., Oettingen, G., 2011. If-then plans benefit delay of gratification performance in children with and without ADHD. Cognitive Therapy and Research 35, 442–455.

Gilbert, S.J., Gollwitzer, P.M., Cohen, A.-L., Oettingen, G., Burgess, P.W., 2009. Separable brain systems supporting cues versus self-initiated realization of delayed intentions. Journal of Experimental Psychology: Learning, Memory, and Cognition 35, 905–915.

Gollwitzer, P.M., 1999. Implementation intentions: strong effects of simple plans. American Psychologist 54, 493–503.

Gollwitzer, P.M., Kirchhof, O., 1998. The willful pursuit of identity. In: Heckhausen, J., Dweck, C.S. (Eds.), Motivation and Self-regulation across the Life Span. Cambridge University Press, Cambridge, England, pp. 389–423.

Gollwitzer, P.M., Marquardt, M.K., Scherer, M., Fujita, K., 2013. Identity-goal threats: engaging in distinct compensatory efforts. Social Psychological and Personality Science 4, 555–562.

Gollwitzer, P.M., Oettingen, G., 2011. Planning promotes goal striving. In: Vohs, K.D., Baumeister, R.F. (Eds.), Handbook of Self-regulation: Research, Theory, and Applications, second ed. Guilford, New York, pp. 162–185.

Gollwitzer, P.M., Sheeran, P., 2006. Implementation intentions and goal achievement: a meta-analysis of effects and processes. Advances in Experimental Social Psychology 38, 69–119.

Guinote, A., 2008. Power and affordances: when the situation has more power over powerful than powerless individuals. Journal of Personality and Social Psychology 95, 237–252.

Heckhausen, H., 1991. Motivation and Action. Springer-Verlag, Berlin.

Higgins, E.T., 2006. Value from hedonic experience and engagement. Psychological Review 113, 439–460.

Kappes, A., Oettingen, G., Pak, H., 2012. Mental contrasting and the self-regulation of responding to negative feedback. Personality and Social Psychology Bulletin 38, 845–856.

Kappes, H.B., Oettingen, G., 2011. Positive fantasies about idealized futures sap energy. Journal of Experimental Social Psychology 47, 719–729.

Locke, E.A., Latham, G.P., 2006. New directions in goal-setting theory. Current Directions in Psychological Science 15, 265–268.

McClelland, D.C., 1985. How motives, skills, and values determine what people do. American Psychologist 41, 812–825.

Oettingen, G., 2012. Future thought and behavior change. European Review of Social Psychology 23, 1–63.

Oettingen, G., Mayer, D., 2002. The motivating function of thinking about the future: expectations versus fantasies. Journal of Personality and Social Psychology 83, 1198–1212.

Oettingen, G., Pak, H.J., Schnetter, K., 2001. Self-regulation of goal-setting: turning free fantasies about the future into binding goals. Journal of Personality and Social Psychology 80, 736.

Oettingen, G., Stephens, E.J., Mayer, D., Brinkmann, B., 2010. Mental contrasting and the self-regulation of helping relations. Social Cognition 28, 490–508.

Oettingen, G., Wadden, T.A., 1991. Expectation, fantasy, and weight loss: is the impact of positive thinking always positive? Cognitive Therapy and Research 15, 167–175.

Poortvliet, P.M., Janssen, O., Van Yperen, N.W., Van de Vliert, E., 2007. Achievement goals and interpersonal behavior: how mastery and performance goals shape information exchange. Personality and Social Psychology Bulletin 33, 1435–1447.

Ryan, R.M., Sheldon, K.M., Kasser, T., Deci, E.L., 1996. All goals are not created equal: an organismic perspective on the nature of goals and their regulation. In: Gollwitzer, P.M., Bargh, J.A. (Eds.), The Psychology of Action: Linking Cognition and Motivation to Behavior. Guilford Press, New York, pp. 7–26.

Stadler, G., Oettingen, G., Gollwitzer, P.M., 2009. Physical activity in women. Effects of a self-regulation intervention. American Journal of Preventive Medicine 36, 29–34.

Stadler, G., Oettingen, G., Gollwitzer, P.M., 2010. Intervention effects of information and self-regulation on eating fruits and vegetables over two years. Health Psychology 29, 274–283.

Tice, D.M., Bratslavsky, E., Baumeister, R.F., 2001. Emotional distress regulation takes precedence over impulse control: if you feel bad, do it! Journal of Personality and Social 80, 53–67.

Webb, T.L., Sheeran, P., 2006. Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. Psychological Bulletin 132, 249–268.

Weiner, B., 1992. Human Motivation. Sage, Newbury Park, CA.

Wicklund, R.A., Gollwitzer, P.M., 1982. Symbolic Self-completion. Erlbaum, Hillsdale, NJ.

Motivation and Emotion: Reflections on the History of an Attributional Approach

Bernard Weiner, University of California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article outlines the history of an attributional approach motivation and emotion from the author's perspective. The article highlights the underlying empirical and theoretical consistency in the apparent face of changes in directions. In addition, the article discusses the legacies that have remained.

In this article, I consider the history of my attribution work, conducted with many outstanding students and colleagues, and focus on the underlying constancy in the apparent face of changes in directions. In addition, I stress the legacies, if any, that were left behind. I found Freudian insights valuable because I realized I had to navigate between three forces: the id-derivative hedonic biasing pushing me to overestimate success, increasing my pride and self-esteem; the superego forces directing me to recognition of failure; and finally, to the realistic position fostered by the ego that some of our work has merit and hopefully will have staying power, whereas other ideas have already been relegated to the dustbin of history. This article is heavily driven by the id forces and attempts to show the underlying empirical and theoretical unity of this work. I begin with some broad theoretical observations and then turn to specific research.

First, another number three in addition to the Freudian taxonomy is pertinent to this overview. I am the product of (at least) three traditions I inherited from my mentor, John Atkinson: one I maintained and the other two I rejected. The tradition I held, or at least came to embrace, is trying to create a so-called 'grand' theory of motivation in the shadows of Hull and Lewin. That is, I eventually wanted to develop a conceptual system, a system of interrelated constructs, which would transcend specific phenomena and identify the 'deep structures' of motivated behavior. Hence, the work of my colleagues and me has addressed not only achievement strivings but also helping, aggression, reactions to the stigmatized, impression formation, punishment, emotions, blame and responsibility, political ideology, coping, change programs, consumer reactions, poverty, and more. I have had less interest in specific observations and data than in the general theory guiding the research. I once gave a talk on an attribution analysis of poverty and was asked by an approving, altruistic-type student how I happened to develop this interest. I think some serious commitment to social justice was the expected response. To her great dismay and disappointment, I replied that I was not particularly interested in poverty and could just as happily written about wealth if it was subject to an equally interesting attribution analysis. This is a lost tradition in motivation: in my day, it was Hull versus Tolman, or Spence versus Atkinson. Theories, not specific research studies, defined the field. I am sorry or disappointed to believe that this contribution of our work has not met with great success.

One of the two traditions I left behind is another specific grand theory – Expectancy/Value theory. I do not doubt that both expectancy and value play important roles in choice and behavior, as has been well documented. Indeed, this may be more consistent with common sense than is attribution theory – what we do depends on what we expect to get and the likelihood of getting it. But it does not, or has not, readily incorporated emotions and the moral aspects of behavior to which I have been drawn.

The second tradition I gave up is a combination of individual differences and projective measurement. I scored thousands of Thematic Apperception Tests (TATs) as a student and a beginning assistant professor, and I left graduate school without any publications. But there was a lot of data in our file drawers. I blame both the measurement instrument and the individual difference approach for the publication failures. Searching for interactions, particularly with individual differences that are considered traits or motives (as opposed to, for example, ideologies or temperaments) generally fails to yield replicable data or capture a great deal of behavioral variance. My preference has been not only to find main effects but main effects that are beyond doubt or question so that replication in front of a classroom is a certainty. I am disturbed by the movement to find or create phenomena that contradict common sense, which so often can't replicate, as now is being discussed regarding the alleged priming effects on behavior. I think in the near future popular phenomena including stereotype threat, ego depletion, value affirmation, implicit self-esteem, implicit attitudes, and a number of others that defy common sense and highlight the unconscious will suffer the same negative fate. We are in an era where the unconscious and the unbelievable have ascended; I am reminded of my work with the TAT.

But if something does not defy common sense, then it often is common sense and findings are not given much credit, labeled as 'bubba psychology,' as though such effects need not be incorporated by theories. I think my goal, or at least one goal, has been to create a conceptual framework that systematizes common sense or the way that the naive person thinks. I inherited this from Fritz Heider. I also believe that the vast majority of our findings are true yet the data are not trivial, particularly when placed within a larger theoretical framework. I must admit that some of these caveats or guiding principles preceded and directed my research, whereas others followed from the research findings. With this background or foundation in mind, let me now turn to some historical developments.

The 1960s Decade: Finding Predictors of Achievement

As already indicated, the bulk of my research as a graduate student with Atkinson was consigned to a no-significant-effects file drawer. Upon leaving graduate school, my focus remained on achievement motivation, although I supplemented that with research on motivation and memory to increase my publication rate. The main issue regarding achievement motivation that I addressed was what psychological variables might predict achievement striving, since I did not think achievement motives as assessed with the TAT did that. My search brought me to Julian Rotter, who in 1966 published his monograph on locus of control. This intuitively seemed a good road to follow: locus of control was a cognitive variable at a time the field was talking about but not specifying the cognitive determinants of behavior; it was assessed with an objective rather than a projective individual difference measure; and his data regarding expectancy change in skill versus luck situations were quite replicable and systematic.

Therefore, I conducted a study in around 1967 with an undergraduate, Penelope Potepan, where we gave the TAT, which was scored for achievement needs, and the Locus of Control Scale and then examined their relations. Some obvious conclusions, primarily about locus of control, dawned on me during this research. First, it was evident that locus of control for success had to be separated from locus of control for failure. One reason for this is that if success is ascribed to high ability, then failure by that individual is not ascribed to low ability. That is, one does not say I succeeded because I am smart but failed because I am dumb. But this was not the case for effort: for the same person, high effort might be construed as a cause of success and low effort a cause of failure. And this resulted in the next obvious fact: ability had to be separated from effort when considering causal locus, although both are internal determinants of achievement. I believe that a good part of my scientific work has been determining the ways in which effort and ability differ, ultimately resulting in a sin versus sickness designation given their absence, seemingly far from where I started but in reality quite close.

Around 1968, so toward the end of the 1960s, a number of events took place of personal research significance. First, I read a book by Richard deCharms, a student of McClelland, in which he mentioned a study of moral evaluation where a person did not repay a debt either because he did not have the money or spent it on something else. I connected this with ability and effort and devised an experiment with Andy Kukla in which we varied effort, ability, and outcome and examined evaluations by others. For example, a person of high ability and low effort failed an exam; evaluate this person. I thought I was doing research on achievement motivation but this had nothing (or little) to do with that topic inasmuch as this is a study of moral evaluation. This research was published in 1970 and became designated a 'citation classic' when it had 100 citations. Today, with Google Scholar, one can get 100 citations just counting letters from your mother.

This publication had a peculiar history. I submitted a three-study manuscript to what is now JPSP, where Bill McGuire, the verbose, critical, and smart, was the editor. He sent me something like a ten-page single-spaced critique, asking for revision. By the time I returned it, the manuscript contained four studies, and he again asked for revision with about a five-page review. Another round yielded an additional experiment and a shorter review. Finally, a six-study manuscript was accepted. It helped my visibility greatly and remains my second most cited empirical paper (exceeded only by the study of stigma I published about 20 years later with Ray Perry). It also taught me the value of replication and multiple-study publications.

Linked with this advance was a six-person conference organized by Hal Kelley that also included Ned or Edward Jones and Dick Nisbett. There I realized I was working within a framework called attribution theory, although I had never seriously (if at all) read Heider or Kelley or Jones. Attribution theorists were social psychologists influenced by quite a different literature than the history that was guiding me, which included Hull, Lewin, Atkinson, and Rotter as well as phenomena related to learning schedules, resistance to extinction, expectancy of success, and the effects of success and failure on achievement performance. I was trained in personality and motivation and not in social psychology.

Another important determinant of my subsequent research life emerged from the Jones-Kelley attribution conference: a paper written with Irene Frieze, Andy Kukla, Henry Reid, Stan Rest, and Bob Rosenbaum. This manuscript contained the beginning of our theoretical analysis of the ability versus effort distinction. There it was stated the reason that luck versus skill tasks have different effects on expectancy, which Rotter had documented and is certainly true, is not because of their locus differentiation but because the perceived causes of the outcomes at these tasks varied on endurance or stability. Luck comes and goes; skill endures. Hence, luck and skill differ in something other than locus. That is, I was expanding the properties of causal beliefs and specified causal stability as the predictor of changes in expectancy of success. I also argued that locus influenced affect (pride) or value, hence trying to remain within the Expectancy/Value framework of Atkinson but placing the data within an attributional framework. More specifically, it was suggested that the relation between expectancy of success and pride presumed by Atkinson is mediated by locus attributions – success at a low expectancy or difficult task gives rise to person attributions and thus high pride, whereas success at a high expectancy of success or easy task promotes situation attributions and hence low pride. These were core theoretical advances for our approach.

The book that emerged from the conference, published in 1972, had a significant impact on the field of social psychology and helped to usher in a nearly two decade era during which attribution theory was at the center of research activity. It temporarily replaced the anti-commonsense dissonance approach. As I already stated, we now have returned once again to the non-commonsense approach (aided by neuromania).

The 1970s: Refinement, Reflection, and Replication

The greater part of the 1970s involved theoretical refinement and some elaboration. First and foremost was that a graduate student, Bob Rosenbaum, introduced the idea of control as a third property of causes. I was initially unresponsive to this because I felt theoretically overwhelmed already and was not

sure how it would be disentangled from locus and stability, which accounted for affect and expectancy. But he proved more than correct. I then began to understand that causes could be placed within a three-dimensional space, with some difficulties, and that they varied quantitatively, not qualitatively. This provided a theoretical foundation that framed two major questions for that period: what are the antecedents of achievement-related causal beliefs and what are their consequences.

Rather than pursuing individual difference antecedents, for example, high need for achievement might give rise to effort beliefs and high locus of control to internal beliefs regarding the determinants of performance, I looked now to social psychology and particularly to Hal Kelley. With Irene Frieze, we explored the informational antecedents of effort and ability beliefs such as consistency and distinctiveness; and with Anna Kun, we considered some of the schemata antecedents, examining the distinction between necessary versus sufficient causality. At that time, it was also firmly established with a number of students that indeed stability rather than locus influences expectancy of success. As well, the ability/effort distinction on evaluation was replicated a number of times: there is no doubt that most reward is allocated to the low-capacity person who overcomes this handicap through hard work and most punishment is given to the high-capacity individual who fails because of not trying. These represent our moral heroes and moral villains. I have to admit some disappointment when critics address the replication issue in social psychology and do not recognize this body of knowledge.

This all became synthesized for a second time (the first was the chapter in the 1972 attribution theory book spearheaded by Jones and Kelley) in a manuscript published in the *Journal of Educational Psychology* in 1979 entitled "A Theory of Motivation for some Classroom Experiences". This article was invited by the editor, Sam Ball; I otherwise would not have undertaken it. On my Google Scholar listing, this is my second most cited journal article.

Then, as the 1970s came to a close, some new research directions emerged that pulled me away from achievement motivation and dominated the remaining years of my career. These shifts were toward emotions and interpersonal motivation, with the latter umbrella including, as earlier listed, help giving; impression management, especially excuses and confession; reactions to the stigmatized; aggression; blame and responsibility; political ideology; poverty; and punishment.

I was glad to leave or at least supplement the achievement research because achievement performance, and especially grade point average, is an overdetermined and complex index, making it difficult to predict and for attributions to carry much predictive variance. I am much more comfortable predicting judgments of others, as in the research with Kukla on evaluation, with Frieze on causal information, and with Kun on causal schemata.

The Late 1970s and Most of the 1980s

I am not exactly sure when the shifts to emotion and interpersonal behavior took place or even if there was a specific time or event. I had long looked on Atkinson's motivation formula with suspicion, not because it was an Expectancy/Value theory but because I did not know of evidence that the primary affect associated with achievement success is pride or that the main failure-produced affect is shame, which he had presumed. Along with Dan Russell and David Lehrman, we conducted some simple studies to examine this question. For example, in an early investigation we merely asked participants to recall a time they succeeded or failed and to recount the emotions that were experienced.

It is not surprising that many emotions were reported. The clear companion to success is happiness and the corresponding affect accompanying failure is unhappiness. This is closer to the Freud/Hull pleasure/pain distinction rather than to pride and shame, as Atkinson had specified. But pride was also reported for success, along with emotions including gratitude, relief, surprise, and many others. The failure-related affects comprised a longer list, including shame but also anger, guilt, hopelessness, helplessness, humiliation, regret, and more. Similar self-report experiments that involved feelings when others succeeded and failed generated additional emotions including admiration, envy, sympathy, pity, and more.

Previously, I regarded causes and causal dimensions as the foundation of our attribution theory and asked what are their antecedents and consequences. But now I shifted focus and asked what are the antecedents of emotions and what are the motivational consequences of feeling states. I limited the antecedents to causal beliefs and/or causal dimensions and began to create a cognitive or attribution taxonomy of emotions, asking what causal dimension placements give rise to what feelings.

I also then realized I was beginning to develop a theory of motivation in which events give rise to causal interpretations that arouse emotions, which goad action. This was a thinking–feeling–acting sequence, hence incorporating the traditional tripartite division of psychology into cognition, emotion, and conation. This temporal sequence captured what I believed to be the 'deep structure' of motivation. It also was becoming clearer to me that this could serve in the 'grand theory' mold of a cognitive theory of motivation, but the key cognition was causal belief and not expectancy of success.

With Sandra Graham, I explored the dimensional antecedents of pity (uncontrollable causality), anger (external, controllable causality), and guilt (internal, controllable causality). I now think there are about 15 or 20 such cause–affect associations that may be described as necessary or sufficient relationships. These include, for example, stable causality for failure with helplessness or hopelessness, unstable causality for failure and hope; controllable causality for other success and admiration; external controllable causality for personal success with gratitude; and, of course, internal causality for success and pride.

We also showed that conveying these emotions provide cues for personal beliefs about causes; that is, there are relations going from causal dimensions to emotions as well as from communicated emotions to causal beliefs. For example, conveying pity tells someone that their failure was uncontrollable, while communicating anger reveals beliefs of controllable causality. This brought more social psychology into the study of emotions. While doing this I also pursued the moral components of achievement strivings with Jonathon Brown,

documenting the positive moral value of succeeding because of high effort.

But my main desire, a full test of the proposed cognition–emotion–action theory of motivation, awaited the detour from achievement performance to a domain where judgments regarding behavior were possible as the dependent variable. I found that in the area of helping behavior.

A series of research investigations was then conducted in the helping domain, primarily with Greg Schmidt and Hector Bentancourt. The paradigm was simple – alter perceptions of perceived control of a need for help, measure affects of anger and sympathy, and get some dependent variable that reflects help. For the latter we often used lending class notes (in those days, just sending the notes over the computer and therefore not losing hold of them was not yet prominent). The pattern of findings was very clear and replicable – control over a need (specifically, going to the beach instead of to class) gives rise to annoyance and neglect by others, whereas lack of control (needing the notes because of eye problems) arouses sympathy and help giving. Conceptually, this was similar to failing a test because of not studying and subsequent punishment versus failing because of lack of capacity and subsequent withholding of punishment, as I had shown with Kukla. But in that research affects were not included and now, about a dozen years later, they were considered as bridging the gap between thinking and doing. I first published the helping data in 1980, in what is my third most cited empirical article.

It was then a small step to recognize that one does not want to be perceived as in control of a need and to not have acted to resolve this problem. It is known that this gives rise to anger in others and neglect: part of our commonsense wisdom in everyday interactions. I then began another series of studies, with Jim Amirkahn, Valerie Folks, Gregg Gold, Sandra Graham, Orli Peter, Seiji Takako, and others on the topic of impression management, managing perceptions of others by using excuses or offering an apology or confession. The positive effects of changing impressions from controllable to uncontrollable, or what is labeled an excuse, or altering person perception by showing remorse and a full confession, are quite powerful and unquestionable. Valerie Folks extended this to reactions of consumers to companies; this field of study now often calls upon attribution theory.

Given these newer advances, I felt it was time for a more comprehensive theoretical statement and wrote a manuscript for *Psychological Review*, published in 1985 and titled "An Attributional Theory of Achievement Motivation and Emotion". This article then became expanded into a monographic book, without achievement in the title. The *Psychology Review* article and subsequent book appeared when I was about 50 years old; they are by far my most cited contributions. This is common in psychology, so for those of you not yet 55 and not sufficiently cited, do not fret!

The Very Late 1980s, the Era of the 1990s

It was evident that I had now shifted attention away from expectancy and the stability dimension, and away from pride and the locus dimension, the core themes of an attributional analysis of achievement strivings, and moved toward other-perceptions of control, their consequences, and how to alter them as exemplified in studies of help giving and impression management. I am again not sure what produced the next less major shift, but I turned from the controllability of a specific action such as needing class notes to the controllability of a condition. The conditions were stigmas, and they included AIDS, homosexuality, obesity, schizophrenia, and others. At that time, in the late 1980s, Ray Perry, Judy Chipperfield, and I initiated research on perception of the causes of stigmas. Once again these were simple judgment studies where instead of varying ability and effort, we described heart disease as due to genetics or an unhealthy lifestyle, or AIDS as due to promiscuous sexual behavior or a blood transfusion, and so on. The results were quite clear – perceived controllability for a stigma has very negative attitudinal and behavioral consequences, whereas uncontrollability is linked with positive other-perception in terms of moral evaluation. In the early 1990s, I wrote a paper for the *American Psychologist* with sin and sickness in the title, and I realized this also could describe the absence of effort (sin) versus the absence of ability (sickness). This fully integrated my achievement research with the studies of interpersonal behavior. Stigma researchers now often turn to these ideas so that this work has had some impact on health psychology, which now recognizes personal responsibility or controllability as the central differentiating property of stigmas.

This line of research also resulted in a more serious consideration of morality and the importance of perceptions of responsibility. I realized I had to move from causal control to personal responsibility because not all controllable transgressions are punished. There are mitigators such as a very young age or mental illness that sever the relation between control and anger and punishment. In 1995, I published a book on personal responsibility and in 2006 followed this with a book focusing on social justice and social emotions that expanded these ideas and considered the concept of responsibility in greater detail.

Into this Century

Again, a few related lines of research were opened by the prior topics that extended into the year 2000 and beyond. Foremost among them was the notion that justice is judged differently by liberals and conservatives. We all know that liberals tend to perceive environmental causality for social ills such as poverty (e.g., lack of educational opportunities), whereas conservatives perceive personal responsibility for poverty (e.g., laziness, lack of saving). In early research with Gail Sahar and later with Lisa Farwell and Danny Osborne, we returned to individual differences, but not involving needs or motives but rather regarding more embracing political ideologies, relating them to perceptions of control and beliefs about welfare (another index of helping). In addition, with Christine Reyna and P.J. Henry, we explored how attributions relate to punishment, differentiating retributive punishment meted out for controllable actions as opposed to utilitarian punishment. Finally, with Udo Rudolph and Scott Roesch, we completed a meta-analysis of nearly 40 research studies that verified the attribution–affect–action conception in both the helping and aggression domains. I have read an aphorism recently that poetically stated: my head said

something was interesting and unusual; my heart said to try it. That almost captures the theory.

Summing up

Hal Kelley once asked me: "So, what are you mucking around with nowadays?" as though I shifted around from one topic to another. But I regard my career as more unified than that, following a rather narrow or circumscribed attribution path that logically progressed from one step to another. I think at each step a brick was laid for a scientific structure, sometimes an empirical and/or sometimes a theoretical residue. I think each individual empirical finding is in accord with common sense. I hope so. I also think that the whole is different from, or other than, the sum of the parts and at the genotypic level transcends common wisdom. The empirical observations should be judged as true or false; the theory should be judged as useful or not.

The common empirical theme throughout much of my work surprisingly (to me) concerns moral evaluation. As I mentioned earlier, my most cited empirical study was with Ray Perry on stigma evaluation and perceptions of control, published in 1988. The second most cited empirical work was published in 1970 with Andy Kukla and relates to achievement evaluation when the causes of success and failure are ability and effort. And my third most cited empirical work was published in 1980 and examines help-giving when classroom-related aid is desired because of controllable or uncontrollable causes. It is clear that these all explore the phenomenon of moral evaluation. The data are without doubt regarding the effects of controllability (bad, immoral) and uncontrollability (good, moral) on the emotions and judgments of others as related to stigmas, achievement evaluation, and desires to help others.

The underlying and unifying theoretical system is that events and states, whether succeeding or failing an exam, having a stigma, or needing help, give rise to specific causal interpretations that are categorized into causal properties. These causal categories, in turn, arouse emotions, that then guide actions. This thinking–feeling–acting system is more easily tested in interpersonal contexts, but applies as well in intrapersonal settings and in achievement contexts. Hence, the underlying theory has generality across motivational domains and to both the self and others.

The data so briefly described above will have to be accounted for by any psychological theory, so even if this attribution conception does not stand the test of time, I believe that the data will endure. On the other hand, I also regard the theoretical representation as having strength.

See also: Avoidance and Approach Motivation: A Brief History; Motivation and Actions, Psychology of; Race and Academic Motivation; Schooling: Impact on Cognitive and Motivational Development; Self-Determination Theory; Temperament and Motivation.

Bibliography

Weiner, B., Frieze, I.H., Kukla, A., Reed, L., Rest, S., Rosenbaum, R.M., 1971. Perceiving the Causes of Success and Failure. General Learning Press, Morristown, NJ.

Weiner, B., 1979. A theory of motivation for some educational experiences. Journal of Educational Psychology 71, 3–25.

Weiner, B., 1985. An attributional theory of achievement motivation and emotion. Psychological Review 92, 548–573.

Weiner, B., 1995. Judgments of Responsibility: A Foundation for a Theory of Social Conduct. Guilford, New York.

Weiner, B., 2006. Social Motivation, Justice, and the Moral Emotions. Erlbaum Press, Mahway, NJ.

Motivation in Australian Aboriginal Populations

Rhonda G Craven and Alexander Seeshing Yeung, Institute for Positive Psychology and Education, Australian Catholic University, Strathfield, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Many Indigenous people worldwide are more disadvantaged compared with their non-Indigenous peers. Australian Indigenous people have been identified as one of the most disadvantaged groups in the world. Consistent with positive psychology approaches, motivational constructs have been identified for non-Indigenous students from diverse cultural backgrounds as critical psychological drivers that facilitate desirable educational outcomes and life opportunities beyond schooling. However, due to limited empirical research with Indigenous populations, less is known about the role of motivational constructs for this population. This article explores the importance of motivational constructs for Indigenous populations with a focus on Indigenous Australian students.

Introduction

Many Indigenous people worldwide are more disadvantaged compared with their non-Indigenous peers as measured by diverse socioeconomic indicators. For example, in Australia and the USA, low numbers of Aboriginal and Native American students complete postsecondary education (Australian Bureau of Statistics (ABS), 2008; US Department of Education, National Center for Educational Statistics, 2008). Given that education predicates life's opportunities, it is of dire concern that generations of Indigenous people from different countries do not achieve commensurate educational outcomes compared with their non-Indigenous peers. The latter contributes to a lack of self-reliance, persistent intergenerational socioeconomic disadvantage, ensuing loss of talent, and undermining nations' prosperity and well-being.

In Australia, national reports and all Australian governments in the last three decades have identified that Indigenous Australians are the most significantly disadvantaged Australians on all socioeconomic indicators including education, and have the highest school dropout rate of any cultural group in Australia (e.g., ABS, 2008; Commonwealth of Australia, 2011). Recent research has also acknowledged that Indigenous Australians are one of the most disadvantaged populations in the world (see Yeung et al., 2013). Many Indigenous Australian students leave school before completing Year 10 of high school, fail to complete the final 2 years of high school (Years 11 and 12), and of those who complete the final years of schooling, fail to gain entry to university. This situation led the Ministerial Council on Education, Employment, Training, and Youth Affairs (MCEETYA, 2006) to conclude that: "These outcomes limit the postschool options and life choices of Indigenous students, perpetuating intergenerational cycles of social and economic disadvantage" (p. 4). Additionally those Indigenous Australians that qualify for university entry have lower completion rates than their non-Indigenous peers. Shergold (2013), who had responsibility for the oversight of Indigenous programs in Australia for over 20 years, has emphasized that, while the governments he served strived to make a positive difference, little progress had been made.

Part of the systemic failure to address Indigenous disadvantage in Australia and elsewhere in the world (e.g., US Department of Education, National Center for Educational Statistics, 2008) can be attributed to research with Indigenous populations employing deficit modeling approaches. Deficit models focus on the failure of Indigenous people, their communities, and social systems to address disadvantage (Riley and Ungerleider, 2012). Deficit approaches result in lower expectations becoming entrenched in individual mindsets, communities, organizations, and policy. For example, Zubrick et al. (2006: vi) have noted that in Australia:

> There has been tacit acceptance of the non-achievement of educational standards by Aboriginal children and young people. The resultant acceptance of this lack of educational success has a cumulative effect. It is based on the belief that Aboriginal children and young people will never reach their full potential and if they fall behind society then welfare will protect them. Their low level of educational success is accepted as a normative expectation.

In contrast, some researchers have attempted to address the issues from a more positive perspective. This perspective has gained increasing attention within educational psychology research with Indigenous populations with a focus on explicating what seeds success. This new direction capitalizes upon a positive psychology approach (or asset-orientated approach) to optimize the psychological resources and well-being of Indigenous students (Craven and Bodkin-Andrews, 2006; McInerney, 2008; McInerney and King, 2013). Positive psychology approaches focus on identifying and emulating how well adjusted and healthy individuals get the most out of life. Fundamental to this new direction in research with Indigenous populations is examining the role of motivational constructs in facilitating desirable educational outcomes and positive pathways beyond schooling that enable Indigenous people to both succeed and flourish. As such, researchers have gone beyond the previous emphasis on the socioeconomic disadvantages these students experience to identify psychological factors that drive academic success. Researchers have started to investigate the psychological well-being of Indigenous students (e.g., Purdie and McCrindle, 2004), their motivation in schoolwork (e.g., McInerney, 2003, 2008), and their self-concept (e.g., Craven and Bodkin-Andrews, 2006). Their new findings have broadened our knowledge about the

disadvantage Indigenous students suffer and have enabled us to identify potentially potent psychosocial tools to inform intervention.

The purpose of this article is to discuss some of these new directions in relation to Indigenous Australians and delineate significant motivational theory and research that have the potential to facilitate Indigenous academic success and well-being. First, we provide an overview of the pervasiveness of Indigenous Australian educational disadvantage. Next, we discuss recent advances in positive psychology approaches and motivational theory. We then provide an overview of some examples of recent research that has capitalized upon positive psychology approaches, motivation and self-concept theory and research to advance our understandings of the importance of adaptive motivational orientations for Indigenous Australian students. Finally, we emphasize the potential of key research-derived motivational constructs for facilitating Indigenous peoples to flourish.

Nature of Indigenous Australian Educational Disadvantage

Recent reports of the Organization for Economic Development and Cooperation (OECD) indicate that students who disengage from school are less likely to find employment or to remain employed during times of economic hardship (OECD, 2011). In comparison to non-Indigenous students, Indigenous Australian students in all age groups from all Australian States and Territories display lower school attendance, retention, and achievement (Craven and Bodkin-Andrews, 2006). For example, the number of Indigenous Australian students progressing into the final year of secondary schooling (Year 12) is much lower (42.9%) than non-Indigenous students (75.6%). Hence it is apparent that Indigenous students generally do less well academically than their non-Indigenous peers. These educational inequities are so pervasive that they extend across preschool through high school (De Bortoli and Thomson, 2010). This is well illustrated by results across three waves of the Program for International Student Assessment (PISA) data (2000, 2003, 2006), whereby 15-year-old Indigenous Australian students' results were well below the OECD average for reading literacy, mathematics, and science (De Bortoli and Thomson, 2010). Furthermore, reflecting the persistent nature of educational disadvantage for Indigenous Australians, over this 6-year period, there was no improvement in Indigenous students' performance.

This educational disadvantage is further exacerbated in relation to tertiary education. While in 2006 Indigenous Australians made up 2.5% (i.e., 517 200 citizens) of the total population of Australia (i.e., 21 017 200 citizens) (ABS, 2008), they were underrepresented in universities (Demosthenous, 2012). Even though university participation, commencement, and completion rates for Indigenous Australians have increased over the past two decades, these rates are not commensurate with the increase in the Indigenous population and the increased participation by non-Indigenous students in tertiary education. For example, in 2009, the gap between the retention of Indigenous and non-Indigenous students was 31.9%. Furthermore, 26% of Indigenous people, in comparison to 53% of non-Indigenous people, had a nonschool qualification in 2006; Indigenous people were also four times less likely to have a Bachelor's degree (5% as compared with 21%) and twice less likely to have an Advanced Diploma or Diploma (9% compared with 4%) (ABS, 2008).

In 2011, an Australian Government's report (Commonwealth of Australia, 2011) noted that the proportion of Indigenous 20- to 24-year-olds who had completed Year 12 of schooling or equivalent was increasing. However, even though Indigenous students who received a Year 12 certificate increased from 20% in 2001 to 26% in 2008, the proportion was far below the non-Indigenous rate of around 56%. Furthermore, demonstrating the international prevalence of this issue, education statistics for Native American students in the United States are similar to those for Indigenous Australian students (US Department of Education, National Center for Education Statistics, 2008). Given these persistent educational inequalities for Indigenous students in different countries, there is a need to identify factors that seed success to ensure that Indigenous students attain commensurate outcomes to their non-Indigenous peers and therefore enjoy similar life opportunities. Improving educational outcomes for Indigenous people also comes with significant socioeconomic benefits. For example, in Australia "if the educational level of the Indigenous population was increased to that of the non-Indigenous population, the value of this hypothetical change [to the nation] is $1.09 billion per annum" (Taylor et al., 2012: viii). Economy-wide advantages are also likely to result from improvements in the quality of life of Indigenous Australians, including more than $10 billion in extra gross domestic product by 2029, an increase in government revenue by $4.6 billion, and a reduction in government expenditure by $3.7 billion (Access Economics, 2008: iv). Hence addressing Indigenous educational disadvantage is a significant international socioeconomic issue of our time.

How to seed success in educational attainment remains unclear, particularly for Indigenous people. Indigenous education is also plagued by a paucity of evidence-based research. For example, Mellor and Corrigan (2004), in reviewing contemporary Indigenous education research, lamented the lack of empirical research noting that: "There is not, in Australia, a research tradition of quantitative measurement in the Indigenous education literature. To ignore such measurement only continues to do injustice to the gravity of the problem" (pp. 46–47). They also noted that research methodology has been limited by factors such as small case studies; focusing on a small subset of the population (e.g., communities with a high Indigenous population); isolating Indigenous education research from the broader discourses in other disciplines such as psychology, sociology, and health; and that the relation between cause and effect has been asserted rather than demonstrated by research. Clearly, there is a dearth of research being undertaken in Indigenous education yet even though such research could be used to improve the educational outcomes for Indigenous populations (Bin-Sallik, 2005). However, as yet advances in motivational theory and research have only just begun to be explored with Indigenous populations. Two such advances involve capitalizing upon

positive psychology approaches and advances in the dominant motivation theory – Expectancy Value Theory (EVT) – which are discussed in the next section.

Advances in Motivational Theory and Research

There is a positive revolution sweeping psychology that emphasizes how individuals get the most out of life, manage to stay healthy and happy in the face of adversity (Craven and Bodkin-Andrews, 2006; Craven and Marsh, 2008). Positive psychology focuses on how healthy, normal, and exceptional individuals can get the most from life; hence positive psychology is the study of optimal functioning that makes life more fulfilling and results in building capability and well-being across the lifespan. The popularity of positive psychology is derived from the research evidence base that focuses on sources of psychological health and well-being as key drivers of multiple desirable outcomes such as adaptive motivation, enhanced performance and academic outcomes, and higher educational aspirations, all of which ultimately contribute materially to societal well-being. However, in order to facilitate the goal of supporting effective intervention within a positive psychology framework, the pathology of maladaptive functioning (e.g., depression, violence, racism, etc.) also needs to be synergistically accounted for; therefore positive psychology research does not preclude inquiries into the negative side of human behavior. Positive psychology approaches have become popular due to criticisms that traditional psychology has focused exclusively on a pathological perspective neglecting the potential of positive self-fulfillment. Instead of concentrating on repairing the damage using a disease model of human functioning, Seligman and Csikszentmihalyi (2000) concluded that an emphasis on the 'disease model' does not enable a focus on how to prevent the problems psychologists aim to cure. Instead, a positive psychology approach would be more powerful in defining effective ways to improve one's well-being by enhancing the self.

Simultaneously, in Australia and throughout the world, there is an increasing emphasis on how to cultivate the well-being and psychological tools of Indigenous peoples (Craven and Bodkin-Andrews, 2006). However, there is a paucity of research on Indigenous populations that could have capitalized on recent advances in positive psychology theory and research. Fresh solutions from a positive psychology perspective are urgently required to address complex issues to successfully cultivate Indigenous students' well-being and address their educational disadvantage. We propose that part of the solution lies in positive psychology approaches that capitalize on advances in motivational theory to result in adaptive Indigenous student motivation that would facilitate schooling engagement, aspirations to participate in higher education, and reaping the benefits thereof. The importance of motivational constructs for education is well demonstrated by Thomson and Hillman (2010) who, based on the Longitudinal Study of Australian Youth (LSAY) extension of PISA 2003, examined the futures of students at age 19 who at age 15 were low achievers. They found that the key predictors at age 15 for long-term success – even controlling for socioeconomic status (SES) – were mathematics motivation, educational and occupational aspirations, enjoyment of school, and a positive school climate. On the basis of these results, they concluded: "The significant influence of motivation on young people's later outcomes is an important message for parents, teachers and policy-makers" (p. 31).

Expectancy Value Theory (EVT), a dominant theory of motivation, explains motivation and decision-making in relation to expectancy of success in a given task (E) and the value of the outcome (V) (Eccles and Wigfield, 2002). Expectancy is conceptualized as a task-specific self-belief about likely success in a future academic task (Eccles and Wigfield, 2002) and is often operationalized as academic self-concept in educational research. Value is hypothesized to have multiple components: importance, interest, usefulness, and cost. Eccles, Wigfield, and their colleagues have been influential motivation theorists for more than a quarter of a century, stimulating research and creating insights into many long-standing real-world problems, leading to fundamental insights into the sociocultural precursors of achievement motivation. Modern EVT has been successful in placing the interplay between expectancy and value into a sociocultural and developmental context. For example, Eccles and Wigfield have noted that EV perceptions are developmental and that educational and employment choices are made in the context of young people's socialization, past educational experiences, and codevelopment with parents, teachers, and peers. As such, EVT provides a useful conceptual framework for understanding student motivation, engagement, aspirations, and long-term outcomes. In the next section, we review some of the emerging motivational research with Indigenous Australians that has begun to capitalize upon these advances in positive psychology and motivational theory and research.

Academic Motivation of Aboriginal Students

Motivation can be conceptualized as students' energy and drive to engage in learning, and is considered to facilitate students' interest and engagement in, and enjoyment of schooling (Craven and Marsh, 2005). Actual academic behaviors that enhance learning (e.g., persistence, on tasks, academic striving behaviors), and achievement are thought to be closely associated with motivation, which is also linked with self-regulation in learning, enhanced academic performance and self-concept, and aspirations for further education (Craven and Marsh, 2005; Eccles and Wigfield, 2002). As such, researchers have advocated motivation as a potentially potent construct for addressing Indigenous educational disadvantage despite only an emerging amount of empirical research with this population (Bodkin-Andrews et al., 2010a), examples of which follow.

In general, some differences were found between Indigenous and non-Indigenous Australian students, but these differences were mostly small (e.g., Yeung et al., 2013). For example, McInerney's (2003, 2008) cross-cultural motivational research program conducted over decades represents a rare sustained program of empirical research that has investigated the structure, measurement, and relation of motivational constructs for Indigenous and non-Indigenous students. In

general, McInerney and colleagues have found little difference between Indigenous and non-Indigenous students' motivational profiles on eight achievement goal dimensions, five achievement goal measures, four sense-of-self measures, and three academic self-concept measures although there were some differences in degree of endorsement. For example, McInerney (2003) examined motivational factors and academic components of self-concept of urban and rural Indigenous and non-Indigenous school students and found that Indigenous students reported significantly higher levels of competition, social power, affiliation, praise, token, social and performance goal orientations, although these variables typically did not predict a range of achievement outcomes for either Indigenous or non-Indigenous groups. More importantly, there were no significant differences on task, effort, social concern, mastery general, global motivation, sense of purpose for education, sense of reliance in learning, positive self-esteem, English self-concept, or math self-concept. Among these, key predictors of a range of achievement outcomes for both the Indigenous and non-Indigenous groups were effort, sense of purpose, mastery general, global motivation, English self-concept, and general self-concept, and there were no significant differences between the groups on these dimensions. The Indigenous students did have significantly lower self-esteem and lower general valuing of motivation, however. These results led McInerney to conclude: "these results tell a positive story regarding the capacity of Aboriginal children to do well at school given the right sort of motivational environment" (2003: 343).

In later research, McInerney demonstrated the important relations of motivation to schooling outcomes. For example, McInerney (2008) found that Anglo Australian, Asian Australian, Lebanese Australian, and Indigenous Australian students all endorsed positive achievement goal motivation constructs, and for all student groups, mastery orientations were positively associated to schooling outcomes. In sum, McInerney's studies over the last two decades with Indigenous Australians have led him to conclude that students even from very remote Indigenous communities have a very similar frame of reference to non-Indigenous students with regard to motivational goals in education. Furthermore, McInerney and King (2013) examined the roles of self-concept and achievement goals in facilitating self-regulation and deep learning strategies for Indigenous Australian ($n = 319$) and Native American high school students ($n = 100$) and found that self-concept was positively related to mastery goals, deep learning, and educational outcomes across both groups and the relations between self-concept and self-regulated learning whether fully or partially mediated through mastery were strong, demonstrating the salience of both for predicting self-regulated learning. This led McInerney and King (2013) to conclude for Indigenous students:

> academic self-concept and mastery goals are critical psychological ingredients that can contribute to optimal learning … as indicated by our results, the strong positive relations between self-concept and positive learning outcomes suggest that educators who are interested in facilitating mastery goal adoption among Indigenous students can draw from the rich literature on improving students academic self-concepts.

Some research has suggested that the largest motivational differences between Indigenous and non-Indigenous groups can be observed across negative motivational constructs such as self-handicapping and disengagement (Bodkin-Andrews et al., 2010a). Bodkin-Andrews et al. (2013) tested the extent to which adaptive and maladaptive motivational constructs, and their relations to educational outcomes, were similar and different for Indigenous Australian ($n = 227$) and non-Indigenous ($n = 1047$) high school students utilizing the Motivation and Engagement Scale – High School (MES) developed by Martin (2009). They found that Indigenous students in comparison to non-Indigenous students had significantly lower levels of self-belief, mastery, and value of schooling, and significantly higher levels of uncertain control, anxiety, failure avoidance, self-handicapping, and disengagement when compared with non-Indigenous students. All 11 factors measured by the MES were significantly associated with at least one schooling outcome for non-Indigenous students, and eight of the 11 had at least one significant association for Indigenous students. These results led Bodkin-Andrews et al. to conclude that Indigenous students may benefit from more adaptive motivational profiles.

Martin et al. (2013) have proposed that academic buoyancy is an important facilitator of academic motivation for both Indigenous and non-Indigenous students. Experiencing academic adversity is theorized to disrupt the motivational processes important for planning and intentions to undertake further education and training. Academic buoyancy is commonly operationalized as a capacity to effectively deal with academic setback, challenge, and adversity (e.g., challenging schoolwork, threats to self-confidence from negative feedback from educators, study pressure, and periods of poor performance). Martin et al. hypothesized that because Indigenous students experience greater academic adversity, academic buoyancy should serve a more salient role in predicting their educational intentions in comparison to non-Indigenous students. They also hypothesized that, because non-Indigenous students do not experience the same extent of educational adversity as Indigenous students, the variance in educational intentions explained by academic buoyancy should not be as large. As such, Martin et al. proposed that academic buoyancy may reduce the higher incidence of academic adversity experienced by Indigenous Australians in comparison to non-Indigenous Australians. They also suggested that academic buoyancy may be associated with positive educational intentions to complete school and engage in further education and training. Martin et al. (2013) examined the role of academic buoyancy in predicting educational intentions for Indigenous ($n = 350$) and non-Indigenous ($n = 592$) Australian high school students. They found that Indigenous students held less positive educational intentions and that academic buoyancy was associated with more positive educational intentions. In addition, the effects of academic buoyancy on educational intentions were found to be stronger for Indigenous than non-Indigenous students. The effects were not significant for junior high students – but they were significant for senior high school students and the effect of academic buoyancy on educational intentions was stronger for Indigenous senior students than non-Indigenous students. These results led Martin et al. to conclude that

academic buoyancy serves a salient role in Indigenous students' educational intentions in senior high school and therefore could be potentially potent to include in psychoeducational interventions.

The above examples evince that recent advances in motivational theory and research are gleaning new insights into research-derived motivational constructs that may inform interventions to seed success in addressing Indigenous educational disadvantage. Self-concept has also been found to be a key aspect of motivation with important implications for Indigenous populations which is discussed in the following section.

Self-concept Research with Indigenous Australian Populations

Academic self-concept refers to one's perceptions about one's academic ability (Craven and Marsh, 2008). Research has demonstrated that academic self-concept is associated with a wide range of adaptive academic outcomes and shares a dynamic and mutually reinforcing reciprocal causal relation with academic achievement (Craven and Marsh, 2008). More importantly, this mutually reinforcing relation is domain-specific such that self-concept in the math domain, for example, influences math outcomes, which also subsequently influence math self-concept; but these mutual relations are specific within the math domain and hence do not transfer to other domains of self-concept (e.g., verbal). Multidimensional self-concepts are also considered to be a powerful component of well-being fundamental to successfully coping with adversity. Self-concepts have also been demonstrated by a body of international research "to impact on a wide range of critical well-being outcomes and serve as an influential platform for enabling full human potential" (Craven and Marsh, 2008: 104). A positive academic self-concept has also been found to influence coursework selection, engagement, intrinsic motivation, subsequent achievement, educational aspirations, and subsequent university attendance (Craven and Marsh, 2008).

Consistent with a positive psychology emphasis, a positive self-concept is valued as a desirable outcome in many disciplines including education. Attesting to this pervasive significance of the self-construct for Indigenous populations and the outcomes that are mediated by it, the NSW Aboriginal Education Consultative Group Incorporated (NSW AECG) notes the failure of education systems to maximize Indigenous children's sense of identity, and to proactively enhance Indigenous children's academic self-concepts so as to ensure that they feel good about themselves (Craven and Bodkin-Andrews, 2006). As such, maximizing self-concept has been proposed as a vital key to unlock Indigenous educational disadvantage (Craven and Bodkin-Andrews, 2006; Craven and Marsh, 2005, 2008). Hence, positive self-belief is valued as a variable that makes good things happen and facilitates the realization of full human potential for Indigenous students. Furthermore, international work undertaken by the OECD has concluded that self-concept needs to be seen alongside with school achievement as an important outcome of schooling. This is because students with a higher academic self-concept are more likely to display better behavior, maintain psychological well-being, and perform well in achievement tests (Craven and Marsh, 2008).

Similarly, Craven and Marsh (2008) claim that enhancing self-concept is a vital goal in and of itself and that self-concept is an important mediating variable that causally impacts on a variety of desirable outcomes including academic achievement. Indeed, many researchers argue that there exists an extensive literature, spanning diverse disciplines and theoretical perspectives, suggesting that high self-concept promotes goals, expectancies, coping mechanisms, and behaviors that facilitate productive achievement and work experiences, and impede mental health problems. Capitalizing on previous advances in theory and empirical research, Marsh and Craven developed a reciprocal effects model (REM) whereby the causal relation between academic self-concept and achievement is conceived as dynamic and reciprocal. In reviewing the available international evidence (see Craven and Marsh, 2008), they found that support for the REM is particularly strong in relation to academic self-concept and school performance whereby increases in specific domains of self-concept lead to increases in associated performance domains and other desirable outcomes and improved performance leads to better self-concepts. These results clearly demonstrate the pervasive significance of the self-concept construct, offer further support for the theory on which this research is based, and affirm that enhancing self-concept has a causal influence on achievement such that enhancing skills alone is not enough; people need also to hold positive self-concepts of their abilities in specific areas. Given recent international research demonstrating the reciprocal effects of achievement and self-concept in academic domains (Craven and Marsh, 2008), Indigenous students may experience double disadvantage in relation to educational academic performance and self-concept. As yet these research advances have not been fully extended to Indigenous students but interesting insights can be gleaned from some of the emerging literature with Indigenous populations in this field.

Craven and Marsh (2005) sought to better understand the nature and structure of Indigenous students' self-concepts and their relation to aspirations and other desirable educational outcomes. The study aimed to characterize Indigenous students' academic and nonacademic self-concepts in comparison to non-Indigenous students; schooling and postschooling aspirations, the relation of SES and academic self-concept to educational outcomes (aspirations, goals for the future, enjoyment of school, attendance, and students' perceptions of their teachers' ratings of their ability), and compare and contrast the pattern of results for Indigenous and non-Indigenous students. They evaluated multidimensional self-concepts, aspirations, ability perceptions, SES, and perceptions of barriers faced in attaining aspirations for Indigenous ($n = 524$) and non-Indigenous ($n = 1149$) Year 7–12 secondary students ($N = 1673$) from urban and rural regions of three Australian States. Marsh and Craven found that Indigenous students, compared with non-Indigenous students, had significantly higher physical, appearance, general self-esteem, and art self-concepts. Indigenous students also had lower academic self-concept (math, verbal, school), peer relations, honesty, and emotional stability self-concept. No significant differences were present for parent self-concept. For both

Indigenous and non-Indigenous students, Craven and Marsh found that a high academic self-concept and high SES are associated with wanting to stay on at school longer, efficacy in achieving future goals, school enjoyment, lower absences, and higher student perceptions of teachers' ratings of ability. Craven and Marsh noted that "these results offer further support for the suggestion that the effects of SES for Indigenous students are partly mediated by academic self-concept" (p. 225) and "imply that academic self-concept may indeed be a potent determinant of a wide variety of desirable educational outcomes" (p. 228). In addition, they noted that since for 7 of 11 facets of self-concept measured, Indigenous students had lower self-concept, and "Given the established importance of enhancing self-concept as a means to advance academic achievement and other desirable educational outcomes, it is suggested that enhancing Indigenous students' academic self-concepts needs to be seen as a priority for intervention" (p. 227).

In regard to aspirations, significantly more Indigenous students in comparison to non-Indigenous students aspired to leave school earlier than non-Indigenous students, reflecting lower retention rates for Indigenous students. Most Indigenous students (37.1%) aimed to get a job after leaving school, whereas most non-Indigenous students aimed to go to university (78.6%), compared with 30.1% Indigenous students. More Indigenous students (24.6%) compared with non-Indigenous students (19.4%) aimed to undertake vocational educational and training postschooling. These results reflect the lower participation of Indigenous students in university education. This pattern of results led Craven and Marsh to conclude that "more effective strategies need to be put in place to assist Indigenous students set and attain higher schooling, future education and employment aspirations" (p. 226). To identify what barriers might impede students from achieving their aspirations, students were asked to rate the extent that they felt nine factors might limit them for achieving their postschooling aspirations (teacher support, family support, absences, achievement, availability of further education facilities, available job opportunities, employer attitudes, career education knowledge they needed to achieve their goals, amount of career advice they had received). Indigenous students in comparison to non-Indigenous students reported higher scores for all nine barriers. Based on these results Craven and Marsh suggested that "a more concerted strategic approach needs to be developed by a diversity of agencies to address barriers Indigenous students perceive as limiting the achievement of their aspirations" (p. 227). Overall, based on the results of this investigation, Craven and Marsh emphasized that: "identifying effective strategies to enhance Indigenous students' self-concepts in specific facets, could prove to be a vital and fundamental key in unlocking Indigenous educational disadvantage" (p. 229). The importance of targeting domain-specific academic self-concepts is also emphasized in the work of Bodkin-Andrews et al. (2010b), who found that across 176 Indigenous Australian students, verbal/English self-concept predicted 7.72% of the variance in English grades, and math self-concept predicted 7.53% of the variance in math grades, over and above the effects of numerous home-educational resources.

Further complicating the issue of potential disadvantages Indigenous students may face is the apparent difference of opportunities in urban and rural school settings. Major concerns for rural Australians include access to educational opportunities, quality of educational provisions, financial issues, and attitudes toward education. Yeung et al. (2013) examined the self-concepts (i.e., school, reading, math, art, and physical abilities), school enjoyment, school participation, and achievement (self-assessment and standardized test) of a sample ($N = 1342$) of Indigenous ($n = 572$) and non-Indigenous ($n = 770$) students from 52 primary schools (Years 3–6) located in rural and urban areas of New South Wales. Employing a 2 (identity: Indigenous vs non-Indigenous) $\times$ 2 (region: urban vs rural) design, they found significant effects of identity for all variables except for art self-concept. For the three academic self-concepts (school, reading, and math), Indigenous students had significantly lower scores than non-Indigenous students. For physical self-concept, Indigenous students had significantly higher scores than non-Indigenous students. However, for art self-concept, Indigenous students did not differ from non-Indigenous students. Indigenous students also had lower scores in enjoyment of school than non-Indigenous students. They also had lower scores in participation than non-Indigenous students. For achievement, Indigenous students had significantly lower standardized achievement scores in both reading and math. Similarly, Indigenous students' self-ratings of school work were significantly lower than for non-Indigenous students. In contrast, the main effects of region were not statistically significant. However, significant identity $\times$ region interaction effects were found in the achievement scores. For reading achievement, Indigenous students in urban schools scored higher than those in rural schools whereas non-Indigenous students in rural schools did better than those in urban schools. A similar pattern was found for math achievement. However, the effect sizes were all very small (all partial $\eta^2 < 0.10$). Therefore, even though there seem to be some subtle differences between Indigenous and non-Indigenous students in urban and rural settings, the authors cautioned that given the small effect sizes, such differences should not be overstated.

Based on these findings, Yeung et al. (2013) concluded that Indigenous students were clearly disadvantaged in terms of academic achievement, and displayed lower school enjoyment, and participation irrespective of region (urban or rural). Yeung et al. also noted that for nonacademic self-concepts the results did not support previous research of a relatively higher art self-concept for Indigenous children based on stereotypical perspectives, but did support previous research demonstrating higher physical self-concept for Indigenous students and lower academic self-concepts. The consistent pattern of Indigenous students displaying lower scores for both achievement and self-concept led Yeung et al. to conclude that Indigenous students were doubly disadvantaged in achievement and self-concept. They also suggested that "the enhancement of both performance in school work and academic self-concepts is a potentially potent intervention to contribute to breaking the vicious cycle of disadvantage experienced by Indigenous students" (p. 420) and "These results imply that school personnel would be well advised to not assume stereotypic differences between Indigenous and non-Indigenous Australian students or assume a great difference between Indigenous students from urban and rural school settings" (p. 405).

To examine the role of self-concept in students' engagement and disengagement in academic work, Bodkin-Andrews et al. (2012) tested structural equation models with longitudinal data obtained from Aboriginal and non-Aboriginal Australian students. The results, with a sample of 264 Indigenous Australian and 947 non-Indigenous Australian high school students, showed that Aboriginal students are more likely to disengage from school, but the causal impact of this disengagement seems to be negated when the causal impact of academic self-concept is also considered. Hence supporting McInerney and King (2013), the Bodkin-Andrews et al. (2012) findings suggest that enhancing academic self-concept may be a useful strategy for addressing patterns of school disengagement that have been repeatedly found for Indigenous Australians.

Summary

This article has outlined that to address the academic disadvantage of Indigenous students, researchers have begun to go beyond the previous emphasis on the socioeconomic disadvantages these students experience and deficit approaches to capitalize upon recent advances in motivational theory and research based on non-Indigenous populations and positive psychology approaches to identify what seeds success. Researchers have started to investigate Indigenous students' motivation, their multidimensional self-concepts, and the drivers of psychological well-being. These new research findings have broadened our knowledge about the disadvantage Indigenous students suffer and have enabled researchers to devise research-derived potentially potent strategies to inform intervention to prevent and address Indigenous educational disadvantage. Indigenous students and non-Indigenous students' motivational orientations and structure of self-concept seem more similar than different which bodes well on extending advances in international motivational theory, research, and practice based on non-Indigenous populations to Indigenous populations. Current research findings also imply that Indigenous students could benefit from more adaptive motivational orientations and less maladaptive orientations, building academic buoyancy, and enhanced academic self-concepts across multiple domains. Overall, it seems important for future theory, research, and practice in Indigenous education research to employ positive psychology approaches and the best available international theory and research to invest in elucidating motivational constructs that seed success and importantly explicate what enables Indigenous people to flourish.

See also: Academic Engagement; Gender and Academic Motivation; Grit; Indigenous Psychology; Interest, Psychology of; Mortality: The Great Historical Decline; Motivation and Actions, Psychology of; Motivation, Learning, and Instruction; Race and Academic Motivation; School Achievement: Motivational Determinants and Processes; Schooling: Impact on Cognitive and Motivational Development; Self-Determination Theory; Self-Regulated Learning: Theories, Measures, and Outcomes.

Bibliography

Access Economics Pty Limited, 2008. An Overview of the Economic Impact of Indigenous Disadvantage Report by Access Economics Pty Limited for Reconciliation Australia viewed 09.07.13. http://www.reconciliation.org.au/getfile?id=81&file=AE_Report_The_economic_impact_of_Indigenous_disadvantage_Aug_08.pdf.

Australian Bureau of Statistics, 2008. Education across Australia. cat. no. 4102.0. ABS, Canberra.

Bin-Sallik, M., 2005. Preface. In: Craven, R.G., Tucker, A., Munns, G., Hinkley, J., Marsh, H.W., Simpson, K. (Eds.), Indigenous Students' Aspirations: Dreams, Perceptions and Realities. DEST, Commonwealth of Australia, Canberra.

Bodkin-Andrews, G., Craven, R.G., Parker, P., Kaur, G., Yeung, A.S., 2013. Motivational cognitions and behaviours for metropolitan Aboriginal and non-Aboriginal Australian students: assessing the relations between motivation and school engagement. In: Liem, G.A.D., Bernardo, A.B.I. (Eds.), Advancing Cross-Cultural Perspectives on Educational Psychology: A Festschrift for Dennis McInerney. Information Age Publishing, Charlotte, NC.

Bodkin-Andrews, G., Dillon, A., Craven, R., 2010a. Bangawarra'gumada – strengthening the spirit: causal modelling of academic self-concept and patterns of disengagement for Indigenous and non-Indigenous Australian students. Australian Journal of Indigenous Education 39, 24–39.

Bodkin-Andrews, G., O'Rourke, V., Craven, R.G., 2010b. The utility of general self-esteem and domain-specific self-concepts: their influence on Indigenous and non-Indigenous students' educational outcomes. Australian Journal of Education 54 (3), 277–306.

Bodkin-Andrews, G.H., O'Rourke, V., Dillon, A., Craven, R.G., Yeung, A.S., 2012. Engaging the disengaged?: a longitudinal analysis of the relations between Indigenous and non-Indigenous Australian students' academic self-concept and disengagement. Journal of Cognitive Education and Psychology 11 (2), 179–195.

Commonwealth of Australia, 2011. Overcoming Indigenous Disadvantage, Report. Author, Melbourne, Victoria.

Craven, R.G., Bodkin-Andrews, G.H., 2006. New solutions for addressing Indigenous mental health: a call to counsellors to introduce the new positive psychology of success. Australian Journal of Guidance and Counselling 16 (1), 41–54.

Craven, R.G., Marsh, H.W., 2005. Dreaming futures: an empirical analysis of Indigenous Australian students' aspirations, self-concepts, and realities. In: Marsh, H.W., Craven, R.G., McInerney, D.M. (Eds.), International Advances in Self Research, vol. 2. Information Age, Greenwich, CT.

Craven, R.G., Marsh, H.W., 2008. The centrality of the self-concept construct for psychological well-being and unlocking human potential: implications for child and educational psychologist. Educational and Child Psychology 25 (2), 104–118.

De Bortoli, L., Thomson, S., 2010. Contextual Factors that Influence the Achievement of Australia's Indigenous Students: Results from PISA 2000–2006. ACER, Camberwell, VIC.

Demosthenous, C.M., 2012. Inclusion/exclusion for Aboriginal and Torres Strait Islander students: understanding how 'we' matters. Journal of Social Inclusion 3 (1), 71–85.

Eccles, J.S., Wigfield, A., 2002. Motivational beliefs, values, and goals. Annual Review of Psychology 53, 109–132.

Martin, A.J., 2009. Motivation and engagement across the academic life span a developmental construct validity study of elementary school, high school, and university/college students. Educational and Psychological Measurement 69 (5), 794–824.

Martin, A.J., Ginns, P., Papworth, B., Nejad, H., 2013. Aboriginal/Indigenous students in high school: understanding their motivation, engagement, academic buoyancy, and achievement. In: Liem, G.A.D., Bernardo, A. (Eds.), Advancing Cross-Cultural Perspectives on Educational Psychology: A Festschrift for Dennis McInerney. Information Age Publisher, Charlotte, NC.

Ministerial Council for Education, Early Childhood Development and Youth Affairs, 2006. Report–Australian Directions in Indigenous Education, 2005–2008. MCEETYA, Carlton South, VIC.

McInerney, D.M., 2003. Motivational goals, self concept, and sense of self – what predicts academic achievement?. In: Marsh, H.W., Craven, R.G., McInerney, D. (Eds.), International Advances in Self Research, vol. 1. Information Age Press, Greenwich.

McInerney, D.M., 2008. Personal investment, culture and learning: insights into school achievement across Anglo, Aboriginal, Asian and Lebanese students in Australia. International Journal of Psychology 43 (5), 870–879.

McInerney, D.M., King, R.B., 2013. Harnessing the power of motivational factors for optimizing the educational success of remote Indigenous students: a cross-cultural study. In: Craven, R., Mooney, J. (Eds.), Seeding Success in Indigenous Higher Education: What Research Says? Diversity in Higher Education Series. Emerald Publishing, USA.

Mellor, S., Corrigan, M., 2004. The Case for Change: A Review of Contemporary Research on Indigenous Education Outcomes. ACER, Melbourne.

Organization for Economic Co-operation and Development, 2011. Education at a Glance: OECD Indicators. OECD, Paris.

Purdie, N., McCrindle, A., 2004. Measurement of self-concept among Indigenous and non-Indigenous Australian students. Australian Journal of Psychology 56 (1), 50–62.

Riley, T., Ungerleider, C., 2012. Self-fulfilling prophecy: how teachers' attributions, expectations, and stereotypes influence the learning opportunities afforded Aboriginal students. Canadian Journal of Education 35, 303–333.

Seligman, M.E.P., Csikszentmihalyi, M., 2000. Positive psychology: an introduction. American Psychologist 55, 5–14.

Shergold, P., 2013. Preface. In: Craven, R.G., Dillon, A., Parbury, N. (Eds.), In Black & White: Australians All at the Crossroads. Connor Court Publishing Pty Ltd, Ballan, VIC.

Taylor, J., Gray, M., Hunter, B., Yap, M., Lahn, J., 2012. Higher Education and the Growth of Indigenous Participation in Professional and Managerial Occupations. Centre for Aboriginal Economic Policy Research, ANU, Canberra.

Thomson, S., Hillman, K., 2010. Against the Odds: Influences on the Post-school Success of 'Low Performers'. National Centre for Vocational Education Research (NCVER), Adelaide, SA.

US Department of Education, National Center for Education Statistics, 2008. The Condition of Education 2008. NCES. 2008-031. NCES, Washington, DC.

Yeung, A.S., Craven, R.G., Ali, J., 2013. Self-concepts and educational outcomes of indigenous Australian students in urban and rural school settings. School Psychology International 34 (4), 405–427.

Zubrick, S., De Maio, J., Shepherd, C., Griffin, J., Dalby, R., Mitrou, F., Lawrence, D., Hayward, C., Pearson, G., Milroy, H., Milroy, J., Cox, A., 2006. The Western Australian Aboriginal Child Health Survey: Improving the Educational Experiences of Aboriginal Children and Young People. Curtin University of Technology and Telethon, Institutes for Child Health Research, Perth.

Motivation in Clinical Interventions

Daniel W McNeil and Laura L Quentin, West Virginia University, Morgantown, WV, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Motivation is a multidimensional drive state that serves as a foundation for most clinical interventions, whether they are psychosocial, pharmacotherapeutic, or some combination. An essential element of health behavior change and maintenance of healthy behaviors, motivation is affected by extrinsic factors such as environmental contingencies and internal factors that are established by learning history. Models of Stages of Change and choice behavior help explicate motivation as a construct. Motivational interviewing is a contemporary, prevalent, and evidence-based approach to assessing and enhancing motivation in clients and patients in health care and social service settings.

Definitions, Conceptualizations, and Contemporary Issues

Motivation is a multidimensional drive state that can be modified, reduced, or enhanced (Delinsky et al., 2011; Hayhurst, 2013), and has been defined as "the impetus that gives purpose or direction to human or animal behavior and operates at a conscious or unconscious level" (VandenBos and American Psychological Association, 2007, p. 594). It can operate at primary (e.g., hunger) or secondary (e.g., individual interests and goals) levels and is affected by forces that are internal (e.g., physical needs such as sleep) or external (e.g., encouraging comments from a coach while playing in a sporting event). The regulation of various systems (i.e., social, biological, and cognitive) also depends upon motivation (Ryan and Deci, 2000). In behavioral terms, motivation is described in terms of establishing operations, which are procedures or conditions "that [change] the effectiveness of a stimulus as a reinforcer or punisher…" and also includes "drive states within the organism" (Catania, 1991, p. G15).

In health care, including medical, dental, psychological, and other therapies involving clinical interventions, patient or client motivation may be necessary both for positive behavior change (e.g., initiating an exercise routine) and for maintenance of health behaviors (e.g., daily toothbrushing) (Madukwe, 2013). As a hypothetical construct (MacCorquodale and Meehl, 1948), motivation can only be inferred and is not directly observable. The term 'motivation,' a label given to a construct, is known through behavior. It can be influenced by social forces, including interactions, inclusive of a patient or client's relationship with his or her clinician (Madukwe, 2013). In fact, it is argued that it is the duty of the clinician to assess and enhance, if necessary, his or her client or patient's motivation in treatment (Madukwe, 2013), although it may best be conceptualized as a shared responsibility of both the patient or client and the clinician.

Internal and External Motivation

Individuals are driven to act in ways and to be involved in activities that are engaging and reinforcing, and to avoid activities that are disinteresting, unpleasant, or otherwise punishing. Most conceptualizations of motivation regard it as something 'inside' a person, that he or she 'has,' and that a person can activate motivation by one's own choice. An oft-told humorous story illustrates this 'internal' view of motivation: "How many mental health professionals does it take to change a light bulb?" and the answer is: "Only one, but the light bulb really has to want to change." Internal motivation has been described as "the inherent tendency to seek out novelty and challenges, to extend and exercise one's capabilities, to explore, and to learn" (Ryan and Deci, 2000, p. 70). According to Cognitive Evaluation Theory, autonomy, self-determination, and competence are critical for internal motivation (Deci and Ryan, 1985). Individuals who are more internally motivated have been found to report more enthusiasm and confidence related to a task, which may lead to greater persistence, creativity, and improved performance on the task. Internally motivated behaviors and goals are associated with greater overall well-being. These views of motivation are that it is intrinsic or internal to the person.

Another view, which typically is portrayed as an opposite or completely unrelated state, is extrinsic or external motivation, in which people are prompted to act by external demands or pressures. External motivation involves "the performance of an activity in order to attain some separable outcome…" (Ryan and Deci, 2000, p. 71). Environmental contingencies that guide and determine ('motivate') behavior are omnipresent in the natural and social environment. Cold weather 'motivates' people to wear warmer clothing and to put on gloves, caps, and coats. Accolades, diplomas, and access to fulfilling future employment 'motivate' students to perform well in classes. And, icy roadways 'motivate' people to drive automobiles at a lower-than-typical speed for safety reasons. These environmental contingencies are present throughout life, and are an important part of social and emotional development through adulthood. As such, these contingencies create a learning history for the person that motivates behavior far into the future, and which can be considered an 'internal' state, one that arises from 'external' forces.

While internal motivation typically is regarded more positively by the general public and health professionals, and is considered more sustainable, in reality, almost all motivation is a combination of internal and external factors. Nevertheless,

of the many theories and viewpoints of motivation, some of the contemporary ones emphasize that it is closely tied to internal forces. Internal motivation requires a supportive environment to flourish, and external motivation that is too demanding or harsh is unlikely to maintain behavior once the external contingencies are absent. While some external rewards, pressure, and imposed restrictions (e.g., deadlines) can be detrimental to internal motivation (Ryan and Deci, 2000), when properly applied, external contingencies allow a person to learn to complete a task, and then to encounter the (hopefully) positive contingencies that will follow, and maintain the task-completing behavior in the future.

In clinical interventions, it often is the case that external contingencies initially are arranged to prompt behavior, with a planned transition to internal motivation, once the client or patient encounters reinforcement that occurs naturally in the environment. For example, a sticker chart for cooperative play in the classroom, with a planned reward for 80% adherence in a single day, might first be used with an 6 year-old child. Then, once those cooperative play behaviors are established, the external motivation typically is gradually eliminated, as natural reinforcement (e.g., attention and positive comments from peers and teachers) takes hold and maintains the behavior. Contingencies in the environment can be set up by teachers, clinicians, and staff, but also can be arranged by clients and patients themselves to motivate their own behavior. Change as a result of psychotherapy, for example, involves a huge challenge, in that, typically, it involves one hour per week (of 168) of clinician contact with the client or patient. Clinicians work with clients and patients to help them become aware of internal motives (e.g., desire to be physically attractive to others on the part of an adolescent female with anorexia) but often also to arrange their own environment to include contingencies that provide external motivation for behavior change (e.g., completing daily food intake charts that are brought to clinic appointments, with social praise from clinicians for positive steps). At some future point, the external motivation, which may appear artificial at times, can be reduced and then eliminated because natural contingencies arise to maintain the behavior (e.g., positive comments from peers about her appearance), and because the person's learning history then is such that it supports continuing healthy behavior.

Transitioning from externally to internally motivations can be advantageous for maintaining behavior over time. Further, the individual may find that internally motivated behaviors are more rewarding and have more meaning, in that the belief may be that "I am doing it" rather than relying on some external force. This transition appears to be a function of regulatory processes and perceived loci of control; it likely is associated with greater self-efficacy (Bandura, 1997). This change is reinforcing to the individual, who then perceives that he or she has control in regard to achieving goals and finds pleasure in working toward them (Ryan and Deci, 2000).

Ultimately, clinicians typically hope to have patients or clients encounter natural contingencies that promote and maintain positive, healthy behaviors. An inherent challenge, however, is that the reinforcement associated with many health behaviors is distant in time (e.g., weight loss after dietary changes). Understanding how individuals can work toward long-term health goals can be informed by an understanding of choice behavior.

Self-Control

Motivation is, in some respects, an issue of self-control and impulsivity. It is easy (and commonplace) to choose an immediate small reward (e.g., eating a candy bar) over a delayed large reward (e.g., social and cardiovascular benefits of maintaining a healthy weight). Delay discounting is a concept that speaks to the issue of valuing immediate rewards more than those for which one must wait (Rachlin and Green, 1972) and has become an important issue across the fields of psychology, economics, and health. The magnitude of delayed, large rewards and how quickly their value becomes discounted impacts whether one might choose rewards that are more immediate or more distal in time. The amount of delay for more immediate and more distal awards also impacts choice. Self-control prompts choosing the delayed larger rewards over more immediate but smaller ones, while impulsivity promotes selecting more immediate rewards. As an example, delay discounting, along with time perspective, helps to predict various health behaviors in college students, such as use of frequency of exercising, and eating breakfast, as well as tobacco, alcohol, and drug use (Daugherty and Brase, 2010). Clinicians can help clients and patients become aware of contingencies that operate distantly in time, and thereby cultivate motivation to make healthier choices.

Can Rewards Undermine Internal Motivation?

Rewards clearly can motivate behavior in the short term, but what about the long term, and maintaining behavior over time? Concerns have been raised about the use of incentives to motivate behavior (Kohn, 1993), suggesting in part that rewards undermine creativity and intrinsic motivation to achieve well. Kohn (1993) specifically attacks the operant system of behaviorism, involving application of rewards and punishments contingent upon behavior. Reitman (1998) responded to these criticisms, arguing that behavioral approaches are solution-focused and based on empiricism. Citing the work of B. F. Skinner, Reitman emphasizes the need to include naturally reinforcing contingencies in schools (and elsewhere). Moreover, both so-called artificial (or better, constructed) incentives and those which are naturally occurring help produce a learning history in people that allows the continuation, and even evolution, of healthy behaviors. The use of incentives clearly is entrenched in modern life. It is difficult to imagine a functioning utopian society in which everyone regularly goes to work without the promise of, and attendant motivation associated with, paychecks. What may be most important is there is an effort to ensure that motivation for positive, healthy behaviors in clinical settings is learned by clients and patients, becomes internalized, and is responsive to reinforcement that occurs naturally in their environment. Similarly, paychecks may become less important for financially secure people who find other satisfactions in their work.

Motivation, Behavior Change, and Maintenance of Behavior

Changing patterns of behavior can require a significant amount of time, effort, and motivation (Noordman et al., 2012). Maintaining lifelong health behaviors (e.g., wearing seat belts, regular exercising, toothbrushing, and flossing), however, is a significant challenge as well. For some clients and patients, and for some behaviors, change can be relatively easy. Typically, however, changing established behavior patterns, some of which may be lifelong, is extremely challenging and fraught with relapse, frustration, and discouragement. Rather than clinicians incorporating persuasive tactics, Hayhurst (2013) argues that conversation, cooperation, and understanding in clinical settings are more likely to influence the motivation of clients and patients. Some general motivational strategies include removing barriers to treatment and progress, providing choices, decreasing the desirability of maladaptive choices, expressing empathy, providing feedback, establishing goals, and providing assistance toward achieving the specified goals (Madukwe, 2013). Various contemporary approaches provide models of, and strategies for, change and sustainability of health behaviors.

Stages of Change

The process of change by clients, patients, and others has been conceptualized as a cyclical process in which people move back and forth between stages based on their readiness for, and engagement in, the change process (Prochaska and DiClemente, 1984). Moving from no awareness of a need to change, across this sequence people recognize the need to change and begin to understand and then implement what is required to achieve change (Madukwe, 2013). The main stages of change are pre-contemplation, contemplation, preparation, and action, although maintenance and relapse can be considered stages as well (Prochaska and DiClemente, 1984).

The pre-contemplation stage is the point at which people do not recognize the presence of maladaptive behavioral patterns and, thus, have no intention to change their behavior. The contemplation stage occurs when one begins to consider the possibility of needing to change maladaptive behaviors. Deciding how to change, and anticipating and planning to change one's behavior, are the main features of the preparation stage. The action stage occurs when the person initiates the change process. During this stage, the patient or client is best served by significant assistance, support, and encouragement from the clinician. The transition from the action to the maintenance stage occurs when new, more adaptive behaviors are established and the threat of older, more maladaptive behaviors is less frequent. Relapse involves the return to previous, maladaptive behavioral patterns, and occurs predominantly in the maintenance stage (Madukwe, 2013). Relapse prevention is a strategy that is used proactively to provide motivation for patients and clients, even in the face of a lapse involving the previously changed problem behavior (e.g., consuming alcohol). This Transtheoretical Model of Stages of Change helps clinicians identify clients' and patients' awareness of and readiness to change, and so serves as helpful 'barometer' for individually assessing motivation during clinical treatment. An approach, however, that can be used, regardless of a client's or patient's stage of readiness to change, is Motivational Interviewing (MI).

Motivational Interviewing

In regard to assessing, eliciting, and promoting motivation among clients and patients in health care settings, there is no more powerful or accepted contemporary method than MI (Miller and Rollnick, 2013). In health care, there is a zeitgeist of activity associated with MI research and clinical application to numerous health and social problems. MI originally was developed to treat substance abuse and addiction, but thereafter was applied to a diverse array of health-related maladaptive behaviors (e.g., anxiety, depression, suicidality, eating disorders, medication compliance, dieting, exercise, smoking cessation) (Arkowitz and Westra, 2009; Martins and McNeil, 2009). MI is a client- or patient-centered approach for enhancing motivation to change (Arkowitz and Westra, 2009).

A theoretical model of MI emphasizes two active, dual roles of a relational aspect involving clinician empathy and 'MI spirit' as well as MI technical skills on the part of the clinician that are designed to evoke and then reinforce change talk by the patient or client (Miller and Rose, 2009). With the acknowledgment that change can be extremely difficult to implement, and perhaps even more challenging to maintain over time, a core feature of MI is great respect for clients and patients as people, acknowledging their autonomous right to make their own decisions. MI spirit consists of a collaborative, working relationship that attempts to strengthen the patient's autonomy and enhance various ideas for successful change (Arkowitz and Westra, 2009). Miller and Rollnick (2013) add principles of acceptance, compassion, and evocation to the concept of MI spirit. Acceptance is conceptualized as the clinician demonstrating "absolute worth, accurate empathy, autonomy support, and affirmation" toward the patient or client during treatment (Miller and Rollnick, 2013, p. 24), while embodying a compassionate, caring approach. Evocation of ambivalence and motivation highlights the belief that the ability to change one's behavior resides within the person, and not in the clinician or with his or her persuasive powers.

Humanistic in spirit, MI also is behavioral in nature, based on the principles of selective (verbal) reinforcement. Using MI, clinicians selectively reinforce which aspects of the client or patient's communication about change to highlight and draw attention to during sessions. With a positive impact on treatment adherence generally (Chanut et al., 2005), MI can be used to support both psychosocial and pharmacotherapeutic approaches in health care. MI has been shown to be applicable across cultural groups and socioeconomic domains. As an approach, it seems to be most effective with individuals who initiate treatment with lower levels of motivation (Chanut et al., 2005).

What Is MI?

Currently defined as "… a collaborative, goal-oriented style of communication with particular attention to the language of

change, MI is designed to strengthen personal motivation for and commitment to a specific goal by eliciting and exploring the person's own reasons for change within an atmosphere of acceptance and compassion" (Miller and Rollnick, 2013, p. 29). Primarily a guiding style (with following and directing elements as well), MI involves enhancing a client or patient's motivation to change by exploring, highlighting, and ultimately helping them to resolve *ambivalence* about change (Arkowitz and Westra, 2009).

MI can be used as a stand-alone therapy in and of itself, or incorporated before or during another form of treatment (Chanut et al., 2005). More than a discrete therapy alone, MI also can guide a conversational style, as evidenced by the current MI definition by its developers (i.e., Miller and Rollnick). This MI conversational style then can be used in very brief encounters with patients and clients, such as in primary medical care.

MI involves the clinician stepping out of the typical 'expert' role and addresses the 'righting reflex' on the part of the clinician, which involves the tendency to 'jump in' and help, to direct the client or patient and to prescribe needed courses of action, and to immediately provide (or even demand?) solutions that are based on their expertise (Chanut et al., 2005, p. 718). Instead, in MI, the client or patient is viewed as the agent of change and is the one who must come to believe that he or she is capable of changing (Arkowitz and Westra, 2009). MI includes the following technical elements: resolving ambivalence, avoiding provoking resistance to change, eliciting self-motivational statements, enhancing overall motivation and commitment to change, and assisting the client or patient through the process of change (Madukwe, 2013). These elements of MI now are conceptualized within four processes: engaging, focusing, evoking, and planning (Miller and Rollnick, 2013).

Engaging

The concept of engaging the patient or client can be defined as "the process of establishing a helpful connection and working relationship" (Miller and Rollnick, 2013, p. 26). In addition to building a meaningful connection with the client or patient, the process of engaging increases the likelihood that the client or patient will continue to attend future treatment sessions and ultimately change their maladaptive behavior(s) (Miller and Rollnick, 2013). In engaging clientele from an MI perspective, clinicians are leery of trying to focus on a goal of treatment too early in the treatment process or trying to assert too much control in the conversational flow, as these behaviors may lead to detrimental effects on the therapeutic alliance and treatment outcomes. The building and maintenance of a therapeutic alliance is critical for the success of MI in eliciting ambivalence, change talk, and ultimately, changes in behavior (Miller and Rollnick, 2013). The engaging process involves, to a great degree, active listening by the clinician, as a precursor to guiding the development of motivation to change.

Focusing

A "process by which you develop and maintain a specific direction in the conversation about change" (Miller and Rollnick, 2013, p. 27), focusing also can be viewed as "agenda-mapping," in which the clinician and the patient or client determine which direction to move forward in treatment (Miller and Rollnick, 2013, p. 110). This agenda mapping often is utilized when changing direction in treatment, going around an impasse, discussing sensitive topics, clarifying the relative roles of clinician and patient or client, and the patient or client coming to resolution or acceptance in regard to some facet of life (Miller and Rollnick, 2013). The focusing process in MI provides a direction for treatment to move as well as a clarification of the goals, aspirations, or outcomes that are desired by the client or patient.

Considerations in utilizing MI are the patient or client, the setting, and the clinician's skills. The focus on the patient or client pertains to the concerns and problems he or she presents in treatment. The focus on setting pertains to treatment provided for a particular issue in a specific facility (e.g., driving while intoxicated, urgent medical care, psychotherapy for depression in a mental health clinic), which can be limiting in relation to magnitude of effect for both the clinician and the patient or client (Miller and Rollnick, 2013). The focus associated with the clinician's expertise is based upon both MI spirit and technical skills. There can be ethical dilemmas that arise when foci and overall goals of treatment differ between the clinician and the client or patient. The main goal should be to blend the clinician's and the client or patient's beliefs about the focus of treatment to enhance the possibility of achieving positive outcomes from treatment (Miller and Rollnick, 2013).

Focusing involves a continuum of communication styles (i.e., directing, following, and guiding) (Miller and Rollnick, 2013). The directing approach involves the clinician controlling the flow of the session as well as focus of treatment. The following or nondirective approach allows the patient or client's concerns and desires to control the focus of treatment. The guiding approach is a combination of both the directing and following approaches, and involves both the clinician and the patient or client working in collaboration and sharing common foci for treatment.

In relation to the guiding approach, Miller and Rollnick (2013) recommend an 'elicit-provide-elicit' sequence in which the provision of information is alternated with eliciting information from the patient regarding treatment and behavior change. This type of sequencing limits the degree to which the clinician is controlling the session and attempts to increase the frequency with which the patient or client speaks in session. Relative 'talk time' is a crucial component of MI consistency; if the clinician talks most of the time, then the client or patient has little opportunity to talk about ambivalence regarding change, and the emphasis in the clinical encounter is on the practitioner and his or her opinions about what is best for the patient or client. In many instances, the clinician should avoid monologues and speak less (in total time) than the client or patient, although that relative percentage may vary across a treatment session, as earlier in an encounter, the clinician may do more reflective listening, and later may provide more information.

Evoking

This process is "…eliciting the client's own motivations for change…" (Miller and Rollnick, 2013, p. 28), and

incorporates the preliminary stages of planning, in which the client or patient generates ideas, thoughts, and feelings about changing his or her behavior, as well as the process by which change can occur. During this time, clinicians listen for and highlight the all-important 'change talk.' Reinforcing change talk (in part, by reflective listening) increases the likelihood of the patient or client ultimately engaging in actual behavior change. During the process of evoking the patient to change his or her behavior, clinicians resist the 'righting reflex' and uninvited advice-giving, and assist patients and clients in finding their own personal motivation(s) for change. Rather than attempting to instill motivation by persuasion, fear tactics, or other measures, clinicians 'nurture' the patient or client's own motivation and commitment to change through the use of technical MI skills. The evoking process of MI is critical for resolving the patient or client's ambivalence regarding behavior change. The process of guiding the client or patient and evoking his or her own motivation for behavior change is one of the defining features of MI (Miller and Rollnick, 2013).

Planning

The process of planning with the client or patient incorporates confirmation of the commitment to change as well as establishing a specific plan (Miller and Rollnick, 2013). Clinical judgment is required in regard to the appropriate time to move from evoking to planning. In planning, there typically first is recapitulation (a summary of the patient or client's change talk), followed by a key question (a simple, direct question regarding thoughts and feelings pertaining to behavior change at the present time). The planning stage predominately focuses on the specific aspects of behavior change and information seeking and provision. This approach is very different from the clinician asking if the patient or client believes he or she should change his or her behavior(s). Creating the plan for change is only the beginning part of the planning phase of MI; it is an ongoing process that requires re-evaluation as the person begins to engage in behavior change (Miller and Rollnick, 2013).

Foundations and Principles of MI

MI spirit is one of two major crucial components of successful treatment in this approach (Miller and Rose, 2009). More specifically, a basic principle is that clinicians using an MI approach suspend judgment, actively listen during treatment, and communicate respect and acceptance to the patient or client. Madukwe (2013) argues that expressing empathy is one of the most important aspects of MI. Further, it is foundational that a collaborative relationship exists between the clinician and the patient or client, one that is supportive, and that creates a safe and open environment for patients or clients to explore their own ambivalence, and to communicate their concerns and feelings relating to change.

A central tenant of MI is that ambivalence regarding altering one's behavior(s) is a normal part of the change process (Miller and Rollnick, 2013). If a clinician attempts to force the idea of change on the patient or client, it is likely that the he or she may resist or even stop attending treatment (Miller and Rollnick, 2013). Some methods for reducing resistance to and ambivalence about change and 'rolling with resistance' include the use of reflections, shifting the focus of the conversation, and reframing (Madukwe, 2013). Patients and clients often express or display hesitancy in regard to the idea of change unless they have confidence in their ability to successfully achieve change (Miller and Rollnick, 2013). Therefore, another principle is that clinicians support the patient or client's self-efficacy (Bandura, 1997). The clinician can assist in increasing confidence through affirmations recognizing and highlighting the person's strengths.

Some argue that, in addition to self-efficacy, the creation of discrepancy between one's values and current behavior is needed for behavior change (Grave et al., 2013; Miller and Rollnick, 2013). Establishing cognitive dissonance between one's values and current behavior patterns is likely to lead to motivation to change one's behaviors to be more in line with one's values. Seeking clarification and exploring the patient or client's concerns relating to the discrepancy between values and current behaviors is pivotal to behavior change.

In MI, it is necessary to draw attention to and reinforce 'change talk.' This form of communication occurs when the patient or client begins to discuss the possibility of changing maladaptive behaviors. During this process, a decisional balance, in which the client or patient examines the pros and cons of change, can be created and explored (Madukwe, 2013).

Overall, the power to decide to change resides with the client or patient (Madukwe, 2013). Therefore, a goal of MI is to assist patients and clients in resolving ambivalence and committing to changing maladaptive behaviors. In addition to MI spirit and underlying principles and foundations, there are specific skills which constitute the second major component of MI approaches.

MI Technical or Micro-Skills

The major technical therapeutic skills of MI are open-ended questions, affirmations, reflective listening, and summarizing. These technical skills area a primary aspect of the other major component of MI, in addition to MI spirit (Miller and Rose, 2009). Open-ended questions keep the conversation moving forward by encouraging the patient or client to freely share after a prompt that is less structured than, say, a closed-ended question. Open-ended questions also communicate that the clinician is listening to the client or patient's concerns and feelings associated with the pros and cons of change.

Affirmations communicate appreciation, respect, understanding, and validation of the patient or client's concerns pertaining to change. Moreover, affirmations assist in building rapport, the therapeutic alliance, and improving the client or patient's self-efficacy and confidence in the ability to successfully change.

Reflective listening involves clinician attempts to discern and state what patients or clients are communicating, thus making the meaning of their statements more salient (Miller and Rollnick, 2013). This skill also can be utilized to further evaluate the person's statements about change and engage him or her in greater thoughtful consideration of behavior change. Clinicians can incorporate paraphrasing, rephasing, as well as a variety of different types of other reflections (e.g., double-sided reflections) to clarify the meaning of the client or

patient's feelings and concerns. Feeling reflections highlight the client or patient's emotions, whereas simple reflections are a direct restatement of the client or patient's verbalizations. Complex reflections involve a combination of drawing attention to feelings and/or thoughts. Overall, reflective listening assists in identifying ambivalence toward change and resolving it, as well as increasing motivation to change behavior (Madukwe, 2013).

Summarizing statements are organizing tools that can link conversational subjects together, as well as emphasize change talk and reduce ambivalence toward change (Chanut et al., 2005). These summaries can be viewed as a 'long reflection' (Miller and Rollnick, 2013, p. 34). Summarizing statements indicate that the clinician is listening carefully to the person's concerns; they increase the client or patient's degree of cognitive processing pertaining to change. Miller and Rollnick (2013) suggest that summarizing statements also can be incorporated as a transitional tool to move from one MI process to another.

Limitations of MI

Treatment utilizing MI can take extra time with clients and patients (although there are very brief interventions, and an MI conversational style), and require additional resources and training (Chanut et al., 2005). Specifically, clinicians must receive training and supervised practice in MI to be able to recognize cues in session as well as to apply the skills effectively (Chatnut et al., 2005). Depending on the context, clinicians may struggle with "stepping out of the expert role" as suggested by MI, based on their professional instincts and training (Chanut et al., 2005). Regarded as 'simple but not easy' (Miller and Rollnick, 2013, p. 303), MI concepts and skills are readily understood but much harder to implement, particularly in 'real time' during interactions with patients or clients. MI may be unnecessary, or even counterproductive, with clients and patients who are ready to change, or already engaging in change behaviors, as it can seem cumbersome, indirect, and time-inefficient. Finally, certain MI concepts, particularly MI spirit, are diffuse constructs and difficult to define.

Summary and Conclusion

Motivation on the part of the client or patient typically is critical for successful clinical interventions, whether they be taking medicines as prescribed, doing homework assignments to practice relaxation, or assessing one's glucose levels at home. Motivation can be influenced by both internal and external factors, and importantly by a person's learning history. Both intrinsic (internal) and extrinsic (external) factors have been shown to have powerful influences on motivation to change. In addition, motivation has been found to be related to choice behavior, and whether a person engages in self-control or is impulsive in decision making. The Stages of Change theory highlights the importance of motivation in terms of behavior change, emphasizing varying levels of preparedness to change.

Motivational interviewing (MI) highlights the importance of encouraging behavior change in clients or patients during treatment. MI is composed of four processes in which the clinician works collaboratively with the patient or client in developing and maintaining motivation in treatment to ultimately decide upon the possibility of behavior change. MI is composed of many specific clinical methods that encompass a person-centered approach to treatment. Originally designed for substance-use disorders, MI has evolved over the years and has been found to be useful as a stand-alone treatment, as a component added to other treatment modalities, or as guide to a conversational style. The overall success and popularity of MI continues to grow, highlighting its promising nature as a treatment for a variety of problem behaviors.

In summary, motivation is a complex and multifaceted drive state that is affected by a person's learning history and environmental contingencies. Specifically, external conditions and prior learning history affect an individual's motivation during clinical interventions and elsewhere. Motivation is an important foundation for behavior change and maintenance of healthy behaviors.

There is a shared responsibility for both the clinician and the client or patient for motivation to change or to maintain health behaviors. Clinicians cannot ethically eschew their duty to help inculcate motivation, or to blame the patient or client for 'lack of motivation.' Similarly, clients and patients cannot rely solely on powerful others (e.g., health care professionals) to provide motivation, and must act in ways that allow ambivalence about change to be identified and resolved, and to structure environments so that change can be supported. Illusive to find, and challenging to maintain over time, motivation is key to positive outcomes in a broad array of clinical interventions in behavioral and general health care.

See also: Avoidance and Approach Motivation: A Brief History; Expectancy-Value-Cost Model of Motivation; Health Self-Regulation, Motivational and Volitional Aspects of; Intrinsic Motivation: A Cultural Perspective; Motivation and Actions, Psychology of; Motivation, Learning, and Instruction; Motivation: History of the Concept; Motivational Development: Cross-Cultural Perspectives; Temperament and Motivation.

Bibliography

Arkowitz, H., Westra, H.A., 2009. Introduction to the special series on motivational interviewing and psychotherapy. Journal of Clinical Psychology 65, 1149–1155. http://dx.doi.org/10.1002/jclp.20640.

Bandura, A., 1997. Self-Efficacy: The Exercise of Control. Freeman, New York.

Catania, A.C., 1991. Glossary. In: Iversen, I.H., Lattal, K.A. (Eds.), Experimental Analysis of Behavior. Elsevier, New York, pp. G1–G44.

Chanut, F., Brown, T.G., Dongier, M., 2005. Motivational interviewing and clinical psychiatry. Canadian Journal of Psychiatry 50, 715–721.

Daugherty, J.R., Base, G.L., 2010. Taking time to be healthy: predicting health behaviors with delay discounting and time perspective. Personality and Individual Differences 48, 202–207.

Deci, E.L., Ryan, R.M., 1985. Intrinsic Motivation and Self-Determination in Human Behavior. Plenum, New York.

Delinsky, S.S., Thomas, J.J., Germain, S.A., Ellison, C.K., Weigel, T.J., Levendusky, P.G., Becker, A.E., 2011. Motivation to change among residential treatment patients with an eating disorder: assessment of the multidimensionality of motivation and its relation to treatment outcome. International Journal of Eating Disorders 44, 340–348.

Grave, R.D., Centis, E., Marzocchi, R., El Ghoch, M., Marchesini, G., 2013. Major factors for, facilitating change in behavioral strategies to reduce obesity. Psychology Research and Behavior Management 6, 101–110. http://dx.doi.org/10.2147/PRBM.S40460.

Hayhurst, C., 2013. A matter of motivation. PT In Motion 5, 18–24.
Kohn, A., 1993. Punished by Rewards: The Trouble with Gold Stars, Incentive Plans, A's, Praise, and Other, Bribes. Houghton Mifflin, Boston.
Madukwe, A.U., 2013. Motivational interview: evidence based strategy in the treatment of alcohol and, drug addiction. IFE Psychologia 21, 174–196.
MacCorquodale, K., Meehl, P.E., 1948. On a distinction between hypothetical constructs and, intervening variables. Psychological Review 55, 97–107.
Martins, R.K., McNeil, D.W., 2009. Review of motivational interviewing in promoting health behaviors. Clinical Psychology Review 29, 283–293.
Miller, W.R., Rollnick, S., 2013. Motivational Interviewing: Helping People Change, third ed. Guilford, New York.
Miller, W.R., Rose, G.S., 2009. Toward a theory of motivational interviewing. American Psychologist 64, 527–537. http://dx.doi.org/10.1037/a0016830.
Noordman, J., van der Weijden, T., van Dulmen, S., 2012. Communication-related behavior change, techniques used in face-to-face lifestyle interventions in primary care: a systematic review of the literature. Patient Education & Counseling 89, 227–244. http://dx.doi.org/10.1016/j.pec.2012.07.006.
Prochaska, J.O., DiClemente, C.C., 1984. The Transtheoretical Approach: Crossing Traditional, Boundaries of Therapy. Dow Jones-Irwin, Homewood, IL.
Rachlin, H., Green, L., 1972. Commitment, choice and self-control. Journal of the Experimental Analysis of Behavior 17, 15–22. http://dx.doi.org/10.1901/jeab.1972.17-15.
Reitman, D., 1998. Punished by misunderstanding: a critical evaluation of Kohn's punished by rewards and its implications for behavioral interventions with children. Behavior Analyst 21, 143–158.
Ryan, R.M., Deci, E.L., 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist 55, 68–78.
VandenBos, G.R., American Psychological Association, 2007. APA Dictionary of Psychology. American Psychological Association, Washington, DC.

Relevant Websites

http://www.motivationalinterviewing.org/ – Motivational Interviewing (MI) Network of Trainers (MINT).
http://www.prochange.com/transtheoretical-model-of-behavior-change – Transtheoretical/Stages of Change Model.

Motivation in Youth Sport and Physical Activity: Developmental Perspectives

Maureen R Weiss and Alison C Phillips, School of Kinesiology, University of Minnesota, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.R. Weiss, volume 6, pp. 3620–3624, © 2001, Elsevier Ltd.

Abstract

Millions of children and youth participate in sports and physical activities, such as organized after-school programs, school physical education, and community classes. We review the developmental sport psychology literature on two major determinants of motivational orientations and behaviors in physical activity settings – perceived competence and social influence by parents, coaches, and peers. Studies reveal robust processes and mechanisms of influence that suggest interventions for optimizing self-determined motivation and participation in physical activity. A physically active lifestyle, in turn, strongly contributes to positive developmental and health-related outcomes among youth.

Introduction

Millions of children and adolescents participate in sports and physical activities worldwide (Weiss and Williams, 2004). The voluntary nature and rates of participation strongly suggest that physical activity contexts are motivating, enjoyable, and viewed as a source for promoting adaptive developmental outcomes (Weiss et al., 2012b). We use *physical activity* as an overarching term to include structured settings such as school and community sport programs, school physical education, and movement-oriented classes (e.g., dance, motor development), as well as unstructured settings such as free play, recreational activities, and active transportation (e.g., walking or riding a bike to school). These varying activity contexts accommodate age-related preferences, range of ability levels, desired intensities of involvement, and preferred modes of movement experiences. The majority of studies investigating the developmental significance of sport and physical activity participation have focused on structured settings, which include scheduled practices or lessons, adult instructors, and instrumental goals such as skill improvement, physical fitness, and health and well-being.

Physical activity motivation can be conceptualized in cognitive and behavioral terms (Weiss and Amorose, 2008; Weiss and Williams, 2004). *Motivational orientations* connote an individual's reasons for initiating and continuing activity involvement. From this cognitive stance, motivation can be conceived as *because* answers to *why* questions: (1) *Why* do some youth choose to participate in sports and physical activities while others do not? (2) *Why* do some youth try hard and persevere in learning and performing, while others give up easily or drop out? and (3) *Why* do some youth adopt a 'love of the game' attitude toward involvement, whereas other youth participate primarily for extrinsic reasons? *Motivational behaviors* refer to initiating involvement, continuing participation over time, demonstrating sustained effort while participating, and engaging in degrees of activity frequency, intensity, and duration. Such adaptive physical activity behaviors contribute to a healthy lifestyle and overall well-being. The literature in developmental sport psychology identifies determinants of motivational orientations and behaviors that, in turn, optimize social, psychological, and physical development (Weiss et al., 2012a,b).

We review the literature on physical activity motivation among children and youth with a focus on processes and mechanisms of influence from a developmental perspective. We first identify and briefly summarize theoretical frameworks that have guided empirical studies of youth physical activity motivation. Second, we review and synthesize robust findings on perceptions of physical competence as determinants and outcomes of participation motivation. Third, we summarize consistent evidence-based findings on social relationships and interactions as determinants of youth physical activity motivation, with attention to mechanisms of influence by parents, coaches, and peers. Finally, we identify directions for future research that can inform researchers and practitioners of how to optimize sport and physical activity as contexts for deriving developmental and health benefits.

Theoretical Frameworks Relevant to the Study of Physical Activity Motivation

Early descriptive studies of why children and youth participate in sport and physical activity uncovered three major categories of motivation: (1) to develop or demonstrate physical competence, (2) to attain social acceptance or approval, and (3) to experience fun and enjoyment (Weiss and Amorose, 2008; Weiss and Williams, 2004). These ingredients of participation motivation led researchers to adopt conceptual approaches that highlight *perceptions of competence*, *social influence*, and *positive affect* as determinants of motivational orientations and behaviors. Over the past 30+ years, theories conceived within developmental, social, and educational psychology have been used to guide studies of motivation in the physical domain – notably competence motivation theory, global model of self-worth, expectancy-value theory, achievement goal theory, self-determination theory, ecological systems, and positive youth development (Horn, 2004a; Weiss et al., 2012a,b).

Some theories embrace an explicit developmental approach (competence motivation theory, global model of self-worth, expectancy-value theory, achievement goal theory) and identify age-related cognitive and physical attributes that explain variations in motivational orientations and behaviors (Horn, 2004a). Self-determination theory has been a popular

perspective in recent years, with its broad appeal and applicability of basic psychological needs for competence, autonomy, and relatedness as antecedents of motivational orientations and behaviors (Stuntz and Weiss, 2010). Ecological systems and positive youth development approaches reflect more recent developmentally oriented frameworks and show great promise in understanding determinants of youth physical activity motivation, participation behaviors, and health and well-being (Horn, 2004a; Weiss et al., 2012b).

Theory-driven studies show strong support for the role of *self-perceptions* (self-esteem, perceived competence, self-efficacy) and *significant adults and peers* on motivational orientations and behaviors in sport and physical activity (Horn, 2004b; Weiss et al., 2012a,b). Because the need to demonstrate physical competence and attain social approval and acceptance have pervaded the youth participation motivation literature, it is perhaps not surprising that *perceived competence* has emerged as the most important individual difference factor, and *parent, coach, and peer relationships* have been the most important social-contextual factors related to physical activity motivation. Moreover, developmental variations in perceived competence and relative influence of significant adults and peers reveal important theoretical and practical guidelines for maximizing motivational orientations and behaviors. In the following sections, we synthesize and consolidate the literature across conceptual frameworks to feature evidence-based findings on perceived competence and social influence as determinants of physical activity motivation among youth.

Perceived Competence and Physical Activity Motivation

Perceived competence is a central construct within all theories mentioned earlier that are relevant to the study of motivation in the physical domain. *Perceived physical competence* is key to understanding variations in physical activity motivation and refers to beliefs, judgments, and feelings about one's physical abilities and competencies in general (e.g., athletic, physically fit) or in a particular domain (e.g., basketball, tennis, golf; Horn, 2004b; Weiss et al., 2008). Youths' *perceptions* of their ability to perform physical tasks are more explanatory of motivational orientations and behaviors than are *actual* skill assessments. Thus, studies have focused on *how* youth come to judge their physical ability as lower or higher and *why* self-evaluations may be accurate or inaccurate relative to objective ratings or scores.

Level and accuracy of perceived physical competence vary developmentally as a function of cognitive maturity and social–environmental changes that occur with age (Horn, 2004b; Weiss et al., 2008). During early childhood (3–6 years), children are very optimistic about, and often overestimate, their abilities. This may be due to the concrete nature of their evaluations, an undifferentiated conception of ability, and the sources of information used to judge physical competencies. During this developmental period, sports and physical activities are typically structured in ways that emphasize trying to do one's best, mastering developmentally appropriate motor skills, having fun, and learning to cooperate and play well with others. These contextual features probably explain why young children primarily use simple task mastery, effort, and parental feedback as information sources to gauge their ability (Horn, 2004b).

During middle and late childhood (ages 7–12 years), level of perceived competence declines and accuracy increases (Horn, 2004b; Weiss et al., 2008). Children no longer overestimate their abilities relative to objective ratings but, rather, subjective and objective ratings are more likely to correspond. Changes in level and accuracy of perceived competence are attributed to increased cognitive maturity (e.g., ability to make self-judgments of ability between and within achievement domains), a shift from more task- to ego-involving motivational climates in organized sport, and increased use of peer comparison, peer evaluation, and coaches' performance feedback as information sources to appraise physical ability. These sources make sense, given the highly public nature of structured sports and physical activities, in which social comparison and evaluation are key contributors to reflected and self-appraisals of ability.

Among adolescents (ages 13–18 years), level of perceived competence may decline, remain stable, or increase depending on the frame of reference for self-evaluating physical ability (e.g., move up in competitive level or not make the team; compare to teammates or all athletes in the school district; Horn, 2004a; Weiss et al., 2008). Variations may also occur because adolescents differentiate and integrate a greater number of information sources to assess their physical abilities. Compared to younger peers, adolescents use self-referenced information more frequently (e.g., skill improvement, goal achievement, speed of learning skills), continue to use social comparison and evaluation by adults and peers, and are capable of judging their abilities both *within* a domain (e.g., softball vs martial arts) and *between* achievement domains (e.g., sports vs music). In sum, developmental differences in level, accuracy, and sources of perceived competence explain variations in physical activity motivation.

A robust finding in the youth sport literature is the strong relationship between perceptions of physical competence and cognitive (e.g., perceived control, performance attributions), affective (enjoyment, anxiety), and behavioral (e.g., effort, participation) variables directly related to physical activity motivation (Kipp and Weiss, 2013; Weiss and Amorose, 2008; Weiss et al., 2012a). Youth who favorably evaluate their physical abilities are more likely to ascribe successful performance to controllable factors, enjoy their experiences, persevere in learning new skills, and be more physically active. These attributes, in turn, increase the likelihood of sustaining involvement in physical activity over time and acquiring social, psychological, and behavioral competencies and health benefits (Weiss et al., 2012b).

Level of perceived physical competence is also related to self-determined motivation such as participating for the inherent joy of movement, for confirming one's identity as an athlete or being physically fit, and for attaining health benefits from involvement (Stuntz and Weiss, 2010; Weiss et al., 2012a). Self-determined motivational orientations are characterized by being freely chosen and self-regulated, in contrast to controlling motivational orientations that include participating out of feelings of guilt or pressure by others. Self-determined motivation, in turn, is strongly related to physical activity level, continued involvement, and psychological well-being.

Studies have also shown a direct link between perceived physical competence and motivated behaviors (Kipp and Weiss, 2013; Weiss et al., 2012a). Participants in youth sport and physical education who favorably appraise their physical abilities also exert greater effort in learning and performing, report greater intention to continue participation, persevere when confronted by challenging skills, and are more physically active. Thus, perceived physical competence is an antecedent of self-determined motivation, which in turn is strongly associated with physical activity behavior and health outcomes (Kipp and Weiss, 2013; Weiss et al., 2012a). In sum, perceived physical competence is an important target of intervention efforts to maintain and enhance physical activity motivation.

Several theory-driven intervention studies have shown that physical activity is a valued context for improving youths' perceptions of physical competence (Kipp and Weiss, 2013; Weiss et al., 2012a,b). Interventions have included coach-training workshops on how to provide effective feedback and reinforcement and shape a climate conducive to enhancing perceived competence, and consequently adaptive physical activity motivation. Other interventions have included team-building challenges, cooperative learning, positive youth development, and health-promotion curricula that target strategies within physical activity settings to promote positive self-perceptions and consequently overall well-being. The key to enhancing perceived competence, self-determined orientations, and motivated behaviors reside in the beliefs and behaviors expressed by significant adults and peers as they interact with youth in physical activity settings.

Social Relationships and Physical Activity Motivation

Children and adolescents name many social reasons for and benefits of participating in sports and physical activities, such as making new friends, affirming friendships, feeling part of a group or team, and receiving approval and support by parents, coaches, and teammates (Weiss and Amorose, 2008; Weiss and Williams, 2004). Social interactions and relationships are very important to youth and, indeed, processes and mechanisms of social influence by adults and peers feature prominently in understanding motivational orientations and behaviors in physical activity settings (Stuntz and Weiss, 2010; Weiss et al., 2012a). Theory-driven research on parents and coaches abound, and peer influence is a rapidly emerging area of study in terms of individual processes (close friendship, friendship quality) and group processes (peer group acceptance, leadership, cohesion). We review the knowledge base on parents, coaches, and peers, respectively, relative to youth participants' motivational orientations and behaviors.

Parents

In physical activity contexts, parental influence is especially salient during the childhood years, when children spend more time with and are dependent on parents for resources and opportunities (Fredricks and Eccles, 2004; Horn and Horn, 2007). Parents play a crucial role in their children's motivational orientations and behaviors by expressing beliefs and expectations for performance as well as modeling attitudes and behaviors. Although the nature of the parent–child relationship may change in the transition from childhood to adolescence, parents nevertheless remain important sources of motivation for teenagers and emerging adults.

Fredricks and Eccles (2004) classified mechanisms of parental influence as providers of experience, interpreters of experience, and role models. Parents serve as *providers of experiences* by affording children opportunities to participate, such as signing up for classes or teams, transporting to and from practices and competitions, purchasing equipment and uniforms, and paying club and travel fees. As *interpreters of experience*, parents verbalize beliefs about their child's sport ability, provide feedback that conveys what they think is important (e.g., trying your best vs winning), and express nonverbally their evaluation of performance attempts (e.g., facial expressions, body language). As *role models*, parents influence their children through behaviors such as participating in physical activity, sharing stories of past athletic experiences, volunteering as a parent-coach, and by demonstrating a positive attitude toward and enjoyment of their activity participation. Collectively all three parent mechanisms – providers of experience, interpreters of experience, and role modeling – are essential in shaping youths' motivational orientations and behaviors in sport and physical activity (Fredricks and Eccles, 2004; Horn and Horn, 2007).

In the youth sport literature, these three mechanisms of parental influence overlap and multiple forms of parental beliefs and behaviors have been examined simultaneously (Weiss et al., 2012a). That is, parents who provide opportunities for their child to participate and express confidence about their child's potential to do well in sports are also modeling attitudes and behaviors about the importance of a physically active lifestyle and sport as an achievement domain. The evidence-based research highlights five specific parental behaviors that made a strong impact on children's participation motivation in physical activity: (1) providing social support, (2) conveying pressure to be successful in sports, (3) modeling attitudes and behaviors, (4) expressing beliefs about the child's physical competence, and (5) expressing beliefs about the value of physical activity (Weiss et al., 2012a).

Parental social support can be conceptualized as companionship, informational, instrumental, emotional, affection, and admiration, among others (Horn and Horn, 2007; Weiss et al., 2012a). Types of support vary developmentally, with companionship more prevalent during childhood, and emotional and instrumental support remaining stable over the adolescent years. Robust findings indicate that youth participants who rate mothers and fathers as more supportive of their participation behaviors report greater perceived competence, enjoyment, self-determined motivation, and participation behaviors than those reporting lower parental support (Fredricks and Eccles, 2004; Weiss et al., 2012a). Similarly, parents who are perceived as controlling or pressuring in their actions relative to performance expectations are associated with youth reporting higher anxiety and lower enjoyment, self-determined motivation, and participation behavior. Because youths' *perceptions* of parental behaviors are what drive their motivational orientations and participation behaviors, it is essential that parents are keenly

aware of how their verbal and nonverbal expressions affect children's thoughts, emotions, and behaviors in sport and physical activity.

Modeling is defined as changes in attitudes, emotions, and behaviors as a result of observing others, and *models* are individuals whose verbal and nonverbal behaviors serve as cues for observers' subsequent thoughts, feelings, and behaviors (Weiss et al., 2008). Thus parents' own physical activity behavior, attitudes expressed toward the value of physical activity, past sport experiences, and involvement in their child's experiences by attending games, volunteering, reading about the child's sport, and watching televised sporting events with their child reflect inclusive ways in which parents serve as physical activity role models. When modeling has been broadly defined as the relationship between parental attitudes, emotions, and behaviors with children's physical activity motivation, strong support is attained for parental modeling effects on children's motivational orientations and behaviors (Fredricks and Eccles, 2004; Weiss et al., 2012a).

Parents' beliefs about their child's physical competencies are strongly related to youths' motivational orientations and participation behaviors (Fredricks and Eccles, 2004; Horn and Horn, 2007). Youth who perceive that their parents think they are physically skilled and expect they will perform well in sports report higher perceived competence, self-determined motivation, and physical activity involvement compared to peers who rate their parents lower in competence beliefs. Similarly, parents' subjective value toward physical activity is strongly related to children's ratings of importance, utility, and interest value in physical activity. Children who report that their parents believe being physically active or being skilled in sports is important, useful, and enjoyable also report high value for physical activity and are more physically active than peers who report lower parental value toward sport and physical activity.

The parent–child connection in competence beliefs and value of physical activity is an example where the three mechanisms of parental influence overlap. Children must interpret their parents' competence beliefs about them through observing verbal and nonverbal behaviors, such as frequency and type of feedback, encouraging or discouraging remarks, and demonstrating supportive or unsupportive behaviors (Fredricks and Eccles, 2004; Horn and Horn, 2007). Children's perceptions, therefore, stem from parents serving as providers of experience, interpreters of experience, and role models. Most researchers have interpreted competence belief findings as a parent-to-child pathway – parents express confidence or lack thereof about the child's sport competence and the child internalizes this belief in self-appraisals of ability. However, some research suggests that a bidirectional or reciprocal relationship may exist. That is, children who show interest, aptitude, and confidence in sport may *elicit* in parents an interest in their child's participation, confidence in their child's ability to perform well, and behaviors that motivate the child to continue their physical activity involvement (Weiss et al., 2012a). Thus, just as parents can influence their children's beliefs and behaviors so can children's interests, attributes, and participation choices influence parents' involvement in their child's physical activity pursuits. This reciprocal relationship has the potential to afford both children and their parents of the many benefits of physical activity involvement.

Coaches

Once parents socialize children into sport and physical activity, coaches assume a central motivational role in practice and competitive settings. Informational and evaluative feedback by coaches represent credible sources of information by which children and adolescents alike judge how physically skilled they are, and these perceptions of competence predict motivational orientations and behaviors (Amorose, 2007; Horn, 2008). Coaches remain a powerful source of motivation over the childhood and adolescent years and even beyond for those individuals who continue to climb the competitive ladder.

Over the past 30+ years, numerous studies have focused on the types of coaching behaviors that positively or negatively impact youth physical activity motivation (Amorose, 2007; Weiss et al., 2012a). Based on the varied theories used to study coaches as significant adults on young people's outcomes, Horn (2008) created a model of coaching effectiveness to provide a framework for identifying (1) antecedents and consequences of coaches' behaviors and (2) moderating variables that explain variations in coaching effectiveness. Antecedents of coaching behaviors include the sociocultural context, organizational climate, coaches' personal characteristics, and coaches' expectancies, values, beliefs, and goals. In turn, frequency and types of behaviors coaches exhibit affect youths' perceived competence, motivational orientations, and participation behaviors (effort, persistence, performance). Moderating variables that directly or indirectly influence the coach–youth relationship include participant characteristics (e.g., age, gender, culture), personality (e.g., self-esteem, trait anxiety), and sport type (e.g., individual-team; independent–interdependent).

Horn's (2008) heuristic model characterizes a parsimonious way of integrating theoretical approaches, empirical findings, and methodologies used to study coaching behaviors, and enables a comprehensive review of coach influence and youth physical activity motivation. Theory-driven studies of coaches' influence on motivational outcomes in youth sport have focused on four behavioral mechanisms – feedback patterns, leadership styles, autonomy-supportive versus controlling behaviors, and motivational climate (Amorose, 2007; Horn, 2008; Weiss et al., 2012a). Robust findings for feedback patterns reveal that *contingent* and *appropriate* reinforcement following successful performance attempts, and encouragement combined with technical instruction following skill errors, are associated with child and adolescent participants who report higher self-perceptions, enjoyment, intention to continue playing, and long-term participation and lower anxiety and attrition rates. Coaches' use of punishment following skill errors and nonreinforcement for desirable behaviors are generally related to lower intrinsic motivation and motivated behaviors.

Leadership styles are characterized as decision-making approaches (democratic and autocratic), motivational behaviors (social support and positive feedback), and technical instruction to improve performance (Amorose, 2007; Horn, 2008). Adolescent athletes, who perceive that their coaches use a democratic style of decision-making, are seen as caring about

athletes as individuals, provide frequent and contingent feedback, and competently teach skill techniques, tactics, and strategies report higher perceived competence, enjoyment, intrinsic motivation, and motivated behaviors. Coaches who are rated higher in an autocratic decision-making style are associated with athletes reporting lower motivational outcomes.

Complementary to decision-making styles and stemming from self-determination theory are the constructs of autonomy-supportive and controlling interpersonal styles (Amorose, 2007; Horn, 2008; Weiss et al., 2012a). Autonomy-supportive behaviors include efforts to empower participants with choices within limits, acknowledge input on skills and strategies, and provide participants with opportunities for initiative, self-regulation, and leadership. A controlling interpersonal style consists of using pressure and other coercive means to elicit behaviors, engaging in overly critical feedback as a form of motivation, and emphasizing social comparison as the standard for evaluating performance. Considerable research has shown that youths' perceptions of coaches' autonomy-supportive behaviors are related to higher perceived competence, autonomy, and relatedness and, in turn, optimize self-determined motivation and physical activity behaviors (Amorose, 2007; Weiss et al., 2012a). By contrast, controlling coach behaviors are related to lower satisfaction of psychological needs, self-determined motivation, and rates of continued participation.

Motivational climate refers to the way in which coaches define successful and unsuccessful performance in the sport environment (Amorose, 2007; Horn, 2008). A mastery or task-involving climate is one in which success is defined in self-referenced terms, such as personal effort, skill improvement, and mastering new tasks (i.e., personal bests), while a performance or ego-involving climate is one in which success is defined in norm-referenced terms, such as favorable comparison to others (i.e., being the best). Coaches shape the climate in the way they structure learning experiences, offer participants opportunities for choice and self-regulation, and use criteria for recognizing and evaluating performance. Youth perceptions of coach-created team climates or sport environments show consistent and strong relationships with self-perceptions, affective responses, and motivational orientations and behaviors (Amorose, 2007; Weiss et al., 2012a). Higher mastery climates are associated with greater perceptions of physical competence, enjoyment, self-determined motivation, and participation behaviors, whereas higher perceived performance climates are negatively related to these outcomes.

Coaches and parents represent the most salient sources of adult influence on children's physical activity beliefs and behaviors. Their behaviors convey beliefs about youths' physical ability and potential for future success, which in turn makes an impact on motivational orientations, participation behavior, and developmental outcomes. As youth seek autonomy from adult authority and the closeness and intimacy of friendships, teammates and sport friends contribute above and beyond parents and coaches to physical activity motivation.

Peers

In sport and physical activity, positive peer relationships increase in salience over the childhood and adolescent years. Youth want to feel accepted as part of the team or group and value close personal friendships (Smith, 2007; Weiss and Stuntz, 2004). Social reasons are frequently cited for initiating and continuing participation, and social comparison and evaluation predominate as sources of competence information. These social constructs are intimately related to motivational orientations and behaviors. The context of sport is indeed a social enterprise, as teammates, classmates, and nonsport friends figure prominently in young participants' participation motivation and developmental outcomes (Weiss et al., 2012b).

Peer group acceptance and close friendship are both important constructs in youths' physical activity experiences (Smith, 2007; Weiss and Stuntz, 2004). Feelings of acceptance and being liked by one's teammates are related to higher perceived competence, enjoyment, self-determined motivation, and physical activity behavior. Being good at sports is a consistent index of popularity among boys and girls alike, compared to boy-only findings in earlier research (Weiss and Stuntz, 2004). Social cohesion among teammates is conceptually similar to peer group acceptance and is defined as perceptions of how well team members get along and like each other (Weiss et al., 2012a). Higher social cohesion is related to greater positive affect, intrinsic motivation, and continued participation. Thus, peer group processes are consistently related to physical activity cognitions, emotions, and motivational behaviors among children and adolescents.

Friendship quality or social support is strongly related to youth participants' motivational orientations and behaviors (Smith, 2007; Weiss and Stuntz, 2004). Children and adolescents who report greater companionship, esteem support, loyalty, intimacy, similarity of interests, and emotional support with sport friends experience greater enjoyment, perceived competence, self-determined motivation, and commitment to their activity participation than those who report lower friendship quality (Weiss and Stuntz, 2004). In middle and late childhood, qualities such as companionship, similar interests, and absence of conflicts are more characteristic of sport friendships, whereas adolescents view psychological and behavioral features such as loyalty, intimacy, admiration, and conflict resolution as more prominent (Weiss et al., 1996). By contrast, greater reported conflicts, betrayal, disagreements, and arguments with sport friends are counterproductive to adaptive motivational and health-related outcomes (Weiss et al., 1996).

Observational learning connotes another strong mechanism of peer influence on youth physical activity motivation (Weiss et al., 2012a; Weiss and Stuntz, 2004). Peer models can be similar in age, gender, and other characteristics (sport type, ability level, personality) and thus elicit selective attention by observers to demonstrated attitudes and behaviors. Peer models are effective for informing and motivating others to learn skills, show confidence, persevere, and perform better. For example, in interventions using peer mastery and coping models, observers' self-efficacy, motivation, and physical skills improved from pre- to postintervention, and improvements remained stable in a retention assessment. Several studies have also shown that youth who report having active friends are more likely to be physically active themselves (Weiss et al., 2012a; Weiss and Stuntz, 2004). Thus, friends are important for encouraging adaptive motivational orientations

and behaviors through a variety of processes, such as group relatedness, social support, and modeling.

Studies consistently demonstrate the motivational significance of peer relationships among youth participants in sport and physical activity. Yet, to date, most interventions to promote youth physical activity motivation have targeted curricular and behavioral training for coaches and teachers, and some have employed family-based models as strategies for invoking physical activity behavior change (Kipp and Weiss, 2013; Weiss et al., 2012a). Implementing interventions that focus on peer groups and friendships as agents of behavior change is a promising direction for future research on youth physical activity motivation.

Future Research Directions on Youth Physical Activity Motivation

In the first edition of the International Encyclopedia of Social and Behavioral Sciences, limited research in developmental sport psychology existed for youth physical activity motivation. Now, just over 10 years later, the literature abounds with theory-driven studies of determinants and outcomes of physical activity motivational orientations and behaviors. Still, much more remains to be known about how to encourage and reinforce adaptive forms of motivation to optimize holistic health and well-being among youth.

Our first recommendation is the need for more intervention studies and evaluation of their effectiveness to invoke behavioral change in physical activity settings (Kipp and Weiss, 2013; Weiss et al., 2012a; Weiss et al., 2012b). The current volume of cross-sectional and longitudinal studies demonstrates robust findings for perceived competence and significant adults and peers as contributors to self-determined motivation and physical activity behaviors. A handful of randomized control and naturalistic intervention studies exist, as well as evaluation research of intact programs, that reveal processes and mechanisms whereby positive change in physical activity motivation and behaviors occurs (Weiss et al., 2012b). These studies importantly identify effective theoretical constructs, and targets of intervention that are susceptible, for invoking change in youths' motivational attitudes and behaviors over time. Continued intervention and evaluation studies are needed to infer causal relationships among processes, mechanisms, and behavioral change.

Our review of developmental perspectives on youth physical activity motivation was slanted toward processes and mechanisms that maximize adaptive orientations and behaviors. However, at higher levels of competitive sport, there is evidence of negative developmental and health-related outcomes, such as injury, stress, burnout, low self-worth, and maladaptive social relationships (Weiss et al., 2012b). Intervention research to address these concerns is needed. On a related note, knowledge is limited about the long-term effects of early sport specialization, a phenomenon that has become commonplace over the last decade (Weiss et al., 2012b). Is starting young and focusing on one sport a necessary condition for developing expertise and high-level performance, or do the downsides outweigh the benefits? Is elite performance even a healthy or realistic goal for the majority of youth? Is a sampling approach, whereby youth are afforded opportunities for learning a wide range of fundamental motor and sport skills, more developmentally appropriate and sustainable for remaining physically active across the life span and optimizing physical and mental health? Answers to these questions await appropriate research designs and methodologies.

Finally, we would be remiss not to mention the role of physical activity to battle the childhood obesity epidemic (USDHHS, 2008). School physical education has the potential to be a strong public health agent in the United States, yet many programs have been eliminated and downsized. Only a handful of states require daily elementary physical education and many high schools specify only one semester of physical education to satisfy graduation requirements. Given the potential of schools to foster greater physical activity and associated health outcomes (Weiss, 2011), it may be timely to institute motivational teaching strategies to foster opportunities for recommended daily physical activity. Children and youth of all ages and ability levels go to school, but not all can afford after-school programs and prohibitive fees that some competitive programs require. Schools can play a huge role in motivating youth and providing developmentally appropriate physical activity opportunities (Weiss, 2011). In doing so, schools can contribute enormously to the health and well-being of our nation's youth. Thus, future research that provides even more definitive evidence of the connections among physical activity and academic, social, mental, and physical health outcomes is key to inspiring change in educational policy that brings back greater frequency and quality of physical education programs at all school levels.

Concluding Remarks

Sport and physical activity contexts are unique compared to other achievement contexts such as school, family, and neighborhood settings. Behaviors and performance are on public display and invite social comparison and evaluation of competencies. As such, understanding determinants and consequences of motivational orientations and behaviors is important for ensuring positive experiences and the potential for youth to derive developmental and health benefits. Perceived physical competence and influence by significant adults and peers are consistent and strong contributors to self-determined motivation and physical activity participation, thus identifying them as susceptible targets of intervention. More intervention and evaluation research is needed in school and out-of-school-time programs to make informed recommendations for how to encourage more youth to initiate, continue, and sustain their activity involvement and thereby accrue the social, mental, and physical health benefits from embracing a physically active lifestyle.

Bibliography

Amorose, A.J., 2007. Coaching effectiveness: exploring the relationship between coaching behavior and motivation from a self-determination theory perspective. In: Hagger, M.S., Chatzisarantis, N.L.D. (Eds.), Intrinsic Motivation and Self-determination in Exercise and Sport. Human Kinetics, Champaign, IL, pp. 209–227.

Fredricks, J.A., Eccles, J.S., 2004. Parental influences on youth involvement in sports. In: Weiss, M.R. (Ed.), Developmental Sport and Exercise Psychology: A Lifespan Perspective. Fitness Information Technology, Morgantown, WV, pp. 145–164.

Horn, T.S., 2004a. Lifespan development in sport and exercise psychology: theoretical perspectives. In: Weiss, M.R. (Ed.), Developmental Sport and Exercise Psychology: A Lifespan Perspective. Fitness Information Technology, Morgantown, WV, pp. 27–71.

Horn, T.S., 2004b. Developmental perspectives on self-perceptions in children and adolescents. In: Weiss, M.R. (Ed.), Developmental Sport and Exercise Psychology: A Lifespan Perspective. Fitness Information Technology, Morgantown, WV, pp. 101–143.

Horn, T.S., 2008. Coaching effectiveness in the sport domain. In: Horn, T.S. (Ed.), Advances in Sport Psychology, third ed. Human Kinetics, Champaign, IL, pp. 240–267.

Horn, T.S., Horn, J.L., 2007. Family influences on children's sport and physical activity participation, behavior, and psychosocial responses. In: Tenenbaum, G., Eklund, R.C. (Eds.), Handbook of Sport Psychology. John Wiley & Sons, Hoboken, NJ, pp. 685–711.

Kipp, L.E., Weiss, M.R., 2013. Physical activity and self-perceptions among children and adolescents. In: Ekkekakis, P. (Ed.), Handbook of Physical Activity and Mental Health. Routledge, London, UK, pp. 187–199.

Smith, A.L., 2007. Youth peer relationships in sport. In: Jowett, S., Lavallee, D. (Eds.), Social Psychology in Sport. Human Kinetics, Champaign, IL, pp. 41–54.

Stuntz, C.P., Weiss, M.R., 2010. Motivating children and adolescents to sustain a physically active lifestyle. American Journal of Lifestyle Medicine 4, 433–444.

U.S. Department of Health and Human Services [USDHHS], 2008. 2008 Physical Activity Guidelines for Americans. Washington, DC.

Weiss, M.R., 2011. Teach the children well: a holistic approach to developing psychosocial and behavioral competencies through physical education. Quest 63, 55–65.

Weiss, M.R., Amorose, A.J., 2008. Motivational orientations and sport behavior. In: Horn, T.S. (Ed.), Advances in Sport Psychology, third ed. Human Kinetics, Champaign, IL, pp. 115–155.

Weiss, M.R., Amorose, A.J., Kipp, L.E., 2012a. Youth motivation and participation in sport and physical activity. In: Ryan, R.M. (Ed.), The Oxford Handbook of Human Motivation. Oxford University Press, New York, pp. 520–553.

Weiss, M.R., Bhalla, J.A., Price, M.S., 2008. Developing positive self-perceptions through youth sport participation. In: Hebestreit, H., Bar-Or, O. (Eds.), The Encyclopaedia of Sports Medicine, The Young Athlete, vol. X. Blackwell Science, Ltd, Oxford, UK, pp. 302–318.

Weiss, M.R., Kipp, L.E., Bolter, N.D., 2012b. Training for life: optimizing positive youth development through sport and physical activity. In: Murphy, S.M. (Ed.), The Oxford Handbook of Sport and Performance Psychology. Oxford University Press, New York, pp. 448–475.

Weiss, M.R., Smith, A.L., Theeboom, M., 1996. "That's what friends are for": children's and teenagers' perceptions of peer relationships in the sport domain. Journal of Sport & Exercise Psychology 18, 347–379.

Weiss, M.R., Stuntz, C.P., 2004. A little friendly competition: peer relationships and psychosocial development in youth sport and physical activity contexts. In: Weiss, M.R. (Ed.), Developmental Sport and Exercise Psychology: A Lifespan Perspective. Fitness Information Technology, Morgantown, WV, pp. 165–196.

Weiss, M.R., Williams, L., 2004. The why of youth sport involvement: a developmental perspective on motivational processes. In: Weiss, M.R. (Ed.), Developmental Sport and Exercise Psychology: A Lifespan Perspective. Fitness Information Technology, Morgantown, WV, pp. 223–268.

Motivation, Familial Influences on

Sandra D Simpkins, T. Denny Sanford School of Social and Family Dynamics, Arizona State University, Tempe, AZ, USA
Jennifer A Fredricks, Human Development Department, Connecticut College, New London, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Grounded in the Eccles' expectancy-value model, we review the literature on how families socialize youths' motivational beliefs and choices. Parental beliefs and behaviors are important precursors to youths' motivation. Parents' beliefs and behaviors are shaped by child characteristics, such as gender, and family characteristics, such as income. Following the review, we provide ideas on future directions for this research.

For years, scholars have tried to understand why people decide to engage in certain endeavors and not others. How do college students select their courses and majors? How do children pick which after-school interests and activities they do? Furthermore, where do youths' motivational beliefs come from? How are they influenced by their immediate surroundings, such as the family? In this article, we provide an overview of how families influence youths' motivational beliefs and a summary of the empirical research to date. First, we outline the Eccles' expectancy-value model, one theoretical model that delineates how families influence motivation. Next, we review the current research on how families either support or undermine motivation. We conclude with recommendations for future research on families and motivation.

Expectancy-Value Model of Achievement-Related Behaviors

Drawing on work associated with decision-making, achievement theory, and attribution theory, Eccles and her colleagues developed the expectancy-value model, a comprehensive theoretical model of achievement-related choices (Eccles-Parsons et al., 1983; see Figure 26.1 for the most recent version). They hypothesized that achievement-related choices would be most directly shaped by individuals' competence beliefs and task values. Individuals' competence beliefs are operationalized as judgments about ones' ability to perform certain tasks, expectations for future success, and self-efficacy. Eccles and her colleagues argue that subjective task value is a function of four components: interest (the enjoyment one gets from engaging in the task), utility value (the instrumental value of the task or activity for helping to fulfill another short- or long-range goal), attainment value (the link between the task and one's sense of self and either personal or social identity), and cost (what may be given up by making a specific choice or the negative experiences associated with each possible choice). According to this model, individuals will select achievement-related activities, such as educational, vocational, and leisure choices, if they have high competence beliefs and task values (Eccles-Parsons et al., 1983).

Box B in **Figure 1** denotes the role that socializers play in shaping youths' experiences and beliefs. Because socializers can include a variety of people, such as teachers and parents, Eccles and her colleagues developed a model that describes specifically how parents can influence their children's achievement-related motivational beliefs and outcomes as shown in **Figure 2** (Eccles, 1993). Beginning at the far left, this model predicts that exogenous, cultural, and demographic characteristics of the family, in combination with specific child characteristics, influence parents' general and child-specific beliefs. In turn, parents' beliefs influence parents' general and child-specific behaviors and practices, which in turn influence children's motivational beliefs and choices. In the next section, we review the current research addressing the various paths depicted in the model.

Child and Family Demographic Characteristics

Families are embedded within a specific sociocultural context. As such, family and parenting processes are shaped by larger cultural norms and expectations that are historically grounded. These broad societal and cultural norms are communicated to youth through more immediate social experiences within the family. Both child characteristics and family demographics determine family and parenting processes.

Child Characteristics

Child gender is one of the central child-level demographic characteristics that filters through the entire expectancy-value model. Gender differences in youths' achievement, engagement, and motivational beliefs are evident when children begin formal schooling and often persist throughout school and even to career selection (e.g., Andre et al., 1999). The early onset of gender differences before schooling suggests that families are an important source of influence.

Parents have expectations of what is appropriate for girls and boys as well as what girls and boys are good at. Parents of sons typically have higher value of traditionally 'male-typed domains' (e.g., sports, math) and believe their child is more skilled in these domains compared to parents of daughters (e.g., Bhanot and Jovanovic, 2005; Fredricks and Eccles, 2005). In contrast, parents of daughters rate their child as more competent in reading than do parents of sons (Bhanot and Jovanovic, 2005; Frome and Eccles, 1998). However, there is great variability in parents' gender-typed beliefs. In fact, other

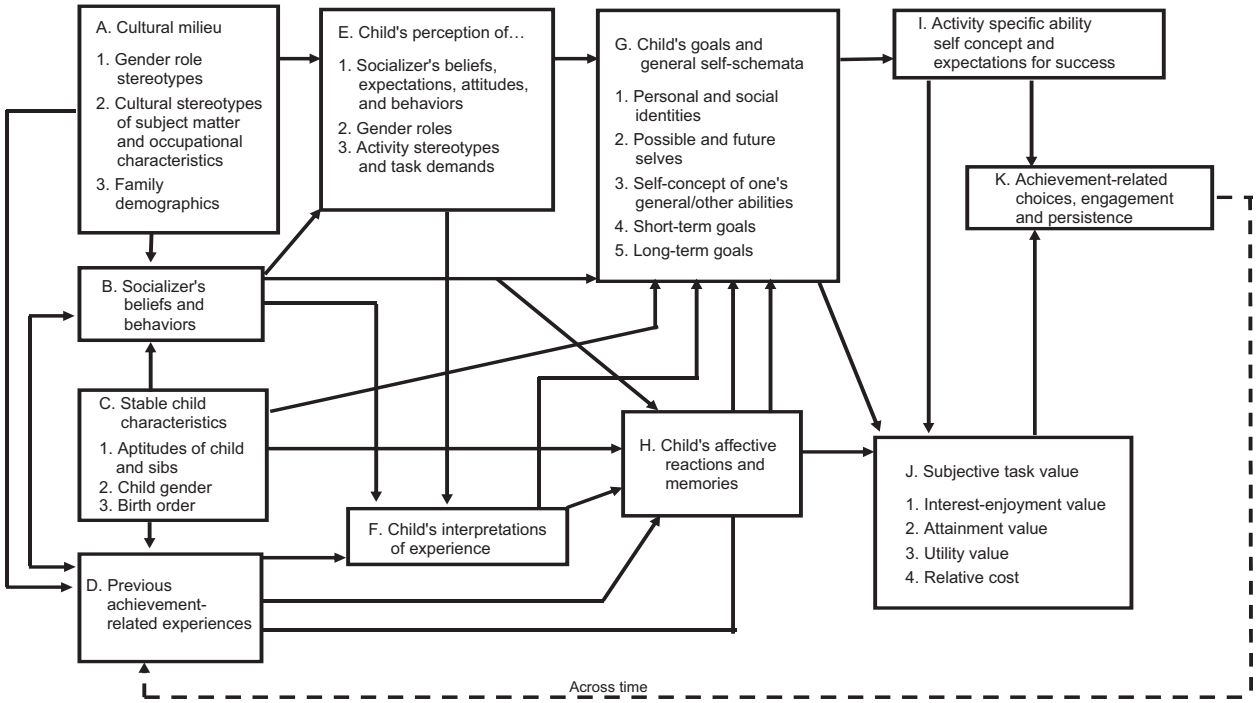

Figure 1 Expectancy-value Model of Achievement Choices.

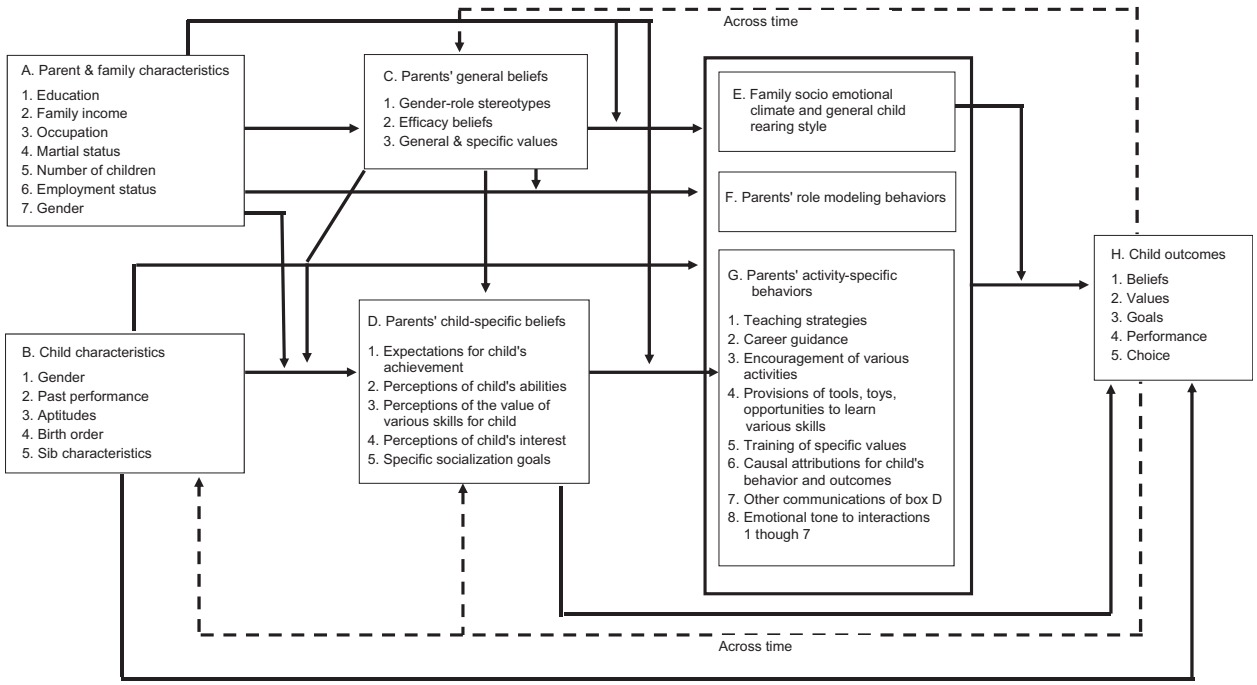

Figure 2 Model of Parents Influence on Children's Achievement Related to Self-Perceptions, Values, and Behaviors.

studies fail to document gender differences in parents' perceptions of children's ability in and the importance of math, sports, and reading (Andre et al., 1999; Bois et al., 2002; Frome and Eccles, 1998). It is likely that some if this variability depends on the extent to which parents hold traditional gender role beliefs.

Parents can also be gender-typed in their behaviors. Emerging evidence suggests that parents may differentially socialize their sons versus their daughters based on their behaviors in traditionally gender-stereotyped domains (Lytton and Romney, 1991). For example, in both childhood and adolescence, parents of sons have been found to provide more

opportunities and materials promoting science and sports, which are traditionally male-typed domains, than parents of daughters (Bleeker and Jacobs, 2004; Fredricks and Eccles, 2005). Furthermore, Tenenbaum and Leaper (2003) found that parents provide richer explanations and discussions about science museum exhibits when with sons than with their daughters.

Not only can child gender shape parents' beliefs and behaviors, but it can also influence development by shaping the relations among these indicators. In other words, a separate and equally important question is whether the precursors of children's motivation and outcomes are the same for boys and girls. It is possible that different parent behaviors may motivate boys compared to girls. Or perhaps, the parents' confidence in children (or lack thereof) really matters for one gender but not for the other. If the precursors vary for boys and girls, then it will be necessary for parents to engage in different behaviors to support the motivation and pursuits of their sons compared to their daughters. Emerging evidence suggests that although the extent to which parents endorse certain beliefs or the frequency that parents engage in certain behaviors may vary by child gender, the precursors are similar. Prior research shows that parents' activity-related behaviors have similar associations with both boys' and girls' beliefs and choices (Bhanot and Jovanovic, 2005; Simpkins et al., 2012; Simpkins et al., 2010). Thus, parent support is related in higher motivation for both boys and girls.

Family Characteristics

Family-level demographic characteristics also determine family socialization and youths' motivation and achievement. Two of the key demographic indicators are socioeconomic status (SES) and culture (Quintana et al., 2006). Family SES and culture could affect children's motivation indirectly through the opportunity structures in the child's surrounding environments as well as parents' beliefs and practices.

Children growing up in high-SES families fare better academically and are involved in more organized activities and skill-based informal activities during the after-school hours than are children growing up in lower-SES homes (e.g., Fredricks and Simpkins, 2012). For children to both participate in and succeed at a domain, they need to have available opportunities, as well as the knowledge and ability to take advantage of those opportunities – both of which are more often limited in low-SES families (Simpkins et al., 2013). Not only do such parents have limited resources to implement whatever strategies they think might be effective, they also have to cope with more external stressors than middle-class families living in stable, resource-rich neighborhoods (Furstenberg et al., 1999; McLoyd, 1990). Furthermore, SES differences may also be a result of differences in family norms and values (Lareau, 2003). Middle- and upper-class families are more likely to engage in a variety of behaviors, such as enrollment in organized activities, because they believe that such 'concerted cultivation' helps their children develop the skills and experiences necessary to be successful.

Other research has explored the role of culture in families and motivation. Culture determines what activities and goals (e.g., career choices) are valued as well as determining general parental child-rearing beliefs, theories, and goals (Rogoff, 2003). For instance, traditional gender roles and familism (e.g., family as referent and strong value of family) are two core values within Latino culture. Latino daughters in families with strong traditional gender roles may be less likely to engage in masculine activities, such as sports, and often find it challenging to attend organized after-school activities as they often are more protected and have more familial duties than sons (Simpkins et al., 2013). Cultural values can also shape the general child-rearing environment. Among Latino families, familism predicts higher parental warmth and protects against parents' increased harsh parenting in the face of neighborhood disadvantage (White et al., 2012). Culture defines parental belief systems and family norms and routines (Grusec et al., 2000).

Summary

There are many ways that child and family demographic characteristics affect the development of children's motivation. It is important to note, however, that even though these demographic characteristics have been linked repeatedly to children's school achievement, their effects are almost always indirect, mediated by their association with parents' beliefs, practices, and resources (Benner et al., 2008; Melby et al., 2008). Parenting (e.g., autonomy, consistency in parenting) can override the effects of even the most stressful circumstances (Van Horn et al., 2009). In order to support families, we need more research with diverse populations examining when and how families overcome stressful circumstances.

Parents' General and Child-Specific Beliefs

The Eccles' expectancy-value model integrates parents' general beliefs and their beliefs about their child specifically. As noted in Box C, parents' general beliefs include their value of various pursuits (in general and for particular groups) as well as their general beliefs about child rearing and child development. Examples of parents' general beliefs include the extent to which parents believe ability is innate or malleable and parents' beliefs regarding what is appropriate for boys and girls. In contrast, parents' child-specific beliefs are beliefs about a specific child in a particular domain. These beliefs are child- and domain-specific, such that parents with multiple children may have unique beliefs for each of their children across different domains. For example, parents may believe that their oldest daughter is a highly skilled artist with weaker athletic skills, whereas their younger daughter is a skilled artist and reader. Parents' child-specific beliefs include perceptions of their child's abilities, value of a domain for that child, and goals for that child.

Evidence from several studies suggests that parents' estimates of their children's competencies and value of a domain for their child are important predictors of children's own self-concepts, values, and performance. As one might expect, parents who believe their child is skilled in a domain or value a domain for their child are more likely to have children who believe in themselves and who also value that domain (Andre et al., 1999; Bhanot and Jovanovic, 2005;

Bleeker and Jacobs, 2004; Fredricks and Eccles, 2005; Frome and Eccles, 1998). These patterns have emerged in several studies of math, English, and sports, with some emerging work in music, although parents' value seems to be a less consistent predictor than their perceptions of children's competencies (Eccles and Harold, 1991; McGrath and Repetti, 2000).

How do parents' beliefs shape children's beliefs? Parents' general beliefs are the basis for their child-specific beliefs (Box D) and provide the lens through which they interpret youths' previous performance and aptitude. For example, mothers who believed males are better at math than females (i.e., the general belief) have been found to have lower confidence in their daughters' math abilities (i.e., the child-specific belief) than mothers with less traditional beliefs (Bleeker and Jacobs, 2004). Parents' general beliefs can also change how they view children's actual performance. In a recent review, Yamamoto and Holloway (2010) found that parents' future educational expectations for their child was more closely aligned with children's previous performance if parents believed ability was innate, they trusted the school, and they had high self-efficacy beliefs.

Parents' general and child-specific beliefs are theorized to influence parents' behaviors. Indeed, parents' general and child-specific beliefs have been found to predict their parenting behavior in expected ways (e.g., Eccles, 1993; Goodnow and Collins, 1990; Sigel et al., 1992). Parents, for example, are more likely to support children's cognitive development and academic pursuits if parents believe academics and schooling are important. However, much of this research has focused on academics and on parents' general beliefs (and not their child-specific beliefs). But, do parents' behaviors explain the correspondence between parents' and children's beliefs? The current evidence is limited and mixed. The evidence is more consistent in the leisure domains (e.g., sports, music) than in the academic domains (e.g., math), though only a few researchers have tested this tenet of the model (Bhanot and Jovanovic, 2005; Simpkins et al., 2012). For example, Simpkins et al. (2012) found that mothers' behaviors mediated the link between mothers' and youths' beliefs in sports, music, and math, but not reading.

Summary

Parents hold general beliefs about child development that provide the basis for their beliefs about each specific child across different domains. It is theorized that parents' beliefs in turn shape their behavior. Although scholars have found that parents' beliefs and children's beliefs are linked, there is less evidence that parents' behaviors account for those links. We discuss one reason why parenting behaviors may not be a strong mediator, namely the operationalized parenting behaviors, under the future directions section.

Parent Behaviors

Parents' behaviors are noted in Boxes E, F, and G in **Figure 2**. Parents' behaviors are grouped into three types: general climate and style, modeling, and activity-specific behaviors.

General Child-Rearing Climate and Style

The affective and motivational climate created by parents can influence youths' motivation in direct and indirect ways. One of the most widely studied aspects of general climate is parenting style. In the United States, well-adjusted youth are most likely to have parents who use an authoritative parenting style, which is characterized by appropriate levels of warmth and structure, compared to other parenting styles (Steinberg et al., 1991). However, the meaning and impact of parenting styles varies across ethnic groups, cultures, and neighborhood contexts (Chao, 2001).

Recently scholars have examined other aspects of the general family climate and parenting. One such dimension is autonomy support versus control (Grolnick, 2003). From a self-determination perspective, a sense of autonomy and lack of feeling controlled are crucial for engagement and learning (Ryan and Deci, 2002). Autonomy-supportive parents allow children to explore their own environment, to initiate their own behavior, to take an active role in solving their own problems, and to express their points of view (Raftery et al., 2012). In contrast, controlling parents exert pressure to regulate children's behaviors through commands, directives, or love withdrawal; deny their children chances to solve problems; and ignore their points of view (Pomeranz et al., 2007). A growing body of research demonstrates the benefits of parental autonomy support for children's motivation and engagement (Raftery et al., 2012). Autonomy support seems to be particularly important for low-achieving children (Ng et al., 2004).

The affective and motivational climate, or the general warmth shared by parent and child can also influence the extent to which a child internalizes the teachings and advice from parents (Grolnick, 2003). A negative emotional climate or a controlling environment can alter the influence of parents' supportive behaviors. For example, mothers' academic support actually had negative consequences for youth who did not share a warm relationship (Simpkins et al., 2006).

Modeling

The second type of parenting behavior emphasized in the expectancy-value model is parental modeling. Parents' vocational and leisure pursuits can influence their children's choices through processes associated both with observational learning and with the desire to be like their parents (Bandura, 1997; Eccles, 1993). Although modeling is often cited as theoretically important for children's acquisition of attitudes and behaviors, the empirical support for role modeling is mixed. Some studies show a significant link between modeling and participation (e.g., Kahn et al., 2008), whereas others show no relation (e.g., Fredricks and Eccles, 2005).

There are several possible explanations for these mixed findings. First, in some of these studies, modeling and other parenting behaviors were included within the same regression-based analysis, and parent modeling may be a weak unique predictor relative to other behaviors (e.g., coactivity). Another possibility is that modeling may only be effective when a behavior or activity can be learned by simply watching someone. Modeling may have a weak influence on

activities that are more cognitive in nature and require instruction, such as completing a math problem or reading a book. Furthermore, parents' engagement in these domains in the presence of their children is likely to vary. We have found that few parents play musical instruments themselves or have a music-related career (Simpkins et al., manuscript under review). Parents' engagement in other domains, like sports, may only influence children if they see their parent exercising or have knowledge of these activities.

Activity-Specific Parenting

The third and final type of parent behavior is their activity-specific behavior (see Box G). These behaviors are unique from the other two types in that these behaviors are specifically geared toward a child in a particular domain. Parents can promote children's motivation and participation in a variety of ways. In some ways, parents interact with youth directly. In other ways, parents design or manage children's environments that indirectly influence motivation.

Parent encouragement and participation in an activity together (i.e., coactivity) are two strategies that parents can use to promote children's activity participation. Parent–child coactivity can occur through informal activities at home, involvement in organized activities such as being a coach, and attending a community event together (Simpkins et al., 2010). Parent–child coactivity provides a context for parents to offer verbal encouragement, convey the value of an activity, provide feedback, and both model and teach skills. Talented adolescents and elite adult athletes, artists, and musicians report that parent encouragement and coactivity are two key dimensions of parenting that shaped their pathways (Bloom, 1985; Csikszentmilayi et al., 1993). Other research has found that both parental encouragement and coactivity predict youths' self-efficacy beliefs, perceptions of importance, career interests, and activity participation in both academic and leisure domains (e.g., Fredricks and Eccles, 2005; LeFevre et al., 2009; Pugliese and Tinsley, 2007).

Another way that parents influence motivation is through the provision of resources inside and outside of the home (Furstenburg et al., 1999). Parents structure children's experiences through the provision of toys, equipment, books, and other learning activities in the home. Parents also determine what environments outside of the home to which their children will be exposed by selecting neighborhoods, schools, and activities. According to sociocultural theory, the people and objects in one's environments expose children to particular experiences and value systems (Vygotsky, 1978). Higher exposure leads to children's comfort, engagement, and learning in a domain, as well as coactivity with their parents (e.g., reading a book with their parents). Existing research supports sociocultural theory; studies link provision of materials to higher motivation, achievement, and participation in sports, music, and reading (e.g., Fredricks and Eccles, 2005; Pugliese and Tinsley, 2007; Whitehurst and Lonigan, 2001).

As can be seen in Box G, many of these behaviors are supportive in nature and focus on promoting children's motivation, engagement, and persistence. However, parents can also play a role in discouraging children's pursuit of particular endeavors by either not encouraging them to pursue a domain or by actively discouraging children. For instance, a child is unlikely to engage in sports if parents do not provide any athletic materials or experiences. Parents may also actively discourage their child from pursuing a domain they deem inappropriate because of gender-typed beliefs or that their child is not skilled at. This discouragement may be intentional or unintentional. For instance, parents who tell their daughters that a science activity is for boys may unintentionally discourage their daughters' pursuit of science. Little is known about parents' active discouragement of motivation. One would expect that parents who discourage children's pursuit of a domain should have children who are less likely to pursue that domain unless they receive support from another source (e.g., friends, teachers).

Summary

Specific supportive behaviors alone cannot enhance motivation. They must be delivered in an autonomous fashion and within a relationship characterized by warmth for those behaviors to have the largest positive impact on motivation. In other words, the three types of parenting behaviors noted in the expectancy-value models work together to impact motivation. Parent support may, in fact, deter motivation if the overall climate is negative or controlling.

Future Directions

In sum, a growing body of research supports the paths outlined in the Eccles' expectancy-value model. This research has provided important insights into how parents can both support and undermine achievement motivation. In the next section, we outline some important future directions for research on families and motivation.

Holistic Models of Parenting

The findings reviewed in this article suggest that multiple parent behaviors influence children. However, few studies have adopted a holistic, integrated view of parenting. Instead, much of the research has focused on only one or two behaviors, one parent, and/or one activity domain. This way of decomposing parenting contradicts qualitative research suggesting that families influence children simultaneously through multiple behaviors. For example, Bloom (1985) found that elite athletes, musicians, and artists came from homes that were 'child-centered,' where parents supported children's talents through multiple strategies at the same time. The reality is that parenting is executed in a holistic manner that incorporates multiple strategies and behaviors. Further, parenting in one domain is usually analyzed as separate from other domains, but this does not reflect most families' lives. Family time and resources are not limitless. Parents simply cannot enroll their child in every activity or support their child in every domain. In reality, parents have to weigh the costs and benefits of various activities and supporting their children in different domains. As a result, parents have to make challenging decisions on what to promote and how (Grusec et al., 2000).

Few researchers, however, have adopted such an integrated view of family influences in their work. Moreover, many studies

use regression techniques that assess the unique contribution of each predictor adjusting for the other variables in the model. Such analyses decompose parenting into separate, compartmentalized behaviors when it is the parenting experience in its entirety that influences child outcomes. One method which is better able to capture parenting as a whole is the creation of indexes, which often include a heterogeneous set of indicators that may have a low inter-item reliability, but nevertheless, should be combined in a singular construct because they are all theorized to cause the same outcome (Bradley, 2004). Classic examples include research on cumulative risk (Sameroff et al., 1998), the HOME observational scale (Bradley, 2004), and family involvement in children's education (Jeynes, 2007). Researchers should also consider the use of latent variables in SEM (Simpkins et al., 2012), and cluster analysis to provide a more holistic view of parental influences, which complements the knowledge gained through a traditional regression analysis.

Direction of Influence

Most research, including this article, addresses the influence of parents on children. The thrust of the expectancy-value model is also on how parents influence children. However, parents are expected to adjust their beliefs and subsequently their behaviors based on children's aptitude and experiences (Eccles, 1993). Parents may enroll their child in athletics because they value athletics due to the opportunities it provides for physical activity and developing life-long skills, like teamwork. Once parents find out that their child is not very coordinated and struggles at sports, they may change their value of athletics for this child and not enroll them in future sport activities. Parents may also follow their child's interests. If their child likes animals, they might read more stories on animals, take their child to the local zoo, or watch television shows on animals to sustain their child's interest in biology. Qualitative work suggests that parents support their adolescents in the organized after-school activity they prefer (Simpkins et al., 2013).

Such bidirectional influences of children on parents are depicted in many theories of parenting (Grusec et al., 2000), but have not been widely researched. The socialization of children is not a unidirectional process by which parents simply shape children. Although the influence of children on parents has been at the forefront of research on some areas of development, such as temperament (Bates, 1987), children's effects on their parents have not received as much attention in the field of achievement motivation.

Family Diversity

In this article, we briefly reviewed the research on how SES and culture influence families. Although these two aspects of family diversity have received the most attention, there are other aspects of family diversity that are equally important, such as immigration and ethnicity (Quintana et al., 2006). For example, in 2011, approximately 40 million people in the United States were immigrants (Pew Hispanic Center, 2013). For each aspect of diversity, researchers will need to understand both the strengths and challenges facing families, and how these experiences impact on child motivation and achievement. It will also be important in future research to move away from using these demographic characteristics as grouping variables and instead, to unpack how these characteristics shape families' development (Benner et al., 2008; Melby et al., 2008; Simpkins et al., 2013; Van Horn et al., 2009). All of these aspects that help define diversity among families work in combination to impact development. For instance, immigration history determines, in part, families' cultural orientation, which in turn can influence parents' beliefs about organized activities and youths' participation (Simpkins et al., 2013). Moving forward, it will be important to address how each aspect of family diversity uniquely impacts development as well as how they codetermine development.

Conclusion

In this article, we reviewed the ways in which families can influence youths' motivation using the Eccles' expectancy-value model as an organizing framework (Eccles, 1993). Parenting is a multidetermined process that is grounded in one's larger sociocultural context and circumstances. Although there are consistencies in parenting over time, it can also fluctuate day to day given people's moods and daily demands. In future research, it will be important to use a range of methodological techniques to collected nuanced data that capture changes in parent behaviors over time. Researchers have made great strides in describing such processes and how parenting may influence youth. As noted in the future directions sections, there is also much more work to complete to address salient gaps in our understanding.

See also: Academic Achievement Motivation, Development of; Academic Engagement; Academic Motivation and Performance: Task Value Interventions; Expectancy-Value-Cost Model of Motivation; Gender and Academic Motivation; Interest, Psychology of; Internal/External Frame of Reference Model; Leisure Activities Choices among Adolescents; Motivation in Youth Sport and Physical Activity: Developmental Perspectives; Race and Academic Motivation; School Achievement: Motivational Determinants and Processes; Schooling: Impact on Cognitive and Motivational Development; Self-Determination Theory; Vocational Interests, Values, and Preferences, Psychology of.

Bibliography

Andre, T., Whigham, M., Hendrickson, A., Chambers, S., 1999. Competency beliefs, positive affect, and gender stereotypes of elementary students and their parents about science versus other school subjects. Journal of Research in Science Teaching 36, 719–747.

Bandura, A., 1997. Self-efficacy: The Exercise of Control. W.H. Freeman, New York.

Bates, J.E., 1987. Temperament in infancy. In: Osofsky, J.D. (Ed.), Handbook of Infant; Development. Wiley, New York, pp. 1101–1149.

Benner, A.D., Graham, S., Mistry, R.S., 2008. Discerning direct and mediate effects of ecological structures and processes on adolescents' educational outcomes. Developmental Psychology 44, 840–854.

Bhanot, R., Jovanovic, J., 2005. Do parents' academic gender stereotypes influence whether they intrude on their children's homework? Sex Roles 52, 597–607. http://dx.doi.org/10.1007/s11199-005-3728-4.

Bleeker, M.M., Jacobs, J.E., 2004. Involvement in math and science: do mothers' beliefs matter 12 years later? Journal of Educational Psychology 96, 97–109. http://dx.doi.org/10.1037/0022-0663.96.1.97.

Bloom, B.S., 1985. Developing Talent in Young People. Ballantine Books, New York.

Bois, J.E., Sarrazin, P.G., Brustad, R.J., Trouiloud, D.O., Cury, F., 2002. Mothers' expectancies and young adolescents' perceived competence: a year-long study. Journal of Early Adolescence 23, 384–406.

Bradley, R.H., 2004. Chaos, culture, and covariance structures: a dynamic systems view of children's experiences at home. Parenting: Science and Practice 4, 243–257. http://dx.doi.org/10.1080/15295192.2004.9681272.

Chao, R., 2001. Extending research on the consequences of parenting style for Chinese Americans and European Americans. Child Development 72, 1832–1843. http://dx.doi.org/10.1111/1467-8624.00381.

Csikszentmihalyi, M., Rathunde, K., Whalen, S., 1993. Talented Teenagers: The Roots of Success and Failure. Cambridge University Press, New York.

Eccles, J.S., 1993. School and family effects on the ontogeny of children's interests, self-perceptions, and activity choice. In: Jacobs, J. (Ed.), Nebraska symposium on Motivation, 1992: Developmental Perspectives on Motivation. University of Nebraska Press, Lincoln, pp. 145–208.

Eccles, J.S., Harold, R.D., 1991. Gender differences in sports involvement: applying the Eccles' expectancy-value model. Journal of Applied Sport Psychology 3, 7–35. http://dx.doi.org/10.1080/10413209108406432.

Eccles-Parsons, J., Adler, T.F., Futterman, R., Goff, S.B., Kaczala, C.M., Meece, J.L., 1983. Expectancies, values, and academic behaviors. In: Spence, J.T. (Ed.), Achievement and Achievement Motivations. Freeman, San Francisco, pp. 75–146.

Fredricks, J.A., Eccles, J.S., 2002. Children's competence and value beliefs from childhood through adolescence: growth trajectories in two male sex-typed domains. Developmental Psychology 38, 519–533.

Fredricks, J.A., Eccles, J.S., 2005. Family socialization, gender, and sport motivation and involvement. Journal of Sport and Exercise Psychology 27, 3–31.

Fredricks, J.A., Simpkins, S.D., 2012. Supporting ethnic minority youth during the after school hours: the potential of organized activity participation. Child Development Perspectives 6, 280–287.

Frome, P., Eccles, J.S., 1998. Parents' influence on children's achievement-related perceptions. Journal of Personality and Social Psychology 2, 435–452. http://dx.doi.org/10.1037/0022-3514.74.2.435.

Furstenberg, F., Cook, T., Eccles, J., Elder, G., Sameroff, A., 1999. Managing to Make It: Urban Families in Adolescent Success. University of Chicago Press, Chicago.

Goodnow, J.J., Collins, W.A., 1990. Development According to Parents: The Nature, Sources, and Consequences of Parents' Ideas. Erlbaum, London.

Grolnick, W.S., 2003. The Psychology of Parental Control: How Well-meant Parenting Backfires. Erlbaum, Hillside, NJ.

Grusec, J.E., Goodnow, J.J., Kuczynski, L., 2000. New directions in analyses of parenting contributions to children's acquisition of value. Child Development 71, 205–211. http://dx.doi.org.ezproxy1.lib.asu.edu/10.1111/1467-8624.00135.

Jeynes, W.H., 2007. The relationship between parental involvement and urban secondary school student academic achievement: a meta-analysis. Urban Education 42, 82–110. http://dx.doi.org/10.1177/0042085906293818.

Kahn, J.A., Huang, B., Gillman, M.W., Field, A.E., Austin, S.B., Colditz, G.A., Frazier, A.L., 2008. Patterns and determinants of physical activity in U.S. adolescents. Journal of Adolescent Health 42 (4), 369–377. http://dx.doi.org/10.1016/j.jadohealth.2007.11.143.

Lareau, A., 2003. Unequal Childhood: Class, Race, and Family Life. University of California Press, Berkeley, CA.

LeFevre, J.A., Skwarchuk, S.L., Smith-Chant, B.L., Fast, L., Kamawar, D., Bisanz, J., 2009. Home numeracy experiences and children's math performance in the early school years. Canadian Journal of Behavioural Science 41, 55–66. http://dx.doi.org/10.1037/a0014532.

Lytton, J., Romney, D.M., 1991. Parents' differential socialization of boys and girls: a meta-analysis. Psychological Bulletin 2, 267–296. http://dx.doi.org/10.1037/0033-2909.109.2.267.

Melby, J.N., Conger, R.D., Fang, S., Wickrama, K.A.S., Conger, K.J., 2008. Adolescent family experiences and educational attainment during early adulthood. Developmental Psychology 44, 1519–1536.

McGrath, E.P., Repetti, R.L., 2000. Mothers' and fathers' attitudes toward their children's academic performance and children's perceptions of their academic competence. Journal of Youth and Adolescence 29, 713–723. http://dx.doi.org.ezproxy1.lib.asu.edu/10.1023/A:1026460007421.

McLoyd, V.C., 1990. The impact of economic hardship on African American families and children: psychological distress, parenting, and socioemotional development. Child Development 61, 311–346.

Mistry, R.S., Vandewater, E.A., Huston, A.C., McLoyd, V.C., 2002. Economic well-being and children's social adjustment: the role of family process in an ethnically diverse low-income sample. Child Development 73, 935–951.

Ng, F.F., Kenney-Benson, G.A., Pomerantz, E.M., 2004. Children's achievement moderates the effects of mothers' use of control and autonomy support. Child Development 75, 764–780.

Pew Hispanic Center, 2013. A Nation of Immigrants: A Portrait of the 40 Million, Including 11 Million Unauthorized. Retrieved from: http://www.pewhispanic.org/2013/01/29/a-nation-of-immigrants.

Pomerantz, E.M., Moorman, E.A., Litwack, S.D., 2007. The how, whom, and why of parents' involvement in children's academic lives: more is not always better. Review of Educational Research 77 (3), 373–410.

Pugliese, J.A., Tinsley, B.J., 2007. Parental socialization of child and adolescent physical activity: a meta-analysis. Journal of Family Psychology 21, 331–343. http://dx.doi.org/10.1037/0893-3200.21.3.331.

Quintana, S.M., Aboud, F.E., Chao, R.K., Contreras-Grau, J., Cross, W.E., Hudley, D., Hughes, D., Liben, L.S., Nelson-Le Gall, S., Vietze, D.L., 2006. Race, ethnicity, and culture in child development: contemporary research and future directions. Child Development 77, 1129–1141. http://dx.doi.org/10.1111/j.1467-8624.2006.00951.x.

Raftery, J.N., Grolnick, W.S., Flamm, E.S., 2012. Families as facilitators of student engagement: towards a home-school partnership. In: Christenson, S.L., Reschly, A.L., Wylie, C. (Eds.), Handbook of Research on Student Engagement. Springer, New York, pp. 343–364.

Rogoff, B., 2003. The Cultural Nature of Human Development. Oxford University Press, Oxford, UK.

Ryan, R.M., Deci, E.L., 2002. An overview of self-determination theory: an organismic-dialectical perspective. In: Deci, E.L., Ryan, R.M. (Eds.), Handbook of Self-determination Theory Research. University of Rochester Press, Rochester, NY, pp. 3–33.

Sameroff, A.J., Bartko, T., Baldwin, A., Baldwin, C., Seifer, R., 1998. Family and social influences on the development of child competence. In: Lewis, M., Feiring, C. (Eds.), Families, Risk, and Competence. Lawrence Erlbaum Publishers, Mahwah, NJ.

Sigel, I.E., McGillicuddy-DeLisi, A.V., Goodnow, J.J. (Eds.), 1992. Parental Belief Systems, second ed. Erlbaum, Hillsdale, NJ.

Simpkins, S.D., Delgado, M., Price, C., Quach, A., Starbuck, E., 2013. Socioeconomic status, ethnicity, culture, and immigration: examining the potential mechanisms underlying Mexican-origin adolescents' organized activities. Developmental Psychology 49, 706–721.

Simpkins, S.D., Fredricks, J., Eccles, J.S. (in press). Parent beliefs to youth choices: mapping the sequence of predictors from childhood to adolescence. Monographs of the Society for Research in Child Development.

Simpkins, S., Fredricks, J.A., Eccles, J.S., 2012. Charting the Eccles' expectancy value model from parents' beliefs in childhood to youths' activities in adolescence. Developmental Psychology 48, 1019–1032.

Simpkins, S.D., Vest, A.E., Dawes, N.P., Neuman, K.I., 2010. Dynamic relations between parents' behaviors and children's motivational beliefs in sports and music. Parenting: Science and Practice 10, 97–118. http://dx.doi.org/10.1080/15295190903212638.

Simpkins, S.D., Weiss, H., McCartney, K., Kreider, H.M., Dearing, E., 2006. Mother–child relationship as a moderator of the relation between family educational involvement and child achievement. Parenting: Science and Practice 6, 49–57.

Steinberg, L., Mounts, N.S., Lamborn, S.D., Dornbush, S.M., 1991. Authoritative parenting and adolescent adjustment across varied ecological niches. Journal of Research on Adolescence 1, 19–36.

Tenenbaum, H.R., Leaper, C., 2003. Parent-child conversations about science: socialization of gender inequities. Developmental Psychology 39, 34–47. http://dx.doi.org/10.1037//0012-1649.39.1.34.

Van Horn, M.L., Jaki, T., Masyn, K., Ramey, S.L., Smith, J.A., Antaramian, S., 2009. Assessing differential effects: applying regression mixture models to identify variations in the influence of family resources on academic achievement. Developmental Psychology 45, 1298–1313.

Vygotsky, L.S., 1978. The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA.

White, R.M.B., Roosa, M.W., Zeiders, K.H., 2012. Neighborhood and family intersections: prospective implications for Mexican American adolescents' mental health. Journal of Family Psychology 26, 793–804. http://dx.doi.org.ezproxy1.lib.asu.edu/10.1037/a0029426.

Whitehurst, G.J., Lonigan, C.J., 2001. Emergent literacy: development from pre readers to readers. In: Neuman, S.B., Dickinson, D.K. (Eds.), Handbook of Early Literacy Research. Guilford, New York, pp. 11–29.

Yamamoto, Y., Holloway, S.D., 2010. Parental expectations and children's academic performance in sociocultural context. Educational Psychological Review 22, 189–214.

Motivation, Learning, and Instruction

Eric M Anderman, The Ohio State University, Columbus, OH, USA
DeLeon Gray, North Carolina State University, Raleigh, NC, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M. Boekaerts, volume 15, pp. 10112–10117, © 2001, Elsevier Ltd.

Abstract

Research on academic motivation offers important implications for instruction. In this article, we review four current and prominent motivation frameworks that are particularly influential: expectancy-value theory, achievement goal orientation theory, self-determination theory, and social cognitive theory. The basic tenets of each theory are reviewed, as well as recent research that focuses on academic outcomes. Implications for learning and instruction are discussed for each framework.

Motivation is a topic that is of great interest to educators because it is related in important ways to student learning. Motivation is extremely important because students' motivation toward various academic subject domains is related to career and life choices. Nevertheless, although the topic is extremely popular with practitioners, there are many misconceptions about what motivation is, and how it is related to student learning. According to Schunk et al. (2008), motivation is "the process whereby goal-directed activity is instigated and sustained" (p. 4). Motivation has to do with how individuals conceptualize their goals, how they carry out activities to attain those goals, how they persist at working toward those goals, and how they evaluate their performance upon goal completion (or noncompletion). As we will review in this article, motivation is a nuanced topic, wherein a variety of theoretical perspectives lend themselves to an array of recommendations to enhance student motivation and achievement.

One of the particularly important distinctions that motivation researchers have made is the distinction between students' *motivational beliefs* and the ways in which teachers' *instructional practices* affect those beliefs (Midgley, 2002). Specifically, it is now very clear that the instructional practices that teachers use in classrooms affect student motivation in important ways (Anderman et al., 2012). In addition, we also know that there is variation in how students will respond to various instructional techniques; whereas the use of a reward might enhance motivation for some students, the same reward might inhibit motivation in others.

In this article, we review four of the most prominent theoretical frameworks in the study of academic motivation. After briefly reviewing the basic tenets of each theory, we will examine current research findings within each, and then discuss implications for learning and instruction. We have selected four theoretical frameworks that have been the focus of much research over the past decade: expectancy-value theory (EVT), achievement goal theory, self-determination theory (SDT), and social cognitive theory (SCT).

Expectancy-Value Theory

EVTs of motivation have a long history in psychology and education. This work originated in theories and studies put forth by scholars including Atkinson (1964) and Lewin (1935). Most current work is rooted in the work by Eccles and her colleagues (most notably Wigfield) (Eccles (Parsons) et al., 1983; Eccles and Wigfield, 2002; Wigfield and Eccles, 2000).

The model proposed by Eccles and her colleagues presupposes that expectancies and values are both important dimensions that underlie academic (and other types of) motivation. The expectancy component refers to individuals' beliefs about their abilities in particular areas or domains and how well they believe that they will perform in these domains both in the present and in the future. The value component consists of four types of beliefs: beliefs about the importance of a task (attainment value), beliefs about the enjoyment experienced by engaging with a task (intrinsic value), beliefs about the usefulness of a task toward reaching goals (utility value), and beliefs about what one must give up in order to engage with a task (costs) (Eccles (Parsons) et al., 1983; Wigfield and Eccles, 2000).

Numerous studies have examined relations between expectancies and values, the roles of expectancies and values in predicting achievement and behavior, and the development of expectancy and value beliefs and found strong support for the model. In general, expectancy beliefs are predictive of achievement and values are predictive of future behaviors (e.g., the intention to take a future course and actual enrollment in those courses in the future) (e.g., Durik et al., 2006). Results of longitudinal studies also indicate that on average expectancies and values decline as children move into and through adolescence (e.g., Wigfield and Eccles, 2001).

More recently, researchers integrated expectancy-value models with other theoretical frameworks. These studies have enhanced our understanding of the relations between specific types of motivational beliefs, and how those relations predict valued outcomes over time. For example, Hulleman and his colleagues (Hulleman et al., 2008) used EVT, achievement goal orientation theory, and work on interest to examine motivation and achievement across a semester in a college psychology course and among high school students enrolled in a summer football camp. Results indicated that when students began the course endorsing mastery (approach) goals, they were more likely to report both intrinsic value and utility value for the course during the middle of the semester; this in turn was predictive of interest in the course and final grades

(as measured at the end of the semester). Thus over time, achievement values mediate the relation between mastery goals and subsequent interest. In other work, Gray (2014) integrated EVT with optimal distinctiveness theory – demonstrating that STEM-focused high school students (i.e., those selected for specialized academic programs in Science, Technology, Engineering, and Mathematics) place more importance on academic tasks that help them both stand out from (uniqueness), and fit in with (similarity), their classmates.

Implications for Learning and Instruction

Researchers in recent years have begun to develop educational intervention studies framed within EVT. Studies focused on increasing expectations for success have existed for many years and have shown consistent support for the power of increasing expectations for success for subsequent academic performance, particularly among lower achieving students (Schunk et al., 2008; also see discussion later on social cognitive theories). More recently interventions have focused on increasing perceived task value. For example, in one recent study by Hulleman et al. (2010), college students were introduced to a method for solving a math problem, and then asked to write either an essay on the potential relevance (i.e., utility value) of a mathematical task, or an essay about two pictures that they were shown. Participants then were given time to work on a problem set, using the new technique. Those participants who wrote essays about relevance experienced enhanced situational interest, and reported being more likely to use the technique again in the future. These effects were particularly strong for participants who entered the study with low performance expectations.

In another study, 188 adolescents and their parents were mailed several brochures about the value of math and science courses and about decision-making during adolescence (Harackiewicz et al., 2012). Access also was provided to a Web site with information about STEM courses and careers. After 15 months, the mothers of the participants reported more positive perceptions of the value of STEM courses; conversations between participants and their parents about STEM courses increased; and the number of STEM courses in which students enrolled during the last 2 years of high school also increased.

Achievement Goal Theory

Students are driven to achieve for different reasons. According to achievement goal theory, the reasons why students strive to achieve pertain to their standards for judging their own competence (Ames and Archer, 1988; Dweck and Leggett, 1988; Elliot and Harackiewicz, 1996; Maehr and Anderman, 1993; Midgley et al., 1998; Pintrich, 2000a,b). Students consequently orient themselves toward meeting these competence standards. Students who endorse *mastery goals* are concerned with learning, growth, or understanding. In a sense, they are immersed in the achievement task itself and preoccupied with their own expertise within that domain. Students who endorse *performance goals* are concerned with their achievement relative to other individuals, or how competent they appear to others. In other words, they use norm-based standards or other individuals as the referent when measuring their own success (for a review, see Kaplan et al., 2002).

More specific forms of achievement goals may be subclassified under the scope of mastery and performance goals (Elliot, 1999). To the extent that individuals vary in their dispositions to approach success and avoid failure, students may endorse performance-approach goals, performance-avoidance goals, mastery-approach goals, and mastery-avoidance goals. A student with a performance-approach goal orientation therefore strives to demonstrate that she or he is smarter, more high-achieving, or more capable than other students. A student with a performance-avoidance goal orientation strives to demonstrate that she or he is no less smart, no less of an achiever, or no less capable than other students. A student with a mastery-approach goal strives to increase her or his understanding, whereas a student with a mastery-avoidance goal strives to avoid a lack of understanding or a loss of expertise. Mastery-approach goals are associated with positive consequences, such as deep cognitive processing of information performance-approach goals are associated with better academic performance, although some mixed findings exist; mastery-avoidance goals are associated with negative consequences, such as greater text anxiety and worry; and performance-avoidance goals are associated with worse academic performance, surface-level processing, and worry (Elliot, 2001; see chapter on Avoidance and Approach Motivation in this volume).

Recent findings have expanded our understanding of achievement goals and their functionality. Sideridis and Kaplan (2011), for example, assessed whether the endorsement of different achievement goals was associated with different patterns of persistence. Study participants were instructed to engage in a series of achievement tasks, which happened to be unsolvable. Though the time participants spent on task deteriorated with each trial, mastery-oriented students appeared most persistent across a series of tasks, and those who were performance-avoidance oriented, appeared the least persistent. After succeeding during one of the trials, however, performance-approach students appeared to demonstrate the strongest motivational recovery – as evidenced by the time they spent engaging in the task. More generally, Huang (2012) conducted a meta-analysis of 151 achievement goal studies and concluded that mastery-approach and performance-approach goals are associated with higher scholastic achievement, but mastery-avoidance and performance-avoidance goals are associated with lower scholastic achievement.

Madjar et al. (2011) reconceptualized mastery-avoidance goals as a two-dimensional construct. The authors describe *task-based mastery-avoidance goals* as a preoccupation with avoiding the loss of competence, and *intrapersonal mastery-avoidance goals* as a preoccupation with avoiding performing worse than one typically performs. In their research these two aspects of mastery-avoidance goals were differentially related with other achievement goals and student engagement outcomes. In particular, intrapersonal mastery-avoidance goals qualified associations between mastery-approach goals and students' reports of positive affect,

negative affect, and maladaptive strategy use. For two students with weak mastery-approach goals, the student with stronger intrapersonal mastery-avoidance goals is estimated to report more positive affect, less negative affect, and lower use of maladaptive learning strategies than the student with weaker intrapersonal mastery-avoidance goals. This effect is even more pronounced when comparing two students who both have strong mastery-approach goals. The findings of Madjar and colleagues highlight that the combination of intrapersonal mastery-avoidance goals and mastery-approach goals may have positive joint impacts on student engagement in ways that the combination of task-based mastery-avoidance goals and mastery-approach goals do not. In addition to these advances in achievement goal theory, further subclassifications of mastery goals also have been proposed (Elliot et al., 2011) – suggesting that individuals may measure success either by success or failure on a task, or in terms to success or failure relative to one's previous performance.

One explanation for why students' achievement goals may differ from context to context is the motivational climate teachers set in their classrooms (Ames, 1992; Ames and Ames, 1984). Teachers who place a premium on learning, effort, and improvement are promoting a *mastery goal structure* in their classrooms. In these mastery-focused environments, students perceive their classroom culture as one in which constant improvement is encouraged and where mistakes are not seen as an indication of incompetence, but are seen as part of the learning process. Students benefit from mastery-focused learning environments – exhibiting fewer self-handicapping behaviors, greater persistence, more effective metacognitive learning strategies, and adaptive help-seeking behaviors (Elliott and Dweck, 1988; Elliot and MacGregor, 2001; Meece and Miller, 2001; Midgley and Urdan, 2001; Ryan and Pintrich, 1997; Stipek and Kowalski, 1989). Teachers who foster a sense of competition among students, or compare high-performing students with other students in the same classroom are promoting a *performance goal structure*. Generally, performance goal structures are linked with a higher prevalence of unfavorable student outcomes including self-handicapping and academic cheating (E. Anderman et al., 1998; Urdan et al., 1998). In these performance-focused environments, students perceive their classroom culture as one in which a student's worth depends on her or his ability or achievement. Messages conveyed by teachers about mastery and performance thus appear to shape learning outcomes directly, as well as indirectly through students' achievement goals (Anderman and Wolters, 2006).

However, students' perceptions of the classroom motivational climate and the classroom social climate appear to be interwoven. Across four studies, Patrick et al. (2011) found that students' perceptions of a mastery goal structure overlapped substantially with their perceptions of teacher emotional and academic support, and the promotion of mutual respect and task-related interaction. Patrick and Ryan (2008) investigated students' interpretations of mastery goal structure items. After responding to mastery goal structure scale items, students were instructed to provide examples of specific teacher behaviors that supported their Likert responses. Results revealed that teachers' pedagogical and affective interactions with students contributed to perceptions of a mastery-focused classroom environment. L. Anderman et al. (2011) took an inductive approach to understand effective teacher practices in the classroom. Over 4000 students were asked to respond to surveys about the instructional climate of their classroom. Survey responses were used to identify teachers who fostered a supportive instructional climate. Classroom observations revealed that these effective teachers supported students' mastery, built and maintained a rapport with students, and managed the classroom. Relatedly, Turner et al. (2013) found that students' perceptions of mastery goal structure and teacher support converge over the course of a semester. The authors proposed a *Changed Perceptions* model of mastery goal structure in which, over the course of a semester, students' perceptions of mastery goal structure would become integrated with their perceptions of teacher support. After following the same students for 2 years across their transition into middle school, the authors found support for this Changed Perceptions model when students were in elementary school, and replicated these results when the same students were in middle school. What the findings of Patrick and colleagues, L. Anderman and colleagues, and Turner and colleagues share is the conclusion that students do not appear to separate perceptions of a mastery-focused classroom environment from the social climate fostered by the teacher.

Implications for Learning and Instruction

There are practical lessons that goal theory brings to bear on the processes of learning and instruction. As a student's perception of goal structures stems from a teacher's messages about the demonstration and development of competence in her or his classroom, goal structures are messages that are delivered alongside the course content itself. For this reason, practices that encourage certain types of goal structures may be manifest in a variety of academic subject areas, and are not necessarily bound to a particular academic subject. For example, conveying to students that their classroom is a safe place to take intellectual risks is a practice that is transferrable to a variety of content areas. Even in nontraditional academic subjects, mastery and performance goal structures are linked with learning. For example, in the context of health education classes, Anderman et al. (2011) found among other outcomes, that higher student perceptions of mastery goal structure were associated with an increase in knowledge about HIV and pregnancy prevention over the course of a semester, and retention of this knowledge a year later. By contrast, performance goal structure was related negatively to students' HIV and pregnancy prevention knowledge both at the end of the semester and at follow-up. Even beyond traditional academic subjects, it appears that instructional practices that foster perceptions of a mastery goal structure can bear directly on important learning outcomes among students.

In terms of the impact of achievement goals on the learning process, Barron and Harackiewicz (2001) noted that mastery goals – when accompanied by performance-approach goals – may facilitate optimal student outcomes. Proponents of

a multiple goals perspective argue that both types of goals are of practical significance in their own right – with mastery goals being related to the use of adaptive learning strategies, and performance-approach goals being related to achievement (Elliot, 1999; Pintrich, 2000b). Proponents of this perspective have recently demonstrated consistent positive correlations between performance-approach goals and achievement, even when controlling for perceptions of competence and measures of ability (Senko et al., 2011). Although it is true that more research is needed in understanding how individuals represent these two types of goals in their minds at the same time, the multiple goals perspective continues to gain currency in contemporary achievement goal theory research.

Self-Determination Theory

Rewards can undermine an individual's natural curiosity and interest. According to SDT (Ryan and Deci, 2000), humans can be drawn to engage in activities because the activity itself is appealing, enjoyable, and fulfilling. This *intrinsic motivation* is seen as an ideal type of motivation in the sense that there are no contingencies associated with the person's engagement in the activity. However, money, recognition, status, and other rewards can shape human behavior such that the same individual who once might have engaged in an activity due to interest now engages in that behavior because of the benefits that are gained by doing so. Rewards can contaminate intrinsic motivation in the sense that the person's level of engagement in the activity would diminish if she or he were no longer able to receive external benefits for engaging in the behavior (Deci and Ryan, 1985; Plant and Ryan, 1985).

The distinction between intrinsic and extrinsic motivation, at times, may represent a false dichotomy with respect to how human behavior is energized and directed (Ryan and Deci, 2008). *Self-determined motivation* (also referred to as autonomous motivation) is a combination of the intrinsic and extrinsic regulatory mechanisms of interest (intrinsic regulation), internalization (integrated regulation), and values (identified regulation). On the other hand, *controlled motivation* represents behavior that is dictated by emotions imposed by others (introjected regulation), such as pride, shame, or guilt, or extrinsic rewards (external regulation). Self-determined motivation is seen as a higher quality form of motivation than controlled motivation in the sense that the forces driving an individual's behavior are all internal in nature, even if some of these forces are secondary to the enjoyment of the task itself (Ryan and Deci, 2000).

Self-determined motivation can be facilitated by contexts that nurture basic human needs for autonomy, competence, and relatedness (Deci et al., 1991). *Autonomy* refers to an individual's perception of choice (i.e., engaging in behaviors because of her or his own volition). *Competence* refers to beliefs in one's capabilities. *Relatedness* refers to one's interpersonal connections with others. In a classroom environment, SDT explains that students whose needs for autonomy, competence, and relatedness are fulfilled through interactions with others in that context will be more self-determined in their achievement behavior.

Recent SDT research has focused on assessing the synergism of different types of motivations. Vansteenkiste et al. (2009) examined profiles of autonomous and controlled motivation among high school and college students. In the high school sample, the authors labeled students, who were high in autonomous motivation yet low in controlled motivation, the *high-quality motivation* profile. Relative to other profiles, this profile of individuals reported greater cognitive, metacognitive, and behavioral outcomes than students with other motivational profiles. In addition to replicating this pattern in the college sample, the authors also found that college students with this profile also perceived their teachers as higher in autonomy support, structure, and teacher involvement. In a similar study of middle school students, Hayenga and Corpus (2010) examined profiles of intrinsic and extrinsic motivation among middle school students. The authors labeled students, who were high in intrinsic motivation yet low in extrinsic motivation, the *high-quality* profile. Relative to other profiles, students in the high-quality profile exhibited superior academic performance. Results also revealed shifts in profile stability over the course of the academic year – with more students shifting toward poor-quality motivation profiles (low-intrinsic, high-extrinsic motivation) over the course of the academic year. In a related study, Wormington et al. (2012) examined profiles of intrinsic, introjected, and external regulation in a sample of high school students. The authors labeled students who were high in intrinsic regulation relative to their introjected and external regulation the *good-quality* profile, and labeled students who were high in intrinsic, introjected, and external regulation the *high-quantity* profile. Membership in these two profiles was associated with high levels of performance and more extracurricular activity participation. However, students who were members of these two profiles did not significantly differ from one another in terms of academic performance and extracurricular activity participation. The authors explain this finding as an advance in SDT research – contending that controlled forms of regulation may perhaps be adaptive in highly controlled high school environments. Overall, work of Vansteenkiste and colleagues, Hayenga and Corpus, and Wormington and colleagues moves toward a clearer understanding of how other types of motivations may bolster or suppress the effect of intrinsic motivation on positive student outcomes.

Implications for Learning and Instruction

Research under the scope of SDT has implications for the extent to which basic psychological needs jointly affect the extent to which students are driven to learn and are able to achieve. Research demonstrates that students consistently benefit when their needs for competence, relatedness, and autonomy are met. For example, students are more motivated to do homework when their basic psychological needs are fulfilled (Katz et al., 2010). Reviews of research on relatedness indicate that students who perceive stronger interpersonal relationships are more motivated to achieve in school (e.g., Martin and Dowson, 2009). Students are more intrinsically motivated when their competence needs are supported (Schunk and Zimmerman, 2006), but the support of competence should be accompanied by support for autonomy (Niemiec and Ryan, 2009). With

respect to the needs for autonomy and competence, students are more likely to achieve, be engaged, be more intrinsically motivated, and experience less negative affect when these needs are fulfilled (Jang et al., 2009). In addition to these joint effects, autonomy support itself is linked with a host of student outcomes including greater academic motivation, positive emotions, persistence, information processing, well-being, and academic performance (Reeve, 2009).

Social Cognitive Theory

SCT, which was popularized by the work of Albert Bandura, is broad in scope, and encompasses a number of concepts and constructs related to the processes of motivation and learning. The theory emphasizes the social interactions between individuals and various aspects of the environments in which they live. The tenets of SCT are far-reaching; the theory and its correlates have served as the basis for work in education, as well as in numerous other fields including human nutrition (Lubans et al., 2012), exercise science (Fuller et al., 2012), and career development (Zikic and Saks, 2009).

Basic Principles of SCT

There are some core ideas that are central to all aspects of SCT. Whereas some of these are more central to learning and instruction than are others, all represent important components of the theory. A thorough understanding of SCT and its implications for learning and instruction are prerequisite to the development of effective SCT-based interventions.

Triadic Reciprocity

First and foremost is the notion of triadic reciprocity (also known as *reciprocal determinism*). Bandura (1986, 1997) posits that there are three main determinants of human agency: behaviors, personal (internal) factors (including cognitions), and the external environment. Human learning and performance are viewed broadly as functions of the interplay and joint influence of these three determinants.

Bandura (1997) cites *depression* as an exemplar of triadic reciprocity. Specifically, depression can be viewed as being caused by interactions between the environment (e.g., experiencing unpleasant or difficult events in one's life), personal factors (e.g., how those unpleasant experiences are interpreted by the individual), and behaviors (e.g., inability to nurture and develop positive relationships with others, which in turn leads to negative reactions from others, which can exacerbate depressive symptoms).

Self-Efficacy

Self-efficacy is a prominent construct within SCT and is quite similar to the expectancy component of EVT. Individuals who are self-efficacious believe that they can be successful at engaging with specific tasks (Bandura, 1997; Pajares, 1996; Schunk, 1989, 1991; Schunk and Pajares, 2002). Self-efficacy is related positively to academic achievement, the use of effective learning strategies, and effort (Schunk and Pajares, 2009). Self-efficacy beliefs can be focused on highly specialized tasks (e.g., one's feelings of self-efficacy at being able to solve a specific algebra problem), or toward more generalized domains (e.g., self-efficacy in mathematics). Researchers have studied the relations of self-efficacy to valued outcomes in numerous domains, including literacy (Schunk and Zimmerman, 2007), mathematics (Fast et al., 2010), and teaching (Ashton and Webb, 1986; Guskey and Passaro, 1994; Tschannen-Moran et al., 1998).

Observational Learning

An important tenet in Bandura's work is that individuals learn from observing others. We can learn from individuals with whom we interact, as well as from individuals whom we are exposed to through television and other media sources, including social media (e.g., Facebook, Twitter, YouTube, etc.). In a classic study, Bandura and his colleagues (Bandura et al., 1963) demonstrated that when children observed models acting aggressively with a doll, those children were more likely to display similar aggressive behaviors when subsequently presented with a doll.

Implications for Learning and Instruction

SCT offers many important implications for learning and instruction. These implications derive from both theoretical tenets of SCT, as well as from years of empirical evidence drawn from applied SCT research. Particularly important are the implications for learning via technology. The major implications are discussed below.

Learning Involves Interactions between Individuals, within Social Contexts

Referring back to the notion of triadic reciprocity, SCT implies that effective learning and teaching involves considerations of the interplay between behaviors, environments, and personal factors. From a design perspective, it is essential to consider all three of these influences. SCT-based work that has emphasized triadic reciprocity has been useful in many domains, including specific areas such as drug and alcohol education (Sharma, 2005), and in more generalized areas, such as studies of self-regulation in learners (Zimmerman and Schunk, 2004).

The implications for instruction are plentiful, and researchers in recent years have begun to examine SCT and triadic reciprocity in new, technology-influenced learning environments (Joe and Lin, 2008; Lin, 2010). Learners interact (behave) within various contexts (environments), and all learners bring their own personal characteristics (beliefs, attitudes, prior knowledge, etc.) into those situations. Notwithstanding its complexity, a consideration of all three of these contexts can allow for the development of optimal learning and teaching opportunities. In particular, when one considers the vastly growing new areas of learning within online and virtual environments, the notion of triadic reciprocity plays an extremely important role, particularly since the environment is now virtual; indeed, the environment is multifaceted, consisting of the software, the hardware, the form of delivery (e.g., video vs audio), the levels of interactivity that are available, and the settings in which the individuals engage in the online learning (e.g., in their homes vs in classrooms).

Learning is Enhanced when Self-Efficacy Beliefs are Supported

As aforementioned, self-efficacy is a strong predictor of learning and achievement. There has been much research indicating that self-efficacy beliefs can be enhanced, thus improving learning and achievement. In recent years, some of the most exciting work in this area has occurred in the medical field. Researchers have demonstrated that self-efficacy beliefs can be enhanced, and that these changes can lead to improved health outcomes. For example, researchers have demonstrated that the use of peer-mentors with newly diagnosed spinal cord injury patients can improve patients' self-efficacy over the course of the first year after diagnosis (Ljungberg et al., 2011). In other studies, researchers have demonstrated that condom self-efficacy in adolescent populations can be enhanced via interventions (Chen et al., 2012).

One of the most well-researched methods for enhancing self-efficacy beliefs is to encourage learners to set proximal (as opposed to distal) goals. Research indicates that when learners set reachable goals and experience success at reaching those goals, efficacy beliefs are enhanced (Bandura and Schunk, 1981). Initial applications of this research were found in the field of special education, where self-efficacy was enhanced in students with learning disabilities when the students were taught to set proximal goals (Schunk, 1985). In recent years, this basic finding has been applied in other fields, including business administration (Brown and Warren, 2009).

Learning is Enhanced through Observation

Much of Bandura's early work on SCT focused on observational learning. Individuals can learn much from observing others. Researchers have applied this work in important ways in recent years. For example, researchers have demonstrated that observation of others' behaviors can lead to improved safety behaviors (Olson et al., 2009). In addition, research indicates that interventions involving observation and imitation can be used to improve outcomes in patients with schizophrenia (Mazza et al., 2010).

Conclusion

In this article, we have examined four theoretical frameworks that have been used in recent years to study and explain student motivation. Each of these perspectives, and others historically addressed in reviews of motivation research (e.g., Weiner's (1985) attribution theory), offers important implications for the design of instruction. Although it is tempting to think of motivation as primarily residing within the student, self-efficacy beliefs, value for learning, self-determination, and achievement goals are informed by a host of factors – including prior experiences with others in a classroom setting and the instructional practices of teachers. Moreover, the study of motivation and learning is left wanting if not considered as being interconnected with identity processes (Eccles [Parsons] et al., 1983; Kaplan and Flum, 2009) and achievement-related emotions (Eccles and Wigfield, 2002; Pekrun, 2006) – two factors that are implicit in existing motivation frameworks, yet they possess a certain theoretical richness in their own right. A broader understanding by educators of the effects of these variables on students may enhance achievement in important ways, and spark and sustain interest in academic domains that have been identified as areas of high priority (e.g., Science, Technology, Engineering, and Mathematics).

See also: Academic Motivation and Performance: Task Value Interventions; Avoidance and Approach Motivation: A Brief History; Gender and Academic Motivation; Interest, Psychology of; Internal/External Frame of Reference Model; Mastery Learning; Motivation and Actions, Psychology of; Motivation in Australian Aboriginal Populations; Race and Academic Motivation; School Achievement: Motivational Determinants and Processes; Schooling: Impact on Cognitive and Motivational Development; Self-Determination Theory; Teacher Motivation; Test Anxiety and Academic Achievement.

Bibliography

Atkinson, J.W., 1964. An Introduction to Motivation. Van Nostrand, Princeton, NJ.
Ames, C., 1992. Classrooms: goals, structures, and student motivation. Journal of Educational Psychology 84, 261–271.
Ames, C., Ames, R., 1984. Student motivation. In: Ames, R. (Ed.), Research on Motivation in Education, vol. 1. Academic Press, Orlando.
Ames, C., Archer, J., 1988. Achievement goals in the classroom: student learning strategies and motivation processes. Journal of Educational Psychology 80, 260–267.
Anderman, E.M., Cupp, P.K., Lane, D.R., Zimmerman, R., Gray, D.L., O'Connell, A.A., 2011. Classroom goal structures and HIV/pregnancy prevention education in rural high school health classrooms. Journal of Research on Adolescence 21 (4), 904–922.
Anderman, L.H., Andrzejewski, C.E., Allen, J., 2011. How do teachers support students' motivation and learning in their classrooms? Teachers College Record 113, 969–1003.
Anderman, E.M., Gray, D.L., Chang, Y., 2012. Motivation and classroom learning. In: Reynolds, W.M., Miller, G.E. (Eds.), Handbook of Psychology, second ed., vol. 7. John Wiley & Sons, Hoboken, pp. 99–116.
Anderman, E.M., Griesinger, T., Westerfield, G., 1998. Motivation and cheating during early adolescence. Journal of Educational Psychology 90 (1), 84–93. http://dx.doi.org/10.1037/0022-0663.90.1.84.
Anderman, E.M., Wolters, C.A., 2006. Goals, values, and affect: influences on student motivation. In: Alexander, P.A., Winne, P.H. (Eds.), Handbook of Educational Psychology. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, pp. 369–389.
Ashton, P.T., Webb, R.B., 1986. Making a Difference: Teachers' Sense of Efficacy and Student Achievement. Longman, New York.
Bandura, A., 1986. Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice Hall, Englewood Cliffs.
Bandura, A., 1997. Self-Efficacy: The Exercise of Control. W.H. Freeman and Company, New York.
Bandura, A., Schunk, D.H., 1981. Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. Journal of Personality and Social Psychology 41 (3), 586–598.
Bandura, A., Ross, D., Ross, S.A., 1963. Imitation of film-mediated aggressive models. Journal of Abnormal and Social Psychology 66, 3–11.
Barron, K.E., Harackiewicz, J.M., 2001. Achievement goals and optimal motivation: testing multiple goal models. Journal of Personality and Social Psychology 80, 706–722.
Brown, T.C., Warren, A.M., 2009. Distal goal and proximal goal transfer of training interventions in an executive education program. Human Resource Development Quarterly 20, 265–284.
Chen, X., Dinaj-Koci, V., Brathwaite, N., Cottrell, L., Deveaux, L., Gomez, P., et al., 2012. Development of condom-use self-efficacy over 36 months among early adolescents: a mediation analysis. The Journal of Early Adolescence 32, 711–729.
Deci, E.L., Ryan, R.M., 1985. Intrinsic Motivation and Self-Determination in Human Behavior. Plenum, New York.
Deci, E.L., Vallerand, R.J., Pelletier, L.G., Ryan, R.M., 1991. Motivation and education: the self-determination perspective. The Educational Psychologist 26, 325–346.
Durik, A.M., Vida, M., Eccles, J.S., 2006. Task values and ability beliefs as predictors of high school literary choices: A developmental analysis. Journal of Educational Psychology 98 (2), 382–393.

Dweck, C.S., Leggett, E.L., 1988. A social-cognitive approach to motivation and personality. Psychological Review 95, 256–273.

Eccles (Parsons), J., Adler, T.F., Futterman, R., Goff, S.B., Kaczala, C.M., Meece, J.L., et al., 1983. Expectancies, values, and academic behaviors. In: Spence, J.T. (Ed.), Achievement and Achievement Motivation. Freeman, San Francisco, pp. 75–146.

Eccles, J.S., Wigfield, A., 2002. Motivational beliefs, values, and goals. Annual Review of Psychology 53 (1), 109–132.

Elliot, A.J., 1999. Approach and avoidance motivation and achievement goals. Educational Psychologist 34, 169–189.

Elliot, E.S., Dweck, C.S., 1988. Goals: An approach to motivation and achievement. Journal of Personality and Social Psychology 54 (1), 5–12.

Elliot, A.J., Harackiewicz, J.M., 1996. Approach and avoidance achievement goals and intrinsic motivation: a mediational analysis. Journal of Personality and Social Psychology 70 (3), 461–475.

Elliot, A.J., McGregor, H.A., 2001. A 2 × 2 achievement goal framework. Journal of Personality and Social Psychology 80 (3), 501–519. http://dx.doi.org/10.1037/0022-3514.80.3.501.

Elliot, A.J., Murayama, K., Pekrun, R., 2011. A 3 × 2 achievement goal model. Journal of Educational Psychology 103, 632–648.

Fast, L.A., Lewis, J.L., Bryant, M.J., Bocian, K.A., Cardullo, R.A., Rettig, M., et al., 2010. Does math self-efficacy mediate the effect of the perceived classroom environment on standardized math test performance? Journal of Educational Psychology 102 (3), 729–740.

Fuller, D.L., Gyurcsik, N.C., Spink, K.S., Brawley, L.R., 2012. Prospective examination of self-regulatory efficacy in predicting walking for active transportation: a social cognitive theory approach. Journal of Applied Social Psychology 42, 2917–2932.

Guskey, T.R., Passaro, P.D., 1994. Teacher efficacy: a study of construct dimensions. American Educational Research Journal 31 (3), 627–643.

Gray, D.L., 2014. Understanding STEM-focused high school students' perceptions of task importance: the role of "standing out" and "fitting in" in mathematics class. Contemporary Educational Psychology 39, 29–41.

Harackiewicz, J.M., Rozek, C.S., Hulleman, C.S., Hyde, J.S., 2012. Helping parents to motivate adolescents in mathematics and science: an experimental test of a utility-value intervention. Psychological Science 23, 899–906.

Hayenga, A.O., Corpus, J.H., 2010. Profiles of intrinsic and extrinsic motivation: a person-centered approach to motivation and achievement in middle school. Motivation and Emotion 34, 371–383.

Huang, C., 2012. Discriminant and criterion-related validity of achievement goals in predicting academic achievement: a meta-analysis. Journal of Educational Psychology 104, 48–73.

Hulleman, C.S., Durik, A.M., Schweigert, S.B., Harackiewicz, J.M., 2008. Task values, achievement goals, and interest: an integrative analysis. Journal of Educational Psychology 100, 398–416.

Hulleman, C.S., Godes, O., Hendricks, B.L., Harackiewicz, J.M., 2010. Enhancing interest and performance with a utility value intervention. Journal of Educational Psychology 102, 880–895.

Jang, H., Reeve, J., Ryan, R.M., Kim, A., 2009. Can self-determination theory explain what underlies the productive, satisfying learning experiences of collectivistically oriented Korean students? Journal of Educational Psychology 101, 644–661.

Joe, S.-W., Lin, C.-P., 2008. Learning online community citizenship behavior: a socio-cognitive model. CyberPsychology & Behavior 11 (3), 367–370.

Kaplan, A., Flum, H., 2009. Motivation and identity: the relations of action and development in educational contexts—an introduction to the special issue. Educational Psychologist 44 (2), 73–77. http://dx.doi.org/10.1080/00461520902832418.

Kaplan, A., Middleton, M.J., Urdan, T., Midgley, C., 2002. Achievement goals and goal structures. In: Midgley, C. (Ed.), Goals, Goal Structures, and Patterns of Adaptive Learning. Erlbaum Associates, Mahwah, NJ, pp. 21–54.

Katz, I., Kaplan, A., Guetta, G., 2010. Students' needs, teachers' support, and motivation for doing homework: a cross-sectional study. Journal of Experimental Education 78, 246–267.

Lewin, K., 1935. A dynamic theory of personality: Selected papers. McGraw-Hill, New York.

Lin, C.-P., 2010. Learning virtual community loyalty behavior from a perspective of social cognitive theory. International Journal of Human-Computer Interaction 26, 345–360.

Ljungberg, I., Kroll, T., Libin, A., Gordon, S., 2011. Using peer mentoring for people with spinal cord injury to enhance self–efficacy beliefs and prevent medical complications. Journal of Clinical Nursing 20, 351–358.

Lubans, D.R., Plotnikoff, R.C., Morgan, P.J., Dewar, D., Costigan, S., Collins, C.E., 2012. Explaining dietary intake in adolescent girls from disadvantaged secondary schools. A test of social cognitive theory. Appetite 58, 517–524.

Madjar, N., Kaplan, A., Weinstock, M., 2011. Clarifying mastery-avoidance goals in high school: distinguishing between intrapersonal and task-based standards of competence. Contemporary Educational Psychology 36, 268–279.

Maehr, M.L., Anderman, E.M., 1993. Reinventing schools for early adolescents: emphasizing task goals. Elementary School Journal 93 (5), 593–610.

Martin, A.J., Dowson, M., 2009. Interpersonal relationships, motivation, engagement, and achievement: yields for theory, current issues, and practice. Review of Educational Research 79, 327–365.

Mazza, M., Lucci, G., Pacitti, F., Pino, M.C., Mariano, M., Casacchia, M., et al., 2010. Could schizophrenic subjects improve their social cognition abilities only with observation and imitation of social situations? Neuropsychological Rehabilitation 20, 675–703.

Meece, J.L., Miller, S.D., 2001. A longitudinal analysis of elementary school students' achievement goals in literacy activities. Contemporary Educational Psychology 26 (4), 454–480. http://dx.doi.org/10.1006/ceps.2000.1071.

Midgley, C. (Ed.), 2002. Goals, Goal Structures, and Patterns of Adaptive Learning. Lawrence Erlbaum Associates, Mahwah, NJ.

Midgley, C., Kaplan, A., Middleton, M., Maehr, M.L., Urdan, T., Anderman, L.H., Anderman, E.M., Roeser, R., 1998. The development and validation of scales assessing students' achievement goal orientations. Contemporary Educational Psychology 23, 113–131.

Midgley, C., Urdan, T., 2001. Academic self-handicapping and achievement goals: a further examination. Contemporary Educational Psychology 26 (1), 61–75. http://dx.doi.org/10.1006/ceps.2000.1041.

Niemiec, C.P., Ryan, R.M., 2009. Autonomy, competence, and relatedness in the classroom: applying self-determination theory to educational practice. Theory and Research in Education 7, 133–144.

Olson, R., Grosshuesch, A., Schmidt, S., Gray, M., Wipfli, B., 2009. Observational learning and workplace safety: the effects of viewing the collective behavior of multiple social models on the use of personal protective equipment. Journal of Safety Research 40, 383–387.

Pajares, F., 1996. Self-efficacy beliefs in academic settings. Review of Educational Research 66 (4), 543–578.

Patrick, H., Kaplan, A., Ryan, A.M., 2011. Positive classroom motivational environments: Convergence between mastery goal structure and classroom social climate. Journal Of Educational Psychology 103 (2), 367–382. http://dx.doi.org/10.1037/a0023311.

Patrick, H., Ryan, A.M., 2008. What do students think about when evaluating their classroom's mastery goal structure? An examination of young adolescents' explanations. The Journal of Experimental Education 77, 99–124.

Pekrun, R., 2006. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. Educational Psychology Review 18 (4), 315–341. http://dx.doi.org/10.1007/s10648-006-9029-9.

Pintrich, P.R., 2000a. The role of goal orientation in self-regulated learning. In: Boedaerts, M., Pintrich, P., Zeidner, M. (Eds.), Handbook of Self-Regulation: Theory, Research and Applications. Academic, San Diego, California.

Pintrich, P.R., 2000b. An achievement goal theory perspective on issues in motivation terminology, theory, and research. Contemporary Educational Psychology 25, 92–104.

Plant, R.W., Ryan, R.M., 1985. Intrinsic motivation and the effects of self-consciousness, self-awareness, and ego-involvement: an investigation of internally-controlling styles. Journal of Personality 53, 435–449.

Ryan, R.M., Deci, E.L., 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist 55, 68–78.

Ryan, R.M., Deci, E.L., 2008. From ego-depletion to vitality: theory and findings concerning the facilitation of energy available to the self. Social and Personality Psychology Compass 2, 702–717.

Ryan, A.M., Pintrich, P.R., 1997. 'Should I ask for help?' The role of motivation and attitudes in adolescents' help seeking in math class. Journal Of Educational Psychology 89 (2), 329–341. http://dx.doi.org/10.1037/0022-0663.89.2.329.

Schunk, D.H., 1985. Participation in goal setting: effects on self-efficacy and skills of learning-disabled children. The Journal of Special Education 19 (3), 307–317.

Schunk, D.H., 1989. Self-efficacy and achievement behaviors. Educational Psychology Review 1, 173–208.

Schunk, D.H., 1991. Self-efficacy and academic motivation. Educational Psychologist 26, 207–231.

Schunk, D.H., Pajares, F., 2002. The development of academic self-efficacy. In: Wigfield, A., Eccles, J.S. (Eds.), Development of Achievement Motivation. A Volume in the Educational Psychology Series. Academic Press, San Diego, CA, pp. 15–31.

Schunk, D.H., Pajares, F., 2009. Self-efficacy theory. In: Wentzel, K.R., Wigfield, A. (Eds.), Handbook of Motivation at School. Routledge, New York, pp. 35–53.

Schunk, D.H., Pintrich, P.R., Meece, J.,L., 2008. Motivation in Education, third ed. Pearson Merrill Prentice Hall, Upper Saddle River, NJ.

Schunk, D., Zimmerman, B.J., 2007. Influencing children's self-efficacy and self-regulation of reading and writing through modeling. Reading & Writing Quarterly: Overcoming Learning Difficulties 23 (1), 7–25.

Senko, C.M., Hulleman, C.S., Harackiewicz, J.M., 2011. Achievment goal theory at the crossroads: old controversies, current challenges, and new directions. Educational Psychologist 46, 26–47.

Sharma, M., 2005. Enhancing the effectiveness of alcohol and drug education programs through social cognitive theory. Journal of Alcohol and Drug Education 49 (3), 3–7.

Sideridis, G.D., Kaplan, A., 2011. Achievement goals and persistence across tasks: the roles of failure and success. The Journal of Experimental Education 79, 429–451.

Stipek, D.J., Kowalski, P.S., 1989. Learned helplessness in task-orienting versus performance-orienting testing conditions. Journal Of Educational Psychology 81 (3), 384–391. http://dx.doi.org/10.1037/0022-0663.81.3.384.

Turner, J.C., Gray, D.L., Anderman, L.H., Dawson, H.S., Anderman, E.M., 2013. Getting to know my teacher: Does the relation between perceived mastery goal structures and perceived teacher support change across the school year? Contemporary Educational Psychology 38 (4), 316–327. http://dx.doi.org/10.1016/j.cedpsych.2013.06.003.

Tschannen-Moran, M., Woolfolk Hoy, A., Hoy, W.K., 1998. Teacher efficacy: its meaning and measure. Review of Educational Research 68 (2), 202–248.

Urdan, T., Midgley, C., Anderman, E.M., 1998. The role of classroom goal structure in students' use of self-handicapping strategies. American Educational Research Journal 31 (1), 101–122. http://dx.doi.org/10.2307/1163453.

Vansteenkiste, M., Sierens, E., Soenens, B., Luyckx, K., Lens, W., 2009. Motivational profiles from a self-determination perspective: the quality of motivation matters. Journal of Educational Psychology 101, 671–688.

Weiner, B., 1985. An attribution theory of achievement motivation and emotion. Psychological Review 92, 548–573.

Wigfield, A., Eccles, J.S., 2000. Expectancy-value theory of achievement motivation. Contemporary Educational Psychology 25 (1), 68–81.

Wigfield, A., Eccles, J.S., 2001. The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In: Wigfield, A., Eccles, J.S. (Eds.), Development of Achievement Motivation. A volume in the Educational Psychology series. Academic Press, San Diego, CA, pp. 21–120.

Wormington, S.V., Corpus, J.H., Anderson, K.G., 2012. A person-centered investigation of academic motivation and its correlates in high school. Learning and Individual Differences 22, 429–438.

Zikic, J., Saks, A.M., 2009. Job search and social cognitive theory: the role of career-relevant activities. Journal of Vocational Behavior 74, 117–127.

Zimmerman, B.J., Schunk, D.H., 2004. Self-Regulating Intellectual Processes and Outcomes: A Social Cognitive Perspective on Motivation, Emotion, and Cognition: Integrative Perspectives on Intellectual Functioning and Development. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, pp. 323–349.

Motivation: History of the Concept

Peter M Gollwitzer, University of Konstanz, Konstanz, Germany; and New York University, New York, NY, USA
Gabriele Oettingen, University of Hamburg, Hamburg, Germany; and New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Originally, motivation was analyzed as a singular determinant of human thoughts, feelings, and actions. It was quickly recognized, however, that motivation operates in concert with other determinants (e.g., cognition, affect, habits). This insight has allowed the psychology of motivation to progress to a very sophisticated level to answer to the question as to which of the human needs are to be differentiated and how these needs manage to guide and energize people. It has also promoted considering the concept of goals (on top of incentives and expectations) which allowed a better understanding of action control; willpower is nowadays analyzed in terms of the effective (effortful and/or automatic) self-regulation of goal pursuit.

Evolutionary theory and the psychoanalytic study of mental illness always were committed to seeing motivation as a primary cause of behavior, and this is also true of behavioral biology and physiological psychology, as they prefer to think in terms of homeostasis (Cofer and Appley, 1964). When hunger occurs, for instance, behavior is instigated (e.g., finding and eating food) that is instrumental to rectifying the imbalance caused by tissue needs and deficits. The named theoretical perspectives have barricaded the simple insight that behavior can occur under externally applied forces as well (e.g., a shove). Even more importantly, the physical structure of the organism, its sensory and perceptual capacities, its cognitive and motor abilities, and so forth, qualify as causal factors. Even habits, once formed, can be seen as such.

Even though motivation has been considered to be the primary cause of people's thoughts, feelings, and actions in evolutionary theory, psychoanalytic theory, behavioral psychology, and physiological psychology, this is not true for the psychology of motivation. There, motivation is just one cause among others, and the focus is on analyzing aspects of thoughts, feelings, and actions conducive to a motivational analysis. These aspects relate to the energization of responses as well as to their selection and persistent implementation. Answers to these questions have changed over time in line with the adhered-to *metamodel* of human functioning. Starting out with mechanistic models (e.g., learning theory), the psychology of motivation moved on to cognitive models (e.g., attribution theory), and has now arrived at self-regulation models that characterize the human as a flexible strategist when it comes to translating motivation into action.

More specifically, the question of the facilitation and energization of certain responses has been analyzed most extensively by Clark L. Hull (1943, 1884–1952) and Kenneth W. Spence (1956, 1907–67), the classic proponents of learning theory. It was suggested that motivational processes as such (assumed to be rooted in general drive or arousal states) do not necessarily control or guide novel forms of behavior but only invigorate innate or learned associative tendencies. Later on, thoughts, feelings, and actions were often characterized as guided, directed, goal-oriented, persistent, or purposive. Such qualities relate to making different choices, but also to unique short-term or long-term efforts in implementing the choices made. Researchers like John W. Atkinson (1957, 1923–2003) and David McClelland (1955, 1917–98) focused on this issue also suggesting a motivational analysis.

To demonstrate how the concept of motivation has changed in history, we describe how two central issues (i.e., basic needs and action control) have been addressed differently by the traditional and the modern psychology of motivation.

Basic Human Needs

Whereas learning theorists, following the lead of Hull, conceived of need as a general, content-free drive, personality psychologists were eager to put content into the concept of need. Sigmund Freud was rather parsimonious by only suggesting two basic needs, namely the life and the death instinct. William McDougall (1932, 1871–1938) listed 18 basic needs which he referred to as instincts (e.g., curiosity, self-assertion, submission). Henry Murray's *Explorations in Personality* (1938, 1893–1988) contains a catalog of 20 psychogenic needs (e.g., need for nurturance, need for succorance) and Abraham H. Maslow's *Motivation and Personality* (1954, 1908–70) proposed a hierarchy of needs whereby the lower need categories related to deficiency needs (e.g., hunger, safety) and the higher ones to growth needs (e.g., need to achieve, need to realize one's potential).

The named approaches excelled in generating sophisticated definitions and descriptions, but did not provide reliable instruments for need assessment. They also failed to answer the question of which of the many needs are activated in a given situation and how such activation processes run off. Moreover, they did not explore the origins of individual differences in the strength of these needs. In later years, Atkinson (1957) and McClelland (1985) addressed these problems with respect to the need for achievement and the power motive.

Still, the question remained: Which of the many conceivable human needs is the most basic? Baumeister and Leary (1995) have raised this issue by applying criteria such as universality, nonderivativeness, satiation, and substitution. According to their analysis, the need to belong in the sense of forming and maintaining strong, stable interpersonal relationships is suggested to be a most basic need. However, this need seems to be in tension with another basic need, the desire

to distinguish oneself and be different from others, as people are found to compete even with their intimates when it comes to performing well on dimensions that are of high personal relevance (Tesser, 1988). Indeed, the need for high self-esteem has been suggested as the ultimate human motive, because it buffers fear of death which is a consequence of humans being conscious of their mortality. Such awareness engenders abject terror that needs to be managed as it could paralyze the ability to act (Pyszczynski et al., 1997). Most recently, human needs have been explored that relate to the cognitive capabilities of a person: the need for competence and autonomy, the need of engaging in and enjoy effortful reasoning, or the need of finding firm answers to open questions (summary by Gollwitzer et al., 2011).

Other recent research on individual differences in terms of motivational orientations has replaced the need (motive) construct with constructs that describe a general goal orientation, such as personal projects, personal strivings, life tasks, or identity goals. Such personal strivings (Emmons, 1996; Gollwitzer and Kirchhof, 1998) are more limited in scope (e.g., to be a honest person) and can be characterized in terms of expectancies of success, complexity, high verses low level of abstraction, avoidance verses approach orientation, degree of conflict between each other, and integration verses fragmentation. The named parameters of personal strivings have been observed to relate to measures of psychological and physical well-being.

Action Control

Early theories portrayed the human as a machine-like reactive organism compelled to act by internal and/or external forces beyond our control (e.g., instincts, needs, drives, incentives, reinforcers, etc.). Prototypical theories are Hull's learning theory and the reformulations by his students. These theories imply that if one just pushed the right button, motivation would result. There is no room for conscious reflections and attempts toward self-regulation. Instead, motivational forces transmit their energy outside of awareness, establishing a state of balance or equilibrium (referred to as arousal reduction, self-preservation, or need satisfaction).

More modern theories have construed human beings as all-just and all-knowing final judges of their actions. For instance, expectancy-value theories (e.g., Atkinson, 1957) assumed that people choose goals in a rational way, based on the comprehensive knowledge of the probability of goal attainment and the goal's expected value. It is proposed that the subjective probability of success and the incentive value of having performed a task (i.e., pride or shame) conjointly affect task choice, both variables being influenced by the perceived difficulty of the task. Elaborations of this model by Heinz Heckhausen (1991, 1926–88) added further expectation-related concepts (e.g., situation-outcome expectancies), and differentiated various aspects of the incentive value (e.g., extrinsic side effects). Attribution theories (e.g., Weiner, 1992) even proposed that the motivational determinants of a person's behavior are causal explanations of prior action outcomes. People are seen as amateur scientists who systematically explore the causes of their behavior. The type of causes discovered are expected to affect a person's readiness to engage in these or related behaviors by influencing affect and expectations.

The motivational importance of control beliefs has also been analyzed systematically. According to Bandura's (1997) self-efficacy theory individuals may hold the firm belief that they possess the potential to execute (i.e., control) the kinds of behaviors that a given task demands. People acquire such beliefs by reflecting on their own relevant past behaviors, observing the behavior of similar others, being evaluated by significant others (e.g., teachers), and observing their own physiological reactions when challenged by a given task. High self-efficacy beliefs are observed to be associated with choosing aspiring goals, exerting strong effort to attain these goals, and persisting in the face of obstacles and hindrances.

Present theories of motivation go beyond conceptualizing humans as all-just and all-knowing. Rather, human beings are construed as flexible strategists. The focus is on the different tasks a person has to perform when transforming wishes into actions (Gollwitzer, 1990, 2012). When choosing goals, people try to live up to the ideal of being an all-knowing and all-just person by processing all the available information in an impartial manner. However, when the implementation of an already set goal is at issue, people are determined to reach the goal and thus become partial, and the desirability and feasibility of a desired outcome are now seen in the most positive light.

Currently, most research on motivation makes use of the concept of goals (Bargh et al., 2010). One line of goal research focuses on the determinants and processes of goal setting, whereas the other line targets goal implementation. With respect to goal setting, for instance, it has been discovered that people who construe their self as an ideal which they intrinsically desire to attain, set themselves promotion goals focusing on establishing and keeping positive outcomes, whereas people who construe their self as an ought which they feel compelled to reach, set themselves prevention goals focusing on avoiding and getting rid of negative outcomes (Higgins, 2006). Moreover, people can regulate the process of goal setting in a more or less productive manner, by the way in which they think about the future outcomes they want to attain. When the desired future is mentally contrasted with negative aspects of impeding reality (e.g., effectively mastering a writing project is mentally contrasted with obstacles standing in its way), relevant expectations of successfully realizing one's fantasies become activated and used. As a consequence, people form goal commitments in a rational manner (i.e., they show strong goal striving when expectations of success are high but leave the field when probabilities of success are low). When people only dream about positive future outcomes or solely ruminate about the negative reality, however, the respective goal commitments stay moderate irrespective of the level of expectations of success (Oettingen, 2000, 2012).

Regarding the determinants of successful goal implementation, how goals are framed makes an important difference. For instance, when achievement goals are framed as learning goals (i.e., goals geared at trying to learn more about how one can successfully carry out the task at hand) as compared to performance goals (i.e., goals geared at trying to

find out how capable one is), failure experiences are coped with more effectively and thus people are more likely to ultimately reach their goals (Dweck, 1999). But it also matters how goal striving is regulated by the individual. For instance, people can protect an ongoing goal pursuit from distractions by making plans on how to deal with them, or they can plan the details (when, how, and where) of the initiation and execution of the goal-directed behavior ahead of time so that any potentially disturbing self-states (e.g., being tired) can no longer interfere (Gollwitzer, 1999; Gollwitzer and Oettingen, 2011). People can also step up their efforts when hindrances are encountered and turn to substitute goals if increased effort still fails to guarantee goal attainment. Other effective action control strategies relate to the regulation of one's emotions, the perceived attractiveness of the goal, and to creating an environment that offers good opportunities for making progress toward goal attainment (Kuhl and Beckmann, 1994).

This recent revival of research on the self-regulation of goal pursuit (Mischel et al., 1996) is reminiscent not only of the mentalists' analysis of willing (James, 1890, 1842–1910) but also of German will psychology (Ach, 1935, 1871–1946; Lewin, 1926, 1890–1947) before the heydays of behaviorism. William James pointed out that any self-regulation either has to do with strengthening a weak tendency to perform a desired behavior (i.e., issues of the obstructed will) or with weakening a strong tendency to perform an unwanted behavior (i.e., issues of the explosive will). William James' analysis of willing was based on the assumption that behavior can potentially be regulated by a person's resolutions (or intentions, subjective goals), even though in certain situations and at certain times this may be difficult.

Kurt Lewin's (1926) experimental work on the willful control of behavior also offers ideas on how such control may come about, and the same is true for the research of Narziss Ach (1935). Lewin suggested that goals assign a valence to objects and events in people's social and nonsocial surroundings. In Lewin's classic example of a person who intends to mail a letter, a mail box entices the person to deposit the letter much as food entices a hungry person to eat. As needs can be satisfied by various types of behaviors which may all substitute for each other in reducing need tension (e.g., eating fruit, vegetables), many different goal-directed behaviors qualify for satisfying the quasineed associated with a set goal. Lewin's tension state metaphor thus effectively accounts for the flexibility of goal striving.

However, Ach's approach to the analysis of willing was quite different. He assumed that the linking in one's mind of an anticipated situation to a concrete intended behavior creates what he called a 'determination' which in turn automatically triggers the intended action when the specified situation is encountered. The strength of the determination was not assumed to relate to the importance of the person's intention or goal but rather to the strength of the specified situation-response link. Ach's assumption that automatic processes may contribute much to a person's goal attainment is supported by recent findings that the self-regulation strategies of both mental contrasting (Oettingen, 2012) and implementation intentions (Gollwitzer and Oettingen, 2011) unfold their beneficial effects on the basis of automatic processes, as well as the discovery that goal striving can be activated (i.e., primed) outside of a person's awareness by subliminally presenting goal-relevant cues (e.g., cues that relate to goal-relevant activities or goal attainment; Bargh, 2006).

See also: Attributional Processes: Psychological; Cognitive Psychology: History; Free Will and Action; Motivation and Actions, Psychology of; Motivation, Learning, and Instruction.

Bibliography

Ach, N., 1935. Analyse des Willens. Urban und Schwarzenberg, Berlin.

Atkinson, J.W., 1957. Motivational determinants of risk-taking behavior. Psychological Review 64, 359–372.

Bandura, A., 1997. Self-efficacy: The Exercise of Control. W.H. Freeman, New York.

Bargh, J.A., 2006. What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior. European Journal of Social Psychology 36, 147–168.

Bargh, J.A., Gollwitzer, P.M., Oettingen, G., 2010. Motivation. In: Fiske, S., Gilbert, D., Lindzey, G. (Eds.), Handbook of Social Psychology, fifth ed. Wiley, New York.

Baumeister, R.F., Leary, M.R., 1995. The need to belong: desire for interpersonal attachments as a fundamental human motivation. Psychological Bulletin 117, 497–529.

Cofer, C.N., Appley, M.H., 1964. Motivation: Theory and Research. Wiley, New York.

Dweck, C.S., 1999. Self-theories. Their Role in Motivation, Personality, and Development. Psychology Press, Philadelphia.

Emmons, R.A., 1996. Striving and feeling: personal goals and subjective well-being. In: Bargh, J., Gollwitzer, P.M. (Eds.), The Psychology of Action: Linking Motivation and Cognition to Behavior. Guilford, New York, pp. 314–337.

Gollwitzer, P.M., 1990. Action phases and mind-sets. In: Higgins, E.T., Sorrentino, E.M. (Eds.), Handbook of Motivation and Cognition: Foundations of Social Behavior, vol. 2. Guilford Press, New York.

Gollwitzer, P.M., 1999. Implementation intentions: strong effects of simple plans. American Psychologist 54, 493–503.

Gollwitzer, P.M., 2012. Mindset theory of action phases. In: Van Lange, P., Kruglanski, A.W., Higgins, E.T. (Eds.), Handbook of Theories of Social Psychology, vol. 1. Sage Publications, London.

Gollwitzer, P.M., Barry, H., Oettingen, G., 2011. Needs and incentives as sources of goals. In: Aarts, H., Elliot, A. (Eds.), Goal-Directed Behavior. Psychology Press, New York.

Gollwitzer, P.M., Kirchhof, O., 1998. The willful pursuit of identity. In: Heckhausen, J., Dweck, C.S. (Eds.), Life-Span Perspectives on Motivation and Control. Cambridge, New York.

Gollwitzer, P.M., Oettingen, G., 2011. Planning promotes goal striving. In: Vohs, K.D., Baumeister, R.F. (Eds.), Handbook of Self-regulation: Research, Theory, and Applications, second ed. The Guilford Press, New York.

Heckhausen, H., 1991. Motivation and Action. Springer-Verlag, Berlin.

Higgins, E.T., 2006. Value from hedonic experience and engagement. Psychological Review 113, 439–460.

Hull, C.L., 1943. Principles of Behavior. Appleton-Century, New York.

James, W., 1890. The Principles of Psychology. Holt, New York.

Kuhl, J., Beckmann, J., 1994. Volition and Personality: Action versus State Orientation. Hogrefe and Huber, Seattle, WA.

Lewin, K., 1926. Vorsatz, Wille und Bedürfnis. Psychologische Forschung 7, 330–385.

Maslow, A.H., 1954. Motivation and Personality. Harper & Row, New York.

McClelland, D.C. (Ed.), 1955. Studies in Motivation. Appleton-Century, New York.

McClelland, D.C., 1985. Human Motivation. Scott Foresman, Glenview, IL.

McDougall, W., 1932. The Energies of Men: A Study of the Fundamentals of Dynamic Psychology. Methuen, London.

Mischel, W., Cantor, N., Feldman, S., 1996. Principles of self-regulation: the nature of will power and self-control. In: Higgins, E.T., Kruglanski, A.W. (Eds.), Social Psychology: Handbook of Basic Principles. Guilford Press, New York.

Murray, H.A., 1938. Explorations in Personality. Oxford University Press, New York.

Oettingen, G., 2000. Expectancy effects on behavior depend on self-regulatory thought. Social Cognition 18, 101–129.
Oettingen, G., 2012. Future thought and behavior change. European Review of Social Psychology 23, 1–63.
Pyszczynski, T.A., Greenberg, J., Solomon, S., 1997. Why do we need what we need? A terror management perspective on the roots of human social motivation. Psychological Inquiry 8, 1–20.
Spence, K.W., 1956. Behavior Theory and Conditioning. Yale University Press, New Haven, CT.
Tesser, A., 1988. Toward a self-evaluation maintenance model of social behavior. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 21. Academic Press, New York.
Weiner, B., 1992. Human Motivation. Sage, Newbury Park, CA.

Motivation: Life Course and Sociological Perspectives

Steven Hitlin, University of Iowa, Iowa City, IA, USA
Matthew A Andersson, Yale University, New Haven, CT, USA
Glen H Elder, Jr., Carolina Population Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G.H. Elder Jr., volume 13, pp. 8817–8821, © 2001, Elsevier Ltd.

Abstract

Recent developments in life course research have enriched the core principle of human agency, moving past personality-based understandings toward an exploration of the ways that human reflexivity interacts with personality to motivate the bounded selection of life course pathways. We build on literature detailing the roles of self-efficacy, personal control, and optimism to suggest economical ways to conceptualize and measure the mutually intersecting influences of individual motivational dispositions and social structures on people's linked lives and life course pathways.

Life course studies involve an interdisciplinary approach to understanding the multifaceted relationships between changing social contexts and human development (Elder and Shanahan, 2006). Fundamentally, the study of the life course adopts a longitudinal perspective, focusing on the interplay between individual biological and social development within changing historical contexts and institutions. Life course theory refers to a set of orienting principles for research on the interplay of social structure and individual lives, reflecting Mills' (1959) notion of the intersection of history and biography (e.g., Elder, 1998). It is fundamentally interdisciplinary, bringing together theory and research across developmental psychology, social psychology, history, and sociology. The past half century has seen the field grow from a fledgling research enterprise into a perspective that influences research across the social sciences, including fields of health, social stratification, social psychology, and developmental science.

The Life Course and Its Paradigmatic Principles

The life course refers to sequential patterns of age-graded events, societal roles, expectations, and social structures bounded by historical time and location. Intersecting trajectories, including marriage, work, education (see Mortimer and Shanahan, 2003) form the backdrop for the contemporary life course. Trajectories involve expected sequences of roles and states (e.g., employed), with some following strongly patterned routes and others involving turning points, redirections of one's life based on circumstances, beliefs, or behaviors. Fundamental to the perspective is that such trajectories are shaped and interpreted by people and institutions within historical eras; the perspective tends toward skepticism about 'timeless' rules of human organization in favor of contextualized understanding of developmental processes and socially structured institutional and cultural influences.

Since its inception a half century ago, the study of the life course has cohered around five general principles that focus less on cohorts and more on individual choices made within historical and social contexts (see Elder et al., 2003): (1) life span development, (2) agency, (3) time and place, (4) timing, and (5) linked lives. We briefly discuss each of these in turn, with an extended discussion of agency later in the article.

The principle of life span development extends the typical focus on childhood development and socialization across the life course. Major changes in individual orientations, physical and mental health capacities, role occupancy, and geography occur throughout one's life. For example, McLeod and Kaiser (2004) demonstrate how early behavioral and emotional problems in children lessen the likelihood of receiving a high school degree and enrolling in college, thus contributing to the reproduction of socioeconomic disparities.

Individual lives are embedded within particular historical eras and locations. While this may appear to be a banal observation, it is often ignored within social science theory and research. The meaning attached to particular transitions or historical events shifts based on era or location; the founding of America means something different depending on one's ancestral history.

Experiences differentially affect individuals based on when they occur within their life course (George, 1993), or the age of an individual experiences a more global event. Elder (1974 [1999]) demonstrates how life course timing informs understanding of how different age cohorts experienced the Great Depression. Younger children experienced more family disruption and difficulty than children who were slightly older, influencing developmental and life course pathways.

Finally, the principle of linked lives explicates the notion that people's lives are lived interdependently through important pairings and social networks. While many statistical models focus exclusively on the individual, in practice, people's life choices are fundamentally shaped by social ties, relationships that can even serve as turning points away from criminal careers (e.g., Sampson and Laub, 1993).

The Emergence of Life Course Theory: An Introduction

Life course theory emerged in response to the growth of longitudinal studies and their conceptual challenges during the last half of the twentieth century (Mortimer and Shanahan, 2003). The challenges called for the replacement of child-based accounts of development with life span concepts and

the contextualization of development over time in an ever-changing society (Elder, 1998). Social roles in relation to the temporal meanings of age have provided this contextual perspective on the social life course of individuals and their development. Age locates people in history through birth year and across their lifetime through age at key life events and transitions, such as first full-time employment. Life course studies now appear with increasing frequency in the social, behavioral, and medical sciences.

These conceptual developments resulted in a life course perspective on Californians who were born at the opposite ends of the 1920s – the Oakland and Berkeley study members (1920–21, 1928–29 birth cohorts) (Elder, 1974[1999]). They experienced hard times in the Great Depression as adolescents and young children, and were followed through the World War II into the postwar era up to the 1970s. These "Children of the Great Depression" were influenced by exposure to hard times in ways that frequently shaped their lives. Conceptual themes from this study depicted key elements of people's lives beginning with a perspective on 'development and aging as lifelong processes.' Their lives were 'linked' to others from family and friends to coworkers. The 'timing of events and social roles,' whether younger or older, in the Depression and World War II, had much to do with the impact of those times on their lives. The 'historical time and place' of the study members profoundly influenced their lives, typically through the choices made and options foregone. These choices were characterized as expressions of 'human agency.'

The central themes of this early study have been reaffirmed in subsequent research as core principles of the life course paradigm. Support for the principle of life span development comes from research that shows the relevance of prenatal development for health in the later years (Kuh et al., 1997). The British national longitudinal studies of birth cohorts (Ferri et al., 2002) are documenting the relationship between the formative years of development and late-life adaptation. The principle of linked lives is expressed through the influence of significant others, whether parents or siblings, close friends or siblings. Laub and Sampson (2003) have shown that wives played an important role in terminating a husband's criminal career.

The principle of life course timing provides insight concerning life chances. Boys who are involved in delinquent activity at an early age tend to persist in this behavior, well into the adult years. Biological dispositions may play a role in this pattern as well as an entrenched life of crime (Moffitt et al., 2002). The principle on historical time and place indicates access to types of life opportunity and risks of disadvantage. Men in the 1920–21 birth cohort of Oakland Study males were exposed to the risks of both Depression hard times and combat in World War II. Very different opportunities and risks were relevant to the male birth cohort of 1940–41.

No paradigmatic principle is more fundamental to the life course and motivation than that of human agency (Haidt and Rodin, 1999). The transition from one activity to another across the life course involves individual initiatives, as well as situational constraints and opportunities. People make choices in life that enable them to exert a degree of control over their life course. In economically deprived families during the Great Depression, young people were confronted by pressing household needs. Some assumed more responsibility over the care of younger brothers and sisters while mother held down a part-time work, while others performed jobs in the community. The employment of mothers illustrates the process by which choices result in 'building a new life course.' No life course is totally shaped by social constraints.

The Agent in Life Course Theory

The problem of understanding individual volition is peculiar to disciplines that typically focus on structural and cultural constraints that enable human action; the issue of 'agency' is only a problematic for scholars that exclusively emphasizes the power of macrosocial forces in shaping observed regularities in human conduct. In the sociological literature, the discussion is often pitched at a level of abstraction that is difficult to translate into empirical research (Hitlin and Elder, 2007). However, notable treatments incorporate the ongoing nature of time into definitions of human agency (e.g., Emirbayer and Mische, 1998; Mische, 2009). For the purposes of this discussion, we bracket those scholars (Loyal and Barnes, 2001; Meyer and Jepperson, 2000) who suggest that human agency is epiphenomenal or sociologically trivial.

In general, the principle of agency suggests the importance of personal factors and choices in the selection and maintenance of various life trajectories. People construct their own courses within social constraints, what Shanahan and Hood (1998) refer to as 'bounded agency.' People navigate their lives as they encounter obstacles and opportunities. Some scholars (e.g., Heinz, 2003; Honneth, 2004) suggest that the modern life course offers greater structural potential for individuals to shape their own life course trajectories.

That individuals make choices is largely unproblematic as social analysis; the devil is in the details. Empirical treatments of agency, less commonly merged with theoretic treatments of the concept, diverge in their focus but can generally be seen to cohere around a few core ideas of individual capacities dealing with potential control over one's life course. Such capacities to exert agency can be subdivided into two broad approaches: (1) a focus on personal skills/attributes that allows people to exert control over their lives and (2) self-beliefs about the possession of those skills (Hitlin and Long, 2009). Prominent life course treatments have focused on both aspects. Shanahan and Hood (1998) hold that a fully specified model of agency requires measurement across multiple dimensions, relationships, and proximal settings that facilitate and mold goal pursuit and macrostructural contexts.

Most notably, Clausen's (1991, 1993) work demonstrates the importance of particular constellations of skills and orientations early in life for understanding later life course outcomes (see also Shanahan et al., 1997). This notion of 'planful competence' draws on psychological Q-sort techniques of members of previously mentioned Berkeley and Oakland studies, rating children on six dimensions combined to assess their skills of self-reflexivity, dependability, and self-confidence. This construct measures how well individuals can make (and adhere to) long-term plans (Shanahan et al., 2000). As Clausen and others demonstrate, young people with more

of these skills evidence more stable work and marital careers and are more capable in dealing with personal setbacks.

A number of related social psychological constructs cohere around the second notion of agency, the motivated significance of self-beliefs in understanding individual pathways across the life course (see Gecas, 2003). From the sociological side, commonly utilized measures include the notion of 'personal control' (e.g., Mirowsky and Ross, 2003, 2007) and 'mastery' (Pearlin, 1983). Psychological traditions offer notions of 'locus of control' (Rotter, 1966) and 'self-efficacy' (Bandura, 1982, 2008; see Gecas, 1989, for sociological appropriations of the construct). Differences exist between these constructs, primarily at the level of scale construction (for an extended discussion of the empirical measures we discuss, see Hitlin and Long, 2009; also Haidt and Rodin, 1999), but they largely agree that individuals carry personal assessments about their ability to persevere in the face of obstacles and influence the events of their lives. A range of studies demonstrate that the possession of higher levels of these beliefs is advantageous for life course outcomes (see Gecas, 2003; Sharkey, 2006; Wheaton, 1980; Christie-Mizell and Erickson, 2007; Frazier et al., 2011; Reynolds et al., 2007), including serving as a buffer between negative life events and individual functioning (e.g., Pudrovska et al., 2005). The ability to persevere in the face of obstacles is itself a social force leading toward advantageous outcomes (e.g., Duckworth et al., 2007). Aging can serve to reduce the sense of agency that people experience (e.g., Pearlin and Schooler, 1978; Schieman and Turner, 1998).

Life Course Motivation

Individuals have a variety of motivations as they navigate the culturally defined sequence of age-graded roles that defines the life course. Chiefly, they are motivated to experience belongingness or relatedness, self-esteem, self-efficacy or competence, autonomy, and a consistent sense of self (e.g., Bandura, 2008; Deci and Ryan, 2000; Leary, 2007).

A consistent sense of self is provided in part through the routine performance of identities associated with social roles and statuses (Owens et al., 2010). Recent advances in life course research have drawn attention to consistency among multiple identities as an antecedent of life course outcomes and as a motivator of life course trajectories. Benson and Elder (2011) performed a cluster analysis of subjective age and psychosocial maturity using nationally representative data, finding support for a pseudo-adult identity type, which is marked by a relatively high subjective age coupled with lower levels of psychosocial maturity; in contrast, an anticipatory adult identity type is marked by levels of maturity that notably exceed subjective age. These inconsistent identity types may well configure role and mental health trajectories in the middle years as adulthood proceeds (e.g., Foster et al., 2008). However, this remains an open area for research. Wang et al. (2012) similarly characterized status inconsistencies among young adults, focusing on mismatches among socioeconomic status, cognitive ability, and high school grade point average in high school. Certain types of mismatches predicted enrollment in the military; military enrollment, in turn, enhanced the attainment of some level of higher education relative to nonveterans with the same status configuration.

Self-consistency is but one motive that shapes agency across the life course. In making sense of the array of motivated dynamics, it is useful to appeal to three overarching temperamental systems that energize motivated action: approach, avoidance, and effortful control (Carver and Connor-Smith, 2010; Caspi, 2000; Caspi et al., 1995). Studies of nonhuman animals, including octopi, goats, cats, marmots, and rhesus monkeys, support a two-factor conception of motivation that comprises exploratory (i.e., approach-oriented) behaviors on the one hand and protective (i.e., avoidance-oriented) behaviors on the other (Elliot and Covington, 2001; Mehta and Gosling, 2006). Procuring valuable resources and avoiding hazardous or risky situations are both vital survival tasks and hence are governed by two independent motivational systems. These motivational systems are orthogonal, which means that activity in one is generally unhelpful in predicting activity in the other and that outputs from both systems combine to produce distinctive patterns of action.

Applied to the life course, the approach system motivates a broadening of one's resources, seeking social support and moving toward positive, rewarding situations. It is indicated in adulthood by personality features such as optimism, extraversion and trait positive emotion (Andersson, 2012a; Hitlin and Elder, 2007). In contrast, the avoidance system motivates self-protection, social withdrawal and staying away from negative, undesired situations (Elliot, 2006; Elliot and Thrash, 2002). Personality indicators of this system include shyness, anxiety, neuroticism, and trait negative emotion (Carver and Connor-Smith, 2010; Caspi, 2000). Individuals vary meaningfully in their typical levels of approach and avoidance motivation; these fundamental temperamental variations are elaborated into personality structures across the life course and predict the successful navigation of key role transitions involving work, family and marriage (Caspi, 2000, 2002; Caspi et al., 1989), the weathering of relationship and work crises (Andersson, 2012b; Carver et al., 2010), and the acquisition of personal resources such as self-esteem and social network diversity (Andersson, 2012a; Caspi, 2000; Hitlin and Elder, 2007). Indeed, approach-oriented individuals seem more likely to accumulate multiple identities that then can serve as sources of resilience during difficult times (Thoits, 2003). Moreover, they should be more adept at handling conflict among role identities.

Meanwhile, effortful control is a superordinate temperament that involves the reflexive management and redirection of outputs from approach and avoidance motivation systems through top-down cognitive functioning (Carver and Connor-Smith, 2010). It is conceptually akin to executive self-control (Baumeister et al., 2007; Mischel et al., 1988). Deliberate, controlled action is important to the stability of occupational trajectories (Clausen, 1991) and to desistance from aggressive and violent behavioral tendencies (Sampson and Laub, 1993). Clausen's (1991) key disposition of planful competence includes elements of approach and effortful control; it is embodied by consistency, self-confidence, and intellectual vigor. Clausen describes the opposite of planful

competence as fearfulness, distrust, and a self-defeating outlook (1991: 836). In this way, it is similar to the inhibited temperament researched by Caspi and colleagues (Caspi, 2000; Caspi et al., 1989).

These three temperamental systems are biologically and genetically shaped. They are expressed from infancy onward, and are elaborated into complex personality systems across the life course through structural circumstances (McLeod and Lively, 2003) and cumulative and interactional continuity mechanisms (Caspi, 2000; Caspi et al., 1989). They exert effects on individual outcomes and life chances indirectly through personality formation and social selection processes and directly through deep emotional responses to particular life situations. Caspi (2000) describes continuity processes linking early temperament to adult personality structure. Adult personality factors of constraint, positive emotion, and negative emotion map onto temperamental variation in control, approach, and avoidance. In sum, Caspi conceptualizes adult personality as a 'social and cognitive elaboration' of temperamental styles (2000: 161).

As suggested by Caspi, the influence of agentic motivations on life course outcomes may reach its apex during periods of role transition. During transitory periods, new processes of role negotiation and learning come into play. Those who show a predominance of avoidance motivation often miss out on developmental opportunities by focusing too rigidly on potential threats and obstacles and reverting to ingrained habits that are unproductive in new role environments. At the same time, some level of avoidant functioning is invaluable to social adjustment, as it predicts refraining from certain delinquent behaviors and poor health behaviors (Hirsh et al., 2009). Indeed, some scholars have suggested that avoidant temperament, while counterproductive in excess, provides the very basis for 'socializability,' whereas approach temperament constitutes a key basis of motivated self-actualization and personal growth (DeYoung, 2006). Conceptual parallels can be found in Piaget's (1970) identification of key developmental tasks including assimilation into the social order and accommodation of shifts in knowledge, what Peterson (1999) interprets as core human tasks of both maintaining social order and adapting to change.

The diverse and wide-ranging impact of approach and avoidance motivation on life outcomes is showcased by its correlates in autobiographic life narratives. In a recent study, Wilt et al. (2011) examined relationships between personality and themes in life stories. They found that avoidant tendencies were linked to construals of threat whereas approach tendencies predicted themes of exploration. In previous research, McAdams (2006) had previously highlighted the importance of personality structure to understanding how individuals recruit and elaborate on basic American themes of contamination and redemption in developing life narratives.

In sum, the approach–avoidance theory of motivation is supported by investigations of animal and human behavior and lends uncanny purchase in understanding the structure and outcomes of motivation across the life course. Approach–avoidance temperament patterns, regulated by effortful control, evolve into complex personality systems and predict the success of key role transitions, engagement in adaptive and delinquent forms of behavior, and the development of various personal resources. Meanwhile, these basic temperaments mold the application of cultural knowledge to the interpretation of singular life events and to the construction of overarching life narratives.

New Directions

The move toward interdisciplinary research across the social sciences in the past decade serves to both validate life course studies' long-held core perspectives and suggests an exciting future for the study of human development in sociohistorical context. As suggested above, there is room for greater understanding of the nature of human motivation and choice as it intersects with structurally bounded changes, especially considering global economic shifts, the results of which are far from known. We suggest a few arenas of profitable life course research.

While there are more studies and data sets available that follow individuals throughout their lives, there is room for more linking of early life temperament and personality with later life motivations and choices. To what extent do early approach–avoidance tendencies exacerbate or mitigate social disadvantage? What factors lead to the development of perseverance in the face of obstacles, and do those forces operate differently across racial, ethnic, gender, and socioeconomic groups? Following up on Elder (2001) in a previous edition of the Encyclopedia, how can we best assess relative influences of individual and structural factors on agentic capacities as well as the employment of those capacities?

The principle of linked lives has rich possibilities for engagement with issues of motivation and agency. What sorts of significant relationships serve to 'push' individuals, or open up structurally blocked pathways? A good deal of research explores the role of marriage in terms of status attainment processes, less work deals with relationships such as mentoring (e.g., McDonald et al., 2007). Additionally, the typically domain-specific nature of much research has, to this point, precluded fruitful work that focuses on overlapping individual trajectories such as the work–family balance (Blair-Loy, 2003).

Many psychological models of motivation suggest a biological component in demonstrated capacities of perseverance as well as personality dispositions that influence situated interaction and life course trajectories. Future work should more fully link neurological, psychological, and sociological levels of analysis for understanding patterned variation and reciprocal influences across these levels of analysis.

Lifelong understandings are becoming more axiomatic in the study of human development, with developmental psychology, social psychology, and mainstream sociology journals regularly publishing longitudinal studies that focus on both individual and structural factors. Similarly, research-linking individual and social structural factors are becoming mainstream across the social sciences (McLeod and Lively, 2003), with advances in statistical modeling techniques enabling research to disentangle individual level processes from different levels of structural influence.

See also: Cultural Views of Life Phases; Self and Identity Development During Adolescence across Cultures; Self-Regulation During Adolescence: Variations Associated with Individual–Context Relations; Work Motivation.

Bibliography

Andersson, Matthew A., 2012a. Dispositional optimism and the emergence of social network diversity. Sociological Quarterly 53, 92–115.

Andersson, Matthew A., 2012b. Identity crises in love and at work: dispositional optimism as a durable personal resource. Social Psychology Quarterly 75 (4), 290–309.

Bandura, Albert, 1982. The self and mechanisms of agency. In: Suls, J. (Ed.), Psychological Perspectives on the Self. Lawrence Erlbaum, Hillsdale, NJ, pp. 3–40.

Bandura, Albert, 2008. Toward an agentic theory of self. In: Marsh, Herbert W., Craven, Rhonda G., McInerney, Dennis M. (Eds.), Self-processes, Learning, and Enabling Human Potential. Information Age Publishing, United States, pp. 15–49.

Baumeister, Roy F., Vohs, Kathleen D., Tice, Dianne M., 2007. The strength model of self-control. Current Directions in Psychological Science 16, 396–403.

Benson, Janel E., Elder Jr., Glen H., 2011. Young adult identities and their pathways: a developmental and life course model. Developmental Psychology 47, 1646–1657.

Blair-Loy, Mary, 2003. Competing Devotions: Career and Family among Women Executives. Harvard University Press, Cambridge, MA.

Carver, Charles S., Connor-Smith, Jennifer, 2010. Personality and coping. Annual Review of Psychology 61, 679–704.

Carver, Charles S., Scheier, Michael F., Segerstrom, Suzanne C., 2010. Optimism. Clinical Psychology Review 30, 879–889.

Caspi, Avshalom, 2000. The child is father of the man: personality continuities from childhood to adulthood. Journal of Personality and Social Psychology 78, 158–172.

Caspi, Avshalom, 2002. Social selection, social causation, and developmental pathways: empirical strategies for better understanding how individuals and environments are linked across the life course. In: Pulkinen, Lea, Caspi, Avshalom (Eds.), Paths to Successful Development: Personality in the Life Course. Cambridge University Press, New York, pp. 281–302.

Caspi, Avshalom, Bem, Daryl J., Elder Jr., Glen H., 1989. Continuities and consequences of interactional styles across the life course. Journal of Personality 57, 375–406.

Caspi, Avshalom, Henry, Bill, McGee, Rob O., Moffitt, Terrie E., Silva, Phil A., 1995. Temperamental origins of child and adolescent behavior problems: from age three to age fifteen. Child Development 66, 55–68.

Christie-Mizell, C.A., Erickson, Rebecca J., 2007. Mothers and mastery: the consequences of perceived neighborhood disorder. Social Psychology Quarterly 70 (4), 340–365.

Clausen, John A., 1991. Adolescent competence and the shaping of the life course. American Journal of Sociology 96, 805–842.

Clausen, John A., 1993. American Lives: Looking Back at Children of the Great Depression. Free Press, New York.

Deci, Edward L., Ryan, Richard M., 2000. The 'what' and 'why' of goal pursuits: human needs and the self-determination of behavior. Psychological Inquiry 11, 227–268.

DeYoung, Colin G., 2006. Higher-order factors of the big five in a multi-informant sample. Journal of Personality and Social Psychology 91, 1138–1151.

Duckworth, Angela L., Peterson, Christopher, Matthews, Michael D., Kelly, Dennis R., 2007. Grit: perseverance and passion for long-term goals. Journal of Personality and Social Psychology 92 (6), 1087–1101.

Elder Jr., Glen H., 1974[1999]. Children of the Great Depression: Social Change in Life Experience, twenty-fifth anniversary ed. Westwood Press, Boulder, CO.

Elder Jr., Glen H., 1998. The life course and human development. In: Lerner, Richard E. (Ed.), Theories of Human Development: Contemporary Perspectives in William Damon (Editor-in-Chief), The Handbook of Child Psychology, fifth ed., vol. 1. Wiley, New York, pp. 939–991.

Elder Jr., Glen H., 2001. Life course. In: Maddox, George E. (Ed.), Encyclopedia of Aging, third ed. Springer, New York, pp. 593–596.

Elder Jr., Glen H., Johnson, Monica K., Crosnoe, Robert, 2003. The emergence and development of life course theory. In: Mortimer, Jeylan T., Shanahan, Michael J. (Eds.), Handbook of the Life Course. Kluwer, New York, pp. 3–19.

Elder Jr., Glen H., Shanahan, Michael J., 2006. The life course and human development. In: Damon, William, Lerner, Richard M. (Eds.), Handbook of Child Psychology, Theoretical Models of Human Development, sixth ed., vol. 1. Wiley and Stone, New York, pp. 665–715.

Elliot, Andrew J., 2006. The hierarchical model of approach-avoidance motivation. Motivation and Emotion 30, 111–116.

Elliot, Andrew J., Covington, M.V., 2001. Approach and avoidance motivation. Educational Psychology Review 13, 73–92.

Elliot, Andrew J., Thrash, Todd M., 2002. Approach-avoidance motivation in personality: approach and avoidance temperaments and goals. Journal of Personality and Social Psychology 82, 804–818.

Emirbayer, Mustafa, Mische, Ann, 1998. What is agency? American Journal of Sociology 103 (4), 962–1023.

Ferri, E., Bynner, J., Wadsworth, M., 2002. Changing Britain, Changing Lives: Three Generations at the Turn of the Century (Bedford Way Papers Series, No. 18). Institute of Education, London.

Foster, Holly, Hagan, John, Brooks-Gunn, Jeanne, 2008. Growing up fast: stress exposure and subjective 'weathering' in emerging adulthood. Journal of Health and Social Behavior 49, 162–177.

Frazier, Patricia, Keenan, Nora, Anders, Samantha P.S., Shallcross, Sandra, Hintz, Samuel, 2011. Perceived past, present, and future control and adjustment to stressful life events. Journal of Personality and Social Psychology 100 (4), 749–765.

Gecas, Viktor, 2003. Self-agency and the life course. In: Mortimer, Jeylan T., Shanahan, Michael (Eds.), Handbook of the Life Course. Klewer, New York, pp. 369–388.

Gecas, Viktor, 1989. The social psychology of self-efficacy. Annual Review of Sociology 15, 291–316.

George, Linda K., 1993. Sociological perspectives on life transitions. Annual Review of Sociology 19, 353–373.

Haidt, Jonathan, Rodin, Judith, 1999. Control and efficacy as interdisciplinary bridges. Review of General Psychology 3 (4), 317–337.

Heinz, Walter R., 2003. From work trajectories to negotiated careers: the contingent work life course. In: Mortimer, Jeylan T., Shanahan, Michael J. (Eds.), Handbook of the Life Course. Kluwer, New York, pp. 185–204.

Hirsh, Jacob B., DeYoung, Colin G., Peterson, Jordan B., 2009. Metatraits of big five differentially predict engagement and restraint of behavior. Journal of Personality 77, 1–17.

Hitlin, Steven, Elder Jr., Glen H., 2007. Understanding 'agency': clarifying a curiously abstract concept. Sociological Theory 25 (2), 170–191.

Hitlin, Steven, Long, Charisse, 2009. Agency as a sociological variable: a preliminary model of individuals, situations, and the life course. Sociology Compass 3 (1), 137–160.

Honneth, Axel, 2004. Organized self-realization: some paradoxes of individualization. European Journal of Social Theory 7 (4), 463–478.

Kuh, D., Power, C., Blane, D., Bartley, M., 1997. Social pathways between childhood and adult health. In: Kuh, D., Ben-Shlomo, Y. (Eds.), A Life Course Approach to Chronic Disease Epidemiology. Oxford University Press, New York, pp. 169–198.

Laub, John H., Sampson, Robert J., 2003. Shared Beginnings, Divergent Lives: Delinquent Boys to Age 70. Harvard University Press, Cambridge, MA.

Leary, Mark R., 2007. Motivational and emotional aspects of self. Annual Review of Psychology 58, 317–344.

Loyal, Steven, Barnes, Barry, 2001. "Agency" as a red herring. Philosophy of the Social Sciences 31 (4), 507–524.

McAdams, Dan P., 2006. The Redemptive Self: Stories Americans Live by. Oxford University Press, New York.

McDonald, Steve, Erickson, Lance D., Johnson, Monica K., Elder Jr., Glen H., 2007. Informal mentoring and young adult employment. Social Science Research 36 (4), 1328–1347.

McLeod, Jane D., Lively, Kathryn J., 2003. Social structure and personality. In: DeLamater, John (Ed.), Handbook of Social Psychology. Kluwer, New York, pp. 77–102.

McLeod, Jane D., Kaiser, Karen, 2004. Childhood emotional and behavioral problems and educational attainment. American Sociological Review 69 (5), 636–658.

Mehta, P.H., Gosling, S.D., 2006. How can animal studies contribute to research on the biological bases of personality. In: Canli, T. (Ed.), Biology of Personality and Individual Differences. Guilford Press, New York, pp. 427–448.

Meyer, John W., Jepperson, Ronald L., 2000. The "actors" of modern society: the cultural construction of social Agency. Sociological Theory 18 (1), 100–120.

Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, New York.

Mirowsky, John, Ross, Catherine E., 2003. Education, Social Status, and Health. Walter de Gruyter, New York.

Mirowsky, John, Ross, Catherine E., 2007. Life course trajectories of perceived control and their relationship to education. American Journal of Sociology 112 (5), 1339–1382.

Mische, Ann, 2009. Projects and possibilities: researching futures in action. Sociological Forum 24 (3), 694–704.

Mischel, Walter, Shoda, Yuichi, Peake, Philip K., 1988. The nature of adolescent competencies predicted by preschool delay of gratification. Journal of Personality and Social Psychology 54, 687–696.

Moffitt, T.E., Caspi, A., Harrington, H., Milne, B.J., 2002. Males on the life-course-persistent and adolescence-limited antisocial pathways: follow-up at age 26 years. Development and Psychopathology 14 (1), 179–207.

Mortimer, Jeylan T., Shanahan, Michael E., 2003. Handbook of the Life Course. Kluwer, New York.

Owens, Timothy, Robinson, Dawn, Smith-Lovin, Lynn, 2010. Three faces of identity. Annual Review of Sociology 36, 477–499.

Pearlin, Leonard I., 1983. Role strains and personal stress. In: Kaplan, H.B. (Ed.), Psychosocial Stress: Trends in Theory and Research. Academic Press, New York, pp. 3–32.

Pearlin, Leonard I., Kohn, Melvin L., 1966. Social class, occupation, and parental values: a cross-national study. American Sociological Review 31 (4), 466–479.

Pearlin, Leonard I., Schooler, Carmi, 1978. The structure of coping. Journal of Health and Social Behavior 19 (1), 2–21.

Peterson, J.B., 1999. Maps of Meaning: The Architecture of Belief. Taylor and Francis, Florence, KY.

Piaget, J., 1970. Piaget's theory. In: Mussen, P. (Ed.), Carmichael's Manual of Child Psychology, third ed. Wiley, New York, pp. 703–732.

Pudrovska, Tetyana, Schieman, Scott, Pearlin, Leonard I., Nguyen, Kim, 2005. The sense of mastery as a mediator and moderator in the association between economic hardship and health in late life. Journal of Aging and Health 17, 634–660.

Reynolds, John R., Burge, Stephanie W., Robbins, Cheryl L., Boyd, Emily M., Harris, Brandy, 2007. Mastery and the fulfillment of occupational expectations by midlife. Social Psychology Quarterly 70 (4), 366–383.

Rotter, J.B., 1966. Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs 80 (1), 1–28.

Sampson, Robert J., Laub, John H., 1993. Crime in the Making: Pathways and Turning Points through Life. Harvard University Press, Cambridge, MA.

Schieman, Scott, Turner, Heather A., 1998. Age, disability, and the sense of mastery. Journal of Health and Social Behavior 39, 169–186.

Shanahan, Michael J., Elder Jr., Glen H., Miech, Richard A., 1997. History and agency in men's lives: pathways to achievement in cohort perspective. Sociology of Education 70 (1), 54–67.

Shanahan, Michael J., Hood, Kathryn E., 1998. Adolescents in changing social structures: bounded agency in life course perspective. In: Silbereisen, Rainer, Crockett, Elizabeth (Eds.), Negotiating Adolescence in Times of Social Change: Cross-national Perspectives on Developmental Processes and Social Intervention. Cambridge University Press, Cambridge, pp. 123–134.

Shanahan, Michael J., Sulloway, Frank J., Hofer, Scott M., 2000. Change and constancy in developmental contexts. International Journal of Behavioral Development 24 (4), 421–427.

Sharkey, Patrick T., 2006. Navigating dangerous streets: the sources and consequences of street efficacy. American Sociological Review 71, 826–846.

Thoits, Peggy A., 2003. Personal agency in the accumulation of multiple role identities. In: Burke, Peter J. (Ed.), Advances in Identity Theory and Research. Springer, New York, pp. 179–194.

Wang, Viktor, Naomi, Spence, Elder Jr., Glen H., December 2012. Status configurations, military service and higher education. Social Forces 91 (2), 397–421.

Wheaton, Blair, 1980. The sociogenesis of psychological disorder: an attributional theory. Journal of Health and Social Behavior 21, 100–124.

Wilt, Joshua, Olsen, Bradley D., McAdams, Dan P., 2011. Higher-order factors of the big five predict exploration and threat in life stories. Journal of Research in Personality 45, 613–621.

Motivational Development, Systems Theory of

Marianne E Schneider, University of Zurich, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 15, pp. 10120–10125, © 2001 Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Systems theory provides a powerful method for the description of homeostatic systems, that is, systems in which feedback-controlled regulation processes occur. Since human goal-directed behavior is regulated by such processes, systems theory is also very useful for psychological research. One of the most elaborated psychological models based on systems theory is the Zurich Model of Social Motivation by Bischof. This model postulates the existence of three basic motives or needs: the needs for security, arousal, and autonomy. Each of these is treated in a specific homeostatic subsystem, which is represented as a negative feedback loop. The model assumes that the attempt to maintain the appropriate amount of security, arousal, and autonomy for the respective need results in specific regulations of the distance to social partners. Furthermore, it is postulated that those three needs, and thus the way of distance regulation, take a specific developmental course. These assumptions are supported by several studies in which the distance regulation behavior of the participants was observed and simulated by means of a parameter estimation.

Systems theory provides a powerful method for the description of systems in which feedback controlled regulation processes occur. Since human goal-directed behavior is regulated by such processes, systems theory is also very useful for psychological research. One of the most elaborated psychological models based on systems theory, the Zurich Model of Social Motivation by Bischof (1985, 1993), will be presented below, together with some empirical evidence for its validity.

Systems Theory

Systems or control theory originated from practical, engineering-related questions concerning regulation processes. Its aim is to analyze the structure of systems and to provide adequate, preferably quantitative, models for them. The basic unit of systems control is a negative feedback loop which consists of the following elements: a detector which provides the system with information about the environment (input), a reference value which indicates what the input ideally should be, and an output function that causes the system in case of a discrepancy between input and reference value to 'behave' in a way that affects the environment in order to reduce the discrepancy (for details see Bischof, 1985, 1995).

The notion that processes such as these not only occur in machines, but also can be found in living systems such as animals or human beings was first introduced into the life sciences by Wiener (1948). Almost at the same time, it was also formulated in German by von Holst and Mittelstaedt (1971). Since then, this idea has appeared in different theories. The models based on systems theory, however, vary considerably in their degree of quantification, their complexity, and their generalizability.

One simple, general, and well-known example of a negative feedback loop used to characterize human behavior is the TOTE unit described by Miller et al. (1960). A more complex attempt to explain the general principles of goal directed behavior in terms of systems theory stems from Vohs and Baumeister (2011).

Since systems theory not only is especially apt for the description of purposive behavior, but also allows for the description of short-term as well as long-term regulation processes, models or theories dealing with more specific topics can be found in the fields of motivational as well as developmental psychology. Examples in the area of motivational psychology are the control theory model of work motivation by Klein (1989) and the control system model of organizational motivation by Lord and Hanges (1987). The former describes individual work motivation as a function of feedback and several cognitive processes. The latter also deals with task performance, but can be applied to individual motivation as well as interpersonal activities in organizations. A more comprehensive model, which is at least partly based on a systemic approach, is presented by Dörner (1994 1998), who intends nothing less than to discover the 'construction plan' of the human soul. In contrast to these authors, Hyland (1988) uses his motivational control theory as a meta-theoretical framework to integrate different, already existing, motivational theories which – according to him – reflect different aspects of a single underlying process which can again be described with a negative feedback loop.

In developmental psychology, a dynamic systems approach is used, for example, by Fogel and Thelen (1987) in their theory of motor development, even if they do not put it into terms of systems theory. The models of Burke (1991) and Bell (1971) are more explicitly based on systems theory. The former describes the process of identity formation as a negative feedback loop. The latter develops a control system to explain reciprocal socialization effects between children and their caregivers. A much broader model which is concerned with the general principles of human development is the developmental systems theory by Ford and Lerner (1992). The best known example for a systems approach in developmental psychology, however, is probably Bowlby's (1969) theory of infant attachment, a phenomenon which he viewed in feedback terms, although he never ventured to formulate it in a mathematical model. This was eventually performed by Bischof (1975,

1985, 1993), who provided a completely quantified model of social motivation. This is referred to as the Zurich Model of Social Motivation (or Zurich Model) and will be described in the following section.

The Zurich Model of Social Motivation

The Zurich Model of Social Motivation by Bischof (1985, 1993, Figure 1) arose from the want of a quantitative model able to explain and predict processes of social distance regulations. Such regulations have been described by attachment theorists such as Bowlby (1969) and also occur during individuation processes in puberty and adolescence. In short, the model postulates three basic motivational systems: the security, the arousal, and the autonomy system. All of these are represented as negative feedback loops and are considered homeostatic. In addition, there is a fourth system, the coping system, which deals with cases in which one of the other systems gets blocked and cannot operate in its usual manner. These four systems will

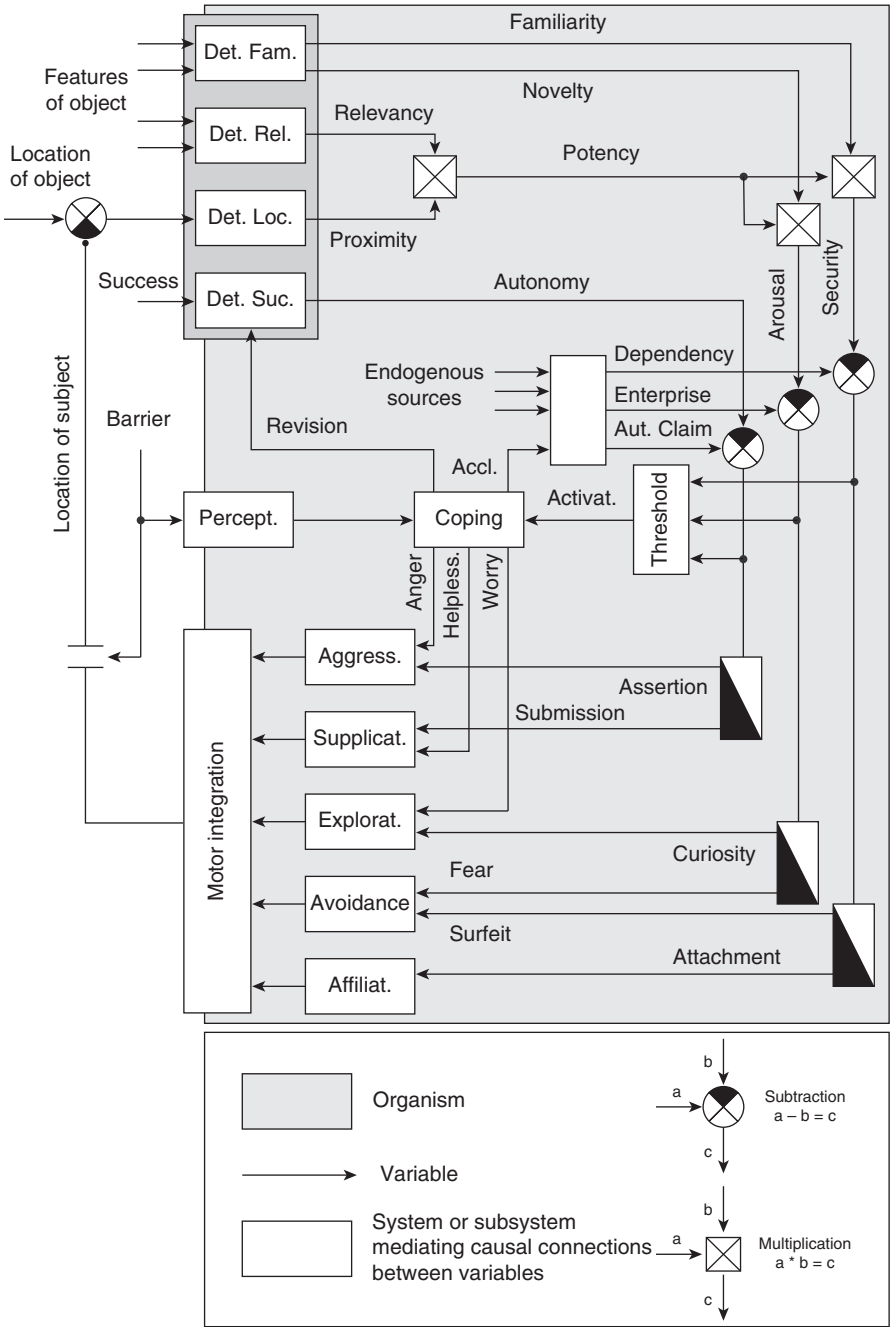

Figure 1 The Zurich Model of Social Motivation (after Bischof, 1993, p. 13).

be described qualitatively, although there also exists a mathematical description (Bischof, 1993; Gubler and Bischof, 1991; Gubler et al., 1994).

The Security System

Security is defined as the feeling of intimacy, warmth, protection, and so forth that one obtains when one is close to a familiar person, typically one's primary caregiver. This feeling has its source in three input variables, all of which are sensed by a specific detector. The detector for familiarity provides the information as to whether a certain object is familiar, in contrast to being new, unknown, or discrepant from one's expectancies. The detector for relevancy assesses the rank and the 'conspecificity' of the object in question. The detector's output is highest when the perceived object is an adult, high-ranking conspecific, and can be reduced, for example, by submissive behavior of the object or by substitution of a human adult by another object such as a child or an animal. The third detector gives information about the distance between oneself and the perceived object. The closer one is positioned to a familiar object the more security one obtains. Hence, the construct of security refers to a hypothetical emotional variable monotonically increasing with familiarity, relevancy, and proximity. Since proximity is the only one among those three variables which the subject can control, the obvious way to provide a homeostasis of security results in distance-regulating behavior.

The reference variable of the security system is the degree of dependency the subject feels. As long as dependency exceeds security, an appetence for security is maintained. This induces the subject to show attachment behavior, that is, to reduce the distance to a person who is able to provide security. The opposite situation, frequently encountered in puberty, results in an aversion against security and consequently in an avoidance of familiar persons (surfeit behavior).

The Arousal System

Arousal is defined as feelings such as interest, fascination, curiosity, as well as feelings of alarm or fear which can occur when one encounters a nonfamiliar person. Basically, the arousal system can be thought of as analogous to the security system, and moreover it can utilize the same three detectors. Like security, arousal is hypothesized to increase with the object's proximity and relevancy. Since the familiarity detector also provides information about how novel an object is, arousal can be connected to the output of this detector as well. Thus, the paradigmatic case for the arousal system is the encounter with a high-ranking, adult human stranger.

The reference value of the arousal system is labeled enterprise. Again, a distinction between an aversive and an appetent behavioral response can be made. Aversion against arousal and subsequent fear behavior result from arousal exceeding enterprise. If arousal falls short of enterprise, an appetence for arousal develops which leads to exploratory behavior.

The security and the arousal systems do not work independently of each other. Both systems respond to the same stimulus situations. Thus, the quantitative processing of the inputs involved leads automatically to effects such as fear-evoking stimuli intensifying attachment behavior towards an available caregiver, or the caregiver serving as a security base for exploratory ventures. Furthermore, the two systems do not develop independently during ontogeny. The reference variables – dependency and enterprise – are assumed to be negatively correlated, and to covary with the individual's age and maturity. Young infants are strongly dependent and only modestly enterprising. They need the presence of a security-providing caregiver in order to start exploring the environment and to acquire competence. Later in development, usually culminating around early adolescence, this pattern changes: growing independence and enterprise now demand and allow for the separation from the familiar partners and the establishment of new relationships. Finally, after adolescence, the two reference values normally level out.

The Autonomy System

The autonomy system can be regarded as analogous to the security and the arousal systems. The input here, however, stems from a different source. The detector which provides the amount of felt autonomy senses experiences of success such as obtaining obedience, acknowledgement, or admiration from other people, solving difficult problems, or being competent in certain subjects. Therefore, autonomy can be described as feeling competent, strong, and powerful, and having influence, prestige, and authority.

The reference variable of the autonomy system is called 'autonomy claim' and is closely related to the power and the achievement motives, as well as the need for prestige and an aspiration for self-confidence. In the case of a discrepancy between the feeling of autonomy and the autonomy claim, however, the model postulates the occurence of behaviors stemming clearly from the domain of power and dominance. If individuals feel an appetence for autonomy, they are assumed to behave assertively by becoming threatening, demanding, or even aggressive. In the opposite case, if individuals have too much autonomy and feel aversive, they will show submissive behaviors such as behaving humbly or servilely.

The autonomy claim is not independent of the other two reference variables, dependency and enterprise. On the contrary, the autonomy claim is postulated to underlie the connection of the security and the arousal system by affecting dependency negatively and enterprise positively.

The Coping System

The coping system operates in cases where one of the other systems gets blocked by an (internal or external) obstacle or barrier. The coping system provides three external and two internal coping strategies, external strategies being active attempts to change the situation and internal ones being efforts to adapt oneself to the situation. The three external coping strategies are labeled invention (searching for a detour), aggression (attempting to destroy the barrier), and supplication (begging someone else to remove the barrier). The two internal strategies are referred to as acclimatization (adaptation of the reference value to the input) and revision (adaptation of the perception of the situation).

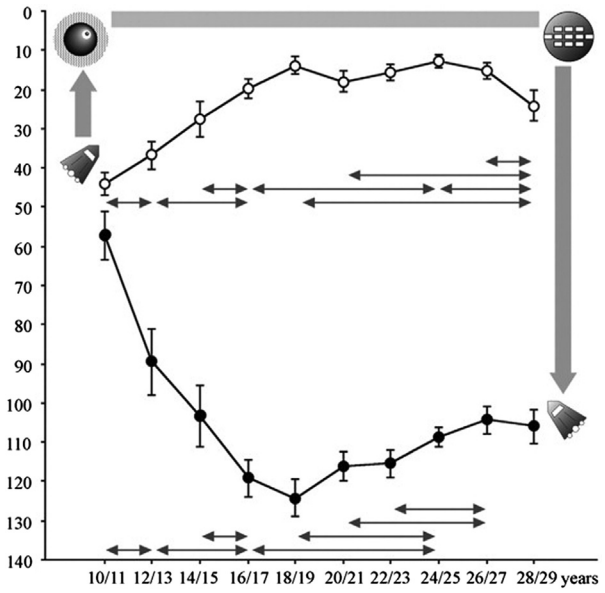

Figure 2 Mean proximity to the security base (black points) and maximal proximity to the source of arousal (white points) in virtual distance units. Error bars indicate standard errors. Horizontal arrows show the nearest age groups which differ with $p \leq 0.05$.

The model does not predict which one of these coping responses is performed in a given context since this is primarily a result of the individual's learning history. However, it is assumed that the internal strategies will succeed if the external strategies do not lead to a success.

Empirical Evidence

The assumptions made by the Zurich Model of Social Motivation were tested by means of several behavior observations and estimation studies. Both types of studies were conducted with either very young children of about one year of age or with older participants ranging from late childhood to adulthood. Since the Zurich Model contains several ideas of attachment theory, settings similar to the Strange Situation Test by Ainsworth et al. (1978) were used for the younger children. For the older participants, a spaceship simulator was developed in which they experienced a space flight which was also designed according to the Strange Situation.

In this simulator, the participants were seated in a cockpit where the virtual space was visible on a computer screen. There were several instruments to manipulate in order to move, orient oneself in space, and protect the spaceship from dangers. The effects of the participant's actions were visible on the screen and the participants could hear sounds, such as their gun firing or the noise of the accelerating engine. In addition, the simulator was moved according to the steering maneuvers of each participant, which ensured a high emotional involvement.

During their virtual flights the participants were confronted with two different objects. One object was considered familiar (either the earth or a home base) and served as an orienting point in space as well as protection. The proximity of the participants to this security base served as a measure for their dependency. A novel object – a strange planet – also appeared during the flight. This planet was fascinating as well as potentially dangerous to the participants. The proximity to this source of arousal was used to assess the enterprise of the participants.

In the late 1980s and throughout the 1990s 14 studies were conducted with this spaceship simulator. Participants were for the most part schoolboys and girls as well as students between 9 and 33 years ($N = 704$, mean AGE = 21.2 years). The data obtained in these studies (Figure 2) strongly support one of the main theoretical assumptions made by the Zurich Model: dependency and enterprise correlate and take an inverse developmental course. The time of the biggest difference is attained at about 18 years, when dependency reaches its minimum while enterprise is at its maximum.

In the spaceship simulator, as well as the laboratory settings used for the younger children, the behavior of the participants was registered continuously. This allowed for investigation methods exceeding simple comparisons of groups by analyzing the distance regulation processes of single participants. In the

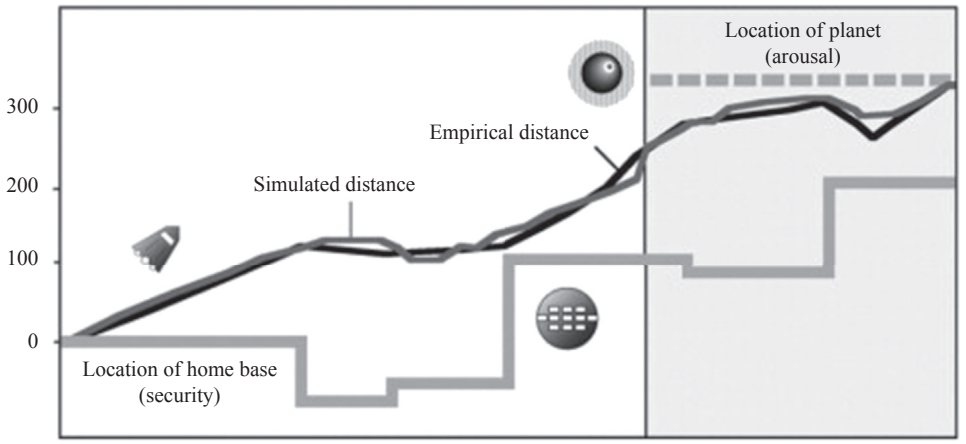

Figure 3 Example of simulated and real distance regulation behavior of one participant (after Gubler et al., 1994, p. 115). The vertical scale indicates virtual distance units, the horizontal scale represents the time of the flight.

estimation studies mentioned above, it was possible to simulate the actual behavior of the participants by means of a parameter estimation based on the mathematical formulation of the Zurich Model (Figure 3). These studies gave further evidence for the validity of the model (for details see Gubler and Bischof, 1991; Gubler et al., 1994).

Future Directions

While the functioning of the security and the arousal systems is sufficiently supported by empirical evidence, there are still several open questions for future research. Although the existing data also support the assumptions made about the autonomy system, the causal link of the autonomy claim to dependency and enterprise still has to be established. Since the Zurich Model was originally conceived from the point of view of a child, more adult motives such as altruism have to be included as well. Furthermore, the study of the developmental course taken by the different reference variables or motives also has to be extended to older age groups in order to get a model explaining the motivational development of the whole life span.

See also: Academic Achievement Motivation, Development of; Motivation and Actions, Psychology of; Motivation, Learning, and Instruction; Motivation: History of the Concept; Self-Determination Theory.

Bibliography

Ainsworth, M.D.S., Blehar, M.C., Waters, B., Wall, S., 1978. Patterns of Attachment: A Psychological Study of the Strange Situation. L. Erlbaum, Hillsdale, NJ.

Bell, R.Q., 1971. Stimulus control of parent or caretaker behavior by offspring. Developmental Psychology 4, 63–72.

Bischof, N., 1975. A systems approach towards the functional connections of attachment and fear. Child Development 46, 801–817.

Bischof, N., 1985. Das Rätsel Oedipus – Die biologischen Wurzeln des Urkonfliktes von Intimität und Autonomie. Piper, Munich, Germany.

Bischof, N., 1993. Untersuchungen zur Systemanalyse der sozialen Motivation I: Die Regulation der sozialen Distanz – Von der Feldtheorie zur Systemtheorie. Zeitschrift für Psychologie 201, 5–43.

Bischof, N., 1995. Struktur und Bedeutung. Eine Einführung in die Systemtheorie. Huber, Bern, Switzerland.

Bowlby, J., 1969. Attachment and Loss: Vol. 1. Attachment. Basic Books, New York.

Burke, P.J., 1991. Identity processes and social stress. American Sociological Review 56, 836–849.

Dörner, J., 1994. Eine Systemtheorie der Motivation. In: Kuhl, J., Heckhausen, H. (Eds.), Enzyklopädie der Psychologie: Motivation, Volition und Handlung. Hogrefe, Göttingen, pp. 329–357, 1996.

Dörner, D., 1998. Bauplan für eine Seele. Rowohlt, Reinbek.

Fogel, A., Thelen, E., 1987. Development of early expressive and communicative action: Reinterpreting the evidence from a dynamic systems perspective. Developmental Psychology 23, 747–761.

Ford, D.H., Lerner, R.M., 1992. Developmental Systems Theory. Sage, Newbury Park, CA.

Gubler, H., Bischof, N., 1991. A systems' perspective on infant development. In: Lamb, M.E., Keller, H. (Eds.), Infant Development: Perspectives from German-speaking Countries. L. Erlbaum, Hillsdale, NJ, pp. 35–66.

Gubler, H., Bischof, N., 1993. Untersuchungen zur Systemanalyse der sozialen Motivation II: Computerspiele als Werkzeug der motivationspsychologischen Grundlagenforschung. Zeitschrift für Psychologie 201, 287–315.

Gubler, H., Paffrath, M., Bischof, N., 1994. Untersuchungen zur Systemanalyse der sozialen Motivation III: Eine Ästimationsstudie zur Sicherheits- und Erregungsregulation während der Adoleszenz. Zeitschrift für Psychologie 202, 95–132.

von Holst, E., Mittelstaedt, H., 1971. The principle of reafference: interactions between the central nervous system and the peripheral organs. In: Dodwell, P.C. (Ed.), Perceptual Processing: Stimulus Equivalence and Pattern Recognition. Appleton, New York, pp. 41–72.

Hyland, M.E., 1988. Motivational control theory: an integrative framework. Journal of Personality and Social Psychology 4, 642–651.

Klein, H.J., 1989. An integrated control theory model of work motivation. Academy of Management Review 14, 150–172.

Lord, R.G., Hanges, P.J., 1987. A control system model of organizational motivation: Theoretical development and applied implications. Behavioral Science 32, 161–178.

Miller, G.A., Galanter, E., Pribram, K.H., 1960. Plans and the Structure of Behavior. Holt, Rinehart, & Winston, New York.

Powers, W.T., 1978. Quantitative analysis of purposive systems: some spadework at the foundations of scientific psychology. Psychological Review 85, 417–435.

Vohs, Kathleen D., Baumeister, Roy F. (Eds.), 2011. Handbook of self-regulation: research, theory, and applications. Guilford Press.

Wiener, N., 1948. Cybernetics. Communication and Control in the Animal and the Machine. Wiley, New York.

Motivational Development: Cross-Cultural Perspectives

Jan Hofer and Holger Busch, University of Trier, Trier, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

People differ in how well the consciously represented, explicit motives they pursue match their implicit motivational system. Such an alignment of explicit and implicit motives has beneficial effects on well-being. The article will first define implicit and explicit motives. Then, consequences resulting from an alignment of motivational systems are highlighted by reviewing research conducted with Euro-American samples as well as with a focus on cross-cultural research. We will demonstrate that implicit motives seem to universally impact the satisfaction derived from goal-pursuance. After summarizing findings on antecedents of motive congruence, an outlook on open questions in research is given.

> "I feel happy when something great was happening. When I am visiting interesting places while travelling or when I am remembering the journeys I've done in the past. I also feel happy when I am at home, when I feel fine, when everybody is well, when I am listening to the radio or when I spend time with my friends." (Elias, 9 years)

It might be odd beginning an article on implicit and explicit perspectives on motivational development across cultures with a boy's statement about happiness. Tackling the central question what makes people happy and flourish (Seligman and Csikszentmihalyi, 2000), we can deduce from his means to be happy that people's motives (e.g., striving for stimulation or close relationships) are crucial in this respect. Conclusive evidence exists that motives energize and direct individuals' behavior and thus shape developmental processes in the long run. If successful, development throughout various stages of life is often seen to be reflected in feelings of happiness and satisfaction (e.g., Erikson, 1959).

Unfortunately, not all motives are made equal. While some motives may be prominent components of the individual's true self (Waterman, 1993), other motivational sources of behavior are rather located outside the individual (Pepitone, 1976). Thus, realization of motives may not always lead to a sense of well-being. Based on the seminal paper by McClelland et al. (1989), we argue that implicit motives are a vital part of the individual's true self. Consequently, they represent a weighing disposition which determines the extent to which certain strivings and activities are considered emotionally pleasant and fulfilling. Above all, individuals able to attune their strivings with their implicit needs are supposed to benefit from success in goal-oriented behavior. We further argue that alignment of different types of motives represents a basic psychological process. Thus, even if culture-bound socialization goals and child rearing practices put varying emphasis on motivational domains, congruence between different motivational components of one's personality is likely to have positive effects on well-being across cultural contexts. In the following, we will first provide a definition of and distinction between implicit and explicit motives. We will then review the evidence for beneficial effects of congruence between implicit and explicit motives and summarize what is known about why individuals differ in how well their implicit and explicit motives are aligned. Finally, we will briefly sketch some open developmental questions in this area.

Dual-Motives Theory: Distinguishing between Implicit and Explicit Motives

In 1989, McClelland, Koestner, and Weinberger offered a compelling solution to a problem that had previously been controversially discussed in motivational psychology: Why were self-report measures of a given motive consistently found to be uncorrelated with thematic apperception measures of the same motive? While this finding had predominantly spawned a debate on the psychometric quality of thematic apperception motive assessment, McClelland and colleagues argued that the two approaches to motive measurement tapped into different motivational systems.

On the one hand, self-attributed or explicit motives (abbreviated *san* for self-attributed need) are those motivational end-states we consciously define as desirable. That is, explicit motives are the goals that we strive for, the plans we set for ourselves, and the values we cherish because we think there is some benefit in attaining them. As people are able to deliberately reflect upon whether they consider a given goal desirable, explicit motives are easily assessed via self-report instruments. The developmental reasoning is that explicit motives are learnt through verbal instructions by parents or other significant socialization agents. Thus, children need to have acquired a certain degree of language development as well as some basic self-concept that motivational concepts such as goals can be integrated into.

On the other hand, implicit motives (abbreviated *n* for need) are less cognitive but rather enduring affective preferences for the approach toward certain incentives and the avoidance of certain disincentives. That is, implicit motives have an impact on whether we seek out a given activity or situation because we expect to derive some pleasure from it (see Hofer and Busch, 2011; on adverse effects on well-being when motives are seen as frustrated). However, this motivational system eludes conscious reflection, although it is not necessarily inaccessible. People cannot report on their implicit motives. Thus, implicit motives need to be assessed through

thematic apperception techniques such as the Picture Story Exercise (PSE; McClelland et al., 1989) which asks participants to make up stories on picture cards depicting ambiguous social situations. Assuming these stories to be motive-toned interpretations of the given picture cues, they are content-coded and used to determine participants' motive strength. Here, the developmental argument is that implicit motives develop early in childhood and are primarily shaped through mechanisms of classical and operant conditioning (McClelland and Pilon, 1983). That is, implicit motives are not integrated into a conscious self-concept but energize and direct behavior on an affective level below conscious representation.

Thus, according to dual-motives theory, explicit and implicit motives represent independent motivational systems. Indeed, as noted before, explicit and implicit measures of the same motive are generally uncorrelated (note that this statistical independence cannot be accounted for by methodological differences between the two measurement approaches alone, cf Thrash, Elliot, & Schultheiss, 2007). There are, however, some individuals whose explicit and implicit motive strength matches well: For example, they score relatively high on both a thematic apperception measure of their implicit achievement motive and a self-report measure of their explicit achievement motive (for a discussion of potential differences between high–high and low–low motive-congruence patterns, see Langan-Fox and Canty, 2010). This phenomenon has been termed motive congruence. For others, there is a discrepancy between their explicit and implicit motive: For example, they might score low in their implicit achievement motive but high in their explicit achievement motive. Individual differences in motive congruence have been demonstrated to have an influence on well-being outcomes. Before we present corresponding evidence, however, we would like to briefly introduce the motives that are usually examined in this context and give a rationale as to why it is instructive to study motive congruence cross-culturally [cf 24 011: cross-cultural psychology].

Implicit motive research has focused on three motives: *n*Achievement, *n*Affiliation-Intimacy, and *n*Power. *n*Achievement is defined as a concern for efficacy and competence; *n*Affiliation-Intimacy is defined as the desire for contact and closeness to others; *n*Power is defined as the desire for having an impact on others' behavior and emotions (see Schultheiss and Brunstein, 2010; for an overview). *n*Affiliation-Intimacy combines two motivational tendencies for interpersonal closeness that are highly correlated but can still have different predictive validity (see Weinberger et al., 2010): Traditionally, *n*Affiliation denotes a concern for closeness for the sake of avoiding solitude, while *n*Intimacy denotes a concern for closeness for the sake of a mutually satisfying relationship.

It has been argued that these motives have evolved because they have increased evolutionary adaptation by increasing the chances of survival and/or reproduction (MacDonald, 1991). Consequently, they ought to be universal properties of human functioning. This does not necessarily mean, however, that all cultures provide the same developmental context for motive development. For example, on the level of explicit motives cultures differ widely in the values they endorse (Schwartz, 1992). Analogously, it seems reasonably safe to assume that cultural variations in parental ethnotheories and subsequent socialization in infancy (Keller, 2007) might lead to culturally diverse shaping of implicit motives. This refers not only to potential cultural differences in motive strength but also to potential differences in how a given motive is expressed. Indeed, there has been a debate about whether *n*Achievement comprises a social-oriented component in cultures stressing group belongingness over individual autonomy because in such cultures individual excellence might entail the disincentive of potential isolation (Maehr and Nicholls, 1980). Still, achievement-related behavior may irrespectively of cultural context be molded by other motives than *n*Achievement alone (see Sorrentino and Shepphard, 1978; for the relation between affiliative motivation and achievement behavior in US-Americans). Hence, we see cross-cultural research on implicit motives as a chance to arrive at insights into aspects of both universal human functioning and culture-specific variation in motivation.

Congruence of Implicit and Explicit Motives: Effects

A variety of studies has confirmed that individuals' well-being is increased when their implicit and explicit motives are in congruence. The general idea is that it is more beneficial in terms of well-being to pursue goals that are in line with affective preferences for certain incentives (i.e., implicit motives) than goals that are not. Pursuance of goals that are not rooted in one's implicit motivational system might also provide some pleasurable experiences (e.g., Hofer et al., 2011), for example through social confirmation. However, generally, scholars agree that it is more beneficial to pursue self-generated goals (e.g., Deci and Ryan, 2000). We argue that how authentically a goal has been self-generated is evidenced in its correspondence to the implicit motivational system.

We will review the evidence on effects of congruence of implicit and explicit motives separately for the three implicit motives in the following. Before we do so, we would like to point out that congruence effects have not only been found for the three single motives but that the broader motivational constructs of agency and communion (Bakan, 1966) have also been examined in this context. Brunstein, Schultheiss, and Grässmann (1998) combined *n*Achievement and *n*Power to form an agency motivation score and used *n*Affiliation-Intimacy as communion motivation score. What they found was that high commitment to motive-congruent goals predicted increases in emotional well-being over four months, while high commitment to motive-incongruent goals was associated with a decrease in well-being. Using the same combinations as Brunstein et al. (1998), Pueschel, Schulte, and Michalak (in press) found that only motive congruent goal progress was associated with less depressive symptoms in their sample of outpatients.

Measuring needs for communion in the context of romantic partnerships on an implicit and an explicit level, Hagemeyer, Neberich, Asendorpf, and Neyer (in press) showed that motive-congruent individuals were more satisfied with their relationship and ran a significantly lower risk of relationship breakup over a one-year period.

With respect to *n*Achievement, Baumann et al. (2005) found that motive congruence was associated subjective well-being. Interestingly, they also found that in a sample of patients diagnosed with heterogeneous psychosomatic and psychiatric disorders achievement motive congruence predicted reduced emotional and somatic complaints. Thus, according to Baumann et al., motive incongruence can be interpreted as a hidden stressor which people are not aware of but still can aggravate psychological symptoms. Similarly, Kehr (2004) hypothesizes and finds in a longitudinal study that motive incongruence – aggregated across all three motives – decreases volitional resources. In support of this notion, Job et al. (2010) demonstrated that motive incongruent individuals have an inclination toward unhealthy eating behavior such as snacking which can be interpreted as a stress-reducing strategy.

The first study to investigate motive congruence effects in a non-WEIRD sample (the acronym WEIRD has been introduced by Henrich et al., 2010; to denote Western, educated, industrialized, rich, and democratic societies which are massively overrepresented in present mainstream psychological literature) was provided by Hofer and Chasiotis (2003). Zambian male adolescents reported on their life-goals and had their implicit motives assessed. Like in all other studies conducted in a non-WEIRD cultural context reported in the following, all instruments were tested on whether they were unbiased in the given cultural sample (see Hofer et al. (2005); for procedural details on testing if a motive-measure is culturally unbiased [cf 23 148: methods of cross-cultural research in developmental psychology]). For *n*Achievement, the expected beneficial effect of motive congruence was identified in that a combination of a high implicit and a high explicit achievement motive was associated with higher levels of life satisfaction.

With respect to *n*Affiliation-Intimacy, two studies on motive congruence have already been referred to: Both Brunstein et al. (1998) and Pueschel et al. (in press) found motive congruence for *n*Affiliation-Intimacy to be beneficial for various measures of well-being, albeit the term communion was used in said studies. Recently, following Weinberger et al.'s (2010) suggestion to analyze *n*Affiliation and *n*Intimacy as two related but yet distinct motives, Busch and Hofer (in preparation a) demonstrated that motive congruence for *n*Intimacy but not for *n*Affiliation was predictive of people's self-reported willingness to settle disagreements with their romantic partner in a cooperative manner.

Particularly interesting from a developmental point of view are results yielded by Hofer et al. (2006): They studied how motive congruence for *n*Affiliation-Intimacy related to interpersonal identity development [cf 23 193: identity formation during adolescence and cultural variations]. According to Marcia's identity status model (e.g., 1966), there are two dimensions characterizing of identity formation: Exploration denotes an active search for identity elements such as beliefs, attitudes, and role perceptions, while commitment denotes the individual's investment in an identity element. Identity diffusion means that there is neither exploration nor commitment; moratorium means that there is exploration but no commitment has yet been made; identity foreclosure means that without any exploration, commitment is made to identity elements that were adopted from parents; finally, identity achievement means that commitment has been made after exploration has taken place. What Hofer, Busch, and colleagues (2006) found was that motive congruent individuals were more likely to have an achieved identity and less likely to have a foreclosed identity in the interpersonal domain (e.g., ideas about friendship, dating). That is, those individuals who set their affiliative goals in line with their implicit affiliation-intimacy motive relied less on others to tell them who they were but found their own answer to this question. Thus, motive congruence is not only linked to positive well-being but also to the outcome of important developmental tasks.

Hofer and Chasiotis (2003) replicated the beneficial motive congruence pattern in their sample of Zambian male adolescents. As was the case with *n*Achievement, a combination of high *n*Affiliation-Intimacy and pronounced importance of intimacy goals was associated with high levels of life satisfaction. Also, in Cameroonian, Costa Rican, and German adults the same beneficial motive congruence effect on general life satisfaction was identified (Hofer et al., 2006): However, explicit motives were not represented by goals but values. In a hierarchy of explicit motives, values are more abstract than goals, thus play a more passive role in influencing actual behavior than goals, and rather denote what a person feels he/she ought to do rather than what he/she wants to do (Emmons, 1989; Kluckhohn, 1951). More specifically, benevolence values were assessed as the concern for others' welfare as a guiding principle in life (Schwartz, 1992). Thus, Hofer et al. (2006) not only demonstrated that motive congruence has positive effects on well-being in very different cultural contexts but also that these effects (at least for *n*Affiliation-Intimacy) hold across various degrees of abstractness of explicit motive formulation.

With respect to *n*Power, only a single study has demonstrated that motive congruence in this domain is associated with increased well-being. Of course, power was included in the agency scores employed by Brunstein et al. (1998) and Pueschel et al. (in press). Power itself, however, was examined only by Hofer et al. (2006) who found that congruence between *n*Power and power values was not related to well-being in their Cameroonian, Costa Rican, and German samples. Following up on this finding, Hofer et al. (2010) could further specify conditions for motive congruence in the power domain to be associated with well-being: In their German and Hong Kong as well as mainland Chinese samples, power values again did not interact with *n*Power to predict well-being. On the other hand, power goals did. As acceptance of institutionalized power is a defining characteristic of cultures (Hofstede, 2001), power values might be too abstract and too devoid of behavioral implications (cf Emmons, 1989) to identify the actual potential of motive congruence effects in this motivational domain. An interesting finding in its own right of Hofer, Busch, Bond, Li, and Law's (2010) study was that power values were positive in terms of individuals' well-being if they were translated into power goals. This suggests that power values might rather be deficiency-oriented, that is, attractive to those who have little

opportunity to experience power, while power goals might rather be approach-oriented.

To conclude, research has consistently found that motive congruence positively affects well-being. That is, such effects have been found for all three basic motives and across various cultural contexts, suggesting that implicit motives may universally be considered as a type of weighing disposition determining whether the pursuance of a particular goal is experienced as satisfying and pleasurable (see Schultheiss et al., 2008). Moreover, motive congruence has been linked to actual behavior (Job et al., 2010) and developmental milestones (Hofer et al., 2006). Thus, motive-congruence seems to be an important factor in individuals' positive development.

Determinants of Congruence of Implicit and Explicit Motives

With motive congruence effects established, researchers have also directed their attention toward the question of why some people pursue motive congruent goals while others pursue rather motive incongruent goals. Experimental studies have shown that imagining the outcomes (e.g., motive-specific positive affect) of motive-directed behavior makes people commit more strongly to motive congruent goals (Job and Brandstätter, 2009; Schultheiss and Brunstein, 1999). That is, people get more willing to pursue goals that allow them to actually experience the emotional states of motive satisfaction they have previously imagined. A second line of research has identified personality characteristics that make some people more likely to set motive congruent goals: Among the traits that were associated with motive congruence were self-determination (Thrash and Elliot, 2002), state orientation (Baumann et al., 2005), private body consciousness, self-monitoring, preference for consistency (Thrash et al., 2007), and referential competence (Schultheiss et al., 2011). These personality characteristics have in common that they denote a person's capacity to gain accurate introspective insight and based on this insight come to conclusions concerning their behavior, including goal setting.

The only study to investigate this issue cross-culturally was conducted by Hofer et al. (2010). Following Thrash and Elliot's (2002) reasoning, they hypothesized self-determination, defined as the ability to get in touch with one's inner needs and direct one's behavior accordingly, to be predictive of achievement motive congruence in Cameroonian, German, and Hong Kong Chinese adults. Results confirmed this assumption, suggesting that introspection universally is a resource of motive congruence. Thus, there is some evidence as to which factors affect motive congruence. More research, however, is urgently needed to get a clearer picture of how individual differences in motive congruence can be accounted for.

Motive Congruence Research: Future Directions and a Developmental Outlook

As reviewed above, motive congruence has consistently proven to be predictive of enhanced well-being and decreased psychological symptoms. Interestingly, it has also been linked to the resolution of a central developmental crisis: motive congruent adolescents were more likely to have an achieved and less likely to have a foreclosed identity status (Hofer et al., 2006). Based on this finding, one might ask if motive congruence has beneficial consequences for personality development in general.

Second, we have reviewed the findings on how individual differences in motive congruence can be accounted for. These findings are meaningful in that they show, with an experimental as well as with a correlational design, that gaining insight into one's implicit motive system allows one to set congruent goals. They do not provide an answer as to how motive congruence develops in the first place. To answer this question, a proper base of knowledge is needed on the development of implicit motives. These two questions, i.e., how motive congruence might foster positive development and how implicit motives develop, will be outlined in some more detail in the following.

Motive Congruence in Positive Development

Implicit motives are known to influence long-term behavioral trends (McClelland, 1987) such as vocational choices (Jenkins, 1994). Moreover, implicit motives have been demonstrated to predict developmental outcomes such as generativity (Hofer et al., 2008; McAdams, 1985). That is, implicit motives may predispose some individuals to successfully resolve certain developmental crises more easily than others. Based on developmental demands, individuals may find themselves confronted with the task of highlighting specific goals, for example generative ones. Given a corresponding implicit motive constellation (high *n*Intimacy as well as high *n*Power; see McAdams, 1985), some people might find these goals more attractive and easier to internalize to begin with. While this is in itself an interesting idea that has received only little attention in psychological research, even less research has been devoted to testing if motive congruence might have an even stronger beneficial effect on positive development. The findings presented by Hofer, Busch, and colleagues (2006) suggest just this for the developmental task of identity formation.

Future research might want to test this assumption for other developmental tasks located at different points across the lifespan. For example, young adults in what Erikson called the intimacy crisis aiming at establishing a long-term, mutually satisfying relationship with a romantic partner might benefit if their intimacy goals are backed up by a corresponding implicit intimacy motive. Also, individuals' life-review might be more positive in old age if people have pursued motive congruent goals in their lives as these can be regarded as reflecting a person's true self (cf Waterman, 1993). In sum, the role of motive congruence in positive development has not been appropriately recognized, let alone complex interactions such as how developmental demands might increase motive congruence. This might be particularly relevant for identity formation in which long-term goals and motivational self-images have to be installed: Implicit motives may be the criterion against which goals and other identity elements are affectively

evaluated and goal commitment may be more likely for those goals that are in line with one's implicit motives.

Development of Implicit Motives

Unfortunately, only little is known about how implicit motives develop. The most influential study has been presented by McClelland and Pilon (1983): They assessed implicit motives in then-adult children of mothers who had been interviewed on child-rearing behavior when their children had been 5 years of age. Relating mothers' description of their child-rearing behavior during the first 5 years of their children's childhood and children's implicit motive scores more than 20 years later, McClelland and Pilon found that *n*Achievement was predicted by the severity of toilet training and *n*Power by mothers' permissiveness for children's sexual and aggressive behavior. Examining *n*Affiliation and *n*Intimacy separately, McClelland and Pilon found the latter to be correlated with the extent of praise by their mothers, while the former was related to mothers' unresponsiveness to their child's crying. These longitudinal data provide information on how maternal parenting shapes implicit motives. However, they rely on mothers' retrospective account of their child-rearing behavior.

In general, beyond these findings, empirically sound knowledge on implicit motive development is scarce. Only few studies have examined implicit motives in children and sought for antecedents of motive strength. Moreover, the studies there often have serious methodological problems such as using picture cues for implicit motive assessment that were originally developed for adults and have not been tested for suitability for children or not using a practice run of the implicit motive assessment for children to get acquainted with the task (e.g., Kagan and Moss, 1959). Recently, a research program has started to assess implicit motives in children with age-adequate picture cues and instructions that present the task in a playful manner. The program has yielded first encouraging results (Busch and Hofer, in preparation b) in that parental child-rearing practices could be identified as predictive of children's motive strength that were comparable to those presented by McClelland and Pilon (1983): For example, *n*Power was highest when parents facilitate both impulsive as well as social behavior in their children.

Thus, PSE or related instruments (e.g., employing preschool versions of the Implicit Association Test which in its adult version has already been used for the assessment of implicit motives; see, e.g., Brunstein and Schmitt, 2004; Cvencek et al., 2011; respectively) for measuring children's implicit motives are currently being developed. With these instruments available much more fine-grained analyses of implicit motive development will be possible which will hopefully lead to a clearer picture of how individual differences in motive strength can be accounted for by childhood experiences.

This established, cross-cultural research provides the opportunity to identify universally valid as well as culture-specific antecedents of implicit motive development. As Greenfield et al. (2003) have argued, across different cultures various developmental pathways can lead to the same developmental outcome. For implicit motive development this could mean that some parenting characteristics might be identified which foster the development of a given motive in various cultural contexts. Still, the way that individuals act out their motivational impulses might be culturally different. For example, as described previously, people in WEIRD societies with their focus on individualism might be more inclined to realize their achievement motive by striving for sticking out as the best. In cultures with a stronger focus on group-belonging, this motive realization might be less well regarded because it entails social isolation. To date, few studies have taken a cross-cultural perspective on which factors might facilitate or impede the realization of implicit motives. For example, on the basis of findings presented by Winter et al. (1998) that behavioral and experiential correlates of implicit motives are moderated by personality traits, Hofer and Busch (in preparation) found personality traits to channel implicit motive realization across cultural groups. While these results indicate cross-cultural similarities, what is nevertheless needed is an open mind for potential cultural variations in how implicit motives are expressed.

To wrap things up, implicit motives have proven to act as weighing dispositions if goal pursuit is perceived as emotionally satisfying. That is, the 9-year-old we quoted at the beginning of the present article might perceive the activities he mentions as making him happy because they match his implicit motive system. Research has shown that motive congruent goal pursuit is associated with enhanced well-being. What is more, this effect has been demonstrated in various cultural contexts. Furthermore, there is first evidence that the resolution of developmental crises can also benefit from motive congruence. This line of research needs to be further elaborated upon, however. Second, research on the development of implicit motives is still in its infancy. Better knowledge on how implicit motives develop is an indispensable basis for understanding in more detail why some people tend to commit to motive incongruent goals and suffer from impaired affective rewards of goal pursuit.

See also: Academic Achievement Motivation, Development of; Cross-Cultural Psychology; Cross-Cultural Research Methods in Psychology; Happiness: History of the Concept; Love and Intimacy, Psychology of; Motivational Development, Systems Theory of; School Achievement: Motivational Determinants and Processes.

Bibliography

Baumann, N., Kaschel, R., Kuhl, J., 2005. Striving for unwanted goals: stress-dependent discrepancies between explicit and implicit achievement motives reduce subjective well-being and increase psychosomatic symptoms. Journal of Personality and Social Psychology 89, 781–799.

Brunstein, J.C., Schmitt, C.H., 2004. Assessing individual differences in achievement motivation with the implicit association test. Journal of Research in Personality 38, 536–555.

Brunstein, J.C., Schultheiss, O.C., Grässmann, R., 1998. Personal goals and emotional well-being: the moderating role of motive dispositions. Journal of Personality and Social Psychology 75, 494–508.

Busch, H., Hofer, J., in preparation a. Intimacy motive-goal congruence predicts relationship satisfaction and cooperative problem-solving – affiliation motive-goal congruence doesn't. Manuscript in preparation.

Busch, H., Hofer, J., in preparation b. Predicting children's implicit motives from parental child-rearing behavior. Manuscript in preparation.

Cvencek, D., Greenwald, A.G., Meltzoff, A.N., 2011. Measuring implicit attitudes of 4-year-olds: the preschool implicit association test. Journal of Experimental Child Psychology 109, 187–200.

Deci, E.L., Ryan, R.M., 2000. The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. Psychological Inquiry 11, 227–268.

Emmons, R.A., 1989. The personal striving approach to personality. In: Pervin, L.A. (Ed.), Goal Concepts in Personality and Social Psychology. Erlbaum, Hillsdale, NJ, pp. 87–126.

Erikson, E.H., 1959. Identity: Youth and Crisis. Norton, New York.

Greenfield, P.M., Keller, H., Fuligni, A., Maynard, A., 2003. Cultural pathways through universal development. Annual Review of Psychology 54, 461–490.

Hagemeyer, B., Neberich, W., Asendorpf, J.B., Neyer, F.J., in press. (In-)Congruence of implicit and explicit communal needs predicts the quality and stability of couple relationships. Journal of Personality.

Henrich, J., Heine, S., Norenzayan, A., 2010. The weirdest people in the world? Behavioral and Brain Sciences 33, 61–135.

Hofer, J., Busch, H., 2011. Satisfying one's needs for competence and relatedness: consequent domain-specific well-being depends on strength of implicit motives. Personality and Social Psychology Bulletin 37, 1147–1158.

Hofer, J., Busch, H., in preparation. The effect of motive-trait interaction on need satisfaction of affiliation among German and Cameroonian adults. Manuscript in preparation.

Hofer, J., Busch, H., Bond, M.H., Kärtner, K., Kiessling, F., Law, R., 2010a. Is self-determined functioning a universal prerequisite for motive-goal congruence? Examining the domain of achievement in three cultures. Journal of Personality 78, 747–779.

Hofer, J., Busch, H., Bond, M.H., Li, M., Law, R., 2010b. Is motive-goal congruence in the power domain beneficial for individual well-being? an investigation in a German and two Chinese samples. Journal of Research in Personality 44, 610–620.

Hofer, J., Busch, H., Chasiotis, A., Kärtner, J., campos, D., 2008. Concern for generativity and its relation to implicit pro-social power motivation, generative goals, and satisfaction with life: a cross-cultural investigation. Journal of Personality 76, 1–30.

Hofer, J., Busch, H., Chasiotis, A., Kiessling, F., 2006a. Motive congruence and interpersonal identity status. Journal of Personality 74, 511–541.

Hofer, J., Busch, H., Kärtner, J., 2011. Self-regulation and well-being: the influence of identity and motives. European Journal of Personality 25, 211–224.

Hofer, J., Chasiotis, A., 2003. Congruence of life goals and implicit motives as predictor of life satisfaction: cross-cultural implications of a study of Zambian male adolescents. Motivation and Emotion 27, 251–272.

Hofer, J., Chasiotis, A., Campos, D., 2006b. Congruence between social values and implicit motives: effects on life satisfaction across three cultures. European Journal of Personality 20, 305–324.

Hofer, J., Chasiotis, A., Friedlmeier, W., Busch, H., Campos, D., 2005. The measurement of implicit motives in three cultures: power and affiliation in Cameroon, Costa Rica, and Germany. Journal of Cross-cultural Psychology 36, 689–716.

Hofstede, G., 2001. Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations across Cultures. Sage, Thousand Oaks, CA.

Jenkins, S.R., 1994. Need for power and women's careers over 14 years: structural power, job satisfaction, and motive change. Journal of Personality and Social Psychology 66, 155–165.

Job, V., Brandstätter, V., 2009. Get a taste of your goals: promoting motive goal congruence through affect-focus goal fantasy. Journal of Personality 77, 1527–1559.

Job, V., Oertig, D., Brandstätter, V., Allemand, M., 2010. Discrepancy between implicit and explicit motivation and unhealthy eating behaviour. Journal of Personality 78, 1209–1238.

Kagan, J., Moss, H.A., 1959. Stability and validity of achievement fantasy. Journal of Abnormal and Social Psychology 58, 357–364.

Kehr, H.M., 2004. Implicit/explicit motive discrepancies and volitional depletion among managers. Personality and Social Psychology Bulletin 30, 315–327.

Keller, H., 2007. Cultures of Infancy. Erlbaum, Mahwah, NJ.

Kluckhohn, C., 1951. Values and value-orientations in the theory of action: an exploration in definition and classification. In: Parsons, T., Shils, E.A. (Eds.), Toward a General Theory of Action. Harper & Row, New York, pp. 388–433.

Kornadt, H.-J., Eckensberger, L.H., Emminghaus, W.B., 1980. Cross-cultural research on motivation and its contribution to a general theory of motivation. In: Triandis, H.C., Brislin, R.W. (Eds.), Handbook of Cross-cultural Psychology, vol. 3. Allyn & Bacon, Boston, MA, pp. 223–321.

Langan-Fox, J., Canty, J.M., 2010. Implicit and self-attributed affiliation motive congruence and depression: the moderating role of perfectionism. Personality and Individual Differences 49, 600–605.

MacDonald, K.B., 1991. A perspective on Darwinian psychology: the importance of domain-general mechanisms, plasticity, and individual differences. Ethology and Sociobiology 12, 449–480.

Maehr, M.L., Nicholls, J.G., 1980. Culture and achievement motivation: a second look. In: Warren, N. (Ed.), Studies in Cross-cultural Psychology, vol. 2. Academic Press, New York, pp. 221–267.

Marcia, J.E., 1966. Development and validation of ego identity status. Journal of Personality and Social Psychology 3, 551–558.

McAdams, D.P., 1985. Power, Intimacy, and the Life Story: Personological Inquiries into Identity. Dorsey Press, Homewood, IL.

McClelland, D.C., 1987. Human Motivation. Cambridge University Press, New York.

McClelland, D.C., Koestner, R., Weinberger, J., 1989. How do self-attributed and implicit motives differ? Psychological Review 96, 690–702.

McClelland, D.C., Pilon, D.A., 1983. Sources of adult motives in patterns of parent behavior in early childhood. Journal of Personality and Social Psychology 44, 564–574.

Pepitone, A., 1976. Toward a normative and comparative biocultural social psychology. Journal of Personality and Social Psychology 34, 641–653.

Pueschel, O., Schulte, D., Michalak, J., in press. Be careful what you strive for: the significanca of motive-goal congruence for depressivity. Clinical Psychology and Psychotherapy.

Schultheiss, O.C., Brunstein, J.C., 1999. Goal imagery: bridging the gap between implicit motives and explicit goals. Journal of Personality 67, 1–38.

Schultheiss, O.C., Brunstein, J.C. (Eds.), Implicit Motives. Oxford University Press, New York.

Schultheiss, O.C., Jones, N.M., Davis, A.Q., Kley, C., 2008. The role of implicit motivation in hot and cold goal pursuit: effects on goal progress, goal rumination, and depressive symptoms. Journal of Research in Personality 42, 971–987.

Schultheiss, O.C., Patalakh, M., Rawolle, M., Liening, S., MacInnes, J.J., 2011. Referential competence is associated with motivational congruence. Journal of Research in Personality 45, 59–70.

Schwartz, S.H., 1992. Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. Advances in Experimental Social Psychology 25, 1–65.

Seligman, M.E.P., Csikszentmihalyi, M., 2000. Positive psychology: an introduction. American Psychology 55, 5–14.

Sorrentino, R.M., Shepphard, B.H., 1978. Effects of affiliation-related motives on swimmers in individual versus group competition: a field experiment. Journal of Personality and Social Psychology 36, 704–714.

Thrash, T.M., Elliot, A.J., 2002. Implicit and self-attributed achievement motives: concordance and predictive validity. Journal of Personality 70, 729–755.

Thrash, T.M., Elliot, A.J., Schultheiss, O.C., 2007. Methodological and dispositional predictors of congruence between implicit and explicit need for achievement. Personality and Social Psychology Bulletin 33, 961–974.

Waterman, A.S., 1993. Two conceptions of happiness: contrasts of personal expressiveness (eudaimonia) and hedonic enjoyment. Journal of Personality and Social Psychology 64, 678–691.

Weinberger, J., Cotler, T., Fishman, D., 2010. The duality of affiliative motivation. In: Schultheiss, O.C., Brunstein, J.C. (Eds.), Implicit Motives. Oxford University Press, New York, pp. 71–88.

Winter, D.G., John, O.P., Stewart, A.J., Klohnen, E.C., Duncan, L.E., 1998. Traits and motives: toward an integration of two traditions in personality research. Psychological Review 105, 230–250.

Motor Control Models: Learning and Performance

Pietro G Morasso, Italian Institute of Technology, Genoa, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The focus of the article is on the variety of attempts that have been investigated for capturing the complexity of purposive action and adaptive behavior, having defined a coordinated action as a class of movements plus a goal. Redundancy is a side effect of this connection, and thus redundancy is necessarily task oriented, something to be managed 'online' and rapidly updated as the action unfolds. The article then analyzes two main computational mechanisms that have been proposed as candidates of how the brain may deal with motor redundancy: (1) the force-field-based solution known as Equilibrium-Point Hypothesis (EPH) and (2) the cost-function-based solution to the degrees of freedom problem, namely Optimal Control Theory. However, both theories apply only to overt actions where force fields and cost functions are directly related to the interaction of the body with the physical world in the course of a real action. Considering that overt actions are just the tip of the iceberg, hiding the vast domain of covert actions that are the skeleton of motor cognition, an extension of EPH is described, Passive Motion Paradigm (PMP). The relationships between PMP, the simulation theory of covert actions, internal models, and the body schema concept are also analyzed. Finally, general learning mechanisms that may support the acquisition of internal computational modules are briefly summarized.

Modeling the way in which humans learn to coordinate their movements in daily life or in more demanding activities is an important scientific topic from many points of view, such as the medical, psychological, kinesiological, and cybernetical. The article analyzes the complexity of this problem and reviews the variety of experimental and theoretical techniques that have been developed for this purpose.

With the advent of technical means for capturing motion sequences and the pioneering work of Marey (1894) and Muybridge (1957) in this area, the attempt at describing, modeling, and understanding the organization of movement has become a scientific topic. The fact that human movements are part of everyday life paradoxically hides their intrinsic complexity and justifies initial expectations that complete knowledge could be achieved simply by improving the measurement techniques and carrying out a few carefully designed experiments. Unfortunately, this is not the case. Each experiment is frequently the source of more questions than answers, and thus the attempt to capture the complexity of purposive action and adaptive behavior, after a century of extensive multidisciplinary research, is far from over.

The conventional view is based on a separation of perception, movement, and cognition and the segregation of perceptual, motor, and cognitive processes in different parts of the brain, according to some kind of hierarchical organization. This view is rooted in the empirical findings of neurologists of the nineteenth century, such as J. Hughlings Jackson, and has a surprising degree of analogy with the basic structure of a modern PC that typically consists of input and output peripherals connected to a central processor. Perhaps the analogy with modern technology justifies why this old-fashioned attitude still has its supporters, in spite of the massive empirical and conceptual challenge to this view and its inability to explain the range of skills and adaptive behaviors that characterize biological organisms.

Let us consider perception, which is the process whereby sensory stimulation is translated into organized experience. That experience, or percept, is the joint product of the stimulation and of the process itself, particularly in the perception and representation of space. An early theory of space perception put forth by the Anglican bishop G. Berkeley at the beginning of the eighteenth century was that the third dimension (depth) cannot be directly perceived in a visual way, because the retinal image of any object is two-dimensional, as in a painting. He held that the ability to have visual experiences of depth is not inborn but can only result from logical deduction based on empirical learning through the use of other senses.

The first part of the reasoning (the need of a symbolic deductive system for compensating the fallacy of the senses) is clearly wrong, and the roots of such misconception can be traced back to the neoplatonic ideas of the Italian Renaissance in general, and to Alberti's window metaphor in particular. Also, the Cartesian dualism between body and mind is just another face of the same attitude and such *Descartes' error*, to quote Damasio (1994), is on a par with the *Berkeley's error* described in the preceding paragraph, and is at the basis of the intellectualistic effort to explain the computational complexity of perception that characterizes a great part of the classic artificial intelligence approach. However, the latter part of Berkeley's conjecture (the emphasis on learning and intersensory integration) is surprisingly modern and agrees, on one hand, with the modern approach to neuropsychological development pioneered by Piaget (1963), and on another hand, with the so-called *connectionist* point of view, originated in the 1980s as a computational alternative to classic artificial intelligence.

An emergent idea is also the motor theory of perception, well illustrated by Berthoz (1997); that is, the concept that perception is not a passive mechanism for receiving and interpreting sensory data but is the active process of anticipating the sensory consequences of an action and thereby binding the sensory and motor patterns in a coherent framework. In computational terms, this implies the existence in the brain of some kind of 'internal model,' as a bridge

between action and perception. As a matter of fact, the idea that the instructions generated by the brain for controlling a movement are utilized by the brain for interpreting the sensory consequences of the movement is already present in the pioneering work of Helmholtz and von Uexküll, and its influence has resurfaced in the context of recent control models based on learning (e.g., Wolpert and Kawato, 1998). The generally used term is 'corollary discharge' (von Holst and Mittelstaedt, 1950), and implies an internal comparison between an outgoing signal (the *efferent copy*) and the corresponding sensory *re-afference*: the coherence of the two representations is the basis for the stability of our sensorimotor world. This kind of circularity and complementarity between sensory and motor patterns is obviously incompatible with the conventional reasoning based on hierarchical structures. A similar kind of circularity is also implicit in Piaget's concept of 'circular reaction,' which is assumed to characterize the process of sensorimotor learning; that is, the construction of the internal maps between perceptually identified targets and the corresponding sequence of motor commands.

An additional type of circularity in the organism/environment interaction can be identified at the mechanical interface between the body and the outside world, where the mechanical properties of muscles interact with the physics of inanimate objects and gravity. This topic area has evolved from the Russian school, with the early work on the nature of reflexes by I.P. Pavlov and the subsequent critical reexamination by Anokhin (1974) and Bernstein (1967). In particular, we owe to Bernstein the seminal observation (the *comparator model*) that motor commands alone are insufficient to determine movement but only identify some factors in a complex equation where the world dynamics has a major influence. This led, among other things, to the identification of muscle stiffness as a relevant motor parameter and the formulation of the theory of equilibrium-point control (Feldman and Levin, 1995; Bizzi et al., 1992).

In general, we may say that, in different ways, Helmholtz's corollary discharge, Piaget's circular reaction, and Bernstein's comparator model are different ways to express the *ecological nature* of motor control; that is, the *partnership* between brain processes (including muscles) and world dynamics. On top of this, another type of circularity from a more cognitive point of view is suggested by the parietofrontal *mirror circuit* (Rizzolatti and Sinigaglia, 2010), which is proposed as the unique mechanism "that allows an individual to understand the action of others *from the inside* and gives the observer a first-person grasp of the motor goals and intentions of other individuals."

On the other hand, these general ideas on motor control could not provide, immediately, mathematical tools of analysis from which to build models and perform simulations. The art and science of building motor control models is a later development and has been influenced by the methods designed by engineers in the field of automatic control and computer science. Most of the techniques are based on linear approximations and explicit schematizations of the phenomena. In particular, two main concepts can be singled out for their influence on the study of motor control: the concept of *feedback* and the concept of *motor program*, both of them originating from the seminal work on cybernetics initiated by Norbert Wiener (1948). Their level of influence in brain theory is certainly determined by the tremendous success of these techniques in the modern technological world. However, their direct applicability to what we may call the *biological hardware* is questionable for three main reasons: (1) feedback control can only be effective and stable if the feedback delays are negligible, but this is not the case for biological feedback signals, where transduction and transmission delays add up to tens of milliseconds; (2) the concept of motor program implies a sequential organization, which only can be effective if the individual steps in the sequence are sufficiently fast, and this is in contrast with the parallel, distributed processing of the brain made necessary by the relative slowness of synaptic processing; and (3) the degrees of freedom (DoF) problem, further analyzed in the following section, implies that engineering modeling techniques, typically conceived for low-dimensional applications, cannot be scaled up easily due to the so-called curse of dimensionality.

Motor Redundancy and the 'Cybernetics of Purposive Actions'

Since the time of Nicholas Bernstein (1967) it has become clear that one of the central issues in the neural control of movements is the 'degrees of freedom problem,' that is, the computational process by which the brain coordinates the action of a high-dimensional set of motor variables for carrying out the tasks of everyday life, usually described and learned in a 'task-space' of much lower dimensionality. Such dimensionality imbalance is usually named 'motor redundancy.' This means that the same movement goal can be achieved by an infinite number of combinations of the control variables and timing patterns, which are equivalent as far as the task is concerned. But in spite of so much freedom, experimental evidence suggests that the brain consistently uses a narrow set of solutions. Consider, for example, the task of reaching a point B in space, starting from a point A, in a given time T. In principle, the task could be carried out in an infinite number of ways, with regards to spatial aspects (hand path), timing aspects (speed profile of the hand), and recruitment patterns of the available DoFs (kinematic DoF). In contrast, it was found that the spatiotemporal structure of this class of movements is strongly stereotypical, whatever their amplitude, direction, and duration: the path is nearly straight (in the extrinsic, Cartesian space, not the intrinsic, articulatory space), and the speed profile is nearly bell shaped, with symmetric acceleration and deceleration phases (Morasso, 1981). That this stereotypicity should be attributed to internal control mechanisms, not to biomechanical effects, is suggested by the analysis of reaching movements in different types of neuromotor impaired subjects and the adaptation of normal subjects to disturbing force fields (Shadmehr and Mussa-Ivaldi, 1994).

A movement, per se, is nothing unless it is associated with a goal and this usually requires recruitment of a number of joints, in the context of an action. Recognizing the crucial importance of multijoint coordination was really a paradigm shift from the classical Sherringtonian viewpoint (typically

focused on single-joint movements), to the Bernsteinian quest for principles of coordination or synergy formation. A coordinated action is a class of movements plus a goal. Redundancy is a side effect of this connection and thus redundancy is necessarily task oriented, something to be managed 'online' and rapidly updated as the action unfolds.

Generally speaking, actions can be considered as operational modules in which descending motor patterns are produced together with the expectation of the (multimodal) sensory consequences. Mounting evidence accumulated in the last decades from different directions and points of view, such as the equilibrium point hypothesis, mirror neurons system, and motor imagery, suggest that in order to understand the neural control of movement, the observation and analysis of overt movements is just the tip of the iceberg, because what really matters is the large computational basis shared by action production, action observation, action reasoning, and action learning.

How the Brain May Deal with Motor Redundancy

Let us consider again the stereotypicity of reaching movements that emerges in a robust way from the redundancy of the human motor system: where is it coming from? Two main alternatives, although with many variants, have been investigated in the last decades: one is based on force fields and nonlinear attractor dynamics and the other on cost functions and optimal control concepts.

The Force-Field-Based Solution to the DoF Problem: The Equilibrium-Point Hypothesis

The best-known example of a force-field-based solution to the DoF problem is the Equilibrium Point Hypothesis (EPH) (Asatryan and Feldman, 1965; Feldman, 1966; Bizzi et al., 1976, 1992; Feldman and Levin, 1995). The force field comes from the elastic properties of muscles and their capacity to store and release mechanical energy.

The elastic properties of muscles are well captured by the so-called λ-model (Feldman and Levin, 1995). In this model, λ is the controllable parameter that sets the activation threshold of the monosynaptic stretch reflex and thus determines the 'rest length' of the 'muscle spring.' Its value is specified by supraspinal motor commands and expresses a combination of the desired levels of muscle length and stiffness. By setting the λ-commands of all the muscles, the brain implicitly codes an equilibrium point, determined by the fact that at this point the spring actions cancel each other. In this way, movement follows as a mechanical consequence of the force field, without a continuous intervention of the brain, and the brain can take advantage of a second type of redundancy (in addition to the 'kinematic redundancy,' determined by the excess number of DoFs), namely 'muscle redundancy,' determined by the excess number of muscles in relation to the number of DoFs.

A crucial feature is that muscles are not linear springs but are characterized by length–tension curves of exponential type, whose stiffness depends on the difference between the λ-command and the actual muscle length. Therefore, the brain has the opportunity of independently controlling two variables: (1) the global equilibrium point, by setting the λ-commands equal to the muscle lengths of a desired posture (the reciprocal component of the λ-commands); and (2) the global stiffness of the joints, by adding to the previous pattern a set of coactivation λ-commands in such a way that the stronger the coactivation, the stronger the stiffness.

Beyond the specific neuromuscular details of the λ-model, EPH is important because it assigns a computational role to the muscles, in addition to its obvious executive action. Summing up, the power of the EPH comes from its ability to solve the DoF problem by positing that the redundant posture of the whole body is not directly controlled by the brain in a detailed way but is the biomechanical consequence of the equilibrium among a large set of muscular and environment forces. In this view, movement is a symmetry-breaking phenomenon, that is, the transition from an equilibrium state to another.

The Cost-Function-Based Solution to the DoF Problem: Optimal Control Theory

Optimal control theory is a classical engineering design technique for controlling complex systems in which infinite solutions are possible, given a desired task or behavior. The general idea is that in order to design the best possible controller of a system, capable to carry out a prescribed task, one should define first a 'cost function,' that is, a mathematical combination of the control variables that yields a single number (the 'cost' of the action). This function is generally composed of two parts: a part that measures the 'distance' of the system from the goal of the action and a part (regularization or penalty term) that encodes the required 'effort.' The design is then reduced to the computation of the control variables that minimize the cost function, thus finding the best possible trade-off between accuracy and effort.

The first attempt to apply this approach to human motor control was carried out by Flash and Hogan (1985), by proposing the 'integrated jerk' as the regularization term of the cost function. 'Jerk' is the time derivative of acceleration, and thus minimizing jerk is equivalent to maximize the smoothness of the generated trajectory. They showed that the solution of such minimization tasks for point-to-point, planar reaching movements was indeed consistent with the spatiotemporal invariances found by Morasso (1981). However, other simulation studies found similar results by choosing different types of cost functions, such as 'integrated torque change' (Uno et al., 1989) or 'motor noise dependent statistics' (Harris and Wolpert, 1998). Thus, it is not clear how to identify the cost function, supposedly used by the brain, because different alternatives tend to yield similar behaviors.

In this line of research, optimal control concepts were used for deriving offline optimal control patterns, to be employed in feed-forward control schemes. A later development (Todorov and Jordan, 2002) suggested using an extension of optimal control theory that incorporates sensory feedback in the computational architecture. In this closed-loop control technique, a block named 'control policy' generates a stream of motor commands that optimize the predefined cost function on the basis of a current estimate of the 'state variables'; this

estimate integrates in an optimal way (by means of a Kalman filter) feedback information (coming from delayed and noise-corrupted sensory signals) with a prediction of the state provided by a forward model of the system's dynamics, driven by an 'efference copy' of the motor commands. One of the most attractive features of this formulation, in addition to its elegance and apparent simplicity, is that it blurs the difference between feed-forward and feedback control because the control policy governs both. On the other hand, the mathematical computations that need to be carried out in order to identify the optimal control policy are quite complex and do not scale up well with dimensionality. Moreover, optimal feedback control requires indeed feedback, and thus is unable to treat overt and covert actions in a uniform way.

Beyond EPH: The Passive Motion Paradigm as a General Approach to Action Generation

In the EPH framework, the computational core of the mechanism is based on force fields and the associated attractor dynamics rather than on cost functions. In the standard formulation, the source of the force fields is physical and is determined by the mechanical properties of muscles. Therefore, this model can only be applied to overt, not covert, movements. However, it is possible to abstract the model to the attractor dynamics of interacting neural representations, and this is proposed by the Passive Motion Paradigm (PMP) (Mussa-Ivaldi et al., 1988). The basic idea can be formulated in qualitative terms by suggesting that the process by which the brain can determine the distribution of work across a redundant set of joints, when the chosen end-effector is assigned the task of reaching a target point in space, can be represented as an internal simulation process that calculates how much each joint would move if an externally induced force (i.e., the goal) pulls the end-effector by a small amount toward the target. This internal simulation in turn causes the incremental elastic reconfiguration of the internal *body schema* involved in generating the action, by disseminating the force field across the kinematic chain (more generally, task-specific kinematic graph) that characterizes the articulated structure of the human or robotic body. The mechanism is labeled 'passive' in line with the EPH. The underlying hypothesis is that the equilibrium point is not explicitly specified by the brain, which just contributes to the activation of 'task-related' force fields. The mechanism is robust and can be easily extended to high-dimensional problems because there is no need to solve ill-posed inverse problems but only to run an internal simulation of the dynamic body schema. This mechanism applies equally well to covert and overt actions, and this is consistent with the mounting evidence from brain-imaging studies in support of common neural substrates being activated during both 'real and imagined' movements. For this reason, it is plausible to posit that also real, overt actions are the results of an internal simulation, where such simulation is a result of the interactions between an 'internal body model' with the attractor dynamics of force fields induced by the goal and task-specific constraints involved during the performance of any action. If the mental simulation converges (i.e., the goal is realized), then the movement can be executed. Otherwise, convergence failure may play the role of a crucial internal event, namely the starting point to break the action plan into a sequence of subactions, by recruiting additional DoFs, employing tools that may allow the realization of the goal, and so forth. In this sense, PMP can be considered a generalization of EPH from action execution (overt actions) to action planning and reasoning about actions (covert actions).

The dynamics of PMP networks is driven by a kind of internal potential energy, associated with the previously described attractor dynamics. The basic idea is not far away from the 'free-energy principle' advocated by Friston (2010) in order to account for action, perception, and learning in a unified way; and it is, at the same time, an alternative to the optimal control theory (Mohan and Morasso, 2011).

Examples of PMP Networks

Let $\vec{q}$ be the set of all the DoFs that characterize the human body or the body of a humanoid robot, possibly extended by including the DoFs of a tool. Any given task identifies one or more end-effectors, and is defined by the motion of one end-effector with respect to another or to some reference point. The natural reference frame for $x(t)$ is linked to the environment (extrinsic) space and not the joint (intrinsic) space. Moreover, the dimensionality of $\vec{q}$ is generally much greater than the dimensionality of $\vec{x}$.

The basic idea of the PMP is to express the goal of an action (e.g., 'reach a target point $\vec{x}_T$') by means of an attractive force field $\vec{F}_{tgt} = K\,(\vec{x}_T - \vec{x})$, centered in the target position and apply it to the body schema, in particular to the task-related end-effector. This force field is mapped into a set of torques applied to the joints according to the Jacobian matrix of the arm: $\vec{T}_{tgt} = J^T K\,(\vec{x}_T - \vec{x})$. The displacement of the articulations is then determined by an admittance matrix A that distributes the motion among the (redundant) set of joints: $d\vec{q} = A\,J^T K\,(\vec{x}_T - \vec{x})$. By integrating over time, this equation and mapping articulations displacements into small displacements of the end-effector ($d\vec{x} = J\,d\vec{q}$), the whole body schema will evolve from the initial equilibrium configuration to a final, desired configuration where the force applied to the end-effector is null because the end-effector has reached the target. This dynamics of the body schema can also be expressed by means of a graph (or kinematic network) such as that in the top panel of Figure 1. The graph also includes a nonlinear gain that allows to reach stability in a prescribed time, according to the theory of *terminal attractors* (Zak, 1988). The relaxation process, from the initial equilibrium state to the final one, is analogous to the mechanism of coordinating the motion of a wooden marionette by means of strings attached to the terminal parts of the body: the distribution of the motion among the joints is the 'passive' consequence of the virtual forces applied to the end-effectors and the virtual compliance of the joints. The extension from a single limb to the whole body of humans or humanoid robots is quite straightforward, as shown in the bottom panel of Figure 1.

This modeling framework can be successfully applied for explaining the formation of Whole Body Reaching (WBR) synergies, that is, coordinated movement of lower and upper limbs, characterized by a focal component (the hand must

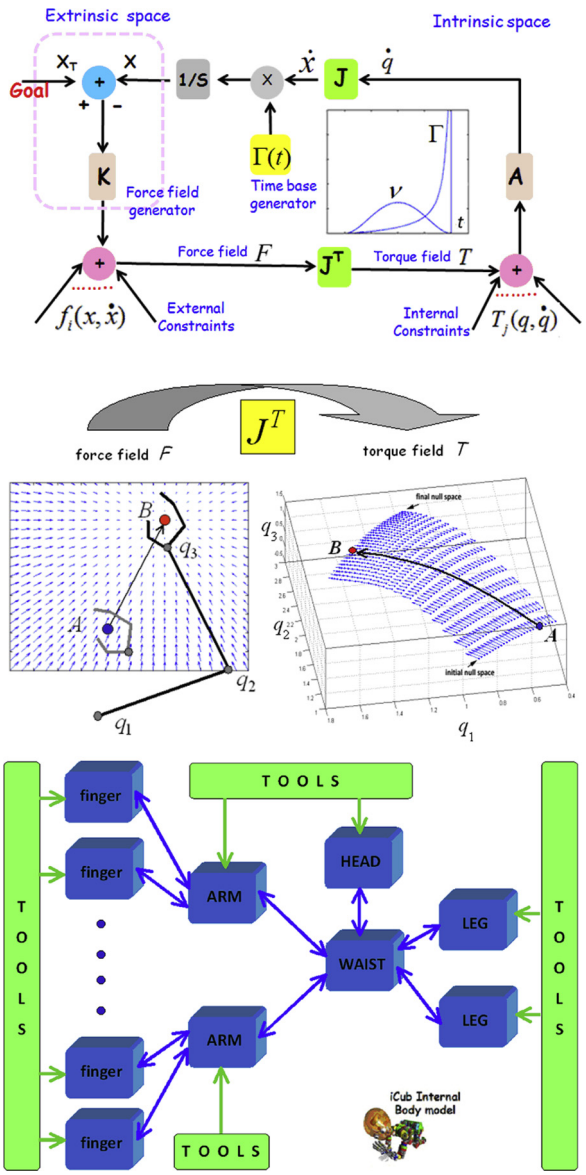

Figure 1 **Top panel**. Basic kinematic network that implements the passive motion paradigm for a simple kinematic chain. It includes the Gamma function, which endows the system with terminal attractor dynamics. This means that equilibrium is not achieved asymptotically but in finite time. (The corresponding speed profile is also shown.) External and internal constraints (represented as task-dependent force/torque fields) bias the path to equilibrium in order to take into account suitable 'penalty functions.' This is a multi-referential system of action representation and synergy formation, which integrates a forward and an inverse internal model. **Middle panel**. The figure illustrates the key element of the architecture of Figure 1 for solving the degrees of freedom problem, namely the mapping of the force field, defined in the extrinsic space and applied to the end-effector, into the corresponding torque field, defined in the intrinsic space and applied to the joints. The mapping is implemented by means of the transpose Jacobian matrix of the kinematic transformation. Dimensionality reduction is obtained implicitly by letting the internal model 'slide' in the torque field. Each point of the trajectory in the extrinsic space corresponds to a whole manifold in the intrinsic space (the 'null space' of the kinematic transformation). The equilibrium point in the force field corresponds to an equilibrium manifold in the torque field. The selection among the infinite number of possible targets is carried out implicitly by the combination of different force/torque fields. **Bottom panel**. Full kinematic network of the iCub robot (53 DoFs; see Sandini, G., et al., 2004). Each blue box corresponds to a basic kinematic network (top panel) for a specific body segment. 'Tools' corresponds to possible interaction points between the external world and the internal body mode; the corresponding green arrows identify potential goals and the related goal-oriented force fields. Such force fields are propagated through the global network (blue arrows), searching for a task-dependent equilibrium configuration. A single or multiple time-base generators control the timing. Sandini, G., Metta, G., Vernon, D., 2004. RobotCub: an open framework for research in embodied cognition. In: Proceed. 4th IEEE/RAS Intl Conf. on Humanoid Robots, LA, CA, 13–32.

reach a target) and a postural component (the center of mass (CoM) must remain inside the support base) (Morasso et al., 2010). The focal component of the task was modeled by means of a simple attractive field to the target, applied to the fingertips of both hands. The postural component was implemented by an additional force field applied to the pelvis region. By simulating the network in various conditions it was possible to show that it exhibits many of the spatiotemporal features found in experimental data of WBR in humans (Stapley et al., 1999; Pozzo et al., 2002; Kaminski, 2007), in particular the fact that the speed profiles of the hand and the CoM are synchronized (Figure 2).

Figure 3 shows the PMP network used for the bimanual coordination of the movements of a humanoid robot. The network, in this case, is composed of three parts, one related to the left arm, a second related to the right arm, and a third to the trunk. Task-specific networks are logically derived from the global whole-body network (the global body schema) by 'grounding' some element of the network, that is, inhibiting the propagation of the task-related force fields beyond a given articulation or body part. In this sense, the body schema represented by PMP networks is not a passive map but is a dynamic structure that recruits at run-time the redundant DoF of the body in the context of task-related force fields.

PMP and the Simulation Theory of Covert Actions

Experimental results (in terms of EEG, fMRI, PET, and NIRS) generally support the idea of common underlying functional networks subserving both the execution and imagination of movements (Kranczioch et al., 2009; Munzert et al. 2009). Marc Jeannerod (2001) went a step forward by formulating the Mental Simulation Theory (2001), which posits that cognitive motor processes such as motor imagery, movement observation, action planning, and verbalization share the same representations with motor execution. The neural activation patterns include not only premotor and motor areas such as the PMC (premotor cortex), SMA (supplementary motor area), and M1 (primary motor cortex) but also subcortical areas of the cerebellum and the basal ganglia. In particular, the presence of activity in the typically motor regions suggests that covert actions are in the same motor format that is required by overt actions. The concurrent activation of descending motor pathways might be involved in the generation of efference copies that propagate upstream

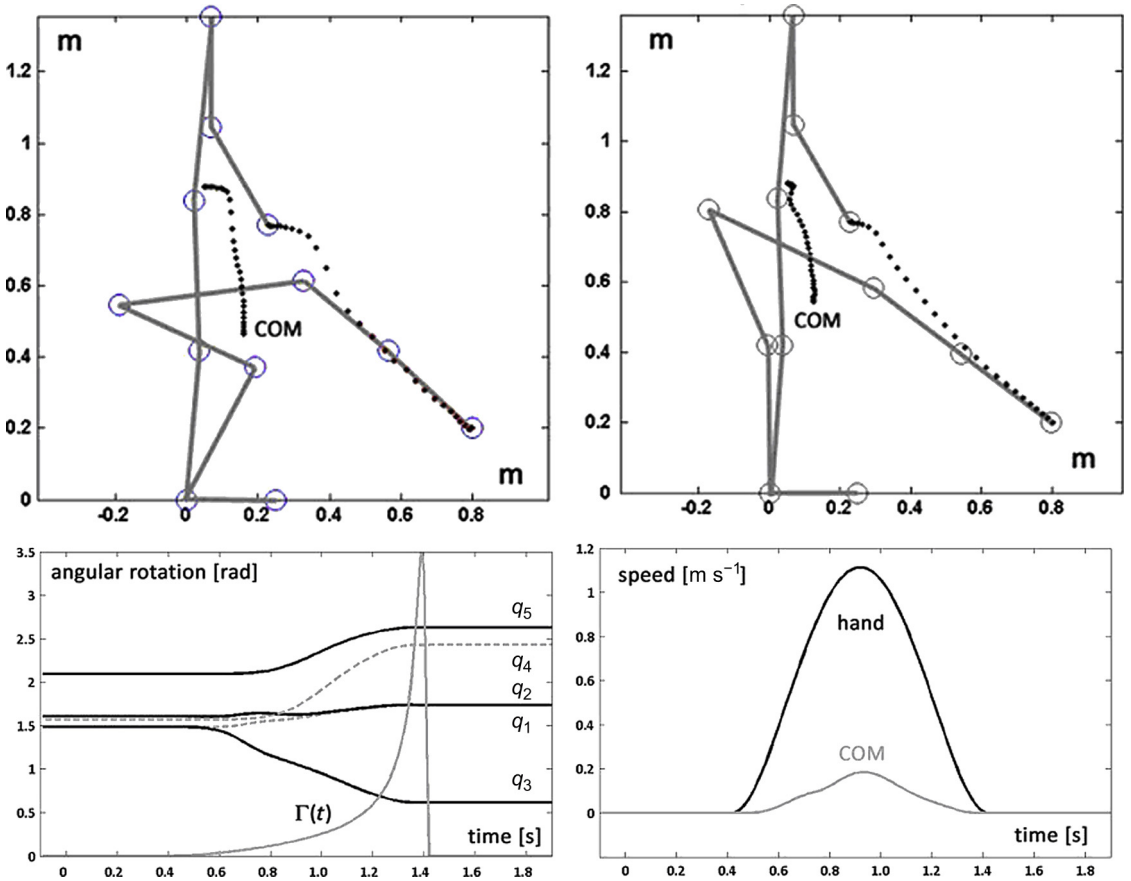

Figure 2 Coordination of whole-body reaching movements by means of the passive motion paradigm. For the same target, the top panel shows two motion patterns obtained with different values of the admittance matrix A, namely the hip admittance is reduced from 2.5 rad s^{-1} Nm (left graph) to 0.1 rad s^{-1} Nm (right graph). The bottom panel shows the time course of the joint rotations, including the time base generator $\Gamma(t)$, (left graph) and the speed profiles of the end-effector and the center of mass.

to parietal and premotor cortices, thus predicting the potential consequences of the planned action. Jeannerod interprets this brain activity as an 'internal simulation' of a detailed representation of action and uses the term *S-states* for the corresponding mental states. The crucial point, in our view, is that if S-states occurring during covert actions are to a great extent quite similar to the states occurring during overt actions, then it is not unreasonable to posit that also real, overt actions are the results of an internal 'simulation' process. This is the basic idea behind the PMP. From this point of view, the simulation of PMP networks is a way to generate what Jeannerod calls 'S-states.'

A closely related issue is that of understanding actions of conspecifics and learning by imitation. One of the explanations proposed is that observation of other's actions would activate, in the observer's brain, the same mechanisms that would be activated, were that action imagined by the observer himself or herself (Gallese and Goldman, 1998). Imitation, as suggested by Iacoboni, could be based on directly matching the observed action onto an internal simulation of that action (Iacoboni, 2009). Internal simulations hence play an important role in allowing the observer to foresee the consequence of an action, predict the intended goal of the actor, and learn to replicate the action through imitation. In a recent study, Mohan et al. (2011) demonstrated how humanoid robots can learn a range of motor skills by observing a teacher, with PMP framework as a central building block. They further demonstrated how abstract motor knowledge acquired by the humanoid robot iCub while learning one skill (drawing) can be reused while learning another, quite different skill (controlling a 2 DoF toy-crane), hence drastically speeding up learning. These results further support the idea of PMP being a unified computational framework for execution, reasoning, understanding, and imitation of action, thus suggesting a strong link between PMP, EPH, simulation theory of covert actions, and mirror neurons systems. In sum, PMP networks can be activated under a variety of conditions in relation to action, either of oneself or observed from other individuals. Their function is not only to shape the motor output during action execution, but also to provide the self with information on the feasibility, consequence, understanding, and meaning of potential actions. What remains out of this article is the set of processes that allow 'action schemas,' generated by a PMP mechanism, to interact with neuromuscular and external world dynamics.

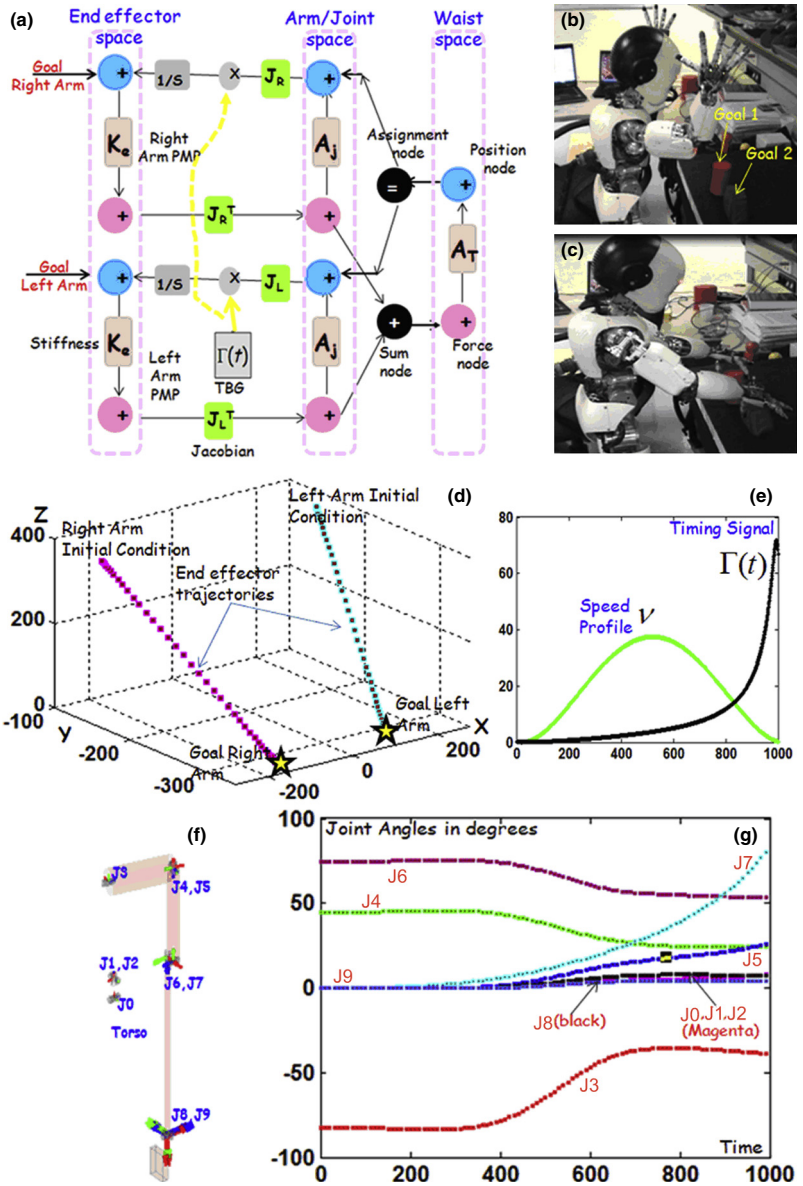

Figure 3 Bimanual coordination task (reaching two objects at the same time) for the iCub robot implemented by means of the passive motion paradigm. **Panel (a)**: kinematic network, with two target goals and a single time-base generator. The network includes three modules: (1) right arm, (2) left arm, and (3) waist. The dimensionality of $\mathbf{J_R}$ & $\mathbf{J_L}$ is 3×10 (this includes the seven DoF's of the arms and the three DoF's of the waist). The dimensionality of $\mathbf{A_j}$ is 7×7 and of $\mathbf{A_T}$ is 3×3. The three subnetworks interact through a pair of nodes ('assignment' and 'sum') that allow the spread of the goal-related activation patterns. This simplified network is obtained from the global network of **Figure 2** by 'grounding' the waist, that is, by inhibiting the spread of activation to the lower limbs. **Panels (b and c)** show the initial and the final postures of the robot and the two cylindrical target objects. **Panels (d and e)** show the trajectories of the two end-effectors and the corresponding speed profiles (together with the output $\Gamma(t)$ of the time base generator). **Panel (f)** clarifies the intrinsic degrees of freedom in the left-arm-torso chain. **Panel (g)** shows the time course of the waist and left-arm joint rotation patterns: J0–J2: joint angles of the waist (yaw, roll, pitch); J3–J9: joint angles of the right arm (shoulder pitch/yaw/roll; elbow flexion/extension; wrist pronation/supination/pitch/yaw).

Learning Paradigms in Neural Networks and Motor Control

At the core of the theories of neural network models is the attempt to capture general approaches for learning from experience tasks that are too complex to be expressed by means of explicit or symbolic models (Arbib, 1995). The mechanism of learning and memory has been an intriguing question after the establishment of the neuron theory at the turn of the nineteenth century (Ramón y Cajal, 1928) and the ensuing conjectures that memories are encoded at synaptic sites (Hebb, 1949) as a consequence of a process of learning. In accordance with this prediction, synaptic plasticity was first discovered in the hippocampus, and nowadays it is generally thought that LPT (long-term potentiation) is the basis of cognitive learning and memory, although the specific mechanisms are still a matter of investigation.

Three main paradigms for training the parameters or *synaptic weights* of neural network models have been identified: (1) 'Supervised learning,' in which a teacher or a supervisor provides a detailed description of the desired response for any given stimulus and exploits the mismatch between the computed and the desired response or error signal for modifying the synaptic weights according to an iterative procedure. The mathematical technique typically used in this type of learning is known as *back propagation* and is based on a gradient-descent mechanism that attempts to minimize the average output error (Rumelhart et al., 1986); (2) 'Reinforcement learning,' which also assumes the presence of a 'supervisor' or a teacher but its intervention is only supposed to reward (or punish) the degree of success of a given control pattern, without any detailed input–output instruction (Sutton and Barto, 1998). The underlying mathematical formulation is aimed at the maximization of the accumulated reward during the learning period; (3) 'Unsupervised learning,' in which there is no teacher or explicit instruction and the network is only supposed to capture the statistical structure of the input stimuli in order to build a consistent but concise internal representation of the input. The typical learning strategy is called Hebbian, in recognition of the pioneering work of D.O. Hebb, and is based on a competitive or self-organizing mechanism that uses the local correlation in the activity of adjacent neurons and aims at the maximization of the mutual information between stimuli and internal patterns.

How to link the learning paradigms above, which have been derived for explaining the plasticity of specific neural networks, to learning paradigms that apply to complex 'naturalistic' behaviors and thus involve a number of networks as well as the dynamics of the outside world, is still an open question that will require significant experimental and theoretical improvements in the coming years.

See also: Cerebral Cortex; Classical Mechanics and Motor Control; Motor Cortex; Self-Organizing Dynamical Systems.

Bibliography

Anokhin, P.K., 1974. Biology and Neurophysiology of Conditioned Reflexes and Their Role in Adaptive Behaviour. Pergamon Press, Oxford, UK.

Arbib, M.A., 1995. The Handbook of Brain Theory and Neural Networks. MIT Press, Cambridge, MA.

Asatryan, D.G., Feldman, A.G., 1965. Functional tuning of the nervous system with control of movements or maintenance of a steady posture. Biophysics 10, 925–935.

Bernstein, N.A., 1967. The Coordination and Regulation of Movement. Pergamon Press, Oxford, UK.

Berthoz, A., 1997. Le sens du mouvement. Édition Odile Jacob, Paris.

Bizzi, E., Polit, A., Morasso, P., 1976. Mechanisms underlying recovery of final head position. Journal of Neurophysiology 39, 435–444.

Bizzi, E., Hogan, N., Mussa Ivaldi, F.A., Giszter, S.F., 1992. Does the nervous system use equilibrium-point control to guide single and multiple movements? Behavioral and Brain Sciences 15, 603–613.

Damasio, A.R., 1994. Descartes' Error. Emotion, Reason and the Human Brain. Putnam Press, New York.

Feldman, A.G., 1966. Functional tuning of the nervous system with control of movement or maintenance of a steady posture, II: controllable parameters of the muscles. Biophysics 11, 565–578.

Feldman, A.G., Levin, M.F., 1995. The origin and use of positional frames of references in motor control. Behavioral and Brain Sciences 18, 723–745.

Flash, T., Hogan, N., 1985. The coordination of arm movements: an experimentally confirmed mathematical model. Journal of Neuroscience 7, 1688–1703.

Friston, K., 2010. Free energy principle: a unified brain theory? Nature Neuroscience 11, 127–138.

Gallese, V., Goldman, A., 1998. Mirror neurons and the simulation theory of mind reading. Trends in Cognitive Sciences 2, 493–501.

Harris, C.M., Wolpert, D.M., 1998. Signal-dependent noise determines motor planning. Nature 394, 780–784.

Hebb, D.O., 1949. The Organization of Behavior. Wiley, New York.

Iacoboni, M., 2009. Neurobiology of imitation. Current Opinion In Neurobiology 19, 661–665.

Jeannerod, M., 2001. Neural simulation of action: a unifying mechanism for motor cognition. Neuroimage 14, 103–109.

Kaminski, T.R., 2007. The coupling between upper and lower extremity synergies during whole body reaching. Gait Posture 26, 256–262.

Kranczioch, C., Mathews, S., Dean, J.A., Sterr, A., 2009. On the equivalence of executed and imagined movements. Human Brain Mapping 30, 3275–3286.

Marey, E.J., 1894. Le mouvement. Édition Masson, Paris.

Mohan, V., Morasso, P., 2011. Passive motion paradigm: an alternative to optimal control. Frontiers in Neurorobotics 5 (Art. 4), 1–28.

Mohan, V., Morasso, P., Zenzeri, J., Metta, G., Srinivasa Chakravarthy, V., Sandini, G., 2011. Teaching a humanoid robot to draw 'Shapes'. Autonomous Robots 31, 21–53.

Morasso, P., 1981. Spatial control of arm movements. Experimental Brain Research 42, 223–227.

Morasso, P., Casadio, M., Mohan, V., Zenzeri, J., 2010. A neural mechanism of synergy formation for whole body reaching. Biological Cybernetics 102, 45–55.

Munzert, J., Lorey, B., Zentgraf, K., 2009. Cognitive motor processes: the role of motor imagery in the study of motor representations. Brain Research Reviews 60, 306–326.

Mussa-Ivaldi, F.A., Morasso, P., Zaccaria, R., 1988. Kinematic networks. A distributed model for representing and regularizing motor redundancy. Biological Cybernetics 60, 1–16.

Muybridge, E., 1957. The Human Figure in Motion. Dover Press, New York.

Piaget, J., 1963. The Origin of Intelligence in Children. Norton Press, New York.

Pozzo, T., Stapley, P.J., Papaxanthis, C., 2002. Coordination between equilibrium and hand trajectories during whole body pointing movements. Experimental Brain Research 144, 343–350.

Ramón y Cajal, S., 1928. Regeneration in the Vertebrate Central Nervous System. Oxford University Press, Oxford, UK.

Rizzolatti, G., Sinigaglia, C., 2010. The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. Nature Reviews Neuroscience 11, 264–274.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536.

Shadmehr, R., Mussa-Ivaldi, F.A., 1994. Adaptive representation of dynamics during learning of a motor task. Journal of Neuroscience 14, 3208–3224.

Stapley, P.J., Cheron, G., Grishin, A., 1999. Does the coordination between posture and movement during human whole-body-reaching ensure center of mass stabilization? Experimental Brain Research 129, 134–146.

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning. MIT Press, Cambridge, MA.

Todorov, E., Jordan, M.I., 2002. Optimal feedback control as a theory of motor coordination. Nature Neuroscience 5, 1226–1235.

Uno, Y., Kawato, M., Suzuki, R., 1989. Formation and control of optimal trajectory in human multijoint arm movement. Minimum torque-change model. Biological Cybernetics 61, 89–101.

von Holst, E., Mittelstaedt, H., 1950. Das Reafferenz prinzip. Wechselwirkungen zwischen Zentral nerven system und Peripherie. Naturwissenschaften 37, 464–476.

Wiener, N., 1948. Cybernetics or Control and Communication in the Animal and the Machine. Hermann & Cie, Paris; MIT Press, Cambridge, MA.

Wolpert, D.M., Kawato, M., 1998. Internal models of the cerebellum. Trends in Cognitive Science 2, 338–347.

Zak, M., 1988. Terminal attractors for addressable memory in neural networks. Physics Letters A 133, 218–222.

Motor Cortex

Steven P Wise, Olschefskie Institute for Neurobiology of Knowledge, Potomac, MD, USA; and Edmond and Lily Safra International Institute of Neurosciences of Natal, Natal, Brazil

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The motor cortex evolved in mammals to supplement an ancestral motor system. It provides three advantages in the control of movements: more accurately graded forces, the generation of novel motor synergies, and the separation of existing motor synergies into individually controlled components. Plasticity in motor cortex neurons underlies learned skills and contributes to recovery of function after brain damage or disease. Recently, neural signals from the motor cortex have been used to control mechanical devices or to directly excite muscles, thus bypassing damaged parts of the motor system and restoring a measure of independence to patients with motor disabilities.

Definition

The motor cortex has been defined in two ways. According to the narrow definition, the motor cortex corresponds to Brodmann's area 4, an area that contains a motor representation of the entire body and sends outputs that control voluntary movements. In this sense, the motor cortex corresponds to the primary motor cortex, often abbreviated as M1 or MI. According to the broad definition, the motor cortex encompasses the primary motor cortex as well as a set of additional motor areas collectively called the premotor cortex or the non-primary motor cortex. This article applies the narrow definition. Article Premotor Cortex on the premotor cortex deals with the motor cortex in the broad sense. Defined in either way, the motor cortex is a part of the cerebral cortex and of the neocortex (*see* Cerebral Cortex). In primates, these areas compose the caudal (posterior) part of the frontal lobe, and in other mammals they are usually located in the rostral (anterior) part of the cerebral cortex (Figure 1).

Identifying Features of the Primary Motor Cortex

Like other parts of the cerebral cortex, the primary motor cortex has a laminated structure. It differs from most of the cerebral cortex, however, in that it lacks a clearly identifiable layer 4, also known as the internal granular layer. For this reason, the primary motor cortex is classed among the agranular areas of neocortex, along with several premotor areas and some caudal parts of the prefrontal cortex. In many primates, the primary motor cortex differs from other areas in that it is significantly thicker than other cortical areas.

In addition to its agranular architecture, three features distinguish the primary motor cortex from other parts of the neocortex. Electrical stimulation of the primary motor cortex evokes movements more easily than it does from other cortical areas. This trait is often measured by passing small amounts of electrical current through an electrode that has been inserted directly into the cortex, an experimental procedure known as intracortical microstimulation. Although some other areas come close, the electrical currents needed to evoke a movement consistently are usually smaller in the primary motor cortex than in other areas. The second defining feature involves thalamic connectivity. The primary motor cortex receives relatively direct inputs, via neural relays in the dorsal thalamus, from both the basal ganglia and the cerebellum. The third defining feature depends on topology, the spatial relationships among structures. The primary motor cortex lies immediately rostral to the somatosensory cortex.

In some primate species, the primary motor cortex has an additional defining feature. The neurons that project to the spinal cord have their cell bodies in layer 5. Some of these neurons reach a conspicuously large size and rank among the largest neurons in the cerebral cortex. Called Betz cells, these giant pyramidal cells can reach 60 μm in width in some monkeys and up to 120 μm in humans.

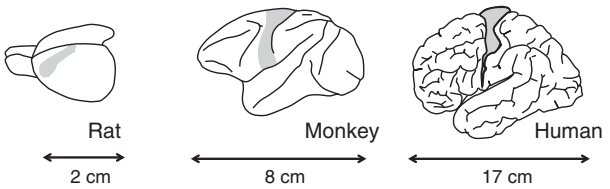

Figure 1 A rat brain, a rhesus monkey brain, and a human brain showing the approximate locations of the primary motor cortex (M1) in gray.

Connections of the Primary Motor Cortex

Cells in the primary motor cortex send outputs to and receive inputs from somatosensory cortex, the posterior parietal cortex, and the premotor cortex. Each of these regions consists of several areas, and nearly all of these projections are reciprocal. In addition to other cortical areas, some cells in the primary motor cortex project to the basal ganglia (specifically, to the striatum), to the cerebellum (via the basilar pontine nuclei), to the red nucleus, to the reticulospinal neurons of the brainstem, and to the spinal cord, among other structures.

At least parts of each of these structures play a role in controlling movements.

The older literature contains discussions of an *extrapyramidal motor system*, one largely dependent on the basal ganglia, as contrasted with a *pyramidal motor system* that emanates from the primary motor cortex. This term refers, not to the pyramidal cells that give rise to the corticospinal projections, but rather to the fiber tract that these and other axons form in the brainstem: the pyramidal tract. The idea of an extrapyramidal motor system has been discarded for several reasons. First, some basal ganglia function is mediated via its connections with the primary motor cortex, including pyramidal-tract outputs from the motor cortex. Second, a key part of the so-called extrapyramidal motor system, the supplementary motor area, was classed as such because it was once said to lack pyramidal-tract and corticospinal projections. Later research, however, showed that this nonprimary motor area has direct projections to the spinal cord that are nearly as extensive as those of the primary motor cortex. In fact, the primary motor cortex only contributes ~40% of the axons that compose the pyramidal tract, with axons from the premotor, somatosensory, and posterior parietal areas contributing the remainder. Third, basal ganglia function extends beyond the control of movements.

History of Motor Cortex Studies

Compared to other parts of the motor system, the primary motor cortex has a short evolutionary history, as the next section explains. As a subject of research, however, it has a very long history. About 2500 years ago, physicians of the Hippocratic school noticed that damage to the cerebral cortex on one side of the brain caused paralysis on the opposite side of the body. By the nineteenth century, the brain damage causing this motor impairment had been localized to a restricted part of the frontal lobe. Independently, G. Theodor Fritsch and Eduard Hitzig in Germany and David Ferrier in England came to believe that the caudal-most part of the frontal lobe controlled voluntary movement, and they set out to confirm this hypothesis experimentally. They used the then-novel phenomenon of electricity to evoke movements by stimulating the cerebral cortex. Unlike the method of intracortical microstimulation, mentioned above, these early researchers used much stronger currents, which they applied to the surface of the brain. The experiments of Fritsch, Hitzig, and Ferrier demonstrated that signals emanating from the motor cortex in dogs and monkeys produced movements on the opposite side of the body. This point is taken up again in the section on the motor map.

At the time, this research raised controversies that have reawakened in recent decades. Antivivisectionists, the forerunners of today's animal rights advocates, condemned the research of Fritsch, Hitzig, and Ferrier and argued then, as their successors do today, that research on dogs and monkeys does not apply to humans. These critics were wrong then, as they are today. The early neuroscience of Fritsch, Hitzig, and Ferrier established the existence of the motor cortex in a way that proved to have as much validity for humans as it did for other mammals. Although the primary motor cortex has become modified during evolution, no one now doubts that the primary motor cortex in dogs, humans, and monkeys, along with many other mammals, such as rats (Figure 1), derives from an ancestral species common to all of them.

Evolution of Motor Cortex

The primary motor cortex evolved in mammals, as did the other parts of the neocortex. As a result, only mammals have a primary motor cortex in the sense used here. Because the motor cortex evolved long after many of the other neural structures that control movement – such as the cerebellum, the red nucleus, parts of the basal ganglia, the reticulospinal system, and the spinal cord – it probably evolved to improve the performance of the remainder of the motor system. After the primary motor cortex developed, the motor system as a whole changed and some of the subcortical motor-control structures became less important.

The exact emergence of the primary motor cortex during mammalian evolution remains less certain than for the primary sensory areas, such as the primary (striate) visual cortex, primary auditory cortex, and primary somatosensory cortex (Kaas, 2011). The reason is that, unlike these and some other sensory areas, it has proven difficult to establish whether there is a separate primary motor cortex in certain groups of mammals. These mammals, specifically marsupials and monotremes, descend from species that diverged from placental mammals relatively early in mammalian history. For many sensory areas of neocortex, it has been possible to identify homologues in monotremes and marsupials, as well as in a broad diversity of placental mammals. For the primary motor cortex, however, no corresponding area has been convincingly identified in monotremes or marsupials, and there is evidence that the primary somatosensory cortex of marsupials has some of the properties that, in other mammals, characterize the primary motor cortex. As a result of these findings, it seems most likely that a separate primary motor cortex first emerged in early placental mammals, although some of its functions might have been performed by the primary somatosensory cortex in marsupial and monotreme mammals.

Although all placental mammals have a primary motor cortex, this area has adapted and changed in each species. For example, in New World monkeys with prehensile tails, much of the primary motor cortex controls tail movements. Likewise, in many primate species, including humans, much of it became specialized for hand use.

In particular, anthropoid primates (monkeys, apes, and humans) evolved a reliance on somatosensory (haptic) processing. As in all placental mammals, the primary motor cortex of anthropoid primates sends a massive corticospinal projection. In most mammals, however, the axons of the corticospinal neurons do not terminate directly on motor neurons or they do so only sparsely. By tracing axonal connections across multiple synapses in rhesus monkeys, it has been possible to show that the rostral part of the anthropoid primary motor cortex shares this characteristic with other mammals (Rathelot and Strick, 2006). Its corticospinal projections terminate mainly on spinal interneurons, not on motor neurons. That is, they lack monosynaptic connections with motor neurons. In contrast to the rostral part of the primary motor cortex, its caudal part projects monosynaptically to spinal motor

neurons, at least in anthropoid primates. This part of the primary motor cortex receives inputs predominantly from sensory receptors in the skin (cutaneous inputs) that have high sensitivity. The appearance of two divisions of the primary motor cortex probably occurred in anthropoid primates, and the existence of these two areas has been confirmed in studies of the human brain. The caudal part, the one dominated by cutaneous inputs, lies in the banks of the central sulcus in most anthropoids and has been called the caudal primary motor cortex in monkeys (abbreviated M1c) or posterior area 4 in humans (abbreviated area 4p).

The importance of this anthropoid specialization appears to relate to hand use. Cutaneous inputs probably play a role in the manipulation of objects by providing feedback concerning skin deformation for movements so fine that they do not change joint angles or muscle lengths very much, if at all. (The section on the function of motor cortex takes up the function of proprioceptive inputs, which signal muscle length.) This anthropoid specialization for haptic exploration likely evolved to provide advantages in fruit selection. The compliance of fruits to manual deformation often provides the most useful and sometimes the only, information about ripeness. In support of this idea, a comparison of nine species of anthropoids showed that the density of Meissner corpuscles correlates with the extent of each species' fruit consumption (Hoffmann et al., 2004). Meissner corpuscles are sensitive cutaneous receptors that respond to small amounts of skin stretching and deformation. The relationship between the density of Meissner corpuscles and a frugivorous diet probably reflects a variable ability to use hardness cues to assess the ripeness of fruits. Low-force finger movements also provide information about the texture of objects. By analogy with the visual system, this specialization of the hand – and especially of the fingers and finger tips – has been called a tactile 'fovea.' Sustained electrical stimulation of the primary motor cortex evokes coordinated movements resembling the manipulation of objects (Graziano, 2005).

Motor cortex function is intimately related to that of the parietal cortex, which has also changed dramatically during evolution. Specifically, the posterior parietal cortex is very small in most nonprimate mammals, including the close relatives of primates such as tree shrews. Thus most of its subdivisions likely evolved in early primates, in which they project to the primary motor cortex and to the premotor cortex. Sustained electrical stimulation of the posterior parietal cortex evokes coordinated reaching, defensive, and grasping movements, much as it does from the motor cortex. Each part of the posterior parietal cortex receives a different set of visual and somatosensory inputs and each projects to primary motor and premotor areas from which stimulation evokes similar movements (Kaas et al., 2011). Accordingly, these paired parietal and frontal regions probably act as functional modules for particular coordinated actions.

The Motor Map

For the primary motor cortex, the existence of a systematic topographic representation of the body has been recognized since the nineteenth century (*see* Topographic Maps). By observing the progression of epileptic seizures, the British neurologist J. Hughlings Jackson discerned the topographic organization of the primary motor cortex. In the motor cortex, it takes the form of somatotopy, a map of the body. From medial to lateral along the primary motor cortex, cells control the leg and foot, the arm and hand, and the face, mouth, tongue, and lips, in an ordered sequence.

The finer organization of the motor cortex is more complex and more difficult to summarize than the well-ordered topography of the somatosensory, visual, and auditory systems. As a general statement, the fingers, toes, and other distal parts of the body are more prominently represented in the caudal part of the primary motor cortex, whereas proximal representations are more prevalent rostrally. There are many exceptions to this tendency, however.

There has been some controversy over the extent of ipsilateral motor control from the primary motor cortex. For primates, it has long been known that the primary motor cortex sends a significant proportion of its corticospinal projection to the ipsilateral spinal cord. Until recently, however, it was not known whether these projections directly synapse upon spinal motor neurons. Recent research has confirmed that nearly all of the relatively direct influences of the primary motor cortex go to muscles on the side of the body opposite to each hemisphere. Stimulation of the primary motor cortex does not reveal any monosynaptic inputs to motor neurons on the ipsilateral side of the spinal cord. Using a technique called spike-triggered averaging, it is possible to show that the discharge of an individual cell in the primary motor cortex of monkeys correlates with a transient increase in muscle activity. This occurs, however, only for muscles contralateral to the hemisphere in which the cell is located. And although some neural discharges seem to correlate with movements of the ipsilateral limb, a careful analysis has shown that this activity can be attributed to small muscle contractions in the contralateral limb (Soteropoulos et al., 2011).

The Function of Motor Cortex

Despite more than a century of research, the function of the primary motor cortex is still not well understood. The earliest understanding came from studying the effects of brain lesions. A number of severe motor disabilities follow such lesions, including weakness, the loss of fine motor control, and in some circumstances complete paralysis. According to one widely held view, the primary motor cortex permits the fractionation of movement. This phrase refers to the independent control of muscle groups that usually work in concert with each other, as synergistic muscles. According to this idea, the fractionation of movement adds flexibility to the motor system by dissociating the components of motor synergies. A related idea is that the motor cortex underlies the construction of novel motor synergies. Discussions of this idea usually involve finger movements, but the primary motor cortex has a representation of the entire body, so its functions must apply more generally. The fine control of face, lip, and tongue movements also reflects flexible synergies, for example. The primary motor cortex does not, however, have any appreciable role in the control of eye movements.

Muscles versus Movements

A longstanding issue regarding the primary motor cortex is known as the muscles versus movement debate. This topic engaged researchers for many years, and it became controversial because it encompassed several separate issues. The question is: does the motor cortex control muscles or movements?

Edward Evarts developed a method to answer one aspect of this question. In the early 1960s, he recorded the activity of individual neurons in the primary motor cortex of monkeys as they made simple limb movements or generated forces against a mechanical device without moving. He found that motor cortex cells had much the same activity in either case. This answered the question from one perspective: the primary motor cortex generates forces, and these forces sometimes (but not always) cause the changes in joint angles that lead to movements. Evarts's findings have been repeatedly confirmed, both in experiments involving forces against variable loads and forces generated against an immovable object. In this sense, the motor cortex controls muscles (forces), not movements per se.

But there are other aspects of the same question, and in this sense it remains an open one. Hiroshi Asanuma suggested that each cell in the motor cortex controlled a single muscle. He used intracortical microstimulation to show that as the current decreases, the stimulation eventually causes the twitch of only one muscle. With more sophisticated methods, however, Paul Cheney and Eberhard Fetz showed that individual cells control the activity of many muscles, often those that act synergistically. They found, using spike-triggered averaging, that the activity of single cells in the primary motor cortex caused increases in muscle activity for several, synergistic muscles at more-or-less the same time. So, in this sense, the motor cortex controls coordinated movements, not (individual) muscles.

Yet another aspect of the question concerns the coordinate frame in which movements are controlled. Some cells in the primary motor cortex reflect a limited amount of information about how these movements occur in visual space, but for the most part the transformation from visual coordinates to motor commands occurs in posterior parietal and premotor areas. This aspect of the debate remains active. Although there is little doubt that the primary motor cortex controls muscles (forces), it may also have a role in controlling movements (joint-angle changes).

Population Coding

Apostolos Georgopoulos and his colleagues showed that neurons in the primary motor cortex encode the direction of hand movements. Each motor cortex cell has its greatest activity for movement in one direction, with systematically less activity as the direction of movement diverges from that *preferred direction*. On the assumption that these cells generate movements in their preferred direction, it is possible to estimate a collective directional signal emanating from a neuronal population. This computation, termed the *population vector*, matches the trajectory of movements very well. The same applies to forces without movement, although most research has involved actual movements of the hand.

Many different kinds of population codes have been shown to have the same basic properties as the population vector, and sometimes it is not necessary to record the activity of isolated neurons to detect something about intended movement direction. Although activity at the single-cell level provides the most information about the trajectory of movements, less refined neural signals also encode this information. In recent years, for example, it has been possible to decode movement direction from the clustered activity of dozens or hundreds of cells and from small electrical potentials, called local field potentials (LFPs), that reflect the activity of thousands of neurons and synapses.

Voluntary and Reflex Movement

The primary motor cortex has long been known to control voluntary movements, but it also plays a role in reflexive movements, called transcortical reflexes. For example, cells in the primary motor cortex integrate sensory information about coordinated shoulder and elbow movements. These signals can counter imposed forces within 50 ms, and exciting the motor cortex with transcranial magnetic stimulation potentiates this reflex response (Pruszynski et al., 2011).

Recent evidence shows that the flow of cutaneous information to the motor cortex is strongly regulated by the brain. All movements cause sensory signals from peripheral receptors, such as the Meissner corpuscles mentioned earlier. Gating mechanisms regulate the flow of this information to the motor cortex. During preparation for a movement, the primary motor cortex, but not the somatosensory cortex or the spinal cord, shows suppression of cutaneous inputs. Furthermore, the greater such suppression, the faster that monkeys can generate the planned movement (Seki and Fetz, 2012). Thus motor planning facilitates movements, in part by suppressing cutaneous inputs.

Motor Learning

During motor learning, such as mastering a sequence of movements or learning to overcome imposed forces in order to reach directly to a target, regional blood flow rates change in several motor areas, including the primary motor cortex. Article Cortical Areas Engaged in Movement: Neuroimaging Methods on the cortical areas engaged in movement, as revealed by neuroimaging methods, discusses these findings.

Single-cell studies in monkeys have revealed changes in the primary motor cortex that mirror the rate of motor learning and its decay. When monkeys experience forces that perturb their arm during reaching movements, several trials of practice enable them to overcome these forces. The number of such perturbations is important, but so too is their frequency (number of perturbations per unit time). The more frequently the monkeys experience the perturbation, the faster they can learn making straight movements to their target. With faster learning, however, comes faster decay of those memories and a reversion of the original prelearning state. Some neurons in primary motor cortex change their activity in a way that reflects the total number of experienced perturbations and not their frequency. Unlike the monkey's behavior, these plastic changes decay at the same rate for faster and slower learning. Other cells, however, depend on both the number and frequency of

perturbations, with faster neuronal changes occurring when learning is faster and a faster decay occurring thereafter, as observed at the behavioral level (Mandelblat-Cerf et al., 2011). These findings support the idea that plastic changes in the primary motor cortex underlie motor learning.

Neural Plasticity and Rehabilitation

In the United States, nearly 1% of the population has suffered a stroke that caused a motor disability. Recovery of function usually proceeds slowly and often remains incomplete. Research has shown that because of brain plasticity, organizational changes occur in the remaining primary motor cortex and in nearby areas that promote the recovery of function (Barbay et al., 2006). This plasticity reflects changes in the balance between cortical inhibition and excitation, which in turn changes the motor maps and level of activation.

A month of heightened plasticity seems to follow lesions of the primary motor cortex. From research on nonhuman animals, it appears that intense rehabilitative training within the first five days or so of brain damage may cause yet additional damage. Waiting for more than a month, however, seems to decrease both the plasticity and the recovery of function. In brain-damaged humans, rehabilitation efforts often come later and with less intensity than indicated by experiments on nonhuman animals. Future improvements in rehabilitation management might therefore involve increases in the intensity and frequency of treatments in a period from one to four weeks after the injury (Krakauer et al., 2012).

Neural Prosthetics

Neural prosthetics promise to assist motor control and communication in patients with damage to the motor cortex or its output pathways to muscles. These devices are intended to restore a measure of independence to these patients, who retain the ability to plan and imagine movements.

The most common approach is to deliver motor commands from the brain to some apparatus, often mediated by a computer. Such brain–machine or brain–computer interfaces can involve an apparatus as simple as a switch or a cursor on a video display, or one as complex as a robotic arm or an artificial limb. Some methods involve recording signals external to the brain, but many have involved the insertion of electrodes directly into the primary motor cortex. The neural commands used to operate these devices similarly range from electroencephalograpic and magnetoencephalographic signals to LFPs and cell recordings.

So far, the clinical applications have been limited, but two paralyzed patients have gained some control over their environment by using motor cortex activity to control various prosthetic devices, including a robotic arm (Hochberg et al., 2012). Another patient has regained some communicative skills (Kennedy et al., 2011).

In addition to mechanical devices and video displays, neural signals can be used to stimulate muscles directly. For example, monkeys have learned to use artificial connections from motor cortex cells to muscles in order to generate movements (Moritz et al., 2008; Ethier et al., 2012). In one of these studies, monkeys simultaneously controlled several neuron–muscle pairs to make hand movements. Likewise, a human quadriplegic has learned to use cortical recordings to command a simulated arm with muscles modeled in a realistic way (Chadwick et al., 2011).

Several challenges remain for developing clinically useful devices. It remains to be demonstrated which of the several competing approaches to the design and implementation of neural prosthetics will prove to be the most beneficial to patients. Furthermore, recording and decoding the intentions of patients depends on stable recordings over a long period. Perhaps most importantly, existing neural prostheses usually use vision, alone, to guide movement, and lack the somatosensory feedback of normal movements. Much of the early work in this field involved the control of cursors or levers along a single dimension. Obviously, a higher level of control is desirable. Like the patients mentioned above, monkeys have been trained to use motor cortex activity to control a simulated arm that has three dimensions of movement, as well as to control a gripper at the end of the arm for feeding movements (Velliste et al., 2008).

Sociological Implications

Research on the motor cortex has at least two sociological implications. First, like the antivivisectionists mentioned in the section on the history of motor cortex studies, contemporary animal rights advocates claim that research on monkeys and other nonhuman animals has no relevance to humans. Contrary to their opinions, the development of devices that enable paralyzed and mute people some control and independence has depended crucially on monkey research. Without monkey research motivated by the muscles versus movement debate and a desire to understand the structure, function, and evolutionary history of the primary motor cortex, there would be no neural prosthetics.

Second, both neural prostheses and improved rehabilitation methods promise to have important effects on society. More than one-third of stroke patients live for over a decade after sustaining brain damage, as do patients with brainstem and spinal injuries. Their contribution to society is much less than it could be. The ability to restore to them a measure of motor control could have effects like those that followed the development of corrective optics in the thirteenth century. In medieval Europe, the development of eyeglasses extended the productivity of artisans by freeing them from the farsightedness that often develops in middle age. Extending the productivity of people with motor disabilities could, likewise, benefit both the affected individuals and society as a whole.

See also: Cerebellum: Cognitive Functions; Cerebral Cortex; Cortical Areas Engaged in Movement: Neuroimaging Methods; Neural Plasticity; Premotor Cortex; Topographic Maps.

Bibliography

Barbay, S., Plautz, E.J., Friel, K.M., et al., 2006. Behavioral and neurophysiological effects of delayed training following a small ischemic infarct in primary motor cortex of squirrel monkeys. Experimental Brain Research 169, 106–116.

Chadwick, E.K., Blana, D., Simeral, J.D., et al., 2011. Continuous neuronal ensemble control of simulated arm reaching by a human with tetraplegia. Journal of Neural Engineering 8, 034003.

Ethier, C., Oby, E.R., Bauman, M.J., Miller, L.E., 2012. Restoration of grasp following paralysis through brain-controlled stimulation of muscles. Nature 485, 368–371.

Graziano, M.S.A., 2005. Arm movements evoked by electrical stimulation in the motor cortex of monkeys. Journal of Neurophysiology 94, 4209–4223.

Hochberg, L.R., Bacher, D., Jarosiewicz, B., et al., 2012. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. Nature 485, 372–375.

Hoffmann, J.N., Montag, A.G., Dominy, N.J., 2004. Meissner corpuscles and somatosensory acuity: the prehensile appendages of primates and elephants. Anatomical Record A 281, 1138–1147.

Kaas, J.H., 2011. Reconstructing the areal organization of the neocortex of the first mammals. Brain Behavior and Evolution 78, 7–21.

Kaas, J.H., Gharbawie, O.A., Stepniewska, I., 2011. The organization and evolution of dorsal stream multisensory motor pathways in primates. Frontiers in Neuroanatomy 5, 34.

Kennedy, P., Andreasen, D., Bartels, J., et al., 2011. Making the lifetime connection between brain and machine for restoring and enhancing function. Progress in Brain Research 194, 1–25.

Krakauer, J.W., Carmichael, S.T., Corbett, D., Wittenberg, G.F., 2012. Getting neurorehabilitation right: what can be learned from animal models? Neurorehabilitation and Neural Repair. March 2012, epub ahead of print.

Mandelblat-Cerf, Y., Novick, I., Vaadia, E., 2011. Expressions of multiple neuronal dynamics during sensorimotor learning in the motor cortex of behaving monkeys. PLoS One 6, e21626.

Moritz, C.T., Perlmutter, S.I., Fetz, E.E., 2008. Direct control of paralysed muscles by cortical neurons. Nature 456, 639–642.

Pruszynski, J.A., Kurtzer, I., Nashed, J.Y., et al., 2011. Primary motor cortex underlies multi-joint integration for fast feedback control. Nature 478, 387–390.

Rathelot, J.A., Strick, P.L., 2006. Muscle representation in the macaque motor cortex: an anatomical perspective. Proceedings of the National Academy of Sciences of the United States of America 103, 8257–8262.

Seki, K., Fetz, E.E., 2012. Gating of sensory input at spinal and cortical levels during preparation and execution of voluntary movement. Journal of Neuroscience 32, 890–902.

Soteropoulos, D.S., Edgley, S.A., Baker, S.N., 2011. Lack of evidence for direct corticospinal contributions to control of the ipsilateral forelimb in monkey. Journal of Neuroscience 31, 11208–11219.

Velliste, M., Perel, S., Spalding, M.C., et al., 2008. Cortical control of a prosthetic arm for self-feeding. Nature 453, 1098–1101.

Motor Skills: Development in Infancy and Early Childhood

Cibelle KMR Formiga, University of Goiás State, Goiás, Brazil
Maria BM Linhares, School of Medicine at Ribeirão Preto, University of São Paulo, São Paulo, Brazil

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes motor development in infancy and early childhood, especially highlighting the major motor skills in the first year of a child's life. The continuing development of motor skills in children means the acquisition of independence and the ability to adapt to the physical and social environment. Motor skills in posture horizontal, vertical, and locomotor skills allow the child greater body control and improvement of social skills and interaction. The motor behavior is the basis for the development of other skills throughout childhood.

Introduction

Child development is a complex interaction process of biological aspects with various environmental influences and experiences. Theories have been developed to explain how changes occur in the body of infants and in their ability to interact with the environment.

The continuing development of motor skills in children means the acquisition of independence and the ability to adapt to the physical and social environment. Motor skills and cognitive processes influence each other and are manifested mostly through behavioral motor modalities (Flehmig, 1992).

Theories of Motor Development

Regarding the evolution of knowledge about the study of child development, one can highlight the emergence of some theories that attempt to explain how the behavioral learning of the infant occurs from the prenatal period until the end of the sixth year of postnatal life. In this respect, the most studied theoretical frameworks are the neuromaturational theory (NT) and dynamical systems theory (DST).

The NT of motor development is the traditional model and remains the most frequently reported theory in textbooks about motor development (Gesell and Amatruda, 1945; McGraw, 1945). The central tenet proposes that changes in gross motor skills during infancy result solely from the neurological maturation of the central nervous system (CNS).

Advancements in the science of embryology led to the discovery that the embryo developed in a symmetrical manner, beginning from cephalocaudal and proximal to distal directions (Gesell and Amatruda, 1945; McGraw, 1945). From these observations, four assumptions have been formulated, which characterize the neuromaturational model: primitive movements for controlled movements, reflex activities for voluntary activities; motor development progresses in a cephalocaudal direction; movement is first controlled proximally and then distally; the sequence of motor development is consistent among infants; and the rate of motor development is consistent for each infant.

The DST emerged in the early 1980s as a new theoretical explanation for the changes that occur over time in motor behavior and motor skills of children (Thelen et al., 1987). The theoretical framework includes all areas of development, being derived from psychological theories, physics, chemistry, and mathematics. Researchers have postulated that when a new behavior is developing, it is dependent on an input of all contributions of the systems. This behavior may have characteristics that could not have been determined by the evolution of individual behaviors.

This observation was transferred to human movement by Bernstein (1967). He noted that the joints and muscles never work in isolation but with a coordinated synergy. It was postulated that the brain controls muscle groups better than individual units and that the muscle synergies themselves are able to autonomously modify an independent movement.

This theoretical approach also recognizes the maturational level of the CNS as an important component for success of the task, but it is not the only factor. Other variables influencing the final motor behavior include the emotional state of the infant, the degree of motivation, cognitive awareness, the infant's posture, muscle strength, and biomechanical leverages. The shape, size, and weight of the toy also determine how the motor skill is executed. In contrast to the neuromaturational model that recognizes only the influence of the cerebral cortex, the dynamic motor theory approach takes into consideration all of the factors impinging on the motor outcome (Thelen, 1989; Thelen et al., 1990).

In summary, the systems approach to motor development represents a holistic and advanced theoretical model. The infant, the environment, and the functional significance of the task cannot be isolated from each other because they represent a synthesized unit and the motor behavior observed as an output is a product of their interactions. The system is capable of autonomously modifying the motor skill, depending on the constraints imposed on the system and the level of functioning of each unit in the system. Elements composing the system can mature in different ways, and any single factor can act as rate limiting, delaying the emergence of a new motor skill (Thelen, 1995).

The theoretical framework presented here provides professionals who work with infants an opportunity to reassess the

traditional paradigm used to describe and understand the motor development.

Motor Skills in Infancy

The motor development of infants is divided into quarters or months, defining the motor behavior of the infants at each stage. However, these steps are not fixed and depend on the interaction with the infant's environment and experiences. The time in which the infant is able to perform various motor acts depends to some extent on the opportunities to rehearse them, varying according to the environment and the way the child was stimulated (Shepherd, 1995).

Each infant demonstrates his or her characteristic development pattern once inherent characteristics suffer constant influence of the interplay between child and environmental context. There is also considerable individual variation among children of different ages, as well as within the same age group. Yet, there are particular characteristics that allow an assessment of the level and quality of the performance (Gallahue et al., 2011).

The acquisition of motor skills is among the most remarkable achievements in the first years of life. Motor milestones such as the emergence of sitting without support or the first independent steps provide a framework for developmental monitoring of children in health supervision visits because these milestones belong to the most salient and best demarcated markers of developmental processes that parents and health care professionals can observe. In clinical practice, general pediatricians, child neurologists, and developmental professionals are often asked to predict future outcomes on the basis of early developmental milestones (Jenni et al., 2013).

Currently, motor development has been studied more as a process than as a product (Tani, 2005). In this sense, the idea of motor milestones does not offer a poor measure to know in detail the complexity of motor skills in the first years of a child's life (Adolph and Robinson, 2013). There is a recent trend in enhancing the quality and control of movements by the child than just whether or not this carries certain motor behavior. Moreover, one cannot fail to consider the cross-cultural influence on motor development. The rate of motor development may differ between children of different cultures, such as Chinese, African, or American. However, the product development appears to be similar and compatible among them (Karasik et al., 2010).

Despite the existence of motor milestones, when studying motor development of the child, it is necessary to verify the cultural context and not just follow Western standards of reference (Harkness et al., 2011). The professional or researcher should also give attention to the type of measuring instrument that is used for the evaluation of the child or parents or both. Each culture has its own organization, language, and habits of life. Therefore, not all instruments are easily applicable in all countries (Gladstone et al., 2009).

Considering the environmental characteristics in development, nutrition is also an important factor for healthy growth and development in childhood. The study by Angulo-Barroso et al. (2011) investigated the gross and fine motor development at 9 months of children from the urban area of China, Ghana, and the United States (African-American) who were iron deficient. The study revealed that African children had better performance in gross motor skills, such as standing with support and walking with support, and fine motor skills. These results are also supported by previous studies in which infants in Africa may have an advantage due to the early stimulation with balance and postural control that is typical of many cultures in Africa (Bril, 1986; Super, 1976; Werner, 1972; WHO Multicentre Growth Reference Study Group, 2006).

Regarding the screening for developmental disabilities, Bornstein and Hendricks (2013) conducted a survey of 172 000 families in 16 developing countries and concluded that developmental disabilities vary by child age and country, and younger children in developing countries with lower standards of living are more likely to screen positive for disabilities.

The development of children in the first year of life has been strongly marked by the explosion of gross motor skills. A device used in evaluating this age range has been the Alberta Infant Motor Scale (AIMS). This scale has been used in children in many cultures and has presented different results in their interpretation of early motor development. To Japanese children, the AIMS percentile ranks of motor development showed results below the expected range (Uesugi et al., 2009). In a study conducted in Brazil with 795 children between 0 and 18 months, the results were almost similar to the Japanese study, in which children had lower mean motor skills according to the standardization sample of the scale for most months evaluated in the first year of life (Saccani and Valentini, 2012). Fleuren et al. (2007) assessed 100 children and concluded that new percentiles should be defined for Netherlands because the motor performance scores they observed were below the Canadian standard. In contrast, Syrengelas et al. (2010) conducted a study with 424 full-term Greek infants and found that their development path was similar to that of the Canadian children, demonstrating that the AIMS reference values could be used without loss of important clinical information.

The study of Tripathi et al. (2008) compared the normal motor development scores of 300 children from Mangalore, India, between birth and 60 months of age, on the Peabody Developmental Motor Scales-2 (PDMS-2) with the normative scores provided with the instrument. The authors founded that the Indian children's scores varied with some differing from the normative sample, whereas others did not differ across age groups and the different subtests of the PDMS-2. The study concluded that it is not possible to develop assessment tools that are culturally sensitive across different geographical regions and environments, but it is necessary to evaluate the cultural sensitivity of such tests for use in a particular region and ethnic group, especially when these assessment tools are being used to diagnose and plan treatment for a child.

To study the development of children above 2 years of age, ethnic or cultural differences seem to be more related not only to motor development, exclusively, but also to other areas such as language, social and cognitive skills, product of the interaction between organism and social environment. In addition, each assessment tool for motor development has its

psychometric properties that need to be taken into account in the analysis of the infant motor development.

In the following, the motor development will be presented according to the process of skill acquisition in accordance with the body posture, rather than motor milestones at ages-keys.

Development in Prone and the Neck Control

In prone position, since the first days after birth the infant begins to lift the head, an act of protective head tilt that is designed to keep the mouth and the nose free. However, the infant demonstrates an improved ability to lift the head when held against the caregiver's shoulder than when lying on the mattress. This demonstrates that even early on, the movement becomes easier under ideal mechanical conditions.

The neonate who spends some time exercising in prone develops extensor muscle strength, and the ability to extend the head and trunk develops rapidly. Within a few weeks, the infant is able to activate the extensors of the neck and the upper portion with enough strength to lift the head and look around. Initially, the baby can lift the head to 45° at 2 months of age, reaching 90° at 3 months. At this age, the maintenance of the head in midline is possible by bilateral contraction of the paraspinal muscles (Bly, 1994).

By 3–4 months of age, the baby can lift the head and chest in prone and push himself up. As head control increases in the prone position, visual attention enhances and the infant can follow an object horizontally to 180°. At 5 months, the baby can keep the head upright and can roll from prone to supine position.

At 6 months of age, the baby can push up on the wrists in prone and begin to perform a new activity in this posture: the pivoting. Pivoting is when the baby moves in the frontal plan with spine lateral flexion, often influenced by visual interest. In this activity, the baby weight shifts on extended arms.

At 7 months, the baby can achieve extended-arm weight bearing and upper extremity weight lifting and reach for toys. There is good pelvic weight shifting and lower extremity dissociation. Increased trunk and pelvic-femoral controls enable the baby to assume quadruped and rock in this position. From 8 months of age on, the baby will experience changes in posture, and will rarely be static when in prone position (Bly, 1994).

Development in Supine and to Roll

At birth, the posture of the arms and legs is predominantly a flexion pattern and such a pattern predominates during the first weeks of life. In a few months, the pattern turns into semiflexion, extension, and finally a posture without a predominant pattern (Flehmig, 1992).

At 2 months of age, functional activities are still somewhat limited, consisting of active head turning and semicontrolled extremity movements. Visual attentiveness and visual reaching are two of the baby's most functional activities. At 2 months, the head is rarely in midline, which may be due to the increased cervical spine mobility. This rotation is accompanied by an increased head extension and chin lifting.

At 3 months of age, the baby demonstrates symmetry and midline head orientation. The upper extremities are often characterized by bilateral flexion, bilateral abduction, and external rotation. There is more alignment of the ribs because there is activation of the abdominal muscles. When the knees approach the chest from triple flexion, there is an increase in abdominal contraction that promotes the lowering of the ribs.

At 4 months, the baby can easily alternate between extension and flexion. Head and trunk symmetry, midline orientation, and bilateral symmetrical extremity movements are dominant and enable the development of coordination between the two sides of the body. At this age, the ocular control is becoming more refined as a result of increased head control and vice versa. The eyes are more active during reaching for objects, though the upper limbs still lack the coordination and control needed for reaching.

Visual 'fixing' increases head stability and ensures its proper orientation in space. The alternating symmetrical movements of the lower extremities and the alternating activation of trunk extensors and flexors facilitate anterior and posterior tilting of the pelvis. These pelvic movements will provide a basis for further normal development of lower extremity movements. The baby is also able to extend the elbows and reach hands to knees when the hips and knees are flexed (Bly, 1994).

At 5 months, the baby can actively roll from supine to side lying. The action is initiated with total symmetrical total flexion, similar to that of the fourth month, but when the baby reaches side lying, the symmetry changes to asymmetry. The lower leg extends while the top leg remains flexed, and the baby momentarily laterally rights the head (laterally flexing against gravity). The baby can bring the feet to the mouth and the hands to the feet, using the control of upper and lower extremities, which helps in the development of body awareness and tactile stimulation.

Increased control for lateral flexion allows the 6-month-old to shift weight in the trunk and pelvis and assume dissociated lower extremity position. This enables the baby to roll toward side lying and maintain the position there. Side lying is a more functional position for upper extremity use. The baby is also beginning to use the lateral and dissociated movements to lift into quadruped from prone. These dissociated lower extremity movements mobilize the lumbar and thoracic spine.

Mobility throughout the spine, pelvis, and knee joint is necessary to achieve reciprocal extremity movements of the extremities during locomotion. Up to 6 months, the baby can pull to sitting when holding the examiner's hand. The baby has sufficient antigravity control and synergistic flex control to flex and lift the head, arms, and legs independently. At the end of the sixth month, rolling from supine to prone is initiated by flexion, rotation, and lateral weight shift. These components occur most frequently in the head and the lower extremities (Piper and Darrah, 1994).

As mentioned in the earlier part of this article, motor development is not exactly the same for all children in all cultures. In terms of motor ability to roll, the study by Nelson et al. (2004) found that Hong Kong Chinese infants roll from supine to prone before they roll from prone to supine. Mean ages of rolling over were 5.1 months for supine to prone and

5.7 months for prone to supine. Age of rolling over from supine to prone was not influenced by usual sleep position, infant's sex, mother's intention to breast-feed infant, number of siblings, marital status, main daytime caregiver, or feeding method over 9 months.

Development of the Baby to Sit

Head support in the sitting position occurs at around 3 months of age. When pulled to sit, the baby uses the optical or cervical righting by fixing the eyes on the examiner as if to reinforce head stability. The upper extremities resist passive extension, being inconsistently active in assisting by pulling into elbow flexion. When elbow flexion occurs, it is usually observed during the first half of the movement. When the head becomes stable and flexes forward, active elbow flexion decreases. During pull to sit, asymmetrical reactions in the extremities are common.

When head and trunk control have improved, head movements do not disturb the balance. The baby will use bilateral scapular adduction to reinforce trunk stability. This response is needed for postural stability, and stability precedes hand use. The baby can visually track objects, but cannot catch it, and will fall forward if left unsupported. There is a minimal resistance in the hips, lumbar, and lower thoracic areas for forward bending. However, the head and upper trunk do respond and resist (Bly, 1994). When held by the forearms and pulled to sit, the baby has the optical righting reaction and tries to reinforce the head-righting ability by visually 'fixing' on the examiner. At the end of the movement, the baby stabilizes the sitting posture after passing through the erect vertical position (Gallahue et al., 2011).

At 5 months, the baby displays increasing balance between the flexors and extensors. Erect sitting at 5 months is possible only when the baby's hands are held or the trunk is supported. The baby can maintain trunk extension, the arms flexed forward, and the scapulae abducted.

The 6-month-old has sufficient trunk and hip control to sit erect without support and uses a ring position of the lower extremities for stability. The upper extremities are freed from the postural system and can be used for reaching, manipulation, or forward protective extension. When sitting, the baby has control of the head and trunk movements on the sagittal plane (flexion and extension). On the frontal and transverse planes, the baby has head control, but no trunk control. When the baby rotates the head, weight is shifted to the same side to which the head turns. This frequently causes the baby to fall to the side.

At 7 months, the baby can sit independently with the back and pelvis straight. Because of increased hip and trunk control, unsupported sitting is becoming a more functional position in which the baby can hold and manipulate toys. However, the baby does not yet have full sitting balance. Some babies at this age pull themselves to stand, accomplishing this by transitioning from quadruped to kneeling while leaning on furniture (Shepherd, 1995).

At 8 months, the baby may use the positional stability in sitting. The femoral-pelvic muscles and trunk muscles are sufficient to stabilize the posture. Consequently, the baby experiments with and uses a variety of lower extremity positions when sitting, such as ring sitting, side sitting, long sitting, and sitting in W position (Bly, 1994; Flehmig, 1992).

At 9 months, sitting is often a transitional state for babies as they continue to explore the environment actively. The baby can also transition to quadruped by rotating the trunk and pelvis over the femur. This transition requires marked pelvic-femoral mobility.

Static sitting is rare for 10-month-olds. When babies are quiet in sitting, they are usually eating or exploring a toy. They spend most of the time moving in and out of sitting, retrieving toys, and transporting them to a new location. Wide abduction of the legs is also possible during long sitting. This posture provides additional positional stability, but it requires marked mobility in the hip adductor muscles (Bly, 1994).

Between 11 and 12 months, the baby may be observed sitting quietly when engaged in dressing, eating, or a fine motor task. Trunk rotation occurs preferably through a greater range than in previous months, and does not interfere with sitting stability. Trunk rotation enables the baby to reach for objects at the side and behind.

Development of Manual Skills

The neonate is able to fixate and track an object briefly as well as track a face from side to side if it is close enough. According to Shepherd (1995), it is likely that the coordinated control of the eye and the hand, i.e., the ability to grasp and to observe the hand is the beginning of the hand functional use.

The audio-visual-cephalic coordinations are present in full-term neonates when they are around 2 months of age. Due to the palmar grip reflex, all neonates display closing of the hands. However, when they are around 2 months of age, grasping movements become more relaxed and more finger extension is observed. The child begins to acquire intersegmentar hand–mouth coordination from the first month of life, becoming more evident in the second month.

The first way to grasp (hand–object coordination) is with one hand, with the three ulnar fingers (fifth, ring, and index fingers) flexed against the palm, where the infant only grabs the object when it touches the hand.

At 3 months, the baby can bring hands together in midline (hand–hand coordination) and flex and reopen them. The grasp reflex is gradually inhibited and at around the fifth month the child can consciously drop the object. The coordination hand-handkerchief on the face, in which a diaper may be put on the baby's face and is withdrawn by him, begins at approximately 4 months of age and the coordination hand–foot is present at 6 months of age.

At 7 months, the infant can also achieve a grip hook pattern (hand grip without thumb opposition). The thumb and index finger participate in a radial-palmar grip and both hands can be used simultaneously. In the next phase of development, the baby will be able to pick up small objects using all fingers, as it is no longer necessary to compress the object against the palm.

Later, the child begins to use a pincer grasp formed by the thumb and index finger to pick up a small object. At 9 months, the child also displays a complete handgrip and does not return an object in order to explore it visually and tactically. At 12 months, the baby will return the object if requested, and up to this age most objects given to the child will be placed in the

mouth. From 9 months on, some babies eat independently with a spoon, but most do it at around 15 months of age (Bly, 1994).

Development of Quadruped Position and Crawling

The quadruped position begins to be experienced by the child some time before crawling, i.e., at around 6–7 months of age. According to Shepherd (1995), not every baby crawls, but for those who do, this method provides the primary means of locomotion to be rehearsed. Crawling requires the child to be capable of making adjustments in kneeling position supported on four points. At 7 months, the baby likes new discoveries and toys that move, as well as small objects and household utensils and initiates transitions from sitting to quadruped and also from prone to quadruped (Bly, 1994).

Once in quadruped, the baby can rock forward, backward, and sideways. Rocking requires sufficient back stability to permit scapular freedom and mobility. Rocking at first utilizes large movements with falling, then it changes to smaller movements without falling. Additionally, rocking provides vestibular, proprioceptive, and kinesthetic stimulation and strengthens the shoulder and hip muscles (Piper and Darrah, 1994).

Diament's studies (1976) indicate that crawling begins at 8 months of age. However, some babies begin to crawl before this age. At 8 months, the baby can move easily from sitting to quadruped, starting this transition flexing the leg and foot under the body. The other leg remains flexed, abducted, and externally rotated, providing stability to the pelvis. The trunk remains symmetrically extended, and movement occurs in the sagittal plane. Marked mobility of the hip joint is necessary for this transition. The ability to crawl is a very efficient way for the baby to move from one place to another. They use the reciprocal extremity movements, which require a diagonal and counterrotation control in the trunk. Crawling and its weight-lifting components provide varied input into the hands, which may contribute to the development of the palmar arches.

Between 7 and 8 months, the baby can roll over, creep, and crawl in a matter of seconds. By the ninth month, the baby has become quite proficient at crawling. Reciprocal extremity movements and trunk rotation continue to be used. Speed and control during crawling are quite refined. The baby can move at varied speeds and can quickly change directions. Crawling is the primary means of locomotion, and the baby uses this skill to explore the environment and to obtain and transport toys (Gallahue et al., 2011).

At 10 months, crawling and climbing are the main activities of the baby. These movements demonstrate and develop the baby's coordination between the trunk and the extremities. The baby now has greater motor and body awareness to maneuver the body over, around, or onto obstacles that may obstruct a toy. If babies encounter obstacles while crawling, they can continue their forward progression by climbing up onto or over the obstacle. Infants encounter obstacles of different weights, sizes, and stability and will thus be challenged to develop problem-solving repertoire through experimental behaviors.

Kneeling without external support is a common occurrence for the 10-month-old. Contraction of the quadriceps is needed to elevate the body. Hip extensors are needed to stabilize the trunk. The 10-month-old infant can easily transition into and play in half-kneeling. Assumption of this posture demonstrates the baby's increased control of lower extremity dissociation (Bly, 1994).

Development of the Baby to Stand and to Walk

Some authors report the onset of orthostatic position at 7 months of age (Diament, 1976). Thus, the child stands independently and is supported at 9 months, and stands unsupported at 11 months. At 10 months, the child is able to stand by holding on to furniture, assumes a wide base of support of the feet, and can go hand in hand with an adult, evolving to one-hand support. With the continuing environment exploration guided by the ability of moving around in space, primarily through creeping and crawling, motor skills to stand and walk begin. The ability of rising to stand requires good muscle activation of the legs and begin at 7 months, but with the help of the upper extremities.

Subsequently, at 8 months, hip control on weight bearing must be sufficient to initiate the weight shift and stabilize the pelvis as the other leg is freed to move. The pelvic-femoral muscles must dynamically stabilize the pelvis and femur in the vertical position, rotate the pelvis over the femur, and maintain external rotation of the weight-bearing leg. If the hip muscles maintain the hip in extension and external rotation, the weight will be transferred to the medial side of the foot. These lower extremity actions resemble and may be precursors to the support and balance phases of gait.

At 10 months, the upper extremities seem to be used more for balance when rising to stand. From half-kneeling, the baby shifts forward on the flexed leg, which demonstrates good concentric control in the quadriceps. Once in standing, the baby uses leg muscles and minimum attendance of one hand to control posture. The arms and hands are free for exploration and manipulation. This eventually enables the baby to stand without upper extremity support. The baby spontaneously relinquishes hand support when presented with a toy that requires two hands. For bilateral hand use to be successful, the baby must preadjust posture for stability. Increased ankle movements are noted in standing, especially active plantar flexion (Flehmig, 1992).

At 10 months, the ability to control their posture and lower extremities in standing enables them to continue to vary the cruising pattern. They may cruise sideways with lower extremity abduction as they did in previous months, or they may turn so that they can face the direction in which they are going. This causes the baby walk forward holding with one hand instead of walking to the side holding it with both hands. Motor planning skills are being developed and practiced in standing as well as in other activities.

At 11 months, standing without external support is a new accomplishment for the baby. It usually occurs automatically when babies become very interested in a toy and they want to hold or explore it with both hands. Wide abduction of the legs assures a wide, stable base of support. These postural adjustments occur automatically before the baby lets go of the hands. With practice, the baby's independence in walking increases quickly.

The development during infancy of independent upright walking is a long and arduous process that requires months

of experience to reach a full flexible and adaptable movement. From a biomechanical point of view, also the gait pattern can be changed depending on the child's experience of walking (Bonneuil and Bril, 2012).

During initial attempts at independent walking, the movement components tend to regress to those used in early supported walking. The baby assumes a wide base of support with the feet, abducts the arms, and flexes the elbows. The upper trunk fixing pattern that the baby uses depends on the goal. The most common pattern of shoulder elevation, scapular adduction, and elbow flexion is used when the baby's goals are more nonspecific, when the baby intends to grab a toy, for instance. Because balance is poor in early walking, the baby moves quickly and usually falls or is caught in the arms of a parent. This does not discourage the baby or the parents, and walking workouts are practiced until the baby is independent (Bly, 1994).

At 12 months, the baby can lower the body with or without external support. To do this, the baby shifts the weight posteriorly, as if sitting down. The baby flexes the hips and knees but not the ankles. The quadriceps, hip extensors, and abdominals are also active in maintaining control. If the weight is shifted posteriorly without the knees being flexed, the baby will fall to sitting.

Standing presents new postural challenges to the baby because the base of support is different. When the baby is sitting, the hips are the base of support. When the baby is standing, the baby's feet are the base of support. Initial postural stability can be achieved through toe curling.

Most babies walk independently by or during the 12th month. Early independent walking usually has the following characteristics: fast speed, short stride length, short step length, high cadence, short swing phase, wide base of support, and no reciprocal arm swing (Flehmig, 1992; Shepherd, 1995).

At around 13–14 months of age, children stand independently. A 15-month-old child can walk and move up the stairs while held by the hands. Toe-off impulse at gait while standing is achieved at about 16 months, and the lifting of the toes while in bipedalism is observed at 17 months of age.

At 18 months of age, the child can climb on a chair and sit, and can also run. These dates are not fixed and maybe some children do not go through the same developmental process of the groups studied by the authors. Therefore, the occurrence of unsupported walking before 12 months of age and shortly thereafter (around 14 months) can be considered within the normal range, taking into consideration that the child was a full-term newborn and has 'no signs' of neurological impairment.

The emergence of the gait is still motivating great discussion among the authors. The act of walking is related not only to the maturation of the nervous system, but also to sensory experiences, and cultural factors related to genetics (Adolph and Robinson, 2013). Every human being has a unique temperament and ability to perform the movements of locomotion and adaptation to the environment can also be related to the psychological characteristics of the developing child.

In addition, the performance of previous motor milestones can be related to the acquisition of independent walking. Study by Kimura-Ohba et al. (2011) examined the relationship between motor development milestones with the age of walking in 290 healthy and term infants born in the district of Osaka City, Japan. Three milestones (rolling over, sitting, and crawling) were observed in the laboratory in infants aged 4 and 9 months by a pediatrician and a developmental psychologist, and the age of walking was confirmed in questionnaires filled in by the parents at 18 and 27 months. The authors found those children who could roll over at 4 months, and sit and crawl at 9 months, walked earlier than children who could not roll over, sit, and crawl. With regard to crawling, children who were creeping had a 1-month delay in walking, and those who could not move forward had a 2-month delay compared to typical crawlers. On multiple regression analysis, these three milestones were positively associated with walking. The study concludes that the age and the patterns of sitting, crawling, and rolling over were all related to the age of independent walking among Japanese infants. Consideration of milestone definition and variations is essential in medical check-up.

Development of Motor Skills after Gait

After walking independently, the child progresses to other skills such as running, jumping, throwing, and receiving, in a successive progression from easier skills to the most difficult ones until being able to combine them all.

Walking has often been defined as the process of continuously losing and recovering balance in erect position. Once independent walking has been achieved, the child progresses rapidly to the elementary and mature stage of walking. The gait can be considered mature in a certain moment in the motor development of children between 4 and 7 years of age (Gallahue et al., 2011).

See also: Child Care and Development across Cultures; Child-Directed Speech: Influence on Language Development; Cognitive Development During Infancy and Early Childhood across Cultures; Cross-Cultural Research Methods in Psychology; Infancy and Human Development; Longitudinal Analyses of Sexual Development through Early Adulthood; Pretend Play and Cognitive Development; Self in Culture: Early Development; Self-Development in Childhood and Adolescence; Social and Emotional Development in the Context of the Family.

Bibliography

Adolph, K.E., Robinson, S.R., 2013. The road to walking: what learning to walk tells us about development. In: Zelazo, P. (Ed.), Oxford Handbook of Developmental Psychology, vol. 1. Oxford University Press, New York, pp. 403–443.

Angulo-Barroso, R.M., Schapiro, L., Liang, W., Rodrigues, O., Shafir, T., Kaciroti, N., Jacobson, S.W., Lozoff, B., 2011. Motor development in 9-month-old infants in relation to cultural differences and iron status. Developmental Psychobiology 53 (2), 196–210.

Bernstein, J.A., 1967. Co-ordination and Regulation of Movements. Pergamon Press, New York.

Bly, L., 1994. Motor Skills Acquisition in the First Year: An Illustrated Guide to Normal Development. Therapy Skill Builders, Tucson, AZ.

Bonneuil, N., Bril, B., 2012. The dynamics of walking acquisition: a tutorial. Infant Behavior & Development 35, 380–392.

Bornstein, M.H., Hendricks, C., 2013. Screening for developmental disabilities in developing countries. Social Science & Medicine 97, 307–315.

Bril, B., 1986. Motor development and cultural attitudes. Themes in Motor Development 35, 297–313.
Diament, A.J., 1976. Evolução Neurológica Do Lactente Normal. Edart, São Paulo.
Eckert, H.M., 1987. Motor Development, third ed. Brown & Benchmark Publisher, USA.
Flehmig, I., 1992. Normal Infant Development and Borderline Deviations: Early Diagnosis and Therapy, fourth ed. G. Thieme Verlag, New York.
Fleuren, K.M., Smit, L.S., Stijnen, T., Hartman, A., 2007. New reference values for the Alberta infant motor scale need to be established. Acta Paediatrica 96, 424–427.
Gallahue, D., Ozmun, J., Goodway, J., 2011. Understanding Motor Development: Infants, Children, Adolescents, Adults, seventh ed. Hardcover.
Gesell, A., Amatruda, C.S., 1945. The Embryology of Behavior: The Beginnings of the Human Mind. Harper & Brothers, New York.
Gladstone, M., Lancaster, G., Umar, E., Nyirenda, M., Kayira, E., van den Broek, N., Smyth, R.L., 2009. Perspectives of normal child development in rural Malawi – a qualitative analysis to create a more culturally appropriate developmental assessment tool. Child: Care, Health and Development 36 (3), 346–353.
Harkness, S., Zylicz, P.O., Super, C.M., Welles-Nyström, B., Bermúdez, M.R., Bonichini, S., Moscardino, U., Mavridis, C.J., 2011. Children's activities and their meanings for parents: a mixed-methods study in six western cultures. Journal of Family Psychology 25 (6), 799–813.
Jenni, O.G., Chaouch, A., Caflisch, J., Rousson, V., 2013. Infant motor milestones: poor predictive value for outcome of healthy children. Acta Paediatrica 102 (4), 181–184.
Kimura-Ohba, S., Sawada, A., Shiotani, Y., Matsuzawa, S., Awaya, T., Ikeda, H., Okada, M., Tomiwa, K., 2011. Variations in early gross motor milestones and in the age of walking in Japanese children. Pediatrics International 53 (6), 950–955.
Karasik, L.B., Adolph, K.E., Tamis-LeMonda, C.S., Bornstein, M.H., 2010. WEIRD walking: cross-cultural differences in motor development. Behavior and Brain Sciences 33 (2–3), 95–96.
McGraw, M.B., 1945. The Neuromuscular Maturation of the Human Infant. Appleton-Crofts, New York.
Nelson, E.A., Yu, L.M., Wong, D., Wong, H.Y., Yim, L., 2004. Rolling over in infants: age, ethnicity, and cultural differences. Developmental Medicine & Child Neurology 46 (10), 706–709.
Piper, M.C., Darrah, J., 1994. Motor Assessment of the Developing Infant. W. B. Saunders Company, EUA.
Saccani, Raquel, Valentini, Nadia C., 2012. Reference curves for the Brazilian Alberta infant motor scale: percentiles for clinical description and follow-up over time. Jornal de Pediatria 88 (1), 40–47.
Shepherd, R.B., 1995. Physiotherapy in Pediatrics, third ed. Butterworth-Heinemann.
Super, C.M., 1976. Environmental effects on motor development: the case of Africa precocity. Developmental Medicine and Child Neurology 18, 561–567.
Syrengelas, D., Siahanidou, T., Kourlaba, G., Kleisiouni, P., Bakoula, C., Chrousos, G.P., 2010. Standardization of the Alberta infant motor scale in full term Greek infants: preliminary results. Early Human Development 86, 245–249.
Tani, G., 2005. Comportamento Motor: aprendizagem e desenvolvimento. Guanabara Koogan, Rio de Janeiro.
Thelen, E., Ulrich, B.D., Jensen, J.L., 1990. The developmental origins of locomotion. In: Woollacott, M.H., Shumway-Cook, A. (Eds.), Development of Posture and Gait Across the Life Span, second ed. University of South Carolina Press, pp. 26–47.
Thelen, E., 1989. The (re) discovery of motor development: learning new things from an old field. Developmental Psychology 25, 946–949.
Thelen, E., 1995. Motor development: a new synthesis. American Psychologist 50 (2), 79–95.
Thelen, E., Kelso, J.A.S., Flogel, J.A.S., 1987. Self-organizing systems and infant motor development. Developmental Review 7, 39–65.
Tripathi, R., Joshua, A.M., Kotian, M.S., Tedla, J.S., 2008. Normal motor development of Indian children on Peabody Developmental Motor Scales-2 (PDMS-2). Pediatric Physical Therapy 20 (2), 167–172.
Uesugi, M., Tushima, E., Shimada, T., 2009. Percentile ranks of the Alberta infant motor scale for Japanese infant. Rigakuryoho Kagaku 24 (1), 15–19.
Werner, E.E., 1972. Infants around the worlds: crosscultural studies of psychomotor development from birth to two years. Journal of Cross-cultural Psychology 2 (3), 111–134.
WHO Multicentre Growth Reference Study Group, 2006. Assessment of sex differences and heterogeneity in motor milestone attainment among populations in the WHO Multicentre Growth Reference Study. Acta Paediatrica Supplement 450, 66–75.

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 16

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser–Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 16

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 16 *xliii*

VOLUME 16

Movement Theory and Constraints in Syntax 1
C-T James Huang and Hiroki Narita

Movements, Social 8
David A Snow and Anna E Tan

Muller, Hermann Joseph (1890–1967) 13
Elof A Carlson

Multicultural Education 18
James A Banks

Multiculturalism 22
Duncan Ivison

Multiculturalism, Anthropology of 28
Thomas Hylland Eriksen

Multidimensional Scaling I 34
Kwanghee Jung and Yoshio Takane

Multidimensional Scaling II 40
Patrick JF Groenen and Ingwer Borg

Multilevel Models in Demography 48
Germán Rodríguez

Multilingual Language Development 57
Alejandro E Brice

Multilingualism 65
Ulrike Jessner-Schmid

Multimodal Therapy in Clinical Psychology 72
Arnold A Lazarus

Multinational Corporations 74
Bruce Kogut and Alicja Reuben

Multiple Comparisons, Statistics of 81
Peter H Westfall

Multiple Imputation 88
John Barnard, Nathaniel Schenker, and Donald B. Rubin

Multisensory Integration, Principles of 94
Benjamin A Rowland, Barry E Stein, and Terrence R Stanford

Multisited Field Studies 103
Mark-Anthony Falzon

Multistate Transition Models in Demography 109
Michel Guillot

Multivariate Analysis: Classification and Discrimination 116
Geoffrey McLachlan

Multivariate Analysis: Discrete Variables (Correspondence Models) 121
Z Gilula, Shelby J Haberman, and Peter GM van der Heijden

Multivariate Analysis: Discrete Variables (Loglinear Models) 125
Shelby J Haberman

Multivariate Analysis: Discrete Variables (Overview) 131
Alan Agresti

Mumford, Lewis (1895–1990) 137
Robert Wojtowicz

Museums 142
Jennifer Barrett

Music: Anthropological Aspects 149
Veit Erlmann

Music, History of 155
Albrecht Riethmüller

Myrdal, Alva (1902–86) 159
Eva S Lyon

Myrdal, Gunnar (1898–1987) 165
Richard Swedberg

Mysticism 168
Dan Merkur

Myth in Religion 172
Robert Segal

Myths and Symbols: Organizational 179
Pasquale Gagliardi

Names and Naming 183
João de Pina-Cabral

Nanotechnology: Social and Cultural Aspects 188
Brigitte Nerlich and Szczepan Lemańczyk

Narrative Approaches to Organizations 194
Hervé Corvellec

Narrative Exposure Therapy 198
Maggie Schauer

Narrative in Social Work 204
Christopher Hall

Narrative Networks 211
Katherine Stovel and Daniel Koski-Karell

Narrative, Sociology of 218
Karin Aronsson

Nation, Sociological Aspects of 223
André-Paul Frognier

National Account Systems 227
Werner Reichmann

National Character 233
Federico Neiburg

National Character, History of 237
Stephen Mennell

National Liberation Movements 241
S Amin

National Security and Domestic Terrorism: The US Case 247
Ashmini G Kerodal, Joshua D Freilich, and Steven M Chermak

National Security Studies and War Potential of Nations 253
Steven Weber

National Socialism and Fascism 257
Hans Mommsen

Nationalism: General 262
Partha Chatterjee

Nationalism, Historical Aspects of: Africa 266
Catherine Coquery-Vidrovitch

Nationalism, Historical Aspects of: Arab World 271
Gudrun Krämer

Nationalism, Historical Aspects of: East Asia 275
Yujiro Murata

Nationalism, Historical Aspects of: South Asia 279
Tapan Raychaudhuri

Nationalism, Historical Aspects of: The West 283
Miroslav Hroch

Nationalism, Sociology of 290
Veljko Vujacic

Nations and Nation-States in History 297
John Breuilly

Nation-State and War 304
Richard Lachmann

Nation-State as Symbolic Construct 309
Bart Bonikowski and Nina Gheihman

Natural Disasters: Health-Related Aspects 315
Jan D Reinhardt and James E Gosney

Natural Law 320
Melissa Moschella and Robert George

Naturalism 325
Daniel Andler

Nature and Society in Geography 334
BL Turner, II

Nature-Nurture Controversy, History of 340
Christopher R Badcock

Navigation 345
Ben Finney

Navigation in Virtual Space: Psychological and Neural Aspects 348
Derek A Hamilton and Daniel Barto

Nazi Law 356
Hubert Rottleuthner

Near Middle East/North Africa Studies: Culture 361
Miriam Cooke

Near Middle East/North Africa Studies: Gender 367
Nilüfer Göle

Near Middle East/North Africa Studies: Politics 371
Steve Heydemann

Near Middle East/North Africa Studies: Religion 376
Dale F Eickelman

Near Middle East/North Africa Studies: Society and History 381
Mohammed Bamyeh

Near-Infrared Spectroscopy 387
Judit Gervain

Needham, Joseph (1900–95) 397
Hans U Vogel

Needs Assessment 402
James W Altschuld

Negation, Linguistics of 409
Johan van der Auwera and Denies Du Mon

Negotiation and Bargaining: Organizational Aspects 415
VK Aggarwal and C Dupont

Negotiation and Bargaining: Role of Lawyers 418
Chris Guthrie and Jennifer K Robbennolt

Negotiation and Conflict, The Psychology of 423
Mara Olekalns

Negotiations: Statistical Aspects 430
James K Sebenius

Neighborhood 437
Naomi Carmon and Efrat Eizenberg

Neighborhood Effects 443
Jacob W Faber and Patrick Sharkey

Neighborhood Revitalization and Community Development 450
WD Keating

Neocolonialism and Social Work 456
Vasilios Ioakimidis

Neoliberalism 462
Vishanthie Sewpaul

Neonatal Bonding 469
Aldo B Lucion

Neonatal Crying Behaviors 475
Kathleen Wermke

Neonatal Imitation 481
Giannis Kugiumutzakis and Colwyn Trevarthen

Neonatal Social Isolation: Effects on Neural Function 489
Takuya Takahashi

Neonatal Temperament: From Phenomenology to Mechanism 493
Hiroyuki Tsuchiya

Neo-Piagetian Theories of Cognitive Development 501
Michael F Mascolo

Neo-Vygotskian Developmental Theory 511
Yuriy V Karpov

Network Analysis 518
Antonio M Chiesi

Network Analysis, History of 524
Philipp Korom

Network Centrality, Measures of 532
Peter V Marsden

Network Models of Tasks 540
Richard Schweickert

Network Paradigm: Applications in Organizational Science 546
Charles C Snow and Øystein D Fjeldstad

Networking for School Improvement 551
Steven Katz

Networks and Cultural Consumption 555
Mark C Pachucki

Networks and Meaning 561
Jan Fuhse and Neha Gondal

Networks, Geography of 567
Barney Warf

Networks in Economics 572
Yves Zenou

Neumann, John von (1903–57) 582
Anatol Rapoport

Neural Basis of Language Development 586
Heidi M Feldman and Vanessa N Kovachy

Neural Basis of Reading 593
Brigitte Stemmer

Neural Correlates of Social Development 600
Daina Crafa

Neural Foundation of Morality 606
Roland Zahn, Ricardo de Oliveira-Souza, and Jorge Moll

Neural Plasticity 619
Jon H Kaas

Neural Representations of Direction (Head Direction Cells) 623
Jeffrey S Taube

Neural Representations of Objects 628
Michael J Tarr and John A Pyles

Neural Synchrony as a Binding Mechanism 634
Wolf Singer

Neuroeconomics 639
Stefano F Cappa and Nicola Canessa

Neuroendocrinology 648
Petra Netter

Neuroesthetics 656
Marcos Nadal and Martin Skov

Neuroethics: Ethics of Science and Science of Ethics 664
Pietro Pietrini and Giuseppina Rota

Neurolinguistic Processing of Psychological Verbs 670
Jonathan Brennan

Neurolinguistics, History of 678
Paul Eling

Neurological Approaches to Agrammatism 690
Dimitrios S Kasselimis and Maria Varkanitsa

Neuromarketing 698
Fabio Babiloni

Neurons and Dendrites, Integration of Information in 703
Bartlett W Mel and Bryan Kolb

Neurophysiological Markers of Language Development 708
Makiko Kaga

Neuroprosthetics 714
Mounia Ziat and Sliman Bensmaia

Neuropsychological Testing 722
D Anne Winiarski and Harry A Whitaker

Neuroscience of Education 728
Andrew Tolmie

Neurosociology 736
David D Franks

Neuroticism 743
Weiting Ng

Neurotransmitters 749
Phillip Grant

Neutrality, Axiological 755
Nicholas Rescher

New Genetics and Race 759
Yulia Egorova

New Institutionalism in the Analysis of Complex Organizations 764
Walter W Powell and Patricia Bromley

New Managerialism and New Public Sector Management 770
Tony Cutler

New Media and Democracy in the Arab World 776
David M Faris

New Media and Social Capital 782
Gustavo S Mesch

New Media and the Digital Divide 787
Alexander van Deursen and Jan van Dijk

New Media, News Production and Consumption 793
Eugenia Mitchelstein and Pablo J Boczkowski

New Media, Political Mobilization, and the Arab Spring 798
Habibul H Khondker

New Religious Movements 805
Eileen Barker

New Urbanism 809
Jill L Grant

News Interview 815
Steven E Clayman

News: General 819
Svennik Høyer

Neyman, Jerzy (1894–1981) 824
Sandy L Zabell

Nietzsche, Friedrich (1844–1900) 829
Volker Gerhardt

Nisbet, Robert (1913–96) 832
Paul Gottfried

Nomads and Nomadism in History 838
Fred Schlolz and Günther Schlee

Non Declarative (Procedural) Memory 844
Leonora Wilkinson and Marjan Jahanshahi

Nonequivalent Group Designs 851
Charles S Reichardt and Melvin M Mark

Nongovernmental Organizations: Anthropological and Historical Aspects 856
Julia Elyachar

Non-Heterosexual British Asian Men: Social Representations, Identity and Social Relations 861
Rusi Jaspal

Nonparametric Regression 867
Martin L Hazelton

Nonparametric Statistics: Advanced Computational Methods 878
Myles Hollander and Jayaram Sethuraman

Nonparametric Statistics: Asymptotics 885
Kei Takeuchi

Nonparametric Statistics: Rank-Based Methods 891
Myles Hollander and Jayaram Sethuraman

Nonparametric Statistics: Records 898
Mohammad Ahsanullah and Valery B Nevzorov

Nonphysician Health Professionals 907
Amy W Stimpfel, Jill M Vanak, and Linda H Aiken

Nonprofit Sector, Geography of 914
Vandana Desai

Nonsampling Errors 919
Elizabeth A Stasny

Nonstandard Reasoning 924
Johan van Benthem

CONTRIBUTORS TO VOLUME 16

V.K. Aggarwal
University of California, Berkeley, CA, USA

Alan Agresti
University of Florida, Gainesville, FL, USA

Mohammad Ahsanullah
Rider University, Lawrenceville, NJ, USA

Linda H. Aiken
University of Pennsylvania, Philadelphia, PA, USA

James W. Altschuld
The Ohio State University, Columbus, OH, USA

S. Amin
Third World Forum, Dakar, Senegal

Daniel Andler
Université Paris-Sorbonne, Paris, France

Karin Aronsson
Dep. of Child and Youth Studies, Stockholm University, Stockholm, Sweden

Fabio Babiloni
University of Rome Sapienza, Rome, Italy

Christopher R. Badcock
London School of Economics, London, UK

Mohammed Bamyeh
University of Rochester, Rochester, NY, USA

James A. Banks
University of Washington, Seattle, WA, USA

Eileen Barker
London School of Economics, London, UK

John Barnard
deCODE genetics, Cambridge, MA, USA

Jennifer Barrett
The University of Sydney, Sydney, NSW, Australia

Daniel Barto
University of New Mexico, Albuquerque, NM, USA

Sliman Bensmaia
University of Chicago, Chicago, IL, USA

Pablo J. Boczkowski
Northwestern University, Evanston, IL, USA

Bart Bonikowski
Harvard University, Cambridge, MA, USA

Ingwer Borg
Hausen o.V., Germany

Jonathan Brennan
The University of Michigan, Ann Arbor, MI, USA

John Breuilly
London School of Economics and Political Science, London, UK

Alejandro E. Brice
University of South Florida St. Petersburg, St. Petersburg, FL, USA

Patricia Bromley
University of Utah, Salt Lake, UT, USA

Nicola Canessa
IUSS Pavia, Pavia, Italy; and San Raffaele Scientific Institute, Milan, Italy

Stefano F. Cappa
IUSS Pavia, Pavia, Italy; and San Raffaele Scientific Institute, Milan, Italy

Elof A. Carlson
Stony Brook, NY, USA; and Institute for Advanced Study, Indiana University, Bloomington, IN, USA

Naomi Carmon
Technion — Israel Institute of Technology, Haifa, Israel

Partha Chatterjee
Columbia University, New York, NY, USA; and Centre for Studies in Social Sciences, Calcutta, India

Steven M. Chermak
Michigan State University, East Lansing, MI, USA; and University of Maryland, College Park, MD, USA

Antonio M. Chiesi
Università degli Studi di Milano, Milano, Italy

Steven E. Clayman
University of California, Los Angeles (UCLA), Los Angeles, CA, USA

Miriam Cooke
Duke University, Durham, NC, USA

Catherine Coquery-Vidrovitch
Université Paris 7 Denis-Diderot, Paris, France

Hervé Corvellec
Lund University, Lund, Sweden

Daina Crafa
Douglas Mental Health Institute, McGill University, Montreal, QC, Canada

Tony Cutler
Centre for Research on Socio-Cultural Change, University of Manchester, Manchester, UK

Vandana Desai
Royal Holloway, University of London, Egham, UK

Denies Du Mon
Universiteit Antwerpen, Antwerpen, Belgium

C. Dupont
Graduate Institute of International Studies, Geneva, Switzerland

Yulia Egorova
Durham, UK

Dale F. Eickelman
Dartmouth College, American University of Kuwait Program, Hanover, NH, USA

Efrat Eizenberg
Technion — Israel Institute of Technology, Haifa, Israel

Paul Eling
Radboud University Nijmegen, Nijmegen, The Netherlands

Julia Elyachar
School of Social Sciences, UC of Irvine, Irvine, CA, USA

Thomas Hylland Eriksen
University of Oslo, Oslo, Norway

Veit Erlmann
University of Texas at Austin, Austin, TX, USA

Jacob W. Faber
New York University, New York, NY, USA

Mark-Anthony Falzon
Clare Hall, Cambridge, UK; and University of Malta, Msida, Malta

David M. Faris
Roosevelt University, Chicago, IL, USA

Heidi M. Feldman
Stanford University School of Medicine, Palo Alto, CA, USA

Ben Finney
University of Hawaii, Honolulu, HI, USA

Øystein D. Fjeldstad
BI-Norwegian Business School, Oslo, Norway

David D. Franks
Virginia Commonwealth University, Richmond, VA, USA

Joshua D. Freilich
City University of New York, New York, NY, USA; and University of Maryland, College Park, MD, USA

André-Paul Frognier
University of Louvain, Louvain-La-Neuve, Belgium

Jan Fuhse
Humboldt University, Berlin, Germany

Pasquale Gagliardi
Catholic University, Milano, Italy; and Fondazione Giorgio Cini, Venezia, Italy

Robert George
Princeton University, Princeton, NJ, USA

Volker Gerhardt
Humboldt University, Berlin, Germany

Judit Gervain
Centre National de la Recherche Scientifique (CNRS), Université Paris Descartes, Paris, France

Nina Gheihman
Harvard University, Cambridge, MA, USA

Z. Gilula
Hebrew University, Jerusalem, Israel

Nilüfer Göle
Ecole des Hautes Etudes en Sciences Sociales, Paris, France

Neha Gondal
The Ohio State University, Columbus, OH, USA

James E. Gosney
International Society of Physical and Rehabilitation Medicine (ISPRM), Geneva, Switzerland

Paul Gottfried
Elizabethtown College, Elizabethtown, PA, USA

Jill L. Grant
Dalhousie University, Halifax, NS, Canada

Phillip Grant
Justus-Liebig-University, Giessen, Hesse, Germany

Patrick J.F. Groenen
Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands

Michel Guillot
Population Studies Center, University of Pennsylvania, Philadelphia, PA, USA

Chris Guthrie
Vanderbilt Law School, Nashville, TN, USA

Shelby J. Haberman
Educational Testing Service, Princeton, NJ, USA

Shelby J. Haberman
Northwestern University, Evanston, IL, USA

Christopher Hall
Durham University, Durham, UK

Derek A. Hamilton
University of New Mexico, Albuquerque, NM, USA

Martin L. Hazelton
Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

Steve Heydemann
Columbia University, New York, NY, USA

Myles Hollander
Florida State University, Tallahasse, FL, USA

Svennik Høyer
University of Oslo, Oslo, Norway

Miroslav Hroch
Charles University in Prague, Praha, Czech Republic

Vasilios Ioakimidis
Durham University, Durham, UK

Duncan Ivison
University of Sydney, Sydney, NSW, Australia

Marjan Jahanshahi
Cognitive Motor Neuroscience Group, University College London, London, UK; and Institute of Neurology, National Hospital for Neurology & Neurosurgery, University College London, London, UK

C.-T. James Huang
Harvard University, Cambridge, MA, USA

Rusi Jaspal
De Montfort University, Leicester, UK

Ulrike Jessner-Schmid
University of Innsbruck, Innsbruck, Austria; and University of Pannonia, Veszprem, Hungary

Kwanghee Jung
University of Texas Health Science Center at Houston, Houston, TX, USA

Jon H. Kaas
Vanderbilt University, Nashville, TN, USA

Makiko Kaga
National Institute of Mental Health, National Center of Neurology and Psychiatry, Kodaira, Tokyo, Japan; and Tokyo Metropolitan Medical Center East for Children/Adults with Developmental Disabilities, Tokyo, Japan

Yuriy V. Karpov
Touro College, New York, NY, USA

Dimitrios S. Kasselimis
University of Crete, Heraklion, Greece; and National and Kapodistrian University of Athens, Athens, Greece

Steven Katz
Ontario Institute for Studies in Education/University of Toronto, Toronto, ON, Canada

W.D. Keating
Cleveland State University, Cleveland, OH, USA

Ashmini G. Kerodal
City University of New York, New York, NY, USA

Habibul H. Khondker
Zayed University, Abu Dhabi, United Arab Emirates

Bruce Kogut
Columbia Business School, Columbia University, New York, NY, USA

Bryan Kolb
University of Southern California, Los Angeles, CA, USA

Philipp Korom
Max Planck Institute for the Study of Societies, Cologne, Germany

Daniel Koski-Karell
University of Washington, Seattle, WA, USA

Vanessa N. Kovachy
Stanford University School of Medicine, Palo Alto, CA, USA

Gudrun Krämer
Free University of Berlin, Berlin, Germany

Giannis Kugiumutzakis
University of Crete, Crete, Greece

Richard Lachmann
University at Albany, State University of New York, Albany, NY, USA

Arnold A. Lazarus†
Rutgers University, New Brunswick, NJ, USA

Szczepan Lemańczyk
University of Nottingham, Nottingham, UK

Aldo B. Lucion
Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil

Eva S. Lyon
Faculty of Arts and Humanities, London South Bank University, London, UK

Melvin M. Mark
Pennsylvania State University, University Park, PA, USA

Peter V. Marsden
Harvard University, Cambridge, MA, USA

Michael F. Mascolo
Merrimack College, North Andover, MA, USA

Geoffrey McLachlan
University of Queensland, Brisbane, QLD, Australia

Bartlett W. Mel
University of Southern California, Los Angeles, CA, USA

Stephen Mennell
School of Sociology, University College Dublin, Belfield, Dublin, Ireland

Dan Merkur
University of Toronto, Toronto, ON, Canada

Gustavo S. Mesch
University of Haifa, Haifa, Israel

Eugenia Mitchelstein
Northwestern University, Evanston, IL, USA

Jorge Moll
D'Or Institute for Research and Education (IDOR), Rio de Janeiro, Brazil

Hans Mommsen
Feldafing, Germany

Melissa Moschella
The Catholic University of America, Washington, DC, USA

Yujiro Murata
University of Tokyo, Tokyo, Japan

Marcos Nadal
University of Vienna, Vienna, Austria

Hiroki Narita
College of Industrial Technology, Nihon University, Chiba, Japan

Federico Neiburg
Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

Brigitte Nerlich
University of Nottingham, Nottingham, UK

Petra Netter
University of Giessen, Giessen, Germany

Valery B. Nevzorov
St. Petersburg State University, St. Petersburg, Russia

Weiting Ng
School of Arts and Social Sciences, SIM University, Singapore

Mara Olekalns
The University of Melbourne, Carlton, VIC, Australia

Ricardo de Oliveira-Souza
D'Or Institute for Research and Education (IDOR), Rio de Janeiro, Brazil

Mark C. Pachucki
Harvard Medical School, Boston, MA, USA

Pietro Pietrini
University of Pisa Medical School, Azienda Ospedaliero-Universiaria Pisana, Pisa, Italy

João de Pina-Cabral
University of Kent, Canterbury, UK

Walter W. Powell
Stanford University, Stanford, CA, USA

John A. Pyles
Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA, USA

Anatol Rapoport
University of Toronto, Toronto, ON, Canada

Tapan Raychaudhuri
Oxford, UK

† Deceased.

Charles S. Reichardt
University of Denver, Colorado, CO, USA

Werner Reichmann
University of Konstanz, Konstanz, Germany

Jan D. Reinhardt
Institute for Disaster Management and Reconstruction of Sichuan University and Hong Kong Polytechnic University, Chengdu, Sichuan, China

Nicholas Rescher
University of Pittsburgh, Pittsburgh, PA, USA

Alicja Reuben
Manhatten College, New York, NY, USA

Albrecht Riethmüller
Free University of Berlin, Berlin, Germany

Jennifer K. Robbennolt
University of Illinois College of Law, Champaign, IL, USA

Germán Rodríguez
Princeton University, Princeton, NJ, USA

Giuseppina Rota
University of Pisa Medical School, Azienda Ospedaliero-Universiaria Pisana, Pisa, Italy

Hubert Rottleuthner
Free University of Berlin, Berlin, Germany

Benjamin A. Rowland
Wake Forest School of Medicine, Winston-Salem, NC, USA

Donald B. Rubin
National Center for Health Statistics, Hyattsville, MD, USA

Maggie Schauer
Clinical Psychology, University of Konstanz & vivo international, Konstanz, Germany

Nathaniel Schenker
Harvard University, Cambridge, MA, USA

Günther Schlee
Max Planck Institute for Social Anthropology, Halle (Saale), Germany

Fred Schlolz
Free University of Berlin, Berlin, Germany

Richard Schweickert
Purdue University, West Lafayette, IN, USA

James K. Sebenius
Harvard Business School, Harvard University, Boston, MA, USA

Robert Segal
University of Aberdeen, Aberdeen, UK

Jayaram Sethuraman
Florida State University, Tallahasse, FL, USA

Vishanthie Sewpaul
School of Applied Human Sciences, Howard College Campus, University of KwaZulu Natal, Durban, South Africa

Patrick Sharkey
New York University, New York, NY, USA

Wolf Singer
Max Planck Institute (MPI) for Brain Research, Frankfurt, Germany; Ernst Srüngmann Institute (ESI) for Neuroscience in Cooperation with Max Planck Society, Frankfurt, Germany; and Frankfurt Institute for Advanced Studies (FIAS), Frankfurt, Germany

Martin Skov
Copenhagen University Hospital Hvidovre, Copenhagen, Denmark; and Copenhagen Business School, Copenhagen, Denmark

Charles C. Snow
The Pennsylvania State University, State College, PA, USA; and BI-Norwegian Business School, Oslo, Norway

David A. Snow
University of California, Irvine, CA, USA

Terrence R. Stanford
Wake Forest School of Medicine, Winston-Salem, NC, USA

Elizabeth A. Stasny
Ohio State University, Columbus, OH, USA

Barry E. Stein
Wake Forest School of Medicine, Winston-Salem, NC, USA

Brigitte Stemmer
Université de Montréal, Montreal, QC, Canada

Amy W. Stimpfel
New York University, New York, NY, USA

Katherine Stovel
University of Washington, Seattle, WA, USA

Richard Swedberg
Cornell University, Ithaca, NY, USA

Takuya Takahashi
Graduate School of Medicine, Yokohama City University, Yokohama, Japan; and Albert Einstein College of Medicine, Bronx, NY, USA

Yoshio Takane
University of Victoria, Victoria, BC, Canada

Kei Takeuchi
Meiji Gakuin University, Yokohama, Japan

Anna E. Tan
University of California, Irvine, CA, USA

Michael J. Tarr
Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA, USA

Jeffrey S. Taube
Dartmouth College, Hanover, NH, USA

Andrew Tolmie
University of London, London, UK

Colwyn Trevarthen
University of Edinburgh, Edinburgh, UK

Hiroyuki Tsuchiya
Fukuda Hospital, Kumamoto, Japan

B.L. Turner
Arizona State University, Tempe, AZ, USA

Jill M. Vanak
University of Pennsylvania, Philadelphia, PA, USA

Johan van Benthem
University of Amsterdam, Amsterdam, The Netherlands

Johan van der Auwera
Universiteit Antwerpen, Antwerpen, Belgium

Peter G.M. van der Heijden
Utrecht University, Utrecht, The Netherlands

Alexander van Deursen
University of Twente, Enschede, The Netherlands

Jan van Dijk
University of Twente, Enschede, The Netherlands

Maria Varkanitsa
University College London, London, UK

Hans U. Vogel
Universität Tübingen, Tübingen, Germany

Veljko Vujacic
Oberlin College, Oberlin, OH, USA

Barney Warf
University of Kansas, Lawrence, KS, USA

Steven Weber
University of California, Berkeley, CA, USA

Kathleen Wermke
University Clinics, University of Würzburg, Würzburg, Germany

Peter H. Westfall
Texas Tech University, Lubbock, TX, USA

Harry A. Whitaker
Northern Michigan University, Marquette, MI, USA

Leonora Wilkinson
Cognitive Motor Neuroscience Group, University College London, London, UK; Institute of Neurology, National Hospital for Neurology & Neurosurgery, University College London, London, UK; and Behavioral Neurology Unit, National Institute of Neurological Disorders and Stroke, Bethesda, MD, USA

D. Anne Winiarski
Emory University, Atlanta, GA, USA

Robert Wojtowicz
Old Dominion University, Norfolk, VA, USA

Sandy L. Zabell
Northwestern University, Evanston, IL, USA

Roland Zahn
Translational Congnitive Neuroscience of Affective Disorders Laboratory, The University of Manchester, School of Psychological Sciences, Manchester, UK

Yves Zenou
Stockholm University, Stockholm, Sweden; and IFN, Stockholm, Sweden

Mounia Ziat
Northern Michigan University, Marquette, MI, USA

Movement Theory and Constraints in Syntax

C-T James Huang, Harvard University, Cambridge, MA, USA
Hiroki Narita, College of Industrial Technology, Nihon University, Chiba, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the considerations that motivate the postulation of movement transformations in generative linguistics, and traces the evolution of movement theory from the early days of transformational grammar through the development of trace theory, modularity, and Move α, to the contemporary theory of Internal Merge. It further describes the development of the theory of movement constraints, starting from the A-over-A Principle and construction-specific island constraints to various subsequent approaches, incorporating notions like cyclicity, headedness, traces, and economy. The notion of covert movement is also briefly reviewed, in light of the syntax–semantics interface and crosslinguistic parametric variations.

Movement in Syntactic Theory

Movement theory is a 'trademark' of Generative-Transformational Grammar (*see* Syntax (General); Generative Grammar; Linguistics: Incorporation). In this framework of natural language syntax, certain sentences are generated not just by combining words or phrases in some appropriate order, but also by moving them from one position to another. Thus the sentences in (1)–(3) are not constructed by directly 'merging' the leading items *will*, *what*, and *John* respectively at the sentence-initial position, but by first introducing them in their 'original' positions (marked by the blanks) and then fronting them to the initial position:

(1) *Will* John __ buy this toy?

(2) *What* has John put __ on the table?

(3) *John* was arrested __ yesterday.

In the Principles and Parameters Theory as implemented in the Government-and-Binding framework (Chomsky, 1981, 1986a,b), the computational system of a grammar generates underlying D(eep)-Structure (DS) representations in accordance with the X′ Theory that satisfy the selectional requirements of lexical items (*see* Valency and Argument Structure in Syntax), and transformational operations (particularly, movement operations) convert them to S(urface)-Structure (SS) representations. These SS representations are in turn mapped onto two levels that form the interfaces between grammar and other cognitive performance systems, PF (Phonetic Form) or PHON (phonetic interface) on the one hand and LF (Logical Form) or SEM (semantic interface) on the other (*see* Syntax–Semantics Interface). Thus for (2), the computational system will carry out movement operations (as indicated by the arrows) that turn its DS to its SS representation, as shown in Figure 1:

It has been assumed that the mappings in SS → PF and SS → LF also involve movement operations.

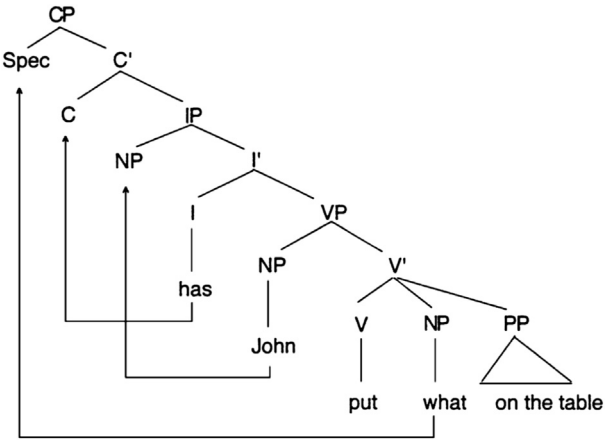

Figure 1 Tree diagram showing the derivation of 'What has John put on the table?'

Motivations for Movement

In some cases the postulation of movement can be justified by intuitive observations. The verbs *put* and *arrest* are transitive in traditional terms, requiring noun phrase (NP) objects following them (cf **John put on the table*; **Bill arrested*; an asterisk preceding a sentence indicates ungrammaticality). In (2) and (3), no NP object appears in the 'expected' postverbal position, despite their grammaticality. This paradox is explained if one says that *put* does have a postverbal object at DS, though the object is fronted at SS (as depicted in Figure 1). The postulation of movement thus enables one to preserve generalizations about the predicate–argument structure of verbs and other lexical items (*see* Valency and Argument Structure in Syntax) that explain grammatical well-formedness.

In other cases, the movement hypothesis has enabled generative syntacticians to unearth interesting phenomena that had escaped introspection in previous descriptions, and gain valuable insight into syntactic structure. For example, it is now widely assumed that, despite their surface similarity, derivation of (4a), but not of (4b), involves movement:

(4) a. Harry seemed to be honest.

b. Harry tried to be honest.

In particular, in (4a) but not (4b), the subject of the main verb *seem* originates as the subject of *be honest* and raises to the higher subject position at SS. Postulating Subject Raising in (4a) but not (4b) explains several differences between *seem*-type and *try*-type sentences that surface upon closer examination, e.g., the contrast between *The shit seems to be about to hit the fan* (under the idiomatic interpretation) but **The shit tries to be about to hit the fan*, etc. Similarly interesting facts have been addressed concerning examples like *Mary is likely to leave* versus *Mary is eager to leave* versus *Mary is easy to leave*.

Still in other cases, the existence of movement cannot be detected by intuitive observations of the type one can make concerning (1)–(3). One early rule, proposed in Chomsky (1955) and popularly known as Affix Hopping, concerned tense marking in English finite sentences. The fact is that tense marking seems to 'migrate' among auxiliary elements and the main verb: it may occur on a modal (*Monica* 'would' *sing*, but **Monica will sang*), on a perfective or progressive aspect (*Monica* 'has' *sung*, *Monica* 'is' *singing*, but **Monica have sings*, **Monica have is singing*, etc.), or directly on the main verb (*Monica* 'sang'). Traditional methods of 'immediate constituent analysis' would face difficulty in describing this range of phenomena in any simple way. Chomsky proposed treating the tense inflectional suffix as an independent syntactic item in DS representation (as the leftmost constituent of the Auxiliary) which, by Affix Hopping, 'docks' on whatever first auxiliary or verbal element follows it: thus, *Monica* PAST *will sing* → *Monica will*-PAST *sing* (=*Monica would sing*), *Monica* PAST *have+en sing* → *Monica have*-PAST *sing-en* (=*Monica had sung*), *Monica* PAST *sing* → *Monica sing*-PAST (=*Monica sang*), and so on. Affix Hopping vastly simplifies the grammar of tense marking, at the same time capturing the observed 'migration' (hopping) property insightfully.

The primary argument for movement is that it allows one to simplify grammatical descriptions, capture significant generalizations, and uncover deep insights into the nature of human language as a 'mirror of the mind.' Such considerations justify the movement hypothesis not only for more observable cases like (1)–(3) and (4a), but also many surprising or less obvious cases. (Some versions of generative syntax incorporate other devices to explain the 'movement phenomena,' including Lexical Functional Grammar Head-Driven Phrase Structure Grammar, Categorial Grammar, Dependency Grammar.)

Evolution of Movement Theory

Early Movement Theory

Chomsky's (1955, 1957) foundational work on English syntax led to a massive explosion of transformational generative studies of various languages and constructions. New syntactic phenomena were unearthed or became theoretically interesting to call for an analysis, and new rules were added to the repertoire of a grammar's transformational component. In addition to the aforementioned Affix Hopping, various other movement rules were also proposed, including Subject–Auxiliary Inversion (1), *Wh*-Movement (2), Passivization (3), Subject Raising (4a), as well as Relativization (5), Topicalization (6), *Tough*-Movement (7), Extraposition from NP (8), etc.

(5) This is the person *who* I have wanted to meet __ always.

(6) *This sort of movie*, I would never watch __.

(7) *John* is tough for us to deal with __.

(8) A book __ has just appeared *on syntactic theory*.

In order to ensure descriptive adequacy, each rule proposed was formulated in such a way as to (a) apply to only the input strings that meet a certain Structural Description (SD), (b) effect precisely the Structural Change (SC) as desired, and (c) only do so under any additional conditions as may be imposed. For example, the Passivization rule of English was defined in terms of the SD of an input structure that contains a postverbal object NP and the instruction for SC that it will (among other things) front the NP to the subject position. Additional conditions are noted for the applicability of the rule, among them the requirements that the main verb be put in the passive participle form, that the movement be carried out to an immediate local subject position, etc. Other rules were defined with similar explicitness and detail. Because the investigations were carried out in the generative (i.e., formal) approach, linguists working on syntactic phenomena were required to formulate precise hypotheses about the form and functioning of movement transformations. The explicit (and therefore falsifiable) formulations led to unprecedented expansion of our explicit knowledge of natural language syntax.

Modularity: From Rules to Principles

While the detailed formulations of movement rules helped ensure observational and descriptive adequacies of proposed grammatical descriptions, linguists increasingly realized that this was achieved at the cost of explanatory adequacy and conceptual beauty. For one thing, the detailed statement of each movement rule obscured the fact that many movement rules have similar properties. By repeatedly stipulating these properties in the statement of every rule, the theory became highly redundant and failed to capture the generality of such properties. In fact, some properties often included in the statement of movement rules are not peculiar to movement, but hold generally of other syntactic phenomena as well. To capture the correct generalizations, then, it was necessary to separate out general properties from the statement of specific rules.

One property of movement that was repeatedly stipulated in the formulation of each rule is that movement is generally *upward*, i.e., that it moves elements to a 'c-commanding' position. Defined in terms of tree diagrams like **Figure 1**, a node

α c-commands β iff neither α nor β dominates (contains) the other, and the first branching node dominating α dominates β. In **Figure 1**, each instance of movement is upward to a position c-commanding the original position. This 'upward movement requirement' should be stated once and for all movement rules, rather than separately for each rule. In fact, Chomsky (1975) noted that this requirement is but a special case of a more general principle that also governs anaphora in natural language. For example, the interpretation of a reflexive pronoun in English requires that it have an antecedent that c-commands it: *John criticized himself*, but **John's mother criticized himself* (the latter is ungrammatical because the intended antecedent *John* is not high enough to c-command *himself*). To capture this common property between movement and anaphora, Chomsky proposed that when a category is moved, it leaves a 'trace' behind, co-indexed with the moved category (as in *John*$_i$ *was arrested* t_i *yesterday* for (3)). The trace, being phonetically empty and having no inherent reference, depends on the moved category as its 'antecedent,' much as a reflexive depends on another NP for its reference. The 'upward movement requirement' thus reduces to a general requirement governing anaphoric dependency and may be eliminated from the statement of each movement rule.

Another common property of movement rules is that they operate only *locally*, as the following sections will show. Further, it has also been observed that movement typically operates in a 'structure-preserving' manner, so that a head (e.g., verb) can only move to a head position, an argument NP to an argument position (A-position), and an operator to an operator position (A′-position). (See again **Figure 1**, illustrating *has* undergoing head-movement, *John* undergoing A-movement, and *what* undergoing A′-movement.) The discovery of these and other general properties finally led linguists to the view that most details formerly attributed to individual rules should be dissociated from them and be ascribed to independent modules of the grammar with general applicability.

Generalizing Movement Operations: Move α and Internal Merge

One approach that culminated in the theory of Government and Binding (Chomsky, 1981) holds that all movement rules can be dispensed with, in favor of a single rule of maximal generality: Move α, α a category. This operation can apply freely and dislocate anything anywhere, irrespective of specific properties of the moving element and the target site, while the well-formedness of the output is to be determined by the other modules of Universal Grammar, such as the aforementioned principles of anaphoric dependency and structure preservation, as well as the locality constraints to be discussed in the 'Constraints on Movement' section.

An even simpler theory of movement, advocated in the contemporary literature, is to further reduce Move α to a special case of the general structure-building operation Merge. Merge is an elementary combinatory operation that recursively combines two syntactic objects, α and β, and forms a constituent comprising of them, thus Merge(α, β) = [α β] (or more precisely a set of α and β, {α, β}, with the label specified in some way or the other; see Chomsky, 1995 *et seq*). One mode of Merge, called External Merge, iteratively combines two independent objects, external to each other, and serves to generate basic structures of sentences from the bottom up. A sample Merge-based derivation of a sentence is given in (9). External Merge is specifically exemplified in (9a–d,f). Moreover, assuming that Merge can apply freely, it may also combine two syntactic objects α, β that are *not* independent of each other, that is, two objects one of which, say α, is internal to the other object β. This mode of Merge-application, called Internal Merge, is exemplified by (9e) and (9g). The underscore marks the relevant α internal to β. As indicated by indices, Internal Merge yields two copies (occurrences) of α, one within β and the other to the edge of β.

(9) A Merge-based derivation of *the apple, John will eat* ___.

a. Merge(the, apple) = [the apple]

b. Merge(eat, [the apple]) = [eat [the apple]]

c. Merge(John, [eat [the apple]]) = [John [eat [the apple]]]

d. Merge(will, [John [eat [the apple]]]) = [will [John [eat [the apple]]]]

e. Merge (*John*, [will [*John* [eat [the apple]]]]) = [John$_i$ [will [John$_i$ [eat [the apple]]]]]

f. Merge(C, [John$_i$ [will [John$_i$ [eat [the apple]]]]]) = [C, [John$_i$ [will [John$_i$ [eat [the apple]]]]]]

g. Merge(*[the apple]*, [C, [John$_i$ [will [John$_i$ [eat *[the apple]*]]]]]) = [[the apple]$_j$, [C, [John$_i$ [will [John$_i$ [eat [the apple]$_j$]]]]]]

Assuming that they undergo deletion at the phonological component, lower copies created by Internal Merge can be thought of as equivalent to traces of earlier movement transformations (the copy theory of movement, replacing the earlier trace theory of movement mentioned above). Thus, applications of Internal Merge in (9e) and (9g) correspond to A-movement of *John* and A′-movement (topicalization) of *the apple*, respectively. In this manner, the contemporary framework of Merge-based syntax (called the theory of bare phrase structure; see Chomsky, 1995 *et seq*) provides a unified theory of bottom-up structure building and movement transformations, incorporating the effect of Move α as a special case of Merge, a further simplification of Universal Grammar.

So long as these highly modular approaches to movement are proven to be successful, there is little or nothing left to do in the formulation of the movement rule itself. Consequently, increasing attention has shifted to the study of other modules of grammar, in particular the study of general locality constraints on movement and on the distribution of empty categories (traces) created by movement.

Constraints on Movement

In generative linguistics, a grammar G of a given language L is a hypothesis about the linguistic phenotype of the speakers of L. Such a hypothesis is descriptively adequate only if it defines all and only the (potential and actual) grammatical sentences of the language. A general linguistic theory or Universal Grammar, a hypothesis of the human linguistic genotype, is likewise adequate only if it defines all and only the possible grammars that are descriptively adequate. In the early stages of generative grammar study, linguists were primarily occupied with providing their grammars with enough descriptive powers to generate all grammatical sentences. In subsequent developments up to the present time, however, major attention has been directed toward restricting the expressive powers of their grammars. The study of movement constraints was motivated not only on empirical grounds of descriptive and explanatory adequacy, but also because it has proven a very productive probe into the nature of the human language faculty.

The A-over-A Principle and Ross's Island Constraints

The need for a general, powerful constraint on movement rules was noted quite early by Chomsky (1964), who proposed the A-over-A Principle to exclude movement of a phrase out of a larger phrase of the same type. The importance of movement constraints gained widespread attention with the emergence of Ross (1967), who showed the A-over-A Principle to be inadequate and proposed, in its place, a module of grammar containing a set of 'island constraints.' According to Ross, although movement rules may appear to displace constituents over an arbitrary distance (e.g., *What$_i$ did Harry say that Maria bought t_i?*), they must be sharply constrained so as not to extract elements out of certain domains that he called 'islands.' The following examples respectively illustrate Ross's island constraints (in each case the relevant 'island' is marked by square brackets): Coordinate Structure Constraint (CSC), Complex NP Constraint (CNPC), and Sentential Subject Condition (SSC). Other familiar constraints include the Subject Condition (SC), the Adjunct Condition (AC), and the *Wh*-Island Condition (WIC):

(10) *What$_i$ did [Harry buy t_i and Maria buy a pencil]? (CSC)

(11) *What$_i$ did Harry talk to [the woman who bought t_i]? (CNPC)

(12) *What$_i$ was [that Maria had bought t_i] surprising? (SSC)

(13) *Who$_i$ did [a picture to t_i] please you? (SC)

(14) ?*What$_i$ did Harry go home [before Maria bought t_i]? (AC)

(15) ?*What$_i$ did Harry wonder [who bought t_i]? (WIC)

The Ross-style constraints properly restrict the applicability of movement, but, as given, they amounted to a list of environments that form 'islands' to the exclusion of those that do not, and thus failed to explain their clustering. To overcome this problem, subsequent research has focused on how such idiosyncratic constraints can be reduced to a more natural set of general principles. Four types of approaches can be identified in this research endeavor, to which we turn one by one.

Cyclicity-Based Accounts: Subjacency, Barriers, and Phases

Chomsky (1973) proposed that most of Ross's island constraints can be subsumed under a general constraint dubbed the Subjacency Condition, which prohibits extraction of a constituent across two or more bounding (i.e., cyclic) nodes, defined as NP or S (the latter corresponding to IP under the generalized X-bar theory of Chomsky (1986b). The theory of Subjacency not only provided a unified account of CSC, CNPC, SSC, and WIC among others (see Chomsky, 1973 for details), but it also put forward the idea that syntactic operations (movement among others) are fundamentally governed by the principle of strict cycle: grammatical rules first operate within the smallest (most deeply embedded) cycle and then apply to successively larger dominating cycles. Under this cyclic model, long-distance dependency (e.g., interclausal movement) can be established only successive-cyclically, spanning over at most one bounding node at a time.

The cyclicity-based approach continued in Chomsky (1986b), where the notion of *barrierhood* was proposed to unify various conditions on movement and other linguistic relations (government, binding, etc.). In the *Barriers* model, movement is allowed to cross at most one barrier at a time, deriving the effect of Subjacency and Huang's (1982) Condition on Extraction Domain (see the 'Headedness-Based Accounts: Condition on Extraction Domain and Its Reformulations' section) while overcoming most of the shortcomings of these earlier proposals. The *Barriers* framework also provided a tool for measuring the degree of (un)grammaticality in incremental terms, capturing the fact that a movement crossing n barriers produces worse results than one that crosses $n - 1$ barriers. In this manner, various constraints defined in terms of barrierhood collectively derive the cyclicity of syntactic computation, whereby the grammar selects movement and other operations that are local, or as local as possible.

The notion of 'strict' cyclicity is further incorporated into the contemporary framework of Phase Theory (Chomsky, 2000, 2001, 2004, 2007, 2008; see also Uriagereka, 1999; Narita, 2014; to be discussed in the 'Headedness-Based Accounts: Condition on Extraction Domain and Its Reformulations' section). It holds that each syntactic derivation is partitioned into a set of smaller periodic units called 'phases,' basically corresponding to clauses (CPs) and transitive verb phrases (v*Ps). Once completed, past phases are eliminated from the derivational workspace and subjected to semantic and phonetic interpretation, and thereby become inaccessible for further computation (Phase-Impenetrability Condition (PIC)). Movement and other syntactic operations should therefore apply phase by phase, yielding the effect of successive cyclicity.

Headedness-Based Accounts: Condition on Extraction Domain and Its Reformulations

The second type of approach to movement constraints capitalizes on the headedness (or endocentricity) of phrase structure. According to the traditional X-bar theory (see, e.g., Chomsky, 1986a,b), every phrase is understood as a 'projection' of some designated lexical item (an atomic element drawn directly from the lexicon), which is called its 'head': thus, NP is a projection of an N, CP is a projection of a complementizer (C), etc. The X-bar theory further defines several structural positions within an XP for each head X. For example, the 'complement' of X is the sister of X: the object NP is the complement of V within $[_{VP}$ V NP] etc. Further, all other phrases within XP are called noncomplements, and specifically, the noncomplement that is immediately dominated (contained) by XP is called the 'specifier' of X: in Chomsky's (1986b) system where the traditional label S(entence) is replaced with 'Infl(ection) Phrase (IP),' the subject NP located within $[_{IP}$ NP $[_{I'}$ I VP]] (with I′ an intermediate projection of I) is called the specifier of I. Finally, all noncomplements that are not specifiers of some head are called 'adjuncts.'

Huang (1982) argued that the SSC, SC, and AC (12)–(14) reflect a general complement versus noncomplement asymmetry and are best subsumed under one general Condition on Extraction Domain (CED), which holds that extraction is possible only from 'properly governed' domains, i.e., complements (see also Kayne, 1981). This condition precludes extraction from noncomplements like specifiers (including subjects) and adjuncts, explaining (12)–(14).

While the notion of government (and proper government) played a pivotal role in the Government-and-Binding framework (Chomsky, 1981, 1986a,b), subsequent studies have tried to dispense with this notion in favor of a maximally simplified theory of Universal Grammar in line with the Minimalist Program (see Chomsky, 1995). As a result, much research has been devoted to possible refinements and reformulations of Huang's CED-based account. One such approach was advocated by Uriagereka's (1999) cyclic Spell-Out model. This theory proposes a unidirectional left-to-right head-complement branching pattern *à la* Kayne (1994), and argues that any structure that deviates from such head-complement recursion, i.e., phrasal specifiers and adjuncts, must be subjected to Spell-Out (mapping to phonetic interpretation) and reduced to a unitary headlike element ('gigantic compound' in Uriagereka's terms). Cyclic application of Spell-Out was assumed to render each noncomplement opaque to further computation (an assumption akin to Chomsky's PIC discussed in the 'Cyclicity-Based Accounts: Subjacency, Barriers, and Phases' section), which successfully accounts for their islandhood without recourse to proper government. However, Uriagereka's theory necessarily implies a strict universal noncomplement–head-complement word order, which poses an insurmountable problem in accounting for cases involving head-final word orders or right-adjunction in various languages. For this matter, Narita (2014) proposes an alternative conception of recursive head-complement branching that is akin to Uriagereka's but free from stipulations on word order. Narita's approach reconciles the basic insights of Huang's CED, Uriagereka's cyclic Spell-Out and Chomsky's PIC, while it also provides some account of certain exceptions to the effects of SC and AC.

Trace-Based Accounts: Empty Category Principle

The third type of movement constraints has to do with the distribution of movement traces. In the earlier framework, proper government played a crucial role in constraining not only extraction *from* subjects and adjuncts (as in CED; see the 'Headedness-Based Accounts: Condition on Extraction Domain and Its Reformulations' section) but also extraction *of* subjects and adjuncts. Sentences (15)–(16) exemplify cases where *long* extraction of a subject or an adjunct is prohibited:

(15) *This is the person who_i John wondered [whether t_i would fix the car].

(16) *This is how_i John wondered [whether Bill would fix the car t_i].

These cases were accounted for by the Empty Category Principle (ECP), which requires that a trace not be separated from its governor by any barrier at all. Due to the ECP, each trace must be governed by a lexical head, or extraction must be local (short distance). Long extraction in (15)–(16) is correctly ruled out by the ECP, because the relevant traces are not governed by any lexical head.

The current status of the ECP becomes somewhat controversial, partially because of subsequent attempts to eliminate the notion of (proper) government, building on various theory-internal considerations. However, the elimination of proper government has, as yet, left a few questions unanswered (e.g., the difference between argument extraction and adjunct extraction out of an island, etc.). Therefore, an alternative characterization of the ECP effect is left for future research.

Economy-Based Accounts: Minimality and Last Resort

The fourth category of locality constraints grows out of the research guided by the so-called Minimalist Program (Chomsky, 1995), which aims to achieve a minimal formulation of grammatical theory. We can approach this goal either by reducing the set of theoretical apparatuses to a bare minimum, or by attributing them to principles of 'least effort' or minimal computation that are more deeply grounded in laws of nature, hence not adding to the burden of Universal Grammar.

One such attempt was Rizzi's (1990) theory of Relativized Minimality, according to which each movement operation of type X (A-, A′-, or head-movement) must always land at the closest potential landing site of the same type X. For example, (17b) and (18c) are ruled out as A-movement and

head-movement skipping an intermediate A-position and head-position, respectively.

(17) a. It seems that *Juma*$_i$ is likely [t_i to win].

b. **Juma*$_i$ seems that it is likely [t_i to win].

(18) a. They could have left.

b. *Could*$_i$ they t_i have left?

c. **Have*$_i$ they could t_i left?

Cases of the WIC (15) are also ruled out as A′-movement violating Relativized Minimality. In Chomsky (1995), minimality is directly incorporated into the definition of Attract F, a reformulation of Move α that seeks and attracts a category with the matching feature F that is closest to the relevant landing site (for example, *wh*-movement of *what* into the specifier of CP = [C *you* Infl *see what*] is reformulated as an instance of Attract F, initiated by C's Q(uestion)-feature attracting the matching Q-feature of *what*). Chomsky (2000) proposes the mechanism of Agree as a further reformulation of Attract F, which just relates an item with a (typically uninterpretable) feature F (the 'probe') to the closest item with a matching F (the 'goal').

Another type of economy condition is the principle of Last Resort, which restricts the otherwise unbounded application of Move α/Merge to only those which are minimally necessary for the derivation to converge (i.e., receive appropriate interpretation) at LF and PF. For example, under the assumption that A-movement is generally driven by the need of an NP to receive abstract Case (Nominative, Accusative, Genitive, etc., which are overtly manifested in, e.g., English pronominal systems, as in *he*, *him*, *his*, etc.) assigned to a relevant A-position (subject, object, possessor, etc.), A-movement from one case position to another case position is superfluous and thus prohibited by Last Resort.

(19) a. It seems that John got a Nobel prize last week.

b. **John*$_i$ seems that t_i got a Nobel prize last week.

c. **A Nobel prize*$_i$ seems that John got t_i last week.

Further considerations are attributed to the general principle of economy along the line with the Minimalist Program. For example, it has been argued that the cyclic derivation by phase also has an important function of periodically reducing memory loads of the active workspace, contributing to computational efficiency (Chomsky, 2000 *et seq*). Another economy condition is the principle of Full Interpretation (Chomsky, 1986a, 1995, 2000, 2004; see also Narita, 2014 for discussion), which requires that every element of PF and LF must receive an appropriate interpretation, forcing all non-interpretable elements (e.g., agreement features) to get checked off and deleted in the course of derivation.

Covert Movement

We have seen a number of cases where movement is overt, with phonetically visible effects, while it has also been widely assumed that certain elements may move covertly as derivations enter LF, producing no phonetic effect. It was specifically assumed in the earlier Government-and-Binding framework that movement may be overt if it occurs before SS, or it may be covert if it occurs in the so-called LF-component (the mapping from SS to LF). Sentences containing quantifiers such as *everyone*, *some student*, *most people* exemplify a movement operation of the relevant sort, referred to as Quantifier Raising (QR). Chomsky (1976) argued that quantifiers are located covertly on the (left) periphery of the clause containing them, binding a variable in their surface position. Thus the surface sentence *John hates every politician* has the predicate-calculus style of representation *For every x*, *x a politician*, *John hates x*. May (1977) suggested that this representation may be obtained by an instance of Move α that raises the quantifier leftward, in a style reminiscent of 'quantifier lowering' rules proposed by generative semanticists in the early 1970s. Important arguments include generalizations concerning the possibilities of interpreting pronouns as 'bound variables' (Higginbotham, 1980).

Another sort of covert movement is that of *wh*-phrases in languages in which they typically occur in their 'original' positions. Unlike English-type languages, Chinese (and many other languages) leave their question words '*in situ*' when forming *wh*-questions. Thus *What did you buy*? in English is simply *ni mai-le shenme* 'you bought what' in (Mandarin) Chinese. Note that the Chinese sentence (20) is ambiguous between a direct and indirect question reading, corresponding to two distinct forms in English (21):

(20) ni Jide Zhangsan mai-le shenme

you remember Zhangsan bought what

(21) a. *What*$_i$ did you remember that John bought t_i? (direct question)

b. You remembered *what*$_i$ John bought t_i. (indirect question)

Despite the overt differences, Huang (1982) argued that the two languages (and all languages) have essentially the same LF representations. In particular, although *wh*-phrases do not move overtly in Chinese, they do so covertly in LF. The ambiguity of (20) follows as a result of the fact that *shenme* 'what' may be moved to the subordinate clause below *jide* 'remember' or to the main clause, deriving LF representations not unlike (21a,b). In addition to capturing certain crosslinguistic generalizations, this hypothesis is supported by the fact that interpretation of *in situ wh*-questions is subject to some restrictions that govern overt movement, e.g., the ECP.

As an instance of Move α, covert movement was not limited to A′-movement like QR or *wh*-movement. Given the *wh*-movement parameter (some languages move *wh*-phrases overtly, others do it in LF), proposals exist in the current literature that both A-movement (e.g., object shift) and head-movement (e.g., V-to-I or I-to-C raising) may occur overtly in some languages but covertly in others. However, the status of the covert LF-component is controversial in the current research under the Minimalist

Program, partially because intermediate levels of representations such as DS and SS got eliminated, and also because the post-SS instances of Move α were not obviously reducible to bottom-up (hence strictly cyclic) applications of Internal Merge. Therefore, a number of researchers currently investigate a reductive approach to covert movement, where the effect of covert LF-movement is relegated either to instances of Agree (feature-matching) without movement (Chomsky, 2000), or to PF deletion of higher (instead of lower) copies of Internal Merge (e.g., by postponing applications of Internal Merge until after Spell-Out; see Nissenbaum, 2000).

Conclusion

The study of movement in syntactic theory has been proven to be a productive probe into the inner workings of the human language faculty. Many discoveries were made and questions addressed since the very beginning of generative grammar: the typology of movement operations (A-, A′-, and head-movement), various constraints on extraction from and of certain structures, the parametric existence of covert movement and its possible reformulations, and various considerations on minimality and computational economy. In current Minimalist inquiries, ever more emphasis is put to the role of Movement/Internal Merge, with new consequences that will continue to intrigue future researchers.

See also: Generative Grammar; Linguistics: Incorporation; Syntax (General); Syntax–Semantics Interface; Valency and Argument Structure in Syntax.

Bibliography

Chomsky, N., 1955(1975). The Logical Structure of Linguistic Theory. Harvard University and MIT; published in part by Plenum Press, New York.
Chomsky, N., 1957. Syntactic Structures. Mouton, The Hague.
Chomsky, N., 1964. Current Issues in Linguistic Theory. Mouton, The Hague.
Chomsky, N., 1973. Conditions on transformations. In: Anderson, S., Kiparsky, P. (Eds.), A Festschrift for Morris Halle. Holt, Rinehart, and Winston, New York, pp. 232–286.
Chomsky, N., 1975. Reflections on Language. Pantheon Books, New York.
Chomsky, N., 1976. Essays on Form and Interpretation. Elsevier, Amsterdam.
Chomsky, N., 1981. Lectures on Government and Binding. Foris, Dordrecht.
Chomsky, N., 1986a. Knowledge of Language: Its Nature, Origin and Use. Praeger, New York.
Chomsky, N., 1986b. Barriers. MIT Press, Cambridge, MA.
Chomsky, N., 1995. The Minimalist Program. MIT Press, Cambridge, MA.
Chomsky, N., 2000. Minimalist inquiries: The framework. In: Martin, R., Michaels, D., Uriagereka, J. (Eds.), Step by Step: Essays on Minimalist Syntax in Honor of Howard Lasnik. MIT Press, Cambridge, MA, pp. 89–155.
Chomsky, N., 2001. Derivation by phase. In: Kenstowicz, M. (Ed.), Ken Hale: A Life in Language. MIT Press, Cambridge, MA, pp. 1–52.
Chomsky, N., 2004. Beyond explanatory adequacy. In: Belletti, A. (Ed.), Structures and beyond: The Cartography of Syntactic Structures. Oxford University Press, New York, pp. 104–131.
Chomsky, N., 2007. Approaching UG from below. In: Sauerland, U., Gärtner, H.-M. (Eds.), Interfaces + Recursion = Language? Chomsky's Minimalism and the View from Semantics. Mouton de Gruyter, Berlin and New York, pp. 1–29.
Chomsky, N., 2008. On phases. In: Freidin, R., Otero, C., Zubizarreta, M.L. (Eds.), Foundational Issues in Linguistic Theory. MIT Press, Cambridge, MA, pp. 133–166.
Higginbotham, J., 1980. Pronouns and bound variables. Linguistic Inquiry 11, 679–708.
Huang, C-T.J., 1982(1998). Logical Relations in Chinese and the Theory of Grammar. Ph.D. dissertation edited and published by Garland Publishing, MIT, New York.
Huang, C.-T.J., 1995. Logical form. In: Webelhuth, G. (Ed.), Government and Binding Theory and the Minimalist Program. Blackwell, New York, pp. 125–175.
Kayne, R., 1981. ECP extensions. Linguistic Inquiry 12, 93–133.
Kayne, R., 1994. The Antisymmetry of Syntax. MIT Press, Cambridge, MA.
May, R., 1977. The Grammar of Quantification. Ph.D. dissertation. MIT.
Narita, H., 2014. Endocentric Structuring of Projection-Free Syntax. John Benjamins, Amsterdam.
Nissenbaum, J., 2000. Investigations of Covert Phrase Movement. Ph.D. Dissertation. MIT.
Rizzi, L., 1990. Relativized Minimality. MIT Press, Cambridge, MA.
Ross, J.R., 1967. Constraints on Variables in Syntax. Ph.D. dissertation. MIT.
Uriagereka, J., 1999. Multiple spell-out. In: Epstein, S.D., Hornstein, N. (Eds.), Working Minimalism. MIT Press, Cambridge, MA, pp. 251–282.

Movements, Social

David A Snow and Anna E Tan, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by L. Tilly, volume 21, pp. 14360–14365, © 2001, Elsevier Ltd.

Abstract

Social movements are important agents of social, political, and cultural change, dating back from the present to at least the far-reaching empires of the Greco-Roman world and the various rebellions and religious movements spawned during that era. This article sketches a general overview of social movements, accenting their role as carriers of social change, providing a working conceptualization, elaborating the varieties or types of social movements, and identifying the focal issues and questions that animate most research on social movements.

Social movements and the associated phenomena they often organize and sponsor, ranging from peaceful protest demonstrations to acts of political violence, from pamphleteering to revolution, and from mass vigils memorializing deceased constituents to boisterous gatherings clamoring for retribution, are collective actions through which aggrieved collectivities give voice publicly to various grievances and press relevant authorities to attend to the associated claims and/or demands. The grievances generally range from political and religious oppression to socioeconomic inequalities to procedural and/or rights-based inequalities to all kinds of lifestyle concerns and perceived social and environmental harms. Although there are other institutionalized forums in which collectivities can express their grievances and concerns, particularly in democratic societies, social movements have functioned as an important vehicle for articulating and pressing an aggrieved collectivity's interests and claims for much of human history. Certainly many of the most significant events and changes throughout the course of human history – such as the rise and spread of Christianity and Islam, the Reformation, and the French, American, Russian, and Chinese Communist Revolutions, for example – cannot be fully understood apart from the operation and influence of social movements. Similarly, many of the world's most influential and famous figures – such as Jesus Christ, the Apostle Paul, Muhammad, Martin Luther, Valdimir Lenin, Mohandas Gandhi, Mao Se Tung, Martin Luther King Jr., and Nelsen Mandela – gained their celebrity and influence, in part, from the movements with which they were associated and/or lead. More generally, one would be hard pressed to identify a drastic shift in any group's social standing, historical or contemporary, that occurred without some kind of organized contention precipitating it. And today, it is difficult to name a contested social issue throughout much of the world with which a social movement is not associated on at least one side of the issue. Whatever the issue – abortion, animal rights, civil rights, human rights, democratization, climate change, gender equality, gun control, immigration, labor and unions, nuclear weapons, religious freedom, war, water conservation and control, and world poverty – one typically finds one or more associated social movements. Whether we live in a 'movement society' or world, as some scholars have suggested (Meyer and Tarrow, 1998), there is little question but that social movements are important social phenomena and that our understanding of the social world, both past and present, is partly contingent on our understanding of social movements and the activities and events with which they are associated.

Conceptualization

As with many concepts in the social sciences, there continues to be little consensus as to how best to define social movements. Some scholars emphasize the strategic, rational, and organized character of social movements (McCarthy and Zald, 1977); others focus on the networked and coalitional character of movements (Della Porta and Diani, 2006); some accent the ideological element of movements (Zald, 2000); still others emphasize the 'manifestly political' character of movements as a form of contentious politics with one or more governments as a claimant or target of claims (McAdam et al., 2001); and some scholars argue for a more inclusive and elastic conceptualization, one that conceives of movements broadly as collective challenges to institutional, organizational, and cultural domains other than just the state or the polity (Snow, 2004; Snow and Soule, 2010). However, one does not have to choose one emphasis or focus over another so long as it is recognized that each conceptualization accents a particular dimension or aspect of social movements. Thus, whatever dimensions of movements a scholar might accent, it can be argued that all social movements share a number of characteristics that together define the phenomenon. First, all movements are change oriented in the sense that they seek or oppose change, although there can be considerable variation in the level and degree of change sought; second, movements are, as a consequence of this change orientation, challengers to or defenders of existing institutional structures or systems of authority, be they political, religious, corporate, educational, or cultural; third, movements are collective rather than individual enterprises; fourth, they act outside of existing institutional or organizational arrangements, although in varying degrees with some movements employing only extrainstitutional tactics and others working within institutional channels while also engaging in some extrainstitutional collective action; fifth, they operate with some degree of organization, ranging from a single social movement organization (SMO) to a network or

coalition of movement organizations that may vary in the degree to which they are tightly or loosely coupled; and sixth, social movements typically display some degree of temporal continuity – they are relatively episodic but rarely fly-by-night fads that are here today and gone tomorrow (see Snow and Soule, 2010).

Varieties of Movements

Although all social movements can be defined in part by the above characteristics, they can also be differentiated in terms of these characteristics. It is worthwhile to consider that even so-called reactionary or conservative movements opposing change must advance their own challenges to authority in responding to real or perceived currents of social change. Therefore, the most common axis of such differentiation concerns the degree or amount of change pursued, which makes sense since the promotion of or resistance to change is the raison de' etra of virtually all social movements. The most basic distinction in this regard is between *reform* and *revolutionary movements*. A more subtle distinction is provided by Smelser's (1962) differentiation between norm-oriented and value-oriented movements. *Norm-oriented movements* are said to seek relatively limited but specific system changes, mainly with respect to rights and rules of access and participation in the various societal institutions. Examples might include movements that have sought to change labor laws, decriminalize marijuana, expand or restrict immigrant rights, and criminalize or decriminalize abortion. *Value-oriented movements*, in contrast, seek more fundamental changes in cultural values and institutional structures and practices, as with movements that seek to redefine the fundamental rights and privileges of personhood and citizenship, like the Gandhi-led Indian Independence Movement and the African-American Civil Rights Movement, or those that have sought to recalibrate the very structure of political governance and association and the values on which they are based, as with the French and American revolutions.

Wallis (1984) provided a similar typology with his distinction between world-rejecting and world-affirming movements. Although developed to capture differences mainly among religious movements, this dichotomy parallels the norm- and value-oriented distinctions. The world-rejecting movement, like the value-oriented movement, condemns the dominant social order, including both its values and institutional arrangements, as in the case of Jim Jones' Peoples Temple that culminated in the mass suicide of its members. World-rejecting movements are not confined to the realm of new religious movements; both the peaceful separatist anarcho-primitivists within the Green Anarchism movement and the anti-industrial agenda of the 'Unabomber' Ted Kaczynski are characterized by their rejection of the existing social order. The world-affirming movement, like the norm-oriented movement, is less contemptuous of the prevailing religious or social order and thus seeks more selective changes in specific segments of that enveloping order, as with organizations such as the People for the Ethical Treatment of Animals or the variety of organizations within the predominately Christian anti-abortion movement. While world-affirming movements may hold forward-looking and/or radical views, their objectives still operate on the assumption that society is redeemable and equipped with the tools for change.

By focusing on the amount or degree of change sought, the above schemes provide one basis for differentiating among movements, but they are incomplete because they neglect the fact that change can have a different focus or occur at different levels. This oversight was addressed by the anthropologist Aberle (1966) in his research on the Peyote religion among the Navaho Indians in the southwest region of the United States. He differentiated among movements based on both the degree of change and the locus or level of change sought, directing attention to the target of change that can vary from the individual level to some aspect of the broader structure. Based on the cross-classification of these two dimensions, Aberle derived four generic types of movements. *Alterative movements* seek partial change in individuals as in the case of the therapeutic and self-help movements that proliferated in the United States in the 1970s and 1980s. *Redemptive movements* seek a more complete, or total rather than partial, change among individuals, which was the objective of various religious movements that flowered in the 1970s, such as the Hare Krishna and Unification Church (Moonies) movements. Characteristics of redemptive movements can also be observed in the strictures of more recent lifestyle-centered movements, such as the vehement rejection of all intoxicants and so-called poisons by adherents of the straight-edge lifestyle. *Reformative movements* seek limited but focused changes in the social system like the normative and world-affirming movements noted above; and *transformative movements*, are akin to value-oriented and revolutionary movements.

An alternative approach to differentiating social movements takes a more historical approach, contrasting movements clustered in different historical areas. Differences in the amount and locus of change sought are not ignored, but the crucial criterion is the affinity between certain historical trends or changes and movement goals and/or strategies and tactics. Perhaps the earliest and most cogent illustration of this approach is found in Mannheim's *Utopia and Ideology* (1946), wherein he identifies four major waves of movements in so-called modern times: *chiliastic movements* expressed in some religious movements and peasant revolts; *liberal humanitarian movements* associated most dramatically with the French and American revolutions; *conservative movements* expressed must fully in fascism and Nazism; and *socialist movements* associated mainly with international communism but also influencing labor movements in the United States, Europe, and elsewhere. Extending Mannheim's historical approach, Turner (1969: pp. 391–395) argued that "major eras in history have differed in the dominant sense of injustice which underlay the major movements of the time and dictated the main direction of social change," contending that both the liberal humanitarian and socialist conceptions of injustice have been exhausted and that "a new revision is in the making." Foreshadowing aspects of new social movement theory, Turner argued that the emerging change preference concerned the yearning for a satisfactory sense of personal worth, dignity, and identity. New social movement theory emerged in the 1980s in Europe to analyze the so-called *new social movements* that began to surface in the 1960s and which were seen as being different

from the older, materially based working-class and labor movements associated initially with Marxism. In contrast, the new social movements – such as the environmental movement, human rights movement, and gay, lesbian, bisexual, and transgender movement – are driven by concerns with procedural rights and various lifestyle issues (Buechler, 2011: pp. 157–173; Kriesi et al., 1995; Larana et al., 1994).

Also acknowledging the link between varieties of movements and historical change, Tilly (1993, 2006) advanced the concept of *repertoires of contention*, identifying particular strategies and tactics for political claims making associated with particular historical and political periods. In particular, Tilly distinguished between an older repertoire associated with the eighteenth century and a newer repertoire associated with the nineteenth century and the development of the nation-state. Tarrow summarizes the distinction as such: The newer repertoire "was cosmopolitan rather than parochial; autonomous rather than dependent on inherited rituals or occasions; and modular rather than particular" (1998: p. 37). "Unlike the older forms, which expressed people's immediate grievances directly against antagonists, the newer repertoire … could be used by a variety of social actors on behalf of a number of different claims and serve as a bridge among them to strengthen their hand and reflect broader and more proactive demands" (Tarrow, 1998: pp. 66–67). However, these repertoires may not simply pass out of existence with the arrival of a new era of contention. This ongoing theme of historical variation in movement characteristics is seen even more recently in the work of scholars who have suggested that the rise of the Internet communication and virtual outlets for activism has prompted the emergence of a new 'digital' repertoire of contention (Earl and Kimport, 2011).

Following Tilly and Tarrow, much work on social movements has focused on state-centered social movements. However, it is important to note that movements are not solely engaged with political institutions but may rather emerge from a host of other institutional interactions including those with religious institutions and other centers of authorities. Young's research (2006), for example, shows that the first sustained national movements in the United States arose in the 1820s and 1830s in response to religious forces, rather than political institutions, calling Americans to bear witness against the sins of intemperance, slavery, and other moral wrongs. These nonstate institutional engagements have persisted throughout history and generated no lack of movements for their own part, as can be seen in the institutional protest seen within the Catholic Church at various stages of its history and the ongoing struggles of women within various institutional sectors (Katzenstein, 1999).

A final basis for differentiating among movements relates to strategic and/or tactical differences. Perhaps the foremost example is the distinction between violent and nonviolent movements. Although the use of violence may occur intermittently in a movement's career, some movements are sometimes defined in part by their strategic use of violence, as with terrorist movements at both poles of the political spectrum, or nonviolence, as with the US Civil Rights Movement and the pacifist Plowshares nuclear disarmament movement (Nepstad, 2008). Strategic or tactical actions are also pivotal in McAdam et al. (2001) distinction between *constrained contention* and *transgressive contention*. In the case of the former, "well established means of claims making are employed" (p. 7), which is to say that they are institutionalized or routinized; whereas in the case of transgressive contention, some parties embroiled in the contention 'employ innovative collective action' – that is, the claims or means are "either unprecedented or forbidden within the regime in question" (p. 8). The selection of tactics and formation of strategy is neither solely a response to a movement's environment nor dictated wholesale by a movement's ideology but rather emerges from a dynamic interplay between multiple factors. Although a movement's tactics will generally accord with its ideological underpinnings, as is the case of the nonviolent sabotage of nuclear missiles undertaken by the pacifist Plowshares movement, SMOs have proven oftentimes to be highly adaptable and innovative in their strategies. Moreover, a movement cannot be characterized as violent or nonviolent, or constrained vs transgressive, in its entirety based upon tactics observed in a sample of protest events alone, as movements will often deploy a broad spectrum of tactics within its various flanks.

Clearly there are a number of alternative ways of differentiating and typologizing social movements. Although much of the extant literature on social movements has focused on left and progressive movements, and/or movements in predominately Western democratic settings, scholars are increasingly attentive to movements on right, far right, and outside of the field of conventional left-wing movements for which earlier conceptualizations may be insufficient (Blee and Creasap, 2010; Hirsch-Hoffer and Mudde, 2013; McVeigh, 2009). What is important is to be cognizant of the different axis of differentiation and to not assume that the conditions that account for the emergence and operation of movements of one type necessarily apply to all types of movements. As well, it is important to keep in mind that all of these theoretical typologies of movements are ideal types of sorts and therefore are likely to be rare, given the significant permeability between the various categories.

Focal Analytic Issues and Questions

As the study of social movements has evolved to the present, most inquires and research are guided by a variant of one or more sets of four focal questions. First, there is the long-standing concern with conditions underlying the *emergence* of social movements. Here there are two associated questions: Since movement adherents are generally stirred to action by a combination of grievances and a sense of efficacy, what are the processes and mechanisms that account for the generation of these mobilizing sentiments? And, since the occurrence of social movements is nested temporally and spatially rather than randomly distributed across time and place (Koopmans, 2004; Tarrow, 1998), what specifically are the sets of contextual conditions that facilitate or constrain the emergence and flourishing of social movements? Early lineages of social movement research, organized under the broader heading of collective behavior, emphasize the role of strains, relative deprivation or structural grievances as generative of

contentious collective action (see Buechler, 2011: pp. 9–106). Although the rise of the political opportunity perspective (see McAdam et al., 1996) has consequently shifted a great deal of scholarly attention to the role of contextual favorability within a movement's political setting, the importance of grievance-generating conditions in movement emergence cannot be overlooked.

The second orienting question concerns the issue of micro-mobilization or *recruitment and participation processes*. Here there is a subset of questions as well: What is the character of movement participation? What are the different ways of participating? And what are the determinants of who participates in social movement activities, and why? Why do only certain individuals come to participate in movement activity while other similarly situated individuals remain on the sidelines? Among the many conditions necessary for participation in a range of movement activities, a rare consensus emerges that one of the single most critical precipitating condition is simply being asked (Schussman and Soule, 2005). Within the wealth of studies attending to this question, scholars have uncovered various permutations of conditions conducive to participation, ranging from the resonance of a movement's frames with a target constituency (Benford and Snow, 2000), affective networks, and relationships as powerful recruitment mechanisms (Diani, 2004) to the spatial ecology of a setting, as observed in the large-scale mobilization of university students in the Tiananmen pro-democracy protests of 1989 (Zhao, 1998).

The third focal question concerns the *dynamics* of social movements. What happens once a movement has surfaced and it is, so to speak, up and running? More concretely, how are social movements organized and led, and how do they go about the business of strategically pressing their claims and dealing with the various relevant actors within their field of operation? What or who are the relevant actors that constitute the movement field of action, and what are their relationships? And what are the various mechanisms and processes associated with the operation and functioning of social movements? Such concerns strike at a broad range of possible processes and conditions within which social movements are enmeshed, necessitating refinement when posing such analytical questions. Generally, scholars of social movements have directed their emphasis variously upon internal dynamics, such as the importance of leadership (Morris and Staggenborg, 2004), mechanisms to sustain participation and persistence over time (Taylor, 1989; Nepstad, 2008), organizational form (Clemens and Minkoff, 2004), and finally SMO resources, as evidenced by the highly generative theory of resource mobilization (McCarthy and Zald, 1977). Among the many dynamics studied, of particularly recent attention has been the study of tactics and strategies. Just as movement organizations may occupy a range of niches within a given social setting, the tactics they deploy will range from highly institutionalized (constrained) to the most transgressive or extrainstitutional. Moreover, movements do not simply settle upon one approach and maintain it for the duration of its activities. As discussed earlier, tactical selection and strategy formation is subject to a host of both internal and external processes, operating at multiple levels within the movement.

The fourth general, focal issue concerns what many scholars and social movement activists consider as the bottom-line question: What difference do movements make, and for whom and in what ways? Do social movements play an important role in challenging authorities and generating impending social change? Are there other functions that they perform? What, in short, are the *consequences or outcomes* of social movements? A number of scholars characterize movements as being both effective and important actors that can be instrumental to lasting social change. Others address movements more cautiously and temper conclusions as to their effects, viewing them as generally ineffective in affecting long-term change or that their influence is predominately a function of already conducive political and social climates as well as favorable political administrations. More recent conceptualizations have expounded upon narrower definitions of movement success and instead focused on a range of outcomes, such as long-time policy feedbacks, cultural consequences, and biographical consequences for the individual participant (Amenta et al., 2010; Giugni et al., 1999; Snow et al., 2013).

Answers to these questions understandably draw on a range of overlapping perspectives and processes, including political process theory, resource mobilization theory, ecological and spatial perspectives, framing processes and theory, cultural perspectives, social psychology, as well as other perspectives and points of view.

See also: Alienation, Sociology of; Charisma, Social Aspects of; Civil Rights Movements; Class Consciousness; Collective Behavior, Sociology of; Collective Identity; Gay and Lesbian Movements; Media and Social Movements; New Religious Movements; Political Protest and Civil Disobedience; Political Sociology; Revolution, Theories of; Revolutions, Sociology of; Social Movements and Political Violence; Youth Movements.

Bibliography

Aberle, D., 1966. The Peyote Religion among the Navaho. Aldine, Chicago.

Amenta, E., Caren, N., Chiarello, E., Su, Y., 2010. The political consequences of social movements. Annual Review of Sociology 36, 287–307.

Benford, R.D., Snow, D.A., 2000. Framing processes and social movements: an overview and assessment. Annual Review of Sociology 26, 611–639.

Blee, K.M., Creasap, K.A., 2010. Conservative and right-wing movements. Annual Review of Sociology 36, 269–286.

Buechler, S., 2011. Understanding Social Movements: Theories from the Classical Era to the Present. Paradigm Publishers, Boulder, CO.

Clemens, E.S., Minkoff, D.C., 2004. Beyond the iron law: rethinking the place of organizations in social movement research. In: Snow, D.A., Soule, S.A., Kriesi, H. (Eds.), The Blackwell Companion to Social Movements. Blackwell Publishing, Oxford, UK, pp. 155–170.

Della Porta, D., Diani, M., 2006. Social Movements: An Introduction. Blackwell Publishing, Oxford, UK.

Diani, M., 2004. Networks and participation. In: Snow, D.A., Soule, S.A., Kriesi, H. (Eds.), The Blackwell Companion to Social Movements. Blackwell Publishing, Oxford, UK, pp. 339–359.

Earl, J., Kimport, K., 2011. Digitally Enabled Social Change: Activism in the Internet Age. MIT Press, Boston, MA.

Giugni, M.G., McAdam, D., Tilly, C. (Eds.), 1999. How Social Movements Matter. University of Minnesota Press, Minneapolis, MN.

Hirsch-Hoefler, Sivan, Mudde, Cass, 2013. Right-wing movements. In: Snow, D.A., Della Porta, D., Klandermans, B., McAdam, D. (Eds.), The Wiley-Blackwell Encyclopedia of Social Movements. Wiley-Blackwell, Malden, MA, pp. 1116–1124.

Katzenstein, M.F., 1999. Faithful and Fearless: Moving Feminist Protest inside the Church and Military. Princeton University Press, Princeton, NJ.

Kriesi, H., Koopmans, R., Duyvendak, J.W., Giugni, M., 1995. New Social Movements in Western Europe: A Comparative Analysis. University of Minnesota Press, Minneapolis.
Koopmans, R., 2004. Protest in time and space: the evolution of waves of contention. In: Snow, D.A., Soule, S.A., Kriesi, H. (Eds.), The Blackwell Companion to Social Movements. Blackwell Publishing, Oxford, UK, pp. 19–46.
Larana, E., Johnson, H., Gusfield, J.R. (Eds.), 1994. New Social Movements: From Ideology to Identity. Temple University Press, Philadelphia.
Mannheim, K., 1946. Ideology and Utopia (L. Wirth, E. Shils, Trans.). Harcourt, Brace and Co., New York.
McAdam, D., McCarthy, J.D., Zald, M.N., 1996. Comparative Perspectives on Social Movements: Political Opportunities, Mobilizing Structures, and Cultural Framings. Cambridge University Press, New York.
McAdam, D., Tarrow, S., Tilly, C., 2001. Dynamics of Contention. Cambridge University Press, Cambridge, UK.
McCarthy, J., Zald, M., 1977. Resource mobilization and social movements: a partial theory. American Journal of Sociology 82, 1212–1241.
McVeigh, R., 2009. The Rise of the Ku Klux Klan: Right-Wing Movements and National Politics. University of Minnesota Press, Minneapolis, MN.
Meyer, D.S., Tarrow (Eds.), 1998. The Social Movement Society: Contentious Politics for a New Century. Roman & Littlefield Publishers, Boulder, CO.
Morris, A., Staggenborg, S., 2004. Beyond the Iron Law: Rethinking the Place of Organizations in Social Movement Research. The Blackwell Companion to Social Movements. pp. 171–196.
Nepstad, S.E., 2008. Religion and War Resistance in the Plowshares Movement. Cambridge University Press, Cambridge, UK.
Schussman, A., Soule, S.A., 2005. Process and protest: accounting for individual protest participation. Social Forces 84, 1083–1108.
Smelser, N., 1962. Theory of Collective Behavior. The Free Press, New York.
Snow, D.A., 2004. Social movements as challenges to authority: resistance to an emerging conceptual hegemony. In: Meyers, D.J., Cress, D.M. (Eds.), Authority in Contention: Research in Social Movements, Conflict, and Change. Elsevier, London/New York, pp. 3–25.
Snow, D.A., Soule, S.A., 2010. A Primer on Social Movements. W.W. Norton & Company, New York.
Snow, D.A., Tan, A.E., Owens, P.B., 2013. Social movements, framing processes, and cultural revitalization and fabrication. Mobilization 18, 225–242.
Tarrow, S., 1998. Power in Movement: Social Movements and Contentious Politics, second ed. Cambridge University Press, Cambridge, UK.
Taylor, V., 1989. Social movement continuity: the women's movement in abeyance. American Sociological Review 54, 761–775.
Tilly, C., 1993. Contentious repertoires in Great Britain, 1758–1834. Social Science History 17, 253–280.
Tilly, C., 2006. Regimes and Repertoires. University of Chicago Press, Chicago.
Turner, R.H., 1969. The theme of contemporary social movements. British Journal of Sociology 20, 390–405.
Wallis, R., 1984. The Elementary Forms of the New Religious Life. Routledge & Kegan Paul, London.
Young, M.P., 2006. Bearing Witness against Sin: The Evangelical Birth of the American Social Movement. The University of Chicago Press, Chicago.
Zald, M.N., 2000. Ideological structured action: an enlarged agenda for social movement research. Mobilization: An International Journal 5, 1–16.
Zhao, D., 1998. Ecologies of social movements: student mobilization during the 1989 prodemocracy movement in Beijing. American Journal of Sociology 103, 1493–1529.

Muller, Hermann Joseph (1890–1967)

Elof A Carlson, Stony Brook, NY, USA; and Institute for Advanced Study, Indiana University, Bloomington, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

H.J. Muller, founder of the field of radiation genetics, won the Nobel Prize in Medicine and Physiology in 1946 for having demonstrated that he could induce mutations with X-rays in fruit flies. Muller was also a contributor to classical genetics in T.H. Morgan's 'Fly Lab.' Unusual for his generation, he left the ivory tower and participated in the social implications of science. He denounced the American eugenics movement as elitist, biased, racist, and sexist but advocated a positive eugenics based on talent, intelligence, leadership, and empathy, which he hoped would be carried out in states favoring social equality. He was sympathetic to the Bolshevik Revolution and edited an underground procommunist newspaper, *the Spark*. He left the United States for Berlin, then Leningrad and Moscow, then fled the USSR during the Lysenko controversy and fought in the Spanish Civil War. He worked at Edinburgh until World War II broke out and then returned to the United States, finding a permanent home in Indiana University. Muller considered himself a reductionist in science, a socialist in political outlook, a critic of totalitarianism in all its forms, and a champion of radiation protection for the public.

Hermann Joseph Muller was born in New York City on 21 December 1890 and he died in Indianapolis, Indiana, on 5 April 1967. He is best known as an American geneticist, the founder of the field of radiation genetics, for which he received the Nobel Prize in Medicine or Physiology in 1946. He was also a controversial critic of the way science is abused by society and he was both a critic and proponent of eugenics (Carlson, 1982). Muller's paternal ancestors came to the United States after the 1848 Revolution. They had lived in the Coblenz area of the Rhine and were Catholics. Muller's father became a Unitarian and Muller abandoned all religions for secular humanism and considered himself an atheist from his high school years on. Muller grew up in New York City and attended Morris High School in the Bronx. His father was co-owner of a metal art shop that produced frames and plaques. From those first Muller settlers in the United States came first cousins Alfred Kroeber, a noted anthropologist and Herbert J. Muller, an English critic known for his best-seller, *The Uses of the Past*. Muller attributed his early interest in evolution and socialism to his father, who died when young Muller was 10 years old. Muller's maternal side, the Lyons family, was of mixed Anglican and Jewish ancestry before they emigrated from England to the United States. Muller excelled in high school and was awarded a scholarship to attend Columbia University.

Founding the Field of Radiation Genetics

The field of radiation genetics emerged after 1927 when Muller published his first articles on the artificial induction of mutations by X-rays in fruit flies (Muller, 1927). Within a few years, geneticists around the world were using X-rays to induce mutations for both basic genetic research and for applied uses, especially new mutations of commercial importance in agriculture. One of the most significant applications was the production and selection for increased yield of penicillin from the fungus *Penicillium* in the later 1940s. Although Muller did virtually all his research using fruit flies, he made applications of that knowledge to working out the major features of classical genetics and to human genetics, evolution, and fundamental findings such as the mechanism that caused radiation sickness among those exposed to high doses of radiation at Hiroshima and Nagasaki from atomic bomb radiation release.

Muller found the courses of Edmund Beecher Wilson particularly influential in introducing him to the chromosome theory of heredity and the relatively new field of Mendelian genetics. He established a biology club with fellow undergraduates Alfred Henry Sturtevant and Calvin Blackman Bridges. They joined Thomas Hunt Morgan's laboratory as it shifted from embryology to genetics using fruit flies. The 'Fly Lab' as Sturtevant called it, established many of the basic tenets of classical genetics, including the finding of X-linked inheritance, the discovery of crossing over among genes on a shared pair of chromosomes, the occasional gain or loss of a chromosome through nondisjunction, and, of particular interest to Muller, the relation of genes to character expression (Carlson, 2007). Muller completed his PhD in 1916 and took a job in Julian Huxley's new Department of Biology at Rice University in Texas. He chaired the Department when Huxley returned to Great Britain to serve in World War I. Huxley and Muller remained life-long friends, sharing an interest in positive eugenics, the differential breeding that favored intelligence, good health, qualities of leadership, and an altruistic empathetic personality.

From 1916 to 1926, Muller developed a theoretical genetics based on the findings of the Fly Lab. He noted that genetic changes involved different processes, some associated with the number and integrity of the chromosomes and some with the individual gene. Muller argued that "variation due to change in the individual gene" was the primary mechanism of Darwinian evolution by natural selection (Muller, 1922). He also argued that the gene was the basis of life because it was a molecule capable of copying itself and its errors, making evolution possible (Muller, 1936a). His focus on the individual gene led him to study the time of origin of mutations and how they were expressed on first appearance and how frequently they were transmitted or arose by spontaneous mutation. This in turn led him to believe the induction of mutations was possible. He began a series of experiments, one using temperature, and the other using X-rays, to show mutations could be artificially induced. What distinguished Muller's success from earlier failures by Morgan and others was his design of genetic stocks

that would detect mutations objectively and quantitatively. In 1927, he published his findings in *Science* and presented his data in detail at the International Congress of Genetics in Berlin.

Muller's Contributions to Classical Genetics and Evolution

Muller's first contribution to evolutionary theory (1918–21) was his study of two mutations, beaded and truncate wings (Muller, 1918; Altenburg and Muller, 1920). In each case, he showed the existence of a chief gene, modifier genes that either boosted or diminished the expression of the wing trait, and environmental factors, such as temperature on the expression. Muller isolated the modifier genes and mapped them and he could combine different modifiers with the chief gene and predict the number of quantitative distributions of wing deformity based on these components. He used this as strong evidence favoring Darwin's thesis of natural selection. His second major contribution was the 'gene doctrine.' In 1936, he studied bar eyes in fruit flies and showed that it was associated with a duplicated tandem segment that arose from unequal crossing over. Muller argued that chromosome length and number could be interpreted through this mechanism and that all genes arise from preexisting genes. Muller believed the gene was the basis of life because it was a molecule that could copy its errors, a feature he believed essential for life to evolve. Muller's third major contribution to evolution in 1932 was his identification of genetic recombination as the basis for the emergence of sex (Muller, 1932). His fourth contribution is called Muller's ratchet. Muller argued as early as 1918 that if a chromosome is prevented from recombination through crossing over, it will accumulate mutations and deletions of its genes. He used this to explain the evolution of the Y chromosome and why it is virtually void of genes except for those associated with sex determination or spermatogenesis.

Muller's Social Views of Genetics: The Eugenics Movement

Muller was a socialist from his high school days. He shifted to bolshevism after the Soviet Union was established. In this role, he favored the social reforms advocated in magazines like *The Masses*. He felt capitalism was unjust and neglected or opposed equal opportunities for women, minorities, and the laboring classes. At the same time in the United States, there was a robust eugenics movement founded by a committee of the American Breeder's Association, chaired by David Starr Jordan and backed by Alexander Graham Bell. They established a eugenics station at Cold Spring Harbor with Charles Davenport as its director. The eugenics outlook favored by this center was called negative eugenics because it stressed the culling of undesirable traits from the American population. They used a multiple approach of institutionalizing the mentally retarded and insane; they backed state compulsory sterilization laws especially by vasectomy and tubal ligation; and they favored marriage laws that restricted who could marry based on considerations of eugenics and race. They also favored restrictive immigration laws, especially against those from southern and central Europe (especially Jews, Poles, Slavs, Serbs, Italians, and those from Balkan nations) who were the majority of immigrants coming to the United States between the 1890s and World War I (Carlson, 2001).

Muller opposed this negative eugenics movement and denounced it at the Third International Congress of Eugenics in 1932, which was held at the American Museum of Natural History. In his paper, 'The dominance of economics over eugenics' Muller accused the movement of being racist, sexist, and elitist, and showed how its basic assumptions were based on wishful thinking and bias rather than science. He felt that eugenics could only be put into practice, if the social conditions were equally favorable to all children, not just those favored by the capitalist values of the Eugenics Record Office. That utopian social outlook he claimed existed only in the USSR. Back at Texas, he was having difficulties with the FBI because he helped edit and distribute an underground unsigned student newspaper, *The Spark*, named for Lenin's *Iskra*. Some trustees wanted Muller investigated and tried by the academic senate for unprofessional behavior. *The Spark* condemned both American society and the state of Texas for its racism against African-Americans, for its failure to provide jobs for graduates because of the Depression (Hoover was still President), and for its hostility to the teaching of evolution in public schools.

Muller left the United States in 1932 and headed to Berlin on a Guggenheim Fellowship. Hitler came to power soon after and Muller was again forced to relocate, this time to the USSR where the head of Soviet agricultural science, N.I. Vavilov, invited him to be a guest investigator. While Muller denounced race hygiene in Germany and other countries in Europe, he promoted his enthusiasm for positive eugenics in a book, *Out of the Night*. Muller's eugenic idealism and socialism are reflected in this work (Muller, 1936a). He believed sperm banks could be used to provide those choosing a eugenic option for children and that the values associated with that choice should be high intelligence, unique valued talents, good health, mental stability, a gregarious personality, and a compassion or empathy for others. He shunned eugenic selection based on notoriety in popular culture and he felt sterilization and restrictive immigration were bad policies based chiefly on discrimination.

Muller's Social Views: The Defense of Science in the USSR

Like many left-wing Americans in the 1920s and early 1930s, Muller felt that the USSR was on the right track to establishing a democratic and just society. In the USSR, there was an English language newspaper (published in Moscow) for those who emigrated there from the United States and for those who were deported by the United States for subversive activity, especially organizing labor unions, condemning racism, or being members in anarchist or communist organizations. Muller established a laboratory first in Leningrad and then in Moscow with numerous graduate students, including two he brought with him from Texas. Muller devoted much of his research to cytogenetics, the study of genes in relation to chromosome changes, and it was here that his gene doctrine emerged. But in

1936, a new movement began in the USSR at the agricultural field station in Odessa. It was led by Trofim D. Lysenko and he called his movement Michurinism. It would be called Lysenkoism in the West. Lysenko was a plant physiologist and practical breeder who believed that he could shatter and retrain the heredity of cereal plants and transform them. His initial claim was that he could shift winter wheat into spring wheat by soaking the grains in ice water. The USSR had wretched weather for its crops because its collective farms were at high latitude and the growing season was short. This led to frequent crop failures. Lysenko promised a revolution in agriculture with his techniques. Soon, he had a sizable following and began challenging classical genetics as slow, pessimistic in its assessments, and reflective of suspect origins from a Moravian monk (Gregor Mendel) whose ideas were central to Nazi race hygiene. Muller found himself spending more of his time defending Vavilov and classical genetics than he desired. A showdown took place in December 1936 in which 3000 participants appeared in Moscow for a debate between Lysenkoists and Vavilov's classical geneticists. When Muller gave his talk he denounced Lysenko as a charlatan, with Michurinism indistinguishable in his views from pseudosciences like alchemy, numerology, and medical quackery or even shamanism. This created a storm of protest from the Lysenkoists and Vavilov told Muller he had better find a way to leave gracefully. Muller did so by joining the International Brigade to fight in the Spanish Civil War. He served in Norman Bethune's blood unit and did experiments to see if cadaver blood from freshly killed soldiers could be used for transfusions. After the siege of Madrid, he was forced to flee again and this time Huxley found him a job in Edinburgh at the Institute for Animal Genetics.

The outbreak of war in 1939 led to a lull in the debate between Lysenko and his detractors. After the war it erupted again in 1948 when Muller denounced Lysenkoism at the International Congress of Genetics in Stockholm. In Moscow, a second showdown debate was scheduled and Lysenko announced that his views were officially endorsed by the Central Committee of the Communist Party. One by one those who were engaged in classical generics studies recanted and pledged allegiance to Michurinism. Until the mid-1960s Lysenkoism became the official doctrine and practice of genetics in the USSR and the Eastern Bloc. It collapsed because Lysenko could not deliver on his promises and those farmers who tried to use Lysenko's techniques complained to their representatives that it did not work. After the collapse of the Soviet Union, Lysenko was revealed as a self-deceived fraud, who bullied farmers into falsifying documents that his techniques worked, by accusing them of sabotage if they failed to do so (Medvedev, 1969).

Muller Fails to Establish a Positive Eugenics Program

In 1957, Muller revived positive eugenics. He believed enough years had elapsed since the war ended and a revival of race hygiene or a movement like the American eugenics movement that Davenport had promoted was not likely to recur. Instead, he proposed what he called 'germinal choice.' It would not be the state or an institution tied to a genetic organization that determined the use of sperm in sperm banks but the users themselves. By allowing couples access to the records of accomplishments and medical records of the donors to sperm banks, Muller believed users would be as serious in their choices as they would for any major decision they were making about their lives, like buying a home, moving to a new location, or choosing a career. Muller was criticized for reviving eugenics by many of his colleagues and by the press, which carried stories of his speeches. They saw all eugenics as doomed to exploitation and leading to bias. Muller argued that our 'load of mutations' was increasing because natural selection was minimized as civilization and modern medicine replaced the ruthlessness of life some 10 000 or more years ago (Muller, 1950). The germ theory shifted infant mortality from 50% in the mid-nineteenth century to less than 1% in the mid-twentieth century in industrialized countries. The best way to circumvent this accumulation was by selective breeding based, he believed, on germinal choice. Muller found only one group interested in applying germinal choice. It was a private group in California calling itself the Foundation for Germinal Choice. Muller visited the sponsors and came away disillusioned because they were only interested in high intelligence and their criterion was the College Entrance Examination Board scores (a top score of 800 on the verbal or mathematical section). Muller felt the people backing this 'sperm bank for geniuses' shared the values of Davenport and his supporters. He forbade use of his name as an endorsement and told his wife that he would rather have no eugenics program than a bad one.

Muller's Concern for Radiation Protection

While many criticized Muller for returning to eugenics in the 1950s, they held him in esteem for raising concern about radiation protection (Muller, 1952). Muller first took an interest in this issue in 1927 when he induced mutations with X-rays in fruit flies. By the 1930s, he and those who were inspired by his work around the world had found that the induction of mutations was roughly linear. Thus a dose of 500 R would give one fourth of the mutations obtained at a dose of 2000 R. A second finding was that when Muller devised a genetic test to detect translocations (an exchange of pieces of nonhomologous chromosomes), they did not obey a linearity of dose to occurrence. Instead they rose exponentially to about the 3/2 power. From these findings Muller concluded that relatively low doses would be of concern for the gene mutations they produced but not for the chromosomal rearrangements they produced because at low doses these were virtually nonexistent. But gene mutations, if extrapolated to lower doses would be cumulative and Muller believed that those exposed chronically to low doses, like physicians, dentists, industrial workers using X-rays to examine for stress fractures in metal, shoe salesmen (in those days shoe size was done by fluoroscope in shoe stores), or chiropractors (who routinely examined spines for displacements) should protect themselves with lead shielding or move away from the radiation source (the dose diminishes rapidly with distance). Many dentists used their fingers to hold the film in the patient's mouth and many physicians stood unshielded by the fluoroscope when examining the patient's internal organs. Muller criticized those who tried to induce ovulation in infertile

women by X-raying their pituitary gland and their ovaries with relatively high doses (about 200 R where a typical chest X-ray then was about 0.1 R).

It was not until 1946 that the health professions began to listen to Muller. They realized the harm from radiation exposure at high doses to a population from descriptions of radiation sickness in Hiroshima and Nagasaki in 1945. After Muller was awarded a Nobel Prize for his work, he was sought as a speaker at universities and by medical groups. He argued that the linearity of dose to mutation production was based on both extrapolation and from studies of attenuated doses compared to acute doses. S.P. Ray-Chaudhuri did experiments while Muller was in Edinburgh that compared 400 R given in 30 min to fruit fly males and to a population of male flies that was exposed to almost a month of radiation that cumulatively equaled 400 R (Ray-Chaudhuri, 1944). This meant that in any 1 min they were receiving less than a chest X-ray's exposure. Yet both groups gave about 2% X-linked lethals, the standard for measurement using a quantitative test that Muller had worked out for his 1927 paper. Untreated control flies gave 1 X-linked lethal per 2000 X chromosomes examined.

Muller and his student, G. Pontecorvo, also determined in 1940 that X-rays induced a condition called 'dominant lethals.' These appeared in abundance when Muller did his 1927 work inducing mutations with X-rays. At high doses (about 4000 R) about 90% of the embryos or eggs laid by females receiving the X-rayed sperm did not hatch larvae. Muller and Pontecorvo designed stocks to reveal that the cause was a chromosomal rearrangement in which chromosomes lacked or contained two centromeres. The centromere is the attachment region of a chromosome when it is pulled to the cell's poles during cell division. Independently, Barbara McClintock found the same phenomenon in maize and she called it a 'breakage-fusion-bridge cycle.' Muller argued after the reports of radiation sickness from the atomic bombings in Japan, that the breakage-fusion-bridge cycle caused the cells to die and since a person in Hiroshima or Nagasaki would have massive chromosome breaks the vulnerable tissues were those that had actively dividing cells such as skin, intestinal lining, bone marrow for red blood cell formation, lining cells of the blood vessels, reproductive tissue, and cells of the immune system. This resulted in the rashes, bleeding through the intestines, drop in blood count, diminished immunity to infections, peritonitis, and other damage reflected in the survivors of radiation sickness.

The concerns over radiation increased as the Cold War developed and the USSR began testing its own atomic weapons, including hydrogen bombs. These released large amounts of radioactive material that were carried in the atmosphere and dropped in rain called 'fallout.' When the United States tested its early hydrogen bombs in the Pacific, a boat with Japanese fishermen many miles away from the test were showered with radioactive fallout that made them sick with the same symptoms as radiation sickness. The United States quickly learned how to detect small doses of radiation (atmospheric fallout) from Soviet tests and a debate erupted over the safety to the civilian population in the United States from these atomic bomb tests. There were three positions that evolved. Edward Teller and Willard Libby took the position that small doses of radiation were harmless or beneficial. In Teller's view the fallout increased the number of mutations, which made evolutionary change possible. Some geneticists cited by the Atomic Energy Commission took the view that heterozygosity induced by radiation was good for a person because it induced hybrid vigor.

Opposed to this view were those who advocated an immediate halt to weapons testing because small doses of radiation were harmful. Chief advocate for this second view was Linus Pauling. He estimated the number of new genetic deaths and health effects of radiation fallout world-wide and they numbered several tens of thousands. Pauling admitted this was a small rise above control rates but felt it was immoral to cause any deaths when weapons testing treaties could prevent them and even unilateral stopping of tests would put pressure on the USSR not to do atmospheric tests. Muller represented a third view. He argued that low doses produced small amounts of damage but one had to balance national security (deterring the USSR from using nuclear weapons) against these sacrifices. He felt it was false to exaggerate low dose effects by either denying their existence or exaggerating their significance to health. He opposed unilateral stopping of nuclear tests and favored bilateral discussions. The debate was never resolved by scientific consensus. Instead, Pauling's efforts to get political action worked and the United States and the USSR worked out a treaty to ban atmospheric testing. No such testing by either power has occurred since.

Muller's Views on the World View of Science

Muller rejected the supernatural in any aspect of science. He fought against those who wished to substitute religious creationism for the origin of life, diversity, and species formation on earth. He was a reductionist who believed science could take any aspect of the natural world, reduce it to its components and show how these work and how the coordinated system functions. He rejected vitalism, holism, and other philosophies of science that invoked some supernatural component (élan vital, enteleche, soul, spirit, or essence). But Muller also believed in ideals to live by. He felt we need them for our careers, our families, and our participation in social life as citizens. He believed human values had an evolutionary history, both cooperative and aggressive behavior being products of mammalian behavior and sometimes subject to endocrine regulation at times of stress or different stages of the life cycle. He believed we can favor and select behavior that fosters respect for differences, tolerance for divergent views, empathy for others, and qualities that we call leadership. It was those components of our personalities that he hoped to cultivate through positive eugenics. Yet even with a eugenic outlet denied by social fears of abuse, he felt this secular humanist outlook was one we should emulate. He became active in the American Humanist Association and served as their President and was named Humanist of the Year in 1957.

Muller lost his faith in communism through his experience in the USSR, even testifying to Congress during the height of the Cold War that he would rather be dead than live his life in a Stalinist society. He never abandoned his belief in the central tenets of democratic socialism such as that practiced in Scandinavia or by the Labor Party in Great Britain. He believed in

government provision of education, health, retirement, and welfare because there was no way a capitalist society could provide them through private charity or private insurance when the wages of laborers would be barely sufficient to provide transportation, food, clothing, and shelter.

Muller's Status after His Death in 1967

Classical genetics has been absorbed by molecular genetics. Muller's idea in 1936 of the gene as a type of crystal that reproduced its variations and which might someday be worked out through X-ray diffraction analysis came true almost 20 years later with the working out of the double helix model of DNA (Muller, 1936b). Muller's view of mutation as a change in the individual gene has been worked out at a molecular level with the type of lesions agents cause within the gene. These are described as changes in the nucleotides or as minute additions or losses or inversions of one or more nucleotides in DNA. The existence of modifier genes and chief genes as a method for generating variations in character expression has been found in many organisms. Muller played a role in getting his post-doctoral student, Charlotte Auerbach, at Edinburgh to perform the first experiments to demonstrate induction of mutations by chemicals. She used mustard gas and hundreds of compounds have been found since then that act as mutagens. Muller's name today is attached to a concept called Muller's ratchet and evolutionary theory is still debating how effective gene loss and chromosome size diminution is associated with this process. While historians of science continue to study Muller's life and career, his name has faded as did that of his own mentor, Thomas Hunt Morgan, and most of the founders of classical genetics between 1900 and 1930.

Muller's name resurfaces when controversies over radiation hazards arise or when discussions of the outcome of new mutations in humans arise. He has been attacked by some of those who favor expanded use of nuclear reactors for electric power and by those who disregard low-dose effects of radiation. Muller fought hard to protect the public in the 1950s by serving on national committees to regulate the emerging nuclear power industry. He argued that as low a dose as possible should be the standard for building and operating nuclear plants because if the initial standards were generous to industry it would be almost impossible to alter those facilities once they were in operation. He also succeeded in getting the health professions to regulate radiation exposure to both practitioners and the public. New technologies of digital X-ray imaging have greatly reduced doses for diagnostic X-rays.

Muller was small in size (about five foot two inches tall), athletic (he enjoyed hiking), well-informed about current events, and he carefully limited his interests to that of his family, national politics, the world of science, and the ideals he chose to live by. He had no hobbies. He rarely relaxed and he worked 7 days a week. He was generous to his students. He allowed them to publish in their own names. He was a demanding mentor who made his students write and rewrite their articles and dissertation until all the criticisms he could introduce were answered.

What is unusual in Muller's career is the amount of attention he applied to the uses and abuses of new scientific findings, especially from his own work. Many scientists adopt an 'ivory tower' role of doing basic science and not concerning themselves with the application of science to society. Muller had enormous personal courage. At a time when two of his Russian students were arrested and later executed during the Stalin Terror years of 1936–38, he took on Lysenko in public debate. He spoke out because he believed that biology necessarily will be at odds with state, commercial, and special interest views of sex, heredity, scientific method, and the integrity of new knowledge.

Muller's papers are chiefly to be found at the Lilly Library at Indiana University in Bloomington. A smaller concentration of papers, many of them his letters to his family, were donated by his son and daughter to the Cold Spring Harbor Laboratory Library archives. The author of this article also donated his Muller papers to the Cold Spring Harbor Laboratory Library archives. The author also wrote a biography of Muller. Muller is a major figure in James Schwarz's book, *In Pursuit of the Gene* (Schwartz, 2008).

See also: Capitalism; Cold War, The; Eugenics, History of; Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; National Socialism and Fascism; Racism, History of; Russian Revolution, The; Social Inequality in History (Stratification and Classes); Socialism.

Bibliography

Altenburg, E., Muller, H.J., 1920. The genetic basis of truncate wing—an inconstant and modifiable character in *Drosophila*. Genetics 5, 1–59.
Carlson, E.A., 1982. Genes, Radiation, and Society: The Life and Work of H.J. Muller. Cornell University Press, Ithaca, NY.
Carlson, E.A., 2007. Mendel's Legacy: The Origins of Classical Genetics. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Carlson, E.A., 2001. The Unfit: A History of a Bad Idea. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Medvedev, Z.A., 1969. The Rise and Fall of T.D. Lysenko. Columbia University Press, NY.
Muller, H.J., 1918. Genetic variability, twin hybrids, and constant hybrids, in a case of balanced lethal factors. Genetics 3, 422–499.
Muller, H.J., 1922. Variation due to change in the individual gene. American Naturalist 56, 32–50.
Muller, H.J., 1927. Artificial transmutation of the gene. Science 66, 84–87.
Muller, H.J., 1932. Some genetic aspects of sex. American Naturalist 66, 118–138.
Muller, H.J., 1934. The dominance of economics over eugenics. In: A Decade of Progress in Eugenics: Proceedings of the Third International Congress of Eugenics. William and Wilkins, New York, pp. 138–144.
Muller, H.J., 1935. Out of the Night. Vanguard Press, New York.
Muller, H.J., 1936a. Bar duplication. Science 83, 528–530.
Muller, H.J., 1936b. The need of physics in the attack on the fundamental problems of genetics. Scientific Monthly 44, 210–214.
Muller, H.J., 1950. Our load of mutations. American Journal of Human Genetics 2, 111–176.
Muller, H.J., 1952. Genetic damage produced by radiation. Science 121, 837–840.
Muller, H.J., 1961. Human evolution by voluntary choice of germplasm. Science 134, 643–649.
Ray-Chaudhuri, S.P., 1944. The validity of the Bunsen-Roscoe law in the production of mutations using radiation of extremely low dose. Proceedings of the Royal Society of Edinburgh 62, 66–72.
Schwartz, James, 2008. In Pursuit of the Gene: From Darwin to DNA. Harvard University Press, Cambridge, MA.

Multicultural Education

James A Banks, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Multicultural education emerged out of the Civil Rights Movement in the United States in which African-Americans demanded that educational and other institutions incorporate their history and culture into the curriculum and make other reforms that would create greater educational equality. The Civil Rights Movement echoed throughout the United States and the world and stimulated other marginalized groups to make similar demands of schools, colleges, and universities, as well as other institutions within their nation-states. This article describes the goals, scope, and implementation of multicultural education in the Western immigrant nations of the United States, Canada, and Australia; in several Western European nations; and in Japan, Korea, and China. It also explains how citizenship education is being conceptualized and implemented in a number of multicultural nations.

The Goals and Scope of Multicultural Education

Multicultural education is an educational reform movement whose major aim is to change schools, colleges, and universities so that students from diverse groups will experience equal educational opportunities. The specific groups that are the focus of multicultural educational reform differ both within and across nations. In the democratic Western nations that view themselves as immigrant nations such as Australia, Canada, and the United States, the groups that are the focus of multicultural education include racial, ethnic, cultural, social-class, linguistic, and religious groups. However, especially within the United States, groups with disabilities as well as students who are gifted and talented often receive attention within multicultural education school, college, and university programs. Since the late 1990s, LGBT (lesbian, gay, bisexual, and transgender) students have become a significant focus of multicultural education in the United States (Mayo, 2013).

Multicultural education first developed in the United States in response to the African American Civil Rights Movement that occurred during the late 1960s and 1970s. African-Americans demanded that schools, colleges, and universities include content about their history and culture in the curriculum and hire African-American teachers, professors, and administrators. Other ethnic groups within the United States, such as Mexican Americans, Asian Americans, and American Indians, made demands similar to those made by African-Americans. Consequently, ethnic studies was the first focus of multicultural education in the United States.

The quests by African-Americans to achieve equal rights in education and within other areas – such as affirmative action in employment and cultural representation in museums and historical sites – echoed throughout the world. Indigenous and ethnic groups in other nations also demanded structural inclusion and educational equality. They called for schools, colleges, and universities to incorporate their cultures and histories into the curriculum and to hire teachers and professors from their racial, ethnic, cultural, and language groups. African Caribbeans in the United Kingdom, First Nations people in Canada, the Australian Aborigines, and the Maori in New Zealand pushed for the inclusion of their histories and cultures into the school, college, and university curriculum. During the late 1990s and 2000s, these groups also asked educational institutions to make teaching in schools more culturally responsive to the unique cultures of the students from their ethnic, cultural, and language communities (Banks, 2009; de Haan, 2012).

Multicultural Education in Western Europe

Many nations in Western Europe – such as France, The Netherlands, Switzerland, Ireland, and England – are facing challenges providing equal educational opportunities for groups that are marginalized because of their racial, ethnic, cultural, linguistic, and religious characteristics. These groups are also challenging notions of citizenship and citizenship education in European nations and forcing a reexamination of these concepts. Schools, colleges, and universities in many European nations have implemented a variety of initiatives and programs to increase the academic achievement of ethnic, linguistic, and religious minority students and to facilitate their participation in mainstream culture and institutions (de Haan, 2012; Osler and Starkey, 2009; Starkey, 2012; Tomlinson, 2009).

When the Civil Rights Movement began in the United States in the 1960s and 1970s, the Western European nations were characterized by ethnic, cultural, racial, linguistic, and religious diversity. Much of this diversity had deep historical roots. The nations in Western Europe had long-standing linguistic and cultural minorities, such as the Basques in France and Spain, the Germans in Denmark, the Danes in Germany, and the Welsh, Scots, and Jews in the United Kingdom. Diversity in European nations was greatly increased when thousands of immigrants from colonial nations came to Europe to improve their economic and social status and to meet labor needs in the years after World War II. These immigrant groups included Asian Indians, Pakistanis, and Jamaicans in England; the Vietnamese and North Africans in France; the Turks in Germany; and the Surinamese in the Netherlands. This series of migrations from former colonial nations was described in a book published in England by the Centre for Contemporary Cultural

Studies at the University of Birmingham in 1982 titled, *The Empire Strikes Back*: *Race and Racism in 70s Britain*.

There is a wide gap between the democratic ideals in European nations and the daily educational experiences of minority groups in schools (Tomlinson, 2004). Ethnic minority students in Europe, as in other nations throughout the world, often experience discrimination because of their cultural, linguistic, and religious differences. Both students and teachers often perceive these students as the 'Other.' When they are marginalized within school and treated as the 'Other,' ethnic minority students – such as Turkish students in Germany and Muslim students in France – tend to emphasize their ethnic identity and to have weak attachments to their nation-state (Mannitz, 2005; Tomlinson, 2009). The extent to which nation-states make multicultural citizenship possible, the recognition and visibility of diverse groups in the curriculum, and the linguistic rights of immigrant and minority groups are among the unresolved and contentious educational issues in European nations.

The academic achievement gap between ethnic minority and majority group students is another salient issue that is being theorized and researched by multicultural education scholars in Europe. African-Caribbean, Pakistani, and Bangladeshi students in Britain and Turkish students in Germany are not achieving on a level equal to that of majority group students in their nations. However, the underachievement of minority students is a complex issue worldwide that defies facile solutions and responses. Indian and Chinese students are the highest-achieving students in Britain (Tomlinson, 2004). In the United States, the academic achievement of Chinese, Japanese, and Asian Indian students exceeds the achievement of White students.

Multicultural Education in East Asian Nations

The construction and evolution of multicultural education in Asian nations such as Japan, Korea, and China has differed in significant ways from its development in the immigrant nations of Australia, Canada, and the United States as well in most European nations. Both Japan and Korea have historically viewed themselves as homogeneous nations with one dominant ethnic group, language, and culture. Little public recognition or attention has been given to the Ainu, the indigenous people of Japan. The Burakumin, which is often referred to as Japan's 'invisible race,' is another marginalized group that has historically experienced discrimination.

Japan and Korea have historically had a *jus sanguinis* notion of citizenship, whereby individuals needed Japanese or Korean 'blood' to become citizens (Hirasawa, 2009; Moon, 2012). This is in contrast to *jus soli*, which automatically grants citizenship to individuals born within a nation, which is the practice in the United States. In the past, citizenship in Germany was also based on the idea of *jus sanguinis*.

Multicultural education in Japan developed after a significant number of immigrants began to settle there in the 1990s (Hirasawa, 2009). In 2008, approximately 2.2 million people with non-Japanese nationality lived in Japan, which was about 1.74% of the population. Migration to Korea has increased significantly since the 2000s. Most of the immigrants to Korea are migrant workers, individuals from other nations who marry Koreans, and people who defect from North Korea (Moon, 2012). The 1 million migrants who were living in Korea in 2008 made up more than 2% of the national population.

China has 56 official ethnic groups, which consist of the dominant or mainstream Han people and 55 ethnic minority groups, which is approximately 123 million people and about 9.44% of the Chinese population (Teng and Zhu, 2012). Ethnic autonomous regions, which make up half of the geographical region of China, are the areas where most ethnic minorities in China live (Postiglione, 2009). The major goal of multicultural education in China is to enable ethnic minority groups to attain literacy in Mandarin Chinese and to acquire the other knowledge and skills needed to function in mainstream Han society without alienating these students from their home languages and cultures. Most ethnic minority students in China enter school speaking an indigenous language or languages and have little proficiency in Mandarin Chinese, which is the national language and the language needed for social mobility and structural inclusion.

The official policy in China is for ethnic minority students to retain their own language or languages while learning Mandarin Chinese. However, the strong assimilationist policy implemented in most schools in China results in many ethnic and linguistically minority students becoming alienated from their home and community cultures and languages (Beckett and Postiglione, 2012; Wang and Phillon, 2009). Important goals of citizenship education in China are for all students to develop an identity with the Chinese national political system and to internalize and respect collectivism (Wan, 2004).

Multicultural Citizenship Education

A significant goal of multicultural education in various nations is to help students from diverse racial, ethnic, cultural, linguistic, and religious groups to attain the citizenship knowledge and skills needed to function effectively within their own cultural communities, across other cultures and communities, within their nation-state, and within the global community (Banks, 2008). This goal is a serious challenge for nations around the world – there is a wide gap in both democratic nations and socialist nations between the ideals that are articulated about diversity and inclusion and the realities within classrooms and schools. Wide gaps between national ideals and actual practices in schools, colleges, and universities exist within the Western democratic nations such as the United States and the United Kingdom as well as in socialist nations such as the People's Republic of China (Wan, 2004) and the Republic of Cuba (Carnoy, 2009). Banks (2004) calls the gap between ideals and practices a 'citizenship education' dilemma.

Multicultural nation-states are faced with the problem of constructing a national culture and identity that reflects and incorporates the diversity of its citizens and yet have an overarching set of shared values, ideals, and goals to which all of its citizens are committed. Citizenship education theorists such as Banks (2007), Gutmann (2004), and Klymlicka (1995) assert that only when a nation-state is unified by inclusive values such as justice and equality can it protect the rights of cultural,

ethnic, religious, and linguistic groups and enable them to experience cultural democracy and freedom. Klymlicka (1995), the Canadian political theorist, and Rosaldo (1997), the US anthropologist, have constructed theories about diversity and citizenship. Both Kymlicka and Rosaldo state that in a democratic society, ethnic and immigrant groups should have the right to maintain their ethnic cultures and languages as well as participate fully in the national civic culture. Kymlicka calls this concept 'multicultural citizenship'; Rosaldo refers to it as 'cultural citizenship.'

In the United States in the 1920s Drachsler (1920) used *cultural democracy* to describe what citizenship education theorists call multicultural citizenship today. Drachsler (1920) and Kallen (1924) – who were Jewish immigrants and advocates for the cultural freedom and rights of the Southern, Central, and East European immigrants in the United States – argued that *cultural democracy* is an important characteristic of a democratic society. They maintained that cultural democracy should coexist with political and economic democracy, and that citizens from diverse groups in a democratic society should participate freely in the civic life of the nation-state and experience economic equality. They should also have the right to maintain important aspects of their community cultures and languages, as long as they do not conflict with the shared democratic ideals of the nation-state. Cultural democracy, argued Drachsler, is an essential component of a political democracy.

Balancing Unity and Diversity

Cultural, ethnic, racial, linguistic, and religious diversity exists in most nations around the world (Banks, 2009). One of the challenges to diverse democratic nation-states is to provide opportunities for different groups to maintain aspects of their community cultures while constructing a nation in which these groups are structurally included and to which they feel allegiance. Citizenship education theorists such as Banks and his colleagues (2005) believe that a delicate balance of diversity and unity is an essential goal of democratic nations and of teaching and learning in democratic societies and that unity should be an important aim when nation-states are responding to diversity within their populations. Gutmann (2004) states that nation-states can protect the rights of minorities and enable diverse groups to participate only when they are unified around a set of democratic values such as justice and equality.

In the past, most nations have tried to create unity by forcing racial, cultural, ethnic, linguistic, and religious minorities to give up their community languages and cultures in order to participate in the national civic culture (Joshee, 2009; Nieto, 2009; Tomlinson, 2009). During the 1950s and 1960s in the Southwest of the United States, Mexican-American students were sometimes punished for speaking Spanish in school. In the nineteenth century in the United States, American-Indian youth were forced to attend boarding schools where their cultures and languages were eradicated (Lomawaima and McCarty, 2006). In Australia, Aboriginal children were taken from their families and forced to live on state missions and reserves (Broome, 1982), a practice that lasted from 1869 to 1969. These children are called 'The stolen generation.' Kevin Rudd, the Australian Prime Minister, issued a formal apology to the stolen generation on 13 February 2008.

Citizenship education theorists believe that in order to embrace the national civic culture, students from diverse groups must feel that it reflects their experiences, hopes, and dreams and that schools and nations cannot marginalize the cultures of groups and expect them to feel structurally included within the nation and to develop a strong allegiance to it.

Multicultural Education across Nations

The major goal of multicultural education is to reform schools, colleges, and universities so that students from diverse groups will experience cultural recognition, educational equality, and structural inclusion. However, multicultural education reflects the national, cultural, and social context of the nation in which it is implemented. Not only do the groups on which multicultural education focuses vary by the national and social context, its approaches, implementation, and assessment are also influenced by contextual and nationwide factors. As the implementation of multicultural cross cultures and nations, it is transformed and mediated. However, its aim of increasing equality and social justice is consistent when it is implemented in different cultures and nations.

See also: Bilingual Education: International Perspectives; Educational Philosophy; Equity and Education; Globalization in Education; School Management.

Bibliography

Banks, J.A., 2004. Introduction: democratic citizenship education in multicultural societies. In: Banks, J.A. (Ed.), Diversity and Citizenship Education: Global Perspectives. Jossey-Bass, San Francisco, pp. 3–15.

Banks, J.A., 2007. Educating Citizens in a Multicultural Society, second ed. Teachers College Press, New York.

Banks, J.A., 2008. Diversity, group identity, and citizenship education in a global age. Educational Researcher 37 (3), 129–139.

Banks, J.A. (Ed.), 2009. The Routledge International Companion to Multicultural Education. Routledge, New York & London.

Banks, J.A., Banks, C.A.M., Cortés, C.E., Merryfield, M.M., Moodley, K.A., Murphy-Shigematsu, S., Osler, A., Park, C., Parker, W.C., 2005. Democracy and Diversity: Principles and Concepts for Educating Citizens in a Global Age. University of Washington, Center for Multicultural Education, Seattle, WA.

Beckett, G.H., Postiglione, G.A. (Eds.), 2012. China's Assimilationist Language Policy: The Impact on Indigenous/Minority Literacy and Social Harmony. Routledge, London; New York.

Broome, R., 1982. Aboriginal Australians: Black Response to White Dominance 1788–1980. George Allen & Unwin, Sydney, Australia.

Carnoy, M., 2009. Social inequality as a barrier to multicultural education in Latin America. In: Banks, J.A. (Ed.), The Routledge International Companion to Multicultural Education. Routledge, New York; London, pp. 512–525.

Centre for Contemporary Cultural Studies, 1982. The Empire Strikes Back: Race and Racism in 70s Britain. Hutchinson & Company and the Centre for Contemporary Cultural Studies, University of Birmingham, London.

Drachsler, J., 1920. Democracy and Assimilation. Macmillan, New York.

Gutmann, A., 2004. Unity and diversity in democratic multicultural education: creative and destructive tensions. In: Banks, J.A. (Ed.), Diversity and Citizenship Education: Global Perspectives. Jossey-Bass, San Francisco, pp. 71–96.

Hirasawa, Y., 2009. Multicultural education in Japan. In: Banks, J.A. (Ed.), The Routledge International Companion to Multicultural Education. Routledge, New York; London, pp. 159–169.

de Haan, M., 2012. Culture and learning. In: Banks, J.A. (Ed.), Encyclopedia of Diversity in Education, vol. 1. Sage Publications, Thousand Oaks, CA, pp. 568–573.

Joshee, R., 2009. Multicultural policy in Canada: competing ideologies, interconnected discourses. In: Banks, J.A. (Ed.), The Routledge International Companion to Multicultural Education. Routledge, New York; London, pp. 96–108.

Kallen, H.M., 1924. Culture and Democracy in the United States. Boni and Liveright, New York.

Klymlicka, W., 1995. Multicultural Citizenship: A Theory of Minority Rights. Oxford University Press, New York.

Lomawaima, K.T., McCarty, T.L., 2006. "To Remain an Indian:" Lessons in Democracy from a Century of Native American Education. Teachers College Press, New York.

Mayo, C., 2013. Queer lessons: sexual and gender minorities in multicultural education. In: Banks, J.A., Banks, C.A.M. (Eds.), Multicultural Education: Issue and Perspectives, eighth ed. Wiley, Hoboken, NJ, pp. 161–175.

Mannitz, S., 2005. Collective solidarity and the construction of social identities in school: a case study on immigrant youths in post-unification West Berlin. In: Luchtenberg, S. (Ed.), Migration, Education and Change. Routledge, London; New York, pp. 64–85.

Wan, M.-G., 2004. Ethnic diversity and citizenship education in the people's republic of China. In: Banks, J.A. (Ed.), Diversity and Citizenship Education: Global Perspectives. Jossey-Bass, San Francisco, pp. 355–376.

Moon, S., 2012. Korea, multicultural education in. In: Banks, J.A. (Ed.), Encyclopedia of Diversity in Education, vol. 3. Sage Publications, Thousand Oaks, CA, pp. 1307–1312.

Nieto, S., 2009. Multicultural education in the United States: historical realities, ongoing challenges, and transformative possibilities. In: Banks, J.A. (Ed.), The Routledge International Companion to Multicultural Education. Routledge, New York; London, pp. 79–95.

Osler, A., Starkey, H., 2009. Citizenship education in France and England: contrasting approaches to national identity and diversity. In: Banks, J.A. (Ed.), The Routledge International Companion to Multicultural Education. Routledge, New York; London, pp. 334–347.

Postiglione, G.A., 2009. The education of ethnic minority groups in China. In: Banks, J.A. (Ed.), The Routledge International Companion to Multicultural Education. Routledge, New York; London, pp. 501–511.

Rosaldo, R., 1997. Cultural citizenship, inequality, and multiculturalism. In: Florres, W.V., Benmayor, R. (Eds.), Latino Cultural Citizenship: Claiming Identity, Space, and Rights. Beacon, Boston, MA, pp. 27–28.

Starkey, H., 2012. Human rights and education. In: Banks, J.A. (Ed.), Encyclopedia of Diversity in Education, vol. 2. Sage Publications, Thousand Oaks, CA, pp. 1115–1118.

Teng, X., Zhu, S., 2012. China, multicultural education in. In: Banks, J.A. (Ed.), Encyclopedia of Diversity in Education, vol. 1. Sage Publications, Thousand Oaks, CA, pp. 342–347.

Tomlinson, S., 2004. The education of migrants and minorities in Britain. In: Luchtenberg, S. (Ed.), Migration, Education and Change. Routledge, London; New York, pp. 86–102.

Tomlinson, S., 2009. Multicultural education in the United Kingdom. In: Banks, J.A. (Ed.), The Routledge International Companion to Multicultural Education. Routledge, New York; London, pp. 121–133.

Wang, Y., Phillon, J., 2009. Minority language policy and practice in China. The need for multicultural education. International Journal of Multicultural Education 11 (1), 1–14 (Online journal).

Multiculturalism

Duncan Ivison, University of Sydney, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Multiculturalism refers to (1) the state of a society or the world in which there exists numerous distinct ethnic and cultural groups seen to be politically relevant; and (2) a program or policy promoting such a society. Political philosophers aspire to deal with the challenges posed by multiculturalism fairly – but there is deep disagreement about what constitutes fair treatment. Multiculturalism might entail the granting of rights (individual or collective) on the grounds of the value cultures have for individuals, or even society as a whole. A multicultural state then might be one in which different forms of recognition are granted to cultural groups to secure the valuable goods cultural membership provides equally and avoid unfairly privileging dominant groups. On the other hand, it can be argued that precisely because of deep social and cultural diversity, the state should remain neutral between groups and ensure that basic liberal rights are fairly and consistently applied across cultural differences. Granting recognition to groups risks undermining the security and welfare of individuals by 'politicizing' forms of cultural identification best kept out of public decision making about the distribution of goods and resources. These views reflect not only conflicting intuitions about the relation between culture and individual welfare, but also deep disagreements about the nature of equality and freedom.

Multiculturalism has come to refer to (1) the state of a society or the world in which there exists numerous distinct ethnic and cultural groups seen to be politically relevant; and (2) a program or policy advocating or promoting such a society. The first is a factual claim, the second evaluative. But like so many claims in the social sciences, the empirical claim involves much more than empirical judgments, and the evaluative claim is contested by as many who promote it (and certainly does not follow from the empirical claim). Controversy has raged over what the best response to the fact of cultural diversity in the world might be, and it is not too much to suggest that the social and political issues raised by multiculturalism continue to be among the most pressing and important of our time. In recent years, multiculturalism also has come under increasing attack from commentators on both the left and the right. Coming to grips with these issues requires asking some difficult questions about the nature of culture and its relation to human welfare, and especially what political consequences follow in light of the answers given. Political philosophers aspire to deal with the challenges posed by multiculturalism *fairly*; the problem is, no one is quite sure what this actually entails.

The Relevance of Culture

Human societies have always been culturally diverse and certainly most organized political societies. Yet for much of the history of Western political philosophy, the ideals of political life that have been promulgated have presumed that citizens shared not only a common political status but also a common descent, language, and culture. Aristotle, for example, one of the central figures in the history of Western political thought, presumed this despite himself being a 'resident alien' of Athens. John Stuart Mill, champion of modern liberty and individual autonomy, thought that free institutions were next to 'impossible' among a people 'without fellow feelings,' or who read and spoke different languages. He argued that the boundaries of a state should coincide "in the main with those of nationalities" (Mill, 1975: p. 392). The presumption that states should be culturally homogeneous was inspired in part, no doubt, by straightforward prejudice against foreigners and outsiders, but it also was linked to a series of issues that continue to bedevil contemporary political philosophers. Among these include the following: Does justice require a common culture to motivate the goodwill and sacrifices it typically demands of citizens? Can a political society be stable if too much emphasis is placed on the diversity of its peoples and cultures? Can democratic institutions work, and be seen to be legitimate, if the citizens of a state speak different languages, have different worldviews, and even perceive the history of their nation differently? Finally, if individuals are fundamentally equal, can we balance the demands of equality with the diverse moral commitments of particular cultures?

Political philosophers from Aristotle to Mill (and beyond) have tended to answer these questions in the negative; too much cultural diversity threatens to undermine the possibilities for a peaceful, prosperous, and ultimately just political community. The problem is, the challenges of cultural diversity are impossible to avoid. The history of the twentieth century, and especially of recent decades, makes this abundantly clear; witness the reemergence of nationalism, the spread of violent ethnic conflict, and the extraordinary movement of peoples across borders fleeing persecution of various kinds and seeking new hope in distant lands.

Before tackling these questions directly, however, it is important to get a sense of what actually is meant by 'culture' and how it has come to occupy a preeminent position in discussions concerning distributive justice and the emergence of 'the politics of difference.'

Diversity of What?

Theorists concerned with multiculturalism agree that culture matters, but they disagree profoundly over how and why it does. Liberal political philosophers, in particular, have struggled to reconcile their commitment to the welfare of

individuals with what seems to be the collective nature of cultural identification. But first, when political philosophers talk about cultural diversity, what are they are actually talking *about*?

The concept of 'culture' has a complex history and, as a result, a set of meanings sometimes at odds with each other (Moody-Adams, 1997; Appiah, 2005). At a basic level, culture refers to the particular set of habits, beliefs, and customary practices of a people or society; to its way of life. In the nineteenth century, culture often was tied closely with quasi-biological notions of race and ethnicity, and thus to judgments about the inherent superiority of one form of culture over another – a tendency that continued to manifest itself in murderous ways throughout the twentieth century. Linking culture to biological or quasi-scientific characteristics thus has come to been seen as extremely dubious, involving both bad science and questionable ethics. Instead, anthropologists (among others) have argued that culture should be understood as a set of ultimately malleable and negotiable cognitive and social frameworks governing human behavior that is complex and dynamic rather than static and that is patched together from a variety of sources rather than woven of the same cloth (Geertz, 1973; Clifford, 1988; Moody-Adams, 1997; Appiah, 2005).

But how are cultures individuated politically? Some insist that the groups that matter politically are those grounded in *ethnicity* (stripped of any biological or scientific overtones): a people united by certain prepolitical allegiances of ancestry, collective memory, homeland, or culture (Smith, 1986). For others, cultural diversity refers to a range of different groups – whether based on ethnicity, race, socioeconomic class, language, lifestyle, or sexual orientation – who are disadvantaged relative to a dominant political culture manifested in the main institutions and 'public reason' of the state in which they reside (Young, 1990). Those shared characteristics or practices, often only perceived as being shared in light of 'mainstream' oppression or rejection, are said (for better or worse) to partly constitute the culture of a group.

Another important account of cultural diversity is provided by Will Kymlicka (1995a, 1998, cf 1995b, 2007). He argues that multiculturalism involves two different kinds of cultural diversity. The first is a diversity of 'societal cultures.' The second is created by individual and familial migration. A societal culture provides its members "with meaningful ways of life across the full range of human activities, including social, educational, religious, recreational, and economic life, encompassing both public and private spheres" (1995a: p. 76). It consists of not only collective values and memories, but usually also a shared language and territory (1995a: pp. 76–80). Note that these are *cultural* groups rather than groups based on race or descent. Membership thus is defined in terms of integration into a cultural community rather than in terms of descent. Aboriginal 'Nations' and the Quebecois, for example, are societal cultures, but also ones in which there is a high degree of racial and ethnic intermixing and thus great internal diversity. Although a state may contain a dominant societal culture, it also might include other minority cultures. Kymlicka called these groups 'national minorities': at the time of their incorporation into a larger state they possessed (1) a distinct 'social vocabulary' embodied in a full set of social practices and institutions, (2) were territorially distinct, and (3) self-governing. The capacity and motivation to form such a culture, for Kymlicka, means it is characteristic of 'nations' or 'peoples.' Migrants, on the other hand (and refugees, although they present a difficult case), essentially have chosen to give up their societal culture when deciding to migrate (1995a: pp. 100–101). So multiculturalism involves (at least) both 'polyethnic' diversity (arising from individual and familial migration) and 'multinational' diversity (arising from the incorporation of previously self-governing cultures into a larger state). Religious groups who seek out marginalization from mainstream society present something of a special case, which will be discussed later. Furthermore, there are groups which cut across ethnic and national lines – such as women, gays and lesbians, and the disabled – who, at different times, make claims based on their disadvantage or difference relative to a dominant societal culture. These claims tend to have more in common with 'polyethnic' rather than 'multinational' claims, however, since they usually concern the terms of integration into society rather than a demand for the preservation or protection of a distinct society (cf Kymlicka, 1998: pp. 90–103).

The reason why so much hangs on how a culture is to be individuated is because of what is said to follow in terms of the *rights* members of that culture – or indeed the group itself – can claim. For Kymlicka, 'national groups' are entitled to strong group rights; that is, those that entail varying degrees of self-government and protection from the decisions of the wider political community. Polyethnic rights are important but much weaker; they are concerned mainly with ensuring that the terms of integration into the dominant societal culture are as fair as possible.

Any set of distinctions generates anomalies and hard cases, and Kymlicka's is no different. Some groups, like certain religious groups, seem to fulfill the criteria for being a 'societal culture' and yet usually are not granted rights associated with national minorities. In some cases however, some 'ethno-religious' sects – such as the Amish or Hutterites in North America – have been granted exemptions from general laws to enable them to maintain certain key features of their religious life. But these are historically special cases; they usually involve a theology that seeks to minimize contact with the modern world. Other groups, such as refugees and some migrant groups (e.g., Muslim communities in the United Kingdom), do not fit comfortably into either category; they are not national minorities, but they also seem more vulnerable to collective harm than the discussion of 'polyethnic' rights might suggest. But how is it that rights follow from claims about culture in the first place?

Cultural Rights

What is striking about recent debates among liberals about multiculturalism is how liberalism's traditional concern with individual well-being has been linked to the value of cultural membership – an ostensibly 'collectivist' value. Not all cultures are individualistic, of course, but liberals generally are concerned with how culture matters for the lives of individuals. But how exactly does culture matter in this way? And even if it does, is it something in relation to which rights should be, or even could be, assigned?

Consider three basic arguments:

1. *The argument from cultural diversity*: Respect for cultural pluralism entails preserving existing cultures. Pluralism is valuable because it provides individuals with alternative ways to live, promotes critical reflection on the culture within which one currently lives, and thus promotes change and growth within cultures generally.
2. *The historical injustice argument*: Certain group rights are justified insofar as they are aimed at redressing or compensating for a history of past injustice against particular historical communities at the hands of dominant majorities.
3. *The value of cultural membership argument*: Certain group rights are justified insofar as membership in a culture has value for individuals, and there is a link between the welfare of the group and welfare of its members. Group rights contribute to the preservation of that culture and thus to the welfare of its members.

All three arguments have played a prominent role in recent debates. However, (1) seems incomplete, because the promotion of cultural diversity per se does not necessarily justify the protection of any *particular* culture. Furthermore, if more cultural diversity is better than less, then either all cultural practices are multiculturalism-enhancing and deserving of toleration or support, or only some are. The former cannot be true (as surely some cultural practices are not so deserving), and the latter forces us to identify those properties, which make some cultural practices valuable and others not. This is no easy task. This leads to consideration of (2) and (3). Perhaps the most powerful case has been one made in relation to (3). The historical injustice argument will be returned to in the Section The Future of Multiculturalism.

One way in which cultures matter to individuals is that they constitute an important aspect of their identity, or as Kymlicka put it, provide a meaningful 'context for choice': "freedom involves making choices amongst various options, and our societal culture not only provides these options, but also makes them meaningful to us" (1995a, 1995b). So in addition to the familiar liberal emphasis on individuals having the freedom of expression, movement, and association, they also need access to a societal culture. A societal culture is a basic 'primary good' for all individuals, and thus relevant to a liberal theory of justice (on primary goods, see Rawls, 1971).

A variation of this argument (but made independently of it) is the relation said to exist between cultural identity and individual self-respect or self-esteem. People's self-respect can be bound up with the way in which their group is perceived by others. Lack of esteem for the group can translate into lack of esteem on the part of individual members and this constitutes harm, the remedy of which may justify granting some form of group rights or rights of self-determination (Taylor et al., 1992; Margalit and Raz, 1990; on the 'social bases of self-respect' see Rawls, 1971: pp. 303, 440).

But do individuals require access to a *particular* culture to secure these benefits, one they can call their own? There are at least three objections to this strategy. Some argue that liberals should not be in the business of preserving particular cultures, since the kind of society they should be promoting is one in which individuals are free to move and adapt *between* cultures as much as within them. So as long as individuals have access to a societal culture – and especially one made up of a diversity of cross-cutting cultural influences – their 'context for choice' is secure. If there is a cultural structure that liberals should promote, it is a cosmopolitan one, a common but diverse 'mélange' of cultural meanings conditioned by the basic liberal freedoms of conscience, expression, association, and exit (Waldron, 1992; Turner, 1993; Appiah, 2005, 2007).

A variation on the cosmopolitan argument is to place greater emphasis on the value of diversity or difference itself. There is no identity without difference, and thus to talk of 'societal cultures,' or even a singular 'cosmopolitan culture,' is to risk missing the inherent multiplicity and hybridity of political and cultural identities. Individuals and groups are constituted by a diversity of moral sources, and the central task of justice should be to remain 'critically responsive' to this fluidity and diversity, rather than reinforce 'found' or extant cultural structures (Bhabha, 1994; Connolly, 1995).

Finally, granting special rights to cultural groups on the grounds of their value for individuals involves a delicate balancing act between the rights of individuals and the measures needed to ensure the survival of the culture. The dangers here are ones to which liberals should be especially sensitive. For some, even to talk about 'societal cultures' or groups as meriting political recognition is to engage in dubious metaphysics and unwise politics. The basic units of social reality cannot be group-based because groups are simply the product of interactions between individuals; there are only associations constituted by individuals and not vice versa (Kukathas, 1997). Thus, there are no grounds upon which to provide political or legal recognition for groups on the basis of their playing a 'constitutive' role in individuals' personal identity. Of course, individuals are free to associate with whom they please. But they are equally free to exit those associations. Moreover, empowering groups works against promoting the kind of social solidarity required for effective democratic societies – either by promoting divisive ethnic 'entrepreneurship,' or undermining a sense of common purpose required for comprehensive social programs (Barry, 2001).

The Politics of Multiculturalism

Disagreement over the social ontology of multiculturalism only intensifies when the more concrete matter of deriving political and institutional arrangements from them arises. The 'politics' of multiculturalism refers to the way in which the consequences of social and cultural diversity are thought to necessitate social and political change.

If the argument is that cultures matter to individuals because they are connected to the capacity for living a good life – whether in terms of providing the 'social bases of self-respect' or as a meaningful 'context for choice' – then does it follow that the *state* has a direct interest in recognizing or promoting a societal culture (or set of cultures)?

For cosmopolitans, it does not follow at all. People are sure to value their cultural attachments, but the state should steer clear of tampering with these commitments. It should treat them the way it treats religious beliefs, as a private matter about

which the state is officially neutral. If it promotes anything, it should be the ability of people to live according to whatever way of life they wish drawn from the many possibilities on offer, consistent with the basic rights of others to do likewise. So the state should remain 'indifferent' toward and generally tolerant of the associations individuals wish to form or join. The costs of a 'politics of difference' – of recognizing and encouraging individuals to organize politically around their cultural identities – outweigh any potential benefits. Politics can be a source of oppression as much as any other social practice, especially given the potential for cultural elites or others to gain control of the resources and institutions set up to protect cultural minorities and use them to take advantage of more vulnerable members or to enrich themselves (Kukathas, 1997: p. 149; Hardin, 1995; Barry, 2001). If individuals are being harmed as a result of their cultural attachments, then the remedy is to ensure a more consistent application of basic liberal rights and protections, not to recognize the groups to which they belong.

The latter view has been referred to as pursuing a policy of 'benign neglect' or 'neutrality.' Cultures neither deserve nor require state assistance to survive, and if they do, they probably are not worth preserving anyway. This view also has been associated with a 'civic' as opposed to 'ethnic' account of the state. A civic state is one in which citizens are members in virtue of a shared commitment to constitutional principles or values that are independent of particular cultural attachments (Habermas, 1993). An ethnic state, on the other hand, presupposes a set of values grounded in a shared cultural framework based on common descent or ethnicity. Hence, also the appeal to civic and ethnic nationalisms, the former involving 'constitutional' patriotism grounded in cosmopolitan principles independent of ethnocultural and historical features, and the latter entailing a particularist patriotism grounded in 'blood and soil.'

Practically speaking, however, the pursuit of neutrality or 'benign neglect' in these ways is extremely difficult to accomplish. States play a crucial role in coordinating social and economic activity, and it is inevitable that, in the process, certain cultural presumptions and practices will be favored over others. Almost every state, for example, has to establish internal boundaries, a common educational system, an 'official' language (or set of languages), and common holidays, and has to enforce a basic legal framework within which citizens will have to act. In other words, the idea that government could be neutral when it comes to the treatment of all cultural groups is not only difficult to imagine, but perhaps even 'incoherent' (Kymlicka, 1995a: p. 108). States have an inescapable cultural component built into them. Moreover, those states that often are invoked as exemplars of civic rather than ethnic nationalism or as having pursued a policy of 'benign neglect' have in fact possessed a distinct societal culture made up of more than just allegiance to cosmopolitan principles and values. Some groups have suffered harm as a result of their difference from this dominant societal culture. Some of these states have granted relatively strong group rights to various cultural groups in light of their disadvantage or via historic agreements in the course of nation building. (All of this seems true of the United States and Canada, for example, two apparently quintessential 'civic' nations.)

Focus on the *harm* people have suffered in relation to their particular cultural attachments has featured prominently in recent debates, mainly because of the appalling treatment just about every state has at some point handed out to minority groups. This touches on the historical injustice argument mentioned previously. Protecting societal cultures may be justified as a means of redressing or compensating members of that culture for a history of past discrimination or oppression at the hands of a dominant majority or minority (Buchanan, 1991; Spinner-Halev, 2012). There are questions surrounding this argument too, however. First, would citizens have to agree on a specific account of their nation's history to fix responsibility for these past injustices? This is difficult to imagine in many places in which intercultural conflict has been particularly brutal. Mutual loathing and distrust undermines cool and impartial reflection on the past.

Second, why should the viability of group rights depend on the group having been oppressed or discriminated against? Tying the granting of group rights to the existence of (past or present) discrimination by or exclusion from a dominant societal culture risks misunderstanding the motivation of groups seeking such rights.

Some societal cultures, it might be argued, are owed special rights independently of any historic pattern of disadvantage they may have suffered (as is argued is the case with Aboriginal peoples; see Kymlicka, 1995a: p. 142; Tully, 1995, 2008). Group rights, in other words, are not necessarily about compensation for exclusion from or discrimination by a dominant culture, but rather about the intrinsic commitment members of particular groups have toward their societal cultures.

But what kind of rights follow specifically in light of the arguments canvassed earlier? It is here that the generality of normative discussions about cultural rights runs up against the complexity of the world in which they are claimed; for it is difficult to imagine a generalizable set of rights that can be applied to all cultural groups equally, given the complex interplay of history, context, and principle in each case. The tensions inherent in the desire to protect both societal cultures and individual well-being emerge with particular force. The very measures needed to protect a societal culture might conflict with the actual freedoms and capacities of the individuals in whose name the protection is being offered in the first place (Hardin, 1995; Okin, 1998; Eisenberg and Spinner Halev, 2005). This is particularly true for those more vulnerable members of a societal culture, such as women and girls (Shachar, 2001; Song, 2007). Having suffered past injustice is no guarantee that relations *within* that group will be just. As a result, it is probably best to think of cultural rights as involving a bundle of different kinds of claims rather than as a complete set of universalizable principles, albeit always conditional on them not to harm the basic interests of individuals. Considered as such, they might include, among other things, the right to be exempted from certain laws or government policies; assistance for the maintenance of linguistic, educational, or cultural activities; various forms of recognition (legal, moral, and political); 'affirmative action' programs, representation rights, self-government rights (within a state or federation); and ultimately secession (Levy, 1997; on secession see Buchanan, 1991). Some of these claims involve negative rights of

noninterference, some positive rights of assistance, and others a combination of the two. Some might be compatible with individual rights (wieldable by individuals as members of that group) and some held exclusively by the group (wieldable by an agent or representative of that group). Much will depend on the context in which they are appealed to. Many inevitably will be controversial.

Aboriginal peoples, for example, have a strong claim for self-government rights because they hold land (or have strong claims to land), possess a distinct societal culture, were previously self-governing, and have suffered a long history of oppression and discrimination at the hands of most settler states. Despite the fact that some Aboriginal practices entail significant restrictions on individual liberty (e.g., rules governing property holding), most liberal theorists think the value and importance to Aboriginal people of maintaining their distinct cultural structure outweigh some of these concerns, or at least that fears of potential abuse are overstated (Kymlicka, 1995a; Tully, 1995; but cf Okin, 1998; Ivison, 2002). Needless to say, when concerning matters such as family law or criminal punishment, opinions begin to differ (Turpel, 1991; Okin, 1998; Pogge, 1997; Shachar, 2001). Self-government rights are the source of considerable debate and social division in countries where they have been claimed or granted. They present a significant challenge to traditional conceptions of equal citizenship and the principle of equality before the law.

Claims by religious groups, on the other hand, generally are not thought of in the same way. They might possess a distinctive societal culture, but save in exceptional circumstances they are not usually thought of as thereby possessing rights to a particular territory or self-government. Their claims more likely will have to do with exemptions from certain state laws, or for state support of education and cultural activities. But in states where religious groups do have extensive exemptions and privileges against the general population (e.g., as in Israel), even these 'weaker' claims can cause considerable social tensions (especially when combined with claims for self-government on the part of another historic community or societal culture living on the very same territory).

The Future of Multiculturalism

It often is suggested that the seemingly irreversible forces of globalization will result in the eventual eradication of the local or particular in favor of a global, although not necessarily cosmopolitan, Western, consumerist culture. Or, that with the end of the Cold War and the 'triumph' of liberal democracy, the great ideological battlelines of the past will be tamed and redrawn within a generally liberal framework (Fukayama, 1992). Recent developments in world politics, however, and especially ongoing ethnic tensions and conflict suggest otherwise. The local and particular has come back with a vengeance. Furthermore, within the established liberal democracies, many are questioning the programs established to promote multicultural identities, especially where they involve providing groups with extra resources or support to maintain their way of life, or as compensation for past discrimination and injustice. This is particularly evident in disputes surrounding the claims of Aboriginal peoples in Canada and Australasia, and reconsideration of 'affirmative action' programs and differentiated citizenship rights in the United States and Europe. The future of multiculturalism, in other words, might appear rather bleak.

Some have laid the blame for disenchantment with multiculturalist policies at the feet of those who were loudest in insisting on the 'recognition of difference' in the first place. Overemphasis on difference has encouraged (perhaps unwittingly) the reification of ethnic and cultural modes of identification, instead of taming them and making them more open-ended and dynamic (Turner, 1993; Hollinger, 1995; Barry, 2001). The constant demand for greater inclusiveness and plurality in every sphere of a nation's cultural life – from its school curricula and national history, to its political parties and assemblies – has sown the seeds of resentment and distrust rather than cross-cultural understanding (Schlessinger, 1991; cf the essays in Taylor et al., 1992; Connolly, 1995; Tully, 1995; Kymlicka, 1998, 2007). The constant focus on cultural, ethnic, and historical identities risks undermining, so it is argued, the necessary conditions for a stable and ultimately just political society, that is, a commitment to a common set of basic individual rights that are justifiable independently of any particular ethnocultural framework (Ignatieff, 1998; Barry, 2001).

Is the future multicultural? In many ways it must be, for the only way to rid societies and states of their diversity is to forcibly remove it, and that entails injustices of the worst kind (with which we are all too familiar). The case against multiculturalism as undermining political stability and the conditions for justice is weak, for it is clearly *not* the case that culturally homogeneous societies are the most just. It is also partly an empirical question, and the evidence so far suggests no strong correlation between multiculturalism and declining support for egalitarian policies (Banting and Kymlicka, 2006). Paradoxically, perhaps, the desire to respect difference and acknowledge alternative sources of moral commitment indeed may lead to the emergence of a kind of cosmopolitan culture 'multiculturalists' often are said to abhor. But if it does, this will not entail having moved 'beyond multiculturalism,' whatever that means. The common culture of justice is multicultural because the people to whom justice must be done and the societies within and between which it must be achieved are multicultural.

See also: Democracy: Normative Theory; Discrimination and the Law; Discrimination: Racial; Diversity and Disagreement in Ethics: Philosophical Implications; Ethnicity and Ethnic Groups: Historical Aspects; Ethnicity, Sociology of.

Bibliography

Appiah, K.A., 2005. The Ethics of Identity. Princeton University Press, Princeton, NJ.
Appiah, K.A., 2007. Cosmopolitanism: Ethics in a World of Strangers. Norton, New York.
Banting, K., Kymlicka, W., 2006. Multiculturalism and the Welfare State: Recognition and Redistribution in Contemporary Democracies. Oxford University Press, Oxford.
Barry, B., 2001. Culture and Equality: An Egalitarian Critique of Multiculturalism. Polity Press, Cambridge.

Bhabha, H.K., 1994. The Location of Culture. Routledge, London.
Buchanan, A., 1991. Secession: The Legitimacy of Political Divorce. Westview Press, Boulder, CO.
Clifford, J., 1988. The Predicament of Culture. Harvard University Press, Cambridge, MA.
Connolly, W.E., 1995. The Ethos of Pluralization. Minnesota University Press, Minneapolis, MN.
Eisenberg, A., Spinner-Halev (Eds.), 2005. Minorities within Minorities: Equality, Rights and Diversity. Cambridge University Press, Cambridge.
Fukayama, F., 1992. The End of History and the Last Man. Free Press, New York.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Habermas, J., 1993. Struggles for recognition in constitutional states. European Journal of Philosophy 1, 128–155.
Hardin, R., 1995. One for All: The Logic of Group Conflict. Princeton University Press, Princeton, NJ.
Hollinger, D., 1995. Postethnic America: Beyond Multiculturalism. Basic Books, New York.
Ignatieff, M., 1998. The Warrior's Honour: Ethnic War and the Modern Conscience. Chatto & Windus, London.
Ivison, D., 2002. Postcolonial Liberalism. Cambridge University Press, Cambridge.
Kukathas, C., 1997. Liberalism, multiculturalism and oppression. In: Vincent, A. (Ed.), Political Theory: Tradition and Diversity. Cambridge University Press, Cambridge, UK.
Kymlicka, W., 1995a. Multicultural Citizenship. Oxford University Press, Oxford, UK.
Kymlicka, W. (Ed.), 1995b. The Rights of Minority Cultures. Oxford University Press, Oxford, UK.
Kymlicka, W., 1998. Finding Our Way: Rethinking Ethnocultural Relations in Canada. Oxford University Press, Toronto, ON.
Kymlicka, W., 2007. Multicultural Odysseys. Navigating the New International Politics of Diversity. Oxford University Press, Oxford.
Levy, J., 1997. Classifying cultural rights. In: Shapiro, I., Kymlicka, W. (Eds.), Ethnicity and Group Rights. New York University Press, New York.
Margalit, A., Raz, J., 1990. National self-determination. Journal of Philosophy 87 (9), 439–461.
Mill, J.S., 1975. On representative government. In: Wolheim, R. (Ed.), Three Essays. Oxford University Press, Oxford, UK.
Miller, D., Walzer, M. (Eds.), 1995. Pluralism, Justice and Equality. Oxford University Press, Oxford, UK.
Moody-Adams, M.M., 1997. Fieldwork in Familiar Places: Morality, Culture & Philosophy. Harvard University Press, Cambridge, MA.
Okin, S.M., 1998. Feminism and multiculturalism: some tensions. Ethics 108, 661–684.
Pogge, T., 1997. Group rights and ethnicity. In: Shapiro, I., Kymlicka, W. (Eds.), Ethnicity and Group Rights. New York University Press, New York.
Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge, MA.
Schlessinger, A., 1991. The Disuniting of America: Reflections on a Multicultural Society. Whittle Communications, New York.
Shachar, A., 2001. Multicultural Jurisdictions: Cultural Differences and Women's Rights. Cambridge University Press, New York.
Smith, A.D., 1986. The Ethnic Origins of Nations. Basil Blackwell, Oxford.
Song, S., 2007. Justice, Gender and the Politics of Multiculturalism. Cambridge University Press, Cambridge.
Spinner-Halev, J., 2012. Enduring Injustice. Cambridge University Press, Cambridge.
Taylor, C., Gutmann, A., Rockefeller, S.C., Walzer, M., Wolf, S., 1992. The Politics of Multiculturalism. Princeton University Press, Princeton, NJ.
Tully, J., 1995. Strange Multiplicity: Constitutionalism in an Age of Diversity. Cambridge University Press, Cambridge.
Tully, J., 2008. Public Philosophy in a New Key. In: Imperialism and Civic Freedom, vol. II. Cambridge University Press, Cambridge.
Turner, T., 1993. Anthropology and multiculturalism: what is anthropology that multiculturalists should be mindful of? Cultural Anthropology 8, 411–429.
Turpel, M., 1991. Home/land. Canadian Journal of Family Law 10, 17–40.
Waldron, J., 1992. Minority cultures and the cosmopolitan alternative. University of Michigan Journal of Law Reform 25, 751–793.
Young, I.M., 1990. Justice and the Politics of Difference. Princeton University Press, Princeton, NJ.

Multiculturalism, Anthropology of

Thomas Hylland Eriksen, University of Oslo, Oslo, Norway

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D.A. Segal, volume 15, pp. 10179–10184, © 2001, Elsevier Ltd.

Abstract

Both a descriptive and a political term, migrating back and forth between politics, journalism, and academic research, multiculturalism cannot easily be defined. However, by common consent it refers simultaneously to ethnically and culturally complex societies and to policies recognizing difference with the aim to achieve equity between the constituent groups of such societies. The relationship between anthropology and multiculturalism is close, but fraught by tension and frictions, since the conceptualization of cultural identities and boundaries presupposed in multiculturalism both emanates from anthropology and has been superseded by theoretical developments in the discipline. Approaches to multiculturalism in anthropology range from studies of indigenous struggles and school curricula in North America to everyday cosmopolitanism in Malaysia and xenophobia in Europe. As an empirical phenomenon it clearly exists, although it is increasingly being challenged; as an analytical concept it is highly debatable.

Introduction

The extent to which historical and even prehistorical societies have been ethnically and culturally complex (and thus 'multicultural') has often been underestimated, but is now well known (Goodenough, 1976; Grillo, 1998). However, in the present context, the term multiculturalism is reserved for contemporary societies where discourses and policies relating to group-based cultural rights are pertinent. In the history of modernity, multiculturalism represents a departure from unilineal views of development, where persistent intrasocietal cultural diversity is acknowledged as a factor worth taking into account in politics.

Nevertheless, multiculturalism is not a simple term with a well-defined meaning. The *Oxford English Dictionary*, tracing its earliest appearance in English to an article about Switzerland published in 1957, defines it as "[t]he characteristics of a multicultural society; (also) the policy or process whereby the distinctive identities of the cultural groups within such a society are maintained or supported." This definition can be a useful starting point, and it notably distinguishes between descriptive and normative aspects of multiculturalism. However, as a matter of fact, very different 'multiculturalisms' are being promoted, which is a main reason why critiques of multiculturalism, usually misleadingly, tend to associate it with either an exaggerated relativistic tolerance of foreign customs and beliefs or an uncritical support of any kind of immigration into the country in question, where immigrants are accorded many rights and few duties. In a review of the term, Stuart Hall (2000) mentions no less than six multiculturalisms: conservative, liberal, pluralist, commercial, corporate, and critical or 'revolutionary' multiculturalism. Each has its approach to the central problem in culturally complex societies, namely, how to reconcile diversity with social solidarity. At the extreme ends of the spectrum are assimilationism (everybody who lives in the same country should have essentially the same culture) and difference multiculturalism (a kind not mentioned by Hall, see Turner, 1993), which demands that society should not be based on one set of values, but should accommodate, recognize the equality of, and indeed celebrate a great variety of cultural values. Vertovec (1998) also distinguishes between eight kinds of multiculturalism, while Delanty (2003) has proposed a list of nine varieties. In practice, most theories of multicultural societies and most state policies concerned with ethnic and cultural diversity try, in different ways, to strike a balance between the extremes of assimilation and segregation. On the one hand, too great a diversity makes solidarity and democratic participation difficult to achieve. On the other hand, total cultural homogeneity is an impossible (and, to most, undesirable) goal to achieve even in ethnically homogenous societies; in any complex society, there will exist religious sects and sexual minorities, to mention only two of the most obvious examples, demanding their right, if possible, to be 'equal but different.'

Seen as a set of policy principles or a political philosophy, multiculturalism is in other words not uniform but diverse, and it is shaped by the specific political, societal, and historical contexts in which it unfolds. One should thus, strictly speaking, talk of multiculturalisms in the plural, but for practical and didactical purposes, it appears reasonable to speak of it in an indefinite or definite singular sense, as the author has chosen to do in the present context. There is a widespread tendency in the literature on multiculturalism, both from its defenders and its opponents, to conflate distinctive social and cultural contexts, and not to distinguish properly between descriptive and prescriptive discourses. Yet, the idea that multiculturalism refers to a homogenous set of ideas and attitudes, is indefensible no matter.

History of Multiculturalism

Although multicultural societies are ancient, and cultural differences have been taken into account by rulers and political élites in many societies, multiculturalism as a coherent ideology is recent. As a tool of political governance, it was initially developed chiefly in Canada in the 1970s as a pragmatic response, influenced by both liberal and communitarian ideas, to rights claims from First Nations and burgeoning separatism among French-Canadians/Québecois, but it was soon applied

to contexts involving immigrants as well. Other countries that began to experiment with multiculturalist policies at an early stage were the United States, the United Kingdom, and Australia. As a rule, countries with considerable ethnic and cultural diversities have more easily adopted multiculturalist policies than others (Banting et al., 2006), since pluralism is part of their political architecture. The strengthening of multiculturalist ideology in many countries, particularly in the 1980s, must also be understood in the context of the rejection of dogmatic versions of Marxism on the political left, which entailed a widespread rejection of class as the main category for analysis as well as political action. A central element in this rejection was the refutation of the view that every form of oppression could be understood on the background of class, and that it would automatically cease to exist in the classless society envisioned by Marxists as the end of history. The new left, which developed from the 1970s, was concerned with a variety of forms of oppression, mobilizing on the basis of different kinds of 'politics of identity' linked to demands of equity and right claims among, for example, indigenous people, regional and immigrant ethnic minorities, women, and homosexuals. With the adoption of multiculturalism, the political left made what could be characterized as a 'cultural turn,' which has contributed strongly to the 'culturalization of politics,' which, according to the philosopher Slavoj Zizek, has marked the last decades (Zizek, 2008). As a consequence, Zizek argues, culture, including religion, has become an increasingly important framework for the understanding of contemporary political and societal development, at the expense of class. Kenan Malik, for his part, describes British multiculturalism as a process that emerged from, and was rooted in, the political and bureaucratic elites (Malik, 2009: p. 41), and Joppke and Lukes (1999: p. 2) claim that only in exceptional cases have oppressed minority persons themselves been the central actors – although one may probably object that the relationship between multiculturalism 'from above' and 'from below' has been shifting and complex in most countries. Canadian philosophers such as Will Kymlicka (1995) and Charles Taylor (1994 [1992]), who proposed intermediate positions or compromises between communitarian and liberal thought, were highly influential theorists in the refining of multiculturalist thought in the 1980s and 1990s.

By the early twenty-first century, departing from a situation where multiculturalism had a central, and in some cases decisive, part in the development of politics and law in the 1990s (Oomen, 2005: pp. 4–13), many countries reached a stage where multiculturalism increasingly was being placed under political pressure and subjected to criticism from various quarters, for contributing to the creation of 'parallel societies' and weakening the collective commitment to shared values and practices (Vertovec and Wessendorf, 2010). Joppke (2004) describes the situation as a total 'retreat' for multiculturalism as theory and policy in liberal states during the last decade. Banting et al. (2006: p. 5) also note that, the critique of and resistance toward multiculturalism formerly appeared to be the domain of the political and intellectual right, whereas it now increasingly appears on the political left, where there are deep concerns as to the effects of multiculturalist policies for socioeconomic equality, social cohesion, women's rights, and the welfare state in contemporary liberal and secular states.

Anthropology's Impact on Multiculturalism

'Multicultural' refers to the coexistence of 'several cultures' within a single society. The concept of culture presupposed in the composite term is the one developed in early to mid-twentieth century cultural anthropology, but with roots in German Romantic thinking, according to which cultures were seen as 'equal but different.' Ninteteenth-century anthropology had drawn upon an evolutionist concept of culture, enabling the ranking of cultures according to their achievements and level of development, whereas social and cultural anthropologists in the next century relied on cultural relativism as a methodological tool. According to this perspective, cultures cannot be ranked on an objective scale, since culturally neutral criteria for such an endeavor do not exist.

At the same time, the evolutionist and relativist perspectives on culture share the premise that human beings are fundamentally shaped by their cultures and that cultures are relatively uniform and bounded. Transposed onto a political canvas, such a conceptualization of culture may underpin nationalism (which it did in nineteenth-century Europe) or multiculturalism. In a context of group-based hierarchical domination, the same concept of culture may form the ideological justification for an ideology like apartheid, and as a matter of fact, anthropologists, notably Werner Eiselen at Stellenbosch University, were among the main intellectual architects of apartheid in South Africa.

The historical connection between the anthropological concept of culture and that of multiculturalist thinkers and policy makers has caused some friction, particularly during the period, since around 1980, when multiculturalism began to grow in significance as an ideological tendency in Western societies. At about the same time as it was being brought into widespread use in policy and politics, the classic concept of culture was increasingly being questioned and criticized within the discipline that made it, namely, anthropology. The criticism concerned the tendency of this concept to create the illusion of 'a global archipelago of cultures' – fixed, timeless, internally coherent – where there was in fact more ambiguity, process, fuzzy boundaries, and change. As a result, a refurbished, some would say poststructuralist, concept of culture (Clifford and Marcus, 1986 is the standard text) could now increasingly be used to criticize the simplifications and unwarranted reifications inherent in multiculturalist conceptualizations (Turner, 1993; Prato, 2009).

Anthropological Approaches to Multiculturalism

Anthropologists working in complex societies have increasingly been compelled to deal with questions pertaining to cultural rights and the politicization of diversity. Research on phenomena such as ethnic resource competition and civil war, cultural property rights, and general processes of exclusion and inclusion in complex societies, although pertaining to cultural diversity, are not considered here, since they are not directly relevant for multiculturalism narrowly defined as a kind of identity politics and/or ideology of governance aiming to achieve equity between culturally different groups without annihilating all cultural differences. This section, therefore,

considers the main areas of concern to anthropologists working on multiculturalist issues.

Postcolonial Societies

Research on multiculturalism has an obvious precursor in the conceptualization and study of plural societies. Originally coined by the geographer Furnivall (1948), the typical plural society was ruled by a colonial empire and peopled by diverse groups, some indigenous, some migrants. The quintessential plural society in the writings of Furnivall, Kuper, and Smith was a plantation society under colonial rule, or at least founded under colonial circumstances, such as could be found in British Burma, Malaya, Natal, Jamaica, or Mauritius (Kuper and Smith, 1969, see also Smith, 1965).

Societal cohesion in the plural society was contingent on external force (the colonial power); the constituent groups had few political rights, and the colonial rulers saw few benefits (for them) in encouraging the growth of a shared national culture. Encounters across ethnic boundaries were limited and largely took place in the market.

Criticized for reifying ethnic boundaries and underestimating the extent of contact across boundaries, the plural society approach nevertheless had its obvious merits in describing a particular kind of colonial society. With decolonization and new demands for broad political participation and citizenship, the coercive colonial ideology was often replaced with democratic slogans and political systems depending on a shared national identity. Attempts to reconcile the existing ethnic pluralism and the vestiges of the colonial, ethnic division of labor, on the one hand, with shared public culture and nationalism on the other, were carried out in many of the erstwhile plural societies, and have been analyzed by anthropologists (see, for example, Nagata, 1979; Eriksen, 1998). The degree to which former plural societies satisfy the criteria of multiculturalism nevertheless varies. To begin with, some are less plural than others; Jamaica is divided by class and color, but to a much lesser extent by ethnicity; Indonesia and Malaysia, although plural in character and multicultural in name, are numerically and politically dominated by one group (the Javanese and the Malay, respectively); Fiji is divided between an indigenous group and an immigrant one (Indians); and so on. However, it has been argued that the original plural colonial societies were precocious experiments in multiculturalism well ahead of the civil rights movements and postwar identity politics of the West. In Mauritius, for example, the French planters were granted cultural rights (to retain their language and religion) by their British conquerors as early as 1810.

Multiculturalism and the New World

There has been considerable anthropological interest in the multiculturalisms of New World countries such as the United States, Canada, and Australia, where questions pertaining to national and cultural identities have long been contested and ambiguous due to the immigration histories of these countries. As pointed out by Hutchinson (1994), major national celebrations of independence anniversaries in Canada (1968), the United States (1976), and Australia (1988) were derailed and subsequently redefined due to massive protests against the Eurocentric and simplistic representations of the national identity first proposed. The growth of cultural rights demands in these societies, dominated by white settlers for centuries, eventually led to calls, sometimes successful, for affirmative action, protective measures to safeguard indigenous cultures, and restructuring of school and university curricula along multiculturalist lines.

In these societies, anthropologists have engaged with the turn toward multiculturalism in interesting and sometimes innovative ways. Notably, for many, working on multiculturalism has represented a rare opportunity to practice engaged anthropology at home. Typically, anthropologists have found themselves in advocacy roles when working with indigenous groups/First Nations or ethnic minorities, but have often taken a more critical stance toward multiculturalist trends in the educational system. Although generally supportive of efforts to make curricula less Eurocentric (and androcentric), anthropologists have often taken exception to tendencies toward essentialization.

Although usually sympathetic toward multiculturalist ideology to the extent that it confirms the antiethnocentric bias inherent in anthropology, anthropologists have rarely defended radical multiculturalist projects involving, for example, polygyny. At least in the North American/Australian context, anthropologists have tended to side with 'critical multiculturalism' aiming to achieve equality for all groups in a multiethnic state, but more rarely with 'difference multiculturalism' (Turner's, 1993, concepts) where the ultimate goal is social, cultural, and moral separatism. It may be noted here that anthropologists in Brazil, a country with a comparable but in important respects different history, have moved in a different direction. Following the lead of Viveiros de Castro (1992) and others, contemporary Brazilian anthropology of cultural difference 'at home' tends to emphasize the continued existence of radical difference and speaks more of land rights than of reading lists.

Migrants in Europe

Owing to important historical differences, the anthropology of multiculturalism in Europe tends to have a different focus from that in the New World. Moreover, as correctly pointed out by Kymlicka (1995), the politics of identity among indigenous peoples is not simply replicated in immigrant groups, many of which represent very different interests from the indigenous movement. Demands for territorial and cultural autonomy have been crucial for most indigenous groups, whereas immigrants typically demand equal treatment and nondiscrimination. Within this ideological discourse, differential treatment (e.g., in the health service or educational system) should ultimately aim. However, especially since around 1990, many immigrant groups in the West, especially those of Muslim religion, have demanded recognition for their cultural specificities as well, and in some countries, issues to do with the language of instruction in schools are vigorously debated.

While it is generally acknowledged in the New World that those societies are historically immigrant societies, and that continued immigration – although it may be stemmed and

limited – will take place, European societies are deeply divided over issues to do with immigration and the integration of immigrants. In fact, much of the current anthropological research on multiculturalism in Europe has racism, discrimination, and xenophobia as its main focus (e.g., Holmes, 2000; Bowen, 2007), and in addition, the 'failed multiculturalism' of Yugoslavia and the subsequent horrors of 'ethnic cleansing' has its own, somber literature (Bringa, 1995; Bowman, 2004). Indeed, the emerging polarization in European societies around migration and ethnic diversity, punctuated by a scattering of terrorist attacks from both militant Islamists and right-wing xenophobes in the 2000s, has led several political leaders to denounce the term multiculturalism. This denunciation has itself spawned an academic literature, notably Lentin and Titley (2011) and Vertovec and Wessendorf (2010), which asks about the causes and effects of the 'crisis of multiculturalism.' While Lentin and Titley focus mainly on the rise of anti-immigrant sentiment, Vertovec and Wessendorf explore the semantics of multiculturalism and the relationship between political rhetoric and media talk on the one hand, and practices on the other. They conclude that most multiculturalist policies continue in spite of the pessimistic turn, but increasingly under the epithet 'diversity' rather than the largely discredited term 'multiculturalism.'

Multiculturalism and Globalization

A different approach to multiculturalism is offered by the literature on globalization. As pointed out repeatedly by anthropologists working in the field (Hannerz, 1996; Appadurai, 1996), globalization does not lead to uniformity, but to a reconfiguration of difference. The phenomenal growth in collective assertions of identity and claims to difference, which has been a worldwide trend since the latter decades of the twentieth century, must nevertheless be understood on the background of perceived challenges to cultural uniqueness and selfhood.

Assertions of cultural uniqueness as responses to encroaching homogenization may find expression in nationalism or collective withdrawal, but may also be expressed as celebrations of diversity and of 'rainbow societies' of various kinds. The latter kind of multiculturalism, with its focus on esthetic and commercial aspects rather than political ones, moreover has a ready market in the global tourism industry, where the demand for a sanitized and harmless 'cultural otherness' is on the rise (cf Comaroff and Comaroff, 2009).

As with the other forms of multiculturalism considered, the culture concept utilized here is reifying, limited, and selective. It presupposes an archipelago metaphor for the world of cultures – bounded, homogenous, and mostly unchanging – where a processual conceptualization emphasizing overlaps, change, and impurities would have been more descriptively appropriate. However, the standard anthropological response to this would not be to castigate multiculturalists of all kinds for using a dated and inadequate concept of culture, but to study how this static and reified culture concept is put to use – and how it gives existential meaning to people, is drawn into political struggles and converted into a marketable commodity.

The Challenge of Normativity

Much of the extant literature on multiculturalism, academic and nonacademic, has a clear normative slant. It is engaged in the sense of trying to make a political difference. Although this is also true of some of the anthropological contributions, explicit normativity is problematic from the perspective of methodological cultural relativism, where the noncommitted, neutral, comparative, and analytical gaze is privileged and seen as virtuous. As Turner points out in his seminal article about anthropology and multiculturalism:

> *Multiculturalism*, unlike anthropology, is primarily a movement for change. To the extent that it has developed a theoretical analysis, it is primarily a conceptual framework for challenging the cultural hegemony of the dominant ethnic group (or the dominant class constituted almost exclusively by that ethnic group) Ö by calling for equal recognition of the cultural expressions of nonhegemonic groups within the educational system. (Turner, 1993: p. 412)

Writing about the implementation of multicultural curricula at Florida State University in the early 1990s, the anthropologists Paredes and Pohl (1995) describe their experience of the university politics involved. While their narrative confirms Turner's general description, they add a dimension that is often missing from the literature on recognition and multiculturalism, namely, class and power. First of all, they point out that multiculturalism, as a political project, is not really about the pursuit of intercultural knowledge, but 'about the political empowerment of historically subordinate segments of our own society or, more precisely, their admission into the ranks of the elite' (Paredes and Pohl, 1995: p. 199, see also Gutmann, 1999). They later conclude that multiculturalism, as it is dealt with in the American academy, is an aspect of high culture, and has nothing to do with the totality of culture, and as such, it differs not only descriptively but also normatively from the project of anthropology, which entails taking all aspects of culture equally seriously to the extent that they are meaningful and important to the people in question.

While this delineation of multiculturalism is narrowly limited to certain discourses in North America, the general points about class and normativity have a broader relevance. And, as has been made clear, because of the misleading conceptualization of culture, many anthropologists who do have a normative commitment still have mixed feelings about the multiculturalist project. It has been described, with reference to multiculturalist state policies in Mauritius, as 'apartheid with a friendly face' (Eriksen, 1997) since a paradox of a multiculturalist ideology aiming to recognize all groups under the heading 'unity in diversity,' is the fact that persons who do not define themselves primarily as members of ethnic groups, have no legitimate place in society. Baumann (1999) addresses this and other 'riddles' of multiculturalism in a book that advocates, on the basis of long-term fieldwork in an ethnically complex suburb of London, a pragmatic solution whereby both nationalists, ethnicists and religious devotees can be recognized as equals. Baumann also addresses the question of the

'cultural hybrids' who do not fit easily into an ethnic or national category.

Conclusion: Beyond Multiculturalism

Especially in North America, but also elsewhere, the classic concept of culture has been used for what it was worth in domestic identity politics, leading in some cases to controversial policies of multiculturalism, where individuals were endowed with particular rights according to their ethnic membership. Critics point out that multiculturalism in some of its versions resemble apartheid and also that by positing a simple one-to-one relationship between ethnic identity and culture, it not only encourages a 'disuniting of America' (Schlesinger, 1998), but also contributes to reifying misleading notions of culture seen as homogeneity, like so many nationalisms writ small. It may seem that multiculturalism has grave shortcomings in two respects: it is descriptively inadequate and misleading; and normatively, it may be a recipe for divisive identity politics through its emphasis on group membership as the basis of identification.

In practice, the potential separatism and groupness implied in multiculturalism tends to be counterbalanced by individualism and overarching practices and ideas. Still, the conceptual and descriptive problems with multiculturalism have prompted anthropologists to refine their conceptual toolbox when describing ethnically complex societies.

One alternative concept, which is gaining currency in local administration and politics in some societies, is interculturalism (cf Asgarally, 2005 with reference to Mauritius). Instead of stressing boundaries, as multiculturalism does, this concept emphasizes connectedness and relatedness.

Another term is superdiversity. Noting that many cities across the world had become, in the early twenty-first century, diverse in new ways, Vertovec (2006) coined this term. His point was that mobility had made the ethnic maps of contemporary cities more unstable and volatile than previously; that temporary workers, students, tourists, asylum seekers, and various intermediate groups, plus the internal mobility within the cities, created a relatively unpatterned urban space that was less stable and regular than terms like 'pluralism' or 'multiculturalism' might imply.

Various conceptualizations of cultural hybridity and ambiguous identities are also posited against the implied essentialism of multiculturalism (Prato, 2009; Hesse, 2000), where impurities, mixing, and even individual idiosyncrasies are emphasized as opposed to the boundedness of the 'many cultures' of multiculturalism.

Finally, there may be sound arguments for replacing 'multiculturalism' with 'diversity.' The latter term is, of course, less specific than the former, but it may turn out to be both more adequate as a descriptive term and more versatile in political contexts, since it does not a priori assume that the social world consists primarily of named cultural groups, but avoids fixation and leaves the options open, to paraphrase Bauman (1996).

See also: Advocacy in Anthropology; Cosmopolitanism; Culture: Contemporary Views; Ethnic, Racial, and Nationalist Movements; Ethnicity: Anthropological Aspects; Globalization, Anthropology of; Nationalism, Sociology of; Race and Racism in the Twenty-First Century.

Bibliography

Appadurai, A., 1996. Modernity at Large: Cultural Dimensions of Globalization. University of Minnesota Press, Minneapolis, MN.
Asgarally, I., 2005. L'interculturel ou la guerre. Editions MSM, Port-Louis.
Banting, K., et al., 2006. Do multiculturalism policies erode the welfare state? An empirical analysis. In: Banting, K., Kymlicka, W. (Eds.), Multiculturalism and the Welfare State: Recognition and Redistribution in Contemporary Democracies. Oxford University Press, Oxford, pp. 49–91.
Bauman, Z., 1996. From pilgrim to tourist; or a short history of identity. In: Hall, S., Du Gay, P. (Eds.), Questions of Cultural Identity. SAGE, London, pp. 18–36.
Baumann, G., 1999. The Multicultural Riddle: Rethinking National, Ethnic and Religious Identities. Routledge, New York.
Bowen, J., 2007. Why the French Don't Like Headscarves: Islam, the State, and Public Space. Princeton University Press, Princeton, NJ.
Bowman, G., 2004. Xenophobia, fantasy and the nation: the logic of ethnic violence in Yugoslavia. In: Goddard, V., Llobera, J., Shore, C. (Eds.), Anthropology of Europe: Identity and Boundaries in Conflict. Berg, London, pp. 143–171.
Bringa, T., 1995. Being Muslim the Bosnian Way: Identity and Community in a Central Bosnian Village. Princeton University Press, Princeton, NJ.
Clifford, J., Marcus, G. (Eds.), 1986. Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley, CA.
Comaroff, J., Comaroff, J.L., 2009. Ethnicity, Inc. University of Chicago Press, Chicago, IL.
Delanty, G., 2003. Community. Routledge, London.
Eriksen, T.H., 1997. Multiculturalism, individualism and human rights. In: Wilson, R.A. (Ed.), Human Rights, Culture and Context. Pluto, London, pp. 49–69.
Eriksen, T.H., 1998. Common Denominators: Ethnicity, Nation-building and Compromise in Mauritius. Berg, Oxford, UK.
Furnivall, J.H., 1948. Colonial Policy and Practice: A Comparative Study of Burma and the Netherlands India. Cambridge University Press, Cambridge, UK.
Goodenough, W.H., 1976. Multiculturalism as the normal human experience. Anthropology and Education Quarterly 7 (4), 4–7.
Grillo, R., 1998. Pluralism and the Politics of Difference: State, Culture and Ethnicity in Comparative Perspective. Clarendon, Oxford, UK.
Gutmann, A., 1999. Democratic Education. Princeton University Press, Princeton, NJ.
Hall, S., 2000. The multicultural question. In: Hesse, B. (Ed.), Un/Settled Multiculturalisms: Diasporas, Entanglements, Transruptions. Zed Books, London, pp. 209–240.
Hannerz, U., 1996. Transnational Connections. Routledge, London.
Hesse, B. (Ed.), 2000. Un/Settled Multiculturalisms: Diasporas, Entanglements, Transruptions. Zed Books, London, pp. 209–240.
Holmes, D., 2000. Integral Europe – Fast Capitalism, Multiculturalism, Neofascism. Princeton University Press, Princeton, NJ.
Hutchinson, J., 1994. Modern Nationalism. Fontana, Glasgow.
Joppke, C., 2004. The retreat of multiculturalism in the liberal state: theory and policy. The British Journal of Sociology 55 (2), S237–S257.
Joppke, C., Lukes, S. (Eds.), 1999. Multicultural Questions. Oxford University Press, Oxford, UK.
Kuper, L., Smith, M.G. (Eds.), 1969. Pluralism in Africa. University of California Press, Berkeley, CA.
Kymlicka, W., 1995. Multicultural Citizenship: A Liberal Theory of Minority Rights. Clarendon Press, Oxford, UK.
Lentin, A., Titley, G., 2011. The Crisis of Multiculturalism? Racism in a Neoliberal Age. Zed Books, London.
Malik, K., 2009. From Fatwa to Jihad: The Rushdie Affair and Its Legacy. Atlantic Books, London.
Nagata, J., 1979. Malaysian Mosaic: Perspectives from a Polyethnic Society. University of British Columbia Press, Vancouver, BC.
Oomen, B., 2005. Chiefs in South Africa: Law, Power and Culture in the Post-apartheid Era. James Currey, Oxford, UK.

Paredes, A., Pohl, M., 1995. Anthropology and multiculturalism in a university curriculum: a case study. Critique of Anthropology 15, 193–202.
Prato, G.B., 2009. Beyond Multiculturalism: Views from Anthropology. Ashgate, London.
Schlesinger, A.M., 1998. The Disuniting of America: Reflections on a Multicultural Society, revised ed. W W Norton, New York.
Smith, M.G., 1965. The Plural Society in the British West Indies. University of California Press, Berkeley, CA.
Taylor, C., 1992. Multiculturalism and the Politics of Recognition. Princeton University Press, Princeton, NJ.
Turner, T., 1993. Multiculturalism and anthropology: what is anthropology that multiculturalists should be mindful of it? Cultural Anthropology 8 (4), 411–429.
Vertovec, S., 1998. Multi-multiculturalisms. In: Martiniello, M. (Ed.), Multicultural Policies and the State. ERCOMER, Utrecht, pp. 25–38.
Vertovec, S., 2006. The Emergence of Super-diversity in Britain. COMPAS Working Papers, 25. Oxford University.
Vertovec, S., Wessendorf, S. (Eds.), 2010. The Multiculturalism Backlash: European Discourses, Policies and Practices. Routledge, London.
Viveiros de Castro, E., 1992. From the Enemy's Point of View: Humanity and Divinity in an Amazonian Society. University of Chicago Press, Chicago, CA.
Zizek, S., 2008. In Defense of Lost Causes. Verso, London.

Multidimensional Scaling I

Kwanghee Jung, University of Texas Health Science Center at Houston, Houston, TX, USA
Yoshio Takane, University of Victoria, Victoria, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J. de Leeuw, volume 20, pp. 13512–13519, © 2001, Elsevier Ltd.

Abstract

Multidimensional scaling (MDS) is a set of data-analytic tools for deriving a graphical representation of objects in a multidimensional space based on proximity relations among them. By the graphical representation, we gain intuitive understanding of the regularity governing the relationships among the objects. In this article, we introduce the basic concept and models of MDS along with its essential ingredients such as distance models, proximity data, fitting criteria, dimensionality selection. To illustrate the use of MDS, three useful MDS procedures (simple MDS, individual differences MDS, and unfolding analysis) are presented with empirical examples.

The Basic Concept of MDS

The notion of proximity (e.g., similarity/dissimilarity) among objects plays an important role in human cognition and development. It is a fundamental principle by which people categorize, generalize, and classify objects (Takane et al., 2009). Multidimensional scaling (MDS) is a set of statistical techniques to derive a spatial map of objects based on a set of distancelike quantities among the objects. MDS represents objects as points in a multidimensional space in such a way that points corresponding to similar objects are located close together, while dissimilar objects are located far apart. The derived map from MDS facilitates our intuitive understanding of proximity relations among the objects by identifying meaningful directions or regions (object clusters) in the space.

To illustrate the use of MDS, let us take a look at similarity rating data among eight US cities presented in **Table 1**. The data were collected from 10 university students, who were asked to indicate the degree of similarity between each pair of cities using an 11-point rating scale; a larger value indicates more similar cities. All participants completed a questionnaire with 28 pairs drawn from the eight US cities. The eight cities were (1) Seattle, (2) New York, (3) Miami, (4) Chicago, (5) San Diego, (6) Atlanta, (7) New Orleans, and (8) Denver. PROXSCAL (Busing et al., 1997) implemented in Statistical Packages for Social Sciences (SPSS) was used to analyze the data, assuming that all participants have an identical object configuration. (this case is called simple MDS),

Figure 1 displays the derived two-dimensional configuration of the eight US cities. The figure shows that the eight cities are roughly classified into four groups, the first group consisting of Seattle and Denver, the second consisting of Chicago and New York, the third consisting of Atlanta and New Orleans, and the fourth consisting of San Diego and Miami. Dimension 1 (the horizontal direction) contrasts New Orleans, Atlanta, New York, and Chicago with Denver, San Diego, and Seattle. We may call the horizontal direction the east–west dimension from right to left. Dimension 2 (the vertical direction), on the other hand, separates Seattle, New York, Chicago, and Denver from New Orleans, Miami, San Diego, and Atlanta. We may call the vertical direction the north–south dimension from top to bottom. It is interesting to note that participants' perception of similarity among the cities seems to be, to a large extent, determined by their geographic locations. This may be because geographic locations are a truly important factor occupying participants' mind in judging similarity among the cities, or it may be because cities located geographically close have similar climate and culture (atmosphere, people's life style, etc.).

Distance Models

As illustrated above, MDS represents interobject similarities/dissimilarities by interpoint distances. The simple Euclidean distance model is most frequently used for this purpose due to its familiarity in our everyday life. This is also the model used to derive the configuration given in **Figure 1**. For other distance models such as the city-block distance model, see Carroll and Arabie (1998). The Euclidean distance between object points i and j (d_{ij}) is calculated by

$$d_{ij} = \left[\sum_{m=1}^{M} \left(x_{im} - x_{jm} \right)^2 \right]^{1/2}, \qquad [1]$$

where x_{im} and x_{jm} denote the coordinates of object points i and j on dimension m, respectively, and M indicates the dimensionality (the number of dimensions) of the space. Once the coordinates of object points are obtained (through MDS), we can locate the objects in the space using a Cartesian coordinate system. Note that the Euclidean distance is invariant over the choice of origin and orientation of coordinate axes. That is, the distance does not change if the origin of the space is shifted, or the coordinate axes are rotated. It is customary to place the origin at the centroid of the object configuration, and set the coordinate axes in the principal axis directions. In principal axis directions, dimension 1 accounts for the largest possible variance among object coordinates, dimension 2 accounts for the next largest variance, given that it is orthogonal to the first dimension, and so on. The principal axis directions are not necessarily most easily interpretable, and if necessary, they may be changed to other coordinate axes having easier interpretations. The configuration displayed in **Figure 1** was in fact rotated about 60° clockwise from the principal axis solution provided by PROXSCAL to facilitate interpretations.

Table 1 Similarity ratings among 8 US cities by 10 participants (1: very dissimilar, 11: very similar)

Participant	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*Participant*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
1	4							6	9						
	6	6							2	9					
	6	9	7						10	11	9				
	6	5	6	6					3	8	11	5			
	6	8	9	7	6				2	10	9	11	9		
	5	2	7	5	5	8			2	8	3	5	7	11	
	3	2	2	5	3	2	2		10	1	1	6	2	5	3
2	4							7	5						
	2	3							1	5					
	6	7	4						3	8	4				
	5	4	7	3					7	6	9	6			
	2	4	6	3	4				1	3	10	8	8		
	3	2	6	4	5	7			2	4	9	5	7	10	
	6	3	1	4	3	4	3		10	7	5	8	4	4	3
3	7							8	7						
	5	6							4	7					
	7	10	8						8	10	6				
	6	5	5	4					8	4	8	5			
	6	5	7	6	5				8	5	8	7	4		
	5	5	7	5	6	8			8	3	9	5	8	8	
	6	6	4	5	6	5	4		8	1	6	3	5	7	7
4	4							9	8						
	3	7							5	6					
	4	10	2						8	9	5				
	5	7	8	4					5	5	7	5			
	4	4	7	5	8				6	7	7	6	6		
	3	3	5	4	5	8			3	4	7	4	5	7	
	5	5	4	6	3	5	3		7	6	6	7	3	4	5
5	4							10	6						
	1	3							5	6					
	2	9	4						9	6	3				
	2	1	6	2					8	5	7	5			
	1	5	3	7	2				7	4	5	9	5		
	1	2	6	2	2	3			3	2	4	5	4	6	
	8	1	1	2	1	2	1		4	3	2	5	4	6	6

The eight US cities are (1) Seattle, (2) New York, (3) Miami, (4) Chicago, (5) San Diego, (6) Atlanta, (7) New Orleans, and (8) Denver.

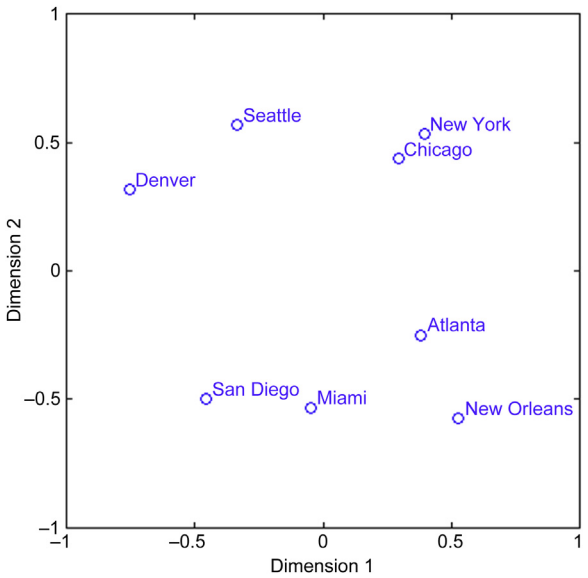

Figure 1 The two-dimensional object configuration of the eight US cities' data.

Proximity Data and Data Transformations

A variety of proximity data may be used in MDS, arising from a variety of data collection procedures, such as similarity ratings, sorting, same–different judgment, frequency of cooccurrences, reaction time, frequency of social interactions, and profile similarity. In similarity rating methods which are used to obtain the data in Table 1, objects are presented in a pair to the participants who are asked to rate the degree of similarity between them on a rating scale. In the sorting method, participants are given a set of objects and are asked to group them into several groups in terms of their similarity. The number of times two objects are put into same groups is counted over a group of participants and is used as a similarity measure between the objects. It is noteworthy that different proximity measures may reflect different aspects of proximity relations among objects, resulting in different configurations of the objects. For more systematic descriptions of the data collection methods and the related issues, see Takane et al. (2009).

Whichever data collection methods may be used, it is unlikely that observed proximity data can be directly

approximated by a distance model. The data may be similarity data (as in the data given in Table 1), which are inversely related to distances. In this case, the observed similarity data must be inversely monotonically transformed in order to be fitted by a distance model. Even if the data are dissimilarity data, they may not be linearly related to underlying distances, and need to be transformed before they can be fitted by a distance model. The problem of how the observed proximity data are related to the underlying distances is generally referred to as the scale level of measurement. Different scale levels require different types of transformations of proximity data.

Four scale levels are typically distinguished in MDS: ratio, interval, log-interval, and ordinal. These distinctions are important because certain MDS procedures are only appropriate for certain scale levels. MDS procedures assuming a ratio, interval, or log-interval scale are called metric MDS, while others assuming an ordinal scale are called nonmetric MDS (Shepard, 1962; Kruskal, 1964a,b). Let o_{ijk} denote the observed proximity between objects i and j for participant k. In the ratio scale, proximity data are assumed approximately related to distances (d_{ij}) by $a_k o_{ijk} \approx d_{ij}$, where a_k is positive (this is called a similarity transformation). In the interval-scaled measurement, it is assumed that $a_k o_{ijk} + b_k \approx d_{ij}$, where a_k is positive if the data are dissimilarity data, and negative if they are similarity data, and b_k is an arbitrary constant (this is called an affine transformation). In the log-interval scale, it is assumed that the observed proximity measures are related to underlying distances by a power transformation, $b_k o_{ijk}^{a_k} \approx d_{ij}$. If the log is taken on both sides, we obtain $a_k \ln o_{ijk} + \ln b_k \approx \ln d_{ij}$, which is a linear relationship between $\ln o_{ijk}$ and $\ln d_{ij}$, thus the name log-interval scale. In the ordinal scale level, o_{ijk} and d_{ij} are assumed to be only monotonically related. That is, $o_{ijk} > o_{i'j'k}$ implies $d_{ij} > d_{i'j'}$ if the data are dissimilarity data, whereas $o_{ijk} > o_{i'j'k}$ implies $d_{ij} < d_{i'j'}$ if the data are similarity data. In both cases, the transformation of the data can be expressed as $t_k(o_{ijk}) \approx d_{ij}$, where t_k is a monotonically increasing function if the data are dissimilarity, and it is a monotonically decreasing function if the data are similarity.

In empirical research, we usually do not know exact scale levels of the observed proximity data in advance. As a practical strategy, we may start with a less strict assumption (e.g., ordinal scale), and then may switch to a stronger assumption (e.g., interval–ratio scale) if appropriate. This may help avoid applying untenable scale level assumptions to the observed proximity data.

In the example given above, the data are similarity rating data. Analysis was carried out under the assumption that the data are ordinal (an inverse monotonic transformation) with the option that tied observations may be untied after the transformation.

Fitting Criteria

In practical data-analytic situations, an exact representation of the observed proximity data by a distance model is usually impossible due to measurement errors. MDS systematically searches for an object configuration and an optimal data transformation that minimizes the discrepancy between the transformed proximity data and the distance predictions. The discrepancy is quantified as a fit index (e.g., the normalized stress), which is also used as a fitting criterion to be minimized for estimating parameters in the model. Two broad classes of discrepancy functions have been used in MDS. One is the least squares (LS) criterion (Kruskal, 1964a,b), and the other is the maximum likelihood criterion (Ramsay, 1977, 1982). Although the maximum likelihood criterion has some appeals for its statistical inference capabilities with data with a large sample size, the LS criterion has been far more predominantly used in MDS for its simplicity and flexibility, requiring no rigid distributional assumptions. The LS criterion is stated as,

$$\varnothing = \sum_{k=1}^{K} \sum_{i<j} \left[t_k(o_{ijk}) - d_{ij} \right]^2, \qquad [2]$$

where t_k indicates a generic data transformation function, whose specific form depends on the scale level of proximity data, as described in the previous section. The LS MDS attempts to find the set of object coordinates x_{im} ($i = 1, \ldots, n$, where n is the number of objects and $m = 1, \ldots, M$), and data transformation parameters that minimize the discrepancy between the transformed proximity data $[t_k(o_{ijk})]$ and predicted distances (d_{ij}). Criterion [2] usually cannot be minimized in closed form, and some sort of iterative algorithms should be used (see Borg and Groenen (2005) for more details of the optimization algorithms used in MDS). The LS criterion used in MDS may have multiple local minima, and iterative optimization procedures may converge to one of them that is not the true minimum of the criterion we wish to find. This is called the local minima problem. To circumvent this difficulty, we may obtain multiple solutions with many different initial estimates and choose the best solution among them. In all results of MDS reported in this article, solutions were obtained with 100 random initial starts.

Dimensionality Selection

Dimensionality refers to the number of coordinates needed to locate objects in a multidimensional space. Two considerations should primarily be taken into account in determining the number of dimensions in MDS. First, the derived object configuration has to fit to the data at hand reasonably well. One practical strategy is to analyze the data under varied dimensionalities, say from 1 to 4, plot the fit value against the dimensionality (called a scree plot), and identify the point where the improvement in fit flattens out (called an elbow). Another important consideration is the interpretability of derived dimensions. We should retain only the meaningful dimensions that contribute to uncovering the underlying structure of proximity relations among objects.

In the example given above, the normalized raw stress values were 0.089 for the unidimensional solution, 0.034 for the two-dimensional solution, 0.013 for the three-dimensional solution, and 0.008 for the four-dimensional solution. The two-dimensional solution was considered optimal, where the improvement in fit flattened out.

Individual Differences MDS

As in the example given above, proximity data are typically collected from a number of participants. In the previous analysis, however, it was assumed that there were no systematic individual differences in the way the participants perceive similarities among the cities. This assumption may not be justified. There may be some interesting differences among the participants in their similarity judgments. Individual Differences MDS (ID MDS) was designed to capture both commonality and individual differences in similarity judgments (Carroll and Chang, 1970). ID MDS assumes that there is a common spatial representation of objects that applies to all individuals, while individual differences arise from differential weighting of common dimensions by different participants. It employs the weighted Euclidean distance model, which is written as,

$$d_{ijk} = \left[\sum_{m=1}^{M} w_{km} \left(x_{im} - x_{jm} \right)^2 \right]^{1/2}, \qquad [3]$$

where d_{ijk} is the distance between objects i and j for individual k, x_{im}, and x_{jm} are, as before, the coordinates of objects i and j on dimension m in the common object configuration, and w_{km} is the weight attached to dimension m by individual k. The d_{ij} in eqn [2] is replaced by d_{ijk} given above. Note that the model [3] reduces to the simple Euclidean distance model in eqn [1], if there is no individual differences in an object configuration, i.e., w_{km} is unity for all k and m. Using the weighted Euclidean distance model, ID MDS estimates both the common object coordinates (x_{im}) and the individual differences weights (w_{km}) in such a way that the predicted distance (d_{ijk}) best matches the optimally transformed proximity data. To eliminate the scale indeterminacy between the common object configuration and the individual differences weights, the former is typically constrained to satisfy $\sum_{i=1}^{n} \frac{x_{im}^2}{n} = 1$ for $m = 1, \ldots, M$ where n is the number of objects. One important characteristic of the weighted Euclidean distance model is that there is no rotational indeterminacy in contrast with the simple Euclidean model. This means that the derived dimensions have intrinsic meanings, and we should interpret them as they are.

For illustration, we applied ID MDS to the eight US Cities data discussed earlier. Again, PROXSCAL was used to analyze the similarity data, but the weighted Euclidean distance model was applied this time. The normalized raw stress also indicated that the two-dimensional solution is best (the normalized raw stress values were 0.089 for the unidimensional solution, 0.031 for the two-dimensional solution, 0.009 for the three-dimensional solution, 0.005 for the four-dimensional solution). The derived two-dimensional common configuration was very similar to the one presented in Figure 1, and thus we omitted the presentation of the common configuration obtained from ID MDS (the only remarkable difference is in the overall size of the configuration due to the normalization constraint discussed above). Recall that the horizontal axis (dimension one) contrasts the east and the west, while the vertical axis (dimension two) contrasts the north and the south. Figure 2 depicts the individual differences weights attached to the two dimensions by different individuals. Although there are overall no big individual differences in this example, three noticeable participants are identified by numbers, 1, 9, and 10. Participant 1 puts more emphasis on dimension 1 than dimension 2, whereas participant 9 puts slightly more emphasis on dimension 2 than dimension 1. Participant 1 is more sensitive to the east–west difference, while participant 9 is opposite. Participant 10 puts almost equal emphasis on both dimensions. As can be seen, the weights may show interesting patterns of individual differences, which in turn may be related to participants' backgrounds (e.g., age, gender). For more examples of interesting applications of the ID MDS, see Takane (2007).

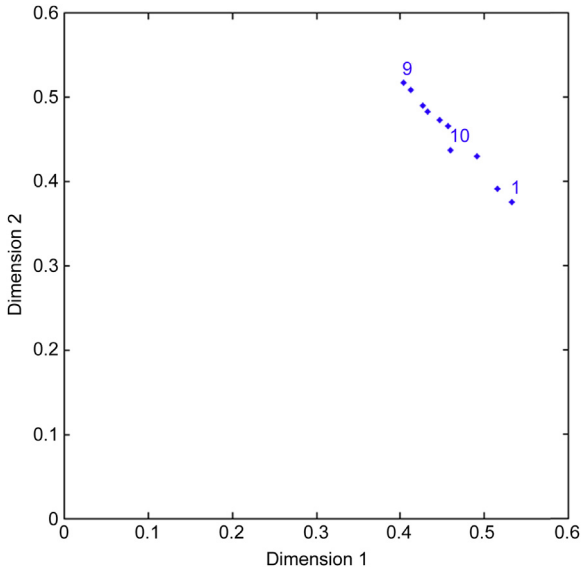

Figure 2 The plot of individual differences weights attached to the two dimensions by the 10 participants.

Unfolding Analysis

Unfolding analysis is a variant of MDS used to analyze individual differences in preference judgments (Coombs, 1964). As an example of preference data, let us look at **Table 2**. A group of 23 participants were asked to rank order 6 different cars from the most preferred to the least preferred. The six cars used were Honda Civic, BMW 328 Sedan, Mercedes C-Class Sedan, Lincoln Navigator, Toyota Camry, and Jeep Grand Cherokee. Tied ranks were allowed by assigning the average of ranks to them. The data are actually antipreference data, where a smaller value indicates more preference.

In unfolding analysis, each participant is assumed to have an ideal object represented as an ideal point in the same space as actual objects are represented. The distances between the ideal point and the actual object points are assumed to be inversely related to the participant's preferences on the objects; more preferred objects are located closer to the participant's ideal points, while less preferred objects are located further apart from the ideal points. Let x_{im} denote the coordinate of object i on dimension m, and y_{km} the coordinate

Table 2 Preference ranking among 6 cars by 23 participants (1: the most preferred, 6: the least preferred)

Participant	*Honda Civic*	*BMW 328 Sedan*	*Mercedes C-Class Sedan*	*Lincoln Navigator*	*Toyota Camry*	*Jeep Grand Cherokee*
1	3	4	6	2	1	5
2	3	6	1	5	2	4
3	5	1	2	6	3	4
4	1	3	4	5	2	6
5	4	2	1	6	3	5
6	1	4	3	5	2	6
7	2	4	3	1	5	6
8	1	3	4	5	2	6
9	5	3	2	4	1	6
10	5	1	3	2	6	4
11	5	1	1	6	6	4
12	5	4	6	3	1	2
13	1	3	5	4	2	6
14	1	3	2	6	5	4
15	5	1	2	6	4	3
16	2	1	3	6	4	5
17	6	3	2	4	5	1
18	1	4	3	5	2	6
19	3	6	5	2	4	1
20	5	1	2	3	4	6
21	2	1	3	6	4	5
22	6	4	3	1	5	2
23	5	2	4	3	6	1

The six cars are Honda Civic, BMW 328 Sedan, Mercedes C-Class Sedan, Lincoln Navigator, Toyota Camry, and Jeep Grand Cherokee.

of individual k's ideal point on dimension m. The Euclidean distance between actual object i and individual k's ideal point is calculated by

$$d_{ik} = \left[\sum_{m=1}^{M} (x_{im} - y_{km})^2 \right]^{1/2}. \qquad [4]$$

The coordinates of the actual objects and ideal points are determined in such a way that the preference values of the objects for particular individual are a decreasing function of the distance between them. This implies that the observed antipreference data are interpreted as dissimilarity data between participants' ideal objects and actual objects. As a result, we obtain two coordinate matrices, one for actual object points, and the other for participants' ideal points. The two coordinate matrices are used to derive a joint plot by superposing actual objects and participants' ideal points in a joint multidimensional space.

The data in **Table 2** were analyzed by PREFSCAL (Busing et al., 2005) implemented in SPSS. An optimal solution was obtained with 100 random initial starts under the assumption of ordinal scale. The normalized stress indicated that the two-dimensional solution is optimal (the normalized stress values were 0.176 for the unidimensional solution, 0.043 for the two-dimensional solution, 0.004 for the three-dimensional solution). **Figure 3** presents the joint MDS configuration of the participants' ideal points and the six cars. The participants were identified by the integers from 1 to 23. Note that a car is located close to participants' ideal

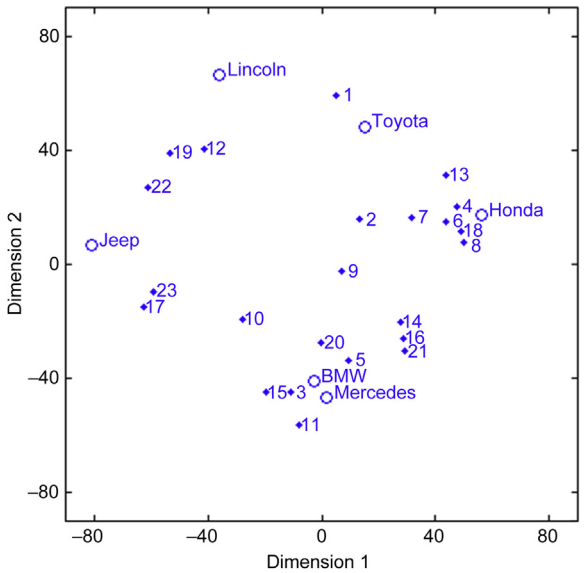

Figure 3 The joint plot of object points and participants' ideal points for the preference data on six cars. The six cars are Honda Civic, BMW 328 Sedan, Mercedes C-Class Sedan, Lincoln Navigator, Toyota Camry, and Jeep Grand Cherokee.

points if it is preferred by the participants. For example, participants 4, 6, 8, and 18 are located close to Honda Civic because these participants preferred Honda Civic to any other cars, while participants 3, 5, 11, 15, and 20 are located close to BMW 328 Sedan and Mercedes C-Class Sedan, preferred to any other cars. **Figure 3** also shows that the six cars are roughly classified into three groups, the first one consisting of Toyota Camry and Honda Civic, the second consisting of BMW 328 Sedan and Mercedes C-Class Sedan, and the third consisting of Lincoln Navigator and Jeep Grand Cherokee. The within-group cars have much in common, compared to the between-group cars. For example, Toyota Camry and Honda Civic are much more similar than others in that they are economic cars produced by Japanese manufacturers. As illustrated, unfolding analysis can be a very useful technique in marketing and consumer behavior research in that it allows us to understand patterns of individual differences in preference judgments on commercial products. The characteristics of individual differences in preference judgments can be further investigated by relating to product features and participants' background information. This sort of analysis may help marketing analysts to develop practical marketing strategies such as brand and product positioning. For more examples of application of unfolding analysis, refer to Takane (2007).

A Summary and Software for MDS

MDS is a class of statistical tools to visualize the relationships among objects in a spatial map, based on proximity and/or preference measures on the objects of interest. In this article, we addressed essential ingredients of MDS, such as the distance model, proximity data, fitting criteria, and

dimensionality selection. We also examined three useful models of MDS and their applications (i.e., simple MDS, ID MDS, and unfolding analysis).

In behavioral and social science research, SPSS has predominantly been used to analyze the proximity and preference data due to its comprehensive data-analytic capabilities and user-friendly graphical interface. For ID MDS, ALSCAL (Takane et al., 1977) implemented in SPSS has widely been used. However, PROXSCAL (Busing et al., 1997), recently implemented in SPSS, is gradually taking over the role of ALSCAL with more flexible visualizing and data-analytic features such as graphing capabilities and multiple random starts. For unfolding analysis, PREFSCAL (Busing et al., 2005), also recently incorporated in SPSS, has primarily been used in empirical research owing to its remarkable feature to overcome the degeneracy problems (e.g., all object points collapsing into one) prevalent in unfolding analysis.

Bibliography

Borg, I., Groenen, P.J.F., 2005. Modern Multidimensional Scaling: Theory and Applications. Springer, New York.

Busing, F.M.T.A., Commandeur, J.J.F., Heiser, W.J., 1997. PROXSCAL: a multidimensional scaling program for individual differences scaling with constraints. In: Bandilla, W., Faulbaum, F. (Eds.), Softstat '97: Advances in Statistical Software. Lucius, Stuttgart, Germany, pp. 237–258.

Busing, F.M.T.A., Groenen, P.J.K., Heiser, W.J., 2005. Avoiding degeneracy in multidimensional unfolding by penalizing on the coeffcient of variation. Psychometrika 70 (1), 71–98.

Carroll, J.D., Arabie, P., 1998. Multidimensional scaling. In: Birnbaum, M.H. (Ed.), Handbook of Perception and Cognition, Measurement, Judgment and Decision Making, vol. 3. Academic Press, San Diego, CA, pp. 179–250.

Carroll, J.D., Chang, J.J., 1970. Individual differences and multidimensional scaling via an *N*-way generalization of Eckart–Young decomposition. Psychometrika 35 (3), 282–319.

Coombs, C.H., 1964. A Theory of Data. Wiley, New York.

Cox, T.F., Cox, M.A.A., 2001. Multidimensional Scaling. Chapman and Hall, Boca Raton, FL.

Everitt, B.S., Rabe-Hesketh, S., 1997. The Analysis of Proximity Data. Wiley, New York.

Kruskal, J.B., Wish, M., 1978. Multidimensional Scaling. SAGE, Beverly Hills, CA.

Kruskal, J.B., 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29 (1), 1–27.

Kruskal, J.B., 1964b. Nonmetric multidimensional scaling: a numerical method. Psychometrika 29 (2), 115–129.

Nosofsky, R.M., 1992. Similarity scaling and cognitive process models. Annual Review of Psychology 43, 25–53.

Ramsay, J.O., 1977. Maximum likelihood estimation in multidimensional scaling. Psychometrika 42 (2), 241–266.

Ramsay, J.O., 1982. Some statistical approaches to multidimensional scaling data. Journal of the Royal Statistical Society, Series A (General) 145 (3), 285–312.

Schiffman, S.S., Reynolds, M.L., Young, F.W., 1981. Introduction to Multidimensional Scaling. Academic Press, New York.

Shepard, R.N., 1962. Analysis of proximities: multidimensional scaling with an unknown distance function, I and II. Psychometrika 27 (2,3), 125–140, 219–246.

Takane, Y., Young, F.W., de Leeuw, J., 1977. Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. Psychometrika 42 (2), 7–67.

Takane, Y., Jung, S., Oshima-Takane, Y., 2009. Multidimensional scaling. In: Millsap, R.E., Maydeu-Olivares, A. (Eds.), Handbook of Quantitative Methods in Psychology. Sage Publications, London.

Takane, Y., 2007. Applications of multidimensional scaling in psychometrics. In: Rao, C.R., Sinharay, S. (Eds.), Handbook of Statistics, Psychometrics, vol. 26. Elsevier B.V., Amsterdam, pp. 359–400.

Young, F.W., Hamer, R.M., 1994. Theory and Applications of Multidimensional Scaling. Erlbaum, Hillsdale, NJ.

Multidimensional Scaling II

Patrick JF Groenen, Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands
Ingwer Borg, Hausen o.V., Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J. De Leeuw, volume 20, pp. 13512–13519, © 2001, Elsevier Ltd.

Abstract

Multidimensional scaling (MDS) is a method for the visualization of dissimilarities between pairs of objects. The representation of the objects is done in a low (usually two)-dimensional space by the distance between the pairs of points. In metric MDS the distances approximate the dissimilarities directly, and in nonmetric (ordinal) MDS the distances should match the rank order of the dissimilarities as well as possible. As distance measure, almost invariably the Euclidean distance is used. The fit is expressed by the Stress value. The interpretation of the MDS configuration is done in terms of the distances between the points.

Introduction

Multidimensional scaling (MDS) is a statistical method that optimally maps proximity data on pairs of objects (i.e., data expressing the similarity or the dissimilarity of pairs of objects) into distances between points in a multidimensional space (usually two or three dimensions). Sometimes, MDS is also used as a container term for all methods that provide a graphical representation of variables and/or objects in a scatter plot. This definition is very broad and contains a wide variety of methods such as correspondence analysis, principal component analysis, canonical correlation analysis, etc. We here limit ourselves to the narrow sense definition of mapping proximities into distances.

To introduce MDS in a research context, we use an example from value research where MDS has always played a cardinal role. Values are guiding principles in people's lives, basic goals a person strives for in many actions and situations. Schwartz (2012) and many others have found 10 universal basic values: power (PO), achievement, hedonism, stimulation, self-direction, universalism, benevolence, tradition (TR), conformity (CO), and security (SE). These values, moreover, always exhibit certain structural relations. For example, persons who attach high importance to achievement also tend to rate benevolence as not so important, and vice versa.

Consider a sample of 151 US adults (Borg and Bardi, in press) who completed the Portrait Values Questionnaire (PVQ, http://wiki.mgto.org/portrait_value_questionnaire_pvq), a questionnaire of 40 items assessing how persons rate the relative importance of the 10 universal values. Value researchers typically first intercorrelate the scores of these items and then subject the correlation matrix to a 'default' MDS. This MDS optimally represents the items as points in two-dimensional space such that the distances of the points are (inversely) ordered just as the correlations of the items. Hence, the closer two points in the MDS solution (Figure 1), the higher the correlation of the items that they represent. The 40 points are coded here in terms of the basic values that the respective items are measuring. For example, 'co1' is item #1 of the PVQ measuring conformity, and 'po2' is item #2 on power. To facilitate interpretation, additional lines were added that cut the space like a cake into wedgelike *regions*. They form a *circumplex*, a circle of regions emanating from a common origin. Each region contains only items of one particular type – except for a few small errors where points (e.g., co4 or tr1) fall into the respective neighboring region.

The MDS solution in Figure 1 is a rather precise representation of the data. MDS comes with a fit index called *Stress*. Simulations show that the Stress of our MDS solution is clearly better than the Stress of any MDS solution for 1000 random permutations of the data. Hence, the MDS solution in Figure 1 is *significant*. Moreover, the circumplex pattern that we see here has been *replicated* in hundreds of similar studies. So, this structure is *robust*.

Rather than scaling items, we may first aggregate the 40 items to yield importance scores for each of the 10 values, and then do an MDS on the 10×10 correlation matrix of these scores. Figure 2 shows that this results in a simple

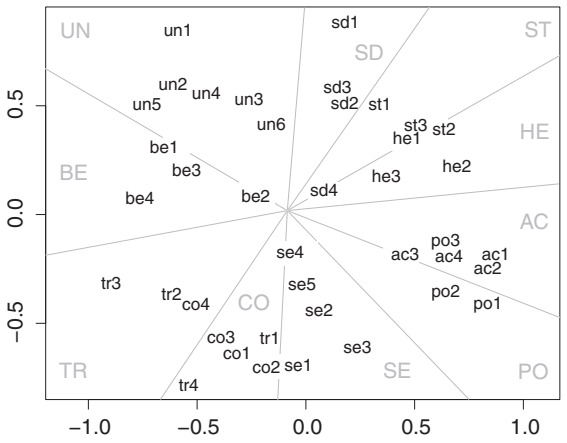

Figure 1 Ordinal multidimensional scaling solution for the intercorrelations of 40 Portrait Values Questionnaire items measuring 10 universal values. PO, power; AC, achievement; HE, hedonism; ST, stimulation; SD, self-direction; UN, universalism; BE, benevolence; TR, tradition; CO, conformity; SE, security.

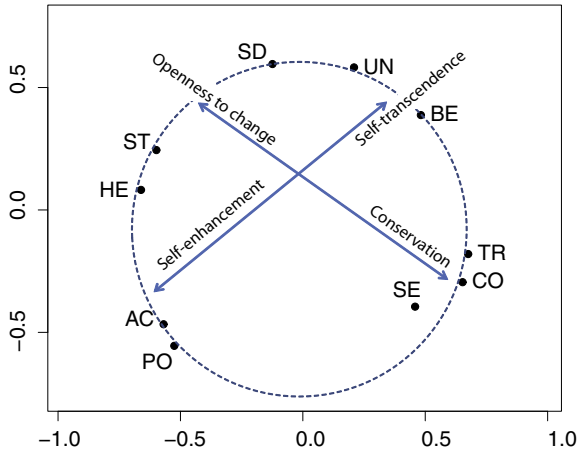

Figure 2 Ordinal multidimensional scaling solution for the intercorrelations of relative importance scores for 10 universal values. PO, power; AC, achievement; HE, hedonism; ST, stimulation; SD, self-direction; UN, universalism; BE, benevolence; TR, tradition; CO, conformity; SE, security.

pattern: The 10 values are all on an almost perfect circle, and their order replicates what has been found in many other studies. The MDS solution is also significant (based on permutation tests). Moreover, repeating the MDS with only 9 of the 10 values (*jackknifing*) leads to very similar configurations. Thus, this MDS solution is also *stable*.

We can go on and ask whether it is possible to force the 10 value points onto a *perfect* circle, so that the values are actually ordered on a *one-dimensional circular scale*. We can impose this theory-driven *external* constraint onto the MDS solution. This turns *exploratory* MDS into *confirmatory* MDS. In this case, it can be done easily using *spherical* MDS. We find that the perfect-circle constraint increases the Stress only insignificantly, and so we should – for reasons of simplicity and theoretical meaningfulness – accept this configuration as the 'final' MDS solution for these data.

Figure 2 also shows a more traditional interpretation in terms of two *dimensions*: self-enhancement versus self-transcendence, and openness to change versus conservation.

Formalizing MDS

In MDS, one deals with three mathematical objects: data that are (or have been converted into) dissimilarities δ_{ij} (for each object pair i and j), where larger values indicate larger dissimilarity; a configuration of n points in m-dimensional space, described by an $n \times m$ matrix $\mathbf{X}$ of coordinate values; and the distances computed on $\mathbf{X}$, written as $d_{ij}(\mathbf{X})$.

When doing MDS, the user has to make some choices. An important feature of MDS is that instead of fitting distances directly to the observed dissimilarities, often a transformation $\widehat{d}_{ij}$ of the observed dissimilarities δ_{ij} are fitted to the distances. The goal of MDS is then to find both $\widehat{d}_{ij}$'s and distances $d_{ij}(\mathbf{X})$'s such that they match as closely as possible. Therefore, one of the choices to be made is to assign a particular scale level to the observed dissimilarities δ_{ij} (ordinal, interval, or ratio). Picking the ratio level amounts to requesting that $\widehat{d}_{ij} = b \cdot \delta_{ij} \approx d_{ij}(\mathbf{X})$, where b is a substantively irrelevant overall positive scaling constant. In interval MDS, one wants $\widehat{d}_{ij} = a + b \cdot \delta_{ij} \approx d_{ij}(\mathbf{X})$, where a and b are chosen optimally as long as the resulting $\widehat{d}_{ij} \geq 0$. And in ordinal MDS, the $\widehat{d}_{ij}$ can be rescaled in *any* way as long as they have the same order as the dissimilarities δ_{ij} so that if $\delta_{ij} \leq \delta_{kl}$ then we must have $\widehat{d}_{ij} \leq \widehat{d}_{kl}$.

A second choice is picking a distance function between rows i and j of $\mathbf{X}$. This is almost always the Euclidean distance that is defined by

$$d_{ij}(\mathbf{X}) = \left(\sum_{s=1}^{m} (x_{is} - x_{js})^2 \right)^{1/2} \qquad [1]$$

and has the simple interpretation of the length of the straight line connecting points i and j.

The core of what MDS attempts to accomplish is expressed by the so-called *Raw Stress* loss function. Raw Stress is the sum of the squared errors of representing a dissimilarity of $\widehat{d}_{ij}$ by the corresponding distance $d_{ij}(\mathbf{X})$, that is,

$$\sigma^2_{\text{raw}}(\widehat{\mathbf{D}}, \mathbf{X}) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} w_{ij} \left(\widehat{d}_{ij} - d_{ij}(\mathbf{X}) \right)^2 \qquad [2]$$

where w_{ij} are nonnegative fixed weights set by the user to weight the importance of error ij (e.g., to handle missing data by setting $w_{ij} = 0$ if δ_{ij} is missing and $w_{ij} = 1$ otherwise) and $\widehat{\mathbf{D}}$ is the matrix with elements $\widehat{d}_{ij}$ that are related to the dissimilarities by one of the specified transformations (e.g., ordinal, interval, or ratio). The loss function [2] formalizes the objective of MDS in terms of a mathematical optimization problem. MDS seeks to minimize Raw Stress by pursuing two objectives: (1) finding an optimal matrix of coordinates $\mathbf{X}$ for given $\widehat{d}_{ij}$'s and (2) finding $\widehat{d}_{ij}$ values within the specified transformation for given distances $d_{ij}(\mathbf{X})$. However, minimization of $\sigma^2_{\text{raw}}(\widehat{\mathbf{D}}, \mathbf{X})$ over $\widehat{\mathbf{D}}$ and $\mathbf{X}$ would yield the trivial solution $\widehat{\mathbf{D}} = 0$ and $\mathbf{X} = 0$ (except when $\widehat{d}_{ij} = \delta_{ij}$). One way to steer away from this trivial solution is to replace $\sigma^2_{\text{raw}}(\widehat{\mathbf{D}}, \mathbf{X})$ by $\sigma^2_{\text{norm}}(\widehat{\mathbf{D}}, \mathbf{X})$ that imposes the explicit normalization restriction $\sum_{i>j} w_{ij} \widehat{d}^2_{ij} = 1$. Alternatively, implicit normalization could be used as in Stress-1 defined by

$$\sigma_1(\widehat{\mathbf{D}}, \mathbf{X}) = \left(\frac{\sum_{i=2}^{n} \sum_{j=1}^{i-1} w_{ij} \left(\widehat{d}_{ij} - d_{ij}(\mathbf{X}) \right)^2}{\sum_{i=2}^{n} \sum_{j=1}^{i-1} w_{ij} d^2_{ij}(\mathbf{X})} \right)^{1/2} \qquad [3]$$

The configurations $\mathbf{X}$ and the $\widehat{d}_{ij}$'s obtained by minimizing normalized Stress and Stress-1 only differ by unimportant overall scaling constants (for proofs, see, e.g., Section 11.1 of Borg and Groenen, 2005). When using MDS, Stress-1 is almost always reported as a fit index, because it is comparable over different solutions due to its normalizing denominator. Hence, when reading 'Stress' in MDS publications, it can be assumed to mean 'Stress-1.'

Similarities, Dissimilarities, and Optimal Transformations

If the proximity data are similarities s_{ij}, they first need to be transformed into dissimilarities (either 'by hand' or within the MDS program). A simple scale reversal is accomplished by

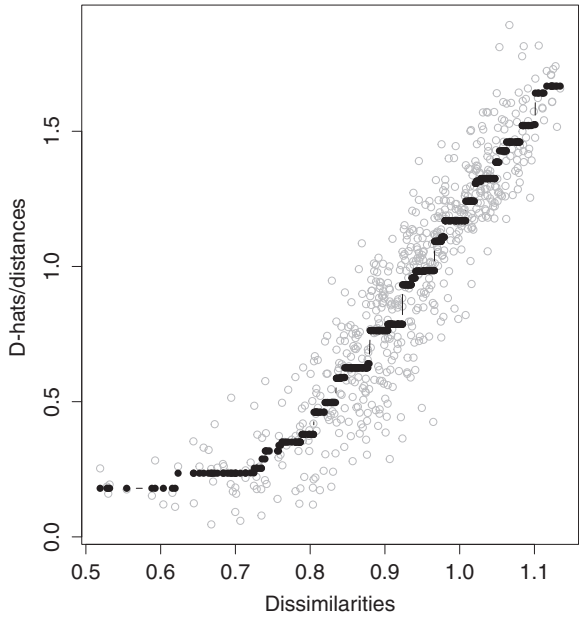

Figure 3 Shepard diagram for **Figure 1**; black points show $(\delta_{ij}, \widehat{d}_{ij})$ pairs; gray open circles show $(\delta_{ij}, d_{ij}(\mathbf{X}))$ pairs; vertical distances of black and gray points are errors of representation.

$\delta_{ij} = c - s_{ij}$, where $c = \max_{ij}(s_{ij})$. A second way of transforming s_{ij} into δ_{ij} exists if all $s_{ij} > 0$. Then choosing $\delta_{ij} = 1/s_{ij}$ is often a good choice. In the value study described above, the correlations r_{ij} were transformed into dissimilarities by $\delta_{ij} = (1 - r_{ij})^{1/2}$, a conversion turning correlations into Euclidean distances.

MDS converts these dissimilarities further into $\widehat{d}_{ij}$'s (d-hats), that is, admissibly rescales them on their scale level so that they differ as little as possible from the distances of the MDS configuration $\mathbf{X}$. The d-hats for the final MDS solution in **Figure 1** are visualized in the *Shepard diagram* in **Figure 3** by the black points. The vertical distance between a gray point (representing the distance $d_{ij}(\mathbf{X})$ in this plot) and the corresponding black point (representing the $\widehat{d}_{ij}$) shows the error of representation of the MDS solution for the object pair ij. If both points (black and gray) are close, then there is little error, and if they are far from each other, the error is large. The Shepard diagram in **Figure 3** shows that there is little scatter of the gray points about the black-point trend. Hence, the MDS solution represents the structural information contained in the proximity data quite well.

A Brief History of MDS

MDS was not invented by statisticians. The main reason for its existence is that psychologists used the distance formula as a model of perception and judgment. They argued that if a judgment on the similarity of two objects is needed, the person first forms a mental representation of the objects in 'psychological space,' a space spanned by the objects' subjective attributes. Dissimilarity judgments are then generated by first assessing the differences of the objects on each attribute, and then summing these intradimensional differences over the dimensions. To the outside observer – and possibly also to the person himself or herself – the psychological space itself may be unknown, but MDS promised to 'uncover' it from the individual's overall dissimilarity judgments. When it became technically feasible to run MDS with ordinal similarity data in the 1960s, this model led to much research.

Ordinal MDS with its flexibility to analyze all kinds of proximity data became popular because computer programs became widely available that solved the complex mathematics of this scaling model. Ordinal MDS dominated the field for many years, but interval and even ratio MDS later recovered considerably in terms of utilization. This had statistical reasons on the one hand (ratio MDS solutions, for example, are often less cluttered), and theoretical reasons on the other hand (the emphasis of measurement foundations had shifted toward cumulative theory construction over replications, away from an almost endless testing of single data sets), and metric MDS solutions often allow for simpler and more robust interpretations than possibly overfitted ordinal solutions. Today, ordinal MDS is but one of several MDS models. Advanced computer programs generate solutions for each of them almost instantaneously, and so they can be tested easily against each other at virtually no costs.

Many of the original motives that led to the development of MDS have become unimportant. What has survived, in particular, is using MDS in attitude and value research, for example, where intercorrelations of survey items are studied for correspondences of the conceptual facets of the items to regions of their spatial representation (as in **Figure 1**). However, most recent applications of MDS serve a much wider purpose, that is, they are done to visualize tables of indices that can be interpreted as (dis)similarity data. For that purpose, MDS is highly useful as it can handle a vast variety of data as long as they are (dis)similarities (e.g., correlations, covariances, co-occurrence data, profile distances); it does not require interval-scaled data but also handles ordinal (and even nominal) data; it is robust against missing data and coarse data; it often serves as a data smoother, showing a structure that is replicable even under conditions of high error; it is easily explained to nonexperts and allows them to explore the solutions without much risk; it is easy to run for nonexperts even though its solution algorithms are rather difficult; and it does not impose a particular interpretation of its solution onto the user but allows the data to speak for themselves.

Data Used in MDS

MDS builds on *proximity* data that express the similarity or the dissimilarity among objects of interest. In psychological research, proximities are often collected directly by asking persons to rate the perceived similarity of objects of interest on a numerical rating scale. A popular alternative is sorting a stack of cards. Each card exhibits exactly one pair of objects. The testees are asked to sort these cards such that the card with the most similar pair is on top of the stack, and the card showing the most dissimilar pair at the bottom. This sorting can be further simplified: The testee is asked to begin by sorting the cards into only two stacks, one for the 'similar' pairs of objects, and one for the 'dissimilar' pairs. For each stack, this two-stack

sorting is repeated several times until the testee feels that he/she cannot reliably split the remaining stacks any further. One then numbers the various stacks from most similar to least similar. Pairs in the same stack receive this stack's number as a similarity score.

Pairwise procedures can lead to a huge number of pairs that must be judged by the subjects. Fortunately, not all these pairs are needed, because MDS also works with missing data. Spence and Domoney (1974) showed that the proportion of missing data can be as high as 80% and (ordinal) MDS is still able to recover an underlying MDS configuration quite precisely, given certain side constraints such as randomly distributed missing values in the data matrix and 'many' points. One can further improve the robustness of MDS by paying particular attention to collecting proximities for pairs of objects that seem very dissimilar rather than similar, because data that correspond to large distances are particularly important for recovering an underlying true MDS configuration (Spence and Graef, 1974).

The labor involved in data collection can also be reduced by simplifying the similarity judgments. Rather than asking for graded ratings on, say, a 10-point scale, one may offer only two response categories, 'similar' and 'dissimilar' for each pair of objects. Summing such data over replications or over respondents leads to *confusion frequencies* or, after dividing by the number of cases, to *confusion probabilities*. Yet, such aggregations are not necessarily required. Green and Wind (1973) showed that robust MDS works well even with using coarse data such as 1-0 data or ratings with few categories (given advantageous side constraints such as scaling in the true dimensionality and having many points).

Today, most applications of MDS are based on proximity indexes derived from pairs of data vectors. The typical case is correlating the objects of interest over their ratings on certain attributes. So, for example, rather than asking a person to rate the global similarity of two countries such as 'Japan and China,' we could ask them to first rate each country on attributes such as size, population, Gross National Product (GNP), etc., and then compute the correlation of these countries over these attributes.

Rather than using correlations, one can also measure their overall dissimilarity by a distance function such as the Euclidean distance. If ordinal MDS is employed, it becomes irrelevant whether product–moment correlations or Euclidean distances are used, because they are monotonically related and the rank order of the data is all that matters in ordinal MDS. If interval or ratio MDS is used, however, or if other distance functions are used, then one has to consider carefully which coefficient is best to express the similarity of profiles in the desired sense.

An interesting special case of proximities are *co-occurrence data*. Here is one example. Coxon and Jones (1978) studied the categories that people use to classify occupations. Their subjects were asked to sort a set of 32 occupational titles (such as barman, statistician, and actor) on the basis of their overall similarity into as many or as few groups as they wished. The result of this sorting can be expressed, for each subject, by a 32×32 *incidence matrix*, with an entry of 1 wherever its row and column entries were sorted into the same group, and 0 elsewhere. The elements of this incidence matrix, thus, are based on directly collected same-different proximities.

Co-occurrence data are typically aggregated over persons. This is done by first adding the various combinations of occurrence, nonoccurrence, and co-occurrence of any two events, X and Y, of the set of events of interest. Each event either occurs ($= 1$), or it does not occur ($= 0$). We denote the frequencies of the various combinations by 11, 10, 01, and 00, respectively. This leads to an amazing number of different similarity measures (Gower and Legendre, 1986). A prominent example is $S_2 = 11/(11 + 10 + 01 + 00)$. Another case is $S_3 = 11/(11 + 10 + 01)$ (Jaccard similarity index). A third example is $S_4 = (11 + 00)/(11 + 10 + 01 + 00)$. Choosing a particular S-index over another such index can have dramatic consequences. Rare events, for example, boost 00 so that S_4 makes two events similar just because they are rare.

Basic Formal Aspects of MDS that the User Should Know

Stress Values

The fit of an MDS solution can be evaluated by considering the Stress value. The dominant Stress measure in the literature is Stress-1 which we report here. The solution in **Figure 1**, for example, has a Stress value of 0.15. (Stress = 0.00 means perfect fit; the worst possible Stress value is 1.00.) This value is often compared with the Stress expected for purely random data. Simulation studies by Spence and Ogilvie (1973) show that for an ordinal MDS configuration of 40 points in two dimensions, the expected Stress is about 0.35, with a standard deviation of 0.003. Hence, our solution in **Figure 1** is highly significant, at least when tested against this criterion, that is, against the "nullest of all null hypotheses" (Cliff, 1973).

In general, when evaluating the Stress of an MDS solution, many factors have to be considered, namely

- The greater the number of points n, the larger the expected Stress.
- The greater the dimensionality of the MDS solution, the smaller the expected Stress.
- The greater the noise in the data, the larger the expected Stress.
- The stricter the transformation, the higher the expected Stress value: ratio > interval > ordinal.
- For ordinal transformations and the primary approach to ties, the more ties (= equal values) in the proximities, the smaller the expected Stress.
- The more proximities are missing, the easier it is to find an MDS solution with small Stress.
- Eliminating particular points that contribute highly to Stress or setting certain proximities as missing (e.g., because they are errors) can reduce the total Stress considerably.

Dimensionality of the MDS Solution

In terms of dimensionality, the vast majority of MDS solutions in the literature are two-dimensional. When exploring a set of data, scaling in two dimensions is always the first option, because such a solution is easily accessible to the researcher's eye and often sufficiently detailed for analyzing the essential (and robust) structure of the proximities. Points that do not fit

at all into this plane ('outliers') could be eliminated at first, and later studied in higher-dimensional MDS spaces or by other statistical methods in terms of how they relate to all the variables that do fit.

Interpreting MDS Solutions

Interpreting an MDS solution can be understood as projecting given or assumed content knowledge onto the MDS configuration. Any feature of the MDS geometry can be interpreted. Three methods can be helpful: directions, regions, and clustering of points. Interpreting directions has been most popular, particularly directions that coincide with dimensions. Often, it is more useful to interpret a direction that is not the same as a dimension. Figure 2 gives an example. The two directions (e.g., self-enhancement vs self-transcendence) are interpreted by studying the projections of the points on these lines from a content perspective.

Even more general is interpreting *regions* as in Figure 1. Regions are subsets of points of an MDS space that are connected (i.e., each pair of points in a region can be joined by a curve whose points lie completely within this region), nonoverlapping, and exhaustive (i.e., each point lies in exactly one region). Such regions may arise when first projecting some codes onto the points, and then *partitioning* the space so that regions emerge where all points have the same code. In Figure 1, all points were first coded by the value that it is related to (power, security, etc.). Then, a set of lines were drawn into this plot so that 10 different regions emerged. Rather than choosing curved lines, all lines are straight here, and all lines emerge from one common origin. This circumplex structure was chosen because it replicates what was found in many previous studies on value research. There are a few points in a 'wrong' (but neighboring) region, but this can be interpreted as due to noise in the data. Three prototypical regionalities that often arise in practice are shown in Figure 4: *axial*, *modular*, and *polar* partitions. Different partitionings can be superimposed to yield two- or multiway partitionings. A prominent example is the *radex*, a combination of a circumplex and a set of concentric regions, all emanating from a common origin. The radex, therefore, is a pattern of regions that looks like a dartboard. It adds to a circumplex a polar facet that explains the points' distance from the origin.

Regions are sometimes confused with *clusters*. Clusters, however, are but special cases of regions. They are often defined as lumps (or chains) of points surrounded by empty space so that each point in a cluster is always closer to at least one point within the cluster than to any point not in the cluster.

Doing It Right

As with other techniques, MDS users need to pay special attention to some practical details.

If an MDS program assumes the wrong polarity for the data, it almost always generates uninterpretable solutions with very high Stress. No computer program can find out by itself whether the proximity data are similarities or dissimilarities. Thus, it is crucial to read in data whose meaning corresponds to the program's default assumption, or click the correct 'similarity' or 'dissimilarity' button when running the program.

As the distances are the most important output of MDS, they should not get distorted in the configuration plots. The MDS user should therefore check that the plot has an aspect ratio of 1, that is, one horizontal unit should be equal to one vertical unit. Many plotting programs try to maximize the spread of the points along the horizontal and vertical axes, thereby destroying the aspect ratio of 1. This even happens in the plots provided by some MDS programs.

As MDS is an iterative method, it can be stopped too early causing the solution to be suboptimal. Therefore, the maximum number of iterations may need to be increased and subsequent Stress-1 values should be smaller than 0.0 001 or subsequent normalized Stress smaller than 10^{-8}. In addition, it is helpful to start the iterations with a good initial configuration. A particularly good one invariably available in MDS programs is the classical MDS start also referred to as 'Torgerson' scaling.

The minimization algorithms can get stuck in a local minimum which is not the global minimum. To guard against such local minima, some programs offer multiple random start configurations. Often, a good initial start also leads to good local minima.

A particularly bad situation of local minima occurs in the case of *unidimensional* scaling ($m = 1$). It can be proven that this is a notoriously hard combinatorial problem. Unless severe precautions are taken (such as a massive number of random starts) the user is advised not to use standard MDS software as it is very likely that it will end in a poor-quality local minimum.

The most uninformative data for MDS consist of all dissimilarities (or d-hats) being the same. It turns out that in this situation, MDS finds special solutions. For $m = 1$, the solution consists of all points equally spaced on a line; for $m = 2$, the points will be distributed in concentric circles; for

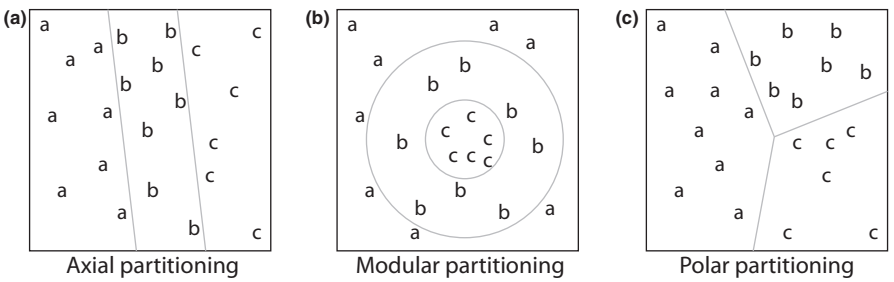

Figure 4 Prototypical partitioning of multidimensional scaling configurations by three facets, each one with three categories (a, b, c).

$m > 2$, the points will be on the surface of a hypersphere. Although constant dissimilarities do not occur without the user knowing this, almost constant d-hats can occur during the optimization process. Therefore, the user should always check the Shepard diagram and consider whether the d-hats found have sufficient spread around their mean. With little spread, the solution in two dimensions will be the uninformative circle of almost randomly spread points. Also, if n is large and the distribution of dissimilarities is more skewed then, too, the MDS solution tends to show the uninformative constant dissimilarity solution.

Weights w_{ij} cannot only be used for handling missing values. Another use is to selectively emphasize or de-emphasize the fit of certain dissimilarities. A useful choice is setting $w_{ij} = \delta_{ij}^{q}$. For large q (e.g., $q = 5$), only the large dissimilarities will determine the shape of the MDS solution. In contrast, choosing q small (e.g., $q = -5$) emphasizes the fit of the smaller dissimilarities so that the distances representing the larger ones may well be poorly represented.

Special weighting patterns exist such that two or more subsets of objects can be formed that are not connected by a single dissimilarity. This situation is called *reducible* because the MDS problem could be split up into as many separate MDS solutions as there are subsets. Such a reducible situation can occur if there are many missing values. As there are no connections between the subsets, there is no use in presenting the objects of different subsets in a single plot and it is better to represent each subset in a separate plot.

Special Cases of MDS

Confirmatory MDS

Standard MDS does not impose restrictions to the configuration **X**. However, if additional information is available or can be theorized, it may be useful to impose constraints on the configuration **X**. Here is a classical example, the MDS of Morse signals. Content information is available on the Morse signals: Each signal has a temporal length (from 0.05 to 0.95 s) and a certain composition of long and short beeps. We code the latter as 'only short beeps' (1), 'more short beeps than long beeps' ($1 > 2$), 'an equal number of short and long beeps' ($1 = 2$), 'more long than short beeps' ($2 > 1$), and 'only long beeps' (2). These two external variables are facets of the signals. An approximate axial partitioning of the unconstrained MDS representation of confusion rates of Morse signals is shown in Figure 5(a). The 'curvy' partitioning lines in this plot were drawn by hand. Figure 5(b) shows the confirmatory solution. Here, the partitioning lines are all straight lines, a property enforced onto the MDS configuration. The axially constrained ('confirmatory') solution has a slightly higher Stress (0.21) compared to the unconstrained solution (0.18), yet it yields a plot that is much easier to interpret in psychophysical terms. Moreover, the linearized structure expresses two simple laws of formation and, therefore, it can be expected to be more robust over replications than the possibly overfitted exploratory MDS pattern with its curvy partitioning lines that were inserted only afterward.

An alternative form of confirmatory MDS is the spherical (in 3D) or the circular constraint (in 2D). The points in Figure 2 are almost on a perfect circle. Therefore, imposing the circular restriction will hardly increase the Stress while making the MDS model more parsimonious.

At this time, the possibilities to impose certain constraints onto **X** are rather limited. For example, it would be desirable to request that the pattern in Figure 1 emerges as a perfect circumplex with no misplaced points, possibly also with a particular order of the various value regions. Yet, solutions or programs for such constraints have not been worked out yet.

In general, the stronger the constraints, the higher the increase in Stress compared to the unconstrained solution. However, if the increase is large then one should doubt

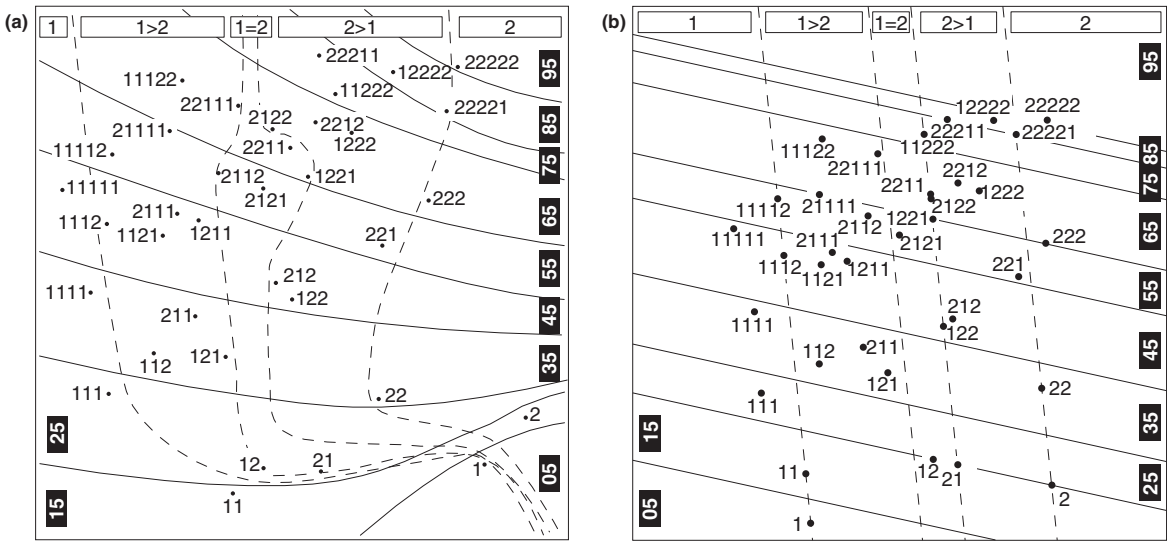

Figure 5 Multidimensional scaling (MDS) solution of ordinal MDS on the Rothkopf confusion data with regions drawn by hand (panel a) and regionally constrained MDS (panel b). The panels in this figure are reproduced from Borg, I., Groenen, P.J.F., 2005. Modern Multidimensional Scaling, second ed. Springer, New York, with kind permission from Springer Science and Business Media B.V.

whether the dissimilarities can be properly modeled by an MDS with the chosen constraints.

Three-Way MDS

Several models have been introduced for dealing with replications of the dissimilarity matrix. One particularly well-known example is the *dimensional weighting model*, often called the INDSCAL model ('individual differences scaling model') of Carroll and Chang (1970).

$$d_{ijk}(\mathbf{X}, \mathbf{w}_k) = \left(\sum_{s=1}^{m} w_{sk} (x_{is} - x_{js})^2 \right)^{1/2}, \quad w_{sk} > 0 \qquad [4]$$

where the parameter $k = 1, \ldots, K$ stands for different individuals or cases. The weight w_{sk} can be interpreted as the *salience* of dimension s for person k. Expressed geometrically, different individuals share a common space $\mathbf{X}$ but they may differ in how they weight the dimensions of this space. Applying the weight w_{sk} in [4] amounts to stretching or compressing $\mathbf{X}$ along dimension s for individual k. Although this is an interesting model, it tends to promise more than it holds. The problem is that one cannot infer from $2 \cdot w_{1k} = w_{2k}$, for example, that individual k attributes twice as much weight to dimension 2 than to 1, because these weights depend on the norming of $\mathbf{X}$: If we stretch or compress $\mathbf{X}$ along its dimensions, the weights change inversely. What can be compared, though, are the weights of different individuals *within* each dimension.

Unfolding

None of the 151 individuals of the introductory example appeared in the MDS solutions in Figures 1 and 2. Rather, all individuals got 'lost' in the correlations that were used for MDS. So, one can ask if the circle of values also exists within individuals. To answer this question, we use *unfolding*, a special case of MDS. In *external unfolding*, we would attempt to represent each individual by adding one particular point to the configuration in Figure 2. This *person point* should be located so that the distances from this point to the *value points* correspond to this person's importance scores. For example, if a person assigns a high importance value to power, his or her person point should be close to the point representing power in Figure 2 (PO). If security is rated as not so important, this person point should be far from the point representing security in Figure 2 (SE).

Unfolding can also be computed directly, without using any previously generated configurations. To do that, we first reverse the importance scores by turning them into dissimilarities by subtracting each importance rating from the largest observed importance score. In *metric unfolding*, we then search for a configuration with 10 points for the 10 values and 151 points for the 151 individuals such that the distances between the person points and the value points directly match the dissimilarities (except for an overall scaling factor). Figure 6 shows the result. The 10 value points are represented by their acronyms, and the 151 person points by small circles. The value points, again, show an approximately circular structure, with values ordered as predicted by the Schwartz value theory.

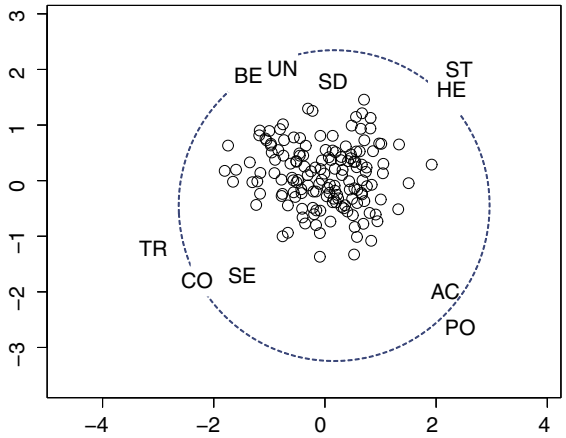

Figure 6 Metric unfolding solution for the relative importance scores given by 151 persons to 10 universal values. PO, power; AC, achievement; HE, hedonism; ST, stimulation; SD, self-direction; UN, universalism; BE, benevolence; TR, tradition; CO, conformity; SE, security.

The unfolding solution has a good fit to the data. Almost all persons are well represented. An exception is a person who rates achievement as not so important and power as very important, for example. This person's point should therefore be far from AC and close to PO. However, in Figure 6, AC and PO are close together, and so this person's ratings cannot be perfectly modeled within the value configuration shown in Figure 6.

Software

MDS can be run in almost all statistical software packages. The two most dominant ones are the smacof package in R and Proxscal in SPSS. For an overview of available MDS software with syntax, see Borg and Groenen (2005) and Borg et al. (2013). The GGvis software within GGobi (either stand alone or callable as an R package) of Buja and Swayne (2002) is one of the very few MDS software packages that is truly interactive. Vosviewer is MDS software (and more) with very good graphics for the analysis of bibliometric data (Van Eck and Waltman, 2010).

Metric unfolding can be done by any MDS program that allows for missing values. The reason is that unfolding can be seen as a special case of MDS where the objects can be split into two sets and the within-set proximities are missing. However, nonmetric unfolding requires special software to avoid degenerate solutions where one of the sets collapses into a single point. At the time of writing, the only software that can handle nonmetric unfolding is Prefscal (in SPSS).

Further Reading

Quite comprehensive treatments of MDS and related techniques can be found in Cox and Cox (2000) and Borg and Groenen (2005). An introductory text on MDS is Borg et al. (2013). A recent paper on MDS for statisticians is Buja et al.

(2008). For the majorization algorithm for MDS (smacof and Proxscal) we refer to De Leeuw and Heiser (1980) and De Leeuw (1988) or Borg and Groenen (2005).

See also: Correspondence Analysis; Multidimensional Scaling I.

Bibliography

Borg, I., Bardi, A. The Value Circle within Persons, in press.

Borg, I., Groenen, P.J.F., 2005. Modern Multidimensional Scaling, second ed. Springer, New York.

Borg, I., Groenen, P.J.F., Mair, P., 2013. Applied Multidimensional Scaling. Springer Briefs in Statistics. Springer, Heidelberg.

Buja, A., Swayne, D.F., 2002. Visualization methodology for multidimensional scaling. Journal of Classification 19, 7–44.

Buja, A., Swayne, D.F., Littman, M.L., Dean, N., Hofmann, H., Chen, L., 2008. Data visualization with multidimensional scaling. Journal of Computational and Graphical Statistics 17, 444–472. http://dx.doi.org/10.1198/106186008X318440.

Carroll, J.D., Chang, J.J., 1970. Analysis of individual differences in multidimensional scaling via an *N*-way generalization of 'Eckart-Young' decomposition. Psychometrika 35, 283–320.

Cliff, N., 1973. Scaling. Annual Review of Psychology 24, 473–506.

Cox, T.F., Cox, M.A.A., 2000. Multidimensional Scaling, second ed. Chapman & Hall, London.

Coxon, A.P.M., Jones, C.L., 1978. The Images of Occupational Prestige: A Study in Social Cognition. Macmillan, London.

De Leeuw, J., 1988. Convergence of the majorization method for multidimensional scaling. Journal of Classification 5, 163–180.

De Leeuw, J., Heiser, W.J., 1980. Multidimensional scaling with restrictions on the configuration. In: Krishnaiah, P.R. (Ed.), Multivariate Analysis, vol. V. North-Holland, Amsterdam, The Netherlands, pp. 501–522.

Gower, J.C., Legendre, P., 1986. Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification 3, 5–48.

Green, P.E., Wind, Y., 1973. Multivariate Decisions in Marketing: A Measurement Approach. Dryden, Hinsdale, IL.

Schwartz, S.H., 2012. An overview of the Schwartz theory of basic values. Online Readings in Psychology and Culture 2. http://dx.doi.org/10.9707/2307-0919.1116.

Spence, I., Domoney, D.W., 1974. Single subject incomplete designs for nonmetric multidimensional scaling. Psychometrika 39, 469–490.

Spence, I., Graef, J., 1974. The determination of the underlying dimensionality of an empirically obtained matrix of proximities. Multivariate Behavioral Research 9, 331–341.

Spence, I., Ogilvie, J.C., 1973. A table of expected stress values for random rankings in nonmetric multidimensional scaling. Multivariate Behavioral Research 8, 511–517.

Van Eck, N.J., Waltman, L., 2010. Software survey: Vosviewer, a computer program for bibliometric mapping. Scientometrics 84, 523–538.

Relevant Websites

http://www.ggobi.org/ – Interactive visualization program that has the GGvis package for dynamic MDS.

http://people.few.eur.nl/groenen/mmds/ – Is the website that goes with the Borg and Groenen (2005) book. It has many MDS data sets and links to MDS software.

http://www.vosviewer.com/ – Is the home page for the VOSviewer software. It has excellent graphics and is particularly suited for the visualizing bibliometric data through MDS.

cran.r-project.org/package=smacof – Refers to the smacof package for MDS in R. It has provisions several MDS models including unfolding with and without restrictions.

Multilevel Models in Demography

Germán Rodríguez, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

We introduce multilevel models in demography, reviewing random intercept and random slope models, cross-level interactions, estimation of the fixed and random parameters, and calculation of subject-specific and population-average predictions. All ideas are discussed in the context of specific applications, a two-level analysis of contraceptive use in 15 countries and a three-level hazard model of infant and child survival in Kenya, with emphasis on the interpretation of results. Throughout the article we include references to other demographic applications of interest.

Introduction

Multilevel models deal with the analysis of data collected at multiple levels of aggregation, such as surveys that collect data on children, their mothers, and the communities where they live. Accounting for the correlation structure of the data is important for correct inference, and multilevel models achieve this aim by introducing stochastic residuals at various levels. The models are also known as *hierarchical*, because the levels are usually nested, and as *mixed*, because they include fixed and random parameters. They also encompass as special cases many models for longitudinal and panel data.

Although there were precursors going back to the 1950s, and significant statistical work in the 1970s, the field of multilevel analysis really took off in the 1980s with the publication of a number of research papers, many motivated by school effectiveness studies, see de Leeuw and Meier (2008). The field of demography was present from the outset, with Wong and Mason (1985) developing a hierarchical logistic regression model to study contraceptive use in 15 countries using data from the World Fertility Survey (WFS), see also Entwisle et al. (1986).

Multilevel analysis came of age in the 1990s, with the development of software packages and book-length treatments by Goldstein (1995) and collaborators, who developed MLn/MLwiN; Bryk and Raudenbush (1992), who developed HLM; and Longford (1993), who wrote VARCL. In the next 20 years, the field continued to advance, and multilevel procedures were incorporated in standard statistical packages, most notably Stata, which includes in version 13 a newly expanded family of multilevel procedures documented in a separate manual (StataCorp, 2013), and R, with functions contributed by Pinheiro and Bates (2000) and Bates (2010).

Estimation in linear mixed models is fairly straightforward, as these are essentially linear models with a patterned variance–covariance matrix, often block diagonal reflecting the hierarchical structure of the data. A common algorithm alternates between estimation of the linear parameters given the variance matrix, and estimation of the parameters in the variance matrix given the linear predictor, using generalized least squares in each step. There are two approaches in current use, maximum likelihood (ML) and restricted maximum likelihood (REML), which differ in the choice of contrasts used to estimate the variance parameters. REML has better small sample properties, but cannot be used to compare nested models that differ in the linear predictor, so ML has wider applicability.

Estimation in generalized linear mixed models, such as multilevel logit or hazard models, is not so straightforward because the resulting likelihoods are intractable. This has lead authors to propose approximate estimation procedures, most notably maximum quasi-likelihood and penalized quasi-likelihood, with linear and quadratic variants. Unfortunately these approximations are not reliable, as shown by Rodríguez and Goldman (1995), see also Lin and Breslow (1996). For models involving a moderate number of random effects the best approach is to use Gaussian quadrature, preferably one of the adaptive variants proposed by Liu and Pierce (1994) and Rabe-Hesketh et al. (2002). For models involving a large number of random effects one must resort to Markov chain Monte Carlo (MCMC) procedures in a Bayesian framework, see Clayton (1996). For a comparison of likelihood and Bayesian procedures, see Browne and Draper (2006).

Most demographic applications of multilevel models involve binary, categorical, count, or survival data, rather than continuous responses. We will, therefore, skip the linear case and focus directly on multilevel generalized linear models. Instead of discussing these models in purely abstract terms, however, we will discuss them in terms of two demographic applications to help ground the ideas: a classic analysis of ever use of contraception in 15 countries using random-coefficient logit models, and an analysis of infant and child mortality using survival models with shared frailty at the family and community levels. Along the way we include pointers to the literature and to other demographic applications. A useful general reference on multilevel modeling is Rabe-Hesketh and Skrondal (2012).

Multilevel Models of Contraceptive Use

We will introduce multilevel logit models in terms of the classic model of Wong and Mason (1985), who studied ever-use of contraception among women aged 40–44 years in 15 WFS countries. We build the model step-by-step, introducing random intercepts, random slopes, and cross-level interactions, before turning to interpretation of the estimates.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31024-8

Let π_{ij} denote the probability of having ever used contraception for the i-th woman aged 40–44 years in the j-th country. We start from a standard logit model

$$\text{logit}(\pi_{ij}) = \alpha + \beta x_{ij}$$

where x_{ij} is an individual-level predictor, such as education or type of place of residence. One problem with this model is that the observations will not be independent if women in each country share unobserved characteristics.

Random Intercepts

We address this problem by introducing a country-level residual and by assuming that conditional on that residual the observations are independent and follow the logit model

$$\text{logit}(\pi_{ij}) = (\alpha + a_j) + \beta x_{ij} \qquad [1]$$

where the random effect $a_j \sim N(0, \sigma_a^2)$ captures variation across countries in levels of contraceptive use, net of the observed covariates.

In this model the observations are still independent across countries but are correlated within each country. The intraclass correlation coefficient can be extended to logit models in terms of a latent variable representing propensity to have used contraception, and is defined as

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \pi^2/3}$$

where $\pi^2/3$ is the variance of the standard logistic distribution. Rodríguez and Elo (2003) proposed measures of intraclass correlation based on observable outcomes or manifest variables.

You may wonder why not introduce a separate dummy variable for each country. This is probably alright if you have just a few countries with large samples on each, but not if you have large numbers of groups with few observations per group, as is often the case with panel data. Consistent estimation of logit models then requires conditioning on a minimal sufficient statistic for the group totals, and leads to a so-called fixed-effects estimator, see Allison (2009).

There is an extensive literature, and more than a fair share of confusion, concerning the advantages and disadvantages of fixed- and random-effect estimators, and a thorough discussion would vastly exceed the scope of this article. A clear advantage of fixed-effect estimators, however, is that they focus on within-group variation and therefore adjust implicitly for any omitted group-level confounders. In contrast, random effects, like all statistical residuals, reflect only variation orthogonal to the observed covariates. As we will see below, however, multilevel models can be used to capture both within- and between-group variation and thus provide a richer modeling framework.

Random Slopes

The next model we will consider allows the effect of an individual-level covariate to vary across contexts, so the slope becomes random as well,

$$\text{logit}(\pi_{ij}) = (\alpha + a_j) + (\beta + b_j)x_{ij} \qquad [2]$$

The level-2 residuals are assumed to have a bivariate normal distribution with unstructured variance–covariance matrix, so the intercept and slope can be correlated

$$\begin{pmatrix} a_j \\ b_l \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}\right)$$

In their analysis of contraceptive use, Wong and Mason (1985) allowed the coefficients of years of education and childhood residence to vary across countries.

You may wonder why not run separate regressions for each country in the survey. Again, this is feasible if you have just a few large groups, but not if you have many small groups, where it is important to improve estimation for each group by borrowing strength from the others. As we will see, even with 15 countries there are substantial gains from a multilevel approach.

Level-Specific Models

An alternative way to write these models uses separate equations for each level, a practice that goes back to Burstein's (1980) slopes-as-outcomes ideas, and popularized by the specialized packages HLM and MLwiN. We specify the level-1 model as

$$\text{logit}(\pi_{ij}) = A_j + B_j x_{ij} \qquad [3]$$

where we use roman letters to indicate that the parameters are random and follow their own models.

In the random-intercept model, the level-2 equation is

$$\begin{aligned} A_j &= \alpha + a_j \\ B_j &= \beta \end{aligned}$$

In the random-slopes model, the level-2 equation is

$$\begin{aligned} A_j &= \alpha + a_j \\ B_j &= \beta + b_j \end{aligned}$$

You should verify that substituting these equations on [3] leads to [1] and [2]. Learning to do this translation is important because statistical packages differ in how models are specified.

Contextual Variables

We now expand the level-2 model to include a group-level predictor z_j, also called a contextual variable. We start from the slopes-as-outcomes model (3) and assume that the intercept and slope satisfy the level-2 models

$$\begin{aligned} A_j &= \alpha_0 + \alpha_1 z_j + a_j \\ B_j &= \beta_0 + b_j \end{aligned}$$

In other words, the baseline level of contraceptive use is a linear function of the contextual variable. If we substitute these equations in [3] we obtain

$$\text{logit}(\pi_{ij}) = (\alpha_0 + \alpha_1 z_j + a_j) + (\beta + b_j)x_{ij}$$

Thus, contextual variables are introduced in the linear predictor just like any other variable. The contextual predictors used in the analysis of contraceptive use were gross national product (GNP), an index of family planning effort (FPE) program, and their interaction.

Between and Within-Group Estimates

An important type of contextual variable is the group mean of an individual-level variable. It is perfectly possible to include in the model both the group mean $\bar{x}_j$ and the individual-level predictor, preferably as a deviation from the group mean $x_{ij} - \bar{x}_j$. In fact, there are often good reasons to do so. Specifically, consider the random-intercept model

$$\text{logit}(\pi_{ij}) = (\alpha + a_j) + \beta_B \bar{x}_j + \beta_W (x_{ij} - \bar{x}_j)$$

We use the subscripts B and W to emphasize that the coefficient of the group mean captures *between* group variation, and the coefficient of the individual predictor reflects *within* group variation. The within-group estimator is similar in spirit to a fixed-effects estimator and is usually very close in value (in fact, in linear models it is exactly the same). Moreover, it can be used in models that have no minimal sufficient statistic on which to base a fixed-effects estimator, including probit and negative binomial models. The random-effects estimator, obtained by omitting the group means, can be shown to be equivalent to a weighted average of the within- and between-estimators, and will obviously be appropriate (and more efficient) when the underlying parameters are the same. A test of $H_0{:}\beta_B = \beta_W$ is similar in spirit to a Hausman specification test; coefficients may be consolidated when the test accepts, or kept separate when the test rejects. For further discussion of this hybrid model, see Allison (2009) and Rabe-Hesketh and Skrondal (2012).

Cross-Level Interactions

The most general model allows the slope of an individual-level predictor to depend on a contextual variable, so the level-2 models become

$$A_j = \alpha_0 + \alpha_1 z_j + a_j$$
$$B_j = \beta_0 + \beta_1 z_j + b_j$$

If we substitute these equations in the basic model (3) we obtain

$$\text{logit}(\pi_{ij}) = (\alpha_0 + \alpha_1 z_j + a_j) + (\beta_0 + \beta_1 z_j + b_j) x_{ij}$$

And rearranging terms to collect the fixed and random coefficients we obtain

$$\text{logit}(\pi_{ij}) = (\alpha_0 + \alpha_1 z_j + \beta_0 x_{ij} + \beta_1 z_j x_{ij}) + (a_j + b_j x_{ij})$$

The fixed part of the model has main effects of the country- and individual-level predictors, as well as a *cross-level interaction* between country and individual, represented by the product $z_j x_{ij}$. The random part of the model depends on x_{ij}.

Wong and Mason (1985) allowed the coefficients of years of education and childhood urban residence to depend on GNP and FPE. Their final model, however, drops GNP and includes only one cross-level interaction between education and program effort, and is based on the following level-specific equations:

$$\text{logit}(\text{ever}_{ij}) = A_j + B_j\,\text{educ}_{ij} + C_j\,\text{urban}_{ij}$$
$$A_j = \alpha_0 + \alpha_1\,\text{fpe}_j + a_j$$
$$B_j = \beta_0 + \beta_1\,\text{fpe}_j + b_j$$
$$C_j = \gamma_0 + c_j$$

where the country-level residual (a_j, b_j, c_j) has a trivariate normal distribution with mean vector zero and unstructured variance–covariance matrix.

In words, the logit of the probability of ever use of contraception depends on the woman's years of education and childhood residence, the level and the education differential depend on the country's FPE, and all three coefficients have residual variation across countries.

Fixed Coefficients

The published estimates of the fixed parameters are reproduced in Table 1.

These coefficients can be interpreted in the usual way, exponentiating the estimates to obtain odds ratios. Thus, women who grew up in urban areas have 63% higher odds of having used contraception than women who grew up in rural areas, when we compare women with the same education who live in the same country. The coefficient of education requires some care because of the interaction with FPE. In countries with no program, each year of education is associated with 24% higher odds of ever use, everything else being equal. The education differential, however, is smaller at higher levels of effort, with an odds ratio of 1.157 at the mean level of effort (13) and an odds ratio of 1.092 at the highest observed level of effort (24). Finally, the odds of using contraception are higher in countries with higher levels of effort, with 9% higher odds per point of effort.

Variance Parameters

The random coefficients capture residual variation across countries in levels and differentials in ever use of contraception. The estimated variance–covariance matrix for these effects is presented in terms of standard deviations and correlations in Table 2.

We see substantial variation in levels of contraceptive use across countries after adjusting for observed covariates. I find it easier to interpret the results if we standardize the random

Table 1 Estimates of fixed coefficients

Variable	*Symbol*	*Coefficient*	*Standard error*	*Odds ratio*
Constant	α_0	−2.3695	0.4277	–
FPE	α_1	0.085	0.02685	1.088
Education	β_0	0.2147	0.03967	1.239
FPE × education	β_1	−0.005296	0.002416	0.995
Urban	γ_1	0.4909	0.1008	1.633

FPE = Family planning effort.

Table 2 Standard deviations and correlations of random effects

	Standard deviations		*Correlations*		
Variable	*Symbol*	*Value*	*Constant*	*Education*	*Urban*
Constant	σ_a	0.9017	1		
Education	σ_b	0.0681	−0.826	1	
Urban	σ_c	0.2419	−0.172	0.273	1

effects, so instead of $a_j \sim N(0, \sigma_a^2)$ we write $a_j = \sigma_a z_j$ where $z_j \sim N(0, 1)$. This shows that σ_a can be interpreted as an ordinary logit coefficient, exponentiating it to obtain an odds ratio. Thus, the odds of ever using contraception in a country that is one standard deviation above the mean are 2.5 times the odds for comparable women in the average country. Educational differentials also vary across countries from net of their level of effort and composition; the odds ratio per year of education in a country that is one standard deviation above the mean is 7% higher than in the average country. We also observe residual variation in urban–rural differentials, which are 27% larger in a country that is one standard deviation above the mean than in the average country, everything else being equal.

The correlations among the random effects are also interesting. Countries that have higher than expected ever use of contraception tend to have lower than expected differentials by education and, to a lesser extent, by childhood residence, net of their level of effort.

Empirical Bayes

We now consider estimating the random effects representing variation in levels and differentials across countries. In the context of our example, it would be useful to assess the extent to which ever use of contraception in a particular country is lower or higher than expected from its program effort and the education and childhood residence of its inhabitants. Similarly, we might be interested in ascertaining the extent to which the effect of education on ever use is larger or smaller than expected given its program effort. We contrast empirical Bayes (EB) and ML estimates.

The EB approach focuses on the posterior distribution of the random effects, which can be obtained from Bayes theorem as the product of the prior multivariate normal and the conditional likelihood or probability of the data given the random effects, which is product Bernoulli. The posterior depends on the regression coefficients and the variance–covariance matrix of the random effects. A full Bayesian analysis would assign prior (or hyper-prior) distributions to these as well. In contrast, an EB approach evaluates the posterior at the ML estimates of these parameters, and then takes the mean or mode. In logit models the calculation requires numerical integration.

An approach closer in spirit to ML uses the estimated linear predictor obtained by multiplying all predictors by their coefficients as an *offset* or known part of the model, and then estimates the random effects by running a separate regression in each group, in our example with education and urban as predictors. If one has few groups and they are reasonably large, it may be possible to simply run a separate regression in each group, which is what was done in this case, but the previous approach has wider applicability. For further details, see Rabe-Hesketh and Skrondal (2012).

Typically the Bayes estimates are closer to the mean than the corresponding ML estimates, a common property of Bayesian estimates known as 'skrinkage.' **Figure 1** illustrates this phenomenon by comparing EB and ML estimates of the urban–rural differential. The figure omits Kenya, where the initial estimate of nearly 2 was pulled all the way down to 0.7, much closer to the average. The EB estimates were also skeptical of the negligible or reverse differentials in Panama and Indonesia and pulled them up. The other countries show smaller differences, with EB estimates usually closer to the mean than ML. Generally speaking, EB estimates represent a compromise between having a separate estimate of the urban–rural differential for each country, and having a common estimate for all, with more weight given to the separate estimates where there is more information.

Predicted Probabilities

In logit models it is often useful to translate results into predicted probabilities, which are more easily understood than odds ratios. With multilevel models, results can be translated into two types of predicted probabilities: subject specific and population average. We illustrate by predicting ever use of contraception by childhood residence.

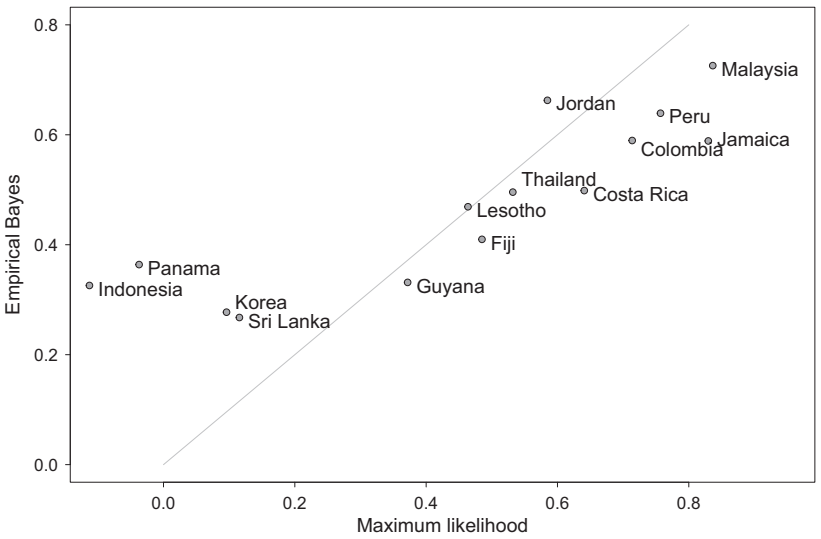

Figure 1 Estimates of urban–rural differential in ever use.

Subject-specific or conditional probabilities apply to an individual at given values of the fixed and random predictors, and can be easily obtained from the fitted equations. Calculations are typically done varying one variable while keeping all other predictors set to their means. In our case, we consider women who grew up in rural and urban areas have average education (3.43 years), and live in an average country, with an effort index of 13 and average unobserved characteristics. The linear predictor, excluding the urban coefficient, is $-2.3695 + 0.085 \times 13 + (0.2147 - 0.0053 \times 13) \times 3.43 = -0.7644$. The predicted urban and rural probabilities are thus

$$\text{logit}^{-1}(-0.7644) = 0.3177 \quad \text{and}$$
$$\text{logit}^{-1}(-0.7644 + 0.4909) = 0.4320$$

Population-average or marginal probabilities apply to an individual at given values of the fixed predictors but averaging over all random effects. In our example, we set childhood residence to rural and urban while keeping education and effort set to their mean values, but average over all country-level residuals. The calculations involve integrals of the form

$$\iiint \text{logit}^{-1}(-0.7644 + a + 3.43b + (0.4909 + c)u\phi(a,b,c)\,da\,db\,dc$$

where u is an indicator of urban childhood residence, and $\phi(a,b,c)$ is the trivariate normal density with mean zero and variance–covariance matrix given in **Table 2**. This integral can be approximated using Gaussian quadrature or by Monte Carlo integration, and leads to ever use estimates of 0.3347 for rural and 0.4399 for urban. In this case, the two kinds of predicted probabilities are very similar. The urban–rural differential in ever use averaged over countries is 0.105, compared to 0.114 in the average country. As usual, the marginal effect is smaller in magnitude than the conditional effect.

Running a logistic regression using generalized estimating equations, or an ordinary logit model that ignores the grouping but corrects the standard errors for clustering, produces population average estimates. A multilevel model can, of course, produce both.

Bibliographic Remarks

There are many applications of multilevel models to categorical outcomes in demography. Pebley et al. (1996) use three-level logit models to study prenatal care and childhood immunization in Guatemala. Rodríguez and Goldman (2001) use the same data as a case study to compare various estimation methods, including MCMC using the Gibbs sampler. Steele et al. (1999) look at the impact of family planning service provision on contraceptive use in Morocco. Ng et al. (2006) illustrate simulated ML using contraceptive prevalence data from Bangladesh. The data are used as an example in the Stata manuals and are therefore available for download. Mills and Begall (2010) look at intention to have a third child in 24 countries using a random-intercept-ordered logit model, which includes a cross-level interaction between a country-level gender gap index and the sex of the previous child. In a somewhat different vein, Schmertmann et al. (2013) estimate total fertility for many small areas in Brazil using a novel combination of EB estimates to smooth age-specific fertility rates with a Brass P/F correction adapted to work when fertility is declining rapidly.

Multilevel Models of Infant and Child Survival

Multilevel survival models are particularly important in demography, extending the classic ideas of unobserved heterogeneity of frailty of Vaupel et al. (1979) to include components shared by members of a group. Many parametric and semiparametric survival models can be extended in this way, but most applications to date have focused on discrete-time models, fitted via the connection with logistic regression (Allison, 1982), or continuous-time models with piecewise constant hazard, fitted via the connection with Poisson regression (Holford, 1980; Laird and Olivier, 1981). These approaches often give similar results, although they focus on probabilities and rates, respectively.

We illustrate the ideas with an analysis of infant and child mortality in Kenya described in more detail in Rodríguez (2008). We use a continuous-time hazard model with frailty shared at the family and community levels, fitted to data from the 1998 Kenya Demographic and Health Survey (KDHS). The analysis focuses on all births in the 10 years preceding the interview, but excludes 170 pairs of twins and one set of triplets, which have higher mortality than singletons. The final sample consists of 10 878 births to 4939 women who live in 530 communities, defined in terms of the ultimate area units in the sample. A limitation of the data is that the community is based on current residence, but 81% of the births occurred to mothers who have never moved.

A Three-Level Hazard Model

Let $\lambda_{ijk}(t)$ denote the instantaneous hazard of death at age t for the i-th child of the j-th mother in the k-th community. We assume that the hazard depends on age t; a set of observed child, family, and community characteristics x_{ijk}; and unobserved family and community effects a_{jk} and a_k, so the conditional hazard is

$$\lambda(t|x_{ijk}, a_{jk}, a_k) = \lambda_0(t)\exp\{x'_{ijk}\beta + a_{jk} + a_k\}$$

where $\lambda_0(t)$ is the baseline hazard, β is a vector of fixed parameters representing effects of observed covariates, and $a_{jk} \sim N(0, \sigma_2^2)$ and $a_k \sim N(0, \sigma_3^2)$ are random effects representing unobserved family and community frailty.

For estimation purposes we assume that the baseline hazard is constant in intervals defined by cutpoints $0 = \tau_0 < \tau_1 < \ldots < \tau_D = \infty$ so that $\lambda_0(t) = \lambda_{0d}$ if $t \in (\tau_{d-1}, \tau_d)$. We have some flexibility in choosing our intervals because the KDHS recorded age at death in days for deaths in the first month, months for deaths in the first year, and years thereafter. Following some exploratory work we settled on 0, 1–5, 6–11, 12–23, and 24–59 completed months, so we capture more detail at ages where the hazard is changing more rapidly. We then split each observation into one episode for each interval visited, and count events and exposure in each episode. The

analysis sample, censored at age five, comprises 48 094 pseudo-observations.

The predictors include one community-level variable, type of place of residence, classified as urban or rural; one family-level variable, years of education of the mother, which we model using linear and quadratic terms centered at seven; and five variables at the level of the child, all well-known risk factors: gender, represented by an indicator for males; cohort, with an indicator for births in 1993 or later; age of mother, represented by linear and quadratic terms centered on age 25; birth order, which we tried modeling using indicators for births of various orders, but was better captured using linear and quadratic terms centered on three; and length of the previous birth interval, which was best captured as $(30 - i)_+$, a linear spline term that is zero for intervals in excess of 30 months and increases linearly as the interval gets shorter.

Fixed Coefficients

The three-level piecewise-exponential proportional-hazards model just defined can be fitted via multilevel Poisson regression, with the counts of events as outcome and the logarithm of exposure time in each episode as an offset, in a direct extension of the results of Holford (1980) and Laird and Olivier (1981). Estimation using ML via adaptive quadrature leads to the results summarized in **Table 3**.

Interpretation of the fixed coefficients proceeds in the usual fashion, exponentiating them so they represent hazard ratios. The shape of the spline and quadratic fits may be better appreciated in **Figure 2**. Obviously, the hazard declines very substantially with age of the child, with the risk at age one being 95% lower than in the first month of life. We estimate a 9% higher risk for males than comparable females at the same age, but this effect is not significant. Children born in 1993, or later, have *higher* risk than children born earlier, after adjusting for all other factors, indicating a deterioration of survival prospects in the late 1990s. The lowest risk by age of mother occurs at age 32, but increases rapidly for older and particularly younger mothers, reaching 40% higher risk at ages 22 or 42, and even higher levels at younger ages. Birth order is not significantly associated with child survival. It is possible that the excess risk often observed for first births has been captured by age of mother. Short birth intervals are strongly associated with poor infant and child survival, with the hazard increasing 4% for each month that the preceding interval falls short of 30 months. This translates into 24% higher risk for children born 2 years after a sibling, compared to those born after an interval of 2.5 years.

Let us turn now to the family and community variables. Mother's education has a strong association with infant and child survival. Years of education ranges from 0 to 19, with quartiles at 4, 7, and 8. The fitted quadratic shows no effect of a few years of education, but a steady and large decline in risk after that. Around the median, each year of education is associated with a 7% decline in risk at any given age, after adjusting for other characteristics. We find no significant differences in survival by residence, with a point estimate of 4% higher risk in the urban sample. The erosion of the traditional survival advantage of urban areas could be associated with higher prevalence of AIDS in the cities.

Table 3 Parameter estimates for infant and child mortality in Kenya

Variable	*Term*	*Coefficient*	*Standard error*	*Hazard ratio*
Fixed coefficients				
Constant	1	−4.588	0.118	–
Age	1–5	−1.642	0.089	0.194
	6–11	−1.998	0.097	0.136
	12–23	−2.822	0.106	0.059
	24–59	−3.362	0.109	0.026
Sex	Male	0.087	0.068	1.091
Cohort	1993+	0.173	0.069	1.189
Mother's age	$a - 25$	−0.047	0.011	0.954
	$(a - 25)^2$	0.003	0.001	1.003
Birth order	$o - 3$	0.043	0.039	1.044
	$(o - 3)^2$	0.004	0.005	1.004
Interval	$(30 - i)_+$	0.036	0.006	1.037
Mother's education	$(e - 7)$	−0.068	0.015	0.934
	$(e - 7)^2$	−0.007	0.003	0.993
Residence	Urban	0.040	0.142	1.041
Variance parameters				
Family	σ_2	0.613	0.086	1.846
Community	σ_3	0.680	0.055	1.973

Random Effects

The most remarkable feature of these results, however, is the large amount of residual variation found at the family and community levels, with estimated standard deviations of 0.613 at the family level and 0.680 at the community level. Just like we did with logit models, we can interpret the standard deviations in essentially the same way as other coefficients, exponentiating them to obtain hazard ratios. To see this point write the family random effect as $a_{jk} = \sigma_2 z_{jk}$, where $z_{jk} \sim N(0, 1)$ is a standard normal random effect, and note that children who live in families with frailty that is one standard deviation above the mean have a log-hazard 0.613 higher than comparable children in the average family. This translates to 85% higher risk, everything else being equal. In the same fashion, we can write the community random effect as $a_k = \sigma_3 z_k$, where z_k is a standard normal random effect, so children who live in a community with frailty that is one standard deviation above the mean have a log-hazard of 0.680, higher than comparable children in the average community. This translates to 97% higher risk.

It is interesting to compare the magnitude of the observed and unobserved effects on infant and child mortality. One way to do this is to combine all the observed risk factors other than age of the child into a single measure of risk, effectively multiplying the terms representing urban residence, mother's education, sex, cohort, age of mother at birth of the child, birth order, and length of previous birth interval, by their coefficients. Using this summary measure, we find that children in the third quartile of relative risk have 61% higher risk than those in the first quartile. In contrast, the results reported for the family and community effects translate into 2.3- and 2.5-fold increases in risk across the interquartile range, respectively. More broadly, the range from the first to the 99th percentile in observed risk factors translates to an eight fold

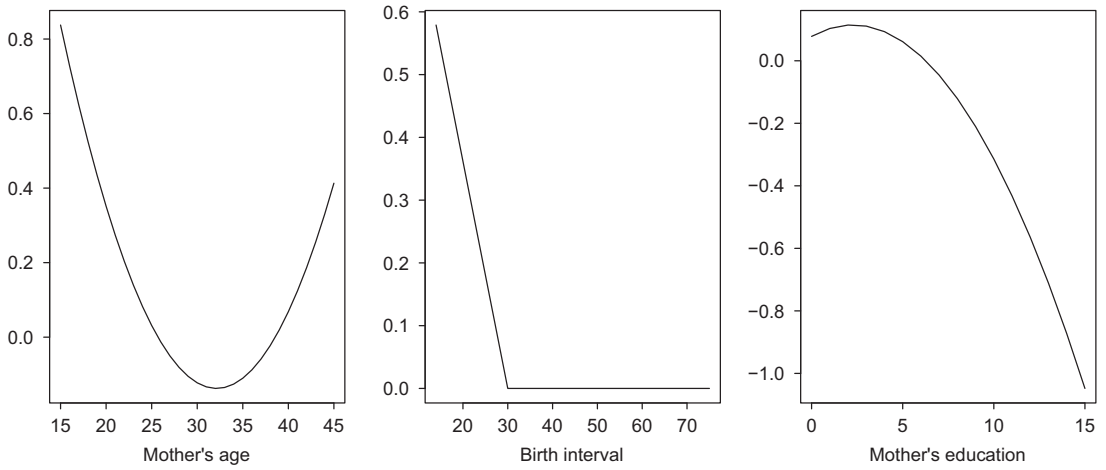

Figure 2 Log risk by mother's age, birth interval, and education.

increase in risk, whereas the equivalent ranges in family and community frailty translate into 17- and 24-fold increases. Clearly, there are very substantial risks associated with family and community characteristics that remain unobserved.

Conditional and Marginal Probabilities

With hazard models it is often useful to convert the results into more familiar survival probabilities. With multilevel survival models we have a choice between conditional and marginal probabilities, which lead again to subject-specific and population-average interpretations of the results.

Consider conditional probabilities first. These are often calculated by setting covariates to their means, but here we will set both the observed and unobserved effects to the first and third quartiles of their distributions. The calculation is based on the fact that in a proportional hazards model the survival function is given by

$$S(t|x_{ijk}, a_{jk}, a_k) = \exp\{-\Lambda_0(t)\exp\{x'_{ijk}\beta + a_{jk} + a_k\}\}$$

We accumulate the estimated baseline hazard up to ages one and five, set the observed relative risk $x_{ijk}\beta$ to the estimated quartiles (which are 0.15 and 0.63), and set the family and community frailties to the estimated normal quartiles $\pm 0.674\sigma_2$ and $\pm 0.674\sigma_3$. Figure 3 shows the results. The conditional probabilities of infant death range from 14 to 119 per 1000, and the probabilities of child death from 22 to 183, with differences by community and family frailty exceeding differences by observed risk. The probabilities of infant death for the first and third quartile of observed risk are 14 and 22 if the family and community are in the first quartile of frailty, but 75 and 119 if they are both in the third quartile of frailty.

We can also calculate marginal or population average probabilities of infant and child death at any given values of the observed covariates by integrating over the family and community frailties

$$S(t|x) = \iint S(t|x_{ijk}, a_{jk}, a_k)\, da_{jk}\, da_k$$

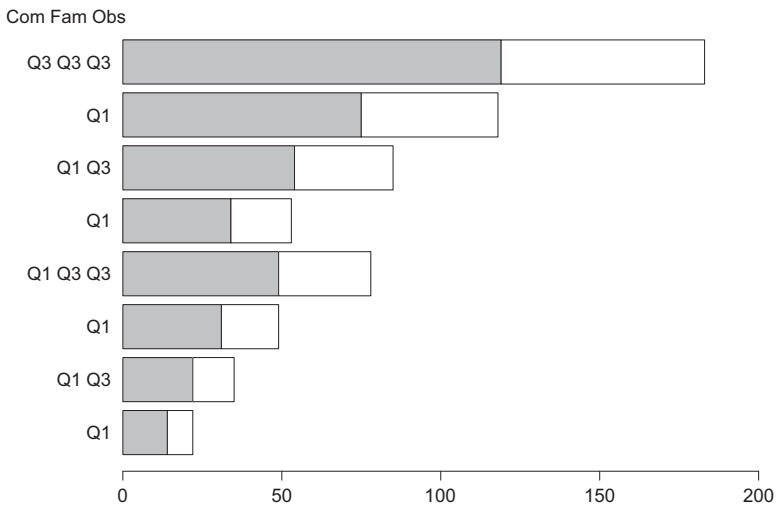

Figure 3 Probability of infant and child death by quartiles of frailty and observed risk.

Evaluation of the integral must be done using numerical integration routines; in this analysis we used 12-point Gauss–Hermite quadrature. We find average probabilities of infant death of 47 and 74 per 1000 for children at the first- and third-risk quartiles, respectively, when averaged across all families and communities. The corresponding probabilities of child death are 74 and 113 per 1000.

Intraclass Correlations

The variance parameters in random intercept models have a direct interpretation in terms of intraclass correlation. In the context of survival models, Oakes (1982) has shown that the variance parameter in two-level models with gamma-distributed frailty is closely related to Kendall's τ, a measure of ordinal association. These results have not been generalized to other settings, and our frailty distribution is in fact log normal, but we can always calculate probabilities of surviving to age one and five based on the results given above, and use these to estimate standard measures of correlation. These probabilities, and hence the derived measures of correlation, depend on the observed covariates. For these data, the Pearson correlations are very low at the lower end of the distribution of risk, but they increase very rapidly at the high end, with the intrafamily correlation reaching as high as 0.18 for infant and 0.21 for child mortality. Even among children of different families in the same community, the intraclass correlation climbs up to 0.10 at high-risk levels. For further details, please see the full analysis in Rodríguez (2008).

Bibliographic Remarks

Not surprisingly multilevel survival models have found many applications in demography. Guo (1993) uses sibling data in a two-level model to estimate child survival in Guatemala. Sastry (1997) uses a three-level model to look at individual, household, and community effects on child survival in Brazil. Steele et al. (1996) study duration of contraceptive use in China using a multinomial version of a discrete hazards model to handle competing risks. Barber et al. (2000) use a discrete-time three-level model to study the initiation of permanent contraceptive use in Nepal. Lillard (1993) was a pioneer in considering joint modeling of marriage and fertility. Following in this tradition, Steele et al. (2006) use multilevel/multiprocess models to study cohabitation and childbearing among young British women, allowing for endogeneity of fertility decisions. Pamuk et al. (2011) use multilevel logistic regression to study the likelihood of infant death using a three-level logit model with random intercepts at the community and country levels.

See also: Generalized Linear Mixed Models; Hierarchical Models: Random and Fixed Effects; Linear Mixed Models; Longitudinal Data: Event–History Analysis in Discrete Time; Statistical Analysis: Multilevel Methods; Survival Analysis: Whether, and If So, When?.

Bibliography

Allison, P.D., 1982. Discrete-time methods for the analysis of event histories. In: Leinhardt, S. (Ed.), Sociological Methodology. Jossey-Bass, San Francisco, pp. 61–98.

Allison, P.D., 2009. Fixed Effects Regression Models. Sage, Newbury Park, CA.

Barber, J.S., Murphy, S.A., Axinn, W.G., Maples, J., 2000. Discrete-time multilevel hazard analysis. Sociological Methodology 30, 201–235.

Bates, D.M., 2010. lme4: Mixed-Effects Modeling with R. http://lme4.r-forge.r-project.org/book/.

Browne, W.J., Draper, D., 2006. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Bayesian Analysis 1, 473–514.

Bryk, A.A., Raudenbush, S.W., 1992. Hierarchical Linear Models: Applications and Data Analysis Methods. Sage, Newbury Park, CA.

Burstein, L., 1980. The analysis of multilevel data in educational research and evaluation. Review of Research in Education 8, 158–233.

Clayton, D.G., 1996. Generalized linear mixed models. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds.), Markov Chain Monte Carlo in Practice. Chapman and Hall, London.

Entwisle, B., Mason, W.M., Hermalin, A.I., 1986. The multilevel dependence of contraceptive use on socioeconomic development and family planning program strength. Demography 23, 199–216.

Goldstein, H., 1995. Multilevel Statistical Models, second ed. Edward Arnold, London. third edition published in 2003.

Guo, G., 1993. Use of sibling data to estimate family mortality effects in Guatemala. Demography 30, 15–32.

Holford, T.R., 1980. The analysis of rates and of survivorship using log-linear models. Biometrics 36 (2), 299–305.

Laird, N., Olivier, D., 1981. Covariance analysis of censored survival data using log-linear analysis techniques. Journal of the American Statistical Association 76, 231–240.

de Leeuw, J., Meier, E., 2008. Introduction to multilevel analysis. In: de Leeuw, J., Meier, E. (Eds.), Handbook of Multilevel Analysis. Springer, New York, (Chapter 1) pp. 1–75.

Lillard, L.A., 1993. Simultaneous equations for hazards: marriage duration and fertility timing. Journal of Econometrics 56, 189–217.

Lin, X., Breslow, N.E., 1996. Bias correction in generalized linear mixed models with multiple components of dispersion. Journal of the American Statistical Association 91, 1007–1016.

Liu, Q., Pierce, D.A., 1994. A note on Gauss-Hermite quadrature. Biometrika 81, 624–629.

Longford, N.T., 1993. Random Coefficient Models. Clarendon Press, Oxford, UK.

Mills, M., Begall, K., 2010. Preferences for the sex-composition of children in Europe: a multilevel examination of its effect on progression to a third child. Population Studies 64, 77–95.

Ng, E.S., Carpenter, J.R., Goldstein, H., Rasbash, J., 2006. Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood. Statistical Modelling 6, 23–42.

Oakes, D., 1982. A model for association in bivariate survival data. Journal of the Royal Statistical Society, Series B 44, 414–422.

Pamuk, E., Fuchs, R., Lutz, W., 2011. Comparing education and economic resources: multilevel analysis of infant mortality in 43 developing countries. Population and Development Review 37, 637–644.

Pebley, A.R., Goldman, N., Rodríguez, G., 1996. Prenatal and delivery care and childhood immunization in Guatemala: do family and community matter? Demography 33, 231–247.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-plus. Springer, New York.

Rabe-Hesketh, S., Skrondal, A., 2012. Multilevel and Longitudinal Modeling Using Stata, third ed. StataCorp LP, College Station, TX.

Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2002. Reliable estimation of generalized linear mixed models using adaptive quadrature. The Stata Journal 2, 1–21.

Rodríguez, G., 2008. Multilevel generalized linear models. In: de Leeuw, J., Meier, E. (Eds.), Handbook of Multilevel Analysis, Chapter 9. Springer, New York, pp. 335–376.

Rodríguez, G., Elo, I., 2003. Intraclass correlation in random-effect models for binary outcomes. The Stata Journal 3 (1), 32–46.

Rodríguez, G., Goldman, N., 1995. An assessment of estimation procedures for multilevel models with binary response. Journal of the Royal Statistical Society, Series A 158, 73–89.

Rodríguez, G., Goldman, N., 2001. Improved estimation procedures for multilevel models with binary response: a case study. Journal of the Royal Statistical Society, Series A 164, 339–355.

Sastry, N., 1997. A nested frailty model for survival data, with an application to the study of child survival in northeast Brazil. Journal of the American Statistical Association 92, 426–435.

Schmertmann, C.P., Cavenaghi, S.M., Assunção, R.M., Potter, J.E., 2013. Bayes plus Brass: estimating total fertility for many small areas from sparse census data. Population Studies 67 (3), 255–273.

StataCorp, 2013. Stata Statistical Software: Release 13. StataCorp LP, College Station, TX.

Steele, F., Curtis, S.L., Choe, M.K., 1999. The impact of family planning service provision on contraceptive use dynamics in Morocco. Studies in Family Planning 30, 28–42.

Steele, F., Diamond, I., Wang, D., 1996. The determinants of the duration of contraceptive use in China: a multilevel multinomial discrete hazards modelling approach. Demography 33, 12–33.

Steele, F., Kallis, C., Joshi, H., 2006. The formation and outcomes of cohabiting and marital partnerships in early adulthood: the role of previous partnership experience. Journal of the Royal Statistical Society, Series A 169, 757–779.

Vaupel, J.W., Manton, K.G., Stallard, E., 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16, 439–454.

Wong, G.W., Mason, W.M., 1985. The hierarchical logistic regression model for multilevel analysis. Journal of the American Statistical Association 80, 513–524.

Multilingual Language Development

Alejandro E Brice, University of South Florida St. Petersburg, St. Petersburg, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Typical language development has far too long been predicated on monolingual development. It is estimated that approximately more than half of the world's population speaks more than one language. Multilingual or multilanguage acquisition differs from monolingual development. It involves more layers that affect ultimate acquisition including second and/or multiple language acquisition factors, environmental factors, individual factors, developmental factors, and, if the child has disabilities, then disabilities are also confounding factors. In addition, the following factors influencing multilingual language acquisition will be discussed: (1) second and/or multiple language acquisition; (2) language dominance and proficiency; (3) language maintenance; (4) language model; (5) language deceleration; (6) cross-linguistic influence, that is, language transference; (7) cross-linguistic influence, that is, language interference; (8) cross-linguistic influence, that is, code switching and code mixing; (9) ultimate attainment; (10) language loss (attrition); (11) language fossilization; (12) environmental factors; (13) individual and developmental factors; (14) biological, neurological aspects of multilingualism; (15) disability factors; and (16) strategies for promoting multilingualism. This article will briefly present these overlaying components and provide suggestions for language professionals in promoting multilingualism.

Introduction

The notion of typical language development has far too long been predicated on monolingual development. It is estimated that approximately more than half of the world's population speaks more than one language (Grosjean, 2010). Multilingualism is the norm worldwide. Thus, the monolingual view to language development is inadequate to describe language acquisition from a global perspective. In addition, the viewpoint that multilingual language development mirrors monolingual language when only one of the child's languages is considered is also a misguided assertion (Brice and Brice, 2009; Grosjean, 1989).

Multilingual or multilanguage acquisition differs from monolingual development (Brice and Brice, 2009). It involves more layers that affect ultimate acquisition including: (1) second and/or multiple language acquisition factors, (2) environmental factors, (3) individual factors, (4) developmental factors, and (5) if the child has disabilities, then disabilities are also confounding factors. This article will briefly present these overlaying components and provide suggestions for language professionals in promoting multilingualism. See **Figure 1**.

Second and/or Multiple Language Acquisition

Learning a second language (L2) or multiple languages is not a straightforward additive process to English. It does not involve a simple one-to-one correspondence. Grosjean (1989) wrote that a bilingual speaker does not consist of two monolinguals speakers. Rather a bilingual speaker is an individual who has access to two languages that interact with each other constantly, even when one language is the primary mode being activated (Grosjean, 1989). When young children acquire two languages simultaneously from birth, their first language (L1) is bilingualism. Their two languages interact in positive and sometimes negative ways from the very first day they are exposed to the two languages (i.e., language transference and interference). Research has indicated that young multilingual children can appropriately code switch as early as 18 months of age when spoken to in L1 or L2 (Brice and Wertheim, 2004/2005; Döpke, 1992; Pearson and Fernandez, 1994; Pearson et al., 1997; Pearson et al., 1995; Wapole, 2000). Hence, no language confusion appears to be occurring for typically developing children.

Learning a third language (L3) differs from learning an L2 (Jessner, 2006). Third language acquisition (TLA) requires different skills than those used by bilingual (L2) learners. Jessner (2006) stated that TLA should be the basis for studying bilingual and monolingual language acquisition, and not vice versa. Cenoz (2000) illustrated that TLA differs from second language acquisition (SLA) in terms of how the languages may be acquired. In SLA, the L2 may be acquired simultaneously with L1, or sequentially with L1, whereas in

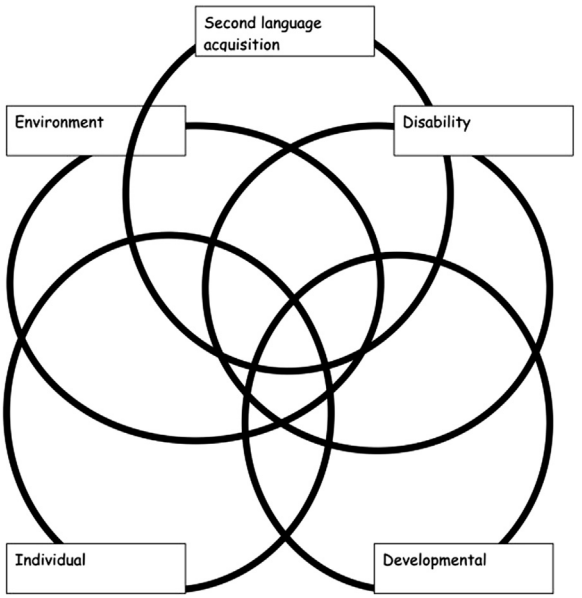

Figure 1 Interaction of bilingual factors.

TLA, four different conditions exist: (1) L1, L2, and L3 may be acquired simultaneously; (2) L1, L2, and L3 may be acquired sequentially; (3) L2 and L3 are acquired after L1; and (4) L1 and L2 are acquired before L3. In addition, the acquisition process of any of the languages may be halted by the process of acquiring any of the other two languages. The interruption process may apply to L1, L2, or L3 at any point in time; consequently, the TLA becomes more complex than either L1 or L2 acquisition. Due to the possible interruption of learning, individuals may also later relearn any of their interrupted languages. Chronological acquisition of L1–L3 does not necessarily equate with proficiency in any of the languages (Hufeisen, 1997). However, Jessner (2006) stated that in TLA, related languages tend to transfer learning more than unrelated languages. Consequently, cross-linguistic influences (CLIs) in TLA are more complex than in SLA.

Language Dominance and Proficiency

The concepts of language proficiency and language dominance are notions that seem to imply that one should consider each language separately instead of in a combined fashion. Hence, proficiency should be investigated in terms of communicative competence (Canale and Swain, 1980; Cazden et al., 1972) and proficiency utilizing a combined approach as is seen in vocabulary studies where total vocabulary and total conceptual vocabulary measures are obtained (Pearson and Pearson, 2004). Dominance varies across speech and language domains (phonetics, phonology, morphology, semantics, pragmatics, etc.) and also seems to vary within each speech domain. Flege and Eefting (1987) demonstrated that multilingual speakers are capable of creating different phonetic categories for voiceless stops across languages. Grosjean's (2001) notions of language modes fits this paradigm. Language modes refer to what extent L1 and L2 are activated and to the continuum of activation among the languages. Bullock et al. (2006) summarize this by stating, "Such research contends that bilinguals' language use is malleable in that they may behave differently according to which language they are producing or perceiving at a given time" (p. 9). Hence, dominance in a multilingual child is not an overall ability assigned to L1, L2, or L3 but one that seems to vary according to language, language domain, and task.

Language Maintenance

> My husband and I adopted twin sons a little over 2 years ago. They were 8 months old when we brought them home from their native country Cambodia, where the only language they had been exposed to was Khmer. My husband and I are both Hispanic in culture, language and heritage. Even though we live in the *Southern USA* [name changed] we speak Spanish at home and with our Hispanic friends as well as, of course, family. We have always wanted to raise our children to be multilingual since that has been both of our experiences. We know that the speech delay they have is very understandable given that they are boys, twins and have been exposed to several languages. A number of English-speaking friends and professionals have, suggested changing to English only at home.

> A few mothers of bilingual children have formed a website and we exchange information we learn out in the 'real world' of raising our own bilingual children. The question is, a few 'professionals', have asked that some parents stop the second language at home to develop the English or whatever the community language is. What is the standard advice? Why do we get this question so often? This really goes against our instincts to not speak our own language to our children.

The above solicitations for advice came from two mothers of multilingual children. The first mother was raising two boys with some speech and motor delays, while the second mother was raising her typically developing multilingual child. Yet, the essential question that they asked was the same, i.e., can we speak the native language at home and raise our children in a multilingual environment? It also appears that they have been given uninformed advice to speak only the outside language at home (English in the first case as this came from the United States and French in the second case as this came from a parent in France). What does the research say and what are the logical consequences of speaking the native language in the home versus speaking the outside language (e.g., English) in the home environment? Research has documented that multilingualism has cognitive and linguistic advantages (Bialystok, 2007; Hakuta, 1986). Research with multilingual children with disabilities has also shown that multilingualism does not have any detrimental effect on language development even with children who have prominent language disabilities as seen in Down syndrome (Kay-Raining Bird et al., 2005). Kay-Raining Bird et al. stated that,

> There was no evidence of a detrimental effect of bilingualism. That is, the bilingual children with DS scored at least as well on all English tests as their monolingual DS counterparts. Nonetheless, there was considerable diversity in the second- language abilities demonstrated by these individuals with DS. (p. 187)

Kohnert (2008) stated that in multilingual homes there exists "… the availability of rich language in the environment and diverse opportunities to develop and use a particular language for meaningful communicative interactions" (p. 10). Thus, maintaining the native language allows parents to communicate with their children in perhaps the best language medium available to them, i.e., their own native language. In addition, speaking in the other language (e.g., English) may diminish communication and possibly provide a less-than-rich experience and inadequate language model. It will take a child between 5 and 6 years to acquire nativelike classroom language abilities in English (Thomas and Collier, 1997). Speaking only English at home will not accelerate English learning and may in fact delay English acquisition due to a less rich language input. Therefore, why do professionals advocate such advice when it is not supported by the research and may go against the parent's instincts? It may be ignorance. For example, according to popular misconception it is believed that when two individuals are speaking in another language they are talking about you. As language specialists,

we must take a broader view (i.e., bilingual or multilingual view) to language use and development. It is our belief that multilingualism is a positive attribute and should be encouraged with all individuals (especially, those with language and learning disabilities). In sum, all aspects of multilingualism need to be understood.

Language Mode

The multilingual speaker has access to both their interacting languages at all times. Hence, the multilingual speaker never truly shuts one language off even when speaking in monolingual fashion. A multilingual speaker has access to two phonologies, two lexicons, two syntactic systems, and two pragmatic manners of interaction. They can choose which system may best convey their message. A more appropriate means of classifying a multilingual speaker's language abilities is to think of neither their L1 nor their L2 abilities but rather, to think of their abilities as a third combined language form that occurs at all times. The combined form is a continual interaction and combination of L1 and L2. The combined form cannot be subtracted from L1 or L2. Sometimes, the combined form involves code switching, or code mixing. At other times, the combined form occurs even when in monolingual mode. When the multilingual speaker is speaking solely L2 (e.g., English) or L1 (e.g., Spanish), their brain is always in the combined mode (i.e., multilingual).

The implication for language professionals is that any assessment of a multilingual child's learning and/or language abilities should include their combined dual language abilities. If the language professional is fortunate to have results from both languages, then these results must not be looked at separately. Assessment and test results should be investigated as combined L1 and L2 efforts (i.e., combined mode). The language specialist should look at both languages and their combined receptive and expressive language abilities in oral and written language, phonology, morphology, semantics, syntax, and semantics. Poplack (1980) postulated that a multilingual speaker has to abide by both languages and their grammar rules when code switching and code mixing. We will postulate that even when the multilingual speaker is communicating in solely one language, he or she is keeping track of both language grammars simultaneously. An example of this is seen in momentary language interference when a multilingual speaker attempts to read an English word as a Spanish word, or when a multilingual speaker displays a broader definition of vocabulary in one language because they have accessed the vocabulary meaning in both languages (e.g., 'father' in English, 'padre' in Spanish, 'padrastro' or stepfather in Spanish; hence, 'father' can also encompass a broader meaning to include 'stepfather').

Consequently, the term 'multilingual' does not accurately reflect the complexity of the language system of the child who has multilingualism as L1. Multilingualism is not simply the use of two languages, but rather, use of three language modes (L1, L2, and their combined languages). A multilingual individual always operates in some varying degree of combined mode because L1 and L2 cannot be completely deactivated or separated. See Figure 2.

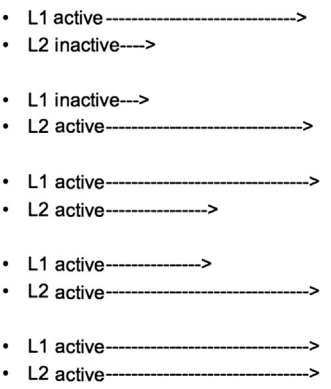
• L1 active---------->
• L2 inactive---->
• L1 inactive--->
• L2 active---------->
• L1 active---------->
• L2 active---------->
• L1 active---------->
• L2 active---------->
• L1 active---------

languages, negative consisting of interference among the languages, or simply interaction of the languages (such as code switching or code mixing) (Jessner, 2006). Language transference benefit occurs when bilingual speakers are able to attain a high level of proficiency in both languages (Cummins, 1984). Regarding L3 learners, high levels of proficiency would need to be obtained among two of their three languages (e.g., L1–L2, L1–L3, or L2–L3). It is also conceivable that transference would occur among all languages if high levels of proficiency were obtained in all languages.

Phonetically speaking, Flege et al. (2003) have shown that the phonetic systems of bilinguals interact. Age of arrival to the country (e.g., early vs late) and amount of L1 use determined the extent to which speech production of English was influenced (vowels in this study). With regard to vocabulary, research by Pearson and Pearson (2004) recommend the use of Total Conceptual Vocabulary so that students will not be identified as having small vocabularies. Therefore, multilingual vocabulary should only be considered when both languages are measured jointly. Montrul (2006) documented that Spanish-English-speaking adults were capable of achieving balanced syntactic abilities in both of their languages. These results indicate that the adults showed robust knowledge of both L1 and L2 syntactic structures, something that would not have been possible if the languages were separate and did not interact. Hence, positive transference seemed to have occurred at the syntactic level.

The question then arises not if dominance and proficiency are suitable theoretical constructs appropriate for multilingual students, but do the bilingual speakers catch up or attain ultimate proficiency in both of their languages?

CLI: Language Interference

Children can and do 'fall between the cracks.' Language interference, language loss, and fossilization can and do occur with SLA. Language interference is the negative influence that a language may have on subsequent languages. It is the imposition of one language onto another. An example is where the Spanish speaker may pronounce a 'th' as a 'd' sound. This is because the 'th' sound does not occur in Spanish of the Americas (excluding Castilian Spanish from Spain). The multilingual speaker will utilize a close approximation from their native language (L1). This interference is normal and will typically disappear given time, adequate English exposure, and perhaps instruction (Cummins, 1984). Language loss is when one of the two languages (typically, the native language) diminishes in growth or disappears. Some of our research with multilingual elementary-aged students (third-, fourth-, and fifth-grade students) indicates that this occurs with typically developing students (i.e., no disabilities present) (Brice et al., submitted). Language loss has been well documented by others (Selinker, 1991). Please note that multilingual children can also shift toward greater English vocabulary with increased English exposure (Kohnert and Bates, 2002; Kohnert et al., 1999).

Brice and Montgomery (1996) showed that language-impaired students and English language learners (ELLs) may display pragmatic problems in the classroom, albeit due to different causes. The ELL students demonstrated pragmatic communicative competence and were in the initial stages of acquiring English. Consequently, the Spanish-English-speaking students were accessing both their Spanish linguistic and pragmatic skills in order to accomplish pragmatic proficiency, yet had not established typically developing levels of monolingual children. Some pragmatic interference errors were noted for the ELL students.

CLI: Code Switching and Code Mixing

Code switching and code mixing are two forms of language alternation exhibited in multilingual speakers' language. It is a typical occurrence and does not signal language difficulties (Brice and Brice, 2009; Grosjean, 2010). Code switching is when the alternation occurs across sentence boundaries (intersentential alternations) as in the sentence, "I'm going to read the directions. Prestes atención? (pay attention)." The switch occurs between the English and Spanish sentences. Code mixing occurs within a sentence (intrasentential alterations). It is developmentally and linguistically more complex of a behavior (Brice and Brice, 2009). An example is when the teacher says, "Open your books to pagina diez (page ten)." Code switching and code mixing behaviors are exhibited along a continuum of language proficiency abilities, i.e., individuals may exhibit low-level language alternation behaviors, midlevel alternation behaviors, or high-level language alternation behaviors.

Poplack (1980) says that a multilingual speaker when code switching/mixing, has to abide by the language rules of all languages. Hence, code switching and code mixing are rule-governed behaviors. A third form of language alternation is the use of loan or borrowed words from another language. However, with loan/borrowed words, one does have to be multilingual as the borrowed word becomes completely incorporated into the guest language (e.g., "I took karate lessons at the dojo").

Ultimate Attainment

Bialystok et al. (2010) stated that:

> Bilingual children need to distribute their language-learning time across two languages and it is likely that some words occur in a context in which they only use one of their languages. In this sense, there is little reason to think that bilingual children are compromised in their expressive ability and every possibility that their **combined vocabulary is equivalent to or greater than the vocabulary of monolingual children** [emphasis added]. (p. 525)

Lugo-Neris et al. (2010) found that children with greater language proficiency in one of their two languages made the most gains from vocabulary intervention. Students with low-level aptitudes in both their languages did not show equal gains as students with better Spanish proficiency. Consequently, ultimate attainment abilities may depend on robust L1 skills. Bialystok et al. (2010) determined that

multilingual students are not hindered in school and academic language and attained ultimate vocabulary attainments for school words. In sum, it appears that ultimate vocabulary attainment is possible given a strong L1, sufficient exposure and adequate vocabulary use, and sufficient time acquiring their two languages (Bialystok et al., 2010; Gathercole, 2010; Lugo-Neris et al., 2010).

Pham and Kohnert (2013) found that bilingual children's English skills overtook their Vietnamese receptive and expressive vocabulary skills at approximately 9 and $7^1/_2$ years of age, respectively. While both languages increased in vocabulary, English became the more dominant language. Expressive skills lagged receptive English proficiency levels. This would indicate that ultimate attainment can be achieved in English; however, this may come at a risk of lesser native language skills (especially if the native language is not fully developed).

Research investigating speech perception abilities of multilingual adults indicates that some bilinguals are capable of attaining equal or ultimate attainment abilities in their L2 (Brice and Brice, 2008). However, this ability was found among only middle bilinguals (i.e., sequential bilinguals who had begin their English acquisition between 9 and 15 years of age and also been exposed to advanced English levels for at least 6 years). It should be noted that most bilinguals achieve, at minimum, functional abilities in both languages. Consequently, it appears that adequate exposure and maintenance to both languages is requisite for L2 development.

Language Loss (Attrition)

Under ideal conditions (e.g., near equal amounts of both languages such as 60% English and 40% Spanish) balanced multilingualism can result (Brice and Wertheim, 2004/2005; Pearson et al., 1997, 1993, 1995). However, language attrition typically occurs under less-than-perfect multilingual conditions (Anderson, 1999; Seliger and Vago, 1991). Given that English is the predominant language in schools, in the media, and for students, it serves as the primary means of communication with teachers and most friends; it seems very plausible that English becomes dominant and that Spanish becomes targeted as a minority language. Consequently, the home language will develop minority status and be at risk for language loss. The acquisition of English while losing the home language has dramatic language learning and educational consequences: (1) the child's language development stalls and resembles a language learning disability, (2) communication with both parents may weaken, (3) it takes a significantly longer time to fully develop English academic language skills (Thomas and Collier, 1997), and (4) positive transference of language skills from L1 to L2 may be lessened (Cummins, 1984, 1998).

Cummins (1984, 1998) stated the existence of a threshold hypothesis whereupon the student must achieve certain proficiency levels in their L1 before language transfer occurs and before the cognitive benefits of balanced multilingualism can be attained. If this threshold is not attained then what may result is language loss (or subtractive multilingualism).

Language loss among typically developing multilingual students may be more common than many educators may suspect. It is to the benefit of the child and the school that multilingual, English learners maintain their home language (i.e., less referrals for speech–language special education interventions as these students may appear to be language learning disabled). Research has demonstrated that promoting multilingualism promotes English language development and acquisition (Brice and Wertheim, 2004/2005; Döpke, 1992; Pearson and Fernandez, 1994; Pearson et al., 1997, 1995; Wapole, 2000). Brice et al. (2006) offered some suggestions in educating ELLs. Some specific language recommendations include: (1) modelling correct English language, (2) providing grammar drills with direct instruction, and (3) allowing for code switching to occur. In conclusion, language professionals should promote multilingualism and encourage teachers, family members, and students to be proud of their native language and to use it often.

Language Fossilization

Language fossilization is where growth in English (or the L2) languishes or stops short (Selinker, 1972). Both language loss and fossilization are real possibilities for multilingual children. Acquiring English and maintaining the native language are not automatic guarantees for multilingual children. Selinker (1972) first defined the concept of fossilization as an end state (final stage) and steady state (continuous stage) of language abilities that are characterized by stagnated growth in L2 development. Selinker (1991) refers to all stages of growth between two languages (until ultimate attainment is achieved) as interlanguage (IL) or intermediate stages of L2 learning. Therefore, fossilization is a permanent cessation of IL learning. Han (2004) identifies several possible causal factors for fossilization: (1) lack of instruction, (2) lack of language input, and (3) satisfactory communicative abilities.

Environmental Factors

According to Wong-Fillmore (1992), multilingualism (i.e., high levels of proficiency and balanced abilities) is possible only when certain conditions are met:

1. *Need to communicate.* The child and others must fulfill communicative intents. Motivation and desire to communicate are integral aspects. Language professionals must provide opportunities for real communication.
2. *Access to language from speakers of that language.* Multilingual individuals must have access to English and this English must be of a high level.
3. *Interaction, support, and feedback from speakers of that language must be present.* This interaction and feedback must be adequate for the learning to occur.
4. *Time.* It takes considerable time to adequately learn an L2, especially beyond just conversational abilities (Cummins, 1984; Thomas and Collier, 2002). It may take from 4 to 6 years or 5 to 7 years to acquire nativelike abilities for classroom language. Thomas and Collier state that this occurs only if the child has a strong base in their L1; otherwise, it may take up to 10–13 years. Again, support of the native language is crucial for English development.

Individual and Developmental Factors

The child's cognitive aptitude and also stages in their language, cognitive, and physical development play roles in L2 learning. It is apparent that gifted and talented students may experience accelerated learning in acquiring English (Brice et al., 2008). Just as developmental stages affect L1 development, these processes and stages also affect SLA. It is beyond the scope of this article to detail all these processes. However, it must be mentioned that the interaction of developmental stages among two developing language systems is more complex and varied than with monolingual development.

In sum, Kohnert (2013) states that bilingual/multilingual students must have the means, opportunities, and motivation to acquire their languages and develop proficiency and proficiency may vary in each of the languages.

Biological and Neurological Aspects of Multilingualism

Language and cognition are multiphasic processes (i.e., language and thinking develop in different stages in brain development) (Sakai, 2005; Uylings, 2006). Most brain growth occurs after birth, i.e., postnatally (Eliot, 2001). Evidence indicates that various aspects of brain development occur into late adolescence and/or early adulthood: (1) white matter increases from 4 to 21 years of age (Guidd, 2004); (2) myelin growth continues into adulthood (Uylings, 2006); (3) synaptogenesis (i.e., creation of new synapses) reaches maturation around 10–15 years of age (Uylings, 2006); (4) neurotransmitters such as dopamine, noradrenaline, and serotonin reach maturation around adolescence or early adulthood, (5) acetylcholinesterase (a key enzyme in the nervous system) matures around 18 years of age or later (Uylings, 2006); and (6) neuron dendrites and synapses and brain neurotransmitters are not expected to reach full development until adolescence or later (Uylings, 2006). Consequently, it appears that many developing brain capabilities occur in late adolescence or early adulthood, which would favor a sensitive and noncritical period to cognition and learning language (meaning that language learning neither reaches a plateau nor ceases around puberty). Neurologically, it appears that L2 learning can occur in adulthood (Grosjean, 2010).

Disability Factors

The most perplexing question among monolingual language professionals is whether bilingual or multilingual children with disabilities can acquire all languages, i.e., near-nativelike abilities. Clinical experiences and research support the notion that these children are fully capable of language and cognitive growth in their two language systems (Brice and Brice, 2009; Kohnert, 2008). The limiting factor appears to be the extent of their disability and not their two languages.

Several studies investigating typically developing monolingual children, typically developing bilingual children, monolingual children with disabilities, and bilingual children with disabilities have demonstrated similar patterns (Brice and Brice, 2009; Windsor et al., 2010). Brice and Brice (2009) found that when comparing phonemic awareness and phonics skills among kindergarten students a descending order of correctness emerged among the following groups: (1) monolingual typically developing students, (2) bilingual typically developing students, (3) monolingual students with disabilities, and (4) bilingual students with disabilities. This order of abilities was also found when comparing receptive syntax and directions for 8-year-olds: (1) English only, (2) bilingual, (3) English only – primary language impairment, and (4) bilingual – primary language impairment.

Strategies for Promoting Multilingualism

Research supports the notion that multilingualism may be an advantage (Hakuta and Bialystok, 1994). Consequently, the following points are for making learning multiple languages an additive process:

1. Language specialists should make learning English in the classroom and in social environments a positive experience in which both languages are valued. High-level language and thinking should be involved. Multilingual children should be encouraged to participate in verbal and literacy problem-solving tasks. Speech-language pathologists (SLPs) should verbalize how they solve problems to provide a cognitive model.
2. Multilingualism should be promoted both at home and in the school. Children should be allowed to speak either Spanish or English when it facilitates their communication (Brice and Perkins, 1997). The children's comfort level for using English should be allowed to mature and not be a forced issue.
3. Children should have well-developed Spanish skills before learning English (Wong-Fillmore, 1992).
4. Opportunities for reading and writing in both Spanish and English should be provided. Parents should be encouraged to read to their children in Spanish. Emergent literacy skills, knowing about the written word and books, are generalizable across languages. Appreciation of books is not language specific.
5. Ample opportunities to interact with native Spanish speakers should be provided to maintain L1.
6. The children should receive appropriate instruction in English. Learning English should reflect natural language usage.
7. Children should be allowed to make errors in English as this is a natural phenomena.

Conclusion

Working with second and multiple language learners with and without disabilities provide extra challenges for language professionals. Some of these challenges are unique, yet, neither teaching nor therapy has been an easy task. Consequently, we must provide high-quality services to all our students, clients, and/or patients. It is in our nature as language professionals to overcome language barriers be it a result of multilingualism, disability, or both.

See also: Child Care and Development across Cultures; First Language Acquisition, Developmental Psychology of; Neurophysiological Markers of Language Development.

Bibliography

Anderson, R., 1999. Impact of first language loss on grammar in a bilingual child. Communication Disorders Quarterly 21 (4), 4–16.

Bialystok, E., 2007. Cognitive effects of bilingualism: how linguistic experience leads to cognitive change. The International Journal of Bilingual Education and Bilingualism 10 (3), 210–223.

Bialystok, E., Luk, G., Peets, K., Yang, S., 2010. Receptive vocabulary differences in monolingual and bilingual children. Bilingualism: Language and Cognition 13 (4), 525–531. http://dx.doi.org/10.1017/S1336728909990423.

Brice, A., Brice, R., 2008. Examination of the critical period hypothesis and ultimate, attainment among Spanish-English bilinguals and English-speaking monolinguals. Asia Pacific Journal of Speech, Language and Hearing 11 (3), 143–160.

Brice, A., Brice, R. (Eds.), 2009a. Language Development: Monolingual and Bilingual Acquisition. Merrill/Prentice Hall, Old Tappan, NJ.

Brice, R., Brice, A., 2009b. Investigation of phonemic awareness and phonic skills in Spanish-English and English speaking kindergarten students. Communication Disorders Quarterly 30 (4), 208–225. http://dx.doi.org/10.1177/1525740108327448.

Brice, A., Leung, C., Gorman, B., submitted. Vocabulary skills of Spanish-English speaking elementary students: A pilot study. Perspectives on Communication Disorders in Culturally and Linguistically Diverse Populations.

Brice, A., Miller, K., Brice, R.G., 2006. Language in the English as a second, language and the general education classroom: a tutorial. Communication Disorders Quarterly 27, 240–247.

Brice, A., Montgomery, J., 1996. Adolescent pragmatic skills: A comparison of Latino students in ESL and speech and language programs. Language, Speech, and Hearing Services in Schools 27, 68–81.

Brice, A., Perkins, C., 1997. What is required for transition from the ESL classroom to the general education classroom? A case study of two classrooms. Journal of Children's Communication Development 19 (1), 13–22. http://dx.doi.org/10.1177/152574019701900103.

Brice, A., Shaunessy, E., Hughes, C., Mchatton, P.A., Ratliff, M.A., 2008. What language discourse tells us about bilingual adolescents: a study of students in gifted programs and students in general education programs. Journal for the Education of the Gifted 32 (1), 7–33.

Brice, A., Wertheim, E., 2004/2005. Language differentiation in young bilingual, children. Tejas. Texas Journal of Audiology and Speech Language Pathology 28, 24–31.

Bullock, B.E., Toribio, A.J., González, V., Dalola, A., 2006. Language dominance, and performance outcomes in bilingual pronunciation. In: O'Brien, M.G., Shea, C., Archibald, J. (Eds.), Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference. Cascadilla Proceedings Project, Somerville, MA, pp. 9–16.

Canale, M., Swain, M., 1980. Theoretical bases of communicative competence to, second language teaching and testing. Applied Linguistics 1 (1), 1–47. http://dx.doi.org/10.1093/applin/1.1.1.

Cazden, C., John, V., Hymes, D., 1972. Functions of Language in the Classroom. Teachers College Press, New York.

Cenoz, J., 2000. Research on multilingual acquisition. In: Cenoz, J., Jessner, U. (Eds.), English in Europe: The Acquisition of a Third Language. Multilingual Matters, Clevendon, pp. 39–53.

Cummins, J., 1984. Bilingualism and Special Education: Issues in Assessment and Pedagogy. College-Hill Press, San Diego.

Cummins, J., 1998. Beyond Adversarial Discourse: Searching for Common Ground in the Education of Bilingual Students. Presentation to the California State Board of Education, Sacramento, CA. Retrieved February 23, 2006 from http://ourworld.compuserve.com/homepages/JCRAWFORD/cummins.htm.

Döpke, S., 1992. A bilingual child's struggle to comply with the 'one parent-one language' rule. Journal of Multilingual and Multicultural Development 6, 467–485.

Eliot, L., 2001. Language and the developing brain. The NAMTA Journal 26 (2), 8–60.

Fabiano-Smith, L., Barlow, J., 2009. Interaction in bilingual phonological acquisition: evidence from phonetic inventories. International Journal of Bilingual Education and Bilingualism 31 (1), 1–17.

Fennel, C.T., Byers-Heinlein, K., Werker, J.F., 2007. Using speech sounds to, guide word learning: the case of bilingual infants. Child Development 78 (5), 1510–1525. http://dx.doi.org/10.1111/j.1467-8624.2007.01080.x.

Flege, J.E., Eefting, W., 1987. Production and perception of English stops by native Spanish speakers. Journal of Phonetics 15, 67–83.

Flege, J.E., Shirru, C., MacKay, I., 2003. Interaction between the native and second language phonetic subsystems. Speech Communication 40, 467–491.

Gathercole, V.C.M., 2010. Interface of face to face? The profiles and contours of bilinguals and specific language impairment. Applied Psycholinguistics 31 (2), 282–293. http://dx.doi.org/10.1017/S0142716409990439.

Genesee, F., 1987. Learning through Two Languages: Studies of Immersion and Bilingual Education. Newbury, Rowley, MA.

Giedd, J.N., 2004. Structural magnetic resonance imaging of the adolescent brain. American New York Academy of Sciences 1021, 77–85. http://dx.doi.org/10.1196/annals.1308.009.

Grosjean, F., 1989. Neurolinguists, beware! The bilingual is not two monolinguals in one person. Brain and Language 36 (1), 3–15.

Grosjean, F., 2001. The bilingual's language modes. In: Nicol, J. (Ed.), One Mind, Two Languages. Beyond Language Processing. Blackwell, Malden, MA, pp. 1–22.

Grosjean, F., 2010. Bilingual: Life and Reality. Harvard University Press, Cambridge, MA.

Hakuta, K., 1986. Mirror of Language. The Debate on Bilingualism. Basic Books, New York.

Hakuta, K., Bialystok, E., 1994. In Other Words: The Science and Psychology of Second-Language Acquisition. Basic Books, New York.

Han, Z., 2004. Fossilization. Five central issues. International Journal of Applied, Linguistics 14 (2), 212–242.

Hufeisen, B., 1997. L3-Specifics. Magazine for Applied Linguistics 26, 83–87.

Jean, M., Geva, E., 2009. The development of vocabulary in English as a second language children and its role in predicting word recognition ability. Applied Psycholinguistics 30, 153–185. http://dx.doi.org/10.1017/S0142716408090073.

Jessner, U., 2006. Linguistic Awareness in Multilinguals. English as a Third Language. Edinburgh University Press, Edinburgh.

Kay-Raining Bird, E., Cleave, P., Trudau, N., Thorodottir, E., Sutton, A., Thorpe, A., 2005. The language abilities of children with Down Syndrome. American Journal of Speech Language Pathology 14, 187–199.

Kohnert, K., 2008. Second language acquisition: success factors in sequential bilingualism. The ASHA Leader 13 (2), 10–13.

Kohnert, K., 2013. Language Disorders in Bilingual Children and Adults, second ed. Plural, San Diego, CA.

Kohnert, K., Bates, E., 2002. Balancing bilinguals II: lexical comprehension and cognitive processing in children learning English and Spanish. Journal of Speech, Language, and Hearing Research 45, 347–359.

Kohnert, K., Bates, E., Hernandez, A., 1999. Balancing bilinguals: lexical-semantic production and cognitive processing in children learning Spanish and English. Journal of Speech, Language, and Hearing Research 42, 1400–1413.

Lugo-Neris, M., Wood-Jackson, C., Goldstein, H., 2010. Facilitating vocabulary acquisition of young English language learners. Language, Speech, and Hearing Services in Schools 41, 314–327. http://dx.doi.org/10.1044/0161-1461(2009/07-0082).

Montrul, S., 2006. On the bilingual competence of Spanish heritage speakers: syntax, lexical-semantics and processing. International Journal of Bilingualism 10 (1), 37–69.

Paradis, J., Genesee, F., 1996. Syntactic acquisition in bilingual children: autonomous or independent? Studies in Second Language Acquisition 18, 1–25.

Pearson, B.Z., Fernandez, S.C., 1994. Patterns of interaction in the lexical growth in two languages of bilingual infants and toddlers. Language Learning 44, 617–653.

Pearson, B.Z., Fernandez, S.C., Lewedeg, V., Oller, D.K., 1997. The relation of input factors to lexical learning by bilingual infants. Applied Psycholinguistics 18, 41–58.

Pearson, B., Fernandez, S., Oller, D.K., 1993. Lexical development in bilingual infants and toddlers: comparison to monolingual norms. Language Learning 43, 93–120.

Pearson, B., Fernandez, S., Oller, D.K., 1995. Cross-language synonyms in the lexicons of bilingual infants: one language or two? Journal of Child Language 22, 345–368.

Pearson, J., Pearson, B.Z., 2004. Bilingual lexical development: influences, contexts, and processes. In: Goldstein, B. (Ed.), Bilingual Language Development and Disorders in Spanish-English Speakers. Brookes, Baltimore, pp. 77–104.

Pham, G., Kohnert, K., 2013. A longitudinal study of lexical development in children learning Vietnamese and English. Child Development (published on-line first July 19, 2013), 1–16. http://dx.doi.org/10.1111/cdev.12137.

Poplack, S., 1980. Sometimes I'll start a sentence in Spanish y termino en español: toward a typology of code-switching. Linguistics 18, 581–618.

Sakai, K.L., 2005. Language acquisition and brain development. Science 310 (4), 815–819.

Seliger, H., Vago, R. (Eds.), 1991. First Language Attrition. Cambridge University Press, Cambridge, MA.

Selinker, L., 1972. Interlanguage. International Review of Applied Linguistics 10 (2), 209–231.

Selinker, L., 1991. Along the way: interlanguage systems in second language acquisition. In: Malavé, L.M., Duquette, G. (Eds.), Language, Culture and Cognition. Multilingual Matters, Bristol, PA, pp. 23–35.

Thomas, P., Collier, V., 1997. School Effectiveness for Language Minority Students. National Clearinghouse for Bilingual Education, Washington, DC.

Thomas, W.P., Collier, V.P., 2002. A National Study of School Effectiveness for Language Minority Student's Long Term Academic Achievement. Center for Research on Education, Diversity and Excellence and the Office of Educational Research Improvement (ERIC Document Reproduction Service No. ED 475048).

Uylings, H.B.M., 2006. Development of the human cortex and the concept of "critical" or "sensitive" periods. Language Learning 56 (1), 59–90.

Wapole, C., 2000. The bilingual child: one system or two. In: Clark, E. (Ed.), The Proceedings of the Thirtieth Annual Child Language Research Forum. The Center for the Study of Language and Information, Stanford, CT, pp. 187–194.

Windsor, J., Kohnert, K., Lobitz, K., Pham, G., 2010. Cross-language nonword repetition by bilingual and monolingual children. American Journal of Speech-Language Pathology 19, 298–310.

Wong-Fillmore, L.W., 1992. When Does $1 + 1 =< 2$? Paper Presented at Bilingualism/Bilingüismo: A Clinical Forum. Miami, FL.

Multilingualism

Ulrike Jessner-Schmid, University of Innsbruck, Innsbruck, Austria; and University of Pannonia, Veszprem, Hungary

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Edwards, volume 2, pp. 1167–1171, © 2001, Elsevier Ltd.

Abstract

Interest in the phenomenon of multilingualism has considerably grown over the last few decades. Since the discovery of cognitive advantages of bilingual children over their monolingual peers in the 1960s positive attitudes have been developed toward the phenomenon, mainly in Applied Linguistics. The field of multilingualism has been studied from both the social and the individual perspective. Current work on multilingualism has focused on the dynamics and complexity of multilingual development and use. The use of English as a *lingua franca* or third language as part of current individual multilingualism will also be addressed.

Introduction

Scholarly interest in multilingualism can be described as a fairly new development in comparison to the phenomenon which has existed for a much longer time in human history. Until rather recently, bi- and multilingual speakers were often treated as schizophrenic and cognitively handicapped speakers with deficits in each of their languages. Ever since Peal and Lambert in the 1960s discovered the cognitive advantages of bilingual children over their monolingual peers the attitudes toward bi- and multilingualism have changed and lately Bialystok's seminal work on cognitive advantages in bilinguals over the life span has definitely shaped our knowledge on emergent cognitive properties in multilingual individuals. Due to the trained flexibility in the bilingual brain she and her collaborators attested elderly bilinguals to be protected against the onset of dementia (for an overview see Bialystok, 2009).

The discussion of the topic of competence in more than one language is complex because multilingualism can be encountered as a societal, discursive, and individual phenomenon. The overview provided in this article can only alert the reader to relevant fields of research. Matters of definition and terminology are dealt with first. Second, problems of assessment of multilingual proficiency are addressed. Third, an overview of linguistic, social, and psycholinguistic aspects of multilingualism is given. Fourth, multilingualism and multilingual development as complex and diverse phenomena are discussed. Finally, a brief outlook into challenges in multilingualism (research) is provided.

Definitions and Terminology

The term multilingualism has been used with increased frequency over the last few years. Yet, a critical analysis of these publications reveals that this term is used to cover a range of meanings, some of which appear to be contradictory. Many still use the term multilingualism to refer to any system with more than one language. As bilingualism is the most commonly studied form of multilingualism, the two terms are also used synonymously. In contrast, an increasing number of researchers argue that bilingualism should be seen as a subtype of multilingualism. Within the same group, third language acquisition (TLA) is not discussed as extended second language acquisition (SLA), but rather as a different language learning process in various respects (cf Jessner, 2008a).

As pointed out by Kemp (2009), the variety of terms and their origins derive from (1) the complex nature of use of various languages in a multilingual context and (2) the researchers' backgrounds, ideologies, and purposes. So whereas from a psycholinguistic perspective bi- and multilingualism, SLA, and TLA are the most common terms in use, scholars with other theoretical backgrounds have recently introduced different terms to describe the use of multiple languages such as polylingualism (Jorgensen et al., 2011), metrolingualism (Otsuji and Pennycook, 2010), or heteroglossia (Bakhtin), to name but a few. In European Union language policy, mainly in educational-oriented research, the term plurilingualism has been established to denote individual multilingualism in contrast to multilingualism on a social level. From an international perspective, multilingualism is the term most prominently used.

With the increase in interest in the field of multilingualism, several problematic issues, in particular with regard to terminology, have arisen. For instance, whereas in traditional SLA research the terms L1 and L2 are clearly defined, research on life span development in multilingual individuals shows that the chronological order in which languages were learnt is not necessarily the most useful factor in defining the L1, in the sense of 'dominant language,' of a multilingual person. An L1 might become an L2 or L3 at a later phase in a multilingual person's life due to a change of linguistic needs. The use of a 'primary/secondary' language system in order to deal with changes in language dominance in a bilingual system has also been suggested (Herdina and Jessner, 2002). Primary learner multilingualism refers to a third language learnt after primary bilingualism, whereas secondary learner multilingualism refers to the acquisition of two languages simultaneously after the primary language system has been acquired (see also Hammarberg, 2010 on the chronological approach to the terms L1, L2, L3, etc.). The strong role that emotion can play in multilingualism might also lead to further complications in terms of individual classification (Pavlenko, 2005). L1 might also denote the language of the heart, e.g., the language somebody shares with her/his partner regardless of the level of proficiency in that language and/or the frequency

of use. Not to mention the problematic defining dimensions of high degrees of proficiency, functionality, identification, and automaticity in the case of the native language(s), as pointed out by Skuttnab-Kangas (1984).

Assessment of Multilingual Proficiency

Guiseppe Mezzofanti, who lived in the nineteenth century, was not only reported to have been fluent in 60 languages, but also to have had translating ability in twice this number. Another linguist living in the same century was Emil Krebs, a German Sinologue who became famous for being literate in 68 languages and capable of dealing with another 111 languages. Ziad Fazah, a born Lebanese, who has acquired most of his languages autonomously, is fluent in 58 languages such as Chinese, Thai, Greek, Indonesian, Hindi, and Farsi. A number of bilingual writers such as the Polish immigrant Joseph Conrad have succeeded in the development of excellent writing abilities in a language they acquired later in life. Elias Canetti whose L1 was Sephardic (Spanish Hebrew) won the Nobel Prize in literature for a book written in German which was actually his fourth language. One of the most well-known actresses Audrey Hepburn was fluent in Dutch, English, French, Italian, and Spanish. She came from a bilingual family and traveled between and lived in a number of countries. Many scholars working in philology claim full competence in several – in many cases typologically related – languages. But how are all these abilities related? What if a person can say 'good morning' in a number of languages? Is s/he multilingual? Is a German-speaking student studying English at university level a bilingual? Can a nonliterate person who is fluent in three languages be called a multilingual? Can a person achieve full mastery of any language if he/she learnt the languages as an adult?

Typologies of bilingualism focusing on competence have taken either a minimalist or maximalist approach. Bloomfield (1933) defined bilingualism as nativelike control of two or more languages, whereas Macnamara (1967) characterizes a bilingual as someone who possesses some proficiency in any one of the four language skills in a language other than his/her mother tongue. Years later Li Wei (2000) came up with 36 types of bilingualism including rarely used categories such as diagonal bilingual for a person "who is bilingual in a nonstandard language or a dialect and an unrelated standard language." Whereas dichotomies are widely used both in bi- and multilingualism as they present clear ways of classification, Valdes (2005) argues for a continuum of bilingual proficiency, ranging from full monolingual proficiency in language A to monolingual proficiency in language B by taking different levels of proficiency into consideration. The conceptualization of a continuum for proficiency in three and more languages still awaits realization.

What has become clear today is that a discussion of balanced, nativelike proficiency in all of a multilingual's languages has been banned. Instead holistic concepts of bilingualism stemming from Grosjean and Cook who portray the bilingual speaker/user as competent but different from a monolingual have turned out to be influential in the development of the field of bi- and multilingualism. According to a holistic or realistic perspective the bilingual person is a competent but special speaker/hearer (Grosjean, 2010) who in contrast to a monolingual person develops different skills, as taken up later by Cook (e.g., 2003) in his construct of multicompetence. They both have argued for the nonapplicability of monolingual norms in multilinguals and consequently in research on bi- and multilingualism.

Codeswitching, Transfer, and Translanguaging

In a multilingual repertoire language systems do not coexist without influencing each other. Weinreich (1953) initiated research on languages in contact. Since then a number of crosslinguistic phenomena have been identified, depending on the fields of research. For a long time SLA research concentrated on interference and transfer, whereas bilingualism research focused on codeswitching and borrowing phenomena. In 1986 Kellerman and Sharwood Smith introduced crosslinguistic influence to cover phenomena studied in both fields.

Research on TLA necessarily has to go beyond the study of two languages in contact. Over the last 15 years research on TLA or multilingualism has been increasingly intensified with the main goal of describing multilingual phenomena in order to investigate differences and similarities between second and third language acquisition (see, e.g., Cenoz et al., 2001, 2003; De Angelis, 2007; Aronin and Hufeisen, 2009). In a multilingual system transfer phenomena not only take place from the L1 to the L2 and vice versa but have also been evidenced from the L1 to the L3 and from the L2 to the L3 and vice versa. Combined crosslinguistic influence occurs "when two or more languages interact with one another and concur in influencing the target language, or when one language influences another, and the already influenced language in turn influences another language in the process of being acquired" (De Angelis, 2007, p. 21). This expansion of transfer possibilities already shows that multilingual acquisition is a far more complex process than SLA, where the role of the L1 in the development of the L2 has been researched extensively. Some of the crosslinguistic interactions can be permanent, such as accent or fossilized use of a syntactic structure, others can be accidental or momentary (see also Grosjean, 2013).

One of the most crucial questions in psycholinguistic research on TLA is related to the status of the L2 in third language learning. In a number of studies it turned out that in L3 production the speakers did not rely on their L1 as expected, but on their L2, and in several studies of learning an L3 of Indo-European origin it could be shown that L3 learners whose L1 is typologically unrelated to the L2 and/or L3 tend to transfer knowledge from their L2 or in the case of bilinguals from the related L1 (for an overview and list of TLA studies see Jessner, 2008b). The activation of prior language knowledge is influenced by factors such as psychotypology (perceived linguistic distance between languages), recency of use, the level of proficiency in the target language, the foreign language effect, the tendency in language learners to activate an earlier second language in L3 performance, and the learner's perception of correctness of

a target word (for an overview see De Angelis and Dewaele, 2009; De Angelis et al., 2015).

Nowadays one focus in multilingualism research, mainly in educational contexts, has shifted to the phenomenon of translanguaging characterized by Li Wei (2011) as both going between different linguistic structures and systems and going beyond them and as including the full range of linguistic performances of multilingual language users. His focus is on opting between language resources, on the coordination of such resources, and on constant decision making in the multilingual individual concerning the achievement of specific communicative effects in social interaction. Garcia introduced the term in 2007 as bilingual pedagogy in order to move away from monoglossic ideologies of bilingualism, also described as flexible bilingualism by Blackledge and Creese (2010). But it should also be noted that in fact in a multilingual context, be it in an institutional or natural setting, languages do not always interact, at least not on a performance level (see also below).

Connectivity between Languages

Research on crosslinguistic phenomena as part of the multilingual mental lexicon is interested in the connectivity between the language systems involved. Although there is plenty of evidence that the languages in a multilingual system dynamically interact, there is also evidence that languages are kept apart in a communicative act. Singleton (2011) argues that "the reality of a degree of crosslinguistic conceptual permeability and transfer does not undermine the other reality of this […] essential differentiation."

Cook (2003) introduces *the language integration continuum* along which possible relationships between the languages are displayed. The continuum does not imply a direction of movement and it does not have to apply to the whole language system or affect all individuals in the same way. This also implies that the relationship between the L1 and the interlanguage within one mind is different from that between the interlanguage in one mind and the L2, when the L2 has the status of an L1 in another mind (Cook, 2006). In that sense, Cook's ideas are rather close to Grosjean's *language mode hypothesis* where he explains the factor of activation of the background languages with his concept of *speech modes*. In brief, Grosjean (2001) asserts that "at any given point in time and based on numerous psychosocial and linguistic factors, the bilingual has to decide, usually quite unconsciously, which language to use and how much of the other language is needed – from not at (*sic*) all to a lot." Thus, a speaker depending on the language mode continuum chooses a base language which is highly activated and puts the other slightly less activated languages aside. According to Grosjean (2001) a trilingual person can function in a monolingual, bilingual, or trilingual mode with various levels of activation, in relation to her or his position on the language mode continuum. Activation of the various languages is strongly influenced, amongst other factors, by the speaker(s)' usual language mixing habits, language proficiency, socioeconomic status, the presence of mono- and bilinguals, the degree of formality, etc.

Cognitive Aspects of Multilingualism

For a long time brain capacity was understood to be finite, with two implications: first, the more capacity you have the more languages you can learn and, second, if you learn something new you will be confronted with the loss of something old. Such a scientific perspective also led to the assumption that growing up with more than one language has to lead to detrimental effects on the cognitive level. Accordingly and exemplary for other academics of the time, Laurie (1890) noted the following: "If it were possible for a child to live in two languages at once equally well, so much the worse. His intellectual and spiritual growth would not thereby be doubled, but halved. Unity of mind and character would have great difficulty in asserting itself in such circumstances" (quoted after Baker, 2011).

Only when Peal and Lambert (1962) reported in their study of bilingual children in the Montreal area that the bilinguals outperformed their monolingual peers in a number of tasks, a more positive attitude toward bilingualism was created. Meanwhile it has become widely known that bi- and multilingual individuals can have tremendous advantages, not only in terms of dual language competence but also in terms of cognitive and social behavior. Hamers and Blanc (1989) argue that cognitive advantages seem to occur mainly at the level of higher creativity and reorganization of information. But with the increase in knowledge a more realistic attitude toward the phenomenon has been developed.

The work by Bialystok and her collaborators (e.g., 2009) has been very influential in the identification of differences between mono- and bilingual subjects across the life span. Many of her recent studies with a focus on executive control (processes which include attention, selection, inhibition, monitoring, and flexibility) in bilingual children and adults have attested a clear benefit to the bilingual subjects. The results of Barac and Bialystok (2012) offer strong support for the claim that bilingualism acts independently of variables such as language similarity, cultural background, and language of schooling.

Intelligence and Age

The expansion of the language repertoire, either actual or potential, is often linked to intelligence and/or age. Young learners are still considered to be advantaged while older learners are considered to have a hard time learning a new language. Such belief is obviously very closely related to the concept of a critical or rather sensitive period which attests better and faster learning in the young because of the plasticity and flexibility of the young brain up to puberty. But thanks to recent work on the interdependence between age and language learning we have learnt that (1) age cannot be studied in isolation from other variables in language learning and (2) there are no clear advantages for a certain age group (Munoz and Singleton, 2011). In fact recent research on neurological aspects of multilingualism has made it clear that although most brain growth occurs right after birth the brain develops into late adolescence and/or early and later adulthood, which would favor a sensitive and noncritical period for cognition and learning language (Uylings, 2006; see also Li, 2013).

Social Contexts of Multilingual Development and Use

In many contexts in the world more than one language is used on a daily basis. It can be assumed that multilingual individuals live in multilingual countries. But we have to distinguish between official and *de facto* multilingualism (Clyne, 1997). Although Switzerland is officially a multilingual nation most people typically grow up with one language, if in the German-speaking cantons we ignore the fact that everyone grows up with Standard German and some variety of Swiss German, an Alemannic dialect. Other examples of such countries are Luxembourg, where German and French are the official languages but Letzebuergesch is the language of daily life, or Singapore, where Standard English, Mandarin, Standard Malay, and Tamil are used for official purposes and complemented by other languages such as Singapore English, Hokkien, Cantonese, and varieties of Malay in everyday communication (Clyne, 1997). Similarly, in Malaysia, for instance, it is common that language skills are developed in various languages for the use in certain domains but that the language proficiency in those languages is rather limited (Kärchner-Ober, 2009). This phenomenon is rather common in multilingual contexts and has led to the redefinition of multilingualism as a phenomenon which views language choice and proficiency as necessarily dependent on use in specific domains of life, thereby moving away from the notion of multiple monolingualism. The fact that "[d]ifferent aspects of life often require different languages" is called 'complementarity principle' by Grosjean (2010, p. 29ff.).

The Dominant Language Constellation (DCD) is the group of the individual's most important languages that function as an entire unit and enable him/her to act in a multilingual environment in such a way as to meet all of his/her needs (Aronin and ÓLaoire, 2004). This constellation operates also on a social level, that is across communities. Examples are speakers of French in France, Canada, and Guiana, or of Russian in Israel, Austria, and Hungary. Typically a DCD includes languages of different status, depending on the context an international language, a regional *lingua franca*, one or more state languages and one or more minority (local or immigrant) languages, for instance a core multilingual repertoire in Austria includes English, German, Croatian, in Germany it includes French, German, Polish, in Hungary it is composed of English/Hungarian/German (see also Aronin and Singleton, 2012, pp. 59–75).

Functions of Languages

All the languages in a multilingual repertoire have certain functions to fulfill. Classifications of the functionality of certain languages are made according to language experience and use. For instance, early multilingualism can have different origins. Hoffmann (2001) offers the following typology by also pointing out that some language users may be placed in more than one category or that categories can overlap, or that due to mobility and other changes in a lifetime users can move from one category into another.

- Trilingual children who are brought up with two home languages which are different from the one spoken by the wider community.
- Children who grow up in a bilingual community and whose home language (either that of one or both parents) is different from the community languages.
- Third language learners, i.e., bilinguals who acquire a third language in the school context.
- Bilinguals who have become trilinguals through immigration.
- Members of trilingual communities.

Other features which could be useful for building typologies are age, acquisition process, acquisition context, and language proficiency, among others (Hoffmann, 2001, p. 19).

English as a Contact Language in Multilingual Contexts

Globalization and the resulting increased mobility has led to a worldwide spread of English. The spread of English has also been linked to its use as a *lingua franca* and as a vehicle for the spread of a culture influenced by the USA. As noted by David Graddol already in 1994 the number of nonnative users of English will correspond to half of the world's population by 2015. He is also convinced that many are scared that English, which has developed the status of a killer language, will grow, but as part of current multilingualism. In fact, nowadays in almost all cases a multilingual repertoire includes English, or in other words, multilingualism with English presents the international norm.

- English as a third language

In a number of international contexts English is not learnt as a second but as a third language. European examples are Belgium where it is learnt besides French and German, Finland where it is learnt besides Finnish and Swedish, or the Austrian province Carinthia where it is learnt besides German and Slovene. This means that in all minority language contexts English is learnt as a third language. But also immigrants whose repertoire includes two other languages very often learn English as a third language at school. An example of this phenomenon is migrant children of Turkish, Polish, Russian, or Hungarian origin in Germany or Sweden. Other contexts such as countries where English is used as a second (such as the French-speaking part of Canada) or first language in which bilingual migrants learn it as a third or even fourth language, which is also the case in the Ladin communities in South Tyrol where pupils learn English from an early age on besides Ladin, German, and Italian. In South and Central America it is the third language for many children who are heritage language speakers of Quechua, Guarani, and Mohawk.

- English as a *lingua franca* (ELF)

English is increasingly used as a *lingua franca* in politics, trade, tourism, media, and science and therefore forms part of most people's lives. In international communication situations English is chosen when the language backgrounds of the communication partners are different but they share English as a common system. This has left its traces on the development of the English language system and it has been argued that ELF should be assigned the status of a variety of its own. Although the use of ELF is as diverse as its users' competences, it has been

tried to identify structures of ELF which are common to all speakers (see, e.g., work by Seidlhofer, 2011).

Dynamics and Complexity of Multilingual Development and Use

Multilingual development is both a dynamic and complex process and nowadays scholars are realizing more and more that in the study of multilingualism, as can be seen from the above discussion, the interaction of a variety of factors needs to be taken into consideration. Individual factors in language learning such as motivation, attitudes, cognitive factors such as personality traits (See e.g., Dewaele, 2007), and also physical traits of multilingual learners (e.g., hearing capabilities) contribute to the complexity of multilingualism (research). Furthermore during the life span of a multilingual person a variety of the factors involved, such as motivation to learn one language, can be subject to change.

In order to be able to judge multilingual development and/or multilingual conversation, all variables which contribute to multilingual development and use would need to be discussed. This means that we have to be aware of the possible acquisition routes whose number increases with the number of languages learnt (Cenoz, 2000). Moreover, forms of learning are considered to play a crucial role in multilingual development. In contrast to SLA, TLA learning routes are much more diverse because very often forms of learning are of a mixed nature, that is multilingual learners acquire their languages in both natural and instructed contexts.

Applying Dynamic Systems Theory to Multilingualism

The investigation of the complexity of multilingualism seems to lend itself to being approached from a dynamic systems theory/complexity theory (DST/CT) perspective, a recently much discussed approach in the field of (second) language acquisition research (see, e.g., Larsen-Freeman and Cameron, 2009; De Bot, 2008).

In the Dynamic Model of Multilingualism (DMM) (Herdina and Jessner, 2002) DST/CT is applied to multilingual acquisition and use. Based on DST/CT principles the development of a multilingual system is characterized by its nonlinearity, reversibility, stability, interdependence, complexity, and change of quality. Multilingual systems are adaptive and dynamic, which means that they are able to change depending on the perceived communicative needs of multilingual individuals. This implies that language choices are psychologically and socially determined and are the driving force of language learning and use (see also language mode above) and, furthermore, that the language development of multilingual individuals includes both acquisition and attrition processes.

Systems stability is related to language maintenance, i.e., the stability of a psycholinguistic system is dependent on the requirements of language maintenance, which implies that the system will erode if not enough energy and time is invested into maintaining a system. In DMM language systems are seen as interdependent and not as autonomous systems, as which they are still perceived in mainstream SLA research.

Multilingual proficiency is defined as the dynamic interaction between the various psycholinguistic systems (LS_1, LS_2, LS_3, LS_n) in which the individual languages (L1, L2, L3, L*n*) are embedded, crosslinguistic interaction, and what is called the M(ultilingualism)-factor. The latter refers to all the effects in multilingual systems which distinguish a multilingual from a monolingual system, regarding all those qualities which develop in a multilingual speaker/learner due to the increase in language contact(s). As mentioned above, language contacts depend on the perceived communicative needs of the individual. In other words, the psycholinguistic systems of the multilingual individual, which are in constant change, interact with each other in a nonadditive but cumulative way.

Crosslinguistic interaction in multilingual individuals, which is described as an umbrella term including not only transfer and interference but also codeswitching and borrowing, is also meant to cover another set of phenomena. These are nonpredictable cognitive effects of the development of the multilingual which determine the development of the systems themselves. Such a view is also related, but not identical, with Cummins' Common Underlying Proficiency (e.g., 1991) and Kecskes and Papp's Common Underlying Conceptual Base (2000) (see also Cook above). According to DMM seemingly identical phenomena of transfer can lead to divergent results in different multilingual systems, even if they are transitionally commanded by the same speaker.

One of the key variables of multilingual development and use which was identified as an emergent property of the multilingual system in the DMM is multilingual awareness.

The M-Factor and Multilingual Awareness

The M(ultilingualism)-factor integrates both language-specific and nonlanguage-specific skills used in language learning, language management, and maintenance. Language management skills refer to the integration and separation of language resources and the act of balancing communicative requirements with language resources (see also translanguaging above).

The key factor of the M-effect – as it might also be referred to since it is difficult to decide whether it constitutes a precondition or a result of multilingualism – is metalinguistic awareness or multilingual awareness. It can be defined as the multilingual's ability to focus attention on languages as objects in themselves and consequently to manipulate languages. The domains which can be affected by metalinguistic ability are linguistic development in general, the development of cognitive, metacognitive, and information-processing abilities, and the development of literacy skills.

Prior language knowledge and awareness of this knowledge influence further language learning, in particular learning a second foreign language (Hufeisen and Jessner, 2009). In contrast to monolingual learners, bilingual children show cognitive advantages in further language learning (Cenoz, 2003). Evidence of this was found in investigations into artificial language learning (Kemp, 2007) and a number of studies mostly in the European minority language context such as Spain (Basque Country, Catalonia) and Finland (for an

overview see Jessner, 2008a). In DMM it is argued that in the process of learning an L3 the learner develops a metasystem which is based on a bilingual norm, unlike SLA, where such a metasystem can only be based on the acquisition of the L1 and therefore different in quality. In DST/CT terms this means that the multilingual system is considered sensitive to initial conditions which makes its development difficult to predict (see also Aronin and Jessner, 2015).

Challenges for Future Research

Multilingualism is a challenging field, both because of the complexity stemming from diverse interdisciplinary interests and because of the number of variables involved in multilingualism. But although by now a more realistic view of multilingualism appears to be established as a norm, the monolingual yardstick is still prevalent in most methodologies used in bi- and multilingualism studies.

As suggested by Aronin and Hufeisen (2009), it is absolutely necessary to cross the second threshold, which implies that following the crucial contributions by Grosjean and Cook on the bilingual view of bilingualism as well as Herdina and Jessner (2002), we now have come to see that multilingualism can only be adequately researched by applying a multilingual norm to multilingual research. Linking research on the social level to research on the individual level presents a big challenge for future studies of multilingualism. Also combining the emic, the multilingual's perspective, with the etic, the researcher's perspective, of multilingualism appears to be the most promising step forward in the development of the field (Todeva and Cenoz, 2009).

Aronin and Singleton (2012) put forward a comprehensive view of multilingualism by pointing to the development of (1) a thesaurus of multilingualism, (2) constructs, (3) models specific to multilingualism, (4) working metaphors, and (5) the philosophical level of conceptualization. As depicted in Jessner and Kramsch (2015) it is also necessary to focus more on the so-called dark side of multilingualism and to face the multilingual challenges ranging from individual crises of identity, suspicions of scientific disloyalty to pervasive anxiety that is related to the transition from a monolingual world of identifiable knowledge, predictable monolingual grammars, and clearly recognizable authorities. This new trend in Applied Linguistics, which follows the very positive attitude that has been developed over the years toward multilingualism, also presents a call for a realistic portrait of the phenomenon in question, be it on the familial, the political, the sociocultural, or the scientific level.

See also: Multilingual Language Development; Second Language Acquisition.

Bibliography

Aronin, L., ÓLaoire, M., 2004. Exploring multilingualism in cultural contexts: towards a notion of multilinguality. In: Hoffmann, C., Ytsma, J. (Eds.), Trilingualism in Family, School and Sommunity. Multilingual Matters, Clevedon, pp. 11–29.

Aronin, L., Hufeisen, B. (Eds.), 2009. The Exploration of Multilingualism. John Benjamins, Amsterdam.

Aronin, L., Jessner, U., 2015. In: Jessner, U., Kramsch, C. (Eds.), The Multilingual Challenge: Interdisciplinary Perspectives. Mouton de Gruyter, Berlin; New York.

Aronin, L., Singleton, D., 2012. Multilingualism. John Benjamins, Amsterdam.

Baker, C., 2011. Foundations in Bilingual Education and Bilingualism, sixth ed. Multilingual Matters, Clevedon.

Barac, R., Bialystok, E., 2012. Bilingual effects on cognitive and linguistic development: role of language, cultural background, and education. Child Development 83, 413–422.

Bialystok, E., 2009. Bilingualism: the good, the bad, and the indifferent. Bilingualism: Language and Cognition 12, 3–11.

Blackledge, A., Creese, A., 2010. Multilingualism: A Critical Perspective. Continuum, London.

Bloomfield, L., 1933. Language. Henry Holt and Co., New York.

Cenoz, J., 2000. Research on multilingual acquisition. In: Cenoz, J., Jessner, U. (Eds.), English in Europe: The Acquisition of a Third Language. Multilingual Matters, Clevedon, pp. 39–53.

Cenoz, J., 2003. The additive effect of bilingualism on third language acquisition: a review. International Journal of Bilingualism 7, 71–87.

Cenoz, J., Hufeisen, B., Jessner, U. (Eds.), 2001. Crosslinguistic Influence in Third Language Acquisition: Psycholinguistic Perspectives. Multilingual Matters, Clevedon.

Cenoz, J., Hufeisen, B., Jessner, U. (Eds.), 2003. The Multilingual Lexicon. Kluwer, Dordrecht.

Clyne, M., 1997. Multiligualism. In: Coulmas, F. (Ed.), The Handbook of Sociolinguistics. Blackwell, Oxfrod, pp. 301–314.

Cook, V. (Ed.), 2003. Effects of the Second Language on the First. Multilingual Matters, Clevedon.

Cummins, J., 1991. Interdependence of first- and second language proficiency. In: Bialystok, E. (Ed.), Language Processing in Bilingual Children. Cambridge University Press, Cambridge, pp. 70–89.

De Angelis, G., 2007. Third or Additional Language Learning. Multilingual Matters, Clevedon.

De Angelis, G., Dewaele, J.-M., 2009. The development of psycholinguistic research on crosslinguistic influence. In: Aronin, L., Hufeisen, B. (Eds.), The Exploration of Multilingualism. John Benjamins, Amsterdam, pp. 63–78.

De Angelis, G., Jessner, U., Kresic, M. (Eds.), 2015. Crosslinguistic Interaction in Multilingual Learners. Bloomsbury, London.

De Bot, K., 2008. Introduction: second language development as a dynamic process. The Modern Language Journal 92, 166–178.

Dewaele, J.-M., 2007. The effect of multilingualism, sociobiographical, and situational factors on communicative anxiety and foreign language anxiety of mature language learners. International Journal of Bilingualism 11 (4), 391–409.

Garcia, O., 2009. Bilingual Education I the 21st Century. A Global Perspective. Wiley-Blackwell, New York.

Graddol, D., 1997. The Future of English. British Council.

Grosjean, F., 2001. The bilingual's language modes. In: Nicol, J. (Ed.), One Mind, Two Languages: Bilingual Language Processing. Blackwell, Oxford, pp. 1–25.

Grosjean, F., 2010. Bilingual. Harvard University Press, Cambridge, MA.

Grosjean, F., 2013. Bilingualism: a short introduction. In: Grosjean, F., Li, P. (Eds.), The Psycholinguistics of Bilingualism. Wiley-Blackwell, London, pp. 5–25.

Hamers, J.F., Blanc, M., 1989. Bilinguality and Bilingualism. Cambridge UP, Cambridge.

Hammarberg, B., 2010. The languages of the multilingual: some conceptual and terminological issues. International Review of Applied Linguistics 48, 91–104.

Herdina, P., Jessner, U., 2002. A Dynamic Model of Multilingualism: Changing the Psycholinguistic Perspective. Multilingual Matters, Clevedon.

Hoffmann, C., 2001. The status of trilingualism in bilingualism studies. In: Cenoz, J., Hufeisen, B., Jessner, U. (Eds.), Beyond Second Language Acquisition: Studies in Tri- and Multilingualism. Stauffenburg, Tübingen, pp. 13–25.

Jessner, U., 2008a. A DST model and the role of metalinguistic awareness. The Modern Language Journal 92, 270–283.

Jessner, U., 2008b. Teaching third languages: findings, trends and challenges. Language Teaching 41 (1), 15–56.

Jessner, U., Kramsch, C. (Eds.), 2015. The Multilingual Challenge: Interdisciplinary Perspectives. Mouton de Gruyter, Berlin, New York.

Jorgensen, N., Karrebæk, M.S., Madsen, L.M., Møller, S., 2011. Polylanguaging in superdiversity. Diversities 13 (2).

Kecskes, I., Papp, T., 2000. Metaphorical competence in trilingual language production. In: Cenoz, J., Jessner, U. (Eds.), English in Europe: The Acquisition of a Third Language. Multilingual Matters, Clevedon, pp. 99–120.

Kellerman, E., Sharwood Smith, M. (Eds.), 1986. Crosslinguistic Influence in Second Language Acquisition. Pergamon Press, Oxford.
Kemp, C., 2007. Strategic processing in grammar learning: do multilinguals use more strategies? International Journal of Multilingualism 4, 241–261.
Kemp, C., 2009. Defining multilingualism. In: Aronin, L., Hufeisen, B. (Eds.), The Exploration of Multilingualism. John Benjamins, Amsterdam, pp. 11–26.
Larsen-Freeman, D., Cameron, L., 2009. Complex Systems and Applied Linguistics. Oxford UP, Oxford.
Li, P., 2013. Succesive language acquisition. In: Grosjean, F., Li, P. (Eds.), The Psycholinguistics of Bilingualism. Wiley-Blackwell, London, pp. 145–155.
Macnamara, J., 1967. The bilingual's linguistic performance: a psychological overview. Journal of Social Issues 23, 58–77.
Munoz, C., Singleton, D., 2011. A critical review of age-related research on L2 ultimate attainment. Language Teaching 41 (1), 1–35.
Otsuji, E., Pennycook, A., 2010. Metrolingualism: fixity, fluidity and language in flux. International Journal of Multilingualism 7 (3), 240–254.
Pavlenko, A., 2005. Emotions and Bilingualism. Cambridge University Press, Cambridge.
Peal, E., Lambert, W., 1962. The relation of bilingualism to intelligence. Psychological Monographs 76 (1), 1–23.
Seidlhofer, B., 2011. Understanding English as a Lingua Franca: A Complete Introduction to the Theoretical Nature and Practical Implications of English Used as a Lingua Franca. Oxford University Press, Oxford.
Skutnabb-Kangas, T., 1984. Bilingualism or Not. Multilingual Matters, Clevedon.
Todeva, E., Cenoz, J. (Eds.), 2009. The Multiple Realities of Multilingualism. Mouton de Gruyter, Berlin.
Uylings, H.B., 2006. Development of the human cortex and the concept of "critical" or "sensitive" periods. Language Learning 56 (1), 59–90.
Valdes, G., 2005. Bilingualism, heritage language learners, and SLA research: opportunities lost or seized? Modern Language Journal 89 (3), 410–426.
Wei, L., 2000. The Bilingualism Reader. Routledge, London.
Wei, L. (Ed.), 2011. The Routledge Applied Linguistics Reader. Routledge, London.
Weinreich, U., 1953. Languages in Contact. Findings and Problems. Humanities Press, New York.

Relevant Websites

http://www.aila.info – Association Internationale de la Linguistique Appliqueé (AILA).
http://www.iamultilingualism.org/ – International Association of Multilingualism.
http://www.tandfonline.com/loi/rmjm20#.U-zLQvl_uyl – International Journal of Multilingualism.

Multimodal Therapy in Clinical Psychology

Arnold A Lazarus[†], Rutgers University, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Multimodal therapy (MMT) arose out of the realization that there were significant treatment omissions in behavior therapy and in cognitive behavior therapy (CBT). CBT rests on trimodal base – affect, behavior, and cognition. In addition, MMT also thoroughly assesses imagery, sensations, interpersonal relationships, and biological factors resulting in a seven-point diagnostic matrix. This broad-spectrum approach addresses response deficits and excesses, and rests on a social and cognitive learning theory base. While eschewing theoretical integration, multimodal therapists may borrow effective strategies from diverse theories without buying into the theories that spawned them – technical eclecticism. Follow-ups indicate that by working multimodally, clinicians are apt to achieve durable results.

When follow-ups on a variety of clients who had received standard behavior therapy revealed that many were apt to suffer setbacks within 6 months to a year, investigations into the reasons behind this lack of durability revealed that treatment omissions (e.g., leaving several response deficits untreated) lay behind most of the relapses (Kwee and Lazarus, 1986). The behavioral treatment repertoire during the 1950s and 1960s glossed over or completely disregarded many cognitive variables, sensory responses, mental images, and interpersonal factors. Thus, agoraphobic clients received relaxation training, imaginal, and *in vivo* desensitization, and between-session homework assignments to consolidate and augment whatever gains ensued. Virtually no attention was paid to cognitive mediation (e.g., negative automatic thoughts, dichotomous reasoning, unrealistic expectations, and other self-defeating beliefs). Similarly, interpersonal processes (other than unassertive behaviors) were given short shrift (e.g., marital discord, familial tensions, the impact of authority figures, or sibling rivalry). Unaddressed and untreated, these issues, when triggered, retained the power to undermine whatever treatment gains had accrued.

Enter Cognitive Restructuring and More

By the mid-1970s, cognitive-behavior therapy had replaced behavior therapy as the accepted designation, and the field became trimodal – emphasizing affect, behavior, and cognition. This remains true to this day. Nevertheless, Lazarus (1971, 2005, 2006) has contended that equal weight should be given to imagery, sensations, interpersonal relationships, and biological considerations thus operating from a seven-point multimodal perspective (behavior, affect, sensation, imagery, cognition, interpersonal relationships, and drugs/biological factors). Using the easy-to-remember mnemonic acronym BASIC I.D. (taken from the initial letters of the foregoing modalities), the multimodal framework emphasizes the discrete and interactive impact of these seven modalities. The 'D' has contended that equal weight should be given to imagery, sensations, interpersonal relationships, and biological considerations thus operating from a seven-point multimodal perspective (behavior, affect, sensation, imagery, cognition, interpersonal relationships, and drugs/biological factors). Using the easy-to-remember mnemonic acronym BASIC I.D. (taken from the initial letters of the foregoing modalities), the multimodal framework emphasizes the discrete and interactive impact of these seven modalities. The modality underscores that prescription drugs are the most commonly used biological intervention, but nonetheless one must not lose sight of the entire medical panoply. Whenever feasible, multimodal therapists will employ empirically supported techniques, or refer out to an expert, when necessary, to implement a specific procedure. The aim, however, is to think in BASIC I.D. terms to ensure thorough and comprehensive assessment and treatment coverage.

Who Might Benefit from This Approach?

Most clients are likely to show positive gains when receiving a broad-spectrum treatment in which salient issues are addressed (Goldfried and Davison, 1994). Follow-up studies that have been conducted since 1973 have consistently suggested that durable outcomes are in direct proportion to the number of modalities deliberately traversed (see Lazarus, 1997, 2005, 2008). Although there is obviously a point of diminishing returns, it is a multimodal maxim that the more someone learns in therapy, the less likely he or she is to relapse. To reiterate, it appears that lacunae or gaps in people's coping responses were responsible for many relapses (Lazarus, 1981).

Contraindications of the Treatment

People in crisis require the therapist to focus immediately on the precipitating circumstances and events. There is often simply no time to dwell on the vicissitudes and nuances of the BASIC I.D. These cases require and may demand instantaneous relief from their untoward anxiety and depression. Less acute issues and much less dramatic problem areas may call for unimodal or bimodal interventions. For example, clients in a smoking cessation group, in terms of treatment adherence, are likely to respond better to structured methods of habit control than to a broad band of assessments and interventions

[†] Deceased.

(see Rudolph, 1985). And when dealing with frankly psychotic (e.g., schizophrenic) individuals, or with those who are in the manic state of a bipolar disorder, the first inroad should probably consist of supportive treatment and biopsychiatric interventions (Mueser and Gingerich 2006). When appropriate medications have taken effect, a multimodal treatment program may then be instituted to remedy excesses and deficits that may precipitate or aggravate the patient's initial or major complaints.

Theory and Mechanisms

The BASIC I.D. or multimodal framework rests on a broad social and cognitive learning theory (e.g., Bandura, 1977, 1986) because its tenets are open to verification or disproof. Instead of postulating putative unconscious forces, social learning theory rests on testable developmental factors (e.g., modeling, observational learning, the acquisition of expectancies, operant and respondent conditioning, and various self-regulatory mechanisms). It must be emphasized that while drawing on effective methods from any discipline, the multimodal therapist does not embrace divergent theories but remains consistently within social-cognitive learning theory. The virtues of technical eclecticism (Lazarus, 1967, 2005, 2008) over the dangers of theoretical integration have been emphasized in several publications (e.g., Lazarus, 1989, 1995; Lazarus and Beutler, 1993).

The major criticism of theoretical integration is that it inevitably tries to blend incompatible notions and only breeds confusion (see Beutler et al., 1995). The main claim is that by assessing clients across the BASIC I.D., one is less apt to overlook subtle but important problems that call for correction, and the overall problem identification process is significantly expedited.

Afterword

The best controlled study on MMT to date was completed in July 2013 in Iran. Dr Faisal Alzarrad and Houssam Alcharif (of Al-Ahliyya University) were the principal investigators who persisted with the study despite what they referred to as "difficult unstable life conditions." The title of their study is "Effectiveness of a clinical multimodal treatment package on a sample of obsessive-compulsive disorder patients." The purpose of their study was to examine the effectiveness of MMT in reducing symptom severity and improving quality of life in a group of obsessive–compulsive disorder patients. In their study, 20 patients were randomly divided into two equal groups wherein 10 patients were treated with a medication regimen, and the experimental group received MMT as well as medication. Results on pre- and posttests indicated significant effects in the group receiving MMT. The data indicated a 60% reduction in symptom severity and a 74.4% increase in quality of life relative to the medication only group. A 1-month follow-up showed stable outcomes. The study was accepted for publication by the Association of Arab Universities in their *Journal for Education and Psychology*.

See also: Behavior Therapy: Background, Basic Principles, and Early History; Concepts and Methods of Cognitive Therapies; Diagnostic Processes in Clinical Psychology and Psychotherapy; Epidemiology, Concepts and Methods of; Intellectual Functioning, Assessment of; Motivation in Clinical Interventions; Personality Assessment; Projective Methods in Psychology; Psychotherapy Integration; Therapist–Patient Relationship.

Bibliography

Bandura, A., 1977. Social Learning Theory. Prentice Hall, Englewood Cliffs, NJ.
Bandura, A., 1986. Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice Hall, Englewood Cliffs, NJ.
Beutler, L.E., Consoli, A.J., Williams, R.E., 1995. Integrative and eclectic therapies in practice. In: Bongar, B., Beutler, L.E. (Eds.), Comprehensive Textbook of Psychotherapy. Oxford University Press, New York.
Kwee, M.G.T., Lazarus, A.A., 1986. Multimodal therapy: the cognitive-behavioral tradition and beyond. In: Dryden, W., Golden, W. (Eds.), Cognitive-behavioral Approaches to Psychotherapy. Harper & Row, London.
Lazarus, A.A., 1967. In support of technical eclecticism. Psychological Reports 21, 415–416.
Lazarus, A.A., 1971. Behavior Therapy and Beyond. McGraw-Hill, New York.
Lazarus, A.A., 1976. Multimodal Behavior Therapy. Springer, New York.
Lazarus, A.A., 1989. The Practice of Multimodal Therapy. Johns Hopkins University Press, Baltimore, MA.
Lazarus, A.A., 1995. Different types of eclecticism and integration: let's be aware of the dangers. Journal of Psychotherapy Integration 5, 27–39.
Lazarus, A.A., 1997. Brief But Comprehensive Psychotherapy: The Multimodal Way. Springer, New York.
Lazarus, A.A., 2006. Multimodal therapy: a seven-point integration. In: Stricker, G., Gold, J. (Eds.), A Casebook of Psychotherapy Integration. APA Books, Washington, DC.
Lazarus, A.A., 2008. Multimodal therapy. In: Corsini, R.J., Wedding, D. (Eds.), Current Psychotherapies, eighth ed. Brooks/Cole, Belmont, CA.
Mueser, K.T., Gingerich, S., 2006. The Complete Family Guide to Schizophrenia. The Guilford Press, New York.
Rudolph, J.A., 1985. Multimodal treatment of agoraphobia: a problem-focused approach. In: Lazarus, A.A. (Ed.), Casebook of Multimodal Therapy. Guilford Press, New York.

Multinational Corporations

Bruce Kogut, Columbia Business School, Columbia University, New York, NY, USA
Alicja Reuben, Manhatten College, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A multinational corporation (MNC) is an organizational vehicle to transfer knowledge from one country to another while preserving cash flow and control rights. The MNC has its origins in the uncertainty of international trade and investment, and its modern form coevolved with the emergence of institutional and political structures that permitted owners to realize and retain profits from the utilization of its organizational knowledge in foreign countries and territories. Since operating overseas incurs costs, the MNC is only viable if its transfer of organizational capabilities permits it to compete effectively against domestic and other foreign competition. With the diffusion of knowledge, local competitors often compete with the MNC in the long run. However, an MNC has the advantage of profiting from arbitraging across markets and leveraging its scale and scope globally. The impact of the revolution in information technologies has consequently an ambiguous effect, both enabling the MNC to operate globally more effectively and enabling markets to contract to more efficiently compete against organizational solutions.

A multinational corporation (MNC) is the organizational form by which the knowledge of *how to coordinate and organize work* is transferred from one national institutional setting to another (Kogut, 1992). It represents, on the one hand, the evolution of organizational mechanisms by which to resolve the historical hazards of long-distance trade and investment. On the other, it is the vehicle for the transfer of organizational knowledge and practices. This transfer initially is from the home country to foreign sites, but eventually can consist of the diffusion of organizational and technological knowledge within a transnational network. In this dual role of providing a private ordering of exchange and as a vehicle for the diffusion of organizational practices, it challenges the power of the state and the resident national institutions and frequently is the intended or unwitting agent of social, economic, and political change in the world economy.

Contrary to the emphasis on the transfer of organizing knowledge, the conventional and complementary economic definition stresses the ability of owners and their managerial agents in one country to *control* the operations in foreign countries. There is a frequent confusion that equates the ability to control with the flow of capital across national borders. Since Hymer's thesis (1976), it is recognized widely that capital flow is not the distinguishing characteristic of a MNC (*see* International Business). Capital can flow from one country to another in expectation of higher rates of return. However, this flow may be invested in the form of bonds, or in equity amounts too insignificant to grant control to foreign owners. In this case, this type of investment is treated as a 'portfolio' investment. The central aspect of 'direct investment' is the ownership claim by a party located in one country on the operations of a foreign firm or subsidiary in another. The MNC is, thus, the product of foreign direct investment (FDI), which is defined as the effective control of operations in a country by foreign owners.

The economic definition, however, does not capture the importance of the MNC as the organizational mechanism by which different social and economic systems confront each other. The MNC, because it usually develops in the cultural and social context of one nation, exports its organizational baggage from one institutional setting to another. In this regard, it plays a powerful role as a mechanism by which to transfer organizational knowledge across borders. However, while being foreign implies that it might serve the valuable role of importing new practices, its foreign status also implies that its practices are likely to conflict with existing institutions and cultural norms. Moreover, since MNCs are often large, they pose unusual challenges to national and regional governments who seek to maintain political autonomy and yet are often anxious to seek the investment, technology, and managerial skills of foreign firms.

There are, thus, economic and sociological definitions of the MNC that differ, and yet complement, each other. In the economic definition, the MNC is the control of foreign activities through the auspices of the firm. In the sociological definition, the MNC is the mechanism by which organizational practices are transferred and replicated from one country to another.

Early History

The MNC is defined in some sense arbitrarily by where frontiers are drawn. In ancient Greece, these frontiers were the borders among city-states. In imperial Rome, the new administrative units of an expanding empire and its external boundaries defined the borders. Political borders correspond only approximately to the distribution of cultural and ethnic dispositions. The history of the MNC is tied closely to the origins of trade in and between cultural communities, and these communities remain important in many sectors in the modern economy.

Early trade was often characterized by the difficulty of transacting across borders. Trading has always instigated from the unequal and varied distribution of resources across geographies. The trading of salt was an important factor in most continents and still can be found in rather ancient forms in east

 http://dx.doi.org/10.1016/B978-0-08-097086-8.73004-2

Africa. Towns, such as Salzburg, owed their origins to their fortuitous access to salt mines.

This unequal distribution incited traders to travel long distances and undergo unusual risks for the hope of gain. The dilemma is quite apparent. How can markets for trade develop between distant cultures? Indeed, there is good evidence that early trading was quite precarious. Brown (1969) notes that Hermes was not only the God of messengers but also the God of theft and trade. He was often worshiped at borders, marking the place of trade. Herodotus describes the practice of 'silent trade' of goods at boundary markings in which one party would deposit its goods at the border and the second party would later deposit a corresponding offer. The original party might accept this bid, take the goods, and leave its initial offer. Clearly, only a strong demand for unusual foreign goods could sustain such a treacherous institution of transaction.

Not surprisingly, Curtin (1984) doubts the accuracy of this account. Instead, he stresses the importance of ethnic groups, such as the Gujarati or overseas Chinese, or the force of imperialism and military power as ways in which 'cross-cultural trading' was permitted to develop. But such solutions remained fragile, with foreign enclaves and ethnic groups subject to local hatred and military rule often confronting competition from other powers.

Yet, trade persisted through the centuries. Indeed, the demand for foreign goods was so strong that incredible profits could be realized by international trade. Since transportation was poor, and technology to hold inventories rudimentary, an agent in the right spot at the right time could earn fortunes. And in this laid the great problem. For in times of great demand, prices could rise astronomically, as Braudel (1973) documents in his history of material life. With such distances between the agent and the principal in the exporting country, an agent could easily, and often did, disappear with the profits.

The early solutions were many and provide insight into one of the great properties of the MNC: its ability to organize transactions within its own organizational boundaries. Great fairs were one solution, where principals met their customers in one place. But fairs were intermittent and often distant from final markets. The further development of Roman commercial law during the medieval and Renaissance periods strengthened the liability of the principal and also the legal recourse in principal–agent conflict. However, the application of law between two parties in different legal jurisdictions has always been a difficult form of dispute resolution. In his study of the Jewish Maghrebi traders of the twelfth century, Greif (1989) emphasizes membership in ethnic 'trading communities.' Since ethnic enclaves in foreign sites could monitor the activities of agents, principals could be informed of malfeasance. Agents could then be excluded from future contracts within the community, thus deterring dishonest behavior. Partnerships were also another way in which agents were also bonded to the firm.

Birth of the MNC

However, another solution was the company, where the archetypical cases were the Dutch West Indies and the British East India Company dating from the 1600s and the 1700s. The stock company form permitted the investing of capital in its ownership by outsiders in promise of future dividends. The capital might be raised for a single voyage or, eventually, could be maintained in the company in the form of stock ownership. These companies were primarily engaged in trading, although increasingly they became involved, economically and politically, in foreign countries.

With the evolution of the limited liability law in the nineteenth century, and the diminished role of the state in restricting the growth of the joint stock company, this organizational form expanded rapidly in many European countries as well as in the United States. Along with the growth of equity markets, the financial resources to invest overseas became more available due to advances in the banking system and also in bond markets. Indeed, the great capital needs of the railway industry created an international market for the sale and purchase of railroad bonds. However, international loans and bonds were risky and it was not uncommon for sovereign governments to default, including state governments within the United States.

The increasing wealth of Western countries, along with constraints on the speed by which national industries could absorb new loans, encouraged massive foreign investments by the end of the nineteenth century and the early twentieth century. This was a period of globalization in terms of the percentages of capital outflows to total capital accumulation. Great Britain is estimated to have exported some 25% of its capital prior to World War I and French capital exports were often greater. This capital went to countries hungry to finance their industrial expansion, including the newly industrializing countries of Russia and especially America.

The British were unusual in directing a considerable amount of their outward capital flows in the form of direct investments. A great deal of this investment went into their colonies. Prior to the mid-1800s, companies operating overseas engaged largely in wholesale operations; they did not run factories, operate mines, or own agriculture. The British direct investments were, thus, different than before, as British came to own and run local operations. Moreover, the British invested in South and North America.

The 'freestanding' company, as labeled by Wilkins (1988), was a peculiar form of investment that was quite prominent in many British colonies but hardly unique; thus the French stock market records indicate that many floated companies had no domestic but only foreign activities (Kogut, 1998). The free-standing firm raised capital in the domestic financial market where its administrative office was located, but it operated no domestic activities. All operations were overseas. These companies were multinational in ownership, but uniquely foreign in operations.

The United States offers an unusual case. A rapidly growing country, it imported more capital than it exported up to World War I. However, whereas it imported largely portfolio investments, its outward flows were dominated by FDIs. In other words, American companies showed an early penchant for expanding overseas (Wilkins, 1970), often competing on their organizing advantages. The Singer Company built within two decades of its founding a large factory employing thousands of workers in Scotland. Oil companies, Kodak, Westinghouse, Ford, and mining and agricultural companies all invested overseas. Companies in oil, mining, and agriculture often invested in poorer markets where there were resources to be

found. These early investors were often involved inextricably in the politics of the foreign governments, and the American military itself intervened aggressively numerous times in the Caribbean, Mexico and Central America, and South America. As in the British case, the history of American direct investment occurred in the context of an expanding military presence of the home government. Moreover, since many of these countries were poor, the MNCs responded to the demands of the host nation, especially in the form of concessionary contracts, to provide public services, for example, hospitals, roads, and power (Robinson, 1964). This complicated legacy of the early history of the MNC in Latin America created hostility on the part of the local population that persisted throughout most of the twentieth century.

National Origins and Institutions

It is important to underscore that the MNC usually evolved in the context of specific national institutions, which left an 'imprint' on its organization and capabilities (Westney, 1993). The MNC is a growing firm whose organizational borders have spilled across borders, but whose national imprint still persists. Moreover, since this large firm is usually tied to a larger domestic network of suppliers and customers, its expansion overseas is accompanied by the coinvestments of these other members. This is a pattern seen in American investments in the United Kingdom in the 1950s and repeated by Japanese MNCs investing in the United Kingdom in the 1980s and the 1990s (Dunning, 1993).

Chandler (1990) noted as well that these MNCs reflected the national characteristics of management. In comparing the cases of the largest firms in the United Kingdom, Germany, and the United States, Chandler found that differences in managerial capabilities, reflecting national institutions, explain their success and failure patterns. He particularly criticized the managerial capabilities of British firms, a point not shared by some British historians (Hannah, 1999). But more importantly, Chandler's thesis assumed that size itself constituted the realization of scale and scope of economies instead of the outcome of success and growth. This observation is especially important for understanding the lack of large MNCs in Italy or in Taiwan, both of which have very successful small firm economies but do not have MNCs comparable to other countries of similar levels of economic wealth. Yet, both countries are relatively wealthy and successful, and their many small companies have achieved high rates of exporting. Even in the case of the United States, the evidence implies that American firms, large and small, came to Europe, riding on the back of the national organizing principles of standardization in work methods (Kogut, 1992). Chandler's larger point of the effect of national systems on firm capabilities is largely accepted; his belief that large firms reflect better management because they achieve scale is disputed far more.

Organization of MNCs

The national origins of the MNC influence their subsequent organizational evolution. In the United States, the great post–World War II expansion of MNCs coincided with the diffusion of their adoption of the multidivisional structure. Indeed, the refinement of 'organizational technologies' permitted American firms to manage their rapidly growing operations on a worldwide basis.

The young American MNC began with little knowledge and experience of the foreign market. Initial foreign sales are exports executed within the existing organizational structure. These structures could be functional (e.g., organized by production and sales) or divisional (e.g., organized by geographic area or product division). As foreign sales increased, an international division was created to make the sale and to provide customer support. In consequence, the domestic organizational unit would sell to the international division its products at an internal 'transfer price.' As this price was set by the domestic market, it was unlikely to reflect the competitive conditions in the foreign country. Moreover, the products were often not designed to the tastes of foreign demand. Because of this internal conflict, the American MNC would replace the international division by transferring responsibilities to area divisions (e.g., Europe, Asia, South America) or to product divisions (Stopford and Wells, 1972). While these structures diminished the internal conflict, they ran the hazard that the divisional managers across areas did not cooperate or that managers across product divisions tended to fall back upon focusing on the home market.

European firms grew up in a different environment. Europe was a highly unstable continent during the twentieth century, with significant political and economic conflicts. Borders sometimes changed, and they also marked the necessity to pay commercial duties. The gradual creation of the European Union in the last third of the twentieth century eliminated tariffs and lessened national differences. This history marked their firms. The organizational technologies in most of these countries were quite different from those found in the United States. Holding company structures were common, banks or other financial institutions played powerful roles in ownership and cross-holdings, and firms were tied together in complex financial webs. These structures, often called 'mother–daughter' organizations, reflect the linguistic terms used in northern Europe to describe headquarters and subsidiary. These mother–daughter structures consisted often of a headquarters that held a portfolio consisting of ownership control over dozens of companies, many of them in the same sector but located in different countries. Reporting relationships among these companies and headquarters was less formal than found in American companies, and control was often exercised through the movement of managers who were lifetime employees. These structures were often seen as inefficient, especially compared to the size and organization of American multinationals that were investing rapidly in Europe at this time. European critics called for the restructuring of the European firms and industry along American lines.

The structure of Japanese companies also reflected their historical origins. In Japan, the trading company performed an extraordinary role in the exports during the first two-thirds of the twentieth century. These companies started primarily as wholesalers, but over time developed their own industrial companies located in Japan and eventually overseas. Though the forced imposition of American antitrust law after World

War II disrupted these patterns, trading companies and other affiliated company networks, called *keiretsu*, reconstituted the earlier pattern of constellations of domestic companies competing in home and, through their trading companies, in foreign markets. As in the American and European cases, Japanese firms varied in their internal structures, with functional and even factory-based organization being the predominant structures. Still, a dominant trend emerged over time as the role of trading companies declined, especially for the exportation of industrial and high-technology goods. This trend consisted of the appendage of an international division to the existing structure (Suzuki, 1991).

These different national structures confronted an increasingly more globally integrated environment in the course of the twentieth century. This environment posed the classic organizational problem of balancing integration and differentiation. For the MNC, this problem posed itself as meeting the demands to achieve global scale against the needs of national markets and governments. Given different national principles of organizing the activities of MNCs, there was not a rapid convergence to a global structure. The growth of MNCs, especially in Europe, was stymied by these traditions. Continuing national differences and the failure to develop harmonized European corporate law deterred the emergence of pan-European MNCs. Many of the most important cross-European mergers in the 1970s failed within a decade. Instead, American MNCs, less bound by these separate national traditions, were able to create integrated European strategies.

Nevertheless, it was in Europe where new organizational structures developed to resolve the conflict between integration and differentiation. MNCs from smaller countries offered the laboratory experiments that influenced the evolution of MNCs from larger nations. Because of the small size of many European countries, large firms in Sweden, Switzerland, and the Netherlands quickly became MNCs, and these corporations had far more assets and employees outside their home countries than within. Over the course of the last two decades of the twentieth century, organizational structures developed, labeled as transnational corporations, that distributed global responsibilities to national subsidiaries, created global teams and projects, and encouraged the transfer of 'best practices' across borders.

Diffusion of Organizational Knowledge

In its evolution the MNC is not without serious contradictions. Evolving from its national context, the MNC employs large numbers of employees of diverse nationalities and ethnicities. Westney (1993) notes that a subsidiary is, thus, caught between the institutional pressures to conform to the company norms and values and to the cultural and social influences of its local national environment. At the heart of the evolution of the multination corporation, thus, lies the tension between national institutions and the fragile emergence of a global culture.

The international evolution of the organizational structures of American multinationals mirrored, as is noted above, the broader diffusion of organizational technologies in the home market. The initial investments by a firm took place, often, on the basis of opportunity and the extension to familiar countries. Much like the ethnic trading communities dating back to the earliest times, the inexperienced MNC preferred countries that are culturally similar to what their managers know at home (Johanson and Vahlne, 1978). In these countries, they often established foreign enclaves where their expatriate managers could live in the simulated familiarity of their home environments.

The relevance of the second definition of FDI as the transfer of organizational knowledge is critical to understanding the powerful conflicts posed by the MNC. The MNC, competing often on superior technologies and managerial capabilities, serves as a conduit of knowledge across borders. Some kinds of technologies can be purchased on markets and the flow of licensing payments attached to the sale of the right to use a technology constitutes a nontrivial flow in international balance of payments. Technology can also flow by the movement of people. Just as English craftsmen were imported by continental Europe in the early industrial period, there exists an international market for managers and skilled workers.

However, some knowledge is embedded in organizations and can only be transferred as an organization across borders. The issue is more than whether knowledge is tacit or explicit, for this pertains to the knowledge held by individuals too. The central feature of the diffusion of knowledge by the multinational is that it transfers the knowledge of how to organize and how to coordinate people and across divisions. These organizing principles then provide the capabilities for the firm to achieve quality, speed products to the market, or lower costs (Kogut, 1992). For this the MNC is required, which through its own activities or in joint ventures with host companies, transfers the organizational knowledge.

The global market for knowledge also extends to consultants. The large American consulting practices often have their origins in their acquisition of American organizing principles that they then transferred around the world. The international diffusion of the divisional structure, for example, reflects the knowledge acquired in American consulting practices that was then sold abroad. Channon (1973), for example, observes that half of the firms he observed as adopting the divisional structure relied upon the consulting services of the same American firm. Similarly, British and then American banks spread throughout the world on servicing the needs of their expanding home clients, and then on transferring their practices into these countries (Jones, 1993).

It is important to realize that once this transition period passed, the MNC and multinational service firms, such as those engaged in consulting, banking, and advertising, had permanently and historically changed the global economy. No longer was the flow from the United States to Europe or from one ethnic community to another. The network of connections developed by the MNC permitted the flow of ideas and practices among and between countries. Thus, whereas it took half a century for American practices to diffuse to Europe, the introduction of quality circles from Japan happened within a decade or two of their original innovation. In Europe, Japanese innovations were often diffused by American firms, including consulting companies.

However, this image of the free flow of knowledge needs to be strongly conditioned on the continuing importance of

national institutions. Nations consist of defined cultures and economic and social institutions, such as unions, financial systems, and religious values. These institutions interact and their complex interactions causally influence behaviors. For example, the German system of centralized bargaining, enterprise clubs, strong banks, and social welfare has, until recent times, composed a national configuration of institutions that influences the capabilities of resident firms to be able to manufacture high-quality goods for export markets (Streeck, 1995; Soskice, 1990). The introduction of Japanese and American teams confronts in Germany the presence of existing work councils. These councils are fairly rigid features of the German environment. Since Japanese methods may not be effective without teams as complementary factors in organization, this institutional refusal can effectively deter the diffusion of these methods. In short, national organizing principles are embedded in the wider social institutions. Inconsistency between these institutions and principles, then, is an empirical question of the bargaining strength of vested powers.

Economics and Politics of MNCs: the Role of CSR

Since the MNC is intrinsic to FDI, theories of FDI must account for why one country invests in another and why this investment is carried out within organizational boundaries of a firm (see Buckley and Casson, 1976; *see* Foreign Investment: Direct). In distinguishing between portfolio and direct investment, Hymer noted that firms operate at a disadvantage in foreign markets and hence they must have an offsetting competitive advantage to compete overseas. These advantages for overseas investments are the same ones that allow a firm to compete and grow in the home market. These observations have important implications. The first is that direct investment is the growth of the firm across borders and hence the firm expands internationally on what it has learned at home. This observation is the basis for the evolutionary theory of the firm. The second observation that Hymer made is that firms that expand overseas, because they have competitive resources, are also likely to be large and to belong to oligopolistic industries.

In these observations, we can understand the ambivalence expressed in popular and policy debates regarding the MNC. Competition among MNCs often is the extension of their home domestic and oligopolistic rivalry that spills across national borders. In many global industries, the same company names dominate each country's list of the largest firms inside their national frontiers. No matter if it is Poland or France, Singapore or Mexico, the same MNCs will be found in the local oligopolistic industries (e.g., consumer goods or automobiles). Because they are large even in their home markets, investments by MNCs can have a large impact on a host country (Caves, 1974).

As a consequence, the MNC has often been the subject of debates concerning national sovereignty and welfare. In recent decades, acquisitions have generally been the primary way by which multinationals invest in wealthy foreign countries, where the vast proportion of direct investment is concentrated. Given the size of a MNC and occasional national importance of the targeted acquisition, even wealthy countries frequently evidence discomfort, if not outright public hostility, to multinational investments. Moreover, MNCs are sometimes the vehicles for foreign policies of their home or host country. The decision, for example, of the United States to embargo technology and investment flows to Cuba, the former Soviet Union, Iran, and other countries periodically has caused conflict with other countries.

MNCs are especially problematic in developing countries. By definition, developing countries are relatively poor, thus both in need of capital and concerned over their loss of independence. As discussed above, the history of MNCs in developing countries is marked by its origins in policies of imperialism and colonialism. Especially in Latin America, where a school of thought labeled *Dependencia* has been influential, the concern over dependence on the United States resulted in efforts to curb the power of the MNCs by restricting the amount of equity ownership a foreign firm could hold in a domestic company or by prohibiting investment in certain sectors. Mexico's constitution forbids foreign investment in the oil industry; Brazil pursued for a long time a policy to restrict foreign participation in the electronics industry.

The other side of the coin is that MNCs bring investment and technology to the foreign country. Vernon (1966) hypothesized that innovations start in wealthy countries. As the market is saturated and as oligopolistic rivalry increases, MNCs are pushed out from their home markets to expand abroad in new markets and to locate less expensive places. Thus, Vernon seized both sides of the debate, not only recognizing the value of the transfer of technology but also emphasizing the oligopolistic nature of multinational investment.

It is, in fact, difficult to draw simple conclusions regarding the relationship of foreign investment and national growth. Countries such as Singapore, Malaysia, and Thailand have encouraged FDI actively. The growth in China's coastal sector is indisputably linked to the massive investments by MNCs. However, historically, Japan and Korea have pursued more cautious policies regarding investments by MNCs. In these countries, the state has often negotiated the terms for entry by MNCs, sometimes requiring licensing to domestic competitors as a price. The efficacy of such policies for these countries is much disputed. However, for many other countries, the intervention of the government in demanding licenses unquestionably leads to internal corruption and to insufficient domestic competition.

MNCs have a natural incentive to collude with corrupt or failed governments or to pursue 'regulatory arbitration' that leads to locating in countries with poor labor and environmental standards. Consequently, the past decades have seen the rise of nongovernmental organizations (NGOs) that seek to pressure MNCs into improving ethical practices into their operations. Pressure tactics consist of publicizing bad practices and thereby damaging the value of brand labels, or lobbying national and international regulatory agencies. Many companies, such as Walmart and Nike in the United States, have claimed to have dramatically improved labor and ethical standards in their own operations and also those in their supply chains. In many cases, NGOs act as partners to MNCs to help them monitor the standards of suppliers or to help them deliver drugs to remote areas in developing country. A particularly interesting partnership is the cooperation between

international agencies, such as the World Health Organization, and private companies for the provision of aid through humanitarian supply chains (Van Wassenhove, 2005).

Institutions and MNCs

There are many channels by which a country can absorb foreign technology and managerial techniques. Most of the evidence shows, however, that prohibitions on the inflows of direct investment can be very costly for many countries. With their domestic industries still to be developed, a developing country requires substantial investment. Some countries, primarily in Asia, have been able to achieve very high savings rates to finance their industries without direct investment. Moreover, high savings rates, plus political stability, create growth, and growth attracts foreign portfolio capital. A poor country that prohibits FDI but does not have high rates of saving entirely depends on portfolio capital. The history of debt and currency crises in the 1990s convinced many poor countries that FDI was a preferable means of attracting capital, because it could not be easily pulled out of a country on short notice in response to a financial crisis.

However, MNCs also respond to the volatility in the global market. This volatility derives from changes in exchange rates, politics, and productivity. Having achieved sufficient experience and having established subsidiaries around the world, the MNC might choose to close a plant in one location and open plants in new locations (Kogut and Kulatilaka, 1994). Of course, such actions might provoke a response by labor, but historically, labor has been organized by national, not by international, organizations (Martinelli, 1975). Yet, there is also the possibility that locations lose some kinds of plants but gain more sophisticated investments. Cantwell (1989) proposed that some regions and countries pull multinational investments. Yet, it has long been noticed that FDI among developed countries flows to high-cost locations. Regions such as Silicon Valley, Baden-Wuertemberg, and Singapore attract multinational investments not because wages are low, but because productivity levels are high and workers are well trained.

MNCs are not only drawn to regions with high agglomeration but also contribute to these clusters through technology transfer and diffusion. For example, Blalock and Gertler (2008) estimate that about a third of output growth in a sample of Indonesian firms is due to gains in technology diffusion to local firms. It is an interesting question though whether MNCs agglomerate in traditional domestic firm clusters or in clusters dominated by MNCs. Alfaro and Chen (2009) analyze this question, finding that MNCs are more strongly attracted to MNC clusters in which technology diffusion plays an important role. This is an intriguing result that suggests that MNCs shape the global economic landscape.

A related and persisting literature concerns the competing effects of multinationals on crowding out domestic firms and technology spillovers. Kosova (2010) found that in the short run, foreign entry does crowd out domestic firms, but that after entry, multinational investment increases demand and the survival of domestic competitors; this work thus also confirms, as discussed above, that FDI is a mode of technology and knowledge transfer. Chang and Xu (2008) analyzed a similar question for China, finding that foreign entry crowds out regional but not national domestic firms. On the contrary, foreign entrants face enhanced competition from 'reformed' Chinese firms, which moved from state ownership to greater private control.

This last observation of the enhanced capabilities of domestic firms in China points of course to one of the most important changes in the composition of FDI. Increasingly, developing countries have given rise to their own MNCs acting in the region and sometimes globally, a trend that has boomed since analyzed by Lall (1983). Often, the MNC acts as a training center in the developmental strategies of emerging economies. Similar to the history of firms from early developed countries, the MNCs from emerging markets grow often aggressively through mergers and acquisitions, such as the acquisition of Jaguar and Rover operations in the United Kingdom by Tata, of Volvo by the Chinese firm Greely, or of IBM laptop computer operations by Lenovo also from China. Firms such as the Brazilian Embraer in aircraft, the Korean electronics firm Samsung, and the Chinese white good manufacturer Haier are major competitors in their respective markets, as are some firms from the so-called 'transition economies,' such as the Russian energy company Gazprom. The analysis of the imprinting of the national origins and their competitive advantage of these firms is a major current research area.

Globalization and the Digital Information Economy

The peculiar conflict in the world economy is the growing trend toward economic integration without a concomitant growth in global political and social institutions to regulate and arbitrate this trend. MNCs are rapidly adopting advanced information technologies to increase the efficiency and capabilities of their operations. Valuable information, such as software or financial services, is transmitted digitally. Governments are often unable to tax the value of these services or control their content. Moreover, information technologies, often supported by private telecommunication networks, permit, for example, a unit in Germany to control the manufacturing operations located in Brazil.

Digital technologies permit a high degree of integration and, in the short run, aggravate the gap between countries that are rich and those that are poor and do not have the infrastructural capabilities. However, these technologies are beginning to have profound effects on some regional economies. In parts of China, India, Israel, and elsewhere, advanced satellite transmission transmits digitally encoded work between their sites and other organizational units of MNCs located elsewhere. Earlier the Indian engineer from Bangalore might have tried to migrate to the United States to bring his human capital to a more attractive labor market, whereas the increasing wages and job prospects in India are encouraging many to stay at home and participate digitally in the world economy.

These trends have unclear effects on MNCs. They permit more easily the development of projects that are in continual development, as work passes from one unit to the next as the day advances. They also allow more easily the coupling of less

expensive labor in one country with more expensive skilled labor in another.

At the same time, the origins of the MNC laid in its ability to organize labor on a worldwide basis on principles other than ethnic and cultural identities. This organization has never been without cost and risk. The growth of a world digital economy permits alternative ways by which labor can cooperate and be coordinated on a world basis. The intriguing question at this point in history is whether the MNC, though still a vital presence in the world economy, nevertheless, will recede relatively in importance as information technologies reduce the meaning of geographic distance.

See also: Development and the State; Diaspora; Globalization: Geographical Aspects; Globalization: Legal Aspects; International Business; International Law and Treaties; International Trade: Economic Integration.

Bibliography

Alfaro, L., Chen, M., 2009. The global agglomeration of multinational firms. NBER Working Papers, #15576.

Blalock, G., Gertler, P., 2008. Welfare gains from foreign direct investment through technology transfer to local suppliers. Journal of International Economics 74, 402–421.

Braudel, F., 1973. Capitalism and Material Life, 1400–1800. Harper and Row, New York.

Brown, N.O., 1969. Hermes the Thief: The Evolution of a Myth, second ed. University of Wisconsin Press, Madison, WI.

Buckley, P.J., Casson, M., 1976. The Future of the Multinational Enterprise. MacMillan, London.

Cantwell, J., 1989. Technological Innovations and Multinational Corporations. Blackwell, London.

Caves, R.E., 1974. Causes of direct investment: foreign firms. Shares in Canadian and United Kingdom manufacturing industries. Review of Economic Statistics 56, 79–93.

Chandler, A.D., 1990. Scale and Scope: The Dynamics of Industrial Capitalism. Belknap Press, Cambridge, MA.

Chang, S.-J., Xu, D., 2008. Spillovers and competition among foreign and local firms in China. Strategic Management Journal 29, 495–518.

Channon, D.F., 1973. The Strategy and Structure of British Enterprises. Harvard University, Boston.

Curtin, P.D., 1984. Cross-cultural Trade in World History. Princeton University Press, Princeton, NJ.

Dunning, J., 1993. The governance of Japanese and US manufacturing affiliates in the UK: some country-specific differences. In: Kogut, B. (Ed.), Country Competitiveness: Technology and the Organizing of Work. Oxford University Press, Oxford, UK.

Greif, A., 1989. Reputation and coalitions in medieval trade: evidence on the Maghribi traders. Journal of Economic History 39, 857–882.

Hannah, L., 1999. Marshal's 'trees' and the global 'forest': were 'Giant Redwoods' different? In: Lamereaux, N.R., Raff, D.M.G., Temin, P. (Eds.), Learning by Doing in Markets, Firms, and Countries. University of Chicago Press, Chicago, pp. 253–286.

Hymer, S., 1976. The International Operations of National Firms: A Study of Direct Investment (Ph.D. thesis). Massachusetts Institute of Technology.

Johanson, J., Vahlne, J.-E., 1978. A model for the decision making process affecting pattern and pace of the internationalization of the firm. In: Ghertman, M., Leontiades, J. (Eds.), European Research in International Business. North Holland, Amsterdam, pp. 9–27.

Jones, G., 1993. British Multinational Banking, 1830–1990. Oxford University Press, New York.

Kogut, B., 1992. National organizing principles of work and the erstwhile dominance of the American multinational corporation. Industrial and Corporate Change 1, 285–325.

Kogut, B., Kulatilaka, N., 1994. Operating flexibility, global manufacturing, and the option of a multinational network. Management Science 40, 123–139.

Kosova, R., 2010. Do foreign firms crowd out domestic firms? evidence from the Czech Republic. Review of Economics and Statistics. 92, 861–881.

Lall, S., 1983. The New Multinationals: The Spread of Third World Enterprise. Wiley, Chichester, UK.

Martinelli, A., 1975. Multinational corporations, national economic policies and labor unions. In: Lindberg, L., Alford, R., Crouch, C., Offe, C. (Eds.), Stress and Contradiction in Modern Capitalism. Heath, Lexington, MA, pp. 425–443.

Robinson, R.D., 1964. International Business Policy. Holt, Rinehart, and Winston, New York.

Soskice, D., 1990. Wage determination: the changing role of institutions in advanced industrialized countries. Oxford Review of Economic Policy 6, 36–61.

Stopford, J.M., Wells, L.T., 1972. Managing the Multinational: Organization of the Firm and Organization of the Subsidiaries. Basic Books, New York.

Streeck, W., 1995. German capitalism: does it exist? Can it survive? In: Crouch, C., Streeck, W. (Eds.), Modern Capitalism or Modern Capitalisms? Pinter, London.

Suzuki, Y., 1991. Japanese Management Structures, 1920–1980. St. Martin's Press, New York.

Van Wassenhove, L.N., 2005. Humanitarian aid logistics: supply chain management in high gear. Journal of the Operations Research Society 57, 475–489.

Vernon, R., 1966. International investment and international trade in the product life cycle. Quarterly Journal of Economics 80, 190–207.

Westney, D.E., 1993. Institutional theory and the multinational corporation. In: Westney, D.E., Ghoshal, S. (Eds.), Organization Theory and the Multinational Corporation. St. Martin's Press, New York, pp. 53–76.

Wilkins, M., 1970. The Emergence of Multinational Enterprise: American Business Abroad from the Colonial Era to 1914. Harvard University Press, Cambridge, MA.

Wilkins, M., 1988. The free-standing company, 1870–1914: an important type of British foreign direct investment. Economic History Review 39, 259–282.

Multiple Comparisons, Statistics of

Peter H Westfall, Texas Tech University, Lubbock, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

An overview of the current state of multiple comparisons is given, with special application to the social sciences. The problems caused by multiple comparisons are detailed, and Bayesian and frequentist solutions are discussed. Special emphasis is given on definition of the 'family' of statistical inferences, as well as on pre-registration of studies.

Introduction

'Multiple comparisons' (hereafter abbreviated MC) refer to making multiple inferences, and the 'MC problem' refers to errors that can arise in that case. MC arise in virtually all statistical studies, because it is extremely rare that a researcher will make only one claim from their data. For example, in multiple regression analysis one considers multiple inferences for the regressor variables X_j on a response Y; in meta-analysis, one considers multiple inferences concerning effects of a single X on a response Y in different settings; in a subgroup analysis, one considers multiple inferences about one or more X variables in data subgroups (male, female, young, old, etc.); and in an experiment, one considers multiple inferences about differences between the means of various treatment groups (A versus B, A versus C, B versus C, etc.). The MC problem arises in variable and model selection as well where there are inferences about literally millions of different models, often performed automatically. MC problems also arise when a researcher performs multiple tests of the same hypothesis – parametric, nonparametric, or covariate-adjusted using different covariates, for example.

MC are not restricted to a single study of data in a researcher's hands: The 'file drawer' effect (Sterling et al., 1995), where the occasional 1 out of 20 erroneously significant studies are published, and the remaining 19 out of 20 correctly insignificant results remain in 'file drawers' of researchers at various universities, is also a manifestation of the MC problem. Neither are MC restricted to concerns over erroneous $p < 0.05$ significances: It is common (but extremely misguided) practice in the social and behavioral sciences to declare regressor effects as 'different' if one is significant ($p < 0.05$) and the other is not ($p > 0.05$), and this is another error that occurs with multiple inferences (Gelman and Stern, 2006; Freedman, 2005).

While the consequences of MC are more broad as suggested above, one of the problems resulting from MC can be illustrated with a simple probability calculation: If k null hypotheses are tested, each at the $\alpha = 0.05$ significance level, then the probability of erroneously rejecting one or more hypotheses is $1 - (1 - 0.05)^k$ under independence. With $k = 14$ hypotheses, the false positive probability is 51%, and with $k = 45$ the probability jumps to 90%. This probability calculation shows how easy it can be to 'rescue' a study where there are truly no effects: By looking at the data from numerous angles, performing multiple tests, erroneous statistical significance is virtually assured.

Another way to view the MC problem, one that does not require an assumption of independence, is to recognize that p-values are uniformly distributed between 0.0 and 1.0 under their respective null hypotheses (e.g., Westfall et al., 2011, p. 4). Thus, one may expect that 5% of them will be less than 0.05 when there is no true effect, no matter what is the dependence structure.

Recently, MC have received much attention in the biological sciences, spurred by problems of multiple inferences with genetic data, where the number of genes evaluated in a study can be easily in the thousands. In this case, even if the genes tested are unrelated to the biological phenomenon, hundreds of p-values will erroneously indicate significance at the 5% level. It is now standard in such analyses to adjust for the MC problem (Efron, 2010). Similarly, with neuroimaging studies where multiple tests are performed at thousands of brain locations, adjustment for MC is considered mandatory (Bennett et al., 2009).

MC have received less attention in the social sciences. Null hypothesis significance testing in general has been disparaged in this literature (Nickerson, 2000), leading many social science researchers to think that multiplicity adjustment is not needed, simply because testing itself is not needed. Indeed, there are several valid reasons for avoiding statistical tests, including concerns about practical versus statistical significance, the implausibility of the precise null hypothesis, and general confusion over the meaning and interpretation of p-values.

But on the other hand, concerns over replicability have recently surfaced in the social and behavioral sciences (Bohannnon, 2014). Multiplicity and replicability are intimately related: If a test erroneously indicates significance, as is likely to occur with MC, then that test result will not replicate in 95% of future studies. But the multiplicity/replicability problem is not restricted to tests. Any evaluation of statistical data from multiple angles, and selection of the 'best' angle for presentation, however 'best' is defined, will necessarily present a biased view, one that will tend not to replicate. This phenomenon is also known as 'regression to the mean.' In the medical literature, Ioannadis (2005) found that even in a selection of randomized comparative trials, the gold standard for biomedical research, results replicated in only 19 of 28 (67.9%) follow-up studies. For observational studies, which are more commonplace in social and behavioral sciences, Ioannadis noted a paltry 1 in 6 (16.6%) replication rate. Bohannon (2014) reports a similarly alarming replication failure rate of

more than 33% in a selected group of popular social psychology studies.

The most important take-home message for researchers in the social sciences about MC is simply this: The MC problem makes it easy to do sloppy, nonreplicable science. It is very easy to get nice-looking results, whether in terms of small p-values (or large p-values, if model confirmation is needed), or large effect sizes, with multiple looks at the data, followed by judicious selection of the results that the researcher wants. The incentives to do so are great (grants, tenure, promotion), while the penalties for not doing so are unfortunately scant, particularly in the social sciences where replication of studies is not practiced as much as it is in the medical literature. Glaeser (2006) and Simmons et al. (2011) note that such researcher incentives are likely to blame for much bad science. Taken to its extreme, the culture of multiple selective inferences includes selective exclusion of data to make results look better; an example was recently reported in which a researcher claimed that "in the area of marketing and (to a lesser extent) social psychology, many consciously leave out data to reach significance without saying so" (Enserink, 2012).

Prespecification of a Family of Inferences

Multiple comparisons procedures (MCPs) are methods to correct for problems that arise when there are multiple inferences. As with all inferential methods in statistics, there are frequentist and Bayesian methods to correct for MC. Methods that control the familywise error rate and false discovery rate (FWER and FDR respectively, defined below) are primarily frequentist. There are many Bayesian methods for dealing with multiplicity, a primary one being hierarchical modeling (Gelman et al., 2012), which attempts to solve the multiplicity problem through shrinkage estimation, discussed further below.

To use any of these methods for MC, the researcher must prespecify a *family* of inferences to be considered in the study. With frequentist methods, one must determine k, the number of inferences in the family, and the adjustments for MC typically involve dividing the error rate α by functions of k. With Bayesian methods, one must specify the joint likelihood and multivariate prior for the k parameters of interest, and then allow the multiplicity adjustment to take place automatically through Bayesian posterior inference.

Prespecification of the family is the most difficult and controversial aspect of MCPs. It is rare in the social sciences to prespecify data analysis plans (DAPs), but there are growing efforts to make pre-registration a prerequisite for publication (Bohannon, 2014). But even with pre-registration, it is difficult to decide exactly which inferences should be included in the MC family. For example, an experiment might consider the effects of three treatment variables on a sample of men and on a sample of women. There are three treatment comparisons within each sex, AB, AC, and BC. Should the family include $k = 6$ comparisons? Or should there be two families of $k = 3$ each, one for men and one for women? Unfortunately, there is no obviously right answer to the question in general, and the consequences of one choice over the other may be great: It can happen that there are no statistical significances when adjusting for MC using a family of $k = 6$, while everything is significant when adjusting for MC using two families of $k = 3$.

The pharmaceutical literature provides some guidance for choice of a family. In clinical trials, it is often required to pre-register a study with the regulatory agency. Such pre-registration includes medical details and statistical details, including sample size determination, statistical models to be used, and even a statement concerning whether and how MCPs are to be performed (ICH, 1998). Westfall and Bretz (2010) provide guidelines as to which medical endpoints should be considered a family, for the purposes of drawing multiple inferences. They state that this family must include all efficacy endpoints (primary and secondary) for which medical or marketing claims will be made, following conclusion of the study.

Applying the lessons from the pharmaceutical industry to research in the social sciences, one might suggest that journals require prespecifications of DAPs as a prerequisite to publication. These DAPs should include similar information as in pharmaceutical DAPs, including a statement as to which tests are considered a 'family,' and how multiplicity adjustment is to be performed. Taking the analogy from the pharmaceutical industry, one might suggest that a research claim of a social science study can be made only if the inference was included in the family, as defined in the pre-registration. Any other claims are disallowed.

Further guidance on choice of a family is given by Westfall et al. (2011, p. 33). There, the authors substitute the word 'question' for 'inference':

- The questions form a natural and coherent unit.
- The questions are considered simultaneously in the decision-making process.
- The questions form a set over which you play 'pick the winner.'

Of the three above, the third provides the clearest guidance. For example, in the description of the experiment above, if the researcher was willing to make (and publish) any claim of a difference between any two treatments, for either sex (even if there was just one such significance), then the family clearly must contain all $k = 6$ inferences across both sexes, even though classical methods such as Tukey's method for MC do not cover this case. Alternative methods are needed in this case, and fortunately they are readily available (Westfall et al., 2011).

FWER and FDR

In statistical hypothesis testing, a Type I error is the outcome of erroneously rejecting a true null hypothesis. The FWER and FDR provide different characterizations of Type I error rates when there are multiple inferences.

Once the family is prespecified, the future outcomes of the inferences appear as shown in **Table 1**. There are k total hypotheses, of which (unbeknownst to the researcher) k_0 are true nulls and k_1 are true alternatives. Suppose the hypotheses are tested by whatever method, multiplicity adjusted or not. Then after the study, the researcher will reject a random number R of hypotheses, and make claims concerning them. Of these R

Table 1 Outcomes of inferences from a prespecified family

		Action		
		'Accept' H_0	*Reject H_0*	*Total*
State of nature	H_0 true		V	k_0
	H_0 false			k_1
	Total	$k-R$	R	k

rejections, a random number V of them will be erroneously rejected. The remaining entries inside the table are deliberately left blank to avoid unnecessary symbolic clutter.

In the setup of **Table 1**, the null hypotheses need not be point nulls such as H_0: $\delta = 0$. They can also be directional nulls (such as H_0: $\delta \leq 0$) or even interval nulls such as H_0: $\delta \leq -0.20$ or $\delta \geq +0.20$, as is used in bioequivalence studies (Berger and Hsu, 1996). Alternatively, one can specify an 'effect size' measure θ (see, e.g., Cohen, 1992) and test a null hypothesis of the form H_0: $\theta \leq 0.20$; rejection would ensure that the effect size is meaningful rather than simply nonzero. Since much of the criticism of hypothesis testing in the social sciences literature is aimed at point null hypotheses, use of more general null hypotheses removes much of the criticism concerning significance testing, and makes MCPs relevant for social science research.

The FWER is defined as $\text{FWER} = \text{P}(V > 0)$. With more technical precision, $\text{FWER} = \sup_{\Theta} \text{P}(V > 0)$, where Θ is a set of relevant nuisance parameters. In words, the FWER is the (maximum) probability of making an erroneous claim for the prespecified family of inferences.

The FDR is defined loosely as $\text{FDR} = \text{E}(V/R)$. However, since R can be zero, the definition must be modified so that the ratio V/R will be counted as zero in the expectation calculation. And again, with more technical precision, $\text{FDR} = \sup_{\Theta} \text{E}(V/R)$, where Θ is a set of relevant nuisance parameters. In loose words, the FDR is the percentage of erroneous claims, among all claims that are made in a study.

Both the FWER and FDR are quantities that the researcher would like to control at a low level, typically 0.05. If $\text{FWER} = 0.05$ for a given procedure, then the researcher may be confident that all significances are in fact real, because only one in 20 studies will result in one or more incorrectly significant results. If $\text{FDR} = 0.05$ for a given procedure, then the researcher cannot be confident that all discovered significances are in fact real, but can be loosely confident that 95% of them are real and 5% are errors. The FDR criterion is looser than the FWER criterion because it allows false positives, hence FDR-controlling methods are usually less conservative than FWER-controlling methods.

To illustrate the difference between FWER and FDR more specifically, consider a hypothetical study in which there are $k = 100$ hypotheses tested, $k_0 = 60$ of which are true nulls and $k_1 = 40$ of which are true alternatives. Suppose also that the true alternatives have large effect sizes and large sample sizes, so that statistical significance is virtually assured, that the tests are independent, and that the tests are all performed at the usual 0.05 significance level, unadjusted for multiplicity. After the study, there will be a random number V of significances among the $k_0 = 60$ true nulls; here V is a binomial random variable with 60 trials and probability 0.05 of 'success.' Hence, $\text{FWER} = \text{P}(V > 0) = 1 - \text{P}(V = 0) = 1 - \{60!/(60! \times 0!)\}\ 0.05^0\ 0.95^{60-0} = 0.954$. In other words, there is a 95.4% chance of finding a false significance in this hypothetical study.

On the other hand, $\text{FDR} = \text{E}(V/R) = \text{E}\{V/(V+40)\} = \sum_{v=0}^{60} \frac{v}{v+40} \times \frac{60!}{(60-v)!v!} 0.05^v. \times 95^{60-v} = 0.068$. In other words, the expected ratio of false significances to total significances is 0.068 for this study.

To control either the FWER or the FDR at the 0.05 level in this hypothetical study, one can use a critical p-value threshold lower than the standard 0.05. To control FWER, one can use a low threshold $1 - 0.95^{1/60} = 0.000854$. But the FDR is already close to 0.05, so not nearly as much adjustment is needed: Direct evaluation of the FDR expression above shows that 0.036 works in this case to control the FDR at the level 0.05. (These threshold calculations are for illustrative purposes only; in practice one cannot use them because the number of true null, k_0 is unknown, and because it is unlikely that all alternatives will be highly significant.)

FWER-Controlling Methods

A method is said to control FWER at level α if $\text{FWER} \leq \alpha$ whenever that method is used. Methods that do not control the FWER can have $\text{FWER} > \alpha$ for certain parameter configurations.

Some notation: The null hypotheses are H_i, with corresponding p-values p_i, $i = 1,\ldots,k$. The ordered p-values are $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(k)}$, corresponding to respective hypotheses $H_{(1)}$, $H_{(2)}$, … ,$H_{(k)}$. The simplest FWER-controlling method is the Bonferroni method, where one rejects all $H_{(i)}$ for which $p_{(i)} \leq \alpha/k$; here α is the desired FWER. Note that the Bonferroni method does not require the ordered p-values, but the stepwise methods defined below do.

Most FWER-controlling methods are quite similar to the Bonferroni method in their extreme dependence on k, with larger k making the procedure more conservative. However, there are several methods for improving the power, while retaining FWER control. A good general reference for the methods given below is Westfall et al. (2011).

Stepwise and testing methods: These methods involve testing the hypotheses sequentially, in order of the p-values. The simplest and most robust of these is Holm's method, where the hypotheses are tested in step-down fashion, from most significant least. If $p_{(1)} \leq \alpha/k$, then $H_{(1)}$ is rejected and the testing continues, otherwise the testing stops. At the second step, if $p_{(2)} \leq \alpha/(k-1)$, then $H_{(2)}$ is rejected and the testing continues, otherwise the testing stops. Testing continues in this fashion until a stop, or until all hypotheses have been rejected. It is clear that the method uniformly improves over Bonferroni, because the denominators $k-j$ of the critical values $\alpha/(k-j)$ are smaller than the denominator k of the Bonferroni method.

There are step-up variations as well, and these typically require assumptions concerning the correlation structure. The simplest and most famous is Hochberg's method, defined as follows: If $p_{(k)} \leq \alpha/1$, then reject all of $H_{(1)},\ldots,H_{(k)}$ and stop. Otherwise 'accept' $H_{(k)}$ and continue testing. At the second step, if $p_{(k-1)} \leq \alpha/2$, then reject all of $H_{(1)},\ldots,H_{(k-1)}$ and stop. Otherwise 'accept' $H_{(k-1)}$ and continue testing. Testing

continues in this fashion until a stop, or until no hypotheses can be rejected.

Hochberg's step-up testing method always finds the same significances as Holm's method, but sometimes more. The improved power comes at a price though: The method can fail to control the FWER in cases of negatively correlated p-values (Westfall et al., 2001).

Closed and logically constrained testing methods: Holm's method is a special case of what are called closed testing procedures; these require enumeration and testing of all subset intersection hypotheses. These methods can be computationally demanding in general because $2^k - 1$ intersection tests are needed, although there are shortcuts in many cases, Holm's method being an example. In cases where the hypotheses satisfy logical constraints, as is the case with all pairwise comparisons, the denominators of the critical values are substantially smaller than those of Holm's method, leading to more power.

Using correlations: Holm's method and its variants do not utilize the correlation structure among the p-values. When there are correlations, these methods can be improved uniformly when such correlations are incorporated into the testing method. Implicitly, the denominators of the critical values are smaller depending on the correlation structure. The higher the correlation, the greater the reduction in the denominator. Tukey's method and Dunnett's method are classical examples of such correlation-utilizing methods (Westfall et al., 2011).

Using non-normal distributional characteristics: When data are highly discrete, the tests involving the more discrete data can be downweighted so as not to contribute to the multiplicity adjustment, further improving power (Westfall and Young, 1993).

The Benjamini–Hochberg FDR-Controlling Method

Like FWER-controlling methods, a method is said to control FDR at level α if FDR $\leq \alpha$ when that method is used. Methods that do not control the FDR can have FDR $> \alpha$ for certain plausible parameter configurations. The most commonly used FDR-controlling method is the Benjamini–Hochberg (B–H) method, a step-up method defined as follows:

- If $p_{(k)} \leq k\alpha/k$, then reject all of $H_{(1)},\ldots,H_{(k)}$ and stop. Else 'accept' $H_{(k)}$ and continue testing.
- If $p_{(k-1)} \leq (k-1)\alpha/k$, then reject all of $H_{(1)},\ldots,H_{(k-1)}$ and stop. Else 'accept' $H_{(k-1)}$ and continue testing.
- If $p_{(k-2)} \leq (k-2)\alpha/k$, then reject all of $H_{(1)},\ldots,H_{(k-2)}$ and stop. Else 'accept' $H_{(k-2)}$ and continue testing.
- … Continue in this fashion until a stop or until no hypotheses are rejected.

The B–H method is similar to Hochberg's FWER-controlling method in that it is a step-up method, and in that it can fail in cases of negatively dependent p-values. But what is most interesting is the comparison of the critical values: Hocherg's method compares $p_{(j)}$ to $\alpha/(k-j+1)$, whereas the B–H method compares $p_{(j)}$ to $\alpha/(k/j)$. Simple math shows that $\alpha/(k/j) \geq \alpha/(k-j+1)$, proving that the B–H method always finds the same significances as the Hochberg method, but possibly more. A more telling story is shown in **Figure 1**. With $k = 100$ tests, the thresholds to be exceeded by the Hochberg method are clearly much smaller than those of the B–H method. It is also apparent that the Hochberg thresholds are similar to the simple Bonferroni 0.05/100 threshold for all but the largest p-values, whereas the B–H thresholds linearly move from the Bonferroni threshold 0.05/100 to the unadjusted threshold 0.05.

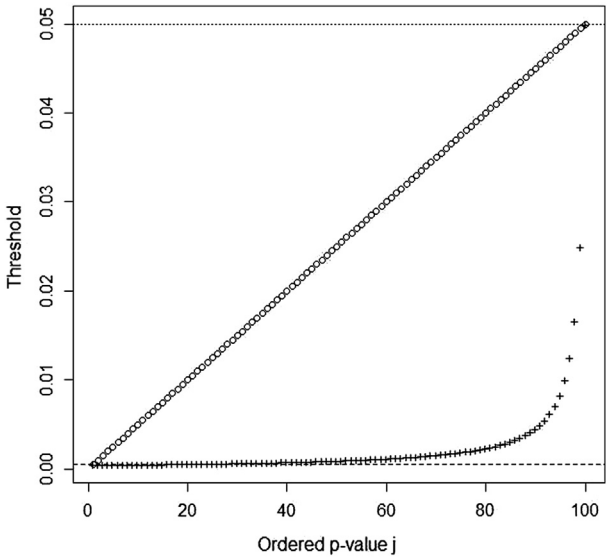

Figure 1 Critical p-value thresholds for Hochberg's method (+'s) and the B–H method (o's) when there are $k = 100$ tests. The Bonferroni threshold 0.05/100 is shown as a dashed line and the unadjusted threshold 0.05 is shown as a dotted line. R code is `t = 1:100`; BHt = 0.05/(100/t − 1); Ht = 0.05/(100 − t + 1); plot(t, BHt, xlab = "Ordered p-value j", ylab = "threshold"); abline(h = 0.05/100, lty = 2); abline(h = 0.05, lty = 3).

There are many variants and extensions of the B–H method; Efron (2010) is an excellent reference.

Adjusted p-values

Adjusted p-values are interpreted just like ordinary p-values in that you simply compare them to 0.05 (or other desired thresholds). The simplest case is the standard Bonferroni method: Significances occur when $p_j \leq 0.05/k$; or equivalently, when $kp_j \leq 0.05$. Thus, kp_j is the Bonferroni-adjusted p-value. Usually, this number is truncated at 1.0, since probabilities can never be greater than 1.0, leading to the definition $\min(kp_j, 1.0)$. For stepwise and correlation-utilizing testing methods the formulas for the adjusted p-values are more complex, but the logic is similar; see Westfall et al. (2011).

If the adjusted p-values are obtained using an FWER-controlling method, then all adjusted p-values ≤ 0.05 correspond to significances at the FWER = 0.05 level. If the adjusted p-values are obtained using an FDR-controlling method, then all adjusted p-values ≤ 0.05 correspond to significances at the FDR = 0.05 level.

Software for performing the analyses above are readily available in the R and SAS software. Both of these packages provide adjusted p-values, in R they are in p.adjust and

elsewhere, in SAS they are in PROC MULTTEST and elsewhere.

The code in Table 2 illustrates how to calculate adjusted p-values for a collection of 100 tests, using both the Hochberg FWER-controlling and B–H FDR-controlling methods. Both the p.adjust function in R and the PROC MULTTEST procedure in SAS require the user to input a collection of raw (unadjusted) p-values, obtained by any means whatsoever. Using R, the code is as shown in Table 2.

Figure 2 shows a comparison of the adjusted p-values obtained using the code in Table 2. Any p-value below the 0.05 threshold indicates significance using the chosen criterion. The most significances (23 in the simulation) are obtained using the raw, unadjusted p-values, but these offer no control over FWER or FDR. Fewer significances (10 in the simulation) are found using the B–H-adjusted p-values; for these the interpretation is 'roughly 5% of the 10 significances are Type I errors.' The fewest significances (four in the simulation) are found using the Hochberg-adjusted p-values; for these the interpretation is that none of the four are Type I errors, because a Type I error will occur in only 1 in 20 studies, and this particular study is not likely to be one of those.

Table 2 Code for obtaining adjusted p-values in R

```
set.seed(12345)  ## to ensure replicability
pvals =          ## Runif is null case, runif^2
  runif(100)^2     provides smaller pvals as under
                 ## the alternative. In practice
                    the pvals come from the study.
                 hoch.adjp = p.adjust(pvals,
                   method = "hochberg") bh.adjp
                   = p.adjust(pvals,
                   method = "BH")
```

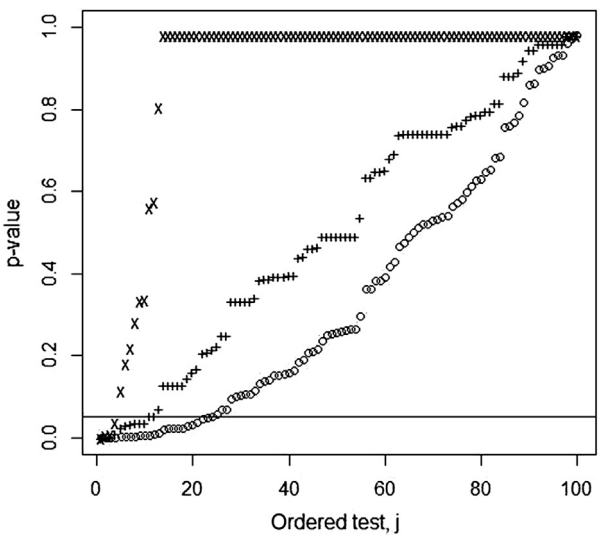

Figure 2 Unadjusted p-values (o's), B–H-adjusted p-values (+'s), and Hochberg-adjusted p-values (X's). Any p-value below the horizontal line at 0.05 indicates significance using the selected Type I error control criterion. R code is plot(sort(pvals), ylim = c(0,1), ylab = "p-value", xlab = "ordered test, j"); points(sort(bh.adjp), pch = "+"); points(sort(hoch.adjp), pch = "x"); abline(h = 0.05).

Bayesian Methods

A rather different approach to multiplicity adjustment is found in the Bayesian paradigm, wherein all inferences, whether individual or multiple, about parameters θ_j are obtained from posterior distribution of the vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, given the data $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_n)$. The posterior distribution is given by $p(\boldsymbol{\theta}|\boldsymbol{\gamma}) = c \times L(\boldsymbol{\theta}|\boldsymbol{\gamma}) \times p(\boldsymbol{\theta})$, where $L(\boldsymbol{\theta}|\boldsymbol{\gamma})$ is the likelihood function, $p(\boldsymbol{\theta})$ is the researcher's prior, and c is a constant of proportionality to ensure integration to 1.0. Modern software allows researchers to sample random vectors $\boldsymbol{\theta}^*$ from $p(\boldsymbol{\theta}|\boldsymbol{\gamma})$ with relative ease; with enough samples, accurate posterior inferences are obtained.

Classical MCPs have been historically disparaged by the Bayesians; however, influential Bayesian statisticians have recently modified their stances. Berger (1985) stated that multiple testing is no more difficult than individual testing, while Berger (2012) stated that failure to adjust for multiplicity is a reason for nonreplicability. Berry (1987) dismissed MC because of their inconsistency with the likelihood principle, but more recently commented "I no longer think Bayesians have an inside track. Multiple comparisons is an example" (reported in Stangl et al., 2012).

As it turns out, there are many correspondences between Bayesian and frequentist MCPs, and the modern approach is to look for methods that have reasonable credibility from both standpoints. One step in this direction was given by Westfall et al. (1997), who showed how the Bonferroni-adjusted p-value kp_j corresponds to the Bayesian posterior probability $\mathrm{P}(\theta_j = 0|\boldsymbol{\gamma})$. The result is obtained by comparing posterior inferences under the case (1) the priors $\mathrm{P}(\theta_j = 0)$ are all set at 0.5, versus the case (2) the prior $\mathrm{P}(\theta_j = 0, \text{all } j)$ is set at 0.5. The comparison of the posterior probabilities $\mathrm{P}(\theta_j = 0|\boldsymbol{\gamma})$ under cases (1) and (2) loosely corresponds to the comparison of the raw p_j with the Bonferroni-adjusted kp_j. An implication is that Bonferroni-style adjustment is reasonable when the joint null hypothesis is plausible, as is the case, for example, in experiments involving extrasensory perception.

A more direct correspondence, one that does not consider adjustments to the prior distribution, is found between FDR-adjusted p-values and Bayesian posterior probabilities (Efron et al., 2001), although the result is not fully Bayesian. More direct correspondences are obtained using a variant of FDR-adjusted p-values called the 'local false discovery rate' (Efron, 2010).

Other correspondences are detailed by Westfall (2005), who noted that classical (e.g., Tukey and Dunnett) simultaneous confidence intervals for parameters θ_j are identical to simultaneous Bayes intervals when using standard vague multivariate priors. Further, as noted by Westfall (1988), Bayesian simultaneous probabilities of rankings of parameters θ_j share similarities to frequentist methods in their conservativeness. For example, if one wants to know whether θ_a is the largest of the collection $\{\theta_j\}$, the Bayesian joint probability $\mathrm{P}(\theta_a \geq \max\{\theta_j\}|\boldsymbol{\gamma})$ will typically be much less than the probability for any individual comparison $\mathrm{P}(\theta_a \geq \theta_j|\boldsymbol{\gamma})$, corresponding loosely to the difference between the more conservative simultaneous (adjusted) inference for the parameters θ_j versus the less conservative individual (unadjusted) inferences for the θ_j found in

frequentist statistics. Westfall et al. (2011: Chapter 20) provide software and examples to obtain these and other Bayesian style inferences.

But what distinguishes Bayesian approaches to multiplicity adjustment most from frequentist methods is that, when a hierarchical prior is used, the parameter estimates themselves are adjusted according to the chosen family of k inferences. An example is shown in Gelman et al. (2013, pp. 109–110), wherein the parameters θ_j are binomial tumor development probabilities for rodents in treatment group j. Using a hierarchical prior $\theta_j \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta)$; $p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}$, they derive a posterior distribution $\{\theta_j|\gamma, \alpha, \beta\} \sim \text{Beta}(\alpha + f_j, \beta + n_j - f_j)$, where f_j and n_j are the tumor frequency and sample size in group j, respectively. The parameters α and β likewise have a posterior distribution given γ; this distribution characterizes the mean and variance in the θ_j that is supported by the data γ. The mean is $\alpha/(\alpha+\beta)$ and the variance is roughly proportional to $1/(\alpha+\beta)$. Thus, in cases where the θ_j are nearly identical, $(\alpha+\beta)$ is large.

For example, suppose $\alpha = 300$ and $\beta = 700$ are supported by the data, suggesting a mean θ_j of 0.300 and a small variance. In this case the raw data $f_j/n_j = 30/50 = 0.600$ are 'shrunken' to a posterior probability estimate around $(300+30)/(1000+50) = 0.314$. On the other hand, with large variance and the same mean, say $\alpha = 0.3$ and $\beta = 0.7$, there is little posterior shrinkage: The same raw data $f_j/n_j = 30/50 = 0.600$ yield a nearly identical posterior estimate around $(0.3+30)/(1+50) = 0.594$. Thus the Bayesian solution to the problem of 'cherry picking' extreme results is to shrink the extreme estimates toward the overall mean, with shrinkage dependent upon the (estimated) true variation in the parameters.

While shrinkage is an attractive solution to the MC problem, it is not a panacea, because there are strong prior assumptions. Hierarchical priors assume that all parameters are different. While this may seem reasonable on its face, there are compelling reasons to consider mixture priors that place a large (say, 50% or more) probability on equality of parameters (or near equality as in Berger and Delampady, 1987) with the remaining prior density continuous and more or less uniform. Such priors are called 'spike and slab' priors, often used for variable selection in regression (Scheipl et al., 2012). In the MCP arena, such priors also can be used to identify homogeneous subsets of means in ANOVA (Gopalan and Berry, 1998). It turns out that posterior inference is highly dependent upon whether there is a spike and slab versus smooth prior: In the unadjusted inference case, Berger and Sellke (1987) find that the posterior probability of a null hypothesis can be more than 0.30 when a p-value is 0.05 and spike and slab priors are used, whereas Casella and Berger (1987) find that the posterior probability of a null is less than 0.05, given a p-value of 0.05, when smooth priors are used. In the multiple inference case, the differences are more extreme because of further dependence on multivariate prior probabilities for joint nulls (Westfall et al., 1997; Gönen et al., 2003). Thus, prespecification and justification of the family of inferences *and* their joint prior is extremely important in Bayesian MCPs, as the posterior conclusions are highly nonrobust to alternative prior specifications.

Conclusion

There are excellent solutions to the multiplicity problem, on both sides of the Bayes/frequentist fence. However, in all cases the methods require prespecification of the family and analysis methods (including priors in the Bayesian case), with solid justification, before data collection. This is best accomplished via pre-registration. Any post hoc construction to solve the MC problem is flawed because there is no clear definition of the family (and of the prior on the family, in the Bayesian case) over which the inferences are to be performed. In the absence of pre-registration, it seems to me that the best solution is to go back to the basics of science, confirming the results not by applying post hoc multiplicity corrections, but instead by replicating the study in fresh contexts.

See also: Bayesian Statistics; Hypothesis Testing in Statistics; Hypothesis Tests, Multiplicity of; Meta-Analysis: Theory; Nonparametric Regression; Probability Theory: Interpretations; Significance, Tests of.

Bibliography

Bennett, C.M., Wolford, G.L., Miller, M.B., 2009. The principled control of false positives in neuroimaging. Social Cognitive and Affective Neuroscience 4, 417–422.

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York.

Berger, J.O., 2012. Reproducibility of Science: P-values and Multiplicity. Available at: http://www.stat.purdue.edu/symp2012/slides/Purdue_Symposium_2012_Jim_Berger_Slides.pdf.

Berger, J.O., Delampady, M., 1987. Testing precise hypotheses. Statistical Science 2, 317–335.

Berger, J.O., Sellke, T., 1987. Testing a point null hypothesis: the irreconcilability of p-values and evidence. Journal of the American Statistical Association 82, 112–122.

Berger, R.L., Hsu, J.C., 1996. Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science 11, 283–319.

Berry, D.A., 1987. Interim analysis in clinical trials: the role of the likelihood principle. The American Statistician 41, 117–122.

Bohannon, J., May 23, 2014. Replication effort provokes praise – and 'bullying' charges. Science 788–789.

Casella, G., Berger, R.L., 1987. Reconciling Bayesian and frequentist evidence in one-sided testing problem. Journal of the American Statistical Association 82, 106–111.

Cohen, J., 1992. A power primer. Psychological Bulletin 112, 155–159.

Efron, B., 2010. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction. Cambridge University Press, New York.

Efron, B., Tibshirani, R., Storey, J.D., Tusher, V., 2001. Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association 96, 1151–1160.

Enserink, M., June 25, 2012. Rotterdam marketing psychologist resigns after university investigates his data. ScienceInsider. Available at: http://news.sciencemag.org/education/2012/06/rotterdam-marketing-psychologist-resigns-after-university-investigates-his-data.

Freedman, D.A., 2005. Linear statistical models for causation: a critical review. In: Everitt, B., Howell, D. (Eds.), The Wiley Encyclopedia of Statistics in Behavioral Science. John Wiley, Chichester, UK.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D., 2013. Bayesian Data Analysis, third ed. Chapman and Hall/CRC (Chapman & Hall/CRC Texts in Statistical Science).

Gelman, A., Hill, J., Yajima, M., 2012. Why we (usually) don't have to worry about multiple comparisons. Journal of Research on Educational Effectiveness 5, 189–211.

Gelman, A., Stern, H., 2006. The difference between "significant" and "not significant" is not itself statistically significant. The American Statistician 60, 328–331.

Glaeser, E.L., 2006. Researcher Incentives and Empirical Methods. Technical Working Paper 329. National Bureau of Economic Research, Cambridge, MA. http://www.nber.org/papers/t0329.

Gopalan, R., Berry, D.A., 1998. Bayesian multiple comparisons using Dirichlet process priors. Journal of the American Statistical Association 93, 1130–1139.

Gönen, M., Westfall, P.H., Johnson, W.O., 2003. Bayesian multiple testing for two-sample multivariate endpoints. Biometrics 59, 76–82.

ICH Harmonised Tripartite Guideline, 1998. E9 Topic: Statistical Principles for Clinical Trials. Available at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf.

Ioannidis, J.P.A., 2005. Contradicted and initially stronger effects in highly cited clinical research. Journal of the American Medical Association 294, 218–229.

Nickerson, R.S., 2000. Null hypothesis significance testing: a review of an old and continuing controversy. Psychological Methods 5, 241–301.

Scheipl, F., Fahrmeir, L., Kneib, T., 2012. Spike-and-Slab priors for function selection in structured additive regression models. Journal of the American Statistical Association 107, 1518–1532.

Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allow presenting anything as significant. Psychological Science 22, 1359–1366.

Stangl, D., Inoue, L.Y.T., Irony, Telba Z., 2012. Celebrating 70: an interview with Don Berry. Statistical Science 27, 144–159.

Sterling, T.D., Rosenbaum, W.L., Weinkam, J.J., 1995. Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. The American Statistician 49, 108–112.

Westfall, P., 1988. Comment on Berry. The American Statistician 42, 88.

Westfall, P.H., 2005. Comment on Benjamini and Yekutieli, 'false discovery rate adjusted confidence intervals for selected parameters'. Journal of the American Statistical Association 100, 85–89.

Westfall, P., Bretz, F., 2010. Multiplicity in clinical trials. In: Chow, Shein-Chung (Ed.), Encyclopedia of Biopharmaceutical Statistics, third ed. Marcel Decker Inc., New York, pp. 889–896.

Westfall, P.H., Johnson, W.O., Utts, J.M., 1997. A Bayesian perspective on the Bonferroni adjustment. Biometrika 84, 419–427.

Westfall, P.H., Ho, S.-Y., Prillaman, B.A., 2001. Properties of multiple intersection-union tests for multiple endpoints in combination therapy trials. Journal of Biopharmaceutical Statistics 11, 125–138.

Westfall, P., Tobias, R., Wolfinger, R., 2011. Multiple Comparisons and Multiple Tests Using SAS, second ed. SAS® Press, Cary, NC.

Westfall, P.H., Young, S.S., 1993. Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment. Wiley, New York.

Multiple Imputation

John Barnard, deCODE genetics, Cambridge, MA, USA
Nathaniel Schenker, Harvard University, Cambridge, MA, USA
Donald B. Rubin, National Center for Health Statistics, Hyattsville, MD, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 10204–10210, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Imputation, that is, filling in a value for each missing datum, is a common approach to handling missing data that has many desirable properties. For example, it allows the application of standard analytic methods that have been developed for complete data sets. Moreover, in the context of the production of public-use data sets, it allows the data producer to incorporate information that might not be available to secondary users of the data, and it ensures consistency of analyses across secondary users since the missing-data problem is fixed in the same way for all of the analyses. Multiple imputation retains the advantages of imputing once for each missing datum, and it provides the added advantage of reflecting the uncertainty about the missing values so that statistically valid analyses may be carried out. With multiple imputation, the missing values are filled in two or more times using an appropriate distribution for the missing values, resulting in two or more completed data sets. The standard complete-data analysis is then applied to each completed data set, and the resulting analyses are combined in a simple way to yield the final inference. Multiple imputation has a Bayesian motivation, but it has been shown to have good frequentist properties.

Missing data are ubiquitous in social science research. For example, they occur in a survey on people's attitudes when some people do not respond to all of the survey questions, or in a longitudinal survey when some people drop out. A common technique for handling missing data is to impute, that is, fill in, a value for each missing datum. This results in a completed data set, so that standard methods that have been developed for analyzing complete data can be applied immediately (*see* Nonsampling Errors; Missing Data).

Imputation has other advantages in the context of the production of a data set for general use, such as a public-use file. For example, the data producer can use specialized knowledge about the reasons for missing data, including confidential information that cannot be released to the public, to create the imputations; and imputation by the data producer fixes the missing data problem in the same way for all users, so that consistency of analyses across users is ensured. When the missing-data problem is left to the user, the knowledge of the data producer can fail to be incorporated, analyses are not typically consistent across users, and all users expend resources addressing the missing-data problem.

Although imputing just one value for each missing datum satisfies critical data-processing objectives and can incorporate knowledge from the data producer, it fails to achieve statistical validity for the resulting inferences based on the completed data. Specifically, for validity, the resulting estimates based on the data completed by imputation should be approximately unbiased for their population estimands, confidence intervals should attain at least their nominal coverages, and tests of null hypotheses should not reject true null hypotheses more frequently than their nominal levels. But a single imputed value cannot reflect any of the uncertainty about the true underlying value, and so analyses that treat imputed values just like observed values systematically underestimate uncertainty. Thus, using standard complete-data techniques will result in standard error estimates that are too small, confidence intervals that undercover, and p-values that are too significant; even if the modeling for imputation is carried out carefully. For example, large-sample results by Rubin and Schenker (1986) show that for simple situations with 30% of the data missing, single imputation under the correct model followed by the standard complete-data analysis results in nominal 90% confidence intervals having actual coverages below 80%. The inaccuracy of nominal levels is even more extreme in multiparameter problems (Rubin, 1987: Chapter 4), where nominal 5% tests can easily have rejection rates of 50% or more when the null hypothesis is true.

Multiple imputation (Rubin, 1987) retains the advantages of single imputation while allowing the data analyst to obtain valid assessments of uncertainty. The key idea is to impute two or more times for the missing data using independent draws of the missing values from a distribution that is appropriate under the posited assumptions about the data and the mechanism that creates missing data, resulting in two or more completed data sets, each of which is analyzed using the same standard complete-data method. The analyses are then combined in a simple generic way that reflects the extra uncertainty due to having imputed rather than actual data. Multiple imputations can also be created under several different models to display sensitivity to the choice of missing-data model.

Theoretical Motivation for Multiple Imputation

The theoretical motivation for multiple imputation is Bayesian (*see* Bayesian Statistics), although the procedure has excellent properties from a frequentist perspective. Formally, let Q be the population quantity of interest, and suppose the data can be

partitioned into the observed values $\mathbf{X_{obs}}$ and the missing values $\mathbf{X_{mis}}$. If $\mathbf{X_{mis}}$ had been observed, inferences for Q would have been based on the complete-data posterior density $p(Q|\mathbf{X_{obs}}, \mathbf{X_{mis}})$. Because $\mathbf{X_{mis}}$ is not observed, inferences are based on the actual posterior density $p(Q|\mathbf{X_{obs}})$, which can be expressed as

$$p(Q|\mathbf{X_{obs}}) = \int p(Q|\mathbf{X_{obs}}, \mathbf{X_{mis}})p(\mathbf{X_{mis}}|\mathbf{X_{obs}})\mathrm{d}\mathbf{X_{mis}} \quad [1]$$

Equation [1] shows that the actual posterior density of Q can be obtained by averaging the complete-data posterior density over the posterior predictive distribution of $\mathbf{X_{mis}}$. In principle, the set of multiple imputations under one model are repeated independent draws from $p(\mathbf{X_{mis}}|\mathbf{X_{obs}})$. Thus, multiple imputation allows the data analyst to approximate eqn [1] by separately analyzing each data set completed by imputation and then combining the results of the separate analyses.

Analyzing a Multiply-Imputed Data Set

The exact computation of the posterior distribution (eqn [1]) by simulation would require that an infinite number of values of $\mathbf{X_{mis}}$ be drawn from $p(\mathbf{X_{mis}}|\mathbf{X_{obs}})$. Fortunately, simple approximations work very well in most practical situations with only three to five imputations.

Inferences for Scalar *Q*

Suppose, as is typically true in the social and behavioral sciences, that if the data were complete, inferences for Q would be based on a point estimate $\hat{Q}$, an associated variance/covariance estimate U, and a normal reference distribution. When data are missing and there are M sets of imputations for the missing data, the result is M sets of complete-data statistics, say, $\hat{Q}_m$ and U_m, $m = 1, \ldots, M$.

The following procedure is useful for drawing inferences about Q from the multiply imputed data (Rubin, 1987; Rubin and Schenker, 1986). The point estimate of Q is the average of the M completed-data estimates,

$$\overline{Q} = \sum_{m-1}^{M} \hat{Q}_m \Big/ M$$

and the associated variance estimate is

$$T = \overline{U} + \left(1 + M^{-1}\right)B$$

where $\overline{U} = \sum_{m-1}^{M} U_m/M$ is the average within-imputation variance, and $B = \sum_{m-1}^{M}(\hat{Q}_m - \overline{Q})^2/(M-1)$ is the between-imputation variance (cf the usual analysis of variance breakdown in 42318). The approximate reference distribution for interval estimates and significance tests is a t distribution with degrees of freedom

$$\nu = \left(\nu_m^{-1} + \hat{\nu}_{obs}^{-1}\right)^{-1}$$

where $\nu_m = (M-1)\hat{\gamma}^{-2}$, $\hat{\nu}_{obs} = \nu_{com}(1-\hat{\gamma})(\nu_{com}+1)/(\nu_{com}+3)$, $\hat{\gamma} = (1+M^{-1})B/T$ is the approximate Bayesian fraction of missing information, and ν_{com} is the complete-data degrees of freedom (see Barnard and Rubin, 1999 for the derivation).

Significance Tests for Multicomponent Estimands

Multivariate analogues of the expressions given in the previous section for scalar Q are summarized by Rubin (1987: Section 3.4). Methods also exist for likelihood-ratio testing when the available information consists of point estimates and evaluations of the complete-data log-likelihood-ratio statistic as a function of these estimates and the completed data.

With large data sets and large models, such as in the common situation of a multi-way contingency table, the complete-data analysis might produce only a test statistic and no estimates (cf Multivariate Analysis: Discrete Variables (Overview)). Rubin (1987: Section 3.5) provided initial methods for situations with such limited information, and improved methods exist that require only the M complete-data χ^2 statistics (or equivalently the M complete-data p-values). These methods, however, are less accurate than methods that use the completed-data estimates.

Creating Multiple Imputations

Ideally, multiple imputations are M independent random draws from the posterior predictive distribution of $\mathbf{X_{mis}}$ under appropriate Bayesian modeling assumptions. Such imputations are called repeated imputations by Rubin (1987: Chapter 3). In practice, approximations often are used and work well.

Modeling Issues

An initial modeling issue in creating imputations is that the predictive distribution for the missing values should be conditional on all observed values. This consideration is particularly important in the context of a public-use database. Leaving a variable out of the imputation model is equivalent to assuming that the variable is not associated with the variables being imputed, conditionally given all other variables in the model; imputing under this assumption can result in biases in subsequent analyses, with estimated parameters representing conditional association pulled toward zero. Because it is not known which analyses will be carried out by subsequent users of a public-use database, ideally it is best not to leave any variables out of the imputation model. Thus, it is desirable to condition imputations on as many variables as possible and to use subject-matter knowledge to help select those variables that are likely to be used together in subsequent complete-data analyses, as by Clogg et al. (1991).

Imputation procedures can be based on explicit models or implicit models, or even combinations (see, e.g., Rubin, 1987: Chapter 5; Rubin and Schenker, 1991). An example of a procedure based on an explicit model is stochastic normal regression imputation, where imputations for missing values are created by adding normally distributed errors to predicted values obtained from a least-squares regression fit. A common procedure based on implicit models is hot-deck imputation, which replaces the missing values for an incomplete case by the values from a matching complete case, where the matching is carried out with respect to variables that are observed for both the incomplete case and complete cases.

The model underlying an imputation procedure, whether explicit or implicit, can be based on the assumption that the reasons for missing data are either ignorable or nonignorable (see Rubin, 1976; Missing Data). The distinction between an ignorable and a nonignorable model can be illustrated by a simple example with two variables, X and Y, where X is observed for all cases whereas Y is sometimes missing. Ignorable models assert that a case with Y missing is only randomly different from a complete case having the same value of X. Nonignorable models assert that there are systematic differences between an incomplete case and a complete case even if they have identical X-values. An important issue with nonignorable models is that, because the missing values cannot be observed, there is no direct evidence in the data to address the assumption of nonignorability. It can be important, therefore, to consider several alternative models and to explore sensitivity of resulting inferences to the choice of model. In current practice, almost all imputation models are ignorable; limited experience suggests that in major surveys with limited amounts of missing data and careful design, ignorable models are satisfactory for most analyses (see, e.g., Rubin et al., 1995).

Incorporating Proper Variability

Multiple imputation procedures that incorporate appropriate variability across the M sets of imputations within a model are called 'proper' by Rubin (1987: Chapter 4) where precise conditions for a method to be proper are also given; also see Meng (1994) and Rubin (1996). Because, by definition, proper methods reflect sampling variability correctly, inferences based on the multiply-imputed data are valid from the standard repeated-sampling (i.e., design-based) frequentist perspective.

One important principle related to incorporating appropriate variability is that imputations should be random draws rather than best predictions. Imputing best predictions can lead to distorted estimates of quantities that are not linear in the data, such as measures of variability and correlation, and it generally results in severe underestimation of uncertainty and therefore invalid inferences.

For imputations to be proper, the variability due to estimating the model must be reflected along with the variability of data values given the estimated model. For this purpose, a two-stage procedure is often useful. Fixing parameters at a point estimate (e.g., the maximum likelihood estimate), across the M imputations and drawing $\mathbf{X}_{\mathbf{mis}}$ from this distribution-generally leads to inferences based on the multiply-imputed data that are too sharp. First drawing the parameters from their joint posterior distributions is preferable.

The two-stage paradigm can be followed in the context of nonparametric methods such as hot-deck imputation as well as in the context of parametric models with formal posterior distributions for parameters. The simple hot-deck procedure that randomly draws imputations for incomplete cases from matching complete cases is not proper because it ignores the sampling variability due to the fact that the population distribution of complete cases is not known but rather is estimated from the complete cases in the sample. Rubin and Schenker (1986, 1991) discussed the use of the bootstrap (*see* Resampling Methods of Estimation) to make the hot-deck procedure proper, and called the resulting procedure the 'approximate Bayesian bootstrap.' The two-stage procedure first draws a bootstrap sample from the complete cases and then draws imputations randomly from the bootstrap sample. Thus, the bootstrap sampling from the complete cases before drawing imputations is a nonparametric analogue of drawing values of the parameters of the imputation model from their posterior distribution before imputing conditionally upon the drawn parameter values. Heitjan and Little (1991) illustrated the use of bootstrap sampling to create proper imputations with a parametric model.

Choice of *M*

The choice of M involves a tradeoff between simulating more accurately the posterior distribution and using a smaller amount of computing and storage. The effect of the value of M on accuracy depends on the fraction of information about the estimand Q that is missing (Rubin, 1987: Chapters 3, 4). With ignorable missing data and just one variable, the fraction of missing information is simply the fraction of data values that are missing. When there are several variables and ignorable missing data, the fraction of missing information is often smaller than the fraction of cases that are incomplete because of the ability to predict missing values from observed values; this is often not true with nonignorable missing data, however.

For the moderate fractions of missing information ($<30\%$) that occur with most analyses of data from most large surveys, Rubin (1987: Chapter 4) showed that a small number of imputations (say, $M = 3$ or 4) results in nearly fully efficient estimates of Q. In addition, if proper multiple imputations are created, then a substantial body of work suggests that the resulting inferences generally have close to their nominal coverages or significance levels, even when the number of imputations is moderate (Clogg et al., 1991; Heitjan and Little, 1991; Rubin and Schenker, 1986; additional publications cited by Rubin, 1996).

Use of Iterative Simulation Techniques

Recent developments in iterative simulation, such as data augmentation (Tanner and Wong, 1987) and Gibbs sampling (Gelfand and Smith, 1990), can facilitate the creation of multiple imputations in complicated parametric models (*see* Markov Chain Monte Carlo Methods). Consider a joint model for $\mathbf{X}_{\mathbf{obs}}$ and $\mathbf{X}_{\mathbf{mis}}$ governed by a parameter, say θ. The data augmentation (Gibbs sampling) procedure that results in draws from the posterior distribution of θ produces multiple imputations as well. (Schafer, 1997) developed algorithms that use iterative simulation techniques to multiply impute data when there are arbitrary patterns of missing data and the missing-data mechanism is ignorable.

Some Recent Applications of Multiple Imputation in Social-Science Research

To illustrate settings in which multiple imputation can be useful, brief descriptions of several recent applications of multiple imputation to social-science research are now given. References to many other examples are provided by Rubin (1996).

Multiple Imputation to Achieve Comparability of Industry and Occupation Codes over Time

The first large-scale application of multiple imputation for public-use data involved employment information from the decennial census. In each census, such information is obtained from individuals in the form of open-ended descriptions of occupations. These descriptions are then coded into hundreds of industries and occupations, using procedures devised by the Census Bureau. Major changes in the procedures for the 1980 census resulted in public-use databases from the 1980 census with industry and occupation codes that were not directly comparable to those on public-use databases from previous censuses, in particular, the 1970 census. This lack of comparability made it difficult to study such topics as occupation mobility and labor force shifts by demographic characteristics.

Because of the size of the 1970 public-use databases (about 1.6 million cases) and the inaccessibility of the original descriptions of occupations, which were stored away on microfilm, it would have been prohibitively expensive to recode the 1970 data according to the 1980 scheme. There existed, however, a smaller (about 127 000 cases) double-coded sample from the 1970 census, that is, a sample with both 1980 and 1970 occupation codes, that had been created for other purposes. Based on Bayesian models estimated from the double-coded sample, five sets of 1980 codes were imputed for the 1970 public-use database as outlined by Clogg et al. (1991). Simulations described by Rubin and Schenker (1987) support the validity of multiple-imputation inferences in this application. Finally, analyses described by Schenker et al. (1993) demonstrated potential gains in efficiency from basing analyses on the multiply imputed public-use sample rather than on the smaller double-coded sample, even though the actual 1980 codes are known on the latter and unknown on the former.

Multiple Imputation for Missing Data in a National Health Survey

The National Health and Nutrition Examination Survey (NHANES) is conducted by the National Center for Health Statistics (NCHS) to assess the health and nutritional status of the United States population and important subgroups. The data from NHANES are obtained through household interviews and through standardized physical examinations. NCHS has recently made available a database from the third NHANES, in which multiple imputation based on complicated multivariate models and iterative simulation techniques was used to handle missing data. This multiple-imputation project is significant because it demonstrates the feasibility of generating proper multiple imputations for a public-use sample obtained by a complex survey design with many observations and variables, varying rates of missingness on key variables of interest, and various patterns of nonresponse.

Documentation available with the database contains descriptions of the models and procedures used to multiply impute data for the third NHANES, as well as descriptions and examples of how to analyze the multiply-imputed data. Earlier work on the project is summarized by Schafer et al. (1996). In particular, a simulation study of the validity of multiple-imputation interval estimates was conducted based on a hypothetical population constructed from data from the third NHANES as well as earlier versions of NHANES. Results indicated that the imputation procedures were successful in creating valid design-based repeated-sampling inferences.

Multiple Imputation in a Study of Drinking Behavior with Follow-up for Nonresponse

The Normative Aging Study was a longitudinal study of community-dwelling men conducted by the Veterans Administration in Boston. In a survey on drinking behavior conducted as part of this study, a large fraction of the men surveyed provided essentially complete information, whereas for a smaller fraction, there was background information available but no information on drinking behavior. Because drinking behavior is a sensitive subject, there was concern that such nonresponse might be nonignorable.

In a follow-up survey, information on drinking behavior was obtained from about one-third of the initial nonrespondents. Glynn et al. (1993) multiply imputed data on drinking for the roughly two-thirds of the initial nonrespondents who remained, under the assumption that, in the stratum of initial nonrespondents, given the follow-up information as well as the background information that had been collected previously, nonresponse was now ignorable. It was found that inferences about the effect of retirement status on drinking behavior (adjusting for age) were very sensitive to whether the multiply-imputed data or only the data for the initial respondents were used.

A Contest: Multiply Imputing the Consumer Expenditure Survey

Raghunathan and Rubin (1997) describe a simulation study to examine the frequentist properties of various methods of generating multiple imputations that was funded by the United States Bureau of Labor Statistics (BLS). The simulation study used real data from the Consumer Expenditure Survey (CES), which is the most detailed source of expenditures, income, and demographics collected by the US government. The objective of the simulation study was to compare frequentist performance of imputation-generation methods that differed by whether they fully accounted for parameter uncertainty and/or whether they conditioned on all observed variables. The recommended approach in the multiple-imputation literature is to generate imputations that account for all sources of uncertainty and condition on all observed information. Some BLS staff questioned this recommendation, preferring to use simpler methods for generating imputations that do not account for parameter uncertainty and exclude some variables, in this case measures of expenditure, because of concerns about endogeneity.

The artificial population from which the simulation samples were drawn consisted of complete reporters from the CES. BLS staff generated 200 independent samples, each of approximate size 500 from the constructed population, then imposed an ignorable monotone missing-data mechanism on each sample, resulting in 200 incomplete samples.

Two variables in the samples had missing data: ethnicity and income. The incomplete samples were then sent to the multiple-imputation team, who were unaware of the exact nature of the missing-data mechanism, for analysis. All the variables used in the missing-data mechanism as well as some other variables were provided to the multiple-imputation team without noting which were which.

The multiple-imputation team generated five imputations for each incomplete sample under a variety of approaches: accounting for parameter uncertainty and using all variables, accounting for parameter uncertainty but excluding some variables (expenditure information), ignoring parameter uncertainty and using all variables, and ignoring parameter uncertainty and excluding some variables. The multiply-imputed samples were sent to BLS for analysis. The results were that the recommended method for generating imputations, accounting for parameter uncertainty and including all variables, produced confidence intervals that had the nominal coverage whereas all of the other methods produced intervals that undercovered, with the method that ignored parameter uncertainty and excluded expenditure information having the worst performance.

Multiple Imputation in the Survey of Consumer Finances

Kennickell (1998) reported on the use of multiple imputation for handling missing data and for disclosure limitation in the Survey of Consumer Finances (SCF). The SCF is a triennial survey conducted by the US Board of Governors of the Federal Reserve System in cooperation with the Statistics of Income Division of the Internal Revenue Service. The SCF is a major source of detailed asset, liability, and general financial information and includes over 2700 questions. The SCF is focused on the details of households' finances, and due to the sensitivity of this topic it suffers from substantial unit and item nonresponse rates. In addition, because the SCF asks many questions about monetary amounts, the skip patterns in the SCF are very complex, resulting in complicated response patterns. These factors make direct analysis of the incomplete SCF data extremely difficult for general users.

To combat the problem of missing data in the SCF, multiply-imputed versions of the SCF have been and continue to be released to the public. The SCF multiple imputations were generated using a novel variation of the Gibbs sampling algorithm that combines great modeling flexibility with computational tractability. Multiple imputation is also being used for disclosure limitation of income information in the SCF Kennickell (1997), based on ideas put forth by Rubin (1993) for protecting sensitive data by creating via imputation completely artificial populations.

Handling 'Don't Know' Survey Responses in the Slovenian Plebiscite

The Republic of Slovenia separated from Yugoslavia on 8 October 1991. One year before, Slovenians voted on independence through a plebiscite. Of eligible voters, 88.5% voted in favor of independence. To predict the results of the plebiscite, the Slovenian government inserted questions into the Slovenian Public Opinion (SPO) Survey that was conducted 4 weeks prior to the plebiscite. The SPO is a face-to-face survey of approximately 2000 voting-age Slovenians containing questions on many aspects of Slovenian life, including education and health. It has been conducted one to two times each year for the past 30 years. The plebiscite-related questions inserted into the SPO had three possible responses: 'yes,' 'no,' and 'don't know' (DK). In some surveys DK is a valid response, but in the plebiscite everyone votes either 'yes' or 'no' to the issue of independence, either by voting in person or by staying home (effectively a 'no' vote). Hence, the DK responses to the SPO survey question can be viewed as missing or masked data, masking either a 'yes' or 'no' plebiscite response.

Rubin et al. (1995) analyzed the SPO results with the goal of predicting the true percentage of voters in favor of independence. They treated the DK responses as missing data and considered ignorable and nonignorable missing-data mechanisms. Their prediction from a straightforward ignorable model was very accurate, while that from their most realistic nonignorable model was substantially far from the truth. The results are particularly interesting because the predictions were obtained before the true percentage was known and indicate that simple ignorable models can produce better answers than more complex and seemingly more plausible nonignorable models.

Conclusion

Multiple imputation is a general approach for handling the pervasive and often challenging problem of reaching statistically valid inferences from incomplete data (*see* Missing Data). It was derived from a Bayesian perspective but produces inference procedures that have excellent frequentist properties. Generating multiple imputations can be difficult, and fortunately computer software is appearing that aids in this task (see below), but analyzing a multiply-imputed data set was designed to be almost as easy as analyzing a complete data set. A multiple-imputation inference is obtained by applying a complete-data inference procedure to each of the multiple data sets completed by imputation and then combining these estimates using simple combining rules. To learn more about multiple imputation (see Rubin, 1987, 1996; Schafer, 1997).

There is currently only a limited amount of software for generating multiple imputations under multivariate complete-data models and for analyzing multiply-imputed data sets (i.e., completing the data sets, running the complete-data analyses, and combining the complete-data outputs), but the situation appears to be rapidly improving. For generating imputations, software to implement the methodology developed by Schafer (1997) has been written for the S-PLUS (Mathsoft, 2001) statistical package and is freely available on the Internet. This software includes programs for multiple imputation in the contexts of incomplete multivariate normal data, incomplete categorical data, and incomplete data under the general location model allowing both categorical and multivariate normal variables.

A revised and expanded version of this software, including facilities for easily analyzing multiply-imputed data, has been incorporated into recent release of S-PLUS. SOLAS (Statistical

Solutions, 2000), the first commercial product for multiple imputation, combines a graphical user interface and integrated environment for analyzing multiply-imputed data with flexible and efficient methods of generating imputations that exploit the computational and modeling advantages of near monotone missing-data patterns. SAS Institute is developing procedures for analyzing and generating multiply imputed data under a variety of parametric models. These procedures are expected to be available in the very near future. The web site www.multiple-imputation.com contains a fairly complete list of software for multiple imputation and serves as source for news and events associated with multiple imputation.

Bibliography

Barnard, J., Rubin, D.B., 1999. Small-sample degrees of freedom with multiple imputation. Biometrika 86, 948–955.

Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B., Weidman, L., 1991. Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. Journal of the American Statistical Association 86, 68–78.

Gelfand, A.E., Smith, A.F.M., 1990. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85, 398–409.

Glynn, R.J., Laird, N.M., Rubin, D.B., 1993. Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. Journal of the American Statistical Association 88, 984–993.

Graham, John W., Olchowski, Allison E., Gilreath, Tamika D., 2007. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prevention Science 8 (3), 206–213.

Heitjan, D.F., Little, R.J.A., 1991. Multiple imputation for the fatal accident reporting system. Applied Statistics 40, 13–29.

Horton, Nicholas J., Lipsitz, Stuart R., 2001. Multiple imputation in practice: comparison of software packages for regression models with missing variables. The American Statistician 55 (3), 244–254.

Kennickell, A.B., 1997. Multiple imputation and disclosure protection: the case of the 1995 Survey of Consumer Finances. In: Record Linkage – 1997: Proceedings of an International Workshop and Exposition. US Office of Management and Budget, Washington, DC, pp. 248–267.

Kennickell, A.B., 1998. Multiple imputation in the Survey of Consumer Finances. In: ASA Proceedings of the Business and Economic Statistics Section. American Statistical Association, Alexandria, VA, pp. 11–20.

King, Gary, Honaker, James, Joseph, Anne, Scheve, Kenneth, 2001. Analyzing incomplete political science data: an alternative algorithm for multiple imputation. American Political Science Review 95 (1), 49–69.

Mathsoft, 2001. S-PLUS User's Manual, Version 6. StatSci, a Division of Mathsoft Inc., Seattle, WA.

Meng, X.-L., 1994. Multiple-imputation inferences with uncongenial sources of input (Disc: P558–573). Statistical Science 9, 538–558.

Raghunathan, T.E., Rubin, D.B., 1997. Roles for Bayesian techniques in survey sampling. In: Proceedings of the Survey Methods Section of the Statistical Society of Canada. Statistical Society of Canada. McGill University, Montreal, PQ, pp. 51–55.

Royston, Patrick, 2004. Multiple imputation of missing values. Stata Journal 4, 227–241.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–590.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

Rubin, D.B., 1993. Comment on 'statistical disclosure limitation'. Journal of Official Statistics 9, 461–468.

Rubin, D.B., 1996. Multiple imputation after 18+ years (Pkg: P473–520). Journal of the American Statistical Association 91, 473–489.

Rubin, D.B., Schenker, N., 1986. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association 81, 366–374.

Rubin, D.B., Schenker, N., 1987. Interval estimation from multiply-imputed data: a case study using census agriculture industry codes. Journal of Official Statistics 3, 375–387.

Rubin, D.B., Schenker, N., 1991. Multiple imputation in health-care databases: an overview and some applications. Statistics in Medicine 10, 585–598.

Rubin, D.B., Stern, H.S., Vehovar, V., 1995. Handling 'Don't know' survey responses: the case of the Slovenian Plebiscite. Journal of the American Statistical Association 90, 822–828.

Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. Chapman and Hall, New York.

Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A., Rubin, D.B., 1996. The NHANES III multiple imputation project. In: ASA Proceedings of the Section on Survey Research Methods. American Statistical Association, Alexandria, VA, pp. 28–37.

Schenker, N., Treiman, D.J., Weidman, L., 1993. Analyses of public use Decennial Census data with multiply imputed industry and occupation codes. Applied Statistics – Journal of the Royal Statistical Society Series C 42, 545–556.

Statistical Solutions, 2000. SOLAS for missing data analysis 2.0. Cork. Statistical Solution Ltd, Ireland.

Sterne, Jonathan AC., White, Ian R., Carlin, John B., Spratt, Michael, Royston, Patrick, Kenward, Michael G., Wood, Angela M., Carpenter, James R., 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ: British Medical Journal 338, b2393.

Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–540.

Multisensory Integration, Principles of

Benjamin A Rowland, Barry E Stein, and Terrence R Stanford, Wake Forest School of Medicine, Winston-Salem, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by B.E. Stein, M.T. Wallace, and T.R. Stanford, volume 5, pp. 3008–3015, © 2001, Elsevier Ltd.

Abstract

Animals have evolved multiple sensory systems in order to detect and analyze different forms of energy in the environment; for example, light and sound. Perception and behavior are greatly enhanced by combining information across the senses. Although this fusion is seemingly effortless and automatic, its neural implementation is complex, as the originating sensory signals are processed in different areas of the brain with very different operating characteristics. In this article, we describe the fundamentals of multisensory integration as they apply to multiple animal species, focusing on the superior colliculus as a model system for its implementation.

One of the most important tasks the brain must perform is interpreting information relayed from its sensory organs. Both individual survival and evolutionary success depend on the speed and accuracy with which such information is processed and transformed into appropriate behaviors. This process is markedly facilitated by the brain's ability to synthesize information from different sensory systems, and the perceptual consequences of this 'multisensory integration' have intrigued philosophers of science since the time of ancient Greece. But it is only recently that we have come to understand some of the principles of multisensory integration by studying its operation at multiple levels, from its consequences on overt behavior to changes in the activity of individual neurons. These endeavors have been greatly facilitated by the study of a model system: multisensory neurons in the superior colliculus (SC). Individual SC neurons are responsive to stimuli appearing within circumscribed regions of space and produce bursts of impulses that drive the head and eye to orient toward that region. They are arranged according to their spatial sensitivity in a topographic map and collectively comprise an elegantly simple architecture in which a sudden stimulus can be registered and initiate an appropriate response toward it. An individual SC neuron can be responsive to multiple sensory modalities, and its sensory-evoked activity can be markedly enhanced or depressed, depending on the spatial and temporal relationships among the cross-modal stimuli it is responding to and the organization of its own sensitivities. These determinants of multisensory integration at the single-neuron level are closely paralleled at the level of SC-mediated overt attentive and orientation behaviors. The capacity for multisensory integration is not present in the newborn, but develops in a protracted postnatal fashion as the animal develops and gains experience with the cross-modal relationships present in the environment it will inhabit. The cross-species similarities in the principles of multisensory integration suggest a remarkable evolutionary constancy in the fundamentals of this process. Computational methods are now beginning to address whether the principles of multisensory integration detailed in the particular circuits and behaviors that have been explored represent the presence of global principles, or if the brain also utilizes specialized principles that are more closely tied to specific circuits and their associated functions.

Multiple Sensory Representations in the Brain

Across species, brain size is generally correlated with body size. Although a species' actual acumen appears more closely related to its ratio of brain mass to body mass (encephalization), this is an imperfect relationship. All species possess multiple sensory systems, and because each requires a good deal of neural circuitry to make them work, much of even the largest and most encephalized brains is devoted to their operation. Thus, the cerebral cortex, generally considered the highest stage of information processing and the seat of perception, appears to have a large portion of its area organized into a succession of higher-order sensory (and motor) processing stations. Thus far, a minimum of 32 visual, 15 auditory, and 8 somatosensory areas have been identified in the primate cortex. Somehow, though no one really knows exactly how, the coordinated activity of these multiple cortical areas across the neural expanse produces a unified perceptual experience.

That so much of the brain is devoted to sensory processing and its associated motor commands is due to the fact that survival depends on the speed and accuracy with which an organism can evaluate external events and properly react to them. Having many senses is generally better than having few, with the additional benefits outweighing the increased metabolic cost of building and maintaining the necessary underlying brain circuits. For one, different sensory modalities can substitute for one another when they are individually compromised (e.g., sight and touch substitute for vision in the dark). For another, because they sample different forms of energy, multiple senses allow the organism to simultaneously obtain different perspectives or estimates about a singular environmental event. This increases the likelihood that a given event may be correctly detected and identified because two events that appear similar to one sense may be disambiguated by another (e.g., two things that sound the same may look different). Information can be combined across the senses to produce a wholly new perspective, and in cases where each sense

imperfectly estimates a feature of the event, their combination can produce a better overall estimate of its properties.

Multisensory Integration in Perception

By synthesizing information from two or more modality-specific inputs the brain can create something that is entirely new. For example, what we think of as 'taste' is not due solely to the signals the brain receives about the food-induced chemical changes on the tongue. Rather, taste represents a synthesis of signals derived from these chemical changes (gustatory) and those from the aroma (olfactory) and texture (somatosensory) of the food. Integrating this information produces a much richer and more diverse sensory experience than would otherwise be possible. It has been known for some time that multiple sensory inputs must be bound together to produce what we experience as a 'unitary' perceptual experience, and that this binding process can have potent perceptual consequences. Yet, we are largely unaware of its actual operation. Perhaps we do not notice the process because it is the normal mode of brain operation. For example, seeing a speaker's face enhances our ability to perceive what is being said (see Sumby and Pollack, 1954), though we are not conscious that we are combining and integrating individual visual and auditory cues. Nevertheless, we might note that we concentrate particularly hard on the speaker's face in a noisy environment. The consequence is an enhancement of the brain activity underlying speech perception (e.g., see Calvert et al., 1999; Sams et al., 1991; Massaro and Stork, 1998) that is induced by the congruent visual (lip movement) and auditory (spoken words) stimuli. Indeed, visual–auditory integration in speech perception is so seamless that we are quite surprised when we experience the so-called 'McGurk effect' wherein combining the sound of one syllable or word with the lip movements of another produces a perceptual product that differs from each. For example, coupling the sound 'bows' with the lip movements used to from 'goes' leads to the perception of 'doze' or 'those.'

These sorts of cross-modal illusions are not limited to meaning, but can affect strongly perception as well. The 'ventriloquism effect' (Howard and Templeton, 1966) is a particularly powerful, yet common, illusion that alters the apparent location of an event. Voices always seem to be coming from the appropriate characters on the television screen, regardless of their location or movement. Yet, all the sounds actually originate from the same audio speakers. There are many other compelling illusions involving different senses, such as the 'parchment skin' illusion, in which the apparent texture of the skin is changed as a consequence of coupling the sensation of rubbing of the fingers with a grating sound that is played through earphones (Jousmaki and Hari, 1998). The oculogravic and somatogravic illusions (Graybiel, 1952; Clark and Graybiel, 1949) are those in which vestibular cues lead to erroneous percepts of visual orientation and body position, respectively, spatial disorientations well known to aviators for their potentially grave consequences.

These illusions are anomalies, but underscore the fact that the brain normally integrates information from different senses in order to increase its sensitivity to environmental events. In some cases, the process creates a signal that is greater than that predicted by simply summing the responses to the individual component stimuli; an excellent biological example of when the whole is greater than the sum of its parts. To make this work properly (and generally it works quite well), the brain must aggregate only those cross-modal cues that are derived from the same event. Though it may use many stimulus features to make this judgment, the spatial and temporal concordance of cross-modal cues has proven to be a particularly effective criterion (see Stein and Stanford, 2008).

The resultant multisensory integration enhances the detection of an event or object, its identification (Giard and Peronnet, 1999; O'Hare, 1991; Angelaki et al., 2009; Stein et al., 1996), and the probability that the brain will initiate appropriate responses to it (Frens et al., 1998; Stein et al., 1989). It is not surprising then, that it also increases the speed with which those responses are programmed (Rowland et al., 2007a) and manifested (Frens et al., 1995; Goldring et al., 1996; Hughes et al., 1994; Bell et al., 2005). In short, it heightens the brain's sensitivity and reactions to environmental events. It is interesting to note that the physiological processes underlying these changes appear to be specific to cross-modal events; that is, multiple concordant signals derived from the same modality do not generate the changes detailed above, thereby distinguishing multisensory integration from unisensory integration. Given two sets of stimuli with approximate equivalence in efficacy, the cross-modal pair will produce much larger enhancements than the within-modal pair (Alvarado et al., 2007, 2008). This difference can also be observed at the behavioral level (Gingras et al., 2009), and is a likely consequence of the fact that in the former case, the signals represent conditionally independent estimates of an event variable, while in the latter case, the same factors that corrupt one of the signals will also significantly corrupt the other. These observations reinforce the utility behind the evolutionary strategy of developing and maintaining sensory systems that are distinct yet able to interact with one another.

Multisensory Integration and the SC

Multisensory integration is already effected at the initial level of information processing in the central nervous system, the single neuron. These multisensory neurons are found at nearly every level in the central nervous system (see Stein and Meredith, 1993), paralleling the distribution of areas devoted to processing information specific to individual sensory modalities.

As a consequence, higher animals have developed a curious duality: some areas of the brain are specialized for processing information on a sense-by-sense basis, and probably give rise to the peculiar subjective impressions that are so closely bound to one modality or another (e.g., the perceptions of color, pitch, tickle, or itch). Conversely, other areas are specialized for information processing regardless of the specific sense or senses from which that information is derived. These latter areas depend on a convergence of inputs from different senses onto a common set of (multisensory) neurons.

One of the best–known structures in which multisensory convergence occurs is a midbrain structure: the SC. In fact, researchers have leaned heavily on the responses of multisensory SC neurons to develop a model that is helpful

in understanding the general neural properties of multisensory integration, and this has proved to be an excellent strategy for a host of reasons. First, the SC is richly endowed with converging visual, auditory, and somatosensory inputs, and thus has many multisensory neurons that can be studied electrophysiologically. Second, the SC is also known to play an important role in producing motor actions that are guided by sensory stimuli. Specifically, it is involved in generating orienting behaviors to stimuli in contralateral space and, thus, for example, mediates rapid shifts of gaze (i.e., eye and head movements) to objects of interest. These factors make it possible to use the same structure to explore the principles governing multisensory integration at the neuronal and behavioral levels.

A multisensory SC neuron has multiple receptive fields, one for each sensory modality to which it is responsive. A receptive field defines the area of sensory space (e.g., visual space, auditory space, somatosensory space) where a given stimulus will be effective in activating the neuron. For multisensory neurons, their different receptive fields are in spatial register with one another. For example, a visual–auditory neuron with a visual receptive field in central space will be responsive to auditory cues in a roughly overlapping region of auditory space as shown in Figure 1. The spatial register among the modality-specific receptive fields of a multisensory neuron plays a critical role in its synthesis of cross-modal cues. Such register is characteristic of multisensory neurons regardless of where they are found in the nervous system, and is maintained even if, for example, one set of peripheral sensory organs moves. Any eye movement can shift visual space relative to auditory (head-axis) or somatosensory (body-axis) space because the retinal receptor cells have moved (Hartline et al., 1995; Jay and Sparks, 1984; Harrington and Peck, 1998). Rather than disrupt receptive field alignment in SC neurons, movement-induced misalignment of sensory reference frames sets into motion compensatory neural mechanisms that attempt to maintain receptive field register. Thus, a shift in eye position to the right produces a rightward shift in the auditory receptive field of an SC neuron (Harrington and Peck, 1998; Jay and Sparks, 1984). Similarly, there is even evidence that the somatosensory receptive fields of visual–somatosensory neurons show compensatory changes as a consequence of eye movements (Groh and Sparks, 1996). Although the mechanisms underlying these dynamic shifts in receptive fields are not well understood, using the retina as a referent for shifting nonvisual receptive fields in the SC reflects its principal role in generating gaze shifts to targets of any modality, including those that are multisensory. This is not the case everywhere multisensory neurons are found. In regions of premotor cortex, where visual–somatosensory neurons are found and whose activity generates limb rather than eye movements, movements of the limb produce compensatory changes in the location of visual receptive fields (Gentilucci et al., 1983; Graziano et al., 1994). Yet, once again, cross-modal receptive field register is a key element in the maintenance of a modality-independent coordinate frame to properly code movement.

Spatial

The enhanced multisensory response of a neuron depends on stimuli having access to its two (or more) receptive fields simultaneously or within a short window of time. Given the spatially overlapping nature of receptive fields in multisensory neurons, the activating stimuli should derive from the same location in space, such as when they originate from the same event. The consequent multisensory response typically is a train of discharges that contains more impulses than one evoked by either of the modality-specific stimuli. Indeed, if the modality-specific responses are themselves weak, the response to the combined stimulus often exceeds the arithmetic sum of the modality-specific responses (Figure 2).

If, however, the stimuli are derived from different locations (e.g., originate from different events), so that one stimulus falls outside its receptive field and is spatially discordant with the other stimulus, no such response enhancement is produced. In fact, under such circumstances there is a good probability that the opposite effect, response depression, will occur. In such cases, even vigorous visual responses can sometimes be

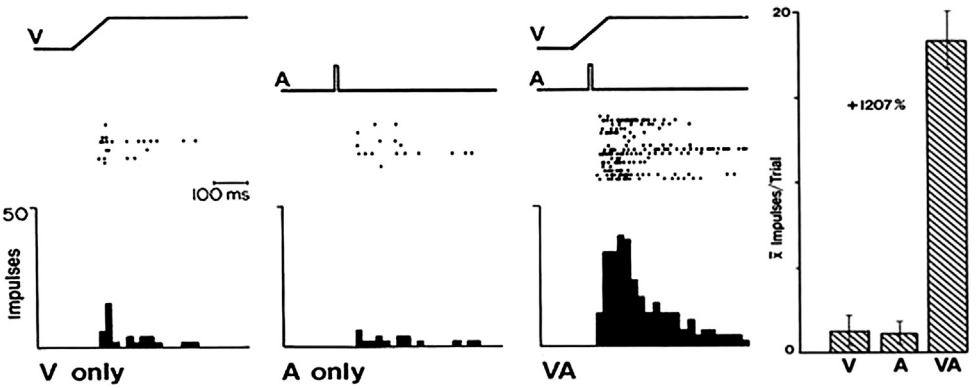

Figure 1 Stimulus traces (top), impulse rasters (middle), and peri-stimulus time histograms illustrating the responses of this neuron to visual (left), auditory (middle), and a spatiotemporally concordant visual–auditory stimulus pair. The multisensory response is clearly elevated over the largest unisensory component response (here, the visual), as illustrated in the summary bar graph on the right. Modified from Meredith, M.A., Nemitz, J.W., Stein, B.E., 1987. Determinants of multisensory integration in superior colliculus neurons: I. Temporal factors. Journal of Neuroscience 10, 3215–3229.

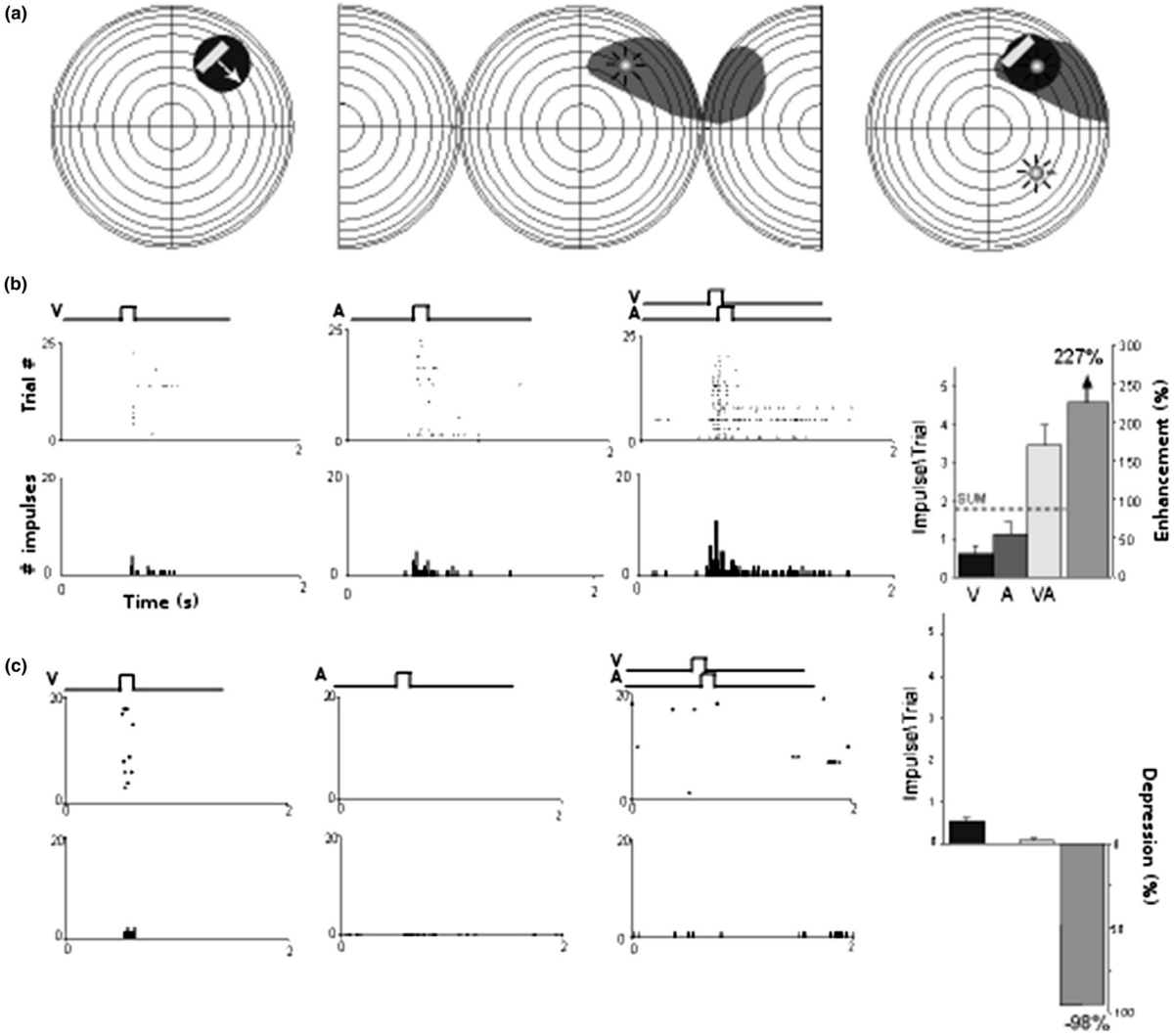

Figure 2 The presentation of two visual–auditory stimulus pairs, that differed only in their spatial coincidence, resulted in very different neural responses. (a) The individual receptive fields of a visual–auditory neuron are in good spatial register. The circles shown here are schematic diagrams of visual (left) and auditory (right) space. The intersection of the horizontal and vertical meridians depicts a point directly in front of the subject, and each concentric circle represents 10°. The left plot shows the location of a visual stimulus (icon) within its receptive field (indicated by shading), the middle plot shows the location of an auditory stimulus within its receptive field (the frontal half of auditory space is shown by the central sphere, and the caudal half is represented by the two hemispheres that have been split and folded forward), and the right plot shows the placement of visual and auditory stimuli within and outside the receptive fields. (b) Multisensory enhancement. Responses to a visual (V only), auditory (A only), and visual–auditory (VA) stimulus placed inside a multisensory neuron's receptive fields. Stimulus traces are illustrated on the top, impulse rasters in the middle, and peri-stimulus time histograms on the bottom. The bar graphs on the right summarize the response magnitudes for each stimulus condition. (c) Multisensory depression (conventions the same as in **Figure 2(b)**). In this example, the auditory stimulus was placed outside of the neuron's receptive field, evoking a depressed response when paired with the visual stimulus.

abolished (Figure 2). These changes at the single neuron level are reflected at the behavioral level as enhanced attentive and orientation responses to spatially coincident cross-modal stimuli, and depressed attentive and orientation responses to spatially disparate cross-modal stimuli (Stein et al., 1989).

These opposing effects of the same stimuli are believed to be due to a neuron receiving synergistic excitatory inputs when its two input channels are stimulated by spatially aligned stimuli (both stimuli fall within the excitatory receptive fields of that neuron), but, because some receptive fields have inhibitory regions at their borders, response depression can be explained by the antagonism between the inhibitory input from the stimulus outside the receptive field and the excitatory input produced by the stimulus within its receptive field (Kadunce et al., 1997).

Temporal

Integrated multisensory responses also depend on the relative timing of the stimuli. When the interval between two different sensory stimuli is too large, they are recognizable in the neuronal response as discrete events. In other words, no integration takes place. Nevertheless, for any given neuron this

temporal 'window' within which they can be integrated may be several hundred milliseconds and the window is generally predictable from the duration of its response to the first stimulus. So, for example, if a somatosensory stimulus elicits a response that lasts 200 ms, a multisensory interaction will occur if the activation produced by a stimulus of a second modality occurs during that period (Meredith et al., 1987; Wallace and Stein, 1997).

This is important because even though different senses may be activated by the same event, the speed with which the modality-specific components activate a multisensory neuron can differ substantially (auditory and somatosensory information reaches SC neurons in 10–25 ms, while visual information requires 40–120 ms). If precise coincidence were required, very few multisensory interactions would take place. Nevertheless, the temporal interaction window is finite and different neurons respond best to different stimulus asynchronies. Generally speaking, the magnitude of the interaction reflects the degree to which the modality-specific activation patterns overlap in time. Thus, the magnitude of the interaction diminishes rapidly as this period of temporal overlap declines (Figure 3). In some cases, temporal intervals that are transitional between those that produce enhanced responses and those that produce two independent responses give rise to response depression. In the case of visual–auditory neurons, the ability to respond differentially to stimuli with different input latencies suggests a tuning to events occurring at different distances from the animal.

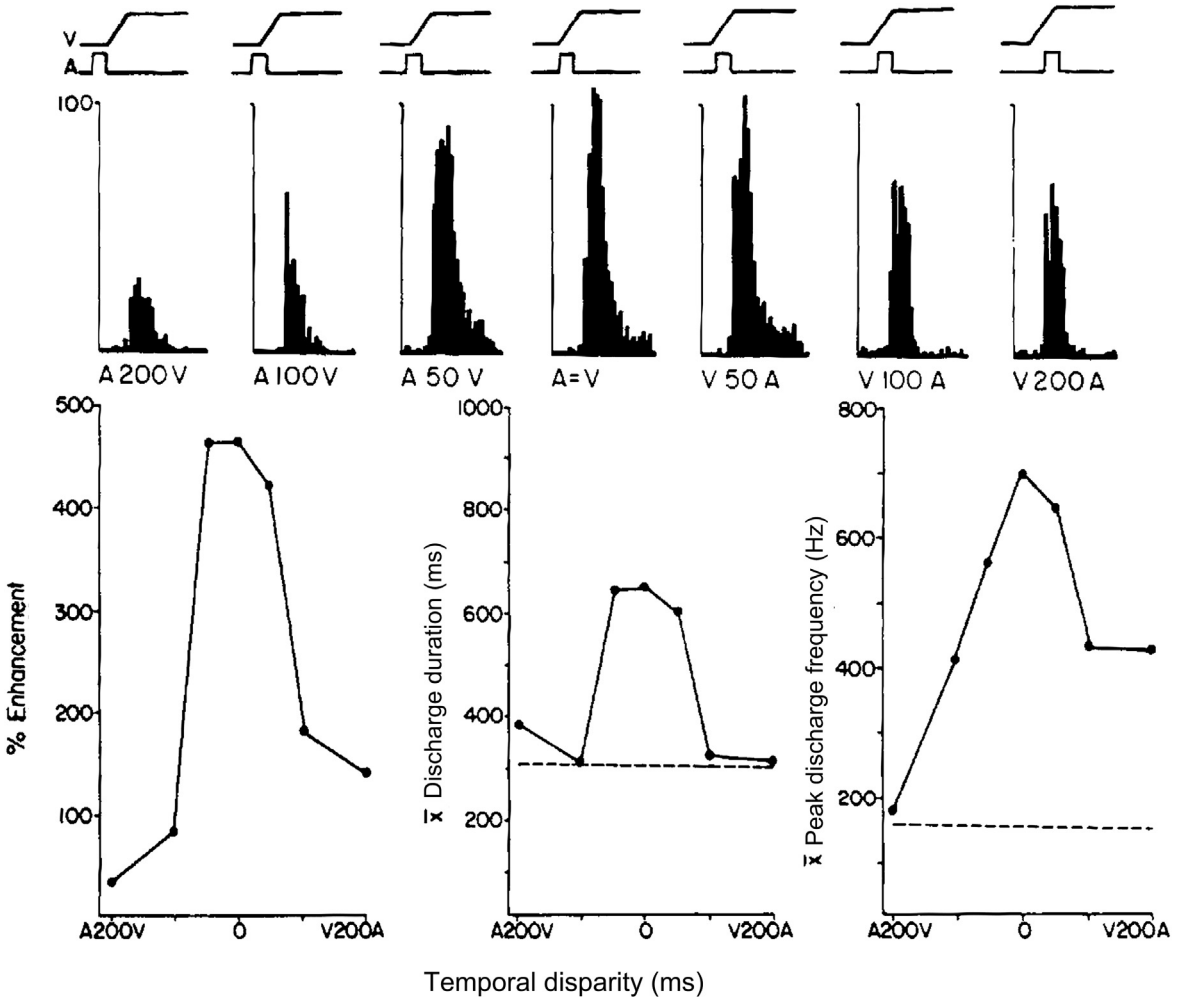

Figure 3 The temporal window of integration for a multisensory interaction in one visual–auditory neuron. The magnitude of the multisensory interaction is plotted here as a function of the interval between the onset of the two stimuli (V = visual, A = auditory). Stimulus traces indicating the relative timing of each stimulus are located above peri-stimulus time histograms of the resulting response. The labels indicate the relative temporal offset between the stimuli; for example, A200V indicates the auditory stimulus onset preceded the visual by 200 ms, while V200A indicates the visual stimulus onset preceded the auditory by 200 ms. Summary statistics at the bottom illustrate % difference between the multisensory and best unisensory response for each temporal offset according to different metrics: response magnitude (% enhancement), mean discharge duration, and mean peak discharge frequency. All metrics show similar trends. In this neuron, simultaneous onset of the visual and auditory stimuli (V = A) produced the greatest multisensory enhancement. Different neurons showed their maximum multisensory response at different cross-modal intervals (generally within 50–100 ms of simultaneity) and had different temporal windows (generally 100–500 ms). Modified from Meredith, M.A., Nemitz, J.W., Stein, B.E., 1987. Determinants of multisensory integration in superior colliculus neurons: I. Temporal factors. Journal of Neuroscience 10, 3215–3229.

Whether there is some systematic organization of these neurons that may be helpful in localizing stimuli in three-dimensional space remains to be determined.

Inverse Effectiveness

The degree to which the response to a cross-modal stimulus exceeds that to each of its modality-specific component stimuli is generally inversely related to the effectiveness of those component stimuli. This makes intuitive sense if we consider the possible functional consequences of multisensory enhancement. By definition, very effective modality-specific stimuli evoke strong neural responses and are very likely to elicit a behavioral response. Thus, if the objective of multisensory enhancement is to increase the probability of detecting and/or responding to an event, there would be little gained from augmenting an already vigorous unisensory response. On the other hand, substantial benefit might be had by combining, and thereby enhancing the influences of weak modality-specific stimuli. The possibility that the system would then be overly susceptible to environmental and/or neural 'noise' is countered by the low probability of random spatial and temporal coincident activation of different sensory channels unless they are linked to the same external event. This logic has been the impetus for many organizations (most prominently, industry and the military) to use this biological system as a model for developing sensor-fusion devices for navigation, air traffic control, and even to search for objects embedded in 'noisy' environments such as the forest or sea floor.

The Essential Circuit for Multisensory Integration in SC Neurons

One might reasonably assume that any SC neuron receiving inputs from two different sensory modalities would have the ability to integrate these inputs. However, this is not the case. Multisensory SC neurons receive converging modality-specific inputs from many sources (see Stein et al., 1993) but cannot integrate their cross-modal inputs in the absence of inputs from a region of association cortex (see Alvarado et al., 2007; Jiang et al., 2001; Jiang and Stein, 2003; Wallace and Stein, 1994). As shown in Figure 4, when the SC is deprived of functional inputs from this region of the cortex (by reversibly deactivating it via cryogenic blockade), SC neurons retain the ability to respond to inputs from different sensory modalities, but lose their ability to integrate them and thereby enhance their

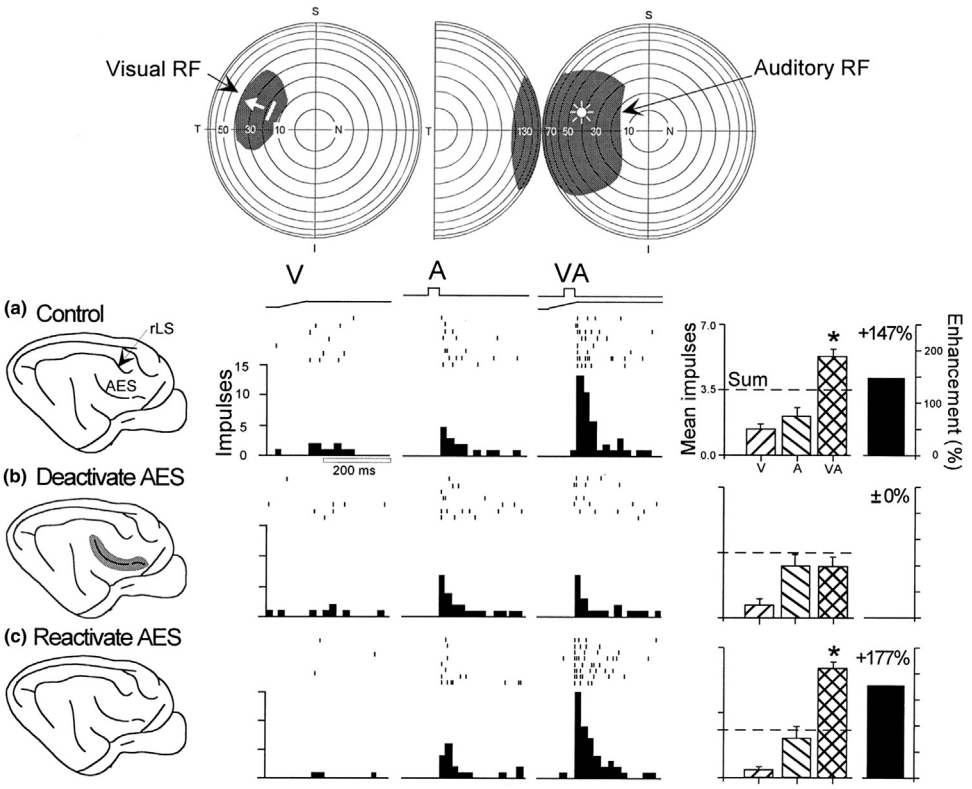

Figure 4 Inputs from association cortex are critical for multisensory integration in the SC. Shown here is a representative example of an SC neuron whose sensory responses were examined before, during, and after temporarily deactivating association cortex by cryogenic blockade (cooling). The top plots show the location of this neuron's visual and auditory receptive fields. (a) Under control conditions, this neuron exhibited a substantial multisensory response enhancement to the visual–auditory stimulus combination. (b) When a region of association cortex (the anterior ectosylvian sulcus, AES) was deactivated, this enhancement was abolished, even though the neuron retained its normal responses to the visual and auditory stimuli when presented individually. (c) Following rewarming of AES, the enhanced multisensory response was reinstated. Modified from Alvarado, J.C., Stanford, T.R., Vaughan, J.W., Stein, B.E., 2007. Cortex mediates multisensory but not unisensory integration in superior colliculus. Journal of Neuroscience 27 (47), 12775–12786.

responses. During cortical deactivation, they respond to combinations of cross-modal stimuli in much the same way as they respond to the most effective of the modality-specific component stimuli.

Presumably, these cortical deactivation-induced physiological changes in SC neurons would disrupt SC-mediated multisensory behaviors. A test of this was made by training animals to respond to visual cues in the presence or absence of auditory stimuli, and then testing them with or without this region of cortex deactivated (Figure 5). When the cortex was temporarily deactivated (with either a local anesthetic injection or cryogenic blockade), there was no change in the animal's responses to the visual cue, but the ability of the auditory cue to enhance (when spatially coincident with the visual) or degrade (when spatially disparate with the visual) these responses was severely disrupted (Jiang and Stein, 2003; Wilkinson et al., 1996). These behavioral observations, coupled with the electrophysiological results described above, demonstrate that the neural circuit in the SC essential for multisensory integration must include inputs from the cerebral cortex.

Development and Plasticity in Multisensory Integration

Although individual neurons in the neonatal brain are responsive to multiple sensory modalities, and in some species even multiple sensory modalities, no species thus far appears capable of multisensory integration at birth. This observation may be derived from the difficulty inherent in developing systems to coordinate activity across sensory systems with unique operational properties and constraints (e.g., time constants and sensitivities), or may reflect an adaptive strategy, whereby the organism postpones instantiating the expected spatial and

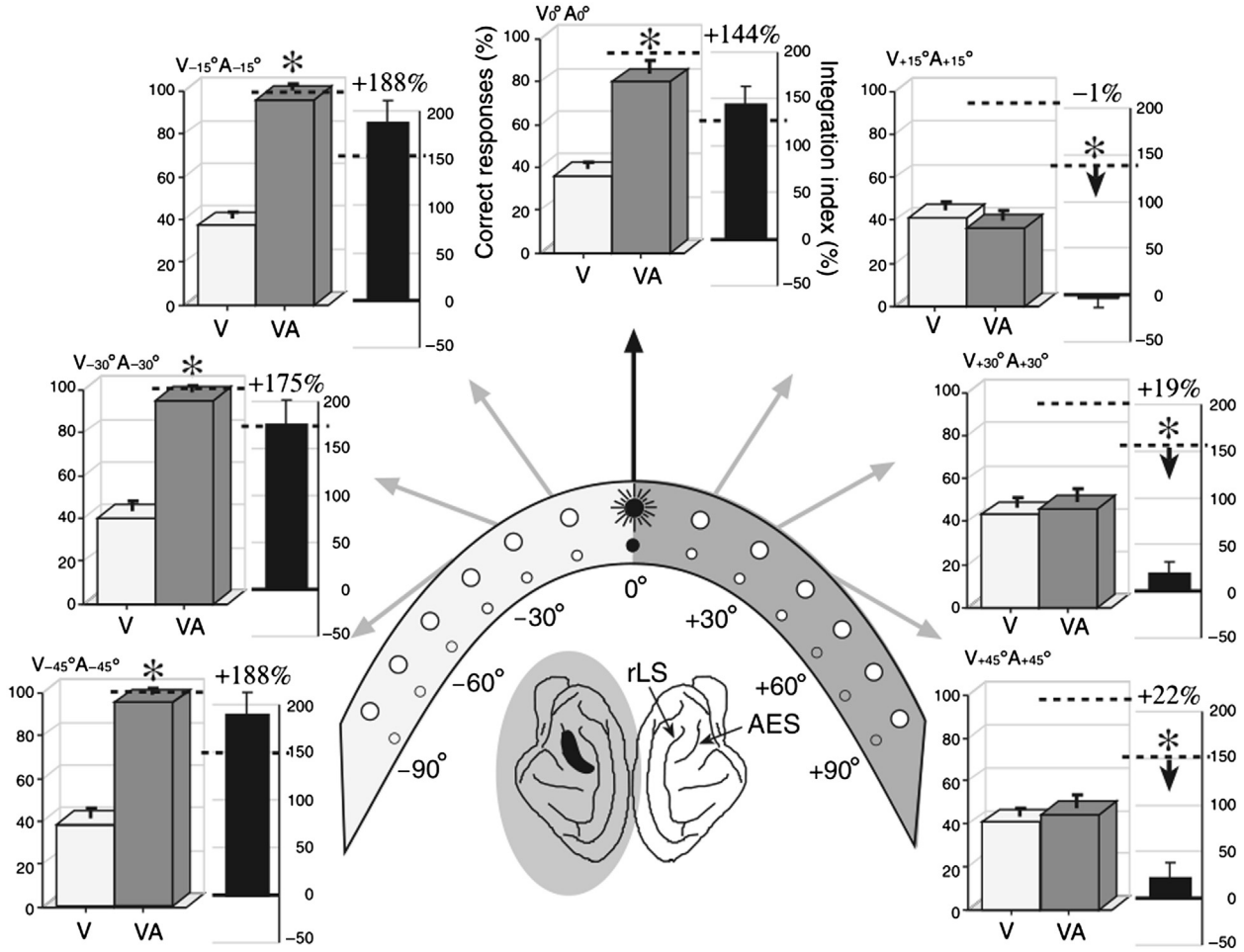

Figure 5 Multisensory orientation behaviors are affected by cortical deactivation as predicted from physiological studies. Three cats were trained in a perimetry device to orient to and approach an illuminated light-emitting diode (LED) for a food reward. During testing, the intensity of each LED was manipulated until each animal responded correctly to that LED on approximately 50% of the trials (white bars, V). The pairing of the auditory (speaker) with the visual stimulus at the same spatial location dramatically enhanced the probability of a correct response (visual + auditory). However, during deactivation of the left AES (shading on brain), these multisensory behavioral response enhancements were compromised in the contralateral hemifield (V + A/AES deactivated). Data are shown for stimulus locations from left to right: 45° left (−30°), 15° left (−15°), directly ahead (0°), 15° right (+15°), and 45° right (+45°). In agreement with the results of physiological studies (see Figure 4), responses to a modality-specific (visual) stimulus alone were unaffected by cortical deactivation (not shown). Modified from Jiang, W., Wallace, M.T., Jiang, H., Vaughan, J.W., Stein, B.E., 2001. Two cortical areas mediate multisensory integration in superior colliculus neurons. Journal of Neurophysiology, 85 (2), 506–522.

temporal relationships between common-sourced cross-modal signals until the unisensory systems themselves become better established and adapted to the environment in which the animal will live postnatally. Whatever the reason, the development of multisensory integration occurs gradually during the postnatal period in normal animals, reaching apparent maturity some time before adolescence. The development of multisensory integration appears to be highly dependent on the experience of the animal with cross-modal cues, such that in the absence of particular cross-modal experiences (e.g., by dark-rearing), neurons do not develop this capacity (Yu et al., 2010; Wallace et al., 2004), and given extensive experience with cross-modal cues having anomalous spatial or temporal relationships, integrate anomalously configured cues as if they belonged to the same event (Yu et al., 2009; Wallace and Stein, 2007).

However, plasticity in multisensory integration does not appear to be strictly limited to a particular developmental time window. Recent evidence suggests that extensive experience with cross-modal cue configurations in the adult can drive similar adaptive changes resulting in altered principles of multisensory integration, both in humans (e.g., Powers et al., 2009) and cats (Yu et al., 2009). In fact, animals deprived of early experience and developing without the capacity for multisensory integration can acquire these capabilities later in life with appropriate spatiotemporal cross-modal experience (Yu et al., 2010). These observations suggest a mechanism by which animals might continuously adjust the way in which they process cross-modal cues to adapt to environmental circumstances as they change.

Multisensory Integration among Different Species and Structures

Although many of the specifics that were presented earlier were drawn from experiments in cats, most of the sensory features that characterize the SC of this animal are also found in every other laboratory species that has been examined thus far. Particularly important in the current context is the spatial register among the different receptive fields of multisensory neurons. This has been found in rodents, carnivores, and primates. The widespread nature of this scheme for coupling inputs from different sensory modalities, and its presence in neurons of the nonmammalian homologue of the SC, the optic tectum (see Stein and Meredith, 1993), suggest that it was preserved during the transition from nonmammalian to mammalian forms and is adaptable to a wide variety of ecological circumstances.

The utility of multisensory integration is not restricted to neurons involved in SC-mediated behaviors, as it is also characteristic of multisensory neurons in many neocortical areas (e.g., Bruce et al., 1981; Duhamel et al., 1991; Fogassi et al., 1996; Graziano et al., 1997; Gu et al., 2008; Ramachandran et al., 1993; Rizzolatti et al., 1981a, 1981b; Stein et al., 1993; Wallace et al., 1992), where such neurons are likely to be involved in functional roles very different from those involving the SC. Although many of the basic principles of multisensory integration (e.g., spatial and temporal congruence) may transcend structure, function, and species, the computational rules that govern its expression are likely to depend on the specific function to be carried out (Stein and Stanford, 2008). For example, multisensory enhancement, a neural correlate of stimulus salience, may not be prevalent, or even advantageous, when multiple modalities of information are combined for the purpose of perceptual evaluation, as for example, when visual and vestibular cues are combined for the purpose of estimating one's movement with respect to the environment (see Angelaki et al., 2009; Stanford, 2012). Future studies will be necessary to reveal the full scope of multisensory representation within the computations that neurons perform.

Computational Approaches to Multisensory Integration

There is substantial interest in developing theoretical frameworks to facilitate our understanding of the similarities and differences that are noted in the operation of multisensory integration in different circuits and species. Computational modeling provides a set of methods on which these links can be founded, in two ways: (a) offering rigorous quantitative formulations of empirical findings, and (b) translating behavioral and physiological observations into a common mathematical language. Recent models of multisensory integration operating at the physiological (Patton and Anastasio, 2003; Knill and Pouget, 2004; Rowland et al., 2007b; Cuppini et al., 2010; Ohshiro et al., 2011) and behavioral (Ernst and Banks, 2002; Colonius and Deiderich, 2008) levels provide insight into how basic computational mechanisms can be used to explain large collections of empirical results. Simultaneously, efforts illustrate how direct links between the neurophysiology and behavioral expressions of multisensory integration can be established through a common mathematical frame of reference (Angelaki et al., 2009). These efforts indicate the inherent power of the computational approach, which will be critical to determining if there exists a common set of principles by which information is pooled across the senses, or whether the brain has developed mechanisms specific to circuits intimately tied to their functional role for this purpose. Ultimately, the goal is to determine whether or not there is some fundamental aspect of the combination of information across the senses that supports a single global theory, or whether there are many.

See also: Neural Synchrony as a Binding Mechanism; Olfactory System; Taste and Smell, Psychology of; Vestibular System; Vision, High-Level Theory of.

Bibliography

Alvarado, J.C., Rowland, B.A., Stanford, T.R., Stein, B.E., 2008. A neural network model of multisensory integration also accounts for unisensory integration in superior colliculus. Brain Research 1242, 13–23.

Alvarado, J.C., Stanford, T.R., Vaughan, J.W., Stein, B.E., 2007. Cortex mediates multisensory but not unisensory integration in superior colliculus. Journal of Neuroscience 27 (47), 12775–12786.

Angelaki, D.E., Gu, Y., DeAngelis, G.C., 2009. Multisensory integration: psychophysics, neurophysiology, and computation. Current Opinion in Neurobiology 19, 452–458.

Bell, A.H., Meredith, M.A., Van Opstal, A.J., Munoz, D.P., 2005. Crossmodal integration in the primate superior colliculus underlying the preparation and initiation of saccadic eye movements. Journal of Neurophysiology 93, 3659–3673.

Bruce, C., Desimone, R., Gross, C.G., 1981. Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. Journal of Neurophysiology 46, 369–384.

Calvert, G.A., Brammer, M.J., Bullmore, E.T., Campbell, R., Iversen, S.D., David, A.S., 1999. Response amplification on sensory-specific cortices during crossmodal binding. NeuroReport 10, 2619–2623.

Clark, B., Graybiel, A., 1949. Linear acceleration and deceleration as factors influencing nonvisual orientation during flight. Journal of Aviation Medicine 20, 92–101.

Cuppini, C., Ursino, M., Magosso, E., Rowland, B.A., Stein, B.E., 2010. An emergent model of multisensory integration in superior colliculus neurons. Frontiers in Integrative Neuroscience 4, 6.

Diederich, A., Colonius, H., 2008. Crossmodal interaction in saccadic reaction time: separating multisensory from warning effects in the time window of integration model. Experimental Brain Research 186, 1–22.

Duhamel, J.-R., Colby, C.L., Goldberg, M.E., 1991. Congruent representations of visual and somatosensory space in single neurons of monkey ventral intraparietal cortex (area VIP). In: Paillard, J. (Ed.), Brain and Space. Oxford University Press, New York.

Ernst, M.O., Banks, M.S., 2002. Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415, 429–433.

Fogassi, L., Gallese, V., Fadiga, L., Luppino, G., Matelli, M., Rizzolatti, G., 1996. Coding of peripersonal space in inferior premotor cortex (area F4). Journal of Neurophysiology 76, 141–157.

Frens, M.A., Van Opstal, A.J., Van der Willigen, R.F., 1995. Spatial and temporal factors determine auditory–visual interactions in human saccadic eye movements. Perception and Psychophysics 57, 802–816.

Gentilucci, M., Scandolara, C., Pigarev, I.N., Rizzolatti, G., 1983. Visual responses in the postarcuate cortex (area 6) of the monkey that are independent of eye position. Experimental Brain Research 50, 464–468.

Giard, M.H., Peronnet, F., 1999. Auditory–visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. Journal of Cognitive Neuroscience 11, 473–490.

Gingras, G., Rowland, B.A., Stein, B.E., 2009. The differing impact of multisensory and unisensory integration on behavior. Journal of Neuroscience 29 (15), 4897–4902.

Goldring, J.E., Dorris, M.C., Corneil, B.D., Ballantyne, P.A., Munoz, D.P., 1996. Combined eye-head gaze shifts to visual and auditory targets in humans. Experimental Brain Research 111, 68–78.

Graybiel, A., 1952. Oculogravic illusion. American Medical Association Archives of Ophthalmology 48, 605–615.

Graziano, M.S., Hu, T., Gross, C.G., 1997. Visuospatial properties of ventral premotor cortex. Journal of Neurophysiology 77, 2268–2292.

Graziano, M.S., Yap, G.S., Gross, C.G., 1994. Coding of visual space by premotor neurons. Science 266, 1054–1057.

Groh, J.M., Sparks, D.L., 1996. Saccades to somatosensory targets. III. Eye-position-dependent somatosensory activity in primate superior colliculus. Journal of Neurophysiology 75, 439–453.

Gu, Y., Angelaki, D.E., DeAngelis, G.C., 2008. Neural correlates of multisensory cue integration in macaque MSTd. Nature Neuroscience 11, 1201–1210.

Harrington, L.K., Peck, C.K., 1998. Spatial disparity affects visual-auditory interactions in human sensorimotor processing. Experimental Brain Research 122, 247–252.

Hartline, P.H., Pandey Vimal, R.L., King, A.J., Kurylo, D.D., Northmore, D.P.M., 1995. Effects of eye position on auditory localization and neural representation of space in superior colliculus of cats. Experimental Brain Research 104, 402–408.

Howard, I.P., Templeton, W.B., 1966. Human Spatial Orientation. Wiley, London.

Hughes, H.C., Reuter-Lorenz, P.A., Nozawa, G., Fendrich, R., 1994. Visual-auditory interactions in sensorimotor processing: saccades versus manual responses. Journal of Experimental Psychology: Human Perception and Performance 20, 131–153.

Jay, M.F., Sparks, D.L., 1984. Auditory receptive fields in primate superior colliculus shift with changes in eye position. Nature 309, 345–347.

Jiang, W., Stein, B.E., 2003. Cortex controls multisensory depression in superior colliculus. Journal of Neurophysiology 90 (4), 2123–2135.

Jiang, W., Wallace, M.T., Jiang, H., Vaughan, J.W., Stein, B.E., 2001. Two cortical areas mediate multisensory integration in superior colliculus neurons. Journal of Neurophysiology 85 (2), 506–522.

Jousmaki, V., Hari, R., 1998. Parchment-skin illusion: sound-biased touch. Current Biology 8, R190.

Kadunce, D.C., Vaughan, J.W., Wallace, M.T., Benedek, G., Stein, B.E., 1997. Mechanisms of within- and cross-modality suppression in the superior colliculus. Journal of Neurophysiology 78, 2834–2847.

Knill, D.C., Pouget, A., 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends in Neurosciences 27 (12), 712–719.

Massaro, D.W., Stork, D.G., 1998. Sensory integration and speechreading by humans and machines. American Scientist 86, 236–244.

Meredith, M.A., Nemitz, J.W., Stein, B.E., 1987. Determinants of multisensory integration in superior colliculus neurons: I. Temporal factors. Journal of Neuroscience 10, 3215–3229.

O'Hare, J.J., 1991. Perceptual integration. Journal of the Washington Academy of Science 81, 44–59.

Ohshiro, T., Angelaki, D.E., DeAngelis, G.C., 2011. A normalization model of multisensory integration. Nature Neuroscience 14 (6), 775–782.

Patton, P.E., Anastasio, T.J., 2003. Modeling cross-modal enhancement and modality-specific suppression in multisensory neurons. Neural Computation 15 (4), 783–810.

Powers III, A.R., Hillock, A.R., Wallace, M.T., 2009. Perceptual training narrows the temporal window of multisensory binding. Journal of Neuroscience 29 (39), 12265–12274.

Ramachandran, R., Wallace, M.T., Stein, B.E., 1993. Distribution and properties of multisensory neurons in rat cerebral cortex. Society for Neuroscience Abstracts 19, 1447.

Rizzolatti, G., Scandolara, C., Matelli, M., Gentilucci, M., 1981a. Afferent properties of periarcuate neurons in macaque monkeys. I. Somatosensory responses. Behavioural Brain Research 2, 125–146.

Rizzolatti, G., Scandolara, C., Matelli, M., Gentilucci, M., 1981b. Afferent properties of periarcuate neurons in macaque monkeys. II. Visual responses. Behavioural Brain Research 2, b47–b63.

Rowland, B.A., Quessy, S., Stanford, T.R., Stein, B.E., 2007a. Multisensory integration shortens physiological response latencies. Journal of Neuroscience 27 (22), 5879–5884.

Rowland, B.A., Stanford, T.R., Stein, B.E., 2007b. A model of the neural mechanisms underlying multisensory integration in the superior colliculus. Perception 36 (10), 1431–1443.

Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O.V., Lu, S.-T., Simola, J., 1991. Seeing speech: visual information from lip movements modified activity in the human auditory cortex. Neuroscience Letters 127, 141–145.

Stanford, T.R., 2012. Multisensory perspectives on perceptual decision making and perceptual learning. In: Stein, B.E. (Ed.), The New Handbook of Multisensory Integration. MIT Press, Cambridge, MA, pp. 469–474.

Stein, B.E., London, N., Wilkinson, L.K., Price, D.D., 1996. Enhancement of perceived visual intensity by auditory stimuli: a psychophysical analysis. Journal of Cognitive Neuroscience 8, 497–506.

Stein, B.E., Meredith, M.A., 1993. The Merging of the Senses. MIT Press, Cambridge, MA.

Stein, B.E., Meredith, M.A., Honeycutt, W.S., McDade, L., 1989. Behavioral indices of multisensory integration: orientation of visual cues is affected by auditory stimuli. Journal of Cognitive Neuroscience 1, 12–24.

Stein, B.E., Meredith, M.A., Wallace, M.T., 1993. The visually-responsive neuron and beyond: multisensory integration in cat and monkey. Progress in Brain Research 95, 79–90.

Stein, B.E., Stanford, T.R., 2008. Multisensory integration: current issues from the perspective of the single neuron. Nature Reviews Neuroscience 9 (4), 255–266.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America 26, 212–215.

Wallace, M.T., Meredith, M.A., Stein, B.E., 1992. Integration of multiple sensory modalities in cat cortex. Experimental Brain Research 91, 484–488.

Wallace, M.T., Perrault Jr., T.J., Hairston, W.D., Stein, B.E., 2004. Visual experience is necessary for the development of multisensory integration. Journal of Neuroscience 24, 9580–9584.

Wallace, M.T., Stein, B.E., 1994. Cross-modal synthesis depends on input from cortex. Journal of Neurophysiology 71, 429–432.

Wallace, M.T., Stein, B.E., 1997. Development of multisensory neurons and multisensory integration in cat superior colliculus. Journal of Neuroscience 17, 2429–2444.

Wallace, M.T., Stein, B.E., 2007. Early experience determines how the senses will interact. Journal of Neurophysiology 97 (1), 921–926.

Wilkinson, L.K., Meredith, M.A., Stein, B.E., 1996. The role of anterior ectosylvian cortex in cross-modality orientation and approach behavior. Experimental Brain Research 112, 1–10.

Yu, L., Rowland, B.A., Stein, B.E., 2010. Initiating the development of multisensory integration by manipulating sensory experience. Journal of Neuroscience 30 (14), 4904–4913.

Yu, L., Stein, B.E., Rowland, B.A., 2009. Adult plasticity in multisensory neurons: short-term experience-dependent changes in the superior colliculus. Journal of Neuroscience 29 (50), 15910–15922.

Multisited Field Studies

Mark-Anthony Falzon, Clare Hall, Cambridge, UK; and University of Malta, Msida, Malta

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the last couple of decades, multisited field studies (usually in their 'multisited ethnography' incarnation) have emerged as a methodological choice that favors a spatially dispersed field – through the use of actual multiple sites and/or the juxtaposition of data – through which social scientists 'follow' people, goods, information, and such. The knowledge that is produced is said to better reflect contemporary thought on space, place, scale, and boundaries, and therefore social formations that are caught up in contexts of transnationalism and globalization. The idea has spawned a considerable methodological–theoretical literature as well as a number of empirical case studies. It has also come under significant criticism, mainly from arguments of holism, lack of depth, and abdication of ethnographic responsibility. These critiques have led to a self-examining 'second-generation' thought and method and multisited studies show no signs of being forgotten.

'Multisited field studies' can mean many things. In the social sciences, the term usually refers to a type of grounded research based on ethnographic principles and techniques. In this standard incarnation, it is known as 'multisited ethnography,' which means that while the two terms are broadly interchangeable, most of the productive theoretical and empirical contributions hinge on multisitedness as applied to ethnography specifically.

Ethnography is an eclectic methodological choice that privileges an engaged, contextually rich, and nuanced type of qualitative social research, in which fine-grained daily interactions constitute the lifeblood of the data produced. With respect to method, it entails the situational combination of field techniques (note taking, audio/visual recording, interviews, examination of indigenous literature, observation, and such) rooted in the ideal of participant observation (to live, to some extent, as the 'natives' themselves do), itself based on relations of trust and a belief that data are produced in and of 'thick' interaction between researcher/s and researched.

Conventionally, ethnography has involved the idea – if not necessarily the practice – of a relatively long-term (typically several months upward) stay in a field site of choice. The site was understood to be the container of a particular set of social relations, which could be studied and possibly compared with the contents of other containers elsewhere. To some extent, the contents might also be generalized into area, regional, or, most optimistically, universal knowledge.

Multisited ethnography purports to break with this convention. The standard reformative thesis was nailed by George E. Marcus to the door of the *Annual Review of Anthropology* (1995). Telegraphically, Marcus argued that multisited ethnography defines as its objective the study of social phenomena that cannot be accounted for by focusing on a single site. Previously, the 'world system' was seen as a framework within which the local was contextualized or compared; it now becomes integral to and embedded in multisited objects of study. The essence of multisited research is to *follow* (a crucial and consequential word in this case) people, connections, associations, and relationships across space, because they are substantially continuous but spatially noncontiguous. Research proceeds by a series of juxtapositions in which the global is collapsed into and made an integral part of parallel, related local situations, rather than something monolithic or external to them. In terms of method, multisited ethnography involves a spatially dispersed field through which the ethnographer moves – actually, via sojourns in two or more places, or conceptually, by means of techniques of juxtaposition of data.

Since 1995, Marcus – sometimes in collaboration with other scholars – has pursued an ongoing project of 'refunctioning ethnography' (Holmes and Marcus, 2004, 2005). The significant number of sophisticated articulations of what multisited ethnography might actually mean, in theory and practice (see, for instance, Hannerz, 2003; Gustavson and Cytrynbaum, 2003), can also be read as developments on the theme. The consensus is that much new ground has been covered since the 'first-generation' multisited ethnography. This openness to self-renewal and self-criticism, spearheaded in many ways by Marcus himself, is what has kept the idea so full of life and potential.

Predictably, reactions to Marcus's program have been mixed. On the one hand, there are those for whom the idea is so much old wine in new bottles. On the other, it has fired the spatial imagination – and enlarged the carbon footprint – of a generation of social scientists, as the converts used it to study things as diverse as global environmentalism (Gatt, 2011), climate change (Krauss, 2009), human immunodeficiency virus/acquired immunodeficiency syndrome (Whyte et al., 2011), and even a type of gourmet mushroom (Choy et al., 2009). Methodologically too, multisited ethnography has been the subject of a great many scholarly contributions; these include at least two edited collections (Falzon, 2009; Coleman and von Hellermann, 2011).

Empirically, the most productive venue has probably been that of mobility studies ('follow the people,' as Marcus would put it). Given the extent to which scholars in the field have largely eschewed 'methodological nationalism' in favor of transnational understandings, this is hardly surprising. As Amelina and Faist put it, the transnational approach "rejects the nation-state as the sole starting point of empirical analysis … [and] … calls for a denaturalization of categories such as 'nation' and 'space'" (2012: p. 2). What that means in practice is that migration scholars increasingly find it uncomfortable to root their research – physically and/or conceptually – in a single site. Amelina and Faist go on to describe how

multisited ethnography has also inspired the 'mobile methods' ethnographic approach, an itinerant strategy that recommends collecting data by 'observing people's mobility,' 'walking with,' 'stalking,' or 'lurking' around others. This supposedly produces an intimate knowledge of, and enters into a dialectical relationship with, geographical and/or virtual mobility. According to this position, transnational scholars would need to go beyond *sui generis* models of state responses to mobility and look at the practices of (potential) migrants, which may result in mobility, immobility, or a combination of both.

Somewhat cheerfully, the idea also seems to have spilled over into disciplines other than anthropology and sociology. 'Multisited archaeology,' for example, is "adapted from multi-sited ethnography to serve archaeologists' interests in comparing locales, multi-scalar connections, and local-global relations" (Ryzewski, 2012: p. 241).

It must be said that déjà vu is not the only finger that has been pointed at multisited ethnography. In the last few years, especially, one has also seen a number of sustained and serious critiques of Marcus's original formulation and its elaborations. Hage (2005), for instance, who mischievously suggests that multisited imaginings may well be a symptom of delusions of innovativeness if not grandeur, argues that with respect to studying, say, migration, the concept of a single geographically discontinuous site is much more useful than that of multisitedness. He also suggests that multisited research may imply a tacit holism, and proposes that a "certain reflexivity concerning the social relations that one is opting not to cover in depth" (Hage, 2005: p. 466) makes for a better definition of one's partiality. His conclusion is startling: "I simply do not think that there can be such a thing as a multi-sited ethnography" (Hage, 2005: p. 465). For Hage, multisited ethnography is a buzzword, since "its signification and ramifications are [not] explored by many of its users … [who] use it mechanically" (Hage, 2005: p. 464).

A second significant recent critique of the multisited imaginary is that by Matei Candea (2007). Like Hage, Candea targets in particular what he sees as a born-again holism implicit (and sometimes explicit) in Marcus's and subsequent formulations. He argues that ethnography is really about setting up 'arbitrary locations' – which in any case it invariably does, in the sense that a landmark work like Garsten's *Apple World* (1994) is actually Apple 'places,' since her ethnography of the multinational company was based on fieldwork in a *limited* number of locations *chosen by herself*. Candea grants that the multisited program has probably served to broaden the range of topics considered suitable for ethnographic study; however, he posits that Marcus's device of 'following,' for instance, could be applied equally well in a local, arbitrary setting – which is what ethnographers have pretty much always done. What the more naïve advocates of multisitedness see as 'incompleteness,' Candea sees as a self-critical methodological decision ('making the cut'), which one 'reflects upon and takes responsibility for' (Candea, 2007: p. 174). Ethnographers need to be more cautious of the seductions of 'limitless narrative possibilities' – which are deceptive in any case – opting instead for 'sensibilities based on self-imposed restrictions' (Candea, 2007: p. 168). Perhaps most importantly, Candea reminds us that siting is, as Coleman and von Hellermann put it, "an epistemological, methodological, and presumably political practice" (2011: p. 7).

A third line of critical thought takes to task Marcus's prescription for field-workers to 'follow.' The problem is that 'following' implies cartographies and trajectories that precede, and would exist quite happily independently of, the fieldwork project (see, for instance, Fairhead, 2011). In this sense, 'following' comes dangerously close to being a passive act and the field to being essentialized and detached from the field-worker – a far cry from say Dilley's (1999) convincing discussion on ethnographic context as a performative act. Recent contributions have tended to get around this by replacing 'following' with terms that are active and agentive. Coleman and von Hellermann (2011), for example, emphasize that there has been a conceptual shift from 'following' to 'collaborating': "We might be tempted to use the term 'fieldmaker' here, but we need to bear in mind that fields are always made, are never 'natural […]'" (p. 3). In fairness, the shift is evident in Marcus's own methodological thought and practice: A work like *Ocasião: The Marquis and the Anthropologist, a Collaboration* (Marcus and Mascarenhas, 2005) is true to its subtitle and indeed breaks new ground in that direction. Besides, once the substitution is made in favor of collaboration, multisited ethnography can quite easily deal with critiques from arbitrary decision making.

In view of these and other critiques, it may be worth remembering that there is a sense in which every fieldwork project is an experiment in siting the object. Rather like the Weberian ideal type (see, for instance, Freund, 1968), an analytical device by means of which the social scientist sets up the boundaries of a definition and proceeds to work within it (thus opening up infinite possibilities of ideal-typic definition and therefore analytical trajectories), any exercise in siting is always one of an infinity of possibilities. Clearly, the process should be based on more than whim; rather, siting proceeds within a triangular space set up by one's preunderstanding (itself a product of the current scholarly literature among other things), the developing researcher-researched collaborative exchange, and – last but not least – scholarly creative choices that are invariably circumscribed by the availability of resources. There are many reasons why this model of siting is useful. First, and pace Hage, it takes multisitedness to be one of a number of possibilities, rather than a messianic calling or a first among equals; this is both realistic and prudent. Second, it allows the notions of arbitrariness, choice, and creative scholarly enterprise happily to coexist with multisitedness in ways that transcend wooden models of 'following.' Third, it accommodates the recent thought on ethnography (including its siting) as a form of active and ongoing collaboration in the field.

The Methodological and Intellectual Background

In order to try to take stock of multisited ethnography, it makes sense to outline some of the reasons why the idea emerged when it did, and why it is thought to be so apt (according to some, necessary) an approach to contemporary research. The question may be phrased as, Why did the localizing strategies

of ethnography come into question, and why in the late twentieth century? Briefly, the author thinks that there are two main reasons.

The first has to do with the notion, and its ethnographic consequences, that space is socially produced. Lefebvre (1991 [1974]) was the first to flesh it out in a sustained way, the story goes, although Foucault was foresighted enough to sense, years earlier, that the 'spatial turn' was round the corner (see Soja, 1989). Definitely by the 1990s, space was all over the social sciences. As summed up by Massey (2005: p. 9), contemporary sensitivity to issues of space rests on three propositions:

> *First*, that we recognise space as the product of interrelations; as constituted through interactions, from the immensity of the global to the intimately tiny … *Second*, that we understand space as the sphere of the possibility of the existence of multiplicity in the sense of contemporaneous plurality; as the sphere in which distinct trajectories coexist; as the sphere therefore of coexisting heterogeneity … *Third*, that we recognise space as always under construction.

There is of course, in ethnography as well as in other social scientific circles, a long tradition of looking at the relation between people and spaces – at the making of places, in other words. What is more recent is the reflection on the methodological corollary of this relation. By 1986, Salzman (1986: p. 528) was asking "Is traditional fieldwork outmoded?", and suggesting among other things research-team models of ethnographic practice. (The suggestion has been taken up by scholars who have explored, often through their own empirical work, some of the ways in which multisited collaborative ethnography involving teams of people might work – see, for instance, Choy et al., 2009; FitzGerald, 2012.) Ten years later, Fog Olwig and Hastrup suggested that "the methodological implications of this insight [that space is socially produced] are still being worked out" (1996a,b: p. 1, the author's parenthesis). The upshot is that social science has had to come to terms with the idea that, logically, if space is produced, there is no reason why the spaces of field studies should be exempt. Marcus's plea may be seen as an attempt to prescribe a methodological antidote to these anxieties.

Aside, the 'spatial turn' has affected fields beyond the social sciences. In literary criticism, for example, recent years have seen a spate of works dealing with the spatialization of the text. Davidson (2007), for example, links free verse, and the way that the shape on the page is produced by the poem, to the Lefebvrian concept that space is not prior to but produced by human activity; Huang (2006) looks at 'spatial negotiation' in Asian-American fiction; and Michelucci (2002) draws on the works of D.H. Lawrence to discuss place as a culturally constructed category that exists in relation to space as a physical and philosophical one.

The second set of forces that inspired Marcus, and that lend allure to multisited ethnography, may analytically be separated into two types: first, the debatable (and debated) idea that contemporary societies are invariably, inevitably, and self-evidently located within larger wholes (see Cook et al., 2009); second, the seemingly obvious corollary that within these wholes, people, information, goods, and ideas are in a constant state of displacement.

There is a sense in which a cautious analytical holism seems to be at the heart of the ethnographic approach. At the same time as ethnography is about the particular, it is also thought to give, as Gay y Blasco and Wardle (2007: p. 43) put it, "further contextual meaning to particular lives by demonstrating their integration within more inclusive social forms." Applying this to space/place, the question of what to do with the local has a long pedigree in ethnographic methodology. For example, during the latter half of the twentieth century, anthropologists – in particular those studying seemingly bounded peasant villages, or 'immigrant communities' – became increasingly uncomfortable with the conventional idea that the local was an adequate form of ethnographic space (Falzon, 2005). Even so, as recently as 2007, Russell (who went on to do multisited fieldwork with the Yakkha people in Tamaphok, Nepal, and various migrant destinations in India and elsewhere) could complain that migrants are people one says 'hello' or 'goodbye' to in their destinations and places of origin, respectively, "but rarely with knowledge … of the travels and travails in between" (2007: p. 362).

The grandest themes of late-twentieth century and contemporary social science seem to revolve around a problematic relation to the local and the search for some larger scale of analysis, and the study of connections between places. As Mintz (1998: p. 117) puts it, "(t)he new anthropology is many things, among them the study of human groups *in motion*. That motion is thought to be more than international; it is *trans*national." World systems theory, transnationalism, migration studies that go beyond classical push–pull and/or integration concerns, diasporas, cosmopolitanism, and so forth – all posit frameworks and scales that invite supralocal understanding and therefore methodology. The most prominent interpretive framework is of course that of globalization, which as a paradigm patterns much of the contemporary thought about people and places. A host of social formations are nowadays seen as being formed and reformed in and of 'global assemblages,' in Ong and Collier's words (2005).

One should add that recent formulations of globalization have moved well away from both the 'global village' model and its less banal if more insidious cousin, namely, the idea that global interconnectedness coexists with local variability. As Massey (2005: p. 88) puts it, "'spatialising globalisation' means recognising crucial characteristics of the spatial: its multiplicity, its openness, the fact that it is not reducible to a 'surface', its integral relation with temporality." The shift is relevant because it raises the cardinal point that while multisited ethnography may be many things, what it certainly is not (or should not be) is an approach that devalues locality and place. On the contrary, multisitedness properly done has the potential to produce a knowledge that highlights the specifics of place and local experience and the hierarchies that these exist within, and that in turn provides the basis for an informed critique of place and location. In other words, multisitedness assumes that place still matters. The alternative would be an aspatial multisitedness, a notion that mocks itself so effectively that hardly any explanation is required.

The point is that the paradigms of globalization and its cousin, transnationalism, no doubt posed the major twentieth-century challenge to ethnographic methods of inquiry and units of analysis by destabilizing the embeddedness of social

relations in particular communities and places (Gille and Ó Riain, 2002). As such, they were also behind the multisited model; indeed, for Hannerz (1998), for example, 'transnational research' is broadly interchangeable with Marcus's own terminology. It is commonly thought that a refusal by ethnography to engage with a type of spatialization associated in the popular (and sometime the scholarly) imagination with modernity would (unacceptably) limit practitioners to an ever-tighter circle of apparently bounded locales.

Clearly, the earlier point about space and the present one on the space of modernity are directly linked. If, as summed up by Marshall Berman's (1983) application of a famous sentence, modernity is about 'All that is Solid Melts into Air,' that includes ethnographic space and field siting. Again, Marcus's and subsequent contributions have provided a useful point of departure. Cook et al. (2009), for example, have suggested that one way of reformulating multisited ethnography is by conceptualizing the field in such a way as to detach it from the concepts of space and place, thus opening up the possibilities of an 'unsited' field within which comparison across theoretically relevant spatial boundaries can take place. They emphasize that unsited fieldwork is a development on, rather than a negation of, multisited ethnography.

Anxieties and Comforts

Much of the discussion on multisited ethnography revolves around the idea that it may well be an oxymoron. That is, there exists a preoccupation that while there is much to be said for researching spatially dispersed objects, a program that proposes to be more routes than roots (see Clifford, 1997) could well end up robbing ethnography of its central tenets as presented earlier, and therefore of its value as the jewel in the crown of qualitative research. This preoccupation is not necessarily born of a purist conservatism (although that element is at times present), but rather of the belief that the ethnographic paradigm – traced to Malinowski by anthropology and to early-twentieth century urban studies by sociologists – has produced some of the richest social scientific insights, and is as such worth preserving. Hage epitomizes this belief: "[A]fter spending so many years examining all kinds of work in the areas of migration and diasporic studies, I have consistently found it to be the case that of all the disciplines deployed in studying globalization, migration and mobility, none are better equipped … than an ethnographic analysis" (2005: p. 474). The two main sources of anxiety (both of which have been referred to earlier) concern depth and ethnographic responsibility.

'Depth' – or, as Geertz (1973) famously put it, 'thick description' – is unquestionably one of ethnography's richest offerings. Its lack is also thought to be a key flaw of the multisited project, which seems to lead field-workers into spreading themselves too thinly. Briefly put, given that this type of research implies moving around and 'following' horizontally, there is little time for staying put and achieving depth by 'following' vertically. As Burawoy put it, "[b]ouncing from site to site, anthropologists easily substitute anecdotes and vignettes for serious fieldwork […]" (2003: p. 673).

In order to address this issue, one must question the process of production of depth/thickness in conventional ethnography; only then can one decide whether or not multisited ethnography measures up. Participant observation is the obvious answer, but what about it? The methodological stance of 'getting off the verandah' or 'hanging out' at street corners is only part, albeit a necessary one, of the story. Crucially, the factor that enables ethnographers to achieve depth and to make informed decisions about the partiality of their work is time. Participant observation has its own time order, which typically runs to several months. Initiates believe that the growth of ethnographic consciousness reveals itself to the field-worker as water boils, that is not only gradually but also through a defining moment, at which one suddenly realizes that one 'understands.' The moment of inspiration itself may be abrupt but it is actually the product of that gradual process. Prospective field-workers are told that, provided they stick it out long enough, their moment will come. In ethnography, therefore, time transforms and makes. We might here take a leaf out of Bourdieu's writing on the implicit (and therefore unquestioned, by researchers and researched alike) periodicity of gift exchange, to the effect that it is not just the nature of the gift that gives it meaning but also its *timing* – "to abolish the interval is to abolish strategy" (1977: p. 6). The contention is that by abolishing the interval, multisited ethnography abolishes depth.

There are at least four counterarguments and/or -proposals. The first argues that multisited fieldwork does indeed present these logistical and methodological challenges, and that the solution is to choose one's sites as carefully as possible in such a way that the periods spent in each site add up to an acceptable interval. In addition, careful field strategies are needed to maximize the dividends in each site, given the limited time. In FitzGerald's words, "[t]he danger of stretching time and resources too thin in multi-sited fieldwork is why it is even more dependent on a clear theoretical orientation and strategic site selection than work in a single site. Fumbling through one field site to ground theory is bad enough; in multiple sites it is disastrous" (2012: p. 3). One supposes that the criteria for site selection will be case specific and will vary according to one's resources, research questions, and ongoing collaboration in the field.

The second is to substitute long-term with very long-term fieldwork, thus enabling one to 'take in' more sites. A reasonable proposition, which, however, makes rather special demands on the field-worker. Leonard (2009), for instance, has shown how she adopted a long-term cumulative strategy by making a particular project her life's work, so to say.

The third argument, which is popularly – and probably unfairly – associated with first-generation formulations, is that the multiplicity of multisitedness makes up for its inadequacies in any single site. That is, as ethnographers move around, it becomes a matter of adding short durations to make a long one. The argument is clearly flawed in that two or more shallownesses do not make a depth; intervals simply do not add up, or at least not unproblematically.

The fourth solution is much more compelling. There is a sense – to some extent implicit in notions of 'time–space compression' and such – in which space and time are methodologically interchangeable. In other words, it is not just time that transforms and makes, but also space. This porosity has always been part of the ethnographic paradigm, not least

since conventional ethnography posits a long stay *in one place* (hence the relevance of space). And historically within anthropology, fieldwork 'away from home' (therefore, spatial displacement) was also seen as a prime route to what textbooks call 'defocusing,' an ethnographic state of mind, which in turn enables the production of data.

Clifford's writings on route-based research and the conventionally backgrounded spatial shift of 'getting there,' are well known (1997). In addition, one might argue that just as classical foundation myths have the *oikists* (founders) mark out a new city's boundaries or sacred temenos by circumambulation, ethnographers have classically set up spatial routines that enabled them to know a people by knowing a place. The ethnographer turns out to be a serial circumambulist, daily retracing their steps and in so doing producing the site and knowledge about it. Spatial routine becomes a route to ethnographic knowledge.

The time/space swap is important because it enables one to suggest that it is precisely the paradigm of participant observation (as the main portal to the native's point of view) that constitutes the strongest case for multisited ethnography. If our object is mobile and/or spatially dispersed, multisited field studies become themselves a form of participant observation by virtue of having those characteristics. As Clifford (1992) puts it, they can be seen as 'fieldwork as travel practice.' And, if conventional depth is hard to come by in unsettled circumstances, that is probably as things should be, in the sense that it represents the way people *themselves* experience the world. Russell, who cites Clifford's work extensively, is himself worth quoting more fully: "[I have] done this [fieldwork as a form of travel practice] through the example of a thirty-six hour sojourn with a historically migrant Yakkha family met fortuitously while I was 'on the road' … Such experiences help to expand, and challenge, the conventional geographical and socio-cultural boundaries of groups such as the Yakkha" (2007: p. 378, the author's parentheses).

This raises an empirical question: If it is indeed the case that multisitedness may itself be seen as a form of participant observation, what sort of tangible on-the-ground things might it help us observe, participate in? The examples that come to mind include the re-/localization dynamics and strategies of mobile groups, the production of located presence in cyberspace (see, for instance, Gatt, 2011), local experiences of geographically broad religious transformations, and the ways in which individuals and groups commute between the various scales (see Herod, 2011; the discussion in Coleman and von Hellermann, 2011) that frame social experience.

In addition, the point has been raised by FitzGerald (2012) that being multisited and having multisited knowledge are excellent ways of gaining access to one's people and eventually building rapport. It is surely relevant that FitzGerald worked with Mexican migrants (again, 'follow the people') – with a group, that is, for whom mobility and knowledge of places are particularly desirable attributes. Likewise, in his fieldwork with Sindhis, Falzon (2005) holds that being mobile and flaunting it turned out to be very effective as a means of earning his informants' respect and trust. Again, however, Sindhis are a group that particularly prizes a kind of multisited spatial fluency.

The second source of anxiety with respect to multisited field studies derives from considerations of ethnographic responsibility. Consider the following piece of postfield recollection:

> Now, if I wanted to carry out research on Gibraltarian Sindhis, where should I do that? Would I have to go to Madrid, where Ramesh Karnani was on the brink of opening a subsidiary of his father's shop? What about cousin Gope, who studied law in London? What about Anusha, whom I met in 1996 and considered a typical 'Gib-girl' of Indian origin, and in 1997 married a Sindhi from Manila, where she lives today? And how to meet Aunt Geetu from Bombay, who has never left India in her 64 years but who is in regular contact with her relatives in Gibraltar, who keeps contact with her family guru in Bombay and who probably knows more about her family in Gibraltar than anyone else living in the colony? (Haller, 2005: p. 155)

One charge that is sometimes leveled at the multisited program is that, in advocating 'following,' it appears to assume a preexisting field – a 'given' space or set of trajectories produced by the people, goods, information, and so on that are being followed. Clearly, that would go against the basic principle that space is socially produced, in this case by both researcher and researched. In other words, the very logic of contemporary understandings of space (as discussed earlier) requires that ethnographers take responsibility for the production of their field sites.

Marcus's line of thought since the original 1995 paper, as well as a number of recent scholarly contributions (see, for example, the discussion in Coleman and von Hellerman, 2011), address this problem by suggesting that the process of siting is essentially a collaborative one (as discussed earlier). It is thus that ethnography retains its ethnographic responsibility, and that practitioners are reminded that they have an important part to play in the dynamics of carving out a site – theirs is a shared responsibility, so to speak. Viewed this way, multisited ethnography is a form of, to use a term from Marcus (1998), 'complicity.' As law teaches us, complicity is still a very real, if partial, responsibility.

Perhaps more importantly, the notion of partial ethnographic responsibility opens up the whole field of looking at ways in which it is produced; surely, this must be one of the more useful corollaries of Marcus's program. As Gille and Ó Riain (2002: p. 289) put it, "[t]he extension of the ethnographic site in space and time sharpens one's sensibilities to the political consequences of defining a site or sites." The methodology of multisited field studies has come a long way from straightforward notions of 'following.'

In conclusion, one might ask whether or not Marcus's call has produced anything of lasting value. It would appear that multisitedness is not by nature a buzzword, although it may be and indeed has on occasion been bandied about as such. Especially when one takes into account the ongoing critiques (including Marcus's own) and increasing theoretical sophistication, it has made at least three contributions to the field:

- first, a heightened methodological and theoretical sensitivity to space, place, scale, and boundaries, and to the ways in which they are manufactured and experienced both by ethnographers and by the people they work with;
- second, a shift toward collaborative field-making and research and therefore toward an ethnographic knowledge

that is better aware of, and intellectually equipped to experiment with, the processes of its coproduction;

- third, an ongoing discussion – and actual empirical results – of some of the practical ways in which ethnographers might go about producing this spatialized and collaborative knowledge.

The place may look familiar but multisited ethnography has thus far turned out to be anything but a methodological cul-de-sac.

Acknowledgment

Sections of this article borrow on the author's Introduction to *Multi-sited Ethnography: Theory, Praxis and Locality in Contemporary Research*. Ashgate, Farnham.

See also: Ethnography; Fieldwork in Social and Cultural Anthropology; Globalization and World Culture; Globalization: Geographical Aspects; Migration: Anthropological Perspectives; Spatial Analysis in Geography; Spatial Thinking in the Social Sciences; Thick Description: Methodology; World Systems Theory.

Bibliography

Amelina, A., Faist, T., 2012. De-naturalizing the national in research methodologies: key concepts of transnational studies in migration. Ethnic and Racial Studies iFirst 35, 1707–1724.
Berman, M., 1983. All That is Solid Melts into Air. Verso, London.
Bernard, H.R. (Ed.), 1998. Handbook of Methods in Cultural Anthropology. Rowman & Littlefield, Walnut Creek, CA.
Bourdieu, P., 1977. Outline of a Theory of Practice, Cambridge Studies in Social and Cultural Anthropology. Cambridge University Press, Cambridge.
Burawoy, M., 2003. Revisits: an outline of a theory of reflexive ethnography. American Sociological Review 68 (5), 645–679.
Candea, M., 2007. Arbitrary locations: in defence of the bounded field-site. Journal of the Royal Anthropological Institute 13 (1), 167–184.
Choy, T., et al., 2009. Matsutake Worlds Research Group. In: Falzon, M.A. (Ed.), Strong Collaboration as a Method for Multi-sited Ethnography: On Mycorrhizal Relations.
Clifford, J., 1992. Traveling Cultures. In: Grossberg, L., et al. (Eds.), Cultural Studies. Routledge, London.
Clifford, J., 1997. Routes: Travel and Translation in the Late Twentieth Century. Harvard University Press, Cambridge.
Collier, S.J., Ong, A., 2005. Global assemblages: technology, politics, and ethics as anthropological problems. In: Ong, A., Collier, S.J. (Eds.), Global Assemblages, Anthropological Problems. Blackwell, Malden, MD.
Coleman, S., von Hellermann, P., 2011. Introduction: queries, collaborations, calibrations. In: Coleman, S., von Hellermann, P. (Eds.), Multi-sited Ethnography: Problems and Possibilities in the Translocation of Research Methods. Routledge, New York and London.
Cook, J., Laidlaw, J., Mair, J., 2009. In: Falzon, M.A. (Ed.), What If There is No Elephant? Towards a Conception of an Un-sited Field.
Davidson, I., 2007. Ideas of Space in Contemporary Poetry. Palgrave Macmillan, Basingstoke.
Denzin, N.K., Lincoln, Y.S. (Eds.), 2005. The Sage Handbook of Qualitative Research, third ed. Sage, Thousand Oaks.
Dilley, R., 1999. Introduction: the problem of context. In: Dilley, R. (Ed.), The Problem of Context. Oxford, Berghahn.
Fairhead, J., 2011. In: Coleman, S., von Hellermann, P. (Eds.), Part C Introduction.
Falzon, M.A., 2005. Cosmopolitan Connections: The Sindhi Diaspora, 1860–2000. Oxford University Press, New Delhi.
Falzon, M.A., 2009. Multi-sited Ethnography: Theory, Praxis and Locality in Contemporary Research. Ashgate, Farnham.
FitzGerald, D., 2012. A comparative manifesto for international migration studies. Ethnic & Racial Studies iFirst 35, 1725–1740.
Fog Olwig, K., Hastrup, K., 1996a. In: Fog Olwig, K., Hastrup, K. (Eds.), Introduction.
Fog Olwig, K., Hastrup, K. (Eds.), 1996b. Siting Culture: The Shifting Anthropological Object. Routledge, London.
Freund, J., 1968. The Sociology of Max Weber. Pantheon, New York.
Garsten, C., 1994. Apple World: Core and Periphery in a Transnational Organizational Culture, Stockholm Studies in Social Anthropology, vol. 33. Almqvist and Wiksell International, Stockholm.
Gatt, C., 2011. Becoming Friends of the Earth: An Anthropology of Global Environmentalism (unpublished Ph.D. dissertation). University of Aberdeen.
Gay y Blasco, P., Wardle, H., 2007. How to Read Ethnography. Routledge, London and New York.
Geertz, C., 1973. The Interpretation of Cultures: Selected Essays. Basic Books, New York.
Gille, Z., Ó Riain, S., 2002. Global ethnography. Annual Review of Sociology 28, 271–295.
Grossberg, L., Nelson, C., Treichler, P.A. (Eds.), 1992. Cultural Studies. Routledge, New York.
Gustavson, L.C., Cytrynbaum, J.D., 2003. Illuminating spaces: relational spaces, complicity, and multi-sited ethnography. Field Methods 15 (3), 252–270.
Hage, G., 2005. A not so multi-sited ethnography of a not so imagined community. Anthropological Theory 5 (4), 463–475.
Haller, D., 2005. Let it flow: economy, spirituality, and gender in the Sindhi network. Anthropological Theory 5 (2), 154–175.
Hannerz, U., 1998. In: Bernard, Russell (Ed.), Transnational Research.
Hannerz, U., 2003. Being there … and there … and there!: reflections on multi-site ethnography. Ethnography 4 (2), 201–216.
Herod, A., 2011. Scale. Routledge, London and New York.
Holmes, D.R., Marcus, G.E., 2004. In: Ong, A., Collier, S.J. (Eds.), Cultures of Expertise and the Management of Globalization: Toward the Re-functioning of Ethnography.
Holmes, D.R., Marcus, G.E., 2005. In: Denzin, N.K., Lincoln, Y.S. (Eds.), Refunctioning Ethnography: The Challenge of an Anthropology of the Contemporary.
Huang, S., 2006. Mobile Homes: Spatial and Cultural Negotiation in Asian American Literature. Routledge, New York, and London.
Krauss, W., 2009. In: Falzon, M.A. (Ed.), Localizing Climate Change: A Mutli-sited Approach.
Lefebvre, H., 1991. The Production of Space. first published in French in 1974 as La production de l'espace, (D. Nicholson-Smith, Trans.). Blackwell, Oxford.
Leonard, K.I., 2009. In: Falzon, M.A. (Ed.), Changing Places: The Advantages of Multi-sited Ethnography.
Marcus, G.E., 1995. Ethnography in/of the world system: the emergence of multi-sited ethnography. Annual Review of Anthropology 24, 95–117.
Marcus, G.E., 1998. Ethnography through Thick and Thin. Princeton University Press, Princeton, NJ.
Marcus, G.E., Mascarenhas, F., 2005. Ocasião: The Marquis and the Anthropologist, a Collaboration. AltaMira, Walnut Creek.
Massey, D., 2005. For Space. Sage, London.
Michelucci, S., 2002. Space and Place in the Works of D.H. Lawrence. McFarland & Co., Jefferson, NC, and London.
Mintz, S.W., 1998. The localization of anthropological practice: from area studies to transnationalism. Critique of Anthropology 18 (2), 117–133.
Ong, A., Collier, S.J. (Eds.), 2005. Global Assemblages: Technology, Politics, and Ethics as Anthropological Problems. Blackwell, Oxford.
Russell, A., 2007. Writing traveling cultures: travel and ethnography amongst the Yakkha of East Nepal. Ethnos 72 (3), 361–382.
Ryzewski, K., 2012. Multiple situated strategies? Multi-sited ethnography and archeology. Journal of Archaeological Method and Theory 19 (2), 241–268.
Salzman, P.C., 1986. Is traditional fieldwork outmoded? Current Anthropology 27 (5), 528–530.
Soja, E.W., 1989. Postmodern Geographies: The Reassertion of Space in Critical Social Theory. Verso, London.
Whyte, M.A., Reynolds Whyte, S., Twebaze, J., 2011. In: Coleman, S., von Hellermann, P. (Eds.), Understanding HIV/AIDS in Uganda: Sites and Positions.

Multistate Transition Models in Demography

Michel Guillot, Population Studies Center, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Courgeau, volume 15, pp. 10210–10214, © 2001, Elsevier Ltd.

Abstract

Multistate transition models provide an analytical framework for situations in which there exists a certain number of 'states' of interest that individuals can occupy, and between which individuals can move. Multistate models allow one to describe these situations by examining patterns of transitions across states and estimating state-specific life expectancies. This article emphasizes the basic properties of multistate transition models, as well as practical strategies for estimating them.

Introduction

In demography, multistate transition models refer to situations in which there exists a certain number of 'states' of interest that individuals can occupy, and between which individuals can move (Rogers, 1975; Land et al., 1982; Schoen, 1988a,b). Examples of 'states' include the following:

- 'Residing in urban areas' versus 'residing in rural areas.' In this example, transitions between states refer to migrations from urban to rural areas and vice versa.
- 'Having no disability' versus 'being disabled.' In this example, transitions between states refer to the incidence of disability (transitions from having no disability to being disabled) and to recovery from disability (transitions from being disabled to having no disability).
- 'Being never married'; 'being married'; 'being divorced'; 'being widowed.' In this example, transitions between states refer to various changes of marital status, including the incidence of first marriage (transitions from 'never married' to 'married'), the incidence of divorce (transitions from 'married' to 'divorced'), etc.

In all the above examples, there necessarily exists an additional state, namely, the state of 'being dead.' Transitions to that state represent mortality. A multistate transition model contains a minimum of two 'living' states, in which individuals can spend a certain amount of time, as well as the state of being dead.

An important feature of multistate transition models is that they allow the study of situations in which individuals can move back and forth between the different 'living' states of interest. This contrasts with single state models, like the classic life table, in which only one transition is possible with no possible return to the state of origin, that is, from 'being alive' to 'being dead.'

Multistate transition models in demography allow one to quantify patterns of transitions between states and study how they vary over age and time. These models also allow one to calculate a number of interesting quantities that result from given transition patterns. For example, if we take the marriage formation model described above, the multistate framework allows one to calculate the following quantities of interest:

- How many years can a newborn expect to live in the 'married' state?
- How many additional years can a person in the 'divorced' state at age 50 expect to live in the 'married' state?
- What is the probability that a marriage will end in a divorce?
- What is the probability of remarriage from divorce or widowhood?
- What is the probability that a newborn will die in the married state?

This article is organized as follows. First, I discuss the simple case of transition models without age. The next section introduces age with a discussion of multistate life tables. The last section discusses multistate population models, i.e., models in which patterns of entry into the multistate system are integrated (for example, via births), generating populations with interesting properties.

Multistate Transition Models without Age

In their simplest form, multistate transition models do not include age: the population of interest is broken down by state but not by age. For example, let us consider a population broken down into two states: urban population versus rural population. A simple multistate model allows one to examine patterns of migration between urban and rural populations and how these populations grow or decline over time as a result of these transitions. Assuming that the overall population is closed to international migration, the following forces are affecting the size of the urban and rural populations:

- For the urban population: the crude rate of natural increase (defined as the difference between the crude birth rate and the crude death rate for the urban population) and the crude rate of out-migration (from urban to rural areas).
- For the rural population: the crude rate of natural increase (here defined for the rural population) and the crude rate of out-migration (here from rural to urban areas).

An interesting feature of this model is the interdependence between the two states: the size of the urban population affects the size of the rural population through out-migration from urban to rural areas, and the size of the rural population affects the size of the urban population through out-migration from rural to urban areas. This is an important feature shared by all multistate transition models.

Another important feature of this model is that if rates of natural increase and out-migration remain constant over time for both urban and rural populations, eventually urban and rural populations will be growing at the same rate, and the proportionate distribution of the population by urban/rural residence will become constant. This eventual state of equilibrium is independent of the initial distribution of the population by urban/rural residence. This is illustrated in Figure 1, which shows trajectories toward equilibrium in populations with proportion urban ranging from 10% to 90% at baseline. Assuming a crude rate of natural increase of 20 per 1000 in urban areas and of 30 per 1000 in rural areas, and a crude rate of out-migration of 10 per 1000 in urban areas and of 30 per 1000 in rural areas, populations will all reach an equilibrium proportion urban of 69.7%, regardless of the initial proportion urban. This equilibrium proportion is entirely the product of the assumed constant rates of natural increase and out-migration.

Multistate (Increment–Decrement) Life Tables

Multistate life tables are a different but related kind of multistate transition model. Instead of following a population over time, like in the above example, a multistate life table follows a cohort through different ages. Members of a cohort are born in a given state, and as they age they may move through different states, until they eventually die. A multistate life table describes and quantifies this process.

Multistate life tables are also called increment–decrement life tables because the number of individuals in one state can increase or decrease with age, depending on the balance of flows in-and-out of that state. This contrast with the classic life table model where the number of people in the state of 'being alive' can only decrease with age.

Multistate Life Tables for a Cohort

Multistate life tables are usually calculated for periods (using synthetic cohorts), but are best understood by looking at a real cohort and tracking cohort survivors over time as they move between different states. Therefore, we first discuss features of a multistate life table by looking at the case of a cohort.

For example, let us examine the simple case of a multistate life table model representing marriage formation, in which there are three states: (1) unmarried; (2) married; and (3) dead. Figure 2 represents the three states and the arrows represent possible transitions between the states.

It can be assumed that all individuals are born in the 'unmarried' state. As individuals age, they may transition between states: they may become married and then return to the 'unmarried' state through divorce or widowhood; they may move back to the state of being married through remarriage. Eventually, due to mortality, all individuals transition to the 'dead' state, either from the 'unmarried' state or from the 'married' state, depending on their marital status at the time of death.

A multistate life table summarizes the history of the cohort with the following columns:

- x = exact age;
- l_x^i = number of individuals in the cohort who are in state i at age x.

For example, in the marriage model described above, we have

l_x^U = number of individuals in the cohort who are *unmarried* at age x;
l_x^M = number of individuals in the cohort who are *married* at age x.

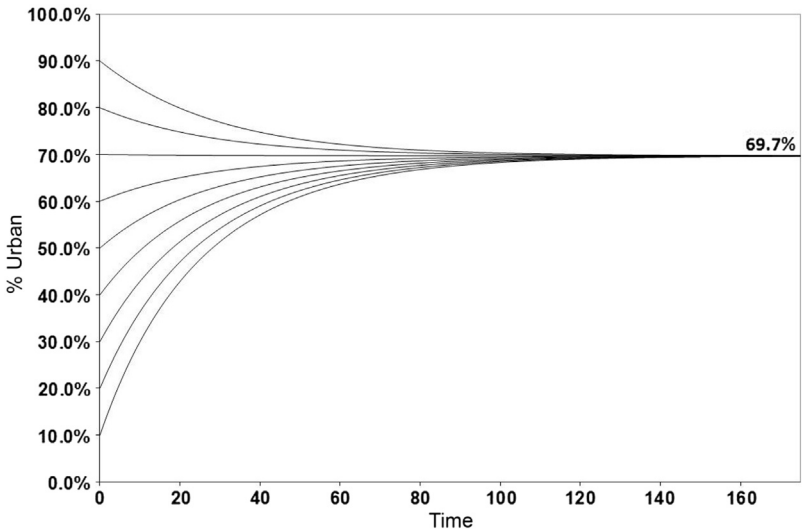

Figure 1 Proportion of the population living in urban areas for populations in which rates of natural increase and rates of out-migration are set to specific values and constant over time. In this example, all populations converge to a proportion urban of 69.7%.

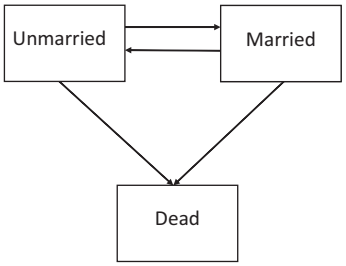

Figure 2 Possible transitions among states in a simple marriage formation multistate model.

Note that l_x^U or l_x^M may increase or decrease with age (which is why, as said earlier, multistate life tables are also called increment–decrement life tables). Note also that $l_x^U + l_x^M = l_x$, the number of cohort survivors at age x. l_0 is the initial size of the cohort (at birth). In this example, all individuals are unmarried at birth, so $l_0^M = 0$ and $l_0^U = l_0$. In other examples, individuals may be distributed across more than one state at birth. For example, in an urban/rural residence transition model, individuals may be born either in urban or rural areas.

- ${}_nd_x^{ij}$ = number of transitions from state i to state j occurring between ages x and $x+n$ among cohort members

For example, in the marriage model described above, we have four different types of transitions:

${}_nd_x^{UM}$ = number of transitions from unmarried to married (i.e., number of marriages) among cohort members aged x to $x+n$;
${}_nd_x^{UD}$ = number of transitions from 'unmarried' to 'dead' (i.e., number of deaths) among unmarried cohort members aged x to $x+n$;
${}_nd_x^{MU}$ = number of transitions from 'married' to 'unmarried' (i.e., through divorce or widowhood) among cohort members aged x to $x+n$; and
${}_nd_x^{MD}$ = number of transitions from married to dead (i.e., number of deaths) among married cohort members aged x to $x+n$.

In a multistate life table, survivors and transitions are related with one another in the following fashion (using the marriage model as an example):

$$l_{x+n}{}^U = l_x{}^U - {}_nd_x{}^{UM} - {}_nd_x{}^{UD} + {}_nd_x{}^{MU}$$

$$l_{x+n}{}^M = l_x{}^M - {}_nd_x{}^{MU} - {}_nd_x{}^{MD} + {}_nd_x{}^{UM}$$

- ${}_nq_x^{ij}$ = probability that an individual who is in state i at age x will be in state j at age $x+n$.

If it can be assumed that only one transition is possible between x and $x+n$, this probability is related to other life table quantities through the following equation:

$${}_nq_x{}^{ij} = {}_nd_x{}^{ij}/l_x{}^i$$

In the marriage example, there are four possible transition probabilities:

${}_nq_x{}^{UM}$ = probability that an *unmarried* individual aged x will be *married* at age $x+n$;
${}_nq_x{}^{UD}$ = probability that an *unmarried* individual aged x will *die* by age $x+n$;
${}_nq_x{}^{MU}$ = probability that a *married* individual aged x will be *unmarried* at age $x+n$; and
${}_nq_x{}^{MD}$ = probability that a *married* individual aged x will *die* by age $x+n$.

- ${}_nL_x{}^i$ = number of person-years lived in state i between age x and $x+n$ by individuals in the cohort

Obviously, person-years can only be lived in a 'living' state, so in the marriage example, there are only two types of person-years:

${}_nL_x{}^U$ = person-years lived in the unmarried state between age x and $x+n$ by individuals in the cohort; and
${}_nL_x{}^M$ = person-years lived in the married state between age x and $x+n$ by individuals in the cohort.

Note that ${}_nL_x{}^U + {}_nL_x{}^M = {}_nL_x$, i.e., the number of person-years lived between x and $x+n$ by all members of the cohort, regardless of marital status.

- ${}_nm_x^{ij}$ = transition rate from state i to state j between age x and $x+n$

This transition rate is related to other life table functions through the following equation:

$${}_nm_x{}^{ij} = {}_nd_x{}^{ij}/{}_nL_x{}^i$$

This is a classic occurrence/exposure rate with events (transitions) in the numerator and person-years of exposure in the denominator. Since only individuals in state i are at risk of experiencing a transition from state i to state j, the exposure term in the denominator is limited to person-years lived in state i. In the marriage example, as in the case of transition probabilities, each age group has four possible transition rates: ${}_nm_x{}^{UM}$; ${}_nm_x{}^{UD}$; ${}_nm_x{}^{MU}$; and ${}_nm_x{}^{MD}$.

- $T_x{}^i$ = number of person-years lived in state i above age x by individuals in the cohort

This quantity corresponds to the sum of ${}_nL_x{}^i$ for ages x and higher. In the marriage example, there will be two types of $T_x{}^i$ values:

$T_x{}^U$ = number of person-years lived above age x in the *unmarried* state; and
$T_x{}^M$ = number of person-years lived above age x in the *married* state.

Note that $T_x{}^U + T_x{}^M = T_x$, the number of person-years lived above age x by all members of the cohort.

- $e_x{}^i$ = life expectancy at age x in state i = number of years that an individual aged x can expect to live in state i

These life expectancies are related to other life table functions through the following equation:

$$e_x{}^i = T_x{}^i/l_x$$

These life expectancies are called 'unconditional' life expectancies, because the denominator includes all survivors at age x (l_x), regardless of state occupancy at age x.

In the marriage example, there will be two types of unconditional life expectancies:

e_x^U = number of years that an individual aged x can expect to live in the *unmarried* state; and
e_x^M = number of years that an individual aged x can expect to live in the *married* state.

Note that $e_x^U + e_x^M = e_x$. That is, a multistate life table apportions the classic life expectancy, e_x, among different states. In turn, state-specific unconditional life expectancies add up to the classic life expectancy.

- T_x^{ij} = number of person-years lived in state j above age x by cohort members who are in state i at age x

Typically, in a multistate transition model, future state occupancy depends highly on the state that an individual occupies at age x. For example, an individual who is married at age 50 cannot expect to spend the same number of years above 50 in the married state as an individual who is not married at age 50. When calculating person-years lived above age x, it is thus useful to calculate person-years lived above age x in a given state conditional on state occupancy at age x. In the marriage example, there will be four types of conditional T_x^{ij} values:

T_x^{UU} = number of person-years lived above age x in the *unmarried* state by individuals who are *unmarried* at age x;
T_x^{UM} = number of person-years lived above age x in the *married* state by individuals who are *unmarried* at age x;
T_x^{MU} = number of person-years lived above age x in the *unmarried* state by individuals who are *married* at age x; and
T_x^{MM} = number of person-years lived above age x in the *married* state by individuals who are *married* at age x.

Note that $T_x^{UU} + T_x^{MU} = T_x^U$ defined earlier. Likewise, $T_x^{UM} + T_x^{MM} = T_x^M$.

These person-years conditional on state occupancy at age x allow one to calculate a second kind of life expectancies, called 'conditional' life expectancies:

- e_x^{ij} = life expectancy at age x in state j, for individuals who are in state i at age x.

These life expectancies are related to other life table functions through the following equation:

$$e_x^{ij} = T_x^{ij}/l_x^i$$

In the marriage example, there are four types of conditional life expectancies:

e_x^{UU} = number of years that an individual who is *unmarried* at age x can expect to live in the *unmarried* state;
e_x^{UM} = number of years that an individual who is *unmarried* at age x can expect to live in the *married* state;
e_x^{MU} = number of years that an individual who is *married* at age x can expect to live in the *unmarried* state; and
e_x^{MM} = number of years that an individual who is *married* at age x can expect to live in the *married* state.

Note that unconditional life expectancies are weighted averages of conditional life expectancies, with the weights being the proportion of survivors by state at age x. For example, in the marriage example

$$e_x^U = l_x^U/l_x \cdot e_x^{UU} + l_x^M/l_x \cdot e_x^{MU}$$

In a multistate life table, the most important quantities are ${}_nq_x^{ij}$ or ${}_nm_x^{ij}$, which summarize the pattern of transitions between states, as well as e_x^i or e_x^{ij}, which summarize mean duration of state occupancy above age x.

Multistate Life Tables in a Period

If individual-level cohort information is available, life expectancies can be readily observed without resorting to any kind of complex computational approach. Person-years spent in each state by cohort members can be directly observed and summed across individuals. Life expectancies in a given state are simply mean person-years lived in a given state, conditional or not on earlier state occupancy.

As a simple numerical example, let us assume that the following information is available on two individuals born the same day: individual #1 died at age 80 with 30 years spent in the unmarried and 50 years spent in the married state, while individual #2 died at age 60 with 46 years spent in the unmarried state and 14 years spent in the married state. It can then be readily estimated that the life expectancy at birth for this cohort of two individuals is $(80+60)/2 = 70$ years, with an unconditional life expectancy at birth in the unmarried state of $(30+46)/2 = 38$ years and an unconditional life expectancy at birth in the married state of $(50+14)/2 = 32$ years. Since in this model all individuals are born in the unmarried state, life expectancies at birth conditional on being unmarried at birth (e_0^{UU} and e_0^{UM}) are equal to the unconditional life expectancies (e_0^U and e_0^M, respectively) and the life expectancies at birth conditional on being married at birth (e_0^{MU} and e_0^{MM}) are undefined. With additional information on ages at transition for cohort members, life expectancies at ages above 0, conditional or unconditional, can also be readily observed without using information on transition rates or probabilities.

Most often, however, analysts are interested in estimating multistate life tables for a period. In a period multistate life table, age-specific transition rates are observed for a given year or period, and the analyst is interested in examining what would happen to a cohort of individuals if they were exposed at each age to the transition rates observed during the given period. This is the classic synthetic cohort approach that is used whenever constructing a classic period life table.

When constructing a period life table, however, person-years and life expectancies by state cannot be readily observed, because the synthetic cohort is not a real cohort and thus cannot be tracked over time. One classic strategy for calculating a period multistate life table starts from calculating transition rates, which can be estimated from period data if information on transitions and population at risk is available:

$${}_nm_x^{ij} \approx {}_nD_x^{ij}/{}_nN_x^i$$

where ${}_nD_x^{ij}$ are observed transitions from state i to state j between age x and $x+n$ during a given year, and where ${}_nN_x^i$ is the midyear population in state i between ages x and $x+n$, used as an estimate of the number of person-years of exposure in state i between ages x and $x+n$ during a given year.

Using these transition rates, a synthetic cohort can then be simulated and corresponding person-years and life expectancies can be estimated. Like in the case of a classic life table construction, a key aspect of period multistate life table construction consists of performing what amounts to a rate-to-probability conversion, so that the synthetic cohort can be simulated. This involves solving a system of equations, with assumptions about how transitions are distributed within the age interval. One classic assumption is to assume that the number of state-specific survivors varies linearly between age x and $x+n$. In the marriage example, this assumption would produce the following system of equations (one system for each age group):

1. $l_{x+n}^{U} = l_x^{U} - {}_nd_x^{UM} - {}_nd_x^{UD} + {}_nd_x^{MU}$
2. $l_{x+n}^{M} = l_x^{M} - {}_nd_x^{MU} - {}_nd_x^{MD} + {}_nd_x^{UM}$
3. ${}_nd_x^{UM} = {}_nm_x^{UM} \cdot {}_nL_x^{U}$
4. ${}_nd_x^{UD} = {}_nm_x^{UD} \cdot {}_nL_x^{U}$
5. ${}_nd_x^{MU} = {}_nm_x^{MU} \cdot {}_nL_x^{M}$
6. ${}_nd_x^{MD} = {}_nm_x^{MD} \cdot {}_nL_x^{M}$
7. ${}_nL_x^{U} = n/2 \cdot (l_x^{U} + l_{x+n}^{U})$
8. ${}_nL_x^{M} = n/2 \cdot (l_x^{M} + l_{x+n}^{M})$

In this system of eight equations, the only known quantities are the following: ${}_nm_x^{UM}$, ${}_nm_x^{UD}$, ${}_nm_x^{MU}$, and ${}_nm_x^{MD}$. In total, there are eight equations and eight unknowns, which is solvable. The algebraic solution for the case of two living states can be found in Schoen (1988a). If there are more than two living states, the system of equations will be more complex and can be solved using matrix algebra (Palloni, 2001).

Like in any life table construction, the rate-to-probability conversion involves certain assumptions. In the example above, the assumption is that the number of survivors in state i (l_x^i) varies linearly between ages x and $x+n$. The most common alternative is to assume that transition rates, in continuous terms, are constant between ages x and $x+n$ (Schoen, 1988a).

Once the system of equations has been solved, the synthetic cohort can be simulated. One simply needs to choose an arbitrary radix for the life table (l_0), and apportion this radix by state (l_0^i). In many cases, the distribution by state of individuals aged 0 is theoretically implicit. For example, in the marriage model, it is assumed that $l_0^U = l_0$. In some cases, however, the analyst needs to decide how to distribute individuals aged 0 by state. A common option is to use actual distributions of births by state as recently observed in the population. We will see later that in multistate models that involve reproduction, the distribution of births by state is entirely determined by the model, so no arbitrary choice needs to be made. Once values of l_0^{i} are chosen, state-specific survivors can be simulated as if they were exposed to the observed period transition rates, and the different life expectancies described above can be calculated.

Note that the construction of a multistate life table typically makes an assumption of homogeneity: all the individuals in state i at age x are assumed to have the same probability of experiencing a transition. That is, transition probabilities depend only on age and current state (a setup also known as a first-order Markov chain in the statistics literature). The past history of state occupancy is not taken into account. This assumption is potentially problematic for period life tables if indeed duration in a state influences transition rates in addition to age.

Statistical Approaches to Estimating Multistate Life Tables

The life table approach discussed above assumes that transition rates can be estimated with precision for well-defined age groups and time periods using survey data or exhaustive population information. In many cases, however, this goal is difficult to attain, especially when dealing with survey data. Transitions of interest to demographers (such as death, marriage, disability, etc.) tend to be relatively rare events. Sample sizes in available surveys are often too small to produce estimates of transition rates that are precise enough for direct use in life table construction. Moreover, longitudinal surveys typically have varying interview dates and individuals are observed at irregular intervals, which create additional difficulties for the calculation of well-defined rates or probabilities. Finally, the approach discussed above is deterministic and does not readily allow statistical inference for transition rates or life expectancies.

The typical data configuration for the statistical estimation of transition rates on the basis of longitudinal data is as follows: a sampled individual is observed at a first interview date, during which information is gathered about age and current state. Later, the individual is reinterviewed, and current state is updated. If the individual has changed states by the time of the second interview, he/she may be asked about the date at which his/her status changed. If the individual died between the two interview dates, information about the exact date of death is sometimes recorded. Sometimes information at a third interview may be available, but some individuals may be missing information at one of the three interviews. Statistical approaches for the estimation of a multistate transition model will use information about sampled individuals with a range of ages at a first interview (for which the exact date may vary across individuals), and then estimate a set of period age-specific transition rates that best agree with the observed patterns of transitions among sampled individuals.

There are many different statistical approaches to estimating period age-specific transitions rates from longitudinal survey data (Willekens and Putter, 2014). One commonly used approach is called Interpolated Markov Chains (IMaCh) and is available as a software (Lièvre et al., 2003; Brouard and Lièvre, 2009). IMaCh is tailored to the estimation of health expectancies (with a transition model involving three states: healthy; disabled; and dead) and makes parametric assumptions about how transition rates vary with age. Specifically, IMaCh assumes that the partial odds of monthly transition probabilities follow a log-linear function of age (Laditka and Wolf, 1998). Using a maximum likelihood approach, the software produces a set of parameters for the log-linear functions that best agrees with the set of observed individual transitions. The estimated transition probabilities are then used to estimate of complete multistate life table, including state-specific life expectancies. Confidence intervals around life expectancies are calculated using the Delta method.

Another software, called SPACE, was published in 2010 by Cai et al. (Cai et al., 2010). One important difference with IMaCh is that life table outputs are calculated by resorting to

micro-simulation. Once transition probabilities are estimated, individual life histories are simulated and then aggregated to produce mean values of the output quantities. Confidence intervals around mean output values are calculated using bootstrapping. The SPACE approach is being increasingly used for the estimation of health expectancies (Payne et al., 2013).

Multistate Population Models

Multistate Stationary Populations

If it can be assumed that there is a constant annual number of births by state (B^i), and that age-specific transition rates are constant over time, then the corresponding population is a multistate stationary population. Like a classic stationary population, this population will have a constant size and constant state and age distribution. Moreover, the population distribution by state will be proportional to the ${}_nL_x{}^i$ column of the corresponding multistate life table:

$$ {}_nN_x^i/N^i = {}_nL_x^i/T_0{}^i, $$

where ${}_nN_x{}^i$ is the population aged x to $x+n$ in state i, and where N^i is the total population in state i; and

$$ N^i/N = T_x{}^i/T_0, $$

where N is the total population size for all states combined.

Thus, in a stationary multistate population, unconditional life expectancies (e_0^i) can be readily calculated by applying the proportion of the total population that is in state i to the overall life expectancy: $e_0^i = e_0 \cdot (N^i/N)$. In the marriage example, if the multistate model is stationary with an overall life expectancy of 75 years and with 40% of the total population in the 'married' state, then the number of years that a newborn can expect to live in the married state if $75 \cdot 0.4 = 30$ years.

This stationary model is a special case of a more general population model in which transition rates are constant over time, and in which the number of births, while varying over time, has a constant distribution by state. That is, $B^i(t)/B(t)$ is constant over time t. This is a useful model, because in many cases, $B^i(t)/B(t)$ is constant over time due to the very nature of the multistate process. In the marriage model, for example, regardless of variations in the number of births each year, 100% of births occur in the unmarried state every year.

In this less restrictive population model, N^i/N is not equal to $T_0{}^i/T_0$, like in the case of a stationary population. However, ${}_nN_x^i/{}_nN_x = {}_nL_x^i/{}_nL_x$. This provides a useful shortcut for calculating a multistate life table in the absence of information on transitions. One simply needs to first calculate an ${}_nL_x$ column for all states combined (i.e., as in a classic life table). ${}_nL_x{}^i$ can then be calculated by applying at each age the observed proportion of the age-specific population that is in state i, using for example, information from a census or a cross-sectional survey:

$$ {}_nL_x^i = {}_nL_x \cdot {}_nN_x^i/{}_nN_x $$

This approach is commonly used for estimating disability-free life expectancy in the absence of information on transitions in-and-out of disability (Cambois et al., 1999). This method, called the Sullivan method (Sullivan, 1971), assumes that: (1) age-specific transition rates in the multistate system have been constant over time; and (2) the proportion of births by disability status has been constant over time. While the second assumption is not problematic (the proportion of individuals who are free of disability at birth is close to 100% and can be assumed to be constant over time), the assumption of constant transition rates can be problematic in populations that have experienced fast mortality declines and changes in the incidence of disability (Rogers et al., 1990; Barendregt et al., 1994, 1997; Lièvre et al., 2003).

Multistate Stable Populations

The multistate stable population model is a model which includes entry into the population through reproduction, in addition to attrition out of the population through mortality (Schoen, 1988a,b). This model makes the following assumptions:

- Age-specific transition rates are constant over time;
- Age-specific fertility rates by state are constant over time.

If these assumptions hold, then a multistate stable population emerges with the following features:

- The proportionate distribution of the population by age and state is constant over time;
- The overall growth rate of the population is constant over time.

Multistate stable populations have the same property of ergodicity as classic stable populations. This implies that any population with multiple states has an underlying 'stable-equivalent' multistate population with an intrinsic growth rate and an intrinsic distribution of the population by age and state. Although it does not include age, the urban/rural residence model presented in **Figure 1** illustrates this property of ergodicity and the concept of intrinsic parameters. In this example, the observed set of rates of natural increase and out-migration produces an 'intrinsic' proportion of the population living in urban areas of 69.7%.

Multistate stable populations have been used in a number of applications. One classic application shows that fertility differentials by IQ levels (with low-IQ persons having above-average fertility), rather than generating decreases in the population's mean IQ level in the long term, generate constant mean IQ levels in the long term that may or may not be lower than the population's current mean IQ level (Preston and Campbell, 1993).

See also: Aging and Health in Old Age; Demographic Models; Families and Households, Formal Demography of; Life Table; Microsimulation in Demographic Research; Period and Cohort Analysis in Demography; Population Dynamics: Classical Applications of Stable Population Theory; Population Dynamics: Mathematic Models of Population, Development, and Natural Resources; Population Dynamics: Momentum of Population Growth.

Bibliography

Barendregt, J.J., Bonneux, L., Van Der Maas, P.J., 1994. Health expectancy: an indicator for change? Journal of Epidemiology and Community Health 48, 482–487.

Barendregt, J.J., Bonneux, L., Van Der Maas, P.J., 1997. How good is Sullivan's method for monitoring changes in population health expectancies? Journal of Epidemiology and Community Health 51, 578–579.

Brouard, N., Lièvre, A., 2009. IMaCh: A Maximum-Likelihood Computer Program Using Interpolation of Markov Chains. http://euroreves.ined.fr/imach/.

Cai, L., Hayward, M., Saito, Y., Lubitz, J., Hagedorn, A., Crimmins, E., 2010. Estimation of multi-state life table functions and their variances using the SPACE program. Demographic Research 22, 129–158.

Cambois, E., Robine, J.M., Brouard, N., 1999. Life expectancies applied to specific statuses: a history of the indicators and the methods of calculation. Population: An English Selection 53, 447–476.

Laditka, S.B., Wolf, D.A., 1998. New methods for analyzing active life expectancy. Journal of Aging and Health 10, 214–241.

Land, K.C., Rogers, A., National Science Foundation (U.S.), 1982. Multidimensional Mathematical Demography. Academic Press, New York.

Lièvre, A., Brouard, N., Heathcote, C., 2003. The estimation of health expectancies from cross-longitudinal surveys. Mathematical Population Studies 10, 211–248.

Palloni, A., 2001. Increment-decrement life tables. In: Preston, S.H., Heuveline, P., Guillot, M. (Eds.), Demography: Measuring and Modeling Population Processes. Blackwell Publishers.

Payne, C.F., Mkandawire, J., Kohler, H.-P., 2013. Disability transitions and health expectancies among adults 45 years and older in Malawi: a cohort-based model. PLoS Medicine 10, e1001435.

Preston, S.H., Campbell, C., 1993. Differential fertility and the distribution of traits: the case of IQ. American Journal of Sociology 98, 997–1019.

Rogers, A., 1975. Introduction to Multiregional Mathematical Demography. Wiley, New York.

Rogers, A., Rogers, R., Belanger, A., 1990. Longer life but worse health? Measurement and dynamics. The Gerontologist 30, 640–649.

Schoen, R., 1988a. Modeling Multigroup Populations. Plenum Press, New York.

Schoen, R., 1988b. Practical uses of multistate population-models. Annual Review of Sociology 14, 341–361.

Sullivan, D.F., 1971. A Single Index of Mortality and Morbidity.

Willekens, F.J., Putter, H., 2014. Software for multistate analysis. Demographic Research 31, 381–420.

Multivariate Analysis: Classification and Discrimination

Geoffrey McLachlan, University of Queensland, Brisbane, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In discriminant analysis, there are a number of distinct groups, and each entity under consideration is assumed to belong to one of them. Typically, the problem is to assign an unclassified entity to its group of origin on the basis of a feature vector, which consists of a number of measurements made on the entity. In order to construct an allocation rule for this assignment, there are so-called training data consisting of feature vectors on a number of entities that are classified with respect to the groups.

Introduction

The need for classification arises in most scientific pursuits. Typically, there is interest in 'classifying' an entity, say, an individual or object, on the basis of some characteristics (feature variables) measured on the entity. This classification is usually undertaken in two main contexts. In the first context, there is a finite number, say g, of predefined distinct populations, categories, classes, or groups, and the entity to be classified is assumed to belong to one (and only one) of these g possible groups. In order to assist with the construction of a classification rule for this purpose, there are usually available so-called training data from each group; that is, these training data comprise the features measured on some entities that are classified with respect to the g underlying groups. In statistical terminology, this classification process is referred to as *discriminant analysis*, while it is referred to as *supervised learning* or *learning with a teacher* in pattern recognition and machine learning.

In the second context, the classification of some entities is undertaken on the basis of their feature data in the absence of any prior knowledge on the existence of distinct groups. In this case, the primary purpose of the classification is to put the entities into a number of clusters so that entities within a cluster tend to have more similar features than those in different clusters. In statistics, this process is referred to as *cluster analysis*. In this article, attention is confined to discriminant analysis.

Discriminant Analysis

It is supposed that an entity belongs to one of g possible groups, denoted by $G_1, \ldots, G_g$, and that there is a (feature) vector $\boldsymbol{x}$ containing p variables that can be measured on the entity. The group-membership of the entity is denoted by the categorical variable z, where $z = i$ implies that the entity belongs to G_i $(i = 1, \ldots, g)$. In this notation, the topic of discriminant analysis may be viewed as being concerned with the relationship between z and the feature vector $\boldsymbol{x}$. Within this broad topic, there is a spectrum of problems that corresponds to the inference-decision spectrum in statistical methodology. At the decision end of the scale, the group-membership of the entity is unknown and the intent is to make an outright assignment of the entity to one of the g possible groups on the basis of $\boldsymbol{x}$. That is, in terms of the present notation, the problem is to estimate z solely on the basis of $\boldsymbol{x}$. An example in which an outright assignment is required concerns the rejection or acceptance of loan applicants by a financial institution. For this decision problem there are two groups with say G_1, referring to applicants who will service their loans satisfactorily, and G_2 to those who do not. The feature vector $\boldsymbol{x}$ for an applicant contains information such as age, income, and marital status. A rule based on $\mathbf{x}$ for allocating an applicant to either G_1 or G_2 (i.e., either accepting or rejecting the loan application) can be formed from an analysis of the feature vectors of past applicants from each of the two groups.

At the other extreme end of the spectrum, no assignment of the entity to one of the possible groups is intended. Rather the problem is to draw inferences about the relationship between z and the feature variables in $\boldsymbol{x}$. An experiment might be designed with the specific aim of providing insight into the predictive structure of the feature variables. For example, a political scientist may wish to determine the socioeconomic factors that have the most influence on the voting patterns of a population of voters. Between these extremes lie most of the everyday situations in which discriminant analysis is applied (Hand, 1997; McLachlan, 1992; Ripley, 1996).

Allocation Rules

Let $r(\boldsymbol{x})$ denote an allocation rule, where $r(\boldsymbol{x}) = i$ implies that an entity with feature vector $\boldsymbol{x}$ is to be assigned to the ith group G_i. The allocation rates associated with this rule $r(\boldsymbol{x})$ are denoted by $e_{ij}(r)$, where $e_{ij}(r)$ is the probability that a randomly chosen entity from G_i is allocated to G_j $(i, j = 1, \ldots, g)$. For a diagnostic test using the rule $r(\boldsymbol{x})$ in the context where G_1 denotes the absence of a disease or condition and G_2 its presence, the error rate $e_{12}(r)$ corresponds to the probability of a false positive, while $e_{21}(r)$ is the probability of a false negative. The correct allocation rates $e_{22}(r)$ and $e_{11}(r)$ are known as the sensitivity and specificity, respectively, of the diagnostic test.

Decision theory provides a convenient framework for the construction of discriminant rules in the situation where an outright allocation of an unclassified entity is required. For given $\boldsymbol{x}$, the expected loss or risk for allocation performed on the basis of the rule $r(\boldsymbol{x})$ is equal to $\sum_{i=1}^{g} \tau_i(\boldsymbol{x}) Q[i, r(\boldsymbol{x})]$ where,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.42150-1

for any u and v, $Q(u, v) = 0$ for $u = v$ and 1 for $u \neq v$, and where $\tau_i(\boldsymbol{x})$ is the posterior probability that the entity belongs to G_i which, by Bayes' theorem, can be expressed as

$$\tau_i(\boldsymbol{x}) = \pi_i f_i(\boldsymbol{x}) / f(\boldsymbol{x}) \quad (i = 1, \ldots, g) \qquad [1]$$

In [1], $f_i(\boldsymbol{x})$ denotes the density of the feature vector $\boldsymbol{X}$ in group G_i, $f(\boldsymbol{x}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{x})$ is the unconditional density of $\mathbf{X}$, and π_i denotes the prior probability that the entity belongs to G_i. The costs of misallocation have been assumed to be zero here for a correct allocation and unity for a misallocation from one group to another.

An optimal rule of allocation can be defined by taking it to be the one that minimizes the conditional risk at each value $\boldsymbol{x}$ of the feature vector. In decision-theory language, such a rule is said to be a Bayes' rule. It follows that the Bayes' rule $r_o(\boldsymbol{x})$ assigns the entity to the group to which it has the highest posterior probability of belonging under the specified utility function.

Sample-Based Allocation Rules

In practice, the group-conditional densities $f_i(\boldsymbol{x})$ are usually unknown. A basic assumption in discriminant analysis is that in order to estimate the unknown group-conditional densities there are entities of known origin on which the feature vector $\boldsymbol{X}$ has been recorded for each. These data are referred to in the literature as either the initial, reference, design, training, or learning data.

The initial approach to the problem of forming a sample discriminant rule, and indeed to discriminant analysis in its modern guise, was by Fisher (1936). In the context of $g = 2$ groups, he proposed that an entity with feature vector $\boldsymbol{x}$ be assigned on the basis of the linear discriminant function $a^T\boldsymbol{x}$, where a maximizes an index of separation between the two groups. The index was defined to be the magnitude of the difference between the group-sample means of $a^T\mathbf{x}$ normalized by the pooled sample estimate of its assumed common variance within a group.

To consider the formation of sample rules in general, we need to consider the sampling designs under which the training data $\boldsymbol{T}$ have been realized. There are two major schemes: joint or mixture sampling and z-conditional or separate sampling. They correspond respectively to sampling from the joint distribution of $\boldsymbol{Y} = (\boldsymbol{X}^T, \boldsymbol{Z}^T)^T$ and to sampling from the distribution of $\boldsymbol{X}$ conditional on z. The first design applies to the situation where the feature vector and group of origin are recorded on each of n entities drawn from a mixture of the possible groups. Mixture sampling is common in prospective studies and diagnostic situations. In a prospective study design, a sample of individuals is followed and their responses recorded. With separate sampling in practice, the feature vectors are observed for a sample of n_i entities taken separately from each group G_i, and so it is appropriate to retrospective studies.

An obvious way of forming a sample-based allocation rule $r(\boldsymbol{x}; t)$ is to take it to be an estimated version of the Bayes' rule $r_o(\boldsymbol{x})$, where the posterior probabilities of group-membership $\tau_i(\boldsymbol{x})$ are replaced by some estimates $\widehat{\tau}_i(\mathbf{x}; t)$ formed from the training data t. A common approach, referred to as the sampling approach, is to use the Bayes' formula [1] to formulate the $\tau_i(\boldsymbol{x})$ through the group-conditional densities $f_i(\boldsymbol{x})$. With the fully parametric approach to this problem, the group-conditional densities $f_i(\boldsymbol{x})$ are assumed to have specified functional forms except for a finite number of parameters to be estimated. A commonly used parametric family for the $f_i(\boldsymbol{x})$ for continuous feature data is the normal with either a linear or quadratic rule in $\mathbf{x}$ being obtained depending on whether the group-covariance matrices are taken to be equal or unequal. A compromise between linear and quadratic discriminant analyses is regularized discriminant analysis (Friedman, 1989).

There is also the direct approach (the diagnostic paradigm) to the estimation of the $\tau_i(\mathbf{x})$, using either nonparametric estimates as with nearest neighbor methods or parametric estimates via the logistic model. With the latter approach, it is the ratios of the group-conditional densities that are being modeled. The fundamental assumption of the logistic approach is that these ratios are linear, which is equivalent to taking the log(posterior) odds to be linear. The linearity here is not necessarily in the basic variables; transforms of these may be taken.

Another method for the direct modeling of the group-posterior probabilities is to use neural networks (Bishop, 1995), which have been coming under increasing attention by statisticians; see Ripley (1996). For example, with a two-layered feedforward neural network with m hidden units in the case of $g = 2$ groups, $\tau_1(\boldsymbol{x})$ is modeled as $\tau_1(\boldsymbol{x}) = \phi(w_{20} + \sum_{k=1}^{m} w_{2k} u_k)$ where $u_k = \phi(w_{10k} + \sum_{h=1}^{p} w_{1hk}(\boldsymbol{x})_h)$ is the value of the kth hidden unit, and the w_{1hk}, w_{2k} are parameters known as weights; $\phi(u)$ is the so-called activation function, often taken to be the sigmoid function, $\phi(u) = 1/1(+e^{-u})$. Neural networks are fitted mostly using *back propagation*, which is essentially gradient descent using the chain rule, but the EM algorithm has been used too recently. By the inclusion of an appropriate number of hidden units in the hidden layer (or layers), neural networks can handle nonlinear boundaries between the groups.

An advantage of the logistic and neural network-based rules is that they can handle discrete feature variables. Also, the multinomial distribution can be used to model the group-conditional probabilities for discrete feature variables. But unless the number m of multinomial cells is very small or the sample size n is extremely large, some of the cells will contain no observations and so some form of smoothing has to be applied. The nonparametric estimates of the multinomial cell probabilities under the assumption of independence of all the p feature variables may impose too high a degree of smoothing on the multinomial estimates. Hence, the aim of smoothing is to provide a degree of smoothing intermediate to that obtained under the extremes of the full multinomial and independence models.

To illustrate the construction of sample-based allocation rules, we consider the construction of a discriminant rule for the problem of assessing loan applications, as mentioned in Section Discriminant Analysis. Here the applicant for a loan is a company and the feature vector contains information on its key financial characteristics. For the purposes of our illustration, we consider the case where there are $p = 2$ feature variables that are the first two principal components of three

financial ratios concerning the debt, liquid assets, and working capital of a company. In Figure 1, we have plotted the boundary

$$\hat{\tau}_1(\boldsymbol{x}_j; t) = \hat{\tau}_2(\boldsymbol{x}_j; t)$$

for the sample version of the Bayes' rule (the sample linear discriminant rule) formed under the assumption that the feature vector has a bivariate normal distribution with a common group-covariance matrix in each of the two groups G_1 and G_2, corresponding to good (satisfactorily serviced) and bad (unsatisfactorily serviced) loans. It can be seen that the overall error rate of this rule reapplied to the training data (its apparent error rate) is not as low as that of the quadratic rule formed without the restriction of a common group-conditional covariance matrix for the feature vector $\boldsymbol{X}$. With its curved boundary, the quadratic rule misallocates only four (all bad) loans. One should be mindful, however, that the apparent error rate provides an optimistic assessment of the accuracy of a discriminant rule when it is applied to data not in the training set (see the next section). We have also plotted in Figure 1 the boundary of the rule obtained by modeling the distribution of the feature vector $\boldsymbol{X}$ by a two-component normal mixture, which results in the misallocation of one bad loan at the expense of misallocating two good loans. The latter rule is an example of a more flexible approach to classification than that based on the assumption of normality for each of the group-conditional distributions, as to be considered now.

The reader is referred to Geisser (1982) for a Bayesian approach to the formation of sample-based rules whereby prior distributions are adopted for the unknown parameters in the postulated model for the class-conditional distributions.

In some applications, the training data are only partially classified with respect to the classes, as considered, for example, in McLachlan (1975); see also McLachlan (1992: Section 2.7) and McLachlan and Peel (2000: Section 2.19). A case study may be found in McLachlan and Gordon (1989).

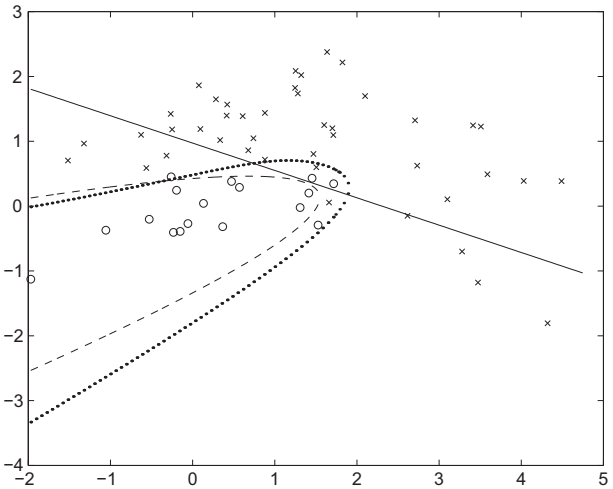

Figure 1 Boundaries of linear rule (solid line), quadratic rule (dashed line), and mixture model-based rule (dotted line); crosses and circles denote good and bad loans, respectively.

Flexible Discriminant Rules

A common nonparametric approach to discriminant analysis uses the kernel method to estimate the group-conditional densities $f_i(\boldsymbol{x})$ in forming an estimate of the Bayes' rule. More recently, use has been made of finite mixture models, mainly normal mixtures, to provide flexible rules of discrimination (Hastie and Tibshirani, 1996). Mixture models provide an extremely flexible way of modeling a density function (McLachlan and Basford, 1988; McLachlan and Peel, 2000) and can be fitted in a straightforward manner via the EM algorithm (McLachlan et al., 1999; McLachlan and Krishnan, 2008). Among other work on flexible discrimination, there is the FDA (flexible discriminant analysis) approach based on nonparametric regression (see Hastie and Tibshirani, 1996). The generic version of FDA based on smoothing splines proceeds by expanding the predictors in a large (adaptively selected) basis set, and then performing a penalized discriminant analysis in the enlarged space using a linear discriminant or a normal mixture model-based rule. The class of nonlinear regression methods that can be used include additive models, the multivariate adaptive regression spline (MARS) model, projection pursuit regression, and neural networks. In machine learning, there has been increasing attention in the case of two groups given to nonlinear rules based on the foundations of support vector machines, as developed by Vapnik (2000). With this approach, the initial feature space is mapped into a higher dimensional space by choosing a nonlinear mapping and then choosing an optimal separate hyperplane in the enlarged feature space.

A rather different approach to the allocation problem as considered up to now is to portray the rule in terms of a binary tree. The tree provides a hierarchical representation of the feature space. An allocation is effected by proceeding down the appropriate branches of the tree. The Classification and Regression Tree (CART) methodology of Breiman et al. (1984) has contributed significantly to the growing popularity of tree classifiers; see also the C4.5 decision-tree learning algorithm of Quinlan (1993). In the context of tree classifiers in particular, there has been growing interest in the use of boosting, which is one of the most important recent developments in classification methodology. Algorithms like the Adaboost algorithm of Freund and Schapire (1996) and the bagging algorithm of Breiman (1996) can often improve performance of unstable classifiers like trees or neural nets by sequentially applying them to reweighted versions of the training data and taking a weighted majority vote of the sequence of classifiers so formed. The test error of the weighted classifier usually does not increase as its size increases, and often is observed to decrease even after the training error reaches zero. Friedman et al. (2001) have exhibited the link of boosting to statistical ideas on additive fitting.

An excellent summary of available software and algorithms for discriminant analysis has been given by the Panel on Discriminant Analysis, Classification and Clustering (1989); see also Huberty (1994) and Michie et al. (1994).

Feature Selection

$$\mathbf{S} = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} \left(y_j - \bar{y}_i \right) \left(y_j - \bar{y}_i \right)^T \Big/ \left(n - g \right) \qquad [2]$$

There is a vast literature on variable selection in discriminant analysis. Some of the methods are applicable only if the number of variables p is well below the sample size n; see, for example, Maugis et al. (2011) in which a general methodology is proposed for selecting predictors for normal class-conditional distributions. When the dimension p is large, some of the commonly used classifiers might be unreliable or not able to be formed in the first instance. For example, Fisher's linear discriminant function requires taking the inverse of the pooled within-class covariance matrix $\mathbf{S}$ as defined by eqn [2]. This is only possible if $p \leq n - 2$. One way to avoid this is to take the class-covariance matrices to be diagonal; that is, to use what is termed as 'naive Bayes' (Hand and Yu, 2001). Bickel and Levina (2004) show that this latter rule can be used effectively in such situations. More recently, Professor Fan and colleagues have demonstrated that if p is ultra high compared to n, then the naive Bayes rule can perform as bad as random guessing in certain situations where the signal-to-noise ratio is weak due to the noise accumulation in the estimation of the class means; see Fan and Lv (2008, 2010) and the references therein. The L_1 or Lasso penalty proposed in Tibshirani (1996) is often applied to penalize the likelihood function in situations where p is very large; see Efron et al. (2004). Recently, Witten and Tibshirani (2011) have proposed a penalized form of Fisher's linear discriminant function by imposing a penalty on the discriminant vectors.

One commonly used method of feature selection is based on the ratio of the between-class sum of squares to the within-class sum of squares. There are other methods such as selecting the variables on the relative size of their weights in the linear kernel in the formation of the support vector machine. Guyon et al. (2002) proposed that the feature variables be selected according to a backward elimination process, which they call recursive feature elimination (RFE). It starts by considering all available features. Each step consists in ranking the features according to the order of magnitude of their associated coefficients (weights) β_i, and then discarding the bottom-ranked variables. The SVM is then refitted to the remaining variables, which are then reranked according to their new weights. Again, the bottom-ranked variables are discarded, and so on. The process can be refined at the expense of greater computational time by deleting fewer variables on each step. One suggested procedure is to apply RFE by removing chunks of features in the first few iterations and then removing one feature at a time once the feature set size reaches a few hundreds.

Projection methods are also very popular for reducing the dimension of the feature space and include matrix factorization (MF) with orthogonality constraints as in principal component analysis (PCA) and singular-value decomposition (SVD) and matrix factorization without orthogonality constraints as in general MF (Nikulin et al., 2011). These projection methods can be implemented in conjunction with penalty constraints as considered by Witten et al. (2009) There is also partial least squares (PLS) as adopted, for example, by Nguyen and Rocke (2002).

There is much ongoing research in the area of variable selection in discriminant and cluster analyses in the case where p is much greater than n. This is because traditional methods are faced with formidable challenges as exemplified by the data piling problems in high-dimensional space observed in Hall et al. (2005); see also Fan and Lv (2008, 2010); Fan et al. (2009). The impact of high dimensionality for distance-based classifiers has been studied by Hall et al. (2008).

Error-Rate Estimation

A guide to the performance of a given discriminant rule can be obtained by applying it to the training from which it was formed and noting the proportion of these observations misallocated (the apparent rate). However, it is well known that the apparent error rate gives too optimistic a view on how the rule might perform on future cases; see McLachlan (1992: Chapter 10) for an account of error-rate estimation. One way of correcting for this bias is to use cross-validation, but it produces an estimate that often is more highly variable than the apparent error rate. This led Efron (1983) to consider his so-called 0.632 estimator, which is a weighted sum of the apparent error rate and the bootstrap error rate at an original data point not in the training set t. Efron (1983) showed that the latter error rate is almost the same as the twofold (half-sample) cross-validated error rate, which previously McLachlan (1977a) had weighted with the apparent error rate in proportions w and $(1 - w)$ with the apparent error rate in order to reduce its bias. Under certain conditions in the case of normality, it was shown that the value of w ranged from 0.6 to 0.7 in order to achieve asymptotically zero bias. Thus, it is very similar to the 0.632 estimator. Efron and Tibshirani (1997) proposed an improved version of the 0.632 estimator known as the 0.632+ estimator. The bootstrap methodology can be used also to estimate the variability in the errors and in other quantities of interest, such as the estimated parameters in the form of the estimated rule.

Conclusion

Discriminant analysis is concerned with the construction of allocation rules for the assignment of unclassified entities to distinct groups in the situation where there are data on entities of known group of origin (training data). In the past, linear rules such as those based on Fisher's linear discriminant function were commonly used in practice. In recent times, with the wide availability of high-speed computers, there has been increasing attention given to the development of more complex nonlinear rules such as those based on flexible parametric models, neural networks, support vector machines, and classification trees.

See also: Bayesian Statistics; Functional Data Analysis; Mixture Models in Statistics; Probability Density Estimation; Screening and Selection.

Bibliography

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

Bickel, P., Levina, E., 2004. Some theory for Fisher's linear discriminant function, "naive Bayes," and some alternatives when there are many more variables than observations. Bernoulli 10, 989–1010.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, California.

Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. Journal of the American Statistical Association 78, 316–331.

Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: the .632+ bootstrap method. Journal of the American Statistical Association 92, 548–560.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression (with discussion). Annals of Statistics 32, 407–499.

Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B 70, 849–911.

Fan, J., Lv, J., 2010. A selective review of variable selection in high dimensional feature space. Statistica Sinica 20, 101–148.

Fan, J., Samworth, R., Wu, Y., 2009. Ultrahigh dimensional feature selection: beyond the linear model. Journal of Machine Learning Research 10, 122013–122038.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188.

Freund, Y., Schapire, R., 1996. Experiments with a new boosting algorithm. In: Saitta, L. (Ed.), Machine Learning: Proceedings of the Thirteenth International Conference. Morgan Kaufmann, San Francisco, pp. 148–156.

Friedman, J.H., 1989. Regularized discriminant analysis. Journal of the American Statistical Association 84, 165–175.

Friedman, J.H., Hastie, T., Tibshirani, R., 2001. Additive Logistic Regression: A Statistical View of Boosting. Technical Report. Department of Statistics, Stanford University, Stanford.

Geisser, S., 1982. Bayesian discrimination. In: Krishnaiah, P.R., Kanal, L. (Eds.), Handbook of Statistics, Classification, Pattern Recognition and Reduction in Dimensionality, vol. 2. North-Holland, Amsterdam, pp. 101–120.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine Learning 46, 389–422.

Hall, P., Marron, J.S., Neeman, A., 2005. Geometric representation of high dimension, low sample size data. Journal of the Royal Statistical Society: Series B 67, 427–444.

Hall, P., Pittelkow, Y., Ghosh, M., 2008. Theoretic measures of relative performance of classifiers for high-dimensional data with small sample sizes. Journal of the Royal Statistical Society: Series B 70, 158–173.

Hand, D.J., 1997. Construction and Assessment of Classification Rules. Wiley, New York.

Hand, D.J., Yu, K., 2001. Idiot's Bayes – not so stupid after all? International Statistical Review 69, 385–398.

Huberty, C.J., 1994. Applied Discriminant Analysis. Wiley, New York.

Hastie, T., Tibshirani, R.J., 1996. Discriminant analysis by Gaussian mixtures. Journal of the Royal Statistical Society: Series B 58, 155–176.

Maugis, C., Celeux, G., Martin-Magniette, M.-L., 2011. Variable selection in model-based discriminant analysis. Journal of Multivariate Analysis 102, 1374–1387.

McLachlan, G.J., 1975. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. Journal of the American Statistical Association 70, 365–369.

McLachlan, G.J., 1977a. A note on the choice of a weighting function to give an efficient method for estimating the probability of misclassification. Pattern Recognition 8, 147–149.

McLachlan, G.J., 1977b. Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. Journal of the American Statistical Association 72, 403–406.

McLachlan, G.J., 1992. Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York.

McLachlan, G.J., Gordon, R.D., 1989. Mixture models for partially unclassified data: a case study of renal venous renin levels in essential hypertension. Statistics in Medicine 8, 1291–1300.

McLachlan, G.J., Basford, K.E., 1988. Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York.

McLachlan, G.J., Krishnan, T., 2008. The EM Algorithm and Extensions, second ed. Wiley, Hoboken, New Jersey.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.

McLachlan, G.J., Peel, D., Basford, K.E., Adams, P., 1999. The EMMIX software for the fitting of mixtures of normal and t-components. Journal of Statistical Software 4 (2).

Michie, D., Spiegelhalter, D.J., Taylor, C.C. (Eds.), 1994. Machine Learning, Neural and Statistical Classification. Ellis Horwood, Chichester.

Nguyen, D.V., Rocke, D.M., 2002. Multiclass cancer classification via partial least squares with gene expression profiles. Bioinformatics 18, 1216–1226.

Nikulin, V., Huang, T.-H., Ng, S.K., Rathnayake, S.I., McLachlan, G.J., 2011. A very fast algorithm for matrix factorization. Statistics & Probability Letters 81, 773–782.

Panel on Discriminant Analysis, Classification, and Clustering, 1989. Discriminant analysis and clustering. Statistical Science 4, 34–69.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo.

Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge Press, Cambridge.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B 58, 267–288.

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory, second ed. Springer, New York.

Witten, D., Tibshirani, R., 2011. Penalized classification using Fisher's linear discriminant. Journal of the Royal Statistical Society: Series B 73, 753–772.

Witten, D., Tibshirani, R., Hastie, T., 2009. A penalized matrix decomposition with applications to sparse principal components and canonical correlation analysis. Biostatistics 10, 515–534.

Multivariate Analysis: Discrete Variables (Correspondence Models)

Z Gilula, Hebrew University, Jerusalem, Israel
Shelby J Haberman, Northwestern University, Evanston, IL, USA
Peter GM van der Heijden, Utrecht University, Utrecht, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 10218–10221, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Correspondence models are a special class of statistical models for the association between categorical variables. The specific parametric structure of such models is based, as in the common (descriptive) correspondence analysis, on the canonical form of joint discrete distributions. Bivariate correspondence models are considered first, based on a simple canonical form involving canonical correlations. Such models are discussed in great detail, where interpretation of parameters and inferential aspects receive special attention. Joint multivariate distributions do not have a universal canonical form, and correspondence models for the multivariate case require therefore a modified methodology. Such methodology is presented next, in which discussion on multivariate correspondence models is divided into two cases. One case is where the variables can be partitioned into two natural groups of variables (response and explanatory, say); the second case is where all variables are treated on an equal footing. Correspondence models for three variables are discussed in detail as a special case from which extensions to a higher number of variables is relatively straightforward. Although the main concern here is with inference from correspondence probability models, attention is also given to the use of the parameters of such models as a descriptive tool.

Introduction

Correspondence models extract information on the association between categorical variables. As in correspondence analysis (*see* Correspondence Analysis), correspondence models are based on a canonical parametrization of the joint distribution of the variables. Traditional correspondence analysis (Benzécri 1973, Greenacre 1984, Hill 1974, Nishisato 1980, Gifi 1990) is concerned with methods used for description and detection of patterns of association between variables via graphical displays, and does not consider the role of random sampling in the generation of the observed data from a population with an associated probability model. In contrast, correspondence models are explicitly concerned with inferential methods for the study of probability models which are fitted to observed data under the assumption that the data have been obtained by random sampling (Goodman 1985, 1986, Gilula and Haberman 1986, 1988). The parameterizations used in correspondence models may still be used to construct descriptive measures and graphs; however, unlike in traditional correspondence analysis, correspondence models involve formal statistical methods such as tests of goodness of fit and approximate confidence intervals for parameters. Development of correspondence models is most straightforward in the case in which the relationship of two categorical random variables is studied (Section Bivariate Correspondence Models). When more than two variables are involved, correspondence models may involve a division of the categorical variables into two groups (Section Two Groups of Categorical Variables). Correspondence models based on multiple correspondence analysis are much more difficult to construct (Section Multiple Correspondence Analysis).

Bivariate Correspondence Models

Let X and Y be two categorical random variables with respective ranges consisting of I and J categories indexed from 1 to I, and from 1 to J, respectively. The joint probability distribution of X and Y may be characterized by the I by J matrix p^{XY} of joint probabilities $p_{ij}^{XY} = p(X = i, Y = j), \quad 1 \le i \le I, \quad 1 \le j \le J$. The marginal probability distributions of X and Y may be characterized by the marginal probabilities $p_i^x = P(X = i), 1 \le i \le I$, and $p_j^y = P(Y = j), \quad 1 \le j \le J$. For simplicity, assume that $p_i^x > 0$ for $1 \le i \le I$, and $p_j^y > 0$ for $1 \le j \le J$. The entire family of correspondence models for bivariate distributions is based on the canonical decomposition (Hirschfeld 1935, Fisher 1940)

$$p_{ij}^{XY} = p_i^X p_j^Y \left(1 + \sum_{m=1}^{M} \rho_{im}\mu_{im}\nu_{jm}\right) \qquad [1]$$

where $M = \min(I-1, J-1), \quad 1 \ge \rho_1 \ge \rho_2 \ldots \ge \rho_M \ge 0$

$$\sum_{i=1}^{I} p_i^X \mu_{im} = \sum_{j=1}^{J} p_j^Y \nu_{im} = 0, \quad 1 \le m \le M \qquad [2]$$

and

$$\sum_{i=1}^{I} p_i^X \mu_{im}\mu_{im'} = \sum_{j=1}^{J} p_j^Y \nu_{im'} = \begin{cases} 1, & 1 \le m = m' \le M \\ 0, & 1 \le m \le m' \le M \end{cases} \qquad [3]$$

In (1), (2), and (3), for each integer m from 1 to M, the parameter ρ_m is the mth canonical correlation, the score μ_{im} is the mth canonical score of category i of variable X, $1 \le i \le I$, and the score ν_{jm} is the mth canonical score of category j of variable Y. Let the mth canonical variable μ_{Xm} associated with X be the random variable with value μ_{im} for $X = i$, and let the mth canonical variable μ_{Ym} associated with Y denote the random variable with value ν_{jm} for $Y = j$. Then μ_{Xm} and ν_{Ym} both have mean 0 and variance 1, and ρ_m is the correlation of the canonical variables μ_{Xm} and ν_{Ym}. In addition, if $1 \le m < m' \le M$, then the canonical variables μ_{Xm} and $\mu_{Xm'}$ of X are uncorrelated, and the canonical variables ν_{Ym} and $\nu_{Ym'}$ of Y are uncorrelated.

For some M^* between 1 and M, graphical display of category i of X may be based on the observed scores μ_{im} for $1 \leq m \leq M^*$ and graphical display of category j of Y may be based on the observed scores ν_{jm} for $1 \leq m \leq M^*$. An alternative approach uses graphical displays for row category i based on the correspondence scores $\rho_m \mu_{im}$ for $1 \leq m \leq M^*$ and graphical displays for column category j based on the correspondence scores $\rho_m \nu_{jm}$ for $1 \leq m \leq M^*$.

Formal inference for canonical models has generally been based on maximum likelihood. Computation of maximum-likelihood estimates has been attempted by use of a cyclic algorithm (Goodman, 1985) and by use of a scoring algorithm for parameters subject to constraints (Gilula and Haberman, 1986). Goodness of fit has typically been examined by use of likelihood-ratio and Pearson chi-square statistics. In typical cases, a chi-square approximation can be applied when the model is true. If the model only assumes that $\rho_m = 0$ for $m > M^*$, then the chi-square approximations have $(I\text{-}M^*\text{-}1)(J\text{-}M^*\text{-}1)$ degrees of freedom.

As a statistical model, (1) is a *saturated* model, for no restrictions are imposed on the joint probabilities p_{ij}^{XY}. Goodman (1985, 1986) and Gilula and Haberman (1986) consider models in which the canonical correlations ρ_m are assumed 0 for $m > M^*$ for some M^* between 0 and M. The case of $M^* = 0$ leads to the conventional model of independence of the variables X and Y. The case of $M^* = M$ is the saturated model in which no restrictions are imposed. The most common nontrivial case has $M^* = 1$. Models are also considered in which linear restrictions are imposed on the scores μ_{im} and ν_{jm} for $m \leq M^*$.

As an example, consider the data in Table 1 (Guttman 1971) on the origin of Israeli adults versus their principal worry. Several cells in the table have small values, so some care must be exercised in interpretation of customary large-sample approximation. Nonetheless, it is worth noting that the model with a single nonzero canonical correlation ($\rho_m = 0$ for $m > 1$) yields a likelihood-ratio chi-square of 29.35 on 18 degrees of freedom, so that the fit is relatively poor. On the other hand, the model with two nonzero canonical correlations ($\rho_m = 0$ for $m > 2$) yields a likelihood-ratio chi-square of only 6.49 on 10 degrees of freedom, so that the fit is quite satisfactory. The estimated canonical correlations are 0.245 and 0.124. These results suggest a modest relationship between the variables under study. The maximum-likelihood estimates of the canonical scores are provided in Table 2. Figure 1 provides correspondence scores for the column variable, and Figure 2 provides correspondence scores for the row variable. Results suggest a contrast between personal worries and other concerns, especially political ones, with subjects of Asian or African background more likely to express personal worries than subjects of European or American background. The second dimension of background appears to involve a contrast between subjects born in Israel with fathers of European or American background and other subjects. Subjects born in

Table 1 Cross-classification of origin by principal worry

Origin	*Principal worry*[a]							
	ENR	SAB	MIL	POL	ECO	OTH	MTO	PER
A/A	61	70	97	32	4	81	20	104
E/A	104	117	218	118	11	128	42	48
fA/A	8	9	12	6	1	14	2	14
fE/A	22	24	28	28	2	52	6	16
fI/I	5	7	14	7	1	12	0	9

[a]Columns are: ENR, enlisted relative; SAB, sabotage; MIL, military situation; POL, political situation; ECO, economic situation; OTH, other; MTO, more than one worry; and PER, personal economics. Rows are: A/A, Africa or Asia; E/A, Europe or America; fA/A, father from Africa or Asia; fE/A, father from Europe or America; and fI/I, father and self from Israel.
Source: Guttman, L., 1971. Measurement as structural theory. Psychometrika 36, 329–347.

Table 2 Estimated canonical parameters for model with two pairs of scores

Principal worry	*Score 1*	*Score 2*
ENR	−0.0573	−0.2380
SAB	−0.0096	−0.2880
MIL	−0.5121	−1.0690
POL	−1.2270	0.6685
ECO	−0.6866	0.0811
OTH	0.0692	1.7794
MTO	−0.5203	−1.1601
PER	2.4428	−0.2684
Origin	*Score 1*	*Score 2*
A/A	1.3472	−0.4105
E/A	−0.8663	−0.4678
fA/A	1.1632	0.7179
fE/A	−0.2830	2.6321
fI/I	0.4120	0.8064

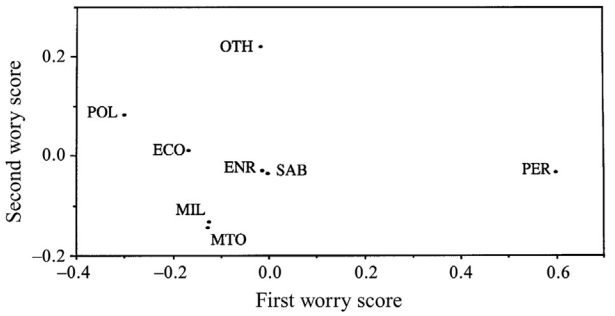

Figure 1 Correspondence analysis scores for worry.

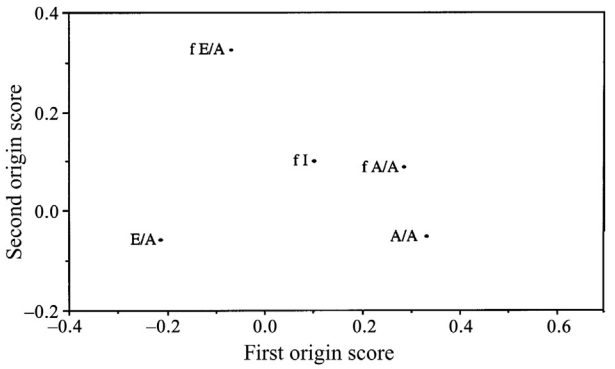

Figure 2 Correspondence analysis scores for origin.

Israel with fathers from Europe or America are more likely to cite more than one worry than are other subjects.

Two Groups of Categorical Variables

Extension of correspondence models to apply to more than two categorical variables is a nontrivial matter. The simplest approach is appropriate if the variables can be divided into two groups (Gilula and Haberman, 1988). In this case, each group of variables can be regarded as a categorical variable with many categories, and correspondence analysis methods for the bivariate case can be applied. For example, in addition to the categorical random variables X and Y of Section Bivariate Correspondence Models let an additional categorical random variable Z be defined with values from 1 to $K \geq 2$. As an example, Gilula and Haberman (1988) consider an example with X equal to education level, Y equal to religion, and Z equal to attitude toward legal nontherapeutic abortions. Each variable in this example involves a classification into three categories, so that I, J, and K are 3. In general, let p_{ijk}^{XYZ} be the joint probability that $X = i$, $Y = j$, and $Z = k$. Then

$$p_{ijk}^{XYZ} = p_{ij}^{XY} p_k^Z \left(1 + \sum_{m=1}^{M} \rho'_m \mu'_{ijm} \nu'_{km}\right), \qquad [4]$$

where M is the minimum of IJ-1 and K-1, the ρ'_m are nonnegative and nonincreasing in m,

$$\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij}^{XY} \mu'_{ijm} = \sum_{k=1}^{K} p_k^Z \nu'_{km} = 0, \quad 1 \leq m \leq M, \qquad [5]$$

and

$$\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij}^{XY} \mu'_{ijm} \mu'_{ijm} = \sum_{k=1}^{K} p_k^Z \nu'_{km'} = \begin{cases} 1, & 1 \leq m = m' \leq M, \\ 0, & 1 \leq m < m' \leq M. \end{cases} \qquad [6]$$

Various restrictions may be imposed on the canonical parameters. For example, if Y and Z are conditionally independent given X, then $\rho_m = 0$ for any $m > (-1)$ and μ'_{ijm} depends only on i for $\rho_m > 0$. For other examples, see Gilula and Haberman (1988).

Multiple Correspondence Analysis

Multiple correspondence analysis, another approach to extension of correspondence analysis to the study of two or more categorical variables, appears in Guttman (1941). Multiple correspondence analysis can be regarded as a special case of correspondence analysis. For illustrative purposes, the case of two and three variables will be considered.

Consider the case of two categorical random variables X and Y defined as in Section Bivariate Correspondence Models. In a multiple correspondence analysis of X and Y, a canonical analysis is used to predict the joint variable (X, Y) from either X or Y, where X or Y is chosen at random. Formally, (X, Y) is predicted by the pair (U, V), where V is a random variable independent of X and Y such that V assumes the values 1 and 2 with equal probability. The variable U is X if V is 1, and U is Y if V is 2. when $V = 1$, then $U = X$, so that prediction of (X, Y) reduces to prediction of Y by X. When $V = 2$, then $U = Y$, so that prediction of (X, Y) reduces to prediction of X by Y. The canonical analysis in this case is closely related to conventional canonical analysis of the variables X and Y.

For multiple correspondence analysis of the three categorical variables X, Y, and Z of Section Two Groups of Categorical Variables the joint variable (X, Y, Z) is predicted by random choice of a marginal variable X, Y, or Z. In this case, let V be a random variable independent of X, Y, and Z which assumes value k with probability $^1/_3$ for $1 \leq k \leq 3$. Let U be the random variable which is X for $V = 1$, Y for $V = 2$, and Z for $V = 3$. Multiple correspondence analysis considers correspondence analysis of (X, Y, Z) and U. If $V = 1$, then $U = X$ and prediction of (X, Y, Z) reduces to prediction of (Y, Z) by X. If $V = 2$, then $U = Y$ and prediction of (X, Y, Z) reduces to prediction of (X, Z) by Y. If $V = 3$, then $U = Z$ and prediction of (X, Y, Z) reduces to prediction of (X, Y) by Z. The analysis differs from that in Section Two Groups of Categorical Variables in that it is symmetrical in the three variables. Thus in the example of Gilula and Haberman (1988), the analysis looks at educational level, religion, and abortion attitude symmetrically. This approach may or may not be appropriate, depending on the context. For instance, one may reasonably argue that abortion attitude reflects educational and religious background, but neither educational background nor religious background is appreciably affected by abortion attitude.

Conclusion

Correspondence models compete in practice with a number of other statistical methods in common use. In their development of canonical models, both Goodman (1985) and Gilula and Haberman (1986) compare canonical models with association models in the case of two variables. Here association models involve a loglinear parameterization similar to the parameterization used in correspondence models. Interpretations involve loglinear interactions rather than the correlations used in correspondence models, but it remains the case that scores are attached to variable categories. As noted in both papers, association models in which scores are fixed rather than estimated are special cases of conventional log-linear models (*see* Multivariate Analysis: Discrete Variables (Loglinear Models)). Goodman (1985) also discusses the relationship of parameters in canonical models to parameters in correspondence models. Gilula and Haberman (1986) and van der Heijden, Gilula and van der Ark (1999) consider the relationship of canonical models to latent-class models (*see* Factor Analysis and Latent Variable Models in Personality Psychology). In the case in which more than two variables are analyzed by division of the variables into two groups, Gilula and Haberman (1988) compare canonical models to association models in cases in which more than two variables are present.

The canonical correlations used in bivariate correspondence analysis may be used to derive formal tests of independence that compete with conventional chi-square statistics and may

be employed to ascertain the nature of departures from independence. For details, see Haberman (1981).

A variety of alternative interpretations of multiple correspondence analysis are available. For example, see Tenenhaus and Young (1985) and Gifi (1990).

See also: Factor Analysis and Latent Variable Models in Personality Psychology; Multivariate Analysis: Discrete Variables (Loglinear Models); Multivariate Analysis: Discrete Variables (Overview).

Bibliography

Barnard, Kobus, Duygulu, Pinar, Forsyth, David, De Freitas, Nando, Blei, David M., Jordan, Michael I., 2003. Matching Words and Pictures. The Journal of Machine Learning Research 3, 1107–1135.

Benzécri, J.-P., 1973. L'Analyse des Données: 2. L'Analyse de Correspondances. Dunod, Paris.

Breiden, Oliver, Mohr, Alexander T., Mirza, Hafiz R., 2006. An Empirical Test of a Correspondence Model of Expatriate Managers' Work Adjustment. The International Journal of Human Resource Management 17 (11), 1907–1925.

Fisher, R.A., 1940. The precision of discriminant functions. Annals of Eugenics 10, 422–429.

Gifi, A., 1990. Non-linear Multivariate Analysis. Wiley, New York.

Gilula, Z., Haberman, S.J., 1986. Canonical analysis of contingency tables by maximum likelihood. Journal of the American Statistical Association 81, 780–786.

Gilula, Z., Haberman, S.J., 1988. The analysis of multivariate contingency tables by restricted canonical and restricted association models. Journal of the American Statistical Association 83, 760–771.

Goodman, L.A., 1985. Analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models, and asymmetry models for contingency tables with or without missing entries. Annals of Statistics 13, 10–69.

Goodman, L.A., 1986. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. International Statistical Review 54, 243–309.

Greenacre, M.J., 1984. Theory and Applications of Correspondence Analysis. Academic Press, London.

Guttman, L., 1941. The quantification of a class of attributes: a theory and method of scale construction. In: Horst, P. (Ed.), The Prediction of Personal Adjustment. Social Science Research Council, New York, pp. 319–348.

Guttman, L., 1971. Measurement as structural theory. Psychometrika 36, 329–347.

Haberman, S.J., 1981. Tests for independence in two-way contingency tables based on canonical correlation and on linear-by-linear interaction. Annals of Statistics 9, 1178–1186.

Hill, M.O., 1974. Correspondence analysis: a neglected multivariate method. Applied Statistics 23, 340–354.

Hirschfeld, H.O., 1935. A connection between correlation and contingency. Proceedings of the Cambridge Philosophical Society 31, 520–524.

Nishisato, S., 1980. Analysis of Categorical Data: Dual Scaling and Its Applications. University of Toronto Press.

Tenenhaus, M., Young, F.W., 1985. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling homogeneity analysis and other methods for quantifying categorical multivariate data. Psychometrika 50, 91–119.

Van der Heijden, P.G.M., Gilula, Z., van der Ark, L.A., 1999. An extended study into the relationships between correspondence analysis and latent class analysis. In: Sobel, M., Becker, M. (Eds.), Sociological Methodology. Blackwell, Cambridge, pp. 147–186.

Zuur, Alain F., Ieno, Elena N., Smith, Graham M., 2007. Analysing ecological data. Springer, New York.

Multivariate Analysis: Discrete Variables (Loglinear Models)

Shelby J Haberman, Educational Testing Service, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Loglinear models are commonly used to analyze relationships involving discrete variables. They include models for independence and conditional independence of polytomous variables, logit models for prediction of binary variables, multinomial response models for prediction of multinomial variables, and Poisson regression models. On the other hand, loglinear models are special cases of exponential families. A unified treatment of loglinear models may be achieved by description of data concerning discrete variables in terms of tables of frequencies. In a loglinear model, each frequency is a random variable with a finite and positive expectation, and the logarithms of the expectations of the frequencies are assumed to satisfy a linear model. The frequencies are assumed to be either independent Poisson random variables or one or more independent multinomial random variables. In loglinear models, exact tests and confidence intervals are available, although such procedures are often difficult to apply. Large-sample approaches based on maximum likelihood are commonly used, and analyses based on weighted regression analysis are encountered in many important special cases. Bayesian methods of analysis have also been proposed.

Introduction

Loglinear models are a class of probability models commonly used to analyze relationships involving discrete variables. They include models for independence and conditional independence of polytomous variables, logit models for prediction of binary variables, multinomial response models for prediction of multinomial variables, and Poisson regression models. They are used both in contingency tables in which all variables under study are categorical, and in cases in which responses are categorical variables but some predicting variables may be continuous. Although numerous probability models in the statistical literature prior to 1969 are examples of loglinear models, an explicit description of a probability model as a loglinear model is not readily found in statistical journals or published books prior to Bishop (1969) and Bishop and Fienberg (1969).

The cell frequencies are random variables with positive and finite expected values, and the logarithms of the expected values of the cell frequencies satisfy a linear model. The frequencies may be independent Poisson random variables (Poisson sampling) or they may have a joint multinomial distribution with known sample size (multinomial sampling). In the case of multinomial sampling, to each cell in the table of frequencies corresponds a cell probability. The linear model for the logarithms of the expected values of the cell frequencies is defined in such a manner that the logarithms of the cell probabilities satisfy the same linear model. In a more complex case (product multinomial sampling), the table of frequencies consists of independently distributed subtables, with each subtable having an independent multinomial distribution with known sample size. Again, each cell in the table of frequencies corresponds to a cell probability for a multinomial distribution, and the linear model is defined so that the logarithms of the cell probabilities satisfy the same linear model as do the logarithms of the expected values of the cell frequencies.

Examples

Several examples of loglinear models provide an indication of their scope.

Example 1. The additive loglinear model for a two-way contingency table. As in Bishop and Fienberg (1969), consider a contingency table with frequencies n_{ij}, $1 \le i \le I$, and $1 \le j \le J$, where there are $i \ge 2$ rows and $j \ge 2$ columns in the table. Let each frequency n_{ij} be a random variable with expectation $m_{ij} > 0$. In an additive loglinear model, it is assumed that the logarithms $\mu_{ij} = \log m_{ij}$ satisfy the additive model

$$\mu_{ij} = \lambda + \lambda_i^A + \lambda_j^B \qquad [1]$$

for $1 \le i \le I$ and $1 \le j \le J$, where λ, λ_i^A, and λ_j^B are unknown parameters such that

$$\sum_{i=1}^{I} \lambda_i^A = \sum_{j=1}^{J} \lambda_i^B = 0 \qquad [2]$$

The constraints in [2] are imposed to identify the parameters λ, λ_i^A, and λ_j^B. The parameter λ is the average of μ_{ij} for $1 \le i \le I$ and $1 \le j \le J$. If $\mu_{i.}$ is the average of μ_{ij} for $1 \le j \le J$ and $\mu_{.j}$ is the average of μ_{ij} for $1 \le i \le I$, then $\lambda_i^A = \mu_{i.} - \lambda$ and $\lambda_j^B = \mu_{.j} - \lambda$. Interpretation of this model depends on the sampling assumptions concerning the frequencies. In the multinomial case, the n_{ij} have a joint multinomial distribution with known sample size $N > 0$ and unknown probabilities $p_{ij} > 0$. Thus, $m_{ij} = N\, p_{ij}$, $\log m_{ij} = \log N + \log p_{ij}$, and

$$\log p_{ij} = \kappa + \lambda_i^A + \lambda_j^B \qquad [3]$$

where $\kappa = \lambda - \log N$. Thus, the logarithms $\log m_{ij}$ of the cell frequencies satisfy the same additive loglinear model as the logarithms $\log p_{ij}$ of the cell probabilities. The probabilities p_{ij} define the joint probability distribution of a row variable A and a column variable B such that $1 \le i \le I$ and $1 \le j \le J$ and p_{ij} is the probability that $A = i$ and $B = j$. The additive model is then equivalent to the model that A and B are independent, so that

$p_{ij} = p_i^A p_j^B$, where p_i^A is the marginal probability that $A = i$ and p_j^B is the marginal probability that $B = j$. If μ^A is the average of the logarithms $\log p_i^A$ for $1 \le i \le I$, then λ_i^A is the difference between $\log p_i^A$ and μ^A. If μ^B is the average of the logarithms $\log p_j^B$ for $1 \le j \le J$, then λ_j^B is the difference between $\log p_j^B$ and μ^B.

In the row-multinomial case, the I rows n_{ij}, $1 \le j \le J$, are mutually independent for $1 \le i \le I$, and the ith row has a multinomial distribution with known sample size $N_i^A > 0$ and probabilities $p_{j\cdot i}^{B\cdot A} > 0$. The mean $m_{ij} = N_i^A p_{j\cdot i}^{B\cdot A}$, so that

$$\log p_{j\cdot i}^{B\cdot A} = \kappa + \kappa_i^A + \lambda_j^B \qquad [4]$$

where

$$\kappa = \lambda - \eta, \quad \eta = I^{-1}\sum_{i=1}^{I} \log N_i^A \qquad [5]$$

and

$$\kappa_i^A = \lambda_i^A - \log N_i^A + \eta \qquad [6]$$

Once again, the logarithms $\log p_{j\cdot i}^{B\cdot A}$ of the probabilities satisfy the same additive linear model as the logarithms $\log m_{ij}$ of the expected values of the cell frequencies. The additive model is equivalent to the homogeneity model that, for any fixed column j, $p_{j\cdot i}^{B\cdot A}$ is a constant p_j^B for each column j. The parameter λ_j^B is defined from p_j^B as in the multinomial case.

In the column-multinomial case, J columns n_{ij}, $1 \le i \le I$, are mutually independent for $1 \le i \le I$, and the jth column has a multinomial distribution with sample size $N_j^B > 0$ and probabilities $p_{i\cdot j}^{A\cdot B} > 0$. The mean $m_{ij} = N_j^B p_{i\cdot j}^{A\cdot B}$, so that

$$\log p_{i\cdot j}^{A\cdot B} = \kappa + \lambda_i^A + \kappa_j^B \qquad [7]$$

where

$$\kappa = \lambda - \eta, \quad \eta = J^{-1}\sum_{j=1}^{J} \log N_j^B \qquad [8]$$

and

$$\kappa_j^B = \lambda_j^B - \log N_j^B + \eta \qquad [9]$$

The additive model is equivalent to the homogeneity model that, for any fixed row i, $p_{i\cdot j}^{A\cdot B}$ is a constant p_i^A for each column j. The parameter λ_i^A is defined from p_i^A as in the multinomial case.

Rasch (1960) considers an additive loglinear model for a contingency table in which each frequency n_{ij} has an independent Poisson distribution with mean m_{ij}. This model is applied to analysis of a reading test in which I subjects read J passages, and n_{ij} is the number of errors made by subject i while reading passage j. The parameter λ_i^A provides a measure of relative subject ability, and λ_i^B measures relative item difficulty. This model is a special case of a general class of models encountered in the study of item responses. For further discussion of the relationship of item-response models and loglinear models, see Kelderman (1984).

The additive loglinear model leads to the quasi-independence model of Goodman (1968) and Bishop and Fienberg (1969) when [1] is only required to hold for pairs (i, j) in a subset K of the pairs (i, j) such that i and j are integers, $1 \le i \le I$, and $1 \le j \le J$.

Example 2. Logit models. As in Bishop (1969) and Haberman (1974: Chapter 8), a logit model may be described by a frequency table with elements n_{ij}, $1 \le i \le 2$, and $1 \le j \le J$. Covariates X_{jk}, $1 \le k \le K$, correspond to each column j. Each column n_{ij}, $1 \le i \le 2$, has an independent multinomial distribution with known sample size N_j^B and unknown positive probabilities $p_{i\cdot j}^{A\cdot B}$, so that $m_{ij} = N_j^B p_{i\cdot j}^{A\cdot B}$ is the expectation of n_{ij}. If $\mu_{ij} = \log m_{ij}$, then the logit $\log(p_{1\cdot j}^{A\cdot B}/p_{2\cdot j}^{A\cdot B})$ corresponding to the conditional probability $p_{1\cdot j}^{A\cdot B}$ is equal to

$$\log\left(m_{1j}/m_{2j}\right) = \log m_{1j} - \log m_{2j} = \mu_{1j} - \mu_{2j}$$

It is assumed that the logits $\log(p_{1\cdot j}^{A\cdot B}/p_{2\cdot j}^{A\cdot B})$ satisfy the linear model

$$\begin{aligned}\mu_{1j} - \mu_{2j} &= \log p_{1\cdot j}^{A\cdot B} - \log p_{2\cdot j}^{A\cdot B} = \log\left(p_{1\cdot j}^{A\cdot B}\Big/p_{2\cdot j}^{A\cdot B}\right)\\ &= \eta + \sum_{k=1}^{K}\beta_k X_{jk}\end{aligned} \qquad [10]$$

for some unknown parameters η and β_k. Thus the logarithms $\log p_{i\cdot j}^{A\cdot B}$ satisfy the same linear model as do the logarithms $\log m_{ij}$ of the expected cell frequencies. Equivalently, if δ is a Kronecker function, then

$$\mu_{ij} = \alpha_j + \eta\delta_{i1} + \sum_{k=1}^{K}\beta_k X_{jk}\delta_{i1} \qquad [11]$$

for unknown parameters α, η, and β_k. Thus,

$$\log p_{i\cdot j}^{A\cdot B} = \gamma_j + \eta\delta_{i1} + \sum_{k=1}^{K}\beta_k X_{jk}\delta_{i1} \qquad [12]$$

where

$$\gamma_j = \alpha_j - \log N_j^B \qquad [13]$$

Example 3. Interactions in a three-way table. As in Birch (1963), consider a three-way table of frequencies n_{ijk}, $1 \le i \le I$, $1 \le j \le J$, $1 \le k \le K$, with respective positive expectations m_{ijk}. Assume that I, J, and K are at least 2. The logarithm $\mu_{ijk} = \log m_{ijk}$ can always be written as

$$\mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC} \qquad [14]$$

where the unknown parameters λ, λ_i^A, λ_j^B, λ_k^C, λ_{ij}^{AB}, λ_{ik}^{AC}, λ_{jk}^{BC}, and λ_{ijk}^{ABC} satisfy the constraints

$$\begin{aligned}\sum_{i=1}^{I}\lambda_i^A &= \sum_{j=1}^{J}\lambda_j^B = \sum_{k=1}^{K}\lambda_k^C = \sum_{i=1}^{I}\lambda_{ij}^{AB} = \sum_{j=1}^{J}\lambda_{ij}^{AB} = \sum_{i=1}^{I}\lambda_{ik}^{AC}\\ &= \sum_{k=1}^{K}\lambda_{ik}^{AC} = \sum_{j=1}^{J}\lambda_{jk}^{BC} = \sum_{k=1}^{K}\lambda_{jk}^{BC} = \sum_{i=1}^{I}\lambda_{ijk}^{ABC}\\ &= \sum_{i=1}^{j}\lambda_{ijk}^{ABC} = \sum_{k=1}^{k}\lambda_{ijk}^{ABC} = 0\end{aligned} \qquad [15]$$

The constraints of [15] are used to identify the model parameters. For example, λ is the average of all the μ_{ijk}. As in the analysis of variance, the parameters λ_i^A, λ_j^B, and λ_k^C may be

termed main effects, the parameters λ_{ij}^{AB}, λ_{ik}^{AC}, and λ_{jk}^{BC} may be termed two-factor interactions, and λ_{ijk}^{ABC} is a three-factor interaction. In the simplest case, the table is a cross-classification obtained from a random sample of size $N \geq 1$ in which the categorical random variables A, B, and C are observed. Here A has values from 1 to I; B has values from 1 to J; and C has values from 1 to K. The count n_{ijk} is the number of observations with $A = i$, $B = j$, and $C = k$. The probability that $A = i$, $B = j$, and $C = k$ is p_{ijk}, so that $m_{ijk} = N\, p_{ijk}$ and

$$\log p_{ijk} = \kappa + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

where $\kappa = \lambda - \log N$. If all two-factor and three-factor interaction parameters are zero, so that the logarithms μ_{ijk} satisfy the additive model

$$\mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C$$

then A, B, and C are mutually independent. Note that the logarithms $\log p_{ijk}$ also satisfy the additive model

$$\log p_{ijk} = \kappa + \lambda_i^A + \lambda_j^B + \lambda_k^C$$

If p_i^A is the marginal probability that $A = i$ and μ^A is the average of the $\log p_i^A$ for $1 \leq i \leq I$, then $\lambda_i^A = \log p_i^A - \mu^A$.

If λ_{ik}^{AC}, λ_{jk}^{BC}, and λ_{ijk}^{ABC} are zero, so that

$$\mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB}$$

then C is independent of the joint variable (A, B). If p_{ij}^{AB} is the joint probability that $A = i$ and $B = j$, then the parameter λ_{ij}^{AB} is a linear function of $\log p_{ij}^{AB}$. If λ_{jk}^{BC} and λ_{ijk}^{ABC} are zero, so that

$$\mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC}$$

then B and C are conditionally independent given A. In a model of no-three-factor interaction, it is assumed that the three-factor interaction terms λ_{ijk}^{ABC} are all 0, so that

$$\mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} \qquad [16]$$

Models of interactions in three-way tables can be applied with other sampling schemes. For example, in the Poisson case, each frequency n_{ijk} has an independent Poisson distribution with mean $m_{ijk} > 0$.

Prior to Birch (1963), the model of no-three-factor interaction can be encountered with somewhat different parameterizations. The models for interactions considered here are examples of hierarchical loglinear models. In these models, the restriction is imposed that higher order interactions are only included in models if corresponding lower order interactions are also included. Thus, if λ_{ijk}^{ABC} is in the model, then all other interactions are included. If λ_{ij}^{AB} is included, then λ, λ_i^A, and λ_j^B must also be included. This family of models is developed in Birch (1963). The models, which are analogous to linear models encountered in the analysis of factorial experiments, have been extensively used in the analysis of multivariate categorical data. Generalizations to contingency tables involving more than three variables are quite common. For example, see Goodman (1970, 1971a, b), Haberman (1974: Chapters 5 and 7, 1978, 1979), and Bishop et al. (1975). Generalizations analogous to quasi-independence models are also readily found. For example, see Haberman (1974: Chapter 7, 1979: Chapter 7).

Maximum Likelihood

The method of maximum likelihood is commonly used for estimation of parameters in a loglinear model. As shown in Haberman (1973b), maximum-likelihood estimates of the expected values of the cell frequencies are uniquely defined if they exist, and they exist if all the observed cell frequencies are positive. The same maximum likelihood estimates apply whether Poisson sampling, multinomial sampling, or product multinomial sampling is used. Maximum-likelihood estimates of expected cell frequencies can also exist if some observed cell frequencies are zero. Haberman (1973b) provides a necessary and sufficient condition for existence of these maximum-likelihood estimates. These conditions do not depend on the particular sampling method employed. In addition, if maximum-likelihood estimates of expected values of cell means do exist, then the estimates do not depend on the sampling method. Thus, the same results are obtained for Poisson sampling as are obtained for multinomial sampling.

Three basic methods are available for computation of maximum-likelihood estimates. The simplest method, if available, involves use of closed-form expressions for the maximum-likelihood estimates. This approach is available in Example 1. The estimates $\widehat{m}_{ij}$ are defined if the row marginal totals $n_i^A = \sum_i n_{ij}$ and the column marginal totals $n_j^B = \sum_j n_{ij}$ are all positive. In this case, the overall total $N = \sum_i \sum_j n_{ij}$ is positive, and $\widehat{m}_{ij} = n_i^A n_j^B / N$. Closed-form maximum likelihood estimates may be obtained for a large class of loglinear models that are commonly termed decomposable models. See Bishop and Fienberg (1969), Goodman (1968, 1970, 1971a), and Haberman (1974: Chapters 5 and 7). These loglinear models normally correspond to combinations of hypotheses concerning equiprobability, conditional independence, and independence. They are closely related to graphical models (Darroch et al., 1980).

A traditional approach to calculation of maximum-likelihood estimates is based on the iterative proportional fitting procedure developed in Deming and Stephan (1940) for use with adjustment of census sample data to conform to census population totals. This approach is readily applied in Example 3. The algorithm generates a series of approximations to the maximum likelihood estimates by means of simple multiplicative adjustments. This approach was used in Goodman (1968, 1970), Bishop (1969), and Bishop and Fienberg (1969) to study hierarchical loglinear models. In typical applications involving decomposable models, convergence is achieved after one cycle of the algorithm. More generally, convergence properties of the algorithm, together with generalizations that apply to all loglinear models, are examined in Darroch and Ratcliff (1972) and Haberman (1974: Chapter 3). A disadvantage of iterative proportional fitting is that it does not help in determination of confidence intervals for parameters.

Several versions of the Newton–Raphson algorithm (Haberman, 1974, 1978, 1979) can be used to obtain maximum-likelihood estimates. The Newton–Raphson algorithm applies to all loglinear models and is currently used in the majority of computer packages that calculate maximum-likelihood estimates for loglinear models. In some instances, the computer packages for loglinear models are contained in packages for generalized linear models. The algorithm normally converges rapidly, and by-products of the calculations may be used to estimate the asymptotic covariance matrix of the model parameters. In some cases of very large contingency tables and models with very large numbers of parameters, the Newton–Raphson algorithm can converge slower than does the iterative proportional fitting algorithm; however, this issue has gradually become less significant as computers become faster and have larger memory capacity.

Large-sample properties of maximum-likelihood estimates for loglinear models have been extensively studied under a variety of conditions. For example, if all expected cell frequencies becomes large and if the model is true (Haberman, 1977), then, for each cell in the table, the ratio between the estimated expected cell frequency and the actual expected cell frequency converges in probability to 1, and a normal approximation is available for the distribution of a linear combination of maximum-likelihood estimates of logarithms of expected values of cell frequencies. This normal approximation is readily employed to construct approximate confidence intervals for linear combinations of logarithms of expected values of cell frequencies. The same normal approximation typically applies for Poisson, multinomial, or product multinomial sampling if, in the case of multinomial or product multinomial sampling, the linear combination of logarithms of expected values of cell frequencies is identical to the corresponding linear combination of logarithms of cell probabilities. Haberman (1977) also examines much more complex cases under which normal approximations can be used.

In cases in which doubts arise concerning the accuracy of the normal approximation used to construct confidence intervals, it is possible, at least in principle, to obtain exact confidence intervals for linear combinations of logarithms of expected values of cell frequencies. The basic approach required is given in Haberman (1974: Chapter 1); however, actual use of exact confidence intervals often involves substantial computational challenges as well as a number of problems concerning the precise way to define an exact confidence interval. Agresti (1992) provides an extensive review of the issues as well as numerous references.

Hypothesis Tests

In the case of loglinear models, the most common tests of goodness of fit are likelihood ratio and Pearson chi-square statistics, although other alternatives exist (Cressie and Read, 1984). For Poisson, multinomial, or product multinomial sampling, these statistics apply to the null hypothesis that the model holds and to the alternative hypothesis that the model is not valid. Central chi-square approximations for the Pearson and likelihood ratio chi-square statistics are available if all expected values of cell frequencies are large and if the model holds. If the model is not valid but all expected cell frequencies are large, then noncentral chi-square approximations may be used.

Cressie and Read (1984) propose a power divergence family of test statistics that includes both the Pearson and likelihood-ratio chi-square and maintains the basic properties of these statistics for the case of large expected cell frequencies.

Two loglinear models may be compared by use of likelihood-ratio chi-squares. Let M and M' represent two nested loglinear models, so that validity of model M implies validity of model M'. Let L^2 be the likelihood-ratio chi-square statistic for goodness of fit of M, and let L'^2 be the likelihood-ratio chi-square statistic for goodness of fit of M'. The difference $L^2 - L'^2$ provides a likelihood-ratio test for the null hypothesis that M holds and the alternative hypothesis that M' holds but M is false. If the expected counts are all large and M holds, then $L^2 - L'^2$ has an approximate central chi-square distribution. If the expected counts are large but M does not hold, then a noncentral chi-square approximation may be considered even in the case in which neither M nor M' holds. Haberman (1977) provides conditions under which these chi-square approximations can also be effective when the sample size is large but individual expected counts are small.

In cases in which the accuracy of large-sample approximations is questionable, exact tests of hypotheses are also available. Such tests are easily implemented in a number of special cases. For example, testing the additive model for a 2 by 2 contingency table or the model of no-three-factor interaction in a 2 by 2 by 2 contingency table is quite straightforward. In other typical cases, the computational labor involved is considerable. Agresti (1992) provides an extensive review of the literature concerning exact tests.

Residual Analysis

Residual analysis provides a basic supplement to formal hypothesis tests. As in regression analysis, residual analysis of loglinear models permits a thorough search for discrepancies between the data and the model. As in regression analysis, residual analysis can involve the use of graphs to consider patterns of residuals and to detect residuals of unusual size. A basic development of residual analysis in loglinear models is found in Haberman (1973a). The raw residual for a cell is the difference between the observed and estimated expected cell frequency. The variability of the raw residual is dependent on the expected cell count, so that it is common to try to rescale the raw residual. For Poisson, multinomial, or product multinomial sampling, a simple rescaling is provided by the standardized residual. The standardized residual for a cell is the ratio of the raw residual and the square root of the estimated expected cell frequency. The sum of the squares of the standardized residuals yields the Pearson chi-square statistic. If the model holds and if the expected cell frequencies are large, then the standardized residuals have approximate normal distributions with means of 0 and variances no greater than 1. The adjusted residual is the ratio of the raw residual and the estimated standard deviation of the raw residual. If

the model holds and if the expected cell frequencies are large, then the adjusted residuals have approximate standard normal distributions.

Alternative definitions of residuals are also available. The Freeman-Tukey residuals (Bishop et al., 1975: pp. 136–141) for a cell with observed count n and estimated expected count $\widehat{m}$ satisfy

$$z = \sqrt{n} + \sqrt{n+1} - \sqrt{4\widehat{m}+1}$$

The Poisson deviance residual D for this cell satisfies the conditions that

$$D^2 = a\left[n\log\left(n/\widehat{m}\right) - n + \widehat{m}\right]$$

and D has the same sign as the raw residual $n - \widehat{m}$. These residuals are similar to the standardized residual. The sum of squares of the Poisson deviance residuals is equal to the likelihood-ratio chi-square statistic. For discussion, see Pierce and Schafer (1986). In the case of generalized residuals, a linear combination of observed cell frequencies is compared to a linear combination of estimated expected cell frequencies. Haberman (1978, 1979) provides a detailed development, references, and examples.

Model Selection

A common issue in the study of loglinear models is the selection of models. This issue is complicated in many common cases due to the very large number of loglinear models that may be considered. As a consequence, problems of multiple tests arise. Consider a case in which $c \geq 2$ loglinear models M_a are under study for $1 \leq a \leq c$, and let L_a^2 be the likelihood-ratio chi-square statistic for goodness of fit of model M_a. Approximate significance levels may be obtained for each L_a^2, and, for cases in which model M_a is nested within M_b, so that validity of model M_a implies validity of model M_b, an approximate significance level may be obtained for the log likelihood ratio statistic $D_{ab} = L_a^2 - L_b^2$ for comparison of models M_a and M_b. Based on these results, various strategies can be devised to select models (Bishop et al., 1975: Chapter 4, Goodman, 1971b, 1973; Haberman, 1978: Chapter 3). Measures of quality of fit have also been proposed by Goodman (1971b), Akaike (1974), Schwarz (1978), and Gilula and Haberman (1994). Measures of quality of fit may be employed to compare models that are not nested.

A significant issue in model selection commonly arises in large samples. The statistic D_{ab} may be very large, but the practical difference between models M_a and M_b may be quite small. To address this issue, Gilula and Haberman (1994) suggest use of information criteria to measure the practical difference between models. Their analysis depends on the information measure H_a, which is minus the ratio of the log likelihood for model M_a and the sample size N. They consider statistics such as $H_a - H_b = D_{ab}/(2N)$ and $1 - H_b/H_a$ to measure model improvement. As sample sizes increase, these measures converge to finite limits, whether or not either of the models is valid, and confidence intervals for measures are readily constructed. Information measures can be employed to assess the relative ability of the models to predict probabilities associated with observations. These measures may be used for comparisons of models that are not nested.

Conclusion

Loglinear models provide a very general approach to the analysis of discrete data. They have the attraction that numerous models can be constructed for complex relationships, maximum likelihood is readily applied, and tests of fit are easily applied. As a consequence, an extensive literature exists concerning loglinear models. Bishop et al. (1975) provide a basic overview of the subject with an emphasis on hierarchical loglinear models. Haberman (1974) provides a rigorous mathematical treatment of loglinear models, while Haberman (1978, 1979) examines such models in the context of research in the social sciences. Goodman (1968, 1970, 1971a, b, 1973) provides a fundamental series of papers on the use of hierarchical loglinear models. For a review of Bayesian approaches to loglinear models, see Agresti and Hitchcock (2005). For further consideration of logit models.

See also: Binary Response Models and Logistic Regression; Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis; Factor Analysis and Latent Structure Analysis: Overview; Factor Analysis and Latent Structure: IRT and Rasch Models; Hypothesis Testing in Statistics; Hypothesis Testing: Methodology and Limitations; Hypothesis Tests, Multiplicity of; Linear Hypothesis: Regression (Basics); Linear Hypothesis.

Bibliography

Agresti, A., 1992. A survey of exact inference for contingency tables. Statistical Science 7, 131–177.
Agresti, A., Hitchcock, D.B., 2005. Bayesian inference for categorical data analysis. Statistical Methods and Applications 14, 297–330.
Akaike, H., 1974. A new look at the statistical identification. IEEE Transactions on Automatic Control 19, 716–723.
Bartlett, M.S., 1935. Contingency table interactions. Supplement to the Journal of the Royal Statistical Society 2, 248–252.
Birch, M.W., 1963. Maximum likelihood in three-way contingency tables. Journal of the Royal Statistical Society, Series B 25, 220–233.
Bishop, Y.M.M., 1969. Full contingency tables, logits, and split contingency tables. Biometrics 25, 383–399.
Bishop, Y.M.M., Fienberg, S.E., 1969. Incomplete two-dimensional contingency tables. Biometrics 25, 119–128.
Bishop, Y.M.M., Fienberg, S.E., Holland, P.W., 1975. Discrete Multivariate Analysis: Theory and Practice. MIT, Cambridge, MA.
Cressie, N., Read, T.R.C., 1984. Multinomial goodness-of-fit tests. Journal of the Royal Statistical Society, Series B 46, 440–464.
Darroch, J.N., Lauritzen, S.L., Speed, T.P., 1980. Markov fields and loglinear interaction models for contingency table. Annals of Statistics 8, 522–539.
Darroch, J.N., Ratcliff, D., 1972. Generalized iterative scaling for loglinear models. Annals of Mathematical Statistics 43, 1470–1480.
Deming, W.E., Stephan, F.F., 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Annals of Mathematical Statistics 11, 427–444.
Gilula, Z., Haberman, S.J., 1994. Models for analyzing categorical panel data. Journal of the American Statistical Association 89, 645–656.

Goodman, L.A., 1968. The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing entries. Journal of the American Statistical Association 63, 1091–1131.

Goodman, L.A., 1970. Multivariate analysis of qualitative data: interactions among multiple classifications. Journal of the American Statistical Association 65, 226–256.

Goodman, L.A., 1971a. Partitioning of chi-square, analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional contingency tables. Journal of the American Statistical Association 66, 339–344.

Goodman, L.A., 1971b. Analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics 13, 33–61.

Goodman, L.A., 1973. Guided and unguided methods for the selection of models for a set of T multidimensional contingency tables. Journal of the American Statistical Association 68, 165–175.

Haberman, S.J., 1973a. Analysis of residuals in cross-classified tables. Biometrics 29, 205–220.

Haberman, S.J., 1973b. Loglinear models for frequency data: sufficient statistics and likelihood equations. Annals of Statistics 1, 617–632.

Haberman, S.J., 1974. The Analysis of Frequency Data. University of Chicago Press, Chicago.

Haberman, S.J., 1977. Loglinear models and frequency tables with small expected cell counts. Annals of Statistics 5, 1148–1169.

Haberman, S.J., 1978. Analysis of Qualitative Data. In: Introductory Topics, vol. 1. Academic, New York.

Haberman, S.J., 1979. Analysis of Qualitative Data. In: New Developments, vol. 2. Academic, New York.

Haberman, S.J., 1981. Tests for independence in two-way contingency tables based on canonical correlation and on linear-by-linear interaction. Annals of Statistics 9, 1178–1186.

Kelderman, H., 1984. Loglinear Rasch model tests. Psychometrika 49, 223–245.

Norton, H.W., 1945. Calculation of chi-square for complex contingency tables. Journal of the American Statistical Association 40, 251–258.

Pierce, D.A., Schafer, D.W., 1986. Residuals in generalized linear models. Journal of the American Statistical Association 81, 977–986.

Rasch, G., 1960. Probabilistic Models for Some Intelligence and Attainment Tests. Nielsen and Lydiche, Copenhagen.

Roy, S.N., Kastenbaum, M.A., 1956. On the hypothesis of no interaction in a multi-way contingency table. Annals of Mathematical Statistics 27, 749–757.

Schwarz, G., 1978. Estimating dimension of a model. Annals of Statistics 6, 461–464.

Multivariate Analysis: Discrete Variables (Overview)

Alan Agresti, University of Florida, Gainesville, FL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 10233–10240, © 2001, Elsevier Ltd.

Abstract

This article deals with discrete multivariate analysis of categorical response variables. A categorical variable is one for which the measurement scale is a set of categories. For studies having two or more categorical variables, a contingency table displays the counts for their cross-classification of categories. Modeling approaches are vital for investigating association and interaction structure. Logistic regression is an analog of ordinary regression for binary response variables. Loglinear models, by contrast, are relevant for analyses analogous to correlation analyses, studying the association structure among a set of categorical response variables. Correspondence analysis is a related descriptive and graphical method. Multivariate analyses are relevant when there are several categorical response variables, such as in repeated measurement (e.g., longitudinal) studies. Recent advances here include models that address marginal components of a multivariate response and models that use cluster-specific random effects to describe the joint distribution. Other recent advances in discrete multivariate analysis include specialized models for ordinal responses and exact small-sample methods. Finally, the unifying concept of generalized linear models connects the primary categorical modeling procedures – logistic regression models and loglinear models – with long-established regression and analysis of variance methods for continuous response variables.

This article describes statistical methods for multivariate data sets having discrete response variables. Section Traditional Analyses of Discrete Multivariate Data summarizes some traditional ways of analyzing discrete data, such as analyses of two-way contingency tables. As in the multivariate analysis of continuous variables, modeling approaches are vital for studying association and interaction structure. Models for a categorical response variable, discussed in Section Regression Modeling of a Categorical Response, describe how the probability of a particular response outcome depends on values of predictor variables. These models, of which logistic regression is most important, are analogs for categorical responses of ordinary regression. The log-linear model, discussed in Section Loglinear Models, provides an analog of correlation analysis; it analyzes association patterns and strength of association. Section Sampling Assumptions discusses sampling assumptions for analyses for discrete data, and Section Alternative Model Types mentions other types of models for discrete data. Section New Methods for Repeated Measurement Data describes recent advances for analyzing data resulting from repeated or clustered measurement of discrete variables. Section Historical Notes provides some historical perspective.

Traditional Analyses of Discrete Multivariate Data

Most commonly, discrete variables are categorical in nature. That is, the measurement scale is a set of categories, such as (liberal, moderate, conservative) for political philosophy. In multivariate analyses with two or more categorical variables, a *contingency table* displays the counts for the cross-classification of categories. For instance, a contingency table might display counts indicating the number of subjects classified in each of the $9 = 3 \times 3$ combinations of responses on political philosophy and political party orientation (Democrat, Republican, Independent). Those combinations form the cells of the table.

Two categorical variables are *independent* if the probability of response in any particular category of one variable is identical for each category of the other variable. This is an idealized situation that is rarely true in practice, although like any statistical model it is sometimes adequate for describing reality. For instance, some surveys of American college students have shown that opinion about whether abortion should be legal is essentially independent of gender; that is, the percentage favoring legalization is about the same for male and female students.

Treating independence as a simple model for a contingency table, we can test its fit using statistics that compare the observed cell counts in the table to those expected (predicted) by that model. The most commonly used test statistics are

$$\text{Pearson statistic} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$\text{Likelihood} - \text{ratio statistic} = 2\sum \text{observed} \log\left(\frac{\text{observed}}{\text{expected}}\right) \qquad [1]$$

These both have large-sample chi-squared distributions. For r rows and c columns, the degrees of freedom (df) for the test equal $df = (r-1)(c-1)$, and the P-value is the right-tail probability above the observed test statistic value. These chi-squared statistics also can compare counts to expected values for more complex models than independence.

In conducting goodness-of-fit tests, one does not seriously entertain the possibility that the model truly holds; any model is false and provides a simple representation that only approximates reality. However, when the chi-squared statistics do not show significant lack of fit, benefits of model parsimony accrue from using the model fit, which smooths the sample cell proportions somewhat. Interpretations are simpler, and estimates of population characteristics such as the true cell proportions are likely to be better using the model-based estimates.

Multidimensional contingency tables result from cross-classifying more than two categorical variables, such as for studying the association separately at levels of control variables race, gender, or educational level. The model of *conditional independence* states that two categorical variables are independent at each level of a third; that is, ordinary independence applies in each *partial table*. Large-sample chi-squared tests of fit of this model, such as the Cochran–Mantel–Haenszel test, combine the information from the partial tables. Most are designed to be relatively powerful when the strength of the association is similar in each partial table. The model of *homogeneous association* states that the strength of association between two variables is identical at each level of a third variable. Goodness-of-fit tests of these models were originally formulated for stratified two-by-two tables, but generalizations exist for stratified *r*-by-*c* tables with ordered or unordered categories.

Goodness-of-fit tests have limited use, and in many circumstances it is more informative to estimate the strength of the associations. Goodman and Kruskal (1979) discussed a variety of summary measures. Those for ordinal variables, such as their *gamma* measure, are correlation-like indices that fall between -1 and $+1$. More attention has been focused on measures that arise as parameters in models. The odds ratio, discussed in the next section, is a parameter in logistic regression and loglinear models and is applicable for two-way or multi-way tables.

For estimating parameters in models for discrete data, the method of *maximum likelihood* estimation is the standard. The probability function for the data (e.g., Poisson, binomial), expressed as a function of the parameters after observing the data, is called the *likelihood function*. The maximum likelihood estimates are values for the parameters for which this achieves its maximum; that is, these estimates are the parameter values for which the probability of the observed data takes its maximum value. Except in a few simple cases, numerical methods are needed to find the maximum likelihood estimates and their standard errors. Software is reasonably well developed in the major statistical packages. It is often most natural to analyze discrete data using procedures for generalized linear models that permit non-normal responses, such as the Poisson and binomial. Examples include PROC GENMOD in SAS and the 'glm' function in S-Plus. For instance, for SAS see Allison (1999) and Stokes et al. (1995), for S-Plus see Venables and Ripley (1999), and for Stata see Rabe-Hesketh and Everitt (2000).

Regression Modeling of a Categorical Response

For multidimensional contingency tables and other forms of multivariate data, modeling approaches are vital for investigating association and interaction structure. *Logistic regression* is an analog of ordinary regression for binary response variables, and it has extensions for multicategory response variables. Like ordinary regression, logistic regression distinguishes between response and predictor variables. The predictor variables can be qualitative or quantitative, with qualitative ones handled by dummy variables.

Let p denote the probability of outcome in a particular response category (e.g., favoring the Democratic rather than Republican candidate for President). For the observations at particular fixed values of predictor variables, the number of responses in the two outcome categories is assumed to have the binomial distribution. With a set of predictors $x_1, x_2, \ldots, x_p$, the logistic regression model is

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad [2]$$

The ratio $p/(1-p)$ is called the *odds*, and the natural logarithm of the odds is called the *logit*.

One reason for using the logit instead of p itself on the left-hand side of this equation is to constrain predicted probabilities to fall between 0 and 1. With a single predictor, for instance, the *linear probability model* $p = \alpha + \beta x$ permits p to fall below 0 or above 1 for many values of x; by contrast, the formula $\log[p/(1-p)] = \alpha + \beta x$ maps out an S-shaped function for which p increases or decreases according to whether $\beta > 0$ or $\beta < 0$ but always stays bounded between 0 and 1. For instance, for the decision to buy/rent a home, suppose the relationship between $p =$ the probability of buying and $x =$ annual family income (in ten thousands of dollars) satisfies the logistic regression model $\log[p/(1-p)] = 4.0 - 0.8x$. Figure 1 shows the plot of p for x between 0 and 10 (i.e., below \$100 000). The plot is roughly linear for p between 0.2 and 0.8.

The *probit model* is another model that has this S-shape. The function probit $(p) = \alpha + \beta x$ has a graph for p (or for $1 - p$ when $\beta < 0$) that has the shape of a normal cumulative distribution function with mean $-\alpha/\beta$ and standard deviation $1/|\beta|$. The probit model can be derived by assuming an underlying latent normal variable (e.g., Long, 1997). In practice, the probit model gives a very similar fit to the logistic regression model. To illustrate, Figure 1 also shows the plot of the equation probit $(p) = 2.35 - 0.47x$, which has the shape of a normal cumulative distribution function with mean 5.0 and standard deviation 2.1; for instance, a decrease in p of 0.68, from

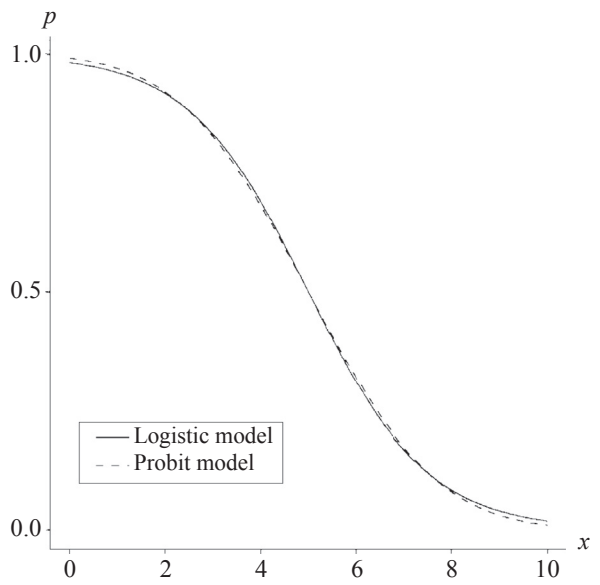

Figure 1 Plot of logistic regressional model $log[p/(1-p)] = 4.0 - 0.8x$ and the probit regression model probit $(p) = 2.35 - 0.47x$.

$p = 0.84$ to $p = 0.16$, occurs between $x = 5.0 - 2.1 = 2.9$ and $x = 5.0 + 2.1 = 7.1$. Although the fits are very similar, estimates of α and β are on a different scale for probit and logistic models, and usually logistic parameter estimates are about 1.6–1.8 times the probit parameter estimates.

With multiple predictors, the coefficient β_k of x_k in the logistic regression model describes the effect of x_k, controlling for the other predictors in the model. The antilog of β_k is an odds ratio representing the multiplicative effect of a one-unit change in x_k on the odds. For instance, in a model for the probability in favoring the Democratic candidate for President, suppose $\beta_k = 0.40$ for the coefficient of $x_k =$ gender, with $x_k = 1$ for females and 0 for males. Then, controlling for the other variables in the model, for females the odds of favoring the Democratic candidate equal $e^{0.40} = 1.49$ times the odds for males; that is, the odds of favoring the Democrat are 49% higher for females than for males.

For quantitative predictors, simple interpretations that apply directly to the probability p rather than to the odds $p/(1-p)$ result from straight-line approximations to the logistic S-shaped curve for p. When p is near 0.5, $\beta_k/4$ is the approximate change in p for a one-unit change in x_k, controlling for the other predictors. To illustrate, suppose $\beta_k = -0.12$ for the coefficient of $x_k =$ annual income (in ten thousands of dollars) in the model for the probability of favoring the Democratic candidate. Then, for an increase of \$10 000 in income near where $p = 0.5$, the probability of support for the Democratic candidate decreases by about $0.12/4 = 0.03$, other variables being kept constant. To test the effect of a term in the model, the ratio of the estimate of β_k to its standard error has an approximate standard normal distribution, under the hypothesis of no effect (H_0:$\beta_k = 0$).

Using the estimated parameters from the logistic regression model, one can also estimate the probability p for given settings for the predictors, as

$$\widehat{p} = \frac{\exp\left(\widehat{\alpha} + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_p x_p\right)}{1 + \exp\left(\widehat{\alpha} + \widehat{\beta}_1 x_1 + \widehat{\beta}_p x_p + \cdots + \widehat{\beta}_p x_p\right)} \qquad [3]$$

When all predictors are categorical, chi-squared tests analyze the adequacy of the model fit by comparing the observed cell counts to values expected under the model.

Generalizations of logistic regression apply to categorical response variables having more than two categories. A multicategory extension of the binomial distribution, called the *multinomial*, applies to the response counts at each setting of the predictors.

For nominal responses, the generalized model forms logits by pairing each response category with a baseline category. For an r-category response, the model has form

$$\log\left[\frac{p(Y = j)}{p(Y = r)}\right] = \alpha_j + \beta_{1j}x_1 + \beta_{2j}x_2 + \cdots + \beta_{pj}x_p \qquad j = 1, 2, \ldots, r-1$$

A separate set of parameters applies for each response category contrasted with the final one. For instance, if Y refers to preferred candidate and has categories (Democrat, Republican, Independent), the effect of gender on the choice between Democrat and Independent may differ from its effect on the choice between Republican and Independent. The choice of category for the baseline (the denominator of the logits) is arbitrary. Regardless of this choice the same inference occurs about comparisons of models with and without certain predictors and about estimated probabilities in the r response categories. For ordinal responses, the *cumulative logit model* describes the effects of predictors on the odds of response below any given level instead of above it (i.e., the model refers to an odds for the cumulative probability; see McCullagh, 1980). It has form

$$\log\left[\frac{p(Y \le j)}{p(Y > j)}\right] = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad j = 1, \ldots, r-1 \qquad [4]$$

For further details about logistic regression models, *see* Multivariate Analysis: Discrete Variables (Overview); also Agresti (1996: Chapter 5), Collett (1991), Hosmer and Lemeshow (2000), and Long (1997).

Loglinear Models

Logistic regression resembles ordinary regression in distinguishing between a response variable Y and a set of predictors $\{x_k\}$. *Loglinear models*, by contrast, treat all variables symmetrically and are relevant for analyses analogous to correlation analyses, studying the association structure among a set of categorical response variables.

For multidimensional contingency tables, a variety of models are available, varying in terms of the complexity of the association structure. For three variables, models include ones for which (1) the variables are mutually independent, (2) two of the variables are associated but are jointly independent of the third, (3) two of the variables are conditionally independent, given the third variable, but may both be associated with the third, (4) each pair of variables is associated, but the association between each pair is homogeneous at each level of the third variable, and (5) each pair of variables is associated and the strength of association between each pair may vary according to the level of the third variable (Bishop et al., 1975).

Relationships describing probabilities in contingency tables are naturally multiplicative, so ordinary regression-type models occur after taking the logarithm, which is the reason for the term *loglinear*. To illustrate, in two-way contingency tables independence between row variable X and column variable Y is equivalent to the condition whereby the probability of classification $P(X = i, Y = j)$ in the cell in row i and in column j depends only on the marginal probabilities $P(X = i)$ and $P(Y = j)$,

$$P(X = i, Y = j) = P(X = i)P(Y = j) \qquad [5]$$

for all i and j. For a sample of size n, the expected count $\mu_{ij} = nP(X = i, Y = j)$ in that cell therefore satisfies

$$\begin{aligned}\log \mu_{ij} &= \log\left[nP(X = i)P(Y = j)\right] \\ &= \log(n) + \log\left[P(X = i)\right] + \log\left[P(Y = j)\right]\end{aligned} \qquad [6]$$

This has the loglinear form

$$\log \mu_{ij} = \alpha + \beta_i + \gamma_j \qquad [7]$$

for which the right-hand side resembles a simple linear model – two-way analysis of variance (ANOVA) without interaction. Alternatively this has regression form by replacing $\{\beta_i\}$ by dummy variables for the rows times parameters representing effects of classification in those rows and replacing $\{\gamma_j\}$by dummy variables for the columns times parameters representing effects of classification in those columns.

Similarly, loglinear model formulas for more complex models such as those allowing associations resemble ANOVA models except for predicting the logarithm of each cell expected frequency rather than the expected frequency itself. Dummy variables represent levels of the qualitative responses, and their interaction terms represent associations. The associations are described by odds ratios. Logistic regression models with qualitative predictors are equivalent to certain loglinear models, having identical estimates of odds ratios and identical goodness-of-fit statistics. For ordinal variables, specialized loglinear models assign ordered scores to the categories and have parameters describing trends in associations.

For further details about loglinear models, see Agresti (1996: Chapter 6), Bishop et al. (1975), and Fienberg (1980).

Sampling Assumptions

Inferential analysis of multivariate discrete data, like that of any type of data, requires assumptions about the sampling mechanism that generated the data. Often, such as in most surveys, the overall sample size is fixed and the multinomial distribution applies as the probability distribution for the counts in a contingency table. In many cases, certain marginal totals are also fixed, such as in experimental designs in which the number of subjects at each experimental condition is fixed. In addition, in regression-type models such as logistic regression, it is common to treat the counts at the combinations of levels of the predictor variables as fixed even if they are not fixed in the sampling design. In these cases, at each such combination the response is usually assumed to have a binomial distribution when the response is binary and a multinomial distribution when the response is multicategory; this replaces the usual normal response assumption for regression models for continuous responses. Other sampling possibilities, less common in practice, are that the sample size is itself random (i.e., no counts are fixed, such as in observing categorical outcome measures for all subjects who visit some clinic over a future time period) or, at the other extreme, all marginal totals are fixed; these lead to Poisson and hyper-geometric distributions.

Although some methods depend on the sampling assumption, estimates and test statistics for many analyses are identical for the common sampling designs. For instance, for two-way contingency tables, the chi-squared statistic, its degrees of freedom, and its large-sample P-value are the same when only the total sample size is fixed, when only the row totals are fixed, when only the column totals are fixed, or when both row and column totals are fixed.

Under standard sampling assumptions, traditional methods of inference for discrete data use statistics having approximate chi-squared or normal distributions for large samples. Alternative methods are available for small samples. Fisher's exact test, due to the famous statistician R. A. Fisher, is a test of independence in 2-by-2 contingency tables. Analogs of this test are available for larger tables and for testing conditional independence and homogeneous association in three-way tables (characterized in terms of homogeneity of odds ratios) as well as testing fit of more complex logit and loglinear models. small-sample confidence intervals for parameters such as proportions, difference of proportions, and odds ratios, result from inverting exact tests for various null values of parameters. Much of the current research in small-sample inference uses simulation methods to closely approximate the true sampling distributions of relevant statistics (e.g., Forster et al., 1996). StatXact (Cytel Software, 1999) is statistical software dedicated to small-sample discrete-data methods. Some of these analyses are also available in modules of software such as SAS and SPSS.

Alternative Model Types

For contingency tables relating several response variables, alternative statistical approaches to loglinear models can describe association among the variables. Foremost among these is *correspondence analysis*, a graphical way of representing associations in two-way contingency tables that has been particularly popular in France (e.g., Benzécri, 1973; see Multivariate Analysis: Discrete Variables (Correspondence Models)). The rows and columns are represented by points on a graph, the positions of which indicate associations between the two. Goodman (1986) developed a model-based version of this approach. He showed it has equivalences with *correlation models*, which determine scores for categories that maximize correlations between categorical variables. He also showed approximate connections with a set of *association models* that he helped to develop, some of which have loglinear model form but with scores or score-like parameters assigned to levels of ordinal variables. Clogg and Shihadeh (1994) surveyed such models.

Another graphical way of representing a multivariate relationship uses *graphical models* that study conditional independence structure among the variables. Roughly speaking, the graph has a set of vertices, each vertex representing a variable. An edge connecting two vertices represents a conditional association between the corresponding two variables. For instance, for four variables, the graph in **Figure 2** portrays

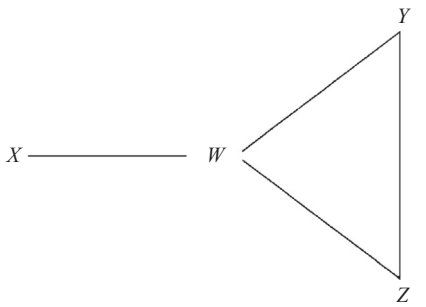

Figure 2 A graphical representation of four variables, with conditional independence between *X* and *Y* and between *X* and *Z*.

a model that assumes that X and Y are independent and that X and Z are independent, conditional on the remaining two variables, but permits association between W and X and between each pair of variables in the set $\{W, Y, Z\}$. The four edges in the graph, connecting W and X, W and Y, W and Z, and Y and Z, represent pairwise conditional associations. Edges do not connect X and Y or X and Z, since those pairs are conditionally independent. Many loglinear models are graphical models. For details, see Lauritzen (1996).

Logistic/probit regression models and loglinear models, together with long-established regression and ANOVA methods for normal response variables, are special cases of a broad family of *generalized linear models* (Gill, 2000). These models are characterized by a choice of distribution for the response variable (e.g., binomial, Poisson, normal), a function of the mean to be modeled (e.g., logit of the mean, log of the mean, the mean itself), and the variables that enter the linear model formula as predictors.

For binary data, the model using the logit transform of the probability p is the most popular generalized linear model. For count data, generalized linear models usually assume a Poisson response, sometimes called *Poisson regression models*. However, more variability often occurs in the counts than the Poisson distribution allows (i.e., there is said to be *over-dispersion*). The Poisson distribution forces the variance to equal the mean; according to it, for instance, at a given mean for Y the variance of the conditional distribution of Y cannot change as predictors are added to or deleted from the model. Methods of dealing with overdispersion include using discrete distributions for which the variance can exceed the mean, such as the *negative binomial*, and adding an extra error term to the model (a *random effect*) to account for unexplained variability due to factors such as unmeasured variables and measurement error.

Extensions of generalized linear models deal with more complex situations. For instance, sometimes the response is a mixture of discrete and continuous, such as when a certain proportion of the responses take value 0 but the rest can be any positive real number (e.g., responses to 'How much time do you spend on exercise each week?'). Sometimes observations are censored – we know only that the response falls above some point or below some point, such as when we count events of a certain type over time but cannot observe beyond the end of the experimental period (Andersen et al., 1993). Models can then focus on the continuous time to response of an event or the discrete counting of events. Depending on the application and the sort of censoring, the models have different names, including survival models, event-history models, and tobit models (e.g., Long, 1997).

New Methods for Repeated Measurement Data

For standard inference with the usual sampling assumptions, the methods discussed above treat separate observations in a sample as independent. In the social and behavioral sciences, multivariate data often results from repeated measurement (e.g., longitudinal) studies. For such studies, this independence assumption is violated. More generally, data sets that consist of clusters of observations such that observations within clusters tend to be more alike than observations between clusters require specialized methods. This was an active research area in the last 15 years of the twentieth century. New methods make it possible to extend traditional repeated measures ANOVA to discrete data and to permit various correlation structures among the responses.

For clustered data, recently developed methods enable modeling binary or categorical responses with explanatory variables, for instance using the logit transform. Two approaches have received considerable attention. In one, models address marginal (separate) components of the clustered response, allowing for dependence among the observations within clusters but treating the form of that dependence as a nuisance that is not a primary object of the modeling. The other approach uses cluster-specific random effects as a mechanism for inducing the dependence.

The marginal modeling approach has the advantage of relative computational simplicity. Estimates are obtained using the methodology of *generalized estimating equations* (see Diggle et al., 1994). With this quasi-likelihood method, rather than fully specify joint distributions for the responses, one need only model how the variance depends on the mean and predict a within-cluster correlation structure. A simple version of the method exploits the fact that good estimates of regression parameters usually result even if one naively treats the component responses in a cluster as independent. Although those parameter estimates are normally adequate under the naïve independence assumption, standard errors are not. More appropriate standard errors result from an adjustment using the empirical dependence in the sample data. Although one makes a 'working guess' about the likely correlation structure, one adjusts the standard errors to reflect what actually occurs for the sample data. An exchangeable working correlation structure under which correlations for all pairs of responses in a cluster are identical is more flexible and usually more realistic than the naïve independence assumption.

With models that apply directly to the marginal distributions, the regression effects are population-averaged. For instance, a fitted probability estimate refers to *all* subjects at the given level of predictors, and the model does not attempt to study potential heterogeneity among the subjects in this probability. Alternatively, one can model subject-specific probabilities and allow such heterogeneity by incorporating subject or cluster random effects in the model. For instance, a model might contain a cluster-specific intercept term that varies among clusters according to some normal distribution. The estimated variance of that distribution describes the extent of heterogeneity. This type of model is called a *generalized linear mixed model* since it contains a mixture of fixed and random effects. Specifying this form of model also has implications about the effects in a marginal model, but the form of model may differ. For instance, with random effects in a logistic regression model, the corresponding model for marginal distributions is only approximated by a logit model and the effects have different size (Diggle et al., 1994).

An alternative approach of dealing with the cluster terms is the fixed effects one of item response models such as the *Rasch model* (see Factor Analysis and Latent Structure: IRT and Rasch Models). In some applications with a multivariate response, it is natural to assume the existence of an unmeasured *latent variable* that explains the association structure. Conditional on the value

of that latent variable, the observed variables are independent. The observed association structure then reflects association between the observed variables and that latent variable. When the latent variable consists of a few discrete classes, such as when the population is a mixture of a few types of subjects, then the resulting model is called a *latent class model*. This is an analog of factor analysis in which the observed and latent variables are categorical rather than continuous. See Heinen (1993) and Factor Analysis and Latent Variable Models in Personality Psychology.

Historical Notes

The early development of methods for discrete multivariate data took place in England. In 1900, Karl Pearson introduced his chi-squared statistic (although in 1922 R. A. Fisher showed that Pearson's *df* formula was incorrect for testing independence) and G. Udny Yule presented the odds ratio and related measures of association. Much of the early literature consisted of debates about appropriate measures of association for contingency tables. Goodman and Kruskal (1979) surveyed this literature and made significant contributions of their own.

In the 1950s and early 1960s, methods for multidimensional contingency tables received considerable attention. These articles were the genesis of substantial research on logistic regression and loglinear models between about 1965 and 1975. Much of the leading work in that decade took place at the universities of Chicago, Harvard, and North Carolina. At Chicago, Leo Goodman wrote a series of groundbreaking articles for statistics and sociology journals that popularized the methods for social science applications. Simultaneously, related research at Harvard by students of Frederick Mosteller (such as Stephen Fienberg) and William Cochran and at North Carolina by Gary Koch and several students and co-workers was highly influential in the biomedical sciences.

Perhaps the most far-reaching contribution was the introduction by British statisticians John Nelder and R.W.M. Wedderburn in 1972 of the concept of generalized linear models. This unifies the primary categorical modeling procedures – logistic regression models and loglinear models – with long-established regression and ANOVA methods for normal-response data.

Texts covering logistic regression are, from least to most advanced, Agresti and Finlay (1997: Chapter 15), Agresti (1996), Hosmer and Lemeshow (2000), Collett (1991), Long (1997), Agresti (1990). Texts covering loglinear models, from least to most advanced, are Agresti and Finlay (1997: Chapter 15), Agresti (1996), Fienberg (1980), Clogg and Shihadeh (1994), Agresti (1990), Bishop et al. (1975). Andersen (1980) and Long (1997) discuss some alternative methods for discrete data.

See also: Experimental Design: Overview; Multivariate Analysis: Classification and Discrimination; Multivariate Analysis: Discrete Variables (Correspondence Models); Multivariate Analysis: Discrete Variables (Loglinear Models).

Bibliography

Agresti, A., 1990. Categorical Data Analysis. Wiley, New York.
Agresti, A., 1996. An Introduction to Categorical Data Analysis. Wiley, New York.
Agresti, A., Finlay, B., 1997. Statistical Methods for the Social Sciences, third ed. Prentice-Hall, Upper Saddle River, NJ.
Allison, P.D., 1999. Logistic Regression Using the SAS System. SAS Institute, Cary, NC.
Andersen, E.B., 1980. Discrete Statistical Models with Social Science Applications. North Holland, Amsterdam.
Andersen, P.K., Borgan Ø, Gill R.D., Keiding, N., 1993. Statistical Models Based on Counting Processes. Springer, New York.
Benzécri, J.-P., 1973. L'analyse des Données. Dunod, Paris.
Bishop, Y.M.M., Fienberg, S.E., Holland, P.W., 1975. Discrete Multivariate Analysis. MIT Press, Cambridge, MA.
Clogg, C.C., Shihadeh, E.S., 1994. Statistical Models for Ordinal Variables. Sage, Thousand Oaks, CA.
Collett, D., 1991. Modelling Binary Data. Chapman and Hall, London.
Cytel Software, 1999. StatXact, Version 4. Cambridge, MA.
Diggle, P.J., Liang, K.-Y., Zeger, S.L., 1994. Analysis of Longitudinal Data. Oxford University Press, Oxford.
Fienberg, S.E., 1980. The Analysis of Cross-Classified Categorical Data, second ed. MIT Press, Cambridge, MA.
Forster, J.J., McDonald, J.W., Smith, P.W.F., 1996. Monte Carlo exact conditional tests for log-linear and logistic models. Journal of the Royal Statistical Society, Series B 58, 445–453.
Gill, J., 2000. Generalized Linear Models: A Unified Approach. Sage, Thousand Oaks, CA.
Goodman, L.A., 1986. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. International Statistical Review 54, 243–270.
Goodman, L.A., Kruskal, W.H., 1979. Measures of Association for Cross Classifications. Springer, New York.
Heinen, T., 1993. Discrete Latent Variable Models. Tilburg University Press, The Netherlands.
Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression, second ed. Wiley, New York.
Lauritzen, S.L., 1996. Graphical Models. Oxford University Press, Oxford.
Long, J.S., 1997. Regression Model for Categorical and Limited Dependent Variables. Sage, Thousand Oaks, CA.
McCullagh, P., 1980. Regression models for ordinal data. Journal of the Royal Statistical Society, Series B 42, 109–142.
Rabe-Hesketh, S., Everitt, B., 2000. A Handbook of Statistical Analyses Using Stata, second ed. CRC Press, London.
Stokes, M.E., Davis, C.E., Koch, G.G., 1995. Categorical Data Analysis Using the SAS System. SAS, Cary, NC.
Venables, W.N., Ripley, B.D., 1999. Modern Applied Statistics with S-Plus, third ed. Springer, New York.

Mumford, Lewis (1895–1990)

Robert Wojtowicz, Old Dominion University, Norfolk, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the life and work of Lewis Mumford (1895–1990), placing special emphasis on his wide-ranging publications. He was the author of more than twenty books and a thousand articles and reviews on subjects, ranging from art and literature to the histories of technology and urbanism. One of America's leading humanists, Mumford defined himself as a 'generalist' who sought an integrated approach to the study and renewal of contemporary society. He is best known as a critic of architecture and urbanism, but he believed his most lasting contributions lay in the field of cultural studies. By choice, Mumford remained largely outside of the Academy, with the result that his ideas reached a broad swath of the educated public.

Introduction

Lewis Mumford was one of the seminal cultural critics to emerge from the United States during the twentieth century. Although known chiefly for his work in architecture and urbanism, he made significant contributions to art and literary criticism, biography and autobiography, the history of technology, sociology, and political science. Largely self-taught, Mumford defined himself as a 'generalist,' one who crossed disciplines and epochs in pursuit of an 'organic' equilibrium in human affairs. This search for 'eutopia' or the good place – to use the Greek spelling and meaning that he preferred – preoccupied him over the course of a writing career that spanned an astonishing six decades (Figure 1).

Life and Work

Background and Intellectual Development

Born on 19 October 1895 in Flushing, Long Island (now part of New York City), Mumford was reared in a middle class, albeit fatherless, household on the Upper West Side of Manhattan. His youth was unremarkable, save for a series of extended walks around New York City that he undertook in the company of his stepgrandfather. He was graduated from Stuyvesant High School in 1912, but ill-health forced him to withdraw from the City College of New York before completing his baccalaureate degree. Rather than languish, Mumford undertook a regimen of self-study and self-improvement under the direct influence of Sir Patrick Geddes, the Scottish biologist, sociologist, and town planner whose ideas Mumford first encountered while enrolled in a college biology course. Mumford began corresponding with Geddes in 1918, and the latter acted as an intellectual mentor and surrogate father to the young man.

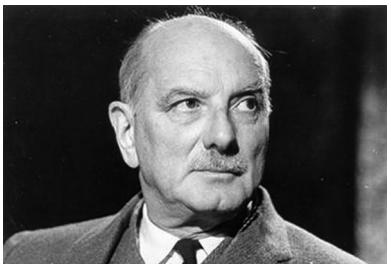

Figure 1 Lewis Mumford. Courtesy of the Lewis Mumford Estate.

Geddes, a remarkable polymath who disregarded disciplinary boundaries, advanced a Spencerian view of evolutionary change in society based upon the integrated study of a given region, including its physical surroundings, the population that lived there, and the work they performed. Over the next several years, Mumford heeded Geddes's call for 'regional survey,' the collection of detailed topographical and architectural data, beginning with the New York metropolitan area and, as his health improved, continuing up and down the eastern seaboard of the United States. This activity, reminiscent of Mumford's childhood walks, helped to focus his interest in the study of architecture and city planning, while still embracing Geddes's emphasis on the social, economic, and cultural aspects of regional development.

At about the same time, Mumford discovered Ebenezer Howard's garden city, which, in its compact form and its balance between urban and rural modes of living, dovetailed almost seamlessly with Geddes's regionalism. Among the numerous other writers who influenced Mumford's development at this early stage were the popular novelist H.G. Wells, the playwright George Bernard Shaw, the Arts and Crafts polemicists John Ruskin and William Morris, and the American Transcendentalists Ralph Waldo Emerson and Henry David Thoreau. This autodidactic phase was interrupted briefly by a stateside tour of duty with the U.S. Navy near the end of World War I. By 1919, Mumford had established himself as a promising journalist and an editor for New York's *Dial* and London's *Sociological Review*. In 1922, he published his first book, *The Story of Utopias*, a wide-ranging survey of utopian literature in which many of the following overarching themes of his career are adumbrated: the investment of artistic form with social meaning; the harnessing of technology for the common good; and the restructuring of the built environment on a human scale in balance with the surrounding region. That same year he married Sophia Wittenberg (1899–1997), who was to be his lifelong intellectual confidante and amanuensis.

Mumford's subsequent career can be divided roughly into the following three chronological stages: an early stage when he concentrated on American cultural history and criticism (*c*.1923–31); a mature stage when his focus broadened to

include all of Western civilization (*c.*1932–70); and a late stage when he reexamined his life and work in a series of autobiographical collections of writings, leading to an incomplete autobiography (*c.*1971–82).

Early Career

The 1920s found Mumford in the company of architects and planners as well as writers. In 1923, he joined the Regional Planning Association of America, a group dedicated to the overlapping ideals of Geddes and Howard, with a practical focus on the improvement of housing, the development of garden cities, and the promotion of regionalism. Its members included the editor Charles Harris Whitaker, the architects Clarence Stein and Henry Wright, the naturalist Benton MacKaye, and the housing reformer Edith Elmer Wood. The Association achieved modest successes in their planned communities at Sunnyside Gardens in the New York City borough of Queens – where Mumford and his wife bought their first home – and at Radburn in suburban New Jersey. The Appalachian Trail, a 2000-mile wilderness trail conceived by MacKaye for the mountainous spine of the Eastern United States, is the Association's best known achievement. Although he was not a designer, Mumford acted effectively as the Association's secretary and publicist.

Mumford continued to earn his living as journalist, writing chiefly criticism on a vast range of subjects for such periodicals as the *Freeman*, the *Journal of the American Institute of Architects*, the *American Mercury*, *Commonweal*, and the *New Republic*. He counted several prominent writers among his friends, including Harold Stearns, Paul Rosenfeld, and Waldo Frank. Under the influence of another close friend and fellow writer, Van Wyck Brooks, Mumford published a series of short, but insightful books on American culture in an effort to recover what Brooks termed the 'usable past.' On a superficial level, *Sticks and Stones* (1924) is an unorthodox, social history of American architecture; on a deeper level, it is a condemnation of modern architectural practice. Mumford argues that, following a halcyon early colonial period when vernacular forms predominated, American architecture became lost in a dense thicket of revival styles in its rapacious drive toward territorial expansion. Consequently, he urges his contemporaries to devise a regionally based modern architecture appropriate for the regionally based modern community. *The Golden Day* (1926) examines five writers from the mid-nineteenth century – Emerson, Thoreau, Nathaniel Hawthorne, Herman Melville, and Walt Whitman – whom Mumford believed had created a specifically American literature in touch with the regional landscape and who might be emulated by his contemporaries.

Mumford's first foray into biography, *Herman Melville* (1929), was a direct outgrowth of *The Golden Day* and one of the first studies to wrestle with the author's troubled psyche. *The Brown Decades* (1931) culminated this early period, and it was Mumford's attempt to redress some critical oversights in both *Sticks and Stones* and *The Golden Day*, particularly the cursory treatment of writer Emily Dickinson, landscape architect Frederick Law Olmsted, and architects H.H. Richardson and Louis Sullivan. Furthermore, architect Frank Lloyd Wright and photographer Alfred Stieglitz are treated as transitional figures, whose creative energies bridge the period from the brown decades to the present. Although surpassed by later, more scholarly works, these four books made a significant contribution to the fledgling discipline of American Studies. Moreover, with the unifying color metaphor of its title and the elegance of its prose, *The Brown Decades* endures as a major work of American letters.

Mature Career

The mature stage of Mumford's career extended from the early 1930s to the early 1970s, with only a brief interruption around World War II. During the Depression, Mumford continued to make his living as a journalist, but he confined his output to a few selected journals and magazines. In 1931, he began an extraordinary, 30-year tenure as the architecture critic for *The New Yorker*. His column, 'The Sky Line,' was an important platform from which he mediated between the often conflicting currents of European and American modernism. Many of his architecture columns have been collected and published in *From the Ground Up* (1956), *The Highway and the City* (1963), and *Sidewalk Critic* (1998; Robert Wojtowicz, editor). Virtually unknown today is his short but influential tenure as the magazine's art critic (1932–37). From cubism to surrealism, Mumford embraced most currents of modern art as they took tentative root in the museums and galleries of New York City and elsewhere, although he was less receptive to nonrepresentational experimentation. A selection of these columns have been published in *Mumford on Modern Art* (2007; Robert Wojtowicz, editor). That artists might lead a cultural renewal in post-World War II society forms the major theme of *Art and Technics* (1952), a published collection of lectures originally delivered at Columbia University.

Mumford, however, wanted to be known primarily as a writer of important books on weightier topics. Fresh from his modest success with the American 'usable past' project, he embarked on an ambitious three-part – later expanded to a four-part – series surveying Western civilization that became known as The Renewal of Life. Inspired by Geddes, the series examines medieval and postmedieval European history and culture from distinct perspectives utilizing an implicitly Christian, salvific paradigm. In Mumford's view, the Middle Ages represented an organic unity that was disrupted by the industrial revolution and its dehumanizing effects. Modern technology, coupled with rational planning and cultural renewal, he believed, could restore that lost unity in the present.

The first volume, *Technics and Civilization* (1934) examines technological developments from the turret windmills of Holland to the hydroelectric dams of the United States. He divides his survey into three periods, using terms borrowed from Geddes: the 'eotechnic,' or preindustrial phase; the 'paleotechnic,' or industrial phase; and the 'neotechnic,' or postindustrial phase. Near the conclusion Mumford focuses on what he terms 'basic communism,' a reorientation of the world economy along communitarian lines that derives more from the writings of Edward Bellamy than Karl Marx. Only through careful, human control, Mumford warns, can technology continue to improve the quality of modern life. *The Culture of Cities* (1938), the second and most widely acclaimed volume of The Renewal of Life, traces concurrent developments in urbanism. The last third of this book is devoted to Mumford's

eutopian vision for a new regionally planned order, citing such New Deal projects as the Tennessee Valley Authority and the Greenbelt Towns as examples of what was already being accomplished in the present. Following the book's publication, Mumford was recognized internationally as an authority on urbanism, his status reified by the placement of his photograph on the cover of *Time* (18 April 1938). Somewhat ironically, Mumford had earlier relocated his family from the urban surroundings of Queens to a small farmhouse in upstate New York. This success of *The Culture of Cities*, in turn, led Mumford to publish *City Development* (1945), a collection of urban-themed essays.

The last two volumes of The Renewal of Life, *The Condition of Man* (1944) and *The Conduct of Life* (1951), were met by less critical and popular success as they were, in some ways, eclipsed by the events of the World War that they bracketed. With tempered optimism, Mumford surveys the impact of politics, science, philosophy, and religion on the human personality and argues that the renewal of the individual is the necessary prelude to the renewal of society as a whole. Spread over a period of 17 years, The Renewal of Life had a less lasting impact than such contemporary multivolume studies of Western civilization as Oswald Spengler's *The Decline of the West* (2 vols, 1918–23) and Arnold J. Toynbee's *A Study of History* (12 vols, 1934–61), even though its first two volumes remain landmark studies in their respective areas. Invitations to write planning reports and to teach university courses resulted from the series, despite Mumford's lack of a baccalaureate degree. Earlier he had been a visiting professor at Dartmouth College (1929–35) and a lecturer at Columbia University (1931–35). In 1942, he was appointed visiting professor at Stanford University but left in 1944, as global events took precedence in his writing. Part-time academic roles would take on an even greater importance for Mumford after the War.

World War II transformed Mumford on a personal and professional level, while raising his political profile significantly. Early to recognize the threat of both Fascism and Nazism, he was one of the first left-leaning American intellectuals to advocate the U.S. intervention in the War, a point, which he argued forcefully and eloquently in *Men Must Act* (1939). *Faith for Living* (1940) offered a hopeful vision of a world freed from the totalitarian menace. Tragedy struck Mumford and his wife Sophia in 1944 when their only son – named, significantly, Geddes – was killed in action on the Italian front. Mumford channeled his grief into the writing of *Green Memories* (1947), a moving biography of his son that opens a fascinating window into the complexities of parenting in the middle of the twentieth century. Alarmed by the U.S. detonation of the atomic bomb and its potential for global destruction, Mumford, the interventionist, became – once the threat of totalitarianism had passed – Mumford, the pacifist. He became an ardent foe of the arms buildup of the Cold War, as advances in technology once again threatened to overwhelm human reason. *In the Name of Sanity* (1954) contains many of his essays on this topic. This same pacifism lay at the core of his vocal and written opposition to the Vietnam War during the 1960s.

Dissatisfied with the reception of The Renewal of Life, Mumford spent the next two decades revising basic assumptions and searching further into the past for clues into modern behaviors. A significant portion of this time was spent within the walls of the Academy, where he benefitted from contacts with colleagues in a myriad of disciplines. Mumford was visiting professor at North Carolina State University (1948–52), professor of city planning and Ford Professor at the University of Pennsylvania (1951–56; 1959–61), and visiting professor at the Massachusetts Institute of Technology (1957–61), among other places. He continued writing at a steady pace. *The Transformations of Man* (1956) was his attempt to condense the core message of The Renewal of Life into a single volume. His conclusion that 'One World man' – one who recognizes and synthesizes the rich traditions of all cultures around the globe – was about to emerge in the latter part of the twentieth century has proven to be both prescient and elusive. This reevaluation of past themes continued with *The City in History* (1961), Mumford's best known book. What he intended as a mere revision of *The Culture of Cities*, gradually evolved into an expansive, mostly new survey of Western urbanism, beginning with the Paleolithic cave and concluding with the sprawling twentieth-century metropolis. Missing was his earlier, eutopian vision of a new regional order; instead he forecasted the inevitable decline of cities unless technology, especially the automobile, were brought under control by designers and planners. The book, which won the National Book Award for Non-Fiction in 1962, solidified Mumford's reputation as America's senior authority on urbanism, although his presumptions regarding the efficacy of rational planning on city design were soon challenged by younger writers such as Jane Jacobs. Its success, moreover, spurred Mumford to publish *The Urban Prospect* (1968), yet another collection of his essays on urbanism.

Building on the momentum of *The City in History*, and at a stage in life when many of his contemporaries would have been content to garner praise and awards, Mumford reposited his views on the history of technology in what may be his most ambitious work, the two-volume *Myth of the Machine* (2 vols, 1967, 1970). In volume one, *Technics and Human Development* (1967), he put forward what he believed was his most original insight: that the ancient rulers of the eastern Mediterranean world exerted insidious control over human labor through the employment of massive work units or 'megamachines.' Although he argued in volume two, *The Pentagon of Power* (1970), that modern technology made such human megamachines obsolete, its all-embracing power, left unchecked by human reason, would eventually lead to global destruction. Critical and popular reaction to *The Myth of the Machine* was mixed, but Mumford's critique of the political and scientific establishment, which he identified as masterminding the modern megamachine, won him wide admiration among a younger, more reform-minded generation.

Late Career

Mumford relished his position as a public intellectual, and he spent the last two decades of his life engaged in various activities related to the stewardship of his considerable legacy. He donated his personal papers to the Van Pelt Library of the University of Pennsylvania in Philadelphia and his graphite,

ink, and watercolor sketches – many of them dating to the regional surveys of his youth – to Monmouth College, now Monmouth University, in West Long Branch, New Jersey. Several collections of his earlier writings were published: *Interpretations and Forecasts: 1922–1972* (1973); *Architecture as a Home for Man* (1975; Jeanne M. Davern, editor); and *The Lewis Mumford Reader* (1986; Donald L. Miller, editor). Two autobiographical collections – *Findings and Keepings* (1975) and *My Works and Days* (1979) – preceded the publication of his autobiography, *Sketches from Life* (1982), which covered only the first half of his long life. In addition, Mumford's correspondence with such key friends and colleagues as Van Wyck Brooks, Frederic Osborn, Patrick Geddes, and Frank Lloyd Wright, have been edited and published, providing a fuller understanding of his formidable but engaging personality. Moreover, he received numerous accolades, including the Royal Institute of British Architects' Gold Medal (1961), the Presidential Medal of Freedom (1964), and the National Medal of the Arts (1986). Mumford died quietly at his country home in Leedsville near Amenia, New York, on 26 January 1990.

Critical Assessment

Mumford's belief that society could be improved through rational and ecologically sound planning, marks him as a committed modernist. Postmodern scholars – who emphasize, among other things, the fragmentation of meaning and the failure of planning – have largely ignored him. Nevertheless, most of Mumford's books remain in print, and his name is beginning to appear with increasing frequency in both popular and scholarly contexts. Mumford's wide-ranging legacy as a generalist has been difficult for any one scholar to address; some have selected one aspect of his work, while others have joined together in delivering group assessments. Examples of the latter include the special issue of the American journal *Salmagundi* (Summer 1980) and *Lewis Mumford: Public Intellectual* (1990; Thomas P. Hughes and Agatha C. Hughes, editors), a collection of essays that focus on Mumford's lifelong involvement with technology. As its title suggests, Donald L. Miller's *Lewis Mumford: A Life* (1989) provides a comprehensive and often insightful overview of Mumford's life. Frank G. Novak Jr., focuses more narrowly on one aspect of Mumford's writing in *The Autobiographical Writings of Lewis Mumford* (1988).

In *Beloved Community* (1990), Casey Nelson Blake examines Mumford's relationship to the 'young Americans,' a group of New York intellectuals searching for cultural renewal in the aftermath of World War I. Shuxue Li's *Lewis Mumford: Critic of Culture and Civilization* (2009) discusses Mumford in the context of Karl Marx and Max Weber. Mumford's contributions to architecture and urbanism are the subject of Robert Wojtowicz's *Lewis Mumford and American Modernism* (1996), while Heinz Tschachler discusses Mumford in a German context in *Lewis Mumford's Reception in German Translation and Criticism* (1994). Mark Luccarelli focuses on Mumford's concern for regionalism and the environment in *Lewis Mumford and the Ecological Region: The Politics of Planning* (1995). Numerous dissertations, moreover, have addressed one or more aspects of Mumford's work. As yet, however, little has been written about Mumford outside of the United States, even in Great Britain, where many of his books were published in separate editions.

See also: Art History; Cultural History; Environmental History; Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; History and Literature; History and the Social Sciences; Intellectual History; Marx, Karl (1818–83); Modernization and Modernity in History; Second World War, The; Spencer, Herbert (1820–1903); Urban History; Weber, Max (1864–1920).

Bibliography

Blake, Casey Nelson, 1990. Beloved Community: The Cultural Criticism of Randolph Bourne, Van Wyck Brooks, Waldo Frank, and Lewis Mumford. University of North Carolina Press, Chapel Hill.

Hughes, Thomas P., Hughes, Agatha C. (Eds.), 1990. Lewis Mumford: Public Intellectual. Oxford University Press, New York and Oxford.

Li, Shuxue, 2009. Lewis Mumford: Critic of Culture and Civilization. Peter Lang, Oxford and New York.

Luccarelli, Mark, 1995. Lewis Mumford and the Ecological Region: The Politics of Planning. Guilford Press, New York.

Miller, Donald L., 1989. Lewis Mumford: A Life. Weidenfeld and Nicolson, New York.

Mumford, Lewis, 1922. The Story of Utopias. Boni and Liveright, New York.

Mumford, Lewis, 1924. Sticks and Stones: A Study of American Architecture and Civilization. Boni and Liveright, New York.

Mumford, Lewis, 1926. The Golden Day: A Study in American Experience and Culture. Boni and Liveright, New York.

Mumford, Lewis, 1929. Herman Melville. Harcourt, Brace, and Company, New York.

Mumford, Lewis, 1931. The Brown Decades: A Study of the Arts in America, 1865–1895. Harcourt, Brace, and Company, New York.

Mumford, Lewis, 1934. Technics and Civilization. Harcourt, Brace, and Company, New York.

Mumford, Lewis, 1938. The Culture of Cities. Harcourt, Brace, and Company, New York.

Mumford, Lewis, 1939. Men Must Act. Harcourt, Brace, and Company, New York.

Mumford, Lewis, 1940. Faith for Living. Harcourt, Brace, and Company, New York.

Mumford, Lewis, 1944. The Condition of Man. Harcourt, Brace, and Company, New York.

Mumford, Lewis, 1945. City Development: Studies in Disintegration and Renewal. Harcourt, Brace, and Company, New York.

Mumford, Lewis, 1947. Green Memories: The Story of Geddes Mumford. Harcourt, Brace, and Company, New York.

Mumford, Lewis, 1951. The Conduct of Life. Harcourt, Brace, and Company, New York.

Mumford, Lewis, 1952. Art and Technics. Columbia University Press, New York.

Mumford, Lewis, 1954. In the Name of Sanity. Harcourt, Brace, and Company, New York.

Mumford, Lewis, 1956a. From the Ground Up: Observations on Contemporary Architecture, Housing, Highway Building, and Civic Design. Harcourt, Brace, and Company, New York.

Mumford, Lewis, 1956b. The Transformations of Man. Harper and Brothers, New York.

Mumford, Lewis, 1961. The City in History: Its Origins, Its Transformations, and Its Prospects. Harcourt, Brace, and World, New York.

Mumford, Lewis, 1963. The Highway and the City. Harcourt, Brace, and World, New York.

Mumford, Lewis, 1967. The Myth of the Machine: I. Technics and Human Development. Harcourt, Brace, and World, New York.

Mumford, Lewis, 1968. The Urban Prospect. Harcourt, Brace, and World, New York.

Mumford, Lewis, 1970. The Myth of the Machine: II. The Pentagon of Power. Harcourt Brace Jovanovich, New York.

Mumford, Lewis, 1973. Interpretations and Forecasts: 1922–1972: Studies in Literature, History, Biography, Technics, and Contemporary Society. Harcourt Brace Jovanovich, New York.

Mumford, Lewis, 1975a. In: Davern, Jeanne M. (Ed.), Architecture as a Home for Man: Essays for Architectural Record. Architectural Record Books, New York.
Mumford, Lewis, 1975b. Findings and Keepings: Analects for an Autobiography. Harcourt Brace Jovanovich, New York.
Mumford, Lewis, 1979. My Works and Days: A Personal Chronicle. Harcourt Brace Jovanovich, New York.
Mumford, Lewis, 1982. Sketches from Life: The Autobiography of Lewis Mumford: The Early Years. Dial Press, New York.
Mumford, Lewis, 1986. In: Miller, Donald L. (Ed.), The Lewis Mumford Reader. Pantheon Books, New York.
Mumford, Lewis, 1998. In: Wojtowicz, Robert (Ed.), Sidewalk Critic: Lewis Mumford's Writings on New York. Princeton Architectural Press, New York.
Mumford, Lewis, 2007. In: Wojtowicz, Robert (Ed.), Mumford on Modern Art in the 1930s. University of California Press, Berkeley, Los Angeles, and London.
Novak Jr., Frank G., 1988. The Autobiographical Writings of Lewis Mumford: A Study in Literary Audacity. University of Hawaii Press, Honolulu.
Tschachler, Heinz, 1994. Lewis Mumford's Reception in German Translation and Criticism. University Press of America, Lanham, Maryland.
Wojtowicz, Robert, 1996. Lewis Mumford and American Modernism: Eutopian Theories for Architecture and Urban Planning. Cambridge University Press, Cambridge, New York, and Melbourne.

Museums

Jennifer Barrett, The University of Sydney, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Museums around the world assume many different forms, yet take as their reference point the museum practice of previous centuries. This article acknowledges past practice before identifying contemporary developments in museums as these institutions respond to challenges, locally and internationally, that affect the museum form. What follows are potential directions for museology, recognizing that museums are increasingly responding to, and sometimes initiating, their changing social context.

Introduction

The museum emerged in the late eighteenth century as a public institution, linked with the 'universal project of democracy,' beginning in 1793, during the French Revolution, with the opening of the Louvre in Paris as the first public museum. Many major museums were reconstituted from private collections to become significant sites tracking material culture, as well as scientific and cultural achievements of communities, cultures, and places through the ages. They came to be *indicative* sites for the international politics of imperialism and the ensuing struggles that followed over the next two centuries. Museums, in this sense, include the building and its roles, activities, collections, engagement with its visitors, and its displays. Museums have been places with which to think about the history of ideas and their relation to society. Museums also link societies, through international networks and practices such as the exchange of ideas, objects, and controversies. Contemporary museum practice, like many academic disciplines, has responded to changes in societies around the world, and continues to do so.

The Origins of Museums

The word 'museum' is Latin from the Greek word '*mouseion*' (Alexander, 1979). The term initially meant 'the abode of the muses' which were believed to be groves on Mounts Parnassus and Helicon. Later these became temples, then universities with many colleges (Dixson, 1919: 3). Some of the earliest museums were in cities such as Alexandria, Athens, and Rome. They were associated with knowledge creation and dissemination (similar to the role of modern universities), or were devoted to displaying captured treasures. With the destruction of the city of Alexandria, the term 'museum' almost disappeared, "and was only revived with the arts and sciences about the middle of the seventeenth century" (Dixson, 1919: 3). While the art collections of royalty and wealthy buyers were the forerunners of public art galleries, and the royal menageries became the predecessors of modern zoological gardens, the museum emerged from the private 'cabinet of curiosities' to become the public collection of history, anthropology, geography, and technology. Debate about the birth of the 'public museum,' and which was the first, has wavered between The British Museum (which developed from the will of Sir Hans Sloane, who died in 1753 and left his massive private collection as the basis for the establishment of The British Museum) and the Museé de Louvre.

The fifteenth-century Medici Palace in Florence is often considered the precursor, with its select invitations to visit, its acquisition and collection practices, displays, and philanthropy. The Medici's use of its collections and its granting of access to them shifted from the private realm and signaled the way in which museums would become signifiers of knowledge and power in new ways for centuries onwards (Alexander, 1979). There are other important developments in the history of museums that are worth acknowledging: notably the first university museum, which was opened in Basel in 1671, and the Ashmolean Museum in Oxford, which was established as the first public natural history museum in 1683 (Barrett and McManus, 2007).

The Evolution of the Modern Museum

Museums in the Eighteenth Century

The 'modern museum,' from which our contemporary museums have emerged, was an important part of changing social values and practices in the nineteenth century. It was also part of the experience of modernity. Not only did the number of cultural institutions increase, but also access to museums and other cultural institutions was more widespread. Moreover, the displays were more likely to reflect the world of the new merchant classes and the everyday experience of the city, rather than the world of royalty.

Prior to the late eighteenth century, 'museums' were generally "collections of valued objects [that] formed a part of the cultural accessories of power in contexts in which it was the organisation and transmission of power within and between ruling strata rather than the display of power before the populace that was the point at issue" (Bennett, 1995: 27). The opening of the Louvre in Paris as the first public museum in 1793 was a significant departure from the curiosity cabinets inside palaces, the collections of the clergy, and the collections of the merchants, which imitated the collections of royalty (see Hooper-Greenhill, 1992). The idea of the Louvre palace being a royal art gallery was advocated much earlier by Lafont de Saint-Yenne in 1747, largely because the royal collection was in

storage and was inaccessible and a royal art gallery would be a monument to the king and the glory of France should it be opened for the display of great works of art (Duncan and Wallach, 1980).

It is important not to mistake the eighteenth-century art museum for the fully formed public, national institution of the nineteenth century. According to Nick Prior, the state was yet to play a formative role in the eighteenth century, and the nation was yet to be brought into line with the state. Furthermore "culture was yet to be subsumed under the latter's socio-political remit" (2002: 35). He adds

> …museums in the eighteenth century were not usually owned by the state on behalf of the people as a corollary to citizenship, governance and democracy. Visitors were subjects not citizens and power was represented 'before' them rather than 'for' them.
>
> Prior, 2002: 35

The change from the princely collections of the late eighteenth century to the public museum of the twentieth century was a journey replete with debate about what constitutes a museum, what role a museum should have in society, who attended museums and what benefit was gained by doing so, and how museum practice should accommodate visitors. Many of these concerns, and some of the answers, predate the 'new museology' that was packaged and presented by Peter Vergo (1989).

Emerging Modern Public Museums in the Nineteenth Century

By the middle of the nineteenth century, the emerging modern public museums were a vital part of industrialization and colonial processes. Museums cataloged and presented socio-economic and technological change to their audiences. The major imperial museums in London, Paris, and later Berlin displayed the wealth and curiosities of each country's dominions to the central cities of those empires, which often had significant environmental and cultural impact on those colonies and territories (Barrett and McManus, 2007). The majority of today's great art museums of the Anglo-Saxon world were created in the Victorian Age; this influenced cultural development in the main metropolitan centers and the development of museum culture all over the world (Barringer, 2006).

The nineteenth century saw a rapid and unprecedented growth in the number of museums. While the study of natural history is much older than the nineteenth century (see Barrett and McManus, 2007), by 1900 there were 150 museums of natural history in Germany, 250 in Britain, and 300 in France (Sheets-Pyenson, 1988). The increasing number of museums was not confined to the imperial powers: Sheets-Pyenson (1988) noted that in 1900 there were 250 natural history museums in the United States, while Thomas Storie Dixson, in 1919, observed in 1903 that there were 31 natural history museums in New York state (population 7 million) and 10 in California (population 1.5 million).

The growth in the number of museums resulted from the confluence of technological progress and ideas about civilized societies. The technological progress of the eighteenth and nineteenth centuries enabled various countries in Western Europe to explore and colonize distant lands. The display of objects from these colonies in the largest cities of the imperial powers not only showed their citizens what they possessed, but also aimed to fulfill the earlier goals of leading scientists, such as Francis Bacon, to broaden knowledge and challenge the intellect. The orderly display of objects, their perusal, and the resultant education of the viewer were intended to engender a fascination with the world, and a respect for authority, particularly from the working classes.

In addition to the large public museums in the most powerful cities, private museums existed where individuals, such as Sir John Soane, collected objects of personal interest. Many of these private museums have not survived, thereby making the Sir John Soane Museum in London a modern-day rarity, much like the Pitt Rivers Museum in Oxford, of interest partly because it is a museum of a museum.

The Educational Value of Museums in the Eighteenth and Nineteenth Centuries

Museum Buildings

The idea that museums and collections had a purpose and value other than being private collections of personal interest of their owner was the foundation for the public museum. The governments of different countries began to acquire by various means the collected treasures of individuals, and established collections that would be open for the public. The buildings that these collections were housed in were often unsuitable for the 'reception and proper display of works of art and archeology.' This influenced the ability of museums to display the objects appropriately, and "made both casual observation and careful study of most of the objects a burden instead of a pleasure" (Dana, 1917 cited in Anderson, 2004: 17).

The critique of museums as being 'gloomy' and 'dull' highlighted some of the concerns in the nineteenth century about the buildings that museum collections were housed in, and the role of the museum. John Wood (1887: 384) wrote that "full of interest to the expert, there is no concealing the fact that to the general public a museum, of whatever nature, is most intolerably dull…." He proposed to create separate museums for the education of pupils and for the higher educated people who would undertake more complex study in the museum.

Relationships with the Public

This concern about the educational levels museum also connects with debates about whom the 'public' was that public museum should be attempting to attract. Much was made of the importance of educating the working man and his family. This involved both the education and the entertainment of the working classes, often to lure them away from public houses (i.e., taverns). Various accounts of late-nineteenth-century museums highlight the challenges that some of these museums faced in being public, while observations about the use of museums by various social classes questioned the value of

claims about the civilizing and educating role of museums for 'the masses.'

In a related vein, Wood wrote of the educational value of museums. He desired to "construct a museum especially adapted to the despised Tom, Dick and Harry, which would amuse them, should be of such a nature as to compel them to take an interest in the subject, and perchance to transform them into the Thomas. H. Huxleys, Richard Owens, and P. Henry Gosses of the next generation" (Wood, 1887: 391). Although some museum curators and boards realized the importance of entertainment and education in the nineteenth century, these functions continued into the twentieth century, well before the rise of the 'new museology.'

The counter argument to this perspective, and one that is sometimes raised about the educational value of contemporary approaches to museum practice, is the writings of Stanley Jevons in the late nineteenth century. Jevons was not against the educational value of museums, but he argued that it varied significantly and should not be assumed. According to Jevons (1883: 55), the "advantage which an individual gets from the visit may vary from nil up to something extremely great," and "the degree of instruction derived is quite incapable of statistical determination." Importantly, Jevons critiqued the notion of visitor studies, which became even more popular in the twentieth century, by noting

> …There seems to be a prevalent idea that if the populace can only be got to walk about a great building filled with tall glass-cases, full of beautiful objects, especially when illuminated by the electric light, they will become civilized. At the South Kensington Art Museum they make a great point of setting up turnstiles to record the precise number of visitors, and they can tell you to a unit the exact amount of civilizing effect produced in any day, week, month, or year. But these turnstiles hardly take account of the fact that the neighbouring wealthy residents are in the habit, on a wet day, of packing their children off in a cab to the so-called Brompton Boilers, in order that they may have a good run through the galleries.
>
> Jevons, 1883: 55

While Jevons highlighted the limitations of equating educational value with the counting of visitors, it was apparent that museums were a form of entertainment as well as supposedly being educational. Concerns about entertainment values of the museums overriding educational values have continued to this day.

The growth of museums in the nineteenth and twentieth centuries can be partly attributed to the notions of educating and civilizing societies, particularly the lower classes. Museums were an integral part of the colonial project during the nineteenth century. Henare (2005: 15), writing about New Zealand in 1840, noted "…the establishment of a learned society with its ubiquitous museum was among the earliest tasks of founding a new settlement." This notion of civilization also operated in rapidly growing cities in the United States. The development of a major public museum announced that the city had 'arrived.' As art historian Carol Duncan (1995: 54) observed, "a big, showy art museum could announce to both national and international business and banking communities the arrival of a city financially and politically." The same author recognized that museums 'conferred social power' and that they reinforced class boundaries while simultaneously appearing "as unifying and even democratizing forces in a culturally diverse society" (Duncan, 1995: 54).

Education as Power

Social and cultural theorist Tony Bennett has made the sharp observation that "while late-nineteenth-century museums were thus intended *for* the people, they were certainly not *of* the people in the sense of displaying any interest in the lives, habits, and customs of either the contemporary working classes or the laboring classes of pre-industrial societies" (Bennett, 1988: 64, emphasis in original). Pierre Bourdieu and Alaine Darbel (1969, cited in Bourdieu et al., 1990) argued that museums, particularly art museums, purport to be accessible to all, but are not accessible to the working classes. According to these authors, even if there is no admission charge to enter, attendance is voluntary and working-class people do not have the cultural capital to feel comfortable in these spaces or to understand how to decode the objects. Bennett's point (1988: 64) extended to the display of objects in museums. One function of museums was to impress upon visitors the power of the state. Initially, this meant that following on from the private collections of wealthy individuals, in the cabinet of curiosities, displays were comprehensive: virtually everything collected was displayed. The criteria for collection and display of material were often uniqueness, distance between the location of the material and its display, and the individual (sometimes aesthetic) properties of the object. In the late nineteenth century, as collections expanded rapidly due to colonization, displays in many major museums became selective. Sir William Flower, Director of the Natural History Museum in London, advocated displaying fewer objects in the public exhibition so that lay visitors had space to view the objects and could gain a general understanding and appreciation of nature (Flower, 1898; Sheets-Pyenson, 1988). The full collection was available to specialist researchers to conduct scientific inquiry.

Museums, along with various organizations, such as the Royal Society in London, were responsible for promoting scientific and cultural advancement, and were influential in the promulgation of ideas about nature. Museum curators taught at universities, and museums were involved in research activity (Yanni, 1999). The modern museum was a product of the late eighteenth and nineteenth conceptions of 'the public' and of citizenship. Notes that the modern museum, as part of a new relationship between the public and the state via public cultural institutions, "cannot be adequately understood unless viewed in the light of a more general set of developments through which culture, in coming to be thought of as useful for governing, was fashioned as a vehicle for the exercise of new forms of power" (Bennett, 1995: 19).

Changes to Museums in the Twentieth Century

Changing Functions of Museums

The early part of the twentieth century saw the continuation of many trends established in the latter part of the nineteenth century. This included the development of new museums and

expanding collections of established museums. During the twentieth century museums continued to increase in numbers throughout the world, with the United States alone having an estimated 17 500 museums in 1995 (American Association of Museums, 1999). Robinson (2006) noted that most American museums have been established since 1970. Visitor numbers have also increased substantially, but perhaps more importantly than quantifying the number of museums and their visitors, it is necessary to discuss the significant changes to museums during the twentieth century.

In terms of museum collections, Theodore Low in 1942 preempted the latter debates about the role of the curator and the idea of popularizing the museum by commenting on the conservative nature of curators, who he saw as focusing on the arrangement of exhibitions without sufficient consideration of the needs of the public and how they would understand the arrangement. Low also critiqued the directors of the museums, who he said, following tradition, "have been much more interested in the building up of collections and in the scholarly prestige of the institution than in making it useful" (Low, 1942 cited in Anderson, 2004: 33).

Other authors saw a major change between the museums of the nineteenth and twentieth century. Lucas (1908 cited in Genoways and Andrei, 2008: 59) wrote about the changes in public museums in the 25 years prior to this time: "one of the great differences between the old museum and the new is that the one displayed objects while the other aims to illustrate ideas." In other words, the culture of collecting was being maintained, but in many museums the culture of display was changing. He cites the example of displaying birds. Rather than having many birds on display, which would have little meaning for the average museum visitor, the new museum would display "a single group showing one of these birds at home, the purpose of which is to show the conditions under which birds live and to interest the beholder in the study of bird life" (Lucas, 1908; cited in Genoways and Andrei, 2008: 59).

Emergence of the Universal Survey Museum

As new cities emerged, they generally developed new museums. This process accelerated later in the twentieth century, partly as a result of increased leisure time, greater awareness of the importance of history, and the ongoing desire of cities and towns to gain recognition and status. Among the museums, the presence of highly visible universal survey museums such as the Metropolitan Museum of Art in New York and the National Gallery of Art in Washington, D.C., remained particularly noteworthy and sometimes controversial.

The Universal Survey Museum, an idea derived from the Louvre and nineteenth-century thought about collecting, displaying, and the role of museums in countries that were politically and economically powerful, reached its zenith in the United States in the twentieth century. Universal Survey Museums were based on the idea that elaborate, dedicated museum buildings housed not only art but also material culture significant to history. According to Duncan and Wallach (1980: 466), the National Gallery of Art in Washington, D.C., which was completed in 1941, is "perhaps the most complete realization of the universal survey idea in America." It is a museum "which could be anywhere in the Western world" and "claims the heritage of Greco-Roman civilization as an abstract and universal value" (Duncan and Wallach, 1980: 466).

The proliferation of museums in recent times is partly linked to the demise of the universal survey museum idea. Museums in the twentieth century became more specialized, and began to focus on the needs of particular audiences and to represent cultures that were absent from the great universal survey museums.

Funding

In addition to the recognition of diverse constituencies (which could be seen as a postmodern approach to what had formerly been called 'the public'), many museums in the late twentieth century engaged with visitors of the museum as consumers and customers. This development arose in part because of the attitude of some governments toward the funding of museums, which, along with other public institutions such as libraries, were expected to be more self-sufficient and able to work within the opportunities and constraints of market forces. One response to this situation was the creation of the 'blockbuster' exhibition, for which visitors would pay and then consume the accompanying merchandise being sold in the museum shop, now expanded and easily accessible as visitors left the museum building. Emma Barker (1999: 127) noted "although large-scale, highly publicized art exhibitions drawing several thousand visitors took place as early as the mid-nineteenth century, the blockbuster show as we know it is a relatively recent phenomenon." She suggested that the main characteristics of the blockbuster exhibition are that it is shown in two or three cities, that it is almost always sponsored by the corporate world, that it has advance booking and should attract a minimum of 250 000 visitors (Barker, 1999). Barker (1999) gives Thomas Hoving, the Director of the Metropolitan Museum of Art in New York from 1967 to 1977, the credit for conceiving the blockbuster format as a way of popularizing the museum and countering claims of elitism, although Prior (2005) dates it earlier to the Fourth Plan (1962–65) and André Malraux in Charles de Gaulle's government in France. The continued existence of the blockbuster is likely due to the combined impact of potentially high financial returns and the ability of the museum to avoid charges of elitism. The blockbuster approach has its downsides, including long queues, financial risk if not well attended, possible closure of other exhibitions, and the potential siphoning of visitors away from smaller, less central museums.

New Museology

Despite, and in some cases because of, these developments in many museums, it is apparent that by the late twentieth century museum practice in places such as Australia, Canada, the USA, New Zealand, and the UK needed to respond to the different experiences of diverse communities they were trying to attract. Many of the critics were coming from outside the museum sector. One response to their critiques, and a major compilation of them, was the emergence of the 'new museology' (Vergo, 1989), which can be characterized as an approach promoting access and participation in the functions of the museum by diverse audiences so as to reflect the democratic

project of the modern museum. New museums emerging in places such as Noumea (Tjibaou Cultural Centre) and Vanuatu (Vanuatu Cultural Centre and National Museum) celebrate distinct local cultural practices (something akin to community cultural centers), and highlight existing cultural heritage while having the capacity to intersect with international museum networks on matters such as repatriation and illegal trafficking of cultural heritage. Diversity and identity politics have highlighted this approach and still tend to inform contemporary museum practices in the early twenty-first century.

Attempts to further democratize the museum – not just in terms of what was represented within museums but how they operated as democratic institutions – would change the public engagement via exhibitions, collecting, and programing. Museums have thus remained a complex and important part of the public sphere that engage not only within the boundary of local communities and nation-states but transgress these territories to engage with international museum networks.

The Contemporary Museum

Museums and Social Change

Within such a framework, the strategies used by museums seek to inspire individuals and publics to reflect deeply on their place in the world and the experience of others, whether it is about the history of imperialism, colonialism, or the resulting development of collections and their contemporary legacy for communities and nations. The complexity of these issues is acknowledged and embraced.

Recognition of the complexity of museums, and the impossibility of being able to be all things to all publics, has resulted in different forms of museology: where in the nineteenth century there was a singular museology, in the twenty-first century there are various approaches such as "appropriate museology" (Kreps, 2008), or "collaborative museology" (Schultz, 2011) and a social inclusion approach to museum practice (Sandell, 2002). The participatory museum (Simon, 2010) has also been promoted by museums that are seeing increased engagement in their virtual form with the use of crowd sourcing to connect communities across the institution and to enhance knowledge of collections and their uses within situated communities and beyond. Developments in new media in the late twentieth century soon went beyond interactivity to the notion of the dialogic museum (Witcomb, 2003). While the Web made accessible information about collections, their location, and basic information for museums and their publics, museum blogs, ways of adding information about collections by the publics and new forms of interaction with audiences have became central to many museum functions.

Human Rights

As discussed previously, museum developments since the twentieth century result not from more traditional forms of collecting and display that characterized the nineteenth century, but from using archives, both personal and state-owned, to produce new experiences of the museum. Such museums are not necessarily founded on the basis of collections, since collections may not exist due to the subject matter, but on principles of justice and democracy, the rationale that museums are effective places to commemorate and memorialize, and that they may also be sites of negotiation and contribute to restoring social justice.

An increasing number of museums and exhibitions are focusing on producing experiences for audiences that require an affective engagement (Witcomb, 2013). Rather than emphasizing their collections or a well-received representation of an event, period, or idea, these museums and their exhibition practices use strategies to produce reflective experiences with a view to inspiring deeper and more critical engagement with subject matter. To some extent, we see this with the rise of human rights museums and with memorials commemorating atrocities, such as genocide.

A number of different kinds of Human Rights museums currently in operation have changed contemporary understandings of what a museum can or should be. A group of museums dedicated to overcoming social injustices, particularly in the area of human rights, includes (according to Terence Duffy in the late 1980s) Museums of Remembrance, Holocaust and Genocide Museums, Museums of Slavery, Museums of African-American Civil Rights, Prison Museums, and Museums of Torture.

In 1999, the International Coalition Of Sites of Conscience, a network of historic sites, was founded by Ruth Abram and dedicated to remembering past infringements on justice while making explicit these infringements' contemporary legacies (see www.sitesofconscience.org). At its founding, the Coalition comprised nine sites on five continents and included in its membership the District Six Museum (South Africa), the Gulag Museum (Russia), the Liberation War Museum (Bangladesh), the Lower East Side Tenement Museum (the USA), and the Maison des Esclaves (Senegal).

Since the founding of many of these museums, a network of like-minded institutions has formed. The origins of institutional members range from those of established democracies (such as the Canadian Museum for Human Rights in Winnipeg, Manitoba) to those of fledgling democracies in postconflict societies, such as the South African Holocaust Foundation. Some institutions display Human rights and Holocaust exhibitions within a museum, e.g., The Imperial War Museum, London.

In 2010, the Federation of International Human Rights Museums (FIHRM), led by the National Museums Liverpool, was founded to build on an earlier initiative of the Coalition of the Sites of Conscience (under the auspice of the International Council of Museums). FIHRM is premised upon two key assertions: first, the roles of museums are changing; and second, museums have acknowledged a growth in their social responsibility (see www.fihrm.org). To meet such expectations, museums may recast existing collections to engage with communities, or use new technologies to link the museum or site to international networks with similar concerns. Digitization of collections and social media allows new ways to link with audiences, educational institutions, and organizations relevant to a museum's strategic goals.

The Future of Museums

The recent proliferation of museums is likely to continue into the future, despite funding constraints in many countries.

The development of new museums whose intended audiences are currently marginalized, the trend toward more commemoration of events, and the desire to attract tourists and inward investment will all contribute to an increase in the number of new museums in the developed countries. Simultaneously, many existing museums will be redeveloped, expanded, and 'rebranded' to achieve relevance (again). This is partly in response to the same forces creating new museums.

There will be more museums in countries that do not share much heritage with the existing museums in developed countries. Some of the newly rich Arabian oil states have consciously planned a development strategy beyond exploiting their finite reserves of oil, and they are using the revenue generated from the sale of this fossil fuel to develop a range of cultural industries. Museums are central to their cultural industries, which include both the establishment of new museums and the importation of branches of existing universal survey museums, such as the Louvre. Newly powerful countries such as China, and perhaps in the future India, will display their power through museums, in similar ways as European countries and the United States in the nineteenth and twentieth centuries. This is already occurring, with the creation of a massive museum complex known as M+ (Museum Plus) on reclaimed land in West Kowloon, in the Hong Kong Special Administrative Region. Earlier conceptions of this development included branches of the universal survey museums and the likelihood of an internationally recognized museum architect, before public pressure led to the plans being abandoned in favor of a locally driven process of consultation, discussion, and development.

Issues of repatriation are also likely to be foregrounded, and highly controversial. Claims for heritage and possession will broaden from a relationship between universal (and sometimes university and city-oriented) museums with indigenous people to include relationships between universal survey museums in the United Kingdom, France, Germany, and the United States of America with powerful museums, strongly supported by government, in the People's Republic of China. These museums are emerging not just in a different era, but also in different cultural contexts from the museums of so-called developed countries. The notions of 'public,' 'being public,' and 'public space' will inevitably be different in relation to these museums.

The engagement with the Internet is both a challenge and opportunity for the future of many museums. The digitization of collections and the use of social media, while appearing to be a gateway for new forms of 'audience engagement,' have revealed the limitations of the methods used in museums to understand audiences and publics. In many ways, technological developments in museums have moved faster than policy and planning. The potential of social media to usher in a new era in museum practice brings new pressures, not only in the securing and allocation of resources, but also in how museums argue that these practices enable the museum to be *more* public than their predecessors. Fundamental questions of engagement remain, such as maintaining inclusivity beyond the voices of the most frequent bloggers. Social media has caused, and will continue to stimulate, changes from the predigital era practices of museums, but it will not be a panacea for the challenges facing museums today and into the future. Nor is it immune to fundamental questions asked of museums. Embracing the complexities inherent in concepts such as the public, public space, public sphere, and community will still be required.

Rather than seeing the passing of the modernist idea of a museum, contemporary museum practice shows its power relationship changing, albeit unevenly, around the world. Museums will continue to reflect dominant systems of knowledge and power in broader society, changing as they do. There are times, as in the work on repatriation in the 1970s and 1980s, when museums were ahead of institutions of the state. As these developments were influenced and in many ways tied to social movements such as civil rights, they worked with these discourses to make change in the museum sector. Collaboration and reciprocity were central and remain so in this field. As disciplines of knowledge have evolved and developed, museums have changed.

There will always be a role for what is becoming known as 'the traditional museum' – a building displaying artifacts for the education and entertainment of visitors. Over time some of these museums will choose to adapt to new technologies and embrace the complexities that this development entails. In fact, there are few examples of museum display techniques and curatorial practices that a curator from the nineteenth century would immediately recognize. In the future museums are likely to be as diverse, if not more so, than the museums that currently exist, with one potential future being a plethora of museums, engaged with their diverse publics and with each other through traditional communications and the innovative use of digital technologies. The challenge will be how to manage the relationship with their diverse communities – virtual and on the physical site. Different modes of access for visitors will not be the only significant change. How the museum operates as a producer of knowledge for and with communities and research purposes will change further. The roles of collecting, collections, and exhibitions within these changing environments are continually being renegotiated. With this come new ethical concerns, about cultural property and its uses. As outlined above, creating reflexive experiences is emphasized.

In conclusion, many ongoing public debates around the world are reexamining the role of the museum in contemporary societies, and causing institutional changes. Contemporary museums are always in a 'period of transition,' and attempts to renew them are continually made. As with previous times in museum history, the present is a complicated and complex period, affected by a range of contradicting demands made on the museum: to be representative, to consider political thinking, collecting, and display practices, issues of representation, and the contemporary world in which we live and which links us in new ways.

See also: Commemoration; Cultural Capital and Education; Cultural Heritage; Organizations and Culture.

Bibliography

Alexander, E.P., 1979. Museums in Motion: An Introduction to the History and Functions of Museums. American Association for State and Local History, Nashville.

American Association of Museums, 1999. Museums FAQ. http://www.aam-us.org/aboutmuseums/abc.cfm#how_many (accessed 29.10.09.).

Anderson, G. (Ed.), 2004. Reinventing the Museum: Historical and Contemporary Perspectives on the Paradigm Shift. Alta Mira Press, Oxford and Lanham, MD.
Barker, E., 1999. Exhibiting the canon: the blockbuster show. In: Barker, E. (Ed.), Contemporary Cultures of Display. Yale University Press in conjunction with The Open University, New Haven and London, pp. 127–146.
Barrett, J., McManus, P., 2007. Civilising nature: museums and the environment. In: Birch, G. (Ed.), Water Wind Art and Debate. University of Sydney Press, Sydney, pp. 319–344.
Barringer, T., 2006. Victorian culture and the museum: before and after the white cube. Journal of Victorian Culture 11 (1), 133–145.
Bennett, T., 1988. Museums and 'the people'. In: Lumley, R. (Ed.), The Museum Time Machine: Putting Cultures on Display. Comedia/Routledge, London, pp. 63–85.
Bennett, T., 1995. The Birth of the Museum: History, Theory, Politics. Routledge, London and New York.
Bennett, T., 1998. Culture: A Reformer's Science. Sage, London.
Bhabha, H., 1994. DISSEMINATION: time, narrative and the margins of the modern nation. In: The Location of Culture. Routledge, London and New York, pp. 139–170.
Bourdieu, P., Darbel, A., Schnapper, D., 1990. The Love of Art. European Art Museums and Their Public (1969) (C. Beattie, N. Merriman, Trans.). Stanford University Press, Stanford, California.
Buckingham, J.S., 1849. National Evils and Practical Remedies. P. Jackson, London.
Chakrabarty, D., 2002. Museums in late democracies. Humanities Research 9 (1), 5–11.
Cuno, J., 2004. The object of art museums. In: Cuno, J. (Ed.), Whose Muse? Art Museums and the Public Trust. Princeton University Press and Harvard University Art Museums, Princeton, NJ, and Cambridge, MA, pp. 49–75.
Dixson, T.S., 1919. Australian Museum, Sydney; Lecture on Its Origin, Growth and Work. Printed copy of a lecture delivered on June 10, 1919. Australian Museum, Sydney.
Duitz, M., 1992. The soul of a museum: commitment to community at the Brooklyn Children's Museum. In: Karp, I., Mullen Kreamer, C., Lavine, S. (Eds.), Museums and Communities. Smithsonian Institution Press, Washington and London, pp. 242–261.
Duncan, C., 1995. Civilizing Rituals: Inside Public Art Museums. Routledge, London.
Duncan, C., Wallach, A., 1980. The universal survey museum. Art History 3 (4), 448–469.
Flower, W., 1898. Reprinted 1972. Essays on Museums (And Other Subjects Connected with Natural History). Books for Libraries Press, Freeport, New York.
Genoways, H., Andrei, M.A. (Eds.), 2008. Museum Origins: Readings in Early Museum History and Philosophy. Left Coast Press, Walnut Creek, CA.
Henare, A.J.M., 2005. Museums, Anthropology and Imperial Exchange. Cambridge University Press, Cambridge.
Hooper-Greenhill, E., 1992. Museums and the Shaping of Knowledge. Routledge, London.
Jevons, W.S., 1883. The use and abuse of museums. In: Methods of Social Reform and Other Papers. Macmillan, London, pp. 53–81.
Kreps, C., 2008. Appropriate museology in theory and practice. Museum Management and Curatorship 23 (1), 23–41.
Lucas, F., 1908. Purposes and aims of modern museums. Proceedings of the Staten Island Association of Arts and Sciences 2, 119–124.
Preziosi, D., 2003. Brain of the Earth's Body: Art, Museums, and the Phantasms of Modernity. University of Minnesota Press, Minneapolis.
Prior, N., 2002. Museums and Modernity: Art Galleries and the Making of Modern Culture. Berg, Oxford.
Prior, N., 2005. A question of perception: Bourdieu, art and the postmodern. The British Journal of Sociology 6 (1), 123–139.
Proctor, N., 2013. Crowdsourcing – an introduction: from public goods to public good. Curator 56 (1), 1–2.
Robinson, F.W., 2006. Learning by looking: the future of museums. In: Genoways, H. (Ed.), Museum Philosophy for the Twenty First Century. AltaMira Press, Lanham, MD, pp. 161–164.
Russo, A., 2011. Transformations in cultural communication: social media, cultural exchange, and creative connections. Curator 54 (3), 327–346.
Sandell, R., 2002. Museums and the combating of social inequality: roles, responsibility, resistance. In: Museums, Society, Inequality. Routledge, pp. 3–22.
Schultz, L., 2011. Collaborative museology and the visitor. Museum Anthropology 34 (1), 1–12.
Sheets-Pyenson, S., 1988. Cathedrals of Science: The Development of Colonial Natural History Museums during the Late Nineteenth Century. McGill-Queens University Press, Kingston and Montreal.
Siegel, J. (Ed.), 2008. The Emergence of the Modern Museum: An Anthology of Nineteenth Century Sources. Oxford University Press, Oxford, pp. 245–246.
Simon, N., 2010. The Participatory Museum. California, Santa Cruz. Museum 2.0.
Taylor, B., 1999. Art for the Nation: Exhibitions and the London Public 1747–2001. Rutgers University Press, New Brunswick.
Vergo, P. (Ed.), 1989. The New Museology. Reaktion Press, London.
Witcomb, A., 2003. Re-imagining the Museum: beyond the mausoleum. Routledge, London.
Witcomb, A., 2013. Understanding the role of affect in producing a critical pedagogy for history museums. Museum Management and Curatorship 28 (3), 255–271.
Wood, J.G., 1887. The dullness of museums. Nineteenth Century 21, 384–396.
Yanni, C., 1999. Nature's Museums: Victorian Science and the Architecture of Display. The John Hopkins University Press, Baltimore.
Zolberg, V., 1995. What price success? Museum International 47 (2), 60–63.

Relevant Websites

www.fihrm.org – Federation of International Human Rights Museums.
www.icom.museum – International Council of Museums (ICOM).
www.sitesofconscience.org – International Coalition of Sites of Conscience.

Music: Anthropological Aspects

Veit Erlmann, University of Texas at Austin, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article surveys the main trends in the anthropological study of music from the 1980s to the present. These trends are as follows: (1) music as culture and the mediation of social process, (2) music as discourse and the nature of musical meaning, (3) music and identity, (4) music and modernization, (5) music and technology, (6) music and knowledge, and (7) music as sonic practice.

Anthropological aspects of music are the main object of ethnomusicology and include all phenomena pertaining to music as a social and cultural activity. The study of the anthropological aspects of music is concentrated in the following main theoretical areas: (1) music as culture and the mediation of social process; (2) music as discourse and the nature of musical meaning; (3) music and identity; (4) music, modernity, and globalization; (5) music, media, and technology; (6) music as knowledge; and (7) music as sonic practice.

Anthropological aspects of music include all those elements pertaining to music as a social and cultural phenomenon. Currently, research on these aspects is concentrated in ethnomusicology. Other disciplines, such as musicology, cultural studies, media studies, and popular music studies, also increasingly take into account anthropological aspects, but have been slow in adopting more context-oriented approaches and ethnographic field methods. Anthropologists and social scientists, in turn, have generally shown only moderate interest in what the members of the societies they study hear, sing, or play. Within ethnomusicology, there exists a considerable plurality of paradigms, models, and methodologies, but the majority of scholars have approached their respective (geographic or historical) area of study from one or several of the following aforementioned perspectives: (1) music as culture and the mediation of social process; (2) music as discourse and the nature of musical meaning; (3) music and identity; (4) music, modernity, and globalization; (5) music, media, and technology; (6) music and knowledge; and (7) music as sonic practice.

Music as Culture and the Mediation of Social Process

Ethnomusicologists in the past have tended to treat music as somehow separate from culture, almost as a thing in itself. They assumed that culture was separated into different spheres such as politics, economics, and so on, each of which could be brought into some sort of relationship with music. According to one approach inspired by British social anthropology, ethnomusicology was the 'study of music in culture' (Merriam, 1964), with culture being little more than some ill-defined social context and music having the function of somehow maintaining the coherence of society. Another approach, modeled on the ideas of Ruth Benedict (among others), saw each society as being constructed around a fundamental principle or pattern that was then imparted to various cultural realms such as religion, art, music, and so on. Thus, Charles Keil, in a classic study of African music, found among the Tiv of Nigeria an 'expressive grid' that manifested itself in the circles, patterns, and structures of Tiv handicrafts, architecture, and songs (Keil, 1979).

Since the 1980s, scholars have begun to study music as culture. Influenced by the work of anthropologist Clifford Geertz and what has been called the 'interpretive turn' in the humanities and social sciences, researchers are thinking about music not so much as a secondary reflection of a more essential social structure or reality than as itself being a social activity mediating other social processes. Instead of gaining its meaning by somehow becoming part of a context, music itself is a context through which other forms of equally primary activity achieve meaning (Herndon and McLeod, 1981).

This new direction had three major consequences for the anthropological study of music. The first is that music makers and their music become dynamic forces, actively intervening in the shaping of cultures that are no longer seen as more or less coherent totalities or realities. The focus on culture as a heterogeneous field of conflicting forces shifts the attention from music as a product, a static expression of collective identity – as in 'the music of the Zulu,' 'music in village X,' and so on – to music as a process, as a mediating agent in the overall process of cultural negotiation. Or, as prominent British ethnomusicologist John Blacking put it, music was not only 'humanly organized sound,' but also 'soundly organized humanity' (Blacking, 1973).

The second consequence concerns the nature of scholarly knowledge and discourse about 'other' musics. For the reformulated critical disciplines, truth could no longer be something that was found from a neutral and distanced position. Following Edward Said's argument in his seminal book *Orientalism* (1979), the point of scholarship is not to deliver the 'true' other. Difference can no longer denote the distance between cultures such as the West and the East, just as it no longer refers back to an overarching scholarly (presumably Western and objective) frame of reference. Rather, difference, in the new disciplines, means the interplay of a multitude of competing and juxtaposed discourses. An example of this approach is Martin Stokes's study of Turkish *arabesk* music (1992). This style of music, popular with the marginalized urban masses, is more than the result of the stylistic development of Turkish folk music. It is part of a 'cultural field' in which different musical styles are best understood as 'different ways of talking about music' (p. 50).

Music as Discourse and the Nature of Musical Meaning

The third consequence, finally, of the 'interpretive turn' in ethnomusicology is the idea that musical meaning is no longer seen as being intrinsic to its sonic structure – a viewpoint still prevalent in musicology as late as the 1990s. Rather, musical meaning, like the meaning of a novel or a poem, materializes only in the act of experiencing music. Like a 'text,' a piece of music must be 'read.' And just like the meanings of a written text are not fixed in the text but are the result of discourses of which authorial voice and intention are one of many, the meanings of a musical performance are constantly changing and emergent, depending on the social structures and power relationships they help to define. The shift in musical scholarship from structural analysis to the interpretation of music as discourse occurred parallel to another significant development in anthropology and sociology in the 1970s: the emergence of theories and ethnographies of performance. On the basis of Erving Goffman's work and Dell Hymes's pioneering essays on the ethnography of speaking, anthropologists like Victor Turner, Richard Bauman, and James Fernandez began to examine the way in which performance contexts, events, and genres shaped communication and social action.

They were followed by numerous ethnomusicologists for whom the question of music's meanings and affective power became intimately tied to its real 'life': its relations with other forms of behavior and expressive culture, the interaction between performers and their audiences, as well as the unpredictability, and often improvisatory and playful nature, of music making. But whereas scholars such as Turner stressed the communal, ordering functions of performance, researchers are now more interested in the potential of musical performance for prompting change and resistance.

In line with these developments, scholars have also increasingly turned to the human body as the primary site for the construction of musical meaning. While earlier studies were based on a notion of the body as a biological given and, hence, stressed the physiological aspects of music making, recent studies posit the body as a cultural phenomenon, no less susceptible to and dependent on meanings than other spheres of human practice (Muller, 1999).

Music and Identity

The waning appeal of discourses that take stable notions of music and culture for granted have led scholars to reexamine the relationship between identity and music in a new, rather more skeptical light. Ironically, it is because of the fact that the growing isolation of the individual in modern society has generated new forms of collective identification – new nations, 'neo-tribes,' and new religions – that the time-honored interest in nationalism and ethnicity as the chief bases for music making has experienced something of a revival.

Of all the articulations of identity currently studied by music scholars, the modern nation-state and various types of nationalist discourse have probably commanded the most attention. Although studies of small-scale, mostly rural communities remain important locales for the ethnographic study of music, it is the contestations and entanglements linking the nation-state with multiply determined cultural practices that have become the object of the most sustained scrutiny undertaken to date. A rich literature on the new popular musics of the developing world – and, to a lesser degree, the advanced countries of the West – demonstrates how musical practices in the post-colonial world often disrupt the putatively homogeneous times and spaces of the nation-state, but at other times exist quite comfortably alongside nationalist agendas. In most cases, however, both positions blend into each other, often producing new exoticisms and articulations of racial or ethnic inclusion or exclusion (Guilbault, 2007; Sutton, 2002). Israeli rock influenced by Arabic music, for instance, celebrates an 'other' quite blatantly at odds with the images of backwardness that the Israeli majority projects of the Palestinian minority, yet at the same time it takes a stable 'Oriental' identity as its reference point (Regev, 1986). Conversely, some music have also been shown to lend themselves to more pliant forms of identity. This is perhaps more typical of the conflicted relationship between the nation-state, nationalist ideology, and cultural practices in diasporic communities. Thus, the music of black Americans is a product of the slaves' encounter with the West and, at the same time, has emerged in opposition to modernity (Gilroy, 1993). It may thus serve as a model for what Gilroy calls anti-antiessentialism, a critique of the oppressive history and racial injuries caused by the 'host' country that refuses to let go of memories of collective identity even where it resists grounding them in a direct genealogy rooted in the African past.

In some contexts, the nation-state may also come to the rescue of cultural diversity. Tanzania, for instance, is one of several countries that have tried successfully to promote 'national' musics by reserving some 85% of the music content on its national broadcasting service for local music products. But here, as in many other cases, the definition of what counts as representative of Tanzanian culture rested in the hands of the Swahili-speaking minority dominating the state apparatus (Wallis and Malm, 1992: 113–114).

Race and ethnicity are perhaps the second most studied forms of identity in ethnomusicology today. Deeply intertwined with nationalism, but also often standing in opposition to nationalist discourses, racial and ethnic identities are frequently seen to be most effectively asserted in the symbolic domain, such as in music. There are many reasons for this close nexus between ethnicity and music, but one of them is the fact that assertions of racial pride are voiced from marginalized social groups whose access to alternative cultural resources and power bases such as economic wealth is limited.

Processes of class formation, by contrast, and the musical construction of class consciousness have lost a great deal of their former appeal, primarily because the evidence gathered does not sustain the Marxist notion of a material base determining a suprastructure of ideology and culture. This lack of evidence was particularly pronounced in those areas where working-class cultures with pristinely defined contours failed to develop, as in much of the 'Third World,' or where they were about to disintegrate, as was the case in most of the advanced economies of post–World War II Europe and North America (Hebdige, 1979). By contrast, recent years have seen a growing interest in the formation of somewhat more fluid identities and

in the permeation of cultures considered to have once been 'premodern' with bourgeois values and practices. Under the impact of Antonio Gramsci's notion of 'hegemony' and Pierre Bourdieu's concept of 'habitus,' ethnomusicologists during the 1980s reassessed colonization and Westernization as projects that entail domination and acquiescence and that do not always occur along neatly separated lines of class or racial identity (Erlmann, 1999).

The notion, finally, that all identities are also gendered identities has greatly invigorated anthropological research on music and dance (Cowan, 1990). Although the idea has been familiar to ethnomusicologists for some time – much more so than to musicologists, whose subject until the 1980s seemed to be an unquestioned male one – recent scholarship has emphasized the negotiated character of sexual identity and the way in which music intervenes in the construction of such identities (Koskoff, 1987). As a recent study of the music of Maghrebian immigrants in France has shown, for example, *rai* performers and audiences construct complex identities around notions of sexuality, romantic love, and adolescence that are inspired by Western liberal ideals of gender equality (Gross et al., 1996). At the same time, these performers reproduce more traditional images of Arabness. While female singers often advocate more freedom for women, the jackets of their cassettes reproduce clichés of a more subdued kind of femininity and the ideal Arab woman.

Finally, an important focus that emerged from the current interest in identity and that undermines even further conventional notions of stable collective and individual identities – and the merely passive role of music in 'expressing' these – is the rather diverse set of practices circumscribed by terms such as diaspora, travel, migrancy, immigration, and displacement. Since the beginning of the 1990s, all these practices have been dealt with to some extent by ethnomusicologists (Turino, 1993), but musicologists and students of popular music have also increasingly committed themselves to exploring the interchange between music and the shifting boundaries of communities and audiences caused by large-scale movements of people. Furthermore, the spaces that people traverse as tourists, migrants, refugees, or touring musicians are all too often depicted as though they were bounded sites that are already filled with sound, and they are much less frequently depicted as multiply overlapping contact zones in which identities and musics are always blurred and emergent.

Music, Modernity, and Globalization

Most scholars now agree that music everywhere have been drawn into the maelstrom of modernization and, more often than not, have themselves become effective agents of modernization. While these scholars have as yet to arrive at a shared definition of modernization and modernity, it seems that the debate has centered on a series of intertwined issues and concepts. The first of these issues is globalization. Much of the discussion about globalization grew out of the concern with displacement and entailed a vigorous rethinking of the interrelationship between music and locality (Stokes, 1994). While previous scholarship tended to assume a relatively static fit between place, spatial identity, and music, an equation that often enabled scholars to produce richly detailed, empirically grounded narratives of local identity and music making (Finnegan, 1989), some researchers have gone one step further in examining locality as something that needs to be created in the first place. In a study of Nigerian *fújì* music and *oríkì* praise poetry, Karin Barber and Christopher Waterman (1995) show that while performers of these genres draw on a multiplicity of external sources to imagine densely local worlds, the emergent sense of locality is not simply stemming from any given body of materials or traditions. Rather, the "ultimate goal of any performance is to intensify the presence, image and prospects of local actors" (p. 243). Although it cannot be ruled out that Yoruba musical productions of place may be driven by the same agendas that underlie movements of national and ethnic purity, Barber and Waterman insist on the hybrid underpinnings of such assertions of local identity.

Moreover, the spaces in which scholars of popular music study and work are no longer necessarily such rather abstract entities as cities and countries. Increasingly, other spaces, such as recording studios, malls, music departments, and taxis, are becoming sites of inquiry.

Another aspect of modernization that dominates current debates is the power of capitalism and commodification to transform 'traditional' cultures, and the capacity of the Western mass media to shape the collective imagination of large populations. Some critics such as French sociologist Jean Baudrillard have tended to think about global cultural flows in rather uniform terms, positing a regime of simulations and simulacra as an all-embracing system. In line with this reasoning, some music scholars have taken a rather sinister view of global musical production at the end of the twentieth century. Steven Feld draws on Gregory Bateson's term 'schismogenesis' to think through a set of escalating relationships marking the production and consumption of music on a global scale. Schismogenesis is a process through which the interplay between essentially dissimilar but mutually appropriative actions may lead to a closer symbiotic interdependence of both sides, to the point that they may even become incapable of self-correction and get caught up in closed circuits of repetition (Feld, 1994). Ultimately, what results from this scenario is a gray out of a new type. As discourses of authenticity and difference assert themselves and activities of commercial appropriation in turn get more overt and outrageous, some kind of 'fusion of the parties for mutual business gain' becomes likely (p. 273).

Others scholars, by contrast, appear to equate 'Third World' cultural practices with antihegemonic agency per se, stressing the potential of music to maintain the heterogeneity of the world's cultures in the face of Western media dominance (Lipsitz, 1994). The majority of scholars, however, have taken a more nuanced position toward the role of culture in reflecting and mediating global relations of power, and numerous theoretical interventions have emerged in recent years that seek to provide new conceptual tools and models with which to grasp the twists and turns of global musical flows (Manuel, 1993). Mark Slobin, for instance, in an attempt to account for the impossibility of representing both global generality and local specificity, has depicted the interplay of various 'micromusics' in the West as a kind of 'interculture,' a constant oscillation between the pressures of the 'supraculture' and the contention of various subcultures (Slobin, 1993).

Finally, a handful of studies have attempted to view the making of globally connected identities as a two-way process in which center and periphery are constantly confused, and in which the West is just as much an 'other' of Africa or Asia as Africans and Asians have been the 'others' of Europeans (Erlmann, 1999). This includes greater attention to the agency of those (predominantly Western) 'global' figures, such as Paul Simon, Peter Gabriel, Quincy Jones, and Mickey Hart, whose practices and discourses are more than just manifestations of some anonymous capitalist world system or the global music industry.

Music, Media, and Technology

If anthropology's time-honored notions of identity and belonging and their intimate association with meaningful and largely homogeneous cultural practices by the 1980s were increasingly being viewed with skepticism, so too did ideas of unmediated individual or collective creativity rooted in genius or tradition come under fire. The emergence of digital technology, the decline of state-controlled media as the sole authority on what constitutes the public good, the legitimacy of power, or the value and meaning of cultural practices, followed by the rapid expansion during the 1990s of deregulated and new media such as the Internet, promised new levels of access, participation, and diversity, in turn leading to a reevaluation of the conditions under which people interact with music. While such broadened arenas of interaction – often all too hastily equated with Western-style liberal democracy per se – may serve to strengthen existing identities and communal boundaries threatened by global flows of capital, frequently they also create new affiliations and new distinctions based on choice and taste rather than state-mandated ideologies and allegiances. Much of this turn toward technology is owed to the pioneering efforts of students of popular music to understand popular music through the prism of its media and technology a priori rather than the structural or historical analysis of works. At the same time, much scholarship within the field of popular music studies prior to the 1990s has been prone to overtheorization, often at the expense of more sustained empirical engagement with embodied uses of technology within specific locations and among historically determined actors. However, by the early 2000s, many scholars working in popular music studies found common ground with ethnomusicologists and anthropologists (and, to a certain degree, music theorists) as these turned to the specifics of music technology as part of socially mediated and locally embedded cultural practices (Greene and Porcello, 2004; Lysloff and Gay, 2003). The result of this convergence of disciplines is a growing body of literature on such diverse topics as the musical impact of the rise of the audiocassette in India (Manuel, 1993), the recording studio as a site of different forms of social interaction in South Africa or Indonesia (Meintjes, 2003; Wallach, 2008), the home studio in the West (Theberge, 1997), the culture of sampling in hip hop (Schloss, 2004), the politics of mixing and sound engineering in Brazil and Jamaica (Moehn, 2012; Veal, 2007), and the history of the MP3 (Sterne, 2012). At the same time, other media that have long been staples of research in social history, media studies, and cultural studies such as radio and television have yet to become the object of sustained attention among ethnomusicologists.

While many of these technologies might be seen as perpetuating the hegemony of Western standards of aesthetic and economic value, in many cases they also support alternative forms of musical creativity, income-generating strategies, audience involvement, and social distinction. At the same time, the rise of MP3 as the standard format for sound reproduction and dissemination fundamentally affected music's role in mediating social structure and practice. While the effortless and instantaneous dissemination of sound files across cultural, social, and national divides opens up a space for envisaging new communities, this creative potential is also severely hampered by new legal frameworks that reduce what had previously been owners to mere 'users.' Increasingly, disputes over aesthetic value, cultural memory, and public access to culture are subsumed under legal battles over intellectual property rights. The ever more forceful intrusion of intellectual property law into every aspect of music making and consumption is the subject of intense debate within media studies and among activists in the West (Burkart, 2010), with anthropologists and ethnomusicologists beginning to consider the impact of international intellectual property regimes on local communities on the fringes of the global economy, attending to issues such as piracy, cultural rights, and heritage (Weintraub and Yung, 2009; Erlmann, forthcoming).

Music and Knowledge

The anthropological study of music, much like other projects within the humanities and social sciences, since the early years of the twenty-first century has come under growing pressure to demonstrate its 'impact' in a global order of knowledge dominated by concepts of 'efficiency' and economic viability. In this climate, music scholars are increasingly looking to the sciences for inspiration and legitimacy. Buoyed by the popular appeal of the neurosciences and narratives about the inherent compatibility of music and mathematics (the 'Mozart effect'), scholars from the 1990s were particularly drawn to the history of science. Having regained widespread acceptance from the 1980s, the notion that discovery and knowledge are culturally contingent and that the arts and the sciences are closely intertwined proved attractive to a field steeped in a variety of epistemological traditions from Buddhism to German idealism and with deep affinities to mathematics and the study of nature at large. Thus, several historians of science have explored the history of acoustics not as the foundation of but as being embedded in musical practice (Gouk, 1999; Jackson, 2008; Erlmann, 2010).

Yet the notion that the relationship of music to things like cognition, consciousness, and affect, instead of being the exclusive domain of psychoacoustics or music psychology, might be culturally determined and that, hence, this relationship would be open to anthropological or ethnomusicological inquiry only took root toward the late 1990s. The work of anthropologist and ethnomusicologist Steven Feld provided a significant stimulus to this line of thinking (Eisenberg, 2010). Departing from and critically responding to Canadian composer Murray Schafer's work in acoustic ecology (Schafer,

1977) and reflecting on his own fieldwork in the rich acoustic environment of the Kaluli living in the rainforest of Papua New Guinea, Feld (1982) in a series of writings and audio-collages (the latter being called, after Schafer, 'soundscapes') intervened in anthropology's and ethnomusicology's ongoing crisis of representation not so much by providing readers and listeners with a sense of what Kaluli music *is* but, rather, by engaging in a written and sounded dialogue with Kalulias they musically and sensorially construct the world around them (Feld, 1991, 1996).

Music as Sonic Practice

One of the most promising developments of the 2000s has been the emergence of sound studies, a field of inquiry that draws much of its rationale from a growing sense of discomfort with 'ocularcentric' constructions of modernity (Jay, 1994) and the hegemony of the visual in a wide range of academic disciplines, but particularly within cultural studies. Concerned with a broad spectrum of sonic phenomena from ring tones to urban noise to the history of sound compression, sound studies has been influential in decentering music from its privileged position at the heart of (Western) romantic and modernist aesthetics. Music instead becomes one among many sonic practices. Although the roots of this 'auditory turn' can be traced to a variety of post–World War I philosophical traditions such as the French Annales and the Histoire des Mentalités schools of thought, other developments from the 1990s also contributed to this trend, such as a shift from the conventional radio feature or *Hörspiel* to experimental sonic practices such as 'soundscapes' and sound art (Samuels et al., 2010), the emergence of film sound studies (Altman, 1992; Chion, 1994), and the turn within anthropology toward the senses (Stoller, 1989; Classen, 1993). Among the latter, the 'denigration of vision' (Jay, 1994) was particularly pronounced, growing out of a sense of disaffection with the interpretive turn led by Clifford Geertz and its emphasis on cultural practice as a form of intertextuality and, hence, on anthropology as 'reading' (Howes, 2003: 17–22).

Anthropologists' rekindled interest in sound and hearing for the most part excluded music, with several influential studies having appeared on sound in Islamic devotional practice (Hirschkind, 2009) and the soundscape of northern Nigeria (Larkin, 2008). Ethnomusicologists, for their part, soon followed the lead of historians and anthropologists in studying listening practices as sites of identity formation, the materiality of acoustic communication, and the politics of noise in urban settings (Erlmann, 2004). Between them, these new directions play a significant role in drawing attention to contemporary music's changing place on a continuum of sound, noise, and multisensory practice, thus widening the scope of the anthropological study of music in a rapidly changing world.

See also: Body: Anthropological Aspects; Dance, Anthropology of; Folklore; Globalization and World Culture; Music, History of; Performance: Anthropological Aspects; Popular Culture; Senses, Anthropology of The.

Bibliography

Altman, R., 1992. Sound Theory, Sound Practice. Routledge, New York.
Barber, K., Waterman, C., 1995. Traversing the global and the local: Fuji music and praise poetry in the production of contemporary Yoruba popular culture. In: Miller, D. (Ed.), Worlds Apart. Modernity through the Prism of the Local. Routledge, London.
Blacking, J., 1973. How Musical Is Man? University of Washington Press, Seattle, WA.
Burkart, P., 2010. Music and Cyberliberties. Wesleyan University Press, Middletown, CT.
Chion, M., 1994. Audio-Vision. Columbia University Press, New York.
Classen, C., 1993. Worlds of Sense: Exploring the Senses in History and across Cultures. Routledge, New York.
Cowan, J.K., 1990. Dance and the Body Politic in Northern Greece. Princeton University Press, Princeton, NJ.
Eisenberg, A., 2010. The Resonance of Place: Vocalizing Swahili Ethnicity in Mombasa, Kenya. PhD thesis, Columbia University.
Erlmann, V., 1999. Music, Modernity and the Global Imagination. South Africa and the West. Oxford University Press, New York.
Erlmann, V. (Ed.), 2004. Hearing Cultures. Essays on Sound, Listening, and Modernity. Berg, New York.
Erlmann, V., 2010. Reason and Resonance: A History of Modern Aurality. Zone Books, New York.
Erlmann, V., forthcoming. Legal Sounds: Intellectual Property Law and Music in South Africa. Duke University Press, Durham, NC.
Feld, S., 1982. Sound and Sentiment. Birds, Weeping, Poetics and Song in Kaluli Expression. University of Pennsylvania Press, Philadelphia, PA.
Feld, S., 1991. Voices of the Rainforest: A Day in the Life of the Kaluli People. Rykodisc, Salem, MA.
Feld, S., 1994. From schizophonia to schismogenesis: on the discourses and commodification practices of 'world music' and 'world beat'. In: Keil, C., Feld, S. (Eds.), Music Grooves. University of Chicago Press, Chicago, IL.
Feld, S., 1996. Waterfalls of song: an acoustemology of place resounding in Bosavi, Papua New Guinea. In: Feld, S., Basso, K.H. (Eds.), Senses of Place. School of American Research Press, Santa Fe, NM.
Finnegan, R., 1989. The Hidden Musicians: Music-making in an English Town. Cambridge University Press, Cambridge, UK.
Gilroy, P., 1993. The Black Atlantic. Modernity and Double Consciousness. Harvard University Press, Cambridge, MA.
Gouk, P., 1999. Music, Science, and Natural Magic in Seventeenth-century England. Henderson, London.
Greene, P., Porcello, T., 2004. Wired for Sound: Engineering and Technologies in Sonic Cultures. Wesleyan University Press, Middletown, CT.
Gross, J., McMurray, D., Swedenburg, T., 1996. Arab noise and Ramadan nights: Rai, Rap, and Franco-Maghrebi identities. In: Lavie, S., Swedenburg, T. (Eds.), Displacement, Diaspora, and Geographies of Identity. Duke University Press, Durham, NC.
Guilbault, J., 2007. Governing Sound: The Cultural Politics of Trinidad's Carnival Musics. University of Chicago Press, Chicago, IL.
Hebdige, Dick, 1979. Subculture: The Meaning of Style. Routledge, New York.
Herndon, M., McLeod, N., 1981. Music as Culture. Norwood, Darby, PA.
Hirschkind, C., 2009. The Ethical Soundscape: Cassette Sermons and Islamic Counterpublics. Columbia University Press, New York.
Howes, D., 2003. Sensual Relations: Engaging the Senses in Culture and Social Theory. University of Michigan Press, Ann Arbor, MI.
Jackson, M., 2008. Harmonious Triads: Physicists, Musicians, and Instrument Makers in Nineteenth-century Germany. MIT Press, Cambridge, MA.
Jay, M., 1994. Downcast Eyes: The Denigration of Vision in Twentieth-century French Thought. University of California Press, Berkeley, CA.
Keil, C., 1979. Tiv Song. University of Chicago Press, Chicago, IL.
Koskoff, E., 1987. Women and Music in Cross-cultural Perspective. Greenwood Press, New York.
Larkin, B., 2008. Signal and Noise: Media, Infrastructure, and Urban Culture in Nigeria. Duke University Press, Durham, NC.
Lipsitz, G., 1994. Dangerous Crossroads: Popular Music, Postmodernism and the Poetics of Place. Verso, London.
Lysloff, R., Gay, L.C., 2003. Music and Technoculture. Wesleyan University Press, Middletown, CT.
Manuel, P., 1993. Cassette Culture. Popular Music and Technology in North India. University of Chicago Press, Chicago, IL.
Meintjes, L., 2003. Sound of Africa! Making Music Zulu in a South African Studio. Duke University Press, Durham, NC.
Merriam, A.P., 1964. The Anthropology of Music. Northwestern University Press, Evanston, IL.
Moehn, F., 2012. Contemporary Carioca: Technologies of Mixing in a Brazilian Music Scene. Duke University Press, Durham, NC.

Muller, C., 1999. Rituals of Fertility and the Sacrifice of Desire: Nazarite Women's Performance in South Africa. University of Chicago Press, Chicago, IL.
Regev, M., 1986. The musical soundscape as a contested area: 'Oriental music' and Israeli popular music. Media, Culture and Society 8, 343–356.
Said, E., 1979. Orientalism. Vintage Books, New York.
Samuels, D., Meintjes, L., Ochoa, A., Porcello, T., 2010. Soundscapes: toward a sounded anthropology. Annual Review of Anthropology 39, 329–345.
Schafer, M., 1977. The Soundscape: Our Sonic Environment and the Tuning of the World. Destiny Books, Rochester, VT.
Schloss, J.G., 2004. Making Beats: The Art of Sample-based Hip-hop. Wesleyan University Press, Middletown, CT.
Slobin, M., 1993. Subcultural Sounds. Micromusics of the West. Wesleyan University Press, Hanover.
Sterne, J., 2012. MP3: The Meaning of a Format. Wesleyan University Press, Middletown, CT.
Stokes, M., 1992. The Arabesk Debate: Music and Musicians in Modern Turkey. Clarendon Press, Oxford, UK.
Stokes, M., 1994. Ethnicity, Identity and Music: The Musical Construction of Place. Berg, Oxford, UK.
Stoller, P., 1989. The Taste of Ethnographic Things: The Senses in Anthropology. University of Pennsylvania Press, Philadelphia, PA.
Sutton, R.A., 2002. Calling Back the Spirit: Music, Dance, and Cultural Politics in Lowland South Sulawesi. Oxford University Press, Oxford.
Theberge, P., 1997. Any Sound You Can Imagine: Making Music/Consuming Technology. Wesleyan University Press, Middletown, CT.
Turino, T., 1993. Moving away from Silence: Music of the Peruvian Altiplano and the Experience of Urban Migration. University of Chicago Press, Chicago, IL.
Veal, M., 2007. Dub: Soundscapes and Shattered Songs in Jamaican Reggae. Wesleyan University Press, Middletown, CT.
Wallach, J., 2008. Modern Noise, Fluid Genres: Popular Music in Indonesia, 1997–2001. University of Wisconsin Press, Madison, WI.
Wallis, R., Malm, K., 1992. Media Policy and Music Activity. Routledge, London.
Weintraub, A.N., Yung, B., 2009. Music and Cultural Rights. University of Illinois Press, Chicago, IL.

Music, History of

Albrecht Riethmüller, Free University of Berlin, Berlin, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, Volume 15, pp. 10259–10263, © 2001, Elsevier Ltd.

Abstract

History of music, or the historiography of music (music history) is a synonym for that segment of musicology referred to in recent decades as 'historical musicology.' Methodological approaches peculiar to the history of music arise from the transitory nature of music as a performed art form, musical notation (which was never obligatory for those belonging to the elitist social group), and the acoustical documentation of voices and sounds since the invention of the phonograph in 1877. The representation of history within the discipline, music's beginnings, the nature of historical musicology and its methods, the concept of the 'relative autonomy' of music history within the scope of art history, and recent developments are examined, in addition to the tension between local and global historiography, the focus on Western art music ('eurocentrism'), and finally political or nationalistic implications.

'History of music' signifies the sum of all musical or music-related events in as much as they have been documented or handed down through time. In particular, it refers to the historiography of music (music history) and is a synonym for the discipline of musicology. In recent decades it has come to designate that part of musicology referred to as 'historical musicology.'

History and Representation

The panorama of cultures, with their divergent concepts of music and history, are too multitudinous to be adequately and equitably considered here. A review of the world's oral and written musical traditions, whether they be Indian, American-Indian, Chinese, Australian-Aboriginal, Japanese, Ewe, Near-Eastern, or Central European, would only culminate in the emptiest of generalizations to compare their individual courses. For reasons of expediency, therefore, the following remarks refer to the history of music as perceived from the perspective of Western culture and are restricted to its tradition.

Three characteristics of music determine and modify questions concerning historical data, documents, and sources, modulating thereby the various approaches to the history of music. (1) By necessity, the transitory nature of music as a performed art form does not engender an immutable artifact for permanent view, and, in contrast to the visual arts, is dependent on reproduction and representation. This contributes to the tendency to perform and listen to music as though it would be permanently present. In fact, performing or listening to music from an earlier period decomposes the historical distance between the period in question and the present time. This impulse to dispossess music of its history may in some instances lead to the radical inclination to regard music as a natural phenomenon rather than as a historical one. (2) Musical notation was an early invention. Two different kinds were already established in Greek antiquity, most likely for educational purposes and not for the preservation of music for future generations or for performances. This also holds true for the earliest stages of polyphony in the Middle Ages, which went hand in hand with the beginning of notational practice and ultimately led to the types of scores now familiar to us. Before printing was invented, it can scarcely be claimed that musical notation produced works as texts comparable to written texts, such as the *The Odyssey* or *The Iliad*. And while illiteracy and education are by nature contradictory, musical illiteracy has remained a socially tolerated phenomenon; someone who is unable to read music, will not be ostracized from an elitist social group. In popular music circles musicians have never been required to possess the ability to decipher musical notation. (3) With Edison's invention of the phonograph, voices and sounds could be acoustically documented for the first time. Due to the omnipresence of music in today's popular culture and the recording industry, one can easily forget that this 'kopernikanische Wendung' in music history took place just one century ago. In 1886 Friedrich Nietzsche asserted that music is the latecomer of every culture (*Menschliches, Allzumenschliches II*, Sect. 171). Although this corresponds to the romantic conviction that music is the youngest of the fine arts it seems to contradict the Western tradition that ranked music theory among the earliest individual academic disciplines, together with arithmetic, geometry, and astronomy. This was the case in the Greek *enkyklios paideia* and later in the *quadrivium* of the medieval *septem artes liberales*. To the degree that music is more often associated with emotional and spiritual qualities than with intellectual or physical ones, the reconstruction of the history of sound is no less precarious than the history of the emotions themselves. Both fields are predominantly dependent on a systematic or experimental approach rather than on a historical one. And in contrast to archeological sites or the visual arts, the historical foundations of music are ultimately of less interest within the context of the essentially personal need for music.

Considering Music History and the Beginning of Music

History has always played a certain role in speculations about the theoretical basis of music. The historical account of music in (pseudo-)Plutarchus's *De musica* from ca. AD 100 weaves together mythological and historical names (that cannot be clarified in each instance) before turning to the matter of 'old

music,' regarded as holy and good, and its succession by 'new music,' seen as degenerate and decadent. Aristophanes and other contemporaries maintained that this transition took place in the late fifth century BC. Twentieth-century historians still adhered to this model. Greek myths and medieval legends based on biblical narrative that once explained and interpreted the origins of music have today been replaced by 'scientific' theories that biologically and genetically construe the earliest stages of the development of music, whether beyond written traditions or in pal eontological respects. One may ask whether modern speculations as to music's prehistoric inception are more reliable than Ovid's report of Pan's invention of the syrinx (*Metamorphoses* I, 689 ss.) that signaled the beginning of another *ars nova*, or more insightful than the lemma handed down since late antiquity that states Pythagoras as the inventor musicae, claiming that he was to have found an explanation for the phenomenon of consonances and therefore to have 'invented' music theory.

By 1700, book titles espousing the history of music appeared throughout Europe (e.g., W. C. Printz, *Historische Beschreibung der Edelen Sing-und Kling Kunst*, Dresden, 1690; G. A. Bontempi, Historia musica, Perugia, 1695; P. Bourdelot and J. Bonnet, *Histoire de la musique*, Paris, 1715). In all cases, the point at which the history of music is seen to begin is dependent upon the historiographical premise that has been taken. Some scholars proceed from early organized polyphony around AD 900; others go back to monophony (Gregorian chant). The belief that ancient Greek music need not necessarily belong to music history's canon is in large part due to the decrease in the knowledge of the Greek language in the twentieth century, with the result that it is subsequently relegated to the fields of ethnomusicology or anthropology (cf West 1992, p. 3). Thus, what was still regarded to be the foundation of European culture and humanistic tradition in the nineteenth century has been cut off from the history of Europe.

The Discipline and its Methods

The rise of historicism in the nineteenth century brought music history (together with art history) as an academic field into the university, a development that began primarily in German-speaking countries. The history of music quickly became an essential component of both the theoretical and practical study of music. Together with specialized research institutes, the academic discipline increased in German universities after World War I, while the wave of emigration following 1933 played an exceptional role in its advancement in the USA. Since World War II the discipline has grown inside and outside Europe, and today is established in universities worldwide. Aside from biographical studies and philological efforts (such as editing scores and theoretical texts on sources), music history's chief methodological approach has been to examine the history of style, or more recently, genre. This is narrowed down further to the history of composition or individual works. Special attention has been given to the history of reception throughout the last decades. Even with respect to essentially synesthetic productions, such as opera or stagings within the multimedia landscape, the score remained the music historian's most important source at hand. Up to the present time, the predominant individuals of interest for historians have been composers, regardless of how prominent a performer, whether singer, conductor, or instrumentalist, may be. This is also the case within the history of performance practice (impulses coming from the UK), which attempts to reconstruct performance style within any given age independently of the work of individual interpreters. Compared to music history's concentration on composers and compositions, adjacent fields such as the history of music theory and of music aesthetics, and the archeology and iconography of music have assumed less prominent positions. This holds true for the relationship between the history of music and history as well. For example, while writings on the history of sacred Christian music are often aligned with related areas such as the history of liturgy, churches, and religion, contextualizations within social and political history remain rare. Supported by the romantic notion of *Kunstreligion* (the substitution of religion by the arts), the ideology that positions music in an apolitical, internalized, subjective domain strictly separated from social and political conditions has not yet been fully overcome.

Boundaries

The history of music has pertained and, in large part, still pertains to Western music, if not exclusively. The periodization of earlier epochs (the Middle Ages, the Renaissance, and Baroque) has remained comparatively stable; thereafter, less so. The history of music is patterned after conceptualizations similar to those used in general historiography (whether in the framework of *Geistesgeschichte*, social, or cultural history); the philosophy of history; and the fundamental concept of permanent progress (innovation and its aesthetic equivalent, originality) and cyclic forms (such as ascent, climax, and decline). In spite of how commonplace the expression 'occidental music' (*abendlän dische Musik*) is, a history of European music in which local, regional, and national traditions are treated as a unity within diversity – in accordance with the standards of the modern European Union – has yet to be written. Historians who specialize in the musical culture of their own regions often lay claim to the concept of occidental music, and yet upon closer scrutiny, the traditions of other countries are comfortably overlooked. In the subtext one can ascertain that the inclusion of music from the Byzantine empire (with consequences for Eastern European music) and from the Americas within the occidental circle is questioned. The post-Wagnerian composer Hans Pfitzner (1926, p. 193) compared music's historical course to the upbringing of a boy whose first nurse was Dutch, who then lived for a season in Italy before making his home in Germany around 1750. Music historiography, dominated by the ascendance of nationalism in the nineteenth century, has been influenced by national historiography. Some music historians' nationalistic stance relies on a conception of music history in which consecutive composers are threaded onto a string of pearls, neither allowing for diversity or pluralistic considerations.

'Relative' Autonomy

Music historians who view the history of music within the scope of art history tend to view favorably the idea of

a '"relative autonomy" of music history' (cf Dahlhaus 1983, Chap. 8). This concept focuses on questions of style, genre, and composition without giving credence to their contextual arena–an idealistic interpretation of art music that boasts of a pragmatic concentration on musical components at the cost of music's isolation from society, economy, and politics. The idea of autonomy maintains that (1) art is independent of life and (2) music history is made up of a sequence of masterpieces written by heroic composers of genius, whose musical innovations and originality are synonymous with progress. This progress is often associated with decline in differing degrees up to the present time: while contemplating the grand summits of art music, no thought should be squandered on the valleys. As can be expected, Marxist writers have explicitly opposed this view (cf Knepler, 1977, p. 462 ss.). Art and social histories were often strictly and programmatically opposed to each other; particularly in the tradition of Hegel, dominant in German musicology, this led to a disengagement between music history and history in general. The historiography of Western music has remained the historiography of an elitist art form. In spite of its universal popularity, Beethoven's music has never belonged to the realm of popular culture (one notable exception being the *Ode to Joy*, used as the anthem for the European Union). Popular music–written for the day, not for eternity–has socially, politically, and historically been the most effective type of music throughout the ages. Music historians seldom participate in the methodological discourses of historians. Likewise, the historian's rare stand on musical matters, owing to a lack of technical skills or detailed knowledge, is similar to the philosopher who eschews discussions on music in their aesthetic writings. None of the famous historians, therefore, focused on music or, particularly, music history as a methodological paradigm, and in spite of the general increase of interdisciplinary research, historical musicology and history rarely cooperate.

Recent Developments

Music history, perceived either as the history of Western music in general or its various national entities, came to be questioned in the second half of the twentieth century. In the USA a discussion about whether its music history should solely be defined from the European perspective has emerged. A front-page article of the *New York Times* entitled 'Ibn Batuta and sitar challenging Columbus and piano in schools' on 12 April 1989 asked whether university professorships for piano should be replaced by those for sitar. Jazz and other forms of popular culture excluded from traditional historiography have been seen by some to pose a threat to the canon of classical music heard in Carnegie Hall and at the Metropolitan Opera. 'Eurocentrism' in music has come under critical scrutiny in recent years. Ethnomusicologists increasingly bemoan the dominance, even hegemony, of Western music as it spreads globally through radio and recordings. This new wave of cultural colonialism is seen as a potential annihilator of the last traces of 'original' music cultures. In extreme cases, ethnomusicologists deny history in music altogether if the related culture in question does not reflect the concept of history, or where a lack of observable change has led to an 'almost absolute stability of the musical cultures of non-Western societies,' a view modified by major scholars (Nettl, 1983, p. 172). Although 'world music' is the slogan of the day, the idea of a universal history of music is seen to be obsolete or, due to the plethora of materials, unmanageable. The relevant UNESCO project begun in 1979, *Music in the Life of Man*, retitled in 1989 *The Universe of Music. A History*, has produced programmatic statements rather than results. Conversely, the controversial search for 'uni versals' in music is generally conducted in a non-historical fashion. 'New musicology,' which began in the USA in the early 1980s, directs its attention to gender issues and the music of minority groups, and emphasizes a contemporary approach to criticism in which the role of history has yet to be fully defined. The phrase 'new historicism' has entered into the discussion of late (cf Stanley, 2001).

Identity Exchanges and Modern Mythology

The current global fashion and passionate search for cultural and national identities presents some dubious consequences for music history. If one attempts to decipher what is specifically 'Italian' or 'American' in a piece of music, the answer will remain arbitrary by definition, since sounds cannot signify unequivocal and indubitable meanings. An example of one such identity crisis can be found in Austria. Viennese classical music, denoting the music of Haydn, Mozart, and Beethoven, has long been considered to be a paradigm of, if not synonymous with, German music. In 1947 Albert Einstein stated that the 'concept of German music' itself has come to rest on the works of these three masters (cf Einstein, 1992, p. 261). Austrian historiographers of music have now come to emphasize their own musical identity as being distinctly different in substance from a German one. This recent dichotomy is all the more curious since Austrians have frequently been eager to discern the German identity in their Austrian composers. Now Beethoven regains the status of an alien immigrant at best (cf Flotzinger, 1997, col. 1201). Hitler's former Reich's Music Chamber President, Richard Strauss, revealed yet another astonishing conviction shortly after World War II. On May 20, 1946 he admitted to his devoted friend and historiographer, Willi Schuh, that in all likelihood cultural historians, German patriots, and the Jewish press would be less than pleased with his tenet that 'political Germany *had* to be destroyed after it had fulfilled its world mission: the creation and perfection of German music' (Strauss 1969, p. 89). Of course, the composer regarded his own music as this ultimate fulfillment, as the end of music history. Uniting the idea of the end of art and the end of politics, his vision placed all future music in a state of a *post histoire*.

Nowadays, music composed just a few years after Strauss' remarkable statement, whether serial, electronic, or chance, is considered classic. The histories of all musics have been subject to various readings throughout the course of time, often stemming from a national, ethnic, or otherwise political intention. The genesis of one of the most significant twentieth-century forms, jazz, has been alternately depicted as originating in Africa, Europe, and America. And it is in the New World that the search for a national musical identity during the first half of the twentieth century failed to produce a national school of composition. 'Each composer found his own way,' and 'perhaps this is the way it should be in America' (Hamm, 1983,

p. 459). Perhaps this maxim could on occasion be applied to the realm of music history as well.

See also: Art, Anthropological Aspects of; Dance, Anthropology of; Music: Anthropological Aspects; Social System.

Bibliography

Allen, W.D., 1962. Philosophies of Music History. A Study of General Histories of Music 1600–1960. Dover Publications, New York.

Dahlhaus, C., 1983. Foundations of Music History. Cambridge University Press, Cambridge, UK.

Eggebrecht, H.H., 1980. Historiography. In: Sadie, S. (Ed.), The New Grove Dictionary of Music and Musicians, vol. 8. Macmillan, London, pp. 592–600.

Eggebrecht, H.H., 1991. Musik im Abendland. Piper, Munich, Germany.

Einstein, A., 1947. Music in the Romantic Era. Norton, New York [original reprinted in 1992 as Die Romantik in der Musik, Metzler, Stuttgart, Germany].

Flotzinger, R., 1997. Österreich. In: Blume, F. (Ed.), Die Musik in Geschichte und Gegenwart, second ed, vol. 7. Metzler, Stuttgart, Germany, pp. 1176–1223. Col.

Hamm, C., 1983. Music in the New World. Norton, New York.

Knepler, G., 1977. Geschichte als Weg zum Musikverständnis. Verlag Philipp Reclam, Leipzig, Germany.

Nietzsche, F., 1878–1879. 1886. Menschliches, second ed. vol. 11. Allzumenschliches. Chemnitz. 546–561.

Nettl, B., 1983. The Study of Ethnomusicology. University of Illinois Press, Urbana, IL.

Pfitzner, H., 1926. Gesammelte Schriften, vol. 1. Benno Filser-Verlag Augsburg, Germany.

Stanley, G., 2001. Historiography. In: Sadie, S. (Ed.), The New Grove Dictionary of Music and Musicians, second ed. Macmillan, London.

Strauss, R., 1969. Briefwechsel Mit Willi Schuh. Atlantis, Zurich.

Treitler, L., 1989. Music and the Historical Imagination. Harvard University Press, Cambridge, MA.

West, M.L., 1992. Ancient Greek Music. Clarendon Press, Oxford.

Myrdal, Alva (1902–86)

Eva S Lyon, Faculty of Arts and Humanities, London South Bank University, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Alva Myrdal was a Swedish social scientist, politician, and feminist who made an intellectual contribution in several fields: child-friendly education, welfare state theory, women in the labor market, international development, and disarmament. A prolific writer and researcher, most of her life was spent working as a representative of the Swedish Social Democratic government in national and international contexts, including a period (1949–55) as Director of Departments at the UN (Social Affairs) and UNESCO (Social Science). Her most famous works are 'Nation and Family' (1941), 'Women's Two Roles' (with Viola Klein, 1956), and 'The Game of Disarmament' (1976). She received the Nobel Peace Prize in 1982.

Alva Myrdal: A Passionate Engagement for Social Justice

Alva Myrdal has been variously described as feminist, social reformer, welfare theorist, diplomat, sociologist, politician, human rights activist, writer, ambassador, cabinet minister, and Nobel Peace Prize winner – indicating the breadth of her extensive contribution to public life. This broad contribution was largely framed by her active, lifelong membership of the Swedish Social Democratic Party during its long period of major welfare and educational reforms; by her leading international positions in various departments of the United Nations and international women's organizations; and by her many contributions, as an international diplomat, to world peace and disarmament. She was also a prolific writer and researcher who wrote extensively in both popular and academic contexts, on social and political issues ranging from childcare and working mothers to nuclear disarmament and superpower politics. All these varied activities in support of a more just and conflict-free social world were underpinned by her passionate faith in the possibility of a scientific, rational understanding of the relationship between the different domains of social and economic life (public and private) and their mutual interdependence. If we could only scientifically better understand what these relationships are and collaboratively resolve their inherent complexity, she argued, life could be free from cruelty and fear and be more openly, fairly, and humanely organized and managed.

A Feminist for Children

Alva Reimer was born in 1902 into a family of modest means. She soon came to share her father's interest in the work of the Social Democratic and Cooperative movements for greater justice for farmers and workers. Her early dreams of independence fostered a commitment to women's emancipation in all its aspects. She was already, as a child, a ferocious reader and writer, who found it a handicap that she was not a boy. Poor access to academic secondary school for girls and with little parental support for further study, she left school early with a qualification in commercial and secretarial skills that enabled her to support herself whilst completing her matriculation at a private tutorial college. She was only 17 years old when she met her future husband, the economist Gunnar Myrdal (1898–1987), with whom she shared a lifelong, intellectual partnership and had three children. At university she studied literature, languages, psychology, pedagogy, and sociology – some of these on grants alongside her husband, as Visiting Scholar in Germany, the United States, Britain, and Switzerland. For the first decades of her married life she also acted as hostess and intellectual sparring partner to her husband's friends in the radical intellectual circle surrounding the economist Gustav Cassel and the Stockholm School of Economics, whose younger members advocated a stronger role for the Government in economic relations. She developed a growing number of her own engagements, including those in the politically important Swedish Social Democratic Women's Federation, an organization, which throughout her life remained for her, an important intellectual and political platform.

In 1929 both Gunnar and Alva Myrdal became recipients of Rockefeller Fellowships – in her case to study social psychology, child development, and social work. Before leaving for the United States they spent a few months studying in London learning fresh (but in their minds, too class-ridden) perspectives on economics and reform. In the United States they were warmly welcomed and introduced to the best minds in their fields. Alva Myrdal drew inspiration from developments in family sociology and in the project-oriented urban sociology of the Chicago School. She was particularly influenced by the sociologists Robert and Helen Lynd, W.F. Ogburn, and W.I. and Dorothy Thomas, the latter becoming long-term friends. All were theoretically and methodologically concerned with the relationship between individual cultural perceptions and industrial and institutional change. The changing status and needs of children in democratic modernity was a topic of particular concern in the United States at the time, as were developments in scientific means of observing them, developed by the Viennese pedagogues Charlotte Buhler and Dorothy Thomas (both at the time at Columbia University Teachers College). There were also many opportunities during the year for networking with the politically influential professional women's organizations of the time. But the year of the visit was also the year of the Wall Street stock market crash and they encountered mass unemployment and large-scale

poverty in a political context of surprising passivity and inaction, despite a 'hands on' culture of political optimism. She returned to Sweden both enthused, and firmer in her conviction that something must be done to prevent such misery and that she could not, like some of the American academics she met, remain politically inactive.

One of her first actions on returning from America was to establish study circles for parents, based on courses she had observed in the United States, to help them become better informed about modern, child-friendly, child rearing practices. When her proposal for publically financed nursery schools was initially rejected by the Social Democrats, she gained the support of housing cooperatives for the creation of her own preschool teacher training institute, the Stockholm Pedagogical Institute for Preschool Teachers, to which she brought the latest international developments in child psychology and pedagogy. Another major engagement at this time became that of housing and the question of how it could be improved to support families, especially working women and children. Stockholm had in 1930 hosted an international exhibition, which saw the breakthrough of functionalism as a new radical architectural approach to the needs created by urban growth and the demands for the modernization of domestic life. Alva Myrdal, like many members of the Swedish radical elite at the time, was enthused and joined one of the architects of the movement, Sven Marklund, using her growing skills as a social scientist and pedagogue to put her own progressive ideas into practice. Her first published article was a polemic in support of collective service housing and its significance for a more democratic form of family life, with both men and women at work. She also wrote a much noted book on the nursery needs of city children and articles on the design of child-friendly pedagogical toys. The collaboration led to the construction of a block of flats in Stockholm, intended as a social experiment, which had communal eating facilities and well-staffed nurseries with modern, child-friendly equipment. The building itself, though much discussed, was not replicated largely due to costs – particularly for providing service staff – but also because of resistance from both men and women to its concept of a professionally organized, expert-led, family life. But the significance of child–friendly, state-supported, affordable housing with communal playgrounds and laundry facilities became generally accepted as a key plank in the Social Democratic Reform agenda. Her activism in the area of housing led to her serving as expert advisor on a major governmental commission on housing in 1935. Her feminism and visions for modern womanhood, expressed in numerous magazine and news paper articles on a range of topics relating to women's right to work, better working conditions whilst having children, and women's education, as well as her activities in various women's organizations, led to her chairing a government commission on women and work. Its 1938 report tried to present a more realistic picture of the needs of women, (particularly married women in the labor market) than was politically customary at the time. Her activities in the International Federation of Business and Professional Women led to her becoming its Vice Chairman between 1938 and 1947, and gave her a foothold amongst politically active women internationally.

Welfare Theorist for Families

Having since her youth publically argued for improved rights for women and children across all social classes, the impact of Alva Myrdal's theorizing became seriously influential through the publication of a book written jointly with her husband, on population policy and the declining birthrate in Sweden during the decades between the two World Wars (1934). The book affirmed her position as a major theorist of the welfare state and the social rights of citizenship. Through closely shared writings, the book brought together the Myrdals' earlier separate writings on economics and the family to form a major critique of liberal laissez faire and voluntary approaches to the maintenance of a basic standard of life for all. It also criticized a narrow Marxist view of historical determinism. Only the State, they argued, could provide financial security for all during periods of economic or individual hardship. But only a democratic State, inclusive of and responsive to its citizenry with respect to values, rights, and duties, could do so in a way that ensured justice and the preservation of individual freedom. In focusing on population decline, the Myrdals painted a picture of a society on the brink of decline with a poorly equipped citizenry, unable to support itself and its children,; and as a consequence a society unwilling to reproduce itself, given the cost of raising children and the need for women to earn wages. A policy of continued restrictions on the availability of birth control, they argued, would only reaffirm social misery and continue the population decline. For social policy to be effective, they argued, both economic and social factors must be preventively addressed through government planning based on up-to-date social and economic specialist expertise. The nation needed to take collective responsibility for children and their well-being. With a wholly man-supported family no longer viable for both economic and political reasons, children were also at risk of being neglected – an unsatisfactory state of affairs when future growth depended on their capacity for independence and learning. Children and youth should be treated as a national asset with their rearing, alongside that of health and education, collectively organized and supported.

Backed by a rich variety of economic and social evidence and influenced by their contact with American-style applied social science, this expose of a poor nation in decline became the focus of a national debate. An immediate bestseller, it proved a focal point for the development of the Swedish Welfare State during the reign of the Social Democratic Party before and after the Second World War. In response to the public debate generated by the book, the Myrdals both became members of the Population Commission established in 1935, which through its work provided the background to specific reforms. They should not, however, be given the full credit for these reforms. Many reforms were already on the agenda of the increasingly successful Social Democratic Party, which in response to the Depression had widened its electoral remit to include interests outside the industrial trade unions, to include farmers and lower-grade white collar workers, ideologically unified under the banner of Sweden as 'the people's home.' The radical enthusiasm with which the ideas were disseminated in a variety of contexts by both the Myrdals proved one of the catalysts for radical change. As

a catalyst, the book should also be understood as a reflection of their urgent wish at the time, to convince the Swedish population that the best weapon against the political instability, violence, and totalitarianism (that was spreading elsewhere in Europe at the time), lay in a shared solidarity for social democratic progress.

From the start their political vision attracted many critics who expressed anxiety about notions of social engineering 'from above'; warned about the over-reliance on professional 'experts' in execution; and deplored the arrogance implied in drawing up detailed norms for the domestic lives of ordinary citizens. Criticism was most forcefully expressed during periods of political opposition to the Welfare State, both from the political right and the left; and it also became particularly directed at Alva Myrdal, herself a housewife and mother. It took its most dramatic form when she was held responsible for the high number of sterilizations undertaken by Swedish doctors during the early postwar decades. Her general recommendations in connection with population planning were those of sterilization as a voluntary means of securing only 'wanted' children. She can be charged with naivety in assuming that when implementing the policy the medical profession would act as democratically and humanely as she envisaged. By the time this unacceptable medical implementation of a largely well-intended law occurred Alva Myrdal had left Sweden, and moved her visions of women's independence and better care for women and children to the international stage. Her trust in 'experts' declined, as her experience widened.

In 1938 Alva Myrdal returned to the United States to accompany her husband, who had become the recipient of a major research grant to investigate American race relations, and began to disseminate information about Swedish welfare reform to American audiences. Her book 'Nation and Family' (1941) which summarized early Swedish reforms in housing, nutrition, care of the disabled, education, contraception, child care issues, and their political rationale soon became a much be supported in a democratic context of individual equality. In this she noted that it was not the women themselves, but marriage as a social and economic institution that was the problem. This was a theme to which she would later return. On the strength of this book she was invited to lecture across the continent, particularly to women's organizations, to explain its major premises. After a brief return to Sweden during the early part of the war, she returned to the United States via a forced stay in Britain in 1941; during this period she wrote Swedish newspaper articles and later a book, about Britain at war, based on wide-ranging interviews and in-depth observations. This work, alongside a book with her husband about their experiences in the United States (1941), should be seen as an intellectual effort to support faith in democracy and free debate during the darkest years of the war.

Educationalist for Democracy

Though often highly critical of American competitive individualism, she maintained a strong belief in the long-term power of the American system of democracy to sustain and improve itself and the role of its common and secular education system. On returning to Sweden later in the war, she became deeply involved with refugee questions, particularly that of the educational needs of orphaned and displaced children. She founded and edited a multilingual newspaper for refugees and continued her regular, sharply prodemocratic, and antiracist contributions to the press. Her international engagement with education was strengthened in her appointment as Founder Chairman of the World Council for Early Childhood Education. She was the Swedish delegate to the ILO conference in Paris in 1945 at which the Declaration of Children's Rights was proclaimed, and she became Swedish delegate to the First General Assembly of UNESCO in 1946.

The combination of party activism, membership of various research commissions and women's organizations, and her growing fame as an expert in education led to Alva Myrdal in 1946 becoming a member of a fresh Social Democratic Commission on Educational Reform – the earlier ones having failed to reach agreement on how to modernize the traditional parallel school system. Her ideas are evidenced in its final, ultimately successful, recommendations of a common school for all Swedish children. This was not a new concept in Swedish educational debates, especially amongst elementary school teachers; but with her arguments drawn from international developments in pedagogy and her access to a variety of political platforms, she became a powerful champion who could speak with the voice of authority and expertise. For a Swedish audience she wrote on the John Dewey-inspired, progressive aspects of the American unified school system and its student-centered, more practical, and civics-oriented pedagogy. She also looked to Great Britain and its reforms to open opportunities for working class pupils. She strongly admired traditional Scandinavian adult education in its more democratic forms of study circles and home learning opportunities. The traditional Swedish school system was highly differentiated, ostensibly on the basis of ability, but in reality on the basis of class, gender, and geographical location. She campaigned forcefully in articles and lectures for undifferentiated, free, nine-year state schooling with a common curriculum of both theoretical and practical subjects. She gained support from women, teachers, and youth organizations in the party. There was also teacher support for better training in new forms of child-friendly pedagogies to foster individual autonomy and growth, especially the skills of democratic collaboration across social and vocational divides. War and its antecedents had made education reform a particularly urgent priority, with the German school system, which had been historically influential in Sweden, discredited for its pedagogy of obedience. Internationalism in the curriculum was to be fostered, including the universal teaching of English. Her originally utopian reform proposal gained acceptance, originally for a period of experimentation. The dual school system was finally abolished in 1957.

As so often before and after, in promoting education policy reforms also, Alva Myrdal chose to become fully engaged in the political process of ensuring the popular success of proposed reforms, enrolling sympathetic experts in her support – a strategy that made her open to the charge of social engineering, a concept that after the war gained increasingly sinister overtones. By the time Alva Myrdal returned to the topic of educational reform in the 1970s, as Chair of a report on equality presented to the Swedish Social Democratic Party, she presented evidence that students from privileged backgrounds,

and particularly boys, continued to make the best use of higher education and the career opportunities it offered, indicating that comprehensive school reform had not fully succeeded in all its aims. There was, she argued, yet more work to be done for educational justice to be achieved for women and working class children.

Social Scientist for International Cooperation

In 1947, Alva Myrdal left Sweden to accompany her husband, who had accepted a position as Executive Secretary of the UN Economic Commission for Europe, a key agency in postwar economic revival. Geneva was at the time a city in which, she as a married woman, was not allowed to work. When in 1949 she was offered the post of Director of the UN Department of Social Affairs, she accepted. This was her first full time employment and the highest post ever held by a woman within the organization. She saw this as a platform from which to continue her activism on behalf of women and children, but also used it as a platform from which to continue to propagate the importance of social science research and evidence. Though often highly critical of the politics of the United States and the failure of its politicians and academics to act in defense of its poorer citizens, she retained her conviction that the combination of its economic strength with its democratic creed placed a special responsibility on Americans to foster improvements in the everyday lives of citizens across the globe; and she saw the agency of the United Nations as a vital organization in the furthering of international cooperation. During the harsh years of the Cold War, US politicians were not always grateful for her loyalty.

After 2 years at the UN headquarters in New York she returned to Europe when she was appointed director of the UNESCO Department of Social Science in Paris. During her stewardship, a network of international social science associations was financially supported and organizationally facilitated with the hope that such associations would help advance an improved culture of social scientific approaches to social problems, particularly the democratization of access to information and its worldwide dissemination. She encouraged bibliographical initiatives and information overviews in specified fields, such as intergroup cooperation, access to education for women, migration, and racism. Her faith in social science research as a 'toolbox' of methods, in principle available to all as a means toward more effective and democratically derived solutions to social problems steeped in entrenched political and value conflicts, was forcefully expressed in a talk she gave at the School of Social Work at Columbia University (1955). In this she expressed her continuing optimistic faith in the power of 'rational' social science, internationally pooled, to act as an important tool in support of both democracy and improved social welfare of all peoples.

Alva Myrdal's growing interest in international development strengthened her ongoing commitment to the cause of women as both workers and mothers. Her UNESCO position gave her a vision of how politically effective an international comparative project on these issues could be. Her aim, as in previous work with government commissions, was to 'take stock' of the situation for women in industrialized nations and make the information available to policy makers. She enrolled the collaboration of the British sociologist Viola Klein, author of a critical study of patriarchal attitudes in social sciences, and like herself, also a member of the International Federation of Professional Women. Myrdal used her wide international political and feminist organizational networks to collect a broad array of statistics on women in Sweden, Germany, France, and the USA; and Klein provided the more academic, conceptual summaries and assembled the data into accessible tables and pictograms. The analysis was framed by the fact of women's increased longevity, which left a smaller proportion of a woman's life devoted to childbirth and child rearing and increased women's demands for better educational and employment opportunities. Both authors believed that children needed their mothers for at least the first years of their lives, and with men's domestic contribution still in its infancy, as they saw it, governments and employers needed to support women's fluctuating career trajectories with subsidized high-quality child care, part-time employment opportunities, and antidiscriminatory employment regulations. In the economic boom after the war there was an underlying assumption that the demand for labor would continue to grow, and this was a demand unlikely to be met unless women's participation in the labor market was made easier. Their book, 'Women's Two Roles' (1956) proved a popular success and a much-used textbook in the emerging field of the sociology of women, and it was translated into six languages. But what was seen as controversial in more catholic contexts, was seen by the new, more radical developments in feminism as 'traditionalist' in prioritizing women's responsibility for children over that of men. The dilemmas outlined in the book and its proposals for child care policies and employment regulations to support the woman's right to choose her own life trajectory, remain as relevant today as when they were formulated in what we can now see as a manifesto in need of permanent reaffirmation.

Ambassador for Peace and Security

In 1955 Alva Myrdal left her post at UNESCO to become Sweden's Ambassador to India, Ceylon, Burma, and Nepal. This assignment and her friendship with Prime Minister Jawaharlal Nehru, a strong advocate of universal education and political nonalignment, further deepened her insight into the need for social reform and improved social indicators of the impact of technological development on families and communities. Again she traveled widely, frequently lecturing on the Swedish reform experience, and especially on the role of cooperative and women's organizations. After her return to Sweden in 1961, she continued to work with the Social Democratic Party as a Member of the Swedish Senate, as a Cabinet Minister, and as Ambassador-at-large, – the last a role which led to a new and intellectually and politically all-consuming task, disarmament. As member of the Swedish UN delegation in New York, in 1962 she became Chief Swedish Delegate to the Geneva Disarmament Committee and spent the next decade working with it as representative of nonaligned states; in her own eyes she saw herself as the only woman at the negotiating table representing women across the world, who were faced with living through the consequences of big power politics. The convention led to many minor disarmament

treaties in the area of, for example, test bans, and an improved role for nonaligned states in inspection processes; but the holistic solutions to banning nuclear weapons that she and the nuclear disarmament movement as a whole had hoped for, did not materialize. In the book in which she summarized her experience of the disarmament negotiations, directed at decision makers, independent experts, and concerned citizens in all countries; she did not mince words when she wrote of her increasing feeling of near despair over the behavior of the superpowers indulging in subterfuges and half truths, with their closest dependent allies following suit or keeping silent (1976). Throughout the process, the competitive race between the two superpowers had steady escalated and the militarization of the economy and national life of all almost countries had intensified.

The guiding principle of her criticism was rationality, she argued, and it was the very principle that had been so absent in the deliberations. National security needs had become defined and exaggerated by vested interest, secretiveness, and falsifications of reality. The ordinary public should, she felt, both demand honest accountability of the policy makers and question their underlying ethics. Having painted a bleak future of an escalating arms race, her own call for rationality came to focus around continuing efforts to inform and educate politicians and the public about the powerful lobbies propelling the arms race: the armament industry, scientific and technological interests in weapons performance, and the military notion of balancing and matching as a security measure. More pragmatically, her own agenda for moving forward included less holistic measures, such as the restriction of arms sales, the prohibition of cruel antipersonnel weapons, the demilitarization of ocean space, and the establishment of verification agencies. She emphasized the need for independent action of the great majority of nonaligned states as a means of compelling change, particularly with respect to information gathering about matters relevant to international security to limit the information control exercised by the superpowers and enable the exposure of the lack of honesty in military affairs. The book was an appeal for improved political morality and democratic accountability, reinforced by her own financial support for its wide dissemination.

One of the more important long-term impacts of Alva Myrdal's activities for disarmament and her commitment to information transparency and accountability, was her role in the creation of the Stockholm International Peace Research Institute (SIPRI) in 1966. She became its first Chair. The institute, established to commemorate Sweden's 150 years of unbroken peace, remains one of the more important independent research resources on global security, dedicated to scientific research into the hard facts of war, military expenditure, the arms trade, disarmament and nonproliferation, and to facilitate dialogue by making such data accessible to all interested parties. It cooperates closely with major intergovernmental organizations such as the UN, as well as with scientific and government delegations on a range of issues; and its annual reports are widely used in international contexts. Its annual yearbooks, monographs, and policy papers have become important academic source material for peace and conflict research. But Alva Myrdal's original vision of a world in which sources of insecurity are identified and understood allowing conflicts thereby to be prevented, remains distant. Given the lack of international solidarity and lack of mutual trust among power blocks, the best a disarmament campaigner can hope for, as Alva Myrdal foresaw, continues to be the power of exposure through the use of well-assembled and trustworthy evidence.

In the last decades of her life, alongside her continuing work for the Swedish government on a variety of policy preparatory commissions, including the topics of equality and the relationship between the Church and the State, she held several visiting professorships in the United States and was the recipient of Honorary Doctorates from both American and European Universities. In 1970 she was, jointly with her husband, awarded The West German Peace Prize. She was the first recipient of The Albert Einstein Peace Prize in 1980. In 1982, shortly before her death, she was awarded the Nobel Peace Prize (with Alfonso Garcia Robles, Mexico) for her work for peace and disarmament. With the prize money she supported research on the naval arms race and peace education. She died in Stockholm in 1986 of a brain tumor, having for a period suffered from heart disease and debilitating aphasia.

Alva and Gunnar Myrdal's three children: Jan (Myrdal), b. 1927, Sissela (Bok), b. 1934, and Kaj (Folster), b. 1936, all gained authorial fame; and all three wrote conflicting biographical narratives of life in the Myrdal family. Alva Myrdal's private papers, alongside those of her husband, were left with the Labor Movement Archive in Stockholm; they included their large mutual correspondence, the public availability of which has led to considerable speculation over her private life as a woman and mother, and its relationship to her intellectual and professional activities. Thus her own long and politically active life became as important a legacy in debates about the nature of the relationship between the public and the private, the personal and the political, as her many publications and organizational activities in support of a fairer and safer world.

See also: Children, Rights of: Cultural Concerns; Education, History of; Family and Kinship, History of; Gender History; Social Movements, History of: General; Welfare State, History of; Welfare: Philosophical Aspects.

Bibliography

The Alva and Gunnar Myrdal Archive at Labour Movement Archives and Library. Stockholm. www.arbark.se.

Bok, S., 1991. Alva Myrdal: A Daughter's Memoir. Perseus Publishing, Cambridge, MA.

Ekerwald, H., 2000. Alva Myrdal: making the private public. Sociology 43 (4), 343–364.

Etzemuller, T., 2010. Die Romantik der Rationalitat: Alva and Gunnar Myrdal – Social Engineering in Schweden. Transcript Verlag, Bielefeld.

Herman, S.R., 1993. From international feminism to feminist internationalism: the emergence of Alva Myrdal, 1936–1955. Peace and Change 18 (4), 325–346.

Hirdman, Y., 2008. Alva Myrdal: The Passionate Mind. Indiana University Press, Bloomington, IN.

Holmwood, J., 2000. Three pillars of welfare state theory: T.H. Marshall, Karl Polanyi and Alva Myrdal in defence of the national welfare state. European Journal of Social Theory 3 (1), 23–50.

Lyon, E.S., 2000. Biographical constructions of a working woman: the changing faces of Alva Myrdal. European Journal of Social Theory 3 (4), 407–428.

Myrdal, A., 1935. Stadsbarn. Kooperative Forbundet, Stockholm.

Myrdal, A., 1941. Nation and Family: The Swedish Experiment in Democratic Family and Population Policy. Harper and Brothers, New York.

Myrdal, A., 1955. A scientific approach to international welfare. In: Myrdal, A., Altmeyer, A.J., Dean, R. (Eds.), America's Role in International Social Welfare. Columbia University Press, New York.
Myrdal, A., 1971. Towards Equality: The Alva Myrdal Report to the Swedish Social Democratic Party. Prisma, Stockholm.
Myrdal, A., 1976. The Game of Disarmament: How the United States and Russia Run the Arms Race. Pantheon Books, New York.
Myrdal, A., Klein, V., 1956. Women's Two Roles: Home and Work. Routledge and Kegan Paul, London.
Myrdal, A., Myrdal, G., 1934. Kris i Befolkningsfragan. Tiden, Stockholm.
Myrdal, A., Myrdal, G., 1941. Kontakt med Amerika. Bonnier, Stockholm.
Myrdal, A., Vincent, P., 1949. Are We Too Many? UNESCO, Food and People. Bureau of Current Affairs, London.
The Nobel Foundation. www.nobelprize.org.
Stockholm International Peace Research Institute. www.sipri.org.
Terling, B., 1987. Alva Myrdal: An Annotated Bibliography 1932–1961 (With Parallel Text in English). Alva och Gunnar Myrdals Stiftelse, Stockholm.
United Nations Intellectual History Project. www.unhistory.org.
Wallensteen, P. (Ed.), 2003. Alva Myrdal in International Affairs. Uppsala Pub. House AB, Uppsala.

Myrdal, Gunnar (1898–1987)

Richard Swedberg, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Gunnar Myrdal (1898–1987) is best known for his work in economics and his advocacy of the welfare state in his native Sweden as well as elsewhere. Trained in neoclassical economics, Myrdal later turned to institutional economics. His two most famous works are *An American Dilemma* (1944) and *Asian Drama* (1968). In the former, he presented a sociological analysis of racism in the United States, which was much discussed at the time. And so was *Asian Drama* with its thesis that a progressive and strong state is central to economic development. Together with his wife Alva Myrdal, Gunnar Myrdal helped to shape the Swedish welfare state in important respects.

Gunnar Myrdal (1898–1987), Nobel Laureate in economics, was a passionate advocate of equality, social reforms, and the welfare state. His work as a young economist was inspired by Knut Wicksell and centered around an attempt to introduce expectations into conventional price theory. It was also through his studies in economics that Myrdal became interested in the role of objectivity in the social sciences, a topic to which he was to devote much attention. Myrdal argued that values always influence social science and that the social scientist should state his or her values in a forthright manner. Myrdal became famous in the United States through *An American Dilemma* (1944) and in the rest of the world through *Asian Drama* (1968). The former can be characterized as a sociological analysis of racism in the United States; and Myrdal argued that there exists a conflict in the heart of the average American between the lofty ideals of 'the American way of life' and the tough realities of everyday life. *Asian Drama* lacks a single thesis and represents rather an attempt to weave together a series of themes, including the proposition that a progressive and strong state is central to economic development and to the introduction of equality and freedom in developing countries.

Gunnar Myrdal, social scientist of encyclopedic range, is best known for his two monumental studies *An American Dilemma* (1944) and *Asian Drama* (1968). He was a founding member of the Stockholm School of Economics and one of the architects of the Swedish welfare state. Throughout his life Myrdal was also deeply concerned with the role that values play in the social sciences.

Early Life and Career as an Economist

Karl Gunnar Pettersson – the father later changed the name of the children to 'Myrdal' – was born on 6 December 1898 in Skattungbyn in the province of Dalarna, Sweden. His mother, Anna Sofia Pettersson (born Karlsson), came from a family of farmers and was soft spoken, religious, and bitterly unhappy in her marriage. The father, Carl Adolf Pettersson, also came from a family of farmers but was secular, restless, and upwardly mobile. The father was a building constructor, and in 1904 he moved the family to Stockholm. Myrdal went to a very good high school, Norra Real, and was mainly interested in the natural sciences. He, however, was also deeply influenced by the ideas of the Enlightenment, to which he was introduced by one of his teachers. All through his life he would display a passionate faith in the power of reason.

After having graduated with excellent grades in 1918, Myrdal became a student of law at Stockholm University (*Stockholms Högskola*). The decision to study law had been influenced by his father, but also by Myrdal's own conviction that law would help him better understand how society operates. Whatever conservatism and class prejudices still dominated his mind, and which he had picked up from his family, were soon to be challenged by Alva Reimer (1902–86), whom he met in the summer of 1919 and married in 1924. Like her husband, Alva Myrdal was soon to become a first-rate social scientist and politician; in 1982 she also received the Nobel Peace Prize. While Myrdal had never enjoyed studying law, he quickly took to economics. In 1927 he got his doctorate through a dissertation for which Gustav Cassel had been the advisor, *Prisbildningsproblemet och föränderligheten* (*The Problem of Price Formation and Change*). In this work Myrdal attempted to develop the beginnings of a dynamic theory of price formation by introducing the idea of expectations. In 1930 he published *Vetenskap och politik i nationalekonomien* (*The Political Element in the Development of Economics*), in which he argued that economists often confuse values and facts in their studies, something which endangers both economics as a science and its potential use for social reform. Based on an article from 1932, Myrdal published a book-length study in 1939 entitled *Monetary Equilibrium*. In this work he attempted to further develop Knut Wicksell's theory of monetary equilibrium by introducing the notion of expectations. In the early 1930s Myrdal also produced a couple of writings for the Swedish Social Democratic government in which he advocated an underbalanced budget as a way of countering a downturn in the business cycle. These works were based on Myrdal's discussions with his friend Erik Lindahl and the Swedish Finance Minister Ernst Wigforss.

All of these writings, except for *Vetenskap och Politik*, are regarded today as important contributions to the Stockholm School of Economics, which is usually defined as a Keynesian type of economics, based, however, on Knut Wicksell's ideas. The value of these works is today mainly historic. They nonetheless played a crucial role in the development of Myrdal's

ideas, especially in his attempt to construct a dynamic theory of causation, based on Wicksell's ideas. *Vetenskap och Politik*, on the other hand, has aged rather well and represents a still valuable contribution to the history of economic ideas. While Myrdal would always approvingly refer to its main thesis – that neoclassical economics is deeply biased through its origin in the natural law tradition and in utilitarian philosophy – he would later criticize his youthful assumption that it was possible to separate out a core of economics that was free from all values.

Political Activities and an American Dilemma

Both Gunnar and Alva Myrdal decided in the early 1930s to become actively involved in politics and were to play a crucial role in the construction of the Swedish welfare state through their advocacy of 'social engineering.' The timing to do so was perfect, since in 1932 the Social Democratic Party had won the election and was eager to transform the poor and class-ridden Sweden into a real 'people's home' (*folkhem*). Gunnar Myrdal not only helped to formulate the economic aspects of the new policy that now came into being, but also its social policy. This he did, in close collaboration with Alva Myrdal, through an avalanche of writings as well as his participation in various royal commissions, appointed by the government to investigate some specific issue. The basic message of the two Myrdals was that social policy had to be fundamentally changed, from focusing on symptoms, to becoming 'prophylactic.'

Gunnar and Alva Myrdal presented their vision for what a new and modern Sweden would look like in *Kris i befolknigsfrågan* (*Crisis in the Population Question*, 1934), which became a political best seller. 'All of society,' they here argued, would have to be changed, so that 'the quality of people' could be improved and better adjusted to the demands of modern, industrial society: living standards, consumption patterns, the way children are brought up, the structure of education, and much, much more. Though permeated by the authors' generosity and enlightenment, *Kris i befolkningsfrågan* also evinces an enormous belief in the authority of science and experts. The Myrdals' advocacy of sterilization of certain categories of people, who were deemed unfit for the new society, has also been criticized.

By the mid-1930s, Myrdal's work had become so dominated by practical and political concerns that he longed for a return to academic studies; and when he was asked in 1937 by the Carnegie Corporation if he would lead a study on the blacks in the United States, he accepted. During the next few years, Myrdal traveled around in the United States like a latter-day Tocqueville, interviewing people from all strata – 'I am from Sweden and I do not know anything about the race problem. What can you tell me?' Through the backing of the Carnegie Corporation, Myrdal put together a small circle of collaborators, and he also engaged, in one way or another, some 70 other scholars, including Samuel Stouffer, Louis Wirth, and E. Franklin Frazier. Fearing many times that he would never be able to synthesize the enormous amount of material that he had commissioned in 1944 Myrdal had the satisfaction of seeing *An American Dilemma* appear in print in two volumes of more than 1400 pages.

Myrdal's work was much commented upon, and most of what was said was positive. *An American Dilemma* was actually to dominate the discussion of ethnicity in the United States from the end of World War II till the 1960s, and it also played an important role in the decision of the Supreme Court in 1954 to declare segregated schools unlawful (*Brown vs Board of Education*). From the mid-1960s till the 1970s, Myrdal's book was criticized for depicting certain aspects of black life as pathological and also for not advocating change from the bottom up; and today, *An American Dilemma* is seen primarily as being of historical interest in ethnic studies.

Myrdal's basic thesis – that there is a moral dilemma in the heart of the average person between 'the American Creed' and everyday behavior toward blacks – may seem old-fashioned and passé today. This, however, should not prevent one from realizing the value of other parts of *An American Dilemma*, especially its appendices on values and on cumulative change (Appendices 1–3). In the first two of these appendices, Myrdal argues that social scientists should always state their value premises since these deeply influence the analysis. In Appendix 3 Myrdal presents the core of his own social science approach that is centered around Wicksell's theory of unstable equilibria or what Myrdal calls 'the principle of cumulation.' An initial change, Myrdal explains, can set off further changes; and these may then move the whole system either in an upward direction or in a downward direction ('vicious circles').

Activities After World War II and Becoming an Institutionalist

After *An American Dilemma* Myrdal returned to a combination of politics and applied research till the mid-1950s, serving first as Minister of Trade in Sweden (1945–47) and then as Executive Secretary in the Economic Commission for Europe of the United Nations (1947–57). Neither of these two experiences were particularly successful for Myrdal, who was more of an intellectual than a politician; and he soon longed to return to academic studies. The topic that he chose to focus on this time was underdevelopment, and a fine introduction to his early efforts in this direction can be found in *Rich Lands and Poor* (1957). It is in this work, for example, that Myrdal introduces the idea that a strong economic development in one area may either impede or accelerate economic growth in nearby areas ('backwash effects' and 'spread effects'). His most sustained effort to understand underdevelopment, however, took 10 years to produce and appeared in 1968 in three volumes of more than 2000 pages: *Asian Drama*.

Like *An American Dilemma*, *Asian Drama* has been much debated. The initial reaction was many times positive, but today Myrdal's work is less often referred to. One reason for this is the fact that his strong advocacy of large-scale state planning and state intervention as well as his clear stance for what he calls 'The Modernization Ideals' do not go very well with today's emphasis on market forces and the prevalent skepticism toward Western ideals. Still, several parts of *Asian Drama* are still of interest, such as the notion of 'the soft state,' the critique of various aspects of neoclassical economics, and Myrdal's attempt to use Wicksell's theory of cumulative causation to

construct a theory of institutionalist economics (see especially Appendices 1–3).

After *Asian Drama* Myrdal kept writing books and articles, and he also gave lectures all over the world. In 1969 he published a small and still useful volume entitled *Objectivity in Social Research*, and in 1974 he was awarded the Nobel Prize in Economics, together with Friedrich von Hayek. Autobiographical elements can be found in *Against the Stream* (1973) as well as in *Hur styrs landet?* (*How is the Country Governed?* 1982). His reference to the superiority of institutionalist economics over neoclassical economics became ever stronger; and the essay *Institutional Economics* (1978) is especially recommended. An attempt to write a sequence to *An American Dilemma*, entitled *An American Dilemma Revisited*, had, however, to be called off in 1985, due to illness. Gunnar Myrdal died on 17 May 1987 in Stockholm. To his very last day he remained a firm believer in the ideals that he shared so strongly with his beloved wife Alva Myrdal, and which for him were symbolized by the one word with which *An American Dilemma* ends: *Enlightenment*.

See also: Constitutionalism, Comparative; Economic Growth; Racial Relations; Racism, History of; Racism, Sociology of; Southeast Asia: Sociocultural Aspects.

Bibliography

Andersson, S., 1986. Alva och Gunnar Myrdals arkiv. Arbetarhistoria 10, 75–78.
Assarsson-Rizzi, K., Bohrn, H. (Eds.), 1984. Gunnar Myrdal: A Bibliography, 1919–1981, second ed. Garland Publishers, New York.
Jackson, W.A., 1990. Gunnar Myrdal and America's Conscience: Social Engineering and Racial Liberalism, 1938–1987. University of North Carolina Press, Chapel Hill, NC.
Myrdal, A., Myrdal, G., 1934. Kris i befolkningsfrågan (Crisis in the Population Question). Albert Bonniers Förlag, Stockholm.
Myrdal, G., 1927. Prisbildningsproblemet och föränderligheten (The Problem of Price Formation and Change). Almqvist and Wicksell, Uppsala, Sweden.
Myrdal, G., 1930. Vetenskap och politik i nationalekonomien. P. A. Nordstedt and Söners Förlag, Stockholm [Trans. 1953 by Paul Streeten as The Political Element in the Development of Economics. Harvard University Press, Cambridge, MA].
Myrdal, G., 1939. Monetary Equilibrium. Hodge, London.
Myrdal, G., 1944. An American Dilemma: The Negro Problem and Modern Democracy. Harper, New York.
Myrdal, G., 1957. Rich Lands and Poor. Harper, New York.
Myrdal, G., 1968. Asian Drama: An Inquiry into the Poverty of Nations. Twentieth Century Fund, New York.
Myrdal, G., 1969. Objectivity in Social Research. Pantheon Books, New York.
Myrdal, G., 1973. Against the Stream: Critical Essays on Economics. Pantheon Books, New York.
Myrdal, G., 1978. Institutional economics. Journal of Economic Issues 12, 771–783.
Myrdal, G., 1982. Hur styrs landet? (How is the Country Governed?) Rebén and Sjögren, Stockholm.
Myrdal, G., 1987. Historien om An American Dilemma (An American Dilemma Revisited). SNS, Stockholm.
Myrdal, G., 2005. The Essential Gunnar Myrdal. New Press, New York.

Mysticism

Dan Merkur, University of Toronto, Toronto, ON, Canada

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 10270–10273, © 2001, Elsevier Ltd.

Abstract

Mysticism consists of a practice of religious ecstasies, together with ideologies, ethics, rites, myths, legends, magics, etc., related to the ecstasies. Mysticism is institutionalized in 90 percent of human societies. Mysticism occurs not only in scripture-based religions, but also in oral religious traditions, where it is known as shamanism, vision quests, spirit possession, spirit mediumship, and prophetism. In oral religions, mystics often assume leadership positions as intermediaries between the numina and the social group. Because book religions invest authority in scriptural exegetes, mystics are marginalized and typically seek politically innocuous ecstasies. In other cases, mystics aspire to authority and become would-be prophets. The contents of mystical experiences may be subdivided into narrative and unitive ecstasies. Both occur in states both of trance and of reverie. Trances reify imaginations as hallucinations and delusions, whose uncritical validation devalues perceptible reality and promotes dissociative philosophies and behavior. Reveries are instead self-evident as heurisms. Mystics have traditionally sought metaphysical transformations, but positive psychological growth has been demonstrated in a few cases.

Definition of Mysticism

Mysticism consists of a practice of religious ecstasies, i.e., of religious experiences during alternate states of consciousness, together with whatever ideologies, ethics, rites, myths, legends, magics, etc., are related to the ecstasies. 'Ecstasy' may be defined as any experience of involuntary belief in the reality of the numinous, i.e., of the holy or demonic. Whenever the numinous is as manifestly real as perceptible reality is during waking sobriety, consciousness has been sufficiently altered that the experience is intrinsically religious (Merkur, 1992).

Since the paradigm shift promoted by the work of Katz (1978, 1983), mysticism is no longer defined as the ideology and practices surrounding the experience of union with the Absolute (Underhill, 1910), or ultimate reality, etc. There is abundant evidence that the people who traditionally have been called mystics have enjoyed many varieties of religious ecstasies. Comparative studies of mysticism currently include the vision quests and shamanism, spirit possessions and spirit mediumship, prophetism and seers, etc., of the oral religious traditions, as well as the visions, voices, mystical unions, enthusiasms, etc., of the scripture-based religions.

Mystical experiences are as varied as the dreams of sleep. Not only personal matters, but also culturally and historically variable religious ideologies contribute to the contents of mystical experiences and their postexperiential interpretations. People in all cultures may potentially experience all possible varieties of religious experience. The world's ecstatic religious traditions each tends to accommodate a considerable diversity of ecstatic experiences, even when certain types are privileged. Conformity is required in matters of doctrine and behavior much more than in experience. The vagueness or nonspecificity of many traditional doctrines permits them to be applied simultaneously to diverse religious experiences.

No sociological research has yet been done that conforms with mysticism's current definition. Collating existing research on shamanism, spirit mediumship, prophetism, mysticism in the book religions, etc., in search of common denominators, discloses few commonalities. People who make religious uses of alternate states of consciousness take their religion more personally and seriously than the general population. They tend to be convinced, and not only hopeful, about the reality of their objects of belief. Religion is not only a social institution for them; it is an integral component of the cosmos. In Allport's (1959) terms, their religiosity is 'intrinsic,' and not only 'extrinsic.'

In other respects, mystics are diverse. Some mystics are religiously and/or socially conservative, but others are revolutionary. Some seek to transcend all perceptible phenomena; others address the immanence of spiritual phenomena within material reality. Some shun their bodies and engage in asceticism. Others celebrate both their bodies and the physical world. Some seek to withdraw from society, but others participate within it. Some mystics prioritize a single type of mystical union, e.g., nothingness, or all-being, or identity with God; but others favor complex systems of multiple types of religious ecstasy that include both visions and unions.

The term 'mysticism' derives historically from the noun *mystes*, 'mystic,' meaning the possessor of a secret, and denoting an initiate of a classical Greek or Hellenistic Mystery. The Mysteries were rites of initiation that induced religious experiences. Technical terms of the Mysteries were taken over by Plato, Philo, and the Christian Church Fathers, who applied the terms to their own practices of religious experiences. By the early medieval period, Christianity officially denied possessing any secrets that had not been publicly divulged through Christ. Hindu yoga, Buddhist *vipassana* meditation, and Confucian meditation, among other mystical traditions, similarly teach their alternate-state practices openly. At the opposite extreme, Taoism, Hindu and Buddhist tantra, theosophical Sufism, the Kabbalah, and Western alchemy are so very secretive that they encode their teachings in secret languages, whose meanings cannot be understood by noninitiates.

Incidence and Varieties of Mystics

In a survey of 488 societies worldwide, Bourguignon (1973) found that 90 percent had institutionalized one or another

religious use of alternate states of consciousness. In different surveys in Western cultures, between one-third and one-half of the population has been found to have had one or more mystical experiences (Greeley, 1975). A considerably smaller part of the population pursues mystical experiences on a frequent or habitual basis. Many factors have contributed to the distinctive historical forms that different mystical traditions have taken.

Shamans

A shaman may be defined as a religious functionary who enters ecstasy and, by commanding the aid of helping spirits, intermediates between the social group and the numina (Hultkrantz, 1973). Shamanism has its origin and major distribution in hunter cultures. Like hunters, shamans' ecstasies are highly mobile. In their visions, shamans may travel out of their bodies to the sky, netherworld, or distant overland locations. Shamans also summon helping spirits to them, sometimes from considerable distances, and they send them abroad on distant errands. Shamans relate to their helping spirits much as hunters relate to half-wild dogs.

A shaman typically is paid for services rendered to a client, such as healing. In their ecstasies, shamans may also find game animals, placate angry deities, guide new ghosts to the afterlife realm, and hex enemies in war. In some hunter cultures, vision quests are pursued more generally by part or all of the society. In these cases of 'democratized shamanism,' shamanism occurs as a private practice, and only mystics who are exceptionally gifted emerge as professionals.

Shamanism persists in agricultural communities in varied forms. Democratized shamanism persisted as a practice among warriors in Indo-European cultures (e.g., Glosecki, 1989). It was also adapted to group possession states in the classical Mysteries of Dionysus. In late antique and medieval Europe, shamans who lacked traditional social locations within their communities emerged as itinerant seers (Brown, 1981; Davidson, 1989), magicians (Smith, 1978), and witches (Ginzburg, 1983). Where seers have a clientele, magicians and witches often engage in shaman-like conjuring of spirits, but privately rather than professionally.

Spirit Mediums

A spirit medium may be defined as a religious functionary who enters ecstasy and, through states of possession, serves as a vehicle for the manifestation of a numen to the social group. Spirit mediumship has its major distribution in clan-based farming villages, both in Africa (Walker, 1972) and in Southeast and East Asia (Elliott, 1955; Paper, 1995). The sedentary life style of farming villages is reflected in the lodging of spirits within mediums. Mediums and their spirits perform no errands that require travel outside the village. Much spirit mediumship is intimately connected with the worship of ancestral ghosts. Deference to family elders, which is economically motivated by the ownership and inheritance of arable land, is continued after the elders' deaths through the agency of mediums, whom the ancestral ghosts possess. The deference is reflected, among other manners, in mediums' claims that all of their activities while possessed are performed by the possessing spirits, while they themselves are passive.

Spirit mediums may also be possessed by gods and spirits that are not ancestral ghosts. Spirit mediumship also occurs in democratized forms, where large numbers of people within a social group engage communally in possession states, and only a few show sufficient virtuosity to function as professionals. Spirit mediumship and possession may persist into more complex societies, e.g., in the theurgy of late antique Hellenism (Luck, 1989), enthusiasm in Protestant Christianity (Knox, 1950), and modern New Age channeling.

Prophets

A prophet may be defined as a religious functionary who enters ecstasy, learns of the will of the god, and later communicates the divine will to the social group. Prophets have their major distribution in kingdoms, whose gods are imagined as kings, and whose prophets are imagined as heralds, spokespeople, or ambassadors of the gods (Cross, 1973). Prophetism had its origin and classical location in urban societies with differentiated labor and trades, both in the ancient Near East and East Asia. Prophetism can be conceptualized as an adaptation of ecstasy to monarchic politics, with the prophet's god outranking the king. Prophets' ecstasies favor verbal revelations from their gods, but may include shamanic and mediumistic forms that have been adapted to monarchic politics.

Anthropologists also use the term 'prophet' in reference to shamans and spirit mediums who have been influenced by missionaries' accounts of Biblical prophets and developed hybrid types of mysticism that reflect their postcontact situations.

Mystics in Scripture-Based Religions

Because book religions invest religious authority in scriptures and their official interpreters, they marginalize professional ecstatics. Mysticism is then often pursued by a leisure class, which may embrace poverty in order to support its practices, as among Hindu forest dwellers, and in Buddhist and Christian monasticism. In other cases, e.g., Taoism, Sufism, and Kabbalah, mysticism has been pursued by individuals sufficiently affluent to devote leisure time to it.

Mystics in the book religions often avoid conflict with religious authorities by favoring varieties of ecstatic experience that are socially and political powerless. Unitive experiences, which are not useful for professional ecstatics, may consequently be privileged. However, other ecstasies are often pursued as well.

Mystics in the book religions may be public religious functionaries, e.g., as leaders or teachers of other mystics. Their leadership roles among nonmystics tend to be limited to religious revival movements, in which mystics proselytize ecstasy among their co-religionists.

Although most mystics are sociopolitically conservative, a minority are radical and potentially revolutionary. In such cases, mystics in the book religions typically seek social roles appropriate to prophets. Rather than seeking ecstasies that are accommodating, inconspicuous, and innocuous, they seek inspirations that claim authority to transform their religions and/or societies. The would-be prophets may then

emerge as charismatic founders of sects, schisms, and/or cults. An apocalyptist may be conceptualized as a would-be prophet in a book religion, who seeks the radical reformation of society, often as an aftermath of catastrophe and revolution. Apocalyptists may found or belong to millennial movements.

The Contents of Mystical Experiences

All ideas and practices that may be entertained and performed during normal waking sobriety can also be thought and done during at least many varieties of ecstasy. Ecstasies facilitate unconscious manifestations that tend not to occur otherwise; but they do not necessarily preclude the activities of normal consciousness. Mysticism may embrace the whole of non-ecstatic religiosity, together with an additional component i.e. contributed by the ecstatic experiences.

Ecstasies may be subdivided on psychological criteria into two major categories. Some ecstasies involve narratives, as do daydreams and night dreams. The narratives may consist of visual images, with characters, sceneries, dialogue, etc. They may instead consist exclusively of words. Narrative ecstasies can also have physical dimensions and take form as automatic speech, automatic writing, automatic dramatization (i.e., possession), and other types of behavior. Somatizations (stigmata) may also occur.

The second category of ecstasies may be termed unitive experiences. All unitive ecstasies involve a sense of separate things becoming unified. What is unified varies. In normal waking sobriety, our sense of self is elastic. Our senses of our 'selves' is limited to our bodies when naked; extends to clothing when dressed; and may include, e.g., even the fenders of a vehicle when we are driving. Romantic partners, children, family, ethnic group, and nation may also be included in the normal fluctuation of the sense of self. Unitive ecstasies include exceptional instances of the elasticity of the self. In some instances of mystical union, the self is co-extensive with the entire physical cosmos; in others, with infinite timeless being; in others again, with God; in still others, with ultimate nothingness. Other commonly encountered unitive ecstacies include a sense of identity with some one particular thing in the cosmos.

Unitive experiences are always unconsciously dyadic. Just as self exists and is defined psychologically only in relation to others, so mystical unity exists and is defined experientially only in relation to the absolutely Transcendent. One cannot experience self as transcendent, unless the ideas of the self and the Transcendent have been brought to union. Conversely, experiences of self and a numinous being in a manifestly dyadic relationship may be categorized as unitive experiences. Self and other are united in a unifying relationship during experiences of the immediate presence of a numinous power, or absolution by a numinous power, etc.

So-called paranormal phenomena, such as precognition, clairvoyance, and meaningful coincidences ('miracles') are further varieties of unitive experience. They unite subjective fantasy with the objectivity of perceptible reality, imparting objective reality to imagined ('extra-sensory') perceptions (Merkur, 1998, 1999).

The Varieties of Alternate States

Most mystical experiences occur during states of trance. Religious trances are instances of either self-hypnosis or hetero-hypnosis. In other cases, mysticism depends on trances that occur as symptoms of hysteria, epilepsy, and other pathologies. In all cases, suggestion – either autosuggestion or hetero-suggestion – is responsible for directing the trance to serve religious purposes (Arbman, 1963).

An important aspect of the trance state is its ability to reify the contents of consciousness (Shor, 1959). For example, a trance state is all that is necessary in order to transform a visual daydream into a waking hallucination. Instead of seeming to be imaginary, daydream images may seem during trance to be objectively existing, external realities. Trance states may similarly transform speculative ideas into delusions that have the force of firm convictions.

Trances are often termed 'dissociative states' because the reification of fantasies leads to a corresponding devaluation and/or denial of sense perceptions. Mystics who use trances uncritically may develop dissociative religious doctrines that reject the perceptible world in favor of transcendent realms or states, otherworldliness, asceticism, devaluation of mortal life, community, and the physical world, and asocial or anti-social behavior (Furst et al., 1976). Mystics may, however, use trance states critically, much as hypnotherapists do.

Some mystical experiences occur in a different class of alternate states that may be termed 'reveries.' Nonecstatic forms of reverie include daydreams, creativity, and hypnagogic states between waking and sleeping. Reveries of ecstatic intensity include 'peak experiences,' which typically occur spontaneously and last only a few seconds or minutes. Sensory deprivation is a means to induce reverie states that may reach ecstatic intensity. Psychedelic drugs, i.e., psilocybin, mescaline, LSD, and THC, the active factor in marijuana, similarly induce reverie states, as distinct from trances (Merkur, 1993).

Both narrative and unitive ecstasies may occur in either trances or reveries. Reveries differ from trances, among other manners, in the ecstatic's knowledge during reverie that his or her experiences are subjective and intrapsychic, rather than objective and external. They are self-evident as heuristic phenomena. Religious reveries seem subjectively to be apperceptions of truth through the medium of imagination. The experiences take form as pseudo-hallucinations and illusions, where the same narrative and unitive experiences have the force of hallucinations and delusions when they occur during trances.

Psychological Transformation

Mystics traditionally have sought transformations that were metaphysical in nature: magical empowerment, salvific perfections of their souls, etc. Since the 1970s, transpersonal psychologists have argued that mystical transformations have powerful psychological benefits, consistent with Maslow's (1964) concept of 'self-actualization.' Some mystical practices have been shown to be beneficial developmentally (Sacks, 1979; Emavardhana and Tori, 1997), but others may be harmful.

See also: Charisma and Charismatic; Christianity, Origins of; Evangelical, Revivalist, and Pentecostal Religions; Healing Practices: Alternative and Complementary; Millennialism; Monasticism: West and East.

Bibliography

Allport, G.W., 1959. Religion and prejudice. Crane Review 2, 1–10.

Arbman, E., 1963. Ecstasy or Religious Trance. Bokforlaget, Stockholm.

Bourguignon, E., 1973. Religion, Altered States of Consciousness, and Social Change. Ohio State University, Columbus, OH.

Brown, J.P., 1981. The Mediterranean seer and shamanism. Zeitschrift fur die Alttestamentliche Wissenschaft 93, 374–400.

Cross, F.M., 1973. Canaanite Myth and Hebrew Epic. Harvard University Press, Cambridge, MA.

Davidson, H.E., 1989. In: The Seer: in Celtic and Other Traditions. John Donald, Edinburgh, UK.

Elliott, A.J.A., 1955. Chinese Spirit-Medium Cults in Singapore. London School of Economics, London.

Emavardhana, T., Tori, C.D., 1997. Changes in self-concept, ego defense mechanisms, and religiosity following seven-day Vipassana meditation. Journal for the Scientific Study of Religion 36, 194–206.

Furst, S.S., Leavy, S.A., Lewis, R.C., Lubin, A.J., Ostow, M., Zales, M.R., 1976. Mysticism: Spiritual Quest or Psychic Disorder? Group for the Advancement of Psychiatry, New York.

Ginzburg, C., 1983. The Night Battles (trans. Tedeschi J, Tedeschi A). Johns Hopkins University Press, Baltimore, MD.

Glosecki, S.O., 1989. Shamanism and Old English Poetry. Garland, New York.

Greeley, A.M., 1975. The Sociology of the Paranormal. Sage, Beverly Hills, CA.

Hultkrantz, Å., 1973. A definition of shamanism. Temenos 9, 25–37.

Katz, S.T. (Ed.), 1978. Mysticism and Philosophical Analysis. Sheldon, London.

Katz, S.T. (Ed.), 1983. Mysticism and Religious Traditions. Oxford University Press, Oxford, UK.

Knox, R.A., 1950. Enthusiasm. Oxford University Press, Oxford, UK.

Luck, G., 1989. Theurgy and forms of worship in Neoplatonism. In: Neusner, J., Frerichs, E.S., Flesher, P.V.M. (Eds.), Religion, Science, and Magic. Oxford University Press, New York.

Maslow, A.H., 1964. Religions, Values, and Peak Experiences. Penguin, Harmondsworth, UK.

Merkur, D., 1992. Becoming Half Hidden. Garland, New York.

Merkur, D., 1993. Gnosis. State University of New York Press, Albany, NY.

Merkur, D., 1998. The Ecstatic Imagination. State University of New York Press, Albany, NY.

Merkur, D., 1999. Mystical Moments and Unitive Thinking. State University of New York Press, Albany, NY.

Paper, J., 1995. The Spirits are Drunk. State University of New York Press, Albany, NY.

Sacks, H.L., 1979. The effect of spiritual exercises on the integration of self-system. Journal for the Scientific Study of Religion 18, 46–50.

Shor, R.E., 1959. Hypnosis and the concept of the generalized reality orientation. American Journal of Psychotherapy 13, 582–602.

Smith, M., 1978. Jesus the Magician, first edn. Harper & Row, San Francisco.

Underhill, E., 1910. Mysticism. New American Library, New York.

Walker, S.S., 1972. Ceremonial Spirit Possession in Africa and Afro-America. Brill, Leiden, Germany.

Myth in Religion

Robert Segal, University of Aberdeen, Aberdeen, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Modern theories of myth, most of which hail from the social sciences, can be classified into two main categories. In the nineteenth century theorists saw myth as the 'primitive,' outdated counterpart to science, which was considered largely modern. It was assumed that myth arose and functioned to explain the physical world. The personalistic, religious explanation was incompatible with the impersonal explanation given by science. One could not consistently have both myth and science. In the twentieth century theorists saw religion as almost anything but the primitive counterpart to science. Either myth was no longer assumed to be about the physical world, or myth was no longer assumed to serve to explain the physical world, or both. Myth now diverged from science and was therefore compatible with it. Myth was no longer automatically tied to religion. In the twentieth-first century the question is whether myth can be brought back to the physical world as an explanation of it, yet in a way compatible with science.

Myth

There is no uniform definition of myth. For some theories of myth, myth need not even take the form of a story, as might be assumed, but can instead be a belief or a conviction. For other theories, mythological stories need not take place in the past but can take place in the present or the future. Myths can take place on earth, in heaven, or in space. The characters in myth need not be gods but can alternatively be humans or even animals. What happens in myth is open-ended. Myth need not, for example, be about the creation of something but can instead be about the destruction of something or be about an event within an existing family, society, or cosmos.

There are many kinds of myths – among them creation myths, myths of paradise, flood myths, myths of nations and of races, and myths of the future. There is diversity even within each kind. For example, in hero myths, heroes are most often human, yet they can also be divine. Animals, especially ones called tricksters, can be heroic as well. Heroes can be ordinary as well as extraordinary persons. Sometimes anyone can be a hero, which is not to say that anyone automatically is a hero. There can be child heroes as well as adult heroes. Heroes can come from any class. And heroes can come from either gender. Finally, some theories of myth classify stories of heroes as legends rather than as myths proper.

Theories of Myth

A myth can be studied either by itself, in isolation from other myths, or in comparison with other myths. Theories of myth compare myths with one another, indeed with all others. Theories are generalizations about myth. They provide answers to the main questions about any myth: What is its origin? What is its function? What is its subject matter? The answer is the same for every myth.

The question of origin asks why and how myth arises. The origin proposed by a theory is recurrent rather than historical. The origin refers not to the time and place of the first appearance of myth, such as the Garden of Eden, but to why and how myth arises whenever and wherever it does. The answer to the 'why' of origin is not a location but a need, which antedates myth and which myth is created to fulfill. What the need is, varies from theory to theory. It can range from the brute need for food to the sublime need for meaningfulness in life.

The 'how' of origin refers to the means by which myth arises. For example, a theory may contend that even with the need for hunger, a myth will not arise without a group. Or a theory may contend that even with a need for explanation of the physical world, a myth will not arise unless there exists the belief in god. Some theories ignore the how of origin.

The question of the function of myth addresses why and how myth lasts. Like the origin, the function is recurrent. The why of the function is the flip side of the why of the origin: whatever need explains the creation of myth also explains the continuation of myth. The how is the means by which myth lasts. For example, a theory may assert that myth lasts only as long as it is tied to a ritual or to a group.

The subject matter of myth refers to the content. The key issue is whether myth is to be read literally or symbolically, and theories vary in their views. If myth is to be read symbolically, what is being symbolized?

Most theories of myth come from the social sciences – from anthropology, sociology, psychology, political science, or economics. Social scientific theories tend to focus on the origin and function of myth. It is not that social scientists confine themselves to behavior and ignore texts. It is that they tend to emphasize why and how texts are created and read (or heard) and not what texts say. In one sense of an admittedly confusing pair of terms, the social scientific focus is on the *explanation* rather than the *interpretation* of myth (see Segal, 1992).

By contrast, theories of myth from the humanities – art, literature, philosophy, and religious studies – focus on the content, or the *interpretation*, of myth. What is myth saying? is the key issue. What are the themes or patterns in myth? Must myth be a story?

Some theorists seek both to explain and to interpret myth. For example, Sigmund Freud and C.G. Jung not only identify who creates and uses myth and why but also offer full glossaries for reading myth, which means for translating the literal meaning into a psychological one. E.B. Tylor insists on a literal

 http://dx.doi.org/10.1016/B978-0-08-097086-8.84023-4

interpretation of myth to match his view of the origin and function of myth. By contrast, Rudolf Bultmann, Hans Jonas, and Albert Camus confine themselves to the content, for which one might use the somewhat confusing synonym "meaning". They translate myth into existentialist terms and just do not consider why myths are created and perpetuated.

In addition to the questions of the origin, the function, and the subject matter of myth, there is also the question of the truth of myth. Theories can judge myths as true or false or take no stand. Tylor takes myth to be false literally and stops there, refusing to consider it symbolically. Eliade takes myth to be true literally. Bultmann, Jonas, Freud, Jung, and Camus take myth to be false literally but true symbolically.

In popular parlance myth is a false belief and a belief taken literally, whether or not the belief is in the form of a story. Books with titles like 'President X: The Man and the Myth,' pit myth against truth. But myth taken as mere myth misses the point. Myth has a grip that goes far beyond its factual status. To label a belief a myth is to elevate, not denigrate, it by noting the hold that it has on those who accept it, no matter what the evidence. A myth is not the false belief that I have ten dollars in my pocket. It is the conviction, hard to dislodge, that President X was courageous, heroic, selfless, or saintly. At the same time a myth need not be false – contrary to phrases like 'the man and the myth.' A myth can equally be true. The power that myth possesses transcends its truth as much as its falsity. When true, myth is not just a true belief but a resolutely held true belief.

Theories of myth are hardly a modern invention. They go all the way back to the pre-Socratics (see Chase, 1949: Chapter 1; de Vries, 1961: Chapter 1). Only in the last century and a half, however, have theories become scientific, for only since then have there existed the social sciences. Some social scientific theories may have counterparts in earlier ones (see Feldman and Richardson, 1972), but earlier theorizing was largely speculative and philosophical. Modern theorizing is far more empirical and testable. Even modern theories from literature, religious studies, and philosophy reflect the influence of the social sciences.

The most straightforward way to compare theories of myth is by discipline. But the differences among disciplines are often blurry. For example, the theories of Tylor, Bronislaw Malinowski, and Claude Lévi-Strauss are as much psychological as anthropological. It is more telling to compare theories by the questions they ask and by the answers they give.

Myth as the 'Primitive' Counterpart to Science

An apt starting point is the theories of Tylor (1832–1917), the pioneering English anthropologist, and of J.G. Frazer (1854–1941), the pioneering Scottish classicist and anthropologist. They epitomize what can be called the nineteenth – century view of myth. For Tylor (1871), myth originates and functions to explain physical events. For Frazer (1911–15), myth originates and functions not only to explain but also to control physical events, such as rainfall. For both, myth is not itself science, even of a 'primitive' variety, which is the position of Lévi-Strauss (1908–2009), the French anthropologist. Rather, myth is the 'primitive' *counterpart* to science, which is exclusively modern. 'Modern myth' and 'primitive science' are contradictions in terms.

By science, Tylor and Frazer mean natural, not social, science. The events explained by myth are primarily those in the external world, such as the rising and setting of the sun, though also explained are biological events, such as birth and death. For Tylor, the payoff of myth is wholly intellectual: "Man's craving to know the causes at work in each event he witnesses … is no product of high civilization, but a characteristic of his race down to its lowest stages" (Tylor, 1871: Volume I, pp. 368–369). For Frazer, the payoff is practical – above all that of getting food by getting crops to grow.

For both Tylor and Frazer, myth and science alike originate through observation, inference, and generalization. Both cover the same range of phenomena. There is nothing in the physical world that myth cannot explain but that science can, or vice versa. Both myth and science are meant literally, though the inconsistent Frazer alternatively reads myth symbolically. The difference between myth and science is in the explanations themselves. Myth, a part of religion, attributes events to the decisions of gods. Science attributes events to impersonal processes.

Myth and science duplicate each other in function but exclude each other in content. Myth and science are not like horses and cars. Once cars were invented, one could still use horses. By contrast, once science arose, which as the commonplace explanation of physical events goes back only a few hundred years, one could no longer use myth. The explanations given by both are mutually exclusive, for both are direct. In myth, gods act in place of scientific laws, not behind them, even if myth in fact presupposes physics to explain how a god creates rainfall, for example. While myth is as rational as science, it is false.

In the twentieth century Tylor's and Frazer's theories were spurned on many grounds: for setting myth against science and thereby precluding modern myths; for subsuming myth under religion and thereby precluding secular myths; for deeming the function of myth explanatory; for deeming myth false; and (half the time for Frazer) for reading myth literally. Postmodernism would dismiss myth as explanation and would treat myth as sheer story. Nevertheless, Tylor's and Frazer's theories remain central to the study of myth, and all other theories can be seen as rejoinders to them.

One response to Tylor and Frazer has been to take the function of myth as other than explanatory, in which case myth diverges from science and can therefore coexist with it. Another response to Tylor and Frazer has been to read myth other than literally, in which case myth does not even refer to the physical world and can therefore likewise coexist with science. The most radical response to Tylor and Frazer has been to alter both the explanatory function and the literal reading of myth.

Myth as More or Other than Explanatory in Function

The most influential reinterpreters of the function of myth have been Malinowski, Lévi-Strauss, and Eliade. It is not clear whether for Malinowski (1884–1942), the Polish-born anthropologist, moderns as well as 'primitives' have myth. It is clear that for him 'primitives' have science as well as myth, so that myth cannot be the primitive counterpart to modern science. For Malinowski, myth is no more the equivalent of applied science – Frazer's view – than the equivalent of

theoretical science – Tylor's view. Primitive peoples use science to explain and, more, to control the physical world. They use myth to reconcile themselves to aspects of the world that cannot be controlled.

Myth reconciles humans to the travails of life by rooting those travails in the primordial actions of gods or humans. Humans age because a god or human once did something that brought old age irremediably into the world: "The longed-for power of eternal youth and the faculty of rejuvenation which gives immunity from decay and age, have been lost by a small accident which it would have been in the power of a child and a woman to prevent" (Malinowski, 1926: p. 104). Myth pronounces the world not the best possible one but, in the wake of irreversible events, the only possible one. Still, the world becomes less capricious than it would otherwise be. In tracing back the origin of phenomena, myth does explain those phenomena, but as a means to an end rather than, as for Tylor, the end itself.

For Tylor and Frazer, myth deals with physical phenomena. For Malinowski, it deals equally with social phenomena like customs and laws. Myth still functions to reconcile humans to the unpleasantries of life, but now to unpleasantries that can be cast off by members of society. Myth spurs members to accept the impositions of society by conferring on them the clout of tradition. In the case of physical phenomena the beneficiary of myth is the individual. In the case of social phenomena it is society itself. If moderns do not have myths, their counterpart is ideology.

The most influential contemporary theorist of myth, the French anthropologist Claude Lévi-Strauss, at first glance seems a throwback to Tylor. For Lévi-Strauss, myth is an exclusively primitive enterprise yet a rigorously intellectual one. He denounces nonintellectualists like Malinowski as vigorously as Malinowski denounces intellectualists like Tylor and would doubtless denounce Levi-Strauss. Indeed, in declaring that primitive peoples, "moved by a need or a desire to understand the world around them, … proceed by intellectual means, exactly as a philosopher, or even to some extent a scientist, can and would do" (Lévi-Strauss, 1978: p. 16), Lévi-Strauss seems indistinguishable from Tylor. But Levi-Strauss in fact argues that primitive peoples think differently from moderns rather than, as for Tylor, think less rigorously than moderns.

Primitive, or mythic, thinking is concrete. Modern thinking is abstract. Primitive thinking focuses on the observable, sensory, qualitative aspects of phenomena rather than, like modern thinking, on the unobservable, nonsensory, quantitative aspects. Yet myth for Lévi-Strauss is no less scientific than modern science. It is simply part of the 'science of the concrete' rather than part of the science of the abstract, to which it is less at odds than askew myth is primitive science, but it is not inferior science.

Myth is an instance of thinking per se for both primitives and moderns because it classifies phenomena. According to Lévi-Strauss, all humans think in the form of classifications, specifically pairs of oppositions, and project them onto the world. Many cultural phenomena express oppositions. Myth is distinctive in tempering oppositions: "the purpose of myth is to provide a logical model capable of overcoming a contradiction" (Lévi-Strauss, 1958: p. 105). Myth resolves a contradiction either by providing a mediating or analogous middle term or by providing an analogous contradiction. Either tactic narrows and thereby alleviates the contradiction, but neither fully resolves it.

Like the contradictions expressed in other phenomena, those expressed in myth are apparently reducible to the fundamental contradiction between 'nature' and 'culture.' Humans experience themselves as at once animal-like and civilized. This split stems from the projection onto the world of the oppositional character of the mind. By diminishing that opposition, myth makes life more bearable and tempers a logical conundrum the payoff is intellectual. It is like that from solving a mathematical puzzle.

In calling his interpretation of myth 'structuralist,' Lévi-Strauss distinguishes it from 'narrative' interpretations, which adhere to the plot of myth. Lévi-Strauss dispenses with the plot and locates the meaning of myth in the structure. The plot is that element – say, event – a leads to event B, which leads to event C. The structure, which is identical with the expression and diminution of contradictions, is either that events A and B constitute an opposition mediated by event C or that events A and B are as opposed to each other as events C and D, an analogous opposition, are opposed.

Lévi-Strauss confines himself to primitive myths, but other structuralists analyze modern myths. In *Mythologies* (1972) the French semiotician Roland Barthes (1915–80) takes as myths various cultural artifacts and shows how they serve to justify the bourgeois outlook of postwar France. The function of myth here is not intellectual but ideological. Myth has nothing to do with natural science.

Where Lévi-Strauss largely analyzes myths independent of their social context – the grand exception is his analysis of the myth of Asdiwal – others, inspired by him have tied myths to their contexts. The most prominent have been the French classicists inspired by not only Lévi-Strauss but also the classicist Louis Gernet, after whom they are called the Gernet School. For these classicists – Jean-Pierre Vernant (1983), Marcel Detienne (1977), Pierre Vidal-Naquet, and Nicole Loraux – the relationship between myth and society is much more malleable and ironic than it is for Malinowski or even Barthes. Myth can as readily challenge as bolster existing ideology.

Unlike Malinowski, the Romanian-born historian of religions Mircea Eliade (1907–86) has no hesitation in making one function of myth explanatory. For him, as for Tylor, myth attributes events to the decisions of gods. Like Malinowski, Eliade includes social phenomena as well as physical ones. Explanation for Eliade is both an end in itself and a means to another end, which is magically to return to the time of the myth, the time of the origin of whatever phenomenon it explains. In this 'primordial time' gods are nearest to humans, as in the biblical case of "the Lord God['s] walking in the garden of the cool of the day" (Genesis 3:8). The return to primordial time reverses the everyday separation from gods, a separation that is equivalent to the fall, and is regenerative spiritually: "What is involved is, in short, a return to the original time, the therapeutic purpose of which is to begin life once again, a symbolic rebirth" (Eliade, 1968: p. 8). The ultimate benefit of myth is proximity to gods, one or more.

Eliade ventures beyond the other respondents to Tylor in proclaiming myth panhuman rather than merely primitive. Where for Malinowski science, but presumably not myth, is

universal, for Eliade myth but not science is. He cites modern novels, plays, and films with the mythic theme of yearning to escape from the everyday world into another, earlier one. If even professedly atheistic moderns thus have myths, then myth must be universal. The only difference between 'primitive,' or 'archaic,' peoples and moderns is that moderns do not know that they have myth.

How modern myths, which do not involve gods, provide entree to gods, Eliade never discloses. How the explanations given by modern myths, which are entirely about social phenomena like customs, inventions, and nations, are compatible with the explanations given by the social sciences, Eliade never discloses either. For him, the fact that moderns have both myth and science, social or natural, is proof that they are compatible. By contrast, Tylor and Frazer would argue that any preservation of traditional myths in the wake of science attests to sheer human inconsistency rather than to any lingering yearning for myth.

Myth as Other Than Literal in Meaning

The most prominent reinterpreters of not the function but the meaning of myth have been the German New Testament scholar Rudolf Bultmann (1884–1976) and the German-born philosopher Hans Jonas (1903–93). Both were students of the early Martin Heidegger and offer existentialist readings of myth. While they limit themselves to their specialties, Christianity and Gnosticism, they presuppose a theory of myth per se.

Bultmann acknowledges that, read literally, myth is about the physical world and is incompatible with science. But unlike Malinowski and Eliade as well as Tylor, Bultmann reads myth symbolically. In Bultmann's excruciatingly misleading phrase, one must 'demythologize' myth, by which he means not eliminating myth but on the contrary extricating its true, symbolic subject matter. Once demythologized, myth is no longer about the external world and is instead about the place of human beings in that world: "The real purpose of myth is not to present an objective picture of the world as it is, but to express man's understanding of himself in the world in which he lives." "Myth should be interpreted not cosmologically, but anthropologically, or better still, existentially" (Bultmann, 1953: p. 10).

Read literally, the New Testament describes a cosmic battle of good and evil anthropomorphic gods for control of the physical world. These gods intervene miraculously not only in the operation of nature but also in human lives. Humans are therefore not responsible for their actions. Demythologized, the New Testament still refers in part to the physical world, but now to a world ruled by a single, nonanthropomorphic, transcendent God. Because God does not act directly in the world and because no evil powers exist, humans are free and therefore responsible for their actions. Because a literal interpretation reduces humans to the pawns of larger, competing forces, it concentrates on those forces, which means on the world. Because a symbolic interpretation pronounces humans free, it concentrates on the actions they choose in response to the world.

Yet myth demythologized is not merely about humans, which would be myth 'psychologized.' It is about the human experience of the world. It is the depiction of a world view. For secular existentialists like Camus and Jean-Paul Sartre, the world experienced is godless and indifferent to humans. For Christians, according to Bultmann, the world experienced is safe and just. But the world can be comforting only insofar as God still intervenes in the world – for example, to save an ill child, which is Bultmann's own example. But then Bultmann's demythologizing proves inconsistent: God does not operate in the world in a way contrary to science. Otherwise moderns could not still have myth. Yet God does operate in ways that demonstrate his concern. Otherwise the world would be godless.

Demythologized, myth ceases to be purely primitive, as for Tylor, and becomes universal, as for Eliade. Myth ceases to be false, as for Tylor, and becomes true. Where Eliade invokes the existence of modern myths as *ipso facto* evidence of the compatibility of myth with science, Bultmann actually tries to reconcile myth with science, successfully or not. Where Eliade claims that moderns have myths of their own, Bultmann claims that moderns can retain biblical myths. He seeks to reconcile the Bible with science rather than, like Eliade, to uncover distinctively modern, seemingly secular myths.

Bultmann's boldest response to Tylor is to circumvent the function of myth. In translating the meaning of myth into terms acceptable to moderns, he sidesteps the issue of why moderns, even if they can have myth, need it. Unlike other symbolic interpreters of myth such as the religious philosopher Paul Ricoeur (1967) and the philosopher Philip Wheelwright (1968), Bultmann never maintains that the meaning of myth is untranslatable into nonmythic terms. Taking the meaning of myth from philosophy, he never asks why myth is needed.

Like Bultmann, Jonas seeks to show that ancient myths have an eternal meaning that speaks to moderns. For both Bultmann and Jonas, myth describes the alienation of humans from the world as well as from their true selves prior to their acceptance of God. Because Gnosticism, unlike mainstream Christianity, is radically dualistic, humans remain alienated from the physical world and from their bodies even after they have found the true God. Gnostics find the true God only by rejecting the false god of the physical world.

Unlike Bultmann, who wants to bridge the divide between Christianity and modernity, Jonas acknowledges the divide between Gnosticism and modernity. In Gnosticism the state of alienation is temporary. In modern, secular existentialism alienation is permanent. Alienation is the human condition, not a fall from it. What Jonas 'demythologizes' is not the source of alienation or the solution to it, for there is neither source nor solution, but the fact of alienation. He translates Gnostic myths into existentialist terms not to make Gnosticism palatable to moderns but only to show the similarity between the Gnostic and the existentialist outlooks: "the essence of existentialism is a certain dualism, an estrangement between man and the world …. There is only one situation … where that condition has been realized and lived out with all the vehemence of a cataclysmic event. That is the gnostic movement" (Jonas, 1963: p. 325). To parallel Gnosticism to modern existentialism, he must ignore the heart of Gnostic mythology – the elaborate cosmogony – and limit himself to the present state of the world. Like Bultmann, Jonas bypasses the function of myth and confines himself to the meaning.

Myth as Both Other than Explanatory and Other than Literal

The most radical departures from Tylor transform both the explanatory function and the literal meaning of myth. The most influential theorists here have been depth psychologists, Sigmund Freud (1856–1939) and C.G. Jung (1875–1961). For both, the subject matter of myth is the unconscious, and the function of myth is to express the unconscious.

The psychologizing of myth does not begin with Freud or with Jung. The recognition that myth involves the projection of human qualities onto gods goes back to at least the pre-Socratic philosopher Xenophanes, who, writing in the sixth century BC, observed that Ethiopians imagine their gods to be swarthy and flat-nosed, whereas Thracians visualize their gods as fair-haired and blue-eyed. Most famously, Xenophanes asserted that if lions, horses, and oxen had gods, their gods would look like them. Jung in particular is eager to trace a psychological understanding of myth all the way back to ancient Gnostics and to medieval alchemists. Going beyond – and against – Xenophanes, he claims that ancient myth tellers themselves often recognized the psychological meaning of their myths.

Nevertheless, for Freud and Jung alike, the key intellectual accomplishment of modernity has been the disentanglement of the psychological from the physical and the metaphysical – the disentanglement of the inner from the outer. Projections onto the outer world, which had taken the form of gods and of myths about gods, have largely been withdrawn. The outer world has come to be recognized as a natural rather than a supernatural domain, to be explained by impersonal scientific laws rather than by the decisions of gods recounted in myths. For both Freud and Jung, the rise of science has spelled the fall of religion and of religious myths.

Because the unconscious for Freud is composed of repressed sexual and aggressive drives, myth functions to release those drives in a disguised, vicarious way, so that the creator and the user of a myth need never confront its meaning and thereby their own nature. Myth, like other aspects of culture, serves simultaneously to reveal and to hide its unconscious contents. The classic psychoanalytic study of myth is that of Freud's one-time disciple Otto Rank (1884–1939). Focusing on myths of male heroes, Rank sees the myths as providing an unconscious, vicarious fulfillment of Oedipal drives. Myth serves neurotic adult males fixated at their Oedipal stage: "Myths are, therefore, created by adults, by means of retrograde childhood fantasies, the hero being credited with the myth-maker's personal infantile [i.e., childhood] history" (Rank, 1914: p. 82).

Contemporary psychoanalysts like Jacob Arlow take myth positively rather than negatively. For them, myth helps solve the problems of growing up rather than perpetuating them. Myth is progressive rather than regressive, and abets adjustment to society and the outer world rather than childish flight from both. Myth serves less to vent repressed drives than to sublimate them: "Accordingly, the myth can be studied from the point of view of its function in psychic integration – how it plays a role in warding off feelings of guilt and anxiety, how it constitutes a form of adaptation to reality and to the group in which the individual lives, and how it influences the crystallization of the individual identity and the formation of the superego" (Arlow, 1961: p. 375).

Jungians have taken myth positively from the outset. For them, the unconscious expressed in myth is not the Freudian repository of repressed, antisocial drives but a storehouse of innately unconscious 'archetypes,' or sides of the personality that have simply not had an opportunity at realization: "Contents of an archetypal character … do not refer to anything that is or has been conscious, but to something essentially unconscious" (Jung, 1968: p. 156). Myth is one means of encountering this Jungian, or 'collective,' unconscious. The function of myth is less release, as for classical Freudians, than growth, as for contemporary ones. But growth means less adjustment to the outer world, as for contemporary psychoanalysts, than cultivation of the 'inner world.' The payoff is self-realization. Some Jungians and Jungian-oriented theorists like the American Joseph Campbell (1904–87) so tout the payoff of myth that it becomes a panacea for all humanity's problems. Jung himself never goes this far. For Jung, myth works best as part of therapy. For Campbell, myth fends off the need for therapy.

Campbell (1949) differs most with Jung over the origin and function of myth. Where for Jung the archetypal contents of myth arise out of the unconscious, only in some works of Campbell's do they do so. Even then, sometimes for Campbell the unconscious is acquired rather than inherited – a view similar to freud's. At other times the contents of myth emerge from the imprint of either recurrent or traumatic experiences. In all of these cases each society creates its own myths – whatever the source of the material it uses. In contrast to Jung, Campbell is sometimes a diffusionist: myths originate in one society and spread elsewhere.

Campbell comes to declare repeatedly that myth serves four distinct functions: to instill and maintain a sense of awe and mystery before the world; to provide a symbolic image for the world, such as that of the Great Chain of Being; to maintain the social order by giving divine justification to social practices like the Indian caste system; and most of all to harmonize human beings with the cosmos, society, and the parts of themselves. Despite conspicuous differences, Campbell stands close to Jung and stands closest in '*The Hero with a Thousand Faces*' (1949), which remains the classic Jungian analysis of hero myths.

By far the most radical innovation in the Jungian theory of myth has been the development of 'archetypal' psychology, associated with James Hillman. Archetypal psychology faults classical Jungian psychology on multiple grounds. By emphasizing the compensatory, therapeutic message of mythology, classical Jungian psychology purportedly reduces mythology to psychology and gods to concepts. In espousing a unified self, or 'Self,' as the ideal psychological authority, Jungian psychology supposedly projects onto psychology a Western, monotheistic, Protestant outlook. The Western emphasis on progress is purportedly reflected in the primacy that Jungian psychology accords hero myths and in the primacy that it accords the ego, even in the ego's encounter with the unconscious: the encounter is intended to abet development. Finally, Jungian psychology is berated for placing archetypes in an unknowable realm, distinct from the knowable realm of symbols.

As a corrective, Hillman and his followers advocate that psychology be viewed as irreducibly mythological. Myth is still to be interpreted psychologically, but psychology is itself to be interpreted mythologically. One grasps the psychological meaning of the myth of Saturn, a favorite Roman god of

Hillman's, by imagining oneself to be the figure Saturn rather than by translating Saturn's plight into clinical terms like depression. At the same time the depressed Saturn represents a legitimate aspect of one's personality. Each god deserves its due. The psychological ideal should be pluralistic rather than monolithic – in mythological terms, polytheistic rather than monotheistic. Insisting that archetypes are to be found *in* symbols rather than outside them, Hillman espouses a relation to the gods in themselves and not to something beyond them. The ego becomes but one more archetype with its attendant kind of god, and it is the 'soul' rather than the ego that experiences the archetypes through myths. The payoff is a esthetic rather than moral: one gains a sense of wonder and contemplation, not a guide to living.

Recent Theories of Myth

The most notable recent theories of myth have been variations on myth as primitive science, theoretical and applied. Cognitivists led by the French anthropologist Pascal Boyer (1994) analyze the mental processes that shape thinking processes far more rigidly than Tylor's open-ended sequence of observation, analogy, and generalization. Cognitivists investigate the constraints on mythic explanations more than the explanations themselves. The German classicist Walter Burkert and the French literary critic René Girard have revived the theory that ties myth to ritual. Where for Frazer and his English follower Lord Raglan (1936) myth is the script for the ritualistic killing of the king, whose death and replacement ensure the rebirth of crops, for Burkert myth reinforces the ritual that commemorates the past hunting of animals. The function of myth is not physical but psychological and social: to cope with the guilt and anxiety that members of society feel toward their own aggression and to unite society by turning that aggression onto outsiders (see Burkert, 1996). Where for Frazer and Raglan the king is heroic because he is willing to die for the sake of the community, for Girard the hero, who can range from the most marginal person to royalty, is killed or sometimes exiled as a scapegoat for violence in society. Rather than directing the ritualistic killing, as for Frazer and Raglan, myth for Girard arises afterward to cover it up by making the victim first a criminal and then a hero. The function of myth is social: to preserve the ethos of sociability by hiding not only the killing but more deeply the violence endemic to society (see Girard, 1972).

Bringing Myth Back to the World

Where theorists of the nineteenth century assumed that myth could not be dislodged from the world and therefore could not be saved from science, theorists of the twentieth century saved myth from science either by removing myth altogether from the world or by removing it as an explanation of the world. The question for the twenty-first century is whether myth can be returned to the world – but in a way still compatible with science. The postmodern dismissal of the authority of science, evinced in the labeling of science as itself mythic, cheapens both myth and science. Twentieth-century theorizing commendably sought to accommodate myth to science rather than to spurn science in the name of myth.

One possible way of bringing myth back to the world is through the application to myth of D.W. Winnicott's notion of play. Winnicott did not himself analyze myth, but his analysis of children's play and of its continuation in adult make-believe provides one road back to the world. For Winnicott, play is *acknowledged* as other than reality: children grant that they are just playing. But play is no mere escapism. It involves the appropriation of reality for oneself. To pretend that a spoon is a train is to take a spoon and to turn it into a train. Far from confusing itself with reality, play demarcates the difference. Play grants itself the right to treat a spoon as a train, and a parent is barred from asking whether the spoon really is a train. Once play is over, the train is again a mere spoon.

As adult extensions of play, Winnicott names gardening and cooking and more deeply art and religion – in all of which one constructs a world out of elements from the outer world: "It is assumed here that the task of reality-acceptance is never completed, that no human being is free from the strain of relating inner and outer reality, and that relief from this strain is provided by an intermediate area of experience which is not challenged (arts, religion, etc.). This intermediate area is in direct continuity with the play area of the small child who is 'lost' in play" (Winnicott, 1982: p. 13).

Myth, or at least some myths, may be considered as adult make-believe (see Segal, 2003: pp. 154–157). For example, in films, film stars, like gods, can do all kinds of things that celebrities who perform in person cannot. And in a movie theater they are gargantuan in size. In all of these ways film stars fit the popular conception of gods, which is one of degree rather than of kind. God, like humans, has a body, but it is bigger. On screen, merely human virtues get magnified into superhuman ones: bravery becomes fearlessness, kindliness becomes saintliness, strength becomes omnipotence, wisdom becomes omniscience. Film stars are 'idolized' and 'worshiped,' and the greatest are called 'gods.' As 'stars,' they shine brightly in a heaven far above us.

It might understandably be argued that even if film stars are godlike in their attributes, worshipping them still does not restore divinity to the physical world. Yet stars who take responsibility for accomplishing things in the social and even the physical world are often more successful than whole nations, such as in canceling Third World debt, responding to famine, and saving the environment.

Against my argument that film stars are the modern, secular version of gods, it might sensibly be observed that these days nobody believes the hype. But this hard-nosed view of present-day fans is the naive one. Fans continue to 'idolize' and 'worship' stars not in ignorance of their flaws but in defiance of them. The flaws are either denied or discounted. But the devotion of fans is not mindless. It is done knowingly. It is, following Winnicott, make-believe not credulity. The myths of movie stars are their films and their idealized biographies. Taking those biographies as make-believe reconnects myths to gods, therefore to religion, and therefore to the world.

See also: Freud, Sigmund (1856–1939); Jung, Carl Gustav (1875–1961); Levi-Strauss, Claude (1908–2009); Religion, Anthropology of; Religion, Psychology of; Science and Religion; Scientific Explanation.

Bibliography

Arlow, J.A., 1961. Ego psychology and the study of mythology. Journal of the American Psychoanalytic Association 9, 371–393.

Barthes, R., 1972. Mythologies (A. Lavers, Trans.). Hill & Wang, New York; Cape, London.

Boyer, P., 1994. The Naturalness of Religious Ideas. University of California Press, Berkeley.

Bultmann, R., 1953. New testament and mythology (1941) (R.H. Fuller, Trans.). In: Bartsch, H.-W. (Ed.), Kerygma and Myth, vol. 1. SPCK, London.

Burkert, W., 1996. Creation of the Sacred. Harvard University Press, Cambridge, MA.

Campbell, J., 1949. The Hero with a Thousand Faces, first ed. Pantheon Books, New York.

Chase, R., 1949. Quest for Myth. Louisiana State University Press, Baton Rouge.

Detienne, M., 1977. The Gardens of Dionysus (J. Lloyd, Trans.). Harvester Press, Sussex.

Eliade, M., 1968. The Sacred and the Profane (W.R. Trask, Trans.). Harvest Books, New York.

Feldman, B., Richardson, R.D., 1972. The Rise of Modern Mythology, 1680–1860. Indiana University Press, Bloomington.

Frazer, J.F., 1911–15. The Golden Bough, third ed., 12 vols. Macmillan, London.

Girard, R., 1972. Violence and the Sacred (P. Gregory, Trans.). Athlone Press, London; Johns Hopkins University Press, Baltimore.

Hillman, J., 1975. Re-visioning Psychology. Harper & Row, New York.

Jonas, H., 1963. The Gnostic Religion, second ed. Beacon Press, Boston.

Jung, C.G., 1968. The Archetypes and the Collective Unconscious, second ed. Princeton University Press, Princeton, NJ.

Lévi-Strauss, C., 1958. The structural study of myth (1955). In: Sebeok, T.A. (Ed.), Myth. Indiana University Press, Bloomington (Reprinted in Lévi-Strauss, 1963. Structural Anthropology (C. Jacobson, B.G. Schoepf, Trans.). Basic Books, New York).

Lévi-Strauss, C., 1978. Myth and Meaning. University of Toronto Press, Toronto.

Malinowski, B., 1926. Myth in Primitive Psychology. Kegan Paul, London; Norton, New York (Reprinted in 1948. Magic, Science and Religion and Other Essays. Free Press, Glencoe, IL).

Raglan, Lord, 1936. The Hero. Methuen, London (Reprinted in Rank, O., et al., 1990. In Quest of the Hero. Princeton University Press, Princeton, NJ).

Rank, O., 1914. The Myth of the Birth of the Hero (F. Robbins, S.E. Jelliffe, Trans.), first ed. Journal of Nervous and Mental Disease Publishing (Reprinted in Rank, O., et al., 1990. In Quest of the Hero. Princeton University Press, Princeton, NJ) (2004 second, 1922 German edition (G.C. Richter, E.J. Lieberman, Trans.). Johns Hopkins University Press, Baltimore).

Ricoeur, P., 1967. The Symbolism of Evil (E. Buchanan, Trans.). Harper & Row, New York.

Segal, R.A., 1992. Explaining and Interpreting Religion. Peter Lang, New York.

Segal, R.A., 1999. Theorizing about Myth. University of Massachusetts Press, Amherst.

Segal, R.A., 2003. Psychoanalyzing myth: from Freud to Winnicott. In: Jonte-Pace, D. (Ed.), Teaching Freud. Oxford University Press, Oxford and New York.

Segal, R.A., 2004. Myth: A Very Short Introduction. Oxford University Press, Oxford and New York.

Tylor, E.B., 1871. Primitive Culture, first ed., 2 vols. Murray, London.

Vernant, J.P., 1983. Myth and Thought among the Greeks. Routledge & Kegan Paul, London.

de Vries, J., 1961. Forschungeschichte der Mythologie. Alber, Freiburg.

Wheelwright, P., 1968. The Burning Fountain, rev. ed. Indiana University Press, Bloomington.

Winnicott, D.W., 1982. Playing and Reality. Routledge, London and New York.

Myths and Symbols: Organizational

Pasquale Gagliardi, Catholic University, Milano, Italy; and Fondazione Giorgio Cini, Venezia, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In what follows, the essential features of 'organizational symbolism' are first described, since this has been the intellectual context within which the constructs analyzed have largely been developed. As this distinctive stream of inquiry, within the field of organization studies, focuses on the expressive – as opposed to the instrumental, pragmatic – dimension of organizational life, the author analyzes the nature of (and the tensions implied by) the interrelations between productive and symbolic practices. Recent developments and methodological innovations in the field are also discussed. The article describes then the main traits and functions of different organizational symbols – language and linguistic constructs (myths), (ritual) behaviors, and objects/artifacts. Special attention is paid to the role of saga – defined as a collection of myths – in the process of institutionalization and in shaping organizational identity. Finally, the origin and function of mediatory myths in managing the dialectic relation between organizational ideology and operational needs are highlighted.

Organizational Symbolism

The constructs 'organizational myth' and 'organizational symbol' first made their appearance in organizational and managerial studies in the second half of the 1970s, at a time of renewed and rapidly spreading interest in the cultural study of organizations. This interest was manifest in two main currents of study: the interorganizational perspective of 'new institutionalism' (Powell and DiMaggio, 1991), which explored the ways in which cultural values shape organizational forms – and induce the organizations that populate a particular 'field' (a territory, an industrial sector, or a distinct sphere of institutional life) to grow similar rather than dissimilar – and the microperspective of 'organizational symbolism' (Dandridge et al., 1980; Turner, 1990), which viewed individual organizations as 'cultures' characterized by distinct paradigms, to be then analyzed not only in their instrumental and economic aspects but also in their ideational and symbolic ones, using holistic and interpretative research models.

Born at the end of the 1970s as a marginal and nonconformist phenomenon, organizational symbolism rapidly grew into one of the main currents of thought in organizational studies. It produced an extraordinary flourishing of studies and research, gaining unusual popularity even outside the academic sphere. Most notable among the numerous surveys of this body of literature are the following: Alvesson and Berg (1992), a particularly thorough treatment which reviews the various theoretical perspectives that have developed internally to the movement, and discusses the relevance of symbolism to managerial practice; the study by Trice and Beyer (1993), which provides a detailed picture of the material set out in numerous empirical research studies; and the work of Jones (1996), which gives an overview of the main methodological issues connected with the study of organizational symbolism.

Although the approach employed by the movement's exponents was often presented as radically innovative, it was in fact deeply rooted in the social sciences. Despite their frequent use of anthropological terms and concepts, scholars of organizational cultures more frequently acknowledge their intellectual indebtedness to certain currents of sociological thought – in particular to phenomenological sociology, symbolic interactionism, ethnomethodology, and above all to the great ethnographic tradition of urban and occupational sociology (e.g., Whyte, 1955; Roy, 1961). However, although an approach inspired by the same theoretical ideas had already been used in the study of formal organizations (see, in particular, Selznick, 1949; Gouldner, 1954; Crozier, 1963), until the early 1970s organizational and managerial studies were indubitably dominated by a rationalist and positivist paradigm which viewed the analysis of relations among objective and measurable variables as the central concern of research and substantially neglected the notion of culture. Ironically, many of these analyses were conducted in the name of Max Weber, regarded as the founding father of organizational theory, but they ignored Weber's invitation to see organizations as the expression of cultural values, interpreting his bureaucratic model instead as the organizational model *par excellence* and with no alternatives (Clegg, 1995). In one of the first essays devoted expressly to organizational symbolism, Dandridge et al. (1980) wrote that: "… a survey of major texts within the field of organizational behavior establishes clearly that there are virtually *no* references to the phenomenon that is the subject of this paper … For all practical purposes … it is as if the phenomenon did not exist or was not important." In effect, in those years the statement that organizations have their own myths and symbols would have seemed at most a paradox, almost an oxymoron, to the extent that formal organizations – especially economic ones – were regarded as utilitarian forms of social aggregation in which individual and collective behaviors were governed mainly by norms of instrumental rationality (*see* Organization: Overview). If modern organizations are, by definition, the domain of a legal-rational ethos, and characterized by pragmatic behaviors shaped by the practical goals to which they are directed, how can these organizations possibly harbor expressive and disinterested behaviors shaped by impulses and emotions? And if these forms of behavior do exist, how can they be anything but secondary, irrelevant, or at any rate dysfunctional and therefore to be eliminated? Organizational symbolism challenges these prevalent opinions and invites us to read organizations in a completely different light. As Turner (1990) observes, "… the current growth of interest in organizational symbolism and corporate culture points to the end of an era. Looking anew

at the organizational world we can see that it is a sensual and emotional realm, replete with its own ceremonies, rites and dramas."

The cultural approach asserts that every productive practice is determined equally by the practical requirements from which it springs and by particular visions of the world. No productive practice is only and exclusively such, for it is simultaneously a symbolic practice, a way to appropriate reality by imposing one's own specific view of what reality is. Accordingly, scholars of organizational cultures have explored the interweaving of technical requirements and expressive needs, the ritualistic significance of apparently rigorously pragmatic behaviors and processes, and the overlapping of ideological prejudice and mythic knowledge with scientific expertise. In other words, they have shown that culture and utility are woven together in even the most strictly utilitarian forms of social aggregation. From this perspective, therefore, organizations – and their environments – are conceived as 'symbolic fields': that is, as intersubjectively negotiated systems of meaning recognizable in a coherent set of symbols or conventional representations of important and complex aspects of organizational and social reality. Decoding an organization's symbolic field involves analysis of its answers to such fundamental questions as, for example, the reliability or unreliability of human beings, the symmetry or asymmetry of social relations, the type of relationship – dominance, harmony, or subordination – with the external environment, and the nature of reality and of truth.

In the early 2000s, the study of organizational symbolism and corporate cultures gradually lost the popularity and central role it had built up in the previous 20 years. At first glance, intellectual movement seems to have become marginal to organizational studies and to have run out of innovative steam. Yet, on closer examination, it has not disappeared but has gradually been replaced by (1) currents of parallel studies that came into being and thrived in a general climate of epistemological renewal encouraged by the symbolist movement; and (2) thematic studies that grew as offshoots from the mature branch of studies focused on organizations as cultures.

Among the parallel movements that organizational symbolism contributed to legitimize it is worth mentioning the narrative approach (Czarniawska and Gagliardi, 2007), which emphasized the importance of the stories and myths produced in organizations and explored the analogy – proposed by Ricoeur (1981) – between text (narrating) and action (organizing): "The raw material of everyday organizational life consists of disconnected fragments, physical and verbal actions that do not make sense when reported with simple chronology. Narrating is organizing this raw and fragmented material with the help of such devices as plot and characters. Simultaneously, organizing makes narration possible, because it orders people, things and events in time and place." (Czarniawska and Gagliardi, 2007, p. VII).

The most striking of the thematic studies that developed within the symbolist movement is so-called organizational aesthetics. This approach sets out to study the role of sensory experience in the daily life of organizations; that is (1) (tacit) aesthetic knowledge as opposed to intellectual knowledge; (2) disinterested, 'expressive' action shaped by the emotions, as opposed to 'impressive' action that can be recognized in terms of the practical aims that it focuses on; and, finally (3) aesthetic communication, seen as the set of modes whereby sensory experience – by nature ineffable – is conveyed and shared. The development of this current of thought has been enhanced by the study of artifacts, one of the three fundamental symbolic systems that can be used – as we see in the next section – to express the beliefs and values of a culture. Although some remarks on the importance of the aesthetic dimension of organizations were conducted slightly earlier (e.g., Ramirez, 1991), there has been a growing production of literature on the subject since the late 1990s (Strati, 1999; Linstead and Hoepfl, 2000; Guillet de Monthoux, 2004; Gagliardi, 2006). Its aim is to explore the methodological implications of a specifically 'aesthetic' approach to the study of organizations, as a distinct research field. Organizational aesthetics is in fact solidly rooted in the cultural-symbolist movement without, however, sharing its predominant cognitivist stance (Gagliardi, 2006).

Organizational Symbols: Types and Functions

The definition of 'symbol' most widely used by scholars of organizational cultures is the one provided by Cohen (1976): "Symbols are objects, acts, relationships or linguistic formations that stand ambiguously for a multiplicity of meanings, evoke emotions, and impel men to action." Organizational symbols normally are grouped into three categories corresponding to the three main systems for expressing beliefs and cultural values within an organization: verbal language, behavioral language, and the language of objects.

Verbal Symbols

Principal among verbal symbols is language itself, with its ability to categorize and articulate experience using specific 'jargons,' which may differ even among organizations operating in the same context and speaking the same language. The category of verbal symbols also comprises linguistic artifacts like legends, stories, slogans, the names given to places, persons, events, and myths. The latter can be defined as dramatized accounts of events whose veracity is asserted as dogma, or is taken for granted, and which have the effect of legitimating and making desirable the behavior enacted in the events narrated and the ideas from which such behavior springs. To the extent that the cadence of organizational life is set by the alternation of crucial experiences – which jeopardize its identity and sometimes its very survival – with periods of relative stability, myths are often the idealization of the specific ways in which those difficulties have been overcome, and they shape the routines that orient behavior during periods of stability. Consequently, the myths most frequently encountered in organizations relate to their birth (myth of the origin) and to the transformations induced by traumatic events (the death of the founder, for example, or a slump in the market). In all organizations, the beliefs and values that myths embody exert a pervasive influence on decisions (*see* Organizational Decision Making), and it is difficult not to discern an element of 'mythic knowledge' even in decisions apparently taken solely on technical–scientific grounds. However, it has been pointed out that the more decisions are operational, and the effects of actions measurable, the more decisions are taken following

principles of instrumental rationality, and the broader the range of options and viable choices available. By contrast, the more choices are general and tied to the base values on which the organization grounds its identity, the more restricted the range of options, and the more mythical, prescientific, and unamenable to the scrutiny of reason and experience becomes the knowledge utilized (Gagliardi, 1986).

Symbolic Actions

Of particular importance among the symbolic manifestations represented by acts and styles of behavior are rites and ceremonials as emotionally charged collective actions – usually performed in a rigorously prescribed sequence – by means of which an organizational community celebrates its successes, heroes, and organizational ideals. Various classifications of organizational rites – which assume a wide variety of connotations – have been proposed, as well as analysis of their manifest and covert cultural consequences (Trice and Beyer, 1984). Among the most frequently observable rites are:

1. rites of passage (e.g., training seminars as a prelude to a promotion), which ease the transition of individuals to new roles and status;
2. rites of enhancement (like the awarding of bonuses for outstanding performance), which consolidate the social identities of the rewarded and induce others to emulate them;
3. rites of degradation (like the replacement of senior managers for some reason deemed unworthy), which solemnly and publicly reaffirm the importance of the social roles compromised by the behavior of the persons degraded;
4. rites of integration (like the celebrations organized by businesses on special occasions like Christmas or the anniversary of the company's foundation), which encourage shared feelings of equality and participation in a common enterprise, temporarily suspending the norms which sanction differences in power and status, but implicitly reaffirm the adequacy of those rules in day-to-day life;
5. rites of renewal (like the periodic drawing up of strategic plans), which reassure the organization's members that it is by keeping up with the times that present problems will be overcome, but at the same time bolster the legitimacy of existing systems of power and authority.

Artifacts and Aesthetic Knowledge

A third system of symbols is constituted by material artifacts or the intentional products of human action which exist in the organization independently of their creator and which can be perceived by the senses: products, images, buildings, furnishings, arrangements of physical space. Given the durability of physical matter, artifacts are able tenaciously and incessantly to transmit particular messages and cultural stimuli, thereby encouraging the diffusion and sharing of special 'modes of feeling' in ways that are all the more efficient because they evade intellectual control (Gagliardi, 1990).

The advent of organizational aesthetics has given/received a great boost to/from the study of corporate artifacts as vehicles of sensory knowledge and what is defined as 'corporate pathos'; that is, the specific 'way of feeling' cultivated and transmitted in the culture of each organization. Artifacts include a category that has recently been the focus of considerable attention: images. This highlights the increasingly important role of visual methods in organizational research (Warren, 2009), a trend in line with the development of "a mass-mediated, image-saturated society where aesthetic appeal is an increasingly important arbiter of economic value" (Warren, 2009, p. 566). Images are above all visual but we can also speak of auditory, tactile, and olfactory images in order to refer to distinctive combinations of sensible experiences recognizable through senses other than sight. The heuristic potential of images – in this wider definition – derives from their capacity to elude the restraints of descriptive language (which only allows us to say one thing at a time) and enables the use of a 'presentational' language (which allows us to say several, even contradictory things simultaneously, thus conveying knowledge by acquaintance, which eliminates the filter of abstraction) (Bruner, 1962; Langer, 1969).

The growing importance attached to the study of artifacts, to the structuring of space and the sensory experiences that they encourage or hinder has led to deep innovations in ethnographic practice. It has also meant breaking the monopoly of traditional fieldwork techniques and has highlighted the indispensable role of the researcher's 'sensuality' as an epistemological disposition and vital tool in understanding reality (Fine, 1996; Stoller, 1997).

The Role of the Saga in the Process of Institutionalization

The organizational myths of origin and transformation are often blended together in a 'saga' or an account of an 'extraordinary' series of events which traverses and unifies the various phases of the organization's evolution. In a saga, historical facts are embellished by subsequent re-elaborations, inconsistencies are eliminated or justified, and the rational explanation of how certain means lead to certain ends acquires an affective connotation which transforms the workplace into a loveable and loved institution. The saga therefore performs an essential role in the process of institutionalization: that is, in the process by which the organization, from a sterile mechanical apparatus for the efficient distribution and coordination of tasks, becomes imbued with values and acquires a distinctive 'character'. This character on the one hand underpins the collective identity and sentiment of belonging of the organization's members, while on the other it legitimates the organization in the social environment (Selznick, 1957). "Whether developed primarily by management or by employees, the story helps rationalize for the individual his commitment of time and energy for years, perhaps for a life-time, to a particular enterprise. Even when weak, the belief can compensate in part for the loss of meaning in much modern work, giving some drama and some cultural identity to one's otherwise entirely instrumental efforts. At the other end of the continuum, a saga engages one so intensely as to make his immediate place overwhelmingly valuable. It can even produce a distortion, with the organization becoming the only reality, the outside world becoming illusion" (Clark, 1972).

To the extent that the saga generates a shared conviction that the organization is unique because it has been the protagonist

of an unusual enterprise crowned by extraordinary success – a conviction which induces individuals to remain within the system, endeavoring to improve it rather than pursuing their interests elsewhere – the saga may constitute a precious organizational resource. Indeed, ensuring the stable participation and constant commitment of its members is one of the crucial problems faced by any co-operative system constructed to pursue specific ends.

Mediatory Myths

In that, as mentioned in the first section, the dialectical relation between expressive needs and pragmatic exigencies unavoidably conditions the lives of organizations; a further crucial organizational problem is the management of the contradictions, tensions, and psychological costs that derive from that relation: mediatory myths are discursive formulations which make contradictions acceptable.

Unlike the myths of origin and transformation, which mainly orient actions to be undertaken, mediatory myths provide retrospective justification for previous actions and states of affairs. The need to justify concrete action arises, first, from the fact that different values may conflict in specific situations – and it is difficult to comply with one principle without being in breach of another – and second from the fact that actions always tend, actually or apparently, to fall short of the ideals that have inspired them, so that there is invariably a gap between 'what we should do' and 'what we are able to do,' between the purity of the ideal and the impurity of the concrete organizational practice. Mediatory myths harmonize conflicting principles, not by means of logical arguments but by means of assertions of fact, and they resolve or reconcile contradictions by simply denying and inverting them, or by holding irreconcilable alternatives in suspension. Abravanel (1983) reports an interesting example of a mediatory myth: that of the 'wild ducks' institutionalized at IBM when T.J. Watson Jr. was chief executive officer. The contradiction between the ideology encouraged by Watson, namely that the company's employees were free to express their individuality, and a corporate reality characterized by rigid training and socialization practices was resolved by institutionalizing the myth that the employees were like wild ducks: they could not be tamed but nevertheless flew in formation and did not choose their migratory route.

See also: Corporate Culture; Information and Knowledge, Organizational; Innovation; Intelligence, Prior Knowledge, and Learning; Leadership in Organizations, Sociology of; Learning: Organizational; Organizational Behavior, Psychology of; Organizational Decision Making; Organizations, Sociology of; Rational Choice and Organization Theory.

Bibliography

Abravanel, Harry, 1983. Mediatory myths in the service of organizational ideology. In: Pondy, Louis R., Frost, Peter J., Morgan, Gareth, Dandridge, Thomas C. (Eds.), Organizational Symbolism. JAI Press, Greenwich, CT, pp. 273–293.
Alvesson, Mats, Berg, Per-Olof, 1992. Corporate Culture and Organizational Symbolism: An Overview. Walter de Gruyter, New York.
Bruner, Jerome S., 1962. Essays for the Left Hand. Harvard University Press, Cambridge, MA.
Clark, Burton R., 1972. The organizational saga in higher education. Administrative Science 17, 178–184.
Clegg, Stewart R., 1995. Of values and occasional irony: Max Weber in the context of the sociology of organizations. In: Bacharach, Samuel B., Gagliardi, Pasquale, Mundell, Bryan (Eds.), Studies of Organizations in the European Tradition. JAI Press, Greenwich, CT, pp. 1–46.
Cohen, Abner, 1976. Two Dimensional Man: An Essay on the Anthropology of Power and Symbolism in Complex Society. University of California Press, Berkeley, CA.
Crozier, Michel, 1963. Le phénomène bureaucratique. Editions du Seuil, Paris (1964. The Bureaucratic Phenomenon. The University of Chicago Press, Chicago).
Czarniawska, Barbara, Gagliardi, Pasquale (Eds.), 2007. Narratives We Organize by. John Benjamins Publishing Company, Amsterdam/Philadelphia.
Dandridge, Thomas C., Mitroff, Ian I., Joyce, William F., 1980. Organizational symbolism: a topic to expand organizational analysis. Academy of Management Review 1, 77–82.
Fine, Gary A., 1996. Kitchens. The Culture of Restaurant Work. University of California Press, Berkeley, CA.
Gagliardi, Pasquale, 1986. The creation and change of organizational cultures: a conceptual framework. Organization Studies 2, 117–134.
Gagliardi, Pasquale (Ed.), 1990. Symbols and Artifacts: Views of the Corporate Landscape. Walter de Gruyter, New York.
Gagliardi, Pasquale, 2006. Exploring the aesthetic side of organizational life. In: Clegg, Stewart R., Cynthia, Hardy, Lawrence, Thomas B., Nord, Walter R. (Eds.), The Sage Handbook of Organizations Studies, second ed. Sage, London, pp. 701–724.
Gouldner, Alvin W., 1954. Patterns of Industrial Bureaucracy. Free Press, New York.
Guillet de Monthoux, Pierre, 2004. The Art Firm. Aesthetic Management and Metaphysical Marketing. Stanford University Press, Palo Alto, CA.
Jones, Michael O., 1996. Studying Organizational Symbolism: What, How, Why? Sage, Thousand Oaks, CA.
Langer, Susanne K., 1969. Philosophy in a New Key. A Study in the Symbolism of Reason, Rite, and Art. Harvard University Press, Cambridge, MA.
Linstead, Stephen, Hoepfl, Heather, 2000. The Aesthetics of Organization. Sage, London.
Powell, Walter W., DiMaggio, Paul J. (Eds.), 1991. New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago.
Ramirez, Rafael, 1991. The Beauty of Social Organization. Accedo, Munich.
Ricoeur, Paul, 1981. Hermeneutic & the Human Sciences. Cambridge University Press, Cambridge.
Roy, Donald. F., 1961. Banana time: job satisfaction and informal interactions. Human Organization 18, 158–168.
Selznick, Philip, 1949. TVA and the Grass Roots. University of California Press, Berkeley, CA.
Selznick, Philip, 1957. Leadership in Administration. Harper & Row, Evanston, IL.
Stoller, Paul, 1997. Sensuous Scholarship. University of Pennsylvania Press, Philadelphia.
Strati, Antonio, 1999. Organization and Aesthetics. Sage, London.
Trice, Harrison M., Beyer, Janice M., 1984. Studying organizational cultures through rites and ceremonials. Academy of Management Review 4, 653–669.
Trice, Harrison M., Beyer, Janice M., 1993. The Cultures of Work Organizations. Prentice Hall, Englewood Cliffs, NJ.
Turner, Barry A., 1990. The rise of organizational symbolism. In: Hassard, John, Pym, Denis (Eds.), The Theory and Philosophy of Organization. Critical Issues and New Perspectives. Routledge, London.
Warren, Samantha, 2009. Visual methods in organizational research. In: Bryman, Alan, Buchanan, David (Eds.), Handbook of Organizational Research Methods. Sage, London, pp. 566–582.
Whyte, William. F., 1955. Street Corner Society: The Social Structure of an Italian Slum. University of Chicago Press, Chicago.

N

Names and Naming

João de Pina-Cabral, University of Kent, Canterbury, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The category 'personal name' describes a centrally important aspect of all known forms of human life, even if the actual way such names are attributed and used has varied immensely from region to region and across time. People start working on their names as they enter language and sociality. Therefore, our names carry within them specific, localized histories of personhood and belonging, both collective and singular. This article focuses on the ways in which names are functionally diverse, how they reflect our collective belongings, how they position us by relation to other people, namely in terms of hierarchic relations, and how they alter during the person's life course.

From an anthropological, comparative point of view it can be securely stated that the analytic category 'personal name' is indispensible. It describes a centrally important aspect of all of the different styles of human existence that anthropologists and historians have recorded over the years. While specific processes of personal naming vary widely from context to context and across time, the near omnipresence of modes of naming personal singularity cannot be doubted. Human ways of life are both social *and* personal, and naming is one of the central aspects through which the two imply each other (Rorty, 1969).

Personhood is the mold within which humans enter humanity – that is, it is the central pathway to reflexive thinking and, therefore, to human sociality (Mauss, 1985[1938]). Persons are called into personhood by other persons with whom they are in close contact (Trevarthen, 1998), and there are no human forms of collective experience that do not rely on the agency of persons who are aware of themselves as somehow singular – as Godfrey Lienhardt put it, "one can lay too much one-sided stress on the collectivist orientation of African ideas of the person. (…) the recognition of the importance of an inner, mysterious *individual* activity, comparable to what is meant by speaking in English of 'what goes on inside' a person is attested by many proverbs" (Lienhardt, 1985: 145).

Self-building is both a condition for language and a result of language. People start working on their own personhood as they enter language. Thus, the diverse modes in which languages differentiate subject positions affect people's awareness of their own personhood. Personal naming cannot be dissociated from the way in which languages use pronouns (Benveniste, 1966: 251–257) and conjugate verbs, as Socrates pointed out so long ago (cf Ademollo, 2011). While there may be occasions in which singular persons remain unnamed, and while there are contexts in which personhood is subordinated to collectivity, the dissolution of personal singularity cannot ever be absolute, as that would imply a step out of language. This means that personal naming is a central aspect of human sociality wherever the latter may be found.

Furthermore, naming systems reflect the fact that humans are called into humanity by other humans in an always previously inhabited world: alterity is always anterior to identity and between the two there is no symmetry (Lévinas, 1996). This means that plurality is recurrent in human life (Marriott, 1976); human singularity is never stable, never complete, and it emerges within intersubjectivity out of the contagion and contrast of different anterior singularities, as Freudian thinkers have long identified (Butler, 2012; Winnicott, 1971).

Naming systems reflect such processes in many, historically differentiated ways. Our names carry within them specific, localized histories of personhood, both collective and personal, but always anterior to the emergence of each one of us as persons. Thus, our names are seldom unique – as they would be if we were named by a conjugation of numbers in an infinite series. As it happens, reificatory modes of naming of this nature have been tried out, but they have been experienced always as being repressive. They work against the person's immersion in anterior pluralities; by negating the person's history, they negate the person.

Naming is, in this way, a deeply ambiguous phenomenon, since we both accept and acknowledge our names. Naming is fully part of the intersubjective experience and is acquired in an interpersonal way. Fully private names (those of which only the named would be cognizant) must be left out of the broad analytic category of personal names.

Since coevals are named in broadly similar ways, they tend to share more or less similar paths of ontogenesis (histories of personal constitution – Toren, 2002). It is because our names echo other people's that they are recognizable as meaningful human names. People whose names somehow fail to conform to the norm, or are incomplete, or absent in significant ways,

are prone to be pulled out of ordinary personhood, even to experience stigma. Naming traditions play on differentiation and identification to create naming systems. Moreover, systematicity in naming may result from denotational as well as connotational processes (Lévi-Strauss, 1966). Most frequently, naming systems take recourse to a conjugation of both. Very seldom are personal names so utterly meaningless as to lack any reference whatsoever to aspects of the world that are seen to complement the person's personhood.

The list of possible personal naming systems must remain open-ended, as the history of human creativity has not yet finished. There are, however, a number of forms of constituting systematicity in personal naming that are very recurrent in the anthropological and historical record. In the sections that follow, I will attempt to outline some of the most common of these.

Functional Diversity

Types of names differ depending on the purposes for which they have been conceived and the manner in which they are supposed to be used, so that it is common for people to be named differently in different contexts, depending on circumstances (either through utterly different names or through different variations of what is considered to be 'the same name' – e.g., as in Jack and John, Sue and Suzanne, etc.). In the great Eurasian region, as in most other social contexts where an administrative apparatus is present, one frequently encounters naming systems that are centrally structured by a major distinction between proximity naming and formal naming (that is, the name that the State condones). Often, the latter is treated as having greater externality, truthfulness, or objectivity (Pina-Cabral, João de, 2010a). Thus, typically, formal names are seen as being more immune from change as opposed to 'nicknames,' 'home names,' 'little names,' etc., which are seen as mutable and prone to varying depending on the interlocutors' particular relation (Segalen, 1980; Manning, 1974).

Hierarchy and kinship play here a central role and are often reflected in forms of name declination: hypocoristics, diminutives, and suchlike. Similarly, *noms de guerre* or professional names are also very commonly encountered. The usual example here is the practice of modern armies to choose a simplified version of each soldier's formal name so that (1) soldiers' names are simple and (2) there are no two soldiers in a battalion called in the same way. In Iberian societies (and in others such as Morocco or Russia) where official names are often perceived as being too long or too formal for daily use, adults are prone to develop *noms de guerre* that they use in their more common daily encounters. There is continuity between this practice and that of taking recourse to functionally specific pseudonyms: that is, names that are used for specific purposes, elaborating a dramatized fiction of split personhood (that this is treated as a fiction remains implicit in the 'pseud-').

Wherever Western European modes of naming have been adopted around the world, there is a tendency to believe that naming has the primary function of individuating the person. This theory, however, can hardly withstand analytic inquiry (Pina-Cabral, João de, 2012). Plurality and singularity are constantly at play in personhood, but plurality is always anterior and singularity never fully achieved.

Collective Naming

Persons often take on a combination of names some of which are more singularizing while others are specifically geared at reflecting collective belonging. In the broad mass of Eurasian societies, it is very common for personal names to be differentiated between proper name and surname (so as to simplify, I use here the English form of the distinction, but it must be noted that the meaning of the terms used to describe the two kinds of names varies immensely, with significant implications, even within Western Europe). In fact, this broad differentiation is far less rigid and homogeneous than it is usually thought to be: surnames can be of many kinds; proper names are formed in many different ways; and between the two there can exist third type of names (as, for example, the interesting English category of 'middle names,' or Russian patronyms, or still Basque household names) (for collections emphasizing naming diversity, see Tooker and Conklin, 1984 or vom Bruck and Bodenhorn, 2006). Surnames as we use them today in our passports are a relatively recent phenomenon and ones that are still not universally applied, although it would seem that the demands of modern travel are quickly changing that. In Western Europe, surname systems differ profoundly, namely between those that are strictly patrilineal (as in the United Kingdom or France – and this is something that is also quickly changing due to feminist activism) and those that are loosely bilateral (as in the regions where Iberian languages are spoken).

People can take on names that reflect house belonging, familial belonging, descent group belonging, village belonging, local group belonging, ethnic group belonging, etc. The relation between such collective naming practices and State-condoned formal names can also differ profoundly from context to context defying generalization. Things like clan names and lineage names can mix with and approximate titles, particularly in elite contexts. Names often are treated as prized possessions.

Again, to the contrary, in the case of deeply subaltern persons (such as slaves and outcasts) the way they are named often reflects their own subalternity (Benson, 2006) – for example, in eighteenth- and nineteenth-century Brazil, slaves used the surnames of the masters and their children inherited such surnames, so that, today, the fact of a black person inheriting a classic upper-class Portuguese surname can be taken to signify slave ancestry. The most extreme example of name oppression through name denial is probably what was attempted by the Austrian State during World War II (1938, the year Mauss delivered his famous article on the person), when all Jewish citizens were denied a surname and forced to adopt the same name, Israel for males and Sarah for females. Their birth registries were stamped over in red with these names.

Aristocrats often use different naming systems from commoners thus creating ready modes of signaling status. In contemporary metropolitan contexts, it is common for different ethnic groups to take recourse to different name series,

thus allowing for ready differentiation. But the use of *noms de guerre*, pseudonyms, and other more informal naming systems also often reflects collective belonging. This may be either through serial linking – as in the broad family of teknonyms (Mary's mother, John's father, etc.) or in group form (as in 'John of Musgrove,' where Musgrove is the name of a village or a household). Such naming processes are usually very sensitive to the interlocutors' relative status. In particular, nicknames (again I take recourse to an English shorthand for what is a very diverse family of informal forms of naming) can be either group names or more individuating names (in the Iberian Peninsula, for example, the way nicknaming is practiced differs profoundly from region to region – Pina-Cabral, João de, 1984). Such names may be comical or somehow demeaning and can be more or less readily accepted by the named person. Hypocoristics and diminutives usually bear within them some mark of the emotional or moral quality of the relation between the interlocutors.

Homonymy and Seriality

Names that describe only a singular person (what in English one normally calls 'proper names' – but again this category cannot be too broadly generalized) can be used to reflect collective belonging either in a person-to-person manner or a person-to-group manner. In some contexts, it is permissible to 'invent' new names (as in urban Brazil); most often it is not. Even where that is the case, however, one very soon meets up with the boundaries of creativity. Proper names are usually chosen from among a relatively limited list or are created through the use of a set of relatively limited techniques, or both. What this means is that homonymy is a recurrent aspect of most naming systems. People are often given names that belonged to other people before them (often relatives, patrons, godparents, etc.). Namesakes have relations between them, which are often quite intense (Pina-Cabral, João de, 2010b). They can be created in a number of ways, so it can be that people 'own' a series of valued names and, in the course of life, they pass these onto younger persons (Maybury-Lewis, 1974). Alternatively, people can receive the name of a dead relative (often a senior relative of the alternate generation – see McCallum, 2001 or Vernier, 1994) or they receive one of a rather limited list of names belonging to a family.

These forms of homonymy can reach paroxysms of repetitiveness, as was often the case in the so-called Celtic Fringe of Europe (from northern Portugal, along the Atlantic coast, to Scotland – see Dorian, 1970). Cases of generalized homonymy, however, should not be read as mistakes or as, somehow, signs of ignorance – as is so often taken for granted by ethnographers who study such contexts. Rather, name repetition must be seen as a perfectly legitimate (albeit statistical and not mechanical) mode of constructing collective belonging (Zonabend, 1984; Segalen, 1980).

People's names often respond to and reflect their particular social roles, as classically explained by Boas for the original populations of British Columbia that he studied (Boas, 1897). But there are also a large number of examples of this in Africa and Asia. The passing on of names between persons of different generations is often accompanied by the transfer of specific responsibilities and by the experience of constitutive surrogacy between the persons involved in the name transmission. In fact, it is generally the case that, even when there is no homonymy strictly speaking, the namer and the named experience a specifically strong relation of copresence (Pina-Cabral, João de, 2010b).

Homonymy and name seriality are somewhat continuous as processes. Very often people are 'named after' someone else merely by means of the linking of the names through a particle. Thus, for example, the initial letter of the name or an aspect of the name is kept on in the name of the younger named person, thus signaling social linking across generations.

This same process can be used to create commonality among members of one single generation. In upper-class Qing China, proper names of males were usually composed of two particles, the first of which was common to a series of brothers or patrilateral cousins (Alleton, 1993). This kind of seriality can also be found in contemporary Brazil and it was known to have occurred in some lower-class French populations (Wateau, 1989).

Naming and Life Course

Within each particular sociocultural context, persons typically follow a broadly recognizable life course. They are prone to change status at different moments in the life course and very often this change corresponds to recognized changes in naming. Such changes are often experienced as sudden, ritualized moments rather than as continuous evolutions. Ethnographers and historians have noted that these processes of personal change are not approached as binary breaks (on/off) but rather as rites of passage – that is, tripartite processes where the principles that effect the break are outlined by the nature of their absence during a liminary or intermediate stage.

One of the most common aspects of personal naming throughout the world is that a child seldom receives a definitive name until the immediate period of danger after birth has passed. In many contexts, recently born children are referred to by means of subterfuge or by offensive or descriptive names. The fear that the child might not be able to resist the forces of evil that confront a fully formed human was reflected in a number of naming practices. Ritual naming (as in baptism or other kinds of initiation rituals), however, usually occurs quite early on in the life of a person. Later on, successive life stages are often marked out by rituals, which characteristically involve name change; either through the actual change of name, or through the accumulation of names, or yet through the attribution of titles or other linguistic markers of distinction.

People may change certain parts of their names only (as in the case of French and British women, who often change surname at marriage – or, alternatively, in the case of monks, who change their proper names upon taking orders) or they can hoard valuable names that they carry with them and redistribute at a later moment. Life course rituals continue well into death and beyond in many instances, so funerary naming is very common. Of course, in historical contexts where Christianity has played a long-term role that is not the case. Since the sixth century, when Christian theories concerning the

spiritual essence of humans became more definitive, the soul is seen as created at conception and to continue forever (Givens, 2010). This means that baptism is conceived as a definitive change in a person's soul, a spiritual rebirth. To that extent, funerary names, which are so common elsewhere, do not make sense in Christian contexts.

The relation between naming and the essence of personhood – what anthropologists often call 'soul' – is very porous. Name magic is very commonly found, both for obtaining positive and negative effects. Indeed, this is the most common root of the widespread practices of name silencing and restriction of use (Parkin, 1989). While personal names cannot in principle be fully private, the fact is that a person's name is seen as somehow continuous with a person's value, so nowhere can names be used in an absurd or irresponsible way without that having implications for the bearer. In many contexts known to ethnographers, the European presupposition that formal names are essentially public does not apply. People may avoid declaring their name, they may not appreciate being called out aloud, they may feel their name is too precious a thing to be used lightly (Ramos, 1995). It must be understood, however, that the silencing of names is always a relative matter, not the same as the withdrawal of a name from the process of sociality.

There are, however, contexts where some people are not subject to direct naming. This may be a mark of subalternity, as used to be the case with Cantonese women (Watson, 1986). But it may also be a mark of distinction. Often, certain people are barred from pronouncing the name of certain other people (relatives, affines, friends, or hierarchical superiors), thus marking the specific nature of their relation by means of the negation of processes of naming that are taken for granted in other relations. Still again, rules of etiquette can limit considerably the way some people can use other people's names. In Imperial Qing China, for example, characters that were part of the Emperor's name were banned from general use for so long as he remained alive.

Conclusion

In the course of writing these pages I have been obliged to overuse words such as 'typically,' 'characteristically,' 'often,' 'may,' 'generally,' 'frequently.' While this may worry some readers for it may seem vague, the fact is we have no option but to learn to live with that kind of open-endedness in anthropological comparison and analysis. If, indeed, human communication is only possible within what Quine and Davidson called interpretive indeterminacy, and since sociality is essentially a communicative process, we have to accept that nothing will ever be absolute, complete or fully determinable in human experience (Davidson, 2005: 211). Thus, a condition for speaking of human life in its broadest outlines is to accept to remain open-ended. The alternative of giving up on broad generalization is as dependent on a dystopia of positive knowledge, as the alternative of believing in the possibility of formulating universalities is dependent on a utopia of positive knowledge. Both have proved equally insufficient for anthropological analysis. There is no margin for doubt that the deethnocentrification process that generations of ethnographers have patiently undertaken has left us with an invaluable legacy of information concerning the human condition in its more general outlines. But indeterminacy and underdetermination will never vanish, that is, history will ever be our condition – both in science and in everyday life. This, however, should not mean that we have failed to learn the great lessons inscribed in our anthropological and historical record.

See also: Affect and Emotion, Anthropology of; Childhood: Anthropological Aspects; Cognitive Development During Infancy and Early Childhood across Cultures; Cohabitation: United States; Comparative Method in Anthropology; Consciousness, Neural Basis of; Embodied Social Cognition; Family: The Anthropology of the Concept and Its History; Identity and Identification, Social Psychology of; Identity in Anthropology; Intentionality During Infancy and Early Childhood, Development of; Kinship in Anthropology; Levi-Strauss, Claude (1908–2009); Linked Lives; Personhood, Anthropology of; Self and Identity Development During Adolescence across Cultures; Self-Development in Childhood and Adolescence; Social Identity in Social Psychology.

Bibliography

Ademollo, Francesco, 2011. The Cratylus of Plato: A Commentary. Cambridge University Press, Cambridge.

Alleton, Viviane, 1993. Les chinois et la passion des noms. Aubier Montaigne, Paris.

Benson, Susan, 2006. Injurious names: naming, disavowal, and recuperation in contexts of slavery and emancipation. In: vom Bruck, G., Bodenhorn, Barbara. (Eds.), The Anthropology of Names and Naming, op.Cit, pp. 177–199.

Benveniste, Émile, 1966. Problèmes de Linguistique Générale, vol. 1. Gallimard, Paris, NRF.

Boas, Franz, 1897. The Social Organization and the Secret Societies of the Kwakiutl Indians. http://www.archive.org/details/cihm_14300. (accessed 01.11.12.).

Butler, Judith, 2012. Taking another's view: ambivalent implications. In: Axel Honneth Reification and Recognition: A New Look at an Old Idea. Oxford University Press, pp. 97–119.

Baxter, Timothy, 1992. The Cratylus: Plato's Critique of Names. E.J. Brill, Leiden.

Davidson, Donald, 2005. Truth, Language, and History. Clarendon Press, Oxford.

Dorian, Nancy C., 1970. A substitute name system in the Scottish Highlands. American Anthropologist 72 (2), 303–319.

Givens, Terryl L., 2010. When Souls Had Wings: Pre-mortal Existence in Western Thought. Oxford University Press, Oxford.

Lévi-Strauss, Claude, 1966. The Savage Mind. University of Chicago Press, Chicago.

Lévinas, Emmanuel, 1996. In: Peperzak, A., Critchley, S., Bernasconi, R. (Eds.), Basic Philosophical Writings. Indiana University Press, Bloomington, IN.

Lienhardt, Godfrey, 1985. Self: public, private. Some African representations. In: Carrithers, Michael, Collins, Steven, Lukes, Steven (Eds.), The Category of the Person: Anthropology, Philosophy, History. Cambridge University Press, Cambridge, pp. 141–155.

McCallum, Cecilia, 2001. Gender and Sociality in Amazonia: How Real People Are Made. Berg, Oxford.

Manning, Frank E., 1974. Nicknames and number Plates in the British West Indies. Journal of American Folklore 87 (344), 123–132.

Marriott, McKim, 1976. Hindu transactions. In: Kapferer, Bruce (Ed.), Transactions and Meaning. Institute for the Study of Human Issues, London, pp. 109–140.

Mauss, Marcel, 1985[1938]. A category of the human mind: the notion of person; the notion of self (W.D. Halls, Trans.). In: Carrithers, Michael, Collins, Steven, Lukes, Steven (Eds.), The Category of the Person: Anthropology, Philosophy, History. Cambridge University Press, Cambridge, pp. 1–25.

Maybury-Lewis, David, 1974[1967]. Akwe-shavante Society. Oxford University Press, New York.

Parkin, David, 1989. The politics of naming among the Giriama. Sociological Review Monograph 36, 61–89.
Pina-Cabral, João de, 1984. Nicknames and the experience of community. Man n.s. 19 (1), 148–150.
Pina-Cabral, João de, 2010a. The truth of personal names. Journal of the Royal Anthropological Society 16 (2), 297–312.
Pina-Cabral, João de, 2010b. Xará: namesakes in Southern Mozambique and Bahia (Brazil). Ethnos 73 (3), 323–345.
Pina-Cabral, João de, 2012. The functional fallacy: on the supposed dangers of name repetition. History and Anthropology 23 (1), 17–36.
Ramos, Alcida Rita, 1995. Sanumá memories: Yanomami ethnography in times of crisis. University of Wisconsin Press, Madison.
Rorty, Amélie O., 1969. Introduction. In: Rorty, A. (Ed.), The Identities of Persons. University of California Press, Berkeley, pp. 1–15.
Segalen, Martine, 1980. Le Nom Caché: la denomination dans le pays bigouden sud. L'Homme XX (4), 63–76.
Tooker, E., Conklin, H., 1984. Naming Systems. The American Ethnological Society, Washington, DC.
Toren, Christina, 2002. Anthropology as the science of all it takes to be human. In: Fox, Richard, King, Barbara (Eds.), Anthropology beyond Culture. Berg, Oxford, pp. 105–124.
Trevarthen, Colwyn, 1998. The concept and foundations of infant intersubjectivity. In: Bråten, Stein (Ed.), Intersubjective Communication and Emotion in Early Ontogeny. Cambridge University Press, Cambridge, pp. 15–46.
Vernier, Bernard, 1994. La génèse sociale des sentiments. Aînés et cadets dans l'île grecque de Karpathos. EHESS, Paris.
vom Bruck, G., Bodenhorn, B. (Eds.), 2006. The Anthropology of Names and Naming. Cambridge University Press, Cambridge.
Wateau, Fabienne, 1989. Gens de l'eau. Structure familiale de la batellerie artesanle du nord de la France. Ethnologie Française 19 (4), 350–361.
Watson, Rubie S., 1986. The named and the nameless: gender and person in Chinese Society. American Ethnologist 13 (4), 619–631.
Winnicott, Donald N., 1971. Playing and Reality. Routledge, London.
Zonabend, F., 1984. Prénom et identité. In: Dupâquier, J., Bideau, A., Ducreux, M.-E. (Eds.), Le prénom: mode et histoire. Entretiens de Malher 1980. Éditions EHESS, Paris, pp. 23–35.

Nanotechnology: Social and Cultural Aspects

Brigitte Nerlich and Szczepan Lemańczyk, University of Nottingham, Nottingham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Nanotechnology is not just an emerging technology but also an emerging cultural phenomenon. From its earliest beginnings nanotechnology and nanoculture have lived in a symbiotic relation, one inspiring the other. Promises and hopes, fears and doubts about nanotechnology have found expression in movies, novels, video games and so on. Nanotechnology and art have inspired each other. In this context public perceptions of nanotechnology have emerged, evolved, and been studied by social scientists. This discussion provides an overview of both the intertwining of nanotechnology and culture and of some of the public perceptions that may have been shaped by this process.

Introduction

The sociologist C. Wright Mills wrote in 1959:

> [I]n their everyday life they [men] do not experience a world of solid fact; their experience itself is selected by stereotyped meanings and shaped by ready-made interpretations. Their images of the world, and of themselves, are given to them by crowds of witnesses they have never met and shall never meet. Yet for every man these images – provided by strangers and dead men – are the very basis of his life as a human being.
>
> (from Alexander, 2009, p. 11)

Nanotechnology involves studying and working with matter on an ultrasmall scale: one nanometer (nm) is one-millionth of a millimeter, or one-billionth of a meter; a single human hair is around 80 000 nm in width. Nanotechnology deals with the creation and use of functional materials, devices, and systems with novel functions and promises to lead to improvements in microelectronics, energy technology, biomedicines, and more. Nanotechnology has a long history as a science and technology, but it only began to penetrate the sociocultural sphere in earnest around the turn of the second millennium. Public perceptions of this emerging technology are shaped by ready-made interpretations and images, as nanoscale phenomena are well beyond the range of what the human eye can see and as promising applications are still mostly in the future.

Stories and images are therefore important to embed this new technology in society and for society to be able to absorb, appreciate, contest, and engage with it. In the following section, we first provide a short cultural history of nanotechnology. This is followed by an overview of how cultural narratives and culturally salient images are used to construct perceptions and visions of nanotechnology. We then go on to sketch various attempts made by social scientists to assess public opinions and perceptions of nanotechnology, and finally we attempt to determine how public opinion may be shaped by the media, as a mediator between science and society and one of the main distributors of 'ready-made interpretations.'

A Short Sociocultural History of Nanoscience/Nanotechnology

Standard histories of nanoscience/nanotechnology typically begin by referring to Richard Feynman's famous 1959 lecture "There's Plenty of Room at the Bottom" (Feynman, 1959). In this afterdinner speech, Feynman, physicist and Nobel Laureate, sketched the possibility of manipulating and controlling matter on a small scale, indeed down to the bottom of the atomic scale. This mention of Feynman's talk is usually followed by a reference to K. Eric Drexler, who, in his 1986 book *Engines of Creation: The Coming Era of Nanotechnology*, speculated not only about controlling matter at the atomic scale but also about molecular assemblers that could potentially self-replicate endlessly, a dystopian scenario that could lead to what he called *gray goo* (Drexler, 1986). Drexler later regretted the use of this phrase, as it was employed in the initial phase of 'nano hype,' around 2004, as a way to induce fear rather than to portray nanotechnology as a revolutionary technology of hope and promise (Bensaude-Vincent, 2004).

Sometimes histories of nanotechnology refer to other events before and after Feynman (Regis, 1996), such as Norio Taniguchi using the term *nanotechnology* in a 1974 paper on ion-sputter machining; Drexler using the concept of molecular nanotechnology at the Massachusetts Institute of Technology in 1977; the first publication of a technical paper on molecular engineering and the invention of the scanning tunneling microscope in 1981; the discovery of the buckyball (C60) molecule, a first nanomaterial, in 1985, and the invention of the atomic force microscope in 1986.

A few years later, in 1989, Don Eigler of IBM Almaden and his colleagues used a scanning tunneling microscope (a microscope that sees through touch, so to speak, to spell out the IBM logo using xenon atoms, an image that put nanotechnology on the map of the public imagination (Eigler and Schwiezer, 1990)), and was followed by even more spectacular images of 'quantum corrals' that became icons of science and at the same time *objets d'art*. This was the beginning of a period of give and take between advances in nanoscience and advances in societal speculations about these advances, including a host of science fiction scenarios explored in popular culture (Nerlich, 2005).

In 2000, Bill Clinton, then president of the United States, launched the US National Nanotech Initiative, and serious financial investment began to pour into the development of nanotechnology. In 2003, US Congressional hearings were held and many surveys, opinion polls, and public engagement and outreach activities were launched on the societal implications of nanotechnology. In 2004, a landmark report by the U.K.'s Royal Society and Royal Academy of Engineering was published under the title Nanoscience and Nanotechnologies: Opportunities and Uncertainties (Royal Society and Royal Academy of Engineering, 2004).

Nanotechnology began to attract the attention of social scientists, who started to study public perceptions (Shapira et al., 2009), to engage the general public with nanotechnology and to reflect on its cultural significance, especially with relation to topics such as art and science, science and fiction, and science and literature – titles of three sections of one of the first books devoted to nano culture (Hayles, 2004).

Nano Facts Meet Nano Fictions

In this context of increasing social science scrutiny, the story or myth of Feynman as the founding father of nanotechnology was challenged by the cultural anthropologist Chris Toumey (Toumey, 2008), who told a more convincing story about the beginnings of nanotechnology. Furthermore, the idea that Feynman was the first to talk about manipulating matter at the atomic sale was challenged by Colin Milburn, an expert in science and the humanities (Milburn, 2008). He traced the idea back to the science fiction writer Robert Heinlein and his novella *Waldo* (see Milburn, 2008). Having come across both Toumey's and Milburn's work, the physicist Richard Jones pointed out on his blog *Soft machines*:

> There's no evidence that Feynman had read "Waldo" himself, but Feynman's friend Al Hibbs certainly had. Hibbs worked at Caltech's Jet Propulsion Laboratory, and he had been so taken by Heinlein's idea of robot hands as a tool for space exploration that he wrote up a patent application for it (dated 8 February 1958). Ed Regis, in his book *Nano* (1996), tells the story, and makes the connection to Feynman, quoting Hibbs as follows: "It was in this period, December 1958 to January 1959, that I talked it over with Feynman. Our conversations went beyond my 'remote manipulator' into the notion of making things smaller … I suggested a miniature surgeon robot …. He was delighted with the notion"
> (Jones, 2010).

In "Plenty of Room at the Bottom" Feynman (1959) did indeed talk about 'hundreds of tiny hands' and also about 'swallowing tiny surgeons.' Both images, whether related to Heinlein/Feynman or not, would flourish in fictional accounts of miniaturization and in semifictional accounts of nanoscience, and we shall come back to their enduring attraction in the next section. Here is what Feynman said in his talk about swallowing surgeons:

> A friend of mine (Albert R. Hibbs) suggests a very interesting possibility for relatively small machines. He says that, although it is a very wild idea, it would be interesting in surgery if you could swallow the surgeon. You put the mechanical surgeon inside the blood vessel and it goes into the heart and 'looks' around. (Of course the information has to be fed out.) It finds out which valve is the faulty one and takes a little knife and slices it out. Other small machines might be permanently incorporated in the body to assist some inadequately-functioning organ.
> (Feynman, 1959, p. 30)

A first quasi-scientific image (an artistic and imaginative rendering) of such a small machine appeared in 1988 on the cover of *Scientific American*: "A nanomachine swimming through a capillary attacks a fat deposit (such as normally may accompany an arteriosclerotic lesion)" (Nerlich, 2005). Coincidentally this was a year after the movie *Inner Space* was screened, which was itself a parody of the 1966 movie *Fantastic Voyage*, which became an enduring trope within nanoscientific and popular nanotechnology discourses (Nerlich, 2005, 2008). In *Fantastic Voyage* a submarine and its crew, including a surgeon, is shrunk to microscopic size and injected into the blood stream.

As late as 2010 one could still read pronouncements such as this: "'We are all motivated toward realizing the vision of the 1966 movie 'Fantastic Voyage' and by the potential to enhance medical treatment' Wang tells Nanowerk" (Berger, 2010). *Fantastic Voyage* is such an enduring trope that even those who want to provide more 'realistic' insights into the workings of nanotechnology in general and nanomedicine in particular end up replicating some of its aspects, as for example in an introduction to nanoscience for kids called *Nanoreisen* (nano voyages) (http://www.nanoreisen.com/; also see *Nanomission* at www.fei.com/resources/nano-games.aspx).

There has always been a singularly close relationship in nanotechnology between science and culture, images and imagination, and facts and fictions, with facts inspiring fictions and fictions inspiring facts (Goodsell, 2009). And, we should not forget that technologies, such as various types of advanced microscopes, are being used to do science and also to produce art (see a yearly nano art exhibition at http://nanoart21.org/).

But what does it mean to produce images of (invisible) nanoscale phenomena or images that popularize nanoscience dealing with nanoscale phenomena? And how do these images work? Most importantly, such images have to be able to tell a story. As Gimzewski and Vesna pointed out:

> In both the philosophical and visual sense, 'seeing is believing' does not apply to nanotechnology, for there is nothing even remotely visible to create proof of existence. … More than in any other science, visualization and creation of a narrative becomes necessary to describe what is sensed, not seen …. The scale becomes too abstract in relation to human experience.
> (Gimzewski and Vesna, 2003, p. 3)

In the following section we shall see how nanoscale phenomena are both made visible and used to create visions of the future.

Images Meet Visions

Nanotechnology deals with phenomena that are invisible to the human eye; they are indeed below the wavelength of light

(Goodsell, 2006, p. 44). This poses problems to image visualization. In addition, many applications of nanoscience still lie very much in the future. This opens up a space for the creation of futuristic visions, both utopian and dystopian ones. This in turn has implications for public acceptance or rejection of this new technology, as well as for attracting funding and investment.

The technological problem of making the invisible visible was solved with the invention of the scanning tunneling microscope and other microscopes of a similar type. The problem of public perception – of bringing nanoscience and its applications to the attention of the general public – was first overcome through the creation of images that make the unfamiliar familiar. As early as 1987 Bruno Latour had written about this issue:

> You feel bad because the nanometers of living cells baffle your mind? But it means nothing for anyone as long as it baffles the mind. It begins to mean something when nanometers are centimeters long on the scaled-up electron photograph of the cell, that is, when the eye sees it at the familiar scale and distance. Nothing is unfamiliar, gigantic or far away in these centers that cumulate traces; quite the opposite, they cumulate so many traces so that everything can become familiar, finite, nearby, and handy. (Latour, 1987, p. 230)

Philosophers of nano returned to this topic more recently in 2011:

> The nano-world is not a fantasy world, nor is it an urban, technical, or cosmic world. Instead we encounter for the most part a domesticated nature that is unsettled but not wild … We are drawn into a world in which we are used to move around that looks manageable and waits to be colonized.
> (Schwarz and Nordmann, 2011, p. 234)

The first to make the invisible visible and the unfamiliar familiar was, as we have seen, Eigler with the creation of the IBM logo written in xenon atoms. At the time this was a very laborious process, a process that has now become so commonplace. But, as Roald Hoffman wrote: "Raw electronic images have no color, only intensity among shades of gray. Wavelength information (color) may be communicated later … Immediately, in the choice of color(s), hue and intensity, one is led to artistic decisions" (Hoffmann, 2007, p. 20). And as Eigler himself, the creator of beautiful 'quantum corrals,' pointed out:

> The orange corral was something that I did specifically for the cover of *Science*. I wanted to create an image with as dramatic a perspective as possible to entice the editors into using the image on the cover of the magazine. I chose the colors, 'lighting conditions' and point of view of the observer to suit the purpose of the moment. … When Dominique Brodbeck created the blue corral image, we had a more educational or scientific purpose in mind, and so the perspective is one in which you get a much better idea of the shape of the corral and the shape of the waves in the interior of the corral. Nonetheless, Dominique's image appeared on the cover of *Physics Today* (November 1993).
> (quoted in Frankel, 2005, p. 263)

In the creation of these and many other nano images, science and culture, technology and art merge in a variety of combinations, depending on the cultural context in which they are used and on the communicative intention behind the merger. Artistic conventions of landscape painting or the painting of fictional scenarios can influence the representation of fact, as when coloring quantum corrals, perhaps in the tradition of magazines such as *Astounding Science Fiction* (Hoffmann, 2007), but fiction can also go it alone and create images of future nanomachines that do not yet exist.

As we have seen, fictional visions and images of nanoscience have been circulating in the discourse of science fiction for a long time. They can be used to shape public perceptions to create hopes or fears; they are producers of nano visions. Some of the most important images and fictions relating to nanoscience are those of 'swallowing the surgeon,' which, as noted earlier, were indicated by Feynman and linked to films such as *Fantastic Voyage* and its various sequels and caricatures (Nerlich, 2005). Here we have an image of hope related to nano medicine and expressed in language and imagery related to 'fighting' disease, a common way of speaking about health and illness (Nerlich, 2012). In this context nanobots are conceptualized and visualized as delivering drugs and healing people. One of the most important incarnations of such a nanobot is the so-called nano-louse, which has been variously studied by social scientists (Nerlich, 2008; Ruivenkamp and Rip, 2010), an artistic, but unrealistic rendering of a medical nanomachine in the tradition of the 1988 one that appeared on the cover of *Scientific American*. This image is however still being used in 2013, for example, in a promissory article on nano medicine (Kainic, 2013), which ends by saying, "[a]s Martel notes in an article in IEEE Magazine, 'There is no shortage of possibilities … a real-life fantastic voyage is just beginning.'"

However, there is also a darker side to nanobots, exploited in stories about 'gray goo' or self-replicating nanobots or nanites, perhaps best incarnated in Michael Crichton's novel *Prey*, published in 2002 just before a surge in public attention to nanotechnology. Nanites also have made their appearance in many other novels, movies and video games and have become a particular vehicle for public perceptions and visions of nanotechnology (Milburn, 2008). These types of images, utopian ones and dystopian ones, as well as merely playful ones, continuously serve to science-fictionalize science fact and blur the boundaries between cultural visions and scientific reality.

Images of nanobots not only switch between portraying promise or peril, they also switch, more regularly, between the biological and the mechanical, especially when nanoscientists talk about not-yet-existing but soon-to-exist nanomachines as if they were as real as the (metaphorical) motors, machines, propellers, pumps, gears, or scaffolds that already float through our bodies in the shape of enzymes, cell membranes, or parts of bacteria – as Lang and Gerber (2009, p. 24) call it, "a template orchestrated by nature." Here nanotechnology is made to appear normal by being seen through the lens of nature and the natural.

Images of flowers and landscapes in particular are often used to illustrate texts on nanotechnology. One of the most iconic nano flowers is the blue flower created by scientists in Cambridge in 2004. A BBC news item pointed out at the time that "the nano-sized plants are 'grown' from tiny droplets of the

liquid form of the metal gallium on a silicon surface" (BBC News, 2004). These flowers and landscapes make the unfamiliar seem familiar. This is also true of the use of nano-size cars or guitars or maps or any object than can be 'drawn' using advanced microscopy instruments, including in 2008, a 'Nanobama' (www.nanobama.com/).

Some of these construction and visualization efforts are linked to the metaphor of the 'master builder,' an over-arching metaphor that has been explored in detail by López (2008). Building the world atom by atom was in effect the title of a 1999 brochure produced by the National Science and Technology Council, whose front cover paints an image of an atomic landscape against a backdrop of space, thereby mapping the mappings of space exploration onto the exploration of the newly discovered nano space (Birkeland and Strand, 2009).

Images of nanoscale objects and the visions of nanotechnology that they convey are still being produced today and have become ubiquitous online, especially in image competitions and image galleries, as well as nano-art galleries that function as hybrid spaces between art, culture, and science (de Ridder-Vignone and Lynch, 2012).

Public Perceptions of, and Attitudes toward, Nanoscience and Nanotechnology

Alongside an ever-expanding and changing creation of cultural images and narratives around nanotechnology in science and in popular culture, there has been an increasing flow of studies into how nanotechnology is perceived when the general public is exposed to such images and visions, whether through traditional and online media or through reading novels or by playing video games (Alvardo, 2012). We return to media coverage of nanotechnology in the next section. In this section we provide a short survey of social science research into public perceptions and attitudes.

Social scientists and humanities scholars began to pay attention to nanotechnology around 2003/2004, after the congressional hearings in the United States and the publication of a Royal Society/Academy of Engineering report that emphasized the need to take account of potential public resistance to nanotechnology (Roco and Bainbridge, 2005). The Economic and Social Research Council (ESRC) in the United Kingdom said at the time that "nanotech looks set to become the next big showdown between science and society" (ESRC Social Science Week, 06/04, quoted in Nerlich et al., 2007, pp. 159–160).

In this context, many countries, especially the United States and the United Kingdom, invested heavily in public engagement with nanotechnology activities, and policy makers began to talk about 'upstream engagement' (Rogers-Hayden and Pidgeon, 2007). The hope was that if nanotechnology was publicly discussed openly and early, there would be a better chance of avoiding the so-called GM debacle, namely the full-scale rejection of a technology, genetically modified food in this case, by great parts of the general public.

Social scientists tried to establish a baseline for public understanding (or rather various aspects of public understanding) of nanotechnology that could then be used to engage sections of the general public more effectively with nanotechnology. Surveys focused in particular on differences between public perceptions in different countries (Gaskell et al., 2005), as such differences in perception that contributed to differences in accepting GM in Europe and the United States. They also compared the views of people holding different religious beliefs, for example, because such beliefs or values might impinge on differences in perception and uptake (Toumey, 2011).

Many studies have shown that (at least in the initial phase of nano awareness, and to some extent, nano hype) public attitudes toward nanotechnology have mostly been positive, especially in the United States (Scheufele and Lewenstein, 2005). Gaskell et al. (2005) point out significant differences in the way nanotechnology is perceived in the United States and European countries, with Europeans being more conscious about the potential nano risks. This research into public perceptions continues and is becoming more sophisticated.

It is generally assumed that some of the knowledge about nanotechnology that such surveys assess is derived from the mass media (Scheufele and Lewenstein, 2005). The media are one of the main purveyors of social and cultural knowledge and shape it in various ways through a process called 'framing' (based on the strategic use of words and images).

Nanotechnology in the Media

Where little direct experience of issues exists, the media play an important role in creating perceptions of such issues, and this is especially the case for nanotechnology, which is only gradually penetrating ordinary people's lives. This gradual penetration is of two kinds. On the one hand, mundane applications of nanotechnology are becoming commonplace but mainly go unnoticed. Such applications include the use of nanoparticles in self-cleaning windows or antibacterial socks; one can call such uses 'normal nano.' On the other hand, there is what one may call 'the nano-fantastic' described in films, novels, and video games. In such situations, where there is quite a gap between the mundane and the fantastic, the media play an even more important role in shaping public perceptions and attitudes. Anderson et al. state that "the press play a potentially crucial role in framing newly emerging issues, mainly by helping to establish the initial parameters of debate, by identifying certain news sources as pertinent and credible, and by providing topic-defining reference points" (Anderson et al., 2005, p. 202). Therefore, studies of the media coverage of nanotechnology have increasingly attracted the interest of social scientists and scholars in media studies (see, for example, Friedman and Egolf, 2005; Stephens, 2005; Anderson et al., 2005). In a study of media coverage of nanotechnology in German newspapers, Zimmer et al. (2008) suggest that discussions about nanotechnology are still in an initial phase and, unlike biotechnology, nanotechnology has not yet been a subject of many concerns, i.e., there is no 'Dolly the sheep' publicity or 'Frankenfood' campaign in connection with nanotechnology, despite some anxieties around toxicity, ground water pollution, penetration

of cell walls by nanoparticles, etc. It seems that so far the 'avoid the GM debacle' strategy has worked.

The Social and Cultural Framing of Nanotechnology

Framing is a way to organize a story or an idea (especially a new one) in such a way as to make it comprehensible to those who are so far unaware of it. Frames are sense-making tools (Scheufele and Tewksbury, 2007). According to a widely quoted definition (very often used in other framing studies) offered by Gamson and Modigliani, a frame is the "central organizing idea or story line that provides meaning to an unfolding strip of events …. The frame suggests what the controversy is about, the essence of the issue" (Gamson and Modigliani, 1987, p. 143). Traditional and online media are public spaces in which frames are created, used, and disseminated. This means that "media frames might help set the terms of the debate among citizens, but rarely, if ever, do they exclusively determine public opinion. Instead, as part of a 'frame contest,' one interpretative package might gain influence because it resonates with popular culture or a series of events, fits with media routines or practices, and/or is heavily sponsored by elites" (Nisbet and Scheufele, 2009, p. 1770).

A number of frames have been used to raise awareness of nanotechnology, to discuss its benefits and risks (Donk et al., 2012; Lewenstein et al., 2005), and just to enjoy the fun of its novelty. Frames can be signaled, triggered, or established through a variety of 'framing devices,' including metaphors (master builder, revolution), images (nanobots, nano flowers, etc.), stock myths and legends (Prometheus, Pandora's box), stock literary characters (Frankenstein, Dr Moreau, Waldo), or stock literary book or movie titles (*Brave New World, Fantastic Voyage*), stock historical references (asbestos, thalidomide), and clichés (playing God, going down the slippery slope, be careful what you wish for). Such literary devices can have positive or negative connotations; for example, a book that provides a dystopian vision of the future can be uncoupled from the original vision and later provide a positive framing, as mostly is the case when websites and headlines talk about the brave new world of nanotechnology.

The choice of frames is therefore important and has been discussed extensively, for example, by the European Commission in a brochure entitled "Understanding Public Debate on Nanotechnologies: Options for Framing Public Policy" (European Commission, 2010).

However, despite the deluge of work on social and cultural aspects of nanotechnology, only a small proportion of that research is devoted to examining media framings of attitudes toward and public perceptions of nanotechnology in non-Western cultures. There are some exceptions (for an overview, see Lemańczyk, 2012), but more work needs to be done, as, nanotechnology as a cultural phenomenon will be influenced by the cultural context in which it develops.

Conclusion

From its very beginning nanotechnology has not only been a scientific and technological phenomenon but also a social and cultural one. Artists have been inspired by advances in nanotechnology to produce new forms of art; novelists and film and video game makers have incorporated ideas derived from nanotechnology to engage readers, viewers, and players with new forms of science and new forms of fiction. Nanoscientists themselves have, conversely, been inspired by art to produce nanotechnology-based artifacts, ranging from letters to logos to flowers and landscapes. This almost continuous give-and-take between fact and fiction, hype and reality was at its height between 2002 and 2004. This is when Richard Jones wrote, "There is an almost surreal gap between what the technology is believed to promise and what it actually delivers" (Jones, 2004). Over the past decade this gap has become narrower. There are now numerous products on the market that contain nanoparticles, from sunscreens to trousers to (increasingly controversial) food. Nanotechnology has become part of reality. However, nanotechnology is also still a fertile ground for popular imagination, fiction, and speculation. It is therefore all around us in our everyday lives and in our everyday imaginations. We shall have to wait and see whether the issue of nano food for example might still become a GM moment in the development of nanotechnology and which cultural resources might be mobilized if it does.

See also: Chemical Science: History and Sociology; Cultural Studies of Science; Engineering Studies; Science, Public Engagement With; Science, Sociology of; Visualization in Science and Technology.

Bibliography

Alexander, R., 2009. Framing Discourse on the Environment: a Critical Discourse Analysis. Routledge, London.

Alvardo, S., 2012. Nanotechnology as Portrayed in Video Games – Part 2: Metal Gear Solid. Gamasutra. http://www.gamasutra.com/blogs/SebastianAlvarado/20130221/187106/Nanotechnology_as_Portrayed_in_Video_Games__The_Crysis_Nanosuit.php (accessed 06.11.12).

Anderson, A., et al., 2005. The framing of nanotechnologies in the British newspaper press. Science Communication 27 (2), 200–220.

BBC News, 22 June 2004. Science Cultivates "Nano-Flowers". http://news.bbc.co.uk/1/hi/sci/tech/3830061.stm (accessed 02.07.13).

Bensaude-Vincent, B., 2004. Two cultures of nanotechnology? In: Schummer, J., Baird, D. (Eds.), HYLE – International Journal for Philosophy of Chemistry 10 (2), 65–82.

Berger, M., 2010. Another Nanotechnology Step Towards "Fantastic Voyage". Nanowerk (accessed 02.07.13.). http://www.nanowerk.com/spotlight/spotid=19378.php.

Birkeland, T., Strand, R., 2009. Understanding nano images. Techné 13 (3), 182–189.

De Ridder-Vignone, K., Lynch, M., 2012. Images and imaginations. An exploration of nanotechnology image galleries. Leonardo 45 (5), 447–454.

Donk, A., et al., 2012. Framing emerging technologies: risk perceptions of nanotechnology in the German press. Science Communication 34 (1), 5–29.

Drexler, E.K., 1986. Engines of Creation. Anchor Press/Doubleday, Garden City, New York.

Eigler, D.M., Schwiezer, E.K., 1990. Positioning single atoms with a scanning tunnelling microscope. Nature 344, 524–526.

European Commission, 2010. Understanding Public Debate on Nanotechnologies: Options for Framing Public Policy. European Commission, Brussels.

Feynman, R., 1959. There is Plenty of Room at the Bottom. Available at: http://www.zyvex.com/nanotech/feynman.html.

Frankel, F.C., 2005. Sightings. American Scientist 93 (3).

Friedman, S., Egolf, B., 2005. Nanotechnology: risks and the media. IEEE Technology and Society Magazine 24 (4), 5–11.

Gamson, W.A., Modigliani, A., 1987. The changing culture of affirmative action. In: Braugart, R. (Ed.), Research in Political Sociology. JAI Press, Greenwich, CT, pp. 137–177.
Gaskell, G., et al., 2005. Imagining nanotechnology: cultural support for technological innovation in Europe and the United States. Public Understanding of Science 14 (1), 81–90.
Gimzewski, J., Vesna, V., 2003. The nanomeme syndrome: blurring of fact and fiction in the construction of new science. Technoetic Arts Journal 1 (1), 3.
Goodsell, D.S., 2009. Fact and fantasy in nanotech imagery. Leonardo 42 (1), 52–57.
Goodsell, D.S., 2006. Seeing the nanoscale. Nanotoday 1 (3), 44–49.
Hayles, K., 2004. Nanoculture: Implications of the New Technoscience. Intellect Books, London.
Hoffmann, R., 2007. Blow-up: images from the nano-world challenge viewers thinking. American Scientist 95 (1), 20.
Jones, R., 2010. Feynman, Waldo and the Wickedest Man in the World. Soft Machines. Available at: http://www.softmachines.org/wordpress/?p=958 (accessed 11.06.12.).
Jones, R., August 2004. The future of nanotechnology. Physics World, 25–29.
Kainic, J., 5 February 2013. Microbots: Using Nanotechnology in Medicine. Yale Scientific. http://www.yalescientific.org/2013/02/microbots-using-nanotechnology-in-medicine/.
Lang, H.P., Gerber, C., 2009. Up close & personal with atoms & molecules. Materials Today 12 (7–8), 18–25.
Latour, B., 1987. Science in Action: How to Follow Scientists and Engineers Through Society. Harvard University Press, Cambridge, MA.
Lemańczyk, S., 2012. Between national pride and the scientific success of "others": the case of Polish press coverage of nanotechnology, 2004–2009. NanoEthics 6 (12), 101–115.
Lewenstein, B., et al., 2005. The Salience of Small: Nanotechnology Coverage in the American Press, 1986–2004. Presented at International Communication Association. New York, 26–30 May, pp. 1–22.
López, J.J., 2008. Nanotechnology: legitimacy, narrative and emergent technologies. Sociology Compass 2 (4), 1266–1286.
Milburn, C., 2008. Nanovision: Engineering the Future. Duke University Press, Durham, NC.
Nerlich, B., 2005. From nautilus to nanobo(a)ts: the visual construction of nanoscience. Journal of Nanotechnology Online 1. http://www.azonano.com/article.aspx?ArticleID=1466 (accessed 02.07.13).
Nerlich, B., 2008. Powered by imagination: nanobots at the Science Photo Library. Science as Culture 17 (3), 269–292.
Nerlich, B., et al., 2007. Risks and benefits of nanotechnology: how young adults perceive possible advances in nanomedicine compared to conventional treatments. Health, Risk and Society 9 (2), 159–171.
Nerlich, B., 2012. Biomilitarism and nanomedicine: evil metaphors for the good of human health? Covalence Magazine: http://www.elca.org/What-We-Believe/Social-Issues/Faith-Science-and-Technology/Covalence/Features/Biomilitarism-and-nanomedicine.aspx.
Nisbet, M.C., Scheufele, D.A., 2009. What's next for science communication? Promising directions and lingering distractions. American Journal of Botany 96 (10), 1767–1778.
Regis, E., 1996. Nano: the Emerging Science of Nanotechnology. Little, Brown, Boston, MA.
Roco, M.C., Bainbridge, W.S., 2005. Societal implications of nanoscience and nanotechnology: maximizing human benefit. Journal of Nanoparticle Research 7 (1), 1–13.
Rogers-Hayden, T., Pidgeon, N., 2007. Moving engagement "upstream"? Nanotechnologies and the Royal Society and Royal Academy of Engineering's inquiry. Public Understanding of Science 16 (3), 345–364.
Royal Society and Royal Academy of Engineering, 2004. Nanoscience and Nanotechnologies. The Royal Society & The Royal Academy of Engineering, London.
Ruivenkamp, M., Rip, A., 2010. Visualizing the invisible nanoscale: study of visualization practices in nanotechnology community of practice. Science Studies 23 (1), 3–36.
Scheufele, D.A., Lewenstein, B., 2005. The public and nanotechnology: how citizens make sense of emerging technologies. Journal of Nanoparticle Research 7 (6), 659–667.
Scheufele, D.A., Tewksbury, D., 2007. Framing, agenda setting, and priming: the evolution of three media effects models. Journal of Communication 57 (1), 9–20.
Schwarz, A., Nordmann, A., 2011. "Hier bin ich Mensch, hier darf ich's sein!" – Partaking in the nanoworld. NanoEthics 5 (2), 233–243.
Shapira, P., Youtie, J., Porter, A.L., 2009. The emergence of social science research in nanotechnology. Scientometrics 85 (2), 595–611.
Stephens, L.F., 2005. News narratives about nano S&T in major U.S. and non-U.S. newspapers. Science Communication 27 (2), 175–199.
Toumey, C., 2008. Reading Feynman into nanotechnology: a next for a new science. Techné 13 (3), 133–168.
Toumey, C., 2011. Seven religious reactions to nanotechnology. NanoEthics 5 (3), 251–267.
Zimmer, R., Hertel, R., Böl, G.-F., 2008. Public Perceptions about Nanotechnology. Federal Institute for Risk Assessment, Berlin. Available here: http://www.bfr.bund.de/cm/350/public_perceptions_about_nanotechnology.pdf.

Narrative Approaches to Organizations

Hervé Corvellec, Lund University, Lund, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents an overview of the narratives in and about organizations. It is structured as follows. The first section introduces the idea that organizational members are *Homo Narrans* or storytelling beings; the second section describes the tenets of narrative enquiry in organizational settings; and the third section discusses the pros and cons of storytelling, which is the systematic use of narratives to manage people.

The notion of narrative does not only refer to novels, films, and other works of fiction. Narrative refers instead to any way of representing a series of events as a sequence that moves from a beginning point to an ending point. These events can be real or imagined, and they can pertain to anything that one has in mind, for example, nature, artifacts, or human activities. There exists an almost infinite diversity of narrative forms, for example, formal or mundane, oral or written, in dance, or in paintings. Narratives can be produced, circulated, and received in private or in public, individually or in a group. They can also serve a wide range of purposes, including incentivizing people to go to work. Narratives crowd politics, news reporting, and scientific theories, as well as sport commentaries, directions about how to get somewhere, and accounts of one's day at the office. There is no group of people without a narrative activity and tradition; narratives are present in every age, in every place, and in every society.

Organizations are no exceptions. Through their budgets, strategic plans, and annual reports to project meetings and wage discussions, narratives pervade organizations. These narratives are essential to organizational life, and studying how organizational narratives are produced, circulated, and received proves to be a fruitful way to study this life.

Part 1: Narrative Anthropology: Organizational Members as *Homo Narrans*

Narratives are essential to how humans create a sense of time for themselves (Ricœur, 1984). By ordering events chronologically, narratives provide this sense of time. Narratives can expand time. For example, a brief announcement of a Chief Executive Officer's (CEO) dismissal can be described in such a wealth of detail about the atmosphere and the attitudes of the people present that it takes a longer clock-time to read this description than it took for the Chairman of the Board to read the announcement. Inversely, narratives can compress time, for example, when an account of the event states that "It took a few weeks for the Board to find another CEO." Thanks to narratives, humans can actually relate to the mechanical clock-time and create an intelligibility of their lived times, that is time as they experience it in their lives.

The ability of narrative to make sense of time is closely connected to the ability of narrative to create a sense of causality. This latter ability is conveyed by the notion of plot. Aristotle defines in the *Poetics* a plot as an arrangement of events into a whole with a beginning, middle, and end. More precisely, a plot expresses how one passes from one state of affairs to another state of affairs that is different. By so doing, a plot provides a sense of progression, coherence, and understandability. For example: "The disclosure that the CEO had lied about having a PhD led to his immediate dismissal by the Chairman of the Board. However, the historical role of the CEO in the past developments of the company and major disagreements among the shareholders about the future orientations of the corporation made it difficult to find someone to replace him. It took a few weeks for the Board to find another CEO." A narrative's plot organizes one's understanding of how the past conditions the present and influences the future. This is particularly true of masterplots: plots that, because they have been told over and over, have acquired a moral force. For example, how the young David could defeat Goliath thanks to strategically directed efforts serves regularly as evidence that skill can prove to be superior to force.

Providing a sense of time and causality, narratives help people make sense of the negotiated order of organizational life. Organizing is akin to plotting (Czarniawska, 1997), and managers are the practical authors (Cunliffe, 2001) of their experience of an organization's life. Narratives have proven to be of practical help not only in making sense of organizational events, introducing change, and gaining political advantage (Boje, 1991), but also in identifying hazards, defining risk, and ascribing blame (Gephart, 1993). Even the celebrated process of drafting corporate strategies proves to be an ad hoc, conflictual, and multivoiced process of negotiating convincing narratives (Barry and Elmes, 1997). Managers expend consequential effort in selecting facts and ordering them into plots to produce a legitimate discourse that can orient their actions (Abolafia, 2010). Alone or with others, managers draw on all of the heroic, tragic, romantic, or ironic narrative styles to orient themselves and others in the ambiguous, changing, and open-ended everyday of organizational life.

The reason is that humans are in their actions and practice essentially storytelling animals (MacIntyre, 1981). Human communication rests on narration, and humans are *Homo Narrans* who depend on narratives to produce and communicate their interpretations of the world (Fisher,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.73118-7

1987). The capacity of humans to engage in collective actions is not limited to a logico-scientific mode of thought (Bruner, 1986): a mode of thought that rests on an ideal of a formal (mathematical) system of description and explanation, on categorization or conceptualization that follows a formal logic of conjunction, disjunction, implication, and aims at providing explanations in terms of causes and consequences that strive toward generalization. Rather, the capacity of humans to build and develop organizations rests as much on a narrative mode of thought (Bruner, 1986): a mode of thought that emphasizes the roles of good stories, gripping dramas, and believable historical accounts; that intends to address the vicissitudes of human or human-like intentions and actions, and more generally human experience; that works through contextualization; and that also emphasizes the singular.

Narratives are building stones of individual and organizational identity constructions. The history of a company posted on its Website is nothing but a selection of events that the current top management team wants the people inside and outside the organization to consider as characteristics of the company. Likewise, a job application emphasizes which past experiences and future expectations the applicant considers as most likely to be relevant to an advertised position. In both cases, it is a matter of building an identity through narration in the sense of defining the narrating subject with the help of a coherent and purposeful sequence of events in the past, present, and future. Organizational gossip and un-told stories function in the same way: they select and emphasize.

"We have evolved from being a waste management company to being an energy company" or "I am project manager with a proven focus on both people and results": such narratives help organizational members answer the existential questions of who they are, why they work, who they can trust, why they should say yes or no, and the like. Narratives are technologies of the self (Tamboukou, 2008) that provide ways for organizational members to work out their sense of organizational and professional identity. Correspondingly, narratives help nonorganizational members elucidate what kind of people and/or organization they are dealing with, and thus help them make sense of their and other people's identities as consumers. Narrative approaches to organizations link up with the thesis that humans realize the meaning of their life through narration (Arendt, 1958). Narration is the characteristic feature of the human condition in that it makes it possible for humans to go beyond a mere biological expression of life (*zôé*) and express a unique human ability not only to act, but also to give meaning to our actions by reflecting and communicating about them (*bios*).

By acknowledging the importance of narrative knowledge for interpretation, communication, and action, organization studies plug into the vast cultural heritage of storytelling. In particular, a narrative approach to organizations makes it possible to "gain access to deeper organizational realities, closely linked to their members' experiences" (Gabriel, 2000: p. 2). An interest in organizational narratives makes it possible for organization studies to (re-)connect to fundamental anthropological aspects of organizing that might otherwise remain ignored, such as the pervasiveness of fantasy in organizational settings as well other aspects such as humor or derision, complacency or contempt, even conflicts and violence. Organization members keep telling narratives to communicate and keep consuming narratives to make sense of their world. It is thus important to develop methodological ways to track and analyze the production and circulation of narratives in organizations. This is the object of organizational narrative enquiry.

Part 2: Narrative Enquiry: Ways of Studying Organizations

Narrative enquiry is not a cohesive field. Ways to collect and analyze narratives are as diverse as narratives are numberless. Narrative enquiry stands for a broad spectrum of fieldwork and reading methods, some of which will be introduced in this section.

Organizational narratives are often transmitted orally, in small talk at coffee breaks or on the job to help solve a task, and they include gossip, rumors, and jokes (Orr, 1996). Many of these narratives are not necessarily yet stabilized, realized, and articulated; they are prospective sensemaking, a kind of narrative bet on the future, and therefore Boje (2001, 2008) calls them antenarratives. Told and retold, antenarratives can become organizational narratives and even corporate legends that act as weaker versions of antique myths (Gabriel, 2000). An example of such a legend is the oft recounted anecdote that Ingvar Kamprad, the billionaire founder and CEO of the Swedish furniture company IKEA, once stopped and chatted with two of his executives sitting in business class in the front part of a plane while he was on his way toward the rear part of the plane to sit in economy class: this narrative perfectly serves IKEA's corporate policy of cost saving out of every detail and Kamprad's image of an ordinary, on the edge of stingy, self-made man. However, many organizational narratives are also put into print, like the CEO's letter to the shareholders, advertising, or job announcements. One can even collect written narratives by providing an arbitrary starting point and, with surprisingly good results, then asking managers to fill in with a narrative (Kostera, 2005).

Collecting written narratives is potentially less challenging, methodologically speaking, than collecting oral narratives as the latter often need to be elicited. Interviews and shadowing are efficient ways to collect narratives since many respondents provide accounts that are spontaneously formed into narrative. One can also prompt narratives by asking "can you recall when you first started to talk about the necessity of reorganizing your department? And what happened next?" (Czarniawska, 2004: p. 51); finally, working with the interview transcript, one can concoct one's own narrative out of what the respondents have said.

The rationale of organizational narrative enquiry is to consider that narration underlies organizing (Czarniawska, 1997), then look for narratives – for example, stories about reforms and change, life and identity, or failures and drama – and, finally, read the material that has been collected as narratives. Classical narratology (Herman and Vervaeck, 2005; Prince, 1982) offers a conceptual framework to analyze the formal characteristics of a narrative. These are the kinds of questions that narratology asks: Who tells the story

(narrator)? To whom is the story told (narratee)? Is the author the same as the narrator or is the story told from a point of view different from that of the author? What is the plot of the narrative? How is time rendered in this plot, and how does this affect the readability of the plot? How are the characters and the settings described? Classic narratology, which is strongly influenced by structuralism, tends to isolate the text from its social context. It focuses on the text's formal aspects, and this can be seen as a limiting trait to its use. Postmodern narratology tries to release narratology from this limitation, for example, coupling narratology to feminist studies, which opens new passageways for organization studies.

In contradistinction to techniques focused on the formal aspects of narratives, other narrative enquiry techniques provide ways to study narratives in their organizational context and help to understand this context. Boje (2001) lists eight such techniques. (1) Deconstruction unfolds the ideological assumptions that underlie a narrative, for example, which voices it privileges or silences. (2) Grand narrative analysis relates the narrative at hand to the metanarratives that characterize a specific society at one time in history, for example, the metanarrative of progress in modernity or that of incredulity in postmodernity. (3) Microstoria analysis looks at the other end of the macro–micro continuum and investigates how a narrative refers to knowledge and events that are very specific to a given time and given place. Microstoria often privileges narratives produced by the 'little people.' (4) Story network analysis aims at mapping, for example, with network diagrams, how a narrative relates to other narratives and how these relations follow or differ from the contact patterns among people or among organizations, and, more generally, the social networks within which organizational actors evolve. (5) Intertextuality analysis shares with story network analysis an interest in how narratives relate to one another, but focuses on how a narrative borrows or distances itself from other narratives, for example, through loans or quotations. (6) Causality analysis focuses on the kind of causes that are invoked to put together the fragments that constitute the narrative, for example physical, psychological, or successionist causality. (7) Also focused on causality, plot analysis focuses instead on how a narrative unfolds the arrangement of events in the text and then relates this arrangement to the four master forms of emplotment: romance, satire, comedy, and tragedy. (8) Finally, theme analysis uncovers how stories make sense of the chaotic everyday of organizational life by bringing events into known themes, for example, bureaucracy. Together, these ways of reading a narrative provide an understanding for how narratives function in the social settings where they are produced, diffused, and received.

Within all these accounts, narrative enquiry offers a rich method to analyze organizational culture. Methods inspired by narratology or other methods of narrative enquiry provide a practical means to critically examine taken-for-granted assumptions and analyze how organizational members and nonmembers perceive and use organizational rituals and symbols. Such methods also unfold the rationales of practices that are specific to the organization. More generally, narrative enquiry provides insights about how organizational members enact and reenact organizational culture.

Organizational scholars with an interest in narrative have also shown the relevance of works of fiction for management practice, management studies, and management education. A key rationale behind the use of works of fiction for the study, teaching, and practice of management is to get closer to the subjective experience of organizational membership (Phillips, 1995). Managers are invited to read good novels – and one could also watch good films – to become better managers (Czarniawska and de Monthoux, 1994) following the argument that works of fiction provide rich case studies that transmit tacit knowledge and render the complex dynamics of organizational life without simplifying it. For example, the classic play *Antigone* offers an opportunity to get public administrators into a discussion of the ethical dimensions of law enforcement (Marini, 1992); a bestselling textbook, *The Goal*, written in the form of a novel in the classic master–teacher–pupil genre, can effectively teach accounting techniques; and *Dilbert* (e.g., Adams, 1997) demonstrates the relevance of humorous narratives to convene, and possibly help support, the absurdities of office life. Paying attention to novels, films, TV series, and other narrative expressions of popular culture provides, moreover, an opportunity for managers and management scholars to reflect on how management is perceived by the public at large.

Part 3: Narrative Leadership: Storytelling in Organizations

The potential that exists in using narrative for a practical managerial purpose has not gone unnoticed by management consultants and other supporters of managers. In a few years, storytelling has established itself as a standard managerial tool. Armstrong (1992) is one of the earliest advocates of storytelling as a managerial tool. 'Storying around,' as he says, is simple – you have done it since childhood. Stories are timeless – old stories will prove as effective as new ones, and they are demographic-proof – everybody likes to listen to them. Stories are memorable, and they are therefore an excellent way to pass along traditions, an efficient form of training, and a wonderful form of recognition. Stories are also a great way to recruit and hire people; moreover, they even work to sell products. For Armstrong, storytelling is simply an effective way to make you a better manager.

There seems to be a story, type of story, or way of telling stories for every managerial situation, purpose, and style. Fables are praised for coming with a moral, sagas with a hero, and detective stories with a clarification. Storytelling is featured as a privileged way to communicate, inspire, mediate knowledge, build trust, support people, overcome problems, and bring along change – in two words: practice leadership. Storytelling is also invoked to turn the history of a company or a product into a good story that conveys an identity, a sense of time and place and values – a story easy to remember. This story can be an embellishment of the past, such as when leaders at the Walt Disney Corporation systematically invoke the genius (or genie?) of Walter Disney (Boje, 1995). But the story can also be faked, such as when two Swedish restaurant designers invent Joe Farelli, a successful restaurant owner from Little Italy in New York

City, to establish the credibility of an eponym mafia-films pitched restaurant in Gothenburg. "Let me tell you a story…" is featured as an all-round managerial tool.

This tool is not innocuous, though. Stories format brains and thoughts as much as they inform; and it is hard to have reservations about narratives since one of their key traits is to deflect questioning, bracket criticism, and neutralize critique. Storytelling is an efficient tool for corporate power: it creates a dream society that facilitates a corporate control on mental representations; a clever plotting has an ability to conceal logical contradictions and lock people in contradictory binds able to destructure work collectives; the depiction of actual persons as larger than life creates unrealistic patterns of explanation and expectations; a steady repetition of success stories establishes performance levels as standards that are basically unrealistic, thereby creating a permanent sense of shortfalls and failure among the workforce; and, puns and derisive stories can turn not only oppositions but even injustice and suffering into mockery (Salmon, 2010). Simple plots and streamlined organizational histories simplify the diversity of organizational life and legitimate ad hoc rewards and punishments (Corvellec, 1997). Storytelling harnesses the age-long art of telling stories to create a fictive work environment but uses this fiction to run real work places in the interest of the storyteller.

Conclusion

The growing returns of narrative anthropology, enquiry, and politics have prompted a vivid interest among organizational practitioners, scholars, and students for narratives from, in, about, and around organizations. This interest has allowed major advances, for example, in understanding how organizational members use links of causality when they provide accounts, the role of temporality in organizing, and, more generally, the richness of organizational members' experiences. Novels, film, and other narrative forms have a proven pedagogical record. Moreover, storytelling has grown into a managerial tool aimed at creating a sense of organizational history, developing the cohesiveness of organizational culture or supporting organizational change.

However, this success raises a corresponding need to develop a critical awareness about narratives in organizations. The point of such a critical awareness is not to reject a narrative mode of thinking, enquiring, and leading in favor of reverting to a mode of understanding organization that would rest exclusively on what Bruner (1986) calls a logico-scientific mode of thought. The narrative mode of thought offers unique epistemological and practical entries into the reality of organizational life, and it would be a real pity to ignore the benefits that it can bring. But at the same time, the use of narratives as a mode of enquiry, a way to present organizational research, or a management tool should be accompanied by a good dose of critical awareness. There are limits to the capacity of narrative to represent things (Kohler Riessman, 1993). A critical sense is needed to explore these limits, and, for example, identify the political agenda, values and priorities, and the worldview that any narrative perpetuates. Narratives can be so seductive that it is difficult to question how they connect the past, present, and future, and what they claim about how people get along. Critical awareness is what is needed to resist the seductive power of narratives.

See also: Corporate Culture; Leadership in Organizations, Sociology of; Organization: Overview; Organizational Climate; Organizations: Negotiated Orders.

Bibliography

Abolafia, Mitchel Y., 2010. Narrative construction as sensemaking: how a central bank thinks. Organization Studies 31 (3), 349–367.
Adams, Scott, 1997. The Dilbert Principle: A Cubicle's-Eye View of Bosses, Meetings, Management Fads & Other Workplace Afflictions. Boxtree, London.
Arendt, Hannah, 1958. The Human Condition. University of Chicago Press, Chicago, IL.
Armstrong, David, 1992. Managing by Storying Around: A New Method of Stewardship. Doubleday Currency, New York.
Barry, David, Elmes, Michael, 1997. Strategy retold: toward a narrative view of strategic discourse. Academy of Management Review 22 (2), 429–452.
Boje, David M., 1991. The storytelling organization: a study of story performance in an office-supply firm. Administrative Science Quarterly 36 (1), 106–126.
Boje, David M., 1995. Stories of the storytelling organization: a postmodern analysis of Disney as "Tamara-land." Academy of Management Journal 38 (4), 997–1035.
Boje, David M., 2001. Narrative Methods for Organizational and Communication Research. Sage, London.
Boje, David M., 2008. Storytelling Organizations. Sage, London.
Bruner, Jerome S., 1986. Actual Minds, Possible Worlds. Harvard University Press, Cambridge, MA.
Corvellec, Hervé, 1997. Stories of Achievements – Narrative Features of Organizational Performance. Transaction Publishers, New Brunswick, NJ.
Cunliffe, Ann L., 2001. Managers as practical authors: reconstructing our understanding of management practice. Journal of Management Studies 38 (3), 351–371.
Czarniawska, Barbara, 1997. Narrating the Organization: Dramas of Institutional Identity. University of Chicago Press, Chicago, IL.
Czarniawska, Barbara, 2004. Narratives in Social Science Research. Sage, London.
Czarniawska, Barbara, de Monthoux, Pierre Guillet, 1994. Good Novels, Better Management: Reading Organizational Realities. Harwood Academic Publisher, Amsterdam.
Fisher, Walter R., 1987. Human Communication as Narration: Toward a Philosophy of Reason, Value, and Action. University of South Carolina Press, Columbia, SC.
Gabriel, Yiannis, 2000. Storytelling in Organizations: Facts, Fictions, and Fantasies. Oxford University Press, Oxford.
Gephart Jr., Robert P., 1993. The textual approach: risk and blame in disaster sensemaking. Academy of Management Journal 36 (6), 1465–1514.
Herman, Luc, Vervaeck, Bart, 2005. Handbook of Narrative Analysis. University of Nebraska Press, Lincoln, NE.
Kostera, Monika, 2005. The Quest for the Self-Actualizing Organization. Liber, Malmö.
MacIntyre, Alasdair C., 1981. After Virtue: A Study in Moral Theory. University of Notre Dame Press, Notre Dame, IN.
Marini, Frank, 1992. The uses of literature in the exploration of public administration ethics: the example of antigone. Public Administration Review 52 (5), 420–426.
Orr, Julian E., 1996. Talking about Machines: An Ethnography of a Modern Job. ILR Press, Ithaca, NY.
Phillips, Nelson, 1995. Telling organizational tales: on the role of narrative fiction in the study of organizations. Organization Studies 16 (4), 626.
Prince, Gerald, 1982. Narratology: The Form and Functioning of Narrative. Mouton, Berlin.
Ricœur, Paul, 1984. Time and Narrative. University of Chicago Press, Chicago, IL.
Riessman, Catherine Kohler, 1993. Narrative Analysis. Sage, Newbury Park, CA.
Salmon, Christian, 2010. Storytelling: Bewitching the Modern Mind. Verso Books, London.
Tamboukou, Maria, 2008. A Foucaldian approach to narratives. In: Andrews, Molly, Squire, Corinne, Tamboukou, Maria (Eds.), Doing Narrative Research. Sage, Los Angeles, CA, pp. 102–120.

Narrative Exposure Therapy

Maggie Schauer, Clinical Psychology, University of Konstanz & vivo international, Konstanz, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Narrative Exposure Therapy (NET; Schauer, Neuner, Elbert 2005/2011) is an evidence-based treatment for trauma spectrum disorders in adult and child survivors of multiple stressors with complex trauma histories. NET was originally developed to treat survivors of war, terror, torture, and abuse. Since oral narratives are an integral part of every human culture worldwide, and imaginal exposure is proven to be a first-line treatment in trauma spectrum disorders, NET ensures healing of trauma symptoms through these components. The testimony approach that is inherent in this treatment enables the survivors to reclaim their life stories and also assists societal repair. Based on these principles, it has been shown that individuals from diverse backgrounds with a high trauma load and broken lifelines, suffering from traumatic stress, significantly benefit from NET. Working through an individual's biography facilitates recognition of the interrelated emotional networks of experiences in cases of multiple, repeated, or continuous trauma and hardship. This helps to build episodic memory and contextual indexing, fosters a sense of identity, and gives deep personal understanding of schemas and social emotions that have evolved across the lifespan. Revisiting and processing socially painful situations whilst supported by active, empathic listening allows for a corrective and healing relationship experience and integration. Subsequently disturbing symptoms significantly decline and the overall health and immune functions as well as social and occupational functionality increase. In Narrative Exposure work survivors can find meaning, testify to human rights violations, regain dignity, and satisfy the need for acknowledgment and validation.

Theoretical Underpinnings for Narrative Exposure Therapy

As humanitarians, we aim at alleviating the plight of war and complex emergencies, of natural and manmade disasters, assisting human beings who experience human-/child-rights violations, abuse, and continuous trauma. When one stressful experience after another adds to the growing associative memory of fear and trauma, the feeling of threat becomes a permanent plight. Given the high prevalence of emotional neglect, and familial, sexual, and organized violence that is not limited to geographic regions shaken by war and crisis but can be found in all countries of the world, low-threshold trauma treatments are needed to overcome these 'building blocks' of traumatic events and allow remembrance. In search for straightforward, disseminable approaches that are not linked to formal education, culture, or age, field-based studies have shown the efficacy of this short-term, evidence-based trauma treatment method, Narrative Exposure Therapy (NET), which can be successfully implemented on a large-scale level and applied by locally trained lay counselors (Schauer et al., 2011; Jacob et al., 2014; Neuner et al., 2004b; Schauer and Schauer, 2010). To heal individuals and communities it is essential to conduct and evaluate trauma treatments in real-world settings and to respond to universal human psychosocial needs (Nickerson et al., 2011; Robjant and Fazel, 2010).

Typically, the construction of an individual's reality involves a dimension of historicization and localization in the continuum of past, present, and future (Schauer and Ruf-Leuschner, 2014). Survivors of trauma, however, often encounter difficulty detailing a biographical account consistent in both content and sequence (Elbert and Schauer, 2002; Ehlers and Clark, 2000). In severe and complex cases, entire lifetime periods can seemingly be lost. Traumatized people fail to sufficiently connect their impulses, bodily sensations, and experiences to the contexts of their biography (Schauer et al., 2005/2011). Individuals who suffer from chronic trauma spectrum disorders have typically been exposed to multiple traumatic stressors such as assaults and persecution in their community or repeated incidences of familial violence including abuse, neglect, and social rejection. Effects are most devastating when the child had to endure repetitive and prolonged stress during development (Elbert and Schauer, 2014; Ruf and Schauer, 2012; Teicher et al., 2006).

As a consequence of their trauma memory, survivors suffer from a range of painful symptoms, leading to severe problems in daily functioning and participation in social life. Symptom remission and achieving functionality requires traumatic experiences to be oriented in the appropriate context, particularly the time and place, in the autobiography, i.e., trauma-focused remembering. Forcing a survivor to simultaneously recall, rank, and prioritize some traumatic events over others in order to participate in an 'exposure exercise' is often not only technically impossible but can even hinge on unethical. The process of having an individual select the 'worst' traumatic event from their trauma history oversimplifies the intricate context surrounding each event and neglects to address developmental influences. However, a viable treatment option that circumvents such complications and is both efficient and effective involves the narrating of the *entire* life story in a manageable period of time. The steady, step-by-step working through an individual's biography facilitates the recognition of interrelated emotional networks of experiences. The aim of NET is to complete the survivor's autobiography by linking *hot* (sensory–bodily–emotional) and *cold* (episodic) memory traces of events, thus contextualizing the experience of major traumata. Narrative Exposure builds on the theory of the dual representation of traumatic memories (Brewin et al., 2010; Elbert and Schauer, 2002; Schauer et al., 2011).

During a traumatic event, mainly sensory and perceptual information is stored in memory. The mind and body become

extremely aroused (rapid heartbeat, sweating, trembling) and brace for actions such as opposing or escaping. Emotional memories are tied together in a network of sensory, cognitive, emotional, and physiological elements. The sensory elements, together with the related cognitive, emotional, and physiological responses, then form associations in memory related to the traumatic experiences (*hot memory*; Metcalfe and Jacobs, 1996; *situationally accessible memory, or sensory perceptual representation*, see Brewin et al., 2010; Schauer et al., 2011). These hot memories are disconnected from the contextual information that constitutes the *cold memory* (*verbally accessible memory*, or *contextual representation*), i.e., the individual cannot remember the event within their contexts, i.e., where and when they have happened. Environmental stimuli and internal cues can easily activate the trauma structure. Multiple trauma is conceptualized as an extension of the fear/trauma network through experience of multiple traumatic events in such a manner that the fear network becomes increasingly strengthened and enlarged and ultimately can be triggered by exposure to a wide range of cues contained associated with the network.

Since the activation of the fear/trauma network serves as a frightening and painful recollection, many patients learn to avoid cues that act as reminders of the traumatic event and are unable to orient the specific fear associated with the events appropriately in time and space or to clearly structure these traumatic events in chronological order (Schauer et al., 2011). Trauma networks connect to response dispositions, which can either be an alarm response with fight and flight behavior or dissociative responding (fright, flag, and faint: e.g., functional sensory de-afferentation, emotional detachment, and numbing while tonic or flaccid immobile). Thus, dissociative amnesia or 'shut-down' can occur, replacing intrusions and hyperarousal with dissociation and passive avoidance (for a more detailed understanding of the biology of surviving and the defense cascade see Schauer and Elbert, 2010). NET is thought to reverse these detrimental conditions by tying down connections to their exact context.

As adversities and stressors cumulate, the trauma network expands, ultimately leading to clinically significant trauma-related suffering. There is a dose–response relationship ('building block') between experiences of traumatic events and trauma spectrum disorders. All symptoms of traumatic stress disorders and depression have repeatedly been shown to correlate in their severity with the cumulative exposure to traumatic stress (Mollica et al., 1998; Schauer et al., 2003; Neuner et al., 2004a; Elbert et al., 2009; Schauer E. and Elbert, 2010). More recently, it has become obvious that childhood adversity is the other major dimension in predicting trauma-related mental illness (Catani et al., 2009, Neuner et al., 2006; Teicher et al., 2003).

We all carry numerous implicit associative neural networks that have formed in response to emotionally arousing experiences, with many never reaching conscious awareness. Therein, mutually excitatory mnemonic representations are interconnected with action dispositions (emotions) and behavioral and physiological responses. Salient discrimination between past and present, between what are memories and what are current sensations, is essential. A well-tolerable way to attain this is working through both the negative and the positive life events chronologically, furnishing a narration about the entire biography. Reorganizing the memory and embedding the traumatic events in one's stream of life, in their appropriate positions on an individual's biographical timeline can achieve acknowledgment and closure of painful experiences. Revisiting positive, resourceful moments and taking the time to explicitly examine them one by one helps empower the individual to be able to look forward to the future (for both see Schauer et al., 2011).

In NET, the individual, with the assistance of the therapist, constructs a chronological narrative of his or her life story with a focus on the traumatic experiences. Fragmented reports of the traumatic experiences will be transformed into a coherent narrative. Empathic understanding, active listening, congruency, unconditional positive regard, and directive perseverance are key components of the therapist's approach for individuals under continuous trauma conditions, after familial, sexual, and organized violence, emotional neglect, or social relational victimization. For traumatic stress experiences the therapist asks in detail for context, emotions, cognitions, sensory information, physiological responses, and probes for respective observations. The narrator is encouraged to relive these emotions while actively maintaining their connection to the 'here and now': using permanent reminders that the feelings and physiological responses result from memories, the therapist links the experiences to episodic facts, i.e., time and place. In this way reprocessing, meaning making, and integration are facilitated. At the end of treatment the documented autobiography may be used for human rights advocacy.

The therapeutic procedure of NET (see **Figure 1**; *for a detailed account of the structure of NET, see the* treatment manual by Schauer *et al.,* (2011); for a clinicians short description see Elbert et al., 2015)

- Part 1 (duration: 1–2 sessions, each about 90–120 min): Structured clinical interview of trauma spectrum disorders including event checklists followed by a brief psychoeducational introduction (trauma memory, symptoms, outline of the treatment rationale and plan, approximate number of sessions)
- Part 2 is optional (duration: 1 session, 90–120 min): Laying out the *Lifeline* as a biographical overview
- Part 3 (duration: about 4–12 sessions, each 90–120 min): *Narrative Exposure* as the core procedure of NET, narrating the life story in several treatment sessions along the chronology of the timeline with an emphasis on high arousing moments, focusing on the reprocessing of traumatic experiences by inviting 'imaginal exposure'.

The Character of NET

To heal individuals and communities NET was tested in ecologically valid, various real-world settings, i.e., in scenarios of conflict and crisis as well as in typical inpatient and outpatient clinical settings (Schauer et al., 2004; Onyut et al., 2004, 2005; Neuner et al., 2004b, 2008a; Bichescu et al., 2007; Maedl et al., 2010; Neuner et al., 2011; Crombach and Elbert, 2014; Pabst et al., 2015; Hensel-Dittmann et al., 2011; Robjant and Fazel, 2010). NET was found to be a robust, low-threshold approach that is not linked to formal education, culture, or age. Field-based studies have shown the disseminability and

Part 1 Diagnostics	Part 2 Lifeline	Part 3 Narrative Exposure	
Diagnostic interview including event checklists + **Psychoeducation**	**Laying out the lifeline** symbolizing highly arousing, important events	• Narration **starts** at the beginning of life, proceeding along the lifeline. Traumatic events are confronted and reprocessed (**imaginal exposure**) until arousal decreases. Notes of key points are taken. • **Between sessions:** therapists structures the own transcript, writes the narration down and takes it to the next session. • In the **following session** in sensu exposure is facilitated again through the rereading of the last part of the narration with active participation of the survivor. Narration gets corrected and more details are added. • The **procedure is repeated across sessions along the timeline,** highlighting important life-events and summarizing others, until a final version of the client's biography is reached.	Closing session: **re-reading** the entire narration *or* portraying a **final lifeline** Survivor and all witnesses (therapist, interpreter, co-therapist) **ritually sign the testimony**. Document is **handed over** to the survivor. Utilization for **human rights work** or juridical purpose is discussed.

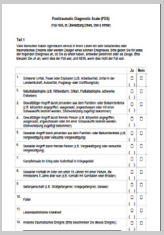

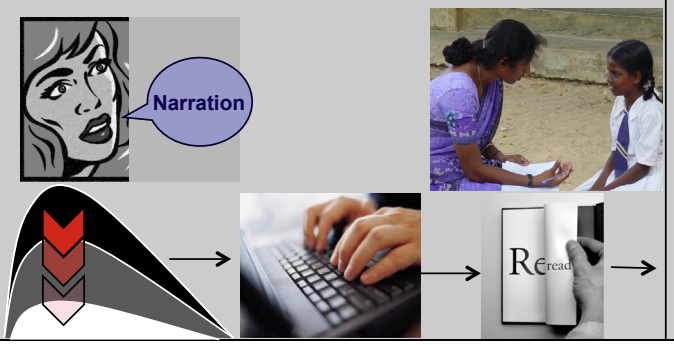

Figure 1 The Narrative Exposure process (Schauer, M., Neuner, F., Elbert T., 2011. Narrative Exposure Therapy: A Short Term Treatment for Traumatic Stress Disorders, second ed. Hogrefe Publishing, Cambridge, MA.).

efficacy of this short-term, trauma-focused treatment module that can be successfully built into large-scale service provision and applied by, e.g., trained counselors, psychologists and psychiatrists, medics, paramedics, social workers, and teachers (Schauer, E. 2008; Neuner et al., 2008b; Jacob et al., 2014). Because narratives are an integral part of every culture, people from diverse backgrounds with broken lifelines suffering from stress caused by events such as, e.g., sexual and physical childhood abuse, forced recruitment or migration, political violence and torture, life-threatening illnesses, disasters, or victimization by peers and attachment figures can benefit from the building of episodic memory and the facilitation of their testimony. NET has a therapeutic and human rights focus.

Therapeutic elements of NET that have proven effective in trauma treatment (from Schauer et al., 2005/2011)

1. Active chronological reconstruction of the autobiographical/episodic memory.
2. Imaginal exposure to the traumatic events ('hot spots') and full activation of the fear memory in order to modify the emotional network through detailed narration and imagination of the traumatic events.
3. Meaningful linkage and integration of physiological, sensory, cognitive, and emotional responses to one's time, space, and life context (i.e., comprehension of the original context of acquisition and the reemergence of the conditioned responses in later life).
4. Cognitive reevaluation of behavior and patterns (i.e., cognitive distortions, automatic thoughts, beliefs, responses), as well as reinterpretation of the meaning – content through reprocessing of negative, fearful, and traumatic events – completion and closure.
5. Revisiting of positive life-experiences to activate resources and to adjust basic assumptions.
6. Regaining of the survivor's dignity through satisfaction of the need for acknowledgment through the explicit human rights orientation of 'testifying.'

For full information about the NET procedure, see the treatment manual:

Schauer, M., Neuner, F., Elbert, T., 2011. Narrative Exposure Therapy: A Short-Term Treatment For Traumatic Stress Disorders, second ed. Hogrefe Publishing, Cambridge, MA.

In addition to the English edition, there is currently a Japanese, French, Korean, Slovak, and Italian version of the treatment manual as well as a Dutch interpretation (see references).

Evidence for NET

Adult and child survivors with multiple traumatizing life-events have been demonstrated to benefit from NET. State-of-the-art *in sensu* exposure and the narrative reprocessing of the trauma memory, the broken lifeline, and ultimately the entire biography effectively reduces suffering and allows the survivor

to foster personal identity (Jongedijk, 2014; Robjant and Fazel, 2010; Dömen et al., 2012). Individuals from diverse backgrounds with lives torn apart by stressful events like childhood abuse and neglect, loss of caregivers, adoption, forced recruitment or migration, political violence and torture, life-threatening illnesses, severe disasters, and many more show significantly reduced clinical symptomatology and an enhanced quality of life and level of occupational and social functioning (Mørkved et al., 2014; Hermenau et al., 2013; Dyregrov and Yule, 2006). NET therefore also provides effective treatment for survivors of organized violence and severe torture experiences with large effect sizes (Hensel-Dittmann et al., 2011; Neuner et al., 2010) or complex trauma survivors, including those with borderline personality disorders (Pabst et al., 2012a,b, in press). Stenmark et al. (2013) showed that with NET, refugees as well as asylum seekers can be successfully treated for post-traumatic stress disorder (PTSD) and depression in a general psychiatric health-care system. Most pronounced improvements are observed at (long-time) follow-up, suggesting that NET elicits an ongoing change in the dynamics of self-perception and self-regulation of the client, commencing a healing process that eventually leads to a sustained improvement in psychopathological symptoms, physical health, functioning, and quality of life. NET has effectively been applied in situations that remain volatile and insecure, like conditions of continuous trauma, i.e., ongoing real threat, e.g., by war or by a perpetrator (Neuner et al., 2014). It effectively reduces PTSD symptoms in the individual whilst bearing witness to the atrocities endured. Reviews identified NET as an evidence-based treatment for different groups of survivors of violence (Robjant and Fazel, 2010; Crumlish and O'Rourke, 2010; McPherson, 2012; Nickerson et al., 2011). Meanwhile, a number of studies showing the effectiveness of NET have been conducted completely independent from the originators of NET (Zang et al., 2013; Hijazi, 2012; Hijazi et al., 2014; Gwozdziewycz and Mehl-Madrona, 2013; Ejiri et al., 2012; Dōmen et al., 2012) and thus the NET procedure has been successfully implemented in a variety of countries (e.g., Zech and Vandenbussche, 2010; Jongedijk, 2012, 2014). NET is available in special adaptations for offenders (FORNET; Elbert et al., 2012; Hermenau et al., 2013; Crombach and Elbert, 2014) and for children and adolescents (KIDNET; Schauer et al., 2005, 2004/2011; Onyut et al., 2005; Catani et al., 2009; Ruf et al., 2010; Hermenau et al., 2012; Neuner et al., 2008a; Ruf and Schauer, 2012).

The effectiveness of NET has been demonstrated not only by measuring the clinical and social symptoms but also in validating these results by means of markers from neurophysiology and molecular biology. NET as a successful psychotherapeutic intervention can reorganize the memory and modify the architecture of the brain on a macroscopic level (Schauer et al., 2007, 2006). The success not only showed up in symptom scores but also in parameters of neuromagnetic activity. During the 6-month follow-up oscillatory brain activity in the NET group, but not in the control group, became more similar to that of healthy controls. Moreover, using magnetic source imaging of the brain, Adenauer et al. (2011) observed that NET causes enhanced cortical top-down regulation, which is associated with the ability to inhibit the fear response.

Various somatic diseases, including chronic pain, cancer, cardiovascular, respiratory, gastrointestinal, and autoimmune diseases (Boscarino, 2004; Seng et al., 2006; Felitti et al., 1998), emerge from traumatizing experiences. Altered immune functions and inflammatory processes are found to be responsible for the poor physical health in individuals with PTSD (Pace and Heim, 2011). NET was shown to turn health parameters (like frequencies of cough, diarrhea, and fever) for the better even under harsh living conditions (Neuner et al., 2008b). In the immune system, T-cells are critical for maintaining balance, regulating the immune response, and preventing autoimmune diseases. Recently, Morath et al. (2014) showed a treatment-related increase in the previously decreased proportion of regulatory T-cells in the NET group at 1-year follow-up. Moreover, NET is able to reverse in individuals with PTSD the pathological levels of DNA strand breaks back to a normal level (Morath et al., 2014b). These findings may have implications for physical health, including carcinogenesis.

A decisive strength of NET is its encouragingly low drop-out rate and, due to its robust nature, the potential for dissemination, including to counselors in low-income countries, war, and crisis regions (Schauer and Schauer, 2010; Catani et al., 2009; Neuner et al., 2008b; Ertl et al., 2011; Jacob et al., 2014).

See also: Mass Trauma: Psychopathological Effects across the Life Span; Post-Traumatic Stress Disorder; Post-Traumatic Stress in Social Work; Torture and its Consequences, Psychology of.

Bibliography

Adenauer, H., Catani, C., Gola, H., Keil, J., Ruf, M., Schauer, M., Neuner, F., 2011. Narrative Exposure Therapy for PTSD increases top-down processing of aversive stimuli – evidence from a randomized controlled treatment trial. BMC Neuroscience 12 (1), 127.

Brewin, C.R., Gregory, J.D., Lipton, M., Burgess, N., 2010. Intrusive images in psychological disorders: characteristics, neural mechanisms, and treatment implications. Psychological Review 117 (1), 210.

Bichescu, D., Neuner, R., Schauer, M., Elbert, T., 2007. Narrative Exposure Therapy of political imprisonment-related chronic trauma-spectrum disorders. Behaviour Research and Therapy 45, 2212–2220.

Boscarino, J., 2004. Posttraumatic stress disorder and physical illness: results from clinical and epidemiologic studies. Annals of the New York Academy of Sciences 1032, 141–153.

Catani, C., Kohiladevy, M., Ruf, M., Schauer, E., Elbert, T., Neuner, F., 2009. Treating children traumatized by war and tsunami: a comparison between exposure therapy and meditation-relaxation in north-east Sri Lanka. BMC Psychiatry 9, 22. http://dx.doi.org/10.1186/1471-244X-9-22.

Crombach, A., Elbert, T., 2014. Controlling offensive behavior using Narrative Exposure Therapy a randomized controlled trial of former street children. Clinical Psychological Science, 2167702614534239.

Crumlish, N., O'Rourke, K., 2010. A systematic review of treatments for post-traumatic stress disorder among refugees and asylum-seekers. Journal of Nervous and Mental Disease 198 (4), 237–251.

Dōmen, I., Ejiri, M., Mori, S., 2012. Narrative Exposure Therapy for the treatment of complex PTSD: an examination of the effect and adaptation. The Japanese Journal of Psychotherapy 13 (1), 67–74.

Dyregrov, A., Yule, W., 2006. A review of PTSD in children. Child and Adolescent Mental Health 11 (4), 176–184.

Ehlers, A., Clark, D.M., 2000. A cognitive model of posttraumatic stress disorder. Behaviour Research and Therapy 38 (4), 319–345.

Ejiri, M., Dōmen, I., Mori, S., 2012. A trial study for introducing Narrative Exposure Therapy into psychiatric practice: an examination of the effect and adaption. The Japanese Journal of Psychotherapy 13 (1), 59–65.

Elbert, T., Schauer, M., 2002. Burnt into memory. Nature 412, 883.

Elbert, T., Schauer, M., 2014. Epigenetic, neural and cognitive memories of traumatic stress and violence. In: Cooper, S., Ratele, K. (Eds.), Psychology Serving Humanity: Proceedings of the 30th International Congress of Psychology, Western Psychology, vol. 2. Psychology Press, East Sussex, New York, pp. 215–227.

Elbert, T., Schauer, M., Schauer, E., Huschka, B., Hirth, M., Neuner, F., 2009. Trauma-related impairment in children – an epidemiological survey in Sri Lankan provinces affected by two decades of civil war and unrest. Child Abuse & Neglect 33, 238–246.

Elbert, T., Hermenau, K., Hecker, T., Weierstall, R., Schauer, M., 2012. FORNET: Behandlung von traumatisierten und nicht-traumatisierten Gewalttätern mittels Narrativer Expositionstherapie. In: Endrass, J., Rossegger, A., Urbaniok, F., Borchard, B. (Eds.), Interventionen bei Gewalt- und Sexualstraftätern: Risk-Management, Methoden und Konzepte der forensischen Therapie. Medizinisch Wissenschaftliche Verlagsgesellschaft, pp. 255–276.

Elbert, T., Rockstroh, B., Kolassa, I.-T., Schauer, M., Neuner, F., 2006. The influence of organized violence and terror on brain and mind – a co-constructive perspective. In: Baltes, P., Reuter-Lorenz, P., Rösler, F. (Eds.), Lifespan Development and the Brain: The Perspective of Biocultural Co-Constructivism. Cambridge University Press, New York, pp. 326–349. http://www.eabct.com/pdf/common_language.pdf.

Elbert, T., Schauer, M., Neuner, F., 2015. Narrative Exposure Therapy (NET) – reorganizing memories of traumatic stress, fear and violence. In: Cloitre, Schnyder (Eds.), Evidence Based Treatments for Trauma Related Disorders: A Practical Guide for Clinicians, in press.

Ertl, V., Pfeiffer, A., Schauer, E., Elbert, T., Neuner, F., 2011. Community-implemented trauma therapy for former child soldiers in Northern Uganda: a randomized controlled trial. JAMA 206, 503–512.

Felitti, M.D., Vincent, J., Anda, M.D., Robert, F., Nordenberg, M.D., Williamson, M.S., James, S., 1998. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: the Adverse Childhood Experiences (ACE) study. American Journal of Preventive Medicine 14 (4), 245–258.

Gwozdziewycz, N., Mehl-Madrona, L., 2013. Meta-analysis of the use of Narrative Exposure Therapy for the effects of trauma among refugee populations. The Permanente Journal 17 (1), 70.

Halvorsen, J.Ø., Stenmark, H., 2010. Narrative Exposure Therapy for posttraumatic stress disorder in tortured refugees: a preliminary uncontrolled trial. Scandinavian Journal of Psychology 51 (6), 495–502.

Hensel-Dittmann, D., Schauer, M., Ruf, M., Catani, C., Odenwald, M., Elbert, T., Neuner, F., 2011. The treatment of traumatized victims of war and torture: a randomized controlled comparison of Narrative Exposure Therapy and stress inoculation training. Psychotherapy and Psychosomatics 80, 345–352.

Hermenau, K., Hecker, T., Ruf, M., Schauer, E., Elbert, T., Schauer, M., 2012. Childhood adversity, mental ill-health and aggressive behavior in an African orphanage: changes in response to traumafocused therapy and the implementation of a new instructional system. Child and Adolescent Psychiatry and Mental Health 5, 29.

Hermenau, K., Hecker, T., Schaal, S., Maedl, A., Elbert, T., 2013. Addressing post-traumatic stress and aggression by means of Narrative Exposure: a randomized controlled trial with ex-combatants in the Eastern DRC. Journal of Aggression, Maltreatment & Trauma 22 (8), 916–934.

Hijazi, A.M., 2012. Narrative Exposure Therapy to Treat Traumatic Stress in Middle Eastern Refugees: A Clinical Trial (Ph.D. thesis) – Wayne State University. Wayne State University Dissertations. Paper 543.

Hijazi, A.M., Lumley, M.A., Ziadni, M.S., Haddad, L., Rapport, L.J., Arnetz, B.B., 2014. Brief Narrative Exposure Therapy for posttraumatic stress in Iraqi refugees: a preliminary randomized clinical trial. Journal of Traumatic Stress 27, 314–322.

Isele, D., Teicher, M.H., Ruf-Leuschner, M., Elbert, T., Kolassa, I.T., Schury, K., Schauer, M., 2014. KERF-An instrument for measuring adverse childhood experiences: construction and psychometric evaluation of the German MACE (Maltreatment and Abuse Chronology of Exposure) scale. Zeitschrift für Klinische Psychologie und Psychotherapie 43 (2), 121–130.

Jacob, N., Neuner, F., Mädl, A., Schaal, S., Elbert, T., 2014. Dissemination of psychotherapy for trauma-spectrum disorders in resource-poor countries: a randomized controlled trial in Rwanda. Psychotherapy & Psychosomatics 83, 354–363.

Jongedijk, R.A., 2012. Hoofdstuk (chapter) 34. Narratieve Exposure Therapie. In: Vermetten, E., Kleber, R.J., van der Hart, O. (Eds.), Handboek Posttraumatische stress stoornissen. De Tijdstroom, Utrecht, pp. 551–564.

Jongedijk, R.A., 2014. Levensverhalen en psychotrauma. Narratieve Exposure Therapie in theorie en praktijk (Life Stories and Psychotrauma. Narrative Exposure Therapy in Theory and Practice). Uitgeverij Boom (Boom Publishers), Amsterdam.

McPherson, J., 2012. Does Narrative Exposure Therapy reduce PTSD in survivors of mass violence? Research on Social Work Practice 22 (1), 29–42.

Metcalfe, J., Jacobs, W.J., 1996. A "hot-system/cool-system" view of memory under stress. PTSD Research Quarterly 7 (2), 1–3.

Mollica, R.F., McInnes, K., Poole, C., Tor, S., 1998. Dose-effect relationships of trauma to symptoms of depression and post- traumatic stress disorder among Cambodian survivors of mass violence. The British Journal of Psychiatry 173 (6), 482–488.

Maedl, A., Schauer, E., Odenwald, M., Elbert, T., 2010. Psychological rehabilitation of ex-combatants in non-western, post-conflict settings. In: Trauma Rehabilitation after War and Conflict. Springer, New York, pp. 177–213.

Morath, J., Moreno-Villanueva, M., Hamuni, G., Kolassa, S., Ruf, M., Schauer, M., Bürkle, A., Elbert, T., Kolassa, I.T., 2014. Effects of psychotherapy on DNA strand break accumulation originating from traumatic stress. Psychotherapy and Psychosomatics 83 (5), 289–297.

Morath, J., Gola, H., Sommershof, A., Hamuni, G., Kolassa, S., Catani, C., Adenauer, H., Ruf-Leuschner, M., Schauer, M., Elbert, T., Groettrup, M., Kolassa, I.T., 2014b. The effect of trauma-focused therapy on the altered T cell distribution in individuals with PTSD. Evidence from a randomized controlled trial. Journal of Psychiatric Research 54, 1–10.

Mørkved, N., Hartmann, K., Aarsheim, L.M., Holen, D., Milde, A.M., Bomyea, J., Thorp, S.R., 2014. A comparison of Narrative Exposure Therapy and Prolonged Exposure Therapy for PTSD. Clinical Psychology Review 34 (6), 453–467.

Neuner, F., Schauer, M., Karunakara, U., Klaschik, C., Robert, C., Elbert, T., 2004a. Psychological trauma and evidence for enhanced vulnerability for PTSD through previous trauma in West Nile refugees. BMC Psychiatry 4, 34.

Neuner, F., Schauer, M., Klaschik, C., Karunakara, U., Elbert, T., 2004b. A comparison of Narrative Exposure Therapy, supportive counseling, and psychoeducation for treating posttraumatic stress disorder in an African refugee settlement. Journal of Consulting and Clinical Psychology 72 (4), 579–587.

Neuner, F., Schauer, E., Catani, C., Ruf, M., Elbert, T., 2006. Post-tsunami stress: a study of posttraumatic stress disorder in children living in three severely affected regions in Sri Lanka. Journal of Traumatic Stress 19, 339–347.

Neuner, F., Kurreck, S., Ruf, M., Odenwald, M., Elbert, T., Schauer, M., 2010. Can asylum-seekers with posttraumatic stress disorder be successfully treated? A randomized controlled pilot study. Cognitive Behavior Therapy 39, 81–91.

Neuner, F., Schauer, M., Elbert, T., 2011. Psychotherapy in war and crisis regions: Narrative Exposure Therapy and beyond. ISTSS Traumatic Stress Points 25 (1), 8–9.

Neuner, F., Catani, C., Ruf, M., Schauer, E., Schauer, M., Elbert, T., 2008a. Narrative Exposure Therapy for the treatment of traumatized children and adolescents (KIDNET): from neurocognitive theory to field intervention. Child and Adolescent Psychiatric Clinics of North America 17 (3), 641–664.

Neuner, F., Onyut, P.L., Ertl, V., Odenwald, M., Schauer, E., Elbert, T., 2008b. Treatment of posttraumatic stress disorder by trained lay counselors in an African refugee settlement: a randomized controlled trial. Journal of Consulting and Clinical Psychology 76, 686–694.

Neuner, F., Schauer, M., Elbert, T., 2014. On the efficacy of Narrative Exposure Therapy: a reply. Intervention 12 (2), 267–278.

Nickerson, A., Bryant, R.A., Silove, D., Steel, Z., 2011. A critical review of psychological treatments of posttraumatic stress disorder in refugees. Clinical Psychology Review 31 (3), 399–417.

Onyut, P.L., Neuner, F., Schauer, E., Ertl, V., Odenwald, M., Schauer, M., Elbert, T., 2005. Narrative Exposure Therapy as a treatment for child war survivors with posttraumatic stress disorder: two case reports and a pilot study in an African refugee settlement. BMC Psychiatry 5, 7.

Onyut, L.P., Neuner, F., Schauer, E., Ertl, V., Odenwald, M., Schauer, M., Elbert, T., 2004. The Nakivale camp mental health project: building local competency for psychological assistance to traumatized refugees. Intervention 2, 90–107.

Pace, T.W., Heim, C.M., 2011. A short review on the psychoneuroimmunology of posttraumatic stress disorder: from risk factors to medical comorbidities. Brain, Behavior, and Immunity 25 (1), 6–13.

Pabst, A., Aldenhoff, J., Schauer, M., Ruf, M., Elbert, T., Seeck-Hirschner, M., 2012a. Behandlung von Borderline Störungen mit komorbider Posttraumatischer Belastungsstörung. Anwendung der narrativen Expositionstherapie (NET). Nervenheilkunde 31, 645–652.

Pabst, A., Schauer, M., Bernhardt, K., Ruf, M., Goder, R., Rosentraeger, R., Elbert, T., Aldenhoff, J., Seeck-Hirschner, M., 2012b. Treatment of patients with borderline personality disorder and comorbid posttraumatic stress disorder using Narrative Exposure Therapy: a feasibility study. Psychotherapy and Psychosomatics 81 (1), 61–63.

Pabst, A., Schauer, M., Bernhardt, K., Ruf, M., Goder, R., Elbert, T., Rosentraeger, R., Robjant, K., Aldenhoff, J., Seeck-Hirschner, M., 2015. Evaluation of Narrative Exposure Therapy (NET) for borderline personality disorder with comorbid posttraumatic stress disorder. Clinical Neuropsychiatry 11 (4), 108–117.

Robjant, K., Fazel, M., 2010. The emerging evidence for Narrative Exposure Therapy: a review. Clinical Psychology Reviews 30, 1030–1039.

Ruf, M., Schauer, M., Neuner, P., Catani, C., Schauer, E., Elbert, T., 2010. Narrative Exposure Therapy for 7 to 16 year olds a randomized controlled trial with traumatized refugee children. Journal of Traumatic Stress 23, 437–445.

Ruf, M., Schauer, M., 2012. Facing childhood trauma: Narrative Exposure Therapy within a Cascade model of care. In: Murray, Jennifer (Ed.), Exposure Therapy: New Developments. Nova Science Publishers, New York, pp. 229–261.

Schaal, S., Elbert, T., Neuner, F., 2009. Narrative Exposure Therapy versus group interpersonal psychotherapy: a controlled clinical trial with orphaned survivors of the Rwandan genocide. Psychotherapy and Psychosomatics 78, 298–306.

Schalinski, I., Schauer, M., Elbert, T., 2014. The Shutdown Dissociation Scale (Shut-D) under review.

Schalinski, I., Elbert, T., Schauer, M., 2011. Female dissociative responding to extreme sexual violence in a chronic crisis setting: the case of Eastern Congo. Journal of Traumatic Stress 24 (2), 235–238.

Schauer, M., Elbert, T., 2010. Dissociation following traumatic stress: etiology and treatment. Journal of Psychology 218, 109–127.

Schauer, E., Elbert, T., 2010. The psychological impact of child soldiering. In: Matz, E. (Ed.), Trauma Rehabilitation after War and Conflict. Springer, New York, pp. 311–360.

Schauer, E., Neuner, F., Elbert, T., Ertl, V., Onyut, P.L., Odenwald, M., Schauer, M., 2004a. Narrative Exposure Therapy in children – a case study in a Somali refugee. Intervention 2 (1), 18–32.

Schauer, E., 2008. Trauma Treatment for Children in War: Build-up of an Evidence-based Large-scale Mental Health Intervention in North-eastern Sri Lanka.

Schauer, M., Neuner, F., Elbert, T., 2005. Narrative Exposure Therapy: A Short-term Intervention for Traumatic Stress Disorders after War, Terror, or Torture. Hogrefe & Huber Publishers.

Schauer, M., Neuner, F., Elbert, T., 2011. Narrative Exposure Therapy: A Short Term Treatment for Traumatic Stress Disorders, second ed. Hogrefe Publishing, Cambridge, MA.

Schauer, M., Neuner, F., Kamnakara, U., Klaschik, C., Robert, C., Elbert, T., 2003. PTSD and the "building block" effect of psychological trauma among West Nile Africans. ESTSS (European Society for Traumatic Stress Studies) Bulletin 10 (2), 5–6.

Schauer, M., Ruf-Leuschner, M., 2014. Die *Lifeline* – Zugang zur Narrativen Exposition (NET) in der Traumatherapie. Psychotherapeut 59, 226–238.

Schauer, E., Neuner, F., Elbert, T., Ertl, V., Onyut, L., Odenwald, M., Schauer, M., 2004. Narrative Exposure Therapy in children-A case study. Intervention 2 (1), 18–32.

Schauer, M., Elbert, T., Gotthardt, S., Rockstroh, B., Odenwald, M., Neuner, F., 2006. Wiedererfahrung durch Psychotherapie modifiziert Geist und Gehim. Verhaltenstherapie 16, 96–103.

Schauer, M., Elbert, T., Neuner, F., 2007. Interaktion von Neurowissenschaftlichen Erkenntnissen und Psychotherapeutischen Einsichten am Beispiel von Angst und traumatischem Stress. In: Becker, R., Wunderlich, H.-P. (Eds.), Wie wirkt Psychotherapie: Forschungsgrundlagen für die Praxis. Thieme-Verlag, Stuttgart/New York, pp. 87–108.

Schauer, M., Neuner, F., Elbert, T., 2005/2011. Narrative Exposure Therapy (NET). A Short-Term Intervention for Traumatic Stress Disorders, second ed. Hogrefe & Huber Publishers, Cambridge/Göttingen. ISBN 978-88937-388-4.

Schauer, M., Neuner, F., Elbert, T., 2014. Naratívna expozičná terapia. Bratislava, Martinus.

Schauer, M., Schauer, E., 2010. Trauma-focused Public mental-health interventions: a paradigm shift in humanitarian assistance and aid work. In: Martz, E. (Ed.), Trauma Rehabilitation after War and Conflict. Springer, New York, pp. 361–430.

Stenmark, H., Catani, C., Neuner, F., Elbert, T., Holen, A., 2013. Treating PTSD in refugees and asylum seekers within the general health care system. A randomized controlled multicenter study. Behaviour Research and Therapy 51 (10), 641–647.

Seng, J.S., Clark, M.K., McCarthy, A.M., Ronis, D.L., 2006. PTSD and physical comorbidity among women receiving Medicaid: results from service-use data. Journal of Traumatic Stress 19 (1), 45–56.

Teicher, M.H., Andersen, S.L., Polcari, A., Anderson, C.M., Navalta, C.P., Kim, D.M., 2003. The neurobiological consequences of early stress and childhood maltreatment. Neuroscience & Biobehavioral Reviews 27 (1), 33–44.

Teicher, M., Samson, J., Polcari, A., McGreenery, C., 2006. Sticks, stones, and hurtful words: relative effects of various forms of childhood maltreatment. American Journal of Psychiatry 163 (6), 993–1000.

Zang, Y., Hunt, N., Cox, T., 2013. A randomized controlled pilot study: the effectiveness of Narrative Exposure Therapy with adult survivors of the Sichuan earthquake. BMC Psychiatry 13 (41).

Zech, E., Vandenbussche, F., 2010. La thérapie par exposition à la narration de Schauer, Neuner et Elbert: Manuel de traitement de l'état de stress post-traumatique après la guerre, la torture et la terreur. Presses universitaires de Louvain.

Narrative in Social Work

Christopher Hall, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The 'narrative turn' has had a major impact across the social sciences, but has been slow to develop in social work research. This article outlines some of the major approaches to narrative theory and analysis in the social sciences, looking at narrative structure, identity, and performance in talk and interaction. Empirical research, which deploys narrative approaches to examine social work topics and narrative analysis is applied to research interviews, texts and everyday social work talk and interaction are identified. Implications for social work practice are also indicated.

Introduction

Narrative is intrinsic to social work. Stories are told to social workers by service users (Hall et al., 2006; Hyden and Overlien, 2005; Marvasti, 2002). Social workers tell stories to other professionals (Crepeau, 2000; White, 2002; Riemann, 2005) and write stories in files (Hall, 1997; Hall et al., 2006; White et al., 2009). Stories are told by service users and social workers to researchers (Hall, 1997; Martin, 1998; Wells, 2010). Narrative is increasingly featured in social work education (Rutten et al., 2010). Social workers are encouraged to think of their practice in narrative ways (Baldwin, 2013), with specific social work methods drawing on narrative (Parton and O'Byrne, 2000). Social work has a strong oral tradition (White, 2002) and social work agencies might be seen as 'storytelling organizations' (Boje, 1991). This article will examine what the study of narrative and storytelling can offer social work theory and practice.

While narrative theory and analysis has traditionally been associated with literary studies, over the last 30 years theory, research and practice drawing on narrative have flourished across the social sciences. Narrative approaches to data collection and analysis are now a constituent part of (mainly) qualitative research methods, some directly applying concepts from literary theory to talk and interaction in a wide variety of contexts. Many aspects of social life draw on, are constrained by, or constituted through narrative. Indeed, some authors (both literary and social science) suggest that it is our ability to make sense of our social world through storytelling, which makes us human. Roquentin says in Sartre's novel *La nausée* (1938: 64):

> 'a man is always a teller of tales, he lives surrounded by his stories and the stories of others, he sees everything that happens to him through them; and he tries to live his own life as if he were telling a story'.

Social psychologists, Gubrium and Holstein (1998: 163) suggest that "more and more, we are learning that life comes to us in the form of stories", and Jameson (1981: 13), a political scientist, talks of 'the all-informing process of narrative which I take to be the central function or instance of the human mind'.

However, the 'narrative turn' in the social sciences has been so far-reaching that writers discuss narrative from different and sometimes contradictory positions. These include: identifying narratives in everyday talk and interaction (e.g., Sacks, 1992); seeing narrative as located in the psychology of the individual 'the storied self' (e.g., McAdams, 1996); viewing narrative as a feature of wider community formations (e.g., Rappaport, 1995); and focusing the analysis on the text itself (e.g., Labov, 1972). While offering a wide range of possibilities for the researcher, it is often hard to compare writers' different theoretical foundations. Furthermore, the popularity of narrative has resulted in publications in which any form of talk that recounts past events, especially interview data, is referred to as a narrative, even when no narrative analysis is attempted (Atkinson, 1997; Riessman, 2008).

This article concentrates on how narrative contributes to social work, particularly research concerned with narrative structure, identity, and performance. First, I outline some of the key features of narrative theory, followed by a discussion of how data for narrative research analysis are collected and analyzed. I give particular consideration to narrative research that examines social work topics. Finally, I examine how narrative helps us better understand the users of social work and social work practices.

Definitions of Narrative

Although different disciplines emphasize different aspects, there are some commonalities. Narrative suggests a temporal aspect to talking, writing, and thinking usually with a beginning, middle, and end. Events are organized in such a way that the resulting narrative has consequences for both the narrator and listener/reader. It implies a move from one state of affairs to another, and in doing so, makes a point. Furthermore, the telling of the story is an occasion in which the narrator and listener attempt to communicate and make sense of their social world. As Riessman and Quinney (2005: 394) put it, "sequence and consequence: events are selected, organized, connected and evaluated as meaningful for a particular audience."

Some writers distinguish between narrative and story, especially in literary theory. In terms of etymology, narrative is

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28059-8

associated with an authoritative telling of events, whereas storytelling is associated with entertaining an audience, setting up an interesting challenge for the audience as to the truthfulness or deception of the story. In the social sciences, the use of these terms is less precise. Johnstone (2001: 639) makes the distinction between narrative as talk about a sequence of events, and story as the act of storytelling, whereas Linde (1993), in contrast sees narrative as a subunit of a person's life story. In this article, I use the terms interchangeably.

Narrative Theory

Some of the key concepts in narrative theory are time, identity and performance and are associated with the philosopher, Paul Ricoeur, although it is not suggested that all narrative research directly flows from his work. Ezzy (1998: 246) links the three concepts together in Ricouer's work:

> 'Narrative identity is coherent but fluid and changeable, historically grounded but 'fictively' reinterpreted, constructed by an individual but constructed in interaction and dialog with other people'.

Ricoeur develops the notion of time by examining the relationship between history and fiction; each draws on the other: "fiction is quasihistorical and history is quasifictive" (Ezzy, 1998: 243). When we recount the past, we inevitably draw on our interpretative imagination shaped by narrative. Ricoeur sees narrative as offering ways in which people make sense of events around them and turn those reflections into action: he asks, "in what way is the ordinary experience of time, borne by daily acting and suffering, refashioned by its passage through the grid of narrative" (1985: 249). The temporal characteristic of narrative is more than organizing past events. Storytellers give significance to those events for the present, which endure in ongoing narrative performances: "if we may speak of the temporal identity of a story, it must be characterized as something that endures and remains across that which passes and flows away" (Ricoeur, 1991: 22). Narrative provides structures for events to be interpreted and communicated. The key concept is 'emplotment': events that might appear unconnected, are taken from the teller's experience and synthesized into a unified and complete story. It is 'drawing configuration out of a succession' (Ricoeur, 1991: 22). Narrative theory influenced by structuralism, concentrates on the internal relations between the elements in the text (Propp, 1968). Ricoeur, however, takes a hermeneutic approach concerned with both the internal form and the lived experience of tellers and listeners.

For Ricoeur narrative enables us to gain a sense of ourselves, "life as an activity and passion in search of a narrative" (1991: 29). By creating and performing stories, we are able to create a coherent sense of who we are and at least for the time being we can create self-understanding, "Subjects recognize themselves in the stories they tell about themselves" (Ricoeur, 1988: 247). Narrative enables us to understand the world around us and also offers ways to act in and upon the world:

> 'Our life, when then embraced in a single glance, appears to us as the field of a constructive activity, borrowed from narrative understanding, by which we attempt to discover and not simply to impose from outside the *narrative identity which constitutes us*' (italics in original) (Ricoeur, 1991: 32).

Such an orientation has been developed by Bruner, "we have no other way to describing 'lived time' save in the form of narrative" (2004: 692). Bruner's work has been particularly important in narrative psychology and life story methods. He applies concepts from literary theory to distinguish three aspects of the story – theme, the main gist or moral of the story; discourse, the way it is told, in particular plot, language, and genre, the cultural conventions of storytelling. Bruner notes that not any story will do. To be successful, storytelling needs to draw on culturally available morals. It is the juxtaposing of (and most often resolving) these human dilemmas, what Bruner calls 'trouble' (2004: 697), which drives the drama of the story.

The third concept associated with narrative is performance, the story as told to an audience, and constructed in terms of 'ratified hearers' (Goffman, 1981). There is a tradition in literary theory for locating the understanding of the text in the interpretation of the reader (Iser, 1974), or as Fish (1980) calls it the 'interpretative community.' In the social sciences, the work of conversation analysts sees the story as constructed in the interaction between the teller and listener. When storytelling is examined in everyday conversation, the formal features of narrative are harder to identify. Edwards (1997) challenges attempts by Bruner and others to identify definitive elements of a narrative. As Goodwin and Heritage (1990: 299) say: "Rather than treating stories as self-contained cultural artifacts, conversation analysts have stressed the way stories are structured with respect to the contingencies of the interaction in which they are embedded."

Sacks (1992) analyzed stories in conversation as one form of talk-in-interaction. A central feature of everyday conversation is its organization in terms of turn-taking; interlocutors generally manage sequences of talk in terms of, for example, questions and answers, invitation and acceptance/rejection. To make space to tell a story in ongoing talk, the storyteller creates an agreement to talk for an extended period without an interruption. This requires a 'story preface,' an announcement of an intention or request to tell a story, for example, 'I must tell you what happened to me the other day.' The listener is therefore constrained from interrupting until the story is completed, for example, with a closing comment like, 'so that's what happened.' The orientation is to study "how stories are told and what conversational actions are accomplished in their telling" (Stokoe and Edwards, 2006).

Collecting Narrative Data

When researchers carry out empirical research with a narrative orientation, the distinction between temporal structure, identity and performance characterizes the collection and analysis of data.

For some, the focus is how people talk about themselves and see the research act as a storytelling occasion, in particular life story approaches. Here, the researcher interviews the respondent and organizes the questions in such a way that story is encouraged, by, for example, asking about past events, with prompts 'what happened next.' In her study of infertility among women in India, Riessman (2008: 31) describes and analyzes interviews in which women describe how their infertility affected their personal history, relationships, and social status. The narrative is the whole interview, which is seen as coconstructed between interviewer and respondent.

Others assume that the respondent already has a story waiting to be heard which they have developed over time to make sense of their circumstances. This is especially the case in illness narratives (Hyden, 1997). For example, Mattieson and Stam (1995: 284) see stories about cancer as different from ordinary talk, "they are more than just 'stories' but the vehicle for making sense of, not an illness, but a life." Others see narrative analysis concerned to display hidden or silenced stories of oppressed or marginalized peoples. Again, their story is assumed to exist outside of the research process (Rappaport, 1995: 799).

Those with an interest in interaction see the narrative as one feature of talk or writing. Narrative is identified as displaying particular persuasive and performative features, as opposed to, for example, question and answer sequences or descriptions. The analysis concentrates on the interactional practices involved in producing a story and the consequences for the ongoing talk. Horton-Salway (2001) examines how people with chronic fatigue syndrome/myalgic encephalomyelitis (CFS/ME) produce an accountable story of the onset of their condition, which is organized to resist versions that question it as a legitimate illness.

There are other versions of narrative. For example, shared narratives tell stories of a group or community, often sacred or mythical, which exist outside the individual or the interaction, but are available to be appropriated for personal identity and community interaction. For example, Humphreys (2000) examines narratives of Alcoholic Anonymous, which provides stories of hope as alternatives to dominant exclusion narratives. In summary, how researchers organize their data collection is strongly influenced by their definition of narrative – as ways of ordering the past, organizing a person's identity, or managing interactional dilemmas.

Narrative Analysis

In the analysis of narrative data, similar orientations apply. Researchers are interested in themes underlying the narrative, the structure of the story and their performance. Riessman (2008) identifies four approaches to analysis: thematic, structural, dialogic, and visual. Wells (2011) suggests narrative content, structure, and contexts.

Thematic

Analyzing the themes of what is said (or written) by respondents in a data corpus is the most widespread, but least distinctive approach to analysis. More traditional approaches to qualitative data analysis code: select themes that most represent the data; identify and compare instances across the data corpus. In contrast, narrative analysis is reluctant to break up the sequence of the narrative as each part is considered to be dependent on its location on the whole. In thematic approaches to narrative analysis, the researcher identifies themes in the way the respondents tell their story. As Riessman (2008: 53) says the analysis concentrates on "the case rather the component themes across the cases." For example, Dominelli et al. (2011) interviewed fathers involved in the child protection system. Their analysis produced a typology of different depictions of how the fathers described themselves, for example, 'misrepresented dad' or 'the mothering dad.' The analysis does not consider how the story is constructed or performed, but presents categories, which display overall themes regarding how fathers describe themselves in the interviews. Jansen and Haavind (2011) interviewed young people in residential care. They identified a distinction between stories characterized as 'If only' or 'Despite all.' They see such narratives as depicting the constraints of subjectivities, which are available in wider discourses to formulate their transition to adulthood.

Wells (2011: 44) considers thematic approaches are particularly appropriate for exploring social work issues, as they link respondents' talk to topics that social work might wish to explore. However, there is a danger that in analyzing narrative data as merely instances of self-evident themes, the complexity of narrative construction and performance is missed (Riessman, 2008: 76).

Structural

Structural approaches concentrate on how the narrative is organized, the form rather than the content. The influence of literary theory is evident with attempts to develop concepts like plot, character, genre, point of view in the analysis of talk and interaction. Narratives are considered to be organized in terms of beginning-middle-end. However, the key feature is how narrative is structured in such a way that it is consequential for the audience. The narrator needs to consider what aspects of the situation are included in the narrative and what are left out, and how the events fit into a plot, which the listener/reader can recognize. For the researcher, studying the underlying structures of the story can reveal the deeper meanings on which the narrator draws.

These approaches tend to concentrate on the narrative as a text, an artifact to be analyzed in itself rather than considering the narrator or storytelling occasion. The narrative is organized in such a way as to address an audience; sometimes specifically addressed and clearly identifiable; other times a general audience. In order to keep the audience involved, the narrative needs to address a number of tasks, for example, to make a point, be intelligible, not be repetitive, and have a sense of drama.

The most celebrated model for examining narrative structure is provided by the sociolinguist, Labov (1972). He was interested in how stories of African-American adolescents could be seen as having a coherent structure. He provides an ideal structure of a 'fully formed narrative,' based on the temporary organization of the clauses in the speech. There

are six elements: an abstract, which introduces the gist of the story to come; an orientation, which sets the scene; a complication, which introduces the problem or rupture in routine affairs; the evaluation, which indicates the significance of the event; the resolution, which deals with the problem and the coda, which rounds off the story. A verbal sequence of clauses in the story is matched to a sequence of events organized temporally. Such a structure provides both causality and credibility to the story, and is organized to counter what Labov calls the 'so what' question. The evaluation in particular is a direct address to the audience that events happened in the way they did and were (and are) important.

Riessman (2008) uses the Labovian model to understand how divorced people talk about relationship breakdown and make sense of their marital problems. She considers that the organization of events in the stories and the significance given to them uncover displays of distress which would not have been possible through thematic analysis. McDaniels (1995) examines how adults struggled to make sense of chaotic circumstances during their childhood by examining the use of various forms of evaluation in narratives to display the significance of events. Usita et al. (1998) uses narrative coherence to examine how people with Alzheimer's disease struggle to organize the chronology of their stories. In contrast to Labov, Gee (1991) provides a broader approach to examining narrative structure. He is concerned with identifying how parts or topics of long stretches of life story interviews fit together using the poetic feature of tone and voice, and displaying shifting temporal structures. His analysis identifies the narrative coherence in a life story told by woman suffering from schizophrenia, in what at first appears to be incoherent talk. Poindexter (2002) shows how different methods of addressing the structure of interview data (Labov, Gee, Mishler) uncovered different meanings. For example, by adding to Gee's model, a concern with the coconstruction aspects of the storytelling (for example, incomplete turns), more complex explanations emerge of the structure of the story, including affiliation between narrator and listener. Applying concepts of narrative structure provides important insights into how people describe and manage stressful situations by revealing the complexity of meanings on which narrators draw.

Performance

In contrast to studies of narrative structure, research on narrative performance concentrates on the telling of the story rather than the story itself. As Gubrium and Holstein (1998: 166) consider: "it casts storytelling as an ongoing process of composition rather than the more-or-less coherent reporting of experience." The evaluation clause in the Labov model implies an interruption in the temporality of the story to directly address the audience. Interactive approaches directly examine the ways in which stories are coconstructed, and their function in ongoing interaction. Storytelling is a practical activity produced for and with the listener to manage the dilemmas of the present encounter. Much of this work is concerned with narratives told in everyday situations, although the approach can also be applied to interview data by examining.

Everyday talk is likely to involve more complex methods of story preface and the negotiation of speaking turns. Conversational stories are rarely personal in the sense of belonging to a particular speaker or bounded by formal structures. Instead, they are coconstructed, contested and constituted in relationships, often characterized by challenges, questions, clarifications and speculations (Ochs and Capps, 2001; Crepeau, 2000). One story is often responded to with another story (Sacks, 1992) as the second narrator demonstrates his/her understanding of the significance of the first story. Likewise, the listener is likely to provide 'back-channeling' (Jefferson, 1984), comments like 'yeah' and 'mm' to encourage the storyteller to continue and responses of agreement or surprise at the denouement of the story. As Gubrium and Holstein (1998: 178) say "listeners are often active coparticipants in both the elicitation and production of stories, working with the machinery of ordinary conversation to shape storytelling."

Apart from negotiation of stories in turn-taking sequences, the storyteller is likely to include in the story features that may directly address the particular listener. These could be appeals to shared understanding of the salience of the story, 'you know what I mean?' They may also display their legitimacy to tell the story, the trusted teller of the tale (Shuman, 2006) by, for example, indicating their presence at the incident or expertise on the topic. Smith (1993) indicates how a student describes her friend's mental illness and provides preliminary instructions as how the story should be heard. For example, she states "I was the last to admit that she was becoming mentally ill," indicating that the mental illness was independent of her recognition, therefore, she was not complicit in its identification. Apart from addressing the potential 'so what' question, the storyteller faces a wide range of interactional challenges in terms of accountability, alignment and credibility, as in all communication (Buttny, 1993).

In institutional storytelling, the allocation of roles of narrator and listener are more likely to be established with particular kinds of stories following partially predetermined storylines. In counseling, welfare offices, medical consultations, stories are elicited with particular forms expected. In court, narratives are restricted in the sense that the storyteller is encouraged to tell the temporary flow of the events but prevented from indicating their significance. Gubrium and Holstein examine questioning by a district attorney of a potentially involuntary psychiatric patient. The district attorney starts a question-and-answer sequence, but then refuses to take the next allotting turn, thereby forcing the potential patient to extend and elaborate her story. As Gubrium and Holstein (1998: 177) summarize "narrative production is necessarily collaborative, even while it is institutionally informed."

Narrative in Social Work

This article has examined narrative theory and methods in general. We now look at research on social work topics, both the stories of service users and social work practices.

Riessman and Quinney (2005) carried out a literature review on narrative in social work and found four main

groups. The first and the majority of the papers were practice-oriented, specifically clinical. They were often case-based and explored how clients' stories were addressed therapeutically. There were also examples of self-narratives of client groups. However, they found that narrative theory was rarely applied to the analysis. The second group was narratives in pedagogy and how students write about their work. The third group was autobiographical accounts, particularly related to reflective approaches. The fourth group was research papers using narrative approaches. Overall the authors were disappointed in the research corpus, both the relatively small numbers and the lack of explorations of narrative theory and methods.

Despite the relative lack of significant exploration of narrative in social work research and theory, narrative can inform social work in a number of important ways. Versions of personal and social identity as constructed through narrative provide a major challenge to traditional notions of the individual in social work. Wilks (2005) provides a strong argument for the significance of narrative ethics, which resists the essentialism associated with social work views of people with fixed positions and definable characteristics. Narrative sees people as becoming and unfinalized and moral reasoning as 'narratively shaped.' It also challenges practices based on principles. He provides an example of 'choice.' In itself, the principle of choice for service users does not help to guide action if different stakeholders have different views of what is the right choice for the service user.

> '[In dilemmas of choice] … we are actually presented with are a series of intersecting stories about a situation. These stories look to both the future and the past, and offer competing outcomes based on different histories. We work with the stories, juxtaposing different accounts – the user's, the worker's, the manager's – looking for commonalities, points where the stories coincide, placing one story with another, the history and the outcome for an immediate choice played out within the larger narrative of a service user's life'
> (Wilks, 2005: 1256).

Interestingly, this contradicts Walsh (2009: 274) who suggests that 'the ultimate value of narrative practice is *empowering* clients' (italics in original). For Wilks (2005: 1260):

> '… the idea of moral sensibility [moves] away from the universalism of the categorical imperative or pursuit of the good life into a localized, temporal and changing sphere, much closer to our experience of the moral world'.

Understanding Clients

A number of researchers have investigated the narratives of the social work clients, using in-depth interviews to provide deeper understandings of the client. Wells (2011) analyses the story of a mother whose child was placed in care and then returned home. She examines how the mother constructs a narrative identity on the basis of the "fundamental binary opposition of the absent frightening mother versus the present protective mother" (p. 446). This study displays the complex and stigmatized identities with which mothers struggle, and are frequently suppressed by the child welfare system. Usita et al. (1998) found that people with Alzheimer's disease were often able to produce 'comprehensive and richly textured life stories' with the help of the research interviewer. They suggest that "it is conceivable that the salient events that were recalled, though limited, were enough to maintain some sense of self" (p. 194).

Martin (1998) describes a project in which young people leaving care were encouraged to tell their story, which the researcher typed (or 'scribed') in their presence. In a second session, they were encouraged to review and edit the story and then share it in a focus group with other participants. Such narrative construction highlighted the dynamic nature of narratives, not only producing a story but reflecting on it and having it questioned. Martin considers that in the process, the young people "found themselves in the telling, experienced themselves as creating themselves and as recovering themselves from the stories that had been told them" (p. 9). In mental illness, researchers examine the importance of narratives of psychotic patients to understand recovery. Roe and Davidson (2005: 94) suggest that narrative offers the opportunity to author "a new story. The process of authoring, in turn, helps consolidate and integrate a sense of self."

All these studies display the complexity of narrative and suggest that the study of narratives can lead to more sympathetic portrayals of social work service users. The view of shifting rather than fixed stories is of particular importance as it questions the requirements of social work systems to produce fixed versions of clients. Such an orientation links to studies of social work interaction in which versions of clients are seen as interactionally produced and managed (e.g., Hall et al., 2013).

Understanding Practices

Work gets done in social work by telling and evaluating stories. Service users approach social services agencies often with stories about their problems and how the agency can help. Marvasti (2002) examines how social workers in a homeless agency examine clients' accounts for their 'service-worthiness,' by editing the story in terms of the institution's mission. Urek (2005) also studies how social workers construct stories to produce a particular, characterization of clients to justify decisions and actions. She follows how a case is constructed and performed as a narrative in various meetings, and how it becomes the institutional narrative to establish the social services' version of the mother and child's character, which appears in a court report. Alternative narratives are undermined and disregarded.

Team and supervision meetings are particularly important occasions for the telling of stories. Pithouse and Atkinson (1988) examine how narrative enables the social worker to produce and justify the invisible craftwork of their trade to colleagues and supervisors. Riemann (2005) looks at social work case discussions and in particular the balance between narrative and argumentation. He notes how when social workers describe difficulties in making sense of their cases, the complexity of the stories are lost as colleagues narrow down the evaluation. Crepeau (2000) examines how narratives are told in multiprofessional meetings.

Professionals use stories of particular interactions with services users to evidence their portrayals to colleagues. How the story is told and how it is supported is crucial in informing subsequent decision-making. These studies examine the performance of the narratives in social work practices but make connections with institutional practices more generally.

Future Developments

Recent publications promote the exciting developments of narrative in social work (Larsson and Sjőblom, 2009; Wells, 2011; Baldwin, 2013). While the narrative turn has had a major impact across the social sciences, especially in medicine (Hyden, 1997), social work has yet to exploit the potential of narrative theory and analysis. For a profession concerned almost exclusively with talk and interaction, it is surprising that so little attention is paid to what happens in everyday social work (Hall and Slembrouck, 2009). Developing narrative approaches within practice intervention is particularly interesting. A range of approaches to narrative therapy have developed (White and Epston, 1990), some of which are applied to social work (Parton and O'Byrne, 2000; Roscoe et al., 2011).

There are, however, a number of criticisms of the popularity of narrative. Riessman (2008: 5) notes that 'narrative is everywhere but not everything is narrative.' Stokoe and Edwards (2006: 57) are particularly critical of biographical approaches, which they describe as "a contrived situation based on the assumption that their 'life stories' and experiences are readily available to 'dump' from memory." Atkinson (1997: 327) sees narrative research as a "mixed blessing … often sentimental and romantic," and warns against privileging the narrative as a special kind of representation. Woods (2011: 401) questions the instruction of the narrative researcher: "Tell us your story, because it is true (to the human condition), because it is yours (an authentic expression of your individual experience), and because it is good for you (as part of the healing process)."

See also: Social Work Education; Social Work Theory.

Bibliography

Atkinson, P., 1997. Narrative turn or blind alley? Qualitative Health Research 7 (3), 325–344.

Baldwin, C., 2013. Narrative Social Work: Theory and Application. Policy Press, Bristol.

Boje, D., 1991. The storytelling organization: a study of performance in an office supply firm. Administrative Science Quarterly 36 (1), 106–126.

Bruner, J., 2004. Life as narrative. Social Research 54 (1), 691–710.

Buttny, R., 1993. Social Accountability in Communication. Sage, London.

Crepeau, E., 2000. Reconstructing gloria: a narrative analysis of team meetings. Qualitative Health Research 10 (6), 766–787.

Dominelli, L., Strega, S., Walmsley, C., Callahan, M., Brown, L., 2011. Here's my story: fathers of 'looked after' children recount their experiences in the Canadian child welfare system. British Journal of Social Work 41 (2), 351–367.

Edwards, D., 1997. Discourse and Cognition. Sage, London.

Ezzy, D., 1998. Theorizing narrative identity: symbolic interactionism and hermeneutics. The Sociological Quarterly 39 (2), 239–252.

Fish, S., 1980. Is There a Text in This Class? The Authority of Interpretative Communities. Harvard University Press, Cambridge, Massachusetts.

Gee, J., 1991. A linguistic approach to narrative. Journal of Narrative and Life History 1 (1), 15–39.

Goffman, E., 1981. Forms of Talk. Basil Blackwell, Oxford.

Goodwin, C., Heritage, J., 1990. Conversation analysis. Annual Review of Anthropology 19 (1), 283–307.

Gubrium, J., Holstein, J., 1998. Narrative practice and the coherence of personal stories. The Sociological Quarterly 39 (1), 163–187.

Hall, C., 1997. Social Work as Narrative: Storytelling and Persuasion in Professional Texts. Ashgate, Aldershot.

Hall, C., Juhila, K., Matarese, M., van Nijnatten, C., 2013. Analysing Social Work Communication: Discourse in Practice. Routledge, London.

Hall, C., Slembrouck, S., Sarangi, S., 2006. Language Practices in Social Work: Categorisation and Accountability in Child Welfare. Routledge, London.

Hall, C., Slembrouck, S., 2009. Communication in child welfare: skills, language and interaction. Child & Family Social Work 14 (4), 461–470.

Horton-Salway, M., 2001. Narrative identities and the management of personal accountability in talk about me: a discursive psychology approach to illness narrative. Journal of Health Psychology 6 (2), 247–259.

Humphreys, K., 2000. Community narratives and personal stories in alcoholics anonymous. Journal of Community Psychology 28 (5), 495–506.

Hyden, L., 1997. Illness and narrative. Sociology of Health and Illness 19 (1), 48–69.

Hyden, M., Overlien, C., 2005. Applying narrative analysis to the process of confirming or disregarding cases of suspected sexual abuse. Child and Family Social Work 10, 57–65.

Iser, W., 1974. The Implied Reader. John Hopkins University Press, Baltimore.

Jameson, F., 1981. The Political Unconscious: Narrative as a Socially Symbolic Act. McMillan, London.

Jansen, A., Haavind, H., 2011. 'If only' and 'despite all': narrative configuration among young people living in residential care. Narrative Inquiry 21 (1), 68–87.

Jefferson, G., 1984. Notes on a systematic deployment of the acknowledgement tokens 'yeah' and 'mm hm'. Papers in Linguistics 17 (2), 197–216.

Johnstone, B., 2001. Discourse analysis and narrative. In: Schiffrin, D., Tannen, D., Hamilton, H. (Eds.), The Handbook of Discourse Analysis. Blackwell, Oxford.

Labov, W., 1972. Language in the Inner City: Studies in the Black English Vernacular. University of Pennsylvania Press, Pittsburgh.

Larsson, S., Sjőblom, Y., 2009. Perspectives on narrative methods in social work research. International Journal of Social Welfare 19 (3), 272–280.

Linde, C., 1993. Life Stories: The Creation of Coherence. Oxford University Press, Oxford.

Martin, F., 1998. Tales of transition: self-narrative and direct scribing in exploring care-leaving. Child and Family Social Work 3 (1), 1–12.

Marvasti, A., 2002. Being constructing the service-worthy homeless through narrative editing. Journal of Contemporary Ethnography 31 (5), 615–651.

Mattieson, C., Stam, H., 1995. Renegotiating identity: cancer narratives. Sociology of Health and Illness 17 (3), 283–306.

McAdams, D., 1996. Personality, modernity and the storied self: a contemporary framework for studying persons. Psychological Inquiry 7 (4), 295–321.

McDaniels, S., 1995. Identity Construction Through Narrative: The Impact of Chaotic Environments and Negative Affective Experiences of Childhood. Doctoral thesis, University of Chicago. http://www.icsw.edu/_dissertations/mcdaniels_1995.pdf.

Ochs, E., Capps, L., 2001. Living Narrative: Creating Lives in Everyday Storytelling. Harvard University Press, Cambridge, Massachusetts.

Parton, N., O'Byrne, P., 2000. Constructive Social Work: Towards New Theory and Practice. Palgrave Macmillan, London.

Pithouse, A., Atkinson, P., 1988. Telling the case: occupational narrative in a social work office. In: Coupland, N. (Ed.), Styles of Discourse. Croom Helm, Beckenham, pp. 183–200.

Poindexter, C., 2002. Meaning from Methods: re-presenting narratives of an HIV-affected caregiver. Qualitative Social Work 1 (1), 59–78.

Propp, V., 1968. Morphology of the Folk Tale. University of Texas, Austin.

Rappaport, J., 1995. Empowerment meets narrative: listening to stories and creating settings. American Journal of Community Psychology 23 (5), 795–808.

Ricoeur, P., 1985. Narrated time. Philosophy Today 29 (4), 259–272.

Ricoeur, P., 1988. Time and Narrative, vol. 3. University of Chicago Press, Chicago.

Ricoeur, P., 1991. Life in quest of narrative. In: Wood, D. (Ed.), On Paul Ricoeur: Narrative and Interpretation. Routledge, London.

Riemann, G., 2005. Trying to make sense of cases: features and problems of social workers' case discussions. Qualitative Social Work 4 (4), 413–430.

Riessman, C., 2008. Narrative Methods for the Human Sciences. Sage, London.

Riessman, C., Quinney, L., 2005. Narrative in social work. Qualitative Social Work 4 (4), 391–412.

Roe, D., Davidson, L., 2005. Self and narrative in schizophrenia: time to author a new story. Medical Humanities 31 (2), 89–94.

Roscoe, K., Carson, A., Madoc-Jones, L., 2011. Narrative social work: conversations between theory and practice. Journal of Social Work Practice 25 (1), 47–61.
Rutten, K., Mottart, A., Soetaert, R., 2010. Narrative and rhetoric in social work education. British Journal of Social Work 40 (2), 480–495.
Sacks, H., 1992. Lectures on Conversation, vol. I. Blackwell, Oxford.
Sartre, J.-P., 1938/1996. La Nausea. Gallimard, Paris.
Shuman, A., 2006. Entitlement and empathy in personal narrative. Narrative Inquiry 16 (1), 148–155.
Smith, D., 1993. Texts, Facts and Femininity: Exploring the Relations of Ruling. Routledge, London.
Stokoe, E., Edwards, D., 2006. Story formulations in talk-in-interaction. Narrative Inquiry 16 (1), 56–65.
Urek, M., 2005. Making a case in social work: the construction of an unsuitable mother. Qualitative Social Work 4 (4), 451–467.
Usita, P., Hyman, I., Herman, K., 1998. Narrative intentions: listening to life stories in Alzheimer's disease. Journal of Aging Studies 12 (2), 185–197.
Walsh, J., 2009. Theories for Direct Social Work Practice. Thomson, Brooks-Cole Publishing, Belmont, CA.
Wells, K., 2010. A narrative analysis of one mother's story of child custody loss and regain. Children and Youth Services Review 33 (1), 439–447.
Wells, K., 2011. Narrative Inquiry. Oxford University Press, Oxford.
White, M., Epston, D., 1990. Narrative Means to Therapeutic Ends. W.W. Norton, New York.
White, S., 2002. Accomplishing a case in paediatrics and child health: medicine and morality in inter-professional talk. Sociology of Health and Illness 24 (4), 409–435.
White, S., Hall, C., Peckover, S., 2009. The descriptive tyranny of the common assessment framework: technologies of categorization and professional practice in child welfare. British Journal of Social Work 39 (7), 1197–1217.
Wilks, T., 2005. Social work and narrative ethics. British Journal of Social Work 35 (8), 1249–1264.
Woods, A., 2011. Post-narrative: an appeal. Narrative Inquiry 21 (2), 399–406.

Narrative Networks

Katherine Stovel and Daniel Koski-Karell, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Narrative networks are a method for representing narrative data that facilitates systematic comparison and interpretation of qualitative accounts of social phenomena. The basic insight behind the narrative network approach is that it is possible to break a narrative into 'elements' and treat these elements as nodes in a network, where links between nodes represent either causal or temporal connection between nodes. Such networks can then be analyzed using standard network measures such as density, centrality, or connectivity. Sometimes the structure of the narrative network will be revelatory in and of itself – as in studies of identity formation – while in other cases narrative accounts can be compared or integrated.

Introduction

Over recent decades, social scientists have increasingly looked to narratives as a source of insight into social and historical processes. This 'narrative turn' (Polletta et al., 2011) has taken a variety of forms, but one important thread produced a series of strategies designed to systematically or formally represent the structure of narratives. Several of these approaches involve applying logical rules to a reduced or abstracted form of a narrative in order to reveal the underlying cause of an event (see Abell, 1987; Corsaro and Heise, 1990). In contrast, narrative networks are a method for representing narrative data that facilitates systematic comparison and interpretation of qualitative accounts of social phenomena. Rather than seeking to determine *why* something happened, the strength of the narrative networks approach is that it distills for the analyst *how the author of the narrative understood the event* (Gerteis, 2002). Sometimes this will be revelatory in and of itself – as in studies of identity formation – while in other cases narrative accounts can be compared or integrated, thereby revealing the extent to which multiple accounts mirror one another.

The basic idea behind the narrative network approach is that it is possible to break a narrative into 'elements' and treat these elements as nodes in a network, where links between nodes represent either causal or temporal connections between nodes. Such networks can then be analyzed using standard network measures such as density, centrality, or connectivity. Introduced in Bearman et al. (1999) and Bearman and Stovel (2000), the approach has been used to study a diverse range of substantive problems, including the local meaning of revolution and counter-revolution (Bearman et al., 1999), various types of personal life histories (Bearman and Stovel, 2000; Smith, 2007), the use of technology (Pentland and Feldman, 2007), and accounts of rape (Quinlan and Quinlan, 2010).

The narrative network approach extends a long tradition of using the lens of network analysis – with its focus on relational data and social structure – to analyze historical events from a social scientific perspective. The challenge of historical social science has always been moving beyond particularistic explanations while still retaining the salience of context. In response, scholars have used the narrative network approach to emphasize the function of particular types of events and produce comparable measures of the structural features of narrative, allowing for rich and systematic analysis of important social processes (Bearman et al., 1999: p. 509).

In the following article (Video Cultures: Television), we describe this approach, emphasizing narrative networks' dual status of theoretical orientation and methodological technique. First we briefly discuss the nature of narratives, highlighting what they tell us about a narrator's identity and perspective as well as how different types of narrative may function as data. Next, we introduce a basic method for analyzing narratives as networks, showing how structure can illuminate meaning. The section that follows traces significant developments in the approach over the past decades, including advancements in how to compare narratives and analyze different forms of discourse. Finally, we discuss several contributions and limitations of the approach.

What Is a Narrative?

Narratives are a particular kind of data, profoundly distinct from the abstracted variables that dominate most quantitative approaches to social science. Narratives offer an "account of a sequence of events in the order in which they occurred to make a point" (Polletta et al., 2011: p. 111). Like stories, narratives have characters (or, as Pentland and Feldman, 2007 call them, *actants*), who can both act and be acted upon (Gerteis, 2002; Polletta et al., 2011). But beyond characters, narratives weave events together into a plot; the choice of which events out of many to include – and how they are connected – reveals much about what point the author is trying to make.

Everyday, our lives are filled with a multitude of events: we wake up, we get up, we brush our teeth, perhaps we make and drink coffee, and some days we take a train to work. Most of these events are background to the story of our lives, though once we know what is important or significant or worthy of attention, some elements may rise to the foreground of a particular narrative. If one were living in Madrid on 11 May 2004, the fact that one took the train to work could turn out to be a very significant event in the story of a city terrorized by a bombing. In constructing a narrative, an author decides not only what to include, but also makes claims about how the elements of the narrative are related to one another. These claims could be causal (e.g., "because I took a later train, I arrived at the station after the bombing") or temporal ("after I

arrived at the station, I observed the carnage"). Taken together, causal and temporal connections create the plot – or theory – of the narrative. An account without causal or temporal links between elements is simply a report, with no narrative beginning, middle, or end (Bearman and Stovel, 2000; cf Polletta et al., 2011).

Types of Narratives

One important type of narrative is the *life story*. Told at a particular moment in time, a life story is first-person account of a life. Some life stories, those that take the form "I was born, I went to school, I married, I enjoy bicycling…" have little narrative, in the sense that they have no plot, and therefore no point. In contrast, a *narrative life story* is an account of how a person came to be the person she believes she is. In narrative life stories, the selection of events in the life, and the temporal and causal claims about how they are related to one another, provide insight into the author's theory of self. While narrative life stories are revelatory of the theory of self, an author's ability to tell her own life story may be constrained by her ability to express herself, her educational background, or the cultural or historical goods she has to work with.

Some narratives, including some narrative life stories, orient toward a particular *script* (Polletta et al., 2011; Gerteis, 2002). Scripts can be produced and reinforced by institutional structures as well as cultural expectations (Polletta et al., 2011: p. 112). When institutions are strong, they channel actors through particular social settings and provide an interpretation of events that can be reflected in a narrative. Similarly, culture, in the sense of shared symbolic meanings, offers cues about how things work and what they mean. Both institutions and culture therefore may suggest the plot of a narrative, regardless of the particular elements the author has to choose from. For instance, in contemporary Western societies, a culturally expected account of how one came to be married might involve details of how two people 'happened' to find each other, and immediately – or slowly – recognized their shared and deep attraction. This expected script differs dramatically from the accounts one would find in societies dominated by institutions of arranged marriages, where the actions, preferences, and interests of the couple would play a minimal role in the narrative. Religious and therapeutic organizations often encourage their members to adopt a particular narrative form to describe their journey toward transformation, e.g., "I was lost, and then I was found." By embracing an accepted narrative structure – and with it the associated theory of how things come to be – an actor can signal to others that she views herself as a member of a particular community (Polletta et al., 2011: p. 118). Yet while conforming to an institutionally or culturally supported script can convey belonging, deviating from expectations about how something came about can also produce powerful narratives. In a collectivist society like Japan, telling a life story that attributes success to hard work and personal accomplishment might shock and confuse (or inspire!) listeners.

Whether a narrative recounts a life history or some other event, it is often possible to find stories told from *multiple perspectives*. The existence of multiple accounts of the same or related phenomena is one of the factors that make narratives of interest to social scientists. While individuals might always be interested in the current version of their life story, it is the possibility of integrating or comparing multiple accounts of an event or process – either told by different people, or at different points in time – that allows us to move beyond the realm of journalism and yield insight into general features of social and political life. Analysis of multiple related narratives may reveal the existence of a normative script (Quinlan and Quinlan, 2010; Gerteis, 2002), may help define the boundaries of an event (Bearman et al., 1999), or may shed light on the process by which individuals come to share a sense of belonging to a common community (Smith, 2007; Lee et al., 2010).

Narrative Networks

The narrative network approach was developed by Peter Bearman and introduced in a short series of papers (Bearman et al., 1999; Bearman and Stovel, 2000). Narrative networks represent a method for distilling the *structure* of a narrative in order to recover its meaning. The approach rests on recognition of a fundamental similarity between narratives and networks, in the sense that both are characterized by distinctive patterns of connection between elements or events. A narrative is distinctive not just because of the particular set of elements it includes, but because of the particular connections the teller imposes on those elements. Similarly, a network is defined not just by who is in it, but also by the pattern of connections that link (or fail to link) the various actors. As Smith puts it, "It is a narrative's structure that allows us to analyze functions and patterns among constitutive units both within and across narratives" (Smith, 2007: p. 23). The idea that one can represent narrative text as a network or graph opens up a whole range of analytic tools that allow us to systematically summarize and compare extended narratives (Mckether et al., 2009).

To construct a narrative network, the scholar identifies each discrete element of the narrative and codes these elements as *nodes* in a network. The discrete elements can reflect persons, places, and things, author's realizations, or other details included in the narrative (Bearman and Stovel, 2000; Lee et al., 2010: p. 3). In most applications of narrative networks, nodes are coded inductively, though Shostak and Fox (2012) consider the advantages of coding for themes identified a priori. Pairs of nodes are connected by a directed link (an *arc)* whenever the author of the narrative imposed a narrative clause between the elements – where a narrative clause is a statement "that is temporally ordered in such a way that moving it involves changing the meaning of the sub-sequence in which it is embedded" (Bearman et al., 1999: p. 512). Thus arcs connect elements in the story only when the author imputes a causal or cognitive connection between them; purely temporal orderings are typically not coded links between nodes.

Bearman and Stovel (2000) offer examples of the process they used to code the life histories of Germans who joined the Nazi party before 1930. For instance, they provide the following segment of narrative text:

> When I was sixteen years old my mother bought me some brown trousers. If she had known the significance of this color, she would

never have bought them for me. I told her that they pleased me, and she also bought some for my brother… Now I felt as though I were half a Nazi." (Bearman and Stovel, 2000: p. 76).

They then transform this text into a set of unique elements, including:

- Mother bought brown trousers.
- Trousers pleased me.
- Told mother the trousers pleased me.
- Mother bought trousers for brother.
- …
- Felt like 'half a Nazi.'

Once the entire story is so coded, the narrative is transformed into an adjacency matrix, where the rows and columns represent elements in the story and cell entries represent narrative connections between nodes. After a narrative has been represented as an adjacency matrix, it is relatively trivial to analyze the narrative formally as a network. Various network measures, including reach, density, component size, and centrality can provide a summary of the structural properties of the network (Wasserman and Faust, 1994).

Following the practice of network analysis, an adjacency matrix can also be visualized as a graph; this graphical representation – which we can think of as a map of the narrative – reflects the structural connections between elements of the narrative. An illustrative graphical representation of a narrative account of becoming a Nazi is shown in **Figure 1**.

In Bearman and Stovel's study of accounts of becoming a Nazi, the primary analysis focuses on comparing the local density, centrality, and narrative reach of elements in the first and second half of the narrative. Narrative plot, they argue, is "locally dense, often cyclic, knotted, and characterized by a redundancy of ties" (Bearman and Stovel, 2000: p. 71). Thus they believe that an author's account of how something came to be (in this case, how a person became a member of the NSDAP) will involve dense subsequences of mutually reinforcing events – a structure of connectivity that is easy to detect by measuring local network density or node-specific reach. "By focusing on [betweenness] centrality, narrative process at the level of the element context is revealed. This process can, and does, differ from the explicit theory proposed by the author, for element centrality may be invisible in the unfolding narrative, as the most important nodes are those that bridge subsequences. … The structure of a narrative network provides insight into a motive, without exogenous imputation" (Bearman and Stovel, 2000: p. 88).

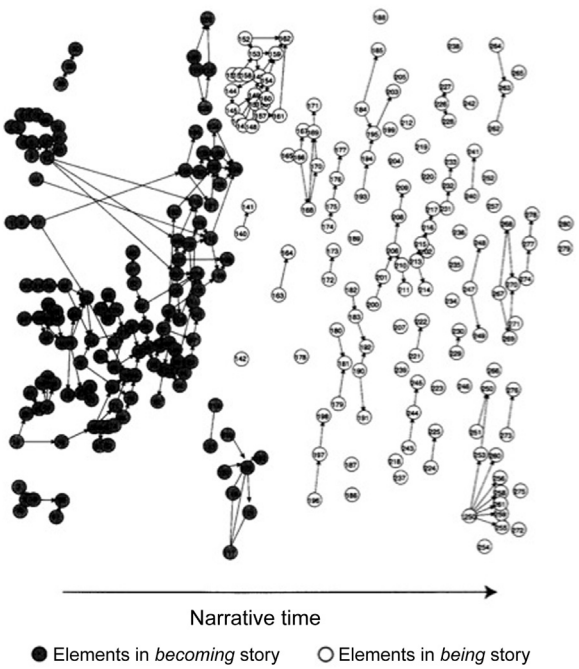

Figure 1 Graphical Representation of a Narrative Network (Figure 2 in Bearman and Stovel 2000).

While the process of identifying nodes and determining the presence of arcs may appear straightforward, many complexities arise, particularly when the aim is to integrate multiple narrative voices, or to compare across narratives. This issue is addressed in different ways in different studies; one aspect of Bearman and Stovel's approach is to classify nodes into categories, and examine the pattern of ties within and between categories (see also Gerteis (2002) for a similar approach). They begin by differentiating nodes based on the function of the element in the narrative. In their study of the life histories of early Nazis, they classify each node as either a macrolevel event (*M*), a local event (*L*), or a cognition (*C*).

Macro-level elements are elements noted in the narrative that occur outside the author's local world. In this narrative, macro-level elements include the 'great war', the putsch, and rising inflation. Local event elements are ordinary events occurring in the author's local world. These are events that are experienced directly. In our illustrative story local elements include experiencing a childhood illness, wearing a swastika to school, and taking a bicycle trip with his sister. Cognitions are aspects of the narrative in which the author questions the meanings of elements, draws conclusions, or reveals a new cognitive understanding about his situation. Cognition elements include 'communists and Hitler both want power', 'recognizes the beauty of nature', and 'felt as if half a Nazi.' (Bearman and Stovel, 2000: p. 83)

Analysis of the distribution and pattern of connection between *M*, *L*, and *C* elements reveals the important role cognitions play in connecting and animating local events during the process of becoming a Nazi. Beyond the function of events in the narrative, Bearman and Stovel also classify events in terms of their content – a strategy that others adopt in later work. In the case of the life stories of early Nazis, events are coded into a series of concepts or themes (see also Mckether et al., 2009) that reflect factors others have identified as important for high-risk mobilization processes. These categories include chaos, exposure to other Nazis, anti-Semitism, order, and the elision of social relations.

Bearman and Stovel's analysis of life histories as networks reveals a profound difference between the structure of the narrative that described the process of *becoming* a Nazi and the portion that described the author's life as a Nazi (the *being* story). While the *becoming* narrative is composed of complex,

dense, and connected subsequences of elements involving cognition, the *being* story is sparse and relatively unconnected. Things just happen, but there is no narrative glue stitching them together. Bearman and Stovel argue that this sparse narrative structure is characteristic of the adoption of a total identity: it is as if for such actors, history stops, and they live their life with the zeal of one who embraces an ideology that answers all questions for them. "Nothing is left to operate the story, except for a mechanical agency…. The authors of the stories we consider can continue to report the facts, but nothing beyond the facts. This is not because there is not a story to tell: they explicitly select facts to tell a story, but it is a banal story of action without actors" (Bearman and Stovel, 2000: p. 89).

Extensions of the Method

Integrating Multiple Narratives

In another early example of narrative networks, Bearman et al. (1999) analyze the relational structure of multiple narratives. Rather than focusing on the processes of identity formation, Bearman and his coauthors use the life stories of Chinese villagers from one small village to better specify the bounds of a historical incident, thereby aiding in the interpretation of the events within the case. This study, despite its dissimilar aim, complements Bearman and Stovel by highlighting how narrative networks help reveal the meaning hidden in personal histories.

Bearman et al. (1999) use narrative networks to address a significant obstacle in historical analysis: the problem of the future. As the authors put it, "only when we can provide the beginning and an end to a sequence of interrelated events can we understand the meaning of an event within the sequence and, by extension, the meaning of the event as a whole" (Bearman et al., 1999: p. 502). Yet, how are we to know when a historical incident ends? In considering this question, these scholars exploit the parallels between narratives and networks, arguing that demarcating the beginning and end of a historical incident, or 'casing' the incident, is similar to specifying the boundary of a network. The boundaries then tell us about the historical incident's fluidity, or its ability to be conditioned or activated by future events. The theoretical argument is that the more redundant or interdependent the relationships between the constitutive events within the boundaries, the less likely it is that the meanings attributed to those events will change.

Bearman et al.'s analysis is innovative in that it integrates multiple narratives describing the 'same' historical event into a master narrative. Using a strategy similar to that described in Bearman and Stovel (2000), they first code the historical events contained within 14 different life stories as nodes. They then assign ties between nodes connected by narrative clauses or clauses that temporally connect events within a narrative. Yet because their data include different narrators describing the same events, the structure of paths between nodes varies, depending on how each narrator understood the historical events. At this stage, Bearman et al.'s approach diverges from the Bearman and Stovel method: rather than analyzing each narrative separately, Bearman et al. build an integrated network via the nodes shared across narrative. In this new narrative network, nodes are connected if there is a tie between them in any of the original networks. Thus the final narrative network captures how the 14 narratives articulate with one another.

The key step in the analysis of casing is identifying a 'bicomponent' in the interrelated network. This is a component, or maximal connected subgraph, where "all nodes are connected by at least two independent paths and where the addition of a node requires that it is connected to two nodes in the subgraph" (Bearman et al., 1999: p. 517). A bicomponent, the authors argue, provides a candidate for a 'cased' historical sequence. If a researcher simulates new networks based on the integrated narrative, and finds that the bicomponent is highly resilient to perturbations, it is reasonable to conclude that the meaning of the events constituting the bicomponent are historically stable, in the sense that the meaning is unlikely to be changed by future events. Researchers are more likely to agree on the bounds of the case and, therefore, more likely to agree on the meaning of events within the case. In contrast, finding fragile networks reduces our confidence in the meanings ascribed to the events. This, in itself, is an important insight to how we interpret networks. While many analysts specify the boundaries of networks by finding partitions between clusters, Bearman et al. note that these connecting arcs may in fact represent important links in a historical case.

Aggregation and Comparison

In a fascinating comparative study of ethnic identities among ethnically Italian and Croatian Istrians, Smith (2007) further develops narrative network analysis by introducing a new strategy for comparing multiple, and sometimes contradictory, narratives. Smith draws on over 200 narratives collected from Istrians living in Italy, Croatia, and New York City to explore "the ambiguous space between two hardened social identities" (Smith, 2007: p. 23). Her analysis shows how narratives of ethnic violence have helped maintain sharp social boundaries between Italian and Croatian Istrians living in Europe, while the narratives of Italian and Croatian Istrians who moved to New York City have blurred ethnic boundaries and reveal the emergence of a shared Istrian ethnic identity. The goal is "to produce an understanding of how meaning emerges from the structure of relations among narrative elements, just as social meaning emerges from the structure of social ties" (Smith, 2007: p. 24).

Smith's key methodological contribution is a method that allows her to identify abstract elements (themes) that are likely to play a role in multiple narratives. Following Gerteis (2002), Smith seeks to identify shared, or collective, narratives, which "emerge when regularly occurring plots connect to key characters in an empirically stable way" (Gerteis, 2002: p. 593). Smith first recognizes that events can be nested within other events and, when grouped and sequenced, these groups of events – which often include a combination of local and cognitive events – form aggregated *narrative elements* (e.g., notions of liberation, victimhood, relocation). She explains, "Aggregate elements are defined by sequences of events that contribute to the element's meaning; an understanding of how aggregate elements are tied together requires an understanding of the array of relations among the events that compose the aggregate element" (Smith, 2007: p. 26). In other words, the meanings of the aggregate elements that provide the

foundations for the narrative's plot depend on how specific events are interpreted by the narrator. And, since different narrators may interpret events differently, the aggregate elements are likely to mean different things across narratives – especially when narrators fall on opposing sides of ethnic conflict. Following Bearman et al. (1999), Smith claims that the robustness of the structure of ties between events reveals when elements are vulnerable to new interpretations. Thus, "the reorganization of relations within discrete aggregate elements is likely to be the locomotive for change in the narrative as a whole" (Smith, 2007: p. 27). Smith's aggregation technique allows her "to identify which relations among the lower-level events have changed in order to produce changes in meaning at the level of abstract element and, subsequently, of the narrative as a whole" (Smith, 2007: p. 42). This approach shows how the meanings of aggregate elements in narratives of ethnic violence demarcate the social boundaries between Italian Istrians and Croation Istrians living in Europe.

At the same time, some aggregate elements are likely to be *shared* by narratives from each side of the social boundary. Smith terms these aggregate elements 'boundary elements' and, while they may differ in meaning, they link narratives together. This can have notable consequences when narrators begin to interpret the boundary elements' constitutive events differently, thereby changing their meaning. When this occurs, boundary elements can fuse together distinct narratives. Smith shows, for example, that when Italian and Croatian Istrians in New York began reinterpreting the events surrounding Istrian victimization and the role of Italian Istrians who remained in Istria (instead of fleeing to Italy), they were able to merge 'the opposing narratives and smooth over differences' (Smith, 2007: p. 39).

This conceptualization of boundaries and the comparative approach more generally leads Smith to propose a different perspective on the relationship between narratives, causality, and identities. Whereas Bearman and Stovel (2000) used the narratives of Nazis to gain insight into the processes of identity formation (i.e., *becoming* and *being* a Nazi), Smith focuses our attention on how differences in narratives can cause a change in the nature of an identity over time. When Italian and Croatian Istrians socialized in New York City, they had reason to reinterpret the events of the original narratives. This resulted in transformed narrative elements, including boundary elements, which then led to a new unified Istrian narrative and a panethnic Istrian identity.

Institutional and Individual Narratives

In another innovative study, Quinlan and Quinlan (2010) advance the use of narrative networks by examining the types of narrative structures that may exist in nonstory forms of discourse (Polletta et al., 2011). In their comparison of a rape narrative told by a victim – what they call the 'lived experience' – with an institutionalized representation of rapes produced through official reports, Quinlan and Quinlan seek to identify the 'line of fault' between different forms of discourse. Their finding reinforces our intuitive understandings of narrative: the victim's story is driven by the recounting of the event, while the institutionalized report is a disjointed account centered on cataloging the identity of the subject, with little plot to glue the account together.

Quinlan and Quinlan begin by collecting the institutional texts – in this instance, forms from sexual assault evidence kits that record data from a series of questions posed to the victim concerning the assault. Roughly following Bearman and Stovel's substantive classification of nodes, they categorize their information into thematic elements (see also Shostak and Fox (2012) for thematic coding). For example, name and address are grouped under the larger thematic element of 'her identity.' These elements serve as the nodes of the network; connecting arcs represent links between elements. As a comparison case, they analyze a published narrative account of a rape and identify themes based on a selection of sentences or phrases that they deem most representative. These themes become nodes, which are then linked based on the thematic flow of the story.

In their analysis of the story of the rape, Quinlan and Quinlan's narrative network reveals the node 'his action' (i.e., the act of rape) is the driver mode, in the sense that it is the node from which all other nodes in the network of the narrative follow. Calculating the centrality score associated with each node, however, reveals that 'terror' is the most centralized node, while, local, lived experiences outside the rape exert no influence on the narrative. In contrast, the institutionalized representation of the rape has no driver node; the institutional text is comprised instead of clusters of thematic elements isolated from one another. With no narrative path through the institutionalized networks, the authors turn to a measure of centrality, and find that the identity of the victim – 'her identity' – is the most central thematic element in the text (even more centralized that the 'assault' node). It is interesting to note that Quinlan and Quinlan find that 'her identity' is the primary theme in the unconnected institutional texts whereas Bearman and Stovel's analysis reveals that the self disappears in the unconnected *being* narrative (Bearman and Stovel, 2000: p. 89).

Quinlan and Quinlan's study presents an approach to comparing narrative networks that emphasizes differences in the narratives (and meanings) that emerge from the stories of 'real,' lived experiences and the abstracted institutionalized renderings of those experiences. The comparison presents us with two models of identity. One is the personal identity, which is constructed via the connection between specific and detailed recollections of an experience; the second is the institutional view of identity, which is constructed using details abstracted from the actual experiences of the person. While the authored narrative has actors and agency, the institutional identity is all about figuring out who is who. This is the 'line of fault' between these two types of text. By contrasting these two types of narratives, Quinlan and Quinlan advance the 'narrative turn's' awareness of institutional context (Polletta et al., 2011), demonstrating the usefulness of narrative networks in revealing the forces by which institutions coordinate our identities and experiences.

Contributions and New Directions

The narrative network approach offers both a methodological tool and a framework for theoretical insights. Methodologically, it is one of a very small number of techniques designed to facilitate what Mario Small refers to as 'crossover analyses'; analyzing qualitative data with formal, quantitative methods

(or vice versa) (Small, 2011). Indeed, the narrative network approach rests on gathering large amounts of qualitative data and conducting systematic comparisons and analysis of network structure derived from the texts. As such, it can be applied to a wide range of substantive problems – as long as they were captured by stories.

At the same time, this approach is theoretically generative. Instead of producing causal models of what factors caused a particular outcome to occur, the narrative networks approach instead offers insight into how actors – or sets of actors – interpret and understand historical events and social phenomena. Especially when it is used in a comparative design, narrative networks' strength is that it makes it possible to identify the structure of common scripts, collective narratives, or changes in narratives, thereby 'fleshing out' a historical period or past experience.

However, as with other social scientific approaches that provide scholars with both methodological tools and theoretical insights, the use of network narratives raises several challenging issues. In this final section, we highlight some key obstacles associated with the narrative network approach. We view these limitations less as conceptual flaws and more as opportunities for further advancement and development.

Methodological Challenges

Sources of Data

For many research problems, the relevant narrative data are more complex than single-authored life stories of the type that dominate the studies described here. As in Quinlan and Quinlan (2010), textual data may come from official records, but they could also reflect second-hand accounts, transcripts of group interviews (Mckether et al., 2009), or organizational representations of events (Gerteis, 2002). Appropriate analyses of such diverse types of narrative will require further developments of narrative network techniques, and perhaps better appreciation of the consequences of selection bias in the sample of narratives. And yet since some types of qualitative data are not 'narratives,' – in the sense that they do not have a plot – the payoff of employing the narrative network for some types of data approach may be unclear. Thus this methodological issue surrounding different types of narrative data raises theoretical questions about what network structure can reveal about collective understandings, and the coercive power of strong narrative scripts.

Coding of Data

The majority of research using the narrative network approach employs a simple formula of coding nodes as events (Bearman et al., 1999; Bearman and Stovel, 2000; Quinlan and Quinlan, 2010), themes, or a combination (Smith, 2007) and arcs as the narratives clauses, causal relationships, or connecting relationships. Yet with each new study, new issues associated with coding emerge. When is a narrative element significant enough to warrant becoming a node? What constitutes a link between nodes? A key assumption of the approach is that arcs represent causality from a narrator's point of view, a perspective that may very well be at odds with other – perhaps more accepted – points of view. To draw from an example given by Lee et al. (2010), when dealing with the statement 'increased political tension in the middle-East drove the gas price up' the traditional narrative network approach would create a path between 'increased political tension in the middle-East' and 'gas price going up,' because the narrator believed it to be true. However, if historical evidence contradicts this, how do we understand the meaning of this arc?

Time

Applying the narrative network method raises many complex issues about time, timing, and temporality. Beyond the question of how to represent temporal flow in the coding of text, working with narrative data may reveal that an actor's narrative account itself changes over time, as her understanding of the situation changes, or as the institutional or cultural context shifts (e.g., Shostak and Fox, 2012). While a narrative network represents the temporal sequencing of events from one vantage point in time, many important questions revolve around how narratives change over time. Bearman et al. (1999) argue that when multiple narrative networks produce a stable bicomponent, the narrative meaning of a historical event is fixed. Does an analogous process operate when an individual tells his life story at multiple moments in time?

Theoretical/Interpretive Challenges

The second major challenge facing scholars using narrative networks concerns how to make sense of the findings. This, of course, is related to how narratives are coded and what network analyses are conducted, but at a more conceptual level, we must consider what narratives can and cannot tell us about the author's world.

We begin by focusing on the author of the narrative – the original subject of narrative networks – and ask what insights narratives give us into the multilayered, hierarchical identities that people have. Bearman and Stovel (2000) fix their gaze on the development of a 'master identity,' a concept that is clearly relevant for those, like Nazis, with strong ideological orientation. But most persons do not claim a master identity; in fact, social life is particularly interesting and fraught with ambiguity when actors must determine what identity to deploy in a particular situation. The tools of narrative networks have not been applied to research problems that focus explicitly on the choice of what identity people use in different settings, and yet it seems reasonable to hope that narrative networks will enable us to gain insight into the layers of identities and meaning that reside in an author's interpretation of their experiences.

Another theoretical question concerns how to connect a narrative to the cultural and institutional context that it is embedded in. Exactly because institutions and cultural context provide the background for narrative, crucial aspects – even, from the narrator's perspective, *causally* crucial aspects – may be omitted from the told tale. These omissions may lead an analyst to fundamentally misunderstand the author's intended plot. One way forward is to follow Bearman et al. (1999) and Smith (2007), and analyze multiple narratives of the same events, with the hope that the integrated narrative will point toward important cultural frameworks guiding the narrative accounts, yet neither of these studies explicitly theorized how context shapes narrative understanding.

Even more challenging is to understand how the telling of narratives – whether by individuals or through the

development of collective narratives – impacts institutions, politics, culture, or social organization. Much research lumped under the broad narrative umbrella claims narrative as a resource that can shape the social environment, and yet to date, the narrative network approach has not been used to explore how narratives can be deployed to effect change. Important questions include Does the extent to which a narrative account is shared matter? Does it matter how a narrative dovetails with empirically verifiable facts? These issues become even more intriguing given the emergence and spread of real-time communication technologies (e.g., social media, cheap and accessible mobile devices), which result in the proliferation of nonnarrative forms of discourse. For instance, future scholars might use narrative networks' approaches to compare revolutionary narratives produced in Egypt (2011) to those written about Nasser's 1952 Egyptian revolution. This and other challenges make us optimistic that the narrative network approach can grow beyond its founding link with historical sociology and more fully engage scholars and disciplines focusing on social change.

See also: Collective Identity; Collective Memory, Psychology of; Collective Memory, Social Psychology of; Culture and Networks; Culture, Cognition and Embodiment; Formal Methods of Cultural Analysis; Narrative in Social Work; Narrative, Sociology of; Network Analysis; Networks and Meaning; Symbolic Boundaries.

Bibliography

Abell, P., 1987. The Syntax of Social Life: The Theory and Method of Comparative Narratives. Clarendon Press, Oxford.

Bearman, P., Faris, R., Moody, J., 1999. Blocking the future: new solutions for old problems in historical social science. Social Science History 23 (4), 501–533.

Bearman, P.S., Stovel, K., 2000. Becoming a Nazi: a model for narrative networks. Poetics 27 (2), 69–90.

Corsaro, W.A., Heise, D., 1990. Event structure models from ethnographic data. Sociological Methodology 20 (1), 1–57.

Gerteis, J., 2002. The possession of civic virtue: movement narratives of race and class in the knights of labor. American Journal of Sociology 108 (3), 580–615.

Lee, Sungmin, Ramenzoni, Nica C., Holme, Peter, 2010. Emergence of collective Memories. PLoS One 5 (9), e12522.

Mckether, W.L., Gluesing, J.C., Riopelle, Kenneth, 2009. From interviews to social network analysis: an approach for revealing social networks embedded in narrative data. Field Methods 21 (2), 154–180.

Pentland, B.T., Feldman, M.S., 2007. Narrative networks: patterns of technology and organization. Organization Science 18, 781–795.

Polletta, F., Chen, P.C.B., Gardner, B.G., Motes, A., 2011. The sociology of storytelling. Annual Review of Sociology 37, 109–130.

Quinlan, E., Quinlan, A., 2010. Representations of rape: transcending methodological divides. Journal of Mixed Methods Research 4 (2), 127–143.

Shostak, S., Fox, N.S., 2012. Forgetting and remembering epilepsy: collective memory and the experience of illness. Sociology of Health & Illness 34 (2), 362–378.

Small, M.L., 2011. How to conduct a mixed methods study: recent trends in a rapidly growing literature. Annual Review of Sociology 37, 57–86.

Smith, T., 2007. Narrative boundaries and the dynamics of ethnic conflict and conciliation. Poetics 35 (1), 22–46.

Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge, UK.

Narrative, Sociology of

Karin Aronsson, Dep. of Child and Youth Studies, Stockholm University, Stockholm, Sweden.

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Narratives are ways of linking together past events, making sense of historical or fictional personae. There exists a multitude of genres that draw on diverse academic traditions, ranging from legal studies and sociology to literature and hermeneutics.

This article explores discursive approaches to narratives, studied as situated *social action*, focusing on storytelling, rather than on the stories as such. In courtroom examinations, medical interviews, everyday blame allocations, and other contexts where stories matter, narrative action involves numerous storytelling devices: e.g., social accounts, reported speech, and social categorizations. The conversational work of narrators also invokes participation frameworks and audience design.

The Narrative Turn and Technologies of Selves

During the last two decades of the past century, social theory has undergone a narrative turn. Before that, Wittgenstein's (1953) work on language games had broken new ground in legitimizing the study of language usage, moving beyond a study of language as a fixed set of structures. Contemporary analyses of narratives feature work – not only on language games, ambiguity, and fuzzy concepts – but also on storytelling procedures, alternative versions, and contested stories. Some of these new directions, including the idea of intertextuality, genre, and audience design can be traced to Bakhtin (1981) and the importation of concepts from literary theory into the social sciences. Simultaneously, a novel focus on sequential unfolding, and micro-politics draws on Goffman (1961, 1967, 1981), Sacks (1992), and new developments within sociology, anthropology, and linguistics. In order to understand a story fully, it is necessary to know something about where, when, why, to whom, and by whom the story was told. In work on narration, analyses of subversion are closely related to dramaturgical aspects of storytelling: as in a focus on participation frameworks, that is, on speaker/listener roles (Goffman, 1981).

In his historical societal analyses, Foucault has discussed how dominant stories have been employed in the colonization of selves. Foucault's discussion of the technologies of selves (1997), presupposes that our public selves are intimately linked to conversational practices. His analysis of the historical roots of self-technologies links such technologies loosely to self-reflexivity practices and to early forms of self-confessions. In the case of early Greco-Roman philosophy, or in monastic confessions in early Christian spirituality, the self was narrated in more or less monological formats, e.g., in letters to friends.

Interviewing, teaching, counseling, and therapy can be seen as prototypical discursive technologies for eliciting what Foucault calls confessions and authoritative responses to such confessions, e.g., judgment, evaluation, celebration, and discipline (Foucault, 1997). In modern society, confessions have become integrated into what Foucault calls *truth regimes* that organize institutional practices in normative patterns for what should be told to whom and in what ways. In modern life, there are several discursive genres that can be related to the formation – and reflection – of moral order and selves in a Foucauldian universe. Such genres are important in discursive practices, but Foucault himself does not detail how such identity work is done on a conversational level.

In folkloristic and ethnographically oriented sociology, life-story interviews have been key methods in the formulation of reflexive analyses of everyday action. Traditionally, interviews have been the main tools for eliciting life stories or story fragments. However, what has been analyzed is normally the life story as such, not the storytelling. But, since the 1990s, ethnographic and anthropological work on life narratives has turned toward procedural aspects; that is, to the storytelling, as well as to how narratives are linked to the formation of selves in different cultures (for reviews, see Gubrium and Holstein, 2008; Ochs and Capps, 1996). In line with the work of Bamberg (2011) the focus is often on small stories rather than on grand narratives. Analyses of such storytelling practices are particularly illuminating, as they combine life-story material with natural data not primarily elicited for the benefit of the interviewer. In a Foucauldian universe, education is seen as one of the technologies of the self. It is not merely an issue for formal schooling, but also for informal schooling in family-life contexts. In work on family dinner conversations, it can be seen that narratives are deeply embedded in family politics that are, in turn, shaped by cultural norms for participation. In their analyses of agoraphobia in ordinary family conversations, Capps and Ochs (1995), for instance, illuminate the ways in which panic is displayed and co-construed within the everyday narratives of a family where one member suffers from agoraphobia. Through her mundane panic stories, the mother reveals her fears, but she also passes on patterns for how to construe the world as a fearful place. In their detailed analyses of this troubled family's dinner conversations, Capps and Ochs (1995) identified narrative patterns, reflecting underlying family hierarchies.

One of the main methodological controversies in recent work on narration concerns the relative role of text and context. The work of Foucault, as well as much ethnographic and traditional sociological work, is primarily focused on contextual aspects of discourse. In contrast, conversation analysis is primarily oriented to the text as such and to intersubjectivity (the dialogic production of narratives). Context is in focus only in those cases where it is actually invoked by the participants

themselves (Schegloff, 1997). In conversation analysis and ethnomethodology (Heritage, 1984), phenomena such as dominant narratives are thus primarily seen as participants' phenomena (see also, Sacks, 1992). It is only if domination is invoked by the participants themselves that it is actually taken up in the researcher's analyses.

Contemporary narrative models (Bruner, 2004) position language and the role of narration in the foreground in the formation of minds. Similarly, narratives are seen as the basic building blocks for imagined communities (Andersen, 1991), in that the very notion of community is often based on historically grounded, or oppositional narratives. In fact, the social sciences are based on local ways of building narratives and on the type of language varieties involved (Rorty, 1980).

Social Action and Technologies for Building Stories

Some contemporary models are deliberately 'mindless' in that they are more oriented toward procedural aspects of narration: how narratives evolve sequentially and how they are co-construed or deconstrued as part of social action (Drew, 1992; Galatolo and Drew, 2006; Goodwin, 2007; Sacks, 1992; Schegloff, 1997). Sacks (1992) investigated the interactional features of a variety of narrative practices, such as telephone calls to a suicide center, problem formulations in therapy, and the telling of jokes. His lectures on conversation primarily discussed the turn-taking machinery, the dialogical and sequential organization of conversational turn-taking (Sacks, 1992). Yet they also cover entitlement to stories, tellability, and what he called 'a technology for building stories.'

More recently, there have been discussions concerning how talk in action, that is, dialogic mechanisms for storytelling and speaking can be seen in courtrooms and other institutional settings (Drew, 1992; Heritage and Clayman, 2010; Lynch and Bogen, 1996). Such interactional analyses of narratives focus on the *storytelling* or the narrating, not on the static structural aspects of traditional narrative analyses of stories. In the realm of interactional theories, the micro-politics of identity work in institutional settings has been documented in a series of analyses of contemporary institutional narratives. There are many genres of such narratives: for instance, couple therapy (Buttny, 1993), and suspects' and defendants' storytelling in police interrogations and courtroom examinations (Drew, 1992; Galatolo and Drew, 2006; Lynch and Bogen, 1996; Stokoe, 2010), as well as in anamnestic contexts (e.g., Heritage and Robinson, 2006). All these genres have implications for identity formation, both for displaying identities and for cocreating identities. This Janus-face nature of conversations and selves is also an essential feature of ethnographic and conversational work on narratives and self-identities.

Narrative Devices

Much research on narratives has centered on the structure of plots in relation to various elements (such as background, future events, introduction of main characters). The temporal order of the narrative has long been a key element of narrative organization as in Labov's (1972) classical analyses of oral narratives of inner-city youth. Minimally, a narrative scheme involves two temporally ordered events. In the more elaborate type of fight stories and other personal narratives that Labov elicited from inner-city boys, it is often possible to discern a more complex pattern involving abstract, orientation, complicating action, evaluation (what is the point?), resolution, and coda (is there a signal that the story is over?). Yet, the sequential order of storytelling need not coincide with the temporal organization of the tale. In a critical discussion of narratives and cognition, Edwards (1997) demonstrates ways in which script models and other structural models (for example, Labovian schemes) in fact do not hold up in the face of the complexities of mundane interaction; they collapse when they are applied to real-life online narration in interactional contexts.

Interactional analyses of storytelling events are therefore important tools for sharpening our reasoning on narration as '*social action*.' Such work focuses on participants' orientations to coparticipants and to the narrated events, rather than primarily or only on the temporal framework of the story. The narrator's affective stance or point of view can often be detected in linguistic cues such as pronoun choice or passive versus active verb forms (Goodwin, 2006, 2007; Ochs and Capps, 1996). In some cases, linguistic resources can be seen to index stances that are closely related to conversational or institutional genres (Bakhtin, 1984; Lynch and Bogen, 1996), such as assessments or complaints. In interactional work on narration, genre is therefore not merely seen as a background variable for talk, but rather as an outcome of the interaction as such. In many institutional encounters, it is, for instance, the powerful party who engages in authoritative moves, such as reformulating what the story is about (Sacks, 1992). Conversely, this means that speakers may position themselves as people in power through exploiting discursive devices that are associated with institutional authority.

The intersubjective architecture of confessions was discussed in Sacks' (1992) lectures on conversation. In telling 'a first story,' a speaker may invite 'a second story,' and categorizations and first assessments often precede second assessments. In everyday conversations between friends, disclosures of secrets or taboo or delicate information may similarly invite disclosures by copresent parties. Disclosures similarly tend to assume a sequential structure, involving fragmented and gradual revelations. Similarly, Goodwin (2006) has demonstrated the complex sequential structure of preadolescent girls' adversative storytelling by showing how they would animate and enact characters, using reported speech in casting someone as blameworthy, while simultaneously securing the alignment of other copresent parties. Moreover, the girls would draw on social categorizations, such as 'tag along' in creating social exclusion. Blame narratives were therefore resources for co-creating inclusion/exclusion and peer group order.

The written record is obviously an important resource in the technologies of selves. Cicourel (1968) revealed ways in which written police records were ultimately the products of talk; that is, the products of extended oral negotiations. Recently, Lynch and Bogen (1996, p. 203) specified many ways in which courtroom dialogue may be 'on the verge of breaking into

writing,' as in the Iran–Contra hearings in which prepared written statements, documentary exhibits, citations of prior testimonies, and written notes played a prominent role in the defendant's 'spontaneous' oral response to questions in the public hearing.

Written narratives tend to look more objective or factual than oral narratives. Yet, many written documents obviously emanate from talk-in-interaction. The delicate interplay between written and spoken language can also be seen in other institutional settings than courtrooms. In his work on answering patterns at a suicide prevention center, Sacks (1992), for instance, showed that the professional's answering sheet influenced the order in which she/he asked questions, illustrating how the design of written forms and records is important for the formation of oral narratives.

In several institutional practices – e.g., medical consultations, trials, and psychotherapy – an important part of conversational action is to identify the problem. Competing problem formulations are therefore key elements in such institutional narrations (e.g., Aronsson and Cederborg, 1994; Aronsson and Osvaldsson, 2014; Lynch and Bogen, 1996; Stivers, 2007). Narrations that are related to self-presentations entail substantial moral work in that they project different positions within a given social order. But in conversational studies of narration, control is not restricted to fixed positions in local hierarchies. Within the medical arena, it has been shown that doctors influence patients' problem presentations in a series of ways, for instance, in how they design their opening questions (Heritage and Robinson, 2006); whether they invite anamnestic narratives through a general inquiry (e.g., 'so tell me about this pain you are getting') or whether they merely ask for anamnestic confirmations. However, patients also influence doctors, and prescriptions of antibiotics can partly be seen as the outcome of social interaction, rather than medical concerns alone (Stivers, 2007).

Social Accounts and Character Contests

In institutional contexts such as therapy talk or interrogations, participants regularly present social accounts (Buttny, 1993; Stokoe, 2010) where they justify or excuse themselves in the light of implicit or explicit blame.

Narratives can be said to be located within epistemic territories (Heritage and Raymond, 2005) where participants recurrently assess story contributions with regard to credibility, entitlement, or other issues related to the local interaction order. Social action is, for instance, accomplished through upgradings and downgradings of narrative assessments, at times combined with tags (inviting coparticipants' agreement). Occasionally, such narratives actually form part of what Goffman (1967, pp. 239–270) has called *character contests* where not only past actions form part of assessments, but also the ways in which a target participant manages to account for mistakes or failure in convincing or compelling ways. In so doing, participants may invoke social categories to accomplish complaints, blame, accounts, or other types of social action (Stokoe, 2010). For instance, Sjöblom and Aronsson (2012) have shown ways in which adolescent boys during computer gaming conversations recurrently categorize each other as 'noobs' ('newbie,' incompetent) when accounting for player death or other category-bound events (linked to incompetent gaming). Conversely, skillful gaming is related to various ways of avoiding failure. What is important is that a player manages to present *social accounts* for past gaming in such a way that he does not risk being categorized as a noob, that is, as someone at the bottom of the local rank order. In their retellings (and accounts) of crucial game events, players are then ultimately assessed, not only with regard to gaming skills, but also with regard to their ways of accounting for or failing to account for unfortunate game performance, that is, their narrative skills.

Resistance and Social Categorizations

Dominance and power are classical sociological themes, but in the work of narratives in interaction, collusion, and subversion are equally important phenomena. Subversive action has been at the heart of sociological work on narratives, not only in the work of Foucault on societal institutions such as prisons, but also in Goffmanian work on mundane everyday negotiations. In many ways, narratives can be seen as battlefields for self-ascriptions and self-projections even if some of the battles are indirect or hidden. In his classical work on asylums, Goffman (1961) discussed the type of categorization work done by total institutions in their diagnoses of patients, as well as the patients' resistance to stigmatization. Ambiguous narrative roles (who speaks on whose behalf?) are also at the core of Goffman's (1981) work on participation frameworks and in Bakhtin's (1981) work on the multivoicedness of narration, but neither of these scholars has in detail documented the sequential unfolding of resistance.

Recently, resistance has, for instance, been analyzed in detail in Stokoe's (2010) work on *category-bound denials* in police interrogation storytelling where defendants claim that they would never touch members of a specific social category, for instance children or the elderly (e.g., 'I am not going to hit a lady'). Negotiations about social categories can also be seen in teenage girls' divergent story versions about 'bullying events,' where the same party can be cast as a victim or a bully, depending on who is telling the story to whom (Karlsson and Evaldsson, 2011).

Narratives in interaction are often contested, as language use is never disinterested – narrators always direct their narratives to someone, and they wittingly or unwittingly speak on somebody's behalf (Bakhtin, 1984; Goffman, 1981), adopting various affective and epistemic stances. Within an epistemology of a multivoiced or polyphonic universe, encounters also involve potentials for secret alliances and collusions. Such collusion can be seen in various institutional contexts, such as in asymmetrical encounters when there are dominant or authoritative voices, on the one hand, and subordinate voices, on the other (see also Heritage and Raymond, 2005).

In detailed analyses of small claim court interactions, it has been shown that narrative genres at times have collusive qualities. For instance, relationship-oriented and informal narratives need not be more egalitarian and democratic than

more formal rule-oriented narratives (Conley and O'Barr, 2005). In fact, inexperienced persons are often more at a loss in informal contexts, in that they are required to construe legal narratives on their own, whereas formal settings offer more support from legal professionals.

Silent opposition can also be found in public events. In legal contexts, witnesses are expected to be compliant. Yet in police interrogations, for instance, compliant witnesses may, through their collaboration, produce the very evidence needed for an eventual accusation (Lynch and Bogen, 1996). Institutional confessions or disclosures, therefore, often involve substantial downgrading of agency through impersonal constructions, vague or evasive responses, and minimizations of the offensive act. In their in-depth analyses of the Iran–Contra hearings, Lynch and Bogen (1996), for instance, demonstrate how the defendant Oliver North cleverly exploited his response turns, changing the genre of the trial during its 6 days on television. Through his military politeness, minimal responses, and strategic repetitions, he displayed an attitude of apparent compliance and respectfulness to the great media audience – the popularity of which was also reflected in an audience response of flowers, supportive letters, and successive transformation of an alleged villain of war into a public hero.

Moreover, the two authors present fine-grained analyses of the conversational management of staged spontaneity and strategic 'forgetting' of crucial events. For instance, Oliver North acknowledged lying in the past, but claimed that he would now tell the truth: "the good, the bad, and the ugly" (Lynch and Bogen, 1996, p. 38). On a deeper level, however, the analyses of Lynch and Bogen (1996) show how a well-prepared witness like Oliver North skillfully avoided responding to the important questions, undermining the interrogator's narrative, thereby indirectly mocking his interrogators. The trial was thereby changed into public entertainment; what Foucault would have called a historical spectacle as opposed to a serious courtroom interrogation.

Storytelling as Storytelling Contests

Competing narratives can, for instance, be found in family-therapy conversations. Aronsson and Cederborg (1994) demonstrate that therapy often involves underlying storytelling contests in which one side ratifies its own story, whereas the other side's story is deauthorized in that the story point is challenged or undermined. The therapist plays a key role as an intermediary – an orchestrator of who talks to whom about what – and as arbitrator about when a narrative is ultimately finished. But the therapists' control in family therapy is more concerned with issuing storytelling rights than with sheer dominance over turn taking. It primarily concerns what is reportable or tellable, as well as story boundaries, that is, when a story is to be told and implicit norms for when it is finished.

A pioneer study of oppositional work in legal settings can be seen in Cicourel's (1968) monograph on juvenile delinquents and the conversational work of parents, probation officers, and experts such as psychiatrists. Cicourel did not detail the conversational work as such, but he discussed the defensive interactional work of parents, showing, for instance, how middle-class parents recruited outside assistance from therapists and teachers, working on recategorizing their sons as depressive rather than as delinquent, or as temporarily influenced by peers rather than as active and willful. The written police record categorizations of 'delinquent' actions were therefore partly the result of extended parent–expert negotiations. In his micro-sociological analyses of parental storytelling in the face of alternative legal versions, Cicourel (1968) simultaneously shows that detailed ethnographic data need not exclude a concern for macro-sociological issues such as class.

In adversarial courtroom proceedings, trials can be seen as storytelling contests, in which each attorney must ask questions in such a way that his/her client tells the most convincing story. Credibility is thus largely a matter of narrative design, performed by the witness or the defendant and masterminded by the attorney. Such storytelling contests underlie various actions in adversarial courtroom contexts. Witnesses are positioned as members of either side's team (the prosecution or defense side), which means that, for instance, rape victims tend to be treated as prosecution team members, and are at times exposed to discrediting and damaging questions by the attorneys of the defense. In an analysis of an American rape trial, Drew (1992) shows that such trials may involve substantial resistance and strategic defense on the victim's part. In his detailed analyses of an extended episode from a rape trial, Drew shows that the rape victim employed alternative descriptions as well as other discursive devices in her story (of being merely an acquaintance of the rapist) as opposed to that of the prosecution (of her being a lover or willing partner). Yet, the two competing stories were told as fragments, sequentially revealed in the attorney's questions and in the witness' defensive and oppositional responses.

In a discussion of the (in)famous Rodney King case, Goodwin (2007) presents detailed accounts of how a victim's bodily movements as *reaction* to police violence can instead be transformed into prosecution side stories about incipient *aggressive actions*. In a recent study of parents' courtroom talk in child custody trials, Ingrids and Aronsson (2014) document a series of such radical transformations of agency, transforming protagonists of opposing story versions from victims into culpable parties and vice versa. Moreover, the analyses (Ingrids and Aronsson, 2014) document the role of exact quotes (direct reported speech) in enacting prior events, building credibility through alternative stories, rendering opinions both more vivid and authentic or 'objective.' These quotes were embedded in repeats, exaggerations, and at times even reported dialogues that embellished the contesting narrative versions. Reported speech simultaneously provided a unique resource for including the absent child's positions in this courtroom context, an important rhetorical device for validating both authenticity and epistemic authority. The testimony reveals that the parent actually *heard* what happened, and it animates the absent party, putting the listener in the child's place through vivid descriptions.

Interaction Order and Identity Work

In everyday life, character is recurrently at stake, first in the ways that participants manage to act in line with the local interaction order and second, in the ways that they manage to explain

failure or success and to account for past action in such ways that their actions will not be contested. These narrative contestations therefore invoke two types of selves, the selves of the participants themselves, on the one hand, and their narrator selves, on the other.

See also: Epistemic Cultures; Foucault, Michel (1926–84); Goffman, Erving (1922–1982); Narrative Approaches to Organizations; Narrative Networks; Narrative in Social Work.

Bibliography

Andersen, B., 1991. Imagined Communities. Verso, London.

Aronsson, K., Cederborg, A.-C., 1994. Conarration and voice in family therapy: voicing, devoicing, and orchestration. Text 14, 345–370.

Aronsson, K., Osvaldsson, 2014. Trust and the contestation of blame narratives: veiled stances in an institutional assessment context. In: Linell, P., Marková (Eds.), Dialogical Approaches to Trust in Communication. Information Age Publishing, Charlotte, NC, pp. 29–49.

Bakhtin, M., 1981. The Dialogic Imagination: Four Essays. University of Texas Press, Austin.

Bakhtin, M., 1984. Problems of Dostoevsky's poetics. University of Minnesota Press, Minneapolis, MN.

Bamberg, M., 2011. Who am I? Narration and its contribution to self and identity. Theory & Psychology 21, 3–24.

Bruner, J., 2004. Life as narrative. Social Research 71, 691–710.

Buttny, R., 1993. Social Accountability in Communication. Sage, London.

Capps, L., Ochs, E., 1995. Constructing Panic: The Discourse of Agoraphobia. Harvard University Press, Cambridge, MA.

Conley, J.M., O'Barr, W.M., 2005. Just Words: Law, Language and Power. University of Chicago press, Chicago.

Cicourel, A., 1968. The Social Organization of Juvenile Justice. John Wiley, New York.

Drew, P., 1992. Contested evidence in courtroom cross-examination: the case of a trial for rape. In: Drew, P., Heritage, J. (Eds.), Talk at Work: Interaction in Institutional Settings. Cambridge University Press, Cambridge, UK, pp. 470–520.

Edwards, D., 1997. Discourse and Cognition. Sage, London.

Foucault, M., 1997. Technologies of the self. In: Rabinow, P., Foucault, Michel (Eds.), The Essential Works, Ethics, Subjectivity and Truth, vol. 1. Penguin Press, Harmondsworth, UK, pp. 223–251.

Galatolo, R., Drew, P., 2006. Narrative expansions as defensive practices in courtroom testimony. Text & Talk 26, 661–698.

Goffman, E., 1961. Asylums: Essays on the Social Situation of Mental Patients and Other Inmates. Penguin Books, Harmondsworth, UK.

Goffman, E., 1967. Interaction Ritual: Essays on Face-to-Face Behavior. Pantheon Books, New York.

Goffman, E., 1981. Forms of Talk. Basil Blackwell, Oxford.

Goodwin, M.H., 2006. The Hidden Life of Girls: Games of Stance, Status, and Exclusion. Blackwell, Malden, MA.

Goodwin, C., 2007. Interactive footing. In: Holt, E., Clift, R. (Eds.), Reporting Talk: Reported Speech in Interaction. Cambridge University Press, Cambridge, England, pp. 16–46.

Gubrium, J.F., Holstein, J.A., 2008. Narrative ethnography. In: Hesse-Biber, S Nagy, Leavy, P. (Eds.), Handbook of Emergent Methods. The Guilford Press, pp. 241–264.

Heritage, J., 1984. Garfinkel and Ethnomethodology. Polity Press, Cambridge.

Heritage, J., Clayman, S., 2010. Talk in Action: Interactions, Identities, and Institutions. Wiley-Blackwell, Malden.

Heritage, J., Raymond, G., 2005. The terms of agreement: indexing epistemic authority and subordination in talk-in-interaction. Social Psychology Quarterly 68, 15–38.

Heritage, J., Robinson, J.D., 2006. The structure of patients' presenting concerns: physicians' opening questions. Health Communication 19, 89–102.

Ingrids, H., Aronsson, K., 2014. Reported speech and reported affect in child custody disputes. Research on Language and Social Interaction 47, 69–88.

Karlsson, M., Evaldsson, A.-C., 2011. "It was Emma's army who bullied that girl": a narrative perspective on bullying and identity making in three girls' friendship groups. Narrative Inquiry 21, 24–43.

Labov, W.A., 1972. Language in the Inner City. University of Pennsylvania Press, Philadelphia, PA.

Lynch, M., Bogen, D., 1996. The Spectacle of History: Speech, Text, and Memory in the Iran-Contra Hearings. Duke University Press, Durham.

Ochs, E., Capps, L., 1996. Narrating the self. Annual Review of Anthropology 25, 19–43.

Rorty, P., 1980. Philosophy and the Mirror of Nature. Princeton University Press, Princeton, NJ.

Sacks, H., 1992. Lectures on Conversation, vols. I and II. Basil Blackwell, Oxford, UK.

Schegloff, E.A., 1997. Whose text? Whose context? Discourse & Society 8, 165–187.

Sjöblom, B., Aronsson, K., 2012. Disputes, stakes and game involvement: facing death in computer gaming. In: Danby, S., Theobald, M. (Eds.), Disputes in Everyday Life: Social and Moral Orders of Children and Young People. Sociological Studies of Children and Youth, vol. 15. Emerald Boks, Cambridge, MA.

Stivers, T., 2007. Prescribing under Pressure: Parent-Physician Conversation and Antibiotics. Oxford University Press, New York.

Stokoe, E., 2010. 'I'm not gonna hit a lady': conversation analysis, membership categorization and men's violence toward women. Discourse & Society 21, 59–82.

Wittgenstein, L., 1953. Philosophical Investigations: Philosophische Untersuchungen. Macmillan, Oxford.

Nation, Sociological Aspects of

André-Paul Frognier, University of Louvain, Louvain-La-Neuve, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of nation and the sociology of nation are discussed in their historical contexts. The distinction is drawn between those authors taking a 'modernist' viewpoint – that the nation is a truly modern invention which takes the form of an imaginary community – and the 'primordialists' concerned mainly with the historical roots of the nation. Globalization and cultural homogenization as well as the financial and debt crisis make the future of the nation uncertain, because it can revive or dissolve in larger communities.

Nation: A Late Political Meaning

Etymologically speaking, a nation, from the Latin *natio*, designates all those linked by birth. In modern language, the word that best expresses that original meaning of Nation is the English word 'native.' In ancient Rome, the word had no particular political meaning. It seemed even to characterize human groups with no political structures. During the Middle Ages, the word nation obtained a linguistic meaning. This explains why students at the first great European universities were regrouped in nations according to the language they spoke and sometimes even clashed along those lines!

In the Renaissance, during the process of state building, the nation again endorsed a more political bent until the French Revolution: for a long time, the nation was identified with those benefiting from a political status because they were close to the crown (the power) or were represented in those pre-democratic parliaments, from which the rest of the population was excluded (Schulze, 1996).

Only during modern times does the nation take its present and more popular meaning. The concept even absorbs almost all the different denominations characterizing political societies. Just think about the League of Nations, or the United Nations organization: this testifies equally to the might of the word and to its largely undetermined character, quasiassimilating the concept of nation with any kind of statist political organization.

From Nation to the Sociology of Nation

The nation as we conceive it today is a product of the nineteenth century, which makes it the main support of the constitution and development of modern states by using the famous 'principle of nationalities.' However, sociology took her own good time to discover the nation.

The founding fathers of sociology who carved their imprint in the nineteenth century, whether Karl Marx, Max Weber, or Emile Durkheim, were far from giving a central place to the nation in their analysis of modern society, contrarily to historians or philosophers, or even artists, who started discussions, as witnessed by diverging opinions from Renan, e Fustel de Coulanges, Fichte, or Herder. Sociology is based on 'society' whose meaning remains largely undefined. As an object research, it deals with the state or the people, but not the nation. However, these notions are often described under the characteristics of a modern nation – as without doubt by Weber in his analysis of society – but never did the concept of nation emerge.

Marx and Engels deal a little more with the notion of nation, but they consider it more as an archaic form than as a new reality pregnant with history. The next generation of Marxist authors will be more concerned by the notion, but for practical more than theoretical reasons, given the attractiveness of nationalistic slogans for the working class. We owe to Bauer, and to the 'Austro-Marxism,' a more sophisticated conceptualization (Bauer, 1907). Stalin himself tried to define exactly what he means by nation in a work where he uses its nonpolitical aspects as a basis – a restrictive one – for the Bolshevik policy of recognition of nationalities (Stalin, 1937).

As pointed out by Hobsbawm, the multiplication of works on the nation dates back mainly to the period 1968–88 when more was written on the subject 'than at any other period, even twice as long' (Hobsbawm, 1990). The concept of nation then gets a truly sociological meaning, though its definition has opened new conflicts that are still quite alive today.

However, in modern times, nation will be unanimously recognized as *the* political community that will ensure the legitimacy of the state over its territory, and moreover transform the state into the state of all its citizens. The notion of nation-state illustrates that new alliance between state and nation. Nationality will bind the citizen to the state, a bond that will be more and more tied to the advantages of a social policy in as much as the welfare state will develop.

Nobody will deny either that the nation finds its ideology in nationalism, where nation and state necessarily are intertwined into a new entity dominating all others. Nationalism will not only mark the building of nations, but also provoke the reaction of those populations who feel suffocated by this process and want to build their own nation-state. The violence of nationalism will depend on the surrounding circumstances and evolve according to the obstacles found on its way.

Where true nation-states will emerge, the nation will be the point of convergence of three historical processes. The first one is the development of states with growing sovereignty over territories whose boundaries are more and more clearly defined, and with growing autonomy vis-à-vis the non-territorial power of the Church, whose prestige and influence declined after the Reformation; this process was faster in

Northern Europe than in the South. The next step is the emergence of democracy in the aftermath of the French Revolution, transferring sovereignty from the state to the people and regrouping all citizens under principles of equality and solidarity. Finally, the development of capitalism increases communications between individuals, social homogeneity, and educational and cultural standardization, and by its very crises leads to the strengthening of states' power and the development of the welfare state endorsing an ever-increasing part of the life of each citizen.

The differences of evaluation among authors focus on the importance to be given to the ties linking the modern nation to the former 'protonational' entities. These differences concern the process of construction of those nations and the role of history in that process. For sociologists like Gellner, Hobsbawm, and Anderson, modern nation is first a rupture, even sometimes a complete one, with the past. This 'modernist' definition is at loggerheads with the definitions, called 'primordialist' by Gellner, of which the best-known representative is Smith, a former pupil of Gellner.

Two Concepts of Nation?

For the modernists, the rise of the nation can only be understood as the product of state action, and more particularly, according to the famous hypothesis of Gellner, as the result of the nationalist action of its elites (Gellner, 1983). Nationalism gives birth to nation and not the opposite. This voluntarist action is made possible, and even necessary, by modernization, that is, by the transition from agrarian society to industrial society and capitalism. Modernization destroys old social hierarchies and loyalties therein associated, and entails homogenization of language and culture throughout the state territory to allow both for mobility of workers and for mass production. Whereas agrarian societies possessed stable hierarchies where culture was linked to social status, modern society, by destroying these hierarchies, transforms culture into a kind of personal appropriation of each individual and creates a longing for harmony between individual culture and the environment: 'What automatically makes people into nationalism' (Gellner, 1999).

In that sense, nation is a truly modern invention that takes the form of an 'imaginary community' (Anderson, 1983). Indeed, nation-states have a past, at least most of them, but modernization is a phenomenon modifying so radically social practices that this past is practically irrelevant for the construction of modern nations. Or, at best, it is so remolded that it loses any true historical characteristic. One example, pointed out by Gellner, of a nation created *quasi ab nihilo* is the case of Estonia. Whereas that nation has no past to speak of, its elites managed to give it life, to develop a distinctive culture, and to impose it among the other nations. In this manner, Estonia is an ideal type of a nation in the modern sense of the word.

At the opposite end, for the primordialists, the development of the nation-state, and therefore of the nation, is understandable only through analyzing its historical roots. Smith calls this approach 'ethno-symbolist.' One cannot but study the role of myths and prenational symbols in order to distinguish nationalisms from one another, and above all to explain why nationalism will be virulent in one case and not in the other (Smith, 1986). The existence of competing nationalisms in Greece cannot be understood if one loses sight of "its dual heritage of byzantine imperial orthodoxy and classical democratic antiquity. Modernism tells half the story," writes Smith. But Gellner answers that "if it tells half the story, that is enough for me!" (Smith, 1999; Gellner, 1999).

In this perspective, Hroch offers an authoritative developmentalist typology of national movements. These movements evolve in three phases. First is that of purely cultural, linguistic, or folklore-related aspirations. Second is that of the rise of militant groups promoting the 'national idea' and starting social trouble around the national theme. Third is one that finds spreading support among the masses for the new national aspirations. The transition from the second to the third phase will often be marked by a 'nationally relevant conflict of interest' (Hroch, 1985).

What are the factors prior to the nation that can legitimize its rise? The most frequently mentioned ones are language and religion. Language will be particularly important as a factor when it has been threatened or has been a basic mode of identification. This is the case, for instance, in Belgium, where Flemish nationalism concentrates in a language that has been the object of assimilation by a dominant language (Delfosse, 1994). However, language is not necessarily the essential legitimizing source of nationalism and nation. The French Revolution does not establish a nation-state on a linguistic basis: the French language is deemed indispensable to proper revolutionary ideals, but is not considered an exclusive and sacred good. It is a kind of tool of citizenship and, through it, of the nation. Completely different is the importance of language in the creation of the German or Italian nation-state. Before its founding, language was the only unifying factor of scattered populations and therefore took a more affective meaning than in the French case.

Another factor of importance is religion. The Reformation played a key role in the rise of modern nation-states by emancipating them from the power of the Church, on the spiritual level as well as on the territorial level. However, for a sociologist like Rokkan, religious criteria seem of less importance than language. For him, language has a more defined territorial characteristic than religion and the notion of nation is linked strongly with a territorial foundation. That last factor is the decisive one: there is no nation without territorial identity. Therefore one must call on intricate processes, ethnicocultural, religious, linguistic, and politico-administrative factors that can go back a long way (Flora, 1999).

Other determinants could be named here, but it seems almost impossible to find a unique definition of the objective elements that would characterize the nation. The primordialist school would tend to offer a specific approach of the creation of each particular nation. Indeed, nations have been built on criteria much too different to allow for the constitution of a 'hard core,' even if modernization in each case has added its own particular homogenizing effect.

This controversy between modernists and primordialists, that is, between those for whom nationalism created the nation and those who look for links between nationalisms and preexisting national characteristics to explain their different development and intensity, is far from ending in the

sociological literature. For Hroch, these two approaches are not necessarily incompatible, for there is interaction and not dominance between nation and nationalism (Hroch, 1999). In this perspective, we can borrow Disraeli's lapidary expression that applies so well to national construction: of art and time.

The Future of the Nation

Since the nineteenth century, the world has become one of nations whose number is ever increasing in the framework of the United Nations. Some of them have been the result of Europe's export of her own model of nation-state through colonialism. That model, despite its artificial character, has been largely 'instrumentalized' by movements of national liberation in order to consolidate their state authority (Badie, 1992). National demands inside even the best-established nation-states, directed against their domination (the so-called 'internal colonialism'), never stopped and the desire for a state sometimes led to ruptures, and sometimes to reorganizations of states on the regionalist or federalist models.

Today, however, nobody can deny that changes are at work on a global scale that put into jeopardy the future of the so-called modern nation. 'Mondialization' and most of all 'globalization' send a major challenge to most of the nation-states. Whereas nation-states drew most of their power from their quasiabsolute control over their territory, and particularly over their national economy, globalization deprives them increasingly of their real competences as well as their resources, while severely affecting their legitimacy, even though this globalization is on a 'shifting scale' according to different domains. At the same time, that evolution has been facilitated largely by the collapse of the Soviet Empire, which, however, gave free rein to an unexpected 'pent-up nationalism.'

The cultural melting pot combined with the prevalence of English as an international language and the mixing of populations and cultures, be it on the virtual level with Internet or in reality with population migrations and immigration to rich countries constitutes an evolution whose effects are particularly difficult to integrate in homogeneous cultures typical of classical nation-states.

Habermas states the problem in terms of a relatively simple alternative, at least on the politicoeconomic level: either the most threatened nation-states will withdraw in neo-protectionist 'tentatives' bound to fail, or they will try to reach 'the same level as the markets,' which entails their entering some sort of supranational integration. The world of culturally homogeneous nation-states will fade away, leaving the stage for new mixed entities, particularly on the level of cultural homogeneity (Habermas, 2000).

The big question of whether the process of globalization will dissolve nation-states into new and necessarily pluralistic entities, or will give rise to renewed nationalistic demands challenging it, or, yet, whether the two phenomena will go hand in hand, remains largely open, even if what can be observed goes in the sense of the third alternative. In this sense, the European Union represents the most advanced form of supranational integration in the world. Since Europe is also the cradle of modern nations, an analysis of the processes taking place there is of the utmost interest for the future of the nation.

During the 1990s and at the beginning of the twenty-first century, for most analysts, this double-faced process was supposed to continue without seriously threatening the advancement of European integration, which still profited from the 'permissive consensus' of the national populations, even if it was slowly declining. Partisans of European integration regarded with great astonishment and even with real fear that the results of the referenda about the Constitutional Treaty were clearly negative in France, the Netherlands, and Ireland (2005). The effect was a crisis in the development of the integration process and the discovery of an antagonism deeper than anticipated before between Europe and important segments of national populations. The financial crisis of 2007 and the following crisis of the public debts of nation-states did not help at all to reduce the gap – even if the only real solution of this latter crisis may lie in a more federal union, a solution which is not supported until now by most of the nation-states. It seems, on the contrary, that, for the present time, most of the populations prefer to find solutions in the safety net offered by their own states. Does this signal mean a revival of the old nationalism of the European peoples? Not exactly – with the exception of some extreme populist movements – but rather a 'sovereignist' (*souverainisme* in French) attitude, which can be conceived as a form of functional nationalism, seeking more protection for maintaining their privileged situation, on social and political grounds, without having to share their resources with neighbors, rather than an aggressive attitude vis à vis the other nations perceived as enemies. It is more than a kind of 'banal nationalism' in the sense of Billig (Billig, 1995), but not very much more.

One can conclude that the future of nations is uncertain. They are subject to contradictory forces. Nations are affected by the globalization processes, producing for mixed populations a new standardized culture, which weakens the importance of territories and consequently of nations and national cultures. But this evolution can also trigger a reaction of these national cultures, with very different kinds of effects, from boosting national pride, but not in an antagonist way, to attitudes close to ethnocentrism and beliefs in 'clash of civilizations' theories (Duchesne and Frognier, 2008). At the same time, as far as Europe is concerned, it is difficult today to predict whether the European Union's efforts to manage the present crisis will succeed, with the implication that national egoism of the rich countries will have to be tempered, or if the European integration process will come to a halt or even will return to earlier, lower degrees of integration. A Pyrrhic victory for the European nations?

See also: Citizenship, Sociological Aspects of; Ethnic, Racial, and Nationalist Movements; Ethnicity, Sociology of; Identity Movements; Ideology, Sociology of; Institutions; Legitimacy, Sociology of; Minorities; Modernity; Modernization, Sociological Theories of; Nationalism, Sociology of; Race Identity; Racism, Sociology of; Social Identity in Sociology.

Bibliography

Anderson, B.R.O.'G., 1983. Imagined Communities. Verso, London.
Badie, B., 1992. Etat importé; (The Imported State). Fayard, Paris.
Bauer, O., 1907. Nationalitätenfrage und die Sozialdemocratie (National Question and Social Democracy). Verlag der Wiener Volksbuchhandlung, Vienna.
Billig, M., 1995. Banal Nationalism. Sage, London.

Breuilly, J., 1992. Nationalism and the State. Manchester University Press, Manchester, UK.
Delfosse, P., 1994. Nationalisme et parlementarisme. Res Publica 36 (2), 179–219.
Duchesne, S., Frognier, A.-P., 2008. National and European identifications: a dual relationship. Comparative European Politics 6 (2), 143–168.
Flora, P. (Ed.), 1999. State Formation, Nation-Building, and Mass Politics in Europe: The Theory of Stein Rokkan. Oxford University Press, Oxford, UK.
Gellner, E., 1983. Nations and Nationalism. Oxford University Press, Oxford, UK.
Habermas, J., 2000. Après l'Etat-nation. Une nouvelle constellation politique. Fayard, Paris (After the Nation State. A New Political Constellation).
Hobsbawm, E., 1990. Nation and Nationalism since 1780: Program, Myth, Reality. Cambridge University Press, Cambridge, UK.
Hroch, M., 1985. Social Preconditions of National Revival in Europe. Cambridge University Press, Cambridge, UK.
Hroch, M., 1999. Real and constructed: the nature of the nation. In: Mortimer, E. (Ed.), People, Nation and State. The Meaning of Ethnicity and Nationalism. Tauris, London, pp. 91–107.
Schulze, H., 1996. Etat et Nation dans l'Histoire de l'Europe. Seuil, Paris (State and Nation in the History of Europe).
Smith, A.D., 1986. The Ethnic Origin of Nations. Blackwell, Oxford, UK.
Smith, A.D., 1999. The nation: real or imagined. In: Mortimer, E. (Ed.), People, Nation and State. The Meaning of Ethnicity and Nationalism. Tauris, London, pp. 36–43.
Stalin, I., 1937. Le marxisme et la question nationale et coloniale. Sociales, Paris (Marxism and the National and Colonial Question).

National Account Systems

Werner Reichmann, University of Konstanz, Konstanz, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article addresses the historical development of the System of National Accounts (SNA) as a process of negotiating and struggling about concepts and definitions. It starts with an explanation of the idea of the SNA, presents briefly some of its predecessors and shows its chronological development in the twentieth century, national differences in the adoption of the SNA, and the main cleavages in the international harmonization process. Then, it goes deeper in the revision of the most recent SNA and, finally, shows its political relevance.

Introduction

The System of National Accounts (for short, SNA) is the most comprehensive tool that aims at describing and analyzing economical activities within a defined geographical area of a defined time period. It is also used to forecast economic growth and produces the perhaps most visible and powerful scientifically produced number: the gross domestic product (for short, GDP). In the second half of the twentieth century the SNA became a powerful tool for policy making and economic planning. Actors on financial markets are also geared to the quarterly published data produced in the framework of SNA (Kenessey, 1994, p. 1). The SNA is "the internationally agreed standard set of recommendations on how to compile measures of economic activity in accordance with strict accounting conventions based on economic principles. The recommendations are expressed in terms of a set of concepts, definitions, classifications and accounting rules that comprise the internationally agreed standard for measuring such items as gross domestic product (GDP)" (SNA, 2008). The SNA gives a systematical overview of how income, production, and distribution of goods and services are related to each other and how they are distributed to consumers, businesses, governments, and foreign nations and considers how taxes and social transfers influence the flow of capital. The SNA is intended for global use and it should be flexible enough that countries of different levels of development can implement it. It is a tool for policy making, for statistical and scientific use, and it is the main information basis that informs a society about its own economic development and the development of other societies. It should provide an overview of the overall economy, its main sectors, and their relationships to each other and should unify the view and the epistemic use of economy on a global level. When economists describe the SNA they tend to use metaphors like the "bird's eye view on the economy" (Gilbert und Jaszi, 1944) indicating that it provides a distant view on the economy, or they use the image of a "radiograph of the economy" (Frentzel-Wagner, 1959), respectively, a "tool for [economic] diaphanoscopy" (Nützenadel, 2005, p. 90) to point to the SNA's ability to look 'into' the economy.

In 1939 Joseph A. Schumpeter wrote about the national income, the predecessor of SNA, that "[t]he familiar difficulties which we all experience in defining National Income are of course due to the fact that it is not a technical term wedded to one definite use but a word of common parlance that is loosely used for a great many purposes which cannot be served equally well by a single definition" (Schumpeter, 1939, p. 561). In contrast to costs that can be measured by prices, the total output in terms of physical entities is a "figment which […] would not as such exist at all, were there no statisticians to create it. We seem indeed to be faced by a meaningless heap." (Schumpeter, 1939, p. 484). Considering Schumpeter's argument, the history of the SNA consists of at least 350 years of lasting struggle about this 'common parlance' and about giving the statistical 'heap' a meaning by defining what counts as economic activities and how they are grouped. The main problem in this struggle is to negotiate and to define accounts and their corresponding economic activities. It is this struggle about the "accounting structure" (Vanoli, 2005, p. 109) that employed a good number of economists.

At least four of them were awarded with the Nobel Prize for works on national accounts: Jan Tinbergen in 1969, Simon Kuznets in 1971, Wassily Leontief in 1973, and Richard Stone in 1984. Stone and Kuznets were decorated by the Nobel committee explicitly for their "pioneering research around national accounting systems" and for their contribution to the development of "methods for calculating the size of, and changes in, national income" (www.nobel.se, last update 31.12.13.).

The idea to measure a nation's total economy and to find out the economic strength of a whole country goes back to the seventeenth century, "a determined antiquarian can almost always find an earlier example, often by recourse to China" (Galbraith, 1980, p. 75). However, modern, systematic, and frequent compilations of the national product were set up in the 1930s, institutionalized in the 1950s, and internationalized in the 1990s. The struggle about definitions and standards is not over at the time of this publication. Still, it has calmed down and is regulated in controlled paths.

Predecessors

The first attempts to compile an overview of the economy as a whole were made by William Petty (1623–1687) in the seventeenth century. In his book from 1662 titled *Political Arithmetick* he expresses social and economic activities in quantitative terms. He was deeply concerned with the creation

of more reasonable methods to fix taxes. In his work he considered three accounts of income (land, estates, and work) and one for expenditures (summarizing clothes, food, housing, and all other necessaries) and the national income was a simple addition of the three income accounts (Studenski, 1958, p. 15; Vanoli, 2005, p. 4). Following Petty's work, a number of authors were engaged with compiling a National Income, such as Gregory King (1648–1712) (Studenski, 1958, p. 16) in England, or Pierre de Boisguillebert (1646–1714) in France, who tried to measure the economic strength of his country to compare it with others.

Several works followed and began to specify national income – the struggle about a "common parlance" (Schumpeter, 1939, p. 561) began. One of the problems was how to value goods. Thomas R. Malthus (1766–1834) argued to value them by market prices, and Adam Smith (1723–1790) wrote about 'natural prices.' Other discussions were about the relationship between national income and welfare. Arthur C. Pigou (1877–1959), for example, invented the notion of 'national dividend' and defined welfare to be a consequence of a high 'national dividend.' Another one was the question what should be included in the national income.

Alfred Marshall is called to be the first one to include not just products but also services and valued them by market prices (Lepenies, 2013, pp. 39–46). In contrast, the French physiocrats even limited the national income to agricultural products. Another milestone was Francois Quesnay's (1694–1774) *Tablau Economique* because he emphasized on the *flow* of goods and money between social groups, that is "an anticipation of what was done in the twentieth century" (Vanoli, 2005, p. 12).

Early attempts to estimate the national income were limited by the access and reliability of data. The "enlightened amateurs" (Vanoli, 2005, p. 13) of the early days tried to gather the biggest possible amount of data and had to be very creative and flexible when compiling accounts and the main aggregates. Most of the estimates about national income included material production and material goods, the trade with these goods and transportation. And there are some good reasons to assume that the motivations were not just in public interest but also political: to pay less taxes, to make the own social class stronger, to show that one's own country is superior to others.

The SNA in the Twentieth Century

The early attempts of compiling the national income were noninstitutionalized initiatives of private persons who had quite diverse motivations to conceptualize and represent economies as a whole (see e.g., the story of Roger Babson in Friedman, 2014). At the beginning of the twentieth century this changed twofold. First, the organizational frame of economic statistics changed from private to government and increased in quality (Speich Chassé, 2013, p. 96). The end of the "hero phase" (Vanoli, 2005, p. 15) started when Canada, as the first country worldwide, began to regularly publish estimates of its national income. That was the time when economic statistics got out of the hands of amateurs and private thinkers and administrative units took over, began to gather data and compiled economic aggregates. The new organizational frame is a result of the economic crisis in 1929 when governments began to think about how to intervene in economic development (Lepenies, 2013, p. 122). It is also connected with John M. Keynes' general theory that established the concept of macroeconomy and provided the theoretical basis to end a laissez-faire policy (Kurabayashi, 1994). And it was also the result of the needs of a war economy and the demand for an epistemic system that documents and legitimates governments' actions.

Second, there was a conceptual shift. The perspective changed from the emphasis of a nation's income to the sum of the production of a nation's economy. The former conceptualized a nation's economy in terms of capital and incomes of businesses and individuals. The latter emphasizes products of economic sectors, summarized in accounts. In the first third of the twentieth century, several authors had the idea to extend double-entry bookkeeping and accounting from a business level to the whole economy. Irving Fisher, Morris A. Copeland, and Robert F. Martin in the United States; André Vincent in France; Ferdinand Grüning in Germany; and Ed van Cleff in the Netherlands suggested different ways to think about national economies in terms of national accounts (Vanoli, 2005, pp. 16–17).

It might be arbitrary to identify one single starting point of the international spread of the SNA. However, Richard Stone's and James Meade's paper from 1941, with their suggestions on how to structure and compile an SNA, is widely accepted as the historical starting point of an international development of adoption and harmonization of the system national economic accounts as well as discussing the problems and limits of a unique SNA (Meade and Stone, 1941; **Table 1**).

This article landed on fertile soil. Even in 1939 the League of Nations had the idea to compare countries by a standardized scheme and started to collect data – but World War II stopped

Table 1 Milestones of the SNA-Development

1941	Richard Stone and James Meade publish their paper on national accounting
1945 (Sept)	United Kingdom, USA, and Canada meet to discuss possible ways to harmonize their national accounts
1945 (Dec)	The League of Nations meet in Princeton where Richard Stone presents his system of national accounts (report published in 1947)
1949–1950	OEEC simplified system
1952	OEEC standardized system
1953	UN published first version of SNA 1953, revised in 1960 and 1964
1968	UN published SNA 1968, preparation 1964–1968
1970	European System of Accounts (ESA), preparation 1964–1970
1971	UN publishes the Basic Principles if the System of Balances
1977	Comparisons of the SNA and the System of Balances
1989	UN publishes a revised version of the System of Balances
1993	UN, IMF, World Bank, OECD and European Union publishes the SNA 1993
1995	European Union publishes the ESA 1995
2008	UN, IMF, World Bank, OECD and European Union publishes the SNA 2008
2010	European Union publishes the ESA 2010

this attempt. Furthermore, in September 1945 Canada, Great Britain, and the USA organized a tripartite conference to negotiate about the possibilities of a common standard about national income estimates and about the establishment of a common SNA (Vanoli, 2005, p. 24).

In December 1945, Richard Stone had the chance to present the ideas summarized in the paper mentioned to a conference organized by the League of Nations in Princeton. As a result, this conference published a report titled "Measurement of National Income and the Construction of Social Accounts" that was fully accepted by the United Nations in 1947. Though its status is challenged – Vanoli (2005, p. 26) argues that it was even a technical paper, Bos (1994) treats it equal to the all SNA recommendations – this report was a hallmark of the international process of the harmonization of the SNA.

The 1947 report was the basis of several follow-up papers. The OEEC, established in Paris in 1948 to organize the recovery program in Europe, published a 'Simplified System' in 1950 and in 1952 the OEEC's 'Standardised System of National Accounts' came out, closely followed by 'System of National Accounts and Supporting Tables,' published by the United Nations. Until 1965 the OEEC compiled its own system of accounts parallel to the United Nations. The main motivation for that was to establish comparable indicators to estimate economic wealth and growth rates of the countries that profit from the Marshall plan. As the economic recovery in Europe succeeded there was no more use for two parallel SNAs (Vanoli, 2005, pp. 89–90).

The SNA was generally revised for several times. On average, a SNA has a tenability of about 18.5 years. General revisions were made in 1968, 1993, and 2008 and include completely new accounts structures. For the first time in 1970, the 'European Communities' published their own accounting system called European System of National Accounts (for short, ESA) with a time gap of 2 years to the SNA (Vanoli, 2005, pp. 91–98). So, whereas the OECD stopped making its own system, the European Union is still compiling its own system of accounts which is a revised and adapted derivate from the SNA.

(Inter-)National Differences

In the 1950s only a small number of member states of the UN completely implemented the recommendations of the SNA 1953. The basic system of accounts was similar in Western countries but differentiated in details. They had in common to emphasize the production side of the accounts and the amount of the GDP (Lepenies, 2013, p. 159). At the same time a global process of harmonization of the national accounts structure started – a process that is not completed nowadays. In Western Europe, the Economic Cooperation Administration (ECA) that coordinated the financial aid for the economic recovery played a crucial role for the standardization of the SNA. It aimed at estimating the economic potential of European countries after World War II to define standards for economic aid. The ECA founded the OEEC and asked it, together with the 'National Accounts Research Unit' at Cambridge, to develop a simple but standardized system of accounts for European countries and trained European economists and administration staff to gather data and to use the account system. Thus, the US help to recover European economies was also a way to harmonize the SNA.

However, the harmonization process was not this successful in other parts of the world. The SNA was not adopted by all the countries at the same speed and in the same manner. As it was a Western concept, mainly developed in Great Britain and the United States, the Soviet Union was skeptical to implement it and went its own way, called 'Material Product System' (MPS) with the main difference that only material products add value to the economy. A predecessor of the MPS was an index for the industrial output compiled by Nikolai D. Kontradieff. It was available for every year until 1885 (Gerschenkron, 1947, p. 145). After the breakdown of the Soviet Union, Eastern Europe implemented the SNA 1993 and so brought its internationalization a step forward (Vanoli, 2005, pp. 100–101). China used the MPS and the western SNA in parallel between 1985 and 1992. Afterwards, China changed to the use of the SNA 1993, and the MPS disappeared globally together with the USSR.

The United States played an ambivalent role in the harmonization process (Speich Chassé, 2013, pp. 149–45). Although US economists and organizations were leading in the conceptual foundation of the SNA, the US administration never implemented the system changes and built its own account structure. The SNA 2008 was the first one completely used by the United States. And USA's commitment in the revision processes was quite different in the 1960s and the 1990s. Though, they were deeply engaged in the development for the 1993-SNA, they never completely implemented it for reasons of law specialties and personnel changes in the administration staff (Vanoli, 2005, p. 125).

Germany was never a driving force in the development and internationalization of SNA – Reich calls it a "latecomer in national income estimating compared to other countries" (Reich, 1994, p. 148). Although there are some early attempts to analyze economies as a whole, German economists especially from the 'Historical School' denied any approach to estimate national income in the 1920s (Reich, 1994, p. 149; Studenski, 1958; Speich Chassé, 2013, pp. 67–80). In the first decade after World War II political actors were skeptical about the SNA as a whole. It was ill-reputed as a tool of illegitimate state intervention and did not fit into the German postwar economic policy. In contrast to that, France very early recognized that SNA was also the "accounts of power" (Nützenadel, 2005, p. 102).

In the 1970s many countries in the southern hemisphere started to implement the 1968-SNA. It was France who helped many third world countries to get systematically informed about their economies. Until today, there are controversies over the SNA's role in less developed countries' economies (Vanoli, 2005, p. 104).

The Main Cleavages during the Process of Harmonization

In his Nobel prize lecture, Richard Stone reported that in the 1950s there was just a small number of statisticians who did fully understand the ideas and conceptions of the SNA. Thus, there were less discussions when nations tried to harmonize the accounts. The duration of the discussion process and the complexities of the system increased with the number of well-informed participants (Lepenies, 2013, p. 132). Still, it is not just the higher number of actors involved, but also their

unclear roles, and their different national backgrounds with different ways to decide on a national level (Vanoli, 2005, p. 87). The number of countries and organizations involved in the revisions of the SNA increased enormously during the twentieth century as did the size of the system that grew from 57 pages in 1953 to 722 pages in 2008 (with a peak of 838 pages in 1993). This growth of about 1200% shows the complexity of an SNA composed for global use.

What are the main conflict lines that can be identified in the history of establishing global conventions for the SNA? Mainly, there is a continuous discussion about what economic activities are and which economic activities should be included in an estimate covering the whole economy. A second general question is, what role the state should play within the concept of the national income and the SNA. Connected to this problem, there is a broad discussion about the aims of economic activities. Is it only to produce more goods or is it to make a society happier? With the focus turned from national income to the economical production side, a further conflict came out: Are the aggregates (e.g., the GDP) or the accounts, i.e. the basis of the aggregates, more important for an SNA (Vanoli, 2005, p. 28)? A quite early discussed problem is the idea to use national income statistics and national accounts to compare countries (e.g., Rosenstein-Rodan, 1944). This question is still open as the analysis of differently structured economies and different natural resources using the same theoretical approach is often criticized (e.g., Morgan, 2011). Gerschenkron's (1947) attempt to analyze the economic growth of Russia until 1885 is an impressive document of the problems to compile and to compare a less industrialized and politically planned economy with capitalistic economies.

Conventional Changes in the SNA

The frequent revisions of the SNA show its conventional character. Some of the recent changes of the ESA exemplify that. The ESA 2010 replaces the one from 1995. It is a general revision and will become res judicata in the second half of 2014 in all member countries of the European Union. In contrast to the SNA, the ESA is obligate to all members of the EU. Both are consistent in their basic concepts, but the latter includes further and more precise rules to gather and compile the data. This is necessary because the ESA is the basis for the calculation of the contributions to the EU budget, it compiles the ratios of deficit and debt and thus, it is a tool to keep the member states under surveillance. This makes the ESA a politically relevant system. In all member countries a revision of the accounting system makes a deep recalculation of the data sets necessary to keep the data from the past comparable with present and future data. Every revision produces costs and needs much organizational and computational work. So, why is it necessary to revise the SNA regularly? These revisions enable the SNA to represent recent economic and social developments adequately and to bring macroeconomic statistics in line with the recent economic environment and the needs of the users (Statistisches Bundesamt, 2013).

There are two kinds of revisions: first, there are technical and methodological revisions most users of the SNA do not notice, for example, new grouping conventions for sectors of economic activities. These revisions do not change the aggregates. Second, and more interesting to the public are changes in the conceptualization of the accounts. Normally, these changes are reactions to developments in economy and society. One of the main changes of the ESA 1995 may exemplify that. Whereas the ESA 1995 defined costs for Research and Development as current expenditures, the ESA 2010 treats them as investments. Especially in knowledge-intensive and digital economies these investments are important and therefore they should be part of the economy's output. The same is with the treatment of military goods and weapon systems. They will also be treated as states' investments in the future. Actually, both reclassification will influence the GDP. Preliminary estimates show that all changes between the systems of 1995 and of 2010 will increase the GDP +3% for the United States and +2.4% for the EU. And they will also increase the analytical capacity of the balance sheets and their fitting to reality. Another change regards the treatment of the flow of stocks and financial products through the world as a reaction to globalization of the economy (SNA, 2008, p. xlviii; see also Vanoli, 2005, pp. 109, 124). And the revision will also consider new data sets to make the accounts more precise and up-to-date, e.g., the statistics of people in paid labor, the data about where and how inhabitants live.

The Political Nature of the SNA

The struggle about definitions and the classification of economic activities are not mere academic questions or nerdy nitpicking. Besides, it is also a political conflict. Even in the early attempts of compiling national income estimates there were actually political conflicts behind the scientifically discussed questions. William Petty's concepts and computations aimed at showing that tax charges were distributed uneven between the classes. He argued that income from labor was higher than it was supposed to be and that it represented a higher percentage of the national income than it was supposed to. Petty aimed at showing that the impact of his own class was smaller and therefore, tax charges should be smaller (Lepenies, 2013, p. 35). Another sign of the political nature of National Income and SNA estimates is that many of the early authors were offended for political reasons. Their books were forbidden and destroyed or authorities decided to keep them secret because of both, the critical topic per se and because of the conclusions the authors draw from their data (Vanoli, 2005, p. 5).

An example from a long-lasting struggle in the 1930s and 1940s shows that different economical classifications lead to contrary political results. In Simon Kuznets conception of the national income the state does not play a crucial role. When he spoke about the national income he meant the 'national income paid out,' that is, the amount of money available to every household. For Kuznets, expenditures for armament, infrastructure, and transport (the costs for the 'economic civilization') are 'evil necessary' and decreases the whole amount of national income because the money spent on these goods is no more to the private households' disposal. For Kuznets, the states' expenditures decrease the national income (Kuznets, 1933). Before the United States was engaged in World War II, the Office of Price Administration and Civilian Supply

(OPACS) argued vehemently for an increase of the war budget but did not succeed with that political position. One of the reasons was that, using Kuznets' conception, it would decrease national income. Only a Keynesian approach could solve this problem. It argues the other way around: the states' investments increase the national income as they are also a form of production of goods, no matter who produces them for what.

The work of Meade and Stone was highly influenced by Keynesian ideas and, through a personal network between Simon Kuznets, Milton Gilbert, Richard Stone, and Richard Gilbert, the cousin of Milton, these ideas entered the 'Defense Economic Section' in the OPACS. With this classification of economic activity, the higher states' investments do not necessarily decrease national income but, actually, increases the 'social product.' The conceptual turn from income to product, that also was the reason for the break up between Simon Kuznets and his scholar Milton Gilbert, shows that different economic theories lead to different types of classifying economic activities leading to how the states' activities are handled and thus, how the state can legitimate its actions. It's classification, stupid! (Lepenies, 2013, p. 100)

Another historical struggle about how to measure economies in total is the one between the East and the West during the Cold War. The battle between two different political systems was also fought on the field of SNA definitions. In 1961 the Yugoslavian economist Branko Horvat reported the Soviet system of accounts, called Material Product System (MPS), and quotes a leading soviet economic textbook: "In Socialism, as in any other system, total social product is created by the labour of workers in branches of material production. Along with manual workers, brain workers (scientists, engineers, etc.), engaged in branches of material production, participate directly in the creation of material wealth. [...] Workers engaged in the non-productive sphere (state administration, culture, welfare, medical service), do not create material wealth. Nevertheless, the labour of workers of non-productive branches is indispensable for socialist society, for material production it represents socially useful labour" (Ostrovitjanov und Vasil'evič 1959 cited in Horvat, 1961, p. 234). Horvat argues that the Soviet definition fits better to measure 'economic' and 'military strength' and also "social welfare" (Horvat, 1961, p. 235) than the American one because it distinguishes between goods and "social waste" (Horvat, 1961, p. 238), e.g., weapons. 'Social waste,' Horvat says, may increase the American social product but would not make its society better. The definition differs from the concepts of both Kuznets and Stone and (Vanoli 2005, pp. 100–101) claims that it did not fit the reality but is the product of ideology.

A recent political debate about the significance of the SNA and especially about the aggregate in the public focus, the GDP, was initiated by the former French president Nicolas Sarkozy. Together with prominent economists he built a commission "to identify the limits of GDP as an indicator of economic performance and social progress" (Stiglitz et al., 2010, p. 7). Quite critical, it recommends to differentiate between mere economic measures, the peoples' well-being, and the evaluation of the sustainability of both. At about the same time, the German parliament built an enquete commission with economists and members of the parliament on the same topic. In April 2013, it published its final report where it agrees that GDP cannot be the only indicator that informs about the condition of and recommends splitting it up to three dimensions: wealth, social development, and ecological factors. Until now, the GDP is still the central economic indicator in political debates in both countries.

Conclusion

In the beginning of the twenty-first century the SNA is central for both economics and economic policy. The history of the SNA is not just one of struggles about scientific classifications and of the competition between political systems. It is also an impressive success story of a scientific concept as it becomes institutionalized in more and more countries and it is dominating the way how societies think and know national economies as well as global economic relations. The main aggregate of the SNA, the GDP, is something like a unique index in the scientific world as it gives a simple and accepted answer to the complex question, how a whole economy is doing. The criticism that is passed on the SNA will not calm down. Still, at the moment, there is no evidence that another economic concept will replace the SNA in the global discourse.

See also: Culture and Economy; Econometrics, History of; Economics, History of; Macroeconomics, History of From 1933 to Present; Political Economy, History of; Post-Socialism: Political Economy; Quantification in the History of the Social Sciences.

Bibliography

Bos, Frits, 1994. Constancy and change in the united nations manuals 1947, 1953, 1968 and 1993. In: Kenessey, Zoltan (Ed.), The Accounts of Nations. IOS Press, Amsterdam, pp. S.198–S.217.

Frentzel-Wagner, Hanno, 1959. Röntgenbild der Wirtschaft – Die volkswirtschaftliche Gesamtrechnung. Offene Welt – Zeitschrift für Wirtschaft, Politik und Gesellschaft 63/64, 520–538.

Friedman, Walter A., 2014. Fortune Tellers. The Story of America's First Economic Forecasters. Princeton University Press, Princeton.

Galbraith, John Kenneth, 1980. The national accounts: arrival and impact. In: Reflections of America. Commemorating the Statistical Abstract Centennial. US Department of Commerce & Bureau of the Census, Washington, pp. S.75–S.80.

Gerschenkron, Alexander, 1947. The rate of growth in Russia. The rate of industrial growth in Russia since 1885. Journal of Economic History 7, 144–174 (Supplement: Economic Growth – A Symposium).

Gilbert, Milton, Jaszi, George, 1944. National product and income statistics as an aid in economic problems. Dun's Review.

Horvat, Branko, 1961. The conceptual background of social product. The Review of Income and Wealth 9, 234–252.

Kenessey, Zoltan (Ed.), 1994. The Accounts of Nations. IOS Press, Amsterdam.

Kurabayashi, Yoshimasa, 1994. Keynes' how to pay for the war and its influence on post-war national accounting. In: Kenessey, Zoltan (Ed.), The Accounts of Nations. IOS Press, Amsterdam, pp. S.93–S.108.

Kuznets, Simon, 1933. National income. Encyclopedia of the Social Sciences 11, 205–224.

Lepenies, Philip, 2013. Die Macht der einen Zahl - Eine politische Geschichte des Bruttoinlandsprodukts. Suhrkamp, Berlin.

Meade, John, Stone, Richard, 1941. The construction of tables of national income, expenditure, savings and investment. Economic Journal 51 (202/203), 216–233.

Morgan, Mary S., 2011. Seeking parts, looking for wholes. In: Daston, Lorraine, Lunbeck, Elizabeth (Eds.), Histories of Scientific Observation. The University of Chicago Press, Chicago, London, pp. 303–325.

Nützenadel, Alexander, 2005. Stunde der Ökonomen - Wissenschaft, Politik und Expertenkulturen in der Bundesrepublik 1949–1974. Vandenhoeck & Ruprecht, Göttingen.

Ostrovitjanov, K.V., Vasil'evič, Konstantin, 1959. Političeskaja Ekonomija. Gospolitizdai, Moskva.

Reich, Utz-Peter, 1994. German national accounts between politics and academics. In: Kenessey, Zoltan (Ed.), The Accounts of Nations. IOS Press, Amsterdam, pp. S.144–S.168.

Rosenstein-Rodan, Paul, 1944. The international development of economically backward areas. International Affairs (Royal Institute of International Affairs 1944–) 20 (2), 157–165.

Statistisches Bundesamt, 2013. Methodische Weiterentwicklung der Volkswirtschaftlichen Gesamtrechnungen – die Revision 2014. https://www.destatis.de/DE/Methoden/VGRRevision/Revision2014.html. last update 01.02.14.

Schumpeter, Joseph A., 1939. Business Cycles: A Theoretical, Historical, and Statistical Analysis of the Capitalist Process. McGraw-Hill Book Company, Inc, London, New York.

Speich Chassé, Daniel, 2013. Die Erfindung des Bruttosozialprodukts. Globale Ungleichheit in der Wissensgeschichte der Ökonomie. Vandenhoeck & Ruprecht, Göttingen.

Stiglitz, Joseph E., Sen, Amartya, Fitoussi, Jean-Paul, 2010. Report by the Commission on the Measurement of Economic Performance and Social Progress. Download. www.stiglitz-sen-fitoussi.fr/.

Studenski, Paul, 1958. The Income of Nations. New York University Press, New York.

SNA, 1953. United Nations. 1953. A System of National Accounts and Supporting Tables, New York quoted as SNA 1953.

SNA, 1993. Commission of the European Communities, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations, and World Bank. 1993. System of National Accounts 1993, Brussels/Luxembourg, New York, Paris, and Washington, D.C., quoted as SNA 1993.

SNA, 2008. European Commission, International Monetary Fund, Organisation for Economic Co-operation and Development. United Nations and World Bank. 2008. System of National Accounts 2008, New York quotes as SNA 2008.

Vanoli, André, 2005. A History of National Accounting. IOS Press, Amsterdam.

National Character

Federico Neiburg, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 10296–10299, © 2001, Elsevier Ltd.

Abstract

National character is an expression which describes forms of collective self-perception, sensibility, and conduct which are shared by the individuals who inhabit modern nation-states. It presupposes the existence of psychological and cultural homogeneity among the citizens of each country, as well as the idea that each nation can be considered a collective individual, with characteristics analogous to the empirical individuals who are its inhabitants. The noun *character* seeks to describe a universal aspect of social life – an internal dimension to the existence of individuals and an external one, observable through collective behaviour. The adjective *national* situates this universal aspect of social life in the specific context of those social units we call nations. Social theory interested in understanding the social force of feelings of national belonging has turned once again to this expression, which was first formulated in Europe during the second half of the eighteenth century. What distinguishes this trend is the fact that there is no attempt at attributing any theoretical status to national character; instead, it is more concerned with it as a practical category used in the discourse and action of the social agents and groups. The aim of this article is to summarize the genealogy of the expression and to discuss its current heuristic value.

National character is an expression which describes forms of collective self-perception, sensibility, and conduct which are shared by the individuals who inhabit a modern nation-state. It presupposes the existence of psychological and cultural homogeneity among the citizens of each country, as well as the idea that each nation can be considered a collective individual, with characteristics analogous to the empirical individuals who are its inhabitants. The noun *character* seeks to describe a universal aspect of social life – simultaneously an internal dimension to the existence of individuals and an external one, observable through collective behavior. The adjective *national* situates this universal aspect of social life in the specific context of those social units we call nations.

Since it was first formulated in Europe during the second half of the eighteenth century, the idea that the inhabitants of each nation possess psychological and cultural characteristics in common, which confer upon them identity and distinguish them from others, has always been a part of the ideology and practice of nationalism. Hence its double status, being simultaneously a descriptive and a prescriptive expression. On the one hand it describes communities of custom, forms of being and feeling, on the other it serves to make legitimate claims for sovereignty and forms of political action. One refers to the essence of each people or nation, the other supposes that collective action can serve to favor or even to mold the national character of a political community in the making.

During the 1940s and 1950s, the expression national character gained scientific legitimacy within the so-called North American anthropological school of Culture and Personality, to the point of designating an object of study and a university specialization: national character studies. From then on, it has been at the center of various academic controversies, where many have questioned its theoretical status, attributing its scientific shortcomings to its close relationship with the essentialism characteristic of nationalist ideologies. Yet recently, social theory studies interested in understanding the social force of feelings of national belonging and the new forms of nationalism have turned once again to the expression. What distinguishes this trend from national character studies is the fact that there is no attempt to attribute any theoretical status to the expression; instead, current studies are more concerned with the expression national character as an object, a practical category used in the discourse and action of the social agents and groups which one seeks to understand (for an argument in favor of the study of the categories of national identification as 'practical categories' see Brubaker, 1996).

National Character and Nation-states

Questions dealing with national self and unself have been historically associated with the origin of nation-states and with international relations. While the processes leading to the emergence of new nation-states served to substitute older national elites for new ones, redefining the borders between countries along the way, the belief in the singularity of *one* national character, also demanded the recognition of a plurality of national characters from which it became necessary to differentiate.

One of the most detailed examples of the relationship between we-ideals and we-images in the formation of nation-states can be read in Norbert Elias' studies of the Germans (see Elias, 1994, 1996). Discussions concerning 'national character' in Germany emerged within the context of an encounter between new intellectual figures and a public capable of understanding and identifying with them. The poetry and philosophy of the second half of the eighteenth century (authors such as Herder, Schiller, or Goethe and movements such as *Sturm und Drang*) celebrated the ethic and aesthetic community grounded in the concept of *Kultur*, whose principal, distinguishing element was the use of the German language in opposition to the French spoken in the German courts by the social elites of the time. One century later, when the heirs to this intellectual tradition, backed by sections of the military and the old imperial nobility, established themselves within the German nation-state, words such as *Kultur* (alongside words

such as *Volksgeist* and *Zeitgeist*) had already become 'nationalized,' thus designating the we-ideals and the we-images of social groups in a larger scale and describing forms of collective self-perception and the 'German National Character.'

The diffusion of the nation-state as a means for organizing social life (see, for example, Balibar, 1991 [1988] and Anderson, 1983) brought about similar processes in other areas of the world. At the start of the twentieth century, and particularly after World War I, the end of the old Empires served to reinvigorate nationalist ideologies in central and eastern Europe. Ideas about the national character of the Hungarians, Poles, Czechs, Romanians, Croats, and Serbs began to inform the unification policies of new states. By then, however, debates about 'national character' were no longer confined to European intellectual circles. In America too, the most renowned authors were those who succeeded in describing the traits of the 'spiritual community' to their compatriots, who then recognized themselves in these. Their texts and ideas served as a guide for state politics, spreading beyond the literate elite and onto other social groups in the process of nation-making (particularly through the expansions of literacy and the school system). In Spanish-speaking America one of the best examples may be the Mexican writer Manuel Gamio, author of well-known statements on what it means to be Mexican, centered on the relations between the descendants of pre-Hispanic peoples and the Spanish (Gamio, 1916). In Portuguese-speaking America the best example is Gilberto Freyre, remembered as the author of one of the most influential representations of Brazilian national character, based on the idea of miscegenation between Whites, Blacks, and Amerindians (see Freyre, 1933).

Among the many similarities between the two authors is the fact that during the 1920s both resided in the USA and studied under the German anthropologist Franz Boas at Columbia University. Boas is largely recognized as one of the intellectual fathers of the Culture and Personality school which later served as an inspiration behind national character studies (on the Culture and Personality School see the articles collected in Stocking, 1986).

National Character Studies: A War-time Anthropology

The Culture and Personality School was marked by two main ideas. The first, based on the concept of primary socialization, sought to explain the relationship between types of culture and types of personality based on the experiences of childhood (see Mead, 1928). The second, centered around the idea of patterns of culture, intended to describe and psychologically profile cultural configurations and the types of personality associated with them: each culture, although made up of fragments from different origins, developed a specific 'style,' capable of bringing about its unification as a synthetic totality and, at the same time, distinguishing it from other cultural units (see Benedict, 1934). Even though the relationship between the concept of patterns of culture and the theories concerning German national character were very explicit (in her books, for example, Benedict mentions Dilthey and Spengler), the use of the concept was primarily restricted to primitive or simple (i.e., non-national) social forms, the traditional object of anthropological inquiry.

The use of these ideas in attempts to understand cultural configurations and personalities within national communities was, in fact, closely tied to the experience of the extreme and intense form of international relations which is war. For it was immediately after North-American intervention in World War II that a group of anthropologists (foremost among them were Gregory Bateson, Ruth Benedict, Geoffrey Gorer, Clyde Kluckhon, Ralph Linton, Margaret Mead, Rhoda Métraux, and George P. Murdock) became involved with the Office of War Information, intending to place the theories and methods of social anthropology at the service of the war effort. Its first goal became to produce a statement about national character which would help maintain the high morale of Americans during the conflict. Soon afterwards, its activities turned towards attempts at understanding allied and, specially, enemy nations.

The end of the war led to what would later be considered the most convincing evidence of the efficacy of national character studies. After Japan's surrender, the American government decided to maintain the Japanese emperor in power, as had been recommended by the Office of War Information in a memorandum signed by Gorer. This helped to ensure stability in postwar Japan and to guarantee that, despite Hiroshima and Nagasaki, a new state of equilibrium between the two countries could be established. This concurrence between US foreign policy and the anthropological analysis of the place of the Emperor in Japanese culture has been cited ever since as a fine example to the kind of 'applied anthropology' which national character studies allows.

But it was in the early 1930s that anthropological solutions to 'practical' problems began to be elaborated in the USA. Up until World War II, applied anthropology had been a matter for internal affairs, aimed at solving problems derived from American cultural and social diversity. In shifting the field of applied anthropology to external affairs, national character studies crossed a significant threshold in the history of anthropology. According to the division of labor established between the disciplines at the time, anthropology had been confined largely to the study of social forms different from those of the observer, and thus different from national societies and cultures. Now, for the first time, anthropologists began to deal explicitly with nation-states as an object of study. The discussions which then followed went beyond the concerns of anthropologists alone, and began to interest all those who sought to comprehend national forms of social and cultural life (for an analysis of the relations between theory and politics in national character studies, see Neiburg and Goldman, 1998).

First the USA and the Americans, then Japan and the Japanese, Germany and the Germans, Great Britain and the British; a little later (already in the context of the Cold War) Russia and the Russians, Poland and the Poles (see, for example, Gorer and Rickman, 1950) – all came to be treated as cultural worlds susceptible to analysis on the basis of the same categories used in the study of so-called 'simple societies,' those to which, until then, the majority of anthropologists had restricted their attention. Frontiers between countries were treated as cultural, and cultural frontiers were treated as national, thereby delimiting new objects for analysis. The context of war and of conflicts between nations demanded, however, a methodological 'development': 'studies of culture at a distance' (see Mead and Métraux, 1953; Mead, 1953). Anthropologists, prevented

from carrying out the traditional task of prolonged participant observation in the field, would now interview immigrants, analyze novels or films made in other countries, or examine media productions, so as to reach those cultural universes which were always qualified as *national*.

So began a tradition of research based on empirical sources which can still be recognized today, and not only in anthropology but also in other disciplinary traditions interested in the study of national cultures or societies (notably the so-called cultural studies). Yet it is interesting to draw attention to the paradoxical nature of the idea of 'studies of culture at a distance' for, in studying nation-states, researchers were socially and culturally 'closer' than traditional anthropologists in relation to their objects. Thus the fact that traditional methods and practices of anthropology (such as participant observation) were rarely made use of seems to have more to do with the specific properties of the object of study (or with the nature of the relations between researcher and object) than with the supposed impossibility of direct access to it.

National Character Reframed

In the theoretical debates concerning the expression 'national character' we can recognize some of the main questions surrounding contemporary discussions on the status of the nation-state as an object of the social sciences. One finds arguments in favor of the need to renew theoretical models in order to conceptualize empirical objects which need to be thought of as a historical novelty – such as those connected with processes of decolonization, localization, and transnationalization of social and cultural life – or denunciating the supposed limitations behind theoretical models whose origin would be a mere reflection of the world of nation-states which these models were developed to explain in the first place.

The main objections to national character studies can be summed up in three points. First, cultural patterns (obtained through the direct observation of empirical behavior) may become the principles through which one then seeks to explain these very patterns – a vicious circle which appears to be characteristic of culturalism in general (for a critique, see Kuper, 1999). Second, observed behavior in a fraction of the society analyzed would then be abstracted and attributed to all nationals, without a critical enquiry into the social position occupied by that fraction within national society. As such, the idea of national character would serve to disguise a more profound sociological reality; it would become an 'ideology' in the Marxist sense. Finally, in its rush to discuss racist theories, the Culture and Personality school (and, by extension, national character studies) may have shifted from a biological essentialism to a cultural or psychological one, which may then lead it to neglect other dimensions of social life, fundamental to an understanding of national realities and the elaboration of feelings of national belonging – such as institutional aspects related to the presence of the state, or the relations of differentiation, force, and domination between social groups (see Leite, 1969 [1954]).

In an article first published in 1942, Gregory Bateson outlined what remains the most convincing argument in favor of studies of national character. First, he argued that the fact that national characters are constructs does not mean that they do not exist and therefore do not serve to emphasize cultural and social differences. Second, he affirmed that to speak of national character does not imply cultural uniformity, but can, on the contrary, stress the existence of a system of differentiations organized in relation to a pattern (a set of stereotypes) recognized as being dominant. In this sense, national borders (political, cultural, and 'character-based') would serve as a good example of how something which is socially generated can lead to certain effects, acting on the forms of self-perception of the very communities which they differentiate and constitute, ultimately affecting ways of feeling, living styles, and moral configurations (for a similar argument see M. Mauss', 1969 [1921]) classical text; good examples of anthropological studies heading in this direction include Handler (1988) on the objectification of national cultures and Herzfeld (1993) on the social productivity of national stereotypes).

Feelings of national belonging and national forms of organizing social life are, like the Roman god Janus, two-faced. One face is that of the solidarity between nationals, between those who stand united through a relatively homogenous social and cultural base. The other reveals itself at the national borders and the attempts at imposing them and re-enforcing them, through exclusion and nationalist violence. The sociological and anthropological understanding of both faces – as well as of the social force of feelings of national belonging, which are, in effect, the basis for their existence – constitute one of the central themes in contemporary social science. A careful examination of the diverse contents of the discussions brought about by the idea of national character shows us how an advance on this understanding must pass through three theoretical frameworks: we must situate feelings of national belonging alongside other modes of social belonging and identification (ethnic, racial, regional, transnational) as a means of discovering what they may owe to (and reveal about) the universal properties and forms of social life; we cannot ignore the historical and sociogenetical processes behind the formation of categories of identity and the homogenization of forms of being and feeling; and, finally, it is essential to situate the comprehension of national characters as a dimension of the processes which objectify realities (cultures, societies, identities, stereotypes, etc.), both national and international.

See also: Anthropological Approaches to the Study of Education: The United States and Beyond; Benedict, Ruth (1887–1948); Cultural History; Cultural Influences on Interpersonal Relationships; Culture and Emotion; Elias, Norbert (1897–1990); Gender and Feminist Studies in Geography; Mead, Margaret (1901–78); Nationalism, Sociology of; Nationalism: General; Psychological Anthropology; Religion: Nationalism and Identity; Sexual Orientation: Biological Influences; Values, Anthropology of.

Bibliography

Anderson, B., 1983. Imagined Communities. Verso, London.

Banac, I., Verdery, K. (Eds.), 1995. National Character and National Ideology in Interwar Eastern Europe. Yale Center for International and Area Studies, New Haven, CT.

Balibar, E., 1991 [1988]. The nation form: history and ideology. In: Balibar, E., Wallerstein, I. (Eds.), Race, Nation, Class. Verso, London.
Bateson, G., 1978 [1972]. Morale and national character. In: Steps to an Ecology of Mind. Granada, London.
Benedict, R., 1934. Patterns of Culture. Houghton Mifflin, Boston.
Benedict, R., 1946. The Chrysanthemum and the Sword. New York and Houghton Mifflin, Boston.
Brubaker, R., 1996. Nationalism Reframed. Nationhood and the National Question in the New Europe. Cambridge University Press, Cambridge, UK.
Elias, N., 1994 [1978]. The Civilizing Process: The History of Manners and State Formation and Civilization. Blackwell, Oxford, UK.
Elias, N., 1996. The Germans. Power Struggles and the Development of Habitus in Nineteenth and Twentieth Centuries. Columbia University Press, New York.
Freyre, G., 1933. Casa-Grande & Senzala (Fromação da Familía Brasileira sob o Regime de Economia Patriarcal). Rio de Janeiro: Maia & Schmidt Ltda (published in English 1946 The Masters and the Slaves (A Study in the Development of Brazilian Civilization). Translated by Samuel Putnam. Afred A. Knopf, New York.
Gamio, M., 1916. Forjando Patria: Pro nacionalismo. Porrua Hermanos, México.
Gorer, G., 1943. Themes in Japanese culture. Transactions of the New York Academy of Sciences 2 (5), 106–124.
Gorer, G., 1948. The American People, first ed. Norton, New York.
Gorer, G., 1955. Exploring English Character. Cresset Press, London.
Gorer, G., Rickman, J., 1950. The People of Great Russia. W. W. Norton, New York.
Handler, R., 1988. Nationalism and the Politics of Culture in Quebec. University of Wisconsin Press, Madison, WI.
Herzfeld, M., 1993. The Social Production of Indifference. University of Chicago Press, Chicago.
Kuper, A., 1999. Culture – The Anthropologists' Account. Harvard University Press, Cambridge, MA.
Leite, D.M., 1969 [1954]. O Caráter Nacional Brasileiro. História de uma Ideologia, second ed. Livraria Pioneira Editora, São Paulo, Brazil.
Mauss, M., 1969 [1921]. La Nation. In: Œuvres III. Minuit, Paris, pp. 571–625.
Mead, M., 1928. Coming of Age in Samoa. Penguin, Harmondsworth, UK.
Mead, M., 1965 [1942]. And Keep Your Powder Dry. An Anthropologist Looks at America. Morrow, New York.
Mead, M., 1953. National character. In: Tax, S. (Ed.), Appraisal Anthropology Today. University of Chicago Press, Chicago.
Mead, M., Métraux, R. (Eds.), 1953. The Study of Culture at a Distance. University of Chicago Press, Chicago.
Neiburg, F., Goldman, M., 1998. Anthropology and politics in studies of national character. Cultural Anthropology 13 (1).
Smith, A.D., 1986. The Ethnic Origins of Nations. Blackwell, Oxford, UK.
Stocking Jr., G.W. (Ed.), 1986. Malinowski, Rivers, Benedict and Others. Essays on Culture and Personality. University of Wisconsin Press, Madison, WI.

National Character, History of

Stephen Mennell, School of Sociology, University College Dublin, Belfield, Dublin, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of 'national character' needs to be used, if at all, with great caution, although its history is of some interest. Discussions of the topic can be traced back at least as far as the writings of Alexis de Tocqueville. The concept was particularly current in the work of anthropologists of the 'culture and personality' school, and in the efforts of other social scientists seeking – during the Second World War and in its aftermath – to understand aspects of the behavior notably of German and Japanese people. Since, it has largely been subsumed into the concept of 'habitus,' and in that sense it has found a place in the writings of Norbert Elias and some of his followers.

The concept of 'national character' is now generally viewed by social scientists with considerable suspicion. It is redolent of the stereotyping of members of a nation, or of a social category, which played a part in some of the great sociopolitical disasters of the twentieth century. More innocuously, notions of national character enter into popular humor: witness jokes of the 'There was an Englishman, a Scotsman and a Welshman …' type, which dates back at least to Shakespeare. Or there are light-hearted but often illuminating popular books such as the Scot A.G. Macdonell's witty observations of the English, or the Hungarian George Mikes's comments on both the English and the Americans, or series like the 'Xenophobes guides' (which are not xenophobic – see for example the *Xenophobe's Guide to the Italians* (Solly, 1995), which is perceptive as well as amusing). But, if used at all by social scientists, the idea of national character needs to be handled with great care and sophistication.

That is not to deny, however, that the concept has a long and – at its best – distinguished history. One can, for example, point to Alexis de Tocqueville's classic work *Democracy in America* (1961[1835–40]), which is a pioneering study of the complex connections between American institutions, culture, and *moeurs*. Moreover, it is a *comparative* study of the *mentalités* of what he conceptualizes perhaps too dichotomously as 'democratic' society in America and 'aristocratic' society in France. Implicitly, indeed, it is a three-way comparison with Britain too, as becomes clear in his other major works, *The Old Regime and the Revolution* (1955), *Recollections* (1970), and *Journeys to England and Ireland* (1958).

The great era for social scientific studies of supposed 'national character' was the Second World War and its aftermath. This was war work for some distinguished social scientists, the aims of whose research included understanding the character and culture of the enemy during the war, and helping to guide the policy of the victorious powers in their occupation of the defeated countries. Thus Ruth Benedict wrote about the Japanese (1946), and Talcott Parsons (1954) about both the Germans and the Japanese. These concerns also lingered on into the early stages of the Cold War.

Shortly after the war, the Canadian social psychologist Otto Klineberg wrote an important survey and critique of the academic literature on 'Personality in relation to nationality' (1950). Dealing first with what he calls 'descriptive accounts' such as Tocqueville's, or such now less well-remembered studies as those by André Siegfried (1927) and Salvador de Madariaga (1928), Klineberg acknowledged that though they often contained interesting suggestions, "they do not carry conviction to the empirically-minded reader … the inference from the facts to the psychological conclusions is not always convincing" (1950: 11). (Of the many selections of Tocqueville's writings, that edited by John Stone and Stephen Mennell – *Alexis de Tocqueville on Democracy, Revolution and Society* (Chicago: University of Chicago Press, 1980) – emphasizes his discovery of several sociological principles that have later been developed more systematically by professional sociologists; but this does not contradict the point made by Klineberg.) They nevertheless contained useful material; for example differences between the accounts of America written by British and French authors might function a little like a projective test, telling as much about aspects of British or French character as about the Americans.

Klineberg took a similar view of the studies written by members of the 'culture and personality' school of anthropology, which include – besides Ruth Benedict's work (1946) on the Japanese – Margaret Mead's comparisons of the Americans (1942), and the Samoans (1927) and Geoffrey Gorer's books on the Americans, the British, and the Russians (1948, 1949, 1955). In particular, Klineberg made the point that such studies in general "give us no indication of the variations among individuals who constitute the national group"; moreover, there are "great differences in the degree and kind of information obtainable about a relatively simple community of 200–300 persons and a complex modern nation" (Klineberg, 1950: 19, 21). The culture and personality school overlapped to a considerable extent with psychoanalytic approaches to the question of national character. Gorer, for instance, paid a good deal of attention to the toilet training of infants (and, in Russia, the practice of swaddling them). Klineberg's criticisms applied to them too; yet he acknowledged valuable insights arising from the careful studies made by Abram Kardiner (1939), who worked in collaboration with a number of anthropologists, from Erich Fromm's study (1942) of how the readiness of Germans to accept Nazism can be understood, and also from the then as yet unpublished work of H.V. Dicks (1972) based on interviews with SS killers.

Turning to vital and social statistics, Klineberg (1950: 25–36) noted their abundance but also the difficulties posed by greatly varying definitions between countries of, for example, mental illness. Even so, he acknowledged that the study of what was a 'typical' mental illness in each country might cast light on its cultures, as might the *absence* of certain known abnormalities. On the other hand, psychiatric interpretations appeared to be of little value (1950: 36–40). As Klineberg observed, there was no evidence at all of the validity of studies of the 'paranoia' supposedly characteristic of German culture: not even the Nazi leaders tried at Nuremberg showed distinctly paranoid characteristics (Kelley, 1947).

Overall, Klineberg observed that there was little consensus among social scientists on the concept of 'national character' or on methods for its investigation, and commented on 'the great complexity' of the problems involved. "A combination of historical study, ethnological analysis, intensive community studies, interviews with a representative sample, and the application of psychological techniques to individuals would seem to be the most promising approach at the present time" (1950: 213, 214).

Whether or not Klineberg's critical survey and noncommittal conclusions had anything to do with it, the concept of 'national character' came to be less and less employed by social scientists in the 1950s and afterward. And yet it does answer to an element in everyone's everyday experience. As Norbert Elias puts it,

> In every country the forms of perception and behaviour, in their full breadth and depth, have a pronounced national tinge. Often one only becomes aware of this in one's dealings with foreigners. In interactions with one's compatriots, *individual* differences usually impinge so strongly on consciousness that the *common* national coloration, what distinguishes them from individuals of other nations, is often overlooked. First of all, one often expects that people everywhere will react to the same situations in the same way as people of one's own nation. When one finds oneself in a situation in which one is compelled to observe that members of different nations often react in a quite different way to what one is accustomed to at home, one mentally attributes this to their 'national character.' As such, it often remains inexplicable. One simply accepts it.
>
> Elias, 2008[1960]: 230, 231

But how is such an everyday notion to be conceptualized for social scientific purposes? If it demonstrated nothing else, Klineberg's survey demonstrated that the problem had not been solved by the mid-twentieth century. One obstacle was the then-prevalent preoccupation of social theorists – notably Talcott Parsons and the numerous scholars whom he influenced in the various social sciences – with drawing rather ridiculous distinctions between personality, social, and cultural systems. Klineberg recognized, but struggled with, the obvious fact that whatever 'national character' might be, it is at once psychological, social, and cultural. The greatest danger, which even he did not entirely avoid, is taking an excessively psychologistic view of national character. That is to say, citizens of a country were often seen as having certain 'personality traits' in common – often in a quasiclinical sense – which were then seen as an independent variable 'causing' cultural and social consequences, such as aggressive militarism and 'tensions affecting international understanding.' In the years after their defeat in the Second World War, the Germans were especially the targets of this kind of approach, as social scientists internationally – and in Germany itself – struggled to understand the rise and fall of the Nazi regime. There was much discussion of what were taken to be the peculiar characteristics of German society, its history, and the psychology of its people. In the most extreme case, a constant 'national character' was traced back as far as Luther (McGovern, 1946). The risk is that 'national character' comes to be treated almost as a prime mover of events, rather than as something that is itself shaped and modified by events, power struggles, and social processes over time. It is better to think about it in terms of *processes* of habitus formation or conscience formation over time. 'Habitus,' a concept that is now especially associated with the work of Pierre Bourdieu but was in fact used earlier by many French and German social scientists, can be understood – as Norbert Elias puts it – simply as 'second nature': it comprises everything that individuals have in fact learned from other people from birth onward but which has become so deeply habituated that it seems even to themselves as something 'natural.' Seeing it in this way makes clear that the attempt to distinguish, even 'analytically,' the levels of personality, social, and cultural characteristics is futile. Many years ago, the anthropologist Clyde Kluckhohn and the psychologist Henry Murray made the point very well:

Every person is in certain respects
1. like *all* other people,
2. like *some* other people, and
3. like *no other* person (1948: 35).

That is to say, there are some characteristics that are common to every human being: every normal person has, for example, the capacity to learn a language. There are usually also ways in which every individual has some distinctive or unique characteristics or quirks of his or her own. But the study of habitus and its formation is concerned with level (2), the level of characteristics that are learned, shared, and transmitted among members of a specific group of people. 'Learned, shared, and transmitted' is one way in which social scientists have defined 'culture,' but habitus is at once cultural, social, and psychological.

It is important to note, however, that processes of habitus formation in this sense are by no means confined to *national* groups and '*national* character.' People may learn, share, and transmit characteristics in any group with a continuing form with which they may identify and be identified – from families to nations and perhaps now to global formations. In complex societies, there are many levels to habitus. It has been hypothesized (Mennell, 1994) that the strongest feeling of identity, and the most general influence on habitus formation, tends to be with what Elias (2010: 182–189, 195–207) called a person's 'survival unit.' By that he meant the largest group within which individuals look for meeting their most important needs – food, shelter, protection, procreation, and orientation. Survival units have tended to become larger in the course of the development of human society, although they can also shrink – for example in times of danger, conflict, and war. In the late-twentieth and early twenty-first centuries, there are have been a number of

instances of supposedly sovereign and unitary nation-states fragmenting into smaller territorial units, often through civil war. Nevertheless, the nation-state has been prominent in the organization of humankind, and the most obvious survival unit with which their members typically identify, at least in the nineteenth and twentieth centuries. That is one reason why, for a time, social scientists tended to focus special attention on *national* habitus.

This strand has most recently been especially prominent in the writings of Norbert Elias. As he remarked in his *magnum opus*, *On the Process of Civilisation*,

> In the conduct of workers in England, for example, one can still see traces of the manners of the landed noblemen and gentry and of merchants within a great trade network, and in France the airs of courtiers and a bourgeoisie brought to power by revolution. In the workers too, we find a stricter regulation of conduct, a type of courtesy more informed by tradition in colonial powers which have for a long period had the function of an upper class within a large network of interdependences.
>
> Elias, 2012[1939]: 426

Elias always stressed the need for *processual* thinking about the intertwining of what he called socio- and psychogenesis. Habitus must be seen as itself the *product* of long-term characteristics of national societies, formed through conflicts and power struggles, both *within* and *between* such societies. In the series of studies published at the end of his life as *Studies on the Germans* (2013), Elias extended this comparison not just to the Germans but also the Dutch:

> The paradoxical situation of having an upper class composed of urban merchants has left deep marks on the habitus of the Dutch. Fostering equality is of prime importance. … In the course of their national development, the Dutch city merchants provided an exemplary case for the solution of the problem of how civilians can protect themselves from violent external attack without being dominated by their own military helpers.
>
> 2013: 15–17

In contrast, the history of Germany produced the opposite result. In the politically fragmented German-speaking lands, the *Bildungsbürgertum* had long aspired to national unification through peaceful means, but after the failure of the revolution of 1848 unification was subsequently achieved by the army on the battlefield – resulting in greater prestige for and deference toward the military caste. Later, Elias emphasized, national defeat in the First World War brought about a profound sense of national humiliation. He made an important connection between political history and events on the one hand and culture and habitus on the other:

> The central question is how the fortunes of a nation over the centuries become sedimented into the habitus of its individual members. Sociologists face a task here that distantly recalls the task that Freud tackled. He attempted to show the connection between the outcome of the conflict-ridden channelling of drives in a person's development and his or her resulting habitus. But there are also analogous connections between a people's long-term fortunes and experiences and their social habitus at any subsequent time. At this layer of the personality structure – let us for the time being call it the 'we-layer' – there are often complex symptoms of disturbance at work which are scarcely less in strength and in capacity to cause suffering than the individual neuroses.
>
> 2013: 24

Others have taken up Elias's ideas on this score. Kuzmics and Axtmann (2007) examined the connections between 'authority, state, and national character' in Austria and England, making especially rich use of literature as a source of sociological evidence. Their book covers the period 1700–1900, and that serves as a reminder that habitus changes: what they depict as characteristic of the English in 1900 now appears almost unrecognizable from the vantage point of the early twenty-first century. "Britain in the recent past," commented Elias (2013: 6) "is a moving example of the difficulties a great power of the first rank has had in adjusting to its sinking to being a second or third-class power." Mennell (2007), writing about the United States, avoided the term 'national character,' and conceptualized the problem as a process of habitus formation. But the strategy was similar: to show the connection between a country's history and the *moeurs* of its people. In particular, Mennell stressed how social and political conflicts internally had over the long term produced a patchy monopolization of the means of violence in the United States; and that, externally, the country's rise to overwhelming power in the world at large had also left its mark on Americans' perceptions of themselves and on their assumptions about and behavior toward non-Americans.

C. Wright Mills (1959: 4) famously defined the sociological imagination as the ability "to grasp history and biography and the relations between the two within society." The strand of thought in sociology, anthropology, and psychology traced in this article stands squarely within Mills's vision. But the important connection between history and biography can generally be explored without using the slightly hazardous concept of 'national character.'

See also: Anthropology, History of; Elias, Norbert (1897–1990); Parsons, Talcott (1902–79); State Formation, Theory of; Tocqueville, Alexis de (1805–1859).

Bibliography

Benedict, Ruth, 1946. The Chrysanthemum and the Sword. Houghton Mifflin, Boston.
Dicks, Henry V., 1972. Licensed Mass Murder: A Socio-psychological Study of Some SS Killers. Chatto & Windus, London.
Elias, Norbert, 2008[1960]. National peculiarities of British public opinion (E. Jephcott, Trans.). In: Essays II: On Civilising Processes, State Formation and National Identity. Collected Works, vol. 15. UCD Press, Dublin, pp. 230–255.
Elias, Norbert, 2010. The Society of Individuals (E. Jephcott, Trans.). In: Collected Works, vol. 10. UCD Press, Dublin.
Elias, Norbert, 2012[1939]. On the Process of Civilisation: Sociogenetic and Psychogenetic Investigations (E. Jephcott, Trans.). In: Collected Works, vol. 3. UCD Press, Dublin.
Elias, Norbert, 2013. Studies on the Germans (E. Dunning and S. Mennell, Trans.). In: Collected Works, vol. 11. UCD Press, Dublin.
Fromm, Erich, 1942. Fear of Freedom. Routledge & Kegan Paul, London.

Gorer, Geoffrey, 1948. The Americans. Cresset Press, London.
Gorer, Geoffrey, 1955. Exploring English Character. Cresset Press, London.
Gorer, Geoffrey, Rickman, John, 1949. People of Great Russia: A Psychological Study. Cresset Press, London.
Kardiner, Abram, 1939. The Individual and His Society. Columbia University Press, New York.
Kelley, Douglas, 1947. Nuremberg Diary. Farrar, Strauss, New York.
Kuzmics, Helmut, Axtmann, Roland, 2007. Authority, State and National Character: The Civilizing Process in Austria and England, 1700–1900 (B. Allen and A. Skinner, Trans.). Ashgate, Aldershot.
Klineberg, Otto, 1950. Tensions Affecting International Understanding: A Survey of Research. Social Science Research Council and UNESCO, New York.
Kluckhohn, Clyde, Murray, Henry A., 1948. Personality in Nature, Society and Culture. A.A. Knopf, New York.
Macdonell, A.G., 1933. England, Their England. Macmillan, London.
Madariaga, Salvador de, 1928. Englishmen, Frenchmen, Spaniards. Oxford University Press, London.
McGovern, William Montgomery, 1946. From Luther to Hitler: The History of Nazi-Fascist Philosophy. Harrap, London.
Mead, Margaret, 1927. Coming of Age in Samoa. Jonathan Cape, London.
Mead, Margaret, 1942. And Keep Your Powder Dry. William Morrow, New York.
Mead, Margaret, 1955. Russian Attitudes towards Authority. Tavistock, London.
Mennell, Stephen, 1994. The formation of we-images: a process theory. In: Calhoun, Craig (Ed.), Social Theory and the Politics of Identity. Blackwell, Oxford, pp. 175–197.
Mennell, Stephen, 2007. The American Civilizing Process. Polity, Cambridge.
Mikes, George, 1946. How to Be an Alien. André Deutsch, London.
Mikes, George, 1952. How to Be a Yank. Wingate, London.
Mills, C. Wright, 1959. The Sociological Imagination. Oxford University Press, New York.
Parsons, Talcott, 1954. 'Democracy and social structure in pre-Nazi Germany' [1942], and 'Population and the social structure of Japan' [1946]. In: Essays in Sociological Theory. Free Press, New York, pp. 104–123, 275–297.
Siegfried, André, 1927. America Comes of Age: A French Analysis. Harcourt, Brace, New York.
Solly, Martin, 1995. Xenophobe's Guide to the Italians. Xenophobes Guides, London.
Stone, John, Mennell, Stephen (Eds.), 1980. Alexis de Tocqueville on Democracy, Revolution and Society. University of Chicago Press, Chicago.
Tocqueville, Alexis de, 1955. The Old Regime and the Revolution (S. Gilbert, Trans.). Doubleday, Garden City, NY.
Tocqueville, Alexis de, 1958. Journeys to England and Ireland (G. Lawrence and K.P. Mayer, Trans.). Yale University Press, New Haven.
Tocqueville, Alexis de, 1961[1835–40] (H. Reeve, Trans.). Democracy in America, 2 vols. Schocken, New York.
Tocqueville, Alexis de, 1970. Recollections (G. Lawrence, Trans.). Doubleday, Garden City, NY.

National Liberation Movements

S Amin, Third World Forum, Dakar, Senegal

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 15, pp. 10309–10314, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

National liberation movements (NLMs) constitute diverse socio political movements which share the aim of establishing an independent state for what they consider their nation within the borders of a territory recognized by the international community. These movements prioritize the liberation of their nation over the many other possible social goals, and therefore demand that their people to unite around that priority. NLMs have indeed been among the major political forces which could be found in Europe in the nineteenth century and later, in the twentieth century, expanded to Asia and Africa. The case of African liberation movements are especially considered here. The NLMs can only be understood by reference to the establishment of global capitalism, and despite links to controversies over religion and national identity, economic forces have always dominated such movements.

National liberation movements (NLMs) along with the labor movements have represented major antisystemic forces compelling capital to adjust to the social interests they represented. The expansion of capitalism on a global scale has always generated unequal development and therefore continuously produced NLMs which pursued the target of catching up with the advanced centers of the global system. Such movements expanded in the Americas in response to the mercantilist stage of capitalism, in Central, Eastern, and Southern Europe in the nineteenth century, and in Asia and Africa in the twentieth century. While the circumstances between 1945 and 1980 encouraged such projects, that page of history has now been turned.

NLMs constitute a large spectrum of socio political movements which share in common the target of struggling for establishing an independent state for what they consider their nation within the borders of a territory recognized by the international community. These movements also share that they privilege the struggle for the liberation of their nation over the many other targets of other social movements, and therefore call their people to unite around that priority. The very existence of such movements assumes that the nation either must be united with a view of constituting a single state, or must be separated from a plurinational state in which it is included and that is in their view dominated by another nation, or has to emancipate from a colonial status and achieve political independence of that territory. NLMs have indeed been among the major political forces which operated mainly in Europe in the nineteenth century and later, in the twentieth century, expanded to Asia and Africa. They have shaped major changes in the political geography of the modern world and possibly will continue to do so in some areas.

There are, of course, different approaches to the study of this set of important social phenomena. Some scholars take for granted the preexistence of the nation which these movements mobilize, and therefore focus on the analyses which help in understanding the reasons for their success (or failure), their strategies, tactics, and ability to mobilize objective realities operating within the society or in the field of international politics which may support their project.

Other scholars focus rather on the relationship between the deployment of NLMs and the major changes which singularize the process of modernization of the nineteenth and twentieth centuries. Along with Eric Hobsbawm, Immanuel Wallerstein, and the Austro-marxists, the author will adopt that specific approach for the analysis of the NLMs. Commodification of production, industrialization, urbanization, expansion of education – and all these dimensions of the process of modernization are deeply related to that of capital accumulation – played, in that respect, a decisive role in the crystallization of a new type of community, usually called nation. The new economy and social life implied an intensified use of written communications, and therefore gave to the language used for education and in the workplace a role far greater than it had been the case in premodern societies.

Centers and Peripheries in the Modern World System

Nations as well as social classes are products of modern times, even if one can discover ancestors to them in the past, which anyway have been radically reshaped by the dominant rationale of capital accumulation operating simultaneously at the various local ('national') and global levels. In that respect, the history of capitalism is that of a continuous conflict between the logic of capital accumulation and that through which various other social interests express themselves. NLMs have their place in that historical frame, just as the other social movements, particularly those of the working classes. In fact, these two families of movements – labor and national liberation – have represented, throughout at least the nineteenth and twentieth centuries the major antisystemic forces, compelling capital accumulation to adjust to the social interests that they meant to represent.

National liberation refers to 'nation,' and that term presupposes certain articulations between this reality, real or alleged, and other realities such as the state, the world system of states, the local economy as related to the world economy, and social classes. We owe these concepts and their articulation into a system to the different social theories developed in the light

of the nineteenth century European historical experience. Within this framework, the elaboration of two sets of theories took place – as it turned out, in counterpoint to one another: on the one hand, socialism and the theory of class struggle; on the other hand, nationalism and the theory of class integration into the nation-state. Both theories take account of many aspects of the immediate reality which is marked both by social struggles ending in deep transformations of capitalism and even in revolutions, and by struggles between nation-states ending sometimes in war as well as NLMs ending in many cases in the creation of new independent states. For protagonists of these theories, they have proven to be potent guides to action.

The expansion of capitalism on a world scale has always been, and continues to be, based on unequal economic and, therefore also, political relations between those areas which constituted themselves as full-fledged centers of the system controlling the global process of accumulation, and those areas which have been submitted to the logic of that accumulation, i.e., the peripheries. Therefore, the relationship between state and nation on one the hand and the economic system on the other is not similar in the centers and in the peripheries of global modern capitalism. In the centers there is a rough coincidence between the state (whether national or multinational) and an autocentric productive system; in the peripheries such a coincidence is always lacking. That coincidence between the state, the nation, the high level of capitalist transformation, and the dominant position of such states in the world capitalist system which led these central partners to form 'autocentered' interdependent economic units enjoying a high degree of autonomy vis-à-vis each other, defines a possible field for 'national' economic policy. The instruments of this policy – the national centralized monetary system, customs, the network of material infrastructures in transport and communications, the unifying effects of a 'national' language, the unified administrative system, and so on – enjoy a definite autonomy in relation to the constraints imposed by an economy operating worldwide. Relations between classes, however wrought, are relegated to and by the national state.

True, there is no Wall of China to separate these national systems from the world system that they constitute. Internal social relationships are partly dependent upon positions occupied by the national states, in the world hierarchy. If social relations permit it, these states can improve their position by pursuing coherent national policies. This effectiveness in turn facilitates social compromise and, without 'abolishing the class struggle,' puts definite limits to conflicts.

However, at closer inspection it appears that these circumstances, pervasive but limited in time to the nineteenth century, are even more limited in space. Around a few 'model' nation-states, the world of the capitalist system – structured by different pasts which in turn lost their legitimacy and efficacy – remains undefined in the light of an uncertain and obscure future. Therefore, the problem changes when we quit the limited framework imposed by the central bourgeois nation-states. For this forces us to examine more closely 'regions' whether they are organized into the states or not. Regions are peripheral in relation to continuously expanding capitalist reproduction. On this level there is only a central state, i.e., a state which masters external relations and submits them to the logic of autocentered accumulation. On different levels, there are only 'countries,' which are administered from outside as colonies or semicolonies; these appear to be independent but are incapable not only of molding the outside according to their needs, but also of avoiding their drift and shaping from outside.

The consequence of that fundamental and continuous contrast between centers and peripheries immanent to the history of really existing global capitalist expansion is that 'national liberation movements' have been continuously produced by the uneven development associated with that expansion. These movements have always pursued the same major target, that of 'catching up' with the advanced centers, through constructing in their turn a modernized independent autocentered nation-state able to act as an autonomous participant in the global capitalist system. They always have nurtured a basic ideology of modernization and national building. In their struggle against the dominant forces in the global system, NLMs have also often attempted to build 'international common fronts' associating the movements involved. Mutual sympathy and common strategic and tactic approaches helped in constructing such common fronts. These common fronts often survived the victory of the NLMs, i.e., after those movements had conquered political independence. Yet these fronts were also gradually eroded by the very success of the NLMs. The new states pursued their own policy targets within the global system, which led to their insisting on their particular interests and specific strategic and tactical policies aiming at helping them to achieve them. The similarity with the other major antisystemic movement – that of the working classes and socialism which have built their own successive 'internationals' – is striking.

Stages of Deployment of the NLMs

Capitalist global expansion has moved through successive stages. For each of those stages specific forms of the relationship between centers and peripheries can be identified. The NLMs produced in each stage have therefore been confronted with specific challenges.

During the mercantilist stage of capitalist expansion (1500–1800) the American continent was conquered and indigenous populations subjugated when not exterminated, Africans were imported as slaves, while migrants from Europe either created autonomous small production systems (in New England, for instance) or constituted themselves as the local ruling class operating as a transmission belt for the metropoles, dominant powers (slave owners and masters of encomiendas). The NLMs in the late eighteenth century (the 'American revolution') and in the early 1800s (in Latin America) were led by those local ruling classes of European origin, disregarding the interests of subjugated peoples (black slaves and indigenous Indians). Haiti is in this respect an exception. The common front which was constituted with a view to excluding European powers from the continent was quickly captured by the new capitalist center which emerged in the USA (that is the meaning of the Monroe declaration in 1823). The pattern of an uneven relation was gradually constructed in that frame between the US center and its Latin American periphery. It was only much later,

in the twentieth century, that nationalist bourgeois movements started questioning that uneven relation, along in some cases with a revival of indigenous Indian peoples movements. The Neozapatistas (Chiapas) movement in Mexico is the most recent expression of that unfinished struggle for national liberation and social progressive change.

A second wave of NLMs developed throughout the nineteenth century in Central, Eastern, and Southern Europe, associated with the expansion of capitalism beyond its northwestern European corner of origin. In two cases, Germany and Italy, these movements led to the constitution of new nations, which became new capitalist centers. Further to the east the three multiethnic autocratic empires (the Russian, the Austro-Hungarian, and the Ottoman) were assaulted by the expansion of capitalism, moved by a conjunction of internal forces (produced by industrialization and the formation of modern social classes – a bourgeoisie and a working class) and the impact of external forces dominating the global system. The dilemma with which those societies were confronted was the following: could capitalist development flourish within the frame of such large states, becoming multinational and democratic, or should they be dismantled on a more or less national base? In these countries more than elsewhere the various class struggles (between the autocratic leading groups of the old regime and the new bourgeoisie, between the latter and the working class, between the large landowners – often of different ethnic origin – and their peasants), the political struggles (for democracy, for socialism), and the 'national question' overlapped and interacted in a complex manner. The major organizations within the socialist movement were rather in favor of maintaining the large multinational state (so was the meaning of the struggle of the Austro-marxists for cultural autonomy). However, the peasant and bourgeois movements favored 'national liberation first,' and that option was finally the one which succeeded in imposing itself.

Simultaneously, during the second half of the nineteenth century, there was an extraordinary expansion of global capitalism which conquered Asia and Africa and colonial (or at least semicolonial) systems were established there in most cases. The national question was moving out of Europe and NLMs of a new type started being constituted, inspired by the local elites and fractions, at least of the new bourgeoisies. The ancestors of those NLMs which played a decisive role in the making of the twentieth century could be identified at least, in China, India, Persia, Egypt, and Syria, where it animated cultural revivals (such as the Arab Nahda) and the first political modern organizations. These new NLMs related to class issues in a variety of ways, but in general they gradually became what has been labeled later as 'antifeudal–anti-compradore' political movements, associating in their struggles for reconquerring the lost national independence and modernization of the 'external enemy' ('imperialism') and its local allies (precisely the 'feudals' and 'compradores'). Yet the dominant force among the international socialist movement – the social democracy of the Second International – ignored those new NLMs, overlooked their importance for the future and shared the dominant European imperialist ideology of the 'civilization mission' of colonialism.

The colonization of Asia and Africa has nationally produced a wave of most important NLMs which finally won their struggle for national independence after World War II and played an important role in the changing of the face of the world since. The NLMs of Africa and Asia were very different in outlook. The past and recent history of their peoples, the social forces, the political and ideological parties and trends which they represented, their visions of the future of the societies to be constructed or to be reconstructed, and its relationships with the West were all subjects on which they differed. However, what brought them together was their common project, which included the completion of the political decolonization of Asia and Africa. Beyond the Afro-Asian States which emerged out of the victories of NLMs after World War II all considered that the political independence they had regained was only the means, the end being the achievement of economic, social, and political liberation. In the opinion of by far the majority 'development' would be possible through interdependence within the world economy. The NLMs and then after the new states they created never intended to 'leave the system,' or 'delink' even if they did not all share the same strategic and tactical vision of 'development.' But, to varying degrees, they considered that the building of an independent, developed economy and society (even within global interdependence) implied a certain amount of 'conflict' with the West in its dominance (the radical wing thought it necessary to put an end to the control of national economies by the capital of foreign monopolies). Moreover, anxious to safeguard the independence which they had regained, they refused to take part in the military struggle at planet level, and to serve as bases for the surrounding of the socialist countries which American hegemony was attempting to impose. However, they also thought that if they refused to be part of the Atlantic military alliance, this did not imply that it would be necessary for them to seek the protection of the latter's adversary, namely the USSR, whence 'neutralism' and 'non-alignment.'

Wilson's 'principle of nationalities' (a state for each nation), which was restricted to Europe after World War I, was thereafter partly recognized to Asian and African peoples by the UN after World War II. Yet it took several more decades for the NLMs to achieve their target. That success was favored to a certain extent by the bipolarity which characterized the international system in those times.

The coming together of the Afro-Asian states had already begun by the setting up of the Arab–Asian group within the United Nations in the late 1940s with a view to defending the independence of the colonies in arms. In 1955, the foremost Heads of State of the countries in Asia and Africa which had regained their political independence met for the first time at Bandung. Nasser, Nehru, Sukarno, and Chou En Lai were to have a considerable influence on their times, even although the experience of the new powers they represented, and to a large extent inspired, was very recent. India and Indonesia had been independent for barely 10 years, communist China for not quite 5, and in Egypt the monarchy had disappeared from the scene only 3 years previously. The struggle for the achievement of the historical task of independence was not over: the first war in Vietnam was barely over when the second one loomed on the horizon, the Korean war had just ended in a status quo, the Algerian war was at its height, the decolonization of Africa south of the Sahara had not even been contemplated, the Palestinian drama was still in its first stages, South Africa was

closing itself in hard apartheid. Bandung had the effect of reinforcing this rapprochement and of giving an impetus to this struggle. Three years later, in independent Accra, Kwame Nkrumah declared that 'Africa must unite.' But it was only to do so once independence had been achieved and after the failure of Nkrumah's pan-Africanism, followed by the demonstration of the impotence of the two groups formed as a result of the Congo crisis (the Casablanca group and the Monrovia group between 1960 and 1963), in the minimal form of the setting up of the OAU in 1963.

From summit to summit, throughout the 1960s and 1970s, 'nonalignment' was nevertheless gradually to change direction from a position of a front of political solidarity based on the support for NLMs and struggles and the refusal of military pacts, to a position of building a common front with a view to reinforcing collective economic demands from the North. It inspired the 'decades of development' in the United Nations system, the creation of UNCTAD in 1965, and a whole series of global negotiations on crucial economic global issues. The struggle for a 'new international economic order' which began in 1975 after the October 1973 war and the revision of oil prices was the culmination of this development and the beginning of its end.

The West had no intentions of accepting the spirit of Bandung as a foregone conclusion on either the political level or on that of the economic struggle. The project of society which constituted a kind of basic charter for the NLMs and thereafter inspired the policies of the Asian and African states did exist, although it was often implicit and imprecise. This project, setting aside its various concrete manifestations and the specificities of its expression at national level, can be defined in terms of the following elements: (1) the desire to develop the forces of production and to diversify production (in particular to industrialize); (2) the desire to ensure that the nation-state controls and directs the process; (3) the belief that 'technical' models constitute 'neutral' data which, once mastered, only have to be reproduced; (4) the belief that the initiative did not imply its being initiated by the people in the first instance but only their support for state actions; and (5) the belief that the process is not fundamentally in contradiction with the participation in exchange within the world capitalist system, even if it implied conflicts at times with the latter.

The circumstances of capitalist expansion between 1955 and 1970 to some extent encouraged the crystallization of this project. The implementation of this national bourgeois project implies that the hegemonic national, bourgeois class, by means of its state, controls a series of processes which include as a minimum:

1. the management of the reproduction of the labor force, which implies relatively full and balanced development so that local agriculture, for example, is capable of providing the elements required for this reproduction in sufficient quantity and at suitable prices to ensure the valorization of capital;
2. the control of national natural resources;
3. the control of local markets and the capacity of entering the world market on competitive conditions;
4. the control of the financial circuits enabling the centralization of surpluses and the directing of its productive usages;
5. the control of the technologies being used at the level of the forces of production concerned.

Seen from this angle, the experiences of the Third World inspired by the ideology of NLMs can be divided into two categories: that of the countries which have attempted nothing more than the acceleration of growth without considering the implementation of the conditions outlined above (Ivory Coast, Kenya, Pakistan, Saudi Arabia, etc.) and that of the countries which have attempted to implement the conditions outlined (Nasser's Egypt, Algeria, Tanzania, India, Brazil, South Korea, etc.). It is obvious that the classification does not necessarily distinguish between regimes concerned with the implementation of a modicum of social justice and reforms, in particular agrarian reforms (e.g., Nasser's Egypt and South Korea) and those which have had no hesitation in accepting the intensification of social inequality (e.g., Brazil). Nor does it necessarily distinguish between the attitudes toward transnational capital (Brazil and Kenya both encourage the latter, but the former attempts to integrate it into its own national policy, whereas the latter simply adjusts to its demands), nor even between the question of the political relationship of conflict or alliance with the East and the West. It is true that there are correlations but the variations in composition as a function of the concrete situation make a special case of each of the Third World countries. However, over and above the variety of experiences, the fact remains that the most coherent implementations have taken place when there has been the conjunction of an intense nationalist struggle and a powerful social revolution. Led by communist parties, these NLMs in China, Vietnam, and Cuba are the most distinct examples of such a radicalization.

It is no longer possible to disregard the inadequacies of all these attempts which collapsed when circumstances became favorable to their overthrow or erosion. The food and agricultural crisis, the external financial debt, the increased technological crisis, the external financial debt, the increased technological dependence, the fragility of the capacity to resist possible military aggressions, and the advance of waste on the lines of the capitalist consumption models and of their ideological and cultural content point to the historical limits of the attempts. Even before the present crisis presented the possibility of an 'offensive from the West' which would succeed in reversing the trends, these inadequacies had already led to an impasse.

In this sense, we may consider that the NLM project under consideration deserves to be described as a national, bourgeois project. The project for the society that the NLMs had in mind was, as we have shown above, that of creating a new state integrated in the global capitalist system. In that perspective the ruling class imagined itself becoming similar to that which controls the process of accumulation in the capitalist centers, and therefore it deserves to be qualified as 'bourgeoisie' (including state bourgeoisie forms), although of course it constituted itself out of different historical origins. Moreover, it has proved to be impossible. In this sense, history has proved that the national bourgeoisie is incapable, in our times, of implementing what it achieved elsewhere, in Europe, in North America, and in Japan in the nineteenth century. A concrete examination of the history of NLM ideologies and the policies they inspired in the Third

World independent countries would illustrate that fact: that of the uninterrupted sequence of attempts by the national bourgeoisie, of their repeated failures and of the submission to the demands of subordination which followed on each occasion; in Latin America since the nineteenth century (we shall only refer here to the more recent examples of the Mexican revolution between 1910 and 1920 and of Peronism in Argentina), in India (e.g., the change in direction since Nehru's 'first plan' until the return to power of the right after the first failure of the Congress), as in numerous Arab and African countries.

The favorable economic circumstances at the end of World War II were indeed exceptional and could not last. On the economic level, strong growth in the 'North' facilitated 'adjustment' in the South. On the political level, peaceful coexistence along with the rise of Soviet power, and its combination with the decline of the old French and British colonialisms and the rise of the Afro-Asian struggles for independence made the Soviet alliance truly efficient.

New Challenges, Problems, and Perspectives

The page of that history of NLMs has now been turned and the global capitalist system has entered in a new stage of its expansion. In this respect it is often said that any attempt at a national development has become irrational and impossible, that the challenges are basically global and therefore cannot find their solutions; but at that level, in other words that there is no more national capital, national bourgeoisies, but only a globalized capital managed by a globalized bourgeoisie. Yet the system is not moving toward reducing global polarization, but is rather deepening it. Uneven development between centers and peripheries and therefore nations and peoples not only continues but aggravates, albeit in new conditions. The global capital and bourgeoisie – if existing – would be restricted to the central triad, whereas in the peripheries they are bound to be compradore transmission belts of domination of the first. Therefore, the national question remains on the agenda; it continues to be a real challenge. What can be said in that respect is that the change marks the end of an epoch which in the case of Asia, Africa, and Latin America can be called the century of the national bourgeoisie, in the sense that it has precisively been marked by successive national bourgeoisie edification attempts. The Third World bourgeoisie now finally sees its own development in terms of the compradore subordination imposed upon it by the expansion of transnational capitalism. A new wave of NLMs should therefore be expected, different in their social targets and methods from the previous ones. The crystallization of the alternative new national popular (not bourgeois) strategy involves a merging of three conditions. First, delinking in the sense of a strict submission of external relations in all areas to the logic of internal choices taken without consideration of criteria relating to world capitalist rationality. Second, a political capacity to operate social progressive reforms through advances of democracy. This political capacity is a condition of delinking – since existing hegemonic classes have no interest in it. Delinking without that transfer of political hegemony has little chance of emerging, and if it did emerge under such conditions, it would lead to an impasse. Third, a capacity for technological absorption and invention, without which the autonomy of decision making acquired could not be realized.

Thus defined, the conditions for a positive response to the challenge of history appear severe, and any merging of such conditions remote. That is not surprising. The response to new challenges never comes spontaneously. It takes time for the social structures to stabilize, for the new forms of social movements to produce efficient strategies, for them to formulate legitimate ideologies in keeping with the real challenges. In the meantime, what we have is rather chaos, and therefore disarray and nostalgias of the past, in other words, false responses to real problems. It is in this frame that we can observe revivals of movements which label themselves 'national liberation,' usually on the basis of ethnicity and religion, denying importance to any other dimension of social reality, such as class. Often presented by the dominant ambiguous ideologies as 'primordial' in the sense of being the expression of a transhistorical fundamental 'difference' and 'specificity,' these movements are in fact more of the nature of being a symptom of the crisis rather than an answer to it.

See also: Capitalism: Global; Environmental History; Imperialism, History of; Nationalism, Historical Aspects of: Africa; Nationalism, Historical Aspects of: Arab World; Nationalism, Historical Aspects of: East Asia; Nationalism, Historical Aspects of: South Asia; Nationalism: General; Social Movements, History of: General; Third World.

Bibliography

Albertini, M., 1978. L'État National. Fédérop, Paris.
Amin, S., 1976. Unequal Development. Monthly Review Press, New York (Chapter 5).
Amin, S., 1994a. L'Ethnie à L'Assaut des Nations. L'Harmattan, Paris.
Amin, S., 1994b. Re-reading the Post War Period. Monthly Review Press, New York (Chapter 5).
Amin, S., 1997. Capitalism in the Age of Globalization. Zed Books, London (Chapter 4).
Anderson, B., 1983. Imagined Communities. Verso, London.
Bauer, O., 1907. Die Nationalitten frage und die Sozialdemobratis Vienna.
Benot, Y., 1975. Les Independences Africaines Idéologie et Realites. Maspero, Paris.
Blunt, J.M., 1987. The National Question, second ed. Zed Books, London.
Bottomore, T., Goode, P., 1978. Austro Marxism. Clarendon Press, Oxford, UK.
Cohen, M., 1994. Ethniciti Politique. L'Harmattan, Paris.
Davis, H.B., 1967. Nationalism and Socialism. Monthly Review Press, New York.
Davis, H.B., 1978. Toward a Marxist Theory of Nationalism. Monthly Review Press, New York.
Fieldhouse, R., 2005. Anti-apartheid: A History of the Movement in Britain: A Study in Pressure Group Politics. Merlin, London.
Gibson, R., 1972. African Liberation Movements. Oxford University Press, Oxford, UK.
Haupt, G., Lowy, M., Weill, C. (Eds.), 1974. Les Marxistes et la Question Nationale 1848–1914: Études et Textes. F. Maspero, Paris.
Hobsbawm, E.J., 1990. Nations and Nationalism since 1780. Cambridge University Press, Cambridge, UK.
Jalata, A., 2004. State Crises, Globalization, and National Movements in North-east Africa. Routledge, London.
Lenin, V.I., 1970. Imperialism, the Highest Stage of Capitalism. Progress Publishers, Moscow.

Mullings, L., 2009. New Social Movements in the African Diaspora: Challenging Global Apartheid. Palgrave Macmillan, Basingstoke.
Maharaj, M., Mandela, N., 2002. Reflections in Prison: Voices from the South African Liberation Struggle. University of Massachusetts Press, Amherst.
Porter, D., Boulouque, S., 2011. Eyes to the South: French Anarchists and Algeria. A K Press, Oakland, CA.
Riazanov, 1923. Le Komintern et l'Orient.
Strasser, J., Pannekock, A., 1974. Nation et Lutte de Classe. Union Generale d'Editions, Paris.
Snow, D.A., 2013. The Wiley-Blackwell Encyclopedia of Social and Political Movements. Wiley-Blackwell, Chichester, West Sussex.
Torreguitar, E., 2009. National Liberation Movements in Office: Forging Democracy with African Adjectives in Namibia. Lang, Frankfurt am Main.
Wallerstein, I., 1974. The Modern World Systems, 3 vols. Academic Press, New York.
Wallerstein, I., 2000. The Essential Wallerstein. New Press, New York.

National Security and Domestic Terrorism: The US Case

Ashmini G Kerodal, City University of New York, New York, NY, USA
Joshua D Freilich, City University of New York, New York, NY, USA; and University of Maryland, College Park, MD, USA
Steven M Chermak, Michigan State University, East Lansing, MI, USA; and University of Maryland, College Park, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Domestic extremists, such as Al Qaeda and associated movements, far-right extremists, and far-left extremists, currently pose a threat to national security. These threats are dynamic and national security policies need to be sensitive to the overall criminal offending patterns of specific movements, characteristics of the specific movement and its supporters, and the unintended consequences of counterterrorism policies. This article provides an overview of the most active domestic extremist movements, reviews current national security policies, and provides suggestion to reduce the risks posed by domestic extremists.

This article provides an overview of the most active domestic extremist movements, reviews current national security policies, and provides suggestion to reduce the risks posed by domestic extremists. The FBI defines domestic terrorism as "domestic acts of violence in furtherance of political, religious, racial or social ideology" (Mueller, 2012). Thus, the perpetrator(s), victim(s), and incident must occur within one country for the act to be deemed domestic terrorism (Parkin et al., 2011). According to Hewitt (2003), more than 3000 terrorist incidents, that took more than 700 lives, were committed in the United States between 1954 and 2000. Most of these incidents were perpetrated by domestic extremists. In fact, LaFree et al. (2006) found that each year usually about 85% of terrorist attacks are domestic. In addition, Freilich and Chermak (2012) found that since 1990, domestic extremists committed more than 650 financial schemes, such as tax avoidance, money laundering, and material support, which resulted in more than $700 million in financial losses.

Smith (1994) reviewed federal prosecutions of terrorists between 1980 and 1989 and subdivided domestic terrorism into Puerto Rican, Jewish, Left wing, Right wing, and special interest or single-issue terrorism. At that time, law enforcement considered antiabortionists – termed 'single-issue terrorists' by Smith (1994) – and right-wing groups the greatest threat to national security (Carlson, 1995; Freilich et al., 2009b; Riley and Hoffman, 1995). Currently, right-wing and single-issue terrorism such as eco/animal rights extremists continue to pose a threat, but Puerto Rican and Jewish terrorism have waned (Parkin et al., 2011). Not surprisingly, Al Qaeda and associated movements (AQAM) is considered by many to be the single largest threat to national security. Freilich et al. (2009b) survey of state police agencies, found that Islamic extremists were ranked as the greatest threat to state and national security, but other domestic groups were also of grave concern. The US Extremist Crime Database (ECDB), which uses open source documents to track violent and financial crimes perpetrated by domestic extremists, found that AQAM supporters have caused the most fatalities (over 3000) since1990. Interestingly though, far-right perpetrators committed more ideologically motivated homicide attacks (150+) than the 30 events committed by AQAM supporters (Freilich et al., 2012).

Domestic extremists' threat to national security expands beyond the potential loss of lives and property damage from acts of terrorism. Since 1990 far-rightists have committed thousands of violent and nonviolent crimes (Freilich et al., 2009b). Domestic extremists engage in a range of preparatory and nonideological or 'routine' crimes (Chermak et al., 2010; Freilich et al., 2014b). Terrorist organizations require funds to finance terror attacks and to cover operational costs, such as living expenses for members who live on the compound, security, pamphlets, and other propaganda materials and recruitment (Chermak et al., 2011). Although some funding is obtained via legitimate business ventures and donations from supporters, terrorist organizations must often resort to illegitimate activities, such as drug smuggling, fraud, forgery, counterfeited goods, and money laundering (Chermak et al., 2011).

This article first describes the range of terrorist and criminal behavior typical of three types of domestic extremist groups: AQAM, far-right, and eco/animal rights extremists. Detailed descriptions of these ideologies may be found in Freilich et al. (2014b) and are available upon request. Next, we discuss current national security policies. The article concludes with policy recommendations that consider the overall criminal trajectory of domestic extremists and their relative risks to national security.

The Scope of the Problem

This section provides a general overview of the types of domestic extremists that currently pose a threat to national security. A brief description of the types of terrorist and criminal behavior commonly committed by each movement is provided. This section provides a framework to situate the discussion of current counterterrorism (CT) policy and their limitations.

Al Qaeda and Associated Movements

Supporters of AQAM include extremists affiliated with groups such as Al Qaeda, Al Fuqra, al-Gamaa al-Islamiya, Pakistani

Taliban, Hamas, and Al Shabaab (Freilich et al., 2012). The Center on Law and Security (2011), based at the New York University School of Law, investigates AQAM cases that are classified as terrorism-related cases by the federal government. The Center on Law and Security identified 578 federal AQAM cases since 11 September 2001 to 11 September 2011. Most of these defendants ($n = 241$) were American citizens. Frequent charges included terrorism, national security, fraud and false statements, immigration violent crimes, weapons possessions, racketeering, and theft and fraud (Center on Law and Security, 2011).

The ECDB has identified 30 acts of ideologically motivated homicide incidents (i.e., the perpetrator killed others), committed by supporters of AQAM since 1995, which claimed over 3000 victims (Freilich et al., 2012). Almost half of these incidents (48%) were committed by lone actors. Since 1990, supporters of AQAM have also committed 12 attempted homicide incidents and were involved in 66 foiled violent plots, which were prevented by law enforcement. About 17% of AQAM supporters who committed an act of homicide also had a history of prior arrests. The homicide incidents ($N = 30$), attempted homicide incidents ($N = 12$), and violent failed or foiled plots ($N = 66$) were committed by 210 individuals, as some incidents involved multiple perpetrators.

Although the period for researching the ECDB by Freilich et al. (2012) was longer (1990–2012 vs 2001–11) than that covered by the Center on Law and Security (2011), the latter identified more perpetrators than the ECDB (578 vs 446). This could be because some of the cases identified in by the Center on Law and Security would not meet the ECDB inclusion criteria (See Freilich et al., 2012 for a detailed description of the ECDB inclusion criteria). Also, since the Center on Law and Security analyses federal terrorism-related cases, they would have excluded financial and routine crimes prosecuted in state courts. However, many AQAM cases appear in both the ECDB and the Center on Law and Security databases.

The ECDB, in addition to the violent acts described above, has also identified an additional 127 financial or material support schemes committed by at least one AQAM supporter since 1990. The ECDB defines a financial scheme as an "illicit financial operation involving a set of activities [i.e., techniques] carried out by one or more perpetrators to obtain unlawful gain or other economic advantage through the use of deliberate deception" (Belli, 2011: p. 64). Significantly, 25% of these schemes were linked to violence. These financial and material support schemes were committed by over 446 perpetrators. Importantly, these schemes were often carried out with nonextremist collaborators and utilized the national and international banking system, that is, wire transfers, cash deposits, and checks (Freilich et al., 2013). Many of these schemes were preparatory crimes, i.e., perpetrated to raise funds for 'jihad' in the US or abroad. This highlights the need to track the criminal behavior of extremists, in addition to their terrorist behaviors.

Extremist, in fact, require funds for their terrorist activities and organizational expenses. Since 2007, the frequency of federal material support charges has been increasing, as compared to other terrorism charges. Material support charges encompass monetary, weapons, or service support provided by others or the perpetrator. In 2007, material support charges accounted for 10% of all federal AQAM cases, but this figure rose close to 90% in 2011 (Center on Law and Security, 2011). This suggests either a change in prosecutorial strategy (attempting to reduce terrorism but cutting off the source of funding for the attacks) or an increase in support or sympathy for AQAM among domestic extremist. This backlash effect is discussed in greater detail in the section on CT and countering violent extremism (CVE) below.

Far-Right

Members of the American far-right movement, who believe in conspiracy theories, are extremely distrustful of federal authority and officials, and fiercely protective of their personal liberties, such as the right to own guns (Freilich et al., 2014b). Since the far-right does not recognize the authority of the federal government, they consider many federal laws and regulations, such as taxes and vehicular licenses to be invalid. Racist, anti-Semitic, and homophobic beliefs are also prevalent among some far-rightists (Freilich et al., 2014b).

The ideology of the far-right may lead some members to engage in (1) acts of terrorism to inspire political, social, or religious change, or (2) criminal acts as a result of their extreme beliefs (Chermak et al., 2010). For example, far-rightists' racists, anti-Semitic, and homophobic beliefs may motivate members of the movement to commit hate crimes, while their antigovernment sentiments could motivate far-rightists to engage in tax avoidance schemes, react negatively or violently to routine law enforcement activities (e.g., a vehicular stop) or actively target government agencies and officials.

According to Hewitt (2003), between 1978 and 2000, more than 250 persons were killed from terrorist attacks by far-rightists. More recently, analysis of ECDB data found that since 1990 far-rightists have been responsible for the loss of over 600 lives, including victims of the Oklahoma City bombing (Freilich et al., 2014b). These lives were claimed in over violent 370 incidents, which includes acts of terrorism, hate crimes, attacks against abortion providers, inter- and intragroup conflicts, and nonideological motivated homicides (Freilich et al., 2014b). Importantly, over 150 of these homicide incidents, that took around 400 lives, were ideologically motivated. Many of the ideologically motivated homicides (35%) were committed by lone actors (Gruenewald et al., 2013). Close to 62% of lone wolf far-rightists and about 50% of nonloner far-rightists who committed a homicide had a history of prior arrest. Not all crimes perpetrated by the far-rightists are motivated by their extreme beliefs. About 20% were nonideological but linked to the movement (e.g., inter- and intragroup conflicts) and approximately 40% were motivated by personal reasons not linked to the movement (Freilich et al., 2014a,b).

Far-rightists also engage in nonviolent crimes. Since 1990, far-rightists have committed more than 540 financial schemes (Freilich et al., 2014b). Unlike the AQAM, far-rightists typically engage in tax-related financial schemes, such as filing false tax returns, refusing to file tax returns, filing false liens, making frivolous tax claims, or creating elaborate offshore tax shelters to hide income from the IRS (Belli, 2011). Although some far-rightists were motivated by greed or a combination of greed and ideology, many far-rightists

were motivated by their extreme antigovernment beliefs (Belli, 2011).

Although many criminal acts committed by far-rightists are committed by individuals who belong on a far-rightist group, there is evidence to suggest that criminal acts by loners or "self-radicalized extremists…with no ties to extremist organizations" (Gruenewald et al., 2013: footnote 1) have been increasing. Gruenewald et al. (2013) argued that since leaders of the far-right movement, such as Louis Beam, Alex Curtis and Tom Metzger, began to promote the concept of 'leaderless resistance' as a strategy to avoid detection by law enforcement, loner attacks might be increasing in prevalence, as evidenced by the decreased number of individuals per case indicted for federal terrorism offenses (see also Damphousse and Smith, 2004). Gruenewald and colleagues used data from the ECDB to compare loner attacks ($n = 47$) to homicides committed by multiple perpetrators ($n = 92$) during 1990–2010. They found that since 9/11, loner attacks were significantly more likely to occur than homicide events committed by nonloners. Loners tended to be older (mean = 35.3) than other far-right homicide perpetrators (mean = 27.2), and more socially isolated. Finally, Gruenewald et al. (2013) found that loners were significantly more likely than other far-right homicide perpetrators to suffer from a mental illness (40% vs 8%).

Environmental and Animal Rights Extremists

Environmental and animal rights extremists also pose an important national security threat (Carson et al., 2012; Freilich et al., 2011; Smith, 1994). A 2008 Department of Homeland Security report named ecoterrorists as the largest terrorist threat to national security. Similarly, the survey of state police agencies ranked environmental extremists as the second largest national threat and animal rights extremists as the third largest national threat (Freilich et al., 2009b). Animal and environmental rights extremists typically believe in biocentrism, which states that everything in nature has inherent value and humans do not hold a superior position in the ecosystem. They also believe that the ecosystem is in imminent danger; the government and corporations are responsible for this danger; and there is a dire need to protect the ecosystem (Carson et al., 2012; Freilich et al., 2011).

Carson et al. (2012: p. 297) also noted that animal and environmental rights activists subscribe to the philosophy of 'peaceful coexistence with nature' – and unlike extremists – generally prefer legal avenues to accomplish their goals. However, some animal and environmental rights extremists perpetrate acts of terrorism and crimes in their attempts to ensure biodiversity. Environmental rights extremist groups, such as the Earth Liberation front (ELF) and The Family are concerned with protecting the entire ecosystem, while animal rights extremist group such as the Animal Liberation front (ALF), are more narrowly focused on ensuring equality of animals (Carson et al., 2012).

Environmental and animal rights extremists engage in acts of property damage (e.g., arson, theft, vandalism, monkey wrenching/sabotage, tree spiking, and releasing animals held in captivity) against organizations and individuals they believe to be harming the environment and/or animals, but rarely target individuals. Thus, some researchers (e.g., Carson et al., 2012; Smith and Damphousse, 2009; Vanderheiden, 2005) are reluctant to use the term 'ecoterrorists' or 'ecoterrorism.' Instead the preferred terms are 'radical ecogroups' to refer to the movement and 'ecotage' – sabotage of property believed to be involved in destruction of the ecosystem or animal life – to describe the criminal activities of the movement. Carson et al. (2012) examined illegal activities from 1970 to 2007 that were motivated to protest abuse or destruction of the environment and/or animals. The authors identified 1069 criminal acts committed by environmental and animal rights extremists. Only 17% of these incidents were classified as acts of terrorism. Although Carson et al. (2012) noted the possibility of nonidentification of incidents that occurred in the 1970s, they argued that federal law enforcement had overestimated the risk posed by radical ecogroups, since only one incident in their database resulted in death. The ECDB has uncovered several additional violent plots not identified by Carson et al. (2012), however, this project is ongoing and results are not yet available (Freilich et al., 2014b).

According to Jarboe (2002, see also Freilich et al., 2011), between 1996 and 2002, environmental and animal rights extremists committed more than 600 criminal acts and have caused more than $40 million in damage. Jarboe's (2002) findings are consistent with a 2008 DHS report noted that ecoterrorists had committed more acts of terrorism in the US than any other group and caused more than $100 million in damages in the 25 years from 1981 to 2005 (DHS, 2008; see also Carson et al., 2012).

Animal and environmental rights extremists tend to be younger, have higher levels of education, and have better jobs than far-right extremists (Smith, 1994). Carson et al. (2012) noted that these individuals tended to be guided by a respect for life (though not for property) and an awareness of the legal consequences of their actions. Many of the extremists interviewed by Carson et al. (2012) mentioned a preference for legitimate advocacy activities and unwillingness to face legal sanctions for illegitimate behavior.

According to Carson et al. (2012), environmental and animal rights extremists generally operate using small cell structures or leaderless resistance, which makes it more difficult for law enforcement to arrest and convict members of radical ecogroups. However, state police officials surveyed by Chermak et al. (2010) did not believe that supporters of ELF/ALF normally commit crimes in small cells or alone. Perhaps this was due to unidentified cases in Carson's et al. (2012) study or law enforcement officials might be unaware of the offending patterns of extremists (Chermak et al., 2010). The aforementioned moral, educational, and demographic characteristics, as well as the typical group structure of radical ecogroups, have implications for desistance policies, which will be discussed in the next section.

Current Domestic Terrorism and Extremism Policies

Initial CT efforts after 9/11 focused primarily on supporters of AQAM, as is common after a focusing event. Since 2011, DHS has attempted to implement evidenced-based practices. Their current policies are based on over 75 assessments

on homegrown extremism, and extends beyond Al Qaeda-influenced individuals and groups (DHS, 2012). Current policies by national, state, and local authorities to address the national security threat of domestic extremists include: new legislation and more stringent penalties; the use of confidential informants to infiltrate terrorist groups (Center on Law and Security, 2011); CT training and CVE programs (DHS, 2012).

New Legislation and Stricter Penalties

Lum et al. (2006) found that increasing legal punishments and efforts to crack down on terrorism have almost no effect or cause a displacement effect as extremists find more vulnerable targets. Excessive compliance measures, especially those that violate civil rights and liberties, have been criticized for having a backlash effect and increasing support for AQAM (Lum et al., 2006). The overt use of this strategy against domestic supporters of AQAM declined several years after 9/11, and attempts have been made to integrate outreach programs in immigrant communities, to reduce perceptions of exclusion and the risk of radicalization (DHS, 2012). This strategy is discussed in greater detail in the upcoming section, CT Training and CVE Programs.

Excessive compliance measures can also have a backlash effect among far-rightists (Freilich et al., 2014a). Sprinzak (1995) and Kaplan (1995) argued that such excessive compliance measures would result in an 'us vs them' outlook by both law enforcement and far-rightists. This mind-set could increase the level of radicalization among far-rightists and, subsequently, the likelihood of attacks made against government officials and law enforcement (Kaplan, 1995; LaFree et al., 2009; Pridemore and Freilich, 2007; Sprinzak, 1995). Support for these claims were found by Gruenewald et al. (2013), which found that more than 50% of loners and other far-rightists who committed acts of homicide from 1990 to 2012 had at least one prior arrest. According to Gruenewald et al. (2013), many of these arrests were for nonviolent crimes. Thus, far-rightists' offending behavior progressed from nonviolent to violent crimes (Gruenewald et al., 2013). Far-rightists' offending also escalated from nonideological to ideological crimes (Freilich et al., 2014b), which further supports the thesis that excessive compliance measures increases radicalization.

Nevertheless, the arrest history of far-rightists provides opportunities for law enforcement to identify extremists before they commit a violent crime or terrorist attack. Chermak et al. (2010) recommend that law terrorism task forces be expanded to include state tax agencies, the IRS, Treasury Department, and state-land use and zoning agencies. Chermak et al. (2009a) caution, though, that penalties, which disrupt the funding sources of extremist groups, could have a backlash effect, as they find alternative, illicit sources of revenue.

However, a backlash effect is not anticipated among environmental and animal rights extremists. In fact, Carson et al. (2012) recommend increasing the penalties for ecotage committed by environmental and animal rights extremists. The authors argued that stricter penalties would not create a backlash effect and instead would reduce likelihood that environmental and animal rights extremists would engage in ecotage (Carson et al., 2012). These extremists tend to be younger, better educated, and have higher status jobs than far-rightists (Freilich et al., 2011). Thus, they risk a greater future economic loss than the typical far-rightist. Members of radical ecogroups interviewed by Carson et al. (2012) were very aware of the legal penalties and repercussion of their actions and expressed a desire to avoid such penalties. An increase in penalties for ecotage would change the cost–benefit assessment and discourage animal/environmental rights extremists from engaging in ideologically motivated offenses, especially if this tactic is combined with the provision of legal avenues to political, social, and economic change (Carson et al., 2012).

The risk of a backlash effect would also be less likely with nonextremist collaborators. As noted by Freilich et al. (2013), many supporters of AQAM and the far-right commit financial crimes to fund their organizations and future violent attacks. These financial schemes are normally accomplished with the aid of professional services provided by nonextremist accomplishes. Since these collaborators are not motivated by a deep and abiding desire to inspire social, political or religious change, they might be more likely to 'flip' if threatened with a strict sentence before being offered a plea in exchange for their testimony against their extremist cooffenders.

Use of Confidential Informants

Undercover federal agents and informants were used in 41% of AQAM terrorism cases identified by the Center on Law and Security (2011) in the 10 years after 9/11 and many of these cases were sting operations. The FBI considers sting operations in AQAM cases to be a preventative law enforcement strategy. This practice is controversial, since the operations are often planned by FBI informants (Center on Law and Security, 2011). A defense of entrapment has never been successful (Center on Law and Security, 2011), despite the fact that several AQAM supporters had a history of mental illness. Indeed, 13% of the 210 AQAM perpetrators identified by Freilich et al. (2012) had a history of mental illness. While this bolsters evidence for claims of entrapment, the law enforcement practice of using informants and sting operations continues. Undercover federal agents have had more success with the far-right movement and some (Chermak et al., 2009; Freilich et al., 2009a) have attributed the decline of far-rightist groups to this practice.

CT Training and CVE Programs

CT training currently focuses on the front line or local officers. These officers are trained to recognize terrorism and related crimes, taught how to record observations in a standardized manner, and taught how to share this information with the Federal Bureau of Investigation-led Joint Terrorism Task Forces (DHS, 2012). More than 200 000 law enforcement officers have been trained to record suspicious activities through the Nationwide Suspicious Activity Reporting Initiative. This initiative was created in collaboration with law enforcement agencies and privacy, civil rights, and civil liberties officials (DHS, 2012). CT strategy also includes the "If You See Something, Say Something" public awareness campaign (DHS, 2012). However, this campaign is listed on different DHS documents as linked to the CVE programs.

CVE takes a more community policing approach. It integrates (1) community involvement and "If You See Something, Say Something" (also known as the 'Blue Campaign') public awareness campaign, (2) training programs for local law enforcement, and (3) attempts to reduce support for terrorism by integrating immigrant populations and reinforcing American ideals/culture (DHS, 2012; see also Freilich and Chermak, 2009a who discuss applying similar strategies to far-right extremists). Although not immediately apparent, goals (1) and (3) can be incongruent: encouraging community members to be vigilant against terrorist supporters risks radicalizing at-risk immigrant populations who may perceive themselves as being unfairly targeted by the Blue Campaign (McCants and Watts, 2012). McCants and Watts (2012) criticized CVE programs for not having a clear definition shared by federal agencies. For example, DHS classifies CT and CVE training programs as distinct, however, the tactics used by both are similar – community involvement that risks ostracizing immigrant populations. Further, the increasing use of informants and sting operations (Center on Law and Security, 2011) could possibly undermine privacy, civil rights, and civil liberties of individuals. Thus, current CVE programs, which focus on potential suspects and sting operations, may at times increase radicalization among AQAM supporters (McCants and Watts, 2012). These strategies may be incompatible with the constantly evolving threats posed by different types of domestic extremists (Chermak et al., 2009b).

The points raised by Chermak et al. (2009b) that CT training programs tended to ignore the relative risks posed by different elements of the domestic extremist movements and primarily focus on AQAM supports remains valid even after the 2011 updates spearheaded by DHS. Both the "If You See Something, Say Something" public awareness campaign and attempts to integrate immigrant populations into American culture (1) ignores groups and individuals in the domestic far-right and environmental and animal rights extremist movements, which also pose a real threat to national security, and (2) unnecessarily risks increasing the radicalization among AQAM supporters (McCants and Watts, 2012).

Domestic Extremists' Offending Patterns and National Security

National security measures need to consider not just the risk of terrorist attacks, but also the criminal trajectory and offending patterns of extremist groups. Since there is widespread disagreement by academics and law enforcement about what constitutes terrorism (Schmid and Jongman, 1988; Chermak et al., 2010), obtaining reliable counts of acts of terrorism is extremely difficult. Lack of reliable data is detrimental to the creation of sound anticrime policies. According to Chermak et al. (2010), domestic extremists commit crimes that vary in magnitude (from nonviolent financial crimes, to brutal hate crimes, and bombings that cause millions of dollars in property damage and countless lives lost) and motivation (e.g., greed, anger, hate). These must also be considered to assess the risks to national security and design appropriate countermeasures. Further, as noted by Chermak et al. (2010), limiting the focus to acts of terrorism also ignores the many precursor crimes that extremists commit to fund their operations and terrorist attacks. These precursor crimes provide opportunities to law enforcement and the criminal justice system to identify extremists and to engage in general and specific deterrence before they engage in an act of terrorism.

Law enforcement actions and civil suits by the government and nonstate actors – such as watch groups – have been known to cause the decline of specific extremist groups (Chermak et al., 2009a; Freilich et al., 2009b; Smith, 1994). Skinhead groups declined in the 1980s due to a combination of 'aggressive federal criminal prosecutions and private civil rights attorneys,' which targeted key personnel (Smith, 1994: p. 91). For example, Richard Butler, the leader of the Aryan Nation, was forced into bankruptcy because of a civil suit in 2000, which lead to the decline of the Aryan Nations as many members defected to other far-rightist groups or formed their own groups.

Research has found that routine traffic stops involving far-rightists can escalate into a violent incident, or even fatality (Freilich and Chermak, 2009a). In addition to deaths or injury caused by the incident, escalation can have far-reaching consequences, e.g., it can radicalize many less extreme far-rightists who become aware of the incident or further radicalize the individual(s) involved in the violent incident (Chermak, 2002). However, the risk of a routine stops escalating into a dangerous incident is not consistent among all members of the far-right. For example, not all militia groups pose the same degree of threat. Duffy and Brantley (1997) created the Militia Threat Assessment Typology, which divides militia groups into not dangerous – category 1 to 2 – and groups that pose a major threat – category 3 to 4. There are also gradations of risk to the public among skinhead groups. Hate motivated skinhead groups tend to commit acts of violence against minorities and are more dangerous, while criminally motivated skinhead groups engage in illicit financial schemes for profit (Chermak et al., 2009b). Chermak et al. (2009b) recommended that law enforcement officials pay more attention to the more dangerous groups to maximize their desistence efforts.

The risk of escalation can be reduced by a combination of nonconfrontational dialogue by law enforcement officials and law enforcement training (Chermak et al., 2009b; Duffy and Brantley, 1997). The former humanizes law enforcement officers, while the latter would help officers to identify far-rightists (e.g., driving with no/strange license plates and referring to himself or herself as a sovereign citizen are indications for far-right ideology) and distinguish between those who belong to high threat groups, as opposed to a group that poses a minor threat. Appropriate law enforcement training would also aid in intelligence gathering (Chermak et al., 2009b). Appropriate law enforcement training would increase the likelihood that officers would know which questions to ask to determine the level of risk posed by the far-rightist. Such intelligence can then be shared with other law enforcement personnel (Chermak et al., 2009b).

Law enforcement should also pay closer attention to older extremist groups. Caspi et al. (2012) found that older far-right groups accounted for more homicides than younger groups. Older groups also had more connections to other groups (Caspi et al., 2012). A focus on networks of groups in the far-right movement would provide opportunities to law

enforcement: arrest of key personnel in older, more connected groups could have more wide reaching deterrence and desistence effects, when compared to arrest or conviction of members of less connected groups.

Conclusion

This article provided an overview of the three domestic extremist movements that pose the greatest threat to national security. However, these threats are dynamic and national security policies need to be sensitive to the overall criminal offending patterns of specific movements and the characteristics of the specific movements and its supporters. Although recent DHS-led CT polices were created in consultation with civil liberties and civil rights organizations, the lack of evidence of extremist ideology among many individuals prosecuted on AQAM cases suggests a net widening strategy that risks radicalizing alienated individuals in migrant communities. Recent DHS-led CT stated goals to counter domestic extremism and its practices that target primarily supporters of AQAM risks ignoring the threats posed by the far-right and far-left, namely environmental and animal rights extremists. Suggested amendments to national security CT efforts include widening the scope of law enforcement training to include strategies that would help identify other types of extremists, inclusion of extremists in the political process and by providing them with legal avenues to attain social, political, and religious change.

See also: Law and Terrorism; Race and Racism in the Twenty-First Century; Racism, History of; Social Movements: A Social Psychological Perspective; Terrorism.

Bibliography

Belli, R., 2011. Where Political Extremists and Greedy Criminals Meet: A Comparative Study of Financial Crimes and Criminal Networks in the United States. Graduate Center/John Jay College; CUNY.

Carson, J.V., LaFree, G., Dugan, L., 2012. Terrorist and non-terrorist criminal attacks by radical environmental and animal rights groups in the United States, 1970–2007. Terrorism and Political Violence 24 (2), 295–319.

Caspi, D.J., Freilich, J.D., Chermak, S.M., 2012. Worst of the bad: violent white supremacist groups and lethality. Dynamics of Asymmetric Conflict 5 (1), 1–17.

Chermak, S.M., Bringuel, A., Freilich, J.D., Shearer, J.K., 2011. Terrorism and counterfeiting: a synopsis of critical issues and research opportunities. In: Bringuel, A., Janowicz, J.C., Valida, A.C., Reid, E.F. (Eds.), Terrorism Research and Analysis Project (TRAP) Volume I: A Collection of Research Ideas, Thoughts and Perspectives. Government Printing Office, Washington, DC, pp. 149–187.

Chermak, S.M., Freilich, J.D., Caspi, D., 2009a. Policymakers and law enforcement must consider the unintended consequences of their proposed responses to extremist and terrorist groups. In: Frost, N.A., Freilich, J.D., Clear, T.R. (Eds.), Contemporary Issues in Criminal Justice Policy: Policy Proposals from the American Society of Criminology Conference. Cengage Learning, Wadsworth, CA, pp. 139–150.

Chermak, S.M., Freilich, J.D., Shemtob, Z., 2009b. Law enforcement training and the domestic far-right. Criminal Justice and Behavior 36 (12), 1305–1322.

Chermak, S.M., Freilich, J.D., Simone Jr., J., 2010. Surveying American State Police Agencies about lone wolves, far-right criminality, and far-right and islamic jihadist criminal collaboration. Studies in Conflict and Terrorism 33, 1019–1041.

Damphousse, K.R., Smith, B.L., 2004. Terrorism and empirical testing: using indictment data to assess changes in terrorist conduct. In: Deflem, M. (Ed.), Terrorism and Counter-Terrorism: Criminological Perspectives. Elsevier Science, Philadelphia, PA, pp. 75–79.

Freilich, J.D., Bringuel, A., Chermak, S.M., Shearer, J., 2011. Animal and environmental rights extremists and crime: setting forth a research agenda. In: Bringuel, A., Janowicz, J.C., Valida, A.C., Reid, E.F. (Eds.), Terrorism Research and Analysis Project (TRAP) vol. I. A Collection of Research Ideas, Thoughts and Perspectives. Government Printing Office, Washington, DC, pp. 263–295.

Freilich, J.D., Chermak, S.M., 2009a. Preventing deadly encounters between law enforcement and American far-rightists. Crime Prevention Studies 25, 141–172.

Freilich, J.D., Chermak, S.M., March 2009b. United States Extremist Crime Database (ECDB), 1990–2008: Preliminary Results. Department of Homeland Security University Network Research and Education Summit, Washington, DC.

Freilich, J.D., Chermak, S.M., 2012. Building the United States extremist crime database (ECDB): lessons learned. Canadian Diversity 9 (4), 18–22. Asking the right questions in the empirical measurement of security, terrorism and counter terrorism. The Association for Canadian Studies.

Freilich, J.D., Chermak, S.M., Belli, R., 2014a. Homegrown terrorism in the United States. In: Bruinsma, G., Weisburd, D. (Eds.), Encyclopedia of Criminology and Criminal Justice. Springer Verlag, New York, pp. 2336–2345.

Freilich, J.D., Chermak, S.M., Belli, R., Gruenewald, J., Parkin, W.S., 2014b. Introducing the United States extremist crime database (ECDB). Terrorism and Political Violence 26 (2), 372–384.

Freilich, J.D., Chermak, S.M., Caspi, D., 2009a. Critical events in the life trajectories of domestic extremist white supremacist groups: a case study analysis of four violent organizations. Criminology and Public Policy 8 (3), 497–530.

Freilich, J.D., Chermak, S.M., Simone, J., 2009b. Surveying American state police agencies about terrorism threats, terrorism sources, and terrorism definitions. Terrorism and Political Violence 21 (3), 450–475.

Freilich, J.D., Gruenewald, J., Chermak, S.M., Parkin, W.S., 2012. Al-Qa'ida-Related Terrorism: Violent Incidents and Foiled Plots, Preliminary Report to Resilient Systems Division, DHS Science and Technology Directorate. START, Maryland.

Gruenewald, J., Chermak, S.M., Freilich, J.D., 2013. Distinguishing "Loner" Attacks from other Domestic Extremist Violence: A Comparison of Far-Right Homicide Incident and Offender Characteristics. Criminology and Public Policy 12 (1), 65–91.

Hewitt, C., 2003. Understanding Terrorism in America: From the Klan to Al Qaeda. Routledge, New York.

Jarboe, J., 2002. The Threat of Eco-terrorism. Testimony before the House Resources Committee, Subcommittee on Forests and Forest Health. http://www.fbi.gov/news/testimony (accessed 27.03.13.).

Kaplan, J., 1995. Right wing violence in North America. Terrorism & Political Violence 7 (1), 44–95.

LaFree, G., Dugan, L., Kort, R., 2009. The impact of British counter terrorist strategies on political violence in Northern Ireland: comparing deterrence and backlash models. Criminology 47, 501–530.

Lum, C., Kennedy, L.W., Sherley, A.J., 2006. The Effectiveness of counter-terrorism strategies. Campbell Collaboration Systematic Reviews 2. http://dx.doi.org/10.4073/csr.2006.2. http://www.campbellcollaboration.org/lib/project/11/.

McCants, W., Watts, C., 2012. U.S. Strategy for Countering Violent Extremism: An Assessment. The Foreign Policy Research Institute. https://www.fpri.org/publications (accessed 27.03.13.).

Mueller, R.S., 2012. Statement before the Senate Committee on Homeland Security and Governmental Affairs. The Federal Bureau of Investigation. http://www.fbi.gov/news/testimony (accessed 28.03.13.).

Parkin, W., Freilich, J.D., Chermak, S.M., 2011. The international implications of domestic terrorism in the United States. In: Natarajan, M. (Ed.), International Crime and Justice. Cambridge University Press, New York, pp. 283–289.

Pridemore, W.A., Freilich, J.D., 2007. The impact of state laws protecting abortion clinics and reproductive rights on crimes against abortion providers: deterrence, backlash, or neither? Law and Human Behavior 31 (6), 611–627.

Schmid, A.P., Jongman, A.J., 1988. Political Terrorism: A New Guide to Actors, Authors, Concepts, Databases, Theories and Literature. North Holland, Oxford.

Smith, B.L., 1994. Terrorism in America: Pipe Bombs and Pipe Dreams. State University of New York Press, New York.

Smith, B.L., Damphousse, K.R., 2009. Patterns of precursor behaviors in the lifespan of a U.S. Environmental Terrorism Group. Criminology and Public Policy 8 (3), 475–496.

Sprinzak, E., 1995. Right-wing terrorism in a comparative perspective: the case of split delegitimization. Terrorism and Political Violence 7 (1), 17–43.

U.S. Department of Homeland Security, 2008. Ecoterrorism: Environmental and Animal Rights Militants in the United States, pp. 1–40. http://humanewatch.org/images/uploads/2008_DHS_ecoterrorism_threat_assessment.pdf (accessed 26.03.13.).

U.S. Department of Homeland Security, Office for Civil Rights and Civil Liberties, 2011. CVE Training Guidance and Best Practices. http://www.dhs.gov/xlibrary/assets/cvetraining-guidance.pdf (accessed 27.03.13.).

Vanderheiden, S., 2005. Eco-terrorism or justified resistance? Radical environmentalism and the 'war on terror. Politics and Society 33 (3), 425–447.

National Security Studies and War Potential of Nations

Steven Weber, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 15, pp. 10314–10318, © 2000, Elsevier Ltd., with revisions made by the Editor.

Abstract

Through much of the 1800s and into the first half of the 1900s, the security of a state depended heavily on the balance among its own war potential and that of its adversaries and allies. In the second half of the 1900s that relationship weakened in the context of cold war bipolarity and a nuclear standoff between superpowers. The relationship is changing again in fundamental ways at the start of the twenty-first century. 'Globalization' brings an increase in the mobility of goods, ideas, and technologies that multiply the speed with which war potential changes, and erases many of the boundaries that linked security conceptually to the war potential of *states*. Research aimed at understanding the war potential of *situations* covers some of these changes, but not all. As the ability of diverse political units to produce and deploy organized violence for political ends spreads, behavior that in the past might have been thought of as terrorism, will increasingly become more like war. The goal – realistically – might be to undermine the foundations of a state or bring down a government. In that scenario, each important term in the phrase 'national security and the war potential of nations' will have taken on a very different meaning.

Klaus Knorr wrote *The War Potential of Nations* in 1956. This book marks a key transition between different approaches toward the study of national security. Through much of the nineteenth and into the first half of the twentieth century, the security of a nation depended heavily on the balance between its own war potential and that of its adversaries and allies. In the second half of the twentieth century that relationship weakened in the context of cold war bipolarity and a massive nuclear standoff between superpowers. The relationship is changing again in fundamental ways at the start of the twenty-first century. The war potential of nations surely will continue to affect national security, but it will not nearly determine it. More importantly, almost all the terms in that statement are coming to mean very different things than they did in 1956.

Historical Perspectives

In the second half of the nineteenth century, 'war potential' was a narrow conceptual proxy for 'measuring' the military power of nations. The foundations of war potential were the economic and technological resources that could be mobilized for military purposes. This made good analytic sense in an era of industrial warfare – when nations maintained in peacetime a small fraction of the military capabilities that they could mobilize for war by simply redirecting production. World Wars I and II were won by nations with greater war potential, not by those with the advantage of existing forces at the start of the war.

Knorr understood well the promise – both analytic and practical – of being able to measure this potential. If national security were some function of objective measures of power, then parsing out the problem (at least) would be made a much simpler task for scholars and policy makers. He distinguished among three critical components of war potential – economic and technological capacity, administrative skill, and the political foundations of military power. The idea was to calculate war potential: *if* each component could be measured with reasonable precision *and* some arguments about their interactions could be built.

Knorr assessed this enterprise carefully enough to see clearly its inherent limitations. Except in the very few cases where nations use sheer force simply to seize something, the purpose of deploying military power was to influence the behavior of another actor. In all cases of influence there clearly were objective components to power. But these were always and everywhere put into play in the context of relationships or anticipated relationships that had many other components of power. Without using formal models or even 'soft' game-theory metaphors, Knorr wrote in rather precise analytic terms about the importance of information asymmetries, domestic political constraints, reputational effects, and the like in relationships of mutual coercion. He argued strongly that net assessment (the detailed counting and analysis of military forces) was by itself a very weak tool. Without reference to particular objectives and contingencies, to study and measure the war potential of nations would be a fragile approach to thinking about and planning for national security.

These arguments today seem obvious, almost trivial. What was less obvious to Knorr, and remains more difficult today, is how to move the intellectual and policy debate forward. The consolidation of a bipolar power distribution during the middle years of the Cold War, along with the deployment of massive nuclear arsenals that essentially overwhelmed any realistic measures of 'war potential,' made that problem more immediate.

Twentieth-Century Strategy

Thomas Schelling's (1960) work *The Strategy of Conflict* proposed one kind of solution. The new studies of strategy focused much less on the war potential of nations and much more on the war potential of *situations*. It did this by abstracting away from

precise measures of capabilities and even more importantly from precise understandings of the intentions and objectives of an adversary – a major analytic move during a period of time when American policy makers (at least) felt they had a poor understanding of the USSR.

At a time when net assessment was becoming intellectually stagnant (and arguably boring), the turn to 'strategic studies' opened up a new and extremely interesting intellectual agenda for national security scholars to explore. It also pulled into the picture a new set of tools that were developing rapidly in mathematics and economics – in particular, mixed motive game-theory models. The combination of newly configured research questions and fast-improving tools led to an explosion of work that clarified powerfully the strategic logic of situations like nuclear deterrence, brinkmanship crises, and coercive diplomacy.

This knowledge came at a cost, of course. The power of abstract-deductive theory to produce generalizable arguments rests in part on a set of assumptions about the characteristics of the actors or players in the 'game.' It has always been a central disagreement among scholars whether (more precisely, when and under what conditions) the benefits of this approach are worth the cost.

The end of the Cold War sharpens both the intellectual and practical consequences of that disagreement. Bipolarity was a significant clarifying force in international politics – while the Cold War was never as 'simple' as some nostalgic views would have it, it did create a marked explicitness about the major lines of conflict that the twenty-first century so far lacks. The national security environment for many nations, and surely for the USA, is now much more fine-grained than it was in 1965.

Future Directions

The problem with flattening out the past is that it makes it harder to see the future in multiple dimensions without becoming frightened and confused. It may be that the USA has enjoyed a post-Cold War honeymoon since 1990 and that the future will be increasingly challenging and unstable. Of course, there are periods of modern history where linear extrapolations of contemporary trends would have worked tolerably well as a foundation for planning the future – for example, between 1950 and 1975. There are other periods – for example 1925 to 1950, or 1975 to 2000 – where discontinuous changes in world politics, technology, social forces, and the economy radically displaced what were assumed to be core 'realities' of security planning. The period around the start of the twentieth century was such a period. Between 1895 and 1920, applications of the internal combustion engine, electricity, and the airplane – along with the rise of the German nation, social Darwinist ideologies, and decolonization – recast fundamental features of the security environment for nations.

Such periods present two distinct kinds of challenges: to master the technology, and to understand the social, economic, and military implications and possibilities that come with it. National security planners recognize the new uncertainties of threats that are smaller, more diffuse, and likely coming from many different directions. From a theoretical perspective, the strategic approach (to build increasingly sophisticated, generalized models of situations) will need at a minimum to be supplemented with renewed attention to the actors and players in the game.

Questions of National Security

The historic literature on war potential of nations is a reasonable place to start. It forces certain questions: What is war going to be? What potentials are relevant? And are nations still the appropriate central focus for the investigation? The answers to these questions were never clean and simple, and they are becoming far more complex in ways that matter deeply for thinking about national security.

The Meaning of 'Security'

One important complicating factor is simply that security has come to mean much more to many people than the core issues of territorial integrity and political autonomy. The discourse of international politics (both popular and scholarly) now embraces concepts like economic security and environmental security. There may be instrumental political reasons to import metaphors from the security debate into these other important areas of national interest, but there are intellectual costs as well.

The broadening of the realm of 'security' makes thinking and theorizing about it as a dependent variable of anything a much more complicated task. The war potential of other nations would play a small (although not insignificant) part in an analysis of US environmental security. Environmental interests are not necessarily any less important than security interests. But the clarity of focusing on security in a narrow sense as territorial integrity and political autonomy is diluted by stretching the definition. Realists argue that security is a prerequisite for almost anything else that nations want to achieve in world politics. Non-realists may disagree with the vehement emphasis on security. But few doubt that military power is still a critical restraint on the freedom of nations to act internationally, and particularly to act in ways that threaten the vital interests of the possessor of military power.

Globalization and Mobility

A second complicating factor comes from a set of transformations that fall under the rough heading of 'globalization.' More precisely, the ability of political units to produce and deploy organized violence for political ends (an updated notion of war potential) is being deeply affected by changes in the global economic environment.

The simplest way to think about globalization is to see it as a story about the causes and consequences of an increase in *mobility*. For most of human history, neither goods, nor capital, nor most people, nor many ideas moved very far from their place of origin. Over the past several hundred years, mobility of many (although not all) things has increased – unevenly across time and with some setbacks.

Political economists recognize that mobility makes possible the expansion of markets beyond the physical and other borders that previously contained them. There is simply no serious theoretical framework in which security is unaffected by

this change. The war potential of nations effectively stretches beyond national borders in a profound way. This is made more conspicuous by the 'spin-on' phenomenon. The commercial sector now leads the military sector in most key technologies that underpin military power. And the logic of the commercial sector is decidedly not national. The so-called revolution in military affairs (RMA) and particularly the increasing information intensity of military operations (which is part of that revolution) cut in the same direction. Information is inherently more mobile than tanks, airplanes, or almost any physical asset. Of course, it is also harder to secure in most settings than an aircraft carrier or a nuclear missile buried deep underground.

The most significant exception to increased mobility is people. People do not move across borders as readily as other technological and economic resources. This may be a key national factor of power that nations still control or contain – although it is not clear for how long that will remain the case. In any case, national security concerns about the increased mobility of people may complicate economic globalization as well as vice versa. Technology can be anywhere, but people still have to be somewhere, and where they are may matter a great deal for individual nation states. The current US lead in information technology owes a great deal to Indian, Taiwanese, and Israeli engineers and entrepreneurs who have immigrated to the USA (or come temporarily on H1B visas) and maintain strong network links to evolving technology clusters in their home nations. At some point nations will experience in a much more acute way the tension between the mobility of people who carry key technological talent and the 'national' security implications of allowing that talent to cross borders freely.

In addition to mobility (which has increased markedly in the past, most dramatically during an earlier period of 'globalization' at the end of the nineteenth century), the next phase of globalization is also about a dramatic increase in *speed*. The digital revolution is reducing the cost of moving a piece of information around the world in real time to near zero. Big, heavy things move much faster as well because they can be moved more intelligently with the use of information. This is unsettling to humans, who think and act slowly. It is even more unsettling to political organizations, because aggregating a bunch of slow humans into a polity usually makes them even slower. An increasing number of those polities now have some form of democratic decision-making system which has the effect (indeed, is designed explicitly in some ways) to slow things down even more.

This disconnect in speed will have profound implications for national security. There is a straightforward and narrow version of the argument. If wars will be faster – that is, time limited by military technology and by the constraints of democratic control – then existing forces may once again become a more important determinant of who wins than 'war potential' in a deeper sense. Yet there are inherent strains – and not just fiscal ones – for democracies maintaining large existing forces during peacetime.

There is a broader version of the argument as well. Balances of power have been an important stabilizing factor in international politics because balancing behavior has prevented the consolidation of hegemony by an aggressor nation. But balances do not form quickly. Napoleon nearly conquered the European landmass before an effective balance formed against him. The Grand Alliance did not fully come together to oppose Hitler until 1942 – quite late in an offensive that actually began almost a decade before. When competing systems move at increasingly asymmetric speeds, the results are likely to be unstable and surprising. Balance of power both as theory and practice has been an important part of the relationship between national security and the war potential of nations. Yet balances may become less stable and more volatile in the future than nations expect.

Size of the War Machine

A third complicating factor, loosely connected to globalization arguments, is the issue of *size*. The tools of large-scale organized violence are getting smaller. Nuclear weapons were the first modern step in this process, but they were hard to engineer and build, even for large nations. Biological weapons are different: a biological agent can be microscopic. More importantly, the equipment to produce custom-designed virulent biological agents will soon be available in suitcase size for US$10 000 or less. An effective computer virus may be even easier and cheaper to manufacture. In more than 50 years no subnational group has yet gained possession of and 'used' a nuclear weapon (including in a bargaining situation) for political purposes. Almost certainly it will not be possible to say the same of the next generation of weapons of mass destruction or disruption 50 or far fewer years hence.

Conclusions

Perhaps this kind of capability does not represent war potential in a traditional sense. One conclusion to draw from this is that it makes sense fundamentally to rethink the term itself, as well as the range of actors who can possess war potential. Changes in mobility, speed, and size have dramatically increased the porosity of national borders. Given that porosity, behavior that in the past might have been thought of as terrorism could increasingly become much more like war, in the sense that the goals of the behavior might not simply be revenge or the placing of pressure on a particular, discrete policy. The goal – realistically – might be to undermine the foundations of a nation or bring down a government. In that scenario, each important term in the phrase 'national security and the war potential of nations' will have taken on a very different meaning.

Editor's Note

Security and security studies in the twenty-first century are discussed in Alan Collins (2013), Collin S. Gray (2011), Jonathon R. White (2012).

See also: Cold War, The; Geopolitics; Military Geography; Military History; Military Sociology; Military, War, and Politics; Warfare in History; Wars among Nation-States: Patterns and Causes.

Bibliography

Art, R.J., Waltz, K.N., 1999. The Use of Force: Military Power and International Politics, fifth ed. Rowman and Littlefield, Lanham, MD.

Bull, H., 1977. The Anarchical Society: A Study of Order in World Politics. Columbia University Press, New York.

Collins, Alan, 2013. Contemporary Security Studies, third ed. Oxford University Press, Oxford.

Collin, S. Gray, 2011. Hard Power and Soft Power: The Utility of Military Force as an Instrument of Policy in the Twenty-first Century. Strategic Studies Institute monograph, Washington.

Jonathon, R. White, 2012. Terrorism and Homeland Security, seventh ed. Wadsworth Cengage.

Katzenstein, P.J., Social Science Research Council (US) Committee on International Peace and Security, 1996. The Culture of National Security: Norms and Identity in World Politics, New Directions in World Politics. Columbia University Press, New York.

Kennedy, P.M., 1987. The Rise and Fall of the Great Powers, first ed. Random House, New York.

Knorr, K., 1956. The War Potential of Nations. Princeton University Press, Princeton, NJ.

O'Rourke, K.H., Williamson, J.G., 1999. Globalization and History: The Evolution of a Nineteenth-century Atlantic Economy. MIT Press, Cambridge, MA.

Ruggie, J.G., 1996. Winning the Peace: America and World Order in the New Era. Columbia University Press, New York.

Schelling, T.C., 1960. The Strategy of Conflict. Harvard University Press, Cambridge, MA.

Walt, S.M., 1999. Rigor or rigor mortis? Rational choice and security studies. International Security 23 (4), 49–80.

Waltz, K.N., 1964. Man, the State, and War: A Theoretical Analysis. Columbia University Press, New York.

Weber, S., 2001. Globalization and the European Political Economy. Columbia University Press, New York.

National Socialism and Fascism

Hans Mommsen, Feldafing, Germany

© 2015 Elsevier Ltd. All rights reserved.

Italian Fascism and National Socialism signify a new type of political mobilization, which came into being after the end of World War I. Both movements were ideologically influenced by integral nationalism and racism but drew their specific political impetus by exploiting the widespread anticommunism and antisocialism among the bourgeois strata of postwar society. Italian Fascism in particular was taken as a model by comparable movements all over Europe, but, with the exception of that in Spain, only Italian Fascism and National Socialism were able to establish a dictatorship.

Italian Fascism

Origins

The origins of Italian Fascism are closely connected with the internal weakness and lack of legitimacy of the liberal regime reflected in a series of protest strikes and factory occupations by rural and industrial workers. Hence, Giovanni Giolitti tolerated the use of violence against organized labor by the fascist *Squadre d'azione*, which resulted in the emergence of a dualistic power structure. It was reflected in Gabriele D'Annunzio's expedition to Fiume in 1919 as well as in the actionism of the Blackshirts terrorizing the countryside and socialist communities. Benito Mussolini seized the opportunity to establish the veterans organization *Facio de Combattimento* and to shape it into a nonartisan mass movement, which he reorganized in November 1922 as *Partito Nazionale Fascista* (PNF). By threatening the governing liberal cabinet by activating the revolutionary potential of quadrism and by unleashing the march on Rome on 27 October 1922, Victor Emmanuel III installed him as Prime Minister in a right-liberal coalition cabinet, although the PNF controlled only a minority of seats in the parliament.

Fascist Dictatorship

Mussolini used his new position to establish his personal dictatorship by skillfully extending the prerogatives of the Prime Minister through legal means. By an ostensibly manipulative reform of the electoral system, he attained an overwhelming parliamentary majority for the governing coalition in the 1924 elections. By a series of 'fascist decrees,' the civil liberties, the freedom of the press, and the space of action for the parliamentary opposition, which formed the antifascist Aventin, were severely curtailed. The ensuing Matteotti Crisis provided a pretext to dissolve the opposing parties and to promulgate a one-party state relying on the extraparliamentary power of the Fascist Great Council. Mussolini thus managed to maneuver both the cabinet and the parliament – which in 1938 was reshaped on a corporatist basis – into the background. The introduction of the fascist calendar and the adoption of the fascist emblems as state symbols completed the creeping process of shaping a fascist state. Mussolini achieved this by carefully evading open clashes with the constitution and achieving a *modus vivendi* with the conservative elite and big business.

Mussolini's tactical flexibility climaxed in concluding the Lateran treaties, which, by comprehensive concessions in social and cultural matters, secured the support of the Catholic Church. The combination of social propaganda and compulsion, as reflected in the program of '*dopo lavoro*,' together with cooperation with big business, enabled him to neutralize the syndicalistic trade unions. The PNF, which in 1943 was composed of 4.75 million members, combined heterogeneous social and political groups and exhibited strong interregional tensions. By isolating the party radicals under Roberto Farinacci, who endorsed the principle of forming a cadre party, the PNF became more or less bureaucratized and was controlled through local and regional administration. Thus, despite terrorist police and the abolition of civil rights by the *lege fascissimi* in 1925–26, the system was not completely totalitarian, particularly in comparison to the Nazi regime. The monarchy, the army, and certainly the Catholic Church preserved a certain independence, and the fascistization of the administrative apparatus was only partially attained. Mussolini's rule relied primarily on a continuous organization of consensus – his domestic policy lacked any clear direction and ideological consistency. The myth of the indispensable role of the Duce for Italian national survival kept everything together, but the disastrous consequences of Mussolini's alliance with Hitler raised resistance among not only the conservative elite and the military but also his own chieftains, who eventually used the instrument of the Great Council to depose him in July 1943.

National Socialism

Origins

The Nationalsozialistische Deutsche Arbeiterpartei (NSDAP) belonged to the series of *völkisch* groupings emerging in the early postwar period in Bavaria, under the impact of the Munich Soviet Republic. The NSDAP originated from the Deutsche Arbeiterpartei (DAP), which was founded by Anton Drexler and Karl Harrer and affiliated to the völkisch Thule Society in Munich. This was a clandestine affiliation of the Pan-German League, which also founded the German *Völkisch* Protection League (Deutschvölkischer Schutz und Trutzbund), designated to draw the labor class into the nationalist camp. A majority of the NSDAP leadership started its political career in the DVST.

In November 1919, Adolf Hitler found a job as political informant for the Bavarian Reichswehr to which he belonged until 1920. Starting as völkisch propagandist, he joined the DAP, which was still an insignificant splinter party among

a number of similar associations. Backed by a group of influential sympathizers in Munich, he first seized power within the DAP in 1921 and shaped it into a fascist party by enforcing the leadership principle and focusing its activities only on continuous propaganda, which connected the fight against the Versailles peace treaty with the mobilization of antirepublican resentments especially in bourgeois strata, together with an extremist racial anti-Semitism. Simultaneously, Hitler succeeded in eliminating intraparty elections and debates.

Even with its stronghold in Bavaria, the Nazi party remained dependent on right-wing conservative support. The projected march on Berlin, according to the model of Benito Mussolini in October 1923, with the aim of establishing a national dictatorship under Erich Ludendorff, the former chief of the imperial army and now leader of the völkisch wing of the German nationalists, was postponed. Hitler, who claimed the political leadership of the right-wing opposition and its paramilitary organizations, ultimately decided to act independently from his conservative partners, but the attempt to proclaim his dictatorship and to gain the approval of the national camp by his propaganda march to the Feldherrnhalle on 9 November 1923 failed completely. The indictment of Hitler for high treason and his imprisonment in the Landsberg penitentiary until December 1924 did not impair his political prestige. The prohibition of the party and its paramilitary wing – the *Stormtroopers* (SA) – weakened, but did not eliminate the successor organizations. Thus, Hitler after his premature dismissal was able to restore the NSDAP and to eliminate competing successor organizations. By the formal dissolution of the NSDAP and its reinvention in 1925, he successfully excluded possible rivals and restored the unrestricted leadership principle, while restricting the party activities exclusively to unrelenting agitation.

In 1936, Hitler succeeded in neutralizing the opposition by the Northwestern group of the Gauleiter, which was controlled by Gregor Straβer and originally was supported by Joseph Goebbels. Their intention to modernize the original party program of 1920, which did not comply with the conditions of the stabilization period, was successfully blocked by Hitler who prevented any constructive programmatic debate. Hence, the left wing led by Otto Straβer left the party in 1931; Hitler's leadership no longer was questioned.

Despite the organizational improvements under Gregor Staβer and the extended propaganda apparatus, the NSDAP even in the Reichstag elections of May 1928 acquired just 2.6% of the votes. But in the following years, the NSDAP profited from the accelerating erosion of the electoral support of the bourgeois middle parties. This trend made itself felt already in the local and regional elections early in 1929 and culminated in the Reichstag elections of September 1930, where the party gained 18% of the votes. Only after this breakthrough the NSDAP became a full-fledged mass movement and, therefore, a serious factor in German politics. Composed originally of mainly protest voters and pensionists, peasants, and former nonvoters, it tended to become an omnibus party, with growing support from the upper middle class. This, however, was connected with an extreme fluctuation among party members as well as voters.

The Seizure of Power

The success of the Nazi party at the polls appeared spectacular, but at its peak in July 1932 it attracted no more than 37.3% of the vote and, in November, suffered a visible setback, while in the 5 March 1933 elections it obtained a slight majority only combined with the conservative German Nationalist Party. Thus, Hitler's nomination as Reich Chancellor in the 'Cabinet of National Concentration' on 30 January 1933 by Reich president Paul von Hindenburg was far from being the outcome of the democratic process, being due to informal pressures exerted by conservative interest groups and the army leadership on the Reich president whose spokesman, Franz von Papen, believed that he could control Hitler within the Cabinet by a majority of conservatives. Papen convinced the reluctant president that the Hitler cabinet was virtually a majority government pretending that the Catholic Centre Party would eventually join the coalition.

Hitler by scheduling premature elections used his majority to get rid of the bourgeois encirclement, compelling his conservative partners to accept the Enabling Law of 23 March 1933. This law, together with the emergency powers of the Decree for the Protection of People and State issued on 28 February after the burning of the Reichstag, invalidated civil liberties and legalized sanctions against the socialist and communist opposition, thereby paving the way for unrestricted dictatorship.

In the first phase of his chancellery, Hitler was preoccupied with the consolidation of his power through a systematic policy of Gleichschaltung (coordination) of public and private institutions: subduing them to party control and placing leading Nazis in high administrative positions without changing the governmental apparatus and its conservative persuasion – except for the installment of Joseph Goebbels as Minister for Public Enlightenment and Propaganda. The radical wing of the party, especially the huge paramilitary organization of the SA under Ernst Röhm, who called for a 'second revolution' and demanded to interfere in the economy as well as take control of the armed forces, was neutralized by the events of 30 June 1934, when Hitler under the pressure of the SS, Hermann Göring, and the army command ordered the immediate liquidation of the Supreme SA leaders. This action led his conservative partners to believe that he was basically a moderate and would put an end to the party's influence on the political system.

The Nazi State

After Hitler had successfully eliminated the influence of his conservative cabinet members, he used the death of Reich President Hindenburg in August 1934 to consolidate his power by combining the presidency with the office of the Reich Chancellor, thereby taking over the command of the armed forces, which in order to prove their loyalty swore as the civil servants an oath of personal allegiance to the Fuehrer. In conjunction with the expanding leadership cult, Hitler disliked to cooperate with the Reich cabinet whose meetings lost their former significance and were finally canceled at all. The contact with the resorts lay in the hand of the Reich Chancellery serving as a clearing agency between Hitler and the governmental apparatus. Actually, Hitler showed less and less interest in

day-to-day politics and refrained from interfering in internal disputes insofar his personal prestige was not involved. Instead of using governmental institutions, Hitler preferred to meet his decisions on the basis of informal information supplied by frequently irresponsible advisers, whereas the Reich Chancellery, the task of which was above all to coordinate the departments, was circumvented by an increasing number of ad hoc agencies established between party and state and derived their authority from Hitler himself. Besides the Stormtroopers (SS), which took over the complete police apparatus, the German Labour Front, which copied the role of the former trade unions, as well as Hermann Göring's Four Year Plan Office were not controlled by the central bureaucracy.

Hence, the Nazi system of government was not at all signified by any 'unity of will' or clear-cut subordination patterns, but by the principle of unlimited freedom of action for the individual position holders, as long as they did not tangle Hitler's personal prestige. Therefore, this system of 'government without administration' instigated a quasi-socio-Darwinist struggle between competing agencies and never-ending personal rivalries among the party chieftains. During the so-called combat period, this system favored the party's dynamism. In the regime phase, where mere propagandists' mobilization became obsolete, internal rivalries and competition turned out to become counterproductive and utterly detrimental for any regular decision-making process, resulting in an accelerating cumulative radicalization. In the long run, the system implied growing disarray and self-destruction. This was supported by the ambition of the party chieftains after 1943, to takeover an all-embracing political control and to continue des party revolution, which had been stopped in 1933–34.

Military Aggression and Racial Annihilation

Analogous to his domestic policy, Hitler dealt with foreign affairs by adapting a 'trial and error' method. Initially he pursued the path of a revision of the Versailles treaty, although being resolved from the start to establish a German hegemony over Europe beyond the territorial limitations of Versailles and to conquer 'living space' (*Lebensraum*) in Eastern Europe. His first step, however, involved breaking the diplomatic isolation of the German Reich, which had been increased by her retreat from the Geneva disarmament conference and the League of Nations in October 1933. By signing the German–Polish nonaggression treaty that ran counter to previous conservative foreign policy, and by concluding a naval agreement with Great Britain, Nazi foreign policy succeeded in undermining the Stresa Front. Despite conflicting interests with respect to Austria's independence, Mussolini entered the Axis alliance in 1936 and joined the Anti-Comintern Pact thereafter.

The domestic corollary of the transition to active foreign policy lay in stepping up the clandestine rearmament, which was already on the way before the seizure of power. Financial and economic policies both gave overall priority to rearmament, culminating in 1936, when the Four Year Plan required the German army to be capable of waging war within a period of 4–6 years. The remilitarization of the Rhineland as well as the introduction of conscription in 1935 was also accepted by the Western powers, despite the fact that they obviously violated the demilitarization provisions of Versailles. The Western powers still believed it was possible to reach an agreement with Hitler and tame his territorial ambitions, even after having acceded to the annexation of the Austrian Republic in March 1938. Not November 1937, but the spring of 1938 became the climacteric of the progress toward open aggression. Although Hitler had played a rather passive role in the Austrian crisis (which had been pushed ahead by Hermann Göring), he was resolved to settle the Sudeten conflict by military means, and his decision to take over the war ministry himself enabled him to act. Neville Chamberlain's decision to continue the policy of appeasement and Mussolini's mediation forced the dictator to reluctantly accept the provisions of the Munich conference in September 1938; however, any expectation that the cession of large parts of Czechoslovakia would curb Hitler's aggressive appetite proved to be utterly wrong – as was proved by the conquest of independent Czechoslovakia and the establishment of the Reich protectorate Bohemia–Moravia in March 1939.

The remilitarization of the Rhineland as well as the introduction of conscription obviously contradicted with the demilitarization provisions of the Versailles peace treaty, but the Western powers did not intervene, still believing that is was possible to reach an agreement with Hitler and to accept his assurance that after the unification of Austria with the Reich there would not be any further territorial demands by Germany. The opposite was true. While Hitler had played a rather passive role with respect to the Austrian crisis, which had been pushed ahead by Hermann Göring, he now was resolved to settle the Sudeten conflict by military means. After the Blomberg crisis Hitler took over the war ministry by himself was resolved to go to war this respect to Czechoslovakia. Neville Chamberlain's intention to avoid a military conflict for the time being as well as Benito Mussolini's intervention that had been endorsed by Göring compelled him to accept the provisions of the Munich conference in September 1938. The expectation that the cession of large territories of Czechoslovakia might curb Hitler's aggressive appetite proved to be utterly wrong and did not prevent the destruction of independent Czechoslovakia in March 1939 and the establishment of the Reich protectorate Bohemia–Moravia.

World War II

Different from the First World War, where the responsibility must be ascribed to the major European powers, the Second World War was the outcome of Hitler's policy to gain for Germany the predominance in Europe and even her rise to a world power. After having achieved the destruction of Czechoslovakia, Hitler was resolved to settle the Polish question either by compelling her to become an anti-Soviet satellite or by using military means. The dictator believed that it would be possible to extinguish the Polish Republic before an intervention by France and Britain. This calculation failed, and Hitler became engaged in a two-front war, which he all the time wanted to avoid. The German–Soviet nonaggression pact strengthened the military and economic position of the Reich, and the ensuing occupation of the Benelux and Baltic states and also the defeat of Greece and Yugoslavia, turned the power balance toward the side of Hitler and the axis powers.

The unexpectedly rapid victory over France did not induce Winston Churchill to accept Hitler's allegedly generous peace offer. The German defeat in the air battle over the English Channel and only short-term successes of the Wehrmacht in North Africa led to a military impasse and would in the long run weaken the German strategic position. Hitler tried to overcome this strategic stalemate by attacking in June 1941 the Soviet Union, which he regarded as the British 'Festlandsdegen,' but already the German defeat in the battle of Moscow in December 1941 destroyed any expectation of an imminent collapse of the Soviet Union, and in the other theaters of war, especially in North Africa and finally Italy, the German army was forced into the defensive.

Hitler was resolved to wage the campaign against the Soviet Union as a war of unrestricted racial and ethnical annihilation. The war should lead not only to the defeat of the Bolshevik regime but also to the complete destruction of the Russian statehood and to a systematic ethnic cleaning of wide parts of the Soviet territory in order to provide living space for the German or Germanic settlers who should build the core of Himmler's gigantic 'General Plan' project. At the end, there was not only the loss of the former Eastern territories and up to 40 million people but also the destruction of the German cities by the Allied Bombing offensive, not to talk about the losses especially of Russian soldiers and civilians. The European war ended with the German capitulation on May 8 and 9 after Hitler had committed suicide in the Fuehrer bunker and resulted in a complete debellatio.

Racial Annihilation

Aggressive anti-Semitism became more and more the center of National Socialist agitation since socialists and communists were no longer their predominant target. The boycott of 1 April 1933 should be a starting point for an anti-Jewish mobilization of the population, but remained restricted to the radical wing of the Nazi movement, while the average people remained distanciated. Hence, the boycott had to be finished after 1 day and Joseph Goebbels decided to utilize instead the strategy of legal restrictions against the Jewish population, which climaxed in the promulgation of the Nüremberg Laws in September 1935. The pogrom of November 1938 was the starting point for the systematic expropriation of the Jewish population. Despite the social segregation and economic expropriation of the Jewish population, the economic pressures prevented any further emigration. The strategy of enforcing Jewish emigration reached an impasse in 1939, although it was supported by the regime until October 1941. By the annexation of Austria and Czechoslovakia and the inclusion of Poland and other East European states, the number of Jews within the German power block had been quadrupled. After the Polish and the Russian campaign, the German leadership planned to replace the emigration concept by creating Jewish reservations, at first at Nisco in Poland, finally by the Madagascar plan. All projects of this kind failed due to strategic or technical reasons and were regarded as provisory steps in the road of achieving a 'final solution' after the war. The role of the SS task units in the occupied Soviet territory started with the liquidation of Soviet functionaries and Jews in leading positions, but since August 1941 they would kill all Jews including women and children. The great majority of the indigenous Jews had already been liquidated by October 1941 when the establishment of annihilation camps and the use of gas vans created the technical facilities for the now intended systematic mass murder, now comprising German and West European Jews as well as the Jews living in Poland and the Southeast European countries under German control in the so-called action Erntefest (harvest festival). The systematic murder of all Jews living under German rule was discussed at the Wannsee Conference on 20 January1942 and then implemented by Adolf Eichmann since April 1942.

The 'Holocaust' was not the outcome of a premeditated plan, but resulted from an interaction between the proceedings on the local level and the Reich Main Security Office under Heinrich Himmler and Reinhard Heydrich. There was no need for any formal order by the dictator whose part consisted mainly in instigating continuous radicalization, without being involved in the concrete destruction process for which the close cooperation between the SS, the Wehrmacht, and the civil administration was indispensable. At the end, about 5.3 million Jews were murdered. Although the news of the anti-Jewish atrocities were distributed by the international press and there were many leaks to reveal the fate of their former neighbors, the knowledge about the dimension of the Shoah among the German population was limited, set aside that many Germans tended to repress the Jewish genocide committed by the Nazi regime.

The Dissolution of the Third Reich

The capitulation of the 6th Army in Stalingrad on 31 January 1943 marks the definite turn of tide against the Axis powers. Because of its lack of manpower, ammunition, and technical equipment, the Wehrmacht did not regain the strength necessary to continue strategic operations and was forced onto the defensive. The breakthrough of the Red army at the middle sector of the Eastern front in the summer of 1944 opened its road into the Reich despite unremitting German attempts to stabilize the front. After the Allied invasion of France and the Allied air offensives, the military situation became hopeless, and it was just a question of time until the Reich was defeated.

The artificial prolongation of the fight relied on the total suppression of domestic opposition, since the army had lost its relative autonomy after the abortive putsch of 20 July 1944. Unlike in Italy, there was no institution left that could take an initiative. The renewed efforts of fanatical party functionaries to take over public administration and to urge for fighting to the last man by means of a progressively tenuous hold over propaganda, and a rule of terror by martial law, signified the last phase of the war – during which Hitler and his chieftains in Berlin had clearly lost any grasp on reality. The end of the Third Reich was composed of not only total military defeat but also the complete collapse of its institutional framework.

European Fascist Movements

Fascist movements originated in most parts of Europe during the interwar period, but only in exceptional cases were they

able to exert any great political influence or establish – as in Italy Germany – a regime of their own. Most of these fascist parties never developed into mass movements, especially within the rural societies of Southeastern and Eastern Europe. Only the Romanian 'Iron Guard' and the Hungarian Nimiety Karat Partial ('Arrows of the Cross') achieved temporarily significant popular support comparable to the PNF. In two cases, fascists were able to rule for a limited period: the Stash under Ante Pevely in Croatia in 1941 and the Arrows of the Cross under Fervency Salami in Hungary in 1944 – both relying on the German occupants. The Link Guards in Slovakia were amalgamated in the authoritarian Assemblymen of Josef Tyson and lost their independence. In Austria, competing fascist movements – National Socialists and the Headwear (Home Guard) – attracted considerable portions of the electorate, but were suppressed (like the National Socialists) or became absorbed by the state-controlled Vaterlandische Front (Patriotic Front).

In Britain, France, Belgium, and the Netherlands, as well as in Norway, fascist groups played an insignificant role. In Spain, the influential fascist Falange was tamed by the authoritarian rule of General Franco and lost its original impetus. Similar developments prevailed in Southeastern Europe where parliamentary democracy was replaced by authoritarian regimes with varying populist leanings. The majority of the fascist movements remained splinter parties and stood in opposition not only to the left but also to the ruling conservative forces and could share power only with the support of the German occupants. Those fascist groups that sympathized with National Socialism – such as the Partie Populaire Française of Jacque Doriot in France, the Front Populaire des Rex of Léon Degrelle in Belgium, the Nationaal-Socialistische Beweging of Adriaan Anton Mussert in the Netherlands, and the National Samling of Vidkun Quisling in Norway – exerted only marginal influence, and this was even more true for Oswald Mosley's British Union of Fascists.

Structural Elements

Italian Fascism provided the model for similar movements, especially for National Socialism that copied techniques, symbols, and political strategies from its Italian counterpart. Although some historians reject the use of a general typology of fascism including National Socialism, the similarities between them cannot be disregarded; however, the extremely destructive and inhuman character of the Nazi experiment appears to be exceptional. The varieties of fascism are characterized by their rigorous antiliberal and anticommunist stance as well as by their extreme nationalism and racism, their specific populist ideology, and voluntarist and decisionist political styles. They can be distinguished from right-wing bourgeois parties by their techniques of exploiting resentments and prejudices of the electorate, especially nationalism and anti-Semitism or – as in the case of Italy – racism, for mass mobilization, and of avoiding any choice between contradictory programmatic targets, thus replacing actual political issues by purely propagandists slogans, although many of their goals are similar to those of the bourgeois right.

The 'cult of the will' – the arbitrary pursuit of utopian visions first articulated by Georges Sorel – and the myth of the unity of the people ('people's community') are specifics for the fascist variety of politics that added a new element to the political process. Although the antiliberal and antienlightenment impetus is a general criterion, there exist differences with regard to the degree of parasitic disruption of the inherited state apparatus, which culminated in the National Socialist regime in Germany, caused by their technique of mobilizing nonsynchronous popular resentments and prejudices.

Fascist movements tend to replace liberal organizational principles by means of amorphous personal legacies culminating in the leadership cult. Mass mobilization is regarded by them as an end in itself, even at the price of a chameleonlike propaganda. As genuine protest movements, their membership and constituency are characterized by an overrepresentation of youth, of marginal social strata, and of social outcasts. Frequently, modernization crises provide a fertile ground for emergent fascist movements. They are specifically a phenomenon of the interwar period, although a series of successor organizations did emerge after 1945, such as Peronism in Argentina.

See also: Antisemitism; Capitalism; Conflict: Organizational; Dictatorship; Discrimination: Racial; Ethnic Cleansing, History of; Ethnic Conflicts; Eugenics as a Basis of Population Policy; First World War, The; Holocaust, The; Imperialism; Judaism; Labor History; Mass Killings and Genocide, Psychology of; Minorities; Nationalism, Sociology of; Political Thought, History of; Racism, History of; Racism, Sociology of; Russian Revolution, The; Second World War, The; Terrorism; Violence, History of; Warfare in History.

Bibliography

Evans, R., 2005–2009. The History of the Third Reich, 3 vols.
De Felice, R., 1965–1997. Mussolini, 8 vols. Turin.
Kershaw, I., 2001. Hitler, 2 vols, p. 24.
Laqueur, W.Z., 1991. Fascism. A Reader's Guide Aldershot.
Mueller, R.D., 2004. Der Zweite Weltkrieg. Stuttgart.
Paxton, R., 2006. The Anatomy of Fascism.
Payne, Stanley G. History of Fascism. p. 204.
Schieder, W., 2008. Faschistische Diktaturen. Studien zu Italien und Deutschland.
Weber, E., 1964. Varieties of Fascism. London.
Woolf, S.J., 1981. Fascism in Europe.

Nationalism: General

Partha Chatterjee, Columbia University, New York, NY, USA; and Centre for Studies in Social Sciences, Calcutta, India

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although nations often claim ancient genealogies, nationalism is connected to the need in modern industrial societies for a standardized language, a homogeneous national culture, and an imagined horizontal community. The modern political conception of popular sovereignty identifies the people with the nation-state. Liberal theorists distinguish between civic nationalism as a voluntary fraternity of equal citizens and ethnic nationalism based on natural organic ties of community. From the Americas in the eighteenth century and Europe in the nineteenth, nationalism spread to the colonized countries of Asia, Africa, and the Caribbean in the twentieth century. Besides its political aspect, nationalism has a major role to play in major aspects of modern cultural life in many countries such as language, literature, art, gender, food, sports, etc. While there have been suggestions that postnational formations such as the European Union and the rise of a global civil society are about to supersede the nation-state, there are opposite pressures as well for autarkic economic policies and chauvinistic ideologies.

Most theorists are agreed that in the specific form in which it is known today, nationalism is closely related to the modern state and its politics. Its rise and spread are linked to the need in modern industrial societies for a standardized language, a homogeneous national culture, and an imagined horizontal community. The modern idea of popular sovereignty was expressed in the identity of the people with the nation-state. This in turn legitimized the rise of anticolonial movements as a necessary unfolding of modernity. But at the beginning of the new millennium, the relations of nationalism to the global order, on the one hand, and to the particularisms of ethnicity, gender, and class, on the other, are facing new challenges.

The Modernity of Nationalism

There is a debate about nationalism's age. Some argue that nations are old, built as they often are around histories of ethnic, linguistic, religious, or dynastic solidarity going back several hundred years before the modern era (Smith, 1986). European countries such as France, Portugal, Denmark, Sweden, or Russia and Asian countries such as China, Japan, or Korea are often portrayed as having old and continuous histories in which a well-defined national identity can be easily perceived. The Jewish people are said to have been a nation, distinguished by ethnicity and religion, since Biblical times. Other analysts hold that even when nations claim ancient lineages through such ethnic and cultural histories, the particular combination of elements that gives to a people the distinct form of a nation is always a product of the modern era (Gellner, 1983; Hobsbawm, 1990). Various dates have been suggested as the point of emergence of nationalism as a modern phenomenon. Some say it was the French Revolution, others the independence movements in Latin America, while still others think that it was the German Romantic movement that supplied the crucial ingredient. But it is generally agreed that by the early nineteenth century, most of the elements that would characterize nationalism were in place within the emergent discourses of politics and society in the Western world. Henceforth, nationalism would be a constituent aspect of modernity itself.

One of the doctrines that nationalism has turned into a feature of common sense is that the world is divided into nations, so that every person, under normal circumstances, is born into a nationality. This common understanding is bolstered by such everyday visual representations as multicolored maps of the world, identity documents such as passports, and international gatherings such as the United Nations or the Olympic Games. These representations are built on another modern doctrine that has become a truism – except for Antarctica, the entire landmass on earth is divided up into national territories, each of which is under the unique sovereign jurisdiction of a state. The two doctrines together add up to a moral claim – that the proper form of modern politics is to be found where the political unit of the state coincides with the cultural form of the nation.

It is not difficult to see the strong connections between these doctrines of nationalism and some of the basic tenets of modern political theory. The definition of national sovereignty as the unique jurisdiction of a single state authority over a territory clearly demarcated on maps and recognized by other state authorities had emerged in Europe in the seventeenth century. What was added in the eighteenth century was the revolutionary idea that the true locus of this sovereignty is the people. Following the course of antiabsolutist thinking in the period of the French and the American revolutions, just as there emerged the idea of the citizen as the autonomous bearer of rights, so also was it asserted that the people, self-constituted and self-determining, are autonomous. This doctrine would supply the moral justification for the nineteenth-century democratic national movements in the territories of the Habsburg, the Ottoman, and the Czarist empires. In the twentieth century, with the rise and success of anticolonial movements in Asia and Africa, this would become recognized as the right of self-determination of nations.

The Morality of Nationalism

Although nationalism shares its genealogy with the modern state and its politics, it has not always found moral approval. The grounds of criticism, however, usually lie in ideological

debates that are part and parcel of modern politics itself. A common criticism is based, for instance, on the distinction between those nationalisms that are more voluntarist in their membership, built around institutions that are more like civic associations, and those that are more organic where membership is ascribed by birth into a collective bound by historic ties of cultural community. The former is said to be the more desirable type, since it respects the right of the individual to choose his or her nationality; the latter is likely to place greater value on the community than on the individual and could thus easily lead to xenophobia and authoritarian politics. This is a distinction that is at the heart of liberal theories of nationalism and usually classifies nations like France, Britain, the Netherlands, Belgium, and the United States as belonging to the first type and those of Central and Eastern Europe, Russia, the Middle East, and most of Asia to the second (Kohn, 1944).

Civic nationalism is meant to think of the political community of the nation as a voluntary fraternity of citizens whose rights of equality and liberty find fruition in the nation-state. The organic idea of the nation, on the other hand, seeks a more natural basis of community, such as, very influentially, in the German Romantic notion of a folk united by the historic associations of a common language and shared myths. The former, the liberal would say, tends to be universalist and inclusive in its political vision, the latter particularist and exclusive (Greenfeld, 1992; Tamir, 1993). In actuality, of course, even liberal states subscribing to the universalist values of civic association have gone through much conflict before extending their citizenship to those who were not regarded as part of the 'original' political community – Catholics in the United Kingdom, for instance, or Afro-Americans in the United States – and the debate still goes on in many Western countries over the status of the new immigrants whose race, language, or religion marks them out as different from the 'normal' member of the political community.

Of modern political ideologies, Marxism in its early years strongly upheld the universalist idea of the political community and regarded nationalism as a limited and transient phenomenon, which, if espoused by the working class, would impede the truly emancipatory path of proletarian revolution. The emerging democratic movements in the territories of the Austro-Hungarian empire, however, impelled the Austrian Marxist Otto Bauer to the view that social democrats should support the demand for independent nation-states. Later, in the context of the nationalities question in Russia and the discussions in the Communist International on the colonial question, Lenin argued for the right of self-determination of nations and Stalin proposed a set of objective criteria such as shared territory, language, and history by which nationalities were to be identified (Nimni, 1991).

The Sociology of Nationalism

Quite apart from the doctrinal terms in which nationalists describe their project, sociologists have tried to provide an account of the sociohistorical conditions that produce nationalism. Ernest Gellner has identified these as the conditions of modern industrial society, which requires free movement of labor and goods throughout the national economy, a general and universal schooling in a standard national language, and a rhetoric of equal access to social mobility. The complexity, interdependence, and mobility of industrial life place an unprecedented emphasis on complex, precise, and context-free communication. This tends to destroy both the exclusiveness of premodern elite high cultures and the provincialism of local peasant cultures and produce in their place an increasingly standardized and homogeneous national culture. The transition to some form of modern industrial society, Gellner argues, is unstoppable. Consequently, nationalism too is an ubiquitous feature of our contemporary world (Gellner, 1983).

If nationalism is indeed the product of certain structural features of modern industrial society, why is it that its particular cultural content is so often the focus of such intense ideological passion and contestation? Benedict Anderson has sought to supply the link between the material and the ideological. The nation, he says, is an *imagined* political community. It is imagined, first, in the sense that, although no single member personally knows all the other members of the nation, in their minds they all live in community. It is imagined, second, as a sovereign community, in line with the modern political idea that the emblem of freedom is the sovereignty of the people. It is also imagined, third, as a limited and particular community, because humankind is divided into many nations. Finally, it is imagined as a community – one that creates a deep, horizontal comradeship. This imagining of the nation, however, becomes possible only through certain specifically modern technologies and institutions. The most crucial is the technology of printing, which, in combination with the new production system of capitalism, gives rise to a publishing industry that produces, through newspapers, textbooks, novels, and bureaucratic documents, a national literature in a standardized national language. Newspapers and novels, in particular, make it possible to imagine lives lived in simultaneity with other real and fictional members of the community, convert events happening to other people at other places into events in the shared life of the community, and thus transform an imagined solidarity into the homogeneity of everyday lives. Censuses, maps, museums, and memorials are other important institutions that contribute to the imagining of nations (Anderson, 1991).

Nation and State

The modern form of the nation is both universal and particular. The universal dimension is represented by the notion of 'the people' as the original locus of sovereignty in the modern state and of all humans ('man,' in the original eighteenth-century revolutionary declaration) as bearers of rights. If this was held to be universally true, how was it to be realized? By enshrining the specific rights of *citizens* in a state constituted by a particular people, namely, a *nation*. Thus, the nation-state became the particular, and normal, form of the modern state. It has been argued that if the basic framework of rights in the modern state was defined by the twin ideas of freedom and equality, the often contradictory pulls of these two ideas had to be mediated by two further concepts: those of property and community. Property sought to resolve the contradictions between freedom

and equality at the level of the individual in relation to other individuals. Community was where the contradictions were sought to be resolved at the level of the whole fraternity. Along the dimension of property, the particular resolutions might be more or less liberal; along the dimension of community, they might be more or less communitarian. But it was within the specific form of the sovereign nation-state that the universal ideals of modern citizenship were expected to be realized (Balibar, 1993).

It needs to be emphasized that just as the sovereign nation-state was a sign of the fulfillment of the universal moral values of modern citizenship, so also was it a concrete entity defined by the practical considerations of viable statehood. The former provided the ideological force in the nineteenth century to the revolutionary democratic movements in many regions of the Habsburg and the Ottoman empires as well as to the movements for German and Italian national unification. But the latter became operative when judging the claims of groups that were considered too small or territorially dispersed or politically immature to have viable nation-states. Thus, the Welsh, Sicilians, Bretons, the many small ethnic groups in the Balkans, or a European colony such as Ireland was considered unfit for independent statehood. This highlights the fact that no matter what the specific cultural content that defines nationhood in a particular nation-state, the requirement of undivided and viable territorial sovereignty will always produce national minorities within that state. It also points to another feature that developed through the nineteenth century to become a basic principle of the modern political world: sovereign statehood depended crucially on recognition from other states within an international states system.

The late nineteenth century was a period of the democratization of politics in the Western world, with the expansion of the suffrage and a wider sense of involvement of the working class in national politics, even if in a role of opposition to the ruling classes. Alongside, there also operated a process of producing national citizens, through a rapidly expanding system of schooling, the imposition of standardized forms of the national language, and not the least by inculcating a popular sense of patriotism by celebrating the achievements of its state leaders and armies. There is little doubt that at the turn of the century, popular nationalism in the United Kingdom, France, Germany, Russia, Italy, Japan, and the United States – regarded as the great powers – was deeply implicated in imperial ambitions, colonial expansion, and military glory.

Anticolonial Nationalism

The emancipatory appeal of modern nationalism lay in its association with the universalist moral claims of the modern state. But, as pointed out before, it was also necessarily conditioned by particular claims to viable statehood within an international states system. The anticolonial settler and creole nationalisms of North and South America in the late eighteenth and nineteenth centuries were, in the many ways, of the same type as the democratic nationalisms of nineteenth-century Europe in that the antiabsolutist element was far more predominant and unambiguous than the anticolonial. The interesting and tragic contrast was the failed Haitian revolution of the 1790s, launched by Black slaves and exslaves inspired by the emancipatory message of the French Revolution and crushed under orders from the revolutionary regime in Paris (Trouillot, 1995). The world was not yet ready for the real wave of anticolonial nationalism.

That would begin a 100 years later. Many of the apparatuses characteristic of the modern state, including its bureaucracies, legal and judicial systems, economic institutions, censuses, communication networks, etc., were introduced into European colonies in Asia, Africa, and the Caribbean as technologies of colonial governance. In some British and French colonies, this even included limited representative institutions in which native elites could participate. Equally significant was the introduction into these countries of modern education in European languages, which opened up for a new group of indigenous intellectuals the possibility of imagining a modern nationhood for their peoples, free from colonial bondage. Curiously, the emancipatory claims of modern statehood were used in the nineteenth and twentieth centuries by colonial apologists to justify Western rule over the rest of the world: those backward regions, they said, had to come under Western tutelage in order to introduce them to modernity and prepare them for responsible government. The nationalist movements of Asia and Africa, by contrast, argued that it was precisely colonial rule that stood in the way of their becoming modern and self-governing nations, fully responsible for their own destinies. Not only was self-government demanded in accordance with the universalist modern doctrine, as a natural right of every nation, but a new economic argument was also added to assert that modern colonialism was inevitably accompanied by the economic exploitation of the colony, and hence, political sovereignty was a necessary condition for the removal of poverty and economic backwardness.

By World War II, anticolonial nationalism had become a significant political force in many countries of South and Southeast Asia, the Middle East, and North Africa. While aspiring to fulfill the project of modernity in their own countries, anticolonial nationalist leaders also launched powerful movements of social and cultural reform, but outside the supervision of the colonial state, attempting to assert their sovereignty over an inner sphere of national life even before the political battle over the economy and the state had been won (Chatterjee, 1986, 1993). This was the period when standard national languages were institutionalized and national literature created and canonized, national histories constructed and archives assembled, and religious doctrines and social practices submitted to a thorough modernist scrutiny and reform. The political process of decolonization became rapid and unstoppable after the end of World War II. In many colonies of Asia, Africa, and the Caribbean, independent states were created through negotiated transfers of power; in others, national liberation was won after armed struggle. It was often argued that independence achieved by peaceful negotiation produced socially conservative and politically conformist regimes, whereas when it was won through armed struggle, the structures of colonial rule, themselves founded on violence, had a greater chance of being socially and politically transformed (Fanon, 1963).

Nation and Culture

One of the consequences of the new theoretical understanding of nationalism, introduced by Anderson, has been the explosion of scholarly interest in the deep involvement of nationalism in areas of culture that apparently have little to do with the political contest over nation-states. The histories of language, literature, historiography, art, music, theater, cinema, sports, and such other modern forms of cultural production are being examined for their many connections with nationalist projects – in their formal properties, in their changing contents, and in the institutions that promote and foster them. Alongside, there has emerged a new literature on family, gender, sexuality, religious practices, food, clothes, and other aspects of everyday life that can be shown to have been shaped and transformed by nationalist cultural projects. A recurrent theme in many of these studies is the active role of the nationalist imagination in the gendering of numerous institutions and practices of modern life. Thus, a common feature is the identification of the inner, the domestic, and the spiritual as the domain of the feminine and that of the outer, the public, and the material as that of the masculine. Another is the imagining of women as the bearers of national tradition. There are interesting variations in these structural relations of gender in the case of anticolonial nationalisms (Abu-Lughod, 1998; Altinay, 2004; Yuval-Davis, 1997).

The other aspect of the relation between culture and nationalism that has received much attention in recent years is that of ethnic politics or, in some descriptions, the politics of identity. A distinction has been made, most notably by Anderson (1998), between nationalism in its classical version and the politics of ethnicity. The former, he argues, imagines community in the form of 'unbound serialities,' i.e., the everyday universals of modern social thought such as nations, citizens, workers, bureaucrats, intellectuals, etc. This form of imagining of universal community is potentially liberating. Ethnic politics, on the other hand, is produced by the 'bound serialities' of governmental classifications, such as censuses. They operate by integers – one can be counted only as one or zero, for example, Black or non-Black, Muslim or non-Muslim, German or non-German, and all partial or mixed affiliation must be ruled out. Such serialities are constricting and inherently conflictual. Others argue that the supposed distinction between classical nationalism and ethnic politics is once again a version of the liberal distinction between good and bad nationalism and that they represent not so much a difference of type as the inadequacy of the classical nation-form to properly represent within itself the full range of aspirations for community in modern political life (Brubaker, 2004).

It has also been said that the apparent proliferation of the politics of identity, occurring in the era of economic globalization, is a symptom of the historical demise of the nation-state. The breaking down of trade barriers between nations, the global spread of the new communications technology, and the unprecedented flow of international migrants are making the sovereign nation-form largely redundant (Hardt and Negri, 2000). The European Union has been frequently cited as an example of a postnational formation superseding the constitutional and ideological limits of the sovereign nation-state (Habermas, 2001). This is accompanied by the increasing assertiveness of a global civil society that often considers the sovereignty of existing nation-states as impediments to the realization of universal ethical goals such as human rights. These conditions, some thinkers have argued, are producing the possibility of the emergence of cosmopolitan forms of government as the more acceptable receptacle of modern political community (Held, 1995; Appiah, 2006). However, many of these arguments produced at the turn of the millennium have now come under challenge with the global financial crisis that has led to a resurgence of autarkic economic pressures and chauvinistic movements against immigrants. Nationalism appears to be alive, if not well, in the twenty-first century.

See also: Colonialism, Anthropology of; Colonization and Colonialism, History of; Globalization and World Culture; Nation, Sociological Aspects of; National Character; National Liberation Movements; Nationalism, Sociology of; Nations and Nation-States in History; Religious Nationalism and the Secular State.

Bibliography

Abu-Lughod, L., 1998. Remaking Women: Feminism and Modernity in the Middle East. Princeton University Press, Princeton, NJ.
Altinay, A.G., 2004. Myth of the Military Nation: Militarism, Gender, and Education in Turkey. Palgrave Macmillan, New York.
Anderson, B., 1991. Imagined Communities, rev. ed. Verso, London.
Anderson, B., 1998. The Spectre of Comparisons. Verso, London.
Appiah, K.A., 2006. Cosmopolitanism: Ethics in a World of Strangers. W W Norton, New York.
Balibar, E., 1993. Masses, Classes, Ideas: Studies on Politics and Philosophy before and after Marx (J. Swenson, Trans.). Routledge, New York.
Brubaker, R., 2004. Ethnicity without Groups. Harvard University Press, Cambridge, MA.
Chatterjee, P., 1986. Nationalist Thought and the Colonial World: A Derivative Discourse? Zed Books, London.
Chatterjee, P., 1993. The Nation and Its Fragments: Colonial and Postcolonial Histories. Princeton University Press, Princeton, NJ.
Fanon, F., 1963. The Wretched of the Earth. Grove Press, New York.
Gellner, E., 1983. Nations and Nationalism. Blackwell, Oxford, UK.
Greenfeld, L., 1992. Nationalism: Five Roads to Modernity. Harvard University Press, Cambridge, UK.
Habermas, J., 2001. Postnational Constellation (M. Pensky, Trans.). MIT Press, Cambridge, MA.
Hardt, M., Negri, A., 2000. Empire. Harvard University Press, Cambridge, MA.
Held, D., 1995. Democracy and the Global Order: From the Modern State to Cosmopolitan Governance. Polity Press, Cambridge, UK.
Hobsbawm, E., 1990. Nations and Nationalism since 1780. Cambridge University Press, Cambridge, UK.
Kohn, H., 1944. The Idea of Nationalism. Macmillan, New York.
Nimni, E., 1991. Marxism and Nationalism. Pluto, London.
Smith, A., 1986. The Ethnic Origins of Nations. Blackwell, Oxford, UK.
Tamir, Y., 1993. Liberal Nationalism. Princeton University Press, Princeton, NJ.
Trouillot, M.-R., 1995. Silencing the Past: Power and the Production of History. Beacon Press, Boston, MA.
Yuval-Davis, N., 1997. Gender and Nation. Sage, Thousand Oaks, CA.

Nationalism, Historical Aspects of: Africa

Catherine Coquery-Vidrovitch, Université Paris 7 Denis-Diderot, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In Africa, nationalism had a tendency to be studied only in the context of modern independent states, by reference to the European paradigm. Nevertheless, as early as precolonial times, true small nation-states had existed, such as the kingdom of Abomey in today's southern Benin, as also the kingdom of Congo in Northern Angola. Another case was the emergence throughout the nineteenth century of large conquering states, Islamic or others, usually based on slave trading, against which preexisting subjugated nationalities reacted. Today, nationalism is often rooted in such a context. As early as the onset of colonialism, the elites among the subjects developed processes of moderate nationalism. In several cases, it turned into struggles for national liberation favoring the emergence of modern nationalism. But recent trends of tribalism reactivate regionalisms inherited from the past. Do they portend the new nationalism to come? In spite of many predictions, this danger appears less serious than a tendency for existing states to generate xenophobia inside their frontiers, even if these states in their current form have been in existence only for a century or so.

Most analyses of the phenomenon of nationalism in Africa were made recently, in the period after independence was achieved. Studies focus on nationalist processes comparable to those which have taken place in Europe, most of them starting in the nineteenth century. The analyses usually began with the era when emerging states started to demand international recognition, as other nations had done earlier. Only recently a few political scientists and historians have given a modest amount of attention to the long history of nationalist movements in Africa (Davidson, 1992; Berman and Lonsdale, 1992). Others have stressed the catalytic role played by the colonial phase (Ranger, 1968; Johnson, 1971; Mamdani, 1996). But Africanist literature had focused primarily on 'ethnic' problems, portraying them as quite different from – even antagonistic to – a 'nationalist' approach. Until recently, no investigator had pointed out that ethnicity can equally be a function of nationalism as much as nationalism was derived from ethnic awareness (Anderson, 1983; Berman and Lonsdale, 1992; Turner, 2000). The widely prevalent view must be considerably revised, because it is the offshoot of a long tradition of biases inherited from the colonial past of ethnographic research (Hobsbawm and Ranger, 1983; Amselle and Mbokolo, 1985; Coquery-Vidrovitch, 1995).

African political history began very early, several centuries before the colonial episode. Therefore, we designate endogenous nationalism antedating colonization as 'older nationalism.' We note that in precolonial days, what we could regard as legitimate small 'nation-states' did indeed exist, although they were not established on the same presuppositions as in the Western world. These preexisting elements were the foundation on which the process of 'modern' nationalism was, more or less convincingly, grafted in the form of independence as defined by international settlements and the United Nations.

With minor exceptions, the African continent has remained independent until very late in history, at least vis-à-vis European powers. Certain limited points on the coast were occupied at a very early date: the port of Loanda in the west and Mozambique Island in the east by the Portuguese; the port of Mombasa on the Kenyan coast, disputed among the Arabs and Portuguese from the sixteenth century; the Senegal island of Saint Louis at the mouth of the river by the French; and Cape Town by the Dutch, starting in the middle of the seventeenth century. A series of coastal outposts was taken in the seventeenth and eighteenth centuries by the British, including Cape Coast Castle on the Gold Coast and, above all, the small Crown colony created in 1807 at the tip of the Sierra Leone Peninsula to receive slaves 'liberated' from ships engaged in smuggling them. However, at least until the 1830s – when the conquest of Algeria by the French began and the occupation of the interior plateaux of southern Africa was commenced by the Boers, soon followed by the British on the coast of Natal (1844) – almost the total population of African peoples lived in independence, except for a conglomeration of east coast territories dominated by the Sultanate of Oman and Zanzibar.

Older Forms of Nationalism

In the past, we come across two forms of 'nationalism': first, the nationalism associated with the institution of actual small nation-states built over a long course of time around a well-defined plan, and the second, associated with the conquering or colonizing forces in the context of sudden emergence of subjugated states that hardly had the time to reorient and instead triggered revolts of preexisting nationalities.

These conquests of states resulted from a series of jihads (holy wars of Islam), which succeeded one another with no interruption in the northern hemisphere from the middle of the eighteenth to the last third of the nineteenth century. They did not arise from nationalism in the usual sense of the term, but they forged a strong cultural bond. This was established in western Africa around the Pular culture (Robinson, 1985). Fouta Djallon is a good example. In this mountainous region containing the sources of the Niger, Senegal, and Gambia rivers, the Muslim Fulbe (advocates of slavery, including shepherds, farmers, and scholars) gained the upper hand over local farmers and indigenous religions. Locally, the main innovation of the Fulbe was using the Arabic alphabet to

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62065-2

transcribe the literature of the Pular language. This led to the popularization of a culture based on educated people who had an awareness of solidarity. Pular migrations all over West Africa helped propagate their feeling of belonging to a common culture. The Pular community, reinforced by Usman dan Fodio in 1804 in northern Nigeria, was thus both an ideological and an economic venture. Its specificity has survived to the present day. The community marvelously managed to resolve the problem common to all Pular theocracies: how to convert a tolerated minority religion, only adopted by chiefs when it seemed to them to be useful, to a religious and imperial state ideology; how to turn a small, strict, geographically isolated Muslim group into a driving force to obtain recognition of western Sudan as Dar Al Islam and so on.

Precolonial Nation-States

Several animist or Muslim states were constituted as a nation much earlier. The oldest examples of these are ancient Egypt and, to a certain extent, Ethiopia from the beginning of the current era. North Africa was reconfigured by the Ottoman Empire: Morocco, Algeria, and Tunisia already had their borders long before the nineteenth century; the first as an independent country, the other two as Turkish provinces. Tunisia and Morocco became protectorates (in 1883 and 1912), Algeria a French *département*. Nationalism was molded in preexisting frameworks in conjunction with the anticolonial struggle, although it triggered reactions in the form of regionalisms established on linguistic and cultural claims. Berber speakers, vigorously repressed by speakers of Arabic in contemporary Algeria, are a case in point.

Outside Islam, the west coast of Africa was experiencing the proliferation of small nation-states at least since the sixteenth century – the term is not exaggerated, as Basil Davidson (1992) has demonstrated, *mutatis mutandis* – if we take this to mean political formations of people who have shared a common history and most often the same language for several generations, or even several centuries and, recognizing the same political system under the rule of a sovereign or chief acknowledged by all, and working to a greater or lesser extent under this authority in the same economic network of productive and commercial activities. Since the time of articulations by Benedict Anderson (1983), we do not hesitate to speak of the genesis of nationalism outside the Western world.

The emergence (toward the end of the seventeenth century), the rise (from 1760 to 1840), and the decline (in the second half of the nineteenth century) of small political organizations focused on slaving – coinciding chronologically with the parallel evolution of the Atlantic slave labor trade, which itself developed simultaneously with the expansion of sugar cane plantations from Brazil to the Caribbean – was no mere coincidence. The establishment of many political organisms correlates closely with the era's dominant international market, at least until the middle of the nineteenth century. The system combined war and trade. During each dry season, the army foraged for prisoners to supply objects of the slave trade. During the rainy season, the soldiers became peasants again, coming back home to toil their lands working in subsistence agriculture.

The kingdom of Abomey (in today's Benin, formerly Dahomey) established a strong political community with a common language and culture, and each generation of children of slaves born on Dahomean soil became Dahomean. This national unity explains why the conquest of Dahomey by the French (1890–94) was the toughest colonial war and the longest in the region. The people of Dahomey gave mass support to the resistance put up by the last king, Behanzin. By the same token, in the heart of present-day Ghana, the Asante Empire, annexed in 1897, did not accept British sovereignty until the Crown resolved, in 1928, to recall the Asantehene and reinstate the sovereign on his Golden Stool, the symbol of national unity.

We could go on and on citing cases of small nation-states which took shape from the eighteenth century onward (Coquery-Vidrovitch, 1999). Other kingdoms appeared during the same period in Central Africa, in the rift lakes region of Rwanda, Burundi, Ankole, and Buganda, where a long collective history had endowed the inhabitants with a common language and, consequently, a common culture. Nationalist excesses in the form of ethnic purges are merely a recent manifestation of a longstanding demographic and political disequilibrium aggravated by the traumas of colonial and postcolonial history.

At the turn of the nineteenth century, the small, militarized Zulu kingdom of Shaka appeared in the northern part of what is now Natal. This realm never submitted to British domination. The last revolt there erupted in 1906. Using this glorious past to justify their political demands paved the way for young militant Christians to found the nationalist Inkatha Party in the early 1920s.

Genesis of Modern Nationalisms

Militant nationalists of the early years of independence attuned their brand of emerging nationalism to the numerous revolts which had developed nearly everywhere during the colonial period. Nevertheless, we should not confuse what have been called 'primary resistance movements' with modern nationalism (Ranger, 1968). Ranger postulated that, slowly but surely, inhabitants of a territory delimited by a colonial power will accept this entity and claim possession of it. By now this 'nationalization' has taken place almost everywhere, despite often poorly understood reactionary movements. A contrasting example is Eritrea, which won its independence for the very reason that Eritrean nationalism has been constructed for a century outside Ethiopia. Ethiopia triumphed in 1896 over the Italian invader, while the province of Eritrea became an Italian colony in 1890, then, as a result of World War II, a British protectorate from 1940 to 1954. The Eritreans never recognized their territory's reversion to Ethiopia, arranged by Western negotiations in 1960.

In sub-Saharan Africa, nationalism engendered by colonialization did not always work out where preexisting nation-states were prevalent. An example is Cameroon, well studied by Mbembe. Another is the Fante nationalism, which developed in the second half of the nineteenth century on the Ghanian coast. Fante communities which had previously existed constituted a sprinkling of small districts, united solely

by language and their occupations because they had long been agents of European merchants on the coast and on privileged terms with the English. Exposed to Western influences more than the rest of the population, they had created a Christianized Creole culture, which made them early adherents of political modernism and aligned them against European intervention from the start. The movement was triggered by British interference starting in 1843. The chiefs, dispersed until then, reacted in 1852 by sanctioning an internal confederation inspired by Western ideas. The Fante Confederation (1867–73) provided for a constitutional regime, with a national assembly of chiefs and an executive parliament. These ideas were propagated by the publication of *West African Countries and People* by the Sierra Leonean physician James Africanus Horton. From 1865 on, to the dismay of the British, Horton advocated the idea of African independence, arguing that the country was more advanced than Liberia 20 years earlier.

A Moderate Colonial Protonationalism

Beginning at the end of the nineteenth century, new economic and social conditions fostered by the conquest provoked an anticolonial reaction of the 'evolved' victims of colonialization. These protests had the paradoxical effect of reinforcing their territorial cohesion within borders which were frequently conceived and imposed entirely by outsiders.

The initial militant nationalists were graduates of missionary schools. The first of these in Sierra Leone were descendants of freed slaves, subsequently known as Krio, which describes both a language and a culture. Others were natives of the Ghanian coast and the mouth of the Niger, often scions of historically dominant classes. A prime example is the Brew clan, established by an eighteenth-century Irish slave trader. His African descendants devoted themselves to trading – first to the slave trade and, when that became impossible, to export commodities. In the second half of the nineteenth century, when British competition was making grave inroads thanks to colonialization, these cultivated bourgeois still sent their children to study in England, or at least to missionary schools on the coast. They bred lawyers and journalists. A dozen of these intellectuals earned PhDs. One Brew was a member of the delegation the Fante sent to Queen Victoria in 1895, when they still believed it was possible to set up a less closed political system than colonial law and order (Priestley, 1969). The next year, three Tswana chiefs traveled from the other end of the continent to London to try to persuade the queen to help their country preserve its independence.

South Africa

African nationalism manifested itself most conspicuously in South Africa. We have already mentioned Zulu regionalism. Another important point is the length of the so-called Xhosa Wars, expressions of resistance which flared up periodically from the end of the eighteenth century to the end of the nineteenth century, reflecting the maturation of a Xhosa awareness among peoples inhabiting Eastern Cape Province. But it was the injection of mining capital (diamonds were discovered in 1867, gold in 1886) that elicited modern forms of black nationalism in the Rand compounds and in Johannesburg. The African National Congress (ANC), ever since the main protagonist of South African nationalism despite several periods of severe repression, was established in 1912. From the start, militant South African nationalists were radicals and often had close clandestine ties to the Communist Party, but most of them were graduates of institutions of higher learning created from the middle of the nineteenth century by various Christian missions (Lovedale, Healdtown, St. Matthew). The first militant feminists appeared on the scene in South Africa and West Africa even before the 1950s. They included Ray Alexander (1941), Ida Mtwana, who assumed leadership of the Women's League of the ANC in 1947, and Liz Abraham, who is renowned for her struggle against apartheid (Walker, 1982).

Evolution From the Years between the World Wars until Decolonialization

It was not until the twentieth century, between the two world wars, that nationalist demands really began to be voiced. This is because severe censorship was never loosened by the French, except by the Front Populaire government (1936). In British Africa the turning point was not until the World War II, when trade unions were first permitted, even before noncolonial political parties. Although its way of life and thinking was more akin to European intellectuals than to the people in local rural communities, the young African bourgeoisie did not wait until the grant of civil rights before demanding, sometimes vigorously, that it be allowed to participate in public affairs. The spearhead of the movement was British West Africa, first centered on Sierra Leone, where an indigenous press developed at an early stage. In the French colonies, political life went on in classical French style in the 'Four Communes' of Senegal (Saint-Louis, Gorée, Rufisque and Dakar), whose inhabitants became French citizens in 1916 (Johnson, 1971). The colony of Dahomey was also a hotbed of protest. There, a merchant bourgeoisie developed alongside the low-ranking civil servants with a French education. It had connections to prominent Afro-Brazilian families and to aristocratic circles of the former kingdoms of Abomey and Porto Novo, and it possessed sufficient financial resources to launch local publications. *La voix du Dahomey*, a newspaper established in 1928 and existing into the 1950s, was its principal mouthpiece. The colonial government was unable to ban the paper, despite a highly publicized trial during the Front Populaire administration.

Nationalists expressed a desire to take the reins of government into their own hands in British-ruled West Africa at an early date. During the 1920s, they obtained the right to participate in municipal government to the extent of voting in local elections. In French colonies, nationalist aspirations were offset for a long time by the French model of assimilation tried out in Senegal: acquiring the citizenship of the colonial power was a way of obtaining rights, based on equality and individual freedom, i.e., everything normally denied to natives, even assimilated groups. In both cases, the political ideal remained reformist. It did not challenge the colonial system. The idea was to fight colonial abuses and to collaborate with the system. This collaboration became a fact after the World War II. Virtually

alone, Dr Danquah demanded the status of a dominion at the Gold Coast Youth Conference in 1940. His platform was still moderate. He advocated the democratization of the legislative council, which then only had nine Africans among the 29 elected members. He proposed a two-chamber system modeled after the British Parliament, one elected and the other composed of chiefs, and he demanded that Africans be granted the right to apply for civil service jobs.

Models, institutions, and trappings of the colonial government fascinated the future leaders of the newly independent states for a long time (Werbner, 1998). Anticolonial radicalism was limited to a very small minority of young Communists, whose militancy had little chance to express itself in Africa. The initial impulse came from George Padmore, a Marxist from Trinidad, who established the Negro Workers Bureau in Hamburg and contacted Lagos militants, undoubtedly thanks to dockers in the port. It was virtually impossible in French Africa and not much easier in Sierra Leone, where the most radical militant was the journalist Wallace Johnson, who established the Sierra Leone Youth League on the eve of the war.

The climate changed only slowly after the World War II. In the French sphere, the nationalist ideas of the elite were sublimated under the cult of Negritude, which was less of a political movement than an ethical ideal. This concept, launched in 1939 by the Antilles poet Aimé Césaire, later author of the explicit indictment *Discours sur le colonialisme* (1955), was developed for the most part after 1947 by the Senegalese poet and politician Léopold Sédar Senghor. In British Africa, intellectuals sued for divorce earlier. They were, nevertheless, likewise steeped in Western culture (the Ghanian Kwame Nkrumah and the Nigerian Azikiwe, future statesmen, had left on the eve of the war to study in the United States, and in 1938 the Kenyan Jomo Kenyata wrote the first thesis in anthropology on his country). But since in the 1950s the British were not yet prepared to budge, the nationalists hardly had any choice but to accept 'multiracialism': one 'race,' one vote, or in Kenya the same number of seats for each of three groups – Whites, Indians, and Blacks – which were nowhere near equal in number.

The Role of National Liberation Movements

Much has been written about the Mau Mau rebellion, the 1952–55 brush fire that swept most of Kenya. It has been argued that the movement was primarily a Kikuyu affair. But the severity of the repression and the blindness of the colonial authorities, whose minds were set in the belief that opposition politicians were the main instigators, contributed to cementing a national, anticolonial sentiment which forced Britain, although victorious on the field of battle, to grant independence (1963). Berman and Lonsdale have written a superb work on the subject (1992).

The anticolonial struggle also formed a national awareness in former Portuguese colonies. After independence (1975), resistance against apartheid South Africa and white Rhodesia sufficed to forge a nation state in Mozambique, a more than 3000-km-long territory which was very disparate at the outset. However, in Angola, the southern Savimbi regionalism was constructed on a partially real, partially manipulated precolonial Ovambo past. This is the case with many contemporary tribalist manifestations; they are exploited primarily for political ends by ambitious, unscrupulous manipulators. For the past 25 years or so it has been demonstrated that one should not lump ethnic group and nation together (Hobsbawm and Ranger, 1983; Amselle and Mbokolo, 1985). But contemporary propaganda resuscitates strong sentiments handed down from the past – for example, the resentment of ravaged peoples against the nations which ravaged them, going back to the black slave trade, as was the case in Angola, or that of social groups dominated and exploited by large landowners of the period, as in the case of the Hutu and Tutsi in Rwanda.

Do these regionalisms of today portend tomorrow's nationalisms? This is unlikely, because in most of the examples we have cited all sides claim the entirety of the country involved. Once again, the struggle against foreign power sparks nationalism. A typical case is the Congo (RDC, former Zaire), where international observers forecast an inevitable explosion. Ironically, it was the widely hated dictator Mobutu who sanctified the national unity of the former Belgian Congo in the name of the national hero and founding father Patrice Lumumba (Turner, 2000).

The risks of regionalist explosions seem to be less of a threat today than do conflicts arising from a chauvinistic approach to nation-states that have become established despite arbitrary delineation imposed more than a century ago (Bayart, 1996; GEMDEV, 1996). The recent Ivory Coast conflict based on an exclusive, xenophobic *Ivoirité* is one of the most ominous signs.

Nevertheless, the partition of Sudan (2011) possibly announces a new era: for the first time, former colonial frontiers reinforced by the Organization of African Unity (OAU) chart (1963) were forgotten, rejected to the past in the name of a general worldwide tendency aiming at regenerating local historical and cultural regionalism.

See also: Africa, Sociocultural Overviews: West Africa; African Studies: Culture; African Studies: History; Central Africa: Sociocultural Aspects; Colonization and Colonialism, History of; East Africa: Sociocultural Aspects; National Liberation Movements; Nationalism, Historical Aspects of: Arab World; Nationalism, Historical Aspects of: The West; Nationalism, Sociology of; Postcoloniality; Southern Africa: Sociocultural Aspects.

Bibliography

Amselle, J.-L., Mbokolo, E., 1985. Au Coeur de l'Ethnie. Decouvert, Paris.
Anderson, B., 1983. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London.
Bayart, J.-F., 1989. L'État en Afrique: La Politique du Ventre. Fayard, Paris. Trans. 1993. The State in Africa: The Politics of the Belly. Longman, London.
Bayart, J.-F. (Ed.), 1996. La Greffe de l'tat. Karthala, Paris.
Berman, B., Lonsdale, J., 1992. Unhappy Valley: Conflict in Kenya; 1. State and Class; 2. Violence and Ethnicity. James Currey, London.
Coquery-Vidrovitch, C., 1995. De la Nation en Afrique Noire. Le Débat 84, 121–138.
Coquery-Vidrovitch, C., 1999. L'Afrique et les Africains au 19è Siècle. Colin, Paris.
Coquery-Vidrovitch, C., 2003. Nation without a state, and state without a nation: the case of Africa south of the Sahara. In: Teichova, A., Matis (Eds.), Nation, State and the Economy in History. Cambridge University Press, Cambridge, pp. 239–250.

Davidson, B., 1992. The Black Man's Burden: Africa and the Curse of the Nation-State. Times Books, New York.
GEMDEV, (Ed.), 1996. Les Avatars de l'État en Afrique. Karthala, Paris.
Hobsbawm, E., Ranger, T., 1983. The Invention of Tradition. Cambridge University Press, Cambridge, UK.
Horton, J.A., 1868. West African Countries and People London.
Johnson, G.W., 1971. The Emergence of Black Politics in Senegal: The Struggle for Power in the Four Communes, 1900–1920. Stanford University Press, Stanford, CA.
Mamdani, M., 1996. Citizen and Subject: Contemporary Africa and the Legacy of Late Colonialism. Princeton University Press, Princeton, NJ.
Mbembe, A., 2000. De la postcolonie. Essai sur l'imagination politique dans l'Afrique contemporaine. Paris.
Priestley, M., 1969. West African Trade and Coast Society: A Family Study. Oxford University Press, London.
Ranger, T.O., 1968. Connections between primary resistance movements and modern mass nationalism. The Journal of African History 9 (3), 437–454, 9 (4), 631–642.
Robinson, D., 1985. The Holy War of Umar Tal: The Western Sudan in the Mid-Nineteenth Century. Clarendon Press, Oxford, UK.
Turner, T., 2000. Ethnogénèse et Nationalisme en Afrique Centrale: Aux Racines de Patrice Lumumba. L'Harmattan, Paris.
Walker, C., 1982. Women and Resistance in South Africa. Onyx Press, Cambridge, MA.
Werbner, R., 1998. Memory and the Postcolony: African Anthropology and the Critique of Power. Zed Books, London.

Nationalism, Historical Aspects of: Arab World

Gudrun Krämer, Free University of Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 15, pp. 10344–10348, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Throughout the twentieth century, Arab nationalism was one of the major political currents in the Arab world, serving as a powerful source of anticolonial struggle and sentiment. The movement of Arab cultural revival (*nahda*) of the late nineteenth century, Arab nationalism of the interwar period and, after World War II, pan-Arabism aiming at Arab unity reflected various understandings of Arab identity, interest, and solidarity. In spite of its secular image, Arab nationalism was marked strongly by (*Sunni*) Muslim references; in spite of the failure of political union, it carried its mobilizing force into the twenty-first century.

For almost a century, Arab nationalism was one of the major political currents in the Arab world; in the 1950s and 1960s, it was undisputedly the dominant one, albeit never uncontested. Like in other parts of what was then still called the Third World, its intellectual origins in Europe did not hinder its spread and success as a principal source of anti-imperialism. Like other nationalisms, it covered a wide range of ideas, images, and aspirations that appealed to and were supported by different social groups and strata; it built on existing self-views and affinities, forming various blends of local, regional, and transnational allegiances, reflecting different shades of religious and secular orientation. Critical scholarship has absorbed key assumptions of the theoretical literature on nationalism and ethnicity more generally, which can be read as a refutation of, or counter-discourse to, nationalist claims and convictions: it stresses the constructedness of Arab like any other identity or identification; its hybrid character rather than the purity of race, culture, language, and design nationalists insist on; fragmentation rather than the unity of intent, action, and objective nationalists assert; fuzziness rather than the clarity of program, boundaries, and orientation they proclaim (Jankowski and Gershoni, 1997; Khalidi et al., 1991; Gelvin, 1998). Arab nationalism thus emerges as multifocal, multifaceted, and multilayered. Measured against the level of theoretical sophistication, knowledge of the historical evolution of Arab nationalism remains highly uneven, and in several respects quite inadequate.

Contingency and Continuity

There was no unilinear movement, or organic growth, from cultural revival (Arabism) to Arab nationalism first in greater Syria and then the remainder of the Arab world, to pan-Arabism and Arab unity. From the beginning, Arab activists interacted, and competed, with advocates of alternative frames of reference that emerged simultaneously in the Middle East and North Africa: if Ottomanism died with the collapse of the Empire, territorial nationalism and political Islam found equally strong resonance among the area's population. While in the Middle East, as in most other parts of the world, there existed a tradition of local pride and attachment that was also reflected in a specific literary genre (*fada'il*) listing the beauties, monuments, and personalities of a specific town or region (*watan*, comparable to the German notion of *Heimat*), this sense of belonging did not necessarily translate into patriotism or nationalism, be it defined in terms of territory (*wataniyya*) or a people/nation (*qawmiyya*). Arabism never constituted more than one option of cultural and political orientation. To insist on choice and contingency, however, cannot imply denying all kinds of continuity between Arabism, Arab nationalism, and pan-Arabism.

Arabism

The intellectual and social roots of Arabism can be traced back to the second half of the nineteenth century, when in the Arab provinces of the Ottoman Empire, more particularly in certain cities of Egypt and the Fertile Crescent, small circles of scholars and intellectuals began to call for a revival or renaissance (*nahda*) of the Arab language, literature, and culture. The Nahda, which in later nationalist writings was represented as an 'awakening' of the Arab nation (*al-umma al-arabiyya*), or of 'Arab consciousness' (*al–wa'i al–arabi*), involved religious scholars, writers, journalists, government officials, and members of the liberal professions, reflecting the spread of education, literacy, and new means of communication, first and foremost an Arab press, that resulted in the creation of a public sphere ranging from clubs and associations of all kinds to salons, tea and coffee houses, theaters, etc. The revival of Arab(-Islamic) culture and literature (in written standard Arabic, *fusha*, not the vernacular used by the vast majority of Arabic-speakers) was to be combined with modern knowledge coming from Europe. While new literary forms gradually developed (novels, theater plays, and later films), poetry retained its central role for cultural and political expression in the Arabic language. The long-held view (e.g., Antonius, 1938) that Arab cultural revival was largely the work of Arab Christians who, through bonds of faith and their education in European schools had been more receptive to European thought including nationalism as it was embraced by their coreligionists in the European provinces of the Ottoman

Empire, has since been challenged and revised (Khalidi et al., 1991; Tauber, 1993; Nafi, 1998). The prominent role of Christians in the early Arab press cannot be denied. However, education on modern lines was not restricted to foreign schools, but equally adopted at Ottoman state schools and private institutions that were established by members of the local elites, Muslims, Christians, and Jews. Most important, Arab cultural revival was at that stage intimately linked to Islamic reform and maintained its link to Islam even later. A second view has been modified (Kayali, 1997): that Arabism evolved as a reaction to a Turkification drive pursued by the Young Turks after their 'revolution' of 1908, resulting in the marginalization of the Arabic language in official use and of Arab delegates in Ottoman representative bodies. The use of Turkish in Ottoman courts, schools, and government offices had been made obligatory since the 1870s. The Young Turks merely accelerated and intensified the centralization policies of the Ottoman reform era (Tanzimat), and as a result Turkish was more widely used in the Arab provinces of the Empire. At the same time, Arabs became less well represented in parliament, the higher echelons of the bureaucracy, and at court. With the decline of a legal and educational system that was based almost exclusively on Islamic learning, whose normative texts were in Arabic, knowledge of the 'sacred language' lost much of its prestige and usefulness for a career in the Ottoman public service. Modernization, then, contributed in more than one way to the rise of Arabism and, later, of Arab nationalism.

Arab political societies that from the 1880s onwards, formed and met in Damascus, Baghdad, Beirut, Istanbul, or Paris mostly worked for greater autonomy within the Ottoman Empire, for constitutional reform or decentralization, not for political independence (Tauber, 1993; Nafi, 1998). Geographically, their horizon was limited to 'greater' or 'natural Syria' (*bilad al-Sham*); Egypt with its distinctive tradition and emerging national movement, Sudan, the Maghreb, and the Arab Peninsula, Hijaz excepted, were not part of their vision. Membership in secret societies such as Al-Ahd (The Covenant, 1909), Al-Fatat (The Young, *c.*1911) or the Ottoman Administrative Decentralization Party (1912) was small; an Arab Congress held in Paris in 1913 was attended by some 25 delegates. Similar to Iranian, Turkish, or Russian opposition groups of the same period, they were made up of urban intellectuals and military officers. In contrast to the literary clubs and associations that also included some women, or addressed themselves to a female audience, they appear to have been exclusively male. Their impact on local society and politics was slight. Still, the secrecy surrounding them and rumors of broad popular support alerted official circles in Istanbul at a time when nationalist feeling was rising not only in the Balkans and among other Christian subject populations like the Armenians, but also in predominantly Muslim territories such as Egypt.

Arab Nationalism

During World War I, the vast majority of Ottoman subjects remained loyal to the Empire. Separatism remained a minority trend among its Muslim population: in 1915, Arab activists in Damascus contacted the Sharif of Mecca, Hussein ibn Ali, a descendant of the Prophet Muhammad in the Hashemite line and *de facto* of the holy city, who, after some hesitation, signaled his willingness to rebel, in the name of Islam, against the sultan-caliph. In a separate move, Hussein obtained British declarations of support for Arab independence under Hashemite leadership in exchange for military revolt. The Ottoman authorities reacted harshly to all stirrings of autonomous political activity including Arabism, which under the circumstances amounted to high treason, and in 1915/16 hanged dozens of activists, creating the first martyrs to the Arab cause. In June 1916, Hussein declared the Arab Revolt and after the fall of Medina in November 1916, proclaimed himself 'king of the Arab lands.' In the course of the years 1917/18, British, Arab, and Jewish forces jointly conquered Palestine and Syria. In October 1918, an 'Arab government' was installed in Damascus under Faisal ibn al-Hussein, one of the Sharif's sons, who with a group of Arab nationalists of diverse origins, attempted to take control of a country ravaged by war, famine, and disease. In March 1920, a Syrian General Congress representing Arab nationalists from Syria, Palestine, and Lebanon proclaimed the independence of 'greater Syria' with Faisal as king. Within Syria proper, their policies were partly assisted and partly resisted by urban, tribal, and village-based popular committees defending their own understandings of collective interest (Gelvin, 1998).

Britain and France were in the meantime taking measures to implement their different wartime arrangements, most of them secret, which included not only the correspondence with the Sharif, but also the Sykes–Picot agreement of May 1916 regarding their respective spheres of interest in the Arab provinces of the Ottoman Empire, and the Balfour Declaration of November 1917 promising British support for the establishment 'in Palestine' of a 'national home' for the Jewish people. The terms of these commitments had somehow to be reconciled; Britain and France did so at the expense of their Arab allies. As a result of World War I, more Arab lands were under European control than ever before, be it as provinces (Algeria), colonies (Aden), protectorates (Morocco, Tunisia, Egypt, Kuwait), or mandates (Lebanon, Syria, Palestine, Transjordan, Iraq) under the newly created League of Nations. Only the Zaidi Imamate of Yemen and the Saudi kingdom in Najd and Hijaz gained or retained full independence. The war and its aftermath defined some of the themes that were gradually woven into the 'grand narration' of Arab nationalism: the crucial role of 'martyrs' for national liberation; the 'Great Arab Revolt' or 'revolution' (*thawra*) of 1916–18 (to be followed by similar 'revolutions' in Egypt 1919, Iraq 1920, Syria 1925–27, the Egyptian officer coup of July 1952, as well as a series of 'corrective revolutions' in the 1970s and 1980s); exposure to 'conspiracies' and intimately linked to it, 'betrayal' of the Arabs by the West.

Within the framework of the emerging Arab state system, Arabism evolved into a political movement of growing impact and complexity, providing one of the major sources of anticolonial struggle and sentiment. The interwar period witnessed the rise of new social and political forces alongside the urban notables that still dominated politics and society (at least outside tribal areas): new urban middle-class parties that could not be considered mere covers for elite families or factions; paramilitary youth movements inspired by the fascist model; Islamic movements of social and political reform; socialist and communist groups, some with links to trade unions, of various

backgrounds and affiliations – all of them responding to Arab references, inflecting their emphases, adapting or rejecting them. If there was a hegemonic formulation of (politicized) Arabism (*al-uruba*) it was provided by *Sati al-Husry* (1880–1968) who as a high government official was able to influence public education in Iraq (then still under British mandate) (Cleveland, 1971; Tibi, 1981). Visibly influenced by the German romantic school, and Herder in particular, Husry defined the Arab nation on the basis of language, history, and destiny, embedding his vision in a narrative of pristine glory, foreign-machinated decline and eventual awakening, with a heavy emphasis on the us/them dichotomy, 'them' being the Arabs' enemies and oppressors from the Byzantines and Crusaders to the Ottoman Turks, Zionist Jews, and European colonial powers. The criteria of inclusion and exclusion remained malleable and contested: Arabness could be interpreted in secular terms to include all men and women of Arabic language and culture regardless of their religious or ethnic background, and thus provide the basis of common nationhood and citizenship; but it was impregnated so strongly with references of a specifically Sunni Muslim nature that it could also be given a distinctly religious (Sunni Muslim) coloring. At the same time it contained enough references to the superiority of the Arab race and genius to make it possible for Kurds, Berbers, or Jews to be marginalized, if not entirely excluded, from the national community.

Pan-Arabism

After World War II, two dimensions were added to nationalist discourse and strategy: Arab socialism addressing the issue of how to organize the growing number of independent Arab states, and pan-Arabism (*al-qawmiyya al-arabiyya*) aiming to transcend the existing regional order. Previous projects of political union had been associated mostly with members of the Hashemite dynasty, who, after losing Syria to the French and Mecca, Medina, and the rest of Hijaz to the Saudis, reigned in Iraq and Transjordan; and they had been restricted to the Fertile Crescent and Arabia (Porath, 1986). The 1930s saw the gradual diffusion of Arab identification to other parts of the Middle East and North Africa. The conflict over Palestine, which in 1936–39 erupted in the Arab revolt or 'revolution,' contributed to widening the sense of a shared Arab destiny. Intellectuals, scholars, and politicians in Egypt, Sudan, and the Maghreb who had not defined themselves previously in terms of Arabness, began to increasingly add an Arab dimension to their agenda, or even adopt pan-Arabism as their ideology. In the late 1940s, the Arab Socialist Ba'ath Party presented the new and enlarged credo of Arab nationalism, revolving around three concerns: unity (*al-wahda*) of the Arab nation both internally and across the existing borders that had been 'artificially' created by the colonial powers and had to be eliminated to create one single Arab state stretching from the Indian to the Atlantic Ocean (*min al-nahr ila l-bahr*); freedom (*al-hurriyya*) from foreign rule and reactionary feudal oppression at home; (Arab) socialism (*al-ishtirakiyya*) which was to be distinguished from Marxism or communism by its belief in God and the rejection of class struggle (Abu Jaber, 1966). The Arab League founded, with British support, in 1945, could at best offer partial fulfillment of this vision: it worked for political, economic, and cultural cooperation among existing Arab states, not for their unification, and it did not adopt Arab socialism (Maddy-Weitzmann, 1991).

In the course of the 1950s, Syria, which until then considered itself the 'beating heart of Arabism,' was overshadowed by Nasserist Egypt as self-proclaimed leader of the Arab nations. The call for Arab unity and solidarity evoked extraordinary responses from broad sections of the population all across the Arab world (the Arab 'masses' constantly evoked in Arab nationalist rhetoric), testifying to the power nationalism, and Arab nationalism in particular, had acquired regardless of its foreign origins. Appeals to Arab solidarity had the obvious function of legitimizing, inter alia, a 'fair share' in 'Arab' oil wealth. The politics of Arab unity equally bespoke the will to control and interfere, in the name of shared Arab concerns, in what rivaling national elites viewed as their domestic affairs. The heyday of pan-Arabism (*c.*1956–67) was also a time of relentless competition among Arab states and intrastate actors (Kerr, 1971). All experiments of merger and fusion (some of them, like Iraq's incorporation of Kuwait in 1990, based on military conquest) failed, most notably the union between the two champions of pan-Arabism, Egypt and Syria, that only lasted for 3 years (1958–61). The failure to solve the Palestine issue, epitomized in the Arab defeat of 1967, was therefore not the only factor weakening pan-Arabism as a political movement of mass mobilization. Yet while, from the 1970s, the 'end of pan-Arabism' was proclaimed repeatedly (Ajami, 1981), not least by the adherents of political Islam and many foreign analysts, reference to Arab identity, interest, and solidarity continued to play a significant role in Middle Eastern politics even at the beginning of the twenty-first century and beyond.

See also: Islam: Middle East; Islamic Fundamentalism; Middle East, North Africa: Sociocultural Overview; Nationalism, Sociology of; Nationalism: General; Near Middle East/North African Culture Private; Near Middle East/North African Studies: Politics; Religion, Nationalism, and Identity.

Bibliography

Abu Jaber, K.S., 1966. The Arab Socialist Ba'th Party. Syracuse University Press, Syracuse, NY.

Ajami, F., 1981. The Arab Predicament. Arab Political Thought and Practice since 1967. Cambridge University Press, Cambridge, UK.

Antonius, G., 1938. The Arab Awakening. Librairie du Liban, Beirut, Lebanon.

Alnasrawi, A., 1991. Arab Nationalism, Oil, and the Political Economy of Dependency. Greenwood Press, New York.

Ashton, N.J., 1996. Eisenhower, Macmillan, and the Problem of Nasser: Anglo-American Relations and Arab Nationalism, 1955–59. St. Martin's Press, New York.

Ayyād, A.-A., 1999. Arab Nationalism and the Palestinians, 1850–1939. PASSIA, Jerusalem.

Cleveland, W.L., 1971. The Making of an Arab Nationalist: Ottomanism and Arabism in the Life and Thought of Sati'al-Husri. Princeton University Press, Princeton, NJ.

Choueiri, Y.M., 2000. Arab Nationalism – a History: Nation and State in the Arab World. Blackwell Pub, Oxford.

Gelvin, J.L., 1998. Divided Loyalties, Nationalism and Mass Politics in Syria at the Close of the Empire. University of California Press, Berkeley, CA.

Hourani, A.H., 1962. Arabic Thought in the Liberal Age 1789–1939. Oxford University Press, Oxford, UK.

Hopwood, D., 2000. Arab Nation, Arab Nationalism. Macmillan Press, Houndmills, Basingstoke, Hampshire.

Jankowski, J., Gershoni, I. (Eds.), 1997. Rethinking Nationalism in the Arab Middle East. Columbia University Press, New York.
Kayali, H., 1997. Arabs and Young Turks. University of California Press, Berkeley, CA.
Kerr, M.H., 1971. The Arab Cold War. Gamal 'Abd Al-Nasir and His Rivals, 1958–1970, third ed. Oxford University Press, Oxford, UK.
Khalidi, R., Anderson, L., Muslih, M., Simon, K.S. (Eds.), 1991. The Origins of Arab Nationalism. Columbia University Press, New York.
Khoury, P.S., 1987. Syria and the French Mandate: The Politics of Arab Nationalism, 1920–1945. Princeton University Press, Princeton, NJ.
Khashan, H., 2000. Arabs at the Crossroads: Political Identity and Nationalism. University Press of Florida, Gainesville.
Khalidi, R., 1991. The Origins of Arab Nationalism. Columbia University Press, New York.
Maddy-Weitzmann, B., 1991. The Crystallization of the Arab State System: 1945–1954, first ed. Cambridge University Press, Cambridge, UK.
Meskell, L., 1998. Archaeology under Fire: Nationalism, Politics and Heritage in the Eastern Mediterranean and Middle East. Routledge, London.
Nafi, B.M., 1998. Arabism, Islamism and the Palestine Question 1908–1948. Ithaca Press, London.
Porath, Y., 1986. In Search of Arab Unity, 1930–1945. Cass, London.
Saunders, B.F., 1996. The United States and Arab Nationalism: The Syrian Case, 1953–1960. Praeger, Westport, CT.
Tauber, E., 1993. The Emergence of the Arab Movements. Cass, London.
Tibi, B., 1981. Arab Nationalism: A Critical Enquiry. St Martin's Press, New York.
Tibi, B., Farouk-Sluglett, M., Sluglett, P., 1997. Arab Nationalism: Between Islam and the Nation-state. St. Martin's Press, New York.

Nationalism, Historical Aspects of: East Asia

Yujiro Murata, University of Tokyo, Tokyo, Japan

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 15, pp. 10348–10351, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Nationalism in East Asia challenged a broad international system known as the Chinese World Order, and modern Japan played the most important role in dissolving this system. In particular, the Sino–Japanese War of 1894–95 and Russo–Japanese War of 1904–05 marked a decisive turning point for the formation of national self-identity in this area. At the same time, a new discourse of Asianism emerged in Japan that had the goal of opposing Western imperialism against all Asian nations by uniting the Asian nations under Japanese leadership. In spite of the lofty principle of Asian cooperation behind the idea, proponents of this concept were never able to take on a point of view critical of Japan's increasing aggression and imperialism. It was not until Japan's defeat in the war in 1945 that these discourses justifying Japanese imperialism virtually disappeared.

Nationalism in East Asia challenged a broad international system known as the Chinese World Order. In particular, nationalism in modern Japan played the most important role in dissolving this system. The concept of the Chinese World Order needs to be unpacked in order to explain this process and the role of Japan in challenging it, which led in turn to the establishment of national liberation movements across East Asia.

The Chinese World Order

The scholar J.K. Fairbank and others (1968) have noted that the Chinese World Order of premodern East Asia was a far-reaching international system characterized by the existence of an emperor considered to have been designated by heaven because of his innate virtue. He was referred to as a 'son of Heaven' and had the 'Mandate of Heaven' to rule. The emperor was considered to rule the world (*tianxia*, 'all under heaven') as a 'child of Heaven,' and his subjects ostensibly enjoyed the benefits of benevolent rule reminiscent of the 'Golden Age' of Chinese civilization described in ancient sacred texts. In reality, the power of the Chinese emperor had a limited reach, and there were a number of foreign tribes on the outskirts of the clearly defined spheres of Chinese civilization. Under the Chinese World Order, however, all subjects were ideally expected to be protected by the emperor, regardless of differences in their race, ethnicity, place of birth, or language – similar to the corpus Christianum of the Middle Ages in Europe.

One basic principle of the Chinese World Order was that it attempted to enfold these foreign tribes in its system and stave off invasion through tributary missions and other rituals. Thus, many foreign rulers from surrounding areas shared and benefited from the common idea of a Chinese World Order. They used the system to their own advantage by sending tributary missions to Chinese dynasties which in turn gave them authority and protection in their own areas. Chinese emperors entered into feudalistic relationships with these rulers under this tributary system, providing them with royal titles, official seal, and a calendar system in return for their loyalty. Foreign rulers, in turn, benefited by securing long-lasting political relations with Chinese dynasties through tributary missions. In addition, the tributary system played a significant role in functioning as a broad network of exchange of people, commodities, and information among different regions (Hamashita, 1990).

There has been academic dispute over when and how the Chinese World Order was developed and maintained. Some argue that it can be dated as far back as the 'Spring and Autumn' and 'Warring States' periods (770–221 BC), while others claim that the prototype of the system can be identified in the age of the Sui-Tang dynasties (581–907), with refinement of the system being completed as late as the Ming-Qing period (1368–1911). These theories notwithstanding, certain characteristics are indisputable: the Chinese World Order constituted one of the primary ways for China to regard its world, and it had a great impact in connecting many disparate and broad regions through their common use of the Chinese written language.

The China-Centered View of China's Neighbors

The philosophy of China as the center of the world (made evident by the Chinese character word for China, 'Middle Kingdom'), which constituted the Chinese World Order, was not always used solely by Chinese emperors. For example, in seventh-century Japan, which was undergoing the formation of a state, the self-consciousness of the people as an independent state was enhanced by their admiration of China and their emulation of Chinese imperial political systems. Later, in the seventeenth century, when the Ming dynasty was destroyed by a non-Han ethnic minority (the Manchus of the Qing Dynasty), a powerful new discourse arose in tributary countries such as Korea and Vietnam which claimed that they, not the Manchus, were the legitimate successors to the Chinese dynasty, given their long history of tributary relations with China.

In Japan, where the tributary system never took root, a number of intellectuals expressed their outright hostility against Chinese civilization, which had entered Japan

through the widespread use of Chinese characters and Chinese literary texts. For instance, in the eighteenth and nineteenth centuries some Confucian scholars claimed that the term 'Middle Kingdom' was meant to refer only to Japan. Other nativist scholars studying Japanese classical literature during the Tokugawa period (1600–1868) claimed Japan was the 'land of the gods,' effectively placing Japan above China. Anti-Chinese sentiment against Qing (Manchu) China prevailed in the peripheral countries such as Korea, Japan, and Vietnam, and played an important role in forming a political and societal basis for national self-consciousness, because of the natural hostility of these nations toward the 'usurper' dynasty (Arano, 1988).

Early modern Japan engaged in the process of forming an independent national consciousness earlier than other East Asian nations. Never having been a tributary of China, Japanese scholars were able to form their own ideology of international relations during this period of relative stability (Toby, 1984). In addition, Japan successfully maintained an individual route for economic trade with Western countries during the peaceful Tokugawa era. In this respect, it can be said that the Japanese state in the early modern age had already begun to separate itself from a China-centered Asia well before the nineteenth century, when the formation of a modern state along the lines of Western nation states was begun.

The Western Threat and the Acceptance of International Law

The Chinese World Order underwent a tremendous change in the nineteenth century when Western nations began incursions into Asia in search of trading opportunities. During the Qing dynasty, a series of wars between Western nations and China in 1840–42 and 1858–60 compelled a small number of intellectuals to promote the urgent necessity of strengthening state defenses. The treaties concluded between China and the Western powers forced China's involvement in the Western nation state system. As symbols of this, the first bureau of foreign affairs, the *Zongli-Yamen* (Office for General Management), was established in the capital city in 1861, and the exchange of permanent missions also began in the 1870s (Bannno, 1964, 1973).

However, high-ranking Qing government officials as well as officials responsible for making foreign policy actually regarded these treaties as an extension of their tributary policy; therefore, they did not regard these new relations with the West as relations among equal, independent nation states, let alone with China as a subordinate. Although Chinese territory was repeatedly invaded, few Chinese were able to reach across societal divisions to form a consciousness of national crisis and national identity. As long as the conventional, historical Chinese view of the world (with China at the center) remained alluring, it was difficult for the Chinese to rally around a nationalism which would necessitate that they actually view their nation as one of many nations in the world.

Meanwhile, nationalism was growing rapidly in Japan – perhaps partially because Japan had never been free of a consciousness of itself as on the periphery of dominant Chinese civilization, and so had long tried to carve out an independent identity. The catalyst for the rapid growth of nationalism and nation-strengthening movements was the Opium War (1840–42) in China, which called Japan's attention to the danger of the Western powers and their interest in East Asia. Japanese nationalism was aimed at strengthening and modernizing Japan along the lines of a Western-style nation state, and to encourage Japan to abide by the system of international law which theoretically supported the formation of a nation state. As soon as the first international treaty was concluded between the USA, the UK, and Russia in 1854, a political movement called *sonno joi* ('revere the emperor, expel the barbarian') was launched to overthrow the bakufu and establish the Emperor as the head of the state. This movement led to the Meiji Restoration of 1868 (*see* Meiji Revolution, The). With the aim of defending itself against the threat of the Western powers, the Meiji government energetically promoted policies aimed at modernizing Japan's industries, military power, and constitutional system. The end goal was to establish a fully fledged modern nation state, with the emperor at its head, which would earn the respect of the Western nations.

Meiji statesmen and bureaucrats learned the principles of sovereignty and the territory through their studies of international laws (*bankoku koho* in Japanese, *wanguo gongfa* in Chinese) which were introduced from China during the late Tokugawa period. Their main goal was to abolish the 'unequal treaties' which Japan had been forced to sign with the West in the 1850s (*see* Peace-Making in History). Japan also sought influence in some East Asian territories in which China had been dominant, such as Taiwan and Korea. Japanese expansionism brought about increasing regional tensions with the Qing dynasty over influence in Taiwan, Ryukyu, and Korea, and the tension finally culminated in military clashes between the two countries.

The Sino–Japanese War of 1894–95 and the Advent of Modern Nationalism

The Sino–Japanese War of 1894–95 was a major catalyst in promoting the growth of nationalism in Japan and transforming the East Asian world order. This war, which broke out over struggles between the two countries over supremacy in Korea, culminated in Japan's unexpectedly thorough victory over China. The tributary system was finally dissolved with the granting of the unconditional independence of Korea, which had been Qing China's closest tributary state. Taiwan also became a colony of Japan. The result was a new East Asian world order with Japan at its center, replacing China.

The mobilization of the Japanese people during the Sino–Japanese (1894–95) and Russo–Japanese (1904–05) Wars marked a decisive turning point for Japan and the formation of national self-identity. In the wake of the Russo–Japanese War, modern Japan began to expand its image of empire with a view to invading the Chinese continent, since Japan was given rights and interests in Korea and Manchuria as part of the spoils of war. The war united patriotic militarists and also promoted an unprecedented sense of alliance among people in Japan from different regions or social classes.

China's defeat during the Sino–Japanese War was a tremendous shock to the Chinese. Moreover, the sense of crisis in China was intensified by the fact that a small country which had long been regarded as a backward area unexpectedly defeated the Chinese army with military power adopted from Western civilization. It was at this momentous time that a true sense of Chinese nationalism first appeared as the old Chinese World Order finally was dissolved. In particular, the news on the military defeat of China rendered the serious impact on lower level bureaucrats and progressive intellectuals, and they attempted to promote nation state building along the lines of Meiji Japan. Although this reform movement (*bianfa*) failed in 1898 under the pressure of conservatives, the Qing government attempted their own top-down reforms, and the transition to a modern constitutional state was launched on a large scale during the first decade of the twentieth century.

During this period, efforts to build up a nation state were also being made in the former tributary countries such as Korea and Vietnam, especially among the lower stratum of elite. In Vietnam and Korea, nationalism came into being through anti-imperialist movements against France in the case of Vietnam and against Japan in the case of Korea. Taiwan's nationalistic movement was born when the Taiwanese opposed Japanese colonial discrimination during the 1920s under the influence of Wilsonian ideas of national self-determination prevailing since World War I. Decades later, a separate form of nationalism in Taiwan gained strength that was based upon a consciousness of Taiwan as separate from the Chinese continental government.

Nationalism and Asianism

As mentioned earlier, modern Japan was the first East Asian country to create a unified nation state successfully, and it struggled to build up a national identity under the strong outside pressures of Western forces and with the new common national goal of 'leaving Asia, entering Europe' (*Datsu-a nyu-ou*). However, the nationalism of the expanding Japanese empire that had already incorporated Okinawa and Hokkaido in its territories took on a new form when Taiwan and Korea were colonized by Japan in 1895 and 1910, respectively. A new discourse of Asianism emerged that had the goal of opposing Western imperialism against all Asian nations by uniting the Asian nations under Japanese leadership. This ideology led to new complications owing to the variety of ethnicities and national loyalties that existed in East Asia (Takeuchi, 1963).

One example of an Asianist thinker was the Meiji intellectual Tokutomi Soho (1863–1957), who promoted an idea of democracy among the common people (*heimin-shugi*) that also stressed the importance of a strong power in Asia to oppose Western powers. He was particularly provoked by the Triple Intervention of Russia, France, and Germany in 1895, which was perceived as the ultimate example of Western arrogance against Japan.

Asianism would only succeed if it accommodated itself to the new regional order, which was constantly changing in response to factors such as the gradual worsening of Japanese–Chinese relations and global changes in the wake of World War I. The possibility of Sino–Japanese cooperation against the West seemed stronger with the publication of works such as Sun Yat-sen's *Greater Asianism* in 1924 (Sun, 1986).

However, in spite of the lofty ideals of the building of a new East Asian order, the Manchurian Incident of 1931 and the start of the Sino–Japanese war in 1937 meant that the autonomy and unification of the Asian nations was not brought to fruition. Certainly, right after the outbreak of hostilities, the idea of 'East Asian Cooperativism' differed from other philosophies such as liberalism or communism as it attempted to promote a principle that would unify Asia during this period of total war. Such ideas were an important test of the possibilities of redefining the traditional boundaries of the state and its people. They were also a criticism of self-serving Japanese policies toward China and of the general attitude toward Chinese nationalism. These were a new manifestation of the possibilities of the principle of Asianism (Hasikawa and Matsumoto, 1970).

However, aggression continued to maintain a stubborn relationship to unity within Asianist thought. The constantly expanding battlefront and the continued use of military aggression against Asian peoples eventually destroyed the hope of Asian unification. Euphemistic slogans such as 'Liberation of the [Asian] people' or 'Building a new order' inevitably became fodder for Japan to deceive itself and justify aggressive war against its neighbors. The wartime concept of the 'Greater East Asian Co-Prosperity Sphere' (*Dai to-a kyoei ken*) was another manifestation of pan-Asianism. In spite of the lofty principle of Asian cooperation behind the idea, proponents of this concept were never able to take on a point of view critical of Japan's increasing aggression and imperialism. It was not until Japan's defeat in World War II in 1945 that these discourses justifying Japanese imperialism disappeared.

Ethnic Nationalism in the Periphery

The Qing dynasty dissolved in the wake of the Revolution of 1911. The new Republic of China aimed at establishing a highly centralized state that would embark on a number of industrial, societal, and educational reforms. Rural skirmishes with warlords and regional military cliques occupied the energies of the government, however, and the unification of the nation was not achieved until the Nanking Nationalist Government was established in 1928. Mongolia and Tibet both virtually declared independence from China on the grounds that the former tributary relationship was severed when the Chinese emperor abdicated in 1912. Undoubtedly, the Chinese central government did not approve of independence, but it was difficult for the government to rule these areas in an effective way since they needed to maintain diplomatic relations with countries such as the USSR and the UK, which enjoyed a number of concessions in those regions. The tendency for non-Han regions to distance themselves from the central government grew stronger (Mancall, 1984). One such example could be found in the independence of the 'Outer Mongolia' in 1924.

During the 1930s and 1940s, there were a number of political movements for independence by Muslims of Turkish descent in the area of East Turkestan (Xinjiang province), as

well as military disputes between Tibetans and Chinese in the co-habitated areas of Tibet. Although areas such as Inner Mongolia, East Turkestan, and Tibet were merged into China under direct governmental control when the Peoples' Republic of China was established in 1949, ethnic nationalism in these areas reportedly has become increasingly vigorous in recent years (Mori, 1998).

See also: East Asian Studies: Politics; Nationalism, Historical Aspects of: The West; Nationalism, Sociology of; Nationalism: General; Nations and Nation-States in History.

Bibliography

Arano, Y., 1988. Kinsei Nihon to Higashi Ajia (The Early Modern Japan and East Asia). Yoshikawa Kobunkan, Tokyo.

Banno, M., 1964. China and the West, 1858–1861: The Origins of the Tsungli Yamen. Harvard University Press, Cambridge, MA.

Banno, M., 1973. Kindai Chugoku Seiji Gaikoshi (The History of Foreign Relations of Modern China). Tokyo Daigaku Shuppankai, Tokyo.

Best, A., 2010. The International History of East Asia, 1900–1968: Trade, Ideology and the Quest for Order. Routledge, London.

Castles, S., 1998. New Migrations: Ethnicity and Nationalism in Southeast and East Asia. University of Oxford, Oxford. Transnational Communities Programme.

Cheng, G., 2012. Nationalism and Human Rights: In Theory and Practice in the Middle East, Central Europe, and the Asia-Pacific. Palgrave Macmillan, New York.

Christiansen, F., Hedetoft, U., 2004. The Politics of Multiple Belonging: Ethnicity and Nationalism in Europe and East Asia. Ashgate, Aldershot, Hants, England.

Fairbank, J.K. (Ed.), 1968. The Chinese World Order; Traditional China's Foreign Relations. Harvard University Press, Cambridge, MA.

Hamashita, T., 1990. Kindai Chugoku no Kokusaiteki Keiki: Choko Boeki Shisutemu to Kindai Ajia (International Moments of Modern China: Tributary System and Modern Asia). Tokyo Daigaku Shuppankai, Tokyo.

Hasikawa, B., Matsumoto, S., 1970. Kindai Nihon Seiji Shisoshi. Yuhikaku, Tokyo (The History of Political Thought in Modern Japan).

Hasegawa, T., Tōgō, K., 2008. East Asia's Haunted Present: Historical Memories and the Resurgence of Nationalism. Praeger Security International, Westport, CT.

Kratoska, P., 2001. South East Asia: Colonial History. Routledge, London etc (South East Asia.).

Lai, Y.M., 2001. The Resurgence of Chinese Nationalism: Implications toward the Regional Security of East Asia. Kota Kinabalu: Centre for the Promotion of Knowledge and Language Learning. Universiti Malaysia Sabah.

Mancall, M., 1984. China at the Center: 300 Years of Foreign Policy. Free Press, New York.

Mori, K., 1998. Shuen kara no Chugoku: Minzoku Mondai to Kokka. Tokyo Daigaku Shuppankai, Tokyo (Ethno-nationalism in Contemporary China).

Nyiri, P., Breidenbach, J., 2005. China inside Out: Contemporary Chinese Nationalism and Transnationalism. Central European University Press, Budapest, Hungary.

South East Asia: 4, 2001. Routledge, London.

Sun Yat-sen, Y., 1986. Sun Zhongshan Quanji (Collected Works of Sun Yat-sen), (vol. II). Zhonghua Shuju, Beijing, China.

Takeuchi (Ed.), 1963. Ajia Shugi (Asianism). Chikuma Shobo, Tokyo.

Toby, R.P., 1984. State and Diplomacy in Early Modern Japan: Asia in the Development of the Tokugawa Bakufu. Princeton University Press, Princeton, NJ.

Nationalism, Historical Aspects of: South Asia

Tapan Raychaudhuri, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 15, pp. 10351–10355, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

This article surveys the origins of nationalism in South Asia in the nineteenth century under colonial rule, its many strands induced by multiethnicity and ideological differences. The emergence of two successor states, India and Pakistan, as the end product of nationalist developments and the breakup of Pakistan into two states are discussed. Also discussed are resulting postcolonial developments and the various mutually contested explanations of the phenomenon.

Nationalism, which first developed in the West as a conscious political ideology, spread to other parts of the world in the course of the nineteenth and twentieth centuries. It sustained political and social-cultural movements of great power in South Asia, i.e., the Indo-Pak subcontinent. The ethnic diversity of the region, the responses to colonial rule which subsumed and overlapped with reactions to the encounter with Western civilization, as well as the shifting definitions of national identity determined by the region's history both under colonial rule and after decolonization, gave nationalism in the area a very distinctive and at times highly idiosyncratic character. The purpose of this article is to identify these diversities, explain their nature and origin, and trace the outlines of their history over time. It will also discuss the wide variety of approaches to the phenomenon which will define and explain it in fundamentally different ways.

Origin and Development of Nationalism in South Asia

The Pan-Indian Scene in the Nineteenth Century

One major consequence of the British rule in India was the unification of the subcontinent under one centralized political and administrative authority, even though two-fifths of the territory remained under the rule of Indian princes who accepted British paramountcy but enjoyed autonomy in all internal matters. A second development, crucial to our understanding of nationalism in the region, was the growth of a class of people who assisted in the functioning of colonial rule as middle- or low-level functionaries or found new means of livelihood like law, modern medicine, journalism, or teaching in the new Western-style educational institutions. This 'colonial middle class' soon had a shared language in English and were participants in a common intellectual discourse, facilitated by the new print culture and the modern means of transport and communication. Their colonial experience and exposure through modern education to Western history and socio-political ideas laid the intellectual basis for nationalism and also helped create a shared body of expectations and grievances, which were sharpened by the experience of racist discrimination (see Sarkar, 1989).

A new upsurge in literary and intellectual creativity and movements of social reform, described by some as India's nineteenth-century Renaissance, laid the foundations of what may be described as protonationalist ideology in the first half of the nineteenth century (Majumdar, 1962–63). The researches of Western, especially British, Orientalists fed into these new concerns helping forge the belief in a great and unique past and thence the yearning for a comparable future glory. One essential component of this glorious future was to be a seat of honor in the comity of nations. The latter implied sovereignty, though clear articulation of such bold expectations was late in coming: In the last decade of the nineteenth century only a small group of militant nationalists began to think in such terms (Majumdar, 1962–63). The educated Indians, a small portion of the total population, who began to share the newly emerging political consciousness accepted, for the most part, the fact of British rule as an enduring political reality which would persist for an indefinite period. The responses to this awareness ranged from enthusiastic faith in the redeeming and providential character of British rule, through hopeless resignation, to a desperate longing to try to end it: nationalist ideology could be and initially was, for the most part, strongly proempire.

The Dominant Strand in Indian Nationalism

All-India nationalist consciousness and ideology found an institutional and programmatic base in the Indian National Congress (INC) established in 1885. Initially blessed by the colonial government as a 'safety valve' for letting out any accumulated grievance against the ruler, the INC soon came to be suspected as a potentially seditious organization, despite its repeated avowal of loyalty to the empire at its annual meetings. But soon a militant wing did develop within the Congress and its most prominent leader, B.G. Tilak, declared in his journal, *Kesari*, that freedom was the birthright of humankind (Wolpert, 1989). When the first political murders took place, the government believed that Tilak was the source of inspiration for these acts. Militancy found expression in a mass agitation when the government decided to partition the province of Bengal in 1905 hoping, inter alia, to weaken the political base of the nationalistic Bengali Hindu middle class whom they identified as highly 'anti-British' (Sarkar, 1973). The consequent resentment also fed into the movement for violent action against the rulers which led to local and India-wide conspiracies during World War I and after. The conflict between the 'extremists' and 'moderates' induced a split in the Congress

(1907) which weakened the organization for years to come (Argov, 1967).

The advent of Gandhi, a world figure since his nonviolent campaign in South Africa against the regime's racist policies, led to a transformation of the nationalist agenda after the war. Gandhi, despite his pacifism, had cooperated with the war effort until 1918. But violent repressive actions in Punjab in 1919 and the way the British-sponsored enquiry into these events tried to play these down were major factors in his decision to launch a nonviolent campaign of noncooperation. The aim was to win *swaraj*, i.e., self-rule (Brown, 1972). A major section of Muslim opinion in India had been alienated by the way the Allies treated their religious head or *Khalifa*, the Sultan of Turkey. The mass movement launched by Gandhi brought in these disgruntled Muslims who sought justice for the Khalifa (Hasan, 1991). One strikingly new feature of this agitation, known as the Noncooperation-Khilafat Movement, was the technique of *satyagraha* or nonviolent struggle which had strong ethical overtones. For Gandhi, the struggle was not merely a campaign to achieve independence but an 'experiment with truth,' the moral purpose being for him at least as important as the political goal, if not more so (Gandhi, 1948). Following a very violent action by the supporters of the movement, he suspended it in 1922. This brought to an end the Khilafat Movement as well, and with it the unity in a common struggle which brought the nationalists and militant Muslims together.

Gandhi led one more mass agitation, the Civil Disobedience Movement (1930–34) and in 1937, when the British granted a new constitution to India which conceded a fair measure of power to elected governments in the provinces, the INC gained a majority in most of the provinces, but had very limited support in the Muslim majority provinces. During World War II, India was declared belligerent without consultation with the Indian political parties and the Congress ministries in the provinces resigned. A British mission led by Sir Stafford Cripps failed to secure the INC's cooperation in the war effort. The Congress leadership, including Gandhi, who were planning to launch another mass agitation, were imprisoned in 1942 and a largely spontaneous mass movement, known by a slogan devised by Gandhi, 'Quit India,' erupted. Meanwhile, in 1940 the Muslim League, which eventually won the support of the majority of Muslims in India, launched a campaign for an autonomous Muslim homeland, Pakistan. The Congress leaders were released in 1945 at the end of the war and negotiations with all political parties led to transfer of power and the partition of India into two successor states, India and Pakistan in 1947 (Moore, 1974; Mansergh, 1970).

While the INC under Gandhi's leadership provided a common platform for nationalists of many hues, Indian nationalism was never a seamless web. Firstly, within the Congress, from a very early stage there were the moderates and extremists, i.e., those who preferred the path of constitutional agitation leading to autonomy within the empire and the others who preferred a more militant program aiming at independence. Later, the 'Right' and 'Left' within the Congress represented these two different approaches, except that Gandhi, usually identified with the Right, preferred caution even though he led two nationwide agitations. 'Right' and 'Left' came to subsume a more basic ideological division: the former was in favor of moderate socioeconomic reforms, while the latter wanted a radical program of change undermining the existing status quo. The Congress developed a socialist wing in the 1930s and even the newly born Communist Party of India was allowed to function from within the Congress until 1942 when they opposed the Quit India movement (Majumdar, 1962–63). The tension between the Congress's 'Right' and 'Left' reached a breaking point when the youthful leader, Subhash chandra Bose, felt compelled to leave the party and form a separate organization, the Forward Bloc. Later he left India during the war and, with the Japanese help, formed the Indian National Army (INA). The INA, which participated in the war against the British in the eastern theater of war made a tremendous impact on Indian nationalist consciousness. The British had to abandon the trial of INA prisoners of war in the face of such strong sentiments (Gordon, 1990)

Secondly, while the Muslim support for the Congress dwindled, especially after 1928 when an All Party Conference failed to secure agreement between the nationalists and Muslim parties seeking future safeguards for Muslims in independent India, sections of Muslims remained staunchly loyal to the nationalist cause within or outside the Congress (Hasan, 1991).

Thirdly, India's multiethnicity meant that modern political consciousness had more than one focus. 'Nationalism' often meant a strong sense of local cultural identity and a consequent pride in an imagined ethnic past. In many cases, pan-Indian nationalism could accommodate such 'local nationalisms' as an integral part of the larger identity. There was no conflict between a Bengali, Assamese, or Gujrati sense of past glory and present loyalty to the language-based culture, and similar sentiments focused on Indianness and an Indian political identity (McCully, 1940; Brass, 1974). The Congress successfully accommodated regional aspirations and managed interregional conflicts of aspiration, both before and after the transfer of power. Only in specific historical circumstances, especially after the transfer of power, did such potentialities of tension and alienation reach unmanageable levels.

Muslim Nationalism and Foundation of Pakistan

The INC and Indian nationalism failed in their effort to absorb one religion-based ethnic identity. In precolonial times, the Hindu–Muslim divide, so called, was a social rather than a political fact. Under the Muslim kings, Hindus always constituted a large, if not the larger, proportion of the ruling class. In the great rebellion of 1857, Hindu and Muslim soldiers, peasants, feudatories, and princes fought together with the ostensible purpose of restoring Mughal rule. In the colonial era, the rulers' policy of securing alliances through the distribution of such resources as they controlled, especially education, jobs, and seats on representative bodies, were often based on an identification of the communities as basic political constituencies. This fact helped politicize community identities: In the case of Muslims this identity was seen to have a pan-Indian dimension. The cultural revival in nineteenth century India focused among other things on a romanticized vision of Islam's past glory in and outside India (Pakistan Historical Society, 1957–[1970]).

As Indian nationalism's image of the past focused initially on Hindu achievements and often portrayed the rule of the

medieval Muslim dynasts as an age of slavery for the Hindus, the political agenda of the Pakistan movement postulated the theory of two nations, ignoring the multiple components of both the Hindu and the Muslim communities in India, as also the absence of any pan-Indian consciousness in either community in the precolonial period. This two-nation theory was endorsed by some Hindu historians as well who treated the history of the Indo-Islamic era as a period of foreign rule (Hardy, 1972; Page, 1987; Majumdar, 1962–63).

Nationalism in Postcolonial South Asia

In divided India, the power of nationalism in the Indian Union was manifest in the success in bringing all the princely states within the union's boundaries into the union, with the active support of the local populations and the very survival of the union for more than 50 years to date. The local aspirations, often manifested as centrifugal tendencies, have been successfully negotiated for the most part. In Punjab, the aspirations of the Sikh community developed into a movement for independence, but a policy of severe repression and conciliation of more moderate elements has proved successful (Brass, 1994). In the north-east, several local groups continue to fight for secession from India as do the Muslim separatists in Kashmir. But the forces of commitment to a united India appear to be stronger than centrifugal tendencies. A number of ethnic movements in different parts of India are concerned more with the formation of autonomous states within India than with dreams of independence (Baruah, 1999).

In Pakistan, the ascendancy of Punjab and the ruling elite's refusal to accommodate the aspirations of the Bengalis in East Pakistan produced a nationalist movement first based on their linguistic-cultural identity. In 1971, the East Pakistan-based Awami League won an absolute majority in Parliament, but their demand for virtual autonomy was answered by massive repression. This led to a war of liberation, and Indian intervention contributed to the defeat of the Pakistani army. The independent state of Bangladesh was thus born. Some saw in this event the inevitable triumph of ethnicity-based centrifugal tendencies, but perhaps the hamhanded policies of Pakistan's military government were more to blame. The movements for autonomy or secession in Sindh, Baluchistan, and the North West have also perhaps more to do with cynical misrule than spontaneous ethnic aspirations (Khan, 1985).

Differing Interpretations of Nationalism in South Asia

There is no consensus on the nature and dynamics of nationalism in South Asia. The imperialist perception repeatedly emphasized the multiethnicity of South Asian society, implying that a nation state was not viable in the context of this 'society of societies.' In this view, nationalism was the false aspiration of a cynical elite in India who tried to mislead the masses, normally loyal to the imperial connection (Strachey, 1888; Chirol, 1910). Opposed to this was the nationalist self-perception as the upholders of the natural rights of the Indian people: They wrested India's independence from the rulers through sustained struggle; a reluctant colonial power conceded it in slow stages in response to successive waves of mass agitation, unique in their allegiance to an ideology of nonviolence (Sitaramayya, 1946–47). An organization originally confined to the elite gradually brought in the masses through successful propaganda by words and actions. The imperialists, by contrast, projected the notion that they gradually trained up the Indian people in the art of self-government. The delay in handing over power was entirely due to the internal squabbles of Indian politics, especially the irreconcilable aspirations of Hindus and Muslims which eventually led to the partition of India (Coupland, 1944).

More recent interpretations have strong resonances of these older views. A group of historians in Cambridge explained the dynamics of Indian politics in terms of collaboration with the Raj and competition between various indigenous groups for a share in the resources made available to the collaborators. The 'have nots' of such shares took to agitation, and nationalism was the rhetoric of their frustrated aspirations intended to win support (Gallagher et al., 1973). The Marxists explain Indian nationalism as the agenda of the colonial bourgeoisie whose expectations remained unfulfilled under colonial rule. While they successfully secured the support of peasants and workers, their agenda both before and after independence denied the masses their legitimate rights (Dutt, 1940). One version of this interpretation suggests that the class basis of the nationalist movement did expand over time, bringing in the vast majority of the Hindu population, but it excluded significant sections of the masses, especially those at the bottom rungs of Indian society (Pandey, 1978).

The emphasis on ethnicity as a factor which will inevitably undermine national cohesion features prominently in the writings of some observers like Neville Maxwell, who saw in the emergence of Bangladesh the beginning of the process of disruption leading to the inevitable breakup of India and Pakistan along ethnic lines. Radical writers like Partha Chatterjee provide indirect support for this view in stating that the basic loyalty of the masses is to the communities who should be empowered. The state, in this view, is an institution superimposed by the dominant classes to exploit the underprivileged (Chatterjee, 1993a,b). This perception is a logical corollary of the thesis developed by the historians and social scientists associated with the publications entitled the *Subaltern Studies*. This thesis suggests that all studies of nationalism hitherto have focused on the elite groups denying the underprivileged their agency, their role as subjects in their own right. The notion that the elite leadership mobilized the ignorant populace is shown to be false: The masses interpreted the message of nationalism in their own light and participated in the action only when they felt it to be in their interest. And their primary loyalty was always to their community. This thesis has been criticized, among other things, on the ground that the idea of a pan-Indian nation did not come spontaneously to the masses: they learnt it from the nationalist leadership and their propaganda by word and action.

Open Questions and Prognosis for the Future

Some features of nationalism in South Asia are indeed unique. The Indian Union, the world's largest democracy representing a population of which 40% are illiterate, is also the largest

multiethnic state in the history of the world. The comparison with the USSR is invalid, because that state was successor to the Tsarist empire and hence did not represent any voluntary union. The Indian state, on the other hand, was the union of willing partners who voted for such a union through a Constituent Assembly. And the remarkable fact is that despite tensions the union has held together and representative democracy, a system unknown to the region's political traditions, has survived for half a century. In Pakistan, the creation of a state based on the identity of faith was an innovation. That identity has not, however, been able to hold the state together and the democratic principle has also been violated repeatedly by the army. In Bangladesh, there is a persistent tension between the language-based Bengali identity and the emphasis on Islam preferred by significant groups in the country (Ziring, 1992). In both countries, however, the electorate have repeatedly proved their allegiance to the imported ideal of representative democracy.

Crystal-gazing into the future of the South Asian states is not a very rewarding exercise. If there are threats to their survival and risks of breakup into smaller units, one needs to ask whether these arise from irreconcilable ethnicities or from the contingencies of bad government and the exploitation of particular groups by others. How far the problems of poverty and ignorance have inhibited the growth of national consciousness among the masses also needs to be analyzed. Another question which will have to be addressed is the implication of globalization for nationalist ideologies and the nation state itself.

See also: Asia, Sociocultural Overviews: South Asia; Nationalism, Sociology of; Nationalism: General; Religion: Nationalism and Identity; Religious Nationalism and the Secular State.

Bibliography

Argov, D., 1967. Moderates and Extremists in the Indian Nationalist Movement. Asia Publishing House, Bombay, India.
Ahmed, T., 2009. Literature and Politics in the Age of Nationalism: The Progressive Writers' Movement in South Asia, 1932–56. Routledge, Taylor & Francis Group, New Delhi.
Baruah, S., 1999. India against Itself: Assam and Politics of Nationality. University of Pennsylvannia Press, Philadelphia, PA.
Brass, P.R., 1974. Language, Religion and Politics in North India. Cambridge University Press, London.
Brass, P.R., 1994. The Politics of India since Independence. Cambridge University Press, Cambridge, UK.
Brown, J., 1972. Gandhi's Rise to Power – Indian Politics 1915–1922. Cambridge University Press, Cambridge, UK.
Bénéï, V., 2005. Manufacturing Citizenship: Education and Nationalism in Europe, South Asia and China. Routledge, London.
Bhagavan, M.B., 2010. Heterotopias: Nationalism and the Possibility of History in South Asia. Oxford University Press, New Delhi.
Bose, S., Jalal, A., 1998. Modern South Asia: History, Culture, Political Economy. Routledge, New York.
Chatterjee, P., 1993a. Nationalist Thought and the Colonial World. University of Minnesota Press, Minneapolis, MN.
Chatterjee, P., 1993b. Nation and Its Fragments: Colonial and Post-Colonial Histories. Princeton University Press, Princeton, NJ.
Chirol, V., 1910. Indian Unrest. Macmillan, London.
Coupland, R., 1944. Constitutional Problem in India. Oxford University Press, New York.
Cederlöf, G., Sivaramakrishnan, K., 2006. Ecological Nationalisms: Nature, Livelihoods, and Identities in South Asia. University of Washington Press, Seattle.
Dutt, R.P., 1940. India To-Day, second ed. Manisha, Calcutta, India.
Fazal, T. (Ed.), 2013. Minority Nationalisms in South Asia, Routledge, New Delhi.
Franke, M., 2009. War and Nationalism in South Asia: The Indian State and the Nagas. Routledge, London.
Gallagher, J., Johnson, G., Seal, A., 1973. Locality, Province and Nation. Essays on Indian Politics, 1870–1947. Cambridge University Press, Cambridge, UK.
Gandhi, M.K., 1948. Gandhi's Autobiography: Story of My Experiments with Truth. Public Affairs Press, Washington, DC.
Gordon, L., 1990. Brothers Against the Raj: A Biography of Subhas and Sarat Bose. Viking, New Delhi, India.
Ghosh, P.S., 2007. The Politics of Personal Law in South Asia: Identity, Nationalism and the Uniform Civil Code. Routledge, New Delhi.
Hardy, P., 1972. The Muslims of British India. Cambridge University Press, Cambridge, UK.
Hasan, M., 1991. Nationalism and Communal Politics in India, 1918–1928. Manohar, New Delhi, India.
Hashmi, T.I., 1992. Pakistan as Peasant Utopia. The Communalization of Class Politics in East Bengal 1920–1947. Westview, Boulder, CO.
Khan, A. (Ed.), 1985. Islam, Politics and the State. The Pakistan Experience. Zed Books, London.
Mansergh, N. (Ed.), 1970. Transfer of Power, 12 vols. HMSO, London.
Majumdar, R.C. (Ed.), 1996. History of Freedom Movement, 3 vols. Classical Publishing Company, Calcutta, India.
Majumdar, R.C. (Ed.), 1962–63. British Paramountcy and Indian Renaissance. Bharatiya Vidya Bhavan, Bombay, India.
McCully, B., 1940. English Education and the Origins of Indian Nationalism. New York.
Moore, R.J., 1974. Crisis of Indian Unity. Clarendon Press, Oxford, UK.
Mallot, J. Edward, 2012a. Memory, Nationalism, and Narrative in Contemporary South Asia. Palgrave Macmillan, New York.
Mallot, J.E., 2012b. Memory, Nationalism, and Narrative in Contemporary South Asia. Palgrave Macmillan, New York.
Page, D., 1987. Prelude to Partition: The Indian Muslims and the Imperial System of Control, 1920–1932. Oxford University Press, Delhi, India.
Pakistan Historical Society, 1957–[1970]. History of Freedom Movement, Being the Story of Muslim Struggle for the Freedom of Hind-Pakistan, 1707–1947. Pakistan Historical Society, Karachi, Pakistan.
Pandey, G., 1978. The Ascendancy of the Congress in Uttar Pradesh 1926–34. A Study in Imperfect Mobilization. Oxford University Press, Delhi, India.
Pandey, G., 2001. Remembering Partition: Violence, Nationalism, and History in India. Cambridge University Press, Cambridge.
Rai, R., Kukreja, S., 2009. Nationalism in South Asia. Ashwin-Anoka Press, New Delhi.
Sarkar, S., 1973. Swadeshi Movement in Bengal 1903–08. People's Publishing House, New Delhi, India.
Sarkar, S., 1989. Modern India, 1885–1947, second ed. Macmillan, Basingstoke, UK.
Sitaramayya, P., 1946–47. History of the Indian National Congress. Padma Publications, Bombay, India.
Strachey, J., 1888. India. London.
Silva, N., 2004. The Gendered Nation: Contemporary Writings from South Asia. Sage Publications, New Delhi.
Wolpert, S., 1989. Tilak and Gokhale: A Revolution and Reform in the Making of Modern India. Oxford University Press, Delhi, India.
Watt, C.A., Mann, M., 2011. Civilizing Missions in Colonial and Postcolonial South Asia: From Improvement to Development. Anthem Press, London.
Ziring, L., 1992. Bangladesh from Mujib to Ershad. An Interpretative Study. Oxford University Press, Karachi.

Nationalism, Historical Aspects of: The West

Miroslav Hroch, Charles University in Prague, Praha, Czech Republic

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article starts from the premise that nationalism has to be studied in relation to nation: an historical aspect of nationalism is the formation of a modern nation and the creation of nation-states in Europe. The study of the phenomenon, however, is complicated by disparate use of the concept of nationalism by various researchers. 'Nationalism' has thus become a hazy and not very suitable tool for analysis, which is also reflected in numerous misunderstandings in evaluations of both the positive and the negative roles of nation and nationalism in history. This article documents that, despite the considerable differences in terminology and evaluation, there are roughly five fundamental causal relations that the various research schools consider crucial for nation – and nationalism – forming: (1) state-political historical foundations; (2) ethnic roots; (3) processes of modernization, including industrialization, institutional education, and social communication; (4) nationally relevant conflicts of interest and unevenness of development; and (5) psychological and emotional manipulation.

The complicated process of the modern nation-forming and the onset of the manifestations of nationalism has been, quite understandably, a subject of scholarly research ever since the nineteenth century. This research has had to struggle with a number of theoretical and semantic problems, and has not achieved consensus over the fundamental issues of terminology. The question of what role was played by 'nationalism' has been answered differently by different authors, primarily depending on how they defined nationalism, nation, nationality, national identity, and ethnic group – and relations between these categories. Therefore, it is most important to investigate through what terms the authors analyzed the processes and how they defined the terms.

The Problem of Definition

The fundamental problem of terminology derives from the fact that it concerns a process resulting in the formation of new communities, and that these communities used a term for themselves that had already originated in the Middle Ages, but was given a different content in different political and cultural conditions. On the one hand, 'nation' in English included all inhabitants living under one government and ruled by the same laws. The first edition of *Grande encyclopedie* also defined '*la nation*' in similar terms, but later the use of a common language was added to this characterization. On the other hand, 'nation' in German and also in Polish, Czech, and so forth was traditionally primarily associated with language, culture, and also sometimes with a shared past. That is why anglophone researchers find it hard to accept speaking about a fully fledged nation in cases where there is no nation-state (Kemilainen, 1964). It was only during the last two decades that the term 'nations without state' was introduced into the scientific discourse (Guibernau, 2007). Another difficulty comes from the fact that 'nationalism' entered the academic vocabulary quite late: it was a neologism originally used in politics and, therefore, laden with political connotations. Nevertheless, the term is also derived from 'nation,' which means that it becomes related to the linguistic connotations mentioned. It follows that anglophone literature will associate nationalism, one way or other, with a state and/or a struggle toward statehood.

Views on Nation, Nationality, and Nationalism in the Nineteenth Century

The study of nation and national consciousness developed alongside the nation-forming processes at least from the mid-nineteenth century and was influenced by the following circumstances: (1) the previous linguistic tradition; (2) the axiom of 'primordiality,' that is, by the idea that the nation *an sich* had existed since the Middle Ages and that people gradually realized their objective belonging to it; (3) the degree of nation-forming of a given nation: English and French researchers presumed the existence of a nation as a nation-state, while for German, Italian, and 'East-European' writers a nation was not commonplace and they were still looking for a suitable definition. It is not a coincidence that the largest numbers of the studies of nation can be found in the German-speaking territory.

It needs to be added that English and later American authors also looked for a definition of a nonstate community. For the most part, they used the words 'nationality' or 'peoples.'

The large number of studies written up until World War II can be divided into two main streams:

1. Nation was defined by objective, empirically researchable characteristics. These usually included, first of all, cultural and linguistic characteristics in combination with various other features – from political ties to territorial, economic, and 'blood' relations. Nevertheless, already at the beginning of the twentieth century some authors concluded that it was impossible to find a universally applicable combination of characteristics, including all entities considered nations at the time. That is why a differentiation between a state and a cultural nation gained support (Meinecke, 1907). The criterion was whether the community that called itself a nation was a state or not. Another approach was to define nation as a 'community with a shared past,' which originally applied only to the members of the ruling

classes and elites, but which integrated the entire population in the course of historical development since the Middle Ages (Bauer, 1907).
2. Nation was defined subjectively – its members were aware of their belonging to the nation and wanted to belong to it. This concept could be expressed in terms of agitation – as a will to form a nation – or existentially, as a sort of a 'plebiscite' – as a decision to belong to an existing national community (Renan, 1888). Typically, the subjectivist concept of nation originally appeared in essays rather than in empirically researched studies. The dividing line between both concepts is not clear cut. On the one hand, supporters of the objectivist definition realized more and more strongly that a nation could not exist without national consciousness.

On the other hand, the subjectivist definition did not preclude a primordialist (or perennialist) concept of a nation. It is symptomatic that national consciousness, or 'nationalism,' became a tool of scholarly analysis as late as in the period between the two World Wars. The pioneer of the study of nationalism, Carlton J. Hayes (1926), presumed that it was necessary to define the object of nationalism, and attempted to define both nation and nationality through objective features.

The Definition of Nation and Nationalism in the Second Half of the Twentieth Century

The sad experience with nationalism during World War II for Europe affected the approaches of historians and social scientists to the study of nation. The political acuteness of this study increased and also with it the political commitment of the authors.

Although opinions continued to vary, most researchers agreed on a number of points in setting priorities in research and in discarding insupportable views:

1. An overwhelming majority of researchers more or less distanced themselves from the view that nation is a timeless category and a blood-related community.
2. Agreement was also reached concerning the impossibility of defining nation only through cultural and linguistic characteristics.
3. Nation was usually accepted as a distinct community only if evidence was available that the members of the community (or at least the prevailing majority of them) considered themselves a nation.
4. Hence the growing emphasis on the study of nationalism as a subjective manifestation or as a precondition for the objective existence of a nation.

Nevertheless, conservative historians perpetuated for a while the traditional idea that a modern nation was a continuation of a medieval nation. After all, studies of national (or proto-national) identities in the Middle Ages and the Early Modern Age are fairly common in Europe.

In the conditions of political bias against German nazism and Italian fascism, the atmosphere in Europe was not favorable for applications of the term 'nationalism' as a neutral tool for historical analysis as it was developed by Hayes (1931). On the contrary, the numbers and importance of authors who considered nationalism a negative extravagance that needed to be overcome increased. According to authors such as Friedrich Hertz, Edward H. Carr, and later also Elie Kedourie (1960), nationalism was invented by irresponsible educated individuals, spread by equally irresponsible demagogues and politicians, and should be eradicated as a dangerous legacy of the past.

A polarized and modified variant of this concept as promoted by Hans Kohn in his numerous publications became more influential. In his view, two kinds of nationalism were to be distinguished: Western, which was rationalist, constitutional, liberal, democratic, and developed in France and the United Kingdom; and Eastern, which was irrational, mythical, racist, and reactionary. Although this dichotomy (Snyder, 1954) was persuasively criticized many times (Kemilainen, 1964; Billig, 1997; Schopflin, 2000), it had a number of adherents.

Most historians and social scientists sought acceptable ways of defining nation as an objectively existing social group during the postwar decades. The approach that it was possible to develop an authoritative set of features every nation had to have survived only as the official (although not Stalinist) doctrine of Soviet research. On the other hand, a number of authors were trying to find a more free definition by combining several kinds of relations. Among these relations, Boyd Shafer (1955) distinguished 'illusions,' such as economic needs, language, and shared history; and objective relations, such as shared government, shared cultural institutions, and social institutions. The sociologist Pitrim Sorokin (1947) saw nation as a 'multibound group' formed by a set of individuals "linked by two or more unibound ties." These could be a territory, a language, a religion, and/or physical conditions. A similar view among sociologists is held by Smith (1991), who characterized nation as a population set sharing the same myths, same historical territory, mass culture, and economy. Another variant of this perspective was a concept of nation as a social group bound by a changing combination of several kinds of historically strengthened, mutually substitutable relations, while civic equality of all members of the group and the consciousness of their shared history were necessary and nonsubstitutable ties (Hroch, 1985).

The problems with finding a common, generally acceptable definition of nation also led to giving up such a definition and favoring the subjective definition of nation: "a nation exists when a significant number of people in a community consider themselves to form a nation" (Seton-Watson, 1977). This change was often associated with the 'discovery' of Renan's concept of nation. The result was an increased frequency of use of the term 'nationalism.' The object of this nationalism was often not defined, or it was characterized as a super-personal entity toward which a nationalist individual feels unconditional loyalty (Lemberg, 1964). This view gradually radicalized: nationalism began to be considered a given reality, while nation was regarded a construct, a myth, or an invention. The theoretical authority became Benedict Anderson (1983) – through no initiative of his own. He called nations 'imagined communities,' by which he meant that a nation as a group could only exist if its members were able to imagine that they belonged to a community of people,

of whom the majority they did not know, and would never know. This view, however, can also be interpreted to mean that nation is just an illusion and an invention and therefore the necessity of its formation can be doubted. The constructivist concepts of nation do not solve the problem of definition, they merely move it to another perspective: they also have to explain the term 'nationalism' in its historical context.

Equally as in the case of nation, there is not a consensus among researchers on the term 'nationalism' either. Quite the contrary, the terminological confusion is augmented by the haphazard use of the term by researchers and without attempts at definition.

However, we can also find important exceptions. John Breuilly (1982) defines nationalism as a politics based on the premise that a nation exists, and which identifies itself with this nation. It demands political autonomy for the nation, usually in the form of a state. Gellner's authoritative definition of nationalism as an effort to obtain the attributes of a state for one's own ethnic entity appears around the same time. Later, Gellner substantially modified his definition: "Nationalism is a political principle which maintains that similarity of culture is the basic social bond" (Gellner, 1997). Recently, some authors propose to understand nationalism as a "discoursive formation" (Calhoun, 2007) or "the way of seeing the world" (Delanty, 2002). This particular form of discourse does not mean a denial of reality (Ozkirimli, str. 2010).

Attention will now be paid to those authors who considered nation as objectively existing, or essential. Karl W. Deutsch (1953) made the first significant attempt at a new approach. He defined nation as "a community of complementary habits of communication." This communication community then forms a basis for a shared national culture and also nationalism. This approach was later more precisely defined as a group of people capable of better mutual than outward communication and cooperation, and which is politically organized and autonomous. Members of the nation then associate community consciousness with certain values and with culture. Thus a foundation was created for the development of nationalism. Another step in this direction is the definition of nation as a "social formation" that is "integrated through the means and relations of communication, production, organization, and exchange" (James, 1996). It needs to be mentioned that the majority of these authors did not know that they were drawing on a concept of Bauer (1907).

The view that nation is primarily defined by statehood is widespread mainly among anglophone authors. J. Breuilly, Stein Rokkan, and Charles Tilly should be mentioned, and also Louis Snyder and Anthony Giddens, who maintain that a nation only exists "when a state has a unified administrative reach over the territory over which its sovereignty is claimed" (Giddens, 1987). Similarly, Eric Hobsbawm concludes that nation is "a social entity only insofar as it relates to a certain kind of modern territorial state" (Hobsbawm, 1989). In these schools of thought, nationalism was closely connected with the state.

Defining nation by statehood runs against an important empirical fact that communities that have had a long way to go to statehood, and sometimes have not even sought it, call themselves nations. Especially in English there is a problem with finding an adequate term for this case.

From the opposite side, cultural activities and traditions are recently again regarded as a factor of central importance defining the modern nation (Leerssen, 2006).

Debate over the Evaluation of Nationalism

As mentioned in section The Definition of Nation and Nationalism in the Second Half of the Twentieth Century, the term 'nationalism' had been a political term until the inter-war period. It was used in both negative and positive senses, referring either to a negative and an antisocial attitude (especially in the political language of the socialist movement) or to a commendable and noble love for one's nation (in the political language of some right-wing movements).

Some authors writing within the Marxist tradition associated nationalism with negative social phenomena. E. Hobsbawm and, less distinctly, B. Anderson are examples of this perspective. It is also true, however, that this same tradition made the important attempt to characterize the 'Janus-face' of nationalism as a phenomenon that could have a negative and oppressive impact on the one hand, and a positive impact as a liberating force in the struggle against national or colonial oppression on the other (Nairn, 1981).

The distinction between the two kinds of nationalism is typical for a large number of researchers, and it is foreshadowed by Hans Kohn's West–East dichotomy mentioned earlier. The distinction between positive and negative nationalisms according to the country or the national context was also used later (Greenfeld, 1992).

Another dichotomy is a distinction between nationalisms according to time periods. In 1956, the influential German historian Theodor Schieder (1991) evaluated nineteenth-century nationalism as a progressive and creative factor, while he criticized the manifestations of the destructive side of nationalism, especially in the first half of the twentieth century. This distinction was later repeated by other authors. Ironically, the official Soviet scholarship also recognized two similar phases, although it did not explicitly use the term 'nationalism.'

A majority of contemporary authors recognize the double face of nationalism in history, but without giving geopolitical meaning. For example, Louis Snyder (1976) considered nationalism as moral, as well as immoral, human as well as inhuman. According to these criteria, it could be evaluated as positive in some cases and as negative in others. He saw the main danger of nationalism in its association with collectivism against individualism. The French philosopher Etienne Balibar and Immanuel Wallerstein (1991) sees the good side of nationalism in that it contributes to the development of state and a community of citizens in general, it derives from the love of people, and it is tolerant. In contrast, the bad side of nationalism is in its expansiveness, hatred, and intolerance. It is symptomatic that the positive potential of nationalism is mostly appreciated by non-Anglo-American authors: apart from the French author mentioned, also the Catalunian historian Josep Llobera (1994) and the Israeli historian Hedva Ben-Israel can be mentioned among many other authors. Liah Greenfeld (2006) finds that nationalism became an important driving force of modernization.

Another group of authors considers the broad meaning of the term 'nationalism' to be a source of imprecisions and

misunderstandings, and recommends the use of the term only to refer to national egoism, preference of one's own nation over others, and placing one's own national interest over the interests of an individual (P. Sugar, L. Doob, O. Dann, W. Norman). Some recommended the use of other terms, such as patriotism (Doob, 1964; Samuel, 1989; Vincent, 2002), love for one's country, national consciousness, and national feelings for the positive manifestations of national identity. Another variant is the recommendation to preserve nationalism as the neutral term and find other terms, such as ethnocentrism (L. Le Vine, D. Cambell), ethnonationalism, xenophobia, or chauvinism for the negative manifestations. A real terminological potential can come from those researchers who take the historical – that is, developmental – perspective. They proceed from the fact that attitudes called nationalism developed from certain previous, developmentally 'lower' attitudes, perceptions, and relations, which are sometimes called identities, at other times loyalty or primordial sentiments. The term 'loyalty' is rarely used in historic research, but it belongs to the terminology of cultural anthropology. It is characterized as "a relation of an individual toward pre-existing given groups, phenomena values – in the social reality in which the individual lives" (Geertz, 1973).

The term 'national identity' now belongs to the common terminology of a number of historical analyses. It is used most consistently by Anthony Smith, who maintains that nationalism needs to be studied as a certain consequence, or as another phase in the development of national identity. At the same time, the consciousness of each individual is determined by multiple identities, which are more or less interrelated. Unlike the term 'nationalism,' the term 'national identity' is not laden or obscured with superficial journalism in the vocabulary of historic analysis. A number of contemporary scholars and projects promote using both terms in different contexts (Guibernau).

The Swedish historian Sven Tagil (1984) suggests a distinction between identity, referring to the simple consciousness of belonging to a group, and identification, referring to the active and committed participation in the work for the nation.

The debate over definitions and evaluation should not be seen as mere academic disputes or noncommittal intellectual exercises. The debate is not self-serving; it is a discussion over whether and how a particular term or concept is suitable for explaining the fact that after 200 years of dramatic and controversial development in Europe, in most cases, the nation-forming process almost everywhere resulted in monoethnic nation-states. Nevertheless, neither national identity nor nationalism disappeared.

Why Nation and Nationalism? Attempts at a Historical Explanation

Every survey of views on the genesis of nations and nationalism is confronted with the contradiction between those who consider nation an invention or a construct – and therefore the product of nationalism – and those who see nation as a result of objective nation-forming processes. Judging from the verbal expressions, the conflict is irreconcilable. However, is that really so, or is the conflict not so fundamental – as Paul James (1996) argues? The work of Ernest Gellner illustrates the productivity of the approach that tries to see reality behind words and aphoristic statements. His apodictic statement about nationalism creating nations is often cited and made absolute by 'modernists.' Few of them realized that Gellner's definition only moved the causal explanation to a different level: the point was to explain why nationalism had gained such a response. At this level, Gellner works mostly with objective social and cultural factors. Similarly, it clearly follows from the statements of Benedict Anderson that nations as 'imagined communities' could only be formed, thanks to particular objective social preconditions. Therefore, a causal interpretation does not necessarily have to dramatically contrast the nation-forming process with nationalism.

Trying to find an easier orientation, Anthony Smith (1998) distinguished several categories of historical explanations of nationalism. Perennialist and primordialist approaches explain the nation building in terms of 'longue durée' and of primordial ties and attitudes (Bergh, Geerts, Fishman, Connor, Hastings). The broad diapason of modernist authors explained nationalism in the context of all kinds of modernization processes, in some cases with stronger emphasis on constructivism, sometimes with elements of perennialism (Anderson, Hobsbawm, Breuilly, Tilly, Nairn, Gellner, Hroch). Between these two main streams Smith positioned his own theory and gave it the name of ethnosymbolism (Armstrong, Hutchinson). Ethnic identities were reinterpreted and used as symbols mobilizing national feelings. Smith considers the postmodernists – who try to find fully new dimensions, partially based on modernist paradigms, partially deconstructing them (Brubaker, Chaterjee, Billig, Ozkirimli) – as more marginal.

The flood of theories of nationalism seems quite justly confusing at first sight, but a closer examination will reveal that most accounts attempting historical explanations of nationalism repeat the same basic relations and factors, although using different vocabulary and concepts. The differences between the authors are then mostly in the emphasis on a particular kind of relation. This is not the case of those who belong to the postmodernist stream, and who are resigned to generalizing explanations of nationalism.

This article will therefore not deal with singular authors but rather it will characterize their evaluation of the relevance of the particular relations and contexts: (1) historical background; (2) ethnic and cultural roots and, within these, also denominational relations; (3) the process of modernization, that is, particularly industrialization – and the ensuing changes of social structure, social communication, and the unevenness of the modernization process; (4) conflicts of interests; and (5) psychological relations.

Historical Preconditions or the Legacy of the Past

The view that nation-forming was an inevitable result of the historical conditions of the past centuries is now considered a traditionalist residue. Despite that, historians as well as social scientists, Anthony Smith the most important among them, still consider the past to be an important component of the causal interpretation of the nation-forming process and nationalism. They differ in the strength of the influence they

ascribe to history and, particularly, in the analytical perspective. To put it simply, the importance of history is seen from two perspectives: from the perspective of institutions or other objectively operating results of the past development, and from the perspective of 'collective memory' or the construct of national history.

In the former the issue is how institutions, relations, and stereotypes surviving from the past influenced nation-forming: it refers to the state, legislation, and also estates, judiciary, borders, internal administration, the extent of centralism, and other concepts. The notion that the right to the recognition of national existence is conditioned by the existence of a national state from the Middle Ages gave rise to the nineteenth-century distinction between historical and nonhistorical nations, that is, nations with or without their own (political) history. Although this distinction is no longer considered relevant, a number of authors still work implicitly with the idea that some nations (such as the French, Dutch, and Portuguese) are a given, natural, and organic part of the state system, while others formed somewhat by accident (the Estonians, Slovenes, Bulgarians, and many others). Authors (for example, Shafer, Breuilly, Hobsbawm, Rokkan, Llobera, Tilly, and Barth) who define nation by statehood or understand nationalism as a struggle for statehood, usually assign an important role to political or state history.

A particular problem in analyzing the role of history is that people expressed views and stereotypes that many researchers would today consider nationalist – that is, manifestations of national consciousness – already in the Middle Ages and especially in the Early Modern Period – a long time before the modern nation-forming processes. Modern historians agree that the pre-Modern manifestations of national consciousness did not include the entire community but only the privileged classes – the nobility and also sometimes the higher burgher categories (Hagen Schulze, Norbert Elias, and others). Bauer was already aware of this in his periodization of national development.

'Collective memory' is the other perspective on the effects of the past on nation-forming. Authors working within discourse analysis particularly emphasize that the strength of past statehood is not important, but the way the past was understood, interpreted, and worked with – and therefore also manipulated in the national interest. Hobsbawm's phrase "invented tradition" is usually quoted in this context, and it usually refers to skepticism toward any real historical continuity. At the same time, it is unthinkable to invent the past without any real evidence. On the other hand, recent research confirmed how important it is to study the creating of the image, the construct of national history under the conditions of nation-forming (Stephan Berger and Chris Lorenz, 2008).

Ethnic Roots of Nationalism

It is understandable that most authors who associate nation and nationalism with state do not ascribe much importance to the ethnic and cultural component of the nation-forming process. Some even consider ethnic consciousness neutral, and not automatically a source of nationalism.

On the other hand, a number of authors consider ethnic identities and cultural particulars an important, if not a key precondition of development toward a modern nation. Anthony Smith (1991, 1986) paid the greatest attention to this relation. He differentiated between the degrees of ethnic identity and considered that an ethnic group could also develop its collective memory. Ethnic groups could also enter into conflictual relations, which in many aspects forestalled national conflicts (Horowitz, 1985). A number of other authors considered ethnic roots of national existence important, despite their differences in terminology: sometimes they spoke of ethnonationalism, sometimes of nationality, and at other times of primordiality or primordial ethnicity.

Unlike the general role of ethnicity, religion is only mentioned in concrete cases (the Balkans, Ireland), and it is not listed among the necessary preconditions of the nation-forming processes. Nevertheless, several authors acknowledge the indirect role of religion as a modernizing factor in education and communication, particularly as an agent of reformation and counterreformation (Anderson, 1983; Armstrong, 1982 and others). The significance of overcoming religiosity with the rise of secularism is also acknowledged. At the same time, some authors see nationalism as a 'national,' secularized religion.

Nation and Nationalism as Manifestations or Consequences of Modernization

The modern nation-forming doubtlessly occurred during the period of civil revolutions, industrializations, and the beginnings of the free market. That is, during the period called modernization by some, overcoming of traditional society by others, and the transition from feudal to capitalist society by other authors. Most authors (unlike, for example, Kohn) consider this simultaneous occurrence more than a coincidence, but they construct the causal relations in different contexts. Already in 1953, Deutsch saw the, industrial revolution and the onset of the free market as crucial for the spread of nationalism. Later, he considered nationalism a response of people on the one hand to the possibilities of improving their status thanks to social mobility, and on the other hand to the loss of ties to the traditional lordships and settlements. Gellner's thesis (1983) about industrialization as the fundamental impulse for the birth of nationalism cannot therefore be seen as original. Gellner, however, does not conceive of industrialization in the literal sense referred to by historians, that is, as the introduction of the industrial mode of production. Rather, he means market economy and social mobility in general as factors opening the way for the principles of equality of citizens and their new role in the transformation of culture. Michael Mann (1988) and other authors, such as Etienne Balibar and Immanuel Wallerstein (1991), work with a similar, although more differentiated, thesis. Carolyne M. Vogler (1985) also shares this perspective when she maintains that the "disjunction between economics and politics" occurred, and that significantly contributed to the dissolution of multinational states and the creation of nation-states. In contrast to the thesis of the role of industrialization, other authors emphasize general modernization changes ensuing from the progressing liberation of the market and people (Hroch, 1985; Llobera, 1994). The emergence of the new forms of enterprise (that is, capitalist enterprise) and the related social mobility are specifically reflected in increasing

social communication, that is, a system of an increasing amount of information being exchanged among an increasing number of people. As mentioned, Karl W. Deutsch ascribed the key importance to this circumstance on the formation of nation and nationalism, but his one-sided emphasis on communication was not generally accepted. His views were nevertheless confirmed within another terminological framework: the thesis about the crucial role of education in the people's ability to accept the notion of a nation as an abstract community became the central argument of Anderson. It was already advocated earlier by Gale Stokes (1974), who proposed that the acceptance of such a notion depended not only on elementary literacy, but also on schooling until the age when a child developed the ability of abstract thinking. The German sociologist Berndt Giesen (1991) connected both of these perspectives: in the conditions of the new type of communication between addressees who did not know each other, a new style of communication and new coding was necessary. The idea of nation became such a code for encoding the new living conditions of modern society.

Numerous researchers argue that the spread of nationalism was boosted by the deepening of the unevenness of development – another consequence of modernization and industrialization. A specific form of unevenness was a 'cultural division of labor between the core and the periphery,' which resulted in the inferior social role of the periphery. This led to the 'internal colonialism' as a breeding ground of national and regional movements opposing the 'center' (Hechter, 1975). Mann (1988) and Nairn (1981) also considered uneven development and the associated sentiments of oppression, important factors of nation-forming processes. The view of the nation-mobilizing role of unevenness ('backwardness') has been criticized frequently and its general validity has been rejected.

Nationally Relevant Conflicts of Interests

This brings the discussion to the factor that – again veiled by the various terms – plays an important role for perhaps still the majority of approaches seeking the explanation of nation-forming and also of the spread of nationalism in objective social changes. The thesis of internal colonialism alone suggested that the conflict between the falling-behind periphery and the center could be expressed in the categories of national interest. In other words, uneven development could be counted among nationally relevant conflicts.

Some authors accentuated the role of international conflicts and above all of wars as a decisive basis of nationalism and national mobilization (Langewiesche, Laitin). There is no doubt that in some parts of Europe – above all in southeast – wars accompanied all cases of nation formation. On the other side, wars used to be rather a consequence and accelerator of militant nationalism instead of being its origin (Hogan, 2009).

All this range of conflicts from wars to socially and status-motivated tensions can be seen as conflicts of interest. This fact makes any generalization difficult, but it also allows for different interpretations. The most frequently listed conflict in the explanations of nationalism is the conflict ensuing from uneven development.

A social conflict between groups or classes that, at the same time, belong to different ethnicities is theorized relatively seldom in contemporary research (Miroslav Hroch, Paul James, Tom Nairn, Eric Hobsbawm), as is the idea of the struggle against an old society in the national interest (Boyd Shafer).

Nationalism as a Result of Psychological, Cultural, and Emotional Manipulation

Most researchers working on the explanations of nationalism also look for its sources in the realm of psychology and emotions. The individual models, however, differ significantly according to which aspect of the human psyche they consider important. Insecurity of a human individual and the need to belong to a group and to identify with it are some of the most frequently quoted aspects. The insecurity is then seen as being caused by perceived (or real) endangerment, while the source of endangerment can be either external danger or internal insecurity resulting from the perception of uprootedness, dissolution of old securities, old faith, and so forth (Miroslav Hroch, Gale Stokes, Anthony Smith, Bernd Giesen, Ernest Gellner). Nationalism is sometimes considered a sick, pathological reaction to such perceived endangerment (Committee on International Relations, 1987), and at other times a response to the inferiority complex (Lemberg, 1964). Anderson speaks of the need to overcome the fear of an individual death by identifying with an immortal nation.

Following partially from Geertz's primordial ties, some researchers tried to apply biological research on animal behavior to nationalism, and explain nationalist attitudes as the 'territorial imperative' (Snyder, 1975). Nevertheless, this approach has remained marginal in the overall context.

'All the stronger' has recently become the effort to explain nationalism as a cultural construct deriving from the activities aiming at the emotional mobilization of human psyche. Thus national rituals are becoming a subject of intense research (Mosse, 1973), as well as symbols, celebrations in the name of nation (E. Francois, Elgenius, 2011), and pilgrimages to places commemorating the national past (Nora, 1986) – in short, activities deliberately aimed at the emotional element of the human psyche. It includes also manipulations through state institutions, media, movies – factors that are regarded as decisive of nationalist mobilization in the eyes of constructivists and postmodernists.

Overall, differences in the explanations of nation formation are not as fundamental as it may appear on the surface from the different terminology and the verbal radicalism of some authors. The understandable efforts for originality and impressive contribution are often a facade of new expressions repeating older conclusions and findings.

Most of the studies on the historical origins of nationalism are flawed in that they ignore the fact that the process of nation-forming was present in two basic types in Europe: the first started with nation-states that had existed since the Middle Ages and began to transform into modern nations at the end of the eighteenth century or later. The point of departure of the second type were nondominant ethnic groups that inhabited the territories of large multiethnic empires, did not have ruling elites, had their own cultural traditions in a codified language, and had either a disrupted or no tradition of statehood. That is why it is important to distinguish between nationalism (whether considered as a part of national identity or merely as

its nationally egoistic variant) developed in the conditions of nation-states and the kind of nationalism known from national movements. The latter type again differs depending on whether it developed under conditions of phase of national agitation (phase B) or a mass national movement (phase C). In any case, nationalism should never be viewed as a compact theory or ideology comparable to liberalism, clericalism, or socialism. It has always been a mixture of variously structured stereotypes that, in the environment of forming a fully fledged nation, had the ability to attach itself to any political ideology and live with it in a symbiotic or parasitic relationship. (Translated by Libora Oates-Indruchova.)

See also: Clan; Cold War, The; Conservatism: Historical Aspects; Environmental History; Ethnicity and Ethnic Groups: Historical Aspects; French Revolution, The; Military History; National Socialism and Fascism; Nationalism, Historical Aspects of: Arab World; Nationalism, Sociology of; Nationalism: General; Nations and Nation-States in History; Power in Society; Rationality in Society; Religion: Nationalism and Identity; Revolutions, History of; State Formation, Theory of; State, History of.

Bibliography

Anderson, B.R., 1983. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London.

Armstrong, J.A., 1982. Nations before Nationalism. University of Carolina Press, Chapel Hill, NC.

Balibar, E., Wallerstein, I., 1991. Race, Nation, Class: Ambiguous Identities. Verso, London, New York.

Bauer, O., 1907. Die Nationalitatenfrage und die Sozialdemokratie (The Issue of Nationality and Social Democracy). Verlag der Wiener Volksbuchhandlung, Vienna.

Berger, S., Lorenz, Ch., 2008. The Contested Nation. Ethnicity, Class, Religion and Gender in National Histories. Palgrave Macmillan, Basingstoke [England]; New York.

Billig, M., 1997. Banal Nationalism. Sage, London.

Breuilly, J., 1982. Nationalism and the State. St. Martin's Press, New York.

Brubaker, R., 1996. Nationalism Reframed: Nationhood and the National Question in the New Europe. Cambridge University Press, Cambridge, NY.

Calhoun, C., 2007. Nations Matter: Culture, History, and the Cosmopolitan Dream. Routledge, London, New York.

Connor, W., 1994. Ethnonationalism: The Quest for Understanding. Princeton University Press, Princeton, NJ.

Committee on International Relations, Group for the Advancement of Psychiatry, 1987. Us and Them: The Psychology of Ethnonationalism. Brunner/Mazel, New York.

Deutsch, K.W., 1953. Nationalism and Social Communication: An Inquiry into the Foundations of Nationality. Technology Press of the Massachusetts Institute of Technology, Cambridge, MA.

Doob, L.W., 1964. Patriotism and Nationalism: Their Psychological Foundations. Yale University Press, New Haven, CT.

Elgenius, G., 2011. Symbols of Nations and Nationalism: Celebrating Nationhood. Palgrave Macmillan, Basingstoke, Hampshire, UK.

Gellner, E., 1983. Nations and Nationalism. Basil Blackwell, London.

Gellner, E., 1997. Nationalism. Weidenfels & Nicolson, London.

Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.

Giddens, A., 1987. The nation-states and violence. In: Giddens, A. (Ed.), Social Theory and Modern Sociology. Stanford University Press, Stanford, CA.

Greenfeld, L., 1992. Nationalism: Five Roads to Modernity. Harvard University Press, Cambridge, MA.

Greenfeld, L., 2006. Modernity and nationalism. In: Delanty, G., Kumar, K. (Eds.), The Sage Handbook of Nations and Nationalism. Sage, London.

Guibernau, M., 2007. The Identity of Nations. Polity Press, Cambridge.

Hayes, C.J.H., 1926. Essays on Nationalism. Macmillan, New York.

Hayes, C.J.H., 1931. The Historical Evolution of Modern Nationalism. R R Smith, Inc., New York.

Hechter, M., 1975. Internal Colonialism: The Celtic Fringe in British National Development 1536–1966. Routledge & Kegan Paul, London.

Hobsbawm, E.J., 1990. Nations and Nationalism since 1780: Programme, Myth, Reality. Cambridge University Press, Cambridge, NY.

Hogan, P.C., 2009. Understanding Nationalism. Ohio University Press, Columbus.

Horowitz, D.L., 1985. Ethnic Groups in Conflict. University of California Press, Los Angeles.

Hroch, M., 1985. Social Preconditions of National Revival in Europe: A Comparative Analysis of the Social Composition of Patriotic Groups among the Smaller European Nations. Cambridge University Press, Cambridge, NY.

James, P., 1996. Nation Formation: Towards a Theory of Abstract Community. Sage, London.

Kedourie, E., 1960. Nationalism. Hutchinson, London.

Kemilainen, A., 1964. Nationalism: Problems Concerning the Word, the Concept and Classification. Studia Historica Jyvaskylansia, vol. III. Jyvaskyla, Finland.

Kohn, H., 1944. The Idea of Nationalism: A Study in Its Origins and Background. Macmillan, New York.

Laitin, D.D., 2007. Nations, States, and Violence. Oxford University Press, Oxford.

Langewiesche, D., 2003. Der Krieg in den Grundungsmythen europaischer Nationen und der USA (The War in the Foundings-Myths of European nations and USA). Campus, Frankfurt M.

Leerssen, J., 2006. National Thought in Europe. Amsterdam University Press, Amsterdam.

Lemberg, E., 1964. Nationalismus, vols. I and II. Rowohlt Reinbeck, Hamburg, Germany.

Llobera, J.R., 1994. The God of Modernity: The Development of Nationalism in Western Europe. Providence, Oxford.

Mann, M., 1988. States, War and Capitalism: Studies in Political Sociology. Blackwell, Oxford, UK.

Meinecke, F., 1907/1911. Weltburgertum und Nationalstaat (The Weltburgertum and the Nation-State). R. Oldenbourg, Munchen-Berlin, Germany.

Mosse, G.L., 1973. The Nationalization of the Masses. Howard Fertig, New York.

Nairn, T., 1981. The Break-Up of Britain: Crisis and Neonationalism, second exp. ed. NL Bond-Verso, London.

Nora, P., 1986. Les Lieux de Me_moire [The Locations of Memory]. La Nation II (1–3). Gallimard, Paris.

Norman, W., 2006. Negotiating Nationalism. Nation Building, Federalism and Secession in the Multinational State. Oxford University Press, Oxford, New York.

Ozkirimli, U., 2010. Theories of Nationalism. A Critical Introduction. Palgrave Macmillan, New York.

Renan, E., 1888 (1947). Qu'est-ce Qu'une Nation? (What Is a Nation?). In: Oeuvres Completes I. Calman Levy, Paris.

Samuel, R. (Ed.), 1989. Patriotism: the Making and Unmaking of British National Identity. Routledge, London and New York.

Schieder, T., 1991. Nationalismus und Nationalstaat: Studien zum nationalen Problem im modernen Europa (Nationalism and nationstate: studies of the national problem in modern Europe). Vandenhoeck & Ruprecht, Gottingen, Germany.

Schopflin, H., 2000. Nations, Identity, Power. Hurst, London.

Schulze, H., 1996. Staat und Nation in der europaischen Geschichte (State and Nation in European History). Verlag C.H. Beck, Munich, Germany.

Seton-Watson, H., 1977. Nations and States: An Inquiry into the Origins of Nations and the Politics of Nationalism. Methuen, London.

Shafer, B.C., 1955. Nationalism: Myth and Reality. Harcourt, Brace and Comp., New York.

Smith, A.D., 1986. The Ethnic Origins of Nations. Blackwell, Oxford, UK.

Smith, A.D., 1991. National Identity. Penguin Books, Harmondsworth, UK.

Smith, A.D., 1998. Nationalism and Modernism: A Critical Survey of Recent Theories of Nations and Nationalism. Routledge, London and New York.

Smith, A.D., 2009. Ethno-Symbolism and Nationalism. Routledge, New York.

Snyder, L.L., 1954. The Meaning of Nationalism. Rutgers University Press, New Brunswick, NJ.

Snyder, L.L., 1976. Varieties of Nationalism: A Comparative Study. Dryden Press, Hinsdale, IL.

Sorokin, P., 1947. Society, Culture and Personality: Their Structure and Dynamics, A System of General Sociology. Harper, New York.

Stokes, G., 1974. Cognition and the function of nationalism. Journal of Interdisciplinary History 4, 525–542.

Tagil, S. (Ed.), 1984. Regions in Upheaval. Lund Studies in International History, vol. 22. Essette Studium, Solna, Sweden.

Vincent, A., 2002. Nationalism and Particularity. Cambridge University Press, Cambridge, New York.

Vogler, C.M., 1985. The Nation State: The Neglected Dimension of Class. Brookfield Gower, Aldershot.

Nationalism, Sociology of

Veljko Vujacic, Oberlin College, Oberlin, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The sociology of nationalism is a fairly new field. Even so, several theoretical approaches to the study of nationalism have emerged: functionalism, modernization theory, neo-Marxism, comparative–historical sociology, social constructionist theories, and rational choice theory. Despite theoretical disagreements about the origins and evolution of nationalism, a degree of consensus has developed: the rise of the nation was intertwined with the emergence of the absolutist state, the idea of popular sovereignty, industrialization, the expansion of cultural communication, and uneven development. The intimate relation between modernity and nationalism contradicts liberal and Marxist expectations of a unified world.

The study of nationalism is a fairly recent development in sociology, a discipline which took for granted the nation-state as the territorial framework of its main subject of investigation – modern societies. Nevertheless, several sociological approaches to the study of nationalism can be distinguished: structural functionalism, modernization theory, neo-Marxism, rational choice theory, and a variety of comparative–historical and social constructionist approaches.

One recurrent theme in sociological debates about nationalism concerns the nature of the relation (historical and conceptual) among several interrelated phenomena: ethnic group, nation, and nationalism. Each one of these phenomena has been seen either as an historical antecedent of the others or somehow accorded priority. Thus, while one of the most influential theorists argues that 'nationalism created nations' (Gellner), another sees nationalism as a 'subjective reflection' of the group reality of the nation on the level of social consciousness (Hroch). In addition, whereas some emphasize the 'ethnic origins of nations' (Smith), others argue that 'ethnic nationalism' is a later-day ideological development (Greenfeld). A second line of sociological debate has separated modernization theorists and neo-Marxists who explain the emergence of nationalism by reference to economic factors or the functional requirements of industrial society, and historical sociologists who question this view on a variety of grounds. An important obstacle to sociology in this field is imposed by the subject itself: as the most salient form of particularism in the modern world, nationalism defies easy generalization.

Nationalism and Classical Sociology

Nationalism was not a central concern in classical sociology. This curious lacuna must be understood in light of the history of a new discipline preoccupied with demarcating its own field of study – society as a reality *sui generis* (Durkheim). A second reason might be sought in the intellectual bias of sociology itself: political phenomena were seen as derivative of more fundamental social realities. Finally, sociological interest was consumed by the dramatic transformations caused by industrialism and capitalism. As a result, concentration on the problems of industrial society overshadowed the concern with its overall territorial framework – the modern nation-state. The rise of imperialism and mass politics at the turn of the century changed this equation.

Max Weber

Weber's views on the nation must be understood in this historical context (Weber, 1978). For him, the unprecedented fusion of culture and politics in the modern world offers the main clue for understanding the nation: its *differentia specifica* as a social group lies in the striving for territorial political power on the basis of a shared culture. This symbiosis of culture and power in the nation-state resulted from a twofold dependency: if the state's legitimacy increasingly depended on nationalist appeals in the age of mass politics, the nation needed the state for the protection of its unique culture. By combining the status concerns of intellectual strata as the bearers of national culture with the Realpolitik ambitions of political elites, nationalism provided the state with a new source of legitimacy, greatly increasing its mobilization potential. The sacralization of the state as the guardian of cultural values, in turn, inevitably transformed national conflicts, both within and between states, into 'a struggle of life and death.'

Second, Weber attempted to explain national solidarity. In contrast to ethnic groups whose members must share a 'subjective belief in common descent,' national solidarity does not depend on a myth of common origin (Switzerland) or objective markers of status differentiation (race, religion, language). Thus, the loyalty of German-speaking Alsatians to France was derived from the historical experience of a revolutionary regime that abolished feudalism. Conversely, the ethnic and linguistic affinity of Serbs and Croats did not obliterate their national differences. Consequently, Weber argued that common historical experiences and shared memories play a decisive role in forging the nation as a community of 'political destiny.'

Weber's third point concerns the mass appeal of nationalism: as the only source of status superiority available to the masses in a secularized world (ethnic or national honor), nationalism cuts across class lines in defiance of socialist expectations. An additional reason for nationalism's appeal to the lower strata has to do with status compensation: In many instances, the denigration of 'inferior' ethnic or racial groups by the lower classes of the dominant group (such as the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32100-6

emancipated black Americans by the poor whites in the Southern United States) is the very precondition of the latter's status superiority. These observations do not amount to a theory of nationalism; however, Weber's main intuition – that the fusion of culture and power constituted the *differentia specifica* of the nation as a new type of sociological community – became a cornerstone of all later theories of nationalism.

Emile Durkheim and Marcel Mauss

In *The Elementary Forms of Religious Life*, Durkheim notes the similarity between the collective effervescence of religious rituals and revolutionary-patriotic festivals: in the Jacobin phase of the French Revolution, the secular values of Fatherland, Liberty, and Reason became the core values of a revolutionary-patriotic religion whose object of worship was society itself (Durkheim, 1995). For Durkheim, this self-worship of society was another confirmation of his theory of religion. The connection with nationalism as a 'secular faith' remained unexplored in his work. On the other hand, as a committed advocate of secular schools in the Third Republic, Durkheim envisaged an important role for civic patriotism in moral education. During World War 1, Durkheim favorably contrasted this ideal of civic patriotism to expansionist German nationalism, whose excesses he interpreted as a sign of collective *anomie*.

A Durkheimian view of the nation was elaborated further by Mauss (1920). Mauss saw nationalism as a modern phenomenon originating in the doctrine of popular sovereignty espoused by the American and French revolutionaries. Sociologically, nationalism's success as a movement was predicated on the presence of three general factors: (1) a strong central power that successfully broke down traditional affiliations of tribe, clan, and estate; (2) a morally integrated society consisting of citizens who identified with the state and its laws; and (3) economic unity. On the other hand, the belief in a unique racial or civilizational heritage was a retrospective elaboration of nationalists: its importance lay in cementing an already existing social bond on the level of consciousness. For Durkheim and Mauss, the nation was first and foremost a community of citizens; their understanding of ethnic nationalism was limited.

Structural Functionalism

A new sociological interest in nationalism developed as a result of the proliferation of independent states after World War 2. The dominant sociological school of the postwar decades – structural functionalism – viewed nationalism as a by-product of the structural strains caused by industrialization.

Thus, Smelser saw nationalism as a response to the state of collective *anomie* induced by structural differentiation (division of labor, role specialization, etc.). As a secular ideology of industrialization, nationalism overcomes the resistance of traditional structures, mobilizes new constituencies, and gives meaning to the sacrifices required for social change. Conversely, once differentiation proceeds apace, nationalism can take the form of a nativist reaction against the corroding influences of further modernization (Smelser, 1966).

This logic was generalized by Shils, who built on Durkheim's idea about secular nationalism as the functional equivalent of religion: the transition from loosely integrated traditional societies to the modern nation-state is predicated on the gradual penetration of the various social and geographical peripheries (kinship, ethnic, religious, and regional) by the 'charismatic center,' seen as the embodiment of a society's core values (Shils, 1975). By allowing the members of society to partake in the institutionalized charisma of a sacralized political center through the exercise of citizenship rights, the nation-state succeeds in breaking the power of tradition. This sacralization of the secular is of special significance in transitional societies. By transposing the charisma of the tribe or religion onto the nation, the leaders of new states are able to attract mass support and legitimate social change.

Functionalist explanations can be criticized on the following grounds: (1) there is no inevitable correlation between structural differentiation and nationalism; (2) teleology: the end-state is posited by the observer and used to explain the past; (3) the desirability of meeting select functional goals is subject to elite contest; and (4) nationalism is a secular ideology more similar to other modern ideologies than to religion.

Modernization Theory

A significantly modified functionalist modernization approach was developed by Karl Deutsch and Ernst Gellner. The formulations of these two theorists have remained influential to this day.

Karl Deutsch

Deutsch (1966) begins with a distinction between society and culture. Society refers to the division of labor and social stratification; culture, on the other hand, is a set of 'stable, habitual preferences' based on values inculcated through socialization. Societies 'produce, select, and channel goods and services'; culture stores information about the past and recombines it in the present, enabling political authority to function. The technological changes and complex division of labor induced by the industrial revolution resulted in a growing need for expanding channels of cultural communication. Simultaneously, the breakdown of traditional status barriers and exposure to the insecurity of the market raised the aspirations of the socially mobilized masses, already agitated by democratic ideas. These developments favored the emergence of the nation as a horizontal functional group linked by complementary communication channels, typically language (e.g., Czechs versus Germans in Bohemia), but also shared memories or historical experiences (e.g., Switzerland). As *the* ideology of popular sovereignty, nationalism held a special attraction for the socially mobilized masses whose claims to status equality and upward social mobility it helped justify.

Deutsch's theory is empirical and predictive: the complementary character of communication channels can be assessed by social-psychological indicators; rates of social mobilization can be measured by urbanization, mass literacy, newspaper circulation, etc.; and assimilation depends on the extent and content of communication between in and out-groups. The theory runs counter to the standard predictions of Parsonian

structural functionalism: the expansion of communication strengthens nationalism, not universalism.

Ernst Gellner

Gellner (1983) starts with an ideal type of agrarian societies; their central feature is an ideologically buttressed functional division of labor that separates the 'high culture' of a hereditary administrative-military ruling class and universal clerisy from the 'low cultures' of socially isolated and illiterate peasant communities. This status segregation has a cognitive dimension as well: the world is experienced as culturally and ontologically heterogeneous.

A profound change occurs with the modern cognitive revolution. The world is now seen as a coherent whole subject to universal laws expressed in a unitary linguistic idiom. The social correlate of this cognitive revolution is industrialism. If intellectual progress presupposes the perpetual exploration of reality, the idea of unlimited growth demands the constant redrawing of traditional social boundaries and roles in line with the functional requirements of the division of labor. These requirements can be met only by a common linguistic idiom transmitted through standardized education. Literacy in a shared language prepares individuals for new functional roles, increases their mobility prospects, and facilitates communication among strangers in an impersonal world. The emergence of a shared culture favors nationalism as a political principle which holds that the state must rest on the foundation of national culture; the state, in turn, acquires a new source of legitimacy (cf. Weber). Thus, industrialism is a necessary condition of nationalism. Ethnicity is secondary; it is nationalism that invents nations.

Finally, nationalism disseminates through uneven development. In a typical imperial (e.g., Habsburg) situation, an ethnic division of labor is present; the carriers of high culture belong to one ethnic group (or groups), those of low culture to another (or others). The social exclusion of the aspiring intelligentsia of the subordinate group leads it to adopt ethnic nationalism as a strategy of collective mobility (e.g., Czech nationalism). In contrast, 'diaspora nationalism' develops among politically excluded high culture pariah groups (e.g., Jews). The ideal case is a mature industrial society in which elites and masses share a standardized idiom and consider themselves to be conationals in a common state.

Gellner's elegant theory has enjoyed enormous influence. It can be criticized on several grounds: (1) the theory does not specify a causal (group) agent, nationalism is seen as a byproduct of the impersonal process of industrialization; (2) nationalism can precede industrialization (e.g., Balkan nationalism); (3) nations can precede nationalism; and (4) nations are rooted in premodern, ethnic identities (see Section Comparative–Historical Sociology).

Neo-Marxism

The classics of Marxism failed to develop a consistent sociological theory of the nation (Szporluk, 1988). Although partially filled by Austro-Marxism (Bauer, Renner) and the Bolshevik ideologists (Lenin, Stalin), this lacuna was addressed seriously only after World War 2. A strong impetus to the Marxist sociology of nationalism was given by Immanuel Wallerstein's world system theory: in the underdeveloped periphery of the world capitalist system, class struggle (indigenous proletariat and peasantry versus the comprador bourgeoisie) and national conflict (the exploited nation versus the world capitalist metropole) were superimposed upon each other.

Class and Peripheral Nationalism

Hechter (1975) extended the metropole–periphery logic to ethnic conflict in the British Isles. In defiance of functionalist logic, the incorporation of the (Celtic) periphery by the (English) center did not result in the creation of a homogenous national society, but a dynamic of 'internal colonialism' instead. The subordination of the underdeveloped periphery was ensured by its economic dependence on a developed center and a rigid ethnic division of labor that enabled the English to monopolize power and status privileges. The persistence of ascriptive status identities and the superimposition of class and ethnic conflicts run counter to the expectations of modernization theory: instead of integration and consensus, uneven development leads to the resurgence of peripheral nationalism.

Studying the same case, Nairn (1977) came to a different conclusion: peripheral 'neonationalism' could be caused by 'relative overdevelopment' as well. Thus, the resurgence of Scottish neonationalism was the result of the relative economic decline of the English center and new opportunities opened by oil exploitation in the North Sea. A newly awakened Scottish bourgeoisie in search of profit appealed to nationalism in its struggle against a parasitic center. In larger historical perspective, nationalism was a product of the ideological reaction of the awakened intelligentsia of the semiperiphery to uneven development and imperialist exploitation of their countries (not yet nations) by the centers of world capitalism (England and France). The nationalism of the center, in turn, developed in 'dialectical counter-mobilization' against its semiperipheral counterparts.

Class and Nation: The Social Base of Nationalist Movements

The relationship between class and nation is a central concern in Hroch's (1985) study of the developmental phases and social base of nineteenth-century national movements in small European nations. For Hroch, the nation is an 'objective social category' rooted in the memory of a shared ethnic past, dense linguistic and sociocultural ties, and a conception of citizenship. National awareness and nationalism are only a 'subjective reflection' of this reality on the level of consciousness (contrast Gellner).

Hroch distinguished three phases in the development of national movements: in phase A (ethnographic), linguists and folklorists rediscover a shared past; in phase B (agitation), intellectuals promote national awareness among the masses; finally, in phase C, a mass national movement arises. While the social base of nationalist movements varied, especially in phase B (teachers in Slovakia; petty bourgeoisie in Norway), the social mobilization of the bourgeoisie and peasantry was a necessary condition for success in phase C: an independent nation

requires 'a fully developed class structure.' Another necessary condition of nationalist movements was the superimposition of ethnic and class conflicts: only when membership in a small nation appeared as a decisive disadvantage could 'social contradictions' be expressed in national terms. Hroch did not conflate class and nation; however, he did argue that a successful national movement must be based on distinct class interests.

Neo-Marxist approaches can be criticized for (1) an unwarranted reduction of national to class conflicts; (2) limited generalizability (internal colonialism); and (3) an inability to explain the nationalism of the metropole of the world system (England, France).

Comparative–Historical Sociology

Historical sociologists have questioned functionalist and Marxist explanations in two ways: (1) by emphasizing the ethnic or early modern roots of nationalism and (2) by demonstrating that nationalism predates industrialization.

Ethnic Groups and Nations

The first type of argument is developed by Smith (1986), who builds on the pioneering work of Armstrong (1982) on 'nations before nationalism.' Smith rejects the view that nations are based on primordial ties (e.g., race); however, they *are* rooted in premodern *ethnie*. *Ethnie*, in turn, are not based on ascriptive criteria, but on intergenerationally transmitted 'constitutive myths' (ethnic *mythomoteurs*). Typically, such constitutive myths include or imply a collective name, a myth of common descent, a shared historical narrative (legends, epics), belief in a unique culture, association with a territory, and solidarity vis-à-vis out-groups. The transformation from *ethnie* to nations occurred under the impact of the triple revolution of administrative centralization, capitalism, and 'cultural coordination.' Capitalism helped bind social classes into a unified economic community; political centralization, warfare, and cultural coordination (the rise of vernacular languages and mass culture) forged the link between state and nation. Whereas nation-building paths varied in accordance with historical timing, religious identifications, the character of the constitutive myth, and its elite carriers (aristocracy, bourgeoisie, intelligentsia), two broad trajectories emerged as a result of uneven development (1) the early Western path in which the nation was defined as a territorial community of citizens bound by laws and (2) the late 'German' (East-European) route in which *ethnie* served as the basis for statehood. However, Western nations also required a constitutive myth: premodern cultural identities were crucial everywhere, and industrialization was only one factor that helped transform *ethnie* into nations.

Nationalism as Ideology

Another approach is exemplified by Greenfeld (1992): nationalism is an antecedent of industrialization. Greenfeld traces the idea of the nation to a long process of semantic transformation: originally applied to groups of students in medieval universities and the elite of a 'community of opinion,' in sixteenth-century England 'the nation' was extended to include the whole people. This momentous change from the medieval (people = rabble) to the modern *Weltanschauung* (people = elite) was caused by the structural crisis of feudal society. In the first nation, England, the decimation of the aristocracy and the decline of the clergy (War of the Roses; separation from Rome) shattered traditional status barriers, opening channels of social mobility to gentry and educated commoners. The profound status inconsistency felt by members of the elite led them to discover the nation: the view that status inequality was secondary to membership in a community of citizens resolved their identity crisis and legitimized social mobility across status lines. English nationalism culminated in popular sovereignty and an individualistic and civic ideological self-definition: the nation was a composite of individuals, and membership was open to all citizens (likewise the United States).

The diffusion of nationalism was a consequence of England's hegemonic status. Thus, the eighteenth-century French nobility, squeezed between the absolute monarchy and wealthy commoners, imitated its English counterpart and competitor, finding in nationalism a solution for its own identity crisis. However, while membership in the French nation was open to all (as in civic nationalism), the nation came to be seen as expressing a single collective will (*volonté générale*); thus, French nationalism was collectivist and civic. A further semantic transformation occurred in Germany and Russia, where the term was associated with ethnicity – the unique character of a people; nationalism became both collectivist and ethnic. The motor force behind this change was *ressentiment* – the profound existential envy felt by the elites of follower societies toward the 'superior' West; hence the compensatory glorification of the *Volk* with all its authoritarian implications.

Subsequently, Greenfeld (2001) extended this theoretical framework to the relationship between nationalism and economic growth. Instead of Weber's proverbial Protestant ethic, Greenfeld sees nationalism as the main cultural source of 'the spirit of capitalism,' i.e., the modern orientation to sustained economic growth. As an inherently egalitarian ideology, nationalism breaks down the status barriers of traditional societies, opens the doors to social mobility (careers open to talent), elevates 'traditionally disparaged occupations' (e.g., trade), and legitimizes profit-making activity (Greenfeld, 2001: pp. 22–23). In the pioneering case of England and other civic-individualist nationalisms (e.g., the United States), economic achievement was defined as a positive national value because it empowered individuals who constituted the nation. However, even in the case of England, the initial emphasis of economic nationalism was on the common good of the nation as a whole rather than on the relationship between the public good and private interests of constituent members. The idea of a 'natural' *homo oeconomicus* was a later development, predicated on the gradual emancipation of the economic sphere from religion, politics, and status considerations, and its reconstitution as an independently legitimate domain of social action. Subsequently, economic nationalism developed as a reaction to the contagious effect of the English model and in competition with it; in nations like Germany or Japan, economic growth was

harnessed to the political emancipation of the nation from foreign domination and propelled by state action. Where economic achievement, however, was not seen as an important national value (e.g., Russia), nationalism was not conducive to economic growth.

Greenfeld's conclusions are original: (1) democracy emerged from individualist and civic nationalisms; (2) ethnic nationalism is a later ideological elaboration; (3) the historical carrier of nationalism is not the bourgeoisie, but aristocracy or intelligentsia (Germany); (4) both capitalism and democracy could flourish only once the idea of the nation delegitimized traditional status barriers; and (5) nationalism is the main source of 'the spirit of capitalism.'

Greenfeld's theory avoids materialism and teleological functionalism; however, it can be criticized on both theoretical and empirical grounds: (1) the dichotomy between civic (Western: whether individualist or collectivist) and ethnic collectivist (Eastern) nationalisms is far from clear-cut, as Western nations were also based on an ethno-symbolic foundation (Smith, 2009), while cultural nationalism can cut across the civic–ethnic divide (Hutchinson, 1987); (2) early national consciousness should not be equated with a fully developed nationalism: English nationalism developed considerably later than the sixteenth century (Kumar, 2003); (3) cultural determinism: some nations (the United States, England, France) appear destined for democracy, while others seem doomed to authoritarianism (Germany, Russia); and (4) nationalist ideology may be a necessary precondition for legitimizing social mobility across status barriers and enabling capitalist development, but it can hardly explain economic growth.

Nationalism as Cultural Idiom

The continued salience of civic versus ethnic national self-definitions is documented by Brubaker (1992). In France, the republican tradition defined the nation as a territorial 'community of citizens'; membership was not tied to ethnicity. In Germany, a territorially divided nation was defined as a 'community of descent': all (and only) Germans were members. The persistence of these cultural idioms (not instrumental interests) helps explain the difference in immigration policies: the French one became assimilationist and inclusive, the German one ethnic and exclusive.

Social Constructionist Approaches

Nation as Invented Tradition

The socially constructed character of all collective and, *ipso facto*, national identity is emphasized by a variety of contemporary authors. One influential version of the argument is developed by Hobsbawm and Ranger (1983): the nation is largely a product of 'invented tradition.' By 'invented tradition' is meant "a set of practices, normally governed by overtly or tacitly accepted rules and of a ritual or symbolic nature, which seek to inculcate certain values and norms of behavior by repetition, which automatically implies continuity with the past" (Hobsbawm and Ranger, 1983: p. 1). The modern nation is a particularly good example of invented tradition: although the nation-state with its well-demarcated territory and standardized national language is an eminently modern phenomenon, nationalists typically claim it be a 'natural community' rooted in the 'remotest antiquity.' In actual fact, national symbols, flags, languages, and histories are a product of nationalist social engineering, while the connection with the remote past is retrospectively established by nationalist imagination. It follows from this that the main task of the historian and sociologist is to study those symbolic and discursive practices that helped establish the particular nation's idea of itself as a natural community in the first place.

Nation as Imagined Community

A related argument is developed by Anderson (1991) who sees the nation as a new type of limited and sovereign 'imagined community.' The nation is an imagined community because most of its members will never know each other; it is imagined as a limited community because it is externally bounded by other nations; and it is seen sovereign because it is historically rooted in the idea of internal and external freedom (popular rule and national independence). Finally, the nation is imagined as a community, because it is 'always conceived as a deep, horizontal comradeship' (Anderson, 1991: p. 7). The emotional resonance of the nation stems from the fact that it replaces kin and religion as the main locus of individual loyalty and self-sacrifice in the modern era.

According to Anderson, the emergence of the nation as a new type of imagined community was predicated on the presence of certain definite cultural and material conditions: (1) linguistic diversity; (2) the increasing relevance of vernacular languages and literacy after the Protestant Reformation; and (3) "print capitalism" – the production of books, newspapers, and journals for mass cultural consumption. Once set in motion, print capitalism gave an unprecedented stimulus to the development of nationalist discourse, albeit under a variety of historical circumstances that require separate study.

Social constructionists correctly attacked the fallacy of group realism. However, their views can be subjected to the following criticisms: (1) while the nation might be invented, this social construction must be based on a preexisting reality, be it the memory of historical statehood (e.g., Croatia, Poland), ethnicity (e.g., Germany), or a more complex combination of various factors (e.g., Austria); (2) in view of the above, modern national identity might be 'invented' but it is not 'fabricated'; (3) political factors (warfare, state centralization) and elite (or class) interests played an equally important role in the emergence of nationalism as 'print capitalism'; and (4) the notion of 'imagined community' applies to other kinds of communities besides the nation (e.g., class, religious denomination). Such objections provided an impetus for revisions of the social constructionist position.

Nationhood as Institutionalized Form, Practical Category, and Cognitive Scheme

Brubaker (1996) challenged the fallacy of sociological "group realism" from the standpoint of the new sociological institutionalism: instead of treating the nation as a "substantial entity, collectivity, or community" that can be objectively defined, sociologists should study 'nationhood' from the point of view of its political and cultural institutionalization, both within and

among states. Thus, the dramatic collapse of the Soviet Union has to be understood as a consequence of the ethnoterritorial institutionalization of nationhood both on the group (ethnic federalism) and individual (personal identity document) levels. Such institutionalization turned ethnicity into an 'organizing principle of social classification,' constituting the interests of social groups and individuals, serving as a basis for political action, and informing practical conduct in everyday life. Thus, nationalism is not so much 'engendered by nations' as it is induced by particular institutional arrangements.

In a further elaboration, Brubaker (2006) argued that ethnic groups, races, or nations should not be thought of as objective entities but rather as 'ways of perceiving, interpreting, and representing the social world,' i.e., as 'practical categories, situated actions, cultural idioms, cognitive schemes, discursive frames, organizational routines, institutional forms, political projects, and contingent events.' This emphasis on ethnicity and nationhood as cognitive frame, social process, and relational category is designed to prevent sociologists from reifying ethnicity while simultaneously avoiding the reductionist pitfalls of methodological individualism. Cognitive approaches can help explain how ethnicity and nationhood are constructed through actions in everyday life, thus avoiding the traditional preoccupation in social constructionist approaches with elite manipulation of nationalist discourse and symbols.

Nationalism as Discourse, Narrative, and Event: Religion and Nationalism

Gorski (2000) challenges the idea that secular nationalism is a functional substitute for traditional religion by arguing that Dutch national consciousness arose in the early modern era on the basis of a Calvinist-inspired 'Hebraic nationalism.' In the course of subsequent religious conflict and the Dutch revolt against Spain (1555–1609) a full-blown nationalism developed: a sense of the cultural uniqueness of the *Netherlandish* nation (religion, language, custom) blended with political revolt, social mobilization (vernacular literacy), and emerging ideas of popular sovereignty (the state should emanate from the 'people/nation'). An analogous development took place in England, where the commemoration of Protestant martyrs was associated with the idea of England as an 'elect nation.' Both in England and in the Netherlands, however, 'Hebraic nationalism' competed with rival nationalist discourses (e.g., republican, monarchist). The distinction between early modern protonationalism and modern nationalism thus is not as sharp as modernist theorists posit. Consequently, historical sociology should focus on the various 'threads' (historically bound discourses) and 'fibers' (nationalist narratives) from which the whole 'fabric of nationalist discourse is composed.' Nationalism is best defined "as any set of discourses or practices that invoke the nation" (Gorski, 2000: p. 1461).

This view is seconded by Zubrzycki (2006): the originally civic nationalism of Poland's 'republic of nobles' was ethnicized as a result of the nation's territorial dismemberment in the late eighteenth century. Religious subjection, shared language, and the appropriation of religious symbols by romantic nationalist intellectuals resulted in the identification of Polish nationality with Catholicism (Poland as 'the Christ of nations'). Religious self-identification thus became integral to national identity at a later date, and nationalism did not become fully secular as modernist theories posit. Once cemented by recurrent historical experiences (e.g., subjection under Soviet communism), nationalist 'narrative structures' and accompanying symbols (the cross; the Black Madonna of Czestochowa) came to constitute the nation as a 'category of practice.' Even so, the nation remains a 'daily plebiscite' (Renan): the salience, meaning, and use of nationalist discourse (including symbols and rituals) vary across time (political context), while contentious events play an important role in evoking 'collective representations' of the nation. This 'eventful sociology' defines the nation as a "discursive space constituted by a web of representations articulated by various communities of discourse around specific events, and constituted also through the practical uses of symbols commonly understood as 'national'" (Zubrzycki, 2006: p. 216).

Rational Choice Theory

Avoiding grand debates, rational choice theory attacks the fallacy of sociological group realism from the point of view of methodological individualism: how do nationalist movements overcome the free rider problem? A good illustration is Hechter's (1987) analysis of nationalism in Quebec: regional managerial positions served as a selective incentive for the Francophone middle class. Once the spoils of office were distributed, however, nationalist politics abated. Consequently, in addition to offering private goods, nationalist elites must prevent the defection of followers. From these two conditions (selective incentives, internal control), a series of predictive hypotheses emerges. Another illustration is offered by Laitin's explanation of variations in the linguistic assimilation of Russian-speaking minorities in four different post-Soviet republics by the 'linguistic calculus' of individuals faced with the prospect of assimilation: economic returns, the status costs of defecting from the in-group, and the degree of out-group openness emerge as key predictors of the likelihood of assimilation. Thus, cultural choices are shaped by rational calculations about material rewards and social honor (Laitin, 1998).

Rational choice theories of nationalism can be criticized on the following grounds: (1) collective ideologies like nationalism cannot be reduced to the aggregate actions of individuals; (2) value choices are not 'mere' preferences: the intergenerational transmission of myths, collective memories, and narratives shapes the outlook of members of the nation and constrains choice; and (3) nationalist appeals contain a powerful symbolic (irrational) dimension that is critical for explaining the intensity of national feelings and nationalist mobilization.

Conclusion

Despite the fact that nationalism did not stand at the center of sociological interest, especially in the classical epoch of the discipline, sociologists have studied the topic from a variety of perspectives. Even so, a considerable area of agreement has emerged. For one, sociologists agree that the nation is not based on primordial ties; rather its emergence as a community of citizens was related to rise of the modern state (administrative

centralization), the novel idea of popular sovereignty, industrialization, and the expansion of cultural communications (vernacular languages, literacy). Uneven development was crucial for the dissemination of nationalism, whether on account of the ethnic division of labor, the reaction of the exploited periphery, or *ressentiment*. The salience of ethnicity (itself not an ascriptive, but subjective category) was greatest in areas where class and ethnic conflicts were superimposed (e.g., Habsburg empire), the putative nation lacked a state (e.g., Poland), and ideology conceptualized the nation as *Volk* (e.g., Germany). Conversely, ethnicity was secondary in established states (e.g., England), or where shared memories and experiences favored other identifications (e.g., Switzerland). The consolidation of national identity everywhere involved a considerable effort at the invention of tradition. However, the protean character of nationalism as an adaptable and malleable set of ideas, collective representations, symbols, and 'cognitive frames' ensures that the 'making of nations' is a continuous process, albeit one that takes place within the constraints imposed by dominant narratives. The intimate relation between modernity and nationalism defies liberal and Marxist expectations of a unified world; this may well be the most provocative insight of the sociology of nationalism.

These areas of agreement point to considerable progress in the sociology of nationalism. This progress is also manifested in the growing number of areas that sociologists have begun to investigate: religion and nationalism, nationalism and collective violence, nationalism and collective memory, nationalism and gender, the nationalist narrative, nationalism and everyday life, nationalist symbols and rituals, national identity and ideology, dominant and peripheral nationalism, diaspora and homeland nationalism, nationalism and immigration, and nationalism and globalization. Still, many sociologists seem reluctant to fully engage with nationalism; this is understandable, since to an extent all individual nationalisms tell their own stories, and this presents obstacles to generalization. Yet, as Max Weber demonstrated, the uniqueness of a phenomenon neither diminishes its importance nor precludes sociological investigation: a conceptually informed comparative–historical sociology can develop middle-range theories while taking into account the peculiarities of individual instances.

Bibliography

Armstrong, John, 1982. Nations before Nationalism. The University of North Carolina Press, Chapel Hill.

Anderson, Benedict, 1991, (1983). Imagined Communities. Reflections on the Origins and Spread of Nationalism. Verso, London.

Breuilly, John, 1993. Nationalism and the State. University of Chicago Press, Chicago.

Brubaker, Rogers, 1992. Citizenship and Nationhood in France and Germany. Harvard University Press, Cambridge, Mass.

Brubaker, Rogers, 1996. Nationalism Reframed. Nationalism and the National Question in the New Europe. Cambridge, England.

Brubaker, Rogers, 2006. Ethnicity without Groups. Harvard University Press, Cambridge, Mass.

Connor, Walker, 1994. Ethno-nationalism. The Quest for Understanding. Princeton University Press, Princeton, New Jersey.

Delanty, Gerard, Kumar, Krishan (Eds.), 2006. The SAGE Handbook of Nations and Nationalism. SAGE Publications, London, Thousand Oaks, and New Delhi.

Deutsch, Karl, 1966, (1953). Nationalism and Social Communication. MIT Press, Cambridge.

Durkheim, Emile, 1995, (1912). The Elementary Forms of Religious Life. Free Press, New York.

Gellner, Ernest, 1983. Nations and Nationalism. Cornell University Press, Ithaca.

Gorski, Philip S., 2000. The mosaic moment: an early modernist critique of modernist theories of nationalism. American Journal of Sociology 105 (5), 1428–1468.

Greenfeld, Liah, 1992. Nationalism. Five Roads to Modernity. Harvard University Press, Cambridge, Mass.

Greenfeld, Liah, 2001. The Spirit of Capitalism. Nationalism and Economic Growth. Harvard University Press, Cambridge, Mass.

Hall John, A. (Ed.), 1998. The State of the Nation. Ernst Gellner and the Theory of Nationalism. Cambridge University Press, Cambridge, England.

Hechter, Michael, 1975. Internal Colonialism. California Press, Berkeley and Los Angeles.

Hechter, Michael, 1987. Nationalism as group solidarity. Ethnic and Racial Studies 10, 4.

Hobsbawm, Eric, Ranger, Terence (Eds.), 1983. The Invention of Tradition. Cambridge, England.

Hroch, Miroslav, 1985. The Social Preconditions of the National Revival in Europe. Cambridge, England.

Hutchinson, John, 1987. The Dynamics of Cultural Nationalism. The Gaelic Revival and the Creation of the Irish Nation State. Allen & Unwin, London.

Kumar, Krishan, 2003. The Making of English National Identity. Cambridge, England.

Laitin, David, 1998. Identity in Formation. The Russian-Speaking Population in the Near Abroad. Cornell University Press, Ithaca.

Mauss, Marcel, 1969, (1920). La nation, vol. 3. Mauss, Oeuvres, Minuit, Paris, pp. 573–639.

Motyl, Alexander, et al., 2001. Encyclopedia of Nationalism, vol. 2. Academic Press, San Diego and London.

Nairn, Tom, 1977. The Break-Up of Britain. Humanities Press, London.

Shils, Edward, 1975. Center and Periphery. Essays in Macrosociology. Chicago University Press, Chicago.

Smelser, Neil, 1966. Essays in Sociological Explanation. Prentice-Hall, New Jersey.

Smith, Anthony, 1986. The Ethnic Origin of Nations. Basil Blackwell, Oxford.

Smith, Anthony, 1998. Nationalism and Modernism. Routledge, London and New York.

Smith, Anthony, 2009. Ethno-symbolism and Nationalism. Routledge, London and New York.

Szporluk, Roman, 1988. Communism and Nationalism. Karl Marx vs. Friedrich List. Oxford University Press, Oxford and New York.

Weber, Max, 1978. Economy and Society, vol. 2. University of California Press, Berkeley and Los Angeles.

Zubrzycki, Geneviève, 2006. The Crosses of Auschwitz. Nationalism and Religion in Post-communist Poland. University of Chicago Press, Chicago.

Nations and Nation-States in History

John Breuilly, London School of Economics and Political Science, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article begins by considering the debate about the existence and significance of nations and nationalism before the rise of the nation-state in the modern period. It then considers the role of nations and nationalism in the formation of nation-states. This begins with Europe, looking at the distinct processes of state reform, national unification, and separatist nationalism in producing different kinds of nation-states. The role of international relations and war is also considered. The following section turns to nation-state formation beyond Europe, especially the waves of new states following the two World Wars and the collapse of the Soviet Union. The article concludes with some reflections on the future of nations and nation-states.

This survey starts by asking whether nations existed before the establishment of nation-states. It then considers the formation of national states before the development of nationalist movements. The next two sections look at nation-state formation in and beyond Europe around 1850 to 1970. A final section considers the place of nations in a world of nation-states.

Nations before Nation-States?

Nationalists answer this question in the affirmative. In the beginning was the nation. As the nation acquired consciousness of itself, there arose explicit demands and movements (nationalism) seeking autonomy, usually a nation-state. The nationalist sequence is therefore nation > nationalism > nation-state.

Nationalist historiography frequently locates the 'nation' in the distant past. Arminius (Hermann) is celebrated by German nationalists for his victory over Roman legions in AD 9. The feudal kingdoms succeeding the Roman Empire are presented as national: the Western Frankish kingdom equals France and the Anglo-Saxon kingdoms equal England. Nationalist etymologies are full of references to the English, French, Germans, Italians, etc.

This view appears plausible for cases like England with a long continuity of institutions bearing the national name. Nationalist historians of subordinate groups in imperial states or of stateless, nonliterate societies have greater difficulties. Political fragmentation also presents problems, even given evidence of cultural identity (e.g., Italy, Germany). There were some premodern societies with sources that seem to claim an ethnic or religious unity between elites and nonelites against foreign rule (e.g., Jewish resistance to Rome). However, powerful premodern states were characterized by divisions between elites of priests, warriors, and officials, usually serving a monarch, and the mass of subjects. Elites referred to ethnicity but this was an aristocratic discourse and subordinated to religious and dynastic values. Subjects were transferred from one ruler to another without reference to their 'nationality.' Indeed, the puzzle would be if elites in highly stratified and undemocratic societies sought to identify with those they ruled.

A more significant argument is that premodern ethnicity is an important condition for modern national identity (Armstrong, 1982; Hastings, 1997; Smith, 1986). Against this is the view that references to ethnic identity do little more than provide modern nationalists with historical material that they can use politically, although that indeed can be important (Breuilly, 2005). There is general agreement that nationalism and the nation-state are modern. However, just *how* modern; what the relationship is between state, nationalism, and nation-state; and why the nation-state has become the dominant political unit of modern times are matters of debate.

Nation-States before Nationalism

In Europe, around 1500, there was little idea that the political order was only legitimate if national (however, see Hirschi, 2012). Germany and Italy were fragmented. By 1600, some voices called for national unity in Italy and the United Provinces against foreign domination. Nevertheless, political loyalty was overwhelmingly to existing states with their republican (city-state), confessional, and dynastic legitimations.

Dynastic empires ruled over different ethnic groups. Henry VIII described his kingdom as an 'empire,' ruling Wales and Ireland, and claiming parts of France. The Habsburg dynasty ruled Austria, much of Southeastern Europe, Spain, Belgium, the Netherlands, and South America. The Valois and Bourbon dynasties claimed territories beyond those of contemporary France, itself inhabited by different language groups. By the seventeenth century, the Romanovs were extending their rule beyond the core of Russia, west and south into Slav regions and north and east into Asiatic ones. Opposition came not from 'Slavs' or 'Asians,' but from other dynasties, e.g., the elective Kings of Poland (Saxons). The core of the Hohenzollern dynasty, Brandenburg-Prussia, included Polish speakers. The Ottoman Empire ruled Slav populations, Muslims, and Christian.

The national idea was unimportant. Historians write of long-term 'winners' emerging as the core of later nation-states (England, France, Spain, Prussia, Russia, and Piedmont). However, these had no stronger, more consciously 'national' characteristics than long-term 'losers' like Poland–Lithuania,

Burgundy, Brittany, Habsburg Austria, the Ottoman Empire, city-states, and small princedoms (Davies, 2011). Beyond Europe, there was little one could call a national state (arguable exceptions are Ethiopia and Japan). Often, there was no distinct state but systems of lineal rule described as stateless or tribal. Again, dynastic empires ruled over many ethnic groups, e.g., in China and India.

Only after 1500 was the national idea used to legitimize states and political movements. In England, the construction of a common legal and administrative system from the late Anglo-Saxon period, the growing importance of Parliament, island situation, increased emphasis on the vernacular language, an economy and polity increasingly dominated by the growth of London, and the national emphasis of the Protestant Reformation combined to make the national idea salient. Shakespeare in *Henry V* (1599) delineates Irish, Welsh, and Scottish stereotypes, although all support the king (Shakespeare only constructed English stereotypes at an elite level in contrast to other Europeans, e.g., the suitors for Portia's hand in *The Merchant of Venice*). The enemy was France, but hostile stereotypes were confined to the elite, not the people. The play includes Henry wooing a French princess; his marriage and military campaigns are designed to maintain power in France. Ethnic and national ideas are prominent but subordinated to the monarchical principle. If England was the most 'advanced' case, this suggests the marginality of the idea elsewhere in 1600 (on stereotypes, Leerssen, 2006).

Four processes increased the importance of national arguments: confessional conflict, absolutism, new political opposition, and intensification of interstate conflict. All developed alongside an extension of European influence beyond Europe and the growth of a commercial economy that undermined privilege and brought into politics 'middling sorts' of people – financiers, lawyers, merchants, commercial farmers, and manufacturers.

The Reformation challenged the dominance of Papacy, Catholic clergy, and the use of Latin. Translations and mass circulation of the *Bible* into vernacular languages along with efforts to extend literacy, because individuals must know the word of God and take responsibility for their souls rather than leaving it to priests, gave a national flavor to Protestantism. Luther challenged the Italian Pope as a German, and Henry VIII as an Englishman. In Spain and France, the monarchy took control over the Catholic Church, creating a 'national' church. Popular religious movements share affinities with later nationalist movements. Nevertheless, religion mattered more than nationality, and only when combined, usually under princely control, did national language figure prominently in politics.

Princes extended their powers by sale of office, appointment of extraordinary commissioners, legal codification, subordination of church, greater taxation, and the establishment of standing armies, processes summarized as 'absolutism.' There was nothing intrinsically 'national' in this, but where the core territory could be regarded as national, e.g., Spain, England, and France, this was justified by national arguments (Anderson, 1974).

This affected political opposition. By 1620, the English Parliament presented itself as a defender of national freedoms. *Parlements* and provincial estates in eighteenth-century France depicted the crown as despotic and themselves as guardians of national liberties. These arguments could escape beyond elite control. In England in the 1640s, national claims were extended to the 'poorest he' by radicals in the Putney Debates. In France, radical values were projected upon the Estates-General summoned in 1789, soon converted into the 'National Assembly.' Monarchy was redefined as at best the servant of the nation, at worst its enemy (Breuilly, 1993: Chapter 3).

Conflict between states had national consequences. Wars in late seventeenth- and eighteenth-century Europe were 'cabinet wars,' the War of Spanish Succession, Austrian Succession, etc. However, they had a global dimension. The central struggle was between France and Britain for world hegemony, using other states as proxies. The struggle extended to India and North America (Schumann and Schweizer, 2008). The wars mobilized the expanding resources of commercializing, imperial economies. They engaged the attentions of an ever-larger part of the population. Commercial growth was associated with the rise of a print culture. In England, the electorate increased in size. In France, the crown and its opponents resorted to pamphleteering, appealing to 'public opinion,' and presenting political conflicts in national terms.

These national arguments accepted the territory of the state but questioned its legitimacy.

> The Nation is essentially the source of all sovereignty; nor can any individual or any body of men, be entitled to any authority which is not expressly derived from it.
>
> (Article III of the French Declaration of the Rights of Man and Citizen, 1789).

The nation was the body of citizens within the existing state. For radicals this meant all adult males. If ideology was logical, this would mean that the national idea was purely political. However, cultural elements were soon introduced.

The idea of a state occupying a national territory associates national virtues with the state. As the French revolutionary state plunged into crisis, it took up these associations. The ethnic and linguistic characteristics of the regions that rebelled against the Jacobins, let alone of opposing states, were linked to counter-revolution:

> We have observed that the dialects called Bas-Breton, the Basque dialect, and the German and Italian languages have perpetuated the reign of fanaticism and superstition, revered the domination of priests and aristocrats, and favoured the enemies of France.
>
> Speech in the National Assembly by Barère, quoted in Carlton J. Hayes, The Historical Evolution of Nationalism.
>
> (New York, 1931: p. 65)

Opponents of the Jacobins presented the 'true' France as Catholic and monarchical in which the historic provinces and the estates enjoyed autonomies and privileges. Edmund Burke depicted nations as carriers of traditions that bound together generations and inhibited the arbitrary and destructive use of reason to alter institutions.

This greater use of national arguments was reinforced by the making of sharp distinctions between 'state' and 'society' and

between state boundaries. Such changes became apparent during the Napoleonic wars. The French revolution abolished distinctions of privilege in France and exported this principle beyond France. This did not herald an era of freedom and equality. New governments were often more autocratic than their predecessors and political rights were restricted to a small minority. Nevertheless, equality before the law was introduced and a clear distinction was made between the state as public power and society as private. When governments appealed for support, they appealed to that 'whole' society rather than to privileged minorities.

Opponents of Napoleon noted that he could tap popular enthusiasm and linked this to new freedoms achieved in France. Reformers advocated the removal of privilege and the extension of political rights to strengthen support for the state. In politically fragmented regions, some invoked the 'nation' as a unifying and mobilizing principle. The nation was presented as a cultural entity from which should rise nation-states, unlike France where the state created national unity. The idea was weak at first, although stronger in Germany than in Italy, probably because in Germany there were national traditions such as the Holy Roman Empire and two powerful states – Prussia and Austria – which had an interest in appealing to the national idea.

Napoleon's defeat was inflicted principally by the conservative regimes he had not conquered – Russia and Britain – rather than by national movements in areas he had. Even there, dynasties, state patriotism, hatred of the French, religious convictions, and opportunism better explain effective resistance. After 1815, there were attempts to restore privileges and old political arrangements. However, the map of Europe after 1815 was closer to that created by Napoleon than what had existed before. States were fewer and larger, more rationally and bureaucratically organized, with clearer boundaries. Governments could suppress national and liberal oppositions, but not the ideas that those movements embodied. Commercialization, even without industrialization, increased social and territorial mobility, which undermined attempts to restore privileges. All this pressured monarchies to shift toward constitutional forms of rule. New politicians emerged who appealed to the people or the nation (Baycroft and Hewitson, 2006; Colley, 1992; Dann and Dinwiddy, 1988; Greenfeld, 1992).

Nation-State Formation in Europe

Nation-state formation took place in three different ways: within preexisting territory (reform), by bringing together smaller states (unification), and by breaking away from a larger state (separation). These are artificial distinctions, but they draw attention to types of political movements and processes and provide a framework for analysis (Breuilly, 1993).

Reform

France and Britain had a national character by 1800, although rebellions in Ireland and the Vendée indicated that not all subjects agreed. Napoleon introduced legal, administrative, and other reforms, which over the nineteenth century helped overcome regional antagonisms, turning 'peasants' into Frenchmen. Vital was the development of mass education and politics. This took some time, marked by intense conflicts. Nevertheless, by 1900, although disagreements remained about the values to attach to the nation-state, all key political actors subscribed to the national idea and the priority of defending France against foreign threats.

In Britain, the 1707 union between England and Scotland, imperial growth, and the stress on Protestant resistance to Catholic threats shifted the national idea from an English to a British level. The Reform Act of 1832 doubled the electorate, established the House of Commons as the national center of power, and entrenched the two-party system. Following further franchise extensions, universal manhood suffrage was instituted in 1918, and universal adult suffrage in 1930. By 1870, elementary education was made compulsory and in the last decades of the nineteenth century there emerged a mass print media. Peasant and Catholic Ireland responded with separatist movements, but in the rest of the kingdom, the effect was to integrate politics and sentiments on national lines.

Similar processes can be observed in other 'state-nations' such as Spain, Belgium, the Netherlands (after the separation of the two territories in 1830), and the Scandinavian states. However, sometimes this was accompanied by regional resistance that took on a national form, e.g., Catalonia and the Basque regions of Spain (Hobsbawm, 1987; Pugh, 1994; Weber, 1979; Cabo and Molina, 2009).

Unification

German and Italian unification was marked by four features: a dynasty opposed to unification, a state promoting unification, a zone of small states, and neutrality or support from major powers. The growth of constitutional and popular politics meant that political movements increasingly appealed to the nation. Such movements were more important in Germany with suprastate institutions (Confederation of 1814; Customs Union of 1834) and more advanced economic development. Even here, national movements lacked mass support or access to state power.

The opposing dynasty was the Habsburg Empire. The revolutions of 1848 made that opposition very clear and revealed the weakness of popular nationalism. Elite nationalists turned to Prussia and Piedmont. The support of Louis Napoleon, British neutrality, and the weakening of Austrian and Russian capacities to resist change after the Crimean War all created the conditions for effective action against Austria. Cavour, Prime Minister of Piedmont, allied with France against Austria in 1859. The decrepit Kingdom of the Two Sicilies collapsed in the face of invasion by Garibaldi, extending unity to the south. The new state had little elite unity or popular support. As D'Azeglio put it: "We have made Italy, now we must make Italians."

Prussia became capable of acting against Austria after the 1859 war, Austrian crisis, and military modernization in Prussia. Yet Prussia as a conservative dynasty was reluctant to take up the national idea with its radical implications. However, Bismarck, Prime Minister of Prussia from 1862, was an unconventional conservative who believed that the national idea could serve the Hohenzollerns. In alliance with Austria he

exploited the mistakes of Danish nationalism to defeat Denmark in war in 1864 and bring Schleswig–Holstein under direct Austro-Prussian control. He maneuvered a war with Austria, helped by an alliance with Italy (engineered by Louis Napoleon), which was swiftly won. Prussian control was extended over northern and central Germany in alliance with liberal nationalists who helped to organize this political creation. French opposition to extension of control over South Germany was removed in the war against France of 1870–71. Italy exploited the wars of 1866 and 1870–71 to bring the Papal States and Rome into the nation-state.

The greater role of 'positive' factors of Prussian leadership and a national liberal movement compared to the 'negative' factors of state collapse and international support meant that regionalism was more quickly overcome compared to the Italian case. Nevertheless, Germany in 1871 was reluctant to embrace the national idea, as evidenced by the absence of a national anthem, the refusal to adopt the national colors of black–red–gold, and the rapid move into opposition of popular parties representing workers and Catholics. Both Germany and Italy required the same processes of industrial and urban growth, mass education, and politics, as noted for France and Britain, before a genuine nation-state could develop (Breuilly, 2011; Riall, 2009).

Separation

Military defeats weakened Habsburg's capacity to resist demands for internal change. In 1867, a new constitution gave autonomy to Hungary, converting the Magyars into a prop of the empire while depriving non-Magyars of support from Vienna. This intensified nationalist opposition, but Magyars maintained control, providing the outlet of individual assimilation into the dominant culture. In the western half of the empire, a Czech national movement had compelled the dynasty by the 1880s to make concessions at the expense of Germans. The Polish province of Galicia enjoyed much autonomy. Nationalist ambitions were checked by a combination of state power, reform, and the sense that full independence would create a small state at the mercy of powerful neighbors. Consequently, although national conflicts intensified in the Habsburg Empire, they did not directly threaten the empire before 1914.

By contrast, in the Ottoman Empire in Europe, national opposition was less the product of conflicts between ethnic/cultural groups and more the work of Christian elites seeking to replace waning Ottoman power. With assistance from European powers, motivated by self-interest and sentimental identification with Christian and national principles, this enabled the creation of separate states: Greece, Romania, Serbia, and Bulgaria. This stimulated a Turkish nationalist response that challenged the imperial Ottoman elite and in turn reinforced nationalist opposition, not only in Ottoman Europe but also among Arab subjects. Serbia supported national movements among 'its' national counterparts in the Habsburg Empire, which in turn stimulated Vienna to crush Serbia.

In Russia, there was an intensifying interaction between 'official' Russian nationalism and non-Russian movements. These remained weak, especially after the most significant, that of the Poles, was repressed in 1863. In Germany, there was Polish and Danish resistance to incorporation into a nationalizing state (Baycroft and Hewitson, 2006; Haddad and Ochsenwald, 1977; Okey, 2000).

Political Transformation

The trigger for transformation did not come from direct nationalist challenges to Habsburg and Romanov power, but from their responses to the new states in former Ottoman Europe. Austria sought to exploit the assassination of Archduke Ferdinand in June 1914 to bring Serbia to heel and Russia supported Serbia. Austria only acted as she did because of assurances of German support. Russia mobilized her alliance with France. Britain entered the war to prevent German domination in Europe. The success of nation building, especially in Britain, France, and Germany, meant that public opinion powerfully supported their nation-states.

Nationalist movements had little impact on the war. Rather, the defeat of Russia in 1917, and Germany, Austria, and Turkey in 1918, catapulted nationalism into prominence. The national idea was the legitimating principle of the Versailles peace settlement, as expressed in the 'Fourteen Point' program of the US President, Woodrow Wilson. Germany lost her Danish, Polish, and French acquisitions. The Ottoman and Habsburg empires were transformed into a series of nation-states in Europe and Allied mandates in the Middle East. Bolshevik Russia maintained control of non-Russian peoples, but incorporated the national principle into its ideology and institutions (Roshwald, 2001; Manela, 2007).

The new states were the product of great power diplomacy. They included national minorities. There were ethnic conflicts, even massacres. The result was a series of weak states that appealed to the core nationalism, persecuted national minorities, and whipped up discontent among their 'own' national minorities in other states.

These problems were compounded by global depression and the recovery of Germany and the USSR. The Allies could not guarantee the post-1918 order in East-Central Europe. Between 1936 and 1941, Germany and the USSR destroyed that order. With the German invasion of the USSR, the way was opened to the fantastic project of racial empire in Europe.

After Allied victory in 1945, the national principle was too entrenched to permit any alternative to the nation-state, although the principle was not applied to Germany, which was partitioned and occupied. However, states were grouped into supranational blocs under US and USSR leaderships. Many observers thought that the era of nationalism (especially when equated with fascism) and sovereign nation-states had ended in Europe, even if nation-state and national sentiment were firmly established (Brubaker, 1996; Dunn and Fraser, 1996; Payne, 1996).

Nation-State Formation beyond Europe

Nationalism and nation-state formation beyond Europe were responses to Europe. The earliest movements by settlers in the Americas utilized eighteenth-century appeals to political liberty and self-determination, but not the nationalist idea of cultural uniqueness. Americans did not yet see themselves as culturally

distinct from Europeans. The formation of distinctive national cultures was here, as in Europe, a nineteenth-century development associated with commercial, industrial, and urban growth. Similar patterns, although not accompanied by political revolt against the imperial state, mark other states dominated by large-scale white settlement. In Australia, New Zealand, and Canada, this led to the peaceful achievement of Dominion status within the British Empire, later membership of the British Commonwealth, and *de facto* independence, although disputes continue about remaining political ties, as evidenced by the 1999 referendum in Australia on the British monarch remaining head of state. In Canada and South Africa, the story was complicated by conflict between two different streams of European settlement – French and English, and English and Dutch, respectively. In Canada, this conflict remains politically important. In South Africa, the conflict became violent at the start of the twentieth century, but after the formation of an autonomous South African state in 1910 was increasingly replaced by conflict along race lines. This in turn ended with the end of apartheid and the achievement of black majority rule in 1994. In the Americas, Australia, and New Zealand, the precolonial populations were marginalized; in South Africa, this was not possible.

Neither was it possible where European domination was less through settlement than government and trade. Officials, missionaries, and merchants outnumbered white farmers, entrepreneurs, and workers. Colonial states were established in India and Africa. In large parts of Asia, notably China, power was indirectly exercised through client rulers.

There is a continuum rather than a sharp distinction between 'direct' and 'indirect' rule and 'formal' and 'informal' empire. Even an elaborate imperial bureaucracy as in India depended on informal agreements with indigenous elites. Areas of indirect rule were often marked by economic exploitation and cultural domination. Nevertheless, to understand the formation of nation-states, the distinction is important. Arbitrary boundaries imposed by Europe largely persist in the twenty-first century.

This suggests that the nation-state was the inheritor of the colonial state. Political movements formed within the colonial state, whether to oppose or collaborate. Such movements, using modern methods of political rallies, electoral contests, and print media, broke from precolonial patterns, appealing to the 'nation,' as subjects of the colonial state and as bearers of a distinct culture to set against European claims to civilizational and racial superiority.

Formal empire was largely constructed between 1858 (India) and 1900. Territory was transferred from the defeated to the victors of World War I. There were hardly any nationalist movements in these vast regions. Some opposition developed in the interwar period. Just as World War I created the basis for independence claims in Europe, so too did World War II beyond Europe. Japanese success in Southeast Asia and China abruptly destroyed myths of white superiority, although only to be replaced by a new and brutal imperial rule. Nationalist movements were divided as whether to ally with the old or the new imperial power.

World War II destroyed the empires of the defeated powers and fatally weakened those of the European victors. The United States and USSR denied that they were empires and promoted antiimperialist values. Imperial rule was restored by European countries, but was desperately weak. The United Kingdom had to quit India in 1947. Arab nationalism threatened British and French mandates in the Middle East. A uniquely complicating factor was the foundation of Israel in 1948. The crisis of 1956, when Israel, the United Kingdom, and France sought control of the Suez Canal, led to the withdrawal of US support for her Western allies and defeat for the United Kingdom and France.

From 1956, European withdrawal from Africa and the Middle East was just a matter of time. Nationalism became a politics of inheritance. Where political structures had promoted indigenous elites using modern political methods, decolonization was negotiated. Where this was blocked by white settlers or overdependence on direct economic exploitation, decolonization took place in chaos and violence, often with fateful consequences for the independent state. Decolonization is not the same as removal of cultural or economic power. Indeed, some regard decolonization as an imperialist trick. New forms of nationalism developed, stressing economic independence and cultural authenticity rather than 'mere' political self-determination.

In the Middle East, Southeast Asia, and elsewhere, the story is complicated by the emergence of the United States as an imperial power, even if indirectly exercising power. This was partly due to fear of communist expansion, partly to support for client states (e.g., Israel), and partly to protect economic interests. Whatever the reasons, it stimulated anti-American nationalism. This remained tied to existing state boundaries. Nationalists in Tunisia, Libya, Algeria, Egypt, Syria, and Iraq espoused Arab nationalism, but jealously maintained state independence. Vietnamese antiimperialists turned to communism, but never ceased to distrust China. This is not to argue that ancient national rivalries predate and trump other sentiments, but rather that the colonial state and its successor provided the framework around which nations formed.

Important regions not subject to formal empire were Japan and China. In Japan, a unique transition took place from feudal/decentralized to imperial/centralized state, similar to the European pattern and partly in reaction to Western threats. The assertion of autonomy was couched in national terms. Japanese modernization was sufficiently advanced to lead to successful war against Russia in 1904. This is perhaps a unique case of modern nation-state formation without European involvement, although Japan's post-1945 history has been shaped by defeat and American occupation.

The subjugation of China to indirect Western rule produced nationalist rejection of the corrupt empire and alien influences, leading to the formation of a republic in 1911. However, nationalists had little control over much of China, which was run by local warlords, and this inhibited appeals to the 'people.' This appeal came from communists, although the 'people' were constructed in class, not in national, terms. The Japanese invasion of China tilted the balance of power from nationalists to communists, enabling the communists to appeal to anti-Japanese sentiment as well as class interest. Thus, Chinese communism had a nationalist element, which in turn shaped the state it established (Anderson, 1998; Breuilly, 2013; Brown, 1985; Darwin, 1988; Antlov and Tonnesson, 1996).

Nations in a World of Nation-States

After 1945, the nation-state appeared both less and more important than previously: more important as the accepted political unit and because forced population transfers, mass murder, and boundary changes since 1918 had created more ethnically homogenous populations and less important because of the discrediting of nationalism and the emergence of supranational blocs.

The result was 'banal nationalism.' National identity was presented as something 'natural,' separate from 'politics.' National rivalry was displaced from politics into cultural contests. How far such sentiments can be politically mobilized is questionable; after a few years of casualties in Vietnam, it was difficult to sustain popular support for the US war effort. Nevertheless, sentiments of national identity are transmitted everyday through mass media. The banality of the idea means that no political actors can disavow national values, but equally they have little specific political content (Billig, 1995).

There are two major exceptions to this. First is the emergence of regional national movements in Europe and Canada. Some have gained regional autonomy, often the goal rather than the dangers of independence, and something central governments felt able to concede, often as part of a more general decentralizing political process. It is difficult to generalize about these movements that do not appear as part of a general trend, but represent specific problems in individual nation-states.

Second is the collapse of the Soviet Union. The Warsaw Pact states gained real independence, while also abandoning communism, if not authoritarian rule. Non-Russian republics in the USSR declared independence. Regions within Russia asserted claims against Moscow. Russian nationalists resented the loss of imperial power and sought to retain and regain territory and offered help to 'Russians abroad.' Has the collapse of the USSR unleashed a new wave of nationalism?

Nationalism arguably contributed little to the collapse of the USSR, which was principally due to economic crisis and destabilizing political reforms. National movements were a response to that instability. The issue was less about national independence than the form of the post-Soviet state. Answers varied along a continuum from old-style communism through reform communism, takeover by popular leaders, dissenting intelligentsia or nationalists, to disappearance (East Germany). Each can change, e.g., as electorates shift between market-oriented liberal reformers and protectionist, socialist restorers. The nation-state is the political unit within which these politics are practiced, including seeking closer association with Western Europe, especially through entry into the European Union. The great exception was the violence that accompanied the collapse of Yugoslavia. Does this show that nationalism was stronger in Yugoslavia than in the USSR? Arguably, the crucial difference was that Russia led by Yeltsin did not resist, but indeed led the dissolution of the USSR, whereas Serbia, led by Milošević, resisted the breakup and, along with Croatia, sought to enforce territorial claims on other former Yugoslav republics.

In non-Russian republics, the first nationalist outbursts leading to independence subsequently subsided. In Ukraine and Moldova, ethnic nationalism gave way to a stress on the national character of the territory and all its inhabitants. Russian minority problems in states like Estonia became matters of negotiation in ways pessimists had not thought possible. In Russia, major conflicts revolve around economic interests and democratization, rather than nationalist issues.

That does not mean nationalism is unimportant. There are non-Russian areas where military-based movements claim autonomy by violent means, notably Chechnya. Russian nationalism has been inflamed by Western intervention. Festering disputes remain between Russia and former republics. Muslim sentiments can fuel separatist movements, especially in mountainous and inaccessible areas. Nevertheless, in much of the region, political geography seems fairly settled.

Just as one can exaggerate nationalist conflict in Eastern Europe, so can one underestimate it in Western Europe. Many features of the 'classic' nation-state are being superseded. European Union nationals are semicitizens in other EU states. Settled immigrant workers possess legal and social rights. National governments shift from policies of cultural assimilation or exclusion to those of multiculturalism. Governments cede powers to EU institutions. Nevertheless, the most important EU institution is the Council of Ministers, that is, delegates of national governments. While most people remain largely confined to their national language and territory, they will regard themselves 'naturally' as national, something routinely reinforced through mass media. Nationalism as the basis of the organized pursuit and use of state power may have passed its peak. Nations as communities and nation-states as their protectors are still regarded as natural and remain of central importance. The current Eurozone crisis inflames nationalist sentiments and this seems likely to grow in importance (King, 2010; Kupchan, 1995; Smith, 2005).

See also: Age, Sociology of; Nationalism: General; State Formation, Theory of; State and Local Taxation; State, History of; State, Sociology of The; State: Anthropological Aspects; States and Civilizations, Archaeology of.

Bibliography

Anderson, B., 1998. The Spectre of Comparisons: Nationalism, Southeast Asia and the World. Verso, London and New York.

Anderson, P., 1974. Lineages of the Absolutist State. New Left Books, London.

Antlov, H., Tonnesson, S. (Eds.), 1996. Asian Form of the Nation. Curzon, Richmond.

Armstrong, J.A., 1982. Nations before Nationalism. University of North Carolina Press, Chapel Hill, NC.

Balakrishnan, G. (Ed.), 1996. Mapping the Nation. Verso, London and New York.

Baycroft, T., Hewitson, M. (Eds.), 2006. What Is a Nation? Europe 1789–1914. Oxford University Press, Oxford.

Billig, M., 1995. Banal Nationalism. Sage, London/Thousand Oaks/New Delhi.

Breuilly, J., 1993. Nationalism and the State. Manchester University Press and Chicago University Press, Manchester and Chicago.

Breuilly, J., 2005. Changes in the political uses of the nation: continuity or discontinuity? In: Scales, Len, Zimmer, Oliver (Eds.), Power and the Nation in European History. Cambridge University Press, Cambridge, pp. 67–101.

Breuilly, J., 2011. Austria, Prussia and the Making of Modern Germany, 1806–1871. Pearson, London.

Breuilly, J. (Ed.), 2013. The Oxford Handbook of the History of Nationalism. Oxford University Press, Oxford.
Brown, J., 1985. Modern India: The Origins of Asian Democracy. Oxford University Press, Oxford, UK.
Brubaker, R., 1996. Nationalism Reframed: Nationhood and the National Question in the New Europe. Cambridge University Press, Cambridge, UK.
Cabo, M., Molina, F., 2009. The long and winding road of nationalization: Eugen Weber's Peasants into Frenchmen in modern European history (1976–2006). European History Quarterly 39 (2), 264–286.
Colley, L., 1992. Britons: Forging the Nation 1707–1837. Yale University Press, New Haven, CT and London.
Dann, O., Dinwiddy, J. (Eds.), 1988. Nationalism in the Age of the French Revolution. Hambledon, London.
Darwin, J., 1988. Britain and Decolonisation: The Retreat from Empire in the Post-war World. Macmillan, London.
Davies, N., 2011. Vanished Kingdoms: The History of Half-Forgotten Europe. Allen Lane, London.
Dunn, S., Fraser, T.G. (Eds.), 1996. Europe and Ethnicity: The First World War and Contemporary Ethnic Conflict. Routledge, London.
Greenfeld, L., 1992. Nationalism: Five Roads to Modernity. Harvard University Press, Cambridge, MA.
Haddad, W., Ochsenwald, W. (Eds.), 1977. Nationalism in a Non-national State: The Dissolution of the Ottoman Empire. Ohio State University Press, Columbus, OH.
Hastings, A., 1997. The Construction of Nationhood: Ethnicity, Religion, and Nationalism. Cambridge University Press, Cambridge, UK.
Hayes, C.J., 1931. The Historical Evolution of Nationalism. Macmillan, New York.
Hirschi, C., 2012. The Origins of Nationalism. An Alternative History from Ancient Rome to Early Modern Germany. Cambridge University Press, Cambridge.
Hobsbawm, E., 1987. The Age of Empire 1875–1914. Weidenfeld and Nicolson, London.
King, C., 2010. Extreme Politics. Nationalism, Violence, and the End of Europe. Oxford University Press, Oxford.
Kupchan, C. (Ed.), 1995. Nationalism and Nationalities in the New Europe. Cornell University Press, Ithaca, NY, and London.
Leerssen, J., 2006. National Thought in Europe. A Cultural History. Amsterdam University Press, Amsterdam.
Manela, E., 2007. The Wilsonian Moment: Self-determination and the International Origins of Anti-colonial Nationalism. Oxford University Press, Oxford.
Mann, M., 1993. The Sources of Social Power: Vol. 2, The Rise of Classes and Nation-states, 1760–1914. Cambridge University Press, Cambridge, UK.
Okey, R., 2000. The Hapsburg Monarchy 1780–1918: The Improbable Empire. Macmillan, London.
Payne, S., 1996. A History of Fascism, 1914. University of Wisconsin Press, Wisconsin.
Pugh, M., 1994. State and Society: British Political and Social History 1870–1992. Edward Arnold, London.
Riall, L., 2009. Risorgimento: The History of Italy from Napoleon to Nation-state. Palgrave Macmillan, Basingstoke.
Roshwald, A., 2001. Ethnic Nationalism and the Fall of Empires: Central Europe, Russia and the Middle East, 1914–1923. Routledge, London.
Schumann, M., Schweizer, K.W., 2008. The Seven Years War: A Transatlantic History. Routledge, New York.
Seton-Watson, H., 1997. Nations and States: An Inquiry into the Origins of Nations and the Politics of Nationalism. Methuen, London.
Smith, A.D., 1986. The Ethnic Origins of Nations. Blackwell, Oxford, UK.
Smith, A.D., 1998. Nationalism and Modernism. Routledge, London.
Smith, J., 2005. The Fall of Soviet Communism, 1987–1995. Palgrave Macmillan, Basingstoke.
Weber, E., 1979. Peasants into Frenchmen: The Modernization of Rural France 1870–1914. Chatto & Windus, London.

Nation-State and War

Richard Lachmann, University at Albany, State University of New York, Albany, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Conscription transformed how wars are fought, making possible far larger armies and bloodier conflicts. Conscription also changed the relationship of citizens to their nation, allowing soldiers and their families to claim civil rights and social benefits and to demand that their sacrifices be recognized and honored, transforming how war dead are memorialized. The history of conscription, which began with the American and French Revolutions, is now ending in the West, with consequences for how wars currently are fought.

War before Nation-States

Wars, for most of human history, were not fought by nation-states. Men (and it almost always has been men who have fought) identified with and went to war for their kin group, tribe, religion, or ethnic group. Such committed fighters were not military professionals; they were farmers or townspeople, and therefore were available to fight only for short periods of time and near their homes.

The limits of wars fought by part-timers can be illustrated by the ancient Greeks, who left better records than most. Their wars have been analyzed in most detail by Victor Hanson (1989). He finds that while the Greeks tried to kill as many of the enemy (soldiers and civilians) as they could, they also wore heavy armor to protect themselves, which exhausted their wearers in a few hours, and the citizen-soldiers returned to their farms or cities, or to the ships that had transported them to another Greek island, literally, at the end of the day. Greek and other premodern soldiers generally fought only when they were not needed on their farms.

If rulers wanted to fight sustained wars over long distances they needed to recruit aristocrats and their retainers or hire mercenaries. Both those avenues for building armies were problematic. Premodern polities were not unified. Rulers for the most part could not command aristocrats within their 'kingdoms' to fight in wars when and where the king commanded. Instead, rulers had to entice aristocrats to join them in battle. Aristocrats, in turn, had to entice lesser aristocrats to bring their armed men to war. In battle, aristocrats kept command of their own armed men. As a result, 'armies' were not cohesive fighting forces and often fought at cross-purposes with disastrous results. If an aristocrat decided he no longer agreed with the goals or strategy of the war, he could take his men and go home.

Aristocrats' low commitment and generally poor fighting ability led rulers to hire mercenaries as a substitute or supplement for the retainers nobles brought to battles. However, mercenaries' willingness to fight depended entirely on a king's ability to pay them. Mercenaries were fungible: they varied little in their level of skill, and with rare exceptions their commanders lacked the strategic brilliance that would allow them to achieve a decisive victory with a smaller army than their opponents. Technological innovations were quickly copied. Thus, money ultimately was the key to victory.

Rulers, before the era of nation-states, had no reliable method for procuring revenues. Aristocrats and clerics were as reluctant to share their incomes with kings as they were to commit their armed retainers to a monarch's war. Rulers lacked bureaucracies that could collect taxes from their nominal subjects. The best way to extract revenues was by using armed men, who could be used to threaten towns and villages with destruction and death unless they met the king's demands for taxes, or who could be sent into battle to conquer new territories that could be looted or incorporated into the rulers' kingdom and then forced to pay regular tribute.

The conundrum was that rulers needed money to hire the mercenaries that were required to collect money. The process of building armed forces and then bureaucracies to collect revenues was slow and uneven. Charles Tilly (1990), whose 'fiscal-military model' best identifies the patterns in this centuries-long process, notes that cohesive polities formed first in areas with commercial activity that could most easily be identified and taxed, and later spread to areas with aristocracies that could be enlisted to tax settled peasant communities. The rulers with the most success at appropriating revenues created armies that were larger than those of neighboring rulers whom they then could defeat and absorb their territories. In this way, Europe was transformed from a continent with over 500 political units in 1500 to one with fewer than 30 in 1990.

The armies premodern rulers cobbled together with nobles and their armed retainers and the mercenaries they could afford to hire were small; the largest had tens not hundred of thousands of troops. Although, at least in Europe, the trend was toward territorial consolidation, most battles and wars did not reach conclusive ends that permanently altered the borders of the combatant polities. Instead, wars often petered out after years or decades, ending in an agreement to return to previous borders, a repatriation of prisoners, or payment of ransom for a captured king or great noble. The famous conquerors of the ancient world, such as Alexander, Genghis Khan, or Tamerlane, in fact did more looting than conquering. Their empires came apart after their deaths, as heirs and generals grabbed portions of the conquered lands for themselves. The Roman and Chinese empires were exceptional in world history for their size and durability. Those two empires achieved success by creating a unified Roman army and Chinese bureaucracy; however, both those institutions were sustained by granting local garrisons and officials a high level of autonomy in return for not

 http://dx.doi.org/10.1016/B978-0-08-097086-8.96041-0

challenging central authority and sending a fraction of their revenues to the imperial capital. Neither empire was a cohesive polity in the way that modern nation-states are.

War throughout most of human history was not an activity in which most men participated, and when they did it was a part-time task. Nobles, who made up 1–2% of the population, and mercenaries were the only professional soldiers. This pattern was to change with the advent of nationalism and nation-states.

Creating National Armies

Conscription was invented out of desperation. The revolutionary American and French governments, both about to be crushed by the overwhelming financial and military power of Great Britain (alone in the former case, and in alliance with all the other major European powers in the latter) issued calls to ordinary citizens requiring them take up arms to defend their new republics. Conscription created nationalism as much as it drew legitimacy from citizens' already existing feelings of national identity.

Service in a national army brought soldiers together with men from elsewhere in the polity, creating intense feelings of membership in a collectivity, the "sense of a deep, horizontal comradeship" (Benedict Anderson [1983]1991, p. 7) sees as the essence of nationalism, and that extended far beyond the lineages, occupational groups, religious communities, and localities that were the limits of almost all humans' feelings of solidarity until the American and French Revolutions. As Meyer Kestnbaum, the sociologist who has studied the origins of conscription, writes, "Citizen conscription … help[ed] consolidate a politically mass-mobilizing regime … by rendering all citizens formally equal … by integrating them and their state into a single polity … by politicizing them, their relations to one another, and to the state" (2002, p. 131). In essence, conscription was a deal between a regime and its subjects. In return for serving in the military, and risking their lives, subjects became citizens, with rights that residents of a polity who were not members of privileged groups had never before received. The process of gaining those rights was contested and uneven, as we will see in the next section.

The commanders of armed forces realized that, in addition to offering citizen rights and benefits (salary, medical care, pensions, education), they needed to instill nationalist sentiment and understanding in recruits if they wanted them to fight with fervor. For most Europeans of the nineteenth century, their first significant interaction with government officials from beyond their localities came when they were drafted into the military. In addition to training for war, recruits and draftees were taught the national language, which became the working language of the armed forces, and often were given lessons in the civics and history of the state they served. Governments in Asia, Africa, and Latin America have undertaken similar efforts to teach and foster nationalism in their soldiers.

Militaries that were able to draw on nationalism to conscript soldiers, and then succeeded in deepening the nationalist sentiments and fighting fervor of their troops, enjoyed a decisive advantage over armies that still relied on aristocrats and mercenaries. First, conscription severed the link between tax revenues and army size. The French revolutionary army was larger than those of all its antagonists combined in the 1790s, and most importantly, even after massive casualties, it could be replenished with new rounds of draftees. Second, draftees were much less likely to desert than mercenaries and aristocrats, who fought only as long as they were paid or while their patron or noble faction remained in alliance with the king. In contrast, conscripts were pulled into the army by legal obligation to serve (which was not the case for either mercenaries or aristocrats and their retainers). While the nationalist fervor that often attracted soldiers into service generally faded under the grind and horror of battle, it was replaced by a desire for vengeance against the enemy who was killing their fellow citizen-soldiers. Above all, once militaries were able to draft soldiers and keep them in the field for months or years, they were then able to move conscript armies far from home. In essence, conscripts were marooned on the battlefield, unable to return home until their army had won or lost the war. As conscripts spent long periods together, they developed loyalties to the fellow soldiers in their company or platoon, and then fought to help each other survive and to avenge the deaths of individual soldiers with whom they lived and fought. Armies came to realize that soldiers fought better if their commanders fostered solidarity at the platoon level.

Conscript armies revolutionized military strategies. Eighteenth century commanders had developed fighting techniques that relied on soldiers who were well trained, frequently drilled, and willing to obey their officers' orders. However, until the advent of conscription those techniques could not really be implemented. Mercenaries and aristocrats generally did not serve long enough, and almost never showed up long enough before battle, to be trained and drilled in the new battle tactics. Nor were mercenaries and especially aristocrats willing to follow orders.

Despite all the advantages of conscript armies, most European countries refused to institute the draft. Aristocrats often had enough political power to prevent reforms that would undermine their control over armies. Kings and aristocrats also feared that subjects who were armed and trained to fight could turn on their rulers. Yet, repeated victories by conscripts over aristocratic armies eventually led governments to adopt the draft. The US Civil War was the first war in human history in which both armies were manned by conscripts. By the time of World War I all European armies relied on conscripts, except for Britain, but even Britain had to institute the draft after its volunteer armies were destroyed in the first years of trench warfare.

Conscripts were willing, or could be compelled, to train, drill, and fight in ways that exposed them to much higher risk of death than aristocrats or mercenaries had been willing to tolerate. This led to a further military revolution, as commanders gave priority to killing the vastly larger enemy armies they now faced. Once an enemy had access to almost limitless numbers of fresh troops, it became essential to kill as many of the enemy as possible. Armies sought to develop weapons able to do that. Militaries invested resources in creating more powerful weapons. The machine gun, for example, as John Ellis (1975) has shown, was first developed to allow a few Europeans to kill and intimidate large numbers of natives in Asian and African colonies. However, the few

available machine guns at the start of World War I proved most effective at killing large numbers of enemy troops and both sides then purchased huge numbers of the guns, which along with poison gas and shells, turned World War I from a war to capture territory into a war to annihilate the enemy's entire armed forces. This logic was carried to its ultimate conclusion in World War II, as the Allies and Axis both targeted civilians. After all, civilian men could be called up in future rounds of the draft, and civilian women could produce babies who could become soldiers in later years. This culminated in the annihilation of entire cities, and remains the logic behind countries' manufacture of and planning to use nuclear weapons.

Citizen Armies and Soldiers' Rights

During and after wars, soldiers and their families made demands on the governments that forced them to fight and that exposed so many of them to injury and death. Benefits such as old age and disability pensions, health care, and other forms of social insurance, which were generally offered first to government employees, were then extended during and after wars to conscripts and veterans. Often the rest of the population had to wait decades to obtain, or still have not received, the benefits given to soldiers.

The US created a system of pensions for veterans and their widows after the Civil War that preceded the universal social security system by seven decades. Theda Skocpol (1992) shows that Civil War pensions became the largest item in the federal budget by the end of the nineteenth century. At the end of World War II, US veterans received subsidized home mortgages, unemployment benefits, and free university educations through the Servicemen's Readjustment Act of 1944 or G.I. Bill. These benefits, which later were offered in modified form to the entire population, made the US the world's leader in home ownership and, until the twenty-first century, in the fraction of young people with university educations. In other domains, benefits offered to veterans have not been extended to others: in the early twenty-first century, soldiers remain the only Americans who have access to guaranteed day care for their children.

Many European, Latin American, and Asian countries also modeled their social programs on benefits first given to soldiers and veterans, although in most of those countries benefits were extended much more quickly to the rest of the population than they were in the US. Often the rest of the population laid claim to social benefits by arguing that they shared in soldiers' sacrifices. Mothers who gave birth to and raised sons had the most compelling claims to civilian benefits.

Soldiers and their families also demanded voting rights. John Markoff (1996) finds that worldwide waves of democratization occurred at the end of both World Wars. This encompassed the extension of civil liberties and voting rights to excluded classes and ethnic and racial groups, women and youth, and expanded rights to those who already were full-fledged citizens. Markoff argues, correctly, that the World Wars empowered women's and labor movements, which won voting rights and other concessions in return for their commitment to munitions factory work and willingness to forgo strikes to ensure the delivery of critical supplies of weapons during the wars. Those practical deals were fortified as well by the moral claims women made to citizenship as the mothers and wives of soldiers.

Conscription, and the extension of social benefits and voting rights changed the meaning of national identity. Conscription turned subjects into citizens. Citizens were obligated to protect the nation of which they were a part, an obligation expressed primarily by years of service in the military and the risk of injury and death in that service. However, in return soldiers and their families were part of the nation and received the benefits of citizenship. Their basic needs – a minimum income when they were unable to work due to age, disability, or unemployment, access to health care, education, housing, child care – were part of a growing list of needs which their nation's government became obligated to provide. As citizens, soldiers, and their families demanded a role, through the right to vote, in determining when and where their nation fought wars, and in all the other decisions, including the provision of social benefits, once made by rulers or narrow elites.

Obviously, conscription was not the only spur to voting rights and social benefits. The US stands as an example of how benefits granted to soldiers are not necessarily quickly or ever granted to the rest of the nation. Voting rights and social benefits also were the product of struggles led by working classes, minorities, women and other popular forces, and organizations. However, the vast reach of social programs and the universal franchise in many countries would not have been possible without the claims made by soldiers and their families and the openings that wartime mobilization created to press those claims against governments that were fighting for their survival and therefore were desperate to secure the consent and willingness to sacrifice of their citizens.

Conscripts and their families also made another sort of claim upon their governments: for recognition of their sacrifices. Before the era of conscription, governments made no efforts to memorialize dead soldiers. Memorials were meant for commanders. Triumphal arches were built for generals; victory parades honored the generals not the soldiers. Nor did governments even attempt to count the number of their soldiers who were killed or injured or to inform families of the fates of their loved ones. When soldiers failed to return from war, their families often did not know if they had died or just had decided to settle elsewhere.

The US Civil War, the first war fought between two conscript armies, began a fundamental transformation in memorializing war dead. For the first time a government assumed the obligation of identifying all the war dead, recovering their bodies, and informing families of those deaths. The dead were then buried in national cemeteries without distinction as to military rank, religion, or ethnicity (although until the middle of the twentieth century African-Americans were buried in separate sections of most US military cemeteries). Britain, France, and Germany continued to separate soldiers by rank in their military cemeteries until the end of World War I. The magnitude of death in that war, as for Americans in the Civil War, led soldiers' families to demand, and governments to recognize, the equality of their sacrifices, treating all dead citizen-soldiers equally.

These American innovations were copied by other nations that conscripted soldiers. The Arc de Triomphe traces the evolution of war commemoration. It was originally designed, following the model of imperial Roman triumphal arches, to mark Napoleon's victory at Austerlitz in 1806, and only generals' names were carved on it. After World War I it became the site for the Tomb of the Unknown Soldier, which in France and in all other countries with such tombs, symbolizes the nation's commitment to honor every one of its war dead.

The commitment to honor all the war dead became an effort to actually find and identify each soldier's body. If soldiers were to be buried, their bodies had to be recovered from the field of battle. The US and other countries have taken increasingly elaborate measures to ensure that all bodies were accounted for. After the difficulty of identifying bodies left on battlefields in the Civil War, the US and other countries made use of 'dog tags' in World War I, giving each soldier a metal disk with their name and other identifying information to be worn at all times so that their bodies could be identified. More recently, DNA testing has allowed even the tiniest body fragments to be identified, and governments have undertaken the task of finding and testing such fragments to determine the identities and place of death of formerly unknown bodies and to link them with soldiers classified as 'missing in action.' The often extreme and expensive efforts to recover bodies and body fragments (the US still is attempting to do so in Vietnam, four decades after the end of the war) are an indication of how seriously soldiers' families take this obligation and how strongly they hold governments to it. Similarly, soldiers are expected to try to recover the bodies of their dead comrades from the field of battle, even at the risk of their own lives. This imperative, which became central to concepts of military honor in countries as they adopted conscription, is yet another way of indicating the fraternity and equality of citizen-soldiers.

The Demise of Conscription

The two-century history of conscription is coming to an end, at least in the West. States' capacities to convince or compel its citizens to risk their lives, or even to give a few years of service, for their country have declined dramatically in Europe and the United States since World War II, and might decline in much of the rest of the world in coming years as well.

Commitment to military service has fallen at the most basic level: countries first moved in recent decades to reduce the length of military service and offer more exemptions to students and conscientious objectors, and then abolished the draft entirely. The status of conscientious objectors has changed most dramatically in Europe: from World War I, when, as Nicoletta Gullace (2002) documents, British men who did not serve were publicly humiliated by women who pinned white feathers on them, fired by employers, and stripped by Parliament of the right to vote at war's end, to the late twentieth century, when beginning in the 1970s the Universal Declaration of Human Rights was reinterpreted to encompass conscientious objection and Amnesty International adopted jailed conscientious objectors as prisoners of conscience.

The draft was abolished entirely in Britain in 1960, Australia and New Zealand in 1972, the US in 1973, Belgium in 1994, France and the Netherlands in 1996, and in much of the rest of Western Europe and most of the former Soviet bloc in the first decades of the twenty-first century. Russia is in the process of abolishing its draft. While Israel still has universal service (and is the only country in the world ever to draft women), a growing percentage of young people are being allowed to leave military service after a few months rather than fulfilling their full legal obligation.

The European and American abolition of the draft is a rejection both of the sense that young men should be obligated to devote time to national service and also of the belief that nations face threats that deserve to be countered at the cost of their citizens' lives. The most common explanation for this rapid and a dramatic change comes from Hanson (2010, p. 242): "Western medical science not only ensures longer life, but, through the use of sophisticated pharmaceutics and artificial body parts, also promises an enjoyable longer life. Those advances will make it ever harder to ask a small minority of our citizens to lose their lives to ensure longer and better ones for the rest of us. Most twenty-first century Westerns do not see death on their streets. They do not butcher the animals they eat. And they do not lose a large minority of their children to disease and hunger before the age of twelve." Leon De Winter (2007) argues: "Today's parents often have no more than two children, some may have only one son. His life is so precious that it has come to seem unbearable for him to be killed in battle." If this theory is correct, we can expect to see an end to conscription, and greater resistance by citizens to allow themselves or even volunteers to fight and die in wars, in countries such as China and Iran which are experiencing rapidly falling birthrates.

Parents' and all citizens' aversion to death is magnified by the increasing attention governments, the media, and citizens pay to the individual soldiers who die in war. The increasingly personalized and elaborate memorialization of the dead, which we examined in the previous section, serves to magnify each death for the public at large and makes it more difficult for voters to agree or governments to demand that citizens submit themselves to conscription and risk their lives in war.

The Future of War

The end of conscription is drastically transforming states' capacities and strategies for fighting wars. Without conscription, states are forced to rely on volunteers. As the supply of soldiers shrinks, strategies that depend on tolerating massive casualties to gain territory or on long wars of attrition are no longer possible. As a result, Western nations practice what Martin Shaw (2005) has labeled "risk-transfer war" in which the cardinal rule is to minimize one's own casualties even if that increases civilian casualties in the foreign lands where Western nations fight.

States have developed technologies and strategies to minimize their own soldiers' casualties. Drones and other remote-controlled smart weapons are the most prominent technological fixes. However, despite the claims of governments that drone attacks are carefully planned to hit only military targets and that civilians are rarely or never hit, in fact

the main purpose of such weapons is to protect soldiers from one's own country, not civilians in the territories under attack.

High-tech weapons are useful only for certain sorts of attacks: ones that target a relatively small number of enemy fighters. They are not effective for controlling territory or subduing large populations. In any case, such weapons are effective only if the attacking force has access to intelligence on the ground, and that requires having live soldiers or spies in place who are then vulnerable to attack. When drones hit civilians they inevitably anger the local population, undermining efforts to win locals' 'hearts and minds.' Thus, risk-transfer strategies make it more difficult if not impossible to win wars that require winning the loyalty or at least acquiescence of the local population.

Armed forces dedicated to minimizing their own casualties become vulnerable to and distracted by insurgent strategies. Soldiers are captured or taken prisoner in almost all wars. In the long era before nationalism, governments only made efforts to ransom kings or high nobles, and cared not at all if common soldiers were imprisoned or even enslaved by their enemy captors. With the advent of conscript armies of national citizens, the recover and decent treatment of prisoners of war became an obligation that governments took seriously. When governments slacked in their efforts to recover prisoners, captured soldiers' families agitated for their release. This gives insurgent forces a powerful source of leverage. When they capture enemy soldiers they can hold them for substantial ransoms. The current Israeli practice of trading 1000 Palestinians, whom the Israeli government considers terrorists, for a single kidnapped or captured Israeli soldier, a practice supported by large majorities of the Israeli population, is emblematic of how efforts at casualty minimization affect governments' military strategies. If troops need to be positioned to ensure they are not captured, battlefield tactics are severely limited, just as medieval armies that gave top priority to protecting their king or noble commander were hampered in war. And of course thousand to one trades counteract an army's goal of neutralizing as many enemy fighters as possible.

Casualty minimization already forces governments to restrict their foreign policy goals to ones that can be accomplished with high-tech weapons, and inevitably will continue to do so in the absence of a return to conscription. The US and European nations' decision not to intervene to stop the Rwandan genocide was an accurate reflection by those governments of their publics' unwillingness to tolerate casualties to save civilians in a distant nation. Similarly, the decision to intervene in Bosnia with high-altitude bombing that sacrificed local civilians for the safety of NATO pilots reflects the same calculus.

Nationalism, thus, elevates the value of one's own citizens at the same time as it devalues the lives of foreigners. In the World Wars, this was expressed in support for massive attacks on civilian populations in enemy countries. Today, it is expressed in extreme solicitousness for the soldiers of one's own nation and an almost absolute unwillingness to sacrifice even a single life of one's countryman to save large numbers of civilians abroad.

If the present emphasis on protecting the lives of one's own soldiers above all other considerations continues governments will fight fewer wars. Militaries will exist mainly as deterrents against improbable attacks upon the nation's home territories and as symbolic representations of the nation. That does not mean wars will end. Nonstate actors, and states where nationalism is weak or where it is not expressed in concern for the lives of each soldier as an individual, will continue to be able to fight wars. Governments concerned with casualty minimization are likely then to intensify their efforts in 'homeland security,' in an effort to keep the violence of nonstate actors at bay. Israel offers a model of this approach. Its security wall and related measures are a highly successful (at least until now) strategy to protect Israelis, in settlements in the occupied territories as well as in Israel proper, from attack. Combined with the threat and use of massive retaliation against its enemies, Israel has managed to deter and deflect attacks despite continuing to hold and build settlements in the West Bank. This is the future of warfare in modern nation-states.

Nationalism, which in the nineteenth and twentieth centuries, made possible warfare and casualties on a scale never seen in human history, today is undermining states' capacities to enlist soldiers and place them at risk of death. The geopolitics made possible by mass armies of draftees is being transformed as conscription and tolerance for casualties disappears from more and more countries.

See also: Military Sociology; Military, War, and Politics; War and Democracy; War, Sociology of; Warfare in History.

Bibliography

Anderson, [1983]1991. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London.

De Winter, Leon, 8 June 2007. My only son. The Wall Street Journal, p. A16.

Ellis, John, 1975. The Social History of the Machine Gun. Random House, New York.

Faust, Drew Gilpin, 2008. This Republic of Suffering: Death and the American Civil War. Knopf, New York.

Gullace, Nicoletta, 2002. The Blood of Our Sons: Men, Women, and the Renegotiation of British Citizenship during the Great War. Palgrave, New York.

Hanson, Victor Davis, 1989. The Western Way of War: Infantry Battle in Classical Greece. Knopf, New York.

Hanson, Victor Davis, 2010. The Father of Us All: War and History, Ancient and Modern. Bloomsbury, New York.

Kestnbaum, Meyer, 2002. Citizen-soldiers, national service and the mass army: the birth of conscription in revolutionary Europe and North America. In: Mjoset, Van Holde (Eds.), The Comparative Study of Conscription in the Armed Forces. JAI, Amsterdam, pp. 117–144.

Markoff, John, 1996. Waves of Democracy. Sage, Thousand Oaks, CA.

Mjoset, Lars, Stephen, Van Holde, 2002. Killing for the state dying for the nation: an introductory essay on the life cycle of conscription into Europe's armed forces. In: Mjoset, Van Holde (Eds.), The Comparative Study of Conscription in the Armed Forces. JAI, Amsterdam, pp. 3–94.

Moller, Bjorn, 2002. Conscription and its alternatives. In: Mjoset, Van Holde (Eds.), The Comparative Study of Conscription in the Armed Forces. JAI, Amsterdam, pp. 277–305.

Mosse, George Lachmann, 1990. Fallen Soldiers: Reshaping the Memory of the World Wars. Oxford University Press, New York.

Shaw, Martin, 2005. The New Western Way of War: Risk-Transfer War and its Crisis in Iraq. Polity, Cambridge.

Skocpol, Theda, 1992. Protecting Soldiers and Mothers: The Political Origins of Social Policy in the United States. Harvard University Press, Cambridge.

Tilly, Charles, 1990. Coercion, Capital, and European States. Blackwell, Oxford.

Nation-State as Symbolic Construct

Bart Bonikowski and Nina Gheihman, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Research on nationalism has been especially preoccupied with those aspects of the phenomenon that are most destabilizing for existing institutions, thereby assuming that in the absence of violent upheavals, nationalism in established democracies is simply a *fait accompli* rather than a source of continued social and political change. In contrast, more recent studies have turned their attention to everyday forms of nationalism, arguing that the primacy of the nation-state as a unit of political governance and collective identification is continually reinforced – and sometimes subtly altered – through routine cognitive and affective orientations that are themselves products of institutional and ritual practices. This article provides an analytical overview of this literature, identifying its contributions, limitations, and potential for achieving a more complete understanding of nationalism in contemporary societies.

Introduction

Nationalism is one of the central ideologies of the modern era. Not surprisingly, its hegemonic rise over the past 200 years has attracted considerable attention from historians, sociologists, and political scientists. Researchers have been especially preoccupied with those aspects of nationalism that have been most disruptive for existing social and political institutions, from the emergence of collective identification in newly forming nation-states and the struggle for statehood by minority groups to the sudden eruptions of nativism and xenophobia in otherwise stable societies. Much of this literature has been based on the implicit assumption that in the absence of violent upheavals, nationalism in established democracies is simply a *fait accompli* rather than a source of continued social and political change.

In contrast, more recent studies have turned their attention to everyday forms of nationalism, arguing that the primacy of the nation-state as a unit of political governance and collective identification is continually reinforced – and sometimes subtly altered – through routine cognitive and affective orientations that are themselves products of institutional and ritual practices. Such work has shown that meanings attached to the nation are not uniform within a given national community, but systematically vary across individuals and social groups and are associated with other socially and politically relevant attitudes and behaviors. This article provides an analytical overview of this burgeoning literature, identifying its contributions, limitations, and potential for achieving a more nuanced understanding of nationalism in contemporary societies.

Traditional Nationalism Scholarship

The majority of research on nationalism falls within three broad traditions that investigate the origins of the nation-state as an institutionalized political form, the rise of new independent nation-states in the post-War period; and radical nationalist ideologies in contemporary societies.

Focusing on the historical emergence of the nation-state, the first approach is grounded in the observation that prior to the eighteenth century, the political authority vested in statelike institutions was rarely absolute or uncontested. Most states ruled over diverse populations sharply divided along lines of ethnicity, cultural tradition, religion, and language. It is only with the rise of the modern nation-state and its constitutive ideology of nationalism that state sovereignty became codified in both a territorial and cultural sense, with state subjects increasingly viewing themselves as members of a community with shared descent, customs, and collective fate (Gellner, 1983; Anderson, 1983). In the process, the nation became the source of the state's legitimacy and the congruence of the state's political and the nation's cultural boundaries became a taken-for-granted reality. Of course, this process was far from teleological; the rise of the nation-state came in fits and starts and was accompanied by protracted conflicts and institutionalized forms of intergroup domination.

The second tradition documents the rise of independent nation-states in the post-War period, most recently during the postcommunist transition in Eastern Europe. This literature is rich in case studies, but has been less effective at formulating general conclusions about nation-state formation and diffusion. Among the exceptions is Hechter's (2000) *Containing Nationalism*, which argues that nationalist movements are mobilized by centralized states' direct rule over culturally heterogeneous but territorially bounded populations. Viewing this as a problem for state stability, Hechter advocates policies that grant autonomy to minority populations. Another important theoretical contribution is Brubaker's (1996) study of ethnic relations in former Soviet republics and satellite states, which he sees as characterized by three distinct types of nationalism: 'nationalizing nationalism' of the newly sovereign states, 'minority nationalism' of domestic minority groups, and 'homeland nationalism' of coethnics in bordering countries. The interaction between these three political forces gives rise to state-driven social exclusion, oppositional struggles for minority rights, and occasional eruptions of violence directed against states and their majority populations.

Finally, the third body of literature focuses on radical nationalist movements. Rather than aiming to reconfigure the political boundaries of existing nation-states, these movements attempt to mobilize members of dominant groups to resist changes to the nation's symbolic boundaries brought on by

immigration or the extension of rights to oppressed native-born populations. These mobilization efforts are typically motivated by racially, ethnically, or religiously essentialist conceptions of the nation that explicitly exclude entire categories of people based on ascribed characteristics. Much contemporary research in this tradition focuses on Europe, which has experienced a resurgence of radical right-wing politics with strong ethnic nationalist tendencies (Taggart, 2000; Mudde, 2007; Bale, 2013). A distinct but substantively related line of research on the United States has focused on the country's long history of exclusionary politics based on nativist, racist, and fundamentalist ideologies (Higham, [1955]2002; Burris et al., 2000; Blee, 2002).

The above three approaches have generated many rich insights about the origins and continued importance of nationalist ideology in modern society. Yet, this classic work has largely ignored the more mundane – but no less important – aspects of everyday nationalism in stable democratic countries.

Nationalism in Settled Times

As a growing number of studies have demonstrated, the political significance of nationalism is not limited to periods of major political transformation or national crisis; nationalism exerts distinct effects on political change in settled times, although its impact is likely to be more subtle (Swidler, 1986). To understand this process, the analytical lens must shift from explicit political ideologies to tacit understandings of the nation's symbolic significance within a given polity. While most citizens of contemporary societies take for granted the legitimacy of the nation-state form, they often disagree in important ways about what their nation-state means to them. This variation – both within and across populations and over time – is an important feature of political culture that influences political behavior and drives policy change. Scholars have explored this dimension of nationalism from two vantage points: first, historical, ethnographic, and interpretive studies have explored the meaning-making processes that produce and maintain national identification; second, survey-based research has sought to map the variation in nationalist attitudes within and across populations and to demonstrate the impact of these attitudes on other political beliefs. Together, these approaches have laid the groundwork for a more complete understanding of nationalism in politically settled times.

Interpretive Studies of Everyday Nationalism

A long line of work in political sociology and political science has focused on how meanings attributed to the nation affect people's self-understanding and how such conceptions relate to a given country's overarching political culture. Classic studies in this tradition were based on a functionalist understanding of culture as a coherent system of agreed-upon values that facilitates social cohesion; this logically implied that the task for nationalism scholars was to uncover each nation's essential and stable cultural characteristics. For instance, in *Continental Divide*, Lipset (1990) argues that the national identity of the United States consists of the central tenets of the American Creed: antistatism, individualism, populism, and egalitarianism. In contrast, Canada – Lipset's comparison case – prioritizes a distinct set of principles, including deference to authority, collectivism, elitism, and group-based particularism. Such large-scale cultural generalizations were typical of post-War scholarship, as exemplified by the work of Myrdal (1944), Hartz (1964), and the broader enterprise of consensus history.

An influential legacy of the functionalist approach has been the distinction between ethnic and civic nationalism, initially made by Friedrich Meinecke ([1907]1970) and subsequently elaborated by Hans Kohn (1944). This binary opposition assumes a stable character to national identity, but differentiates between two alternatives: the first based on ascriptive criteria such as race, ethnicity, ancestry, religion, or language and the second on elective criteria, such as commitment to the country's core ideology, subjective identification with a national community, and respect for the nation's laws and traditions. This dichotomy has been used to classify the central ideologies of specific nations – with Germany as the prototype of ethnic nationalism and France of civic nationalism – and occasionally entire world regions.

The view of national culture as stable and homogenous has been challenged by more recent research. Rogers Smith's (1997) work on citizenship law in the United States, for instance, reveals a layered and often contradictory patchwork of legislation and court decisions informed by three distinct ideological perspectives (i.e., liberalism, republicanism, and 'ascriptive Americanism'), which have competed for dominance over the course of American history. Others have critiqued the ethnic–civic distinction on theoretical and empirical grounds for attributing essential properties to entire countries and regions and glossing over considerable within-country heterogeneity (Kaufmann, 2000; Shulman, 2002; Brubaker, 2004; Ceobanu and Escandell, 2008).

Despite its limitations, functionalist research on national identity has generated a number of valuable insights. Perhaps most importantly, it has helped scholars recognize that even though nationalism has become a hegemonic ideology in modern society, the content of nationalist beliefs can be highly variable. This work has also highlighted the inherent tendency of nationalism toward social exclusion: given that nationalism is predicated on a fundamental belief in the unique characteristics of each nation, it inevitably draws sharp symbolic and social boundaries around national communities based on a range of arbitrary criteria. Of course, the need for distinguishing between members and nonmembers is necessary for every state's ability to fulfill its core functions, such as generating tax revenue, managing economic development, providing social programs, and ensuring national security, but how such distinctions are made and maintained is an important object of study.

While the functionalist approach attempted to identify the shared attributes of a given political culture, more recent work in cultural sociology has focused on national narratives, which serve as focal points for collective identification. This tradition is inspired by the classic Durkheimian insight that in order to maintain social solidarity, collectivities must engage in ritualistic practices that imbue shared symbols with moral meaning. Thus, studies have paid particular attention to ritualistic events such as parades, concerts, festivals, and sporting events

(Waldstreicher, 1997) as well as commemorative practices, such as the construction of national memorials and monuments (Collins, 2012). These empirical sites often reveal not only widely shared narratives, but also processes of contestation over the dominant interpretation of a nation's past, its core values, and its future aspirations. Once such symbolic conflicts subside, the symbols produced in the process come to serve as tacit reminders of the primacy of the nation-state in everyday life. This is as true of memorials and holidays, as it is of more banal manifestations of the nation-state, such as currency, commemorative street names, or flags waving over private homes and businesses (Billig, 1995). These ritual practices are not merely by-products of collective identification, but are essential for the reproduction of the national community.

Survey-Based Studies of Everyday Nationalism

One of the key insights of interpretive research on national identification is that the meaning of the nation-state is not uniform across a national population. Case studies of specific commemorative conflicts, however, are unable to map the full variation of national beliefs and determine their association with other social and political attitudes and behaviors. This task is more appropriately handled with survey data that enable population-level inferences based on representative samples of respondents. Most of this work has relied on a handful of multiple-wave, representative-sample surveys systematically administered across a wide range of countries.

Survey researchers face a considerable challenge, however, in operationalizing nationalism in a way that makes it amenable to quantitative analysis. This task has proven even more difficult, given the general lack of consensus in the literature about what nationalism actually entails. In an effort to organize the varied survey-based literature, we identify four ways of defining nationalism: as national attachment; national identity; national pride; and national hubris. In our view, these four labels represent different aspects of a single overarching phenomenon: the repertoire of meanings attached to the nation-state, through which modern-day nationalism maintains its hegemonic status. Under certain circumstances, these meanings also drive political change, although typically within the largely unquestioned auspices of the nation-state.

National attachment refers to identification with the nation over and above other collective entities. Social identity theorists have long argued that individuals possess multiple identities that are shaped and reproduced by institutions (Stryker and Burke, 2000). The salience of particular identities varies over time for any given individuals but it is also likely to vary, on average, across individuals. Given that the former is notoriously difficult to study outside of the laboratory setting, most scholars interested in national attachment focus on the latter, typically using basic survey questions that explicitly ask respondents about the absolute or relative importance of the nation in their lives (e.g., "How close do you feel to America?").

Some scholars view national attachment as a particularly useful measure of nationalism, because it is largely devoid of ideological content: whether a nation is important for an individual tells us little about the meanings that the individual ascribes to the nation (Huddy and Khatib, 2007). Furthermore, national attachment is related to but distinct from other concepts of interest to nationalism scholars, such as attitudes toward minorities, sentiments toward state institutions, and feelings of national superiority.

Studies have demonstrated that heightened national attachment is systematically associated with measures of political knowledge and participation (Huddy and Khatib, 2007) and prosocial behavior (Rahn, 2000). At the same time, greater attachment can activate ethnic forms of national identity, thereby promoting animosity toward minority groups (Sidanius et al., 1997). The nature of the relationship between national attachment and antipathy toward perceived out-groups often depends on how a particular issue is framed in public discourse. A national crisis, such as 9/11, can raise the relative salience of the nation in people's identity hierarchies, but whether the same occurs for ethnic identities depends on whether the narratives that mediate people's interpretation of collective events are framed in ethnically essentialist terms (Li and Brewer, 2004).

Ethnic identities, however, are only one form of collective affiliation with which national attachment interacts; people are simultaneously members of local, national, and in some cases, supranational communities, and these collective affinities often exhibit mutual complementarities and contradictions. Indeed, given the growing importance of subnational and transnational communities in the era of globalization, such place-based identities have become increasingly complex and multifaceted. Much of the research on these multilayered identification processes has been conducted within the context of the European Union (EU), the largest supranational integration project in recent history. For many years, the consensus in the EU literature has been that European and national identities are inherently opposed and that full European cultural integration depends on the gradual erosion of national loyalties. Recent studies, however, have shown that the two sets of identities can coexist, and in fact often reinforce one another, particularly among educated cosmopolitan elites (Díez Medrano and Gutiérrez, 2001; Fligstein et al., 2012).

The second dimension of nationalism – national identity – consists of specific criteria of inclusion and exclusion that are seen as legitimately defining the nation's symbolic and social boundaries (Lamont and Molnár, 2002; Wimmer, 2008). This aspect of nationalism has generated considerable attention, because it is directly tied to processes of social stratification. Groups that are viewed as external to the national community are frequently denied fundamental rights and excluded from equal access to resources and opportunities. At the individual level, exclusionary conceptions of national identity are associated with anti-immigrant sentiments, as well as support for restrictive immigration policies and cultural protectionism (Ceobanu and Escandell, 2008; Schildkraut, 2003).

Scholars typically distinguish between forms of national identity based on ascriptive (ethnic) and elective (civic) traits. There is some tension, however, surrounding the analytical utility of the ethnic–civic distinction, because of its origins in research on essential national character and the persistent finding that national identity is in fact heterogeneous within countries. Scholars have employed two strategies to resolve this tension: they have used the ethnic–civic binary at the individual level – allowing for within-country variation – or they have replaced the dichotomy altogether with more nuanced typologies.

In a study emblematic of the former approach, Jones and Smith (2001) demonstrate that both civic and ethnic forms of nationalism are found within the countries in their sample, although there is some variation in the degree to which religion is valued as a criterion of national belonging (see also Citrin et al., 1990; Hjerm, 2003). Representative of the latter approach, Citrin et al. (1994) build on the ethnic–civic binary by identifying three distinct understandings of American national identity: cosmopolitan liberalism, nativism, and multiculturalism. Similarly, Schildkraut (2003, 2011) identifies four distinct varieties of national identity, which are highly predictive of individuals' policy preferences and political behavior (for a related approach at the cross-national level, see Bonikowski, 2013).

In addition to classifying conceptions of national identity, researchers have also attempted to identify the factors that affect the prevalence of these beliefs within national populations. At the individual level, respondents who hold ascriptive views of national identity tend to have lower levels of education and lower income, be more religious and older, belong to ethnic majority groups, and have resided in the country for a longer period of time (Hjerm, 2001; Jones and Smith, 2001; Kunovich, 2009; Wright, 2011a). At the national level, countries with a lower prevalence of ascriptive national identity are typically more developed, democratic, urbanized, and culturally globalized, and are characterized by greater internal ethnic differentiation (Jones and Smith, 2001; Kunovich, 2009; Green et al., 2011; but see Ariely, 2012 on globalization); they also tend to have more generous welfare state policies (Wright, 2011b).

While national identity reflects beliefs about criteria of legitimate membership in the nation, national pride refers to the affective evaluation of the nation-state. Some researchers prefer the term patriotism to distinguish this ostensibly innocuous sentiment from a more invidious nationalism. In support of this distinction, studies have found that pride in the nation-state's institutions is uncorrelated with out-group prejudice (e.g., De Figueiredo and Elkins, 2003), while others go further to argue that pride actually leads to more tolerant attitudes (Green et al., 2011). This evidence is contradicted, however, by studies that show patriotism to be linked with a variety of authoritarian attitudes (Staub, 1997). To overcome this conceptual ambiguity, Schatz et al. (1999) distinguish between 'constructive' and 'blind' patriotism and argue that blind patriots are more chauvinistic and feel more threatened by minorities, while constructive patriots have a greater sense of political efficacy (Hjerm, 1998; Parker, 2007, 2010).

In contrast to research that views national pride as unidimensional (whether positive or negative), other studies argue that national pride is itself heterogeneous and domain specific, consisting of distinct sentiments toward national institutions and the nation's cultural heritage (Hjerm, 1998; Bonikowski, 2011). Indeed, this distinction is consistent with historical research on nationalism, which has sought to understand how political units (i.e., states) come to be viewed as legitimate authorities over imagined communities of shared fate (i.e., nations). The two varieties of national pride appear to have distinct associations with other aspects of nationalism, as demonstrated by Ariely (2011): pride in the nation – what he calls "cultural patriotism" – is correlated with ethnocultural conceptions of national membership, while pride in political institutions (i.e., 'political patriotism') lends itself to civic forms of national identity.

The final dimension of nationalism prevalent in existing scholarship is national hubris – or more conventionally, chauvinism. While the other aspects of nationalism are inward oriented, hubris is concerned with judgments of the nation's position vis-à-vis the rest of the world (much as individual hubris involves perceptions of superiority to other people). For some scholars, particularly in political psychology, hubris is coterminous with nationalism as a whole, typically in juxtaposition with the ostensibly more benign patriotism.

The primary reason for the scholarly interest in national hubris is its association with support for bellicose foreign policy (Kosterman and Feshbach, 1989) and authoritarian politics (Blank, 2003). A heightened sense of national superiority logically lends itself to the belief that the nation should pursue its interests against all odds, even if this places it in conflict with others. Not surprisingly, such views go hand-in-hand with uncritical support for the state that is a hallmark of 'blind patriotism.' It is likely that the effect of hubris on aggressive policy preferences is partly mediated by general out-group antipathy, as measures of hubris are also correlated with prejudice against minority groups (Blank and Schmidt, 2003; De Figueiredo and Elkins, 2003).

Limitations and Future Directions

As this article demonstrates, research on nationalism is rich in empirical findings and, increasingly, in theoretical sophistication. The literature does, however, suffer from conceptual confusion and a tendency to conflate specific aspects of nationalism with the concept as a whole. In what follows, we identify possible areas for further theoretical and methodological elaboration that may help overcome some of these limitations.

The first step for future research is to take seriously the role of meaning-making in people's collective understandings of the nation. Drawing on insights from cultural sociology and cognitive psychology, scholars should shift their focus from specific attitudes to the relational structures of meaning that constitute collective representations of the nation. People are not merely patriots or chauvinists – their schemas of the nation are likely to combine the four dimensions of nationalism (attachment, identity, pride, and hubris) in complex webs of meaning. Mapping these cultural networks and understanding their causal role in political behavior and policy change is an important task that is beginning to be taken up by nationalism scholars (e.g., Bonikowski, 2011, 2013; Centeno and Abascal, 2012).

A second important objective for future research is to gain a better understanding of the relationship between cognitive and affective components of nationalism. Nationalism has long been understood as an ideology that evokes a complex array of emotions, such as fervent pride toward the nation-state, revulsion against perceived out-groups, or superiority compared to other national populations. Indeed, without

the deeply emotional content of nationalist beliefs, it would be difficult to explain the more extreme forms of behavior fueled by nationalist ideology, such as people's willingness to risk their lives at wartime in the name of the national community. But emotions are likely to play an important role in routine forms of nationalism as well with important consequences for political behavior and intergroup relations (Feinstein, 2013). As the empirical study of emotions becomes increasingly prevalent in cultural sociology and political sociology, the timing is opportune for a more explicit focus on affect in nationalism studies as well.

Third, nationalist beliefs need to be studied over time. For much of its development, the literature on nationalism has treated its object of study as temporally stable. This is as true in studies of country-level nationalist ideologies, where essential features of national character are presumed to date back to a given nation's founding moment, as it is in survey analyses, which implicitly view individuals' responses as indicative of their time-invariant attitudinal orientations. Yet, these stability assumptions are in tension with a growing volume of research that shows the intensity and content of nationalist ideas to change in reaction to exogenous shocks to the nation, shifts in political discourse, and life-course events. To resolve these contradictions, more work needs to be done on the dynamic aspects of nationalism and particularly on situational factors that increase the salience of national identities for individuals and collectivities (Collins, 2012).

Finally, in addition to studying nationalism itself, researchers should also gain a better understanding of the relationship between nationalism and other political belief systems. For instance, a large literature in political science and political sociology has become interested in the rise of right-wing extremism in Western European polities (Taggart, 2000; Mudde, 2007). These studies often refer to the radical parties using seemingly interchangeable labels, such as populist, nationalist, conservative, xenophobic, and anti-immigrant. It is clear that supporters of these political movements subscribe to particular conceptions of the nation-state and that the rhetoric of the movement leaders is intended to resonate with such beliefs. At the same time, however, nationalism is only one element of the multifaceted ideological strategy utilized by extreme parties. Just how nationalist concerns interact with other arguments that do not necessarily evoke the nation is an important topic for future study.

Conclusion

Despite pressures from globalization, the nation-state remains a predominant unit of political governance, cultural integration, and collective identification. Indeed, the international nation-state system is so thoroughly institutionalized that most people view it as a natural way of organizing human society. Yet, the meanings attributed to the nation vary considerably within and across country borders. Over the past two decades, scholars have begun to investigate this variation from a number of complementary perspectives. The resulting studies have produced a wide range of consistent findings about the effects of multiple aspects of nationalism on intergroup relations and public policy preferences, as well as the impact of sociodemographic and country-level predictors on nationalist attitudes. Although this work has its limitations, it has opened the way toward new approaches that can better account for the multidimensionality, dynamism, and emotional content of nationalism, as well as its ongoing interaction with other salient political ideologies. Given these developments, the future of research on contemporary nationalism looks promising.

See also: Collective Identity; Commemoration; Culture, Cognition and Embodiment; Formal Methods of Cultural Analysis; Nationalism, Historical Aspects of: The West; Nationalism, Sociology of; Nationalism: General; Networks and Meaning; Organizations and Culture; Religion: Nationalism and Identity; Religious Nationalism and the Secular State; Symbolic Boundaries; War and Nationalism.

Bibliography

Anderson, Benedict, 1983. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, New York.

Ariely, Gal, 2011. Constitutional patriotism, liberal nationalism and membership in the nation: an empirical assessment. Acta Politica 46, 294–319.

Ariely, Gal, 2012. Globalization, immigration and national identity: how the level of globalization affects the relations between nationalism, constructive patriotism and attitudes toward immigrants? Group Processes & Intergroup Relations 15, 539–557.

Bale, Tim, 2013. More and more restrictive—but not always populist: explaining variation in the British conservative party's stance on immigration and asylum. Journal of Contemporary European Studies 21, 25–37.

Billig, Michael, 1995. Banal Nationalism. Sage Publications, Thousand Oaks, CA.

Blank, Thomas, 2003. Determinants of national identity in East and West Germany: an empirical comparison of theories of the significance of authoritarianism, anomie, and general self-esteem. Political Psychology 24, 259–288.

Blank, Thomas, Schmidt, Peter, 2003. National identity in a United Germany: nationalism or patriotism? An empirical test with representative data. Political Psychology 24, 289–312.

Blee, Kathleen, 2002. Inside Organized Racism: Women in the Hate Movement. University of California Press, Berkeley, CA.

Bonikowski, Bart, 2011. Toward a Theory of Popular Nationalism: Shared Representations of the Nation-State in Modern Democracies (Ph.D. dissertation). Department of Sociology, Princeton University, Princeton, NJ.

Bonikowski, Bart, 2013. Varieties of Popular Nationalism in Modern Democracies: An Inductive Approach to Comparative Research on Political Culture. WCFIA Working Paper. Weatherhead Center for International Affairs, Harvard University, Cambridge, MA.

Brubaker, Rogers, 1996. Nationalism Reframed: Nationhood and the National Question in the New Europe. Cambridge University Press, New York.

Brubaker, Rogers, 2004. Ethnicity without Groups. Harvard University Press, Cambridge, MA.

Burris, Val, Smith, Emery, Strahm, Ann, 2000. White supremacist networks on the Internet. Sociological Focus 33, 215–235.

Centeno, Miguel A., Maria Abascal, 2012. The Geography of American Nationalism. Paper Presented at the 2012 ASEN Conference on Nationalism, Ethnicity, and Boundaries, London, UK.

Ceobanu, Alin M., Escandell, Xavier, 2008. East is West? National feelings and anti-immigrant sentiment in Europe. Social Science Research 37, 1147–1170.

Citrin, Jack, Reingold, Beth, Green, Donald P., 1990. American identity and the politics of ethnic change. Journal of Politics 52, 1124–1154.

Citrin, Jack, Haas, Ernst B., Muste, Christopher, Reingold, Beth, 1994. Is American nationalism changing? Implications for foreign policy. International Studies Quarterly 38, 1–31.

Collins, Randall, 2012. Time-bubbles of nationalism: dynamics of solidarity ritual in lived time. Nations and Nationalism 18, 383–397.
De Figueiredo, Rui J.P., Elkins, Zachary, 2003. Are patriots bigots? An inquiry into the vices of in-group pride. American Journal of Political Science 47, 171–188.
Díez Medrano, Juan, Gutiérrez, Paula, 2001. Nested identities: national and European identities in Spain. Ethnic and Racial Studies 24, 753–778.
Feinstein, Yuval, 2013. Nationalism, Emotions, and Support for Military Action: Evidence from a Survey-Based Experiment (unpublished manuscript).
Fligstein, Neil, Polyakova, Alina, Sandholtz, Wayne, 2012. European integration, nationalism and European identity. Journal of Common Market Studies 50, 106–122.
Gellner, Ernest, 1983. Nations and Nationalism. Cornell University Press, Ithaca, NY.
Green, Eva G.T., Sarrasin, Oriane, Fasel, Nicole, Staerklé, Christian, 2011. Nationalism and patriotism as predictors of immigration attitudes in Switzerland: a municipality-level analysis. Swiss Political Science Review 17, 369–393.
Hartz, Louis, 1964. The Founding of New Societies. Harcourt, Brace & World, New York.
Hechter, Michael, 2000. Containing Nationalism. Oxford University Press, New York.
Higham, John, 2002[1955]. Strangers in the Land: Patterns of American Nativism, 1860–1925. Rutgers University Press, New Brunswick, NJ.
Hjerm, Mikael, 1998. National identities, national pride, and xenophobia: a comparison of four Western countries. Acta Sociologica 41, 335–347.
Hjerm, Mikael, 2001. Education, xenophobia, and nationalism: a comparative analysis. Journal of Ethnic and Migration Studies 27, 37–60.
Hjerm, Mikael, 2003. National sentiments in Eastern and Western Europe. Nationalities Papers 31, 413–429.
Huddy, Leonie, Khatib, Nadia, 2007. American patriotism, national identity, and political involvement. American Journal of Political Science 51, 63–77.
Jones, Frank L., Smith, Philip, 2001. Diversity and commonality in national identities: an exploratory analysis of cross-national patterns. Journal of Sociology 37, 45–63.
Kaufmann, Eric, 2000. Ethnic or civic nation? Theorizing the American case. Canadian Review of Studies in Nationalism 27, 133–154.
Kohn, Hans, 1944. The Idea of Nationalism: A Study in its Origins and Background. Macmillan, New York.
Kosterman, Rick, Feshbach, Seymour, 1989. Towards a measure of patriotic and nationalistic attitudes. Political Psychology 10, 257–274.
Kunovich, Robert M., 2009. The sources and consequences of national identification. American Sociological Review 74, 573–593.
Lamont, Michèle, Molnár, Virág, 2002. The study of boundaries in the social sciences. Annual Review of Sociology 28, 167–195.
Li, Qiong, Brewer, Marilynn, 2004. What does it mean to Be an American? Patriotism, nationalism, and American identity after September 11. Political Psychology 25, 727–739.
Lipset, Martin S., 1990. Continental Divide: The Values and Institutions of the United States and Canada. Routledge, New York.
Meinecke, Friedrich, 1970[1907]. Cosmopolitanism and the National State. Princeton University Press, Princeton, NJ.
Mudde, Cas, 2007. Populist Radical Right Parties in Europe. Cambridge University Press, Cambridge, UK.
Myrdal, Gunnar, 1944. An American Dilemma: The Negro Problem and American Democracy. Harper & Row, New York.
Parker, Christopher S., 2007. On the Meaning and Measurement of Patriotism: A Group-Based Assessment (unpublished manuscript). Robert Wood Johnson Foundation. University of California, Berkeley, CA.
Parker, Christopher S., 2010. Symbolic versus blind patriotism: distinction without difference? Political Research Quarterly 63, 97–114.
Rahn, Wendy, 2000. Affect as information: the role of public mood in political reasoning. In: Lupia, A., McCubbins, M.D., Popkin., S.L. (Eds.), Elements of Reason: Cognition, Choice, and the Bounds of Rationality. Cambridge University Press, New York, pp. 130–150.
Schatz, Robert T., Staub, Ervin, Lavine, Howard, 1999. On the varieties of national attachment: blind versus constructive patriotism. Political Psychology 20, 151–174.
Schildkraut, Deborah, 2003. American identity and attitudes toward official-English policies. Political Psychology 24, 469–499.
Schildkraut, Deborah, 2011. Americanism in the Twenty-First Century: Public Opinion in the Age of Immigration. Cambridge University Press, New York.
Shulman, Stephen, 2002. Challenging the civic/ethnic and West/East dichotomies in the study of nationalism. Comparative Political Studies 35, 554–585.
Sidanius, Jim, Feshbach, Seymour, Levin, Shana, Pratto, Felicia, 1997. The interface between ethnic and national attachment: ethnic pluralism or ethnic dominance? The Public Opinion Quarterly 61, 102–133.
Smith, Rogers, 1997. Civic Ideals: Conflicting Visions of American Citizenship in U.S. History. Yale University Press, New Haven, CT.
Staub, Ervin, 1997. Blind versus constructive patriotism: moving from embeddedness in the group to critical loyalty and action. In: Bar-Tal, D., Staub., E. (Eds.), Patriotism in the Lives of Individuals and Nations. Nelson-Hall, Chicago, IL, pp. 213–228.
Stryker, Sheldon, Burke, Peter, 2000. The past, present and future of an identity theory. Social Psychology Quarterly 63, 284–297.
Swidler, Ann, 1986. Culture in action: symbols and strategies. American Sociological Review 51, 273–286.
Taggart, Paul, 2000. Populism. Open University Press, Buckingham.
Waldstreicher, David, 1997. In the Midst of Perpetual Fetes: The Making of American Nationalism, 1776–1820. University of North Carolina Press, Chapel Hill, NC.
Wimmer, Andreas, 2008. The making and unmaking of ethnic boundaries: a multilevel process theory. American Journal of Sociology 113, 970–1022.
Wright, Matthew, 2011a. Diversity and the imagined community: immigrant diversity and conceptions of national identity. Political Psychology 32, 837–862.
Wright, Matthew, 2011b. Policy regimes and normative conceptions of nationalism in mass public opinion. Comparative Political Studies 44, 598–624.

Natural Disasters: Health-Related Aspects

Jan D Reinhardt, Institute for Disaster Management and Reconstruction of Sichuan University and Hong Kong Polytechnic University, Chengdu, Sichuan, China
James E Gosney, International Society of Physical and Rehabilitation Medicine (ISPRM), Geneva, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Natural disasters are sudden ecological disruptions or threats exceeding the adjustment capacity of affected communities. Natural disasters impact on public health directly by causing death and traumatic injuries and indirectly through compromise of public infrastructure, disruption of health care delivery, and displacement of affected populations. Health consequences of natural disasters include fatalities, traumatic injuries, communicable disease, psychological disorders and distress, and long-term disability. Public health in disaster-affected regions can be managed with a cycle encompassing response, recovery, mitigation, and preparedness phase.

Natural disasters are sudden ecological disruptions or threats that exceed the adjustment capacity of the affected community and require external assistance (Lechat, 1979; WHO, 1980). Natural disasters can be broadly classified into categories including geophysical such as earthquakes and volcanic eruptions; hydrological such as floods; meteorological such as hurricanes; climatological such as heat and cold waves and droughts; and biological such as epidemics (Guha-Sapir et al., 2012). Natural disasters considerably impact the health of affected populations directly – by injuries sustained due to collapsing buildings in an earthquake, for example – and indirectly through compromise of public health infrastructure and services with loss of basic needs such as drinking water, nutrition, sanitation, and shelter. Health workers and officials may also be killed or displaced (WHO et al., 2011d).

Health Consequences

Public health consequences of natural disasters include mass fatalities and casualties, communicable disease, long-term physical impairments, and mental health conditions including anxiety, posttraumatic stress disorder, and depression. Preexisting chronic illnesses and disabilities may also be exacerbated.

Fatalities

From 2003 to 2013, over 1.1 million people died due to natural disasters worldwide with earthquakes accounting for more than half of fatalities (see **Figure 1**). Main causes of death vary by type of disaster and include direct causes such as trauma due to building collapse (earthquakes) or flying debris (hurricanes) and drowning (floods, hurricanes); indirect causes include motor vehicle accidents (Doocy et al., 2013a,b,c). Evidence on risk factors for mortality due to disaster is mostly inconclusive. Children, the elderly, and persons with disability are vulnerable groups (Doocy et al., 2013a,b,c; Reinhardt et al., 2011). Males have been reported to be at higher risk of death in developed countries while females are exposed to an increased risk in developing countries (Doocy et al., 2013a,b,c).

Natural disasters can result in large numbers of fatalities, which exceed the local capacity for mortuary management,

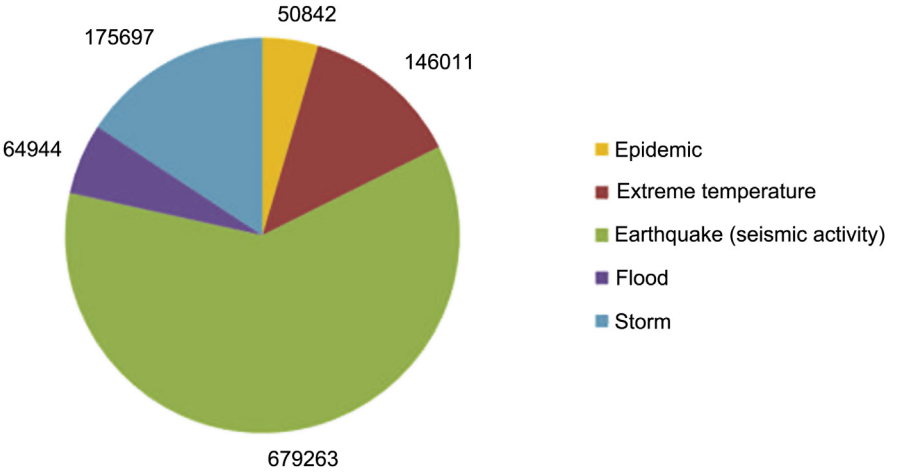

Figure 1 Absolute numbers of deaths from natural disasters by disaster type; 2003–13. Data extracted from the Centre for Research on the Epidemiology of Disasters online database EM-DAT: The International Disaster Database. http://www.emdat.be/database (accessed 02.02.14.).

necessitating mass fatality management (WHO et al., 2011c). Immediate health risks to the general public from dead bodies are often overestimated, however. Secondary health risks to injured populations, victims' families, and health care workers include injuries sustained when moving bodies or debris and psychological distress which may be exacerbated when funeral rites are delayed or not held. Diversion of health workers from treatment of living victims to management of fatalities also compromises public health in the wake of a disaster (WHO et al., 2011c).

Casualties

The number of persons injured exceeds deaths in most natural disasters and a significant increase in the ratio of injuries to deaths over time has been reported. This trend may primarily reflect advances in disaster preparedness and response (Reinhardt et al., 2011). Casualties following natural disasters are mainly due to blunt trauma, crush injuries, near drowning, and psychological sequelae (WHO et al., 2011b). Severe traumatic include fractures, spinal cord injury, and traumatic brain injuries; less-serious soft tissue injuries such as lacerations and muscle strain are common. Since survival probability from serious injury significantly decreases after 24 h, rapid extrication, on-scene stabilization, and transportation to a definitive care facility is critical to survival. Appropriate surgical and medical intervention also helps ensuring survival and preventing long-term consequences including permanent disability and mental disease (Leaning and Guha-Sapir, 2013; Li et al., 2012). Mass casualty triage, the prioritization of treatment of victims based on severity of injury, chances of survival, and long-term health prognosis, is required given limited medical resources and other health constraints in a disaster (Iserson and Moskop, 2007).

Communicable Disease

Largely misreported by the media, communicable disease transmitted by dead bodies is in fact a myth (Watson et al., 2007). The risk of infectious disease in natural disasters may be significantly increased by disrupted or nonexistent public health services and increased limited access, especially in rural areas. Lack of adequate predisaster disease surveillance and control programmes, delayed launch of post-disaster programmes, and low levels of immunity to vaccine-preventable diseases also contribute strongly. Population displacement with resulting overcrowding, lack of clean water, inadequate nutrition, and poor sanitation are a significant source of infectious disease in natural disasters. Evidence for outbreaks of communicable disease after disaster has been best documented for floods (Watson et al., 2007; WHO et al., 2011a). Communicable diseases transmitted in natural disasters are categorized in **Table 1**.

Mental Health

Psychological distress resulting from adverse health consequences, loss of home and livelihood, and death of family and friends result in a high incidence of post-traumatic stress disorder and other disaster-related mental conditions including anxiety disorders, depression, psychosis, and substance abuse in disaster populations. Persons may also experience exacerbation of preexisting conditions due to additional disaster-related distress complicated by disruption of medication and psychotherapy treatment plans as well as loss of general social support. Adequate treatment of postdisaster mental health needs is extremely challenging given often underdeveloped predisaster mental health systems, the high burden of mental disease, difficulty in establishing postdisaster programs, and the complexity of treatment plans including length of treatment time required (Jones et al., 2007; WHO, 2013a).

Long-Term Physical Disability

Long-term physical disabilities due to natural disasters are primarily related to physical impairments resulting from traumatic injuries sustained in the disaster, which have persistent or otherwise long-lasting effects. Inadequate management of other physical injuries and communicable disease as well as mental health conditions can also result in long-term physical impairments. While the focus of medical disaster response has

Table 1 Communicable disease categories associated with natural disaster

Category	*Cause*	*Examples*
Waterborne diseases	Lack of safe water, drinking contaminated water, poor sanitation, increased proximity to rats	Diarrheal disease (i.e, cholera, hepatitis A/E) Leptospirosis
Vectorborne diseases	Increased vector population (i.e, new mosquito-breeding sites due to standing water), increased risk due to weakened immunity due to malnutrition	Malaria Dengue fever Japanese encephalitis
Diseases due to crowding	Increased person-to-person transmission in overcrowded habitation sites, lack of available vaccines, increased risk due to weakened immunity due to malnutrition	Measles Neisseria meningitidis Acute respiratory infections
Other vaccine-preventable diseases	Contact with infectious agents including bacteria while lacking immunity	Tetanus Polio Diphtheria Pertussis

been mostly on surgical and primary care, chronic and preexisting disabling health conditions may lead to a larger burden of disease in the long term (WHO et al., 2011d). 'Disability' in this context indicates impairment in body structure and function and also includes limitation in daily activities and restriction in participation in society (WHO, 2001). Activities limited include functional activities of daily living such as walking, grooming, and toileting and societal restrictions compromise education, employment, and other opportunities for socialization. Significant traumatic disabling injuries include spinal cord injury, traumatic brain injury, burn injury, and extremity injuries such as amputations, bone fractures, and peripheral nerve injury (Reinhardt et al., 2011; Gosney et al., 2011). Exacerbation of existing physical impairments is commonplace in disasters as persons with disabilities are especially vulnerable (Reinhardt et al., 2011; WHO, 2013b). Characterization of long-term physical impairment following a disaster reflects the prevalence of injury type and therefore the primary mechanism of injury as well as demographic variables unique to the disaster including its severity, proximity to urban population centers, and time of occurrence. The 2010-Haiti earthquake, for example, resulted in such catastrophic mortality, morbidity, and burden of disability due to the high number of buildings collapsing in the densely congested capital city of Port-au-Prince during the workday (Gosney et al., 2013). Long-term physical impairments due to natural disasters can be minimized by early use of physical rehabilitation as demonstrated by improved positive rehabilitation outcomes and reduced preventable secondary complications for disabling injuries including spinal cord injury and amputation (WHO, 2013c; Knowlton et al., 2011; Zhang et al., 2013).

Emergency Management Cycle and Health Actions

Health consequences of natural disasters and other public health emergencies can be managed using a comprehensive emergency management model consisting of a cycle of response, recovery, mitigation, and preparation phases (**Figure 2**) (Hoyle, 2009). Management of mass fatalities and casualties, communicable disease, mental health conditions, exacerbation of preexisting conditions, and associated long-term physical impairments during each phase of the response depends on unique disaster characteristics and medical resource availability among other factors. Whereas a less severe disaster in a developed region with a strong public health infrastructure may be effectively managed with local resources only, a more severe disaster in a developing region results in a greater negative health impact on individuals and society and generally requires international assistance. Regardless of disaster context, effective management of health consequences across the disaster cycle of health action requires strong coordination and collaboration between responders and responding organizations at all levels of emergency management.

During the response phase, immediate medical management priorities include on-site triage of severe traumatic injuries with life-saving surgical and medical care provided at an appropriate level medical facility. Patients receive continuing care including

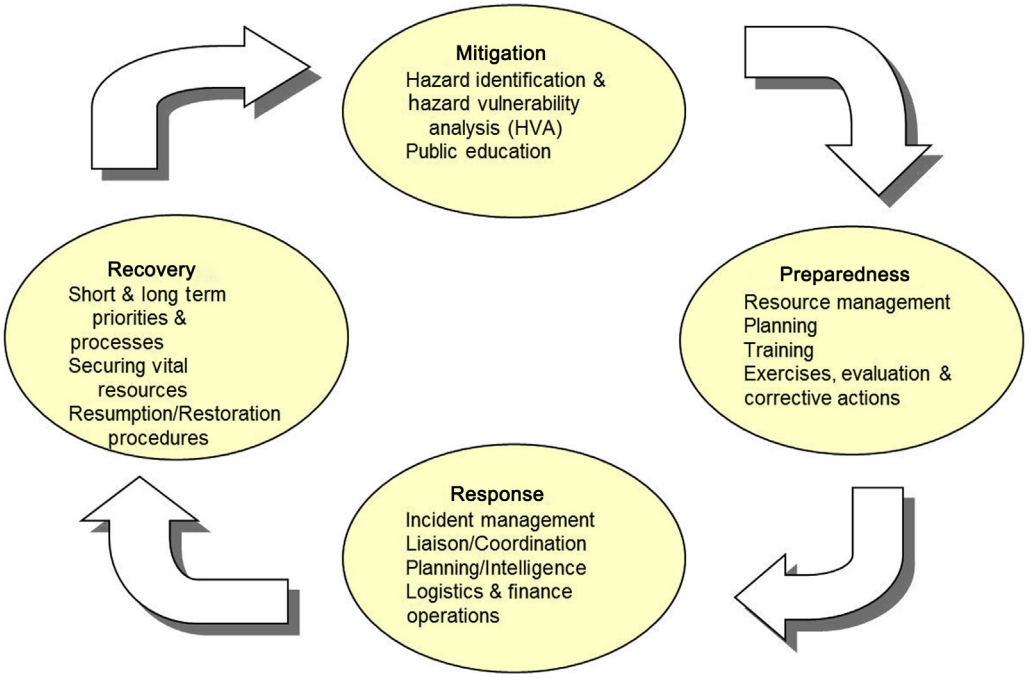

Figure 2 Disaster cycle of health action. Hoyle, J., 2009. Health care facility disaster management. In: Koenig, K., Schultz, C. (Eds.), Koenig and Schultz's Disaster Medicine Comprehensive Principles and Practices, Cambridge University Press, Cambridge, pp. 285–311. Copyright © 2010 Kristi L. Koenig and Carl H. Schultz. Reprinted with the permission of Cambridge University Press.

physical rehabilitation as needed until sufficiently stable for discharge. Outpatient treatment including proper wound care and rehabilitation is necessary to maximize functional outcomes and prevent complications. Community-based supports and services including educational, vocational, and social rehabilitation are also required to optimize patient integration into the community and resume a full, meaningful life (WHO, 2010). Routine population emergency and primary care health requirements must be concurrently addressed following the disaster as well, further complicating postdisaster health delivery. Concomitant with clinical activities focused on individual care, nonclinical health care response activities may include: assessment of evolving casualty patterns and medical resource requirements; management of patient triage, discharge, referral, and tracking systems; data collection, management and analysis; collaboration with other health care service providers; and coordination with emergency system and host health system and government managers.

During the postdisaster recovery phase, the additional health care burden of direct disaster injuries as well as the ongoing provision of routine emergency and primary care should be optimized, necessarily requiring strengthening of health care system capacities. Disaster-related communicable disease as well as mental health conditions and long-term physical impairments are also treated. Planning for future long-term physical impairment needs to reduce the total population burden of disability is critical to community recovery.

Mitigation of health consequences in disasters includes predisaster and emergency activities intended to limit the severity and number of casualties. Emergency planning for persons with disabilities must consider medication and equipment availability as well as evacuation and sheltering strategy to reduce secondary casualties in a disaster, for example (WHO, 2013b).

Disaster preparedness measures focus on improving health care and community-based rehabilitation system sustainability overall and disaster responsiveness specifically. Absence of disaster plans and policy has been shown in fact to compound earthquake severity of impact on elements of the local health care system (Bremer, 2003). Public health disaster response plans should address all major areas of health consequences aforedescribed, outline triage policies, and strategically map required human and physical resources including designated facilities for expansion/conversion to accommodate casualties. Plans should be refined through simulation exercises to maximize their efficacy during the actual disaster.

After a large scale natural disaster, usually funds are available from donations and nongovernmental organizations, governmental actors may be more open to innovations, and technological know-how is provided by national and international aid agencies. While natural disasters cause significant mortality and morbidity, they thus also provide opportunities for rethinking, rebuilding, and strengthening health systems including mental health and physical rehabilitation services, which are often less developed before the disaster occurs (WHO, 2013a; Zhang et al., 2013).

See also: Disability: Social and Psychological Aspects; Environmental Health and Safety: Social Aspects; Functioning, Disability and Health, International Classification of; Globalization and Health; Hazards and Disaster Research in Contemporary Anthropology; Health Interventions, Community-based; Injuries and Accidents: Psychosocial Aspects.

Bibliography

Bremer, R., 2003. Policy development in disaster preparedness and management: lessons learned from the january 2001 earthquake in Gujarat, India. Prehospital and Disaster Medicine 18, 372–384.

CRED, 2014. Centre for Research on the Epidemiology of Disasters. EM-DAT: The International Disaster Database (online). http://www.emdat.be/database (accessed 02.02.14.).

Doocy, S., Daniels, A., Murray, S., Kirsch, T.D., 2013a. The human impact of floods: a historical review of events 1980–2009 and systematic literature review. PLoS Currents 5.

Doocy, S., Daniels, A., Packer, C., Dick, A., Kirsch, T.D., 2013b. The human impact of earthquakes: a historical review of events 1980–2009 and systematic literature review. PLoS Currents 5.

Doocy, S., Dick, A., Daniels, A., Kirsch, T.D., 2013c. The human impact of tropical cyclones: a historical review of events 1980–2009 and systematic literature review. PLoS Currents 5.

Gosney, J., Reinhardt, J.D., Haig, A.J., Li, J., 2011. Developing post-disaster physical rehabilitation: role of the World Health Organization liaison sub-committee on rehabilitation disaster relief of the international society of physical and rehabilitation medicine. Journal of Rehabilitation Medicine 43, 965–968.

Gosney, J.E., Reinhardt, J.D., Von Groote, P.M., Rathore, F.A., Melvin, J.L., 2013. Medical rehabilitation of spinal cord injury following earthquakes in rehabilitation resource-scarce settings: implications for disaster research. Spinal Cord 51, 603–609.

Guha-Sapir, D., Hoyois, P., Below, R., 2012. Annual Disaster Statistical Review 2012: The Numbers and Trends (online). http://cred.be/download/download.php?file=sites/default/files/ADSR_2012.pdf (accessed 12.01.14.).

Hoyle, J., 2009. Health care facility disaster management. In: Koenig, K., Schultz, C. (Eds.), Koenig and Schultz's Disaster Medicine Comprehensive Principles and Practices. Cambridge University Press, Cambridge, pp. 285–311.

Iserson, K.V., Moskop, J., 2007. Triage in medicine, part I: concept, history, and types. Annals of Emergency Medicine 49, 275–281.

Jones, L., Asare, J., Elmasri, M., Mohanraj, A., 2007. Mental health in disaster settings. British Medical Journal 335, 679–680.

Knowlton, L.M., Gosney, J.E., Chackungal, S., Altschuler, E., Black, L., Burkle Jr., F.M., Casey, K., Crandell, D., DEMEY, D., Di Giacomo, L., Dohlman, L., Goldstein, J., Gosselin, R., Ikeda, K., Le Roy, A., Linden, A., Mullaly, C.M., Nickerson, J., O'connell, C., Redmond, A.D., Richards, A., Rufsvold, R., Santos, A.L., Skelton, T., Mcqueen, K., 2011. Consensus statements regarding the multidisciplinary care of limb amputation patients in disasters or humanitarian emergencies: report of the 2011 humanitarian action summit surgical working group on amputations following disasters or conflict. Prehospital and Disaster Medicine 26, 438–448.

Leaning, J., Guha-Sapir, D., 2013. Natural disasters, armed conflict, and public health. New England Journal of Medicine 369, 1836–1842.

Lechat, M.F., 1979. Disasters and public health. Bulletin World Health Organization 57, 11–17.

Li, Y., Reinhardt, J.D., Gosney, J.E., Zhang, X., Hu, X., Chen, S., Ding, M., Li, J., 2012. Evaluation of functional outcomes of physical rehabilitation and medical complications in spinal cord injury victims of the Sichuan earthquake. Journal of Rehabilitation Medicine 44, 534–540.

Reinhardt, J.D., Li, J., Gosney, J., Rathore, F.A., Haig, A.J., Marx, M., Delisa, J.A., 2011. Disability and health-related rehabilitation in international disaster relief. Global Health Action 4, 7191.

Watson, J.T., Gayer, M., Connolly, M.A., 2007. Epidemics after natural disasters. Emerging Infectious Diseases 13, 1–5.

WHO, 1980. Emergency care in natural disasters. Views of an international seminar. WHO Chronicle 34, 96–100.

WHO, 2001. International Classification of Functioning, Disability, and Health (ICF). WHO Press, Geneva.

WHO, 2010. Community-based Rehabilitation: CBR Guidelines – Introductory Booklet (online). http://whqlibdoc.who.int/publications/2010/9789241548052_introductory_eng.pdf (accessed 20.01.14.).

WHO, 2013a. Building Back Better: Sustainable Mental Health Care after Emergencies (online). http://apps.who.int/iris/bitstream/10665/85377/1/9789241564571_eng.pdf (accessed 10.02.14.).

WHO, 2013b. Guidance Note on Disability and Emergency Risk Management for Health (online). http://www.who.int/hac/techguidance/preparedness/disability/en/ (accessed 14.10.13.).

WHO, 2013c. International Perspectives on Spinal Cord Injury (IPSCI). WHO Press, Geneva.

WHO, United Kingdom Health Protection Agency, Partners, 2011a. Disaster Risk Management for Health: Communicable Disease (online). http://www.who.int/hac/events/drm_fact_sheet_communicable_diseases.pdf (accessed 12.01.14.).

WHO, United Kingdom Health Protection Agency, Partners, 2011b. Disaster Risk Management for Health: Management of Mass Casualties (online). http://www.who.int/hac/events/drm_fact_sheet_mass_casualty_management.pdf (accessed 20.01.14.).

WHO, United Kingdom Health Protection Agency, Partners, 2011c. Disaster Risk Management for Health: Mass Fatalities/dead Bodies (online). http://www.who.int/hac/events/drm_fact_sheet_mass_fatalities.pdf (accessed 20.01.14.).

WHO, United Kingdom Health Protection Agency, Partners, 2011d. Disaster Risk Management for Health: Overview (online). http://www.who.int/hac/events/drm_fact_sheet_overview.pdf (accessed 14.01.14.).

Zhang, X., Reinhardt, J.D., Gosney, J.E., Li, J., 2013. The NHV rehabilitation services program improves long-term physical functioning in survivors of the 2008 Sichuan earthquake: a longitudinal quasi experiment. PLoS One 8, e53995.

Relevant Websites

http://www.handicap-international.org/ – Handicap International.

http://www.isprm.org/collaborate/who-isprm/rehabilitation-disaster-relief/ – International Society of Physical and Rehabilitation Medicine: Disaster Relief.

http://www.redcross.org/what-we-do/disaster-relief – American Red Cross: Disaster Relief.

http://www.who.int/hac/en/ – World Health Organization: 10 years recovery from the 2004 tsunami.

http://www.who.int/disabilities/cbr/guidelines/en/ – World Health Organization: Community-based rehabilitation guidelines.

Natural Law

Melissa Moschella, The Catholic University of America, Washington, DC, USA
Robert George, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R.D. Schwarz, volume 15, pp. 10388–10392, © 2001, Elsevier Ltd.

Abstract

The natural law tradition of thought about morality and jurisprudence has its roots in ancient Greek and Roman philosophy, was developed and systematized by Scholastic thinkers, particularly Thomas Aquinas, and continues to be influential to the present day. This article focuses on the central tradition of natural law thinking, outlining its basic features with particular attention to its political and legal implications as well as its relation to the social and behavioral sciences.

The natural law tradition of thought about morality and jurisprudence has its roots in ancient Greek and Roman philosophy, was developed and systematized by Scholastic thinkers, particularly Thomas Aquinas, and continues to be influential to the present day. While the concept of natural law played an important role in early modern moral and political theory, particularly in the works of Grotius, Pufendorf, Hobbes, and Locke, these thinkers' understanding of natural law represented a radical departure from the central tradition of moral and political thought that reached a high point in the work of Aquinas. This article, therefore, will focus on natural law theory within the central tradition, outlining its basic features with particular attention to its political and legal implications as well as its relation to the social and behavioral sciences.

The core of natural law theory is the view that, through rational reflection, human beings can discern principles of right action that direct us to order our wills toward integral human fulfillment, the overall well-being of all human beings as persons and communities. These principles hold universally in all times and cultures because the basic requirements of human well-being are rooted in an essentially unchanging human nature, although many concrete applications of these principles may vary with circumstances. Natural law thus includes a theory of value – a substantive account of human well-being in its various dimensions – as well as a theory of right action – moral rules and principles that guide and structure ethical decision making in concrete cases. It also involves an account of virtues understood as the habits of mind and character that reliably dispose one to choose and act in accordance with moral norms and that in themselves constitute an important aspect of human flourishing, but the focus here will be primarily on the first two aspects.

Natural Law as a Theory of Value

Natural law theories seek to provide objective and universal accounts of the basic and constitutive elements of human well-being, while at the same time recognizing that human good, although determinate, is multifaceted, thus leaving room for a diversity of good lives. What are these basic human goods or values that constitute a flourishing life? How are they identified? What is their relation to human nature? While there is dispute among natural law theorists about the precise answer to these questions, a common starting point is Question 94, Article 2 (q. 94, a. 2), of Thomas Aquinas's *Summa Theologiae*. There, Aquinas argues that the natural law has various precepts because it orders human beings toward a variety of goods that reason apprehends as intrinsically valuable, that is, as choiceworthy in themselves and not merely as means toward further ends. These goods correspond to and are perfective of the various dimensions of human nature. Because humans are bodily beings, life and health are inherent aspects of our well-being. Because we are social beings, friendship, the formation of communities, marriage, and family life all have intrinsic value for us. Because we are rational beings, our fulfillment involves the acquisition of knowledge, both practical and theoretical, including knowledge of the truth about God.

Contemporary natural law thinkers, such as Germain Grisez, John Finnis, and Joseph Boyle, have attempted to offer a fuller explanation of the nature of these basic goods, as well as a more complete list of them, since Aquinas mentions them only in passing as if by way of example, without any pretensions to an exhaustive treatment of the topic. Thus, in addition to the goods listed here, which are all more or less explicitly mentioned by Aquinas, Finnis speaks of the good of performative excellence, the goods of integrity (inner harmony among the reasoning and desiring aspects of the self) and authenticity (harmony between one's judgments and one's actions), as well as the good of harmony with God understood as the transcendent source of existence and meaning (Finnis, 2011).

There is some controversy among contemporary natural law theorists regarding how these goods are known and how they relate to human nature. Aquinas speaks of these basic goods as "those things to which man has a natural inclination" (q. 94, a. 2), leading some to claim that natural law bases moral norms on sub-rational desires inherent in human nature, thus denying the Humean fact–value distinction by implying that an *ought* can be derived from a mere *is*. This is one of the criticisms levied against natural law theories by H.L.A. Hart in *The Concept of Law*. Some contemporary natural law thinkers, such as Ralph McInerny and Alasdair MacIntyre, argue that the fact–value distinction is false and that moral norms are indeed derived from facts about human nature (MacIntyre, 1999, 2007; McInerny, 1982).

Even these accounts, however, are far from the crass naturalism of which critics like Hart accuse them, for they have a sophisticated metaphysical understanding of nature that distinguishes only certain essential aspects as normatively relevant. On this view, nature, and human nature in particular, is normative not 'as it happens to be' but 'as it could be if it realized its goal or *telos*.' And the *telos* or perfection of human nature is, ultimately, discovered by reason.

Other contemporary natural law thinkers, including Grisez, Finnis, and Boyle, accept the fact–value distinction and deny that natural law, including Aquinas's own formulation of it, runs afoul of that distinction when properly understood. For Aquinas, human nature is defined by rationality; what is most natural to human beings, in Aquinas's sense of the term 'natural,' is to act in accordance with reason. The natural inclinations to which Aquinas refers, as corresponding to the human goods to which the first principles of practical reason direct our choices, are not the inclinations of sub-rational desire but rather the inclinations of the will, a rational appetite that tends toward intelligible goods as apprehended by reason (Brock, 2005). Thus, while goods like life, knowledge, and friendship are intrinsic to the flourishing of all human beings because of the nature that we have and because they are perfective of that nature, Grisez, Finnis, and others who follow their account argue that practical reason directly grasps the choiceworthiness of these goods, rather than claim that our knowledge of what is inherently valuable for human beings is derivative of our prior understanding of anthropological facts. In other words, the choiceworthiness of these goods is grasped immediately by one who properly understands them and is not impeded in understanding by prejudice, partisanship, or some other factor (George, 1999).

On this view, the social and behavioral sciences can offer detailed descriptions of human attitudes, inclinations, and behaviors, their typical consequences, and the biological or cultural factors that typically engender them, but these descriptions in themselves cannot be the basis for value judgments. Nonetheless, the social sciences are not irrelevant to discerning the basic and constitutive aspects of human well-being. Rather, as Finnis puts it, they are helpful insofar as they offer "reminders of the range of possibly worthwhile activities and orientations" (Finnis, 2011: p. 81). And, while the mere fact of universal or near-universal agreement about a value is insufficient to establish it as such, comparative anthropology does provide evidence of broad cross-cultural and transhistorical appreciation of the values that Aquinas and contemporary natural law theorists point to as basic and constitutive forms of human well-being – as one would expect if, indeed, basic human goods are as these theorists describe them.

Natural Law as a Theory of Right Action

Recognition that goods like life, knowledge, and friendship are choiceworthy in themselves as distinctive aspects of human flourishing corresponds to what natural law thinkers call the first principles of practical reason, principles of the form 'X is to be done and pursued, and its contrary to be avoided,' where X is any one of the basic goods. They are first principles because they are the starting point for all practical reasoning, reasoning about how to choose and act. In this sense, they are analogous to the first principles of theoretical reasoning, such as the principle of noncontradiction (a thing cannot both be and not be at the same time and in the same respect), which all reasoning necessarily presupposes. These first practical principles in themselves, however, distinguish only intelligible from unintelligible (irrational) choices and actions, not morally right from morally wrong actions. Any action in pursuit of a basic good is intelligible (and, to that extent, rational) but not necessarily moral. A surgeon, for instance, might perform unnecessary operations on her patients for the sake of practicing and improving her skills. Although obviously immoral, such conduct is hardly irrational or unintelligible, as would be the conduct of someone who for no reason at all devoted his life to hopping on one foot all day in the corner of his room.

How, then, does the theory of value discussed in the section Natural Law as a Theory of Value form the basis for concrete norms that guide morally significant choice and action? As a preliminary matter, it is worth mentioning here that one of the core differences between natural law thinking in the central tradition and the natural law theories typical of the early modern and modern periods is that, in the central tradition, law's force (i.e., its binding power or obligatoriness) lies essentially in its reasonableness, rather than in the will or command of a superior – of any authority, human or divine, understood as external to reason itself. As Aquinas puts it, law is essentially an 'ordinance of reason'; its reasonableness is the source of its binding power *as law* (as opposed to brute force) (q. 90, a. 1, 4).

Thus, according to Finnis, the first and most general moral principle is "be fully reasonable," where reasonableness is understood not merely instrumentally as a matter of choosing efficient means toward pregiven ends but also, rather, as including within itself a substantive account of the ends worth choosing, of the various distinct and irreducible goods that are constitutive of human flourishing. To choose and act reasonably, then, are to choose and act in ways that respect the intrinsic value (choiceworthiness) of *all* of the basic goods for *all* human beings – that is, to choose and act in ways that are compatible with the ideal of integral human fulfillment (Finnis, 2011: p. 451).

A number of concrete moral norms flow from this master moral principle. Since, for instance, practical reason directs us toward each of the basic goods and away from their contraries, a direct choice to damage human well-being in any of its basic forms is always unreasonable, and therefore immoral – i.e., against the natural law. This norm accounts for the existence of absolute moral prohibitions that are binding always and everywhere, such as the prohibition on direct killing of innocent human beings. (Some, including Finnis, hold that this prohibition includes *all* direct killing, but all natural law thinkers agree that the natural law strictly prohibits at least direct killing of the innocent.) It is important to note here that these absolute moral prohibitions apply only to choices and actions *directly* against the basic goods, not choices that give less priority to one good rather than other, or that accept damage to one good as a foreseeable side effect of pursuing another good. Indeed, since human beings have limited time and energy, and the

basic goods are irreducibly distinct, acting for any one good necessarily means forgoing the opportunity to act for another. Choosing to spend an afternoon visiting a sick friend entails choosing not to spend that afternoon studying or going to the gym. Such a choice, however, is not *directly* against the good of knowledge or health, even if one knowingly risks catching the friend's illness by visiting.

Being fully reasonable, of course, involves constraints not only on what one may directly will or choose to do, but also on the acceptance of the foreseeable side effects of one's actions. Since each good is a distinct and, in the abstract, equally choiceworthy form of human well-being, and since each good is grasped as inherently valuable not only for oneself but for all human beings, arbitrariness in prioritizing one good over another and in favoring the well-being of one person over another is always unreasonable. Hence the moral requirement, for instance, to develop and follow a rational plan of life that takes into account one's talents, dispositions, realistic possibilities, obligations to other people, and so forth, and to establish one's priorities accordingly; hence also the moral requirement of fairness, traditionally expressed by the Golden Rule.

In part because friendship and sociability are themselves intrinsically valuable as well as instrumentally indispensable for human well-being, and in part because the basic forms of human well-being are grasped as valuable not only for oneself but for all human beings, full practical reasonableness also implies a moral requirement to favor and foster the common good of one's communities. It is important to point out here that a natural law understanding of the common good is completely different from the utilitarian notion of the greatest good (or 'happiness') for the greatest number. Indeed, on the natural law view, this utilitarian ideal is not only wrong but also incoherent, because the distinctive forms of human well-being are irreducible to one another or to any more basic value. Thus, they are in a certain sense incommensurable as they figure as options in morally significant choosing. Even if the choice to visit a sick friend rather than go to the gym is morally right, that does not mean that the goods one participates in (the good of friendship, among others) by doing so include all of the 'goodness,' and more, that one would have gained by going to the gym. There simply is no common currency in which the basic goods can be added to or subtracted from and weighed against one another, as utilitarianism would require. Nor, on the natural law view, are individual human beings fungible repositories of 'goodness' or 'utility' such that one person's good can be weighed against the good of others in a way that could justify, say, killing (i.e., intending the death of) one person in order to save 1000. Rather, according to natural law theory, the good of each individual is a constitutive element of the common good, and the common good is just the set of conditions within which all members of the community, as individuals and in association with one another, can pursue their well-being. Further, contributing to the common good is itself a constitutive element of the well-being of any individual, and in the concrete this generally means fulfilling one's particular obligations to others (as individuals and groups) to and for whom one has a particular responsibility by virtue of one's professional, social, familial, religious, and/or civic ties.

Natural Law and Human Rights

By providing (1) a substantive account of the basic and fundamental forms of human well-being, (2) moral norms that require respect for each of these basic values both in oneself and in others, and (3) a conception of the common good in which the good of each individual is a constitutive aspect, natural law theory offers a firm basis for human rights, including rights that are absolute (i.e., universal, inalienable, and inviolable), avoiding an undue proliferation of ungrounded rights claims as well as the putative antagonism between individual rights and the common good. According to natural law, the language of rights is simply a way of expressing what is right or just – that is, of expressing the requirements of justice – from the perspective of the person or group to whom something is owed. Thus, absolute rights are, on this view, simply the flip side of absolute moral prohibitions. The claim that human beings have an absolute right to life and bodily integrity (i.e., a right against being killed, maimed, or tortured as an end in itself or as a means to some end), for instance, is simply a translation into rights-language of the claim that there is an absolute duty to refrain from direct killing (or maiming) of innocent human persons. Similarly, absolute negative rights (rights that require only noninterference on the part of others) correspond to all of the basic aspects of human well-being – hence the right of association, the right to pursue one's profession, the right to marry and raise one's children, the right to practice one's religion, as well as rights of conscience more broadly. These rights are absolute insofar as they involve an absolute immunity from actions on the part of others (individuals or communities) whose object is precisely to impede one's participation in the relevant good. Further, acting in ways that *indirectly* harm others with respect to their participation in basic values – i.e., not directly intending such harm but only accepting it as a side effect – is a violation of rights insofar as there is a lack of fairness – i.e., insofar as one would not be willing to accept the harm in question if it were to fall on oneself or one's own.

For example, one implication of the right to life is the norm of noncombatant immunity in war. Governments can run afoul of this norm – and thus violate human rights – not only by *directly* targeting noncombatants but also by a lack of fairness in their willingness to accept noncombatant deaths as a side effect of attacks on legitimate military targets. Such a lack of fairness would be involved if, for instance, the government were to order a daytime attack on a munitions factory located next door to an elementary school, when it would refrain from such an attack if, instead of the factory being located near an elementary school, it were located near a prisoner-of-war camp. In other words, the government would be acting unfairly by treating the lives of 'enemy' children as much less important than those of its own soldiers being held prisoner.

Natural law also provides the basis for so-called 'positive' rights, rights that are not simply immunities but that involve a claim of entitlement to receive certain benefits or services from others. Given the understanding of human well-being that is essential to natural law theory, as well as natural law theory's view of human beings as naturally social due in part to human interdependence and vulnerability, it makes sense to speak of human rights in regard to goods like healthcare, adequate food

and shelter, education, and so on. Yet such positive rights are much more indeterminate than negative rights. While negative rights involve a clear claim on *everyone* to refrain from violating the right in question, to make positive rights meaningful it is necessary to specify precisely *who* has the obligation to provide the benefits or services in question, and precisely *how* to do so, and these questions cannot be answered in the abstract, for they will vary depending on social, cultural, political, and legal factors, as well as individual circumstances. The principles of natural law thus leave much room here for prudential judgment on the part of both individuals and political authorities.

Natural Law and Authority

Natural law theories also offer an account of legitimate authority and of the obligation to obey the law that avoids some of the well-known and seemingly intractable difficulties of basing legitimate authority on actual or hypothetical consent. Beginning with early modern social contract theorists such as Pufendorf, Hobbes, and Locke and continuing into the contemporary period, most famously with the hypothetical contract theory proposed by John Rawls in a *Theory of Justice*, there is an assumption that to be a part of an authoritative community – a community with some degree of decision-making power over its members – is, although ultimately necessary for survival and well-being, fundamentally unnatural to human beings. For authoritative communities to exist legitimately (rather than merely on the basis of brute force), they must be based upon the consent of their members, who are understood as free and equal individuals subject to no preconsensual authority structures.

By contrast, natural law theories within the central tradition view human beings as naturally enmeshed within a web of authoritative communities at different levels – the family, religious communities, fraternal and civic associations, and the political community – whose legitimacy is not based primarily on consent but on their inherent connection to the common good, which natural law requires each individual to protect and promote. On the natural law view, justifying political authority is not fundamentally a matter of consent but, rather, of recognizing that solving coordination problems among individuals, families, and particular associations is necessary for the common good, and that the only practicable way to solve such problems is through authority. This would be true, on the natural law view, even in a society of saints, because the practical problems involved in coordinating to pursue common aims admit many reasonable solutions. What practical reasonableness requires is only that *one* of the many possible solutions be adopted and followed by all, and this can be done only if authority is established (be it a person, a group, or a particular decision-making procedure such as majority vote). Thus, reason requires obedience to authority insofar as that authority is on balance actually helping to solve coordination problems in view of the common good and has not degenerated into tyranny. Consent, on this account, need not be a part of the justification of authority at all. All that is necessary to generate an obligation to obey political authority (or any authority) is the recognition that such authority is in fact conducive to the common good.

Natural Law and Positive Law

It is a famous natural law dictum, articulated by Augustine and reiterated by Aquinas, that "an unjust law is no law at all." Legal positivists, including H.L.A. Hart, have harshly criticized this view, arguing that law is simply a matter of social fact and that laws have no necessary connection to morality, because patently unjust laws clearly can and do function as laws insofar as they are accepted as law and treated as binding by both officials and community members (Hart, 2012). Yet the natural law claim that an unjust law is no law at all does not conflict with the claim that, as a matter of social fact, unjust rules do function as laws. Nor does the natural law view deny that there is value in providing a purely descriptive sociological account (as Hart himself proposes to do) of what it means for a particular social rule to function as a law. The natural law theorists' claim, rather, is that the unjust social rule that recognizably functions as a law is not a law in the fullest – i.e., moral – sense of the term, and thus that it is possible to criticize that law as failing precisely to fulfill the justifying purpose of law as such (i.e., as failing to advance the common good, in part by not establishing a fair balance of benefits and burdens among the community's members) (George, 2007).

According to the natural law view, a positive law is just to the extent that it is in accordance with the natural law. This does not mean, however, that positive laws are simply straightforward deductions from the natural law, that the positivity of a law adds nothing to its moral force, or that the positing of a law cannot sometimes itself create moral obligation where it would not otherwise exist. With regard to the mode of their derivation from natural law, positive laws can be divided into two categories. The first category consists of positive laws that flow directly from the requirements of natural law, such as laws against murder. The second category consists of positive laws that relate more indirectly to natural law by way of what Aquinas calls the 'determinations' of the legislator. As Aquinas explains, such laws are derived from natural law not, as opposed to those in the first category, like "conclusions deduced from general principles," but like "implementations [*determinationes*] of general directives" (q. 95, a. 2). Thus, for instance, natural law requires that some resources be pooled for the sake of public goods such as bridges and roads, but natural law does not indicate any specific way of doing this. Indeed, the prudent legislator will take into account the particular social and cultural conditions of the community in an attempt to devise the most just and expedient means of pooling resources. What results – a set of tax laws, for instance – is a determination or specification of the natural law requirement to favor and foster the common good (in this case, through the contribution of material resources) that is binding only for the members of a particular community. In this sense, such laws are binding entirely because they are posited, even though ultimately their normative force comes from the permanent principles of practical reason. The choice (determinatio) of rulers will thus determine what is uniquely just for those under their rule.

Precisely because the normative force of all laws ultimately flows from their relationship to the requirements of natural law, laws fail *as laws* to the extent that they are in conflict with the requirements of natural law. Hence the claim that an unjust law is not law in the full moral sense of the term, even though it

may indisputably function as such as a matter of social fact. According to Thomas Aquinas in the *Summa Theologiae* (q. 96, a. 4), there are several ways in which laws can be unjust. The first is procedural: laws can be unjust if they are not enacted in accordance with the established decision-making procedures or go beyond the limits of the lawmaker's rightful authority. Laws can also be unjust on substantive grounds in two ways. First, they can directly contradict the natural law, as did genocidal laws in Nazi Germany or laws maintaining the institution of slavery in the pre–Civil War United States. Second, they can fail to fairly distribute the benefits and burdens of common life. Note that established decision-making procedures themselves can also be substantively unjust on grounds of fairness.

A famous application of this understanding of the proper relationship between natural law and positive law is found in Martin Luther King's 'Letter from a Birmingham Jail,' in which he defends his acts of civil disobedience against those who criticize him for failing to show sufficient respect for the law (King, 1963). Citing Thomas Aquinas, King states that "an unjust law is a human law that is not rooted in eternal law and natural law." Glossing this in a way that highlights the inherent connection between natural law and human flourishing, and paralleling the connection that should exist between positive law and the common good, King writes, "Any law that uplifts human personality is just. Any law that degrades human personality is unjust". He then goes on to point out the obvious truth that segregation laws are a clear instance of the latter, and thus are unjust and do not merit respect as laws. Aside from his criticism of segregation laws as directly contradicting the natural law, King also points to lack of fairness as an independent ground for the injustice of segregation laws: "An unjust law is a code that a numerical or power majority group compels a minority group to obey but does not make binding on itself." He argues further that the laws in question are procedurally unfair because "devious methods are used to prevent Negroes from becoming registered voters." In other words, King makes a comprehensive case that the laws against which he is protesting are truly not laws at all in the morally relevant sense because they violate all of the criteria, both substantive and procedural, for just laws.

Yet King also follows Aquinas and the natural law tradition more broadly in recognizing that, even when faced with patently unjust laws that cannot be obeyed in good conscience, respect for the system of law as a whole (provided the system is not completely corrupt) – and thus for legitimate authority as conducive on the whole to the common good – must nonetheless be maintained. That is why King stipulates that "one who breaks an unjust law must do so openly, lovingly, and with a willingness to accept the penalty." He thus distinguishes himself from the anarchist by arguing that his acts of civil disobedience – in which he openly defies a patently unjust law and accepts imprisonment for doing so, with the aim of "[arousing] the conscience of the community over its injustice" – are "in reality expressing the highest respect for the law." Only because he holds this view of positive law as deriving its binding character *as law* (as opposed to brute force) from its essential relationship to the natural law can King make such a claim.

Indeed, at the heart of the natural law tradition's criticism of legal positivism is the argument that, by denying any essential connection between law and morality (between positive law and natural law), legal positivists make it ultimately impossible to distinguish law from brute force, or to distinguish the civil disobedience of someone like King from mere anarchism (Simon, 1965). Further, as Hart points out, any sophisticated sociological understanding of law must be able to account for the fact that those who administer and are subject to laws view them as essentially different from mere commands backed by threats of force. By conceiving of law as inherently aimed at the common good and as having normative force precisely insofar as it achieves this purpose, natural law theorists claim to offer a more accurate description of law as social fact, as well as to provide broader moral standards according to which the justice of any particular law or legal system can be critically assessed.

Natural law reasoning has been applied to contemporary controversies ranging from bioethical issues like abortion (Kaczor, 2011; Lee and George, 2009), euthanasia (Keown, 2002), and embryonic stem cell research (George and Tollefsen, 2008) to marriage policy (Girgis et al., 2012), capital punishment (Grisez, 1970), torture (Tollefsen, 2008), nuclear deterrence (Finnis et al., 1987), and property and welfare rights (Boyle, 2001).

See also: Civil Liberties and Human Rights; Injustice: Legal Aspects; Justice and Law; Legal Positivism; Rights: Legal Aspects.

Bibliography

Aquinas, Thomas. Summa Theologiae, vol. I–II, q. 90–97.
Boyle, Joseph, 2001. Fairness in holdings: a natural law account of property and welfare rights. Social Philosophy and Policy 18 (1), 206–226.
Brock, Stephen L., 2005. Natural inclination and the intelligibility of the good in Thomistic natural law. Vera Lex 6, 57–78.
Finnis, John, 2011. Natural Law and Natural Rights, second ed. Oxford University Press, New York.
Finnis, John, Boyle, Joseph, Grisez, Germain, 1987. Nuclear Deterrence, Morality and Realism. Oxford University Press, New York.
George, Robert, 1999. In Defense of Natural Law. Oxford University Press, New York.
George, Robert, 2007. Natural law. American Journal Of Jurisprudence 52, 55–75.
George, Robert, Tollefsen, Christopher, 2008. Embryo: a Defense of Human Life. Doubleday, New York.
Girgis, Sherif, Anderson, Ryan, George, Robert, 2012. What is Marriage? Man and Woman: a Defense. Encounter Books, New York.
Grisez, Germain, 1970. Toward a consistent natural law ethics of killing. American Journal Of Jurisprudence 15, 64–96.
Hart, H.L.A., 2012. The Concept of Law, third ed. Oxford University Press, New York.
Kaczor, Christopher, 2011. The Ethics of Abortion. Routledge, New York.
Keown, John, 2002. Euthanasia, Ethics and Public Policy. Cambridge University Press, Cambridge.
King Jr., Martin Luther, April 16, 1963. Letter from a Birmingham Jail. American Friends Service Committee, Philadelphia.
Lee, Patrick, George, Robert, 2009. Body-Self Dualism in Contemporary Ethics and Politics. Cambridge University Press, New York.
MacIntyre, Alasdair, 1999. Dependent Rational Animals. Open Court, Chicago.
MacIntyre, Alasdair, 2007. After Virtue, third ed. University Of Notre Dame Press, Notre Dame.
McInerny, Ralph, 1982. Ethica Thomistica. The Catholic University Of America Press, Washington, DC.
Simon, Yves, 1965. The Tradition of Natural Law. Fordham University Press, New York.
Tollefsen, Christopher, 2008. Biomedical Research and Beyond: Expanding the Ethics of Inquiry. Routledge, New York.

Relevant Website

http://www.nlnrac.org – Natural Law, Natural Rights and American Constitutionalism.

Naturalism

Daniel Andler, Université Paris-Sorbonne, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Naturalism is a deep trend in philosophy, rather than a precise doctrine, with multiple connections in many areas of science and culture. Its starting point is the belief that nothing exists outside nature. It has been developed throughout history along two main directions. Ordinary naturalism relies on commonsense and everyday human practices to uncover reality, while scientific naturalism, dominant today, takes nature to be what physics tells us exists, and natural science to be the sole provider of genuine knowledge. Cognitive science and evolutionary biology attempt to account for mental and social phenomena in naturalistic terms.

Naturalism

Naturalism is such a protean concept that calling it a doctrine is already committing a Procustean reduction. A bewildering array of philosophers and philosophies are regarded, or consider themselves as exemplifying naturalism, with hardly any further commonality, or so it seems, between them. Yet naturalism is also sharp enough an indicator that just about anyone in philosophy or in the sciences of man is sensitive to the naturalistic valence of whatever theory, author, approach they happen to be confronted with. Many authors see it as so vague and equivocal as to be essentially useless; yet the same authors use the term in their own productions. This need not be an inconsistency: in context, naturalism may take on a meaning precise enough for theoretical purposes. Still, there is a peculiar versatility of the term, which is often obfuscated in expositions that focus right away on a certain situated understanding, and give no indication of there being other meanings, deemed just as central in different contexts. Nor is this versatility a mere accident: the issue of naturalism is inescapable and brings out an indefinite range of responses from thinkers in philosophy and the sciences of man from all walks of academic life.

At the highest level of generality, naturalism is a stance; the stance leads to one of two broad heuristic principles; each of these eventuates in any one of a large variety of doctrines, which in turn range from abstract speculations to well-articulated research programs. This article will focus on what happens to be, in the present intellectual mood, the most salient variety of naturalism, viz. scientific naturalism, seen as a cohesive set of theoretical commitments giving rise to, and getting support from, a cohesive set of research programs in philosophy and the sciences of man (which are taken here to include the social sciences). But first the broader landscape will be briefly sketched; it is the proper background against which the major positions are being debated and will continue to be in the foreseeable future, as the present dominance of scientific naturalism is increasingly contested while the research programs that it inspires and that support it show no sign of weakening.

Naturalism in General: One Stance, Two Branches, Three Positions

As a stance, naturalism presents a united front. To adhere to naturalism, one must, first, believe or be willing to bet that nothing exists outside of nature and grant that conjecture a central position among one's theoretical commitments; second, have the courage to rebel against received views; third, to have at least a conjecture regarding how to pursue the goal of showing that nature exhausts reality. Naturalists are rebels because at all times, in one or another realm of thought, the received view promotes some form of dualism, one branch of which escapes nature – as if the struggle to establish the unity of reality were never definitively won.

Of course, both nature and reality are contentious concepts, and to equate the two is merely the starting point of a reflection which branches out in many directions. Still, one thought that is common to all forms of naturalism is this: only by looking at nature, at what surrounds us and stretches under our eyes, can one ascertain what actually exists. It is a metaphor for sure, and the central difficulty emerges as one tries to cash out the visual element. Nature is not antecedently given: it is gradually revealed to the inquiring subject. The naturalist seems faced with a dilemma. Either she claims to know what inquiry is, or she does not. If she does, she has a conceptual grasp on nature, as that which inquiry gradually reveals, but she has already forsaken naturalism: inquiry is not 'visible' in any extended sense, so her knowledge must stem from, or belong to another realm. If she does not, she has no conceptual grasp on nature, because whatever nature is, it includes both visible or more broadly perceptively salient things, but not all such things, as well as a wealth of nonperceptively salient things, and only inquiry can sort all existing things from the rest. But without an epistemic grasp on nature, it is hard to see what sense can be made of the claim that nature is all that exists.

There is a way out of the dilemma, championed by Quine (1969). The naturalist does not start from scratch: she has a pretty secure hold on parts of nature, and she is acquainted with some procedures of inquiry. The way to go then is to lean on the means of inquiry one already possesses and trusts in order to broaden one's knowledge of nature, and to lean on one's broadened understanding of nature in order to augment

one's understanding and means of inquiry. Success of this bootstrapping strategy is not guaranteed. In fact, it may be argued that it rests on a postulate of continuity: what we do not yet know must in some way resemble what we do know, as it is our present means of inquiry that are to reveal it, and by the same token, the new means of inquiry must in some sense be continuous with those we already have. The continuity postulate sits well with naturalism, which favors unity and connectedness over all forms of essential duality. And it clashes with some received views: of nature as composed of semi-independent orders, and (more starkly) as inquiry evolving by leaps, under the combined effect of criticism, whereby a method of inquiry abruptly appears as unacceptable, dragging in its fall entire patches of previously held beliefs about nature, and of creative thought, bringing into being novel means of inquiry that in turn disclose unsuspected areas of nature.

It is customary to distinguish ontological from methodological naturalism: the former (sometimes called metaphysical naturalism) is a claim on what (actually) exists, as above, the latter (also sometimes called epistemic naturalism) a claim concerning the (legitimate) way of getting to know reality. As we have just seen, the issues of reality and of inquiry are intertwined; yet it is possible to consider them separately. The possibility arises once a starting point for the bootstrapping process has been chosen. And this is where a major parting of the ways occurs. The crucial question is: Where do we seek guidance as we start off? There are two plausible answers: commonsense or everyday experience, on the one hand, science on the other. Each of these operates as a heuristic. Neither uniquely determines a research program, a theory, or a doctrine. Rather, they orient our thinking in one or the other direction.

When the starting point is the natural stance toward the world and the human experience, consisting in mankind's shared perceptual faculties, cognitive resources, linguistic, and other practices, the heuristic is what might be called ordinary naturalism. An early representative is Thomas Reid, who based his well-known commonsense naturalism, opposed to Hume's, on a conception of commonsense as epistemically privileged. But as a general orientation this kind of naturalism need not be dependent on a Reidian view of commonsense, and focus instead on practice and/or language. It accommodates a wide variety of philosophers, often but not systematically linked to pragmatism or ordinary-language philosophy, from Goethe and Nietzsche to Wittgenstein and W. Sellars, and exemplified today by such authors as Huw Price, John McDowell, Robert Brandom, perhaps Thomas Nagel, and others (see De Caro and Macarthur, 2004).

What those thinkers have in common are their reservations with respect to the other main branch of naturalism, dominant today. More cannot be said in this article. The focus from now on will be that other branch, scientific naturalism. Not that ordinary naturalism lacks relevance for the sciences of man. But its main contributions are critical and foundational: it is in part a rebellion against the excesses of a higher-order dichotomy, one that pits natural science against other forms of understanding. First-order research programs pursued today are overwhelmingly on the side of scientific naturalism, and a large part of contemporary debates concern the value and scope of these proposals.

Scientific naturalism is the heuristic that recommends that the bootstrapping process start with the current natural science as method of inquiry, and natural-scientific knowledge as knowledge base. Ontological (scientific) naturalism claims that all that exists is that whose existence is granted by the natural sciences; methodological (scientific) naturalism claims that natural science is the sole reliable method of inquiry regarding the realm of existing things (for a classic and detailed exposition, see Papineau, 1993).

As remarked earlier, the ontological and methodological claims may seem to be two sides of the same coin, and indeed that is how they are viewed by many of their proponents, be it of the 'ordinary' or scientific persuasion. But as announced, they can be pursued separately, once the starting point of the bootstrap has been fixed. So, for example, someone might be convinced that natural science provides the only access to genuine knowledge without taking a stand on what exists and how that might be determined. And someone may be wedded to the thought that natural science is the arbiter of existence, but be indifferent to the means of access to truth (except of course regarding existence). Indeed, indifference may turn into rejection: the first thinker might actually affirm the existence of a realm unverified and unverifiable by natural science (and thus not susceptible of genuine knowledge), or simply deny that the issue makes sense; similarly, the second thinker might recognize sources of genuine knowledge other than natural science. Parallel possibilities arise for ordinary naturalists.

The 'ordinary' naturalist and the scientific naturalist need not disagree on all counts. The former is willing to accept natural science and its pronouncements regarding some phenomenon as long as they do not clash too sharply with the common view. The latter, in such cases, will accept the common view as an approximation of the correct view procured by science; even in the case of a serious clash, she will try to account for the existence and persistence of the common view. But she will not hesitate to reject it in favor of the current scientific view. The commonsense naturalist, absent a decisive argument against the common view, will treat the scientific naturalist's insistence on granting epistemic priority to the scientific view *also in a contested case* as a *reductio*, showing that natural science has exceeded its scope and that scientific naturalism has turned into scientism.

Among the contested cases are the usual suspects: mind, thought, meaning, consciousness, self, numbers and other abstract objects, norms, culture, practices, language, history, and so on. These are generally regarded as the bones of contention between naturalists and antinaturalists. It might be thought then that ordinary naturalists, reticent as they are to unquestionably accept the authority of natural science, are actually antinaturalists. This need not be the case. Ordinary naturalists mean to preserve naturalism against what they see as the scientific naturalists' mistake of, so to speak, putting all of one's eggs in the scientific basket; while antinaturalists see the whole enterprise as deeply misguided. To be sure, it is often a fine line separating the two groups. The situation is clarified somewhat with the introduction of a new category, one which includes ordinary naturalists next to thinkers who, like them, share a naturalist commitment and believe that the scientific naturalist overstates science's case, but do not necessarily espouse a well-articulated brand of ordinary naturalism. At this

point the various heuristics and orientations must solidify into more precise doctrines, on pain of confusion. The present state of the field is then best characterized as structured around three clusters of positions: 'liberal' or 'broad' naturalism, scientific naturalism, and antinaturalism, with the first standing somewhere between the other two.

Scientific Naturalism

Scientific naturalism ('scientific' will henceforth be omitted), in a nutshell, relies on science, and science alone, to provide a correct fundamental understanding of reality. Science is to be understood as the set of disciplines modeled after the natural sciences as we know them. Physics plays a dual role here: it is the paradigm of natural science, it also serves, in some versions of the doctrine, as the basic science in a reductive sense. However the life sciences, chemistry, and other so-called special sciences such as geology or meteorology are regarded as legitimate providers, whether or not physics is thought to be able in principle to reconstruct their findings. In practice, a vast majority of naturalistic programs would not get off the ground were it not acceptable, from the naturalist's standpoint, to take biology as an alternative model, next to physics. In fact, the main, though not the only reason why the issue of naturalism is notably different today than it was at the beginning of the twentieth century is the rise of biology: on the one hand, it now stands on equal footing with physics as a respectable science, and on the other it is much more hospitable to a natural science of man.

As indicated above, there are two sides to naturalism. Ontological naturalism concerns the 'furniture' – the constituents – of the world. Methodological naturalism concerns the legitimate ways of bringing to light the make-up of reality.

It might appear that ontological naturalism simply follows from methodological naturalism: after all, whether something is or is not part of the world is one question among many that can be asked regarding reality. But it is a very special kind of question, one which requires an answer before science can start working on the answer: science needs to know what to examine, or where to look, in order to get going. As it progresses, it revises its initial ontological commitments and in the fullness of time delivers a (fallible) pronouncement. But that result seems to depend on the starting point, the choice of which calls for a separate examination. This is one reason why ontological naturalism is a semiautonomous debate, and one which, in contradistinction with methodological naturalism, concerns mostly philosophers.

The path leading from the historical starting point of ontological naturalism to the ontological commitments of contemporary natural science is complex, and cannot be sketched here. At this point in history, according to the best-developed version of ontological naturalism, physicalism, the reconstructed starting point is this: existing entities are those located in space and time that can exert a causal influence on, or as some authors put it make a causal difference for (other) existing entities. As for the end result, the main proposal is that physics delivers the ultimate ontological judgments. In other words, there exist only entities expressly warranted by physics or made up of such entities. The main argument in favor of this starting point is the causal closure of the physical world: as our understanding of the material world has progressed, the likelihood of a case where part of the physical world would be affected by a nonphysical factor has become vanishingly small (Papineau, 1993). Moreover, if our epistemic capacity itself is natural, resulting from processes that we can account for in naturalistic terms, then in particular we can only get acquainted with entities that can have a causal impact on our cognitive apparatus. Again, there is a circle, but a virtuous one according to the naturalist: as our understanding of mind and brain progresses, the natural-causal picture of cognition becomes clearer and more persuasive, and that in turn strengthens the naturalistic premise and its conclusion.

How then do prima facie nonphysical things fit in the physicalist's world? The answer varies according, first, to the kind of thing under consideration, and, second, to the kind of naturalism one espouses. Things, on the one hand, can be entities, properties, or facts; on the other, they may belong to various realms: supernatural (in a sense whose paradigm are deities, spirits, occult forces, or principles, and so on), mental, abstract, normative, social, and so on. Entities can variously be eliminated altogether (naturalists, also of the liberal persuasion, are unanimous in eliminating supernatural entities; for example, some, but not all, scientific naturalists eliminate entities attested by folk psychology, such as beliefs or values), or else be shown to be identical to natural entities (for example, a mental state – believing that 7 is a prime number – might be shown to be identical to a state of the brain, or a family of such states; Wall Street culture in the 2010s might be shown to be identical to an evolving network including mental representations in members of a population and public representations including texts, images, tools, practices). Properties might be shown to be natural properties, either basic properties (on the model of water being identical to the natural kind of molecular composition H_2O) or disjunctions of such properties (on the model of a corkscrew being identical to one of a number of mechanical devices); alternatively, they may be shown to *supervene* on natural properties, in such a way that two entities with the exact same natural properties cannot differ with respect to the property under consideration: just as it stands to reason that two physically identical faces cannot differ in their degree of beauty, the proposal is that if one state of the brain, say, underlies resentment, another state of the brain identical, neuron for neuron so to speak, to the first cannot but underlie resentment (while resentment, on the other hand, might have nonidentical realizations as brain states). Finally, facts can be either identified, empirically, with natural facts (a ghostly apparition, for example, might be identified with a certain visual effect produced by the interaction of a natural scene and the witness, or the experience of anger with a rush of adrenalin, or again a reasoning episode with a brain process amounting to some manipulation of material symbols); or they can be seen as nonfacts, either destined for the dustbin of science (the destruction of Troy did not result from Olympian rivalries; people do not succeed because they were born under a lucky star; things do not burn by losing their phlogiston), or reinterpreted (as when the putative moral fact that killing is wrong is regarded as expressing a repugnance toward killing).

Physicalism raises several objections, four of which will be mentioned. (1) Some argue that it is not true to the facts: some entities, despite being assemblies of physical entities obeying

physical laws, exhibit 'emergent' properties, and thus deserve a place of their own in the natural ontology. The fallback is a nonphysicalist, or pluralistic naturalism, which postulates a disjunctive ontology usually comprising the ontologies of the physical and life sciences. But why stop there, it then may be asked? Why not include the ontology of the sciences of man, and take on board their load of prima facie 'suspect' entities, thus all but depriving naturalism of its substance? (2) Physicalism is charged with indeterminacy, on two grounds. (2a) Hempel's dilemma (Hempel, 1980) concerns the physics we take as our ontological guide. If we mean today's physics, we run the risk of having to change our ontology next time physics undergoes a revolution; this is a particularly worrisome prospect for those interested in the mind–body problem: all our present attempts to integrate the mind into today's physical ontology could turn out to be worthless, if, for example, quantum mechanics and cosmology were absorbed tomorrow into a totally novel theory of everything. If on the other hand we mean to refer to the ontology of physics as it will turn out in the fullness of time, then we are referring to nothing at all: who knows what that ontology will even remotely look like? One response to Hempel's dilemma is to throw in doubt the likelihood of such a revolution, and to take the bet that the present attempts to 'physicalize' the mind will not be overthrown by some fanciful sea change in physics (Bokulish, 2011). Another response goes by the name of via *negativa*: the basic intuition driving physicalism is that nonphysical entities have no place in the world as we know it (this is what two or three millennia of experiencing and experimenting seem to have taught us). We know what the main categories of prima facie nonphysical entities are: the supernatural, the spiritual, the mental, the abstract, the normative, the social, and so on. So let us rephrase physicalism as stating that the world consists solely of the nonmental, nonsupernatural, and so on, and that this latter domain is causally closed (Montera and Papineau, 2005). It is not clear that this move strengthens the case of physicalism, which seems to boil down to a conviction, based on the history of natural science, that there is no need for mental (or supernatural, and so on) stuff to account for the workings of the natural world, and that seems to beg the question. (2b) Physicalism seems to rest on an unanalyzed notion of causality, while there is at present little consensus on how causality is to be understood, and how it fits in the picture of fundamental physics, which is often said to be acausal. (3) Finally, the supervenience strategy for fastening 'suspect' entities onto the natural world, while avoiding the pitfalls of reductionism, is too weak a foundation for a full-blooded naturalism, one that would put the 'suspect' entities on the same footing as the regular ones, making them in particular susceptible to the same kind of direct control.

Methodological naturalism stands on firmer ground. First, it is relatively well defined: although there is no agreed upon set of criteria for methods acceptable in the natural sciences, there is little dispute about concrete cases: scientists and philosophers alike are generally in agreement regarding a given natural-scientific procedure – when they see one they know one. Second, it escapes the analog of Hempel's dilemma. On the one hand, methods condoned by the natural sciences are an open-ended, evolving collection, so that there is no danger of the claim dissolving whenever a new method (such as sophisticated software applied to colossal quantities of automatically generated data) comes into force. Nor does this open-endedness make the claim vague: natural-scientific methods form a well-structured set whose elements bear a strong family resemblance, and which moreover strongly support one another; no new member is admitted without having been thoroughly tested. Third, in strong contrast with ontological naturalism, methodological naturalism is, so to speak, self-referentially closed: it claims that only natural-scientific methods should be used, and it applies that demand to itself, insofar as the decision whether or not to accept a candidate method is left entirely to natural science, with no help from outside sources except perhaps philosophy, which naturalists accept only insofar as it is continuous with natural science. Fourth, is does not give any specific discipline pride of place: physics plays no privileged role, so that methodological naturalism need not impose on inquiry the unreasonable demand of conforming to physics. Finally, the demand it does make is rather clear: naturalism is *opposed* to methodologies frequently and sometimes exclusively used today in many branches of these disciplines. Only methodologies that are clearly similar to, or continuous with those used in the natural sciences, are deemed acceptable. This is not to deny that the sciences of man have their proprietary domain, or to call for the elimination of their entities and phenomena in favor of neurons, cells, molecules, or bosons. The demand rather is that no special treatment be granted to those entities and phenomena except as motivated for precise reasons and to a limited extent. The mere fact that they are attached to human beings or to human populations does not qualify them for special treatment. Cells and organisms are different from stones and cascades, yet to the extent that they give rise, in science, to domain-specific methods and concepts, that is not due to an initial decision from which the inquiry proceeds; on the contrary, it is a consequence of that inquiry, operating under the ground rules of natural science. The same goes, according to naturalism, for human beings, minds, cultures, societies: they start out as generic entities, and as the inquiry progressively identifies some of their specific features, the necessary adjustments are made, without in any way relaxing the standards of natural science. This requirement is known as the 'unity of science,' which is not (contrary to popular views) a call for reduction to physics (although some authors have indeed hoped to accomplish this ideal via a reduction of all science, regardless of subject matter, to physics) or to biology (a more popular orientation today). Naturalism, seen in this light, is simply the rejection of a specific duality, that between the natural sciences and the sciences of man, very much ingrained to this day in mainstream academic culture, despite the best efforts of naturalists over the last half-century.

In What Ways Does Scientific Naturalism Make a Difference?

That (scientific) naturalism should make a large difference is intuitively obvious. Yet there are both a variety of ways in which it does, and ways in which it does not. Naturalism makes a difference on at least five distinct issues, only one of which is directly relevant to the present volume.

The first is the existence of supernatural entities, the most prominent of which in our culture is God: naturalism sides

with atheism, and this makes a difference for theists. Yet even theists, when they put on their scientific goggles, fall back on a special kind of methodological naturalism, one which recommends the scientific stance in the daily conduct of one's scientific life: the scientist who believes in God brackets her religious beliefs when she acts as a scientist. This extends to a vast majority of philosophers, as Barry Stroud has emphasized, regardless of the fact that nonbelievers are far from constituting a vast majority among philosophers. It has long been clear that God could well have chosen to proceed precisely by those means that science can identify, and that, until recently, the (in)existence of God made a difference at most in areas where science happens to be silent – a situation which is undergoing rapid, though not necessarily lasting changes, with some radical naturalists' attempts to refute theism (Harris, 2004; Dawkins, 2006; Dennett, 2006).

The second is the quest for the ultimate reality: what are the fundamental components of the world? This is of concern to a small contingent of philosophers and physicists. Naturalism, which claims exclusive authority for natural science on this issue, and attempts to draw conclusions from the present state of fundamental physics, makes little or no difference in most other contexts, whether academic or not.

The third is the role of science in knowledge. Naturalists (of the scientific branch of course) claim an epistemic privilege for science. The weak form of the claim is that wherever science has a say, what it says has final authority, superseding all other beliefs on the same topic. The strong form is that science has 'universal coverage': anything that can be known at all can only be known by science. The weak form is relevant in contexts where rationalism is challenged: it is not an issue *within* the rationalist camp, which includes all scientists and most philosophers. The strong form is solely of philosophical import: many things are at present unknown by science, so the claim is that we are left right now in a situation of total ignorance regarding those things, which cuts no ice for those who, unlike philosophers, are not preoccupied by the issue of what counts as genuine knowledge. Both the weak and the strong forms of the claim leave us with the more pressing demand of drawing the right distinctions within the extensive set of things unknown to science, that the naturalist labels nonknowledge: those of which we have a certain kind of grasp, those which we can conceive but are uncertain about, those which we cannot conceive with any clarity, or at all, those about which we are wrong, etc.

The fourth is the nature of philosophy: here naturalism takes the form of another 'continuity thesis,' according to which philosophy is continuous with natural sciences: it shares or should share their goal, which is to acquire an understanding of the world, it aims or should aim at taking on board the lessons of science, both its results and its methods, and it does not or should not avail itself of anything like a philosophical method differing in essence from the scientific method. Insofar as philosophy remains distinct from any given science, it is due, in the naturalist's perspective, to the kind of question that it asks: too general, or too imprecise at the present stage, to be taken on by any research program in one of the established sciences.

Closely linked to this view of philosophy is the naturalistic turn in epistemology. In a seminal article, Quine (1969) defended the view that knowledge is a natural process, the output of human cognitive activities, which transform sensory stimulations into a certain kind of theoretical output. The externalist conception of knowledge sharpens this insight by attributing the difference between mere belief and genuine knowledge to the reliability of the natural processes leading to the formation of a given thought (Ramsey, 1931; Goldman, 1967).

Naturalism in this sense (sometimes also called methodological) is of direct concern only to philosophers, although it does bear on the relations that philosophy can or should establish with other theoretical and practical enterprises, including the sciences of man.

The sciences of man are the fifth issue in which naturalism makes a difference, and, indeed, it is the one that we shall now be concerned with. It inspires two distinct, though related, projects. The first is to 'naturalize' the mind: to show that the mind and all things mental are natural (Armstrong, 1968; Dennett, 1969; Fodor, 1975; Dretske, 1995). The second is to produce a naturalistic social science (Mill, 1843; Rosenberg, 2008).

On the negative, 'rebellious' side, naturalism takes the form, as was said earlier, of a rejection of the well-entrenched dichotomy between two kinds of science. This raises an immediate perplexity, akin to the one we encountered in relation to what constitutes genuine knowledge: naturalism rules out of bounds a large amount of research conducted in the sciences of man. If we assent to naturalism, for whatever reasons, we are left with the problem of disposing of the body of rejected research, both methods and results. Three main strategies can be deployed: wholesale elimination, wholesale reconstruction, or sorting and correction. Wholesale elimination is sometimes explicitly proposed: whether mainstream research is taken to be mistaken in its methodology – *verstehen* or interpretivism, hermeneutics, historicism, and so on – or in its basic assumptions – cultural relativism, social constructionism, omnipotence of the mind (its alleged capacity to develop any set of thoughts whatsoever), and so on – its output is consigned for the most part to the dustbin of history. But more often elimination proceeds tacitly: the results of mainstream research are essentially put to the side, and the ground is cleared for starting afresh, by picking the low-hanging fruit, accessible from the present stage of development of the natural sciences of sociality, and proceeding stepwise from there. Wholesale reconstruction rests on the thought that social scientists have unwittingly been going down the naturalist's path, while consciously following some official non-naturalist methodology. The aim is to redeploy all or most of the corpus of social science in a proper naturalistic framework, where it would blend seamlessly with results obtained directly from a naturalistic base. Finally, a naturalist social scientist may feel that throwing away the fruits of decades or centuries of work by mostly excellent minds is unreasonable, yet have no idea of how to reconstruct the entire field from first principles: she would then retain certain results that she has independent reasons to find at least approximately true or plausible, and attempt to correct and complete them with the help of natural science. This in fact is what is beginning to happen at the interface of the cognitive and social sciences (a good example is Elster, 2007).

Still, the main task of the naturalist, whether philosopher or social scientist, is to develop a naturalistic research program in the sciences of man, or, in other words, move from proclamation, conviction, or a priori arguments to existence proof: to show how and that a naturalistic account can be given of entities, processes, phenomena that have heretofore been regarded as out of reach of a naturalistic science. This is the stage where the naturalistic stance turns into a form of anchored naturalism, one that aims at making a scientific difference.

Naturalizing the Mind: Cognition, Brain, Evolution

The single most important reason why the discussion of scientific naturalism does not proceed today as it did in previous epochs is that biology has both matured into a full-fledged natural science and become hospitable to the human realm. Or so it was claimed earlier (Section Scientific Naturalism). It is important at this point to qualify this statement: without the rise of the formal sciences, biology would arguably not have developed into a bona fide natural science, and it would unquestionably not have become hospitable to the human realm.

The formal sciences are understood here to cover logic, in the broadest sense of the word, information and computer science, control and signal theory, theoretical linguistics and extend today to probability and statistics, dynamical systems, and mathematical modeling techniques in general. The unity and scope of this vast area is a topic beyond the reach of this article. The decisive point is the development of logic and information theory in the late 1930s, which led in short succession to the computer model of the mind, to generative linguistics, and to the information-theoretical approach in molecular biology and genetics. A few years earlier, statistics developed in such a way as to revolutionize and revive evolutionary theory. By the early 1960s cognitive science was born (though not yet baptized); initially understood as a computational theory of mental capacities, it rapidly expanded in several directions: first, it incorporated developmental psychology, so that the ontogenesis and the mature state of human capacities were seen as connected; second, the scope of cognitive science extended beyond the initial repertoire of learning, memory, reasoning, language, motor control, and absorbed emotions, consciousness, self, social skills; third, non-normal cognition and nonhuman cognition became integral parts of the study of cognition; fourth, partly as a consequence of the last point, neuroscience joined the federation, so that cognitive science crystallized in the mid-1970s into a multilevel, multidisciplinary integrative approach of the 'mind/brain.'

Two further additions were made to this structure beginning in the 1990s. First, the mutual dependence of the cognitive and social sciences became evident; second, evolutionary theory, which for many years was seen as no more than a lateral aspect of the problem situation, came to the fore, to the extent where an 'evolutionary social cognitive neuroscience' is seen by some as the inevitable result, in the coming years, of an ongoing process of multiple convergences. The present situation, though fluid, is more accurately described as a multipronged naturalization program of the human realm.

When the focus is on the individual, member of the *Homo sapiens* species, three avenues are open for investigation. First, the human individual has a complex of mental or cognitive capacities that can be characterized at the functional level of information processing; this is the mission of such fields as developmental psychology, psycholinguistics, experimental psychology; the computational models produced and tested may in principle have counterparts in artificial systems developed in artificial intelligence and robotics. Second, these capacities are undergirded by the brain: not only would there be no mind if there were no brain, but the complex structure of the mind depends on, and in fact faithfully reflects, the complex structure and workings of the brain, as the mind is seen as nothing over and above an integrated system of functions of the brain. An account of the latter amounts to a naturalization of the former. Third, these capacities are by and large functional in the biological sense, and are thus the outcome of biological evolution. Three distinct paths are thus thought to naturalize the mind, providing in the long run, respectively, (1) an abstract mechanical account of the information processing system; (2) a concrete biological account of the material system embodying the system; (3) an evolutionary account of the cognitive organ, described either at the functional, information-processing level (the evolved mind) or at the neural level (the evolved brain).

Are these programs bound to converge? The naturalist certainly thinks so. After all, the human body (and organisms in general) can be studied at the abstract functional level (models of the bodily functions can be constructed), at the anatomical, physiological, and metabolic level, and as a product of evolution by natural selection, and we do expect that these perspectives, as they are filled in, gradually blend into one unique, multidimensional picture. Indeed, this is the horizon in which many cognitive scientists frame their efforts. However there is a major difference: in the case of the human body, naturalization is not, or is no longer, an issue, and because we are antecedently assured of its natural status, these various investigations do not bear the burden of proving it, and by the same token we expect that they will converge in the long run. The case of the mind is different: each of the research programs is meant as a partial vindication of the claim that it is natural, and their convergence, if and when it is achieved, would likely be regarded as settling the issue, and also as showing how the respective ways in which the mind is 'naturalized' fit with one another. In the present stage of progress, no clear-cut conclusions are forthcoming: although our understanding has made impressive progress on many fronts, such that what may be called the 'natural profiles' of the mind are gradually coming into focus, the ultimate goal, namely proving beyond reasonable doubt that the mind is natural through and through in the operational sense of being fully accessible by natural science, remains elusive.

Indeed, how much of the mind is targeted by the naturalization program – what should be included in its final, completed 'natural profile' – is not entirely clear. Strict naturalism sees the mind as a natural system, and relies on natural science to account for both its invariant structure (allowing for development from birth to adulthood) and its particular processes. A (standard) physical system has invariant features and follows certain trajectories under various environmental

conditions, and both the features and the trajectories are under the jurisdiction of physics. Science however does stop short of singular happenings: a leaf falling from a particular tree at a particular time, thoroughly physical an event as it is, does not qualify for treatment by physics. Where to draw the line at which the mission of science ends is a vexed question even for physics (Cartwright, 1983). To a first approximation, the physics of a system stops where the contribution of the environment to the interaction overwhelms, in terms of complexity, that of the system: it includes types of trajectories under a variety of fairly simple environmental conditions, but not specific trajectories in highly complex situations.

When it comes to the mind, we would not expect science to account for the details of every episode of everyone's mental life, but only for the regularities that all such episodes exhibit, and for the shared functional architecture of the systems that produce them. A large part of the task is to define the basic concepts in terms of which to describe both. One proposal stemming from the formal sciences is to regard the mind as a set of information-processing functions operating on inner representations; in this framework, the aim of cognitive science is to characterize the types of mental states and the types of transformations on those states. Both states and processes raise a number of perplexities.

One essentially unsolved problem of particular importance is to provide a naturalistic account of the intentionality of mental states, that in virtue of which they represent (real or fictional) states of affairs – in brief, the challenge is to provide a naturalistic account of meaning (Millikan, 1984; Dretske, 1995). Another unresolved issue is that of consciousness, with many rival contenders vying with limited success for recognition as a naturalistic account (Block et al., 1996). A third conundrum is raised by the status of reasons in actions and behavior: reasons are prima facie distinct from causes, insofar as they obey incommensurable laws, yet enter in the commonsense explanations of action; naturalism demands that actions, and behavior generally, be explained in causal terms; so either, conceptual analysis notwithstanding, my reason for X-ing causes my X-ing, or it does not, despite the fact that absent my reason for X-ing, my X-ing is no longer an action but a mere movement (Davidson, 1980). Meaning and reasons both involve normativity: the mind seems to have a fundamental grasp of normativity, and to keep track of particular norms while orchestrating behavior that does or does not conform to the relevant norms. Making naturalistic sense of normativity appears to some philosophers as the most daunting task (Wedgewood, 2007; De Caro and Macarthur, 2010).

Important as they are to philosophers, these issues are less central for cognitive and social scientists, who are more anxious to find a principled way of drawing a boundary, within the dynamics of the mind, between what is due to the system and what to the environment. The nature/nurture debate, though transformed by what we have learned about development and plasticity, remains a bone of contention between naturalists and various strands of antinaturalists: if the mental dynamics owes a large part of its specific features to contingent or stable features of the history of interactions of the individual with her, mostly social, environment, then naturalizing the mind is an enterprise of limited scope, or so it seems (but see below). Nobody today doubts that there are natural constraints on what the mind can achieve, i.e., constraints on the mental dynamics; but the naturalists insists, and her opponents deny, that these constraints are of great import.

Most cognitive scientists involved in hands-on research, eager to avoid controversies which they see as fruitless, are in fact content with defending a modest naturalism, one which takes no stand on how far the naturalist account will eventually go, and what slack it will leave for the contingencies of the individual's life course, and/or the environmentally transmitted constraints originating in the social sphere. The real challenge lies in discovering the conceptual repertory in which to couch the natural constraints, not in attempting to second guess the extent of their reach.

Naturalizing the Social Sphere

Postulating, as has just been done, a clean separation between individual cognition and the social sphere is a gross oversimplification, and one that prejudges the issue of naturalism regarding the latter. And in fact naturalism is perhaps nowhere more salient as a challenger to established dichotomies than in the social sciences. Indeed, in some circles the mere fact that some phenomenon is under the influence of social factors implies *by stipulation* that it is not natural. This clearly assumes some particular character of *human* society, for what ants or bees do is regarded as no less natural for their being social species. Perhaps humans, also a social species, owe their natural character to both their individual and social constitutions. That is precisely what naturalism proposes; indeed, the most refined forms now under discussion aim to show that the naturalness of the human individual and the naturalness of human sociality are deeply intertwined.

Naturalism deploys three main strategies, either independently or in combination as the case may be, in its attempts to establish the natural character of the social sphere. The first is informational: society as a general phenomenon, as well as particular societies, are to a very large extent determined by the informational dynamics that they harbor. Information should be understood in a theoretical sense, shared with cognitive science, and in particular not limited to linguistic communication: many vehicles, including purely mental ones, carry and propagate meaningful tokens that in part determine both individuals' behaviors and the modification of the informational landscape. Game theory, which purports to provide a formal characterization of strategic interactions between agents, is a typical achievement of this strategy, one in fact which aspires to provide a unifying framework for the social sciences (Gintis, 2009).

The second strategy relies on cognitive science: by offering a detailed characterization of the individual mind, it narrows the set of possible social arrangements and facts. This need not presuppose a strong form of methodological individualism: a principle of individualistic transmission (Nelson, 1990), according to which any putative specifically social factor can have an impact only by impinging on individual members of the social group, is enough to make cognitive science relevant; in fact, it is irrelevant only on a strong form of social holism, claiming that the individual level of causality is screened out, and that the members of a group are indiscernible tokens that act as mere vehicles of social forces. Still, naturalism and

methodological individualism of one or another shade are regular bedfellows. Cognitive science, it is argued, substitutes genuine knowledge of human mental capacities, particularly memory, learning, reasoning, decision making, for intuitive notions or philosophers' speculations. As a result, behavior that is called irrational when it fails to conform to the latter yields to explanation in the light of the former. A new field, misleadingly labeled neuroeconomics, aims at providing an account of decision making based on cognitive neuroscience. Naturalism also harbors the hope to improve people's behaviors, in the sense of making them cohere better with their true preferences, for example by engineering the setting in which they make their choices so as to orient their natural decision processes in a more favorable direction (Thaler and Sunstein, 2009).

The third strategy rests on a general theory of evolution, with two distinct branches, biological and cultural. Biological evolution is called upon to account for the existence of cognitive mechanisms specifically geared to allow social interactions among humans to develop, stabilize, and evolve in particular ways, and which together constitute our shared 'social cognition.' Cultural evolution accounts for the genesis of particular forms of social processes, which generally (though not always) allow human groups to acquire new capabilities and capitalize on ones acquired at an earlier stage (Tomasello, 1999), at a considerably faster rate than biological evolution would permit. Two additional phenomena have recently enriched the explanatory toolkit of evolutionary approaches. The first is group selection, a highly controversial issue: the conjecture is that Darwinian selection can operate at the group level, so that one population may acquire as a whole a gene pool allowing it to outcompete another. Long widely regarded as discredited, the conjecture is alive again (Sober and Wilson, 1999). The second device is gene-culture co-evolution: a given gene pool favors the spread of certain cultural forms, that in turn favor individuals bearing certain genes, giving rise to a new gene pool (Richerson and Boyd, 2005). Niche construction, a phenomenon not limited to, but spectacularly exploited by humans, provides an illustration: the genetic endowment of a population leads it to alter its local environment in such a way that it become hospitable to variants which would otherwise be at a disadvantage. The variants thus flourish and go to fixation, leading to a genetically different population, well adapted to the niche that led to its selection (Odling-Smee et al., 2003). A niche need not be mainly material: those that matter most for humans are cultural.

Group selection and/or gene-culture co-evolution are harnessed for the resolution of several enigmas in standard biological evolution, bringing heretofore resistant phenomena within the purview of naturalism. The most important is cooperation, a central characteristic of humankind. Biological evolution in the Darwinian sense would seem to select exclusively self-regarding behaviors that destabilize in short order any cooperative system in which individuals contribute to the welfare of nonrelatives without expecting the favor to be returned. Another singular feature is the ability of human populations to adjust to an immense variety of habitats, which only a certain type of evolving culture can achieve; in turn, the selection process leading to inscribing in the human genome the ability to adapt to all sorts of environments rather than to a narrow range seems to require a co-evolutionary process. These examples are representative of a new orientation, within evolutionary psychology and anthropology, that anchors the study of culture in biology while granting it semiautonomous status (Richerson and Boyd, 2005; Sterelny, 2012; Wilson, 2002), rather than essentially reducing culture to a direct expression of our genetic endowment (Tooby and Cosmides, 1992; Pinker, 2002).

Naturalism inspires a large array of research programs in the social sciences, whose ambitions range from modest to revolutionary: they aspire to complement, reconstruct, or replace mainstream anti- or non-naturalistic approaches. A large part of the current efforts, and accompanying controversies, partake of methodological naturalism. A resolution of the age-old issue of *verstehen vs erklären* seems within reach, insofar as interpretation can be understood naturalistically: the meaning of human actions and events may in fact be objective properties, which social science can bring to light without renouncing the explanatory goal of natural science. A full reconciliation would seem to depend on several moves that are resisted by many (Taylor, 1971; Geertz, 1983) the discussion of which vastly exceeds the scope of this article. What does rather clearly stand out is the emergence of a significant body of naturalistic theory in the social sciences, which bypasses rather than confronts the interpretative tradition, by setting its own agenda and foregoing, in accordance with the lessons of biology, the demands of lawlike generality and predictability.

Fewer attempts are made to vindicate ontological naturalism. Still, the nature of social entities has come under sustained attention in the last quarter century. Social representations, institutions, norms, joint action, collective intentionality have been the focus of several proposals: Dawkins's theory of memes (Dawkins, 1976), Sperber's (1996) epidemiology of representations, Searle's (1995) theory of institutions as constitutive rules, the competing account of institutions as game-theoretic equilibria, pursued in the wake of Lewis's seminal *Conventions* (1969) by Skyrms (2010), Bicchieri (2006), and others, Gilbert's (1989) notions of plural subject and attached predicates such as collective belief, shared intention, etc., competing with accounts by Bratman (2014), Searle, Tuomela (2007), and others, Goldman's (1999) program for social epistemology, and so on. Here again, philosophical and formal theorists are under increasing pressure to share the field with empirical scientists, who bring data from developmental psychology, behavioral ecology, neuroscience, experimental economics, etc. that naturalism is committed to regard as relevant. Much remains to be done, however, before a clear picture of the viable options emerges.

Here as elsewhere, scientific naturalism in action, much like the sciences themselves, gradually refines its self-conception as it pursues the multiple theoretical, formal, and empirical inquiries inspired by its initial commitments: what 'nature' refers to, what are the admissible ways in which the entire repertory of entities attested in our practices, linguistic and otherwise, is in turn to be inscribed in nature, these are the terms of both the problem and the partial solutions proffered along the way. The process and its fallouts may well be more significant than any putative final vindication.

See also: Causation: Physical, Mental, and Social; Conventions and Norms: Philosophical Aspects; Evolution, Natural and Social: Philosophical Aspects; Evolutionary Approaches to Human Behavior: Philosophical Aspects; Functional Explanation: Philosophical Aspects; Geertz, Clifford (1926–2006); Individualism versus Collectivism: Philosophical Aspects; Joint Action; Mental Causation; Methodological Individualism: Philosophical Aspects; Physicalism and Alternatives; Pragmatism: Philosophical Aspects; Reference and Representation: Philosophical Aspects; Searle, John (1932–); Shared Belief: Philosophical Aspects; Social Ontology; Taylor, Charles (1931–); Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Armstrong, D.M., 1968. A Materialist Theory of the Mind. Routledge and Kegan Paul, London.

Bicchieri, Cr., 2006. The Grammar of Society: The Nature and Dynamics of Norms. Cambridge University Press, Cambridge.

Block, N., Flanagan, O., Guzeldere, G. (Eds.), 1996. The Nature of Consciousness: Philosophical Debates. MIT Press, Cambridge, MA.

Bokulish, P., 2011. Hempel's Dilemma and domains of physics. Analysis 71, 646–651.

Bratman, M., 2014. Shared Agency: A Planning Theory of Acting Together. Oxford University Press, New York.

Cartwright, N., 1983. How the Laws of Physics Lie. Oxford University Press, Oxford.

Davidson, D., 1980. Essays on Actions and Events. Oxford University Press, Oxford.

Dawkins, R., 1976. The Selfish Gene. Oxford University Press, Oxford.

Dawkins, R., 2006. The God Delusion. Bantam Press, London.

De Caro, M., Macarthur, D. (Eds.), 2004. Naturalism in Question. Harvard University Press, Cambridge, MA.

De Caro, M., Macarthur, D. (Eds.), 2010. Naturalism and Normativity. Columbia University Press, New York.

Dennett, D., 1969. Content and Consciousness. Routledge and Kegan Paul, New York & London.

Dennett, D., 2006. Breaking the Spell: Religion as a Natural Phenomenon. Penguin Group, New York.

Dretske, F., 1995. Naturalizing the Mind. MIT Press, Cambridge, MA.

Elster, J., 2007. Explaining Social Behavior: More Nuts and Bolts for the Social Sciences. Cambridge University Press, Cambridge.

Fodor, J.A., 1975. The Language of Thought. Crowell, New York.

Geertz, C., 1983. The Interpretation of Cultures. Basic Books, New York.

Gilbert, M., 1989. On Social Facts. Routledge, London.

Gintis, H., 2009. The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences. Princeton University Press, Princeton.

Glimcher, P.W., Fehr, E. (Eds.), 2013. Neuroeconomics: Decision Making and the Brain, second ed. Academic Press, London, New York.

Goldman, A.I., 1967. A causal theory of knowing. The Journal of Philosophy 64, 357–372.

Goldman, A., 1999. Knowledge in a Social World. Oxford University Press, Oxford.

Harris, S., 2004. The End of Faith: Religion, Terror, and the Future of Reason. Norton, New York.

Hempel, C.G., 1980. Comments on Goodman's ways of worldmaking. Synthese 45, 193–199.

Kim, J., 2008. Physicalism, or Something Near Enough. Princeton University Press, Princeton.

Kitcher, P., 1992. The naturalists return. The Philosophical Review 101, 92–114.

Lewis, D., 1969. Convention. Harvard University Press, Cambridge, MA.

List, C., Pettit, P., 2013. Group Agency: The Possibility, Design, and Status of Corporate Agents. Oxford University Press, Oxford.

Mill, J.S., 1843. A System of Logic. J.W. Parker, London.

Millikan, R.G., 1984. Language, Thought, and Other Biological Categories. MIT Press, Cambridge, MA.

Montera, B., Papineau, D., 2005. A defence of the via negativa argument for physicalism. Analysis 65, 233–237.

Nelson, A.J., 1990. Social science and the mental. Midwest Studies in Philosophy XV, 194–209.

Odling-Smee, F.J., Laland, K.N., Feldman, M.W., 2003. Niche Construction: The Neglected Process in Evolution. Princeton University Press, Princeton.

Papineau, D., 1993. Philosophical Naturalism. Blackwell, Oxford.

Pinker, S., 2002. The Blank Slate: The Modern Denial of Human Nature. Viking.

Quine, W., 1969. Epistemology naturalized. In: Quine, W. (Ed.), Ontological Relativity and Other Essays. Columbia University Press, New York, pp. 69–90.

Ramsey, F.P., 1931. Knowledge. In: Braithwaite, R.B. (Ed.), The Foundations of Mathematics and Other Essays. Harcourt Brace, New York.

Richerson, P., Boyd, R., 2005. Not by Genes Alone: How Culture Transformed Human Evolution. University of Chicago Press, Chicago.

Rosenberg, A., 2008. Philosophy of Social Science, third ed. Westview Press, Philadelphia, PA.

Searle, J.R., 1995. The Construction of Social Reality. The Free Press, New York.

Skyrms, B., 2010. Signals: Evolution, Learning, and Information. Oxford University Press, Oxford.

Sober, E., Wilson, D.S., 1999. Unto Others – the Evolution & Psychology of Unselfish Behavior. Harvard University Press, Cambridge, MA.

Sperber, D., 1996. Explaining Culture: A Naturalistic Approach. Blackwell, Oxford.

Sterelny, K., 2012. The Evolved Apprentice. MIT Press, Cambridge, MA.

Sterelny, K., Joyce, R., Calcott, B., Frasen, B. (Eds.), 2013. Cooperation and its Evolution. MIT Press, Cambridge, MA.

Stroud, B., 1996. The charm of naturalism. Proceedings of the American Philosophical Association 70, 43–55. Reprint in: De Caro, M., Macarthur, D. (Eds.), 2004. Naturalism in Question. Harvard University Press, Cambridge, MA.

Taylor, C., 1971. Interpretation and the sciences of man. Review of Metaphysics 25, 3–51.

Thaler, R.H., Sunstein, C.R., 2009. Nudge: Improving Decisions about Health, Wealth and Happiness. Penguin, New York.

Tomasello, M., 1999. The Cultural Origins of Human Cognition. Harvard University Press, Cambridge, MA.

Tooby, J., Cosmides, L., 1992. The psychological foundations of culture. In: Barkow, J., Cosmides, L., Tooby, J. (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, New York, pp. 19–136.

Tuomela, R., 2007. The Philosophy of Sociality: The Shared Point of View. Oxford University Press, Oxford.

Wedgewood, R., 2007. The Nature of Normativity. Oxford University Press, Oxford.

Wilson, D.S., 2002. Darwin's Cathedral: Evolution, Religion, and the Nature of Society. University of Chicago Press, Chicago.

Nature and Society in Geography

BL Turner, II, Arizona State University, Tempe, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

With a pedigree traced to antiquity, nature–society or human–environment inquiry vied for the identity of the modern discipline of geography as it emerged in nineteenth-century Germany. This identity involved the study of the human impact on and adjustment to the physical environment. For a brief period in the early twentieth century, a particular definition of the relationship, environmental determinism, dominated geographic education in the United States. The excesses and limitations of determinism relegated the nature–society identity in geography relative to a spatial–chorological one. This other and equally old identity remained largely unchallenged from the 1930s through the 1940s and onward. By the late 1970s, however, nature–society studies returned to geography in a significant way, building from interests that gave rise to cultural ecology and risk-hazard studies as well as several other subfields of study. Entering the twenty-first century, nature–society research in geography has grown in ways reflecting a range of perspectives found in the human sciences and humanities. Its future role in the discipline will be determined in part by the rise of new, interdisciplinary human–environment research communities, such as sustainability science.

Nature–society or human–environment relationships have been part of geographic inquiry since antiquity. The modern foundation of the themes of human impact on and adjustment to the physical environment, however, was articulated by nineteenth-century German geographers, with various claims that these relationships, broadly interpreted, constituted the identity of the discipline. For a brief period in the early twentieth century, a particular definition of the relationship, environmental determinism or the geographic factor, formally dominated geographic education in the United States. The excesses of determinism discredited nature–society studies in the discipline, with various attempts to embed human–environment relationships within the dominant spatial–chorological identity. By the late 1970s, however, nature–society or human–environment scholarship had returned to geography in a significant way, building from interests that helped to give rise to cultural ecology and risk-hazard studies as well as several other interdisciplinary subfields. Entering the twenty-first century, nature–society geography has grown significantly in the number of its practitioners and research influence, exploring the range of perspectives crossing from the sciences to the humanities. This rebirth coincides with renewed academic and public interests in the sustainability of the human–environment condition, whether such interests focused on climate change, ecosystem and biotic integrity, or sustainable development, and raises questions about the role of nature–society interests in the discipline of geography and in the academy at large.

Roots of the Theme in Geography

Ancient geography (Greek *geographia*, meaning writing about the Earth) involved descriptions of the Earth, serving to compile knowledge of places not experienced in everyday life and as found in travel and exploration. All places contain 'nature,' and most are occupied or used in some capacity by our species. It is not surprising, therefore, that 'complete' geographies documented the human and environmental character of places (Glacken, 1967). This substantive interest, in various intellectual forms, remains throughout the history of geography, although it continually struggles with other identities of the discipline in search of its intellectual niche.

German Foundations of Modern Geography

The nineteenth-century organization of knowledge within European academe shaped modern geography at a time in which the discipline was especially vibrant in Germany. The competing views expounded by German scholars during that century remain essential, if contested, elements of the discipline's identity in the twenty-first century.

The Spatial–Chorological Science

Immanuel Kant (1724–1803) laid the foundation for a spatial or chorological identity: geography provided integrative understanding through a focus on the spatial attributes and relations of phenomena as they were arranged in areas, regions, or space (Tatham, 1951: p. 40). This vision was elaborated in three ways by Karl Ritter (1779–1859), Ferdinand von Richthofen (1833–1905), and Alfred Hettner (1859–1941), among others. Geography was the science of areal or regional differentiation (and, hence, the need to bound or demarcate areas to inform understanding), the history and particularities of places (chorography), and distribution studies. These various identities, here collectively labeled chorology, generated a discipline based largely on a shared approach to, more than on a phenomenon of, study. This 'unique' quality, shared with history (Kant's chronological science), marked geography as an exception to the 'systematic' (science) disciplines, given its long-standing insistence that it is or large parts of it are a science.

The Nature–Society Science

Building on the natural histories of Comte de Buffon (1707–88), Johann R. Forster (1729–98), and Johann G.A. Forster (1754–95), an alternative identity of the discipline was championed by other nineteenth-century geographers, if not always explicitly. Alexander von Humboldt (1769–1859) set the stage for this vision of geography through such works as *Cosmos* (*Sketch of a Physical Description of the Universe*), which documented the 'unity in diversity' in nature, or how diverse phenomena, including humans, give rise to ordered, functioning landscapes (akin to the modern use of landscape in ecology). Humboldt was not especially interested in defining geography or articulating approaches within it, but he recognized three divisions of knowledge about nature, one of which he called geognosy (Earth science) and which dealt with the spatial distribution and relationships of phenomena (Dickinson, 1969: p. 24). This recognition along with his research approach, which involved bounding a landscape (thereby creating an areal unit), was consistent with chorology. Humboldt, however, had a phenomenon of study, the landscape, including the human alterations of it.

Unlike Humboldt, Frederich Ratzel (1844–1904) was keenly interested in the identity of geography and championed the view that the discipline should be a systematic science, complete with a phenomenon of study, the nature–society relationship (Dickinson, 1969: p. 65, Tatham, 1951: p. 64). In *Anthropogeography: Introduction to the Application of Geography to History* (1882), Ratzel advanced a science of reciprocal relationships between nature and society in which the distribution of cultures (his students employed 'cultural area' before the term emerged in anthropology) was predicated on a Darwinian-influenced notion of the capacity of different societies to adapt to their physical surroundings, also referred to as landscape. Ratzel focused on human adjustments to and manipulation of natural resources and environmental conditions, although subsequently others would reverse this orientation.

Other German geographers of the time followed partially, if not fully, the Ratzelian view. Otto Schlüter (1872–1952) found fault with the environmental orientation in the emerging nature–society thought (Dickinson, 1969: p. 126), but he saw the scope of geography as understanding the 'visible' landscape (*landscahft*), not as a bounded area but as the nature–society relationships shaping it (Livingstone, 1993: p. 264). Albrecht Penck (1858–1945), among others, found merit in this view, favoring a landscape-led geography akin to Schlüter's view (Dickinson, 1969: p. 110). From this point onward, landscape would hold different meanings to different geographers. Regardless of the specific definition used, the landscape concept invariably implied a collection of organic and inorganic phenomena (or an aggregate phenomenon) whose interrelationships gave rise to some meaningful areal identity on the Earth's surface, be it a coastline, system of cultivation, or urban settlement.

Ties between the Two Views

Chorology and nature society were linked historically. Human–environment proponents did so by focusing on landscapes bounded as areal units. For example, Vidal de la Blache (1845–1918) and, to a lesser extent, his student, Jeans Brunhes (1869–1930), both influenced by Ratzel, organized French geography around nature–society relationships as they were manifested in small areas and recorded in regional monographs. German chorologists, in contrast, invariably called for the examination of the place, area, or region, which axiomatically included human–environment relationships. Ritter's own work, *Erdkunde*, is exemplary, as registered in the subtitle (*Earth Science: in Relation to Nature and the History of Man*) and in the text: "The earth and its inhabitants stand in the closest reciprocal relations, and one cannot be truly presented in all its relationships without the other" (in Tatham, 1951: p. 44).

The Geographic Influence or Geographic Factor

The nature–society position vied for the sole identity of the discipline as it emerged in the United States during the early part of the twentieth century. There, geography's birth was influenced by William Morris Davis (1850–1934), a geomorphologist of no small repute, who created a geography program within geology at Harvard and was the force behind the creation of the Association of American Geographers (1904), serving as its president three times. He increasingly came to view geography as the study of inorganic (nature) controls over the organic (society). In his 1905 Presidential Address to the Association of American Geographers, he not only made the case for the study of this relationship as the defining element of geography, but also disposed of the chorological position in two brief sentences.

> There is so little support for this narrow view of the subject [location] to be found in modern geographical books that it need not be further considered [and] … if geography were only the science of distribution, that is, the regional aspect of other subjects, it would be hardly worth while to maintain the study of geography apart from that of the subject whose regional aspect it considers. (in Johnson, 1954: p. 9)

The strength of Germanic influences on Davis is unclear. He referenced and critiqued German geographers, but his views of geography may have been more strongly affected by the Darwinian and neo-Lamarckian fervor of the time and the search for organic analogy (Livingstone, 1993: pp. 203–205). His position was shared, even anticipated, by a number of highly productive geographers who were well versed with the German traditions, none more important than Ellen Churchill Semple (1863–1932) of Clark University and the University of Chicago. Mentored by Ratzel, she championed the theme of *geographic influence* or the *geographic factor* as articulated in *Influences of Geographic Environment* (1911). She was joined in this vision in the United States by such luminaries as Ellsworth Huntington (1876–1947) at Yale, a prolific author who sought to understand the history of regions through the lens of climate (e.g., *Civilization and Climate*, 1915), and abroad by Griffith Taylor (1880–1963), among others, who attempted to explicate the controlling constraints of environment on human occupation of different

lands (e.g., *Australia, A Study of Warm Environments and Their Effect on British Settlement*, 1941).

Taken as a whole, this group of geographers paid little, if any, attention to alternative or linked views of the identity of the discipline. Enhanced by the elite private institutions in which Davis, Huntington, and Semple were positioned, their moment of power penetrated the very foundations of American education. The Conference on Geography in 1892 adopted the Davisian identity of geography, which in turn was accepted by the Committee of 10 of the National Educational Association (1892), whose role it was to define precollege educational programs and college entrance requirements (James and Martin, 1981: p. 290). The discipline was thus defined as the study of inorganic controls over the organic, or the geographic influence factor, and its K–12 textbooks were organized accordingly.

The ascendancy of this human–environment identity was short lived. The geographic factor narrowed the human–environment relationship to the search for the causal power of nature over society – an extreme form of human adjustment or paucity thereof to nature. Alternative views of the relationship, such as the possibilism of the French geographers or by the American, George Perkins Marsh (*Man and Nature or, Physical Geography as Modified by Human Action*, 1864) on the environmental impacts of human activity were unheeded. The geographic factor became *environmental determinism*, which, in its lay applications, not only led to unfounded simplifications about nature's influence on humankind but also were drawn on as support for social and political views proposing the superiority of northern environments and peoples over all others.

The extremes of determinism devastated geography and seriously wounded the nature–society position. The stigma attached to the geographic factor was so strong that by the 1940s human–environment interests were relegated to the margins of the discipline and were even dismissed by some historians of the field (Livingstone, 1992: p. 6). For much of the middle of the twentieth century, the central intellectual debate in US geography centered on the kind of spatial–chorological science geography should be: a place-based and areal differentiation science (Hartshorne, 1961), or a science of spatial relationships (National Research Council [U.S.] 1965). The former focus was open to human–environment relationships (above 2.3), whereas the latter focus was indifferent to them. By the 1960s, the latter focus gained intellectual hegemony among the leadership of UK and US geography, remaining in power for two decades before it succumbed to challenges that reinserted the role of the contingencies of place and regions (*see* Geography: Overview). These challenges notwithstanding, geography's identity remains linked primarily to the 'why of where' question, and much of human geography dismisses nature–society studies or sees them as falling outside the 'human' domain of the discipline's study (*see* Geography: Overview).

Alternatives to the Geographic Influence

If the 'battle' for the identity of geography were lost by proponents of the geographic influence, strong geographic interest in nature–society themes persisted. At least two alternatives to environmental determinism arose before its collapse, each providing the foundation for contemporary nature–society studies in geography. Carl O. Sauer (1889–1975) sketched his alternative vision in his 1925 essay, 'The Morphology of the Landscape.' Inspired by German scholarship, Sauer sought to bring the discipline back to the landscape (an area of interrelated phenomenon), whose qualities as a whole (form, structure, and function) constituted a reality not captured by the constituent parts separately (in Leighly, 1963: p. 321). The landscape was a nature–society relationship, and the morphology was a method of synthesis, a "massing and ordering phenomena as forms … integrated into structures" (1963: p. 326). He argued that there was "a strictly geographic way" of entertaining these structures, and it was through the material imprints that society made upon area (1963: p. 326). Thus, Sauer's nature–society relationship was the human impact on the 'physical whole' (i.e., nature and the built environment) which created the landscape.

Whatever Sauer's views on the ties between nature–society and spatial–chorological geography, he and the research prodigy of the 'Berkeley school' produced a large number of works focused on the human use of and impact on the Earth, including Sauer's own role in *Man's Role in Changing the Face of the Earth* (Thomas, 1956). Exemplary was Philip Wagner's *The Human Use of Earth: An Examination of the Interactions Between Man and His Physical Environment* (1960), which provides only passing references to alternative identities of geography before exploring a discipline devoted to the human-made landscape. Such nature–society interests were taken up outside the United States under the rubric of cultural or historical geography (e.g., Darby, 1940; Williams, 1974).

Harlan Barrows (1877–1960) laid the foundation for the other nature–society theme in his 1922 Presidential Address to Association of American Geographers. He proposed geography as the study of human ecology or the adjustment that humankind makes to the occupied environment (Barrows, 1923). Here was an explicit statement that nature–society relationships were the foundation of the discipline, and that this foundation need not be deterministic. Barrows' call had no immediate discernible impacts, but some 30 years later his former student, Gilbert F. White (1911–2006), returned to Chicago to pioneer risk-hazards geography. White was captured by the service role that academic research could provide for solving social problems and was not particularly responsive to academic debates about the discipline's definition. Nevertheless, the defining quality of his interests and the geographic work he advanced was a form of human–environment relationship, specifically human adjustments to environmental perturbations, articulated in Ian Burton, Robert W. Kates, and Gilbert F. White, *The Environment as Hazard* (1978) (*see* Gilbert F. White).

Thus, geography entered the 1960s with two strong research communities addressing nature–society themes: human use and impact on nature, and human responses and adjustment to environmental change and perturbations. Arguably, a large number, perhaps even a majority, of these practitioners were drawn to geography primarily out of its human–environment traditions and the paucity of them in most other disciplines (Kates, 1987). For the most part, however, the nature–society

practitioners were content to do their work and made minimal attempts to challenge the identity of geography as a chorological science.

Contemporary Nature–Society Subfields in Geography

By the 1970s, nature–society research in geography was prospering, but under newly emerging or expanding subfields to which geographic organizations were partitioning themselves. These subfields initially were stimulated by systems and ecosystems approaches combined with theory-driven, behavioral analysis (Harden, 2012; Turner and Robbins, 2008; Zimmerer, 2010). The *environmental hazard* subfield was in full swing, having taken on virtually every type of environmental perturbation to human occupation and use of the environment, from drought to volcanic eruptions. Research began to expand into non-Western world contexts, and a search for generic templates of human responses to such perturbations was under way. Hazards research spawned risk assessment and technological hazards and, more recently, informed by critique of it, vulnerability assessments (e.g., *see* Environmental Risk and Hazards; Environmental Vulnerability). A complementary subfield, *environmental perception and behavioral geography*, moved toward issues of the bounding the human–environment experience and its meaning for hazard adjustments and landed resource use more generally.

The *cultural ecology* subfield in geography arose from the cross-fertilization of practitioners whose legacy may be seen in the substantive interests in the landscape but with the goals and aims articulated by Harold C. Brookfield in Australia on household and community behavior and Karl W. Butzer in the United States (via Canada and Germany) on systems (e.g., *see* Ecology, Cultural Geography). Increasingly, the work of cultural ecologists in geography and anthropology blurred as attested in work by, for example, Robert McC. Netting (1934–95), Emilio Moran, and William M. Denevan.

By the 1980s, both environmental hazard and cultural ecology were strongly affected by Marxist and critical approaches, and in the 1990s by social theory broadly conceived, with various impacts on the two subfields (Turner and Robbins, 2008). *Political ecology* emerged (Blaikie and Brookfield, 1987), directing attention to local nature–society conditions as a product of social differentiation (e.g., power, wealth, class, gender) and of linkages to distant decisions and processes. Many proponents were interested in both the drivers of and responses to environmental change and linking the two, dissolving some of the distinction that existed between cultural ecology and risk-hazard research. In addition, political ecology research in its various renditions broke down the partitions in nature–society research between 'pure' and 'applied' work, as did its risk-hazard counterpart, (e.g., *see* Feminist Political Ecology; Ecology, Political).

Contemporary agriculture and rural land use and *water resource* subfields emerged to tackle current nature–society problems, primarily in the developed world (e.g., *see* Water Resources). Both experienced similar intellectual influences that gave rise to political ecology. The *human dimensions of global environmental change* emerged in the 1990s as a hybrid of the other subfields – one that appealed to global environmental change and sustainability themes (Turner et al., 1990). More recently, subfields such as land change science have emerged (Turner et al., 2007).

Mirroring the cacophony of perspectives in play in the natural and human sciences and humanities, nature–society practitioners have explored a wide range of frameworks for understanding (Turner and Robbins, 2008): postpositivistic science, to critical and social theory, to the search for meaning in the symbolic landscape (e.g., see Cultural Landscape in Environmental Studies). An appreciation for such diversity and the ability to move among the various frameworks have long been hallmarks of geographic work. After a period in which different communities sought to advance competing frameworks or perspectives for understanding human–environment systems, much nature–society geography has now refocused, informing concepts and problems in the way that its intellectual ancestors might have imagined, if not the kinds of questions and issues addressed. That work is strongly empirical and increasingly quantitative, but it is responsive to qualitative data and methods and focused on the material relationships between humankind and the physical environment, whether it is human-induced land degradation or equity in opportunities to respond to it. Explanatory models seek to treat more equitably than before actor-based and structural-based approaches and to consider the roles of context, idiosyncratic agents, and historicity (path dependence). Additionally, the environment as an active force within the relationship has been reinstated, albeit examined through such themes as nonstationarity, complexity, system thresholds, and coadaptation. The former barriers between human use–impact and human response–adjustment studies have all but disappeared. In addition, nature–society practitioners may be poised to tear down the blinders of established explanatory perspectives, seizing the moment of 'hybrid explanatory vigor' to formulate improved understanding of the human–environment relationship (Turner and Robbins, 2008).

Geography as Nature–Society Science?

Do these trends suggest that geography is poised to redirect its identity to nature–society relationships? Probably not, although Stoddart (1987) came very close to making such an appeal, and Peet (1998) argues for a variant of such a move. To adopt this identity, retain the spatial–chorological one, or, better, to unite the two, requires resolution of several intellectual and pragmatic considerations confronting a restructuring academy.

What Other Disciplines Concede

May (1970: p. 186) argued that no other discipline claims the nature–society relationship and most concede the subject to geography. As well, three of five texts granted the status as paradigmatic shifts in geographic practice by Harvey and Holly (1981) belong to Ratzel, Vidal de la Blache, and Sauer, all of which can be interpreted as nature–society visions as much as, if not more so, than spatial–chorological visions. May's observations, however, ran counter to those by Hartshorne

(1961: pp. 124–125), who argued that the other human sciences 'must' address nature–society relationships "to understand the different social conditions of different lands and places." Save for anthropology and resource and environmental economics, however, the human sciences have largely ignored these relationships (e.g., Goldman and Schurman, 2000), at least until the recent rise of public and government concerns about them. If Hartshorne erred, was May correct? Human–environment geographers have played a major role in the creation of interdisciplinary research institutes and international organizations and programs dedicated to climate and global change and sustainability; their impact in these endeavors has been far out of proportion to their numbers in the academy at large. The very interdisciplinary qualities of these activities, however, raises questions about how much the 'core' of nature–society studies is embedded in any one discipline.

Unified Understanding

Does a nature–society identity provide unity to geographic understanding? Again, Hartshorne was adamant that it did not, drawing on Hettner's argument that the "logical unitary structure of geography [is] … destroyed" by the dualism inherent in the nature–society position, whereas "if the concept of nature and man proceeds from the chorological viewpoint, it is homogeneous in all important points" (in Hartshorne, 1961: p. 123).

Hartshorne provided little support for Hettner's claim, as if the weight of history were sufficient to prove the point. But this history is not necessarily borne out in contemporary geography in which the divisions between human and physical geography have grown, and frameworks for unification drawn from the spatial–chorological position do not necessarily speak to nature–society interests (but see Hanson, 1999). Contemporary attempts to articulate the unity of geography through the nature–society position are few. In principle, however, there is no reason why a focus on the relationship should create the dualism noted, and various reasons why it should lead to more integrative or unified views of humankind and the physical world. Indeed, this last rationale is precisely that used to justify the investment in the many interdisciplinary sustainability programs, centers, and institutes emerging worldwide.

Nature–Society and Geography as Practiced

What passes for accepted practice in large measure shapes the identity of disciplines. Pattison (1964) identified four traditions or domains of research practiced in geography in the mid-1990s, which are still appropriate in the twenty-first century, allowing for name changes: mapping sciences, physical or environmental geography, space–society studies, and human–environment studies. Privileging the nature–society domain as the identity of geography encounters problems rationalizing the existence of the other three domains as independent subjects of study. Barrows and other proponents of the nature–society position of yesteryear noted this predicament and dealt with it by dismissing the other domains, save as they were needed to understand nature–society relationships. Peet (1998: pp. 1–3) linked nature–society and spatial–chorological geography by insisting that nature–society relationships could be understood only in real-world relational space. In this interpretation, space is the 'veneer' of nature–society relationships, such that to examine spatial relationships axiomatically involves nature–society relationships. Alternatively, Sack (1997) explored a way in which nature and society can be reconciled without the reductionism employed by the natural and some social science approaches. This treatment is accomplished by focusing on places because, according to Sack, place (a bounded area) is the instrument that humankind uses to combine nature and society, or the forces and process of each. Without place, understanding privileges nature, society, or a particular explanatory form drawn from one or the other. With place, however, a 'less privileged' understanding is possible. Should this understanding be forthcoming, perhaps the long-standing tensions between the two visions of geography might dissipate. It is noteworthy that these attempts to link the two identities mimic those that have been made in past, but with different language.

Nature–Society Geography Entering the Twenty-First Century

What is and ought to be humankind's relationship with nature? This question remains one of the most fundamental and vexing confronting our species (Kates, 2001). Geography has long claimed much of this question as its own, and perhaps has devoted more attention to its various elements than any other formal discipline of study. This tradition, strongly muted during the mid-twentieth century, perhaps reflecting geography's self-image vis-à-vis environmental determinism, revived significantly by the 1990s in concert with the rise internationally of multidisciplinary research foci variously labeled global environmental change and sustainability, and the emergence of new perspectives and methods for nature–society studies. Rebuilding on its past, human–environment geography may offer the largest number and array of researchers from any discipline experienced in bridging the natural and human science components of these concerns.

Reinvigorated in this way, the contributions of the nature–society domain of geography appear to stand on equal footing with those of all other geographic domains of study (National Research Council, 1997, 2010).

This circumstance, however, has come at a moment when the academy at large has discovered integrated human–environment studies, referred to in the literature as social–ecological systems, human and natural systems, and human–environment system – all forms of complex adaptive systems. International programs of research, such as the Global Land Project, DIVERSITAS, the Program on Ecosystem Change and Society, and the Resilience Alliance, are under way to advance human–environment research, while the International Council of Science reconfigures the global change research program accordingly under the rubric of Future Earth. In tandem, institutions of higher learning internationally are creating degree programs to foster this integrated understanding and which tend to maintain the label 'sustainability' in their titles. A series of new journals

(e.g., *International Journal of Sustainability*, *Sustainability: The Journal of Record*) and sections of established ones (e.g., *Proceedings of the National Academy of Sciences, USA*) serve the science and scholarship emanating from sustainability science and related studies. All of these and other efforts may be creating a new, integrated human–environment science independent of established disciplines, such as sustainability science, which take the human–environment system as its phenomenon of study. The outcome of these dynamics will determine the role of nature–society geography within and without the discipline.

See also: Cultural Geography; Environmental Determinism; Geographic Thought, History of; Geography: Overview; Sauer, Carl Ortwin (1889–1975); Political Geography.

Bibliography

Barrows, H.H., 1923. Geography as human ecology. Annals of the Association of American Geographers 13, 1–14.

Blaikie, P., Brookfield, H., 1987. Land Degradation and Society. Methuen, London.

Darby, H.C., 1940. The Draining of the Fens. Cambridge University Press, Cambridge, UK.

Dickinson, R.E., 1969. The Makers of Modern Geography. Praeger, New York.

Glacken, C.J., 1967. Traces on the Rhodian Shore: Nature and Culture in Western Thought from Ancient Times to the End of the Eighteenth Century. University of California Press, Berkeley, CA.

Goldman, M., Schurman, R.A., 2000. Closing the "great divide": new social theory on society and nature. Annual Reviews in Sociology 26, 563–584.

Hanson, S., 1999. Isms and schisms: healing the rift between the nature-society and space-society traditions in human geography. Annals of the Association of American Geographers 89, 133–143.

Harden, C.P., 2012. Framing and reframing questions of human-environment interactions. Annals of the Association of American Geographers 102, 737–747.

Hartshorne, R., 1961 [1939]. The Nature of Geography: A Critical Survey of Current Thought in the Light of the Past. Association of American Geographers, Lancaster, PA.

Harvey, M.E., Holly, B.P. (Eds.), 1981. Themes in Geographic Thought. Croom Helm, London.

James, P.E., Martin, G.J., 1981. All Possible Worlds: A History of Geographical Ideas, second ed. Wiley, New York.

Johnson, D.W. (Ed.), 1954. Geographical Essays by William Morris Davis. Dover, New York.

Kates, R.W., 1987. The human environment: the road not taken, the road still beckoning. Annals of the Association of American Geographers 77, 525–534.

Kates, R.W., 2001. Queries on the human use of the earth. Annual Review of Energy and Environment 26, 1–26.

Leighly, J. (Ed.), 1963. Land and Life: A Selection from the Writings of Carl Ortwin Sauer. University of California Press, Berkeley, CA.

Livingstone, D.N., 1992. The Geographical Tradition: Episodes in the History of a Contested Enterprise. Blackwell, Oxford, UK.

May, J.A., 1970. Kant's Concept of Geography and Its Relation to Recent Geographical Thought. University of Toronto Press, Toronto, Canada.

National Research Council, 1997. Rediscovering Geography: New Relevance for Science and Society. National Academy Press, Washington, DC.

National Research Council, 2010. Understanding the Changing Planet: Strategic Directions for the Geographical Sciences. National Academy Press, Washington, DC.

National Research Council (U.S.) Committee on Geography, 1965. The Science Of—Geography: Report. National Academy of Sciences–National Research Council, Washington, DC.

Pattison, W.D., 1964. The four traditions of geography. Journal of Geography 63, 211–216.

Peet, R., 1998. Modern Geographical Thought. Blackwell, Oxford, UK.

Sack, R.D., 1997. Homo Geographicus: a Framework for Action, Awareness, and Moral Concern. Johns Hopkins University Press, Baltimore, MD.

Stoddart, D.R., 1987. To claim the high ground: geography for the end of the century. Transactions of the Institute of British Geographers 12, 327–336.

Tatham, G., 1951. Geography in the nineteenth century. In: Taylor, G. (Ed.), Geography in the Twentieth Century. Methuen, London, pp. 28–69.

Thomas, W.M. (Ed.), 1956. Man's Role in Changing the Face of the Earth. University of Chicago Press, Chicago.

Turner II, B.L., Clark, W.C., Kates, R.W., Richards, J.F., Mathews, J.T., Meyer, W.B. (Eds.), 1990. The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere Over the Past 300 Years. Cambridge University Press, Cambridge, UK.

Turner II, B.L., Lambin, E., Reenberg, A., 2007. The emergence of land change science for global environmental change and sustainability. Proceedings of the National Academy of Sciences, of the United States of America 104, 20666–20671.

Turner II, B.L., Robbins, P., 2008. Land change science and political ecology: similarities, differences, and implications for sustainability science. Annual Reviews in Environment and Resources 33, 6.1–6.22.

Williams, M., 1974. The Making of the South Australian Landscape: A Study in the Historical Geography of Australia. Academic Press, London.

Zimmerer, K.S., 2010. Retrospective on nature-society geography: tracing trajectories (1911–2010). Annals of the Association of American Geographers 100, 1076–1094.

Nature-Nurture Controversy, History of

Christopher R Badcock, London School of Economics, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The nature/nurture distinction originated with Francis Galton, who was also the first to realize that twin studies could provide a crucial test of the contributions of each. Both behaviorists and social scientists made extreme claims for nurture during the twentieth century. In the Soviet Union Lysenkoism enshrined nurture over nature as a dogma of Marxist biology while in the West scientists who failed to subscribe to a less dogmatic equivalent were subject to campaigns of casuistic criticism and character assassination. Findings from experiments of nature such as feral children and attempted gender-assignment contrary to chromosomal sex, along with recent evidence from autism, Turner's syndrome, and the link between monoamine oxidase A and antisocial behavior suggest that Galton was substantially correct in thinking nature to be the dominant factor. But the recent discovery of epigenetic factors, which for example explain genetic differences between identical twins, suggests caution in accepting the traditional terms of the debate at face value.

Origins of Nature/Nurture: Francis Galton

Sir Francis Galton (1822–1911) introduced *nature* and *nurture* as understood with their modern connotations in his book of 1874, *English Men of Science: Their Nature and Nurture*. According to Galton,

> The phrase "nature and nurture"… separates under two distinct heads the innumerable elements of which personality is composed. Nature is all that a man brings with himself into the world; nurture is every influence from without that affects him after birth.

Galton was careful to add that "Neither of the terms implies any theory: natural gifts may or may not be hereditary; nurture does not especially consist of food, clothing, education or tradition, but it includes all these and similar influences whether known or unknown" (Galton, 1874: p. 12).

He went on: "It is needless to insist that neither is self-sufficient; the highest natural endowments may be starved by defective nurture, while carefulness of nurture can overcome the evil tendencies of an intrinsically bad physique, weak brain, or brutal disposition" (Galton, 1874: pp. 12–13). Galton concludes that "In the competition between nature and nurture, when the differences in either case do not exceed those which distinguish individuals of the same race living in the same country under no very exceptional conditions, nature certainly proves the stronger of the two" (Galton, 1874: p. 16).

Galton noticed that "There are twins of the same sex so alike in body and mind that not even their own mothers can distinguish them" (Galton, 1874: p. 13), and made a distinction between what we would now call identical, or monozygotic (MZ) twins (those which developed from a single fertilized egg cell) and fraternal, or dizygotic (DZ) twins (which develop from different eggs). Although Galton remarked that "It occurred to me that the after-history of those twins who had been closely alike as children, and were afterward parted… would supply much of what was wanted…" (Galton, 1869: pp. 294–295) he did not propose that the resemblance of MZ twins be compared to the resemblance of DZ pairs to assess genetic influence. Only much later in 1924, did Curtis Merriman in the USA and Hermann Siemens in Germany realize that MZ twins separated at birth provided what appeared to be an acid test of nature versus nurture because it was presumed that their common genetics held nature constant while their different environments varied the factor of nurture (Rende et al., 1990).

Making a Case for Nurture: Behaviorism

Galton was candid enough to admit "that my evidence may seem to prove too much, and be discredited on that account, as it appears contrary to all experience that nurture should go for so little" (Galton, 1874: p. 16).

The behaviorist John Broadus Watson (1878–1958) famously claimed that: "we no longer believe in inherited capacities, talent, temperament, mental constitution, and characteristics. Give me a dozen healthy infants, well-formed, and my own specified world to bring them up in and I'll guarantee to take any one at random and train him to become any type of specialist I might select – doctor, lawyer, merchant-chief and yes, even beggar-man and thief, regardless of the talents, penchants, tendencies, abilities, vocations and race of his ancestors." He admitted that "I am going beyond my facts and I admit it," but added, "so have the advocates of the contrary and they have been doing it for many thousands of years" (Watson, 1930: p. 82).

Experiments by John Garcia showed that rats could be conditioned to associate the size of a food item with an electric shock, but that rats cannot be conditioned to associate a shock with the taste of food. Again, when paired with X-rays causing illness 1 h later, rats learn to associate the taste of food with illness, but cannot learn to associate the size of food items with illness caused in the same way. Clearly, insects with stings and poisoned baits that rats naturally encounter in their habitat might account for why the animals associate size with shocks, and taste with poisoning. But taste would seldom if ever be naturally associated with an instantaneous insult, and the size of a food item is no clue to its toxicity. In a similar way, pigeons

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03136-6

cannot be conditioned to associate odor or sound with food, but will readily associate its appearance. And even though pigeons quickly learn to peck an illuminated key before food is delivered because pecking is a natural food-finding activity, attempts to condition them not to peck the key by making pecking it prevent the food reward have no effect. On the contrary, once established, pecking persists, even if the pigeon starves to death as a result (Garcia, 1981, 1990).

In short, there are definite limits to conditioning, even in the case of classic laboratory animals such as rats and pigeons, and it would be astonishing if the same were not even more true of human beings.

The Cultural Claims for Nurture

Parallel claims to those of behaviorists for the overriding importance of nurture over nature were made by numerous social science writers, notably Franz Boas (1858–1942) and his colleague, Alfred Louis Kroeber (1876–1960). But these claims remained mere assertions until Boas's student, Margaret Mead (1910–78) wrote *Coming of Age in Samoa*, which purported to provide ethnographic proof that nurture was indeed the dominant factor – at least in Samoa (Mead, 1928).

However, the anthropologist, Derek Freeman, followed in her footsteps in the 1960s, and published a series of books suggesting otherwise. For example, Mead claimed Samoans were "one of the most amiable, least contentious, and most peaceful peoples in the world" but Freeman reports that serious assault in mid-1960s Western Samoa was 67% higher than in the USA, 494% higher than in Australia, and 847% higher than in New Zealand, while common assault was 500% that of the USA. In Samoa according to Mead "love between the sexes is a light and pleasant dance," male sexuality "is never defined as aggressiveness that must be curbed," adding that "the idea of forceful rape… is completely foreign to the Samoan mind." But Freeman reports that rape convictions in Samoa of the 1960s were twice the level of those in the USA and 20 times those of the UK. At the time Mead was in Samoa, rape was the third-most common criminal offense. Whereas Mead claimed that young girls had "as many years of casual love-making as possible" before marriage, there was in fact a cult of virginity; and contrary to her portrayal of it, the pattern of adolescent crime was much the same as anywhere else. Eventually, Freeman obtained a confession from one of Mead's informants and concluded that Margaret Mead's credulity for nurture over nature had led to her being hoaxed by her Samoan informants (Freeman, 1983, 1998, 2000).

According to Judith Rich Harris, the usual view of cultural transmission – that the culture is passed down from the parents to the child – is inadequate and misleading. An example is language in the USA. The children of immigrants who speak English with a heavy accent end up speaking English with no foreign accent at all if they grow up in a neighborhood of native-born Americans. Harris thinks that other aspects of culture are transmitted in the same way as language. In developed societies the parents start the process at home, but whether children retain what they learned at home will depend on what they find outside. Indeed, she claims that there are many societies where the parents hardly talk to their babies at all, and the babies do not learn language until they graduate from their mother's arms into the local playgroup. They learn both the language and how to behave from the older children in such groups, making peers the most important factor in nurture (Harris, 1999).

Politics Against Nature: Lysenkoism in the USSR and Parallels in the West

Mead's close contemporary, Trofim Denisovitch Lysenko (1898–1976) was promoted through positive discrimination in favor of peasants by Marxist revolutionaries in the USSR, and ultimately became Stalin's biology czar. Lysenko took the nurture-not-nature ideology of official Marxism to bizarre and extreme lengths, claiming that crops could be 'educated' to grow in winter (so-called *vernalization*), and that such learning was heritable. Lysenko denounced Mendelian genetics as a 'capitalist' and 'clerical conspiracy,' and denied the existence of genes, hormones, and viruses. Lysenko also unconditionally accepted the inheritance of acquired characteristics, denounced natural selection as 'Darwin's greatest mistake,' claimed species could transform into one another without intermediate steps, and believed that fish in puddles arose from spontaneous generation (Medvedev, 1969)!

Astonishingly, Lysenkoism became the official genetic doctrine of the USSR until the mid-1970s, and after 1948, thousands of scientists were purged, tortured and murdered in detention, or died of starvation and exposure exiled in Siberia. The teaching of Mendelian genetics and criticism of Lysenko's environmentalist inheritance became a crime, and Soviet agriculture was forced to adopt Lysenko's methods, with disastrous consequences that contributed to widespread famines and food shortages (Gershenson, 1990; Medvedev, 1969).

But such proscription of genetics was not limited to the USSR: in a less official form it also affected the USA, as the extract from the following letter, signed by 50 scientists and published in *American Psychologist* in July 1972, eloquently attests. After citing "periods when scientific research or teaching was censured, punished, or suppressed for nonscientific reasons, usually for seeming to contradict some religious or political belief," and naming Galileo, Darwin, Einstein, and 'Mendelian biologists, in Stalin's Russia,' as cases, the letter continues to point out that

> Today, a similar suppression, censure, punishment, and defamation are being applied against scientists who emphasize the role of heredity in human behavior. Published positions are often misquoted and misrepresented; emotional appeals replace scientific reasoning; arguments are directed against the man rather than against the evidence… And a large number of scientists, who have studied the evidence and are persuaded of the great role played by heredity in human behavior, are silent… it is virtually heresy to express a hereditarian view, or to recommend further study of the biological bases of behavior. A kind of orthodox environmentalism dominates the liberal academy, and strongly inhibits teachers, researchers and scholars from turning to biological explanation or efforts.
>
> Page (1972)

Examples of the abuses mentioned in this quotation are character assassinations of leading researchers who continued the start made by Galton on testing intelligence and its links to heredity or to physical factors such as brain-size or race. In a press campaign reminiscent of those whipped up by Lysenko, the leading British researcher, Sir Cyril Burt (1883–1971) was accused of wholesale scientific fraud after a biographer had revealed some of complexities of Burt's character and sometimes unconventional working methods (Hearnshaw, 1979). But in 1989 and 1991, the psychologist Robert Joynson (Joynson, 1989) and the sociologist Ronald Fletcher (Fletcher, 1991) independently showed that the attack on Burt had no factual basis. On the contrary, Burt's claim that identical twins reared apart have IQs correlated at 0.77 has been vindicated by the Minnesota Twin Study, which found 0.78 (Rushton, 1994).

According to Stephen J. Gould in his influential book *The Mismeasure of Man*, Henry H. Goddard's (1866–1957) pioneering work on IQ in the USA was aimed at "preventing the immigration and propagation of morons." But Goddard's own conclusion was the opposite: he believed that immigrants' low IQ scores would rise in better social conditions in the USA and asserted that "we may be confident that their children will be of average intelligence and if rightly brought up will be good citizens…" Again, contrary to Gould's portrayal of the facts, Lewis Terman (1877–1956) reported high IQs of Jewish and oriental immigrants, and IQ testers such as Goddard and Terman did not advocate restrictions on US immigration like those of the 1924 act. In fact, Congress took little or no notice of IQ findings in framing the legislation, none of the major testers testified, and their publications were not cited in the legislative record. And although Gould claimed that Samuel G. Morton (1799–1851) "unconsciously" doctored cranial capacity measurements to endorse Caucasian "racial superiority," remeasurement by John S. Michael in 1988 found few errors – and those not in the direction asserted by Gould. Michael concluded that Morton's research "was conducted with integrity" and that Gould was "mistaken" (Rushton, 1997).

Test Cases: Sex Determination

The reports of Mead and others were often interpreted as evidence for nurture against nature, and another example might be the claims of John Money (1921–2006), a psychologist at Johns Hopkins University, who held that sex was socially determined in childhood and could be reassigned up to age three. He based these claims principally on the case of 'Brenda' Reimer, one of a normal pair of twin males who was castrated following a botched circumcision and reassigned female by Money in the 1960s – the first ever attempt to do so with an otherwise normal male. Indeed, the case became paradigmatic as an apparent proof of the power of nurture over nature until later revelations revealed an altogether different story of a conflicted child who refused to wear dresses from the start, rejected girls' toys, was described as "boyish" by teachers and who changed his official name to Bruce. Bruce Reimer revealed the coercive side of the nurture dogma when he described himself as a David versus the Goliath of the medical profession and society and eventually married, adopted children and worked in an all-male occupation until he committed suicide in 2004 following the break up of his marriage and suicide of his twin 2 years earlier (Colapinto, 2001).

In the medical literature as a whole "there is no known case where a 46 chromosome, XY male, unequivocally so at birth, has ever easily and fully accepted an imposed life as an androphilic female regardless of physical and medical intervention" (Diamond and Sigmundson, 2000). On the contrary, a study of 18 cases of pseudohermaphroditism due to 5-α-reductase deficit in the Dominican Republic raised as girls who underwent normal male puberty found that 15 made a full sex-role change and were living with women at the time of the study (Imperato-McGinley et al., 1979). Similar findings were reported regarding the same pseudohermaphroditic syndrome among the Simbari Anga of Papua New Guinea who had a radically different culture from that of the Dominican Republic: after male initiation rites prior to puberty the two sexes are kept rigorously separate, and ritualized oral sex occurs between men from puberty until premarital age (Imperato-McGinley et al., 1991; LeVay, 1993: p. 134).

Experiments of Nature: Feral Children

A natural nature/nurture experiment is presented by feral children who provide potential answers to the question of how a child would develop if not nurtured at all, or nurtured by animals, and a surprisingly large number of cases are known. A classic example is The Wild Boy of Aveyron, who was discovered aged 12, living in woods in the late 1700s and found to be mute – but not deaf. He was taught some sign language by a local doctor who was a follower of Rousseau and a believer in the power of nurture over nature, but never spoke or developed social skills or attachments despite the best efforts of his mentor (Losure, 2013).

A modern equivalent is provided by the case of Genie, who was kept chained to a stool in a darkened room in a suburb of Los Angeles for years. Thought to be six on discovery, she was found to be a 49 lb, 3′8″ teenager who was permanently bent over, unable to walk, incontinent, and unable to focus on distant objects. Genie was fostered with psychologists who were enthusiastic about her prospects of rehabilitation, but she never acquired any real language ability, despite their best efforts (Rymer, 1993). However, better results were obtained with Justin, a boy who was raised as a dog. Justin was left with his grandmother aged 2 months until she died 11 months later, and was then raised by her boyfriend along with his dogs in a kennel. Discovered at age five unable to walk or talk, he was fostered and made a remarkably rapid recovery, doing well at school and speaking normally (Perry and Szalavitz, 2006).

Of course, all such cases are anecdotal, but a larger sample was provided by 165 Romanian orphans who were brought to the UK in the 1980s. Most were under a year and had suffered extreme neglect in state orphanages, but by age four, most had recovered. Eleven older children showed signs of autism, but by age six most symptoms had ameliorated and their IQs increased by 20 points compared to an average increase of 7 points in the rest of the group (Frith, 2003: pp. 50–51).

The Lesson of Autism and Turner's Syndrome

Autism is another test case, since this early onset illness evidenced by deficits in social and mentalistic skills (often with severe withdrawal, language delay or complete mutism) was claimed to be caused by nurturing – or the lack of it – by "refrigerator mothers" (Bettelheim, 1967).

Today, however, autism is widely accepted to be a genetic – or at least epigenetic – disorder, and demonstrably runs in families, with high concordance in identical twins (Aitken, 2008). Nevertheless, and despite the fact that all autistic children show some kind of learning difficulties and that many are severely retarded, so-called *autistic savants* demonstrate remarkable abilities, invariably in mechanistic, computer-like skills, most notably and frequently calendar calculation, maths, and music (Treffert, 2000). To the extent that nurture equates to teaching, such cases show that although such children can learn and indeed perform remarkably, they do not do so in ways their parents or teachers wish, but quite independently, and to internal directions of their own: what you might call natural, rather than nurtured, talent (Badcock and Crespi, 2006).

Another crucial experiment of nature that argues the same case for the ability to be nurtured being determined by a child's nature is Turner's syndrome. Turner's affects 2–3 in 5000 births and some Turner's cases tend to be socially insensitive, easily upset, impulsive and difficult to control, with some diagnosed autistic. Turner's is caused by presence of a single X sex chromosome without the second X that would normally be found in a female or the Y that would be found in a male. A study by David Skuse and colleagues found that suffers whose X came from the mother were much more likely to show autism-like social deficits by comparison to those whose X came from the father (Skuse, 1999, 2000, 2002, 2007; Skuse et al., 1997). Given that family backgrounds of the sufferers were not otherwise much different, the finding argues the same conclusion as autism in general: *nurture via nature*, so to speak.

Monoamine Oxidase A and Antisocial Behavior

An even stronger case for the same conclusion comes from the Dunedin Multidisciplinary Health and Development Study which tracked 1037 children born in New Zealand in 1942 and assessed them at two-yearly intervals, with the original sample 96% intact 26 years later. The study notes that childhood maltreatment increases risk of later criminality by approximately 50%, and that the earlier it occurs, the more likely is it to have this effect. Nevertheless, most maltreated children do not become delinquents or criminals, but the reasons why are not known.

Monoamine oxidase A (MAOA) is an enzyme that degrades neurotransmitters such as serotonin, norepinephrine and dopamine. Disruption of the gene in mice increases concentration of these neurotransmitters along with aggression. Evidence that this gene might be implicated in violent behavior in man was found in a Dutch family where eight male members who had been convicted of assault, incestuous rape, arson, and attempted homicide were found to lack MAOA entirely. Of 442 males in the Dunedin sample, 12% who both were maltreated and had low-activity MAOA accounted for 44% of all convictions for violence in the sample; 85% with low MAOA activity showed some antisocial behavior; 45% of those with high-activity MAOA and maltreatment showed antisocial behavior; and 25% of nonmaltreated were antisocial irrespective of MAOA expression. The authors conclude that

> Maltreated children with a genotype conferring high levels of MAOA expression were less likely to develop antisocial problems. These findings may partly explain why not all victims of maltreatment grow up to victimize others, and they provide epidemiological evidence that genotypes can moderate children's sensitivity to environmental insults.
>
> Caspi et al. (2002)

Conclusion: X-Linkage, Epigenetics, and the New View of Nature and Nurture

Duchenne muscular dystrophy is a classic X-linked recessive disorder that affects approximately 1 in 3500 males and results in muscle wastage, paraplegia, and eventual death. But cases are known in which one of a pair of MZ female twins has the disease, and the other has not. The reason is that because males have only a single X-chromosome, they are effectively homozygous for the disorder, while females, with two Xs, will usually be heterozygous, and protected by the non-Duchenne X. On average, half of the cells in a woman's body express one X, and the other half express the other. But because cell lineages inherit their pattern of X-inactivation from very early precursors during embryonic development, and because the splitting that can result in MZ twins occurs at much the same time, it is possible that one of a pair of such twins may predominantly express just one parental X and the other the other one, explaining the remarkable fact that such female twins can be discordant for this classic genetic disorder (Bainbridge, 2003: pp. 145–154).

There are at least 150 genes linked to intelligence on the X-chromosome, and verbal IQ is definitely known to be X-linked (Goodman and Anderton, 1997). A recent study found that compared to male, female identical twins vary more on measures of social behavior and verbal ability thanks to differential X-inactivation. As a consequence of this, the heritability of intelligence has probably been systematically underestimated (Loat et al., 2008). Indeed, the fact that males have only one X-chromosome to rely on – the one they got from their mother – almost certainly explains not simply why so many more males than females have IQs below 70, but also why 20% more males than females have IQs above 140. Greater variability is inevitable if only one X-chromosome is being expressed as compared with the two on which a woman can normally rely (Lehrke, 1997).

As a result of such findings as these, a recent study of DNA expression profiles in MZ and DZ twins pointed out that "molecular mechanisms of heritability may not be limited to DNA sequence differences." Indeed, the authors speculate that, because identical twins reared together or apart are generally quite similar on measures such as brain-imaging, IQ and other psychometrics, "epigenetic differences between identical twins – in other words, those affecting gene expression – are

much more important than environment" (Kaminsky et al., 2009). And of course, if this is true, it may well be true of many other factors that demonstrably run in families, show high twin concordances, but do not obey classical Mendelian rules of inheritance – most importantly perhaps, mental illnesses such as autism or schizophrenia, where Crespi and Badcock recently proposed a general theory in terms of such epigenetic mechanisms (Badcock and Crespi, 2008; Crespi and Badcock, 2008).

At the very least, these recent findings suggest that Galton – and indeed most other writers on the subject since – was wrong to think that effects of nurture could be assumed to explain any and all differences between MZ twins and that nature explained all the similarities. It now seems that variations in the expression of genes like those found in MZ female twins could also explain observed differences quite independent of any environmental or nurturing factor. Furthermore, the emergence of epigenetics and the discovery of a possible regulatory role for 80% of the nonprotein coding DNA which makes up 97% of the human genome suggests that on the nature side too there remains much to be learnt (Ecker et al., 2012). Indeed, it has been argued that the nature/nurture controversy is itself little more than a cultural expression of more profound genetic conflicts originating in the extreme asymmetry of parental investment in human beings (Badcock, 2000: pp. 262–268).

See also: Autism Spectrum Disorders; Epigenetic Inheritance; Epigenetics of Brain Plasticity and Behavior; Galton, Sir Francis (1822–1911); Psychology, History of (Early Period); Psychology, History of (Twentieth Century); Social Epigenetics; Twin Studies and Epigenetics; Twin Studies in Demography.

Bibliography

Aitken, K.J., 2008. Intersubjectivity, affective neuroscience, and the neurobiology of autistic spectrum disorders: a systematic review. Keio Journal of Medicine 57, 15–36.

Badcock, C.R., 2000. Evolutionary Psychology: A Critical Introduction. Polity Press, Cambridge.

Badcock, C.R., Crespi, B., 2006. Imbalanced genomic imprinting in brain development: an evolutionary basis for the etiology of autism. Journal of Evolutionary Biology 19, 1007–1032.

Badcock, C.R., Crespi, B., 2008. Battle of the sexes may set the brain. Nature 454, 1054–1055.

Bainbridge, D., 2003. The X in Sex: How the X Chromosome Controls Our Lives. Harvard University Press, London.

Bettelheim, B., 1967. The Empty Fortress: Infantile Autism and the Birth of the Self. Free Press, New York.

Caspi, A., Mcclay, J., Moffitt, T.E., Mill, J., Martin, J., Craig, I.W., Taylor, A., Poulton, R., 2002. Role of genotype in the cycle of violence in maltreated children. Science 297, 851–854.

Colapinto, J., 2001. As Nature Made Him: The Boy Who Was Raised as a Girl. Harper Perennial, New York and London.

Crespi, B., Badcock, C., 2008. Psychosis and autism as diametrical disorders of the social brain. Behavioral and Brain Sciences 31, 241–320.

Diamond, M., Sigmundson, H.K., 2000. Sex reassignment at birth. In: Ceci, S.J., Williams, W.M. (Eds.), The Nature-Nurture Debate: The Essential Readings. Blackwell, Oxford.

Ecker, J.R., Bickmore, W.A., Barroso, I., Pritchard, J.K., Gilad, Y., Segal, E., 2012. ENCODE explained. Nature 489, 52–55.

Fletcher, R., 1991. Science, Ideology and the Media: The Cyril Burt Scandal. Transaction Publishers, New Brunswick, USA.

Freeman, D., 1983. Margaret Mead and Samoa: The Making and Unmaking of an Anthropological Myth. Harvard University Press, Cambridge, MA.

Freeman, D., 1998. The Fateful Hoaxing of Margaret Mead: A Historical Analysis of Her Samoan Researches. Westview Press, Boulder, CO.

Freeman, D., 2000. Was margaret mead misled or did she mislead on Samoa? Current Anthropology 41, 609–622.

Frith, U., 2003. Autism: Explaining the Enigma. Oxford, Blackwell.

Galton, Sir F., 1869. Hereditary Genius: An Inquiry into its Laws and Consequences. Macmillan, London.

Galton, F., 1874. English Men of Science: Their Nature and Nurture. Macmillan, London.

Garcia, J., 1981. Tilting at the paper mills of academe. American Psychologist 36, 149–158.

Garcia, J., 1990. Learning without memory. Journal of Cognitive Neuroscience 2, 287–305.

Gershenson, S.M., 1990. The grim heritage of Lysenkoism: four personal accounts IV difficult years in Soviet genetics. Quarterly Review of Biology 65, 413–479.

Goodman, J.D.T., Anderton, R.B., 1997. X-linkage, lyonization and a female premium in the verbal IQ results of Orkney school children, 1947–1975. Journal of Biosocial Science 29, 63–72.

Harris, J.R., 1999. The Nurture Assumption: Why Children Turn Out the Way They Do. Free Press, New York.

Hearnshaw, L.S., 1979. Cyril Burt, Psychologist. Hodder and Stoughton, London.

Imperato-McGinley, J., Miller, M., Wilson, J.D., Peterson, R.E., Shackleton, C., Gajdusek, D.C., 1991. A cluster of male pseudohermaphrodites with 5α-reductase deficiency in Papua New Guinea. Clinical Endocrinology 34, 293–298.

Imperato-McGinley, J., Peterson, R., Gautier, T., Sturla, E., 1979. Androgens and the evolution of male-gender identity among male pseudohermaphrodites with a 5α-reductase deficiency. New England Journal of Medicine 300, 1233–1237.

Joynson, R.B., 1989. The Burt Affair. Routledge, New York.

Kaminsky, Z.A., Tang, T., Wang, S.-C., Ptak, C., Oh, G.H.T., Wong, A.H.C., Feldcamp, L.A., Virtanen, C., Halfvarson, J., Tysk, C., McRae, A.F., Visscher, P.M., Montgomery, G.W., Gottesman, I.I., Martin, N.G., Petronis, A., 2009. DNA methylation profiles in monozygotic and dizygotic twins. Nature Genetics 41 (2), 240–245. doi: http://www.nature.com/ng/journal/v41/n2/suppinfo/ng.286_S1.html http://dx.doi.org/10.1038/ng.286.

Lehrke, R., 1997. Sex Linkage of Intelligence: The X-factor. Praeger, Westport, CT and London.

LeVay, S., 1993. The Sexual Brain. The MIT Press, Cambridge, MA.

Loat, C.S., Haworth, C.M.A., Plomin, R., Craig, I.W., 2008. A model incorporating potential skewed X-inactivation in MZ girls suggests that X-linked QTLs exist for several social behaviours including autism spectrum disorder. Annals of Human Genetics. http://dx.doi.org/10.1111/j.1469-1809.2008.00470.x.

Losure, M., 2013. Wild Boy: The Real Life of the Savage of Aveyron. Candlewick Press, Somerville, MA.

Mead, M., 1928. Coming of Age in Samoa: A Psychological Study of Primitive Youth for Western Civilisation. Jonathan Cape, London.

Medvedev, Z.A., 1969. The Rise and Fall of T. D. Lysenko. Columbia University Press, New York.

Page, E.B., 1972. Behavior and heredity. American Psychologist 27, 660–661.

Perry, B., Szalavitz, M., 2006. The Boy Who Was Raised as a Dog: And Other Stories from a Child Psychiatrist's Notebook–What Traumatized Children Can Teach Us about Loss, Love, and Healing. Basic Books, New York.

Rende, R.D., Plomin, R., Vandenberg, S.G., 1990. Who discovered the twin method? Behavior Genetics 20, 277–285.

Rushton, J.P., 1994. Victim of scientific hoax. (Cyril Burt and the genetic IQ controversy). Society 31, 40–45.

Rushton, J.P., 1997. Race, intelligence and the brain: the errors and omissions of the 'revised' edition of S.J. Gould's *The Mismeasure of Man* (1996). Personality and Individual Differences 23, 169–180.

Rymer, R., 1993. Genie: A Scientific Tragedy. Harper Collins, New York.

Skuse, D., 1999. Genomic imprinting of the X chromosome: a novel mechanism for the evolution of sexual dimorphism. Journal of Laboratory and Clinical Medicine 133, 23–32.

Skuse, D., 2 May 2002. Why is autism a male disorder: is the answer in the amygdala? In: Behavioural and Brain Sciences Unit Seminars. The Institute of Child Health.

Skuse, D.H., 2000. Imprinting, the X-chromosome, and the male brain: explaining sex differences in the liability to autism. Pediatric Research 47, 9–16.

Skuse, D.H., 2007. Rethinking the nature of genetic vulnerability to autistic spectrum disorders. Trends in Genetics 23, 387–395.

Skuse, D.H., James, R.S., Bishop, D.V.M., Coppin, B., Dalton, P., Aamodtleeper, G., Bacaresehamilton, M., Creswell, C., Mcgurk, R., Jacobs, P.A., 1997. Evidence from Turner's syndrome of an imprinted X-linked locus affecting cognitive function. Nature 387, 705–708.

Treffert, D.A., 2000. Extraordinary People: Understanding Savant Syndrome. iUniverse.com, Lincoln, NE.

Watson, J., 1930. Behaviorism. University of Chicago Press, Chicago.

Navigation

Ben Finney, University of Hawaii, Honolulu, HI, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 10410–10413, © 2001, Elsevier Ltd.

Abstract

Deep-sea navigation began some 3500 years ago when Stone Age voyagers first began expanding into the Pacific, using naked-eye observations of the stars, winds, ocean swells, and other natural signs to guide their canoes. However, contemporary, science-based navigation developed much later from seafaring traditions practiced around the linked continents of Africa, Europe, and Asia that were initially more oriented toward coastal trade and travel than open ocean sailing. In the fifteenth-century AD Portuguese mariners combined methods from these traditions with new techniques for exploiting global wind patterns to begin sailing beyond long established routes and eventually around the world, which in turn led to the establishment of the first global maritime empires. In following centuries, further application to navigation of astronomy, mathematics, and other disciplines enabled navigators to guide their ships more and more expeditiously and safely to any port in the world, a crucial requirement for the current world system. More recent developments in electronic navigation, now epitomized by the satellite-based Global Positioning System, have brought increasingly greater precision to what long ago began as simple wayfinding. The development of navigation has played a key role in the globalization of the human community, and now provides the means for extending the human reach into space.

'Navigation,' from the Latin *navis*, a ship, and *agere*, to drive, originally referred to the art of sailing a vessel from one point to another, but is now also applied to guiding aircraft and spacecraft. Beginning with early evidence for deep sea voyaging, this article examines the worldwide development of navigation and touches upon the implications for humankind of navigating freely around the globe and into space.

'Navigation,' from the Latin *navis*, a ship, and *agere*, to drive, originally referred to the art of sailing a vessel from one point to another, but is now also applied to guiding aircraft and spacecraft. Beginning with early evidence for deep sea voyaging, this article examines the worldwide development of navigation and touches upon the implications for humankind of navigating freely around the globe and into space.

Early Navigation in the Pacific

Although most accounts of the development of navigation begin with early seafaring in the Mediterranean and the Indian Ocean, recent research points to the Pacific as the region where humans first sailed routinely far out to sea. Around 1500 BC Stone Age seafarers began expanding eastward from island Southeast Asia toward the open Pacific, where, over the next 2500 or so years, they and their descendants discovered and settled all the inhabitable islands of remote Oceania, the vast island world composed of eastern Melanesia, Micronesia, and Polynesia.

Sailing eastward into the ocean was a formidable task as the trade winds blow from the east and inter-island distance increase until they reach more than 3500 km in the far reaches of Polynesia. To accomplish this, pioneering seafarers sailed in outrigger and double canoes, exploited seasonal wind shifts, and navigated without instruments by observing the stars, sun, and other celestial bodies, ocean swells, winds, birds, and other phenomena. They oriented on a conceptual compass of up to 32 points labeled with the rising and setting points of key stars and constellations or the distinctive winds that blew from particular directions. Although they might outline their compass on the ground for teaching purposes, navigators sailed only with its image in mind, along with memorized compass bearings to and from the islands within their voyaging range. They steered on stars and sun when low on the horizon, or on the dominant swells and prevailing wind, all calibrated to their compass points. Their system did not include latitude and longitude, although some navigators may have judged north-south position by what stars passed directly over them, and from the height of Polaris and other circumpolar stars. Progress en route was estimated in terms of the shifting bearings of an unseen reference island off to one side of the course, and other mental plotting methods. To detect land before it could be seen, navigators looked for land-nesting birds fishing out to sea, disruptions in the ocean swells caused by islands, and other signs. That virtually every Oceanic island is part of an archipelago both facilitated discovery of new islands and made navigating between widely separated ones easier. A navigator could make landfall on any island within the chain containing the target island, then reorient toward the final destination.

However, although this Pacific initiative demonstrates that long, navigated voyages can be made without instruments, charts or mathematical calculations, it did not lead to the global, science-based navigation methods employed today.

Navigating Offshore Europe, Asia, and Africa

While canoe voyagers were expanding over the Pacific, other maritime peoples were sailing along the shores of the linked continents of Africa, Europe, and Asia, and to and from adjacent islands. Sea traders and raiders had long sailed in the Mediterranean, the Indian Ocean, and the China Sea. By Roman times – if not before – trade was flowing between the Mediterranean world and China via the Indian Ocean. Judging

from Greek, Arab, and Chinese sources, although much voyaging was coastal, and thus involved navigating primarily by landmarks and soundings, some routes took mariners well away from land, and therefore required at a minimum the ability to sail by steady winds and observations of the stars. Voyages made between Arabia and India by means of alternating monsoon winds provide a case in point.

Deep Sea Navigation and Global Voyaging

Although Herodotus mentioned a Phoenician expedition made around Africa, and much later Viking sailors were able to travel from one island to another across the North Atlantic, regular sailing beyond the linked continents did not begin until the fifteenth-century AD, when the Portuguese developed methods for deep-sea navigation. Early in that century Portuguese mariners began to explore down the African coast, but found that whereas it was easy to sail south with the prevailing northerly winds, tacking back against these was most difficult. Experience gained in sailing to and from the Azores and islands scattered down the African coast had acquainted them with the global winds, the easterly trade winds of the tropics and the westerlies of the temperate latitudes. In sequence, these offered a way to sail back from Africa with fair winds all the way. By heading directly out to sea, these mariners learned that they could sail in a long slant northwest across the easterly trade winds, and, upon reaching the belt of westerlies, could then turn east to sail downwind to Portugal.

However, methods worked out in the Mediterranean for navigating with compass, rudimentary charts, and dead reckoning (estimating position by recording compass heading and distance run) proved inadequate on these long, roundabout crossings. Accordingly, the navigators adapted an astronomical technique for measuring the height of Polaris above the horizon to monitor their progress northward until they were directly west of their destination in Portugal. Before departing from, for example, Lisbon, they marked the altitude of Polaris as seen from that port on a quadrant, a simple instrument for measuring the height of stars above the horizon. (As Polaris was then located about 3.5° from the north celestial pole, with corrections its angular altitude indicated the latitude of the observer.) Then, on the return voyage as they headed west from Africa and then started curving to the north, they watched Polaris rise higher and higher in the sky. Using their quadrant they determined when Polaris had risen to the same height as seen from Lisbon, and then turned east to sail downwind with the westerlies, keeping the Pole Star at the same height in order to reach the port directly.

The crisis caused when, upon crossing the Equator, Polaris dropped below the horizon was resolved in the 1480s. In one of the first examples of a government bringing science to bear on an applied problem, King João II enlisted astronomers, mathematicians, and other scholars to develop a method whereby the height of the sun could be used in place of that of Polaris to determine latitude. With the resultant 'Rule of the Sun,' and tables of the sun's daily position at various latitudes, navigators were able to sail farther and farther down the African coast. In 1497–98 Vasco da Gama sailed in a long arc across the trade winds to the South Atlantic westerlies, then turned east to sail around Africa and up its east coast. From there, under the direction of a local navigator (arguably either an Arab or a Gujerati), he reached India, thereby opening a direct route from Europe to this fabled source of spices, jewels, and other riches.

In 1492 Christopher Columbus used techniques learned from the Portuguese to sail with the trade winds to the Bahamas (which he insisted were outliers of Asia), and then returned by heading north to catch the westerlies. Then in 1522 Juan del Cano brought the *Victoria* back to Spain, completing the first circumnavigation begun by his fallen commander, Ferdinand Magellan. At last the 'world encompassed,' thanks to Iberian seamanship in exploiting global winds and in applying navigational innovations pioneered in other seafaring traditions. For example, a good case can be made that the Chinese developed the magnetic compass, and employed it at sea at least several centuries before its appearance in the Mediterranean in the twelfth century. Furthermore, although the Portuguese use of the height of Polaris to judge progress northward from Africa was a genuine breakthrough, Arab navigators had long before mastered the art of sailing across the Indian Ocean with the aid of Polaris measurements to locate headlands and ports.

Further Navigational Improvements and the World System

The Iberians exploited their maritime achievements by establishing maritime empires, but soon were outgunned and surpassed in both nautical innovation and techniques of colonial exploitation by the Dutch and English. As ship design and seamanship advanced, quantitative navigation evolved through the further application of mathematics and astronomy, and developments in instrumentation and cartography. Astronomical tables multiplied, eventually leading to the publication from 1678 onward of nautical almanacs. Instruments for measuring the altitude of celestial bodies improved, leading to the reflecting octant and sextant in the eighteenth century. Magnetic variation was mapped globally, and improved compasses and ways of measuring distance sailed were introduced, along with charting innovations, including the Mercator projection that enabled mariners to plot long courses with reasonable accuracy.

The last great innovation of this era of sail was the marine chronometer. Whereas navigators could determine latitude, they could only estimate their longitude, resulting in uncertainty, delay, and sometimes maritime tragedy. Because the earth turns on its axis, to determine the longitude of a ship, that is its position east or west of a reference point such as England's Greenwich Observatory, navigators had to keep the time of that reference point. With the development in the late eighteenth century of marine chronometers that could keep time accurately on transoceanic voyages, and their widespread adoption in the nineteenth century, they could at last calculate longitude. The celestial navigation system that had been evolving over the centuries was finally complete. Given clear weather, a navigator could expeditiously and safely direct a ship to any port over the globe. With the conversion from sail to steam, the nautical

means for efficient global maritime trade were in place, and the modern world system erected upon it boomed.

Navigating Electronically

However, the weak point in navigating by magnetic compass, chronometer, and stellar observations remained the changing weather. When clouds, rain, or fog prevented navigators from seeing the stars they had to fall back upon dead reckoning, which, even with instruments for measuring direction of travel and distance run, suffered from cumulative error. Radio direction finding pioneered in the 1910s was an early but only partial solution. However thankful fogbound navigators were to gain radio bearings, these were not very accurate, and available only along the coasts of industrial nations. The first truly all-weather position finding system, LORAN, was developed during World War II for wartime navigation. The reception of synchronized pulses from two pairs of transmitters located at carefully surveyed positions resulted in accurate navigational fixes. The subsequent development of inertial navigational systems was stimulated by further military requirements: guiding missiles and providing bombers and submarines with position information needed for targeting. Employing accelerometers and gyroscopes to determine changes in velocity and direction, these systems provide a continuous indication of position, subject, however, to cumulative error. The most accurate and revolutionary of the electronic navigational technologies is the Global Positioning System (GPS), a US military system made accessible to civilian users, though initially degraded in accuracy. (Russia has its own *Glonass* system.) The GPS system depends upon a constellation of 24 satellites in low earth orbit, which transmit signals that cheap hand-held receivers can process to determine latitude, longitude, and altitude with great accuracy. It was fully operational by the early 1990s, and is now widely used around the world, though not without some misgivings over military control.

Global Navigation and Beyond

Efforts to develop techniques of way finding that began when seafarers first sailed beyond sight of land resulted eventually in a global system of navigation whereby mariners employed the starry points of the celestial sphere to guide their vessels. Now ships, aircraft, and even lone hikers are guided precisely over the globe within an artificial sphere of orbiting satellites. Sophisticated navigational tools also direct robotic and crewed spaceships through interplanetary space, and eventually will be employed to guide them to the stars. Considering the globalizing impacts on human society that have flowed from navigating with increasing accuracy around the planet, thought must now be given to what transformations await humankind as they spread beyond the natal planet.

Bibliography

Casson, L., 1991. The Ancient Mariners, Seafarers and Sea Fighters of the Mediterranean in Ancient Times, second ed. Princeton University Press, Princeton, NJ.

Finney, B., 1998. Traditional navigation and nautical cartography in Oceania. In: Woodward, D., Lewis, G.M. (Eds.), The History of Cartography: Cartography in Traditional African, American, Arctic, Australian, and Pacific Societies, vol. 3, pt. 2. University of Chicago Press, Chicago, pp. 443–492.

Finney, B.R., Jones, E.M. (Eds.), 1985. Interstellar Migration and the Human Experience. University of California Press, Berkeley, CA.

Getting, I.A., 1993. The Global Positioning System. IEEE Spectrum 30 (12), 36–47.

Hourani, G.F., 1963. Arab Seafaring in the Indian Ocean in Ancient and Early Medieval Times. Khayats, Beirut, Lebanon.

Lewis, D., 1972. We the Navigators. University of Hawaii Press, Honolulu, HI.

Nadvi, S.S., 1966. The Arab Navigation. Sh. Muhammad Ashraf, Lahore, Pakistan.

Needham, J., 1971. Science and Civilization in China, vol. 4. Physics and Physical Technology, Part III Civil Engineering and Nautics. Cambridge University Press, Cambridge, UK.

Parry, J.H., 1981. The Discovery of the Sea, first ed. University of California Press, Berkeley, CA.

Taylor, E.G.R., 1956. The Haven-Finding Art, A History of Navigation from Odysseus to Captain Cook. Hollis & Carter, London.

Waters, D., 1992. Columbus's Portuguese Heritage. The Mariner's Mirror 78, 385–405.

Williams, J.E.D., 1992. From Sails to Satellites, the Origin and Development of Navigational Science. Oxford University Press, Oxford, UK.

Navigation in Virtual Space: Psychological and Neural Aspects

Derek A Hamilton and Daniel Barto, University of New Mexico, Albuquerque, NM, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by N. Burgess, volume 15, pp. 10417–10422, © 2001, Elsevier Ltd.

Abstract

Virtual, computerized spatial navigation tasks have become increasingly popular and important tools for the study of behavioral and cognitive processes involved in spatial navigation and the underlying neural systems. This article provides a broad overview of (1) commonly used virtual navigation tasks, (2) basic behavioral observations relevant to strategies, learning processes, and attention, (3) effects of sex and age, (4) performance in individuals with neurological disorders or brain damage, (5) relationships between measures of brain structure and performance, and (6) findings from functional neuroimaging studies of virtual navigation. The use of virtual navigation approaches in nonhuman animals is also briefly summarized.

Introduction

Three-dimensional computer-generated space, also commonly referred to as 'virtual space,' 'virtual reality,' or a 'virtual environment,' has become an increasingly popular research tool in the experimental investigation of behavioral and cognitive processes involved in navigation and the corresponding neural systems that support these processes. Visual aspects of a virtual environment are typically presented on one or more displays or via a head mounted display (HMD) that can provide the viewer with more immersive experience of the virtual environment. Virtual environments are most commonly rendered from the viewer's perspective (i.e., the first-person perspective); however, third-person perspectives in which the virtual environment is rendered such that the viewer sees an actor (avatar) from a fixed virtual distance and angle are also common. In the case of the first-person perspective the orientation of the viewer can be controlled via a joystick, keyboard, or mouse, or – in the case of HMDs – sensors that detect rotational movement of the head. The term 'virtual navigation' refers to the movement from one location in a virtual environment to another, which can be controlled by the viewer via manipulation of a device (e.g., joystick and keyboard) that results in directional movement within the virtual environment.

Benefits, Appeal, and Limitations of Virtual Navigation in the Behavioral Sciences

The study of navigation by nonhuman animals in mazes and other spatial navigation tasks has a long and rich history in experimental psychology and has been central to research on the neurobiology of spatial behavior and cognition, including contemporary approaches that emphasize the use of spatial navigation tasks to better understand the neurobiology of learning and memory. Since the first studies published in the 1990s, the use of virtual navigation tasks has been increasingly utilized to investigate the behavioral and cognitive processes and underlying neural systems in humans. In addition to allowing laboratory investigations of human navigation, virtual navigation tasks and the virtual environments utilized in them offer other benefits, including excellent control over the visual aspects of the environment, control over how the viewer interacts with the environment, and the ability to compose environments that might otherwise be impractical or impossible to realize in the real world.

The benefits of using virtual navigation tasks do, however, come with some limitations. Perhaps most notably, the lack of realistic movement from one location to another as well as the vestibular, proprioceptive, and haptic (touch) stimulation that accompanies movement is often recognized as a major limitation in generalizing the results of virtual navigation studies to real-world navigation (Taube et al., 2013). Despite these limitations on the range and quality of sensory and perceptual experiences in virtual navigation, the transfer of knowledge acquired in the virtual domain to the real-world has been documented in several reports (Brooks et al., 1999; Foreman et al., 2003). These limitations can also be overcome, for example, through the use of haptic stimulation or immersive environments in which the participant's movements from one location to another in real space are translated to movements in virtual space. It is, further, important to recognize that behavioral and cognitive processes in virtual navigation tasks are genuine processes that reflect the capacities of participants to navigate based on the available information. With the increased use of virtual environments in everyday life and specialized training applications, understanding the behavioral and cognitive processes involved in performance in virtual tasks is an important and challenging undertaking quite apart from the potential generality to real-world navigation processes. Characterization of the basic similarities and differences in the behavioral, cognitive, and neurobiological processes involved in navigation in real and virtual space will be an important topic of study for future research.

Types of Virtual Navigation Tasks

Virtual navigation tasks can be grouped broadly into two categories based on visual content, task demands, and scientific questions: (1) those that attempt to replicate laboratory spatial navigation tasks used in nonhuman animals in the virtual domain and (2) those that attempt to more closely replicate environments encountered by people during typical day-to-day

navigation such as cities or the interior of buildings. Many of the early virtual navigation studies published in the late 1990s involved simulating tasks such as the Morris water task (MWT) for rodents (Morris, 1981). In this task, rodents (typically rats) are placed into a circular pool of cool, opaque water in which a small, hidden (submerged) platform that allows escape from the water is located at a particular spatial location. Visual cues placed in the room beyond the boundary of the pool provide the major source of information used by the animals to swim to the escape platform. Animals that have mastered this task execute more or less direct swims to the hidden platform from multiple start locations and will persist in searching at the platform location if it is removed for a probe trial. The MWT has become the gold standard for measurement of spatial learning and memory in rodents, owing to its simplicity, its unambiguous measurement of spatial navigation, and the clear demonstrations that learning and memory in this task critically depend upon the hippocampus (Morris et al., 1982; Sutherland et al., 1982). Several laboratories have developed and utilized virtual navigation tasks that attempt to capture the important features of behavior, learning, and control of behavior by visual stimuli in the MWT (Astur et al., 1998; Hamilton and Sutherland, 1999; Jacobs et al., 1998; Sandstrom et al., 1998; Gandhi et al., 2007; Piper et al., 2011). The major environmental features of a virtual MWT (VMWT) are a local frame of reference such as a pool or other notional apparatus that includes a conspicuous, visible boundary that provides no disambiguating information regarding spatial location, a hidden goal location (e.g., a hidden platform) that can be made visible to provide a visual cue co-localized with the goal, and a constellation of conspicuous visual cues in the virtual space beyond the apparatus boundary that provide information that can be used for spatial localization and orientation. An example view and layout of a VMWT are shown in **Figure 1**.

The popular radial arm maze (RAM) has also served as a model for virtual navigation tasks that measure reference and working memory in the spatial domain (Astur et al., 2004; Bohbot et al., 2007; Braun et al., 2012). In a typical RAM or virtual RAM (VRAM) there are eight arms that extend out from a central platform, half of which have some item of value to the participant at the end of the arm (e.g., food in the RAM and points in the VRAM), while the other arms hold nothing of value. The maze remains in the same position and orientation relative to visual cues beyond the maze, so the participant can learn to navigate to the locations where there are items of interest and avoid empty arms over the course of training. This type of learning and memory is commonly referred to as 'reference memory,' whereas keeping track of which arms have been visited during a trial is considered a 'working memory' component of the task. The time required to complete a trial and the number of errors (entries into empty arms) are measured. An example view in a VRAM and layout are shown in **Figure 2**. Mazes that require turns at choice points ranging in complexity from the relatively simple T-mazes to mazes with multiple choice points have also been employed (Grön et al., 2000; Moffat et al., 1998; Shore et al., 2001; Baker and Holroyd, 2009; Schmitzer-Torbert, 2007). Canovas et al. (2008) also developed a hole board comparable to its rodent counterpart in which participants must locate multiple rewards among a set of possible goal locations. Virtual navigation tasks have also been utilized to assess learning about geometric features of an environment (Redhead and Hamilton, 2007, 2009; Wilson and Alexander, 2010, 2008). All of these tasks share in common the use of first-person perspective and three-dimensional presentation to simulate viewing real space, and that they were designed to measure performance in tasks that attempt to closely mirror common laboratory tasks. Virtual environments that represent real-world settings such as cities or interiors of buildings are also commonly used to either train participants to navigate in a particular environment or to evaluate navigation in a situation that more closely mirrors commonly utilized navigation skills and strategies in a real-

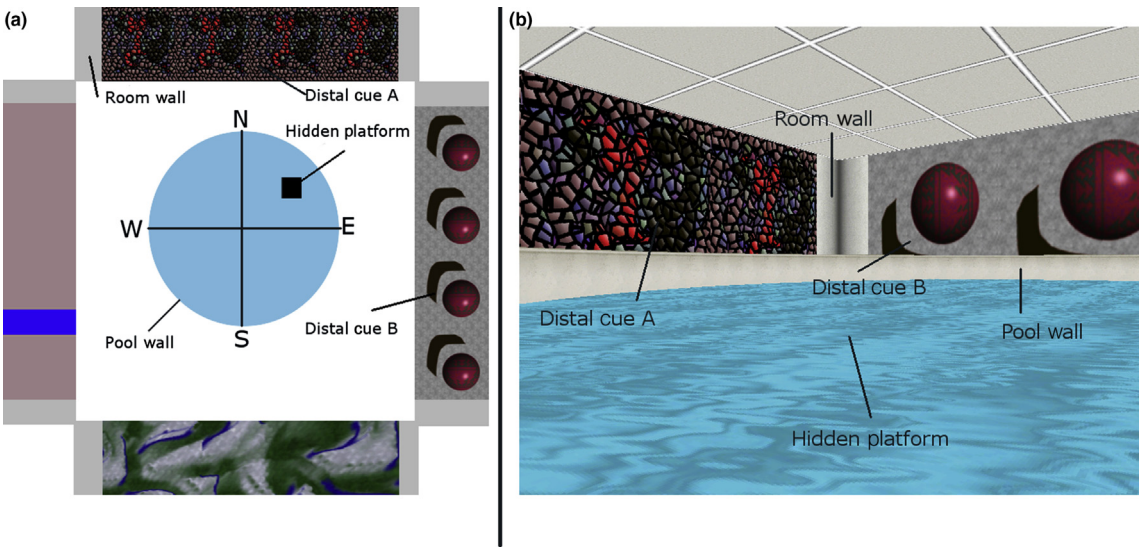

Figure 1 The virtual Morris water task environment. Panel (a) depicts an aerial view of the virtual environment, with the four vertical walls flattened. The circular pool is positioned in the center of the room. The hidden platform is shown as a black square within the pool. Panel (b) is a first-person view of the virtual environment that would be presented to the participant.

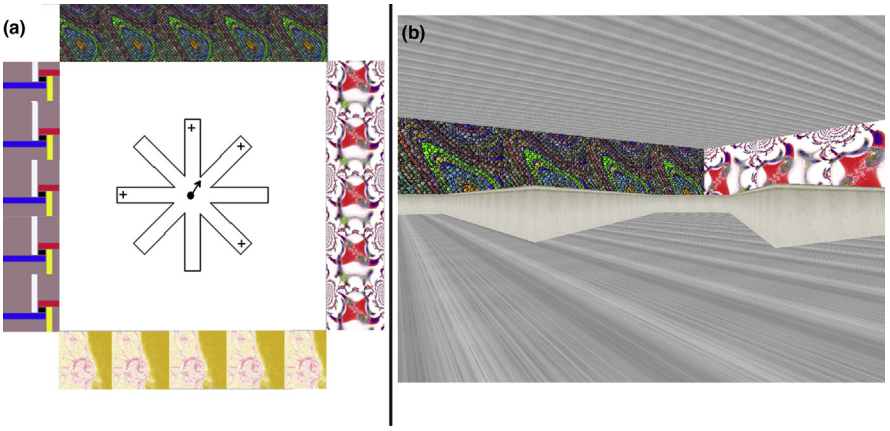

Figure 2 A virtual eight-arm radial arm maze (VRAM). Panel (a) depicts an aerial view of the VRAM, with the four vertical walls flattened. The maze is positioned in the center of the room. '+' indicate the "reward" locations. Panel (b) is a first-person view of the virtual environment.

world setting (Brooks et al., 1999; Maguire et al., 1998; Spiers and Maguire, 2007b; Thibault et al., 2013; Wilson et al., 2004).

Psychological Aspects

Navigation Strategies, Behavior, and Cognitive Processes

A number of studies have utilized virtual navigation tasks to investigate the strategies used by humans to navigate under various stimulus conditions and task demands. The conceptual dissociation between allocentric – (world-based) in which the viewer is thought to possess a representation of the external world in which the topographical relationships among the features of the environment are maintained – versus egocentric – (self-based) strategies in which the individual executes movements by way of reference to one's current position or orientation has been widely utilized in research on spatial navigation. In particular, the clearly established link between allocentric navigation and the hippocampus has been a major driving force in the literature on the neurobiology of spatial navigation and spatial cognition (O'Keefe and Nadel, 1978; Knierim and Hamilton, 2011). Accordingly, a considerable amount of research using virtual environments has been aimed at better understanding allocentric navigation, or navigation to a particular location based on its spatial relationship to visual cues, heretofore referred to as 'place navigation.' A variety of distinct behavioral and cognitive processes associated with orientation in the spatial domain have been postulated and measured in virtual navigation tasks (Arnold et al., 2013; Hamilton et al., 2009). With respect to behaviors controlled by external stimuli, place navigation is commonly contrasted with 'cued navigation' (navigation to a location based on the presence of one or more cues located at that location), which can be classified as a form of 'simple guidance,' or 'complex guidance' if sequences of simple guidance responses are involved. 'Response strategies' as in egocentric navigation or selecting movements based on the directional left/right response to be made relative to one's current position, and more recently 'directional responding' (moving in the direction of a goal location by selecting an orientation based on visual features of the environment) are also contrasted with place or allocentric navigation strategies.

An early study by Jacobs et al. (1998) found that participants utilized the topographical relationships between visual cues in the distal room environment in a task similar to the VMWT. Astur et al. (1998) characterized the performance of young university students in a VMWT and noted that many participants took more or less direct trajectories to the goal location from multiple release points and would persist in searching at this location if the goal was removed for a probe trial, consistent with use of a spatial or place navigation strategy, in which the available cues are used to guide navigation to a particular location in the environment. In a VMWT some participants adopt non-spatial strategies that involve non-direct paths to the goal that often include systematic searching of the environment until the goal is encountered (Hamilton et al., 2009). Such strategies usually involve navigating based on features of the local (apparatus) reference frame provided by the circular boundary of the pool that provide no disambiguating spatial information to guide navigation. For example, a participant may learn that the goal is located at a particular distance from the boundary of the local reference frame and perform a circuitous swim around the apparatus at the appropriate distance from the boundary until the goal is encountered.

Hamilton et al. (2002) composed a VMWT in which access to some regions of the virtual pool was restricted such that the participants could not navigate through some regions of the environment, while they could see the entire environment. These authors found that lack of experience in navigating through a region of the environment on a trajectory to the goal location disrupted later navigation through the restricted regions. The authors suggested that familiarity with the views experienced while navigating was important for accurate performance. The results were also highly similar to those observed in rats by Sutherland et al. (1987) using similar manipulations, suggesting a basic similarity in these species in responses to manipulation of distal visual cues in their respective tasks.

A variant of the VMWT in which the goal location is made visible allows for measurement of 'cued navigation' which is

commonly utilized to assess navigation when a spatial strategy is not required and is not the most efficient strategy to support navigation to the goal. Chamizo et al. (2003) employed a VMWT in which both distal room cues and a cue co-localized with the goal were present simultaneously and found a reduction in performance when the cue marking the goal was removed for a test trial, requiring the participant to navigate using the distal cues. That is, the presence of a cue at the goal interfered with learning about the distal cues, a phenomenon referred to as 'overshadowing.' Similar virtual navigation tasks have been utilized to determine whether other types of spatial information such as geometries formed by the apparatus or room are also interfered with or are impervious to the presence of cues that support non-spatial or non-geometric strategies (Alexander et al., 2009; Redhead and Hamilton, 2007, 2009; Redhead et al., 2013; Wilson and Alexander, 2008, 2010). Although the pool in the VMWT is circular and, therefore, does not provide information capable of disambiguating spatial location, Hamilton et al. (2009) hypothesized that the frame of reference provided by the boundary of the pool was, nonetheless, an important determinant of behavior. These authors systematically reduced the height of the pool boundary (ranging from a typical height as shown in **Figure 1** to completely invisible) and found that performance systematically declined as the height of the pool boundary was reduced. One conclusion offered by these authors was that the distal cues disambiguated the otherwise spatially ambiguous pool reference frame, and as the signal provided by the latter frame of reference was reduced, the performance declined as well. Importantly, these findings also indicate that distal cues are not sufficient for spatial navigation in the VMWT, as the reduction in the height of the boundary also reduced the degree to which the distal cues were obscured by the boundary.

Attention

Several groups have integrated virtual navigation tasks with eye-tracking technology to investigate how attention to visual features of the virtual environment, as assessed via gaze measurements, is related to strategy selection or performance. Mueller et al. (2008) used a VMWT to examine sex differences in visual sampling of the virtual environment in relation to sex differences in strategy selection. Eye-tracking data were obtained during the initial portion of the trials in which participants started at several distinct locations in the virtual pool. Males tended to visually sample more of the virtual space than females, who had longer gaze durations directed at particular locations or cues. Livingstone-Lee et al. (2011) also examined visual gaze during the initial portion (1 s) of each trial and observed that attention to the available cues indicated which of two strategies a participant would employ during the trial. Using a VMWT Hamilton et al. (2009) recorded gaze location throughout each trial while participants navigated to a hidden goal and characterized attention to distal visual cues and the local apparatus (pool) boundary by university students who either learned to solve task using a spatial strategy or who adopted a non-spatial strategy. Participants who adopted a spatial strategy initially looked at distal cues and then shifted their attention to the precise location in the pool to which they were navigating for the remainder of the trial. The authors suggested that initial orientation was based on distal visual cues while goal-directed navigation was based on local apparatus features, comparable to shifts that occur during active navigation by rats in the MWT when both local and distal cues are available (Hamilton et al., 2004). In contrast, participants who adopted non-spatial strategies tended to keep their gaze restricted to the region within the pool boundary with little sampling of distal visual cues and no particular focus of the gaze toward any particular location. Andersen et al. (2012) combined eye-tracking with an eight-arm VRAM and found that females spent more time visually sampling landmarks compared to males, while male sampling of visual landmarks decreased with experience. Because information about visual attention can be gathered as individuals navigate in virtual environments, the merging of these technologies will likely play an important role in future investigations concerned with characterizing how visual information is utilized in the service of navigation and how attention to visual information in the spatial domain is influenced by stimulus parameters and modified by experience.

Sex Differences and Aging

Sex differences in spatial abilities have been well characterized and with the advent of virtual navigation tasks several studies have assessed sex differences in ability and strategy selection. Astur et al. (1998) reported a large sex difference in place navigation using a VMWT, with males performing better than females. Driscoll et al. (2005) found that males outperformed females in a VMWT even well into adulthood (>60 years) and that this sex difference across the lifespan was partially explained by circulating levels of testosterone. Mueller et al. (2008) also reported that males performed better than females in a VMWT and, as described in the previous section, tended to sample a larger range of the virtual space as assessed via eye-tracking. One possibility noted by these authors is that males may simply choose to use a spatial (allocentric) strategy more than females. Sandstrom et al. (1998) utilized a similar virtual navigation task and found that females tended to rely on using landmarks while males utilized both geometric features of the environment as well as landmarks. Woolley et al. (2010) used a VMWT and found that the most robust sex difference was evident in the accuracy of the initial trajectory to the goal, whereas only a nonsignificant numerical difference was observed for measures of search behavior. Sex differences have also been observed in complex virtual mazes. Moffat et al. (1998) found that males learn to complete a complex, multiple choice-point virtual maze faster than females and commit fewer errors. Cimadevilla et al. (2011) also reported that males performed better than females in a virtual labyrinth maze. van Gerven et al. (2012) utilized a VMWT in which both place navigation or an egocentric strategy were equally effective for solving the task and found that females adopted a spatial strategy at a higher rate than males. Quantitative measures of performance such as latency to locate the goal or persistence in searching during a probe still favored males, leading these authors to conclude

that sex differences observed in virtual navigation tasks may be more a matter of efficiency in performance rather than strategy selection. One possibility is that differences in experience playing video games that utilize virtual space may contribute to sex differences (Astur et al., 1998; Hamilton et al., 2002). Some studies have also failed to detect sex differences in a VMWT in the same university population that yielded robust sex differences in earlier studies (Hamilton et al., 2009), which might reflect a narrowing of this experience gap.

Several studies have also examined the relationship between aging and performance in virtual navigation tasks. Driscoll et al. (2003, 2005) reported deficits in spatial learning in a VMWT in older adults (60–85 years) compared to younger adults (20–59 years). These age-related declines were accompanied by decreases in hippocampal neurometabolites and volume as well as circulating levels of testosterone. Similar deficits in virtual spatial navigation with age were observed by other groups using a VMWT or comparable task (Iaria et al., 2009; Laurance et al., 2002; Moffat and Resnick, 2002).

Neural Aspects

Virtual Spatial Navigation in Patient Populations

The well-established dependence of spatial navigation in rodents, as in the MWT, on the hippocampus and related circuitry (Sutherland et al., 1982; Morris et al., 1982) motivated several laboratories to examine the performance of patients with damage to medial temporal lobe structures in virtual spatial navigation. Amnesic patients with medial temporal lobe resections are impaired in the VMWT (Astur et al., 2002; Goodrich-Hunsaker et al., 2010) and a VRAM (Bohbot et al., 2004; Goodrich-Hunsaker and Hopkins, 2010). Bartsch et al. (2010) subsequently demonstrated deficits in a VMWT in patients with focal damage to the CA1 subfield of the hippocampus. A number of laboratories have also reported deficits in virtual navigation tasks in patient populations in which hippocampal pathology provided the motivation to evaluate place learning and navigation in these populations. These include deficits in schizophrenic patients (Hanlon et al., 2006), individuals with traumatic brain injuries (Skelton et al., 2000), children with fetal alcohol syndrome (Hamilton et al., 2003), patients with unilateral or bilateral vestibular failure (Brandt et al., 2005; Hüfner et al., 2007), temporal lobe epileptics (Cimadevilla et al., 2011), pediatric anxiety patients (Mueller et al., 2009), individuals with depression (Cornwell et al., 2010), and individuals with posttraumatic stress disorder (Tempesta et al., 2012).

Virtual Navigation and Structural Neuroimaging

Several studies have linked structural measurements in brain regions involved in real-world navigation and performance in virtual navigation tasks. Driscoll et al. (2003) reported a strong positive correlation between time spent searching at the goal location during a no-platform probe trial in the VMWT and hippocampal volume. Iaria et al. (2008) also reported that navigation performance in a virtual environment is related to structural features of the hippocampus. Using diffusion tensor imaging (DTI) these authors found that individuals with high fractional anisotropy, a measure of axonal density and myelination, in the right hippocampus, were faster at constructing a map-like representation of the virtual environment and better at using this information to orient. Hippocampal atrophy related to vestibular loss has also been linked to poor spatial performance in the VMWT (Brandt et al., 2005). Segmentation analyses of gray matter density have also been correlated with strategy selection. In a VRAM Bohbot et al. (2007) found that individuals who adopt a spatial strategy had higher levels of gray matter in the hippocampus and lower levels of gray matter in the caudate nucleus compared to individuals who adopted strategies that involve navigation based on individual cues. Evensmoen et al. (2013) also found that volume of specific hippocampal subregions varied with the resolution of an individual's spatial representation of a virtual environment, with more coarse representations correlated with anterior hippocampal volume and more fine-grained representation associated with posterior hippocampal volume. Correlations between performance in a VMWT and gray matter in the inferior parietal lobule have also been observed in preadolescent males (Mueller et al., 2011).

Virtual Navigation and Functional Neuroimaging

Neuroimaging experiments have consistently yielded evidence that the hippocampus, parahippocampal gyrus, and parietal cortex are engaged during virtual spatial navigation. In an early fMRI investigation of virtual navigation Aguirre et al. (1996) scanned participants while they navigated in a series of virtual corridors and observed increased activation of parahippocampal cortex bilaterally. In a positron emission tomography (PET) study Maguire et al. (1998) found that exploration of virtual rooms that contain salient landmarks result in increased activation of the right parahippocampal gyrus compared to rooms devoid of conspicuous landmarks. Aguirre and D'Esposito (1997) also demonstrated that the parahippocampal region is recruited when participants are asked to recognize a location in a virtual city or select the direction of another location. The inferior parietal cortex was activated to a greater degree in the direction selection task than the recognition task, an observation consistent with the distinction between spatial localization as in an allocentric strategy and orientation/trajectory selection. Spiers and Maguire (2007a) reported dissociation between egocentric direction and distance to locations in an fMRI experiment, with the former correlated with parietal cortex activation and the latter positively correlated with frontal cortex activity and negatively correlated with the activity in the subiculum and entorhinal cortex, the major neocortical input to the hippocampus. Grön et al. (2000) measured brain activation with fMRI while males and females navigated in a maze that included multiple choice points. Both males and females displayed activation of right hippocampus, parahippocampal gyrus, and parietal cortex. Males displayed higher levels of activation in the left parahippocampal gyrus and left hippocampus, whereas females displayed higher levels of activation in the right parietal and right frontal cortices. These authors reasoned that the observed differences reflect sex differences in strategy selection, with females relying

more on landmarks than geometric features of the environment (Sandstrom et al., 1998).

Iaria et al. (2003) found that participants who used spatial information provided by configurations of landmarks displayed increases in hippocampal activity, whereas following more extensive experience participants who utilized non-spatial strategies displayed sustained activity within the caudate nucleus, an aspect of the basal ganglia implicated in well-established behavioral responses and habit. Bohbot et al. (2004) similarly found that the use of a spatial strategy resulted in increased hippocampal activity, whereas the use of a simple response strategy resulted in an increased activity in the caudate nucleus. Doeller et al. (2008) used a circular arena surrounded by cues, similar to the configuration of a VMWT, but which also contained local landmarks, to dissociate brain regions involved in navigation based on boundary features of an environment and landmarks. Boundary information was associated with hippocampal activation whereas landmark processing was related to activity in the dorsal striatum, which includes the caudate nucleus. Evensmoen et al. (2013) reported dissociations between anterior and posterior hippocampus based on whether participants possessed a local, fine-grained, or global, coarse, representation. Anterior hippocampal activity was associated with global representation of the virtual environment whereas posterior hippocampal activity was related to both local representation of the environmental features and the use of non-spatial strategies. The common regions activated during virtual navigation, namely the hippocampus, parahippocampal gyrus, and parietal cortex have also been shown to display decreased activation in elderly individuals compared to young participants during virtual navigation (Moffat et al., 2006).

In the rat, movements related to navigation from one location to another are associated with low-frequency oscillations in the 3–15 Hz (rhythmic slow activity) or theta (~4–8 Hz) range, which has been linked to processing of spatial information during active navigation (Okeefe and Recce, 1993). Kahana et al. (1999) recorded electroencephalograms (EEGs) from depth electrodes implanted in the hippocampi of epileptic patients while they performed a virtual navigation task. Theta activity in the temporal cortex was elevated in more complex virtual mazes. Cornwell et al. (2008) recorded magnetoencephalography (MEG) signals while participants navigated in a VMWT and found increased theta activity in the left hippocampus during performance of the VMWT compared to a task that did not involve goal-directed movements. These observations indicate the possibility of a basic similarity between the role of low-frequency oscillations in rats and humans during performance of real and virtual navigation tasks (O'Keefe and Burgess, 1999; Watrous et al., 2013).

Collectively, the available data on the role of the hippocampus and related structures in human virtual navigation is highly consistent with data from rodent studies showing that place navigation is impaired following hippocampal damage, and the existence of hippocampal place cells that fire selectively when the animal is in a particular location in an environment (O'Keefe and Dostrovsky, 1971). More recently, the so-called 'grid cells' in the medial entorhinal cortex, a major neocortical input to the hippocampus, have been described in rats, bats, and monkeys. Grid cells fire at regular spatial intervals as an animal moves through an environment such that when recorded over a long period of exploration, the spatial plots of firing rates are organized in a grid (Hafting et al., 2005). One possibility is that these cells provide information about the spatial scale of movements and location within an environment based on both internal and external cues. Jacobs et al. (2013) found cells in patients implanted with depth electrodes that display grid-like firing rate properties when participants were performing a virtual navigation task. Jacobs et al. (2010) also characterized a population of entorhinal cortex neurons that selectively coded for the direction of movement in a virtual environment, which is consistent with observations in rats showing that some grid cells have directional firing properties (Sargolini et al., 2006). Collectively, these observations indicate that many aspects of the neural bases of navigation characterized using invasive studies in rodents are also observed in humans at various levels of analysis.

Virtual Navigation in Nonhuman Animals

Among the major considerations that contribute to the appeal of using virtual navigation technology is the capacity to measure spatial navigation in humans in a laboratory setting. Other appealing features of virtual navigation tasks such as increased control over environments and the ability to compose environments in numbers or scope that would be impractical, however, would also be beneficial in studies that examine navigation in nonhuman animals. Several laboratories have utilized virtual environments to examine navigation by monkeys (Furuya et al., 2014; Hori et al., 2005, 2003; Sato et al., 2004; Washburn and Astur, 2003), allowing for the simultaneous recording of individual neurons during virtual navigation. For example, Furuya et al. (2014) found that neurons in the parahippocampal gyrus showed reliable responses across three environments in which the scaling of the environment was manipulated in relation to the distal visual cues, while neurons in the hippocampus showed reliable place responses across environments. In both regions, the firing properties of the neurons were modified based on environment scale, with reductions in scale associated with reductions in the size of place fields. Virtual navigation tasks have also been utilized in non-primates. Miyata and Fujita (2010) trained pigeons to follow routes based on optic flow information provided via an LCD display. Virtual navigation tasks have also been used successfully in rats (Holscher et al., 2005; Youngstrom and Strowbridge, 2012) and mice (Dombeck et al., 2010; Harvey et al., 2009). The use of virtual navigation tasks in nonhuman animals will allow for comparative analyses that are less confounded by differences between real-world and virtual navigation, and for recordings of individual neural activity during ongoing behavior to be characterized and incorporated with the rich literature on individual neural activity during real-world navigation in standard laboratory tasks. Such studies will provide information about the basic differences between real-world and virtual navigation crucial for

understanding the limitations on generalizations that can be made from results obtained in virtual navigation tasks.

See also: Dysexecutive Amnesia; Eye Movement, Control of: Oculomotor Control; Hippocampal Amnesia; Hippocampus and Related Structures; Human Spatial Orientation, Neural Basis of; Memory Models: Quantitative; Parietal Lobe; Spatial Cognition During Infancy and Early Childhood Across Cultures, Development of; Spatial Navigation; Temporal Lobe; Virtual Reality and Spatial Cognition.

Bibliography

Aguirre, G.K., D'Esposito, M., 1997. Environmental knowledge is subserved by separable dorsal/ventral neural areas. Journal of Neuroscience 17, 2512–2518.

Aguirre, G.K., Detre, J.A., Alsop, D.C., Desposito, M., 1996. The parahippocampus subserves topographical learning in man. Cerebral Cortex 6, 823–829.

Alexander, T., Wilson, S.P., Wilson, P.N., 2009. Blocking of goal-location learning based on shape. Journal of Experimental Psychology Learning Memory and Cognition 35, 694–708.

Andersen, N.E., Dahmani, L., Konishi, K., Bohbot, V.D., 2012. Eye tracking, strategies, and sex differences in virtual navigation. Neurobiology of Learning and Memory 97, 81–89.

Arnold, A., Burles, F., Krivoruchko, T., Liu, I., Rey, C.D., Levy, R.M., Iaria, G., 2013. Cognitive mapping in humans and its relationship to other orientation skills. Experimental Brain Research 224, 359–372.

Astur, R.S., Ortiz, M.L., Sutherland, R.J., 1998. A characterization of performance by men and women in a virtual Morris water task: a large and reliable sex difference. Behavioural Brain Research 93, 185–190.

Astur, R.S., Taylor, L.B., Mamelak, A.N., Philpott, L., Sutherland, R.J., 2002. Humans with hippocampus damage display severe spatial memory impairments in a virtual Morris water task. Behavioural Brain Research 132, 77–84.

Astur, R.S., Tropp, J., Sava, S., Constable, R.T., Markus, E.J., 2004. Sex differences and correlations in a virtual Morris water task, a virtual radial arm maze, and mental rotation. Behavioural Brain Research 151, 103–115.

Baker, T.E., Holroyd, C.B., 2009. Which way do i go? Neural activation in response to feedback and spatial processing in a virtual T-maze. Cerebral Cortex 19, 1708–1722.

Bartsch, T., Schonfeld, R., Muller, F.J., Alfke, K., Leplow, B., Aldenhoff, J., Deuschl, G., Koch, J.M., 2010. Focal lesions of human hippocampal CA1 neurons in transient global amnesia impair place memory. Science 328, 1412–1415.

Bohbot, V.D., Iaria, G., Petrides, M., 2004. Hippocampal function and spatial memory: evidence from functional neuroimaging in healthy participants and performance of patients with medial temporal lobe resections. Neuropsychology 18, 418–425.

Bohbot, V.D., Lerch, J., Thorndycraft, B., Iaria, G., Zijdenbos, A.P., 2007. Gray matter differences correlate with spontaneous strategies in a human virtual navigation task. Journal of Neuroscience 27, 10078–10083.

Brandt, T., Schautzer, F., Hamilton, D.A., Bruning, R., Markowitsch, H.J., Kalla, R., Darlington, C., Smith, P., Strupp, M., 2005. Vestibular loss causes hippocampal atrophy and impaired spatial memory in humans. Brain 128, 2732–2741.

Braun, J.M., Lucchini, R., Bellinger, D.C., Hoffman, E., Nazzaro, M., Smith, D.R., Wright, R.O., 2012. Predictors of virtual radial arm maze performance in adolescent Italian children. Neurotoxicology 33, 1203–1211.

Brooks, B.M., McNeil, J.E., Rose, F.D., Greenwood, R.J., Attree, E.A., Leadbetter, A.G., 1999. Route learning in a case of amnesia: a preliminary investigation into the efficacy of training in a virtual environment. Neuropsychological Rehabilitation 9, 63–76.

Canovas, R., Espinola, M., Iribarne, L., Manuel Cimadevilla, J., 2008. A new virtual task to evaluate human place learning. Behavioural Brain Research 190, 112–118.

Chamizo, V.D., Aznar-Casanova, J.A., Artigas, A.A., 2003. Human overshadowing in a virtual pool: simple guidance is a good competitor against locale learning. Learning and Motivation 34, 262–281.

Cimadevilla, J.M., Canovas, R., Iribarne, L., Soria, A., Lopez, L., 2011. A virtual-based task to assess place avoidance in humans. Journal of Neuroscience Methods 196, 45–50.

Cornwell, B.R., Johnson, L.L., Holroyd, T., Carver, F.W., Grillon, C., 2008. Human hippocampal and parahippocampal theta during goal-directed spatial navigation predicts performance on a virtual Morris water maze. Journal of Neuroscience 28, 5983–5990.

Cornwell, B.R., Salvadore, G., Colon-Rosario, V., Latov, D.R., Holroyd, T., Carver, F.W., Coppola, R., Manji, H.K., Zarate Jr., C.A., Grillon, C., 2010. Abnormal hippocampal functioning and impaired spatial navigation in depressed individuals: evidence from whole-head magnetoencephalography. American Journal of Psychiatry 167, 836–844.

Doeller, C.F., King, J.A., Burgess, N., 2008. Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. Proceedings of the National Academy of Sciences (USA) 105, 5915–5920.

Dombeck, D.A., Harvey, C.D., Tian, L., Looger, L.L., Tank, D.W., 2010. Functional imaging of hippocampal place cells at cellular resolution during virtual navigation. Nature Neuroscience 13, 1433–U1180.

Driscoll, I., Hamilton, D.A., Petropoulos, H., Yeo, R.A., Brooks, W.M., Baumgartner, R.N., Sutherland, R.J., 2003. The aging hippocampus: cognitive, biochemical and structural findings. Cerebral Cortex 13, 1344–1351.

Driscoll, I., Hamilton, D.A., Yeo, R.A., Brooks, W.M., Sutherland, R.J., 2005. Virtual navigation in humans: the impact of age, sex, and hormones on place learning. Hormones and Behavior 47, 326–335.

Evensmoen, H.R., Lehn, H., Xu, J., Witter, M.P., Nadel, L., Haberg, A.K., 2013. The anterior hippocampus supports a coarse, global environmental representation and the posterior hippocampus supports fine-grained, local environmental representations. Journal of Cognitive Neuroscience 25, 1908–1925.

Foreman, N., Stanton, D., Wilson, P., Duffy, H., 2003. Spatial knowledge of a real school environment acquired from virtual or physical models by able-bodied children and children with physical disabilities. Journal of Experimental Psychology-Applied 9, 67–74.

Furuya, Y., Matsumoto, J., Hori, E., Boas, C.V., Anh Hai, T., Shimada, Y., Ono, T., Nishijo, H., 2014. Place-related neuronal activity in the monkey parahippocampal gyrus and hippocampal formation during virtual navigation. Hippocampus 24, 113–130.

Gandhi, M., Skelton, R., Livingstone, S., Gillingham, S., 2007. Gender differences in spatial navigation in a virtual water maze: negation by landmarks. Canadian Journal of Experimental Psychology-Revue Canadienne De Psychologie Experimentale 61, 366.

van Gerven, D.J.H., Schneider, A.N., Wuitchik, D.M., Skelton, R.W., 2012. Direct measurement of spontaneous strategy selection in a virtual Morris water maze shows females choose an allocentric strategy at least as often as males do. Behavioral Neuroscience 126, 465–478.

Goodrich-Hunsaker, N.J., Hopkins, R.O., 2010. Spatial memory deficits in a virtual radial arm maze in amnesic participants with hippocampal damage. Behavioral Neuroscience 124, 405–413.

Goodrich-Hunsaker, N.J., Livingstone, S.A., Skelton, R.W., Hopkins, R.O., 2010. Spatial deficits in a virtual water maze in amnesic participants with hippocampal damage. Hippocampus 20, 481–491.

Grön, G., Wunderlich, A.P., Spitzer, M., Tomczak, R., Riepe, M.W., 2000. Brain activation during human navigation: gender-different neural networks as substrate of performance. Nature Neuroscience 3, 404–408.

Hafting, T., Fyhn, M., Molden, S., Moser, M.B., Moser, E.I., 2005. Microstructure of a spatial map in the entorhinal cortex. Nature 436, 801–806.

Hamilton, D.A., Driscoll, I., Sutherland, R.J., 2002. Human place learning in a virtual Morris water task: some important constraints on the flexibility of place navigation. Behavioural Brain Research 129, 159–170.

Hamilton, D.A., Johnson, T.E., Redhead, E.S., Verney, S.R., 2009. Control of rodent and human spatial navigation by room and apparatus cues. Behavioural Processes 81, 154–169.

Hamilton, D.A., Kodituwakku, P., Sutherland, R.J., Savage, D.D., 2003. Children with fetal alcohol syndrome are impaired at place learning but not cued-navigation in a virtual Morris water task. Behavioural Brain Research 143, 85–94.

Hamilton, D.A., Rosenfelt, C.S., Whishaw, I.Q., 2004. Sequential control of navigation by locale and taxon cues in the Morris water task. Behavioural Brain Research 154, 385–397.

Hamilton, D.A., Sutherland, R.J., 1999. Blocking in human place learning: evidence from virtual navigation. Psychobiology 27, 453–461.

Hanlon, F.M., Weisend, M.P., Hamilton, D.A., Jones, A.P., Thoma, R.J., Huang, M.X., Martin, K., Yeo, R.A., Miller, G.A., Canive, J.M., 2006. Impairment on the hippocampal-dependent virtual Morris water task in schizophrenia. Schizophrenia Research 87, 67–80.

Harvey, C.D., Collman, F., Dombeck, D.A., Tank, D.W., 2009. Intracellular dynamics of hippocampal place cells during virtual navigation. Nature 461, 941–U196.

Holscher, C., Schnee, A., Dahmen, H., Setia, L., Mallot, H.A., 2005. Rats are able to navigate in virtual environments. Journal of Experimental Biology 208, 561–569.

Hori, E., Nishio, Y., Kazui, K., Umeno, K., Tabuchi, E., Sasaki, K., Endo, S., Ono, T., Nishijo, H., 2005. Place-related neural responses in the monkey hippocampal formation in a virtual space. Hippocampus 15, 991–996.

Hori, E., Tabuchi, E., Matsumura, N., Tamura, R., Eifuku, S., Endo, S., Nishijo, H., Ono, T., 2003. Representation of place by monkey hippocampal neurons in real and virtual translocation. Hippocampus 13, 190–196.

Hüfner, K., Hamilton, D.A., Kalla, R., Stephan, T., Glasauer, S., Ma, J., Bruning, R., Markowitsch, H.J., Labudda, K., Schichor, C., Strupp, M., Brandt, T., 2007. Spatial memory and hippocampal volume in humans with unilateral vestibular deafferentation. Hippocampus 17, 471–485.

Iaria, G., Lanyon, L.J., Fox, C.J., Giaschi, D., Barton, J.J.S., 2008. Navigational skills correlate with hippocampal fractional anisotropy in humans. Hippocampus 18, 335–339.

Iaria, G., Palermo, L., Committeri, G., Barton, J.J.S., 2009. Age differences in the formation and use of cognitive maps. Behavioural Brain Research 196, 187–191.

Iaria, G., Petrides, M., Dagher, A., Pike, B., Bohbot, V.D., 2003. Cognitive strategies dependent on the hippocampus and caudate nucleus in human navigation: Variability and change with practice. Journal of Neuroscience 23, 5945–5952.

Jacobs, J., Kahana, M.J., Ekstrom, A.D., Mollison, M.V., Fried, I., 2010. A sense of direction in human entorhinal cortex. Proceedings of the National Academy of Sciences of the United States of America 107, 6487–6492.

Jacobs, J., Weidemann, C.T., Miller, J.F., Solway, A., Burke, J.F., Wei, X.-X., Suthana, N., Sperling, M.R., Sharan, A.D., Fried, I., Kahana, M.J., 2013. Direct recordings of grid-like neuronal activity in human spatial navigation. Nature Neuroscience 16, 1188–U1140.

Jacobs, W.J., Thomas, K.G.F., Laurance, H.E., Nadel, L., 1998. Place learning in virtual space II. Topographical relations as one dimension of stimulus control. Learning and Motivation 29, 288–308.

Kahana, M.J., Sekuler, R., Caplan, J.B., Kirschen, M., Madsen, J.R., 1999. Human theta oscillations exhibit task dependence during virtual maze navigation. Nature 399, 781–784.

Knierim, J.J., Hamilton, D.A., 2011. Framing spatial cognition: neural representations of proximal and distal frames of reference and their roles in navigation. Physiological Reviews 91, 1245–1279.

Laurance, H.E., Thomas, K.G.F., Newman, M.C., Kaszniak, A.W., Nadel, L., Jacobs, W.J., 2002. Older adults map novel environments but do not place learn: findings from a computerized spatial task. Aging Neuropsychology and Cognition 9, 85–97.

Livingstone-Lee, S.A., Murchison, S., Zeman, P.M., Gandhi, M., van Gerven, D., Stewart, L., Livingston, N.J., Skelton, R.W., 2011. Simple gaze analysis and special design of a virtual Morris water maze provides a new method for differentiating egocentric and allocentric navigational strategy choice. Behavioural Brain Research 225, 117–125.

Maguire, E.A., Frith, C.D., Burgess, N., Donnett, J.G., O'Keefe, J., 1998. Knowing where things are: parahippocampal involvement in encoding object locations in virtual large-scale space. Journal of Cognitive Neuroscience 10, 61–76.

Miyata, H., Fujita, K., 2010. Route selection by pigeons (columba livia) in "traveling salesperson" navigation tasks presented on an LCD screen. Journal of Comparative Psychology 124, 433–446.

Moffat, S.D., Elkins, W., Resnick, S.M., 2006. Age differences in the neural systems supporting human allocentric spatial navigation. Neurobiology of Aging 27, 965–972.

Moffat, S.D., Hampson, E., Hatzipantelis, M., 1998. Navigation in a "virtual" maze: sex differences and correlation with psychometric measures of spatial ability in humans. Evolution and Human Behavior 19, 73–87.

Moffat, S.D., Resnick, S.M., 2002. Effects of age on virtual environment place navigation and allocentric cognitive mapping. Behavioral Neuroscience 116, 851–859.

Morris, R.G.M., 1981. Spatial localisation does not require the presence of local cues. Learning and Motivation 12, 239–260.

Morris, R.G.M., Garrud, P., Rawlins, J.N.P., O'Keefe, J., 1982. Place navigation impaired in rats with hippocampal damage. Nature 297, 681–683.

Mueller, S.C., Jackon, C.P.T., Skelton, R.W., 2008. Sex differences in a virtual water maze: an eyetracking and pupillometry study. Behavioural Brain Research 193, 209–215.

Mueller, S.C., Merke, D.P., Leschek, E.W., Fromm, S., Grillon, C., Cornwell, B.R., Vanryzin, C., Ernst, M., 2011. Grey matter volume correlates with virtual water maze task performance in boys with androgen excess. Neuroscience 197, 225–232.

Mueller, S.C., Temple, V., Cornwell, B., Grillon, C., Pine, D.S., Ernst, M., 2009. Impaired spatial navigation in pediatric anxiety. Journal of Child Psychology and Psychiatry 50, 1227–1234.

O'Keefe, J., Burgess, N., 1999. Theta activity, virtual navigation and the human hippocampus. Trends in Cognitive Sciences 3, 403–406.

O'Keefe, J., Dostrovsky, J., 1971. The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. Brain Research 34, 171–175.

O'Keefe, J., Nadel, L., 1978. The Hippocampus as a Cognitive Map. Oxford University Press, Oxford, England.

O'keefe, J., Recce, M.L., 1993. Phase relationship between hippocampal place units and the EEG theta-rhythm. Hippocampus 3, 317–330.

Piper, B.J., Acevedo, S.F., Edwards, K.R., Curtiss, A.B., McGinnis, G.J., Raber, J., 2011. Age, sex, and handedness differentially contribute to neurospatial function on the Memory Island and novel-image novel-location tests. Physiology & Behavior 103, 513–522.

Redhead, E.S., Hamilton, D.A., 2007. Interaction between locale and taxon strategies in human spatial learning. Learning and Motivation 38, 262–283.

Redhead, E.S., Hamilton, D.A., 2009. Evidence of blocking with geometric cues in a virtual watermaze. Learning and Motivation 40, 15–34.

Redhead, E.S., Hamilton, D.A., Parker, M.O., Chan, W., Allison, C., 2013. Overshadowing of geometric cues by a beacon in a spatial navigation task. Learning & behavior 41, 179–191.

Sandstrom, N.J., Kaufman, J., Huettel, S.A., 1998. Males and females use different distal cues in a virtual environment navigation task. Cognitive Brain Research 6, 351–360.

Sargolini, F., Fyhn, M., Hafting, T., McNaughton, B.L., Witter, M.P., Moser, M.B., Moser, E.I., 2006. Conjunctive representation of position, direction, and velocity in entorhinal cortex. Science 312, 758–762.

Sato, N., Sakata, H., Tanaka, Y., Taira, M., 2004. Navigation in virtual environment by the macaque monkey. Behavioural Brain Research 153, 287–291.

Schmitzer-Torbert, N., 2007. Place and response learning in human virtual navigation: behavioral measures and gender differences. Behavioral Neuroscience 121, 277–290.

Shore, D.I., Stanford, L., MacInnes, W.J., Klein, R.M., Brown, R.E., 2001. Of mice and men: Virtual Hebb–Williams mazes permit comparison of spatial learning across species. Cognitive Affective and Behavioral Neuroscience 1, 83–89.

Skelton, R.W., Bukach, C.M., Laurance, H.E., Thomas, K.G.F., Jacobs, W.J., 2000. Humans with traumatic brain injuries show place-learning deficits in computer-generated virtual space. Journal of Clinical and Experimental Neuropsychology 22, 157–175.

Spiers, H.J., Maguire, E.A., 2007a. A navigational guidance system in the human brain. Hippocampus 17, 618–626.

Spiers, H.J., Maguire, E.A., 2007b. Neural substrates of driving behaviour. NeuroImage 36, 245–255.

Sutherland, R.J., Chew, G.L., Baker, J.C., Linggard, R.C., 1987. Some limitations on the use of distal cues in place navigation by rats. Psychobiology 15, 48–57.

Sutherland, R.J., Kolb, B., Whishaw, I.Q., 1982. Spatial mapping: definitive disruption by hippocampal or frontal cortical damage in the rat. Neuroscience Letters 31, 271–276.

Taube, J.S., Valerio, S., Yoder, R.M., 2013. Is navigation in virtual reality with fMRI really navigation? Journal of Cognitive Neuroscience 25, 1008–1019.

Tempesta, D., Mazza, M., Iaria, G., De Gennaro, L., Ferrara, M., 2012. A specific deficit in spatial memory acquisition in post-traumatic stress disorder and the role of sleep in its consolidation. Hippocampus 22, 1154–1163.

Thibault, G., Pasqualotto, A., Vidal, M., Droulez, J., Berthoz, A., 2013. How does horizontal and vertical navigation influence spatial memory of multifloored environments? Attention Perception & Psychophysics 75, 10–15.

Washburn, D.A., Astur, R.S., 2003. Exploration of virtual mazes by rhesus monkeys (*Macaca mulatta*). Animal Cognition 6, 161–168.

Watrous, A.J., Lee, D.J., Izadi, A., Gurkoff, G.G., Shahlaie, K., Ekstrom, A.D., 2013. A comparative study of human and rat hippocampal low-frequency oscillations during spatial navigation. Hippocampus 23, 656–661.

Wilson, P.N., Alexander, T., 2008. Blocking of spatial learning between enclosure geometry and a local landmark. Journal of Experimental Psychology Learning Memory and Cognition 34, 1369–1376.

Wilson, P.N., Alexander, T., 2010. Enclosure shape influences cue competition effects and goal location learning. Quarterly Journal of Experimental Psychology 63, 1552–1567.

Wilson, P.N., Foreman, N., Stanton, D., Duffy, H., 2004. Memory for targets in a multilevel simulated environment: a comparison between able-bodied and physically disabled children. British Journal of Psychology 95, 325–338.

Woolley, D.G., Vermaercke, B., Op de Beeck, H.P., Wagemans, J., Gantois, I., D'Hooge, R., Swinnen, S.P., Wenderoth, N., 2010. Sex differences in human virtual water maze performance: novel measures reveal the relative contribution of directional responding and spatial knowledge. Behavioural Brain Research 208, 408–414.

Youngstrom, I.A., Strowbridge, B.W., 2012. Visual landmarks facilitate rodent spatial navigation in virtual reality environments. Learn Memory 19, 84–90.

Nazi Law

Hubert Rottleuthner, Free University of Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Following the exposition of two basic principles of the Nazi state and law – the leader principle and the people's community based on racial homogeneity – the development of Nazi law within three dimensions of a legal order is described: legislation in public, penal, and civil law; the legal staff; and the societal legal culture. Nazi law can be characterized by (1) a permanent combination of normal routine and terror culminating in the Holocaust, (2) a mixture of modernity and regression, and (3) morally laden concepts that were used to justify every legal measure and illegal ones as well. Finally, one can ask whether there was a 'legal order' at all during the Nazi era.

It would not be appropriate to look at Nazism only in light of the Holocaust, without considering everyday law and the administration of justice. It is important not to neglect the routine work, the 'normality' of the judicial system with its millions of cases and unpublished decisions. A high degree of normality, of daily routine, is indispensable in such a system in order to secure the conformity of the legal staff, while continuous intervention would lead to a feeling of uncertainty.

The term 'Nazi law' could refer broadly to all legal norms and court decisions that were issued in Germany during the Nazi era between 1933 and 1945. In a narrow sense, Nazi law means only those norms and decisions that can be interpreted as an expression of basic principles of a crude Nazi ideology.

Basic Principles

The basic Nazi principles of the organization of law and state can be characterized by an ambivalent juxtaposition of the leader principle (*Führerprinzip*) and a racial–ethnic substance of the people's community (*Volksgemeinschaft*). On the one hand, we find the 'Führer' as the head of a unitary, authoritarian, and militaristic state whose decisions are law. On the other hand, we have the homogeneous nation of the German people. On the political level, there is no separation of powers, no federal structure, only one party; hence there is no democratic, parliamentary, and pluralistic struggle for power. The leader acts beyond the ligatures of the rule of law. On the societal level, class struggle is replaced by racial integration. Foreign elements – non-Aryans, political enemies (such as communists and social democrats), 'asocial' people, criminals, and so on – are excluded by legal measures, by eugenic procedures, or by the arbitrary power of the police and the *Schutzstaffel* (SS). The common good is authoritatively defined according to the 'sane feeling of the people' (*gesundes Volksempfinden*). The distinction between state and society, the boundaries between public and private, are dissolved by the totalitarian equalization and unification (*Gleichschaltung*) of intermediary organizations such as parties, trade unions, the media, universities, and the arts (only the churches could maintain some kind of relative autonomy). This system of paradoxical 'substantial decisionism' collapsed after its belligerent expansion.

Legislation

On the ideological level of legal philosophers, legislation was understood – and in fact it was a misunderstanding of Savigny's notion of the *Volksgeist* – as the 'expression' of the popular substance, that is, of a homogeneous *Volksgemeinschaft*. But how did the 'law on the books' come into the books in the first place? Legislative products, even during the Nazi era (1933–45), reflected a compromise of conflicting forces. So one has to look behind propaganda and find out how they came into being. There exist only a few studies on the origin of Nazi legislation, for example, in the field of marriage law, labor law, and the law against the misuse of legal advice.

In terms of legislative activity in Nazi Germany, a huge amount of routine work was done after 1933. There were not only the Nuremberg Laws (of 15 September 1935), including the 'law for the protection of German blood and German honor' (*Gesetz zum Schutz des deutschen Blutes und der deutschen Ehre*) and the 'Reich citizenship law' (*Reichsbürgergesetz*), but also regulations concerning chimney sweeps, seeds, and wine corks.

The height of legislative regulation in twentieth century Germany – as far as the number is concerned – was reached between 1938 and 1939 (they were mainly measures to prepare for a war economy); this finding might be of interest for those who talk about today's 'regulation flood.' This vast amount of legislation indicates an instrumental use of norms, not an expression of a *völkische* (racial–ethnic) will as a fundamental principle of Nazi ideology. As far as the forms of regulation are concerned, we find a variety of norms ranging from traditional laws to single orders of the Führer, sometimes circulated via radio broadcasts, and even unpublished ones like the command of September 1939 to start the euthanasia program.

Public Law

After 1933 it was an open question whether the Weimar constitution (drafted in 1919) was still valid. Although it was

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86085-7

never formally invalidated, it was displaced by new 'basic laws,' including the program of the Nazi Party. Basic rights were eliminated through the *Verordnung des Reichspräsidenten zum Schutz von Volk und Staat* of 28 February 1933. There was no separation of powers after the *Ermächtigungsgesetz* of 24 March 1933, which empowered the government or single ministries and other authorities to issue laws. The federal structure of the Weimar Republic was destroyed. Political power was centralized in the Reich authorities with Hitler on top as the Führer and Reichskanzler in one person after 1 August 1934. The will of the Führer was the supreme and unlimited source of law. His will could be manifested in a variety of norms (*Gesetz, Verordnung, Erlaß, Einzelbefehl, Gesamtauftrag*, etc.). Below the level of the Führer there existed no clear competences, but rather a structural chaos that has been labeled a polycratic regime.

Penal Law

In penal law the principle of *nulla poena sine lege* (no punishment without a law) was abandoned by introducing the 'analogous' application of the law according to the *gesundes Volksempfinden*, and replaced by the principle of *nullum crimen sine poena* (no crime without a punishment). Various retroactive laws were issued. The penal system no longer focused on the criminal act defined as an offense against the *Volksgemeinschaft*. Instead the 'types' of the criminal (*Tätertypen*) became important. There was the *Gewohnheitsverbrecher*, the habitual criminal against whom the Nazis introduced the 'three strikes rule' with capital punishment as an obligatory sanction after 1941; the violent criminal (*Gewaltverbrecher*); the murderer as a distinct type (*Mörder*); and the 'people's blight,' a person like a pestilential insect that has to be exterminated (*Volksschädling*). Instrumental use of penal law such as the 'war-time penal code' (*Kriegssonderstrafrechtsverordnung*) (1939), or in its extreme instance the so-called penal code for Poles (*Polenstrafrechtsverordnung*) (1941), became normatively unlimited. It is characterized by extremely vague norms that confer broad judicial discretion, and by a terroristic severity of sanctions.

Civil Law

In civil law there were only minor changes in the legislation, except in agricultural law, family law, and labor law. Using old rules, the judiciary invented new doctrinal formulas that enabled the judges to intervene in the formerly private sphere of autonomous individuals by the interpretation of a contract as a *Gemeinschaftsverhältnis* (relationship of a community, an organic union). Through the 'law ordering national labor' (*Gesetz zur Ordnung der nationalen Arbeit*) of 20 January 1934, the autonomy of trade unions and employer organizations to reach collective agreements was abolished. Strikes were illegal. The collective actors were integrated in the *Deutsche Arbeitsfront* and put under the control of state executors. The worker became a popular figure of Nazi propaganda, while the labor market, including wages and mobility, was intensively regulated long before World War II. In other spheres of the economy, the state favored the interests of factions of big business and their trusts; state intervention increased during the war.

The Legal Staff

During the first years after 1933, court organization was not substantially changed. Special courts (*Sondergerichte*) (1933), the *Volksgerichtshof* (1934), the *Reichskriegsgericht* (1936), and the *Reichsverwaltungsgericht* (1941) were established. Under the new regime, about 15% of judges and prosecutors were dismissed on political and/or racial grounds. For various reasons, the caseload decreased. A substantial change took place in the form of *Verreichlichung*, that is, the centralization of the court administration under the department of justice of the Reich that after 1935 controlled the jurisdiction of the courts. After 1936, the judiciary lost its competence in 'political' issues (mainly matters concerning the secret police). In an agreement, with the ministry of justice in 1937, the Gestapo gained permission to torture (*verschärfte Vernehmung*). Step by step the courts were excluded from prosecuting and sentencing Jews, homosexuals, and other enemies of the German *Volk*. In an agreement between Thierack (Minister of Justice) and Himmler (Chief of the Police and of the SS, respectively) of September 1942, 'asocial elements,' Jews, gypsies, Russians, Ukrainians, and so on, had to be handed over from the prisons to Himmler's myrmidons.

All these measures were executed, accepted, or tolerated by members of the legal staff. A basic problem in dealing with Nazi law is how to explain the degree of conformity among judges, prosecutors, attorneys, and legal scholars. One explanation given after 1945 (Radbruch, 1946) was that they acted on the basis of a 'positivistic' attitude by simply adhering to the law. Instead, it was a positive attitude to the new regime that was hailed by the organizations of judges and prosecutors. Judges were overzealous in their duty, as in dismissing Jews from their legal positions as employees or tenants. Faced with the atrocities against Jews and political opponents, they displayed a passive onlooker's attitude. Courts used racist arguments without being obliged to do so. They imposed excessively severe sentences: estimates of the number of executions, including those imposed by military courts, range from 20 000 to 50 000. There was almost no judicial opposition; instead there was widespread voluntary obedience. The explanation that legal positivism was responsible for the conformity of the legal staff is no longer accepted today.

Legal Culture

There exists very little evidence about 'legal culture,' in the sense of knowledge and opinions about law among the populace, during the Nazi era. Systematic examination of facts that illustrate people's attitudes toward legal norms, judgments, legal institutions, and professional groups does not reveal a homogeneous *Volksgemeinschaft*, but a variety of differing opinions and attitudes. Many findings, such as reports by the *Sicherheitsdienst* of the SS, indicate that a deeply rooted trust in

the courts (and in the Nazi regime in general) persisted until 1945.

Generally, only a small fraction of social problems or conflicts reach courts; criminal prosecution and private invocation of the courts are extremely selective, so it is significant for the Nazi era that a considerable portion of the activities of the Gestapo was instigated and supported by private informers.

Normality and Terror

An essential feature of the Nazi regime was the combination of open terror and everyday routine. Legally 'normal' and terrorist measures alternated, and terrorism occurred under the guise of the 'normal.' Three events exemplify this alternating dance between terrorism and normality:

1. The murders of Röhm (the leader of the *Sturmabteilung* (SA)) and his followers in June and July 1934, which was followed by doctrinal justification (in particular by Carl Schmitt, at that time a professor for public law in Berlin), as an act of government self-defense.
2. The elimination and extermination of persons of so-called inferior quality, starting with the 'law on the prevention of offspring afflicted with a hereditary disease' (*Gesetz zur Verhütung erbkranken Nachwuchses*) of 14 July 1933 (introducing coercive sterilization), followed by the 'law on the protection of the hereditary health of the German people' (*Gesetz zum Schutze der Erbgesundheit des deutschen Volkes*) of 18 October 1935 (preventing marriage in the case of certain diseases), and finally the transition to secret euthanasia actions, authorized by Hitler, started in January 1940.
3. The legal persecution and finally the extinction of Jews, which had been ideologically and administratively prepared by cumulative anti-Semitic legal regulations as well as nonlegal measures.

The mixture of normality and terror, together with strong elements of continuity, helps explain the high degree of conformity among the legal staff as well as among the general populace.

Law and the Holocaust

The legal persecution of Jews advanced step by step, interrupted and radicalized by brutal, direct measures. First came the 'law for the reconstitution of officialdom' (*Gesetz zur Wiederherstellung des Berufsbeamtentums*) of 7 April 1933, excluding Jews – with a few exceptions – from the civil service. The concept 'Jew' was not defined according to criteria of race; instead the religion of the grandparents was adopted as a criterion that could be easily used administratively. Step by legal step, the scope of social activities and the positions available to Jews were reduced. In September 1935 (through the Reich citizenship law) Jews were denied full citizenship. The new crime of '*Rassenschande*' (prohibition of sexual intercourse between Jews and non-Jews) was introduced (in the 'law for the protection of German blood and German honor'). Legal 'aryanization' of Jewish property and exclusion from professions were pursued continuously by the legislators. The yellow star (*Judenstern*) was introduced in September 1941. In December 1941, Poles and Jews in the occupied territories in Eastern Europe were put under special penal regulations, with the threat of capital punishment for almost every kind of disobedience. After the January 1942 Wannsee Conference, at which the plans for the extermination of the Jews were formulated, their further persecution and extinction was by no means carried out extralegally but in a legal–bureaucratic manner. From 1 July 1943, crimes committed by Jews were persecuted and sanctioned by the police only, and no longer by courts. The last published formal directive, of 11 November 1944, barred Jews in Berlin from using heated public rooms.

Apart from this kind of legally sanctioned persecution, there were repeated 'wild' actions, starting with the anti-Semitic excesses in March and April 1933. Such persecutions reached their first climax in the anti-Semitic pogroms in November 1938, followed by a legal sequel (obligation to pay a tribute; collection of insurance benefits). The same pattern holds for the legal exclusion of other victims of Nazism: Sinti, Roma and other *Fremdvölkische*, homosexuals, criminals, the insane, asocials, communists, social democrats, Jehovah's Witnesses, and so forth.

Injustice did not arise all of a sudden, but gradually. The step-by-step character of the persecution of people not belonging to the *Volksgemeinschaft* contributed to the high degree of conformity among the legal staff. By accepting the first steps, officials became unable to resist the further steps, even though those might have seemed unjust.

Hitler and the Law

The remarks about law in Hitler's *Mein Kampf* indicate a purely instrumentalist understanding of law's function in society: law is deprived of any intrinsic value and only conceived as an instrument to promote the Aryan race, to suppress and exclude all kinds of enemies, and to secure the life of the German people. A few deprecatory remarks made by Hitler about jurists are reported. Sometimes he intervened arbitrarily in court cases, most famously in the Schlitt case (in 1942), which he used to have himself declared – for the second time after the Röhm murders – the supreme judge. After 1942 he was no longer concerned with individual cases.

Continuity and Discontinuity, Modernity and Regression

The Nazi era must neither be seen as entirely continuous with or totally isolated from what came before and what happened afterward, that is, its Weimar prehistory and the coming to terms with Nazism after 1945. The moments of continuity include both personnel and substantive factors. Personnel issues include the continued activity – and loyalty – of most lawyers within the judiciary and among legal scholars before and after 1933 as well as (at least in the Western zones of Germany) after 1945. During the last decade, a couple of studies were published concerning the postwar continuity of organizations, state institutions, and big firms that were

involved in the exploitation of forced workers during World War II. Data on career patterns of higher members of the legal staff before and after 1945, revealing a high degree of continuity, were collected and analyzed (Rottleuthner, 2010). Monographs deal with prominent Nazi legal scholars like Carl Schmitt, Ernst Forsthoff, and Karl Larenz, who after 1945 disguised their Nazi past in a cloud of silence.

Substantively, in the legal domain the Nazis and their legal lackeys were not very inventive. With the possible exception of agricultural law – and after 1939 the regulations concerning the war economy – they retained the traditional legal corpus. The vast majority of legal norms remained valid after 1933, but they were given new interpretations. In large measure they continued trends that had been evident before 1933, such as in penal law. With regard to eugenics, too, the Nazis referred back to much older concepts and also claimed to rely on similar practice abroad, for example, in the United States and Scandinavia. In civil law, too, one can find continuity, as in the so-called 'materialization' tendencies in private law, that is, the limiting of contractual autonomy.

In some areas, the Nazi period achieved solutions for old disputes, as in the corporation act (*Aktiengesetz*) of 1937. From a wider historical perspective some reforms can be interpreted as 'progressive': protective regulations concerning pregnancy and motherhood as well as juvenile workers, regulation of working hours, and so on. In divorce law (1938), the grounds for divorce were broadened from the guilt of one spouse to include marriage breakdown (this again was an idea dating from before 1933). An innovation in legal doctrine was E. Forsthoff's concept (introduced in 1938) of *Leistungsverwaltung*, reflecting the change from the traditional intervention state (*Eingriffsverwaltung*) to a state administration that provides, in addition, welfare services and infrastructure.

Justifications

All of the Nazi regime's official legal measures – starting from the *Machtergreifung* between January and March 1933 – were justified by legal scholars, in legal philosophy as well as in legal doctrine. Carl Schmitt, for example, justified the Röhm murders in 1934; Karl Larenz and others provided reasons for legal discrimination against Jews. It was no problem to accept eugenics; the presidents of the regional appeal courts and the chief public prosecutors were informed by the ministry of justice about the euthanasia program in April 1941; in turn, they showed no opposition. For the excessive use of the death penalty, one could refer to the saying that hard times require hard measures.

Nazi law – legislation, court decisions, and legal doctrine as well – was full of moral notions. In all legal domains, notions such as honor (*Ehre*), faith (*Treue*), and grace (*Anstand*) were applied. Likewise morally laden were the negative counterparts such as malice (*Heimtücke*), infamy (*Schande*), inferior motives (*niedrige Beweggründe*), and blighter (*Schädling*) (see Gross (2010) for the morality of the Nazis). Hitler stated in 1933, addressing the *Juristentag* (the traditional public meeting of jurists), that law and morality were no longer separated.

Was It Law?

In general, we may doubt – from a legal–philosophical point of view – whether Nazi law was law at all, whether a legal system existed under the Nazis. It is not necessary to rely on natural law in order to deny the term law to what the Nazis called law, as did Gustav Radbruch after 1945 in the case of extreme injustice. From a rather empirical point of view, Ernst Fraenkel (Fraenkel, 1941) called the regime a 'dual state,' combining a 'normative state' and a 'prerogative state.' The normative state – traditional state agencies, in particular the courts – acted 'according to the law' in dealing with nonpolitical issues; while the prerogative state, the 'organs of dictatorship,' acted 'according to given conditions' in what they defined as political matters. While the normative state contributed to the rational calculability required in a capitalist society, the prerogative state aimed at the oppressive creation of the *Volksgemeinschaft*, the dissolution of the institutions of the working class, and the persecution of political foes. Fraenkel finished his manuscript in 1938 before he left Germany. In American exile, Franz Neumann (Neumann, 1942/44) came to the conclusion that Nazi Germany was a 'non-state.' Conflicts between various power groups – parts of the state administration, party cliques, big business interests, the military forces, and the SS–police complex – resulted in a chaos of competences and a polycentric dissolution of legality as the traditional mode of state organization.

In his pragmatic jurisprudence Lon Fuller (Fuller, 1969: p. 39) describes "eight distinct routes to legal disaster": no rules at all but only ad hoc decisions; failure to publicize; retroactive legislation; failure to make rules understandable (e.g., extremely vague rules); contradictory rules; rules that require conduct beyond the powers of the affected party; too frequent changes in the rules; and no congruence between the rules as announced and their actual administration. According to Fuller, the Nazis did not create a bad system of law; instead, their attempts resulted in "something that is not properly called a legal system at all" (Fuller, 1969: p. 33).

See also: Antisemitism; Compliance and Legal Authority; Dictatorship in History; Dictatorship; Holocaust, The; Injustice: Legal Aspects; Judaism and Gender; Legal Culture and Legal Consciousness; National Socialism and Fascism; Race and the Law; Rule of Law (and Rechtsstaat); Totalitarianism.

Bibliography

Fraenkel, E., 1941. The Dual State: A Contribution to the Theory of Dictatorship. Oxford University Press, New York. (Original German version Der Doppelstaat. In: Fraenkel, E., 1999. Gesammelte Schriften, Band 2. Nomos, Baden-Baden, Germany.)

Fuller, L.L., 1969. The Morality of Law, revised ed. Yale University Press, New Haven, CT.

Gross, R., 2010. Anständig geblieben. Nationalsozialistische Moral. Fischer, Frankfurt am Main, Germany.

Gruchmann, L., 2001. Justiz im Dritten Reich 1933–1940. Anpassung und Unterwerfung in der Ära Gürtner, third ed. Oldenbourg, Munich, Germany.

Hirsch, M., Majer, D., Meinck, J. (Eds.), 1984. Recht, Verwaltung und Justiz im Nationalsozialismus. Bund, Cologne, Germany.

Müller, I., 1987. Furchtbare Juristen. Die unbewältigte Vergangenheit unserer Justiz. Kindler, Munich, Germany. (Translated into English, 1991. Hitler's Justice: The Courts of the Third Reich. Harvard University Press, Boston.)

Neumann, F., 1942/44. Behemoth. The Structure and Practice of National Socialism 1933–1944. Oxford University Press, New York.

Radbruch, G., 1946. Gesetzliches Unrecht und übergesetzliches Recht, Süddeutsche Juristen-Zeitung, pp. 105–108.

Rottleuthner, H., 2010. Karrieren und Kontinuitäten deutscher Justizjuristen vor und nach 1945. Berliner Wissenschaftsverlag, Berlin, Germany.

Rüthers, B., 1989. Entartetes Recht. Beck, Munich, Germany.

Stolleis, M., 1998. The Law under the Swastika. University of Chicago Press, Chicago.

Near Middle East/North Africa Studies: Culture

Miriam Cooke, Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Fiction, drama, filmmaking, and art play a significant role in the Middle East and North Africa (MENA), where intellectuals are often considered spokespersons for the people to the regime. From Morocco to Iran, they have a moral authority that promises influence, imposes responsibility but also invites manipulation by those in power. This article provides a historical examination of the production of culture from the end of the nineteenth century to the revolutions of the twenty-first century. Language reform, war, Palestine, gender justice, Islam, and decoloniality have figured importantly on the MENA cultural scene. During the past 20 years systematic translation of MENA literature and new film festivals featuring MENA cinema have helped to globalize MENA culture.

Culture in postindependence Middle East and North Africa (MENA) is the ensemble of literary and artistic production that reflects the norms and values of daily life. Fiction, drama, filmmaking, and art play a significant role in the public sphere where intellectuals are often viewed as spokespersons for the people to the regime. They have a moral authority that promises influence, imposes responsibility but also invites manipulation by those in power. This article examines the production of culture from the end of the nineteenth century to the revolutions of the twenty-first century.

The MENA have come to be conjoined as the Arab world, with Turkey, Iran, and, more recently, Israel as non-Arab neighbors. Scholars have traditionally treated the region as a separate and more or less isolated block, yet it never was. Framing the southern and eastern shores of the Mediterranean, this area is geographically, historically, and culturally connected with Europe.

In his influential *Orientalism* (1978), Edward Said argues convincingly that eighteenth- and nineteenth-century military and political incursions from the northern rim were accompanied by an equally violent discursive campaign that created an "ontological and epistemological distinction between the 'Orient' and (most of the time) 'the Occident'. [Orientalism was a] Western style for dominating, restructuring and having authority over the Orient" (Said, 1979: 2, 3). Apparently objective orientalist scholarship in most disciplines in fact served the political ends of their governments and shaped domestic practices and policies. Postcolonial scholars have revealed that European portrayals of MENA as primitive, barbaric but ripe for conversion are colonizing projections. Yet orientalists' representations had the force of truth at the time, and they were remarkably influential. Arab, Turkish, and Iranian intellectuals writing during the nineteenth and early twentieth centuries often built on such positivist foundations. 'Progress' became the leitmotiv of the civilizing discourse and missions. Failure to modernize adequately was predicated on the inability to imitate the West.

Nineteenth-Century Nahda

To understand the role of culture and especially literature in the Middle East we must first glance at its history. Although Middle Eastern literatures have become privileged sites for recording and engaging with sociopolitical tensions and conflicts, this was not always the case. In the medieval and early modern periods, Middle Eastern literatures had been the preserve of the elite, providing them with literary frames in which they might elaborate an already familiar tradition and demonstrate formal skills and erudition. Popular literature such as *A Thousand and One Nights* also relied on the already known. Social commentary was generally downplayed.

The arrival in the late eighteenth and nineteenth centuries of the French and British colonial expeditions, with their technological, cultural, and intellectual institutions, forced a new look at society and culture. Contact with European modernity cast Islamic cultures in a poor contrastive light. Egyptians in particular, because they were the first to experience French colonialism, became concerned to learn about Western culture and science. Under such Ottoman governors as Muhammad Ali Pasha and his son Ibrahim Pasha, groups of male Egyptian scholars traveled to Western Europe, and particularly to France, to study scientific texts and to translate them into Arabic. Because of the enormity of the task, these scientific missions might stay for months on end in Europe where they were introduced to theater and modern literature. The newest and most intriguing genre was the novel, and these men returned with translations of European masterpieces by Honoré de Balzac, Gustave Flaubert, Leo Tolstoy, and Ivan Turgenev, to mention only a few. Some intellectuals came to believe that the writing and function of literature in the Middle East had to change to accommodate their new realities. They also knew that conservatives might resist the introduction of new genres, fearing the great classical literatures of the Arabs, Persians, and Turks might not survive the invasion of the trivia of the modern world. Others saw in the dialogue between their own cultural traditions and those of others the possibility of reimagining national identity within a common modernity.

The Question of Language

Whether they were closed or open to the outside world, Middle Eastern intellectuals considered the question of language to be vital in the process of modernization. Some interrogated the boundary between high-culture languages and the vernaculars,

while others considered the implications of importing foreign words. In Iran, the key was to "disassociate from the Islamic past and to project a 'pure' national origin. … The invention of an idealized past was contemporaneous with the restyling of language, which was achieved in a dialogic relationship with Iran's Arab-, European-, and (the often ignored) Indian-Other. … The Persian language was reconstituted as the essential component of Iranian national identity" (Tavakoli-Targhi, 1990: 77, 86). History could not be reconceived without a radical transformation of the language and its grammar.

The Turkish and Arabic linguistic situations were more extreme than the Iranian because of the distance between the written and spoken languages. Indeed, in both cases there had emerged a kind of bilingualism. The written language of the Turks was so different from the spoken that it was even named differently: it was called Ottoman Turkish. Written in the Arabic script and filled with Persian and Arabic vocabulary, Ottoman Turkish was a distinct language with its own rules of grammar (Halman, 1982: 36). Paradoxically, the poetry of the fourteenth-century mystic Yunus Emre, written before the Ottomans' importation of a foreign lexicon, is more accessible to Turks schooled in the post-Atatürk era than the literature they learned from their grandparents. Mustafa Kemal Atatürk's Westernization campaign, erasing Islamic elements in the language and introducing the Latin alphabet, put an end to the 'bilingual' situation that had pertained in Turkey for centuries. There has been no comparable revolution in the Arab world, and so the split between the classical and the many local colloquial Arabic languages persists. The codification of the spoken language is still tentative because the Arabic of the Qur'an retains its place of honor as the literary medium of expression. Without consensus over the codification of the colloquial, some fear a tower of Babel in which multiple, mutually incomprehensible local literatures would vie for a small spot on the grand stage of what had once been a unified literary tradition. The answer has been to create a modern standard Arabic used by the media and understood from the Atlantic Ocean to the Arabian Gulf.

Anxieties about language were compounded by the imposition of colonial languages. French had become the lingua franca among Middle Eastern intellectuals, even in countries like Turkey that the French had not colonized. French, English, and Russian writers in French translation provided models of how to represent current crises and concerns. Translations of nineteenth-century European literature were followed by literary experimentation.

A New Public Sphere

The literary hub of the Arab Middle East in the early twentieth century was Egypt. Intellectuals censored at home moved to Cairo. Literary salons convened men and sometimes even women to discuss the latest developments in politics and literature. At a time when most middle- and upper-class women were confined to their homes, these gatherings of intellectuals were almost the only places where women could appear with men. Literary schools in Egypt, such as the Diwan School, the New School, and the Apollo School (none of which included women among their numbers), encouraged the production of new kinds of writing that were then often published in their own journals and newspapers. These debates and their publication produced a new public sphere. Discussions about national identity and the new roles women and men were expected to play in the future nation became matters concerning everyone and were no longer restricted to a small community of scholars, often religious scholars. In the process, new conceptions of literary criticism appeared.

The new critics' insistence on the centrality of the imagination in works of art and the need for the work to interact with and hopefully to change society influenced a new generation of critics and writers. The notion that criticism should not be descriptive but rather investigative led to some radical challenges. Taha Husayn (1889–1973), the Egyptian doyen of letters and author of dozens of books, including novels, works of philosophy, and his own autobiography, titled *The Days* (1929), led the charge. New literary critical tools opened up old texts and revolutionized literary production.

Simultaneous with this literary activity in Egypt, two revolutions broke out in Turkey and Iran. Young Turks and Iranians brought up on Enlightenment ideals rebelled against autocratic, self-indulgent rule by the last rulers of the Ottoman and the Qajar Empires. Between 1919 and 1922 Mustafa Kemal (1881–1938, known as Atatürk after 1933) waged war against the British, French, Italians, and Greeks. Expulsion of foreigners was followed by a series of revolutions that continued beyond the 1923 founding of the Turkish republic. After abolishing the Ottoman caliphate in 1924, Atatürk implemented his secularization program and introduced a new Europeanized dress code.

In 1905, Iran was rocked by its constitutional revolution. The period was rich with new kinds of writing. These liberal and often socialist writings depict the turmoil of the period and react against the social irrelevance of the Qajar writers. After his coup in 1921, Mohammad Reza Shah became absolute monarch. He introduced strict censorship and centralized control, but he also concerned himself with the position of women. In 1936, perhaps in response to Atatürk's recent prohibition on the wearing of the veil, the shah enforced unveiling.

Such legislation against veiling should be seen against the background of late nineteenth-century debates about feminism going on throughout the Middle East and beyond. Claims on behalf of women's education were entering the public sphere. Research by feminist scholars both in the Middle East and elsewhere has revealed that it was women themselves who were the first to raise the issue of women's absence in decision-making positions at the national or local levels. Yet it was not until the men took up the 'woman question' that it became a common concern. Atatürk's fascination with the West, where feminists were beginning to attract attention, had predisposed him to advocate women's rights. After coming to power, he gave them the vote.

During the 1930s, some writers began to look for an indigenous, premodern greatness to rival the cultures and power of the Europeans. Marginalizing Islamic influences, intellectuals turned in the 1930s and 1940s to the Pharaonic, Babylonian, Ugaritic, ancient Persian, and pre-Ottoman civilizations. Arab critical schools such as Diwan (1920s) and Apollo (1930s) brought together poets and critics schooled in

England and France who saw the need to knit Western theory into local literary traditions. Discussions about national identity and the new roles women and men should play in the future became matters of general concern.

Independence Struggles

During this same period, the European Zionist movement was becoming active and more Jews emigrated from Europe to settle the land of Palestine. Zionism at that time was primarily a socialist ideology owing more to Russian-inspired communism than to Judaism. As such it looked much like the *missions civilisatrices* that the French and the British had been dispatching to the Middle East since the eighteenth century. In 1948, when most MENA countries had either achieved their independence or were close to it, Palestine was recolonized and transformed into the home of Jews flooding into the region out of fascist Europe.

The region was in turmoil, and MENA intellectuals began to address the effects of rapid change. Saudi exile Abd al-Rahman Munif (1933–2004) situated his five-volume novel *Cities of Salt* (1980s) in a recently oil-rich desert state. As a petroleum economist, he understood the environmental and demographic impact of oil exploitation. The reader witnesses the growing brutalization of the bedouin as they recognize that in this new world of competition and profit, there is more to be gained by individual enterprise and greed than in tribal solidarity. The Iranian writer Al-e Ahmad (1923–69), son of a Muslim Shiite cleric, coined the term 'westoxification' that became a slogan used against the Shah. For him, as for many of his contemporaries, the cultural, economic, and political subordination of MENA countries to the West was anathema. The future had to be modern but not westernized. It was this antioccidentalism that led to the Islamic Revolution of 1978. Scholarship of this period has emphasized growing self-confidence and identification with movements opposed to EuroAmerican neocolonialism.

War Stories

Conflict and violence are so much a part of MENA history and culture that no consideration of its culture can afford to ignore them. The Palestinian struggle against Israeli hegemony since 1948, the Algerian War of Independence (1954–62), the various Arab–Israeli wars but especially the 1967 war, the Lebanese Civil War (1975–92), the Iranian Revolution (1978–79), the Iran–Iraq War (1980–88), the Gulf War (1991), the 2003 US invasion of Iraq, and the 2009–12 uprisings in Iran, Tunisia, Egypt, Libya, Yemen, and Syria are the best known of the MENA conflagrations. Each war produced its own coherent body of literature with women also actively engaged.

After the establishment of the Israeli state in 1948, Palestinians wrestled with their new loss of independence. Although many left, a few, like poet Mahmud Darwish (1942–2008), did stay behind what came to be called the Green Line. Darwish's family had left in the first exodus but soon stole back in. They returned after the period for registering Palestinians had passed, and so the poet's writings dwell on the lack of identity papers. His poem 'Identity Card' (1965) became emblematic of the Palestinian condition. Founder of the binational Israeli Communist Party, Emile Habibi (1921–96) won prizes for his 1974 *The Secret Life of Saeed, the Ill-fated Pessoptimist: A Palestinian Who Became a Citizen of Israel* that tells the picaresque tale of a hapless Palestinian stumbling through life in Israel. Two years later, from Nablus inside the Occupied Territories, Sahar Khalifa (b. 1941) published the first of three novels on the experience of Israeli occupation. Her heroines demonstrate how women's ways of fighting are more effective than those of the men. In the post-1967 period, some Palestinians within Israel wrote in Hebrew. Anton Shammas (b. 1941) wrote *Arabesques*, an autobiographical novel that has been hailed by some as a masterpiece of Hebrew prose. Yet despite the critical acclaim, the novel has not been fully accepted as part of the canon of Hebrew literature. Shammas has described this work as his challenge to the Israeli claim that Palestinians living under Israeli rule could be integrated as Israeli citizens. No matter how assimilated they tried to become in Israeli society, no matter how excellent the Palestinians' command of Hebrew, they could never become Israeli citizens in the full sense of the word. Shammas was long in conversation with A.B. Yehoshua (b. 1936) considered one of Israel's most important fiction writers and social critics. His fiction unveils the Arab in the Jew. In his first novel *The Lover* (1977), he points to the dreaded other that is always there, lying in wait for the opportunity to reveal itself. The figure of the Arab lies at the core of Israeli identity, disturbing the unmarked Europeaness of Israeli identity. The more it is repressed the more violent its ultimate eruption into consciousness. Iraqi Israeli Sasson Somekh (b. 1935) narrates the challenges faced by Jews in Arab countries who lived in the interstices between local and colonial systems that both privileged them as Jews and emphasized their outsider status.

Women's versions of their experiences of national violence have challenged the familiar frame narrative that characterizes men's war stories. They have insisted on the importance of their contributions to the national welfare. In the 1980s, the Algerian novelist and historian Assia Djebar (b. 1938) started to publish her semiautobiographical quartet. Through the eyes of women she radically revises Algerian colonial history, linking her own experience with the stories of neglected nineteenth-century women resisters with those of living women who had been active and then silenced during the war of independence from the French. The 1990s attacks by Islamic fundamentalists on intellectuals, and particularly women, in Algiers perpetuate a war that was not resolved in the early 1960s.

In Lebanon the Beirut Decentrists, a school of women writers, wrote about the civil war. Some analysts have argued that political anarchy relaxed social norms so that the prohibition of women writing about war, an experience they have always been said not to have had, was lifted. Beirut became emblematic of postmodern wars that erase the boundaries between binaries, like home and front, which have been so necessary to the effective telling of the War Story (Cooke, 1997). Critical of participants in the senseless fighting, they also exposed the literary recklessness of those who wanted to represent the war as having any moral sense. Emily Nasrallah (b. 1938) wrote about the devastating effects of the civil war and the massive emigration it precipitated. Hanan al-Shaykh's (b. 1941) *The Story of Zahra* (1980) traces the growth in consciousness of a woman whom the war empowers.

In 1980, while the war in Lebanon was at its height, the Iran–Iraq War broke out. Scholars from the region as well as those from outside have pointed out that Saddam Hussein, like many dictators, has had an ambivalent relationship with the intellectual elite of his country. He needed them and therefore he feared them. Throughout the war, he coerced the writers and artists to glorify the war in ink and paint. The Ministry of Culture established literary series and organized festivals in honor of the war to which scholars of Arabic literature from all over the world were invited. Many Iraqi writers and artists did as they were told. Some did not. Under the very eye of the censor, some writers managed to articulate criticism of a war they were supposed to praise (Cooke, 1997). Not only writers and artists, but literary critics also were commissioned to describe and analyze the cultural effects and products of the war. So prolific were they that by the late 1980s books of criticism about the literary criticism of war literature were being published.

Ironically, the Islamic government of Iran was doing exactly the same as its enemy, namely sponsoring praise of the war and censuring whatever was seen to be oppositional to its Islamic interests. Some writers complied with the government's will; others protested. A. Rahmani's 'Short Hike' acts out a moment of defiance between an Islamic guard and the woman he has reprimanded for revealing a strand of hair. Restrictions forced women to speak out (Milani, 1992). Many fled to London, Paris, Los Angeles, and New York that have become MENA exile capitals. Observing their native lands from a distance, creative artists and critics publish works of political, social, and cultural analysis as well as fiction and memoirs.

In March 2003, a year and a half after the 9/11 attacks on the World Trade Center in New York and the Pentagon in Washington DC, the US military invaded Iraq. The pretext was Iraq's alleged complicity in 9/11 and the goal was to eliminate the weapons of mass destruction that until today have not been found. After toppling Saddam Hussein, the American forces stayed another 8 years. Sectarian violence spun out of control. Young bloggers were the first to inform the world about life in Iraq under American occupation. Two have achieved renown: Salampax and Riverbend. Both of them have published books of their blogs; indeed, this melding of blog and literary text has become a trend in contemporary Arabic literature. In 2007, Sinan Antoon published his *I'jam: An Iraqi Rhapsody*, an imagined prison memoir that recollects the anxiety and fears that ruled Iraqis' lives since the war with Iran. The US military pulled out in 2012 having overseen the destruction of the country.

Writers, filmmakers, and artists, especially grafitti artists, were active in the Arab uprisings of 2010–12. Like the Tunisian muralists, the Egyptian Ganzeer painted his commentary on to Cairo's walls and inspired the young people to continue their stand against the government. Short story writers and filmmakers collaborated to produce films like '18 Days' that within months of the resignation of President Mubarak were making the rounds of international film festivals.

Political Commitment and Prison

Political independence in MENA countries demanded political and moral vision that would owe as little as possible to European norms. Intellectuals set out to 'revolutionize the revolutions'. In 1946, the Iran-Soviet Society sponsored the first congress of Iranian writers, which called for the centrality of ideology to literature. Seven years later in Lebanon, the first editorial of the influential monthly periodical *Al-Adab* declared that all forms of writing must be politically engaged. The time of art for art's sake had passed to be replaced by a revolutionary ethic in all creative activities.

The project of cultural production from inside authoritarian regimes is fraught with danger: just how much criticism is too much, how much too little? Nevertheless, despite the dire consequences for articulating opposition, public intellectuals rarely flinch from what they consider to be their responsibility. Some are killed for their words. Many are imprisoned. So many have published books about their prison experiences that it is possible to talk about a body of MENA prison literature that is beginning to receive the critical attention it deserves. The fact that intellectuals have had such experiences that they then publish them gives them a moral authority others may never achieve.

For many MENA intellectuals prison is a badge of honor, exile a terrible necessity, cooptation an ever-threatening possibility. Public intellectuals in many MENA countries are closely watched. Since books, plays, paintings, and films may assume the force of a political manifesto, governments as well as some nongovernmental associations try to monitor cultural production before, during, and after its completion. Intellectuals judged to be dissident may be eliminated, exiled, or coopted. Cooptation is a widespread phenomenon in MENA, making arduous the decision to stay in one's country. Filmmaking is particularly susceptible. Like theater, film watching is a communal experience and as such it allows individuals to share with an anonymous group a sense of justified anger, or to laugh at the expense of oppressors. The power of such cathartic moments has not been lost on repressive regimes: national film institutes sponsor, censor, and retain monopoly of the finished product. Since the Islamic Revolution, Iranian filmmakers, like the internationally recognized Mohsen Makhmalbaf, have had to submit their scripts, rushes and final films to the government censors, who are especially concerned with Islamizing the representation of women. During the latter part of Hafiz Assad's rule, Syrian filmmakers were sponsored to produce films that were remarkably critical of the regime yet rarely shown. Films like Muhammad Malas' *The Night* (1991) are only shown during international film festivals. People know of their existence and long to watch them. In his *Historical Miniatures* (1993), the Syrian playwright Saadallah Wannus (1941–98) examines the triangular relationship among state, society, and public intellectuals through the controversial meeting of the fourteenth-century Mongol tyrant Tamerlane with the North African historian Ibn Khaldun. Mamduh 'Adwan's (1941–2005) play *The Ghoul* (1996) focuses on the cruelty of the last Ottoman governor sent to Syria. He warns at the end of the play that if the tyrannical system is not crushed it will recur, and it has with the rise of Bashar Asad. The 2011–12 mass revolt of the Syrians against the regime seems to respond to Adwan's call to the people not to cut off the hand of the jailor but to destroy the prison. The viciousness of the government's response to the people's initially peaceful requests for reform has shocked the world.

Exile

For some the cost of staying at home is too much to bear, and so they leave. Citizens of repressive regimes flee so as to find a place where they might breathe and speak freely. They have congregated in major cities like Paris, London, Buenos Aires, New York, and Los Angeles, where they have founded publishing houses and launched newspapers and magazines.

The Palestinian case has become emblematic of the dilemma of many MENA intellectuals: they cannot live at home, they scarcely survive outside. Should the committed intellectual fight or write or do both or neither? In 1963 Ghassan Kanafani (1936–72) published his *Men in the Sun*, which was turned into a film entitled *The Dupes*. This novella explores the possible outcome of the decision to leave one's home and to seek refuge in Arab countries. The fugitives suffer multiple humiliations before they perish on their way to Kuwait.

Those who write in exile often feel farther from home than the miles that mark that distance. Yet in a world whose borders are increasingly porous, where migration is a necessary part of many people's lives, a new kind of cultural nationalism enhanced by social networking is emerging. It allows those who are living far from their land of birth or ancestry to retain very real ties of identity. Turkish novelist Habib Bektas (b. 1951) is a good example. Since 1972 he has been living and writing in Germany. In 1997, his book *The Smell of the Shadow* won the Turkish Inkilab Kitabevi Press annual prize; it was the first time this coveted prize had been awarded to a nonresident Turk. Bektas explained that his Turkishness did not consist in residence in Turkey but in language: he lives in Turkish. This adoption of the language as nation is a strategy many writers in exile invoke.

Islam and Revolution

Religion has come to play an increasing role in MENA societies and culture. The success of the Shiite Islamic Revolution in 1979 Iran marked a radical turning point in regional politics with Muslim ideologues leading the charge. Scholars, especially Western-trained anthropologists, have asserted that the wide distribution of cassette tapes of charismatic preachers' Friday sermons provide even the illiterate with a new vocabulary and syntax. Whether the Islamization of the public sphere is regarded as a good or a bad thing, it has created a new kind of public sociability that entails a mandate for public intellectuals, even the most secular of them, to learn its language. The attack on the Twin Towers in New York City on 9 September 2001 announced to the world the founding of al-Qaeda, an Islamist terrorist organization under the leadership of the Yemeni Saudi Osama Bin Ladin committed to opposing Western capitalism and ungodliness in the world.

Public intellectuals are under constant surveillance by the government and vigilante groups. The case of Salman Rushdie and Ayatollah Khomein's 1989 *fatwa* against him for his 'heretical' *Satanic Verses* (1988) has entered the annals of world history. In the mid-1990s Algeria, the Islamic Salvation Front was voted into office by the people. When the government canceled the election, the Islamists turned their anger on intellectuals. Films have been made on the senseless killings.

Moral censure of literature has escalated to sanction for deadly assaults on the authors. In 1992, Egyptian Islamists assassinated Farag Foda, a journalist, for his critique of their discourse, especially in his *Before the Fall* (1992). That summer, they published in paperback format the proceedings of a trial of the Egyptian writer, physician, and activist Nawal El Saadawi (b. 1931) they had conducted, titled *Nawal El Saadawi in the Witness Stand*. With its cover picture of a wild, white-haired woman staring out from behind bars, the book was so widely distributed that it could be found on the blankets of Cairo sidewalk book vendors, selling for a few piasters. She had called for the 'Lifting of the Veil from the Mind' in her Arab Women's Solidarity Association (founded 1982), in her lectures given all over the world, and in novels like *The Fall of the Imam* (1987). For that they condemned her to death.

Feminists from within the region are analyzing the situation, forming contingent coalitions, and proposing new ways to challenge the policing of women's actions, behavior, and speech. They have recognized the dangers inherent in Islamists' declarations that women are the symbolic and moral center of their society. What they do and how they act have become men's affairs. It is within this context that the veil, abandoned or even banned in most places during the first half of the twentieth century, has reassumed such central importance. In places like Afghanistan, Iran, Algeria, and Sudan, the unveiled woman may be reflexively labeled wanton. Although the veil is not Islamic, it is seen by many to be highly recommended. There are several problems with this highlighting of the veil as an emblem of the society's moral rectitude. Feminists in the late nineteenth and early twentieth century fought against the veil as a symbol of men's ownership of women. To go back to wearing this patriarchally loaded symbol is to renounce the gains of their grandmothers. Another problem with the blanket mandate that women veil is that not all women in MENA countries are Muslims. At the other end of the political spectrum are Islamist women like Zaynab al-Ghazali (b. 1912) whose prison memoirs, *Days from My Life* (1977), enact a model of women's empowerment within orthodox Islam (see Cooke, 2000).

Translation and Recognition

MENA literature, art, and films are now enjoying international recognition, as they enter the mainstream of world culture. Three writers have won the Nobel Prize in literature: the Israeli Shmuel Yosef Agnon (1888–1970) in 1966; the Egyptian Naguib Mahfouz 1988; the Turkish Orhan Pamuk in 2006. In 1989, the Moroccan Tahar Benjelloun won the Prix Goncourt, the first Arab to be awarded this coveted French prize. In 1997, the Turk Yashar Kemal was awarded Italy's most prestigious prize, the Nonino Literary Award. In that same year four Iranian films were awarded major prizes at Cannes, Locarno, Tokyo, and Montreal. Many countries host annual film festivals with latest additions being Dubai, Abu Dhabi, and Qatar.

Arabic, Hebrew, Persian, and Turkish fiction and poetry are being translated into European languages. In 1981, the Palestinian poet and literary critic Salma al-Khadra' al-Jayyusi founded her Project for Translation from Arabic in the United States. Whereas the first translations found homes in marginal presses dedicated to publishing works by writers from the

global South, by the late 1980s major European, American, and even Japanese publishing houses started to adopt literary series as well as individual authors. The growth of the translation industry is significant for readers, who now have unprecedented access to the cultural imaginary of people they had previously known through the stories of scholars and travelers only. It is important above all for the intellectuals who are writing with the awareness that their works will be translated, that what they say about their culture is being consumed far from the local readership they had originally targeted. Writing thus entails consequences that must be anticipated. MENA intellectuals are seeing their books and films becoming part of a global cultural project in which they should play an increasingly visible role.

See also: Education in the Middle East and North Africa; Islam: Middle East; Middle East, North Africa: Sociocultural Overview; Middle Eastern and North African Legal Traditions; Nationalism, Historical Aspects of: Arab World; Near Middle East/North Africa Studies: Gender; Near Middle East/North Africa Studies: Politics; Near Middle East/North Africa Studies: Religion; Near Middle East/North Africa Studies: Society and History; New Media and Democracy in the Arab World; New Media, Political Mobilization, and the Arab Spring; Sexual Politics in the Arab World.

Bibliography

Cooke, Miriam, 1997. Women and the War Story. California University Press, Berkeley.

Cooke, Miriam, 2000. Women Claim Islam: Creating Islamic Feminism through Literature. Routledge, New York.

Cooke, Miriam, 2007. Dissident Syria: Making Oppositional Arts Official Durham. Duke University Press, Durham.

Halman, Talat Sait (Ed.), 1976. Modern Turkish Drama: An Anthology of Plays in Translation. Bibliotheca Islamica, Minneapolis.

Halman, Talat Sait (Ed.), 1982. Contemporary Turkish Literature. Associated University Presses, London.

Milani, Farzaneh, 1992. Veils and Words. The Emerging Voices of Iranian Women Writers. Syracuse University Press, Syracuse.

Tavakoli-Targhi, Mohamad, 1990. Refashioning Iran: language and culture during the constitutional revolution. Iranian Studies 23, 1–4.

Talattof, Kamran, 2000. The Politics of Writing in Iran: A History of Modern Persian Literature. Syracuse University Press, Syracuse.

Yudkin, Leon Israel, 1984. 1948 and after: Aspects of Israeli Literature. University of Manchester Press, Manchester.

Near Middle East/North Africa Studies: Gender

Nilüfer Göle, Ecole des Hautes Etudes en Sciences Sociales, Paris, France

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 10435–10439, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The question of women in the Middle East is not a secondary but an agenda-setting issue for the broader historical, political, and cultural processes that have shaped these societies. Women's status and agency were central in the shaping of movements of modernization, state-building processes, nationalist ideology, and Islamist protest movements. Women's visibility and participation in public life and emerging gender consciousness can be distinguished as a motor of change, having an impact on not only the lives of women, changing male–female subjectivities, but also on the course of Middle Eastern history. In a similar way that modernist experience is transmitted by unveiled bodies and secularized ways of life, Islamist identity and difference are embodied in women's veiling and disciplining the religious self. Women's position and agency in society are decisive in shaping the interplay between Islam and modernity: new definitions of self, modern, and Muslim are emerging from this conflictual but creative encounter. Understanding the distinct trajectories and identity formations that Muslim women follow, thereby changing the interpretations and interactions between religion, traditions, and modernity can provide us with new intellectual tools for a critical reflection on the general field of gender studies.

Introduction

The study of women and gender has opened up a new perspective in our approaches to Middle Eastern history and society, in particular interactions between women's issues and such major cultural and political movements as national independence, social reform, and modernization movements in the nineteenth century and Islamist movements revived by the advent of Islamic Revolution in Iran in 1979. It observes changing approaches to the study of women and history in the Middle East, and their implications on gender studies in general.

'Gender' studies began to appear in the 1980s, first in the United States and soon afterward in Europe and parts of the Middle East, in the wake of the new discipline of women's studies, which was itself an academic and intellectual outcome of the feminist movements of the 1960s and 1970s. It was a new attempt to reveal unspoken inequalities and power relations between men and women. In contrast to the biological differences highlighted by reference to 'sexes,' the term 'gender' depicts socially constructed, culturally specific forms of masculinity and femininity that are subject to social interpretation, negotiation, and change. Furthermore, gender is used, not exclusively in relation to women's issues, but as a constitutive element of all aspects of human society and organizations, such as the family, economy, science, medicine, state, and politics. The analysis of the ways in which different aspects of social life, history, and culture are gendered, namely embedded by culturally perceived differences within a field of power relations, has become a major concern for social sciences. Few concepts have transcended cultural and geographical boundaries so powerfully during the 1980s and 1990s as gender. This has contributed to pioneering work on gender within the field of Middle Eastern history (Keddie and Baron, 1991). Gender awareness has changed ways of studying the history of colonization, state building, nationalism, and Islamic identity in the Middle East. At the same time, tracking the distinct trajectories and identity formations followed by Middle Eastern women, both Muslim and non-Muslim, renders a new map for critical reflection in gender studies. As concepts travel in two directions, they create new avenues of learning, dialogue, and interaction between the West and other cultures, ranging from academy and law to citizenship. This new dialogical concern revitalizes an interest in the study of local cultures and in cross-cultural comparative perspectives.

For at least a century, the problematic link between modernization and cultural identity in the Middle East put women at the heart of political and societal debates. Discussions on the appropriate position for women in Muslim countries continue to underpin any political discourse aiming at the reorganization of these societies. Women are situated at the crossroads between two distinct sets of values and affinities, namely between universalism, secularism, and feminism on the one hand, and cultural authenticity, religion, and Islam on the other. Or, to put it succinctly, in Middle Eastern societies, gender issues are embedded in civilizational issues in the sense that women's position in society is the determining factor both in framing and intertwining what were posited as the existing dualities, such as traditional/modern, religion/secularism, Islam/the West, equality/difference, and interior/exterior. Studies on women in the Middle East not only follow common and overlapping trends with Western feminist thought but also reflect the distinctive centrality of gender issues in Middle Eastern societies.

Women's Civilizing Mission

The social position of women began to be discussed with the onset of modernization attempts in the nineteenth century. As modernization seemed to imply Westernization, the ultimate question that reformists needed to answer was whether Islam was compatible with Western civilization and more particularly with women's rights and freedom. Different answers to that question – ranging from national modernists and Islamist

reformists to religious conservatives – continue to put their stamp on the intellectual and political history of the Middle East. Analysis of these historical debates from the point of view of women's questions has shown discursive parallels in countries such as Egypt, Turkey, and Iran (Ahmed, 1992; Badran, 1995; Gole, 1996; Najmabadi in Kandiyoti, 1991). For progressive nationalists, the emancipation of women from 'erroneous' religious prescriptions and traditional ways of life was the prerequisite for national progress. Islamist reformists asserted that Islam was not a barrier to progress and that it was possible in Islam to embrace modern ideals, such as gender equality and human rights. But for conservatives, any change in the direction of Western modernity, especially pertaining to women's position in society, would have meant a loss in cultural identity and consequently, to preserve moral integrity and social cohesion, gender relationships were to be framed according to the Islamic religious law (*Shari'a*) as they understood it.

As markers of cultural identity, women, therefore, set the agenda for different and competing social projects. Furthermore, the most disputed issues throughout the nineteenth and twentieth centuries, namely polygamy, veiling, gender segregation, women's education, employment, and participation in public life, remained on the public agenda of Middle Eastern countries in the 1980s and 1990s.

The Middle Eastern model of early modernization, namely the era of colonial, postcolonial, and nationalist state building, can best be summarized by a formula that equates women's emancipation and national progress. With striking similarities between Turkey, Iran, and Egypt, and also with the North African countries, women's public visibility and gender equality are considered to be the yardsticks of progress and civility. Turkish Kemalist nationalism articulated and implemented gender equality as an intrinsic part of its modern identity and societal reforms. Women gained full political rights in 1934, and the removal of family law from the jurisdiction of religion with the enactment in 1926 of a secular Civil Code (adapted from Swiss law) in Turkey can be considered as the most radical example of engagement in secular feminism. In Iran, too, women's rights, free education, and compulsory unveiling in 1936 were implemented by the secular nationalist politics of Reza Shah. In Egypt, alongside nationalism, the Islamic reformist movement (especially in the works of Muhammed Abduh and Qasim Amin) paved the way for women to gain access to public life, to claim a public voice, and to demand educational, work, and political rights. Veiling was criticized as a sign of women's confinement and as an obstacle to women's participation in public life. Thus, the removal of the veil and the head covering of women in public became a symbol of women's emancipation as well as national progress. A well-known Egyptian feminist, Huda Sharawi, announced the start of the feminist movement when, returning from an international feminist meeting in Rome in 1923, she disembarked from her train in Cairo and unveiled her face in public. Therefore, public visibility of women was put forward by the first wave of feminist activism and writing in the Middle East. A pre-Islamic past, whether Anatolian, Zoroastrian, Pharaonic, or Berber, was idealized as egalitarian in gender relations and celebrated in the construction of a new nationalist idiom. Women, especially upper and middle class women, mostly coming from a traditional universe but exposed to modern ideals and education, were called to participate in this nation-building process, as citizens, educators, and role models for a 'new life' in the family as well as in the public sphere. The limits of women's participation in public life and social agency were framed within the terms set by the nationalist movements whether backed up by secular or religious reformist discourse. Feminist scholarship in the Middle East revealed these issues pertaining to gender and nationalism and developed in distancing itself from and in criticism of nationalist idiom (Kandiyoti, 1991; Moghadam, 1994).

Women and Islamist Movements

The revival of Islamist movements marks the second turning point in contemporary Middle Eastern history where women's issues and agencies once again play a central role, but this time as markers of Islamic differences rather than as transmitters of Western-centered modernity.

Although Islamic movements are far from being monolithic, one can still discern common features in the contemporary Islamist movements that have developed since the end of the 1970s in almost all Muslim countries. Their aim is to transform the whole of society along Islamic lines, ranging from government, law, and science to ways of life, gender relations, and faith. The Iranian Islamic revolution in 1979 provided the political source of reference and legitimacy for such an ideal. Actors of Islam, men and women alike, come mostly from provincial towns and lower social classes, but are an outcome of urban migration and modern education. The pressures generated by the encounter with a modern universe create a need to reinterpret religion and re-adapt gender roles and divisions. A call for a return to the original, unadulterated, fundamental sources of Islam (the Qur'an and the Sayings of the Prophet) has engendered the revival of debates on Islamic knowledge. The interpretation of religious precepts and texts is opened up to public debate in which not only the religious specialists (*ulema*), but also new Islamist figures, both men and women, participate. In a paradoxical way, contemporary Islamism has instigated the democratization of religious knowledge whereby various actors of Islamism can lay a claim on the interpretation of religion and Islamic jurisprudence. Issues related to the segregation of, and discrimination against, women, such as compulsory veiling (*hijab*); religious marriages and polygamy; women's obedience and legal subordination to the husband's authority; inequalities in divorce, custody, or inheritance; and interdictions put on Muslim women's access to religious positions, praying in mosques, marriage to non-Muslim men are no longer issues settled under the monopoly of religious *ulema* but are the subjects of public controversy. Fatima Mernissi has made a pioneering contribution to understanding the specificities of Islamic constructions of female sexuality and spatial social organization, and giving voice to religious women figures in public life and history (Mernissi, 1975). Other academic work followed and contributed to new readings of women's position in religious and legal texts (Stowasser, 1994; Yamani, 1996; Mayer, 1999).

In its criticism of essentialism and reflections on the constructed nature of identity and difference politics, feminist

theory has contributed to the development of new approaches to Islamic identity. Islamist identity is distinct from Muslim identity to the extent that it expresses a refutation of given definitions, accepted ways of 'being a Muslim,' 'born being Muslim' categories, handed down from generation to generation, and asserts a collective and countercultural attitude. Islamism expresses the problematization of Muslim identity both by a return to the origins of religion and by a non-assimilative attitude toward Western modernity. Islamism shares a similar pattern of collective and personal empowerment with other contemporary social movements in the West, namely that of appropriation of their difference (Muslim, woman, or black), the very difference which accounted for their exclusion and stigmatization. Islamism is a collective manifestation of an exacerbated form of Muslim difference. A search for an alternative in Islam endows especially newly urbanized and educated Muslim actors with a very strong quest for an authentic indigenous identity. The covering of women, the 'veiling,' is the most visible symbol of this claim for Islamic authenticity and difference.

In all Muslim societies, the Islamist movements gained public visibility through the veiling of women. On the one hand, the veil recalls the traditional definitions of Muslim womanhood based on values of modesty and virtue, the segregation of sexes, and the interdictions put on women with respect to their participation in social and public life. But, on the other hand, the new veiling signifies the political participation and the active voluntary reappropriation of Islam identity. As such, the new veiling does not represent the traditional images of Muslim women: uneducated, docile, passive, and devoted to family life. On the contrary, young, urban, educated groups of Muslim girls are politically demanding and publicly visible. The way they readopt the Islamic outfit is different also from its traditional outlook.

Traditional Islamic ways of covering change from one region to another, manifesting diverse forms in different national settings (the Iranian *chador*; the 'kerchief' covering the lower part of the face in North Africa; the face mask of the Gulf region; and the Afghan *burqa*). In contrast, contemporary Islamic dress is a new style, almost like a uniform, independent of the national context. Although there is a wide range of styles in Islamic dress, it can be distinguished by a headscarf totally covering the hair and falling to the shoulders, combined with a sort of long and ample overcoat (made of pale-colored industrial fabric) to hide the forms of the body. Overall, the Islamic outfit is a modern reappropriation and reinterpretation of Islamic modesty; the silhouette of women with widened shoulders, carrying big bags, walking in the streets, taking buses, going to the universities, hints at the emergence of a new profile of Muslim women: that of urban, combative women pursuing a career and mixing with men. In short, women are actors of Islamist movements, but they are also following a multifold life and professional strategies, and making part of a modern universe. And this, of course, cannot avoid creating a tension between Islamic precepts, which define the 'ideal Muslim woman' as a homemaking, self-sacrificing mother and wife, and the new individual and professional aspirations of these women. Acquired public visibility, social mixing with men, and intellectual and professional aspirations force them into a tension between their collective role as agents of Islam, as markers of 'authentic Islamic identity,' and their emerging individualist and feminist consciousness (Gole, 1996). Paradoxical it may appear, in promoting the veil as the symbol of their cultural difference, but Islamic movements unintentionally encourage women's participation and visibility in public life, which, in turn, constitutes a subversive potential for gender consciousness and activism.

Women voice publicly their demands for access to education, employment, and political life (Adelkhah, 1991). In the post-Khomeini Iran, women's periodicals proliferate in which women defend gender equality in their readings of religious texts and criticize Islamic jurisprudence. They argue for women's rights such as their eligibility to engage in religious studies, to hold juridical positions, to participate in the lawmaking process, to stand for public office, to have custody of children after divorce, to have equal consideration with men in testimony. In these periodicals, Iranian women argue for a women's rights perspective in their critical readings and interpretations of Islamic law and power.

In Egypt, where there has been a rich tradition of modernist Islamic thought, there is an increasing concern for the feminist consciousness to be religiously rooted in Islam. Whereas in Turkey, where secularism and Islam were radically separated, Islamic women's claim to have the right both to wear the headscarf and to enter into secular public spaces (to attend the universities, to be admitted to public service, and to be elected to the Parliament) addresses not only such questions with new urgency to Islam, but also foremost to secular definitions of feminism and politics. The homogeneous and secular constitution of the public sphere is thus challenged by the emerging claims of religion and difference.

Gender and Multiple Modernities

The entrance of the newly veiled women into the public sphere across the Middle East by political, educational, and professional means stands in continuity with the process begun by the upper and middle class women who took off the veil almost a century ago. Both are products and agents of social change, transforming the relations between tradition and modernity, and religion and secularism. If the first generation of women acted for accommodation and aspiration to modernity, framed by nationalism, the second generation searches resources in religion to act upon modernity. Islamism is yet another moment in the indigenization of modernity in the Middle East; it represents a critical reflexive mode of the appropriation of modernity.

Religious and cultural traditions, rather than being immutable, static, and marginalized, emerge as a vibrant force in the processing of modernity (Fernea, 1994; Gocek and Balaghi, 1994). Traditions are selectively reappropriated and transformed in the light of modern dilemmas that women face at the forefront. Women's employment, political engagement, and public presence, have an impact on the necessity to reconsider the family sphere, the place of motherhood in society, and the values of modesty and primacy of community. Rather than assimilating to the liberal values of modernity and encouraging every move from the private to the public in the name of liberation Islamic politics advances cultural difference

as a form of resistance and as a way to articulate an alternative project. Women's bodies are sites where this politics of difference is inscribed; the new veiling displays the embodied difference. In a similar way that the modernist experience was carried out on unveiled bodies and secularized ways of life, the Islamist difference is crystallized through women. New Muslim subjectivities are worked out by politics of difference, celebrating and disciplining religious self and bodily practices in a modern universe. Muslim women are situated at the crossroads between assimilating modernity and subjugating traditions; gender consciousness provides women with a critical reflexivity, a compass to map out a new trajectory. The interplay between Islam and modernity blurs the distinctions between secular and religious, private and public. New definitions of self, modern, and Muslim are emerging from this conflictual but creative encounter. Women's agency is decisive in determining the course of this ongoing, dynamic process.

The new combinations and crossbreeding between cultural difference and modernity, Islam and gender equality necessitate a new perspective for understanding simultaneous but different experiences of modernity. Approaches in terms of insufficient and delayed modernity do not account for the centrality of women in the Middle East. Studying gender in the Middle East inevitably reinforces local and a cross-cultural comparative perspective.

See also: Islamic Fundamentalism; Nationalism, Historical Aspects of: Arab World; Near Middle East/North Africa Studies: Culture; Near Middle East/North Africa Studies: Politics; Near Middle East/North Africa Studies: Religion; Near Middle East/North Africa Studies: Society and History.

Bibliography

Abu-Lughoud, L. (Ed.), 1998. Remaking Women. Feminism and Modernity in the Middle East. Princeton University Press, Princeton, NJ.

Adelkhah, F., 1991. La revolution sous le voile. Karthala, Paris.

Ahmed, L., 1992. Women and Gender in Islam. Yale Univesity Press, New Haven, CT.

Badran, M., 1995. Feminists, Islam and Nation. Gender and the Making of Modern Egypt. Princeton University Press, Princeton.

Fernea, E. (Ed.), 1994. Women and the Family in the Middle East: New Voices of Change. University of Texas Press, Austin, TX.

Gocek, F.M., Balaghi, S. (Eds.), 1994. Reconstructing Gender in the Middle East. Tradition, Identity and Power. Columbia University Press, New York.

Gole, N., 1996. The Forbidden Modern. Veiling and Civilization. University of Michigan Press, Ann Arbor, MI.

Haddad, Y.Y., Esposito, J.L. (Eds.), 1997. Islam, Gender and Social Change. Oxford University Press, Oxford, UK.

Kandiyoti, D. (Ed.), 1991. Women, Islam and the State. Macmillan, London.

Keddie, N.R., Baron, B. (Eds.), 1991. Women in Middle Eastern History. Shifting Boundaries in Sex and Gender. Yale University Press, New Haven, CT.

Mayer, A.E., 1999. Islam and Human Rights. Tradition and Politics. Westview, Boulder, CO.

Mernissi, F., 1975. Beyond the Veil: Male–Female Dynamics in a Modern Muslim Society, rev. ed. Indiana University Press, Bloomington, IN.

Moghadam, V.M. (Ed.), 1994. Gender and National Identity. Women and Politics in Muslim Societies. Zed Books, London.

Najmabadi, A. (Ed.), 1990. Women's Autobiographies in Contemporary Iran. Harvard University Press, Cambridge, MA.

Stowasser, B.F., 1994. Women in the Qur'an, Traditions, and Interpretation. Oxford University Press, Oxford, UK.

Yamani, M., 1996. Feminism and Islam. Legal and Literary Perspectives. Ithaca Press, UK.

Near Middle East/North Africa Studies: Politics

Steve Heydemann, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 10442–10447, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

This article reviews the development of Middle East politics as a field of study. It identifies prominent themes and questions that have shaped research on Middle East politics in the United States, Europe, and the Middle East, focusing on comparative politics and comparative political economy. The article emphasizes how the relationship between the study of Middle East politics and the social sciences has changed over time. It concludes with a brief speculation about where this relationship is headed at the turn of the twenty-first century. It concludes with a brief overview of the Arab Spring events that shocked the politics of this region.

The study of Middle East politics has deep historical roots in the social sciences. The Middle East – defined here to include the region that stretches from Morocco in the west to Afghanistan in the east, and from Turkey in the north to the Arab Peninsula in the south – has figured prominently in a tradition of political and sociological-comparative research stretching from Max Weber to Karl Wittfogel and Immanuel Wallerstein. Broadly defined, this tradition has sought to understand why capitalism and democracy emerged first in the West, and how the initial appearance of these forms of economic and political organization in the West affected the developments of non-Western parts of the world. For well over a century, social scientists have drawn, in particular, on the experience of the Ottoman Empire to study political institutions, social structures, and economic arrangements perceived as explicit counterpoints to those that sparked the rise of capitalism and the transition from absolutist to republican systems of rule in the West (Anderson, 1979).

This tradition, in which the experiences of the Ottoman Empire and the Arab states that formed in its wake are used as counterpoints to European experiences – often negative counterpoints – has had a lasting impact on the study of Middle East politics. It established a focus on absolutism, 'predatory rule,' despotism, and authoritarianism as central themes in social science research on the Middle East. The legacy of this tradition has been decidedly mixed. On the one hand, authoritarianism and personalistic modes of governance remain prominent issues in the study of Middle East politics today – appropriately so, given political conditions in the region (Heydemann, 1999; Weeden, 1999). Less positively, this legacy is expressed through research in which individual behavior and collective outcomes are understood in terms of deeply embedded cultural constraints – an essentially authoritarian political culture, for example – that reinforce patrimonialism, undermine prospects for democracy, and impede the development of institutions capable of securing economic growth. Such culturalist assumptions reinforce the impression of Arab (or Muslim) politics as exceptional. They diminish possibilities for integrating Middle Eastern cases into broader comparative frameworks or applying to social and political processes in the Middle East theories and methods that have been used productively in other settings. In practice, moreover, culturalist research has rarely escaped the tautological quality of many theories that rely on culture as an explanatory variable. Largely for this reason, culturalist arguments are less prominent in the study of Middle East politics today than they were in the 1960s and 1970s.

The Post-World War II Period

After World War II, with the rise of the USA as a superpower, and the not unrelated emergence of comparative politics as an established subfield of political science in the USA, the study of Middle East politics made significant contributions to a number of important research programs in the social sciences. These included theories of political development, political decay, and praetorianism (military rule), as well as theories of dependent and peripheral development (Keyder, 1987; Abdel-Malek, 1968). As early as 1956, for example, economic historian Charles Issawi studied the relationship between economic development and democracy in the Middle East, predating the more general work on this topic by Seymour Martin Lipset (1960). Daniel Lerner's *The Passing of Traditional Society: Modernizing the Middle East* (1958) also influenced the subsequent development of modernization theory. His study was cited by Lipset and was noted, as well, in Almond and Coleman's *The Politics of the Developing Areas* (1960), a volume that exerted considerable influence in the USA on the study of politics in the developing world for much of the 1960s and 1970s, both as a model to be followed and as the object of criticism. Samir Amin's *Accumulation on a World Scale: A Critique of the Theory of Underdevelopment* (1974) was an influential contribution to debates about the role of the industrialized North in imposing conditions of underdevelopment on the postcolonial states of the South. Edward Said's *Orientalism* (1978) originated outside the social sciences, but has had a profound and widespread impact on the study of Middle Eastern politics and is central to the rise of postmodernism within the social sciences more generally. Said's work, along with the writings of postmodern theorists such as Michel Foucault, has generated significant interest in discursive and

interpretive methods in the study of Middle East politics (Mitchell, 1988), and revitalized research on power and authority in relations between the Middle East and the West. Other important areas of research in which the Middle East has figured prominently include the study of nationalism, state formation, and state–society relations in the developing world, religion and politics, and social revolution – themes that continue to inform research on Middle East politics today.

For much of the post-War period, therefore, the study of Middle East politics, especially in the USA, developed in a close and interactive relationship with the field of comparative politics and with other disciplines in the social sciences and humanities. During these years scholars of Middle East politics, like their colleagues who worked on other regions, made considerable progress in exploring such concerns as the experience of colonialism and how it influenced the trajectory of political and social life after independence, state formation, the construction of national societies, and the formation of national markets; the political economy of industrialization in the Middle East and the region's incorporation into a global capitalist system; the role of new military elites and the rise of single-party authoritarian regimes; the incorporation of workers and peasants into the political arena and its effects on coalition building and regime formation; the rise of import substitution industrialization as the dominant economic development strategy in the region and the growing role of the state in the economy, as well as the relationship between private sectors and regimes that often pursued, and coercively implemented, highly redistributive social policies; the effects of regime type on a wide range of economic, social, and political processes; the rise of state-corporatist models of interest representation; and more recently the revitalization of 'civil' societies as sites of political contestation.

While these issues were prominent in social science research on Middle East politics within the USA and provided the empirical foundations for theoretical developments in the field, research programs in Europe followed a somewhat different trajectory, less with respect to themes and issues – European scholars, like their colleagues in the USA, engaged issues concerning the state, modes of governance and power, the politics of economic development, and processes of social formation – than to methods and conceptual approaches. Disciplinary boundaries loomed less large among European scholars of Middle East politics, who preserved closer links between history and political science, than became the case in the USA, especially from the 1980s onward. It is also the case, however, that European traditions of linguistic–philological and Orientalist scholarship, as well as the legacy of having ruled the Middle East as colonial powers, were more evident in the concerns of European research programs than in those that developed within the USA. In addition, during the post-War period, the work of European scholars reflected a more assertive engagement with various strands of Marxist political theory than was evident in the Cold War atmosphere of the USA. This was manifest most productively in the development of a rich tradition of labor studies among European, especially French, scholars of Middle East politics. More recently, the presence of growing Middle Eastern communities within Europe has sparked considerable interest in issues such as the dynamics of migration from Muslim societies to Europe, the experience of immigrant life among Muslims in Europe, the politics and political economy of a greater Mediterranean region that links, in particular, North Africa and southern Europe, as well as the impact of expanding Muslim populations on social relations and political processes in largely Christian European states.

Within the Middle East, as well, research trends were and remain heavily influenced by the experience of colonialism. Higher education systems tend to be organized along lines similar to those found in the country that exercised colonial authority in a given place, with scholars in North Africa and Lebanon socialized into French modes of training and scholarship – as well as French academic networks – while training, research, and professional socialization of scholars in Egypt and Jordan reflects (though to a more modest degree) the influence of British rule. However, the most powerful influence on the study of contemporary Middle Eastern politics by scholars working in the Middle East is the repressive political environment found in most Arab states, and the resulting lack of financial support, data resources, and independent institutions capable of functioning as sponsors and promoters of social science research. The research environment in the region varies, however, and scholars in the Arab world continue to work, against the odds, to produce high-quality social science. In recent years, active and vibrant research communities in Morocco, Turkey, Egypt, and Lebanon (among others) have established significant research programs on democratization, Islam and politics, civil society and the role of new social movements, the politics of economic liberalization, as well as regional and international relations. The state of social science research on Middle East politics in Israel reflects the influence on Israeli scholars of trends in American social science, including an emphasis on quantitative and formal methods. Work on Middle East politics also benefits from a more open research environment than elsewhere in the region. However, Israeli research programs also reflect the effects of the protracted and persistent Arab–Israeli and Palestinian–Israeli conflicts, both as a filter that shapes the perspectives of researchers and as a focal point of research programs themselves.

Politics and Society

Since the early 1990s, and still in keeping with broader trends in the social sciences, scholars of Middle Eastern politics have focused attention on issues of gender, economic and political informality, local-level politics, identity politics, civil society, and the dynamics of social movements (Singerman, 1995; Norton, 1996). They have sought to integrate the experience of the Middle East into larger debates in the social sciences around such questions as authoritarianism and democratization (al-Naqib, 1991; Salame, 1994) and the comparative political economy of development (Waterbury, 1993; Waldner, 1999). Inevitably, however, the study of Middle East politics has also emphasized that which is distinctive about the region. Three aspects of Middle East politics have received particular attention from researchers over the past two decades.

First, the relationship between Islam and politics, and the dynamics of politics in Muslim societies more generally have been and remain subjects of considerable interest (Eickelman

and Piscatori, 1996). These became especially prominent in the wake of the Iranian Revolution of 1979 and the increasing visibility of militant Islamist movements and organized Islamist political parties from Morocco to Afghanistan. Second, the social, economic, and political consequences of the oil boom of the 1970s on Middle Eastern states and societies have also sparked a considerable body of research in Middle East comparative politics and comparative political economy that focuses on the fiscal sociology of the state. Perhaps the most enduring contribution of this research can be found in theories of the rentier or distributive state (Luciani and Beblawi, 1987; Chaudhry, 1997) that explore the political, institutional, and social implications of circumstances in which states and governments generate a significant share of their revenue from the sale of natural resources and are not dependent on the direct taxation of their citizens. Whether representative forms of governance can emerge where taxation (the direct extraction of resources from citizens) is low remains a matter of considerable interest in the study of Middle East politics. With the decline of oil prices in the 1980s and 1990s, the question of whether, and under what conditions, states that have not historically developed the capacity to tax might do so has also received some attention.

Third, and not entirely unrelated to work on the political economy of oil, researchers in a number of disciplines and fields, writing from a variety of (often contending) perspectives, have developed research programs focused on the state in the Middle East. Again, this work has been situated largely within political science, and in the subfields of comparative politics and international relations, but has been a core concern of scholars in the Middle East and in Europe as well as in the USA. Here too, moreover, the state as a conceptual category has been approached from a wide range of perspectives, including many that have subjected the category to sustained critical scrutiny. Within comparative politics, one persistent debate has addressed the question of the exceptionalism of the Arab state; whether culture and history (the absence of an experience comparable to the Enlightenment, and the resulting lack of differentiation between religion and state are most often cited in this regard) combined to produce a distinctive form of state, one that defies ready comparison with states in other regions and, moreover, requires distinctive conceptual tools and a special analytic vocabulary if we are to make sense of it. Increasingly, however, researchers tend to operate on the contrary assumption and view the Arab state as simply one variant among a range of state forms. States in the Middle East are shaped by and subject to the same social, economic, and political processes that affect states everywhere, and there is a recognition that these processes are filtered through and experienced within particular historical contexts, institutional legacies, and cultural frames of reference that give state formation in the Middle East a distinctive quality (Zubaida, 1989; Ayubi, 1995). This starting point permits the development of research programs that draw on and contribute to general theoretical frameworks while accounting for both intra- and interregional variations.

In addition, questions about and critiques of the state in the Middle East have also served as the basis for a number of productive research programs. A concern with the fixity or fluidity of state borders in the case of Israel and Palestine served as the basis for Lustick's general theory of how state boundaries change (1993). Mitchell (1991) has used Middle Eastern cases to explore how the impression of a boundary separating the state and society operates to construct stateness as a general attribute, and how this attribute is mirrored in (and, in his view, distorts) research on the state in political science. Migdal (1987) also drew on a nuanced understanding of the state in the developing world, including the Middle East, to explore how societies maintain a measure of autonomous power in the face of states that seem to possess an extraordinary capacity to intervene in and shape societies.

International Relations

In international relations, the contingent quality of state sovereignty in the Middle East and the impact on state consolidation of regimes that were at times highly responsive to transnational, pan-Arab ideologies pose important challenges to neorealist theories of the international system. Pointing to the fluid, fragmented, and contested quality of stateness in the Middle East, scholars of international relations, both those more sympathetic and those more critical of neorealism, have used Middle Eastern cases to question the neorealist assumption that states are like units that operate as unified actors intent on maximizing their security in an anarchic international system. They have worked to develop theories of interstate behavior that account for variability in the nature of state interests and extend more weight to economic considerations, and to the role of norms, ideas, and institutions as determinants of policy making (Barnett, 1998; Mufti, 1996).

At the beginning of the twenty-first century, however, the study of Middle East politics stands in uneasy relationship to the social sciences. Within political science, sociology, and certain subfields of economics, significant research programs have emerged to explore processes of democratization, the transition from state- to market-based modes of economic governance, and the deepening of economic globalization. Yet, as noted earlier, the Middle East remains largely authoritarian, despite limited political openings in some states. The transition to markets has been partial and slow. Linkages between domestic and global markets vary considerably but typically are weak. Economic performance also varies widely, and though trends in the 1990s were positive, overall performance levels remain low. The effect of these conditions has been to limit opportunities for integrating Middle East cases into some current research agendas, notably but not exclusively in the subfield of comparative politics. In contrast to general trends in the discipline, therefore, the study of Middle East politics at the dawn of the twenty-first century has tended to focus less on democratization and democracy than on understanding how persistent authoritarian regimes adapt to changing domestic and international conditions, less on the transition to and consolidation of markets than on the limits of economic reform, less on the impact of global economic integration than on barriers to globalization, and less on developmental success than on persistent obstacles to development. Work on these questions has been profoundly important in explaining political outcomes in the Middle East and holds out significant theoretical value for the disciplines. Yet, they are, nonetheless,

reflective of how the everyday-lived experience of politics in the Middle East has moved the study of politics along paths that differ from those that are grounded in the experiences of other regions in the developing world.

Methodological issues have also constrained the inclusion of Middle East cases into broader research agendas. Generally, the study of Middle East politics rests on the use of case-specific, macrohistorical methods rather than the formal or quantitative methods that are now more broadly applied to the study of politics in other regions. The reasons for this are several, but the methodological orientation of research on Middle East politics results, in no small measure, from the lack of data resources that are needed to undertake quantitative research and a research climate that is not conducive to data collection and dissemination. Thus, while the continuing development of various institutionalist approaches holds out considerable potential for linking region and discipline, and for cross-regional comparative research that incorporates Middle East cases, prospects for building robust research programs that make use of quantitative and formal methods for the study of Middle East politics are limited, at least at present.

Despite these concerns, current tensions in the relationship between the study of Middle East politics and the social science disciplines should not be overstated. Certainly, it should not be permitted to overshadow the presence of a large number of creative, analytically focused, and highly productive research programs both inside and outside the USA. As noted, researchers on Middle Eastern politics have made good use of and contributed to the development of new institutionalist methods. Comparative macrohistorical research coexists comfortably with case-specific but theoretically generalizable research programs that apply a wide range of methods – from archival, discursive, and interpretive to public opinion polling; from network and social movement theories to moral economy and rational choice – all of which are being used to ask important questions about contemporary political life in the Middle East and to engage the analytic concerns of the disciplines. Through attention to the dynamics of political institutions and institutional change; the organization and dynamics of local-level politics and social networks; business–state relations; the relationship between gender, power, and politics; the impact of new information technologies on political life; the formation and transformation of identities; new approaches to the study of culture and politics, and to the role of norms and ideas in political behavior – a list that is far from complete or comprehensive – scholars of Middle Eastern politics are pursuing research programs that not only resonate across regional boundaries but express increasingly sophisticated theoretical perspectives. On this basis, there is reason for optimism that the relationship between the study of Middle Eastern politics and the social sciences is now (again) becoming interactive, reciprocal, and mutually constructive.

A Postscript: Arab Spring

On 17 December 2010, Mohamed Bouazizi committed an act of self-immolation in Tunisia, which led to a series of protests that eventually unleashed widespread political conflict and change throughout the region, an event broadly called 'Arab Spring.' Some countries were affected only mildly. Saudi Arabia faced only moderate protests, while larger protests forced Jordan's King Abdullah to dismiss the government – not for the last time, as the king would have to dismiss several more governments in the face of mobilized public pressure. The Arab Spring had a more fundamental impact elsewhere, perhaps the most dramatic being the overthrow of Muammar Gaddafi's long-lived regime in Libya and Gaddafi's eventual death in the civil war that erupted there. Ultimately, all countries of this region witnessed some degree of political contention, and in four countries leaders and regimes were brought down (Egypt, Libya, Tunisia, and Yemen). Strikes and organized demonstrations have been one fairly common trait of this wave of political change, and social media have been particularly prominent as tools for mobilization and strategic organization.

Much like the collapse of state socialist regimes in East Europe in 1989, which also had a wavelike quality and appeared quite suddenly, Arab Spring was grounded in a combination of deeper institutional causes and proximal conditions. One quality shared by countries in this region was long-simmering public dissatisfaction with leading politicians and parties. Corruption, repressive politics, and other forms of injustice had marked political systems that had various authoritarian traits (with some, such as Libya, relatively more populist than others but still authoritarian). Further, the 2008 global economic crisis had also affected markets and employment in the region, leaving many average civilians unemployed or not well-off economically. These political and economic strains were exacerbated by inability of these authoritarian regimes to address social pressures, either through sufficiently robust welfare measures or through the kind of democratic politics (e.g., elections and turnover in leadership) that can act as an outlet for social discontent. Finally, these were not all docile societies. For example, there had already been stirrings of protest in Egypt and Tunisia before 2010; in Egypt, in particular, trade unions had become active, defending workers' rights, calling strikes or other forms of industrial actions before 2010. Thus, in several countries there already was an existing organizational foundation for mobilizing and directing action, and experience with collective action. Social media facilitated exchange of ideas and information to aid that mobilization, although social media alone did not provoke or shape protest.

The trigger that turned potential political contention into real social protest and action was Bouazizi's act in December 2010, which he committed in response to his treatment by a local inspector and general discontent with corrupt police. His subsequent death in January 2011 brought together a wide swath of people (rights activists, lawyers, students, trade union representatives, etc.) also disaffected with Tunisia's state. Protests in Tunisia in January 2011 began to spread elsewhere, in part because of a demonstration effect: citizens in one country might view successful protest elsewhere, and sharing similar dissatisfaction, become more likely to take action themselves. Protests in Tunisia quickly spread to Oman, Egypt, Morocco, and Syria, and by February 2011 Hosni Mubarak had stepped down from the presidency of Egypt (a position he had held through rigged elections since the 1980s), and Libya's eventual civil war had begun with conflict in Benghazi. By the end of 2011, the army was maintaining some order in Libya and Egypt, although protests and violence continued to break

out in urban areas and organized protest groups called for the military to transfer political authority to elected bodies and officials.

Except in Syria, where a bloody civil war was dragging out, the politics of many countries of the Arab Spring had settled down to an extent; some had a change in government, while in others (e.g., Algeria) major protest had been repressed but had forced some government concessions (e.g., more social provisions or some political reforms). In Libya and Egypt, social discontent continue to break out on occasion, for example, the attack of the American diplomatic mission in Benghazi in September 2012, or the political crisis in Egypt in 2012 and 2013 during the brief tenure of President Mohamed Morsi, who was overthrown by the military in July 2013. The worst violence of the Arab Spring was in Syria, where Bashar al-Assad responded to protests with violent repression. This transformed the protest into armed resistance, and as a result Syria was engulfed in a civil war that continues to this writing and has so far claimed more than 120 000 victims.

The ultimate effect of Arab Spring is impossible to predict at this point in time, given the newness and rapidity of events. While there have been changes in specific leaders, deeper institutional and structural changes have thus far been far weaker. Militaries continue to play important roles in these countries, and democratic procedures have not yet taken deep roots (although it is far too early to expect such a result). The Syrian civil war also contributed to the creation of the radical group ISIS, which controls territory within Syria and Iraq and has proven to be a major threat to all actors in the region. As for any revolutionary event, Arab Spring has had its romantic portrayals, and the upheaval may ultimately be more superficial than real – but the effects of such revolutions take time to emerge as well. What is clear is that the region is still not entirely stable and more change likely awaits elites and citizens before long.

See also: Islam: Middle East; Nationalism, Historical Aspects of: Arab World; Near Middle East/North Africa Studies: Culture; Near Middle East/North Africa Studies: Religion; Near Middle East/North Africa Studies: Society and History; Orientalism.

Bibliography

Abdel-Malek, A., 1968. Egypt, Military Society: The Army Regime, the Left, and Social Change under Nasser. Vintage Books, New York.

Almond, G.A., Coleman, J. (Eds.), 1960. The Politics of the Developing Areas. Princeton University Press, Princeton, NJ.

Amin, S., 1974. Accumulation on a World Scale: A Critique of the Theory of Under-Development. Monthly Review Press, New York.

Anderson, P., 1979. Lineages of the Absolutist State. Verso Editions, London.

Ayubi, Nazih NM., 1995. Over-Stating the Arab State: Politics and Society in the Middle East. I B Tauris, London; New York.

Barnett, M.N., 1998. Dialogues in Arab Politics: Negotiations in Regional Order. Columbia University Press, New York.

Chaudhry, K.A., 1997. The Price of Wealth: Economies and Institutions in the Middle East. Cornell University Press, Ithaca, NY.

Eickelman, D.F., Piscatori, J.P., 1996. Muslim Politics. Princeton University Press, Princeton, NJ.

Heydemann, S., 1999. Authoritarianism in Syria: Institutions and Social Conflict, 1946–1970. Cornell University Press, Ithaca, NY.

Issawi, C.P., January 1956. Economic and social foundations of democracy in the Middle East. International Affairs 27–42.

Keyder, C., 1987. State and Class in Turkey: A Study in Capitalist Development. Verso, London.

Lerner, D., 1958. The Passing of Traditional Society: Modernizing the Middle East. Free Press, New York.

Lipset, S.M., 1960. Political Man: The Social Bases of Politics. Doubleday, Garden City, NY.

Luciani, G., Beblawi, H. (Eds.), 1987. The Rentier State. Croom Helm, London.

Lustick, I.S., 1993. Unsettled States, Disputed Lands: Britain and Ireland, France and Algeria, Israel and the West Bank–Gaza. Cornell University Press, Ithaca, NY.

Migdal, J.S., 1987. Strong Societies and Weak States: State–Society Relations and State Capabilities in the Third World. Princeton University Press, Princeton, NY.

Mitchell, T., 1988. Colonising Egypt. Cambridge University Press, New York.

Mitchell, T., 1991. The limits of the state: beyond statist approaches and their critics. American Political Science Review 85 (1), 77–96.

Mufti, M., 1996. Sovereign Creations: Pan-Arabism and Political Order in Syria and Iraq. Cornell University Press, Ithaca, NY.

Norton, A.R. (Ed.), 1996. Civil Society in the Middle East. Brill, Leiden, Holland; New York.

al-Naqib, K.H., 1991. al-Dawlah al-tasallutiyah fi al-Mashriq al-'Arabi al-mu'asir: dirasah bin'iyah muqaranah. Markaz Dirasat al-Wahdah al-'Arabiyah, Beirut, Lebanon.

Said, E.W., 1978. Orientalism. Vintage Books, New York.

Salame, G. (Ed.), 1994. Democracy without Democrats?: The Renewal of Politics in the Muslim World. I B Tauris, London.

Singerman, D., 1995. Avenues of Participation: Family, Politics, and Networks in Urban Quarters of Cairo. Princeton University Press, Princeton, NY.

Waldner, D., 1999. State Building and Late Development. Cornell University Press, Ithaca, NY.

Waterbury, J., 1993. Exposed to Innumerable Delusions: Public Enterprise and State Power in Egypt, India, Mexico, and Turkey. Cambridge University Press, Cambridge/New York.

Weeden, L., 1999. Ambiguities of Domination: Politics, Rhetoric, and Symbols in Contemporary Syria. University of Chicago Press, Chicago.

Zubaida, S., 1989. Islam, the People and the State: Essays on Political Ideas and Movements in the Middle East. Routledge, London/New York.

Near Middle East/North Africa Studies: Religion

Dale F Eickelman, Dartmouth College, American University of Kuwait Program, Hanover, NH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The adherents to Islam, Christianity, and Judaism in the Middle East, while making universal claims for their respective faiths, increasingly recognize the significance of the highly varied local and particular contexts in which their religious symbols, ideas, and practices are understood. This 'local knowledge' also profoundly influences the understanding of ideas and practices considered global and universal. Since the midtwentieth century, the rise of mass education, the proliferation of new communications media, and the greater ease of travel have reconfigured how people think about their own and other religious traditions. These understandings are often distorted, but the fact of their existence is now foregrounded in the consciousness of many believers. Local traditions of belief and practice, earlier taken for granted, increasingly are seen as systems of meaning and action to be articulated and reformed, argued and explained.

The Changed Setting for Religious Pluralism

Islam is the predominant religion in all the countries of the Middle East and North Africa with the obvious exception of Israel, where non-Jews make, primarily Arab Muslims and Christians, up 24% of the population. There are significant Christian and Jewish minorities throughout the region and the lives, doctrines, and practices of the carriers of all these faiths are intricately linked. As with Islam, the practice and understanding of Judaism and Christianity in the region have been highly diverse both in history and in the present, showing an ongoing creative tension between universalism and particularism. The labels 'Muslim', 'Christian', and 'Jew' – not to mention Zoroastrianism in Iran – aggregate the identities of distinct and often autonomous sectarian groups. Christians form significant minorities in Egypt, Lebanon, Iraq, Syria, and Turkey, and European settlers gave the North African states sizable Christian populations until the end of the colonial era in the midtwentieth century.

At the beginning of the twentieth century, Christians and Jews were widely distributed throughout the Middle East and North Africa (Planhol, 1997; Courbage and Fargues, 1997; Sanasarian, 2000). Until the late 1940s in northern Iraq, for example, Sunni Arab Muslims formed a majority of the population in Mosul, with Jews living in both Mosul and surrounding agricultural communities. Shii Turkomen tribes resided in outlying villages, as did communities of Monophysite, Assyrian, and Nestorian Christians. Likewise, Lebanon has 23 recognized Christian and Muslim religious sects, each with separate community identities. At the beginning of the twentieth century there were also thriving Jewish communities distributed widely throughout the region. Some of these remained vital for decades after the declaration of Israel as a state in 1948, and Morocco's Jewish community, although reduced in numbers, remains vital to this day, with Morocco's Jewish heritage formally recognized in its constitution (Gottreich and Schroeter, 2011).

Theologians and ideologues often highlight the distinctiveness of their respective faiths, but Muslims, Christians, and Jews in the region often share implicit background understandings and practices. Sacred spaces shared by all three faiths are found primarily only in Jerusalem and nearby places such as Hebron/al-Khalil, and shared Muslim and Jewish places of regional pilgrimage dot the Moroccan landscape and are equally venerated by Muslims and Jews. Such shared shrines continue to play a vital role in popular belief and practice. Thus in North Africa, festivals for Muslim saints (*musim*s) and pilgrimages to regional shrines (*ziyara*s) have their counterpart in festivals (*hillula*s) for Jewish saints, with some shrines shared by the two faiths (Albera and Couroucli, 2012).

Variations in doctrine, religious practice, and traditions of shared memory also distinguish Muslim communities from one another. These include Sunni Muslims, the Shia (or 'sectarians', who comprise about 10% of Muslims worldwide and are further divided into numerous groups), and Ibadi Muslims of Oman, East Africa, and parts of Tunisia, Libya, and Algeria. The intellectual and educated elite often deny or downplay the significance of local, regional, and sectarian differences in faith and practice, but these practices often are powerfully linked to the local social order and continue to play a major role in the popular imagination.

'Modernity' in the form of uniform state schooling and mass communications often alters these practices, but does not erode them. Nonetheless, rising educational levels and the proliferation of communications media, including satellite television unrestrained by state authorities, are tilting the balance in favor of ideas and practices that can be explained, defended, and foregrounded in public.

Explaining Religious Diversity and Change

Islam, like Judaism and Christianity, is rich and varied in its creativity and cultural expression. It is a challenge for Muslim believers and scholars, both Muslim and non-Muslim, to account for the variations of belief and practice as they have been expressed for over 1400 years in richly diverse cultural and historical traditions. As a response to this complexity, some scholars dismiss as non-Islamic or as incorrect those practices they consider not in accord with 'central' Islamic truths, even though people who hold such beliefs consider themselves fully Muslim. Some deal with these issues by

seeing 'foundation' texts at the core of each religious tradition – the Qur'an in the case of Islam – and then seeing how successive generations understand and articulate them. If such texts and ideas are too closely bound to their original settings, no matter how persuasive for their original audience, they risk failing to attract wider audiences. Such texts must remain relevant to the social and political contexts of succeeding generations.

When first introduced in the late 1940s, the notion of 'Great' and 'Little' traditions offered one way of coming to terms with the universal and the particular in religious expression and practice. The Great/Little tradition contrast rekindled an interest in how popular understandings of religion – the 'little', localized, traditions – were related to elite, literary ones. It also encouraged attention to the specifics of who advocated or explained particular religious orientations and practices, whether they were influential in their own society or marginal, appealing primarily to a particular social group or category (Von Grunebaum, 1955). In some settings, individuals claiming descent from the family of the Prophet Muhammad were given precedence. The concept of the Great/Little tradition encouraged exploring the links among religious ideas and practices, authority, and influence.

Often, however, the concept was employed only to juxtapose statements of 'essential' Islamic principles, as elaborated in standard texts and by educated Muslims, with inventories of local religious practices. For example, the study of trance and possession states (called *zar* cults in the Sudan and East Africa), often indicative of the affective side of religious traditions, was often frequently relegated to the 'little' tradition side of the divide and attributed primarily to women.

Yet such ideas and practices also shed light on ideas of the person and how both women and men in societies with such practices express and explore alternative social realities (Boddy, 1989). Mosque-centered activities such as Friday prayers, almost the exclusive domain of men, often have been seen as part of the 'great' tradition. Women, who often predominate in visits to local shrines and in *zar* cults, are then seen as sustainers of 'local' traditions. This gendered division is made all the more plausible by the fact that until recent decades women had lower levels of education and mobility than men. An alternative explanation, however, is to see mosque-oriented and shrine-oriented practices as integrally complementary, and which must be 'read' together as a unit in order to understand religious practice (Tapper and Tapper, 1987).

In North Africa, Hammoudi (1993) suggested the explanatory utility of linking the canonical Muslim sacrificial feast associated with the annual pilgrimage to Mecca with practices such as *Bûjlûd*, a carnival-like masquerade in Morocco's Berber highlands that follows a few days after the canonical feast. Colonial ethnographers regarded the two events as separate, as do Muslim intellectuals. In the masquerade, a young male villager dons the skin of a sacrificial animal and assumes the character of *Bûjlûd* ('the wearer of the skins'), and with his cohorts cavorts noisily throughout the village, sacrilegiously mocking established religious and social conventions, entering homes, and bantering obscenely with women. Hammoudi argues that the masquerade evokes the contradictions implicit in the cultural classifications that regulate ordinary behavior, including gender roles, directing caricature and self-directed irony against local social hierarchies, tradition, and innovation. The canonical Muslim sacrifice, on the other hand, embodies the ideal community. The canonical sacrifice and masquerade inherently conjoin ritual piety and impiety, embodying an ongoing contest over significance and interpretation in which existing and alternative social orders are embodied and intertwined.

Ritual and Belief in Morocco

In the late 1960s, Morocco became 'good to think with' for understanding religion in complex societies, just as the Trobriand Islands had earlier entered the anthropological canon for understanding exchange systems and India for understanding caste and stratification. The ethnographies of Morocco's colonial era did not lend themselves to comparative studies or lead to disciplinary classics. The publication of Clifford Geertz's *Islam Observed* (1971), originally a series of lectures delivered to historians of religion, marked a breakthrough in linking religious studies in general to the ethnographic study of specific cultural traditions. Geertz took for granted the unity of the Islamic tradition, but emphasized the highly divergent ways in time and space that it articulated with social experience. His point of departure was Islam as experienced, interpreted, and practiced from the fifteenth century to the present in Indonesia and Morocco, the two antipodes of the Muslim majority world. For Geertz, Islamic symbols and meanings are culturally defined and socially shared, like all other systems of meaning. If commonsense gives shared meaning to immediate practical experience, religion offers a higher and more general interpretive order. These two systems of meaning – commonsense and religion – are intricately intertwined, both shaped by historical experience and giving shape to it.

A key element of Geertz's argument was to explore the constructed meanings of sainthood and Sufism (mysticism) in the two contrasting settings. In Morocco, individuals with highly varied attributes – reputations for piety, scholarship, miracle working, or other distinctions – were popularly imputed with saintly authority. Since the current Moroccan social order took hold in the fifteenth century, Geertz argued, the cultural order in which Moroccan ideas of saintly authority was sustained was characterized by a 'strong man' politics of zealous warriors, rulers, and saints enforcing their will and inducing clients to act on their behalf. Their authority was charismatic, based on personal dominance. In contrast, the Indonesian view of sainthood is hierarchical. Saints incrementally achieved prominence through years of meditation and Indonesia's settled agricultural villages placed high value on cooperation and order.

In the contexts of both Indonesia and Morocco, the idea of saintly authority took significantly divergent forms because of the contrasting nexus of commonsense perceptions of the world and the different historically constructed senses of meaning through which Indonesians and Moroccans interpreted Islam. These commonsense perceptions and background understandings form the necessary point of departure for understanding religious belief, experience, and practice. Indeed, the perspectives of religion and commonsense are

intertwined in a dynamic, often unstable, dialectic (Zein, 1977; Eickelman, 2005).

Theologians and 'common' people interpret religious texts, including the 'foundation' texts of religious traditions such as the Qur'an, in multiple and contested ways. Thus, the most common term for saint in contemporary Morocco is 'pious one' (*salih*), although 'marabout', the term used by Geertz and which has made its way into French and English usage, is also understood. 'Pious one' conveys the multivalence of Moroccan usage. For educated Moroccans, the term is reserved solely for individuals of outstanding piety, and many speak of saint, or maraboutic, cults as 'pre-Islamic' vestiges. For many other Moroccans, however, both in towns and in rural areas, the 'pious ones', living or dead, are imputed with mystical powers, including the ability to secure positive outcomes for individuals and groups who have maintained a connection with them through gifts or donations to their descendants (Eickelman, 1976). Moroccan monarchs and senior government officials continue publicly to visit the shrines of pious ones and make gifts toward their maintenance. Elsewhere in the Muslim world there is a similar interplay between different notions of piety, and there is a similar dialectic of ideas of sainthood among Jews of North African origin (Weingrod, 1990).

The emphasis Geertz placed on the cultural dimension in understanding religious experience suggested an alternative to characterizing religious traditions in terms of an essential core prevalent in all places and for all historical periods. His approach instead stresses how systems of religious meaning and practice, like all cultural experience, are contested and dynamic, and therefore not reduced to a single, cohesive set of principles. Some conceptions of Islam are significantly intertwined with, although not limited to, the local social order. However, other conceptions are more universalistic and global, amenable to articulation with a wide range of contexts. These universalistic and particularistic conceptions are copresent and in a dynamic, often volatile, tension with one another.

Geertz's approach, like other efforts to understand religious expression and practice in Moroccan society and elsewhere, including Ernest Gellner's *Muslim Society* (1981) – note the singular – had significant implications for understanding religious expression in earlier historical periods and outside the Muslim majority world. For example, Brown (1981) notes that his understanding of the moral order of early Latin Christianity, the 'lasting, urgent problems of any civilized community – friendship and protection, justice, power and mercy' – were in part inspired by ethnographies of contemporary North African saints. Similarly, Cornell (1998) shows how medieval Moroccan arguments on the nature of sainthood – cultural understanding of and debates among mystics in both urban and rural settings about concepts related to God's love and authority from the eleventh through sixteenth centuries – contributed to the Sufi tradition elsewhere in the Muslim world. Cornell's argument also compares the North African experience with sainthood as understood and practiced in Europe in the same period. Writings on Judaism in North African and Middle Eastern societies likewise indicate a reemphasis on the cultural dimension and explore the significance of local cultural contexts (Deshen, 1989; Goldberg, 1996; Gottreich and Schroeter, 2011).

The Iranian Revolution and Religious Politics

The 1978–79 Iranian revolution was one of several major events that underscored the continued significance of religion in politics and public life, notwithstanding its neglect in most midtwentieth century discussions concerning modernization and modernity. In Iran itself, the revolution led to the creation of a new, rapidly changing religious idiom for interpreting religious, social, and political events – and to contesting them (Khosrokhavar, 1997; Adelkhah, 2012). The debates in Iran continue unabated over the form that an Islamic state should take and the role that religion should play in society, including gender relations (e.g., Mir-Hosseini, 1999; Adelkhah, 2000). Successive incremental decisions and the unanticipated outcomes of earlier ones meant that the revolution looked different in the decades after the event both to many of those who participated in it and to the majority of Iranians not even born when it happened. Some observers use the term 'post-Islamism' to describe Iranian society in the 1990s, emphasizing the significantly different way that many Iranians, including many clerics, began to argue publicly for a more limited role of religious institutions in Iranian politics and public life (Khosrokhavar and Roy, 1999).

In spite of contemporaneous hopes for some and fears for others at the time of the Iranian revolution, it did not trigger similar political movements in other Muslim majority countries. It did, however, compel regimes elsewhere in the Middle East, both secular (including Egypt, Turkey, Syria, and Algeria) and religiously oriented (including Morocco and Saudi Arabia) to become more attentive to popular understandings of social justice and religion in society. From the first decades of the twentieth century, new religious movements emerged that had distinctly modern political forms. Egypt's Muslim Brotherhood, founded in the late 1920s and regarded by the midtwentieth century as Egypt's only genuinely popular mass political movement, foreshadowed the growth of similar later Islamic movements, many of which benefit from their ability to attract leaders with strong skills in organization and communication. Distinctly modern in organizational form, with a clear hierarchy and (at least at the outset) formal criteria for membership, the Brotherhood's leadership rested not with traditionally educated religious scholars but with those effective in organizing their followers and articulating their values, providing social and medical services as well as religious inspiration in many communities. After the fall of the Mubarak regime in Egypt in February 2011, Egypt's Muslim Brotherhood reemerged as a powerful political force, but with strong competition from other secular and religious groups. As after many changes in political authority, those kept by force from exercising political authority and rights in Egypt and elsewhere in the region show inexperience in managing the state machinery once in office. The result has been to engender significant competition from both religious groups, whether or not formally organized for political purposes, and secular ones.

As dramatic as the events of the 2011 'Arab Spring' were in some parts of the Middle East, they were in many ways the culmination of long-term shifts in economy, society, and religious expression and practice. As early as the 1960s, for example, the Muslim Brotherhood and similar organizations looked conservative in contrast to the proliferation of

movements advocating the Islamization of politics and society. These 'Islamist' movements ranged from small, radical groups such as the one responsible for the assassination of Egypt's Anwar Sadat in 1981 to larger ones sometimes organized like political parties, such as Hizbullah of Lebanon, the Virtue Party of Turkey (succeeded in 2001 by the Justice and Development Party (AKP)), or the Islamic Salvation Front of Algeria. Some of these Islamist movements had their origins as 'spokesmen for the despised', and had Jewish and Christian analogues (Appleby, 1997; Marty and Appleby, 1995). In general, the influence of extremists, having taken off in the late 1960s, appeared to have waned by the 1990s (Kepel, 2000), only to have subsequently regained momentum. Most movements work within existing national boundaries, even if they sometimes receive external or transnational support. Some movements have advocated violence to achieve their goals, but others have been strong advocates of tolerance and civil society, and many movements have altered as they enter the mainstream of national political life.

Religion as System and Ideology

Since the latter half of the twentieth century, mass education and the proliferation of new communications media have profoundly altered how religious knowledge and practice are produced and understood in the Middle East. In the Arab world, for example, large numbers of people now have the level of literacy needed to follow religious and political debates from multiple sources and the command of formal language needed to participate effectively in public life. The religious lessons learned from state schooling may be regarded with skepticism, but like other aspects of state schooling, they inculcate the habits of thought and language needed to learn about and elaborate alternative views of the role of religion in society.

Ironically, the ideas of many radical Islamists mirror competing secular ideologies in form and structure. The writings of Sayyid Qutb, executed by Nasser's Egypt in 1966, attracted a wide audience among white-collar workers, professionals, and students, the beneficiaries of modern education. In Arabic and in translation, Qutb's writings continue to have widespread appeal because they offer 'Islamic' explanations for contemporary political and economic developments and for the perceived injustices of existing regimes. More contemporary religious ideologues, like Morocco's 'Abd al-Salam Yasin (1928–2012), argue that contemporary Muslim societies have been de-Islamized by imported ideologies and values. Yet the style and content of these writers, like others, imply an audience already familiar with the secular, imported ideologies against which they argue. This is the case even for the writings and speeches of Osama bin Laden (1957–2011) (Euben and Zaman, 2009). Implicitly, they also convey a sense of Islam as an ideology and objectified system of meaning, which can be conveyed as one subject among others in schools, and in catechism-like pamphlets. In this sense, religion forms one of several systems of meaning in modern society, even as some believers argue that Islam is a 'complete' way of life. Yet the argument of 'completion' does not mean closure. As Iqtidar (2011) argues in a key study on a leading Islamic group in Pakistan that has moderated its views and practices as it becomes more mainstream, the contents of *shari'a* and 'tradition' are constantly contested and adapted to changing social and economic circumstances.

The awareness of alternatives to immediate social circumstances is facilitated by the increased availability of contemporary forms of communication, ranging from the press and broadcast media to readily duplicated DVD (which have largely replaced audio cassettes), ubiquitous mobile telephones, and the Internet. In an earlier era, state authorities easily regulated the media. By the end of the twentieth century, the top-down asymmetry of broadcast radio and television was reversed by newer media that are readily adaptive, interactive, and difficult for states effectively to control. These developments have contributed to the fragmentation of religious and political authority by allowing large numbers of people to participate in an increasingly public debate over the symbolic language of Islam (Eickelman and Anderson, 2003). Combined with political change and the growth of diaspora communities, these developments indicate a changed social and cultural environment for religious expression and practice in which a mass and anonymous public has given way to one characterized by an increased multiplicity of horizontal relationships. The shift to a network society has created communities and communicative spaces in which large numbers of people, while not abandoning their ties of family and community, discover and create alternative ways of thinking about the role of the sacred in society and politics.

See also: Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns.

Bibliography

Adelkhah, F., 2000. Being Modern in Iran. Columbia University Press, New York.
Adelkhah, F., 2012. Les mille et une frontierres de l'Iran: Quand les voyages forment la nation. Karthala, Paris.
Appleby, M. (Ed.), 1997. Spokesmen for the Despised: Fundamentalist Leaders of the Middle East. University of Chicago Press, Chicago, IL.
Albera, D., Couroucli, M. (Eds.), 2012. Shaping Sacred Spaces in the Mediterranean: Christians, Muslims, and Jews at Shrines and Sanctuaries. Indiana University Press, Bloomington, IN.
Boddy, J., 1989. Wombs and Alien Spirits: Women, Men, and the Zar Cult in Northern Sudan. University of Wisconsin Press, Madison, WI.
Brown, P., 1981. The Cult of the Saints: Its Rise and Function in Latin Christianity. University of Chicago Press, Chicago, IL.
Cornell, V.J., 1998. Realm of the Saints: Power and Authority in Moroccan Sufism. University of Texas Press, Austin, TX.
Courbage, Y., Fargues, P., 1997. Christians and Jews under Islam. Tauris, London.
Deshen, S., 1989. The Mellah Society: Jewish Community in Sherifian Morocco. University of Chicago Press, Chicago, IL.
Eickelman, D.F., 1976. Moroccan Islam: Tradition and Society in a Pilgrimage Center. University of Texas Press, Austin, TX.
Eickelman, D.F., Anderson, J., 2003. Redefining muslim publics. In: Eickelman, D.F., Anderson, J. (Eds.), New Media in the Muslim World: The Emerging Public Sphere, second ed. Indiana University Press, Bloomington, IN.
Eickelman, D.F., 2005. In: Geertz, Clifford, Shweder, R.A., Good, B., et al. (Eds.), Clifford Geertz and Islam. University of Chicago Press, Chicago, IL.
Euben, R.L., Zaman, M.Q., 2009. Princeton Readings in Islamist Thought: Texts and Contexts from al-Banna to Bin Laden. Princeton University Press, Princeton, NJ.
Geertz, C., 1971. Islam Observed: Religious Development in Morocco and Indonesia. University of Chicago Press, Chicago, IL.

Gellner, E., 1969. Saints of the Atlas. University of Chicago Press, Chicago, IL.
Gellner, E., 1981. Muslim Society. Cambridge University Press, Cambridge, UK.
Goldberg, H. (Ed.), 1996. Sephardi and Middle Eastern Jewries: History and Culture. Indiana University Press, Bloomington, IN.
Gottreich, E.B., Schroeter, D.J. (Eds.), 2011. Jewish Culture and Society in North Africa. Indiana University Press, Bloomington, IN.
Hammoudi, A., 1993. The Victim and Its Masks: An Essay on Sacrifice and Masquerade in the Maghreb. University of Chicago Press, Chicago, IL.
Iqtidar, H., 2011. Secularizing Islamists? Jama'at-e-Islami and Jama'at-ud-Da'wa in Urban Pakistan. University of Chicago Press, Chicago, IL.
Kepel, G., 2000. Jihad, expansion et déclin de l'Islamisme. Gallimard, Paris.
Khosrokhavar, F., 1997. Anthropologie de la révolution Iranienne: La réve impossible. L'Harmattan, Paris.
Khosrokhavar, F., Roy, O., 1999. Iran: Comment sortir d'une révolution religieuse. Seuil, Paris.
Marty, M., Appleby, R.S. (Eds.), 1995. Fundamentalisms Compared. University of Chicago Press, Chicago, IL.
Mir-Hosseini, Z., 1999. Islam and Gender: The Religious Debate in Contemporary Iran. Princeton University Press, Princeton, NJ.
Planhol, X., 1997. Minorités en Islam: Géographie politique et sociale. Flammarion, Paris.
Sanasarian, E., 2000. Religious Minorities in Iran. Cambridge University Press, Cambridge, UK.
Tapper, N., Tapper, R., 1987. The birth of the prophet: ritual and gender in Turkish Islam. Man (n.s.). 22, 69–92.
Von Grunebaum, G.E. (Ed.), 1955. Unity and Variety in Muslim Civilization. University of Chicago Press, Chicago, IL.
Weingrod, A., 1990. The Saint of Beersheba. State University of New York Press, Albany, NY.
Zeghal, M., 2008. Islamism in Morocco: Religion, Authoritarianism, and Electoral Politics. Markus Weiner, Princeton, NJ.
Zein, A.H., 1977. Beyond ideology and theology: the search for the anthropology of Islam. Annual Review of Anthropology 6, 227–254.

Near Middle East/North Africa Studies: Society and History

Mohammed Bamyeh, University of Rochester, Rochester, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 15, pp. 10451–10458, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article explores how the study of Middle Eastern society and history has moved gradually from a focus on homogeneity and continuity to more elaborate discussions of the impact of many determining factors. It is now commonly accepted that due to its historical centrality to historical world systems and its contemporary geopolitical significance, Middle Eastern society and history cannot be understood apart from global trends which influence the region and are influenced by it as well. The studies are most revealing when they go above or below contemporary state jurisdictions and emphasize instead the historically autonomous mode of self-organization of civil society, cultural communities, travel, and trade. Middle Eastern society has historically been defined more so by the hybridity and diversity resulting from the above factors than by the social engineering of states.

Study of the 'Middle East'

The term 'Middle East' first came into use in the early twentieth century, and after World War II it gradually supplanted the nineteenth century term 'Near East' as a geopolitical and geocultural designation. Both terms have been repeatedly criticized for their Eurocentric origin and point of emphasis, yet the term 'Middle East' continues to survive as a term for lack of better alternatives. It supplanted older designations such as the Levant, which designated the eastern Mediterranean and included present-day Turkey and Greece. Geographically, the Middle East encompasses West Asia, including Iran and Turkey, and North Africa, including Sudan. Some scholarly journals extend its boundaries further to include Afghanistan and the Horn of Africa. While in modern history the region has been divided among more than two dozen States, many historians and social scientists assume its various parts to contain roughly homogenous cultural patterns and sociological profiles, shaped by a shared history. Middle Eastern culture is frequently discussed in religious rubrics, and religiosity (specifically Islamic) and is frequently thought to capture what is essential about Middle Eastern culture in general. An earlier tradition in Western social science regarded Middle Eastern society culturally as tradition bound and sociologically as a tribally identified and ethnically segmented premodern mosaic kept together normally by an authoritarian State.

The social scientific study of the region we now call the Middle East is quite old, although its history is uneven. Ibn Khaldun (1332–1406), who spent most of his life in North Africa, was the first person to use the term 'sociology,' which he understood as the science of the life cycles of urban civilization. His theories, which emphasized the role of social solidarity and the weakening of that solidarity over time as civilization became more entrenched and oriented toward surplus and luxury, anticipated by centuries of European theories of civilizational cycles, expressed most loudly by Arnold Toynbee and Oswald Spengler. Ibn Khaldun's work must also be understood in the context of his time, which was characterized by conflicting claims of legitimacy, unstable political orders, recurrent Mongol invasions, the political fragmentation of the Muslim world, and the demonstrated incapacity of sovereigns to claim monopoly over the loyalties of the faithful everywhere, as had been roughly the case in the first three centuries of Islam.

An important feature of Ibn Khaldun's perspective consisted of combining the study of society and history, in the underlying belief that a sociology of civilization could only be based on an adequate philosophy of history. This continues to be the spirit characterizing much of the sociological investigation done in the Middle East itself, where many practitioners are just as likely to be historians or philosophers, and where questions of culture, religion, heritage, and modernity are inseparable from a myriad of contemporary social issues.

The modern professional study of the Middle East began in the nineteenth century from two directions. In the East, there arose an intellectual movement which came to be known as the *nahda*, whose significance has been pointed out in a recent flow of important studies by Albert Hourani, Hisham Sharabi, Aziz Al-Azmeh, and Sadiq Jalal al-'Azm, among many others. That movement was essentially an eclectic mix of reflexive native commentators whose thought was in tune with urban middle-class sentiments and in communication with European modernity. They sought to revive what they regarded to be a dormant spirit of science and liberalism in Islam, while seeking politically to modernize the Ottoman State. In Europe, by contrast, the study of the Middle East was far more scholastic and was confined to a small circle of orientalist scholars who, in the course of the nineteenth century, became increasingly better acquainted with the classic texts and archaeology of the region, especially following Napoleon's Egyptian campaign. The orientalist approach was highly antisociological and antitheoretical. It focused almost exclusively on ancient history and disregarded contemporary society, which was thought to be largely static in terms of its cultural patterns, and thus accessible through knowledge of the classic texts founding it. Influential trends within orientalism viewed oriental society culturally as static, antirational, unphilosophical, and antiscientific; sociologically as mosaic of floating tribal and ethnic solidarities; and politically as predisposed to authoritarianism.

This approach gained further ascendance within sociology itself with Max Weber's ventures into the study of oriental

cultures, including Islam. In common with the reigning orientalist tradition, Weber's approach to the question as to why Muslim societies (along with Buddhist and Hindu ones) had not developed capitalism as Europe had done, referred to the Qur'an and sayings attributed to Muhammad, and assumed the cause of social change (or lack of it) to be internal rather than external to the culture. While Weber was more interested in social theory than were orientalist scholars, who were preoccupied with philological and textual concerns, Weber nonetheless replicated the basic image of oriental society as a self-enclosed and distinct cultural cocoon, interpretable by its classical texts and definable in terms of its difference from the European experience of liberal capitalism.

The subsequent rise of modernization theory, closely associated with the Parsonian paradigm in sociology, set aside for some time the question of culture, emphasizing instead on the normative convergence of societies along the path of modernization. Daniel Lerner's *The Passing of Traditional Society* (1958) is a classic work in this genre with respect to the Middle East. While steeped in orientalist assumptions that Middle Eastern society, culture, and personality were largely tradition-bound and antirational, Lerner's work sought to detect venues of the modernist challenge arriving to society from without. Like many theorists of modernization, Lerner understood it as simply 'Westernization.' Another important text of modernization theory was C.A.O. van Nieuwenhuijze's *The Sociology of the Middle East* (1971), a detailed and lengthy study which portrayed Middle Eastern society as a set of parallel societies converging upon a cultural core. This further justified the old view of culture as an independent variable when approaching Middle Eastern societies.

Contemporary Approaches to Near Eastern/North African Societies

With decolonization, the increased presence of the United States in the region, the Cold War, the creation of the State of Israel, the rise of Arab nationalism, and the Arab defeat in 1967, Middle East studies itself became a veritable battleground. A new generation of scholars was determined to contest what they regarded to be a distorted image of the region and its peoples in both academic and public settings. The rallying cry was issued by Edward Said's *Orientalism* (1978), which immediately galvanized the field and continues to be referenced today as a founding text in postcolonial studies, generally. A sophisticated critical appraisal of the academic and high culture representations of largely Middle Eastern societies, *Orientalism* was conceived by Said as part of a trilogy, which included two subsequent books on the media coverage of Islam and the Palestine question. Said sought to reveal classical orientalism, which modernization theory did little to contest, as the intellectual regurgitation of common stereotypes based less on social and historical realities than on the tendency of European intellectuals to imagine a distinction between Europe and the rest of the world.

While Said's analysis was directed largely at mainstream trends in Western scholarship of the Middle East, another set of scholars set out around the same time to contest the heritage of orientalism in left-leaning scholarship. The main point of contention here was Karl Marx's thesis of an 'Asiatic mode of production,' distinct from European feudalism in its propensity to stagnate after a founding phase, needing a strong central State for its maintenance, and incapable of producing mercantile surplus that could found a basis for capitalism. While his method was fundamentally different from Weber's, Marx nonetheless reached the same conclusion regarding the foreignness of capitalism to the orient in general, and thus positing the need for an external stimulus, such as European colonialism, to introduce the bases of bourgeois accumulation and, in turn, the precondition of proletarian socialism. In spite of the brevity and inconclusive nature of Marx's own remarks on this theme, which addressed sporadically the encounter between India and Algeria with European colonialism, his assessment of the lack of internal capacity of oriental societies for capitalism and thus revolution has reigned well into our times. For example, Shlomo Avineri, a Marxist Israeli political theorist, has argued that Israel, by introducing into 'traditional' Middle Eastern society the same forces of destruction of 'Asiatic' patterns, would also add a revolutionary stimulus to the sum total of productive forces in the region.

This line of thinking was contested in a number of widely-referenced works from the recent past. They include Maxime Rodinson's *Islam and Capitalism* (1973), which took both Weber and Marx to task for failing to appreciate the paramount role of the urban merchant class throughout the economic history of the Middle East, for their propensity for partial appraisal of Islamic economic doctrines, for their tendency to confuse religion and culture, and for treating Islam, in particular, as a monolithic force. Rodinson argued that the Muslim world in general had a historically strong merchant capitalist class, and that, if anything, Islam was more of an aid than an obstacle to its interests. Mahmood Ibrahim took the argument further, arguing in *Merchant Capital and Islam* (1990) that the rise of Islam itself must be understood in terms of its connection to the interests of merchant capital and long-distance trade. The argument was also made in Janet Abu-Lughod's *Before European Hegemony* (1989) where, using a world-system perspective, she shows both the centrality of the Middle East to premodern world systems and the evolution of common modes of capitalist thinking and planning among commercial classes throughout the world well before the rise of the West.

The academic attacks on the leftist variety of orientalism were sealed in a number of other works which sought to demonstrate the capitalist vitality of Middle Eastern societies in early modern times. These include Peter Gran's *Islamic Roots of Capitalism: Egypt, 1760–1840* (1979) and Surayah Faroqhi's *Towns and Townsmen of Ottoman Anatolia* (1984). In *Marx and the End of Orientalism* (1978), Bryan Turner attempts to revalidate Marxist analysis and the mode-of-production approach to Middle East studies by repudiating the Asiatic mode of production thesis altogether and presenting a more tenable historical and theoretical alternative. An overarching treatment of the subject matter from a much broader perspective is Halil Inalcik's *Economic and Social History of the Ottoman Empire: 1300–1914* (1994).

Those studies also occasion the beginning of the application of a political economy framework to the study of Middle Eastern history and society. In his pioneering *The Middle East in World Economy: 1800–1914* (1981) and subsequent works,

Roger Owen identifies major global economic factors at work in the final century of Ottoman domination, factors which reshuffled the social structure within the region. Owen argues that Europe's increased demand for agricultural commodities during that period strengthened coastal parts of the Middle East at the expense of internal ones, and produced small local elites tied to European commerce. Those elites failed, nonetheless, to invest back in agriculture and industry, thus causing a development in the direction of dependency rather than modern, locally rooted capitalism. Owen's thesis marks an avowed application of dependency theory, hitherto common in Latin American studies, to Middle Eastern studies. A similar approach which situates the history of Iran, omitted from his analysis by Owen, in the context of the world system is John Foran's *Fragile Resistance* (1993). Other political economists such as Yusif Sayigh (1979), Abbas Alnasrawi, and Jacqueline Ismail do the same for other parts of the region.

Some works in this genre stand out because they attempt to connect political economic analysis to historical formations in the culture at large. In *The Arab Nation* (1978), Samir Amin outlines the basis of a pan-Arab nation in a historical, far-flung urban network maintained by a commercial warrior class that generated the surplus upon which it depended from long-distance trade rather than from the adjoining peasantry, which remained unincorporated. This feature meant that the fate of the Arab nation would always be dependent on factors external to it, and thus explained why the Arab nation readily entered the modern era in the logic described by dependency theory.

Of particular concern in studies of the political economy of the Middle East is the question of the nature and the role of the State. The fact that for much of its history the Middle East was governed by loose, territorially expansive States rather than confined city-states meant that local civil society, social networks, and even identities could exist and evolve at a distance from State doctrines. The historian Roy Mottahedeh, for example, has shown how medieval Muslim society in Buyid Iran could function without the trappings of the State as such. While in general the term 'nation-state' would not adequately describe Middle Eastern polities, there were notable attempted exceptions. These involved projects aiming at building nation-states along the European model, as in nineteenth century Egypt and Kemalist Turkey following the dissolution of the Ottoman Empire in the aftermath of World War I. The State that Mehmed Ali set out to build in Egypt in the early nineteenth century paralleled the European State in its extension of the range of its social roles, its curtailment of clerical prerogatives, the strengthening of direct state presence in the countryside, its reliance on regular levies and taxes, and the introduction of educational and health reforms. As in Europe, that kind of State began to assume for itself many functions that had historically been within the domain of civil society.

But such state experiences tended to be the exceptions, and even the Egyptian State was unable to demonstrate credibly its sovereignty in the face of colonialism. On the other extreme of experiences of state formation is the Persian Gulf region, whose political history until modern times had known only ephemeral States at best. In *State and Society in the Gulf and the Arab Peninsula* (1990) Khaldoun H. Al-Naqeeb identified the pre-eighteenth century political economy of the Gulf as a 'natural economy,' because all its components, including long-distance trade across the Indian Ocean, lacked the services and supervision of a centralizing state apparatus. Al-Naqeeb's thesis was that such a natural economy was eventually destroyed by British imperialism, which supported the rise of new dynasties and authoritarian, rentier States. Such States thus had to be run (until today) as official or unofficial protectorates.

This reality, coupled with the newness of all Arab States in the Levant and North Africa, has led some commentators like Michael Hudson in *Arab Politics: The Search for Legitimacy* (1977) to identify the question of legitimacy as the central problem of postcolonial polities in the region. This stance essentially dismisses the relevance of much of the common practice in political science writing about the region, which generally takes States as normative categories of analysis, focuses attention on diplomatic history and games theory, and pays little attention to ideologies, social movements, political parties and organizations, and also to such questions as political alienation and, above all, legitimacy.

New directions of research thus began in the early twenty-first century to address more complicated issues regarding the nature of the State in the region. At least three directions of research can be identified in this connection. One involves projects oriented toward studying the historical sociology of the State in the Middle East, spearheaded by Elia Zureik, Talal Asad, Huri Islamoglu-Inan, and Khaled Fahmy. Another concerns civil society which, in the context of Middle Eastern studies, is usually seen as both an antidote to the State and a factor in democratizing it. Major works in this direction are led by Saad Eddin Ibrahim's series of reports on civil society and democracy in various countries of the region, and similar collections of reports edited by Jillian Schwedler and Augustus R. (*c.*1995). A third direction of research focuses on state rituals and symbolism, a good example of which is Peter Chelkowski and Hamid Dabashi's study of the Iranian post-revolutionary State, *Staging a Revolution* (1999). Also significant in this genre is Lisa Wedeen's *The Ambiguities of Domination* (1999), which employs ethnographic methods to study the nuances of popular responses to state symbolism in Syria. Earlier important studies in this genre by Sami Zubaida (1989), Ernest Gellner (1981) and Clifford Geertz focus on the state appropriation of popular beliefs, and demonstrated how the State tends to cause religious ideology and symbols to become more rigid and orthodox when it adopts them for its own use.

These studies have helped move the center of gravity in the sociological study of the Middle East away from a focus on formal institutions and historical continuities, to a focus on the determining relations between social structures and social movements. Hanna Batatu's classic *The Old Social Classes and the Revolutionary Movements of Iraq* (1978), as well his more recent work on Syria, combines Marxist and Weberian perspectives to show how modern social movements could be understood in terms of the transformation in the basis of status groups from tribal, religious, and official dignitaries to landed aristocratic and merchant capitalist classes.

Batatu's landmark study hails from a new trend which rejects a well-established view of religion as pre-constitutive and internal to the culture. That view is not only common in orientalist tracts, but also in a tradition of sociological commentary ranging from Reuben Levy to S.N. Eisenstadt,

and is well represented in the contemporary *Fundamentalisms Project*, edited by Martin Marty and Scott Appleby. More recent trends by such anthropologists as Talal Asad (1993) and Ralph Coury treat both Islam and culture as dependent variables, and stress the historical dynamism and multiplicity of both. With respect to modernity, Talal Asad shows how religion itself can foster modernizing tendencies, and Ernest Gellner, who had otherwise stressed the similarities across Muslim societies and the pre-constitutive role of religion, also comes to argue the special closeness of Islam to the spirit of modernity, and its ability to encompass various types of social structures.

Many observers, however, continue to be deeply interested in the dynamics operating between more permanent historical social structures and contemporary ideologies and social movements. Sami Zubaida's work stands out in this regard as an exposition of the complexity of determining forces, and for demonstrating the need for an integrated analytical grid of cultural, sociological, comparative, and historical knowledge to apprehend the nuances of modern movements. Other scholars emphasize the determining role of specific sub-State social formations, such as the Ottoman *Millet* system, whereby various religious communities related to the State as recognized and legally autonomous corporate bodies. Main research questions in this regard concern the extent to which such a system gave rise to or inhibited the full expression of modern movements, such as nationalism, sectarianism, or secularism.

Scholars of the region are increasingly inclined to regard seemingly 'premodern' identities expressed in the political culture of the region as actually *modern* inventions. They also tend to stress the rootedness of modern social movements in external stimuli, such as colonialism and foreign influence. An important recent example of this scholarly trend is Ussama Makdisi's *The Culture of Sectarianism* (2000). Focusing on nineteenth-century Lebanon, Makdisi shows how confessional identities along which the modern Lebanese system was built were manifestations of various processes of modernity rather than expressions of reaction to it, and how they were fostered by Western intervention rather than emerging on their own. Elizabeth Thompson's *Colonial Citizens* (2000) shows how it was the intrusion of colonial powers, rather than indigenous traditionalism, that had been detrimental to both women and liberal thought, generally in Syria and Lebanon. Lacking roots in the colonies and conceiving of them as largely tradition-bound territories, colonial powers weakened indigenous liberalism, because the latter adhered to anticolonial nationalism, and strengthened, through its compensatory alliance with them, the more conservative forces, whose influence had been declining before the advent of colonialism.

Many commentators also view the emergence of modern nationalism, in particular, in the context of colonialism rather than that of indigenous developments. Because of its extremity, the case of the Palestinians is usually cited as the most evident, but by no means the only, illustration of this centrality of external intervening factors to the framing of modern identities. An important study in this regard is Rosemary Sayigh's *Palestinians: From Peasants to Revolutionaries* (1979), which uses oral history to show the growth of nationalism due to external colonizing pressure. A more recent work along these lines is Rashid Khalidi's *Palestinian Identity* (1997), which shows the gradual emergence of that identity in reaction to Jewish nationalism in Palestine. Similarly, in examining the Palestinian *Intifida*, Samih Farsoun and Jean Landis reject the relevance of most trends in social movement theory, emphasizing instead on the global context of the Palestinian struggle.

The pivotal geopolitical status of the modern Middle East, which entailed massive foreign intervention and a plethora of movements of resistance as well as political fragmentation, gave rise to a rich tradition of engaged scholarship on social movements. Khalil Nakhleh (1977) a Palestinian sociologist, has rejected outright any possibility of nonreflexive, detached, 'normal,' or 'nonnative' sociology under current circumstances in the region. Other scholars have sought to outline oppositional social movements in terms which endow them with an under-explored sense of agency. Examples of this approach include Joel Beinin and Zachary Lockman's *Workers on the Nile* (1987), and Anouar Abdel-Malek's *Contemporary Arab Political Thought* (1983), which moves away from cultural and structural determinism altogether and stresses the priority of agency under modern circumstances.

Social Movements and Social Change

Apart from questions of colonial encounter and political economy, an important line of emphasis in studying social movements in the Middle East concerns how they make use of cultural symbols. Julia Clancy-Smith's *Rebel and Saint* (1994) shows how resistance to colonialism in North Africa used icons and authorities from traditional settings whose function in society had until then been quite apolitical. In a nuanced analysis of the new veil and work environments in Cairo, Arlene Elowe Macleod (1991) shows how the veil today, far from being simply a patriarchal imposition, is actually used proactively by women to express various messages combining both accommodation and protest.

These studies stem from a very weighty orientation in Middle East studies, particularly in Middle Eastern countries, toward questions of heritage and its connection to modernity. Many important contemporary commentators, such as Fazlur Rahman, Seyyed Hossein Nasr, Syed M.N. al-Attas, and Ismail al-Faruqi have revisited classical Islamic philosophy and science in a quest to reveal a specifically Islamic philosophy of knowledge and science in tune with modernity. More politically inclined contemporary philosopher-activists (frequently Shi'ites), such as Ali Shariati, Mohammed Baqr al-Sadr, and Morteza Motahhari, have also revisited classical sources extensively in order to show the compatibility of Islam with modern ideologies, notably socialism and anticolonial nationalism. The same is true of Tayyib Tazzini's and Hussein Muruwwah's secular reading of the heritage. The voluminous contributions of the latter who was killed by Amal fighters in Beirut in 1985 were geared toward showing how classical Islamic philosophy and thought had always housed 'materialist' perspectives and methods in tune with Marxist dialectics.

Many other commentators are equally interested in voluminous, encyclopedic surveys of classical thought, but emphasize organizing principles as discursive, semiological, and hermeneutic rather than being socio-historical approaches. These include Muhammad Abid al-Jabiri's analysis over many volumes of the structures of knowledge in Arab culture and

poet Adonis's (Ahmad Ali Sa'id) equally critical study of imitation and originality in Arab culture, *The Permanent and the Changeable* (1974). While many of these studies seek to discover through historical investigation a root cause of contemporary 'stagnation' or 'backwardness,' many more seek to reveal through the same method the dynamism and adaptability of the heritage to various social structures and ideological claims. This latter approach to the heritage had required an abandonment of the philological and preconstitutive approach of orientalism. Among the most important founders of this critical approach is Muhammad Arkoun, whose pioneering studies of the Qur'an have emphasized the value of other than purely philological approaches to the holy text, such as anthropological, sociological, and semiological approaches. Mohammed Bamyeh's *The Social Origins of Islam* (1999) builds on that tradition of combined methods of reading, arguing that the growth of Islam was precisely due to an ability exhibited at its founding phase to accommodate the worldviews of different and even discordant social classes and types of social structure. Abdul-Kabir al-Khatibi echoes a similar theme when he argues that Islam must be seen as a sort of 'compromise' religion, and that its historical vitality stems from that attitude rather than from an alleged orthodox base and orientation.

In light of these new perspectives, the question arises again as to why this demonstrated dynamism of the heritage has not lent itself to a more vocal communion with liberal European modernity. Many responses to this question tend to dismiss it outright, citing as its faulty premise a Eurocentric tendency to outline various historical trajectories as more sharply differentiated than they actually were. In his critique of Eurocentrism, Samir Amin builds on and extends Martin Bernal's thesis in *Black Athena* (1987). Amin contests European fabrication of a monopolistic claim on Greek heritage, as well as the construction of European history through a direct line that flows from Greece to Rome. This perspective, according to him, ignores the larger Mediterranean contexts of classic civilization and the continuity and further development of Greek political and philosophical thought, as well as mathematics, medicine, and science in the Islamic world during Europe's Middle Ages.

Criticism of the tendency to view Middle Eastern or Islamic history in terms too distinct from other world histories has also been levied by the urban historian and sociologist Janet Abu-Lughod (1989) who, after her initial study of Cairo, came to repudiate the notion that there was anything specifically Islamic about the 'Islamic city.' For others, such as Jabiri and Hichem Djait (1974) the point became less relevant, as they identify the important lines of demarcation with respect to modernity not along large East/West types of categories. Rather, they prefer subregional categories, such as Arab Mashreq (Western Asia and Egypt), versus Arab Maghreb (rest of North Africa). Djait argues that the relatively more intense and enduring colonialism experienced by the Maghreb caused a more cultural proximity to modernity, since all illusions about self and society were destroyed through a colonial uprooting that had not been as severe in the Mashreq. The latter, being politically more advanced, had obtained its independence earlier than the Maghreb had and proceeded to be governed by dictatorial regimes. That condition, according to Djait, favored the construction of more illusions about self and society in the Mashreq, and thus more relative distance from the spirit of modernity.

Within discussion of the proper attitude to modernity, the question of gender has had a particular resonance. The feminist sociologist Fatima Mernissi attained fame by blaming Islam itself for perpetuating a retrograde attitude toward women. An alternative trajectory within feminist writings on Islam, represented in the works of Leila Ahmed, Nawal Sa'dawi, and Denise Spellberg, offers more complex readings of the role of women in Islamic history and places more emphasis on the agency of women. More recent trends have de-emphasized the alleged role of traditional religion even further as an explanatory factor. Deniz Kandiyoti's (1991) work has contributed to shifting attention away from religion as such and into patriarchal structures as sources of the oppression of women. Hisham Sharabi's pioneering study *Neopatriarchy: A Theory of Distorted Change in Arab Society* (1988) singles out structural changes within the Arab World as well as elsewhere within the past century that had led to 'neopatriarchy' instead of modernity. Neopatriarchy is explained by Sharabi as a form of patriarchy produced by dependent capitalism and social change lacking in a sense of authenticity. Thus neopatriarchy tends to distort modernization and nervously presents itself as the only valid option.

Given the magnitude of social, political, economic, and cultural transformations in the Middle East over the twentieth century, many social scientists will find some literary works very useful in apprehending change in certain areas. Three particular novelists stand out for their ability to reconstruct history and society in more overarching and encyclopedic ways than has so far been possible in the social sciences, and thus are worthy of greater attention. Adbelrahman Munif's (1989) multivolume epic *Cities of Salt* reconstructs in elaborate and patient detail the coming into being of Saudi Arabia in the 1930s, following the discovery of oil in the country by Western companies. The novel documents transformations in social relations and customs, the introduction of wage labor on top of an austere, seminomadic economy, the emergence of a dynasty and a State, the resulting tensions between authenticity and inauthenticity in public imagination, and other fascinating details which are usually glossed over in more dispassionate sociological and historical writings.

Gamal Ghitani's historical novel *Zayni Barakat* (1990), set at the time of Ottoman's takeover of Egypt from the Mamluks early in the sixteenth century, reconstructs elements of continuity and change in political structure as well as in the relationship between spiritual and political authority over the time-span of the two systems. Finally Nobel laureate Naguib Mahfuz's prolific career is dedicated to an elaborate and highly nuanced reconstruction of social psychology and its connection to mores and traditions. Characters in Mahfuz's works, who frequently mimic middle- and lower-middle class inhabitants of Cairo's neighborhoods, are usually shown as agents of their own destinies rather than as victims of established structures, and are capable of using a wide reservoir of proverbs and customary allowances of deviance to justify their actions. This holds true even in their attitude toward fate. In analyzing one famous story by Mahfuz, Halim Barakat shows that when characters claim what they are doing to be ordained by some fate, they are usually using the language of fatalism as a tool of

self-assertion, rather than as an explanation of their powerlessness with respect to established patterns of relations and social expectations. At both the macro and micro levels, these works combine research orientations central to understanding Middle Eastern society and history. These orientations involve political economy, social and institutional structures and mentalities, and the flexible morality governing everyday street-level interactions.

Conclusion

The study of society and history of the Middle East has gradually moved from focus on homogeneity and continuity to more elaborate discussions of the impact of many determining factors. It is now commonly accepted that due to its historical centrality to historical world systems and its contemporary geopolitical significance, Middle Eastern society and history cannot be understood apart from global trends which influence the region and are influenced by it as well. The studies are most revealing when they go above or below contemporary state jurisdictions and emphasize instead on the historically autonomous mode of self-organization of civil society, cultural communities, travel, and trade. Middle Eastern society has historically been defined more so by the hybridity and diversity resulting from the above factors, than by the social engineering of States.

See also: Islam: Middle East; Islamic Fundamentalism; Modernization, Sociological Theories of; Nationalism, Historical Aspects of: Arab World; Near Middle East/North Africa Studies: Culture; Near Middle East/North Africa Studies: Politics; Near Middle East/North Africa Studies: Religion; Orientalism.

Bibliography

Abu-Lughod, J.L., 1989. Before European Hegemony. Oxford University Press, New York.
Al-Azmeh, A., 1993. Islams and Modernities. Verso, London.
Al-Naqeeb, K.H., 1990. Society and State in the Gulf and Arab Peninsula. Routledge, London.
Amin, S., 1978. The Arab Nation. Zed Press, London.
Arkoun, M., 1982. Lectures du Coran. Editions G.-P. Maisonneuve et Larose, Paris.
Asad, T., 1993. Genealogies of Religion. Johns Hopkins University Press, Baltimore, MD.
Bamyeh, M.A., 1999. The Social Origins of Islam. University of Minnesota Press, Minneapolis, MN.
Barakat, H., 1990. Beyond the always and the near: a critique of social psychological interpretations of Arab society and culture. In: Sharabi, H. (Ed.), Theory, Politics and the Arab World: Critical Responses. Routledge, New York.
Batatu, H., 1978. The Old Social Classes and the Revolutionary Movements of Iraq. Princeton University Press, Princeton, NJ.
Beinin, J., Lockman, Z., 1987. Workers on the Nile: Nationalism, Communism, Islam, and the Egyptian Working Class, 1882–1954. Princeton University Press, Princeton, NJ.
Chelkowski, P., Dabashi, H., 1999. Staging a Revolution: The Art of Persuasion in the Islamic Republic of Iran. New York University Press, New York.
Clancy-Smith, J.A., 1994. Rebel and Saint: Muslim Notables, Populist Protest, Colonial Encounters. University of California Press, Berkeley, CA.
Djait, H., 1974. La Personnalité et le Devenir Arabo-islamiques. Editions du Seuil, Paris.
Fahmy, K., 1997. All the Pasha's Men: Mehmed Ali, His Army, and the Making of Modern Egypt. Cambridge University Press, Cambridge, UK.
Faroqhi, S., 1984. Towns and Townsmen of Ottoman Anatolia. Cambridge University Press, Cambridge, UK.
Foran, J., 1993. Fragile Resistance: Social Transformation in Iran from 1500 to the Revolution. Westview Press, Boulder, CO.
Ghitani, G., 1990. Zayni Barakat. Penguin Books, New York.
Gellner, E., 1981. Muslim Society. Cambridge University Press, Cambridge, UK.
Gran, P., 1979. Islamic Roots of Capitalism: Egypt, 1760–1840. University of Texas Press, Austin, TX.
Hourani, A., 1991. History of the Arab Peoples. Belknap Press of Harvard University Press, Cambridge, MA.
Hudson, M., 1977. Arab Politics: The Search for Legitimacy. Yale University Press, New Haven, CT.
Ibrahim, M., 1990. Merchant Capital and Islam. University of Texas Press, Austin, TX.
Inalcik, H., 1994. An Economic and Social History of the Ottoman Empire: 1300–1914. Cambridge University Press, Cambridge, UK.
Islamoglu-Inan, H. (Ed.), 1987. The Ottoman Empire and the World-Economy. Cambridge University Press, Cambridge, UK.
Kandiyoti, D. (Ed.), 1991. Women, Islam, and the State. Temple University Press, Philadelphia, PA.
Khalidi, R., 1997. Palestinian Identity. Columbia University Press, New York.
Lerner, D., 1958. The Passing of Traditional Society. Free Press, Glencoe, IL.
Makdisi, U.S., 2000. The Culture of Sectarianism. University of California Press, Berkeley, CA.
Macleod, A.E., 1991. Accommodating Protest: Working Women, the New Veiling, and Change in Cairo. Columbia University Press, New York.
Mottahedeh, R.P., 1980. Loyalty and Leadership in an Early Islamic Society. Princeton University Press, Princeton, NJ.
Munif, A., 1989. Cities of Salt. Vintage Books, New York.
Nakhleh, K., 1977. Anthropological and sociological studies on Arabs in Israel: critique. Journal of Palestine Studies 6 (4), 41–70.
van Nieuwenhuijze, C.A.O., 1971. Sociology of the Middle East. Brill, Leiden, The Netherlands.
Owen, R., 1981. The Middle East in World Economy: 1800–1914. Methuen, London.
Rodinson, M., 1973. Islam and Capitalism. Pantheon Books, New York.
Said, E.W., 1978. Orientalism. Pantheon Books, New York.
Sayigh, R., 1979. Palestinians: From Peasants to Revolutionaries. Zed Press, London.
Sharabi, H., 1988. Neopatriarchy: A Theory of Distorted Change in Arab Society. Oxford University Press, New York.
Thompson, E., 2000. Colonial Citizens: Republican Rights, Paternal Privilege, and Gender in French Syria and Lebanon. Columbia University Press, New York.
Turner, B., 1978. Marx and the End of Orientalism. Allen & Unwin, London.
Wedeen, L., 1999. The Ambiguities of Domination: Politics, Rhetoric, and Symbols in Contemporary Syria. University of Chicago Press, Chicago.
Zubaida, S., 1989. Islam, the People and the State: Essays on Political Ideas and Movements in the Middle East. Routledge, London.
Zureik, E., 1981. Theoretical considerations for a sociological study of the Arab state. Arab Studies Quarterly 3 (3), 229–257.

Near-Infrared Spectroscopy

Judit Gervain, Centre National de la Recherche Scientifique (CNRS), Université Paris Descartes, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Near-infrared spectroscopy (NIRS) is a relatively new and increasingly widespread brain imaging technique, particularly suitable for use with young infants. The technique employs near-infrared light to assess the concentration changes of oxygenated and deoxygenated hemoglobin, accompanying local brain activity. The basic physical, physiological, and neural principles underlying the use of NIRS and some of the existing developmental studies are reviewed. Issues concerning technological improvements, parameter optimization, possible experimental designs, and data analysis techniques are also discussed and illustrated.

Introduction

One of the greatest advances in the recent history of the social and cognitive sciences has been the refinement and increasingly more versatile application of brain imaging techniques. Different methods measuring the electrophysiological (e.g., electroencephalography (EEG), magnetic encephalography (MEG)) or metabolic (e.g., magnetic resonance imaging (MRI), positron emission tomography (PET), near-infrared spectroscopy (NIRS)) aspects of neural activity have been used with a wide range of healthy and clinical populations to explore brain organization and function noninvasively in behaving participants.

NIRS is one of the most recently developed such technique. It consists of using near-infrared light to assess the concentration changes of oxygenated and deoxygenated hemoglobin (oxyHb and deoxyHb, respectively), accompanying local brain activity. Although the principles of using light to image the brain had been first laid down almost 40 years ago (Jobsis, 1977), the technique has only been used in cognitive neuroscience research for the last 15–20 years. While it can be applied to participants of any age, NIRS is most commonly used for brain imaging in the youngest developmental populations, including preterm and full-term newborns and young infants. Several of its features make it an ideal method for these age groups. Infants' tissues, e.g., skin, skull, etc., are thinner than those of older children and adults, allowing NIRS to probe deeper into the brain in these young participants. NIRS is also highly noninvasive and safe: it does not require the use of any isotope, carrier substance, gel, or other contact agent. It is noiseless and involves no strong magnetic field. It is easy to use and apply: the headgear containing the optical sensors is usually a comfortable, soft cap which can be placed on participants' heads in a matter of a few seconds. It is also highly tolerant to movement and motion artifacts, a constant problem for imaging participants who cannot easily follow verbal instructions. Most NIRS systems are small and often portable (some even have stand-alone headgear that wirelessly transmits information to the data acquisition system), allowing bedside or in-the-home testing. Mainly relying on the high motion tolerance and portability of NIRS, additional applications include clinical monitoring and bedside testing of brain function in patients who cannot be transferred to other imaging facilities, like a PET or an MRI scanner, as well as probing brain function in participants practicing sports or performing other physical action.

While a relatively recent imaging technique, NIRS has already provided significant empirical contributions to our understanding of cognitive and neural development from birth through infancy and early childhood. These, reviewed in detail below, include the origins of the brain specialization for speech and language processing (e.g., Gervain et al., 2008; Minagawa-Kawai et al., 2011a; Peña et al., 2003; Telkemeyer et al., 2009), the neural correlates of the development of face perception and social cognition (e.g., Blasi et al., 2007; Csibra et al., 2004; Lloyd-Fox et al., 2009b; Otsuka et al., 2007), multisensory processing (e.g., Bortfeld et al., 2007; Taga and Asakawa, 2007), and the assessment of brain function and oxygenation in preterm infants and infants having suffered perinatal trauma (e.g., Chen et al., 2002; Mahmoudzadeh et al., 2013).

Below, the basic principles and different technological implementations of near-infrared optical imaging will be discussed first. NIRS will then be compared to other brain imaging modalities. This will be followed by an overview of the existing NIRS literature, focusing mostly on developmental work. Next, technical issues relating to the practical application of NIRS will be discussed. This will be followed by a discussion of experimental designs and data analysis techniques suitable for NIRS. Finally, future directions will be outlined.

The Basic Principles of NIRS

Neural activity is accompanied by changes in blood oxygenation, which can be detected by near-infrared light. Optical imaging is thus an indirect measure of neural activity (with the exception of fast NIRS, see below), like fMRI, but unlike more direct electrophysiological measures such as EEG and MEG. The following review articles provide more detailed discussions about the physical and physiological principles underlying optical imaging (Aslin and Mehler, 2005; Chance et al., 1992; Ferrari et al., 2004; Gervain et al., 2011; Hiraoka et al., 1993; Lloyd-Fox et al., 2010; Minagawa-Kawai et al., 2008; Villringer and Chance, 1997). Below, we provide a brief overview of the most basic principles.

Continuous Light Optical Imaging

The most commonly used NIRS technique relies on continuous wave (CW), monochromatic near-infrared light to monitor the changes in the concentration of oxyHb, deoxyHb, or cytochrome *c* oxidase in brain tissue, related to neural activity.

When monochromatic light travels through a medium, some of it is absorbed, some of it is scattered, and some of it is transmitted, i.e., continues its trajectory unaffected by the medium. Exactly what proportions of the light are absorbed, scattered, and transmitted depend on the properties of the medium (e.g., absorption coefficient, concentration, etc.) and the light (e.g., wavelength, etc.). Biological tissues like the skin, the skull, and the brain are highly scattering media, causing light to travel in a nonstraight line. A physical law, known as the modified Beer–Lambert law, describes the relation between the relevant factors:

$$A = -log(I/I_0) = (c \times \varepsilon_\lambda \times l \times DPF) + G \qquad [1]$$

where A is absorbance, I is the intensity of the transmitted light, i.e., the light after the medium, I_0 is the intensity of the incident light, i.e., the light before the medium, c is the concentration or density of the medium, ε_λ is the molar extinction coefficient characteristic of the medium for a light of wavelength λ, l is the length of the medium in the dimension along which light travels, DPF is the differential path length factor that accounts for the nonlinear trajectory of light in biological media, and G is the scatter. Given this equation, the concentration of the medium can be obtained by measuring the intensity of the light that leaves the medium, assuming that the original light intensity, the length of the medium, and the molar extinction coefficient are known. Since the concentrations of oxyHb and deoxyHb in brain tissue are indicators of neural activity, we can use the above equation to calculate their concentrations by shining light of an appropriate wavelength and a known intensity on the head and measuring the intensity of the exiting light.

Using CW techniques, the actual scatter and the DPF cannot be measured directly. As a consequence, absolute values for the concentrations of oxyHb and deoxyHb cannot be obtained. However, the scatter is constant and can thus be mathematically eliminated when changes in the concentration of oxyHb and deoxyHb are calculated.

$$\Delta A = \left(\left(\Delta c_{oxy} \times \varepsilon_{\lambda_{oxy}}\right) + \left(\Delta c_{deoxy} \times \varepsilon_{\lambda_{deoxy}}\right)\right) \times l \times DPF \qquad [2]$$

The molar extinction coefficients of oxyHb and deoxyHb are shown in Figure 1.

The red and near-infrared range of the spectrum, encircled in Figure 1, contains the most adequate wavelengths for the optical imaging of biological media, because intervening tissues (e.g., skin, bone) are relatively transparent to light at these wavelengths. As mentioned before, the CW technique cannot directly measure the DPF, although estimates exist for the newborn and infant brain (Duncan et al., 1996; Wyatt et al., 1990). In addition, the distance that the light travels in the brain depends on l, which, in the case of NIRS applied to the human head, corresponds to the separation between the light source and the light detector locations on the surface of the scalp. Earlier NIRS systems usually applied one source and one detector, or optode, pair, constituting one measurement channel. Since then, multichannel systems have appeared using several sources and detectors.

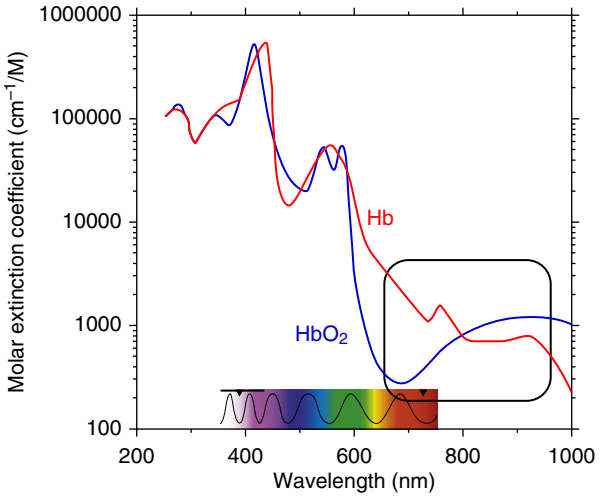

Figure 1 The molar extinction coefficients of oxyHb and deoxyHb for different wavelengths of light. Encircled in black is the region of the spectrum that is most relevant for NIRS, with the isosbestic point of oxyHb and deoxyHb.

The availability of multiple channels has given rise to two distinct types of techniques: optical topography and optical tomography. The former provides a two-dimensional sampling from the surface of the cortex, whereas the latter, rarely used in experimental research due to technical limitations, allows a three-dimensional reconstruction of the hemodynamic signals from the brain. In topography systems, the coupled sources and detectors forming channels are located at a distance of a few centimeters from one another and the light travels through a banana-shaped trajectory (Figure 2), penetrating the surface of the cortex at a depth of a few millimeters. In young infants, whose surface tissues are thin, NIR light can penetrate as deep as 30 mm into the head (measured from the scalp surface), thus 10–15 mm into the cortex at a source–detector separation of 3 cm, whereas in adults, the penetration into the cortex is only about 3–5 mm deep with the same separation.

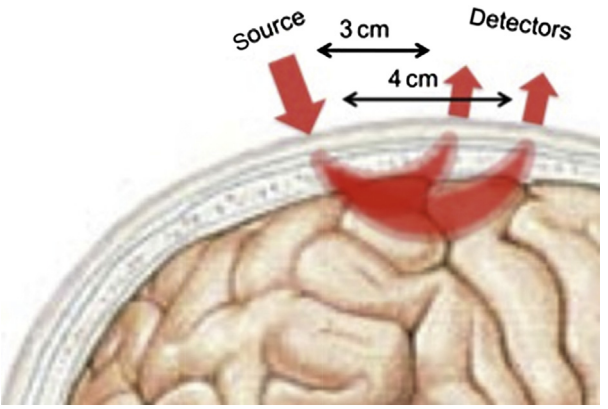

Figure 2 A schematic illustration of the trajectory of near-infrared light in the human head at different source–detector separations.

CW systems thus use changes in oxyHb, deoxyHb, and totalHb (the sum of oxyHb and deoxyHb) concentrations as an indirect measure of brain activity. This is possible because there is a relationship between focal brain activity and regional cerebral blood volume and flow. This neurovascular coupling means that increased activity, requiring additional metabolic supplies, e.g., oxygen and glucose, is accompanied by local vasodilation, increased blood flow, and oxygenation. Crucially, the increase in blood flow and oxygen exceeds the demands, resulting in a local and transient excess of oxygen. This metabolic response, known as the hemodynamic response function (HRF), follows local neural activity by several seconds and the exact physiological mechanisms relating the HRF to neural firing are not fully understood (Logothetis et al., 2001; Magistretti, 2006). While the shape of the HRF for adults is well described (an increase in oxyHb concentration several seconds after stimulation, a somewhat later-peaking decrease in deoxyHb concentration, and a slow return to baseline, with possible over- and undershoots, for both Hb species; Boden et al., 2007; Obrig et al., 2000), less is known about its shape in developmental populations. The HRF might be delayed, slower to peak, and/or slower to return to baseline in infants than in adults (Meek, 2002). The slowness of the HRF needs to be taken into account when designing NIRS experiments (see below).

Measuring relative oxyHb and deoxyHb concentrations is usually sufficient for research-related applications of NIRS. However, whenever absolute concentration values are relevant, e.g., in clinical settings, non-CW techniques, developed to estimate the path length and/or the scatter, need to be used. Two such methods include the intensity-modulated or frequency-resolved technique and the time-of-flight or time-resolved technique.

Frequency-Resolved Optical Imaging

In frequency-resolved NIRS, the intensity of the light is modulated at a certain frequency. While traveling through tissues, the intensity is attenuated and the phase of the modulation is shifted due to scattering. By measuring the attenuation and the phase-shift, the path length and the scatter can be determined, allowing the calculation of absolute oxyHb and deoxyHb concentration values (Chance et al., 1990). There are commercially available systems based on this approach (for an overview of commercially available systems, see Lloyd-Fox et al., 2010; Wolf et al., 2007). However, they typically have lower sampling rates and higher costs than the CW systems.

Time-Resolved Optical Imaging

Another approach for determining the path length of light in biological tissue is to use single light pulses instead of continuous light. When a very short pulse of light is sent through tissues, the time distribution of individual photons that leave the head can be registered by a detector or photon-camera. The average time-of-flight of the photons is then multiplied by the speed of light to determine the mean path length (Delpy et al., 1988; Lloyd-Fox et al., 2010; Wolf et al., 2007). This technique offers slower data acquisition than most CW systems.

Fast Optical Imaging

Unlike the previous three techniques, which measure the metabolic correlates of brain activity, fast optical imaging (also known as event-related optical signal, EROS) detects the changes in the optical properties of neurons and surrounding tissue during firing (Gratton and Fabiani, 2003, 1998; Tse et al., 2007). As this measure is a correlate of the electrophysiological activity of neurons and not of metabolism, it is a much faster signal, with a latency in the millisecond range, comparable to that of electrophysiological techniques. This fast signal and the slow metabolic response can be detected using the same equipment, which could provide an ideal method to characterize neural activity with high temporal and spatial resolution at the same time. However, the reliability and strength of this fast optical signal as measured noninvasively in behaving adult subjects may be limited, at least with the currently available technologies (Steinbrink et al., 2005). The remainder of this article will concentrate on the three metabolic NIRS techniques and will mostly ignore fast NIRS.

NIRS as Compared to Other Imaging Techniques

When compared to other brain imaging and electrophysiological techniques, NIRS has several clear advantages, in particular when used with infants. Unlike the magnetic gradients used in MRI, NIRS is completely silent, providing a nonintrusive environment and allowing for an easy presentation of auditory stimuli. No strong magnetic field or radio frequency pulses are involved, alleviating safety concerns. In terms of the obtained signal, NIRS has the advantage of measuring both oxyHb and deoxyHb changes, providing a physiologically plausible measure of blood flow and volume, while the blood oxygen level dependent (BOLD) response in fMRI is related to deoxyHb only. The costs of NIRS are also considerably lower than those of MRI.

As compared to EEG, NIRS provides better spatial localization, as most topographic implementations of NIRS are not subject to the inverse problem associated with source localization of electrical potentials from the scalp. NIRS is also less sensitive to motion artifacts than both MRI and EEG. While the time resolution of NIRS is lower than that of EEG, the sampling rate of most CW machines is around 10 Hz compared to 0.5 Hz in fMRI, allowing for an improvement in temporal resolution, once mathematical algorithms are in place to adequately analyze event-related data. PET, single-photon emission computed tomography (SPECT), and other nuclear imaging techniques require the use of isotopes and carrier substances, raising safety concerns for vulnerable populations.

The key limitation of NIRS is that it only probes the surface layers of the cortex. Consequently, brain structures that lie deeper in the cortex or below it are not visible in NIRS topography techniques. In addition, the temporal resolution of NIRS is lower than that of EEG, rendering the detection of fast responses to single events more challenging. Similarly, the spatial resolution of NIRS is inferior to that of MRI.

An exciting new avenue to overcome some of the shortcomings of NIRS, e.g., its low temporal resolution, while taking full advantage of its strengths, is to combine it with other measures, such as EEG. Indeed, several recent NIRS–EEG

studies (Mahmoudzadeh et al., 2013; Telkemeyer et al., 2011, 2009), mainly investigating speech perception in newborns and infants, have proven the feasibility of such corecording approaches.

What NIRS Has Taught Us: A Brief Summary of the Existing NIRS Literature

As discussed above, the advantages of NIRS make it an ideal tool for brain imaging in young infants, whereas adults can be safely and successfully imaged using techniques that are more constraining, but offer better spatial or temporal resolution. Consequently, NIRS has mainly been used in developmental neuroscience research in the last 15 years. A brief and non-exhaustive summary of this body of literature is provided below.

The Origins of the Functional Specialization of the Brain: Localization of Perceptual Abilities

The first studies used NIRS to establish the developmental origins of the brain specialization for different perceptual modalities. In one of the earliest studies with newborns, Sakatani et al. (1999) registered increased activation in the bilateral frontal lobes while infant participants listened to music. NIRS has also proven useful to register orbitofrontal brain activation in newborns during olfactory stimulation (Bartocci et al., 2000). Newborns between 6 h and 8 days old showed increased response to a vanilla scent as well as to the odor of their mother's colostrum as compared to water. The intensity of the response to colostrum attenuated with post-natal age, demonstrating the sensitivity of NIRS to detect a graded response. Probing visual perception, Taga et al. (2003) found an increased response in the occipital areas of 2–4-month-old infants while the infants were watching checkerboard pattern reversals. When measuring infants' responses to multiple sensory stimulation, studies (Bortfeld et al., 2007; Taga and Asakawa, 2007) found that NIRS was sensitive enough to detect brain responses specific to auditory stimulation while a simultaneous visual stimulus was also present, suggesting that NIRS has sufficient selectivity to localize neural activity triggered by different perceptual modalities. Similar results were later obtained by Bortfeld et al. (2009), who showed that the response to speech, even when presented together with visual stimulation, was left-lateralized in 6–9-month-old infants.

Early Lateralization and Functional Specialization for Language

The origin and developmental trajectory of the left lateralization of language, found in most right-handed adults (Kimura, 1967), is another functional, localizational question that has been addressed using NIRS. In a pioneering study, Peña et al. (2003) compared newborns' responses to three blocks of stimuli: forward-going infant-directed speech in the native language, the same stimuli played backwards, and silence. Newborns showed increased activation in the left temporal area in response to forward speech, but not to backward speech or silence. Similar results were obtained in an fMRI study with 3-month-old infants (Dehaene-Lambertz et al., 2002). Whether this lateralization is the reflection of familiarity with the native language due to exposure *in utero* or whether the left hemisphere is prepared to process language stimuli in general, including unfamiliar, nonnative languages, is currently heatedly debated. One line of research addressing this issue tested the processing of native and nonnative languages in newborns using NIRS. Some of these studies replicated the left lateralization for the native language, but found bilateral activation to a nonnative language (Sato et al., 2012), whereas other studies obtained bilateral activation for both the native and a nonnative language, with a difference only in response intensity in favor of the native language (May et al., 2011). A second line of research has focused on the physical/acoustic properties of the speech signal that might drive specialization. Different theoretical accounts (Hickok and Poeppel, 2007; Poeppel, 2003; Zatorre and Gandour, 2008; Zatorre et al., 1992) suggest that the temporal and spectral properties of speech might play an important role: the right hemisphere is believed to integrate over long time windows and/or spectral information, whereas the left hemisphere or both hemispheres might be responsible for precise temporal processing. Recent NIRS studies in young infants have provided mixed support for these hypotheses. Newborns, 3- and 6-month-old infants showed bilateral activation for fast-modulated stimuli and a right hemisphere advantage for slowly modulated sounds in an NIRS–EEG corecording study (Telkemeyer et al., 2011, 2009). However, no significant response was obtained for spectrally modulated stimuli (Minagawa-Kawai et al., 2011b). Thus the developmental origins of the brain specialization for speech processing remain an open question.

A different set of studies has addressed the lateralization of individual linguistic properties and language-specific phonological contrasts. These have been found to take several months of experience with the native language to lateralize, mirroring the behavioral transition in language perception during the first year of life from general, broad-based abilities toward enhanced sensitivity to the specifics of the native language (known as perceptual attunement). A study by Minagawa-Kawai et al. (2007) illustrates this attunement using the short–long vowel contrast in Japanese. Infants show a U-shaped discrimination curve, distinguishing the two categories at around 6–7 months and after 13–14 months, but not at around 10–11 months. Importantly, left lateralization in response to these stimuli, which is the typical adult pattern, appears only at around 13–14 months and approximates the exact adult spatial distribution after 25–28 months. These results suggest that left lateralization for the short–long vowel contrast emerges as the perception of the contrast is fine-tuned, shifting from a simple acoustic difference to a language-relevant phonological property. Another example comes from work by Sato et al. (2010), who investigated the development of the behavioral and neural response to Japanese pitch-accent patterns and their pure tone equivalents, which served as a nonlinguistic control. Behaviorally, both 4- and 10-month-old infants could discriminate the different pitch-accent patterns (high-low vs low-high). However, at 4 months, infants did not

show different NIRS responses to real Japanese words and their pure tone equivalents, suggesting that processing is acoustic/nonlinguistic at this age. At 10 months, by contrast, infants showed greater activation to a change in word pitch-accent patterns than to a change in the corresponding pure tone patterns, and this differential response was localized in the left hemisphere, indicating that processing was linguistic in nature.

A similar trend was found for the perception of prosody (the melodic property of speech that includes variations in pitch, intensity, and duration). Results with newborns (Saito et al., 2007a) indicate that they show increased frontal response to infant-directed speech as compared to adult-directed speech, allowing infants to tune into speech addressed to them early on. A key difference between the two speech registers is the highly variable, exaggerated prosody in infant-directed speech, which might explain the more prominent activation found for this stimulus. Indeed, in another study, the same authors (Saito et al., 2007b) showed that newborns can discriminate between normally intonated prosody and monotone speech with flat prosody, as they show increased activation in the bilateral frontal areas to the former, but not to the latter. While the frontal areas do not show lateralization at this early stage, Homae et al. (2007, 2006) found that at 3 months, infants show greater activation in the right temporoparietal areas to normally intonated speech as compared to speech with flattened prosody. Wartenburger et al. (2007) provided converging evidence for a hemispheric specialization to different aspects of speech processing. Testing 4-year-old children, they found that when prosody is presented devoid of linguistic content, triggering nonlinguistic processing, the right frontotemporal areas are involved, whereas presenting the same prosody together with the corresponding linguistic content engages the left hemisphere.

Investigating when the ability to process the structural properties of language begin, Gervain et al. (2008) found increased neural activity in the bilateral temporal and left frontal areas of newborns when they listened to speech that followed a structural rule (ABB: e.g., 'mubaba') as compared to unstructured speech (ABC: 'mubage'). The activation was more pronounced in the left than in the right temporal area, reproducing previous lateralization results. These results have been expanded by Gervain et al. (2012), who investigated the perception of sequence-initial versus sequence-final immediate repetitions (AAB vs ABB) and found that they were discriminated by the newborn brain.

Recent studies on language processing have also used NIRS systems with a large number of channels to establish functional connectivity and to identify the language network within the brain. In two studies, Homae et al. (2011, 2010) have tracked the development of global cortical networks at rest, as well as during and after speech stimulation at birth, 3 and 6 months. They have found that resting state connectivity increases within and between hemispheres in the occipital, parietal, and temporal lobes, but reduces in the frontal lobe over development. During speech stimulation, 3-month-olds showed increased connectivity between the channels in the temporal and temporoparietal, in particular within the same hemisphere. Interestingly, higher connectivity was maintained between the frontal and temporal areas during rest following speech stimulation as compared to the resting period before stimulation, indicating the existence of a memory/storage network.

Cognitive Development

Nonlinguistic cognitive functions have also been explored using NIRS. One of the first studies (Baird et al., 2002) investigated the neural correlates of object permanence in a longitudinal study with 5–12-month-old infants. Infants were tested behaviorally for the presence of object permanence while neural activity in their frontal cortex was recorded using NIRS. NIRS data from the session where the infant first showed object permanence behaviorally were compared to NIRS data from the preceding session. Increased activation was observed in the object permanence sessions as compared to the prepermanence ones. Object processing was also assessed by NIRS (Wilcox et al., 2009, 2008). Infants were presented with physically possible and impossible events involving object movement and identity. The possible event involved two different objects emerging from either side of a wide occluder, whereas the impossible event involved the same two objects emerging from behind an occluder that was too narrow to hide two objects. The featural differences between the objects were systematically manipulated. The objects either differed in multiple features, or only shape or color, or they were identical. In the occipital areas, all four conditions gave rise to similar activations, while in the inferior temporal areas, only the multifeatural and shape change conditions evoked increased responses.

In the social domain, Csibra et al. (2004) found a difference between neural activation in the occipital cortex in response to faces versus visual noise stimuli in adults and 4-month-old infants. A second study (Blasi et al., 2007), which repeated this paradigm, but used a multiple channel array (rather than a dual channel array), allowed a wider cortical region to be investigated. This study replicated the results of the adult study, finding significant increases in oxyHb concentration to the face stimuli and not to visual noise. A later study by Otsuka et al. (2007) found a brain signature for the face inversion effect, i.e., the fact that while upright faces are special, preferred stimuli, processed by a dedicated brain area from a very young age, inverted faces are not privileged. The authors compared activation in the left and right temporal areas while 5–8-month-old infants were watching upright faces, inverted faces, or objects. They found increased activation in the right temporal areas in response to upright, but not to inverted faces. The same group of authors (Nakato et al., 2009) also explored the neural mechanisms underlying the development of view-invariance in face perception. Five-month-olds showed increased activation in the previously identified right temporal areas for front views of faces only, not for profile views, but by 8 months, increased activation emerges for the latter type of stimuli as well, indicating the emergence of view-invariance between these two ages.

Investigating the perception of dynamic visual social cues, Lloyd-Fox et al. (2009) found that 5-month-olds showed increased cortical activation in response to dynamic social stimuli (such as 'Peek-a-boo') in the bilateral superior

temporal and inferior frontal regions of the cortex. Importantly, this response was not found to dynamic nonsocial stimuli (such as mechanical toys), suggesting that the response was to the social component and not simply to movement. A study (Grossmann et al., 2008) investigating eye gaze perception with an accompanying smile cue in 4-month-olds found larger activation in infants' prefrontal cortex and right superior temporal region in response to a gaze shift toward, but not away from, the infant.

NIRS in Clinical Populations

NIRS is also increasingly used in clinical pediatric practice, providing a useful tool for the functional assessment of infants with different medical conditions or at risk for developmental disorders. An early example of such work comes from Chen et al.'s (2002) study, who compared healthy newborns' brain response to auditory stimulation with that of newborns suffering from hypoxic-ischemic encephalopathy. Normal newborns showed increased oxyHb and totalHb concentrations in the frontal areas during stimulation, whereas about two-thirds of the newborns with hypoxia/ischemia exhibited a decrease in oxyHb and totalHb concentrations, the extent of which negatively correlated with the severity of the pathology. More recently, NIRS has also been used to assess brain functions in preterm newborns. Mahmoudzadeh et al. (2013) have found that when exposed to speech, the brain of preterm infants born at 28–32 weeks gestational age already showed important similarities with the functional organization of the adult brain. This is all the more remarkable, as at this early stage of preterm development, not all cortical layers of the brain are yet formed. NIRS has also proven useful in assessing developmental disorders such as autism in young infants. As autism spectrum disorder has a strong genetic component, the younger siblings of children with autism are themselves at risk, but a stable diagnosis is not possible before 3 years of age. A recent NIRS study, however, has found that already during the first 6 months of life, infants at risk for autism show a reduced neural response to visual and auditory social stimuli as compared to typical controls (Lloyd-Fox et al., 2013), suggesting it might be possible in the future to identify brain signatures for certain developmental disorders even before overt behavioral markers become observable.

Technical and Methodological Considerations

As noted before, NIRS is an increasingly popular, but still relatively recent imaging technique. As a consequence, some technical, technological, and methodological issues are not fully standardized, and different practices coexist. Below, some of these points are discussed, with the aim of providing a useful starting point when planning or interpreting NIRS studies.

Probe Design and Source–Detector Separation

Unlike fMRI and PET, but like EEG, NIRS systems are equipped with a headgear that can be flexibly positioned on the participant's head, and in many systems, the headgear itself is also configurable, e.g., in terms of the number of optodes used or the relative distance between light sources and detectors. This flexibility allows experimenters to adapt to a large number of different research questions and populations. At the same time, finding the optimal design requires careful consideration of a number of factors.

One such factor is the distance between sources and detectors. In NIRS systems, a measurement point or channel is not an actual physical entity. Rather, it is the coupling of a light source and a detector, which sample from the biological tissues under the space between the two optodes. The source–detector separation is an important parameter of NIRS systems, as it determines the depth of penetration, as well as the spatial resolution. Larger separations sample from areas deeper in the cortex (the arc of the banana-shape sampled from is larger, see **Figure 2**), providing more information about neural activity. However, because the trajectory of the light is longer and it goes through more nonbrain tissues, the signal-to-noise ratio typically decreases with increasing separation. Further, larger separations decrease the spatial resolution of the measurements and allow for fewer channels to be placed on the participant's head (which is particularly relevant for young infants, whose head size is relatively small). The optimal source–detector separation is also influenced by technical parameters, such as the type of optical fibers used as well as by anatomical factors, such as the brain's distance from the scalp in a given head/brain region, the thickness of the skin and skull. Given all these factors, the optimal trade-off between penetration depth and spatial resolution in most systems is a source–detector separation between 2 and 5 cm. Most studies have used the same separation between all source–detector pairs, but a few studies have used several different separations (Blasi et al., 2007; Cristia et al., 2014; Gervain et al., 2011).

The regions of interest, where brain activity is expected to be obtained, is another relevant point to consider. Unlike EEG, NIRS is not subject to the inverse problem. In other words, the localization of brain activity as measured by NIRS is precise, even though spatial resolution is not particularly high (a few centimeters) and is mostly limited to the surface of the cortex. Existing fMRI studies can often be informative when determining the regions of interest to be targeted.

One final issue regarding the choice of headgear is the actual physical design. In young infants, who cannot follow instructions and cannot be expected to stay motionless for extended periods, the headgear plays a quintessential role in optimizing data quality. It needs to be easy to apply, light and comfortable to wear and, most importantly, it needs to ensure a stable attachment to the head such that the infant's movements should not displace the headgear with respect to the head. Additionally, it should allow the researcher to move hair out of the way, when necessary. Many NIRS systems use stretchy caps or soft, elastic straps to meet these criteria, but designing a good-fitting headgear remains one of the major challenges when working with young infants.

What Measure: OxyHb, DeoxyHb, and TotalHb

NIRS has the advantage of measuring both oxyHb and deoxyHb. The sum of the two, totalHb, is also often reported,

as it corresponds to the physiological measure of regional cerebral blood volume. Given the typical healthy adult HRF, oxyHb and deoxyHb should show concomitant changes in the opposite direction, reflecting the neurovascular coupling characteristic of brain activity.

However, in developmental studies, it is often the case that only one hemoglobin species is reported or shows significant changes as a function of stimulation. Some technical parameters, such as the material of the optical fibers used or the exact light wavelengths, certainly influence an NIRS system's differential sensitivity to the two hemoglobin species. Indeed, many infant studies, conducted with the currently available NIRS technologies, often report significant changes in the concentration of oxyHb only or report changes in the non-canonical direction for one or both hemoglobin species, i.e., an increase in deoxyHb or a decrease in oxyHb. Since the exact shape of the hemodynamic response in infants is not fully known, and to allow for better comparisons across studies, the best practice is to report both oxyHb and deoxyHb results, even when not significant.

A better understanding of the neurovascular coupling and the physiological basis of the hemodynamic response, as well as comparisons with the BOLD response measured by fMRI, originating from the paramagnetic properties of deoxyHb, will be necessary to fully clarify this question.

Brain Development

When working with infants and children, the populations most often tested with NIRS, brain development needs to be taken into account. Factors such as head size, synaptogenesis, and synaptic pruning as well as the myelination of the brain change over time. Vascularization and the relative position of the brain compared to the skull also change. As all these factors influence which brain structures are visible to NIRS at a given point in development, it is indispensable for researchers to formulate anatomically informed hypotheses. Further, comparisons across different age-groups require caution.

These problems are often exacerbated by the fact that NIRS probes are placed on the head using surface landmarks only, such as the ears or the vertex, since structural images of the participants' brains are typically not available.

One possible solution comes from the growing body of available MRI images of developmental brains. Several large-scale studies have recently been conducted resulting in searchable archives of MRI brain images from birth to adulthood (see list of relevant Web sites below). These allow developmental change along a relevant dimension to be measured and quantified, and can also serve as a basis for the assessment of individual variability within a developmental age. (Although, of course, the assessment of the actual variability among the participants of any specific study can only be precisely measured if brain scans are available for those individuals.)

Experimental Designs

The HRF is a slow, metabolic correlate of brain activity, peaking several seconds after the actual neural response. It is often assumed that in developmental populations, the response is even slower than in adults. This needs to be taken into account when designing NIRS experiments. The most straightforward and so far most popular way to circumvent this problem has been the use of block designs, also common in fMRI experiments. In a block design, a short stimulus is repeated or a longer stimulus is presented for a duration of up to several, often 10–20 s. To obtain a good signal-to-noise ratio, each experimental condition is presented in several blocks. Blocks are typically separated by extended baseline periods, e.g., 10–20 s, to allow the hemodynamic response to return to baseline. This design provides a robust signal, due to the additivity of the responses to the individual stimuli within a block, and it does not require mathematically sophisticated analyses. However, its temporal resolution is obviously quite low, and habituation/suppression or other task-related effects might occur due to the frequent repetition of the same or similar stimuli.

As NIRS machines typically have a relatively high sampling frequency (often around 10 Hz or above), it is also possible to conduct studies with event-related designs, allowing a much better temporal resolution than block designs. In event-related designs, single stimulus tokens are presented at random intervals. The interstimulus interval needs to be sufficient long to allow the response to return to baseline, but a few seconds are often enough, as the single-stimulus response itself is weaker and shorter than the block-response. This design is able to capture single-stimulus responses and fast events. At the same time, the obtained signal is relatively weak, which makes this paradigm somewhat less suited for infants, whose data are often noisy and heavily artifacted. In addition, data analysis with this design often requires complex mathematics, with assumptions about the shape of the hemodynamic response, which is not fully known in infants. Consequently, few developmental NIRS studies have used event-related designs.

Data Analysis and Interpretation

The analysis of imaging data typically consists of two steps: preprocessing, seeking to improve data quality, and statistical analysis, testing the experimental hypothesis. NIRS is no exception.

Although NIRS is more resistant to movement artifacts than other imaging methods, preprocessing is still an important step, especially when working with infants and children. NIRS data are most often band-pass filtered. The low-pass filter serves to exclude high-frequency physiological signals (around 1 Hz and above), such as the heartbeat, as well as high-frequency noise due to hair, etc., whereas the high-pass filter blocks slow baseline drifts and low-frequency oscillations due to systemic changes not related to experimental stimulation, such as changes in blood pressure, etc. Manual or automatic artifact rejection can also be necessary. It is facilitated by the fact that the hemodynamic response is relatively slow, while movement and other related artifacts are typically abrupt changes. Baseline correction is also a common procedure, whereby the mean signal intensity of a baseline period, recorded at the beginning of the entire time series or at the beginning of each individual trial, is subtracted from the data. This is another way to correct for overall shifts in systemic parameters influencing blood flow.

Two main methods have been used for the statistical analysis of experimental hypotheses. Mostly inspired by EEG, one option is time series averaging, whereby the responses to individual blocks/trials within one condition are averaged, and then the averages between conditions are statistically compared using t-tests or analyses of variance. Parameters most often compared are the amplitude, latency, or spatial localization of the responses, either to baseline or between the conditions. The other possibility, borrowed from fMRI research, is the use of a general linear model or statistical parametric mapping type of approach. The assumption behind these methods is that the NIRS data (Y) can be mathematically described as the sum of the weighed (β) contribution of different predictors (X) plus an error term (noise, ε), i.e., $Y = \beta_1 X + \beta_2 X + \ldots + \beta_n X + \varepsilon$. Predictors can be chosen as appropriate for the study design and research question, but minimally, they include a predictor for the response obtained as a boxcar function (sequences of 'on' and 'off' periods) representing the stimulation convolved with the HRF. One problem here is that the infant HRF is not well known, so studies often use the adult HRF instead, a less than ideal solution.

More recently, other statistical methods such as cross-correlations or phase information have also been used, e.g., to assess functional connectivity.

The canonical HRF, with an initial dip and a subsequent increase in oxyHb and a decrease in deoxyHb, results from a local and transient surplus of oxygen delivered by the vascular system to active brain tissue. For the correct interpretation of any NIRS data, it is thus important to realize that NIRS measures a metabolic and hence indirect correlate of neural activity. In other words, it is the metabolic 'cost' or effort that is assessed, not an actual neural response. Different NIRS signals, i.e., different metabolic costs, often imply different underlying neural mechanisms. The contrary is not necessarily true: similar HRFs do not necessarily indicate that the underlying neural mechanisms are the same. Relatedly, an activation seen in NIRS, i.e., an increase in oxyHb and a concomitant deoxyHb, might correspond to an excitatory or an inhibitory connection/relation at the neural level, as both excitation and inhibition have metabolic 'costs.' Thus, an activation pattern seen in NIRS cannot discriminate between the two.

A 'deactivation' pattern, i.e., a decrease in oxyHb and an increase in deoxyHb, is also sometimes observed. A noncanonical HRF should always be treated with caution. After all, a decrease in the oxygen supply of the brain is associated with traumatic events such as ischemia or a stroke. Therefore, we can only talk about genuine deactivation if there is reason to believe that a greater activation level was achieved prior to the measurement period, with respect to which a lower activation level, i.e., a relative decrease in oxyHb, is then observed during measurement. One situation in which this can occur is when the baseline period contains control stimuli that evoke a stronger response in a given brain region than the actual experimental stimuli. It is thus quite important to carefully consider the nature of the baseline or control stimuli. Other forms of noncanonical responses, e.g., when both hemoglobin species change in the same direction, are even harder to interpret from a physiological perspective, although such responses are sometimes observed in developmental populations. More research will be needed to fully understand nonadult HRF patterns.

Future Directions

NIRS is a relatively new, but increasingly popular brain imaging technique, which has found widespread use in developmental neuroscience. One important challenge that the NIRS community now faces is the need for more standardized procedures and methods, allowing for easier comparisons across different studies and populations. Coregistration with other imaging and experimental methods also offers an important new avenue for research. The investigation of the physiological basis of the NIRS signal, in particular in infants and children, will also be necessary for a better understanding of the complex relationships between neural, vascular, psychological, and behavioral processes.

See also: Brain Development: Ontogenetic Neurobiology of; Cerebral Cortex; Cognitive Development in Infancy, Neural Mechanisms of; Language Acquisition; Language Processing, Functional Magnetic Resonance Imaging of; Neural Basis of Language Development.

Bibliography

Aslin, R.N., Mehler, J., 2005. Near-infrared spectroscopy for functional studies of brain activity in human infants: promise, prospects, and challenges. Journal of Biomedical Optics 10, 11009.

Baird, A.A., Kagan, J., Gaudette, T., Walz, K.A., Hershlag, N., Boas, D.A., 2002. Frontal lobe activation during object permanence: data from near-infrared spectroscopy. Neuroimage 16, 1120–1125.

Bartocci, M., Winberg, J., Ruggiero, C., Bergqvist, L.L., Serra, G., Lagercrantz, H., 2000. Activation of olfactory cortex in newborn infants after odor stimulation: a functional near-infrared spectroscopy study. Pediatric Research 48, 18–23.

Blasi, A., Fox, S., Everdell, N., Volein, A., Tucker, L., Csibra, G., Gibson, A.P., Hebden, J.C., Johnson, M.H., Elwell, C.E., 2007. Investigation of depth dependent changes in cerebral haemodynamics during face perception in infants. Physics in Medicine and Biology 52, 6849–6864.

Boden, S., Obrig, H., Kohncke, C., Benav, H., Koch, S.P., Steinbrink, J., 2007. The oxygenation response to functional stimulation: is there a physiological meaning to the lag between parameters? Neuroimage 36, 100–107.

Bortfeld, H., Fava, E., Boas, D.A., 2009. Identifying cortical lateralization of speech processing in infants using near-infrared spectroscopy. Developmental Neuropsychology 34, 52–65.

Bortfeld, H., Wruck, E., Boas, D.A., 2007. Assessing infants' cortical response to speech using near-infrared spectroscopy. Neuroimage 34, 407–415.

Chance, B., Maris, M., Sorge, J., Zhang, M.Z., 1990. A phase modulation system for dual wavelength difference spectroscopy of hemoglobin deoxygenation in tissues. Proceedings of SPIE 1204, 481–491.

Chance, B., Villringer, A., Dirnagl, V., Einhaupl, K.M., 1992. Optical imaging of brain function and metabolism. Garmisch-Partenkirchen, 21–22 October 1991. Journal of Neurology 239, 359–360.

Chen, S., Sakatani, K., Lichty, W., Ning, P., Zhao, S., Zuo, H., 2002. Auditory-evoked cerebral oxygenation changes in hypoxic-ischemic encephalopathy of newborn infants monitored by near infrared spectroscopy. Early Human Development 67, 113–121.

Cristia, A., Minagawa-Kawai, Y., Egorova, N., Gervain, J., Filippin, L., Cabrol, D., Dupoux, E., 2014. Neural correlates of infant dialect discrimination: a fNIRS study. Developmental Science 17 (4), 628–635. http://dx.doi.org/10.1111/desc.12160.

Csibra, G., Henty, J., Volein, A., Elwell, C., Tucker, L., Meek, J., Johnson, M.H., 2004. Near infrared spectroscopy reveals neural activation during face perception in infants and adults. Journal of Pediatric Neurology 2, 85–89.

Dehaene-Lambertz, G., Dehaene, S., Hertz-Pannier, L., 2002. Functional neuroimaging of speech perception in infants. Science 298, 2013–2015.

Delpy, D.T., Cope, M., van der Zee, P., Arridge, S., Wray, S., Wyatt, J., 1988. Estimation of optical pathlength through tissue from direct time of flight measurement. Physics in Medicine and Biology 33, 1433–1442.

Duncan, A., Meek, J.H., Clemence, M., Elwell, C.E., Fallon, P., Tyszczuk, L., Cope, M., Delpy, D.T., 1996. Measurement of cranial optical path length as a function of age using phase resolved near infrared spectroscopy. Pediatric Research 39, 889–894.

Ferrari, M., Mottola, L., Quaresima, V., 2004. Principles, techniques, and limitations of near infrared spectroscopy. Canadian Journal of Applied Physiology Revue Canadienne de Physiologie Appliquée 29, 463–487.

Gervain, J., Berent, I., Werker, J.F., 2012. Binding at birth: the newborn brain detects identity relations and sequential position in speech. Journal of Cognitive Neuroscience, 1–11.

Gervain, J., Macagno, F., Cogoi, S., Pena, M., Mehler, J., 2008. The neonate brain detects speech structure. Proceedings of the National Academy of Sciences of the United States of America 105, 14222–14227.

Gervain, J., Mehler, J., Werker, J.F., Nelson, C.A., Csibra, G., Lloyd-Fox, S., Shukla, M., Aslin, R.N., 2011. Near-infrared spectroscopy: a report from the McDonnell infant methodology consortium. Developmental Cognitive Neuroscience 1, 22–46.

Gratton, G., Fabiani, M., 1998. Dynamic brain imaging: event-related optical signal (EROS) measures of the time course and localization of cognitive-related activity. Psychonomic Bulletin & Review 5, 535–563.

Gratton, G., Fabiani, M., 2003. The event-related optical signal (EROS) in visual cortex: replicability, consistency, localization, and resolution. Psychophysiology 40, 561–571.

Grossmann, T., Johnson, M.H., Lloyd-Fox, S., Blasi, A., Deligianni, F., Elwell, C., Csibra, G., 2008. Early cortical specialization for face-to-face communication in human infants. Proceedings of the Royal Society B Biological Sciences 275, 2803–2811.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nature Reviews Neuroscience 8, 393–402.

Hiraoka, M., Firbank, M., Essenpreis, M., Cope, M., Arridge, S.R., van der Zee, P., Delpy, D.T., 1993. A Monte Carlo investigation of optical pathlength in inhomogeneous tissue and its application to near-infrared spectroscopy. Physics in Medicine and Biology 38, 1859–1876.

Homae, F., Watanabe, H., Nakano, T., Asakawa, K., Taga, G., 2006. The right hemisphere of sleeping infant perceives sentential prosody. Neuroscience Research 54, 276–280.

Homae, F., Watanabe, H., Nakano, T., Taga, G., 2007. Prosodic processing in the developing brain. Neuroscience Research 59, 29–39.

Homae, F., Watanabe, H., Nakano, T., Taga, G., 2011. Frontiers: large-scale brain networks underlying language acquisition in early infancy. Frontiers in Psychology 2.

Homae, F., Watanabe, H., Otobe, T., Nakano, T., Go, T., Konishi, Y., Taga, G., 2010. Development of global cortical networks in early infancy. Journal of Neuroscience 30, 4877–4882.

Jobsis, F.F., 1977. Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. Science 198, 1264–1267.

Kimura, D., 1967. Functional asymmetry of the brain in dichotic listening. Cortex 3, 163–178.

Lloyd-Fox, S., Blasi, A., Elwell, C.E., February 2010. Illuminating the developing brain: the past, present and future of functional near infrared spectroscopy. Neuroscience & Biobehavioral Reviews 34 (3), 269–284.

Lloyd-Fox, S., Blasi, A., Elwell, C.E., Charman, T., Murphy, D., Johnson, M.H., 2013. Reduced neural sensitivity to social stimuli in infants at risk for autism. Proceedings of the Royal Society B Biological Sciences 280.

Lloyd-Fox, S., Blasi, A., Volein, A., Everdell, N., Elwell, C.E., Johnson, M.H., 2009. Social perception in infancy: a near infrared spectroscopy study. Child Development 80, 986–999.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., Oeltermann, A., 2001. Neurophysiological investigation of the basis of the fMRI signal. Nature 412, 150–157.

Magistretti, P.J., 2006. Neuron–glia metabolic coupling and plasticity. Journal of Experimental Biology 209, 2304–2311.

Mahmoudzadeh, M., Dehaene-Lambertz, G., Fournier, M., Kongolo, G., Goudjil, S., Dubois, J., Grebe, R., Wallois, F., 2013. Syllabic discrimination in premature human infants prior to complete formation of cortical layers. Proceedings of the National Academy of Sciences of the United States of America 110, 4846–4851.

May, L., Byers-Heinlein, K., Gervain, J., Werker, J.F., 2011. Language and the newborn brain: does prenatal language experience shape the neonate neural response to speech? Frontiers in Psychology 2, 222.

Meek, J., 2002. Basic principles of optical imaging and application to the study of infant development. Developmental Science 5, 371–380.

Minagawa-Kawai, Y., Cristià, A., Dupoux, E., 2011a. Cerebral lateralization and early speech acquisition: a developmental scenario. Developmental Cognitive Neuroscience 1 (3), 217–232.

Minagawa-Kawai, Y., Cristià, A., Vendelin, I., Cabrol, D., Dupoux, E., 2011b. Assessing signal-driven mechanisms in neonates: brain responses to temporally and spectrally different sounds. Frontiers in Psychology 2, 135.

Minagawa-Kawai, Y., Mori, K., Hebden, J.C., Dupoux, E., 2008. Optical imaging of infants' neurocognitive development: recent advances and perspectives. Developmental Neurobiology 68, 712–728.

Minagawa-Kawai, Y., Mori, K., Naoi, N., Kojima, S., 2007. Neural attunement processes in infants during the acquisition of a language-specific phonemic contrast. Journal of Neuroscience 27, 315–321.

Nakato, E., Otsuka, Y., Kanazawa, S., Yamaguchi, M.K., Watanabe, S., Kakigi, R., 2009. When do infants differentiate profile face from frontal face? A near-infrared spectroscopic study. Human Brain Mapping 30, 462–472.

Obrig, H., Wenzel, R., Kohl, M., Horst, S., Wobst, P., Steinbrink, J., Thomas, F., Villringer, A., 2000. Near-infrared spectroscopy: does it function in functional activation studies of the adult brain? International Journal of Psychophysiology. Official Journal of the International Organization of Psychophysiology 35, 125–142.

Otsuka, Y., Nakato, E., Kanazawa, S., Yamaguchi, M.K., Watanabe, S., Kakigi, R., 2007. Neural activation to upright and inverted faces in infants measured by near infrared spectroscopy. Neuroimage 34, 399–406.

Peña, M., Maki, A., Kovacic, D., Dehaene-Lambertz, G., Koizumi, H., Bouquet, F., Mehler, J., 2003. Sounds and silence: an optical topography study of language recognition at birth. Proceedings of the National Academy of Sciences of the United States of America 100, 11702–11705.

Poeppel, D., 2003. The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. Speech Communication 41, 245–255.

Poeppel, D., Idsardi, W.J., Van Wassenhove, V., 2008. Speech perception at the interface of neurobiology and linguistics. Philosophical Transactions of the Royal Society B: Biological Sciences 363, 1071–1086.

Saito, Y., Aoyama, S., Kondo, T., Fukumoto, R., Konishi, N., Nakamura, K., Kobayashi, M., Toshima, T., 2007a. Frontal cerebral blood flow change associated with infant-directed speech. Archives of Disease in Childhood Fetal Neonatal Edition 92, F113–F116.

Saito, Y., Kondo, T., Aoyama, S., Fukumoto, R., Konishi, N., Nakamura, K., Kobayashi, M., Toshima, T., 2007b. The function of the frontal lobe in neonates for response to a prosodic voice. Early Human Development 83, 225–230.

Sakatani, K., Chen, S., Lichty, W., Zuo, H., Wang, Y.P., 1999. Cerebral blood oxygenation changes induced by auditory stimulation in newborn infants measured by near infrared spectroscopy. Early Human Development 55, 229–236.

Sato, H., Hirabayashi, Y., Tsubokura, H., Kanai, M., Ashida, T., Konishi, I., Uchida-Ota, M., Konishi, Y., Maki, A., 2012. Cerebral hemodynamics in newborn infants exposed to speech sounds: a whole-head optical topography study. Human Brain Mapping 33, 2092–2103.

Sato, Y., Sogabe, Y., Mazuka, R., November 2010. Development of hemispheric specialization for lexical pitch-accent in japanese infants. Journal of Cognitive Neuroscience 22 (11), 2501–2513.

Steinbrink, J., Kempf, F.C., Villringer, A., Obrig, H., 2005. The fast optical signal–robust or elusive when non-invasively measured in the human adult? Neuroimage 26, 996–1008.

Taga, G., Asakawa, K., 2007. Selectivity and localization of cortical response to auditory and visual stimulation in awake infants aged 2 to 4 months. Neuroimage 36, 1246–1252.

Taga, G., Asakawa, K., Maki, A., Konishi, Y., Koizumi, H., 2003. Brain imaging in awake infants by near-infrared optical topography. Proceedings of the National Academy of Sciences of the United States of America 100, 10722–10727.

Telkemeyer, S., Rossi, S., Koch, S.P., Nierhaus, T., Steinbrink, J., Poeppel, D., Obrig, H., Wartenburger, I., 2009. Sensitivity of newborn auditory cortex to the temporal structure of sounds. Journal of Neuroscience 29, 14726–14733.

Telkemeyer, S., Rossi, S., Nierhaus, T., Steinbrink, J., Obrig, H., Wartenburger, I., 2011. Acoustic processing of temporally modulated sounds in infants: evidence from a combined near-infrared spectroscopy and EEG study. Frontiers in Psychology 1, 62.

Tse, C., Lee, C., Sullivan, J., Garnsey, S.M., Dell, G.S., Fabiani, M., Gratton, G., 2007. Imaging cortical dynamics of language processing with the event-related optical signal. Proceedings of the National Academy of Sciences of the United States of America 104, 17157–17162.

Villringer, A., Chance, B., 1997. Non-invasive optical spectroscopy and imaging of human brain function. Trends in Neurosciences 20, 435–442.

Wartenburger, I., Steinbrink, J., Telkemeyer, S., Friedrich, M., Friederici, A.D., Obrig, H., 2007. The processing of prosody: evidence of interhemispheric specialization at the age of four. Neuroimage 34, 416–425.

Wilcox, T., Bortfeld, H., Woods, R., Wruck, E., Armstrong, J., Boas, D., 2009. Hemodynamic changes in the infant cortex during the processing of featural and spatiotemporal information. Neuropsychologia 47, 657–662.

Wilcox, T., Bortfeld, H., Woods, R., Wruck, E., Boas, D.A., 2008. Hemodynamic response to featural changes in the occipital and inferior temporal cortex in infants: a preliminary methodological exploration. Developmental Science 11, 361–370.

Wolf, M., Ferrari, M., Quaresima, V., 2007. Progress of near-infrared spectroscopy and topography for brain and muscle clinical applications. Journal of Biomedical Optics 12, 062104.

Wyatt, J.S., Cope, M., Delpy, D.T., van der Zee, P., Arridge, S., Edwards, A.D., Reynolds, E.O., 1990. Measurement of optical path length for cerebral near-infrared spectroscopy in newborn infants. Developmental Neuroscience 12, 140–144.

Zatorre, R.J., Evans, A.C., Meyer, E., Gjedde, A., 1992. Lateralization of phonetic and pitch discrimination in speech processing. Science 256, 846–849.

Zatorre, R.J., Gandour, J.T., 2008. Neural specializations for speech and pitch: moving beyond the dichotomies. Philosophical Transactions of the Royal Society B: Biological Sciences 363, 1087–1104.

Relevant Websites

http://pediatricmri.nih.gov/nihpd/info/index.html – Archive of Developmental MRI Images.

http://fnirs.org/ – The fNIRS Community Web Site.

Needham, Joseph (1900–95)

Hans U Vogel, Universität Tübingen, Tübingen, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 10458–10462, © 2001, Elsevier Ltd.

Abstract

Joseph Needham (1900–95) was not only an excellent biochemist, but also an eminent historian. He became famous for his impressive achievement embodied in the massive series *Science and Civilisation in China*, published in many volumes since 1954. Although the exploration of the history of traditional Chinese science, technology, and medicine had begun already before him, he is certainly the scholar who contributed most to the academic institutionalization of these studies, both in East and West. One of his basic convictions was that modern science and technology, as they arose in the West, owed much to ideas and inventions originating in East Asia or China. He conceived of modern science as an outcome of a universalistic process, to which various cultures had made their contributions. In his opinion, the reason that modern science did not arise in China was because it was hindered by external factors, especially the structure of Chinese feudal-bureaucratic society. Needham's approaches provide important elements for a comparative historical and differential sociology of civilizational complexes and intercivilizational encounters. He not only ascribed social values to the notion of universality of modern science, but also perceived it as a sharp tool of critique directed against Eurocentric internalism.

Joseph Needham was not only an excellent scientist, but also an eminent historian. He became famous for his impressive achievement embodied in the massive series *Science and Civilisation in China* (SCC), published in many volumes since 1954. His work on this subject has deeply influenced not only research on the history of China, but also studies in the history of science, technology, and medicine in general. Although the exploration of the history of traditional Chinese science, technology, and medicine had begun already before him, he is certainly the scholar who contributed most to the institutionalization of these studies, resulting, both in East and West, in the foundation of specific academic research disciplines and institutes. One of his basic convictions was that *modern* science and technology, as they arose in the West, owed much to ideas and inventions originating in East Asia or China. He conceived of modern science as an outcome of an universalistic process, like a sea into which the rivers of various cultures had flowed, making their contributions. In his opinion, the reason that modern science did not arise in China was because it was hindered by external factors, that is, the structure of Chinese feudal-bureaucratic society.

Life

Needham was born in London on 9 December 1900, the only son of a physician and a musically talented mother. In 1918 he went to Gonville and Gaius College in Cambridge to study biochemistry, where he became a student and collaborator of Frederick Gowland Hopkins, a subsequent Nobel Prize winner in medicine. In 1924, he married the biochemist Dorothy Moyle. In 1931 Needham published *Chemical Embryology* in three volumes, a description, definition, and historical account of a new scientific discipline in which he had increasingly specialized. This was followed, in 1942, by *Biochemistry and Morphogenesis*. Already at an early stage in his career, Needham engaged in public debates about the relationships between science, religion, philosophy, history, and politics. In the period from 1925 to 1946, he wrote or edited such works as *Science, Religion, and Reality* (1925), *The Sceptical Biologist* (1926), *The Nazi Attack on International Science* (1941), *History is on Our Side* (1946), etc., the latter two attacking the Nazi abuse of science.

In 1937 a group of young Chinese scientists came to Cambridge, among them Lu Gwei-Djen, who later became one of his collaborators and, 2 years after the death of Dorothy Moyle in 1987, his second wife. These Chinese scientists aroused in him a deep interest in and fascination for Chinese culture and language. Needham started to learn Chinese privately with the help of a Cambridge professor of Sinology. In 1942, the British government commissioned him to go to China to establish the Sino-British Science Co-operation Office in Chongqing, the city to which the Chinese government had withdrawn in the face of the Japanese invasion. He spent more than 3 years in China, where, apart from his official duties, he started to collect information, literature, and source material on the history of science, technology, and medicine in China. From 1946 to 1948 Needham was appointed director of the department for science at UNESCO in Paris. In 1952 he was severely criticized for joining an international scientific committee, which on the invitation of the governments of North Korea, the People's Republic of China, and the communist-sponsored World Peace Council was commissioned to investigate charges of US biological warfare in the Korean War. The committee concluded that the US military had experimented with biological weapons, and Needham was subsequently reproached as having been deluded by communist propaganda and his love for China. As a consequence, he was shunned by many in the academic world, and was not permitted to visit the US in the coming years.

After his return to Cambridge in 1948 Needham all but gave up his scientific research, and turned instead to his great project SCC. The first volume was published in 1954. In the course of time, more than twenty collaborators from all parts of the world joined Needham to participate in this project. In 1968 Needham founded in Cambridge the East Asian History of Science Trust to preserve his library and continue his work. By 1987, funds had been raised to build a home for his library, the

Needham Research Institute, which was finally completed in 1991. Together with his first wife he was elected Fellow of the Royal Society (science) in 1941. Thirty years later, he also became Fellow of the British Academy (humanities). In 1968 the Society for History of Science and the Society for History of Technology honored him with the George Sarton Medal and the Leonardo da Vinci Medal, respectively. He was bearer of the Order of the Brilliant Star conferred on him by the People's Republic of China. Finally, in June 1992, his achievements were given official recognition when he was made Companion of Honour by Her Majesty the Queen. He died in Cambridge on 24 March 1995, at the age of 94.

Contribution to Knowledge and Intellectual Environment

Based on his ecumenical and universalistic view of the history of science and technology, in which the Old World must be taken as a whole, one of Needham's crucial and seemingly paradoxical problems was why modern science, the mathematization of hypotheses about Nature, with all its implications for advanced technology, made its meteoric rise only in the West at the time of Galileo. Of equal importance for him was the question of why it was that between the second century BC and the sixteenth century AD East Asian culture was much more efficient than the European West in applying human knowledge of Nature to useful purposes. How could it have been that the weakness of China in theory and geometrical systematization did not prevent the emergence of technological discoveries and inventions often far in advance of contemporary Europe? What were the inhibiting factors in the Chinese case that prevented a rise of modern science in Asia analogous to that which took place in Europe from the sixteenth century onward? And finally, how was it that Chinese backwardness in scientific theory coexisted with the growth of an organic philosophy of Nature, closely resembling that which modern science (in Whitehead's sense) had adopted after three centuries of mechanical materialism?

In order to find answers to these questions, Needham carried out enormous comparative research, not only dedicated to internalist interpretations of science and technology, but also taking into account influences from philosophy, language, religion, and ideology as well as social, economic, and geographical conditions. It was likewise important for him to fix dates of discoveries and inventions, aiming to 'titrate' the great civilizations one against another and to find out and give credit where credit was due. In many cases where Needham and his collaborators were able to ascertain an earlier date of a discovery or invention in China than in the West, he concluded that this must have been a case of direct borrowing, indirect transmission, or diffusion stimulus, even if no other concrete evidence was available. Basically, he shifted the onus of proof onto those people who doubted that such a transfer had taken place. Generally speaking, however, Needham was never thinking in simplistic terms, as he did, for instance, admit the possibility of independent discoveries and inventions. From very early on he concluded that until the Scientific Revolution in the late Renaissance the West was not only deeply influenced in the field of technology by Chinese or East Asian technological inventions and scientific discoveries, but also deeply affected by them in its social structure. For instance, the adoption in the West of the Chinese invention of gunpowder and firearms dealt a fatal blow to castles and hence to Western military aristocratic feudalism.

Needham drew a clear distinction between modern science on the one hand and ancient and medieval sciences on the other. Modern science, as it developed only in Western Europe at the time of Galileo, is based on five fundamental characteristics, namely: the application of mathematical hypotheses to Nature, the full understanding and use of the experimental method, the distinction between primary and secondary qualities, the geometrization of space, and the acceptance of a mechanical model of reality. Primitive and medieval sciences distinguish themselves clearly from the modern type, as, for instance, their intrinsic and essential vagueness always made them incapable of proof or disproof. As far as numerical figures entered into them, these numbers were manipulated in forms of numerology or number-mysticism constructed *a priori*, and not employed as the stuff of quantitative measurements compared *a posteriori*.

During his years of study and research in Cambridge, Needham's worldview was characterized by four conceptual clusters that substantially influenced his work for SCC:

1. Needham had a strong antipathy against the mechanical concepts upheld by many physicists of his time. His philosophy of science was influenced by Whitehead's organicism. He was, however, not a rigid defender of Neo-vitalism, but together with Lancelot Hogben argued for its 'mechanical' or 'materialistic' aspects. Needham considered Whitehead's theory of organicism a synthesis of mechanism and vitalism.
2. Needham espoused the idea of evolution as a concept of wide application. For him evolution was not only a phenomenon of the biological order, but could also be observed in social organization. Social organizations necessarily followed a linear course of evolution, for which the Marxist stage theory of social development was an indispensable explanative tool. Social development cannot be explained by biological theory, but needs its own unique approach. Needham's concept of evolutionary development of social organizations did, however, not deal with a socio-Darwinist 'survival of the fittest,' but aimed at overcoming the excesses of capitalism. 'Collectivism,' as it would manifest itself in a worldwide communist society, was seen by Needham as the next higher stage of social organization.
3. Needham sympathized with Marxist dialectic materialism. Most of the researchers in Hopkins's laboratory were politically left wing. This was typical of many British biologists of the twenties and thirties, like J. B. S. Haldane, J. G. Crowther, and J. D. Bernal. During his youth Needham was impressed by the events of the Russian Revolution. Moreover, in 1925 he read Engels's *Dialektik der Natur* [Dialectics of Nature] which fascinated him. During one of his stays at the Roscoff Marine Biological Station Needham and his wife met Louis Rapkine, a biochemist born in Lithuania who grew up in Canada. Rapkine introduced Needham to the Marxist classics. Moreover, he learned much of him

about the living conditions of the lower strata of the working society for which Needham had always entertained great sympathy. Like Julian Huxley and the adherents of a 'scientific humanism,' Needham expanded his interests from biological problems into the social field. Due to his rational, antireductionist attitude toward science, he was convinced that scientists also had social and ethical responsibilities. Democratic, socialist, and populist elements were the main constituents of his political stance. Needham, however, did not count himself an orthodox Marxist because he was not dogmatic with regard to the question of which aspects of human action constituted foundation or superstructure. He belonged to a generation of the 1920s who adhered to Marxism because of ideology and romantic ideals, in contrast to the generation of the 1930s which was characterized by a more realistic attitude toward politics and tactics. In his later years, Needham never gave up his hopes for the realization of an 'ideal socialism.'

4. Needham stressed synthesis over analysis. It was his basic conviction that dialectic materialism was the synthesis between mechanical materialism and idealism. Religion was for him something that need not be got rid of in favor of materialism, but which should be absorbed by it. On the one hand, he stated in 1936 that communism possessed a higher sense of the numinous than any of the traditional religions, and that, on a higher stage, a synthesis has to be carried out between religion and Marxism. Communism could provide the moral theology of our time. On the other hand, Christianity was necessary for him to help redeem the ruthlessness of science. In an autobiographical account published in 1973 under the pseudonym 'Holorenshaw,' Needham informed the public about three antitheses which had determined his life. The first is the antithesis between deep religiousness and science, the second between deep religiousness and his sympathy for socialism, especially in its Marxist form, and the third referring to the differences between Eastern and Western cultures. He declared it to be his duty to transform these antitheses into syntheses. This also shows that Needham was not an adherent of scientism. He was of the opinion that different molds of human experience (e.g., religion, art, history, philosophy, and science) coexist and all have claims for validity, and that science could not do without ethics.

During their tracing of the history of discoveries and inventions Needham and his collaborators unearthed an impressive array of early Chinese achievements in technology and science. One example is the mastery of cast iron in China, a millennium and a half before it was known in the West. This Chinese expertise which started in the fourth century BC and which required substantial empirical knowledge of the properties of iron was based on a highly developed bronze casting technology. Cast iron was not only mass-produced for weapons, agricultural tools, and cooking utensils, but also put to such innovative uses as chain suspension bridges (first century AD) and deep-drilling for brine (eleventh century AD) and, later, natural gas. The double-acting piston bellows producing a constant flow of air both during the pushing and pulling of the piston was probably a crucial device enabling the Chinese to achieve superiority in metallurgy for so many hundreds of years. In metallurgical processing also the crank and the piston rod were developed. In agriculture, row cultivation and intensive hoeing were adopted from about the sixth century BC. By the third century BC frame-plows were equipped with cast-iron plowshares and moldboards which substantially reduced friction and allowed the cultivation of heavy and waterlogged soils. These plows were far in advance to the ards used in the Western world.

About the same time, the equine breast-strap and collar harnesses emerged by which the efficiency of animal power was improved. Other agricultural innovations adopted centuries prior to the West included various types of rotary-fan winnowing machines for the efficient separation of grain and chaff as well as multitube seed drills especially in row cultivation. Both these technical devices were known since the second century AD. All these inventions spurred agricultural production and provided the economic basis for social differentiation and state building.

Among the nautical innovations of the Chinese, apart from the well-known compass, the sternpost rudder (first century AD) and watertight compartments in ships (second century AD) should be mentioned. The Chinese were also great masters in the building of irrigation works, canals, and canal pound-locks. Some of these ancient infrastructure achievements can still be admired today, like the Dujiangyan irrigation system north of Chengdu in Sichuan Province built in the third century BC. Some of these feats of engineering, however, fell into disuse and were forgotten, like the modern-type canal pound-locks with two gates of the tenth century, which were later replaced by the much more ordinary flash-gates with one gate only.

In the military field, Needham noted that gunpowder had arisen from the investigations of Taoist alchemists in about AD 850. About two centuries later it was used for fireworks, bombs, grenades, land mines, and sea mines. In the eleventh and twelfth centuries gunpowder served the production of rockets, and only about a century later it provided the propelling force in guns, cannons, and mortars. In China, however, gunpowder and firearms were not further developed to occupy a position as dominant as that which they did later in European warfare.

Woodblock printing on paper and silk arose in China around the seventh century AD. This allowed the mass production of all kinds of scriptures and texts. However, only few experiments were made with movable-type printing. Paper, another revolutionary product, was invented in China already in about the first century BC. It spread from China to Korea and Japan in the seventh century and to Central Asia, the Arabs and Europe one century later. Another famous Chinese invention that in the eighteenth century was imitated in the West was porcelain.

The decimal system was already in use in archaic China. It also seems that the Chinese were the first to use a blank space for zero and to recognize and use negative numbers as well as decimal fractions. They were also in the possession of a refined value for *pi* (third century AD) and were the first to note the hexagonal structure of snowflakes (second century BC), to mention only the most important inventions and discoveries. Many of the Chinese findings described by Needham and his collaborators in SCC have stood the test of time and thus have to be considered historical facts.

Assessment and Impact

Needham's work in the framework of SCC was frequently the object of criticism that can be summarized as follows: because Needham took modern science in its final result as a guideline, there is a tendency in his work to highlight those Chinese developments and events best suited to be interpreted as antecedents to modern science. Those currents, however, which did not show universalistic characteristics were neglected by him or even considered obstructive to the development of modern science, without his having undertaken a serious attempt at interpreting them as culturally specific and integral systems of thought and action. It has been pointed out that in his earlier works his linguistic capabilities were not of the highest standard, sometimes reading modern scientific ideas and virtues into traditional Chinese texts. Some reproached him for not having differentiated clearly enough between science and technology, because he sometimes regarded technology as nothing more than applied science. Clear conceptual differentiation was also lacking with regard to such concepts as the Scientific Revolution and the Industrial Revolution, as well as invention and innovation.

Others criticized the encyclopedic structure of SCC. The first volumes, at least, were said to be largely characterized by a timeless, even static, sense of history, to provide only a blurred picture of the world of thought and social interaction in the fields of science and technology, and to be inadequate in pointing out temporal and spatial differences. Discussions arose also with regard to individual findings of Needham. This in particular concerned his history of traditional Chinese astronomical clockworks and their linkwork escapement mechanism that he interpreted as the missing link in the development from clepsydras to mechanical clocks, the production of crystalline synthetic sexual hormones from the urine of pubescent boys by Chinese alchemists, or his identification of traditional Chinese concepts of Nature as precursors of Whitehead's organicist view. Moreover, it was shown that cross- and intracultural transmissions of ideas and devices may have been more complex processes than imagined by Needham.

More fundamental criticism addressed Needham's explanations of the socioeconomic and political reasons why the Scientific Revolution did not take place in China. Nathan Sivin repudiated this as an inadmissible contrafactual question that would diminish in interest as soon as it became clear in which social, economic, political, and cultural environments, and for which purposes traditional Chinese scientists carried out their activities. Benjamin Nelson stated that Needham had underestimated the complexity of social, political, cultural, and intellectual processes leading to the Scientific Revolution in the West. Needham's juxtaposition of Europe's mercantile bourgeoisie and commercial capitalism as conducive to the rise of modern science and China's Confucian 'feudal bureaucracy' as obstructive was considered by many critics as simplistic – as was his characterization of the Taoists as promoters of science and the Confucians as exerting a rather negative influence on it. In conclusion, Needham's work is often considered to be a mixture consisting to a large part of solid findings, plausible, but not always incontestable interpretations, personal speculations, and, sometimes, errors.

In spite of these criticisms, Needham's work undoubtedly embodies a number of tremendous achievements. First of all, due to his work not only is much more now known about the history of science, technology, and medicine in China, but it also triggered a dramatic increase in the research activities and institutionalization of the academic disciplines concerned. At the same time, a specialization took place, as it is hardly possible anymore for an individual researcher to deal with all the various disciplines to the same depth. Furthermore, as Nelson pointed out, Needham was much more interested in historical processes than Max Weber, one of the great scholars who had attempted to explain the uniqueness of the Western rise of modern science. Needham's work also underlines the importance of intercivilizational encounters and points to the inevitable drive toward universality and thus the overcoming of all local particularities. This is another difference to Weber who was much more interested in a differential typology and the establishment of ideal types, and thus more or less treated each civilization as a discrete unit. These and other insights of Needham therefore provide important elements for a comparative historical and differential sociology of civilizational complexes and intercivilizational encounters (Nelson, 1986). And although much criticism was voiced against his notion of the universality of modern science because potentially this might be invoked to justify a blanket modernization, one should not forget that Needham not only ascribed social values to this notion, but also perceived it as a sharp tool of critique directed against Eurocentric internalism.

See also: Asia, Sociocultural Aspects: China; Biomedical Sciences, History and Sociology of; Chemical Science: History and Sociology; East Asian Studies: Culture; East Asian Studies: Economics; East Asian Studies: Gender; Historical Thought and Historiography: Christian Tradition; Historical Thought and Historiography: East Asia; Human Sciences, History of; Nationalism, Historical Aspects of: The West; Physical Sciences: History and Sociology; Science, History of; Science: Constructivist Perspectives, History of; Urban Planning in China.

Bibliography

Blue, G., 1997. Joseph Needham: A publication history. Chinese Science 14, 90–132.

Brook, T., 1996. The sinology of Joseph Needham. Modern China 22 (3), 340–348.

Elvin, M. (Ed.), 1980. Symposium: The work of Joseph Needham. Past and Present 87, 17–53.

Gazagnadou, D., 1991. Joseph Needham, un taoïste d'honneur, autobiographie. De l'embryologie *à* la civilisation chinoi.se. Éditions du Féiin-UNESCO, Paris.

Habib, S.I., Dhruv, R. (Eds.), 1999. Situating the History of Science: Dialogues with Joseph Needham. Oxford University Press, Delhi, India.

Holorenshaw, H., [pseudonym of Needham], 1973. The making of an honorary Taoist. In: Teich, M., Young, R. (Eds.), Changing Perspectives in the History of Science: Essays in Honour of Joseph Needham. Heinemann, London.

Li, G.H., Zhang, M.W., Cao, T.Q. (Eds.), 1982. Explorations in the History of Science and Technology in China: A Special Number of the 'Collections of Essays on Chinese Literature and History,' Compiled in Honour of the Eightieth Birthday of Dr. Joseph Needham. Shanghai Classics Publishing House, Shanghai, China.

Nakayama, S., Sivin, N. (Eds.), 1973. Chinese Science: Explorations of an Ancient Tradition. MIT, Cambridge, MA.

Needham, J., 1970. Clerks and Craftsmen in China and the West: Lectures and Addresses on the History of Science and Technology. Cambridge University Press, Cambridge, UK.

Needham, J., 1979. The Grand Titration: Science and Society in East and West, Second ed. Allen & Unwin, London.

Needham, J., et al., 1954–2001. Science and Civilisation in China, 16 and more vols. Cambridge University Press, Cambridge, UK.

Nelson, B., 1986. Der Ursprung der Moderne: Vergleichende Studien zum Zivilisationsprozeß. Suhrkamp, Frankfurt, Germany.

Sivin, N., 1983. Chinesische Wissenschaft: Ein Vergleich der Ansätze von Max Weber und Joseph Needham. In: Schluchter, W. (Ed.), Max Webers Studie über Konfuzianismus und Taoismus: Interpretation und Kritik. Suhrkamp, Frankfurt, Germany.

Sivin, N., 1988. Science and medicine in imperial China—the state of the field. Journal of Asian Studies 47 (1), 41–90.

Vogel, H.-U., 1993. Naturwissenschaften und Technik im vormodernen China: Historie, Historiographie und Ideologie. Studienbrief für die FernUniversität—Gesamthochschule Hagen, Fachbereich Erziehungs-, Sozial- und Geisteswissenschaften.

Needs Assessment

James W Altschuld, The Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The goal of this article is to describe in a straightforward manner what needs are and their assessment. It begins with a brief overview of the nature of needs and the process for determining them. Then, it deals with origins of the concept, the modern history of needs assessment, definitions, some guiding models for working with needs along with a discussion of their key premises, and a look at commonly utilized methods. It also covers topics such as why needs should be determined, their importance for meaningful program development and improvement, factors affecting the implementation of assessments, and what the future might hold for this area of endeavor.

Origins

Social scientists have a sense of what needs are and what needs assessment is; yet surveys of the training of evaluators who often deal with such entities indicate that relevant coursework is just not there. The implication seems to be that just knowledge of methods is required to do one. The premise of this article is that there is a rich and established base of literature, which makes for more sophisticated practice. A brief look at the history of needs assessment is given followed by definitions of terms, structures to guide assessment, some key issues, and what might be on the horizon for the enterprise.

Where did the concept come from, where did it originate? Wedman (in press) speculated that it was in the Stone Age when our ancestors dealt with the determination of needs. There are no records to support this view but at a later time the Egyptian *Book of the Dead* (British Museum, 2011) talks about what the deceased would require in the afterlife. The proposition that needs are embedded in human existence was demonstrated in Hansen's classic 1991 study. He noted that across fields when people work on problems or issues they think of discrepancies between what is and what should be, that is they start with needs as a foundation for their work.

Recently Atschuld and Watkins (2014) developed a time line of what has taken place since the 1960s until the present as shown in **Table 1**. Several entries in the table should be highlighted. One is that writings about needs and how to identify and use them have been there for years and have accelerated since 2010 although research studies are less in evidence. There is a strong tie between evaluation and needs assessment as exemplified in the writings of Stufflebeam, Scriven, and Roth, and others, and by the national needs assessment group of the American Evaluation Association. Pointed criticisms began in the 1970s, periodically reemerged, and rose to a bit of a crescendo by the early 1990s. Toward the end of the decade and up to the present a new phase of practice has been noticeable, that of a hybrid between asset/capacity building and needs assessment with some major cases appearing.

Definition and Some Concepts

There is a basic idea that underlies all thinking about needs. It is of a discrepancy between what is (current condition) and what should be or what is desirable and that the gap is important in regard to program planning and evaluation. Yet, that simple definition can mask complexities, e.g., there are many ways to define the 'what should be' condition such as what is likely, what is feasible, what is ideal, and so forth. In the field of Human Performance Technology (more in the business and private sectors) 'what should be' might be cast as a high level of performance of an individual or group against lower levels that are observed. Yet, no matter the discipline, deficits are to be identified, examined, and if of high priority they are to be rectified.

The definition requires that current and desired states be measured. That is important for in the past-published studies labeled as needs assessments did not always contain data on both conditions. Since the early 2000s there has been movement toward more appropriate recognition of the components of need. Things subsumed under the definition are different types of needs and the virtually hidden notion of the counterfactual state.

The types of needs would include short- and long-term needs, maintenance needs of equipment and skills, severe versus slight needs, and others. Consider that power blackouts routinely occur in some cities – the value of maintenance takes on meaning. Or think of three levels of need (1 – recipients of service, 2 – providers of service, 3 – the support system that enables level 2 to assist level 1) and it is apparent that the idea of a discrepancy or gap is a rubric for a much more sublime tapestry. Add into the mix the counterfactual state that asks what would happen if nothing were done about a discrepancy. Might it resolve over time of its own accord? A sprained ankle will tend to go away perhaps after several weeks and no direct intervention is needed, although for professional athletes alacrity would be pertinent. So needs have to be prioritized to see if they warrant immediate or longer-term action and the expenditure of resources (money, organizational effort) to improve the situation. In some instances the answer might be no. With the basic premise in hand what are models for assessing need?

Some Guiding Models

Two approaches are well known. The three-phase model of assessment originally sketched out by Witkin in 1984 and in its latest incarnation by Altschuld and Kumar in 2010 is portrayed in **Table 2**.

Table 1 Overview of modern needs assessment (NA)

Time period	*Nature of event(s)*	*Some contributors*
Mid-1960s (the early years)	Elementary and Secondary Education Act (ESEA) of 1965 called for the determination of needs for programs and projects Context, Input, Product, and Process (CIPP) evaluation model in 1965	Congress of the United States Stufflebeam develops his CIPP model with the C part related to needs
Mid-1960s into the early 1980s (the formative period)	Many NAs being done especially in public schools Writings and presentations of some prominent authors Prominent criticisms appear	Witkin's nationwide survey of the practices of school districts Warheit et al.'s (1979) book on NA in mental health Kaufman presents on the topic and develops his model Scriven and Roth paper on the Big Bob idea of needs (see below) Kamis catalogs criticisms of NA
1980s through the early 1990s (coming of age but headwinds)	A heyday of watershed events, at least four major books are written and gain wide acceptance Small number of courses devoted to NA at several universities Small national NA group joins American Evaluation Association (AEA) Criticisms again toward the latter part of the period	Witkin, Kaufman, Neuber, Lauffer, McKillip books on aspects of needs Courses at Southern Illinois, Utah State, and Ohio State The models of Witkin and Kaufman take full form Kaufman publishes numerous papers
Early 1990s through the early 2000s (enhancements but challenges)	Rejoinders to criticisms Attack on NA from an asset/capacity building stance Asset/capacity building efforts take off based on starting from the positive as opposed to a needs, discrepancy-oriented position Major new books on NA are published	Witkin takes on the critics Kretzmann and McKnight (1993) harshly point to the negatives of assessment Altschuld and Witkin partner for two books, Gupta generates one oriented toward business settings Kaufman continues to publish Scriven and Roth paper reissued by the American Evaluation Association
Early 2000s through nearly the end of the decade (continued work and improvement)	Major papers about the placement of NA in organizations, Performance Pyramid, and NA and Evaluation survey Persistence of NA-Topical Interest Group at AEA as a viable entity Hybrid NA and asset/capacity building efforts appear More books about NA	See papers by Altschuld, Wedman and Watkins, Watkins and Guerra Good source on hybrid thinking for asset/capacity building and NA is the CADISPA website The five volume Needs Assessment Kit (Altschuld et al.) and book by Gupta et al.
2010 to the present (toward maturity)	More books, applications in international venues More examples of hybrid assessment in the literature	Gupta, Sleezer, and Russ-eft revise their earlier text, Watkins et al. (2012), Kaufman and Guerra-Lopez (2013), Altschuld (2014)

Adapted from Altschuld, J.W., Watkins, R. Needs Assessment: Trends and toward the Future. New Directions for Evaluation, in process.

Another widely used model is that of Kaufman (Kaufman and Guerra-Lopez, 2013). The first author has proposed that there are sequenced dimensions of needs and one works from the highest down to the lowest in organizations, businesses, and institutions. First are Mega needs, those of the society denoted by a healthy population, more production than consumption, and citizens who are employed and have the skills necessary to be successful in their work. This is necessary for a productive society. From there, needs assessors would progress to the Macro dimension which in education would consist of graduates with the prerequisite skills and understandings for job performance or societal involvement. Finally at the lowest level (Micro) needs assessors would determine if the resources, educational personnel, instructional materials, and so forth are there to satisfy macro needs and in turn would lead to the mega needs being resolved.

Kaufman's approach is based heavily on gaps at each step of the sequence which in essence is an interconnected system. In contrast, Witkin's original three phases are about the process of doing an assessment. No matter which approach is used there will be some overlap across the two frameworks.

Methods for Assessing Needs

In 1995, Witkin and Altschuld described more than 20 methods in categories aligned with the three phases; similarly Watkins et al. (2012) described in some detail numerous techniques employed in needs assessment. Assigning exactly where they fit in the conduct of an individual assessment is difficult because of highly varied use when studies of needs are done. A closer examination of procedures suggests a set of common methods. They are often employed in a complementary manner with the use of several in conjunction with each other as the recommended course of action. Most often quantitative and qualitative strategies are paired, with reliance on a single strategy generally to be avoided.

Table 2 The three phases of Needs Assessment (NA)

Phase	*Nature*	*Methods overview*
Phase 1 – pre-assessment	Reconnaissance to see what needs are there and of sufficient importance be pursued further Usually done under the aegis of an organization or community via a Needs Assessment Committee (NAC) The phase relies mainly on existing sources before going to the expensive, time consuming Phase 2 Phase 1 leads to the decisions to ● do nothing, needs are not really there ● go to Phase 2 entailing more extensive data collection, or ● go to Phase 3, action plans to resolve the needs uncovered	Data in records Databases Information from the literature Information routinely kept internally or by external agencies Evaluation reports if available Other reports Informal interviews Observations Additional methods as appropriate
Phase 2 – assessment	If Phase 1 did not provide enough understanding about needs, enter Phase 2 Intensified data collection via in-depth surveying, full analyses of existing sources, seeking more information, and more exploration of the literature Also might prioritize and causally analyze needs to identify potential solution strategies Decisions are to go no further or enough has been learned and prioritized to plan and initiate solutions (Phase 3)	Formal surveying Individual and focus group interviews More literature review about needs and solutions Causal analyses Prioritization strategies Other procedures
Phase 3 – post-assessment	From what was learned in Phases 1 and 2, begin actions to rectify needs Phase 3 overlaps with the latter part of Phase 2 Develop and implement action plans as well as evaluate how well they were working Lastly, include strategies for evaluating the needs assessment itself	Many of the procedures indicated above especially causal analyses and prioritization Action plan-related activities such as benchmarking, multiattribute utility theory, etc. Formative and summative evaluation procedures

Procedures commonly observed are surveys, epidemiology or database studies, and focus groups or individual interviews. Surveys for needs are unique in that they employ, double or triple scales with the former being more notable. As an example, respondents are asked to indicate the importance of a topic, the extent to which it is being achieved or attained, and, on occasion, the frequency of its occurrence. If triple scaled, a need would be an item that is rated high in importance, low in achievement, and as frequently occurring. Five point (Likert) ratings scales are most prevalent and sometimes there are response categories such as don't know (DK) or not applicable (NA). When the latter are chosen they indicate that the respondents do not have much knowledge about a topic.

Epidemiology as a way to study diseases or other issues is almost a perfect fit with needs assessment in its focus on the prevalence (present time occurrence) of a disease and incidence (what will happen if it spreads unchecked) and the gap between the two conditions. This aligns with the definition of need. The AIDs epidemic in the 1980s is an illustration of epidemiology and the investigative work it entails. What was causing the affliction, was a specific segment of the population being affected, what is its progression, and where did it originate were elements of what had to be studied to determine what must be done to stem the progression of the disease. Fortunately, those looking into the situation were able to identify causes, and (although sadly many were lost early to the ravages of the scourge) their efforts paved the way for helpful drug therapies. Another example would be what is the current rate of retirement of teachers and what might that rate be if the pension system were to include more incentives for leaving the workforce?

Epidemiology almost seamlessly blends qualitative and quantitative methodology and is aligned with current needs assessment practice. Hard data are absolutely essential as an indicator of a problem and its rapid escalation, otherwise the researchers would be lost. That data by itself is insufficient to find out about other aspects of what was happening or is likely to happen. Direct contact via intensive interviews with the infected population or those having the potential of contracting a disease is critical in the process. It is equivalent to what police do on a crime scene when they perform a canvas of the immediate area. This is time consuming and requires extensive resources but there is no substitute for the activity.

Notice that interviews are used. Indeed that is one of the key methods that are part of the needs assessment process, either focus group or individual interviews. What are groups thinking about, what might be the issues on their minds, do others in the group have similar or different views, and so forth. Individual interviews might be done with key informants, those that know the pulse beat of a community or organization. What can they tell us about community views in an in-depth questioning session, and if patterns emerge across responses, can be quite informative and cost effective as compared to large-scale surveys of communities.

Another version of the procedure is rapid or community-based appraisal. Members of a community are recruited and trained to conduct quick and short interviews of a convenience sample in the locality. If there are 10 'interviewers' each doing 10–15 rapid interviews, much can be gained in a short period of time. Community members are often easily trained for the task and is much cheaper due to the substitution for more expensive professional personnel. There may be some loss of quality but not that much in what is derived from the procedure, and there is a positive compensating dimension in the rapport between those providing perceptions and those collecting them. The use of this approach applies both to urban and rural settings.

Before leaving this topic, several other methods merit comment. Many times numerous important needs will be identified but attending to all of them will extend and exhaust resources beyond the breaking point. Priorities must be established and the causes of discrepancies understood. Assessors must be familiar with how to set priorities (not all needs require immediate attention and are of equal importance and meaning to decision-making groups). Causal analysis is imperative to look into the nature of what factors have led to the need. Such activities take place in Phase 2 of the process or early in Phase 3 and will be briefly addressed in the next section.

Table 3 Reasons for doing a Needs Assessment (NA)

Reason	*Explanation*
Human nature	It is how we conceptualize and attack problems It is just part of what we do as people
Systematic strategy for planning based on needs	Requires identification of needs, prioritizing them, analyzing their causes, and determining possible ways to improve predicated on causes It makes sense as a way to move forward
Establishing the basis for formative evaluation of selected solution(s)	Understanding causes leads to better solutions being implemented and where to direct formative evaluation resources May indicate where a solution will most likely fail Leads to a targeted and more selective strategy for rectifying needs
Overall improved program accountability	If the above activities are done well, they help to uncover quality solutions and make for meaningfully being able to demonstrate accountability Adds an enhanced level of sophistication to what we do

Why Is It Important to Deal with Needs?

Now that it has been established what the needs assessment is, why should we be concerned with the whole idea, what is the rationale for doing one (see Table 3)? Hansen (1991) explored, what people do when attacking an issue, a problem, or a concern, and found that despite differences in terminology across fields and contexts there was a core strategy revolving around what is and what should be, and the discrepancy between them which is the very heart of what a need is. The consideration of needs is inherent in planning and human exploration of how to proceed. Therefore one reason is our human makeup and what writers about the topic have simply created a reconstructed logic of what the process is like (see Table 3).

A systematic way to think about planning is the second reason that needs assessment is important. It means that the assessors have thoroughly (not superficially) investigated the topic of concern. They have identified and prioritized needs of most importance for a group, organization, agency, or business to focus on and to which resources should be devoted. It shows dedication to detailed probing of the problem or issue. And if causal analysis has occurred it is likely that a better solution will have greater likelihood of being chosen for implementation and it should have enhanced potential for resolving the underlying problem. Overall the solution should be more in tune with the nature of the discrepancy.

The third reason that assessment is important is that since we have learned what causes the problems and how causal factors operate, we anticipate where the process of change and development may tend to fail. This would tip off where to focus and direct formative evaluation efforts and resources given that not everything can be evaluated. If things were not going right we should be able to see them early, not late, in the process of change and take appropriate corrective actions or modifications. This should increase the probability of success.

Collectively the three reasons point to the conclusion that there is an improved chance of producing programs and solution strategies that achieve the goal of resolving needs. This implies higher program accountability. Thus it seems advantageous to always do a needs assessment, though frequently that is not the case.

What Affects the Implementation of Needs Assessment

Reflect on needs assessment as an exploration of an agency, a community, a company, and so forth. Like a research study, it is very dependent on appropriate conceptualization. But the goal of research is to enhance understanding of phenomena and fundamental propositions about them. Needs assessment is different from research in the use of multiple methods, involving members of the group for which it is done, resulting in change and concrete organizational actions, requiring more of a facilitative stance on the part of the individual or group leading it, and consisting of many steps where it could go awry – lots of things could negatively impact an assessment. Also, depending on the scope of what is being examined (extent and seriousness of the problem, how many people are affected by it), and other factors, needs assessment can be time consuming, expensive to conduct, and usually will end up in resources being directed to one need over another. The latter part of the statement points to how political the

enterprise can be. Someone or some segment of the organization will win and by the same token some will lose.

Focusing so that the endeavor is not so broad that nothing will be accomplished, or so narrow that the bigger picture is missed, is also an important feature. Needs assessment is a bit of an art and thinking through how to start is glossed over in practice, resulting in less than stellar results. More attention should be paid to how to proceed in what can be a complex undertaking so that meaningful outcomes are achieved. Refer back to Table 2 and note Phase 1, see it as the make or break point for success. Has the heart beat of the community or agency been uncovered? Is the assessment dealing with issues that a community values and would the assessment lead to careful thought being given to outcomes? Are the variables examined or to be examined attuned to the group, and would the group be interested in what has been found? If questions like these are not dealt with at the beginning, the probability of change tends to decrease.

This relates to another concern, short-term assessments are estimated to require about 3 months not including getting the solution strategy off the ground. Major needs will take perhaps a year for study or even longer before a solution is in place. In the latter case and with the involvement of a NAC it can be anticipated that maturation (getting bored, losing patience, tiring of the endeavor), as a Campbell and Stanley source of invalidity will arise. If the imagination of the NAC is not captured, motivation probably will deteriorate. Most groups are action oriented and with time-consuming ventures they may become impatient. Extra effort will be necessary to keep the group involved and there is not enough discussion in the literature about maintaining interest and facilitating a work-oriented group (see Spaeth et al., 2013). These are the ingredients that produce a savory dish instead of one that nobody wants to eat.

Other factors affecting the quality of a needs assessment are double- or triple-scaled instruments and multiple methods/sources of data. What on the surface appears to be a simple proposition is not. Altschuld and Hamann (2012) identified problems in wording, prioritizing needs, improper analyses, not fully understanding response options, and so on. Add in that we might want to ascertain the risks inherent in meeting needs or not doing anything about them (that counterfactual state). If not attended to, might risk dissipate and no action is required? Are they really worthy of the investment of time and resources?

Compounding the process is that assessors should employ multiple methods consisting of existing records and databases, interviews, observations, surveys, and group processes, and the difficulty of the task magnifies exponentially. Do assessors have the requisite skills and understandings of diverse techniques? Are they equally adept at all of them? Even if they are, do they have experience in pulling the data collected into a comprehensive summary that integrates findings? What if the results do not agree or in the extreme are contradictory? What approaches guide those doing the work in establishing priorities from such an array of sources? How should they engage with community-based committees to guide them in decision-making?

These concerns are not easy to deal with, but on the other hand they are the excitement of needs assessment. They raise the possibilities for research into the process and for ways to move our work with needs to a higher level. Offering them here sensitizes those doing the assessment to the subtle aspects of needs and the in-depth thought they require.

What of the Future for Needs Assessment

Technology and methodology is evolving. Recently the author ran across a study on health being conducted in the US in which approximately 1 000 000 individuals were given simple I-pad devices to record daily readings on their health. This might appear expensive due to the costs of the devices and for respondent involvement, but putting it in perspective it would produce vast amounts of information that would likely generalize to the entire population and quickly lead to an amazing database readily available for study. The costs, while not unnoticeable, would be relatively small for being able to extrapolate to in excess of 300 000 000 people. Earlier and not so advanced versions of such studies, but still expensive ones, were the Epidemiological Catchment Study of the 1980s, the West of Scotland Study (WOSCOPs), and NHANES (Centers for Disease Control) which is an ongoing investigation of children's health and well-being.

Consider the Geographic Information Systems (GISs) and smartphone/table applications (app) that are now available. The author's son and his family were planning a vacation and found a book about the intended destination that offered an app about recreational sites for an additional eight dollars. It included ties to the GPS so they could easily navigate their travel. If a group was thinking about solutions to their needs that included recreational opportunities, this was a godsend and due to changes in technology it was at their fingertips for a nominal cost. Another advancement is the GIS that Hites et al. (2013) creatively used in conjunction with a qualitative technique (focus group interviews, FGIs) to investigate needs about crime in the vicinity of a large university in a major southern city. They included database information and across sources were able to present concrete portrayals of day- and night-time problems. Aside from the FGIs the other information already existed and simply had to be extracted from such sources.

Yet a different technique is photovoice that has been part of a number of studies. It was used in two ways to uncover needs and/or assets in a community. Inexpensive cameras (cell phones would now be even better) are provided to members of a community or organization and they, with some limited guidance, go out and photograph exemplars of problems, assets (strengths), or both. Once this is done they are assisted in writing descriptions of their photos and the pictures with explanations become the emphasis of discussion in a large meeting involving the broader community. The generation of information and its interpretation are largely under the aegis of the community, not those guiding or facilitating the assessment (the voice of the people is prominent, a concept that has traction in recent work on joint asset/capacity building and needs assessment). It is a very different approach for collecting data and from a perspective perhaps not captured by more traditional investigators or investigations.

Lastly, this text would be remiss if it did not refer to a very sharp challenge to the premise of needs assessment itself.

A counterview emphatically proposed by Kretzmann and McKnight (1993) in their asset-oriented book is that starting with needs is a negative, 'downer,' not an upbeat activity and will not lead to progress, change, and improvement. In the worst case it generates a dependency mindset, particularly for impoverished communities; looking through the needs lens, they will never recognize their strengths and never develop a sense of self-reliance.

A need represents a problem, something that is wrong and we as humans do not plan on that basis. Rather we should look to our assets and resources and view them as the building blocks for moving forward. If we focus on what do we bring to the table to a high degree, we can be in control of our own destiny. Whether we adhere to a needs stance or an asset one, there is merit in the criticism of the above two authors. But is it best to begin with strengths or needs (a rhetorical question if there ever was one)?

A practical alternative is beginning to be seen in the literature. Since about 2000 hybrid studies are appearing in public health, community development, and other areas. They start with assets and resources that a community or group can tap into and utilize. The positive foundation of assets was abundantly clear to the authors of the pertinent articles and was often directly stated as a governing principle for their investigation.

At the same time the authors were not oblivious to the importance of needs and were aware of the prominent role they play in contemplating new directions and in leading to developments that enhance the community or organization. Such hybrid studies are more complex than a stand-alone needs assessment or an asset/capacity building effort and will take considerable time. That is reasonable given issues such as crime prevention, poverty, gun control, health of groups, agricultural services, etc. It is in such contexts that the hybrids are seen. In other words the time and resources to do both and pull together more complicated data is justified because of the complexity of the underlying problems. We cannot and should not expect easy solutions to long-standing social issues; joint asset/capacity building and needs assessment endeavors are a meaningful and reasonable way to proceed.

Summary

What has been provided is an overview of a deceptively simple concept that masks what lies beneath the surface. There are many variations in meaning and ways to frame ideas about needs, and how to assess them so the picture quickly becomes cloudy and the vision gets blurry. Therefore it is to be stressed to keep discrepancy foremost when thinking about needs and that they are fundamental aspects of human activity whether that has been stated or is tacit.

Implicitly, needs are at the heart of everything we do but it is important to distinguish those for individuals from those of organizations, agencies, businesses, government, and the like. This article is about the latter and that opens the door to a myriad of questions such as how are needs determined across multiple groups in different organizations and settings (multifaceted businesses, state and local agencies, and school systems). What happens when perspectives do not coalesce (often they will not) and how can differences be reconciled? What should we do with many important needs that are all competing for finite and precious resources? How are decisions made as to which ones are attended to and which are not?

This is the deeper subtle side of assessing needs. This makes it exciting and a challenge. It is not a cut-and-dried cookie-cutter approach – follow the steps in the process and the needs will be identified and rectified. Every context has uniqueness, so a general approach is fine but must be tempered and tailored to specific nuances and parameters. This makes the assessment of needs more of an art where one has to navigate rough waters and many turns and twists along the way. The assessor must be ever vigilant to constantly changing dimensions. It is never dull.

That being said, the goal was not to deter you from dealing with how needs are assessed and resolved. Rather it was to identify issues and concerns about the journey so that you will be attuned to the hazards that likely will be encountered. The hope is that it has whetted your appetite for digging more deeply into the topic. Many references have been given as guideposts as you begin on your way.

See also: Applied Social Research, History of; Cost-Benefit and Cost-Effectiveness Analyses in Evaluation Research; Epidemiology, Concepts and Methods of; Ethical Questions in Social and Behavioral Sciences, History of; Evidence-Based Practice; Implementation Science; Methodology of the History of the Social and Behavioral Sciences; Policy Analysis; Prevention Research; Qualitative Research; Quantification in the History of the Social Sciences; Resilience; Social Epidemiology.

Bibliography

Altschuld, J.W., 2014. Bridging the Gap between Asset/Capacity Building and Needs Assessment: Concepts and Applications. Sage Publications, Thousand Oaks, CA.

Altschuld, J.W., Watkins, R., 2014. Needs Assessment: Trends and toward the Future. New Directions for Evaluation.

Altschuld, J.W., Watkins, R. (Eds.), 2014. A primer on needs assessment: more than 40 years of research and practice. In: Needs Assessment: Trends and Toward the Future. New Directions for Evaluation. Chapter 1.

Altschuld, J.W. (Ed.), 2010. The Needs Assessment Kit. Sage Publications, Thousand Oaks, CA.

Altschuld, J.W., Hamann, M.S., October 2012. Overlooked dimensions in needs assessment surveys. In: Demonstration Session at the Annual Conference of the American Evaluation Association, Minneapolis, MN.

Altschuld, J.W., Kumar, D.D., 2010. Needs Assessment: An Overview. Sage Publications, Thousand Oaks, CA.

Altschuld, J.W., Witkin, B.R., 2000. From Needs Assessment to Action: Transforming Needs into Solution Strategies. Sage Publications, Thousand Oaks, CA.

British Museum, 2011. Journey through the Afterlife: Ancient Egyptian Book of the Dead. Retrieved from: http://www.britishmuseum.org/channel/exhibitions/2011/book_of_the_dead.aspx.

The CADISPA Trust, 2013. About Us. Retrieved from: http://www.cadispa.org/.

Carnegie UK Trust, 2011. Exploring Community Resilience in Times of Rapid Change. Fiery Spirits Community of Practice, Dunfermline, Scotland.

Centers for Disease Control and Prevention/National Center for Health Statistics, 2013. About the National Health and Nutrition Examination Survey. Retrieved from: http://www.cdc.gov/nchs/nhanes/about_nhanes.htm.

Gupta, K., Sleezer, C.M., Russ-Eft, D.F., 2007. A Practical Guide to Needs Assessment. Pfeiffer, San Francisco, CA.

Hansen, D.J., 1991. An Empirical Study of the Structure of Needs Assessment. Unpublished doctoral dissertation. The Ohio State University, Columbus.

Hites, L., Fifolt, M., Beck, H., Su, W., Kerbawy, S., Wakelee, J., 2013. A geospatial mixed methods approach to assessing campus safety. Evaluation Review 37, 347–369.

Kamis, E., 1979. A witness for the defense of need assessment. Evaluation and Program Planning 2 (1), 7–12.

Kaufman, R., October 1987. A needs assessment primer: a ten-step approach to help even beginners with this necessary, yet often dreaded task. Performance and Instructional Journal 78–83.

Kaufman, R., Guerra-Lopez, I., 2013. Needs Assessment. ASTD, Alexandria, VA.

Kretzmann, J.P., McKnight, J.L., 1993. Building Communities from the Inside Out. ACTA Publications, Chicago, IL.

Lauffer, A., 1982. Assessment Tools: For Practitioners, Managers, and Trainers. Sage Publications, Beverly Hills, CA.

Lee, J., 2013. Heart Study to Enroll 1 Million Smartphone Users Worldwide. The Week in Healthcare, Technology. Retrieved from: http://www.modernhealthcare.com/.

McKillip, J., 1987. Needs Analysis: Tools for the Human Services and Education. Sage Publications, Thousand Oaks, CA.

Neuber, K.A., Associates, 1980. Needs Assessment: A Model for Community Planning. In: Human Services Guide, vol. 14. Sage Publications, Beverly Hills, CA.

Scriven, M., Roth, J., 1990. Needs assessment: concepts and practice. Reprint in Evaluation Practice 11 (2), 135–144.

Spaeth, T., Larson, N., Wagner-Lahr, J., 2013. Overview: Research and promotion council forum. Agricultural Utilization Research Institute, Minnesota.

Stevahn, L., King, J.A., 2010. Needs Assessment: Taking Action for Change. Sage Publications, Thousand Oaks, CA.

Stufflebeam, D.L., 1968. Evaluation as Enlightenment for Decision Making. Ohio State University Evaluation Center, Columbus.

Warheit, G.J., Bell, R.A., Schwab, J.J., 1979. Needs Assessment Approaches: Concepts and Methods. National Institute of Mental Health, U.S. Department of Health, Education, and Welfare, Rockville, MD.

Watkins, R., Guerra, I., 2002. How do you determine whether assessment or evaluation is required? ASTD T&D Sourcebook 131–139.

Watkins, R., West Meiers, M., Visser, Y.L., 2012. A Guide to Assessing Needs. International Bank for Reconstruction and Development/International Development Association or the World Bank, Washington, DC.

Wedman, J. Needs assessment in the private sector (Chapter 4). In: Altschuld, J.W. Watkins, R. (Eds), Needs Assessment: Trends and toward the Future. New Directions for Evaluation, in press.

West of Scotland Coronary Prevention Group (WOSCOPS), 1996. West of Scotland Coronary Prevention Study: identification of high risk groups and comparison with other cardiovascular intervention trials. Lancet 348 (9038), 1339–1342.

Witkin, B.R., 1984. Assessing Needs in Educational and Social Programs: Using Information to Make Decisions, Set Priorities, and Allocate Resources. Jossey-Bass, San Francisco, CA.

Witkin, B.R., Altschuld, J.W., 1995. Planning and Conducting Needs Assessments: A Practical Guide. Sage Publications, Thousand Oaks, CA.

Negation, Linguistics of

Johan van der Auwera and Denies Du Mon, Universiteit Antwerpen, Antwerpen, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abbreviations

ACC accusative
AFF affirmative
COMP completive
DISTPST distant past
FFNLV fixed form of negative lexical verb
GEN genitive
M masculine
NEG negation
OBJ object
PFV perfective
PL plural
POT potential
PROH prohibitive
PRS present
PRT particle
PST past
SBJ subject
SG singular
SUBJ subjunctive
T tense

Abstract

Negation refers to a kind of meaning, most intuitively paraphrased in terms of a speaker denying or contradicting something. This article deals with several aspects of negation from a typological point of view. It discusses the meaning and the form of negation. It distinguishes between scope and focus in the meaning section and contrariness and contradiction. The article also offers an overview of some different strategies for expressing negation across languages.

Meaning and Form

The term 'negation' refers to a kind of meaning, most intuitively paraphrased in terms of a speaker denying or contradicting something.

(1) The King of France is *not* bald.

If one has any occasion to use sentence (1), one will use it to deny or to contradict the view that the King of France is bald. One communicates this negative meaning with a linguistic sign, viz., with the word *not*. This word can be called a 'negation,' a 'negative marker/adverb/particle' or a 'negator.' Obviously, the meaning and the form are related: one could say that the English word *not* is dedicated to the expression of a negative meaning. However, the dedication is not complete. Not every use of a negative marker actually conveys a negative meaning.

(2) Je crains que Jean *ne* vienne.
I fear that Jean NEG come.SUBJ.PRS.3SG
'I fear that John is coming.'

In French (2), the complement of a verb of fearing may contain the negative element *ne*, but what the speaker fears in (2) is not that John would not be coming, but rather that he would. Conversely, not every transmission of a negative meaning needs a unique dedicated negative marker. In (3) the same negative meaning allows two strategies, and neither is more dedicated than the other.

(3) (a) I saw *nothing*.
(b) I did*n't* see *anything*.

In what follows the focus will first be on meaning, and then on form. In the section on form, some attention will also go to the diachrony of negation. A more extensive, recent, multiauthored discussion of negation is Horn (ed.) (2010).

Meaning

Scope and Focus

In this section, we differentiate between the scope and the focus of negation. Negation can have scope over a number of constituents or over just one constituent. In Shakespeare's *Julius Caesar*, Brutus has to explain to the citizens of Rome why he killed Caesar. His speech includes the following words:

(4) Not that I loved Caesar less …

Sentence (4) starts with the negator *not* and is said to be 'within the scope of negation.' It is a statement that some citizens of Rome might have made in the context of Brutus' murder of Caesar: if one murders someone, one probably does not care for the victim too much. We here hit upon a typical feature of negation: when a proposition is negated, it is usually entertained in the context.

Note that though *not* has scope over an entire sentence, the negation does not really concern each constituent of the

sentence. What is denied is only the degree of Brutus' love of Caesar, compared to his love of Rome. For this more restricted sense of aboutness, one often uses the term 'focus' of the negation, and so in (4) the focus is the degree of Brutus' love and this aboutness is not reflected in the syntax. In (4) the syntax is simply vague with respect to the focus of the negation. Even when we integrate the negation into the sentence, as in (5), the focus may still be syntactically vague.

(5) I did not love Caesar less,
  (a) … I loved Rome more
  (b) … Cassius did
  (c) … I loved Antony less

Although in (5) the negation is syntactically associated with the verb, it does not necessarily negate the verb. It remains a 'sentence negation.' This is different in (6), which is part of what Mark Antony has to say to the Romans.

(6) I come to bury Caesar, *not* to praise him.

In (6) the scope is still a whole sentence: Mark Antony denies the truth of the *I came to praise Caesar*. However this time the focus of the negation is much clearer: it is the constituent *to praise Caesar*. This use of the negation can be called 'constituent negation.'

The study of negation and its relation to the truth of the sentence is not only a concern for linguists. It has been on the agenda of logic for as long as the discipline exists. Within logic, one can approach negation from the perspective of scope. One is then necessarily doing propositional logic, whereby negation will be a propositional operator. When one looks at the focus of negation, however, one can enter the internal structure of the proposition and one can enrich the study of quantifiers (*all*, *some*) with the study of negative quantifiers (*not all* or *no*, *not some*). A term for the part of the scopal proposition that is out of the focus of negation is 'presupposition.' The term was introduced into twentieth-century philosophy of language by Strawson (1950). It was a much debated notion in the 1970s and attempts were made to turn it into a pivotal notion in the study of the interface between semantics and pragmatics (e.g., Levinson, 1983). As to pragmatics, note that the quote in (4) is only part of what Brutus said: the full utterance is (7). After identifying the lack of love as the wrong motive for the assassination, Brutus goes on to identify the real motive, viz., the love of Rome.

(7) Not that I loved Caesar less, but that I loved Rome more.

Without the part of the sentence that starts with *but*, the audience would not be satisfied. In general, people are more interested in what is the case than in what is not the case, and in the context at hand, the Romans are more interested in what motivated Brutus to kill Caesar than in what did not motivate him. As compared with positive sentences, negative sentences are marked and relatively infrequent.

Contrariness and Contradiction

(8) (a) Mark Antony was *unhappy*.
  (b) Mark Antony was *not happy*.

In both versions of the sentence in (8) we are arguably dealing with constituent negation: in both the speaker denies the appropriateness of the predicate 'happy.' However, the two versions can differ in meaning. In (8a) Mark Antony is reported as being sad. (8b) is vague: Mark Antony could indeed be unhappy too, but he could also be neither happy nor unhappy, or he could even be 'very happy.' Both 'unhappy' and 'neither happy nor unhappy' are instantiations of 'not happy,' and 'not happy' is called the 'contradictory negation' of 'happy.' On the other hand, 'unhappy' is the 'contrary' negation of 'happy.'

Like the distinction between sentence negation and constituent negation, the one between contradiction and contrariness exists since the beginnings of logic, most prominently since the formulation of the Aristotelian 'Square of Oppositions,' implicit in Aristotle but explicit first in the work of the Apuleius of Madaura (second century BC, Londey and Johanson, 1987). A version of this square, as applied to quantifiers, is shown in Figure 1.

Just like 'unhappy' is the contrary of 'happy,' 'all not' is the contrary of 'all,' and just like 'not happy' and 'happy' exhaust all logical possibilities and are thus each other's contradictories, so 'not all' and 'all' are contradictories.

Notions such as contrariness and contradictoriness have been a concern to both logicians and linguists. What interests linguists is that these notions offer a framework for the study of lexicalization patterns. Just like English *unhappy* is a lexicalization of the contrary of *happy*, English *all* has its contrary lexicalized as *no*. The contradictory 'not happy' does not allow for a lexicalization, and neither does 'not all.' What these examples illustrate is that contrariness lexicalizes more easily than contradictoriness (at least in English), which is not to say that the latter does not lexicalize at all (thus English *non-*, different from both *un-* and *in-*, often produces contradictories, e.g., *nonscientific*). Linguists have also made it clear that contradictory notions often change their meanings into contrariness. A classic case, known since Tobler (1882), is French *il ne faut pas*. It used to mean 'it is not required,' which is the contradiction of 'it is required,' but it now only means 'it is forbidden,' which is the contrary of 'it is required.' The linguist that has done most for making the logical tradition relevant for the linguistics of negation and for discussing issues of lexicalization and polysemy is Laurence R. Horn (esp. 1989). Horn has also looked at what has come to be known as 'negative raising' from this perspective. What happens with negative raising is that a negation whose meaning is felt to belong to

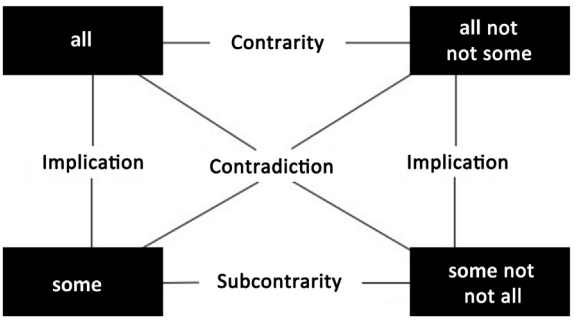

Figure 1 Aristotelian square – square of oppositions.

a complement clause is actually expressed with the complement taking main verb. Consider (9).

(9) (a) I do*n't* suppose that you believe him.
(b) I suppose that you do*n't* believe him.

Using a transformational perspective, the negation can be seen as 'raised' from the lower clause (*you believe him*) to the higher clause (*I suppose*). The point that is relevant here is that the two versions of the sentence can be characterized in terms of the distinction between contrariness and contradictoriness. The raised (9a) is literally a contradictory negation (either one supposes something or one does not), but it is normally equivalent with the nonraised (9b) and it then conveys a contrary meaning (instead of supposing *p* and supposing *not p*, one could also suppose neither *p* nor *not p*).

Form

Negation can take many forms, both morphological and syntactic, and occupy different positions within the sentence. This section deals with this matter. The distinction between *unhappy* and *not happy*, just touched upon, obviously has a formal dimension too. In *unhappy* the negation is expressed with morphology, more specifically with a prefix. In *not happy* the negative marker is an independent word, an adverb or particle – one can speak about 'syntactic negation.' Semantics and form are related. With *unhappy* versus *not happy*, affixal negation is associated with contrariness, and syntactic negation is associated with contradictoriness. The relation is not very tight, however. The syntactically negated *il ne faut pas que*, after all, expresses a contrary meaning in present-day French. Conversely, in *nonscientific*, the prefixal *non*-strategy results in contradictoriness.

Crosslinguistically, negating with morphology and negating with a separate syntactic element are the two basic strategies. Sound or word order differences may have a function too, but they rarely or never mark negation by themselves. Both morphological and syntactic negation exhibit subtypes. The linguist's increased understanding of the typology of morphological and syntactic negation is directly related to the increased importance of typology in general, in the wake of Greenberg (1966). Important works in the typology of negation are Dahl (1979), Dryer (1988), Forest (1993), Honda (1996), Kahrel (1996), Miestamo (2005) and, somewhat indirectly, Haspelmath (1997). The great variation in the expression of negation, both in one and the same language and across languages, has also been a challenge for formal models of grammar, e.g., in generative grammar (Haegeman, 1995), lexical functional grammar (Knüppel, 2001), or optimality theory (de Swart, 2010).

As to syntactic negation, there are two subtypes: either the negator is an uninflected element (adverb or particle) or it is a verb (and, depending on the criteria used, an auxiliary verb). English *not* is an example of an uninflected element. For an example of a negative verb, one can turn to the Tungusic language, Evenki. Evenki has a negative verb, whose stem is *e-*, which shows agreement with the subject in person and number, and which furthermore inflects for tense, aspect, and mood. It combines with what Nedyalkov (1994) calls the 'fixed form of the negated lexical verb' ('FFNLV').

(10) Si min-e *e-che-s* doldi-ra.
you I-ACC NEG-PST-2SG hear-FFNLV
'You did not hear me.'
(Nedyalkov, 1994: 16)

Negative verb forms are arguably also the English form *don't* as in (11) or the Classical Latin negative imperative construction with *noli* in (12).

(11) I *don't* like reggae.

(12) Noli venire.
'Don't come.'

The Evenki negative verb differs from the English and Latin ones, however, in that a morphological analysis of the latter will allow one to identify a morphological negator, too. The element *don't* contains the morphological (enclitic) negator *n't*, and *noli* is the imperative of the negative *nolle* 'not want,' which is a combination of *velle* and a prefixal (or proclitic) *ne*. This does not detract from the fact that *don't* and *noli* count as syntactic negators. To negate (13) and (14), the morphological negator cannot attach to the main verb, nor can it be replaced by a negative particle.

(13) (a) I like reggae.
(b) *I like*n't* reggae
(c) *I like *not* reggae
(d) *I *not* like reggae.

(14) (a) Veni '
Come!'
(b) **Ne*veni.
(c) **Non* veni.
(d) *Veni *non*.

It has been argued that if a language has a special negative verb, it will frequently be an existential verb or a verb of wanting, and that these specialized constructions may lose their verbal properties, thus turning into particles (Payne, 1985: 222; Croft, 1991). It has also become clear (esp. through Honda, 1996) that when a language employs a special verb to express negation, this special verb does not itself have to be negative. In the Carib language, Hixkaryana, negation can be expressed with an affix on the verb, but this construction further requires a positive copula 'be.'

(15) (a) Koso w-ono-ye
deer 1SBJ/2OBJ-eat-DISTPST
'I ate the deer.'
(b) Koso y-ono-*hra* w-exe-ye.
deer 3-eat-NEG 1-be-DISTPST
'I didn't eat the deer.'
(Derbyshire, 1979: 48)

A literal paraphrase of (15b) is 'I was not/without eating the deer,' and there is a sense in which the sentence is not really negative.

Note that each of the syntactic negators illustrated thus far precedes the main verb. There seems to be a consensus that languages generally prefer to have their negation in front of the main verb, an observation that was perhaps first made by Jespersen (1917: 5), named the 'NEG FIRST' principle by

Horn (1989: 293, 449–452), and found confirmation in large-scale typology in Dahl (1979) and Dryer (1988). However, a postverbal position is by no means rare, as in Dutch (16), or as when negation is sentence-final, i.e., post-everything, and hence also postverbal, e.g., in the Chadic language, Ngizim, in (17).

(16) Ik hou *niet* van reggae.
I hold NEG of Reggae
'I don't like reggae.'

(17) Ii naa duuka *bai*.
I have horse NEG
'I don't have a horse.'
(Thompson, 1998: 328)

Whether the NEG FIRST preference also holds for morphological negation is not clear (for conflicting views see Dahl, 1979: 82 and Bybee, 1985: 177). In any case, morphological negation can be prefixal, already illustrated with Latin *noli* in (12), as well as affixal, as in English *isn't*.

Negation, whether syntactic or morphological, may also involve a part that precedes and a part that follows. This is often called 'discontinuous' or 'embracing' negation. It is illustrated with French (18) and with Cairene Arabic (19) (Bernini and Ramat, 1996: 47).

(18) Je *ne* le lui donne *pas*
I NEG it him give NEG
'I don't give it to him.'

(19) *Ma*-katab-m̄-š.
NEG-write. PFV-3PL-NEG
'They didn't write.'

Discontinuous negation is a special case of double negation. It is particularly perspicuous and well-studied in European languages (Bernini and Ramat, 1996), less so in the languages of Africa (Devos and van der Auwera, 2013) and the Pacific (Vossen and van der Auwera, 2014). The *locus classicus* is Jespersen (1917), and what has come to be known – through Dahl (1979: 88) – as 'Jespersen's cycle', is the view that languages may go from single negation to double negation and then back to single negation. The double negation is usually taken to result from some kind of emphasis, though the pragmatic notion of emphasis is itself not all that clear (Schwenter, 2006; Hansen and Visconti, 2009), after which the new single negation may result from the bleaching of the emphatic meaning. The phenomenon was already described before Jespersen (1917); very clearly so by Meillet (1912), who spoke of this cycle as a 'spiral' and used it to illustrate the more general process of grammaticalization. We now know that the multiple exponence of negation need not be restricted to double negation, for languages sometimes manifest triple and rarely even quadruple negation. Nese, a language of Vanuatu, normally has double, embracing negation, with a realis–irrealis alternation in the preverbal negative prefix, with *se* being for realis and *be* for irrealis, but the two preverbal prefixes can be concatenated, thus filling the first slot twice, and as a result we get three markers.

(20) Nese (Crowley, 2006: 70)
Ø-se-be-yes-*te*
3SG.NEG-NEG-walk-NEG
'(S)he did not walk.'

In Lewo, another Vanuatu language, older speakers may double up their preverbal realis prefixes (*pe*) too, but this time with the particle that is normally postverbal, which may still occur in its usual postverbal place, and there may even be a clause final negation – thus resulting in four negations.

(21) Lewo (Early, 1994: 420)
pe-re a-pim *re* *poli*
NEG-NEG 3SG.come NEG NEG
'They didn't come.'

(21) also shows that multiple exponence may both involve the repetition of a negator (*re*) or the occurrence of more than one different negator (*pe*, *re*, and *noli*). An example of a language that doubles, like French, but through repetition, unlike French, is Afrikaans.

(21) Ek sal *nie* kom *nie*
I shall NEG come NEG
'I won't come.'

Note also that language need not have double negation in the French style to have more than one negation. Quite often languages negate declarative and imperatives with different markers (van der Auwera et al., 2005), and sometimes there is more than one marker for each environment. Mandarin, for instance, has at least three (and arguably five) types of declarative negation, viz., *bu* and *méi* (*yǒu*), and at least two prohibitive negations, viz., *bié* or *buyào* (Li and Thompson, 1981: 417, 418, 455).

(22) (a) Tā *bu* niàn shū
3SG NEG study book
'(S)he does not study.'
(b) Tā (*méi*)*yǒu* kāi mén.
3SG NEG open door
'(S)he didn't open the door.'
(c) *Bié*/*buyào* dòng!
PROH move
'Don't move!'

The negation of existential, locative, or adjectival predicates also often needs special negations. In Serbian, for example, the normal negation is *ne*, but, at least in the present, in an existential proposition a special negation is used, a univerbation containing the normal negation (Veselinova, 2010: 199–200).

(23) (a) Meri pev-a,
Mary sings
'Mary sings.'
(b) Meri *ne* pev-a.
Mary NEG sings
'Mary does not sing.'
(c) Ima divlj-ih mačak-a
has wild-GEN.M.PL cat- GEN.M.PL
'There are wild cats.'
(d) Nema divlj-ih mačak-a
not-has wild- GEN.M.PL cat- GEN.M.PL
'There aren't any wild cats.'

It has been claimed that special negations of this type may take over from the normal negation and that we thus have

another type of negative cycle, different from Jespersen's cycle (Croft, 1991; Veselinova, 2010).

A rare manifestation of negation, even rarer than quadruple negation is 'zero negation.' In South Dravidian languages, negation can be expressed by the mere absence of a tense suffix. Havyaka Kannada, spoken in the coastal areas of the Indian state of Karnataka, makes a distinction between indicative and subjunctive, and between past and nonpast. In (23a–b) the verbal forms have a tense mood slot filled by the nonpast indicative and subjunctive forms. In (22c) this slot is unfilled, and this in itself marks the structure as negative subjunctive (van der Auwera and Bhat, 1999).

(23) (a) ma:śTrakko kate o:du-tt-avu
teachers story read-IND.NPST-3PL
'The teachers are reading a story.'
(b) ma:śTrakko kate o:du-g-u
teachers story read-SUBJ.NPST-3PL
'The teachers may be reading a story.'
(c) ma:śTrakko kate o:d-Ø-avu
teachers story read-SUBJ.NEG-3PL
'The teachers may not be reading a story.'

Zero negatives may be rare, but they are not exclusive to Dravidian. They have also been reported for the Tupí language Karitiána, spoken in the Amazon Basin. According to Landin (1984), affirmative sentences mark verbs with a preverbal particle or clitic expressing affirmation. In negative sentences, however, there is no prefix or clitic.

(24) (a) Y ta-oty-j ỹn
I AFF-bathe-T I
'I will bathe.'
(b) Y oty ỹn
I bath I
'I will not bathe.'
(Landin, 1984: 237)

A general discussion of this phenomenon is offered by Miestamo (2010).

Note that the verbs in the Karitiána and Havyaka Kannada negative sentences lack tense markers. This touches upon a large area of investigation: the relation between negation and other categories. A forceful illustration of the intricacy of this relation is provided by the Mixtecan language, Copala Trique. The illustration makes the following points: (1) a language may have more than one negative marker, (2) the choice between them may be related to the choice of other categories, in this case a tense-aspect-mood ('TAM') category, and (3) the compositionality of joining the negative marker and the TAM marker need not be very transparent. In (25a) the prefix *g*-marks completive TAM. The language has two negative markers, *ne̥*^{3} and *ze̥*^{4}. The construction in (25a) combines only with *ze̥*^{4}. But then the TAM changes and the negated completive becomes what Hollenbach (1976: 126) calls 'potential.' Note that the superscripts in the examples mark tone contours.

(25) (a) g-ucu̥h^{34} žini^{3} yuwe^{21} âh
COMP-lay boy palm.mat PRT
'The boy laid the palm mat down.'
(b) *ze̥*^{4} g-ucu̥h^{34} žini^{3} yuwe^{21} âh
NEG COMP-lay boy palm.mat PRT
'The boy won't lay the palm mat down.'

To reach the meaning 'The boy didn't lay down the palm mat,' one has to start from a potential aspect.

(25) (c) *ne̥*^{3} g-ucu̥h^{4} žini^{3} yuwe^{21} âh
NEG POT-lay.POT boy palm.mat PRT
'The boy didn't lay the palm mat down.'

Negation may also relate to the shape and the meaning of referring expressions. In Russian, for instance, objects that take Accusative cases in positive environments may allow or prefer Genitive case in negative environments (Timberlake, 1993: 869).

(26) (a) On soxranil podlinnik pis'm-a
he preserved original.ACC letter-GEN
'He preserved the original of the letter.'
(b) On *ne* soxranil podlinnik/ podlinnik-a pis'm-a.
he NEG preserved original.ACC/ original-GEN letter-GEN
'He didn't preserve the original of the letter.'

Or consider English (27): the word *something* is normally replaced by (c) *not ... anything*.

(27) (a) I heard *something*.
(b) ?I didn't hear *something*.
(c) I did*n't* hear *anything*.

That the addition of a negation to a positive utterance may involve changes to other parts of the sentence has been given a typological study by Miestamo (2005). He introduces the concepts of symmetry and asymmetry: when the negation leaves the positive utterance unchanged, the negation is called 'symmetric' and if there is a change, the negation is 'asymmetric.' The specific form of asymmetry illustrated with the switch from *something* to *anything* has received a lot of attention from semanticists. It concerns 'polarity,' both the positive polarity of *something*, which prefers positive contexts, and the negative polarity of *anything*, which prefers either negative context or negative-like contexts, such as questions or conditionals (e.g., van der Wouden, 1997).

(28) Did you hear anything?
(29) If you hear anything, let me know.

The study of what unites negative and negative-like contexts has received various answers, prominent ones being those of Ladusaw (1980) and Giannakidou (1994).

Note that English allows two alternatives to (27c).

(27) (d) I heard *nothing*
(e) I did*n't* hear *nothing*.

In (26d) the indefinite itself is negative – a strategy that seems typical for Europe (Kahrel, 1996; Haspelmath, 1997) and the

Americas (Van Alsenoy and van der Auwera, 2014), and in (26e), which is substandard English, there is both a negative indefinite and a clausal negation. This is a kind of double negation, similar to the one illustrated with French (17), but different enough to deserve its own label, viz., that of 'negative concord.'

See also: Grammaticalization; Linguistic Presupposition; Psychiatric Assessment: Negative Symptoms.

Bibliography

van der Auwera, Johan, 2009. The Jespersen cycles. In: van Gelderen, Elly (Ed.), Cyclical Change. Benjamins, Amsterdam, pp. 35–71.

van der Auwera, Johan, Bhat, D.N.S., 1999. Havyaka kannada: negation and modality. Indian Linguistics 60, 1–26.

van der Auwera, Johan, Lejeune, Ludo, Goussev, Valentin, Pappuswamy, Umarani, 2005. The prohibitive. In: Haspelmath, Martin, Dryer, Matthew, Gil, David, Comrie, Bernard (Eds.), World Atlas of Language Structures. Oxford University Press, Oxford, pp. 290–293.

Bernini, Giuliano, Ramat, Paolo, 1996. Negative Sentences in the Languages of Europe. Mouton de Gruyter, Berlin.

Bybee, Joan L., 1985. Morphology. A Study of the Relation between Meaning and Form. Benjamins, Amsterdam.

Croft, William, 1991. The evolution of negation. Journal of Linguistics 27, 1–27.

Crowley, Terry, 2006. Nese: A Diminishing Speech Variety of Northwest Malakula (Vanuatu). Australian National University, Canberra.

Dahl, Östen, 1979. Typology of sentence negation. Linguistics 17, 79–106.

Derbyshire, Desmond C., 1979. Hixkaryana. North-Holland, Amsterdam.

De Swart, Henriëtte, 2010. Expression and Interpretation of Negation. An OT Typology. Springer, Dordrecht.

Devos, Maud, van der Auwera, Johan, 2013. Jespersen cycles in Bantu: double and triple negation. Journal of African Languages and Linguistics 34, 205–274.

Dryer, Matthew S., 1988. Universals of negative position. In: Hammond, Michael, Edith, Moravcsik, Wirth, Jessica (Eds.), Studies in Syntactic Typology. Benjamins, Amsterdam, pp. 93–124.

Early, Robert, 1994. A Grammar of Lewo, Vanuatu. Doctoral dissertation. Australian National University.

Forest, Robert, 1993. Négations: Essai de syntaxe et de typologie linguistique. Klincksieck, Paris.

Giannakidou, Anastasia, 1994. Polarity Sensitivity as (Non-)veridical Dependency. Benjamins, Amsterdam.

Greenberg, Joseph H. (Ed.), 1966. Universals of Grammar. MIT Press, Cambridge, Mass.

Haegeman, Liliane, 1995. The Syntax of Negation. Cambridge University Press, Cambridge.

Hansen, Maj-Britt Mosegaard, Visconti, Jacqueline, 2009. On the diachrony of "reinforced" negation in French and Italian. In: Rossari, Corinne, Ricci, Claudia, Spiridon, Adriana (Eds.), Grammaticalisation and Pragmatics: Facts, Approaches, Theoretical Issues. Emerald, Bingley, pp. 137–171.

Haspelmath, Martin, 1997. Indefinite Pronouns. Oxford University Press, Oxford.

Hollenbach, Barbara E., 1976. Tense-negation interplay in Copala Trique. International Journal of American Linguistics 42, 126–132.

Honda, Isao, 1996. Negation: A Cross-Linguistic Study. Doctoral dissertation. University of New York at Buffalo.

Horn, Laurence R., 1989. A Natural History of Negation. The University of Chicago Press, Chicago.

Horn, Laurence R. (Ed.), 2010. The Expression of Negation. De Gruyter Mouton, Berlin.

Jespersen, Otto, 1917. Negation in English and Other Languages. A.F. Host, Copenhagen.

Kahrel, Peter, 1996. Aspects of negation. Doctoral dissertation. University of Amsterdam.

Knüppel, Veronika, 2001. Die Syntax der Negation im Französischen: eine lexikalisch-funktionale Analyse. Niemeyer, Tübingen.

Ladusaw, William A., 1980. Polarity Sensitivity as Inherent Scope Relations. Garland, New York.

Landin, David J., 1984. An outline of the syntactic structure of Karitiâna sentences. In: Dooley, Robert A. (Ed.), Estudos sobre linguas tupi do Brasil. Summer Institute of Linguistics, Brasilia.

Levinson, Stephen C., 1983. Pragmatics. Cambridge University Press, Cambridge.

Li, Charles N., Thompson, Sandra A., 1981. Mandarin Chinese. A Functional Reference Grammar. University of California Press, Berkeley.

Londey, David, Johanson, Carmen, 1987. The Logic of Apuleius. Including a Complete Latin Text and English Translation of the Peri Hermeneias of Apuleius of Madaura. Brill, Leiden.

Meillet, Antoine, 1912. L'évolution des formes grammaticales. Scientia 12, 384–400 (Reprinted in Meillet, Antoine (1926). Linguistique historique et linguistique générale, 130–148. H. Champion, Paris.).

Miestamo, Matti, 2005. Standard Negation. The Negation of Declarative Verbal Main Clauses in a Typological Perspective. Mouton de Gruyter, Berlin.

Miestamo, Matti, 2010. Negatives without negators. In: Wohlgemuth, Jan, Cysouw, Michael (Eds.), Rethinking Universals: How Rarities Affect Linguistic Theory. de Gruyter Mouton, Berlin, pp. 169–194.

Nedyalkov, Igor, 1994. Evenki. In: Kahrel, Peter, Van den Berg, René (Eds.), Typological Studies in Negation. Benjamins, Amsterdam.

Payne, John R., 1985. Negation. In: Shopen, Timothy (Ed.), Language Typology and Syntactic Description, Clause Structure, vol. 1. Cambridge University Press, Cambridge, pp. 197–242.

Strawson, Peter F., 1950. On referring. Mind 69, 320–344.

Schwenter, Scott, 2006. Fine-tuning Jespersen's cycle. In: Birner, Betty, Ward, Gregory (Eds.), Drawing the Boundaries of Meaning: Neo-Gricean Studies in Pragmatics and Semantics in Honor of Laurence R. Horn. Benjamins, Amsterdam, pp. 327–344.

Thompson, Sandra Annear, 1998. A discourse explanation for the cross-linguistic differences in the grammar of interrogation and negation. In: Siewierska, Anna, Jung, Song. Jay (Eds.), Case, Typology and Grammar. In Honor of Barry J. Blake. Benjamins, Amsterdam, pp. 309–341.

Timberlake, Alan, 1993. Russian. In: Comrie, Bernard, Corbett, Greville C. (Eds.), The Slavonic Languages. Routledge, London, pp. 827–886.

Tobler, Alfred, 1882. Vermischte Beiträge zur französischen Grammatik. Erste Reihe. Hirzel, Leipzig.

Van Alsenoy, Lauren, van der Auwera, Johan, 2014. On the relation between double clausal negation and negative concord. In: Mosegaard-Hansen, Maj-Britt, Visconti, Jacqueline (Eds.), The Diachrony of Negation. Benjamins, Amsterdam, pp. 13–46.

Veselinova, Ljuba, 2010. Standard and special negators in the slavonic languages: synchrony and diachrony. In: Björn, Hansen, Grkovic-Major, Jasmina (Eds.), Diachronic Syntax of the Slavonic Languages. Wiener Slawistischen Almanach, Vienna, pp. 197–210.

Vossen, Frens, van der Auwera, Johan, 2014. The Jespersen cycles seen from Austronesian. In: Mosegaard-Hansen, Maj-Britt, Visconti, Jacqueline (Eds.), The Diachrony of Negation. Benjamins, Amsterdam, pp. 47–83.

van der Wouden, Ton, 1997. Negative Contexts: Collocation, Polarity and Multiple Negation. Routledge, London.

Negotiation and Bargaining: Organizational Aspects

VK Aggarwal, University of California, Berkeley, CA, USA
C Dupont, Graduate Institute of International Studies, Geneva, Switzerland

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 10473–10477, © 2001, Elsevier Ltd, with an updated Bibliography section supplied by the Editor.

Negotiation and bargaining is a process in which two or more parties seek a mutual agreement through an explicit or implicit exchange of views. The focus of this article is on international negotiation and bargaining with a special emphasis on the role of international institutions in facilitating cooperative bargains. In particular, the article discusses institutional design in connection with the nature of the problem that parties face, viewed through the lens of the provision of different types of goods.

The Diversity of Bargaining Situations

Reaching a mutual agreement may take very different roads depending on the characteristics of the bargaining situation. This section discusses the key distinction between positive verses zero-sum situations. It then relates this basic difference to a typology of goods that parties may seek to provide through a process of negotiation.

Pure Conflicts of Interest: Zero-Sum Situations

A widespread assumption is that bargaining is, foremost, a process in which parties seek to reconcile contradictory interests and values. The aim is to agree on a particular split of the differences between parties. Such situations have been labeled 'zero-sum situations' and are particularly salient when negotiators value outcomes in relative terms. In such cases, any gain by one party comes at the expense of a loss by another party. Bargaining is essentially a process of (re)distribution. In international politics, an archetype of a zero-sum situation is the division of a given territory (for instance, the carving-up of the Antarctic into different national sectors). In economics, the process of allocating market shares under nongrowing market assumptions is also a zero-sum situation.

Mixed-Motive Situations

Whereas distribution is often a key concern of bargaining, parties may also share common interests. Agreements can create value, as well as distribute it. Negotiation situations with both distributive and creative dimensions are labeled 'mixed-motive situations.' Figure 1 illustrates the two dimensions for a hypothetical bargaining case between negotiators Alpha and Beta aiming at establishing market shares.

Moving along the line [1, 1] corresponds to the case of pure conflict of interests mentioned in Section Pure Conflicts of Interest: Zero-Sum Situations. Negotiator Alpha and Beta fully control a steady market and have to mutually agree on a distribution of market shares. Consider, however, the case where Alpha and Beta are not the only market actors and thus will not be able to fully control the market unless they come to a mutual agreement. In Figure 1, point A is the outcome in case of no agreement, with Alpha and Beta joint market shares lower than one. Bargaining aims at reaching total control of the market, thus creating value collectively. The collective gain must be redistributed, along segment [B, C] in Figure 1. Another mixed situation arises when negotiators seek to expand the total size of the market, from 1 to $(1+x)$ in Figure 1. This does not eliminate the initial distributive aspect of bargaining but makes the potential conflict less severe.

In sum, in mixed-motive situations there may be a trade-off between individual and collective rationality. Mixed-motive situations dominate when negotiators not only care about bargaining outcomes in relative terms but also in absolute terms (Bazerman, et al., 2000).

Bargaining Situations and Types of Goods

Bargaining situations may be characterized by different combinations of positive and zero-sum features. In particular, it is useful to distinguish four types of objects or goods that negotiators may deal with, namely public goods, common pool resources (CPRs), inclusive club goods, or private goods (see Aggarwal, 1998; Cornes and Sandler, 1996; Snidal, 1979).

Differences among goods can be characterized along two dimensions: jointness, which refers to the extent to which goods are affected by consumption; and by the possibility of exclusion, which refers to whether noncontributors to the provision of the good can be kept from consuming it (see Table 1).

The two dimensions – jointness and exclusion – largely influence the mixture of interests in a bargaining process. In the case of public goods, there is an important value-creating potential – given that actors can consume the same amount of the good without affecting the consumption of others (as in

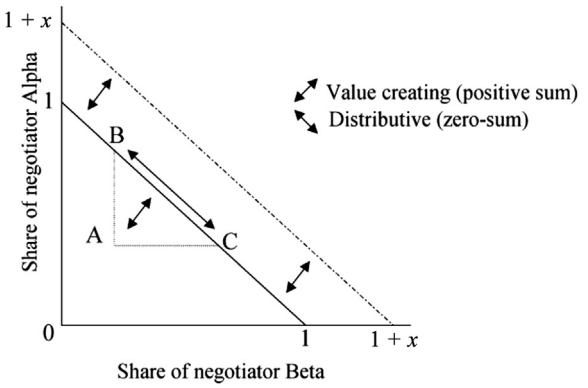

Figure 1 A mixed-motive bargaining situation.

Table 1 Type of goods

		Jointness? YES	Jointness? NO
Exclusion possible?	NO	Public	Common pool resources
	YES	Inclusive club	Private

the case of national defense). Yet, the possibility of cost sharing creates a distributive aspect to this good. Given the impossibility of excluding noncontributors, there is in fact a dual distributive problem – among those who pay for the good and between those who pay and those who do not.

Inclusive club goods refer to the case of goods that exhibit jointness (not diminished, by use), but where exclusion is possible (for instance satellite transmission of television). In this situation, bargaining reveals a strong value-creating feature and the problem of distributing costs is less problematic than in the case of public goods because of the possibility of excluding noncontributors from the consumption of the good. The availability of exclusionary mechanisms may help generate revenue, which further reduces the initial problem of cost distribution.

The positive potential of cooperation to collectively produce private goods tends to be complicated by the distributive problem created by the lack of jointness (for instance, the provision of a system of irrigation that delivers private amounts of water to different field owners). Yet, the availability of exclusionary mechanisms can alleviate the problem of distributing the costs. By contrast, in the case of common pool resource goods, bargaining is essentially a distributive process among those that are willing to pay for the good and between them and the noncontributors.

Bargaining Situations and the Role of Institutions

When and how do institutions affect patterns of negotiations? We begin with a definition of institutions, and then turn to an examination of the role they might play. The most readily evident role for institutions could be for situations of market failure, described nicely by the Prisoners' dilemma. In such cases, institutions may help to overcome suboptimality and ensure that actors reach a mutually superior outcome by cooperating. But as noted in Section Institutional Designs for Different Goods, other bargaining situations might also benefit from institutions. Section Bargaining to Create Institutions briefly examines some theories that help explain the bargaining process leading to the formation of institutions themselves.

Definitions: Metaregimes, Regimes, and International Institutions

International institutions have been defined in many ways, and considerable disagreement remains over the differences among international organizations, international institutions, and international regimes. International organizations are generally seen to be more formally specified institutions, with a secretariat, permanent office, and the like. International regimes have been defined as sets of principle, norms, rules, and decisions upon which actors' expectations converge (Krasner, 1983). It is useful to distinguish between principles and norms on the one hand, and rules and procedures on the other. This allows us to delineate two aspects of institutions: metaregimes and regimes (Aggarwal, 1998). Whereas metaregimes represent the principles and norms underlying international arrangements; international regimes refer specifically to rules and procedures. Regimes can be examined in terms of their characteristics: their strength, nature, and scope. Strength refers to the stringency of the multilateral rules that regulate national behavior; nature (in an economic context) refers to the degree of openness promoted by the accord; and scope refers to two aspects: (1) the number of issues incorporated in the regime, or issue scope; and (2) the number of actors involved (bilateral or multilateral), or institutional scope. To summarize, international institutions consist of metaregimes and regimes; if they are highly formalized, we can refer to them as international organizations.

Institutional Designs for Different Goods

Institutions affect the strategies available to actors and the payoffs linked to various bargaining outcomes. Systematic and specific links between bargaining situations and institutions have been explored through strategic analysis of actors' behavior (Aggarwal and Dupont, 1999; for earlier efforts, see Snidal, 1985; Stein, 1982). Aggarwal and Dupont (1999) model negotiation processes over the provision of diverse types of goods with game-theoretic structures that correspond to different assumptions about the costs and benefits of these goods and variation in actors' capabilities. They derive a variety of games (including the situations known as Prisoners' Dilemma, Chicken, Assurance, Harmony, Deadlock, or Battle of the Sexes; for a taxonomy of situations see Rapoport and Guyer, 1966) that can be solved using the Nash equilibrium concept (Nash, 1951). Issues of institutional design are then discussed in connection with the evaluation in terms of efficiency of the predicted outcome(s) of the various games. This analysis yields the following findings on the role of institutions in bargaining situations.

First, institutions may help secure positive outcomes in bargaining processes. When actors have large resources and when they face public and CPR goods, conflict may prevent the provision of such goods. Institutions can play the role of a third party in such cases and enforce a mutual agreement, making the parties more willing to settle on a bargaining outcome in the first place. To successfully overcome the tendency of players to defect, institutions should be strong and formalized. When actors have fewer resources, free riding is less of an option and the problem is one of developing mutual confidence in the ability of the other actors to implement a bargain. In such cases, institutions may provide insurance schemes against defection.

Second, institutions may reduce distributive problems. They may eliminate some sharply asymmetric outcomes and may provide focal point solutions for both cost sharing (which applies to all types of goods) and benefit splitting (when there

is no jointness of supply, that is, for private and CPR goods). Institutions with a firmly and widely established metaregime tend to perform these tasks well.

Third, institutions are not useful only for their help in the initial provision of various types of goods, but also may provide ongoing benefits. For instance, institutions may be useful even if there is some powerful political entrepreneur, or hegemon that provides public goods. Hegemons wish to coerce weaker actors to move away from free riding on the unilateral provision of goods and wish to have them carry some of the burden. They try to do so by linking the provision of the good with other issues. This can be done through institutions that connect different issues.

Fourth, institutions help clarify the mechanism of exclusion (Scott and Davis, 2007). Goods are rarely naturally given but must often be constructed. Actors in international relations use institutions to construct and enforce exclusionary mechanisms. Careful institutional design might, therefore, transform problematic goods such as public or CPR goods into less-problematic bargaining situations such as club goods.

Bargaining to Create Institutions

While institutions may help facilitate or alter the costs and benefits arising from the bargaining process, institutions themselves are the result of a bargaining process. We can consider three schools of thought that have been used to explain the formation of institutions (see Aggarwal, 1998 for references on these three schools of thought). First, from a neorealist power-based tradition, states in the international system compete for security in an anarchic international system. For analysts in this school, international institutions are simply reflections of the existing balance of power, and institutional outcomes have distributional consequences, that is, benefits from cooperation may be unequal.

The neoliberal tradition, on the other hand, emphasizes the incentives that states have to cooperate. In the neoliberal framework, states create 'rules of the game' (i.e., institutions) based on certain agreed-upon norms. Institutions foster sustained cooperation as well as lower information and transaction costs in a particular issue-area. International institutions are seen to be 'sticky' – that is, existing institutions constrain future institutional change – and new institutions must be designed to be either 'nested' within, parallel, or independent from existing institutions (Aggarwal, 1998). Nested institutions are hierarchically ordered. Parallel institutions refer to an institutional division of labor, while independent institutions are simply unconnected since their mandates do not overlap.

The cognitive perspective stresses the role of expert consensus and the interplay of experts and politicians in the formation of institutions. Those within the cognitive tradition are concerned with how the development of knowledge can foster learning on a particular issue area that may drive policy innovation (Hodgkinson, 2003).

Future Research Directions

The issue of institutional design, both at the international and domestic level, will remain a key research and policy challenge for the future. Analysts are increasingly recognizing the need to precisely specify the types of bargaining situations for which institutions might be needed. Rather than debating if institutions are necessary or useless, future work will provide us with a more precise understanding of the nature of goods and their relationship to institutions. Together with insight into the factors that contribute to the creation of institutions themselves, we will achieve a better understanding of the conditions under which institutions facilitate and contribute to the bargaining process.

See also: Conflict and Consensus; Conventions and Norms: Philosophical Aspects; Dispute Resolutions in Economics; Game Theory: Noncooperative Games; Game Theory; International Arbitration; International Communication: History; Negotiation and Bargaining: Role of Lawyers; Prisoner's Dilemma, One Shot and Iterated.

Bibliography

Aggarwal, V.K., 1998. Reconciling multiple institutions: bargaining, linkages, and nesting. In: Aggarwal, V.K. (Ed.), Institutional Designs for a Complex World: Bargaining, Linkages, and Nesting. Cornell University Press, Ithaca, NY, pp. 1–3.

Aggarwal, V.K., Dupont, C., 1999. Goods, games, and institutions. International Political Science Review 20, 393–409.

Bazerman, M.H., Curhan, J.R., Moore, D.A., Valley, K.L., 2000. Negotiation. Annual Review of Psychology 51, 279–314.

Cornes, R., Sandler, T., 1996. The Theory of Externalities, Public Goods, and Club Goods. Cambridge University Press, New York.

Hodgkinson, G.P., 2003. The interface of cognitive and industrial, work and organizational psychology. Journal of Occupational and Organizational Psychology 76 (1), 1–24.

Kaul, I., Grinnberg, I., Stern, M.A. (Eds.), 1999. Global Public Goods. Oxford University Press, New York.

Krasner, S. (Ed.), 1983. International Regimes. Cornell University Press, Ithaca, NY.

Nash, J., 1951. Non-cooperative games. Annals of Mathematics Journal 54, 286–295.

Rapoport, A., Guyer, M., 1966. A taxonomy of 2×2 games. General Systems 203–214.

Scott, W.R., Davis, G.F., 2007. Organizations and organizing: rational, natural and open system perspectives. Pearson Education, Upper Saddle River.

Snidal, D., 1979. Public goods, property rights, and political organization. International Studies Quarterly 22, 532–566.

Snidal, D., 1985. Coordination vs prisoner's dilemma: implications for international cooperation. American Political Science Review 79, 923–942.

Stein, A.A., 1982. Coordination and collaboration: regimes in an anarchic world. International Organization 36, 294–324.

Negotiation and Bargaining: Role of Lawyers

Chris Guthrie, Vanderbilt Law School, Nashville, TN, USA
Jennifer K Robbennolt, University of Illinois College of Law, Champaign, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C. Guthrie, volume 15, pp. 10477–10481, © 2001, Elsevier Ltd.

Abstract

Negotiation is a communication process that lawyers use to resolve disputes and to complete transactions on behalf of their clients. Two models are likely to guide lawyers' negotiations: a distributive model, which posits that negotiation is primarily a zero-sum game; and a problem-solving model, which posits that negotiation is a collaborative search for joint gains. Regardless of the negotiation model(s) employed, lawyer-negotiators find that predictable strategic, psychological, and structural barriers often arise to prevent them from obtaining optimal negotiation outcomes for their clients.

Negotiation is an interactive communication process that occurs whenever we want something from someone else or another person wants something from us (Shell, 1999). Whether functioning as litigators handling civil disputes; as transactional or corporate lawyers putting together deals; as policymakers or advisors proposing and revising rules and regulations; or as prosecutors or defense counsel litigating criminal cases, lawyers negotiate.

Negotiation Models

Scholars have identified two basic models of, or approaches to, negotiation. The first, and perhaps most familiar, is variously labeled the 'adversarial,' 'distributive,' 'positional,' or 'competitive' model of negotiation (Gifford, 1985; Lewicki et al., 1999). This distributive model of negotiation posits that negotiation is a zero-sum game in which any gain one party receives necessarily occurs at the expense of the other party. Because it views negotiation as a zero-sum game, this model counsels negotiators to attempt to claim as many of those gains as possible through the use of such tactics as extreme opening positions, few (and small) concessions, commitments to positions, and the withholding of relevant information (Gifford, 1985). The distributive model, in short, places a primacy on claiming value in negotiation.

The downside to the distributive model, at least according to its critics, is that it is often harmful to the relationship between the parties and can easily lead to impasse or to negotiated agreements that leave value on the table. To address these shortcomings, scholars have developed a second approach to negotiation, currently fashionable in American law schools (as well as in business, public policy, and international relations schools), called the 'problem-solving,' 'integrative,' 'interest-based,' or 'principled' model of negotiation (Fisher and Ury, 1991; Menkel-Meadow, 1984; Walton and McKersie, 1965). The problem-solving model of negotiation posits that negotiation is not necessarily a zero-sum game. Rather, most negotiations are (or at least have the potential to become) collaborative problem-solving exercises in which the parties work together to devise creative, 'win–win' outcomes. Given this view of negotiation, the problem-solving model counsels negotiators to focus on the parties' underlying interests in negotiation, to devise creative options that help satisfy both parties' interests, and to develop and use legitimate, objective criteria that both parties view as applicable to the negotiation (Fisher and Ury, 1991). The problem-solving model, in short, emphasizes both creating and claiming value in negotiation.

Neither of the primary negotiation models is sufficient to capture the complexity of legal negotiations as an empirical matter, and in all likelihood neither is sufficient as a prescriptive guide, either. Lawyer-negotiators no doubt do, and probably should, borrow liberally from either or both negotiation model(s) depending upon the nature of the negotiation, relationship between the clients, practice area, norms in the community, and relevant legal standards and rules (see also Korobkin, 2008).

Available empirical evidence suggests, however, that lawyer-negotiators tend to rely somewhat more heavily on one model or the other when negotiating. According to one empirical study, for instance, lawyer-negotiators evaluating their adversaries rated 65% of them as cooperative or problem-solving negotiators, 24% as competitive or adversarial negotiators, and only 11% as incapable of categorization (Williams, 1983). A more recent study obtained similar results, with 64% of lawyer-negotiators classified as problem-solving negotiators and 36% classified as adversarial (Schneider, 2002). An empirical study of American civil litigators found that most express a preference for the problem-solving model of negotiation but tend to behave more consistently with a more distributive or adversarial model of negotiation (Heumann and Hyman, 1997).

Negotiation in the Legal System

Lawyers perform many roles in America and around the world. Broadly speaking, though, lawyers' work typically falls into one of two categories: deal-making and dispute resolution. Negotiation is central to both.

Deal-Making

Many lawyers function as deal-makers. Corporate lawyers, mergers and acquisitions specialists, and securities lawyers,

among others, are deal-makers. Deal-making lawyers are 'transaction cost engineers' who use differences between the parties (e.g., one party wants to buy a good, while one wants to sell), noncompetitive similarities (e.g., licensee and licensor, both of whom want a technology to succeed), economies of scale (e.g., pooling production or consumption facilities to lower costs of a good), and the like to create value for their clients (Gilson, 1984; Gilson and Mnookin, 1995). Deal-making lawyers also strive to claim value for their clients by negotiating advantageous divisions of the value they have helped to create.

Dispute Resolution

Other lawyers function as dispute-resolvers. Civil litigators and criminal defense attorneys, for instance, are dispute-resolvers. Dispute-resolvers spend much of their professional lives helping their clients settle disputes through negotiation.

Most disputes do not require lawyers, however, because most disputants either choose to 'lump it' or to settle their disputes on their own through negotiation (Miller and Sarat, 1980–81). Once disputants have invoked the formal legal system by hiring a lawyer and filing suit, their negotiation is no longer shaped merely by informal norms but also by the law and legal system. The disputants and their lawyers negotiate, in other words, 'in the shadow of the law,' because the consequence of failing to reach a negotiated agreement is possible imposition of a binding decision by a third party, that is, a trial (Mnookin and Kornhauser, 1979).

Trials, however, are quite rare in the American civil justice system (and their occurrence is declining, Galanter, 2004), because most disputants settle their cases through negotiation. The conventional explanation of this phenomenon is economic. According to the economic model of suit and settlement, disputants and their lawyers seek to maximize value in litigation. To do so, they calculate the expected value of trial (i.e., the anticipated judgment times, and the probability it will occur, less trial costs), compare it to any settlement offer on the table, and select whichever promises to give them more value. Because the costs of trial are so much greater than the costs of settlement, most disputants opt to settle their cases through negotiation because settlement generally promises more value than trial. Trial, according to this economic account of litigation, results only when the disputants and their lawyers develop widely divergent predictions about the likely outcome at trial (Priest and Klein, 1984). As described below, however, settlements may break down for other reasons as well.

Barriers to Negotiation

Whether functioning as deal-makers or dispute-resolvers, lawyer-negotiators are likely to find that predictable barriers or impediments arise during the negotiation process to prevent them from obtaining optimal negotiation outcomes for their clients. These barriers or impediments fall roughly into three categories: strategic, psychological, and structural (Arrow et al., 1995; Mnookin, 1993).

Strategic Barriers

Strategic barriers arise where lawyer-negotiators adopt strategies and tactics, like hard-bargaining techniques, that backfire. Strategic barriers are likely to arise because lawyer-negotiators and their clients find themselves trapped in the so-called negotiator's dilemma (Lax and Sebenius, 1986). The negotiator's dilemma, like the prisoner's dilemma upon which it is based, focuses on the tension between distributive or value-claiming moves on the one hand and problem-solving or value-creating moves on the other.

Every negotiation involves something of value, which the parties can simply divide in some way (i.e., value claiming) or which they can attempt to expand and divide (i.e., value creating and claiming). Negotiators who make value-claiming moves may succeed in obtaining relatively more of the existing value, but their competitive moves are likely to inhibit the problem-solving needed to create additional value for the parties. The negotiator's dilemma refers to negotiator's tendency to focus narrowly on claiming value at the expense of creating additional value for the parties to divide.

The negotiator's dilemma is inherent in negotiation, but it is particularly likely to confound lawyer-negotiators and their clients under certain conditions. For example, because value creation depends upon common information, the negotiator's dilemma is more likely to impede successful negotiation when the parties face informational asymmetries (Gilson and Mnookin, 1995). When information is asymmetric, parties must agree to share their private information to create value, but parties may be reluctant to do so for fear that their adversaries will exploit the information they have divulged. This is particularly likely to be a problem in litigation.

Psychological Barriers

Psychological barriers arise due to predictable cognitive and motivational biases that influence the way lawyer-negotiators and their clients make predictions, interpret information, assess their adversaries, and weigh their options. Of the many psychological barriers that may arise, some of the most significant are self-serving and confirmation biases, overconfidence, reactive devaluation, the fixed-pie bias, loss aversion, and anchoring.

Lawyer-negotiators and their clients are likely to fall prey to self-serving and confirmation biases when making judgments about ensuing negotiations because people generally have difficulty interpreting information in a neutral fashion. To illustrate, researchers asked experimental participants to assess damages in a tort case arising from an automobile-motorcycle collision in which the plaintiff had sued the defendant for $100 000. The researchers assigned some participants to play the role of plaintiff and others the role of defendant, but they provided both sets of participants with identical information about the case. Despite reviewing identical information, participants interpreted it in self-serving ways. Participants evaluating the case from the perspective of the plaintiff predicted that the judge would award $14 527 more than the defendant-participants predicted, and identified a fair settlement value for the case that was $17 709 higher than the value selected by defendant-participants (Loewenstein et al., 1993).

These self-serving assessments are 'an important determinant of bargaining impasse' (Babcock and Loewenstein, 1997).

In addition to self-serving bias, lawyer-negotiators and their clients are also susceptible to overconfidence. In a recent study, attorneys were asked to describe a case they were handling, to describe their minimum goals for the representation, and to indicate how confident they were that they would meet or exceed those goals. When compared with the outcome of the cases once they were resolved, attorneys were overconfident in their estimates. They did not, however, seem to realize it – more attorneys reported meeting their original goals than actually met them (Goodman-Delahunty et al., 2010).

Not only are lawyer-negotiators and their clients likely to interpret their own positions in self-serving and overconfident ways, they are also likely to discount or devalue their adversaries' positions due to a phenomenon psychologists have labeled 'reactive devaluation' (Ross and Stillinger, 1991). Proposals offered by an opponent often seem less attractive to a negotiator than if the same proposal had been offered by another source (Ross and Stillinger, 1991; Korobkin and Guthrie, 1994b). Parties may, therefore, decline proposals that might otherwise be agreeable.

Reactive devaluation may stem, in part, from the fixed-pie bias – the assumption that the interests of the two sides are inevitably incompatible. Parties who believe the pie is fixed are likely to emphasize value claiming over value creation, and are therefore likely to miss opportunities to expand the pie (Thompson and Hrebec, 1996).

Lawyer-negotiators and their clients can also be influenced by 'anchors' – numbers that are available in the negotiator's mind, that provide a convenient starting place for judgment, and that influence outcomes, but that may not be relevant to the decision at hand (Tversky and Kahneman, 1982). Legal negotiators have been shown to be influenced by media reports of case outcomes, information about settlements in other cases, initial offers or demands, damage caps, and their own articulated aspirations (Babcock and Pogarsky, 1999; Depoorter, 2010; Korobkin, 2002; Korobkin and Guthrie, 1994; Robbennolt and Studebaker, 2003).

Finally, loss aversion is a significant psychological barrier that can arise during the negotiation process. Psychologists have found that negotiators do not experience equivalent gains and losses the same way. Rather, losses loom larger, 'generally twice as large' as corresponding gains (Kahneman et al., 1991; Kahneman and Tversky, 1984). Loss aversion may manifest itself in negotiation in a couple of different ways.

First, loss aversion is likely to affect negotiators' valuations of goods in transactional negotiations. Sellers are likely to place a higher value on goods than buyers because sellers experience the transfer of the good as a loss while buyers experience the transfer of the good as a gain. In an often-replicated study, researchers divided experimental subjects into a sellers group and a buyers group. The researchers gave members of the sellers group a coffee mug and the members of the buyers group $6.00 in cash. The researchers asked the sellers to indicate the minimum amount they would accept to sell the coffee mug, asked the buyers to indicate the maximum amount they would pay for a mug, and informed all of the subjects that they would execute trades based on the responses given. The researchers found that the two groups valued the mugs quite differently. Sellers, who would experience a loss if they parted with their mug, assigned a median price of $5.25, while buyers, who would experience the attainment of the mug as a gain, assigned a median price of $2.25 to $2.75 (Kahneman et al., 1991). In short, sellers tend to value goods much more highly than buyers, which may prevent lawyer-negotiators from obtaining optimal outcomes for their clients in transactional negotiations.

Second, loss aversion may influence the way negotiators perceive concessions. Negotiators view concessions they make as losses but concessions they receive from the other side as gains. Because losses are weighted more heavily than gains, negotiators are likely to view the concessions they make as more significant than the equivalent concessions their adversaries make. This 'concession aversion' may also undermine successful negotiation (Kahneman and Tversky, 1995).

Each of these psychological barriers can create a serious problem, particularly when the parties distrust one another, as is often the case in negotiation. Indeed, simply being in an adversary posture can distort negotiation judgments. For example, people tend to attribute disagreement to bias on the part of their counterpart – rather than to simple disagreement. Such attributions can incline parties to negotiate more competitively (Kennedy and Pronin, 2008).

Structural Barriers

The third type of barrier that may arise in negotiation is structural. Structural barriers are a product not of individual negotiators' calculations or cognitive processes but rather of the way in which conflict and conflict resolution are managed. The most significant structural impediment to successful negotiation faced by lawyer-negotiators and their clients arises out of the lawyer–client relationship.

The lawyer–client relationship is a classic principal–agent relationship (Pratt and Zeckhauser, 1985), in which the client is the principal who possesses substantive decision-making authority, while the lawyer is an agent who is ethically obligated to use legal means to carry out the client's objectives. The structure of this relationship can facilitate settlement as well as pose barriers to agreement. Lawyers, as third parties to a dispute or deal, can sometimes help clients avoid psychological barriers to agreement (Gilson and Mnookin, 1994; Korobkin and Guthrie, 1997). But, as with any principal–agent relationship, incentives are seldom perfectly aligned between principal and agent. This misalignment of incentives can have a detrimental impact on negotiation.

To give one concrete illustration, consider a typical personal injury lawsuit in the United States. In such a case, the plaintiff is usually a one-shot player with limited resources who hires a lawyer on a contingency fee basis, while the defendant is often a repeat-player organization of some type that hires a lawyer on an hourly fee basis (Galanter, 1974; Gross and Syverud, 1996; Kritzer, 2002). Principal–agent problems may arise for both plaintiffs and defendants.

The plaintiff's lawyer who takes the case on a contingency basis recovers a fee only if the plaintiff succeeds in obtaining a settlement or judgment. If the plaintiff is successful, the plaintiff's lawyer will receive a percentage of the recovery. Because of this financing scheme, the plaintiff's lawyer may have an incentive to settle quickly so as to maximize his hourly

return, but the plaintiff may fare better if the plaintiff's lawyer settles later or takes the case to trial. Suppose, for instance, that a plaintiff's lawyer on a 25% contingency settles a case for $10 000 after 10 h of work. In this instance, the plaintiff receives a net recovery of $7500, and the plaintiff's attorney receives $2500 for his efforts. On an hourly basis, the plaintiff's attorney earns a $250 per hour wage. Now suppose instead that the plaintiff's lawyer settles the case for $16 000 after conducting 40 h of work. In this instance, the plaintiff does better, obtaining a net recovery of $12 000. By contrast, the plaintiff's attorney does worse, earning only $100 per hour for his efforts (i.e., $4000 for 40 h of work). In short, the plaintiff's attorney would do better to settle earlier, while the plaintiff would do better to settle later.

Defense counsel working on an hourly fee basis faces a very different incentive structure than the plaintiff's lawyer. Defense counsel's hourly wage is constant regardless of the number of hours worked. Thus, defense counsel has a financial incentive to work the case as long as possible because that will maximize earnings. The client, in contrast, is likely to fare better if the case settles earlier. Suppose, for instance, that defense counsel charging $200 per hour settles a case for $10 000 after working 10 h on the case. In this instance, settlement costs the defendant $12 000 (i.e., $10 000 to plaintiff and $2000 to defense counsel) and earns defense counsel $2000. Now suppose, instead, that defense counsel settles for $6000 after spending 40 h on the case. In this instance, settlement costs the defendant more (i.e., $6000 to plaintiff and $8000 to defense counsel) but earns defense counsel more (i.e., $8000 vs $2000). In short, the defense counsel would do better to settle later, while the defendant would do better settling earlier.

Other considerations also play a role in how attorneys advise clients about when and how to settle cases – considerations including concern for reputation and referrals, the demands and promise of other cases, and many more (Kritzer, 2002). Legal negotiators must constantly weigh all of these competing considerations (see Robbennolt et al., 2010).

The Future of Negotiation Research

Because of its importance in law, business, foreign affairs, and interpersonal relationships, negotiation will continue to attract the attention of scholars and practitioners trying to shed light on the process. Three research trends are likely. First, scholars will propose additional models of negotiation that combine or challenge the existing distributive and problem-solving models of negotiation. Second, scholars will explore opportunities that new communications technologies create (as well as problems they cause) for negotiation (see, for example, Thompson and Nadler, 2002). Finally, scholars will continue to conduct empirical and experimental investigations into barriers to optimal negotiation outcomes. Ideally, scholars will not only identify additional barriers but will develop some appreciation of the magnitude of these barriers, learn more about how the barriers might interact with one another, and devise prescriptions for overcoming the barriers. Continued investigation of the negotiation process is important because negotiation will continue to be the process by which most disputes are resolved and most deals get done.

See also: Contracts: Legal Perspectives; Dispute Resolutions in Economics; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Negotiation and Bargaining: Organizational Aspects; Plea Bargaining: A Comparative Analysis.

Bibliography

Arrow, K.J., Mnookin, R.H., Ross, L., Tversky, A., Wilson, R. (Eds.), 1995. Barriers to Conflict Resolution. W.W. Norton, New York.
Babcock, L., Loewenstein, G., 1997. Explaining bargaining impasse: the role of self-serving biases. Journal of Economic Perspectives 11, 109–119.
Babcock, L., Pogarsky, G., 1999. Damage caps and settlement: a behavioral approach. Journal of Legal Studies 28, 341–370.
Depoorter, B., 2010. Law in the shadow of bargaining: the feedback effects of civil settlements. Cornell Law Review 95, 957–987.
Fisher, R., Ury, W. (Eds.), 1991. Getting to Yes: Negotiating Agreement without Giving In, second ed. Houghton-Mifflin, Boston.
Galanter, M., 1974. Why the 'Haves' come out ahead: speculations on the limits of legal change. Law & Society Review 9, 95–160.
Galanter, M., 2004. The vanishing trial: an examination of trials and related matters in federal and state courts. Journal of Emprical Legal Studies 1, 459–570.
Gifford, D.G., 1985. A context-based theory of strategy selection in legal negotiation. Ohio State Law Journal 46, 41–94.
Gilson, R.J., 1984. Value creation by business lawyers: legal skills and asset pricing. Yale Law Journal 94, 239–313.
Gilson, R.J., Mnookin, R.H., 1994. Disputing through agents: cooperation and conflict between lawyers in litigation. Columbia Law Review 94, 509–566.
Gilson, R.J., Mnookin, R.H., 1995. Foreword: business lawyers and value creation for clients. Oregon Law Review 74, 1–14.
Goodman-Delahunty, J., Granhag, P.A., Hartwig, M., Loftus, E.F., 2010. Insightful or wishful: lawyers' ability to predict case outcomes. Psychology, Public Policy, & Law 16, 133–157.
Gross, S.R., Syverud, K.D., 1996. Don't try: civil jury verdicts in a system geared to settlement. U.C.L.A. Law Review 44, 1–64.
Heumann, M., Hyman, J.M., 1997. Negotiation methods and litigation settlement methods in New Jersey: 'You can't always get what you want'. Ohio State Journal on Dispute Resolution 12, 253–310.
Kahneman, D., Knetsch, J.L., Thaler, R.H., 1991. Anomalies—The endowment effect, loss aversion, and status quo bias. Journal of Economic Perspectives 5, 193–206.
Kahneman, D., Tversky, A., 1984. Choices, values, and frames. American Psychologist 39, 341–350.
Kahneman, D., Tversky, A., 1995. Conflict resolution: a cognitive perspective. In: Arrow, K.J., Mnookin, R.H., Ross, L., Tversky, A., Wilson, R. (Eds.), Barriers to Conflict Resolution. W.W. Norton, New York, pp. 44–60.
Kennedy, K.A., Pronin, E., 2008. When disagreement gets ugly: perceptions of bias and the escalation of conflict. Personality & Social Psychology Bulletin 34, 833–848.
Korobkin, R., 2002. Aspirations and settlement. Cornell Law Review 8, 1–61.
Korobkin, R., 2008. Against integrative bargaining. Case Western Reserve Law Review 58, 1323–1342.
Korobkin, R., Guthrie, C., 1994. Psychological barriers to litigation settlement: an experimental approach. Michigan Law Review 93, 107–192.
Korobkin, R., Guthrie, C., 1997. Psychology, economics, and settlement: a new look at the role of the lawyer. Texas Law Review 76, 77–141.
Kritzer, H., 2002. Lawyer fees and lawyer behavior in litigation: what does the empirical literature really say? Texas Law Review 80, 1943–1983.
Lax, D.A., Sebenius, J.K., 1986. The Manager as Negotiator: Bargaining for Cooperation and Competitive Gain. The Free Press, New York.
Lewicki, R.J., Litterer, J.A., Minton, J.W., Saunders, D.M., 1999. Negotiation, second ed. Irwin, Burr Ridge, IL.
Loewenstein, G., Issacharoff, S., Camerer, C., Babcock, L., 1993. Self-serving assessments of fairness and pretrial bargaining. Journal of Legal Studies 22, 135–159.
Menkel-Meadow, C., 1984. Toward another view of legal negotiation: the structure of problem solving. U.C.L.A. Law Review 31, 754–842.
Miller, R.E., Sarat, A., 1981. Grievances, claims, and disputes: assessing the adversary culture. Law and Society Review 15, 525–566.
Mnookin, R.H., 1993. Why negotiations fail: an exploration of barriers to the resolution of conflict. Ohio State Journal on Dispute Resolution 8, 235–249.

Mnookin, R.H., Kornhauser, L., 1979. Bargaining in the shadow of the law: the case of divorce. Yale Law Journal 88, 950–997.
Pratt, J.W., Zeckhauser, R. (Eds.), 1985. Principals and Agents: The Structure of Business. Harvard Business School Press, Boston, MA.
Priest, G.L., Klein, B., 1984. The selection of disputes for litigation. Journal of Legal Studies 13, 1–55.
Robbennolt, J.K., MacCoun, R.J., Darley, J.M., 2010. Multiple constraint satisfaction in judging. In: Klein, D., Mitchell, G. (Eds.), The Psychology of Judicial Decision Making. Oxford University Press, pp. 27–39.
Robbennolt, J.K., Studebaker, C.A., 2003. News media reporting on civil litigation and its influence on civil justice decision making. Law & Human Behavior 27, 5–27.
Ross, L., Stillinger, C., 1991. Barriers to conflict resolution. Negotiation Journal 7, 389–404.
Schneider, A.K., 2002. Shattering negotiation myths: empirical evidence on the effectiveness of negotiation style. Harvard Negotiation Law Review 7, 143–233.
Shell, G.R., 1999. Bargaining for Advantage: Negotiation Strategies for Reasonable People. Viking, New York.
Thompson, L., Hrebec, D., 1996. Lose-lose agreements in interdependent decision making. Psychological Bulletin 120, 396–409.
Thompson, L., Nadler, J., 2002. Negotiating via information technology: theory and application. Journal of Social Issues 58, 109–124.
Tversky, A., Kahneman, D., 1982. Judgment under uncertainty: heuristics and biases. In: Kahneman, D., Slovic, P., Tversky, A. (Eds.), Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, pp. 3–22.
Walton, R.E., McKersie, R.B., 1965. A Behavioral Theory of Labor Negotiations: An Analysis of a Social Interaction System. McGraw-Hill, New York, pp. 126–183.
Williams, G.R., 1983. Legal Negotiation and Settlement. West Publishing, St. Paul, MN.

Negotiation and Conflict, The Psychology of

Mara Olekalns, The University of Melbourne, Carlton, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although conflict is associated with higher levels of emotion than negotiation, the two processes share many similarities. Both disputants and negotiators can opt for a more cooperative or competitive approach to resolving their differences. The approach that they choose and its consequences are affected by the social context within which conflicts and negotiations take place, the social processes that emerge as conflicts and negotiations unfold, and individuals' thoughts and feelings. This article explores how variables such as power, gender, culture, communication, and emotion influence the reconciliation of conflicting interests.

Negotiation and conflict are linked by the strategic options available to individuals. In both domains, two-dimensional typologies set out options based on the relative importance that individuals place on protecting their interests and outcomes *or* preserving their relationships with the other party. Typologies such as Pruitt's Dual Concern Model (Pruitt, 1981) and Rahim's Conflict Styles (Rahim, 1983) identify four strategic options: competition (protects outcomes, damages relationship); accommodation (erodes outcomes, preserves relationship); avoiding (erodes both outcomes and relationship); and problem solving (preserves or strengthens both outcomes and relationships). Of these, problem solving offers the most constructive path forward in both negotiations and conflicts. However, finding and following this path is difficult for negotiators and disputants alike. Consequently, negotiation and conflict researchers have focused on identifying the conditions that facilitate or impede a problem-solving approach. In both domains, researchers have explored the role of the social context, cognition, and emotion in influencing the strategies that individuals favor. Increasingly, researchers are also interested in how negotiators and disputants can turn a dysfunctional competitive process into a more constructive problem-solving one, and in the conditions that support this change.

The Psychology of Negotiation

Negotiators strive to balance two conflicting goals, the goal of maximizing individual outcomes and the goal of maximizing joint outcomes. The maximization of individual outcomes requires negotiators to claim value, that is, to persuade the other party to make favorable concessions and to yield to their demands. However, if negotiators are too unrelenting in their demands, they risk triggering escalatory cycles of argumentation and deadlock. In contrast, the maximization of joint outcomes requires that negotiators create value by sharing information, revealing their own interests and uncovering those of their counterparts. While value creation offers the benefit of increasing resources, the openness that underpins value creation also exposes negotiators to the risk of exploitation by their counterparts. Nonetheless, without creating value, negotiators are unlikely to reach a mutually acceptable agreement. Recognizing the tension between these processes, negotiation researchers have focused on identifying the factors that lead negotiators to emphasize value creation over value claiming, or vice versa. Four broad themes have emerged from this research: social context, social processes, social perception, and cognition and emotion.

Social Context

The social context within which negotiations take place shapes the way that individuals approach their negotiation. Social context is defined by factors considered to be 'outside' of the negotiation, that is, factors other than the issues to be negotiated. A unifying theme in this research is the representation of the underlying negotiation relationship, which may emphasize either independence or interdependence. Negotiators are less likely to believe that their fate (and outcome) is intertwined with that of an opponent when the social context emphasizes independence than when it emphasizes interdependence. As a result, negotiators with an independent perspective are more likely to favor value claiming than those with an interdependent perspective. Conversely, negotiators with an interdependent perspective are more likely to favor value creating than those with an independent perspective (Gelfand et al., 2006). Among the variables that can push negotiators to a more independent or interdependent representation of their relationship with an opponent, four have dominated negotiation research: social motives, power, gender, and culture.

Social Motives

Social motives capture the goals that negotiators have at the outset of a negotiation. These goals can be either *individualistic* and emphasize the maximization of individual outcomes or *cooperative* and emphasize the maximization of both negotiators' outcomes. One of the most enduring findings in negotiation research is that an individualistic motive encourages competition whereas a cooperative motive encourages problem solving (De Dreu et al., 2000). Negotiators with an individualistic orientation favor strategies such as argumentation, threats, and demands and consequently claim a greater share of resources for themselves. Negotiators with a cooperative orientation favor strategies such as information exchange, relationship building, and concession exchange, consequently creating greater value than individualistic negotiators. Differences in social motives maybe the result of explicit instructions

given to negotiators or may occur 'naturally' as the result of individual differences.

Power

Power resides both with the individual negotiator and with the negotiating dyad. Individuals may start negotiations with either high or low personal power, typically defined as the number of alternative options available to negotiators (Kim et al., 2005). Negotiators who have many available alternatives also have high power because of the greater ease with which they can exit an unsatisfactory negotiation. As a result, they are less concerned than low-power negotiators about preserving their relationship with an opponent. Tactically, this lack of concern is evident in their willingness to make more extreme offers and to give fewer concessions than low-power negotiators. As a result, high-power negotiators typically claim a greater share of resources than do low-power negotiators.

Power cannot, however, be fully understood at the individual level. It is also necessary to consider how power is distributed between two negotiators because power distribution determines negotiators' level of dependence. When power is evenly distributed (both negotiators have many or few alternatives), they are mutually dependent. Under conditions of mutual dependence the advantages of high power and the disadvantages of low power are reduced: having many alternative options is an advantage only if the other party has few and having few alternative options is a disadvantage only if the other party has many. When power is asymmetrically distributed (one negotiator has many and the other has few alternatives), they are nonmutually dependent. One consequence is that contention rapidly escalates: both the high- and low-power negotiators use more threats, personal attacks, and persuasive arguments. A further consequence is that negotiators are more likely to create value when negotiators are mutually dependent than when they are nonmutually dependent.

Gender

Early research found gender differences in negotiation to be somewhat elusive. Although women and men obtained markedly different outcomes in their negotiations, meta-analyses showed that only a small proportion of the variance in outcomes could be attributed to differences in the level of competitiveness that they displayed (Walters et al., 1998). More recent research shows that women's poorer negotiations are the result of both how women approach negotiations and the responses that they elicit from their opponents. A fundamental difference is women's greater reluctance to initiate negotiations. The failure to 'ask' means that women deprive themselves of the opportunity to improve their outcomes or meet their needs. Even when they do ask, they typically negotiate in a more narrow domain than men, seeing less of the world as negotiable. Those women who do initiate negotiations underestimate their worth: they set lower targets; are more likely to accept the first offer that they receive; and are less likely to walk away from an offer that does not meet their needs. These actions on the part of women are complemented by those of their opponents, who make lower offers to women based on the assumption that these offers will be accepted. Each of these actions erodes women's economic outcomes. Recent research shows that women obtain better outcomes when they believe that problem solving is linked to effective negotiation, and when there is transparency about what is negotiable.

The negotiation landscape is further complicated because women need to be mindful of both their economic and social outcomes. While women who initiate negotiations can improve their economic outcomes, this improvement carries a social cost. Irrespective of how they approach a negotiation, the initiation of a negotiation in its own right results in women being perceived as more pushy, less likable, and less desirable work colleagues. Women can offset these social costs by negotiating on behalf of others, by attributing the initiation of a negotiation to a mentor, or by harnessing female gender stereotypes (Kray and Thompson, 2005; Kulik and Olekalns, 2012).

Culture

Two dimensions dominate research linking cultural differences to negotiation tactics and outcomes, communication style, and individualism–collectivism (Brett, 2007). The first of these dimensions reflects cultural preferences for direct (low-context) vs indirect (high-context) communication; the second of these dimensions differentiates cultures based on the emphasis placed on individuals or groups. Drawing on these dimensions, researchers have shown that low-context negotiators favor giving and seeking explicit information about their underlying priorities and preferences, and that high-context negotiators are more likely to learn about each others' priorities and preferences indirectly through the exchange of multiple offers. These strategies work equally well to create value in intracultural negotiations but these different communication preferences impede value creation in intercultural dyads.

The structure of communication is shaped by communication styles. Negotiators are more likely to reciprocate emotional appeals in the opening moments of a negotiation when they are both from a low-context culture, than when they are from a high-context culture or when they have different communication styles. They are more likely to reciprocate offers over time when both negotiators have the same communication style than when they do not. And, when at least one negotiator is from a low-context culture, negotiators are progressively less likely to respond to emotional appeals with direct information about priorities as the negotiation progresses.

In a comparison of Asian cultures, researchers showed that Chinese negotiators are less likely than Japanese or Korean negotiators to use rational influence or share information about underlying priorities, and that Korean negotiators are less likely than Japanese or Chinese negotiators to make emotional appeals in order to improve their outcomes. These findings demonstrate that we need to focus not just on a single cultural dimension, but on a cluster of cultural dimensions, in order to understand how culture shapes negotiations.

Social Processes

Not all of the variance in negotiators' outcomes can be attributed to the social context within which negotiations occur: negotiators in the same objective conditions do not obtain identical outcomes. For example, not all negotiators with an individualistic social motive fail to create value and not all

negotiators with a cooperative social motive succeed in creating value. This observation has led negotiation and communication scholars to focus on the 'black box' of negotiation, that is, to try and link what negotiators say and do as a negotiation unfolds to their outcomes.

Strategies and Tactics

To unpack the black box, researchers have focused on the tactics that negotiators use and when they use them. Tactics can be classified according to the information that they provide (positional vs priority), the kind of relationship that they signal (dominance vs affiliation), the management of offers and concessions, and the management of the negotiation process. To understand the relationship between these tactics and negotiation outcomes, negotiation researchers have focused on the frequency with which they are used, how they are sequenced, and when in a negotiation they occur (Adair and Loewenstein, 2013).

The frequency with which strategies are used provides important information about a negotiator's dominant strategic orientation. As strategies accumulate over time, they convey either a dominant cooperative or a dominant competitive orientation. Negotiators who give and ask for information about underlying needs and priorities, who emphasize affinity and an equal relationship, and who are willing to intervene and redirect a negotiation away from competitive tactics signal their intention to cooperate. Conversely, negotiators who favor a pattern of argument and counterargument, who assert power and dominance, and who consistently make demands and threats signal their intention to compete. Unsurprisingly, whereas a broadly cooperative strategy supports value creation, a broadly competitive strategy supports value claiming.

Focusing on how strategies are sequenced, that is what one negotiator says and how his/her opponent responds, provides insight into the moment-to-moment tactical decisions that negotiators make as they strive to move toward their goals. Strategy sequences fall into one of three broad categories: reciprocal, complementary, or transformational. Both reciprocal sequences, in which negotiators exactly match each others' tactics (e.g., a demand from one negotiator elicits a demand from the other), and complementary sequences, in which negotiators match each others' overall strategic orientation (e.g., a demand from one negotiator elicits a threat from the other) preserve the overall strategic tone of a negotiation. They hold negotiators in a broadly competitive or cooperative dynamic. Transformational sequences, in which negotiators mismatch tactics (e.g., a demand from on negotiator elicits a question about underlying needs and interests from the other), offer negotiators the opportunity to change the dominant strategic tone of their negotiation. The analysis of strategy sequences demonstrates that negotiators who create competitive cycles by continually reciprocating each others' competitive tactics at best obtain very poor outcomes and at worst fail to reach agreement. Conversely, negotiators who reciprocate each others' cooperative strategies are better able to create value. Finally, this research shows a complex link between social context, strategy sequences, and paths to value creation. It shows that the strategies and strategy sequences that facilitate successful value creation vary as function of negotiators' social motives, their power, and their culture.

Phases and Interruptions

As negotiations unfold over time, there are shifts in the dominant tactics and strategies that negotiators use. Stage models of negotiation represent these shifts as occurring in an orderly way, and as necessary to move negotiations from beginning to end. These models describe a process in which negotiators start by positioning in order to identify the constraints on the negotiation, move to information exchange in order to create value, and then shift back to a more competitive approach to claim value. Phase models represent the process as a more fluid one in which strategy switches occur more frequently as negotiators strive to navigate specific obstacle to agreement (Brett et al., 2004).

Although both stage and phase models capture the emergent aspect of negotiations, they do not address the question of what triggers stage or phase shifts. The answer to this question lies in turning points (Druckman and Olekalns, 2013), events that introduce an element of surprise into negotiations because they violate negotiators' expectations. A key feature of turning points is that they are highly salient events that stand out from the flow of an ongoing negotiation, inviting greater scrutiny of an opponent's actions, and triggering a search for explanations. As a result, they can alter the underlying relationship or the entire negotiation process. For example, when turning points lead to negative attributions about an opponent's intentions they erode trust whereas when they lead to positive attributions about an opponent's intentions they build trust. Turning points may be triggered by internal processes, often occurring when one negotiator violates the expectations of an opponent. However they may also be triggered by external events such as policy and leadership changes, third-party interventions, or the introduction of new information. In either case, they provide an opportunity to change the overall flow of the negotiation either by changing the process to facilitate or inhibit agreement.

Social Perception

The opening moments of a negotiation are characterized by uncertainty and ambiguity. While negotiators are fully cognizant of their own goals and intentions, at best they have only partial information about the goals and intentions of their opponents. Because negotiators' strategy choices are, in part, determined by their assessment of whether the other party will treat them fairly or attempt to exploit them, this uncertainty needs to be resolved early in the negotiation. Consequently, negotiators rely on thin-slicing, the use of limited information and intuition, to form an impression of others' intentions. This process is critical to how negotiations unfold, with the opening moments of a negotiation predicting final outcomes (Curhan and Pentland, 2007).

Trust and Trustworthiness

The first judgment that negotiators make is about their opponents' trustworthiness. Determining an opponent's trustworthiness is a key task for negotiators: a trustworthy opponent invites problem solving and information exchange, whereas an untrustworthy opponent invites contention and argumentation. Like other impressions, the assessment of trustworthiness is made rapidly at the start of a negotiation and creates a frame for interpreting others' actions. It also sets

in motion a self-perpetuating cycle of trust or mistrust. Because negotiators who believe their opponent is trustworthy implement cooperative strategies, they invite cooperation from opponents establishing a virtuous cycle in which cooperation begets cooperation and builds trust. Conversely, negotiators who believe their opponents are untrustworthy implement competitive strategies, inviting competition from their opponents, and establishing a vicious cycle in which competition begets competition and erodes trust. The social context contributes to negotiators' initial assessments of trustworthiness: competitive negotiators trust others less than cooperative negotiators, and high-power negotiators are perceived as less trustworthy than low-power negotiators.

Microethical Judgments

A second judgment made by negotiators centers on exploitation. Negotiators ask not only whether they are likely to be exploited by an opponent, but may also assess the costs of acting exploitatively. The question of exploitation and exploitability leads us to the issue of how negotiators make ethical judgments, and the conditions under which they might choose to act unethically. Two models, moral pragmatism and opportunistic betrayal, provide insight into the factors that shape negotiators' ethical judgments. Moral pragmatism licenses negotiators to deceive their opponents under specific conditions. The central tenet of this model is that of self-protection. Negotiators are advised to be open and honest if they believe that their opponents will negotiate in good faith. However, if they believe that their opponents will try to exploit them, they are encouraged to withhold or misrepresent information in order to protect their interests. Moral pragmatism thus encourages a less ethical course of action when negotiators face adverse conditions. Opportunistic betrayal provides a counterpoint to this position. According to an opportunistic betrayal model, individuals including negotiators assess the relative costs and benefits of pursuing a less ethical course of action. Negotiators choose a less ethical path when benefits outweigh costs, implying that they may act less ethically when they believe the other party is well intentioned and will negotiate in good faith. Both models have been supported by research, which also shows that negotiators' decisions to withhold or misrepresent information are influenced by social context: social motives, power, regulatory focus, and gender all interact with trustworthiness to influence when and how negotiators deceive their opponents.

Cognition and Emotion

At the heart of a negotiation we find the individual. Although individuals' actions are shaped by the social context within which negotiations take place, what opponents say and do and the intentions that they convey are also shaped by individuals' thoughts and feelings. How negotiators represent their task, how they process information, and how they use and respond to emotion displays, all influence their strategy choices and outcomes.

Negotiators' Cognition

One of the earliest directions in negotiation research was the integration of negotiation theory with behavioral decision-making theory (Neale and Bazerman, 1991). Drawing on Prospect Theory, this research showed that negotiators' are influenced by their decision frames. Negotiators who represent their outcomes as losses are more risk seeking, make higher opening offers, and give fewer concessions than negotiators who represent their outcomes as gains. Continuing to draw on behavioral decision theory, researchers have also demonstrated that negotiators are overconfident and that they will obtain a better outcome than their opponents. When they open negotiations, they are able to anchor their opponents' expectations and so shift the negotiation in their favor. Like other decision makers, negotiators also show an egocentric bias in relation to their own costs and reactive devaluation of opponents' costs (Thompson and Hastie, 1990). The focus on decision biases and heuristics also identified two negotiation-specific biases: the fixed-pie bias, capturing the assumption that negotiators' interests are necessarily opposed, and the incompatibility bias, capturing negotiators' failure to recognize points of commonality.

Drawing on social cognition, researchers have also shown negotiators to be motivated information processors (De Dreu and Carnevale, 2003). Given the central role of accurate information in the value-creation process, negotiators' motivation to undertake an exhaustive search for information is of particular interest. The willingness to search for information is determined by individuals' epistemic motivation. Individuals with high epistemic motivation report less certainty and confidence than individuals with low epistemic motivation. As a result, they are more likely to engage in extensive information than to rely in heuristics to guide their decision making. Negotiators with high epistemic motivation spend more time searching for information about their opponents' underlying priorities and preferences and are consequently better able to create value. Epistemic motivation is sensitive to the negotiation context, and is higher when individuals have competitive goals, when they represent outcomes in terms of losses, and when they have low power. Koning and Van Dijk (2013) conclude that epistemic motivation decreases as negotiators' cognitive load increases.

Negotiators' Emotion

Although negotiators' cognitions and emotions work in parallel to shape negotiations, the focus on emotions in negotiation is relatively recent (Barry et al., 2004). The first stream of research focused on negotiators' felt emotion and investigated the impact of positive or negative emotions on information processing. This research showed that positive emotions such as happiness and optimism increase negotiators' concerns about their relationship with opponents, and as a result they favored cooperation and concessionary behavior over competition. This effect on negotiators' strategy choice was underpinned by the different expectations that positive and negative emotions induce. Whereas negotiators who report negative emotions also report low trust and the expectation that negotiations will be difficult, those who report positive emotions also report high trust and expect to reach agreement with their opponents. Recent research also links felt emotions to ethical judgments: negotiators who experience positive effect are less likely to deceive their opponents than those who experience negative effect.

Expressed emotions shape the reactions that negotiators elicit from their opponents. Researchers have predominantly focused on anger and consistently show that expressions of anger elicit concessions from opponents. The greater willingness to give concessions in the face of anger is triggered by the inference that anger signals high limits and because anger induces fear in negotiators. A smaller body of research explores other negative emotions: negotiators who experience guilt or regret are more likely to cooperate, whereas negotiators who feel disappointed are less likely to cooperate. Expressed positive emotions, on the other hand, increase the likelihood both that an agreement will be reached and that it will be implemented, and preserve future relationships. In high stakes negotiations such as hostage negotiations, the reciprocation of positive emotions by negotiators and hostage-takers facilitates a successful conclusion.

The Psychology of Conflict

Conflicts arise because one or more people (or groups) believe that they have incompatible goals. There are several factors that might lead to these incompatibilities: individuals believe others are blocking their plans and goals, scarce resources need to be divided, individuals rely on others to complete tasks, or they disagree about the best path to achieving their goals. Two factors differentiate conflicts from negotiation. The first factor is that, by the time conflicts surface, communication between individuals or groups has broken down. This means that individuals need to reestablish communication and trust before they can resolve the substantive conflict. The second factor is the higher negative emotion that typically accompanies conflicts, making it more difficult to reestablish communication.

Research in both field and experimental settings has focused on identifying the factors that trigger and escalate conflict. Many of the themes that characterize negotiation are also evident in conflict research: social context, cognition, and emotion all shape how individuals characterize and react to conflicts. Unlike negotiation researchers, conflict researchers have also focused identifying and understanding different kinds of conflicts.

Typology of Conflicts

Conflict can occur because individuals have different ideas and opinions about a task (task conflict), about the responsibilities assigned to different team members in order to accomplish the task (process conflict), or may be unrelated to the task and instead highlight differences in personal values or individuals' styles (relationship conflict). For some time researchers believed that task conflicts could be beneficial because they stimulated performance, whereas relationship conflicts would always block progress toward task completion. Although an early meta-analysis showed that team satisfaction and performance decrease as *either* task or relationship conflict increased (De Dreu and Weingart, 2003), more recent meta-analyses of the role that task, relationship, and process conflict play in achieving group goals show the relationship between conflict type and outcomes is more complex. All three types of conflict reduce trust and group member commitment. Task and relationship conflict reduce organizational citizenship behavior and counterproductive workplace behavior. In the case of relationship conflict, this effect is moderated by relational closeness. Relationship and process conflict reduce group member satisfaction and group performance, and relationship conflict reduces group member cohesion and positive affect. Turning to antecedents of conflict, stretch goals are found to intensify relationship conflicts unless there is a strong interpersonal justice climate.

Two meta-analyses suggest that the negative relationship between task conflict and performance is moderated by task type: decision-making teams and top-level management teams appear to benefit from task conflict (De Wit et al., 2012; O'Neill et al., 2013). Recent research also shows that psychological safety climate and the importance of the issues under dispute may be important moderators of the task conflict–performance relationship. Moreover, the co-occurrence of task and relationship conflicts moderates the relationship between task conflict and group performance. Finally, in analyzing how different types of conflict emerge over time, researchers have shown that task conflict leads to relationship conflict because it generates negative affect whereas relationship conflict leads to task conflict because it reduces trust in group members.

Social Context

The social contextual variables that influence negotiations also influence conflicts. In particular, research has focused on how gender and culture influence individuals' preferences for different conflict resolution styles. The underlying assumption in both streams of research that an interdependent relational orientation will lead disputants to favor relationship-preserving strategies such as avoidance and accommodation, whereas an independent relational orientation will lead disputants to favor self-oriented strategies such as competition (Gelfand et al., 2006).

Gender

Paralleling the relationship between gender and negotiation styles, gender also affect the conflict styles that individuals use and how those styles are perceived by others (Olekalns, 2014). There is consistent evidence that women favor constructive approaches to conflict resolution such as problem solving and compromise whereas men favor a more relationally damaging competitive approach. Organization-based research shows that, in response to a conflict, men express more anger than women. These gender-congruent approaches carry over from direct conflicts to situations that include a mediator. In divorce mediations, women continue to favor conciliatory strategies, focusing on their contributions to the relationship and the sacrifices that they have made whereas men focus on legal and rational arguments to justify their claims. As mediators, women favor process-oriented interventions but men favor directive action. How others respond to conflict styles is also influenced by gender. For example, mediators' trustworthiness is affected by gender: impartiality predicts trust in female mediators but empathy predicts trust in male mediators.

Culture

Research exploring cultural differences in conflict resolution typically compares individualistic and collectivist cultures.

An underlying assumption is that individualistic cultures are less concerned about preserving social harmony and face than collectivist cultures, and this difference leads them to favor different conflict resolution strategies, specifically leads individuals from collectivist cultures to favor avoidance. In reality, the relationship between culture and conflict resolution is more nuanced. First, the cultural groups frame conflicts differently. For example, whereas Japanese frame conflicts in terms of compromise and obligation, Americans frame them in terms of winning and infringements of self. Second, they respond to conflicts differently. Recent research shows that conflicts that incur loss of face are more likely to lead to anger and relational damage in collectivist than in individualistic cultures. In managing workplace conflicts, collectivists (Chinese) report they are more likely to embarrass their colleagues and to try and teach them a moral lesson, whereas individualists are more likely to report hostility and a desire for revenge. The two cultural types also report different preferences for third-party interventions: whereas collectivists favor interventions that help them to treat each other respectfully, individualists (Dutch) favor interventions that provide emotional comfort and convey a willingness to listen. Third, the consequences of conflict also differ: individualists report psychological strain in response to direct conflict whereas collectivists (Chinese) report physical symptoms of stress in response to indirect conflicts.

Cognition and Emotion

Unlike analyses of negotiators' cognition, which focus on individual information processing, analyses of disputants' cognition focus on the broad perspective through which disputants filter information. These conflict frames shape both how disputants approach a conflict and their expectations about an acceptable resolution. Finding this solution is more challenging because of the negative emotions that imbue disputes and block communication.

Disputants' Cognition

The cognitive processes that shape our understanding of conflicts have been analyzed at two distinct levels. The first level focuses on the broad perspective that individuals bring to their conflicts, that is, their conflict frames. Individuals may view their conflicts from one of three frames: power, which tests whether one disputant has greater power than the other to impose a settlement; rights, in which disputants focus on contractual obligations as a mechanism for resolving conflicts; or interests, in which disputants use a problem-solving strategy to identify a mutually acceptable solution. Both power and rights frames result in a more competitive and escalatory process than does an interest frame, especially when disputants reciprocate each other's frame. A central question is how disputants can break out of reciprocal power-rights spirals and redirect the dispute to an interest-based frame (Brett et al., 1998).

More recent research explores how individuals' process information based on their group membership. This research shows that information processing is affected by whether or not individuals who represent their groups in a conflict resolution process perceive themselves to be central to their group. Group centrality is thought to affect individuals' sense of security, and as a result affect their epistemic motivation. Recent research reports that peripheral members report higher levels of epistemic motivation, recall more information, and are more likely to identify mutually beneficial solutions, but only when they are accountable to group members (Van Kleef et al., 2013).

Disputants' Emotion

Anger functions differently in disputes: whereas expressed anger elicits concessions in negotiation, in disputes expressed anger elicits retaliation and reciprocal expressions of anger because disputants perceive it to be unfair. Unsurprisingly, expressions of anger work against agreement delaying or even stalling settlement. The negative consequences of anger can be offset if the anger elicits a sincere apology or softens the tone of communication. Venting, on the other hand, can increase anger if it is directed at third parties rather than individuals directly involved in a conflict. Other emotions affect the strategies that disputants choose to resolve conflicts. Cooperative strategies such as integrating, compromising, and yielding are facilitated by positive emotions in intragroup conflicts and by guilt in divorce mediations. Competitive strategies such as forcing and dominating are elicited by negative emotions in intragroup conflicts and by shame in divorce mediations. Finally, at least in child custody disputants, when husbands mirror their wives positive emotions, settlement becomes more likely.

Anger also influences how intergroup conflicts unfold. Contrary to the more usual perception that expressing anger is dysfunctional, recent research shows that it may reduce intergroup conflict if it is interpreted as signaling the value of the long-term relationship between the groups thereby evoking empathy. This effect may depend on timing. If expressed when a conflict is escalating, anger triggers negative consequences; if expressed after sustained attempts to de-escalate conflict it results in positive consequences. Finally, in intergroup conflicts, felt emotions are shaped by context. In intergroup conflicts, out-group members who are perceived as less warm elicit anger, whereas those who are perceived as less competent elicit contempt. These effects are underpinned by the attribution of negative intentions in the case of low warmth, and increased concerns about lack of control in the case of low competence.

See also: Conflict and Conflict Resolution, Social Psychology of; Cooperation and Competition; Culture and the Self: Implications for Psychological Theory; Heuristics in Social Cognition; Peace Promotion, Psychology of; Social Dilemmas, Psychology of.

Bibliography

Adair, W., Loewenstein, J., 2013. Talking it through: communication sequences in negotiation. In: Olekalns, M., Adair, W. (Eds.), Handbook of Research on Negotiation. Edward Elgar, London.

Barry, B., Fulmer, I.S., Van Kleef, G., 2004. I laughed, I cried, I settled: the role of emotion in negotiation. In: Gelfand, M.J., Brett, J.M. (Eds.), The Handbook of Negotiation and Culture. Stanford University Press, Stanford, CA, pp. 71–94.

Brett, J., 2007. Negotiating Globally: How to Negotiate Deals, Resolve Disputes, and Make Decisions across Cultural Boundaries, second ed. Jossey Bass, San Francisco, CA.

Brett, J.M., Shapiro, D.L., Lytle, A.L., 1998. Refocusing rights- and power-oriented negotiators toward integrative negotiations: process and outcome effects. Academy of Management Journal 41, 410–424.

Brett, J., Weingart, L., Olekalns, M., 2004. Baubles, bangles and beads: modeling the evolution of negotiating groups over time. In: Blount, S., Mannix, B., Neale, M. (Eds.), Research on Managing Groups and Teams: Time in Groups, vol. 6. JAI Press, Greenwich, CT.

Curhan, J., Pentland, A., 2007. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. Journal of Applied Psychology 92, 802–811.

De Dreu, C.K.W., Carnevale, P.J.D., 2003. Motivational bases of information processing and strategy in conflict and negotiation. Advances in Experimental Social Psychology 35, 235–291.

De Dreu, C.K.W., Weingart, L.R., 2003. Task versus relationship conflict, team effectiveness, and team member satisfaction: a meta-analysis. Journal of Applied Psychology 88, 741–749.

De Dreu, C.K.W., Weingart, L.R., Kwon, S., 2000. Influence of social motives on integrative negotiation: a meta-analytic review and test of two theories. Journal of Personality and Social Psychology 78, 889–905.

De Wit, F., Greer, L., Jehn, K., 2012. The paradox of intragroup conflict: a meta-analysis. Journal of Applied Psychology, 360–390.

Druckman, D., Olekalns, M., 2013. Punctuated negotiations: transitions, interruptions and turning points in negotiation. In: Olekalns, M., Adair, W. (Eds.), Handbook of Research on Negotiation. Edward Elgar, London.

Gelfand, M.J., Major, V., Raver, J.L., Nishii, L.H., O'Brien, K., 2006. Negotiating relationally: the dynamics of relational self in negotiations. Academy of Management Review 31, 427–451.

Kim, P.H., Pinkley, R.L., Fragale, A.R., 2005. Power dynamics in negotiation. The Academy of Management Review 30, 799–822.

Koning, L., Van Dijk, E., 2013. Motivated cognition in negotiation. In: Olekalns, M., Adair, W. (Eds.), Handbook of Research on Negotiation. Edward Elgar, Northhampton, MA, pp. 46–74.

Kray, L.J., Thompson, L., 2005. Gender and negotiation. Research in Organizational Behavior 26, 102–182.

Kulik, C., Olekalns, M., 2012. Negotiating the gender divide: lessons from the negotiation and organizational behavior literatures. Journal of Management 38, 1387–1415.

Neale, M.A., Bazerman, M.H., 1991. Cognition and Rationality in Negotiation. Free Press, New York.

Olekalns, M., 2014. Natural-born peacemakers? Gender and the resolution of conflict. In: Coleman, P., Deutsch, M., Marcus, E., Dillard, J. (Eds.), The Handbook of Conflict Resolution: Theory and Practice, third ed. Jossey-Bass, San Francisco, CA.

O'Neill, T.A., Allen, N.J., Hastings, S.E., 2013. Examining the "pros" and "cons" of team conflict: a team-level meta-analysis of task, relationship, and process conflict. Human Performance 26, 236–260.

Pruitt, D.G., 1981. Negotiation Behavior. Academic Press, New York.

Rahim, M.A., 1983. A measure of styles of handling interpersonal conflict. Academy of Management Review 26, 368–376.

Thompson, L., Hastie, R., 1990. Social perception in negotiation. Organizational Behavior and Human Decision Processes 47, 98–123.

Van Kleef, G., Steinel, W., Homan, A., 2013. On being peripheral and paying attention: prototypicality and information processing in intergroup conflict. Journal of Applied Psychology 98, 63–79.

Walters, A.E., Stuhlmacher, A.F., Meyer, L.L., 1998. Gender and negotiator competitiveness: a meta-analysis. Organizational Behavior and Human Decision Processes 76, 1–29.

Negotiations: Statistical Aspects

James K Sebenius, Harvard Business School, Harvard University, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Negotiation analysis' seeks to develop prescriptive theory and useful advice for negotiators and third parties. It generally emphasizes the parties' underlying interests, alternatives to negotiated agreement, approaches to productively manage the inherent tension between competitive actions to 'claim' value individually and cooperative ones to 'create' value jointly, as well as efforts to change perceptions of the game itself. Since advice to one side does not necessarily presume the full game-theoretic rationality of the other side(s), negotiation analysts often draw on the findings of behavioral scientists and experimental economists. Further, this approach does not generally assume that all the elements of the 'game' are common knowledge. It tends to de-emphasize the application of game-theoretic solution concepts or efforts to find unique equilibrium outcomes. Instead, to evaluate possible strategies and tactics, negotiation analysts generally focus on changes in perceptions of the 'zone of possible agreement' and the (subjective) distribution of possible negotiated outcomes conditional on various actions. It has been used to develop prescriptive advice for the simplest bilateral negotiations between monolithic parties, for negotiations through agents, or with linked 'internal' and 'external' aspects, for negotiations in hierarchies and networks, as well as for more complex coalitional interactions.

Introduction

The term 'negotiation analysis' encompasses an emergent body of work that seeks to generate prescriptive theory and advice for parties wishing to negotiate in a 'rational' manner, without necessarily assuming that the other side(s) is a strategically sophisticated, rational actor.

To what bodies of theory and empirical work might one turn for advice in a negotiation over selling a car or forging a global warming treaty? Decades of psychological studies of people negotiating offer many powerful insights but generally lack a prescriptive theory of action. Instead, this behavioral approach has largely sought accurate descriptions and analytic explanations of how negotiators do act. What they should do often remains an implicit or ad hoc implication of this careful work.

Decision analysis, the practically oriented cousin of decision theory, would seem a logical candidate to fill this prescriptive void. It suggests a systematic decomposition of the problem: structuring and sequencing the parties' choices and chance events, then separating and subjectively assessing probabilities, values, risk, and time preferences. A von Neumann–Morgenstern expected utility criterion aggregates these elements in ranking possible actions to determine the optimal choice. This approach is well suited to decisions 'against nature,' in which the uncertainties – such as the probability that a hurricane will strike Miami in August – are not affected by the choices of other involved parties anticipating one's actions.

Yet when decision-making is interactive – as is true in negotiation, where each party's anticipated choices affect the other's and vice versa – assessment of what the other side will do qualitatively differs from assessment of 'natural' uncertainties. Of course, the theory of games was developed to provide a logically consistent framework for analyzing such interdependent decision-making. Full descriptions of the courses of action open to each involved party are encapsulated into 'strategies.' Rigorous analysis of the interaction of strategies leads to a search for 'equilibria,' or plans of action such that each party, given the choices of the other parties, has no incentive to change its plans. A great deal of analysis by game theorists seeks conditions for unique equilibria among such strategies (von Neumann and Morgenstern, 1944; Luce and Raiffa, 1957; Aumann, 1989).

Game theory has been especially useful for understanding repeated negotiations in well-structured financial situations. It has offered useful guidance for the design of auction and bidding mechanisms, has uncovered powerful competitive dynamics, has usefully analyzed many 'fairness' principles, and now flourishes both on its own and in applications such as microeconomic theory. Despite signal successes, the dominant game-theoretic quest to predict equilibrium outcomes resulting from the strategic interactions of fully rational players often lacks prescriptive power in negotiations.

Three major aspects of mainstream game theory, discussed at length in Sebenius (1992), contribute to this 'prescriptive gap.' First, on standard assumptions, there are often numerous plausible equilibrium concepts, each with many associated equilibria – and no a priori compelling way to choose among them. Second, even where one party wishes to act rationally, the other side may not behave as a strategically sophisticated, expected utility-maximizer – thus rendering conventional equilibrium analyses less applicable. A large and growing body of evidence – especially in 'behavioral game theory' and experimental economics – suggests that people systematically and significantly violate the canons of rationality. Third, the elements, structures, and 'rules' of many negotiating situations are not completely known to all the players, and even the character of what is known by one player may not be known by another. The frequent lack of such 'common knowledge' limits – from a prescriptive standpoint – much equilibrium-oriented game analysis, though recent work has begun to relax the widespread understanding that common knowledge of the situation was essential to game models (Aumann and Brandenburger, 1995). Even where it is possible to bend such a situation

into the form of a well-posed game, and gain insights from it, the result may lose prescriptive relevance.

The Negotiation Analytic Response

If descriptive psychological approaches to negotiation lack a prescriptive framework, if decision analysis is not directly suited to interactive problems, and if traditional game theory presupposes too much rationality and common knowledge, then 'negotiation analysis' represents a response that yokes the prescriptive and descriptive research traditions (Sebenius, 2009). Using Howard Raiffa's (1982) terms, unlike the 'symmetrically prescriptive' approach of game theory, wherein fully rational players are analyzed in terms of what each should optimally do given the other's optimal choices, the 'asymmetrically prescriptive/descriptive' approach typically seeks to generate prescriptive advice to one party conditional on a (probabilistic) description of how others will behave. This need not mean tactical naiveté; as appropriate, the assessments can incorporate none, a few, or many rounds of 'interactive reasoning.'

Works that embody the spirit of this work can be found as early as the late 1950s. While Luce and Raiffa's *Games and Decisions* (1957) was primarily a brilliant synthesis and exposition of game theory's development since von Neumann and Morgenstern's (1944) classic work, Luce and Raiffa began to raise serious questions about the limits of this approach in analyzing actual interactive situations. Perhaps, the first work that could be said to be in the spirit of 'negotiation analysis' was *The Strategy of Conflict* by Thomas Schelling (1960). Its point of departure was game-theoretic, but it proceeded with less formal argument and the analysis had far broader direct scope. Though nominally in the behavioral realm, Walton and McKersie's *Behavioral Theory of Labor Negotiation* (1965) drew on Schelling's work as well as rudimentary decision and game theories.

The first overall synthesis of this emerging field appeared with Raiffa's (1982) *The Art and Science of Negotiation*, elaborated in Raiffa et al. (2002). Building on Sebenius (1984), this approach was systematized into an overall method in Lax and Sebenius' *The Manager as Negotiator* (1986) and *3D Negotiation* (2006). *Negotiation Analysis* (1991), edited by Peyton Young, furthered this evolving tradition, which was characterized and analytically evaluated in Sebenius (1992). Further contributions in the same rationalist vein were included in Zeckhauser et al. (1996), Arrow et al. (1995), Subramanian (2011), and, adding insights from organizational and information economics, in Mnookin et al. (2000).

Meanwhile, another group of researchers was coming to a negotiation analytic view from a behavioral point of departure. With roots in the cognitive work of behavioral decision theorists, e.g., Bell et al. (1988), behavioral scholars began in the early 1990s to explicitly link their work to that of Raiffa and his colleagues. In particular, Neale and Bazerman's *Cognition and Rationality in Negotiation* (1991), Bazerman and Neale (1992), and Thompson (2011) pulled together and developed a great deal of psychological work on negotiation in an asymmetrically prescriptive/descriptive framework. These efforts began systematically to build up more structure on what had been, in the works of Raiffa et al. a largely ad hoc descriptive side of the ledger.

Elements of a Negotiation Analytic Approach

Full negotiation analytic accounts (e.g., Sebenius, 1992, 2000) generally consider the following basic elements: the actual and potential parties; their perceived interests; alternatives to negotiated agreement; the linked processes of 'creating' and 'claiming' value; and the potential to 'change the game' itself.

Parties

In the simplest negotiation, two principals negotiate with each other. Yet potentially complicating parties such as lawyers, bankers, and other agents may be present as may multiple internal factions with very different interests. The crucial first step for an effective negotiation analysis is to map the full set of potentially relevant parties in the context of the decision processes.

Interests

The next step is to probe deeply for each relevant party's or faction's underlying interests and to assess their trade-offs carefully. When individuals or groups with different concerns constitute a negotiating 'side,' it is no longer in general possible to specify overall tradeoffs; however, carefully tracing which set of interests is ascendant according to the internal bargaining process of given factions may continue to provide insights. Raiffa (1982) offers an extended discussion of assessing trade-offs in negotiation, building on extensive work by Keeney and Raiffa (1976; Sebenius, 1992). This assessment is radically subjective in the sense that less-tangible concerns for self-image, fairness, process, precedents, or relationships can have the same analytic standing as the 'harder' or 'objective' interests such as cost, time, and quality that are common to traditional economic approaches.

It is often useful to distinguish the full set of the parties' underlying *interests* from the *issues* under negotiation, on which *positions* or stands are taken. The connections among positions on issues and interests are rarely simple. Sometimes a focus on deeper interests can unblock a stubborn impasse over incompatible positions that relate only partially to the parties' real concerns. For example, the nominal issue under negotiation may be the base salary amount with incompatible positions (the boss says lower while the employee demands higher). Yet if the employee seeks a salary boost to pay for advanced technical training (an interest), the boss may realize that providing such training directly could be cheaper and in the company's interest as well. In other cases, emphasizing interests will only generate hopeless conflict when mutually beneficial agreement on certain overt positions could be reached.

Alternatives to Negotiated Agreement

People negotiate in order to satisfy the totality of their interests better through some jointly decided action than they could otherwise (typically by unilateral action such as shifting to an

alternative supplier, going to court, or striking). Thus, for each side, the basic test of a proposed joint agreement is whether it offers higher subjective worth than that side's best course of action absent agreement. In examining a negotiation, one should analyze each party's perceptions of its own – and the others' – valuations of their alternatives to negotiated agreement.

Alternatives to agreement may be certain, with a single attribute: an ironclad competing price quote for an identical new car. They may be contingent and multiattributed: going to court rather than accepting a negotiated settlement can involve uncertainties, trial anxieties, costs, time, and precedents that contrast with the certain, solely monetary nature of a pretrial accord. No-agreement alternatives may also involve potentially competing coalitions, threats, and counterthreats or the necessity to keep negotiating indefinitely. Various analytic tools from decision analysis to optimal search theory may aid in assessment of such alternatives (see Sebenius, 1992 for references).

While this evaluation provides a strict lower bound for the minimum worth (the 'reservation price') required of any acceptable settlement, alternatives to agreement also play tactical roles. The more favorably negotiators portray their best alternative course of action and willingness to 'walk away,' the smaller is the ostensible need for the negotiation and the higher the standard of value that any proposed accord must reach. Moves 'away from the table' that shape the parties' alternatives to agreement can strongly affect negotiated outcomes. Searching for a better price or another supplier, cultivating a friendly merger partner in response to a raider, or preparing an invasion should negotiations fail to resolve a contested issue, may influence negotiations more than sophisticated tactics employed 'at the table' such as clever opening offers or patterns of concession.

Representing the Structure

Imagine that two negotiators have thought hard about their underlying interests in different possible settlements of the apparent issues. Further, suppose that they have a relatively clear, if possibly changing, assessment of their trade-offs and have compared them to the value of their best no-agreement alternatives. Each has a sense of any 'rules of engagement' that structure their interaction. From the viewpoint of each party, a set of possible agreements has been envisioned. Assume that an analyst was privy to the results of these evaluations by each side, along with the (likely asymmetric) distribution of information about interests, beliefs, no-agreement options, and possible actions; these evaluations need not be common knowledge of the parties. The situation might be familiarly represented as in Figure 1.

The origin represents the value of failing to reach agreement: each side's best alternative to agreement implies the location of this point. The 'Pareto frontier' in the northeast part of the graph represents the evaluations of the set of those possible agreements on the issues that could not be improved on from the standpoint of either party without harming the other. In general, neither side knows the location of the frontier, only theoretically that it is there. The entire shaded region – bounded by the two axes and the frontier – is the 'zone of possible agreement.' In general, each party has its own

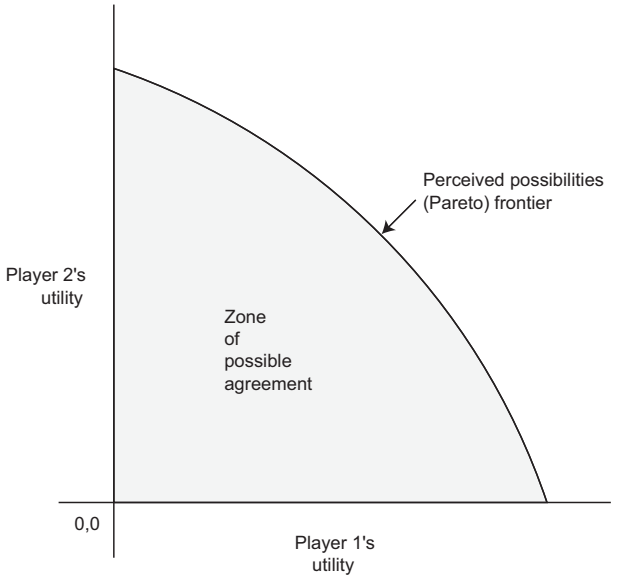

Figure 1 Representation of the zone of possible agreement.

perceptions of it. (In a purely distributive bargain, with no room for joint gains beyond the fact of agreement, the shaded region would collapse to a diagonal frontier along which better terms for one party imply worse terms for the other – even though both parties prefer such deals to impasse.) Since this representation is quite general, it can in principle encompass the whole range of possible interests, alternatives, and agreements.

From Structure to Outcome: Favorable Changes in the Perceived Zone of Possible Agreement

This is the point at which game-theoretic and negotiation analytic approaches tend to diverge. A game theorist would typically tighten the above specification of the situation, presume common knowledge of the situation and strategic rationality of the parties, and, by invoking a rigorous concept of equilibrium, investigate the predicted outcome(s) of the interaction. Indeed, as Rubenstein's insightful commentary noted, "for forty years, game theory has searched for the grand solution" that would achieve a

> prediction regarding the outcome of interaction among human beings using only data on the order of events, combined with a description of the players' preferences over the feasible outcomes of the situation.
>
> 1991: p. 923

Even with as powerful and ubiquitous a concept as that of the Nash equilibrium in noncooperative games, however, it is often impossible, even with the imposition of increasingly stringent requirements or refinements to limit a game's equilibrium outcomes to a unique or even small number of points. Often there is an infinitude of such outcomes. As Tirole noted when explaining why "we are now endowed with nearly a dozen refinements of perfect Bayesian equilibrium," the

> leeway in specifying off-the-equilibrium-path beliefs usually creates some leeway in the choices of equilibrium actions; by ruling out some potential equilibrium actions, one transforms other actions into equilibrium actions.
>
> 1988: p. 446

Frequently this implies an infinite number of perfect Bayesian equilibria.

Despite insights into how rational players might select from among multiple Nash equilibria (Harsanyi and Selten, 1988), the rationale for a particular choice may ultimately seem arbitrary. As Kadane and Larkey remark,

> we do not understand the search for solution concepts that do not depend on the beliefs of each player about the others' likely actions, and yet are so compelling that they become the obvious standard of play for all those who encounter them.
>
> 1982: pp. 115–116

This seems especially apt in light of their observation that "solution concepts are a basis for particular prior distributions" and hence

> the difficulty in non-zero sum, N-person game theory of finding an adequate solution concept: no single prior distribution is likely to be adequate to all players and all situations in such games.
>
> 1982: pp. 115–116

Since each party should accept any settlement in the zone of possible agreement rather than no agreement, Schelling (1960) made the potent observation that the outcome of such a situation could only be unraveled by a 'logic of indeterminate situations.' Without an explicit model or formal theory (equilibrium-based or other) adequate to map structure and tactics onto bargaining outcomes with confidence, how can an individual negotiator or interested third party decide what to do? In the (often implicit) view of many negotiation analysts, the negotiator's subjective distribution of beliefs about the negotiated outcome conditional on using the proposed tactics must be compared with his subjective distribution of beliefs about the outcome conditional on not using them. The tactic is attractive if the former distribution gives him higher expected utility than the latter.

Such 'improvement' has a subjective basis analogous to the Rothschild and Stiglitz (1970) characterization of a subjectively perceived 'increase' in risk. Specifying these distributions may require an internalized and subjective model of the bargaining process, since no such general model exists; where there is a well-developed and applicable game-theoretic model, of course, it should be used. Of course, the 'better' the empirical and theoretical basis for the assessment, the 'better' the subjective distributions of outcomes.

Much negotiation analytic work consists in improving the basis for assessing such outcome distributions. To do this, analysts pay attention to the special structure and dynamics that derive from the joint decision-making among parties, some of whom may be 'boundedly' rational.

Improving the Basis for Assessing Outcome Distributions I: Fundamental Processes of Negotiation

At bottom, negotiation processes involve both (1) actions to enhance what is jointly possible through agreement ('creating value') and (2) actions to allocate the value of agreement ('claiming value').

1. *Creating Value*. In most negotiations, the potential value of joint action is not fully obvious at the outset. 'Creating value' – that is, reaching mutually beneficial agreements, improving them jointly, and preventing conflict escalation – requires an approach often associated with 'win–win,' 'integrative,' or 'variable sum' encounters: to share information; communicate clearly; be creative; and productively channel hostilities. Yet regardless of whether one adopts a cooperative style or not, it is useful to have an analytic guide as to the underlying substantive bases for joint gains and how to realize them by effective 'dealcrafting.'
   First, apart from pure shared interests, negotiators may simply want the same settlement on some issues. Second, where economies of scale, collective goods, alliances, or requirements for a minimum number of parties exist, agreement among similar bargainers can create value. Third – and most interestingly – though many people instinctively seek 'common ground' and believe that 'differences divide us,' it is often precisely the differences among negotiators that constitute the raw material for creating value. Each class of difference has a characteristic type of agreement that makes possible its conversion into mutual benefit. For example, differences in relative valuation suggest joint gain from trades or from 'unbundling' differently valued attributes. Differences in tax status, liquidity, or market access suggest arbitrage. Complementary technical capacities can be profitably combined. Probability and risk aversion differences suggest contingent agreements or bets. Differences in time preference suggest altering schedules of payments and other actions. Sebenius (1984) formally characterizes such optimal betting, risk sharing, and temporal reallocations. Much work on optimal contracting (e.g., Hart, 1995) as well as classical economic study of gains from trade and comparative advantage directly bear on the kinds of agreements that create value on a sustainable basis. Sebenius (1992) and Lax and Sebenius (2002, 2006) offer extensive illustrations of these 'dealcrafting' or 'deal design' principles in practice.
2. *Claiming Value*. Crucial aspects of most negotiations, however, are primarily 'distributive,' 'win–lose,' or 'constant-sum,' that is, at some points in the process, increased value claimed by one party implies less for others. Several broad classes of tactics used for 'claiming value' in these kinds of bargains have been explored (e.g., Schelling, 1960; Walton and McKersie, 1965; Raiffa, 1982; Lax and Sebenius, 1986, 2006). Such tactics include: shaping perceptions of alternatives to agreement; making commitments; influencing aspirations; taking strong positions; manipulating patterns of concessions; holding valued issues 'hostage'; linking issues and interests for leverage; misleading other parties; and exploiting cultural and other expectations. By means of these tactics, one party seeks advantage by influencing another's perceptions of the zone of possible agreement.

Managing the Tension between Creating and Claiming Value: *The Negotiators'* Dilemma

In general, the benefits of cooperation are not fully known at the outset of a negotiation. Colloquially, the parties often do not know how large a value pie they can create. The way in which they attempt to expand the pie often affects its final division, while each side's efforts to obtain a larger share of the pie often prevents its expansion in the first place – and may lead to no pie at all, or even to a fight. Conclusion: creating and claiming value are not in general separable processes in negotiation. This fact undermines much otherwise useful advice (that, for example, presumes 'win–win' situations to have no 'win–lose' aspects, or 'integrative' bargains to be unrelated to 'distributive' ones).

Each negotiating party tends to reason as follows: if the other parties are open and forthcoming, I can take advantage of them and claim a great deal of value; thus I should adopt a value-claiming stance. By contrast, if the other parties are tough and adopt value-claiming stances, I must also adopt such a stance in order to protect myself. Either way, a strong tendency operating on all parties drives them toward hardball. Since mutual hardball generally impedes value creation, competitive moves to claim value individually often drive out cooperative moves to create it jointly. Outcomes of this dynamic include poor agreements, deadlocks, and conflict spirals. This tendency, closely related in structure to the famous Prisoner's Dilemma, was dubbed the 'Negotiator's Dilemma' (Lax and Sebenius, 1986: Chapters 2 and 7). Much negotiation advice is aimed at productively managing the inherent tension between creating and claiming value, especially to do so on a sustainable basis (Lax and Sebenius, 1981; Raiffa, 1982).

Improving the Basis for Assessing Outcome Distributions II: Behavioral Insight

A fully rational 'baseline' analysis helps to understand the possible responses of a rational other side. Urging consistent, if not fully rational, behavior on the subject of one's advice is often wise. After all, well-structured, repeated negotiations may penalize departures from rational behavior. Yet many negotiating situations are neither well-structured, repeated, nor embedded in a market context. While negotiators normally exhibit intelligent, purposive behavior, there are important departures from the 'imaginary, idealized, superrational people without psyches' (Bell et al., 1988: p. 9) needed by many game-theoretic analyses.

Considerable empirical work – such as that cited above by Bazerman, Neale, Thompson, and their colleagues – offers considerable insight into how people actually behave in negotiations. Excellent reviews of the psychological side of negotiation can be found in Bazerman et al. (2000a) and, focusing especially on developments on the social psychological side, in Bazerman et al. (2000b). Complementing this work is the burgeoning research in experimental economics (Kagel and Roth, 1995) and what Colin Camerer (1997) described as a 'behavioral game theory.' This work blends game-theoretic and psychological considerations in rigorous experimental settings. Two related levels are consistently important, the individual and the social:

1. *Individual negotiating behavior*. As negotiators, people have different histories, personalities, motivations, and styles. Systematic characteristics of individual cognitive processes can both help and hinder the process of reaching agreement. For example, negotiators may be anchored by irrelevant information, subject to inconsistencies in the way they deal with uncertainty, hampered by selective perception, obsessed by sunk costs and past actions, prone to stereotyping and labeling, susceptible to influence tactics and manipulation by the way in which equivalent situations are framed, liable to see incompatible interests when joint gains are possible, and use a variety of potentially misleading heuristics to deal with complexity, ambiguity, and conflict.
2. *Social behavior in negotiation*. In groups of two or more, especially where there is some perceived conflict, a variety of social psychological dynamics come into play that may enable or block negotiated outcomes. For example, a powerful norm toward reciprocity operates in most groups and cultures. Tentative cooperative actions by one side can engender positive reactions by the others in a cycle that builds trust over time. By contrast, social barriers can involve aspects of the interactive process that often lead to bad communication, misattribution, polarization, and escalation of conflict, as well as group dynamics that work against constructive agreements. Such dynamics may be especially pronounced when negotiations involve players of different genders, races, or cultures. Group dynamics can involve pressures for conformity, a hardening of approach by a representative before an 'audience' of constituents, bandwagon effects, and the individual taking cues from the behavior of others to decide on appropriate actions in the negotiation.

This experiment-based work is developing an empirical grounding for the behavioral basis of much a priori theorizing in economics and game theory. For negotiation analysis, these experimental approaches to actual behavior help to remedy a key defect of prior game-theoretic work. While for the most part not prescriptively framed (but see Malhotra and Bazerman, 2008), this body of work also provides rigorous evidence and theory on how people in fact are likely to behave – to inform assessments of outcome distributions and against which to optimize as appropriate.

Changing the Game

Much existing theory proceeds from the assumption of a well-specified and fixed situation within which negotiation actions are taken. Yet purposive action on behalf of the parties can change the very structure of the situation and hence the outcomes. Often actions can be understood as a tacit or explicit negotiation over what the game itself will be. First investigations into this phenomenon gave rise to the terms 'negotiation arithmetic' or 'adding' and 'subtracting' issues and parties (Sebenius, 1984). This means that a perfectly legitimate and potentially valuable form of analysis may involve a search for ways to change the perceived game – even though the menu of possibilities may not be common knowledge.

Issues can be linked or separated from the negotiation to create joint gains or enhance leverage. Parties may be 'added' to a negotiation to improve one side's no-agreement alternatives, as well as to generate joint gains or to extract value from others. Though perhaps less commonly, parties can also be 'subtracted' – meaning separated, ejected, or excluded – from larger potential coalitions. For example, the Soviets were excluded from an active Middle East negotiating role in the process leading up to the Camp David accords that only involved Israel, Egypt, and the United States. The process of choosing, then approaching and persuading others to go along may best be studied without the common assumption that the game is fully specified at the outset of analysis; Sebenius (1996) dissects and offers many examples of the process sequencing to build or break coalitions. Walton and McKersie (1965) focused on how negotiators seek to change perceptions of the game by what they called 'attitudinal restructuring.' Similarly, Lax and Sebenius (2006) systematically analyze many classes of purposeful changes in the 'setup' of a negotiation. In the context of competitive strategy and thinking, Brandenburger and Nalebuff (1996) develop a powerful, analogous logic for 'changing the game' that provides an overall approach and many ingenious examples of this phenomenon.

Refer back to **Figure 1**, an improvement in Player 1's no-agreement alternative shifts the vertical axis to the right, leaving the bargaining set generally more favorable to that side. If Player 2's no-agreement alternative worsens, the horizontal axis shifts down, worsening its prospects. A successful commitment to a bargaining position cuts off an undesired part of the zone of possible agreement for the party who makes it. A new, mutually beneficial option (e.g., suggestion of a contingent, rather than an unconditional, contract) causes the frontier to bulge upward and to the right, reducing the parties' 'conflict of interest.' When issues change or other basic aspects of the game vary, each side's perceptions of the basic picture in **Figure 1**, the zone of possible agreement, will be transformed. These possibilities add evolutionary elements to the analysis.

The Approach as a Whole

Figure 1 can now be seen to summarize visually the extended negotiation analytic 'model' of possible joint action. Parties determine the axes; interests provide the raw material and the measure; alternatives to agreement imply the limits; agreements hold out the potential. Within this configuration, the process consists of creating and claiming value, which gives rise to characteristic dynamics; yet, the elements of the interaction may themselves evolve or be intentionally changed. In this sense, the elements of the approach form a logically consistent, complete whole oriented around perceptions of the zone of possible agreement.

In the skeptical view of Harsanyi, this negotiation analytic approach might boil down to

> the uninformative statement that every player should maximize expected utility in terms of his subjective probabilities without giving him the slightest hint of how to choose these subjective probabilities in a rational manner.
>
> 1982

Yet, as described above, distinct classes of factors have been isolated that appear to improve subjective distributions of negotiated outcomes. Understanding the dynamics of creating and claiming value can improve prescriptive confidence. Psychological considerations can help as can cultural observations, organizational constraints and patterns, historical similarity, knowledge of systematic decision-making biases, and contextual features. Less than full-blown, game-theoretic reasoning can offer insight into strategic dynamics as can blends of psychological and game-theoretic analysis. When one relaxes the assumptions of strict, mutually expected, strategic sophistication in a fixed game, Raiffa's conclusion is appealing: that some "analysis – mostly simple analysis – can help" (1982: p. 359).

Conclusions and Further Directions

Naturally, there are many other related topics ranging from game-theoretic concepts of fairness for purposes of mediation and arbitration to various voting schemes. More elaborate structures are under study. For example, where negotiation takes place through agents, whether lawyers or diplomats, or where a result must survive legislative ratification, the underlying structure of a 'two-level game' is present (see Putnam, 1988). Negotiations also take place in more complex multilevel and coalitional structures (Raiffa, 1982; Lax and Sebenius, 1986, 1991; Sebenius, 1996). In many cases, the individual negotiation is not the appropriate unit of analysis. Instead, the study of 'negotiation design' – for example, of international conferences, alternative dispute resolution procedures, grievance negotiation processes – as well as analysis of multifront 'negotiation campaigns' seeks to maximize the chances of constructive outcomes from different structures and processes (Sebenius, 1991, 2010; Lax and Sebenius, 2012).

While game theorists and behavioral scientists will continue to make valuable progress in understanding negotiation, a complementary prescriptive approach has been developing that conditions its prescriptions on the likely behavior of the other side, fully 'rational' or not, and regardless of whether the 'game' is fixed and entirely common knowledge. In describing the logic of negotiation analysis and the concepts and tools that can facilitate it, this discussion has not stressed the many useful ideas that arise from focusing on interpersonal and cultural styles, on atmosphere and logistics, on psychoanalytic motivation, on communication, or on other aspects. Yet because the logic is general, it can profitably accommodate insights from other approaches as well as from experience. The basic elements of this logic – parties' perceptions of interests, alternatives, agreements, the processes of creating and claiming value, and changing the game – become the essential filters through which other factors must be interpreted for a meaningful assessment of the zone of possible agreement and its implications for the outcome.

See also: Data Bases and Statistical Systems: Security and Conflict; Decision and Choice: Bounded Rationality; Decision and Choice: Economic Psychology; Decision and Choice: Heuristics; Experimental Laboratories: Psychological; Experimental Laboratories: Social and Economic; Game Theory: Noncooperative Games; Game Theory; Models of Power in Exchange Networks; Multidimensional Scaling I; Negotiation and Bargaining: Organizational Aspects; Negotiation and Bargaining: Role of Lawyers; Organizations: Negotiated Orders; Probabilistic Thinking; Trade Unions, Economic Behavior of.

Bibliography

Arrow, K.J., Wilson, R., Ross, L., Tversky, A., Mnookin, R.H., 1995. Barriers to Conflict Resolution. W.W. Norton, New York.

Aumann, R.J., 1989. Game theory. In: Eatwell, J., Milgate, M., Newman, P. (Eds.), Game Theory. Norton, New York, pp. 1–53.

Aumann, R.J., Brandenburger, A., 1995. Epistemic conditions for Nash equilibrium. Econometrica 63, 1161–1180.

Bazerman, M.H., Curhan, J., Moore, D., 2000a. The death and rebirth of the social psychology of negotiations. In: Fletcher, G., Clark, M. (Eds.), Blackwell Handbook of Social Psychology: Interpersonal Processes. Blackwell, Oxford, UK, pp. 196–228.

Bazerman, M.H., Curhan, J., Moore, D., Valley, K.L., 2000b. Negotiations. Annual Review of Psychology. Annual Reviews, Palo Alto, CA.

Bazerman, M.H., Neale, M.A., 1992. Negotiating Rationally. Free Press, New York.

Bell, D.E., Raiffa, H., Tversky, A., 1988. Decision Making: Descriptive, Normative, and Prescriptive Interactions. Cambridge University Press, Cambridge, UK.

Brandenburger, A.M., Nalebuff, B.J., 1996. Co-Opetition. Currency Doubleday, New York.

Camerer, C.F., 1997. Progress in behavioral game theory. Journal of Economic Perspectives 11 (4), 167–188.

Harsanyi, J.C., 1982. Subjective probability and the theory of games: comments on Kadane and Larkey's paper. Management Science 28 (2), 120–124.

Harsanyi, J.C., Selten, R., 1988. A General Theory of Equilibrium Selection in Games. MIT Press, Cambridge, MA.

Hart, O., 1995. Firms, Contracts, and Financial Structure. Clarendon Lectures in Economics. Oxford University Press, Oxford, UK.

Kadane, J.B., Larkey, P.D., 1982. Subjective probability and the theory of games. Management Science 28 (2), 113–120.

Kagel, J., Roth, A.E., 1995. The Handbook of Experimental Economics. Princeton University Press, Princeton, NJ.

Keeney, R., Raiffa, H., 1976. Decisions with Multiple Objectives: Preferences and Value Tradeoffs. John Wiley, New York.

Lax, D.A., Sebenius, J.K., 1981. Insecure contracts and resource development. Public Policy 29, 417–436.

Lax, D.A., Sebenius, J.K., 1986. The Manager as Negotiator: Bargaining for Cooperation and Competitive Gain. Free Press, New York.

Lax, D.A., Sebenius, J.K., 1991. Thinking coalitionally: party arithmetic, process opportunism, and strategic sequencing. In: Young, H.P. (Ed.), Negotiation Analysis. University of Michigan Press, Ann Arbor, MI, pp. 153–193.

Lax, D.A., Sebenius, J.K., 2002. Dealcrafting: the substance of three dimensional negotiation. Negotiation Journal 18 (1), 5–28.

Lax, D.A., Sebenius, J.K., 2006. 3-D Negotiation: Powerful Tools to Change the Game in Your Most Important Deals. Harvard Business School Press, Boston.

Lax, D.A., Sebenius, J.K., 2012. Deal making 2.0: a guide to complex negotiations. Harvard Business Review 90 (11), 92–100.

Luce, R.D., Raiffa, H., 1957. Games and Decisions. Wiley, New York.

Malhotra, D., Bazerman, M., 2008. Negotiation Genius. Bantam, New York.

Mnookin, R.H., Peppet, S.R., Tulumello, A.S., 2000. Beyond Winning: Negotiating to Create Value in Deals and Disputes. Harvard University Press, Cambridge, MA.

Neale, M.A., Bazerman, M.H., 1991. Cognition and Rationality in Negotiation. Free Press, New York.

von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ.

Putnam, R.D., 1988. Diplomacy and domestic politics: the logic of two-level games. International Organization 4Z, 427–460.

Raiffa, H., 1982. The Art and Science of Negotiation. Belknap Press, Harvard University Press, Cambridge, MA.

Raiffa, H., Richardson, J., Metcalfe, D., 2002. Negotiation Analysis: The Science and Art; of Collaborative Decision Making. Belknap Press of Harvard University Press, Cambridge, MA.

Rothschild, M., Stiglitz, J.E., 1970. Increasing risk, I: a definition. Journal of Economic Theory 2, 225–243.

Rubinstein, A., 1991. Comments on the interpretation of game theory. Econometrica 59, 909–924.

Schelling, T.C., 1960. The Strategy of Conflict. Harvard University Press, Cambridge, MA.

Sebenius, J.K., 1984. Negotiating the Law of the Sea: Lessons in the Art and Science of Reaching Agreement. Harvard University Press, Cambridge, MA.

Sebenius, J.K., 1991. Designing negotiations toward a new regime: the case of global warming. International Security 15 (4), 110–148.

Sebenius, J.K., 1992. Negotiation analysis: a characterization and review. Management Science 38 (1), 18–38.

Sebenius, J.K., 1996. Sequencing to build coalitions: with whom I should talk first? In: Zeckhauser, R., Keeney, R., Sebenius, J.K. (Eds.), Wise Decisions. Harvard Business School Press, Cambridge, MA, pp. 324–348.

Sebenius, J.K., 2000. Dealmaking Essentials: Creating and Claiming Value for the Long Term. Harvard Business School, Boston. Publishing Item 2-800-443.

Sebenius, J.K., 2009. Negotiation analysis: from games to inferences to decisions to deals. Negotiation Journal 25 (4), 449–465.

Sebenius, J.K., 2010. Beyond the deal: wage a 'negotiation campaign'. Negotiation 13 (11), 1–2.

Subramanian, G., 2011. Dealmaking: The New Strategy of Negotiauctions. Norton, New York.

Thompson, L., 2011. The Mind and Heart of the Negotiator, fifth ed. Prentice-Hall, Upper Saddle River, NJ.

Tirole, J., 1988. The Theory of Industrial Organization. MIT Press, Cambridge, MA.

Walton, R., McKersie, R., 1965. A Behavioral Theory of Labor Negotiations. McGraw-Hill, New York.

Young, H.P. (Ed.), 1991. Negotiation Analysis. University of Michigan Press, Ann Arbor, MI.

Zeckhauser, R., Keeney, R., Sebenius, J.K. (Eds.), 1996. Wise Decisions. Harvard Business School Press, Cambridge, MA.

Neighborhood

Naomi Carmon and Efrat Eizenberg, Technion – Israel Institute of Technology, Haifa, Israel

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by N. Carmon, Volume 15, pp. 10488–10494, © 2001, Elsevier Ltd.

Abstract

The article provides a brief summary of selected publications about neighborhood/s by scholars in the social and behavioral sciences as well as by urban planners and architects. Neighborhoods have been part of the human mode of life since the dawn of civilization. In spite of the current decreasing importance of distance, following the accelerated development of telecommunication, it seems that neighborhoods will continue to exist in the next decades and fulfill important social functions for many persons in at least some of their time on earth.

Introduction

The term neighborhood is rooted in the verb neighbor. 'To neighbor' has a double meaning: (1) to live in the vicinity of one another and (2) to be friendly, to render mutual favors or assistance (Webster, 1983). This double meaning expresses the essence of the neighborhood: continuous physical proximity among people together with some social attitudes, such as friendliness, and/or special behaviors, such as mutual assistance. A neighborhood is always a part of a town or city, a part that is defined through an agreement among its residents, regular users, and others in its vicinity regarding its borders and its special physical and social characteristics. Frequently, neighborhoods are not officially defined and not everybody agrees about the size and borders of a neighborhood or its special identity.

This article is a brief summary of scientific and professional writings of the twentieth century about neighborhood/s by scholars in the social and behavioral sciences and by planners and architects. Its last section is devoted to contemporary discussions that shape the research of neighborhoods at the beginning of the twenty-first century.

The Neighborhood Concept in the Social Sciences

Neighborhoods have been considered by many scholars as a basic pattern of social organization, which is universal and timeless. Cooley (1909) argued that the neighborhood is 'a natural phenomenon' that exists wherever a group of people produce a dwelling place. Archeologists identified clusters of living that may be considered neighborhoods in kingdoms of the ancient world. The ancient neighborhoods were arranged by family or clan lines and later on, by religious, ethnic, or professional lines. In modern times, the divisions are largely by cultural–ethnic origin and/or socioeconomic status.

Neighborly relations among people who live in each other's vicinity, including intimate acquaintance and numerous interactions, were the normal way of life in rural communities prior to the industrial era (Tonnies, 1887). Sociologists of the early twentieth century, especially those from the Chicago school, claimed that such *primary relations* among neighbors are essential for maintaining the social order, because they create an effective informal control system (Cooley, 1909; Park et al., 1925). According to this view, isolation from a neighborhood portends an individual's alienation, while disintegration of neighborhoods threatens the social order (Durkheim, 1893).

The socioeconomic process of urbanization, which followed the technological-economic process of industrialization, disrupted social cohesion in rural communities and entailed the creation of distressed neighborhoods in the industrial cities. Among the first to describe them were Fredric Engels (1844), who demarcated the physical structures and daily life in Manchester and Salford in the United Kingdom, and Jacobs Riis (1890), who illustrated depiction of New York City's slum neighborhoods. The urban way of life was considered by many writers as a source of evil. Wirth (1938) argued that the size, density, and heterogeneity that characterize cities create changes that reduce primary relationships, including neighborly relationships, and increase loneliness and social deviance. Wirth represents the nostalgic yearning for healthy village life and the antiurban tone that was prevalent among writers of his generation.

A kind of 'neutral' description of neighborhoods was presented in the first half of the twentieth century by the Chicago 'ecological school,' headed by Robert Park, Ernest Burges, and R.D. McKenzie. They applied concepts and principles derived from plant and animal ecology to the analysis of human communities (Park et al., 1925). Empirical, sociological, and anthropological studies of urban neighborhoods in the middle of that century portrayed a more diverse urban picture (Glass, 1948; Janowitz, 1952). Within large industrial cities the researchers found residential areas that were *Gemeinschaft*s by Tonnies's typology; the intimate spontaneous relationships between the neighbors in such areas were based on proximity as well as on kinship, ethnic identity, and/or sharing a work place or other common denominator. Gans (1962) provided a vivid analysis of the life of Italian-Americans in a Bostonian neighborhood; he described these people, who managed to preserve their special family life, peer groups, and communal institutions within the modern city of Boston, as 'urban villagers.' Jacobs (1961) claimed that the traditional neighborhood such as the one described by Gans has lost its meaning and function in the modern big cities, where the lively street, the quarter (~100 000 residents), and the city as a whole are the relevant units. Suttle (1972) argued that in the modern city, one can find various different types of

neighborhoods, from the ones that are based on a form of organization that is almost a primary group to expanded communities of 'limited liability.'

A special stream of research was devoted to the impact of architectural design on social attitudes and behavior in the neighborhood context. Researchers frequently found strong relationships between architectural design and social variables: proximity and planned paths seemed to influence close friendship (Festinger et al., 1950), while large project size and height of buildings were considered as causes of increased delinquency and crime (Newman, 1972). Later critics, who named this line of reasoning 'physical determinism,' claimed that these studies failed to control for important intermediate variables, especially those related to the sociocultural and the socioeconomic characteristics of the residents. Keller (1968) urged architects to tailor their design by the variety of socioeconomic characteristics and cultural values of the users.

In response to the unidirectional view on the influence of the built environment on people, the school of man–environment research, consisting mainly of psychologists who cooperate with architects, emphasizes a two-way man–environment interaction. Their ecological approach to human interaction centered on understanding individuals in the context of their surroundings (Barker, 1968) and allows for personal-subjective definitions of neighborhoods. According to this perspective, the neighborhood is considered as one of the contextual ecological systems that a person is part of, including the nuclear family, peer group, and workplace (Bronfenbrenner, 1979).

Thus, neighborhoods have been defined through consolidating both social and physical spaces, i.e., when physical boundaries, social networks, local facilities, and special symbolic and emotional connotations are congruent in people's mind (Rappoport, 1977). According to this approach, people relations with their neighborhood coevolve through a process in which a space is transformed into a 'place' – a setting with a specific meaning and value (Tuan, 1974). Emotions and other psychological and social mechanisms (i.e., belonging and identification, sense of community and social capital) play a role in the development of a neighborhood and at the same time influence the lives of individuals; one is constantly affected by the other. For example, physical conditions and opportunities for social interaction in the neighborhood affect personal attitudes such as satisfaction and identification with the neighborhood, which, in turn, play a role in residents' willingness to be involved and improve their neighborhood (Ahlbrandt, 1984).

In addition to their social and psychological roles, urban neighborhoods often function as administrative units for various ends, including political organization and provision of social and commercial services. For census purposes, residential areas are divided into statistical units, the borders of which are frequently drawn by the boundaries of traditional neighborhoods. Social scientists have made intensive use of this statistical division, especially for developing empirical social area analysis that concentrates on changes over time in residential areas (Shevky and Bell, 1955; Abu-Lughod, 1969).

Neighborhood change was a focus of interest of social scientists throughout the twentieth century. The Chicago school invented the invasion–succession model, inspired by the world of plant and animals; 'invasion' illustrates the entrance of a different social group – racially, ethnically, or socioeconomically different – to a populated area, and the creation of competition and conflict within the invaded neighborhood that ends either by withdrawal of the invaders or by their triumph in the form of 'succession' (Park et al., 1925). The study of neighborhood change flourished in the second half of the twentieth century. Taeuber and Taeuber (1965) produced a comprehensive study of racial neighborhood change in American cities. Several economists added a considerable amount of supporting empirical evidence, especially for explaining neighborhood deterioration (Grigsby et al., 1987). A review of the neighborhood change literature of the time was provided by Schwirian (1983).

Another model of neighborhood change – *the life cycle model* – was suggested by Hoover and Vernon (1959), who identified five stages: development, transition, downgrading, thinning out, and renewal. These stages are related to the quantity and status of the residents, the intensity and kind of land use, and the quality of housing. Not all neighborhoods go through each stage; some may skip one or stay indefinitely in another.

The *political economy school* offers a different approach. They argue that the fate of a neighborhood is determined by the interests and the complex relationships among the political and economic public institutions and the strong actors in the business market (banks, developers, and others). Most of these forces play from outside the borders of the neighborhood, and hence, the ability of local people to influence them is very limited (Molotch, 1976).

Neighborhoods in the Eyes of Planners and Architects

Modern urban planning took its shape in the last quarter of the nineteenth century on the background of the socioeconomic movements of industrialization and urbanization. Large populations of workers and their families lived in extremely poor and crowded conditions in the industrial cities. The first planned neighborhoods in Europe were workers' neighborhoods, built by owners of industrial plants or by philanthropic organizations.

All the early city planning initiatives – The City Beautiful, The City Efficient, and The Garden City – sought to make the city and its neighborhoods safe, sanitary, economically efficient, and socially attractive. The leading idea was to use physical planning in order to create a socially better world. This central motive accompanies the efforts of architects and planners throughout the twentieth century. Several critics say that this is clear evidence that good intentions may sometimes lead to hell.

A prominent figure from the onset was Ebenzer Howard; his small-size book, *Garden City of Tomorrow* (1902), influenced urban planning in the twentieth century more than any other publication. He advocated an urban environment that combines the advantages of the city, such as modern occupations and services, with those of the village – human scale and many local social contacts. His model covered the entire scale from the dwelling through the neighborhood to the region. Howard's ideas had a decisive

impact on the New Towns Movement, which gave rise in the middle of the twentieth century to 30 new towns in Great Britain, a similar number in Israel, several *villes nouvelles* around Paris, others around Stockholm, as well as a few new towns in the United States. Within each of these, some version of Howard's plan for the 'ward' (his version of a neighborhood) was implemented. However, none of the many towns and neighborhoods that followed Howard's physical suggestions (at least partly) followed his social vision. As Hall (1988: p. 87) put it: "They see him as a physical planner, ignoring the fact that his garden cities were merely the vehicles for a progressive reconstruction of capitalist society into infinity of co-operative commonwealths."

In the United States, a rather similar (physically) model for a neighborhood unit was proposed by Clarence Perry (1929). He suggested a design formula with specific guidelines for spatial distribution of residences, community services, roads, open spaces, and businesses; he believed in using physical design tools for achieving a social vision of neighborliness and community building. A strong supporter of the concept of the neighborhood unit was the journalist–sociologist Louis Mumford (1954). He saw neighborhood planning as an opportunity and leverage for fostering feelings of belonging and for supporting amiable behavior among neighbors. He contributed to the idealization of the neighborhood as a framework for preserving positive social values and institutions in the modern city. This idealization was among the factors that convinced the American Public Health Association to adopt the neighborhood unit model as the basis for planning residential environments. Subsequently, it was adopted by other professional organizations and public agencies and guided much of the post-World War II construction of middle-class suburbs. All this happened in spite of the severe criticism of the model that had been prevalent since the 1940s. Social scientists disagreed with its premises and planners questioned its unintended consequences (Banerjee and Baer, 1984). Even though theoretically the model called for social mix and social integration, in practice, it encouraged segregation. An example is the early settlement in the neighborhoods of the new town of Radburn (New Jersey, US); in 1934, there were no blue-collar workers among the residents, and the realtors kept out Jews and Blacks as well (Schaffer, 1982).

What was planned and built in the middle of the twentieth century in Europe was highly influenced by the architects of *Les Congres Internationaux d'Architecture Modern*, an organization that was established in 1928 and lived for 30 years (Kallus and Law-Yone, 1997). These architects were deeply influenced by rigid socialistic concepts on the one hand and by new technical options that opened up through the industrial advancement on the other. They created the international style of architecture with the principles of minimalism (planning a house by the minimal needs of health and convenience of the residents), standardization (disregarding differences of culture, climate, etc.), and separation of land uses. Le Corbusier was the most famous architect who promoted these concepts. Thousands of neighborhoods, which were built in Europe following World War II, were influenced by these principles. However, many of their inhabitants disliked the minimal and standard blocks that the architects built for them; better-off households left them as soon as they could afford it. The unattractive design, together with severe maintenance problems, played a central role in the processes of social and physical deterioration and segregation, which characterized large residential areas of the post-World War II residential construction in European and other countries (see Ch. 7 in Hall, 1988).

An architectural school that opposes the standardization of international architecture as well as the typical design of neighborhoods-suburbs in the United States is the *New Urbanism*. It developed in the 1980s, as an umbrella term that encompassed two design schemes: the traditional neighborhood development, actually, back to the neighborhood unit concepts, and the transit-oriented development. The common principles of these two are small (compact) size that creates walkable distances (typically no more than a quarter of a mile from center to edge), clearly defined edges and centers, mixed land uses (residences, shops, schools, and workplaces) and mixed 'market segments' and priority to public buildings and spaces (Katz, 1994). In some magic way, this design is expected to create not only a distinct physical unit but also social identity and community. Underneath the detailed design instructions of the New Urbanism lies the same old belief of architects that by means of proper design of the physical environment one can create a better social world.

In addition to a common belief in *physical determinism*, most of the neighborhood models suggested by architects are characterized by disregard of the users' voice. For Le Corbusier this was obvious; he argued that planning is too important to let ignorant people be involved in it. Other architects continued the paternalistic approach without considering issues of residents' participation. One of the exceptions was the English architect John Turner (1976), the father of 'housing as a verb' rather than a noun. He argued that when dwellers control the major decisions concerning the design, construction, and management of their home and neighborhood, both the process and the environment produced stimulate individual and social well-being. Turner based his approach on his experience with communities in which there was mutual commitment and partnership between the members, mostly traditional communities. Is consideration of users' point of view in designing and managing the built environment applicable also to regular neighborhoods in urban Western societies?

A positive answer to this question evolved within a new breed of planners. In the middle of the twentieth century, social scientists – sociologists, economists, human geographers, and lawyers with social sensitivity – transformed the discipline/profession of urban/city planning. The sociologist Herbert Gans (1968) played a central role in this process of change; years later, he phrased an important part of the transformation by saying that "America's major urban problems were poverty, scarcity of decent jobs, racism … more serious and wide-ranging than the land-use-related problems the city planners of the time were dealing with" (Gans, 1991: p. IX). Scholars and practitioners of urban planning developed the ideas and practices of citizen participation in planning and community development (Arnstein, 1969; Bratt and Rearden, 2013), of Advocacy Planning (Davidoff, 1965), Progressive Planning (Angotti, 1993), and Equity Planning (Krumholz and Forester,

1990; Carmon and Fainstein, 2013). All these frameworks of thought and action devoted much of their work to urban neighborhoods of minorities and low-income households.

The different approaches of planners who had only architectural education vis-à-vis those who were widely exposed to the social sciences were clearly expressed in the development of programs for rehabilitating distressed neighborhoods (Carmon, 1990). Three successive 'generations' of urban/neighborhood renewal policies were identified and analyzed by Carmon (1999), policies that were implemented at about the same time in Western countries on both sides of the Atlantic. The first generation – the era of the bulldozer, which was dominated by architects, by ideas of physical determinism and activities of slum clearance—started in the United Kingdom in 1930, in the United States in 1937 (see also the law of 1949), in Canada in 1948, and in several European countries in the 1950s. The slum clearance projects, in spite of big differences among them, usually generated the same kind of criticism for hurting individuals and households, eliminating healthy communities, and, in general, bringing about more negative than positive human and urban consequences (Wilmott and Young, 1957; Belash and Hausknecht, 1967; Gibson and Langstaff, 1982). In the second generation of policies, when urban planners with education of social sciences became influential, large-scale governmental programs for both social and physical neighborhood rehabilitation were adopted and citizen participation in planning became common; examples are Model Cities in the United States (mid-1960s), Neighborhood Improvement Program in Canada (started in 1973), Neighborhood Social Development in France (started in 1981), and Project Renewal in Israel (started in 1977). Despite the abundance of goodwill and large sums of money, the results were at best mixed: where implemented, these combinations of social and physical programs frequently benefited people (local residents) but not their neighborhoods, i.e., they could not change the negative image and the low status of the target neighborhoods. The third generation is related to the rise of the neoliberal ideology from 1970s to 1990s and emphasizes economic considerations in neighborhood and urban regeneration. Public–private initiatives of planned gentrification that encouraged better-off households to enter low-income neighborhoods, usually in the city center, became popular in many large and small cities. Many of those have physically improved parts of old cities and neighborhoods, but were severely criticized for causing direct and indirect displacement of incumbent residents (Marcuse, 1985; Smith, 2002).

Neighborhood Research on the Brink of the Twenty-First Century

We began this article by saying that neighborhoods have been part of the human mode of life since the dawn of civilization. An ancient verse – "better is a neighbor that is near than a brother far off" (Proverbs 27:10) – illustrates the important social functions that neighborhoods and neighbors had in traditional societies. With the development of communication technologies, people became much more mobile, their world widened considerably, and they became less dependent on physical proximity and on constant relationships. The decreasing importance of distance has been enhanced with the accelerated development of telecommunication in recent years. Do these changes mean that the time of neighborhoods as territories that combine physical and social closeness is over?

The answer is that neighborhoods will continue to exist in the next decades and fulfill important social functions for many persons in at least some of their time on earth. In general, those who are attached to their neighborhoods are people on the ends, on both ends of the life cycle – children and elderly, and on both ends of the social ladder – disadvantaged and well-to-do households. People whose mobility is limited, either because modern communication technologies have not reached their environment, or because they are too young, too old, or too poor to use the technologies, will continue to be dependent on their neighbors and neighborhoods as sources for fulfilling functional as well as social and emotional needs. In addition, research findings show that housing proximity is associated with the existence of important joint interests. This is very clear among minorities and immigrant groups, which tend to cluster geographically and create *urban enclaves*; in their enclaves, they can support each other; defend and develop their specific social, economic, and sometimes political interests; and when they feel necessary, can organize to fight contradicting interests (Abrahamson, 1996). This is also common among exclusive neighborhoods of affluent citizens who create gated communities (Blakely and Snyder, 1997; Atkinson and Blandy, 2005). Mobile middle-class households may not be inclined to be part of neighborhoods-communities, but where and when they feel a need to promote a local interest, and, especially, when some changes in the social or physical environment is perceived as a threat to the value of their home, they too function as a community of interests on a local basis.

Thus, in recent years, research into 'neighborhood effects' has become common; its point of departure is frequently an assertion that neighborhood characteristics play a significant role in shaping opportunities, quality of life, and social norms of their residents, especially immigrants and low-income residents. The current renewed interest in these effects is traced back to the book *The Truly Disadvantaged* (Wilson, 1987). Since its publication, an explosion of work has emerged testing the hypothesis that living in a neighborhood of concentrated poverty has negative effects on economic self-sufficiency, violence, drug use, low birth weight, cognitive ability, and more. Especially salient are the many studies devoted to evaluating the *Moving to Opportunity* housing mobility experiment (Briggs et al., 2010). In general, the findings have been disappointing; the move to better neighborhoods has not been found as having considerable positive effect on self-sufficiency or on physical health; yet, many claim that the failure was built into the implementation of the experiment and the research design (Sampson, 2008). Disappointing results regarding impacts on labor market outcomes were reported also from Britain (van Ham and Manley, 2010). However, research of other kinds of 'effects' did find positive ones: health researchers have found significant positive impacts of neighborhood characteristics on individual

physical and mental health (Diez Roux, 2004), especially among children and adolescents (Leventhal and Brooks-Gunn, 2000). Kim and Kaplan (2004) showed how neighborhood design features, new urbanism style, have direct and indirect effects on residents' sense of community. When the physical design of a neighborhood (Bothwell et al., 1998) and/or participation in neighborhood activity lead to development of social capital, especially 'bridging' social capital, the range of opportunities for disadvantaged residents is increased (Gittell and Vidal, 1998).

In the context of a globalizing world, researchers identified the two contrasting yet complementary components of the global and the glocal. Relevant examples are global marketing of neighborhoods to attract capital investment and residents of the creative class (Mele, 2000), and building bottom–up organizations to counter the top–down controls of global bodies. In general, it seems that the perceived importance of the local is gathering force in parallel with the global trend. Neighborhoods are sites for engagement of residents in the practice of planning their environment (Innes and Booher, 2004) and for establishment of active citizenship (Brannan et al., 2006) aiming at democratizing social relations and promoting social and environmental justice. In Europe, this orientation is implemented in local bodies through the 'new localism' (Lowndes and Sullivan, 2008), while the Community Development Corporations are continuing their activity across the country in the United States (Bratt, 2006), and the Local Agenda 21 is spreading around the world (RIO+20, 2012). The outburst of protests in 2011, protests that occurred at about the same time in close to a thousand cities in 82 countries (Rogers, 2011), is a demonstration of the current new relationships between local communities and global developments.

Because the local context of a neighborhood continues to play an important role in the life of many individuals and groups, social scientists will continue to be engaged with understanding neighborhoods in the foreseeable future. Following Herbert Gans (2002), the authors suggest focusing on causal relations between space/place and society: the few but important ways in which a place – a neighborhood in this case – influences persons and collectivities, and the many ways in which collectivities turn places into social spaces and shape their uses.

See also: Applied Sociology; Cities: Internal Structure; Environmental Determinism; Gentrification; New Urbanism; Planning Theory: The Good City and Its Governance; Poverty, Geography of; Social Capital; Social Geography; The Right to the City; Urban Geography; Urban Poverty in Neighborhoods.

Bibliography

Abrahamson, M., 1996. Urban Enclaves: Identity and Place in America. St. Martin's Press, New York.
Abu-Lughod, J., 1969. Testing the theory of social area analysis: the ecology of Cairo, Egypt. American Sociological Review 34 (2), 198–212.
Ahlbrandt, R., 1984. Neighborhoods, People and Community. Plenum Press, New York.
Arnstein, S.R., 1969. A ladder of citizen participation. Journal of the American Institute of Planners 35 (4), 216–224.
Angotti, T., 1993. Metropolis 2000: Planning, Poverty and Politics. Routledge, London and New York.
Atkinson, R., Blandy, S., 2005. Introduction: international perspective on the new enclavism and the rise of gated communities. Housing Studies 20 (2), 177–186 (special issue on Gated Communities).
Banerjee, T., Baer, W.C., 1984. Beyond the Neighborhood Unit. Plenum Press, New York.
Barker, R.G., 1968. Ecological Psychology: Concepts and Methods for Studying the Environment of Human Behavior. Stanford University Press, Stanford, CA.
Belash, J., Hausknecht, M. (Eds.), 1967. Urban Renewal: People, Politics and Planning. Doubleday, Anchor Books, New York.
Blakely, E.J., Snyder, M.G., 1997. Fortress America: Gated Communities in the US. Brookings Institute Press, Washington, DC.
Bothwell, S.E., Gindorz, R., Lang, R.E., 1998. Restoring community through traditional neighborhood design: a case study of Diggs Town public housing. Housing Policy Debate 9, 89–114.
Brannan, T., John, P., Stoker, G., 2006. Active citizenship and effective public services and programs: how can we know what really works? Urban Studies 43, 993–1008.
Bratt, R., Rearden, K., 2013. Beyond the ladder: new ideas about resident roles in contemporary community development in the US. In: Carmon, N., Fainstein, S. (Eds), Policy, Planning and People: Promoting Justice in Urban Development. University of Pennsylvania Press, Philadelphia, PA.
Bratt, R.G., 2006. Community development corporations: challenges in supporting the right to housing. In: Bratt, R.G., Stone, M., Hartman, C. (Eds.), The Right to Housing. Temple University Press, Philadelphia, PA, pp. 340–359.
Briggs, X.S., Popkin, S.J., Goering, J., 2010. Moving to Opportunity: The Story of an American Experiment to Fight Ghetto Poverty. Oxford University Press, Oxford, NY.
Bronfenbrenner, U., 1979. The Ecology of Human Development. Harvard University Press, Cambridge, MA.
Carmon, N. (Ed.), 1990. Neighborhood Policy and Programmes: Past and Present. Macmillan, London.
Carmon, N., 1999. Three generations of neighborhood remedies: research-based analysis and policy implications. Geoform 30 (2), 145–158.
Carmon, N., Fainstein, S. (Eds), 2013. Policy, Planning and People: Promoting Justice in Urban Development. University of Pennsylvania Press, Philadelphia, PA.
Cooley, C.H., 1909. Social Organization. Charles Scribner's Sons, New York.
Davidoff, P., 1965. Advocacy and pluralism in planning. Journal of the American Institute of Planners 31 (4), 331–338.
Diez Roux, A.V., 2004. Estimating neighborhood health effects: the challenges of causal inference in a complex world. Social Science and Medicine 58 (10), 1953–1960.
Durkheim, E., 1893/1964. The Division of Labor in Society. Free Press, New York.
Engels, F., 1844/2009. The Condition of the Working Class in England. Penguin Books, New York.
Festinger, L., Schachter, S., Back, K., 1950. Social Pressure in Informal Groups: A Study of Human Factors in Housing. Harper, New York.
Gans, H.J., 1962. The Urban Villagers: Group and Class in the Life of Italian Americans. Free Press, New York.
Gans, H.J., 1968. People and Plans: Essays on Urban Problems and Solutions. Basic Books, New York.
Gans, H.J., 1991. People, Plans, and Policies: Essays on Poverty, Racism, and Other National Urban Problems. Columbia University Press, New York.
Gans, H.J., 2002. The sociology of space: a use-centered view. City and Community 1 (4), 329–339.
Gibson, M.S., Langstaff, M.J., 1982. An Introduction to Urban Renewal. Hutchinson, London.
Gittell, R., Vidal, A., 1998. Community Organizing: Building Social Capital as a Development Strategy. Sage Publication, Thousand Oaks, CA.
Glass, R., 1948. The Social Background of a Plan: A Study of Middlesborough. Routledge & Kegan Paul, London.
Grigsby, W., Baratz, M., Galster, G., Maclennan, D., 1987. The dynamics of neighborhood change and decline. Progress in Planning 28, 1–76.
Hall, P., 1988. Cities of Tomorrow. Basil Blackwell, Oxford, UK.
Hoover, E.M., Vernon, R., 1959. Anatomy of a Metropolis. Harvard University Press, Cambridge, MA.
Howard, E., 1902/1985. Garden Cities of Tomorrow. Attic Books, London.
Innes, J.E., Booher, D.E., 2004. Reframing public participation: strategies for the 21st century. Planning Theory and Practice 5 (4), 419–436.
Jacobs, J., 1961. The Death and Life of Great American Cities. Random House, New York.

Janowitz, M., 1952. The Community Press in an Urban Setting. Free Press, Glencoe, IL.
Kallus, R., Law-Yone, H., 1997. Neighborhood – the metamorphosis of an idea. Journal of Architectural and Planning Research 14, 107–124.
Katz, P., 1994. The New Urbanism: Towards and Architecture of Community. McGraw-Hill, New York.
Keller, S., 1968. The Urban Neighborhood: A Sociological Perspective. Random House, New York.
Kim, J., Kaplan, R., 2004. Physical and psychological factors in sense of community: new urbanist Kentlands and nearby Orchard Village. Environment and Behavior 36 (3), 313–340.
Krumholz, N., Forester, J., 1990. Making Equity Planning Work: Leadership in the Public Sector. Temple University Press, Philadelphia, PA.
Leventhal, T., Brooks-Gunn, J., 2000. The neighborhoods they live in: the effects of neighborhood residence on child and adolescent outcomes. Psychological Bulletin 126 (2), 309–337.
Lowndes, V., Sullivan, H., 2008. How low can you go? Rationales and challenges for neighborhood governance. Public Administration 86 (1), 53–74.
Marcuse, P., 1985. Gentrification, abandonment and displacement: connections, causes and policy responses in New York. Journal of Urban and Contemporary Law 195, 195–240.
Mele, C., 2000. Selling the Lower East Side: Culture, Real Estate, and Resistance in New York City. University of Minnesota Press, Minneapolis, MN.
Molotch, H., 1976. The city as a growth machine: toward a political economy of place. American Journal of Sociology 82 (2), 309–332.
Mumford, L., 1954. The neighborhood and the neighborhood unit. Town Planning Review 24, 250–270.
Newman, O., 1972. Defensible Space: Crime Prevention Through Urban Design. Collier Books, New York.
Park, R.E., Burgess, E.W., McKenzie, R., 1925/1984. The City: Suggestions for the Study of Human Nature in the Urban Environment. University of Chicago Press, Chicago, IL.
Perry, C.A., 1929. The Neighborhood Unit: A Scheme of Arrangement for the Family Life Community. Regional Plan of New York and Its Environs, New York.
Rappoport, A., 1977. Human Aspect of Urban Form: Towards a Man-Environment Approach in Urban Form and Design. Pergamon Press, Oxford, NY.
Riis, A. Jacob, 1890/1997. How the Other Half Lives. Penguin Books, New York.
RIO+20, 2012. United Nation Conference on Sustainable Development. http://www.uncsd2012.org/.
Rogers, S., 2011. Occupy Protests Around the World. The Guardian, 14 Nov. http://www.guardian.co.uk/news/datablog/2011/oct/17/occupy-protests-world-list-map.
Sampson, R.J., 2008. Moving to inequality: neighborhood effects and experiments meet social structure. American Journal of Sociology 114 (1), 189–231.
Schaffer, D., 1982. Garden Cities for America: The Radburn Experience. Temple University Press, Philadelphia, PA.
Schwirian, K.P., 1983. Models of neighborhoods. Annual Review of Sociology 9, 83–102.
Shevky, E., Bell, W., 1955. Social Area Analysis; Theory, Illustrative Application and Computational Procedures. Stanford University Press, Stanford, CA.
Smith, N., 2002. New globalism, new urbanism: gentrification as global urban strategy. Antipode 34 (3), 427–450.
Suttle, G., 1972. The Social Construction of Communities. University of Chicago Press, Chicago, IL.
Taeuber, K.E., Taeuber, A.R., 1965. Negroes in Cities. Aldine, Chicago, IL.
Tonnies, F., 1887/1957. Translated from German. In: Loomis, C.P. (Ed.), Community and Society (Gemeinschaft und Gesellschaft). Michigan State University Press, East Lansing, MI.
Tuan, Y.F., 1974. Space and place: humanistic perspective. In: Gale, S., Ollson, G. (Eds.), Philosophy in Geography. Springer, New York, pp. 387–427.
Turner, J.F.C., 1976. Housing by People: Towards Autonomy in Building Environments. Pantheon Books, New York.
van Ham, M., Manley, D., 2010. The effect of neighborhood housing tenure mix on labour market outcomes: a longitudinal investigation of neighborhood effects. Journal of Economic Geography 10 (2), 257–282.
Webster, 1983. Webster's New Universal Unabridged Dictionary, second ed. Dorset and Baber, Simon and Schuster, Cleveland, OH.
Wilmott, P., Young, M., 1957. Family and Kinship in East London. Routledge and Kegan Paul, London.
Wilson, W.J., 1987. The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy. University of Chicago Press, Chicago, IL.
Wirth, L., 1938. Urbanism as a way of life. American Journal of Sociology 44, 1–24.

Neighborhood Effects

Jacob W Faber and Patrick Sharkey, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social scientists have long been concerned with the role of space in systems of stratification. While scholars in the field of 'neighborhood effects' have typically focused on how a community affects the life chances of its residents, we argue for a broader view of neighborhood effects that considers how spatial stratification serves to maintain and reproduce inequality across multiple dimensions. This article outlines major theoretical arguments exploring how local residential contexts affect social and economic outcomes at the level of individuals and communities, drawing attention to the empirical challenges to measuring neighborhood effects.

Introduction

The term 'neighborhood effects' has come to be associated primarily with research that considers how characteristics of communities affect the life chances of individuals within them. This strand of research represents only a narrow view of the role that residential settings play in systems of stratification, however. A long-standing tradition of research begins with the idea that systems of stratification are organized, in part, along spatial lines. The neighborhood, from this perspective, is a core dimension of inequality. Building on this assumption, we argue for a broader view of the neighborhood effects literature, one that includes all research attempting to move toward a more complete understanding of how spatial stratification serves to maintain and reproduce inequality across multiple dimensions.

This view of neighborhood effects research is difficult to summarize, and we do not attempt an exhaustive review. Instead, we describe some of the most common approaches to studying neighborhood effects, and we offer examples of research supporting our more expansive perspective. We begin with a focus on theoretical conceptions about the processes through which neighborhood inequality affects other dimensions of inequality. Next, we consider the challenges in moving from theory to empirical evidence on neighborhood effects, and we explore some of the evidence base pertaining to neighborhood effects on several domains of economic and social life. We conclude with a discussion of where advancements in data collection, methods, and theory may bring the field.

Theory and Mechanisms

The resurgent interest in the role of neighborhoods as a core dimension of stratification often is traced back to the publication of William Julius Wilson's *The Truly Disadvantaged* (1987). In this book, Wilson laid out a theory to explain the transformation of urban poverty in the post-Civil Rights era, and argued that the new concentration of poverty in central city neighborhoods helped explain changes in the prevalence of other social outcomes such as joblessness, teenage parenthood, welfare dependence, and violent crime (see also Sampson and Wilson, 1995). Wilson's theory is complemented by the research of Douglas Massey and several colleagues, who focused attention on the importance of racial segregation as the mechanism by which changes in economic conditions translate into concentrated poverty at the community level (Massey and Denton, 1993).

What is most notable about these arguments is their focus on the spatial organization of inequality and the consequences for urban phenomena and social outcomes at the community level. In most reviews of the literature, the ecological approach to neighborhood effects research receives less attention than research focusing on individual outcomes. However, Sampson (2003, 2006, 2012; Sampson et al., 2002) points to a long tradition of research on community organization, health, and violence in arguing that the ecological approach aligns most closely with classic sociological research on spatial stratification (for example, see Bursik and Grasmick, 1993; Shaw and McKay, 1942). Building on this tradition, Sampson et al. (1997) argued that the relationship between concentrated poverty and violence is mediated by community-level social processes arising from social cohesion, trust, and community organization (see also Sampson et al., 2002). This research is one example of a large body of work that considers the spatial organization of social, economic, and political phenomena and the consequences for community-level outcomes related to health, crime, violence, and well-being (Massey and Fischer, 2003; Sampson et al., 2002).

A larger set of theoretical mechanisms have been proposed as plausible ways in which neighborhoods may be linked to individual social and economic outcomes. In an early review of the empirical literature, Jencks and Mayer (1990) outlined several sets of mechanisms by which individuals' residential contexts (including both neighborhoods and schools) might affect resident outcomes in both positive and negative ways. *Epidemic models* focus on the ways that peers influence one another, *collective socialization models* focus on adult influence on children other than their own, and *institutional models* focus on how children's outcomes may be affected by neighborhood institutions. *Relative deprivation models* focus on how people see themselves in relation to people within and outside the community, *cultural conflict models* focus on the ways that groups create a common culture to deal with collective failure, and *competition models* posit conflict over scarce resources make affluent neighbors a liability.

Since the influential article by Jencks and Mayer (1990), other reviews of the literature have added nuance to the question of how growing up in a particular place might matter for individuals. Several of these articles argued that aspects of local context may have different effects across life stages. Ellen and Turner (1997), for example, suggested that infants and preschool children are most directly affected through parents, whereas schools become increasingly relevant for elementary school-aged children. Leventhal and Brooks-Gunn (2000) argued that environmental pollution may have a stronger effect on an individual in earlier stages of development. Ellen and Turner (1997) and Small and Newman (2001) argued that social life within the residential environment may have the greatest influence on the lives of adolescents, who are at risk of becoming involved in risky behaviors or activities through their peers and role models in the community. For adults, neighborhood effects may operate through the prevalence of services, information, social networks, and employment opportunities (Ellen and Turner, 1997; Small and Newman, 2001).

As the empirical literature has evolved and expanded, so too have theoretical models of mechanisms by which the local residential context could potentially shape individual outcomes. Recent contributions argue that neighborhood effects should be viewed as an interaction between the duration of exposure to neighborhood characteristics and the vulnerability of the individual exposed (Galster, 2012; Harding et al., 2011; Small and Feldman, 2012). Incorporating these ideas into empirical studies of neighborhood effects allows for a distinction between an individual who, for example, spent her entire childhood in a high-poverty neighborhood and another who only experienced a brief period of exposure. Similarly, this more refined approach suggests that neighborhoods should be operationalized at different scales depending on the specific exposure or process under study (Galster, 2008; Logan, 2012; Sampson et al., 2002). Whereas the school catchment area may be most relevant for understanding children's academic success (Schwartz, 2010), shocks in the demand for labor over a wider geographic area may be more relevant for understanding the effects of the residential setting on economic outcomes (Ananat et al., 2011; Jargowsky, 1997).

From Theory to Evidence

There are several challenges to translating theories of neighborhood effects into persuasive evidence, but virtually every review of the empirical literature on neighborhood effects has focused primary attention on the problem of selection bias (Ellen and Turner, 1997; Jencks and Mayer, 1990; Leventhal and Brooks-Gunn, 2000; Ludwig et al., 2008; Sampson et al., 2002). The challenge of selection bias arises because families may end up in disadvantaged neighborhoods due to a range of unmeasured factors that are likely to be correlated with key outcomes of interest, and the magnitude and direction of this bias often are unclear (Ludwig et al., 2008).

Several approaches have been used to confront selection bias in the neighborhood effects literature. Sampson et al. (2008), for example, took advantage of rich data on family characteristics typically not measured in survey data to develop more refined models for selection into advantaged and disadvantaged neighborhoods in Chicago, and used these models to estimate the effect of exposure to concentrated disadvantage on cognitive development. Harding (2003) used a propensity score-matching approach to estimate the effect of neighborhood poverty on high school dropout and teenage pregnancy, combined with a sensitivity analysis to assess how robust estimates were in the presence of potential confounders (see also Sharkey and Elwert, 2011). Several studies have used family fixed effects methods or sibling and neighbor correlations to identify the effect of neighborhood conditions on educational and economic outcomes (e.g., Aaronson, 1998; Page and Solon, 2003; Plotnick and Hoffman, 1999; Vartanian and Buck, 2005). These approaches typically rely on variation in neighborhood conditions among siblings within the same families, and thus account for unobserved, stable characteristics of families that are shared by all siblings.

Whereas all of the empirical approaches discussed to this point rely on variation in neighborhood conditions derived from the residential choices of families, an alternative set of methods rely on variation derived from natural experiments, or controlled experiments, that lead families into different residential settings. As an example, Kirk (2009) leveraged displacement caused by Hurricane Katrina as a natural experiment that affected the geographic locations to which Louisiana state prisoners returned in the aftermath of the hurricane, and found that prisoners forced to move to a community other than that from which they originated had lower rates of subsequent incarceration. The more common approach is to exploit variation in neighborhood conditions arising from experimental or quasi-experimental residential mobility programs. The first example of this form of research came about in the aftermath of a court desegregation lawsuit. The Gautreaux Assisted Housing Program was the result of a settlement that required the Chicago Housing Authority to provide housing to eligible families in neighborhoods across the entire Chicago metropolitan area, including affluent and predominantly white suburbs. Early research found that families that moved outside Chicago's city limits experienced remarkable changes in adults' economic outcomes and children's educational attainment (Rubinowitz and Rosenbaum, 2000). Subsequent analyses of families in the Gautreaux program have documented problems with the original studies and with the overall research design; however, after adjusting for observed differences between families in the program, virtually all of the research that has been conducted on Gautreaux has found that families benefited substantially when they left the deeply segregated, high-poverty neighborhoods of Chicago and moved to more integrated, less-poor communities (Mendenhall et al., 2006; Votruba and Kling, 2009).

The Moving to Opportunity (MTO) Program was a federally funded social experiment implemented specifically to test whether moving into low-poverty neighborhoods affected the social and economic outcomes of families living in five cities (Baltimore, Boston, Chicago, Los Angeles, and New York) (Goering and Feins, 2003). In each city, families that volunteered for the program were randomized into one of three groups: an experimental group that received housing vouchers that could only be used to rent in low-poverty neighborhoods; a group that received standard Section 8

vouchers without requirements on where the voucher could be used; and a control group that received no voucher at all. Initial evaluations of MTO were mixed (Ellen and Turner, 2003; Ludwig et al., 2008; Sampson, 2008). Ten to fifteen years after random assignment, parents in the experimental group reported substantial improvements in mental and physical health and overall well-being, but no improvements in economic outcomes (Ludwig et al., 2011). Treatment effects on children varied substantially by gender. Girls in the experimental group showed improved mental health and were less likely to participate in some risky behaviors, whereas boys experienced an increase in some risky behaviors and no changes in other indicators of successful development (Sanbonmatsu et al., 2006).

Interpreting the results from MTO is challenging because the study was conducted at a unique historical moment when employment opportunities were expanding very quickly in central city neighborhoods, public housing projects were being torn down in many cities, and violent crime was dropping rapidly (Sharkey, 2013). The experimental treatment provided a short-term change in families' neighborhood environments that faded away for most families. For instance, 10–15 years after random assignment, families in the experimental group lived in neighborhoods with poverty rates just 3% points lower than families in the control group (Ludwig et al., 2012). This feature of the experiment is highly relevant to interpreting the results, as Turner et al. (2012) showed that families that remained in low-poverty neighborhoods for longer periods of time experienced meaningful improvements in outcomes across multiple domains. Further, the experiment played out very differently across the five sites where it was implemented (Briggs et al., 2010), with children in some cities experiencing substantial improvements in developmental outcomes and children in other cities showing null results (Burdick-Will et al., 2011).

Analyses of other quasi-experimental residential housing programs have generated results that are complex, with results that vary depending on where the program was conducted, the type of change in housing and neighborhood conditions induced by the program, and the outcome under study. In Montgomery County, MD, for example, Schwartz (2010) exploited the random assignment of public housing recipients to units that fall within a wide range of school catchment areas. Her analysis found that children falling within the boundaries of lower-poverty schools outperformed children assigned to housing in higher-poverty areas on math and reading tests. In Denver County, CO, low-income families were offered housing via the 'Dispersed Housing Program' when they reached the top of a waiting list. Cutsinger et al. (2011) argued that this process created exogenous variation in the units offered to families, and found that residence in neighborhoods with higher crime, lower homeownership rates, and lower levels of social capital were linked with lower rates of high school graduation. Ludwig et al. (2010) also use data from families on a wait list for subsidized housing in Chicago to estimate the effects of change in neighborhood conditions on standardized test performance. Families were randomly assigned a number on the wait list for housing, which created exogenous variation in the timing of when families were able to move into new apartments in lower-poverty neighborhoods than those from which they came. The researchers found that moving into lower-poverty neighborhoods resulted in substantively large improvements in children's performance on reading and math assessments. Carlson et al. (2012) also leveraged a waiting list process and used propensity scores to match Section 8 recipients in Wisconsin to similarly situated social service recipients who did not receive housing support. While those receiving vouchers moved to somewhat improved neighborhoods, this neighborhood change did not translate into improved economic circumstances on average – there was little effect on unemployment and a negative effect on earnings that faded over time. However, more vulnerable participants (i.e., those with less education and minorities) showed better labor market outcomes. Massey et al. (2013) analyzed the effects of neighborhood change resulting from the court-mandated construction of affordable housing in the relatively affluent Philadelphia suburb of Mt. Laurel. Using propensity scores to match program participants to those on the waiting list who were not offered housing, they found that those who moved had higher earnings and employment rates than those who did not, although there was no effect on welfare receipt. In contrast to the Wisconsin Section 8 study, movers with more education experienced greater economic benefits.

Although the problem of selection bias has received the greatest amount of attention in the literature, it is only one of many challenges to the identification of unbiased neighborhood effects. Disentangling the effects of the neighborhood setting from the effects of family characteristics is extremely difficult in observational studies because family circumstances and neighborhood conditions are reciprocally related over the life course and over generations of family members (Jencks and Mayer, 1990; Sharkey and Elwert, 2011). Identifying the specific dimension of the residential environment that is most salient for a given individual or ecological outcome is challenging because of the way that different dimensions of spatial advantage and disadvantage come bundled together in the neighborhood setting (Jencks and Mayer, 1990; Ludwig et al., 2008). Sampson (2008) argued that estimates of neighborhood effects from observational studies may be biased toward zero because of 'included variable bias,' which arises in studies where controls are added that may represent mechanisms by which neighborhoods affect individual outcomes. For example, if neighborhood conditions affect mental or physical health (Klebanov et al., 1994; Ludwig et al., 2011; Wheaton and Clarke, 2003), then estimating neighborhood effects while adjusting for parental health will lead to overly conservative, biased estimates.

Experimental studies of neighborhood effects are not immune to problems of identification (Sampson, 2008). Estimates derived from experimental residential mobility programs such as the MTO rely on the assumption that families in the treatment and control groups are isolated from each other and have no interactions that may 'contaminate' the estimated treatment effect of the program (Ludwig et al., 2011; Sampson, 2008; Sobel, 2006). Most empirical studies of neighborhood effects, whether observational or experimental, make the implicit assumption that the impact of the neighborhood is linear (Jencks and Mayer, 1990). This assumption is

questionable from both a theoretical standpoint and an empirical standpoint (Crane, 1991). For example, research exploiting variation across the five sites where the MTO program was implemented found that youths from the most intensely disadvantaged neighborhoods (i.e., Baltimore and Chicago) may have benefited most from the intervention, a finding that is suggestive of nonlinear neighborhood effects (Burdick-Will et al., 2011; see also Klebanov et al., 2014).

Various neighborhood risk factors may also work in concert with other neighborhood and individual characteristics to have the most detrimental effect on outcomes of interest, presenting additional empirical challenges (Evans and Kantrowitz, 2002; Jencks and Mayer, 1990). For example, numerous studies have documented substantial differences in neighborhood effects by gender. Women and girls face unique risks in high-poverty, violent neighborhoods (Popkin et al., 2008), which may make them more responsive to changes in the local residential setting (Clampet-Lundquist et al., 2011; Elliott, 2001; Kling et al., 2005; Miller, 2008; Sanbonmatsu et al., 2006; Zimmerman and Messner, 2010; Zuberi, 2012). The evidence is less conclusive regarding differences in neighborhood effects across economic strata (Ananat et al., 2011; Massey et al., 2013; Wodtke et al., 2012) or racial/ethnic groups (Crosnoe, 2009; Crowder and South, 2003; Sanbonmatsu et al., 2006; Turley, 2003). Similarly, while there is not a consistent pattern in the research on differences in neighborhood effects by age (Ananat et al., 2011; Sharkey et al., 2013; Wodtke et al., 2012), negative consequences may compound over the life course. For example, people born into disadvantaged communities not only start life at lower weight (Morenoff, 2003), but experience sharper, unhealthy increases in body mass index as young children (Klebanov et al., 2014).

Beyond the problem of identification, the neighborhood effects literature has struggled with challenges of measurement. The most common approach to measuring neighborhood characteristics is to define the neighborhood through administrative boundaries, typically the census tract, and to measure aggregate characteristics of the population within the boundary, such as the poverty rate (Sampson et al., 2002). Compositional characteristics of census tracts such as poverty, economic inequality, racial and ethnic composition, class composition, and unemployment are all derived from this approach to the operationalization of neighborhoods. Although this is the default approach to studying neighborhood effects, it is important to acknowledge that many theoretical models described in the previous section of this article do not align well with this approach to operationalizing the neighborhood (see, for instance, Burton et al., 1997; Durlauf, 2004; Rosenblatt and DeLuca, 2012).

Research on institutions such as schools, for example, has used school catchment areas as the relevant boundary for studying the link between school quality and academic achievement (Schwartz, 2010). Research exploring environmental pollution has used proximity to sources of air pollution or noise pollution to understand how these aspects of the environment influence individual or aggregate outcomes of children (Bronzaft and McCarthy, 1975; Cohen et al., 1980; Currie et al., 2009; Downey and Van Willigen, 2005; Evans, 2006; Evans and Maxwell, 1997; Hambrick-Dixon, 1985; Ransom and Pope, 1992; Stansfeld et al., 2005). Studies of economic shocks have measured changes in labor markets across wider spatial scales, such as counties or Native American reservations, to study the effect of such shocks on adult employment outcomes and children's academic outcomes (Ananat et al., 2011; Wolfe et al., 2012). Conversely, studies of the detrimental effects of lead (Lanphear et al., 2005; Reyes, 2012) may require an even finer geographic scale than the census block if exposure varies from building to building. These examples reflect a more flexible approach to the study of residential contexts in which the appropriate scale for the measurement of 'neighborhoods' is dependent on the specific theoretical mechanism under study (Galster, 2008).

Operationalizing individuals' exposure to the neighborhood represents an additional challenge. Although much of the empirical literature measures neighborhood conditions at a specific point in time or over a few years of an individual's life, empirical work on the rigidity of neighborhood inequality documents the stability of neighborhood conditions over the life course and across generations (Sharkey, 2008). A recent set of studies has found that the duration of exposure to the neighborhood environment is crucially important in understanding the effects of neighborhood disadvantage on educational attainment, cognitive development, and health (Pope et al., 2002; Sampson et al., 2008; Sharkey and Elwert, 2011; Wodtke et al., 2011).

The Next Stage of Research on Neighborhood Effects

Sharkey and Faber (2014) argue that much of the empirical literature on neighborhood effects has been dominated by the question of whether neighborhoods matter. The emphasis on this question has led to a narrow view of neighborhood effects, a view that is driven by the effort to standardize the measurement of neighborhoods and to confront selection bias in order to identify the causal effect of exposure to poor neighborhoods. Although the focus on methodological rigor has moved the field forward, we argue that the next stage of research on neighborhood effects should be driven by a more expansive view of the various ways in which the stratification of residential settings serves to maintain and reproduce inequality across multiple dimensions.

This approach to neighborhood effects places primary attention on the mechanisms by which residential stratification affects the spatial distribution of social, political, and economic phenomena as well as the life chances of different groups and populations. Neighborhood poverty is only one of many salient dimensions of a community, yet concentrated poverty and affluence have been the focus of much of the empirical literature. Research on the consequences of exposure to other dimensions of the immediate social and physical environment has received much less attention, despite the evidence that has been gathered on the effects of exposure to violence and the importance of air, noise, and soil pollution to child development (e.g., Ransom and Pope, 1992; Reyes, 2012; Sharkey, 2010). Research analyzing processes that operate in space frequently is overlooked in reviews of the neighborhood effects literature, yet is essential to understanding how phenomena such as poverty, illness, and violence come to be concentrated in certain communities and not others (Sampson et al., 2002).

The expanded availability of data on institutions, social and economic phenomena, the built environment, and the physical environment provides researchers with the capacity to generate evidence on a wider range of mechanisms for neighborhood effects (Downey and Van Willigen, 2005; Odgers et al., 2012; Pope et al., 2002; Schwartz, 2010; Sharkey et al., 2013; Small and Stark, 2005). Similarly, new forms of data on networks, use of space, and time use allow researchers to measure how individuals and groups use social space over time (e.g., see Almeida, 2005; Entwisle et al., 2007; Kwan, 2009; Papachristos et al., 2012; Wikström et al., 2012). These improvements in measurement have the potential to lead to a more refined view of the variation in individuals' exposure to different aspects of their residential settings and heterogeneity in individuals' vulnerability to the residential environment.

Methods for the study of neighborhood effects must continue to grapple with the challenge of selection bias, omitted variable bias, and the identification of causal effects. However, relying on exogenous variation in families' residential environments that arises through residential mobility programs is only one limited approach to these challenges. Several studies have exploited exogenous variation in the timing of specific events occurring in individuals' residential environments, such as local homicides or suicides (Hansen et al., 2013; Sharkey 2010), or shocks in local labor markets (Ananat et al., 2011; Lee et al., 2013; Wolfe et al., 2012), in order to identify the effects of changes in the residential environment on outcomes measured at the individual or community level. Another strand of research exploits place-based randomized controlled trials (e.g., Braga, 2005) or experimental audit studies (e.g., Besbris et al., 2014) to identify the effects of the neighborhood setting, or changes in that setting, on social and economic outcomes.

This research reflects a broader view of neighborhoods and neighborhood effects. To understand the role of neighborhoods it is important to move beyond the idea that concentrated poverty can stand in for every salient dimension of the residential setting, or that interviewing children after they have moved out of public housing is a sufficient way to determine whether neighborhoods matter. Our overarching argument is that researchers must take advantage of data on environmental conditions, social processes, institutions, and other dimensions of local context in order to uncover the many ways that residential settings affect the lives of individuals and reproduce patterns of social and economic inequality. In combination with new sources of data and more precise measurement, a set of innovative and rigorous analytic approaches is necessary to develop a more complete understanding of the role that neighborhoods play in the American stratification system.

See also: Neighborhood; Spatial Analysis in Geography; Spatial Thinking in the Social Sciences; Urban Internal Spatial Structure; Urban Poverty in Neighborhoods.

Bibliography

Aaronson, D., 1998. Using sibling data to estimate the impact of neighborhoods on children's educational outcomes. Journal of Human Resources 33, 915–946.

Almeida, D.M., 2005. Resilience and vulnerability to daily stressors assessed via diary methods. Current Directions in Psychological Science 14, 64–68.

Ananat, E.O., Gassman-Pines, A., Gibson-Davis, C.M., 2011. The effects of local employment losses on children's educational achievement. In: Duncan, G., Murnane, R. (Eds.), Whither Opportunity: Rising Inequality, Schools, and Children's Life Chances. Russell Sage, New York; Spencer Foundation, Chicago, pp. 299–315.

Besbris, M., Faber, J.W., Rich, P., Sharkey, P., 2014. An experimental study of neighborhood stigma and the penalty of place: how housing location affects economic transactions. Working Paper.

Braga, A.A., 2005. Hot spots policing and crime prevention: a systematic review of randomized controlled trials. Journal of Experimental Criminology 1 (3), 317–342.

Briggs, X., Popkin, S.J., Goering, J., 2010. Moving to Opportunity: The Story of an American Experiment to Fight Ghetto Poverty. Oxford University Press, New York.

Bronzaft, A.L., McCarthy, D.P., 1975. The effect of elevated train noise on reading ability. Environment and Behavior 7 (4), 517–527.

Burdick-Will, J., Ludwig, J., Raudenbush, S.W., Sampson, R.J., Sanbonmatsu, L., Sharkey, P., 2011. Converging evidence for neighborhood effects on children's test scores: an experimental, quasi-experimental, and observational comparison. In: Duncan, G., Murnane, R. (Eds.), Whither Opportunity: Rising Inequality, Schools, and Children's Life Chances. Russell Sage, New York; Spencer Foundation, Chicago, pp. 255–276.

Bursik Jr., R.J., Grasmick, H.G., 1993. Economic deprivation and neighborhood crime rates, 1960–1980. Law and Society Review 27, 263.

Burton, L.M., Price-Spratlen, T., Spencer, M., 1997. On ways of thinking about and measuring neighborhoods: implications for studying context and developmental outcomes for children. In: Duncan, G., Brooks-Gunn, J., Aber, L. (Eds.), Neighborhood Poverty, Policy Implications in Studying Neighborhoods, vol. 2. Russell Sage, New York, pp. 132–144.

Clampet-Lundquist, S., Edin, K., Kling, J.R., Duncan, G.J., 2011. Moving teenagers out of high-risk neighborhoods: how girls fare better than boys. American Journal of Sociology 116 (4), 1154–1189.

Carlson, D., Haveman, R., Kaplan, T., Wolfe, B., 2012. Long-term earnings and employment effects of housing voucher receipt. Journal of Urban Economics 71 (1), 128–150.

Cohen, S., Evans, G.W., Krantz, D.S., Stokols, D., 1980. Physiological, motivational, and cognitive effects of aircraft noise on children: moving from the laboratory to the field. American Psychologist 35 (3), 231–243.

Crane, J., 1991. The epidemic theory of ghettos and neighborhood effects on dropping out and teenage childbearing. American Journal of Sociology 96 (5), 1126–1159.

Crosnoe, R., 2009. Low-income students and the socioeconomic composition of public high schools. American Sociological Review 74 (5), 709–730.

Crowder, K., South, S.J., 2003. Neighborhood distress and school dropout: the variable significance of community context. Social Science Research 32, 659–698.

Currie, J., Hanushek, E., Kahn, E.M., Neidell, M., Rivkin, S.G., 2009. Does pollution increase school absences? Review of Economics and Statistics 91 (4), 682–694.

Cutsinger, J.M., Galster, G.C., Santiago, A.M., 2011. Neighborhood effects on high school outcomes for Latino and Black youth: evidence from a natural experiment in Denver. In: Presented at the Annual Meetings of the Urban Affairs Association. New Orleans.

Downey, L., Van Willigen, M., 2005. Environmental stressors: the mental health impacts of living near industrial activity. Journal of Health and Social Behavior 46 (3), 289–305.

Durlauf, S.N., 2004. Neighborhood effects. In: Henderson, V., Thisse, J.F. (Eds.), Handbook of Regional and Urban Economics, vol. 4. Elsevier, Amsterdam, pp. 2173–2242.

Ellen, I.G., Turner, M.A., 1997. Does neighborhood matter? Assessing recent evidence. Housing Policy Debate 8 (4), 833–866.

Ellen, I.G., Turner, M.A., 2003. Do neighborhoods matter and why. In: Goering, J.M., Feins, J.D. (Eds.), Choosing a Better Life: Evaluating the Moving to Opportunity Social Experiment. The Urban Institute, Washington, pp. 313–338.

Elliott, M., 2001. Gender differences in causes of depression. Women and Health 33 (3–4), 163–177.

Entwisle, B., Faust, K., Rindfuss, R.R., Kaneda, T., 2007. Networks and contexts: variation in the structure of social ties. American Journal of Sociology 112, 1495–1533.

Evans, G.W., 2006. Child development and the physical environment. Annual Review of Psychology 57, 423–451.

Evans, G.W., Kantrowitz, E., 2002. Socioeconomic status and health: the potential role of environmental risk exposure. Annual Review of Public Health 23, 303–331.

Evans, G.W., Maxwell, L., 1997. Chronic noise exposure and reading deficits: the mediating effects of language acquisition. Environment and Behavior 29 (5), 638–656.

Galster, G., 2008. Quantifying the effect of neighborhood on individuals: challenges, alternative approaches, and promising directions. Schmollers Jahrbuch 128, 1–42.

Galster, G., 2012. The Mechanism(s) of Neighborhood Effects: Theory, Evidence, and Policy Implications. In: van Ham, M., Manley, D., Simpson, L., Bailey, N., Maclennan, D. (Eds.), Neighbourhood Effect Research: New Perspectives. Springer, Dordrecht, pp. 23–56.

Goering, J.M., Feins, J.D., 2003. Choosing a Better Life: Evaluating the Moving to Opportunity Social Experiment. The Urban Institute Press, Washington.

Hambrick-Dixon, P.J., 1985. Effects of experimentally imposed noise on task performance of Black children attending day care center near elevated subway trains. Developmental Psychology 22, 259–264.

Hansen, B., Holden, K., Lindo, J., 2013. Tragedies, student attendance, and academic achievement. In: Paper Presented at New York University, "Violence and Schools" Conference.

Harding, D.J., 2003. Counterfactual models of neighborhood effects: the effect of neighborhood poverty on high school dropout and teenage pregnancy. American Journal of Sociology 109 (3), 676–719.

Harding, D.J., 2009. Collateral consequences of violence in disadvantaged neighborhoods. Social Forces 88 (2), 757–782.

Harding, D.J., Gennetian, L., Winship, C., Sanbonmatsu, L., Kling, J., 2011. Unpacking neighborhood influences on education outcomes: setting the stage for future research. In: Duncan, G., Murnane, R. (Eds.), Whither Opportunity: Rising Inequality, Schools, and Children's Life Chances. Russell Sage, New York; Spencer Foundation, Chicago, pp. 277–296.

Jargowsky, P.A., 1997. Poverty and Place: Ghettos, Barrios, and the American City. Russell Sage Foundation, New York.

Jencks, C., Mayer, S., 1990. The social consequences of growing up in a poor neighborhood. In: Lynn, L.E., McGeary, M.G.H., National Research Council (U.S.) (Eds.), Inner-City Poverty in the United States. Commission on Behavioral and Social Sciences and Education. Committee on National Urban Policy. National Academy Press, Washington, DC, pp. 111–186.

Kirk, D.S., 2009. A Natural Experiment on Residential Change and Recidivism: Lessons from Hurricane Katrina. http://asr.sagepub.com/content/74/3/484 (accessed 14.06.14.).

Klebanov, Pamela Kato, Brooks-Gunn, Jeanne, Duncan, Greg J., 1994. Does neighborhood and family poverty affect mothers' parenting, mental health, and social support? Journal of Marriage and the Family 56 (2), 441–455.

Klebanov, P.K., Brooks-Gunn, J., Evans, G.W., 2014. Poverty, ethnicity, and risk of obesity among low birth weight infants. Journal of Applied Developmental Psychology 35, 245–253.

Kling, J.R., Liebman, J.B., Katz, L.F., 2005. Experimental analysis of neighborhood effects. NBER Working Papers 11577.

Kwan, M.P., 2009. From place-based to people-based exposure measures. Social Science and Medicine 69 (9), 1311–1313.

Lanphear, B.P., Hornung, R., Khoury, J., Yolton, K., Baghurst, P., Bellinger, D.C., Canfield, R.L., Dietrich, K.N., Bornschein, R., Greene, T., Rothenberg, S.J., Needleman, H.L., Schnaas, L., Wasserman, G., Graziano, J., Roberts, R., 2005. Low-level environmental lead exposure and children's intellectual function: an international pooled analysis. Environmental Health Perspectives 113 (7), 894.

Leventhal, T., Brooks-Gunn, J., 2000. The neighborhoods they live in: the effects of neighborhood residence on child and adolescent outcomes. Psychological Bulletin 126 (20), 309–337.

Lee, D., Brooks-Gunn, J., McLanahan, S.S., Notterman, D., Garfinkel, I., 2013. The Great recession, genetic sensitivity, and maternal harsh parenting. Proceedings of the National Academy of Sciences, 1312398110.

Logan, J.R., 2012. Making a place for space: spatial thinking in social science. Annual Review of Sociology 38, 507–524.

Ludwig, J., Liebman, J.B., Kling, J.R., Duncan, G.J., Katz, L.F., et al., 2008. What can we learn about neighborhood effects from the moving to opportunity experiment? American Journal of Sociology 114 (1), 144–188.

Ludwig, Jens, Jacob, Brian A., Johnson, Michael, Duncan, Greg J., Rosenbaum, James E., 2010. Neighborhood Effects on Low-income Families: Evidence from a Randomized Housing Voucher Lottery. University of Chicago. Working Paper.

Ludwig, J., Sanbonmatsu, L., Gennetian, L., Duncan, G.J., Katz, L.F., Kessler, R.C., Kling, J.R., Lindau, S.T., Whitaker, R.C., McDade, T.W., 2011. Neighborhoods, obesity and diabetes – a randomized social experiment. New England Journal of Medicine. [Internet] 365, 1509–1519.

Ludwig, J., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kessler, R.C., et al., 2012. Neighborhood effects on the long-term well-being of low-income adults. Science 337 (6101), 1505–1510.

Massey, D.S., Fischer, M.J., 2003. The geography of inequality in the United States 1950–2000. In: Gale, W.G., Pack, J.R. (Eds.), Brookings-Wharton Papers on Urban Affairs 2003. Brookings Institution (with Mary Fischer), Washington, DC, pp. 1–40.

Massey, D.S., Albright, L., Casciano, R., Derickson, E., Kinsey, D., 2013. Climbing Mount Laurel: The Struggle for Affordable Housing and Social Mobility in an American Suburb. Princeton University Press, Princeton, NJ.

Massey, D.S., Denton, N., 1993. American Apartheid: Segregation and the Making of the Underclass. Harvard University Press, Cambridge, MA.

Mendenhall, R., DeLuca, S., Duncan, G., 2006. Neighborhood resources, racial segregation, and economic mobility: results from the Gautreaux program. Social Science Research 35 (4), 892–923.

Miller, J., 2008. Getting Played: African American Girls, Urban Inequality, and Gendered Violence. NYU Press, New York.

Morenoff, J.D., 2003. Neighborhood mechanisms and the spatial dynamics of birth weight. American Journal of Sociology 108, 976–1017. http://dx.doi.org/10.1086/374405.

Odgers, C.L., Caspi, A., Bates, C.J., Sampson, R., Moffitt, T.E., 2012. Systematic social observation of children's neighborhoods using google street view: a reliable and cost effective method. Journal of Child Psychology and Psychiatry 53 (10), 1009–1017.

Page, M.E., Solon, G., 2003. Correlations between brothers and neighboring boys in their adult earnings: the importance of being urban. Journal of Labor Economics 21 (4), 831–855.

Papachristos, A.V., Braga, A.A., Hureau, D.M., 2012. Social networks and the risk of gunshot injury. Journal of Urban Health 89 (6), 992–1003.

Plotnick, R.J., Hoffman, S., 1999. The effect of neighborhood characteristics on young adult outcomes: alternative estimates. Social Science Quarterly 80, 1–18.

Pope III, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., et al., 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Journal of the American Medical Association 287 (9), 1132–1141.

Popkin, S., Theodos, B., Roman, C., Guernsey, E., 2008. The Chicago Family Case Management Demonstration: Developing a New Model for Serving 'Hard to House' Public Housing Families. The Urban Institute, Washington, DC. http://www.urban.org/UploadedPDF/411708_public_housing_familes.pdf (accessed 14.06.14.).

Ransom, M.R., Pope III, C.A., 1992. Elementary school absences and PM pollution in Utah Valley. Environmental Research 58 (1), 204–219.

Reyes, J.W., 2012. Lead policy and academic performance: insights from Massachusetts. NBER Working Paper 18327.

Rosenblatt, P., DeLuca, S., 2012. "We don't live outside, we live in here": neighborhood and residential mobility decisions among low-income families. City and Community 11 (3), 254–284.

Rubinowitz, L.S., Rosenbaum, J.E., 2000. Crossing the Class and Color Lines: From Public Housing to White Suburbia. University of Chicago Press, Chicago.

Sampson, R.J., Morenoff, J., Gannon-Rowley, T., 2002. Assessing neighborhood effects: social processes and new directions in research. Annual Review of Sociology 28, 443–478.

Sampson, R.J., 2003. Neighborhood-level context and health: lessons from sociology. In: Kawachi, I., Berkman, L. (Eds.), Neighborhoods and Health. Oxford, New York, pp. 132–146.

Sampson, R.J., 2006. How does community context matter? Social mechanisms and the explanation of crime. In: Wikström, P., Sampson, R.J. (Eds.), The Explanation of Crime: Context, Mechanisms, and Development. Cambridge University Press, New York and Cambridge, pp. 31–60.

Sampson, R.J., 2008. Moving to inequality: neighborhood effects and experiments meet social structure. American Journal of Sociology 114, 189–231.

Sampson, R.J., 2012. Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago Press, Chicago.

Sampson, R.J., Raudenbush, S.W., Earls, F., 1997. Neighborhoods and violent crime: a multilevel study of collective efficacy. Science 277, 918–924.

Sampson, R.J., Sharkey, P., Raudenbush, S., 2008. Durable effects of concentrated disadvantage on verbal ability among African-American children. Proceedings of the National Academy of Sciences 105, 845–852.

Sampson, R.J., Wilson, W.J., 1995. Toward a theory of race, crime, and urban inequality. In: Hagan, John, Peterson, Ruth D. (Eds.), Crime and Inequality. Stanford University Press, Stanford, CA, pp. 37–56.

Sanbonmatsu, L., Kling, J.R., Duncan, G.J., Brooks-Gunn, J., 2006. Neighborhoods and academic achievement: results from the moving to opportunity experiment. NBER Working Paper 11909.

Schwartz, H., 2010. Housing Policy Is School Policy: Economically Integrative Housing Promotes Academic Success in Montgomery County, MD. Rand Corporation. http://tcf.org/assets/downloads/tcf-Schwartz.pdf (accessed 14.06.14.).

Sharkey, P., 2008. The intergenerational transmission of context. American Journal of Sociology 113, 931–969.
Sharkey, P., 2010. The acute effect of local homicides on children's cognitive performance. Proceedings of the National Academy of Sciences 107, 11733–11738.
Sharkey, P., 2013. Stuck in Place: Urban Neighborhoods and the End of Progress toward Racial Equality. University of Chicago Press, Chicago.
Sharkey, P., Elwert, F., 2011. The legacy of disadvantage: multigenerational neighborhood effects on cognitive ability. American Journal of Sociology 116, 1934–1981.
Sharkey, P., Faber, J.W., 2014. Where, when, why, and for whom do residential contexts matter? Moving away from the dichotomous understanding of neighborhood effects. Annual Review of Sociology 40, 27.1–27.21.
Sharkey, P., Schwartz, A.E., Ellen, I.G., Lacoe, J., 2013. High stakes in the classroom, high stakes on the street: the effects of community violence on students' standardized test performance. New York University Institute for Education and Social Policy Working Paper, #03–13.
Shaw, C., McKay, H., 1942. Juvenile Delinquency and Urban Areas. University of Chicago Press, Chicago.
Small, M.L., Feldman, J., 2012. Ethnographic evidence, heterogeneity, and neighborhood effects after moving to opportunity. In: Van Ham, M., Manley, D., Bailey, N., Simpson, L., Maclennan, D. (Eds.), Neighborhood Effects Research: New Perspectives. Springer, Dordrecht, pp. 57–77.
Small, M.L., Newman, K., 2001. Urban poverty after the truly disadvantaged: the rediscovery of the family, the neighborhood, and culture. Annual Review of Sociology 27, 23–45.
Small, M.L., Stark, L., 2005. Are poor neighborhoods resource deprived? A case study of childcare centers in New York. Social Science Quarterly 86 (s1), 1013–1036.
Sobel, M.E., 2006. What do randomized studies of housing mobility demonstrate? Journal of the American Statistical Association 101 (476), 1398–1407.
Stansfeld, S.A., Berglund, B., Clark, C., Lopez Barrio, I., Fischer, P., Öhrström, E., Haines, M.M., Head, J., Hygge, S., van Kamp, I., Berry, B.F., 2005. Aircraft and road traffic noise and children's cognition and health: a cross-national study. The Lancet 365, 1942–1949.
Turley, R.N.L., 2003. When do neighborhoods matter? The role of race and neighborhood peers. Social Science Research 32, 61–79.
Turner, M.A., Comey, J., Kuehn, D., Nichols, A., 2012. Residential Mobility, High-opportunity Neighborhoods, and Outcomes for Low-Income Families: Insights from the Moving to Opportunity Demonstration. U.S. Department of Housing and Urban Development, Office of Policy Development and Research, Washington, DC.
Vartanian, T.P., Buck, P.W., 2005. Childhood and Adolescent Neighborhood Effects on Adult Income: Using Siblings to Examine Differences in Ordinary Least Squares and Fixed-effect Models. http://repository.brynmawr.edu/cgi/viewcontent.cgi?article=1002&context=gsswr pubs (accessed 14.06.14.).
Votruba, M.E., Kling, J.R., 2009. Effects of neighborhood characteristics on the mortality of black male youth: evidence from Gautreaux, Chicago. Social Science & Medicine 68 (5), 814–823.
Wheaton, B., Clarke, P., 2003. Space meets time: integrating temporal and contextual influences on mental health in early adulthood. American Sociological Review, 680–706.
Wikström, P.O., Oberwittler, D., Treiber, K., Hardie, B., 2012. Breaking Rules. The Social and Situational Dynamics of Young People's Urban Crime. Oxford University Press, Oxford.
Wilson, W.J., 1987. The Truly Disadvantaged. University of Chicago Press, Chicago, IL.
Wolfe, B., Jakubowski, J., Haveman, R., Courey, M., 2012. The income and health effects of tribal casino gaming on American Indians. Demography 49 (2), 499–524.
Wodtke, G.T., Harding, D.J., Elwert, F., 2011. Neighborhood effects in temporal perspective: the impact of long-term exposure to concentrated disadvantage on high school graduation. American Sociological Review 76 (5), 713–736.
Wodtke, G., Elwert, F., Harding, D.J., 2012. Poor Families, Poor Neighborhoods: How Family Poverty Intensifies the Impact of Concentrated Disadvantage on High School Graduation. PSC Research Report No. 12–776. http://www.psc.isr.umich.edu/pubs/pdf/rr12-776.pdf (accessed 14.06.14.).
Zimmerman, G.M., Messner, S., 2010. Neighborhood context and the gender gap in adolescent violent crime. American Sociological Review 75, 958–980.
Zuberi, A., 2012. Neighborhood poverty and children's exposure to danger: examining gender differences in impacts of the moving to opportunity experiment. Social Science Research 41 (4), 788–801.

Neighborhood Revitalization and Community Development

WD Keating, Cleveland State University, Cleveland, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Neighborhood revitalization efforts and programs, both public and private, are reviewed. This includes early efforts at elimination of slums led by settlement houses, housing reformers, and social workers. This was followed by government programs like housing and health codes, public housing, urban renewal (both slum clearance and rehabilitation), model cities, neighborhood planning and redevelopment, and, most recently in the United States, the Empowerment Zone and Enterprise Communities program. The gentrification phenomenon is discussed. Various techniques of neighborhood revitalization are reviewed, as well as critiques of their impact. Likewise, community development programs, both public and private, are reviewed. Efforts undertaken in the United States by private foundations ('intermediaries') are reviewed as well. The role of community organizations like community development corporations are discussed. The impact of the mortgage foreclosure crisis on urban neighborhoods and responses to it are reviewed. Emphasis is primarily on the United States, with reference to North America and Western Europe.

Introduction

Neighborhood revitalization efforts and programs, both public and private, are reviewed. These include early efforts at elimination of slums led by settlement houses, housing reformers, and social workers. This was followed by government programs like housing and health codes, public housing, urban renewal (both slum clearance and rehabilitation), model cities, neighborhood planning and redevelopment, and, most recently, the Empowerment Zone and Enterprise Communities program. Various techniques of neighborhood revitalization are reviewed, as well as critiques of their impact. Likewise, community development programs, both public and private, and the role of community development corporations (CDCs) are reviewed. The primary emphasis is on the United States.

The Progressive Era: Tenement Reform and Settlement Houses

From the mid-nineteenth century through the beginning of World War I, the population of US cities swelled as a result of mass immigration, primarily from Europe. Most poor immigrants were concentrated in ethnic enclaves. In an era in which there were no governmental social and welfare services other than the poorhouse, poor immigrants lived in overcrowded and unsanitary housing and neighborhood conditions. In cities like New York, they lived in tenements in wretched conditions.

Around the turn of the century, the Progressive Era saw the rise of reform movements. These included the efforts of housing reformers to address the conditions prevalent in slum neighborhoods. New York City was in the forefront, led by a Tenement House Committee of reformers like Lawrence Veiller. The journalist Jacob Riis exposed slum conditions in his graphic descriptions and photographs of 'how the other half lives.' This led to the enactment of the New York Tenement House Law of 1901, designed to eliminate the worst abuses. It became a model for similar legislation in other states. However, reformers like Veiller were loath to invoke the government and were especially opposed to government-subsidized housing for the poor. Instead, they eventually turned to the adoption of municipal housing and health codes, setting minimum regulatory standards for residential housing (Friedman, 1968). A few philanthropists sponsored limited dividend model housing for workers but this failed to prove to be a viable alternative to the housing provided by private landlords (Birch and Gardner, 1981).

During the same period, a settlement house movement began, based upon similar previous efforts in England. Reformers supported by private philanthropy sought to alleviate social problems in slum neighborhoods. They established settlement houses to provide social services to the poor. Largely apolitical, they sought to uplift the poor, while trying to improve their living conditions. Leaders included Jane Addams in Chicago and Robert Woods in Boston (Fisher, 1984; Halpern, 1995).

Slum Clearance: Public Housing

It was not until the Great Depression of the 1930s that public housing appeared in the United States, decades after its adoption in Europe. The Depression exacerbated already existing problems in poor urban neighborhoods. The reforms of the Progressive Era had done little to alleviate the underlying causes of the slums and substandard housing. A new generation of activist housing reformers argued for intervention of the federal government to replace slum housing with government-financed, -owned, and -operated public housing. This was a radical departure from US policy.

Reluctantly, the New Deal administration of President Franklin D. Roosevelt began a program of building public housing as part of a public works slum clearance program aimed at providing work to unemployed workers in the building trades. The real estate industry fought hard against the creation of public housing, arguing instead that private landlords should provide for the housing needs of low-income persons. The 1937 legislation provided federal financial assistance to those local governments that chose to participate

 http://dx.doi.org/10.1016/B978-0-08-097086-8.74018-9

in a program to provide decent, affordable housing for the poor. Local public housing authorities were to construct and operate the public housing projects. The real estate lobby was successful in limiting public housing to the poor and imposing limits on the amounts allowed for public housing units. In addition, public housing could only be built to replace substandard units. These limitations would later prove fateful in undermining the goals of the reformers and dooming the expansion of public housing. Despite the attempts of a few localities, racial segregation of public housing prevailed, reflecting racist attitudes and existing housing and neighborhood patterns.

The public housing program began slowly was interrupted by the emergency of World War II and barely survived conservative opposition in the housing act of 1949. It was adopted primarily by large cities with a preponderance of the poor, which received most of the funding and engaged in the construction of large-scale public housing projects. Even with its limitations, public housing was a great improvement over the slum housing that it replaced. In its early stages, public housing served as housing for the working poor.

However, as the public housing grew older, its population in the 1960s became poorer (Meehan, 1979). The housing projects were often physically isolated, in poor condition, and racially segregated – despite increased federal subsidies and court mandates. Except for housing for the elderly poor, new construction of public housing virtually ended in the 1970s. Instead, the US government switched to providing subsidies to enable poor tenants to rent private rental units. Beginning in the 1990s, some of the most notoriously unsuccessful public housing projects were either demolished or reduced in scale and rebuilt as mixed-income housing through the HOPE VI program. Through its fiscal year 2010, the US Department of Housing and Urban Development (HUD) reported that 96 017 'distressed' public housing units had been demolished and 83 951 replacement units were built, of which 82 425 had been occupied. Affordable housing advocates expressed concern about the loss of low-rent units and the lack of decent, affordable relocation housing for those displaced (Popkin et al., 2004; Bennett et al., 2006; Goetz, 2010, 2011a,b). Critics have also termed the impact of HOPE VI as a form of gentrification (Goetz, 2011a,b). For those displaced public housing tenants who did not return after the sites were redeveloped, a majority were relocated in other public housing projects. Most of the rest received vouchers and mostly moved to nearby neighborhoods. They found these to be better in some respects (e.g., safety) but more expensive. Loss of social ties offset relocation to these neighborhoods, which were generally racially segregated (as were their former public housing projects) (Fraser et al., 2012).

After the administration of President George W. Bush failed to terminate the HOPE VI program, its successor phased it out in favor of a Choice Neighborhoods program. Intended to broaden the impact to 'distressed' urban neighborhoods by linking improved housing to schools, public transit, and employment opportunities, planning grants were awarded in 2011 to 13 communities.

Despite the bad reputation of several projects nationally, for the most part public housing in urban areas did contribute to the elimination of the worst slum conditions.

Slum Clearance: Urban Renewal

Planning for postwar redevelopment in US cities began in the early 1940s. Liberals favored large-scale slum clearance, replacing them largely with public housing (Gelfand, 1975). In 1949, Congress approved an urban redevelopment program, which would eliminate slums and 'blight.' However, as it unfolded, prime downtown real estate became the focus of urban renewal in many cities. Instead of public housing, this land was reused for commercial and industrial facilities, luxury housing, and institutional uses. Despite assurances of relocation and replacement housing, hundreds of thousands of poor slum dwellers saw their neighborhoods and housing destroyed but their housing conditions worsened instead of improving as little new low-rent public housing was constructed on urban redevelopment sites (Greer, 1965). New York City under the leadership of Robert Moses provided some of the most egregious examples of this process, although Moses also built many of the best public housing projects in the United States (Caro, 1974).

The program was renamed as 'urban renewal' in 1954. The newly elected Republican administration of President Dwight Eisenhower, while not opposed to urban renewal, was concerned about its cost to the federal government and the possibility that much of the cleared land was not proving to be attractive to private, for-profit developers. It tried to change the focus of urban renewal in part to conservation and rehabilitation, which it saw as less expensive. Urban renewal clearance and redevelopment policies would later be condemned as bad planning by critic Jacobs (1961) in a famous critique.

In the 1930s, the realtors opposed federal intervention in the housing market and proposed neighborhood associations as the primary vehicle for neighborhood conservation. This was aimed primarily at homeowners' associations. This often had an exclusionary aim, with white homeowners organizing to prevent any entry by African Americans and other minorities. This is also the era in which the newly created Federal Housing Administration (FHA) adopted the racist standards of the real estate industry to rate neighborhoods for the purpose of making and insuring home mortgage loans. This became the basis for the 'redlining' of older urban neighborhoods and contributed to the exodus from these neighborhoods to the suburbs after World War II.

Many urban neighborhoods, mostly poor, also were destroyed to make way for central city links to the interstate highway system created in 1956. As with urban renewal, promises of decent, affordable replacement housing often went unfulfilled. Not all neighborhoods designated as slums for urban renewal or highway construction were necessarily blighted (Gans, 1962). Where residents did organize to oppose the demolition of their neighborhood, typically they failed because they did not have political influence and had no legal standing in court to challenge these plans.

This finally changed in the late 1960s, following summers of riots in US cities from 1965 through 1968. Citizen participation finally was mandated and neighborhood residents at least had an opportunity to present their views. Advocacy planning arose to provide neighborhood residents with the capacity to develop alternate plans. The courts finally recognized the right of those to be displaced to challenge plans,

typically on the ground that there was insufficient replacement housing available to meet federal requirements (Hartman, 1974). Community resistance led to the demise of large-scale clearance for urban renewal and highways in the central cities by the early 1970s (Klemek, 2011).

Model Cities

It was felt that displacement of poor and minority residents for these purposes contributed to the resentment that sparked so many of the urban riots of the 1960s. Resident participation in planning led to more serious consideration of rehabilitation of blighted neighborhoods rather than clearance and redevelopment. The first major example of this policy shift was embodied in the Model Cities program launched by a newly created HUD in 1967. In contrast to urban renewal slum clearance, this reaction to the urban riots proposed to concentrate federal, state, and local resources to improve targeted poor neighborhoods. Housing rehabilitation, social services, job training, and community participation were hallmarks of this approach. Due to insufficient funds, intergovernmental coordination problems, and a change from the Johnson to Nixon administration after the 1968 presidential election, Model Cities failed to fulfill its promises (Haar, 1975).

Community Development Block Grants

In 1974, the urban renewal and Model Cities programs, among others, were ended and replaced by the Community Development Block Grant (CDBG) program. The CDBG program was intended to give local governments more discretionary control over spending and to involve more communities (Hays, 1995). However, CDBG funds were supposed to benefit primarily low- and moderate-income residents. This caused controversy in many localities and often led to wider dispersal of funds, to the detriment of those neighborhoods most in need of assistance (Rich, 1993). This era marks a shift in federal policy that persisted through the Clinton administration, emphasizing devolution of responsibility for many federal aid programs and resulting in cutbacks in many HUD housing assistance programs.

Federal Recognition of Neighborhoods

Federal policy since the 1930s did not usually recognize neighborhoods. HUD and its predecessor agencies made grants and loans to cities and states, and developers and lenders, and provided insurance to individuals (homebuyers). In 1977, President Jimmy Carter made the unprecedented move of appointing an Assistant HUD Secretary for Neighborhoods and created a small program that for the first time awarded grants directly to neighborhood development organizations (Mayer, 1984). In 1979, Carter appointed a presidential commission to advise him on neighborhood policies and programs but then disregarded its recommendations. His successor President Ronald Reagan immediately dismantled Carter's modest neighborhood initiatives at HUD.

Redlining and Community Reinvestment

One of the major reasons for the decline of older urban neighborhoods was their 'redlining.' Lenders, insurers, and realtors avoided those neighborhoods that had a significant minority population on the basis that their presence jeopardized real estate values and would cause white flight. Federal agencies, most notably the FHA, followed suit. This in turn almost assured that eventually these neighborhoods would become all or predominantly minority occupied. 'White flight,' triggered by unscrupulous real estate practices, then occurred in many neighborhoods (Jackson, 1985).

Chicago, where such segregationist practices as 'racial covenants' originated and long the most racially segregated city and metropolitan area in the United States, was the focus of organizing in the 1970s against lender disinvestment in urban neighborhoods. Neighborhood organizers demanded that lenders cease redlining and invest in neighborhoods. Racial discrimination in housing and lending had been previously outlawed, although enforcement of this legislation has yet to end such practices. The antiredlining movement spread and eventually led to the enactment of federal legislation in 1975 and 1977 requiring federally regulated lenders first to disclose generally where their loans were made and then to serve all the communities within their territory. The US Congress rejected more restrictive proposals. The movement then used this Community Reinvestment Act (CRA) legislation to document lender discrimination against older neighborhoods and to require lender commitments to again lend in these neighborhoods. This has led to major lender commitments in many cities (Squires, 1992; Schwartz, 1998). This is a critical component of neighborhood revitalization, especially when the federal government was reducing its role in aiding urban areas. During the mortgage foreclosure crisis, some conservatives attempted to blame the CRA for the high rate of foreclosures in poorer neighborhoods, a claim which has been refuted (Immergluck, 2009).

Gentrification of Urban Neighborhoods

Beginning in the late 1970s, a contrasting phenomenon to disinvestment began. In many large cities, middle- and upper income newcomers began to discover and move into older neighborhoods in decline near downtowns (Palen and London, 1984). These 'urban pioneers' renovated buildings that were typically in disrepair. As their numbers increased, these neighborhoods saw other changes, particularly in the quality of services, both private and public, reflecting the demands of the affluent newcomers.

City governments viewed this process as benign, with improvements in the physical quality of neighborhoods and increased tax revenue. However, in many of these neighborhoods, lower income residents were displaced, directly or indirectly, because they were relocated for building renovation or could not afford rehabilitated housing. There was also often a clash of lifestyles between the residents and the newcomers. While the magnitude of gentrification-based displacement was debated (Schill and Nathan, 1983), many localities did institute land use and housing regulations to provide some

protection to residents against displacement and to ensure the replacement of lower priced housing units (Keating and Smith, 1996). The debate over the impact of gentrification has continued (Brown-Saracino, 2010).

The Emergence of Community Development Corporations

In the 1960s, the US Congress funded directly the CDCs. They are nonprofit community-based organizations operating in distressed, mostly urban and central city neighborhoods. Many originally derived from community organizing. Many are church sponsored. They attempt to address both social and physical problems. They have expanded greatly in number, totaling several thousands by the 1990s (Vidal, 1992). In larger cities, there are now citywide networks of CDCs.

CDCs engage in a variety of activities, typically housing and commercial development, job training, and, to a lesser degree, social services (Erickson, 2009). They are funded from a wide variety of sources, including all levels of government, philanthropic foundations, and corporations (Vidal, 1992). The impact of CDC activities has been rated positively (e.g., Rohe, 1998; Walker, 1993). Nevertheless, they continue to face organizational and financial challenges (Bratt and Rohe, 2004).

However, it must be conceded that, despite the often heroic efforts of CDCs and companion social service agencies, measurable levels of distress such as poverty and crime have increased rather than decreased in many urban neighborhoods since the advent of CDCs (Rusk, 1999). CDCs are simply not enabled or funded to deal with the entire myriad of social and physical problems in distressed urban neighborhoods. Ultimately, governments in partnership with nonprofit organizations, residents, and so-called intermediary funders are responsible for the revitalization of neighborhoods.

Neighborhood Revitalization Techniques

A variety of techniques have been used in the revitalization of neighborhoods. Resident attitudes toward their neighborhood have been surveyed (e.g., Ahlbrandt and Cunningham, 1979). Neighborhoods have been classified according to their physical and social characteristics (Keating and Smith, 1996). Public programs have often targeted neighborhoods, based upon their profile and degree of need for assistance. Housing has been newly constructed and rehabilitated for both lower and moderate-to-middle-income range occupants (Rohe et al., 1998; Walker, 1993). The low-income housing tax credit, which enables private corporations to profitably invest, is a critical factor in the provision of much affordable housing. In badly deteriorated neighborhoods, an urban homesteading approach has been tried to encourage self-help rehabilitation (Varady, 1986). Existing and new businesses have been subsidized to improve their buildings, expand their operations, and assist in the training and employment of neighborhood residents. Neighborhood plans have been developed, with and without resident participation, to identify serious problems and programs and policies to alleviate them (Rohe and Gates, 1985). A favored technique is the mapping of neighborhood assets as cornerstones for neighborhood revitalization. Neighborhood organizing remains a key ingredient in successful neighborhood redevelopment, with block clubs often being the cornerstone of neighborhood organizations (Medoff and Sklar, 1994). Neighborhood organizing can be confrontational or cooperative, depending upon the circumstances and the style of organizing (Fisher, 1984).

The People versus Place Policy Debate

A major debate in neighborhood revitalization has been termed people versus place. In 1968, the US national commission on urban riots posed the dilemma of whether to rebuild poor ghetto areas or try to disperse and integrate their mostly minority residents into mainstream society. More than four decades later, this debate continues (Goetz and Chapple, 2010). Downs (1994) has posed alternate strategies aimed at either improving the social fabric and economic and physical conditions of impoverished inner city neighborhoods or providing mobility in work and housing to enable their residents to move to better neighborhoods in central cities and the surrounding suburbs. As previously discussed, the HOPE VI program was intended to replace demolished distressed high-rise public housing projects with mixed income, low-rise housing that would improve the lives of the low-income tenants. There have been a few small-scale housing mobility experiments to disperse the poor, first in Chicago related to the desegregation of public housing and then as a demonstration program (Moving to Opportunity) in five cities (Polikoff, 2006; de Souza et al., 2010). However, this has not become a national policy. [M]ore emphasis has been placed on the revitalization of inner city neighborhoods with the goal of their improvement for the existing residents.

The latest example in the United States of the latter approach is embodied in the Empowerment Zone and Enterprise Community program that was launched in 1994. It was a belated response to the 1992 civil disturbance in south central Los Angeles. Using a combination of employer tax credits, subsidized loans for businesses, and special funding for social services, including job training, the aim was to provide employment primarily in designated high-poverty areas in more than 100 communities (Gale, 1996). The results from this short-lived program were only mixed at best (Rich and Stoker, 2010).

The Impact of the Mortgage Foreclosure Crisis on Neighborhoods and Responses

In addition to existing problems like substandard housing, many poor urban neighborhoods began to experience a new problem in the early part of the twenty-first century in the midst of a housing bubble: 'predatory lending.' Ironically, in the past, many of the same neighborhoods had been 'redlined' by lenders who refused to lend, claiming unacceptable borrower risk. Their residents were now targeted by mortgage brokers offering mortgage loans, including refinancing, to those

otherwise previously deemed either only marginally eligible or ineligible due to low incomes or poor credit ratings. These 'subprime' loans, often financed indirectly by major lenders, contained high interest rates and terms unfavorable to borrowers, who often did not understand these loan terms. In the most egregious cases, borrower incomes were exaggerated, down payments were arranged by brokers, and only low or no down payments were required. Worse, so-called liar loans contained false information regarding borrower income or employment.

As a result of these predatory loans, many of the borrowers defaulted, leading to growing foreclosure rates. As the lenders foreclosed and the homeowners departed, these vacant homes became targets for those who stripped them of any valuable amenities. In a recessionary economy and a housing crash, lenders found that buyers became scarce. State foreclosure laws sometimes contributed to the time that these vacant homes became vulnerable to vandalism. As economic problems escalated, many homeowners with regular, 'prime rate' mortgages were also unable to meet their obligations and lost their homes to foreclosure. Increasingly, millions of US homeowners found themselves 'under water,' i.e., the plummeting value of their homes was worth less than the amount ('principal') owed to the mortgage lender, leading many to simply walk away and abandon their home. Cities found themselves spending much more to maintain vacant homes abandoned by these owners, and not being maintained by foreclosing lenders. The entry of so-called bottom feeding speculators purchasing these homes at auction or in bulk from lenders and then 'flipping' them, i.e., reselling them quickly after cosmetic repairs, further exacerbated the situation.

The responses of the federal government beginning in 2008 have failed to save millions of homeowners from foreclosure. This has included a program to counsel homeowners in order to prevent foreclosure by persuading lenders to modify mortgages to make them more affordable. The federal government also placed the two bankrupt government-sponsored enterprises into conservatorship and has had to bail them out at huge cost to the taxpayer. They now own and finance most of the residential mortgages in the United States. The so-called robo-signing scandal led to a 2012 agreement between the state attorneys general and five major lenders to make changes in mortgage modification processing and provide some modest restitution to victims of predatory lending.

To help localities to demolish abandoned nuisance properties, assist homeowners, and build affordable housing in targeted neighborhoods, the federal government created a Neighborhood Stabilization Program. While three rounds of funding assisted many localities, the amount of funding fell far short of the need (Mallach, 2009). Therefore, without additional federal aid, until the housing market revives, localities face the prospect of a continuation of the cycle of foreclosures, abandonment, demolitions, and a large number of blighted properties plaguing many of their neighborhoods. However, in many of these communities, local governments and community organizations have created land banks to address abandoned housing and initiated the reuse of vacant land (e.g., urban agriculture) in neighborhoods hit hard by foreclosures (Mallach, 2010).

Neighborhood Revitalization in Canada and Western Europe

Urban revitalization and regeneration programs in Canada and Western Europe are similar to approaches in the United States described above. However, the context is much different. Governmental systems typically feature national and regional policies, programs, and agencies. Municipal authorities may have less authority than in the United States. With few exceptions, the levels of poverty and distress do not compare with the United States, with the exception of industrial cities in sharp decline. Finally, there has not been nearly as much racial conflict in the cities of Canada and Western Europe as in the United States, again with some exceptions, and backlash against immigrants in some countries. The urban regeneration programs of the United Kingdom have been similar in many respects to those in the United States (Jones and Evan, 2008; Lawless, 2010). Urban regeneration programs in Europe vary widely but with an emphasis on sustainability and addressing social exclusion affecting poorer populations (Colantonio and Dixon, 2011; Couch et al., 2003). Beginning in 1994, the European Union launched an URBAN Community Initiative, aimed at the regeneration of deprived neighborhoods. The Regenera network, financed by the European Urbact program, focuses on physical urban renewal; integration of ethnic minorities, accessibility, and transport; and the organization of local communities in 13 cities in 8 European countries (Regenera, 2007).

For an earlier comparative perspective on urban regeneration in North America and Europe, see Judd and Parkinson (1990). For an earlier comparative perspective on neighborhood revitalization in Australia, Canada, and Western Europe, see Gale (1984). More recently, some Western European countries also experienced a housing bubble and then a crash, leaving an escalating number of homeowners facing foreclosure and vacant houses unsalable in a depressed housing market, most notably in Ireland and Spain. Recent higher levels of unemployment resulting from an economic recession have added to the problems of poorer urban areas.

See also: Development and Urbanization; Gentrification; Planning Ethics; Planning Issues and Sustainable Development; Strategic Planning for Public and Nonprofit Organizations; Urban Life and Health.

Bibliography

Ahlbrandt, R.S., Cunningham, J.V., 1979. A New Public Policy for Neighborhood Preservation. Praeger, New York.

Bennett, L., Smith, J.L., Wright, P.A., 2006. Where Are the Poor People to Live? Transforming Public Housing Communities. M.E.Sharpe, London.

Birch, E.L., Gardner, D.S., 1981. The seven percent solution: a review of philanthropic housing, 1870–1910. Journal of Urban History 7, 403–438.

Bratt, R.G., Rohe, W.M., 2004. Organizational changes among CDCs: assessing the impacts and navigating the challenges. Journal of Urban Affairs 26, 197–220.

Brown-Saracino, J., 2010. The Gentrification Debates: A Reader. Routledge, London.

Caro, R.A., 1974. The Power Broker: Robert Moses and the Fall of New York. Alfred A. Knopf, New York.

Colantonio, A., Dixon, T., 2011. Urban Regeneration and Social Sustainability: Best Practice from European Cities. Wiley-Blackwell, Oxford, NY.

Couch, C., Fraser, C., Percy, 2003. Urban Regeneration in Europe. Blackwell Publishing, Oxford, UK.
de Souza, X., Popkin, S.J., Goering, J., 2010. Moving to Opportunity: The Story of an American Experiment to Fight Poverty. Oxford University Press, London.
Downs, A., 1994. New Visions for Metropolitan America. The Brookings Institution and Lincoln Institute of Land Policy, Washington, DC and Cambridge, MA.
Erickson, D.J., 2009. The Housing Policy Revolution: Networks and Neighborhoods. Urban Institute Press, Washington, DC.
Fisher, R., 1984. Let the People Decide: Neighborhood Organizing in America. Twayne, Boston, MA.
Fraser, J., DeFilippis, Bazuin, 2012. In: Bridge, G., Butler, T., Lee, L. (Eds.), HOPE VI: Calling for Modesty in Its Claims. Policy Press, Chicago, IL.
Friedman, L.M., 1968. Government and Slum Housing: A Century of Frustration. Rand, McNally, Chicago.
Gale, D.E., 1984. Neighborhood Revitalization and the Postindustrial City. Lexington Books, Lexington, MA.
Gale, D.E., 1996. Understanding Urban Unrest: From Reverend King to Rodney King. Sage, Thousand Oaks, CA.
Gans, H.J., 1962. The Urban Villagers: Group and Class in the Life of Italian–Americans. Free Press, New York.
Gelfand, M.I., 1975. A Nation of Cities: The Federal Government and Urban America, 1933–1965. Oxford University Press, Oxford, NY.
Goetz, E., 2011a. Where have all the towers gone? The dismantling of public housing in U.S. cities. Journal of Urban Affairs 33, 267–287.
Goetz, E., 2011b. Gentrification in black and white: the racial impact of public housing demolition in American cities. Urban Studies 48, 1581–1604.
Goetz, E., 2010. Better neighborhoods, better outcomes? Explaining relocation outcomes in HOPE VI. Cityscape 12, 5–31.
Goetz, E.G., Chapple, K., 2010. Dispersal as an anti-poverty policy. In: Davies, J.S., Imbroscio, D.L. (Eds.), Critical Urban Studies: New Directions. SUNY Press, Albany, NY.
Greer, S., 1965. Urban Renewal and American Cities: The Dilemma of Democratic Intervention. Bobbs-Merrill, Indianapolis, IN.
Haar, C.M., 1975. Between the Idea and the Reality: A Study in the Origin, Fate and Legacy of the Model Cities Program. Little Brown, Boston, MA.
Halpern, R., 1995. Rebuilding the Inner City: A History of Neighborhood Initiatives to Address Poverty in the United States. Columbia University Press, New York.
Hartman, C., 1974. Yerba Buena: Land Grab and Community Resistance in San Francisco. Glide Publications, San Francisco, CA.
Hays, R.A., 1995. The Federal Government and Urban Housing: Ideology and Change in Public Policy. State University of New York, Albany, NY.
Immergluck, D., 2009. Foreclosed: High-Risk Lending, Deregulation, and the Undermining of America's Mortgage Market. Cornell University Press, Ithaca, NY.
Jacobs, J., 1961. The Death and Life of Great American Cities. Random House, New York.
Jackson, K.T., 1985. Crabgrass Frontier: The Suburbanization of the United States. Oxford University Press, Oxford, NY.
Jones, P., Evan, J., 2008. Urban Regeneration in the UK. Sage, Thousand Oaks, CA.
Judd, D., Parkinson, M., 1990. Leadership and Urban Regeneration: Cities in North America and Europe. Sage, Newbury Park, CA.
Keating, W.D., Smith, J., 1996. Neighborhoods in transition. In: Keating, W.D., Krumholz, N., Star, P. (Eds.), Revitalizing Urban Neighborhoods. University Press of Kansas, Lawrence, KS.
Klemek, C., 2011. The Transatlantic Collapse of Urban Renewal: Postwar Urbanism from New York to Berlin. University of Chicago Press, Chicago, IL.
Lawless, P., 2010. Urban regeneration: is there a future? People, Place and Policy Online 4 (1), 24–28.
Mallach, A., 2009. Stabilizing Communities: A Federal Response to the Secondary Impacts of the Foreclosure Crisis. The Brookings Institution, Washington, DC.
Mallach, A., 2010. Bringing Buildings Back: From Abandoned Properties to Community Assets. National Housing Institute, Montclair, NJ.
Mayer, N.S., 1984. Neighborhood Organizations and Community Development: Making Revitalization Work. The Urban Institute Press, Washington, DC.
Medoff, P., Sklar, H., 1994. Streets of Hope: The Fall and Rise of an Urban Neighborhood. South End Press, Boston, MA.
Meehan, E.J., 1979. The Quality of Federal Policymaking: Programmed Failure in Public Housing. University of Missouri Press, Columbia, MI.
Palen, J.J., London, B., 1984. Gentrification, Displacement and Neighborhood Revitalization. State University of New York Press, Albany, NY.
Polikoff, A., 2006. Waiting for Gautreaux: A Story of Segregation, Housing, and the Black Ghetto. Northwestern University Press, Evanston, IL.
Popkin, S.J., Katz, B., Cunningham, M.K., Brown, K.D., Gustafson, J., Turner, M.J., 2004. A Decade of HOPE VI: Research Findings and Policy Challenges. Urban Institute Press, Washington, DC.
Regenera, 2007. Urban Regeneration of Deprived Areas across Europe. Sharing Experiences. Final Report.
Rich, M.J., 1993. Federal Policymaking and the Poor: National Goals, Local Choices, and Distributional Outcomes. Princeton University Press, Princeton, NJ.
Rich, M.J., Stoker, R.P., 2010. Rethinking empowerment evidence from local empowerment zone program. Urban Affairs Review 45, 775.
Rohe, W.M., 1998. Do community development corporations live up to their billing? A review and critique of the research findings. In: Koebel, C.T. (Ed.), Shelter and Society: Theory, Research and Policy for Nonprofit Housing. State University of New York Press, Albany, NY.
Rohe, W.M., Gates, L.B., 1985. Planning with Neighborhoods. University of North Carolina Press, Chapel Hill, NC.
Rohe, W.M., Quercia, R.G., Levy, D.K., Biswas, P., 1998. Sustainable Nonprofit Housing Development: An Analysis of Maxwell Award Winners. Fannie Mae Foundation, Washington, DC.
Rusk, D., 1999. Inside Game Outside Game: Winning Strategies for Saving Urban America. Brookings Institution Press, Washington, DC.
Schill, M.H., Nathan, R.P., 1983. Revitalizing America's Cities: Neighborhood Reinvestment and Displacement. State University of New York Press, Albany, NY.
Schwartz, A., 1998. Bank lending to minority and low-income households and neighborhoods: do community reinvestment agreements make a difference? Journal of Urban Affairs 20, 269–301.
Squires, G.D. (Ed.), 1992. From Redlining to Reinvestment: Community Responses to Urban Disinvestment. Temple University Press, Philadelphia, PA.
Varady, D.P., 1986. Neighborhood Upgrading: A Realistic Assessment. State University of New York Press, Albany, NY.
Vidal, A.C., 1992. Rebuilding Communities: A National Study of Urban Community Development Corporations. New School for Social Research, New York.
Walker, C., 1993. Nonprofit housing development: status, trends, and prospects. Housing Policy Debate 4, 361–414.

Neocolonialism and Social Work

Vasilios Ioakimidis, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Arguments about the colonial and imperialist character of international social work initially emerged in the 1970s as part of the broader postcolonial movement in social sciences. Since then these approaches have remained on the sidelines of mainstream social work literature. However, the recent reinvigoration of the indigenization debate has provided an important context for social work to reexamine its political character and scope as a 'global profession.' This article critically evaluates the evolution of the postcolonial debates in social work. The author argues that the process of social work 'decolonization' requires broad alliances with the oppressed populations on the basis of challenging neoliberal capitalism.

Introduction

Postcolonialism refers to a set of theoretical concepts, approaches, and interventions, which deal with the diverse effects of the interaction between the colonizer and the colonized. The conceptualization and focus of the postcolonial analyses is very diverse and informed by different schools of thought. Postmodern and Marxian approaches have significantly influenced the formation of the most important postcolonial discourses. These origins provide a distinct critical edge in the postcolonial analysis and often encourage specific actions toward the multilevel liberation and emancipation of communities, which have suffered from isolation, oppression, and discrimination as a result of colonialism. The vast differences between postmodernism and Marxism create a diverse, rigorously debated and inherently self-contradictory postcolonial analysis. Even within the broad realm of critical theory, certain aspects of postcolonialism are targeted by objections and criticisms. The lack of an – even loosely – homogenous discourse and the fragmentation of oppressed peoples' experiences in the name of an ill-defined diversity, have created space for a heated debate within the communities of theorists, scholars, activists, and social movements (Barlow, 1993; Madsen, 1999).

Despite this diversity in origin and knowledge base, one could broadly identify the major domains of discourse that are of interest for postcolonialism as follows:

1. *The historical context of colonialism.* Postcolonial analyses identify the beginning of 'colonialism' in the period of rapid and violent colonial expansion of European powers. The need for natural resources, quest for global control, and competition between European empires introduced a new epoch of exploration, expansion, and exploitation. Such expansion not only involved military might and physical violence but also the forceful conversion of local 'indigenous' populations into the cultural, philosophical, political, and religious principles of the colonial powers. Slavery, racism, and orientalism are creations of this period (see Hobsbawm, 1975). The industrial revolution did not end colonialism, but merely changed its form and practices. Liberation movements and the eventual independence of colonies in the twentieth century generated an intense debate attempting to identify and analyze new forms of colonialism.
2. *The impact of colonial domination on (primarily) colonized populations and the effect of this process on the diverse population of the colonial powers.* Along with the evident and well-documented explicit policies/practices of inhumane treatment and exploitation of the colonized peoples, indirect methods of subordination are often absent or neglected from academic discussions. These methods include the suppression of local cultures, extinction of languages, and degradation of local knowledge and value systems. The main reasons behind the lack of analysis on these practices is linked to a history of aggressive, sophisticated, and persistent colonial assimilation, making oppressed communities weak in resisting such powerful domination. 'Divide and rule' has been a particularly popular strategy among colonial powers (Callinicos, 2009). Consequently, oppressed peoples have internalized oppression and consider the colonial and postcolonial domination as a 'natural order' (Freire, 1970).
3. *The creation of a theory and praxis deconstructs 'given realities' and attempts to emancipate oppressed peoples.* Postcolonial approaches have provided an important tool in analyzing or reconstructing power relations and multiple layers of colonial domination. Such analysis has informed and encouraged a number of social movements aiming at the emancipation and self determination of oppressed groups. Indigenous and antiracist movements have been particularly influenced by postcolonialism.

The term 'postcolonialism' in the literature highlights the historical realities and sociopolitical challenges, which emerge in parts of the world *after* (post) the end of the colonial era. According to this approach, postcolonialism deals with complex issues related, inter alia, to identity, culture, ethnicity, language, and democracy, in countries which had experienced the colonial domination. This claim is fundamentally opposed by leading postcolonial theorists, postcolonialism as it significantly reduces the scope, historical significance, and complex power relations characterizing the interaction between colonial countries and colonized societies (see Ashcroft et al., 1995). The latter approach favors a holistic over a chronological or culture-specific one. In this respect, postcolonial issues do not emerge *after* the conclusion of the colonial. Postcolonial issues, defined by these power relations, arise instantly at the point of interaction between the colonial power and colonized

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28061-6

populations and happen simultaneously at both ends. Fundamental aspects of such dialectic interaction exist and persist even after the 'conclusion' of the colonial era. As Ashcroft et al. (1995: p. 117) explain

> Post colonialism … does not mean post-independence, or after colonialism, for this would be to falsely ascribe an end to the colonial process. Post-colonialism, rather begins from the very first moment of colonial contact. It is the discourse of oppositionality which colonialism brings into being. In this sense, post-colonial writing has a very long history.

In this excerpt, Ashcroft et al. skillfully identify some of the major antinomies within debates on defining postcolonialism. The first, which I already addressed, deals with the boundaries of postcolonial discourse with regards to the binary interaction between *time* and *space* (i.e., *when* and *where* do 'postcolonial affairs' occur?).

The second theoretical constraint identified by Aschroft et al. is of intellectual eclecticism. As I suggest later, postmodernism, mistakenly considered the only source of epistemological analysis within postcolonialism, has produced some theoretically inconsistent discourses. Some of the most important postcolonial scholars have developed rigorous and robust critiques on the inadequacies of postmodern approaches. Moreover, postcolonial analysis precedes the postmodern one, and thus – both ideologically and epistemologically – distinctions need to be highlighted.

Debates about the representation of oppressed populations and the (in)ability of postcolonialism to give voice to these groups provide interesting examples of the aforementioned tensions. Gayatri Spivak, a celebrated scholar in postcolonial studies, has epitomized the ongoing soul searching within postcolonialism asking the vital question 'Can the subaltern speak?' In her homonymous book, Spivak (1988) explains that any person or group who has become educated through the dominant/mainstream institutions and is able to communicate ideas through conventional 'Westernized' channels of communication cannot claim that they are in a position to represent the oppressed 'native' population. Inspired by Gramscian analysis, she utilized the term 'subaltern' to describe exactly this population of indigenous peoples who live and operate beyond the sphere of Westernized systems and institutions. The question, "who qualifies to give voice to the 'subaltern'?" highlights fundamental political and ethical issues for social workers who often keep for themselves the roles of representation and advocacy. Spivak's critique demonstrates the potential fallacy of postcolonialism according to which the West could use its premises, language, value systems, and research tools while speaking *on behalf of* the 'subaltern.'

Part of the problem of representation is linked to the imaginary, 'mythical,' and biased Western views on understanding and describing the *other*. Edward Said's work on this domain has been of unparalleled importance. In his seminal work *Orientalism*, Said (1978) produces a discourse analysis of the Western impressions of the 'Orient.' He focuses on the nineteenth and twentieth centuries and mostly refers to the Arabic populations of the Middle East. Said's (1978) work eloquently demonstrates how the romanticized and prejudiced Western ideas of the 'Orient' have been historically part of the broader imperial agenda. In this sense, art, culture, literature, and aesthetics can play a major role in carving a certain image about oppressed populations and eventually justifying certain geopolitical interventions based on such prejudices. According to Said (2008: p. 8), "orientalism is never far (…) from a collective notion identifying 'us' Europeans against all those 'non-Europeans' and indeed it can be argued that the major component in European culture is precisely what made that culture hegemonic both in and outside Europe: the idea that European identity is a superior one in comparison with all the non-European peoples and cultures."

A Brief Reflection on the Colonial and Imperial History of Social Work

Within social work literature and practice, there is a noteworthy contradiction. Even though social work in many parts of the world has been a well-calculated product of colonial and imperial conditions, only a fraction of mainstream debates, literature, and practices has dealt with these serious issues. Questions about the imperialist nature of social work, emerged relatively early in the history of international social work, but never gained prominence. Postcolonial and antiimperialist analyses have existed in the margins of social work literature (see Dominelli, 2012; Midgley, 1981). A recent reemergence of these debates has been driven by mainstream postmodern analyses on culture and identity issues and largely focuses on the indigenization agenda. As we discuss below, this debate is rather narrow in scope and predominantly focuses on First Nations in Canada, Aboriginals in Australia, Maori in New Zealand, and Native Americans in North America, excluding – at least in Anglo-American literature – a very diverse history of indigenous movements in other parts of the world, which has engaged with broader alliances, class struggles, and radical theories. There are, of course, some notable exceptions to this rule such as Dominelli (2012), Ferguson et al. (2004), Lavalette and Ioakimidis (2011).

Much of social work history is inextricably linked to the emergence and perpetuation of the broader politics of colonialism and imperialism. The process of international social work development emerged in two main chronological waves. During the 1930s and early 1940s social work was imported to countries experiencing the British colonial rule. People in India, Cyprus, and the West Indies witnessed the emergence of a 'new' and rather 'alien' profession, which focused its marginal work on poor families and 'antisocial behavior' of children and young people. The development of social work at the time seemed peripheral and it was detached from any notion of adequate welfare state similar to that debated and eventually developed in the UK. In a climate of broader political repression, colonial social work was largely influenced by the toughest Victorian principles of social order and moralism (Ioakimidis, 2011; Midgley, 1981), with the forms of social work introduced at that time, a caricature of the metropolitan British system. The lack of a welfare system – even in an embryonic stage – the absence of democratic institutions and the violence of colonial rule allowed little space for the

development of a social work relevant to local needs. In Cyprus, for example, only the most punitive and prescriptive dimensions of British social work and youth probation systems were imported from the UK (see Triseliotis, 1977). Other, more inclusive and universal forms of welfare provision never made it to this island.

The second wave of international social work development emerged in the immediate aftermath of World War II. The creation of a global 'bipolar' power system and the subsequent tension between the main superpowers initiated a long period of antagonism known as the 'Cold War.' The United States of America immediately introduced a double-edged foreign policy aiming at tightening control over its proclaimed spheres of influence while actively suppressing the 'threat' of communism at a global level. In pursuing this aggressive policy, the US not only utilized its overwhelming military might, but also promoted economic, cultural, and social domination (see Hoogvelt, 2007). The need for postwar reconstruction in Europe and other parts of the world seemed to be a unique opportunity for the USA to extend its imperial control and win 'hearts and minds' of the people within its sphere of influence. The gigantic Marshal Plan in Europe and the manifold operation of United Nations Relief and Rehabilitation Administration and United States Agency for International Development in the Middle East, Africa, and Asia provide characteristic examples of imperialism through 'humanitarian assistance.'

Social work at the time seemed a very important tool in expanding and promoting this agenda. American social workers were sent to countries of US interest as part of missions that – not infrequently – included military personnel and policy technocrats. Greece and Turkey were the first countries to witness such top-down social work development as part of the postwar imperial project. In fact, in Greece, social work emerged in the middle of a vicious Civil War between the US backed National army and the socialist Democratic army. As Ioakimidis (2011: p. 516) indicates

> the main paradox of social work is that at times it can be both humane and also lacking in compassion or a sense of humanity. In the Greek case, international social work had no reservations in working closely with an authoritarian regime. The idea that social work can be used as a sophisticated and professionally advanced method of making the dominant socio-economical system operate more efficiently has long informed the development of international social work.

Korea, Vietnam, Italy, and Latin America provide similar examples of countries or regions where social work development was wedded to the specific agenda of US imperialism.

The Social Work Indigenization Question

An inevitable outcome of the aforementioned political dynamics was the creation of a colonial social work, which was, at best, irrelevant to local populations and, at worst, was explicitly oppressive and often racist. Shaped by these colonial interests, social work had to engage with extremely oppressive institutions and regimes; notwithstanding its much celebrated commitment to human rights. Racist incarceration and segregation systems (South Africa), harsh and politically motivated child care and adoption policies (UK–Australia), collaborations with dictatorial regimes (Greece, Chile, Argentina) are only few examples of the unpalatable dark and colonial side of social work history.

As national liberation movements and the movement of 'Third Worldism' (see Worsley, 1967), gained momentum in international politics in the early 1970s, anticolonial approaches became increasingly influential in social sciences. A major strand within this realm was the indigenization movement (Atal, 1981), which despite its original dynamism, by the beginning of 1980 had lost its momentum. Within social work, however, indigenization retained some impetus and during the past three decades and reemerged as a central theme in the late 1990s.

One of the reasons, behind the interest in social work indigenization is the very nature of the profession in the developing world and its use as a tool for humanitarian and political intervention. Despite the generous international funding that accompanied the creation of social work in much of the developing world as well as the constant 'scholar support' from Anglo-American academics, local welfare officers immediately recognized the tensions, contrasts, and ultimately the unfeasibility of the new profession. In order for social work to be established and successful in the long term its content and scope had to be relevant to the local context and needs. Antiintellectualism, limited academic debate, marginal job opportunities, and professional unrecognizability are only persistent symptoms of the artificial, alien, and 'top-down' construction of social work in these regions.

Within this context, it does not come as a surprise that one of the first recorded discussions of social indigenization was written by a United Nations regional welfare officer operating in Africa. In this early analysis, Shawky (1972) identified the failures in transferability of social work expertise. Unless 'adapted to fit local needs,' social work could be of little benefit for the overall process of modernization in the developing world. Therefore, he suggested that the United Nations as well as the newly developed social work profession should ensure that the imported technical knowledge is tailored in a meaningful fashion for the indigenous populations. Four years later, Resnick (1976: p. 22), adopting a rather similar approach, described indigenization as "(the) process of relating social work function and education to the cultural, economic, political and social realities of a particular country."

It was not until 1981 that the first detailed and well-documented critique of the colonial, and thus potentially oppressive, nature of international social work emerged. Midgley's (1981) seminal work on the reality of professional imperialism embedded in much of international social work activities still remains a key text in understanding the antinomies of international social work. In this book, Midgley (1981) scrutinized the – rather limited at the time – international social work literature and utilized empirical data from personal communications with social workers in Africa. Midgley, demonstrated the inadequacy and irrelevance of social work curricula in the developing world. Most importantly, he moved beyond the descriptiveness and

antiintellectualism of mainstream international social work literature attempting to demonstrate the importance of the sociopolitical context within which social work operates. His most important achievement and contribution to social work scholarship was the ability to highlight the clear connections between international social work and the broader geopolitical context.

At roughly the same time, a much more politicized discourse emerged in Latin America. Latin American academics and students, influenced by the success of regional revolutionary movements, adopted a Marxist-socialist critique on the nature and character of social work. At the core of this critique was the interconnection between geopolitical imperialism and the international social work project. In this sense, Latin American academics broadened their analysis, dismissing much of the previous 'indigenization' discussions as developmentalist and abstract. According to their analysis, what the social work needed was not only a departure from its imperialist past but the creation of a new '*authentic*' Latin American theory and practice; one that meets the local needs of oppressed classes and promotes the idea of revolutionary change in society. Wolfe (1972) tried to capture this dynamic debate within Latin American social work explaining that

> There is apparent, a generalized and serious attempt on the part of university schools and professional associations to develop social work theories with authentic Latin American orientation and meaning, and to discover from their practice, which is rich but never as yet documented nor analyzed nor conceptualized, techniques more applicable to the problems under the unique situations of the countries of the region.
>
> Wolfe, 1972: p. 33

The idea of 'authentization' dynamically reappeared in social work literature the in late 1980s by Walton and Abo El Nasr. In their analysis of Egypt, the authors (1988) included some of the elements of the Latin American argument for liberation from both the orthodoxies of international social work and the triviality of the indigenization debate. Moreover, they suggested that there was a further need for 'authentization' of social work, which goes beyond simply 'modifying' and adapting Anglo-American social work. Walton and Abo El Nasr argued that

> The pattern of social work is modified by the nature and circumstances of Egyptian society. Also we have to search or look for the authentic roots in the local system to create or build a domestic model for social work in the light of the needs, problems and circumstances prevailing in Egypt (1988:143).

While highlighting the need of searching and documenting social work practices that are relevant to local cultural needs, Walton and Abo El Nasr, distanced themselves from the Latin American paradigm of politicized social work. Their analysis adopted developmentalist approaches, which failed to address the big political questions of inequality at a domestic level and imperialism at an international level, and despite its originality and importance, implied a rather apolitical social work process.

Nevertheless, Walton and Abo El Nasr's references to the importance of local culture were to be further appreciated and explored toward the end of the twenty-first century. A rise in academic interest on the contested theme of First Nations, along with the rapid expansion of social work in China, renewed the discussion about the nature of international social work theory, practice, and educational exchanges (see Yunong and Xiong, 2008). This was a particularly fertile soil for the growth of postmodern analyses in the field of social work.

The nature, epistemology, and scope of postmodern approaches suggested that the debate on social work should drift away from the, vaguely defined, politics of 'modernity' and focus on culture, identity, and diversity. 'Eurocentrism' along with the 'grand narratives' of Modernity (an equally obscure and problematic jargon) were deemed irrelevant to the experiences of the 'local populations.' Decolonizing social work meant that social work had to adopt a strategy for 'the liberation project of Indigenous Peoples and a call to liberate social work from its modernist shackles' (Briskman, 2008: 83). However, as we will see below, this call was rather confusing and abstract as it neither specified the 'constraints of modernity' nor acknowledged the structural inequalities within 'indigenous societies.' Interestingly, part of the criticism against the indigenization of social work came from China. Yunong and Xiong (2008) in a well-argued response against mainstream postmodern approaches highlighted that proponents of the indigenization agenda blindly adhere to existing indigenous structures, ignoring whether such structures advance the betterment of the people or not. Moreover, they suggest that in a rapidly globalized world a fixation with locality is rather monolithic and unable to grasp the dynamism and complexities of the interconnected global and local political debates.

As epistemological, ideological, and theoretical tensions on the issue of indigenization rise, the social work field can only benefit from this rigorous examination of some of the most persistent and inherent contradictions of the profession.

Rethinking 'Indigenization': Toward an International Social Work of Liberation

Discussions on the oppressive nature of 'top-down' Anglo-American social work practice in the 'developing world' are very important in highlighting the need for alternative theories and practices. However, as I also explained in this article, the narrowness of the social work indigenization debate has not allowed for the exploration of concrete theoretical alternatives, which can address the fundamental and structural inequalities caused by neoliberal capitalism and imperialism.

Part of this problem, is linked to the distinctive theoretical relativism and obscure language, which is so characteristic of postmodern analyses. Claiming that social work should be liberated from the restrictions of modernity and challenge all 'fundamental truths of Modernity' might sound as a call for resistance but in reality it is rather call to apathy. 'Are social-political and economic relations of modernity a thing of the past?' is a fundamental question that much of the indigenization debate has failed to address. Modernity has never been a well-defined term within social work debates. Institutions and systems that were prevalent in this ill-defined epoch are still operational and dominant. Capitalist modes of production and the geopolitics of imperialism are still driving forces

behind oppression, exploitation, and war. To claim that these systems do not exist or that they belong in a past era, can only lead to further confusion with regards to the nature of oppression and discrimination within 'indigenous' peoples experiences. As Ferguson and Lavalette (1999: p. 29) explain

> It is not at all clear why the absence of 'universal truths' should be celebrated. For if all dis- courses are equally valid, then presumably there is no basis for distinguishing between the account of the racist to his black victim, of the abused wife to her abusing partner, or of biomedical psychiatry to a mental health users' movement.

Uncritical celebrations of the 'cultural differences' over structural systemic inequalities, do little justice to the local and global movements, which engage with the struggle against neoliberalism. It would be an oversimplification merely to suggest that international social work was imposed from colonial and imperial powers and as a result oppresses all sections of local populations *equally*. In fact, most of the mainstream international social work literature unwillingly indicates the opposite; colonial social work was welcomed by certain social classes of the population and loathed by others. For example, Ioakimidis (2011) explains how the import of social work in Greece received active support from the Greek ruling classes and at the same time it was used as a tool to suppress and control the working classes.

> The Greek state welcomed and actively supported the 'transplantation' of the new profession. Queen Frederica personally adopted the project, becoming the main patron of the profession's development (…). Systematic efforts of local and international patrons to transplant Anglo-American knowledge have been historically combined with eradicating existing grassroots welfare networks' (2011: p. 508).

Interestingly, much of social work interventions, which were used within this oppressive process were celebrated as 'indigenous' and 'localized.'

Therefore, the question of class divisions and the role of state interaction with colonial or postcolonial societies are of the outmost importance and cannot be ignored. Most of these societies also experience sharp internal divisions primarily related to social class and race/gender. Celebrating them as 'united in their cultural uniqueness' would be a gross idealization and oversimplification of the complex sociopolitical antagonism embedded in these societies. Such divisions became deeper during the colonial/imperial era while local elites were nurtured by global imperialism and the metropolis of Western capitalism. Many states created in postcolonial period often mimicked the colonial power relations simply masking oppression under the veil of nationalism or 'indigenization' (see Midgley, 1981).

Within this context, social work has been used as a tool to promote specific state-sponsored policies, which are often oppressive and stigmatizing. Jordan, Saudi Arabia, and Egypt offer characteristic examples of such contradictions. International social work literature routinely celebrates the advancement of professionalism and localization claiming that the culturally relevant forms of social work in these countries has produced a much needed and progressive profession (see Al-Makhamreh and Sullivan, 2012; Villa, 2011). However, such literature primarily focuses on the cultural aspects or the need for 'professionalization' and it ignores the vast inequalities within these countries reflected on the ongoing struggle for change, which has inspired the Arab Spring.

Despite the 'evolution' of imperialist intervention, its core mission still remains intact: geopolitical struggle for influence and exploitation of global markets, natural resources, and workforces. Over past decades, the emphasis of imperial interventions has shifted away (but not completely) from open military invasions to a more sophisticated form of violence; this of forced 'liberalization of economies' and free market reforms. As Callinicos (2009: p. 9) explains

> the imperialism of free trade is exactly what the EU, as well as the US, have been pressing on other countries, particularly since the establishment of the World Trade Organization.

Inevitably, First Nations cannot be seen as unaffected from this process. The idea that, somehow, culturally relevant social work can facilitate the 'liberation' of these populations is futile unless we identify the nature of the institutions and systems from which oppressed groups should break free. 'Naming the beast' has long been a taboo issue within social work literature. However, without clearly exposing the widening inequalities and stark injustices, which are a direct product of neoliberal capitalism, social work would be powerless in the struggle for the liberation of the oppressed peoples; part of which are First Nations communities. A focus on 'culture' and 'diversity' only obscures the common experiences of the working classes and oppressed groups. Creating indigenous 'islands of independence' within capitalist and oppressive states is an impossible mission. The example of Latin America indicates that an isolated indigenous movement cannot prevent capitalist expansion in to the land, natural resources, culture, and autonomy of the First Nations. The struggle of the indigenous people of Chiapas in Mexico provides a powerful case. Zapatista Army of National Liberation (EZLN; Zapatistas) have long fought for the autonomy of the indigenous people, attempting to create a space free from state oppression. However, after a long struggle for 'indigenous autonomy' the iconic leaders of the movement admitted that the struggle for indigenous liberation cannot be isolated from the broader movements within Mexico. In 2005, Zapatistas announced a change in their strategy for liberation. In a declaration known as the 'Sixth Declaration of the Selva Lacandona' they declared that a main priority for the movement should be the creation of broader national and international alliances in order to facilitate global anticapitalist struggle.

> We are going to continue fighting for the Indian peoples of Mexico, but now not just for them and not with only them, but for all the exploited and dispossessed of Mexico (...) we are going to go about building, along with those people (…) a program of struggle, but a program which will be clearly of the left, or anti-capitalist, or anti-neoliberal, or for justice, democracy and liberty for the Mexican people.
>
> EZLN, 2006 http://enlacezapatista.ezln.org.mx/sdsl-en/

This example vividly reveals that many of the problems indigenous people face are shared by millions of other oppressed people across the world, but is often ignored by 'indigenization' advocates. In many cases, global issues (poverty, inequality, environmental disasters) affect both social workers and the people we work with alike. Dominelli (2012) has categorically suggested that social work cannot remain indifferent and apathetic to neoliberal policies, which result in continuous oppression, major injustices, and irreversible natural disasters; issues that need to be addressed as interconnected and not isolated phenomena.

Ferguson and Lavalette (2006: p. 315) have demonstrated that engagement with the broad anticapitalist movement can provide social work with a great opportunity to rethink social work theory and practice and most importantly demonstrate that 'another social work is possible.'

> The injunction from the anti-capitalist movement to think global, act local provides a fresh basis for social workers to actively make the connections between the neoliberal global agenda and their day-to-day experience as social work teachers, workers and clients, and also provides us with the basis for a new, global social work theory and practice.

Historically, social work has been characterized by a major contradiction. It has been developed as means of care for the most vulnerable people in our societies while being used as a tool for social control, targeting the working classes and demonizing the poorest people. This contradiction, which is also evident in much of international social work, indicates that claims about social work as a profession *inherently and unconditionally* committed to social justice are only empty words unless such language leads to specific and tangible actions. Recognizing this dual nature of social work is only the first step in engaging with emancipatory practice. Developing alliances and *acting* for social change is the next and foremost one.

See also: Afrocentric Approaches to Social Work; Colonialism, Anthropology of; Globalization and Transnational Corporations; Globalization and World Culture; Globalization, Anthropology of; Globalization: Welfare and Social Work; Imperialism, History of; Imperialism; Indigenous Knowledge and Intellectual Property Rights; Indigenous Psychology; Indigenous Rights; Orientalism; Postcolonial Geography; Postcoloniality; Social Justice: Historical and Theoretical Considerations.

Bibliography

Al-Makhamreh, S., Sullivan, M.P., 2012. Professionalising social work in the Hashemite Kingdom of Jordan: processes and implications. European Journal of Social Work. http://dx.doi.org/10.1080/13691457.2012.724385.

Ashcroft, B., Griffiths, G., Tiffin, H., 1995. The Post-Colonial Studies Reader. Routledge, New York.

Atal, Y., 1981. The call for indigenization. International Social Science Journal 33 (1), 189–197.

Barlow, Tani (Ed.), 1993. Colonial Modernity. Duke UP, Durham, NC.

Briskman, L., 2008. Decolonizing social work in Australia: prospect or illusion. In: Gray, M., Coates, J., Yellow Bird, M. (Eds.), Indigenous Social Work around the World: Towards Culturally Relevant Education and Practice. Ashgate, Aldershot, pp. 83–93.

Callinicos, A., 2009. Imperialism and Global Political Economy. Polity Press, Cambridge.

Dominelli, L., 2012. Green Social Work. Polity Press, Cambridge.

EZLN, 2006. Sixth Declaration of the Selva Lacandona. http://enlacezapatista.ezln.org.mx/sdsl-en/ (accessed 10.02.13.).

Ferguson, I., Lavalette, M., 1999. Social work, postmodernism, and marxism. European Journal of Social Work 2 (1), 27–40.

Ferguson, I., Lavalette, M., Whitmore, B. (Eds.), 2004. Globalisation, Global Justice, and Social Work. Taylor and Francis, London.

Freire, P., 1970. Pedagogy of the Oppressed. Herder and Herder, New York.

Gray, M., Coates, J., 2008. From indigenization to cultural relevance. In: Gray, M., Coates, J., Yellow Bird, M. (Eds.), Indigenous Social Work Around the World: Towards Culturally Relevant Education and Practice. Ashgate, Aldershot, pp. 13–29.

Hobsbawm, E., 1975. The Age of Capital. Cardinal, London.

Hoogvelt, A., 2007. Globalisation and imperialism: wars and humanitarian intervention. In: Dominelli, L. (Ed.), Revitalising Communities in a Globalising World. Ashgate, London, pp. 17–42.

Ioakimidis, V., 2011. Expanding imperialism, exporting expertise; international social work and the Greek project (1946-1974). International Social Work 54 (4), 505–519.

Lavalette, M., Ioakimidis, V., 2011. Social Work in Extremis, Lessons for Social Work Internationally. Policy Press.

Madsen, D.L., 1999. Post-Colonial Literatures: Expanding the Canon. Pluto Press, London.

Midgley, J., 1981. Professional Imperialism: Social Work in the Third World. Heinemann Educational Books, London.

Resnick, R.P., 1976. Conscientization: an indigenous approach to international social work. International Social Work 19 (2), 21–29.

Shawky, A., 1972. Social work education in Africa. International Social Work 15 (1), 3–16.

Spivak, G., 1988. Can the subaltern speak? In: Nelson, C., Grossberg, L. (Eds.), Marxism and Interpretation of Culture. University of Illinois Press, Urbana, pp. 271–313.

Sullivan, M.P., Forrester, D., Al-Makhamreh, S., 2010. Transnational collaboration: evaluation of a social work training workshop in Jordan. International Social Work 53 (2), 217–232.

Triseliotis, J., 1977. Social Welfare in Cyprus. ZENO, London.

Villa, R., 2011. The winds of change. Social Dialogue 1 (1), 20.

Walton, R., Abo El Nasr, M., 1988. Indigenization and authentization in terms of social work in Egypt. International Social Work 31 (2), 135–144.

Wolfe, V.A., 1972. The dilemma of Latin American social work. International Social Work 15 (3), 48–56.

Worsley, P., 1967. The Third War. Weidenfeld and Nicholson.

Yunong, H., Xiong, Z., 2008. A reflection on the indigenization discourse in, social work. International Social Work 51 (5), 611–622.

Neoliberalism

Vishanthie Sewpaul, School of Applied Human Sciences, Howard College Campus, University of KwaZulu Natal, Durban, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article teases out the concept of neoliberalism as ideology, as a form of governmentality, and as a policy paradigm. It provides a brief history of neoliberalism and then discusses the consequences of neoliberalism in respect of unfair trade, the destruction of domestic production and markets, poverty, and inequality. It concludes with the argument that despite the ideological hegemony of neoliberalism, the current crisis and trends of global social justice movements, which social workers can be engaged in, hold potential to create alternative socioeconomic and political systems in support of a more just world.

Introduction

Neoliberalism has its ideological roots in classical liberalism articulated by early theorists such as John Locke, Adam Smith, and David Ricardo and in the twentieth-century theories of Friedrich von Hayek and Milton Friedman (Harvey, 2005; Pierson, 1991; Steger and Roy, 2010). The neoliberal revolution was consolidated during the regimes of Margaret Thatcher in the United Kingdom and Ronald Reagan in the United States from the early 1980s onward. A related significant event during this period was the opening up of the economy of China, the world's emerging superpower, by Deng Xiaoping in 1978. According to Harvey (2005), events in the United Kingdom, the United States, and China during this period "spread and reverberated to remake the world around us in a totally different image" (p. 1). The breakdown of Soviet communism also played a decisive role in entrenching neoliberalism as the dominant ideology. The past few decades have witnessed the increasing dominance of neoliberalism as discourse and practice on a global level, which has remolded health, education and welfare, and global flows of finance, trade, and labor (Abramovitz and Zelnick, 2010; Schram and Silverman, 2012; Sewpaul, 2008; Rowden, 2009).

Classical liberalism is characterized by a global laissez-faire that holds the market to be sacrosanct, self-regulating, and entirely rational. This was the appeal of Adam Smith's idea that "markets and the pursuit of self-interest would lead, as if by an invisible hand, to economic efficiency" (Stiglitz, 2006: p. xiv). Referring to contemporary neoliberalism, the collapse of markets; the implosion of the American financial system, beginning with the subprime mortgage crisis that began in August 2007; and the bail-out of banks, Gray (2009) reiterated his belief of 1998 that: "Today's regime of laissez-faire will be briefer even than the *belle époque* of 1870 to 1914, which ended in the trenches of the Great War" (p. xii). Obama's (2009) inaugural address in which he warned that, "without a watchful eye, the market can spin out of control" also made some people believe that the end of neoliberalism was imminent. Yet, the "world failed to shake the 'Washington Consensus'" (Bond, 2010: p. 59) – a consensus between the International Monetary Fund (IMF), the World Bank (WB), and the United States Treasury that fostered fiscal austerity, privatization, and market liberalization (Stiglitz, 2006) that are defining features of neoliberalism. However, there is now hope as the majority of the world's dispossessed (symbolically represented as the 99%) is reacting to global capitalism as seen in the recent Arab spring, the Occupy movement, and protests in many parts of the world. This article teases out the concept of neoliberalism, provides a brief history of it, and then discusses the consequences of neoliberalism. It concludes with the argument that despite the ideological hegemony of neoliberalism, the current crisis and trends of global social justice movements hold potential to create alternative socioeconomic and political systems in support of a more just world, which constitutes a central goal of social work education, research, and practice.

Conceptualization of Neoliberalism

Neoliberalism refers to a combination of socioeconomic and political discourse and policy choices based on the values of an unregulated market; the rarefaction of individual freedom and choice; as well as the faith that market fundamentalism, and not state intervention, would promote economic growth, efficiency, progress, and distributional justice, primarily through trickle-down effects. Harvey (2005) defines neoliberalism as "a theory of political economic practices" (p. 2), rather than an unadulterated political ideology. Although neoliberalism has its roots in classic liberalism, it has come to span diverse political leanings and is associated with center-left politics, moderate right-wing conservatism, and autocratic dictatorships, embraced by and/or imposed on almost all countries across the globe. The author draws on the works of Larner (2000) and Steger and Roy (2010) to conceptualize neoliberalism as an ideology, as a form of governmentality, and as a policy paradigm. These dimensions are mutually constitutive and overlapping.

Neoliberalism as Ideology

Duménil and Lévy (2011) maintain that the neoliberal revolution was engineered by class and imperialist objectives, serving the interests of a privileged minority and a few powerful countries, with the United States occupying a unique

hegemonic power. They argue that ideology was not the engine of the neoliberal revolution but that "the hegemony of the upper classes was deliberately restored […] A neoliberal ideology emerged, the expression of the class objectives of neoliberalism" (p. 18). Whether a precursor to and/or an outcome of the neoliberal revolution, ideology has come to play a huge role in its reproduction. While the crisis that began in 2007, with the crash in October 2008, held the potential to create cracks in the consciousness of people about the moral and pragmatic limits of a system regarded as inviolate (Sewpaul, 2008), the opportunities seem to be lost. Although Stiglitz (2006) claimed that there is little intellectual defense of market fundamentalism left, a claim supported by a reading of the vast body of literature that almost unanimously speaks of the negative consequences of neoliberalism, there is yet a marked failure to disrupt the Washington Consensus. Given the power of ideology, the disjuncture between popularly held neoliberal values and empirical data, and the lack of political debate on neoliberalism in mainstream politics, this is understandable.

Fortunato Jr. (2005) described the "manufacture of consent" (p. 4) by the mass media that serves the interests of the corporate world and political elites and that ensures that capitalism succeeds through ideological control of consciousness designed to make us believe that neoliberalism is in our interests and is inevitable. Neoliberalism was engineered and gained ground primarily on account of the lure of the language of liberal theory with its emphases on individualism, ownership, choice, flexibility, and competition that mask "the grim realities of the restoration or reconstitution of naked class power" (Harvey, 2005: p. 119). It is not only the manufacture of consent that is created by the media but also the "manufacture of desire" (Leonard, 1997: p. 37) as the market seduces people into constructing themselves as good consumers where their moral worth is determined by their ability to make the right purchasing choices (Bauman, 1992). Bauman (Bauman and Tester, 2001: p. 87) speaks of the "disguise of the individual's freedom of choice," which is so taken for granted that it "seldom has a chance of being examined and questioned […] an old sinister temptation […] left to be exploited by commercial markets." These views are suggestive of Gramsci's (1977) thesis of ideological hegemony exercised by the state and civil society to control the consciousness of people, not through repression or coercion but by gaining popular consent in the reproduction of class relations.

While most theorists are quick to comment on the dangers of neoliberalism, few are willing to critique the liberal democracy that underpins it. There is a general taken-for-granted assumption of a convergence between the market and democracy, with American liberalism and capitalist hegemony touted as the source of morality and democratic practice, which is furthest from the truth (Amin, 2001; Fortunato Jr., 2005; Harvey, 2005). Samir Amin (2001), a postcolonial writer, cogently raises a number of critical questions in relation to liberal democracy and neoliberalism, concluding that "Convergence theory – the notion that the market and democracy converge – is today pure dogma with much of the talk about democracy reflecting the imposition of policy makers who have usurped power in the United States" (p. 9). Gray (2009) declares that "democracy and the free market are competitors rather than partners" (p. 213), a view supported by Stiglitz (2006) who proclaims that neoliberal globalization is not consistent with democratic principles.

Citing several groups of power elites that serve as the main advocates of neoliberalism Steger and Roy (2010: pp. 11–12) argue that: "These individuals saturate the public discourse with idealized images of a consumerist, free-market world [...] neoliberal decision-makers function as expert designers of an attractive ideological container for their market-friendly political agenda."

Neoliberalism does not get reproduced in a vacuum. We are all complicit. "We all consume […] we all participate in hierarchies of race, class gender and privilege. No one is a pure victim in this economic system, though almost everyone is ultimately a loser" (Haiven, 2011: p. 1). Harvey (2005) describes how neoliberalism penetrates daily consciousness so much so that it is normalized and naturalized, and it is considered necessary for social order despite the gross inequalities and poverty that it engenders.

Neoliberal Governmentality

The treatise on neoliberalism as a mode of governance borrows from Foucault's concept of *governmentalities* that provides a means of understanding the relationships between knowledge, power, and technologies of the self (Foucault, 1984; Leonard, 1997; Lemke, 2002). From this perspective, neoliberalism refers to "a political rationality that tries to render the social domain economic and to link a reduction in (welfare) state services and security systems to the increasing call for 'personal responsibility' and 'self-care'" (Lemke, 2002: p. 203). While neoliberalism guarantees personal and individual freedom, each person is held responsible for his or her own actions and well-being. Thus, poverty and hardships are seen as personal failings rather than being attributed to any structural barriers to unemployment and the consequences of exclusions based on race, class, caste, gender, and/or disability. Steger and Roy (2010) assert that:

> A neoliberal governmentality is rooted in entrepreneurial values such as competitiveness, self-interest, and decentralization [...] Rather than operating along more traditional lines of pursuing the public good (rather than profits) by enhancing civil society and social justice, neoliberals call for the employment of governmental technologies that are taken from the world of business and commerce. (p. 12)

Thus, along with neoliberalism, there is entrenching of new managerialism into all spheres of life, including social work and welfare. New managerialism emphasizes getting the job done, with checks and balances in the shortest space of time as well as with the imposition of economy, efficiency, and effectiveness manifesting in such tangibles as performance measurement, inputs and outputs, outcomes, and targets (Dominelli, 1996). One example of this, in the South African context, is the Financing Policy for Welfare, which speaks of business plans, contracts, affordability, efficiency, outputs, performance audits, outsourcing, venture financing, and service purchasing, thus effectively reconstructing users of welfare services as customers (Sewpaul and Hölscher,

2004). This trend, replicated in many parts of the world, found its origins in Britain and the United States. The 1989 Community Care White Paper in Britain (cited in Sewpaul and Hölscher, 2004) promoted choice, flexibility, and innovation, which was to be achieved through a prospering private welfare sector and outsourcing of government services. The introduction of competition between providers is aimed at increasing value for money, efficiency, accountability, and quality of the services provided. Banks (2011), writing within the British context, observes an intensification of "procedures, measurement and centrally defined targets" (p. 10) and the adoption of depersonalized ethics that focus on predetermined standards, contracts, and procedural manuals. New managerialism relegates the importance of other ethical values such as autonomy, critical reflection, care, equality, solidarity, interdependence, reciprocity, respect, and trust in favor of narrowly defined economic priorities (Bottery, 2000).

Neoliberalism as Policy

As a policy package, neoliberalism is associated with spread of global capitalism and consumerism that has engineered the attacks on, and dismantling of the welfare state (Chomsky, 1999; Harvey, 2005; Steger and Roy, 2010), expressed in what Steger and Roy (2010: p. 14) call the "D-L-P formula," which is deregulation of the economy, liberalization of trade, and privatization of state assets, which provided the ideological basis for the restructuring, privatization, and retrenchment of social policy and welfare programs, and for managerial approaches to social work practice (Clarke, 2007). Neoliberalism is a discourse and practice, which rarefies individual interests where transnational corporations (TNCs) and a few, rich powerful elites dominate the market primarily interested in profit making. Following this individualistic logic, the state uses neoliberal socioeconomic policies to allow it to abdicate its responsibilities in relation to the welfare of its people. Evocative of Foucault's (1984) governmentality and technologies of the self, the most dominant neoliberal discourses in social work and welfare have been on the promotion of self-reliance, the importance of individuals and families taking responsibility for their own well-being, and shifting responsibility from the state to local communities. This neoliberal policy doctrine is reflected in the voice of a manager in the South African context: "The State is an entrepreneur of its own. It must make a profit where it can […]. The state must be minimalist, it must really do the least and the last. Civil society, empowered civil society is to do the most" (Sewpaul and Hölscher, 2004: pp. 84–85), a doctrine that had become quite universal in the 1990s, reflecting the hegemony of neoliberalism. Apart from the use of ideological manipulation that engineered this, there was the use of military force and sanctions, or threats of sanctions by the United States against governments in the developing world that deviated from the US ideal of liberal capitalist democracy (Harvey, 2005; Sewpaul, 2006; Steger and Roy, 2010). In some countries like the United Kingdom, there continues to be the strategic use of budgets and legislation, thus avoiding political discourses on the nature of the welfare state and of social work.

Neoliberalism: A Brief History

There are three distinct periods of lasting and deep structural crises in the history of modern capitalism: the crisis of the 1890s, the Great Depression, and the crisis of the 1970s, with the contemporary crisis of neoliberalism ushering in a new phase (Duménil and Lévy, 2011).

Mid-Victorian England saw the demolition of social markets (which combined a regulated market with issues of social justice, with, for example, universal health care and extensive social security) and the introduction of deregulated, free markets that functioned independently of social needs, based on the theories of classical liberals like Adam Smith and David Ricardo (Gray, 2009; Steger and Roy, 2010) who were convinced that economic difficulties were the consequence of government failures, particularly from state interference. People were seen to be creatures of economic rationality and, in accordance with modernist thinking, as opposed to premodern conceptions of an unquestionable divine order (Howe, 1994), evolved the idea of the human being as subject, "an autonomous self who is the source of his or own action" (Sewpaul and Hölscher, 2004: p. 18), thus imposing no duty on the part of the public to provide, or any right on the part of the destitute to claim any relief.

Gray (2009) argues that the removal of agricultural protectionism, free trade, and the Poor Laws of 1834 formed the cornerstones of the unregulated market of mid-Victorian times that has influenced all subsequent neoliberal policies. The Poor Law Act of 1834 was one of the most demeaning and stigmatizing pieces of legislation based on *laissez-faire* ideology, reflected in the activities of the nineteenth-century charities, particularly that of the Charity Organisation Society whose activities "rested on a belief in moral character as the decisive feature in people's circumstances – no structural issues of power, unemployment or marginalization entered the moral calculus" (Clarke, 1993: p. 9), hence "individuals were solely responsible for their own welfare" (Gray, 2009: p. 9). The *laissez-faire* ideology reduced the role of the state to that of "night watchman" (Steger and Roy, 2010: p. 5) where the state takes responsibility only for the most basic aspects of public order, and it entrenched casework as the basis of social work while marginalizing the more structural approaches to the profession, as reflected in the work of the Settlement House Movement.

This *laissez-faire* governance was legislated out of existence from the 1870s onward and, by the end of World War I, markets had become, in the main, reregulated in the interests of public health and welfare. However, individualism and free trade characterized Britain until the Great Depression of the 1930s that paved the way for Keynesian economics underscored by egalitarian or embedded liberalism, which held sway from the mid-1940s until the 1970s (Gray, 2009; Stiglitz, 2006; Steger and Roy, 2010). This was not the product of a natural evolution. Leonard (1997) posits that the political masses played a central role in establishing the Keynesian welfare state in the West, including the United States, although working class consciousness was weaker and more transitory there than in the social democracies of Europe. Keynesian egalitarianism involved state intervention, regulation of the market, the involvement of organized labor to promote full

employment and growth, and some state ownership of crucial national enterprises like railroads, public utilities, and energy companies (Keynes, 1933). Keynes in 1933 (unpaged) argued that, "Ideas, knowledge, science, hospitality, travel – these are the things which should of their nature be international. But let goods be homespun whenever it is reasonably and conveniently possible, and, above all, let finance be primarily national." The salience of his reflections for our contemporary world is expressed in the following:

> Experience accumulates to prove that most modem processes of mass production can be performed in most countries and climates with almost equal efficiency [...]. National self-sufficiency [...] though it costs something, may be becoming a luxury, which we can afford, if we happen to want it.

Karl Polyani also made significant contributions to economic thinking during this period. In 1944, combining issues of economic and environmental justice, he prophetically wrote: "To allow the market mechanism to be the sole director of the fate of human beings and their natural environment [...] would result in the demolition of society" (Polyani, 1957: p. 73). Polyani, like Keynes, believed that sustainable economic relationships needed to be embedded in a framework of regulations, institutions, and social norms.

In the three decades following World War II, Keynesianism, which inspired the American New Deal, the British welfare state, and the Scandinavian Social Democratic welfare model, produced impressive economic growth, low inflation, high wages, and exceptional levels of human well-being and social security. The benefits of Keynesianism, however, eluded most of Africa (Harvey, 2005), which continued to be the biggest loser throughout history and as neoliberalism intensified into the 1990s. Keynesianism came to a halt with the economic crises of the 1970s, with the oil shocks that raised the price of petrol, and induced high inflation and rising unemployment (Harvey, 2005; Steger and Roy, 2010; Stiglitz, 2006). The structural crisis of the 1970s allowed for the reassertion of the dominance of finance (Duménil and Lévy, 2001) revitalizing the earlier creed of classical liberalism under conditions of globalization.

The welfare state in Britain was one of the casualties of the imposition of IMF conditions in 1976 when Britain borrowed US$3.9 billion. The loan came with a heavy price as it introduced policy changes away from full employment and social welfare toward inflation and public expenditure controls (The Cabinet Papers 1915–82). The early 1980s saw the consolidation of first wave neoliberalism, characterized by conservative *Thatcherism* and *Reaganomics*, based on reducing taxes, liberalizing exchange rate controls, deregulation, privatization, and diminishing the power of labor unions, "interlaced with the geopolitical imperative to stop the spread of communism and socialist developmentalism in the Third World" (Steger and Roy, 2010: p. 47), directed primarily at the maintenance of US hegemony through its imperialist roles in the WB, the IMF, and the economies of Latin America, Africa, and Asia (Stiglitz, 2006; Gray, 2009). The introduction of structural adjustment policies and programs, the changes in the US monetary policies toward price stability before full employment, and their associated debt crises produced untold burdens for Third World countries (Stiglitz, 2006; Duménil and Lévy, 2011): "A strategy for development [...] transformed into an actual catastrophe by a decision emanating [...] within leading capitalist countries, with a total indifference for the hardship imposed on the Third World" (Duménil and Lévy, 2001: p. 595). Second-wave neoliberalism has been identified with the policies of Tony Blair and Bill Clinton, that began in the 1990s and sought to strengthen "social solidarity without dropping the neoliberal ideal of market-oriented entrepreneurship" (Steger and Roy, 2010: p. 50), with a corporate-led globalization that entrenched further inequalities within countries and between the North and the South (Sachs, 2005; Saul, 2006).

Despite its success, Keynesian economics was vilified by classical liberals like Von Hayek and Milton Friedman, whose doctrines dominated the 1944 Bretton Woods conference, which saw the introduction of a system of free trade with fixed exchange rates tied to the US dollar and the establishment of the International Bank for Reconstruction and Development, later renamed the WB, and the IMF (Harvey, 2005; Stiglitz, 2006), initially designed to provide loans for Europe's post-war reconstruction but later expanded to fund various industrial projects in developing countries. The General Agreement on Tariffs and Trade (GATT) was established in 1947 in order to enforce multilateral trade agreements. In 1995, the World Trade Organization (WTO) was founded as a successor to GATT.

In the mid-twentieth century, given that national economies were still relatively closed, governments were able to regulate capital movements, which allowed for a period of relative stability and economic growth that enabled the implementation of Keynesian policies that formed the structural basis of the welfare state (Sewpaul and Hölscher, 2004). But following the recession, in 1971, the United States denounced the fixed exchange rate linked to a fixed gold price, a move followed in 1973 by the introduction of flexible exchange rates and the abolition of capital controls by the United States in 1974, followed by Britain in 1979, and by the early 1990s, most of the other 32 Organisation for Economic Cooperation and Development (OECD) countries (Pierson, 1991). The increase in international trade provoked greater financial speculation and fluctuation of exchange rates, with a steady growth of TNCs. These developments altered the economic context within which the welfare state operated in the context of OECD countries, yet they failed to stem the recession that they were aimed at addressing. Pierson (1991) avers that: "Governments throughout the developed West were simultaneously failing to achieve the four major economic policy objectives – growth, low inflation, full employment and balance of trade – on which the post-war order had been based" (p. 145). With monetary policy increasingly being set by the WB, the IMF, and the United States; the ability of capital to move freely; and the growing presence of TNCs that provided capital with an exit option from one national economy to another, the bargaining power of governments was weakened. Yet, the neoliberal agenda was and continues to be a product of strong governments and cannot function without them.

The attack on the World Trade Center on 11 September 2001 led to another phase where the profoundly "anti-democratic

nature of neoliberalism [is] backed by the authoritarianism of the neoconservatives" (Harvey, 2005: p. 205) and where neoliberal discourses are increasingly tied up with the penalization of poverty (Wacquant, 2001) and with antiterrorism and Islamophobic discourses. On 11 September 2001, radical Islamic forces attacked what they considered to be the symbols of economic oppression – the World Trade Centre and the Pentagon in the world's most neoliberal society. Steger and Roy (2010) rightly concluded that "Bin Laden articulated a criticism of neoliberal market globalism that was shared by opponents of neoliberalism across the Left/Right ideological divide – though the principles and horrific methods of al-Qaeda have been unambiguously denounced by the leaders of the global justice movement" (p. 123). Despite growing inequality and crises engendered by neoliberalism, and the global manifestations of resistance against it, neoliberalism endures largely on account of the taken-for-granted assumptions that are reproduced among people on a daily basis.

Consequences of Neoliberalism

The current crises, which is seeing marked decreases in the living standards of people, has reverberating effects in all parts of the globe, as reflected in the crises of Wall Street, the deepening of the Eurozone crisis, the economic slowdown of some of the Asian countries including India and China, and the entrenching of uneven development and greater inequality within and across countries. Duménil and Lévy (2011) contend that, as a class project, neoliberalism has been exceptionally successful as it has worked in the interests of privileged minority capitalist classes with the divides between the rich and the poor having deepened. Neoliberalism has contributed to distorted development between the North and the South, to greater levels of inequality within nation-states, to the further marginalization of women and to a greater feminization of poverty, especially among Black women (Dominelli, 2002; Gibson, 2009; Lara, 2011; Sewpaul, 2005; Stiglitz, 2006; Gray, 2009; Steger and Roy, 2010). The World Commission on the Social Dimensions of Globalization (2004) concluded that the levels of global imbalances were "morally unacceptable and politically unsustainable" (p. x), while some authors like Dominelli (2012) and Nixon (2010) warn that the great threat is also against environmental sustainability. In a survey of 73 countries, the Commission found that apart from South Asia, the United States, and the European Union, unemployment increased between 1990 and 2002, with global unemployment reaching a high of 185.9 million, and that 59% of the world's people were living in countries with growing inequality.

While countries in Africa, Asia, and Latin America have long been the biggest losers within global capitalism with austerity measures imposed as conditionalities of structural adjustment programs (Dominelli, 2008; Rowden, 2009; Bond, 2005; Stiglitz, 2006; Gray, 2009), patterns of capital accumulation and consumption are now becoming more complex, with many of the rich core capitalist countries now being subject to stringent austerity measures. There are new configurations of global power with, for example, a growing Chinese imperialism that challenges the hegemony of the United States. Also, while many companies have been relocating production to low-wage countries for many years, India has become very successful in attracting high-skilled jobs in computer programming and customer services. The North–South divide is giving way to a new international division of labor where the "familiar pyramid of the core-periphery hierarchy is no longer a geographic but a social division of the world economy" (Hoogvelt, 1997: p. xii). The current manifestation of these is the rising levels of unemployment and inequality and the severe austerity measures sweeping across the globe, with the United States and several European countries like Italy, Greece, Portugal, Britain, and France being extremely hard hit. In the United States, between 2000 and 2008 household incomes declined 1%, while corporate profits rose 70%; the gap between the rich and the poor is higher than at any time since 1930; and the rates of poverty and those living without health insurance has increased by 20% (Nixon, 2010). Economic growth in itself does not translate into a better life for the majority of people, and the 'trickle-down' notion of capitalism has proved to be a myth. The International Labor Organization (ILO, 2009) predicted that the crisis would wipe out at least 20 million jobs by the end of 2009, bringing global unemployment to an unprecedented number of over 200 million.

Accompanying neoliberal policy choices are the weakening of trade unions and labor power; tax cuts for big business and the rich; the erosion of meaningful and egalitarian democracy; cutbacks in state expenditure in sectors such as health, education, and welfare; the relocation of labor to areas of least production costs; and the creation of sweatshops and economic processing zones that compromise workers' rights (Gray, 2009; Stiglitz, 2006; Klein, 2000; Sewpaul, 2006; Steger and Roy, 2010). Trade liberalization and unfair trade has often resulted in massive imports, which undermine local production and prices, and increased unemployment, which places enormous strains on poor people, and contributes to children dropping out of school, women and child trafficking, engaging in dangerous work, starvation, and xenophobic outbursts, issues that social workers engage with on a daily basis. The WTO remains a site of marked resistance on account of unfair trade practices, with the US government using it to protect the market for its multinational corporations with devastating effects on smaller-scale business within the developing world (Gray, 2009; Nixon, 2010; Stiglitz, 2006; Steger and Roy, 2010). The subsidies, what Nixon (2010: p. 260) calls *perverse subsidies* that amount to about US$2 trillion a year, include subsidies from rich country governments like the United States, Europe, and Japan that produce cheaper agricultural imports against which farmers in the developing world cannot compete. This practice exposes the hypocrisy of rich countries that argue no subsidies for agricultural products while doing precisely that in their own countries. Another player on the global bloc, threatening domestic production and employment, is China, with an influx of Chinese trade across the globe setting ablaze anger at the destruction of domestic livelihoods (Reuters, 2012). Local producers cannot compete against a range of cheap Chinese imports of clothes, shoes, toys, crockery, and electronics. Chinese goods accounted for 18% of South Africa's imports in 2012, up from 2% in 1995. With the

Eurozone crisis and a declining European market, China has been increasingly turning its trade interest to Africa, with Africa now accounting for about 70% of the revenue of some Chinese companies (Reuters, 2012).

With growing intolerance for lack of government regulation, people are taking measures into their own hands with violence and threats of violence. One Malawian merchant was reported to have said: "One day there will be blood" (Reuters, 2012: p. 13). In South Africa, xenophobic attacks, linked to socioeconomic deprivations, has seen blood and this seems to be spilling over to Chinese migrants. In November 2011, four Chinese were murdered in South Africa (Reuters, 2012). In August 2012, Kenyans protested with placards that said: 'Chinese must go.' If no attempts are made to regulate trade, it is likely that there will be increased backlashes against cheap Chinese imports. Yet, political leaders go against their own professed understandings. In July 2012, Jacob Zuma in Beijing talked about the importance of exercising caution "when entering partnerships with other economies" and he said that the trade relationship was "unsustainable in *the long term*" (Hook, 2012: p. 1, my emphasis). But it is in the *immediate and short term* that people have to feed, clothe, educate, and care for themselves and their children. The consequences of these discourses and policies for social work practice are self-evident, as social workers are at the forefront of dealing with the various manifestations of poverty and inequality.

Conclusion

The austerity measures imposed by the WB and the IMF and liberalization of trade continue to have devastating effects upon people, culminating in the current neoliberal crisis, which heralds a new multipolar world with the emerging economies of China, India, and Brazil being less dependent on the United States (Duménil and Lévy, 2011), and which has generated a range of oppositional movements. From small-scale protest actions in various parts of the world to the Arab Spring and the international Occupy Movement, lies the core-underlying theme of citizen intolerance of political authoritarianism, poverty and inequality, and exclusion in its various forms (Tormey, 2012; Wight, 2012). However, whether or not the current neoliberal crisis will allow for a transition toward a new world order beyond neoliberalism is open to question. Steger and Roy (2010: p. 137) predict that if the crisis continues or deepens "both third-wave neoliberalism (of a more moderate kind than its two predecessors) and a global new deal (built on Keynesian principles) are distinct possibilities for the second decade of the twenty-first century" (p. 137), while Duménil and Lévy (2011) are optimistic that the crisis might "favor a transition evocative of the New Deal" (p. 333). Given the hegemony of neoliberalism and its entrenched institutions and practices in favor of capitalist elites, who are not willing to give up their privileged positions, this will require a very strong counterhegemonic discourse to neoliberalism, envisioning another and a better world based on egalitarianism and emancipatory politics (Stiglitz, 2006; Nixon, 2010) and ensuring that popular global justice movements continue the pressure in challenging global capitalism. The hope for a better world is borne out of the ongoing struggle of the Arab revolution and in people becoming aware of the legitimating power of neoliberal capitalism, as seen in the international Occupy Movement.

See also: Disposable Youth in America in the Age of Neoliberalism; Governmentality; Liberalism, Anthropology of; Liberalism: Historical Aspects; New Managerialism and New Public Sector Management; Policy Knowledge: New Liberalism; Poverty Policy; Poverty, Geography of; Poverty, Sociology of; Trade Unions and Social Work: Lessons from Canada; Welfare Economics; Welfare Reform; Welfare Retrenchment; Welfare State, History of.

Bibliography

Abramovitz, M., Zelnick, J., 2010. Double jeopardy: the impact of neoliberalism on care workers in the Unites States and South Africa. International Journal of Health Services 40 (1), 97–117.
Banks, S., 2011. Ethics in an age of austerity: social work and the evolving new public management. Journal of Social Intervention: Theory and Practice 20 (2), 5–23.
Bauman, Z., 1992. Intimations of Postmodernity. Routledge, London.
Bauman, Z., Tester, K., 2001. Conversations with Zygmunt Bauman. Blackwell Publishers, Oxford.
Bond, P., 2005. Elite Transition: From Apartheid to Neoliberalism in South Africa. University of KwaZulu-Natal Press, Scottsville.
Bond, P., 2010. Limits to class apartheid. In: Maharaj, B., Desai, A., Bond, P. (Eds.), Zuma's Own Goal: Losing South Africa's War on Poverty. Africa World Press, Trenton, pp. 59–69.
Bottery, M., 2000. Education, Policy and Ethics. Continuum, New York.
Chomsky, N., 1999. Profit over People – Neoliberalism and Global Order. Seven Stories Press, New York.
Clarke, J., 1993. A Crisis in Care? Challenges to Social Work. Newbury Park, London.
Clarke, J., 2007. Subordinating the social? Neoliberalism and the remaking of welfare capitalism. Cultural Studies 21 (6), 974–987.
Dominelli, L., 1996. Deprofessionalising social work: anti-oppressive practice, competencies and postmodernism. British Journal of Social Work 26, 153–175.
Dominelli, L., 2002. Anti-Oppressive Social Work Theory and Practice. Palgrave Macmillan, Basingstoke.
Dominelli, L., 2008. The new world order, scarcity and injustice: social work fights back. Caribbean Journal of Social Work 6 (7), 6–15.
Dominelli, L., 2012. Green social Work: From Environmental Crises to Environmental Justice. Polity Press, Cambridge.
Duménil, G., Lévy, D., 2001. Costs and benefits of neoliberalism: a class analysis. Review of International Political Economy 8 (4), 578–607.
Duménil, G., Lévy, D., 2011. The Crisis of Neoliberalism. Harvard University Press, London.
Foucault, M., 1984. The Foucault Reader. Pautheon, New York.
Ferguson, I., Lavalette, M., 2006. Globalisation and global justice: towards a social work of resistance. International Social Work 49 (3), 309–318.
Gramsci, A., 1977. Selections from Political Writings 1910–1920. Lawrence and Wishart, London.
Gray, J., 2009. False Dawn: The Delusions of Global Capitalism. Granta Publications, London.
Harvey, David, 2005. A Brief History of Neoliberalism. Oxford University Press, Oxford.
Howe, D., 1994. Modernity, postmodernity and social work. British Journal of Social Work 24, 513–532.
Hoogvelt, A., 1997. Globalization and the Postcolonial World: The New Political Economy of Development. Macmillan, London.
Klein, N., 2000. No Logo. HarperCollins Publishers, Hammersmith.
Larner, W., 2000. Neo-liberalism: policy, ideology, governmentality. Studies in Political Economy 63, 5–26.

Lemke, T., 2002. Foucault, governmentality and critique. Rethinking Marxism 14 (3), 49–64.
Leonard, P., 1997. Postmodern Welfare: Reconstructing an Emancipatory Project. Sage Publications, London.
Nixon, B., 2010. A Better World is Possible: What Needs to Be Done and How We Can Make It Happen. O-Books, Winchester.
Pierson, C., 1991. Beyond the Welfare State? The New Political Economy of Welfare. Polity, Cambridge.
Polyani, K., 1957. The Great Transformation: The Political and Economic Origins of Our Time. Beacon Press, New York.
Reuters, September 13, 2012. Chinese Test Limits of Friendship. Daily News.
Rowden, R., 2009. The Deadly Ideas of Neoliberalism: How the IMF Has Undermined Public Health and the Fight against AIDS. Zed Books, London.
Sachs, J., 2005. The End of Poverty: How We Can Make it Happen in Our Lifetime. Penguin Books, London.
Saul, J.S., 2006. Development after Globalization: Theory and Practice for the Embattled South in a New Age Imperialism. UKZN Press, Scotsville. Zed Books Ltd, London.
Schram, S.F., Silverman, B., 2012. The end of social work: neoliberalizing social policy implementation. Critical Policy Studies 6 (2), 128–145.
Sewpaul, V., Hölscher, D., 2004. Social Work in Times of Neoliberalism: A Postmodern Discourse. Van Schaik Publishers, Pretoria.
Sewpaul, V., 2005. Feminism and globalization: the promise of Beijing and neoliberal capitalism in Africa. Agenda 64, 104–113.
Sewpaul, V., 2006. The global-local dialectic: challenges for African scholarship and social work in a post-colonial world. British Journal of Social Work 36, 419–434.
Sewpaul, V., 2008. Social work education in the era of globalization. Caribbean Journal of Social Work 6 (7), 16–35.
Steger, M.B., Roy, R.K., 2010. Neoliberalism: A Very Short Introduction. Oxford University Press, New York.
Stiglitz, J., 2006. Making Globalization Work. Penguin Books, London.
Tormey, S., 2012. Occupy wall street: from representation to post-representation. Journal of Critical Globalisation Studies 5, 132–137.
Wacquant, L., 2001. The penalization of poverty and the rise of neoliberalism. European Journal on Criminal Policy and Research 9, 401–412.
Wight, C., 2012. Riot, why wouldn't you? Journal of Critical Globalisation Studies 5, 161–166.

Relevant Websites

African Development Bank, 2011. The Revolution in Tunisia: Economic Challenges and Prospects. Available online: http://www.afdb.org/fileadmin/uploads/afdb/Documents/Publications/North%20Africa%20Quaterly%20Analytical%20Anglais%20ok_North%20Africa%20Quaterly%20Analytical.pdf (accessed 3.06.12.).
Amin, S., June 2001. Imperialism and globalization. Monthly Review 53 (2). Available online: http://www.monthyreview.org/0601amin.htm (accessed 3.10.12.).
Fortunato Jr., S.J., July–August 2005. The soul of socialism: connecting with the people's values. Monthly Review 57 (3). Available online: http://www.monthlyreview.org/0705fortunato.htm (accessed 30.06.12.).
Gibson, C.C., 2009. Neoliberalism and Dependence: A Case Study of the Orphan Care Crisis in Sub-Saharan Africa. Graduate School Theses and Dissertations. Available online: http://scholarcommons.usf.edu/etd/1983 (accessed 1.11.12.).
Haiven, M., 2011. From NYC: "Occupy Wall Street Has No Agenda" is an Alibi for Apathy, Halifax Media Co-op. Available online: http://www.mediacoop.ca/blog/max-haiven/8378 (accessed 1.11.12.).
Hook, L., 2012. Zuma Warns on Africa's Ties to China. Available online: http://www.ft.com/cms/s/0/33686fc4-d171-11e1-bbbc-00144feabdc0.html#axzz27MqYq8Us (accessed 26.09.12.).
International Labour Organization Report, 2009. Available online: http://www.ilo.org/wcsmp5/groups/public/dgreports/dcomm/documents/publication/wcms/101461.pdf.
Keynes, J.M., 1933. National self-sufficiency. The Yale Review 22 (4), 755–769. Available online: https://www.mtholyoke.edu/acad/intrel/interwar/keynes.htm (accessed 25.09.12.).
Lara, M.S., 2011. Kinship Care Policy: Exacerbating Women's Oppression Through Neoliberal Familialization. Open Access Dissertations and Theses. Paper 6375. Available Online: http://digitalcommons.mcmaster.ca/opendissertations/6375 (accessed 23.07.12.).
Obama, B., 2009. Inauguration Address. Available online: http://www.nytimes.com/2009/01/20/us/politics/20text-obama.html.
The Cabinet Papers 1915–1982. Available online: http://www.nationalarchives.gov.uk/cabinetpapers/themes/sterling-devalued-imf-loan.htm (accessed 19.07.13.).
World Commission on the Social Dimensions of Poverty, 2004. A Fair Globalization: Creating Opportunities for All. International Labour Organisation, Geneva. www.ilo.org/public/english/fairglobalization/report/index.htm (accessed 25.09.12.).

Neonatal Bonding

Aldo B Lucion, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Bonding' and 'attachment' characterize affiliative behaviors of the mother–infant interaction. These behaviors are crucial for the survival, providing the innate needs, and development of the infant in the interaction with the environment and sociability. Neonatal bonding involves a particular neuroendocrine milieu in the infant and the mother. Both infant and mother are hyporesponsive to stress, a condition that seems necessary for the approach and bonding. In rats, olfactory learning is the basic mechanism for the infant attachment to the mother. The maternal behavior would be the reciprocity to the infant needs. Oxytocin and vasopressin facilitate social recognition through the olfactory bulb. In rodents, oxytocin is highly involved in several social interactions, including social recognition, pair bonding, and maternal behavior. Neonatal bonding seems to establish the basic parameters for later social interactions.

Definitions: Bonding and Attachment

The neonatal period is a very particular moment in the lives of the mother and the infant. 'Bonding' is defined as the formation of a close personal relationship (as between a mother and a child), often through frequent or constant association (Webster's Ninth New Collegiate Dictionary Merriam-Webster 1991). According to the online Dictionary Thesaurus, maternal–infant bonding is the attachment that forms between an infant and its mother beginning at birth, and this bonding "influences the child's psychological and physical development."

In the human literature, the concept of mother–infant 'bonding' was used in *Maternal–Infant Bonding* by Klaus and Kennell (1976). This publication had considerable public impact and was controversial among neonatology scholars and professionals. The authors' primary concepts were based on animal studies that described sensitive periods in which certain behaviors were most likely to be formed (Bateson, 1979). It was proposed that in humans, there is a sensitive period necessary for the establishment of the bond between mother and infant, and this period extends to a few days after delivery. In some species, such as sheep, there are narrow time windows when the mother–infant bonding takes place. According to these ideas, infant bonding involves innate attachment mechanisms. Bonding defines a special social relationship; however, it does not reveal the full nature of the attachment (Crouch and Manderson, 1995).

The proposal of a specific period and stimuli that trigger a behavioral pattern has ethological roots. The concept of parent to infant 'bonding' has broad application, involving emotions between the mother or caretaker and the infant that are triggered by specific stimuli during a sensitive period; these emotions exert long-lasting effects on the future social interactions of the infant. The triggering stimulus is the skin to skin mother–infant contact. Should this bond be prevented by some interference, the infant may develop psychopathologies (Higashida et al., 2010; Jaffee and Price, 2012).

Studied by Nobel Prize ethologists K. Lorenz, N. Tinbergen, and K. von Frisch, the imprinted behaviors and instincts of animal species influenced the concept of human bonding. The classic picture of K. Lorenz being followed by a flock of geese is emblematic; he was the first figure that appeared after the eggs hatched. Further, several studies on imprinted behaviors revealed maternal features that trigger the infant behavior. Naturally, those features of the mother's body are characteristic of a particular species, and thus, the imprinted behaviors are species specific. Preprogrammed behaviors need a specific stimulus to be activated. In the natural world, imprinting decreases the number of individuals approached by the infant. It can also influence other aspects of sociability such as preferences for social interaction (Bateson, 1979). 'Sensitive' learning periods are also described for song learning in birds. Song learning is necessary for social communication and interaction. Learning species-specific songs is essential for the development of birds (Nelson and Marler, 1994). This learning process involves important neuroanatomical changes and synaptic plasticity that incorporate new song-related neurons (Nordeen and Nordeen, 1990).

In the literature, the terms mother–infant attachment, mother–infant relationship, and maternal–infant bonding are used interchangeably. Bonding and attachment may be linguistically synonymous, but the definitions and characterizations of attachment are predictive of future development, while bonding is an affective state between a pair (Bicking Kinsey and Hupcey, 2013).

'Bonding' and 'attachment' characterize affiliated behaviors in the interaction of a mother and infant. Bowlby introduced the concept of attachment (Bowlby, 1969). Attachment is a set of behaviors involved in the interaction between the neonate and the caretaker, usually the mother. This behavioral pattern is crucial for the survival, innate needs, development, and sociability of the infant. Furthermore, these behaviors are required for the infant to learn how to interact with the environment. Bowlby conceived this pattern as an internal working model, which is based on the internal representation of the environment. According to Bowlby, human beings are born with an innate psychobiological system (the attachment behavioral system) necessary for inducing social proximity to the attachment figures. The interactions with the attachment figures are the bases for future emotion-regulation strategies (Mikulincer and Shaver, 2012).

The concept of attachment has been largely used in Psychiatry to analyze and define patterns of the mother–infant relationship. Standardized instruments to operationally define the mother–infant relationship were created. The 'Strange Situation' paradigm was developed by M. Ainsworth et al. (1978), and multiple attachment patterns were defined as follows: (1) secure attachment, in which the infant shows signs that he or she misses the parent but calms down easily and is active in greeting the parent when the two are reunited; (2) avoidant attachment, in which the infant does not cry when separation occurs and ignores the parent when they are reunited; the infant shows an intense interest in the toys around him throughout the procedure; (3) resistant ambivalent attachment, in which the infant is focused on his parent throughout the procedure; the infant alternates between searching for contact and resistance or passivity and is difficult to calm down during the procedure. A fourth type of infant–parent attachment is a disorganized pattern, in which the infant's behavior is confused in the presence of the parent. For example, the infant may show freezing behavior. Another paradigm used in adults to infer attachment-related states is the 'Adult Attachment Interview' (Main et al., 1985), which is highly correlated with the Strange Situation protocol (Botbol, 2010). Thus, the attachment state characterized early in life by the Strange Situation protocol is useful to predict emotional outcomes in adulthood. Moreover, early experiences with attachment figures may also affect the development and maturation of the brain and have long-lasting effects on brain structures and function (Dana and Raz, 2012).

The importance of mother–infant attachment in the development of the infant is highlighted by the 'parent–infant bonding.' Attachment facilitates several practices during maternity that stimulate mother–infant interaction. As an indirect result of the early publication by Klaus and Kennel, research has increased significantly regarding the long-lasting effects of early life environment on the development of offspring. Several experimental approaches have been developed to analyze the mechanisms involved in infant 'bonding.'

Neuroendocrine Basis in Neonatal Bonding

Neonatal bonding involves a particular neuroendocrine milieu in the infant and the mother or caretaker. Mammals have adapted to attach to a caretaker during the neonatal period. This short period occurs after birth and is variable depending on the species. At this stage, pups approach any type of stimulus, including ones that are potentially dangerous. During the critical period, pups do not have a developed inhibitory system; they do not seem afraid. They are also hyporesponsive to stress stimuli. In the mother, hormonal changes occur in preparation for delivery and maternity. Both the infant and the mother are hyporesponsive to stress during this period, a condition that seems necessary for the approach and bonding.

The life of a mammal pup depends on the caretaker. Just after birth, the pups need to maintain the close relationship that they had *in utero*. Thus, pups tend to keep in close proximity to a potential caretaker. Rat pups are stressed by unknown adult males. Their physical presence induces a stress response in the pup, likely due to the smell of the conspecific from an unfamiliar environment. However, other stimuli do not induce aversion in the pups. The reports of abusive mothers are examples of attachment to a nonsupportive caretaker. In this case, an asymmetric relationship is observed. Pups approach the mother, while she behaves in a manner not conducive to the pups' well-being. Nevertheless, pups maintain such relationships.

It is interesting that bonding increases the physical proximity of the animals, including body contact, without increasing aggression. The reduction in physical distance may elicit several aversive reactions. Social proximity may be aversive. From the first days after birth until around postnatal day 12, rat pups show no averseness to the environment. Remarkably, pups may approach devices that deliver electric shocks. They may bond to abusive mothers. In humans, this is also observed. The studies of R.M. Sullivan et al. (Moriceau et al., 2009) clearly show that the amygdala is activated by an increase in corticosterone around postnatal day 12. After this event, the pups do not show olfactory learning to conditioned odors, and the pups begin to separate from the mother.

It is unknown what factors induce fighting or bonding in a social interaction. In both the cases, the animals are in physical proximity to each other; however, the most likely consequence of fighting is the disclosure of the participants, while behavior that induces bonding tends to strengthen and stabilize social contact.

Before one can infer the nature of these behaviors, whether they are innate or learned, one must unravel the mechanisms involved in neonatal attachment. Crucial mechanistic insights come from studies on the olfactory learning mechanism in rats and the discovery of the role of oxytocin in pair bonding. These findings demonstrate how animals can establish long-lasting and stable relationships. Attachment involves complex behaviors between at least two participants. These behaviors are dynamic and adaptable over time, making them more difficult to study. In contrast, the behaviors and mechanisms involved in activities that tend to reduce the proximity of participants, such as aggression, are studied extensively. Thus, the stimuli and neuroendocrine factors that underlie behaviors that increase proximity are just beginning to be addressed.

Oxytocin and Vasopressin

Several studies investigate the neural circuits and hormones related to maternal behavior. The mother is usually the primary caretaker, but the father also impacts the development of the infant. However, few studies analyze the father–infant relationship during the first months of life (Feldman, 2012).

Oxytocin is critical for infant attachment to the mother very early in life. In rodents, oxytocin is highly involved in several social interactions, including social recognition, pair bonding, and maternal behavior (Shahrokh et al., 2010; Young et al., 2001). Breast-feeding studies show that autistic infants breast-feed significantly shorter than control infants. Oxytocin and vasopressin facilitate social recognition through the olfactory bulb. These neuropeptides directly act on the nervous system

regions that are important for social information, and they may "modulate the activity of other neurotransmitter systems to increase the salience of social olfactory cues" (Hammock and Young, 2006). Lack of adequate levels of oxytocin in infants affects brain development, indicating that mother–infant attachment is important for the long-lasting development of social bonding later in life (Higashida et al., 2010). In birds, oxytocin has also been implicated in pair bonding (Klatt and Goodson, 2012).

The role of oxytocin on social behaviors early in life is not fully understood. However, there is evidence that lack of adequate exogenous oxytocin during infancy affects the normal development of the brain in genetically susceptible infants, thereby increasing the risk of autism. It is unclear how oxytocin in milk is transported into the brain through the blood–brain barrier; however, it is well established that breast-feeding is important for infant survival and brain development (Higashida et al., 2010). Vasopressin is another important neuropeptide involved in maternal behavior (Bosch and Neumann, 2011). Both oxytocin and vasopressin facilitate maternal care in lactating rats. In the olfactory bulb, vasopressin modulates the activity of mitral cells during social recognition (Tobin et al., 2010).

Olfactory Learning in Rats

In a series of elegant studies, Sullivan and coauthors investigate the mechanisms involved in the infant–mother attachment in rats. In the rat pup, the mechanism involves brain structures including the locus coeruleus, the olfactory bulb, and the amygdala. Olfactory learning seems to be the basic mechanism of infant attachment to the mother (Moriceau et al., 2009). The olfactory learning mechanism involves the neuroendocrine system in rat pups, which responds to proper environmental stimulation. This neural circuit is activated by maternal odor and is the basis for the attachment of the infant to the mother; this is essential for pup survival. The neonatal olfactory learning in rats involves approach behavior by the pup and the mother and a coincident attenuation of aversion. As in many mammals, this learning process occurs in a specific period. Reduced corticosterone seems to be crucial for the establishment of neonatal bonding. Indeed, after the natural increase in corticosterone that occurs with age, rat pups tend to show behavioral inhibition. At the same time, pups withdraw from the mother and show detachment behaviors, such as environmental exploratory activities. At this time (around the end of the second postnatal week), feeding is not necessarily dependent on the mothers; the pups' eyes have opened, their incisive teeth have erupted, and they can independently forage. Thus, the necessity of a caretaker (usually the mother) naturally decreases with age. However, the early social interactions with the mother appear to shape future social interactions. The author discusses this point later in this article.

In rat pups, licking by the mother stimulates the locus coeruleus, which has a different pattern of firing in infants compared to adults. The locus coeruleus is more active during the first days after delivery, which stimulates the olfactory bulb. The eyes and ears of rat pups are closed until the second postnatal week; thus, olfaction is crucial in the neonatal period to establish mother–infant attachment. However, olfaction is still in development at this age. In the olfactory bulb, noradrenaline from the locus coeruleus activates beta-adrenergic receptors to depolarize the mitral cells. The activation of beta receptors induces intracellular signaling such as phosphorylation of cAMP response element-binding protein (CREB), which is thought to be required for olfactory memory of the maternal odor (Moriceau et al., 2009). The specific odor of the mother that induces infant attachment seems to be in the mother's feces, which increase in volume due the increased food ingestion.

Humans: Mother–Infant Attachment

Similarly to other mammals, the neonatal period in humans is crucial for survival. The attachment to a caretaker figure is essential (Fonagy et al., 1991). Infants display preferences for social stimuli, suggesting the evolutionary pressure of those stimuli. However, the mechanisms involved in mother–infant attachment are relatively poorly understood. Nevertheless, infant approach behavior and nipple suckling also seem to be determined by olfactory stimuli. Reaching the source of milk is the first step of mother–infant attachment. The capacity to initiate sucking varies considerably among newborns. For social bonding to occur, individuals must have a capacity for social recognition, which includes the ability to detect as well as discriminate familiar individuals. Animal and human studies have shown that infants attach to a maternal figure, even if abusive behaviors are perpetrated by the caregiver. It can be inferred that it is better for an altricial animal to remain attached to an abusive caregiver than to receive no care (Raineki et al., 2010).

To analyze infant bonding, it is also necessary to analyze maternal behavior (Stoesz et al., 2013). Infant bonding is an attachment behavior that involves reciprocity by the partner, the caretaker. The mother displays maternal behavior, which reflects the infant's needs. Thus, maternal behavior is essential to trigger infant bonding. The licking behavior and physical mother–infant contact in rats are important maternal behaviors. In humans, the direction of the infant to the nipple to stimulate suckling is an important behavior that facilitates the development of social interactions. Maternal behavior is more than a response to the infant's behavior; mothers tend to induce infant bonding. The mother–infant early social interaction involves active approach behaviors by both the mother and the infant. The behaviors of a neonatal altricial mammal are limited due to the motor developmental stage. Nevertheless, altricial pups are essentially prone to approach behavior. They approach the mother and littermates. In some species, pups emit sounds to attract the caretaker. This step of infant attachment involves approaching behavior of the mother or caretaker. As for olfactory learning in rats and other rodents, maternal odorants seem crucial for this primary interaction. Odorants in the nipple and areola direct the attachment behavior of the newborn. However, there is variability among species and individuals in the establishment of mother–infant attachment. With time, the complexity of the mother–infant interaction increases. Other salient social cues

become important, such as the sight and voice of the mother. These stimuli contain information regarding the feelings of the mother toward the infant, which are important for the quality of the mother–infant attachment. Here again, the behavior of the mother is necessary for timely interaction with the infant and the establishment of a synchronized mother–infant interaction (Feldman, 2007; Reeb-Sutherland et al., 2012; Schaal et al., 2009).

Several mammalian species depend on maternal care to survive during the first days after birth. This vital relationship evolved from a different reproduction strategy. This strategy involves the spreading of large amounts of eggs by the progenitors, and no further parent–offspring interaction is developed. In contrast, mammals deposit fewer eggs but intensely invest in the offspring, although this varies among species. Thus, from the first moments after delivery, the pups interact with the mother. In fact, they must establish this new form of interaction with the mother to survive. In mammals, the mother and infant bond during the uterine gestational period, which in many cases is very long. During the gestational period, the fetus interacts with the mother and indirectly with its environment. The restricted and self-contained uterine environment imposes the first close-contact mother–infant interaction. However, after birth, the perspectives for the newborn increase dramatically. The physical mother–fetus interaction during gestation is replaced by a broader environment. The gestational close-contact interaction with the mother is disrupted and a different one must be urgently established. In some cases, such as in sheep, time is crucial; the mother–infant contact must be established within a few hours after birth, otherwise the mother will reject the newborn. The extrauterine environment offers enormous possibilities to the pup. In contrast to the *in utero* situation, a relationship with the mother has to be established. In some species such as rats, the mother–infant bond must be learned, for example, by an olfactory learning mechanism.

Innateness of Bonding

Some authors characterize 'bonding' as an innate behavior; the mother has a maternal instinct (Minde et al., 1980; Martyn et al., 2012). It would be more important to understand the strategies and mechanisms involved in this vital social interaction for each species. Bonding could be considered an intervenient variable for a situation poorly understood. The question of innateness is a long-standing debate among behavioral scientists. Innateness does not involve learning processes; the behavior is ready to be performed at the moment of birth. The animal is born with fixed action patterns, or 'instincts.' Thus, the infant needs no learning to engage in typical behaviors that occur after birth. For example, pups remain close to and follow the mother. The idea of innate, prewired behaviors is attractive. If these behaviors are innate, then the animals would not need environmental or maternal stimuli. The behaviors would initiate and form before birth. The stimuli to generate such 'innate' actions could occur intrauterus. Studies show that the mother–infant interaction begins before birth. The taste and the food spices eaten by the mother cross the placenta and induce a 'signature' in the fetus. After birth, the pups expect the same taste in the mother's milk, which is a crucial characteristic to identify the mother. In this case, the infant is 'imprinted' by the mother's behavior. The association of the stimulus and the object is an essential element in the learning processes. In mammals, it is quite difficult to separate the mother and infant. The close associative relationship between maternal stimuli and the offspring is strong, before and after birth. The prebirth bonding could infer a prebirth associative learning. In rats, the olfactory learning mechanism in pups is the basis for the infant attachment to the mother. This is clearly a learning process; indeed, the formation of attachment utilizes the main biochemical mechanisms involved in learning, such as CREB phosphorylation in the olfactory bulb.

The reciprocal interaction between the caregiving environment and the infant's genetic variations causes individual differences in infant attachment behavior. Variations in 5-serotonin-transporter-linked polymorphic region may shape how behavioral security or insecurity is manifested. Only a handful of studies have investigated genetic contributions to infant attachment (Luijk et al., 2011; Spangler et al., 2009; Raby et al., 2012).

The interaction between the environment and genetics in the programming of behaviors is increasingly studied. The influence of maternal behavior on the genotype of the mother and child and vice versa are areas of future study in the field. Several studies demonstrate the vulnerability of some individuals to environment stressors (Labonté et al., 2013). The gene–environment interaction is complex. Genes are affected by the environment; this may result in changes in behavior and may increase vulnerability to psychopathologies. On the other hand, genes may also affect the environment. A critic might take the position that beyond such generalizations, this research does not tell us anything one did not know already. The idea that person–environment correlations are important in shaping development has been recognized in developmental psychology and psychopathology (Jaffee and Price, 2012).

Long-Lasting Effects of Neonatal Bonding: Concluding Remarks

Several studies have shown that early life events can determine the personality and physiology of mammals (Sheriff et al., 2010). The early infant–caretaker bonding seems to establish the basic parameters for later social interactions. Infant bonding is the embryonic development of the social brain. Indeed, the experience with the mother in early life seems to shape the relationship of an infant with its own offspring. Studies in rats that display high or low licking behavior are quite conclusive. The pattern of high or low licking behavior is 'copied' from mothers to daughters. Thus, the daughters of high-licking mothers will also be high-licking mothers with their own pups. Memories from early life are long lasting compared to the myriad of other stimuli that humans experience. Indeed, these other stimuli do not seem to impact the brain in the same way as early life stimuli. This is likely related to the critical period during which the animal is most vulnerable to environmental stimuli. Other close social relationships, such as pair bonding, follow the same principle and possibly have the same neuroendocrine

basis. Bonding involves special cognitive procedures that affect specific brain areas, neurotransmitters, and neurohormones. Alternatively, bonding may stem from the same areas and neurohormones as other social behaviors but show a different synchronization of activity. In this case, the structural bases are the same, but they result in different functions. Several studies show the importance of maternal care in regulating the development of offspring. Indeed, the quality and quantity of the mother–infant relationship seem to shape the future social competence and ability of the infant to face stress (Pedersen, 2004).

Whether this transgenerational transmission is true for other behaviors or physiological parameters is not fully known (Shah et al., 2010). For instance, the epigenetic transmission of glucocorticoid expression in the hippocampus is well documented. Thus, epigenetic mechanisms could explain the transgenerational transmission of the bonding phenotype. The expression of a receptor, for instance, the glucocorticoid receptor, may have an important impact on the stress response. However, innate behavior implies the existence of a complex pattern of activities. In mammals, there is no evidence of innate behaviors that are capable of producing a complex social interaction like bonding. Pups are prepared to respond to vital stimuli. Moreover, at this early age, the behavioral repertoire of the pups is restricted. One may hypothesize that the neuroendocrine system of young mammals is directed to specific stimuli necessary for survival after birth. Among the several environmental stimuli that the pups experience at birth, many more than in the uterus, mammalian pups respond promptly to the mother. It is important to emphasize the role of the mother in directing the behavior of the pups. Pups bond to the mother in part because the mother stimulates the infant to develop and deepen the interaction with her.

Nevertheless, mechanisms involved in the mother–infant relationship seem to be unraveling by analyzing carefully human subjects as well as using animal models. The importance of the early-life mother–infant bonding has been clearly demonstrated in programming behavior strategies later in life in several species. In a certain way, the neonatal mother–infant bonding would be an extension of the intrauterine environment enabling a full neural development of the animal in an environment closely related to the one that will be found throughout life. Neonatal bonding occurs in several animal species, not only mammals, indicating its evolutionary importance on the development of a social brain. Most importantly, this early life bonding would be the main reference for future social interaction in which animals tend to approach each other in opposition to inhibitory sociability or even social disaggregation by aggression. The social signs are essential for durable social relationship and also the establishment of social hierarchies, both crucial for the species survival. Neonatal bonding seems to establish the basis for the development of the social abilities. The caretaker, usually the mother, would guide that development, which has to occur outside the uterus in mammals or the egg in birds, in close contact with the environment that the infant will probably face in life. Thus, one may infer that the neonatal bonding has its innate roots and also seems to need plasticity in order to face the variability of the environment. The research challenges are various; however, it seems crucial to understand how the environment 'gets into the skin' of the infant and what are the first, second, third, and '*n*' order messengers that convey environmental information to the brain of the infant. Moreover, the neuroendocrine changes induced by those stimuli are necessary to be stated. With that, the origins of some important psychopathologies would be known and thus better treated.

See also: Neonatal Imitation; Neonatal Social Isolation: Effects on Neural Function; Neonatal Temperament: From Phenomenology to Mechanism; Stress: Effects on Neonatal and Child Development.

Bibliography

Ainsworth, M.D.S., Blehar, M.D., Waters, E., Wall, S., 1978. Patterns of Attachment. Erlbaum, Hillsdale.

Bateson, P., 1979. How do sensitive periods arise and what are they for? Animal Behaviour 27, 470–486.

Bicking Kinsey, C., Hupcey, J.E., Feb 26, 2013. State of the science of maternal–infant bonding: a principle-based concept analysis. Midwifery 29, 1314–1320. http://dx.doi.org/10.1016/j.midw.2012.12.019 pii: S0266-6138(13)00007-7.

Bosch, O.J., Neumann, I.D., 2011. Both oxytocin and vasopressin are mediators of maternal care and aggression in rodents: from central release to sites of action. Hormones and Behavior 61, 293–303.

Botbol, M., 2010. Towards an integrative neuroscientific and psychodynamic approach to the transmission of attachment. Journal of Physiology – Paris 104, 263–271.

Bowlby, J., 1969. Attachment and Loss, vol. 1. Basic Books, New York, NY.

Crouch, M., Manderson, L., 1995. The social life of bonding theory. Social Science & Medicine 41, 837–844.

Dana, O., Raz, S., 2012. Adult attachment and emotional processing biases: an event-related potentials (ERPs) study. Biological Psychology 91, 212–220.

Feldman, R., 2007. Parent–infant synchrony and the construction of shared timing; physiological precursors, developmental outcomes, and risk conditions. Journal of Child Psychology and Psychiatry 48, 329–354.

Feldman, R., 2012. Oxytocin and social affiliation in humans. Hormones and Behavior 61, 380–391.

Fonagy, P., Steele, M., Steele, H., 1991. Maternal representations of attachment during pregnancy predict the organization of infant–mother attachment at one year of age. Child Development 62, 891–905.

Hammock, E.A.D., Young, L.J., 2006. Oxytocin, vasopressin and pair bonding: implications for autism. Philosophical Transactions of The Royal Society B 361, 2187–2198.

Higashida, H., Lopatina, O., Yoshihara, T., Pichugina, Y.A., Soumarokov, A.A., Munesue, T.Y., Minabe, Y., Kikuchi, M., Ono, Y., Korshunova, N., Salmina, A.B., 2010. Oxytocin signal and social behaviour: comparison among adult and infant oxytocin, oxytocin receptor and CD38 gene knockout mice. Journal of Neuroendocrinology 22, 373–379.

Jaffee, S.R., Price, T.S., 2012. The implications of genotype–environment correlation for establishing causal processes in psychopathology. Development and Psychopathology 24, 1253–1264.

Klatt, J.D., Goodson, J.L., 2012. Oxytocin-like receptors mediate pair bonding in a socially monogamous songbird. Proceedings of the Royal Society B 280, 1–7.

Klaus, M.H., Kennell, J.H., 1976. Maternal–Infant Bonding: The Impact of Early Separation or Loss on Family Development. Mosby, St Louis.

Labonté, B., Suderman, M., Maussion, G., Lopez, J.P., Navarro-Sánchez, L., Yerko, V., Mechawar, N., Szyf, M., Meaney, M.J., Turecki, G., 2013. Genome-wide methylation changes in the brains of suicide completers. American Journal of Psychiatry 170, 511–520.

Luijk, M.P., Tharner, A., Bakermans-Kranenburg, M.J., van Ijzendoorn, M.H., Jaddoe, V.W., Hofman, A., Verhulst, F.C., Tiemeier, H., 2011. The association between parenting and attachment security is moderated by a polymorphism in the mineralocorticoid receptor gene: evidence for differential susceptibility. Biological Psychology 88 (1), 37–40.

Main, M., Kaplan, N., Cassidy, J., 1985. Security in infancy childhood and adulthood: a move to the level of representation. Monographs of the Society for Research in Child Development 50, 66–107.

Martyn, A.C., Choleris, E., Gillis, D.J., Armstrong, J.N., Amor, T.R., McCluggage, A.R.R., Turner, P.V., Liang, G., Cai, K., Lua, R., 2012. Luman/CREB3 recruitment factor regulates glucocorticoid receptor activity and is essential for prolactin-mediated maternal instinct. Molecular and Cellular Biology 32 (24), 5140–5150.

Mikulincer, M., Shaver, P.R., 2012. An attachment perspective on psychopathology. World Psychiatry 11, 11–15.

Minde, K.K., Marton, P., Manning, D., Hines, B., 1980. Some determinants of mother-infant interaction in the premature nursery. Journal of the American Academy of Child Psychiatry 19 (1), 1–21.

Moriceau, S., Shionoya, K., Jakubs, K., Sullivan, R.M., 2009. Early-life stress disrupts attachment learning: the role of amygdala corticosterone, locus ceruleus corticotropin releasing hormone, and olfactory bulb norepinephrine. Journal of Neuroscience 29, 15745–15755.

Nelson, D.A., Marler, P., 1994. Neurobiology selection-based learning in bird song development. Proceedings of the National Academy of Sciences of the United States of America 91, 10498–10501.

Nordeen, E.J., Nordeen, K.W., 1990. Neurogenesis and sensitive periods in avian song learning. Trends in Neurosciences 13, 31–36.

Pedersen, C.A., 2004. Biological aspects of social bonding and the roots of human violence. Annals of New York Academy Sciences 1036, 106–127.

Raby, K.L., Cicchetti, D., Carlson, E.A., Cutuli, J.J., Englund, M.M., Egeland, B., 2012. Genetic and caregiving-based contributions to infant attachment: unique associations\ with distress reactivity and attachment security. Psychological Science 23 (9), 1016–1023.

Raineki, C., Moriceau, S., Sullivan, R.M., 2010. Developing a neurobehavioral animal model of infant attachment to an abusive caregiver. Biological Psychiatry 67 (12), 1137–1145.

Reeb-Sutherland, B.C., Levitt, P., Fox, N.A., 2012. The predictive nature of individual differences in early associative learning and emerging social behavior. PLoS One 7 (1), e30511. http://dx.doi.org/10.1371/journal.pone.0030511.

Schaal, B., Coureaud, G., Doucet, S., Delaunay-El Allam, M., Moncomble, A.-S., Montigny, D., Patris, B., Holley, A., 2009. Mammary olfactory signalisation in females and odor processing in neonates: ways evolved by rabbits and humans. Behavioural Brain Research 200, 346–358.

Shah, P.E., Fonagy, P., Strathearn, L., 2010. Is attachment transmitted across generations? The plot thickens. Clinical Child Psychology and Psychiatry 15, 329–345.

Shahrokh, D.K., Zhang, T.-Y., Diorio, J., Gratton, A., Meaney, M.J., 2010. Oxytocin-dopamine interactions mediate variations in maternal behavior in the rat. Endocrinology 151, 2276–2286.

Sheriff, M.J., Krebs, C.J., Boonstra, R., 2010. The ghosts of predators past: population cycles and the role of maternal programming under fluctuating predation risk. Ecology 91, 2983–2994.

Spangler, G., Johann, M., Ronai, Z., Zimmermann, P., 2009. Genetic and environmental influence on attachment disorganization. Journal of Child Psychology and Psychiatry 50 (8), 952–961.

Stoesz, B.M., Hare, J.F., Snow, W.M., 2013. Neurophysiological mechanisms underlying affiliative social behavior: insights from comparative research. Neuroscience & Biobehavioral Reviews 37, 123–132.

Tobin, V.A., Hashimoto, H., Wacker, D.W., Takayanagi, Y., Langnaese, K., Caquineau, C., Noack, J., Landgraf, R., Onaka, T., Leng, G., Meddle, S.L., Engelmann, M., Ludwig, M., 2010. An intrinsic vasopressin system in the olfactory bulb is involved in social recognition. Nature 464 (7287), 413–417.

Young, L.J., Lim, M.M., Gingrich, B., Insel, T.R., 2001. Cellular mechanisms of social attachment. Hormones and Behavior 40, 133–138.

Neonatal Crying Behaviors

Kathleen Wermke, University Clinics, University of Würzburg, Würzburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Neonatal crying is far more than an innate biological alarm mechanism employing acoustic signals and reserved exclusively for the communication of emergency and distress. As they have been more closely identified in recent years, the principles of cry melody development allow us to class infant cries with other forms of non-cry vocalizations as essential precursors to language acquisition. The differentiation of cry melodies alongside the parallel development of increasingly finely regulated intensity contours over the first weeks of life distinguishes the crying of human neonates from that of nonhuman primate infants. The latest research indicates a marked divergence between human and nonhuman primate infants in even the earliest stages of vocal development. This makes human infants and their early vocalizations particularly suitable candidates for research on spoken language acquisition.

Crying as a Communicative Behavior

For a long time neonatal crying has primarily been seen as a behavioral state in contrast to other states such as sleep, drowsiness, and wakefulness, and one that points to a physiological state of distress associated with physiological and behavioral changes other than vocalization itself (Wolff, 1987). This behavior, typical of the first months of life, is however done more justice when cries are characterized as a social signal and crying as a communicative behavior. In the first 2 months of life it is primarily crying and fussing that assume a predominant role in the expressive repertoire of the neonate. (Cry intensity may vary between whimpering (fussing), mitigated crying, and vigorous crying. These expressions may differ with regard to their rhythmical features, fundamental frequency and intensity features, as well as to the nature and rapidity of maternal responses that they provoke. However, here the focus lies on the function and general characterization of neonatal crying (and the unpleasant bodily and emotional states it elicits) and a clearly drawn boundary may not be relevant for the purposes of this article. A detailed differentiation can be found elsewhere, e.g., Wolff, 1987; Barr et al., 2000.) The neonate may also produce a few quiet, short non-cry sounds from the beginning of life. The latter play a comparatively minor role in social interaction, however, owing to the affective and communicative power of crying, a feature directly related to its biological function. Beginning in the third month of life, the infant's expressive acoustic repertoire significantly broadens, such that now in stress-free situations pronounced and clearly audible comfort sounds (cooing), capable of eliciting a clear response in the listener, are used to communicate.

As a result of its physical helplessness, the human infant is dependent upon close social contact with its mother from birth. (The term 'mother' is used here to imply the neonate's primary caretaker.) It follows that biological signals would help to facilitate in this interaction. Crying – acoustic signaling – is the most powerful signaling available to the neonate. As a distal cue, it is effective not only in immediate proximity to or physical contact with the mother, as with proximal cues such as smiling and gaze, but also when the mother is outside of the visual range of the neonate or in conditions of darkness. Acoustic signals thus assume a special role in the development of attachment in comparison with attachment-promoting visual signals.

Human neonatal crying is often functionally equated with nonhuman primate infant crying, insofar as both are regarded primarily as 'separation calls,' which function exclusively to minimize distance to the mother. However, this model likely applies to a far greater degree, if not exhaustively, to nonhuman infant crying (Todt, 1988; Bard, 2000). Human neonatal crying has a much broader functional spectrum. It serves primarily interaction-regulating purposes by expressing emotions and feelings (Keller et al., 1996; Falk, 2009; Wermke and Mende, 2011).

Neonatal cries are heavily loaded with emotions. The neonate sends messages regarding his or her sensations, feelings, and wants and these are above all coded in the cry melody. These messages are 'body-near,' in that they immediately express bodily conditions and emotional states, and they are in large part intuitively 'intelligible' to the mother. It is therefore hardly surprising that there exist several diverse approaches seeking, by way of acoustic properties, to grasp the underlying emotional content of an infant's crying and to differentiate, for example, between cries of hunger, pain, and frustration (Wasz-Höckert et al., 1968; Lester and Boukydis, 1985; Wolff, 1987). Over a developmental period of several years leading to full acquisition of the native language, the messages encoded in a child's acoustic utterances will become increasingly complex and differentiated.

A long evolutionary adaptation process has enabled the mother and her newborn baby to be a priori tuned to each other. Characteristics of both infant and mother help shape and ultimately determine their interactions with one another. A 'mistuning' within this biologically preprogrammed system may produce a cumulative feedback effect ultimately leading to behavioral changes on the part of the infant (e.g., extensive or excessive crying, the avoidance of eye contact, or the deliberate turning away of the head) and negative emotion on the part of the mother (e.g., feelings of guilt, failure, and frustration) (Keller et al., 1996). The latter holds all the more true, as the mother is often well-aware that she is the one who more strongly determines the nature and manner of this interaction during the neonatal period.

Crying as Part of the Typical Behavioral Repertoire of all Neonates

Crying belongs to the normal behavioral repertoire of any neonate. This does not mean, however, that all infants cry to the same degree or with the same frequency. As in all forms of human behavior, crying behavior varies between individuals and develops with age. There are several reasons for variations in the amount and daily frequency of crying between neonates. For example, individual personality factors (e.g., temperament, sensitivity to pain, irritability, perinatal and postpartum discomfort experiences), as well as inter-individual differences in brain maturation, may contribute to variability in crying behavior. Apart from general maturational effects, furthermore, all physiological and medical states affect neonatal crying behavior. In this respect, it is not only the amount of crying that may differ between healthy neonates and those neonates who are ill, born prematurely or with certain congenital conditions, or exposed to potentially painful medical treatments or procedures, but also the acoustic character of this crying (Wasz-Höckert et al., 1968; Lester and Boukydis, 1985). This clinical perspective is outlined in more detail below.

Through the process of biological feedback and 'tuning' between mother and child outlined above, maternal personality characteristics and patterns of behavior in relation to the neonate can also influence the amount and frequency of neonatal crying. Observations indicate that the amount and frequency of neonatal crying as well as the response behavior of mothers varies culturally and across a spectrum encompassing irritation, anger, and sympathy (Eibl-Eibesfeldt, 1983; Kruger and Konner, 2010). In pre-industrialized societies, where neonates are often provided with intense physical contact and in which neonatal crying is responded to immediately and by frequent nursing, newborns were found to cry less and rarely in inconsolable bouts (Kruger and Konner, 2010). However, more systematic cross-cultural cry studies, including recorded acoustic data from hunter–gatherer and forager cultures, are needed in order to establish a 'normal neonatal crying curve' or a 'normal amount of daily crying' in neonates cross-culturally. Thus far the extant data are limited, suggesting that claims such as crying of 6.7 h per day during the first 3 days of life of crying is typical, and the mean duration per day for the first 3 months can reach 120 min or postulations as to a peak in the daily amount of crying around 6 weeks of age, should at least not be accepted uncritically. The range of data suggested by several authors, spanning cry periods rarely in excess of 6 min in the first month of life, or 3 min in the second and third months, to cry periods of two to 3 h per day over the first 6 months of life cautions against generalization. (The effort to define the 'normal' amount and duration of neonatal crying was launched by pediatricians faced with cases of excessive crying and related parental complaints. This is, however, a special aspect of cry research that is thoroughly described elsewhere; see, for example, Barr et al., 2000.)

Already in 1987, Peter Wolff pointed out that it is difficult to establish a mean duration or normal quantity for crying. This is because both depend not just upon the neonate and its particular mental state, but also to a significant degree upon maternal behavioral responses, which determine whether crying is terminated "before it [has] run its course" (p. 80).

Anatomical Basis and Neurophysiological Control

The coupled audio-vocal system represents the most developed neuromuscular functional system of the neonatal central nervous system (CNS). Crying may thus represent the most highly coordinated and finely regulated motor behavior available to the neonate. A prerequisite of neonatal cry production is a well-coordinated interaction between respiratory and laryngeal mechanisms. The muscles of the larynx function as a part of the respiratory system before birth, and like all other respiratory muscles they have undergone considerable use prior to birth (Harding, 1984). Crying is produced during the expiratory phase of the breathing cycle; some neonates also vocalize during the subsequent inspiration.

The neonatal larynx features a soft cartilaginous, skeletal structure and a relatively narrow lumen. According to an investigation by Savkovic et al. (2010) in infants between 0 and 12 months of age, the larynx has an average length of 11.9 ± 0.3 mm and an average width of 17.7 ± 0.5 mm. Both measures were significantly correlated with the infants' body length. A layered structure or vocal ligament is absent in the neonate larynx. This and further identified specifics of the fine structure of the human newborn and infant vocal fold mucosa led scientists to conclude that the vocal fold mucosa lack adequate viscoelasticity for phonation and that the newborn glottis hence seems to serve more respiratory than phonatory function. However, the astonishingly vast phonatory range and complexity reflected in neonatal cry properties suggests otherwise, as does the fact that neonates are able to cry loudly and intensively over hours without displaying noticeable signs of exhaustion or even developing aphonia. A comparable performance, involving the prolonged maintenance of a maximum vocal load, would hardly be possible for adults. Through their study of the neonatal larynx, particularly the distribution of hyaluronic acid within the neonatal larynx, Schweinfurth and Thibeault (2008) have contributed to solving this paradox. The differences in the laryngeal structure between neonates and adults seem to explain the neonate's capacity to sustain relatively high-pitched crying and high volumes over a long period of time. Hyaluronic acid, an interstitial polysaccharide, is seen as the primary determinant of vocal fold viscosity and water content. The apparent 'immaturity' of neonatal laryngeal structure thus yields, upon closer observation, a survival advantage by supporting sustained, load crying when necessary. This supports the older findings of Lind and Truby (1965), who, with the help of detailed cineradiographic investigations, demonstrated that the neonate is capable of well-coordinated laryngeal performances.

In order to align subglottic, glottic, and supraglottic functions, all of which are involved in phonation, a comprehensive neurophysiological coordination is demanded. To this end, information is required from several regions of the brain. In human infants, studies of brain function during crying do not exist. Neural circuits mediating crying and their maturation have been identified in animal studies (cf Newman, 2007).

In mammals, the neural structures involved in phonation form an extensive system spanning all levels of the brain (Newman, 2007). Jürgens (2009) described a model of vocal control, based mainly on research in the squirrel monkey, which consists of two hierarchically organized pathways. The first of these controls the readiness to vocalize. It runs from the anterior cingulate cortex via the periaqueductal gray (PAG) into the reticular formation of pons and medulla oblongata, and from there to the phonatory motoneurons. The second pathway was found from the motor cortex via the reticular formation to the phonatory motoneurons. It includes two feedback loops, involving the basal ganglia and the cerebellum, respectively, which provide the motor cortex with information needed to generate final motor commands.

These and other studies challenge the assumption of a so-called 'brainstem model' of crying, which excludes the possibility that structures rostral to the midbrain play a role in cry production (cf Newman, 2007).

Neonatal Crying from a Clinical Perspective

The preceding remarks suggest that the neonatal larynx and indeed the whole phonatory system responsible for crying is not only mature enough for, but also ideally suited to, its function in neonates. The early maturity of this system, the complexity of the underlying neurophysiological mechanisms as well as the high speed at which the vocal control mechanism acts, make neonatal cry properties a suitable indicator for the identification of neurophysiological dysfunctions. In addition to direct studies of the brain, dysfunction in complex cerebral systems can only be identified by studying the performance of the corresponding efference system, such as phonation. Scandinavian pediatricians observed this in the early 1960s and systematically investigated neonatal crying with the then-available technique of signal analysis (Wasz-Höckert et al., 1968; Michelsson, 1971). The value of these pioneering studies cannot be overestimated, even if they were initially centered on the descriptive analysis of crying in neonates already diagnosed with medical conditions. These and a number of more recent studies have shown that, of the numerous theoretically definable signal parameters of a neonate cry, some have proven suitable as risk-markers that most directly reflect dysfunctions in vocal control. These are entities that characterize the sound source (larynx), such as the mean fundamental frequency, the fundamental frequency contour (melody), and related parameters and temporal parameters. As the supralaryngeal structures and their modifying function in vocal production mature, both the importance of vocal tract resonance characteristics and their variability in assessing neurophysiological performance increase. This is nonetheless not to be expected before the third month of life (Wermke and Mende, 2011).

However, clinical cry studies have thus far offered little in terms of new diagnostic information for individual neonates. Apart from the absence in these studies of clinically suitable reference values for the normal range of cry properties among healthy neonates, previous studies have also disregarded the age effect. Investigations conducted in our laboratory have yielded ample evidence for the modification, at times remarkably rapid, of cry properties over the first weeks of life. Additional known factors that can influence cry parameters and which must be taken into account both in the clinical interpretation of cry studies as well as in the establishment of normative values include gestational age, mode of delivery, somatic state (small-for-gestational-age vs appropriate-for-gestational-age neonates), the individual course of the postnatal adjustment phase, neonatal hearing performance, maternal drug abuse, and not least, the influence of maternal behavioral factors. Moreover, in the majority of previous cry studies from a clinical perspective the analyses were based on pain cries, which differ in their typical features from spontaneously uttered cries (see below).

I see at least two clinical areas in which a revival of cry diagnosis would benefit future work. The first of these concerns sensory-neural hearing-impaired infants. A systematic clinical study of changes in cry parameters in hearing-impaired neonates under strong control of all clinical data, including amplification data, may, for example, help to identify neonates with progressive hearing loss or may provide complementary information for an individually optimized treatment from an early age. Secondly, given the complex neural pathways underlying neonatal crying, maturational delays in a single circuit may cause imbalances or dysfunction of the network. In the future it may become possible to evaluate the developmental status of specific brain circuits underlying human neonatal crying.

In any case it will be necessary to establish suitable reference databases of cry characteristics and their range of variability, as well as age-dependent changes in healthy infants, before cry diagnosis can be employed as a clinical method.

Characteristics of Neonatal Cries

The basic acoustical signature of neonatal cries is primarily determined by processes that characterize the sound source (phonation). Because neonate populations worldwide do not differ in terms of the anatomical structures and functional mechanisms of phonation, the essential acoustic attributes of neonatal cries are similar everywhere. Neonatal cries are characterized by a mean fundamental frequency between one and two octaves higher than that of an adult voice. The mean fundamental frequency of spontaneous cries of healthy neonates varies between about 350 and 650 Hz. Investigations as to whether this range is slightly modified in neonates with an ambient tonal language are still pending. Neonatal cries, except pain cries, which are described below, exhibit an undisturbed spectral pattern consisting of a fundamental frequency which varies with time (melody), and its integral multiples called harmonics. Nonstationary frequency spectra provide information on spectral changes over cry time and represent a visualization of the perceived acoustic spectrum (**Figure 1**a–c, right side).

Neonatal cries can exhibit fundamental frequency instabilities, as in the form of sudden jumps of the fundamental frequency (shift) or vibrato-like phenomena. In neonates, these are often signs of a temporarily immature vocal control. Only if they occur repeatedly and over a longer time period may they indicate an at-risk or even pathological condition. It is important to note that this applies only to spontaneous cries that were not elicited by a pain stimulus.

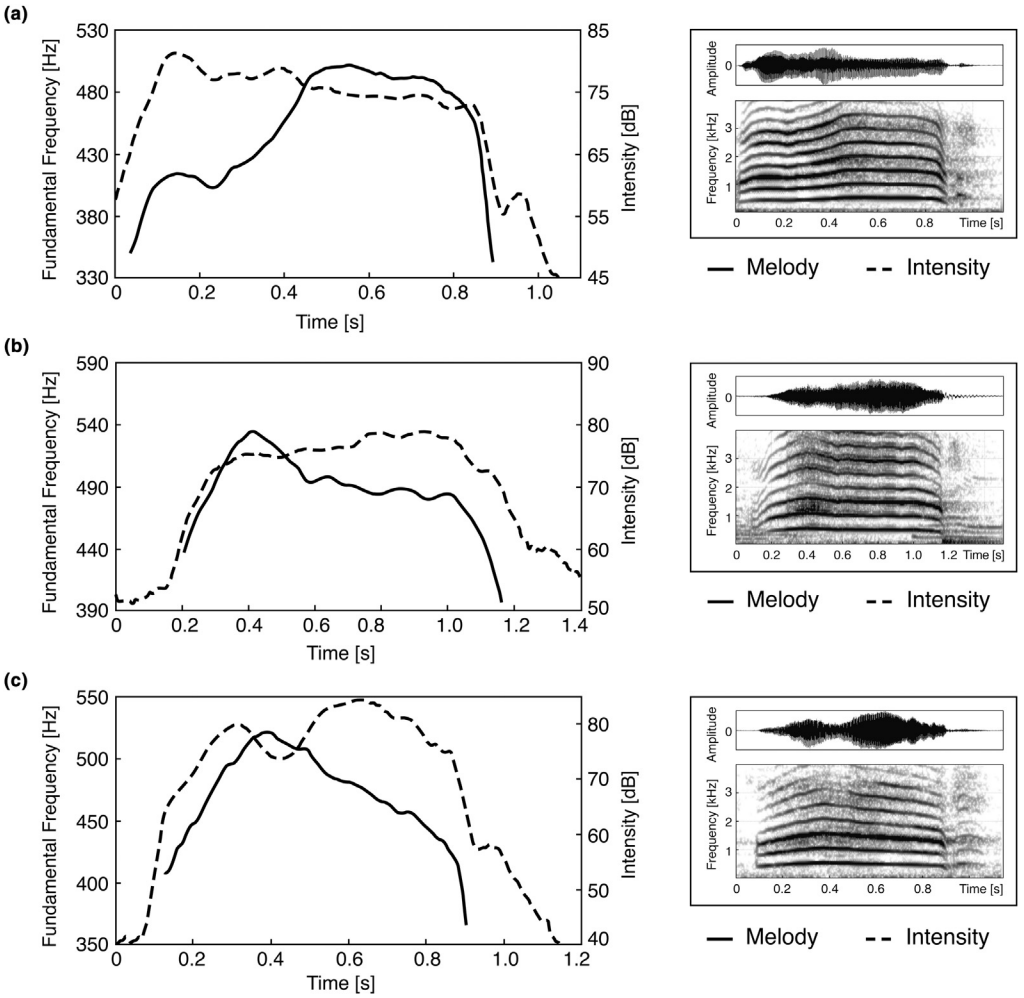

Figure 1 a–c: Melody/intensity contour diagrams of 3-day-old newborns' cries.

The frequency spectra of single cries may also contain subharmonics or noisy elements; both are found in healthy neonates and neonates with medical conditions. Mende et al. (1990) were the first to show that at least some of the phonatory noise phenomena in neonates' crying are caused by nonlinearity in vocal fold mechanics and can be interpreted as low-dimensional chaos. As the distress level increases and crying becomes more intense, these phonatory noise phenomena grow more pronounced, likely a result of the inability of the tiny vocal folds to withstand the enormous subglottal pressure during crying (cf Titze, 1993). Over the first 4 months of life, the neonate learns to avoid these phenomena and the frequency of occurrence of such subharmonic and noise phenomena drastically declines in healthy infants, without, however, disappearing entirely. In subsequent stages of vocal development, noise-like phenomena are intentionally produced where required by the specific features of the native language.

The fundamental frequency contour (melody) of a cry reflects the oscillation pattern of the vocal folds as it varies over time. It is only roughly recognizable on the basis of the contour of higher harmonics in the frequency spectrum, and is therefore best represented in special melody/intensity diagrams (Figure 1a–c, left side).

The melody consists of single rising-then-falling arcs exhibiting a rising contour (a) and a falling contour (b, c), respectively. The corresponding spectrograms are also displayed. In all three examples melody contour is relatively independent from intensity contour in respect to monotony behavior.

The fundamental frequency of a cry is produced through a coordinated interplay of laryngeal muscles. CNS mechanisms must regulate this interplay in order to enable the efference of precisely reproducible melody contours. It is often claimed that neonatal crying merely reflects the infant's respiratory cycle or even represents a reflexive manner of vocalizing in which the infant has lost vocal control. Of course there must be a principal coupling between respiratory activity and laryngeal cry production, but as our research very clearly indicates this is an elastic coupling. Neonates are already capable of an independent control of fundamental frequency and intensity (see Figure 1a–c). The elastic coupling, or the relative independence of melody from intensity, powerfully suggests that neonatal cries are not simply the byproducts of aerodynamic processes. These findings clearly contradict older descriptions that claimed

that cry melody was constrained by the respiratory cycle and emphatically refute the assumption that the crying of human neonates is reflexive and uncontrolled.

In the course of the first months of life, a neonate's increasing mastery of fundamental frequency modulation (melody) develops in close concert with growing control and the intentional regulation of intensity. From around the fourth week of life one can often observe how, in an almost playful manner, melody and intensity contour are woven into the most diverse vocal combinations (Wermke and Mende, 2011). These variations give the impression of a preparatory exercise, in which it is possible that the groundwork is being laid for acquisition of language-specific patterns of accent and emphasis. This decoupling is, alongside observed forms of systematic melody development (see below), an example of the specificity of the human developmental pathway. The investigation of the interplay of melody and intensity contour could itself prove a fruitful area of future research, both in the field of clinical cry research as well as from the perspective of language acquisition. In this area there remain considerable gaps in our knowledge. In conclusion, I would like to underscore that neonatal crying does not only differ in amount and duration, but also in acoustic properties from neonate to neonate.

The Specifics of Pain Cries in Neonates

Crying is a highly effective means by which a neonate can express acute pain. Due to their salience, cry features are used in several pain-scoring systems for neonates. The identification of pain-related specifics of cry properties has been a matter of intensive research. Cries uttered immediately after an acute pain stimulus may be characterized as having an increased mean fundamental frequency, a short rise time of the melody contour up to its maximum, an often very irregular frequency pattern and an extended duration. After the acute phase of pain sensation, the fundamental frequency declines to 'normal' values, cry duration decreases, and cry melody changes its shape from a falling toward a typical rising-falling pattern (cf Wasz-Höckert et al., 1968; Lester and Boukydis, 1985). Thus, cry parameters change with the diminishing of a pain sensation. It is important to note here that there is no single type of pain-stimulated cry (prototypical pain cry), just as there is no single hungry, pathological, or stereotypical 'normal' cry. When one compares across different neonates the first few cries uttered after receiving the same pain stimulus (e.g., heel-lancing), considerable differences can appear in the cry parameters. This is, however, to be expected, owing to the complexity of the underlying neurophysiological mechanisms. One must take account of the uniqueness of pain cries when investigating neonatal crying.

Developmental Changes of Cry Properties: A Special Note on Melody

The first week of life is generally characterized by individual adaption processes. These processes are reflected only marginally in the cry melody pattern of these neonates, however. They are more reflected in the short-time variability of the fundamental frequency (Wermke and Mende, 2011). This points to the particular importance of cry melody among the full range of cry parameters.

Neonatal cry features undergo significant changes in the first weeks of life. Fluctuations in volume or variations in melody or rhythm, as in the insertion of pauses between single melody arcs, can be observed already in the second week of life. The richness of variability among melodic and rhythmic elements is already in these first weeks remarkably high (Wermke and Mende, 2011).

Neonates are particularly sensitive to musical elements. Even in utero, the fetus is especially receptive to prosodic acoustic features, particularly to the melody of human speech. The dense functional coupling of auditory and vocal systems furthermore allows the neonate to actively produce and rehearse melodic-rhythmic elements in his or her own crying from the moment of birth. The early rehearsal of melodic-rhythmic elements appears to be essential for the undisturbed process of language acquisition in the first year of life. We postulated that neonates' melody skills could mean much more than a special acoustic capability; they might even be a language learning imperative (Wermke and Mende, 2011). Adopting this perspective, a young infant's melodic skills should predict that same child's language skills much later, because both reflect the same capacity of language learning.

Our longitudinal studies demonstrated a continuous coherent unfolding of cry melody features during the first 28 weeks of life which we interpret as a developmental program (Wermke and Mende, 2011). We found a relatively smooth unidirectional development from simple single-arc melody structures, prevailing in neonatal crying of the first week of life, to an increasing number of double- and multiple-arc melody structures in cries of the following weeks. Complex cries may consist of two, three, four, or more melody arcs within a single expiratory cry signal, and can be understood as being composed of a concatenation of simple arcs as building blocks. Following this interpretation, melody development is seen as a primitive step in a coding hierarchy creating the first combinatory assemblies of melody arcs, coding emotions, affects, and needs (Wermke and Mende, 2011). This development starts immediately after birth, that is, long before babbling or any symbolic word use occur. This coding is tightly coupled with bodily processes in neonates and is therefore much less arbitrary than the semantic configurations that appear in the utterances of older infants, children, or adults. Melody development in infants' vocalizations lays the groundwork for a later divergence of separate, but closely related, developmental pathways culminating in the musicality and linguistic competence of children. Over the course of the second 6 months of life, this process of divergence first begins, leading on the one hand to the complex phonological structures and stable semantics of language, and on the other to the complex tonal structures of music. Music and language appear to emerge from the same basic emotional framework.

Crying from an Evolutionary Perspective

From an evolutionary perspective, there is a further supportive argument for a potential link between crying and language provided by Dean Falk (2009). She assumes that the persistent adult-directed crying of human infants and the typical infant-directed speech of human mothers are complementary behaviors that represent prelinguistic substrates that paved the way for the eventual emergence of protolanguage.

Mother–child interaction, in which mental states, feelings, and needs are to a large degree vocally communicated, perhaps provided a motor for the evolution of spoken language in the prehistoric development of humankind. With the appearance of physiologically 'premature' birth among our hominid ancestors, in the context of the evolution of upright posture and concomitant skeletal changes affecting the maternal pelvis, hominid babies lost the ability to cling independently to their mothers' bodies (cf Falk, 2009). Temporary separation from and insufficient physical contact with the body of its mother very likely provoked protest on the part of the child. According to the anthropologist Dean Falk (Falk, 2009), the special, very melodic form of vocal communication that exists between a mother and her infant (*motherese*) could well have co-evolved in this prehistoric time. Falk hypothesizes that in the appearance of early forms of *motherese* the decisive step was made toward the evolution of spoken language ('putting the baby down' theory).

All human neonates cry during their first months of life. Nonhuman primate infants may also show a similar behavior. They will respond with 'fussing' ('hoo' vocalizations and whimpers) and crying (or screaming) upon separation from their mother, or upon the perception of physical discomfort and pain. Chimpanzee infants may exhibit, even for extended periods, distress vocalizations from birth that are accompanied by typical facial expressions, where different patterns of vocal-facial expressions are indicative of different degrees of distress (Bard, 2000). However, there is a decisive difference: while human mothers exhibit a variety of behavioral responses to their infants' crying, such as picking up the infant, rocking the infact gently and rhythmically, or nursing, they will also respond vocally by talking or singing to the neonate. A chimpanzee mother never responds in this way (cf Falk, 2009). This makes of human infants and their early vocalizations particularly suitable candidates for future research on the evolution of spoken language.

See also: Neonatal Bonding; Neonatal Imitation; Neonatal Social Isolation: Effects on Neural Function; Neonatal Temperament: From Phenomenology to Mechanism; Stress: Effects on Neonatal and Child Development.

Bibliography

Bard, K.A., 2000. Crying in infant primates: insights into the development of crying in chimpanzees. In: Barr, R.G., Hopkins, B., Green, J.A. (Eds.), Crying as a Sign, a Syptom, and a Signal. Mac Keith Press, Distributed by Cambridge University Press, London, pp. 157–175.

Barr, R.G., Hopkins, B., Green, J.A. (Eds.), 2000. Crying as a Sign, a Symptom, and a Signal. Mac Keith Press, Distributed by Cambridge University Press, London.

Eibl-Eibesfeldt, I., 1983. Patterns of parent-child interaction in a cross-cultural perspective. In: Oliverio, A., Zappella, M. (Eds.), The Behavior of Human Infants. Plenum Publishing Corporation, pp. 177–217.

Falk, D., 2009. Finding Our Tongues. Basic Books, A member of the Perseus Books Group, New York.

Harding, R., 1984. Function of the larynx in the fetus and newborn. Annual Review of Physiology 46, 645–659.

Jürgens, U., 2009. The neural control of vocalization in mammals: a review. Journal of Voice 23 (1), 1–10.

Keller, H., Chasiotis, A., Risau-Peters, J., Völker, S., Zach, U., Restemeier, R., 1996. Psychobiological aspects of infant crying. Early Development and Parenting 5 (1), 1–13.

Kruger, A.C., Konner, M., 2010. Who responds to crying? Maternal care and allocare among the !Kung. Human Nature 21, 309–329.

Lester, B.M., Boukydis, C.F.Z., 1985. Infant Crying. Theoretical and Research Perspectives. Plenum Press, New York.

Lind, J., Truby, H.M., 1965. Newborn Infant Cry. Almqvist&Wiksells Boktryckeri AG, Uppsala.

Mende, W., Herzel, H.P., Wermke, K., 1990. Bifurcation and chaos in newborn infant cries. Physics Letters A 145 (8/9), 418–424.

Michelsson, K., 1971. Cry analyses of symptomless low birth weight neonates and of asphyxiated newborn infants. Acta Paediatrica Scandinavica Supplement 216, 1–45.

Newman, J.D., 2007. Neural circuits underlying crying and cry responding in mammals. Behavioural Brain Research 182 (2), 155–165.

Savkovic, A., Delic, J., Isakovic, E., Ljuca, F., Nuhbegovic, S., 2010. Age characteristics of the larynx in infants during the first year of life. Periodicum Biologorum 112 (1), 75–82.

Schweinfurth, J.M., Thibeault, S.L., 2008. Does hyaloronic acid distribution in the larynx relate to the newborn's capacity for crying? The Laryngoscope 118, 1692–1699.

Titze, I.R., 1993. Vocal Fold Physiology: Frontiers in Basic Science. Singular Publishing Group, San Diego.

Todt, D., 1988. Serial calling as a mediator of interaction processes: crying in primates. In: Todt, D., Goedeking, P., Symmes, D. (Eds.), Primate Vocal Communication. Springer, Berlin, Heidelberg, New York, pp. 88–107.

Wasz-Höckert, O., Lind, J., Vuorenkoski, V., Partanen, T., Valanne, E.H., 1968. The Infant Cry: A Spectrographic and Auditory Analysis. Heinemann, London.

Wermke, K., Mende, W., 2011. From emotion to notion: the importance of melody. In: Decety, J., Cacioppo, J.T. (Eds.), The Oxford Handbook of Social Neuroscience. Oxford University Press, New York, pp. 624–648.

Wolff, Peter H., 1987. Behavioral States and the Expression of Emotions in Early Infancy. New Proposals for Investigation. The University of Chicago Press, Chicago and London.

Neonatal Imitation

Giannis Kugiumutzakis, University of Crete, Crete, Greece
Colwyn Trevarthen, University of Edinburgh, Edinburgh, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Imitation is essential for cultural learning, language and conventional reasoning, but its development been controversial for millennia, and remains problematic. Many psychologists deny newborns imitate, believing that months of learning will need a self-awareness to construct representations of mental states for perceiving others' intentions and emotions. Nevertheless, microanalysis of naturally motivated engagements has proved (1) that a newborn moves with an integrated and emotional self-awareness by proprioception, and (2) can imitate face expressions, vocalizations, and gestures made by another person. Neonatal imitation displays innate alteroceptive readiness for interpersonal dialogic awareness, which is regulated by relational emotions of interest and pleasure.

The Nature of Human Intelligence at Birth

Movements of newborn infants are complex, coordinated, and characteristically human in the adaptations of the whole body for acquiring experience of the outside world with many senses (Trevarthen, 1986). Their movements are emotionally regulated and expectant for sympathetic engagement with the rhythmic expressions of motives and feelings of other persons who offer intimate communication (Trevarthen, 1979). These are adaptations for intelligent life in a human community. They are prepared for in prenatal development of the anatomy of a human body and brain (Trevarthen, 2001), and are made evident by the sensitively directed actions and emotional expressions of the fetus (Trevarthen and Delafield-Butt, 2013). Knowledge of the spontaneous behaviors, narrative awareness, and affective sensibilities of young infants, of their response to the desires and interests aroused in a loving parent, and of age-related developments in communication before language, gives us a natural science of cultural learning, social collaboration and mental health in human families and communities, and of the functions of language.

A Controversial Topic in Philosophy and Psychology

A newborn human body is very immature and its brain, though distinctly human, is small. Many life sustaining functions must change to make use of the new environment outside the mother's body (Nagy, 2011). For many months the infant's life and its changes will depend on maternal attention, support, and care. Nevertheless, the newborn has an ingenious *motor intelligence* and may show adventurous animation with finely coordinated and purposeful actions that direct senses in selective, intentional ways. Within minutes of an easy birth, while accepting gentle and affectionate responses in intimate contact, a baby may gaze with intense interest at the other human being, making delicate hand gestures and sharing smiles or small vocal sounds, sometimes by direct imitation (Maratos, 1973; Meltzoff and Moore, 1977; Kugiumutzakis, 1998, 1999; Butterworth, 1999; Nagy and Molnár, 2004; Reddy, 2008). Postures, gestures, face expressions, and the voice sounds are expectant of sharing experience and emotion in human company. The emotional exchange of parent and child is one of an instinctive reciprocal, playful 'sympathy' with complementary and constructive feelings, not the one-sided pickup of feeling now being identified as 'empathy' (Reddy and Trevarthen, 2004; Kugiumutzakis et al., 2005).

And yet, the oldest records we have of the views of philosophers and physicians deny that an infant can have either self-consciousness or awareness of another person. Respect for the importance of education in fact-based explanation and in the conventions of language, and for the need to cultivate conventions of ethical social behavior, leads infants to be thought of as nonpersons who require instruction to be engaged with other persons' motives and emotions. Medical science has represented infants as delicate organisms needing nourishment and protection, with therapeutic treatment of disorders and capable only of reflex responses. Conscious awareness of self and of others and their purposes, ideas, and feelings are perceived to require weeks, months, or years of learning. Any communication of awareness will wait until the child learns representations of other persons and their actions, and these concepts will be absent or rudimentary until the child can speak and reflect with words on what others' actions mean.

How imitation develops persists as a problem in modern developmental psychology (Reddy and Trevarthen, 2004; Reddy, 2008; Nagy, 2011). Many who experiment with how infants respond to stimuli or master use of objects have denied neonates can imitate on the grounds that appreciation of others' intentions and feelings requires months of learning to construct a basic self-world awareness – how to move as one agent guided by acquired conceptions of objects gained with 'pleasure in mastery' (Piaget, 1962). Once a coherent sense of being in an objective time and space is built, then the actions of others, and their feelings, can be appreciated and adopted as a consciousness of their mental states, or a 'theory of mind.'

The Developmental Psychobiology of the Infant Social Mind

From analysis of detailed recordings of newborns' behaviors in appropriately responsive company that accepts their initiatives and expectations, and with new evidence on how body, brain, and behavior develop before birth, it is evident that a child achieves rational and representational intelligence of a culture and its language on the foundation of innate powers of

self-creation (Trevarthen and Delafield-Butt, 2013). Intentions to move with feeling of the body, and cognitive appreciation of the consequences of being in touch with objects, are regulated from the start by emotions adapted to appreciate the risks of acting in that world (Trevarthen, 2001). The sensitive actions of the fetus are adapted for social cooperation and learning with other persons (Piontelli, 2010).

There is proof that, after a premature birth, smiles, vocalizations and gestures can be ready for engagement with the feelings of life in another person 2 months before term (Trevarthen and Delafield-Butt, 2013). We gain further evidence from the development of affective communication through infancy, regulated by age-related changes in motives for cooperation in play and for sharing practical use of objects (Trevarthen, 1977, in Bråten, 1998).

That neonatal imitation is real and necessary for attachment in parental care and for development of human social and cultural intelligence is supported by advances in the neuroscience of intentions and emotions (Ammaniti and Gallese, 2014). Detailed descriptions of how a young infant can take initiative in communication with parents have transformed the management of perinatal medical care of infants and their mothers (Brazelton, 1979), and analysis of the dynamics of play with infant's motives are leading to new ways of conceiving early education for the development of cultural skills and language, and psychotherapy.

In spite of these advances, it remains difficult for scientific psychology, with its rational bias, to accept that a newborn has sufficient awareness to imitate another person's expressive movements. This disbelief has a long history.

The History of Thinking about Newborn Awareness and Communication

The Ancient Dispute about Neonatal 'Mimesis'

According to Plato and Aristotle, mimesis is an innate human ability, but neonates cannot imitate. They are in a state of 'initial adualism' (Kugiumutzakis, 1998). The Platonic embryo has the inborn 'wisdom' of life, but forgets at birth the essential awareness of its nature, including the sense of *sameness* and *difference* in experience. In a chaotic state, the neonate cannot recognize the other or imitate. Aristotle declared that at birth we are unable to distinguish the self from the other. Because the baby lacks memory and cannot discriminate similarities and differences, it has no positive emotions to express to a parent. "… parents love their children as soon as these are born, but children love their parents only after time has elapsed and they have acquired understanding or the power of discrimination by the senses" (Aristotle, *The Nicomachean Ethics*, 1161b).

Two hundred years after Plato and Aristotle, in the second century AD, the Stoics taught that human newborns posses one of the three innate impulses of affection, called *oikeiosis* or 'familiarity of being' – an affinity for being an organism, an innate orientation to self-love and self-preservation, and thus self-perception. Hierocles was the first to oppose the Aristotelian view with a thesis of innate *dualism* or self–other awareness, for humans and for other animals. "'The first thing that animals perceive is their own parts … both that they have them and for what purpose they have them, and we ourselves perceive our eyes and our ears and the rest. So whenever we want to see something, we strain our eyes, but not our ears, toward the visible object … Therefore the first proof of every animal's perceiving itself is its consciousness of its parts and the functions for which they were given" (Hierocles, quoted by Long and Sedley, 1987, *The Hellinistic Philosophers*, vol. 1, p. 347).

Hierocles' observations were ignored by science for 2000 years. Ideas of initial adualism prevailed in Developmental Psychology until the last quarter of the twentieth century, with notable exceptions among philosophers of the Scottish Enlightenment, Frances Hutcheson, David Hume, Adam Smith, and Thomas Reid, who believed we relate by 'innate sympathy.' Charles Darwin, James Mark Baldwin, Martin Buber and John Macmurray also rejected Cartesian Dualism and opened the way to a new interest in innate relational awareness, now supported by both psychological studies of infants (Reddy, 2008: Stern, 2010), and research on the brain mechanisms of self-regulation and other-awareness (Nagy, 2011; Ammaniti and Gallese, 2014).

Fathers' Diaries of Infants' Early Life

Since the seventeenth century there have been many attempts to make natural descriptions of early human behavior and its embodiment in imaginative moving. These lead to ideas opposing the rational dualism of René Descartes. In particular, there are attempts to understand the nature of immediate interpersonal awareness, about which Descartes had less to say.

In 1787, in his diary of his infant's development Dietrich Tiedemann concludes that the desire to communicate with others 'is rooted in human nature.' Charles Darwin, observing his firstborn son, William, felt that the infant's awareness of others during the first few months is limited, before the emergence of 'practical reasoning or rational, intentional action.' In 1877, he reported that his son, "… had no distinct evidence of his distinguishing and recognising anyone, until he was nearly four months old." Darwin suggested that, "The imitation by articulate sounds of musical cries might have given rise to words expressive of various complex emotions" (On the evolution of language from imitation of expressive vocalizations see Malloch and Trevarthen, 2009).

Tiedemann and Darwin attributed their son's *neonatal* behavior to reflexes. For Tiedemann, while an initial dualism is necessary for the development of intelligent imitation, neonatal imitation of emotional expressions such as the smile may be a primary, sympathetic response. "And herein lies the first source of sympathy; for by looking upon others, who are stirred by some emotion, the soul projects itself into the same situation, wherein it has once experienced that feeling" (Tiedemann, 1787: p. 17, quoted in Butterworth, 1999: p. 73). This was the understanding of Frances Hutcheson and his students, Hume, Smith, and Reid.

William Thierry Preyer observed his infant son with special knowledge of early brain development. He, like Tiedemann, recorded imitation of the mother's smile 1 month after birth. Preyer did not consider this to be 'real' imitation. He believed that the loose organic connection of the spinal cord and the brain in early infancy leads to two 'egos absolutely isolated from each other.' The ego of the spinal marrow (that only feels) is different from the ego of the brain (that feels, sees, hears, tastes,

smells, and speaks). Each one of the imperfectly developed, uncoordinated senses has its separate ego – each sense 'perceives for itself.' Later on their coordination, representations, and concepts (the 'I'-makers) will lead to the formation of the abstract, 'corporate I.' This synthesis is reached *late in the second year*, after many 'playful experimentations' with the infant's own body, with other persons and with objects.

Modern Developmental Psychology: From Unconscious Emotionality and Schemas for Objects to Innate Intersubjectivity

James Mark Baldwin (1897) developed a comprehensive theory of the cognitive and social development of organisms, which has informed subsequent research on the genesis of infant self and its intelligence. However, the richness of Baldwin's understanding of the generative powers of organismic action with awareness, also considered of basic importance to psychology by William James and James Herbert Mead, was dismissed in the United States with the advent of behaviorism and learning theory, which dominated teaching and research on 'psychological mechanisms' for five decades after James Watson defined its method in 1913.

Baldwin recorded the development of social reactions in four infants, including his two daughters. He accepted repetition of actions as essential to the development of intelligence in the brain, and his developmental theory assumed that to be social a person must be imitative. His explanation of imitation is based on his theory of 'circular reactions,' the performance of stimulus-repeating movements, found in ontogeny, phylogeny, and heredity. A stimulus triggers a motor process that tends to reproduce itself and the same sensory feedback, generating repeating psychological sequences of 'reality, image, movement.'

According to Baldwin, all organic selection tests its awareness, retaining the vitally good and pleasurable relations with the environment and suppressing bad and unpleasant stimulation. It is called 'circular' because its action tends to repeat itself to reproduce stimulation *caused by movement*. It is 'conservative,' since it keeps up old movements, creating habits, but it is also 'playful,' open to new adaptations by the overproduction of movement variations. Play permits change by selection. Both evolution and development require cooperation between the organism and the environment. A way of life is selected, by 'intelligent adaptation.' The theory resembles that of the creative relation of an organism to its 'life world' presented by Jacob von Uexküll (1957), which inspired the development of the theory of animal semiotics or social signaling of Thomas Sebeok and the ethologists.

Baldwin proposes that the only way a new movement can arise with purpose is playfully, by making the movements already possible so excessively and with so many varieties of combination that new, useful, and pleasurable adaptations may occur. Circular reactions or organic imitations appear at birth and they grow by distinguishing and selecting pleasant, useful self-imitations, and, eventually, become heteroimitations of other's actions. Unlike other species, infants' imitations are not limited by fully functional instincts. For humans, experimental imitation, play, is the main instrument of learning and leads to the acquisition of functions from the social environment.

Baldwin assumes that for the neonate, the self, other persons, and inanimate things are experienced as *objects*. The deliberate movements of other persons begin to attract attention and acquire meaning during the first 6 months. After 6 months, with new cerebral development there appears *persistent imitation* of others' actions, which makes the baby feel, *like them*. By imitating them, the infant acquires the 'germ of his *sense of agency*' and starts to discriminate the 'characters' of the others.

Psychoanalytic Theory of Neonatal Identification with the Vital Impulses of the Other

In his works on primary identification, Sigmund Freud proposes that identification is "… the earliest … the original expression of an emotional tie with another person …" (1921: pp. 134, 137, 140), but in his writing about shared feeling of the Id he makes no reference to imitation. With most psychoanalysts who followed him, Freud accepts the concept of initial adualism of Baldwin. Before the emergence of the ego, food intake and sexual activity are undifferentiated as a source of organic pleasure, and the neonate and the young infant is unable to discriminate his bodily self from the mother as source of comfort and pleasure. Exactly when, during the oral stage of primary narcissism or autoerotism and primary identification with the mother, imitation appears, is not clear in the writings of Freud.

Later the psychoanalysts Fenichel and Gaddini proposed that a primitive imitation is possible from the start. These early imitations are involuntary, reflexlike, emotional, immediate reactions, obeying the pleasure principle, and felt as 'undifferentiated wholes.' Gaddini argued that, during the adualisic, narcissistic period, imitation is a core factor in early development of the ego and adaptation. Early, 'realistic' identification presupposes early imitation, which is concerned 'only with unconscious fantasy' in the absence of the gratifying objects.

In Psychoanalysis, any acceptance of the idea of the neonatal imitation came late. Early infant imitation was conceived as an involuntary, unconscious, adualistic experience regulated by primary emotional systems of the Id. The neuropsychoanalitic theory of primal or lower brain emotional consciousness presented by Jaak Panksepp (2009) is compatible with many elements of Freud's theory of unconscious emotions.

The psychoanalyst and developmental psychologist Daniel Stern, integrating findings on infant communication of the preceding 30 years, in *The Interpersonal World of the Infant* first published in 1985, challenged the traditional account of the original state of the infant's psyche, and of the stages of ego development. He proposed that the human infants: "… never experience a period of total self/other undifferentiation. There is no confusion between self and other in the beginning or at any point during infancy … There is no symbiotic-like phase … The period of life from roughly nine to eighteen months is not primary devoted to the developmental tasks of independence … It is equally devoted to the seeking and creating of intersubjective union with another" (Stern, 2000: p. 10). Stern recognized that "In light of the new evidence on other-centered-participation shown by infants in their many forms of imitation, as well as the new findings on mirror neurons

and adaptive oscillators, I am now convinced that early forms of intersubjectivity exist from almost the beginning of life" (Stern, 2000: p. xxii). Stern describes a layered model of development in which the first phase of intersubjective awareness, and all subsequent phases of its elaboration, remain as foundations throughout life, and can be invoked in the therapeutic process with adult patients (Stern, 2000: Revised Figure 2.2, p. xxv).

Piaget's Rejection of Neonatal Imitation

For Piaget (1962), who by his tests of the construction of object-conceiving abilities of infants, laid the foundation for developmental cognitive psychology, the neonate cannot imitate. While there is an innate tendency to imitate, the 'technique' of imitation must be learned or practiced. The innate element is termed *reproductive* or *functional* assimilation, a mechanism that motivates *self-repetition* of a sensory motor schema, recalling Baldwin's 'circular reaction.' This develops by integration of new 'external elements' i.e., acquired, tested, and confirmed experiences, which become schemas for useful objects (Piaget, 1962: p. 5).

With Baldwin, Piaget assumes that *self-imitation precedes and leads to heteroimitation*, that imitation has to be, at first, an *intrapersonal phenomenon*, which gradually becomes *interpersonal* in infancy, then at 2 years of age becomes again intrapersonal as interiorized imitation or thinking about it, as in the contemporary 'theory of mind.' In his words:

> Our study of the beginnings of representation in the child will mainly be in those fields where the individual processes of mental life dominate the collective factors, and we shall emphasize these individual processes particularly in the case of imitation, which though it leads to inter-individual relationships does not necessarily result from them.
>
> 1962: p. 2

For Piaget, imitation during the first 8 months of life is a circular reaction because the infant experiences the actions modeled by another person as a continuation of self-related activity. 'True,' spontaneous imitation, based upon active, spontaneous assimilation appears later in development and lasts. Early 'pseudoimitation' of a facial model does not last, says Piaget, except under the influence of continual stimulation. After the eighth month, Piaget invokes the passive process of *association* for the further development of imitation, because the active processes of *assimilation* and *accommodation* cannot explain the appearance of facial imitation. In other words, he disregards intracorporal proprioceptive sensibility, which the physiologist Charles Sherrington discovered and called 'the felt Me.' Piaget, echoing the ideas of Preyer, notes that "training in imitation is necessary, especially when it is a case of imitation of movements which the child cannot see himself make" (1962: p. 41). The newborn infant, 'amid the chaos of impressions by which he is beset' (Piaget, 1962: p. 83) during the first 30 days, does nothing more than exercise his 'pure' reflexes – exercise which, after the first month, will lead gradually to imitation.

The Role of Emotions in Early Imitation: The Love of Shared Knowledge

The ancient Greeks believed mimesis could not occur between emotionless persons. Over the last 100 years, the emotional or 'sympathetic' aspect of early mimesis has been accepted by some developmentalists, notably Baldwin, Freud, Guillaume, and Wallon, but the hypothesis of a strong link between emotion and imitation was subsequently dismissed under the influence of Piaget's rational philosophy (Kugiumutzakis, 1998: pp. 81–88; Kugiumutzakis et al., 2005: pp. 161–163).

Piaget made only four observations of his newborn infant between the first and the sixth days after birth, 1% of his corpus of observations on infant imitation. The infant was asleep, drowsy or crying, states of self-expression that disfavor imitation, and he only tested for imitation of crying (Piaget, 1962: pp. 7–8). He gives equal status to cognitive and emotional aspects of action, assuming that the cognitive provides structure, and the emotional the 'energetic force' of behavior, stressing that in imitation the emotion of 'interest' has a purely cognitive origin.

Piaget's views on the 'emotional' aspect express his radical cognitivism. His observations of infants' curiosity and object concepts cannot explain the sympathetic companionship of artful, humorous play and cultural learning, or the primary sympathetic self–other awareness that enables a newborn infant and a parent to enjoy their affection by reciprocal imitation.

Recent Research on Neonatal Imitation

The theory that imitation of another person's face expression is impossible for a newborn was shown incorrect by Zazzo in 1957, and disproved experimentally by Olga Maratos in her PhD thesis presented to Piaget in Geneva in 1973, and later by tests made independently by Meltzoff, Kugiumutzakis, Vinter, Field, Reissland, Heimann, and Nagy, all of whom used microanalysis of film or television recordings to record the infants efforts. Maratos's work, which included consideration of prenatal motor coordination, persuaded Piaget, who had rejected Zazzo's work, to change his mind and say, 'Indeed they imitate' (see Reddy, 2008: p. 46).

It is important to remember that conscious anticipation of the stimulations that arise with any organized movements within one's own body depends on the regulated rhythm of movements (Trevarthen, 1986). Live communication requires coregulation of the timing and dynamic modulations in movements between subjects (Stern, 2010). Condon and Sander, studying 'self-synchrony' of movements of body parts in infants and adults, published evidence in 1974 of 'inter-synchrony' of neonates climaxes of arm movement with speech of the mother, suggesting that the infant's impulses to move is 'entrained' by the mother's behavior. Trevarthen interpreted films of mother–infant communication as revealing 'mutual regulation,' "infant and mother appear to achieve a more complex co-operation which only occasionally results in synchrony of individual acts when mother and infant overlap to the same beat. Usually their acts alternate or complement one another" (Trevarthen, 1977: p. 241). This implies shared circular sensory motor reactions that test both self- and interregulations of

movement in experimental 'playful' ways, through periods of time that are shared with prospective control, as in music (Malloch and Trevarthen, 2009).

Meltzoff and Moore (1977) confirmed Maratos's findings, demonstrating that newborns aged 2 to 3 weeks can imitate tongue and lip protrusion, mouth opening, and finger movements. In cross-sectional and longitudinal studies with newborns and older infants, Meltzoff and collaborators added new findings on the development of vowel imitation between 3 and 5 months, on deferred imitation, and on the ability of infants at 18 months to infer goals from people's unsuccessful attempts to manipulate objects. Meltzoff and Moore's claims generated intense debates. Failures to replicate the findings, using strict controls of stimulation insensitive to the infant's initiative, supported convictions that the movement was a primitive reflex controlled by an 'innate releasing mechanism,' an 'arousal' response to the stimulus, or a product of associative learning of contingent acts (see Heimann in Bråten, 1998).

Meltzoff hypothesized that newborn facial imitation involves an 'active intermodal mapping' (AIM mechanism). "Infants' self-produced movements provide proprioceptive feedback that can be compared to the representation of the observed act by vision. AIM proposes that such comparison is possible because the observation and execution of human acts are coded within a common framework. We call it a 'supramodal act space'" (Meltzoff, 2002: p. 24). This is offered as the foundation of the 'theory of mind' proposed by Baron-Cohen to explain joint attention to purposes of object use developed at the end of the first year. Trevarthen (1977) showed evidence that acts directed with 'prospective control' from birth, by hands or by head and eyes to objects seen or heard outside the body, are coordinated in a single integrated 'behavior space' defined by the prenatal mapping of body form, sensory fields and movement space in integrated neuronal nets of the brain (Trevarthen, 2001). This innate intentional subjectivity is the foundation for neonatal intersubjectivity or communication of intentions (Trevarthen, 1979).

There has been much discussion about whether newborns have 'feelings of emotion' and a capacity to imitate their expression. Tiffany Field and colleagues showed that newborns can imitate an actor's expressions of happiness, surprise, and sadness, and they studied changes in attentiveness and imitation of infants from 2 to 6 months of age (Field et al., 1982). Imitation of other infants' crying, called 'contageous' has been interpreted as an automatic response 'confusing' self-stimulation with stimulation from another. Close observation of expressions of the face while newborns imitate a variety of expressions of different parts of the body supports the theory that they feel emotions of interest and pleasure in the 'intersubjective, imitative game' (Kugiumutzakis, 1998: p. 78; **Figure 1**).

Jacqueline Nadel and collaborators have demonstrated that a newborn infant can imitate two-dimensional images of tongue protrusion from a mouth seen on a screen, or the movements of a robot protruding a tongue (Nadel, 2014). The infants move with effort, selectively attending to and matching the artificial visual image of part of a human being to their own mouth moving. The dynamics of the images must be important, but clearly the infant has, within the first few days, a capacity to focus attention on an isolated feature of the face, and identify it with their own corresponding organ. The infant brain has 'feature detectors' for eyes, mouth, and tongue. Dynamic configurations of a smile or scowl on the whole face can also be detected as signals of different emotions.

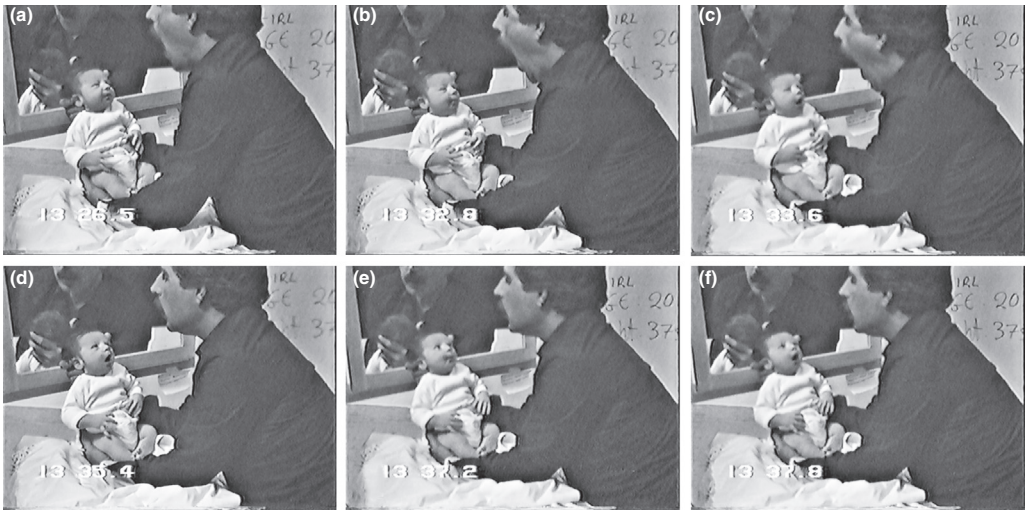

Figure 1 A cycle of imitations of mouth opening with a female infant 20 min after birth. Recorded in a maternity hospital in Herakleion, Crete in 1983 by Giannis Kugiumutzakis for his PhD research at the University of Uppsala. (a) (0 s) The researcher presents a wide open mouth for the first time to the attentive infant, focusing on his mouth, and with slightly closed eyes and pursed mouth. (b) (+6.3 s) The researcher opens his mouth for the fourth time. The neonate continues to observe his mouth with evident interest. The right hand moves up. (c) (+0.8 s) The researcher opens his mouth for the fifth time. The neonate imitates him once, synchronously while watching his mouth. The right hand closes. (d) (+1.8 s) The infant imitates a second time, looking up at the researcher's eyes as he waits. (e) (+1.8 s) Both pause, waiting. The infant is still looking at his eyes. (f) (+0.6 s) The infant makes a third large imitation while looking at the researcher's mouth.

Using a combination of naturalistic or spontaneously timed interactions and experimental or controlled methods of observation, Kugiumutzakis examined imitation by studying 170 infants within 10–45 min after full term or preterm births, most natural but some by Caesarian section. He presented tongue protrusion and mouth opening, blinking of his eyes, and three sounds, /a/, /m/, and /ang/ to alert infants (**Figure 1**). His findings are based on what he experienced in engagement with the infants while presenting the models, and on analysis of the videos. Some responses were also acts in spontaneous behavior of the infants, but these were produced significantly more often in the presence of the model. All the models of face movements seen were imitated by full terms or preterms, regardless of the method of birth. No differences were found between male and female infants. The babies tried to imitate the sound /a/, with mouth opening and occasional sounds, but they did not try to imitate the sounds, /m/ and /ang/.

Most neonates made a visible effort to direct their attention, first by staring, then by visual exploration directed to the moving part of the experimenter's face, or to the open mouth a sound. The baby attended with *frowning, a movement that* Darwin said shows the mind to be 'intent on one object' (Darwin, 1938 Notebook, cited in Gruber, 1974: p. 341). The periods of visual exploration and frowning of premature neonates were longer than those of the full-term infants. During the presentation of the sound a model, the majority of the imitators turned their heads to localize the sound, with eye widening while the brows were held high, a sign of attentive listening. A few full-term neonates glanced at a face movement briefly, then reproduced the model in 3 or 4 s. Some responded immediately after the first or the second presentation of the /a/ sound. Those that imitated the blinking model, attended more slowly (range 8–19 s) than the imitations of tongue protrusion and mouth opening.

None of these behaviors are automatic reflexes or fixed action patterns. Sometimes the neonates reproduce the models once, matching the model's action accurately. The majority of imitations of the eye-movement were of this kind. More often, with tongue protrusion and mouth opening, the model movement was repeated several times as a series of attempts becoming progressively more accurate, sometimes beginning with exploratory movements of the tongue inside the buccal cavity. This confirms observations of Maratos (1973) and Meltzoff and Moore (1977) that the infants 'try' to improve their imitations.

In some cases, the first imitative effort was a clear reproduction of the model, with additional 'efforts' becoming less accurate as if interest was lost. Tongue protrusion was imitated more easily than mouth opening, and they imitated these oral models more frequently than the eye blinking, which evoked awkward movements of the eyelids. Imitative responses to the sound /a/ were strained, usually by an intense explosion of a prolonged and unstructured /a/ sound accompanied by stretching hand movements and closed eyes. A considerable number of cases failed to imitate, again indicating that the responses are not simple reflexes.

A major advance in experimentation with the motives and emotions of neonates for communication was made by a Hungarian medical doctor and psychologist Emese Nagy who, in her PhD research with Peter Molnár in a hospital in Szeged, offered a receptive interpersonal approach in tests that proved the infants imitated to take part in a dialogue (Nagy and Molnár, 2004). She obtained imitation of an index finger extension, then paused looking at the infant. Newborns, while waiting for the experimenter's response, spontaneously produced the previously imitated gesture, initiating a 'provocation.' Measurements of heart beats revealed that the effort of imitation was accompanied by heart rate increase, a sign of excitement, while presentation of a gesture as provocation was accompanied by heart rate deceleration, indicative of attention. Infants are not only capable of imitating a model gesture, but can also invite an imitative response, thus sustaining an emotionally charged interaction. These findings constitute a laboratory demonstration of the first dialogue (Nagy, 2006, 2011), and they show how human 'imprinting' or affectionate attachment begins. Neonatal imitation is a precursor of the active intersubjective regulation of protoconversations with alert infants in the second month (Trevarthen, 1977, 1979, 1986; Malloch and Trevarthen, 2009).

The Key Features of Neonatal Imitation

Not all newborns imitate. Imitation is favored by an alert awake state with comfortable posture and sensitive engagement with the other person. Repeated stimulations in experimental tests of imitation of intended signs may inhibit the infant's readiness to 'play the game of imitation.' Premature infants have difficulty regulating their state of attention and body coordination, and their immaturity restricts their capacity to imitate. The actions that are imitated may also be performed spontaneously by the newborn, with no incentive from a 'model.' The important fact is that such actions expressive of the infant's motive state may be significantly increased and developed through imitative exchange in 'felt immediacy' (Bråten, 2009).

Some actions, such as mouth opening, tongue protrusion, hand movement, are imitated often and quickly, sometimes accurately in the first performance. Others, including eye closing, simple back vowels, different finger extensions, are less frequent and often slower with progressive improvement of accuracy of reproduction; more complex vocalizations requiring control of movements of the front of the mouth and tongue are not imitated.

Imitations are accompanied by movements indicative of effort and concentration. General body activity, staring, frowning, grimacing, emotional expressions of focused 'interest' and 'pleasure,' and complex hand gestures, precede and accompany imitations (**Figure 1**). Infants search for reactions to their communicative expressions, and this leads to repetition of imitated acts as 'provocations' for the other to imitate. Nagy's measures of heart beat show that neonates are physiologically aroused by emotions, when they are sharing imitations as purposeful expressions that appreciate response – imitation being preceded by heart rate acceleration, and provocation by heart deceleration.

Finally, in all natural, attentive and sympathetic engagements of expressive movements with alert newborn infants there is *reciprocal imitation* in *rhythmically patterned interchange*. Parents happily imitate too, and do so *more than the infants*.

These behaviors, timed to achieve intersynchrony between intentions and experiences, are necessary for the development of cooperative involvement of purposes with objects and their affective appraisal, for constructive projects, and for creating conventional narratives in art and with symbols. They are adapted for cultural learning by 'intent participation' in *sympathetic shared action and experience*.

Neurobiology of Neonatal Imitation or 'Primary Dialogue'

The brain mechanisms for self-awareness in body movement and body form are, with the corresponding peripheral sensory motor organs, formed for shared self–other awareness before birth (Trevarthen, 2001), when the embryo and fetus are in intense physiological contact with the mother's body and in 'amphoteronomic' or shared autonomic regulation (Trevarthen et al., 2006). Different motor and sensory capacities form at different stages of development. The first sensations of movement are by within-body proprioception and by self-touching of the hands and with the face and mouth (Piontelli, 2010; Trevarthen and Delafield-Butt, 2013). Discriminating awareness of another individual is first by taste and touch, then, in the last trimester by hearing. Sight is functional before term and may contribute to both self-awareness and other-awareness within minutes of birth. The sensory motor systems for self-regulation, of eyes, face, voice and hands, are adapted for transformation to serve dialogic or synrhythmic self–other regulation from birth (Trevarthen et al., 2006; Nagy, 2011). The movements most readily imitated, notably round the mouth, are those with larger representation in the newborn brain and with better developed muscular and sensory systems.

The intermodal matching between proprioception of a face movement of the self and vision of that movement in another person, the AIM hypothesis, is necessary but not sufficient to explain imitation, which depends on matching organized intentional motor systems that seek confirmation by different forms of reafference. Motor impulses are regulated in time by a hierarchy of rhythms generated in the brain and body (Malloch and Trevarthen, 2009), by parameters of active proprioception and exproprioception leads to 'alteroception' or awareness of the motives and expectations of another individual in 'felt immediacy' (Bråten, 2009).

An intention generates anticipation of correctly timed sensory feedback as 'experience.' In dialogue, the form of the movement displays this intention and leads apprehension of the experience of the other in 'synrhythmia' (Trevarthen et al., 2006), which includes both synchrony and concordant alternation or shared serial ordering of intentions, with affective appraisal. This results in the two-way communication of 'sympathy,' which must be an interchange of complementary affective states, not the unilateral understanding of 'empathy' (Reddy and Trevarthen, 2004).

Conclusion

Our abstract notes that, "imitation is essential for cultural learning, language and conventional reasoning, but its development been controversial for millennia, and remains problematic …" If attention is restricted to the artificial products of human collaboration and invention, to cultural history, the biological process of development that generates shared experience is ignored. Evidence we have reviewed indicates all that cultural intelligence achieves, and the well-being of individuals in human relationships, depend upon the environment-expectant motor intelligence we are born with, and the ability and infant has to sense, respond to, and share the motives and feelings implicit in the movements of another person. This ability is impulsive within the bodies and minds of infant and adult, with the matching rhythms and sympathetic affective regulations. They join in the hypermotility and 'vitality dynamics' of human actions, with story making imagination of 'communicative musicality,' a combination of gestural and vocal semantics, of dance and song (Malloch and Trevarthen, 2009; Stern, 2010).

A neonate's imitation of an expression offered by another person with attention for response, establishes a first cooperative enterprise in shared awareness of life. With no interest in technical artifacts or symbols, and limited by the infant's immaturity and lack of world knowledge, it confirms intimate engagement with a person who is present in the infant's awareness as company. This individual may have articulate assumptions or beliefs that preclude acceptance of the infant's movements as acts motivated for communication – a philosopher or medical doctor may believe the infant has a sensitive body but no mind; a psychologist may look only for signs of physiological arousal.

The infant's imitation may, however, be more warmly accepted with intuitive affection – as a request for milk by a mother wanting to breast feed, as a sign of interest in telling her something by a proud grandmother, as a joke by a sibling seeking a playmate. It becomes a two-way mental event, an act of primary intersubjectivity expecting confirmation – not a sensory motor circular reaction within the body of the Self, but one coupled to another's awareness of Self in an intersubjective loop of 'vitality dynamics' with emotional 'attunement' (Stern, 2000).

From the start, imitation supports collaborative adventures in moving, even though it has limited scope for imagination of goals or objects for action. The infant's mimesis is playful, motivated by interest and pleasure in sharing (Turner, 1982; Donald, 2001). It is made not just to be a single act of shared 'like me' experience, but becomes part of a 'story' or 'plot' that 'grows meaning' over tens of seconds in mutually sustained expectation of a rhythmic project of communication, engaging two persons in a nonverbal 'narrative' (**Figure 1**). Neonatal imitation is both 'autopoetic' and innately dialogical, self-making activity of a human being who has need for company – a biological capacity for the growth of a sociable 'consensuality' upon which all cultural understanding and all languages of human communities depend (Maturana and Varela, 1987). It is 'heteropoetic,' originating from a mind seeking, cocreating, and sharing an intersubjective imitative game adapted to make relationships in a community.

Most reports of neonatal imitation have been made with Caucasian infants in Europe or the United States, in technical and rational European cultures, which value intelligence that is guided by laws of behavior to promote practical and social conformity among autonomous individual minds, rather than the intuitive prosocial life of community (Bierhoff, 2002).

Kugiumutzakis (1985) in his research in Crete observed facial imitation by two black Afro-American infants in the first day after birth – a 1-month premature neonate, and a baby born full term. A study made by Reissland (1988) in the Himalayas of Nepal, with Indo-Aryan infants of Maithil parentage who imitated lip movements of the author within the first hour after birth, confirms that the infant's motives and the adult human response to them transcend differences in race, and also in beliefs about the nature of human intelligence and traditional rules of life to sustain acquired rituals of technical creativity and social responsibility.

See also: Biological Preconditions for Language Development; Conscious Control During Childhood, Development of; Cultural Views of Life Phases; Embodied Social Cognition; Emotion and Expression; Face Perception Development During Early Childhood; Imitation During Infancy and Early Childhood; Implicit Social Cognition; Infancy and Human Development; Infant Gaze Behaviors; Intentionality During Infancy and Early Childhood, Development of; Meltzoff's "Like Me" Theory; Mirror Neurons, Theory of; Motivation and Actions, Psychology of; Motivational Development: Cross-Cultural Perspectives; Motor Skills: Development in Infancy and Early Childhood; Neonatal Bonding; Neonatal Temperament: From Phenomenology to Mechanism; Neural Correlates of Social Development; Piaget's Theory of Cognitive Development; Prosocial Behavior During Infancy and Early Childhood: Developmental Patterns and Cultural Variations; Self in Culture: Early Development; Self-Determination Theory; Self-Efficacy; Selfconscious Emotions, Psychology of; Wisdom, Psychology of.

Bibliography

Ammaniti, M., Gallese, V., 2014. The Birth of Intersubjectivity: Psychodynamics, Neurobiology, and the Self. Norton, New York.

Baldwin, J.M., 1897. Mental Development in the Child and the Race. Macmillan, New York.

Bierhoff, H.-W., 2002. Prosocial Behaviour. Psychology Press, Hove, East Sussex.

Bråten, S. (Ed.), 1998. Intersubjective Communication and Emotion in Early Ontogeny. Cambridge University Press, Cambridge.

Bråten, S., 2009. The Intersubjective Mirror in Infant Learning and Evolution of Speech. John Benjamins, Amsterdam/Philadelphia.

Brazelton, T.B., 1979. Evidence of communication during neonatal behavioural assessment. In: Bullowa, M. (Ed.), Before Speech: The Beginning of Human Communication. Cambridge University Press, London, pp. 79–88.

Butterworth, G., 1999. Neonatal imitation: existence, mechanisms and motives. In: Nadel, J., Butterworth, G. (Eds.), Imitation in Infancy. Cambridge University Press, Cambridge, pp. 63–88.

Donald, M., 2001. A Mind so Rare: The Evolution of Human Consciousness. Norton, New York/London.

Field, T., Woodson, R., Greenberg, R., Cohen, D., 1982. Discrimination and imitation of facial expressions by neonates. Science 218 (8), 179–181.

Gruber, H.E., 1974. Darwin on Man: A Psychological Study of Scientific Creativity. Wildwood, London.

Kugiumutzakis, G., 1985. The Origin, Development and Function of the Early Infant Imitation. Department of Psychology, Uppsala University, Ph.D. Thesis (Acta Universitatis Uppsaliensis, 35). University of Uppsala, Uppsala.

Kugiumutzakis, G., 1998. Neonatal imitation in the inter-subjective companion space. In: Braten, S. (Ed.), Intersubjective Communication and Emotion in Ontogeny. Cambridge University Press, New York, pp. 63–88.

Kugiumutzakis, G., 1999. Genesis and development of early infant mimesis to facial and vocal models. In: Nadel, J., Butterworth, G. (Eds.), Imitation in Infancy. Cambridge University Press, Cambridge, pp. 127–185.

Kugiumutzakis, G., Kokkinaki, T., Markodimitraki, M., Vitalaki, E., 2005. Emotions in early mimesis. In: Nadel, J., Muir, D. (Eds.), Emotional Development: Recent Research Advances. Oxford University Press, Oxford, pp. 161–182.

Malloch, S., Trevarthen, C., 2009. Communicative Musicality: Exploring the Basis of Human Companionship. Oxford University Press, Oxford.

Maratos, O., 1973. The Origin and Development of Imitation in the First Six Months of Life. Ph.D. Dissertation. University of Geneva.

Maturana, H.R., Varela, F.J., 1987. The Tree of Knowledge. The Biological Roots of Human Understanding. Shambhala, Boston/London.

Meltzoff, A.N., 2002. Elements of a developmental theory of imitation. In: Meltzoff, A.N., Prinz, W. (Eds.), The Imitative Mind: Development, Evolution, and Brain Bases. Cambridge University Press, Cambridge, pp. 19–41.

Meltzoff, A.N., Moore, M.K., 1977. Imitation of facial and manual gestures by human neonates. Science 198, 75–78.

Nadel, J., 2014. How Imitation Boosts Development, in Young Infants and in Children with Autism Spectrum Disorders. Oxford University Press, Oxford.

Nagy, E., 2006. From imitation to conversation: the first dialogues with human neonates. Infant and Child Development 15, 223–232.

Nagy, E., 2011. The newborn infant: a missing stage in developmental psychology. Infant and Child Development 20, 3–19. http://dx.doi.org/10.1002/icd.683.

Nagy, E., Molnár, P., 2004. Homo imitans or homo provocans? The phenomenon of neonatal initiation. Infant Behavior and Development 27, 57–63.

Panksepp, J., 2009. Brain emotional systems and qualities of mental life: from animal models of affect to implications for psychotherapeutics. In: Fosha, D., Siegel, D.J., Solomon, M.F. (Eds.), The Healing Power of Emotion: Affective Neuroscience, Development, and Clinical Practice. Norton, New York, pp. 1–26.

Piaget, J., 1962. Play, Dreams and Imitation in Childhood. Routledge and Kegan Paul, London.

Piontelli, A., 2010. Development of Normal Fetal Movements: The First 25 Weeks of Gestation. Springer-Verlag, Wien and New York.

Reddy, V., 2008. How Infants Know Minds. Harvard University Press, Cambridge, MA and London.

Reddy, V., Trevarthen, C., 2004. What we learn about babies from engaging with their emotions. Zero to Three 24 (3), 9–15.

Reissland, N., 1988. Neonatal imitation in the first hour of life: observations in rural Nepal. Developmental Psychology 24, 464–469.

Stern, D.N., 2000. The Interpersonal World of the Infant: A View from Psychoanalysis and Development Psychology, second ed., with new Introduction. Basic Books, New York.

Stern, D.N., 2010. Forms of Vitality: Exploring Dynamic Experience in Psychology, the Arts, Psychotheraphy, and Development. Oxford University Press, Oxford.

Tiedemann, D., 1787. Beobachtungen uber die Entwicklung der Seelenfahigeiten bei Kindern. Altenburg (English Translation by S. Langer, C. Murchison, 1927). Observations on the development of the mental faculties of children. Pedagogical Seminary and Journal of Genetic Psychology 34, 205–230. In: Dennis, W. (Ed.), Historical readings in developmental psychology. Appleton Century Crofts, New York (1971).

Trevarthen, C., 1977. Descriptive analyses of infant communication behavior. In: Schaffer, H.R. (Ed.), Studies in Mother-Infant Interaction: The Loch Lomond Symposium. Academic Press, London.

Trevarthen, C., 1979. Communication and cooperation in early infancy. A description of primary intersubjectivity. In: Bullowa, M. (Ed.), Before Speech: The Beginning of Human Communication. Cambridge University Press, Cambridge, London, pp. 227–270, 321–347.

Trevarthen, C., 1986. Development of intersubjective motor control in infants. In: Wade, M.G., Whiting, H.T.A. (Eds.), Motor Development in Children: Aspects of Coordination and Control. Martinus Nijhof, Dordrecht, pp. 227–270.

Trevarthen, C., 2001. The neurobiology of early communication: intersubjective regulations in human brain development. In: Kalverboer, A.F., Gramsbergen, A. (Eds.), Handbook on Brain and Behavior in Human Development. Kluwer, Dordrecht, The Netherlands, pp. 841–882.

Trevarthen, C., Delafield-Butt, J., 2013. Biology of shared experience and language development: regulations for the inter-subjective life of narratives. In: Legerstee, M., Haley, D., Bornstein, M. (Eds.), The Infant Mind: Origins of the Social Brain. Guildford Press, New York, pp. 167–199.

Trevarthen, C., Aitken, K.J., Vandekerckhove, M., Delafield-Butt, J., Nagy, E., 2006. Collaborative regulations of vitality in early childhood: stress in intimate relationships and postnatal psychopathology. In: Cicchetti, D., Cohen, D.J. (Eds.), Developmental Psychopathology, Developmental Neuroscience, second ed., vol. 2. Wileys, New York, pp. 65–126.

Turner, V., 1982. From Ritual to Theatre: The Human Seriousness of Play. Performing Arts Journal Publications, New York.

Von Uexküll, J., 1957. A stroll through the worlds of animals and men. In: Schiller, C.H. (Ed.), Instinctive Behavior – the Development of a Modern Concept. International Universities Press, New York, pp. 5–80.

Neonatal Social Isolation: Effects on Neural Function

Takuya Takahashi, Graduate School of Medicine, Yokohama City University, Yokohama, Japan; and Albert Einstein College of Medicine, Bronx, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Neonatal stressful environment could influence the establishment of neuronal circuit formation and result in the various psychiatric disorders such as depression, anxiety disorders, and personality disorders. Despite their fundamental importance, the molecular and cellular mechanisms underlying how neonatal stress affects nervous system remain poorly understood. Here I review about how neonatal social isolation stress, an environment caused by neglect, alters developmental circuit formation in the cortex. Neonatal social isolation disrupts experience-dependent synaptic AMPA receptor delivery in the somatosensory cortex, leading to disturbance in somatosensory functions. These recent findings could shed light on the understanding of social stress-induced mental illness.

Introduction

Neonatal mental stress can have profound effects on the nervous system (*see* Stress in Childhood: Effects on Development). In particular, social isolation early in life, an environment caused by neglect (a form of child abuse), can induce mental stress and lead to various psychiatric disorders such as depression, drug addiction, and borderline personality disorders (Insel and Fernald, 2004; Khonicheva et al., 2006). A number of studies have reported a positive relationship between child maltreatment and personality disorders such as borderline personality disorder, which exhibits prominent deficits in social behaviors (Allen et al., 2013; Bierer et al., 2003; Johnson et al., 2000; Tyrka et al., 2009; Widom et al., 2009) (*see* Borderline (Patient) Personality). Impulsive behaviors and emotional instability are common in borderline personality disorders; depression, aggression, substance or alcohol abuse, eating disorders, indiscriminate sex with multiple partners, quitting jobs or personal relationship, and self-injury (Sadock and Sadock, 1996). Borderline personality disorder affects 1–2% of population (Sadock and Sadock, 1996). Due to the complicated regulation of neuronal functions by stress (Stranahan et al., 2006; Wallace et al., 2009) (*see* Stress in Adolescence: Effects on Development), elucidation of how stress affects the nervous system at the molecular, cellular, and circuit levels is a crucial step toward the effective treatment of stress disorders by neonatal maltreatment.

Early postnatal experience establishes normal functioning cortical circuits (Buonomano and Merzenich, 1998; Fox, 2002). Sensory experience through whiskers, an important tool of social communication in rodents (Ahl, 1986; Strozik and Festing, 1981), leads to a prominent increase in excitatory synaptic transmission in the barrel cortex during first 2 weeks after birth (Takahashi et al., 2003).

Information from a neuron to other neurons is mediated by the structure, called synapse. α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA)-type glutamate receptors mainly mediate excitatory synaptic transmission in the central nervous system (Bredt and Nicoll, 2003; Hollmann and Heinemann, 1994). Upon stimulation of presynaptic neurons, glutamate is released from the presynaptic terminal and binds to glutamate receptors including AMPA receptors. AMPA receptors form ion channels, and binding of glutamate opens channel and induces ion flux. The ion flux through AMPA receptors changes the membrane potentials and conveys information to postsynaptic neurons. The number of synaptic AMPA receptors determines the synaptic responses of postsynaptic neurons and regulates neuronal activity (Barry and Ziff, 2002; Bredt and Nicoll, 2003; Kessels and Malinow, 2009; Malinow and Malenka, 2002; Scannevin and Huganir, 2000). Tetanic synaptic stimulation elevates synaptic responses. This phenomenon is called Long-term potentiation (LTP) and has been suggested to be the cellular basis of neuronal plasticity including learning and memory. LTP-inducing stimuli deliver GluA1-containing AMPA receptors into synapses, increase the number of synaptic AMPA receptors, and elevate synaptic response *in vitro* (Hayashi et al., 2000; Zhu et al., 2000). *In vivo*, whisker experience drives GluA1 (or GluA4)-containing AMPA receptors into synapses and potentiates synaptic responses in the developing barrel cortex, which is a cortical area receiving input from whiskers (Clem and Barth, 2006; Miyazaki et al., 2012a; Takahashi et al., 2003). Synaptic trafficking of GluA1-containing AMPA receptors contributes to behavioral modification in several systems such as hippocampus-dependent learning (Cui et al., 2008; Engblom et al., 2008; Hu et al., 2007; Humeau et al., 2007; Jitsuki et al., 2011; Kessels and Malinow, 2009; Lee et al., 2003; Mitsushima et al., 2011, 2013; Reisel et al., 2002; Rumpel et al., 2005; Tada et al., 2013; Vialou et al., 2010).

Previous studies have shown that stressful environments such as social isolation affect LTP (Kehoe and Bronzino, 1999; Kim and Diamond, 2002). While some stress paradigms attenuate LTP in the hippocampus (Diamond et al., 1994a; Diamond and Park, 2000; Foy et al., 1987; Xu et al., 1997), other protocols, such as brief exposure to a predator, enhance LTP (Hu et al., 2007). Thus, the effect of stress on synaptic plasticity may be dependent on the stress protocol. However, the molecular mechanism of how behavioral stress attenuates synaptic plasticity was still unclear.

Rodents and humans may display similar effects of neonatal maltreatment on the nervous system (Diamond

et al., 1994b; Etherton et al., 2009; McGowan et al., 2009). In this section, we review recent findings that neonatal isolation perturbs experience-driven synaptic AMPA receptor delivery in the rat barrel cortex and discuss about the potential mechanisms of how neonatal maltreatment induces mental disorder in human.

Social Isolation Disrupts Whisker Experience-Dependent Synaptic Delivery of AMPA Receptors

In order to study the effect of social isolation on receptor trafficking, we isolated rat pups from their mother and other siblings for 6 h day^{-1} from postnatal (P)7 to P11 (Zimmerberg et al., 2003). We expressed green fluorescent protein (GFP)-GluA1 in the layer 2/3 of the barrel cortex of either isolated or nonisolated siblings from the same litter at P12 with virus. Recombinant GluA1 forms homomeric receptors which show little outward current at positive potential (while endogenous AMPA receptors display current at both negative and positive potentials) and thus synaptic incorporation of recombinant AMPA receptors is detected as an increased rectification (ratio of response at −60 to +40 mV) of AMPA receptor-mediated synaptic transmission. Litters were maintained in normal condition after viral injection and acute coronal brain slices were prepared at P14. Animals were maintained with whiskers intact throughout the experiment. Whole-cell recordings of layer 2/3 infected pyramidal neurons and nearby noninfected neurons were conducted and evoked synaptic transmission by stimulating layer 4. GFP-GluA1 infected neurons from nonisolated control siblings showed increased rectification compared with nearby noninfected neurons (Miyazaki et al., 2012b). However, no increase of rectification was detected from GFP-GluA1 expressing neurons in isolated animals (Miyazaki et al., 2012b). Therefore, experience-driven synaptic delivery of GluA1 was prevented by social isolation at this early postnatal period. In order to examine if synaptic delivery of endogenous AMPA receptors was attenuated by social isolation, we monitored the ratio of AMPA to *N*-Methyl-D-aspartic acid (NMDA) responses (A/N). If the number of synaptic AMPA receptors is large, A/N ratio is high. In brain slices from P14 animals isolated between P7 and P11 for 6 h day^{-1}, the A/N ratio was significantly lower than that observed in nonisolated animals of the same age, indicating that social isolation from P7 to P11 disrupted synaptic delivery of endogenous AMPA receptors (Miyazaki et al., 2012b). Interestingly, mother separation for 6 h day^{-1} (but several pups were kept together and thus social interaction was maintained) or transient blockade of whisker input for 6 h day^{-1} during P7–P11 did not prevent synaptic AMPA receptor delivery (Miyazaki et al., 2012b). This suggested that isolation-induced prevention of AMPA receptor trafficking is not due to the lack of nutrition from the dam or potential reduction of sensory deprivation by isolation.

Whisker experience drives the delivery of GluA4 but not GluA1 to layer 4–2/3 pyramidal synapses in the barrel cortex during P8–P10, while GluA1 but not GluA4 is delivered to these synapses during P12–P14 (Miyazaki et al., 2012a). Social isolation for 6 h day^{-1} from P4 to P7 disrupted experience-driven synaptic GluA4 delivery during P8–P10 as well as GluA1 trafficking during P12–P14 (Miyazaki et al., 2012b, 2013).

Consistent with these results, social isolation for 6 h day^{-1} either during P4–P7 or P7–P11 prevented LTP induction at P14 (Miyazaki et al., 2012b). This also indicates that social isolation disrupts biochemical pathways necessary for the trafficking of AMPA receptors.

Thus, neonatal social isolation severely affects synaptic maturation throughout early postnatal development.

Glucocorticoid Mediates Social Isolation-Induced Disruption of AMPA Receptor Delivery into Synapses

Corticosterone is a major stress hormone in rodents. Since the active form of corticosterone is free corticosterone (rather than that bound to corticosterone binding globulin), we investigated the free blood corticosterone level. During social isolation, prolonged rise of free blood corticosterone level was observed (Miyazaki et al., 2012b).

In order to test if the rise of corticosterone during social isolation mediates the prevention of AMPA receptor delivery, we injected RU486, an antagonist of glucocorticoid receptor (GR), during 6 h day^{-1} social isolation from P7 to P11. We performed the same electrophysiological assay as described above and found that systemic injection of RU486 blocked the isolation-induced attenuation of experience-driven synaptic GluA1 delivery at layer 4–2/3 pyramidal synapses of the barrel cortex during P12–P14 (Miyazaki et al., 2012b). Application of RU486 also prevented isolation-induced blockade of GluA4 delivery during P8–P10 (Miyazaki et al., 2013). Further, local injection of RU486 into the barrel cortex during social isolation procedure prevents isolation-induced blockade of synaptic AMPA receptor delivery (Miyazaki et al., 2012b). When Glucocorticoid receptor (a receptor of stress glucocorticoid hormone) is activated, it is translocated to the nucleus. Neonatal isolation translocates glucocorticoid receptor into the nucleus of the layer 2/3 neurons of the barrel cortex, suggesting the activation of glucocorticoid signaling in the barrel cortex by neonatal social isolation (Miyazaki et al., 2012b). These indicate that corticosterone mediates isolation effect on synaptic AMPA receptor delivery directly at the barrel cortex. Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII), a crucial kinase for synaptic AMPA receptor delivery (Lisman et al., 2002), is also activated in the barrel cortex during neonatal social isolation in the glucocorticoid activation-dependent fashion (Miyazaki et al., 2012b). Interestingly, injection of propranolol, a blocker of β-adrenergic receptors (noradrenaline is another stress hormone), during social isolation did not prevent isolation-induced blockade of GluA1 delivery, indicating that noradrenaline does not mediate isolation effect on synaptic AMPA receptor delivery (Miyazaki et al., 2012b).

In order to elucidate further, the role of increased glucocorticoid on experience-dependent synaptic delivery of AMPA receptors, we examined if injection of glucocorticoid mimics the effect of social isolation. Animals injected with corticosterone from P7 to P11 were housed in normal environment. This procedure also blocked experience-driven synaptic GluA1 delivery (Miyazaki et al., 2012b). Injection

of corticosterone during P4–P7 also blocked GluA4 delivery (Miyazaki et al., 2013).

These results indicate that elevated glucocorticoid level mediates isolation-induced prevention of experience-driven AMPA receptor trafficking to synapses via the activation of CaMKII.

Whisker–Barrel Function Is Attenuated by Social Isolation

The disruption of synaptic GluA1 and GluA4 delivery in the barrel cortex could result in an attenuation of the sensitivity of each barrel to the corresponding whiskers. To examine this, we measured the sensitivity of individual neurons in response to the deflection of whiskers with *in vivo* tetrode recordings from a single barrel. Whiskers with the highest sensitivity were defined as principal whiskers. First, we measured the response to principal whiskers deflections of rats either with or without social isolation at juvenile age. We found that the response by the deflection of principal whiskers of isolated animals was significantly smaller than those of control animals, indicating that the sensitivity of the barrel to its principal whisker was attenuated (Miyazaki et al., 2012b). This effect of isolation is due to the attenuation of AMPA receptor delivery, since the expression of GluA1 delivery blocking peptide in the layer 2/3 of the developing rat barrel cortex also exhibited the same effect as social isolation (Miyazaki et al., 2012b).

We next examined the effects of the attenuated whisker sensitivity on behavior. We assessed the disruption of whisker sensitivity with the ability of an animal to detect distances using a modified version of the gap-crossing task, which is known to be dependent on the detection ability of whiskers (Barneoud et al., 1991; Hutson and Masterton, 1986). In this task, rats were initially trained to find a reward at the opposite end of a runway. After this training phase, a gap was inserted in the runway and was widened at 0.5 cm intervals. The gap was widened until rats could not step across it. This gap size was defined as the upper threshold (SU). Subsequently, the gap was narrowed at 0.5 cm intervals until rats could cross it. This gap size was defined as the lower threshold (SL). The difference between SU and SL was calculated (defined as ΔS). ΔS reflects the ability of animals to discriminate the distance. For this task, we used juvenile rats which experienced neonatal social isolation. We found that isolated animals exhibited higher value of ΔS than control animals, indicating that distance detection ability is attenuated by neonatal social isolation (Miyazaki et al., 2012b). This behavioral effect with isolation is glucocorticoid signaling dependent, since application of RU486 during neonatal social isolation rescued this behavioral deficit (Miyazaki et al., 2012b).

These findings suggest that neonatal social isolation results in the deficits in whisker sensitivity and whisker-dependent behavior.

Conclusion

While numerous studies have revealed profound effects of stress on neural function, few have comprehensively related molecular, synaptic, and circuit deficits to behavioral aberrations (Feder et al., 2009; Kim and Diamond, 2002; Roozendaal et al., 2009). In series of our recent studies, we found that the stress produced by neonatal social isolation disrupts synaptic delivery of AMPA receptors, circuit formation in the cortex, and behaviors related to somatosensory function (Miyazaki et al., 2012b, 2013). We focused on plasticity in the barrel cortex since it is likely to play an important role in the integration of social information. This region is the primary sensory cortical area receiving input from whiskers, which are important for social interaction (Ahl, 1986; Strozik and Festing, 1981). Future studies should be focused on other brain areas such as medial prefrontal cortex which could process social information.

Previous human studies proposed the relation between the child maltreatment such as neglect and personality disorders which exhibit various deficits in social behaviors (Allen et al., 2013; Huang et al., 2012; Minzenberg et al., 2006). Interestingly, majority of patients with borderline personality disorders, who often experienced childhood maltreatment, exhibit the deficits in somatosensory processing (Pavony and Lenzenweger, 2014), potentially related to frequent self-injury. Our recent findings show that traumatic experiences during development, such as social isolation, can cause disruption of experience-dependent synaptic delivery of AMPA receptors, aberrant circuit formation in the somatosensory cortex, and behavioral abnormalities. It will be interesting to determine how and if miss-sensory processing leads to the emotional deficits of patients with the history of childhood maltreatment. These studies could lead to the effective treatment for psychiatric patients with the history of childhood maltreatment.

See also: Emotional Development, Effects of Parenting and Family Structure on; Neonatal Bonding; Neonatal Crying Behaviors; Neonatal Imitation; Neonatal Temperament: From Phenomenology to Mechanism; Parenting Attitudes and Beliefs across Cultures; Prosocial Behavior, Effects of Parenting and Family Structure on; Stress: Effects on Neonatal and Child Development.

Bibliography

Ahl, A.S., 1986. The role of vibrissae in behavior: a status review. Veterinary Research Communications 10, 245–268.

Allen, B., Cramer, R.J., Harris, P.B., Rufino, K.A., 2013. Borderline personality symptomatology as a mediator of the link between child maltreatment and adult suicide potential. Archives of Suicide Research 17, 41–51.

Barneoud, P., Gyger, M., Andres, F., van der Loos, H., 1991. Vibrissa-related behavior in mice: transient effect of ablation of the barrel cortex. Behavioural Brain Research 44, 87–99.

Barry, M.F., Ziff, E.B., 2002. Receptor trafficking and the plasticity of excitatory synapses. Current Opinion in Neurobiology 12, 279–286.

Bierer, L.M., Yehuda, R., Schmeidler, J., Mitropoulou, V., New, A.S., Silverman, J.M., Siever, L.J., 2003. Abuse and neglect in childhood: relationship to personality disorder diagnoses. CNS Spectrums 8, 737–754.

Bredt, D.S., Nicoll, R.A., 2003. AMPA receptor trafficking at excitatory synapses. Neuron 40, 361–379.

Buonomano, D.V., Merzenich, M.M., 1998. Cortical plasticity: from synapses to maps. Annual Review of Neuroscience 21, 149–186.

Clem, R.L., Barth, A., 2006. Pathway-specific trafficking of native AMPARs by in vivo experience. Neuron 49, 663–670.

Cui, Y., Costa, R.M., Murphy, G.G., Elgersma, Y., Zhu, Y., Gutmann, D.H., Parada, L.F., Mody, I., Silva, A.J., 2008. Neurofibromin regulation of ERK signaling modulates GABA release and learning. Cell 135, 549–560.

Diamond, D.M., Fleshner, M., Rose, G.M., 1994a. Psychological stress repeatedly blocks hippocampal primed burst potentiation in behaving rats. Behavioural Brain Research 62, 1–9.

Diamond, M.E., Huang, W., Ebner, F.F., 1994b. Laminar comparison of somatosensory cortical plasticity. Science 265, 1885–1888.

Diamond, D.M., Park, C.R., 2000. Predator exposure produces retrograde amnesia and blocks synaptic plasticity. Progress toward understanding how the hippocampus is affected by stress. Annals of the New York Academy of Sciences 911, 453–455.

Engblom, D., Bilbao, A., Sanchis-Segura, C., Dahan, L., Perreau-Lenz, S., Balland, B., Parkitna, J.R., Lujan, R., Halbout, B., Mameli, M., et al., 2008. Glutamate receptors on dopamine neurons control the persistence of cocaine seeking. Neuron 59, 497–508.

Etherton, M.R., Blaiss, C.A., Powell, C.M., Sudhof, T.C., 2009. Mouse neurexin-1alpha deletion causes correlated electrophysiological and behavioral changes consistent with cognitive impairments. Proceedings of the National Academy of Sciences of the United States of America 106, 17998–18003.

Feder, A., Nestler, E.J., Charney, D.S., 2009. Psychobiology and molecular genetics of resilience. Nature Reviews Neuroscience 10, 446–457.

Fox, K., 2002. Anatomical pathways and molecular mechanisms for plasticity in the barrel cortex. Neuroscience 111, 799–814.

Foy, M.R., Stanton, M.E., Levine, S., Thompson, R.F., 1987. Behavioral stress impairs long-term potentiation in rodent hippocampus. Behavioral and Neural Biology 48, 138–149.

Hayashi, Y., Shi, S.H., Esteban, J.A., Piccini, A., Poncer, J.C., Malinow, R., 2000. Driving AMPA receptors into synapses by LTP and CaMKII: requirement for GluR1 and PDZ domain interaction. Science 287, 2262–2267.

Hollmann, M., Heinemann, S., 1994. Cloned glutamate receptors. Annual Review of Neuroscience 17, 31–108.

Hu, H., Real, E., Takamiya, K., Kang, M.G., Ledoux, J., Huganir, R.L., Malinow, R., 2007. Emotion enhances learning via norepinephrine regulation of AMPA-receptor trafficking. Cell 131, 160–173.

Huang, J., Yang, Y., Wu, J., Napolitano, L.A., Xi, Y., Cui, Y., 2012. Childhood abuse in Chinese patients with borderline personality disorder. Journal of Personality Disorders 26, 238–254.

Humeau, Y., Reisel, D., Johnson, A.W., Borchardt, T., Jensen, V., Gebhardt, C., Bosch, V., Gass, P., Bannerman, D.M., Good, M.A., et al., 2007. A pathway-specific function for different AMPA receptor subunits in amygdala long-term potentiation and fear conditioning. Journal of Neuroscience 27, 10947–10956.

Hutson, K.A., Masterton, R.B., 1986. The sensory contribution of a single vibrissa's cortical barrel. Journal of Neurophysiology 56, 1196–1223.

Insel, T.R., Fernald, R.D., 2004. How the brain processes social information: searching for the social brain. Annual Review of Neuroscience 27, 697–722.

Jitsuki, S., Takemoto, K., Kawasaki, T., Tada, H., Takahashi, A., Becamel, C., Sano, A., Yuzaki, M., Zukin, R.S., Ziff, E.B., et al., 2011. Serotonin mediates cross-modal reorganization of cortical circuits. Neuron 69, 780–792.

Johnson, J.G., Smailes, E.M., Cohen, P., Brown, J., Bernstein, D.P., 2000. Associations between four types of childhood neglect and personality disorder symptoms during adolescence and early adulthood: findings of a community-based longitudinal study. Journal of Personality Disorders 14, 171–187.

Kehoe, P., Bronzino, J.D., 1999. Neonatal stress alters LTP in freely moving male and female adult rats. Hippocampus 9, 651–658.

Kessels, H.W., Malinow, R., 2009. Synaptic AMPA receptor plasticity and behavior. Neuron 61, 340–350.

Khonicheva, N.M., Loseva, E.V., Chabak-Garbach, R., Loriya, M.V., Airapetyants, M.G., 2006. A special case of learning disorder in isolant rats as a model of disintegration. Neuroscience and Behavioral Physiology 36, 597–603.

Kim, J.J., Diamond, D.M., 2002. The stressed hippocampus, synaptic plasticity and lost memories. Nature Reviews Neuroscience 3, 453–462.

Lee, H.K., Takamiya, K., Han, J.S., Man, H., Kim, C.H., Rumbaugh, G., Yu, S., Ding, L., He, C., Petralia, R.S., et al., 2003. Phosphorylation of the AMPA receptor GluR1 subunit is required for synaptic plasticity and retention of spatial memory. Cell 112, 631–643.

Lisman, J., Schulman, H., Cline, H., 2002. The molecular basis of CaMKII function in synaptic and behavioural memory. Nature Reviews Neuroscience 3, 175–190.

Malinow, R., Malenka, R.C., 2002. AMPA receptor trafficking and synaptic plasticity. Annual Review of Neuroscience 25, 103–126.

McGowan, P.O., Sasaki, A., D'Alessio, A.C., Dymov, S., Labonte, B., Szyf, M., Turecki, G., Meaney, M.J., 2009. Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse. Nature Neuroscience 12, 342–348.

Minzenberg, M.J., Poole, J.H., Vinogradov, S., 2006. Adult social attachment disturbance is related to childhood maltreatment and current symptoms in borderline personality disorder. Journal of Nervous and Mental Disease 194, 341–348.

Mitsushima, D., Ishihara, K., Sano, A., Kessels, H.W., Takahashi, T., 2011. Contextual learning requires synaptic AMPA receptor delivery in the hippocampus. Proceedings of the National Academy of Sciences of the United States of America 108, 12503–12508.

Mitsushima, D., Sano, A., Takahashi, T., 2013. A cholinergic trigger drives learning-induced plasticity at hippocampal synapses. Nature Communications 4, 2760.

Miyazaki, T., Kunii, M., Jitsuki, S., Sano, A., Kuroiwa, Y., Takahashi, T., 2013. Social isolation perturbs experience-driven synaptic glutamate receptor subunit 4 delivery in the developing rat barrel cortex. European Journal of Neuroscience 37, 1602–1609.

Miyazaki, T., Kunii, M., Tada, H., Sano, A., Kuroiwa, Y., Goto, T., Malinow, R., Takahashi, T., 2012a. Developmental AMPA receptor subunit specificity during experience-driven synaptic plasticity in the rat barrel cortex. Brain Research 1435, 1–7.

Miyazaki, T., Takase, K., Nakajima, W., Tada, H., Ohya, D., Sano, A., Goto, T., Hirase, H., Malinow, R., Takahashi, T., 2012b. Disrupted cortical function underlies behavior dysfunction due to social isolation. Journal of Clinical Investigation 122, 2690–2701.

Pavony, M.T., Lenzenweger, M.F., 2014. Somatosensory processing and borderline personality disorder: pain perception and a signal detection analysis of proprioception and exteroceptive sensitivty. Journal of Personality Disorders 5, 164–171.

Reisel, D., Bannerman, D.M., Schmitt, W.B., Deacon, R.M., Flint, J., Borchardt, T., Seeburg, P.H., Rawlins, J.N., 2002. Spatial memory dissociations in mice lacking GluR1. Nature Neuroscience 5, 868–873.

Roozendaal, B., McEwen, B.S., Chattarji, S., 2009. Stress, memory and the amygdala. Nature Reviews 10, 423–433.

Rumpel, S., LeDoux, J., Zador, A., Malinow, R., 2005. Postsynaptic receptor trafficking underlying a form of associative learning. Science 308, 83–88.

Sadock, B.J., Sadock, V.A., 1996. Kaplan & Sadock's Synopsis of Psychiatry: Behavioral Sciences/Clinical Psychiatry, ninth ed.

Scannevin, R.H., Huganir, R.L., 2000. Postsynaptic organization and regulation of excitatory synapses. Nature Reviews Neuroscience 1, 133–141.

Stranahan, A.M., Khalil, D., Gould, E., 2006. Social isolation delays the positive effects of running on adult neurogenesis. Nature Neuroscience 9, 526–533.

Strozik, E., Festing, M.F., 1981. Whisker trimming in mice. Laboratory Animals 15, 309–312.

Tada, H., Kuroki, Y., Funabashi, T., Kamiya, Y., Goto, T., Suyama, K., Sano, A., Mitsushima, D., Etgen, A.M., Takahashi, T., 2013. Phasic synaptic incorporation of GluR2-lacking AMPA receptors at gonadotropin-releasing hormone neurons is involved in the generation of the luteinizing hormone surge in female rats. Neuroscience 248, 664–669.

Takahashi, T., Svoboda, K., Malinow, R., 2003. Experience strengthening transmission by driving AMPA receptors into synapses. Science 299, 1585–1588.

Tyrka, A.R., Wyche, M.C., Kelly, M.M., Price, L.H., Carpenter, L.L., 2009. Childhood maltreatment and adult personality disorder symptoms: influence of maltreatment type. Psychiatry Research 165, 281–287.

Vialou, V., Robison, A.J., Laplant, Q.C., Covington 3rd, H.E., Dietz, D.M., Ohnishi, Y.N., Mouzon, E., Rush 3rd, A.J., Watts, E.L., Wallace, D.L., et al., 2010. DeltaFosB in brain reward circuits mediates resilience to stress and antidepressant responses. Nature Neuroscience 13, 745–752.

Wallace, D.L., Han, M.H., Graham, D.L., Green, T.A., Vialou, V., Iniguez, S.D., Cao, J.L., Kirk, A., Chakravarty, S., Kumar, A., et al., 2009. CREB regulation of nucleus accumbens excitability mediates social isolation-induced behavioral deficits. Nature Neuroscience 12, 200–209.

Widom, C.S., Czaja, S.J., Paris, J., 2009. A prospective investigation of borderline personality disorder in abused and neglected children followed up into adulthood. Journal of Personality Disorders 23, 433–446.

Xu, L., Anwyl, R., Rowan, M.J., 1997. Behavioural stress facilitates the induction of long-term depression in the hippocampus. Nature 387, 497–500.

Zhu, J.J., Esteban, J.A., Hayashi, Y., Malinow, R., 2000. Postnatal synaptic potentiation: delivery of GluR4-containing AMPA receptors by spontaneous activity. Nature Neuroscience 3, 1098–1106.

Zimmerberg, B., Rosenthal, A.J., Stark, A.C., 2003. Neonatal social isolation alters both maternal and pup behaviors in rats. Developmental Psychobiology 42, 52–63.

Neonatal Temperament: From Phenomenology to Mechanism

Hiroyuki Tsuchiya, Fukuda Hospital, Kumamoto, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Each neonate has a different personality/temperament that gradually develops during the perinatal period. I report that the crying duration at the first bathing is peculiar to each neonate. Neonates who cry longer at their first bathing cry longer at bathing on subsequent days. The crying tendency relates to the withdrawal behavior of neonates, which later will turn into fear and anxiety.

Fear and anxiety are personality/temperament traits that are, in part, regulated by the serotonin system. Genetic polymorphisms in proteins that participate in the synthesis, transport, and degradation of serotonin; and its receptors can alter neurotransmitter levels, especially in the amygdala, which is a central locus of fear and anxiety regulation. Analyzing the effects of these polymorphisms on traits is called the candidate gene approach.

Three representative approaches can be used to assess the role of genetic regulation of the personality/temperament traits. The first approach is the candidate gene approach, which is exemplified by the case of the serotonin system. The second approach includes whole genome sequence, microarray, and genome-wide association studies. The third approach is from the different standpoints: epistasis (gene–gene interaction), gene–environment interaction, and epigenetics. So far, none of the methods adequately explain the heritability of personality/temperament.

Analyzing the relationship between personality/temperament and genetic variability in infants is advantageous because infants' temperaments are simpler than adults' personalities and are not yet strongly affected by the environment. It is not clear whether any of the earlier approaches will elucidate the genetic mechanism of temperament or whether we will have to wait until other methods are developed. However, clarifying this mechanism will increase our understanding of neonates and human personality and will help us raise healthy children in the future.

Introduction

Are babies born with blank brains, or are they already prepared for life in the world? This was a question until several decades ago. According to recent knowledge, babies are born in the latter condition (Nolen-Hoeksema et al., 2009). Babies acquire genes from their parents; their neural network is thus already more or less prepared, their physical and mental development is affected by the environment *in utero*, and their genes and environment influence each other.

There are many works of substance on the developmental psychology of infants. On the other hand, recent developments in neuroscience have brought us important knowledge of human character based on brain imaging and neurogenetics. At least some of the phenomenology in developmental psychology could be explained by neuroscience. However, there have been few studies connecting these two approaches.

I have found (Tsuchiya, 2011, 2013) that there are some differences in the activity of babies when they encounter a new situation and have concluded that these differences may be connected to their neuroscientific background. Here, I review the relationships of neonatal temperament, and preceding studies on developmental psychology, neural circuitry, and neurogenetics. Personality is composed of both temperament and character traits. Temperament refers to an individual's innate emotional predisposition, while character is what we make of ourselves intentionally (Cloninger, 1998). However, in this review, the words personality, temperament, and character are used not according to these strict definitions.

Understanding neonates can provide important insight when we aim to comprehend what is human. We can appreciate our position in evolution when we compare the developmental position of neonates with the positions of other animals. Therefore, the study of the psychology of neonates is of interest to everyone in our society, not only to scholars of developmental psychology.

Birds are a species that has evolved to have individual differences in temperament/character (Barber, 1993). If individual birds have a different temperament/character, is it not possible that every human neonate has a different temperament/character?

The Emergence of Neonatal Temperament

At a large obstetric hospital, I observed the bathing of many neonates by trained nurses every day. Neonates were first bathed on day 2 after birth and were then bathed every morning until day 5 (the day they were discharged). I noticed that there was a tendency for the neonates who cried during the first bathing to cry on the following days. To investigate this hypothesis, I studied the duration of crying in individual neonates under as uniform conditions as possible. This study revealed that neonates who cried during their first bathing were apt to cry on the following days. The correlation coefficient of crying duration between days 2 and 3 was $r = 0.410$, $p < .001$, $n = 100$, and between days 3 and 4, it was $r = 0.506$, $p = .001$, $n = 38$ (Tsuchiya, 2011).

Crying is a negative mood expression classified as a withdrawal reaction of 'approach or withdrawal,' which is one of nine categories of infant behavioral styles and temperaments characterized by Chess and Thomas (1996). From day 2 to day 3, the number of neonates who did not cry during bathing increased, and the latency before crying increased. These observations may be related to another category of infant temperament, 'adaptability' (Chess and Thomas, 1996). Accordingly, even neonates who were 2 days old have their own innate temperament. How is this temperament designed

or inscribed in neonates? In the following section, I review the present status of research on neonatal temperament.

Observations from Developmental Psychology

In *Atkinson and Hilgard's Introduction to Psychology* (Nolen-Hoeksema et al., 2009), perspectives in psychology are summarized in five areas: the behavioral, cognitive, psychoanalytic, subjectivist, and biological perspectives. I start from the behavioral perspective and connect it to the biological perspective. The biological perspective explains and shows the background of human behavior. This approach may also reveal fundamental information related to psychiatric disorders.

Researchers have endeavored to clarify human character and characteristic differences among people. The five trait factors, often called the Big Five, are one of the achievements in this field. The five factors are Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism, and they cover all aspects of personality (Nolen-Hoeksema et al., 2009). However, the factors are only applicable to personality in adults; until relatively recently, there has not been any appropriate approaches on temperament in children.

Chess and Thomas (New York Longitudinal Study)

One of the prominent works in child psychology was produced by Chess and Thomas (1996). They examined the characteristics of infants during the naturalistic observation of the infants' behavior (New York Longitudinal Study) and summarized the children's temperaments into nine categories: activity level, rhythmicity or regularity, approach or withdrawal, adaptability, threshold of responsiveness, intensity of reaction, quality of mood, distractibility, and attention span and persistence.

Using these categories, Chess and Thomas (1996) classified children into three groups: easy, difficult, and slow-to-warm-up. Approximately 40% of the subjects were easy children who were characterized by regularity in diurnal life, positivity to new stimuli, and high adaptability to change. Approximately 10% of children were difficult and exhibited irregularity, negativity, and nonadaptability. Slow-to-warm-up children (15%) showed negative responses to new stimuli but gradually showed positive responses to the same stimuli when they were encountered again. The remaining children (35%) were not categorized into any of the three groups.

Kagan and Colleagues

Kagan and colleagues studied the reactivity of infants to unexpected events from 4 months to 20 years of age (Kagan, 2010). They classified infants into four categories: high-reactive children (20% of the subjects) who were very excited when they encountered the unexpected stimuli, low-reactive children (40%) who were calm, distressed children (25%) who cried out loud but did not thrash their hands and legs, and aroused children (10%) who thrashed their hands and legs but did not cry out loud. The temperament of an infant, high or low reactivity, is apt to be kept until adolescence.

The response to unknown stimuli is regulated by the amygdala and the prefrontal cortex. The major functional roles of the amygdala are social behavior and the processing of novelty, emotion, and reward learning. The primate amygdala is a complex collection of 13 nuclei. The amygdala has an intricate internal architecture of discrete nuclei and widespread external connections (Schwartz et al., 2012). There are neuronal clusters in the amygdala that respond to any unexpected or unfamiliar event and alert other, nearby neurons to react appropriately. If the prefrontal cortex evaluates an event as being safe, it sends signals to the amygdala to stop the preparations for a fight or flight reaction.

Using functional magnetic resonance imaging (fMRI), Schwartz et al. (2012) showed that 4-month-old infants with high reactivity had greater activity in the amygdala in response to novel faces two decades later as adults. Infants with low reactivity showed the complementary pattern. The authors suggested that this temperamental difference already observed in early infants relates to individual differences in vulnerability and resistance to psychiatric disorders.

Rothbart and Colleagues

Rothbart (2011) summarized the development of temperament elements. Negative emotionality and orienting appear from the newborn period, frustration/anger and approach appear from 2 months, fear appears from 6 months, and effortful control appears from 10 months. These elements interact with each other. Differences in newborn orienting, activity, distress proneness, and soothability can be observed early in life and show some continuity with later temperament.

Continuity of Temperament in Life

The easy children (40% of children) in the Chess and Thomas study resemble the low-reactive children (40%) in the Kagan study; the difficult children (10%) of the former study look like the high reactives (20%) of the latter study. Temperament seems to persist from the infant period to at least adolescence. Roberts and DelVecchio (2000) reviewed 152 longitudinal studies, compiled 3217 test–retest correlation coefficients and concluded that "trait consistency increased from 0.31 in childhood to 0.54 during the college years, to 0.64 at age 30, and then reached a plateau around 0.74 between ages 50 and 70." The heritability of personality traits has been studied for a long time. Tellegen et al. (1988) reported heritability was estimated in the range 0.39–0.58.

From Phenotype to Genotype and Neural Circuitry: The Candidate Gene Approach

Cloninger's Model

Cloninger (1987) presented a hypothesis that related personality dimensions to neurotransmitters. There are three personality dimensions that are considered to be related to the brain's system of behavior: novelty seeking related to behavioral activation; harm avoidance related

to behavioral inhibition; and reward dependence related to behavioral maintenance. The corresponding neurotransmitters for these dimensions are dopamine, serotonin, and norepinephrine. This hypothesis showed a new avenue to understand the neural circuitry and neurotransmitters behind a behavior.

Hariri's Model

After Cloninger's proposal, Hariri (2009) reviewed the neurologic background of human behavior. He connected the behavioral trait, brain locus, neurotransmitter system, and functional polymorphism of the related genes. He identified three behavioral traits: trait anxiety, impulsivity, and trait anxiety and impulsivity. The responsible locus for trait anxiety is the amygdala, the neurotransmitter is serotonin, and the functional polymorphism of the serotonin system regulates this behavioral trait. Similarly, for impulsivity, the responsible locus is the ventral striatum, the neurotransmitter is dopamine, and the polymorphism of the dopamine transporter affects the dopamine level. For the combined anxiety and impulsivity trait, the responsible loci are the amygdala and ventral striatum; the polymorphism of the fatty acid amide hydrolase gene, which encodes an enzyme that mediates the hydrolytic breakdown of anandamide, reduces the enzyme activity and increases endocannabinnoid signaling.

Serotonin System

Among the personality traits, trait anxiety and its relationship to psychiatric status have been studied extensively, including the neurotransmitter involved and its genetic polymorphism and neural circuitry. Genetically manipulated mice models have also been examined (Fernandez and Gaspar, 2012; Li et al., 2012).

Neonates who are apt to cry at bathing (a new stimulation/situation) may be related to the difficult children in Chess and Thomas' study, Kagan's high-reactive children, and Rothbart's negative emotionality and trait anxiety. Accordingly, I review the serotonin (5-HT) system, which is a basis for trait anxiety.

5-HT is produced in neurons (Fernandez and Gaspar, 2012). In a presynaptic neuron, the amino acid tryptophan is hydroxylated to 5-hydroxy-tryptophan (5-HTP) by tryptophan hydroxylase 2 (TPH2). TPH2 is the rate-limiting enzyme for 5-HT synthesis in the brain (Berger et al., 2012). Then, 5-HTP is decarboxylated to 5-HT by aromatic amino acid decarboxylase. 5-HT is stored at the terminal of the presynaptic neuron, to be released by neurogenic stimuli in the neuron. At stimulation, 5-HT is released to the synaptic gap, where it binds to 5-HT receptors on the postsynaptic neuron. There are at least 17 distinct receptors, including 5-HT_{2A} (or expressed as HTR2A), 5-HT_{2C}, and 5-HT_3, that are located at the postsynaptic neuron surface. 5-HT_{1A} is located at both presynaptic (autoreceptor) and postsynaptic neurons (**Figure 1**).

Another protein involved in 5-HT neurotransmission is the 5-HT transporter (5-HTT). 5-HTT is located at the surface of presynaptic neurons and works to reuptake 5-HT released from the presynaptic neurons (Hariri and Holmes, 2006). 5-HTT removes 5-HT from the synaptic gap and returns it to the presynaptic neuron where 5-HT can be degraded or released later (Karg et al., 2011). Genetic polymorphisms of these proteins in the 5-HT neurotransmission system have been widely examined (Saudino and Wang, 2012). There are discrepancies among the reports: some show a correlation between the polymorphism and temperament or psychiatric conditions, and others do not. Among animal experiments, the congenic mouse model is a powerful tool in which the influences of a specific polymorphism of the target gene can be analyzed (Berger et al., 2012). However, the 5-HT system does not always produce the same results in different species: the elevation of 5-HT results in aggression in mice, but the consequences are the reverse in rats and silver foxes (Popova, 2006).

Tryptophan Hydroxylase 2

The effect of polymorphism of the TPH2 gene was described by Brown et al. (2005). They related the genotyping of the TPH2 gene to fMRI of the amygdala during a face-processing task in healthy human subjects, paying attention to the G(-844T) polymorphism, which has a high minor allele frequency (T allele = 38%). They found significantly higher bilateral dorsal amygdala activity in T allele carriers (G/T or T/T) than in G/G homozygotes.

The association of single-nucleotide polymorphism (SNP) in the TPH2 gene with the anxiety trait and with depression was studied in mice (Berger et al., 2012). The behavioral and neurochemical consequences of a C1473G polymorphism in the TPH2 gene were examined in the congenic mouse strain. Mice with a 1473 G/G polymorphism showed (1) a significant reduction of 5-HT synthesis, (2) a functional desensitization of 5-HT_{1A} autoreceptors without a reduction in basal or stress-induced 5-HT release, and (3) increased anxiety-related behavior without an alteration in depressive-like behavior.

Serotonin Receptor

5-HT_{1A} is expressed as a presynaptic autoreceptor on raphe neurons and as a major postsynaptic receptor in the hippocampal, cortical, and hypothalamic regions (Le François et al., 2008). The C(-1019)G 5-HT_{1A} promoter polymorphism is associated with serious depression and suicide. In depressed patients, the homozygous G(-1019) allele was enriched twofold vs controls (Lemonde et al., 2003). With this polymorphism, the function of the transcriptional factor NUDR, a 5-HT_{1A} receptor repressor in serotonergic raphe cells, is abrogated. Therefore, it has been proposed that the increased 5-HT_{1A} receptor expression, both pre- and postsynaptically decreased serotonergic neurotransmission, which is implicated in depression and suicide (Bismark et al., 2010). It is predicted that a significant portion (9.2%) of the individual differences in the trait anxiety is explained by this polymorphism (Hariri, 2009).

A relationship between the C(-1019)G 5-HT_{1A} promoter polymorphism and frontal brain electrical asymmetry has been shown (Bismark et al., 2010). In humans, the 5-HT_{1A} high-risk group (allele type GG) is associated with greater

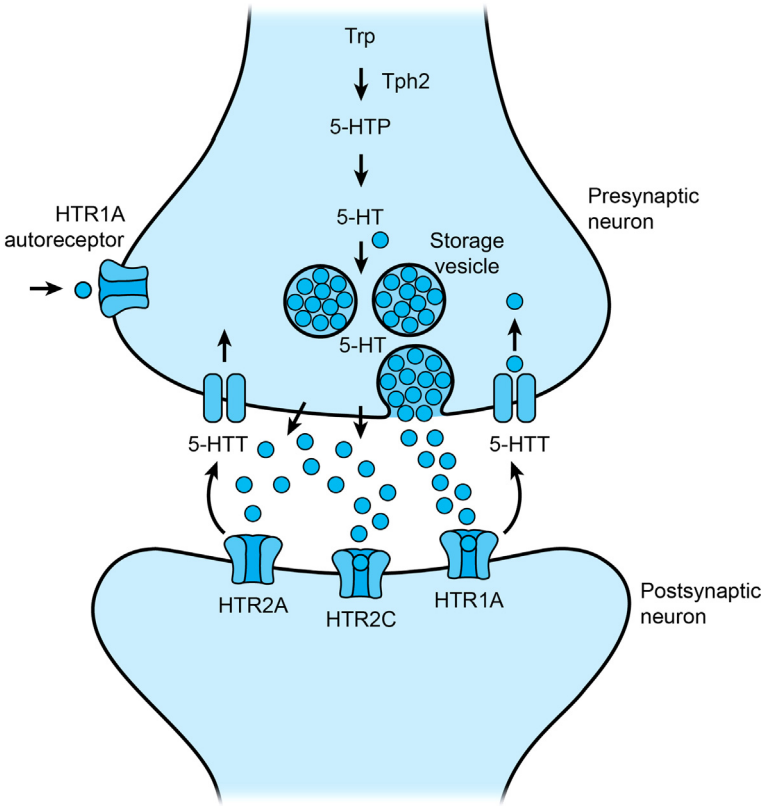

Figure 1 A diagram showing the serotonergic synapse. In a presynaptic neuron, serotonin is produced from tryptophan via 5-hydroxy-tryptophan by tryptophan hydroxylase 2. Serotonin is released from the presynaptic neuron and binds to postsynaptic serotonin receptors. The re-uptake of serotonin into the presynaptic neuron is mediated by the serotonin transporter. Accordingly, the serotonergic signals from presynaptic neurons are regulated by the activities of tryptophan hydroxylase 2, serotonin receptors on presynaptic neurons (5-HT1A and 5-HT1B; autoreceptors) and post-synaptic neurons (e.g., 5-HT1A, 5-HT2A, 5-HT2C, and 5-HT3), and the serotonin transporter. 5-HT1A is located at the cell bodies, and 5-HT1B is located at the exon terminals (McDevitt, R.A., Neumaier, J.F., 2011. Regulation of dorsal raphe nucleus function by serotonin autoreceptors: a behavioral perspective. Journal of Chemical Neuroanatomy 41, 234–246). Gene polymorphisms in these proteins affect serotonergic signal transmission (see the main text). For serotonin receptors, only the representative receptors are shown. Adapted from Fernandez, S.P., Gaspar, P., 2012. Investigating anxiety and depressive-like phenotypes in genetic mouse models of serotonin depletion. Neuropharmacology 62, 144–154 and Hariri, A.R., Holmes, A., 2006. Genetics of emotional regulation: the role of the serotonin transporter in neural function. Trends in Cognitive Sciences 10, 182–191. For a further review, see the figures of Cools, R., Roberts, A.C., Robbins, T.W., 2007. Serotoninergic regulation of emotional and behavioural control processes. Trends in Cognitive Sciences 12, 31–40 and Canli, T., Lesch, K.P., 2007. Long story short: the serotonin transporter in emotion regulation and social cognition. Nature Neuroscience 10, 1103–1109. Trp, tryptophan; Tph2, tryptophan hydroxylase 2; 5-HTP, 5-hydroxy-tryptophan; 5-HT, serotonin; HTR1A, serotonin receptor 1A; 5-HTT, serotonin transporter.

right frontal activity compared with the low-risk groups (allele types CC and CG). Greater right frontal activity is also associated with individuals who have a history of major depression disorder compared with individuals without such history. Kishi et al. (2011) showed an association between C(-1019)G 5-HT_{1A} SNP and bipolar disorder by meta-analysis.

In mouse experiments, Li et al. (2012) injected recombinant adenoviruses containing a 5-HT_{1A} antisense sequence or a 5-HT_{2C} sense sequence into the amygdala. In the 5-HT_{1A} antisense sequence-injected mice, the reduction of 5-HT_{1A} receptors in the amygdala increased anxiety-like behaviors. In the 5-HT_{2C} sense sequence-injected mice, the increase of 5-HT_{2C} receptors in the amygdala also increased anxiety-like behaviors.

A relationship between different types of thinking (i.e., holistic vs analytic) and the C(-1019)G 5-HT_{1A} polymorphism has also been reported (Kim et al., 2010). These researchers suggested that cultural differences influence the effect of polymorphism. Koreans with the GG genotype showed higher attention to context than Koreans with the CC and CG genotypes. In contrast, European Americans with the GG genotype showed lower attention to context than European Americans with the CC and CG genotypes. These researchers concluded that the same genetic predisposition can result in divergent psychological outcomes, depending on an individual's cultural context and that this result shows a gene by culture interaction effect, which is an example of gene–environment interaction described later. Kim et al. (2010) also highlighted the difference in the allelic distribution between Koreans and European Americans. A higher proportion of the population has the GG allele in Koreans than in European Americans, which is consistent with previous reports.

The crystal structure of 5-HT receptor has been reported (Wang et al., 2013).

Serotonin Transporter

Meta-analysis has shown that the promoter region of the 5-HTT gene (5-HTTLPR; 5-HTT gene-linked polymorphic region) genotype polymorphism affects the activation of the amygdala (Munafò et al., 2008). The 5-HTTLPR polymorphism is a short (S) allele comprised of 14 copies of a 20–23 base pair repeat unit and a long (L) allele comprised of 16 copies. This polymorphism affects the transcription rate of the gene: the S allele is transcriptionally less efficient than the L allele. People with the short (SS and SL, combined) genotype and a European ancestry showed higher amygdala activation than those with the long (LL) genotype when emotional facial expressions, pictures, and words were used as stimuli. The neuroimaging technologies that were used in the studies included in the meta-analysis were fMRI, positron emission tomography, and perfusion MRI. However, this association, i.e., higher activation in the S genotype group than in the LL group, may be opposite (the LL genotype is associated with higher amygdala activation) in people of Asian ancestry. The effects of genes and environment in 5-HTTLPR genotype polymorphism need further examination.

Furthermore, 5-HTTLPR polymorphism alters the neural circuitry regulating emotion. The S allele form is associated with increased negative emotionality and stress vulnerability, which is related to the diminished communication between the amygdala and specific regions of the prefrontal cortex, namely, the sub- and supragenual anterior cingulate cortex. Almost 30% of the variance in harm avoidance is predicted by the functional connectivity of the amygdala and the rostal subgenual anterior cingulate cortex (Pezawas et al., 2005; Hariri and Holmes, 2006). Meta-analysis has shown that 5-HTTLPR moderates the relationship between stress and depression, with the S allele being associated with an increased risk of developing depression under stress (Karg et al., 2011).

As shown earlier, neonatal/infant temperament especially withdrawal, anxiety, and/or fear may be explained by neurotransmitters and their genomic variants in serotonin system. Research has been performed on the role of the serotonin system components in temperament; however, no definitive conclusions have been reached, as some studies show positive results, and other studies show negative results. A meta-analysis showed that a role for variants in the serotonin system could have some value (Saudino and Wang, 2012).

Clinical Significance of Serotonin Transmission

The aim of this review is not to express the pathological condition of temperament, but it is still important to mention this aspect. The 5-HT system is not simple. Different effects of 5-HT are observed in different species (Popova, 2006; discussed in the Serotonin System section), and changes in 5-HT transmission result in anxiety (increased 5-HT) or depression and panic disorder (decreased 5-HT) (Cools et al., 2007). Anxiety is treated with benzodiazepines to reduce 5-HT transmission; depression and panic disorder are treated with selective serotonin reuptake inhibitors to increase 5-HT transmission.

Dopamine System

The association between polymorphism of dopamine system genes and personality traits in healthy Caucasians has been reported (Kazantseva et al., 2011). The transporter gene (SLC6A3) polymorphism has been associated with persistence, and the dopamine D2 receptor gene (DRD2) has been associated with neuroticism. Concerning the neural circuitry in dopamine system, it has been suggested that decreased Dopamine D2 receptor availability in the striatum causes a predisposition to reward-related traits and disorders associated with DRD2 gene polymorphism.

Gene–Environment Interaction and Epigenetics

Gene–Environment Interaction

The role of environment in the development of personality has been clarified. Some authors classified approaches in this area, particularly influences on adolescent depression, as genetic, environmental, and gene–environmental (Eley et al., 2004). Environmental influences may be divided into child-specific and family-general risks. The former includes life events and peer relationship problems; the latter includes parental psychopathology, low socioeconomic status, and family-based stressful life events. Gene–environmental interaction may also be divided into two approaches: an indirect approach, where phenotype heritability varies as a function of environmental risk, and a direct approach, e.g., the interaction between adverse life events and the S allele of the 5-HTTPLR on depression and suicide. 5-HTTPLR allele analysis revealed an association with adolescent depression: for females within the high environmental risk group for depression, an increase in the number of S alleles (SS, SL, or LL) increased the frequency of the depression (Eley et al., 2004). Comparing the high and low environmental risk groups, girls with SS alleles in the former group exhibited a 2 times greater risk of depression than the latter group. This observation may be an example of a direct effect of gene–environmental interaction. If the environmental risk is low, even girls with SS alleles (genetically high risk adolescents) do not get depression with high frequency.

In addition, in rhesus monkeys, an interaction between rh5-HTTLPR (a rhesus monkey ortholog of human 5-HTTLPR) and environment has been reported (Champoux et al., 2002). Mother- and nursery-reared animals with the heterozygous 5-HTTLPR form SL demonstrated an increased affective response relative to LL homozygotes. Nursery-reared SL infants exhibited lower orientation scores than their LL counterparts (Champoux et al., 2002).

Gene–environment interaction is observed in many areas. The relationship between body mass index (BMI) and physical activity is one example. The heritability of BMI is moderated by physical activity: the higher the level of physical activity, the smaller the genetic influence on BMI (van Dongen et al., 2012).

Epigenetic Regulation of the Glucocorticoid Receptor

Meaney and colleagues (Weaver et al., 2004; Meaney and Szyf, 2005; Meaney, 2010) reported the epigenetic regulation of the glucocorticoid receptor (GR) gene. There are two types of mother rats in the Long-Evans outbred strain: mother rats that rear their pups with lots of licking and grooming (LG) and mother rats that do not. The female offspring of each type of mother grow up to be mothers with the same style of child-rearing as their mothers. Increased LG by mother rats in the first week of life alters the DNA structure at a GR gene promoter in the hippocampus of the offspring. A difference in the DNA methylation at a GR gene promoter in the offspring of high- and low-LG mothers emerges over the first week of life: low methylation is observed in the pups of high-LG mothers, and high methylation is observed in the pups of low-LG mothers; cross-fostering reverses the methylation states. In the neonatal pups, increased LG is associated with hippocampal 5-HT turnover and the activation of a 5-HT_7 receptor, which is coupled to the induction of the transcription factor NGFI-A and its cofactor CREB-binding protein. The CREB-binding protein enhances the ability of NGFI-A to bind and activates GR gene transcription. In addition, increased histone acetylation, which is a marker of active chromatin, and greater binding of NGFI-A protein to the hippocampal exon 1_7 GR gene promoter are observed in the adult offspring of high-LG mothers compared with the adult offspring of low-LG mothers. The high-LG offspring show significantly increased hippocampal GR mRNA expression, enhanced glucocorticoid negative-feedback sensitivity, and decreased hypothalamic CRH mRNA levels. As adults, the offspring of the high-LG mothers show increased anxiolytic behavioral responses to stress. McGowan et al., (2009) revealed that the same mechanism worked in humans who were abused in childhood, which suggests a common effect of parental care on the epigenetic regulation of hippocampal GR expression.

An adverse environment during infancy affects at least the serotonin system and glucocorticoid receptor regulation. The influence in the infant period appears even in later life.

Discussion – Roles of Heritability on Personal Traits

Cloninger (1987) presented a hypothesis on the relation of personality dimensions to neurotransmitters, which is a milestone in the biological study of human personality. Primary characters such as vigilance and wariness are highly beneficial and are conserved in animals (Nettle, 2006). These character traits may be related to approach and withdrawal in human neonates, fear and anxiety in infants, and neuroticism, one of the Big Five trait factors in adults. Neonatal traits appear via a biological mechanism. Approach and withdrawal may be regulated by the genes of the serotonin system (e.g., tryptophan hydroxylase gene, serotonin receptor gene, serotonin transporter gene) and by specific brain loci (amygdala and prefrontal cortex). The appearance of a neonatal trait may be controlled by the regulation of neurotransmitter genes and the neural circuitry where the neurotransmitters work. Other traits expressed by Chess and Thomas (1996) may be not essential to life, and elucidating a similar responsible set for the trait (the neurotransmitter and the neural circuitry) may thus be more difficult than for the case of approach and withdrawal.

Tooby and Cosmides (1992) suggested that from an evolutionary perspective, personality variation in human individuals is neutral. Nettle (2006) noted that heritable variation is ubiquitous in animals, including humans, and that each of the Big Five personality factors has its own benefits and costs (trade-offs). An example of animal behavioral variation is the exploitation observed in the great tit (*Parus major*), which is consistent within individuals and is heritable in the same range as human personality dimensions.

To examine the relationship between children's early family experience and their later intellectual growth, the vocabularies of children were studied from 1 to 3 years of age (Hart and Risley, 1995). The results showed that the most important factors for language acquisition are the economic advantages of the child's home and the frequency of the child's language experiences. Children born into homes with fewer economic resources learn fewer words, have fewer experiences with words through interactions with other people, and acquire a vocabulary of words more slowly. Although the genetic participation rate may differ from character and intellectual growth, such as language, this report suggests that intervention in character development may also be possible. If infants have high reactivity, we may intervene in their development to not reinforce that character because extremely high reactivity may not only make them unhappy but it may also be related to morbid status, such as unipolar major depression or bipolar disorder.

Schwartz et al. (2012) reported the role of gender in predicting the effect of infant temperament on amygdala reactivity in adulthood. Male adults who were high-reactive infants showed high levels of reactivity to novel faces in the amygdala. In contrast, females showed less reactivity, which may be explained by the effect of estrogen. The testosterone level is high even in neonatal males and may affect neonatal traits, for example, through dopamine neurotransmission (Kazantseva et al., 2011). Men have markedly greater dopamine release than women in the striatal regions (Munro et al., 2006).

It is generally accepted that approximately 50% of human character is controlled by heritability (e.g., Tellegen et al., 1988; Nolen-Hoeksema et al., 2009). There are several different approaches to assess the role of heritability (Plomin, 2013). The primary approach is through candidate genes, as in the case of the serotonin system in fear and anxiety. The second approach includes whole genome sequence, microarray, and genome-wide association studies. These methods are based on the idea that an accumulation of very small genetic differences (e.g., SNPs) can affect character traits. The third approach includes epistasis (gene–gene interaction), gene–environment interaction, and epigenetics.

A large genome-wide meta-analysis study of association on temperament scales involving more than 11 000 adults was reported (Service et al., 2012). However, in this large study, less than 1% of the variability in any temperament dimension

was due to a risk score derived from SNPs showing association to the temperament dimensions, although the temperament scales are considered heritable.

Each genetic method for assessing the heritability of human character has its limitations. For example, in the candidate gene approach, we often do not have strong hypotheses as to which genes are candidates. A personality trait may be a function of hundreds of different genes (Rutter, 2011). These genes are both specific and redundant and can participate in cross-talk with each other. According to Rothbart (2011), there are interactions (either facilitation or inhibition) among temperament elements in infants. So far, neither method can adequately explain heritability. This situation is observed for other human traits that are considered to be multifactor inheritance (e.g., body height). It is called 'missing heritability' (Maher, 2008; Manolio et al., 2009).

A comparison between an infant's temperament assessed using any of Chess and Thomas's, Kagan's, or Rothbart's classifications and the results of polymorphism in relevant genes, as in the case of approach or withdrawal and the serotonin system, may be very interesting. Analyzing the relationship between temperament/personality and genetic variability in infants is advantageous because infants' temperaments are simpler than adults' personalities and are not yet strongly affected by environment. Assessing crying time in neonates has further merits: the method is simple, not painful, and does not depend on self-reported inventories.

It is not known whether any of three approaches described earlier will elucidate the genetic mechanism of temperament or whether we will have to wait until other methods are developed. However, clarifying this mechanism will increase our understanding of neonates and human personality and will help us raise healthy children in the future.

See also: Big Five Factor Model, Theory and Structure; Child and Adolescent Psychiatry, Principles of; Defiant Behavior During Infancy and Early Childhood; Early Emotional Development and Cultural Variability; Endocrinology and Psychiatry; Evidence-Based Psychiatry; Five Factor Model of Personality, Assessment of; Five Factor Model of Personality, Facets of; Five Factor Model of Personality, Personality Disorder; Five Factor Model of Personality, Universality of; Imaging Methods in Clinical Psychology; Infancy and Human Development; Longitudinal Analyses of Sexual Development through Early Adulthood; Neonatal Bonding; Neonatal Crying Behaviors; Personality Assessment: Overview; Personality Assessment; Personality Changes During Adolescence Across Cultures; Personality Development: Systems Theories; Personality Differences and Development: Genetic and Environmental Contributions; Personality, Biological Models of; Personality, Trait Models of; Qualitative Methodology in Developmental Psychology; Stress in Childhood: Effects on Development; Stress in Infancy and Early Childhood: Effects on Development; Stress: Effects on Neonatal and Child Development; Stress: Psychological Perspectives; Temperament Development, Theories of; Temperament and Human Development; Temperament.

Bibliography

Barber, T.X., 1993. Human Nature of Birds: A Scientific Discovery with Startling Implications. St. Martin's Press, New York.

Berger, S.M., Weber, T., Perreau-Lenz, S., et al., 2012. A functional Tph2 C1473G polymorphism causes an anxiety phenotype via compensatory changes in the serotonergic system. Neuropsychopharmacology 37, 1986–1998.

Bismark, A.W., Moreno, F.A., Stewart, J.L., et al., 2010. Polymorphisms of the HTR1a allele are linked to frontal brain electrical asymmetry. Biological Psychology 83, 153–158.

Brown, S.M., Peet, E., Manuck, S.B., et al., 2005. A regulatory variant of the human tryptophan hydroxylase-2 gene biases amygdala reactivity. Molecular Psychiatry 10, 884–888.

Canli, T., Lesch, K.P., 2007. Long story short: the serotonin transporter in emotion regulation and social cognition. Nature Neuroscience 10, 1103–1109.

Champoux, M., Bennett, A., Shannon, C., et al., 2002. Serotonin transporter gene polymorphism, differential early rearing, and behavior in rhesus monkey neonates. Molecular Psychiatry 7, 1058–1063.

Chess, S., Thomas, A., 1996. Temperament: Theory and Practice. Brunner/Mazel, Inc., New York.

Cloninger, C.R., 1987. A systematic method for clinical description and classification of personality variants. A proposal. Archives of General Psychiatry 44, 573–588.

Cloninger, C.R., 1998. A psychobiological model of temperament and character. In: Richter, J., Eisemann, M., Bollow, K., Schläfke, D. (Eds.), The Development of Psychiatry and Its Complexity. Waxmann, Münster, pp. 1–16.

Cools, R., Roberts, A.C., Robbins, T.W., 2007. Serotoninergic regulation of emotional and behavioural control processes. Trends in Cognitive Sciences 12, 31–40.

Eley, T.C., Sugden, K., Corsico, A., et al., 2004. Gene–environment interaction analysis of serotonin system markers with adolescent depression. Molecular Psychiatry 9, 908–915.

Fernandez, S.P., Gaspar, P., 2012. Investigating anxiety and depressive-like phenotypes in genetic mouse models of serotonin depletion. Neuropharmacology 62, 144–154.

Hariri, A.R., 2009. The neurobiology of individual differences in complex behavioral traits. Annual Review of Neuroscience 32, 225–247.

Hariri, A.R., Holmes, A., 2006. Genetics of emotional regulation: the role of the serotonin transporter in neural function. Trends in Cognitive Sciences 10, 182–191.

Hart, B., Risley, T.R., 1995. Meaningful Differences in the Everyday Experience of Young American Children. Paul H Brookes Pub Co., Baltimore.

Kagan, J., 2010. The Temperamental Thread: How Genes, Culture, Time, and Luck Make Us Who We Are. Dana Press, New York.

Karg, K., Burmeister, M., Shedden, K., Sen, S., 2011. The serotonin transporter promoter variant (5-HTTLPR), stress, and depression meta-analysis revisited: evidence of genetic moderation. Archives of General Psychiatry 68, 444–454.

Kazantseva, A., Gaysina, D., Malykh, S., Khusnutdinova, E., 2011. The role of dopamine transporter (SLC6A3) and dopamine D2 receptor/ankyrin repeat and kinase domain containing 1 (DRD2/ANKK1) gene polymorphisms in personality traits. Progress in Neuro-psychopharmacology & Biological Psychiatry 35, 1033–1040.

Kim, H.S., Sherman, D.K., Taylor, S.E., et al., 2010. Culture, serotonin receptor polymorphism and locus of attention. Social Cognitive and Affective Neuroscience 5, 212–218.

Kishi, T., Okochi, T., Tsunoka, T., et al., 2011. Serotonin 1A receptor gene, schizophrenia and bipolar disorder: an association study and meta-analysis. Psychiatry Research 185, 20–26.

Le François, B., Czesak, M., Steubl, D., Albert, P.R., 2008. Transcriptional regulation at a HTR1A polymorphism associated with mental illness. Neuropharmacology 55, 977–985.

Lemonde, S., Turecki, G., Bakish, D., et al., 2003. Impaired repression at a 5-hydroxytryptamine 1A receptor gene polymorphism associated with major depression and suicide. Journal of Neuroscience 23, 8788–8799.

Li, Q., Luo, T., Jiang, X., Wang, J., 2012. Anxiolytic effects of 5-HT_1A receptors and anxiogenic effects of 5-HT_2C receptors in the amygdala of mice. Neuropharmacology 62, 474–484.

Maher, B., 2008. The case of the missing heritability. Nature 456, 18–21.

Manolio, T.A., Collins, F.S., Cox, N.J., et al., 2009. Finding the missing heritability of complex diseases. Nature 461, 747–753.

McDevitt, R.A., Neumaier, J.F., 2011. Regulation of dorsal raphe nucleus function by serotonin autoreceptors: a behavioral perspective. Journal of Chemical Neuroanatomy 41, 234–246.

McGowan, P.O., Sasaki, A., D'Alessio, A.C., et al., 2009. Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse. Nature Neuroscience 12, 342–348.

Meaney, M.J., 2010. Epigenetics and the biological definition of gene × environment interactions. Child Development 81, 41–79.

Meaney, M.J., Szyf, M., 2005. Environmental programming of stress responses through DNA methylation: life at the interface between a dynamic environment and a fixed genome. Dialogues in Cinical Neuroscience 7, 103–123.

Munafò, M.R., Brown, S.M., Hariri, A.R., 2008. Serotonin transporter (5-HTTLPR) genotype and amygdala activation: a meta-analysis. Biological Psychiatry 63, 852–857.

Munro, C.A., McCaul, M.E., Wong, D.F., et al., 2006. Sex differences in striatal dopamine release in healthy adults. Biological Psychiatry 59, 966–974.

Nettle, D., 2006. The evolution of personality variation in humans and other animals. American Psychologist 61, 622–631.

Nolen-Hoeksema, S., Fredrickson, B.L., Loftus, G.R., Wagenaar, W.A., 2009. Atkinson & Hilgard's Introduction to Psychology, fifteenth ed. Cengage Learning EMEA, Hampshire, UK.

Pezawas, L., Meyer-Lindenberg, A., Drabant, E.M., et al., 2005. 5-HTTLPR polymorphism impacts human cingulate–amygdala interactions: a genetic susceptibility mechanism for depression. Nature Neuroscience 8, 828–834.

Plomin, R., 2013. Child development and molecular genetics: 14 years later. Child Development. 84, 104–120.

Popova, N.K., 2006. From genes to aggressive behavior: the role of serotonergic system. Bioessays 28, 495–503.

Roberts, B.W., DelVecchio, W.F., 2000. The rank-order consistency of personality traits from childhood to old age: a quantitative review of longitudinal studies. Psychological Bulletin 126, 3–25.

Robinson, M., Oddy, W.H., Whitehouse, A.J., et al., 2013. Hypertensive diseases of pregnancy predict parent-reported difficult temperament in infancy. Journal of Developmental and Behavioral Pediatrics 34, 174–180.

Rothbart, M.K., 2011. Becoming Who We Are. Guilford Press, New York.

Rutter, M., 2011. Gene–environment interplay: scientific issues and challenges. In: Dodge, K.A., Rutter, M. (Eds.), Gene–Environment Interactions in Developmental Psychopathology. Guilford Press, New York, pp. 3–17.

Saudino, K.J., Wang, M., 2012. Quantitative and molecular genetic studies of temperament. In: Zentner, M., Shiner, R.L. (Eds.), Handbook of Temperament. Guilford, New York, pp. 315–346.

Schwartz, C.E., Kunwar, P.S., Greve, D.N., et al., 2012. A phenotype of early infancy predicts reactivity of the amygdala in male adults. Molecular Psychiatry 17, 1042–1050.

Service, S.K., Verweij, K.J.H., Lahti, J., et al., 2012. A genome-wide meta-analysis of association studies of Cloninger's Temperament Scales. Translational Psychiatry 2, e116.

Tellegen, A., Lykken, D.T., Bouchard Jr., T.J., et al., 1988. Personality similarity in twins reared apart and together. Journal of Personality and Social Psychology 54, 1031–1039.

Tooby, J., Cosmides, L., 1992. The psychological foundations of culture. In: Tooby, J. (Ed.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, New York, pp. 19–136.

Tsuchiya, H., 2011. Emergence of temperament in the neonate: neonates who cry longer during their first bath still cry longer at their next bathings. Infant Behavior & Development 34, 627–631.

Tsuchiya, H., 2013. Personality in neonates. In: Morris, E.F., Jackson, M.-A. (Eds.), Psychology of Personality. Nova Science Publishers, Inc., New York, pp. 23–40.

van Dongen, J., Slagboom, P.E., Draisma, H.H.M., Martin, N.G., Boomsma, D.I., 2012. The continuing value of twin studies in the omics era. Nature Reviews Genetics 13, 640–653.

Wang, C., Jiang, Y., Ma, J., et al., 2013. Structural basis for molecular recognition at serotonin receptors. Science 340, 610–614.

Weaver, I.C., Cervoni, N., Champagne, F.A., et al., 2004. Epigenetic programming by maternal behavior. Nature Neuroscience. 7, 847–854.

Neo-Piagetian Theories of Cognitive Development

Michael F Mascolo, Merrimack College, North Andover, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Neo-Piagetian theories of cognitive development emerged as attempts to preserve core theoretical and empirically supported aspects of Jean Piaget's seminal theory of intellectual development while addressing criticisms leveled against the theory. Neo-Piagetian theories preserve three basic ideas from Piaget's theory: (1) the unit of cognitive analysis is the *scheme* or *psychological structure*; (2) psychological structures undergo *qualitative transformation* over time; and (3) higher order structures develop through the *differentiation* and *coordination* of lower level structures. After a brief discussion of similarities and differences among prominent neo-Piagetian theories, one representative approach (dynamic skill theory) is discussed in depth. The discussion concludes with a description of recent advances in neo-Piagetian systems theory.

In his theory of intellectual development, Jean Piaget articulated a series of questions and ideas that launched the field of cognitive development. Piaget's legacy is most keenly reflected in neo-Piagetian theories of development. Neo-Piagetian models emerged in the 1970s and 1980s address criticisms leveled against Piaget's theory while simultaneously maintaining core principles (Case, 1985; Commons et al., 1998; Demetriou et al., 2010; Fischer and Bidell, 2006; Halford, 1993; Pascual-Leone, 1970).

Piaget's Theory of Intellectual Development

According to Piaget, knowing is a form of acting (de Ribaupierre, Piaget's Theory of Cognitive Development). The basic unit of cognitive analysis is the *scheme*, which refers to a structured system of action or thought. *Schemes* function through the dual processes of *assimilation* and *accommodation*. *Assimilation* refers to the process by which to-be-known objects are incorporated into existing schemes; *accommodation* reflects the simultaneous processes of modifying a scheme to incorporate the object. For example, to grasp a rattle, an infant incorporates (assimilates) the rattle into her organized grasping action (scheme) while simultaneously modifying (accommodating) her scheme around the particular contours of the incorporated object.

Piaget maintained that the structure of a child's schemes undergoes transformation in a series of broad-based *stages* over time. For Piaget, stages exhibit several important properties. First, each stage reflects a qualitatively distinct type of thought or action. Children are not simply little adults; they think and act in fundamentally different ways. Second, stages form a *hierarchical progression* with later stages building upon earlier ones. Third, the stages form a *universal* and *unidirectional* sequence regardless of culture or the content of thought. Fourth, Piagetian stages form *structures d'ensemble* (i.e., 'structures of the whole') – generic structures that have wide application to broad ranges of cognitive tasks and conceptual domains. Although Piaget acknowledged the phenomenon of *décalage* – the finding that structurally similar abilities often develop at different times, he nonetheless held that within a stage, systems of thought form broad, integrated wholes.

Piaget suggested that in infancy, schemes operate within the *sensorimotor stage* of development. Sensorimotor schemes consist of organized systems of *action* on objects. During this time, infants cannot form representations (images) of events without direct sensory contact. Symbolic thought emerges between 18 and 24 months of age with the onset of the *preoperational stage*. At this stage, children form representations of events (e.g., words, images) but cannot manipulate them in logical or systematic ways. Children gain the capacity for pretend play and deferred imitation. During the concrete operational stage (around 6–7 years), with the capacity to think logically using concrete images and representations, children can successfully perform a variety of logical tasks (conservation, class inclusion, seriation, transitivity, etc.). In the formal operational stage (adolescence onward), individuals gain the capacity to think logically using abstract ideas free from particular concrete content.

Although Piaget is most closely associated with his stage theory, the concept of *equilibration* provides the foundation of his theory. Equilibration refers to an inherent, self-regulating, compensatory process that motivates the construction of novel patterns of thought over time. *Equilibrium* occurs when an individual is able to assimilate an object to a scheme while accommodating the scheme to the object. *Disequilibrium* occurs when assimilation and accommodation fail; a mismatch arises between object and scheme. Disequilibrium prompts reorganization (differentiation and integration) of schemes in an attempt to restore equilibrium. For example, recognizing an object as a *bunny* involves assimilating a rabbitlike object into an existing *bunny* scheme. Although initial attempts to assimilate a *kitty* into the *bunny* scheme may be successful, eventually, such attempts will fail as children sense the conflict between the kitty's features and their representation of a bunny. One way to restore equilibrium involves differentiating the initial *bunny* scheme into two separate schemes – one for bunnies and one for kitties. Developmental change occurs through *equilibration*.

Critiques of Piaget's Theory

Table 1 describes eight critical issues that have emerged with regard to Piaget's theory. Regarding Piaget's stage theory, Demetriou (2006) stated, "whenever a replication study

Table 1 Four neo-Piagetian theories

Issue	*Piagetian principle*	*Problem*	*Pascual-Leone*	*Case*	*Fischer*	*Demetriou*
Unit of psychological analysis	Figurative and operative schemes basis in action and internalized action	Difficulty specifying scheme independent of task; task ability changes with task	First-order subjective operators (schemes) and second-order *metasubjective* operators	Executive control structures	Skill: Capacity to control elements of acting, skill and feeling within particular contexts, and psychological domains	Processing potentials Hypercognitive system Specialized structures of thought
Nature of change	Qualitative structural change	Evidence of higher order abilities at lower stages	Qualitative changes in structure; quantitative changes in working memory capacity	Qualitative changes in structure; quantitative changes in working memory capacity	Qualitative and quantitative transformation of skills in context	Structural change in specialized structures of thought; quantitative change in processing potentials
Structures D'Ensemble	Broad-based stages of structures of logicomathematical Intelligence	Decalage between emerging abilities; age of emergence varies with area, task, context, etc.	Task analysis yields structure of actions needed to perform a given task	Development organized around series of different central conceptual structures	Development tied to local domains, tasks, and contexts; variation is rule rather exception; task analysis indicates structure	Specialized structures of thought (e.g., class, causality, number, space, verbal, social domains of thought) each mediated by their own symbol systems
Hierarchical progression	Logical skills involve hierarchical progression	Predicted sequences do not always hold	Tied to task and nature of task sequence	Hierarchical progression within development central conceptual structures	Higher order skills build on lower level skills; multiple paths to any particular form of skill	Hierarchical progression within development central conceptual structures
Universality	Single universal sequence of stages	Variation in particular trajectories by culture; final stages not always reached	Tied to task and nature of task sequence	Different central conceptual structures develop through fixed and universal stage structures	Alternative pathways of structural change in different domains, tasks, contexts, cultures, and individuals	Different central conceptual structures develop through fixed and universal stage structures
Change processes	Equilibration Differentiation and integration to adapt to cognitive conflict; *maturation* of physical structure; active *experiencing* of world; *social transmission* of cultural knowledge	Difficulty specifying how process of differentiation arises from cognitive conflict	*Constructive operators* foster change: *Content operator* (*C*) produces accommodative change as schemes encounter conflict; *learning operator* (*L*) coordinates schemes into higher order units; *mental energy* (*M*) indicates central computing space; *interrupt operator* (*I*) inhibits irrelevant schemas	Hierarchical coordination of executive control structures; changes in working memory engendered by brain development and myelinization	Differentiation and coordination of lower level actions into higher order skills; social scaffolding and contextual support; discontinuities in brain growth and development	Change in speed of processing, control of processing, and representational capacity; metarepresentation and symbolic individuation
Role of affect	Affect merely provides energetics of adaptive action; schemes do the work of adaptive action	Affect more differentiated and plays stronger role than Piaget suggested	*Affect operator* (*A*): Affective schemes activate cognitive schemes, and vice-versa.	Cognition and affect are generated by different systems that influence each other over time	Affect plays central role in organizing consciousness, psychological structures, and the pathways through which they develop	Self-oriented processes generate emotional feelings
Role of context and culture	Sensory and social conflict stimulates development	Change context, change skill; scaffolding	Change arises through dialectical exchange between organism and physiosocial context	Children must have access to cultural activities to construct central conceptual structures	Context and cultural tools play direct role in formation and functioning of skill	Quantitative differences in processing potentials and content among cultures

remained close to the original Piagetian conditions, the same phenomena were observed in more or less the same way at more or less the same age." However, when researchers modified the nature or complexity of Piagetian tasks, children showed evidence of logical abilities earlier than Piaget indicated (Bryant and Trabasso, 1971). Research shows that children often pass structurally similar tasks at different ages (e.g., mass is conserved before volume; Gelman and Baillargeon, 1983). Studies show that children function at higher levels when provided with contextual support than in the absence of support (Fischer et al., 1993; Vygotsky, 1978). Cross-cultural research indicates variation not only in the age at which children are able to perform Piagetian tasks, but also in the form of thinking itself (Saxe, 1991). Other theorists suggested that Piaget's concepts of assimilation and accommodation were vague and difficult to assess (Klahr and Wallace, 1976). Still others suggested that the theory failed to account for the importance of affect (Fischer et al., 1990) and language (Matusov and Hayes, 2000) in the process of development.

Neo-Piagetian Models of Development

Neo-Piagetian models of development preserve three core aspects of Piaget's theory: (1) knowing involves the invocation of schemes or psychological structures; (2) that psychological structures undergo *qualitative transformation* in development; and that (3) higher order forms of psychological structure arise through the *differentiation and hierarchical coordination* of lower level schemes. Table 1 and Figure 1 show similarities and differences among prominent neo-Piagetian approaches.

Pascual-Leone's Theory of Constructive Operators

Juan Pascual-Leone (1970) launched the traditional of neo-Piagetian theories. Drawing on information processing psychology, Pascual-Leone proposed his *theory of constructive operators* (*see* Conscious Control during Childhood, Development of; Executive Functions during Childhood, Development of). Pascual-Leone differentiated two basic categories of operations. These include *subjective operators* (akin to Piagetian schemes) and *metasubjective operators*. In any given situation, a variety of first-order operative and figurative (content) schemes are activated. In the traditional conservation task, activate schemes for understanding height and width of the beakers, the fluidity of liquid, and so forth. The multiplicity of schemes sets the stage for the production of cognitive conflict. For example, a child may notice that when pouring water from a tall and thin beaker to a short and wide one, the amount of water appears to change. Metasubjective operators are second-order processes that operate on first-order operators (schemes). Pascual-Leone differentiated a series of metasubjective operators, the most prominent including the *Content* (*C*), *Learning* (*L*), *Mental Energy* (*M*), and *Interrupt* (*I*) operators. (Other

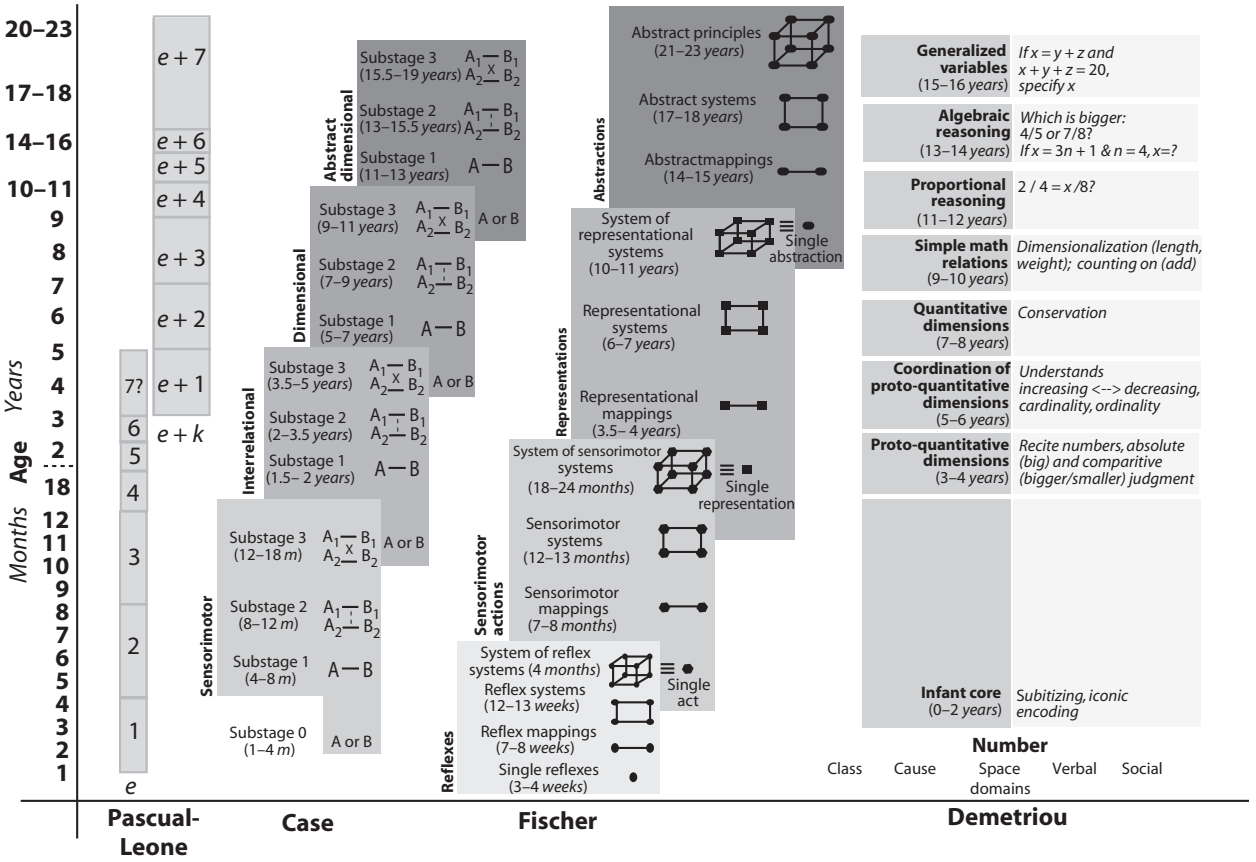

Figure 1 Comparison of levels hypothesized by Neo-Piagetian theories of development.

operators include the Content (*C*), Space (*S*), Field (*F*), and Affect (*A*) operators.) Metasubjective operators function to create new schemes in order to resolve contradictions among competing schemes. For example, the *C* operator adjusts the content of schemes to accommodate to novel experience. The *L* operator creates higher order schemes by coordinating two or more activated schemes (e.g., coordinating schemes for judging *height* and *width* into an understanding that changes in height are compensated by changes in width). The *I* operator *inhibits* the activation of schemes that are irrelevant to a given task (e.g., inhibiting the tendency to rely on perceptual differences in height to make judgments about volume in the conservation task). These operators work together adaptively in the context of cognitive conflict to construct novel schemes that function to resolve cognitive contradiction.

Perhaps the most frequently invoked construct proposed by Pascual-Leone is the concept of mental energy or *M space*. *M* places a limit on the amount of information that can be held in mind at any given time. Pascual-Leone proposed that maturation of the capacity of the *M* operator is the primary source of cognitive developmental change (*see* Short- and Long-Term Memory during Infancy and Early Childhood across Cultures). In development, increases in M capacity allow children to hold increasing numbers of schemes in mind while performing cognitive tasks. According to Pascual-Leone and Johnson (1991), between birth and 3 years of age, sensorimotor M space (e) increases by one scheme every 4 months. For symbolic schemes, the capacity of the M operator (k) increases by one scheme every 2 years from 3- to 15-years of age. Thus, as indicated in **Figure 1**, total *M* space is calculated as the sum of *M* space required for sensorimotor (e) and symbolic (k) schemes ($M = e + k$).

To understand developmental changes in children's task performance, it is necessary to understand the nature and number of schemes that children must hold in mind in order to perform different types of tasks. To achieve this goal, a researcher can perform a task analysis, which involves identifying the component operations and representations required to complete a task. Having performed a task analysis, a researcher can identify the number of schemes necessary to perform the task successfully. Developmental change does not occur in a holistic, stagelike matter. At any given point in time, a child's cognitive abilities are defined by a child's available schemes, the capacity of *M* space, and the cognitive demands of particular tasks.

Case's Staircase Model

Case's (1985, 1991) model of cognitive development builds upon and extends Pascual-Leone's approach. Case's model contains three main components: The theory of *executive control structures*, the invocating of *working memory capacity* as the source of conceptual change, and developmental change organized around *central conceptual structures*. For Case, the primary unit of thinking is the executive control structure, which consists of "an internal mental blueprint, which represents a subject's habitual way of construing a particular problem situation, together with his or her habitual procedure for dealing with it" (Case, 1985: p. 68.). The concept of executive control structure incorporates both the figurative and operative aspects of thinking into a single structure. Similar to Piaget, Case proposed that executive control structures develop through four basic stages. As indicated in **Figure 1**, these include the *sensorimotor* (birth to 18 months), *interrelational* (1½–5 years), *dimensional* (5–11 years), and *vectoral* (or abstract dimensional) stages (11 years through adulthood). Within each stage, control structures develop through four levels: *predimensional*, *unidimensional*, *bidimensional*, and *integrated bidimenional*. The substages develop in a cyclic and iterative way in that the *last* substage within each stage (the integrated bidimensional substage) corresponds to the *first* substage (*predimensional*) of the next broad stage of thought. (The cyclic nature and development of executive control structures will be explained further in the discussion of Fischer's (1980) dynamic skill theory.) Case proposed that developmental progressions are organized around different *central conceptual structures* (Case et al., 1996). These consist of "networks of semantic nodes and relations that have an extremely broad (but not system-wide) domain of application" (Case et al., 1996: p. 5). Central conceptual structures organize a suite of executive control structures around a common core. Case and his colleagues amassed evidence for the existence of a series of central conceptual structures, including those for *quantity*, *space*, *social relations*, *narrative*, *motor action*, and *music*.

Like Pascual-Leone, Case held that changes in attentional capacity are responsible for stage progression. Case suggests that processing capacity involves the integration of two basic components: *operating space* and *short-term storage space*. Operating space refers to the capacity required to execute *goal-related operations*. Short-term memory space corresponds to the number of *representational schemes* on which an individual can focus in a single act of centration. For example, the act of counting the number of items on a grocery list is composed of two integrated components: the goal-related *operations* corresponding to counting each item, and the monitoring of representations that arise from successive counts ('1 item,' '2 items,' etc.). Counting acts require *operational* space whereas the monitoring the representational outcomes requires *short-term storage space*. Case held that *total* processing capacity (i.e., operating space plus short-term storage space) does *not* change in ontogenesis. Instead, he suggested that the ability to execute operations becomes increasingly efficient over time, which frees processing space for short-term storage. Case suggested that increasing efficiency of executive operations results from changes in brain structures that mediate the processing of information, as well as increasing myelinization of neuronal axons.

Halford's Relational Complexity Model

Halford's (1993; Halford et al., 2012) relational complexity model of cognitive development is based on the idea capacity limitations in working memory are not simply a function of the number of representational units that can be retained at any given time, but instead reflect the complexity of the *relations between* representational units. Halford's relational complexity metric is defined in terms of *arguments* and *relations*, and can be represented as follows: RELATION (argument$_i$, argument$_j$, etc.). Binary relations include relations between two arguments;

ternary relations involve three arguments; quaternary relations involve four interrelated arguments, and so forth. For example, the relation LARGER-THAN (dog, cat) expresses the binary relation 'dog is larger than cat'; the ternary relation ADDITION (1, 3, and 5) indicates an additive relation between three numbers. Halford and his colleagues have assessed children's capacity to perform different types of representational tasks (e.g., transitivity, class inclusion, hierarchical classification, cardinality, etc.) that vary in relational complexity. They found that relational complexity was directly related to the age of acquisition of target skills (Andrews and Halford, 2002). Halford has shown that unlike performance on Piagetian tasks, (which show decalage in different conceptual domains) children acquire the capacity to perform tasks at the same level of relational complexity at the same time, regardless of domain or task variables (Halford and Andrews, 2006). Halford's relational complexity theory is not a stage theory in the sense of Piaget's theory. Instead, like Pascual-Leone's theory, Halford is able to track gradual developmental changes in children's thinking by identifying the relational complexity of the tasks that they are able to perform. However, like Case's (1985) theory, Halford's theory predicts that tasks reflecting the same level of relational complexity will be passed around the same age.

Demetriou's Analysis of the Developing Architecture of Mind

For over 25 years, Andreas Demetriou et al. (2010) have worked toward a comprehensive model of the architecture of the developing mind. As indicated in **Figure 2**, Demetroiu suggests that the human 'mind' is organized into three levels of interacting systems: *Specific structural systems*, *hypercognitive* systems, and *core capacities*. Specific structural systems consist of organized domains of representation and operational abilities and are constituted by different symbol systems. These include *categorical, verbal, quantitative, causal, spatial,* and *social* systems of operations. Hypercognitive systems consist of systems of higher order control structures (sometimes called meta-cognitive structures) that govern self-understanding, self-monitoring, and self-regulation. Hypercognitive processes map elements of specific structural systems onto each other in the constructive integration of everyday activity. Core capacities include *processing speed*, *control of processing,* and *storage*. Processing speed refers to the tie needed for the system to process the meaning of information; control functions regulate the activation and inhibition of relevant and irrelevant information in any given task. Storage is mediated by a multidimensional system for retaining and operating on information in real time Baddeley's (2012). Information is retained in specialized information processing buffers (e.g., phonological, visual) where they are operated on by central executive processes.

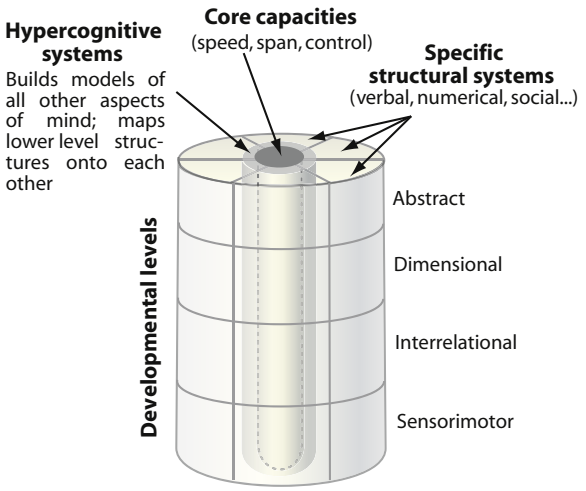

Figure 2 Demetriou's model of cognitive development.

Demetriou suggests that each of these three systems undergoes systematic change in development. Specific domains of thought (structural systems) undergo qualitative change in accordance with Case's (1985) model. To illustrate, developmental changes in the quantitative structural system are indicated in **Figure 1**. Hypercognitive processes also change. For example, children under the age of about 7–8 years show limited awareness of the nature and functioning of their own psychological processes. After about 7–8 years, children use their awareness increasingly of their psychological states to predict, plan, and regulate their thinking and acting. Finally, core capacities show steady change over time. Demetriou and his colleagues (Demetriou and Mouyi, 2011) have shown that core capacities (processing speed, control processing, and information storage) increase systematically with age. While such provide the broad resources for the construction of higher order thought structures, the processes of *metarepresentation* and *symbolic individuation* mediate their actual construction. Metarepresentation processes identify and abstract similarities between mental experiences (integration); symbolic individuation processes tag novel mental experiences (differentiation).

Dynamic Skill Theory: A Neo-Piagetian System Model Psychological Development

The discussion now turns to an in-depth illustration how neo-Piagetian theory – in particular, Fischer's dynamic skill theory (Fischer, 1980; Fischer and Bidell, 2006; Mascolo and Fischer, 2015) – accounts for structural changes in psychological development. In Fischer's (1980) approach, a *skill* refers to an individual's capacity to control elements of thinking, feeling, and acting within specified contexts and within particular task domains. As in Case's (1991) theory, a skill is a type of *control structure*; it refers to the configuration of action over which an individual exerts control within a given context. The concept of skill differs from the Piagetian notion of scheme. Unlike the concept of scheme, skills are not properties of individuals. Instead, they are properties of *individuals-within-contexts*. Changes in context typically result in change in behavior, and vice-versa. Action is always a joint product of person and context.

Contexts differ in how they contribute to individual activity. Skill theory differentiates three types of contexts*: Low support, high support,* and *scaffolding*. By orienting attention to important features of a task, high support contexts support the production of a person's *optimal level* of functioning – the *highest* level one

can perform on one's own. In low support contexts, individuals operate at their everyday *functional* levels, typically several levels lower than their optimum. For example, students who are able to understand complex ideas while attending a lecture (high support) often find that their understanding declines when studying at home (low support). This occurs because at home, students do not benefit from the support of a lecture to orient their attention toward central ideas.

A third type of context provides individuals with *scaffolding*. Scaffolding occurs when a more accomplished individuals break down, structure, direct and assist a less expert person in completing a task (Mascolo, 2005). Scaffolding differs from high support. Under conditions of scaffolding, another person carries part of the load of a task; in high-support situations, individuals complete tasks *on their own*, albeit under conditions that orient attention. Under scaffolded conditions, individuals perform at levels that exceed their optimal levels, thus fostering learning and development (Vygotsky, 1978) (*see* Vygotsky's Theory of Cognitive Development; Neo-Vygotskian Developmental Theory).

A second way in which the concept of skill departs from Piagetian theory is that skills are not general structures. There are no general, decontextualized, or all-purpose skills. Skills that operate within different conceptual domains (e.g., conservation, classification, reading words, social interaction, etc.) develop somewhat independently and at different rates. Development in domain (e.g., conservation) does not necessarily predict development in other domains (e.g., classification), or even in conceptually similar tasks (conservation of number vs conservation of volume). It is only possible to track developmental sequences for particular skills within particular domain, social contexts, and assessment conditions.

The Web of Development

Given the above, it follows that individuals never operate at any single level of development. Instead, they operate within a *developmental range* – a series of levels that vary with task, domain, context, emotional state, and so forth. Given such dynamic variation, there can be no broad-based stages of development. It is thus not helpful to think of a person or a person's abilities as being 'in a stage' of development. Development does not move through a series of fixed steps; development operates more like a *constructive Web*. As indicated in Figure 3, a constructive Web is composed of multiple converging and diverging pathways or strands. Such strands can represent the developmental courses of different individuals, different domains of thought in groups of children, or different domains of skill within the same individual (e.g., writing, arithmetic, social skills). Within particular individuals, skills in different domains develop along different pathways. For example, writing and arithmetic skills tend to develop along relatively independent pathways. However, skills for responding in writing to math problems require convergence among different reading, writing, and arithmetic skills.

Tracking the Development of Psychological Structures

Skills develop through the hierarchical coordination of lower level action systems into higher order structures. Figure 1 presents the levels of hierarchical organization of a developing skill based on Fischer's dynamic skill theory (Fischer, 1980; Fischer and Bidell, 2006). Skills develop iteratively through four broad tiers: *Reflexes* consist of innate action elements (e.g., sucking a nipple placed in the mouth); *sensorimotor actions* refer to acquired actions directed toward objects (e.g., reaching for a seen ball). *Representations* consist of symbolic meanings about concrete aspects of the world (e.g., 'Eating candy is fun'); *abstractions* consist of higher order representations of intangible and generalized aspects of the world (e.g., "Conservation is the idea that the quality of something remains the same despite a change in appearance"). Within each broad tier, skills develop through four levels: A single *set* refers to a single organized reflex, action, representation, or abstraction. *Mappings* refer to coordinations between two or more sets, whereas *systems* consist of coordinations of two or more mappings. A *system of systems* arises from the intercoordination of at least two systems and is the equivalent of the first level of the next broad tier of skills. For example, a *system of sensorimotor systems* creates a single *representational set*. Dynamic skill theory identifies four broad qualitatively distinct tiers comprising a total of 13 specific levels.

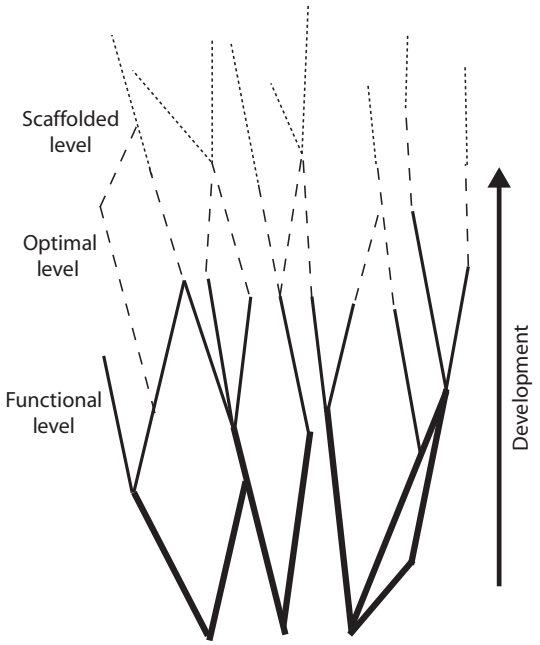

Figure 3 The web of development.

Development in Infancy

Development in infancy spans two broad developmental tiers – reflexes and sensorimotor actions (*see* Motor Skills: Development in Infancy and Early Childhood). Reaching for objects is an example of a sensorimotor action. After a series of levels of prereaching develop during the reflex tier, beginning at around 15–17 weeks of age, infants enter the sensorimotor tier of development. As a result, they gain the capacity to construct single *sensorimotor actions*. Sensorimotor actions consist of controlled actions on objects in the environment. Using sensorimotor actions, in high support contexts, an infant can produce the first successful reaches

toward objects. When posture is supported, an infant can direct arm movements toward a seen object, grasp a toy placed near her hands, or move a grasped toy toward her mouth. Such single sensorimotor acts are represented as follows (see below figure).

Beginning around 7–8 months of age, infants gain the capacity to bring together two sensorimotor acts into a sensorimotor mapping. At this level, an infant is able to fully coordinate distinct acts of looking and reaching into a seamless action. In so doing, an infant can *reach* for an object to *grasp* it, or reach for an object in order to look at it (see below figure).

At this level, infants can reroute their reaches around barriers placed between themselves and target objects. If an adult tries to block a child's reach, he can redirect the reach around the obstacle. Beginning around 11–12 months of age, infants are able to coordinate two sensorimotor mappings into a *sensorimotor system* (S3). In so doing, infants can coordinate multiple acts of looking and reaching in order to explore an object from a variety of angles. Such coordinations are the equivalent of Piaget's concept of tertiary circular reaction (see below figure).

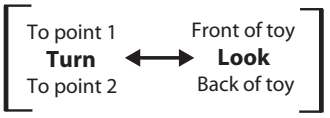

Development through the Representational Tier

Beginning around 18–24 months of age, children gain the capacity to coordinate two sensorimotor systems into a *system of sensorimotor systems*, which is the equivalent of the first level of the next tier of development – *single representational sets*. Using single representations, children form symbolic meanings in the absence of their referents. For example, using single representations, a child can pretend that a teddy bear can walk. In so doing, a toddler can use one sensorimotor system of actions (e.g., holding and moving the teddy bear) to regulate a second such system (e.g., making the bear move forward) in order to represent the symbolic act of walking (see below figure).

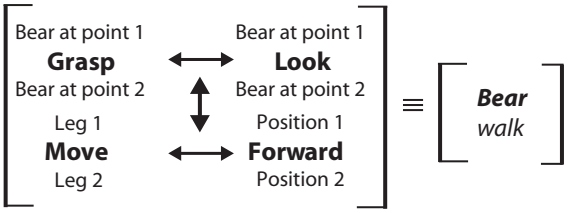

Single representations take different forms in many different modalities (e.g., pretend play, imagery, word use). Single representations are the rough equivalents of a single declarative sentence. Beginning around 3 ½ to 4 years of age, children gain the capacity to coordinate two single representations into a *representational mapping*. At this level, children begin to represent different relations between ideas (e.g., cause/effect; before/after; bigger/smaller; part/whole; reciprocity, and so forth). For example, children can compare the *relative height* of liquid poured into two glasses or the relative width of liquid poured into two glasses, but cannot keep both in mind simultaneously. Given a tall thin glass and a short wide glass filled with the same amount of liquid, a child can correctly indicate that the liquid in one glass is higher or wider than the other (see below figure).

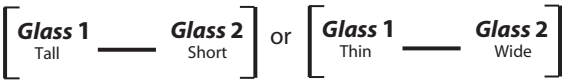

By 6–7 years of age, children can begin to coordinate two representational mappings into a seamless *representational system*. At this level, a child can hold in mind a relationship between two relationships. In the traditional Piagetian conservation task, children are able to relate one representational mapping for comparing changes in height from tall to short to another mapping connecting changes from in width from thin to wide (see below figure).

Using this structure, children solve the conservation task by understanding that the amount of liquid remains the same in both glasses because changes in the height of the liquid in the glasses are compensated for by changes in width.

Developing through the Abstract Tier

Throughout the representational tier, a child's thinking remains concrete. Beginning around 10–11 years of age, in high support contexts, preteens gain the capacity to abstract what is common across multiple representational systems and coordinate them into a single abstraction. Using a single abstraction, a youth can represent intangible aspects of a single idea free from concrete content. For example, a teen can generalize across commonalities between conservation of mass and conservation of volume tasks to construct an abstract (content free) understanding of *conservation*: Despite a change in appearance, the quality of a phenomenon remains the same (see below figure).

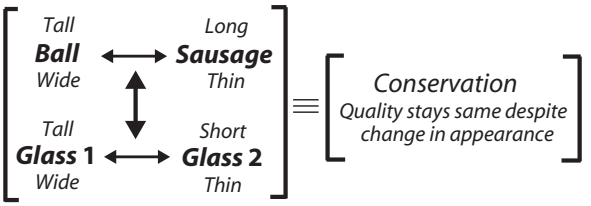

Abstractions develop through the same iterative process that occurred in previous tiers (*see* Self-Regulation during

Adolescence: Variations Associated with Individual ↔ Context Relations*). Around 14–15 years of age, using *abstract mappings*, a teen can hold in mind the relationship between two abstractions. For example, a teen can represent the concept of conservation in terms of the inverse relationship between the loss of *kinetic energy* is lost (e.g., the slowing of the velocity of a ball thrown into the air) and the gain in *potential energy* (gravitational potential) (see below figure).

Using abstract systems, 17- to 18-year-olds begin to coordinate two abstract mappings into an abstract system. At this point, for example, a youth can represent relationships between two abstract aspects of one concept and two abstract aspects of another. One can, for example, represent the concept that *matter can neither be created nor destroyed* in terms of Einstein's principle relating the abstract concepts of *energy* and *matter* ($e = mc^2$) (see below figure).

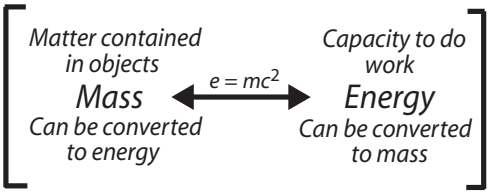

The final level of skills emerges in early adulthood with the capacity to coordinate two or more abstract systems into *abstract principles* – highly abstract representations that organize multiple abstract ideas and large bodies of knowledge. Such skills only arise in individuals with extensive experience in particular domains of activity (e.g., advance training in a discipline).

Recent Advances: The Dynamic Self-organization of Integrative Psychological Structures

Epigenesis refers to the idea that psychological and behavioral structures emerge in development as a product of dynamic coactions between genes and a series of nested environments. Genes and environment are inseparable as causal processes in development. Although Piaget's equilibration model is an epigenetic one, the role he assigned to the environment was rather limited. In postulating a universal stage sequence, Piaget did not allow that schemes could develop in significantly different ways in different sociocultural contexts. According to contemporary epigenetic systems models of development, different developmental pathways self-organize as a product the interplay among diverse aspects of the person–environment system over time (Gottlieb, 2002; van Geert, 2000). One such model of the person–environment system is displayed in the top panel of **Figure 4** (Mascolo and Fischer, 2015). In this model, psychological structures are the emergent product of coactions that occur among systems that operate both within and between individuals. The integrative organization of individual action is indicated in the left panel of **Figure 4**. In any given context, (1) nonconscious *appraisals* of one's relation to one's environs (e.g., judgments of danger, goal blockage) generate (2) *emotional feelings* (e.g., fear, frustration) and (3) *action tendencies* (flee, fight). Thereupon, emotional feelings (4) provide feedback that *amplifies* the importance of appraised events and *selects* them for (5) conscious awareness within working memory. In this way, psychological structures that operate in working memory arise from the interplay of nonconscious cognitive and emotional processes (*see* Emotion in Cognition; Emotions, Psychological Structure of). Thus, although psychological structures regulate action, they do not do so autonomously; they self-organize as a product of the dynamic interplay among processes that operate both within and between individuals (*see* Language Development: Emergentist Theories).

To illustrate the emergent self-organization of psychological structures in real time, consider two brothers engaged in parallel play. After B referred to L as his 'little brother,' the following interaction occurred:

L: (Playing) No. I'm not a *little* brother. I'm big.
B: (Gently) I know you're not little, but (in a slightly sing-song 'teaching' voice) you're. When someone is younger than another brother, they call that a little brother.
L: Oh. (Slightly whining) I'm not (with rising intonation) *little* (looks toward O).

The bottom panel of **Figure 4** shows the emergent structure of thinking, feeling, and acting as it operated both *within* and *between* the two brothers. The left and right panels identify the structure of B's and L's activity, respectively. For each boy, the diagram identifies (1) the *appraisal* that organized affective and cognitive activity, (2) the *emotional expressions* produced by the appraisal, and (3) the structure of the child's *skilled action*. The symbol between the two panels identifies the ways in which the boys adjusted their behavior in relation to each other. In this situation, the little brother's slightly sad complaint that "I'm not little; I'm big" operated at the level of *representational mappings* (Fischer's, 1980). Responding empathically to the little boy's plea, the big brother modified the affective quality of his action, and, operating at the level of *representational systems*, attempted to console his brother. As indicated in this example, psychological structures are not isolated cognitive systems encapsulated within discrete individuals. Instead, they are integrative structures of thinking, feeling, and action that self-organize as interlocutors adjust their processes to each other over time.

Psychological Structures Are Dynamic Constructions

Psychological structures self-organize in both real and developmental time. To say that they self-organize implies that there is no single, fixed or isolated genetic, psychological or sociocultural plan that directs the course of development. The pathways of development are neither fixed nor predetermined. Instead, they emerge over time as a product of richly interactive person–environment interactions. Novel skills coevolve with the formation of novel forms of cultural life. Rather than thinking of development in terms of fixed

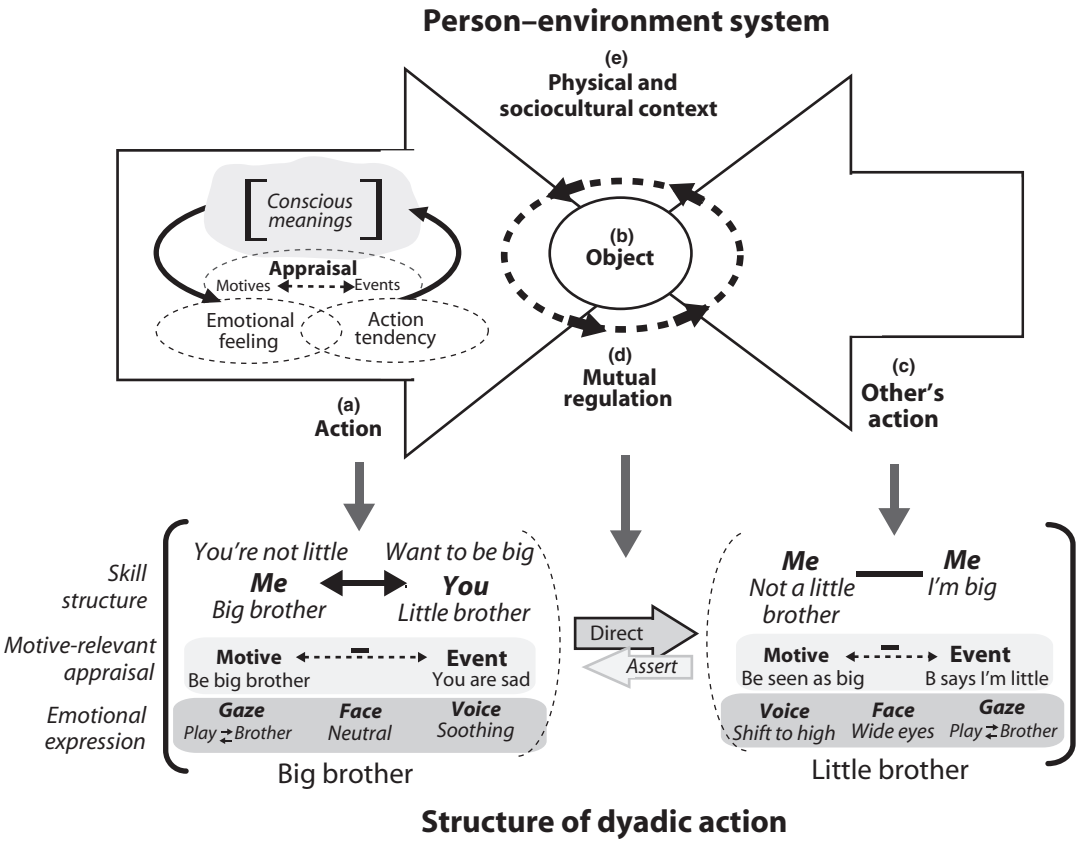

Figure 4 The joint construction integrative structures of action.

pathways, it is better to think of developmental pathways themselves as emergent outcomes. That is, while we sometimes walk along pathways that have already been paved for us, in development, we typically forge our own unique paths as we walk.

See also: Conscious Control During Childhood, Development of; Emotion in Cognition; Emotions, Psychological Structure of; Executive Functions During Childhood, Development of; Language Development: Emergentist Theories; Motor Skills: Development in Infancy and Early Childhood; Neo-Vygotskian Developmental Theory; Piaget's Theory of Cognitive Development; Piaget's Theory of Human Development and Education; Piaget, Jean (1896–1980); Self-Regulation During Adolescence: Variations Associated with Individual–Context Relations; Short- and Long-term Memory During Infancy and Early Childhood across Cultures; Vygotsky's Theory of Cognitive Development; Vygotsky's Theory of Human Development and New Approaches to Education; Vygotsky, Lev Semenovic (1896–1934).

Bibliography

Andrews, G., Halford, G.S., 2002. A cognitive complexity metric applied to cognitive development. Cognitive Psychology 45, 153–219.

Baddeley, A., 2012. Working memory: theories, models, and controversies. Annual Review of Psychology 63, 1–29.

Bryant, P.E., Trabasso, T., 1971. Transitive inference and memory in young children. Nature 232, 456–458.

Case, R., 1985. Intellectual Development: Birth to Adulthood. Academic Press, New York.

Case, R., 1991. The Mind's Staircase: Exploring the Conceptual Underpinnings of Children's Thought and Knowledge. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, England.

Case, R., Okamoto, Y., Griffin, S., McKeough, A., Bleiker, C., Henderson, B., Stephenson, K.M., 1996. The role of central conceptual structures in the development of children's thought. Monographs of the Society for Research in Child Development 61, 1–2. Serial No. 246.

Commons, M., Trudeau, E., Stein, S., Richards, F., Krause, S.R., 1998. Hierarchical complexity of tasks shows the existence of developmental stages. Developmental Review 18, 237–278.

Demetriou, A., 2006. Neo-Piagetische Theorien kognitver Entwicklung. In: Schneider, W., Wilkening, F.(Hrsg.). (Eds.), Theorien, Modelle, und Methoden der Endwicklungpsychologie (Enzyklopadie der Psychologie, Serie Entwicklungspychologie). Hogrefe, Gotingen, pp. 191–263.

Demetriou, A., Mouyi, A., 2011. Processing efficiency, representational capacity, and reasoning: modeling their dynamic interactions. In: Barrouillet, P., Gaillard, V. (Eds.), Cognitive Development and Working Memory: A Dialogue between Neo-Piagetian Theories and Cognitive Approaches. Psychology Press, New York, US, pp. 69–103.

Demetriou, A., Spanoudis, G., Mouyi, A., 2010. The development of mental processing. In: Overton, W.F., Lerner, R.M. (Eds.), The Handbook of Life-span Development, Cognition, biology, and methods, vol. 1. John Wiley & Sons, Hoboken, NJ, pp. 306–345.

Fischer, K.W., 1980. A theory of cognitive development: the control and construction of hierarchies of skills. Psychological Review 87, 477–531.

Fischer, K.W., Rotenberg, E.J., Bullock, D.H., Raya, P., 1993. The dynamics of competence: how context contributes directly to skill. In: Wozniak, R.H., Fischer, K.W. (Eds.), Development in Context: Acting and Thinking in Specific Environments. Erlbaum, Hillsdale, NJ, pp. 93–117.

Fischer, K.W., Bidell, T.R., 2006. Dynamic development of psychological structures in action and thought. In: Lerner, R. (Ed.), Damon, W. (Series Ed.), Handbook of Child Psychology, fifth ed. vol. 1. Theoretical Models of Human Development, Wiley, New York.

Fischer, K.W., Shaver, P.R., Carnochan, P., 1990. How emotions develop and how they organise development. Cognition and Emotion 4, 81–127.

Gelman, R., Baillargeon, R., 1983. A review of some Piagetian concepts. In: Flavell, J.H., Markman, E. (Eds.), Cognitive Development, Handbook of Child Development, vol. 3. John Wiley, New York, pp. 167–230.

van Geert, P., 2000. The dynamics of general developmental mechanisms: from Piaget and Vygotsky to dynamic systems models. Current Directions in Psychological Science 9, 64–68.

Gottlieb, G., 2002. Individual Development and Evolution: The Genesis of Novel Behavior. Lawrence Erlbaum, Mahwah, NJ.

Halford, G.S., 1993. Children's Understanding: The Development of Mental Models. Erlbaum, New York.

Halford, G.S., Andrews, G., 2006. Reasoning and problem solving. In: Kuhn, D., Siegler, R. (Eds.), Handbook of Child Psychology: Cognitive, Language and Perceptual Development, sixth ed., vol. 2. John Wiley & Sons, Hoboken, New Jersey, pp. 557–608.

Halford, G.S., Andrews, G., Wilson, W.H., Phillips, S., 2012. Computational models of relational processes in cognitive development. Cognitive Development 27, 481–499.

Klahr, D., Wallace, J.G., 1976. Cognitive Development: An Information-Processing View. Lawrence Erlbaum, Oxford England.

Mascolo, M.F., 2013. Developing through relationships: a coactive systems framework. In: Lerner, R., Benson, J. (Eds.), Embodiment and Epigenesis: Theoretical and Methodological Issues in Understanding the Role of Biology within the Relational Developmental System. Elsevier, New York.

Mascolo, M.F., 2005. Change processes in development: the concept of coactive scaffolding. New Ideas in Psychology 23, 185–196.

Mascolo, M.F., Fischer, K.W., 2015. Dynamic development of thinking, feeling and acting. In: Overton, W., Molennar, P. (Eds.), Handbook of Child Psychology and Developmental Science, vol. 1. Theory and Method. John Wiley, New York, pp. 113–161.

Matusov, E., Hayes, R., 2000. Sociocultural critique of Piaget and Vygotsky. New Ideas in Psychology 18, 215–239.

Pascual-Leone, J., 1970. A mathematical model for the transition rule in Piaget's developmental stages. Acta Psychologica 32, 301–345.

Pascual-Leone, J., Johnson, J., 1991. The psychological unit and its role in task analysis. A reinterpretation of object permanence. In: Chandler, M., Chapman, M. (Eds.), Criteria for Competence: Controversies in the Assessment of Children's Abilities. Eribaum, Hillsdale, NJ, pp. 153–187.

Saxe, G.B., 1991. Culture and Cognitive Development: Studies in Mathematical Understanding. Erlbaum, Hillsdale, NJ.

Vygotsky, L.S., 1978. Mind in Society. Harvard University Press, Cambridge, MA.

Neo-Vygotskian Developmental Theory

Yuriy V Karpov, Touro College, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Lev Vygotsky's Russian followers have elaborated his ideas into an innovative developmental theory, in which each period in children's development is characterized by their leading activity. In the context of this activity, adults mediate children's mastery of new psychological tools, which results in the development of new mental processes in children. As another outcome of adult mediation, children develop a new motive. As a result, children transit to the new leading activity that is specific to their next age period. Vygotsky's followers have used this theoretical model to analyze children's development in industrialized societies from birth through adolescence.

After Lev Vygotsky's premature death in 1934 at the age of 37 years, his sociocultural developmental theory became the general theoretical foundation for the scientific research and discourse of his Russian colleagues and followers (Daniil Elkonin, Piotr Galperin, Alexei Leontiev, Alexander Luria, Alexander Zaporozhets, and others). They, however, were far from considering Vygotsky's theory a dogma; rather, their scientific efforts were aimed at overcoming the shortcomings of Vygotsky's theory and elaborating those of his ideas that were just sketched by Vygotsky.

Neo-Vygotskian Theoretical Ideas

In their scientific discourse, the neo-Vygotskians repeated all the major steps that Vygotsky undertook when developing his theory. At each step, however, they refined and elaborated the original ideas of Vygotsky.

Psychological Tools

The key idea of Lev Vygotsky's theory is his analogy between practical and mental activities of humans. He held that the major characteristic of human mental processes is that they, just like human labor, are mediated by tools. But, these are special, psychological tools such as language, concepts, signs, and symbols. Humans are not born with these tools, just as they are not born with tools of labor; rather, these tools reflect the accumulated experience of human kind that children acquire and master. Having been mastered by children, psychological tools come to mediate their mental processes. Specifically, human mental processes that are mediated by tools were called by Vygotsky higher mental processes to distinguish them from lower mental processes, with which children are born, and which are specific to both young children and animals.

Having appropriated from Vygotsky the notion of psychological tools as mediators of human mental processes, his Russian followers, however, have considerably elaborated upon this notion. For Vygotsky, psychological tools were semiotic tools (language, concepts, signs, and symbols). Vygotsky, to be sure, realized that a child's learning a new word, concept, sign, or symbol does not automatically make this word, concept, sign, or symbol a psychological tool that comes to mediate this child's mental processes. Discussing, for example, scientific concepts as psychological tools that come to mediate mental processes of school-aged children, Vygotsky (1934/1986) noted that "the difficulty with scientific concepts lies in their *verbalism*" (p. 148). He also pointed out that "scientific concepts … just start their development, rather than finish it, at a moment when the child learns the term or word-meaning denoting the new concept" (Vygotsky, 1934/1986: p. 159). Vygotsky, however, never elaborated these reservations. Therefore, his theory is traditionally (and, in general, correctly) associated with the notion of semiotic tools as mediators of human mental processes (see, e.g., Kozulin, 1986).

Could, however, words, concepts, signs, or symbols by themselves serve as psychological tools mediating human mental processes? From Vygotsky's (1981) perspective, a sign "alters the entire flow and structure of mental functions. It does this by determining the structure of a new instrumental act, just as a technical tool alters the process of a natural adaptation by determining the form of labor operations" (p. 137). The point is, however, that a practical (or technical) tool by itself does not determine the form or structure of a human operation, since the possession of a tool does not lead automatically to the mastery of the procedure for the use of this tool. For example, mere giving chopsticks to one who never used them before will not make it possible for him to use them. The procedure for the use of this tool as a tool for eating will still be missing. Similarly, giving a spoon to an infant does not change the structure of his or her manual operation, and it is not sufficient for the use of this tool in the right way. To use the spoon as a tool for eating, the infant should master the procedure for the use of this tool. Therefore, "the mastery of a tool does not simply mean the possession of the tool, but it means the mastery of the procedure for the use of this tool" (Leontiev, 1959: p. 213).

The same is true of psychological tools. The mastery of a psychological tool requires that the child has mastered the procedure for the use of this tool (Galperin, 1969; Leontiev, 1959, 1983). As Leontiev (1983) indicated, "In order for a child to develop the highest generalization (a concept), it is necessary to develop in him the system of psychological operations [procedures – Y.K.] that are relevant to this highest generalization" (p. 347). Thus, the acquisition of Archimedes' law or the concept of perpendicular lines implies not only that

students are able to repeat this law or concept, but that they have also mastered the procedures that make it possible for them to apply this law or concept to solve subject-domain problems. The elaboration by Vygotsky's followers of his notion of psychological tools has been especially important for the development by them of an innovative approach to instruction, which will be discussed below.

Internalization of Psychological Tools

In Lev Vygotsky's writings, a major aspect of mastery of a psychological tool is its internalization, which, in line with his view of the nature of psychological tools, he understood as "the ingrowing of signs into the individual mental processes" (Arievitch and Van der Veer, 1995: p. 114). For example, from Vygotsky's perspective, a major phenomenon in memory development takes place when "the external sign (e.g., a string tied around the finger – Y.K.) that school children require has been transformed into an internal sign produced by the adult as a means of remembering" (Vygotsky, 1978: p. 45).

From the neo-Vygotskians' perspective, as discussed, the mastery of a psychological tool requires that the child has mastered the procedure for the use of this tool. Therefore, in their view, internalization of a psychological tool involves the internalization of the procedure of the use of this tool. As Alexei Leontiev (1959) wrote, "The mastery of mental procedures, which underlies the acquisition by an individual of knowledge and concepts accumulated by human kind, necessarily requires the transition from the performance of external procedures to … gradual internalization of the procedures, which results in their transformation into abridged mental procedures" (p. 305). Thus, in case of memory development, the difference between school children and adults is not that, as Vygotsky held, in order to memorize and recall, children use external signs, and adults use internal signs. Rather, the difference is in the level of internalization of their mnemonic procedures. The mnemonic procedures of school children are partially exteriorized and therefore cannot be performed without an external support (such as a string tied around the finger), whereas adults can use internal mnemonics.

Mediation as the Determinant of Mental Development

In contrast to Jean Piaget, who believed that children develop new mental processes as a result of independent exploration of the environment, Lev Vygotsky held that, being products of human culture, psychological tools should be explicitly taught to children by representatives of this culture. The process of teaching children psychological tools has been called mediation. In the course of mediation, an adult presents to the child a new psychological tool in the form of an external device and monitors the child's appropriation and mastery of this tool. As the child increasingly masters the tool, it gets internalized and turns into an internal mediator of the child's mental process. Simultaneously, the adult is getting less and less involved in monitoring the child's mastery of this tool. As a result, the child transits from the use of the external psychological tool as mediated by the adult, to the independent use of the internal psychological tool, which indicates the completion of the development of a new higher mental process.

Having enthusiastically adopted Vygotsky's general notion of mediation as the major determinant of mental development, the neo-Vygotskians have significantly elaborated on Vygotsky's understanding of the nature of mediation. Vygotsky's view on psychological tools as semiotic tools determined his understanding of the process of mediation as child–adult verbal communication. As Alexei Leontiev and Alexander Luria characterized Vygotsky's view on mediation, it leads to the position that "whatever is of major importance for the development of individual consciousness, is introduced into it through social consciousness" (Leontiev and Luria, 1968: p. 353).

For the neo-Vygotskians, as noted, the mastery of a psychological tool requires that the child has mastered the procedure for the use of this tool. But, whereas, for example, Archimedes' law can be taught to children via verbal communication, the procedures for application of this law to solve subject-domain problems cannot be just verbally described. Rather, these procedures have to be modeled by an adult and taught to children, and then mastered and internalized by children in the course of their use in problem solving under adult guidance. Therefore, for the neo-Vygotskians, mediation has to take place in the context of child–adult joint activity.

Motivational Factors in Child Development

Stressing mediation as the major determinant of development, Lev Vygotsky was far from viewing children as passive recipients of adult mediation. Discussing, for example, school instruction, which he considered to be the major avenue for mediation during the period of middle childhood, Vygotsky (1984/1998) indicated that "children … never learn without interest" (p. 24). Thus, according to Vygotsky, children's motivation is a necessary prerequisite for their engagement in mediational interactions with adults. As children transit from one developmental stage to the next, they have to become engaged in new mediational interactions, which requires that they have developed new motives. The development of new motives, according to Vygotsky "is the content of the social-cultural development of the child to a much greater degree than of his biological formation" (1984/1998: p. 11). He never elaborated this contention; since, however, Vygotsky viewed the social-cultural development of the child as an outcome of mediation, it can be assumed that he considered mediation to be mainly responsible for the development of new motives in children.

For the neo-Vygotskians, who stress the role of child–adult joint activity as the major avenue for mediation, the problem of children's motivation has become of even greater importance than it was for Vygotsky. Indeed, "there can be no activity without a motive"; a motive serves as "an activity's object that gives it a specific direction" (Leontiev, 1972/1981: p. 59). Therefore, it is the development of a new motive that propels the child to become engaged in a new activity.

Discussing the mechanism of the development of new motives in children, the neo-Vygotskians describe it as the process of ripening of a new motive within the context of a current activity (Leontiev, 1959). For example, if a person is engaged in job-related learning in order to improve his job

performance, his learning is driven by a motive of successful job performance rather than by learning motivation. It may happen, however, that this person at a certain point develops a strong intrinsic interest in learning within a certain domain. Then, the new learning motive will propel this person to become engaged in a new activity (e.g., to enroll in a college program).

In the above example, the ripening of a new motive has happened spontaneously. The neo-Vygotskians hold, however, that adults, rather than relying on the chance of spontaneous formation of new motives in children, should actively mediate this process in the context of their joint activities with children (Elkonin, 1971/1972, 1989; Leontiev, 1959, 1972/1981; Zaporozhets, 1978/1997; Zaporozhets and Elkonin, 1964/1971).

The Notion of a Leading Activity as the Cornerstone of the Neo-Vygotskian Model of Child Development

The term leading activity was initially used by Lev Vygotsky in reference to preschoolers' play as leading their development in this age period (1966/1976: p. 537). For him, however, this term was a figure of speech rather than a well-defined concept. To characterize different periods in children's development, Vygotsky preferred to use the term "the social situation of development," which he vaguely defined as a "unique relation, specific to the given age, between the child and reality, mainly the social reality, that surrounds him" (1984/1998: p. 198). It is in the context of the present social situation of development that, according to Vygotsky, adults engage children in age-specific mediational interactions. As a result, children develop new motives and new mental processes, which leads to their transition to the next social situation of development, that is, to the next developmental period.

For the neo-Vygotskians, who stress the role of children's activity as mediated by adults in their development, it has been a logical move to abandon Vygotsky's vague term "social situation of development" and to adopt instead the notion of a leading activity. Introducing this notion, Alexei Leontiev (1959) suggested that among different activities in which children are engaged in a given age period, there is one activity, the leading one, which plays the major role in children's development during this age period and prepares them for the transition to the next age period. What activity is leading in the given age period is crucially determined by the society in which children live. For example, in industrialized societies, in which children are required to learn in educational settings during the period of middle childhood, the leading activity of preschoolers would be the one that results by the end of preschool period in the development of children's school readiness. Sociodramatic play, according to Leontiev, meets this requirement the best. Therefore, sociodramatic play is the leading activity of preschool children in industrialized societies.

The notion of a leading activity has been enthusiastically embraced by Daniil Elkonin, Alexander Zaporozhets, and other neo-Vygotskians, and it has become the cornerstone of their theoretical model of child development (Elkonin, 1971/1972, 1989; Zaporozhets, 1978/1997; Zaporozhets and Elkonin, 1964/1971), which can be presented as follows. In each age period, children are engaged in the leading activity, which is driven by their leading motive (i.e., the most important one among all their motives in this age period). The children's present mental processes are not sufficient for their independent performance of this activity; therefore, this activity is performed as joint child–adult activity. In the context of this activity, adults mediate children's mastery of new psychological tools, which results in the development of new mental processes in children. As another outcome of adult mediation, children develop a new strong motive, which becomes their new leading motive. As a result, children transit to the new leading activity, which is specific to their next age period.

Neo-Vygotskian Analysis of Developmental Periods

The above-described theoretical model has been used by the neo-Vygotskians to analyze developmental periods in industrialized societies and to explain the mechanism of children's transition from one developmental period to the next.

First Year of Life: Emotional Interactions with Caregivers as the Leading Activity of Infants

In discussing the first year of life, Lev Vygotsky (1984/1998) emphasized infants' 'biological helplessness' (p. 215), inability to satisfy their vital physiological needs by themselves, which results in their dependence on adults. This dependence, according to Vygotsky, becomes the basis for the development of infants' emotional ties to caregivers (in contemporary psychology, these ties are often referred to as attachment).

Having embraced the aforementioned ideas, the neo-Vygotskians, at the same time, argue that infants' attachment to caregivers is not a direct and necessary outcome of the gratification of their physiological needs, as Vygotsky (1984/1998) was inclined to believe. Rather, the formation of attachment is the result of emotional interactions, in which caregivers engage infants in the context of gratification of their physiological needs by smiling at them, talking to them, and fondling them. As a result, infants, initially not responsive to these adult efforts, gradually start, first, to respond to adult initiations of emotional interactions, and then, to initiate these interactions themselves by smiling, vocalizing, and using body gestures. Infants' taking initiative in the emotional interactions with caregivers indicates that they have developed a new motive, which comes to drive their new activity of emotional interactions with caregivers. This activity, according to the neo-Vygotskians, becomes infants' leading activity during the first year of their lives, because it results in the major developmental accomplishments of this period, with attachment being the most important of them (Elagina, 1977; Elkonin, 1971/1972, 1989; Lisina, 1986; Zaporozhets and Lisina, 1974). The other developmental accomplishments (many of which are built upon attachment) are as follows.

First, infants' affection to caregivers is 'transferred' to toys and objects that caregivers present to them in the context of emotional interactions with encouragement to manipulate these objects. As a result, infants develop interest in object-centered explorations, which, by the end of the first year,

becomes much stronger than their interest in emotional interactions with adults and turns into infants' new leading motive.

Second, attachment results in the 'acceptance' of caregivers by infants as mediators of their relations with the external world. In particular, this 'acceptance' reveals itself in infants' using primary caregivers as the 'reference point' in an uncertain situation, as well as in their actively seeking for caregivers' help with new objects and toys by the end of the first year.

Third, the advancement of infants' object-centered explorations increases their need for caregivers' help with new objects and toys, which requires that they develop means of communication with caregivers. Accordingly, infants willingly acquire the means of gestural communication that caregivers model for them. The most important of them, the pointing gesture, in addition to its communicative value, becomes the foundation for later language development and the development of self-regulation in children.

Thus, by the end of the first year of life, as a result of adult mediation, infants develop all the prerequisites for their transition to the next age period, in which object-centered joint activity with adults becomes their new leading activity.

Second and Third Years of Life: Object-Centered Joint Activity with Adults as the Leading Activity of Toddlers

To be sure, toddlers learn much about the physical world through independent explorations, such as shaking and throwing objects, pushing small objects into different holes, and so on. However, since these explorations do not require adult assistance, they do not create a context for adult mediation. Therefore, the neo-Vygotskians give major importance to toddlers' actions with social objects, including their play with toys. As opposed to physical characteristics of objects, their social meanings are not 'written' (Elkonin, 1989: p. 48) on objects and, therefore, cannot be discovered by children independently. Thus, children's learning how to use social objects (e.g., how to use a spoon to eat rather than banging it on the table) can take place only within child–adult joint activity, which provides adults with a context for mediation. That is why the neo-Vygotskians hold that object-centered joint activity with adults becomes the leading activity of toddlers during the second and third years (Elkonin, 1978, 1989; Zaporozhets and Lisina, 1974).

Discussing the advancement of children's object-centered activity, the neo-Vygotskians stress the role of adults who introduce new objects and toys to children and model their proper use (Bugrimenko and Smirnova, 1994; Elkonin, 1978). The neo-Vygotskians especially stress the fact that it is adults who initially introduce object substitutes to children and teach them how to use these substitutes in play (e.g., by suggesting to the child to use a stick as a substitute for a spoon when feeding a doll), which marks children's transition from realistic to symbolic play.

The developmental outcomes of object-centered joint activity (which make this activity the leading one in this age period) are as follows. First, as this activity advances, it comes to require that children develop more sophisticated means of communication with adults than, for example, a pointing gesture. As a result, children, with caregiver guidance and encouragement, acquire language remarkably fast (Elkonin, 1989; Lisina, 1986). This is especially important since language is not only the major tool for social communication, but it is also a psychological tool that comes to mediate children's mental processes. In particular, children become capable of self-regulation by the use of private (self-directed) speech (Luria, 1961).

Second, the use of object substitutes (which represents "a vital transitional stage to operating with meanings" (Vygotsky, 1966/1976: p. 548)), and the acquisition of language (which, in particular, makes it possible for children to rename object substitutes for missing objects) promotes the development in children of symbolic thought (Elkonin, 1978).

Third, with adult help, children discover the exiting world of social roles and relationships that is 'hidden' under the world of social objects (Elkonin, 1978, 1989; Leontiev, 1959). As a result, the imitation of social roles and relationships becomes children's new strong motive (e.g., when playing a mother–daughter game, a 2-year-old girl will imitate the action of feeding a doll with a spoon, whereas a 3-year-old girl will imitate mother–daughter relationships, such as love, care, etc.)

Thus, by the end of the third year, as a result of adult mediation, children develop all the prerequisites for their transition to the next age period, in which sociodramatic play becomes their new leading activity.

Three- to Six-Year-Olds: Sociodramatic Play as the Leading Activity during the Period of Early Childhood

The Vygotskian view on sociodramatic play (which is interchangeably referred to as 'play' in this section) is just the opposite of the common belief that play is children's free spontaneous activity in which they do whatever they want, liberating themselves from any rules and social pressure (Vygotsky, 1966/1976; Elkonin, 1978, 1989; Leontiev, 1959; Zaporozhets, 1978/1997). Rather, following Lev Vygotsky, the neo-Vygotskians hold that by the age of 3 years children develop a strong interest in the world of social roles and relations; that is why they start to imitate and explore these roles and relations in play. Adults facilitate and promote sociodramatic play by presenting and explaining different social roles to children. Such mediation of sociodramatic play by adults is especially important since play leads to major developmental outcomes of this age period.

First of all, in play, the child "looks at himself from the point of view of the role he has taken … and discovers that he is not an adult yet" (Elkonin, 1978: p. 277). Thus, in play children come to realize that, in order to become adults, they should study at school rather than playing adults. As a result, by the age of 6 years, children develop learning motivation as their new leading motive (it is important to note that, although this neo-Vygotskian assumption has been supported by observations (Bozhovich, 1968), it is still lacking strong empirical support).

Second, Vygotsky (1966/1976) and his followers (Elkonin, 1978; Leontiev, 1959) emphasize that, contrary to 'common wisdom,' children are not free in play. In play, children are supposed to act in accordance with their roles, and all playmates are strictly monitoring each other's proper following of his or her play role. But, each role contains some implicit rules; a rule "is hidden … under the role" (Elkonin, 1978: p. 248). Thus, as

a result of mutual monitoring in the context of their play, children learn to suppress their impulsivity and to follow rules, that is, develop the ability to self-regulate their behavior.

Third, in play, children overcome their cognitive egocentrism, that is, inability to take into consideration the position of another person, or to look at a certain event from different perspectives. Indeed, play requires that children coordinate their play actions with the play actions of their playmates; therefore, in play the child's "ability to coordinate his point of view with other possible points of view develops" (Elkonin, 1978: p. 282).

Finally, in the course of sociodramatic play, children use object substitutes and perform within an imaginary situation (develop the plot of the game, plan their play actions, etc.), which further promote the development of their symbolic thought (Elkonin, 1978; Vygotsky, 1966/1976).

The neo-Vygotskians hold that the above developmental accomplishments are crucially important for children's successful learning at school (Bozhovich, 1968; Elkonin, 1978; Leontiev, 1959). Thus, in the course of sociodramatic play, children develop the major components of school readiness, which is a necessary prerequisite for their transition to learning at school as their new leading activity during the period of middle childhood.

The Period of Middle Childhood: Learning at School as Children's Leading Activity

According to Lev Vygotsky (1978, 1934/1986), school instruction is the major avenue for mediation and, therefore, the major contributor to children's development during the period of middle childhood. He emphasized, however, that such a developmental effect will take place only if instruction is organized in the proper way: "The only good kind of instruction is that which marches ahead of development and leads it" (Vygotsky, 1934/1986: p. 188). The major reason for developmental outcomes of school instruction relates to students' acquisition of scientific concepts, which, having been acquired and internalized, become the tools of children's thinking and problem solving; "instruction in scientific concepts plays a decisive role in the child's mental development" (Vygotsky, 1982/1987: p. 220). Specifically, "reflective consciousness comes to the child through the portals of scientific concepts" (Vygotsky, 1934/1986: p. 171). As a result, students' thinking becomes much more independent of their personal experience, and they become capable of exercising formal-logical thought (Luria, 1974/1976; Tulviste, 1991).

The neo-Vygotskians have embraced the aforementioned ideas of Vygotsky and have characterized learning in educational settings as the leading activity of children in industrialized societies during the period of middle childhood; they, however, substantially elaborated Vygotsky's views of school instruction (Davydov, 1972/1990, 1999; Elkonin, 1989; Galperin, 1974/1989; Leontiev, 1959; Talyzina, 1975/1981). As already noted, the neo-Vygotskians repudiated the vision of 'good' instruction as providing children with scientific concepts in the course of verbal communication (the position to which Vygotsky's writings sometimes could lead); rather, they argue, students should be presented with a combination of scientific concepts and relevant procedures. Neo-Vygotskian studies have demonstrated that the procedures that underlie the major subject-domain concepts are methods for scientific analysis in a given subject domain. For example, the procedures that underlie Archimedes' law are the methods of calculating the density of different objects and comparing these densities with the density of water. Similarly, the procedures that underlie the concept of perpendicular lines are the methods of identifying within a given pair of lines those attributes that are necessary and sufficient for associating (or not associating) this pair of lines with the concept of perpendicular lines. The mastery and internalization of scientific concepts and procedures by students should take place in the course of their use of this knowledge under teacher guidance for solving subject-domain problems. The implementation by the neo-Vygotskians of these ideas in educational practices has become known as the theoretical learning approach to instruction.

As evaluation studies have demonstrated, the use of theoretical learning improves dramatically both the learning and developmental outcomes of school instruction (for reviews in English, see Arievitch and Stetsenko, 2000; Davydov, 1972/1990; Talyzina, 1975/1981). One of its major developmental outcomes relates to the facilitation of the development of formal-logical thought in students to a much greater extent than happens under 'traditional' instruction. These findings are especially important since, as will be discussed in the next section, formal-logical thought is a necessary prerequisite for children's transition to the next period of development, that is, the period of adolescence.

The Period of Adolescence: Interactions with Peers as the Leading Activity of Adolescents

According to Lev Vygotsky (1984/1998), the change of adolescents' interests is "the key to the whole problem of the psychological development of the adolescent" (p. 3). This change not only has its roots in physiological (especially, sexual) maturation, but also "there is a reconstruction and formation of interests from the top, from the aspect of the maturing personality and the world view of the adolescent" (Vygotsky, 1984/1998: p. 23). As for the development of adolescents' 'personality and world view,' it becomes possible because of their transition to formal-logical thought; as Vygotsky (1984/1998) noted, "an adolescent appears before us primarily as a thinking being" (p. 30). Because of the new ability to do formal-logical thought, "a whole world with its past and future, nature, history, and human life opens before the adolescent" (Vygotsky, 1984/1998: p. 42). Formal-logical thought also makes adolescents capable of self-analysis, the analysis of their feelings, and their place in the world. The tools for such analysis are social norms and values, which are provided to adolescents by the society; therefore, in Vygotsky's (1984/1998) words, "self-consciousness is social consciousness transferred within" (p. 182).

According to Russian followers of Vygotsky, the major shortcoming of his view of adolescence was overlooking of the role of adolescents' activity in their development. Applying the concept of leading activity to the analysis of adolescence, the neo-Vygotskians suggested that interaction with peers is the leading activity during this period (Elkonin, 1971/1972). To regulate their interactions within a peer group, adolescents use

social norms and values, which they have appropriated from their caregivers, as standards for the behavior of their peers, as well as reflect on their peers' use of social norms and values as standards for their own behavior (Bozhovich, 1968; Elkonin, 1989). As a result, they test, master, and internalize these norms and values, which leads to the development of their self-consciousness (or personal identity, in Erikson's terms) – the major accomplishment of this period that prepares adolescents for the transition to adulthood.

The strongest point of the neo-Vygotskians' explanation of adolescence is that they repudiated the vision of an adolescent as a passive recipient of social norms and values, or as a 'theorist' who suddenly becomes interested in the development of self-consciousness – the vision to which some of Vygotsky's writings may lead. In the neo-Vygotskian theoretical model, the major developmental accomplishments of adolescence (just as the major developmental accomplishments in other periods) take place only because they are required by the activity that is leading during this period.

The major flaw in the neo-Vygotskians' view of adolescence is that they have not answered the question, what motive propels adolescents to become engaged in interactions with peers? A possible answer to this question can be drawn from Vygotsky's idea (abandoned by his Russian followers) that sexual maturation lies "at the base of the whole change in the system of interests in the adolescent" (1984/1998: p. 16). Indeed, there is some evidence that it is sexual motives that propel adolescents to become engaged in interactions with peers (Cole and Cole, 2001: p. 626).

Neo-Vygotskian Developmental Theory: Accomplishments and Shortcomings

Russian followers of Vygotsky have elaborated his theoretical ideas into a logical and internally consistent theory that integrates cognitive, motivational, and social aspects of child development, while successfully overcoming the limitations of Vygotsky's theory. In their theory, adult mediation in the context of children's leading activity results in their cognitive and motivational development, which, in turn, results in their transition to the new leading activity that is specific to the next period of development. Thus, the neo-Vygotskians have achieved the goal that Vygotsky formulated but fell short of reaching; they have presented the process of child development as "a single process of self-development" (1984/1998: p. 189).

To avoid possible misunderstandings of neo-Vygotskian developmental theory, it is important to make two reservations. First, the leading activities described by the neo-Vygotskians are those that are specific for children in industrialized societies. In preliterate societies, neither sociodramatic play is the leading activity during the period of early childhood (Elkonin, 1978), nor is learning in educational settings the leading activity during the period of middle childhood. Second, even in industrialized societies, children may not be engaged in the activity that is leading in a given period; this, however, will result in serious developmental problems. For example, contemporary preschoolers are not much engaged in sociodramatic play, which has been shown to result, in particular, in very poor quality of their self-regulation (Bodrova and Leong, 2006).

Although the neo-Vygotskians' developmental theory has found strong empirical support in the studies of contemporary researchers, some of their notions can be disputed. Their most disputable notion is a disregard of the role of physiological maturation in children's development; as discussed, this disregard, in particular, has weakened the neo-Vygotskian analysis of adolescence.

See also: Cognitive Development, Educational Theories of; Piaget's Theory of Cognitive Development; Vygotsky's Sociocultural Theory; Vygotsky's Theory of Cognitive Development.

Bibliography

Arievitch, I.M., Van der Veer, R., 1995. Furthering the internalization debate: Galperin's contribution. Human Development 38, 113–126.

Arievitch, I.M., Stetsenko, A., 2000. The quality of cultural tools and cognitive development: Gal'perin's perspective and its implications. Human Development 43, 69–92.

Bodrova, E., Leong, D.J., 2006. Tools of the Mind: The Vygotskian Approach to Early Childhood Education, second ed. Prentis Hall, Upper Saddle River, NJ.

Bozhovich, L.I., 1968. Lichnost i ee formirovanie v detskom vozraste [Personality and its Development in Childhood]. Prosveschenie, Moscow.

Bugrimenko, E., Smirnova, E., 1994. Paradoxes of children's play in Vygotsky's theory. In: Cupchick, G., Laszlo, G. (Eds.), Emerging Visions of the Aesthetic Process. Cambridge University Press, Cambridge, pp. 286–299.

Cole, M., Cole, S.R., 2001. The Development of Children, fourth ed. Worth, New York.

Davydov, V.V., 1990. Types of Generalization in Instruction. National Council of Teachers of Mathematics, Reston, VA (Original work published 1972).

Davydov, V.V., 1999. What is real learning activity? In: Hedegaard, M., Lompscher, J. (Eds.), Learning Activity and Development. Aarhus University Press, Aarhus, Denmark, pp. 123–138.

Elagina, M.G., 1977. Vliyanie nekotorykh osobennostei obscheniya na vozniknovenie aktivnoi rechi v rannem bozraste [The influence of some characteristics of communication on the emergence of active speech in early childhood]. Voprosy Psikhologii 2, 135–142.

Elkonin, D.B., 1972. Toward the problem of stages in the mental development of the child. Soviet Psychology 10, 225–251 (Original work published 1971).

Elkonin, D.B., 1978. Psikhologiya igry [Psychology of Play]. Pedagogika, Moscow.

Elkonin, D.B., 1989. Izbrannye psikhologicheskie trudy [Selected Psychological Works]. Pedagogika, Moscow.

Galperin, P.Y., 1969. Stages in the development of mental acts. In: Cole, M., Maltzman, I. (Eds.), A Handbook of Contemporary Soviet Psychology. Basic Books, New York, pp. 34–61.

Galperin, P.Y., 1989. Organization of mental activity and the effectiveness of learning. Soviet Psychology 27, 65–82 (Original work published 1974).

Kozulin, A., 1986. The concept of activity in Soviet psychology: Vygotsky, his disciples and critics. American Psychologist 41, 264–274.

Leontiev, A.N., 1959. Problemy razvitiya psikhiki [Problems of Mental Development]. Izdatelstvo APN RSFSR, Moscow.

Leontiev, A.N., 1981. The problem of activity in psychology. In: Wertsch, J.V. (Ed.), The Concept of Activity in Soviet Psychology. Sharpe, Armonk, NY, pp. 37–71 (Original work published 1972).

Leontiev, A.N., 1983. Ovladenie uchaschimisya nauchnymi poniatiyami kak problema pedagogicheskoi; psikhologii [Mastering scientific concepts by students as a problem of educational psychology]. In: Leontiev, A.N. (Ed.), Izbrannye psikhologicheskie proizvedeniya, Tom 1. Pedagogika, Moscow, pp. 324–347.

Leontiev, A.N., Luria, A.R., 1968. The psychological ideas of L. S. Vygotskii. In: Wolman, B.B. (Ed.), Historical Roots of Contemporary Psychology. Harper & Row, New York, pp. 338–367.

Lisina, M.I., 1986. Problemy ontogeneza obscheniya [Problems of the Ontogenesis of Communication]. Pedagogica, Moscow.

Luria, A.R., 1961. The role of speech in the regulation of normal and abnormal behavior. Pergamon Press, Oxford, England.
Luria, A.R., 1976. Cognitive Development: Its Cultural and Social Foundations. Harvard University Press, Cambridge, MA (Original work published 1974).
Talyzina, N.F., 1981. The Psychology of Learning. Progress, Moscow (Original work published 1975).
Tulviste, P., 1991. The Cultural-historical Development of Verbal Thinking. Nova Science, Commack, New York.
Vygotsky, L.S., 1976. Play and its role in the mental development of the child. In: Bruner, J.S., Jolly, A., Sylva, K. (Eds.), Play: Its Role in Development and Evolution. Basic Books, New York, pp. 537–554 (Original work published 1966).
Vygotsky, L.S., 1978. In: Cole, M., John-Steiner, V., Scribner, S., Souberman, E. (Eds.), Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA.
Vygotsky, L.S., 1981. The instrumental method in psychology. In: Wertsch, J.V. (Ed.), The Concept of Activity in Soviet Psychology. Sharpe, Armonk, New York, pp. 134–143.
Vygotsky, L.S., 1986. Thought and Language. MIT Press, Cambridge, MA (Original work published 1934).
Vygotsky, L.S., 1987. In: Rieber, R.W. (Ed.), The Collected Works of L. S. Vygotsky, Vol. 1: Problems of General Psychology. Plenum, New York (Original work published 1982).
Vygotsky, L.S., 1998. In: Rieber, R.W. (Ed.), The Collected Works of L. S. Vygotsky, Vol. 5: Child Psychology. Plenum, New York (Original work published 1984).
Zaporozhets, A.V., 1997. Principal problems in the ontogeny of the mind. Journal of Russian and East European Psychology 35, 53–94. (Original work published 1978).
Zaporozhets, A.V., Elkonin, D.B. (Eds.), 1971. The Psychology of Preschool Children. MIT Press, Cambridge, MA (Original work published 1964).
Zaporozhets, A.V., Lisina, M.I. (Eds.), 1974. Razvitie obscheniya u doshkolnikov [The Development of Communication in Preschoolers]. Pedagogika, Moscow.

Network Analysis

Antonio M Chiesi, Università degli Studi di Milano, Milano, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Network analysis can be regarded as a set of techniques with a shared methodological perspective, which allow researchers to depict relations among actors and to analyze the social structures that emerge from the recurrence of these relations. The basic assumption is that better explanations of social phenomena are yielded by analysis of the relations among entities. This analysis is conducted by collecting relational data organized in matrix form. If actors are depicted as nodes, and their relations as lines among pairs of nodes, the concept of social network changes from being a metaphor to an operative analytical tool that utilizes the mathematical language of graph theory and linear assumptions of matrix algebra.

The Network Approach in Social Sciences

The notion that social phenomena can be interpreted as networks of relations, that society itself can be viewed as a web, and that social action can be seen as the outcome of constraints and opportunities emerging from interpersonal relations enjoys increasing consensus among social and behavioral scientists. The reasons for this success are in line with a view that fosters the use of the term 'network' as a metaphor for society: the diffusion of the Internet as an emblematic representation of globalization, the evolution of the economy into being based on increasingly open and flexible firms whose efficiency is related to their weblike structure, the emergence of a third sector based on informal relations as opposed to a declining welfare state.

In recent decades the success of relational approaches in the empirical social sciences has been undoubtedly supported by the rapid development of network analysis (NA) tools.

NA is a set of integrated techniques to depict relations between actors and to analyze social structures emerging from the recurrence of these relations. The basic assumption is that better explanations of social phenomena should be found in the analysis of the relations between entities, which is carried out through the collection of relational data organized in matrix form, rather than in their social characteristics, which are usually surveyed on a set of individuals sampled from a larger population.

Traditional sample surveys isolate each actor and reduce his or her social relations to individual characteristics, in order to analyze statistical relations between variables. These causal relations can be only interpreted as useful indicators of underlying mechanisms, which are in fact any kind of encounters and exchanges between real actors who display their everyday life strategies.

NA concentrates its analytical attention directly on these phenomenological encounters, while traditional surveys can only identify causal connections without explaining effective operating mechanisms. For instance, a statistical relation between social origin and status attainment can in no way depict the actual mechanism of social mobility; it is only a signal of social relations both among relatives and within the larger social milieu of the respondent. The shape and intensity of the network of these relations constitute the explanation of the persistence or changes of social status through generations. Although deterministic approaches usually emphasize the point that NA enables a study of how the social structure of relationships around a person, group, or organization affects behaviors and attitudes, social networks are in turn constructed by individual or corporate actors, to the extent that they can mobilize social resources to choose partners or counterparts and to organize coalitions. In short, structurally bounded purposive actions can interplay with social structure, and vice versa.

If actors are depicted as nodes and their relations as lines connecting pairs of nodes, the concept of social network changes from an evocative metaphor, as it has been for a long time since the beginning of the twentieth century, to an operative analytical tool that can take advantage of the mathematical language of graph theory and of matrix and relational algebra. Concepts and theorems of graph theory can be meaningfully applied to NA methods. Matrix algebra is a powerful tool for managing relational data, for measuring distances between nodes, and for computing centrality and centralization statistics (Wasserman and Faust, 1994).

Since the mid-1970s, the development of algorithms (which have subsequently been integrated in statistical packages) and the availability of greater computational capability in the decade that followed have enabled the successful application of these techniques to structures larger and more complex than the personal networks studied previously. (The best known and most widely used packages are Pajek, UCINET, and StOChastic NETworks (StOCNET). The last of these includes the SIENA program, which allows dynamic analysis of longitudinal network data. An extensive comparative analysis of 31 programs and software toolkits for NA is made available by Huisman and van Duijn in Carrington et al. (2005). Graphical techniques and visualization of networks programs are also presented and compared by Freeman in Carrington et al. (2005).) The foundation of specialized journals such as *Social Networks* and *Connections* and the publication of NA manuals in different languages (Scott, 1991; Wasserman and Faust, 1994; Degenne and Forsé, 1994; Chiesi, 1999) has helped to institutionalize this approach, so that NA now finds increasing acceptance in the literature of such disciplines as sociology, economics, anthropology, social psychology, history and political science, as well as communication science, and even

biology (application to animal behavior) and epidemiology (human immunodeficiency virus diffusion patterns), while statistics and mathematics contribute to the development of analytical tools. The activity of an interdisciplinary community of scholars is evidenced by the founding of the International Network for Social Network Analysis in the late 1970s.

These developments have induced some scholars, mostly structural analysts, to identify a new scientific paradigm. Indeed, NA has features typical of what Thomas Kuhn – in his well-known book on scientific revolutions – defined as a paradigm: the formation of a community of scholars who share a special language, publish journals, and meet regularly at conferences to discuss specialized topics concerning applications of technical tools and their theoretical interpretation. There is evidence of the cumulative nature of this activity, in spite of the heterogeneity of the disciplines involved and of its fields of application.

However, NA has long suffered from a lack of shared theoretical assumptions and of theories derived from these assumptions. The idea that society is itself a set of relations and that individual actors are embedded in social networks is still too loosely defined for it to become a theory. For these reasons, NA has been sometimes defined as a technique in search of a theory. One may accordingly regard it as a set of techniques with a shared methodological perspective, rather than a new paradigm in the social and behavioral sciences. These techniques make it possible to dismiss a metaphor and to undertake substantive research, which yields good descriptions of complex structures and generates theoretical interpretations of processes that imply exchanges, information flows, power relations, or social interpersonal support at individual level, as well as interorganizational relations of influence, coordination, and control.

Origins and Development

The straightforward idea that any relational phenomena can be interpreted as network structures, that society itself can be interpreted as a network, and that social action can be explained as the outcome of constraints imposed and the opportunities afforded by social relations has been recurrent since Simmel's formal sociology and his studies on the relations between actors in dyads and triads at the beginning of the twentieth century. Nevertheless, it is only since the 1930s that attempts have been made in various disciplines to move from theoretical programs to empirical applications (Scott, 1991).

The first approach, influenced by gestalt theory, was developed in the United States by a group of social psychologists who had emigrated to that country from Germany. Jacob Moreno, Kurt Lewin, and Fritz Heider stressed the role played by the structure of interpersonal relations in the construction of cognitive schemes, and they studied information flows and the formation of shared ideas. The main tool that they used for the description of empirical networks was the *sociogram* – nothing more than a graphic representation of links (the interpersonal relations) between nodes (the actors) – together with analytical concepts like symmetrical–asymmetrical relation, centrality and periphery, as well as reciprocity and cognitive balance.

Another approach is directly linked to structural functionalism and it was used to recast the classic research study on work organization in the wiring room of the Western Electric plant at Hawthorne by Fritz Roethlisberger and William Dickson during the 1930s. In that study, observation techniques drawn from anthropology had highlighted the importance of informal relations in work performance. Some years later, George Homans reorganized the same data in matrix form to conduct analysis very similar to what came to be called 'block model' in the 1970s.

The so-called Manchester School of Social Anthropology, whose main figures are Max Gluckman, John Barnes, Clyde Mitchell, and Elizabeth Bott, also started from structural functionalism. But in contrast to the traditional emphasis on social integration and cohesion, this school developed NA concepts to study conflict and power relations. Its shift of research subject from African communities to next-door neighbor relations in European cities led to the discovery of the importance of informal relations in urban settings. The Manchester School applied concepts like density, connectivity, and reachability as well as parameters related to the intensity and strength of ties, but its exclusive concern with informal relations and its merely descriptive approach have contributed to the decline of this school since the 1970s.

By this time the main technical developments had moved to Harvard, where Harrison White devised mathematical concepts of structural analysis, following the pioneering work of Frank Harary and Dorwin Cartwright during the 1950s. Thereafter, the mainstream program would be to model the formal properties of social structures on the basis of graph theory and matrix algebra. The approach was, therefore, analytical and deductive rather than descriptive and inductive. Different types of graphs representing networks were defined and their properties studied. Field research and applications generated statistical debate on centrality, centralization, and the identification of cliques and other types of subgroups, while important concepts such as 'structural equivalence' were proposed.

The main theoretical assumption that fostered these technical developments was structural analysis, which theorized that an actor's behavior and attitudes can be interpreted in terms of his or her structural position in the network. Since the end of the 1970s, however, this deterministic approach has been criticized, prompting a second generation of authors to investigate the structural opportunities available to actors when they use relations that can be mobilized from specific positions. In sociology especially, this reorientation has yielded results pertaining both to substantive theories and to analytical models, usefully employed to suggest explanations of social phenomena (Cook et al., 1983, for instance).

As regards the former of these domains, different analytical types of ties, as well as concepts such as point centrality, network centralization, and structural balance, have been proposed, together with numerous criteria for subgroup detection (*see* Social Network Models: Statistical). The concept of structural equivalence, introduced by Francis Lorrain and Harrison White in the 1970s, has been inspired directly by role theory.

Recent Results

As regards substantive research, a number of middle-range theories can be grouped into five different fields of empirical

applications, while a sixth interdisciplinary approach is specialized in the detection of common structural features of large networks.

Network Structure and Actor Behavior

The relation between network structure and actors' behaviors and attitudes has generated results demonstrating that the density of ego-centered networks is correlated with mental disorders (Kadushin, 1982), the efficiency of task-oriented groups depends on their structure (Bavelas, 1950), members of cohesive networks tend to share common opinions (Epstein, 1969), and innovation diffusion is influenced by personal relations between actors (Coleman et al., 1957, Valente in Carrington et al., 2005).

Networks and Social Inequalities

The role played by networks in preserving and creating social inequalities has been investigated by research on social mobility, where macro-outcomes at the level of social stratification and microstrategies are analyzed jointly by means of network techniques. Moreover, studies on the labor market have evidenced that people acquire information about job opportunities mainly from informal relations, and 'strong' or 'weak' interpersonal ties determine differential chances of entering employment according to the segmentation of the labor market based on skills (Granovetter, 1973; Grieco, 1987) or gender (Lin, 2004).

Networks and Economic Behavior

The function played by networks in economic behavior and market mechanisms has inspired studies of interlocking directorates (Stokman et al., 1985; *see* Interorganizational Relationships and Networks), and theoretical interpretations of the market as a structure of roles in which the strategy of each producer is based on the observed actions of all others (White, 1981). In spite of a long-standing tradition of interlocking directorates studies, there still exist a number of competing theories about the way these interlocks work (*see* Interorganizational Relationships and Networks).

Relational Aspects of Power

The relational concept of power – derived from Max Weber's classical work – is particularly suitable for NA applications in the study of political phenomena. It has been found that in negatively connected exchange networks power is not a function of centrality, and exchange choices may generate structural transformations of the overall network because of changes in power dependency relations (Cook et al., 1983). In historical analysis, it has been convincingly proposed (Padget and Ansell, 1993) that the rise of the Medici in Renaissance Florence and their increasing political power came about because of network disjunctures within the elite, which the Medici alone spanned. In this study, network structures are identified by block model analysis of different types of interpersonal relations based on marriages, friendship, common economic interests, exchanges, and political alliances.

Measuring Social Capital

The idea of measuring social capital through NA is more recent and has generated renewed interest in these techniques. Extensive empirical research on social capital began in political science in the early 1990s, thanks to Putnam's analysis of civic traditions in modern Italy (1993), and subsequently developed in other fields through the analysis of its economic consequences (Durlauf and Fafchamps, 2005). This preliminary approach extensively used a variety of indicators as proxies for social capital and as variables measuring degrees of civic integration, sociability, participation in voluntary associations, and even opinions about generalized or interpersonal trust taken from subsequent waves of the World Values Survey. In spite of interesting results, this traditional approach has been criticized because it is not based on a straightforward application of the concept of social capital as such. Scholars agree on a relational definition of social capital as "a set of interpersonal relations that are used by actors to fulfill their goals" (Coleman, 1988). Consequently, authors like Lin (2001), Flap (1999), Flap and Völker (2004) have directly applied NA techniques to study the performance of interpersonal relations structures in the realization of one's goals. Chiesi has shown (2007) that structural features of personal networks and their changes over time predict the business success of entrepreneurs. Dufhues et al. (2011) have found a positive effect of social capital on loan repayment behavior. This approach to social capital has also proved fruitful in the inclusion of previous research on strong and weak ties in a single theoretical perspective, as well as other research fields such as social support, ethnic communities, and migration studies. As strong ties tend to be reciprocal and clustered, the term 'bonding social capital' has been proposed in relation to common identity, feelings of belonging, and sometimes promotion of equality, while weak ties, which are typical of 'bridging social capital,' are mainly instrumental for the fulfillment of one's goals.

Large Networks and Their Common Structural Features

A recent interdisciplinary approach, inspired by Barabasi and enhanced by a group of mathematicians, physicists, computer scientists, sociologists, and biologists, has developed a new field aiming at developing a general theory of networks, which is proposed as a Science of Networks in its own. Barabasi and his colleagues (Newman et al., 2006) have pointed out that the study of networks in the field of social sciences has been mainly descriptive rather than analytical, because it has avoided large-scale modeling, due to the difficulty in getting high-quality large-scale network data. Some limits of field research (i.e., validity and reliability of name generator techniques) can be overcome by the recent availability of large-scale computerized sources of relational information coming from online communities, the World Wide Web, mobile phone calls, or transport movements. In this perspective, the analysis of large network data of many kinds in different disciplines has given some evidence of common features and dynamics, such as the power law degree distribution, which is a feature of free-scale networks.

Analytical Concepts and Techniques

Social network measures are based on the representation of a set of points (actors) and a set of lines (ties between actors), so that many concepts and theorems of graph theory can be meaningfully applied. Simple graphs can be depicted in a binary symmetric adjacency matrix. Direct graphs take into account the direction of the arcs and use binary asymmetric adjacency matrices. Signed graphs are used to depict like/dislike relations and use matrices with positive or negative signs in cells. To study weighted relations data sets loose the algebraic properties of binary matrices but gain substantial information on the intensity of ties. Staked matrices are used to represent multiplex ties with multigraphs, while the representation of hypergraphs through asymmetric affiliation matrices allows the study of membership relations.

For each of these, ties different networks can be studied, with the families of descriptive statistics illustrated in **Figure 1**.

Available statistics can be related to the relative position of the actor in the network and to the structure of the network itself. As far as the actor's position is concerned, two subcategories have been developed. The first is devoted to a rich choice among centrality indexes according to the range of the centrality, the kind of network, and the type of graph (Wasserman and Faust, 1994). The underlying theoretical hypothesis is that central actors may have greater opportunities to fulfill their goals than peripheral ones. Measures based on the degree indicate the direct contacts of a point. This is a local centrality measure because the overall structure of the network is not taken into account. In directed graphs, centrality in terms of outdegree can be interpreted as a measure of influence, while in terms of indegree, it indicates prestige or popularity. The concept of neighborhood and its application to centrality measures allows the researcher to choose a given range of distance from the actor, according to the substantive meaning that can be given to the structure of relations. Distance-based measures take account of the relative position of the actor in regard to all other points in the network. For information networks, specific centrality measures based on betweenness have been devised to take into account how a point is important for the transmission of information between any other couple of points.

The second category of actor's position has been developed around the above-mentioned concept of structural equivalence, which is focused on the detection of similar positions of points in the network (Breiger et al., 1975; Doreian et al. in Carrington et al., 2005). The underlying hypothesis is that actors with similar positions have the same constraints and opportunities structure, because they are subject to the same role expectations.

Network structure measures can be subdivided into four categories: size, density, centralization measures, and subset detection. While the network's size is equal to the number of nodes, density is the likelihood that a random pair of points in a network will be joined with a line. Research has shown that dense networks are more effective in controlling and sanctioning than in spreading information. In general, they generate strong ties, redundant information, and constraints, and are therefore not suitable for innovation development and diffusion (Burt, 1992). Network centralization statistics deal with the distribution of ties in the overall network and are based on the analysis of variance of point centrality measures. The more the structure of a connected network is hierarchical, the higher the variance in the distribution of any kind of point centrality measure. Unlike density, whose values are not directly comparable for networks of different size, centralization measures have been standardized through the study of expected variance for different null models, given the number of points and lines (Snijders, 1981).

Subset detection is an area that numerous ad hoc criteria have been developed, according to the substantial meaning attributed to networks. The standard concepts are component (the maximal connected simple subgraph), strong and weak components (in directed graphs, respectively, maximal strongly or weakly connected subgraphs), *n*-clique (the maximal complete subgraph at distance *n*), and blocks (the subset of points with similar structural equivalence). If a multiplicity of ties is involved, the nested graphs technique (Stokman et al., 1985) can detect denser clusters in a network.

Network partition techniques have to do with the definition and analysis of affiliation networks and bipartite graphs. The former is presented in a two-mode matrix, whose rows refer to actors or participants or members and columns refer to events or associations. Groups of actors are defined by their common

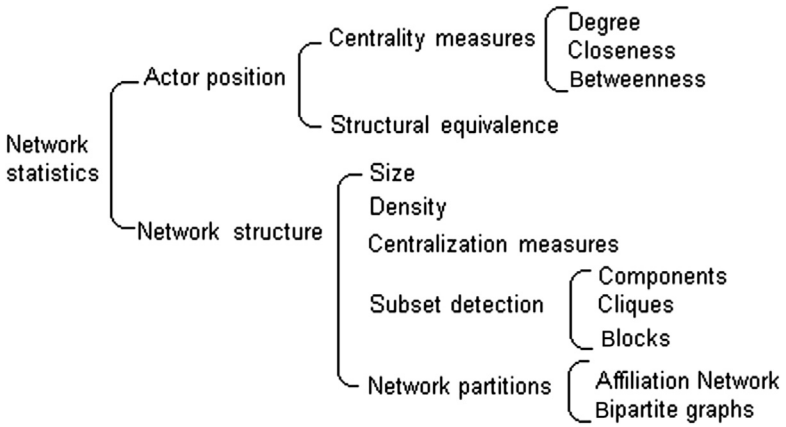

Figure 1 A scheme of different types of descriptive network statistics.

participation or membership to the same events or associations, while events and associations can be clustered according to the proportion of their overlapping participants. Traditionally, this approach is typical of interlocking directorates studies and, more recently, of social movements (Diani and McAdam, 2003). The latter is a graph whose nodes can be divided into two disjoint sets such that every tie connects a node in set 1 with a node in set 2. It is another way to represent two modes networks when substantive considerations suggest treating relations in a symmetrical way, without giving priority to one of the two sets.

Relative Merits, Limits, and Perspectives

NA techniques make it possible to specify empirical indicators of relational concepts, to control field hypotheses through the definition and measurement of traditional catch-all concepts like social structure and cohesion and other more recent ones like social capital. The development of network analytical tools has helped to highlight the structures underlying the apparently inextricable complexity of relations among actors in everyday life. From this point of view, the NA approach has the same analytical purpose as other techniques: for instance, those developed in social mobility studies, whose main goal is to simplify descriptions and to standardize measures to use in comparison. Nevertheless, NA offers a wider range of potential applications than traditional surveys, given that it is hard to conceive a social phenomenon that cannot be analyzed as a set of relations among individual or corporate actors. Moreover, research problems can be addressed at both micro- and macrolevels, without changing the research design and techniques, because corporate actors can simply be interpreted as networks of networks.

Major drawbacks counterbalance these advantages, however. It has often been pointed out that NA's abundance of analytical and technical apparatus contrasts with a somewhat limited scope of theoretical and substantive results, which are mostly descriptive or indeed artifacts generated by simplified simulations. While these critics are perhaps applicable to surveys as well, the relative scarcity of empirical results in NA is largely determined by the specific difficulties of field research. Unlike traditional surveys based on standardized questionnaires on the opinions of anonymous respondents, the collection of relational information involves sampling problems (Frank in Carrington et al., 2005) as well as difficulties relating to the reliability of intrusive questions on personal contacts. This is why network analysts often prefer to collect data from documents and repertoires rather than ask direct questions.

Notwithstanding these difficulties, specific network data collection techniques have been extensively used to test specific hypotheses. Ego-centered networks enable the insertion of network information into traditional surveys, provided that for each interviewee his or her personal network at distance 1 is obtained through a combination of the name generator question and the 'who knows whom' question. This procedure yields important insights into the structure of actor's interpersonal relations, but it cannot give an idea of the entire network and it is difficult to assess the validity of name generator instruments, because criterion data from other sources are unavailable. Moreover, the accuracy of data collection depends more on interviewer effects than in normal survey interviews (Marsden, 2005).

To test the more general societal network structure, so-called 'small world experiments' have been devised to probe the connectedness of large networks and structural holes across age, gender, social class, or ethnic groups. Originally, the research problem was put in this way by Travers and Milgram (1969): given any two people in the United States, chosen randomly, persons A and Z, how many intermediate acquaintance links are needed before A and Z are connected? To trace the chains of acquaintances, folders were used, which produced distributions of acquaintance chains. Despite the success of this approach and its recent adaptations to the specificities and potentials of the Web, the methodological debate is still open about the great proportion of uncompleted chains originated by dropouts.

Available archival data can often overcome the problem of representativeness and completeness, but they often suffer from problems related to their actual relevance and to their interpretation. The already-mentioned literature on interlocking directorates is an instance, together with that on scientific citations. Data on interlocks are highly reliable and complete, but their meaning is still a matter of debate in spite of a long-standing tradition that goes back to the beginning of the last century (Stokman et al., 1985). Data on scientific citations are equally reliable and complete but the actual significance of citations and their emerging network has been a matter of debate since the seminal work by de Solla Price in 1965 until the most recent contributions by Moed (2005) in the field of research evaluation.

Large, complete, and reliable data sets on interpersonal communications and exchanges can be obtained from the World Wide Web. For instance, e-mail interactions can be traced and have been more recently studied in their evolution (Kossinets and Watts, 2006), taking advantage of the availability of software programs for the analysis of very large networks. It has been found that the evolution of these networks is less dependent on spatial and social proximity of actors than physical encounters, but mechanisms of homophily and triadic closure (the tendency of choosing new acquaintances who are friends of friends) are equally present. However, the easy accessibility of these databases – which should always follow privacy protection rules – cannot be interpreted as a shortcut to a more general knowledge of human interaction networks. E-mail exchanges – as well as other interaction media like Short Messaging Service, Facebook, or Twitter – are only one part of the different media people use in personal interactions and further research is needed to assess how these are complementary or alternative to face-to-face encounters.

Since the 1980s, the availability of a complete set of structural indicators has encouraged network comparison in order to explain differences in space and changes over time. Two tendencies are emerging. One concerns the development of statistical models that enable comparison between actual networks and models generated by probabilistic assumptions, in order to test goodness of fit hypotheses (Snijders, 2005). The other implies a systematic combination of relational information with attributes of nodes and ties and has yielded promising

results showing that NA and traditional surveys are not necessarily alternative research techniques, as some 'orthodox' network analysts maintain. The advantage of this integrated approach consists in moving from analytical description to causal explanations, where network statistics can play the role of both independent and dependent variables.

As a matter of fact, in the 1970s and 1980s, structural deterministic approaches have stressed the importance of networks in conditioning actors' behavior, attitudes, and well-being, so that static analysis was required because the network was treated as the independent variable. In the past two decades, the most popular assumption has been that actors actively contribute to the formation of their network by building, strengthening, maintaining, and breaking up relationships. From this perspective, the cross-fertilization of NA with social exchange theory seems potentially fruitful, as well as with game theory.

See also: Boundaries and New Organization Forms; Innovation; Interorganizational Relationships and Networks; Network Paradigm: Applications in Organizational Science; Policy Networks; Social Capital in Organizations; Trust, Sociology of.

Bibliography

Bavelas, A., 1950. Communication patterns in task-oriented groups. Journal of the Acoustical Society of America 22, 271–282.

Breiger, R.I., Boorman, S.A., Arabie, P., 1975. An algorithm for clustering relation data, with applications to social network analysis and comparison with multidimensional scaling. Journal of Mathematical Psychology 12, 328–383.

Burt, R.S., 1992. Structural Holes: The Social Structure of Competition. Harvard Univ. Press, Cambridge, MA.

Carrington, P.J., Scott, J., Wasserman, S. (Eds.), 2005. Models and Methods in Social Network Analysis. Cambridge University Press, Cambridge.

Chiesi, A.M., 1999. L'analisi dei reticoli. Franco Angeli, Milano.

Chiesi, A.M., 2007. Measuring social capital and its effectiveness. European Sociological Review 23, 437–453.

Coleman, J.S., 1988. Social capital in the creation of human capital. American Journal of Sociology 94, 95–120.

Coleman, J.S., Katz, E., Menzel, H., 1957. The diffusion of an innovation among physicians. Sociometry 20, 253–270.

Cook, K.S., Emerson, R.M., Gillmore, M.R., Yamagishi, T., 1983. The distribution of power in exchange networks: theory and experimental results. American Journal of Sociology 89, 275–305.

Degenne, A., Forsé, M., 1994. Les reseaux sociaux: une analyse structurelle en sociologie. Colin, Paris.

Diani, M., McAdam, D., 2003. Social Movements and Networks. Relational Approaches to Collective Action. Oxford University Press, Oxford.

Dufhues, T., Buchenrieder, G., Quoc, H.D., Munkung, N., 2011. Social capital and loan repayment performance in Southeast Asia. Journal of Socio-economics 40, 679–691.

Durlauf, S.N., Fafchamps, M., 2005. Social capital. In: Aghion, P., Durlauf, S.N. (Eds.), Handbook of Economic Growth, 1. Elsevier, Amsterdam, pp. 1639–1699. Part B.

Epstein, A.L., 1969. Gossip, norms and social network. In: Mitchell, J.C. (Ed.), Social Networks in Urban Situations. Manchester University Press, Machester, pp. 117–127.

Flap, H.D., 1999. Creation and returns of social capital: a new research program. La Revue Tocqueville 20, 1–22.

Flap, H.D., Völker, B.G.M., 2004. Creation and Returns of Social Capital. Theory, Research and Measurement. Routledge, London.

Freeman, L.C., 1978. Centrality in social networks: conceptual clarification. Social Networks 1, 215–239.

Granovetter, M., 1973. The strength of weak ties. American Journal of Sociology 78, 1360–1380.

Grieco, M., 1987. Keeping It in the Family: Social Network and Employment Chance. Tavistock, London.

Kadushin, C., 1982. Socail density and Mental Health. In: Marsden, P.V., Lin, N. (Eds.), Social Structure and Network Analysis. London, Sage.

Kossinets, G., Watts, D.J., 2006. Empirical analysis of an evolving social network. Science 311, 88–90.

Lin, N., 2001. Social Capital. A Theory of Social Structure and Action. Cambridge Univ. Press, Cambridge.

Lin, N., 2004. Job search in urban China: gender network chains and embedded resources. In: Flap, H., Völker, B. (Eds.), Creation and Returns of Social Capital. Routledge, London, pp. 145–177.

Marsden, P.V., 2005. Recent developments in network measurement. In: Carrington, P.J., Scott, J., Wasserman, S. (Eds.), Models and Methods in Social Network Analysis. Cambridge University Press, Cambridge, pp. 8–30.

Marsden, P.V., Lin, N. (Eds.), 1982. Social Structure and Network Analysis. Sage Publications, Beverly Hills, London; New Delhi.

Moed, H.F., 2005. Citation Analysis in Research Evaluation. Springer, New York.

Newman, M., Barabasi, A.-L., Watts, D.J., 2006. The Structure and Dynamics of Networks. Princeton University Press, Princeton and Oxford.

Padget, J.F., Ansell, C.K., 1993. Robust action and the rise of the Medici, 1400–1434. The American Journal of Sociology 98, 1259–1319.

Scott, J., 1991. Social Network Analysis. A Handbook. Sage Publications, London.

Snijders, T.A.B., 1981. The degree variance: an index of graph heterogeneity. Social Networks 3, 163–174.

Snijders, T.A.B., 1996. Stochastic actor-oriented models for network change. Journal of Mathematical Sociology 31, 149–172.

Snijders, T.A.B., 2005. Models for longitudinal network data. In: Carrington, P.J., Scott, J., Wasserman, S. (Eds.), Models and Methods in Social Network Analysis. Cambridge University Press, Cambridge, pp. 215–247.

Stokman, F.N., Ziegler, R., Scott, J. (Eds.), 1985. Networks of Corporate Power. Polity Press, Cambridge.

Travers, J., Milgram, S., 1969. An experimental study of the small world problem. Sociometry 32, 425–433.

Wasserman, S., Faust, K., 1994. Social Network Analysis. Methods and Applications. Cambridge University Press, Cambridge.

White, H., 1981. Where do markets come from? American Journal of Sociology 87, 517–547.

Network Analysis, History of

Philipp Korom, Max Planck Institute for the Study of Societies, Cologne, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

While social scientists had been drawing on the abstract idea of social networks clearly since the nineteenth century, which becomes most evident in Georg Simmel's formal sociology, a social network analysis (SNA) grounded in computational models and graphic imagery emerged within the field of small group research in the 1930s. It was Jacob Moreno who introduced the idea of depicting social structure as a network diagram ('sociometry'). Kurt Lewin was an early contributor to the promotion of mathematical models of group relations, and Fritz Heider focused on triads to theorize on what throws groups out of balance. Mostly independent of these ideas, the anthropologist Lloyd Warner adopted a network approach in the study of informal relations between workers and of communities. SNA was further applied by the Manchester school of anthropologists to enhance ethnographic description. Advancing mathematical-formal aspects of SNA at Harvard in the 1970s, Harrison C. White and his collaborators contributed to the establishment of the discipline as a recognized paradigm. In the late 1990s, physicists began to publish work on social networks. Today, SNA has become a multidisciplinary research specialty with distinct theoretical concepts and data-analytic techniques.

The Origins of Social Network Ideas

Social structure has been one of the early key concepts in the social sciences. Social Network Analysis (SNA) is a recently developed set of formal methods for the study of social structures that draws on graph theory in which individuals and other social actors, such as groups and organizations, are represented by points and their social relations are represented by lines. The main theoretical underpinning of SNA is, as Wellman (1988) pointed out, that the structure of relations among actors and the location of individual actors in the network have important attitudinal, behavioral, and perceptual consequences both for individuals and the social structure as a whole. Visual imagery has played a significant role in SNA since its inception. Mathematical and computational models are at the base of more current applications.

Only in the first part of the twentieth century did a handful of social scientists begin to systematically theorize social relationships. Society as a whole was conceptualized as the very tissue of relations. Ferdinand Tönnies (1855–1936) categorized social ties as being either personal and direct ('community') or formal and instrumental ('society'). Herbert Spencer (1820–1903) made a similar distinction between premodern and modern societies by referring to 'ordinary' and 'secondary relations.' Emile Durkheim (1858–1917) theorized the specific structuring of groups as being responsible for the quality of emergent laws and morality, and Gustav Le Bon (1841–1931) was the first to examine the phenomenon of crowd behavior. It was, however, Georg Simmel (1858–1918), a German sociologist whose work stood out against macrolevel theories of scholars such as Max Weber (1864–1920) and Karl Marx (1818–83), who pioneered most explicitly the analysis of dyads (relationships between two persons) and triads (groups composed of three people) as building blocks of social life. Although he never used the term 'social network' as such, his ideas about microlevel structures prove to be a source of inspiration even for current day SNA. Simmel held the view that sociology was no more and no less than the study of interweaving actions in social encounters (Simmel, 1908/2009). He suggested that a consideration of social dynamics caused by the simple addition of a third person could provide insights on society at large, that is, how large structures constrain individuals. While isolated dyads are characterized by individuality and intimacy, triads have a superindividual property: (informal) social pressures are activated, and the variance of behavior as well as interpersonal idiosyncrasies are reduced. Third-party effects may also involve two actors forming a coalition against a third, one actor disturbing the alliance between two others, or one actor even taking advantage of a conflict between two others ('rejoicing third'). It was Simmel's student Leopold von Wiese (1876–1969) who adopted a contemporary terminology of points, lines, and connections to describe social relations (von Wiese, 1924/1932).

The origins of structural research are, however, also located outside the disciplinary boundaries of sociology. Especially the work of scholars in educational and developmental psychology, who were interested in the ways small-group structures affected individual perceptions and actions had a structuralist flavor very early. Helen Bott, for example, set out to document every form of social interaction that occurred among preschool children (Bott, 1928). In fact, she was one of the first to collect ego-centered kinship network data and calculate even network density measures. It is, however, commonly agreed that Jacob Moreno, a student of psychiatry from Vienna who immigrated to the United States in 1925 and championed the field of 'sociometry,' was the main driving force, together with his collaborator, Helen Jennings, in establishing SNA.

Moreno was deeply influenced by Gestalt psychology (Gestalt translates as 'form') developed by mostly German psychologists such as Kurt Koffka (1886–1941), Max Wertheimer (1880–1943), and Wolfgang Köhler (1887–1967) as a protest against behaviorist theories of their day. This school of thought maintained that innate and self-organizing operations of the brain influence the way we see. With respect to the visual recognition of figures, we tend to perceive, for example, whole forms instead of just a collection of lines and curves. Moreno transferred the idea that 'the whole is greater than the sum of its

parts' to the interplay between the social embeddedness of an individual and this person's well-being. He believed that many psychological problems stemmed from failed interactions and that the position of individuals within groups is highly significant for their mental health; Moreno and Jennings (1938) thus spoke of 'psychosocial networks.' Moreno's concern with social structure is mirrored in therapy forms called 'sociodrama' and 'psychodrama,' which aimed at exploring conflicts inherent in social roles. 'Sociometry' was developed as a powerful tool for assessing group dynamics and eliciting graphically subjective feelings between persons. **Figure 1** illustrates typical 'sociograms' featured in the book *Who Shall Survive?* (Moreno, 1934), in which Moreno (and Jennings) probed the causes of runaways at the Hudson School for Girls in upstate New York. The crux of their argument was that the location in the social networks primarily determined whether and when someone ran away.

At first, Moreno's idea to make social structure tangible garnered a great deal of interest, although this turned out to be short-lived. By the 1940s, American social scientists had returned to their focus on the characteristics of individuals (Freeman, 2004).

In the 1940s and 1950s, research on social networks advanced on more than one front, mostly independently of each other. In fact, SNA can be seen as emerging from certain disciplinary trajectories (Prell, 2012), even if scholars working in the field of SNA always crossed disciplines. In the following, the developmental pathways of SNA are outlined in the key disciplines of psychology, sociology, and social anthropology.

Disciplinary Trajectories

(Social) Psychology Trajectory

Kurt Lewin, a German-Jewish émigré and a contemporary of Freud, who after arriving in the United States first held an academic position in Iowa, provided an important intellectual foundation for the development of SNA. His most cited work is *Field Theory in Social Science* (Lewin, 1951), in which he puts forward the idea that behavior is embedded in a psychophysical field consisting of 'valances,' which are analogous in their dynamics to gravitational forces: they pull one toward or push one away. Lewin's field theory was supposed to provide a method of analyzing causal relations among mutual interdependent facts in a concrete situation as perceived by the individual. Lewin tried to formalize his topological concepts: vectors describe the resolution of forces, the lengths represent their strengths, and arrows indicate the direction of field forces. The 'life space,' which contains a person and this person's environment, was divided into regions (family, work, etc.), paths, and barriers. In 1945, Lewin left Iowa to form the Research Center for Group Dynamics at Massachusetts Institute of Technology (MIT). After his sudden death in 1947, most of his coresearchers moved to the University of Michigan ('the Michigan group') and continued to make important contributions to SNA.

Dorwin Cartwright and Frank Harary (1956) extended Fritz Heider's social balance theory (Heider, 1946) which discusses the relations among individuals based on sentiments. In the case of triads, Heider postulated that a balanced state between the three (focal individual P, another agent O, object X) can only exist, if the algebraic multiplication of signs in the triad relation has a positive value. With three sets of possible relationships, each taking one of two values (+, −), eight possible states exist. A triad is harmonious, for example, if there are two negative and one positive relation (e.g., −, +, −): "I don't like person B. B has a dog. I don't like the dog either." One of Heider's propositions is that individuals tend to choose states of balance in their interpersonal relations, which is why the adages 'your friends are my friends' or 'the enemy of my enemy is my friend' adequately summarize core implications of social balance theory. Building on Lewin's work, Cartwright and Harrary translated Heider's diagrams into graph theory and covered more than triadic relationships by asking the following question: What pattern would the interpersonal relations of a group of individuals have if they were balanced? Important among their findings is the 'structure theorem' that "an s-graph [signed graph] is balanced if and only if its points can be separated into two mutually exclusive subsets such that each positive line joins two points of the same subset and each negative line joins points from different subsets" (Cartwright and Harary, 1956: p. 286; **Figure 2**). Subsequently, Davis (1967) generalized balance to clusterability, which allows any number of clusters such that positive arcs appear within clusters and negative arcs between clusters. Davis and Leinhardt (1972) further extended balance theory by including status, which was motivated by the fact that people not only cluster into groups but also adhere to social rankings. They show how directed sentiment relations could generate a structure that is incorporated in a system of hierarchically arranged cliques. These classic papers gave the initial spark for the development of many techniques for identifying cohesive subgroups. One of the most innovative approaches to microstructures was the conceptualization of dyads as a stochastic process in which distribution can be statistically modeled (Holland and Leinhardt, 1977).

In the 1950s and 1960s, Alex Bavelas, another student of Lewin, conducted experiments at the MIT together with colleagues such as Harold Leavitt. He was especially concerned with communication in small groups and hypothesized a relationship between structural centrality and information diffusion processes. In Bavelas (1950), the hierarchy steepness within five-person groups working on a joint task was moderated by installing either a 'wheel' structure, in which all messages could only flow between peripheral members by passing a central actor, or decentralized configurations without limited communication channels. A tentative finding was that centralization was beneficial in the case of simple tasks and detrimental for complex ones. Bavelas was the first to develop a centrality measure, which received much attention and provided an impetus for the refinement of mathematical concepts regarding network structure. At MIT Leon Festinger, R. Duncan Luce, and Albert Perry especially strived for a mathematical formalization of SNA. In a seminal article, Luce and Perry (1949) used graph-theoretic terminology to define cliques as maximal complete subgraphs. What is meant by 'complete' is that every individual is tied directly to every other. Parallel to this, Festinger (1949), who was primarily concerned with the development of group standards, applied somewhat different matrix algebra to identify cliques. In either case, the

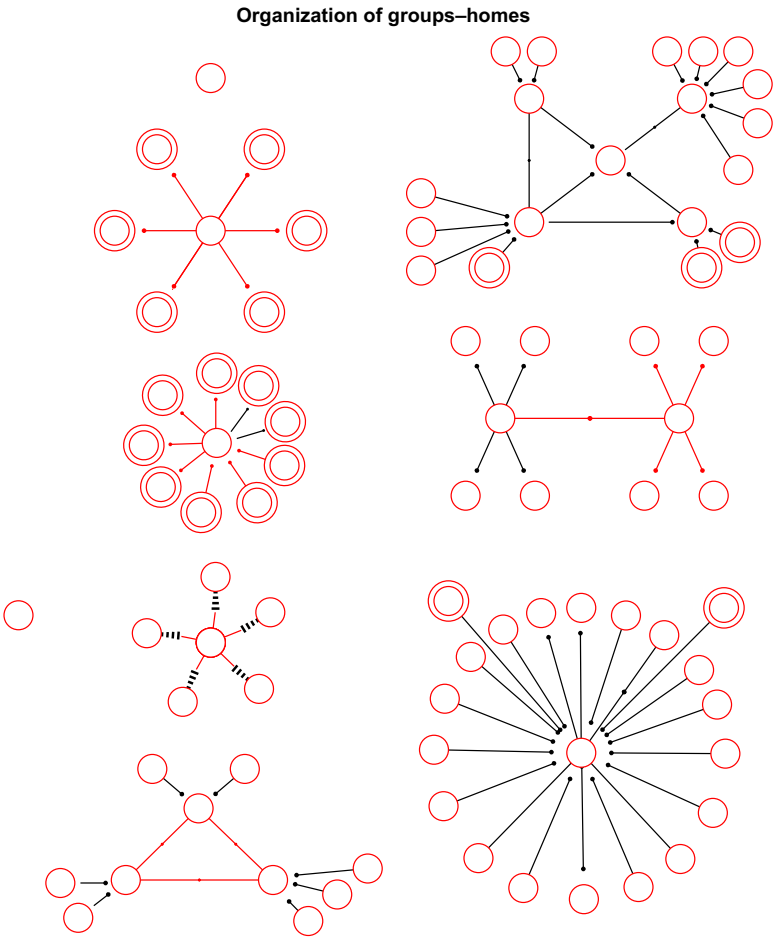

Attractions and repulsions take the form of isolation.

Isolation. Subject is attracted to six individuals outside of his group (outside individuals are symbolized by a double circle) who do not reciprocate.

Isolation. Subject is attracted to four individuals outside of his group and rejects two more; they do not reciprocate; three others who are attracted to him he does not reciprocate.

Isolation. Subject is attracted to five individuals within his group; they respond with indifference.

Mutual attractions between three individuals take the form of a triangle but each of the subjects is otherwise rejected and isolated within his own group; the result is an isolated and rejected triangle of persons.

Five subjects each isolated and rejected within his own group reject and isolate each other.

Two subjects each otherwise isolated in his own group form a pair of mutual attraction; the result is an isolated pair.

Subject rejects six and is rejected by fifteen individuals within his own group; is rejected further by two individuals outside of his own group. The result is an isolated and rejected individual.

Figure 1 Example of sociograms. Source: Moreno, J.L., 1934. Who Shall Survive? Nervous and Mental Disease Publishing Company, Washington, DC, p. 116.

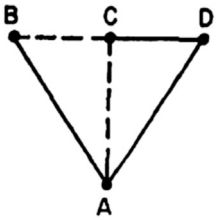

A signed graph of four points and five lines. Solid lines have a positive sign and dashed lines a negative sign. If the points stand for people and the lines indicate the existence of a liking relationship, this s-graph shows that *A* and *B* have a relationship of liking, *A* and *C* have one of disliking, and *B* and *D* have a relationship of indifference (neither liking or disliking).

Figure 2 Structural balance. Source: Cartwright, D., Harary, F, 1956. Structural Balance: A Generalization of Heider's Theory. Psychological Review 63, 283.

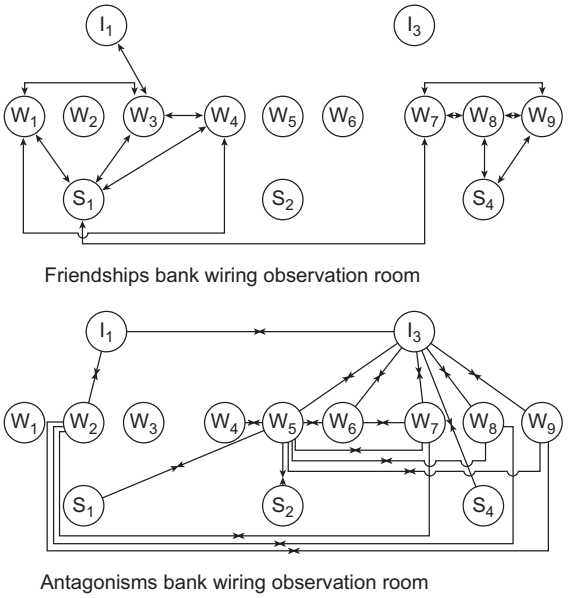

Figure 3 Patterns of friendship in the bank wiring observation room. Source: Roethlisberger, F.J., Dickson, W. J., 2003. Management and the Worker. Routledge, London, p. 367 (originally published in 1939).

implicit assumption motivating clique analysis was that actors who maintain cohesive bonds were prone to act similarly. Subsequently, Cartwright and Harrary proposed alternative ideas for clique detection, such as the maximal strong component idea, which implied loosening the clique concept by allowing for indirect links of various lengths. Thus a maximal strong component is a network subgroup in which each actor can reach other actors directly or indirectly, and no further actor can be added without reducing the mutual reachability. At the time these formalizations of SNA were put forward, the development of large computers had already commenced, although computational methods for findings cliques were still rare.

The approach of Bavelas and his colleagues at MIT spurred innovative approaches to studying the social fabric in the post-World War II era. Ithiel de Sola Pool, a political scientist, and Manfred Kochen, a mathematician at IBM, drew up mathematical models for tackling the vexing 'small-world problem': If two persons are selected at random from a population, how many acquaintances are needed to create a chain between them? While de Sola Pool and Kochen turned to mathematical estimations because of a lack of data, the social psychologist Stanley Milgram came up with an ingenious experiment. Randomly chosen 'starters' in Nebraska were given a letter to deliver to 'target persons' in Massachusetts by sending it on to acquaintances whom they thought to be closer to the target. The idea was to test empirically how many intermediaries were needed to reach someone. Milgram's initial results (Milgram, 1967) led to the now popular notion of 'six degrees of separation,' which touched off a storm of further work across numerous disciplines. Replications, however, provided rather mixed findings on the validity of the experimental setting.

Sociology Trajectory

Although only collaborating for 6 years in the Anthropology Department and the Business School at Harvard, the anthropologist W. Lloyd Warner and the psychologist Elton Mayo produced work that was to become a historical landmark. Industrial psychologists relied on Warner and Mayo when faced with inconsistent effects of the physical work environment on worker productivity. The two men had discovered that improving lighting conditions boosted worker morale, but restoring lighting to its original state did not lead to deterioration in productivity. The Harvard team turned its initial psychological focus to the social structure that was systematically observed in a small group of workers putting together telephone switching equipment in the 'bank wiring room' of Western Electric Corporation in Cicero, Illinois ('Hawthorne Works'). This field study, described in great detail by Roethlisberger and Dickson (1939/2003), was the first to use sociograms to describe formal and informal relations between workers and is today among the classic readings in sociology (see **Figure 3**).

Observations revealed that workers were more responsive to social forces within their (informal) peer groups than to physical incentives. There was, however, no attempt to use graphs to identify sociometrically defined 'cliques.' John Scott (2013: p. 22) even concludes that the authors "appear to lack any theoretical understanding of how social networks might shape the behaviour of individuals" (Scott, 2013: p. 22).

Later on, Lloyd Warner (Warner and Lunt, 1941) helped disenchant the American myth of a classless society by detecting social divisions and rank orders of social participation in Newburyport, Massachusetts ('Yankee City'). Recognizing that the research site was ill-suited to shed light on racial demarcation in social stratification, Warner and his team embarked in further studies on an anthropological investigation of the 'Deep South' (Davis et al., 1941). Unlike in the Hawthorne study, the authors used matrices extensively to study the structural embedding of actors. Generating a data set that captured the

participation of 18 women in 14 events enabled them, for example, to identify cliques and social status among small groups. One of their findings was that social cliques (in colored societies) are composed of different layers: a 'core' of actors in the center of nearly all clique gatherings, a 'primary circle' of those who jointly participate with core members but do not stick together within one subgroup, and a 'secondary circle' of less-intimate friends who rarely participate.

This field research in the precomputer age emerged in parallel with advances in sociometry concerned especially with relational patterns in small groups. Harvard professor George Homans was the first to systematically review and synthesize the bulk of experimental and observational studies dealing with the subject. His path-breaking book *The Human Group* (Homans, 1950) is packed with much-cited conceptual schemes, such as the distinction between what he called an 'external system' (structure that prescribes human interaction, e.g., kinship) and an 'internal system' (processes within groups). Internal structural buildup is generated through interactions based on feeling or participations in common activities, whether or not this is dictated by the external system. Activities and common sentiments would generally lead to clique formation and instill social solidarity from within. Homans also reanalyzed data from the 'Hawthorne' and 'Deep South' studies. By rearranging the rows and columns of *women x events* matrix used by Davis et al. (1941), he aimed at detecting the existence of cliques – a method that is analogous to what has subsequently come to be called blockmodeling. However, he was far from advancing mathematical thinking in SNA.

A further strand of structural thinking within sociology developed in the 1950s at Columbia University, where the lifelong collaborators Paul Lazarsfeld and Robert K. Merton supervised many students, such as Peter Blau, James Coleman, and Charles Kadushin, who advanced the standards of modern SNA. Blau, an exponent of social exchange theory, refined the notion of homophily in SNA and established empirically that multigroup memberships 'crosscutting social circles' promote intergroup relations. Coleman established social capital, the resources inherent in the relations between and among actors, as a key concept in SNA. In a pioneering study on diffusion in social networks, he showed, together with Elihu Katz and Herbert Menzel, how doctors' willingness to prescribe new antibiotics spread through professional contacts. Building on Simmel's notion of social circle, Kadushin explained people's interest in consulting a psychotherapist not by individual influences but rather by memberships in intersecting social groups ('knowing others in therapy, knowing others with similar problems, asking friends for a referral'). Lazarsfeld himself applied newly improved survey technologies to investigate the influence of mass media. One of his main findings was that voters were not directly influenced by media during election time but rather by so-called opinion leaders in medium-sized communities who pass on media information through interpersonal communication (two-step flow of communication, Lazarsfeld et al., 1944).

Social Anthropology Trajectory

A further key contribution to the development of SNA was the foundation of the Department of Social Anthropology and Sociology in the 1950s at the University of Manchester, headed by Max Gluckman. In Manchester, a dynamic group of researchers, including Elisabeth Bott (also affiliated with the London School of Economics), John Barnes, J. Clyde Mitchell, and Siegfried Nadel, popularized the application of social network methods in ethnography. The Manchester school was deeply influenced by Alfred R. Radcliffe-Brown's structural perspective, which they sought to take in a novel direction. Radcliffe-Brown (1940) argued that anthropologists should investigate social phenomena by using methods similar to those in physics, which essentially meant relying on mathematics and analyzing structures such as kinship in terms of their function for maintaining societies across the globe. While Gluckman and his collaborators dismissed functional ideas of social equilibrium and acknowledged conflict as an inherent part of societies, they nevertheless refined the idea that social relations are embedded in wider networks.

In the decade after World War II, these anthropologists mostly carried out fieldwork in what was then British Central Africa (now Zambia, Zimbabwe, and Malawi), which gave theoretical force to such concepts as intercalary roles or crosscutting ties. Gluckman, for example, developed the intercalary theme when analyzing the authority roles of village headmen in rural societies, who were subject to conflicting pressures from fellow villagers and political superiors. At the conceptual level, societies were no longer seen as being monolithic, but rather as a pattern of relationships "obtaining between actors in their capacity of playing roles relative to one another" (Nadel, 1957: p. 12). The initial stimuli to direct study toward social networks were provided especially by Barnes' paper and Bott's book (Barnes, 1954; Bott, 1957). When studying a Norwegian fishing village, Barnes neglected the traditional anthropological concepts of geography, politics, and economics and focused instead on primordial relations such as kinship, friendship, and neighborhood ties, which combine to constitute what he called a 'network.' He is often regarded as the first person to have conceptualized the term, which he imagined as consisting 'of a set of points some of which are joined by lines' (Barnes, 1954: p. 237). Bott (1957) applied this concept to the study of the conjugal roles in families from London. She argued that the degree of segregation in the role-relationship is a function of the configuration of the network of friends and relatives associated of each spouse. For example, if husbands and wives are embedded in joint networks, the members of these external social networks can develop norm consensus and influence how equally spouses interact and share domestic household tasks. She was one of the first to theorize that dense and multiplex networks of ties have a strong norm-enforcement effect on network members, and she made use of sociograms to illustrate differences between dispersed and connected family networks (**Figure 4**).

Mitchell, who continued the Manchester tradition of working in British colonial Africa, made important contributions to the formalization of SNA. In a widely cited introduction (Mitchell, 1969), he suggested a series of conventions, still valid today, on how to model social networks using graph theory and distinguished between total and partial networks. The total network of society is composed of "the general ever-ramifying, ever-reticulating set of linkages that stretches within and beyond the confines of any community or

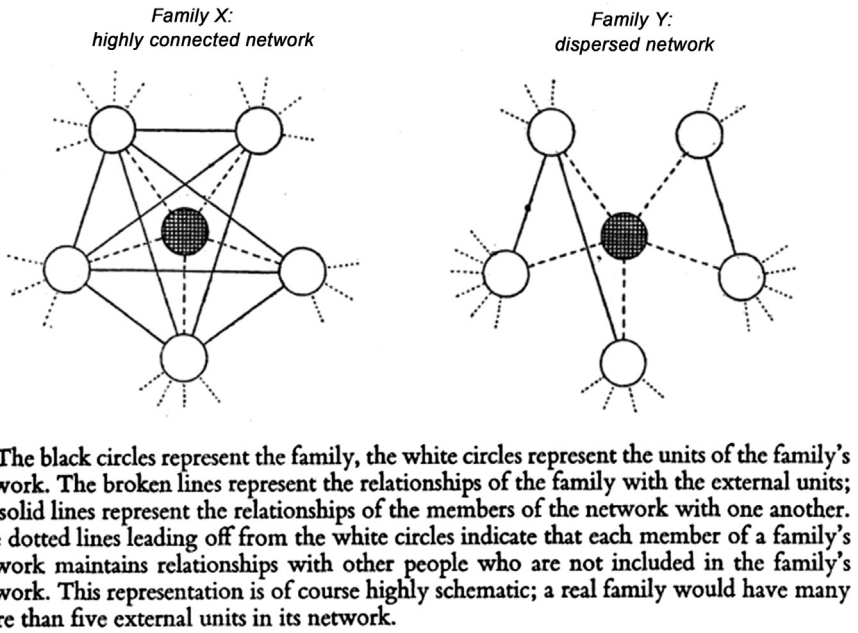

Figure 4 Schematic comparison of the networks of two families. Source: Bott, E., 1955. Urban families: conjugal roles and social networks. Human Relations 8, 348.

organization" (Mitchell, 1969: p. 12). In actual research, it is therefore imperative to select particular aspects of total networks, which means either focusing on ego-centered networks around particular individuals or on a particular aspect of social activity, such as friendship.

Scott (2013) argues that the reason the breakthrough leading to the study everywhere of social networks in all fields of social life was not made in Britain was the great emphasis on the informal side of communications in mostly ego-centered networks. However, the knowledge of the British anthropologists diffused quickly across the Atlantic and was quite commonly incorporated into a more sociological analysis of social structure, because British empiricism fit well with the American penchant for quantitative measurement and statistical analysis (Wellman, 1988: p. 23).

The Harvard Breakthrough

In the 1970s, the center of gravity in network research shifted to sociology. By the time that Mitchell (1969) published his timely review article, a group of American researchers led by Harrison White had begun to revolutionize the formal methodology of SNA. Reflecting his mixed background in mathematics, physics, and sociology, White's teaching and research at Harvard revolved around algebraic models of social structure. A key advance was the CONCOR algorithm, developed by White and his students, to identify a social network's structurally equivalent actors. An equivalence strategy (Lorrain and White, 1971) clusters individuals on the basis of relational similarity, that is, individuals are said to be 'structurally equivalent,' if they are characterized by identical patterns of relationship with other network members (but do not necessarily have mutual awareness of one another). Edward O. Laumann, White's first Harvard PhD student, was curious about the way people form, maintain, and dissolve relationships with others and about the impact of such networks on identities and behavior. Consequently, he pioneered multidimensional scaling (MDS) in community studies. Laumann and Louis Guttman (1966) asked 422 male residents of Cambridge and Belmont, Massachusetts, about the occupations of their seven alters (interalia the three closest friends). Then, an early form of MDS was applied to a table featuring the respondents' occupations in rows and the alters' occupations in columns. The most fitting solution appeared to be three-dimensional, the first dimension being interpreted as indicating occupational prestige. Two other students, Nancy H. Lee and Mark Granovetter, chose a less algebraic social network approach to the studies of illegal and labor markets. Lee figured out the condition under which women searching for an (illegal) abortion successfully activate social networks to gain indirect contacts with others who might help. Granovetter made the point that the economist's idea of a 'perfect labor market' misses the fact that labor-market information is often transmitted as a byproduct of other social processes. What matters are contacts. His idea that weak ties (corresponding to acquaintances) are superior to strong ties (corresponding to friends) for providing support in getting a job, inspired a vast research program on the role of social networks in the labor market (Granovetter, 1973). It also sparked the interest of sociologists to study the social embeddedness of economic phenomena in general, as, for example, the consequences of network positions of firms for their performance. The Harvard setting proved to be stimulating to many other researchers, among them Peter Bearman, Phillip Bonacich, Kathleen M. Carley, Joel Levine, and Barry

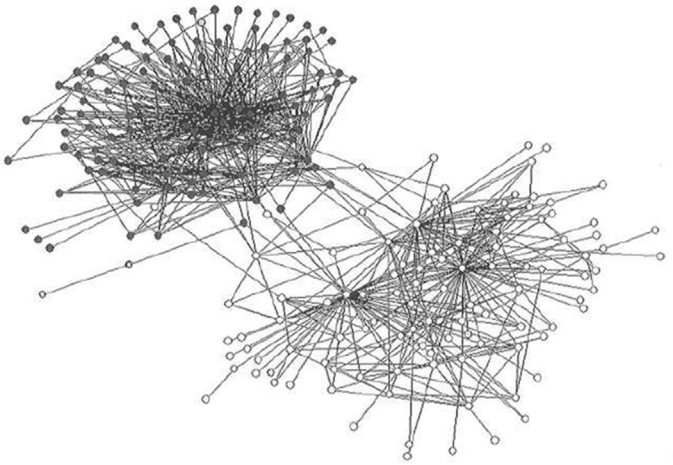

Figure 5 Small-world publications, 1950–2004; physicists are displayed as black points. Source: Freeman, L.C., 2011. The development of social network analysis. In: Scott, J., Carrington, P.J. (Eds.), The SAGE Handbook of Social Network Analysis. Sage, Los Angeles, p. 30.

Wellman, who remained at the forefront of the field. It was this diverse group of US-American researchers that established SNA as a formal method in social science research and influenced work across the globe.

Over a long period, SNA ideas have more or less evolved in the different fields referred to above. The late 1970s, however, witnessed the emergence of a unified research field. Mathematics, the dominant force in the field, helped advance commonly understood and shared terms. Computer programs, above all UCINET developed by Lon Freeman, Martin Everett, and Steve Borgatti, standardized the analysis of social network data. Annual conferences facilitated the communication between different schools, and the foundation of the International Network of Social Network Analysis (INSNA) and its newsletter *Connections* completed the integration of the field. Today different academic journals, such as *Social Networks*, provide an institutionalized forum for representatives of different disciplines, ranging from anthropology to communication science.

The Application of SNA across Disciplines

Having become a standard methodology, SNA experienced changes, the most striking development being the growth of interest apparent among physicists in applying network ideas to social phenomena. First, Watts and Strogatz (1998) published on small worlds. Within the social sciences, small worlds were first studied in depth by Milgram, whose experiments on forwarding letters (see above) established theories on 'six degrees of separation.' Watts formalized the features of small-world networks, in which most nodes are not neighbors but can be reached from every other node by a small number of steps, as the coincidence of high local clustering and short global separation. Watts and Strogatz (1998) show small worlds to be widespread not only in biological and man-made networks, but also in social ones. Barabási (2002) studied the distribution of network connections that grew because new nodes were added; to do this, he used different examples such as sites in the World Wide Web or links between screen actors. He discovered networks not to be random but to contain hubs – nodes with a very high number of links. In such networks, the distribution of linkages follows a power law in that most nodes have but a few connections while others have a huge number of links. In that sense, the network has no 'scale.' Scale-free networks prove to be widespread in the social world and seem to be robust against accidental failures.

Especially the work of Watts and Strogatz steered much interest in network analysis within the physics community. Within few years physicists produced more articles on small worlds than the social network community had turned out in almost a half century. However, Freeman's analysis of citations clearly shows that physicists rarely cite work of social network analysts, and social scientists are reluctant to engage with the models of physicists (**Figure 5**).

Many other disciplines have a heightened interest in SNA. Political scientists use the small-world concept to study information cascades, map out terrorist networks, study advocacy networks, or the influence of social relationships on legislative behavior. Epidemiology is turning toward SNA to test contagion theories on phenomena as diverse as obesity, smoking, or happiness. Communication science uses (semantic) network analysis to examine differences between blog posts and professional journalism or the structure of international news flows. Computer science researchers study changes in trust, migration, and mobility in online network-like communities. Witnessing decades of explosive growth, SNA academic journals such as *Network Science* and research centers emerged that aim at bridging and bonding disciplinary bonds. More than ever SNA has become a multidisciplinary endeavor.

Conclusion

The roots of SNA are diverse. Network theorizing in the social sciences dates back to the work of Georg Simmel, from the last decade of the twentieth century (Simmel, 1908/2009).

Moreno's network visualizations that displayed up to 435 individuals (Moreno, 1934) marked the beginning of sociometry, the precursor of SNA. The approach to represent social relations by graphs of points and lines was especially taken up by three disciplines: Social psychology, sociology, and social anthropology. Heider's balance theory that aimed at describing 'equilibrium' properties of interpersonal relations inspired the work of Cartwright and Harary, who studied more complex cases of structural balancing (i.e., groups with more than three elements). Alex Bavelas examined the impact of different communication network structures on the ability of groups to solve problems. In general, early social psychological work mostly focused on (positive and negative) interpersonal relations within smaller groups and their implication for behavior in social contexts. The network approach to social structures such as status systems in typical American communities (Warner and Lunt, 1941) provided insights that were more amenable to sociology. Social anthropology was primarily concerned with kinship networks. While most of the pioneering studies in SNA transcended disciplinary boundaries only to a limited extent, the pocketing of social network concepts and theories within single disciplines is today seen less. Especially, the newly established field of 'network science' avoids compartmentalization, yet consolidating a common language and methodological toolbox (Vedres and Scotti, 2012).

While SNA is experiencing rapid growth in participants, controversy surrounds the issue whether SNA is 'just' a methodology or also a theory with a set of clearly defined propositions. Borgatti and Halgin (2011) identify two models that underlie network theorizing: the network flow model and the bond model. The first being identified, for example, with the strength of weak ties theory (Granovetter, 1973) that predicts how information flows through a given network and generates outcomes for the nodes. The second being best represented by theories on power in exchange network, which underlying theme is how network ties align and coordinate action in order to enable many nodes to act as a single node with greater capabilities. What is rather uncontroversial is that SNA comes with a set of powerful heuristics that help to analyze as diverse phenomena as corruption or the spread of diseases.

See also: Simmel, Georg (1858–1918); Social Capital; Social Network Analysis; Social Networks.

Bibliography

Barabási, A.-L., 2002. Linked. The New Science of Networks. Perseus Publishing, Cambridge, MA.

Barnes, J.A., 1954. Class and committees in a Norwegian Island Parish. Human Relations 7, 39–58.

Bavelas, A., 1950. Communication patterns in task-oriented groups. The Journal of the Acoustical Society of America 22, 725–730.

Borgatti, S., Halgin, D.S., 2011. On network theory. Organization Science 22, 1168–1181.

Bott, E., 1957. Family and Social Network. Tavistock Publications, London.

Bott, H., 1928. Observations of play activities in a nursery school. Genetic Psychology Monographs 4, 44–88.

Cartwright, D., Harary, F., 1956. Structural balance: a generalization of Heider's theory. Psychological Review 63, 277–293.

Davis, J.A., 1967. Clustering and structural balance in graphs. Human Relations 20, 181–187.

Davis, A., Gardner, B.B., Gardner, M.R., 1941. Deep South. University of Chicago Press, Chicago.

Davis, J.A., Leinhardt, S., 1972. The structure of positive interpersonal relations in small groups. In: Berger, J., Zelditch Jr., M., Anderson, B. (Eds.), Sociological Theories in Progress, Vol. 2. Houghton Mifflin, Boston, pp. 218–251.

Festinger, L., 1949. The analysis of sociograms using matrix algebra. Human Relations 2, 153–158.

Freeman, L.C., 2004. The Development of Social Network Analysis. A Study in the Sociology of Science. Empirical Press, Vancouver.

Granovetter, M., 1973. The strength of weak ties. American Journal of Sociology 78, 1360–1380.

Heider, F., 1946. Attitudes and cognitive organization. The Journal of Psychology: Interdisciplinary and Applied 21, 107–112.

Holland, P.W., Leinhardt, S., 1977. A dynamic model for social networks. Journal of Mathematical Sociology 5, 5–20.

Homans, G.C., 1950. The Human Group. Harcourt, Brace & World, New York.

Laumann, E.O., Guttman, L., 1966. The relative associational contiguity of occupations in an urban setting. American Sociological Review 31, 169–178.

Lazarsfeld, Paul F., Berelson, Bernard, Gaudet, Hazel, 1944. The People's Choice. Columbia University Press, New York/London.

Lewin, K., 1951. Field Theory in Social Science. Selected Theoretical Papers. Harper, New York.

Lorrain, F., White, H.C., 1971. Structural equivalence of individuals in social networks. Journal of Mathematical Sociology 1, 49–80.

Luce, R.D., Perry, A.D., 1949. A method of matrix analysis of group structure. Psychometrika 14, 95–116.

Milgram, S., 1967. The small-world problem. Psychology Today 1, 61–67.

Mitchell, C.J., 1969. The concept and use of social networks. In: Mitchell, C.J. (Ed.), Social Networks in Urban Situations. Manchester University Press, Manchester, pp. 1–50.

Moreno, J.L., 1934. Who Shall Survive? Nervous and Mental Disease Publishing Company, Washington, DC.

Moreno, J.L., Jennings, H.H., 1938. Statistics of social configurations. Sociometry 1, 342–374.

Nadel, S.F., 1957. The Theory of Social Structure. Cohen and West, London.

Prell, C., 2012. Social Network Analysis. Sage Publications, Los Angeles/London.

Radcliffe-Brown, A.R., 1940. On social structure. The Journal of the Royal Anthropological Institute of Great Britain and Ireland 70, 1–12.

Roethlisberger, F.J., Dickson, W.J., 2003. Management and the Worker. Routledge, London (Originally published in 1939).

Scott, J., 2013. Social Network Analysis, third ed. Sage Publications, Los Angeles.

Simmel, G., 2009. Sociology. Inquiries into the Construction of Social Forms. Brill, Leiden (Originally published in 1908).

Vedres, B., Scotti, M., 2012. Introduction: applied network science. In: Vedres, B., Scotti, M. (Eds.), Networks in Social Policy Problems. Cambridge University Press, Cambridge, pp. 1–11.

Warner, W.L., Lunt, P.S., 1941. The Social Life of a Modern Community. Yale University Press, New Haven, London.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of "small-world" networks. Nature 393, 440–442.

Wellman, B., 1988. Structural analysis: from method and metaphor to theory and substance. In: Wellman, B., Berkowitz, S.D. (Eds.), Social Structures: A Network Approach. Cambridge University Press, Cambridge, pp. 19–61.

von Wiese, L., 1932. Systematic Sociology. Adapted and Amplified by Howard Becker. Wiley & Sons, New York (Originally published in 1924).

Relevant Websites

http://www.insna.org/pubs/connections – Connections Publishes Original Empirical, Theoretical, and Methodological Articles that use Social Network Analysis.

http://www.insna.org/ – International Network for Social Network Analysis (INSNA).

journals.cambridge.org/NWS – Network Science is a Relatively New Journal that Features Articles using the Network Paradigm.

http://barabasilab.neu.edu/networksciencebook/ – Network Science Brook Project Aims to Produce an Interactive Textbook for Network Science.

http://www.journals.elsevier.com/social-networks – Social Networks is an Interdisciplinary and International Quarterly.

Network Centrality, Measures of

Peter V Marsden, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Measures of network centrality assess the importance of elements in arrays of relationships linking entities, such as persons, organizations, or documents. Important entities may have many contacts, be related to other important entities, be near other entities, or be positioned between others. Most centrality measures refer to entities, but some refer to relationships. Network-level centralization statistics measure variation in centrality across entities within a network. This article presents centrality measures for common forms of network data, including undirected binary networks, valued networks, directed networks, and two-mode networks of relationships that connect distinct types of entities.

Overview

Network analysis examines relationships among entities, such as persons, organizations, or documents. Operating at multiple levels, it describes and makes inferences about relational properties of individual entities, of subsets of entities, and of entire networks. Measures of network centrality are widely applied indices for describing positions of individual entities within networks.

Centrality is a covering term, referring generally to an element's 'importance' within a network. Visual depictions of networks usually place important elements in central locations. Different centrality measures correspond to distinct ways in which entities or relationships are involved in connectivity patterns: having many contacts, being near other entities, or being positioned between others. Social science applications use centrality measures to represent diverse theoretical concepts including power, influence, status, and prestige, among others.

This article begins by introducing network concepts that are helpful in understanding centrality measures. It then presents major centrality measures for the 'undirected, binary-valued' network data that many measures assume. Following that are variations on those measures suitable for other common data types, including valued, directed, and 'two-mode' arrays. For more comprehensive discussions of network centrality measures, see Koschützki et al. (2005) or Newman (2010).

Foundations

An undirected network data array consists of a set of entities and a set of pairwise relationships among them. Each relationship signifies only that two entities are somehow affiliated (e.g., that John and Sue communicate), making no distinction between 'sending' and 'receiving' entities.

Formally, entities are termed *vertices*, and indexed from 1 to N. Relationships in undirected networks are called *edges*. A binary-valued indicator variable x_{ij} indicates whether an edge between vertices i and j exists ($x_{ij} = x_{ji} = 1$) or not ($x_{ij} = x_{ji} = 0$). Most applications treat self-relations x_{ii} as undefined, setting them to 0 by convention, so an N-vertex network may include as many as $N(N-1)/2$ edges. When relationships are valued, weights w_{ij} give their intensity.

Figure 1 depicts an undirected, binary-valued network with 11 vertices labeled A–K. Lines indicate 12 pairwise relationships or edges; for example, a relationship exists between vertices A and B, but not between A and D.

Many centrality measures involve features of *indirect* relationships consisting of two or more concatenated edges. A *walk* – any sequence of edges beginning at one vertex and ending at another – is a very general type of indirect relationship. A walk may pass through a given vertex more than once and/or traverse a given edge repeatedly. In a *path*, no vertex or edge appears more than once. The number of edges in a path is its *length*, and a shortest-length path between two vertices is called a *geodesic*. The number of geodesics linking vertices i and j is symbolized g_{ij}; the length of the geodesic(s) linking them gives their *geodesic distance* d_{ij}. If no path of any length connects two vertices, their geodesic distance is infinite.

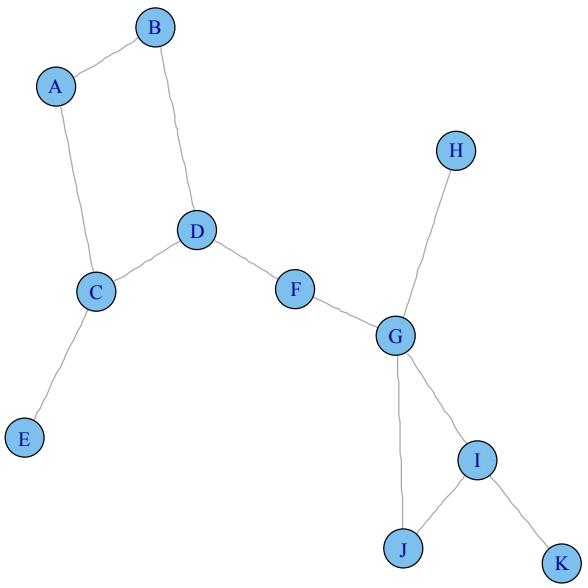

Figure 1 An undirected, binary-valued network. Constructed using igraph (Csárdi, Gábor, Nepusz, Tamás, 2006. The igraph software package for complex network research. InterJournal Complex Systems 1695).

In Figure 1, the vertex sequence (A, B, D, F, D, C) gives a walk of length 5 from vertex A to vertex C. It passes through vertex D twice, so it is not a path. The sequence (G, I, J) is a length-2 path from G to J; it is not, however, a geodesic: a length-1 path (i.e., an edge) also links G and J. Two ($g_{AD} = 2$) geodesic paths – (A, B, D) and (A, C, D) – connect vertices A and D, at geodesic distance $d_{AD} = 2$.

Centrality Measures for Binary-Valued Undirected Networks

Freeman (1979) inventoried measures of network centrality for vertices, elucidating their conceptual foundations. These measures were first developed for studying communication and related group processes; most were formulated for undirected, binary-valued network data. Borgatti and Everett (2006) highlight two primary ways in which a vertex can be structurally important, distinguishing between *radial* and *medial* measures. Radial measures focus on the direct and indirect relationships that originate from (or end at) a vertex, and vertices linked to others via short-distance relationships. Medial measures instead call attention to vertices situated in intermediary positions along walks or paths that connect pairs of others.

Radial Measures

Degree Centrality

The simplest centrality measure is a vertex's *degree*, its number of direct relationships to other vertices. For communication networks, Freeman (1979) suggests that degree centrality reflects a vertex's activity level. Degree centrality captures a vertex's importance in its locality only, ignoring its involvement in indirect relationships. The degree centrality of vertex i in an undirected network is

$$C_D(i) = \sum_{j=1}^{N} x_{ij}$$

Degree centrality counts the number of edges incident to a vertex, so $C_D(i)$ values can be greater when a network has more vertices. Normalized centrality measures adjust for network size by expressing a vertex's centrality as a fraction of its maximum possible level, such that the largest possible normalized centrality is 1. A normalized version of degree centrality is

$$C'_D(i) = \frac{C_D(i)}{N-1}$$

since a vertex can be linked to at most $N - 1$ others.

Table 1 reports scores for vertices in Figure 1 on several centrality measures discussed here. Vertex G's degree centrality is largest (4), while several peripheral vertices (E, H, and K) have the lowest degree (1). Vertex G could be linked to as many as 10 others, so its normalized degree centrality is 0.40.

Degree centrality gives the number of other vertices in a vertex's immediate neighborhood or 'first-order zone.' This idea generalizes readily to a measure of the size of a vertex's 'mth-order zone': the number of other vertices it can reach via geodesics of length m or less:

$$C_{Dm}(i) = \sum_{\substack{j=1 \\ (i \neq j)}}^{N} I(d_{ij} \leq m)$$

where $I(.)$ is an indicator variable. For $m = 1$, this yields degree centrality. If m equals the 'diameter' of a network (i.e., the length of the longest geodesic between any pair of vertices), one obtains a *reachability* index, the number of others that a vertex can reach via paths of any finite length. In Figure 1, vertex F can reach 8 of the other 10 vertices using paths of length 1 or 2, so $C_{D_2}(F) = 8$. $C_{D_6}(i) = 10$ for all vertices, since each lies within geodesic distance 6 or less of all the others.

Distance-Based Centrality Measures

Other radial measures take account of a vertex's proximity to all others, not just those that are nearby. Many such measures depend on geodesic distances, on the reasoning that if it is costly to activate relationships or secure the cooperation of intermediaries, transmitting information or concluding resource transactions using shortest paths will be most efficient.

Table 1 Selected centrality scores for vertices in Figure 1

	Centrality measure						
Vertex	Degree $C_D(i)$	Shortest-path closeness $C_C(i)$	Eccentricity $C_E(i)$	Shortest-path betweenness $C_B(i)$	Reflection (eigenvector)[a] $C_R(i)$	β-Centrality (β positive)[a] $C_\beta(i)$, $\beta = 0.31$	β-Centrality (β negative)[a] $C_\beta(i)$, $\beta = -0.31$
A	2	0.029	0.167	1.0	0.143	0.225	0.219
B	2	0.034	0.200	3.5	0.165	0.237	0.220
C	3	0.037	0.200	12.5	0.196	0.304	0.400
D	3	0.045	0.250	25.0	0.272	0.353	0.398
E	1	0.028	0.167	0.0	0.078	0.123	0.081
F	2	0.048	0.333	25.0	0.325	0.320	0.084
G	4	0.045	0.250	27.0	0.547	0.493	0.654
H	1	0.032	0.200	0.0	0.217	0.182	0.002
I	3	0.036	0.200	9.0	0.442	0.388	0.353
J	2	0.034	0.200	0.0	0.393	0.331	0.098
K	1	0.027	0.167	0.0	0.176	0.149	0.096

[a]Scaled such that sum of squared scores is 1.

In this view, central vertices are those linked to others by short distances. Such vertices can reach others more expeditiously, and hence are attractive as loci at which innovations or interventions can be introduced into a network. They require cooperation from fewer go-betweens to disseminate something widely, and hence are more autonomous. Since the geodesic distance d_{ij} gives the minimum number of edges separating two vertices i and j, the sum of all geodesic distances from a given vertex (i) to others $\left(\sum_{\substack{j=1\\(i\neq j)}}^{N} d_{ij}\right)$ gives the total number of edges between it and the rest of the network. The sum is smaller for more central vertices; inverting it (so that more central vertices have higher scores) gives a *shortest-path closeness* centrality measure:

$$C_{\mathrm{C}}(i) = \frac{1}{\sum_{\substack{j=1\\(i\neq j)}}^{N} d_{ij}}$$

When edges link a vertex to all others, the sum $\sum_{\substack{j=1\\(i\neq j)}}^{N} d_{ij}$ takes its minimum value, $N-1$. Hence the reciprocal of the average geodesic distance from a vertex yields a normalized closeness measure:

$$C'_{\mathrm{C}}(i) = \frac{N-1}{\sum_{\substack{j=1\\(i\neq j)}}^{N} d_{ij}}$$

By using the length of geodesics as a distance index, $C_{\mathrm{C}}(i)$ assumes that vertices locate and use geodesic paths when transacting or communicating with one another. This is not always tenable (Borgatti, 2005), especially for unplanned flow processes. Social actors may be aware only of their local ties and unable to identify shortest paths. Moreover, they do not always deliberately seek to route information or resources to a specific location.

One might posit that flows in networks instead follow *random walks* from one vertex to another. A random walk from vertex i to vertex j begins along a randomly selected edge from i to another vertex k; it next traverses a randomly selected edge emanating from k, continuing on until a randomly selected edge reaches vertex j. Though the length of such walks varies by chance, the expected number of edges m_{ij} in a random walk from i to j can be calculated (Koschützki et al., 2005) and regarded as the distance separating them. Parallel to $C_{\mathrm{C}}(i)$, a *random-walk closeness centrality* measure is the inverse of the sum of the expected random-walk distances from vertex i to all others:

$$C_{\mathrm{CRW}}(i) = \frac{1}{\sum_{\substack{j=1\\(i\neq j)}}^{N} m_{ij}}$$

Calculating closeness centrality measures entails complications in 'disconnected' networks where at least one pair of vertices is not linked by any path (or walk), so that their geodesic (or expected random-walk) distance is infinite. Closeness centrality is then 0 because the sums in the denominators of $C_{\mathrm{C}}(i)$ and $C_{\mathrm{CRW}}(i)$ are infinite if a vertex has no indirect connection to even one other. Many therefore suggest calculating closeness centrality only within 'connected components' of a network, i.e., subsets of vertices within finite geodesic distances of each other. Others recommend imputing some finite value (such as N) to infinite distances, but imputing a value for infinity is necessarily arbitrary. Still another approach (Newman, 2010) redefines the closeness centrality of a vertex as the sum of the inverse geodesic distances (rather than the inverse of the sum):

$$C^{*}_{\mathrm{C}}(i) = \sum_{\substack{j=1\\(i\neq j)}}^{N} \frac{1}{d_{ij}}$$

here, length-1 geodesic distances add most to a vertex's closeness centrality, while infinite ones contribute nothing. The random-walk closeness measure $C_{\mathrm{CRW}}(i)$ can be modified analogously.

A distinct distance-based centrality measure introduced by Hage and Harary (1995) depends on a vertex's *eccentricity*: the maximum geodesic distance emanating from it. Vertices with low eccentricity lie within a comparatively short distance of all others, and hence are central:

$$C_{\mathrm{E}}(i) = \frac{1}{\max_{j}(d_{ij})}$$

In networks representing routes linking places, vertices having high $C_{\mathrm{E}}(i)$ scores are attractive locations for facilities designed to respond to unpredictable events that might occur anywhere.

In **Figure 1**, vertex F has the highest shortest-path closeness. The sum of geodesic distances to F is 21, so $C_{\mathrm{C}}(\mathrm{F})$ is 0.048, with normalized closeness 0.478. Vertices D and G, each with closeness 0.045, come next. Vertex F also has the highest eccentricity closeness; it is within geodesic distance 3 of the other 10 vertices, so $C_{\mathrm{E}}(\mathrm{F}) = 0.33$. For the other vertices, eccentricity centrality is 0.25 or less.

Measures Involving Reflection/Feedback

The intuition that an entity acquires importance through affiliations with other important entities is widespread; one's reputation reflects that of one's associates, and vice versa. Bonacich's (1972) centrality measure is one of several that rest on this idea. It posits that a vertex's centrality is proportional to the centrality-weighted sum of its connections to others. That measure, $C_{\mathrm{R}}(i)$, satisfies the equation

$$C_{\mathrm{R}}(i) = \frac{1}{\lambda}\sum_{j=1}^{N} x_{ij} C_{\mathrm{R}}(j)$$

It outwardly resembles degree centrality, but edges linking a vertex to central vertices convey more importance than those connecting it to peripheral ones. The expression is circular: vertex i's centrality depends on that of its contacts, but their centrality in turn rises with i's. The relative magnitudes of the reflection centrality scores implied by this feedback process can be obtained using the largest

'eigenvalue' (λ) derived from the set of $N(N-1)/2$ network variables $\{x_{ij}\}$ in an undirected network. The $C_R(i)$ values are proportional to elements of the associated 'eigenvector'; they may be rescaled as long as this proportionality is maintained. $C_R(i)$ therefore is often termed *eigenvector* centrality.

With $C_R(i)$ scores scaled such that their squared values sum to 1, vertex G in Figure 1 has the highest reflection centrality, 0.55. By virtue of their connections to G, several other vertices also have relatively high reflection centrality: $C_R(F) = 0.33$, $C_R(I) = 0.44$, and $C_R(J) = 0.39$. Vertex D's reflection centrality is lower, at 0.27; although its degree (like I's) is 3, D's relationships link it to less central others.

The reflection or eigenvector measure $C_R(i)$ assumes that connections to central others enhance a vertex's centrality. Under some circumstances – notably in some bargaining situations – being near central others may be a disadvantage. Bonacich's (1987) flexible *β-centrality* measure generalizes $C_R(i)$, allowing the dependence of a centrality score on those of its associates to vary in both direction and intensity. β-centrality is defined as

$$C_\beta(i) = \sum_{j=1}^{N} (\alpha + \beta C_\beta(j)) x_{ij}$$

Parameter β governs the interdependence among β-centrality scores, while α controls their scaling. Analysts must supply values for both parameters; the magnitude of β should be less than $1/\lambda$.

Bonacich (1987) expresses β-centrality as a weighted sum of the number of walks (of any length) from a vertex, giving less weight to longer walks. The measure takes more account of longer walks when β is larger; for β near $1/\lambda$, it approaches reflection centrality. When β is smaller, $C_\beta(i)$ measures a more local centrality: if $\beta = 0$, $C_\beta(i)$ is proportional to degree centrality. A negative β means that direct connections (and odd-length walks) to central vertices reduce a vertex's β-centrality, while length-2 (and even length) indirect connections to central vertices heighten it.

With $\beta > 0$, β-centrality values for vertices in Figure 1 resemble those for reflection centrality. For $\beta = -0.31$, however, vertices G (0.65), C (0.40), and D (0.40) have high β-centrality. Vertex F, directly linked to both D and G, has a much lower value, $C_\beta(F) = 0.10$. Note, in contrast, that being connected to G increased F's reflection centrality.

Medial Measures

Vertices positioned between the endpoints of paths or walks have the potential to create – or limit – connectivity between others, a distinct form of structural importance. By virtue of their location, they can regulate flows within a network, and hence can be go-betweens, gatekeepers, bottlenecks, brokers, and the like. Centrality indices sensitive to this control capacity are known as medial measures.

Freeman (1979) introduced a widely used medial centrality measure based on geodesics. A focal vertex k lies between two other vertices i and j if it is an intermediary on one or more geodesics connecting them; $g_{ij}(k)$ denotes the number of times k is between i and j. Comparing this to g_{ij}, the total number of geodesics between i and j, gives the fraction of the i-j relationship that k is in a position to control:

$$b_{ij}(k) = \frac{g_{ij}(k)}{g_{ij}}, k \neq i, j$$

Summing across all pairs of vertices leads to a measure of vertex k's *shortest-path betweenness centrality*:

$$C_B(k) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} b_{ij}(k)$$

$C_B(k)$ reaches its maximum possible value, $(N^2 - 3N + 2)/2$, when vertex k lies along the only geodesic linking each of the $(N-1)(N-2)/2$ pairs of other vertices. A normalized betweenness centrality measure is therefore:

$$C'_B(k) = \frac{2C_B(k)}{N^2 - 3N + 2}$$

As noted for closeness measures, assuming that geodesics can be located and used can be problematic. Newman (2005) developed an alternative betweenness measure based on the premise that flows in networks follow random walks rather than geodesics. It first determines the expected number of times that a random walk between vertices i and j passes through vertex k. Aggregating that across all vertex pairs gives random-walk betweenness centrality.

In Figure 1, vertex F lies on the only geodesic linking D and G, so $b_{DG}(F) = 1$; B is on one of the two geodesics between A and D, so $b_{AD}(B) = 0.5$. Vertex G has the highest shortest-path betweenness centrality, with $C_B(G) = 27$, i.e., it is in a position to control the equivalent of 27 geodesics connecting other vertices; G's normalized score is 0.60. Vertices D and F have slightly lower normalized betweenness (0.56), while several vertices (E, H, J, K) lie along no geodesic paths and so have 0 betweenness.

Edge Centrality

Centrality measures usually describe the network positions of vertices, but they can be calculated for edges as well. Most widely applied is a medial measure, edge betweenness, that identifies edges that are most crucial to maintaining a network's connectivity.

By analogy to $C_B(k)$, an edge's betweenness is the number of times it lies on geodesics connecting pairs of vertices.

$$C_B(x_{km}) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} b_{ij}(x_{km}) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{g_{ij}(x_{km})}{g_{ij}}$$

where $g_{ij}(x_{km})$ is the number of times an edge linking vertices k and m lies on a geodesic linking vertices i and j. In Figure 1, both the (D, F) and (F, G) edges are on geodesics between 30 pairs of vertices (55% of all 55 pairs). Removing either edge would partition this network into two disconnected components.

Centralization

Corresponding to each vertex-level centrality measure is a network-level *centralization* statistic. Centralization measures the

extent of inequality among vertex centrality scores. A generic definition of centralization sums the differences between each vertex's centrality score and that of the most central vertex:

$$C_X = \sum_{i=1}^{N}\left(\max_k (C_X(k)) - C_X(i)\right)$$

where X references a particular centrality measure. When all vertices have identical centrality, centralization is 0; it grows larger as differences among scores increase. Dividing by the largest value C_X can take in any network having N vertices yields a normalized centralization score that varies between 0 (for equal centrality values) and 1 (for the largest possible differences from the maximum score). The largest possible C_X usually arises for a 'star' network with $N-1$ edges linking one central vertex to each other; for the reflection/eigenvector measure $C_R(i)$, C_X is maximal in a graph with a single edge.

In Figure 1, degree centrality is largest (4) for vertex G; summing the differences between this and the other 10 values of $C_D(i)$ yields raw centralization $C_D = 20$. In a star network with 11 vertices, one vertex has degree 10 while the 10 others have degree 1, so maximal degree centralization is 90; thus the normalized degree centralization value is 0.22. The corresponding normalized centralization values for shortest-path closeness and shortest-path betweenness are 0.27 and 0.43, so betweenness scores are somewhat more dispersed than the others.

Centrality in Valued Networks

While centrality measures are usually calculated using binary-valued data, versions of several exist for data that record variations in strength among edges, indexed by weights w_{ij}. Reflection centrality $C_R(i)$ and β-centrality $C_\beta(i)$ may be calculated directly after substituting the weights w_{ij} for the corresponding binary network variables x_{ij} in the formulas for those measures. Opsahl et al. (2010) suggest the total strength of a vertex's connections as a parallel to degree centrality for valued data:

$$C_D^W(i) = \sum_{j=1}^{N} w_{ij}$$

Developing closeness and betweenness centrality measures for valued network data require a distance measure d_{ij}^W that takes the edge weights into account. One algorithm for doing so defines the 'weighted length' of a valued edge as the reciprocal of its weight, i.e., $1/w_{ij}$, and that of a valued path as the sum of the weighted lengths of the edges composing it. For example, the weighted length of a path including two edges having weights of 2 and 4 is $1/2 + 1/4 = 0.75$, less than that of a single-edge path with weight 1. The distance d_{ij}^W between two vertices then may be defined as the minimum weighted length of any valued path connecting them. A closeness centrality measure for valued data analogous to $C_C(i)$ follows:

$$C_C^W(i) = \frac{1}{\sum\limits_{\substack{j=1\\(i\neq j)}}^{N} d_{ij}^W}$$

Constructing a betweenness measure $C_B^W(k)$ for weighted data begins by defining g_{ij}^W as the number of minimum weighted length paths connecting vertices i and j, and $g_{ij}^W(k)$ as the number of those paths that include vertex k as an intermediary. Then

$$C_B^W(k) = \sum_{i=1}^{N}\sum_{j=i+1}^{N} b_{ij}^W(k) = \sum_{i=1}^{N}\sum_{j=i+1}^{N} \frac{g_{ij}^W(k)}{g_{ij}^W}$$

Flow-betweenness centrality (Freeman et al., 1991) is a distinct medial measure for valued data. It assesses a vertex's importance to a network's capacity for transmitting information or resources. Flow between two vertices may proceed simultaneously along multiple channels (not only shortest paths) that share no edges. The capacity of an individual edge is its weight w_{ij}, while that of an indirect path is the minimum weight of any edge it includes. The potential flow (f_{ij}) between vertices i and j sums the capacities of one or more edge-distinct paths; the flow between i and j that must pass through an intermediary vertex k is denoted $f_{ij}(k)$. A vertex's flow-betweenness centrality is the ratio of the sums of these quantities over all pairs of vertices:

$$C_{FB}(k) = \frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N} f_{ij}(k)}{\sum_{i=1}^{N}\sum_{j=i+1}^{N} f_{ij}}$$

$C_{FB}(k)$ gives the extent to which the total flow between other vertices in a network depends on a particular vertex k. Koschützki et al. (2005) refer to it as a 'vitality' measure because it gives the proportional reduction in total network flow to be anticipated if vertex k (and all edges involving it) were to be removed from a network.

Centrality in Directed Networks

Many networks involve relationships in which something moves *from* one party *toward* another, and such flows need not be two way. Sentiments such as liking or admiration may be unrequited, for instance; likewise, authority relationships are usually one directional. 'Directed' data arrays represent networks of such relationships.

In directed networks, relationships (known as *arcs*) originate at one entity and end at another (e.g., John reports to Sue). A binary-valued network variable x_{ij} is 1 if an arc from origin vertex i to destination vertex j exists, and otherwise 0; a reciprocal arc from j to i may, but need not, be present. A network with N vertices may contain as many as $N(N-1)$ arcs. Weights w_{ij} may accompany arcs; here we consider binary data, however.

Definitions of indirect connectivity in directed networks take the orientation of arcs into account. A *directed path* is a sequence of concatenated and consistently oriented arcs leading from an origin vertex to a destination one, in which no vertex or arc appears more than once. The geodesic distance from one vertex to a second (d_{ij}) need not equal that from the second to the first (d_{ji}) in a directed network.

Figure 2 displays a directed, binary-valued network including 11 vertices and 20 arcs. Arrowheads indicate the orientation of arcs; an arc runs from vertex 5 to vertex 8, for example. Double-headed lines (e.g., that between vertices 1

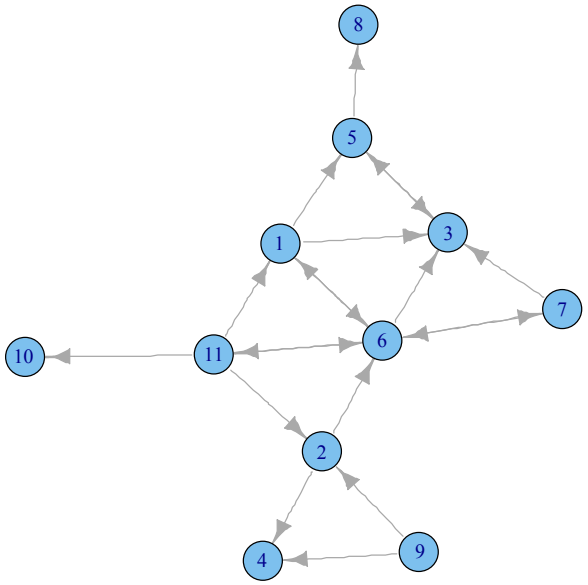

Figure 2 A directed, binary network. Constructed using igraph (Csárdi, Gábor, Nepusz, Tamás, 2006. The igraph software package for complex network research. InterJournal Complex Systems 1695).

and 6) denote relationships with arcs in both directions. A directed path of length 3 from 7 to 8 (via 3 and 5) exists, but no directed path from 8 to 7 is present, so the geodesic distance d_{87} is infinite.

Radial centrality measures for directed networks typically distinguish a vertex's importance as an origin of relationships from that as a destination. Vertices that score high as destinations are often deemed to be prestigious or to hold high status. This leads, for example, to separate measures of outdegree (O) and indegree (I) centrality:

$$C_{\mathrm{D}}^{O}(i) = \sum_{j=1}^{N} x_{ij} \quad \text{and} \quad C_{\mathrm{D}}^{I}(i) = \sum_{j=1}^{N} x_{ji}$$

Outdegree centrality – the number of outgoing arcs from a vertex – is sometimes termed 'expansiveness,' while indegree centrality – the number of incoming arcs – is known as 'popularity.'

Likewise, distinct outgoing and incoming geodesic distances lead to separate out-closeness and in-closeness measures. Counterparts to shortest-path closeness are

$$C_{\mathrm{C}}^{O}(i) = \frac{1}{\sum\limits_{\substack{j=1\\(i\neq j)}}^{N} d_{ij}} \quad \text{and} \quad C_{\mathrm{C}}^{I}(i) = \frac{1}{\sum\limits_{\substack{j=1\\(i\neq j)}}^{N} d_{ji}}$$

Analogous pairs of measures can be constructed using other distance concepts, including random-walk distance or eccentricity.

A generalization of reflection/eigenvector centrality also leads to separate outgoing and incoming measures of a vertex's importance. It defines a central incoming vertex as one targeted by arcs emanating from central outgoing vertices, and a central outgoing vertex as one that sends arcs toward central incoming vertices. Incoming reflection centrality is proportional to a sum of incoming arcs, weighted by the outgoing reflection centrality of their sources:

$$C_{\mathrm{R}}^{I}(i) \sim \sum_{j=1}^{N} x_{ji} C_{\mathrm{R}}^{O}(j)$$

Similarly, outgoing reflection centrality is proportional to a sum of outgoing arcs, weighted by the incoming reflection centrality of their targets:

$$C_{\mathrm{R}}^{O}(i) \sim \sum_{j=1}^{N} x_{ij} C_{\mathrm{R}}^{I}(j)$$

Kleinberg (1999) used these measures to identify important vertices in citation networks linking Web pages. 'Authorities' have high $C_{\mathrm{R}}^{I}(i)$ values; they are prominent sources of knowledge about a topic. 'Hubs,' on the other hand, provide portals leading to authorities; they have high $C_{\mathrm{R}}^{O}(i)$ values.

Incoming reflection centrality scores may be found by calculating reflection centrality $C_{\mathrm{R}}(i)$ for a valued 'cocitation' network with elements

$$s_{ij}^{I} = \sum_{k=1}^{N} x_{ki} x_{kj}$$

that give the number of sources that send arcs to both vertices i and j. Outgoing reflection centrality scores can be obtained via similar calculations for a valued network with elements

$$s_{ij}^{O} = \sum_{k=1}^{N} x_{ik} x_{jk}$$

giving the number of common destinations for arcs that emanate from both vertices in a pair.

In directed networks, a vertex may lie between two others in one direction, but not the other. Medial measures take account of this by counting betweenness for directed pairs of vertices. A directed version of shortest-path betweenness, for example, is

$$C_{\mathrm{B}}(k) = \sum_{i=1}^{N} \sum_{j=1}^{N} b_{ij}(k) = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{g_{ij}(k)}{g_{ij}}$$

where g_{ij} and $g_{ij}(k)$ refer to the number of shortest directed paths linking vertices i and j, and the number of those including vertex k as an intermediary.

Table 2 shows that vertices 6 and 11 in Figure 2 have the highest outdegree, 4; likewise, they also have relatively high outgoing closeness and reflection (hub) centrality. Because infinite geodesic distances separate many pairs of vertices in Figure 2, closeness scores in Table 2 were calculated using Newman's (2010) alternate definition. Vertex 1 has the highest outgoing reflection (hub) centrality because it sends arcs to two highly authoritative vertices, 3 and 6. The latter two vertices have the highest indegree and in-closeness. Vertex 6 ranks highest in terms of betweenness, positioned such that it can control the equivalent of 25.5 geodesics connecting other vertices; its normalized betweenness is 0.28.

Table 2 Selected centrality scores for vertices in **Figure 2**

	Centrality measure						
	Degree		Shortest-path closeness[a]		Shortest-path betweenness	Reflection (hub/authority)[b]	
Vertex	(Out) C_D^O (i)	(In) C_D^I (i)	(Out) C_C^{*O} (i)	(In) C_C^{*I} (i)	C_B (i)	(Out) C_R^O (i)	(In) C_R^I (i)
1	3	2	5.4	3.3	5.5	0.498	0.330
2	2	2	4.9	3.2	12.0	0.268	0.218
3	1	4	1.5	5.3	5.0	0.074	0.577
4	0	2	0.0	3.3	0.0	0.000	0.143
5	2	2	2.0	4.1	7.0	0.239	0.206
6	4	4	6.2	4.5	25.5	0.441	0.600
7	2	1	4.8	2.8	0.0	0.424	0.159
8	0	1	0.0	3.5	0.0	0.000	0.086
9	2	0	4.5	0.0	0.0	0.130	0.000
10	0	1	0.0	2.8	0.0	0.000	0.171
11	4	1	6.3	2.8	11.0	0.475	0.159

[a]Calculated using Newman, M.E.J., 2010. Networks: An Introduction. Oxford University Press, New York definition as the sum of reciprocal geodesic distances.
[b]Scaled such that sum of squared scores is 1.

Centrality in Two-Mode Networks

A special but common type of network involves relationships between two distinct 'types' of entities. Often these are affiliations of individuals with groups, e.g., memberships of directors on corporate boards, or of nations in international treaties. In these '*two-mode*' networks, an entity's type is fundamental: entities of a given type are never linked to one another directly, only indirectly via their shared affiliations with entities of the other type.

A two-mode network includes N_A vertices of type A, indexed from 1 to N_A, and N_B vertices of type B, indexed from $N_A + 1$ to $N_A + N_B = N$. Relationships x_{ij} in such networks are usually binary-valued and undirected. All relationships (edges) link vertices of different types:

$$x_{ij} \equiv 0, \text{ if } i \le N_A \text{ and } j \le N_A, \text{ or if } i > N_A \text{ and } j > N_A$$

The degree, closeness, betweenness, and reflection/eigenvector vertex centrality measures for undirected two-mode networks can be calculated as discussed earlier. Constructing normalized measures requires care, however: maximum possible centrality values depend on vertex type, and must take into account the restriction that no edges can connect vertices of the same type. For degree centrality, for example, a vertex of type A can be linked to all N_B type B vertices (not all $N - 1$ other vertices), while one of type B can be linked to as many as N_A type A vertices. So vertex type determines the appropriate adjustment factor used in obtaining a normalized degree centrality measure:

$$C'_D(i) = \frac{\sum_{j=N_A+1}^{N_A+N_B} x_{ij}}{N_B}, \text{for } i \le N_A [\text{type A vertices}], \text{and}$$

$$C'_D(i) = \frac{\sum_{j=1}^{N_A} x_{ij}}{N_A}, \text{for } i > N_A [\text{type B vertices}]$$

Everett and Borgatti (2005) give similar adjustments for normalized closeness and betweenness centrality in two-mode networks.

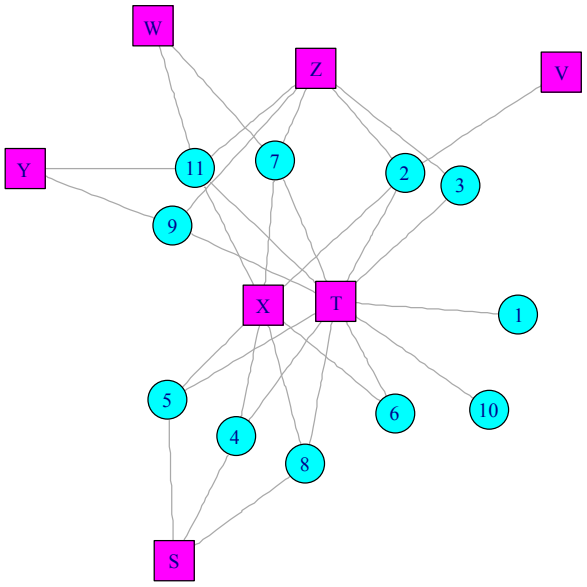

Figure 3 A two-mode network. Constructed using igraph (Csárdi, Gábor, Nepusz, Tamás, 2006. The igraph software package for complex network research. InterJournal Complex Systems 1695).

Reflection centrality scores for the distinct types of vertices in two-mode networks are interrelated in the same way that incoming and outgoing reflection scores are in directed networks. Vertices of one type acquire high reflection centrality via edges that connect them to high-centrality vertices of the other type:

$$C_R(i) \sim \sum_{j=N_A+1}^{N_A+N_B} x_{ij} C_R(j), \text{for } i \le N_A [\text{type A vertices}], \text{and}$$

$$C_R(i) \sim \sum_{j=1}^{N_A} x_{ij} C_R(j) \text{ for } i > N_A [\text{type B vertices}]$$

Table 3 Selected centrality scores for vertices in **Figure 3**

		Centrality measure			
Vertex	*Type*	*Degree* $C_D(i)$	*Shortest-path closeness* $C_C(i)$	*Shortest-path betweenness* $C_B(i)$	*Reflection (eigenvector)*[a] $C_R(i)$
1	A	1	0.026	0.0	0.108
2	A	4	0.030	19.0	0.259
3	A	2	0.027	1.2	0.165
4	A	3	0.029	5.3	0.221
5	A	3	0.029	5.3	0.221
6	A	2	0.027	0.6	0.189
7	A	4	0.030	9.9	0.273
8	A	3	0.029	5.3	0.221
9	A	3	0.029	6.5	0.189
10	A	1	0.026	0.0	0.108
11	A	5	0.032	19.9	0.297
S	B	3	0.021	1.0	0.146
T	B	11	0.043	73.4	0.495
V	B	1	0.020	0.0	0.057
W	B	2	0.022	0.4	0.125
X	B	7	0.032	21.9	0.369
Y	B	2	0.022	0.6	0.107
Z	B	5	0.027	9.7	0.259

[a]Scaled such that sum of squared scores is 1.

A type A vertex is central when it is related to central type B vertices, and vice versa.

Figure 3 displays a two-mode network including 11 vertices of type A (indicated by numbers and circles) and 7 vertices of type B (indicated by letters and squares). All 31 edges connect vertices of different types. **Table 3** reports several centrality measures for this network. Among the type A vertices, 2, 7, and 11 are most central; vertex 11 has the highest normalized degree centrality, 0.71. Vertex T is the most central type B vertex, with normalized degree centrality 1.0; vertex X is next. The less central type B vertices near the periphery of **Figure 3** are affiliated with smaller subsets of type A vertices.

Conclusion

Beyond the measures of network centrality covered here are numerous others. Among these are indices of degree, closeness, and betweenness centrality for sets of two or more vertices (Everett and Borgatti, 2005). A distinct approach (e.g., Gómez et al., 2003) synthesizes network analysis and game theory, conceiving of a vertex's importance as the incremental value accruing when the structure of a network delimits opportunities for exchange or coalition formation. Centrality then depends on both network structure and the rules and payoff structure of a particular game.

Though conceptually distinct from one another, the various centrality measures discussed here tend to assign similar relative scores to vertices in many empirical instances, as illustrated in **Tables 1–3**. Some definitional dependencies among them exist: a vertex with high degree has a large number of length-1 geodesics linking it to others, for example; its average distance to others tends to be lower, and it hence tends to score higher in terms of closeness. Likewise, because a vertex must have degree of at least 2 in order to be between any pairs of others, degree and betweenness tend to be associated. Notwithstanding this, the different measures are not interchangeable: they are sensitive to different features of structural importance within a network.

When selecting a centrality measure from among the many available, the conceptual or theoretical purpose to be served – that is, the way in which a particular centrality measure corresponds to one's notion of (e.g.,) power, status, or influence – is a vital consideration. One important aspect of this is the basic distinction between radial measures of a vertex's capacity to access others and medial measures of its potential as a transmitter. Friedkin (1991), for example, derives measures resembling reflection/eigenvector, closeness, and betweenness centrality from a single process model of social influence; these tap distinct components of a vertex's influence. As well, a study's substantive setting and the nature and content of measured relationships among entities merit attention. Borgatti (2005) offers some guidance here based on the match between centrality measures and different assumptions about the kinds of flow processes taking place within a network structure.

See also: Clustering and Cohesion in Networks: Concepts and Measures; Core–Periphery Patterns in Networks; Models of Influence in Networks; Models of Power in Exchange Networks; Network Analysis, History of; Network Analysis; Network Paradigm: Applications in Organizational Science; Social Network Algorithms and Software; Social Networks; Status and Role: Structural Aspects.

Bibliography

Bonacich, Phillip, 1972. Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology 2, 113–120.
Bonacich, Phillip, 1987. Power and centrality: a family of measures. American Journal of Sociology 92, 1170–1182.
Borgatti, Stephen P., 2005. Centrality and network flow. Social Networks 27, 55–71.
Borgatti, Stephen P., Everett, Martin G., 2006. A graph-theoretic perspective on centrality. Social Networks 28, 466–484.
Csárdi, Gábor, Nepusz, Tamás, 2006. The igraph software package for complex network research. InterJournal Complex Systems 1695.
Everett, Martin G., Borgatti, Stephen P., 2005. Extending centrality. In: Carrington, Peter J., Scott, John, Wasserman, Stanley (Eds.), Models and Methods in Social Network Analysis. Cambridge University Press, New York, pp. 57–76.
Freeman, Linton C., 1979. Centrality in social networks: conceptual clarification. Social Networks 1, 215–239.
Freeman, Linton C., Borgatti, Stephen P., White, Douglas R., 1991. Centrality in valued graphs: a measure of betweenness based on network flow. Social Networks 13, 141–154.
Friedkin, Noah E., 1991. Theoretical foundations for centrality measures. American Journal of Sociology 96, 1478–1504.
Gómez, Daniel, González-Arangüena, Enrique, Manuel, Contrado, Owen, Guillermo, del Pozo, Mónica, Tejada, Juan, 2003. Centrality and power in social networks: a game-theoretic approach. Mathematical Social Sciences 46, 27–54.
Hage, Per, Harary, Frank, 1995. Eccentricity and centrality in networks. Social Networks 17, 57–63.
Kleinberg, Jon M., 1999. Authoritative sources in a hyperlinked environment. Journal of the Association for Computing Machinery 46, 604–632.
Koschützki, Dirk, Lehman, Katharina A., Peeters, Leon, Richter, Stefan, Tenfelde-Podehl, Dagmar, Zlotowski, Oliver, 2005. Centrality indices. In: Brandes, Ulrik, Erlebach, Thomas (Eds.), Network Analysis: Methodological Foundations. Springer-Verlag, Berlin, pp. 16–62.
Newman, M.E.J., 2010. Networks: An Introduction. Oxford University Press, New York.
Newman, M.E.J., 2005. A measure of betweenness centrality based on random walks. Social Networks 27, 39–54.
Opsahl, Tore, Agneessens, Filip, Skvoretz, John, 2010. Node centrality in weighted networks: generalizing degree and shortest paths. Social Networks 32, 245–251.

Network Models of Tasks

Richard Schweickert, Purdue University, West Lafayette, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Processes executed in performing a task can often be represented in a network. Each process requires time and has a probability of producing correct output. When networks are used in scheduling, process durations and organization are known. The network is used to calculate the task completion time. A psychologist has the inverse problem. The time to complete the task is observed. The problem is to infer the organization of the mental processes and their durations. A similar inverse problem arises if accuracy is observed. These inverse problems can be solved by manipulating experimental factors that selectively influence processes.

Network Models of Tasks

A person performing a task such as judging the emotion portrayed in a face carries out processes such as perceiving, deciding, and responding. The processes can often be represented in a task network, as in Figure 1. Vertex o is the start of the task (typically, stimulus presentation). Each arc represents a process. Associated with each arc is the duration of the process and the probability the process produces the correct output for the input given. The processes leaving a vertex all begin when a release condition is met by the processes entering the vertex. A process leaving an AND gate vertex begins execution when all the processes entering the vertex are completed. A process leaving an OR gate vertex begins execution as soon as any process entering the vertex is completed. Vertex r is the end of the task, typically, response execution.

Representation of Processes

The bar chart (or Gannt chart) in Figure 2 illustrates a model of the psychological refractory period (Davis, 1957; Pashler, 1998). Two stimuli, o_1 and o_2, are presented, requiring responses r_1 and r_2, respectively. For stimulus o_1, perceptual processing is denoted as a_1, response selection is denoted as b_1, and motor preparation is denoted as c_1. Processes for stimulus o_2 are denoted analogously. The interval between the onsets of the stimuli, the stimulus onset asynchrony, is denoted as SOA. The length of the bar representing a process indicates its duration. The arrow from the end of b_1 to the beginning of b_2 indicates that b_1 must be completed before b_2 can start, in accordance with the single channel hypothesis that only one response can be selected at a time.

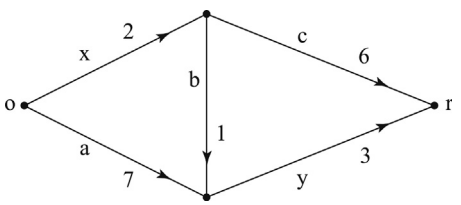

Figure 1 A task network. This is a Wheatstone bridge.

Process durations are easily apprehended from the bar chart, but precedence relations between the processes are not. The bar chart in Figure 2 can readily be converted to a task network, as in Figure 3. Precedence relations are more easily seen in the task network, but visual representation of the process durations is lost.

We begin by discussing the use of a network to represent time; later we discuss accuracy. In the theory of scheduling a task, network is used to calculate the time required to complete the task from known durations of directed paths through the network. A directed path from a vertex u to a vertex z is formed by starting at u, going along an arc leaving u to another vertex v, and so on, ending at the vertex z. There is no directed path from the starting vertex of a process to itself, that is, no directed path is a cycle. Hence, a task network is a directed acyclic network. The duration of a path is the sum of the durations of all the arcs on it. If the vertices are AND gates, the network is called a PERT (Program Evaluation and Review Technique) network, and the longest path from o to r is called the critical path. The time to complete the task is the duration of the critical path. If the vertices are OR gates, the time to complete the task is the duration of the shortest path from o to r. Usually

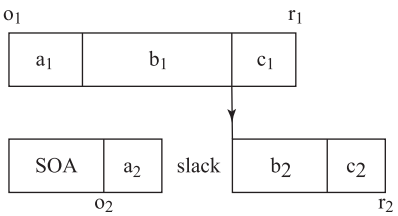

Figure 2 Bar chart for a double stimulation task.

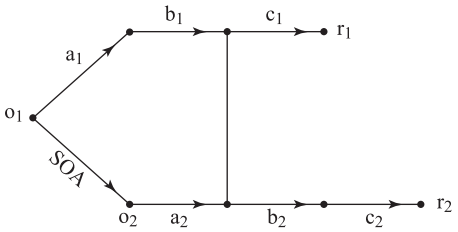

Figure 3 Bar chart in Figure 2, redrawn as a task network.

in tasks performed by humans, the time from stimulus onset to response execution is the reaction time.

Constructing the Network

One way to construct a network representing the processes in a task is through task analysis, that is, systematically observing someone perform the task and diagramming those processes that can be observed and those that can be inferred as logically necessary. Durations of observable processes can be found by timing them, and estimates of durations of unobservable processes, such as switching attention to a tone, can sometimes be found in the literature.

A successful example is a task analysis done by Gray et al. (1993). While new machines for telephone operators were under construction, videos of operators using the old machines were studied. PERT networks were drawn for performance of various tasks with the old machines and with the planned new machines. The new machines were designed to make certain individual processes faster. However, the PERT network analysis predicted that the overall time required for tasks with the new machines would be longer than with the old machines. The reason was that the faster processes were not always on the critical path for a task, so speeding them up did not always shorten the completion time for the task. On the other hand, with the new machines certain processes were added to the critical path, thus increasing the time needed to complete the task. When the actual new machines were installed, tasks indeed required more time.

The processes involved in a task are often unobservable mental processes. The analyst faces an inverse problem. The completion time of the task may be known under various conditions, and the problem is to infer the form of the task network and the durations of the processes. The key is to find experimental factors that selectively influence processes, a technique pioneered by Sternberg (1969) in his additive factor method. Suppose one experimental factor (say, degrading stimulus quality) prolongs one process, leaving the others unchanged. Suppose another experimental factor (say, increasing response complexity) prolongs another process, leaving the others unchanged. Suppose the processes required for the task are executed one after the other, in series. Then the increase in completion time, produced by prolonging both processes, will be the sum of the increases produced by prolonging them individually. This prediction can easily be tested. If it is satisfied, a network with processes in series can account for the data.

When process durations are stochastically dependent, factors influencing processes can have unexpected consequences (Townsend and Thomas, 1994). Complications arise when a factor prolongs a process p, and the duration of process p is correlated with the duration of a process q; then the duration of process q is changed when the level of the factor is changed. Nonetheless, processes whose durations are dependent on one another can be selectively influenced by factors producing well-behaved effects. A crucial assumption is needed when the process durations are conditionally independent, that is (for two processes), the duration of process p is independent of the duration of process q given a value of a third random variable c. For details, see Dzhafarov (2003).

Prolonging a Single Process

If all the processes are in series, two factors selectively influencing different processes will have additive effects on reaction time. But suppose some processes are concurrent with others. Then, when a process p is prolonged by an amount Δp, the response time need not increase by Δp. To see this, suppose all vertices are AND gates. In Figure 1 process x can be prolonged by up to 2 units without increasing the response time. The longest time by which a process p may be prolonged without delaying the response is called the total slack for p, denoted $s(p, r)$. The first $s(p, r)$ units of the prolongation are used to overcome the total slack for p, and the remainder increases the response time. Let $T(0)$ denote the response time when no processes are prolonged and let $T(\Delta p)$ denote the response time when process p is prolonged by Δp. Then $T(\Delta p) = T(0) + \max\{0, \Delta p - s(p, r)\}$. The numbers on the arcs in Figure 1 can be considered values of the process durations for a single trial in an experiment. If the process durations and the prolongation of p are random variables, then, taking expected values, $E[T(\Delta p)] = E[T(0)] + E[\max\{0, \Delta p - s(p, r)\}] \geq E[T(0)]$.

To see how slack arises, consider the SOA in Figure 2. Suppose processes a_1 and b_1 have long durations. Then increasing the duration of the SOA a little will have no effect on the time at which the response to the second stimulus is made.

Prolonging Two Processes

When two factors selectively influence two different processes, the effects on response time depend on how the processes are related. Relations between processes are determined by directed paths. If there is a directed path from the head of a process p to the tail of a process q, then p must be completed before q can start, and p precedes q. Processes p and q are called sequential if either p precedes q or q precedes p. If processes p and q are not sequential, they are concurrent. Concurrent processes p and q can be executed in any order. That is, p and q may be executed simultaneously, p may be completed before q starts, or q may be completed before p starts.

Assume (1) no process precedes itself (irreflexivity) and (2) if a process p precedes a process q, and process q precedes a process t, then p precedes t (transitivity). In other words, precedence is a partial order.

Let the response time when p is prolonged by Δp and q by Δq be denoted $T(\Delta p, \Delta q)$; the notation is analogous for other conditions. Suppose p and q are concurrent processes. Then the response time when both p and q are prolonged is the larger of the response time when p is prolonged and the response time when q is prolonged, that is,

$$T(\Delta p, \Delta q) = \max\{T(\Delta p, 0), T(0, \Delta q)\}$$

Considering expected values,

$$E[T(\Delta p, \Delta q)] = E[\max\{T(\Delta p, 0), T(0, \Delta q)\}]$$

A quantity sensitive to the relation between two prolonged processes is the interaction contrast (Schweickert, 1978). For concurrent processes x and y, this quantity is negative or 0, that is,

$$E[T(\Delta p, \Delta q)] - E[T(\Delta p, 0)] - E[T(0, \Delta q)] + E[T(0, 0)] \leq 0$$

Suppose p and q are sequential processes. When p and q are both prolonged the response time is a complicated function of their respective prolongations, Δp and Δq. When the prolongations are long enough to overcome the relevant slacks, the interaction contrast takes the following simple form,

$$E[T(\Delta p, \Delta q)] - E[T(\Delta p, 0)] - E[T(0, \Delta q)] + E[T(0, 0)] \\ = E[s(p, r) - s(p, q)]$$

Here $s(p, q)$ is the slack from p to q, the longest time by which p can be prolonged without delaying the start of process q. The quantity on the right hand side of the above equation is an important parameter of the network, called the coupled slack between p and q, denoted as $E[K(p, q)] = E[s(p, r) - s(p, q)]$. Its use has been described later.

Ordinarily, the total slack for p is greater than or equal to the slack from p to q, and consequently the interaction contrast for sequential processes is zero or positive. The one exception is when x and y are on opposite sides of a Wheatstone bridge, as in Figure 1. For this network, $s(x, r) = 2 \leq s(x, y) = 4$, and the interaction contrast is negative. Then processes on opposite sides of a Wheatstone bridge can mimic concurrent processes, because both arrangements of processes lead to negative interaction contrasts. The main difference is that for the Wheatstone bridge, when the prolongations are long, the value of the interaction contrast does not change as the prolongations become still longer. This is not true for the interaction contrast corresponding to concurrent processes.

When vertices are OR gates, the main difference is a reversal of signs. For concurrent processes, the interaction contrast is positive. For sequential processes not on opposite sides of a Wheatstone bridge, the interaction contrast is negative or zero. For sequential processes on opposite sides of a Wheatstone bridge, the interaction contrast is positive.

Information beyond that provided by interaction contrasts of mean reaction times can be obtained from interaction contrasts of their cumulative distribution functions (or, equivalently, survivor functions). For details, see Townsend and Nozawa (1995) and Schweickert et al. (2012).

Task Network Inference

By analyzing results of selectively influencing processes one at a time and in combinations, one obtains information about how processes are arranged in the network. As an example, consider the data in Table 1, from an experiment by de Jong (1993). Two stimuli were presented to the subject on each trial, a tone and a letter. The tone was either low or high pitched, and the subject responded to it with the left hand, one button to indicate the low-pitched tone and another button to indicate the high-pitched tone. The letter was either a B or a D, and the subject responded to it by pressing a button with the right hand in one condition and by pressing a pedal with the left or the right foot in another condition. In some blocks of trials the subject pressed one button (or pedal) for B and another button (or pedal) for D, so there were two alternatives for the response to the letter. In other blocks of trials, the subject pressed the same button (or pedal) for both B and D, so there was only one response alternative for the letter. The interval between presentation of the tone and of the letter (the SOA) was 25, 250, or 800 ms. Table 1 gives the mean time at which the response to the letter was made, measured from the onset of the tone. (If we let RT2 denote the interval between the presentation of the letter and the response to the letter, then Table 1 gives RT2 + SOA.) The times were read from Figure 4 of de Jong (1993). The reader is referred to the paper for further data and to Schweickert et al. (2012) for further analysis; just enough is presented here to work out the example.

Table 1 Mean RT2 + SOA in de Jong's (1993) experiment 2

		SOA	
Response 2 alternatives	*Response 2 limb*	*25*	*800*
2	Foot	672	1224
2	Hand	587	1138
1	Foot	545	1071
1	Hand	529	1000

All times are in milliseconds.
Data read from Figure 4 of de Jong, R., 1993. Multiple bottlenecks in overlapping task performance. Journal of Experimental Psychology: Human Perception and Performance 19, 965–980.

The data give clear evidence that there is slack for the SOA. From the bottom row of the table, when the SOA was increased from 25 to 800 ms, an increase of 775 ms, the time to respond to the letter increased by only 471 ms.

There are three factors, SOA, the number of response 2 alternatives, and the response 2 limb. Suppose these selectively influence processes SOA; central processing, b; and motor preparation, c, respectively.

Suppose all the prolongations are long enough to overcome the relevant slacks. Then the coupled slack from SOA to c is

$$K(\mathrm{SOA}, c) = T(\Delta\mathrm{SOA}, 0, \Delta c) - T(\Delta\mathrm{SOA}, 0, 0) - \mathrm{T}(0, 0, \Delta c) \\ + T(0, 0, 0) \\ = 1071 - 1000 - 545 + 529 = 55$$

The expression $T(0, 0, \Delta c)$ denotes the observed mean time for the response when process c is prolonged but no other process is; other expressions are analogous.

The coupled slacks for the other two pairs of processes can be calculated in a similar manner. They are

$$K(\mathrm{SOA}, b) = 80 \text{ and } K(b, c) = 69$$

From these coupled slacks the time at which the response is made when all three processes are prolonged can be predicted. (Derivation is beyond the scope of this article, see Schweickert, 1978.) The predicted value is

$$T(\Delta\mathrm{SOA}, \Delta\mathrm{b}, \Delta\mathrm{c}) = T(\Delta\mathrm{SOA}, 0, 0) + K(\mathrm{SOA}, b) + T(0, \Delta b, 0) \\ + K(b, c) + T(0, 0, \Delta c) - 2T(0, 0, 0)$$

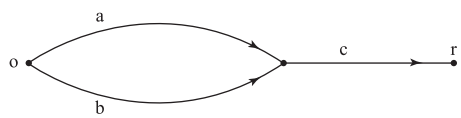

Figure 4 A task network.

The observed value (left hand side) is 1224. The predicted value (right hand side) is 1223. The observed and predicted values agree, support for the model. The outcome not only is a test but also provides information about the process order. There are three coupled slacks, but only two appear in the above equation. The coupled slack $K(\text{SOA}, c)$ is not used; it corresponds to the first and last of the three processes. It can be seen that process b is in the middle. One does not learn which of SOA or c is first from the equation, but logically SOA comes first.

In the example, something is learned about process order through parameter estimates. For a complicated task, it would not be practical to estimate all the parameters in the same experiment, and parameter estimates from different experiments may not be commeasurable. Fortunately, results from different experiments can be combined, without the need for precise parameter estimates. Because concurrent processes and sequential processes behave differently, one can determine which pairs of processes are concurrent and which are sequential. This information, possibly from different experiments, is sufficient for constructing a directed acyclic graph. One begins with a comparability graph in which each process is represented by a vertex, and two vertices are joined by an undirected edge if and only if the corresponding processes are sequential. All possible directed acyclic graphs can then be constructed with the transitive orientation algorithm (Golumbic, 1980).

Estimating Process Durations

Suppose process duration distributions are exponential or gamma, a reasonable assumption. Then the response time can be derived from an order-of-processing (OP) diagram (Fisher, Goldstein, 1983). Figure 4 gives a PERT network in which processes a and b are concurrent and both are followed by process c. Figure 5 gives the OP diagram made from the PERT network. Each vertex in the OP diagram is a state, in which certain processes are underway (upper list) and certain others are completed (lower list). The task begins with a and b both underway. In the following states, either a is completed or b is completed. Following either of these states, process c is underway. In the final state, process c is completed, and the task is finished.

Suppose processes have independent exponential distributions, with the mean of process a denoted as $1/\lambda_a$, and so on (λ_a is the rate of process a). The expected value of the response time can be read off the OP diagram. The first state is completed as soon as the faster of processes a and b is completed. The expected value of the minimum of two independent exponentially distributed processes is itself exponentially distributed, so the expected value of the duration of the first state is $1/(\lambda_a + \lambda_b)$. The probability that a is the first process to finish is $\lambda_a/(\lambda_a + \lambda_b)$. Because b is exponentially distributed, the expected value of the time remaining for b to complete given that a has already completed is $1/\lambda_b$. After repeated application of these rules, the expected value of the response time is the sum over all states of the expected duration of each state times the probability of reaching that state. For Figure 4,

$$E[T] = 1/(\lambda_a + \lambda_b) + [\lambda_a/(\lambda_a + \lambda_b)][1/\lambda_b] + [\lambda_b/(\lambda_a + \lambda_b)] \times [1/\lambda_a] + 1/\lambda_c$$

From the expressions for response times, one can use a nonlinear parameter estimation program to estimate the unknown rates, given the observations from enough conditions.

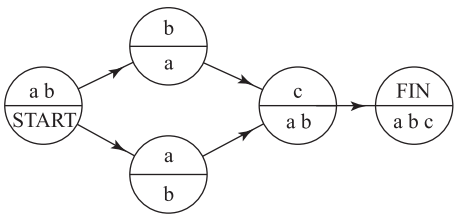

Figure 5 OP diagram for the task network of **Figure 4**.

A Generalization

OP diagrams can represent more general arrangements of processes than those in directed acyclic networks. Consider Figure 2. Suppose b_1 and b_2 are sequential, but either could be executed first. The situation cannot be represented with a single directed acyclic network, but can be represented with one OP diagram.

Processing Tree Representations

The accuracy of a task is often described through a processing tree model (for review see Batchelder and Riefer, 1999). In these models, a process is represented by a vertex and the outcomes of the process by arcs descending from the vertex; see Figure 6. Associated with each arc is the probability that the outcome it represents occurs. When the task starts, processing begins at the root, a vertex with no arc entering it. One of the outcomes of this process occurs, and the arc corresponding to this outcome is followed to its ending vertex. The process at this vertex is executed, one of its outcomes occurs, and so on, until a vertex with no descending arc is reached. Such a vertex is called a terminal vertex. Each terminal vertex is associated with a type of response, for example, a correct or wrong response. When a terminal vertex is reached, a response of its type is made.

A response of a certain type is made when a path is followed from the root to a terminal vertex for this response type. The probability of such a path is the product of the probabilities associated with the arcs on the path. The probability of a response of a certain type is the sum of the probabilities of all the paths resulting in that type of response.

The Inclusion/Exclusion Paradigm

A popular use of processing trees is to model the inclusion/exclusion paradigm developed by Jacoby (1991). Typically in the paradigm a subject is presented with a list of words to memorize, presented, say, visually. Then the subject is presented with another list to memorize, presented, say, auditorily. A recognition test follows. On each trial a test word is presented. Subjects in the inclusion condition are instructed to

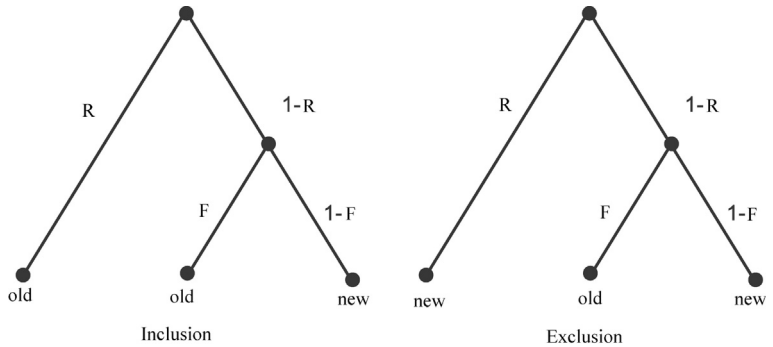

Figure 6 Processing trees for the inclusion/exclusion paradigm.

say the word is old if it appeared in either list. Subjects in the exclusion condition are instructed to say the word is old only if it appeared in the second list.

According to the model, subjects in the exclusion condition say a word is old in two ways. One way is that they vividly recall learning the word, a state called recollection. In that case they will remember the way the word was presented. The other way they will say a word is old is if they do not recollect learning it, but the word looks familiar. Let R denote the probability of recollecting a word and let F denote the probability a word looks familiar, given it was not recollected.

Consider a test word originally presented in list 2. A subject in the exclusion condition recollecting it will remember the way it was presented, conclude it was not presented in list 1, and say new. The only way a subject in the exclusion condition would say a test word from list 2 is old is if the subject does not recollect the word, but the word seems familiar. The probability is $(1-R)F$; see Figure 6.

A subject in the inclusion condition will say a word is old if the subject recollects learning it, or if the subject does not recollect learning it but the word looks familiar. A subject in the inclusion condition will say a word from list 2 is old with probability $R+(1-R)F$. The two equations above can be solved for the two unknown quantities, R and F. (They are not uniquely determined.)

To test the model, experimental factors are manipulated that are expected to change one parameter but not the other, that is, to selectively influence recollection or familiarity. As an example, subjects in a divided attention condition learned words while concurrently performing a distracting task. Subjects in a control condition did not perform the distracting task. The distracting task was expected to change the probability R of recollection, but not the probability F that a word would look familiar, and this was indeed the result.

Tree Inference

An investigator often begins by specifying a plausible processing tree for a task, guided by task analysis and previous findings in the literature. The model is tested by estimating parameters and evaluating goodness of fit. Another approach, tree inference, is to construct a processing tree by examining the effects of factors thought to selectively influence processes in the tree (Schweickert and Chen, 2008). Consider the case where there are two response types, say, correct and wrong. Suppose the investigator manipulates two experimental factors. Consider the hypothesis that the subject performs the task using a processing tree and each factor changes parameters on arcs that descend from a single vertex, a different vertex for each factor.

There are two possibilities for the two vertices. One possibility is that there is no path from the root to a terminal vertex for a correct response that passes through the two vertices; the other possibility is that there is such a path. Let i be a level of one factor and j be a level of the other. A particular processing tree with no path that passes through the two vertices gives the probability of a correct response as

$$p_{ij} = ax_i + (1-a)y_j$$

This tree is called the standard tree for unordered processes.

A particular processing tree with a path that passes through the two vertices gives the probability of a correct response for the combination of factor levels i and j as

$$p_{ij} = w_i + x_i y_j$$

This tree is called the standard tree for ordered processes. (It is for the case when the factor with levels indexed by i selectively influences the vertex that comes first on the path from the root. For the reverse order, one would switch i and j in the above equation.)

These two trees can easily be distinguished with data. Given data, parameters for the first equation can be found if and only if the factors have additive effects on the probability of a correct response. Parameters for the second equation can be found if and only if the following two conditions hold.

The levels j can be numbered in order so that if $j > j'$ then for every level i, $p_{ij} \geq p_{ij'}$.

There exist levels i^* and j^* such that for every i there is a number $r_i \geq 0$ with $p_{ij} - p_{ij^*} = r_i(p_{i^*j} - p_{i^*j^*})$.

If neither standard tree fits the data, then there exists no processing tree in which each factor selectively influences a parameter on arcs descending from a single vertex. Moreover, suppose the subject employs an arbitrary processing tree in which each factor selectively influences a parameter on arcs descending from a single vertex. Then for every combination of levels of the factors the processing tree used by the subject makes exactly the same predictions as one of the two standard trees.

See also: Discrete-State Models of Information Processing; Information Processing Architectures: Fundamental Issues; Mathematical Psychology; Memory Models: Quantitative; Modularity and Organizations; Partial Orders.

Bibliography

Batchelder, W., Hand Riefer, D.M., 1999. Theoretical and empirical review of multinomial process tree modeling. Psychonomic Bulletin and Review 6, 57–86.

Davis, R., 1957. The human operator as a single channel information system. Quarterly Journal of Experimental Psychology 9, 119–129.

de Jong, R., 1993. Multiple bottlenecks in overlapping task performance. Journal of Experimental Psychology: Human Perception and Performance 19, 965–980.

Dzhafarov, E.N., 2003. Selective influence through conditional independence. Psychometrika 68, 7–25.

Fisher, D.L., Goldstein, W.M., 1983. Stochastic PERT networks as models of cognition: derivation of the mean, variance, and distribution of reaction time using order-of-processing (OP) diagrams. Journal of Mathematical Psychology 27, 121–151.

Golumbic, M., 1980. Algorithmic Graph Theory and Perfect Graphs. Academic Press, NY.

Gray, W.D., John, B.E., Atwood, M.E., 1993. Project Ernestine: validating GOMS for predicting and explaining real-world task performance. Human Computer Interaction 8, 237–309.

Jacoby, L.L., 1991. A process dissociation framework: separating automatic from intentional uses of memory. Journal of Memory and Language 30, 513–541.

Pashler, H., 1998. The Psychology of Attention. MIT Press, Cambridge, MA.

Schweickert, R., 1978. A critical path generalization of the additive factor method: analysis of a Stroop task. Journal of Mathematical Psychology 18, 105–139.

Schweickert, R., Chen, S., 2008. Tree inference with factors selectively influencing processes in a processing tree. Journal of Mathematical Psychology 52, 158–183.

Schweickert, R., Fisher, D.L., Sung, K., 2012. Discovering Cognitive Architecture by Selectively Influencing Mental Processes. Singapore: World Scientific.

Sternberg, S., 1969. The discovery of processing stages: extensions of Donders' method. In: Koster, W.G. (Ed.), Attention and Performance II. North Holland, Amsterdam, pp. 276–315.

Townsend, J.T., Nozawa, G., 1995. Spatio-temporal properties of elementary perception: an investigation of parallel, serial, and coactive theories. Journal of Mathematical Psychology 39, 321–359.

Townsend, J.T., Thomas, R.D., 1994. Stochastic dependencies in parallel and serial models: effects on systems factorial interactions. Journal of Mathematical Psychology 38, 1–34.

Network Paradigm: Applications in Organizational Science

Charles C Snow, The Pennsylvania State University, State College, PA, USA; and BI-Norwegian Business School, Oslo, Norway
Øystein D Fjeldstad, BI-Norwegian Business School, Oslo, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Network theory and analysis have played a prominent role in the field of organization science for more than six decades. Previous articles of the organizational networks literature have discussed the attributes, antecedents, and consequences of networks. Our article focuses on the organizational functions of networks. Specifically, we discuss how networks enable exploration, exploitation, and organization.

Network is a valuable concept in the natural, biological, social, and engineering sciences. It has been used to study a wide variety of topics including gene sequencing, terrorist cells, computer software, job search, organizational structure and innovation, and the global economy (Barabási, 2002). A network is a set of *actors* (nodes) and the *ties* (links) between them. In a social network, an actor can be an individual, group, firm, organization, or national government, so the scope and complexity of a network analysis can be considerable (cf Castells, 2010; Johnson, 2014). Network analysis has been utilized in the organization sciences for more than six decades (Simmel, 1950; White, 1963), and arguably it has attained the status of a paradigm (Kuhn, 1962). In this article, we focus on organizational networks and examine them from a functional perspective – how they enable organizations to explore, exploit, and organize.

Properties of Networks

Network research has examined the attributes, antecedents, and consequences of networks (for reviews see Borgatti and Foster, 2003; Borgatti and Halgin, 2011; Brass et al., 2004). Networks have tie properties and structural properties. Ties are characterized by the type of content they carry, such as goods, information, or moral support, and by their strength. Tie strength is a function of various factors, including time and reciprocity, and tie strength affects how a network operates. Strong ties provide support among actors and enable the transfer of rich information and tacit knowledge. Weak ties, on the other hand, are sources of diverse resources and novel information (Granovetter, 1973).

The basic network structure is a dyad: two actors connected by a common tie. Dyads are the building blocks of larger network structures. A triad consists of three actors and at least two ties (Simmel, 1950). If each actor in the triad is directly connected – that is, the triad consists of three dyads – then the triad is closed. A closed triad confers the benefit of trust among actors in the network. When a triad is open, Simmel (1950) called the actor with ties to the two unconnected actors the *tertius gaudens*, which means 'the third who benefits'. A *tertius gaudens* can serve as a broker between actors and may benefit from playing them off against each other, thereby becoming powerful within the network. Triads, in turn, are the building blocks of clusters, which are agglomerations of densely connected actors.

Networks are unbounded in the sense that everyone is directly or indirectly linked to everyone else. Researchers, therefore, usually study particular subsets of networks. An ego network is the network surrounding a focal actor. The ego network is a source of both opportunities and constraints. Three important ego-network measures are network size, actor centrality, and network density. Centrality is a measure of the importance of an actor in a network, and centrality measures differ in the weights given to direct and indirect ties. Centrality confers status, influence, and access to resources in the network. Density is the ratio of actual to possible ties among a set of actors. High density in an ego network provides the focal actor with closure benefits, such as trust and shared understanding, and it constrains opportunistic actions and reduces costly monitoring processes. As a result, a closed network provides a basis for cooperation and collective action.

Large networks consist of interconnected clusters. Such networks are characterized by their degree of centralization, which measures the extent to which there are central actors in the network; degree of clustering, which measures the extent to which triads in a network are closed; and average path length, which measures the average number of links in the shortest path between any two actors in the network. Degree of clustering and average path length are used to characterize how actors in large networks have access to closure benefits as well as the benefits that come from network integration, such as the rapid diffusion of knowledge and information. Small worlds are non-centralized, clustered networks with short average path lengths (Milgram, 1967; Watts, 1999). Small worldliness can increase network performance by allowing resources to develop locally and flow efficiently between the clusters (Uzzi and Spiro, 2005).

Network size and structure are determined by the interaction of technological, social, and institutional factors. Katz and Shapiro (1985) showed that expectations of future network size influence potential actors' decisions to join the network. Moreover, a network's initial properties create path dependence in both current network attractiveness and the actors' expectations about future network size (Arthur, 1989). Finally, the actions of individual actors, guided by 'generative rules' (Kogut, 2000) concerning interaction with others, affect both network size and structure. Thus, as the network evolves, its size and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.73107-2

shape result from past decisions, future expectations, and generative rules.

Functions of Networks

In order to survive and prosper, organizations need to continuously learn and adapt. Networks enable organizations and their members to both shape and adapt to their environments. Organizations make ongoing decisions about *exploring*, the development of new knowledge, capabilities, and opportunities (March, 1991); *exploiting*, the diffusion, utilization, and refinement of existing knowledge, capabilities, and operations (March, 1991); and *organizing*, the accumulation and arrangement of resources into an activity system (Miles and Snow, 1978). Those decisions affect both the short-term performance and the long-run survival of organizations. Exploring, exploiting, and organizing are three core activities of organizations that are facilitated by various types of networks and network processes (Lazer and Friedman, 2007).

Exploring

Organizations arise in response to opportunities. Getting organized implies the exploration of opportunities and the development of capabilities to pursue them. Organizations learn from search and experimentation, and they require absorptive capacity (Cohen and Levinthal, 1990) in order to develop their capabilities. Networks are sources of information, opportunities, knowledge, and capabilities, and different types of network ties and structures play complementary roles in exploration. See **Table 1** for a summary of how networks enable organizations to explore.

Table 1 How networks enable exploration

Network attribute	*Function*
Weak ties	Low-intensity, sporadic relationships support exploration by providing access to novel information and diverse resources.
Strong ties	Intimate, long-lasting relationships facilitate the transfer of rich information and tacit knowledge, thus enabling joint problem solving. However, strong ties, coupled with path dependence, may inhibit sensitivity to disruptive innovations.
Structural hole	Brokers and others who span structural holes can enhance exploration initiatives because of their special access to information, knowledge, and resources. However, they can also divert exploration if they use their proprietary knowledge to engage in power plays or competitive maneuvering.
Centrality	Centrality provides more influence and greater access to information and resources.
Clustering	Direct connections among actors provide them with closure benefits such as shared norms, trust, and reciprocity.
Small worldliness	Dense clustering and short path lengths provide both diversity and closure benefits, and small worldliness positively affects creativity and innovation.

Weak Ties

Network actors typically have many weak ties, which connect them to a variety of other actors. Therefore, weak ties help to integrate social systems, providing access to novel information and diverse resources (Granovetter, 1973). Novel information is important to exploring such things as job opportunities, new technologies, and market opportunities.

Strong Ties

Strong ties facilitate the transfer of tacit or non-codified knowledge as well as the process of joint problem solving (Dyer and Nobeoka, 2000). Strong ties are embedded in social contexts such as long-term personal relationships among the actors in a network (Uzzi, 1996) or in relationships that the actors maintain with others outside the network (Granovetter, 1985). Strong ties facilitate exploration in complex, dynamic environments because of the bandwidth they provide for the transmission of rich, continuously changing information. However, strong ties may inhibit the exploration of discontinuous or disruptive innovations if they limit the scope of search processes. In such cases, firms may need to create new networks particularly suited to enabling discontinuous innovation.

Structural Hole

By extending the notion of an open triad to larger network structures, Burt (1992) developed the concept of a structural hole. A structural hole exists when an actor provides the only connection between other actors. Such an actor spans a structural hole and can benefit in two ways: (1) access to information, knowledge, and resources that are not universally held and (2) the ability to play actors, or groups of actors, off against each other in the competition for resources. The spanning of structural holes has been shown to affect careers and compensation, patenting, alliance performance, corporate acquisition performance, market entry choices, and firm financing and growth.

Centrality

Actors with high centrality have direct or indirect ties to a large portion of the network and therefore have links to otherwise unconnected actors. Overall, centrality provides greater access to resources (Cook and Emerson, 1978), and it can be parlayed into gains in individual performance and promotions, organizational innovations, and organizational performance.

Clustering

Degree of clustering affects actors' ability to monitor other actors' conduct and to sanction undesirable behavior. Actors create and maintain relationships within the network that provide them with collective closure benefits in the form of shared norms, trust, and reciprocity (Coleman, 1990). Clustered networks can transmit sensitive and rich information, mobilize collective action, and increase group innovation.

Small Worldliness

Both structural holes and closure are important attributes of social capital. Small world networks combine high clustering and short path lengths (Watts, 1999). A small world network may offer actors, as a collective, both diversity and closure

benefits, and small worldliness has been shown to positively affect creativity and innovation.

Exploiting

Exploitation is the diffusion, utilization, and refinement of existing knowledge, capabilities, and operations. Organizations benefit from both the internal and external diffusion of valuable innovations resulting from exploration. An organization benefits directly when customers adopt its products and services, and it benefits indirectly when knowledge and practices are adopted by employees and others with whom the organization interacts (Katz and Shapiro, 1986). Adopters are influenced by exposure to innovations from members of their professional and social networks, and successful diffusion is a function of resource value, the properties of sources, and the nature of adopters and their relationships with resource providers. Actors vary in their willingness and ability to transfer information and knowledge to others. Whereas early diffusion research focused on the attributes of early adopters and their networks (Rogers, 1962), recent studies have highlighted the importance to diffusion of actors' knowledge transfer capacity (Martin and Salomon, 2003), the degree of complementarity between the transmitter and the recipient (Aral and Walker, 2012), and the recipient's perception of the trustworthiness of the source (Szulanski et al., 2004). See **Table 2** for a summary of how networks enable organizations to exploit.

Tie Strength

Knowledge transfer, even within organizations, is difficult. Strong embedded ties are conducive to knowledge transfer because they are associated with trust and provide common frames of reference. Tie strength has been shown to be important in the transfer of product development knowledge, knowledge about market trends affecting firm survival, and best practices among units in multinational corporations.

Actor Position

Central actors have broad exposure to others and enjoy high status. As a result, information and innovations rapidly diffuse through their networks (Rogers, 1962). Their central position also gives them a 'vision advantage' (Burt, 2004), which favors their early adoption of innovations. However, Aral and Walker (2012) found that highly influential individuals tend not to be susceptible to influence by others. The diversity of information available to central actors may prevent information from reaching a threshold level required for adoption (Granovetter, 1978). Overall, central actors aid diffusion through their networks, but they are not necessarily good targets for the initial adoption of innovations.

Table 2 How networks enable exploitation

Network attribute	*Function*
Tie strength	Strong ties are conducive to knowledge transfer because of trust and common frames of reference.
Actor position	Central actors rapidly diffuse information and innovations. Because of their high status and central position, they are able to strongly influence other actors. However, central actors tend to be less susceptible to influence by others.
Connectivity	A moderate amount of connectivity is required for effective diffusion.
The network effect	For products, practices, or knowledge exhibiting a network effect, the benefits of joining a network increase with the adoption by others.

Connectivity

Studies of infectious diseases show that dense connectivity (multiple pathways among actors) increases spreading (Watts, 2004). However, dense connectivity has a different impact on the diffusion of innovations or information. When the quality of information is unknown, actors tend to seek social support for adoption, and if the actors are densely connected they may not be affected by new information because of competition from other sources. Therefore, diffusion takes place only when the connectivity of the network is 'neither too sparse nor too dense' (Watts, 2004: 261).

The Network Effect

The network effect refers to an actor's benefit from adoption being a function of the adoption by others. It arises in a number of situations such as the common adoption of institutions that enable societal specialization and integration, practices that facilitate knowledge sharing and collaboration, and new communication technologies. For products, practices, or knowledge exhibiting a network effect, the benefits of joining a network increase with the extent of adoption (Katz and Shapiro, 1985). A strong network effect can lead to winner-take-all situations (Shapiro and Varian, 1999) or fewer competing technologies in an industry (Katz and Shapiro, 1986).

Organizing

Organizing is the accumulation and arrangement of resources into an activity system. Activity systems are dynamic, with some parts of the system changing more frequently than others. An organization is a goal-directed system that must be controlled and coordinated. Networks facilitate control and coordination both within and across organizations. Tie properties as well as structural properties contribute to the organizing function. See **Table 3** for a summary of the organizing functions of networks.

Formal Ties

The formal structure of an organization is a network. Structure reflects how activities have been divided among organizational actors (division of labor), and activities are integrated through control and coordination mechanisms. Formal ties are the static, designed relationships typically shown in the organization chart, including reporting relationships and communication channels. Traditional organizational forms employ hierarchical mechanisms as the primary means of control and coordination. In hierarchical network structures, central actors have the authority to resolve conflicts at the periphery in part because they have a broader view of the organization and its environment. Hierarchies also allow higher-level actors to

Table 3 How networks enable organizing

Network attribute	*Function*
Formal ties	Formal relationships are official designations of the division of labor as well as the specification of authority and communication systems.
Informal ties	Boundary-spanning and other informal relationships complement formal ties to enhance control and coordination.
Dyadic interfirm ties	Dyadic interfirm linkages connect resources held by the respective organizations, creating a stock of relational assets.
Interlocking boards	Interlocking boards can serve four organizing functions: cooptation, resource access, control, and legitimation.
Multifirm ties	The main organizing benefits of the multifirm network are flexibility, the variety of capabilities that can be assembled, and the economies of scale and experience that can be leveraged in each activity.
Community	A community enables self-organization around shared goals and common resources. The community organizational model offers members the opportunity to collectively develop solutions and capabilities, allowing members to accomplish more than they could on their own.

control the goals and/or actions of subordinate actors in cascading principal–agent relationships. When the organization is strictly hierarchical, the network structure is centralized, with formal ties being mostly vertical. Less hierarchical organizations, such as those using matrix structures, create lateral ties to complement vertical ties.

Informal Ties

In addition to the formal organization, the informal organization is a supplementary mechanism that helps get tasks performed properly (Krackhardt and Hanson, 1993). Informal social networks are valuable in organizing the innovation process (Kleinbaum and Tushman, 2007). Boundary spanners (Tushman and Scanlan, 1981) – actors who operate at the boundaries of formal organizational units – help connect the organization to its environment by gathering, interpreting, and diffusing information within their informal networks.

Dyadic Interfirm Ties

A firm's critical resources, as well as the activities using and contributing to them, may span organizational boundaries. Dyadic interfirm linkages between boundary-spanning individuals connect resources held by their respective organizations, creating a stock of relational assets (Dyer and Singh, 1998). The primary function of dyadic ties is to increase the efficiency and effectiveness of network actors by linking their activities and leveraging their relational assets.

Interlocking Boards

Corporations are required by law to have a board of directors that oversees the management of the corporation on behalf of its owners and ensures legal and regulatory compliance. Boards frequently include senior executives and directors from other organizations. The linking of those individuals' multiple affiliations is referred to as board interlocks. Boards are networks of individuals whose ego networks constitute bridges to the other organizations of which they are members. As a result, multiple organizations are linked by way of the boards on which their members serve.

Interlocking boards can serve four organizing functions: cooptation, access to resources, control, and legitimation. One means of dealing with potentially disruptive elements in the environment is through cooptation. By assimilating external actors into the organization via membership on the board of directors, outside directors may come to share the organization's point of view and may even act on the organization's behalf. Organizations depend on access to resources from their environment, and they move proactively to manage such dependence. Interlocking boards provide the organization with relational assets, such as trust and shared knowledge, and board members may be able to provide direct access to needed resources. Interlocking boards are also a means of corporate control and monitoring. For example, bankers sit on corporate boards in order to protect capital at risk in financially troubled firms. Their presence on the board of directors allows them to be informed about the firm's financial status and actions being contemplated by management, and they may be able to influence or even control company actions that might affect their ability to recover loans. Finally, the composition of the board of directors affects the firm's legitimacy and reputation. Interlocking boards composed of high-status directors help legitimate the firm to potential sources of capital, labor, and other resources.

Multifirm Ties

Firms in a variety of industries have chosen to focus on their core activities and to outsource noncore activities to external providers. As a result, activities that could be performed by a single firm are instead performed by multifirm networks (Miles and Snow, 1986). The main organizing benefits of the multifirm network design are flexibility, the variety of capabilities that can be assembled, and the economies of scale and experience that can be leveraged in each activity. The typical multifirm network organization is also hierarchical, organized around a lead firm that works dynamically with its network providers to produce and deliver products and services (Dyer and Nobeoka, 2000). The design of a multifirm network consists of four key choices: forming the right types of external relationships, which companies to partner with, how to structure the relationship between partners and suppliers, and determining the kinds of supporting policies that are needed (Galbraith, 2014). By networking with external suppliers and partners, a few lead firms can organize an entire industry (Mathews, 2012).

Community

A community is a collective of actors who share common values and interests. As a means of organizing, the community provides a platform for the independent members to collaborate with one another by committing to a set of shared values and norms that exert moral suasion to regulate member behavior. By substituting resource commons, protocols, and infrastructure for hierarchical leadership and control, a community can be largely self-governing (Fjeldstad et al.,

2012). The community form of organizing is prevalent in knowledge-intensive sectors where the knowledge base underlying products and services is complex, growing, and widely diffused, requiring interorganizational collaboration to assure the full utilization of continuously developing knowledge (Powell et al., 1996). The community organizational model offers members the opportunity to collectively develop solutions and capabilities, thereby achieving a collaborative advantage.

Conclusion

Network is a valuable concept in the organization sciences. Among organization theories, network theory has been one of the most useful in advancing our understanding of organizations and their environments. Organizations are networks of formal and informal relationships as well as nodes in the larger societal networks in which they are embedded. As organizations and their environments continue to coevolve, network theory and analysis will be one of the major tools used to study them.

See also: Behavioral Economics; Boundaries and New Organization Forms; Business Platforms; Design and Form: Organizational; Industrial Cluster Dynamics; Innovation and Technological Change, Economics of; Interorganizational Relationships and Networks; Learning: Organizational; Modularity and Organizations; Network Analysis; Networks, Geography of; Organization: Overview.

Bibliography

Aral, S., Walker, D., 2012. Identifying influential and susceptible members of social networks. Science 337 (6092), 337–341.
Arthur, W.B., 1989. Competing technologies, increasing returns, and lock-in by historical events. Economic Journal 99 (394), 116–131.
Barabási, A.L., 2002. Linked: The New Science of Networks. Perseus Publishing, Cambridge, MA.
Borgatti, S.P., Foster, P.C., 2003. The network paradigm in organizational research: a review and typology. Journal of Management 29 (6), 991–1013.
Borgatti, S.P., Halgin, D.S., 2011. On network theory. Organization Science 22 (5), 1168–1181.
Brass, D.J., Galaskiewicz, J., Greve, H.R., Tsai, W., 2004. Taking stock of networks and organizations: a multilevel perspective. Academy of Management Journal 47 (6), 795–817.
Burt, R.S., 1992. Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge, MA.
Burt, R.S., 2004. Structural holes and good ideas. American Journal of Sociology 110 (2), 349–399.
Castells, M., 2010. The Rise of the Network Society, second ed. Wiley-Blackwell, Chichester, UK.
Cohen, W.M., Levinthal, D.A., 1990. Absorptive capacity: a new perspective on learning and innovation. Administrative Science Quarterly 35 (1), 128–152.
Coleman, J.S., 1990. Foundations of Social Theory. Harvard University Press, Cambridge, MA.
Cook, K., Emerson, R.M., 1978. Power, equity and commitment in exchange networks. American Sociological Review 43 (5), 721–738.
Dyer, J.H., Nobeoka, K., 2000. Creating and managing a high-performance knowledge-sharing network: the Toyota case. Strategic Management Journal 21 (3), 345–367.
Dyer, J.H., Singh, H., 1998. The relational view: cooperative strategy and resources of interorganizational competitive advantage. Academy of Management Review 23 (4), 660–679.
Fjeldstad, Ø.D., Snow, C.C., Miles, R.E., Lettl, C., 2012. The architecture of collaboration. Strategic Management Journal 33, 734–750.
Galbraith, J.R., 2014. Designing Organizations, third ed. Jossey-Bass, San Francisco, CA.
Granovetter, M.S., 1973. The strength of weak ties. American Journal of Sociology 78, 1360–1380.
Granovetter, M.S., 1978. Threshold models of collective behavior. American Journal of Sociology 83 (6), 1420–1443.
Granovetter, M.S., 1985. Economic action and social structure: the problem of embeddedness. American Journal of Sociology 91 (3), 481–510.
Johnson, J., 2014. Hypernetworks in the Science of Complex Systems. Imperial College Press, London, UK.
Katz, M.L., Shapiro, C., 1985. Network externalities, competition and compatibility. American Economic Review 75 (3), 424–441.
Katz, M.L., Shapiro, C., 1986. Technology adoption in the presence of network externalities. Journal of Political Economy 94 (4), 822–841.
Kleinbaum, A.M., Tushman, M.L., 2007. Building bridges: the social structure of interdependent innovation. Strategic Entrepreneurship Journal 1 (1/2), 103–122.
Kogut, B., 2000. The network as knowledge: generative rules and the emergence of structure. Strategic Management Journal 21 (3), 405–425.
Krackhardt, D., Hanson, J.R., 1993. Informal networks. Harvard Business Review 71 (4), 104–111.
Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago, IL.
Lazer, D., Friedman, A., 2007. The network structure of exploration and exploitation. Administrative Science Quarterly 52 (4), 667–694.
March, J.G., 1991. Exploration and exploitation in organizational learning. Organization Science 2 (1), 71–87.
Martin, X., Salomon, R., 2003. Knowledge transfer capacity and its implications for the theory of the multinational corporation. Journal of International Business Studies 34 (4), 356–373.
Mathews, J.A., 2012. Design of industrial and supra-firm architectures: growth and sustainability. Journal of Organization Design 1 (2), 42–63.
Miles, R.E., Snow, C.C., 1978. Organizational Strategy, Structure, and Process. McGraw-Hill, New York.
Miles, R.E., Snow, C.C., 1986. Organizations: new concepts for new forms. California Management Review 28, 62–73.
Milgram, S., 1967. The small world. Psychology Today 2 (1), 60–67.
Powell, W.W., Koput, K.W., Smith-Doerr, L., 1996. Interorganizational collaboration and the locus of innovation: networks of learning in biotechnology. Administrative Science Quarterly 41 (1), 116–145.
Rogers, E.M., 1962. Diffusion of Innovations. The Free Press, New York.
Shapiro, C., Varian, H., 1999. Information Rules: A Strategic Guide to the Network Economy. Harvard Business School Press, Boston, MA.
Simmel, G., 1950. The Sociology of Georg Simmel. The Free Press, New York.
Szulanski, G., Cappetta, R., Jensen, R.J., 2004. When and how trustworthiness matters: knowledge transfer and the moderating effect of causal ambiguity. Organization Science 15, 600–613.
Tushman, M.L., Scanlan, T.J., 1981. Boundary spanning individuals: their role in information transfer and their antecedents. Academy of Management Journal 24 (2), 289–305.
Uzzi, B., 1996. The sources and consequences of embeddedness for the economic performance of organizations: the network effect. American Sociological Review 61 (4), 674–698.
Uzzi, B., Spiro, J., 2005. Collaboration and creativity: the small world problem. American Journal of Sociology 111 (2), 447–504.
Watts, D.J., 1999. Networks, dynamics, and the small-world phenomenon. American Journal of Sociology 105 (2), 493–527.
Watts, D.J., 2004. The "new" science of networks. Annual Review of Sociology 30, 243–270.
White, H.C., 1963. The Anatomy of Kinship. Prentice-Hall, Englewood Cliffs, NJ.

Networking for School Improvement

Steven Katz, Ontario Institute for Studies in Education/University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Networks of teachers, administrators, schools, and districts have become a pervasive mechanism in education for reaching beyond traditional boundaries. Some networks are formally organized around a common goal, whereas others unite schools into a common organizational structure. Some networks are intentional in their origins and organization, and some are informal, arising in spontaneous ways. Although many positive characteristics are attributed to networks, the question of how networks 'work' has been difficult to answer. This article takes up that question, with an appeal to a research-based theory of action that elucidates how successful networks function as innovative knowledge communities.

Why Networked Learning Communities?

For decades, numerous school improvement models have attempted to reform the thinking and practices of practitioners with the explicit intent of increasing student learning and achievement in schools. Introducing reforms into classrooms and schools has generally accomplished superficial changes to practices that have not easily translated into sustainable improvement for student success (Hargreaves, 2003). In an effort to intentionally create the level of deep learning necessary for practitioners to make meaningful changes in their classrooms, professional networks are increasingly being promoted as mechanisms for knowledge creation and mobilization, that can lever the kinds of changes that make a difference for students.

In a knowledge society, lifelong learning is a cornerstone of the flexibility necessary for highly skilled and educated people to take on novel tasks and continuously adapt to new and changing environments. As we exit the industrial age, characterized by a 'finite' conception of resources, a 'controllable' conception of information, and a 'sequential and task-specific' conception of learning, the notion of networks takes on increased relevance (Allen and Cherrey, 2000). Specifically, networks provide an operational construct for educational provision and a new vehicle for achieving change.

In this knowledge society, practices for facilitating knowledge creation and mobilization are considered to be key tenets of educational provision. Knowledge is the most critical resource for social and economic development (Hakkarainen et al., 2004). Change-directed improvement comes in the form of creating new knowledge or adding value to existing knowledge rather than simply appropriating existing knowledge resources. A fundamental challenge for education then is to organize work with knowledge in a way that facilitates ongoing knowledge building and sharing among members of the community. As Hakkarainen et al. (2004) explain, members of the community need to develop competencies that allow them to function as 'knowledge workers.'

In *Working Laterally*, Hargreaves (2003) describes the demands of knowledge creation (and its supportive competencies) in terms of innovation. Knowledge creation (or transformation) is – in a word – innovation. Innovation, for teachers, is about learning to work differently in order to work better. For teachers, most innovation is the creation of new professional knowledge about their work.

Historically, the route to understanding (and supporting) innovation has been subject to individualistic bias. This bias has paid more attention to mental processes and events at the level of individual agents than to activities that are distributed across several agents in particular contexts and are dependent on characteristics of social organization. Alternatively, many proponents of the knowledge society focus almost exclusively on the organizational level and examine those organizational practices that elicit knowledge creation to the exclusion of cognitive competencies, without which participation in this society is impossible. Hakkarainen et al. (2004) argue that both are necessary. Networks, as a dynamic organizational form that can mediate between the personal and social worlds, have the potential to capture the complex and reciprocal relationship between individual and collective competencies.

How Do Networked Learning Communities 'Work?'

The question of function, or how networks 'work,' has been difficult to answer. At the broadest level, they support diverse goals, like broadening opportunities and sharing resources (Muijs et al., 2010), but the route to school improvement is not always clear. Judith Chapman and David Aspin (2003) suggest the following possible pathways of function:

- Networks offer a means of assisting in the policy implementation process by linking policy both horizontally and vertically.
- Networks provide a process for cultural and attitudinal change, embedding reform in the interactions, actions, and behavior of a range of stakeholders.
- Networks provide for an opportunity for shared and dispersed leadership and responsibility, drawing on resources in the community beyond education.
- Networks can be capacity building insofar as they are able to produce new knowledge and mutual learnings that can then offer feedback and inform public policy.
- Networks can move attention away from a preoccupation with micro-level change at the individual site and function at the meso-level to strengthen interconnections and spread innovation across all levels – micro, meso, and macro.

The educational landscape is populated by networks with a multiplicity of variations. In England, government sponsored networks developed to encourage and support continuous cooperative learning at all levels of the education system. England's Networked Learning Communities (NLC) program was the largest government sponsored school-to-school networking initiative of its kind in the world. In the United States, the growth of organic networks of teachers and administrators has taken place over the past 30 years without a formal government-sponsored infrastructure. Some networks join teachers and/or schools together on a national level. The National Writing Project, for example, is a teachers' national network of writing (Lieberman and Wood, 2002). Other networks are bound by jurisdiction, such as the Consortium for Educational Change, a network of school districts in Illinois, created to improve student achievement by assisting member districts and schools to become collaborative, high-performing organizations. Likewise, the Bay Area School Reform Collaborative (BASRC) is a network of schools in the San Francisco Bay Area that collaborates to achieve equity-minded school reform (Center for Research on the Context of Teaching, 2003).

Understanding networked learning communities is best accomplished from a research-based and empirically-validated theory of action perspective that has to begin with the end in mind (Katz and Earl, 2010). Rather than networking for the sake of networking, the explicit and intended impact of networked learning communities is to enhance student learning and engagement (the ultimate outcome), the antecedent of which is a change in teacher practice (an intermediate outcome). Changed teacher practice, in turn, is preceded by a protocol of knowledge creation and sharing (intermediate outcome). It is here – knowledge creation and sharing – that the collaborative underpinnings of networked learning communities as a dynamic organizational form find expression and effect.

The task of 'enabling' knowledge creation and sharing can be conceptualized as the outgrowth of a community of practice (COP) (Veugelers and O'Hair, 2005) or – to better underscore the key ingredient of intentionality – an innovative knowledge community (IKC) (see Paavola et al., 2004, for a description). That is, while networks differ in the way they go about doing their work, they share the general function of bringing participants to build IKCs. In 'theory of action' terms, networks, as collaborative organizational forms, support particular ways of working that find expression as these IKCs within two distinct organizational units – the network itself and the schools that are represented.

Several studies have found positive associations between the presence of COPs (or IKCs) within particular schools and the schools' success in enhancing student engagement and achievement (Bryk and Schneider, 2002; McLaughlin and Talbert, 2001). Research (Katz and Earl, 2010) suggests that networks can foster the development of school IKCs by: (1) linking newly emerging ones to their more developed counterparts and (2) creating new IKCs at the network level that enable knowledge creation and sharing. What becomes important then is the strength of attachment between schools and networks such that established school-level IKCs can upload their key feature practices into the network – thus creating a network IKC – that in turn allows other school IKCs to download and instantiate these enablers of knowledge creation and sharing.

What connects schools to networks (and networks to schools) are, ultimately, people – whether it be directly through active participation or through the construction of artifacts that serve boundary spanning functions. The study of individuals as the connective tissue of networks takes place at the individual level and is typically accomplished through social network analysis (see Hite et al., 2005, for an example). Social network analysis considers the nature and roles of particular individuals as they connect to others, both within and across organizational boundaries. But beyond such social interactions, a networked learning theory of action needs to consider the organizational (school and network) products (in terms of IKC practices that *enable* knowledge creation and sharing) of these interactions.

What Are the 'Key Enablers' of Successful Networked Learning Communities?

Research (e.g., Katz and Earl, 2010) has identified six key *enablers* of successful networked learning communities, each of which operates within the schools and the network: *purpose and focus*, *relationships*, *collaboration*, *inquiry*, *leadership*, and *capacity building and support*. Each of these constitutes a complete area of inquiry, with an associated research base, on its own. Networked learning communities 'work' because knowledge creation and sharing are enabled by a constellation of practices defined by these enablers. While a fulsome treatment of each is beyond the scope of this article, below is a brief initial orientation to each of the enablers.

Purpose and Focus

Having a fundamental and clear organizational purpose is critical to the success of networked learning communities. A learning focus is likely to have a more direct impact if it is focused in ways that are concrete and useful (Timperley and Robinson, 2003); high leverage in fostering student learning (Marzano et al., 2001); compelling, challenging, and shared (Lieberman and Grolnick, 1996; Bryk et al., 1999); and appropriate to the context (Marzano et al., 2001). A compelling and high leverage learning focus is based on evidence that it can have significant impact on teaching practices and student learning. A challenging focus is one that requires teachers to reconceptualize, unlearn, or make changes to existing practice and structures, legitimating the change process by making the status quo more difficult to protect, and offering opportunities for joint attention to issues that are larger than any one school could address alone (Timperley, 2004). Finally, the learning focus should be 'right' for the participating schools, given their particular context, history, and needs.

Relationships

Relationships are the 'connective tissue' of networked learning communities (Allen and Cherrey, 2000) and provide the social capital that allows people to work together over time and exceed what any of them could accomplish alone

(West-Burnham and Otero, 2004). Relationships create a common language and a sense of shared responsibility, provide channels for communicating and disseminating information to one another about network members' expertise, and develop readiness to trust one another (West-Burnham and Otero, 2004; Veugelers and O'Hair, 2005).

Trust is a key condition of productive relationships. Indeed, Bryk et al. (1999) found that social trust among members of a group was the strongest facilitator of professional community. They propose that a base level of such trust may be necessary for a professional community to emerge, but working and reflecting together can also build trust and strengthen relationships. In relationships, conflict is inevitable – and indeed valuable – but robust and trusting relationships among network members can allow them to work together even when they have different orientations and views (Lieberman and Grolnick, 1996).

Collaboration

Collaboration encompasses much more than relationships. It is intensive interaction that engages educators in opening up their beliefs and practices to investigation and challenge. When colleagues engage in a dynamic process of interpretation and evaluation of practice, they enhance their own practice and that of the profession. This kind of collaboration allows people to address tough problems of teaching (Firestone and Pennell, 1997), build commitment through group understanding (Lieberman and Grolnick, 1996), solve issues of mutual concern (Wohlstetter and Smith, 2000), and spread innovations beyond individuals and single sites (Smith and Wohlstetter, 2001).

Collaboration can be a powerful mechanism for changing ideas and practices, particularly when it involves joint work that includes a balance of personal support with critical inquiry about present practice and future direction (Borko, 2004) and sustained scrutiny of practice, but is not always easy. In fact, moderate conflict is essential for the development of high joint benefit, and the desire to avoid conflicts can undermine this outcome (Engestrom, 1999).

Inquiry

Systematic analysis and professional reflection are regarded as core activities for both individual and collective construction of meaning, and are hallmarks of an 'inquiry habit of mind' (Earl and Katz, 2006). Knowledge creation, especially when people are involved in changing their practice, requires that individuals consider explicit knowledge and share, question, and possibly adapt their tacit knowledge, in order to create new collective explicit knowledge. Little (2005) references a large body of research suggesting that conditions for improving learning and teaching are strengthened when teachers collectively question ineffective teaching routines, examine new conceptions of teaching and learning, find generative means to acknowledge and respond to difference and conflict, and engage actively in supporting one another's professional growth. The inquiry processes of questioning, reflecting, seeking alternatives, and weighing consequences promote the 'transparency' of what otherwise might remain unobservable facets of practice, making tacit knowledge visible and open to scrutiny.

Collaborative inquiry creates an opportunity for educators to work together searching for and considering various sources of knowledge (both explicit and tacit) in order to investigate practices and ideas through a number of lenses, to put forward hypotheses, to challenge beliefs, and to pose more questions. When educators have an inquiry habit of mind, they have developed a way of thinking that is a dynamic iterative system for organizing ideas, seeking out information, and moving closer and closer to understanding some phenomenon together.

Leadership

Networked learning communities include many levels of leadership – both formal and informal. Networks need some system of leadership to direct the work of the network itself, which usually coexists alongside the formal leadership of administrators in schools. Although the leadership literature continues to emphasize the role of principals in successful change and instructional improvement, leadership models are increasingly focusing on what Rowan (1990) called 'network' patterns of control, where leadership activities are distributed across multiple people (Heller and Firestone, 1995). Formal leaders (e.g., principals) provide leadership by encouraging and motivating others, setting and monitoring the agenda, sharing leadership, providing support, and building capacity. At the same time, networked learning communities encourage distributed leadership in schools and across the network, with many people with and without formal positions of authority providing a range of leadership functions such as leading particular initiatives, participating in collaborative groups, supporting colleagues learning, and sharing their knowledge with others.

Building Capacity and Support

Harris (2001) defines capacity building as being concerned with creating the conditions, opportunities, and experiences for collaboration and mutual learning. Years of school improvement research have shown that improving schools are ones that take charge of change, rather than being controlled by it (Stoll and Fink, 1996). As Senge (1990) describes it, a learning organization is one that is continually expanding its capacity to create its future. In networked learning communities the focus is on creating the conditions to support individual and collective learning through all of the processes described in the previous key enablers. Building capacity depends on intentionally fostering and developing the opportunities for members to examine their existing beliefs and challenge what they do against new ideas, new knowledge, new skills, and even new dispositions. When networks are focused on learning, they intentionally seek out and/or create activities, people, and opportunities to push them beyond the status quo within their school and network developments.

Putting It All Together

In summary, networked learning communities represent complex interactions between structures and activities that are

intentionally dedicated to creating new knowledge and sharing it broadly. A number of features have been shown to create the conditions that foster knowledge creation. When educators participate in focused relationships for collaborative inquiry, they open themselves to new ideas and to challenging their views through interaction with others and with formal knowledge. Leaders, both formal and distributed across roles, can provide vision, accountability, and capacity building opportunities for engagement in the activities that move professional learning and knowledge creation forward.

See also: Communities of Practice; Continuing Professional Development; Educational Effectiveness, The Field of; Equity and Education; School Effectiveness Research; School Improvement; School Management.

Bibliography

Allen, K., Cherrey, C., 2000. Systemic Leadership: Enriching the Meaning of Our Work. University Press of America, Boston.

Borko, H., 2004. Professional development and teacher learning: mapping the terrain. Educational Researcher 33 (8), 3–15.

Bryk, A., Schneider, B., 2002. Trust in Schools: A Core Resource for Improvement. Russell Sage Foundation, New York.

Bryk, A., Camburn, E., Louis, K.S., 1999. Professional community in Chicago elementary schools: facilitating factors and organizational consequences. Educational Administration Quarterly 35, 751–781.

Center for Research on the Context of Teaching, 2003. Bay Area School Reform Collaborative Summary Report: Phase I: 1996–2001. Author, Stanford, CA.

Chapman, J., Aspin, D., 2003. Networks of learning: a new construct for educational provision and a new strategy for reform. In: Davies, B., West-Burnham, J. (Eds.), Handbook of Educational Leadership and Management. Pearson, London, pp. 653–659.

Earl, L., Katz, S., 2006. Leading in a Data-Rich World: Harnessing Data for School Improvement. Corwin Press, Thousand Oaks, CA.

Engestrom, Y., 1999. From iron cages to webs on wind: three theses on themes and learing at work. Lifelong Learning in Europe 4 (2), 101–110.

Firestone, W., Pennell, J., 1997. State-initiated teacher networks: a comparison of two cases. American Educational Research Journal 34 (2), 237–268.

Hakkarainen, K., Palonen, T., Paavola, S., Lehtinen, E., 2004. Communities of Networked Expertise: Professional and Educational Perspectives. Elsevier, Amsterdam.

Hargreaves, D., 2003. Working Laterally: How Innovative Networks Make an Education Epidemic. Demos/NCSL, London.

Harris, A., 2001. Building the capacity for school improvement. School Leadership and Management 21 (30), 261–270.

Heller, M., Firestone, W., 1995. Who's in charge here? Sources of leadership for change in eight schools. Elementary School Journal 96 (1), 65–86.

Hite, J., Williams, E., Baugh, S., 2005. Multiple networks of public school administrators: an analysis of network content and structure. International Journal of Leadership in Education 8 (2), 91–122.

Katz, S., Earl, L., 2010. Learning about networked learning communities. School Effectiveness and School Improvement 21, 27–51.

Lieberman, A., Grolnick, M., 1996. Networks and reform in American education. Teachers College Record 98 (1), 7–45.

Lieberman, A., Wood, D., 2002. Inside the National Writing Project: Connecting Network Learning and Classroom Teaching. Teachers College Press, New York.

Little, J.W., 2005. Nodes and Nets: Investigating Resources for Professional Learning in Schools and Networks. Unpublished Paper for National College for School Leadership, Nottingham, UK.

Marzano, R., Pickering, D., Pollock, J., 2001. Classroom Instruction that Works: Research-based Strategies for Increasing Student Achievement. Association for Supervision and Curriculum Development, Alexandria, VA.

McLaughlin, M., Talbert, J., 2001. Professional Communities and the Work of High School Teaching. University of Chicago Press, Chicago.

Muijs, D., West, M., Ainscow, M., 2010. Why network? Theoretical perspectives on networking. School Effectiveness and School Improvement 21, 5–26.

Paavola, S., Lipponen, L., Hakkarainen, K., 2004. Models of innovative knowledge communities: three metaphors of learning. Review of Educational Research 74 (4), 557–576.

Rowan, B., 1990. Commitment and control: alternative strategies for the organizational design of schools. Review of Research in Education 16, 353–392.

Senge, P., 1990. The Fifth Discipline: The Art and Practice of the Learning Organization. Doubleday, New York.

Smith, A., Wohlstetter, P., 2001. Reform through school networks: a new kind of authority and accountability. Educational Policy 15 (4), 499–519.

Stoll, L., Fink, D., 1996. Changing Our Schools: Linking School Effectiveness and School Improvement. Open University Press, Buckingham, UK.

Timperley, H., 2004. Situating Networked Learning Communities in International Research Coherence and Networked Learning Communities: A Distributed Leadership Perspective. Paper commissioned by Aporia Consulting, Toronto, ON.

Timperley, H., Robinson, V., 2003. Workload and the professional culture of teachers. In: Kydd, L., Anderson, L., Newton, W. (Eds.), Leading People and Teams in Education. Sage, London, pp. 151–168.

Veugelers, W., O'Hair, M., 2005. Networking for learning and change. In: Veugelers, W., O'Hair, M. (Eds.), Network Learning for Educational Change. Open University Press, London, pp. 211–221.

West-Burnham, J., Otero, G., August 2004. Educational Leadership and Social Capital. In: Incorporated Association of Registered Teachers of Victoria Seminar Series no. 136.

Wohlstetter, P., Smith, A., March 2000. A different approach to systemic reform: network structures in Los Angeles. Phi Delta Kappan 81, 508–515.

Networks and Cultural Consumption

Mark C Pachucki, Harvard Medical School, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Scholars of culture have been increasingly concerned with the roles played by interpersonal social networks in relational cultural processes. This article highlights the ways that social network composition and dynamics can shape individuals' cultural consumption choices and the ways that consumption can shape the structure of individuals' social networks. Attention is given to symbolic aspects of cultural consumption as well as practical aspects of consumption. While much of the research on this topic emerges from the sociology of culture, a broader focus illustrates productive links with health-based and cognitive perspectives.

Definition and Intellectual Context

Scholars of culture have been increasingly concerned with the roles played by interpersonal social networks in relational cultural processes (for recent reviews, see Mische, 2011; Pachucki and Breiger, 2010). In this article, the term 'social networks' refers to the web of human associations and interactions that constitute an individual's social group (Breiger, 2009; Scott, 2000). The term 'consumption' largely refers to cultural consumption – the ways that individuals broadly enact their cultural tastes and leisure-time activities (i.e., fashion, music, art, television and film-viewing, reading, religious attendance, political engagement, and sporting activities) (Bourdieu, 1984). However, consumption can also refer to physiological ingestion (i.e., food and other substances) as a type of cultural consumption.

This article highlights the ways that social network composition and dynamics can shape individuals' cultural consumption choices, and to a lesser degree it also documents the idea that consumption choices can shape the structure of individuals' social networks. A brief discussion of the history of this topic is followed by a review of current theory and research, concluding with an identification of challenges and future directions. Because of space limitations, this article is illustrative of key works rather than exhaustive of the corpus of growing literature. Most attention is accorded to research that empirically measures some aspect of social network structure and attempts to discern the relationship between social ties between individuals and these same individuals' cultural consumption practices.

Though research in this area is centered primarily around topics of musical taste and arts participation in the sociology of culture, there have been other emergent clusters including the ingestion of food and drink as a form of cultural consumption that affects health; different forms of cultural consumption involving online and offline social media; and consumption of knowledge and the academic sphere. By bringing these conceptually related but often separate intellectual contexts into dialogue with one another, we might identify productive areas for investigation in this growing subfield.

History

The idea that cultural consumption is shaped by forms of socialization has early roots in the work of Thorsten Veblen (2007[1899]), whose well-known work on lifestyle habits generated the phrase 'conspicuous consumption' to refer to the practice of signaling social status through one's consumption choices. Yet more formal discussion of cultural consumption and social network interaction would not follow for some time. It could be argued that some of the earliest sociologists to focus informally on networks in the context of cultural consumption were Harrison White and Cynthia White (1965) in their research on the contexts of artistic creation in French salon society, as well as Howard Becker's (1982) classic treatise on art worlds. Neither text employs formal measurement of social networks. However, the authors' discussions of how artist socialization shapes artwork and audiences clearly evince a sympathetic understanding of how network processes shape cultural consumption, and vice versa.

The oeuvres of both Pierre Bourdieu and Richard Peterson have been highly influential in sociologists' understanding of how cultural consumption is shaped by social status position. Bourdieu (1984) extensively investigated cultural consumption and lifestyle, and his ideas of 'field' could be considered a distant relative of the idea of 'social network' in that he envisioned a highly connected, mutually reinforcing 'field of action' where human and institutional actors and their dynamic relational circumstances constantly re-molded the contours of a domain of activity (Bourdieu, 1993). Peterson and colleagues' explorations of how high-status individuals differentiate themselves through their breadth of consumption – the concept of 'omnivorousness' – offers an alternative mechanism of cultural reproduction (e.g., Peterson and Kern, 1996; Peterson and Simkus, 1992). Like the earlier relational tendencies of White & White and Becker, neither Peterson nor Bourdieu operationalize the idea of social networks beyond their theoretical frames. Regardless, the tension between Bourdieu's and Peterson's visions of how cultural inequality is reproduced has been fruitful for paving the way for scholars to investigate possible roles for network mechanisms. For example, Lizardo (2013) finds that individuals who engage in

a broad palette of cultural participation are more likely to make use of their network contacts to find a job.

Current Theory and Research

Social Networks, Cultural Participation, and Tastes

Investigating possible links between network diversity (or heterogeneity) and symbolic cultural consumption has been a fruitful line of inquiry for scholars. One of the earliest studies to empirically analyze network affiliation and cultural consumption in a rigorous fashion is Erickson's (1996) study of genre knowledge of security workers in Toronto. To assess genre knowledge, Erickson probed sports, art, books, restaurant, and magazine consumption; for network variety, the author asked respondents if they knew someone working in 19 occupation categories, and whether that person was kin or a friend. Greater network diversity – defined as having contact with people in different occupation class locations – was linked with greater variety in forms of cultural consumption. In a similar vein, Relish (1997) studies music genre preferences using the 1993 US General Social Survey (GSS). He measures network complexity by counting the number of associations to which a respondent belongs, geographic mobility since youth, and community type. He finds that a strong association between education and breadth of genre preference is diminished with the introduction of network information, suggesting that networks play a significant role in the development of musical preferences.

In their research on general cultural participation, Warde and Tampubolon (2002) describe trends in association membership and frequency of leisure-time activities using two waves of the British Household Panel Survey spanning 1997–99. Using association membership as a proxy for social capital, they report on cross-tabulations and find group membership to be associated with greater frequency of leisure-time activities. The authors also construct a measure of social distance between best friends using similarity in sex, age, and location and observe that having a highly similar best friend increases the likelihood of recreational activity of various types. By enumerating personal networks of university students, Kane (2004) discovers a gendered dimension to cultural participation in a range of activities. In contrast to earlier studies, Kane's questions to respondents were designed to elicit information about specific network alters, following a more sophisticated network design. By calculating respondents' network density and heterogeneity measures of race, religion, and gender in one's set of reported alters, she reveals that network density is positively associated with attending sports events and negatively associated with museum and gallery attendance. Greater network heterogeneity increased the likelihood of high culture participation, and network effects were greater for women than men. In contrast to the prior research that seeks to understand how networks shape cultural consumption, Lizardo (2006) employs indicators of cultural participation to explore the opposite possibility: that cultural tastes also shape network structure. Using the 2002 module of the GSS, he examines personal network size and tie strength to show that highbrow forms of consumption lead to larger networks of strong ties, while popular consumption leads to larger networks of weak ties.

The analysis of music tastes as a proxy for consumption habits has been productive for many network scholars. Like Relish (1997), Schultz and Breiger (2010) use the 1993 GSS to examine musical genres liked by respondents. However, their starting point is the idea that theories involving the strength of social ties, and sociability more generally, should take into account cultural objects and relations. The authors posit that strong ties emerge from 'weak culture' – from venues of small talk, from tentative efforts to engage common ground with others, and from the efforts required to maintain a relationship. They find that the more music genres a respondent 'likes rather than loves' (their measure of weak culture), the less likely are they to perceive Americans as generally fragmented in their values, compared with those respondents who like fewer genres. In an innovative examination of how network social influence shapes musical tastes, Salganik et al. (2006) developed an online experiment in which participants downloaded and rated music while in some cases aware of others' online expressed preferences. The authors find that knowledge of others' preferences increased inequality in song ratings: knowledge of others' preferences translated into popular songs becoming more popular while unpopular songs would be even more unpopular. Finally, Crossley (2009) merges historical accounts of the Manchester punk music scene between 1976 and 1980 with network mapping techniques to trace connections between key social actors, among them musicians, industry figures, record labels, and fan clubs. Using qualitative historical data to give depth to purely structural accounts of network dynamics, Crossley identifies shared interests, the propensity for redundant information among those with strong ties, and the diffusion of reputation as relevant mechanisms.

Longitudinal research on networks and cultural consumption has been rare due to intense data and analytic demands, though this is quickly changing. A promising line of research involves a class of actor-oriented statistical models that seek to discern mechanisms responsible for observed network structures and consumption patterns using simulation techniques. One example is the adolescent preference research of Steglich et al. (2006), who use data on friendship ties, substance use, and music genre preferences collected from a cohort of Scottish secondary school students. They find that certain types of consumption behaviors shape friendship patterns. More specifically, listening to rock music increases one's popularity in the network, classical music listeners tend to affiliate with one another, and those who drink alcohol cluster together in the network. In a study of the preferences of college Facebook users at an American university, Lewis et al. (2012) examine the relationship between Facebook friendships and music, movie, and literature consumption preferences. In contrast to prior research that examines genre preferences in culture consumption, the authors were able to analyze respondent similarity in the much more precise level of performer preferences (in the case of music), specific movie titles (in the case of film), and specific authors (in the case of literature). They model network and behavioral coevolution using an actor-based approach in order to separate social selection and network influence. For music taste selection effects, they find that 'lite/classic rock' and 'classical/jazz' fans form and maintain friendships with one another more than their peers. For film selection effects, they

observe that 'dark satire' and 'raunchy comedy/gore' viewers are more likely to be friends. Yet of all the possible consumption behaviors, only classical/jazz listeners are likely to influence others to adopt that taste. An overview of the authors' broader study design, scope, and data-gathering approach is described elsewhere (Lewis et al., 2008).

Social Networks, Media Usage, and Knowledge Consumption

Comparative studies of different types of media usage are key themes for a range of scholars across different national contexts. Cardon and Granjon (2005) examine personal networks of middle-class students in France, motivated by the paradox that while studies of cultural practices often involve measurement of collective activities, these same studies rarely explore relational context, instead treating actors as autonomous individuals. They begin with ego-network enumeration of a small sample of young men and women, followed by interrogation of their media consumption as well as whom they engage in these activities with. Their descriptive network analysis – and importantly, reports of discussions of the resulting sociograms with the original respondents – gives a rich sense of different types of cultural consumers. They find three general types of young adult media users: those who engage in cultural activities and keep their groups of alters separate, those who engage with different sets of alters and spread their tastes across these various groups, and those who maintain a core network based on cultural activities and who draw other alters into that core. In contrast, Kraaykamp et al. (2007) focus on the dyad as a precursor to network processes and examine how a person's social status and their significant other's social status affect media consumption – specifically, book reading and television watching. Relying upon multivariate analysis of a large longitudinal survey of Dutch adults, they discover a significant relationship between the partner's status and ego's media consumption. Those with partners of higher status tend to consume more literature, while those with partners of lower status consume more television.

A growing number of scholars are also beginning to investigate the roles that the Internet plays in our cultural consumption habits. Kayahara and Wellman (2007) conducted a study of information search behavior in order to compare the online and offline ways in which individuals learn about cultural activities. Using in–home interview data from Toronto residents, the authors find that interpersonal relationships are the primary source for learning about activities but that the Internet functions to enhance and confirm their understandings. Information search across different modes of interaction is also a concern for Tepper and Hargittai (2009), who study methods of discovery about new music by college students. Through fielding surveys to students at three different universities, they compare the relative efficacy of interpersonal social networks, mainstream media, and digital media in informing new music consumption. The authors find that music listeners with omnivorous tastes are nearly four times more likely to learn about new music from people in their social network as opposed to the other pathways, and that Internet methods of discovery tend not to be disproportionately influential in shaping music discovery.

Another emerging area of research concerns the university as a site for cultural consumption. Research in this area investigates the ways that academic socialization processes shape students' experiences. For example, Kane (2011) seeks to understand how the transition from high school to college affects students' personal network composition. Using a combination of surveys and interviews of incoming students at an elite university at arrival and 1 year later, she examines the density of personal networks, as well as various types of network heterogeneity (ethnic, gender, regional, and religious). Over time, Kane finds network density decreased and network heterogeneity of all types increased. For the young men in her sample, group-based affiliations (such as sports teams) formed the primary basis for social tie development, while women were more likely to spend time cultivating relationships outside a formal group structure. Diversity is also a central interest of Benediktsson (2012), who uses the National Longitudinal Study of Freshmen to study how cultural consumption is related to the racial heterogeneity of incoming college students' social networks at 27 universities. Through the analysis of prior and current arts, religious, political, and sports participation, he finds that cross-racial contact appears to vary by the type of cultural activity. More specifically, religious group participation is associated with higher levels of racial homophily than the other cultural activities, though students are more likely to encounter students of different races (heterophily) through affiliating with arts groups. In a very different type of investigation concerning both the consumption and production of knowledge in academia, Rawlings and Bourgeois (2004) use the network concept of structural equivalence to measure how categories of specialization emerged in the field of agriculture. By combining a historical approach to tracing the development of academic credentialing processes with network tools, they show how symbolic boundary processes and organizational dynamics shaped agriculture as an academic field.

Social Networks and Physiological Consumption (Ingestion)

A much more literal definition of consumption that has rarely been integrated with discussions of symbolic forms of cultural consumption is the voluminous health literature on food consumption and substance use. Here, reported ingestion of certain food items or beverages and miscellaneous drug-related behaviors are often analyzed as risky behaviors associated with poor health status. Scholars researching these topics are beginning to productively identify how certain of these forms of consumption are patterned in social networks.

Alcohol consumption has been a recurrent theme of investigation for network scholars. Among the earliest pieces on this topic is a demographic study of Norwegians' drinking during the nineteenth and twentieth centuries. Skog (1986) tests a hypothesis that long waves of cyclical alcohol consumption may be recurrent throughout history using time-series analyses of trends. Finding some evidence of this cyclical pattern in the time period evaluated, he formally models network interactions to evaluate how microlevel interactions shape chains of indirect influence and changes in drinking culture. In a study of the social contexts of adults' risky health behaviors, Latkin et al. (1996) examined the influence of peers'

alcohol consumption on egos' drug-related risky behaviors (injectable cocaine, injectable heroin, crack cocaine, alcohol, having multiple sexual partners, and having casual sexual partners). Peer alcohol consumption was observed to be associated with egos' drinking behaviors, as well as egos' sexual behaviors.

In a large-scale alcohol consumption study using the National Longitudinal Study of Adolescent Health (Add Health), Gaughan (2006) argues that gender effects are not well understood in peer influence models of adolescent health behavior. The author examines social influence in alcohol consumption of seventh to twelfth graders between two waves of the study, evaluating several types of possible influence – general 'friend to friend' influence, and more specific 'female to female,' 'male to male,' 'female to male,' and 'male to female' directions. While Gaughan finds mutual influence among same-sex friends, in mixed-gender friendship there is evidence of only boy to girl influence, not the reverse. In another notable study based on the Add Health cohort, Kreager and Haynie (2011) specify actor–partner interdependence models in a longitudinal setting to help disentangle selection from influence. While the authors primarily focus on the effect of romantic partners' drinking on an ego's drinking, they also consider the role played by ego's friends, and that of the friends of romantic partners as well. Their findings suggest that partners' drinking is associated with ego's future binge drinking, and that the drinking behaviors of the friend group of a partner is independently associated with ego drinking. While couples tend to become similar in their drinking behaviors over time, there is little evidence of assortative mating with respect to drinking among adolescents. In a large network study of alcohol consumption among adults, Rosenquist et al. (2010) explore how alcohol consumption spreads through relationships over time in a social network. Using data from the Framingham Heart Study, a well-known prospective cohort study designed to study cardiovascular risk, the researchers discover evidence of spread that extended to three degrees of separation from ego. Additionally, they find that female alters are more likely than male alters to influence alcohol consumption, and that while being surrounded by heavy drinkers increases an ego's likelihood of being a heavy drinker, being tied to abstainers also increases an ego's likelihood of abstaining.

Research on food consumption and networks across the social sciences illustrates how eating is patterned among socially connected individuals. Psychological research on dyadic similarity in eating preferences has shed important light on networks considered as chains of dyadic interactions. In one example, Rozin et al. (2004) examined dyadic similarity in food preferences among peers in two different age groups – elementary school children and freshman college roommates – in order to assess the likelihood of peer influence. In order to discount the role of shared culture, the authors compare food preferences of actual pairs of friends with randomly assigned friends. Across both age tiers, the authors find little evidence of mutual influence beyond what would be expected in same-gender randomly assigned peers. In a longitudinal study of several thousand adults from the Framingham Heart Study, Pachucki et al. (2011) examined similarity in self-reported food consumption among peers of different relationship types – spouses, siblings, and friends – while controlling for network position in the greater social network. The authors find seven different eating patterns in the population, and substantial concordance in spouse eating patterns, some concordance in sibling eating patterns, and little concordance in friend eating. By specifying lagged regression models in which the relationship between an alter's prior food consumption and ego's current consumption is assessed, the results provide evidence of a social influence process. Finally, in a recent study of several thousand Minneapolis adolescents across multiple schools, Bruening et al. (2012) calculated measures of healthy eating behaviors in one's peer group (i.e., ego network). Specifically, the investigators examined the separate contributions of ego's friend group and best friends' behaviors on the frequency of breakfast, servings per day of fruit, vegetables, whole grains, and dairy foods. While the propensity to eat breakfast and consume whole grains and dairy were each significantly associated within friend groups and best friend pairs, there was less evidence that adolescents share similarities in levels of fruit and vegetable consumption. In terms of unhealthy foods, a rigorous sociocentric investigation using a stochastic actor-based framework reveals that adolescent intake of low-nutrient, energy-dense foods is predicted by the intake of socially connected friends (De la Haye et al., 2013).

Cigarette smoking as a consumption practice has also been the topic of a great deal of social network research. In one of the first studies to link smoking with peer group structure, Ennett and Bauman (1993) studied the smoking habits of a cross-sectional panel of ninth graders in North Carolina. By analyzing clique membership, liaison positions, and social isolates, they find that smoking is most strongly linked with social isolation. In a longitudinal study, Alexander et al. (2001) examined two waves of the Add Health study to examine the role of peer influence on smoking behavior among seventh to twelfth graders. Through measurement of peer smoking exposure, the authors discovered that students with a peer network in which more than half of one's alters smoked tobacco were almost twice as likely to be current smokers. The ego was also twice as likely to be a current smoker if ego's best friend had been a smoker. Additionally, ego popularity contributed to a small but notable risk of being a current smoker. Valente et al. (2005) seek to extend this last finding concerning popularity, and ask if popularity predicts subsequent smoking initiation among middle-school students at several schools. For the students in the study, popularity was strongly predictive of increased susceptibility to start smoking and also having ever smoked; the authors also found that different measures of network centrality were associated with smoking initiation.

More methodologically sophisticated treatments of peer influence and smoking over time involve the analysis of multiple panels of adult data in an observational setting (Christakis and Fowler, 2008) and actor-based modeling approaches of adolescents (Mercken et al., 2009). In a study of more than 30 years of smoking behavior among more than 12 000 Framingham adult cohort members, Christakis and Fowler discover that while smoking behavior clusters among interconnected actors in the study, entire subgroups tend to quit at nearly the same time, and the relative centrality of smokers in the network declines over time. In contrast, Mercken and collaborators employ a series of stochastic actor-based models in order to distinguish the roles of peer selection

from influence in the smoking behavior of adolescents across multiple high schools in six European countries during more than 2 years. The authors found strong evidence for selection in all six countries, but evidence for influence in only two.

Challenges and Future Directions

Studies of consumption that span topical domains have potential to inform more incisive theory by extending their scope beyond a narrow focus on only one type of consumption. For instance, in an unusual population-based health study, Bygren et al. (1996) examined links between Swedish lifestyle habits and health while controlling for the presence of a strong social network. They discovered that greater cultural consumption in performing arts attendance, reading, and making music is associated with reduced risk of mortality. Though by today's standards a very rough social network indicator was used (a binary measure indicating frequent versus sparse friend group contact), the researchers found that among men the presence of one's social network elevated one's mortality risk, while among women the absence of a social network elevated risk. Another example illustrates how the conceptual distinction between cultural consumption and production can be illusory (Giuffre, 2001). To wit, while art critics function as bridges between artists and the consuming public and are thus neither producers nor consumers of art, it can be reasonably argued that as a group of highly interested cultural consumers equipped with professional expertise, art critics constitute a peculiar audience class unto themselves. This nuance is pertinent to a study of fine art photographers and reviews of their gallery exhibitions in which network ties between artists and galleries classify structurally equivalent groups of relationships (Giuffre, 2001). In examining the links between these artist–gallery groups and the artists' critical reviews, Giuffre discovers that the type of artist discerned through analysis of artist–gallery relations corresponds to the frequency and content of reviews by critics. Research such as these two examples enriches our knowledge by illustrating the links between consumption across ostensibly separate life domains, as well as the shrinking distance between producer and consumer in modern life.

One of the challenges for consumption research is to integrate how we theorize the intersection of the symbolic and pragmatic aspects of cultural consumption. Research on arts participation, for instance, has largely concerned itself with how taste preferences (a symbolic dimension) shape or reflect participation, and issues surrounding access to and constraints on participation (a more pragmatic dimension) have been more peripheral. Networks can affect the development of taste as well as enhance or constrain access. Certainly in research on food consumption, issues of taste have been more interesting for sociologists of culture, while practical issues surrounding food as nutritive fuel have been the domain of health research. A closer integration stands to help us better theorize how biological processes are related with social processes. Growth in this area will benefit from an interdisciplinary focus on cognition and decision-making processes, areas of scholarship that have traditionally been the purview of social psychology.

Another key challenge for future scholars in this field is to integrate study of online and offline cultural consumption and social networks. With several exceptions, this article has largely been concerned with understanding social networks as built from face-to-face human relationships, even though online social media are beginning to transform the very notion of how we understand a social network. Questions surrounding how online social networks shape cultural consumption are increasingly of interest to scholars and change how we think about the nature of relationships. In a contemporary society where individuals have multiple and overlapping online and offline identities, this is not a trivial concern.

See also: Cultural Participation, Trends In; Cultural Production in Networks; Culture and Networks; Leisure and Cultural Consumption: US Perspective; Networks and Meaning; Symbolic Boundaries.

Bibliography

Alexander, Cheryl, Piazza, Marina, Mekos, Debra, Valente, Thomas, 2001. Peers, schools, and adolescent cigarette smoking. The Journal of Adolescent Health 29, 22–30.
Becker, Howard S., 1982. Art Worlds. University of California Press, Berkeley.
Benediktsson, Mike O., 2012. Bridging and bonding in the academic melting pot: cultural resources and network diversity. Sociological Forum 27, 46–69.
Bourdieu, Pierre, 1984. Distinction: A Social Critique of the Judgement of Taste. Harvard University Press, Cambridge, Massachusetts.
Bourdieu, Pierre, 1993. The Field of Cultural Production: Essays on Art and Literature. In: Johnson, Randal (Ed.). Columbia University Press, New York.
Breiger, Ronald L., 2009. The analysis of social networks. In: Hardy, Melissa, Bryman, Alan (Eds.), Handbook of Data Analysis. Sage Publications, London.
Bruening, Meg, Eisenberg, Marla, Maclehose, Richard, Nanney, Marilyn S., Story, Mary, et al., 2012. Relationship between adolescents' and their friends' eating behaviors: breakfast, fruit, vegetable, whole-grain, and dairy intake. Journal of the Academy of Nutrition and Dietetics 112, 1608–1613.
Bygren, Lars O., Konlaan, Boinkum B., Johansson, Sven-Erik, 1996. Attendance at cultural events, reading books or periodicals, and making music or singing in a choir as determinants for survival: Swedish interview survey of living conditions. British Medical Journal 313, 1577–1580.
Cardon, Dominique, Granjon, Fabien, 2005. Social networks and cultural practices – a case study of young avid screen users in France. Social Networks 27, 301–315.
Christakis, Nicholas A., Fowler, James H., 2008. The collective dynamics of smoking in a large social network. New England Journal of Medicine 358, 2249–2258.
Crossley, Nick, 2009. The man whose web expanded: network dynamics in Manchester's post/punk music scene 1976–1980. Poetics 37, 24–49.
De la Haye, K., Robins, G., Mohr, P., Wilson, C., 2013. Adolescents' intake of junk food: processes and mechanisms driving consumption similarities among friends. Journal of Research on Adolescence 23, 524–536.
Ennett, Susan T., Bauman, Karl E., 1993. Peer group-structure and adolescent cigarette-smoking – a social network analysis. Journal of Health and Social Behavior 34, 226–236.
Erickson, Bonnie H., 1996. Culture, class, and connections. American Journal of Sociology 102, 217–251.
Gaughan, Monica, 2006. The gender structure of adolescent peer influence on drinking. Journal of Health and Social Behavior 47, 47–61.
Giuffre, Katherine, 2001. Mental maps: social networks and the language of critical reviews. Sociological Inquiry 71, 381–393.
Kane, Danielle, 2004. A network approach to the puzzle of women's cultural participation. Poetics 32, 105–127.
Kane, Danielle, 2011. The gendered transition to college: the role of culture in ego-network evolution. Poetics 39, 266–289.
Kayahara, Jennifer, Wellman, Barry, 2007. Searching for culture-high and low. Journal of Computer-Mediated Communication 12 (3), 824–845.
Kraaykamp, Gerbert, Van Eijck, Koen, Ultee, Wout, Van Rees, Kees, 2007. Status and media use in the Netherlands – do partners affect media tastes? Poetics 35, 132–151.
Kreager, Derek A., Haynie, Dana L., 2011. Dangerous liaisons? Dating and drinking diffusion in adolescent peer networks. American Sociological Review 76, 737–763.

Latkin, Carl A., Mandell, Wallace, Vlahov, David, 1996. The relationship between risk networks' patterns of crack cocaine and alcohol consumption and HIV-related sexual behaviors among adult injection drug users: a prospective study. Drug and Alcohol Dependence 42, 175–181.

Lewis, Kevin, Kaufman, Jason, Gonzalez, Marco, Wimmer, Andreas, Christakis, Nicholas, 2008. Tastes, ties, and time: a new social network dataset using Facebook.com. Social Networks 30, 330–342.

Lewis, Kevin, Gonzalez, Marco, Kaufman, Jason, 2012. Social selection and peer influence in an online social network. Proceedings of the National Academy of Sciences of the United States of America 109, 68–72.

Lizardo, Omar, 2006. How cultural tastes shape personal networks. American Sociological Review 71, 778–807.

Lizardo, Omar, 2013. Variety in cultural choice and the activation of social ties. Social Science Research 42, 321–330.

Mercken, Liesbeth, Snijders, Tom A.B., Steglich, Christian, de Vries, Hein, 2009. Dynamics of adolescent friendship networks and smoking behavior: social network analyses in six European countries. Social Science & Medicine 69, 1506–1514.

Mische, Ann, 2011. Relational sociology, culture, and agency. In: Scott, John, Carrington, Peter J. (Eds.), The Sage Handbook of Social Network Analysis. Sage Publications, London.

Pachucki, Mark A., Breiger, Ronald L., 2010. Cultural holes: beyond relationality in social networks and culture. Annual Review of Sociology vol. 36, 205–224.

Pachucki, Mark A., Jacques, Paul F., Christakis, Nicholas A., 2011. Social network concordance in food choice among spouses, friends, and siblings. American Journal of Public Health 101, 2170–2177.

Peterson, Richard A., Simkus, Albert, 1992. How musical tastes mark occupational status groups. In: Lamont, Michèle, Fournier, Marcel (Eds.), Cultivating Differences: Symbolic Boundaries and the Making of Inequality. University of Chicago Press, Chicago.

Peterson, Richard A., Kern, Roger M., 1996. Changing highbrow taste: from snob to omnivore. American Sociological Review 61, 900–907.

Rawlings, Craig M., Bourgeois, Michael D., 2004. The complexity of institutional niches: credentials and organizational differentiation in a field of US higher education. Poetics 32, 411–437.

Relish, Michael, 1997. It's not all education: network measures as sources of cultural competency. Poetics 25, 121–139.

Rosenquist, J. Niels, Murabito, Joanne, Fowler, James H., Christakis, Nicholas A., 2010. The spread of alcohol consumption behavior in a large social network. Annals of Internal Medicine 152, 426–W141.

Rozin, Paul, Riklis, Jordanna, Margolis, Lara, 2004. Mutual exposure or close peer relationships do not seem to foster increased similarity in food, music or television program preferences. Appetite 42, 41–48.

Salganik, Matthew J., Dodds, Peter S., Watts, Duncan J., 2006. Experimental study of inequality and unpredictability in an artificial cultural market. Science 311, 854–856.

Schultz, Jennifer, Breiger, Ronald L., 2010. The strength of weak culture. Poetics 38, 610–624.

Scott, John, 2000. Social Network Analysis: A Handbook. Sage Publications, London; Thousands Oaks, California.

Skog, Ole-Jørgen, 1986. The long waves of alcohol-consumption – a social network perspective on cultural-change. Social Networks 8, 1–32.

Steglich, Christian, Snijders, Tom A.B., West, Patrick, 2006. Applying SIENA: an illustrative analysis of the coevolution of adolescents' friendship networks, taste in music, and alcohol consumption. Methodology 2, 48–56.

Tepper, Steven J., Hargittai, Eszter, 2009. Pathways to music exploration in a digital age. Poetics 37, 227–249.

Valente, Thomas W., Unger, J.B., Johnson, C.A., 2005. Do popular students smoke? The association between popularity and smoking among middle school students. Journal of Adolescent Health 37, 323–329.

Valente, Thomas W., Ritt-Olson, Anamara, Stacy, Alan, Unger, Jennifer B., Okamoto, Janet, et al., 2007. Peer acceleration: effects of a social network tailored substance abuse prevention program among high-risk adolescents. Addiction 102, 1804–1815.

Veblen, Thorstein, 2007[1899]. The Theory of the Leisure Class. Oxford University Press Inc., Oxford; New York.

Warde, Alan, Tampubolon, Gindo, 2002. Social capital, networks and leisure consumption. Sociological Review 50, 155–180.

White, Harrison C., White, Cynthia A., 1965. Canvases and Careers; Institutional Change in the French Painting World. Wiley, New York.

Networks and Meaning

Jan Fuhse, Humboldt University, Berlin, Germany
Neha Gondal, The Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social networks are intricately interwoven with forms of meaning like roles categories, identities, and relationship frames. This article offers a comprehensive overview of the recent strand in sociological network research (around Harrison White) that studies this interlinkage, and of its most important theoretical arguments. Social networks are conceptualized as patterns of interaction that are governed by an underlying structure of expectations. The meaning of networks comprises the identities of the actors in the network, the relationships at play, the categories, roles, and institutions structuring the network symbolically, and the styles that develop in a network.

Introduction

The traditional approach in social network analysis is to study the structure (pattern/regularities) of relations. Relationships in these early studies were typically measured in binary terms as existing or not. The meanings of ties, on the other hand, were intentionally ignored or bracketed. While proving fruitful in a number of important studies, this approach displays particular limits because it reduces social phenomena to a-cultural structures. As a response to these limits, recent developments in North American network research around Harrison White have taken a 'cultural turn.' Juxtaposed against a traditional focus on structure, authors like White, Ronald Breiger, Ann Mische, Mustafa Emirbayer, John Mohr, and Paul McLean contend that meaning ought to be treated as important in social network analysis. More specifically, the structure of networks is seen not only as a conduit for the distribution (or bottlenecking) of cultural forms, but as itself based on forms of meaning. This includes the construction of identities of actors involved in interaction, the stories that make and are made in networks, the patterning of ties by institutions, roles, and categories, the imprint of relationships with frames like 'love,' 'friendship,' and 'enmity,' and the emergence of various styles in networks. This article provides a systematic overview of these arguments. It first offers a brief overview of the concept of meaning, delineating it against the parallel concept of culture (1). The most important approaches linking social networks with a concern for meaning are sketched next (2). The last section discusses the core concepts of identities, stories, roles, categories, relationship frames, and styles with regard to their role in networks (3).

Conceptualizing Meaning

Over the last 100 years, meaning has continuously remained one of the core concepts in sociology. How exactly meaning has to be conceptualized, however, differs dramatically in the most important approaches, from Max Weber's theory of action and symbolic interactionism, to Pierre Bourdieu's theory of practice, neo-institutionalism, and Niklas Luhmann's systems theory. Weber treats meaning as *subjective* and places it in the heads of actors. In this tradition, social phenomena have to be traced to individual action, with individuals acting on the basis of subjective meaning. Social structures (from ethnic groups to state bureaucracies) are treated as 'social relations' resulting from and consisting in the mutual attunement of subjective meaning that leads to regularities in social behavior. Talcott Parsons paradigmatically traces this attunement to an alter-ego situation where both actors develop expectations on the basis of each others' actions. Following this tradition, social relationships and networks would have to be grounded in subjective dispositions. They consist of the expectations that actors have of other actors' actions; and all action leads to the stabilization or the modification of these relational expectations.

Symbolic interactionism underscores the *intersubjective nature* of meaning, postulating that meaning emerges out of interaction. Close-knit groups of primary relations develop common 'definitions of the situation' that differ from those in other groups. Such definitions of the situation include the sharp distinction from other groups, but also the adoption and playful variation of roles. While symbolic interactionism mostly focuses on group cultures, Gary Alan Fine and Sherryl Kleinman propose a combination with network analysis, arguing that both are compatible in that the negotiation of the meaning of relationships in networks follows the basic interactionist model (1983).

In stark contrast with the action theoretical tradition, Luhmann conceptualizes meaning as mainly *communicated*. While meaning is also thought in psychic systems, sociologically relevant meaning is processed supraindividually in social systems like encounters, organizations, and functional subsystems (economy, law, politics, etc.). In this perspective, social relationships and networks would result from the self-referential operating of communication, with the subjective meaning of actors remaining fundamentally unknown to others (including social scientists). Again they would consist of expectations. But here the term 'expectations' stands for the meaningful channeling of subsequent communication by communicative events.

Bourdieu's theory of practice to a certain degree proposes a middle ground between action and communication. Here, meaning is enacted in *social practices*, guided by the habitus

acquired in interaction in social fields. Practices differ by position in different types of capital and mostly lead to the reproduction of existing structures. Thus social networks, as a form of capital, would yield habitus and practices according to network position, and thus remain relatively stable.

All of these approaches have implicitly or explicitly informed thinking about networks as structures of meaning. It is neither necessary nor suitable to side with one particular approach here. But the concept of meaning has to be demarcated from that of culture, which is closely related and treated in another article (*see* Culture and Networks). Culture broadly refers to forms of meaning (values, symbols, understandings) shared in a social context (country, city, social group, organization). This article focuses on forms of meaning that are commonly and consensually understood, but that are not shared – they affect actors in networks differently. Roles, categories, identities, stories, and styles are likely to be similarly interpreted by the people connected in a network; but those people do not necessarily belong to the very same categories (e.g., gender) – they occupy different roles, feature distinctively in stories, and do not share the same individual identities. These differentiated forms of meaning are instrumental in relating actors to each other.

Meaning in Social Network Analysis

In a sense, social network analysis has always struggled with meaning. The classical structuralism of network modeling and the social capital tradition deliberately bracket the self-understandings of involved actors and the meanings attached to and negotiated in social relationships. Culture and meaning are treated as passive results of network organization, if at all. On the fringes of this emerging network analytical mainstream, various approaches from social anthropology to relational sociology have incorporated meaning (in the various senses sketched above) into their theoretical formulations and empirical studies. This section presents a short overview of their accounts of meaning in networks.

British social anthropology around J.A. Barnes, Elizabeth Bott, and J. Clyde Mitchell invented and elaborated the network concept in the 1950s and 1960s to get away from a one-sided emphasis on meaning in cultural anthropology. In these early studies, self-perceptions and descriptions of social structures by the participants were seen as defective and misleading. Therefore, these perceptions were questioned and superseded by structural analysis: the focus on the 'concrete' relations between actors. While lay descriptions of social structures typically focus on clear-cut divisions between groups and classes, social structure is 'really' to be found in the complex mesh of social relations – the network. However, to bridge the divide between social and cultural anthropology, Mitchell called for the systematic incorporation of a concern with meaning into social anthropology (1969: pp. 20ff.). According to Mitchell, the structural analysis of networks should be complemented with investigation into actors' understandings of social relationships and into the norms governing interaction. These calls have been taken up recently by anthropologists like Russell Bernard, Ulf Hannerz, and Thomas Schweizer.

Mainstream social network analysis had, until recently, followed the path of early social anthropology and largely ignored meaning. Social structures were reduced to patterns of relations without regard to their meaning for participants. Of particular interest are the SNA technique of *blockmodeling* and its treatment of *positions in networks*. Blockmodeling was developed in the 1970s. It analyzes networks by partitioning them into groups of actors with similar relations to other groups of actors. These groups are called blocks or positions. The blocks may consist of actors not connected to each other, but related to other blocks in similar ways. Actors occupying a given position are termed 'structurally equivalent,' and they are seen as playing the same role (elite/clients, etc.) in the network (White et al., 1976).

This technique of partitioning actors into positions followed in the footsteps of Siegfried Nadel, who postulated roles to comprise both cultural and relational aspects (1957). But blockmodeling followed only one side of this lead by emphasizing relational patterns. Methodologically, blockmodeling provides the prime example of the anticategorical imperative because meaning in the network is inferred from the structure of relations rather than posited to exist independently of the network.

Important impulses with regard to the question of meaning come from *neo-institutionalism*. Notably the works of Paul DiMaggio as well as John Mohr feature networks as a core concept and build on the concepts and implications of blockmodeling (DiMaggio, 1986; Mohr and Friedland, 2008). These systematic patternings mirror institutions – sets of rules for the behavior of actors (in their roles) toward each other, encapsulated in 'practical action.' The concrete roles that actors occupy in a network are governed by institutions. These emerge isomorphically in fields. Network analysis, in particular blockmodeling, thus constitutes a suitable technique for the study of the cultural rules in a field.

The development of blockmodeling, the Anglo-American reception of Bourdieu, and neo-institutionalism paved the way for *relational sociology*, sometimes also called the New York school of relational sociology (Mische, 2011). Harrison White's 1992 book *Identity and Control* (a second, thoroughly reformulated edition was published in 2008) marks the programmatic and theoretical core of this endeavor. The most important theoretical argument of the book for our purpose is that *networks are structures of meaning* ('phenomenological realities'). More precisely, networks are seen as social structures linking identities through stories. The identities themselves are not essential building blocks, but result from their relational construction in stories. Thus language becomes an important factor for the construction of networks.

These basic arguments have been taken up and expanded by a number of scholars around White, notably by Charles Tilly, John Padgett, Ann Mische, Mustafa Emirbayer, Stephan Fuchs, David Gibson, Paul McLean, and John Mohr. Relational sociology benefits from a close connection between theoretical elaboration and empirical studies. Its particular conjectures about the connection between networks and meaning are discussed in the next section. Another approach linking networks

with meaning is actor–network theory around Bruno Latour, Michel Callon, and John Law. This approach remains mostly separate from the advances in relational sociology and is discussed in another article (*see* Culture and Actor Network Theory).

Forms of Meaning in Social Networks

This section discusses various forms of meaning with regard to their embeddedness in and their impact on the creation and maintenance of social networks. It draws on the various theoretical treatises and empirical studies of relational sociology without discussing them in detail. The aim is to present a coherent overview of the role of identities, stories, categories, relational frames, and styles in networks and of the rooting of both these forms of meaning and of the corresponding network structure in the process of interaction.

1. Following Harrison White, the relational definition and negotiation of the identities of the involved actors lies at the heart of social networks (2008: pp. 2ff., 9ff.). His definition of *identities*, however, differs from everyday notions of selfhood or group memberships. All interaction involves uncertainties of differing degrees. Identities emerge out of this soup of contingency as actors try to cope with uncertainties. Identities are the 'footings' or points of view that guide action in uncertain situations. Both the bearer of an identity and others react to this footing, thus leading to the formation of ties between individuals. These constructed identities, not preexisting individuals, are related meaningfully in networks.

For example, a group of persons find themselves in close physical proximity at a party and a discussion starts. Initially, the uncertainty is high because the topic of conversation is undecided after introductions have been made. One person could react to the uncomfortable situation by raising the topic of, say, local vegetarian restaurants, thus adopting the identity of conversation-starter. Other individuals now act on the basis of this new information. Someone else may offer a critical view of a new 'farm to table' restaurant in the neighborhood (adopting the identity of local food 'critic'). There is more talk about local restaurants and some people exchange telephone numbers leading to the creation of new ties. Some of those ties may amount to nothing more than telephone entries. But some others may evolve into lasting friendships.

Note that evolving situations lead to changes in identities. For example, two people who go out to dinner now adopt, say, date identities to deal with different types of contingencies that arise in that situation. Identities are thus dynamic constructs that structure and are structured in the course of interaction.

This relationship between identity and networks has been demonstrated in empirical research. In Randall Collins's study of networks in philosophy, the construction of particular authors as important or peripheral to the development of a field is shaped by her position in networks of intellectual influence (1998). But this identity construction in turn makes for a rearrangement of ties, for example, when an author becomes the proponent of a new school of thought gathering around her or when an important author becomes viewed as even more important through a process of preferential attachment.

The continuous negotiation of the identities of academic authors as well as those of other people constitutes an intricate part of network formations. Identities are always 'footed' in relation to other identities. For example, the academic star would be nothing without her followers, and vice versa. The development of philosophy – as those of other disciplines and cultures – is heavily shaped by shifting network constellations.

Generally, the adoption and attribution of identities is a risky and shaky process. As identities are relational, the adoption of one identity relies on others to play their part – as in the case of the academic leader and her followers. The unilateral adoption of any identity confers power to the other party to accept or reject the claim. For example, in the beginning of a romantic interaction, neither party may want to appear more or less committed to the interaction. A signal of greater commitment from one person leaves open the possibility that the other person signals less commitment awarding more power to that person. In such situations, identities (and their relations to each other) emerge in the course of inherently ambiguous interactional sequences.

The emergence of a collective identity (as among the workers and nonworkers in the Paris Commune studied by Roger Gould, 1995) similarly relies on and triggers a particular structure of social networks. These have to be highly divided along this partition. In turn, the symbolic identification of this partition foments greater social conflict and reinforces categorical differences, thus fostering network segregation.

2. Identities are defined and related to each other in *stories* (White, 2008: pp. 26ff.; Somers, 1994). These stories fluctuate in network ties, but they also structure them. They formulate accounts of the qualities and the dispositions of the actors involved and of their relations to other actors, coalescing into a meaningful, but contested description of the network. In turn, this 'phenomenological reality of the network' structures interaction in the network by rendering particular interactions likely and others unlikely (Fuhse, 2009: pp. 52f.).

For example, if a network features a story of how A cheated on B, this will affect how A and B behave toward each other and others behave toward them. B will be seen as hurt and easily fooled. Others might offer consolation or help against A's action. A will be seen as not trustworthy. Therefore others will probably shy away from engaging in risky enterprises with him or her. B might try to retaliate (with or without the help of others), and A could try to make it up to B, if only to regain a reputation of trustworthiness. Now all of this was presented as if A, B, and the others were rationally calculating agents, making sense of the situation and reacting to it accordingly. According to White, Tilly, and Fuchs, the behavior of the identities involved follows more from circumstances than from calculation and dispositions (Fuchs, 2001: pp. 251ff.). The overall process can thus be attributed to the actions of the identities involved, or seen as recalibration of the network, following the implications of such stories, both in structure and in the construction of meaning.

Everyday storytelling generally adopts an action-theoretical perspective: Events are accounted for as the results of deliberate individual actions, driven by inner motives and interests (Tilly, 2002: pp. 26ff.). As in the example above, such everyday storytelling leads to the construction of identities on the basis of assumed personal qualities (easily fooled) and inner motives (selfish); they are related to each other when these dispositions are directed at particular other identities (B's cheating reveals a lack or a loss of personal interest in A); and they often relate more than two actors to each other, structuring not only dyadic relationships (between A and B) but also constellations (A, B, C, and D). And stories are dually relational by not only linking actors to actors, but more directly events to other events. In any case, such relational stories reduce social contingency by making particular interactions likely and others unlikely.

Stories affect interactions through the formation of expectations (see Section Conceptualizing Meaning). Stories about events lead to expectations about future events – and it is these expectations that effectively condition future events. It is not necessary to conceptualize these expectations as individual and subjective: The expectations are part of the network – they are the meaning of the network; and individuals adopt these expectations and routinely conform to them. Such expectations can develop in the network out of observing events. But they can also derive from institutionalized cultural repertoires, as in the case of relationship frames.

3. Even though social network analysis has often bracketed the question of relational content, opting instead for a binary representation of zeroes and ones, the *type of tie* connecting individuals has always been relevant. That is, individuals are not simply related to one another, but rather related in particular ways – through friendship, advice, economic transactions, sexual intercourse, dating, etc. The type of tie has important and significant implications for network structure and outcomes. Friendship ties, for example, are likely to be reciprocal (if I am your friend, then you are mine) and closed (you are likely to be friends with your friend's friends). Romantic ties, on the other hand, typically defy such extended closure – if person A is romantically involved with both B and C, it is unlikely that B and C are also romantically linked.

The distinction between these types of tie is a symbolic one. Friendship ties are different from romantic ties by virtue of the cultural expectations tied to these. Friendship and love are cultural models (*relationship frames*) for ties. Such models vary over time and social context (Yeung, 2005). They prescribe what people can expect of one another in these relationships, and they suggest a particular structuring of social relations around the tie in question. For example, love relationships are expected to remain exclusive, but also to mesh in the couple's respective friendship networks. Thus, the meanings associated with singular relationships have consequences for their network structure.

Because such meanings are continuously in flux, relational structure also has to be conceived of as dynamic rather than static. For example in Renaissance Florence, patronage seekers discursively frame relationships drawing on cultural models like 'honor,' 'friendship,' 'affection,' and 'service' in the hopes of building networks and gaining favor (McLean, 2007). The use of these discursive framing techniques come from already achieved positions in networks but also alter existing relational structures.

While any given relationship may be based upon cultural blueprints, communication between interactants will also lead to modification of those expectations. Relationships may also switch between frames. Such changes in the meanings of relationships have cascading effects, altering the structure of networks in which those ties are embedded. In a group of friends, for example, the breaking up of one romantic relationship may put strains on other ties as well. In similar spirit, while many relationships can be defined by type (such as love or friendship), others may not be amenable to such stable categorization. A given relation may be imbued with different meanings by different network participants. This diversity in tie-meaning has implications for the network's micro- and macrostructure and can be linked to actors' differential exposure to distinctive social contexts and differential embeddedness of their ties in other networks (Gondal and McLean, 2013).

4. Individual identities, stories, and relationship frames will be strongly affected by the partitioning of the identities in a network by *social categories.* In school classes, friendships predominantly form within-gender, love relationships (still) mostly cross-gender. Thus if A asks B to go to the movies together, this will be seen very differently if they belong to the same gender or not, evaluating this behavior against typical scripts for the categories involved. The same holds true if A happens to be the teacher and B a pupil.

Categorical ascriptions shape the identities adopted by and attributed to actors, thus reducing interactional uncertainty and affecting the storytelling about particular events. But it also directly impinges on the probabilities of events to take place at all. A is unlikely to ask B to go to the movies if B is in a (heterosexual) relationship with C *and* of the opposite sex to A. Similar interactional norms often apply to interracial or interethnic relationships. Categories like ethnic descent or gender thus make for a patterning of social ties (Tilly, 1998: pp. 62ff.). This patterning can lead to constellations where categories divide a network into *cohesive subgroups*. Ethnic categories, for example, usually increase the probability of forming intraethnic personal relationships. Interethnic relationships, in contrast, are unlikely.

This propensity of forming personal relationships within a category is called 'homophily' (McPherson et al., 2001). However, this tendency is often supported by other factors, for example, when minorities live in separate neighborhoods without much interethnic contact to start from, or when minority members share similar cultural traditions and values that increase the likelihood of forming personal relationships. In general, social categories that bound densely connected subgroups allow them to bar outsiders from resources and opportunities – a mechanism Charles Tilly terms 'opportunity hoarding' (1998: pp. 153ff.).

The genders also form interaction groups, but only when we consider friendship ties. These are mostly confined to same-sex relationships, especially among adolescents in school classes (studied in much network research). Love relationships,

however, are predominantly cross-gender. Categories thus do not necessarily separate cohesive subgroups. Rather they structure networks into positions of *structural equivalence*, with subgroups as a special case. For example, the category 'tenant' circumscribes a position in feudal social structure. But tenants are rarely connected – they resemble each other not in their connectivity, but in their relations to the nobility. We see here that categories are related to each other by specific types of tie – the relationship frames of 'friendship,' 'love,' and 'feudal tenure' imply a specific network constellation between the categories related by them (Fuhse, 2009: pp. 62f.).

Not only individuals, but also collective and corporate actors are positioned in networks by category. For example, the network of public discourse allocates very different roles to political parties, to social movements, to unions and churches, and to corporate interest groups. Whether a collective actor gets classified as a voluntary organization or as a corporate interest group has a great impact on their relations to other corporate and collective actors. Therefore this classification is frequently contentious. As with other forms of meaning, social categories for actors in networks are often derived from institutionalized cultural repertoires. But they can also develop in the network to map network positions by appropriate categories like subcliques or hangers-on. Once in place, social categories tend to stabilize relations of structural equivalence in networks.

These considerations lead back to the concepts of roles and institutions. Recall (from Section Meaning in Social Network Analysis) that institutions have been defined as habitually perpetuated, typified rule-systems governing the behavior of actors toward one another. *Roles* in networks such as subcliques or hangers-on are classified into categories, as soon as names or other symbols for these roles develop. These categories institutionalize to the extent that they crystallize as meaningful descriptions of roles and relations across network contexts. Such institutions provide cultural blueprints for the expectations attached to actors of a particular category, thus bearing an imprint on network patterns.

For example, the gender category is an institution that governs the formation of social constellations in a wide variety of social contexts. However, the precise meaning of gender, and of other institutions (and the resulting network structure), can vary considerably by context, for example, by region, class, age, ethnic group, organization, etc.

Apart from role categories, such institutions can entail expectations about relationship formation (relationship frames), or even constellations of a number of roles, for example, family roles in a kinship network. Roles and relationship patterns in a network form local structures that are often heavily imprinted by wider cultural models.

5. A last form of meaning to consider is *styles*. In art, as in science, fashion, or in politics, the notion of styles designates recognizable bundles of symbols as distinct from other styles (White, 2008: pp. 112ff.). Styles often emerge from dense contact within a subclique of a network, symbolically separating this subclique from other groups and individuals. Historical examples from art include impressionism, expressionism, Dadaism, and almost any other art school. The style here acts as the 'culture' of the network, forming a repertoire of symbols shared within the clique. This basic view has to be appended here in three regards, leading from culture to relational meaning:
   a. Art schools like impressionism or expressionism always developed against other artistic traditions (White, 1993). Their styles not only unified the network of the subclique, but also separated it from and related it to an overall network configuration of contending styles and schools in the field of art. The same holds true for many other fields: science, fashion, politics, etc. Styles symbolize as much about identity as difference.
   b. New styles – creativity – seldom develop from scratch, simply out of increased internal contact of a group. Rather, styles from different network clusters are routinely tried for combination at their intersection in networks, bridging 'structural' as well as 'cultural holes' (Collins, 1998: pp. 131ff.; Pachucki and Breiger, 2010). These attempts only crystallize to form a new style if a new clique clusters around it, forming a new subclique separate from the two (or more) clusters from which it emerged.
   c. Styles do not always pertain to cohesive subcliques. They can also correspond to network position, especially if the stylistic identity (and difference) concerns relational practices. Ann Mische demonstrates in her work on Brazilian youth movement how leaders develop specific styles depending on whether they are firmly embedded in one social movement, or whether they bridge several movements (2008: pp. 241ff.). Thus the structural equivalence of network position in different context can make for stylistic differences.

Conclusion

Meaning in networks can be studied in different ways (Fuhse and Mützel, 2011). 'Qualitatively,' it has been analyzed on the basis of rich ethnographic work, on the basis of observations and archival data, or on the basis of in-depth interviews. Yet, meaning also has a tradition of being studied 'quantitatively' in social network analysis. Here, cultural networks between forms of meaning (e.g., concepts) are constructed out of documents or interviews, and analyzed using formal techniques of social network analysis. A prominent technique used by many authors (John Mohr, Charles Tilly, Peter Bearman, Ron Breiger, John Levi Martin, and Paul DiMaggio) is blockmodeling, identifying 'roles' of concepts or other forms of meaning in relation to each other. Other formal analytical methods used to study networks of meaning include Galois lattices (Yeung, 2005) and exponential random graph models (Gondal, 2011).

Overall, this 'cultural turn' in sociological network research leads to a view where network structures are much more than coagulations of 0s and 1s. Rather, they are seen as composed of relations laden with meaning, of nodes that, while embedded in cultural and resource contexts, have identities and roles that are endogenously constructed in a dynamic process of network formation, and whose behavior is often strategic and steeped in interpretive work. This general perspective has been fruitfully applied in a wide variety of social contexts, yielding important advances in the fields of economic and historical sociology,

social movement studies, social inequality, research on institutions, and the sociology of science.

Recent advances in the field point to the importance of microprocesses of interaction or communication in the development, stabilization, and change of social networks. In particular, how roles and other forms of meaning crystallize in conjunction with network structures in the course of communication (Gibson, 2005) and how networks as sociocultural configurations change over time are important themes. Drawing on recent advancements in the modeling of social networks – like exponential random graph models – distinct mechanisms that lead to the construction of networks can be evaluated. For example, these models can be useful for parsing the effects of different types of homophily (gender, race, ethnicity, etc.) from other mechanisms of category-based clustering such as triadic closure in the analysis of, say, school friendships. As before, the development of relational sociology progresses in close conjunction with complex theoretical models and elaborated empirical studies.

See also: Cultural Production in Networks; Culture and Actor Network Theory; Culture and Networks; Culture, Cognition and Embodiment; Formal Methods of Cultural Analysis; Narrative Networks; Network Analysis, History of; Networks and Cultural Consumption.

Bibliography

Collins, Randall, 1998. The Sociology of Philosophies. Belknap, Cambridge/Massachusetts.

DiMaggio, Paul, 1986. Structural analysis of organizational fields: a blockmodel approach. Research in Organizational Behavior 8, 335–370.

Fine, Gary Alan, Kleinman, Sherryl, 1983. Network and meaning: an interactionist approach to structure. Symbolic Interaction 6, 97–110.

Fuchs, Stephan, 2001. Against Essentialism. Harvard University Press, Cambridge/Massachusetts.

Fuhse, Jan, 2009. The meaning structure of social networks. Sociological Theory 27, 51–73.

Fuhse, Jan, Mützel, Sophie, 2011. Tackling connections, structure, and meaning in networks: quantitative and qualitative methods in sociological network research. Quality & Quantity 45, 1067–1089.

Gibson, David, 2005. Taking turns and talking ties: networks and conversational interaction. American Journal of Sociology 110, 1561–1597.

Gondal, Neha, 2011. The local and global structure of knowledge production in an emergent research field: an exponential random graph analysis. Social Networks 33, 20–30.

Gondal, Neha, McLean, Paul, 2013. Linking tie-meaning with network structure: variable connotations of personal lending in a multiple-network ecology. Poetics 41, 122–150.

Gould, Roger, 1995. Insurgent Identities; Class, Community, and Protest in Paris from 1848 to the Commune. Chicago University Press, Chicago.

McLean, Paul, 2007. The Art of the Network: Strategic Interaction and Patronage in Renaissance Florence. Duke University Press, Durham.

McPherson, Miller, Smith-Lovin, Lynn, Cook, James, 2001. Birds of a Feather: homophily in social networks. Annual Review of Sociology 27, 415–444.

Mische, Ann, 2008. Partisan Publics: Communication and Contention across Brazilian Youth Activist Networks. Princeton University Press, Princeton.

Mische, Ann, 2011. Relational sociology, culture, and agency. In: Scott, John, Carrington, Peter (Eds.), Sage Handbook of Social Network Analysis. Sage, Thousands Oaks, pp. 80–97.

Mitchell, J. Clyde, 1969. The concept and use of social networks. In: Mitchell, J. Clyde (Ed.), Social Networks in Urban Situations. Manchester University Press, Manchester, pp. 1–50.

Mohr, John, Friedland, Roger, 2008. Theorizing the institution: foundations, duality, and data. Theory and Society 37, 421–426.

Nadel, Siegfried, 1957. The Theory of Social Structure. Cohen & West, London.

Pachucki, Mark, Breiger, Ronald, 2010. Cultural holes: beyond relationality in social networks and culture. Annual Review of Sociology 36, 205–224.

Somers, Margaret, 1994. The narrative constitution of identity: a relational and network approach. Theory and Society 2, 605–649.

Tilly, Charles, 1998. Durable Inequality. California, Berkeley.

Tilly, Charles, 2002. Stories, Identities, and Political Change. Rowman & Littlefield, Lanham.

White, Harrison, 1993. Careers & Creativity. Social Forces in the Arts. Westview, Boulder.

White, Harrison, 2008. Identity and Control, second ed. Princeton University Press, Princeton.

White, Harrison, Boorman, Scott, Breiger, Ronald, 1976. Social structure from multiple networks. I. Blockmodels of roles and positions. American Journal of Sociology 81, 730–779.

Yeung, King-To, 2005. What does love mean? Exploring network culture in two network settings. Social Forces 84, 391–420.

Networks, Geography of

Barney Warf, University of Kansas, Lawrence, KS, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Geography's recent post-structuralist turn has included a pronounced emphasis on relational conceptions of space. Central to this shift has been the use of networks to conceptualize geographical relations in an age of ferocious globalization and time–space compression. Empirically, networked geographies have been most commonly found in the analysis of economic issues. This article outlines several approaches to the theorization of networks: the rhizomes of Deleuze and Guattari, Latour's actor-network theory, the space of flows as put forth by Castells, commodity chain analysis, and Massey's power-geometries.

Economic, political, and cultural geographers, in acknowledging the massive changes unleashed by post-Fordist capitalism and flexible production, have increasingly emphasized mobilities, flows, and networks as the defining features of contemporary spatiality rather than discreet locales (Larsen et al., 2006). As a result, absolute conceptions of space have largely given way to relational perspectives, particularly concerning the space economy (Ettlinger, 2001; Bathelt and Glückler, 2003; Yeung, 2005; Glűckler, 2007; Sunley, 2008). Networks have become significant in analysis of flows of information, knowledge, and technologies over space and time, transnational corporations, globalization, commodity chains, transportation patterns, and the Internet. Other topics including social networks, and networks associated with differential spatial development.

This change in perspective has important intellectual and political ramifications. As Kirby (1996, p. 71) put it, "The distinctions between widely separated geographical entities lose their meaning as disparate sites themselves come together in a plastic network making proximity and separation relative and mutable." Relational space "is seen as an undulating landscape in which the linkages established in networks draw some locations together while at the same time pushing others further apart" (Murdoch, 2006, p. 86). Taken to the extreme, Doel (1999) follows Deleuze and Guattari in emphasizing the origami-like nature of space as it is repeatedly folded and refolded, fissured, cracked, and fractalized through a series of difference-producing repetitions.

If surfaces were the predominant metaphor that defined the geographies of modernity (e.g., the nation-state, urban commuting fields), their postmodern, post-structuralist equivalent is networks (Warf, 2008), which Ruggie (1993, p. 141) compares to the "economic equivalent of relativity theory." Geographies in this reading exhibit a "fibrous, thread-like, wiry, stringy, ropy, capillary character that is never captured by the notions of levels, layers, territories, spheres, categories, structures, or systems" (Paasi, 2004, p. 541). It is no coincidence that the discovery that space is socially produced, and the shift to relational space, arose in tandem with mounting emphasis on networks. The spatiality of surfaces and regions buttressed assumptions that space was fixed and given, asocial in character. Networks, in contrast, call attention to the social origins of spatial relations, the power interests vested in them, and their political consequences. Networks have been central to ushering in views of geography as topological and non-Euclidean in nature. For example, globalization, in linking spatially separated producers and consumers, has, in Sheppard's (2002) view, created global power-geometries that transcend rigid hierarchies of spatial scale to generate 'wormholes' that connect widely separated places in complex and often unexpected ways. Not coincidentally, the invocation of networks has accompanied geography's mounting intellectual respectability and the 'spatial turn' that has unfolded across numerous social sciences and humanities.

The analysis of space via networks has a remarkably long, if unacknowledged, history. This article charts the intellectual foundations of this process, beginning with the seminal works of Deleuze and Guattari. Next, it charts the dimensions and spatiality of actor-network theory as it originated in the works of Bruno Latour and Michael Serres. Third, it points to two explicit ways in which networks penetrated the discipline of geography, Castells's space of flows, commodity chain analysis, and power-geometries conceptualized by Doreen Massey. It should be emphasized that no uniformity of opinion exists regarding the nature and importance of networks. While most of these perspectives draw from post-structuralist accounts of space, knowledge, and power, there are important differences among them. However, to varying degrees, all of these lines of thought portray geographies in terms of the relational linkages among places rather than absolute location; they share a common concern that space is defined functionally rather than in terms of physical distance; and all transcend the traditional notion of spatial scale as fixed and given. In short, networks call attention to the ontology of relations rather than things.

Deleuze and Guattari: Geography as Rhizomes

Despite the notorious complexity of their writing, Deleuze and Guattari (1987) in many respects pioneered early notions of relational space. Their work suggested a self-consciously 'schizophrenic' view of space that foregrounds movement and constant flux rather than the usual dimensions of distance and direction.

In this conception, social and spatial relations may be traced to the rise of the state during the Neolithic era, in which geographically fixed agricultural societies began to tame and

dominate surround nomadic tribes roaming the surrounding plains. For Deleuze and Guattari, the state embodies all the essential characteristics of class society, including stability, hierarchy, control, classes, and the extraction of surplus value. As a result it produces striated spaces that are ordered and bordered, and thus easily controlled. The territories of the state are essentially arboreal in nature, defined by roots and nested scales and layers, all of which are amenable to panopticonic control. Intimately associated with this process were the construction of time as a linear sequence and the rise of what they call 'arboreal thought' that privileges one vantage point over others. Given their anarchist tendencies, class society and striated spaces are intimately associated with the reigning in of desire, or control over human subjects. In contrast, they juxtapose an idealized model of nomads and nomadic life. Whereas the state represents, for them, stasis and being (in the Aristotelian sense), nomads are paragons of unstructured flows, free movement, and becoming (in the way suggested by Heraclitus). Nomadic life produces what they call smooth spaces that are not easily controlled, allowing desire to unfold unfettered, and affirming the centrality of location in the generation of multiplicity, plurality, variation, and difference. Smooth nomadic space is folded, like origami, as agents produce and reproduce their worlds. Such a view emphasizes space as emergent, forever coming into being through connections, flows, interactions, and movements.

Deleuze and Guattari's most famous geographical contribution was the model of the rhizome, the network with no clear core or periphery, with multiple entry points and exits. Rhizomes exist in nature, of course, including fungi, crabgrass, potatoes, rodent burrows, ant colonies, and the synapses of the brain. Social equivalents are not hard to find, including terrorist networks, traffic jams, patterns of diffusion, and the Internet. Knowledge, too, is rhizomatic, ranging from the circulation of gossip to their condemnation of psychoanalysis as arboreal (and the suggestion that 'schizoanalysis' is emancipatory). The DeleuzeGuattarian sense of rhizomes is deeply political, as they are intimately associated with a world devoid of centralized political control, one in which desire is untamed, allowing emergent possibilities to come to fruition. Rhizomes in this light are arenas in which the possible becomes the actual, spheres of becoming rather than being: they blur the line between what is and what might be. Echoes of this way of thinking can be found in the literature on complex adaptive structures, self-organizing systems, and the path dependencies of evolutionary economics. The metaphor of the rhizome, so different from the transparent spaces of Euclideanism, views of space as interconnected sets of networks that are nonhierarchical in nature, in which connections among locales rather than their absolute positionality is the dominant characteristic.

Actor-Network Theory

A second pillar in the emergence of network models of space is actor-network theory (Murdoch, 1997, 1998; Law and Hassard, 1999). Inspired by the work of Bruno Latour (1993) and Michael Serres (1997) (see also Serres and Latour, 1995), actor-network theory incorporates sociological understandings of structuration theory (Giddens, 1984) with a post-structuralist social constructivist philosophy of science (Bingham and Thrift, 2000). Actor-network theory departs from the Enlightenment focus on dualities such as individual and society, people and nature, human and nonhuman, Western and non-Western, urban and rural, micro and macro, local and global. Rather, it takes as its point of departure the linkages among these categories rather than their essences as actors draw upon and combine them in various forms of hybridity.

Actor networks involve the mobilization of rules, resources, and power, including information, in order to accomplish tasks, creating a net of intended and unintended consequences that stretch across the spatiotemporal boundaries of the network. To maintain network functionality, actors must perform by being engaged with one another recursively, interpreting and translating one another's behavior. Actors and networks are thus twin, mutually presupposing aspects of one phenomenon, simultaneously enabling and constraining actions in time and space. Thus, it is not simply actors in everyday life who constitute the primary focus here, but their relative positionality and powers within integrated systems of power and information that matter most. Such a perspective has helped to humanize even the most abstract of economic processes by revealing them to be the products of agents enmeshed in webs of power and meaning, not disembodied processes that operate independently of the people who create them (Law, 1994). The strategy of embodiment goes a long way toward demythologizing teleological interpretations of globalization, which present it as natural and inevitable, and reveal global processes to be the contingent outcomes of decisions made by human actors tied up in networks that cross multiple spatial scales.

Actor-network theory generated considerable controversy because of the view it offered of humans as actors. In reducing people to positions in networks, critics allege that it was deeply antihumanistic in nature, denying any causal power to consciousness and robbing people of any ability to shape the world. 'Agency,' in this reading, is distinct from consciousness: actors have effects, and their intentions are essentially irrelevant. Moreover, because actor-network theory strives to overcome the artificial boundaries between culture and nature (Latour, 1993), actors in this sociotechnical seamless 'nature-culture' nexus need not be human, but may include inanimate objects such as books, papers, or computer systems (Bingham, 1996; Murdoch, 1997), which are necessary to the maintenance and operation of networks. For example, in order for a bank or other corporation to function, people within them must operate in interaction with computers, documents, buildings, and other objects: privileging the human over the nonhuman makes no sense. Actor-network approaches thus challenge the long-standing anthropocentrism of the social sciences, opening conceptual space for actors to be animals or things.

Because actor networks are deeply spatial, they offer important implications for spatial analysis. Such a view elides the conventional focus on spatial scale, for networks operate across many scales simultaneously, creating as Latour (1993, p. 121) put it, "an Ariadne's thread that allows us to pass with continuity from the local to the global, from the human to the nonhuman. It is the thread of networks of practices and instruments, of documents and translations." By forcing us to rethink how time and space are produced – that is,

topologically rather than in terms of conventional Cartesian and Kantian views of space that have dominated geography – actor-network theory becomes "a machine for waging war on Euclideanism" (Law, quoted in Murdoch, 1998, p. 357). Not surprisingly, this view has become popular among those advocating a 'flat ontology' devoid of scale (Marston et al., 2005) and relational views of space (Murdoch, 2006). Scale, in this view, is constructed, not given, and represents only varying lengths and sizes of actor networks. Distinctions such as global and local lose their significance: Latour famously asked, for example, whether the trains he saw from his Parisian office were 'local' or 'global.' The answer, of course, is that they are both. Like all post-structuralist views, actor-network theory rejects either/or dichotomies in favor of inclusive both/and views that celebrate hybridity.

The Space of Flows

One of the earliest and more influential steps toward a geography of networks was the space of flows proposed by Manuel Castells (1996, 1997). This analysis distinguished earlier *information* societies, in which productivity was derived from access to energy and the manipulation of materials, from later *informational* societies that emerged in the late twentieth century in which productivity is derived primarily from knowledge and information. In his reading, the time–space compression of postmodernism included the three overlapping layers of transportation and communication infrastructure, the cities or nodes that occupy strategic locations within these networks, and the social spaces produced by the global managerial class:

> Our societies are constructed around flows: flows of capital, flows of information, flows of technology, flows of organizational interactions, flows of images, sounds and symbols. Flows are not just one element of social organization: they are the expression of the processes dominating our economic, political, and symbolic life. … Thus, I propose the idea that there is a new spatial form characteristic of social practices that dominate and shape the network society: the space of flows. The space of flows is the material organization of time-sharing social practices that work through flows. By flows I understand purposeful, repetitive, programmable sequences of exchange and interaction between physically disjointed positions held by social actors.
>
> 1996, p. 412

For Castells (1996, p. 470), the space of flows consists of a series of overlapping rhizomatic networks: "Networks are appropriate instruments for a capitalist economy based on innovation, globalisation, and decentralised concentration; for work, workers and firms based on flexibility; for a culture of endless deconstruction and reconstruction; for a polity geared towards the instant processing of new values and public moods; for a social organisation aiming at the supersession of space and the annihilation of time." He argues, for example, that while people live in places, postmodern power is manifested in the linkages among places, their interconnectedness, as personified by business executives shuttling among global cities and using the Internet to weave complex geographies of knowledge invisible to almost all ordinary citizens. This process was largely driven by the needs of the transnational class of the powerful employed in information-intensive occupations; hence, he writes (1996, p. 415) that "Articulation of the elites, segmentation and disorganization of the masses seem to be the twin mechanisms of social domination in our societies." Flows thus consist of corporate and political elites crossing international space on transoceanic flights; the movements of capital through telecommunications networks; the diffusion of ideas through organizations stretched across ever-longer distances; the shipments of goods and energy via tankers, container ships, trucks, and railroads; and the growing mobility of workers themselves. In this light, the space of flows is a metaphor for the intense time–space compression of post-Fordist capitalism. Through the space of flows the global economy is coordinated in real time across vast distances, i.e., horizontally integrated chains rather than vertically integrated corporate hierarchies. In the process, it has given rise to a variety of new political formations, forms of identity, and spatial associations.

Flows by definition involve more than one place; hence, in a networked world, places have little meaning as isolated entities. Places are not locales as much as they are processes in which different types of activities are embedded and different forms of interconnection are established. As they become increasingly connected, the repercussions of actions in one area inevitably spiral out to shape other places, so that discreet boundaries have less and less significance as they are permeated with mounting ease. Decisions made by hedge fund managers in New York, for example, reach out to affect the lives of millions of people in locations as distant as East Asia. The space of flows wraps places into highly unevenly connected chains that center on connections among powerful elites, thus typically benefiting the wealthy at the expense of marginalized social groups. However, such networks are hardly randomly distributed over the earth's surface: rather, it reflects and reinforces existing geographies of power concentrated within specific nodes and places, such as global cities, trade centers, financial hubs, and headquarter complexes. Thus, far from being abstract networks divorced from history, such systems are only made comprehensible by embedding them in history (Cresswell, 2006).

Central to the rise of globalized, postmodern capitalism was the construction of a vast web of telecommunications networks, which were crucial to the hegemony of increasingly information-intensive capitalism. The core of the global telecommunications infrastructure is a seamless network of fiber-optic lines that serve as the backbone of the Internet, a geographical network *par excellance*. Because the implementation of fiber lines reflects the powerful vested interests of international capital, these systems may be seen quite literally as 'power-geometries' (Massey, 1993, 1999) that ground the space of flows within concrete material and spatial contexts. Financial services firms were at the forefront of the construction of high-capacity fiber networks in large part because they facilitated the use of electronic funds transfer systems, which comprise the nervous system of the international financial economy, allowing banks to move capital around a moment's notice, arbitrage interest rate differentials, take advantage of favorable exchange rates, and avoid political unrest. Cyberspace allows for the reconstruction of 'communities without

propinquity,' groups of users who share common interests but not physical proximity, although the ability of virtual communities to substitute for face-to-face ones is debatable. The implications of this process are sobering. As Graham and Aurigi (1997, p. 26) note, "Large cities, based, in the past, largely on face-to-face exchange in public spaces, are dissolving and fragmenting into webs of indirect, specialized relationships." More generally, cheap, instantaneous, and ubiquitous communications have increasingly made the notion of place as a discreet, bounded entity increasingly problematic by allowing people to be in several places simultaneously. In this context, absolute distance is much less relevant than relative distance, i.e., space measured in terms of time, cost, and social access. Conventional geographic preoccupations with proximity mean little in this case, for the topologies of global finance are formed and deformed at the speed of light in ways that the language of location theory cannot capture.

Commodity Chains

Another step in the theorization of space as networks is the notion of commodity chains (Gereffi and Korzeniewicz, 1994; Dicken et al., 2001; Bair, 2005). Drawing from the earlier French tradition of *filieres* or value chain analysis, commodity chains may be defined as networks of labor and production processes that give rise to a particular commodity from raw material to processing, delivery, and consumption. Commodity chains include flows of goods and information among different points or nodes, varying labor relations across the length of the chain, and different constellations of production and governance at each segment. Such tools are useful in dissecting, for example, the spatiality of transnational corporations. Under globalization, commodity chains became increasingly larger in scope and length, linking widespread places through an increasingly complex division of labor in which distant strangers became ever more reliant upon one another.

Gereffi (1996) famously distinguished between producer-driven and buyer-driven chains, in which the power to control transactions lay at different ends of the concatenated linkages. In overcoming the artificial separation between production and consumption, commodity chains help to expose the widespread commodity fetishism prevalent in advanced economies, exposing commodities, *a la* Marx, as more than just things but as embodiments of social processes and thus helping to expose the unequal power relations among places that lie in the creation of goods. Chains are hence simultaneously geographical, economic, political, and cultural. They lie at the boundaries of the tangible and the intangible, incorporating sets of meanings, or as Hartwick (1998, p. 424) elegantly put it, they "conjoin the representational with the geomaterial." Moreover, they are temporally situated, constantly fluctuating in location, composition, and length over time.

Power-Geometries

Drawing upon, overlapping with, and advancing the view of geography as networks is the enormously influential work of Doreen Massey (1993, 1999, 2005). Massey scathingly criticizes traditional notions that represent places as islands of stability in the constantly shifting oceans of capitalist change, arguing that in minimizing the interconnectedness of peoples and places such a characterization is reactionary. Networks, in contrast, highlight how changes in one location reverberate to shape other locations. Instead, Massey promotes a progressive sense of place that firmly links places to other places, a view in which places constantly change, producing and receiving changes through their constant, simultaneously determinant interactions with one another. A relational politics of place calls into question easy distinctions like inside/outside, near/far, space/place, and global/local, artificial differentiations that are always embedded in each other and mutually constituted.

More recently, Massey (2005) argues passionately that Cartesian conceptions of space as a passive surface inevitably deemphasize the temporal flux that are always an inherent part of geographies, and simultaneously create a false dichotomy between the local and the global. As an alternative, she suggests three maxims: (1) that space be seen as the product of interrelations, i.e., of embedded social practices in which identities and human ties are coconstituted; (2) that space be understood as the sphere of multiple possibilities, i.e., as a contingent simultaneity of heterogeneous historical trajectories; and (3) that space must be conceived as always under construction, in the process of forever being made, implying a continual openness to the future. These steps are fundamental to an appreciation of the deeply political nature of geography: "Conceptualising space as open, multiple and relational, unfinished and always becoming, is a prerequisite for history to be open and thus a prerequisite too, for the possibility of politics" (p. 59). This line of thought goes far to injecting an explicit concern for politics in the perspective of geography as networks.

Conclusions

Post-structural geography, in emphasizing the embodied nature of social life as situated practices, drew attention away from Euclidean conceptions of space as an inert container and toward interconnected sets of places as manifolds that are continuously folded and pleated, stretched, distorted, and shredded. This perspective has served to underscore how, unlike traditional chorology, with its emphasis on static places, relational geographies are always dynamic, incomplete, forever coming into being, and perpetually in flux, giving rise to ever-changing patterns of centrality and peripherality. Geography consists of the contingent networks or power-geometries generated by social interaction rather than a homogenous plane that preexists coherent, well-ordered societies. In short, space is emergent rather than existing a priori, it is composed of relations rather than structures.

Theorizations of networks take many forms, but all of them share, to one degree or another, an emphasis on the interconnectedness of places and peoples. In so doing, network geographies help to circumvent deeply entrenched dualities such as global/local that tend to reify spatial scale and obfuscate the political dynamics at work in the reproduction of uneven development. Globalization, for example, can be fruitfully

theorized as a series of long actor-networks that transcend political borders and render scale obsolete, or as commodity chains that unite producers, consumers, and intermediaries, calling attention to the ways in which some places shape others. In an age of intense, rampant time–space compression, networks are a useful and meaningful way of theorizing the social production of space in all its messy complexity, sidestepping conventional infatuations with proximity.

See also: Consumption, Geography of; Critical Realism in Geography; Economic Geography; Globalization: Geographical Aspects; Information Society, Geography of; Knowledge, Geography of; Location, Absolute and Relative; Place in Geography; Scale in Geography; Space and Social Theory in Geography; Spatial Interaction; Spatial Thinking in the Social Sciences; Time-Space in Geography; Transportation Geography.

Bibliography

Bair, J., 2005. Global capitalism and commodity chains: looking back, going forward. Competition and Change 9 (2), 153–180.

Bathelt, H., Glückler, J., 2003. Toward a relational economic geography. Journal of Economic Geography 3 (2), 117–144.

Bingham, N., 1996. Object-ions: from technological determinism towards geographies of relations. Environment and Planning D: Society and Space 14, 635–657.

Bingham, N., Thrift, N., 2000. Some new instructions for travelers: the geography of Bruno Latour and Michael Serres. In: Crang, M., Thrift, N. (Eds.), Thinking Space. Routledge, London, pp. 281–301.

Castells, M., 1996. The Information Age. In: The Rise of the Network Society, vol. I. Blackwell, Oxford.

Castells, M., 1997. The Information Age. In: The Power of Identity, vol. II. Blackwell, Cambridge.

Cresswell, T., 2006. On the Move: Mobility in the Modern Western World. Routledge, New York.

Deleuze, G., Guattari, F., 1987. A Thousand Plateaus: Capitalism and Schizophrenia. University of Minnesota Press, Minneapolis.

Dicken, P., Kelly, P., Olds, K., Yeung, H., 2001. Chains and networks, territories and scales: towards an analytical framework for the global economy. Global Networks 1, 89–112.

Doel, M., 1999. Poststructuralist Geographies: The Diabolical Art of Spatial Science. Rowman and Littlefield, Boulder.

Ettlinger, N., 2001. A relational perspective in economic geography: connecting competiveness with diversity and difference. Antipode 33, 216–227.

Gereffi, G., 1996. Global commodity chains: new forms of coordination and control among nations and firms in international industries. Competition and Change 1, 427–439.

Gereffi, G., Korzeniewicz, M. (Eds.), 1994. Commodity Chains and Global Capitalism. Praeger, Westport.

Giddens, A., 1984. The Constitution of Society: Outline of the Theory of Structuration. University of California Press, Berkeley.

Glűckler, J., 2007. Economic geography and the evolution of networks. Journal of Economic Geography 7 (5), 619–634.

Graham, S., Aurigi, A., 1997. Virtual cities, social polarization, and the crisis in urban public space. Journal of Urban Technology 4, 19–52.

Grannovetter, M., 1985. Economic action and social structure: the problem of embeddedness. American Journal of Sociology 91, 481–510.

Hartwick, E., 1998. Geographies of consumption: a commodity-chain approach. Environment and Planning D: Society and Space 16, 423–437.

Kirby, K., 1996. Indifferent Boundaries: Spatial Concepts of Human Subjectivity. Guilford, New York.

Larsen, J., Urry, J., Axhausen, K., 2006. Mobilities, Networks, Geographies. Ashgate, London.

Law, J., 1994. Organising Modernity. Blackwel, Oxford.

Latour, B., 1993. We Have Never Been Modern. Harvester Wheatsheaf, London.

Law, J., Hassard, J., 1999. Actor Network Theory and after. Blackwell, Oxford.

Leslie, D., Reimer, S., 1999. Spatializing commodity chains. Progress in Human Geography 23 (3), 401–420.

Marston, S., Jones, J.P., Woodward, K., 2005. Human geography without scale. Transactions of the Institute of British Geographers 30, 416–432.

Massey, D., 1993. Power-geometry and a progressive sense of place. In: Bird, J. (Ed.), Mapping the Futures: Local Cultures, Global Change. Routledge, New York, pp. 59–69.

Massey, D., 1999. Imagining globalization: power-geometries of time-space. In: Brah, A., Hickman, M., Mac an Ghail, M. (Eds.), Global Futures: Migration, Environment and Globalization. Macmillan, Basingstoke, pp. 27–44.

Massey, D., 2005. For Space. Sage, London.

Murdoch, J., 1995. Actor-networks and the evolution of economic forms: combining description and explanation in theories of regulation, flexible specialization and networks. Environment and Planning A 25, 731–757.

Murdoch, J., 1997. Towards a geography of heterogeneous associations. Progress in Human Geography 21, 321–337.

Murdoch, J., 1998. The spaces of actor-network theory. Geoforum 29, 357–374.

Murdoch, J., 2006. Post-Structuralist Geography: A Guide to Relational Space. Sage, London.

Paasi, A., 2004. Place and region: looking through the prism of scale. Progress in Human Geography 28, 536–546.

Ruggie, J., 1993. Territoriality and beyond: problematizing modernity in international relations. International Organisation 47, 139–174.

Serres, M., 1997. The Troubador of Knowledge. University of Michigan Press, Ann Arbor, MI.

Serres, M., Latour, B., 1995. Conversations on Science, Culture and Time. University of Michigan Press, Ann Arbor, MI.

Sheppard, E., 2002. The spaces and times of globalization: place, scale, networks, and positionality. Economic Geography 78, 307–330.

Sunley, P., 2008. Relational economic geography: a partial understanding or a new paradigm? Economic Geography 84 (1), 1–26.

Warf, B., 2008. From surfaces to networks. In: Warf, B., Arias, S. (Eds.), The Spatial Turn: Interdisciplinary Perspectives. Routledge, London, pp. 59–76.

Yeung, H., 2005. Rethinking relational economic geography. Transactions of the Institute of British Geographers 30, 37–51.

Networks in Economics

Yves Zenou, Stockholm University, Stockholm, Sweden; and IFN, Stockholm, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

We provide an overview on networks in economics. We first look at the theoretical aspects of network economics using a game-theoretical approach. We derive some results on games on networks and network formation. We also study what happens when agents choose both links and actions. We then examine how these models can be used to address some applied and empirical relevant questions by mainly focusing on labor market networks and crime networks. We provide some empirical evidence on these two types of networks and address some policy implications of the models.

Introduction

The study of social networks was initiated by sociologists more than a century ago and has grown to be a central field of sociology over the past 50 years (see, e.g., Wasserman and Faust, 1994). Over that same period, a mathematical literature on the structure and properties of graphs has been developed and extensively studied (see, e.g., Bollobás, 1998). A recent awakening of interest in social networks has occurred in the computer science and statistical physics literature, mainly over the past 15 years (see Albert and Barabási, 2002; Newman, 2010; for an overview of these studies). While the importance of embeddedness of economic activity in social settings has been fundamental to sociologists for some time, it was largely ignored by economists until the last decade. This is surprising given that nonmarket interactions, i.e., interactions between agents that are not mediated by the market, are crucial to explain different economic phenomena such as stock market crashes, growth, education, religion, crime, etc. The studies of networks with economic perspectives and using game-theoretic modeling techniques have only emerged over the last decade (see Goyal, 2007; Jackson, 2008).

In the present article, we would like to survey the recent literature on networks in economics. As in all surveys and because of space constraint, we will not be able to cover all aspects of the literature and therefore we refer to other overviews such as Goyal (2007), Jackson (2008), de Martí and Zenou (2011) and Jackson and Zenou (forthcoming). (See also Jackson and Zenou (2013) for a collection of the most important published papers in the economics of networks.) We will expose what we believe are the most important aspects of the economics of networks and will illustrate each aspect by a simple theoretical model.

This article is divided in two parts. In the first one, we will look at the theoretical aspects of network economics using a game-theoretical approach. In this part, we will look first at games on networks, which take networks as given and focus on the impact of their structure on individuals' outcomes. We will then analyze network formation and explain the way links between agents are formed. We will also study what happens when agents choose both links and actions. In the second part of this paper, we will see how the theoretical models can be used to address applied and empirical relevant questions. We will mainly focus on labor-market networks and crime networks. We will also provide some empirical evidence on these two types of networks and address some policy implications of the models.

Networks in Economics: Theoretical Aspects

Games on Networks

We begin with a class of canonical and widely applicable games; specifically, games where there is a fixed and given network of interactions. Links indicate which players' strategies affect which others' payoffs. In particular, a given player's payoff depends only on the play of the neighbors. Of course, this still results in network effects since there may be chains of influence.

Network definitions

We provide some basic definitions on networks.

A set of players $N = \{1,\ldots, n\}$ are connected by a network.

A graph or network is a pair $(N, \mathbf{g})$, where $\mathbf{g}$ is a network on the set of nodes N. To each network $\mathbf{g}$, we associate its adjacency matrix $\mathbf{G} = [g_{ij}]$. A graph is undirected if $\mathbf{G}$ is required to be symmetric so that $g_{ij} = g_{ji}$, and is directed otherwise. It is useful to use the notation $ij \in \mathbf{g}$ to indicate that $g_{ij} = 1$ and $ij \notin \mathbf{g}$ to indicate that $g_{ij} = 0$, and one can represent a graph by the set of links that are present (so one could alternatively represent $\mathbf{g}$ by its set of links). A weighted adjacency matrix, $\mathbf{G}$, is when g_{ij} can take values other than 0 and 1.

A relationship between two nodes i and j, represented by $ij \in \mathbf{g}$, is referred to as a link. Links are also referred to as edges or ties in various parts of the literature, and sometimes also directed links, directed edges, or arcs in the specific case of a directed network.

A walk in a network $(N, \mathbf{g})$ refers to a sequence of nodes, i_1, i_2, $i_3,\ldots, i_{K-1}$, i_K, such that $i_k i_{k+i} \in \mathbf{g}$ for each k from 1 to K. The length of the walk is the number of links in it, or $K - 1$. A path in a network $(N, \mathbf{g})$ is a walk in $(N, \mathbf{g})$, i_1, i_2, $i_3,\ldots, i_{K-1}$, i_K, such that all the nodes are distinct.

The distance $d(i, j)$ between two nodes i and j in the same component of a network is the length of a shortest path (also known as a geodesic) between them.

The neighbors of a node i in a network $(N, \mathbf{g})$ are denoted by $N_i(\mathbf{g})$, so that

$$N_i(\mathbf{g}) = \{j | ij \in \mathbf{g}\}$$

The degree of a node i in a network $(N, \mathbf{g})$ is the number of neighbors that i has in the network, so that $d_i(\mathbf{g}) = |N_i(\mathbf{g})|$. Unless otherwise stated, let us suppose that $g_{ii} = 0$, so that nodes are not linked to themselves.

The kth power $\mathbf{G}^k = \mathbf{G}^{(k \text{ times})}\mathbf{G}$ of the adjacency matrix $\mathbf{G}$ keeps track of indirect connections in $\mathbf{g}$. More precisely, the coefficient $g_{ij}^{[k]}$ in the (i, j) cell of $\mathbf{G}^k$ gives the number of paths of length k in g between i and j. In particular, $\mathbf{G}^0 = \mathbf{I}$. Note that, by definition, a path between i and j need not follow the shortest possible route between those agents. For instance, when $g_{ij} = 1$, the sequence $ij \rightarrow ji \rightarrow ij$ constitutes a path of length three in g between i and j.

Game on network definitions

When the network is fixed, there are two types of games: games with strategic complementarities and games with strategic substitutes. Let us define them formally. Players have effort spaces X_i and payoff functions $u_i = X_i \rightarrow \mathbb{R}$. The action spaces are finite sets or subsets of a Euclidean space. A given player's payoff depends on other players' behaviors, but in particular only on those to whom the player is connected in the network. More formally, i's payoff depends only on x_i and $\{x_j\}$ $X_j \in N_i$ (g) and can be written as $u_i(x_i, x_j)$, where $x_i \in X_i$ is the effort of player i. We have the following definitions.

A game exhibits strategic complements if it exhibits increasing differences, i.e., for all i, j, with $i \neq j$, $x_i \geq x_i'$ and $x_j \geq x_j'$:

$$u_i(x_i, x_j) - u_i(x_i', x_j) \geq u_i\left(x_i, x_j'\right) - u_i\left(x_i', x_j'\right).$$

A game exhibits strategic substitutes if it exhibits decreasing differences, i.e., for all i,j, with $i \neq j$, $x_i \geq x_i'$ and $x_j \geq x_j'$:

$$u_i(x_i, x_j) - u_i(x_i', x_j) \leq u_i\left(x_i, x_j'\right) - u_i\left(x_i', x_j'\right).$$

These notions are said to apply strictly if the inequalities above are strict whenever $x_i > x_i'$ and $x_j > x_j'$.

Because of the lack of space, we will only expose games with strategic complementarities since they have been more studied, have nice properties, and have nicer and more natural applications in economics. (Prominent papers studying games with strategic substitutes are Bramoullé and Kranton (2007) and Galeotti and Goyal (2010). For an overview of the literature on games on networks, see Jackson and Zenou (2013).)

A simple model with strategic complementarities

We assume perfect information, i.e., the players know everything about the network and other players' actions. For games on networks with incomplete information, see Calvó-Armengol and de Marti (2009), Galeotti et al. (2009), Hagenbach and Koessler (2011) and Acemoglu et al. (2011).

The game: Consider a game in which n agents, linked to each other in a network $\mathbf{g}$, must decide how much effort to exert in some activity. We denote by x_i the effort level of agent i and by $\mathbf{x} = (x_i, \ldots, x_n)^{\mathrm{T}}$ the vector of efforts of all agents (^{T} means transposition). Each agent i selects an effort $x_i \geq 0$, and obtains a payoff $u_i(\mathbf{x}, \mathbf{g})$ that depends on the effort profile $\mathbf{x}$ and on the underlying network $\mathbf{g}$, in the following way:

$$u_i(\mathbf{x}, \mathbf{g}) = ax_i - \frac{1}{2}x_i^2 + \phi \sum_{j=1}^{n} g_{ij}x_ix_j \qquad [1]$$

where a, $\phi > 0$. In this linear-quadratic utility function, agents are ex ante homogeneous in terms of observable characteristics (i.e., they all have the same a) and their heterogeneity only stems from their position in the network. The first two terms of eqn [1], $ax_i - (1/2)x_i^2$, give the individual benefits and costs of providing effort x_i. The last term of this utility function, $\phi \sum_{j=1}^{n} g_{ij}x_ix_j$, reflects the influence of direct links on own action. This peer effect component can be heterogeneous, and this endogenous heterogeneity reflects the different locations of individuals in the network $\mathbf{g}$ and the resulting effort levels. More precisely, bilateral influences are captured by the following cross derivatives, for $i \neq j$:

$$\frac{\partial^2 u_i(\mathbf{x}, \mathbf{g})}{\partial x_i \partial x_j} = \phi g_{ij} \geq 0 \qquad [2]$$

When i and j are directly linked, i.e., $g_{ij} = 1$, the cross derivative is $\phi > 0$ and reflects strategic complementarity in efforts. When i and j are not direct friends, i.e., $g_{ij} = 0$, this cross derivative is zero.

The Katz–Bonacich network centrality

Define

$$\mathbf{M}(\mathbf{g}, \phi) = (\mathbf{I} - \phi\mathbf{G})^{-1} = \sum_{k=0}^{+\infty} \phi^p \mathbf{G}^p$$

where $\mathbf{I}$ is the identity matrix. Denote by $\mathbf{1}$ the column vector of 1. We have the following definition:

Definition 1 (Katz, 1953; Bonacich, 1987): Given $\phi > 0$, a small enough scalar, the vector of Katz–Bonacich centralities of parameter ϕ in network $\mathbf{g}$ is defined as

$$\mathbf{b}(\mathbf{g}, \phi) = (\mathbf{I} - \phi\mathbf{G})^{-1}\mathbf{1} = \mathbf{M}(\mathbf{g}, \phi)\mathbf{1} = \sum_{p=0}^{+\infty} \phi^p \mathbf{G}^p \mathbf{1}. \qquad [3]$$

The Katz–Bonacich centrality of node i is thus $b_i(\mathbf{g}, \phi) = \sum_{j=1}^{n} m_{ij}(\mathbf{g}, \phi)$, and counts the total number of walks in $\mathbf{g}$ starting from i. It is the sum of all loops $m_{ii}(\mathbf{g}, \phi)$ from i to i itself, and of all the outer walks $\sum_{j \neq i} m_{ij}(\mathbf{g}, \phi)$ from i to every other player $j \neq i$, i.e.,

$$b_i(\mathbf{g}, \phi) = m_{ii}(\mathbf{g}, \phi) + \sum_{j \neq i} m_{ij}(\mathbf{g}, \phi).$$

By definition, $m_{ii}(\mathbf{g}, \phi) \geq 1$, and thus $b_i(\mathbf{g}, \phi) \geq 1$, with equality when $\phi = 0$.

Nash equilibrium

In game theory, the Nash equilibrium (named after the Nobel-prize winner John F. Nash) is a solution concept of a game involving two or more players, in which no player has anything to gain by changing only her own strategy (i.e., by changing unilaterally). If each player has chosen a strategy and no player can benefit by changing her strategy while the other players keep theirs unchanged, then the current set of strategy choices constitute a Nash equilibrium.

Let us now characterize the Nash equilibrium of the game where agents choose their effort level $x_i \geq 0$ simultaneously. Denote by $\mu_1(\mathbf{G})$ the spectral radius of $\mathbf{G}$. Ballester et al. (2006) have shown the following result:

Proposition 1: *If* $\phi\mu_1(\mathbf{G}) < 1$, the game with payoffs has a unique interior Nash equilibrium in pure strategies given by

$$\mathbf{x}^* = a\mathbf{b}(\mathbf{g}, \phi) \qquad [4]$$

This results shows that the Katz–Bonacich centrality is the right network index to account for equilibrium behavior when the utility functions are linear-quadratic. (There are many other centrality measures such as the betweeness or the closeness centrality. See Wasserman and Faust (1994) for an overview on the different possible centrality measures.) In eqn [1], the local payoff interdependence is restricted to direct network contacts. At equilibrium, however, this local payoff interdependence spreads all over the network through the overlap of direct link clusters. The Katz–Bonacich centrality precisely reflects how individual decisions feed into each other along any direct and indirect network path. The condition $\phi\mu_1(\mathbf{G}) < 1$ stipulates that local complementarities must be small enough compared to own concavity, which prevents multiple equilibria to emerge and, at the same time, rules out corner solutions. The condition $\phi\mu_1(\mathbf{G}) < 1$ also guarantees that $(\mathbf{I} - \phi\mathbf{G})$ is invertible and its series expansion well defined.

Proposition 1 has characterized the equilibrium effort x_i^* of each agent i as equals to the Bonacich centrality. We can now characterize the equilibrium utility of i as a function of the Bonacich centrality. It is easily verified that

$$u_i(\mathbf{x}^*, \mathbf{g}) = \frac{1}{2} x_i^{*2} = \frac{1}{2} [ab_i(\mathbf{g}, \phi)]^2$$

In other words, the equilibrium utility of each agent i is positively related to the Katz–Bonacich centrality.

Network Formation

In the previous section, we have seen how the position in the network of each player (i.e., the Katz–Bonacich centrality) affects outcomes when the network is fixed. We now look at network formation but agents cannot decide how much effort to exert in some activity. They only choose with whom they want to form a link.

Static models of network formation

Myerson (1977, 1991) provides an early formulation of a network formation game. The structure of the game is simple: agents have to decide about their potential partners and their strategies consist in naming those with whom they want to form a link with. For a link to be formed, it has to be that two individuals name each other, i.e., there needs to be mutual consent in link creation. The Nash equilibrium concept defined above can then be used to find out which strategy profiles are stable and, hence, which networks are the possible outcomes of the game.

The main problem of using the Nash equilibrium concept is that it exacerbates the coordination problems that arise when all agents are simultaneously deciding about links. In particular, the empty network (i.e., nobody forms a link) is always a Nash equilibrium of this game since no deviation is profitable. Indeed, deviating means here to name someone as a friend but this will not generate a link because of mutual consent. In general, with the Myerson game, there are many equilibria, which reduces the attractiveness of this approach.

A solution to this problem has been proposed by Jackson and Wolinsky (1996). In their seminal paper, they introduce an alternative solution concept for network formation games, namely, pairwise stability, which provides a network equilibrium notion. Let us define it.

Consider some payoff function $u(\mathbf{g}) = (u_1(\mathbf{g}), \ldots, u_n(\mathbf{g}))$ that assigns a payoff to every agent in N as a function of the underlying network $\mathbf{g}$ connecting them.

Definition 2: A network $\mathbf{g}$ is pairwise stable for the payoff function u if and only if

1. for all $ij \in \mathbf{g}$, $u_i(\mathbf{g}) \geq u_i(\mathbf{g} - ij)$ and $u_j(\mathbf{g}) \geq u_j(\mathbf{g} - ij)$
2. for all $ij \notin \mathbf{g}$, if $u_i(\mathbf{g}) < u_i(\mathbf{g} + ij)$, then $u_j(\mathbf{g}) > u_j(\mathbf{g} + ij)$

In other words, a network is pairwise stable if (1) no player gains by cutting an existing link, and (2) no two players not yet connected both gain by creating a direct link with each other. Pairwise stability thus only checks for one-link deviations. (This weak equilibrium concept is often interpreted as a necessary condition for stronger stability concepts.) It requires that any mutually beneficial link be formed at equilibrium but does not allow for multilink severance. (There are other network equilibrium concepts, which are often refinements from the pairwise-stability equilibrium notion. See Bloch and Jackson (2006) for a complete overview on this issue. In particular, they compare variations on three types of definitions: those based on a pairwise-stability notion, those based on the Nash equilibria of a link formation game, and those based on equilibria of a link formation game where transfers are possible.) This notion takes into account the individual incentives to create and sever links and the necessary mutual consent between both sides for a link to be formed. (Bala and Goyal (2000) extend the Myerson model described above by only considering directed networks, i.e., individuals form links unilaterally without requiring the consent of the other party to create a link.)

Pairwise-stable networks can be interpreted as the limiting graphs of a dynamic procedure of network formation. Suppose, indeed, that players myopically add or sever links to improve their current status, and that only one link is added or removed at a time. When this process converges, the networks ultimately reached are pairwise stable. (Note that network transitions only concern one link at time. The dynamics may thus lock in (the set of limiting graphs need not be unique) or cycle. See Jackson and Watts (2002) for more details.)

Let us now define strong efficiency.

Definition 3: A network $\mathbf{g}^E \subset \mathbf{g}^N$ is strongly efficient if

$$W(\mathbf{g}^E) \equiv \sum_{i \in \mathbf{g}^E} u_i(\mathbf{g}^E) \geq W(\mathbf{g}) \equiv \sum_{i \in \mathbf{g}} u_i(\mathbf{g}) \quad \text{for all } \mathbf{g} \subset \mathbf{g}^N$$

Observe that this concept of efficiency is utilitarian and not Paretian since this definition says that the socially efficient network is the one for which the sum of utilities of all agents is the highest among all possible networks with n agents.

A simple model: the connections model

Jackson and Wolinsky (1996) have proposed a simple model where the utility function of each individual i in a network $\mathbf{g}$ is given by

$$u_i(\mathbf{g}) = \sum_{j \neq i} \delta^{d(i,j)} - \sum_{j \in N_i(g)} c = \sum_{j \neq i} \delta^{d(i,j)} - d_i(\mathbf{g})c \qquad [5]$$

where $0 < \delta < 1$ gives the value of the link between two individuals i and j, $d(i,j)$ is the geodesic distance between i and j, i.e., the length of any shortest path between i and j, and finally c is the cost of forming a link. In this formulation, when individual i has a direct link with j, he/she first benefits from this link by obtaining δ (since, for any direct link, $d(i, j) = 1$) but then he/she may benefit from the direct links of j by obtaining $\delta^{d(i,j)}$ (indirect connections, which are weighted by a decay factor).

Observe that if two agents create a link between them it is never detrimental to a third party (positive externalities). This is because creating a new link does not affect the cost of a third player (he/she still has the same number of direct links) but can affect his/her gains since a new link can shorten the distance between a third player and another one. Here is an example (Figure 1) of utilities for the following star network.

We have the following result for the connections model (Jackson and Wolinsky, 1996):

Proposition 2: In the connections model in which the utility of each player $i = 1, \ldots, n$ is given by eqn [5], we have

1. A pairwise stable network has at most one (nonempty) component.
2. For $c < \delta - \delta^2$, the unique pairwise stable network is the complete graph g^N.
3. For $\delta - \delta^2 < c < \delta$ a star encompassing all players is pairwise stable, but not necessarily the unique pairwise stable graph.
4. For $\delta < c$, any pairwise stable network that is nonempty is such that each player has at least two links.

This proposition characterizes the pairwise-stable equilibria. In particular, it says that if $\delta - \delta^2 < c < \delta$, then the star-shaped is a pairwise stable equilibrium. Let us check if this is true for the network described in Figure 1. If the star agent, player 2, wants to delete a link (he/she cannot create any new link), his/her net utility is $c - \delta$. If $c < \delta$, this will never happen. If a peripheral agent (say player 1) deletes a link, his/her net utility is $c - \delta$ while if he/she creates a link (with player 3), his/her net utility is $\delta - c - \delta^2$. Any of these two actions will never take place if $\delta - \delta^2 < c < \delta$. It is easily verified that this same condition holds for a star-shaped network with n agents to be pairwise stable. As can be seen in this proposition, one of the main problems of pairwise-stable networks is that it is very difficult to provide a full characterization of the set of equilibrium networks. (It is indeed well known that noncooperative games of network formation with nominal lists of intended links are plagued by coordination problems (Myerson, 1991; Jackson, 2008). Cooperative-like stability concepts solve them partially, but heavy combinatorial costs still jeopardize a full characterization.)

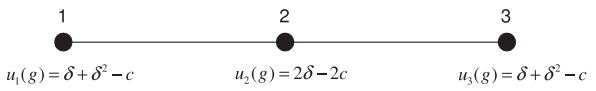

Figure 1 Utility values in a star-shaped network.

Let us now give a result in terms of efficiency.

Proposition 3: The unique strongly efficient network in the symmetric connections model, where the utility of each player $i = 1,\ldots,n$ is given by eqn [5], is

1. the complete graph g^N if $c < \delta - \delta^2$.
2. a star encompassing all players if $\delta - \delta^2 < c < \delta + ((n-2)/2)\delta^2$
3. no links (the empty network) if $\delta + ((n-2)/2)\delta^2 < c$.

There is therefore a tension between stability and efficiency. Take part (4) of Propositions 2. In the high cost range ($c > \delta$), the only (nondegenerated) pairwise-stable networks are those that are overconnected from an efficiency perspective. For example, when $c > \delta$, the star cannot be pairwise stable but it is an efficient network.

A second model: the coauthor model

In this case, for $n_i > 0$ (it is assumed that for $n_i = 0$, $u_i(\mathbf{g}) = 0$), the utility function of a player $i = 1,\ldots, n$ in network $\mathbf{g}$ is given by

$$u_i(\mathbf{g}) = \sum_{j \in N_i(\mathbf{g})} \left[\frac{1}{d_i} + \frac{1}{d_j} + \frac{1}{d_i d_j}\right]$$

where $d_i \equiv d_i(\mathbf{g})$ is the number of direct links player i has (or equivalently the number of projects player i is involved in). Here, each player is a researcher and a link ij is a joint project (or joint paper or collaboration) between players i and j.

Each researcher has a fixed amount of time to spend on research and so the time each researcher i spends on a given project is inversely related to the number of project d_i he/she is involved in. The synergy between two researchers i and j are captured by the term $1/(d_i d_j)$. Here, the more projects each researcher is involved with, the lower the synergy that is obtained per project. There is no direct cost of forming a link but an indirect cost because of congestion.

Contrary to the connections model, when two agents create a link, it is never beneficial to a third party (negative externalities). This is because creating a new link does affect the time the researcher will spend on other projects with the researchers he/she is already connected to.

Proposition 4: In the coauthor model,

1. if n is even, then the strongly efficient network is a graph consisting of $n/2$ separate pairs (dyads), and
2. a pairwise-stable network can be partitioned into fully intraconnected components, each of which has a different number of members. If k is the number of members of one such component and k' is the next largest in size, then $k > (k')^2$.

Choosing Both Actions and Links

There are different models that have combined actions' and links' choices. For example, Bloch and Dutta (2009) proposed a model where both link intensities and network formation are modeled. For games with strategic substitutes, Galeotti and

Goyal (2010) have developed a static model where players choose both actions and links while López-Pintado (2008) investigates a dynamic one. Because we have focused on games with strategic complements, we would like now to expose the static model of Cabrales et al. (2011). (See also Konig et al. (forthcoming) for a dynamic model of network formation with strategic complements in efforts.)

Consider a simultaneous move game of network formation (or social interactions) and investment. $N = \{1,\ldots, n\}$ is a finite set of players, and $T = \{1, \ldots, t\}$ is a finite set of types for these players. We let n be a multiple of t, i.e., $n = mt$ for some integer $m \geq 1$, so that there is the same number of players of each type. The case $n = t$ is referred to as the baseline game and the general case $n = mt$ as the m–replica of this baseline game. For each player $i \in N$, we denote by $\tau(i) \in T$ his/her type. In an m–replica game, there are exactly m players of each type $\tau \in T$. This replica game allows us to take limits as the population becomes large without having to specify the types of the new individuals that are added.

Let $c > 0$. Player i's utility is equal to

$$u_i(\mathbf{x}, \mathbf{s}) = a_{\tau(i)}x_i + \phi \sum_{j=1, j\neq i}^{n} g_{ij}(\mathbf{s})x_i x_j - \frac{1}{2}cx_i^2 - \frac{1}{2}s_i^2 \qquad [6]$$

where $x_i \geq 0$ is the productive effort taken by player i, with $\mathbf{x} = (x_1,\ldots, x_n)$ being a profile of productive efforts while $s_i \geq 0$ is the socialization effort of player i, with $\mathbf{s} = (s_1,\ldots, s_n)$ being a profile of socialization efforts. This utility is very similar to the one defined in eqn [1] with the difference that the socialization effort s_i is also included. In eqn [6], the returns to the investment are the sum of a private component and a synergistic component. The private returns are heterogeneous across players and depend on their type. We denote by $\mathbf{a} = (a_1,\ldots, a_t)$ the profile of these private returns, where $0 < a_1 < a_2 < \ldots < a_t$. Payoffs have nonnegative cross effects, i.e.,

$$\frac{\partial^2 u_i(\mathbf{x}, \mathbf{s})}{\partial x_i \partial x_j} = \phi g_{ij}(\mathbf{s}), \quad \text{for all } i \neq j, \qquad [7]$$

reflecting strategic complementarities in productive investments ($\phi \geq 0$ corresponds to the level of synergistic returns). As in eqn [1], the size $\phi g_{ij}(\mathbf{s}) \geq 0$ of these complementarities depends on the profile of socialization efforts, and varies across different pairs of players.

The key innovation is $g_{ij}(\mathbf{s})$, which determines link formation. Players i and j interact with a link intensity given by

$$g_{ij}(\mathbf{s}) = \frac{s_i s_j}{\sum_{j=1}^{n} s_j} \qquad [8]$$

By definition, links are symmetric, i.e., $g_{ij} = g_{ji}$. We also allow for self-loops (when $i = j$). From eqn [8], we can determine the adjacency matrix $\mathbf{G}$ of this network where each cell ij is $0 < g_{ij} < 1$.

A key innovation of this article is that the synergistic effort $\mathbf{s}$ is generic within a community – a scalar decision. Socializing is not equivalent to elaborating a nominal list of intended relationships, as in the literature on network formation surveyed by Jackson (2008) and exposed above. We have seen (see Proposition 2) that this leads to too many equilibria and that it is very difficult to provide a full characterization of the set of equilibrium networks. To avoid this problem, here, network formation is not the result of an earmarked socialization process so that agents put efforts in socialization that may result in links between agents. This choice of a model without earmarked socialization greatly improves the tractability of the analysis. Unlike with richer models of link formation a la Jackson and Wolinsky (1996), we can resort to off-the-shelf Nash equilibrium analysis without being burdened by the extreme (combinatorial) multiplicity problems of the other models. As a result, the authors can perform a standard type of equilibrium analysis that equates marginal costs and benefits of both production and socialization. Define:

$$\Phi(\mathbf{a}) = \phi \frac{\sum_{\tau=1}^{t} a_\tau^2}{\sum_{\tau=1}^{t} a_\tau}. \qquad [9]$$

Cabrales et al. (2011) demonstrate the following result.

Proposition 5: Suppose that $2\,(c/3)^{3/2} > \Phi(\mathbf{a}) > 0$. Then, there exists an m^* such that for all m–replica games with $m \geq m^*$, there are exactly two stable interior pure strategy Nash equilibria. These pure strategy Nash equilibria are such that, for all players i of type τ, the strategies (s_i, x_i) converge to $(s^*_{\tau(i)}, x^*_{\tau(i)})$ as m goes to infinity, where $s^*_{\tau(i)} = a_{\tau(i)}s$, $x^*_{\tau(i)} = a_{\tau(i)}x$, and (s, x) are positive solutions to

$$\begin{cases} s = \Phi(\mathbf{a})x^2 \\ x[c - \Phi(\mathbf{a})s] = 1 \end{cases} \qquad [10]$$

In other words, when $\Phi(\mathbf{a})$ is small enough compared to the inframarginal cost for a productive investment, the system of eqn [10] with two unknowns has exactly two positive solutions. As m gets large, each such solution gets arbitrarily close to a pure strategy Nash equilibrium of the corresponding m–replica game. We get two approximate Nash equilibria. Besides, as m gets large, every pure strategy Nash equilibrium gets arbitrarily close to a solution of eqn [10]. The equilibrium multiplicity identified in Proposition 5 reflects an intertwinned coordination problem in the socialization process and in the production technology.

The authors are then able to compare the actions and payoffs of players across the two approximate equilibria characterized in Proposition 5. They show that the equilibrium actions can be ranked component-wisely and the equilibrium payoffs can be Pareto-ranked accordingly. There is a Pareto-superior approximate equilibrium $(\mathbf{s}^*, \mathbf{x}^*)$ and a Pareto-inferior approximate equilibrium $(\mathbf{s}^{**}, \mathbf{x}^{**})$ while the socially efficient outcome lies in between the two equilibria. Formally, $(\mathbf{s}^*, \mathbf{x}^*) \geq (\mathbf{s}^E, \mathbf{x}^E) \geq (\mathbf{s}^{**}, \mathbf{x}^{**})$ and $\mathbf{u}(\mathbf{s}^E, \mathbf{x}^E) \geq \mathbf{u}(\mathbf{s}^*, \mathbf{x}^*) \geq \mathbf{u}(\mathbf{s}^{**}, \mathbf{x}^{**})$, where $\geq$ is the component-wise ordering. This simple model has shown how the set of equilibria can be reduced and equilibria can be characterized when agents do not direct their links but instead choose some effort in socialization.

Networks in Economics: Applications and Empirical Aspects

We would like now to study two important applications of economic networks: labor market and crime networks.

Labor-market Networks

There is a host of evidence showing that social networks are pervasive in the labor market. For instance, Holzer (1987, 1988) documents that, in the United States, among 16- to 23-year-old workers who reported job acceptance, 66% use informal search channels while only 11% use state agencies and 10% newspapers. (See also Granovetter (1995) for additional evidence.) Topa (2001) argues that the observed spatial distribution of unemployment in Chicago is consistent with a model of local interactions and information spillovers, and may thus be generated by an agent's reliance in informal methods of job search such as networks of personal contacts. Similarly, Bayer et al. (2008) document that people who live close to each other, defined as being in the same census block, tend to work together, i.e., in the same census block. (For a survey of the literature on social interactions and the labor market, see Ioannides and Loury (2004).) Let us now provide a simple model that captures these facts.

The model

Let us describe the model of Calvó-Armengol and Jackson (2004). Time evolves in discrete periods indexed by t. The vector σ_t describes the employment status of the workers at time t. If individual i is employed at the end of period t, then $\sigma_{it} = 1$ and if i is unemployed then $\sigma_{it} = 0$.

A period t begins with some agents being employed and others not, as described by the vector $\boldsymbol{\sigma}_{t-1} = (\sigma_{1t-1},\ldots,\sigma_{nt-1})$ that gives the employment status of all workers from the last period. Next, information about job openings arrives. In particular, any given individual hears about a job opening with probability γ i.e. between 0 and 1. This job arrival process is independent across individuals. If the individual is unemployed, then he/she will take the job. However, if the individual is already employed then he/she will pass the information along to a friend, picked at random among unemployed friends. As stated above, the graph or network $\mathbf{g}$ summarizes the links of all agents, where $g_{ij} = 1$ indicates that i and j know each other (strong tie), and share their knowledge about job information, while $g_{ij} = 0$ indicates that they do not know each other. Finally, the last thing that happens in a period is that some agents lose their jobs. This happens randomly according to an exogenous breakup rate, δ, which is between 0 and 1. We are able to write the probability $\mathbb{P}_{ij}$ of the joint event that individual learns about a job and this job ends up in individual j's hands. It is equal to

$$\mathbb{P}_{ij}(\boldsymbol{\sigma}) = \begin{cases} \gamma & \text{if} \quad \sigma_i = 0 \text{ and } i = j \\ \gamma / \sum_{k:\sigma_k=0} g_{ik} & \text{if} \quad \sigma_i = 1,\ \sigma_j = 0, \text{ and } g_{ij} = 1 \\ 0 & \text{otherwise} \end{cases} \qquad [11]$$

where the vector σ describes the employment status of all the individuals at the beginning of the period. In eqn [11], γ is the probability of obtaining a job information without using friends and relatives. Three cases may then arise. If individuals i and j are unemployed ($\sigma_i = \sigma_j = 0$), then the probability that j will obtain a job is just γ since individual i will never transmit any information to j. If individual i is already employed and his/her friend j is not ($\sigma_i = 1$, $\sigma_j = 0$), then individual i transmits this job information to all his/her direct unemployed neighbors, whose total number is $\sum_{k:\sigma_k=0} g_{ik}$. We assume that all unemployed neighbors are treated on equal footing, meaning that the employed worker who has the job information does not favor any of his/her direct neighbors. As a result, the probability that an unemployed worker j is selected among the $\sum_{k:\sigma_k=0} g_{ik}$ unemployed direct neighbors of an employed worker j is given by $\gamma / \sum_{k:\sigma_k=0} g_{ik}$. Finally, if individual j is employed, then he/she does not need any job information, at least in the current period.

The first result obtained by Calvó-Armengol and Jackson (2004) is the following.

Proposition 6: The higher $d_i(\mathbf{g})$, the number of strong ties (direct friends) individual i has, the higher is his/her individual probability of finding a job.

Indeed, if an individual has more strong ties (direct friends), then he/she is more likely to hear on average about more jobs through them but his/her chance of finding a job directly does not increase since γ is not affected by the size of the network. This result is quite intuitive since, when the number of direct connections increases, the source of information about jobs is larger and people find it easier to obtain a job through their friends and relatives. This is the first prediction of this model, which implies that, the greater chance workers have of finding a job, the higher is the number of their strong ties. Observe that the individual probability of finding a job through strong ties for individual j is obviously not given by eqn [11] since $\mathbb{P}_{ij}(\mathbf{s})$ is the probability that only one individual, i, who hold a strong tie with j, and who is aware of some job, will transmit this information to individual j. To determine the individual probability of obtaining a job for j, one has to do the calculation for all the direct friends of i.

We would now like to study the impact of weak ties (indirect friends or path-connected friends) on the individual probability of finding a job. Calvó-Armengol and Jackson (2004) show that, in steady state, there is a positive correlation in employment status between two path-connected workers. This result is not at all easy to obtain since, in the short run, the correlation is negative. Indeed, in a static model, if an employed worker is directly linked to two unemployed workers, then if he/she is aware of a job, he/she will share this job information with two unemployed friends (see eqn [11]). These two persons, who are path connected (path of length two), are thus in competition and one (randomly chosen) will obtain the job and be employed while the other will stay unemployed. So their employment statuses will be negatively correlated (see Calvó-Armengol, 2004).

Let us now give the intuition why this negative correlation result does not hold in a dynamic labor-market model. Consider the star-shaped network described in Figure 2 with three individuals, i.e., $n = 3$ and $g_{12} = g_{23} = 1$. Suppose the employment status of these three workers from the end of the last period is $\sigma_{t-1} = (0,1,0)$. In the figure, a black node represents an employed worker (individual 2), while unemployed workers (1 and 3) are represented by white nodes. Conditional

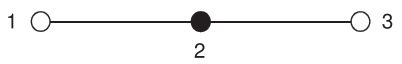

Figure 2 Employment correlations in a star-shaped network.

on this state $\boldsymbol{\sigma}_{t-1}$, the employment states σ_{1t} and σ_{3t} are negatively correlated. As stated above, this is due to the fact that individuals 1 and 3 are 'competitors' for any job information that is first heard by individual 2.

Despite this negative (conditional) correlation in the short run, individual 1 can benefit from individual 3's presence in the longer run. Indeed, individual 3's presence helps improve individual 2's employment status. Also, when individual 3 is employed, individual 1 is more likely to hear about any job that individual 2 hears about. These two aspects counter the local (conditional) negative correlation, and help induce a positive correlation between the employment status of individuals 1 and 3.

Proposition 7: Under fine enough subdivisions of periods, the unique steady-state long-run distribution on employment is such that the employment statuses of any path-connected agents are positively correlated.

The proposition shows that, despite the short-run conditional negative correlation between the employment of competitors for jobs and information, in the longer run any interconnected workers' employment is positively correlated. This implies that there is a clustering of agents by employment status, and employed workers tend to be connected with employed workers, and vice versa. The intuition is clear; conditional on knowing that some set of agents are employed, it is more likely that their neighbors will end up receiving information about jobs, and so on. The benefits from having other agents in the network outweigh the local negative correlation effects, if we take a long-run perspective.

Proposition 8: The longer the length of two-path connected individuals (i.e., weak ties), the lower is the correlation in employment statuses between these two individuals.

Indeed, the correlation between two agents' employment is (weakly) decreasing in the number of links that each an agent has, and the correlation between agents' employment is higher for direct compared to indirect connections. The decrease as a function of the number of links is due to the decreased importance of any single link if an agent has many links. The difference between direct and indirect connections in terms of correlation is due to the fact that direct connections provide information, while indirect connections only help by indirect provision of information that keeps friends, friends of friends, etc., employed. In other words, the longer the path in the social network between two individuals, the weaker is the effect of job transmission.

Empirical test

Using individual-level data from the UK Labour Force Survey, Patacchini and Zenou (2012a) test this model by looking at the employment prospects of ethnic minorities. To the traditional determinants of employment rates (sex, age, education, years since arrival in the United Kingdom, percentage of high-skilled leaving nearby, etc.), they add the local ethnic employment density bands based on travel time between areas.

Because they do not have direct data on social networks, Patacchini and Zenou (2012a) conjecture that the social space is highly correlated to the physical space for ethnic minorities in relatively small areas. This approximation is based on the fact that, in relatively small areas, ethnic minorities of a given group (say Indians or Pakistanis) are likely to interact with other ethnic minorities from the same group and thus exchange information about jobs. In other words, ethnic employment density is interpreted as a proxy for the strength of social contacts in delivering information about jobs.

Patacchini and Zenou (2012a) test Propositions 6 and 8 described above, which show that the individual probability of finding a job increases with the number of strong ties and weak ties, and the longer the length of weak ties, the lower is this probability.

They find that the higher is the percentage of employed workers from a given ethnic group living nearby, the higher is the probability of finding a job through social networks (Proposition 6). This effect decays, however, very rapidly with distance, losing significance beyond approximately an hour travel time (Proposition 8). They argue that local social interactions between people of the same ethnicity can explain this positive relationship and its spatial trend.

Let us now investigate our second application: criminal networks.

Criminal Networks

It is well established that delinquency is, to some extent, a group phenomenon, and the source of crime and delinquency is located in the intimate social networks of individuals (see, e.g., Sutherland, 1947; Sarnecki, 2001; Warr, 2002). Indeed, delinquents often have friends who have themselves committed several offenses, and social ties among delinquents are seen as a means whereby individuals exert an influence over one another to commit crimes.

A simple model

Glaeser et al. (1996) were among the first to model criminal social interactions. In their model, criminal interconnections act as a social multiplier on aggregate crime. They impose, however, a specific network structure, the circle, for the location of criminals. Following Calvó-Armengol and Zenou (2004) and Ballester et al. (2010), we would like to propose a more general model that can encompass any social network. For that, we will use the model described above by reinterpreting it in terms of criminal activities.

Denote by x_i the delinquency effort level (i.e., how often they commit crime) of delinquent i, and by $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ the population delinquency profile. Delinquents in network $\mathbf{g}$ decide how much effort to exert. Each agent i selects an effort $x_i \geq 0$, and obtains a payoff $u_i(\mathbf{x}, \mathbf{g})$ that depends on the effort profile $\mathbf{x}$ and on the underlying network $\mathbf{g}$, in the following way:

$$u_i(\mathbf{x}, \mathbf{g}) = \underbrace{\alpha x_i}_{\text{Proceeds}} - \underbrace{\frac{1}{2} x_i^2}_{\text{Cost of crime}} - \underbrace{p \cdot f \cdot x_i}_{\text{Cost of being caught}} + \underbrace{\phi \sum_{j=1}^{n} g_{ij} x_i x_j}_{\text{Positive peer effects}} \qquad [12]$$

where $\phi > 0$. This utility has a standard cost/benefit structure (as in Becker, 1968). The proceeds from crime are given by αx_i and are increasing in own effort x_i. The costs of committing

crime are captured by the probability of being caught $0 < p < 1$ times the fine fx_i, which increases with own effort x_i, as the severity of the punishment increases with one's involvement in crime. Individuals have a moral cost of committing crime equal to $(1/2)x_i^2$, which is also increasing in own crime effort x_i. Finally, the new element in this utility function is the last term $\phi \sum_{j=1}^{n} g_{ij}x_ix_j$, which reflects the influence of friends' behavior on own action. Denote $a \equiv \alpha - pf > 0$. Then eqn [12] can be written as

$$u_i(\mathbf{x}, \mathbf{g}) = ax_i - \frac{1}{2}x_i^2 + \phi \sum_{j=1}^{n} g_{ij}x_ix_j \qquad [13]$$

which is exactly the same as eqn [1]. We can thus apply Proposition 1 and show that, if $\phi\mu_1(\mathbf{g}) < 1$, there exists a unique Nash equilibrium in crime effort given by

$$\mathbf{x}^* = a(\mathbf{I} - \phi\mathbf{G})^{-1}\mathbf{1} = a\mathbf{b}(\mathbf{g}, \phi)$$

Policy implications

One interesting aspect of criminal networks is that we can propose alternative policies than those advocated by Becker (1968) and others. In particular, because of the network aspect, we can implement a key player policy, which consists in finding and getting rid of the key player, i.e., the delinquent who, once removed, leads to the highest aggregate delinquency reduction.

Given that delinquent removal has both a direct and an indirect effect on the group outcome, the choice of the key player results from a compromise between both effects. In particular, the key player need not necessarily be the one exerting the highest delinquency effort or, equivalently, the one with the highest Katz–Bonacich centrality measure. The planner's objective is thus to generate the highest possible reduction in aggregate delinquency level by picking the appropriate delinquent. Formally, the planner's problem is the following:

$$\max\left\{x^*(\mathbf{g}) - x^*\left(\mathbf{g}^{[-i]}\right) | i = 1, \ldots, n\right\},$$

where $x^*(\mathbf{g}) = \sum_i x_i^*(\mathbf{g})$ is the total equilibrium level of crime in network $\mathbf{g}$ and $\mathbf{g}^{[-i]}$ is i the network when individual i has been removed. In terms of adjacency matrix, network $\mathbf{g}^{[-i]}$ corresponds to the adjacency matrix $\mathbf{G}^{[-i]}$, which is equal to $\mathbf{G}$ for which the row and column of individual i has been removed. The program above is equivalent to

$$\min\left\{x^*\left(\mathbf{g}^{[-i]}\right) | i = 1, \ldots, n\right\} \qquad [14]$$

From Ballester et al. (2006, 2010), we now define a new network centrality measure of player i, the intercentrality index, which solves this program.

$$c_i(\mathbf{g}, \phi) = \frac{b_i(\mathbf{g}, \phi)^2}{m_{ii}(\mathbf{g}, \phi)} \qquad [15]$$

where $m_{ii}(\mathbf{g}, \phi)$ is the ith element on the diagonal of $\mathbf{M} = (\mathbf{I} - \phi\mathbf{G})$. The Katz–Bonacich centrality of player i counts the number of paths in $\mathbf{g}$ stemming from i; the intercentrality counts the total number of such paths that hit i. It is the sum of i's Katz–Bonacich centrality and i's contribution to every other player's Katz–Bonacich centrality. Holding $b_i(\mathbf{g}, \phi)$ fixed, $c_i(\mathbf{g}, \phi)$ decreases with the proportion of i's Katz–Bonacich centrality due to self-loops, $m_{ii}(\mathbf{g}, \phi)/b_i(\mathbf{g}, \phi)$.

Let us now provide an example showing that the key player is not always the most active delinquent in a network.

Example: Consider the following network $\mathbf{g}$ (Figure 3).

We distinguish three different types of equivalent actors in this network, which are the following:

Type	*Players*
1	1
2	2, 6, 7, and 11
3	3, 4, 5, 8, 9, and 10

For $a = 1$, the following table computes for the network described in Figure 3 the value of delinquency efforts x_i, centrality measures $b_i(\mathbf{g}, \phi)$, and intercentrality measures $c_i(\mathbf{g}, \phi)$ for two different values of ϕ, 0.1 and 0.2. In each column, a variable with a star identifies the highest value.

ϕ	*0.1*			*0.2*		
Player type	x_i	$b_i(\mathbf{g},\phi)$	$c_i(\mathbf{g},\phi)$	x_i	$b_i(\mathbf{g},\phi)$	$c_i(\mathbf{g},\phi)$
1	0.077	1.75	2.92	0.072	8.33	41.67*
2	0.082*	1.88*	3.28*	0.079*	9.17*	40.33
3	0.075	1.72	2.79	0.067	7.78	32.67

First, note that type 2 delinquents always display the highest Bonacich centrality measure. These delinquents have the highest number of direct connections. Besides, they are directly connected to the bridge delinquent 1, which gives them access to a very wide and diversified span of indirect connections. For low values of ϕ, the direct effect on delinquency reduction prevails, and type 2 delinquents are the key players – those with the highest intercentrality measure d_i. When ϕ is higher, however, the most active delinquents are not anymore the key players. Now, indirect effects matter a lot, and eliminating delinquent 1 has the highest joint direct and indirect effects on aggregate delinquency reduction. Note that the network $\mathbf{g}^{[-1]}$ has 20 different links, while $\mathbf{g}^{[-2]}$ has 19 links. In fact, when ϕ is small enough, the key player problem minimizes the number of remaining links in a network, which explains why type 2 delinquents are the key player when $\phi = 0.1$ in this example.

Empirical test

Testing the idea of key players in crime is quite complicated since we need very detailed information about networks and crime outcomes. Fortunately, there is a unique database on

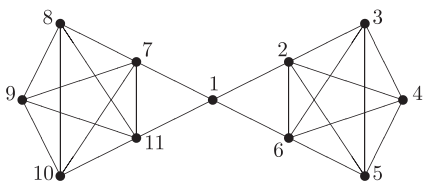

Figure 3 Network with 11 delinquents.

friendship networks from the National Longitudinal Survey of Adolescent Health (AddHealth). (The AddHealth Web site http://www.cpc.unc.edu/projects/addhealth describes survey design and data in details.) It collects data on students in grades 7–12 from a nationally representative sample of roughly 130 private and public schools in years 1994–95. In terms of networks, the most interesting aspect of the AddHealth data is the information on friendships. Indeed, the friendship information is based upon actual friends' nominations. Pupils were asked to identify their best friends from a school roster (up to five males and five females). As a result, $g_{ij} = 1$ if either i or j or both have nominated each other. Otherwise, $g_{ij} = 0$.

Using this dataset, Liu et al. (2011) try to identify key players for juvenile crime in the United States. First, as in the example above, they find that a little bit more than 20% key players are not the most active delinquents. This is because they have a crucial position in the network in terms of betweenness centrality. They also find that, compared to other criminals, 'key' criminals are less likely to be a female, are less religious, belong to families whose parents are less educated, and have the perception of being socially more excluded. They also feel that their parents care less about them, are less likely to come from families where both parents are married, and have more trouble getting along with teachers. An interesting feature is that key players are more intelligent (i.e., higher mathematics scores) than the average criminal and are more likely to have friends who are older (i.e., in higher grades), more religious and whose parents are more educated. Also, even though key players themselves do not have a better self-esteem, they are not more physically developed nor are they more likely to be urbanites than other criminals; their friends are.

The authors also try to determine how efficient a key-player policy as opposed to a random-player policy is, i.e., a policy that removes a criminal at random from the network. Figure 4 displays the results.

To plot this figure, Liu et al. (2011) put together networks of the same size and the average crime reduction for this size of networks under two policies. For example, for all networks of size 4 (horizontal axis), the average crime reduction is 29.94% when the key-player policy (vertical axis blue curve) is implemented while it is 23.86% when a random-target policy is implemented (vertical axis red curve). The difference in crime reduction between these two policies can be large, especially for big networks. This can justify why a key-player policy, although expensive, could be implemented.

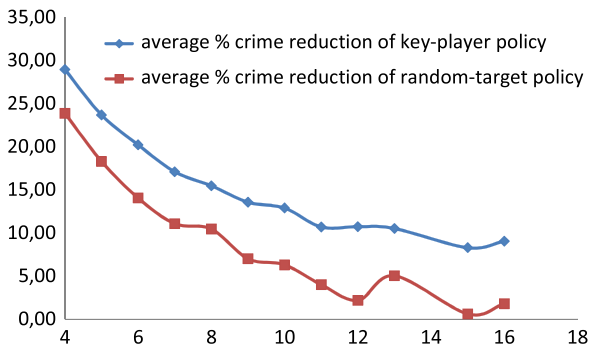

Figure 4 Difference between a key-player policy and a random-target policy.

Concluding Remarks

The study of networks in economics is still at an early stage and much more research needs to be done. In this article, we have provided an overview on this literature by exposing its recent developments. We have seen that the models of network formation are often plagued by too many equilibria and that the full characterization of all possible equilibria is a very difficult task. On the contrary, the study of games on networks, where the network is fixed and the impact of the network structure on individuals' outcomes is analyzed, is much more developed since the existence, uniqueness, and characterization of equilibrium has been established. We have also seen how these models can easily be used to analyze issues related to labor and crime networks (and also education, R&D, cities, etc.; see Jackson and Zenou, forthcoming).

We believe that, in the future, more effort should be devoted to the formation of networks in a dynamic framework using some microfoundations based on economic choices. Indeed, following the physics literature, some economic models have been used to model dynamic network formation but links are usually formed in a random or probabilistic way. (In these models, usually, one agent at a time is selected to form or create links. This greatly simplifies the analysis and avoids the coordination problems highlighted above where all agents could delete or create links simultaneously.) Some recent papers have, however, tried to provide more microfounded models where links between agents are formed based on a careful cost and benefit analysis so that agents maximize their utility (see, in particular, Christakis et al., 2010; Mele, 2011; König et al., forthcoming).

We also believe that more work should be done to estimate empirically economic networks. We have shown how labor and crime networks could be estimated and what kind of policy implications they imply. Recent papers (Bramoullé et al., 2009; Calvó-Armnegol et al., 2009) have shown how econometric issues such as the reflection problem (Manski, 1993) and correlated effects can be solved using a network approach. This is a growing field (see, in particular, Patacchini and Zenou, 2012b; Banerjee et al., 2012) and we hope that many more tests on different economic aspects where networks are prevalent such as crime, microfinance, religion, smoking, teenage pregnancy, etc. will be performed in the future.

See also: Core–Periphery Patterns in Networks; Deterrence Theory: Crime; Economics and Crime; Game Theory; Models of Influence in Networks; Models of Power in Exchange Networks; Network Centrality, Measures of; Prevention of Crime and Delinquency.

Bibliography

Acemoglu, D., Dahleh, M.A., Lobel, I., Ozdaglar, A., 2011. Bayesian learning in social networks. Review of Economic Studies 78, 1201–1236.

Albert, R., Barábasi, A.-L., 2002. Statistical mechanics of complex networks. Reviews of Modern Physics 74, 47–97.
Bala, V., Goyal, S., 2000. A noncooperative model of network formation. Econometrica 68, 1181–1230.
Ballester, C., Calvó-Armengol, A., Zenou, Y., 2006. Who's who in networks: wanted the key player. Econometrica 74, 1403–1417.
Ballester, C., Calvó-Armengol, A., Zenou, Y., 2010. Delinquent networks. Journal of the European Economic Association 8, 34–61.
Banerjee, A., Chandrasekhar, A.G., Duflo, E., Jackson, M.O., 2012. The diffusion of microfinance. NBER Working Paper No. 17743.
Bayer, P., Ross, S.L., Topa, G., 2008. Place of work and place of residence: informal hiring networks and labor market outcomes. Journal of Political Economy 116, 1150–1196.
Becker, G., 1968. Crime and punishment: an economic approach. Journal of Political Economy 76, 169–217.
Bloch, F., Dutta, B., 2009. Communication networks with endogenous link strength. Games and Economic Behavior 66, 39–56.
Bloch, F., Jackson, M.O., 2006. Definitions of equilibrium in network formation games. International Journal of Game Theory 34, 305–318.
Bollobás, B., 2001. Random Graphs, second ed. Academic Press, New York.
Bonacich, P., 1987. Power and centrality: a family of measures. American Journal of Sociology 92, 1170–1182.
Bramoullé, Y., Djebbari, H., Fortin, B., 2009. Identification of peer effects through social networks. Journal of Econometrics 150, 41–55.
Bramoullé, Y., Kranton, R., 2007. Public goods in networks. Journal of Economic Theory 135, 478–494.
Cabrales, A., Calvó-Armengol, A., Zenou, Y., 2011. Social interactions and spillovers. Games and Economic Behavior 72, 339–360.
Calvó-Armengol, A., 2004. Job contact networks. Journal of Economic Theory 115, 191–206.
Calvó-Armengol, A., de Marti, J., 2009. Information gathering in organizations: equilibrium, welfare and optimal network structure. Journal of the European Economic Association 7, 116–161.
Calvó-Armengol, A., Jackson, M.O., 2004. The effects of social networks on employment and inequality. American Economic Review 94, 426–454.
Calvó-Armengol, A., Patacchini, E., Zenou, Y., 2009. Peer effects and social networks in education. Review of Economic Studies 76, 1239–1267.
Calvó-Armengol, A., Zenou, Y., 2004. Social networks and crime decisions: the role of social structure in facilitating delinquent behavior. International Economic Review 45, 935–954.
Christakis, N., Fowler, J., Imbens, G.W., Kalyanaraman, K., 2010. An empirical model for strategic network formation. NBER Working Paper No. 16039.
de Martí, J., Zenou, Y., 2011. Social networks. In: Jarvie, I., Zamora- Bonilla, J. (Eds.), Handbook of Philosophy of Social Science. SAGE Publications, London, pp. 339–361.
Galeotti, A., Goyal, S., 2010. The law of the few. American Economic Review 100, 1468–1492.
Galeotti, A., Goyal, S., Jackson, M.O., Vega-Redondo, F., Yariv, L., 2009. Network games. Review of Economic Studies 77, 218–244.
Glaeser, E.L., Sacerdote, B.I., Scheinkman, J.A., 1996. Crime and social interactions. Quarterly Journal of Economics 111, 508–548.
Goyal, S., 2007. Connections: An Introduction to the Economics of Networks. Princeton University Press, Princeton.
Granovetter, M., 1995. Getting a Job: A Study of Contacts and Careers, second ed. University of Chicago Press, Chicago.
Hagenbach, J., Koessler, F., 2011. Strategic communication networks. Review of Economic Studies 77, 1072–1099.
Holzer, H.J., 1987. Informal job search and black youth unemployment. American Economic Review 77, 446–452.
Holzer, H.J., 1988. Search method used by unemployed youth. Journal of Labor Economics 6, 1–20.
Ioannides, Y.M., Datcher-Loury, L., 2004. Job information networks, neighborhood effects and inequality. Journal of Economic Literature 424, 1056–1093.
Jackson, M.O., 2008. Social and Economic Networks. Princeton University Press, Princeton.
Jackson, M.O., Watts, A., 2002. The evolution of social and economic networks. Journal of Economic Theory 106, 265–295.
Jackson, M.O., Wolinsky, A., 1996. A strategic model of social and economic networks. Journal of Economic Theory 71, 44–74.
Jackson, M.O., Zenou, Y. Games on networks. In: Young, P., Zamir, S., (Eds.), Handbook of Game Theory, vol. 4. Elsevier Publisher, Amsterdam, forthcoming.
Jackson, M.O., Zenou, Y., 2013. Economic Analyses of Social Networks, The International Library of Critical Writings in Economics. Edward Elgar Publishing, London.
Katz, L., 1953. A new status index derived from sociometric analysis. Psychometrika 18, 39–43.
König, M.D., Tessone, C., Zenou, Y. Nestedness in Networks: A Theoretical Model and Some Applications. Theoretical Economics, forthcoming.
Liu, X., Patacchini, E., Zenou, Y., Lee, L.-F., 2011. Criminal Networks: Who Is the Key Player? CEPR Discussion Paper No. 81–85.
Lopez-Pintado, D., 2008. Diffusion in complex social networks. Games and Economic Behavior 62, 573–590.
Manski, C.F., 1993. Identification of endogenous effects: the reflection problem. Review of Economic Studies 60, 531–542.
Mele, A., 2011. A Structural Model of Segregation in Social Networks. Unpublished manuscript. Johns Hopkins University.
Myerson, R., 1977. Graphs and cooperation in games. Mathematics of Operations Research 2, 225–229.
Myerson, R., 1991. Game Theory: Analysis of Conflict. Harvard University Press, Cambridge, MA.
Newman, M.E.J., 2010. Networks: An Introduction. Oxford University Press, Oxford.
Patacchini, E., Zenou, Y., 2012a. Ethnic networks and employment outcomes. Regional Science and Urban Economics 42, 938–949.
Patacchini, E., Zenou, Y., 2012b. Juvenile delinquency and conformism. Journal of Law, Economics, and Organization 28, 1–31.
Sarnecki, J., 2001. Delinquent Networks: Youth Co-offending in Stockholm. Cambridge University Press, Cambridge.
Sutherland, E.H., 1947. Principles of Criminology, fourth ed. J.B. Lippincott, Chicago.
Topa, G., 2001. Social interactions, local spillovers and unemployment. Review of Economic Studies 68, 261–296.
Warr, M., 2002. Companions in Crime: The Social Aspects of Criminal Conduct. Cambridge University Press, Cambridge.
Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge.

Neumann, John von (1903–57)

Anatol Rapoport, University of Toronto, Toronto, ON, Canada

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 10514–10518, © 2001, Elsevier Ltd.

Abstract

Immense breadth and profound depth distinguish John von Neumann's contributions to mathematics. They range from the most abstract – for example, foundations of mathematics (axiomatization), through mathematical physics (quantum mechanics) to concrete applications in artificial intelligence and military technology. Together with economist Oscar Morgenstern, von Neumann laid the foundations of the theory of games, the powerful mathematical tool of contemporary economics. In particular, von Neumann was the first to prove the minimax theorem, on which the concept of rational strategy, fundamental in the theory of games, rests. Von Neumann may have stimulated the reductionist trend in the social sciences by his attempts to explain a wide range of psychological and social phenomena by considerations of self-interest; a psychology ascribed to *homo economicus*. Although von Neumann, and especially his followers, extended game theory to cooperative games, on which concepts associated with collective rationality (e.g., Pareto optimality) play a part, von Neumann did not appreciate the break-out of the theory of noncooperative games from the minimax paradigm, following the discovery of Prisoner's Dilemma and other noncooperative games in which the cardinal difference between individual and collective rationality is demonstrated.

John von Neumann was born Janos Neumann on 28 December 1903. He was the eldest son of Miksa (Max) Neumann and Margit (Margaret) Kann. The family prospered in the last decades before World War I. Max Neumann was a lawyer who was well known in banking circles, and he was later a banker. He was subsequently ennobled, hence the 'von' in John's name.

Max Neumann instilled his ardent love of high culture into his children. In the atmosphere in which John spent his childhood, his phenomenal precociousness in mathematics was soon discovered. His gymnasium teacher pointed out to John's father that there was no point in John's studying mathematics in school. Mature mathematicians were engaged as private tutors. His first paper, written jointly with one of them, appeared in 1922. John was recognized as a professional mathematician before he was 18.

After receiving his *matura*, von Neumann enrolled as a mathematics student in the University of Budapest. However, he did not attend classes and he only appeared at the end of each semester to take examinations. He was simultaneously enrolled first at the University of Berlin, Germany then in the *Eidgenossische Technische Hochschule* in Zurich, Switzerland. He received his doctorate in mathematics from the University of Budapest and the degree of *Diplomingenieur* in chemistry from the *Eidgenossische Technische Hochschule* almost simultaneously. The study of chemistry was probably motivated by the increasing importance of its applications in military technology. Toward the end of his life, von Neumann's most brilliant contributions were closely related to the development of weaponry.

Von Neumann's professional career falls roughly into three periods. He spent 5 years in Weimar, Germany (the period between the monarchy and the Third Reich) as *Privatdozent* in Berlin and Hamburg. Recognizing the impending Nazification of Germany, he accepted a visiting professorship at Princeton University and emigrated. Subsequently, he joined the newly founded Institute for Advanced Studies. The depression decade roughly marks the second period, when he made his most brilliant contributions to mathematics and to mathematical physics. In the last period, emphasis shifted to applied mathematics, particularly to aerodynamics and the physics of explosions and implosions, the development of nuclear weapons, and of electronic computers. His academic affiliation remained with the Institute of Advanced Studies. However, the locale of his most intense preoccupation was Washington, where his contributions to US military might were fully appreciated. On one of his last public appearances President Eisenhower presented him with the Medal of Freedom.

In Norman Macrae's biography of von Neumann (1992), the last chapter is aptly entitled 'With Astounding Influence.' The final scene is at the Walter Reed Hospital in Washington.

The great and important now flocked to his bedside. It was known that Johnny's illness might lead to his talking in his sleep, and Colonel Ford was told to place soldiers to ensure that he did not shout out military secrets. Von Neumann died on 8 February 1957.

An obituary by S. Ulam, a close friend and colleague, in the *Bulletin of the American Mathematical Society* lists titles of 125 publications and of 17 abstracts of papers presented to the American Mathematical Society (1958). The list ranges over all major fields of mathematics and some of physics, for example quantum theory and statistical mechanics. Von Neumann's doctoral dissertation was on axiomatization of mathematics based on set theory, a subject he taught at age 11- to 12-year-old future Nobel Laureate Eugene Wigner. This branch of mathematics seems to provide access to the way von Neumann viewed the functioning of the human mind at its best.

A prevalent view of mathematics (among non-mathematicians) is that mathematical concepts are essentially abstractions of experiences with the physical world. Von Neumann has on occasions expressed that view; and this seems strange, since his earliest dedication, manifested in his doctoral dissertation, was to the foundations (or axiomatization) of mathematics, which represents, perhaps more than any other branch, a complete break with concrete, sense-induced experiences. All universally recognized mathematical knowledge is gained from deduction alone, which is from chains of reasoning about what one assertion implies about another. Empirics plays no role whatsoever in this process.

David Hilbert, arguably the most outstanding mathematician of the nineteenth century, 'bequeathed' to mathematicians 23 problems to be solved in the course of the twentieth century. One of them was the 'axiomatization' of all mathematics, that is, a follow-up of Euclid's feat in geometry. This task appeared challenging, because of difficulties that arose in the highly original work of Georg Cantor on the arithmetic of so-called 'transfinite' numbers. It seemed as if paradoxes or, perhaps, even gross-contradictions appeared in that theory. The only way to clear them up was to axiomatize the entire realm of mathematics on the foundation of its most basic concept – that of a set, thus reducing all mathematical propositions to statements about relationships among sets.

Hilbert, who was markedly enthusiastic about Cantor's work, was convinced that such axiomatization could be accomplished. Other distinguished mathematicians, notably Hermann Weyl and L.E.J. Brouwer, thought it could not and recommended a revision of mathematics that would 'delegitimize' some of its most important branches (mostly dealing with continua). Von Neumann sided with Hilbert (who, in turn, recognized young von Neumann as a rising star) and undertook the task, which Bertrand Russell and Alfred North Whitehead had undertaken at the turn of the century and failed. Von Neumann also failed. In 1931 the problem was solved, if an impossibility proof can be regarded as a solution, by Kurt Gödel, who showed that any mathematical system of complexity comparable to that of elementary arithmetic or greater must contain propositions that cannot be shown to be either true or false. Yet it was von Neumann's unsuccessful effort to solve the problem bequeathed by Hilbert that in a way helped him lay the theoretical foundations of computer technology. He adopted Gödel's cataloguing of logical statements as numbers by cataloging computer instructions as numbers.

Von Neumann's preoccupation with the foundations of mathematics suggests that his philosophy of scientific cognition (it seems he recognized no other kind) was based on 'reductionism.' The merging of conceptual repertoires and methodologies of two disciplines, whereby one with fewer concepts absorbs the one with more, exemplifies reductionism. The best known example is the reduction of chemistry to physics. For example, the fundamental conceptual building blocks of classical chemistry were originally the elements. These differed qualitatively. Their combinations into compounds and, conversely, the break up of compounds into elements were the principal events constituting chemical processes. With the discovery of the periodic table and of the internal structure of atoms, the properties of elements were described in quantitative terms (e.g., numbers of electrons, protons, etc., characterizing the atoms of elements), and the basic events were explainable by physical laws. Thereby, much of chemistry was 'reduced' to physics, a more rigorous science. Eventually physiology was reduced to biochemistry, parts of psychology to physiology. In economics, reduction of macroeconomics (behavior of large economic systems) to microeconomics, that is, the behavior of *homo economicus* (a postulated 'atom' of economic activity) is in the spirit of reductionism.

Von Neumann's work in this spirit was particularly fruitful. For example, the conception of a human brain as a computer is certainly a reductionist step, since computer science is much more rigorous, that is, more mathematical than traditional cognitive psychology or the physiology of the nervous system. In general, the 'mathematization' of any discipline is in the spirit of reductionism. So it was with von Neumann's most important contribution to the social sciences, the discipline of which he undisputedly laid the foundation, namely, the theory of games.

Game theory can be clearly seen as a branch of decision theory in which more than one rational decision-maker (or 'actor') are involved. The prototypes of games examined in the theory comprise so-called 'games of strategy' such as chess, go, bridge, poker, etc., that is, contests in which facility in formal reasoning is decisive. Of special interest are 'two-person constant sum games,' in which the sum of the payoffs to the respective players is constant in all outcomes: the more one wins, the less the other wins or the more he loses. Hence, the interests of the two players are diametrically opposed.

It is in the spirit of reductionism that von Neumann built up the theory of economic behavior, on a foundation provided by the theory of the two-person constant sum game. In his treatment of so-called 'cooperative' n-person games ($n > 2$), some of the players may form a coalition and coordinate their strategies so as maximize their joint payoff. It is then in the interest of the remaining players to form their own coalition. If the n-person game is constant sum, then the game between the two coalitions is essentially a two-person constant sum game. If the sum is not constant, a 'dummy' player can be added whose winnings or losses in each outcome compensate for the differences, making the game constant sum. In this way, a theory of the 'cooperative' n-person game reduces to a game of completely opposed interests.

Von Neumann's great achievement was the proof of the so-called 'minimax theorem.' It asserts that every two-person constant sum game, in which a finite number of strategies are available to each player has a 'solution,' expressible as the outcome resulting from an independent choice by each player of a particular strategy, pure or mixed. A pure strategy is essentially an adoption of a plan to make a particular 'move' among those permitted by the rules in each of the possible situations that may occur in the course of the game. A mixed strategy is a probabilistic mixture of pure strategies. This concept was already introduced by E. Borel in 1921, but illustrated only in special cases. Von Neumann proved the general existence theorem in 1928. *Theory of Games and Economic Behavior*, coauthored with Oscar Morgenstern (1944) very soon attracted the attention of the American military establishment as a possible foundation of a mathematically precise rational conduct of conflict.

The minimax corresponds to what military strategists call the 'worst-case scenario,' the assumption that the enemy doesn't necessarily act so as to achieve the best possible outcome for himself but rather to impose the worst possible outcome on you. In situations analogous to the two-person constant sum game the two goals are identical, since confrontation of enemies is most naturally modeled by a constant sum game. The situation in so-called 'neoclassical' economics is similar. At the foundation of its reduction is theoretical framework is 'methodological individualism,' which P.H. Wicksteed (1933), perhaps to avoid the pejorative connotations of 'egoism,' called 'non-tuism' – the overriding goal of maximizing one's own payoff. As a result of impersonal

market forces (Adam Smith's 'invisible hand') an economic system so modeled is shown with mathematical rigor to tend toward an 'equilibrium,' assuming this to be the best attainable condition for everyone concerned. In this way, von Neumann's minimax theorem appeared to provide not only an idealized model of 'rationally conducted conflict,' but also a justification of a Hobbsian 'competition of every one with every one else.'

To the extent that some references are made in military doctrines to human behavior, these can be arguably included in the subject matter of the social sciences. Economics is unequivocally regarded as a branch of social science, indeed the most theoretically rigorous one. Thus, von Neumann's and Morgenstern's treatise can be justifiably celebrated as a most impressive contribution to social science. Paradoxically, an even more impressive contribution can be discerned in the pursuit of answers to questions that were not raised in the treatise but were inspired by it, namely, in the course of analyzing non-constant sum games.

As has been said, the US military establishments quickly recognized the potential of game theory in developing a 'science of rationally conducted conflict.' Generous support of further research was not long in forthcoming. Workers in the field were invited to the think tank established by the Rand Corporation. Attention was focused on experimental verification of theoretical predictions generated by game theory. Note that laboratory experimentation with a real economic system is practically impossible, while such experimentation on people playing specially designed games is easy, cheap, and, most importantly, reproducible.

The results of the early experiments with nonconstant sum games were disappointing. They in no way confirmed systematically the predictions of game theory based on the minimax principle. Neither did they systematically refute them in a way that would suggest a generalization of the theory to explain the discrepancy. However, a result was observed demonstrating a 'serendipity' effect – finding something more valuable than what was originally sought. Namely, a game was discovered that was soon called Prisoner's Dilemma, now a byword in the vocabularies of economics, social psychology, sociology, political science, management science, peace research, peace education, and philosophy. The game was discovered by M.M. Flood (1952) and M. Dresher and christened by A.W. Tucker, followers of von Neumann's leads. The name of the game stems from the following anecdote.

Two men are arrested on suspicion of burglary. There is not enough evidence to convict them unless at least one confesses. They can, however, be convicted of a lesser crime – possession of stolen goods, which carries a lighter sentence. If only one confesses, he is not prosecuted, while the other is given the most severe sentence. If both confess their sentence is reduced. If neither confesses, both get the lightest sentence (for the lesser crime). It turns out that it is in the interest of each to confess regardless of whether the other confesses or not. However, if both confess, both are given longer prison terms than if neither confesses.

The paradox brings out the distinction between individual and collective rationality. There are innumerable situations where this distinction is dramatically demonstrated. If a fire breaks out in a crowded theater, it is in every one's interest to get to an exit as quickly as possible. If all act 'rationally,' all may perish. It is in the interest of each fishing fleet to maximize its catch. If each acts 'rationally' there may soon be no fish to catch. It is in each of two hostile powers' interest to be stronger than the other is. An escalating arms race may lead to a disastrous war, which both wanted to avoid.

There is no indication that von Neumann was aware of the 'paradox of rationality' revealed by Prisoner's Dilemma, although the basic idea was expressed over a century before the publication of *Theory of Games and Economics Behavior* (1944) by von Neumann and Morgenstern. Having determined the equilibrium point of a market controlled by a duopoly, A.A. Cournot wrote in 1838:

> ... si les deux producteurs *s'etaient entendue* [emphasis in the original], la valeur de ... [price calculated on the basis of the agreement] aurait assigné à chacun un revenu plus grand que celui qu'ils obtiendront avec la valeur de ... [price determined by the duopolists independently]. Cournot, 1838

According to Merril Flood, who called von Neumann's attention to the paradox, the latter admitted that the Prisoner's Dilemma was a challenge to the minimax principle but did not take seriously the results of the Flood-Dresher experiment, which did not support the minimax (equilibrium) solution.

Two years after von Neumann's death, systematic experimentation with games similar to the Prisoner's Dilemma began in the USA and spread worldwide. It gave rise to a current contrary to that originating in the so-called 'rationality paradigm' which sought to reduce all aspects of human behavior to manifestations of self-interest and all 'rational' behavior to cost-benefit analysis. Von Neumann evidently thought within the framework of that paradigm. In his opinion, the fact that 40 000–50 000 annual traffic fatalities are tolerated in the USA indicates that this is an acceptable price to pay for the privilege of driving private automobiles. In the same spirit, he advocated acceleration of the arms race, on the grounds that if the Soviet Union got ahead, it would be more likely to launch a preventive nuclear strike against the USA. Apparently the opposite hypothesis (that the party behind might be more likely to launch a preventive war) did not occur to him.

How, then, can the suggestion be justified that the theory of games launched a movement in the social sciences transcending the reductionist model of the human condition? Recall the seemingly unassailable finality of the Newtonian paradigm resting on the metaphysics of a Euclidian three-dimensional space, independent flow of time, and an all-pervasive ether, defining an absolute system of coordinates. This paradigm was overthrown in the course of a few years by the relativity and quantum revolutions. The point is that the success of non-Newtonian physics was due to the clear axiomatization (except for implicit metaphysical postulates, such as independence of space and time) of the *Newtonian* paradigm. It was clear just what 'axioms' had to be dropped, generalized, or supplanted by others to account for phenomena inexplicable within the old paradigm. For example, the (implicit) postulate of an absolute system of coordinates was supplanted by the (explicit) postulate of the constancy of

the speed of light with respect to all systems of coordinates. In quantum mechanics, the independence of physical events from acts of observation (implicit in the Newtonian system) was supplanted explicitly by the Uncertainty Principle, incompatible with that assumption. Thus, the 'revolution' triggered by the Prisoner's Dilemma was fueled by the clear axiomatization of the rationality paradigm as manifested in von Neumann's and Morgenstern's theory of games. In particular, Pareto optimality (maximization of common rather than individual gain) supplanted the 'non-tuism' of neoclassical mathematical economics.

Also noteworthy is the 'reflexive' model of man developed by Vladimir Lefebvre (1998), in which the (implicit) assumption of 'non-tuism,' namely that the thought processes of interacting individuals are independent, is supplanted by a 'mirror image' assumption. An actor is assumed to be able to mentally reconstruct another actor's thoughts and feelings and, regarding him as his own 'mirror image,' acts accordingly. In the light of this model, the cooperative solution of the Prisoner's Dilemma and similar games is rigorously deduced.

In this way, Max Weber's conception of 'empathic (sic!) understanding' as an epistemological principle in the social sciences was revived. The dogma of 'value-free' social science was revealed as an analog of the Newtonian axiom of independence of observations from the act of observation. It appears that social science, no matter how rigorous, cannot be value free, because social self-observation plays a crucial part in its epistemology and inevitably generates value-oriented conclusions.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Axiomatic Theories; Charity and Philanthropy, Economics of; Crime, Geography of; Decision and Choice: Economic Psychology; Economics, History of; Game Theory: Noncooperative Games; Game Theory; Individualism versus Collectivism: Philosophical Aspects; Individuals in Society: History of the Concept; Interest, Psychology of; Prisoner's Dilemma, One Shot and Iterated; Rational Choice Explanation: Philosophical Aspects; Rational Choice and Organization Theory.

Bibliography

Borel, E., 1921. La théorie des jeux et les équations intégrales à niveau symmétrique. Comptes rendus Academic des Sciences 173, 1304–1308.
Cournot, A.A., 1838. Recherches sur les principes mathématiques de la richesse. Hachette, Paris.
Flood, M.M., 1952. Some Experimental Games. Research Memorandum RM-789. Rand Corporation, Santa Monica, CA.
Heims, S.J., 1982. John von Neumann and Norbert Wiener. MIT, Cambridge, MA.
Lefebvre, V.A., 1998. Sketch of reflexive game theory. Proceedings of Workshop on Multi-reflexive Models of Agent Behavior ARL-SR-63. Army Research Laboratory.
Luce, R.D., Raiffa, H., 1957. Games and Decisions. Wiley, New York.
Macrae, N., 1992. John von Neumann. Pantheon Books, New York.
Poundstone, W., 1992. Prisoner's Dilemma. Doubleday, New York.
Rapoport, A., Chammah, A.M., 1966. Prisoner's Dilemma. University of Michigan Press, Ann Arbor, MI.
Ulam, S., 1958. John von Neumann. Bulletin of the American Mathematical Society Part 2, 64 (3), 1–49.
Von Neumann, J., 1925. Eine Mathematisierung der Mengenlehre. Acta Universitatas Szegedsienses 1, 219–240.
Von Neumann, J., 1928. Zur Theorie der Gesellschaftsspiele. Mathematische Annalen 100, 295–320.
Von Neumann, J., 1932a. Mathematische Grundlagen der Quantenmechanik. Springer, Berlin.
Von Neumann, J., 1932b. Zur Operatorenmethode in der klassischen Mechanik. Annals of Mathematics 33, 587–562.
Von Neumann, J., 1966. (posthumous) Theory of Self-reproducing Automata. University of Illinois Press, Urbana, IL.
Von Neumann, J., Fakete, M., 1922. Über die Lage der Nullstellen gewisser Minimalpolynome. Berichte des Deutschen Mathematischen Vereins 31, 125–138.
Von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ.
Von Neumann, J., Burks, A.W., Goldstein, H.H., 1946. Report prepared for US Army Ord. Dept. under Contract W-36-034-ORD-7481.
Wicksteed, P.H., 1933. The Common Sense of Political Economy. Routledge, London.

Neural Basis of Language Development

Heidi M Feldman and Vanessa N Kovachy, Stanford University School of Medicine, Palo Alto, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The neural basis of language development is understood through studies of children with brain injuries and functional imaging of healthy children. Children with injuries to cortical areas that serve language in adults demonstrate that many different brain networks can support language development. However, children with injuries to cortex, white matter, or cerebellum may be deficient in multiple subdomains of language, indicative of the limits of plasticity and recovery. Healthy children show structural and functional lateralization for language processing in infancy. Increasing left lateralization occurs over development and is affected by age, language proficiency, and the nature of the task.

Introduction

The neural basis of language development is not fully understood. Evolving concepts about the neurobiology of adult language has influenced the experimental approaches and interpretations of findings about the neural basis of child language. Early studies in the child literature, as in the adult literature, inferred the neural basis of language from the neuropsychology of individuals with brain injuries. Recent research draws on structural and functional brain imaging in individuals with neural injuries and in children developing typically. Unanswered questions about the neural basis of language development await the further integration of multiple behavioral and imaging methodologies.

Neurobiological Basis of Mature Adult Language Functioning

Cortical Regions

The 'classical' theory of the neural basis of language emerged in the late nineteenth century based on clinical observations of adults with acquired brain injuries and aphasia. In that theory, the critical language substrate was localized to two distinct cortical regions in the left cerebral hemisphere and a white matter (a component of the nervous system comprised of the axons of cells, covered by myelin; insulates the axon and increases the rate of signal transduction; appears white, giving this brain compartment its name) tract running between them: (1) Wernicke's area (a region of the cortex of the brain that is located on the superior temporal gyrus on the left cerebrum near the junction of the temporal and parietal lobes, described as the posterior part of Brodmann area 22, and in conjunction with Broca's area, is implicated in language and speech processing), the posterior superior temporal area and supramarginal gyrus of the parietal lobe adjacent to the auditory cortex, was thought to be the neural substrate for speech perception and comprehension; (2) Broca's area (a region of the cortex of the brain that is located in the inferior frontal lobe, on the pars triangularis and pars opercularis, or Brodmann areas 45 and 44, and in conjunction with Wernicke's area, is implicated in language and speech production), the inferior frontal operculum, was thought to be the neural basis of speech production; and (3) the arcuate fasciculus, a white matter fiber tract coursing from posterior superior temporal to frontal lobe, was thought to be the main link for information flow within the system.

New experimental methods, including functional brain imaging that assesses the neural substrate of language in alive, awake, and processing individuals, has resulted in a 'contemporary' theory, distinct from its predecessor in several ways. First, classical brain areas are involved in various language functions: Wernicke's area has been implicated in phonological processing, interpretation of written words, and lexical (the vocabulary of a language) and semantic (the rules regulating the creation of meaning in a human language) aspects of language comprehension; Broca's area appears to be involved in syntactic processing, verbal short-term memory, and phonological processing. Second, adult language functioning uses a complex, interconnected, and overlapping set of networks, working in parallel in both hemispheres. A ventral stream that involves structures in the superior and middle portions of the temporal lobe appears to be involved in processing speech for recognition and comprehension (Hickok and Poeppel, 2007). A dorsal stream that involves structures in the posterior dorsalmost aspect of the temporal lobe and parietal operculum and posterior frontal lobe appears to be involved primarily in translating speech signals into articulatory representations for speech production (Hickok and Poeppel, 2007). However, these areas are also activated in individuals who are deaf and use sign language for communication, indicating that they are used in higher language processing as well as speech production (Petitto et al., 2000). Other networks come on line flexibly and dynamically, based on the modality of the inputs, top-down predictions from prior knowledge, linguistic and nonlinguistic context, task demands, such as comprehension versus judgments, and task difficulty (Price, 2010). Specific task requirements, such as suppressing unintended or irrelevant speech, may activate regions of the frontal lobe, anterior cingulate, and bilateral basal ganglia (Price, 2010).

The contemporary theory also shows that many but not all language functions show left lateralization. Prelinguistic auditory processing begins with bilateral activation and becomes left lateralized as speech processing begins (Price, 2010). Sentence-level comprehension results in activation of bilateral superior temporal sulci, whereas semantic retrieval

activates left angular gyrus and pars orbitalis. Left lateralization characterizes word retrieval, articulatory planning, motor cortex for initiation, and execution of speech (Price, 2010).

White Matter Pathways

White matter pathways link the various language networks. White matter pathways can now be dissected and analyzed using modern structural imaging techniques, such as diffusion tensor imaging (DTI, a magnetic resonance imaging method that uses the rate and direction of water diffusion to characterize the volume and properties of different compartments of the brain, and also the method of choice for visualizing and characterizing the white matter of the brain), a technique that differentiates white matter from cerebral spinal fluid and gray matter on the basis of differential water diffusion patterns in the three compartments (Basser et al., 1994). Asymmetries have been found in the volume of white matter pathways connecting the superior and middle temporal gyri to frontal lobes, suggesting that white matter shows left lateralization. Adult DTI data find that signals carried along the arcuate fasciculus and neighboring superior longitudinal fasiculus may be essential for semantic verbalization, syntactic comprehension, and skilled reading (Friederici, 2009). Another pathway, ventral to the arcuate, connects the superior temporal lobe to the inferior frontal lobe and is hypothesized to be important for language comprehension (Saur et al., 2008). The integrity of this ventral pathway may explain variations in language abilities and the degree of recovery after damage to the arcuate fasciculus.

Cerebellum

The cerebellum is also involved in language processing and organization. Two primary reciprocal pathways link the cerebellum to the frontal lobe via crossing fibers, such that the right cerebellum connects to left frontal lobe (Murdoch, 2010). These pathways coordinate linguistic ability with frontal lobe executive functions. Increased blood flow to the cerebellum, particularly the right cerebellum, occurs during complex language tasks, such as word generation and novel naming tasks, but not during simple language tasks (Murdoch, 2010). The left hemisphere of the cerebellum, though less active than the right has been shown to be involved in high-level linguistic skills, including sentence formulation, lexical-semantic operations, and organization and monitoring of responses in collaboration with the frontal lobes (Murdoch, 2010).

Findings from Neuropsychological Studies of Children's Language

Behavioral Studies of Children with Focal Injuries

Research on the neural basis of language development has focused on to what extent neural systems for language are organized prior to language learning. Before the modern era of neuroimaging, studies to address this question assessed language and cognitive function in children with early brain injuries. Children may suffer from strokes or other focal injury to the brain early in life. Children with vascular or other destructive lesions sometimes undergo a surgical procedure called hemispherectomy (the surgical procedure where one cerebral hemisphere is removed completely) that leaves them with a single functioning cortical hemisphere. These injuries and conditions can be conceptualized as naturalistic experiments that allow researchers to determine how the early injury affects the rate and pattern of language development and ultimately the status of mature language functioning (Feldman, 2005).

In early research on children with neural injuries, results were interpreted in terms of the 'nature–nurture' controversy. The 'nature' position argued that the brain is specialized for language based on genetic mechanisms. According to this view, injury to classical left hemisphere brain regions prior to language learning should result in severe delays and disabilities in language. The 'nurture' position argued that brain specialization results from environmental input and the experiences using language. In its extreme form, all cortical regions should be equally adequate to subserve language function, a position known as 'equipotentiality' (Lenneberg, 1967). According to this view, injury to classical left hemisphere brain regions prior to language learning should not seriously delay or disturb language learning because other regions and pathways could assume language functions. An intermediate position has been called 'developmental specialization.' It posits that particular brain regions have inherent anatomical, physiological, or computation predispositions for language learning and that if these regions are injured other areas may be able to serve language function. According to this position, predictions about the outcomes after focal injury are uncertain, depending upon multiple factors, including the degree and location of injury, the health of the remaining areas of cortex, and the internal structure of regions that assume language processing. The need to rely on pathways not usually associated with language may result in delays in development or deficits in mature language functioning (Feldman, 2005).

Children with focal injury to the left hemisphere at birth do not develop aphasia. Such children begin to speak in the toddler/preschool era (Feldman, 2005). By age 4 years, they master the complex morphosyntactic structures of language, even as complex as Hebrew (Levy et al., 1994). They outperform adults with similar injuries (Dick et al., 1999). These results demonstrate that the brain has impressive plasticity for language learning, and suggest that various brain regions can serve language function.

Children with focal injuries to either the left or the right hemisphere experience mild-to-moderate developmental delays in many language and cognitive functions compared to children without injury. Affected domains include prelinguistic skills, including the onset of babbling and communicative gestures, and linguistic skills, including the onset of vocabulary development, the use of word combinations, and the interpretation of syntactically complex sentences (Feldman, 2005). Their rate of developmental progress in vocabulary and sentence length in the preschool era once they begin to speak is comparable to children developing typically. However, they do not necessarily catch up to their peers. At school age, compared to children developing typically, they show conversational competence, and near average

scores on receptive vocabulary and syntax (the rules regulating sentence structure and organization in a particular language) (Demir et al., 2010). However, they show deficits in several areas: basic skills underlying language processing, such as speed of recognition and discrimination of sounds (MacWhinney et al., 2000); use of word order and animacy cues to determine the agent of the action in sentence comprehension (Feldman et al., 2002); interpretation of syntactically complex sentences (Dick et al., 1999); word learning, nonword repetition, and immediate serial recall (Gupta et al., 2003); and creation of narrative discourse (Demir et al., 2010). The maturity of language in children with focal injuries to either hemisphere is associated with features beyond lesion characteristics, including the presence of seizures and the quality of the parental input (Rowe et al., 2009). These findings indicate that plasticity after early brain injury is limited and that the integrity of networks in both hemispheres is important to child language. If any of the multiple pathways involved in language processing are damaged, the result is usually relative difficulties in language abilities, particularly with higher level of language processing skills.

It is still disputed whether children with left hemisphere focal injury have greater impairment in language skills than children with other lesions, including focal right hemisphere lesions. Studies of children with hemispherectomy for intractable epilepsy find that children with an isolated left and right hemisphere have similar potential for developing an adequate level of receptive vocabulary (Liegeois et al., 2008). Greater deficits after left than right hemisphere injury have been found in multiple studies, but the findings are restricted to specific ages, subdomains of language, and/or measures of language functioning. For example, before age 3 years, children with perinatal injury to left posterior brain regions were found to show greater delays in the development of expressive language skills than children with injuries in other locations (Thal et al., 1991) but this difference was no longer present at age 5 years (Vicari et al., 2000). Several studies of children with unilateral injury and with hemidecortication report that children with left hemisphere injury or an isolated right hemisphere have weaker skills than children with right hemisphere injury or an isolated left hemisphere in the domain of syntax, such difficulties comprehending sentences in which the only clue to meaning is syntactic structure (Dennis, 2009).

In children with injuries acquired after the early stages of language development, children with left hemisphere lesions, including left hemispherectomy have greater deficits than those with right hemisphere lesions. They show slower and less accurate understanding of syntactically difficult sentences in comparison to children with right hemisphere injuries of the same general type (Dennis, 2009). Those with lesions to the posterior left hemisphere have word finding difficulties and use semantic paraphasias and circumlocutions (Dennis, 2009). In a sample of children after hemispherectomy for epilepsy acquired on average at age 20 months, left hemisphere injury resulted in substantial language deficits, limitations in short-term verbal memory, and, to a lesser extent, in verbal intelligence (Liegeois et al., 2008). Such findings suggest that left lateralization of basic and complex language functions progresses after the initiation of language learning in the second year of life such that loss of the left hemisphere results in persistent language deficits. However, plasticity after a late destructive left hemisphere injury has been documented (Hertz-Pannier et al., 2002).

Children with a wide range of other central nervous system disorders, such as seizure disorder, traumatic brain injury, or tumor show deficits in language development (Dennis, 2009). Word finding difficulties are common in many different conditions, characterized by long pauses, creation of sequences that are not actual words (neologisms), circumlocutions, and excessive and inefficient productions (paraphasias). Children with traumatic brain injury in whom many brain pathways are injured show deficits in many aspects of language, from mastery of social functions of language (pragmatics) to accuracy of speech sounds (Campbell et al., 2007). These findings again suggest the importance of the integrity of the entire brain for competent language function.

Neuroimaging Studies of Children with Focal Injuries

Neuroimaging studies of children with early injuries provide additional insights into the neural organization of language development. Functional magnetic resonance imaging (fMRI – a magnetic resonance imaging method that directly measures changes in blood flow to brain areas to infer regions of brain activity during tasks or at rest) studies comparing children with focal left hemisphere injury to adults and typically developing children have found a rightward lateralization of language functions in those with injury. Activations in right hemisphere homologues of language areas during language tasks have been documented in children with early left middle cerebral artery stroke (Holland et al., 2007) and in a child with late left hemispherectomy (Hertz-Pannier et al., 2002). However, lateralization of language skills could not be reliably inferred from location of injury on an individual basis in children with epilepsy from early injury (Liegeois et al., 2004). Lesions in or near Broca's area were found to be more highly associated with peri-lesional activation within the damaged left hemisphere, whereas lesions remote from the classical language areas were more often associated with non-left language lateralization in several cases (Liegeois et al., 2004). In a study of children with early injury and their siblings, the level of both expressive and receptive language skills in children with early brain injuries correlated with the degree of activity in the left inferior frontal regions during a category fluency task, suggesting the relative advantage of left hemisphere networks for language function (Raja Beharelle et al., 2010). In addition, children with brain injury who showed bilateral activation in left and right superior temporal-inferior parietal regions had better language function than participants with either predominantly left- or right-sided unilateral activation.

Taken together, these findings support the interpretations from behavioral studies of the same population. Children with left hemisphere lesions activate nonclassical brain areas in language tasks, including regions of the right hemisphere. It is not yet clear whether the regions activated are those used in mature adult language functioning under other circumstances or whether these regions represent completely different networks. Predictions about which regions are activated after injury are difficult to make on the basis of lesion location or

size. However, activation of left hemisphere regions appears to confer an advantage to language function in children with early injuries.

Language Development after Injury to White Matter Pathways in Children

Neuropsychological studies also implicate the importance of white matter integrity in language development. Children born preterm, especially those born before 32 weeks gestation, are at high risk for injury to the white matter of the brain. DTI now reveals evidence of subtle white matter changes in a high proportion of children born preterm. Many studies document that performance on language assessments in this population remains below performance of well-matched controls (Barre et al., 2011; van Noort-van der Spek et al., 2012), even after controlling for intelligence quotient (Lee et al., 2011).

DTI studies in children born preterm confirm that the integrity of the white matter network affects language functioning. Multiple clusters on bilateral white matter tracts in both hemispheres were found to be associated with verbal IQ, linguistic processing speed, syntactic comprehension, receptive vocabulary, verbal memory, and reading in children born preterm (Feldman et al., 2012). The uncinate fasciculus connects the temporal lobes to the frontal lobes. Measures of both the left and right uncinate fasciculi have been found to correlate with receptive vocabulary (Mullen et al., 2011). The inferior frontal-occipital fasciculus runs from posterior occipital cortex through parietal and temporal cortex to the frontal lobe. The arcuate fasciculus connects the superior posterior temporal lobe and parietal lobe to the frontal lobe. Measures of the left and right arcuate fasciculi correlated with rapid naming scores in adolescents born preterm (Mullen et al., 2011). A recent case study identified a 12-year-old girl born preterm that was missing an arcuate fasciculus presumably from injuries associated with prematurity. This child nonetheless learned to speak and read. Further analysis of her white matter pathways found preserved ventral pathways (Yeatman et al., 2009). These findings suggest that alternative white matter pathways may support speech and reading, at least in selected cases. Finally, adolescents with evidence of injury to the corticospinal and corticobulbar pathways in the left hemisphere following preterm birth have been found to be at high risk for persistent difficulties in speech and oromotor control (Northam et al., 2012) though almost a third of the group did not have any such problems.

Language Characteristics after Cerebellar Injuries in Children

Children with surgical removal of cerebellar tumors in the right hemisphere show decreased verbal fluency, verbal intelligence, and syntax comprehension. Even 10 years after the surgery, such children often show speech deficits including consonant imprecision, reduced rate, and lack of variation in pitch and loudness. Poorer speech outcomes are more highly associated with right than left cerebellar tumors, implicating disruption of the reciprocal right cerebellar to left frontal pathways (Morgan et al., 2011). Cerebellar involvement in linguistic skills has also been shown in children born preterm. In this group, the cerebellum may be significantly reduced in volume and differentiation. Injuries in the cerebellum are associated with problems in language comprehension, expressive and receptive language, and syntax (Limperopoulos et al., 2007).

A syndrome called mutism with subsequent dysarthria has been described in children with severe lesions acquired later in development, such as Herpes simplex encephalitis and tumors of the posterior fossa. The total loss of speech is typically transient with subsequent speech characterized by slow rate, articulatory imprecision, and variable pitch. Though improvement over time is not uncommon, residual findings may be present a decade later (Dennis, 2009).

Genetic Abnormalities and Neural Basis of Language Development

The recent identification of a three-generation family in whom half of the members have a distinctive speech disorder and the other half of whom are entirely normal has opened a new chapter in the understanding of the connections of genes, neuroanatomy, neural function, and language (Vargha-Khadem et al., 2005). The autosomal-dominant speech disorder has been categorized as verbal dyspraxia; the affected and unaffected family members were found to differ primarily on the ability to repeat words, nonwords, and oral–facial movements. A genetic mutation in affected family members was localized to the FOXP2 gene on the long arm of chromosome 7. Detailed neuroanatomical studies of the brain have confirmed that the genetic mutation is associated reductions in the volume of the caudate nucleus, a subcortical gray matter structure within the basal ganglia that is typically associated with motor function. Other neuroanatomical findings include reductions in the volume of Broca's area, the precentral gyrus or motor strip, and the cerebellum. Affected family members showed less activation in Broca's area and more activation in posterior parietal and occipital regions during a language task than unaffected family members, linking the structural findings to abnormal function. The studies suggest that the development of fronto-striatal and fronto-cerebellar networks may be important for not only manual and motor sequences, but also speech motor control (Vargha-Khadem et al., 2005).

Neuroimaging in Typically Developing Infants and Children

Prelinguistic Language Processing in Infants

Modern neuroimaging has now opened the opportunity to study the neurobiology of language networks in intact, healthy, and typically developing infants and children, expanding understanding of the neural basis of language. fMRI and event-related potentials (ERPs – a measurement of the electrophysiological brain responses to a direct cognitive or sensory event) are the methods of choice for imaging healthy and typically developing infants and young children because they are noninvasive and avoid ionizing radiation.

Many behavioral studies document that human infants have impressive capacities to discriminate and categorize all of the phonetic distinctions within all human languages in early infancy, before they have acquired the ability to speak. Infants' phonemic representations can be modified without

training or implicit instruction based on the statistical distributions of speech input in order to converge to the phonemic categories of their primary language (Dehaene-Lambertz and Gliga, 2004). By the first birthday, children can no longer make distinctions between the speech sounds that are not part of their native language. This perceptual narrowing characterizes American Sign Language as well as verbal language (Kuhl and Rivera-Gaxiola, 2008). These abilities appear to involve general perceptual and cognitive abilities rather than language-specific abilities. To accomplish these abilities, infants appear to use the same neural structures as adults use in speech processing. fMRI and ERP studies find that an adult-like left-lateralized network of perisylvian areas become active when 3-month-old sleeping infants listen to sentences in their native language. The response was comparable for forward and backward speech stimuli, suggesting that the left hemisphere may be more responsive than the right to any auditory stimulus or to fast temporal auditory transitions characteristic of language rather than to language intonation (Dehaene-Lambertz et al., 2006). However, functional lateralization is not as strong in young children as it is in adults. Progressive neural commitment results in both the inability to detect nonnative phonetic contrasts and the concurrent ability to process larger language units, such as words and sentences. Using ERP, neural discrimination of phonetic contrasts at 7½ months of age correlated with language abilities at 2 and 2½ years of age (Kuhl and Rivera-Gaxiola, 2008).

Activations while listening to the mother's voice as opposed to other speech stimuli was modulated in several areas, including areas involved in emotional processing (amygdala, orbito-frontal cortex), and a large extent of the left posterior temporal lobe, suggesting that the mother's voice has a special status in the active engagement of posterior language areas. The fetus has the ability to hear its mother's voice more clearly than any other voice. These results suggest that early biological constraints on the language processing system interact with environmental inputs, particularly exposure to mother's voice, in the emergence of cortical networks for language processing in humans.

Broca's area activation during infancy can be found when stimuli are repeated, suggesting engagement of this region for verbal memory prior to any substantial verbalization. When a series of syllables change in either the characteristics of the voice (e.g., male to female) or the phonemes (e.g., /ba/ to /ga/), ERP studies find different scalp topographies and different latencies for the mismatch response. This finding suggests that, as in adults, different networks are involved in representations of different features of the stimulus (Dehaene-Lambertz et al., 2006). In addition, certain brain regions, such as the angular gyrus and precuneous, activate more during forward than backward speech, stimuli that vary primarily in intonation patterns rather than segmental features, suggesting sensitivity to prosodic patterns of the native language in the infant. Activation of the right dorsolateral prefrontal cortex in awake but not sleeping infants suggests that this region may be involved in active processes such as memory.

The development of white matter pathways in language-related temporal and frontal regions may be coincident with the development of vocabulary (Pujol et al., 2006). Rapid myelination of sensorimotor pathways begins almost immediately after birth, whereas rapid myelination of language-related pathways begins later at about 5 months of age. However, the rate of maturation was similar for temporal and frontal regions suggesting that the entire network acquires myelin simultaneously.

Functional Imaging in Toddlers and Preschoolers

Using ERP techniques on normal children learning language, a picture emerges that initial comprehension of language requires a wide and bilateral neural network. Progressive neural specialization and lateralization occurs in association with language experience rather than strictly with age. Specific ERP components in toddlers 13–20 months of age demonstrate differences between words that the child understands and words that are incomprehensible. Word comprehension shows both age and experience components on the ERP. At 13 months of age, the comprehension effects can be found over both the right and left hemispheres and over anterior and posterior regions. By 20 months of age, the effects are detectable only over the posterior left hemisphere. However, younger toddlers who are more advanced in language development show this laterality effect at a younger age than do less advanced children (Kuhl and Rivera-Gaxiola, 2008). In the toddler-preschool era phonological processing and word learning influence each other; children with strong phonetic skills learn vocabulary more quickly than children with weak phonetic skills and the more words children learn, the more they need to pay attention to phonetic units that distinguish one word from another (Kuhl and Rivera-Gaxiola, 2008). In terms of sentence processing, toddlers show adult-like components on ERP in response to syntactic and semantic errors but the differences in latency and scalp distributions of components differ compared to adults.

Functional Imaging in School-Aged Children and Adolescence

By the end of the preschool era, left lateralization can be identified on fMRI in many language functions: listening to speech versus backwards speech. Holland and colleagues completed a series of studies in which they used simple and complex syntactic and semantic tasks in children 5–18 years old (Holland et al., 2007). Simple tasks produced less activation than complex tasks. Syntactic tasks were less lateralized than semantic tasks. As children aged, the language tasks produced greater anterior change than posterior change. The greatest change in development was found in the verb generation task in which children generate as many verbs as possible for each noun presented (Holland et al., 2007). Comparisons of adult and school-aged children in reading single words out loud found that changes in left hemisphere activation patterns relate to both age and proficiency (Schlaggar et al., 2002). Thus, patterns of activation during language tasks continue to mature through childhood and into adulthood and are influenced by individual differences in skills as well as age.

White matter tracts within the language networks are late to mature. A DTI study that compared fiber tracts in adults and children found that dorsal pathways were less mature in school-aged children than in adults. This finding is consistent with functional imaging and behavioral data showing that

language functions remain relatively immature in school-aged children despite their impressive conversational abilities (Friederici, 2009).

Future Directions

Though much has been learned in the last decade about the neural basis of language development, many questions remain unanswered. How do abilities to perceive and categorize speech sounds lead to the production of speech? How do individual variations in anatomy of gray matter and white matter affect the development of language in individual children? How can we explain language impairments as features of other conditions or occurring in isolation?

As the field progresses, the usual taxonomy of language abilities into subdomains of phonology (the study of the organization and execution of sounds and phonemes in language and phonemes are the smallest unit of speech sounds that change the meaning of words), semantics, syntax, and pragmatics may need to be reconsidered in light of genetic and neurological studies. Performance factors, such as acoustic processing characteristics, motor planning, linguistic processing speed, and memory, may underlie apparent differences in language subdomains. Current methods of neuroimaging are well suited for the study of language in groups. However, vast individual differences are currently not well-understood using current methods. New methods under development that analyze structure and function in individuals will catapult the field. Future studies will also need to integrate multiple aspects of language functioning and multiple methods for understanding neural structure and function to broaden understanding about language as a complex and distinctly human function.

See also: Associative Learning in the Cerebellum; Biological Preconditions for Language Development; Cerebellum: Cognitive Functions; Language Development: Emergentist Theories; Language Processing, Functional Magnetic Resonance Imaging of.

Bibliography

Barre, N., Morgan, A., Doyle, L.W., Anderson, P.J., 2011. Language abilities in children who were very preterm and/or very low birth weight: a meta-analysis. Journal of Pediatrics 158 (5), 766–774 e761.

Basser, P.J., Mattiello, J., LeBihan, D., 1994. MR diffusion tensor spectroscopy and imaging. Biophysical Journal 66 (1), 259–267.

Campbell, T.F., Dollaghan, C., Janosky, J.E., Adelson, P.D., 2007. A performance curve for assessing change in percentage of consonants correct revised (PCC-R). Journal of Speech Language & Hearing Research 50 (4), 1110–1119.

Dehaene-Lambertz, G., Gliga, T., 2004. Common neural basis for phoneme processing in infants and adults. Journal of Cognitive Neuroscience 16 (8), 1375–1387.

Dehaene-Lambertz, G., Hertz-Pannier, L., Dubois, J., 2006. Nature and nurture in language acquisition: anatomical and functional brain-imaging studies in infants. Trends in Neurosciences 29 (7), 367–373.

Demir, O.E., Levine, S.C., Goldin-Meadow, S., 2010. Narrative skill in children with early unilateral brain injury: a possible limit to functional plasticity. Developmental Science 13 (4), 636–647.

Dennis, M., 2009. Language disorders in children with central nervous system injury. Journal of Clinical and Experimental Neuropsychology 32 (4), 417–432.

Dick, F., Wulfeck, B., Bates, E., Saltzman, N., Naucler, N., Dronkers, N., 1999. Interpretation of compelx syntax in aphasic adutls and chidlren with focal lesions or specific learning impirment. Brain and Language 69, 335–336.

Feldman, H.M., Lee, Eliana S., Yeatman, Jason D., Yeom, Kristen W., 2012. Language and reading skills in school-aged children and adolescents born preterm are associated with white matter properties on diffusion tensor imaging. Neuropsychologia under review.

Feldman, H.M., 2005. Language learning with an injured brain. Language Learning and Development 1 (3–4), 265–288.

Feldman, H.M., MacWhinney, B., Sacco, K., 2002. Sentence processing in children with early unilateral brain injury. Brain & Language 83 (2), 335–352.

Friederici, A.D., 2009. Pathways to language: fiber tracts in the human brain. Trends in Cognitive Sciences 13 (4), 175–181.

Gupta, P., MacWhinney, B., Feldman, H.M., Sacco, K., 2003. Phonological memory and vocabulary learning in children with focal lesions. Brain & Language 87 (2), 241–252.

Hertz-Pannier, L., Chiron, C., Jambaque, I., Renaux-Kieffer, V., Van de Moortele, P.F., Delalande, O., Le Bihan, D., 2002. Late plasticity for language in a child's non-dominant hemisphere: a pre- and post-surgery fMRI study. Brain 125 (Pt 2), 361–372.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nature Reviews Neuroscience 8 (5), 393–402.

Holland, S.K., Vannest, J., Mecoli, M., Jacola, L.M., Tillema, J., Karunanayaka, P.R., Byars, Anna W., 2007. Functional MRI of language lateralization during development in children. International Journal of Audiology 46 (9), 533–551.

Kuhl, P., Rivera-Gaxiola, M., 2008. Neural substrates of language acquisition. Annual Review of Neuroscience 31, 511–534.

Lee, E.S., Yeatman, J.D., Luna, B., Feldman, H.M., 2011. Specific language and reading skills in school-aged children and adolescents are associated with prematurity after controlling for IQ. Neuropsychologia 49 (5), 906–913.

Lenneberg, E.H., 1967. Biological Foundations of Language. John Wiley & Sons, Inc, New York.

Levy, Y., Amir, N., Shalev, R., 1994. Morphology in a Child with a Congenital, Left-Hemisphere Brain Lesion: Implications for Normal Acquisition Constraints on Language Acquisition: Studies of Atypical Children. Erlbaum, Hillsdale, NJ.

Liegeois, F.J., Connelly, A., Cross, J.Helen, Boyd, S.G., Gadian, D.G., Vargha-Khadem, F., Baldeweg, T., 2004. Language reorganization in children with early-onset lesions of the left hemisphere: an fMRI study. Brain 127 (Pt 6), 1229–1236.

Liegeois, F.J., Cross, J.H., Polkey, C., Harkness, W., Vargha-Khadem, F., 2008. Language after hemispherectomy in childhood: contributions from memory and intelligence. Neuropsychologia 46 (13), 3101–3107.

Limperopoulos, C., Bassan, H., Gauvreau, K., Robertson Jr, R.L., Sullivan, N.R., Benson, C.B., duPlessis, Adre J., 2007. Does cerebellar injury in premature infants contribute to the high prevalence of long-term cognitive, learning, and behavioral disability in survivors? Pediatrics 120 (3), 584–593.

MacWhinney, B., Feldman, H.M., Sacco, K., Valdes-Perez, R., 2000. Online measures of basic language skills in children with early focal brain lesions. Brain & Language 71 (3), 400–431.

Morgan, A.T., Liegeois, F., Liederkerke, C., Vogel, A.P., Hayward, R., Harkness, W., Vargha-Khadem, F., 2011. Role of cerebellum in fine speech control in childhood: persistent dysarthria after surgical treatment for posterior fossa tumour. Brain & Language 117 (2), 69–76.

Mullen, K.M., Vohr, B.R., Katz, K.H., Schneider, K.C., Lacadie, C., Hampson, M., Ment, L.R., 2011. Preterm birth results in alterations in neural connectivity at age 16 years. Neuroimage 54 (4), 2563–2570.

Murdoch, Bruce E., 2010. The cerebellum and language: historical perspective and review. Cortex 46 (7), 858–868.

Northam, G.B., Liegeois, F., Chong, W.K., Baker, K., Tournier, J., Wyatt, J.S., Morgan, A., 2012. Speech and oromotor outcome in adolescents born preterm: relationship to motor tract integrity. Journal of Pediatrics 160 (3), 402–408 e401.

Petitto, L.A., Zatorre, R.J., Gauna, K., Nikelski, E.J., Dostie, D., Evans, A.C., 2000. Speech-like cerebral activity in profoundly deaf people processing signed languages: implications for the neural basis of human language. Proceedings of the National Academy of Sciences of the United States of America 97 (25), 13961–13966.

Price, C.J., 2010. The anatomy of language: a review of 100 fMRI studies published in 2009. Annals of the New York Academy of Sciences 1191, 62–88.

Pujol, J., Soriano-Mas, C., Ortiz, H., Sebastian-Galles, N., Losilla, J.M., Deus, J., 2006. Myelination of language-related areas in the developing brain. Neurology 66 (3), 339–343.

Raja Beharelle, A., Dick, A.S., Josse, G., Solodkin, A., Huttenlocher, P.R., Levine, S.C., Small, S.L., 2010. Left hemisphere regions are critical for language in the face of early left focal brain injury. Brain 133 (Pt 6), 1707–1716.

Rowe, M.L., Levine, S.C., Fisher, J.A., Goldin-Meadow, S., 2009. Does linguistic input play the same role in language learning for children with and without early brain injury? Developmental Psychology 45 (1), 90–102.

Saur, D., Kreher, B.W., Schnell, S., Kummerer, D., Kellmeyer, P., Vry, M., Weiller, C., 2008. Ventral and dorsal pathways for language. Proceedings of the National Academy of Sciences of the United States of America 105 (46), 18035–18040.

Schlaggar, B.L., Brown, T.T., Lugar, H.M., Visscher, K.M., Miezin, F.M., Petersen, S.E., 2002. Functional neuroanatomical differences between adults and school-age children in the processing of single words. Science 296 (5572), 1476–1479.

Thal, D.J., Marchman, V., Stiles, J., Aram, D., Trauner, D., Nass, R., Bates, E., 1991. Early lexical development in children with focal brain injury. Brain & Language 40 (4), 491–527.

van Noort-van der Spek, Inge L., Franken, Marie-Christine J.P., Weisglas-Kuperus, Nynke, 2012. Language functions in preterm-born children: a systematic review and meta-analysis. Pediatrics 129 (4), 745–754.

Vargha-Khadem, F., Gadian, D.G., Copp, A., Mishkin, M., 2005. FOXP2 and the neuroanatomy of speech and language. Nature Reviews Neuroscience 6 (2), 131–138.

Vicari, S., Albertoni, A., Chilosi, A.M., Cipriani, P., Cioni, G., Bates, E., 2000. Plasticity and reorganization during language development in children with early brain injury. Cortex 36 (1), 31–46.

Yeatman, J.D., Ben-Shachar, M., Bammer, R., Feldman, H.M., 2009. Using diffusion tensor imaging and fiber tracking to characterize diffuse perinatal white matter injury: a case report. Journal of Child Neurology 24 (7), 795–800.

Neural Basis of Reading

Brigitte Stemmer, Université de Montréal, Montreal, QC, Canada

Published by Elsevier Ltd.

Abstract

Skilful reading is an astonishing accomplishment that implicates extracting sound and meaning from visual notations. This contribution investigates the neural substrates involved in reading and addresses questions such as whether there are neural systems in the brain dedicated to reading and whether reading affects brain organization in adults, in children learning to read, in the illiterate, and in the blind.

Introduction

In 1892, the neurologist J.J. Déjerine reported on a Parisian merchant who one morning woke up without being able to read a single word or letter despite being able to speak well and distinguish letters, objects, and persons around him (Déjerine, 1892 reported in Cohen, 2012). The patient was thus perfectly able to see and had all his cognitive capacities preserved but letters had become an abstract entity to which he could not assign any meaning. This story was taken as an indication that our brain possesses a system that is uniquely dedicated to reading. Since then researchers have tried to identify such a system or systems. In what follows, we will discuss the neural substrates that have been associated with sighted and non-sighted adult reading and with the process of learning to read in children, illiterate people, and the blind. Although most of this research has investigated alphabetical script languages there is a growing body of research that also looks at different script systems such as Chinese.

Neural Systems Involved in Adult Reading

One issue that has excited researchers is the observation that when reading scripts of languages as different as Japanese, Hebrew, or English, there is a small region in the brain that consistently becomes activated. This region has been called the visual word form area (VWFA) and lies within the left fusiform gyrus near the visual cortex in the occipital lobe (**Figure 1**). If this region is impaired, the patient may not be able to read anymore as was the case with Déjerine's patient. This was taken to support the notion that there are specific brain networks in the left occipital-temporal cortex (including the fusiform gyrus and the VWFA) dedicated to the reading process. However, as is frequently the case in research, what first seemed to be relatively clear-cut has turned out to be more complex. It has been found that the VWFA not only responds to written letters and words, but to all sorts of visual stimuli such as objects, faces, houses, and so on.

This observation has raised some controversy over the role of the VWFA and other regions activated during reading. While there is consensus that skilful reading requires the interaction of orthographic, phonological, and semantic processes, no consensus has been reached as to the neural substrates that underlie these processes (Cattinelli et al., 2013; Vogel et al., 2013). For example, one meta-analysis (that is a standardized analysis of a large amount of studies addressing a common topic) of 35 neuroimaging studies identified three neural network systems related to word and pseudoword reading (Cattinelli et al., 2013). One system, the *difficulty-modulated network*, was related to item difficulty and recruited brain regions in the left frontal region (Broca's region), the left superior parietal region, and the right cingulum. Broca's region was interpreted as contributing to translating orthographic representations into phonological ones, a process that is sensitive to item difficulty. The parietal and cingulum regions were thought to be related to attentional processes that are recruited, if stimuli are effortful. Another system, *the word-related network*, recruited left temporoparietal regions (including the fusiform gyrus). And a third system, the *pseudoword-related network*, comprised left and right parietal regions. These findings are in contrast to another meta-analysis based on five neuroimaging studies that did not identify distinct reading-related regions (Vogel et al., 2013). These researchers argued that the neural networks activated during word reading perform general processing. They *may* be used for reading but they are not used *predominantly for* and are *not specific to* reading (Vogel et al., 2013). The same meta-analysis also did not find neural regions for reading that *specifically* emerged with development. The question whether there are neural systems dedicated specifically to reading thus remains currently an open one. In addition, none of the meta-analytical approaches provide clear support for the classical models of reading such as the dual route model of reading or the connectionist single-mechanism account of reading.

Learning to Read

Being able to write and read what another person or you have written is a remarkable achievement. Despite different writing systems like Chinese or Japanese kanji, which depicts whole words, Japanese hiragana, which represents syllables, or English, which is an alphabetic system, we are able to acquire them all. Alphabetic writing systems rely on the fact that there is some correspondence between a letter (grapheme) and its sound (phoneme) (see **Figure 2**). In an ideal world, each phoneme would correspond to a single letter of the alphabet. Children learning to read would thus only have to learn the sonority that goes along with one letter. But unfortunately this is not the case, the correspondence between grapheme and phoneme is not always regular. Just consider the different

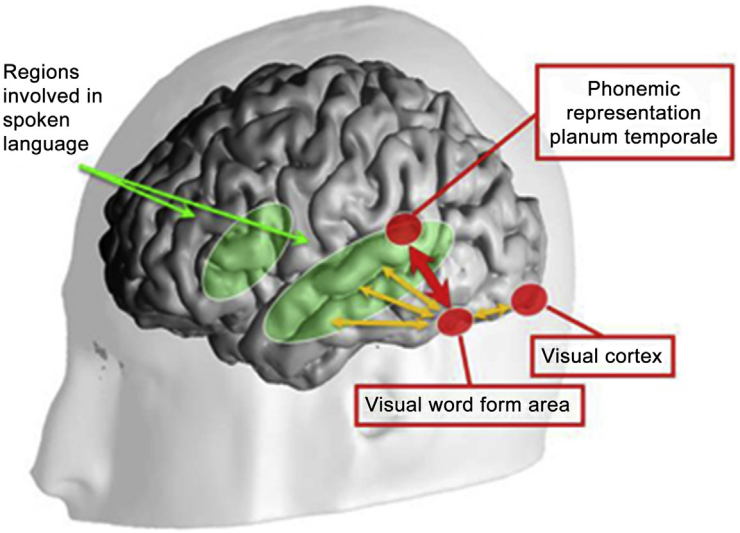

Figure 1 A vast brain circuit is active when people read or learn to read. As described in the text, the regions shown in red increase their activation and specialization during the acquisition of literacy. When readers become more skilled, activation decreases. In addition to these regions, there are also massive fiber bundles that link visual areas of the left hemisphere with areas in temporal brain regions. The figure is taken from Dehaene, S., 2013. Inside the Letterbox. How Literacy Transforms the Human Brain. Online publication: https://www.dana.org/Cerebrum/2013/Inside_the_Letterbox__How_Literacy_Transforms_the_Human_Brain/. The legend is phrased by Stemmer, B., based on Dehaene, 2013.

pronunciations of 'ou' in *soup*, *young*, or *thought*. Some languages like English are more irregular in this regard than others such as German or Italian.

Word-Level Reading

In order to be able to read, a child needs to consciously match the structures of the oral language with the visual code of the letters. This is not as simple as one may think. On the one hand, letters such as *e*, *o*, or *c* need to be distinguished as different letters although their form only differs minimally. On the other hand, letters such as 'A,' 'a,' or '*a*' need to be recognized as the same letters although their forms clearly differ. It has been suggested that reading acquisition partially recruits or 'recycles' preexisting cortical systems such as VWFA for novel tasks (such as reading) (Dehaene and Cohen, 2011). During the learning-to-read process, VWFA becomes more and more responsive to letters and letter sequences. fMRI studies have shown that activity in VWFA increases when orthographic strings (letters, words, and nonwords) become more familiar (**Figure 3**). Interestingly, the VWFA shows reduced activation in developmental dyslexia (*see* Dyslexia, Developmental). (For a summary, see Dehaene, 2011.) Growing expertise in reading is not only reflected in the modification of specific brain regions as shown by fMRI but also by the time course as reflected in EEG activity. When words or nonwords are presented on a screen, within 160–180 ms the brain registers this difference, which is reflected as a negative wave form in the EEG, the N170. The N170 does not show in illiterate children, but only after they have received instruction in reading (reported in Hruby and Goswami, 2011).

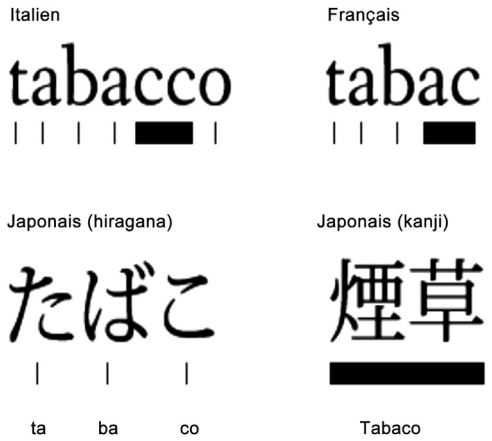

Figure 2 Different scripts and the relation of the written elements to their sounds. Some languages operate at the phoneme level (Italian, French), others at the syllable (Japanese hiragana) or whole word (Japanese kanji) level. Figure taken from Dehaene, S., 2011. Apprendre à lire. Odile Jacob, Paris. www.odilejacob.fr, p. 17.

Besides distinguishing and recognizing the form of letters and words, reading also entails developing efficiency in matching letters and the sound of language. It is thus not surprising that systematic correlations between visual and auditory brain areas and word decoding have been observed (Hruby and Goswami, 2011). One brain region that has been activated during these processes is the planum temporale, which is in close proximity to the primary auditory area. It has thus been suggested that this region is implicated in coding the sounds of a language.

Generally, word processing in children correlates with left-hemisphere activity. Increasing reading skills correlate with increasing neural activation in temporoparietal and occipito-temporal regions (see **Figure 1**). These correlations seem to depend on the developing visual expertise with letter strings

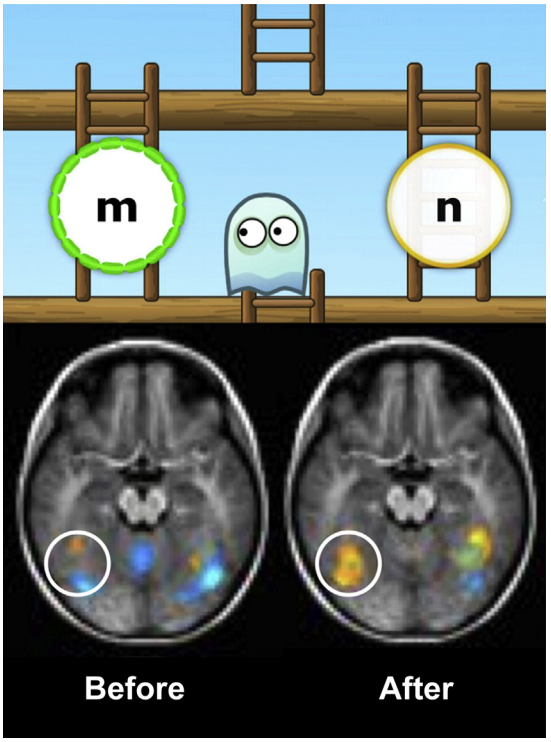

Figure 3 The upper part of the figure shows a sample stimulus from the Graphogame that was used to teach kindergarten children letter–speech sound associations. The brain of the children was investigated with fMRI before and after training (lower part of the figure). Increased activation in the visual word form area (VWFA) was seen in response to reading letters versus nonletters. Legend by Stemmer, B., based on McCandliss, B.D., 2010. Educational neuroscience: the early years. Proceedings of the National Academy of Sciences 107 (18), 8049–8050. Figure taken from McCandliss et al., 2010, which is based on work by Brem, S., et al., 2010. Brain sensitivity to print emerges when children learn letter–speech sound correspondences. Proceedings of the National Academy of Sciences 107 (17), 7939–7944.

such as words and developing skills in integrating letter and sound (Hruby and Goswami, 2011). The child learning to read alphabetic scripts needs to become aware that the spoken words of a language are made up of basic sounds (phonemes). Developing phonemic awareness in alphabetic language systems has been identified as a key stage in learning to read. In fact, language games that make the child manipulate syllables (e.g., rhyming games) have been shown to efficiently prepare the child for reading. Teaching a child to focus attention on letters and letter sequences that correspond to phonemes within the written word is essential. There is also some indication that systematically teaching grapheme–phoneme correspondence is more beneficial than teaching reading through global word forms without explicitly raising phoneme awareness. In fact, the latter strategy seems to activate inappropriate neural circuits in the right hemisphere and may hamper efficient learning (For a summary see Dehaene, 2011).

An interesting observation in children is that they frequently confuse letters that are mirror images of each other (such as *p* and *q* or *b* and *d*) when beginning to read. This has been related to the fact that the child's brain generally does not distinguish mirror images of an object and this difficulty is also reflected in distinguishing mirror letters.

Text-Level Reading

Most developmental reading studies focus on letter and word reading. However, reading is much more than reading words and implicates reading texts, comprehending syntactic structures, assigning meaning to and interpreting what has been read. Most generally, the language network is left lateralized by age four and overlaps largely with the network of older children and adults. Most differences are found with regard to higher order skills such as comprehension that continue to be fine-tuned. One of the few developmental studies comparing reading and listening comprehension of stories in children found that reading comprehension tasks activated a broader network in anterior frontal, temporal, and occipital (visual processing) regions (see Figure 4). These regions were also activated to a higher degree. This has been explained by different cognitive demands and strategies related to reading comprehension of stories. Further, the reading tasks activated more bilateral regions than the listening tasks although up to 25% of children had inconsistent laterality. When reading skills improve and phonological processing becomes more automatic, the magnitude of activation in posterior brain regions decreases.

Reading in Different Script Systems

We have already mentioned that different script systems follow different rules for assigning sounds to graphemes. An interesting question is thus whether different script systems implicate similar or different brain networks during reading. The answer is yes and no. Although reading procedures and the neural networks that underlie them are partly universal, the demands that alphabetic (e.g., English) and nonalphabetic (e.g., Chinese) scripts makes on reading and learning to read seem different. For example, Chinese writing differs from English alphabetic writing not only in the way graphic units are mapped to linguistic units (phoneme-level mapping in English and syllable-level mapping in Chinese), but also in its visual form (such as spatial configuration of stroke sequences in Chinese). Both, alphabetic and Chinese reading engage phonology as universal components but the procedures that lead to its pronunciation depend on the linguistic units inherent in the writing system (For a summary see Perfetti et al., 2013). While the retrieval of meaning from written units thus seems universal across writing systems, the retrieval processes may differ across writing systems. For example, in Chinese retrieving the meaning for a word may be triggered by a semantic radical that is part of the character or by the whole character.

Do the neural systems for reading reflect this duality, that is, universality on the one hand and writing system differences on the other? Contrasting English and Chinese reading has indicated that visual processing activates posterior (occipital and occipital-temporal) areas for both languages. However, while alphabetic writing shows strong left-lateralized effects in visual

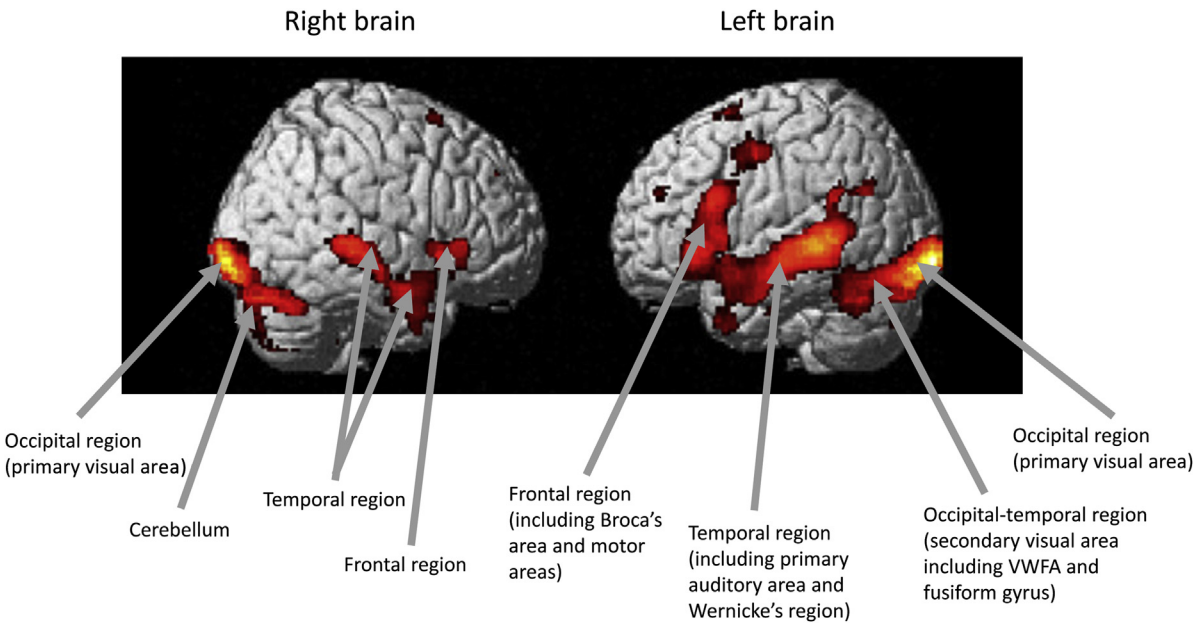

Figure 4 The result of brain activations in response to reading comprehension tasks based on small stories for children aged 4–12 years (VWFA, visual word form area). Figure from Berl, M.M., et al., 2010. Functional anatomy of listening and reading comprehension during development. Brain and Language 114 (2), 115–125; Legend and labeling by Stemmer, B., based on Berl et al., 2010.

areas (including the VWFA), Chinese writing shows additional activation of visual areas in the right hemisphere (see Figure 5). The latter effect has been interpreted as reflecting the spatial configuration of Chinese character components. In addition, Chinese reading consistently shows greater activation in left frontal brain regions (including Broca's area) (Perfetti et al., 2013; Wu et al., 2012). This has been interpreted as reflecting higher demands of character composition or lexical selection through an orthographic memory in Chinese reading (Perfetti et al., 2013).

The Illiterate Brain

According to the United Nations Educational, Scientific and Cultural Organization's global monitoring report for 2013, 4774 million adults are still illiterate, two-thirds of which are women, with the highest numbers in India followed by China. Most Western-industrialized countries show a high literacy rate, although exact data are missing. According to the US Department of Education, in 2003 (the latest year data were available) 14% of US adults, that is about 32 million, lacked basic prose

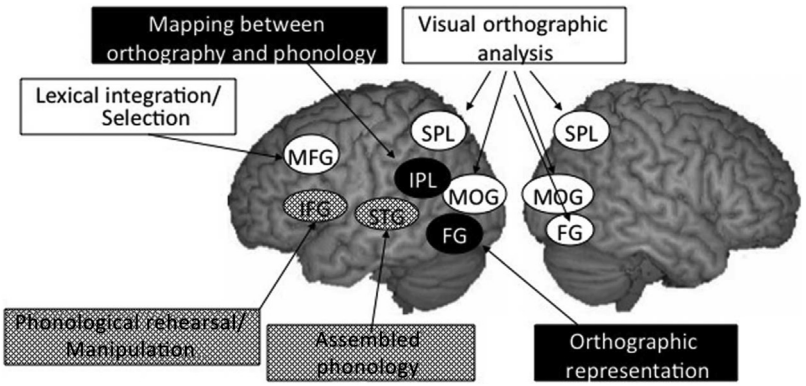

Figure 5 Brain areas that support reading in English and Chinese. Activation for both systems is seen in left hemisphere posterior regions that support visual identification and left-hemisphere parietal region that may support graphic-to-linguistic mapping. Differences between the two systems include greater activation for Chinese of right-hemisphere posterior areas that support character identification and greater activation in frontal area that may support selection of character connections to meaning and pronunciation. Cross-hatches mark areas more involved in English than Chinese reading (IFG and STG). Areas with white backgrounds are involved in Chinese more than English. Areas with black backgrounds (IPL, FG) are involved in both Chinese and English for adults. Figure and legend taken from Charles Perfetti, et al., 2013. Specialization and universals in the development of reading skill: how chinese research informs a universal science of reading. Scientific Studies of Reading 17 (1), 5–21. doi: 10.1080/10888438.2012.689786.

literacy skills. A question, one may ask is whether the brain of those who have not learned to read at a young age and remained illiterate is organized differently compared to literate people. One of the few studies addressing this issue compared illiterate Brazilian and Portuguese adults who had never learned to read with ex-illiterates who had received adult alphabetization classes and achieved a modest level of literacy. Another comparison group was literate adults of the same socioeconomic background (Dehaene et al., 2010). It was observed that the VWFA was massively affected by literacy, to the extent that by simply measuring the brain activity, the researchers could predict a large proportion of the variance in the number of words the people could read per minute (see **Figure 6**) (Dehaene, 2013; Dehaene et al., 2010).

During the acquisition of literacy, activation in the planum temporale, the VWFA and the visual cortex increased. At the same time, activation also increased in occipital regions in response to all sorts of other nonreading-related visual stimuli. This was interpreted as indicating that other visual processes seem to profit from the acquisition of literacy and become more accurate and efficient. Another observation was that the network for spoken language also changed under the influence of reading in the sense that less activation was observed. This was interpreted as indicating that reading facilitates the comprehension of complex sentences and requires less effort in literate people. It was further suggested that a pathway (the arcuate fasciculus) connecting the posterior temporal regions of the brain (including the VWFA) with inferior parietal and frontal brain regions may link visual and phonological areas and play a role in grapheme-to-phoneme conversion (Dehaene, 2013). Studies like this are difficult to perform and it is thus not surprising that only a relatively small number of people were investigated. Replication of the study with a larger number of people is needed to strengthen the authors' arguments.

Reading in the Blind

In sighted people, reading relies on visually identifying letters. When blind people use Braille, visual letter identification is replaced by a tactile system (raised dots arranged in columns). Braille reading involves the control of finger movements, the

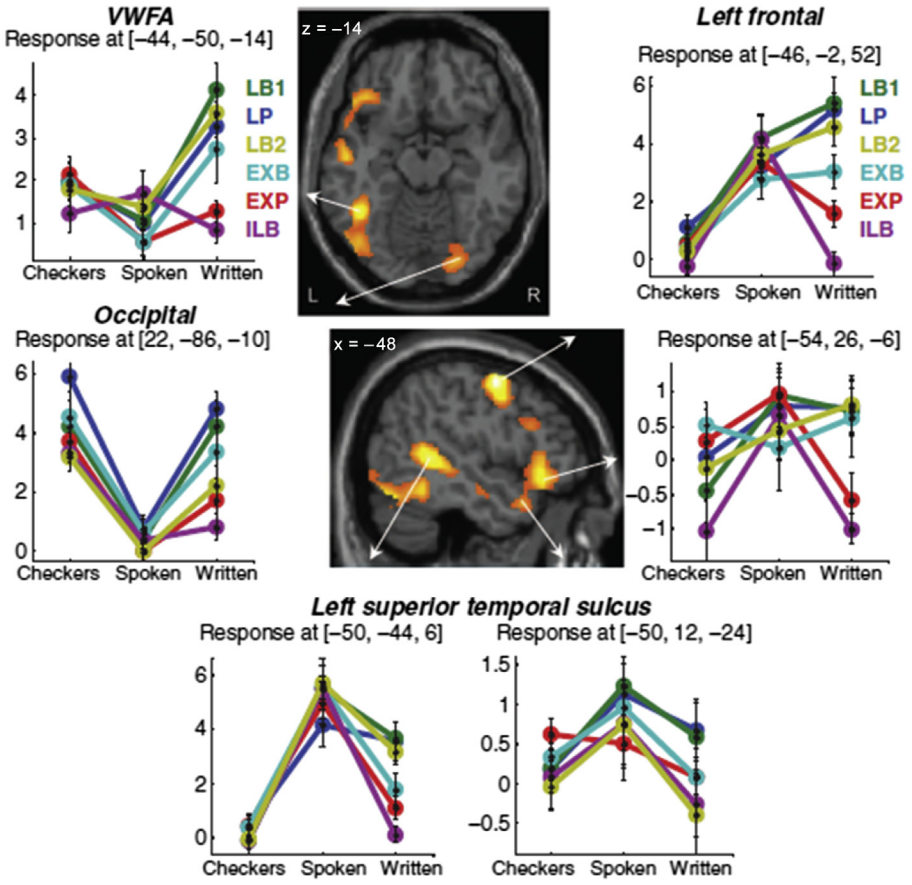

Figure 6 The figure shows the effects of literacy on the brain in response to written sentences. The horizontal and sagittal brain slices show the regions that were modulated by reading performance during the viewing of simple sentences. A large effect can be seen in the left occipital-temporal site (including the classical VWFA) and right occipital sites. Reading performance also modulated a vast left-hemisphere language network. The plots report activation to visual checkers, spoken and written language relative to rest in arbitrary units. Colored labels refer to participant groups and are ordered according to reading performance (LB1, LP, LB2 = literates; EXB, EXP = ex-illiterates; ILB = illiterates; VWFA = visual word form area). Figure taken from Dehaene, S., Pegado, F., Braga, L.W., Ventura, P., Filho, G.N., Jobert, A., et al., 2010. How learning to read changes the cortical networks for vision and language. Science 330 (6009), 1359–1364. Figure on p. 1360. Legend by Stemmer, B., based on Dehaene et al., 2010.

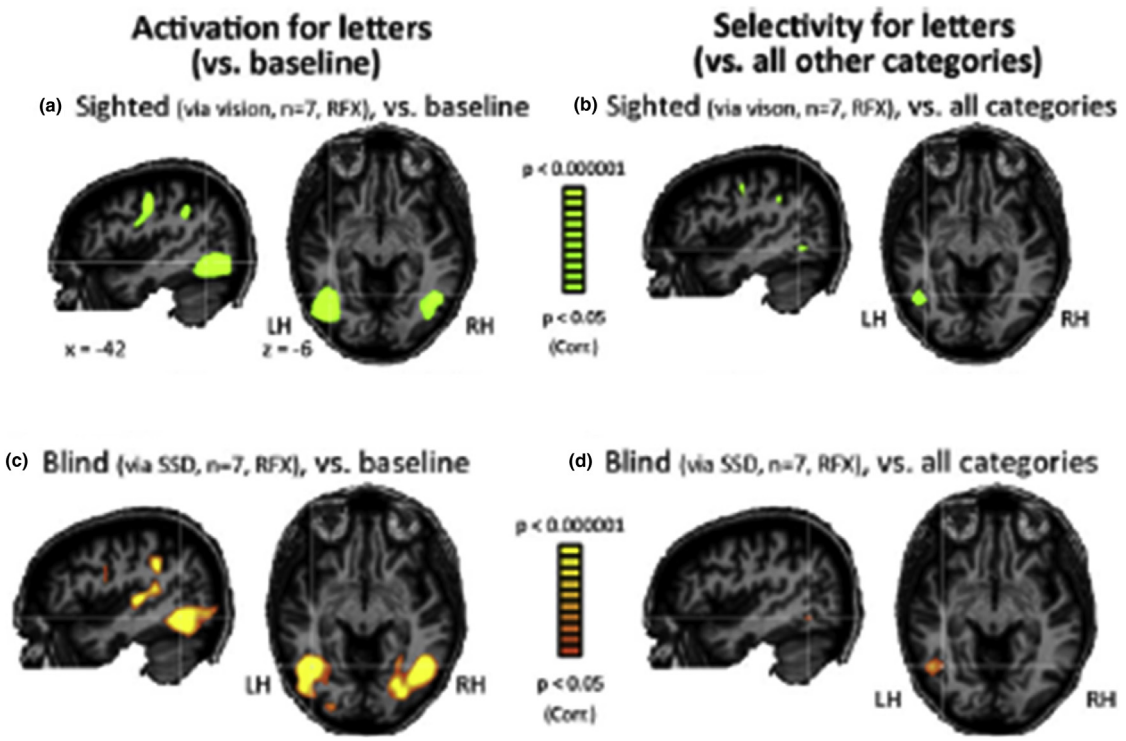

Figure 7 Panels (a) and (b) show the brain activation during normal reading of sighted individuals. Panels (b) and (c) show the brain activation when congenitally blind individuals were reading using soundscapes. Extensive bilateral activation of the ventral visual cortex including visual word form area (VWFA) can be seen when letter reading was compared to baseline (a and c), and left activation of the ventral visual cortex including VWFA when letter reading was compared to all other categories of visual stimuli (b and d). Figure taken from Striem-Amit, E., et al., 2012. Reading with sounds: sensory substitution selectively activates the visual word form area in the blind. Neuron 76 (3), 640–652. Legend by Stemmer, B., based on Striem-Amit et al., 2012.

perception of raised dots and the recognition of the pattern in which these dots are arranged. The tactile information must then be converted into meaningful patterns that have lexical and semantic properties (Sadato, 2005). Processing of tactile information usually recruits the somatosensory region of the frontal-parietal cortex while the processing of visual information (such as reading) involves the primary and secondary visual areas in the occipital cortex. When people who are blind from birth read Braille (that is use their fingers to touch the raised dots), the major activity detected by functional brain imaging is not, as one may expect, in the somatosensory area where touch is usually processed, but in the visual areas of the occipital lobe (for a summary, see Amedi et al., 2005). A similar observation was made for the auditory modality. Usually, sounds activate the auditory cortex in the temporal lobe. However, when congenitally blind individuals were 'reading' using the auditory modality via a system that converted the environment to auditory sounds, so-called soundscapes, regions in the ventral visual cortex, including the VWFA, were activated (Striem-Amit et al., 2012) (see **Figure 7(b) and 7(d)**). One explanation for these transfer phenomena is that the traditionally visual areas create new sensory associations and connectivity patterns and are thus recruited for some other function such as processing touch or auditory input. An alternative explanation is that the occipital lobe is inherently equipped for various functions but that these functions are normally inhibited in sighted people and are only revealed by visual loss (Amedi et al., 2005).

See also: Chinese Literacy Acquisition: A Multidimensional Puzzle; Dyslexia, Developmental; Dyslexias, Acquired and Agraphia; Illiteracy, Sociology of; Literacy and Illiteracy, History of; Reading Skills, Acquisition of: Cultural, Environmental, and Developmental Impediments.

Bibliography

Amedi, A., Merabet, L.B., Bermpohl, F., Pascual-Leone, A., 2005. The occipital cortex in the blind: lessons about plasticity and vision. Current Directions in Psychological Science 14 (6), 306–311.

Cattinelli, I., Borghese, N.A., Gallucci, M., Paulesu, E., 2013. Reading the reading brain: a new meta-analysis of functional imaging data on reading. Journal of Neurolinguistics 26 (1), 214–238.

Cohen, L., 2012. Le cerveau lecteur. Revue Neurologique 168 (10), 771–772.

Dehaene, S., 2011. Apprendre à lire. Des sciences cognitives à la salle de classe. Odile Jacob, Paris.

Dehaene, S., 2013. Inside the Letter Box: How Literacy Transforms the Human Brain. Cerebrum. https://www.dana.org/Cerebrum/2013/Inside_the_Letterbox__How_Literacy_Transforms_the_Human_Brain.

Dehaene, S., Cohen, L., 2011. The unique role of the visual word form area in reading. Trends in Cognitive Sciences 15 (6), 254–262.

Dehaene, S., Pegado, F., Braga, L.W., Ventura, P., Filho, G.N., Jobert, A., et al., 2010. How learning to read changes the cortical networks for vision and language. Science 330 (6009), 1359–1364.

Déjerine, J., 1892. Contribution à l'étude anatomo-pathologique et clinique des différentes variétés de cécité verbale. Compte Rendu Hebdomadaire des Séances et Mémoires de la Société de Biologie 4, 61–90.

Hruby, G.G., Goswami, U., 2011. Neuroscience and reading: a review for reading education researchers. Reading Research Quarterly 46 (2), 156–172.

Perfetti, C., Cao, F., Booth, J., 2013. Specialization and universals in the development of reading skill: how chinese research informs a universal science of reading. Scientific Studies of Reading 17 (1), 5–21.

Sadato, N., 2005. How the blind "see" Braille: lessons from functional magnetic resonance imaging. Neuroscientist 11 (6), 577–582.

Striem-Amit, E., Cohen, L., Dehaene, S., Amedi, A., 2012. Reading with sounds: sensory substitution selectively activates the visual word form area in the blind. Neuron 76 (3), 640–652.

Vogel, A.C., Church, J.A., Power, J.D., Miezin, F.M., Petersen, S.E., Schlaggar, B.L., 2013. Functional network architecture of reading-related regions across development. Brain and Language 125 (2), 231–243.

Wu, C.-Y., Ho, M.-H.R., Chen, S.-H.A., 2012. A meta-analysis of fMRI studies on Chinese orthographic, phonological, and semantic processing. Neuroimage 63 (1), 381–391.

Neural Correlates of Social Development

Daina Crafa, Douglas Mental Health Institute, McGill University, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D.A. Snow, volume 4, pp. 2212–2219, © 2001, Elsevier Ltd.

Abstract

Humans are social animals and the human brain is simultaneously adapted to facilitate social interaction and to be shaped by it. Many brain processes and regions contribute to social development. This article will discuss anatomical changes that occur during human development as well as functional roles of key brain regions.

Introduction: The Situated Brain

For the duration of our lives, our brains are influenced by the social world. Neonates begin learning about the world by watching their parents. These interactions help shape their brain formation. In fact, sociocultural influences play a large role in human brain development and processing across the life span. Under extreme circumstances, this influence can be deterministic and limiting. For example, infants kept in prolonged social isolation exhibit underdeveloped brain formation and fail to acquire basic motor skills or language (Rees, 2008; Watts-English et al., 2006; Lin et al., 2005; Hildyard and Wolfe, 2002; Perry, 2002; Perry and Pollard, 1997).

However, even under normal circumstances, the social environment influences epigenetic and endocrine processes (Curley et al., 2011; Champagne, 2010) and shapes neural pathways (Rule et al., 2013). Clarke and Chambers (1998) posit that the mind is 'extended,' meaning that it is embodied, enacted, situated, and distributed. The same can be said about the brain's relationships with the body and the world. Biological processes influence the brain's responses, potentials, and limitations, while social interactions and the greater environment provide stimuli for the brain to respond to. In each of these cases, a rich interplay between converging social and biological processes influences neural function and response.

The Culture-Brain-Behavior Interaction model (Crafa and Nagel) characterizes this interplay in terms of a feedback loop between social, neural, cognitive, and behavioral processes. When people encounter new social situations or information, learning occurs. This learning, however subtle, has the potential to alter, relax, or constrain psychological, neural, and biological processes. Likewise, these processes filter what is learned. This bidirectional interaction produces traces of culture in these processes but also accounts for individual variability.

This article will focus on two aspects of that feedback loop, specifically the relationship between human social development and the brain. When relevant, related biological processes will be discussed. Since other article in this encyclopedia address language development, this article will primarily discuss neural processes related to nonverbal social interactions and other fundamental features of human development that make social interaction possible, as well as some complex processes, such as those underlying empathy and identity. The relationship between neural processes and culture will also be discussed.

Structural Brain Development: Infancy through Early Adulthood

Human brain development begins during gestation, when the first neurons form and migrate toward different areas of the developing brain, and eventually differentiate and specialize in key neurophysiological processes (Perry, 2002; Rakic and Lombroso, 1998). At birth, neonates already have approximately 100 billion neurons, nearly the same as adults. However, the brain only weighs approximately 25% of what it will weigh in adulthood. During the first 3 years of life synapses, the connections between neural cells rapidly develop and the connectivity between neurons and brain regions exponentially increases. Myelination, the development of fatty tissue that insulates and lubricates neural processes, also begins. During the first decade of life, the human brain also undergoes substantial anatomical growth and changes in gray matter and white matter (Blakemore, 2008; Webb et al., 2001). Synaptic pruning occurs during childhood and adolescence, during which time nearly half of the synapses developed during early childhood are discarded while new synapses are formed. Myelination, on the other hand, continues to develop into adulthood. These changes produce reductions of gray matter and increases in white matter volume and integrity into adulthood, normally stabilizing around the mid-20s (Giorgio et al., 2010; Blakemore, 2008).

Social experiences play distinct roles in human brain development. Children who are raised in conditions of extreme neglect or isolation exhibit smaller brain volumes and fail to develop normal motor and social skills (Perry, 2002). Both synaptic and myelin development are affected by life experiences (Shonkoff and Phillips, 2000). Adequate nutrition and guardian attention are required for normal brain development. Some studies demonstrate that maternal touch during early infancy may improve cognitive processes and benefit brain development (Weiss et al., 2004). Other social interactions also help the child make necessary developmental progress, such as acquiring adaptive motor and language skills (Meltzoff, 2007). For example, the more a child is exposed to speech, the stronger these neural connections become; whereas, less exposure can cause these neural connections to diminish (Perry, 2006). Accordingly, acute isolation, emotional trauma, or stress can

play inverse roles in brain development, potentially inhibiting key processes (Shonkoff and Phillips, 2000).

The Social Brain: Functional Brain Processes

The 'social brain' commonly refers to brain regions that are involved in social cognition. Certain injuries can cause social impairment or disrupt normal social functioning, leading researchers to conclude that social cognition is unique from other types of neurocognitive processes (Grossmann and Johnson, 2007). Key areas of the social brain include the superior temporal sulcus (STS), anterior cingulate cortex (ACC), medial prefrontal cortex (mPFC), inferior frontal gyrus (IFG), as well as the anterior insula and the amygdala (Blakemore, 2008; Frith and Frith, 2007.).

These brain regions play functional roles in social cognition. Human infants are born ready to learn through social experiences with other humans (Csibra and Gergely, 2006). Neonates are born with multiple tendencies that facilitate social interaction, such as natural abilities for eye contact, face recognition, and emotional processing, as well as attentional preference for biological motion. Other social abilities, such as perception of others, joint attention, and prosociality, develop during infancy and early childhood. Adolescence is a time of rapid change in the social brain, affecting basic social abilities like facial recognition, self-awareness, and theory of mind. These abilities correspond with specific neural activity that appears to simultaneously facilitate and accommodate the social environment.

Recognition of Others: STS

Beginning at birth, human neonates demonstrate attentional preferences for human faces and eye contact (e.g., Farroni et al., 2005). This innate preference for the eyes and the face helps infants to bond with their parents to learn salient behaviors. Studies in adult humans have shown that the STS (**Figure 1**) is active when viewing faces (Blakemore, 2008). The STS has also been implicated in processing biological motion. Like eye and face gaze, having a preference for biological motion is advantageous to the infant for identifying important social and environmental information. Likewise, studies have demonstrated similar responsiveness of the STS when viewing biological motion (Blakemore, 2008; Saygin, 2007). Together, these studies point to an important role of the STS in processing these conspecifics, and in the recognition of other humans more generally. These processes are considered foundational for human social interaction, in part because they are innate abilities and in part because they act as building blocks for more complex social processes.

One of these processes is mentalizing, i.e., the ability to understand the internal states of others. Studies recording brain activity during mentalizing frequently report STS activity (Vander Wyk et al., 2009; Blakemore, 2008). Currently, there is some debate about how to precisely interpret this activity (Blakemore, 2008). On the one hand, given the role of the STS in processing conspecifics, it follows that it may play a similar role in mentalizing. Specifically, Saxe (2006) proposes that the STS may interpret the goals of others based on their body movements. Frith (2007), on the other hand, proposes that the role of the STS is limited to the anticipation of future movements of these conspecifics. While this predictive role facilitates mentalizing, it may not directly interpret the actions it is predicting. STS activity in mentalizing is stronger in U.S. children than Japanese children during word-based stimuli; whereas, Japanese children show higher STS during cartoon mentalizing tasks (Han and Northoff, 2008). Some researchers suggest that this could be due to the different role of disposition across cultures. While U.S. cultures tend to value disposition, environmental factors and social context appear to play larger interpretive roles in many East Asian cultures (Ames and Fiske, 2010).

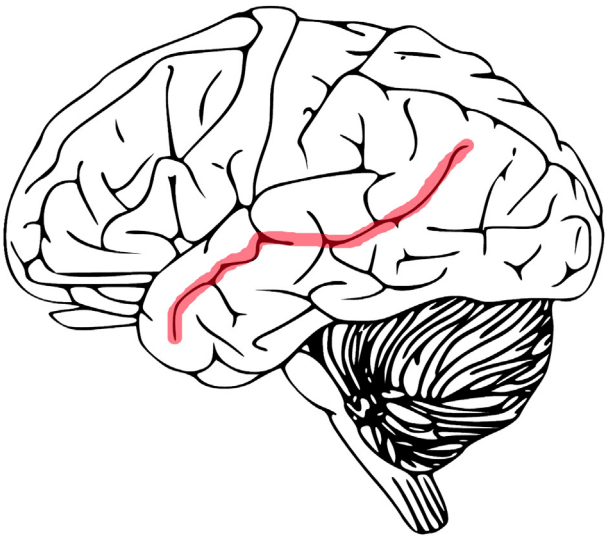

Figure 1 The superior temporal sulcus (STS), indicated here in red, is the crevice on the surface of the temporal lobe that separates the temporal gyrus from the middle temporal gyrus. "Brain 01" (http://openclipart.org/detail/35791/brain-01-by-anonymous) by Anonymous is licensed under CC BY 2.0 and modified incompliance with this license agreement.

Emotion and Response Selectivity: ACC

The ACC (**Figure 2**) has also been implicated in understanding internal states of other, particularly in terms of emotion processing. Many studies of the ACC demonstrate that it is active when processing the emotions of other as well as when generating emotional responses. Early studies of the ACC suggest that it helps to determine the intensity of an emotional experience and ties it to emotional response regulation (Lane et al., 1998; Wilson and Chang, 1974). More recently, it has been associated with many high-level processes, including attention and behavioral response selection. Lane et al. (1998) propose that the ACC's role in emotion processing may be a by-product of these high-level processes. For example, the ACC appears to become more active when information is personally relevant and positively valenced (Moran et al., 2006). Findings such as these demonstrate that emotion processing is implicit to ACC activity, and the observed response selectivity implicates the

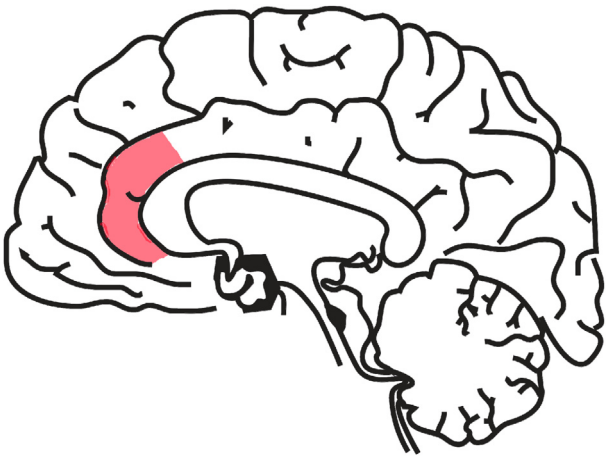

Figure 2 The anterior cingulate cortex (ACC), highlighted in red, occupies the frontal portion of the cingulate cortex. "58294main_-The.Brain.in.Space-page-157-brain" (http://openclipart.org/detail/38533/-by-rejon-38533) by rejon is licensed under CC BY 2.0 and modified incompliance with this license agreement.

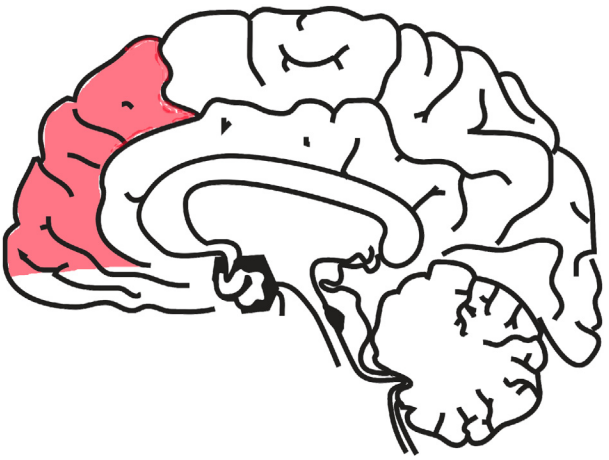

Figure 3 The medial prefrontal cortex (mPFC) in red sits adjacent to the anterior cingulate cortex (ACC) and occupies the highlighted portion of the prefrontal cortex, which makes up a majority of the frontal portion of the brain. "58294main_The.Brain.in.Space-page-157-brain" (http://openclipart.org/detail/38533/-by-rejon-38533) by rejon is licensed under CC BY 2.0 and modified incompliance with this license agreement.

ACC as a key player in social reward. Other studies demonstrate a bidirectional response to valence. For example, ACC activity in adults increases during social rejection and corresponds with the strength of reported pain the rejection has caused (Eisenberger et al., 2003, 2007). Such findings demonstrate that the ACC may selectively respond to any emotion that is personally relevant. Likewise, it selectively chooses how to respond. Given its selectivity, ACC activity appears contingent on the social norms and expectations of an individual. Considering that norms and expectations are often heavily influenced by social context and cultural background, ACC activity likely also varies partially as a function of culture, which may be an interesting topic for future research.

Knowledge of Self and Others: mPFC

Like the ACC, the mPFC (**Figure 3**) has also been implicated in emotion processing. The mPFC helps to determine the appropriateness of an emotion based on contextual cues, and control the degree and longevity of its expression (Etkin et al., 2011). However, the mPFC is best known for its role in 'self' processes more generally, such as self-control and self-identification. The mPFC has been shown to preferentially respond to personal preferences and values. For example, activity in the mPFC increases when identifying behaviors or values that describe oneself (Lombardo et al., 2010; Chiao et al., 2009). However, mPFC activity appears to be contingent on personal values. Studies comparing people who value individualism with those who value collectivism demonstrate that, when someone values collectivism, the mPFC responds to questions pertaining to both self and a close other (e.g., one's mother) with similar activity increases. In contrast, the mPFC preferentially processes only self-information when a person values individualism. During adolescence, a critical time for the development of personal identity, mPFC activity is often linked with decisions whether or not to engage in risky behavior. Several studies report that teens who take more risks have heightened mPFC activity during risk-taking experiments. Other studies show that teens who refuse to take risks show greater mPFC activity. This apparent contradiction potentially suggests that the decision to take risks may relate to personal identity.

Activity in the mPFC is theorized to facilitate interpretation of others by using oneself as a reference. One study, for example, showed that seeing oneself performing an action facilitates interpretation of the emotions of others (Li and Tottenham, 2011). In this study, participants who viewed their own faces before viewing the faces of others showed increased emotion recognition. These findings support the idea that self-information facilitates mentalizing about others. Additionally, these findings show that embodied self-representations facilitate other-mentalizing, which suggests that some high-level areas may be employed when viewing embodied information. In order to better understand the role of self-information in learning and mentalizing, the underlying processes must be further explored.

Self–Other Mapping: IFG

The IFG (**Figure 4**) is part of the mirror neuron system, a network of brain areas that are hypothesized to facilitate social cognition and action imitation. Like the ACC and mPFC, the IFG appears to play a role in interpreting others. In contrast with these two other brain regions, the IFG appears to specialize in social perception pertaining to movement and action imitation. Disruption of IFG activity makes imitation more difficult, and also reduces accuracy when interpreting the actions of others (Keuken et al., 2011). The IFG is also selectively active in response to facial emotional cues, and

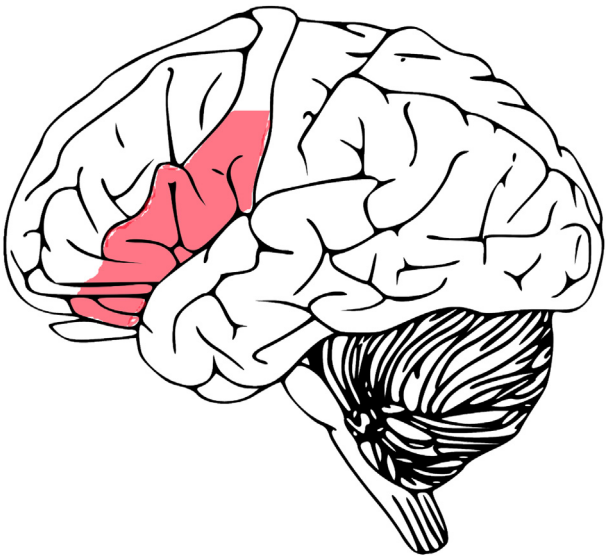

Figure 4 The inferior frontal gyrus (IFG), shown here in red, is a cleft in the prefrontal cortex, which is bordered by the inferior frontal sulcus, the lateral fissure, and the inferior precentral sulcus. "Brain 01" (http://openclipart.org/detail/35791/brain-01-by-anonymous) by Anonymous is licensed under CC BY 2.0 and modified incompliance with this license agreement.

appears to be essential for processing motion-based social information (Dapretto et al., 2006).

Specialized Emotion Areas: Anterior Insula and Amygdala

Both the anterior insula and the amygdala (**Figure 5**) are involved in emotion processing and are usually active when interpreting any emotional cue (Lamm and Singer, 2010). However, these regions are perhaps best known for being particularly sensitive to certain emotions. Multiple studies link anterior insula activity with expressions of disgust, and one study even showed that the amplitude of its activity corresponds with the expressed magnitude of disgust (Phillips et al., 1997). The amygdala, on the other hand, specializes in fear and aggression, and hyperactivity of the amygdala can cause difficulty in interpreting emotional faces (Schultz, 2005). Disgust and threat are both fundamental emotions that help to extend longevity, by influencing what we avoid eating and when we protect ourselves. Although these basic emotions can be found in every culture, the defining features of gross and threatening things vary widely. For example, eating the eyes or brains of an animal is widely accepted in some parts of the world and considered disgusting in others. Similarly, hitchhiking is considered safe in some countries but can be very dangerous in others. Such cultural variability suggests that anterior insula and the amygdala activity likely vary according to a person's social context and cultural background.

Neural Plasticity and Sociocultural Learning

Neural plasticity is a key component of human brain development, and necessary for social learning. Neural plasticity refers to the brain's innate ability to form new synaptic connections as well as to alter or eliminate existing ones. Learning that occurs from new social experiences leads to these synaptic changes (Black, 1998). This ability is so dynamic that the brain is sometimes able to compensate for severe damage, such as traumatic injury or stroke, by recruiting other brain areas to perform the functions the damaged areas are normally responsible for. This flexibility is ubiquitous in the brain; although, it is less often observed in areas that control basic functions such as heart rate. Neural plasticity theoretically makes it possible for people to learn appropriate sociocultural behaviors during childhood as well

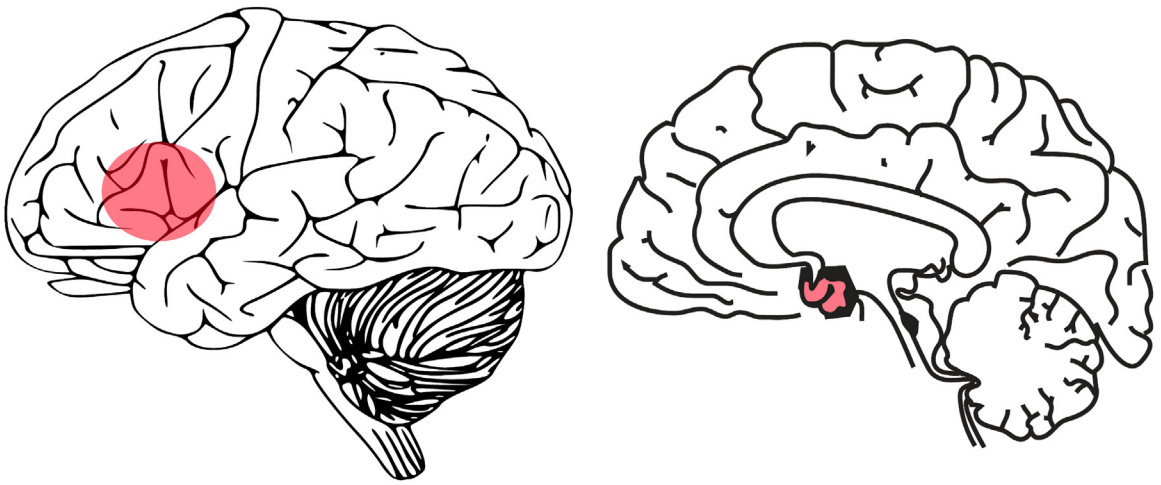

Figure 5 The anterior insula (left) cannot be easily seen via diagrams of the brain because it is positioned under the surface layer of the cerebral cortex, in approximately the highlight location. The amygdala (right) is located in the temporal lobe. "Brain 01" (http://openclipart.org/detail/35791/brain-01-by-anonymous) by Anonymous is licensed under CC BY 2.0 and modified incompliance with this license agreement; and "58294main_The.Brain.in.Space-page-157-brain" (http://openclipart.org/detail/38533/-by-rejon-38533) by rejon is licensed under CC BY 2.0 and modified incompliance with this license agreement.

as to adapt to new customs and changing social environments as an adult. Reduced plasticity has been observed in people with social and mental impairments, suggesting that plasticity's role is fundamental for normal human social development (for review, see Zatorre et al., 2012).

Challenges and Opportunities for Developmental Social Neuroscience

Social neuroscience is a prominent and expanding field, within which developmental research is a crucial subset. Despite the vast body of literature that has accumulated since the field's inception, a majority of this research comes from North America and Western Europe. Considering that social cognition and related brain activity are largely contingent on immediate social context as well as individual cultural background, social neuroscience research is potentially very specific to the population being studied. The underrepresentation of many global populations in research studies may limit the generalizability of certain findings. Known cases of this have been discussed throughout this article; however, cultural variations in neural processes are only beginning to be studied. This may be, in part, for practical reasons. Until recently, conducting cross-cultural research posed many logistical challenges, including inadequate resources as well as communication barriers, and often required expensive travel. However, international collaborations are occurring with increasing frequency, and communications tools such as Skype help close the distance. Future research should focus on representing populations outside North America and Western Europe, with an eye toward understanding human diversity in the social brain.

Closing Remarks

Despite the challenges it may face, developmental social neuroscience provides numerous insights into the neurobiological processes that facilitate social interaction. These processes are fundamental for interpreting the actions, intentions, and emotions of other people. They situate our brain within the social world, allowing us to learn from others and others to learn from us. These are fundamental mechanisms underlying some of the most precious moments in life. Through understanding them, we come closer to understanding ourselves and our species.

See also: Breast-feeding: Effects on Cognitive and Neural Development; Child Care and Child Development; Early Emotional Development and Cultural Variability; Imitation During Infancy and Early Childhood; Infant Gaze Behaviors; Mirror Neurons, Theory of; Neonatal Social Isolation: Effects on Neural Function; Self and Identity Development During Adolescence across Cultures; Self in Culture: Early Development; Self-Development in Childhood and Adolescence; Social and Emotional Development in the Context of the Family; Stress in Adolescence: Effects on Development; Stress in Childhood: Effects on Development; Stress in Infancy and Early Childhood: Effects on Development.

Bibliography

Ames, D.L., Fiske, S.T., 2010. Cultural neuroscience. Asian Journal of Social Psychology 13 (2), 72–82.
Black, J.E., 1998. How a child builds its brain: some lessons from animal studies of neural plasticity. Preventive Medicine 27 (2), 168–171.
Blakemore, S.J., 2008. The social brain in adolescence. Nature Reviews Neuroscience 9, 267–277.
Champagne, F.A., 2010. Epigenetic influence of social experiences across the lifespan. Developmental Psychobiology 52 (4), 1–13.
Chiao, J.Y., Blizinsky, K.D., 2010. Culture–gene coevolution of individualism–collectivism and the serotonin transporter gene. Proceedings of the Royal Society B: Biological Sciences 277 (1681), 529–537.
Chiao, J.Y., Harada, T., Komeda, H., Li, Z., Mano, Y., Saito, D., Parrish, T.B., Sadato, N., Iidaka, T., 2009. Dynamic cultural influences on neural representations of the self. Journal of Cognitive Neuroscience 22 (1), 1–11.
Clarke, A., Chambers, D., 1998. The extended mind. Analysis 58 (1), 7–19.
Cole, S.W., 2009. Social regulation of human gene expression. Current Directions in Psychological Science 18 (3), 132–137.
Crafa, D., Nagel, S., In press. Traces of culture: the feedback loop between brain, behavior, and disorder. Transcultural Psychiatry.
Csibra, G., Davis, G., Spratling, M.W., Johnson, M.H., 2000. Gamma oscillations and object processing in the infant brain. Science 290 (5496), 1582–1585.
Csibra, G., Gergely, G., 2006. Social learning and social cognition: the case for pedagogy. In: Munakata, Y., Johnson, M.H. (Eds.), Processes of Change in Brain and Cognitive Development. Attention and Performance XXI. Oxford University Press, Oxford, pp. 249–274.
Curley, J.P., Jensen, C.L., Mashoodh, R., Champagne, F.A., 2011. Social influences on neurobiology and behavior: epigenetic effects during development. Psychoneuroendocrinology 36 (3), 352–371.
Cushing, B.S., Kramer, K.M., 2005. Mechanisms underlying epigenetic effects of early social experience: the role of neuropeptides and steroids. Neuroscience & Biobehavioral Reviews 29 (7), 1089–1105.
Dapretto, M., Davies, M.S., Pfeifer, J.H., Scott, A.A., Sigman, M., Bookheimer, S.Y., et al., 2006. Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. Nature Neuroscience 9 (1), 28–30.
Eisenberger, N.I., Lieberman, M.D., Williams, K.D., 2003. Does rejection hurt? An FMRI study of social exclusion. Science 302 (5643), 290–292.
Eisenberger, N.I., Way, B.M., Taylor, S.E., Welch, W.T., Lieberman, M.D., 2007. Understanding genetic risk for aggression: clues from the brain's response to social exclusion. Biological Psychiatry 61 (9), 1100–1108.
Etkin, A., Egner, T., Kalisch, R., 2011. Emotional processing in anterior cingulate and medial prefrontal cortex. Trends in Cognitive Sciences 15 (2), 85–93.
Farroni, T., Johnson, M.H., Menon, E., Zulian, L., Faraguna, D., Csibra, G., 2005. Newborns' preference for face-relevant stimuli: effects of contrast polarity. Proceedings of the National Academy of Sciences of the United States of America 102 (47), 17245–17250.
Frith, C.D., 2007. The social brain? Philosophical Transactions of the Royal Society B 362 (1480), 671–678.
Frith, C.D., Frith, U., 2007. Social cognition in humans. Current Biology 17 (16), 724–732.
Giorgio, A., Watkins, K.E., Chadwick, M., James, S., Winmill, L., Douaud, G., et al., 2010. Longitudinal changes in grey and white matter during adolescence. NeuroImage 49 (1), 94–103.
Grossmann, T., Johnson, M.H., 2007. The development of the social brain in human infancy. European Journal of Neuroscience 25 (4), 909–919.
Han, S., Northoff, G., 2008. Culture-sensitive neural substrates of human cognition: a transcultural neuroimaging approach. Nature Reviews Neuroscience 646 (9), 646–654.
Harada, T., Li, Z., Chiao, J.Y., 2010. Differential dorsal and ventral medial prefrontal representations of the implicit self modulated by individualism and collectivism: an fMRI study. Society for Neuroscience 5 (3), 257–271.
Hildyard, K.L., Wolfe, D.A., 2002. Child neglect: developmental issues and outcomes. Child Abuse & Neglect 26 (6–7), 679–695.
Keuken, M.C., Hardie, A., Dorn, B.T., Dev, S., Paulus, M.P., Jonas, K.J., et al., 2011. The role of the left inferior frontal gyrus in social perception: an rTMS study. Brain Research 1383, 196–205.

Lamm, C., Singer, T., 2010. The role of anterior insular cortex in social emotions. Brain Structure and Function 214 (5–6), 579–591.

Lane, R.D., Reiman, E.M., Axelrod, B., Yun, L.S., Holmes, A., Schwartz, G.E., 1998. Neural correlates of levels of emotional awareness: evidence of an interaction between emotion and attention in the anterior cingulate cortex. Journal of Cognitive Neuroscience 10 (4), 525–535.

Li, Y.H., Tottenham, N., 2011. Seeing yourself helps you see others. Emotion 11 (5), 1235–1241.

Lin, S.H., Cermak, S., Coster, W.J., Miller, L., 2005. The relation between length of institutionalization and sensory integration in children adopted from Eastern Europe. American Journal of Occupational Therapy 59 (2), 139–147.

Lombardo, M.V., Chakrabarti, B., Bullmore, E.T., Wheelwright, S.J., Sadek, S.A., Suckling, J., MRC AIMS Consortium, Baron-Cohen, S., 2010. Shared neural circuits for mentalizing about the self and others. Journal of Cognitive Neuroscience 22 (7), 1623–1635.

Meltzoff, A.N., 2007. 'Like me': a foundation for social cognition. Developmental Science 10, 126–134.

Moran, J.M., Macrae, C.N., Heatherton, T.F., Wyland, C.L., Kelley, W.M., 2006. Neuroanatomical evidence for distinct cognitive and affective components of self. Journal of Cognitive Neuroscience 18 (9), 1586–1594.

Paulesu, E., McCrory, E., Fazio, F., Menoncello, L., Brunswick, N., Cappa, S.F., Cotelli, M., Cossu, G., Corte, F., Lorusso, M., Pesenti, S., Gallagher, A., Perani, D., Price, C., Frith, C.D., Frith, U., 2000. A cultural effect on brain function. Nature Neuroscience 3, 91–96.

Perry, B.D., 2002. Childhood experience and the expression of genetic potential: what childhood neglect tells us about nature and nurture. Brain Mind 3 (1), 79–100.

Perry, B.D., 2006. Applying principles of neurodevelopment to clinical work with maltreated and traumatized children: the neurosequential model of therapeutics. In: Webb, N.B. (Ed.), Working with Traumatized Youth in Child Welfare. The Guilford Press, New York, pp. 27–52.

Perry, B.D., Pollard, D., 1997. Altered Brain Development Following Global Neglect in Early Childhood. Society For Neuroscience: Proceedings from Annual Meeting, New Orleans.

Phillips, M.L., Young, A.W., Senior, C., Brammer, M., Andrew, C., Calder, A.J., et al., 1997. A specific neural substrate for perceiving facial expressions of disgust. Nature 389 (6650), 495–498.

Rakic, P., Lombroso, P., 1998. Development of the cerebral cortex: forming the cortical structure. Journal of the American Academy of Child & Adolescent Psychiatry 7 (1), 116–117.

Rees, C., 2008. The influence of emotional neglect on development. Paediatrics & Child Health 18 (12), 527–534.

Rule, N.O., Freeman, J.B., Ambady, N., 2013. Culture in social neuroscience: a review. Society for Neuroscience 8 (1), 3–10.

Saxe, R., 2006. Uniquely human social cognition. Current Opinion in Neurobiology 16, 235–239.

Saygin, A.P., 2007. Superior temporal and premotor brain areas necessary for biological motion perception. Brain 130 (9), 2452–2461.

Schultz, R.T., 2005. Developmental deficits in social perception in autism: the role of the amygdala and fusiform face area. International Journal of Developmental Neuroscience 23 (2–3), 125–141.

Shonkoff, J.P., Phillips, D.A., 2000. From Neurons to Neighborhoods: The Science of Early Childhood Development. National Academy Press, Washington, DC.

Turner, R., 2002. Culture and the human brain. Anthropology Humanities 26 (2), 167–172.

Vander Wyk, B.C., Hudac, C.M., Carter, E.J., Sobel, D.M., Pelphrey, K.A., 2009. Action understanding in the superior temporal sulcus region. Psychological Science 20 (6), 771–777.

Watts-English, T., Fortson, B.L., Gibler, N., Hooper, S.R., De Bellis, M.D., 2006. The psychobiology of maltreatment in childhood. Journal of Social Issues 62 (4), 717–736.

Webb, S.J., Monk, C.S., Nelson, C.A., 2001. Mechanisms of postnatal neurobiological development: implications for human development. Developmental Neuropsychology 19 (2), 147–171.

Weiss, S.J., Wilson, P., Morrison, D., 2004. Maternal tactile stimulation and the neurodevelopment of low birth weight infants. Infancy 5 (1), 85–107.

Wilson, D.H., Chang, A.E., 1974. Bilateral anterior cingulectomy for the relief of intractable pain. Connia Neurology 36, 61–68.

Zatorre, R.J., Fields, R.D., Johansen-Berg, H., 2012. Plasticity in gray and white: neuroimaging changes in brain structure during learning. Nature Neuroscience 15, 528–536.

Neural Foundation of Morality

Roland Zahn, Translational Congnitive Neuroscience of Affective Disorders Laboratory, The University of Manchester, School of Psychological Sciences, Manchester, UK.
Ricardo de Oliveira-Souza and Jorge Moll, D'Or Institute for Research and Education (IDOR), Rio de Janeiro, Brazil

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Moral behavior is one of the most sophisticated human abilities. Many social species behave altruistically toward their kin, but humans are unique in their ability to serve complex and changing societal needs. Cognitive neuroscience has started to elucidate specific brain mechanisms underpinning moral behavior, emotion, and motivation, emphasizing that these ingredients are also germane to human biology, rather than pure societal artifacts. The brain is where psychosocial learning and biology meet to produce the rich individual variability in moral behavior. This article discusses how cognitive neuroscience improves the understanding of this variability and associated suffering in neuropsychiatric conditions.

Introduction

The terms moral and ethical derive from their respective Latin and Greek roots (*moralis* and *ethikos*). Literally, they would probably be best translated as in accordance with societal customs. Based on this, the authors have suggested an operational definition of moral behavior as behavior that is in accordance with other people's needs or sociocultural norms (Zahn et al., 2011). Moral behavior requires at least two components: (1) knowing about sociocultural norms and the needs of others and (2) being motivated to act on this knowledge (Zahn et al., 2011). Distinguishing moral from selfish motivations in real life is often difficult, and cooperation in societies is usually not sustained on the basis of moral motivations alone. Instead, such morally motivated behaviors intermingle with self-serving ones that sustain individual survival in a complex interplay that often helps stabilize productive relations in social groups. Such behaviors essentially include codes of reciprocity (described in the economic and evolutionary literature, as a direct or indirect gain for somebody for cooperating in a society (Nowak and Sigmund, 2005)). Societies have therefore developed external reward and punishment systems, nowadays enshrined in law, work ethics, and other organizational regulations, to enforce behavior in accordance with communal welfare.

Computational models have shown that individual variability in moral behavior may have sustained cooperation in small social groups that preceded our modern societies (Boyd et al., 2003; Boyd and Richerson, 2009). These studies suggest that in the absence of law enforcement systems, cooperation can only be sustained if at least a fraction of individuals are motivated to risk substantial resources in order to punish other individuals who behave immorally. Were these individuals perfectly economically 'rational' and purely selfishly motivated, there would be no incentive to do so in that they would usually lose more than they would win. Such costly behaviors have been referred to as 'altruistic punishment.' In these models, selfishly motivated individuals, known as 'defectors,' are kept in check by the 'altruistic punishers.' A third type of agents will contribute own resources to the community ('contributors'). The debate in evolutionary sciences is how the genes of people who sacrifice their own resources could have survived evolutionary pressures. This is easy to explain in the small groups from which we are thought to originate, based on kinship-based genetic relatedness mechanisms. Thus, everything that was beneficial to the survival of the group would benefit the duplication of the gene pool, relative to individuals belonging to other small-scale societies. However, many evolutionary theorists need to explain sacrifices of one's own resources in large-scale modern societies by nonmoral motivations, reverting to at least indirect benefits of cooperation for oneself or one's family (West et al., 2007). This is because they disagree with the argument of group selection benefits, which is supported by Boyd et al., whose model proposes that cooperative societies enhance the survival chances of their genes in competition with noncooperative ones even in the absence of intragroup genetic relatedness. This view is also adopted by researchers performing neuroeconomic experiments to demonstrate altruistic behavior among strangers in the absence of direct or indirect benefits (Gintis et al., 2008).

While it is beyond the scope of this chapter to delve into this fascinating debate further, important implications for the neuroscience of morality can be derived:

1. If gene selection has promoted moral motivations within societies, then these genes must have influenced brain mechanisms that support moral behaviors, because biological influences on human behavior are mediated by the brain.
2. If this was indeed the case, then those brain mechanisms must have also allowed for adaptation through cultural learning that allowed individuals to adapt to changing sociocultural environments, because the same behavior can be adaptive and maladaptive depending on cultural and situational contexts (e.g., 'killing in war' and 'killing in peacetime').
3. If individual variability in moral motivations was necessary to provide group survival advantages, then one should be able to identify different types of brains supporting different inclinations toward different balances of moral and selfish

motivations that are further shaped by cultural learning. This would then allow for individuals ranging in their motivational inclinations from selfish to different degrees of preparedness to moral punishment or cooperation.

This article will provide an overview of the human brain mechanisms that appear to be selective for moral as compared with selfish motivations (Section Cognitive-Anatomical Basis of Moral Motivation) thereby addressing the first question. The authors further devote a section to brain mechanisms underpinning social knowledge to address the second question (Section Cognitive-Anatomical Basis of Social Knowledge). They take the perspective here, that social knowledge can be used for moral as well as selfish purposes (e.g., manipulation and reputation gain) and is therefore not selective for the moral domain. To address the third question, the authors discuss evidence supporting individual differences in the neuroanatomy of moral motivations and relate this to the question of how these may contribute to mental health and suffering (Section Implications for Differential Vulnerability in Psychiatry). A large proportion of us suffer from psychiatric disorders at least once during our lifetime, for example, depression is estimated to affect 151 million people around the world (Funk, 2010). Our risk is significantly influenced by genetic factors as shown by studies comparing mono- and dizygotic twins (Kendler, 2006). Given the large number of affected people, it is tempting to speculate whether some psychiatric risk genes have carried advantages for psychosocial functioning and survival in order to be selected for by evolutionary pressures. If it is the variability between individuals within societies that makes these function most adaptively, this could explain the variability in how people navigate in the social world. This variability could then explain vulnerability to different types of mental health problems depending on the demands of one's social environment. Some evolutionary thinkers support the idea that risk genes for psychiatric disorders could have adaptive functions through group selection (Wilson and Wilson, 2007). Others, in contrast, have argued that the polygenetic nature of these risk genes and their complex interactions with environmental factors can explain the high frequency of psychiatric disorders, despite their proposed evolutionary maladaptiveness (Keller and Miller, 2006).

Brain mechanisms that influence social cognition and behavior are based on macroanatomical, microstructural, and neurochemical components that are employed to learn and adapt to complex information in our psychosocial environment. The advent of moral neuroscience was the discovery that disruptions of the macroanatomical structure of brains could lead to marked changes in moral behavior (Section Moral Behavior in Patients with Macroanatomical Brain Lesions). Microstructural changes were documented in postmortem studies of conditions such as autism (Casanova et al., 2002), a disorder of social cognition in general rather than moral cognition more specifically. Due to the difficulty in obtaining such data, however, there is not enough evidence on microstructural changes tied to moral behavior to be included here. Functional magnetic resonance imaging (fMRI) has been employed since the 2000s to study moral cognition and emotion. While this has provided us with a rich set of data, its interpretation needs cautious validation from lesion studies. This is because the results of functional activation-based neuroimaging depend as much on the psychological task conditions as they do on brain structure and function. fMRI, however, provides a unique opportunity to measure psychobiological interaction thereby capturing the results of brain adaptations to psychosocial learning (Section Functional Neuroimaging of Moral Cognition and Emotion). Electrophysiological properties of the brain are probably closest to human experience and novel techniques of analyzing electroencephalography and magnetic encephalography have opened up great possibilities. The evidence base using these techniques in moral cognition is still quite narrow and will therefore not be reviewed here. Transcranial magnetic stimulation is a way to simulate brain lesions and has been employed in moral neuroscience, but its use is restricted to areas near the skull surface and thereby does not allow for a systematic study of several of the brain regions known to be relevant from patient lesion studies. Neurochemical influences on moral cognition and emotion have been reviewed elsewhere and are beyond the scope of this article (Moll and Schulkin, 2009). Finally, the authors will not review correlations between gene polymorphisms with neuroimaging data. This is because the focus of this article is on understanding the neural rather than the genetic prerequisites of moral behavior. Section Cognitive-Anatomical Models of Moral Cognition and Emotion briefly summarizes the conflicting neurocognitive models of moral cognition and emotion and lays the ground for future directions for research.

Moral Behavior in Patients with Macroanatomical Brain Lesions

In 1888, Leonore Welt, a Swiss physician, reviewed a series of patients with damage to the frontal cortex, who showed marked changes in moral character as judged by their jocular childish or aggressive behavior (Welt, 1888). Interestingly, their general intelligence was preserved. Phineas Gage, the now famous US American patient (Damasio et al., 1994), with an iron-bar injury to his medial frontal lobes was one of the cases reviewed by Dr Welt. Based on postmortem neuropathology and injury pattern, she concluded that the most consistent structure damaged in all cases was the right medial orbitofrontal cortex. While this was probably the first systematic evidence for a causal role of specific brain structures for enabling moral behavior, it also marked the beginning of a controversy about the function of the frontal lobes. This is because systematic studies of patients with head injuries during the First World War confirmed earlier case reports of some patients with frontal lesions and no behavioral change (Teuber, 2009). This lack of impairment in some patients could of course be explained by insufficient characterization of the lesion or lack of sensitivity of clinical observation. However, it could also point to individual variability in cognitive-anatomical architecture leading some patients to be more sensitive than others to frontal brain lesions.

In the 1980s, brain computed tomography became available in clinical settings and led to a renaissance of neuropsychological case studies. Eslinger & Damasio sparked new

interest in the neural basis of behavior counter to moral rules by describing EVR, a patient with ventral frontal lobe damage due to tumor resection (Eslinger and Damasio, 1985). He exhibited difficulty in making daily life decisions and made choices that led to a ruin of his financial fortunes as well as his marriage. At the same time, he showed normal functioning on classical tests of executive functioning such as the Wisconsin Card Sorting test, which is usually impaired in patients with left dorsolateral frontal lesions, which was not damaged in EVR.

In the same decade, a neurodegenerative disease, originally described by Arnold Pick in the late nineteenth century, became recognized as a relatively common form of early-onset dementia that could be reliably diagnosed before death (Snowden et al., 2001). These patients exhibited varying degrees of frontal and anterior temporal atrophy captured in the term 'frontotemporal lobar degeneration.' This led to cumulating evidence on inappropriate social behavior in these patients (Bozeat et al., 2000) that was independently associated with right anterior temporal and ventromedial frontal atrophy (Liu et al., 2004). Patients with right anterior temporal lesions behave as inappropriately as patients with ventral frontal atrophy. For example, in their clinic, the authors saw a patient (further referred to as 'the singing lady') with isolated right anterior temporal lobe atrophy, who lost her job, because of kissing and singing to customers at work, who also sang happy birthday to her parents on the cemetery.

Studies in patients with brain lesions were often hampered by the lack of neuropsychological tasks that could probe cognitive functions relevant to the observed behavioral changes and largely relied on carer reports. Loss of guilt was reported by carers of patients with ventromedial (prefrontal cortex) lesions (Koenigs et al., 2007) and neurodegeneration in this area, which included the subgenual cingulate cortex, was associated with loss of interpersonal warmth as judged by carers (Sollberger et al., 2009). Inappropriate behavior in these patients often goes against societal needs or needs of others and therefore clearly has moral implications. However, the disruption of which cognitive and emotional components leads to this behavioral change is disputable. Patients with ventral frontal lesions were found to show intact social knowledge when asked about their behavior (Saver and Damasio, 1991; Eslinger and Damasio, 1985). Similarly, Mendez et al. found in their patients with frontotemporal lobar degeneration that knowledge of moral rules was intact when assessed on a questionnaire, but nevertheless patients failed to act on these rules (Mendez et al., 2000). This may lead to the hypothesis that moral motivation and knowledge may be at least partly dissociable. The hypothetical causes of socially inappropriate behavior are

1. Loss of (access to) knowledge of appropriate social behavior (Zahn et al., 2009b).
2. Loss of the moral motivation to act upon intact social knowledge (Moll et al., 2005b).
3. Loss of the ability to suppress urges/drives (Brutkowski, 1965) with intact social knowledge of their inappropriateness and intact moral motivation to act upon this knowledge.

In order to test these different possibilities, one needs to probe moral motivations experimentally while controlling for social knowledge and vice versa. Two studies in patients with frontotemporal lobar degeneration have used this approach. One study showed that patients with damage of the right anterior temporal lobe displayed selective deficits on abstract knowledge of social behavior (i.e., social conceptual knowledge, e.g., what it means to act 'tactfully') relative to nonsocial conceptual knowledge (e.g., what it means to be 'nutritious,' Zahn et al., 2009b). Furthermore, selective deficits on social concepts were associated with inappropriate social behavior. The social concept discrimination task used to probe this particular type of social knowledge did not require moral motivations and controlled for the effects of positive or negative emotional valence. Interestingly, the right anterior temporal lobe atrophy patient, 'the singing lady,' exhibited a selective impairment on social concepts with relatively intact nonsocial concepts when using this test (Zahn et al., 2012).

Another study showed that patients with frontotemporal lobar degeneration and relatively stronger involvement of the septal area were selectively impaired on a task probing guilt and pity, relative to other negative emotions such as embarrassment, disgust, and anger (Moll et al., 2011). The task controlled for the degree of social knowledge required by using similar statements to describe social situations in which one usually feels a particular moral feeling (e.g., "Your mother calls you one night, telling you she was not feeling well. You did not take her seriously and the next day she died" associated with guilt in 80% of a normative sample). Neurodegeneration of the frontopolar cortex was associated with selective loss of guilt, pity, and embarrassment compared with disgust and anger.

These studies, together with evidence from fMRI reviewed below, indicate that moral motivations and social conceptual knowledge are associated with dissociable brain regions and can thus independently contribute to inappropriate social behavior. The third and most popular explanation for such behavior, however, is that these patients know about other people's needs and are motivated to act on them, but that they have lost the ability to suppress selfish urges in subcortical regions that counteract their ability to act morally.

This frontal-subcortical suppression view probably became influential because of the way conditioning experiments in nonhuman animals were interpreted (Brutkowski, 1965). To the authors' knowledge, there is, however, no anatomical evidence supporting strictly unidirectional, inhibitory connections from frontal cortical to subcortical areas, which would support such a suppression mechanism. Furthermore, there is no evidence that selfish motivations are more dependent on subcortical than cortical mechanisms. On the contrary, there is evidence from case reports in patients with subcortical lesions that these may be associated with severe antisocial behavior. Subcortical lesion cases included hypothalamus, septal area, amygdala, and basal ganglia (Moll et al., 2003), but it is unclear how regularly these associations can be found. A more plausible version of the frontal-subcortical suppression model is mentioned below in which the frontal cortex is thought to control (Miller and Cohen, 2001) rather than suppress other areas of the brain.

In principle, one must consider another explanation for 'immoral' behavior after ventral frontal lesions, namely, that moral behavior requires more complex cognition and will thereby be nonspecifically impaired in large brain lesions. If this was the case, patients with similar degrees of brain damage and impairment should also show similar degrees of deviations from moral behavior. Interestingly, this is clearly not the case in that patients with Alzheimer's dementia in similar mild to moderate stages of the disease do show intact interpersonal behavior in the overwhelming majority of cases (Bozeat et al., 2000). Alzheimer's dementia typically leads to neurodegeneration of medial temporal, as well as medial and lateral parietal brain regions in these stages (Herholz, 2003). This means that moral behavior in a daily life context appears to be independent of these brain regions.

Functional Neuroimaging of Moral Cognition and Emotion

Introduction

When describing the function of a brain area, one can either describe (1) which type of task it is involved in carrying out (i.e., process-dependent description), (2) which information content it stores (i.e., content-dependent description), or (3) in which format it stores information (i.e., format-dependent description). The brain regions considered to be important for moral cognition and emotion were found to be involved in tasks and contents of different formats (e.g., visually presented pictures and written statements, reviewed in Moll et al., 2005b). However, because most people assume format independence of moral cognition and emotion, this was not rigorously investigated for most brain regions. Most fMRI studies of moral cognition and emotion have either taken a process- or a content-based approach to the design of their experiments. Describing a brain area as involved in a certain process or representation of a certain content is not very helpful in itself, unless one can make the case that there is some degree of selectivity of its involvement in that particular process or content compared with other processes or contents and other brain areas. This is because the less specifically a process or content is defined and tested, the more brain areas and cognitive components will be involved in it. So one could design experiments testing very broad functions, which will then involve a large set of brain areas, but this will not be very informative (i.e., add new information to what was known before). The authors have therefore focused on reviewing studies that chose experimental and control conditions in such a way that a high degree of selectivity of a given brain region for the function of interest could be derived. Claims of process selectivity are often not supported by generalizing across different types of contents and likewise content selectivity is often not supported by being tested across different tasks (processes). Future studies are needed to fill this gap. The broader involvement of a network of frontal, subcortical, and anterior temporal regions in moral cognition and emotion tasks, however, is generally agreed upon when considering the lesion and functional imaging literature (Moll et al., 2005b, 2008). In fMRI studies, posterior superior temporal, temporoparietal junction, and precuneus are also regularly activated in moral cognition and emotion tasks. The fact that patients with posterior cortical atrophy and Alzheimer's disease do not usually show grossly abnormal social behavior (see above) implies that the involvement of these regions must be more subtle and it is unlikely that they will show the same degree of selectivity for moral cognition and emotion.

Fronto-temporo-subcortical networks activated in response to morally relevant stimuli are highly overlapping with brain responses to other social cognition tasks (e.g., when having to think about mental states compared with physical inferences ('theory of mind') (Frith, 2007)). It is therefore important to define what distinguishes moral from social contents or processes. The authors define in this article social behavior as any type of behavior that has potential consequences for others. Thereby, any type of social behavior will have moral implications, subtle as they might be, in that it affects other individuals or society in general. Lonesome behaviors such as paper recycling at midnight when nobody watches are very good examples of social behaviors with moral implications in that the action of recycling motivating the action is driven by societal needs. When observing social behavior from the outside, there is therefore no way of distinguishing social from moral behavior. Only when hypothesizing about the different motivations driving these behaviors, it is that we are able to distinguish moral from selfish social behaviors. For example, unobserved behavior that supports societal needs such as a man recycling paper at night is more likely to be morally motivated than the same behavior with relevant others watching. For example, the recycling man could be motivated by trying to charm an ecologically conscious women observer whom he is sexually interested in. Thus, according to our definition here, what is special about moral relative to social contents and processes is their specificity for morally relative to selfishly motivated behavior. Selfish social behaviors require the same amount of mental state inferences as morally motivated social behaviors if they require manipulating other people. For both cases of morally or selfishly motivated behaviors, however, the level of engagement of social cognition mechanisms will depend on the level of the complexity of predictions about the effects of one's behavior on others and associated social interactions. Nevertheless, there is quite a large repertoire of selfish behaviors that are nonsocial and thus require much less complex cognition, because social interaction is generally more complex and contextualized. This may explain why impairments of general social cognition appear to impair moral behavior more than selfish behavior in general. One example is the right anterior temporal lobe atrophy patient, 'the singing lady,' mentioned before, who showed a selective impairment on social conceptual knowledge, which impaired her understanding of social situations and thereby led her to violate social rules with a relatively intact ability to help herself to food, play Sudoku puzzles, and appear quite happy in her mood. When reviewing the functional imaging literature, the authors will therefore include brain regions that show some selectivity for social compared with nonsocial cognition even if they are most probably not special for moral functions proper. The authors will not, however, focus on brain regions such as the left lateral parietal and left dorsolateral prefrontal cortices whose

specialization for social cognition is unlikely given the intact clinical appearance of social behavior in patients with focal lesions to these areas, such as aphasic stroke patients (Zahn et al., 2004).

Content-Based Subdivisions

At the broadest level, morally relevant contents were compared with nonmorally relevant contents. This included verbal statements or pictures rated by participants according to their moral relevance before the fMRI experiment. Obviously, it is important to control for the emotional relevance in those experiments, which was done in at least some of the studies. Taken together, frontopolar, ventral frontal, subcortical, and anterior and posterior temporal sulcus were most consistently activated when comparing morally relevant with morally irrelevant materials (reviewed in Moll et al., 2005b). The difficulty lies in identifying the functional contribution of each of these areas in that one could argue that socially relevant materials have been found to engage similar brain areas and it is unclear to what extent any of these areas is selective for moral cognition and emotion. Should they be involved in more general social functions, it is also unclear which exact contribution they are making to moral cognition and emotion. It is a matter of debate, how to best subdivide different categories of morally relevant contents. One approach is to distinguish between harm-, disgust-, and dishonesty-related (Parkinson et al., 2011) or justice- and care-related (Ochsner et al., 2004) moral contents and these were shown to be associated with differential brain activation patterns. The difficulty, however, is to demonstrate that complex moral scenarios used in these studies differed mainly with regard to these categorical distinctions, or whether there were associated differences in other nonmeasured variables.

The eighteenth century German philosopher Immanuel Kant distinguished the ability to judge what is morally right and wrong ("*principium diiudicationis*") from the motivation ("*principium motivationis*") to act morally (Kant, 1786/1903). He thereby anticipated a dissociation later found in some patients with ventral frontal brain lesions whose knowledge of how they should behave seems to be better preserved than their motivation to act accordingly (Eslinger and Damasio, 1985). The reverse dissociation can be observed as well. For example, the patient with a focal right anterior temporal lobe atrophy ('the singing lady') was selectively impaired on conceptual knowledge of social behavior as described before. Interestingly, she was among a minority of patients with frontotemporal lobar degeneration in which carers agreed that she appeared able to feel guilty. The authors probed this experimentally by giving her 22 sentences where she was described acting badly toward her best friend (negative self-agency condition: e.g., 'You act in a stingy way toward Jane'), in the negative other-agency condition, the authors just reversed the agents and recipients thereby keeping the social knowledge required to understand the sentences equal (e.g., 'Jane acts in a stingy way toward you'). Often, the patient did not understand what the concept (e.g., 'stingy') meant, but she was generally able to say whether it was a good or a bad thing to do (the most general form of social knowledge). After each sentence she was asked to say out aloud the feeling she would feel in this situation and was given a selection from guilt, anger, pride, gratitude, and none/other. This allowed her to respond with guilt in 73% of the negative self-agency trials, rarely responding with anger (5%) in this condition. In the negative other-agency condition, however, she showed the reverse pattern (68% of anger and 5% of guilt). It thus appears that if she was given the right cues, she was able to experience guilt, but because of her impaired social conceptual understanding, she would often miss the cues of the situation in real life. The sections below review fMRI studies that have contributed to a better understanding of the interlocking but partly independent neural systems underpinning social knowledge and moral motivation.

Cognitive-Anatomical Basis of Social Knowledge

Social knowledge has been defined as knowledge of one's own and other people's minds (Adolphs, 2009). Because 'mind' or 'mental states' are hard to break down neuropsychologically, the authors prefer defining social knowledge as denoting nonepisodic (i.e., semantic) knowledge of social sensory properties and social behavior (i.e., functions, see **Figure 1**).

There is growing consensus among neuroscientists investigating semantic knowledge of objects, that this requires not only the representation of modality-specific sensory and motor features of objects (e.g., 'the color of a robin,' 'the feel of its feathers'), but also a more abstract representation of the core aspects of what one means by, e.g., 'robin.' These core aspects are particularly important for understanding the behavior or functions of the things surrounding us. Based on evidence

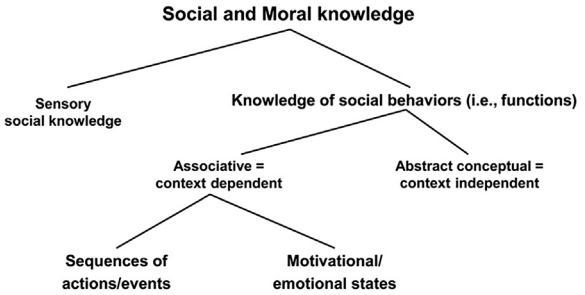

Figure 1 The hypothesized neurocognitive components of social and moral knowledge are illustrated (modified from Zahn, R., Oliveira-Souza, R., Moll, J., 2011. The neuroscience of moral cognition and emotion. In: Decety, J., Cacioppo, J. (Eds.), The Handbook of Social Neuroscience. Oxford University Press). The authors refer to non-episodic long-term memory here and distinguish between sensory social knowledge presumably represented in the posterior superior temporal cortex and knowledge of social behaviors (i.e., functions). The latter is subdivided into two anatomically separable systems: (1) abstract conceptual (i.e., context-independent) knowledge of social behavior represented in the anterior temporal lobes (aTL) and (2) associative (i.e., context-dependent) knowledge of social behavior represented in frontosubcortical networks. Associative knowledge of social behavior is subdivided into a component representing knowledge of sequences of social actions (events) within frontopolar cortex (FPC) and ventral prefrontal cortex (PFC) (representing associations of action/event sequences with motivational/emotional states) and a subcortical mesolimbic component representing 'free-floating' motivational/emotional states (see also **Figure 4**).

from fMRI, repetitive transcranial magnetic stimulation, and a neurodegenerative disease of the anterior temporal lobes called semantic dementia (a subtype of frontotemporal lobar degeneration), the anterior temporal lobe has been identified as hosting this core representation of concepts that is independent of the context of the type of task, modality, or arbitrary associations (Ralph and Patterson, 2008). For example, 'greeting colleagues' is associated with 'cheek kissing' in Brazil, but people in Britain would not strongly associate the two at least in the context of British work life. Such context-dependent associations are important for context-appropriate social behavior.

Wood and Grafman provide an overview on evidence linking the frontal cortex with one particular type of associative context-dependent knowledge, namely, the knowledge of sequences of events and actions (Wood and Grafman, 2003). More specifically, they hypothesized that the ventral frontal cortex represents socially relevant knowledge of sequences of events and actions. This knowledge about the context of sequences of actions enables us to appropriately choose an action in one sequential context (e.g., 'cheek kissing of colleagues in Brazil'), but refrain from it in another (e.g., 'Britain'). This context-dependent appropriateness of many of our actions could explain why patients with ventral frontal lesions, one of the earliest and most consistently affected areas in frontotemporal lobar degeneration, do something that is inappropriate in a particular context, but not necessarily totally out of character. For example, the authors saw an 80-year-old lady with frontotemporal lobar degeneration from Manchester, who patted bottoms of men as soon as they bent down. This was despite not feeling any sexual desire. When the men turned around, she said: "Oh, I am sorry, I was just checking where your wallet was." Interestingly, her husband told us that she was patting other men's bottoms at parties when younger, but now this behavior had generalized to nonparty contexts.

fMRI studies have confirmed the hypothesis that regions within the anterior temporal lobes are selective for social versus nonsocial concepts and that this area stores these concepts in a context-independent way. Right superior anterior temporal activations were found irrespective of the context of emotional valence (negative, positive (Zahn et al., 2007, Zahn et al., 2009c)), agency (self or other (Zahn et al., 2009c)), and task (semantic (Zahn et al., 2007) or emotional judgment (Zahn et al., 2009c)). The localization of activations concurred with the lesion evidence reported above. Another study found a more middle anterior temporal area on the left side, which was shown to be more active for social concepts relative to nonsocial concepts as well as overlapping with a mentalizing task (Ross and Olson, 2010). Thus, anterior temporal lobe activations during mentalizing tasks (Frith, 2007) could be explained by accessing social conceptual knowledge to interpret the intentions of other people. Subsequently, the right superior anterior temporal lobe localization for social compared with nonsocial concepts has been independently confirmed (Skipper et al., 2011).

Direct evidence regarding knowledge of sequences of events and actions within the ventral frontal lobes from functional neuroimaging is still emerging. The medial frontopolar cortex (BA10), however, was activated when participants had to decide whether actions/events (e.g., 'order the food,' 'ask for the menu') were in the right or wrong sequential order. This activation was relative to a control condition of deciding whether the events were printed in the same or different font (Krueger et al., 2007). This task-dependent activation does not prove the existence of sequence-specific contents in this area, but an additional analysis in which the frequency with which events/actions occurred revealed that different subsectors of the medial frontopolar cortex were activated more strongly depending on frequency of occurrence in daily life. This is better explained by content-, rather than process-specific models, because response time did not explain these differences. Another study confirmed these results by showing that the medial frontopolar area was the only brain region to linearly increase activation with increasing complexity of event sequences (Krueger et al., 2009). One conundrum is how to explain relatively intact social knowledge in patients with ventral frontal lesions that often encompass the frontopolar cortex. A possible explanation is that the intact anterior temporal lobe representation of context-independent social knowledge may help these patients answer theoretical questions about appropriate behavior, but that in order to probe this knowledge of complex consequences of behavior as required for the execution of real actions, one will need to develop more refined ways of experimental testing.

Cognitive-Anatomical Basis of Moral Motivation

The term moral motivation was used by the eminent moral philosopher Francis Hutcheson (Bishop, 1996), who played an important role in highlighting the importance of what were later called 'moral sentiments' by Adam Smith (Smith, 1759/1966) – now mostly referred to as 'moral emotions.' Morally relevant materials will always be emotionally relevant (Haidt, 2001) unless one tries to take the perspective of an impartial spectator (Smith, 1759/1966). The same network of brain regions mentioned before to be selective for morally relevant compared with morally less relevant materials is therefore the starting point for the search of neural correlates of moral sentiments. Here, the authors summarize the evidence on those moral sentiments that were studied in more than two fMRI studies.

Guilt

The anticipation of guilt is thought to play an important role in preventing us from acting immorally and to motivate reparative action (Tangney et al., 2007; Eisenberg, 2000). The experience of guilt has been associated with activations of the dorsal cingulate in an early study (Shin et al., 2000), but this did not use an emotional control condition and does therefore not allow conclusions about guilt selectivity. Frontopolar activations were found most reproducibly for guilt while using different control conditions (other-critical feelings such as indignation (Moll et al., 2007; Zahn et al., 2009c); anger toward self (Kedia et al., 2008); embarrassment (Takahashi et al., 2004); regret with no consequences for others (Morey et al., 2012); and sadness (Basile et al., 2011)). The subgenual cingulate cortex (including the posteriorly adjacent septal area in some studies) was found to be selectively activated for guilt compared with indignation

toward others, but only when modeling individual variability in empathic concern (Zahn et al., 2009a), or guilt-proneness (Zahn et al., 2009c; Green et al., 2012). Interestingly, subgenual cingulate activations for guilt were independently reproduced (Morey et al., 2012; Basile et al., 2011). In contrast, another fMRI study found right lateral orbitofrontal and posterior dorsomedial prefrontal activations to be selective for guilt compared with shame and sadness (Wagner et al., 2011).

Pity (Compassion, Sympathy)

Emotional empathy and pity are very closely related and empathy has been extensively studied using neuroscience methods (Decety and Grezes, 2006). Empathy is usually conceived of as simulating other people's experiences, whereas sympathy requires an extra step of feeling for the other person (de Vignemont and Singer, 2006). The frontopolar cortex was activated most consistently for compassion versus a neutral control condition (Immordino-Yang et al., 2009), as well as compared with other-critical feelings (Moll et al., 2007) and compared with self-directed anger (Kedia et al., 2008). None of these studies investigated individual differences in the experience of pity and this may have led to a failure in detecting septal or subgenual cingulate regions engaged by empathic concern while feeling guilt (Zahn et al., 2009a). Interestingly, a study in adolescents and adults found septal/subgenual cingulate activations to be associated with how wrong people perceived depicted immoral actions, but their empathic sadness was associated with paracingulate cortex activation instead, among other regions (Decety et al., 2012).

Other-Critical (other-Blaming) Moral Sentiments

Moral disgust (highly associated with contempt in the authors' pilot studies) and moral anger (highly associated with indignation) are important to motivate altruistic punishment, which is of immense evolutionary importance as discussed. The most reproducibly activated region for these feelings is the lateral orbitofrontal cortex (other-critical vs prosocial moral sentiments (Moll et al., 2007); indignation/anger when being the victim of an action vs guilt when being the agent (Zahn et al., 2009c)). It remains unclear, however, whether regions within the lateral orbitofrontal cortex are selectively involved in other-critical feelings. This is because one study found moral disgust to activate this region more strongly than nonmoral disgust (Moll et al., 2005a), whereas another study found the lateral orbitofrontal cortex activated for both moral and nonmoral disgust compared with a neutral condition (Borg et al., 2008). Another study did report the frontopolar cortex for anger toward others when compared with anger toward self, but there was no comparison with prosocial moral sentiments to directly relate the results to the studies above (Kedia et al., 2008). Loss of anger and disgust were selectively associated with neurodegeneration of posterior dorsomedial prefrontal, dorsal anterior cingulate, and amygdala, when controlling for performance on pity, guilt, and embarrassment (Moll et al., 2011).

While the above studies tested the neural correlate of the subjective experience of specific moral sentiments, another approach has been to probe altruistic decisions compared with selfish decisions. If moral sentiments play a role in motivating these decisions, then one should find partly overlapping areas of brain activation.

Moral versus Selfish Decisions

Figure 2 describes how people were asked to donate to charitable organizations while being in the fMRI scanner (Moll et al., 2006). They were presented with different payoffs for the organization and for themselves and could choose whether they wanted to accept the payoff or not. This allowed the identification of specific neural signatures within the septal-subgenual cingulate region of deciding to donate to charity compared with keeping money for oneself. The frontopolar and anterior orbitofrontal cortex were associated with sacrificing one's money to either donate or punish an organization (i.e., costly decisions). Both monetary rewards and donation decisions were associated with ventral striatum activations. Interestingly, the lateral orbitofrontal cortex was associated with punishing organizations people disliked (e.g., the US National Rifle Association). The septal part of the nucleus accumbens, which is its medial aspect and often overlooked in imaging studies by describing this as ventral striatum activation, was subsequently detected for donation decisions in two fMRI studies (Hsu et al., 2008; Harbaugh et al., 2007). The subgenual cingulate region was also found to be more active in people with higher empathic concern while they made decisions to sacrifice some of their money to help others to avoid an electroshock (Feldman Hall et al., 2012).

Affiliative Motivational States

One basic ingredient for empathic moral sentiments such as pity, guilt, as well as moral decisions based on these feelings is the ability to feel attached to other people (Moll et al., 2008; Moll and Schulkin, 2009). Social attachment such as pair bonding and mother–offspring bonding has been linked to subcortical structures in nonhuman animals (Insel and Young, 2001; Stack et al., 2002). The core subcortical areas implicated in social attachment in nonhuman animals include the preoptic-anterior hypothalamic area, septal region, amygdala, ventral striatum, and the bed nucleus of stria terminalis (Stack et al., 2002). These regions are rich in neuropeptide and monoaminergic receptors (oxytocin, vasopressin, dopamine, and opioid) that have been shown to influence attachment-related behaviors. Human fMRI studies on maternal and romantic love found activation of these key subcortical regions (Bartels and Zeki, 2004; Aron et al., 2005; Swain et al., 2007). A recent study confirmed the prediction that there are regions selectively involved in affiliative emotional/motivational states within the hypothalamic-septal region irrespective of positive or negative emotional valence and irrespective of the context of agency (Moll et al., 2012). This is in keeping with the fronto-temporo-subcortical integration model of moral motivation as arising from decontextualized (i.e., free floating) motivational states in subcortical areas that are bound together with conceptual and sequential action information in anterior temporal and frontal areas to produce the complex experience of moral sentiments and

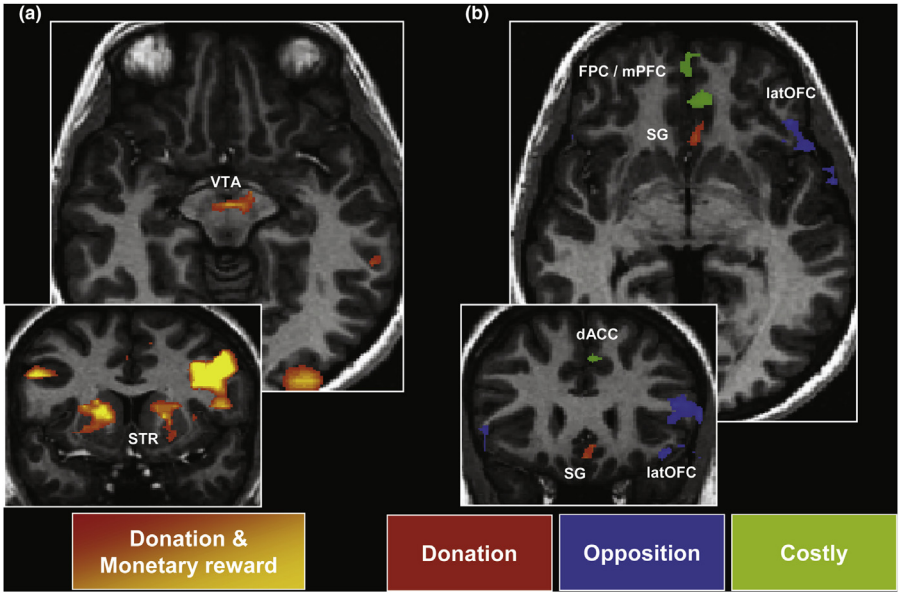

Figure 2 Brain regions activated when participants donated or were opposed to donate to charitable organizations during fMRI (Moll et al., 2006; figure adapted from Moll, J., Schulkin, J., 2009. Social attachment and aversion in human moral cognition. Neuroscience and Biobehavioral Reviews 33, 456–465). (a) Both pure monetary rewards and decisions to donate (with or without personal financial costs) activated the mesolimbic reward system, including the ventral tegmental area (VTA) and the ventral and dorsal striatum. (b) The septal-subgenual region (SG), however, was selectively activated by decisions to donate, as compared with pure monetary rewards (both by costly and noncostly decisions, conjunction analysis). The lateral orbitofrontal cortex (latOFC) was activated by decisions to oppose charities. This activation extended to the anterior insula and to the inferior dorsolateral prefrontal cortex (PFC) and was present both for costly and noncostly decisions (conjunction analysis). The frontopolar cortex (FPC) and ventral medial PFC (mPFC) were activated for costly decisions (when voluntarily sacrificing one's own money either to donate to a charity or to oppose it (conjunction analysis).

motivations (which always entail a goal rather than just the free-floating motivational state, see **Figure 4**).

Moral Values

Moral sentiments are often tied to abstract values described by social concepts such as 'honesty.' They thereby form moral values, which motivate behavior in a more stable way than the context-dependent rules about certain types of behaviors we feel guilty or proud of in particular situations. The right superior anterior temporal lobe was shown to be equally activated irrespective of the type of moral sentiment (guilt, pride, gratitude, and indignation) associated with sociomoral values, whereas different frontal-subcortical regions were associated with the different moral sentiments (**Figure 3**, (Zahn et al., 2009c)). This provides a neural architecture that allows social conceptual knowledge to be integrated with moral sentiments to give rise to the experience of moral values and thereby motivate behavior.

Process-Based Subdivisions

Frontopolar and anterior dorsolateral prefrontal cortices were consistently found for moral judgment/reasoning tasks (Greene et al., 2004, 2001; Moll et al., 2001). It is unclear, however, whether these activations were selective for moral relative to nonmoral judgment (as compared, for example, in (Moll et al., 2001)), because moral judgments were carried out on moral contents and nonmoral judgments on nonmoral contents, thus confounding process and content. In other approaches to moral judgments, different types of moral dilemmata were compared that differed in contents and were designed to probe different types of processes (Greene et al., 2004, 2001), but again confounding contents with process. The different theoretical interpretations of these findings will be briefly discussed in Section Cognitive-Anatomical Models of Moral Cognition and Emotion. These data do not, however, allow the conclusion that moral reasoning as a process has a specific neuroanatomical substrate.

Other investigators have sought to identify differences in implicit and explicit (Harenski and Hamann, 2006) moral cognition or real versus imagined moral decisions (Feldman Hall et al., 2012). While these may undoubtedly exist, these studies did not investigate whether the difference between these types of processing were selective for moral or social processing, which is the focus of this article.

Cognitive-Anatomical Models of Moral Cognition and Emotion

Anterior dorsolateral prefrontal and frontopolar cortex were found to be activated in moral judgment tasks (Greene et al., 2004, 2001; Moll et al., 2001), but as discussed, there is no evidence for their selective involvement in moral versus nonmoral judgment. The frontopolar cortex does show selectivity for guilt compared with other-critical feelings

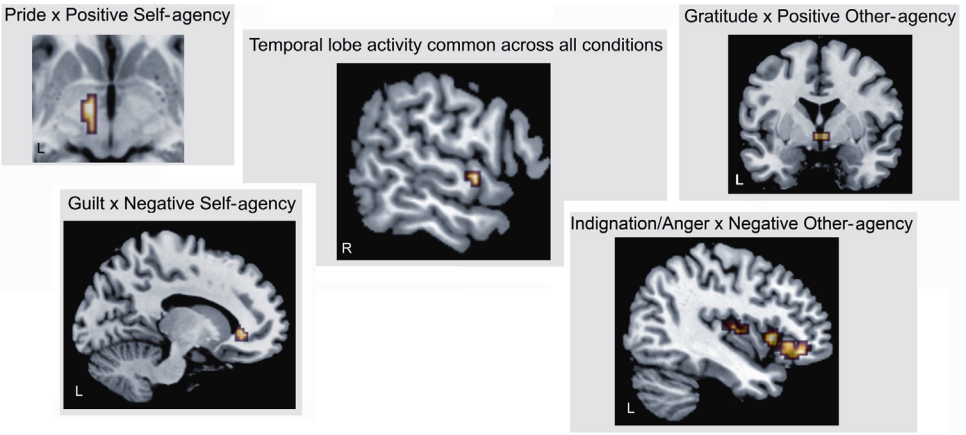

Figure 3 Using fMRI in healthy participants this study investigated the neuroanatomical basis of abstract moral and social values (figure adapted from Zahn, R., Moll, J., Paiva, M.M.F., Garrido, G., Krueger, F., Huey, E.D., Grafman, J., 2009c. The neural basis of human social values: evidence from functional MRI. Cerebral Cortex 19, 276–283; Zahn, R., Oliveira-Souza, R., Moll, J., 2011. The neuroscience of moral cognition and emotion. In: Decety, J., Cacioppo, J. (Eds.), The Handbook of Social Neuroscience. Oxford University Press). Participants had to imagine actions in accordance with or counter to a value described by a written sentence and to decide whether they would feel pleasantly or unpleasantly about the action. After the scan, they rated the unpleasantness/pleasantness on a scale and chose labels that best described their feelings (the analysis compares each moral sentiment vs visual fixation and vs two other moral sentiments; only selective effects were reported). There were four experimental conditions: (1) positive self-agency: "Tom (first name of participant) acts generously toward Sam (first name of best friend)" – pride in this condition was associated with ventral tegmental, septal, and ventral medial FPC activation (not depicted); (2) positive other-agency: "Sam acts generously toward Tom" – gratitude in this condition was associated with hypothalamic activation; (3) negative self-agency: "Tom acts stingily toward Sam" – guilt in this condition was associated with subgenual cingulate cortex as well as ventral medial FPC activation (not depicted and only when modeling individual frequency of guilt trials); and (4) negative other-agency: "Sam acts stingily toward Tom" – indignation/anger in this condition was associated with lateral orbitofrontal/insular activation. In the center, one can see the right superior anterior temporal lobe (aTL) region showing equally strong activation during all moral sentiment and agency contexts; this region showed increased activity with increasing richness of conceptual details describing social behavior and is identical to the activation found in a semantic judgment task (Zahn et al., 2007). These results confirmed the right superior anterior temporal lobe as a context-independent store of social conceptual knowledge that allows us to understand the core meaning of social and moral values regardless of what exact feelings or actions one ties to the value.

(see above) and was also activated for affiliative versus nonaffiliative feelings (Moll et al., 2012). It is therefore unlikely to be involved in the process of judgment itself. Unpublished secondary data analyses show that frontopolar activations in the latter study were associated with the number of consequences people thought about during scanning as based on postscanning ratings. This confirms the prediction of the fronto-temporo-subcortical integration model that frontopolar activations contribute to moral cognition by representing consequences of one's actions (**Figure 4**), a crucial ingredient of affiliative moral sentiments (i.e., guilt and pity).

It remains controversial how to interpret frontal activations in moral cognition and emotion; the different models are briefly summarized in **Table 1**. Further details of the fronto-temporo-subcortical integration model of moral cognition and emotion are found in **Figure 4**. This model currently provides the most parsimonious explanation for different cortical activation patterns for different types of moral sentiments (guilt vs indignation). Furthermore, it is so far the only model predicting subcortical and cortical structures with selective involvement in moral motivations. This is consistent with the evidence presented in previous sections. The authors have provided a more detailed discussion of how different models explain the evidence regarding moral dilemma choice paradigms elsewhere (Moll et al., 2008).

Implications for Differential Vulnerability in Psychiatry

The application of moral neuroscience to the study of psychiatric disorders is still in its infancy. It is likely, however, that moral neuroscience will play an important role in understanding differential vulnerability to certain psychiatric disorders. Here, the authors will focus on major depressive disorder (MDD), especially, its melancholic subtype, on the hypermoral, and psychopathy on the hypomoral end of a putative spectrum. This is a highly speculative perspective, but may prove fruitful for designing future experiments to probe this further.

Psychopathy has been long conceived of as a disorder of moral motivation with a lack of guilt and pity while leaving the ability to manipulate others for one's selfish needs intact (Cleckley, 1976). Callousness in patients with psychopathy (Hare, 2003) was indeed found to be associated with gray matter volume reduction in a network of brain regions including the frontopolar and subgenual cortex (de Oliveira-Souza et al., 2008). The anterior subgenual cortex was also found to be less activated in response to moral violation-depicting pictures in incarcerated people with psychopathy compared with a control group (Harenski and Hamann, 2006). The right anterior temporal lobe was another area of abnormal activation in this study and showed gray matter volume loss in a further study in people with psychopathy

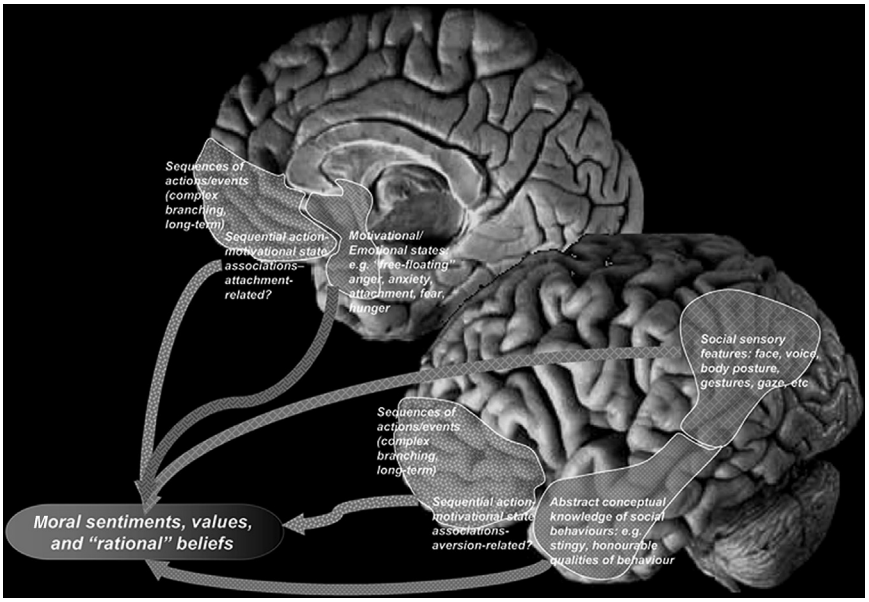

Figure 4 Brain regions involved in moral functions based on evidence from brain lesions and functional neuroimaging (adapted from Zahn, R., Oliveira-Souza, R., Moll, J., 2011. The neuroscience of moral cognition and emotion. In: Decety, J., Cacioppo, J. (Eds.), The Handbook of Social Neuroscience. Oxford University Press; Moll, J., and Schulkin, J., 2009. Social attachment and aversion in human moral cognition. Neuroscience and Biobehavioral Reviews 33, 456–465; Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., Grafman, J., 2005b. Opinion: the neural basis of human moral cognition. Nature Reviews Neuroscience 6, 799–809). Cortical region: frontopolar cortex (FPC), medial and lateral ventral prefrontal cortex (PFC), right anterior dorsolateral PFC, anterior temporal lobes (aTL), and posterior superior temporal sulcus (pSTS). Subcortical structures include the extended amygdala, hypothalamus, basal forebrain (especially the preoptic and septal regions), basal ganglia, and midbrain regions. Integration across these corticolimbic structures gives rise to event-feature-emotion complexes (EFEC) by temporal binding according to the fronto-temporo-mesolimbic integration model (Moll et al., 2005b). The hypothesized cognitive-anatomical components are the following: (1) sequential knowledge of actions/events represented within PFC subregions. FPC: complex branching of consequences of actions and ventral PFC regions representing associative knowledge of motivational/emotional states embedded into sequential event/action contexts; (2) social sensory features stored in pSTS and abstract (i.e., context-independent) conceptual knowledge of social behavior stored in the anterior temporal cortex, especially in the superior sectors; (3) central motive or basic emotional states, such as 'free-floating' anger, attachment, sadness, and sexual arousal (represented by the subcortical structures listed earlier).

Table 1 Alternative models of the role of frontal cortex in social and moral behavior

Hypothesis label	*Main claim*
Top-down frontal cognitive control	Frontal cortex does not store linkages between stimulus and response, but it represents goals that control the information flow in other cortical and subcortical areas when an automatic response needs to be overcome (Miller and Cohen, 2001).
Bottom–up subcortical control	Ventral (medial) frontal cortex stores linkages of subcortical 'somatic markers' and action knowledge in posterior brain areas. Explains problems with rapid and complex decision making after ventral (medial) frontal lesions (Bechara et al., 2000).
Reciprocal fronto-temporo-subcortical integration	Context-dependent knowledge of sequences of social actions/events is stored in ventral frontal (and frontopolar) cortex necessary to enable context-appropriate social behavior (Moll et al., 2005b). In this model, the frontal cortex stores sequential information linking stimuli and responses in different contexts.

Three alternative models of frontal involvement in social and moral cognition are summarized. The influential cognitive control model has subsequently been extended to the moral domain as the dual-process model of moral cognition and emotion (Greene et al., 2004). The assumption of this model is that 'rational' or 'cognitive' processes relying on dorsolateral prefrontal and parietal cortex can be separated anatomically from 'emotional' processes relying on subcortical and ventromedial frontal areas. Therefore, the model predicts that rational and emotional processing of moral dilemmata leads to different choices. The authors have pointed out that the model fails to explain how different types of emotions (e.g., guilt and indignation) and different types of motivations (e.g., selfish vs moral) can be accounted for. The reciprocal fronto-temporo-subcortical integration model of moral cognition and emotion is based on representations of different types of contents within each brain area of the network. This allows for explaining dissociations between different types of contents relying on different parts of one larger brain area (Moll et al., 2008).

(Muller et al., 2008). Frontopolar and anterior temporal lobe gray matter volume loss were reproduced as the distinctive feature of people with psychopathy compared with people with antisocial personality disorder, but no psychopathy (Gregory et al., 2012). The right uncinate fascicle, a white matter tract connecting temporal poles, ventral frontal cortices, and subcortical structures, was found to show decreased structural connectivity in people with psychopathy (Eslinger et al., 2004), which may be a result of the cortical changes described above or vice versa.

MDD is associated with overgeneral forms of empathy-based interpersonal guilt (e.g., 'feeling responsible for everything,' (O'Connor et al., 2002)), which remained detectable on remission of symptoms in a group of primarily melancholic subtype patients (Green et al., 2013) suggesting this as a trait vulnerability factor. One hypothesis is that increased guilt proneness in MDD may have an influence on helping behavior outside of depressive episodes (O'Connor et al., 2002). This link has so far, however, not been conclusively demonstrated and the overgeneral and uncontrollable nature of self-blame as proposed in an influential cognitive model of MDD (Abramson et al., 1978) could lead to the opposite prediction of a propensity to withdraw from helping others (Tangney et al., 2007). Overgeneralization of guilt (i.e., feeling guilty for everything and thus hating oneself as a result) was associated with functional disconnection of the right superior anterior temporal lobe and the septal-subgenual cingulate region in people with MDD whose symptoms had subsided (Green et al., 2012). This functional disconnection was only present for guilt and not for indignation, thus confirming the selectivity for overgeneral self-blame, explaining the relatively infrequent occurrence of violence in people with pure major depression (Green et al., 2013). This confirmed the prediction of frontotemporal integration as critical for adaptive moral motivations (Moll et al., 2005b). Further studies of neural correlates of moral sentiments in affective disorders are needed to confirm and extend these findings.

Conclusions and Future Directions

The authors have presented evidence that the septal-subgenual cingulate and septohypothalamic area are currently the best candidates for hosting subregions with selective involvement in moral versus selfish motivations. The authors have further reviewed the reproducible involvement of the frontopolar cortex for prosocial moral sentiments relative to other-critical sentiments. None of the other brain areas activated in moral cognition and emotion tasks were yet shown to be credible candidates for selective involvement in moral relative to nonmoral motivations. Social knowledge as a critical prerequisite of moral cognition and emotion is only partly understood in that its context-independent (i.e., conceptual) aspects can be localized within the right anterior temporal lobe. Future studies are needed to confirm these findings and to extend the understanding of the functional contribution of different frontal brain regions. An exciting future of applying these new insights to an improved pathophysiological understanding and treatment of psychiatric disorders lies ahead.

See also: Cortical Areas Engaged in Movement: Neuroimaging Methods; Depression; Emotion in Cognition; Emotion, Perception and Expression of; Heuristics in Social Cognition; Limbic System; Moral Development, Cultural Differences In; Moral Development, Theories of; Morality: Evolution of; Prefrontal Cortex Development and Development of Cognitive Function; Prefrontal Cortex; Psychopathy.

Bibliography

Abramson, L.Y., Seligman, M.E.P., Teasdale, J.D., 1978. Learned helplessness in humans - critique and reformulation. Journal of Abnormal Psychology 87, 49–74.

Adolphs, R., 2009. The social brain: neural basis of social knowledge. Annual Review of Psychology 60, 693–716.

Aron, A., Fisher, H., Mashek, D.J., Strong, G., Li, H., Brown, L.L., 2005. Reward, motivation, and emotion systems associated with early-stage intense romantic love. Journal of Neurophysiology 94, 327–337.

Bartels, A., Zeki, S., 2004. The neural correlates of maternal and romantic love. Neuroimage 21, 1155–1166.

Basile, B., Mancini, F., Macaluso, E., Caltagirone, C., Frackowiak, R.S., Bozzali, M., 2011. Deontological and altruistic guilt: evidence for distinct neurobiological substrates. Human Brain Mapping 32, 229–239.

Bechara, A., Damasio, H., Damasio, A.R., 2000. Emotion, decision making and the orbitofrontal cortex. Cerebral Cortex 10, 295–307.

Bishop, J.D., 1996. Moral motivation and the development of Francis Hutcheson's philosophy. Journal of the History of Ideas 57, 277–295.

Borg, J.S., Lieberman, D., Kiehl, K.A., 2008. Infection, incest, and iniquity: investigating the neural correlates of disgust and morality. Journal of Cognitive Neuroscience 20, 1529–1546.

Boyd, R., Gintis, H., Bowles, S., Richerson, P.J., 2003. The evolution of altruistic punishment. Proceedings of the National Academy of Sciences USA 100, 3531–3535.

Boyd, R., Richerson, P.J., 2009. Culture and the evolution of human cooperation. Philosophical Transcations of the Royal Society of London. Series B, Biological Sciences 364, 3281–3288.

Bozeat, S., Gregory, C.A., Ralph, M.A., Hodges, J.R., 2000. Which neuropsychiatric and behavioural features distinguish frontal and temporal variants of frontotemporal dementia from Alzheimer's disease? Journal of Neurology, Neurosurgery and Psychiatry 69, 178–186.

Brutkowski, S., 1965. Functions of prefrontal cortex in animals. Physiological Reviews 45, 721–746.

Casanova, M.F., Buxhoeveden, D.P., Switala, A.E., Roy, E., 2002. Minicolumnar pathology in autism. Neurology 58, 428–432.

Cleckley, H.M., 1976. The Mask of Sanity. Mosby, St. Louis, MO.

Damasio, H., Grabowski, T., Frank, R., Galaburda, A.M., Damasio, A.R., 1994. The return of Phineas Gage: clues about the brain from the skull of a famous patient. Science 264, 1102–1105.

de Oliveira-Souza, R., Hare, R.D., Bramati, I.E., Garrido, G.J., Azevedo Ignácio, F., Tovar-Moll, F., Moll, J., 2008. Psychopathy as a disorder of the moral brain: fronto-temporo-limbic grey matter reductions demonstrated by voxel-based morphometry. NeuroImage 40, 1202–1213.

de Vignemont, F., Singer, T., 2006. The empathic brain: how, when and why? Trends in Cognitive Sciences 10, 435–441.

Decety, J., Grezes, J., 2006. The power of simulation: imagining one's own and other's behavior. Brain Research 1079, 4–14.

Decety, J., Michalska, K.J., Kinzler, K.D., 2012. The contribution of emotion and cognition to moral sensitivity: a neurodevelopmental study. Cerebral Cortex 22, 209–220.

Eisenberg, N., 2000. Emotion, regulation, and moral development. Annual Review of Psychology 51, 665–697.

Eslinger, P.J., Damasio, A.R., 1985. Severe disturbance of higher cognition after bilateral frontal lobe ablation: patient EVR. Neurology 35, 1731–1741.

Eslinger, P.J., Flaherty-Craig, C.V., Benton, A.L., 2004. Developmental outcomes after early prefrontal cortex damage. Brain and Cognition 55, 84–103.

Frith, C.D., 2007. The social brain? Philosophical Transcations of the Royal Society of London. Seires B, Biological Science 362, 671–678.
Funk, M., 2010. Mental Health and Development: Targeting People with Mental Health Conditions as a Vulnerable Group. WHO.
Gintis, H., Henrich, J., Bowles, S., Boyd, R., Fehr, E., 2008. Strong reciprocity and the roots of human morality. Social Justice Research 21, 241–253.
Green, S., Lambon Ralph, M.A., Moll, J., Deakin, J.F., Zahn, R., 2012. Guilt-selective functional disconnection of anterior temporal and subgenual cortices in major depressive disorder. Archives of General Psychiatry 69, 1014–1021.
Green, S., Moll, J., Deakin, J.F., Hulleman, J., Zahn, R., 2013. Proneness to decreased negative emotions in major depressive disorder when blaming others rather than oneself. Psychopathology 46, 34–44.
Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., Cohen, J.D., 2004. The neural bases of cognitive conflict and control in moral judgment. Neuron 44, 389–400.
Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D., 2001. An fMRI investigation of emotional engagement in moral judgment. Science 293, 2105–2108.
Gregory, S., Ffytche, D., Simmons, A., et al., 2012. The antisocial brain: psychopathy matters: a structural MRI investigation of antisocial male violent offenders. Archives of General Psychiatry 69, 962–972.
Haidt, J., 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological Review 108, 814–834.
Hall, Oriel F., Dalgleish, T., Thompson, R., Evans, D., Schweizer, S., Mobbs, D., 2012. Differential neural circuitry and self-interest in real vs hypothetical moral decisions. Social Cognitive and Affective Neuroscience 7, 743–751.
Harbaugh, W.T., Mayr, U., Burghart, D.R., 2007. Neural responses to taxation and voluntary giving reveal motives for charitable donations. Science 316, 1622–1625.
Hare, R.D., 2003. The Hare Psychopathy Checklist-Revised. Multi-Health Systems, Toronto.
Harenski, C.L., Hamann, S., 2006. Neural correlates of regulating negative emotions related to moral violations. Neuroimage 30, 313–324.
Herholz, K., 2003. PET studies in dementia. Annals Nuclear Medicine 17, 79–89.
Hsu, M., Anen, C., Quartz, S.R., 2008. The right and the good: distributive justice and neural encoding of equity and efficiency. Science 320, 1092–1095.
Immordino-Yang, M.H., Mccoll, A., Damasio, H., Damasio, A., 2009. Neural correlates of admiration and compassion. Proceedings of the National Academy of Sciences of the United States of America 106, 8021–8026.
Insel, T.R., Young, L.J., 2001. The neurobiology of attachment. Nature Reviews Neuroscience 2, 129–136.
Kant, I., 1786/1903. Grundlegung zur Metaphysik der Sitten. Kgl Preussische Akademie der Wissenschaften, Berlin.
Kedia, G., Berthoz, S., Wessa, M., Hilton, D., Martinot, J.L., 2008. An agent harms a victim: a functional magnetic resonance imaging study on specific moral emotions. Journal of Cognitive Neuroscience 20, 1788–1798.
Keller, M.C., Miller, G., 2006. Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? Behavioral and Brain Sciences 29, 385–404 discussion 405–452.
Kendler, K.S., 2006. Reflections on the relationship between psychiatric genetics and psychiatric nosology. American Journal of Psychiatry 163, 1138–1146.
Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., Damasio, A., 2007. Damage to the prefrontal cortex increases utilitarian moral judgements. Nature 446, 908–911.
Krueger, F., Moll, J., Zahn, R., Heinecke, A., Grafman, J., 2007. Event frequency modulates the processing of daily life activities in human medial prefrontal cortex. Cerebral Cortex 17, 2346–2353.
Krueger, F., Spampinato, M.V., Barbey, A.K., Huey, E.D., Morland, T., Grafman, J., 2009. The frontopolar cortex mediates event knowledge complexity: a parametric functional MRI study. Neuroreport 20, 1093–1097.
Liu, W., Miller, B.L., Kramer, J.H., Rankin, K., Wyss-Coray, C., Gearhart, R., Phengrasamy, L., Weiner, M., Rosen, H.J., 2004. Behavioral disorders in the frontal and temporal variants of frontotemporal dementia. Neurology 62, 742–748.
Mendez, M.F., Chow, T., Ringman, J., Twitchell, G., Hinkin, C.H., 2000. Pedophilia and temporal lobe disturbances. Journal of Neuropsychiatry and Clinical Neurosciences 12, 71–76.
Miller, E.K., Cohen, J.D., 2001. An integrative theory of prefrontal cortex function. Annual Review of Neuroscience 24, 167–202.
Moll, J., Bado, P., de Oliveira-Souza, R., Bramati, I.E., Lima, D.O., Paiva, F.F., Sato, J.R., Tovar-Moll, F., Zahn, R., 2012. A neural signature of affiliative emotion in the human septohypothalamic area. Journal of Neuroscience 32, 12499–12505.
Moll, J., de Oliveira-Souza, R., Eslinger, P.J., 2003. Morals and the human brain: a working model. Neuroreport 14, 299–305.
Moll, J., de Oliveira-Souza, R., Garrido, G.J., Bramati, I.E., Caparelli-Daquer, E.M.A., Paiva, M.M.F., Zahn, R., Grafman, J., 2007. The self as a moral agent: linking the neural bases of social agency and moral sensitivity. Social Neuroscience 2, 336–352.
Moll, J., de Oliveira-Souza, R., Moll, F.T., Ignacio, F.A., Bramati, I.E., Caparelli-Daquer, E.M., Eslinger, P.J., 2005a. The moral affiliations of disgust: a functional MRI study. Cognitive and Behavioral Neurology 18, 68–78.
Moll, J., de Oliveira-Souza, R., Zahn, R., 2008. The neural basis of moral cognition: sentiments, concepts, and values. Annals of the New York Academy of Sciences 1124, 161–180.
Moll, J., Eslinger, P.J., Oliveira-Souza, R., 2001. Frontopolar and anterior temporal cortex activation in a moral judgment task: preliminary functional MRI results in normal subjects. Arq Neuropsiquiatr 59, 657–664.
Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., Grafman, J., 2006. Human fronto-mesolimbic networks guide decisions about charitable donation. Proceedings of the National Academy of Sciences of the United States of America 103, 15623–15628.
Moll, J., Schulkin, J., 2009. Social attachment and aversion in human moral cognition. Neuroscience and Biobehavioral Reviews 33, 456–465.
Moll, J., Zahn, R., de Oliveira-Souza, R., Bramati, I.E., Krueger, F., Tura, B., Cavanagh, A.L., Grafman, J., 2011. Impairment of prosocial sentiments is associated with frontopolar and septal damage in frontotemporal dementia. Neuroimage 54, 1735–1742.
Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., Grafman, J., 2005b. Opinion: the neural basis of human moral cognition. Nature Reviews Neuroscience 6, 799–809.
Morey, R.A., Mccarthy, G., Selgrade, E.S., Seth, S., Nasser, J.D., Labar, K.S., 2012. Neural systems for guilt from actions affecting self versus others. Neuroimage 60, 683–692.
Muller, J.L., Ganssbauer, S., Sommer, M., Dohnel, K., Weber, T., Schmidt-Wilcke, T., Hajak, G., 2008. Gray matter changes in right superior temporal gyrus in criminal psychopaths. Evidence from voxel-based morphometry. Psychiatry Research 163, 213–222.
Nowak, M.A., Sigmund, K., 2005. Evolution of indirect reciprocity. Nature 437, 1291–1298.
O'connor, L.E., Berry, J.W., Weiss, J., Gilbert, P., 2002. Guilt, fear, submission, and empathy in depression. Journal of Affective Disorders 71, 19–27.
Ochsner, K.N., Ray, R.D., Cooper, J.C., Robertson, E.R., Chopra, S., Gabrieli, J.D.E., Gross, J.J., 2004. For better or for worse: neural systems supporting the cognitive down- and up-regulation of negative emotion. Neuroimage 23, 483–499.
Parkinson, C., Sinnott-Armstrong, W., Koralus, P.E., Mendelovici, A., Mcgeer, V., Wheatley, T., 2011. Is morality unified? Evidence that distinct neural systems underlie moral judgments of harm, dishonesty, and disgust. Journal of Cognitive Neuroscience 23, 3162–3180.
Ralph, M.A.L., Patterson, K., 2008. Generalization and differentiation in semantic memory - insights from semantic dementia. Year in Cognitive Neuroscience 1124, 61–76.
Ross, L.A., Olson, I.R., 2010. Social cognition and the anterior temporal lobes. NeuroImage 49, 3452–3462.
Saver, J.L., Damasio, A.R., 1991. Preserved access and processing of social knowledge in a patient with acquired sociopathy due to ventromedial frontal damage. Neuropsychologia 29, 1241–1249.
Shin, L.M., Dougherty, D.D., Orr, S.P., Pitman, R.K., Lasko, M., Macklin, M.L., Alpert, N.M., Fischman, A.J., Rauch, S.L., 2000. Activation of anterior paralimbic structures during guilt-related script-driven imagery. Biological Psychiatry 48, 43–50.
Skipper, L.M., Ross, L.A., Olson, I.R., 2011. Sensory and semantic category subdivisions within the anterior temporal lobes. Neuropsychologia 49, 3419–3429.
Smith, A., 1759/1966. The Theory of Moral Sentiments. Kelly, New York.
Snowden, J.S., Bathgate, D., Varma, A., Blackshaw, A., Gibbons, Z.C., Neary, D., 2001. Distinct behavioural profiles in frontotemporal dementia and semantic dementia. Journal of Neurology Neurosurgery and Psychiatry 70, 323–332.
Sollberger, M., Stanley, C.M., Wilson, S.M., Gyurak, A., Beckman, V., Growdon, M., Jang, J., Weiner, M.W., Miller, B.L., Rankin, K.P., 2009. Neural basis of interpersonal traits in neurodegenerative diseases. Neuropsychologia 47, 2812–2827.
Stack, E.C., Balakrishnan, R., Numan, M.J., Numan, M., 2002. A functional neuroanatomical investigation of the role of the medial preoptic area in neural circuits regulating maternal behavior. Behavioural Brain Research 131, 17–36.
Swain, J.E., Lorberbaum, J.P., Kose, S., Strathearn, L., 2007. Brain basis of early parent-infant interactions: psychology, physiology, and in vivo functional neuroimaging studies. Journal of Child Psychology and Psychiatry 48, 262–287.

Takahashi, H., Yahata, N., Koeda, M., Matsuda, T., Asai, K., Okubo, Y., 2004. Brain activation associated with evaluative processes of guilt and embarrassment: an fMRI study. Neuroimage 23, 967–974.

Tangney, J.P., Stuewig, J., Mashek, D.J., 2007. Moral emotions and moral behavior. Annual Review of Psychology 58, 345–372.

Teuber, H.L., 2009. The riddle of frontal lobe function in man. Neuropsychology Review 19, 25–46. Reprinted from: The frontal granular cortex and behavior, p. 410 (1964).

Wagner, U., N'diaye, K., Ethofer, T., Vuilleumier, P., 2011. Guilt-specific processing in the prefrontal cortex. Cerebral Cortex 21, 2461–2470.

Welt, L., 1888. Über Charakterveränderungen des Menschen. Deutsches Archiv Fur Klinische Medizin 42, 339–390.

West, S.A., Griffin, A.S., Gardner, A., 2007. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. Journal of Evolutionary Biology 20, 415–432.

Wilson, D.S., Wilson, E., 2007. Rethinking the theoretical foundation of sociobiology. The Quarterly Review of Biology 82, 327–348.

Wood, J.N., Grafman, J., 2003. Human prefrontal cortex: processing and representational perspectives. Nature Reviews Neuroscience 4, 139–147.

Zahn, R., de Oliveira-Souza, R., Bramati, I., Garrido, G., Moll, J., 2009a. Subgenual cingulate activity reflects individual differences in empathic concern. Neuroscience Letters 457, 107–110.

Zahn, R., Drews, E., Specht, K., Kemeny, S., Reith, W., Willmes, K., Schwarz, M., Huber, W., 2004. Recovery of semantic word processing in global aphasia: a functional MRI study. Brain Research. Cognitive Brain Research 18, 322–336.

Zahn, R., Green, S., Burns, A., Moll, J., Caine, D., Lambon Ralph, M.A., 2012. Double dissociations of conceptual knowledge of social behaviour and the knowledge of its consequences in frontotemporal dementia International Conference on Frontotemporal Dementias Manchester. Dementia and Geriatric Cognitive Disorders.

Zahn, R., Moll, J., Iyengar, V., Huey, E.D., Tierney, M., Krueger, F., Grafman, J., 2009b. Social conceptual impairments in frontotemporal lobar degeneration with right anterior temporal hypometabolism. Brain 132, 604–616.

Zahn, R., Moll, J., Krueger, F., Huey, E.D., Garrido, G., Grafman, J., 2007. Social concepts are represented in the superior anterior temporal cortex. Proceedings of the National Academy of Sciences of the United States of America 104, 6430–6435.

Zahn, R., Moll, J., Paiva, M.M.F., Garrido, G., Krueger, F., Huey, E.D., Grafman, J., 2009c. The neural basis of human social values: evidence from functional MRI. Cerebral Cortex 19, 276–283.

Zahn, R., Oliveira-Souza, R., Moll, J., 2011. The neuroscience of moral cognition and emotion. In: Decety, J., Cacioppo, J. (Eds.), The Handbook of Social Neuroscience. Oxford University Press.

Neural Plasticity

Jon H Kaas, Vanderbilt University, Nashville, TN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Once it was thought that the human brain is immutable; but now it transpires that neural systems in human brains are highly plastic because the brain circuits are not fully specified by the genes. They are plastic in the sense that they are capable of changing in form and function in response to changes in the environment or damage to neural systems. Normal and altered brain development, learning and imprinting, and recoveries from brain injury and sensory loss all depend on this capacity, which is mediated by the growth and formation of new neural connections in the brain. Neural plasticity is provoked by genetic and cellular responses to changes in neural activity patterns.

Neural plasticity refers to the ability of neurons and neuron circuits to change in form and function in response to alterations in their environment. Neurons function as parts of local circuits in the brain, and each neuron can change its functional role in a circuit by altering how it responds to inputs or influences other neurons. Neuronal interactions for computational functions largely depend on synaptic relationships, and neural plasticity largely results from changes in the strengths of synaptic connections between neurons and the formation of new connections. Neural plasticity occurs as neurons respond to the activities of neighboring neurons that are spontaneously active or are activated by events in the external environment, and to trophic and guidance factors released in the local environment. Neural plasticity is essential for the normal development of brain circuits, creating the differences in those circuits that make us individuals. Neural plasticity mediates the acquisition of knowledge and skill, and brain repair after injury. Plasticity may also lead to misperceptions and pain, and maladaptive behavior. An understanding of neural plasticity might help us promote this plasticity when it is useful, and control this plasticity when it has unfortunate consequences.

Neural Plasticity during Brain Development

The basic circuitry of the brain gradually emerges during development. While the mature brain remains plastic, developmental changes are especially impressive. For example, a short period of sensory loss, such as produced by having an eye covered by a patch early in childhood, can have a devastating effect on vision of that eye. The brain circuits for vision become dominated by the unpatched seeing eye so that connections and synaptic strengths change, deprived neurons fail to grow and axon branches withdraw, and a deficit in vision for the closed eye occurs that persists even after the removal of the eye patch (Antonini and Stryker, 1993). A similar period of deprivation in adults has little effect. The developing brain is especially plastic because it is using information from the environment to help construct itself (Rauschecker, 1991; Schatz, 1990). Slight changes in the instructions from the environment, and the timing of those instructions can lead to quite different outcomes. In the mature brain, the circuits have been built, and further modifications are usually less profound.

One of the important accomplishments of developmental plasticity is to create functionally adaptive circuits in the brain. Developmental plasticity refers to the capacity of the brain to have different developmental outcomes. The development of some brain circuits, some of those in simple invertebrates, for example, may be highly fixed. Thus, the developing circuits usually achieve the same adult form. But, this is not necessarily the case for the complex circuits of the higher brain centers of mammals, especially those in humans with long periods of development. The human brain is highly plastic because the brain circuits are not fully specified by the genes. All of the necessary information for proper development does not reside in the genes or gene products. Thus, it is necessary to use information from the environment to direct development. Neurons are designed to respond to sensory stimuli and each other with action potentials or spikes. Self-generated spontaneous activity and evoked activity are very important in mediating normal development (Schatz, 1990). Neurons also influence the development of other neurons through the release of trophic and guidance factors (Conner and Dragunow, 1998).

Consistencies in the internal and external environments during development typically lead to the development of similar brains with very similar brain circuits. Because the developmental outcomes can be so similar, the role of the environment may not always be obvious. Thus, one way to study developmental plasticity is to rear animals in abnormal environments (Rauschecker, 1991). As examples, the development of the visual system can be altered by rearing mammals in the dark, with one eye closed, or under strobe lights. These experiments indicate that sensitivity to the environment is an important component of the normal development of brain circuits. In an abnormal environment, development can be misdirected, often with unfortunate functional consequences.

Sensitivity to the environment also allows the developing nervous system to become tailored to specifics of the environment of the individual. The newly hatched bird imprints on its parents, rather than birds in general, although even a moving box or balloon will do in the absence of tending parents. The young bird can become completely attached to such obviously false parents. The child that practices piano or basketball becomes the great piano or basketball player. Humans learn the sounds of their own language, and as adults

find it hard to process the sounds of other languages. Thus, neural plasticity allows the brain to adapt to an uncertain environment, and permits specializing the formation of neural circuits within individuals for individual needs.

Because neuronal circuits are designed to be altered by aspects of the environment at certain times in development, or subsequently during critical points – a period in brain development when certain brain circuits are highly capable of modification as a result of experience – in the lifecycle, there are changes in circuits that are possible only at certain times, and these changes are often difficult or impossible to reverse. These times are called critical or sensitive periods. Eye patching for a period of days to weeks, for example, will greatly alter brain circuits, only during a fixed early period of visual development. Earlier or later periods of deprivation have little effect. There may be a critical or sensitive period for songbirds learning songs, for human language to be easily and flawlessly acquired, and for learning perfect pitch.

In summary, neural plasticity is essential for normal brain development, especially for the development of large complex brains that mature slowly. The normal development of the human brain is especially dependent on neuronal plasticity. Many circuits require information from the environment in order to develop properly. In addition, neuronal plasticity allows the development of the individual differences, including individual skills and abilities.

Neural Plasticity and the Lifecycle

Neural plasticity also allows brain circuits to function differently at different phases of the lifecycle. Hormones play a major role in adjusting neural circuits. Songbirds, for example, adjust and even add new neurons to the neural circuits used in producing and acquiring songs used during the breeding season (Konishi, 1985). At puberty, humans and other mammals alter brain circuits that mediate aspects of sexual behavior, and later brain circuits are adjusted to promote maternal and paternal behaviors. Thus, plasticity permits rapid and functionally important changes in brain circuits as life styles change with puberty and parenthood.

Neural Plasticity and Learning

Learning is based on neural plasticity. Each individual acquires a great deal of useful information about the environment through experience, and improves and perfects a great number of motor skills. The importance of learning varies with species, and it is probably most important in humans. There are several types of learning, but they all depend on modifying the functional circuits of neurons.

The neuronal mechanisms of learning are not fully understood, although they have been under intensive study for many years. It is likely that two types of induced changes in synaptic effectiveness between neurons are important in most learning situations. When a source of input to a neuron or group of neurons is intensely stimulated for a brief period of time, subsequent stimulations might be more effective and produce a greater response. This change in synaptic effectiveness, which may persist for minutes to hours or longer, has been called long-term potentiation (LTP; see Bliss and Lomo, 1973). Of course, new learning involves forgetting, so that connection strengths need to be weakened as well. Patterns of less intense, more distributed stimulation may reduce synaptic effectiveness. This reduction is called long-term depression (LTD; see Nelson and Turringiano, 1998). Possibly, LTP depends on reversing LTD.

Both LTP and LTD are likely to be critical components of memory formation and the acquisition of skills. LTP and LTD have been most studied in the hippocampus, both in a tissue slice preparation, which offers clear technical advantages, and in the intact animal. LTP and LTD are components of what has been called Hebbian plasticity (see Brown et al., 1988; Rauschecker, 1991) after the neuropsychologist, Donald Hebb. In an effort to account for learning, Hebb theorized that the repeated activations of local neuron circuits changed the effectiveness of pathways so that some inputs came to activate the circuits more easily and more effectively. If neuron A takes part in the firing of neuron B, the future effectiveness of neuron A in firing neuron B is increased. The cellular mechanisms of Hebbian plasticity have been studied, and they involve the activation of the NMDA (N-methyl-D-aspartate) glutamate receptor and a calcium influx into the postsynaptic neuron when presynaptic activity is paired with postsynaptic depolarization (Bear, 1995; Rauschecker, 1991). The hippocampus is known to have a critical early role in the formation of long-term memories, but it is not generally thought to be the site of permanent storage. The amygdala encodes the emotional component of memories formed during fear conditioning, but like the hippocampus, the neuronal plasticity that occurs in the amygdala may modulate the consolidation of memories formed elsewhere.

Most memories and skills depend on neocortex. LTP and LTD also have been demonstrated in neocortex circuits where these types of synaptic plasticity are likely to be important in the consolidation and storage of information. Finally, plastic circuits in the basal ganglia are used in the acquisition of motor skills (Graybiel, 1995).

Neural Plasticity and Brain Repair

Neural plasticity is also important in brain reorganization and repair after injury. The effects of sensory loss have been extensively studied in the sensory systems of adult mammals (Kaas, 2001; Kaas et al., 2008). Sensory loss of one type or another commonly occurs during adult life. For example, humans typically lose auditory receptors and, especially, high-frequency hearing as they age. Studies in monkeys, cats, and rodents show that auditory cortex reorganizes after such partial hearing loss so that more neurons are devoted to processing the remaining inputs. The functional significance of this reorganization is not clear, but it may help individuals adjust to the hearing loss and more effectively use the remaining inputs. Similar changes occur when small portions of the retina are damaged. The regions of visual cortex that are deprived of their normal sources of activation by such damage soon come to be responsive to surrounding intact parts of the retina.

Finally, any restricted loss of tactile inputs from the skin, such as those caused by nerve damage, is followed by a

reorganization of somatosensory cortex so that deprived neurons become responsive to preserved tactile afferents. Such findings indicate that the brains of even adult mammals have considerable capacity for functional reorganization so that neurons that are deprived of a normal source of activation gradually acquire new sources of activation. Unresponsive neurons are not of much use, and the reactivation of deprived neurons, as a common feature of brain damage, may greatly contribute to behavior recovery and compensation. The mature brain may well be designed to compensate for some neuronal loss and damage, and reorganization may lead to more effective use of remaining neurons and neural circuits. Humans and other mammals compensate remarkably well from cortical damage that is relatively restricted. In part, this may be the result of reorganization within the damaged area of cortex. Sensory and motor representations in cortex reorganize after damage to recover some of the lost parts of the representation. Other intact areas of cortex may reorganize to help compensate for the loss. During some periods of brain development, some parts of the brain may be capable of regenerating lost neurons (Kolb et al., 1998). Impressive recoveries after brain damage may reflect design features involving plasticity, or even the fortunate utilization of mechanisms built in by selection for normal development and adult learning. However, plasticity is not always beneficial. The plasticity that is there for one set of purposes may be evoked in other sets of circumstances where the results are detrimental and unfortunate. As one example, individuals with amputated limbs typically feel that the missing limb is still there (Ramachandran and Blakeslee, 1998). This may be the result of neurons in the brain, deprived of normal sources of activation by afferents from the missing limb, becoming more sensitive to intrinsic signals or other inputs. The growth of new connections in the brainstem and cortex seems to add the further complexity of trigger zones, locations on the face and stump that, if touched, evoke a sensation in the missing limb. Fortunately, this type of misperception is not very disturbing. However, severe pain also might be felt in the missing limb, apparently as a result of the plasticity of circuits in the thalamus. This pain is very difficult to treat as it is not generated in the normal way. As another example of maladaptive plasticity, individuals that over practice a skill may develop motor disorders or dystonias (Bly et al., 1996). Neural plasticity is essential for developing motor skills, but musicians and writers may over practice or overuse a skill so that performance deteriorates rather than improves. The practice may coactivate neurons so frequently that they become so strongly interconnected that they fail to respond independently when needed in other neuronal circuits. Thus, the musician loses control of the skilled hand movements needed to play, and the writer develops writer's cramp. Recoveries depend on other training and experiences that rebalance the motor circuits.

Mechanisms of Neural Plasticity

Brain circuits are plastic because new connections form, unused connections are lost, and existing connections increase or decrease in effectiveness. Developing brains are especially plastic because it is a time of great neuronal growth. Activity-dependent genes (Kaczmarek and Chaudhuri, 1997) within neurons function to promote and direct new growth, or stop growth, and cause retractions of axons and dendrites and even cell death. The developing nervous system overproduces neurons and connections, and uses activity patterns induced in neurons by the external and internal environments to select the useful, more active neurons and connections and eliminate other less active neurons and connections. By altering activity patterns through lesions, drugs, and experience, outcomes can be changed so that normally lost neurons are saved, normally lost connections are preserved, and even new growth occurs. One of the early discoveries was that rats reared in a complex rather than a simple environment (Greenough et al., 1985) developed a thicker cortex with larger neurons and more complex dendritic arbors. As a more unusual, more dramatic, example of developmental plasticity, rodents reared after a major loss of auditory inputs to the forebrain grow visual inputs into the auditory thalamus, so that auditory cortex is activated by visual stimuli (Sur et al., 1988). The potential for the growth as well as the loss of new connections is great in the developing brain. Neuronal growth and death can be a feature of normal lifestyle changes as well. Neurons may increase in size and connectional complexity in brain regions related to sexual behavior or bird song during appropriate parts of the breeding season, and then shrink and disconnect subsequently (Konishi, 1985). Neurons may even be born and added to such circuits, to subsequently die. While neuronal growth and even death are major components of programmed plasticity for expected developmental and lifecycle events, much less new growth or retraction might be expected at other times in the mature brain, which has developed brain circuits for a period of stable performance. Indeed, except for learning and learning mechanisms, the mature brain has been considered to be rather fixed in organization and not plastic. Now, considerable evidence has accumulated to show that the functional organization of the mature brain is quite mutable (Kaas, 2001), and that at least some of the changes depend on the growth of new connections, including dendritic branches and spines, and axonal sprouting (Darian-Smith, 2009). In cases of severe sensory loss, such as with brain or spinal cord lesions, or peripheral nerve damage, growth of axons over considerable distances occurs to reactivate deprived neurons in the spinal cord, brainstem, and cortex. While extensive new growth in the mature brain is often likely to be detrimental, and glial cells in the central nervous system produce factors that inhibit such new growth, the local growth of axons and dendrites is probably ongoing and common. In addition, the cellular factors for inhibiting more extensive growth may be overcome by extreme deprivation. It remains to be demonstrated how much new growth contributes to behavioral recoveries after brain damage, but a major contribution by new growth seems likely. In addition, new neurons may be added to injured and disconnected parts of the brain throughout life (Darian-Smith, 2009). While neurogenesis had been thought to be limited to the period of brain development, recent evidence for the addition of neurons to the mature primate brain raises the possibility

that such new neurons contribute to brain plasticity. A common and possibly more powerful factor in neural plasticity is increasing or decreasing the effectiveness of existing connections. This can be done by increasing or decreasing the number, sizes, and positions of synapses, the amount of neurotransmitter released, and the receptor sites responsive to the transmitter. The important roles of LTP and LTD have already been discussed. These mechanisms may be potentiated by the appropriate release of neuromodulators that increase neural activity for short but behaviorally significant periods of time. Thus, stressful and other motivating events activate brainstem neurons that release the neurotransmitter, acetylcholine (ACh) throughout the brain where synaptic strengths are likely to be modified. This makes the changes in neural circuits more likely, so that the motivating events can be retained in memory and skills can be acquired more readily. The permissive effects of ACh release on neural plasticity can even be obtained artificially by electrically stimulating the brainstem neurons that project to cortex and release ACh during learning (Kilgard and Merzenich, 1998).

There are many other ways of promoting neural plasticity. Neurons tend to self-regulate through the actions of activity-dependent genes to reach a set level of activity. Neurons with lowered levels of activity tend to express lower levels of the inhibitory neurotransmitter, gamma-aminobutyric acid (GABA), or become less sensitive to GABA by making fewer GABA receptors (Hendry and Jones, 1988). More active neurons make more GABA and more GABA receptors. Thus, deprived neurons become more sensitive and respond to inputs that were previously subthreshold and ineffective, while overactive neurons lose responsiveness to weak but previously effective inputs. Other activity-dependent genes probably regulate neuron growth through the release of trophic and guidance factors. In addition, it is possible to selectively promote the formation of new connections in the brain by temporarily digesting the perineuronal nets that form to inhibit the growth of new connections as the nervous system matures (Bowes et al., 2012). Thus, there are treatments that can be used to facilitate recoveries from sensory loss and damage to the brain.

Conclusions

Neural plasticity is important for the normal development and maintenance of neural circuits. This plasticity adjusts mature brain circuits to mediate individual perceptual and motor skills, and adaptive adjustments to an uncertain environment. Neural plasticity is especially important in storing information. Plasticity also functions in brain repair. Because some outcomes of neuronal plasticity are highly desirable, while others are unfortunate, a comprehensive understanding of the mechanisms of plasticity and how to control them would be clinically and therapeutically beneficial.

See also: Depression; Hippocampal Amnesia; Hippocampus and Related Structures; Long-Term Potentiation and Long-Term Depression; Neural Plasticity; Occipital Lobe (Visual Cortex): Functional Aspects.

Bibliography

Antonini, A., Stryker, M.P., 1993. Rapid remodeling of axonal arbors in the visual cortex. Science 260, 1819–1821.

Bear, M.F., 1995. Mechanism for a sliding synaptic modification threshold. Neuron 15, 1–4.

Bliss, T.V.P., Lomo, T., 1973. Long-lasting potentiation of synaptic transmission in the dentate area of the anesthetized rabbit following stimulation of the perforant path. Journal of Physiology 232, 331–356.

Bly, N.N., Merzenich, M.M., Jenkins, W.M., 1996. A primate genesis model of focal dystonia and repetitive strain injury: I: learning-induced dedifferentiation of the representation of the hand in the primary somatosensory cortex in adult monkeys. Neurology 47, 508–520.

Bowes, C., Massey, J.M., Burish, M., Cerkevich, C., Kaas, J.H., 2012. Chondroitinase ABC promotes selective reactivation of somatosensory cortex in squirrel monkeys after a cervical dorsal column lesion. Proceedings of the National Academy of Sciences of the United States of America 109, 2595–2600.

Brown, T.H., Chapman, P.F., Kairiss, E.W., Keenan, C.L., 1988. Long-term synaptic potentiation. Science 242, 724–728.

Conner, B., Dragunow, M., 1998. The role of neuronal growth factors in neurodegenerative disorders of the human brain. Brain Research Reviews 27, 1–39.

Darian-Smith, C., 2009. Synaptic plasticity, neurogenesis, and functional recovery after spinal cord injury. Neuroscientist 15, 129–165.

Graybiel, A.M., 1995. Building action repertoires: memory and learning functions of the basal ganglia. Current Neurology 5, 733–741.

Greenough, W.T., Hwang, H.M., Gorman, C., 1985. Evidence for active synapse formation or altered postsynaptic metabolism in visual cortex of rats reared in complex environment. Proceedings of the National Academy of Sciences of the United States of America 82, 4549–4552.

Hendry, S.H., Jones, E.G., 1988. Activity-dependent regulation of GABA expression in the visual cortex of adult monkeys. Neuron 1, 701–702.

Kaas, J.H., 2001. In: Kaas, J.H. (Ed.), The Mutable Brain. Harwood Academic Publishers, Amsterdam.

Kaas, J.H., Qi, H.X., Burish, M.J., Gharbawie, O.A., Onifer, S.M., Massey, J.M., 2008. Cortical and subcortical plasticity in the brains of humans, primates, and rats after damage to sensory afferents in the dorsal columns of the spinal cord. Experimental Neurology 209, 407–416.

Kaczmarek, L., Chaudhuri, A., 1997. Sensory regulation of immediate-early gene expression in mammalian visual cortex: implications of functional mapping and neural plasticity. Brain Research Reviews 23, 237–256.

Kilgard, M.P., Merzenich, M.M., 1998. Cortical map reorganization enabled by nucleus basalis activity. Science 279, 1714–1718.

Kolb, B., Gibb, R., Grazyna, G., Whishaw, I.Q., 1998. Possible regeneration of rat medial frontal cortex following neonatal frontal lesions. Behavioural Brain Research 91, 127–141.

Konishi, M., 1985. Birdsong: from behavior to neuron. Annual Review of Neuroscience 8, 125–170.

Nelson, S.B., Turringiano, G.G., 1998. A key payer in the cortical balancing act. Nature Neuroscience 1, 539–541.

Ramachandran, V.S., Blakeslee, S., 1998. Phantoms in the Brain. Morrow, New York,.

Rauschecker, J.P., 1991. Mechanisms of visual plasticity: Hebb synapses, NMDA receptors, and beyond. Physiological Reviews 71, 587–615.

Schatz, C.J., 1990. Impulse activity and the patterning of connectivity during CNS development. Neuron 5, 745–756.

Sur, M., Garraghty, P.E., Roe, A.W., 1988. Experimentally induced visual projections into auditory thalamus and cortex. Science 242, 1437–1441.

Neural Representations of Direction (Head Direction Cells)

Jeffrey S Taube, Dartmouth College, Hanover, NH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

To navigate successfully animals require knowledge about their location and directional orientation before embarking on a route. Cells in several limbic-system brain areas discharge in relation to the animal's directional heading in the azimuthal plane independent of location and are referred to as head direction (HD) cells. Each HD cell is tuned to a particular direction, which can be controlled by a visual landmark. However, directional firing is maintained in the absence of visual landmarks by using idiothetic cues, suggesting that the cell's response represents the animal's perceived directional heading. The vestibular system is critical for generating the directional signal.

The Head Direction Signal

To navigate successfully animals require knowledge about their location and directional heading before embarking on a route. While location information may be provided by place cells in the hippocampus, which discharge with respect to the animal's location, or grid cells in the entorhinal cortex, which discharge in a repeating tessellated hexagonal pattern (*see* Human Spatial Orientation, Neural Basis of), another group of limbic system neurons discharge in relation to the animal's directional heading in the azimuthal plane, independent of the animal's behavior and location (Taube et al., 1990a; Wiener and Taube, 2005; Taube, 2007). For example, a head direction (HD) cell might discharge whenever the animal points its head northeast, no matter where it is in the environment. The direction at which the cell fires maximally is referred to as the preferred firing direction. The position of the animal's trunk does not affect the cell's firing, nor does pitch or roll of the animal's head up to 90°. HD cell firing is motion-independent – cells will discharge whether the rat is moving or remains still, as long as the animal is pointing its head in the proper direction. Very little, if any, adaptation of firing rate is noted when the head remains fixed at the preferred firing direction.

HD cells are tuned to different directional headings, and within a particular brain area the cell population collectively encodes all 360° directions. Each HD cell can be characterized by its tuning curve, constructed by plotting firing rate as a function of directional heading. **Figure 1** shows a typical tuning curve for an HD cell that fired maximally at ~150. For each HD cell several properties can be characterized from its tuning curve: (1) the cell's preferred firing direction; (2) the maximal (peak) firing rate of the cell at the preferred direction; and (3) the angular range over which cell firing is elevated from background firing rate. Some HD cells have low peak firing rates (5–10 spikes per second) while other HD cells have high peak firing rates (100 spikes per second). Intermediate peak firing rates are also observed. The functional significance of different cells having different peak firing rates is unknown. The firing rate of each HD cell decreases linearly as one moves away from the preferred firing direction. The range of directional headings where firing is elevated above background firing levels is about 90° for most cells. The firing characteristics of an HD cell usually remain stable indefinitely, as long as the environmental parameters remain unchanged. Passive rotation of a head-fixed animal in the horizontal plane does not lead to any significant changes in cell firing properties, indicating that active head movements per se are not necessary to drive cell firing (Shinder and Taube, 2011).

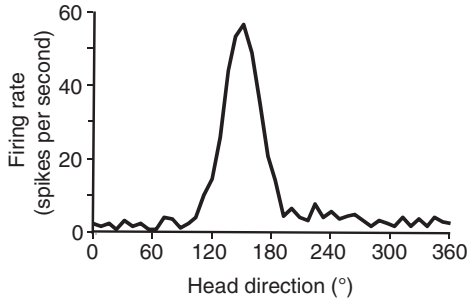

Figure 1 Firing rate versus head direction plot for a typical HD cell.

Where Are HD Cells Located in the Brain?

HD cells were originally identified in the dorsal portion of the presubiculum in rats (often referred to as the postsubiculum) by James Ranck. They have subsequently been identified in several limbic system areas in the rat, including the anterior dorsal, anterior ventral, and lateral dorsal thalamic nuclei, lateral mammillary nuclei, parasubiculum, and retrosplenial and medial entorhinal cortices (Taube, 2007). Note that many of these areas are considered part of the classical Papez circuit (**Figure 2**). HD cells have also been identified in some nonlimbic structures such as the dorsal striatum and medial granular cortex in the rodent's frontal cortex. The functional significance of having the signal present in so many different brain areas is presently unclear.

HD cells have been observed in a number of different species: mice, guinea pigs, chinchillas, and nonhuman primates. Importantly, the directional firing in the presubiculum of nonhuman primates has been shown to be independent of eye position, suggesting that the signal

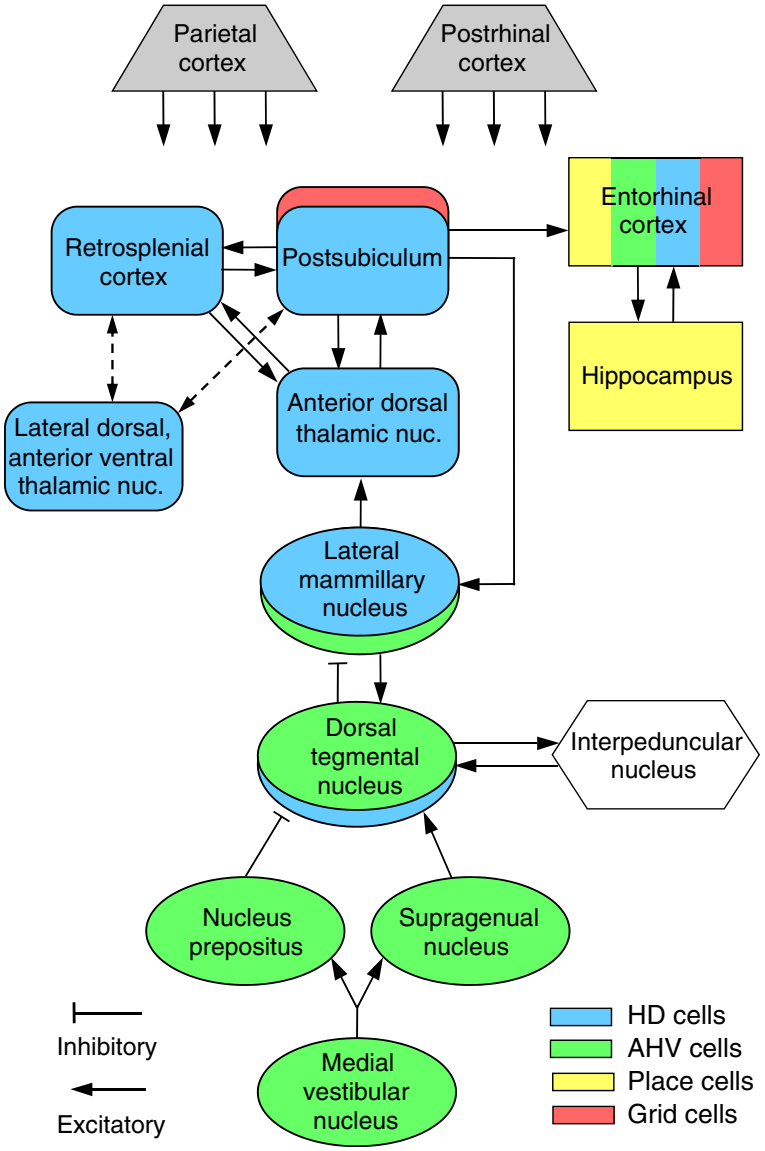

Figure 2 Head direction network.

represents the true directional heading of the animal and not its gaze direction (Robertson et al., 1999). With few exceptions, HD cells in each of the different brain areas have similar tuning curves, and behave similarly following manipulations of the animal's environment. However, HD cells in the anterior dorsal thalamus and lateral mammillary nuclei appear to discharge optimally to the animal's future directional heading, while HD cells in the dorsal presubiculum discharge maximally to the animal's current directional heading (Blair and Sharp, 1995). Similarly, in contrast to HD cells in the dorsal presubiculum, HD cell firing in the anterior dorsal thalamus and lateral mammillary nuclei is secondarily correlated to the rat's angular head velocity (AHV), where firing rates within the cell's preferred direction are higher for faster head turns than slower head turns (Taube, 1995). It is also noteworthy that there are distinct populations of cells in the lateral mammillary nuclei that, independent of directional heading, encode for either pitch of the animal's head or AHV (Stackman and Taube, 1998).

HD Cell Responses in 3D

How HD cells respond when the animal is oriented outside the yaw (Earth horizontal) plane has also been studied during the past decade. HD cells maintain their directional firing properties in the vertical plane, but usually lose them when they are inverted and locomote upside-down on a ceiling (Taube, 2011). When an animal actively locomotes from the horizontal plane into the vertical plane while moving in the direction of the cell's preferred firing direction, the cell will continue to discharge in the vertical plane, but if the vertical plane locomotion begins with the animal's head orientation

outside the cell's preferred direction, then the cell will not discharge on the vertical wall until the animal turns its head to the proper orientation. One model consistent with these findings is that HD cells define the horizontal reference frame as the animal's plane of locomotion, but this leaves open the question of which cue source(s) determine the horizontal reference frame: vestibular, somatosensory, visual, motor cues, or some combination of these sources. Alternatively, HD cells may define their reference frame in a fixed earth-based coordinate system, where their three-dimensional firing characteristics, in terms of yaw and pitch and using polar coordinates, can be represented as the surface of a hemitorus. In contrast to stable directional firing in the vertical plane, if the animal continues to move from the vertical plane onto a ceiling in an inverted (upside-down) position, then directional cell firing will cease. Why directional firing is disrupted when the animal is inverted is unclear, but recent findings have shown that the vestibular otolith organs, which are sensitive to the direction of gravity, play an important role in HD cell firing (Yoder and Taube, 2009).

Both Landmark Cues and Internally Derived Self-Motion Cues Affect HD Cell Firing

While HD cells bear a similarity to a navigational compass, they are not dependent on the Earth's magnetic field. They respond to several different types of sensory stimuli, some of which are external to the animal and others of which are self-generated. Visual landmarks that are stable over time play the predominant role in controlling the HD cell's preferred firing direction. For example, if an animal is first disoriented and then returned to a familiar environment where all the visual landmarks have been rotated, the HD cell's preferred firing direction will also shift a similar amount (Taube et al., 1990b). The postsubiculum plays a critical role in processing visual landmark information as lesions of it will impair the ability of a landmark to control HD cell preferred firing directions and hippocampal place cell fields (Yoder et al., 2011). Interestingly, neither disruption of the dorsal or ventral visual streams, via lesions of the parietal or entorhinal/postrhinal cortices, respectively, impair landmark control of HD cells, suggesting that the direct pathway from visual cortex to postsubiculum may convey information about visual landmarks to the limbic system.

Although visual landmarks exert strong control over their preferred firing directions, HD cells continue to respond normally when the animal moves about in darkness or when the prominent visual cues are removed, indicating that self-motion cues are capable of maintaining directional firing as well as the cell's preferred firing direction. HD cell firing is not environment- or context-specific, and the cells will discharge at some direction in any environment. These characteristics suggest that the HD cell network resembles more a population coding scheme rather than a sparse coding scheme. If an animal locomotes from a familiar to a novel environment, the preferred firing direction will be similar in both environments (Taube and Burton, 1995). These findings indicate that HD cells receive information concerning the animal's movements, either through vestibular, proprioceptive, or motor corollary discharge systems. This movement information is integrated over time so that the animal can keep track of its directional heading given an initial reference point. When an animal is passively transported from a familiar to a novel environment the preferred firing direction of a cell will often shift substantially, indicating the important role that motor systems play in maintaining a stable preferred firing direction (Stackman et al., 2003). Under this condition, the animal is deprived of the normal motor, proprioceptive, and kinesthetic cues that accompany self-locomotion, but vestibular cues are left intact. Thus, while the vestibular system may play a crucial role in the generation of the HD cell signal (see below), these findings indicate that vestibular cues alone are not sufficient to allow maintenance of a stable preferred direction and suggest a critical role for motor information for updating the signal over time.

Most studies have shown that when spatial information derived from landmarks conflicts with information derived from internally generated self-movement, the landmark information usually predominates and overrides the spatial signals developed through self-generated movements. Thus, when a salient, familiar cue is reintroduced into an animal's environment into a position that conflicts with the cell's current preferred firing direction, HD cells will shift their preferred direction to reflect their originally established orientation with the cue.

When two or more HD cells are monitored simultaneously in the same animal, the effects of an environmental manipulation on the preferred direction for one cell are similar to the responses observed in the other cell(s) (Taube et al., 1990b). This finding indicates that afferent input driving one HD cell will similarly influence other HD cells, and indicates that HD cells within a particular brain area behave as a network and that their preferred directions always remain a fixed angle apart (in-register) from one other.

Generation of the Directional Signal

Given that HD cells have been identified in several brain areas, many of which are interconnected with one another, the question arises as to which areas are critical for generating the directional signal. By recording from one brain area containing HD cells, and lesioning various afferent areas, a number of studies have concluded that (1) the vestibular system plays a critical role in generating and updating the directional signal (Stackman and Taube, 1997; Muir et al., 2009), and (2) that there is a general hierarchical flow of information that originates in the vestibular nuclei and then projects rostrally in the following sequence (see **Figure 2**): medial vestibular nuclei → supragenual nuclei and nucleus propositus (and possibly the paragigantocellularis reticular nucleus - not pictured in **Fig. 2**) → dorsal tegmental nuclei → lateral mammillary nuclei → anterodorsal thalamus → postsubiculum → medial entorhinal cortex (Sharp et al., 2001a; Taube, 2007). The importance of the vestibular system for generating the directional signal is also consistent with findings demonstrating the presence of cells sensitive to the animal's AHV throughout many of the brainstem nuclei mentioned above (Bassett and Taube, 2001; Sharp et al., 2001b).

Lesions of the anterior dorsal thalamus or the postsubiculum do not abolish location-specific firing in

hippocampal place cells, indicating that spatial representations concerning location can be generated independently of HD cell information. In contrast, recent evidence suggests that grid cell firing in the entorhinal cortex is dependent on HD signals, as lesions of the anterodorsal thalamus abolish the grid cell signal (Clark et al., 2011). The vestibular lesion data suggest that the neural code for directional bearing is critically dependent upon vestibular information, and that the loss of HD cell information may account for the orientation and navigational deficits observed following vestibular dysfunction in humans. Attempts to model HD cell firing have used attractor-type neural networks. These models have been able to replicate most, but not all, of the empirical data.

In animals with lesions of the hippocampus each HD cell's preferred firing direction remained stable across days, even when the animal was placed into a novel environment post-lesion (Golob and Taube, 1997). This result suggests that extrahippocampal structures are capable of creating and maintaining a novel representation of the animal's spatial context – at least in regards to the animal's directional heading. This association is considered a form of perceptual learning, as opposed to episodic learning.

Relationship of HD Cell Activity to Behavior

One important issue concerning HD cells is what role they play in an animal's behavior (Muir and Taube, 2002). Are HD cells actively involved in guiding navigational behavior? Evidence suggesting a link between HD cells and behavior has been produced using a radial arm maze in a spatial reference memory task. Animals were trained to run down a particular maze arm relative to a visible landmark cue to obtain a water reward. Following rotation of the cue, the animal's maze arm selection and the HD cell's preferred direction shifted in tandem relative to the salient visual cue. Furthermore, when the HD cell's preferred direction did not shift relative to the cue, neither did the animal's behavioral response. When an HD cell was recorded over several days as the animal learned the task, there was little change in the firing properties of the cell. These findings indicate that the coding of directional heading information is independent of the neural systems involved in task acquisition, and that over time, an animal learns to use information provided by the directional system to guide its behavioral responses.

Other experiments that have tested this hypothesis more rigorously have surprisingly found that HD cell activity and behavioral choice were not linked together on two different types of associative spatial tasks. One task was a spatial reference memory task where animals were required to generalize appropriately between a square and a rectangular enclosure to receive a reward. The other task involved a spatial working memory task. In both cases HD cell activity in the anterior dorsal thalamus did not correlate very well with the animal's behavioral responses. Recent experiments have shown that the spatial information from HD cell activity closely matched an animal's response in the food-hoarding task, a task that requires path integration (Valerio and Taube, 2012). In addition, this study found that there were two types of correction processes that occurred when the animals made an error in their homeward path. With small behavioral errors, the cells' preferred directions were reset to their original orientations upon the animal's return to the home base. In contrast, when behavior errors were large (>50°) animals 'remapped' their spatial reference system to the cells' new orientations.

More recent studies have utilized the finding that direction-specific firing is disrupted when the animal is inverted and shown that the presence of the directional signal is necessary if the animal is to perform a spatial task that requires a flexible representation of space (Gibson et al., 2013). Although animals were able to perform a spatial task from an inverted orientation when they were released from one of the two locations, their performance fell to chance when released from a novel location. Simultaneous to these behavioral outcomes, direction-specific firing in HD cells was impaired throughout the task, suggesting that their firing was necessary when a spatial mapping strategy was needed (i.e., cognitive map), but not when the task relied more on a relatively simple beacon or response strategy.

Summary

In summary, HD cells represent a high-level spatial signal that encodes the animal's directional heading. HD cells have been identified in several brain areas – mostly within the limbic system, but the functional role served by each brain area remains to be elucidated. HD cells receive multimodal sensory and motor information. It appears that the signal arises from the lateral mammillary nuclei, or subcortical structures afferent to it, and that both vestibular and motor information play important roles in generating the signal. Visual landmarks can exert control over the preferred firing direction of an HD cell. The nature of the HD cell signal makes it an ideal candidate for studying the neural mechanisms underlying navigation.

See also: Human Spatial Orientation, Neural Basis of; Short-Term Memory: Psychological and Neural Aspects; Spatial Cognition.

Bibliography

Bassett, J.P., Taube, J.S., 2001. Neural correlates for angular head velocity in the rat dorsal tegmental nucleus. Journal of Neuroscience 21, 5740–5751.

Blair, H.T., Sharp, P.E., 1995. Anticipatory head direction signals in anterior thalamus: evidence for a thalamocortical circuit that integrates angular head motion to compute head direction. Journal of Neuroscience 15, 6260–6270.

Clark, B.J., Valerio, S., Taube, J.S., 2011. Disrupted Grid and Head Direction Cell Signal in the Entorhinal Cortex and Parasubiculum after Lesions of the Head Direction System. Program No. 729.11. 2011 Abstract Viewer/Itinerary Planner. Society for Neuroscience, Washington, DC (Online).

Gibson, B.M., Butler, W.N., Taube, J.S., 2013. The head-direction signal is critical for navigation requiring a cognitive map but not for learning a spatial habit. Current Biology 23, 1536–1540.

Golob, E.J., Taube, J.S., 1997. Head direction cells and episodic spatial information in rats without a hippocampus. Proceedings of the National Academy of Sciences USA 94, 7645–7650.

Muir, G.M., Taube, J.S., 2002. The neural correlates of navigation: do head direction and place cells guide spatial behavior? Behavioral and Cognitive Neuroscience Reviews 1, 297–317.

Muir, G.M., Brown, J.E., Carey, J.P., Hirvonen, T.P., Della Santina, C.C., Minor, L.B., Taube, J.S., 2009. Semicircular canal occlusions disrupt head direction cell activity in the anterordorsal thalamus of chinchillas. Journal of Neuroscience 29, 14521–14533.

Robertson, R.G., Rolls, E.T., Georges-François, P., Panzeri, S., 1999. Head direction cells in the primate pre-subiculum. Hippocampus 9, 206–219.

Sharp, P.E., Blair, H.T., Cho, J., 2001a. The anatomical and computational basis for the rat head-direction cell signal. Trends in Neuroscience 24, 289–294.

Sharp, P.E., Tinkelman, A., Cho, J., 2001b. Angular velocity and head direction signals recorded from the dorsal tegmental nucleus of gudden in the rat: implications for path integration in the head direction cell circuit. Behavioral Neuroscience 115, 571–588.

Shinder, M.E., Taube, J.S., 2011. Active and passive movement are encoded equally by head direction cells in the anterodorsal thalamus. Journal of Neurophysiology 106, 788–800.

Stackman, R.W., Golob, E.J., Bassett, J.P., Taube, J.S., 2003. Passive transport disrupts directional path integration by rat head direction cells. Journal of Neurophysiology 90, 2862–2874.

Stackman, R.W., Taube, J.S., 1997. Firing properties of head direction cells in rat anterior thalamic neurons: dependence upon vestibular input. Journal of Neuroscience 17, 4349–4358.

Stackman, R.W., Taube, J.S., 1998. Firing properties of rat lateral mammillary nuclei single units: head direction, head pitch, and angular head velocity. Journal of Neuroscience 18, 9020–9037.

Taube, J.S., 1995. Head direction cells recorded in the anterior thalamic nuclei of freely moving rats. Journal of Neuroscience 15, 70–86.

Taube, J.S., 2007. The head direction signal: origins and sensory-motor integration. Annual Reviews of Neuroscience 30, 181–207.

Taube, J.S., 2011. Head direction cell firing properties and behavioural performance in 3D space. Journal of Physiology 589, 835–841.

Taube, J.S., Burton, H.L., 1995. Head direction cell activity monitored in a novel environment and during a cue conflict situation. Journal of Neurophysiology 74, 1953–1971.

Taube, J.S., Muller, R.U., Ranck, J.B., 1990a. Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. Journal of Neuroscience 10, 420–435.

Taube, J.S., Muller, R.U., Ranck, J.B., 1990b. Head direction cells recorded from the postsubiculum in freely moving rats. II. Effects of environmental manipulations. Journal of Neuroscience 10, 436–447.

Valerio, S., Taube, J.S., 2012. Path integration: how the head direction signal maintains and corrects spatial orientation. Nature Neuroscience 15, 1445–1453.

Wiener, S.I., Taube, J.S., 2005. Head Direction Cells and the Neural Mechanisms of Spatial Orientation. MIT Press, Cambridge, MA.

Yoder, R.Y., Clark, B.J., Taube, J.S., 2011. Origins of landmark encoding in the brain. Trends in Neuroscience 34, 561–571.

Yoder, R.M., Taube, J.S., 2009. Head direction cell activity in mice: robust directional signal depends on intact otoliths. Journal of Neuroscience 29, 1061–1076.

Neural Representations of Objects

Michael J Tarr and John A Pyles, Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.J. Tarr, volume 15, pp. 10563–10567, © 2001, Elsevier Ltd.

Abstract

Large regions of occipito-temporal cortex are recruited during the visual perception of objects. Here we review data from single unit (individual neurons), and neuroimaging studies to consider how objects are encoded in cortex, and the organization of the cortical machinery supporting object perception.

representation, *n.*[1]

> *… Something which stands for or denotes another symbolically; an image, a symbol, a sign. Chiefly with of ….*
>
> Oxford English Dictionary online edition; http://www.oed.edu

Within cognitive science, mental representations are taken to be high-level, manipulable entities that 'stand for' things in the world. More narrowly, in the study of vision, object representations are taken to be *spatial* in that they encode the approximate shape properties of physical objects (Pinker, 1985). In either case, although we (and the field) may have strong intuitions about what we mean by 'representation,' specifying the actual content of mental representations is rather difficult, with even fundamental concepts such as what constitutes a visual 'feature' being severely under-constrained. When discussing the *neural representation* of objects, things become murkier still.

What does it mean to neurally 'represent' a visual object? Can a neural signal emanating from one neuron or millions of neurons ever 'stand for' an object? It is our contention that the neural responses our field refers to as 'object representations' are nothing of the sort. For example, single neuron spikes of 'face cells' do not stand for faces (Tsao et al., 2006) anymore than the summed responses of millions of neurons, as measured indirectly by fMRI, across the 'extrastriate body area' stand for human bodies (Downing et al., 2001). Although others have argued against such results on theoretical grounds – in particular, suggesting that these nominally domain-specific responses are actually part of more domain-general processing systems (Tarr and Gauthier, 2000) – our argument here is a philosophical one: mental constructs cannot be reduced to localized patterns of neural activity. That is, in reviewing how neural structures support object perception and processing, we should always consider experimental results in a context whereby neurons, populations of neurons, and neural systems carry and encode information, but in which this information, in and of itself, is not about/of objects.

How Small Numbers of Neurons (May) Code Information about Objects

One of the most commonly measured units of 'representation' in the brain is the neuron. Somewhat surprisingly, some functional information about objects seems to be coded at this incredibly fine scale. Single-unit recordings of neurons in the ventral visual cortex reveal activity that is often tuned to individual objects or classes of objects. That is, a particular stimulus object presented in the receptive field of a given neuron will produce significantly greater activity (more spikes per second) in that neuron as compared to any other tested stimulus (typically a varied assortment of natural and artifactual objects). For example, neurons in many cortical areas along the ventral pathway – the part of the primate brain that is thought to support high-level vision, including object recognition – are highly selective for object categories that have been previously presented (Sheinberg and Logothetis, 2001). Similarly, cortical neurons may be specifically selective to individual objects within a category, at least when the stimuli are very familiar to the monkey, having been seen hundreds of times previously (Logothetis and Sheinberg, 1996). And lastly, there is extensive evidence that many different regions (or 'patches') along the primate ventral pathway are selective for faces (Tsao et al., 2006). At the same time, even putting aside the question of how this single-neuron selectivity translates into a visual understanding of objects, few researchers believe that single neurons code for individual objects. Rather, it is commonly held that large populations of neurons carry information objects or object classes in a distributed fashion. This is nicely demonstrated in a recent study that explored how long-term experience with objects affects the response properties of both excitatory and inhibitory neurons within the inferior temporal region of the ventral pathway. Critically, experience had opposite effects on the two types of neurons, with excitatory neurons showing an increase in their maximal responses to their preferred highly-familiar objects and inhibitory neurons showing a decrease both in their maximal response to highly-familiar objects and in their average rates of response (Woloszyn and Sheinberg, 2012). Taken together, these two effects indicate that information about familiar objects is carried by a population of both excitatory and inhibitory neurons, whereby a 'balance' is maintained between excitatory neurons narrowly encoding information about object familiarity and inhibitory neurons broadly encoding information about object novelty. A different sort of evidence for population codes – somewhat indirect – is provided by single-neuron studies that systematically explore which separable object features yield neural activity equivalent to the maximal response observed for the complete object. In many instances,

the maximal response is maintained when highly schematic features are presented, suggesting that individual neurons must work in concert (Tanaka, 1996).

To the extent that individual neurons carry information derived from (images of) objects, there remains the question of format (Marr and Nishihara, 1978). Information can be more or less invariant over changes in the appearance of the object in the image. Ideally, perception and recognition should remain robust when an object changes its configuration, or is moved, rotated, or illuminated differently. Many object-selective neurons exhibit some invariance: the level of response to a preferred object is not dramatically affected by changes in the size or the spatial position of the object for which a given neuron is selective (Desimone et al., 1984). On the other hand, these same neurons often show great sensitivity to changes in pose or illumination direction. For example, a large percentage of object-selective neurons are also 'view-tuned' in that they show the highest level of activity to familiar objects in specific viewpoints (Logothetis and Sheinberg, 1996). Face-selective neurons may respond most strongly when presented with a frontal view of a face, but show a progressively diminished response as the face rotates away from this viewpoint (Perrett et al., 1985). Similarly, in the visual cortex of monkeys taught to recognize novel objects, neurons that become selective for these objects are typically highly tuned to the particular viewpoints used during training (Logothetis and Pauls, 1995). These findings are consistent with much of the behavioral data on object recognition, where recognition performance is often strongly dependent on the familiarity of a given viewpoint (Tarr and Bülthoff, 1998). Thus, at the finest scale of analysis, single neurons participate in a code that is generally invariant over size and location, but are sensitive to viewpoint and other changes in the image.

Hierarchical Processing in the Visual Cortex

How can single neurons or populations of neurons code the complex patterns and feature conjunctions present in almost all objects? Many models of object recognition assume that the answer is that object- or feature-tuned neurons are the endpoints of a hierarchy of progressively more and more complex visual codes (*see* Visual Perception, Neural Basis of). Supporting this approach, neurons in the earliest visual areas (*see* Occipital Lobe (Visual Cortex): Functional Aspects) respond to very simple, more generic properties of objects, for instance, oriented edges or blobs of color. As one moves 'up' the visual pathway (*see* Visual Streams: Dorsal and Ventral) the coding becomes more specific, with neurons at higher levels generally showing selective responses to more complex properties of objects than the level(s) below them (Van Essen and Maunsell, 1983). The principle behind this hierarchy is straightforward: from one level to the next, neurons respond only if they receive input from multiple neurons, thereby creating selectivity that is necessarily a product of a given neuron's inputs. This many-to-one principle not only pools information in the straightforward sense, but because of the pattern of connectivity from one layer to the next, assembles new and more complex features as one moves up the hierarchy. Given the strong evidence for this sort of feed-forward processing in the visual system, many computational models, even those that are quite dissimilar in other respects, have adopted a hierarchical processing architecture to capture complex object features (Marr and Nishihara, 1978; Riesenhuber and Poggio, 1999). Although the hierarchical approach is appealing, some theorists have pointed out that assembling complex features in this fashion may lead to overly selective neurons. That is, given too precise information about an object image, a neuron or population of neurons might be incapable of generalizing to new viewing conditions, even with only slight changes in the image.

To address this need for invariance over changes in the image, some computational models have adopted a hierarchical architecture in which there is an alternation between conjunctions of features – the standard hierarchical model – and disjunctions of features (Riesenhuber and Poggio, 1999). For example, one layer might consist of neurons that respond only if they receive input from a *collection* of neurons coding for simpler features, while the next layer might consist of neurons that respond if they receive input from *any* of several neurons. This sort of hierarchical coding is exemplified by a single neuron that codes for a three-dimensional object regardless of its orientation, but only by virtue of the fact that this view-invariant neuron is driven by any one of several view-tuned neurons, each coding for a different view of the same object (Logothetis and Sheinberg, 1996).

The Building Blocks of Object Selectivity

The hierarchical organization of the visual cortex suggests that information about objects takes the form of a 'vocabulary' of features constructed from simpler features (Marr and Nishihara, 1978). This assumption implies that somewhere between orientation-selective neurons in early vision and object-selective neurons at the highest levels of the ventral visual pathway, are neurons encoding information about the critical features that distinguish one object from another. Under one view, such features should look something like object parts; intuitively the first level of decomposition when we perceive objects (Marr and Nishihara, 1978). There is, however, little evidence to indicate that neurons coding for parts, as we typically construe them, are the precursors to object-selective neurons. When the responses of neurons in intermediate and high-level visual areas are systematically decomposed, they show consistent responses to patterns that are more complex than edges, but less complex than objects. However, these 'features' often appear arbitrary, coding for odd, almost random shape properties. For example, one neuron might respond most strongly to a lollipop shape, while another neuron might respond to an eight-point star (Tanaka, 1996). While such features may appear to be random or, at best, schematic versions of certain objects, they are not arrayed haphazardly across the visual cortex. Rather, as with most visual areas, there is a columnar organization in which neurons within a column all code for the same feature and adjacent columns tend to code for similar features, albeit with columnar boundaries sometimes representing dramatic changes in feature selectivity (Tanaka, 1996).

In contrast, more recent work has investigated responses of IT neurons in a principled way by morphing three-dimensional novel stimuli to map the tuning of neurons (Yamane et al., 2008). These data suggest that shape-selective

neurons may be capturing three-dimensional aspects of objects through selectivity to curvature-based features, as well their configurations. More recently, a study following up on this result found evidence for neural selectivity to both medial axis shape and surface properties (Hung et al., 2012), suggesting that high-level selectivity to objects may be based on an object's skeletal shape and surface appearance. Interestingly, this account is consistent with earlier computational models of object representation in which an object's medial axis and surface features comprised the essential 'building blocks' for describing objects (Marr and Nishihara, 1978; Blum, 1967).

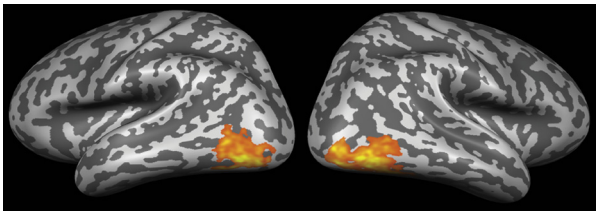

Figure 1 A representative map on an inflated cortical reconstruction showing BOLD activity in the lateral occipital complex (LOC) selective for objects over scrambled objects.

How Large Numbers of Neurons Code Information about Objects

While significant advances in our understanding of high-level vision have arisen from studying the responses of single neurons, most theorists hold that the process of object perception is not captured by either individual neurons or small numbers of neurons. Object perception and recognition almost surely rely on population codes comprised of ten thousands or more neurons, with still larger populations recruited for the performance of specific tasks. Ideally we would like to measure the responses of many neurons spread across visual cortex simultaneously. Unfortunately, it is impossible to make such measurements using present-day single-unit or multi-unit recording methods. As an alternative, to explore the responses of large populations of neurons during object perception, functional MRI (fMRI) – which indirectly measures the activity of large populations of neurons using a blood oxygen level-dependent (BOLD) signal – has become a popular neuroimaging technique applicable to human subjects. The entire human brain can be imaged simultaneously, collecting signals from 100 000 to 400 000 'voxels' (volume pixels), the resolution of each voxel being between 2 and 3 mm^3 – a volume that covers approximately 10^6 neurons.

Object Selectivity in Human Visual Cortex

To investigate whether there exist populations of object-selective neurons within the human ventral visual cortex researchers have compared the BOLD signal evoked when viewing common objects to the signal evoked when viewing scrambled versions of those same objects or texture patterns (Malach et al., 1995; Grill-Spector et al., 1998). A typical result is that a region of lateral occipital cortex located somewhat posterior and inferior to motion-selective area MT/V5 exhibits a greater response to intact objects as compared to scrambled objects. In contrast, lower-level visual areas such as V1 show no such difference. This object selective region, identified consistently in many other studies (e.g., Kourtzi and Kanwisher, 2000), has been referred to as the lateral occipital complex (LOC) (**Figure 1**). But what is the nature of information about objects encoded within the LOC and what role does such information serve in the context of the larger object processing hierarchy? Although single neurons are often sensitive to image variations such as viewpoint or lighting, behaviorally we are capable of recognizing an object across such variations. As the pooled response of 1 000 000 of neurons, it is possible that LOC functionally encodes information about objects in an invariant manner that reflects our phenomenological experience (although it is worth noting that even when our recognition seems effortless, we are often slower or less accurate across novel image variations; Tarr and Bülthoff, 1998). Using fMRI 'adaptation' techniques in which reduced responses over repeated trials denote sensitivity to the repeated stimulus property (or lack of sensitivity to a varying stimulus property), it has been found that regions of LOC are both size and position invariant (Grill-Spector et al., 1999). Similarly, LOC appears to be invariant with respect to whether an object is presented as a gray scale photograph or as a line drawing (Ishai et al., 2000; Kourtzi and Kanwisher, 2000). However, whether LOC is invariant over changes in viewpoint is less clear-cut, with the degree of adaptation being dependent on the magnitude of the pose change (Grill-Spector et al., 1999; James et al., 2002) – a result that is actually quite consistent with the single neuron data in which larger changes in viewpoint lead to larger decreases in a neuron's spike rate. Finally, adaptation has also revealed that LOC responds in an invariant manner to moving objects (Pyles and Grossman, 2009), and to changes in lower-level features such as contours (Kourtzi and Kanwisher, 2001). Overall, there is accumulating evidence that LOC encodes critical information that is integral to our perceptual experience of a stable world in which changes in our viewing position or in the configuration of the environment do not perturb our perception of objects.

Category Selectivity within the Ventral Visual Cortex

LOC appears to encode object information in domain-general manner – that is, without regard to the semantics associated with a given object (e.g., its category and its function). In contrast, one of the most consistent findings across neuroimaging studies of human ventral visual cortex has been that there exist localized regions that respond preferentially to specific object categories. The most studied of these category selective areas is a region of the mid-fusiform gyrus (mFus) that shows significantly higher responses when observers view faces as compared to when they view common, non-face objects (Kanwisher et al., 1997; Sergent and Signoret, 1992). Similar functionally-defined, category-selective cortical regions have been identified for letters, places/scenes, houses, human body parts, tools, and moving objects (Epstein and Kanwisher, 1998; Cohen et al., 2000; Downing et al., 2001; Grossman and Blake, 2002).

Thus ventral visual cortex seems to show an organization based for at least some object categories, on high-level visual semantics. Moreover, as with single-neuron studies, fMRI likewise reveals that some of these categories – notably faces and body parts – consistently evoke preferential responses in multiple brain areas and that these areas have consistent anatomical locations and spatial relationships with one another (Weiner and Grill-Spector, 2011; Pyles et al., 2013). Reiterating our earlier points, the existence of multiple category-selective brain regions for a single object category indicates that localized selectivity in and of itself does not denote phenomenological association with the perception of that category. Rather it seems likely that each localized brain region realizes a category-critical computation or computations within a larger functional network.

Large-Scale Coding of Objects

Beyond the idea that category-selective brain regions should be considered as participants in larger functional networks, the entire concept of category selectivity, as measured by fMRI, may be misleading. In particular, evidence for localized category-selective brain regions typically comes from comparing the responses associated with one object class to the responses associated with a second object class; for example, comparing face-related responses to house-related responses. Such comparisons, almost by definition, give rise to clustered, localized differences that mask the overall pattern of responses across much of the ventral visual pathway (Friston et al., 2006). In truth, *any* object category produces a complex pattern that is almost always different – on a voxel-by-voxel basis – from the response patterns arising for any other categories of objects (otherwise how else could we tell them apart?). Of course, although logically present, these differences are often subtle and relatively small compared to the large differences seen between highly-overlearned visual categories, such as faces, scenes, and words. Yet such subtle differences are likely to be critical elements in the neural code. Multivariate pattern analysis (MVPA) addresses this issue, considering the pattern of neural responses associated with an object or object category (Norman et al., 2006). Beyond enabling exploration of more subtle aspects of object coding, MVPA allows one to move beyond studying neural responses only in functionally localized or neuroanatomically-defined brain regions, instead using systematic 'searchlight' methods to interrogate the neural coding of information across the brain (Kriegeskorte et al., 2006). For example, one study using MVPA observed complex response patterns across large portions of ventral visual cortex that reflected different high-level properties of objects, such as animacy, natural kind status, and human/not human (Kriegeskorte et al., 2008). Still more ambitious studies have mapped out the neural encoding of high-level semantic space for a wide variety of object and action categories (Huth et al., 2012). Interestingly, in the former study, using the same object images in both human fMRI and primate neurophysiology, neurons in primate visual cortex showed a similar pattern of featural organization to that revealed using fMRI (Kriegeskorte et al., 2008). More generally, exploring the subtle response patterns of large-scale neural populations in an unrestricted manner across the entire brain has enabled the identification of functional regions involved in object perception that might have otherwise been overlooked. For example, a region of anterior temporal lobe was recently identified as playing an important role in the processing of identity in human faces using MVPA (Nestor et al., 2008, 2011). Overall, the application of modern machine learning and statistical methods – as exemplified in MVPA – promises to shed more light on the featural basis of object perception as realized both throughout the brain and, particularly, within the ventral visual cortex.

Organizing Principles of Object Selectivity

One question that is rarely asked with respect to the organization of ventral visual cortex is *why* it is organized as it is? That is, why do some object categories give rise to category-selective responses, while other categories appear to produce only more distributed patterns of responses? Moreover, in that humans, almost all exhibit a similar functional organization across ventral visual cortex, why do particular category-selective regions consistently localize to approximately the same neuroanatomical locations across individuals? There are at least two accounts for this common neural organization across individuals. Under one account, different category-selective brain regions are functionally fixed as domain specific. That is, regions are category-selective for faces, bodies, or scenes because this high-level selectivity is intrinsic to our neural architecture (Kanwisher, 2010). In contrast, an alternative account posits that a variety of low-level constraints shape which visual inputs are processed in different regions of the ventral visual cortex. That is, given the structure of the input, task demands, and connectivity constraints, particular visual object categories are processed and encoded in brain regions that are biased toward those image structures, tasks, and connectivity to support further processing.

For example, one recent fMRI study found that face-selective regions, including mFus, are recruited when observers view displays with predominantly curved contours, but that scene-selective brain regions are recruited when observers view displays with predominantly straight contours (Tootell et al., 2012). Thus, low-level image statistics, including curved versus straight, appear to play an important role in determining which high-level brain mechanisms ultimately end up processing different object categories, for instance, faces and scenes. With respect to task, a second recent fMRI study manipulated whether observers were trained to recognize novel objects at the individual or the basic (categorical) levels. Critically, individuation training produced an increase in object-selective neural responses within the mFus, while categorical training produced an increase in object-selective neural responses along the medial portion of the ventral occipito-temporal cortex (Wong et al., 2009). Thus, the particular task associated – by default – with a specific object category appears to also play some role in determining which high-level brain mechanisms end up processing those categories. Finally, there is evidence that the visual word form area – a brain region selective for written letter strings – and the face-selective mFus localize to the left and the right hemispheres, respectively, because of competition between the two domains (Behrmann and Plaut, 2013). In particular, connectivity to left-lateralized brain areas

specialized for language processing seems to dictate that brain regions supporting visual reading skill are also co-localized to the left hemisphere. Thus, the ways in which different high-level brain mechanisms are functionally connected to other neural systems helps to determine which visual object categories are likely to end up being processed by those mechanisms. In sum, the answer to the question of why particular category-selective regions are located where they are in the brain may have more to do with a host of relatively low-level biases within the architecture of the visual system rather than any pre-specified high-level constraints on which domains are processed where.

The Functional Role of Localized Brain Regions

Beyond asking why the ventral visual cortex is organized as it is, we must ultimately address the question *how* these many neural structures combine to produce complex visual behavior. As we have discussed, neuroimaging methods are useful for localizing *where* in the brain particular computations occur. Moreover, neuroimaging has increasingly focused on the functional nature of these computations – that is, the specific roles different neural substrates play in producing a behavior. This more recent trend builds on neuropsychological methods that have been historically used to explore deficits in behavior as a direct consequence of brain injuries to particular brain regions. For example, the existence of category-selective brain regions sensitive to faces was predicted by earlier neuropsychological research in which damage to regions such as mFus were observed to impair face recognition. Similarly, wider damage within the ventral visual cortex can result in the loss of one's ability to visually recognize many or all object classes – a syndrome known as *agnosia* (*see* Agnosia). Although these patients are grouped under this single label, agnosia includes damage to many different regions of the visual cortex and many different visual recognition deficits.

One of the ways in which neuropsychology has informed our understanding of the neural encoding of objects, has been through the deficits associated with damage to specific brain regions. As with the neuroimaging literature, the identification of category-specific visual recognition deficits has been taken as evidence for computationally 'modular' brain structures dedicated to singular visual domains. To take one example, if this modularity argument is correct for face processing, then damage to the brain region specialized for faces should result in deficits that are exclusive to face recognition and no deficits for non-face object recognition. Such a pattern would imply that the damaged region was functionally domain-specific (for faces). Thus, knowledge regarding the location of a brain injury, in tandem with a pattern of sparing and loss in behavior, may help reveal the functional roles of intact and damaged brain regions (*see* Cognitive Neuropsychology, Methods of). As mentioned, for face processing, patients with injuries to their ventral visual cortex often suffer a specific form of agnosia referred to as *prosopagnosia* (*see* Prosopagnosia) in which face recognition appears to be disproportionally impaired relative to object recognition (Farah, 1990). At first blush, the existence of prosopagnosia would seem to reinforce the idea of category-specific brain regions exclusively sensitive to faces. Indeed, it is known that damage to the right mFus in particular will typically give rise to impairment of visual face recognition. However, prosopagnosia can arise as a consequence of damage to brain regions other than mFus (Barton, 2008). Even more telling is the fact that developmental prosopagnosia – that is, prosopagnosia that does not occur as a result of brain damage – can also arise due to structural connectivity differences within the ventral visual cortex (Thomas et al., 2009). Indeed, individuals with developmental prosopagnosia typically show *normal* mFus responses as measured by fMRI, yet impaired ability to individuate faces. Thus, one must be cautious in attributing category-specific impairments to damage to putatively category-specific brain regions. Reinforcing the need for caution in interpreting neuropsychological patterns of sparing and loss, there are studies demonstrating that when prosopagnosic patients perform difficult individual-level discriminations with non-face objects (e.g., snowflakes), they show equal impairment with faces and non-faces (Gauthier et al., 1999). Thus, the nominally face-specific mechanisms damaged in cases of prosopagnosia are actually implicated in the processing of non-face objects, suggesting that, as with the category-selective responses seen in neuroimaging, category-specific deficits should be considered in a larger context of tasks and more complex brain networks.

See also: Agnosia; Cognitive Neuropsychology, Methods of; Occipital Lobe (Visual Cortex): Functional Aspects; Prosopagnosia; Visual Perception, Neural Basis of; Visual Streams: Dorsal and Ventral.

Bibliography

Barton, J.J.S., 2008. Structure and function in acquired prosopagnosia: lessons from a series of 10 patients with brain damage. Journal of Neuropsychology 2 (Pt 1), 197–225.

Behrmann, M., Plaut, D.C., 2013. Distributed circuits, not circumscribed centers, mediate visual recognition. Trends in Cognitive Sciences 17 (5), 210–219.

Blum, H., 1967. A transformation for extracting new descriptors of shape. In: Wathen-Dunn, W. (Ed.), Models for the Perception of Speech and Visual Form. MIT Press, Cambridge, MA, pp. 362–380.

Cohen, L., Dehaene, S., Naccache, L., Lehericy, S., Dehaene-Lambertz, G., Henaff, M.A., Michel, F., 2000. The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. Brain: A Journal of Neurology 123 (Pt 2), 291–307.

Desimone, R., Albright, T.D., Gross, C.G., Bruce, C., 1984. Stimulus-selective properties of inferior temporal neurons in the macaque. The Journal of Neuroscience 4 (8), 2051–2062.

Downing, P.E., Jiang, Y., Shuman, M., Kanwisher, N., 2001. A cortical area selective for visual processing of the human body. Science 293 (5539), 2470–2473.

Epstein, R., Kanwisher, N., 1998. A cortical representation of the local visual environment. Nature 392 (6676), 598–601.

Farah, M.J., 1990. Visual Agnosia: Disorders of Object Recognition and what They Tell Us about Normal Vision. MIT Press, Cambridge, MA.

Friston, K.J., Rotshtein, P., Geng, J.J., Sterzer, P., Henson, R.N., 2006. A critique of functional localisers. NeuroImage 30 (4), 1077–1087.

Gauthier, I., Behrmann, M., Tarr, M.J., 1999. Can face recognition really be dissociated from object recognition? Journal of Cognitive Neuroscience 11, 349–370.

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., Malach, R., 1999. Differential processing of objects under various viewing conditions in the human lateral occipital complex. Neuron 24 (1), 187–203.

Grill-Spector, K., Kushnir, T., Hendler, T., Edelman, S., Itzchak, Y., Malach, R., 1998. A sequence of object-processing stages revealed by fMRI in the human occipital lobe. Human Brain Mapping 6 (4), 316–328.

Grossman, E.D., Blake, R., 2002. Brain areas active during visual perception of biological motion. Neuron 35 (6), 1167–1175.

Hung, C.-C., Carlson, E., Connor, C., 2012. Medial axis shape coding in macaque inferotemporal cortex. Neuron 74 (6), 1099–1113.

Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76 (6), 1210–1224.

Ishai, A., Ungerleider, L.G., Martin, A., Haxby, J.V., 2000. The representation of objects in the human occipital and temporal cortex. Journal of Cognitive Neuroscience 12 (Suppl. 2), S35–S51.

James, T.W., Humphrey, G.K., Gati, J.S., Menon, R.S., Goodale, M.A., 2002. Differential effects of viewpoint on object-driven activation in dorsal and ventral streams. Neuron 35 (4), 793–801.

Kanwisher, N., 2010. Functional specificity in the human brain: a window into the functional architecture of the mind. Proceedings of the National Academy of Sciences of the United States of America 107 (25), 11163–11170.

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. The Journal of Neuroscience 17 (11), 4302–4311.

Kourtzi, Z., Kanwisher, N., 2000. Cortical regions involved in perceiving object shape. The Journal of Neuroscience 20 (9), 3310–3318.

Kourtzi, Z., Kanwisher, N., 2001. Representation of perceived object shape by the human lateral occipital complex. Science (New York) 293 (5534), 1506–1509.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America 103 (10), 3863–3868.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., et al., 2008. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60 (6), 1126–1141.

Logothetis, N.K., Pauls, J., 1995. Psychophysical and physiological evidence for viewer-centered object representation in the primate. Cerebral Cortex 3, 270–288.

Logothetis, N.K., Sheinberg, D.L., 1996. Visual Object Recognition. Annual Review of Neuroscience. Annual Reviews, Palo Alto, CA.

Malach, R., Reppas, J.B., Benson, R.R., Kwong, K.K., Jiang, H., Kennedy, W.A., et al., 1995. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proceedings of the National Academy of Sciences of the United States of America 92 (18), 8135–8139.

Marr, D., Nishihara, H.K., 1978. Representation and recognition of the spatial organization of three-dimensional shapes. Proceedings of the Royal Society of London B, Biological Sciences 200, 269–294.

Nestor, A., Plaut, D.C., Behrmann, M., 2011. Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. Proceedings of the National Academy of Sciences of the United States of America 108 (24), 9998–10003.

Nestor, A., Vettel, J.M., Tarr, M.J., 2008. Task-specific codes for face recognition: how they shape the neural representation of features for detection and individuation. PLoS ONE 3 (12).

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences 10 (9), 424–430.

Perrett, D.I., Smith, P.A.J., Potter, D.D., Mistlin, A.J., Head, A.S., Milner, A.D., Jeeves, M.A., 1985. Visual cells in the temporal cortex sensitive to face view and gaze direction. Proceedings of the Royal Society of London B, Biological Sciences 223, 293–317.

Pinker, S., 1985. Visual Cognition. The MIT Press, Cambridge, MA.

Pyles, J.A., Grossman, E.D., 2009. Neural adaptation for novel objects during dynamic articulation. Neuropsychologia 47 (5), 1261–1268.

Pyles, J.A., Verstynen, T.D., Schneider, W., Tarr, M.J., 2013. Explicating the face perception network with white-matter connectivity. PLoS ONE 8 (4).

Riesenhuber, M., Poggio, T., 1999. Hierarchical models of object recognition in cortex. Nature Neuroscience 2, 1019–1025.

Sergent, J.J., Signoret, J.L.J., 1992. Functional and anatomical decomposition of face processing: evidence from prosopagnosia and PET study of normal subjects. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 335 (1273), 52–55.

Sheinberg, D.L., Logothetis, N.K., 2001. Finding familiar objects in real-world scenes: the role temporal cortical neurons in natural vision. The Journal of Neuroscience 21, 1340–1350.

Tanaka, K., 1996. Inferotemporal Cortex and Object Vision. Annual Review of Neuroscience. Annual Reviews, Palo Alto, CA.

Tarr, M.J., Bülthoff, H.H., 1998. Object Recognition in Man, Monkey, and Machine. MIT Press, Cambridge, MA.

Tarr, M.J., Gauthier, I., 2000. FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. Nature Neuroscience 3, 764–769.

Thomas, C., Avidan, G., Humphreys, K., Jung, K.-J., Gao, F., Behrmann, M., 2009. Reduced structural connectivity in ventral visual cortex in congenital prosopagnosia. Nature Neuroscience 12 (1), 29–31.

Tootell, R.B., Nasr, S., Yue, X., 2012. Selectivity for Different Shapes Drives 'Face Selective' and 'Scene Selective' Areas in Human Visual Cortex (Annual Meeting of the Society for Neuroscience. New Orleans, LA).

Tsao, D.Y., Freiwald, W.A., Tootell, R.B., Livingstone, M.S., 2006. A cortical region consisting entirely of face-selective cells. Science 311 (5761), 670–674.

Van Essen, D.C., Maunsell, J.H.R., 1983. Hierarchical organization and functional streams in the visual cortex. Trends in Neuroscience 6, 370–375.

Weiner, K.S., Grill-Spector, K., 2011. Neural representations of faces and limbs neighbor in human high-level visual cortex: evidence for a new organization principle. Psychological Research.

Woloszyn, L., Sheinberg, D.L., 2012. Effects of long-term visual experience on responses of distinct classes of single units in inferior temporal cortex. Neuron 74 (1), 193–205.

Wong, A.C., Palmeri, T.J., Rogers, B.P., Gore, J.C., Gauthier, I., 2009. Beyond shape: how you learn about objects affects how they are represented in visual cortex. PloS One 4 (12), e8405.

Yamane, Y., Carlson, E.T., Bowman, K.C., Wang, Z., Connor, C.E., 2008. A neural code for three-dimensional object shape in macaque inferotemporal cortex. Nature Neuroscience 11 (11), 1352–1360.

Neural Synchrony as a Binding Mechanism

Wolf Singer, Max Planck Institute (MPI) for Brain Research, Frankfurt, Germany; Ernst Srüngmann Institute (ESI) for Neuroscience in Cooperation with Max Planck Society, Frankfurt, Germany; and Frankfurt Institute for Advanced Studies (FIAS), Frankfurt, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Binding operations, the dynamic establishment of semantic relations between distributed computational results, are key functions of cognitive processes in the brain. A common binding strategy is anatomical convergence of feature-encoding neurons onto conjunction-specific cells that respond only if the respective constellation of features is present. It is argued that this binding mechanism needs to be complemented by a more flexible mechanism that allows the brain to define relations on the fly in order to account for the nearly unbounded ability of evolved brains to represent the ever-changing constellations of elementary features characterizing perceptual objects. Evidence is reviewed which suggests that such dynamic encoding of relations is achieved by synchronization of temporally structured neuronal responses and hence by convergence in the time domain. A hallmark of neuronal responses is their oscillatory patterning that concentrates discharges to a particular phase of the oscillation cycle. By self-organized phase locking of these oscillatory responses the discharges of neurons can be synchronized. This in turn jointly enhances the saliency of the synchronized responses and establishes unequivocal relations among them. Neuronal oscillations cover a broad frequency range but here mainly the high-frequency oscillations in the beta- and gamma-frequency range are considered as these have been investigated most thoroughly in the context of binding functions.

The mammalian brain is a highly distributed system in which a large number of different processes are executed in parallel and there is no singular center for the joint evaluation of the numerous, simultaneously obtained computational results. This raises the question of how partial results of the distributed computations are bound together in order to support the generation of coherent percepts, the convergence of decision processes, the coordination of executive actions and last but not least the unity of conscious experience. Such binding problems present themselves at all levels of processing, both within and across modalities. The reason is that the brain's connectome (the comprehensive map of neural connections in the brain) has small world properties, allowing for very efficient communication between nodes irrespective of their anatomical distance. The prevailing objectives are reciprocity of connections both within and across processing streams (van den Heuvel and Sporns, 2011; Markov and Kennedy, 2013). In principle, there are two complementary strategies for the implementation of binding functions, i.e., for the encoding of relations between neuronal responses. The first relies on spatial convergence of afferents (neurons that carry signals from the sense organs to the central nervous system) onto common target cells. In this case relations between the responses of the feeding neurons are encoded by the conjunction-specific responses of the target cells. The second uses time as coding space for the expression of relations. In this case relations are expressed by temporal coordination of distributed responses. Thus, relations can be encoded by either conjunction in space or conjunction in time. The two mechanisms coexist and complement one another.

The binding of signals that occur very often in conjunction or whose contingent occurrence is of particular behavioral relevance is achieved mainly through conjunction-specific neurons that are formed by anatomical convergence in hierarchically organized architectures. The binding of infrequent or novel constellations of signals is preferentially achieved by the formation of temporally coherent assemblies of neurons whose configuration changes rapidly and in a context-dependent way due to dynamic modulation of the coupling between neurons. Perceptual objects are defined by a unique constellation of features whereby the diversity of possible constellations is virtually unlimited. This requires versatile binding operations that are difficult to realize with conjunction in space because of the combinatorial explosion of required connections between feeder cells and conjunction-specific neurons. Combinatorial problems of similar magnitude have to be solved for the programming of movements. Although the elementary components – the contraction of individual muscle fibers – are limited in number, the diversity of movements that can be generated by combining the elementary components is again virtually infinite. Evolved brains have acquired an extraordinary competence to solve such combinatorial problems and it appears that this is achieved by the combination of binding operations that exploit the options to encode relations both by conjunction in space and time.

Two Complementary Binding Strategies

In the primary visual cortex of mammals, relations among the responses of retinal ganglion cells are evaluated and represented by having the output of selected arrays of ganglion cells converge in diverse combinations onto individual cortical neurons. Distributed signals are bound together by conjunction-specific neurons that signal elementary feature constellations (Hubel and Wiesel, 1962). Through iteration of this strategy neurons in higher cortical areas acquire increasingly sophisticated response properties, representing complex constellations of features and, in special cases, even whole perceptual objects (Gross, 1992; Tanaka, 1997; Quian

 http://dx.doi.org/10.1016/B978-0-08-097086-8.55038-7

Quiroga et al., 2005). A similar organization is manifested in each of the other sensory modalities.

However, the strategy to represent relations explicitly by responses of specialized conjunction cells is constrained by the combinatorial explosion of the number of required binding units. This combinatorial problem is further accentuated by the fact that perceptual objects are often defined by conjunctions of features encoded in different sensory modalities. Most objects are characterized by a specific combination of visual, haptic (touch), or acoustic properties. Finally, difficulties arise if novel objects with unfamiliar feature constellations need to be represented because it appears only marginally feasible if not impossible to a priori generate a pool of conjunction-specific cells for the unfamiliar or to rearrange anatomical convergence on the fly in response to a novel object. These problems can be overcome by encoding relations through temporal conjunction because it permits utilization of the same neurons for the representation of different objects (Hebb, 1949). Here, the constellation of features characterizing perceptual objects is thought to be represented by the temporally coordinated activity of a dynamically associated ensemble of cells, whereby each of the participating neurons represents explicitly only subsets of features that characterize a particular perceptual object. Different objects can then be represented by recombining in various constellations neurons tuned to elementary features or their characteristic, frequently occurring combinations. The advantage of such distributed codes is that a particular cell can participate at different times in different assemblies. This sharing of neurons economized on neuron numbers and copes effectively with the huge combinatorial space occupied by real-world objects.

For assembly coding two constraints need to be met, however. First, a selection mechanism is required that permits dynamic, context-dependent association of neurons into distinct, functionally coherent assemblies. Second, grouped responses must get labeled so that they can be distinguished by subsequent processing stages as components of one coherent representation and do not get confounded with other unrelated responses.

Such labeling can be achieved by jointly raising the saliency of the selected neurons, because this ensures that they are processed and evaluated together at the subsequent processing stage. Neuronal systems have three options to increase the relative saliency of selected responses: First, non-grouped responses can be inhibited; second, the discharge frequency of the selected responses can be enhanced; and third, the selected cells can be made to discharge in precise temporal synchrony. Inhibition simply eliminates potentially confounding signals, enhanced firing raises the impact of the selected responses through temporal summation and synchronization raises the saliency of responses through spatial summation of synaptic potentials and exploits the exquisite sensitivity of neurons to coincident excitatory inputs (Singer, 1999). Modulation of discharge rates is a common and well-characterized signaling mechanism but it has certain disadvantages in the context of distributed coding. First, discharge rates of neurons depend on a large variety of variables and hence cannot signal unambiguously whether a cell has become part of an assembly. Increases in discharge rate resulting from changes in stimulus intensity and context would not be distinguishable from those signaling membership of a cell in an assembly. Ambiguities could also arise when different assemblies need to be configured simultaneously at the same processing stage. This may be the case when different objects and their relations to one another need to be represented simultaneously. If the corresponding assemblies were defined solely by enhanced discharge rate of their member neurons it would be difficult to distinguish which of the enhanced responses belongs to which assembly. These ambiguities can be reduced if the cells participating in an assembly synchronize their discharges with high temporal precision because synchrony can be adjusted independently of rate fluctuations (Lima et al., 2011). Moreover, relations between responses can then be defined with very high temporal resolution because synchronization enhances only the saliency of those discharges that are precisely synchronized and generate coincident synaptic potentials in target cells at the subsequent processing stage. In case of response selection by synchrony the selected events are individual spikes or brief bursts of spikes and not the overall increase in discharge rate. Thus, relations can be defined on short timescales. This permits temporal intercalation of different assemblies whereby the rate at which assemblies can be multiplexed without getting confounded is limited only by the duration of the interval over which synaptic potentials are integrated. This interval depends on the time constraints of neuronal membranes and for cortical neurons is in the range of 10–15 ms. These integration windows can be further shortened if neuronal networks engage in oscillations. In this case cells are exposed to periodic volleys of inhibition, which further shorten the window over which cells can effectively integrate excitatory inputs. With high oscillation frequencies (>40 Hz) these windows of susceptibility may be as short as a few milliseconds. Thus, different assemblies can coexist within the time span of perceptual presence without getting confounded.

Synchrony as a Signature of Relatedness

The discharge of neurons in the cerebral cortex reflects with great precision the temporal structure of stimuli. Synchronously presented stimuli evoke synchronous responses. This makes it possible to examine with psychophysical experiments whether synchrony of responses is indeed interpreted as a signature of relatedness. The results of biophysical studies indicate that the synchronous responses evoked by simultaneously presented stimuli are indeed bound perceptually and get segregated from the temporally offset responses evoked by asynchronously presented patterns of elements. In the visual system the temporal resolution for the perceptual binding of synchronous responses and the segregation of asynchronous responses is better than 10 ms (for review, see Singer, 1999). This suggests that precise temporal relations among the discharges of distributed neurons are exploited for perceptual grouping and segregation, respectively, and that offset intervals are used for segregation that are considerably shorter than those that allow one to determine the sequence order of rapidly presented stimuli.

In the mid-1980s, it was discovered that neurons distributed across the visual cortex can synchronize their discharges

with millisecond precision when activated by a single visual object (Gray and Singer, 1987, 1989). This synchronization is caused by dynamic interactions between cortical neurons that lead to phase locking of oscillations. Cortical networks get entrained in such oscillatory activity when in the appropriate state and activated by sensory stimuli: Typically, these stimulus-induced oscillations occur in the frequency range of 30–90 Hz, the so-called gamma-frequency band – and are due to local reciprocal interactions between excitatory and inhibitory neurons. Based on the theoretical considerations expounded above it had been proposed that this internal synchronization serves to establish relations between distributed responses by jointly raising their saliency and favoring their processing at subsequent processing stages. Since then numerous experiments have been performed in order to test predictions derived from this hypothesis (for review see Singer, 1999; Engel et al., 2001; Fries, 2009). This required demonstration that (1) synchronous activity is computationally relevant and has a stronger impact on target structures than nonsynchronous activity, (2) synchrony can be established and changed sufficiently rapidly to be compatible with processing speed, and (3) synchronization is involved in cognitive and executive processes requiring dynamic selection and binding of responses. Examples for such functions are feature binding and perceptual grouping, cross-modal integration, selection and binding of pattern elements by attention, temporary association of stored items in working memory, and the coordination of novel movements. All these functions have in common that relations have to be encoded on the fly and in ever-changing constellations between responses of widely distributed neurons on the backbone of the fixed network of anatomical connections – a function difficult to realize with conjunction-specific neurons that are specialized for the encoding of distinct relations.

Rapid Synchronization

Both simulation studies with spiking neurons and recordings from the brain revealed that networks could undergo very rapid transitions from uncorrelated to synchronized states (e.g., Gerstner, 1996; Vicente et al., 2008). One factor contributing to rapid synchronization is phase locking of oscillatory activity, which is an emergent property of networks consisting of reciprocally coupled neurons. Another mechanism enabling very fast and context-dependent synchronization of responses to sensory stimuli is based on correlated spontaneous oscillations. In the visual cortex neurons with related feature preferences are reciprocally coupled and this has the effect that they tend to oscillate synchronously even without stimulation (Fries et al., 2001a). These self-generated patterns of synchronized oscillations reflect the architecture and the weights of the intracortical association connections, which are held responsible for the grouping and binding of responses. Spontaneous activity can thus be considered as the reflection of a continuous readout of the grouping criteria that reside in the functional architecture of cortical connections. Because of these ordered patterns of self-generated oscillations grouping by synchronization can be extremely fast and occurs as a function of both the prewired associational dispositions and the current functional state of the cortical network. Upon arrival of retinal signals, cells already engaged in coherent oscillations instantaneously generate precisely synchronized discharges.

Relation to Perception

Stimulus-induced synchronization does not simply reflect anatomical connectivity but depends critically on the configuration of the stimuli. It is the result of a dynamic and context-dependent self-organization process that follows some of the Gestalt rules that govern perceptual grouping. In the visual domain this has been shown for the criteria of continuity, proximity, similarity in the orientation domain, colinearity, and common fate (Gray et al., 1989; Castelo-Branco et al., 2000; for review, see Singer, 1999; Uhlhaas et al., 2009).

Most of the early experiments on response synchronization were performed in anesthetized animals but more recent evidence indicates that internally generated synchrony is even more pronounced in the awake brain, exhibiting similar sensitivity to context (Kreiter and Singer, 1996; Buschman et al., 2012), and becoming particularly precise when the electroencephalogram (EEG) is desynchronized (Munk et al., 1996) and the animals are attentive (Roelfsema et al., 1997; Steinmetz et al., 2000; Fries et al., 2001b).

Direct relations between synchronization of oscillatory responses and cognitive functions have been found in strabismic amblyopia ('lazy eye') (Roelfsema et al., 1994), a developmental impairment of vision. Here synchronization of responses from the amblyopic eye is disturbed which accounts well for the reduced visual acuity, the perceptual distortions and the impaired figure-ground segregation in amblyopic vision, given that binding by synchrony serves perceptual grouping and disambiguation of contour ownership. Moreover, this finding agrees well with observations that entrainment of cortical networks in synchronous gamma oscillations reduces the variance of responses and increases signal-to-noise ratios (Rodriguez et al., 2004).

Another close correlation between response synchronization and perception has been found in primary visual cortex for the phenomenon of binocular rivalry (Fries et al., 1997). Responses from the eye that won in interocular competition increased their synchronicity while the reverse was true for responses from the eye that became suppressed. Interestingly, however, there were no rivalry-related changes in the discharge frequency of the neurons. This supports the notion that the saliency of responses can be enhanced either by increasing discharge rate or synchrony. The perceptual equivalence of the two saliency-enhancing mechanisms has recently been demonstrated (Biederlack et al., 2006). Evidence from noninvasive EEG and magneto-encephalographic (MEG) recordings in human subjects indicates further that access to consciousness is gated not only by enhanced sensory evidence and stimulus saliency but also – and to a crucial extent – by the synchronization of neuronal responses in extended cortical networks (Melloni et al., 2007; Gaillard et al., 2009).

The State Dependence of Response Synchronization

Converging evidence from human and animal studies indicates that synchronization of high-frequency oscillations in the beta- and gamma-frequency range is exquisitely state-dependent and increases with arousal, attention and expectancy to act. Preparation for the execution of sensory-motor tasks is associated with the formation of widely distributed cortical networks characterized by phase synchronization in the beta- and/or gamma-frequency range (Roelfsema et al., 1997; Womelsdorf et al., 2007). The expectancy of having to react to a visual stimulus leads to a massive enhancement of synchronized gamma oscillations in the primary visual cortex (Lima et al., 2011) and focusing attention is accomplished by the formation of synchronized networks comprising prefrontal and parietal cortical areas (Gregoriou et al., 2009, 2012; Buschman and Miller, 2007). In early visual areas responses to attended stimuli get entrained in gamma oscillations and phase lock with responses in subsequent areas of the processing hierarchy (Fries et al., 2001b; Bosman et al., 2012). It is assumed that this attention-dependent synchronization serves response selection by enhancing the saliency of selected responses and facilitates propagation of these responses across subsequent processing stages by phase locking of oscillatory activity (Fries, 2005).

Disturbed Synchronization in Psychiatric Diseases

Some of the cognitive deficits characteristic for mental diseases such as schizophrenia and autism spectrum disorder suggest abnormalities in binding functions and large-scale coordination of cortical networks. Motivated by the evidence that synchronization of oscillatory activity has numerous functions in sensory processing and also plays a major role in the formation of functional networks supporting working memory (Miltner et al., 1999) and that these functions are severely impaired in these diseases, patients were studied with EEG and MEG in search of abnormalities in oscillations and synchrony. The results obtained so far are unequivocal. Both diseases are associated with marked anomalies of these dynamic variables (for review see Uhlhaas and Singer, 2006, 2012). Patients show deficits in the ability to generate high-frequency oscillations and/or to synchronize oscillatory activity across cortical areas and these deficits correlate with the severity of the clinical symptoms. It is hoped that further study of the electrophysiological signatures will shed new light on the pathophysiology of these mental diseases and eventually lead to the identification of reliable endophenotypes.

Epilogue

The discovery of stimulus-induced response synchronization and the proposal that synchronization might serve as a neuronal code for the definition of relations have solicited highly controversial debates in the scientific community. Arguments raised in this context have been summarized in a series of reviews published in the journal *Neuron* (1999, vol. 24: pp. 7–125). These contributions focus on the putative role of synchronization in visual perception as the early studies have investigated mainly sensory processes. In the meantime oscillatory patterning and synchronization of responses have been identified as ubiquitous phenomena not only in and between areas of the cerebral cortex but also in the cerebellum and in subcortical structures such as the hippocampus and the basal ganglia. Oscillation frequencies vary across different structures and often one observes cross-frequency coupling; i.e., a fast oscillation is modulated in phase and/or amplitude by a slower oscillatory process (Canolty et al., 2010). It is suggested that such concatenation of rhythms serves the encoding of nested relations.

Taken together complementary evidence from animal studies and electrophysiological recordings from human subjects suggest that an oscillatory patterning of neuronal activity in the beta- and gamma-frequency band and the synchronization of this oscillatory activity serve a wide diversity of functions, such as feature binding and perceptual grouping, attention-dependent input selection, the dynamic formation of functional networks in the context of attentional control, working memory, and the coordination of movements (von der Malsburg et al., 2010; Uhlhaas et al., 2009).

At present, most of the evidence for the functional role of oscillations and synchrony is still correlative in nature as is the case also for the relations between other electrophysiological variables and behavior. The reason is the difficulty to selectively interfere with synchronization without affecting other response variables. It is undeniable, however, that the new approach – to register neuronal activity simultaneously from different sites and to search for temporal relations among the activity of distributed neuronal populations, rather than measuring only the rate variations of individual cells – provides new and fascinating insights into the dynamics of neuronal processes. In a number of investigated paradigms, changes in temporal correlations among distributed neuronal responses turned out to be a better predictor of stimulus configurations, of attentional states and of the outcome of a perceptual process than the rate changes of individual neurons. This suggests that the nervous system exploits the option to encode information in the precise temporal relations between the discharges of distributed neurons and the phase relations of oscillations. It is highly likely that massive parallel recording from large numbers of neurons will reveal that synchrony is merely a special state of the complex, nonlinear dynamics of the brain. It is one of the great challenges for systems neuroscience to decipher the alphabet of temporal codes in the high dimensional time series generated by dynamically interacting neuronal assemblies.

See also: Neural Representations of Objects.

Bibliography

Biederlack, J., Castelo-Branco, M., Neuenschwander, S., Wheeler, D.W., Singer, W., Nikolic, D., 2006. Brightness induction: rate enhancement and neuronal synchronization as complementary codes. Neuron 52, 1073–1083.

Bosman, C.A., Schoffelen, J.M., Brunet, N., Oostenveld, R., Bastos, A.M., Womelsdorf, T., Rubehn, B., Stieglitz, T., De Weerd, P., Fries, P., 2012. Attentional stimulus selection through selective synchronization between monkey visual areas. Neuron 75, 875–888.

Buschman, T.J., Denovellis, E.L., Diogo, C., Bullock, D., Miller, E.K., 2012. Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. Neuron 76, 838–846.

Buschman, T.J., Miller, E.K., 2007. Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. Science 315, 1860–1862.

Canolty, R.T., Ganguly, K., Kennerley, S.W., Cadieu, C.F., Koepsell, K., Wallis, J.D., Carmena, J.M., 2010. Oscillatory phase coupling coordinates anatomically dispersed functional cell assemblies. Proceedings of the National Academy of Sciences of the United States of America 107, 17356–17361.

Castelo-Branco, M., Goebel, R., Neuenschwander, S., Singer, W., 2000. Neural synchrony correlates with surface segregation rules. Nature 405, 685–689.

Engel, A.K., Fries, P., Singer, W., 2001. Dynamic predictions: oscillations and synchrony in top–down processing. Nature Reviews Neuroscience 2, 704–716.

Fries, P., 2005. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. Trends in Cognitive Sciences 9, 474–480.

Fries, P., 2009. Neuronal gamma-band synchronization as a fundamental process in cortical computation. Annual Review of Neuroscience 32, 209–224.

Fries, P., Neuenschwander, S., Engel, A.K., Goebel, R., Singer, W., 2001a. Rapid feature selective neuronal synchronization through correlated latency shifting. Nature Neuroscience 4, 194–200.

Fries, P., Reynolds, J.H., Rorie, A.E., Desimone, R., 2001b. Modulation of oscillatory neuronal synchronization by selective visual attention. Science 291, 1560–1563.

Fries, P., Roelfsema, P.R., Engel, A.K., König, P., Singer, W., 1997. Synchronization of oscillatory responses in visual cortex correlates with perception in interocular rivalry. Proceedings of the National Academy of Sciences of the United States of America 94, 12699–12704.

Gaillard, R., Dehaene, S., Adam, C., Clémenceau, S., Hasboun, D., Baulac, M., Cohen, L., Naccache, L., 2009. Converging intracranial markers of conscious access. PLoS Biology 7, e1000061.

Gerstner, W., 1996. Rapid phase locking in systems of pulse-coupled oscillators with delays. Physical Review Letters 76, 1755–1758.

Gray, C.M., Singer, W., 1987. Stimulus-specific neuronal oscillations in the cat visual cortex: a cortical functional unit. Society of Neuroscience Abstracts 13, 404.3.

Gray, C.M., Singer, W., 1989. Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. Proceedings of the National Academy of Sciences of the United States of America 86, 1698–1702.

Gray, C.M., König, P., Engel, A.K., Singer, W., 1989. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization, which reflects global stimulus properties. Nature 338, 334–337.

Gregoriou, G.G., Gotts, S.J., Desimone, R., 2012. Cell-type-specific synchronization of neural activity in FEF with V4 during attention. Neuron 73, 581–594.

Gregoriou, G.G., Gotts, S.J., Zhou, H., Desimone, R., 2009. High-frequency, long-range coupling between prefrontal and visual cortex during attention. Science 324, 1207–1210.

Gross, C.G., 1992. Representation of visual stimuli in inferior temporal cortex. Philosophical Transactions of the Royal Society of London Series B – Biological Sciences 335, 3–10.

Hebb, D.O., 1949. The Organization of Behavior. John Wiley & Sons, New York.

Hubel, D.H., Wiesel, T.N., 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. Journal of Physiology (London) 160, 106–154.

Kreiter, A.K., Singer, W., 1996. Stimulus-dependent synchronization of neuronal responses in the visual cortex of the awake macaque monkey. The Journal of Neuroscience 16, 2381–2396.

Lima, B., Singer, W., Neuenschwander, S., 2011. Gamma responses correlate with temporal expectation in monkey primary visual cortex. The Journal of Neuroscience 31, 15919–15931.

Markov, N.T., Kennedy, H., 2013. The importance of being hierarchical. Current Opinion in Neurobiology 23, 187–194.

Melloni, L., Molina, C., Pena, M., Torres, D., Singer, W., Rodriguez, E., 2007. Synchronization of neural activity across cortical areas correlates with conscious perception. The Journal of Neuroscience 27, 2858–2865.

Miltner, W.H.R., Braun, C., Arnold, M., Witte, H., Taub, E., 1999. Coherence of gamma-band EEG activity as a basis for associative learning. Nature 397, 434–436.

Munk, M.H.J., Roelfsema, P.R., König, P., Engel, A.K., Singer, W., 1996. Role of reticular activation in the modulation of intracortical synchronization. Science 272, 271–274.

Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C., Fried, I., 2005. Invariant visual representation by single neurons in the human brain. Nature 435, 1102–1107.

Rodriguez, R., Kallenbach, U., Singer, W., Munk, M.H.J., 2004. Short- and long-term effects of cholinergic modulation on gamma oscillations and response synchronization in the visual cortex. The Journal of Neuroscience 24, 10369–10378.

Roelfsema, P.R., Engel, A.K., König, P., Singer, W., 1997. Visuomotor integration is associated with zero time-lag synchronization among cortical areas. Nature 385, 157–161.

Roelfsema, P.R., König, P., Engel, A.K., Sireteanu, R., Singer, W., 1994. Reduced synchronization in the visual cortex of cats with strabismic amblyopia. European Journal of Neuroscience 6, 1645–1655.

Singer, W., 1999. Neuronal synchrony: a versatile code for the definition of relations? Neuron 24, 49–65.

Steinmetz, P.N., Roy, A., Fitzgerald, P.J., Hsiao, S.S., Johnson, K.O., Niebur, E., 2000. Attention modulates synchronized neuronal firing in primate somatosensory cortex. Nature 404, 187–190.

Tanaka, K., 1997. Mechanisms of visual object recognition: review of monkey and human studies. Current Opinion in Neurobiology 7, 523–529.

Uhlhaas, P.J., Singer, W., 2012. Neuronal dynamics and neuropsychiatric disorders: toward a translational paradigm for dysfunctional large-scale networks. Neuron 75, 963–980.

Uhlhaas, P.J., Singer, W., 2006. Neural synchrony in brain disorders: relevance for cognitive dysfunctions and pathophysiology. Neuron 52, 155–168.

Uhlhaas, P.J., Pipa, G., Lima, B., Melloni, L., Neuenschwander, S., Nikolic, D., Singer, W., 2009. Neural synchrony in cortical networks: history, concept and current status. Frontiers in Integrative Neuroscience 3, 1–19.

Van den Heuvel, M.P., Sporns, O., 2011. Rich-club organization of the human connectome. The Journal of Neuroscience 31, 15775–15786.

Vicente, R., Gollo, L.L., Mirasso, C.R., Fischer, I., Pipa, G., 2008. Dynamical relaying can yield zero time lag neuronal synchrony despite long conduction delays. Proceedings of the National Academy of Sciences of the United States of America 105, 17157–17162.

Von der Malsburg, C., Phillips, W.A., Singer, W., 2010. Dynamic Coordination in the Brain. From Neurons to Mind. MIT Press, Cambridge (Frankfurt/M.: FIAS).

Womelsdorf, T., Schoffelen, J.-M., Oostenveld, R., Singer, W., Desimone, R., Engel, A.K., Fries, P., 2007. Modulation of neuronal interactions through neuronal synchronization. Science 316, 1609–1612.

Neuroeconomics

Stefano F Cappa, IUSS Pavia, Pavia, Italy; and San Raffaele Scientific Institute, Milan, Italy
Nicola Canessa, IUSS Pavia, Pavia, Italy; and San Raffaele Scientific Institute, Milan, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Neuroeconomics is a recent, fast developing area of cognitive neuroscience, investigating the brain mechanisms involved in economic decision-making. It is an intrinsically interdisciplinary field, taking advantage of several mature areas of research including behavioral economics, neuroscience of reward, and reinforcement learning theories. This is a selective review of experimental evidence related to some key concepts of this vast domain, starting from early neuropsychological evidence of the role of the prefrontal cortex in decision-making and adaptive learning from experience. The role of the reward system, delineated by animal studies and explored in humans by means of functional neuroimaging methods as a key mechanism responsible for behavioral learning, has been extensively investigated in decision-making situations involving uncertainty, temporal delay, and the exploration of potential rewards. The theoretical framework of prospect theory has been used to investigate the neural mechanisms involved in anomalies of choice, in particular in loss aversion. Another important area of neuroeconomic research deals with the neural mechanisms involved in social contexts, i.e., those underpinning theory of mind and reciprocal trust. These multiple approaches are contributing to the rigorous study of individual differences in decision-making and of their neurobiological foundations.

Introduction

Economic decision-making, i.e., the identification of values and uncertainties involved in the production, distribution, and consumption of goods and services, is a central component of human life. The investigation of the brain mechanisms responsible for the set of cognitive processes involved in the selection of one course of action among the manifold available options informs the area of research called neuroeconomics. This field, as any other area of cognitive neuroscience, is intrinsically interdisciplinary and takes advantage of different methodologies and research traditions (among others, experimental and behavioral economics, cognitive and social psychology, computational modeling) (Levallois et al., 2012). These multiple approaches converge on the notion of *adaptive behavioral learning*, the core issue of economic behavior (Ullsperger et al., 2014).

Neuropsychological Insights

The history of neuroeconomics is relatively short, although its beginnings can be traced back to the classical neuropsychological case study of Phineas Gage, a survivor of extensive damage to the medial prefrontal cortex with extensive changes in personality, which resulted, among other consequences, in faulty decisions in everyday life (Macmillan, 2002). The implications of this case were developed in particular by Antonio Damasio and Antoine Bechara (Damasio, 1994), who studied individuals in whom similar lesions, involving in particular the ventromedial prefrontal cortex (vmPFC), were associated with impaired decision-making in the absence of other relevant cognitive disorders (Bechara et al., 1997). In what would become a very famous experiment, the subjects were asked to play a gambling task based on four decks of cards. Unbeknownst to the player, two 'disadvantageous' decks were associated with high winnings in the first trials but even higher losses in the long run, while the two other ('advantageous') decks resulted in higher winnings overall. While the control subjects began to prefer the latter decks even before becoming consciously aware of such difference (i.e., an implicit behavioral learning), the patient group continued to select the former, even after they explicitly understood the effective strategy. This observation was one of the foundations of the 'somatic marker hypothesis,' i.e., the proposal that choices are driven not only by rational and explicit computations, but also by unconscious biases, resulting from previous emotionally laden experiences. As discussed later, this observation connects this first neuropsychological foundation of neuroeconomics with the neurophysiological approach, concerned with the investigation of the neural mechanisms of behavioral learning in the mesocorticolimbic dopaminergic pathway (see Schultz, 2007, 2013). It is noteworthy, indeed, that such a differential learning in the two groups reflected in opposite affective reactions to risky choices and their anticipated outcomes. While the normal subjects developed a psychogalvanic response before drawing a card from 'disadvantageous' decks, the patients failed to display any change in skin conductance even after being consciously aware of their risky nature. These clinical observations opened the avenue for the investigation of the interacting roles of cognitive computations and emotional responses during decision-making in healthy individuals, particularly with neuroimaging techniques. Patient studies, however, continue to provide important insights into the necessary role of specific brain structures in decision-making, even beyond the strictly 'economic' domain. This is the case of the neural bases of moral, rather than purely economic, decision-making, which have been shown to involve overlapping regions (Shenhav and Greene, 2010). In other words, the computation of the expected 'moral value' of a decision (number of lives lost/saved) engages the same regions activated by monetary rewards, and vmPFC damage results in 'pure

utilitarian' choices in personal moral dilemmas characterized by a high emotional charge (i.e., inflicting serious harm to one person to save five) (Koenigs et al., 2007). That is, in choices characterized by a privileged consideration of the aggregate benefit, ignoring the aversive reactions displayed by control subjects when evaluating the negative consequences of one's own choice for another's welfare.

Expected Utility, Evaluation, and the Reward System

The classical approach of economic theory is normative, i.e., it is based on what can be considered as the behavior of a fully rational decisional agent. According to the expected utility theory (EUT; von Neumannn and Morgenstern, 1953), the ideal economic decision-maker, when confronted with a choice situation, computes the expected utility of each option by integrating the (positive or negative) subjective value of its possible outcomes, weighted by their probabilities. This computation is followed by a comparison among the valuation results, and by the selection of the option associated with the highest utility. Making decisions thus entails learning, estimating, predicting, and updating the *values* associated with available options. This notion underpins both the interdisciplinary nature of neuroeconomics, and its potential explanatory power well beyond the strictly economic field. Indeed, the strict relationship, both in behavioral and neurobiological terms, between the notion of value and that of *reward* allows the grounding of neuroeconomics in one of the most advanced areas of neuroscience, i.e., the neurobiology of learning and motivation (e.g., Doya, 2008). This is a crucial point that has been missed by some of the critics of the application of neuroscientific methods to the economic and social sciences (e.g., Legrenzi and Umiltà, 2011). Far from being speculative, and based on loose and neophrenological interpretations of the functional role of a given brain region, neuroeconomics builds on extremely solid anatomical and physiological foundations.

Indeed, the functional anatomy of the reward system is well known, largely on the basis of extensive evidence from animal studies (for a comprehensive review see Haber and Knutson, 2010). Pharmacological, physiological, and behavioral studies have clearly indicated that the mesocorticolimbic dopaminergic pathway (see **Figure 1**), originating in the midbrain ventral tegmental area (VTA) and projecting to the nucleus accumbens (NAcc), amygdala and medial prefrontal cortex, is the crucial component of the valuation system (Rolls, 2000; Schultz, 2007, 2013). Early imaging studies in humans, based on pharmacological stimulations by cocaine, highlighted an extensive network of areas, extending, beyond the NAcc, to the amygdala and prefrontal cortex, activated by the *experience* of motivationally relevant stimuli (Breiter et al., 1997). A comparable network was then shown to be engaged by a crucial reward for humans such as monetary gains (Thut et al., 1997). In particular the vmPFC, implicated by neuropsychological studies, has been shown to represent the subjective value of different goods (food, nonfood, monetary gambles), when a decision to purchase is made (Chib et al., 2009; Plassmann et al., 2007), or when a trade-off between crucial choice parameters such as quality versus quantity is computed (Padoa-Schioppa and Assad, 2006). The interplay of this region with other frontal

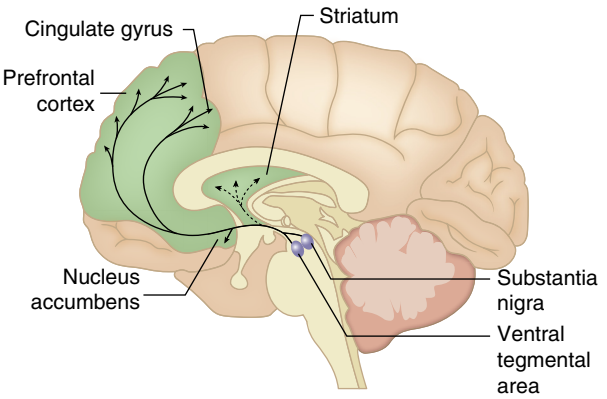

Figure 1 Brain reward circuits. The major dopaminergic projections to the forebrain that underlie brain reward are shown superimposed on a diagram of the human brain: projections from the ventral tegmental area to the nucleus accumbens (ventral striatum), and prefrontal cerebral cortex. Also shown are projections from the substantia nigra to the dorsal striatum (caudate and putamen and related structures) that play a role in habit formation and in well-rehearsed motor behavior, such as drug seeking and drug administration. Copyright Elsevier: http://www.elsevierimages.com/image/25905.htm; reprinted from Hyman, S.E., Malenka, R.C., Nestler, E.J. 2006. Neural mechanisms of addiction: The role of reward-related learning and memory. Annual Review Neuroscience 29, 565–598.

areas, and particularly the dorsomedial–dorsolateral frontal networks involved in cognitive control (Koechlin et al., 2003), was demonstrated by the observation that self-control while making choices is mediated by modulation of this region by the dorsolateral prefrontal cortex (dlPFC) (Hare et al., 2009).

Via the experience (or 'nonexperience') of rewards and punishments, valuation plays a central role in their anticipation, and accordingly in *behavioral learning*. An epochal step in the understanding of the functional properties of the reward system has been the development of the notion of 'reward prediction error' (RPE), based on the crucial concept of temporal difference learning (see Sutton and Barto, 1998). This notion derives from the identification of the response properties of midbrain dopaminergic neurons by Wolfram Schultz. The firing properties of these neurons are compatible with a process of comparison between expectation and outcome, a process that is at the basis of reinforcement learning (Schultz et al., 1997). A positive RPE (an outcome larger than predicted) increases firing rate and elicits a phasic response. A negative RPE (an outcome smaller than predicted) decreases firing rate and elicits a tonic response. A perfect match of expected and actual rewards (null RPE) does not engage the midbrain dopaminergic system. Learning to predict rewards is a crucial requirement of effective decision-making, and there is ample evidence that the dopaminergic encoding of the difference between expected and actual rewards computed in the NAcc is a critical mechanism of human reward-based learning (Schultz, 2007, 2013). Another crucial step in the interdisciplinary integration among the subfields of neuroeconomics was the discovery that reinforcement learning, mediated by RPE, accounts for both Pavlovian (stimulus–stimulus) and instrumental (stimulus–action) learning (Barto, 1995). In the former case, the signals in the ventral striatum (VS) and in the orbitofrontal cortex,

assessed on the basis of the neural response during Pavlovian conditioning in humans, paralleled the results of single-cell recording in monkeys, reflecting the properties of the RPE signal (O'Doherty et al., 2003). Instrumental conditioning is based on the acquisition of the contingencies between the organism's actions and the resulting positive and negative outcomes. One component process, the 'critic,' computes the temporal difference error and updates predictions about future rewards via the same neural system activated by Pavlovian conditioning, centered on the VS. A second component, the 'actor,' updates action plans in order to maximize reward via approach versus avoidance behaviors, in the dorsal striatum. In human subjects, indeed, the left anterior caudate nucleus, a component of the dorsal striatum, shows stronger prediction error-related responses in instrumental, compared with Pavlovian, conditioning (O'Doherty et al., 2004). The phasic prediction error signal provided by the midbrain dopaminergic neurons is also involved in situations in which the predictions are based on statistical inferences or logical reasoning (model-based learning), rather than on immediate reward contingencies (model-free learning) (see Daw et al., 2011). While the role of the dopaminergic system in the acquisition of models is unknown, the response properties clearly highlight the modulation of response by model-based predictions (Schultz, 2013). There is evidence that, in humans, providing explicit knowledge about cue-reward probabilities, which results in improved performance, has an impact on dlPFC activation, correlated to a decrease in VS–vmPFC activity (Li et al., 2011). The computation of what has been called a 'state prediction error' in model-based learning (the state that is expected on the basis on the previous states and choices) has been shown to involve the lateral prefrontal cortex and the intraparietal sulcus (IPS) (Glascher et al., 2010).

Besides the amount of reward, another crucial determinant of expected utility is its *probability*. Probability results in *uncertainty*, which is maximal at $p = 0.5$, and zero at $p = 0$ and $p = 1$. Animal studies show that the midbrain dopamine system separately codes probability (by a monotonically increasing response, when the conditioned stimulus predicting reward is presented) and uncertainty (by a gradual increase in activity, peaking for $p = 0.5$, between the conditioned stimulus and potential time of reward) (Fiorillo et al., 2003). Neuroimaging studies in humans have highlighted, at the whole-brain level, the neural mechanisms of outcome anticipation in uncertain situations, showing that uncertainty is correlated with activity in the lateral orbitofrontal cortex (Tobler et al., 2007). A further distinction, which reflects in different patterns of brain activity, concerns *risk* and *ambiguity*, i.e., when probabilities are, respectively, known or unknown. Ambiguous, compared with risky, situations recruit inferior frontal and posterior parietal brain regions, both when ambiguous outcomes are passively perceived (Bach et al., 2009), and when a decision has to be made (Huettel et al., 2009). The latter study reported an interesting dissociation between these regions, highlighting the role of *individual differences* in risk preference on decision-making. Preference for ambiguity was predicted by activity on the lateral prefrontal cortex, and was inversely related to impulsivity. In contrast, risk preference was predicted by activity in the posterior parietal cortex. Neural signatures of risk-seeking versus risk aversion have been found during the anticipation phase of high-risk gambles (Rudorf et al., 2012). Risk adverse subjects, compared with risk-seeking ones, displayed a stronger response in the VS and anterior insula.

The timing of reward, and particularly the willingness to defer immediate gratification, is another important source of individual differences in decision-making, predicting academic achievement and other desirable outcomes (Duckworth and Seligman, 2005). A useful metrics of this behavioral trait is *delay discounting*, i.e., the decline of subjective value as a function of reward delay. Activity in the VS and vmPFC tracks delay discounting, as indicated by the decline in BOLD response paralleling the decrease in subjective value (Pine et al., 2010). Individual differences in the brain correlates of thinking about the future have been shown to predict delay discounting, suggesting that basic differences in temporal perception may be responsible for this crucial dimension of individual behavior (Cooper et al., 2013).

Behavioral learning, i.e., updating the value of stimuli and actions on them, requires also exploring the environment to gather information and thus decrease uncertainty. Exploring is intrinsically riskier than exploiting available knowledge, but the resulting novel information has a value of its own. The neural mechanisms subserving what has been called the *exploration–exploitation dilemma* have been studied using tasks in which the outcomes of the available options change over time, such as the classical 'four-armed bandit task.' This is a widely used decision-making task in which, on every trial, participants have to choose one among four slot-machines, whose payoffs differ and change slowly but continuously over time. By selecting a bandit and attending its current payoff, subjects can thus estimate and update its expected value at a given timepoint. The task highlights the value of exploring the decisional space because, although less risky in principle, choosing repeatedly the same bandit (i.e., exploiting) has a cost, i.e., the impossibility to learn the value of other options, as well as the possibility that previously collected knowledge is not up-to-date anymore. Optimal performance thus requires the ability to balance exploitative choices (i.e., choosing the bandit associated with the highest estimated expected value in a given timepoint) and explorative choices (i.e., choosing another bandit). At the neural level, the choice between exploiting current knowledge, and exploring in order to collect new one, involves a trade-off in the activity of two brain networks. While abundant evidence supports the role of the dopaminergic frontostriatal system in learning to exploit (Beeler et al., 2010), explorative decisions engage the IPS and frontopolar cortex (FPC) (Daw et al., 2006; Laureiro-Martinez et al., 2014a, 2014b; see **Figure 2** – top). Recent studies show that also the ability to timely switch between exploitation and exploration is subject to individual differences, involving the FPC. Activity in this region, particularly in the right hemisphere, tracks the value of alternative choices and predicts effective switching between exploitation and exploration (Boorman et al., 2009). Supporting evidence comes also from a recent study comparing brain activity of matched groups of entrepreneurs (i.e., people managing the organization they founded based on their venture idea) and managers (involved in making strategic decisions but with no

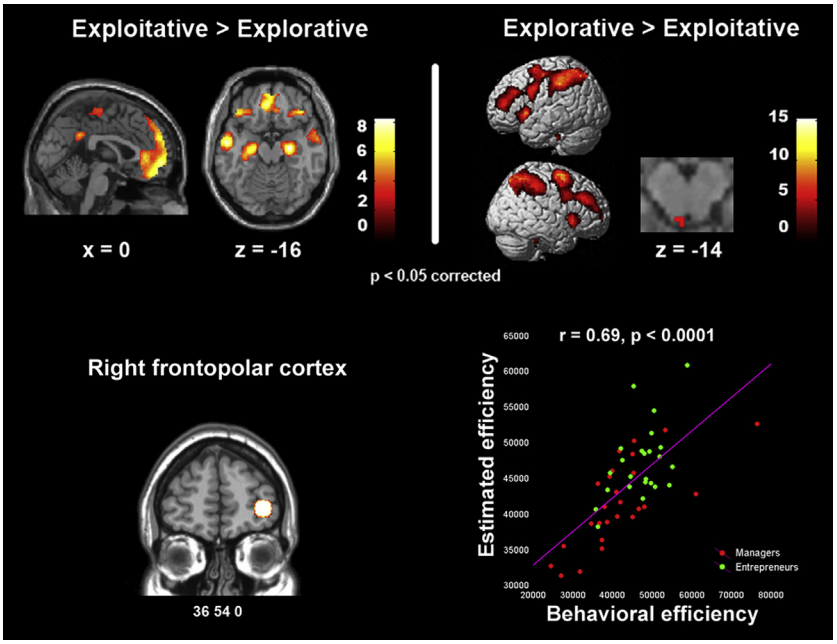

Figure 2 Top: The brain regions that were more strongly activated by exploitative than explorative choices (left), and by explorative than exploitative choices (right). Bottom: The relationship between behavioral decision-making efficiency (payoff divided by response time) at the 'four-armed bandit task' and estimated efficiency, based on the activity of regions which, in previous studies, showed a significant difference between exploitative and explorative choice (ventromedial prefrontal cortex, bilateral frontopolar cortex and intraparietal sulcus dorsal, sector of anterior cingulate cortex).

venture experience) engaged in the four-armed bandit task (Laureiro-Martinez et al., 2014a). Compared with managers, entrepreneurs showed higher decision-making efficiency (i.e., comparable payoff in shorter time), related to stronger activity in the same sector of the right FPC described by Boorman et al. (2009) (**Figure 2**-bottom), a finding that may be compatible with specific experience or with biological predispositions (see also Coates et al., 2009).

Prospect Theory and Anomalies of Choice

The advancements of neuroeconomics in mapping the critical nodes of the decisional brain, as well as the functional description of their role in terms of approach versus avoidance behaviors modulated by reinforcement learning, now provide a neurological interpretation to many of the human deviations from the normative rationality assumed by classical economic theories. Indeed, an important line of research at the crossroad between experimental economics, psychology, and decision science is based on a *descriptive*, rather than normative approach to decision-making, i.e., on the notion that many critical assumptions of EUT are systematically violated when real people make real choices (Tversky and Kahneman, 1974). Many resulting anomalies of choice have been investigated, including loss aversion (the higher sensitivity to losses than equivalent gains (Kahneman and Tversky, 1984)) and the framing effect (the different attitude toward risk when outcomes are framed as gains vs losses (Tversky and Kahneman, 1981)). A cognitive interpretation of both these anomalies is given by the doctrinal corpus of prospect theory (PT) (Kahneman and Tversky, 1979), which departs from EUT in several respects. First, in PT outcomes are separately coded, via specific value functions, as gains or losses with reference to a neutral reference point which is not the absolute zero. Second, the value function defined over losses is steeper than that defined over gains, which explains loss aversion. Third, both value functions display a logarithmic trend (concave for gains and convex for losses), which accounts for risk aversion in losses and risk-seeking in gains (i.e., the framing effect). While PT originally provided a cognitive interpretation of these anomalies of choice, i.e., in terms of cognitive functions translating normative values and probabilities in, respectively, subjective values and probability weights the increasing integration between such behavioral description and neuroscientific evidence extends this view, by highlighting individual differences in the balance among cognitive, emotional, and motivational factors involved in the processes of anticipation and valuation. Several studies, indeed, investigated the neural bases of human decision-making by focusing on the regions involved in the anticipation of prospective outcomes. The first study assessing brain responses to the expectation and experience of monetary gains and losses was reported by Breiter et al. (2001). This study firmly established the role of the mesolimbic 'reward' pathway, involving the NAcc, amygdala, and hippocampus, in tracking the monetary value of prospects. Subsequent studies related individual differences in the disposition toward specific anomalies of choices, such as framing effect and loss aversion, to the strength of activation in some of these brain structures.

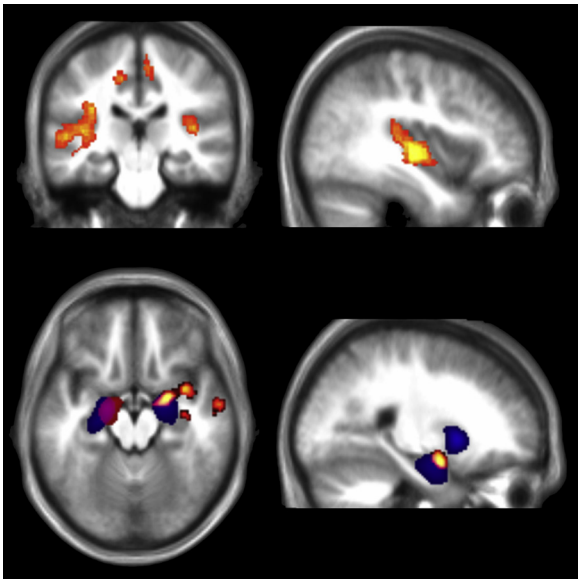

Figure 3 The functional and structural neural basis of individual differences in loss aversion (see Canessa et al., 2013). Yellow to orange colors highlight the brain regions where brain activity is positively correlated with potential losses in decision-making under risk (bilateral posterior insula and middle cingulate cortex in the top part of the figure, right amygdala in its bottom part). Blue colors indicate a structural network including bilateral amygdala, thalamus, and striatum, in which source-based morphometry highlighted a positive correlation between loss aversion and gray matter volume. Red colors indicate the left amygdala cluster in which voxel-based morphometry highlighted a positive correlation between loss aversion and gray matter volume.

The neural bases of the framing effect were investigated by De Martino et al. (2006). Following seminal studies by Kahneman and Tversky, in every trial their subjects were asked to choose between a safe option ($p = 1$) and a gamble. In different trials, prospective outcomes were presented either as sure gains (e.g., keep 20 USD) or losses (lose 30 USD), or probabilistic gains (keep all with $p = 0.4$) or losses (lose all with $p = 0.6$), with respect to an initial endowment (e.g., 50 USD). The 'rational' expected values were thus identical across the four types of trials (i.e., safe-gain, safe-loss, gamble-gain, gamble-loss). According to EUT, rational decisions by a utility maximizer should thus be independent from the way prospects are presented. This, however, is not the case in the real world, and the framing of outcomes affected significantly the choice, with the 'loss' frame resulting in higher risk-taking than the 'keep' one. At the neural level, the activation of the amygdala was specifically associated with the framing effect, both when the sure option choice was chosen in the gain frame, and when the gamble was chosen in the loss frame. In contrast, when subjects were choosing against such general tendency increased activity was observed in the anterior cingulate and in the dlPFC. As it could be expected from previous behavioral evidence, the susceptibility of individual subjects to the framing effect was quite variable. This was reflected, in the imaging data, by a correlation between a 'rationality' index and activity in vmPFC, likely inhibiting aversive signals generated by the amygdala (Gottfried and Dolan, 2004).

The latter structure is also critical in generating individual differences in loss aversion when anticipating prospective outcomes, as shown by both lesion evidence in brain-damaged patients (De Martino et al., 2010) and neuroimaging data from healthy individuals (Canessa et al., 2013). Different neural mechanisms, however, appear to underlie the disproportionate weighting to losses than gains. One such mechanism specifically involves the VS, which displays a bidirectional response of activation for gains and deactivation for losses, with the steeper degree of deactivation versus activation reflecting individual differences in loss aversion (Tom et al., 2007). A recent study by Canessa et al. (2013) shows that loss aversion additionally reflects the activity of another mechanism, specific to the anticipation of losses and centered on the amygdala. In the latter study, subjects were asked to accept or reject mixed gambles offering equal (50%) chances of gaining or losing different amounts of money. The results highlighted both bidirectional and gain/loss specific responses while evaluating risky gambles, with amygdala and posterior insula specifically tracking the magnitude of potential losses. At the individual level, loss aversion was reflected both in limbic *f*MRI responses and in gray matter volume in a structural amygdala–thalamus–striatum network, in which the volume of the 'output' centromedial amygdala nuclei mediating avoidance behavior was negatively correlated with monetary performance (see **Figure 3**).

Outcome anticipation and ensuing loss aversion thus appear to involve multiple neural systems, showing functional and structural individual variability directly related to the actual financial outcomes of choices.

Social Decision-Making: Economic Exchanges, Cooperation, and Social Learning

Individual profiles of loss aversion, susceptibility to the framing effect, and intertemporal preference can be thought to summarize the essential features of a decision-maker, and to be able to predict her/his choice behavior. This model has been defined as a 'two dimensional' perspective (Rustichini, 2009). Economic decision-making, however, typically takes place in the context of social interaction, rather than in the isolated condition of the lottery gamble. The social nature of the human species thus reflects in the 'strategic' nature of many human choices, i.e., in decisional processes which need to take into account others', besides one's own, utility when evaluating prospective outcomes. While emotions and motivations play a key role in individual decision-making, the social context has been shown to enhance the weight of affective drives (Bault et al., 2008, 2011), as well as to recruit processes of social cognition associated with the understanding of others' intentions and emotions. This is the case, for instance, of 'trust' and 'cooperation' games which, not surprisingly, involve the ability to infer each other's mental states, i.e., Theory of mind or mentalizing (Premack and Woodruff, 1978). The study by McCabe et al. (2001) provided the first insights into the neural mechanisms involved in interpersonal exchange in a trust game. In this

study, subject A has to decide whether to select move x, which ends the game, or move b, which gives to subject B the possibility to move. In the former case, both subjects receive the same sum (e.g., 45). In the latter, the final outcome depends on B, who again has two choices: one gives to both increased winnings (e.g., 180 to A and 225 to B), while with the second she/he wins a large sum (e.g., 405) and the first subject gets nothing. Of course the choice of A depends crucially on her/his trust that B will reciprocate, while the choice of B depends on her/his perception of commitment toward reciprocation. The different payoffs were variable, in order to increase or decrease the incentive for cooperation. Crucially, subject B could be a human counterpart, or a computer deciding on the basis of fixed probabilities that were communicated to subject A. The main result is that most subjects decided to cooperate, and their prefrontal cortex was more active when playing with human subjects than with a computer. This difference was not seen in the minority of subjects who did not reciprocate.

Besides understanding another's intentions, sharing her/his emotional states has also an impact on decision-making. For example, different studies highlighted behavioral adaptations (i.e., increased vs decreased risk-taking) induced by both experienced and attended regret. These studies were based on a gambling task composed of several consecutive trials, requiring participants to choose between two gambles (Mellers et al., 1999). Gambles are depicted as 'wheels-of-fortune' in which different probabilities of variable amounts of gain or loss are represented by the relative size of colored sectors of the wheel. Immediately after the choice, the gambles are played and the results are shown. In different trials, participants are shown only the results of the chosen gamble (and thus experience satisfaction for a gain or disappointment for a loss) or the results of both the chosen and unchosen gambles (and thus experience relief or regret for a foregone outcome that might have been worse or better, respectively). Both lesional (Camille et al., 2004) and neuroimaging (Coricelli et al., 2005, 2007) evidence shows that the experience of regret, and the resulting behavioral adaptations, involve the vmPFC alongside the anterior cingulate cortex and the hippocampus. Canessa et al. (2009) used the same gambling task to test whether a 'resonant' neural mechanism is activated both when personally experiencing and when attending in someone else a complex emotion such as regret or relief. The participants either chose one of the two gambles, resulting in real gains or losses, or observed the same sequence of events (gamble evaluation, decision, and outcome evaluation) experienced by another individual playing the same task in a nearby room. The baseline was a condition in which the computer randomly chose one of the gambles, either for the participant or for the other player. Observing the regretful outcomes of someone else's choices activates the same regions that are activated during a first-person experience of regret, i.e., vmPFC, anterior cingulate cortex, and hippocampus (Coricelli et al., 2005). This finding suggests that the understanding of others' regret is mediated by the reactivation of the same brain regions that induce the feeling of regret in the beholder during a first-person experience. In a subsequent study, the same authors investigated whether this mechanism is also involved in learning from others' previous outcomes (Canessa et al., 2011). At the behavioral level the subjects changed their risk-aptitude in line with the outcomes of regret/relief of their previous decision, showing increased risk-seeking after 'relief for a risky choice' and 'regret for a nonrisky choice,' and reduced risk-seeking after 'relief for a nonrisky choice' and 'regret for a risky choice.' The same behavioral adaptation to previous experience was observed also after the *other player's* previous outcomes, while no significant change was observed after an outcome resulting from a random-choice by the computer. The latter finding indicates that the behavioral influence induced by one's own and others' previous outcome result from an emotional response associated with a *sense of responsibility*. This behavioral learning was reflected in cerebral regions specifically coding the effect of previously experienced regret/relief when making a new choice. Activity in subgenual cortex and caudate nucleus tracked the outcomes that increased risk-seeking (relief for a risky choice, and regret for a nonrisky choice). These regions were also more strongly activated by final risky, compared with nonrisky, decisions, and their conjoint activity is likely to reflect the motivational drive arising from previous outcomes that highlighted the reward-value of risky options. In contrast, activity in the vmPFC, amygdala, and periaqueductal gray matter reflected the outcomes reducing risk-seeking (relief for a nonrisky choice, and regret for a risky choice). All these regions, along with the anterior insula, were also more strongly activated while making nonrisky versus risky choices. The vmPFC likely reflects adaptive learning from past emotional experiences reducing risk-seeking and, via connections with amygdala, insula, and periaqueductal gray matter (Reynolds and Zahm, 2005) activates the negative feeling associated with regret and its anticipation. A subset of these regions reflected both first and third person previous outcomes when making new choices. This finding, that fits with the influence from others' outcomes highlighted by behavioral data, extends the concept of emotional resonance to the decisional domain, where such a shared response might act as one of the neural mechanisms underlying social learning. This mechanism would entail the mapping of the emotional consequences of others' choices on the same emotional states that are experienced as a first person, through the reactivation of the same cerebral regions that are involved in their direct experience.

The role of complex emotions in economic decision-making is also supported by studies based on the Ultimatum Game (UG) (Sanfey et al., 2003). This task involves two players that have to split a sum of money. The proposer decides how to split the sum, while the respondent can either accept the split or reject it. In the latter case, neither player receives anything. The theoretical implications of the UG are straightforward. The EUT prediction is that the respondent should accept any split, as any monetary outcome is better than nothing, and proposers should thus offer the smallest possible sum. Humans, however, do not behave this way, in that most proponents reject unfair divisions (see Camerer, 2003). The discussion about the possible reasons for this 'breach of rationality' by normal subjects is extensive, and open to different interpretations (Civai et al., 2013). The corresponding neural signature, however, is consistent, and involves emotional brain areas, in particular the anterior

insula, associated to negative emotional reactions such as those involved in pain and disgust (Wicker et al., 2004). Interesting converging evidence for a crucial role of this region in economic social exchange is provided by a study of borderline personality disorder (BPD), a clinical condition characterized by disruptive behavioral features which can be at least in part ascribed to defective cooperation. During a cooperative game based on multiple rounds of investment by healthy 'trustor' subjects, healthy trustees continued to cooperate, while BPD trustees failed to repay after the early rounds (King-Casas et al., 2008). This behavior was associated with a different activation profile in the anterior insula region. In healthy subjects, insular activity increased both with the sum received and the sum paid back, while only the latter response was observed in BPD.

Conclusions

From the economist's vantage point, the crucial question is whether the large and increasing neuroscientific evidence, of which we have given some selected examples, can be expected to have any impact on economic theories (Farb, 2013). As neuroscientists we cannot enter into this lively debate, but we can summarize what we consider as the most important implications of neuroeconomics for cognitive neuroscience. In the first place, the application of the rigorous and computationally explicit models of experimental economics to neuroimaging studies of human subjects provides a powerful tool for the investigation of individual differences in decision-making. The possibility to quantify variability in parameters such as risk aversion, loss aversion, delay discounting, explorative behavior, or cooperation is an important contribution of neuroeconomics research to the understanding of the neural bases of human behavior. In turn, evidence at the neurological level raises further questions whose implications exceed the domain of cognitive neuroscience. For instance, are these differences in behavior and neural activation the consequence of natural predispositions, or rather of brain plasticity phenomena consequent to training and experience? The former view is supported by studies of genetic polymorphisms, which, once again, typically highlight the role of genes affecting neurotransmission in regions associated with reward processing (e.g., dopamine modulation (Krugel et al., 2009)), amygdala reactivity (e.g., serotonin transporter 5-HTTLPR; (Crisan et al., 2009; Roiser et al., 2009)) and executive functioning (e.g., COMT gene; Frank et al., 2009; Golimbet et al., 2007). Studies of training effects are however needed to assess the possible role of neuroplasticity in the causation of individual and group-level differences, a topic of great interest also in terms of possible practical applications. A second, exciting area of application is translational neuroeconomics, i.e., the use of models and results of neuroeconomic research to increase our understanding of human disease. Many psychiatric and neurological disorders involve reward processing and decision-making, from addiction to personality disorders (Sharp et al., 2012), from Parkinson's disease (Schott et al., 2007) to frontotemporal dementia (Chiong et al., 2013). The possible impact of neuropharmacological manipulation has been already demonstrated for the dopamine system (Chowdhury et al., 2013), as well as for sexual hormones (Bos et al., 2010) and oxytocin (Declerck et al., 2010), while alternative ways to modulate the 'decisional brain' are already represented by neurostimulation techniques such as transcranial magnetic stimulation and transcranial direct current stimulation (Miniussi and Ruzzoli, 2013). As in most lively scientific fields, each empirical answer raises novel research questions, and this is one of the reasons why the interdisciplinary effort underlying neuroeconomics is one of the most exciting scientific challenges in contemporary cognitive neurosciences.

See also: Brain Stimulation Reward; Emotion, Neural Basis of; Neuromarketing; Reinforcement: Neurochemical Substrates.

Bibliography

Bach, D.R., Seymour, B., Dolan, R.J., 2009. Neural activity associated with the passive prediction of ambiguity and risk for aversive events. J Neurosci 29 (6), 1648–1656. http://dx.doi.org/10.1523/JNEUROSCI.4578-08.2009.

Barto, A.G., 1995. Adaptive critic and the basal ganglia. In: Houk, J.C., Davis, J.L., Beiser, D.G. (Eds.), Models of Information Processing in the Basal Ganglia. MIT Press, Cambridge, MA, pp. 215–232.

Bault, N., Coricelli, G., Rustichini, A., 2008. how social ranking affects choice behavior. PLoS One 3 (10), e3477. http://dx.doi.org/10.1371/journal.pone.0003477. Interdependent utilities.

Bechara, A., Damasio, H., Tranel, D., Damasio, A.R., 1997. Deciding advantageously before knowing the advantageous strategy. Science 275, 1293–1295.

Beeler, J.A., Daw, N., Frazier, C.R., Zhuang, X., 2010. Tonic dopamine modulates exploitation of reward learning. Frontiers in Behavioral Neuroscience 4, 170.

Boorman, E.D., Behrens, T.E., Woolrich, M.W., Rushworth, M.F., 2009. How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. Neuron 62, 733–743.

Bos, P.A., Terburg, D., Van Honk, J., 2010. Testosterone decreases trust in socially naive humans. Proceedings of the National Academy of Sciences of the United States of America 107, 9991–9995.

Breiter, H.C., Aharon, I., Kahneman, D., Dale, A., Shizgal, P., 2001. Functional imaging of neural responses to expectancy and experience of monetary gains and losses. Neuron 30, 619–639.

Breiter, H.C., Gollub, R.L., Weisskoff, R.M., Kennedy, D.N., Makris, N., Berke, J.D., Goodman, J.M., Kantor, H.L., Gastfriend, D.R., Riorden, J.P., Mathew, R.T., Rosen, B.R., Hyman, S.E., 1997. Acute effects of cocaine on human brain activity and emotion. Neuron 19, 591–611.

Camerer, C., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press, Princeton.

Camille, N., Coricelli, G., Sallet, J., Pradat-Diehl, P., Duhamel, J.R., Sirigu, A., 2004. The involvement of the orbitofrontal cortex in the experience of regret. Science 304 (5674), 1167–1170. http://dx.doi.org/10.1126/science.1094550.

Canessa, N., Crespi, C., Motterlini, M., Baud-Bovy, G., Chierchia, G., Pantaleo, G., Tettamanti, M., Cappa, S.F., 2013. The functional and structural neural basis of individual differences in loss aversion. Journal of Neuroscience 33, 14307–14317.

Canessa, N., Motterlini, M., Alemanno, F., Perani, D., Cappa, S.F., 2011. Learning from other people's experience: a neuroimaging study of decisional interactive-learning. Neuroimage 55, 353–362.

Canessa, N., Motterlini, M., Di Dio, C., Perani, D., Scifo, P., Cappa, S.F., Rizzolatti, G., 2009. Understanding others' regret: a FMRI study. PLoS One 4, e7402.

Chib, V.S., Rangel, A., Shimojo, S., O'doherty, J.P., 2009. Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. Journal of Neuroscience 29, 12315–12320.

Chiong, W., Hsu, M., Wudka, D., Miller, B.L., Rosen, H.J., 2013. Financial errors in dementia: testing a neuroeconomic conceptual framework. Neurocase.

Chowdhury, R., Guitart-Masip, M., Lambert, C., Dayan, P., Huys, Q., Duzel, E., Dolan, R.J., 2013. Dopamine restores reward prediction errors in old age. Nature Neuroscience 16, 648–653.

Civai, C., Rumiati, R.I., Rustichini, A., 2013. More equal than others: equity norms as an integration of cognitive heuristics and contextual cues in bargaining games. Acta Psychologica (Amst) 144, 12–18.

Coates, J.M., Gurnell, M., Rustichini, A., 2009. Second-to-fourth digit ratio predicts success among high-frequency financial traders. Proceedings of the National Academy of Sciences of the United States of America 106, 623–628.

Cooper, N., Kable, J.W., Kim, B.K., Zauberman, G., 2013. Brain activity in valuation regions while thinking about the future predicts individual discount rates. Journal of Neuroscience 33, 13150–13156.

Coricelli, G., Critchley, H.D., Joffily, M., O'Doherty, J.P., Sirigu, A., Dolan, R.J., 2005. Regret and its avoidance: a neuroimaging study of choice behavior. Nature Neuroscience 8 (9), 1255–1262. http://dx.doi.org/10.1038/nn1514.

Coricelli, G., Dolan, R.J., Sirigu, A., 2007. Brain, emotion and decision making: the paradigmatic example of regret. Trends in Cognitive Science 11 (6), 258–265. http://dx.doi.org/10.1016/j.tics.2007.04.003.

Crisan, L.G., Pana, S., Vulturar, R., Heilman, R.M., Szekely, R., Druga, B., Dragos, N., Miu, A.C., 2009. Genetic contributions of the serotonin transporter to social learning of fear and economic decision making. Social Cognitive Affective Neuroscience 4, 399–408.

Damasio, A.R., 1994. Descartes' Error. Emotion, Reason and the Human Brain. Avon Books, New York.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J., 2011. Model-based influences on humans' choices and striatal prediction errors. Neuron 69, 1204–1215.

Daw, N.D., O'doherty, J.P., Dayan, P., Seymour, B., Dolan, R.J., 2006. Cortical substrates for exploratory decisions in humans. Nature 441, 876–879.

De Martino, B., Kumaran, D., Seymour, B., Dolan, R.J., 2006. Frames, biases, and rational decision-making in the human brain. Science 313, 684–687.

De Martino, B., Camerer, C.F., Adolphs, R., 2010. Amygdala damage eliminates monetary loss aversion. Proceedings of the National Academic Sciences of the USA 107 (8), 3788–3792. http://dx.doi.org/10.1073/pnas.0910230107.

Declerck, C.H., Boone, C., Kiyonari, T., 2010. Oxytocin and cooperation under conditions of uncertainty: the modulating role of incentives and social information. Hormones and Behavior 57, 368–374.

Doya, K., 2008. Modulators of decision making. Nature Neuroscience 11, 410–416.

Duckworth, A.L., Seligman, M.E.P., 2005. Self-discipline outdoes IQ in predicting academic performance of adolescents. Psychological Science 16, 939–944.

Farb, N.A.S., 2013. Can neuroimaging inform economic theories of decision-making? Neuroscience and Neuroeconomics 2, 1–10.

Fiorillo, C.D., Tobler, P.N., Schultz, W., 2003. Discrete coding of reward probability and uncertainty by dopamine neurons. Science 299, 1898–1902.

Frank, M.J., Doll, B.B., Oas-Terpstra, J., Moreno, F., 2009. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nature Neuroscience 12 (8), 1062–1068. http://dx.doi.org/10.1038/nn.2342.

Glascher, J., Daw, N., Dayan, P., O'doherty, J.P., 2010. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66, 585–595.

Golimbet, V.E., Alfimova, M.V., Gritsenko, I.K., Ebstein, R.P., 2007. Relationship between dopamine system genes and extraversion and novelty seeking. Neurosci Behav Physiol 37 (6), 601–606. http://dx.doi.org/10.1007/s11055-007-0058-8.

Gottfried, J.A., Dolan, R.J., 2004. Human orbitofrontal cortex mediates extinction learning while accessing conditioned representations of value. Nat Neurosci 7 (10), 1144–1152. http://dx.doi.org/10.1038/nn1314.

Haber, S.N., Knutson, B., 2010. The reward circuit: linking primate anatomy and human imaging. Neuropsychopharmacology 35, 4–26.

Hare, T.A., Camerer, C.F., Rangel, A., 2009. Self-control in decision-making involves modulation of the vmPFC valuation system. Science 324, 646–648.

Huettel, S.A., Stowe, C.J., Gordon, E.M., Warner, B.T., Platt, M.L., 2006. Neural signatures of economic preferences for risk and ambiguity. Neuron 49 (5), 765–775. http://dx.doi.org/10.1016/j.neuron.2006.01.024.

Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 97, 263–291.

Kahneman, D., Tversky, A., 1984. Choices, values and frames. American Psychologist 39 (4), 341–350.

Keysers, C., Wicker, B., Gazzola, V., Anton, J.L., Fogassi, L., Gallese, V., 2004. A touching sight: SII/PV activation during the observation and experience of touch. Neuron 42 (2), 335–346.

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., Montague, P.R., 2008. The rupture and repair of cooperation in borderline personality disorder. Science 321, 806–810.

Koechlin, E., Ody, C., Kouneiher, F., 2003. The architecture of cognitive control in the human prefrontal cortex. Science 302, 1181–1185.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., Damasio, A., 2007. Damage to the prefrontal cortex increases utilitarian moral judgements. Nature 446, 908–911.

Krugel, L.K., Biele, G., Mohr, P.N., Li, S.C., Heekeren, H.R., 2009. Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. Proceedings of the National Academy of Sciences of the United States of America 106, 17951–17956.

Laureiro-Martinez, D., Brusoni, S., Canessa, N., Zollo, M., 2014a. Understanding the exploration–exploitation dilemma: an MRI study of attention control and decision-making performance. Strategic Management Journal.

Laureiro-Martinez, D., Canessa, N., Brusoni, S., Zollo, M., Hare, T., Alemanno, F., Cappa, S.F., 2014b. Frontopolar cortex and decision-making efficiency: comparing brain activity of experts with different professional background during an exploration-exploitation task. Frontiers in Human Neuroscience 7, 927.

Legrenzi, P., Umiltà, C., 2011. Neuromania: On the Limits of Brain Science. Oxford University Press, Oxford.

Levallois, C., Clithero, J.A., Wouters, P., Smidts, A., Huettel, S.A., 2012. Translating upwards: linking the neural and social sciences via neuroeconomics. Nature Reviews Neuroscience 13, 789–797.

Li, J., Delgado, M.R., Phelps, E.A., 2011. How instructed knowledge modulates the neural systems of reward learning. Proceedings of the National Academy of Sciences of the United States of America 108, 55–60.

Macmillan, M., 2002. An Odd Kind of Fame: Stories of Phineas Gage. The MIT Press, Cambridge, MA.

McCabe, K., Houser, D., Ryan, L., Smith, V., Trouard, T., 2001. A functional imaging study of cooperation in two-person reciprocal exchange. Proceedings of the National Academy of Sciences of the United States of America 98, 11832–11835.

Miniussi, C., Ruzzoli, M., 2013. Transcranial stimulation and cognition. Handbook of Clinical Neurology 116, 739–750.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., Dolan, R.J., 2004. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304, 452–454.

O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H., Dolan, R.J., 2003. Temporal difference models and reward-related learning in the human brain. Neuron 38, 329–337.

Padoa-Schioppa, C., Assad, J.A., 2006. Neurons in the orbitofrontal cortex encode economic value. Nature 441, 223–226.

Pine, A., Shiner, T., Seymour, B., Dolan, R.J., 2010. Dopamine, time, and impulsivity in humans. Journal of Neuroscience 30, 8888–8896.

Plassmann, H., O'doherty, J., Rangel, A., 2007. Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. Journal of Neuroscience 27, 9984–9988.

Premack, D., Woodruff, G., 1978. Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences 4, 515–526.

Reynolds, S.M., Zahm, D.S., 2005. Specificity in the projections of prefrontal and insular cortex to ventral striatopallidum and the extended amygdala. Journal of Neuroscience 25 (50), 11757–11767. http://dx.doi.org/10.1523/JNEUROSCI.3432-05.2005.

Roiser, J.P., De Martino, B., Tan, G.C., Kumaran, D., Seymour, B., Wood, N.W., Dolan, R.J., 2009. A genetically mediated bias in decision making driven by failure of amygdala control. Journal of Neuroscience 29, 5985–5991.

Rolls, E.T., 2000. The orbitofrontal cortex and reward. Cereb Cortex 10 (3), 284–294.

Rudorf, S., Preuschoff, K., Weber, B., 2012. Neural correlates of anticipation risk reflect risk preferences. Journal of Neuroscience 32, 16683–16692.

Rustichini, A., 2009. Neuroeconomics: what have we found, and what should we search for. Current Opinion in Neurobiology 19, 672–677.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., Cohen, J.D., 2003. The neural basis of economic decision-making in the Ultimatum Game. Science 300, 1755–1758.

Schott, B.H., Niehaus, L., Wittmann, B.C., Schutze, H., Seidenbecher, C.I., Heinze, H.J., Duzel, E., 2007. Ageing and early-stage Parkinson's disease affect separable neural mechanisms of mesolimbic reward processing. Brain 130, 2412–2424.

Schultz, W., 2007. Behavioral dopamine signals. Trends in Neurosciences 30, 203–210.

Schultz, W., 2013. Updating dopamine reward signals. Current Opinion in Neurobiology 23, 229–238.

Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science 275, 1593–1599.

Sharp, C., Monterosso, J., Montague, P.R., 2012. Neuroeconomics: a bridge for translational research. Biological Psychiatry 72, 87–92.

Shenhav, A., Greene, J.D., 2010. Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. Neuron 67, 667–677.

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.

Thut, G., Schultz, W., Roelcke, U., Nienhusmeier, M., Missimer, J., Maguire, R.P., Leenders, K.L., 1997. Activation of the human brain by monetary reward. Neuroreport 8, 1225–1228.

Tobler, P.N., O'doherty, J.P., Dolan, R.J., Schultz, W., 2007. Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. Journal of Neurophysiology 97, 1621–1632.

Tom, S.M., Fox, C.R., Trepel, C., Poldrack, R.A., 2007. The neural basis of loss aversion in decision-making under risk. Science 315 (5811), 515–518. http://dx.doi.org/10.1126/science.1134239.

Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. Science 185, 1124–1131.

Tversky, A., Kahneman, D., 1981. The framing of decisions and the psychology of choice. Science 211 (4481), 453–458.

Ullsperger, M., Danielmeier, C., Jocham, G., 2014. Neurophysiology of performance monitoring and adaptive behavior. Physiological Reviews 94, 35–79.

von Neumann, J., Morgenstern, O., 1953. Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ.

Neuroendocrinology

Petra Netter, University of Giessen, Giessen, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.V. Martin, volume 16, pp. 10584–10588, © 2001, Elsevier Ltd.

Abstract

Neuroendocrinology refers to the interaction of hormones related to brain functions, but also to the release and inhibition of other hormones affecting somatic functions. Basic principles of structure, transport, metabolism, and functions are elucidated, as well as the multiple interactions between hormones synthesized within the nervous system and those produced in peripheral glands and tissue cells of the organism. It is demonstrated that most hormones produced in the periphery also have receptors and functions in the brain. Furthermore, feedback loops, structural and functional relationships between hormone systems are described, as well as the role of hormones as indicators of transmitter functions.

Introduction

Neuroendocrinology refers to the interaction between the nervous and the endocrine system and can therefore be viewed as representing a link between endocrinology, concerned with effects of hormones released from endocrine glands on somatic functions of organs and diseases, and psychoneuroendocrinology, relating the mutual influences between the nervous system and hormones to psychological processes and dispositions. Therefore, some 'somatic' as well as 'psychological' aspects of the neuroendocrine system will be mentioned.

Definitions

The nervous (neurological) system is composed of the brain, spinal cord, and peripheral nervous system divided into the sympathetic and parasympathetic parts. The endocrine system refers to hormones secreted from glands as well as from neurons in the brain and peripheral nerves, and also includes hormones produced in tissue cells distributed in different organs. Hormones may act directly on specific target cells, may serve as releasers and inhibitors of other hormones, or may act as transmitters of electrical impulses in neurons or as cotransmitters (neuromodulators). Hormones, like the nervous system, serve the function of information processing in the organism and have in common that they are synthesized, secreted, and the ones produced remote from their target cells have to be transported by the bloodstream or by neurons. This gave rise to their name hormones (Greek hormao = I move). For exerting their action, most hormones have to bind to specific protein structures of the target cells called receptors.

History

Modern Neuroendocrinology may be viewed as a renaissance of the ancient Greek theory of the physician Hippocrates that body fluids are responsible for diseases as well as for temperament (blood = sanguinics, black bile = melancholics, yellow bile = cholerics and mucous = phlegmatics).

This was revived in the nineteenth century when interest arose in 'internal secretion' represented by the works of Claude Bernard, Thomas Addison, and Charles-Edouard Brown-Séquard, when the role of the thyroid in 1836, the adrenals in 1840, and the parathyroid organs in 1890 were first described (Cowadias, 1941). The role of certain substances isolated from blood and urine as indicators of functions of endocrine glands became evident when animal research revealed that extirpation of glands like the ovaries, the pancreas, or the thyroid led to certain deficiencies and diseases. Therefore, with growing interest in diseases related to internal glands, the relevance of these substances for therapeutic efforts increased. It was first pointed out by Brown-Sequard in 1889, although in the beginning only extracts from total glands could be used for therapy. The term 'hormone' was only created in 1905 by Ernest Henry Starling. In the first half of the twentieth century, most of the important hormones like glucocorticoids, gonadal hormones, growth hormone, and thyroid hormones had been identified. The interaction between hormones, their role in metabolism and, in particular, their control by the nervous system, and relations to psychological processes were only discovered and analyzed in detail during the first half of the twentieth century. With the analysis of their chemical structures, it became possible to use synthetic hormones and their derivatives for therapy most of which only became available after World War II.

Classification and Biochemical Properties of Hormones

Hormones may be classified according to their *structure* (polypeptides, glycoproteins, amino acid derivatives, and steroid hormones), their *site of synthesis* (neurohormones = produced in neural tissue, glandular hormones = produced in endocrine glands, and hormones produced in their target tissues (paracrine secretion)), and according to their *function*. The latter refers to growth, maturation reproduction, metabolism, circulation, regulation of temperature, balance of water and electrolytes, as well as to feeding, sleep, pain, affective states, and personality traits. There is large overlap between functions of different hormones and most of them are responsible for several different functions. They are often divided into endocrinologically acting hormones and cotransmitters, which are

Table 1 Glandular hormones according to sites of production and chemical classes

Organ or tissue	*Hormone or neuropeptide released*	*Chemical class*
Pituitary: anterior (glandular) part posterior part (=neuropituitary)	LH; FSH, ACTH, TSH, STH, PRL ADH (=AVP) OXT (=oxytocin)	Peptides, composed of amino acids
Parathyroid gland	Parathyroid hormone (parathyrin)	
Pancreas, Langerhans islets	Insulin, glucagon, somatostatin	
Pineal gland	Melatonin	
Thyroid	Thyroxin (T4); triiodothyronin (T3)	Tyrosine derivatives
Adrenal medulla	Adrenalin, noradrenaline, dopamin	
Adrenal cortex	Glucocorticosteroids, mineralocorticosteroids, androsteroids	Steroids, derived from cholesterol
Ovary	Estrogens, gestagens	
Testes	Androgens	

LH = luteinizing hormone, FSH = follicle stimulating hormone, ACTH = adrenocorticotropic hormone TSH = thyroid stimulating hormone, STH = somatotropic (=growth) hormone, PRL = prolactin, ADH = antidiuretic hormone = AVP = arginin vasopressin, OXT = oxytocin.

released in combination with transmitters and exert local paracrine functions by supporting the action of transmitters. Table 1 provides an overview according to their site of biosynthesis and class of chemical structure (for details, *see* Hypothalamus; Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of)

The hormones are listed according to the three major classes of chemical structures: peptides (which include polypeptides and the glycoprotein hormones FSH, LH, and TSH = peptides carrying a carbohydrate side chain), tyrosine-based hormones (comprising catecholamines and thyroid hormones), and steroids (glucocortico- and mineralocortico-steroids and gonadal steroids produced in the adrenal cortex) and gonads.

The major biochemical and physiological differences between these classes of hormones are listed in Table 2.

It becomes evident that, although peptides and catecholamines belong to different classes of structures, they may follow similar principles. They are both released by exocytosis (ejection of the substance from vesicles in the cell) as opposed to the thyroid hormones and steroids, which leave the site of synthesis by diffusion. Similarly, the groups differ in their transport in fluids, which occurs as the free hormone or bound to a transport protein. Type of transportation may vary within the same hormone, depending on the type of fluid (steroids are bound to transport proteins in plasma, but are found in their unbound state in saliva, which is the effective form). The first two groups also differ from the third and fourth classes with respect to their ability to pass the blood–brain barrier, a structure that enables lipophilic substances like the thyroid hormones and steroids to enter the brain, while the hydrophilic classes of peptides and catecholamines are blocked. Similarly, differences in location of respective receptors and receptor

Table 2 Biochemical properties of the major classes of hormones

Property	*Peptides*	*Catecholamines*	*Thyroid hormones*	*Steroids*
Type of secretion	Exocytosis		Diffusion	
Transport	Free compound		Bound to plasma proteins	
Passage through blood–brain barrier	No		Yes	
Receptors				
Localization	Cell membrane		Nucleus	
Mode of action	Activation of second messengers		Control of transcription and mRNA	
Half life	Minutes–hours	Seconds	Days	Hours
Duration of action	Minutes–hours	Seconds–minutes	Days	Hours–days
Metabolic degradation	Proteolysis in plasma	By enzymes COMT, MAO	Glucoronidation and sulfatation in the liver	

mRNA = messenger ribonucleic acid, COMT = catechol-*o*-methyltransferase, MAO = monoamine oxidase.
Modified from Voigt, K.H., 2001. Endokrines system. In: Klinke, R., Silbernagel, S. (Eds.), Lehrbuch der Physiologie, third ed. Thieme, Stuttgart, p. 449.

functions are identical for peptides and catecholamines, on the one hand, and thyroid hormones and steroids, on the other. Yet, all classes differ in half-life (time until the concentration of the active compound measured in plasma has been reduced to half of its initial value). So catecholamines can only be measured for seconds after their release, followed by peptides, while steroids and thyroid hormones can be assessed hours or even days after release. These are important limitations in psychophysiological research, when activity of hormones is investigated by stimulating their release by stressors or drugs. There are also differences in catabolism: Cleavage of peptides to single amino acids is achieved by proteolysis, the catecholamines are metabolized by the monoamine-oxidizing enzymes MAO A and MAO B in brain and sympathetic nerves, and the thyroid hormones and steroids are metabolized by liver enzymes.

Functions of Glandular Hormones

The Pituitary

Most hormones have many different functions and most functions are mediated by the interplay and cooperation of several hormones. So hormones of an identical class, according to Table 1, may serve very different functions: Within the peptides released from the anterior lobe of the pituitary, the four glandotropic hormones *ACTH, FSH, LH, and TSH* stimulate the release of hormones in the adrenal cortex, the gonadal, and the thyroid gland (see Figure 1). ACTH is also important, because fractions of its amino acids (e.g., ACTH 4–9) have been shown to enhance memory (de Wied et al., 1993). *PRL* stimulates growth of the mammalian gland and, after rising at the end of pregnancy stimulates lactation after delivery, but it is also produced in males, responds to stressors, and is inhibited by dopamine. The most diversified functions are observed for

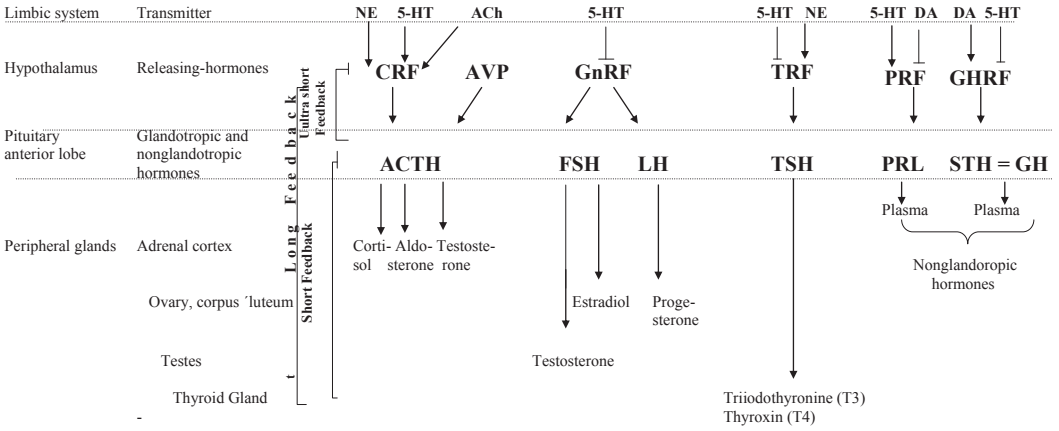

Figure 1 Regulation of hormones by limbic–hypothalamo–pituitary processes (↓ = stimulation; ⊥ = inhibition). *Transmitter level*: NE = norepinephrine, 5-HT = serotonin, DA = dopamine, Ach= acetylcholin. *Hypothalamic level*: CRF = corticotropin releasing factor, AVP = arginine vasopressin, ADH = antidiuretic hormone, GnRF = gonadotropin releasing factor, TRF = thyreotropin releasing factor, PRF = prolactin releasing factor, GHRF = growth hormone releasing factor. *Pituitary level:* ACTH = adrenocorticotropic hormone, FSH = follicle stimulation hormone, LH = luteinizing hormone, TSH = thyroid stimulating hormone, PRL = prolactin, STH = Somatotropic hormone, GH = growth hormone. Modified from Netter, P., 2005. Endokrines System und Persönlichkeit. In: Hennig, J., Netter, P. (Eds), Biopsychologische Grundlagen der Persönlichkeit. Spektrum Akademischer Verlag/Elsevier, p. 293.

the growth hormone *STH* (*GH*). It is responsible for growth of the bones during puberty and hence for protein synthesis, which needs support from the insulin-like growth factor IGF1 produced in the liver. STH also promotes lipolysis (breakdown of lipids) and may stimulate insulin production in pancreatic B-cell in a first step, but in a second step may inhibit transport of glucose into cells (antagonism to insulin).

The peptides *adiuretin* (*vasopressin* (*AVP*)) and *oxytocin* (*OXT*) produced in the posterior lobe of the pituitary (the neuropituitary consisting of neural tissue) and directly released into the bloodstream, again have different functions: AVP is responsible for water retention in the kidney and in higher concentrations for vasoconstriction, while oxytocin stimulates uterus contractions during delivery and participates in stimulation of lactation. Because of their particular relevance to behavior, these two peptides deserve a more thorough presentation. These two peptides are synthesized in the hypothalamus and transported to the neuropituitary by axonal transport from, where they are released into circulation. *Adiuretin* and oxytocin consist of nine amino acids each, differing only in two amino acids. Yet, there are several differences in functions: AVP, besides affecting reabsorption of water via V2 receptors at the site of the kidneys and exerting vasoconstrictive actions via V1 receptors at the arterial vessels, it also reaches the hypothalamus via the eminentia mediana, a paraventricular organ, where it supports CRF in stimulating ACTH as a hypothalamic releasing hormone of the HPA axis (see **Figure 1**), and thereby influencing its various psychological functions in relation to stress and emotions. In addition, AVP, like some other peptides, has been shown to enhance memory by acting on the ventral hippocampus, a process, which, however, is highly dependent on the time of application (at the beginning or at the end of the learning process) (Alescio-Lautier et al., 2000). AVP shares this observation with oxytocin and a number of other peptides and also with the ACTH fraction of amino acids ACTH 4–9 (de Wied et al., 1993). In addition, vasopressin has been investigated together with *oxytocin* with respect to social behavior. Oxytocin, known to be responsible for uterus contractions during labor and for ejection of milk from the mammarian gland, has been shown to be responsible for social bonding in animals (Insel et al., 1998) and humans (Uvnes-Moberg,1996; Heinrichs et al., 2009; Ebstein et al., 2012). This concerns mother–child interaction, sexual partnership as well as interpersonal trust in experimentally induced social interactions and social perception. It has always been discussed, if the two hormones, due to their structural similarity, act in the same direction or display opposite effects, oxytocin eliciting interpersonal trust, reducing anxiety, counteracting stress induced cortisol increases, perceiving threatening social stimuli as less aversive, and adiuretin being associated with anxiogenic interpretation of stimuli and aggressive feelings and behavior. Any reported differences may depend on the dimorphic actions of the two hormones, because some experimental findings obtained in males do not apply to females and vice versa. Nowadays, findings have partly been validated by neuroimaging studies (demonstrating e.g., that behavioral responses to the peptides are associated with changes in the activity of the amygdala) as well as by molecular genetics relating behavioral findings to polymorphisms of the OXT and the AVP receptors (Meyer-Lindenberg et al., 2011)

The Pineal Gland

Melatonin does not actually belong to the peptides because it is derived from serotonin, a transmitter based on the amino acid tryptophan and, unlike peptides, is lipophilic. Furthermore, the pineal gland, like the neuropituitary part, belongs to neuronal tissue. It is part of the circumventricular organs of the brain (see Section The Adrenal Cortex). Melatonin is inhibited by light and secreted at night. It regulates the day–night rhythm acting on the M1 and M2 receptors at the suprachiasmatic nucleus in

the hypothalamus (*see* Hypothalamus). Therefore, it has been used as a drug for compensating jet lag or disturbances of other rhythms underlying psychiatric diseases like depression. In addition, multiple actions of melatonin on many other tissues and systems like the immune system have contributed to its unjustified labeling as an antiaging substance.

The Parathyroid Gland

The peptide of the parathyroid gland, *parathyrin*, is responsible for the calcium–phosphate homeostasis regulating the stability of bones in cooperation with the tissue hormones calcitonin and calcitrole derived from vitamin D, but, as opposed to insulin, is not relevant for brain activities.

The Pancreas

The pancreatic cells type B of the Langerhans islets produce *insulin*, predominantly released upon consumption of carbohydrates and high levels of blood glucose. This is responsible for transport of glucose from the blood into somatic cells thus decreasing blood glucose, but it also has anabolic effects on fat and protein metabolism. In addition, insulin has receptors in the hypothalamus, where in combination with the tissue hormone leptin it may reduce food intake (satiety) and induce energy metabolism when both insulin and leptin reach high brain levels, while the reverse is observed for low brain levels (→hunger and increased food intake). Pancreatic A cells produce *glucagon*, an antagonist of insulin, which is released when protein rich food is consumed and blood glucose levels are low. It catabolizes glycogen stored in liver and muscles thereby increasing blood glucose. It can be modified by activation of catecholamine receptors forming a link between pancreatic activity and the autonomic nervous system and is also found in brain. Finally, pancreatic D cells produce *somatostatin*, which inhibits both insulin and glucagon when produced in the pancreas, but is also released from the hypothalamus as an inhibitor of STH in the pituitary.

The Thyroid Gland

The hormones released from the thyroid (tyrosine-based hormones) *thyroxin* and *triiodothyronine* (*T4 and T3*) need iodine for their synthesis and are transported protein bound to receptors into almost all cells of the organism, where T4 is changed into T3. This increases basic metabolism in interaction with the steroids cortisol and aldosterone as well as with the catecholamines leading to sympathetic stimulation of heart rate, temperature, and general physiological arousal.

The Adrenal Medulla

The *catecholamines noradrenalin, adrenalin, and dopamine,* released from the adrenal medulla (and noradrenine additionally from sympathetic neurons) *are* also synthesized from the amino acid tyrosine and structurally similar to each other. Besides their sympathetic action on circulation and metabolism in the periphery, they are synthesized in neurons of the brain where they act as neurotransmitters (*see* Hypothalamus) and neuromodulators by their effects on the hypothalamic releasing and inhibiting hormones of the pituitary (see Figure 1).

The Adrenal Cortex

Similarly, the steroids *cortisol and aldosterone*, produced in the adrenal cortex by stimulation of ACTH from the pituitary and the gonadal steroids *estrogen and testosterone* stimulated by LH and FSH respectively (see Figure 1) have many different functions in reproduction and metabolism, which, however, have been described in the article Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of.

Most hormones secreted from peripheral glands also have receptors in the brain. This refers to the steroids from the adrenal cortex, thyroid hormones, the peptides of the pituitary, melatonin from the pineal gland, and insulin secreted from the pancreas. Steroids from plasma can pass the blood–brain barrier due to their lipophilic structure, some of the peptides released from the pituitary may reach the brain via receptors at the circumventricular regions (locations not surrounded by the blood–brain barrier, which cannot be penetrated by peptides) This means that besides their functions in metabolism, circulation, reproduction, or their role as releasers or inhibitors of other hormones, they can modify mental functions and behavior. There is an abundant bulk of research concerning the role of hormones in behavioral, emotional, and cognitive responses to stress, in personality traits, or in psychiatric diseases, published in specialty Journals like *Psychoneuroendocrinology*.

Neuropeptides, Tissue Peptides, and Neurosteroids

Besides the peptides listed as hormones from the neuropituitary, ADH and OXT, there are *neuropeptides* produced in neurons of the hypothalamus acting as releasers or inhibitors of the pituitary hormones (CRF, GnRF, TRF, GHRF) (see Figure 1). Of these, the *corticotropin releasing factor* (*CRF*) deserves special interest, because apart from being part of the HPA axis by releasing proopiomelanocortin (POMC) in the pituitary from which ACTH, α-MSH, and the β-endorphin are derived, CRF has direct effects on sympathetic activity and the adrenal medulla thus being responsible for release of catecholamines. Moreover, it reduces food and water intake and increases locomotion and anxiety and is therefore an essential hormone in stress responses.

Similarly multiple functions are exerted by the inhibiting hormone of STH, *somatostatin (growth hormone inhibiting factor GHIF)* an antagonist of the releasing factor GHRF depicted in Figure 1 Somatostatin also inhibits TSH and ACTH in the pituitary, renin in the kidneys, a number of gastric peptides, and the pancreatic peptides described above, but also inhibits release of acetylcholine in the brain and hence electroencephalographic (EEG) activity (Klinke and Silbernagel, 2001: p. 466).

There is a large number of additional brain-derived peptides involved in several central nervous functions, like α-melanotropin (α-MSH, responsible for reduction of food intake) and β-endorphin (reduction of pain), both released by CRF in the

hypothalamus (see above), or the enkephalins (also acting as endogenous opioids), and many others responsible for nerve growth and neuroprotection. But, in addition, an ever growing number of the so-called *tissue hormones* produced in tissues of the gut, lung, liver, heart, kidney, and adipose tissue, has been detected, which partly act on the surrounding tissue (paracrine action) or by stimulation and inhibition of their own synthesis and release (autocrine action). Many of them have additionally been found in the brain, predominantly in the hypothalamus regulating, for instance, hunger and satiety (e.g., leptin, neuropeptide Y (NPY)). Also substance P and the vasoactive intestinal polypeptide (VIP) are produced in neurons, acting as neurotransmitters for stimulation or inhibition of gastrointestinal secretion or muscle contractions (substance P, VIP). Even neuropeptides, which are only produced in the gut and not in neurons (cholecystokinin (CCK) and neurotensin (NT)) may, in addition to their role in digestion, have psychological effects like regulation of feeding and appetite. Several of these act as cotransmitters (neuromodulators, i.e., enhancing or reducing transmitter activity). For example, NPY is a cotransmitter of noradrenaline, adrenalin, and GABA; CCK of GABA; substance P and VIP of acetylcholine.

In addition to neuropeptides, the so-called *neuroactive steroids* are produced in the brain. They are either synthesized in neurons of the brain or enter the brain as metabolites of adrenocortical or gonadal steroids related to each other as depicted in Figure 2. Pregnenolone, the general precursor of the four major steroids of the adrenal cortex as well as dehydroepiandrosterone (DHEA) and derivatives of progesterone have been identified as neuroactive steroids. Their role in neuroactivity is modulation of neuronal excitability, for instance, by potentiating or reducing GABA-receptor mediated activity. So, for example, two derivatives of progesterone of very similar structure act antagonistically on the GABA A receptor and have been found to be increased or decreased in depression, respectively (Schüle et al., 2011). Thus, these compounds have been used therapeutically in the treatment of depression and anxiety. Like glandular steroids, neurosteroids act on intracellular receptors and can, therefore, also regulate gene expression (Dubrovsky, 2006).

Feedback Systems

Hormones, like many other systems of the body governed by the principle of homeostasis (e.g., blood glucose, body temperature), are characterized by *feedback* processes. The most popular one is that peripheral levels of a hormone signal to the site of synthesis that further production should be reduced (negative feedback) and vice versa (positive feedback). This is the case with paracrine actions in tissue hormones in somatic cells as well as with transmitters in the brain, where reduction or increase of hormone release are reflected not only by regulation of synthesis and release, but also by up- and downregulation of respective receptors (low levels → upregulation = increase in number of receptors, high levels → downregulation = reduction in number of receptors, *see* Neurotransmitters). This principle also applies to the feedback loops depicted in Figure 1.

In Figure 1 the three major systems organized along a four step axis: the hypothalamo–pituitary–adrenal (HPA) axis (*see* Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of), the hypothalamo–pituitary–gonadal (HPG) axis, and the hypothalamo–pituitary–thyroid (HPT) axis were depicted. In addition, the regulation of the two hormones of the posterior lobe of the pituitary, adiuretin and oxytocin, were included, which do not act on a subordinate gland, but are also governed by releasing and inhibiting hormones from the hypothalamus. The reason for calling the system as a four step system is that neurotransmitters from other parts of the brain, mostly from the limbic system, have been identified as stimulating and inhibiting factors of the hypothalamus, which gave rise to change of names into LHPA-, LHPG- and LHPT-axes. For reasons of space, the inhibiting factors of prolactin (prolactostatin (PIF)) and growth hormone (somatostatin (SIH)) produced in the hypothalamus have not been illustrated. As shown in the figure, feedback loops can be observed from the pituitary level to the hypothalamus (ultrashort feedback), from the peripheral gland to the stimulating hormones from the pituitary (short feedback), and directly from plasma levels of the peripheral hormones to the hypothalamus. Feedback principles like these are used for diagnostic tests assessing disturbances in hormone release as in the case of the dexamethasone or the dexamethasone–CRH test used for testing the feedback of cortisol responses to CRF sensitivity (*see* Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of).

Rhythms

In addition to this demand adapted regulation, most hormones are also governed by rhythmic release concerning amplitude of levels as well as number of pulsatile secretions. These may follow an ultradian, circadian, or lunar pattern. *Ultradian cycles* refer to changes of several times within a day. Pulsatile rhythms of secretion of about 12–15 min are reported for insulin, a rhythm of about 90 min (called rest activity cycle, RAC, characteristic of many systems), for instance, applies to GnRF releasing LH, which follows at the same 90 min intervals.

This rhythm may also exist in addition for hormones governed by *circadian cycles* like ACTH, cortisol, or melatonin as well as to pituitary peptides like β-endorphins, which change according to a 24 h (circadian, day–night) rhythm, not necessarily depending on sleep–waking stages. These cycles may be synchronized or desynchronized by external stimuli-like daylight or temperature. Other rhythms are rather coupled to patterns of sleep like the pituitary peptides STH, PRL, and TSH as well as LH and FSH. Acrophases (highest concentrations) for ACTH, cortisol, and β-endorphins are observed between 6 and 9 a.m. and for the sleep dependent hormones between midnight and 2 a.m.

A *lunar cycle* is represented by the hormones of the gonadal axis, where estradiol and in consequence LH are low during the menstrual phase, rise in the follicular phase, and have their peak at ovulation on day 12–14 of the menstrual cycle. Estradiol releases LH, and drops much more slowly than LH during the luteal phase, when progesterone released from the corpus luteum rises to higher levels than estradiol.

But even longer periods of several months or years organized according to seasons like the gonadal hormones in animals, or rhythms depending on the phase of life are

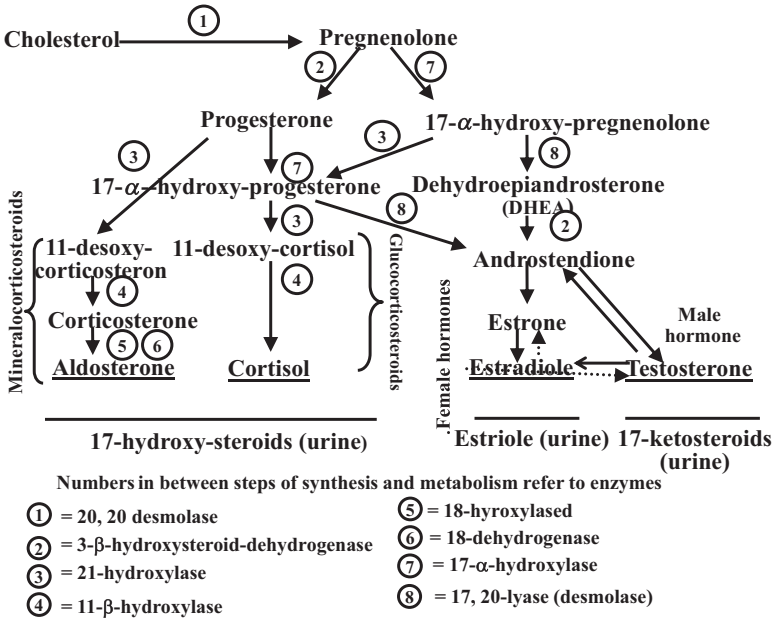

Figure 2 Relationships between steroid hormones of the HPA axis (left) and the HPG axis (right). Modified from Netter, P., 2008. Neurochemische und endokrine Systeme in der experimentellen Emotionsforschung: Forschungsansätze und Grundlagen. In: Janke, W., Schmidt-Duffy, M., Debus, G. (Eds.), Experimentelle Emotionsforschung. Pabst Science Publishers, Lengerich, Berlin, p. 628.

observed: E.g. STH and testosterone both have their acrophase around 20 years of age, STH decreasing thereafter, while testosterone only drops at the age of about 40–50 in males. In pregnancy, extra rhythms are determined by hormones produced in the placenta, where human chorionic gonadotropin (HCG) shows a steep rise until week 6 of gestation (which can be used as a test of pregnancy in urine), while estradiol and progesterone as well as human placental lactogen (HPL) start rising from about week 12 to the end of gestation. Disturbances of these rhythms represent symptoms of certain diseases, like depression or infertility. Even more relevant are disturbances in amplitude of hormone production, since in practically all major hormones hyper- as well as hypoactivity can cause specific endocrinological diseases.

Interactions between Hormone Systems

It must be pointed out that the hormones representing specific functions in metabolism, reproduction, and behavior are heavily interrelated with each other and with additional hormones and transmitters. Interactions between the two systems most relevant for behavior, the HPA and the GPA axes, may be selected for demonstrating that hormones may not only be functionally, but also structurally related. Figure 2 demonstrates structural relationships between the two major products of the HPA axis, the glucocorticoid cortisol and the mineralocorticoid aldosterone, and the two final compounds of the GPA axis, the female hormone, estradiol and the male hormone, testosterone.

Belonging to the class of steroids, they are all derived from the basic molecule cholesterol and end up in different metabolites in urine, which can be used for assessing endocrine effects from experimental manipulations or endocrine diseases. The similarity of action of the steroids may be explained by the fact that progesterone, the hormone of the corpus luteum, and adrenal cortex can be transformed into practically all of the final hormones (see Figure 2). Similarly, testosterone (T) can be transformed into estradiol (E2) in fatty tissue, which is relevant for the observation that older males may develop female breasts (the reverse process (T into E2) is still questioned).

Yet, when considering the functional interaction of the HPA and HPG axes, they may exhibit antagonistic properties This is due to (1) direct inhibitory effects of CRH on GnRF, (2) an inhibitory action of cortisol on GnRF, FSH, and LH, (3) an inhibitory effect of the CRH-stimulated catecholamines noradrenaline and adrenalin on GnRF, and (4) an inhibitory effect of CRH stimulated endogenous opiates on GnRF. Furthermore, 5-HT stimulates release of CRF, but inhibits release of GnRF. This becomes evident by increase of cortisol (C) and suppressions of testosterone (T), when serotonergic substances are applied, which may explain the antagonistic responses of cortisol and testosterone to controllable and uncontrollable stressors (controllable stress: decrease or no change of C and increase of T, uncontrollable stress: increase of C and decrease of T, (Henry, 1986)).

Another example of interaction between two neuroendocrine systems is the mutual influence of the adrenal medulla and adrenal cortex: cortisol locally enhances the synthesis of noradrenaline and adrenalin in the adrenal medulla by stimulating three of the enzymes involved in transforming the amino acid tyrosin across dopa and dopamine into noradrenaline, which is finally converted into adrenalin, and vice versa, the catecholamines act on cortisol via stimulation of the HPA axis (see Figure 1).

Table 3 Principles of transmitter challenge tests, classes of substances used, and corresponding hormone responses

Principle of action	*Serotonin (5-HT)*		*Dopamin (DA)*		*Noradrenaline (NE)*	
	Substance (example)	*Hormone response*	*Substance (example)*	*Hormone response*	*Substance (example)*	*Hormone response*
Precursor for synthesis	L-Tryptophan	PRL↑ Cortisol (↑) GH (↑)	L-Dopa	(PRL ↓)		
Blockade of synthesis	Tryptophan depletion	PRL ↓	AMPT	(HVA ↓)		
Release	D-Fenfluramine	PRL ↑ Cortisol (↑)	Amphetamine	(PRL ↓)	Reserpine	?
Reuptake inhibition	Citalopram, Fluoxetine	Cortisol ↑	Mazindole	(PRL ↑)!!* GH ↑	Reboxetine	Cort ↑ GH ↑
Receptor agonism	Ipsapirone 5 – HT_{1a} receptor	Cortisol ↑	BromocriptineD_2 receptor Apomorphine D_2D_1 receptor	PRL ↓ GH (↑) PRL ↓ GH ↑	α_2:Clonidine	TSH ↑ GH ↑
Receptor antagonism	Cyproheptadine – HT_{2a} receptor	Cortisol ↓	Fluphenazine D_2D_1 receptor	PRL ↑		

!!* = dopamine increase would normally suppress PRL, here: opposite effect, PRL = prolactin, Cort = cortisol, GH = growth hormone, HVA = homovanillic acid, AMPT = α-methylparatyrosin, TSH = thyroid stimulating hormone, 5-HT = serotonin, D = dopamine; ? = no hormone response confirmed, weak or contradictory response or response only with higher doses, ↑ increase, ↓ decrease.
Modified from Netter, P., 2008. Neurochemische und endokrine Systeme in der experimentellen Emotionsforschung: Forschungsansätze und Grundlagen. In: Janke, W., Schmidt-Duffy, M., Debus, G., (Eds.), Experimentelle Emotionsforschung, Pabst Science Publishers, Lengerich, Berlin, p. 625.

Hormones as Indicators of Neurotransmitter and Hypothalamic Activity

The catecholamines produced in the adrenal medulla mentioned above, do not only act as hormones in the periphery, but are also produced in neurons in the brain stimulating or inhibiting synthesis or release of hypothalamic and pituitary hormones as shown in Figure 1. Therefore, hormones can also be used as indicators of transmitter activity by the application of the so-called *neurotransmitter challenge tests:* a specific pharmacological substance acting on the noradrenergic, dopaminergic, or serotonergic system is applied, which either interferes with synthesis, release, receptors, or metabolism of that particular transmitter, and specific hormone responses on the pituitary level (PRL, GH, TSH) or on the level of glandular responses (mostly cortisol) can be measured in plasma or saliva (in urine only after a delay of several hours, but contaminated by different intermediate effects). Hormone responses have to be measured before and several times after the challenge. Parameters of hormone responses are either measures of amplitude of hormone released (area under the response curve or highest difference from baseline) or parameters of slope of increase and decline (see Reuter et al., 2002) or time of onset of the hormone response. Challenge tests have been used in clinical as well as in experimental research to test for disturbances and individual differences in sensitivity of certain neurotransmitter systems, which are particularly relevant in psychiatric diseases, since there is no direct access to the transmitter systems in the brain in humans. Examples of challenge tests with their substances for testing transmitter activity and corresponding hormone responses are listed in Table 3.

As can be seen from Table 3, tests may refer to synthesis, release, reuptake, and receptor agonistic and antagonistic effects demonstrating different levels of interference with the specific transmitter. Unfortunately, not all hormones suitable as possible indicators of transmitter actions have been tested for all transmitter principles of action and for all transmitter systems, and even, if so, their effects on hormones may not be positively correlated to each other (see e.g., receptor agonists of dopamine). This may be due to the fact that pharmacological challenges by drugs may affect rate of synthesis, rate of release, sensitivity and number of receptors, sensitivity of the transporter (responsible for reuptake of the remaining transmitter from the synaptic cleft after binding of the transmitter to the postsynaptic receptors), and enzymatic catabolism of the transmitter. So, unfortunately, when monitoring a hormone response, one is never sure to which of these processes the hormone response may be attributable. This is even more the case when psychological responses, like performance, emotional reactions, or social behavior are the target of interests, because then the challenge substance itself, the hormone elicited, and the genetic and personality dispositions of the person may be responsible for variations in hormone responses. Yet, challenge tests have revealed insight into the relevance of serotonin, dopamine, and noradrenaline for personality related transmitter responsitivity and psychiatric diseases, most of which are based on disturbances of transmitter systems.

Similarly, hormones of the pituitary can be tested by application of respective hypothalamic stimulating hormones. For example, TRH application can indicate either increased sensitivity of TSH represented by elevated TSH responses demonstrating hypofunction of the thyroid hormones (observed with the disease of myxedema) or, vice versa, reduced TSH responses indicating high thyroxin and T3 levels observed in hyperfunction of the thyroid as observed with Morbus Basedow (an autoimmune disease). By a more indirect

pathway, the application of the corticosteroid dexamethasone can indicate a disturbed feedback loop of the HPA axis (*see* Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of). Many more tests like this make use of the stimulation and feedback loops discussed above.

See also: Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of; Hypothalamus; Limbic System; Neurotransmitters.

Bibliography

Alescio-Lautier, B., Paban, V., Soumireu-Mourat, B., 2000. Neuromodulation of memory in the hippocampus by vasopressin. European Journal of Pharmacology 29, 63–72.

Cowadias, A.P., 1941. History of endocrinology. Proceedings of the Royal Society of Medicine 34, 303–308.

Dubrovsky, B., 2006. Neurosteroids, neuroactive steroids and symptoms of affective disorders. Pharmcology, Biochemistry and Behavior 84, 644–655.

Ebstein, R.P., Mankuta, D., Chew, S.H., Lai, P.S., 2012. The contributions of oxytocin and vasopressin pathway genes to human behavior. Hormones and Behavior 61, 359–379.

Heinrichs, M., von Dawans, B., Domes, G., 2009. Oxytocin, vasopressin, and human social behavior. Frontiers of Neuroendocrinology 30, 448–557.

Henry, J.P., 1986. Neuroendocrine patterns of emotional response. In: Plutchik, R., Kellerman, H. (Eds.), Emotion: Theory, Research, and Experiences, vol. 3. Academic Press, San Diego, pp. 37–60.

Insel, T.R., Winslow, J.T., Wang, Z., Young, L.J., 1998. Oxytocin, vasopressin, and the neuroendocrine basis of pair bond formation. Advances in Medicine and Biology 449, 215–224.

Klinke, R., Silbernagel, S., 2001. Lehrbuch der Physiologie, third ed. Thieme, Stuttgart.

Meyer-Lindenberg, A., Domes, G., Kirsch, P., Heinrichs, M., 2011. Oxytocin and vasopressin in the human brain: social neuropeptides for translational medicine. Nature Reviews in Neuroscience 12, 524–538.

Netter, P., 2004. Personality and hormones. In: Stelmack, R.M. (Ed.), On the Psychobiology of Personality. Elsevier, New York, pp. 353–377.

Netter, P., 2005. Endokrines System und Persönlichkeit. In: Hennig, J., Netter, P. (Eds.), Biopsychologische Grundlagen der Persönlichkeit. Spektrum Akademischer Verlag/Elsevier, pp. 291–396.

Netter, P., 2008. Neurochemische und endokrine Systeme in der experimentellen Emotionsforschung: Forschungsansätze und Grundlagen. In: Janke, W., Schmidt-Duffy, M., Debus, G. (Eds.), Experimentelle Emotionsforschung. Pabst Science Publishers, Lengerich, Berlin, pp. 617–630.

Reuter, M., Siegmund, A., Netter, P., 2002. The trigonometric responder approach: a new method for detecting responders to pharmacological or experimental challenges. Pharmacopsychiatry 35, 182–189.

Schüle, C., Eser, D., Baghai, T.C., Nothdurfter, C., Kessler, J.S., Rupprecht, R., 2011. Neuroactive steroids in affective disorders; target for novel antidepressant or anxiolytic drugs? Neurosciences 191, 55–77.

Uvnes-Moberg, K., 1996. The Oxytocin Factor. Da Capo Press, Cambridge.

Voigt, K.H., 2001. Endokrines system. In: Klinke, R., Silbernagel, S. (Eds.), Lehrbuch der Physiologie, third ed. Thieme, Stuttgart, pp. 443–492.

de Wied, D., Diamant, M., Fodor, M., 1993. Central nervous system effects on the neurohypophyseal hormones and related peptides. Frontiers of Neuroendocrinology 14, 251–302.

Neuroesthetics

Marcos Nadal, University of Vienna, Vienna, Austria
Martin Skov, Copenhagen University Hospital Hvidovre, Copenhagen, Denmark; and Copenhagen Business School, Copenhagen, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Neuroesthetics is a subfield of cognitive neuroscience that studies the biological mechanisms and psychological processes evoked in the creator or the spectator when adopting an esthetic orientation toward an artistic or nonartistic object in the course of interacting with it. These psychological processes are related to perception, cognition, emotion, evaluation, social, and contextual aspects. Here, we outline the scope of neuroesthetics and summarize its historical background. We thereafter sketch three current approaches to neuroesthetics, and examine recent developments in three areas: emotions, context, and expertise. We finish with an exploration of the potential for future inquiry in neuroesthetics.

Introduction: Aims and Scope of Neuroesthetics

Stated briefly, neuroesthetics is a scientific field that aims to understand the neural mechanisms that underlie esthetic behavior. Despite possessing the virtue of conciseness, such a characterization is unsatisfactory because it ignores important conceptual subtleties. First, neuroesthetics can properly be viewed as a subfield of cognitive neuroscience. As such, it combines neuroscience techniques and cognitive science methods (Gazzaniga, 1984), bringing the cognitive and neural levels of explanation of behavior together into a single approach (Churchland and Sejnowski, 1988). Second, rather than regarding esthetics as an abstract concept, neuroesthetics conceives it as a form of human cognition, and thus, it focuses on particular esthetic experiences, esthetic creation, esthetic preference, esthetic choice, and so on. Third, neuroesthetics assumes that such forms of esthetic cognition arise from the interaction of perceptual, emotional, and evaluative processes with social and contextual factors. In fact, understanding such interaction is one of the field's main goals. Fourth, the creation and appreciation of artworks are within the scope of neuroesthetics. However, just as the field of esthetics is concerned with other objects in addition to artworks, so is neuroesthetics. To understand the biological underpinnings of the human capacity for art production and appreciation is the goal of the cognitive neuroscience of art (Seeley, 2011). Both fields overlap, but they are not identical (**Figure 1**). Neuroesthetics deals with a broader set of objects of esthetic interest, such as utensils, commodities, or graphic and industrial designs. The cognitive neuroscience of art, in turn, sets out to understand a broader set of issues related to art, beyond the purely esthetic ones, such as grasping the meaning of an artwork, or understanding its significance in its art-historical or art-critical contexts. Both fields overlap when attempting to understand how esthetic qualities are used to produce and appreciate the content and form of an artwork.

Neuroesthetics, thus, is not concerned with a particular class of objects. Rather, it focuses on a certain kind of way objects can be experienced when people approach them with an esthetic orientation (Cupchik, 1992). From this perspective, neuroesthetics can be conceived as the study of the biological mechanisms and psychological processes evoked in the creator or the spectator when adopting an esthetic orientation toward an artistic or nonartistic object in the course of interacting with it. These psychological processes and their biological underpinnings are related to perception, cognition, emotion, evaluation, as well as to social and contextual aspects (Skov and Vartanian, 2009).

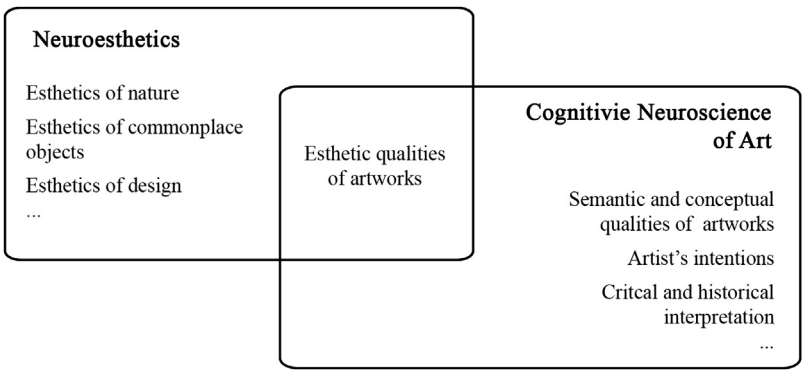

Figure 1 Representation of the relation between neuroesthetics and the cognitive neuroscience of art and their research topics.

Historical Background of Neuroesthetics

The term neuroesthetics was coined at the end of the twentieth century (Zeki, 1999). Scientists and humanists, however, have been interested in the neurobiological substrates of esthetic experience for at least 250 years. Burke (1757) argued that such experiences did not rely on specialized biological mechanisms, that the beautiful and the sublime were mediated by the same physiological mechanisms as common emotions like love and fear. Objects, landscapes, and other people were regarded as beautiful – he believed – because they caused the same relaxation of the nerves as the emotions of love or tenderness. Objects and events that agitated and tensed the nerves, in a similar way to pain, fear, or terror, were experienced as sublime. Two main issues became the focus of scientists' attention during the second half of the nineteenth century. First, the relation between pleasure, pain, and esthetics was extensively discussed. Whereas Marshall (1893) went as far as considering psychological esthetics as a branch of hedonics, and Allen (1877) made the bodily processes of pain and pleasure the cornerstone of his *Physiological Aesthetics*, James (1890) believed that what defined esthetic feelings was precisely the absence of the physiological component of common emotions. Second, the issue of the adaptive value of art and esthetics spurred hypothetical scenarios – based on natural selection or sexual selection – that hoped to explain the origin and evolution of art and esthetics (Darwin, 1871/1998).

It was Fechner's (1876) *Vorschule der Aesthetik*, however, that marked the inception of an empirical science of esthetics. Fechner developed a suite of basic principles that aimed to explain esthetic appreciation. His methodological innovations were, nevertheless, his most lasting and influential contributions. It was he who introduced the now familiar practice of measuring the responses of samples of participants, representative of certain populations, to large numbers of objects. He devised three main methods to elicit his participants' responses: the method of choice, the method of production, and the method of use. The first of these became dominant in the experimental psychology of esthetics, while the use of the other two declined, to the point that most of what empirical esthetics has shown about people's esthetic experience is based on participants' choices or ratings (Westphal-Fitch et al., 2013).

During the twentieth century, the incipient neuroesthetics explored the effects of brain lesions and neurological illnesses on artistic creation. The focus of much of this work was the relation between aphasia and music, painting, and literature (Alajouanine, 1948). Because such reports often described artistic change in single cases after the onset of lesions or illnesses in ambiguous or imprecise terms, it was not easy to arrive at meaningful conclusions or to identify general trends (Chatterjee, 2004). The neuroimaging techniques that became available toward the end of the twentieth century allowed scientists to explore their hypothesis experimentally in controlled situations, with the participation of healthy subjects, and to correlate the appreciation and enjoyment of music, painting, architecture, sculpture, and dance, with neural activity.

Approaches to Neuroesthetics

The biological underpinnings of esthetic behavior have been explored using a broad variety of methods. One approach, which Chatterjee (2011) refers to as parallelism, aims to illustrate how artists throughout the centuries have devised techniques and used resources that catch the attention of spectators, interest them, and appeal to them by engaging certain neural processes related with reward and arousal (Zeki, 1999; Ramachandran and Hirstein, 1999).

Relying on clinical cases, other researchers have described the impact of brain damage and neural degeneration on the production and appreciation of art and on esthetic experiences. This second approach has shown that artists are vulnerable to the same visual, motor, auditory, and cognitive neuropsychological deficits that affect other people, despite their proficient perceptual and motor skills, though they manifest these deficits in strikingly eloquent ways (Chatterjee, 2004). Most artists suffering from neurological disorders continue to be artistically motivated, productive, and expressive after the onset of their condition (Zaidel, 2005). There usually is, however, a noticeable change in the work of most artists who have suffered a stroke. Those sufferings from neurodegenerative disorders, such as Alzheimer's disease, seem to gradually loose the ability to represent the world with precision, but are able to make use of color and form in esthetically appealing ways (Miller and Hou, 2004). Neurological disorders can also affect esthetic appreciation. Depending on the nature of the disorder, people might have difficulties perceiving objects' particular sensory features, recognizing or recollecting them, or engaging emotionally with them. In most cases, however, patients are capable of recognizing and experiencing the esthetic qualities of art and other objects in a meaningful and consistent way (Halpern and O'Connor, 2013). Neurological studies suggest that esthetic preference is surprisingly resilient in the face of neurological conditions, and argue against the existence of specialized brain mechanisms underlying the experience of art (Zaidel, 2005).

The third approach to neuroesthetics noted by Chatterjee (2011) is experimental neuroesthetics. This approach has mostly – though not exclusively – used neuroimaging methods to study the role of different brain regions in esthetic experiences. When applied to neuroesthetics, such methods have usually required participants to judge the beauty, the attractiveness or the appeal of stimuli presented to them, or to state how much they like or prefer them, while their brain activity is being registered. Neuroimaging studies have revealed that neural activity in the reward circuit is a key component of esthetic experience, but the picture is a complex one. Positive or negative esthetic experiences are the result of a complex interplay of neural processes related to reward representation, prediction and anticipation, affective self-monitoring, emotions, and the generation of pleasure, that take place in cortical (i.e., anterior cingulate, orbitofrontal, and ventromedial prefrontal) and subcortical (i.e., caudate nucleus, substantia nigra, and nucleus accumbens) regions, as well as some of the regulators of this circuit (i.e., amygdala, thalamus, and hippocampus) (Cupchik et al., 2009; Harvey et al., 2010; Ishizu

and Zeki, 2013; Kirk et al., 2009b; Lacey et al., 2011; Salimpoor et al., 2011; Salimpoor et al., 2013; Vartanian and Goel, 2004).

However, neuroimaging studies have shown that there is more to esthetic appreciation than pleasure and reward (Nadal and Pearce, 2011). Such studies have also shown that esthetic experiences involve an enhancement of activity in cortical regions related to sensory processing. These regions include the bilateral fusiform gyri, angular gyrus, and the superior parietal cortex in visual esthetic experiences (Cela-Conde et al., 2009; Cupchik et al., 2009; Lengger et al., 2007; Vartanian and Goel, 2004), primary and secondary auditory cortices during musical esthetic experiences (Brown et al., 2004; Salimpoor et al., 2013), and the extrastriate body area of the occipital cortex and the ventral premotor cortex during esthetic appreciation of dance (Calvo-Merino et al., 2010). The analysis of functional connectivity suggests that this enhancement of perceptual processes owes primarily to the effects of attention (Lacey et al., 2011). Finally, viewing artworks or designs that are liked, found appealing, or beautiful, is also accompanied by an increase in the activity throughout a network of cortical regions related to evaluative judgment, the allocation of attentional resources, and the retrieval of information from memory to contextualize the stimuli and judgment, including the dorsolateral and ventrolateral prefrontal cortex, anterior medial prefrontal cortex, temporal pole, posterior cingulate cortex, and precuneus (Cela-Conde et al., 2004; Cupchik et al., 2009; Jacobsen et al., 2006; Lengger et al., 2007).

Main Current Research Topics

Emotion

What are the biological mechanisms that underlie the affective and emotional aspects of esthetic experience? What does it mean, in biological terms, to be moved by such esthetic qualities as beauty, grace, or sublimity? These questions can be answered only in part when there still is much debate surrounding such fundamental issues as whether there are specific esthetic emotions, whether esthetic experiences can involve common emotions – such as joy, anger, or sadness – or whether emotion is even a necessary ingredient of esthetic experience.

Is there such a thing as esthetic emotions? Some researchers have argued that there are indeed certain classes of emotions that are intrinsic to esthetic experiences. Konecni (2005), for instance, conceived esthetic awe as a specific esthetic emotion, a prototypical response to the sublime. In his view "It is the most pronounced, the ultimate, esthetic response, in all ways similar to the fundamental emotions" (Konecni, 2005: p. 31). Keltner and Haidt (2003), in contrast, suggested that awe might accompany a variety of experiences, including the appreciation of art, though only under certain circumstances. It is art's reliance on depictions of vastness, exceptional actions or moments, or just its sheer size – Keltner and Haidt (2003) argue – that makes it possible to produce profoundly moving emotions of awe that go beyond mere esthetic pleasure.

Lazarus (1991) has espoused the opposite view, namely, that there are no exclusive or prototypical esthetic emotions. In his view, an 'esthetic emotion' is a common emotion experienced in response to art or any other object of esthetic contemplation. Thus, esthetic emotions are elicited by portrayals of emotional content to which people can relate. The strength of such esthetic emotion depends fundamentally on the extent to which spectators identify with the representation and the personal significance of the depicted content. Emphasizing the role of appraisals, Silvia (2006) has also argued that a broad range of common emotions can be a part of an esthetic experience. Appraisals constitute the key mechanism underlying the elicitation of all kinds of emotions in response to objects of esthetic contemplation, and specifically to artworks. Interest, confusion, and surprise, that is to say, the knowledge emotions, are elicited by appraisals in terms of novelty, complexity, familiarity, and coping potential. Appraisals of goal-incongruence, intentionality, harmfulness, or contamination can lead to such hostile emotions as anger, disgust, or contempt. People can also feel self-conscious emotions, such as pride, shame, and embarrassment, while engaging with art. They are related to appraising the congruence of artworks with one's own values, self-image, or goals, the degree personal responsibility, and the consistency with one's standards (Silvia, 2009).

Scherer (2004), however, argued that esthetic emotions – understood as common emotions experienced in response to art – differ from utilitarian emotions, that is to say, common emotions experienced in response to physical and social events occurring outside any artistic context. Esthetic emotions lack the appraisals of goal relevance and coping potential, common to utilitarian emotions: "an aesthetic experience is one that is not triggered by concerns with the relevance of a perception to my bodily needs, my social values, or my current goals or plans, nor with how well I can cope with the situation, but one where the appreciation of the intrinsic qualities of a piece of visual art or a piece of music is of paramount importance" (Scherer, 2004: p. 244). Although esthetic emotions can come with strong phenomenological feelings, they exhibit weaker physiological manifestations than utilitarian emotions, and they lack the arousal and action-oriented responses observed in the latter: when physiological responses – in the form of chills, shivers, or moist eyes – accompany esthetic emotions, they are not aimed at preparing the organism for any specific adaptive action.

Finally, others have maintained that the emotional facet of esthetic experiences is dual. It is not only the content and its personal significance, that drives the emotional response in an esthetic episode (Frijda, 1986; Tan, 2000). The response to Scherer's (2004) intrinsic qualities constitutes a separate and discrete emotional aspect of the esthetic experience. Thus, the represented content arouses one kind of emotions, called 'complementing emotions' (Frijda, 1986) or 'represented emotions' (Tan, 2000). These could be any of the common emotions. In addition, the style, the medium, or the process itself of achieving understanding of the artwork or the object, can elicit a different sort of emotions, called 'responding emotions' (Frijda, 1986) or 'artifact emotions' (Tan, 2000). The enormous expressive potential of art and esthetics allows multiple combinations of these classes of emotions. Artists throughout the centuries have, in fact, exploited the effects of

the incongruence of emotions elicited by content and style (Eco, 2007). Some Hellenistic artists, for instance, abandoned idealist representations in favor of the accurate depiction of profane and everyday topics, which sometimes involved portraying the visible manifestations of old age, violence, or suffering (Stokstad and Cothren, 2011) (**Figure 2**).

Neuroesthetics has contributed to this discussion, mainly through the use of neuroimaging and psychophysiological techniques, in two different ways. First, it has provided psychologists with a new level of analysis. Second, it has allowed asking fascinating new questions. For instance, Skov (2010) sought to characterize the neural mechanisms that enable us to deal with the aforementioned conflicting emotional responses elicited by images' content and formal qualities. He showed participants a series of 300 pleasant, neutral, and unpleasant photographs from the International Affective Picture System, while in the functional magnetic resonance imaging (fMRI) scanner, and asked them to rate them as beautiful, neutral, or ugly. His comparison of brain activity while viewing images rated as beautiful and ugly, regardless of their affective valence, confirmed previous studies that had revealed the involvement of perceptual processing (i.e., superior occipital gyrus, fusiform gyrus, among other regions), reward processing (i.e., orbitofrontal cortex, cingulate cortex), and executive processing (i.e., several prefrontal regions). By examining the interaction between the affective valence of the images' content and participants' ratings, he identified brain regions whose activity was different when participants rated unpleasant images as beautiful and when they rated pleasant images as beautiful. This comparison revealed activity in additional regions of the brain, especially in high-level perceptual processing areas, such as the inferior and superior temporal sulci, and certain regions related with reward, such as the caudate nucleus. These results suggest that specific perceptual and affective mechanisms come into play when we appreciate the esthetic qualities of stimuli depicting emotionally negative content.

Figure 2 Statue of an old market woman (first century AD). She constitutes the opposite of the composed and ideally represented beauty of the Greek Classical period. Her dropped lower jaw, unsteady gait, foggy gaze, and indifference toward her exposed breast, suggest she might be a debauched Dyonisian follower on her way to an offering. Technically, however, the statue constitutes a masterful rendering of complex textures and forms that create the striking contrasts of light and shadow that characterize much of Hellenistic sculpture (Stokstad and Cothren, 2011). Original height: 49 5/8 in. (125.98 cm). The Metropolitan Museum of Art, Rogers Fund, 1909 (09.39). Image © The Metropolitan Museum of Art. Reprinted with permission.

Salimpoor et al.'s (2011, 2013) studies on the neural underpinnings of positive emotional responses to music constitute excellent examples of how neuroesthetics can open new avenues of inquiry. Salimpoor et al. (2011) examined the release of dopamine in different brain regions while people listened to musical pieces that they either did or did not deeply enjoy with a special interest in participants' arousal peaks. Their study revealed an unforeseen functional dissociation. Whereas the caudate was more active during the anticipation of peak emotional experiences, the actual experiences were associated with dopaminergic activity in the nucleus accumbens, showing how the emotional experience of music is mediated by two distinct anatomical pathways that play two different, but complementary, roles in anticipating and generating pleasurable feelings. But how is this positive emotional response tied into the overall esthetic experience? Salimpoor et al. (2013) subsequently used a bidding paradigm, in which participants were asked to listen to unfamiliar fragments of music and allocate amounts of money to listen to them again if they wished. The degree of activity in the nucleus accumbens and an increase in functional connectivity between this region and the auditory cortex, the amygdala, and the ventromedial prefrontal cortex, predicted the amount of money participants were willing to pay to listen to their preferred fragments again (**Figure 3**). These results attest to the importance of the interactions among brain regions involved in reward, sensory processing, and valuation in the generation of the esthetic experience.

Context

With *Fountain* (1917), Marcel Duchamp revolutionized the debate about what constitutes an artwork, and attempted to demonstrate that esthetic qualities could not be part of necessary criteria (**Figure 4**). However, even when his readymade aimed to divorce esthetics from art, some critics were unable to detach themselves from their esthetic orientation toward artworks, believing that "Duchamp was drawing attention to the white gleaming beauty of the urinal" (Danto, 1997), and even stating that "A lovely form has been revealed, freed from its functional purpose, there a man has clearly made an esthetic contribution" (Danto, 1997). A fundamental question arises, thus: what are the contextual conditions that foster such esthetic admiration for everyday objects whose esthetic qualities go otherwise unnoticed?

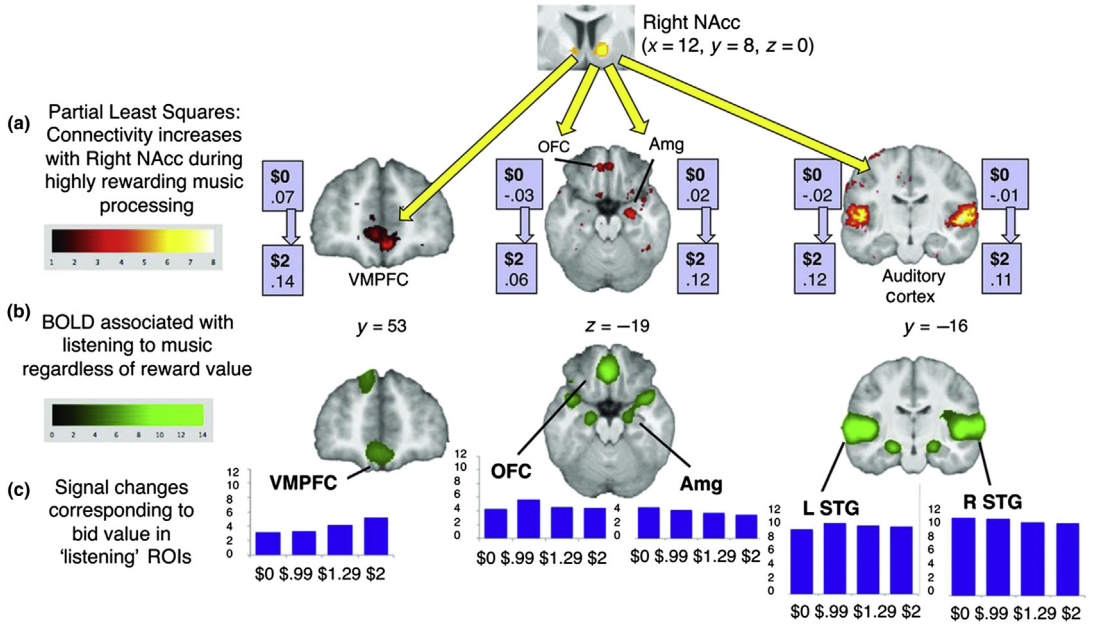

Figure 3 Changes in NAcc functional connectivity associated with increasing desirability of music. (a) Partial least-squares analysis revealed robust increases in connectivity between the nucleus accumbens (NAcc) and other subcortical and cortical regions when individuals hear music they consider highly desirable, compared with music they do not want to hear again. The boxes show changes in correlation as a function of amount bid between the NAcc and each region. A subset of these regions (shown in panel a) overlap with areas that are recruited during music listening compared with rest (panel b). These areas show equally high activity during all music valuation conditions compared with rest (panel c), but their interactions with the NAcc increase as items become more desirable. Amg, amygdala; ROI, region of interest. Copyright 2013 by the American Association for the Advancement of Science. Reprinted with permission.

Figure 4 Marcel Duchamp '*Fountain*, 1917'/© Succession Marcel Duchamp/VBK, Vienna 2013. Photo: Alfred Stieglitz '*Fountain, photograph of sculpture by Marcel Duchamp* (1917)'/© VBK, Vienna 2013. Reprinted with permission.

The conditions under which certain contextual features modulate neural activity underlying esthetic preference are, in fact, a major topic for neuroesthetics. Several studies have shown that personal attitude toward the object presented and beliefs about it have strong effects. Cupchik et al. (2009) demonstrated that people's orientation toward paintings modulated brain activity associated with esthetic preference. They asked participants to focus on the artworks with an esthetic orientation, attending to their esthetic experiences and the works' esthetic qualities, or to view them merely to obtain information from them and take note mostly of the depicted content. The esthetic orientation led to a greater activity in prefrontal brain regions, suggesting a greater degree of cognitive control. The pragmatic orientation, in contrast, was related to an increase in the activity of occipital regions, suggesting an enhancement of perceptual processes. These results show how certain parts of the network of brain regions involved in computing esthetic preference are modulated by the attitude with which we approach an object.

Kirk et al. (2009a) showed how different semantic contexts, which prompted different expectations on behalf of the participants, had a modulating effect on the activity of certain brain regions involved in esthetic appreciation. Participants viewed a series of abstract images while in the fMRI scanner. They were led to believe that some of these images, which were accompanied by the label 'gallery,' were reproductions of artworks exhibited in a renowned gallery. They were told that the other images, presented with the label 'computer,'

had been created by the experimenters using an image editing software. In reality, all stimuli had been created by the experimenters for the experiment. Despite this, participants preferred the images they believed were taken from a gallery more than those they believed were computer generated. Moreover, the authors found that the activity of the medial orbitofrontal cortex, a brain region known to play a key role in the processing of reward value, increased significantly when participants were viewing the images under the gallery-label condition. A similar study was performed by Huang et al. (2011). They asked their participants to view a set of Rembrandt portraits, and led them to believe that some were authentic and others were copies. Their results showed that believing that the images were copies increased the activity in the frontopolar cortex, which in turn modulated activity in the visual cortex, which is consistent with participants' reports that in this condition they had attempted to identify the cues indicating that those portraits were copies. When the portraits were accompanied by the label 'authentic,' activity in the orbitofrontal cortex increased, similarly to Kirk et al. (2009b) finding. This sort of framing-effects appears to influence neural activity as soon as 200 ms after the presentation of the stimuli (Noguchi and Murota, 2013). These studies clearly show how beliefs and expectations generated using semantic framing with simple words, such as 'authentic,' 'fake,' or 'copy,' can influence neural processes related with the affective processes involved in esthetic preference. These contextual effects, however, can be overridden by participants' expertise (Kirk et al., 2011).

Expertise

Understanding how humans acquire, store, and use knowledge is one of the central aims of cognitive science. Because experts excel at these tasks within their own domains, they have attracted the attention of cognitive scientists. Within their particular domain, experts arrive at the best solution faster and more accurately, they are able to select adequate strategies more successfully, they detect features that novices cannot, they represent problems at a deeper level, they have more accurate self-monitoring strategies, and they execute skills with greater automaticity and less effort (Chi, 2006). There are three fundamental reasons why researchers within the field of neuroesthetics have turned to the study of art or esthetic expertise. In the first place, it provides an optimal strategy to understand how art and esthetic knowledge is acquired, organized, and used. Second, it allows understanding how prior knowledge and experience interact with cognitive and affective processes to produce the esthetic experience or to modulate esthetic preference (Leder et al., 2004). Finally, given that expertise and long-term training are related to modifications in brain processes underlying domain-relevant tasks, expertise can be regarded as a 'natural model' of neural plasticity (Münte et al., 2002).

Studies on the neural foundations of expertise in the domains of art and esthetics have yielded four main findings. First, when making judgments about their objects of expertise, experts tend to show greater activity than nonexperts in brain regions related to conceptual association, memory retrieval, and the activation of referential contexts (Kirk et al., 2009a; Wiesmann and Ishai, 2010), illustrating how experts rely to a greater extent on stored knowledge than nonexperts, who rely more on subjective impressions and momentary feelings.

Second, experts in the domains of art and architecture seem to process their objects of expertise with greater 'neural efficiency,' as suggested by a reduced neural activity in certain regions when compared with nonexperts (Berkowitz and Ansari, 2010; Pang et al., 2013). This finding mirrors the decreases in the extent and intensity of brain activation observed when experts in many other domains perform tasks they are specialized in, suggesting that training and practice reduces the extent or intensity of brain activation by increasing neural efficiency, defined as a sharpening of neural responses in task-relevant processing networks (Kelly and Garavan, 2005).

Third, studies have shown that experts are able to filter out the less relevant information for the task at hand. Expert dancers, for instance, rely less than nonexperts on visual information, and emphasize proprioceptive information (Jola et al., 2011). When evaluating their own objects of expertise, musicians, architects, and art experts integrate reward-related processes in a different way to nonexperts, and are able to ignore irrelevant contextual information biasing such processes (Kirk et al., 2009a, 2011; Müller et al., 2010). For instance, Müller et al. (2010) found that, when asked to rate musical stimuli esthetically, laypeople relied on their affective reactions to the stimuli more than experts, who were able to set aside such reactions and focus on the esthetic qualities of the stimuli to a greater degree. Kirk et al. (2011) demonstrated similar emotional regulation in art experts. Whereas monetary sponsorship enhanced laypeople's appreciation of artworks, and increased activity in the ventromedial prefrontal cortex, art experts were not affected by sponsorship. Results showed that, in the case of experts, the dorsolateral prefrontal cortex downregulated activity in the ventromedial prefrontal cortex, suppressing the effects of the monetary favor on the appreciation of art.

Finally, although expertise is by definition domain-specific, experts outperform nonexperts in tasks that are not directly related to their domain of expertise, and several studies have begun exploring the neural underpinnings of these effects. Musicians, for instance, are able to detect small local deviations in foreign language pitch better and earlier, and such ability is related with underlying neural processes related to the categorization of pitch contours (Marques et al., 2007). Electrophysiological data have shown that musicians are also better than nonmusicians at learning linguistic structures, presumably because of their training in segmentation and memory of auditory streams (Francois and Schön, 2011). The application of abilities learned in one domain to a related one has also been observed in the area of visual arts (Pang et al., 2013). This sort of studies suggests that art and music experts transfer their strategies and abilities to nonartistic and nonmusical stimuli.

Future Avenues for Research in Neuroesthetics

Probably, the greatest challenge neuroesthetics facing today is to expand beyond its reliance on neuroimaging and its focus on brain regions involved in esthetic experience. The tools and

methods afforded by cognitive neurogenetics, for instance, could foster a deeper understanding of the biology of esthetic experience below the level of networks of regions, clarifying the role of hormones, neurotransmitters, or genes. Alterations in neurotransmitter function in disorders such as Parkinson's, depression, or anxiety, can have clear effects on value-based decision making in other domains (Ramsøy and Skov, 2010). Increased dopamine levels can interfere with the integration of emotional and cognitive processes, a crucial mechanism underlying decision making in normal conditions. High concentrations of serotonin, on the other hand, can shift choice strategy, from searching for gains to avoiding loss, risk, or ambiguity.

As Green et al. (2008) argued, however, studying genetic variation as a determinant factor of brain function is not only relevant to understanding the origin and development of pathologies. It also holds great potential to clarify healthy cognitive function. In fact, the aforementioned effects linked to certain disorders only reflect the extremes of a continuum of individual variability in naturally occurring neurotransmitter levels. Such variation is, in turn, the result of the interaction between genetic differences among individuals and particular life experiences (Green et al., 2008; Ramsøy and Skov, 2010). Individual differences in specific genes lead to quantifiable differences in the synthesis of proteins. Because genetic variation can occur in genes related to neurotransmitter function, they can produce differences in neurotransmitter synthesis, transport, postsynaptic uptake, presynaptic reuptake, or breakdown. In turn, this sort of variation in molecular function can affect the degree or location of neural activity underlying certain cognitive operations, which might even translate into differences in overt behavior or performance (Green et al., 2008; Ramsøy and Skov, 2010).

Although these facts have been shown for domains unrelated to art or esthetics, they constitute firm grounds for testing hypotheses regarding esthetic appreciation. It is conceivable that genetic imaging could be used in future studies to point to certain genomic variations that are linked to emotional or cognitive processes involved in esthetic appreciation, and even characterize neural connectivity patterns associated with such processes in participants that differ as a function of the underlying genetic variation. Is this a viable direction for research within neuroaesthetics? Will this sort of strategies lead to an integral understanding of the biology of esthetic experience? "Well, we must wait for the future to show" (Woolf, 1930).

See also: Cognitive Neuroscience; Emotion, Neural Basis of; Human Cognition, Evolution of; Neuroeconomics; Neuromarketing; Prefrontal Cortex; Visual Perception, Neural Basis of.

Bibliography

Alajouanine, T., 1948. Aphasia and artistic realization. Brain 71, 229–241.

Allen, G., 1877. Physiological Aesthetics. Henry S. King & Co, London.

Berkowitz, A.L., Ansari, D., 2010. Expertise-related deactivation of the right temporoparietal junction during musical improvisation. Neuroimage 49, 712–719.

Brown, S., Martinez, M.J., Parsons, L.M., 2004. Passive music listening spontaneously engages limbic and paralimbic systems. Neuroreport 15, 2033–2037.

Burke, E., 1757. A Philosophical Enquiry into the Origin of Our Ideas of the Sublime and the Beautiful. Dodsley, London.

Calvo-Merino, B., Urgesi, C., Orgs, G., Aglioti, S.M., Haggard, P., 2010. Extrastriate body area underlies aesthetic evaluation of body stimuli. Experimental Brain Research 204, 447–456.

Cela-Conde, C.J., Marty, G., Maestú, F., Ortiz, T., Munar, E., Fernández, A., Roca, M., Rosselló, J., Quesney, F., 2004. Activation of the prefrontal cortex in the human visual aesthetic perception. Proceedings of the National Academy of Sciences of the United States of America 101, 6321–6325.

Cela-Conde, C.J., Ayala, F.J., Munar, E., Maestú, F., Nadal, M., Capó, M.A., del Río, D., López-Ibor, J.J., Ortiz, T., Marty, G., 2009. Sex-related similarities and differences in the neural correlates of beauty. Proceedings of the National Academy of Sciences of the United States of America 106, 3847–3852.

Chatterjee, A., 2004. The neuropsychology of visual artistic production. Neuropsychologia 42, 1568–1583.

Chatterjee, A., 2011. Where there be dragons: finding the edges of neuroaesthetics. American Society for Aesthetics Newsletter 31, 4–6.

Chi, M.T.H., 2006. Two approaches to the study of experts' characteristics. In: Ericsson, K.A.N., Charness, P., Feltovich, J., Hoffman, R.R. (Eds.), The Cambridge Handbook of Expertise and Expert Performance. Cambridge University Press, New York, pp. 21–30.

Churchland, P.S., Sejnowski, T.J., 1988. Perspectives on cognitive neuroscience. Science 242, 741–745.

Cupchik, G.C., 1992. From perception to production: a multilevel analysis of the aesthetic process. In: Cupchik, G.C., László, J. (Eds.), Emerging Visions of the Aesthetic Process: Psychology, Semiology and Philosophy. Cambridge University Press, New York, pp. 83–99.

Cupchik, G.C., Vartanian, O., Crawley, A., Mikulis, D.J., 2009. Viewing artworks: contributions of cognitive control and perceptual facilitation to aesthetic experience. Brain and Cognition 70, 84–91.

Danto, A.C., 1997. After the End of Art. Contemporary Art and the Pale of History. Princeton University Press, Princeton, NJ.

Darwin, C., 1871/1998. The Descent of Man, and Selection in Relation to Sex. Prometheus Books, Amherst, NY.

Eco, U., 2007. On Ugliness. Harvill Secker, London.

Fechner, G.T., 1876. Vorschule der Ästhetik. Breitkopf und Härtel, Leipzig.

Francois, C., Schön, D., 2011. Musical expertise boosts implicit learning of both musical and linguistic structures. Cerebral Cortex 21, 2357–2365.

Frijda, N.H., 1986. The Emotions. Cambridge University Press, Cambridge.

Gazzaniga, M.S. (Ed.), 1984. Handbook of Cognitive Neuroscience. Plenum Press, New York.

Green, A.E., Munafò, M.R., DeYoung, C.G., Fossella, J.A., Fan, J., Gray, J.R., 2008. Using genetic data in cognitive neuroscience: from growing pains to genuine insights. Nature Reviews Neuroscience 9, 710–720.

Halpern, A.R., O'Connor, M.G., 2013. Stability of art preference in frontotemporal Dementia. Psychology of Aesthetics, Creativity and the Arts 7, 95–99.

Harvey, A.H., Kirk, U., Denfield, G.H., Montague, P.R., 2010. Monetary favors and their influence on neural responses and revealed preference. Journal of Neuroscience 30, 9597–9602.

Huang, M., Bridge, H., Kemp, M.J., Parker, A.J., 2011. Human cortical activity evoked by the assignment of authenticity when viewing works of art. Frontiers in Human Neuroscience 5, 134.

Ishizu, T., Zeki, S., 2013. The brain's specialized systems for aesthetic and perceptual judgment. European Journal of Neuroscience 37, 1413–1420.

Jacobsen, T., Schubotz, R.I., Höfel, L., von Cramon, D.Y., 2006. Brain correlates of aesthetic judgment of beauty. Neuroimage 29, 276–285.

James, W., 1890. The Principles of Psychology. Henry Holt, New York.

Jola, C., Davis, A., Haggard, P., 2011. Proprioceptive integration and body representation: insights into dancers' expertise. Experimental Brain Research 213, 257–265.

Kelly, A.M.C., Garavan, H., 2005. Human functional neuroimaging of brain changes associated with practice. Cerebral Cortex 15, 1089–1102.

Keltner, D., Haidt, J., 2003. Approaching awe, a moral, spiritual, and aesthetic emotion. Cognition and Emotion 17, 297–314.

Kirk, U., Harvey, A., Montague, P.R., 2011. Domain expertise insulates against judgment bias by monetary favors through a modulation of ventromedial prefrontal cortex. Proceedings of the National Academy of Sciences of the United States of America 108, 10332–10336.

Kirk, U., Skov, M., Christensen, M.S., Nygaard, N., 2009a. Brain correlates of aesthetic expertise: a parametric fMRI study. Brain and Cognition 69, 306–315.

Kirk, U., Skov, M., Hulme, O., Christensen, M.S., Zeki, S., 2009b. Modulation of aesthetic value by semantic context: an fMRI study. Neuroimage 44, 1125–1132.

Konecni, V.J., 2005. The aesthetic trinity: awe, being moved, thrilled. Bulletin of Psychology and the Arts 5, 27–44.
Lacey, S., Hagtvedt, H., Patrick, V.M., Anderson, A., Silla, R., Deshpande, G., Hu, X., Sato, J.R., Reddy, S., Sathian, K., 2011. Art for reward's sake: visual art recruits the ventral striatum. Neuroimage 55, 420–433.
Lazarus, R.S., 1991. Emotion and Adaptation. Oxford University Press, New York.
Leder, H., Belke, B., Oeberst, A., Augustin, D., 2004. A model of aesthetic appreciation and aesthetic judgments. British Journal of Psychology 95, 489–508.
Lengger, P.G., Fischmeister, F.P.S., Leder, H., Bauer, H., 2007. Functional neuroanatomy of the perception of modern art: a DC-EEG study on the influence of stylistic information on aesthetic experience. Brain Research 1158, 93–102.
Marques, C., Moreno, S., Castro, S.L., Besson, M., 2007. Musicians detect pitch violation in a foreign language better than nonmusicians: behavioral and electrophysiological evidence. Journal of Cognitive Neuroscience 19, 1453–1463.
Marshall, H.R., 1893. Hedonic aesthetics. Mind 2, 15–41.
Miller, B.L., Hou, C.E., 2004. Portraits of artists. Emergence of visual creativity in Dementia. Archives of Neurology 61, 842–844.
Müller, M., Höfel, L., Brattico, E., Jacobsen, T., 2010. Aesthetic judgments of music in experts and laypersons - an ERP study. International Journal of Psychophysiology 76, 40–51.
Münte, T.F., Altenmuller, E., Jäncke, L., 2002. The musician's brain as a model of neuroplasticity. Nature Reviews Neuroscience 3, 473–478.
Nadal, M., Pearce, M.T., 2011. The Copenhagen neuroaesthetics conference: prospects and pitfalls for an emerging field. Brain and Cognition 76, 172–183.
Noguchi, Y., Murota, M., 2013. Temporal dynamics of neural activity in an integration of visual and contextual information in an esthetic preference task. Neuropsychologia 51, 1077–1084.
Pang, C.Y., Nadal, M., Müller, J.S., Rosenberg, R., Klein, C., 2013. Electrophysiological correlates of looking at paintings and its association with art expertise. Biological Psychology 93, 246–254.
Ramachandran, V.S., Hirstein, W., 1999. The science of art: a neurological theory of aesthetic experience. Journal of Consciousness Studies 6, 15–51.
Ramsøy, T.Z., Skov, M., 2010. How genes make up your mind: individual biological differences and value-based decisions. Journal of Economic Psychology 31, 818–831.
Salimpoor, V.N., Benovoy, M., Larcher, K., Dagher, A., Zatorre, R.J., 2011. Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. Nature Neuroscience 14, 257–262.
Salimpoor, V.N., van den Bosch, I., Kovacevic, N., McIntosh, A.R., Dagher, A., Zatorre, R.J., 2013. Interactions between the nucleus accumbens and auditory cortices predict music reward value. Science 340, 216–219.
Scherer, K.R., 2004. Which emotions can be induced by music? What are the underlying mechanisms? and how can we measure them? Journal of New Music Research 33, 239–251.
Seeley, W.P., 2011. What is the cognitive neuroscience of art…and why should we care? American Society for Aesthetics Newsletter 31, 1–4.
Silvia, P.J., 2006. Artistic training and interest in visual art: applying the appraisal model of aesthetic emotions. Empirical Studies of the Arts 24, 139–161.
Silvia, P.J., 2009. Looking past pleasure: anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions. Psychology of Aesthetics, Creativity and the Arts 3, 48–51.
Skov, M., 2010. The pleasure of art. In: Kringelbach, M.L., Berridge, K.C. (Eds.), Pleasures of the Brain. Oxford University Press, Cambridge, MA, pp. 270–283.
Skov, M., Vartanian, O., 2009. Introduction: what is neuroaesthetics. In: Skov, M., Vartanian, O. (Eds.), Neuroaesthetics. Baywood, Amityville, NY, pp. 1–7.
Stokstad, M., Cothren, M.W., 2011. Art History, fourth ed. Prentice Hall, Upper Saddle River, NJ.
Tan, E.S., 2000. Emotion, art, and the humanities. In: Lewis, M., Haviland-Jones, J.M. (Eds.), Handbook of Emotions, second ed. Guilford Press, New York, pp. 116–134.
Vartanian, O., Goel, V., 2004. Neuroanatomical correlates of aesthetic preference for paintings. Neuroreport 15, 893–897.
Westphal-Fitch, G., Oh, J., Fitch, W.T., 2013. Studying aesthetics with the method of production: effects of context and local symmetry. Psychology of Aesthetics, Creativity and the Arts 7, 13–26.
Wiesmann, M., Ishai, A., 2010. Training facilitates object recognition in cubist paintings. Frontiers in Human Neuroscience 4, 11.
Woolf, V., 1930. To the Lighthouse. Hogarth Press, London.
Zaidel, D.W., 2005. Neuropsychology of Art: Neurological, Cognitive, and Evolutionary Perspectives. Psychology Press, Hove, England.
Zeki, S., 1999. Inner Vision. An Exploration of Art and the Brain. Oxford University Press, Oxford.

Relevant Websites

http://www.association-of-neuroesthetics.org – Association of Neuroesthetics.
http://neuroesthetics.org – Institute of Neuroesthetics.
http://www.brams.org/en/ – International Laboratory for Brain, Music and Sound Research.
http://neuroaesthetics.net – International Network for Neuroaesthetics.

Neuroethics: Ethics of Science and Science of Ethics

Pietro Pietrini and Giuseppina Rota, University of Pisa Medical School, Azienda Ospedaliero-Universiaria Pisana, Pisa, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Scientific advancements over the last decades have led to powerful tools to investigate the neural underpinnings of mental function as well as to novel strategies to treat brain disorders. These unprecedented acquisitions pose questions that fall far beyond the clinical field itself as they affect also the social, ethical, and law domains. The term *neuroethics* refers to (1) the ethical issues that arise in neuroscientific research and its clinical applications and (2) the study of the neurobiological underpinnings of ethical concepts and moral judgment. In this article, we discuss the main issues comprised within this dual definition of neuroethics.

Introduction

The word *neuroethics* was used for the first time in 1973 by Anneliese A. Pontius, a Harvard neuropsychiatrist, in an article published in the journal *Perceptual and Motor Skills* to describe the negative consequences associated with early walking as an attempt to accelerate motor development in infants (Pontius, 1973). Additional meanings have been associated to this word in the following years. In 1989, for example, the neurologist Ronald Cranford used it to refer to ethical issues in clinical neurology (Cranford, 1989). In 1991, Patricia Churchland defined *neuroethics* as the biological approach to the study of human morality (Churchland, 1991), which can be achieved through modern techniques for the visualization of *in vivo* brain activity.

The birth of neuroethics as a new discipline may be traced back to May 2002. In this year, for the first time, the *New York Times* journalist William Safire used the word *neuroethic*s with its current meaning. In 2002, an important meeting held in San Francisco under the title *Neuroethics: Mapping the Field* gathered scientists and philosophers to discuss ethical issues related to the advancements in the applications of neuroscience (*see* Cognitive Neuroscience; Computational Neuroscience). Since then, the term *neuroethics* has been used to define both ethical issues in neuroscientific research and its clinical applications as well as the investigation of the neural underpinnings and correlates of ethical concepts such as moral decision making and free-will (Roskies, 2002).

In general, this discipline aims to clarify the way neuroscience findings and advances in neuroscientific methods and technology affect society and culture.

The *ethics of neuroscience* focuses primarily on (1) the way research is planned and the procedures that are followed in neuroscience, and (2) the effects and the influence of neuroscientific findings on medical treatments, social policy, the legal system, and so on. The first group of issues includes principles that research in neuroscience should observe, such as following approved protocols, obtaining an informed consent, and respecting autonomy and privacy rights of patients and healthy volunteers. Technological advances make new forms of interventions into the brain available. This, in turn, poses many ethical problems as to whether these practices may alter the personal identity or carry adverse side effects for health and whether they may be considered as morally acceptable in light of their implications for society.

The *neuroscience of ethics* is concerned with the use of neuroscientific methods to investigate topics that have been traditionally the center of philosophical investigations, such as free-will or moral judgment. This area investigates the neurobiological correlates of moral cognition, and specifically the way moral reasoning and decision making are represented in the human brain.

Most topics of investigation of neuroethics have important reverberations and implications for society. For example, is it ethically acceptable to take advantage of drugs to enhance mental and/or physical abilities? Which consequences may have learning that a specific genetic predisposition exposes us to a significantly greater risk of developing a given illness or to manifest aggressive behavior in response to a provocative situation? A deeper understanding of the etiopathology of mental disorders may lead to redefine the notion of free-will, with consequent implications for the concept of personal responsibility in court. How will the ability to predict behavior in a probabilistic way affect society? These and many others are the questions that challenge researchers in neuroethics. In this article, we discuss the main topics that are embraced by neuroethics.

The Ethics of Neuroscience

Over the last decades, novel powerful technologies for brain research have been developed. Modern neuroimaging methodologies, including high-field magnetic resonance imaging (MRI), enable a fine visualization of the morphological and functional architecture of the brain in a noninvasive manner in humans. On a clinical ground, these fine methodologies have made it possible to achieve earlier and more accurate diagnoses, as well as to perform sophisticated interventions, suffice to think of electrode implantation into structures that are deep in the brain or the application of neuroprostheses to restore lost sensory or motor functions. Psychopharmacology research may allow enhancement of individual skills, compensation for congenital or acquired cognitive or affective deficits in impaired individuals, and potentiation of existing skills in healthy population. Although very promising and potentially very useful for the treatment of neurological and psychiatric disorders, these methods raise many ethical issues. Here, we discuss the main ethical aspects related to the distinct domains.

Psychosurgery

In the history of medicine, many interventions on the brain have been devised. In ancient times, for example, holes were drilled into the skull as a means to relieve pain due to hematomas and other neurological accidents. The term *psychosurgery*, that is neurosurgery applied to cure mental disorders, was first introduced to describe a type of brain intervention also known as *prefrontal leucotomy*, which consisted in interrupting the connecting fibers from the prefrontal cortex (*see* Prefrontal Cortex) and other brain areas by sectioning subcortical white matter and by injecting alcohol (Moniz, 1937). In transorbital lobotomy, the disconnection of white matter tracts in the frontal lobes was achieved by inserting a metallic tool called *lobotome* into the skull through the eye orbits and by moving this instrument back and forth. This procedure, called *lobotomy*, was often performed in the 1940s and 1950s to treat refractory psychiatric illnesses. Almost invariably, lobotomy resulted in a significant neurological damage and was accompanied by neurological sequelae and personality alterations. Cingulotomy is another type of neurosurgery that has been used to treat severe obsessive compulsive disorder (OCD). This procedure disconnects the cingulate cortex from prefrontal regions. As it disrupts the brain pathways that convey the emotional and affective experience associated to nociceptive stimulation, cingulotomy has been used to alleviate intractable pain. A significant improvement has been observed in around 30% of patients treated with this procedure (see, for example, data collected in 30 years of medical praxis at the Massachusetts General Hospital, Dougherty et al., 2002).

Psychosurgery is in use nowadays for the treatment of symptoms associated to severe and refractory depression and OCD, which are not responsive to other medical interventions. Modern interventions rely on MRI-guided stereotactic procedures, which greatly increase surgical precision. The ability to specifically target selective brain regions can significantly increase effectiveness of neurological and psychiatric treatments. The risk of side effects always needs to be taken into account and evaluations of the ratio cost/benefits have to be carried out carefully for each individual. Nevertheless, relieving psychiatric symptoms may greatly increase subjective well-being and individual social integration, especially when those symptoms are particularly detrimental for daily life routines, as in the case of massive obsessive compulsions. Psychosurgery, by directly altering the functioning of neuronal substrates of psychiatric and neurological disorders, may be used as an effective treatment. Nevertheless, such invasive procedures may result in significant loss of cognitive functions and major personality changes. Thus, the application of these methods is not uncontroversial from the ethical point of view, as they may completely alter the sense of the self of the patient, leading to permanent changes in personal identity. This raises the question whether the effects of psychosurgery in these cases may not be worse than the illness itself.

Neurostimulation

Contrary to more invasive means of brain interventions, neurostimulation does not result in permanent brain lesions. Neurostimulation is usually performed by using microelectrodes that are implanted deeply in the brain, such as in subcortical structures. This technique is also known as *deep-brain stimulation* (DBS) and works by releasing electrical current directly in targeted brain regions, whose activity is altered in a number of neurological disorders. DBS has been employed successfully in the treatment of Parkinson's disease to allow patients to recover lost movement abilities. DBS has been shown to be particularly useful for patients with intractable tremor and with complications due to long-term treatment with levodopa (Volkmann, 2004). DBS has been used to inhibit hyperactivity within (*see* Epilepsy) epileptic foci in the treatment of drug-resistant epileptic patients. Also, severe depression has been shown to be responsive to this method of intervention (Mayberg et al., 2005). Finally, DBS is a promising technique for the treatment of refractory OCD (Lakhan and Callaway, 2010).

Importantly, DBS allows patients to actively control the level of stimulation that they receive by switching electrodes on and off. Reports of improper use of brain implants in patients with Parkinson's disease, including for erotic self-stimulation, have been described (e.g., Morgan et al., 2006). Abuse of stimulation with brain implants may also lead to compulsive gambling and financial problems in these patients.

The implanted electrodes are removable. Nevertheless, implanting electrodes carries the risk of lesioning brain areas that are adjacent to the targeted ones. Achieving a high precision in brain spatial placement is crucial in order to obtain the desirable therapeutic effects.

Potential neurological sequelae of DBS may vary according to the functions of the brain areas where implants are located and must be taken into careful consideration prior to each intervention. Also, stimulation of the intended brain portion may result in unforeseen side effects that may be due to alterations in its activity as well as to changes in activity of brain sites that are functionally connected to it.

Cognitive Enhancement

In human history, various types of substances including caffeine, nicotine, and cocaine have been used to enhance cognitive performances and reduce fatigue. In recent years, there has been a growing interest in the potentiation of attention and other cognitive functions. Cognitive enhancement can be achieved through the administration of different pharmaceutical products now available in the market. Both methylphenidate (Ritalin) and modafinil (Provigil) are examples of cognitive enhancers that are widely used to treat disorders of attention and vigilance. Modafinil was first introduced for treating narcolepsy and it can be employed for curing the shift-work sleep disorder and problems related to apnea during sleep. Modafinil is effective in the regulation of the sleep–wake cycle and circadian rhythms and greatly alleviates daytime sleepiness and reduces the rate of accidents in time-shift working people. As a performance enhancer, this drug has been used to increase alertness in healthy volunteers with no dysregulation of circadian rhythms who need to maintain a sharp cognitive functioning for prolonged periods of time. Cognitive functions that show an improvement as an effect of modafinil intake include pattern recognition and a reduction in reaction times (Scholey et al., 2012). The use of

drugs such as modafinil has many potential applications outside of those strictly clinical. For instance, modafinil may allow military soldiers to manage highly stressful and demanding tasks without losing concentration or facilitate the work of medical services that are active in high-risk emergency situations. Modafinil seems to selectively target the functioning of the hypothalamus (*see* Hypothalamus) in the regulation of the wake–sleep cycle. No addictive effects have been associated to modafinil so far. Nevertheless, (*see* Sleep and Memory) sleep plays an important role in crucial cognitive functions, such as consolidation of memory (*see* Learning and Memory, Neural Basis of) traces and synaptic (*see* Synaptic Mechanisms and Protein Synthesis in Learning, Memory, and forgetting) homeostasis. Sleep deprivation may lead to a higher incidence of hypertension and thus affect health in ways that may be unforeseen on a short time-period window.

Herbal extracts, such as from Guaraná, Sage, Ginseng, and Bacopa, are also widely employed for their effects as cognitive enhancers. In healthy young adults, for example, intake of Panax Ginseng has been shown to improve memory performances and reduce fatigue associated with prolonged cognitive effort (Scholey et al., 2012). Energy drinks that contain caffeine are also included in this category and are being popularly used especially by young individuals.

Inhibiting sleep or increasing the capacity to maintain alertness for prolonged periods of time by using pharmacological means may not come without negative side effects for health. Is it ethically acceptable to do so? Is it safe to use cognitive enhancers even though the biochemical mechanisms of their functioning on the central nervous system are still largely unclear? Epidemiological longitudinal studies should be carried out to gain a closer understanding of the short- and long-term effects of cognitive enhancers, both investigating positive and adverse side effects on health. Public institutions should implement informative campaigns that target the medical sector, schools, and families in order to give hints on adverse effects that these artificial aids may carry for health.

Finally, leaving aside medical considerations, it is important to keep in mind that social consequences are likely to be associated with the use of performance enhancers. The recourse to neuroenhancers is stimulated by contemporary cultural phenomena, including the strong competitiveness and the need to maximize immediate productivity and efficiency that are often seen in working environments. In theory, if the use of cognitive enhancers becomes too generalized, then everybody will need to assume them merely to achieve equal performances. Furthermore, cognitive enhancers may be available primarily to those who possess greater economic resources, thus contributing to increased social inequity. Finally, in high-risk jobs, the inability to maintain high performances for longer times without potentiation may expose to greater risks when enhancers cannot be used and the same performances cannot be achieved. Thus, many are the issues that should be taken into consideration as far as the use of cognitive enhancers is concerned. Among them, it is important to mention the value of equal opportunities, merit due to effort and hard work as well as fairness in society.

Brain–Computer Interfaces and Neuroprosthesis

Since the 1960s, researchers have investigated the effects of EEG (*see* Electroencephalography: Basic Principles and Applications)–biofeedback–trainings on behavioral performances. Clinical applications include the treatment of the attention deficit/hyperactivity disorder, the suppression of epileptic seizures, and the restoration of communication abilities in locked-in patients. More recent applications have shown the feasibility of near-infrared spectroscopy, (*see* Near-Infrared Spectroscopy) functional magnetic resonance imaging (fMRI), and magnetoencephalography (MEG)-based brain–computer interfaces (BCIs) to enable individuals to gain volitional control over circumscribed brain areas. Compared to the EEG counterpart, neurofeedback based on these methods provide real-time information on brain activity levels with high spatial resolution. As such, it can be used to study behavioral effects of activity modulation in areas of interest (e.g., Rota et al., 2009).

BCIs may rely on signals that are derived from brain activity as inputs to control electrical external devices. Clinical applications include rehabilitation and restoration of movement in patients with stroke and brain damages (*see* Brain Damage: Neuropsychological Rehabilitation) (see Birbaumer et al., 2008 for a review). Invasive BCIs rely on implanted brain electrodes and have been mostly investigated in animal models. Neuroprostheses (*see* Neuroprosthetics) allow animals to perform movements, as well as to reach and grasp objects by self-eliciting variations in brain activity.

The development of more and more sophisticated neuroprostheses raises several concerns about their potential misuse. Do the benefits of neuroprostheses and artificial body implants justify the risk of adverse health effects due to invasive surgical procedures? Is it desirable to achieve a permanent enhancement of human physical and cognitive capabilities with these methods? Will cyborgs and human/machine hybrids remain obscure characters of science fiction scenarios or is there a true potential for abuse?

Privacy Rights and Informed Consent

In recent years, attention has increased for the respect of personal autonomy and privacy rights (Beauchamp and Childress, 2009) of patients and healthy volunteers who participate in research studies. Each developed country has created a corpus of laws for the protection of privacy rights. In clinical trials and scientific studies, patients and healthy volunteers are required to receive all information concerning the procedures that they will undergo and the risks involved and to read and sign an informed consent. The ability to give an informed consent is grounded on the capacity to make a decision, which in turn relies on a number of cognitive functions such as attention, working memory, understanding, and general reasoning. The ability of the researcher to take the perspective of the consenting individual and to recognize his/her current emotional states is crucial for understanding whether he/she is able and willing to give an informed consent. Capability to give full consent to participation to a study is a particularly sensitive issue when it comes to patients with neurodegenerative disorders, *in primis* dementia, that affect cognitive abilities, or with mood disorders, that may impair

judgment and will. Given the importance of properly assessing decisional capacity for obtaining an informed consent, in recent years a number of clinical tests have been devised.

The Neuroscience of Ethics

In the 1990s, the so called *Decade of the Brain*, an amazing technological advancement in the investigation of the *in vivo* functioning of the human brain took place. Novel neuroimaging methodologies have made it possible to map complex cognitive and emotional processes (*see* Emotion, Neural Basis of) in the brain, thus paving the way to the study of the neural and molecular underpinnings that sustain the complexity of human behavior (Pietrini, 2003). The main neuroimaging technologies include positron emission tomography, single photon emission computed tomography, MEG, MRI, and fMRI. Neuroimaging can be employed to uncover the neuronal correlates of brain function both under healthy and pathological conditions. Also, direct information on the functioning of neural substrates of mental activity may provide psychiatrists with a direct feedback on the therapy's effectiveness allowing them to adjust their intervention accordingly and to prevent relapses, thus greatly improving subjective well-being.

Advancements in the methods of investigation and visualization of brain activity allow researchers to identify the neuronal underpinnings of morality and free-will. Also, molecular biology is shedding novel light on the genetic factors that modulate behavior, including control of impulses and expression of aggression. In the following sections, we briefly review the findings concerning moral judgment and free-will.

Moral Judgment

Moral evaluations (*see* Neural Foundation of Morality) are a crucial component of decision-making processes and are typically engaged when choosing among two or more competing options that may advantage ourselves or a third party. Imaging studies have relied on ethical dilemmas to study the neural underpinning of moral evaluations. Moral dilemmas typically consist of short stories that present a conflict of choices and the question as to whether a given action may be morally acceptable in a given situation. This paradigm relies on script-driven mental imagery, and it is typically applied for addressing complex emotional and cognitive processes that otherwise would be impossible to investigate in a scientific manner (see, for instance, the investigation of the neural correlates of forgiveness by Ricciardi et al., 2013).

Reasons and emotions are core elements for moral judgments. Plato was the first prominent thinker who described the dichotomy of reason and emotion in humans and advocated the role of reason in governing passions and guiding righteous actions. However, reason does not alone determine our moral choices. Evidence from neuroimaging studies shows that a strong interplay of reason and emotion takes place when people judge the acceptability of controversial moral choices (Greene et al., 2001). The trolley and the footbridge dilemmas are illustrative of the moral impasse created in such situations. In the first dilemma, subjects imagine viewing a runaway trolley on which the conductor lost control that threatens to kill five workers along the runaway if no action is taken. The only way for the subjects to save the five workers is to pull a lever that will divert the trolley onto an alternative set of tracks on which there is only one worker. This man will die, but the other five will be spared. In the footbridge dilemma, subjects have to imagine being on a bridge that overlooks the tracks and seeing the runaway trolley directing toward the five men. The only means they have to stop the trolley is by pushing a stranger next to them off the bridge onto the tracks, using his body to interrupt the trolley's run. Both dilemmas confront people with the same utilitarian reasoning: deciding to sacrifice one life to save five others. However, most people who engage in those judgments accept one person's death in the trolley but not in the footbridge dilemma. Greene et al. (2001) showed that the judgment of moral acceptability in the trolley dilemma is sustained by activation in brain areas (i.e., the dorsolateral prefrontal cortex (*see* Cerebral Cortex)) involved in utilitarian reasoning only (i.e., based on a cost/benefits calculation). On the contrary, when evaluating the footbridge dilemma, a conflict between reason and emotion arises and areas of the brain that are involved in emotional processing (i.e., the ventromedial prefrontal cortex) strongly become activated. This result suggests that emotion and rationality are intertwined aspects of moral decision making and that their role may become more or less relevant if the context in which the same moral question is framed changes. A number of factors have been shown to affect emotional involvement and brain responses in moral judgment, including the level of personalness and/or directness of the harmful action (e.g., whether there is spatial proximity between the agent and the victim and whether muscular force and direct contact are used by the agent, as, for instance, by pushing the victim off the bridge, or by activating a mechanical device that mediates the action, Greene et al., 2004, 2001).

Research in neuroethics allows for bridging deficits in emotions and empathy (*see* Empathy: A Social Neuroscience Perspective) with deficits in moral reasoning that may lead to impairments in moral judgment. In the antisocial personality disorder, (*see* Psychopathy) both emotional life and empathy are impaired (Pietrini and Bambini, 2009). Morphological alterations of brain structures crucial for moral judgments and impulse control, such as the medial prefrontal cortex (MPFC) and the dorsolateral prefrontal cortex, have been documented in psychopathic subjects who committed violent crimes (Boccardi et al., 2011). Consistently, Koenigs et al. (2007) have found that patients with damage in the ventral portion of MPFC engaging in moral evaluations highly prefer utilitarian choices, that is, choices that permit the achievement of a greater good even though they may be harmful to others.

Social categorization and stereotype also affect moral evaluations. The acceptability of causing harm to other people in order to achieve a greater good is influenced by their membership to the social group of the dilemma's judger (Cikara et al., 2010). Cikara and colleagues showed that individuals of extreme out-groups (e.g., homeless people) are more easily sacrificed for the survival of in-group members. This study specifically tested the Stereotype Content Model according to which our emotional reactions are affected by the levels of warmth and competence that we perceive in the

other. For example, elderly and children belong to high-warmth and low-competence categories, whereas police officers belong to a high-competence and high-warmth category. Social categories with high warmth are likely to elicit feelings of benevolence and prosocial behavior whereas low-warmth categories are more likely to elicit negative emotions, such as disgust. A prefrontal network linked to the resolution of complex tradeoffs that comprises medial, dorsolateral, and orbitofrontal regions is activated when volunteers decide to sacrifice low-warmth, low-competence targets to save in-group members. Thus, intergroup biases heavily affect neural reactivity during moral choices.

Intentionality in causing harm to others affects acceptability of the action itself. A great body of literature demonstrated a higher acceptability of harm that is the consequence of good intention, that happens accidentally, or that is the consequence of inaction (see Christensen and Gomila, 2012 for a review).

Ecological validity of experimental stimuli that are used for eliciting moral evaluations is an important condition that should be carefully taken into consideration while planning experimental tasks. Obviously, imagining scenarios in which soldiers combat is described as an easier and a more realistic task by individuals who had first-hand or vicarious experiences related to those scenes.

Finally, certainty of events is an important variable that so far has been barely controlled for. While full certainty about the future course of real-life events is rare, events depicted in moral dilemmas are usually presented in a way that they are bound to follow a certain course with no uncertainty. While this condition is due to experimental purposes, it may also represent a limitation to the ecological validity of the results.

Neurobiological Correlates of Free-Will and Potential Forensic Implications

Free-will has been the center of philosophical disputes over centuries and it is broadly understood as the ability to freely choose to engage in a given action over other ones that are available. Great interest for this topic has been stirred by the findings of Libet et al. (1983). The researchers investigated neuronal responses and subjective reports of individuals who freely performed motor acts, (*see* Cortical Areas Engaged in Movement: Neuroimaging Methods) showing that cerebral activity is detectable in the supplementary motor area (*see* Motor Cortex) a few hundred milliseconds before experiencing the subjective intention to move. Soon et al. (2008) used an fMRI motor-decision paradigm to show that activity in the prefrontal and parietal cortex is measurable up to 10 s before subjects become aware of their motor intention. Taken together, these findings suggest that unconscious brain activity influences an individual's volitional act long before it enters awareness. This argument has been used by some to challenge free-will (e.g., Wegner, 2002). For others, these data merely suggest that the conscious experience of an intention needs a longer time to build up (see Guggisberg and Mottaz, 2013 for a review).

Free-will also implies the ability to exercise volitional will-power on one's own choices. As such, free-will may not be properly conceptualized as a binary 'on–off' variable. It may rather consist of a scalar variable that comes in degrees. Thus, individuals on one side of the spectrum are completely deprived of volitional control on their behavior and of criminal responsibility for their acts. On the other side, individuals fully possess the ability to inhibit impulsive behavior, especially in situations where they are exposed to social triggers that may lead others to aggressiveness and violence.

A number of applications of neuroscientific methods in forensic settings, such as for lie and memory detection, have stirred great interest in the scientific community (e.g., Sartori et al., 2011). The use of diagnostic brain imaging methods may have important implications in court as these exams may provide information regarding anatomical alterations that may compromise the functioning of brain areas that are crucial for the control and inhibition of impulsive behavior (Pietrini and Bambini, 2009). Scientific evidence has linked lesions and hypofunctionality of the prefrontal cortex to impulsive, violent, and antisocial behavior (e.g., Pardini et al., 2011). These data have also been supported by consistent findings in healthy controls with no history of violence (Pietrini et al., 2000). Reduced gray matter in prefrontal regions and increased activation in the amygdala, a region that controls emotional reactivity, have also been documented in violent and antisocial behavior (Raine et al., 2000; Davidson et al., 2000).

Behavioral genetics has expanded our understanding of genetic influences on antisocial behavior. For instance, previous research has linked some alleles of a few genes as MAOA, 5-HTTLPR, COMT, and DRD4 to aggressiveness and violent behavior (see Iofrida et al., 2014 for a review).

Specific genetic variants have been associated to variations in functionality and morphology in brain areas critical for the inhibition of aggressive impulses. Are these findings moving us toward a *molecular* ethics (Pietrini, 2003)? What are the consequences of this evidence for the notion of free-will?

A recent sentence in Italy has stirred great international interest as neuroscientific evidence, including brain imaging and genetic evaluations, was used in court to account for a reduced ability of the defendant to control impulsive behavior. In this case, the presence of genetic variants linked to violent behavior and neuroimaging evidence of poor control of impulsive behavior led the judge to shorten the sentence of jail for the offender (Pellegrini and Pietrini, 2010). In the future, evaluations of criminal responsibility may be greatly facilitated by a deeper understanding of the neural signatures on normal and abnormal behavior. Importantly, neuroscientific and behavioral genetic investigations may greatly improve psychiatric assessment by providing objective data on the neuronal bases of a criminal's mental disorder (Rigoni et al., 2010; Sartori et al., 2011). This is very relevant as it may lead to the development of interventions for rehabilitation that will be tailored to each individual's capacities and potentials.

Conclusions

From its first introduction in 1973 the word *neuroethics* has come to indicate on one hand the ethical issues that arise in many fields of neuroscience research and their clinical applications and on the other the neurobiological bases of ethical functions. Innovative structural and functional brain imaging methodologies have made it possible to enter the era of a fascinating journey through the human brain in search of the

mind, so that ancient philosophical questions can now be approached from a neuroscientific perspective. The results of these studies are shedding new light on our understanding of brain functions, including moral judgment and decision-making processes, with implications that fall far beyond the mere field of neuroscience, as they affect ethics, philosophy, and even the law itself. At the same time, the development of more and more sophisticated tools is leading to unprecedented opportunities to treat brain disorders or restore lost functions through the use of neuroprostheses and brain implants. As neuroscience is advancing with no pause, neuroethics will witness greater and greater challenges in the years to come.

See also: Behavioral Genetics: Psychological Perspectives; Bioethics in the Post-genomic Era; Bioethics: Genetics and Genomics; Bioethics: Philosophical Aspects; Brain Damage: Neuropsychological Rehabilitation; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Cognitive Neuropsychology, Methods of; Cognitive Neuroscience; Emotion, Neural Basis of; Empathy: A Social Neuroscience Perspective; Epigenetics of Brain Plasticity and Behavior; Ethical Practices: Institutional Oversight and Enforcement; Genetics and Society; Genetics: Legal Aspects; Informed Consent of Research Subjects; Intelligence, Genetics of: Cognitive Abilities; Internet Research: Ethical Concerns; Media Ethics; Neural Foundation of Morality; Neuroprosthetics; Objectivity of Research: Ethical Aspects; Prefrontal Cortex Development and Development of Cognitive Function; Privacy of Individuals in Social Research: Confidentiality; Research Ethics, Cross-Cultural Dimensions of; Research Ethics: Research.

Bibliography

Beauchamp, T.L., Childress, J.F., 2009. Principles of Biomedical Ethics, sixth ed. Oxford University Press, New York/Oxford.

Birbaumer, N., Murguialday, A.R., Cohen, L., 2008. Brain–computer interface in paralysis. Current Opinion in Neurology 21 (6), 634–638.

Boccardi, M., Frisoni, G.B., Hare, R.D., Cavedo, E., Najt, P., Pievani, M., Rasser, P.E., Laakso, M.P., Aronen, H.J., Repo-Tiihonen, E., Vaurio, O., Thompson, P.M., Tiihonen, J., 2011. Cortex and amygdala morphology in psychopathy. Psychiatry Research 193 (2), 85–92.

Churchland, P.S., 1991. Our brains, our selves: reflections of neuroethical questions. In: Roy, D.J., Wynne, B.E. (Eds.), Bioscience and Society. Wiley, New York.

Cikara, M., Farnsworth, R.M., Harris, L.T., Fiske, S.T., 2010. On the wrong side of the trolley track: neural correlates of relative social valuation. Social Cognitive and Affective Neuroscience 5 (4), 404–413.

Cranford, R.E., 1989. The neurologist as ethics consultant and as a member of the institutional ethics committee. Neurologic Clinics 7, 435–437.

Christensen, J.F., Gomila, A., 2012. Moral dilemmas in cognitive neuroscience of moral decision-making: a principled review. Neuroscience and Biobehavioral Reviews 36 (4), 1249–1264.

Davidson, R.J., Putnam, K.M., Larson, C.L., 2000. Dysfunction in the neural circuitry of emotion regulation – a possible prelude to violence. Science 289, 591–594.

Dougherty, D.D., Baer, L., Cosgrove, G.R., Cassem, E.H., Price, B.H., Nierenberg, A.A., Jenike, M.A., Rauch, S.L., 2002. Prospective long-term follow-up of 44 patients who received cingulotomy for treatment-refractory obsessive-compulsive disorder. American Journal of Psychiatry 159, 269–275.

Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., Cohen, J.D., 2004. The neural bases of cognitive conflict and control in moral judgment. Neuron 44 (2), 389–400.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D., 2001. An fMRI investigation of emotional engagement in moral judgment. Science 293 (5537), 2105–2108.

Guggisberg, A.G., Mottaz, A., 2013. Timing and awareness of movement decisions: does consciousness really come too late? Frontiers in Human Neuroscience 7 (385), 1–11.

Iofrida, C., Palumbo, S., Pellegrini, S., 2014. Molecular genetics and antisocial behavior: where do we stand? Experimental Biology and Medicine.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., Damasio, A., 2007. Damage to the prefrontal cortex increases utilitarian moral judgements. Nature 446, 908–911.

Kohlberg, L., 1964. Development of moral character and moral ideology. In: Hoffman, M.L., Hoffman, L.W. (Eds.), Review of Child Development Research, vol. 1. Russell Sage Foundation, New York.

Lakhan, S.E., Callaway, E., 2010. Deep brain stimulation for obsessive-compulsive disorder and treatment-resistant depression: systematic review. BMC Research Notes 3, 60.

Libet, B., Gleason, C.A., Wright, E.W., Pearl, D.K., 1983. Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. Brain 106 (Pt 3), 623–642.

Mayberg, H.S., Lozano, A.M., Voon, V., McNeely, H.E., Seminowicz, D., Hamani, C., Schwalb, J.M., Kennedy, S.H., 2005. Deep-brain stimulation for treatment-resistant depression. Neuron 45, 651–660.

Moniz, E., 1937. Prefrontal leucotomy in the treatment of mental disorders. American Journal of Psychiatry, 1844–1944. American Psychiatric Publishing; 1994. ISBN 978-0-89042-275-5.

Morgan, J.C., diDonato, C.J., Iyer, S.S., Jenkins, P.D., Smith, J.R., Sethi, K.D., 2006. Self-stimulatory behavior associated with deep brain stimulation in Parkinson's disease. Movement Disorder 21 (2), 283–285.

Northoff, G., 2006. Neuroscience of decision making and informed consent: an investigation in neuroethics. Journal of Medical Ethics 32, 70–73.

Pardini, M., Krueger, F., Hodgkinson, C., Raymont, V., Ferrier, C., Goldman, D., Strenziok, M., Guida, S., Grafman, J., 2011. Prefrontal cortex lesions and MAO-A modulate aggression in penetrating traumatic brain injury. Neurology 76, 1038–1045.

Pellegrini, S., Pietrini, P., 2010. Siamo davvero liberi? Il comportamento tra geni e cervello. Sistemi Intelligenti 2, 281–294.

Pietrini, P., Bambini, V., 2009. Homo ferox: the contribution of functional brain studies to understanding the neural bases of aggressive and criminal behavior. International Journal of Law and Psychiatry 32 (4), 259–265.

Pietrini, P., 2003. Toward a biochemistry of mind? American Journal of Psychiatry 160 (11), 1907–1908.

Pietrini, P., Guazzelli, M., Basso, G., Jaffe, K., Grafman, J., 2000. Neural correlates of imaginal aggressive behavior assessed by positron emission tomography in healthy subjects. American Journal of Psychiatry 157 (11), 1772–1781.

Pontius, A.A., 1973. Neuro-ethics of 'walking' in the newborn. Perceptual Motor Skills 37, 235–245.

Raine, A., Lencz, T., Bihrle, S., LaCasse, L., Colletti, P., 2000. Reduced prefrontal gray matter volume and reduced autonomic activity in antisocial personality disorder. Archives of General Psychiatry 58, 119–127.

Ricciardi, E., Rota, G., Sani, L., Gentili, C., Gaglianese, A., Guazzelli, M., Pietrini, P., 2013. How the brain heals emotional wounds: the functional neuroanatomy of forgiveness. Frontiers in Human Neuroscience 7, 839.

Rigoni, D., Pellegrini, S., Mariotti, V., Cozza, A., Mechelli, A., Ferrara, S.D., Pietrini, P., Sartori, G., 2010. How neuroscience and behavioral genetics improve psychiatric assessment: report on a violent murder case. Frontiers in Behavioral Neuroscience 4, 160.

Roskies, A., 2002. Neuroethics for the new millennium. Neuron 35 (1), 21–23.

Rota, G., Sitaram, R., Veit, R., Erb, M., Weiskopf, N., Dogil, G., Birbaumer, N., 2009. Self-regulation of regional cortical activity using real-time fMRI: the right inferior frontal gyrus and linguistic processing. Human Brain Mapping 30 (5), 1605–1614.

Sartori, G., Pellegrini, S., Mechelli, A., 2011. Forensic neurosciences: from basic research to applications and pitfalls. Current Opinion in Neurology 24 (4), 371–377.

Scholey, A., Stough, C., Verster, J.C., 2012. Cognitive enhancement: are we barking up the wrong tree? Current Drug Abuse Reviews 4, 255.

Soon, C.S., Brass, M., Heinze, H.J., Haynes, J.D., 2008. Unconscious determinants of free decisions in the human brain. Nature Neuroscience 11 (5), 543–545.

Volkmann, J., 2004. Deep brain stimulation for the treatment of Parkinson's disease. Journal of Clinical Neurophysiology 21, 6–17.

Wegner, D., 2002. The Illusion of Conscious Will. MIT Press, Cambridge, MA.

Neurolinguistic Processing of Psychological Verbs

Jonathan Brennan, The University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Psychological verbs are used to communicate how we feel by describing mental states such as 'love,' 'amuse,' and 'bother.' These verbs denote a particular class of concepts and have been found across many languages to have unique syntactic and semantic properties that offer a window in to how meaning is grammatically encoded. Aphasiology, electrophysiology, and neuroimaging studies have leveraged the unique conceptual and grammatical properties of these verbs to shed light on neural mechanisms involved in representing concepts, identifying argument roles, and building complex compositional meanings. A recurring theme across this diverse field is the challenge of linking the representational properties of these verbs, which remain debated in linguistics, with specific neurobiological mechanisms.

Introduction

Language is commonly used to communicate how we feel. Verbs such as 'love,' 'appreciate,' 'amuse,' and 'bother' convey the mental, or psychological, state of the participants in a sentence (see Levin, 1993: 188–192 for a comprehensive list). 'Psych verbs,' or PVs, contrast with verbs that describe actions, activities, and other things that individuals *do*. This difference is one of lexical/conceptual semantics, or word meaning. Neuroscientists have used the unique semantics of this class of verbs to gain insight into how the brain represents abstract concepts like that denoted by the word 'love.' Linguistic research has also found that PVs have numerous interesting syntactic and compositional semantic properties. Thus, PVs also offer a window in to how the meaning of a sentence relates to its syntactic structure and accordingly provide insight into the neural mechanisms that underlie grammar. The sections below discuss the relevant linguistic properties of PVs and then provide a comprehensive review of the cognitive neuroscience research that has used these verbs.

Linguistic Properties of Psychological Verbs

Also called 'experiencer verbs' or 'mental verbs,' PVs have numerous lexical, syntactic, and compositional semantic properties that set them apart from other verbs (Belletti and Rizzi, 1988; Grimshaw, 1990; Baker, 1997; Pesetsky, 1995; Landau, 2009). This section reviews three key properties most important for neurolinguistic research in this area: (1) the grammatical argument structures specific to these verbs, (2) their compositional semantics, and (3) their lexical semantic properties.

Most discussed in the linguistic literature is the syntactic argument structure of PVs. Whereas most transitive verbs associate the subject of a sentence with the 'doer' or thematic agent of the action described by the verb, as in the example sentences in (1), PVs vary as to whether the thematic role of the subject is the 'experiencer' of the mental state denoted by the verb, or whether it is the object of the sentence that experiences the mental state. The sentences in (2) illustrate so-called 'subject experiencers' (SE) while those in (3) illustrate 'object experiencers' (OE).

1. a. Mary kissed John.
   b. Mary drove a truck.
2. a. Mary adored John.
   b. Mary distrusted John.
3. a. Mary amused John.
   b. Mary frightened John.

The thematic variation found in psychological verbs contrasts with other classes. For example, no English verbs encode the agent role in the object position. This fact is one of many that suggests there are grammatical restrictions on how thematic roles are syntactically encoded (Baker, 1997).

There is significant debate over the sorts of mental representations necessary to capture PV argument structure. One view holds that both SEs and OEs share a common syntactic core in which the 'experiencer' is always the subject, but for OE verbs, an additional layer of syntactic structure is necessary which leads to the observed word order (Belletti and Rizzi, 1988; Grimshaw, 1990). The word ordering comes about because the nonexperiencer argument of the verb has 'moved' to the syntactic position of the sentential subject to form a long-distance dependency. To get an intuition of this kind of theory, consider sentences such as *the boat sank* compared with *the torpedo sank the boat*, and observe that the subject of the first sentence has the same thematic role as the object in the second. To say that 'the boat' moves from the object to subject position in the first sentence captures the fact that this noun phrase has properties of both a subject and (thematic) object (e.g., Burzio, 1986). Others, noting nuanced semantic differences between OE and SE verbs, contend that OE verbs contain a more complex verb phrase which denotes a complex multi-part event (Pesetsky, 1995). This can be seen by contrasting the thematic role played by 'news' in the two sentences in (4a, b). While (4a) is well-paraphrased by saying that the news 'caused' John to be angry, even if the anger is not directed at the news (e.g., there was a political event that John did not like), in (4b) the object of anger must be the show itself (e.g., the format of the show was upsetting). Further, it is infelicitous to deny a causal connection for OE verbs (4c), but in contrast, it is easy (if stylistically awkward) to do so for SE verbs (4d). The causal meaning in (4a) is argued to require more complex verb phrase syntax, attributed either to a multilayered verb phrase

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54044-6

(Pesetsky, 1995; Pylkkänen, 2000) or a more refined verbal ontology (Arad, 1998).

4. a. The news show angered John.
   b. John was angry at the news show.
   c. # The news show angered John thought it did not make him angry.
   d. John trusted the news, though the news did not make him trust it.

To summarize, some theories attribute syntactic differences to syntactic movement (Belletti and Rizzi, 1988; Grimshaw, 1990), others to complex verbal structure (Pesetsky, 1995; Pylkkänen, 2000) or a different verbal ontology (Arad, 1998), while still others remain agnostic as to whether there is a difference in syntactic complexity (Dowty, 1991; Baker, 1997). Thus, while many (though not all) linguistic accounts take OEs to be more complex syntactically than SEs, the source of that complexity remains debated.

Independent of these syntactic debates, there is general agreement that there are semantic differences between these verb classes which may (Pesetsky, 1995) or may not (Baker, 1997) have syntactic consequences. These two kinds of PVs are distinguished both by the kinds of events that they denote and by the complexity of their entailments (Brennan and Pylkkänen, 2010), a domain of grammar called 'aspect.' The present discussion will focus on the difference between state-denoting phrases, such as 'be asleep,' 'is tall,' or 'loves tomatoes,' and non-states such as 'walk around the lake,' 'arrive at the store,' or 'build a house,' setting aside numerous finer distinctions. The term 'eventualities' will be used to describe both states and non-states, while the latter alone will be labeled 'events' (Bach, 1986).

States can be distinguished from events in English based on how they behave in the present tense. Present tense eventive verb phrases carry a habitual meaning, as shown in (5), such that the event need not be occurring at the time of utterance (Dowty, 1979). In contrast, stative verbs lack an habitual interpretation in the present tense, rather the denoted eventuality must hold at the time of utterance (6). (The hash mark '#' indicates sentences with a pragmatically odd or infelicitous meaning.) The interjective 'Look!' provides further evidence for this distinction as it forces a nonhabitual interpretation and thus is only appropriate when the eventuality holds at the time of utterance (7) (Brennan and Pylkkänen, 2010). According to these diagnostics, the sentences containing SE verbs in (8) and (9) lack a habitual implication, and thus pattern like the state-denoting predicates in (6) and (7b) and unlike the event-denoting predicates in (5) and (7a).

5. Mary runs in the park, though she is not running at the moment.
6. # Mary is asleep, though she is not asleep right now.
7. a. # Look! Mary runs in the park.
   b. Look! Mary is asleep.
8. a. Look! The child trusts the teacher.
   b. Look! The baby likes the toy.
9. a. # The children trust the teacher, though they do not at the moment.
   b. # The baby likes the kitty, though she does not like it today.

These diagnostics show that SEs are stative. The aspectual properties of OEs are more complex; essentially, they sometimes pattern as state-denoting, and sometimes as event-denoting predicates. For example, the sentences in (10) show that OE verbs are interpreted habitually in the present tense and, accordingly, need not entail that the mental state holds at the time of utterance.

10. a. The students amuse the coach, but she is not amused right now.
    b. Those horses frighten the dog, but he is not frightened right now.

In this way, OEs differ from SEs (compare the sentences in (10) with those in (9)). However OEs appear to be stative in other contexts, as can be seen with the 'Look!' test (11).

11. a. Look! The students amuse the coach.
    b. Look! Those horses frighten the dog.

It appears, then, that OEs can be interpreted as states or as events as dictated by the context, a flexibility in aspectual interpretation dubbed 'coercion' (Moens and Steedman, 1988). Note, however, that the event denotation of SE verbs is not entirely inflexible. The sentence in (12) demonstrates that SE verb meanings may also be coerced in certain limited circumstances.

12. Within 10 minutes, the children trusted the teacher.

This sentence is interpreted to mean that the children have begun to trust the teacher by the end of the time interval described. When combined with a durative modifier, SE sentences denote a change of state event, rather than a simple state.

SE and OE verbs also differ in the complexity of their entailments (Pesetsky, 1995): whereas OEs with living subjects entail that the subject *caused* the mental state to hold of the experiencer in object position, SEs contain no such entailment. Note that the presence of this entailment should be considered separately from whether it affects the syntactic argument structure of these sentences (see the discussion surrounding example (4) above). The sentences in (4) above illustrate the causative entailment of OEs. Importantly, it is infelicitous to explicitly reject a causal connection for OE verbs (4c) but not for SE verbs (4d). The absence of causal semantics in SEs is what permits, considering the example sentence in (4d), the absence of causal semantics in SEs an interpretation by which 'John' can trust the news even if the news show has never done anything to earn that trust (perhaps John is a trusting person).

To summarize, OE and SE psychological verbs differ in at least three relevant respects: (1) OEs contrast with SEs in having a noncanonical argument structure that arguably involves a more complex syntactic structure (2) OEs more flexibly denote different kinds of eventualities than SEs, and (3) OEs have more complex entailments than SEs. All of these properties will bear on the extant neuroimaging findings.

Finally, of course, by definition PVs differ from other classes of verbs in the concepts they denote. By denoting mental states, not observable actions or events, PVs serve as a window into so-called 'embodied' theories of concepts: to what degree are concepts for words represented in terms of perceptual and

motor-based neural systems. PVs lack concrete visual, auditory, or other sensory properties and are also not strongly associated with canonical motor actions or movements. Thus, they offer a class of linguistic stimulus to compare with verbs that do have concrete perceptual or motoric associations.

Neurolinguistic Research Using Psychological Verbs

This section divides neurolinguistic research using psychological verbs into three domains based on the types of questions being asked. One body of literature, drawing on aphasia studies, electrophysiology and neuroimaging, has investigated brain mechanisms involved in identifying thematic roles and verbal argument structure using PVs. A second literature has probed compositional semantic processing with PVs. Lastly are studies using patient populations and neuroimaging that have focused on conceptual representations.

PVs and Argument Structure

The syntactic properties of PVs have been leveraged in the study of the language disorder of agrammatical aphasia. There has been debate over how best to characterize the grammatical deficit(s) that underlie this disorder (see e.g., Grodzinsky, 2001 and associated commentaries). One hypothesis points to difficulties representing long-distance dependencies, such as the relationship between the subject and the verb in a passive sentence such as *The truck was driven by Mary*. See the discussion on p. 1. If such a dependency cannot be represented by the patient then they cannot recover the proper thematic role, raising the question of if and how any thematic roles are interpreted.

Grodzinsky (1995) argues that passive SE sentences provide a test for whether patients with agrammaticism use nonsyntactic heuristics. In (13) *t* indicates the grammatical object position which forms a long distance dependency with the subject in passive sentences. (The discussion here elides the additional dependencies required by the verb-internal subject hypothesis (Sportiche, 1988).) Failing to represent dependencies, Grodzinsky argues, these patients employ a heuristic which assigns the agent role to the first noun phrase. Considering the example in (13), an 'agent-first' heuristic will grant the agent role to 'the man' and a theme role to 'the woman.' Thus, the sentence will be interpreted as an active sentence with thematic reversal (e.g., the man will be doing the desiring, opposite the intended meaning). Consequently, agrammatic aphasics should show below-chance performance on passive SE sentences in comparison to active SE sentences and non-PV controls.

13. (The man)$_{THEME}$ was desired *t* by (the woman)$_{EXPERIENCER}$.

Grodzinsky (1995) reports results from a sentence–picture matching task in which four agrammatic patients with left frontal, frontotemporal, and/or frontoparietal damage that are in accordance with these predictions. These results appear to support the use of the 'agent-first' comprehension heuristic in this patient population.

Beretta and Campbell (2001) argue against this conclusion by comparing comprehension of OE and SE verbs in active and passive sentences with two agrammatic patients. They observe chance, not below-chance, performance for SE passives and report above chance performance on all other conditions. They take the contrast between chance performance on SE verbs and higher performance on passive OE verbs to reflect syntactic differences between these verb classes Following (Belletti and Rizzi, 1988), Beretta and Campbell assume that OE verbs form so-called 'adjectival passives' which contain fewer long-distance dependencies than passive sentences with SE verbs (but cf Pesetsky, 1995: chapter 2); these constructions are thus less affected by a deficit specific to long-distance dependencies. Beretta and Campbell reconcile the differences with Grodzinsky's findings by noting that the patients in the earlier study had relatively extended lesions and that certain PVs might have been difficult to visualize for the picture–sentence matching task (e.g., 'understand', 'trust'), thus inflating task difficulty.

The same pattern of comprehension results is reported by Thompson and Lee (2009) in a study of eight agrammatic patients with left hemisphere lesions. Thompson and Lee also conducted an elicitation study with comparable results, and find that OE verbs showed more errors in the active compared to passive voice. They attribute the degraded active voice performance to either increased syntactic complexity (Belletti and Rizzi, 1988) or added semantic complexity (Pesetsky, 1995) in these verbs.

The results from Beretta and Campbell and Thompson and Lee appear to argue against the operation of a nonlinguistic 'agent-first' heuristic. However, to explain the increased performance for OE passives compared to actives, Thompson and Lee, following Ferreira (1994), appeal to an alternative heuristic which prefers for arguments with prominent thematic roles to precede those with less prominent roles (in accordance with a 'thematic hierarchy' e.g., Jackendoff, 1987). Thus, passive OEs would be more frequently produced by agrammatic patients because the higher-ranked experiencer occurs first and the lower-ranked theme argument argues second.

These two findings need reconciliation: (1) chance performance for SE passives but not SE actives and (2) increased performance for OE passives compared to OE actives; while (1) precludes a strong 'agent-first' heuristic, (2) calls for some sort of heuristic which prefers for experiencers to precede themes. Note that this pattern cannot reflect syntactic differences between OE and SE passives as such differences would predict equal performance for OE passives and actives. One possible reconciliation retains weak heuristic that biases towards the inclusion of an agent role, but holds that agent and experiencer roles are not ordered with respect to each other. A lack of ordering may be plausible given that these roles never occur within the same clause in English. Under this view passive SE sentences like that in (13), when comprehended by agrammatic aphasics, contain two unordered roles neither of which is linked to a noun phrase, thereby leading to observed chance performance.

Argument ordering heuristics have also been argued to play a role in syntactic parsing in unimpaired populations. Researchers have suggested that sentences with noncanonical orderings (like OE verbs) might require additional argument structure processing, and therefore might shed light on the neural mechanisms involved in determining such structure. Electrophysiology and neuroimaging research that has pursued this aim is discussed next.

Bornkessel and colleagues have leveraged the unique properties of PVs in a series of experiments using event-related potentials (ERPs) and functional magnetic resonance imaging (fMRI) (Bornkessel et al., 2003, 2005). In verb-final German clauses, a parsing heuristic that assigned thematic roles relying simply on word order would require revision upon encountering an OE verb with reversed thematic roles (the higher-ranked experiencer role occurs in the object position). Such a mismatch would increase processing costs. These experiments use two properties of German to examine what kinds of heuristics might influence thematic role assignment. German permits the arguments of a verb to be ordered freely with respect to each other in certain circumstances (a phenomenon called 'scrambling'). Further, definite noun phrases carry grammatical case, which mark each noun's grammatical role irrespective of word order. The key question is whether OE verbs elicit a systematic neural response when the verb is encountered that patterns with a mismatch between the the argument structure inferred from word order, and the structure required by the verb. Examples of the stimuli are given in (14) (Bornkessel et al., 2003: 279–280). In these examples, NOM stands for Nominative case, which indicates the subject of a sentence and typically has the more prominent thematic role; DAT stands for dative case which is used for the direct or indirect object and typically has a less prominent role.

14. a. Active (NOM-DAT)
    … dass der Priester dem Gärtner folgt und …
    … that the$_{NOM}$ priest the$_{DAT}$ gardener follows and …
    '… that the priest is following the gardener.'
    b. ObjExp (NOM-DAT)
    … dass der Priester dem Gärtner imponiert und …
    … that the$_{NOM}$ priest the$_{DAT}$ gardener impresses and …
    '… that the priest impresses the gardener.'
    c. Active (DAT-NOM)
    … dass dem Priester der Gärtner folgt und …
    … that the$_{DAT}$ priest the$_{NOM}$ gardener follows and …
    '… that the gardener is following the priest.'
    d. ObjExp (DAT-NOM)
    … dass dem Priester der Gärtner imponiert und …
    … that the$_{DAT}$ priest the$_{NOM}$ gardener impresses and …
    '… that the gardener impresses the priest.'

Using ERPs, Bornkessel et al. (2003) find an increased early (200–350 ms) positivity for DAT-NOM sentences at the presentation of the second noun. They take this to indicate syntactic reanalysis: the initial dative noun phrase led the reader to temporarily believe that the phrase was going to be in the passive voice. At the verb they report a later positive effect (300–600 ms) for OE verbs compared to Active controls, which they take to reflect a mismatch between the thematic roles assigned prior to encountering the verb and the argument structure of the OE verb. Noting that both positivities occur earlier than the familiar P600 component associated with standard syntactic reanalysis (Osterhout and Holcomb, 1992; Kutas et al., 2006), Bornkessel et al. suggest that the effect reflects a nonsyntactic 'thematic' reanalysis, and thus favors nonsyntactic heuristics in thematic role assignment. However, the results are equally compatible with a predictive syntactic parsing algorithm (e.g., Sedivy et al., 1999; Chambers et al., 2002) for which the most likely verbal structure is postulated based on the noun phrase arguments and case marking. Such predictions would allow the assignment of thematic roles, but necessitate reanalysis when an OE verb was encountered. This alternative view is compatible with early effects of syntactic reanalysis associated with the diagnosis, rather than the repair, of a mismatch (e.g., Friederici and Mecklinger, 1996) as well as the absence of any interaction with word order, which would have been predicted if a word-order based heuristic influenced thematic role assignment.

To identify the neural generators of these effects, Bornkessel et al. (2005) examine stimuli similar to those in (14) using fMRI. In contrast to the ERP results, they report a network of left hemisphere regions where activity showed an interaction between verb type and word order, including inferior frontal gyrus (LIFG), inferior frontal junction (LIFJ), ventral premotor cortex, posterior superior temporal sulcus (LPSTS), and the intraparietal sulcus (LIPS). While they relate the LIFJ and LIPS activation to task-specific demands, they distinguish two different activation patterns in the LIFG the LPSTS. In the LIFG, the interaction reflects an increase in activation for noncanonical (DAT-NOM) ordered sentences with Active verbs paired with the reverse pattern for OE verbs; that is, these regions showed increased activation whenever the thematic order of the arguments was mismatched to the preferred order of the verb. The LPSTS, on the other hand, while showing a similar pattern for Active verbs, showed no effect for experiencer verbs.

The mismatch between the interaction effects observed in this fMRI study and the two main effects, separated in time, seen in the ERP study challenge a simple interpretation of the-processes indexed by these activations. The limited time resolution of fMRI means that the interaction results reflect processing stages that occur over many seconds and the observed interaction may reflect a summation of preverbal word-order effects and verb-driven reanalysis effects. Bornkessel et al. offer an alternative perspective, suggesting that the LPSTS pattern reflects processes sensitive to the match between the result of a word order-based heuristic and the thematic structure of the verb, while LIFG pattern reflects linearization processes that are taxed most by noncanonical arguments with Active verbs, and less so by variable argument ordering in OEs. This interpretation is compatible with the aphasia results reviewed above where left fronto(temporal) damage correlated with declined performance with passive-voice SE and active-voice OE verbs, both of which conflict with a word order-based heuristic. Given, however, the mismatch between the fMRI patterns and the ERP results above, however, the precise interpretation of the underlying cognitive processes remains unclear.

The above discussion has focused on syntactic argument structure, yet the contrasting verb types used in these studies differ in more than the order of thematic roles. As discussed above, OEs also have distinct semantic characteristics including causativity and a capacity for coercion. While neither semantic property is unique to just these verbs, the studies under consideration do not provide the set of activity verbs used, leaving open the role semantic differences may play in the observed neural differences.

PVs and Compositional Semantics

The discussion of argument structure has reached limited conclusions in part because the relationship between syntactic argument analysis and semantic composition remains poorly understood. Two studies of thematic role violations with PVs illustrate the difficulty in addressing this issue. Bourguignon et al. (2012) contrast English SE sentences and agentive sentences (e.g., "the boys have eaten the cake") containing either animate or inanimate subject noun phrases, the latter of which are semantically implausible. For agentive verbs, animacy violations lead to the so-called 'semantic P600' (sP600) (e.g., Kim and Osterhout, 2005; Kuperberg, 2007). The sP600 has received significant attention because these animacy violations might have been expected to lead to an N400 associated with lexical semantic implausibility (Kutas and Hillyard, 1980; see Lau et al., 2008 for a review). Interestingly, inanimate subjects with SE verbs do, in fact, lead to an N400 followed by a P600, in contrast to inanimate subjects with agentive verbs. Furthermore, Paczynski and Kuperberg (2011) test OE verbs against control verbs that select for animate 'patient' objects (e.g., *the guards punished the convict*) and they find an N400–P600 effect for infelicitous inanimate objects for both verb types, and no interaction effect. In sum, animacy violations lead to an N400 effect for experiencer subjects, experiencer objects, and patient objects but lead to the sP600 for subjects.

This pattern is puzzling if the sP600 reflects a mismatch between semantic composition and non-syntactic thematic heuristics that are guided by noun-phrase animacy (Kuperberg, 2007): why would animacy affect interpretation differently in object versus subject position for agentive verbs but not PVs. A possible answer lies in how semantic composition is balanced with syntactic reanalysis during parsing. The semantic mismatch between animacy and subject experiencers is of a different kind than with inanimate agents: inanimate noun phrases often have quasi-agent roles (e.g., "the car hit the wall") but they cannot be granted mental states without a semantic shift, for instance by anthropomorphizing the noun's denotation. Perhaps this semantic difference licenses distinct repair mechanisms for the two types of animacy mismatches when they occur in subject position. When mismatch is in object position, in contrast, the parser may have committed more fully to the argument structure, restricting syntactic reanalysis and thus leading to the expected semantic mismatch for inanimate objects regardless of verb type. Further work testing animacy violations against other semantic mismatches may help tease out the relative weight compositional factors play in argument structure analysis.

Neural mechanisms involved in resolving semantic mismatches have been the subject of an independent line of research in which PVs have also played a role. As described above, OE and SE verbs differ semantically in at least two respects: OEs contain a causative entailment absent in SE verbs, and OEs can be more easily coerced to refer to different kinds of events. The former is a difference in lexical semantics, while the latter taps into compositional semantics. Brennan and Pylkkänen (2010) manipulated these two properties to explore the neural mechanisms behind these two distinct kinds of semantic processing in the brain.

By using PVs, Brennan and Pylkkänen sought to maximize the similarity between target expressions. They compared triplets such as those in (15). Comparing SE with OE verbs (15a vs 15c) allowed testing of the lexical semantic differences between these classes of verbs. For coercion, the durative adverbial modifier in (15b), but not the nontemporal modifier in (15a) in combination with a stative SE verb results in a coerced change of state interpretation (see the discussion of example (12)).

15. a. Without a doubt, the child cherished the precious kitten.
    b. Within a few minutes, the child cherished the precious kitten.
    c. Without a doubt, the child scared the precious kitten.

Earlier literature suggests that both types of complexity engender a behavioral processing cost (Gennari and Poeppel, 2003; McKoon and Macfarland, 2000; Frisson and Frazier, 2005; McElree et al., 2001; Traxler et al., 2002, 2005; Piñango et al., 1999, 2006; Brennan and Pylkkänen, 2008), though no studies have directly compared the two. Further, coercion has been linked with a specific neural signature in a series of magnetoencephalography (MEG) studies (see Pylkkänen et al., 2011 for a review): an increase in anterior neuromagnetic activity spanning the midline within 200–500 ms after encountering a word that requires coercion. With this background, the contrasts in (15) using PVs served to connect the otherwise independent prior studies of each type of semantic complexity.

Brennan and Pylkkänen examined these contrasts behaviorally, using a self-paced moving window reading protocol, as well as using MEG. In self-paced reading, both kinds of complexity, lexical semantic complexity (OE vs SE) and coercion (SE states vs SE changes of state), led to increased reading times at the word immediately following the verb (see also Cupples, 2002; Ferreira, 1994). Using MEG, Brennan and Pylkkänen found an increase in left anterior temporal and left ventromedial brain activity for coercion sentences compared to control sentences within 200–400 ms following verb presentation, but found no differences for lexical semantic complexity. These results support a neural distinction between mechanisms responsible for processing lexical semantic and those responsible for processing compositional semantic information engaged by coercion. Left open, though, is what neural resources are engaged by lexical semantic complexity.

A notable limitation of these results is that the lexical semantic manipulation may covary with possible syntactic differences between these two verb classes as well as by the difference in semantics that permit OE verbs to more flexibly denote different kinds of events. Both of these factors may play a role in explaining the behavioral slowdown, though the null MEG results for lexical semantics do not provide any data that bear on this collection of differences. This null result is, furthermore, somewhat surprising in the context of the aphasia and fMRI data reviewed above which showed effects in the left inferior frontal gyrus for OE compared to SE verbs. The different results may be due to temporal properties of the underlying processes. For example, extended or offline interpretive processing that differs between verbs may be

apparent in the fMRI data but not in the rapid MEG signal. The absence of an MEG effect should not be overinterpreted, however, as it may reflect reduced sensitivity of MEG to the relevant cortical signals, among other possibilities.

PVs and Conceptual Representation

The last domain for review addresses the neural bases of conceptual representation. A long-standing debate concerns the degree to which conceptual representation is embodied, or grounded necessarily in motor and/or perceptual systems (see Pulvermüller and Fadiga, 2010; Barsalou, 2008 for reviews). An important test of this hypothesis is whether the comprehension of verbs that are associated with concrete actions and/or with specific visual characteristics, such as high motion, requires motor and/or perceptual neural systems, or whether comprehension relies on amodal neural resources. In this large literature, PVs have recently emerged as an important class of control stimuli as mental states typically lack concrete motor and sensory associations. Only a few studies have used PVs for this purpose, but the findings have rather interestingly shown only limited neural differences between verbs that denote mental states and those with more concrete denotations. A note of caution is warranted as extant studies offer only minimal description of the class of words they count as PVs ('mental' or 'abstract' verbs) and do not exhaustively list their stimuli, thus the linguistic details of these stimuli are not well characterized. If, however, this pattern stands up to further scrutiny, it will have important consequences for theories of embodied conceptual representation.

Using fMRI, Tettamanti et al. (2005) examined the brain response to spoken sentences from four categories, those with verbs describing oral actions (e.g., 'bite'), hand actions ('grasp'), foot actions ('kick'), and so-called abstract verbs (the SE 'appreciate' is given as an example). Increased activity for the three 'action' conditions compared to the 'abstract' (PV) condition was observed in the left inferior frontal gyrus (pars opercularis). Comparing each action condition individually to the abstract condition revealed additional anterior inferior frontal (pars triangularis) activity for mouth sentences, left precentral and left inferior temporal gyri for hand sentences, and left dorsal premotor cortex for leg sentences. These patterns are suggestive of motor-related activity for hand- and leg-verbs, but not mouth-verbs. Further, they do not bear on whether such activation reflects processing necessary for comprehension, or whether it reflects associative activation (cf Hickok, 2013).

Kemmerer et al. (2013) test whether motor neural systems play a causal role in conceptual processing by compare the comprehension of action (running, hitting, cutting, and speaking verbs) and nonaction (change-of-state verbs and PVs) verbs by participants with and without Parkinson's Disease (PD). PD is a neurodegenerative disease involving disruptions to motor and premotor regions causing severe movement disorders. The strong embodiment hypothesis holds that impaired motor systems will negatively impact conceptual processing of verbs with motor-related semantic features, but not verbs that denote concepts without motor features. This hypothesis is not supported by their results, which show equal performance by patients with PD and neurotypical controls in a semantic matching task for action verbs and abstract verbs such as PVs.

Turning to perceptual embodiment, further evidence that PV and action verb representation relies on similar neural substrates comes from Bedny et al. (2008, 2012). These studies focus on the degree to which concepts that imply motion engage brain regions involved in motion perception. Using fMRI, Bedny et al. (2008) compare high-motion, medium-motion, and low-motion nouns and verbs; the low-motion verb class is comprised of PVs. Consistent with previous research, high-motion verbs compared to low-motion nouns engage a posterior region of the left temporal lobe that has been implicated in motion perception. To test whether this activation reflects motion processing per se, or more abstract grammatical processing, Bedny et al. compare low-motion verbs (PVs) with high-motion nouns (words denoting animals) and find that the target region shows greater activation for PVs, consistent with an account of this region based on grammatical class (i.e. verb vs noun) and in contrast an account based on motion perception. In the same vein, Bedny et al. (2012) used similar stimuli and compared processing between congenitally blind and sighted individuals. Examining the same posterior temporal area, the results replicate an effect of grammatical class (verb vs noun), and the absence of an effect for motion (e.g., high-motion nouns vs low-motion PVs). Furthermore, in both region-of-interest and whole-brain analyses they find no regions that differentiate high-motion action verbs from PVs, in contrast to Tettamanti et al.'s findings described above, nor did they find any activation that was sensitive to the motion contrast and differed between blind and sighted individuals. In other words, no activity differentiated high from low-motion verbs, and visual experience did not explain the activation patterns that distinguished high-motion nouns from low-motion psychological verbs.

These studies provide limited evidence for embodiment; the results are compatible with abstract amodal conceptual representations. Follow-up work should consider the fine-grained as well as coarse-grained lexical semantic differences between PVs and other verb classes. For instance, the causal meaning present in OEs could lead to activation not found in noncausal verbs (e.g., verbs of speaking) that could be mistaken for conceptual-related differences. Using SE verbs would address this one concern but introduce a possible confound between verb class and event aspect as states are less complex than activities (Gennari and Poeppel, 2003; Brennan and Pylkkänen, 2010). To serve as an ideal control stimulus, PVs must be contrasted with other verb classes that denote similar kinds of events, differing only in the degree to which the latter contain concrete perceptual and/or motor associations.

Conclusions and Future Directions

This article has focused on three areas of neurolinguistic research that have been studied using psychological verbs: argument structure processing, semantic complexity, and conceptual representation. While PVs have played a role in

several key debates, a theme throughout the article has been the challenge posed by correlations between lexical semantic, compositional semantic, and syntactic properties of these verbs. As such, extant findings cannot neatly tease out syntactic from semantic effects in terms of neural activity or behavioral patterns in patients.

This difficulty is amplified by the linguistic debates surrounding these predicates, in particular there is real disagreement over the syntactic representation of the OE subclass of PVs. Attempts to reconcile poorly understood linguistic representations with neural findings risk limiting progress as uncertainty in both domains may lead to ill-formed research questions (a progression labeled 'mutual sterilization' by Poeppel and Embick, 2005). To meet this challenge, studies must take in to account the full range of conceptual, semantic, and syntactic properties of PVs. To give one concrete example, research testing the mechanisms involved in reconciling noncanonical argument ordering by comparing SE with OE verbs might address the confound between argument ordering and the presence of causative meaning in OEs by including a control condition that contained causal meaning with otherwise canonical argument ordering, as found with causative verbs such as in *The torpedo sank the boat* or *The hot sun melted the ice-cream*. Identifying more granular comparisons that isolate specific representational properties of PVs will facilitate efforts to increase the precision with which neural mechanisms can be linked with linguistic processes.

See also: Lexical Semantics; Movement Theory and Constraints in Syntax; Semantics; Syntax (General); Syntax–Semantics Interface; Valency and Argument Structure in Syntax.

Bibliography

Arad, M., 1998. Psych-notes. In: Harris, J., Iten, C. (Eds.), UCL Working Papers in Linguistics, vol. 10. UCLWPL, London, pp. 203–223.

Bach, E., 1986. The algebra of events. Linguistics and Philosophy 9, 5–16.

Baker, M.C., 1997. Thematic roles and syntactic structure. In: Haegeman, L. (Ed.), Elements of Grammar. Kluwer, Dordrecht, pp. 72–137.

Barsalou, L.W., 2008. Grounded cognition. Annual Review of Psychology 59, 617–645.

Bedny, M., Caramazza, A., Grossman, E., Pascual-Leone, A., Saxe, R., 2008. Concepts are more than percepts: the case of action verbs. Journal of Neuroscience 28 (44), 11347–11353.

Bedny, M., Caramazza, A., Pascual-Leone, A., Saxe, R., 2012. Typical neural representations of action verbs develop without vision. Cerebral Cortex 22 (2), 286–293.

Belletti, A., Rizzi, L., 1988. Psych-verbs and θ-theory. Natural Language & Linguistic Theory 6 (3), 291–352.

Beretta, A., Campbell, C., 2001. Psychological verbs and the double-dependency hypothesis. Brain and Cognition 46 (1–2), 42–46.

Bornkessel, I., Schlesewsky, M., Friederici, A., 2003. Eliciting thematic reanalysis effects: the role of syntax-independent information during parsing. Language and Cognitive Processes 18 (3), 269–298.

Bornkessel, I., Zysset, S., Friederici, A.D., von Cramon, D.Y., Schlesewsky, M., 2005. Who did what to whom? The neural basis of argument hierarchies during language comprehension. NeuroImage 26 (1), 221–233.

Bourguignon, N., Drury, J.E., Valois, D., Steinhauer, K., 2012. Decomposing animacy reversals between agents and experiencers: an ERP study. Brain and Language 122 (3), 179–189.

Brennan, J., Pylkkänen, L., 2008. Processing events: behavioral and neuromagnetic correlates of aspectual coercion. Brain and Language 106 (2), 132–143.

Brennan, J., Pylkkänen, L., 2010. Processing psych verbs: behavioral and MEG measures of two different types of semantic complexity. Language and Cognitive Processes 25 (6), 777–807.

Burzio, L., 1986. Italian Syntax. Reidel, Dordrecht.

Chambers, C.G., Tanenhaus, M.K., Eberhard, K.M., Filip, H., Carlson, G.N., 2002. Circumscribing referential domains during real-time language comprehension. Journal of Memory and Language 47, 30–49.

Cupples, L., 2002. The structural characteristics and on-line comprehension of experiencer-verb sentences. Language and Cognitive Processes 17 (2), 125–162.

Dowty, D., 1979. Word Meaning and Montague Grammar. Kluwer, Dordrecht.

Dowty, D., 1991. Thematic proto-roles and argument selection. Language 67 (3), 547–619.

Ferreira, F., 1994. Choice of passive voice is affected by verb type and animacy. Journal of Memory and Language 33 (6), 715–736.

Friederici, A.D., Mecklinger, A., 1996. Syntactic parsing as revealed by brain responses: first-pass and second-pass parsing processes. Journal of Psycholinguistic Research 25 (1), 157–176.

Frisson, S., Frazier, L., 2005. Carving up word meaning: portioning and grinding. Journal of Memory and Language 53 (2), 277–291.

Gennari, S., Poeppel, D., 2003. Processing correlates of lexical semantic complexity. Cognition 89 (1), B27–B41.

Grimshaw, J., 1990. Argument Structure. MIT Press, Cambridge, MA.

Grodzinsky, Y., 1986. Language deficits and the theory of syntax. Brain and Language 27 (1), 135–159.

Grodzinsky, Y., 1995. Trace deletion, θ-roles, and cognitive strategies. Brain and Language 51 (3), 469–497.

Grodzinsky, Y., 2001. The neurology of syntax: language use without Broca's area. Behavioral and Brain Sciences 23 (01), 1–21.

Hickok, G., 2013. Do mirror neurons subserve action understanding? Neuroscience Letters 540, 56–58.

Jackendoff, R., 1987. The status of thematic relations in linguistic theory. Linguistic Inquiry 18 (3), 369–411.

Kemmerer, D., Miller, L., Macpherson, M.K., Huber, J., Tranel, D., 2013. An investigation of semantic similarity judgments about action and non-action verbs in Parkinson's disease: implications for the embodied cognition framework. Frontiers in Human Neuroscience 7, 146.

Kim, A., Osterhout, L., 2005. The independence of combinatory semantic processing: evidence from event-related potentials. Journal of Memory and Language 52, 205–225.

Kuperberg, G.R., 2007. Neural mechanisms of language comprehension: challenges to syntax. Brain Research 1146, 23–49.

Kutas, M., Hillyard, S.A., 1980. Reading senseless sentences: brain potentials reflect semantic incongruity. Science 207, 203–205.

Kutas, M., Petten, C.K.V., Kluender, R., 2006. Psycholinguistics electrified II (1994–2005). In: Gernsbacher, M.A., Traxler, M.J. (Eds.), The Handbook of Psycholinguistics, second ed. Elsevier Press, New York.

Landau, I., 2009. The Locative Syntax of Experiencers. MIT Press, Cambridge, MA.

Lau, E.F., Phillips, C., Poeppel, D., 2008. A cortical network for semantics: (de)constructing the N400. Nature Reviews Neuroscience 9 (12), 920–933.

Levin, B., 1993. English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press, Chicago.

McElree, B., Traxler, M.J., Pickering, M.J., Seely, R.E., Jackendoff, R., 2001. Reading time evidence for enriched semantic composition. Cognition 78, B15–B25.

McKoon, G., Macfarland, T., 2000. Externally and internally caused change of state verbs. Language 76 (4), 833–858.

Moens, M., Steedman, M., 1988. Temporal ontology and temporal reference. Computational Linguistics 14 (2), 15–28.

Osterhout, L., Holcomb, P.J., 1992. Event-related potentials elicited by syntactic anomaly. Journal of Memory and Language 31, 785–806.

Paczynski, M., Kuperberg, G.R., 2011. Electrophysiological evidence for use of the animacy hierarchy, but not thematic role assignment, during verb argument processing. Language and Cognitive Processes 26 (9), 1402–1456.

Pesetsky, D., 1995. Zero Syntax: Experiencers and Cascades. MIT Press, Cambridge, MA.

Piñango, M.M., Winnick, A., Ullah, R., Zurif, E., 2006. Time-course of semantic composition: the case of aspectual coercion. Journal of Psycholinguistic Research 35 (3), 233–244.

Piñango, M.M., Zurif, E., Jackendoff, R., 1999. Real-time processing implications of enriched composition at the syntax-semantics interface. Journal of Psycholinguistic Research 28 (4), 395–414.

Poeppel, D., Embick, D., 2005. Defining the relation between linguistics and neuroscience. In: Cutler, A. (Ed.), Twenty-First Century Psycholinguistics: Four Cornerstones. Lawrence Erlbaum, Mahwah, NJ.

Pulvermüller, F., Fadiga, L., 2010. Active perception: sensorimotor circuits as a cortical basis for language. Nature Reviews Neuroscience 11, 351–360.

Pylkkänen, L., 2000. On stativity and causation. In: Tenny, C., Pustejovsky, J. (Eds.), Events as Grammatical Objects. CSLI Publications, Stanford, pp. 417–422.

Pylkkänen, L., Brennan, J., Bemis, D.K., 2011. Grounding the cognitive neuroscience of semantics in linguistic theory. Language and Cognitive Processes 26 (9), 1317–1337.

Pylkkänen, L., McElree, B., 2007. An MEG study of silent meaning. Journal of Cognitive Neuroscience 19 (11), 1905–1921.

Sedivy, J.C., Tanenhaus, M.K., Chambers, C.G., Carlson, G.N., 1999. Achieving incremental semantic interpretation through contextual representation. Cognition 71 (2), 109–147.

Sportiche, D., 1988. A theory of floating quantifiers and its corollaries for constituent structure. Linguistic Inquiry 19 (3), 425–449.

Tettamanti, M., Buccino, G., Saccuman, M.C., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S.F., Perani, D., 2005. Listening to action-related sentences activates fronto-parietal motor circuits. Journal of Cognitive Neuroscience 17 (2), 273–281.

Thompson, C.K., Lee, M., 2009. Psych verb production and comprehension in agrammatic Broca's aphasia. Journal of Neurolinguistics 22 (4), 354–369.

Traxler, M.J., McElree, B., Williams, R.S., Pickering, M.J., 2005. Context effects in coercion: evidence from eye movements. Journal of Memory and Language 53, 1–25.

Traxler, M.J., Pickering, M.J., McElree, B., 2002. Coercion in sentence processing: evidence from eye-movements and self-paced reading. Journal of Memory and Language 47, 530–547.

Neurolinguistics, History of

Paul Eling, Radboud University Nijmegen, Nijmegen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The history of neurolinguistics is described, starting with the modern period, beginning in the 1960s. Subsequently, an overview of studies and theories on the effects of brain lesions on language behavior is presented. Up to the nineteenth century, mainly single case descriptions have been recorded. Franz Joseph Gall stimulated the search for the localization of mental functions in the brain, in particular language. Originally, language was primarily viewed as a production and perception of words. In the twentieth century, aphasiologists began to investigate syntactical and phonological issues in relation to aphasia. Interest in linguistic aspects of language resulted in the emergence of neurolinguistics as a scientific discipline.

Birth

Levelt (2012) has written an extensive and insightful history of psycholinguistics. A remarkable feature of this history is that it pertains only to the *pre-Chomskyan* period: it ends with Chomsky's opposition to Skinner's behavioristic explanation of language behavior (Chomsky, 1959). Most psycholinguists would argue that, feat actually marked the *beginning* of psycholinguistics. Levelt argues that the psychology of language has a long history, starting at the end of the eighteenth century, concerns various domains of research, including also the study of the consequences of localized brain lesions on language or aphasiology in the nineteenth century. In a similar way, the birth of neurolinguistics as a scientific discipline in the 1960s should not be considered as the beginning of something completely new. As with psycholinguistics, there is a very long history, going back to antiquity, of scholars writing about language and the brain, or rather language impairments following brain disorders; it overlaps to a large extent with that of psycholinguistics. This historical overview will therefore deal more extensively with the period *before* neurolinguistics 'officially' emerged around the 1960's, than the time since its 'birth'.

While Chomsky's review of Skinner's book is for many the marker of the birth of psycholinguistics, there is no precise, commonly recognized starting point for neurolinguistics. Chomsky's publications probably played an important role as well, be it indirectly by stimulating researchers to incorporate linguistic notions in their study of language disorders, in particular of aphasia. One could say that neurolinguistics emerged, due to various attempts to provide the study of language disorders a stronger theoretical basis. That neurolinguistics was not a completely new discipline, but rather the continuation of the various disciplines involved in the study of language and the brain, was evident to the founders of 'modern' neurolinguistics.

Eric Lenneberg (1960) in a book review of Penfield and Robert's (1959) well-known *Speech and brain mechanisms* (see below) introduced the term *aphasiology* to indicate the field of study that 'has been with us *de facto* for a good hundred years'. Edith Trager (1960) mentioned Lenneberg's proposal, but suggested the term *neurolinguistics* for the interdisciplinary field of research, the substance of which was the relationship between the human nervous system and language. Dingwall and Whitaker (1974) in their review of the literature, titled *Neurolinguistics*, simply referred to Trager's definition of the area and discussed studies in different sections: studies on the localization of language functions and studies on neurological and linguistic analysis of aphasia. A third section was called *manipulative studies of brain function*, dealing with different methodologies of examining the central question, such as neurosurgical studies, split brain studies, pharmacological deactivation, and electrical stimulation of the brain. Interestingly, they noted that some years earlier, an attempt had been made to introduce the term *biolinguistics*, but it failed to gain acceptance. A journal with that title started in 2007 (Boeckx and Grohmann, 2007) and the core questions apparently are more or less the same as those of neurolinguistics in the 1970s.

David Caplan (1987) wrote the first comprehensive introduction to the emerging fields of neurolinguistics and linguistic aphasiology (see **Figure 1**). This distinction indicates that neurolinguistics encompassed more than aphasiology. Neuroimaging studies, examining patients and healthy individuals,

Figure 1 David Caplan.

were quickly becoming popular. He distinguished between four 'forms of arguments' that are relevant for developing a theory on language and the brain: (1) arguments from linguistics and psycholinguistic structures to neural structures; (2) arguments from neural structure and function to the nature of language and neural processing of language; (3) arguments from parallel structures; and (4) analysis of the results of disease of the brain. Furthermore, he described five techniques of study in neurolinguistics: (1) stimulation of the brain; (2) recording neurophysiological correlates of language; (3) lateralized stimulus presentation; (4) anesthetization of a single hemisphere (i.e., the Wada test; see below); and (5) metabolic scanning (i.e., blood flow scanning). This list illustrates that although aphasiology and neurolinguistics deal to a very large extent with the same issues, there is more under the sun than aphasic patients that can provide answers concerning the relationship between language and the central nervous systems. In particular, the technical developments concerning the various brain imaging techniques have played a crucial role in the last two decades and dramatically changed the situation: aphasia was practically the main source of information for centuries, but we can now examine the core question *in vivo* in intact human individuals in a very controlled way, looking at reliable functional changes occurring within milliseconds and in very small regions.

Neurolinguistics as a Scientific Discipline

A defining feature of a discipline is the existence of written records, in particular handbooks and scientific journals. The American Harry Whitaker, trained as a linguist, played a key role in this process (see **Figure 2**). As an editor he started a book series entitled *Perspectives in Neurolinguistics and Psycholinguistics*, 27 volumes large, dealing with various issues, such as neurolinguistics, lateralization, development, and bilingualism. He was also the founding editor of the journal *Brain and Language*, introduced in 1974. The interdisciplinary study of brain and language was the central motive for starting this journal. With Brigitte Stemmer, Whitaker edited the *Handbook of Neurolinguistics* (1998), which opened with a chapter on the history of neurolinguistics from his hand. And the modern development is reflected in the *Handbook of the Neuroscience of Language* (2008) from the same editors.

The *Journal of Neurolinguistics* was founded in 1985 by Fred Peng. Not much later, in 1987, Chris Code and David Muller started the journal *Aphasiology*. Looking at the editorial statements, it seems that the *raison d'etre* of these journals was not a new area of scientific questions, but rather the need for publishing a large number of *interdisciplinary* articles, for which apparently there was not enough room in more neurologically oriented journals, like *Brain*, or more general journals like *Neuropsychologia* or *Cortex* (in which studies on language and the brain also have been published).

Russian Roots

One major root of the 'new' field of neurolinguistics lies in the United States, in particular the Boston Aphasia Unit, established by the neurologist Norman Geschwind (1926–84) and the psychologist Harold Goodglass (1920–2002) (see below). A second root originated in Russia. According to a communist view, everybody is equal and has equal chances, given the proper environmental circumstances. Lev Vygotsky (1896–1934) formulated a developmental psychology based on this fundamental assumption. Language plays a crucial role in this system, as it provides the means of social communication. Alexander Luria (1902–77; see **Figure 3**) adopted Vygotsky's views and applied these in his work as a clinical neuropsychologist. His vast experience with individuals with brain lesions allowed him to develop his well-known model of the working of the brain (Luria, 1973). According to this model, language acts as a control mechanism: it is used to regulate and

Figure 2 Harry Whitaker.

Figure 3 Alexander Luria.

organize behavior. The study of language disorders was very important for understanding human behavior and Luria described his ideas in his *Basic problems of neurolinguistics* (1976). However, linguistics here did not refer to anything like Chomskyan principles (Buckingham, 2007). On the contrary, it was based on the Aristotelian association principles that had also been used in the classical conditioning theory, developed by the Russian physiologist Ivan Pavlov (1849–1936).

Finally, the Russian linguist Roman Jakobson (1896–1982) was a contemporary of Vygotski and Luria and they most likely knew each other (Akhutina, 2003). He settled in the United States in 1941. He applied his linguistic ideas on language development in children also to aphasia (Jakobson, 1941). He relied on Kurt Goldstein's descriptions of aphasic phenomena, but argued that they should be analyzed within a linguistic framework, specifically the structuralist approach from Ferdinand de Saussure, rather than the holistic psychological framework. Jakobson's work played an important role for the introduction of linguistics in aphasiology in the United States, specifically in the analysis of phonological disturbances (Blumstein, 1973; Friedrich, 2006).

We have witnessed the rise of many 'new' disciplines since the 1960s, many carrying the prefix 'neuro' (for instance, neuropsychology, neurolinguistics, and neurophilosophy), and there were also new branches within 'linguistics' (for instance, sociolinguistics, anthropological, and historical linguistics). However, due to the arrival of the enormous technical facilities to study, *in vivo*, structure and function of the brain, scientists of these many different strands currently collaborate under the flag of *cognitive neurosciences*. This pattern of specialization and integration indicates that what counts is the work that is done, rather than its flag. Rather than focusing on these more recent activities in the field of neurolinguistics since the 1960s, this history will deal with the description and analysis of language disorders. First, the history of aphasia will be described. Subsequently, the issue of hemispheric differences will be discussed, followed by a number of related issues. The emphasis will be more on the various approaches, research areas, and significant milestones than on concrete findings.

Antiquity

When examining early descriptions of language disorders associated with brain lesions, it is important to realize that language was conceptualized in a different way than we currently do. The common conceptualization of mental functions was, broadly speaking in terms of perception, evaluation or judgment, and memory, also referred to as the inner senses (see also Whitaker, 1998a). Language was not a separate function in this scheme. Knowledge of words was regarded to be an aspect of memory and for centuries, language disorders that we now might interpret as aphasia were referred to as the loss of the memory for words. Only in the nineteenth century, this picture changed dramatically, but even at the end of the nineteenth century, aphasia was interpreted as simply a form of *partial memory loss (amnésie partielle)* (Sollier, 1892: p. 61).

Also the conceptualization of the brain and brain functioning was completely different from current views. Related to

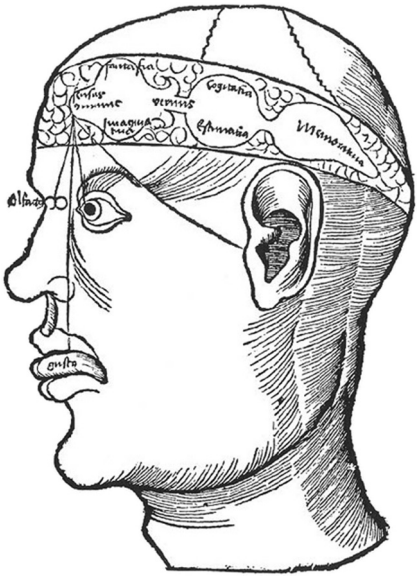

Figure 4 The inner senses.

the model of the inner senses mentioned above was the *ventricular* or *cell theory*: the inner senses, representing the soul or psyche, were associated with the three ventricles in the brain (see **Figure 4**; Clarke and Dewhurst, 1996). The white matter played only a secondary role, while often the cortex was compared to the rind of a tree. The significance of the cortex became clear at the beginning of the nineteenth century and a systematic division and nomenclature of cortical regions was developed by Louis Pierre Gratiolet (1854). With this in mind, we now turn to these early descriptions of language disorders.

An early medical document is the *Edwin Smith Papyrus*, found in Egypt. It is dated from around 1700 BC, but is obviously a copy of a much older papyrus and contains 48 case descriptions and advices for physicians how to deal with particular disorders (Breasted, 1930). Among these cases, there are five with language disorders. A typical description of such a case is Case 20:

> One having a wound in his temple, penetrating to the bone, perforating his temporal bone; while he discharges blood from both his nostrils, he suffers with stiffness in his neck, he is speechless. An ailment not to be treated.
> (Breasted, 1930: p. 286)

The Egyptians adhered to the so-called cardiocentric system, implying that the soul was associated to the heart and not the brain. Although the Greek scholar Aristotle (384–322 BC) also subscribed to this cardiocentric view, many others assumed that the brain was the organ associated with the soul or mind. Among these were Plato (428–347 BC) and Herophilus (335–280 BC), who is considered by many as the father of anatomy and the founder of the above mentioned ventricular theory. While Aristotle's view on the localization of the mind was wrong, he had an enormous impact on many scientists, working in different disciplines, for many centuries. His theory of the functioning of memory, still frequently cited, is

expressed in his well-known association laws: concepts are associated because of *similarity*, *contiguity*, and *contrast*. The organization of an individual's knowledge, including language, presumably is based on these rules. Aristotle had a detailed view on language (McKeon, 1946), but it is his description of speech disorders that has been incorporated in the medical literature (Wollock, 1997). Speech was described in terms of the instruments used (lungs, air, larynx), the modulation of the noise into vowels and consonants, and a third important feature was the sequential order of sounds, reflecting Aristotle's law of contiguity. In line with this view of speech production, different speech disorders were recognized. The reduction of control over the production of speech sounds was termed *traulotes*. The absence of sounds or syllables was named *psellotes* and the incorrect connection of speech sounds, syllables, the duplication of syllables, or the lengthening of sounds was named *ischophonia*. The latter type seems to reflect what we would call stuttering. This classification has been used until the Renaissance.

In addition to these mere general descriptions of brain functioning and speech, a number of descriptions can be found in medical and nonmedical (e.g., the Bible) documents of individuals suffering from language problems (see O'Neill, 1980). However, these are purely descriptive, often very short with few details, and they usually do not reveal much from underlying notions of brain functioning or the nature of language disorders.

Renaissance

O'Neill (1980) identified a turning point at the beginning of the sixteenth century. Giovanni de Vigo's (1450–1525) *Practica Copiosa in Arte Chirurgia* (1514) contains a case of traumatic speechlessness. Similar observations were made by Giacomo de Berengario (1460–1530), Nicola Massa (1504–89), William Clowes (1540–1604), Hieronymus Mercurialis (1530–1606), and Johann Schenck von Grafenberg (1530–98). These works reflect a shift in medicine during the sixteenth century from mere description toward an analysis of mental functions in relation to the brain, a change presumably due to the fact that physicians started to register their own anatomical observations, rather than relying on classical texts.

Benton and Joynt (1960; see also Benton, 2000) reviewed Renaissance descriptions. Two cases from the seventeenth century, from Johann Schmidt (1624–90) in 1676 and Peter Rommel (1634–1708) in 1683, attained a level of analysis that was superior to those that had hitherto appeared. Schmidt noted that his patient substituted one word for another, that he could not recognize letters or read words, but that he was able to write to dictation. Rommel (1683) presented his case under the title *On a rare aphonia*, apparently surprised by the fact that his patient could not speak spontaneously but could still recite her prayers, an early description of a dissociation between voluntary and automatic speech. The Swiss physician Johannes Jakob Wepfer (1620–59) collected his observations in several works, notably *Observationes medico-practicae de affectibus capitis internis & externis*, posthumously published in 1727 (Luzzatti and Whitaker, 1995). At least 13 cases showed a clear language disorder, which Wepfer described as the loss of memory. In point of fact, the classical authors typically interchanged the terms 'memory' and 'language' (the latter always to be understood as speech or expressive language).

Giovanni Battista Morgagni's (1682–1771) famous *De sedibus et causis morborum per anatomen indagatis* (Morgagni, 1762) contained a number of cases of speechlessness associated with apoplexy and head trauma. He frequently indicated that although the patient could not speak, he was able to understand spoken language. Indeed, in none of his cases a speech comprehension deficit is observed. It should be noted that a specific speech comprehension deficit was not recognized as such until the late 1860s; comprehension problems were interpreted as intellectual deficits (see also Boller, 1978). Morgagni, like Wepfer, observed on many occasions the occurrence of a speech problem in the presence of a *right-sided* hemiplegia and a lesion in the *left* hemisphere.

The first major work devoted to the subject of aphasia was Johann Gesner's (1738–1801) chapter *Die Sprachamnesie*, in volume 2 of his *Samlung von Beobachtungen aus der Arzneygelahrtheit* (Gesner, 1770; Luzzatti, 2002; see **Figure 5**). Gesner tried to analyze the nature of the impairment more precisely and concluded that it consisted of an inability to associate images or abstract ideas to their expressive verbal symbols. Gesner's theoretical analysis is a first attempt to provide a functional explanation of aphasia. A somewhat similar attempt was made by Alexander Crichton (1763–1856) also attempted to use speech disorders as a means to understand underlying mental processes and impairments. Like Gesner, he suggested that aphasia may be regarded as a very singular defect of (verbal) memory, specifically, "a defect of that principle by which ideas, and their proper expressions, are associated" (Crichton, 1798).

Figure 5 Johann Gesner.

The Nineteenth Century

The study of language and the brain changed dramatically in the nineteenth century. The German physician Franz Joseph Gall (1758–1828; see **Figure 6**) originally studied the anatomical structure of the brain. He then decided to analyze *human nature* in relation to the brain. He formulated a number of principles that together implied the notion of localization of faculties or *organs* of mental functions: he believed that faculties were inborn and every faculty had its own place on the cerebral cortex. Next to his revolutionary (heretic, according to many) view on localization of the mind, he also had different view on mental functions. He did not believe in a unified mind, the *res cogitans*, like Descartes, but argued that faculties were naturally acquired mechanisms to interact with the world; there are various lower order faculties, important for survival, and other functions, relevant for interaction and communication and these also cause the individual differences in character. These faculties must be inferred by systematic examination, they cannot be defined deductively. One way to collect supportive evidence for the existence of specific faculties was to find 'bumps' on the skull systematically associated with specific well-developed faculties, in humans as well as in animals. Another useful approach was to study the consequences of focal lesions on mental faculties. The study of the localization of lesions leading to language disorders became the prototypical case in this discussion on localization of function. Gall (1822–1825) himself argued that there were two areas associated with language, situated in the cortex just behind the eyes, one area responsible for language in general and the other for the memory of words.

The physiologists in Paris opposed heavily against Gall's localization theory, in particular Jean Pierre Flourens (1794–1867) in his *Examen de la Phrenologie* (1842). However, Jean-Baptiste Bouillaud (1796–1881) used this latter approach, which was referred to as the clinicoanatomical method, to argue that Gall's *principle of localization of faculties* was essentially correct: based on an analysis of lesion sites of aphasic patients language appeared to be localized in the anterior part of the brain (a more systematic classification of lobes, gyri, and sulci was only developed around the 1850s). Bouillaud's (1825, 1848) attempts were not really successful, until Paul Pierre Broca (1824–80) presented the case of a patient, *monsieur Leborgne* and nicknamed *Tan* after the syllable he used to communicate, to members of the Anthropological Society (Broca, 1861; see **Figure 7**). Only recently Domanski (2013) was able to reveal some biographical details of *monsieur Leborgne*.

Within a period of 4 years, and on the basis of a limited number of patients, Broca convinced many of his colleagues that a very specific mechanism, the memory for the articulation of words, was situated at the foot of the third frontal gyrus, and only in the left hemisphere (Harris, 1991). Broca assumed that the language faculty consisted of several layers of processes, from a general faculty to use language to more specific faculties to express oneself in a specific modality: speech, writing, and gesticulation (see Eling, 1984). A detailed description by Dominique Larrey (1814) of a patient, with a language disorder similar to that of Tan, and a verified postmortem localization of the lesion in the left frontotemporal region had gone unnoticed: the time was not ripe (Feinsod and Aharon-Peretz, 1994).

The next step was taken by the German physician Carl Wernicke (1848–1905; Eggert, 1977). Having analyzed the pathway of nerve fibers from the ear to the temporal cortex, he looked for patients with an impairment in understanding language (Wernicke, 1874; see **Figure 8**). For Broca, *aphemie* (the term aphasia was introduced by Armand Trousseau (1864) and generally accepted) implied an inability to articulate words, while intelligence and comprehension are intact. Wernicke argued that there are also patients with a severe problem in understanding language. The lesion responsible for this condition would be situated in the temporal lobe. Wernicke then went one step further: he argued that there are language *centers*: centers with word images, images for recognizing spoken words, for speaking words, but also for reading and writing words. (This switch from faculties to centers has never been analyzed and

Figure 6 Franz Joseph Gall.

Figure 7 Paul Broca.

Figure 8 Carl Wernicke.

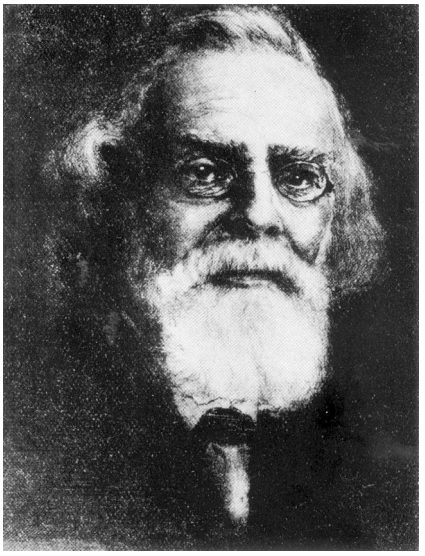

Figure 9 Heymann Steinthal.

clarified.) These centers are also interconnected: an interruption of the connection, a disconnection of the center for understanding words and the center for speaking words results also in a language disorder, conduction aphasia, where patients can understand and speak more or less normally, but they have problems repeating sentences spoken to them.

The German psychiatrist Carl Finkelnburg (1832–96) in 1870 (see Duffy and Liles, 1979) also opposed the idea that in aphasic patients only speech is affected. He described five cases and argued that in most cases different forms of language production (speech and writing) and language perception (hearing and reading) are involved, essentially all processes that convey meaning through learned sensory signs – that is through symbols. He spoke of the symbolic faculty or the *facultas signatrix* as it was termed by the German philosopher Immanuel Kant.

While Wernicke is traditionally regarded as the person who introduced the idea of a language system, consisting of centers and their interconnections, the English physician Henry Charlton Bastian (1837–1915) had developed a rather similar diagram in 1869. Lorch (2013) has shown that his model, and in particular his clinical protocol for analyzing the nature of language disorders in various modalities had been very influential for many decades in England.

Developing diagrams, based on the analysis of intact and impaired language functions, became a popular activity among students of the brain. The site of the lesions was frequently a point of discussion and could only be verified postmortem and this information was therefore not always available. Few words were spent on the psychological theory of language used in these studies.

There were a few important exceptions, for instance, the German linguist Chajim Steinthal, the German internist Adolf Kussmaul, and the English neurologist John Hughlings-Jackson. Steinthal (1823–99; see **Figure 9**; see Eling, 2006) has been called the first psycholinguist. Unlike his fellow linguists who focused on comparative linguistics, he was also interested in language as a psychological process and developed a theory of producing and understanding an utterance. Having formulated the first psycholinguistic theory (Steinthal, 1871), he demonstrated the usefulness of it in an appendix in which he applied his ideas to aphasia. An important point was his differentiation between *aphasia* and *acataphasia*. In aphasia, the problem is at the word level, a word memory retrieval problem, while acataphasia refers to a problem at the sentence level, the relation of words among each other, representing the actual meaning of the utterance. His theory was rejected by physicians in Berlin, arguing that he had no knowledge of aphasic patients.

In contrast to Steinthal, Adolf Kussmaul (1822–1902) was respected by his colleagues and his monograph on *Sprachstörungen* (1877) was regarded as probably his best scientific contribution. While Wernicke stressed the need to localize centers in the brain, Kussmaul focused on language as a functional system and his model is a (psychological) process model, not a neuroanatomical model. Like Wernicke he distinguished different modalities. But he also distinguished between levels of language, referring to Steinthal (Levelt, 2012). He distinguished between disorders of articulation (*dysarthria*) and cortical language disturbances, which he subdivided in *dysphasias* or aphasias (real language problems) and *dyslogias* or *dysphrasias* (due to intelligence problems). Kussmaul introduced the notion of *agrammatism*, referring to grammatical problems and here also he distinguished various forms: aphasic acataphasia (omission of words in a sentence), grammatical acataphasia (which we would now refer to as telegraphic speech), and syntactic aphasia and we would call that paragrammatism, the combination of sentences or sentence fragments. However, Kussmaul's ideas on sentence processing were not generally adopted by his colleagues at that time. His model on word processing, however, was cited frequently.

Jackson (1878) did not subscribe to models that localize language centers, arguing that *to localize a symptom is not the same*

as localizing a function. Moreover, he indicated that the fundamental unit of analysis should not be a word, but a *proposition*. His opposition against strict localization was important, but he himself did not formulate a psychologically or linguistically oriented framework to explain aphasic symptoms.

The Twentieth Century

The first half of the twentieth century can be characterized by the opposition to the rather simplistic localization approach, in two ways with Pierre Marie (France; see **Figure 10**), Von Monakov (Switzerland), Goldstein (Germany), and Head (England) as key figures. First, the idea of strict localization was contested. Brain areas are interconnected and the brain functions more as a whole than as a set of specialized functions. Although this argument against the localizationist approach sounded plausible, it did not lead to a new and more productive description of the functioning of the brain. Secondly, language or rather communication was more than words, to study the relation between language and the brain by associating words to brain regions, would be to miss a crucial feature of language, the meaning of an utterance based on the relations among words and word forms. This approach is usually characterized as *holistic*. The contest between the parties led to the famous *Aphasia Discussion* in Paris in 1908 (Lecours et al., 1992). This development is illustrated in various ways (see York, 2009).

A number of scholars studied grammatical problems in the beginning of the twentieth century. Among these were Arnold Pick, Karl Bonhoeffer, Karl Kleist, Kurt Goldstein, Erich Salomon, and Max Isserlin. Henry Head, Hughlings-Jackson's pupil and successor, was familiar with this work and referred to them as the *psychological school* (Howard, 1985).

The German–Hungarian neuropsychiatrist Arnold Pick (1851–1924; see **Figure 11**) broke with the associationistic tradition of Wernicke and adopted the approach developed by Oswald Külpe in Würzburg, also known as *Thinking Psychology*. Karl Buhler (1897–1963), also working there, attempted to identify different stages in thought formulation on the road from thinking to speech, which culminated in an early textbook on psycholinguistics: *Sprachtheorie* (Language Theory, 1934). Pick had collaborated with Buhler, and he used these ideas to analyze speech problems in aphasic patients. One of his first papers was on agrammatism (Pick, 1898), the most comprehensive formulation of his theory of speech production in aphasia appeared posthumously in 1934 (Pick, 1931; see De Bleser, 1987 for an extensive discussion).

Figure 11 Arnold Pick.

Figure 10 Pierre Marie.

Henry Head (1861–1940) played a prominent role in England, first by promoting Hughlings-Jackson's work, secondly by criticizing many of the early aphasiologists by calling them *diagram makers*, and thirdly by presenting his view on aphasia in his two-volume monograph *Aphasia and Kindred Disorders of Speech* (Head, 1926; Jacyna, 2005). Head assumed that in some sense, aphasia was also a disorder of intelligence, though not general intelligence but of thought as far as it depends on language. He distinguished between *verbal, syntactical, nominal,* and *semantic defects*. A verbal defect is primarily characterized by a problem in finding the right words; understanding of individual words and simple sentences is intact. A syntactic defect is characterized by a severe problem with rhythm and syntax. The patient may speak rather quickly, produces jargon, prepositions, conjunctions, and articles may be omitted. A nominal defect is not a pure linguistic phenomenon; the patient can produce all sorts of words, but has difficulty to find the right words for what he wants to say. A semantic defect occurs when a patient fails to understand the full significance of words and phrases. These categories represent theoretical entities

that rarely occur in a pure form in clinical cases. Clealry, Head's criticism of the classical aphasiologists has had a greater impact than his new taxonomy.

The German neuropsychiatrist Kurt Goldstein (1878–1965), who emigrated to the United States in 1935, rejected the theory and classification of his teacher Carl Wernicke, and was influenced by the Gestalt psychology, becoming popular at that time (Eling, 2012). Brain lesions result, according to Goldstein, in problems with *abstraction*: the concrete features of the situation determine perception and thought. Patients may also suffer from catastrophic reactions and their language behavior may also reveal attempts to readjust. Although this approach was sympathetic because of its emphasis on the patient as a person, it had little explanatory power and could not simply be tested in an objective manner.

Perhaps Head was quite correct when he judged the situation in the middle of the twentieth century as *chaos*. In contrast to the two theoretical camps, on the one hand the rather strict localizationist and associationist approach started by Wernicke and the more global, holistic, psychological approach of Pick, Goldstein, and Head, a new approach was tried in the United States by the neurologist Theodore Weisenburg (1876–1934) and the psychologist Katherine McBride. They studied aphasia from a clinical point of view, collected data from large numbers of aphasic patients on a battery of subtests, addressing various modes of language behavior. On the basis of these kinds of observations, they suggested that patients could be classified as predominantly expressively disturbed, receptively, mixed, or as amnestic. This approach, however, was not convincing from a theoretical point of view.

The chaotic situation had led to a diminished interest in aphasia in general in the middle of the twentieth century. The American neurologist Norman Geschwind early in his career became intrigued by the works of Carl Wernicke and he was convinced that the fundamental issue of relating psychological function to brain (areas) should be addressed by adopting Wernicke's approach, by looking for cortical centers and connections, or rather, *disconnections* (Geschwind, 1965a,b). Geschwind started with the psychologist Harold Goodglass at the Boston Aphasia Unit. This unit attracted the interest of many physicians and (neuro)psychologists from all over the world. This was an important factor for the development of neuropsychology in general and a specific interest in neurolinguistics as well.

Acquired Reading and Writing Disorders

Early language models, as used by, for instance, Broca and Bastian, assumed independent memory stores for word images for separate input and output modalities: for hearing and seeing a word, for speaking and writing a word. Attempts have been made to localize these centers, in particular by examining the occurrence of specific reading and writing disorders. Originally, physicians used the term *alexia* and *agraphia*; *dyslexia* (introduced by the German physician Rudolph Berlin in 1887) and *dysgraphia*, first used to refer to a *disturbance* (or a developmental problem) rather than a complete *loss* of function, have replaced these terms. Also the term *word blindness* was used frequently. In the nineteenth century, this study was hampered by the fact that reading and writing were not common in the general population. Starting with the case description of Johann Schmidt of patient with a reading problem in 1676 (see above), various other case descriptions appeared in the literature, also in the nineteenth century (Benson and Cummings, 1985). Jules Joseph Dejerine (1849–1917; see **Figure 12**) published reports of two patients, a man with alexia and agraphia and *monsieur C.*, showing alexia *without* agraphia and no difficulties with spontaneous speech (Dejerine, 1891, 1892; Bub et al., 1993 for a detailed analysis). Subsequently, Dejerine described a third form of alexia: in this type, what Dejerine called the language zone (a more general language function, encompassing the more specialized areas of Broca and Wernicke) is disrupted, resulting in a disturbance of *inner speech*. Postmortem studies indicated that the first form is associated with a lesion in the area of the angular gyrus (the store of visual word forms, necessary for reading and writing), the second form with a lesion in pathways from the special word form area to the general language area that separates it from the angular gyrus. And the third form results from cortical lesions in the language zone. Dejerine's interpretation was quickly adopted by others.

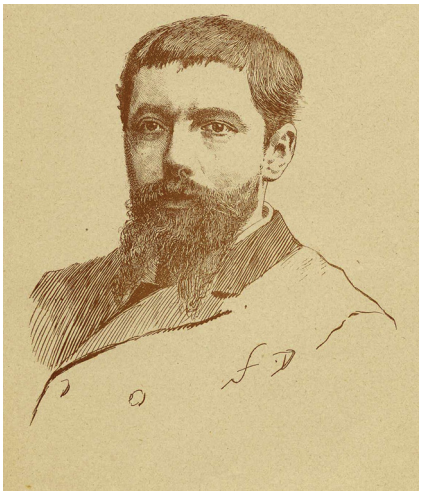

Figure 12 Jules Dejerine.

The study of reading disorders became particularly of particular interest in the context of learning problems in children without obvious neurological disorders, also referred to as *developmental dyslexia*. New interest in acquired reading disorders was aroused by the analysis of reading errors of neurological patients by John Marshall (1939–2007) and Freda Newcombe (1925–2001) (Marshall and Newcombe, 1966, 1973). Discovering different patterns of errors, they argued that reading of words may follow different routes in the cognitive system: the visual route and the phonological route. Borrowing word recognition models from psycholinguistic studies, they demonstrated the usefulness of this way of modeling functional disorders (e.g., Coltheart et al., 1980; Patterson et al., 1985) and this became a strong impetus for the rise of cognitive neuropsychology in the 1980s. In essence, these models were similar to the classical diagrams, although more elaborated and with a better empirical foundation.

Louis-Victor Marcé (1828–64) provided clinical case descriptions of patients with dissociated patterns of spoken and writing language disorders in 1856. The first description of a case of pure agraphia is from Albert Pitres (1848–1928) (Pitres, 1884; Lorch and Barriere, 2003). The localization of a writing center is commonly ascribed to Exner (1881), but actually he did not have proper empirical basis for that and only speculated about this possible localization (Roux et al., 2010). Almost all aphasic patients have some degree of agraphia and it is usually a mixture of language and nonlanguage factors (Benson and Cummings, 1985) and since pure agraphia is very rare, the phenomenon is hard to analyze.

Related Areas

Laterality

When Broca convinced his colleagues in the early 1860s that the principle of localization of function might be correct at least for some functions, he also introduced another quite revolutionary idea. He suggested that the faculty of articulated speech was restricted to the left hemisphere. Immediately, Gustav Dax claimed priority for this observation for his father, who presumably gave a lecture at a conference in Montpellier in 1836 (Finger and Roe, 1999). Another fact, suggesting an asymmetry was hand preference. Broca explicitly denied a connection between these two phenomena, but the two were almost immediately linked to each other. The general idea (formulated by Broca) was that the left hemisphere had a little advantage in development because it received blood with somewhat more oxygen. This results in a dominance of the left hemisphere and this theory was referred to as the cerebral dominance theory. According to this theory, the two hemispheres are basically identical but the left develops quicker and becomes the dominant hemisphere in which language function emerges (Harris, 1991).

The cerebral dominance theory was replaced by the hemispheric specialization theory in 1960. It was the result of a research program, undertaken by Roger Sperry, Joe Bogen, Michael Gazzaniga, and collaborators, on the consequences of the split-brain operation technique for patients with intractable epilepsy. Special experimental paradigms, for instance, the visual half-field presentation paradigm and the dichotic listening test, revealed that the right hemisphere had its own special functions. Moreover, the right hemisphere appeared to be active in language processes, although it appeared to lack grammatical functions.

Hemispherectomy

Hemispherectomy, the removal of one hemisphere under the assumption that functioning remained possible with the remaining parts of the brain, was introduced in 1928 by the American neurosurgeon Walter Dandy (Goodman and Whitaker, 1985). It has been used for children with brain tumors and severe forms of epilepsy. According to the cerebral dominance theory, specialization develops only later in life. Thus, it should be possible for the remaining hemisphere, even the right one, to acquire language. Originally, cognitive functioning after hemispherectomy was surprisingly good and language functioning at first glance seemed intact. However, more detailed studies have shown that the right hemisphere can acquire language only to some extent. Subtle syntactic deficiencies remain (Dennis and Whitaker, 1976).

Brain Stimulation

Yet another surgical procedure, used for the treatment of epilepsy, has revealed important information with respect to language representation in the brain. Focal sources of epileptic attacks often occur in the temporal lobe, close to areas that are involved in language processing. Removal of tissue to stop these attacks may be effective but the loss of language function should be prevented. Cortical stimulation is a technique, applied during brain surgery, to localize the area responsible for the epilepsy and also for exploring which areas are important for language (see Whitaker, 1998b).

Penfield and Roberts (1959) have recorded their extensive experiences with this procedure (see **Figure 13**). Similar studies were performed by Ojemann and Whitaker (1978). While highly discrete areas were found, indicative of specific language centers, a rather large variability of localization of these centers was observed among individuals.

Wada Test

Before a surgery is performed in severe cases with intractable epilepsia, a *Wada test* is performed to determine which hemisphere is primarily responsible for language. In this test, named after the Japanese–Canadian neurosurgeon Juhn Wada, each hemisphere is deactivated for a short period, that is, some minutes: language functions remain intact if the nondominant hemisphere is activated (Wada, 1997; Snyder and Harris, 1997; see **Figure 14**). The general picture arising from studies with this procedure is that language is localized primarily in the left hemisphere, even for left-handed and ambidextrous individuals. However, language functions are examined in a rather general manner and the procedure is not used for analysis of specific linguistic processes.

Figure 13 Wilder Penfield.

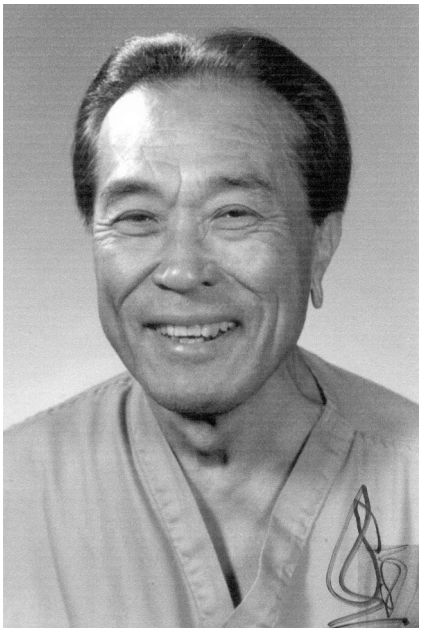

Figure 14 Juhn Wada.

Language and the Right Hemisphere

The various areas of research mentioned above have played a crucial role in our understanding of the complementary roles of the two hemispheres in language processes. Language disorders typically result from lesions in the left hemisphere, but subtle effects have been observed following right hemisphere lesions. Hughlings-Jackson (1864) agreed with Broca that speech was specifically associated with the left hemisphere but argued that the right hemisphere played a significant role in language as well (which is not in contradiction with Broca's position!). Importantly, Hughlings-Jackson (1874) claimed that the basic function was to propositionize, to express a message; this could be in the form of a spoken statement but it could also be expressed in the form of gesture and mime. The relation between verbal and nonverbal communication is an old one in the aphasia literature (Feyereisen and Seron, 1982a,b; Lindell, 2006). Prosody, but also aspects such as metaphorical language and the understanding of jokes, and even sign language (Corina, 1998) play a role in this discussion. Hughlings-Jackson (1878) noted that some aphasic patients could speak when they became emotional, frustrated. He also observed that speech was flat, lacked emotional tone following right hemisphere lesions. The study of the role of right hemisphere in language was specifically triggered by a paper by Monrad-Krohn (1947) who argued that the right hemisphere plays a role in emotional tone or prosody of speech. This view was triggered by analyzing the speech of a woman who recovered from aphasia but her-speech sounded like that of a German speaker. A short, similar observation had been made by Pierre Marie in 1907 and Whitaker (1982) introduced the notion of foreign accent syndrome for this clinical picture. In later years, a distinction was made between different functions of prosody: more linguistic functions (meaning, sentence structure) and emotional functions with a larger role of the right hemisphere in the latter functions (Ross and Monnot, 2008).

Bilingual Brain

Many people are able to communicate in more than one language. The study of how the second (and other) languages are represented in the brain and are affected by brain lesions is a relatively young branch of neurolinguistics. Yet, a number of case descriptions have appeared in the literature. An early paper was by Lordat (1843), describing a priest who became aphasic, lost the ability to speak French, but was able to communicate in a dialect. Michel Paradis has collected and translated these case reports published before 1935 (Paradis, 1983). The general picture arising from such aphasiological studies is that all kinds of patterns may occur: an earlier language may be lost while a later learned language may remain relatively intact and vice versa, and this also holds for patterns of recovery. Importantly, dissociations occur, suggesting that languages do not rely on the same underlying mechanisms. The representation of multiple languages in the brain can also be studied in healthy individuals. Albert and Obler (1978) provided the first overview of psychological, neuropsychological (e.g., using dichotic listening and visual half-field paradigms), and neurolinguistic aspects of bilingualism and concluded that language is represented more bilaterally in the bilingual brain than in the monolingual brain.

Modern Times

The study of language and the brain changed dramatically with the introduction of neuroimaging techniques. The techniques to represent anatomical structures as well as physiological processes improved very quickly. Extent of lesions and neural activity could be examined with great precision and in healthy subjects! However, these developments fall outside the scope of this historical overview.

See also: Aphasia; Bilingual Aphasia; Dyslexia, Developmental; Dyslexias, Acquired and Agraphia; Interhemispheric Interaction in Language and Cognitive Processes; Lateralization of Language as Demonstrated by Brain Imaging Procedures.

Bibliography

Akhutina, T.V., 2003. L.S. Vygotsky and A.R. Luria. Foundations of neuropsychology. Journal of Russian and East European Psychology 41, 159–190.

Albert, M., Obler, L., 1978. The Bilingual Brain: Neuropsychological and Neurolinguistic Aspects of Bilingualism. Academic Press, New York.

Bastian, H.C., 1869. On the various forms of loss of speech in cerebral disease. British and Foreign Med-Chirurg Review 43, 209–236, 470–492.

Benson, D.F., Cummings, J.L., 1985. Agraphia. In: Frederiks, J.A.M. (Ed.), Handbook of Clinical Neurology, Clinical Neuropsychology, vol. 45. Elsevier, Amsterdam.

Benton, A., 2000. Exploring the History of Neuropsychology. Selected papers. Oxford University Press, New York.

Benton, A., Joynt, R.J., 1960. Early descriptions of aphasia. Archives of Neurology 3, 205–222.

Berlin, R., 1887. Eine besondere Art von Wortblindheit (Dyslexie) (A Special Kind of Word Blindness (Dyslexia)). Verlag von J.F. Bergmann, Wiesbaden.

Blumstein, S., 1973. A Phonological Investigation of Aphasic Speech. Mouton, The Hague.

Boeckx, C., Grohmann, K., 2007. The biolinguistics manifesto. Biolinguistics 1, 1–8.
Boller, F., 1978. Comprehension disorders in aphasia: a historical review. Brain and Language 5, 149–165.
Bouillaud, J.-B., 1825. Recherches cliniques propres à démontrer que la perte de la parole correspond à la lésion des lobules antérieurs du cerveau, et à confirmer l'opinion de M. Gall, sur le siège de l'organe du langage articulé. Archives of General Medicine 3 (8), 24–45.
Bouillaud, J.-B., 1848. Recherches cliniques propres à démontrer que le sens du langage articulé et le principe coordinateur des mouvements de la parole résident dans les lobules antérieurs du cerveau. Bulletin de l'Académie nationale de médecine, 1er trimestre, 699–719.
Breasted, J.H., 1930. The Edwin Smith Surgical Papyrus. University Chicago Press, Chicago.
Broca, P., 1861. Perte de la parole, ramollissement chronique de destruction partielle du lobe antérieur gauche du cerveau. Bulletins et Mémoires de la Société d'anthropologie de Paris, 235–238.
Bub, D., Arguin, M., Lecours, A.R., 1993. Jules Dejerine and his interpretation of pure alexia. Brain and Language 45, 531–559.
Buckingham, H., 2007. Alexander Romanovich Luria: the man and his science. In: Koertge, Noretta (Ed.), New Dictionary of Scientific Biography. Charles Schribner's Sons Thomson Gale, Farmington Hills.
Bühler, K., 1934. Sprachtheorie. Fischer, Stuttgart.
Caplan, D., 1987. Neurolinguistics and Linguistic Aphasiology: An Introduction. Cambridge University Press, Cambridge.
Chomsky, N., 1959. A review of B.F. Skinner's verbal behavior. Language 351, 26–58.
Clarke, E., Dewhurst, K., 1996. An Illustrated History of Brain Function. Imaging the Brain from Antiquity to the Present, second ed. Norman Publishing, San Francisco.
Coltheart, M., Patterson, K., Marshall, J.C. (Eds.), 1980. Deep Dyslexia. Routledge & Kegan Paul, London.
Corina, D., 1998. The processing of sign language: evidence from aphasia. In: Stemmer, B., Whitaker, H.A. (Eds.), Handbook of Neurolinguistics. Academic Press, New York, pp. 313–329.
Crichton, A., 1798. An Inquiry into the Nature and Origin of Mental Derangement, Comprehending a Concise System of the Physiology and Pathology of the Human Mind and a History of Passions and Their Effects. Cadell & Davies, London.
De Bleser, R., 1987. From agrammatism to paragrammatism: German aphasiological traditions and grammatical disturbances. Cognitive Neuropsychology 4, 187–256.
De Vigo, G., 1514. Practica Copiosa in Arte Chirurgia. In: The Most Excellent Works of Chirurgie (Bartholomew Traheron, Trans.). London, p. 1571.
Dejerine, J., 1891. Sur un cas de cécité verbale avec agraphie, suivi d'autopsie. Comptes Rendus des Séances et Mémoires de la Société de Biologie 43, 197–201.
Dejerine, J., 1892. Contribution a l'étude anatomo-pathologique et clinique des differentes varietés de cécité verbale. Memoires de la Societé de Biologie 4, 61–90.
Dennis, M., Whitaker, H.A., 1976. Language acquisition following hemidecortication: linguistic superiority of the left over the right hemisphere. Brain and Language 3, 404–433.
Dingwall, W.O., Whitaker, H.A., 1974. Neurolinguistics. In: Annual Review of Anthropology, vol. 3. Annual Reviews, Inc, Palo Alto, pp. 323–356.
Domanski, C.W., 2013. Mysterious "Monsieur Leborgne": the mystery of the famous patient in the history of neuropsychology is explained. Journal of the History of the Neurosciences 22 (1), 47–52.
Duffy, R., Liles, B., 1979. A translation of Finkelnburg's 1870 lecture on aphasia as "asymbolia" with commentary. Journal of Speech and Hearing Disorders 44, 156–168.
Eggert, G.H., 1977. Wernicke's Works on Aphasia: A Sourcebook and Review. Mouton, The Hague.
Eling, P., 1984. Broca on the relation between handedness and cerebral speech dominance. Brain and Language 22, 158–159.
Eling, P., 2006. The psycholinguistic approach to aphasia of Chajim Steinthal. Aphasiology 20, 1072–1084.
Eling, P., 2012. Neurognostic answer. Journal of the History of the Neurosciences 21, 119–125.
Exner, S., 1881. Untersuchungen über die Localisation der Functionen in der Groshirnrinde des Menschen (Studies on the Localization of Functions in the Cerebral Cortex). Braumüller, Wien.
Feinsod, M., Aharon-Peretz, J., 1994. Baron Larrey's description of traumatic aphasia. Journal of the History of the Neurosciences 3, 45–52.
Feyereisen, P., Seron, X., 1982a. Nonverbal communication and aphasia: a review. I. Comprehension. Brain and Language 16, 191–212.
Feyereisen, P., Seron, X., 1982b. Nonverbal communication and aphasia: a review. II. Expression. Brain and Language 16, 213–236.
Finger, S., Roe, D., 1999. Does Gustave Dax deserve to be forgotten? The temporal lobe theory and other contributions of an overlooked figure in the history of language and cerebral dominance. Brain and Language 69, 16–30.
Finkelnburg, F.C., 1870. Bericht: Niederrheinische Gesellschaft In Bonn. Medizinische Sektion. Sitzung vom 21. Marz 1870. Berl Klin Wochenschr 7, 449–450, 460–462.
Flourens, P., 1842. Examen de la phrénologie. Baillière, Paris.
Friedrich, J., 2006. Psychopathology and the essence of language: the interpretation of aphasia by Kurt Goldstein and Roman Jakobson. History of Psychiatry 17 (4), 419–436.
Gall, F.J., 1822–1825. Sur les fonctions du cerveau et sur celles de chacune de ses parties, 6 Bandes. Ballière, Paris.
Geschwind, N., 1965a. Disconnexion syndromes in animals and man. Brain 88, 237–294.
Geschwind, N., 1965b. Disconnexion syndromes in animals and man. Brain 88, 585–644.
Gesner, J., 1770. Samlung von Beobachtungen aus der Arzneygelahrtheit. Zweyter Band. Neue verbesserte Auflage. Beck, Nordlingen.
Goodman, R., Whitaker, H.A., 1985. Hemispherectomy: a review (1928–1981) with special reference to the linguistic abilities and disabilities of the residual right hemisphere. In: Best, C. (Ed.), Hemispheric Function and Collaboration in the Child. Academic Press, New York, pp. 121–155.
Gratiolet, L.P., 1854. Mémoire sur les plis cérébraux de l'homme et des primates. A. Bertrand, Paris.
Harris, L.J., 1991. Cerebral control for speech in right-handers and left-handers: an analysis of the views of Paul Broca, his contemporaries, and his successors. Brain and Language 40, 1–50.
Head, H., 1926. Aphasia and Kindred Disorders of Speech, 2 vols. Macmillan Company, Cambridge.
Howard, D., 1985. Introduction to "on agrammatism" ueber agrammatismus, by Max Isserlin, 1922. Cognitive Neuropsychology 2, 303–307.
Jackson, J.H., 1864. Hemiplegia on the right side, with loss of speech. British Medical Journal 2, 572–573.
Jackson, J.H., 1874. On the nature of the duality of the brain. In: Taylor, J. (Ed.), (1958) Selected Writings of John Hughlings-Jackson, vol. 2. Basic Books, New York, pp. 129–145.
Jackson, J.H., 1878. On affections of speech from disease of the brain. In: Taylor, J. (Ed.), (1958) Selected Writings of John Hughlings Jackson, Evolution and Dissolution of the Nervous System, Speech, Various Papers, Addresses and Lectures, vol. 2. Basic Books, New York, pp. 155–204.
Jacyna, S., 2005. Starting anew: Henry Head's contribution to aphasia studies. Journal of Neurolinguistics 18, 327–336.
Jakobson, R., 1941. Kindersprache, Aphasie and allgemeine Lautgesetze. Universitets Arsskrift, Uppsala.
Kussmaul, A., 1877. Die Störungen der Sprache. Versuch einer Pathologie der Sprache. Vogel, Leipzig.
Larrey, D.J., 1814. Memoirs of Military Surgery and Campaigns of the French Armies, Classics of Surgery Library, 1985. reprint of Joseph Cushing.
Lecours, A.R., Chain, F., Poncet, M., Nespoulous, J.L., Joanette, Y., 1992. Paris 1908: the hot summer of aphasiology or a season in the life of a chair. Brain and Language 42, 105–152.
Lenneberg, E., 1960. Review of W. Penfield and L. Roberts: speech and brain mechanisms. Language 36, 97–112.
Levelt, W., 2012. A History of Psycholinguistics the Pre-Chomskyan Era. Oxford University Press, Oxford.
Lindell, A., 2006. In your right mind: right hemisphere contributions to language processing and production. Neuropsychology Review 16, 131–148.
Lorch, M., 2013. Examining language functions: a reassessment of Bastian's contribution to aphasia assessment. Brain 136, 2629–2637.
Lorch, M., Barriere, I., 2003. The history of written language disorders: reexamining pitres case 1884 of pure agraphia. Brain and Language 85, 271–279.
Lordat, J., 1843. Analyse de la parole pour servir à la theorie de divers cas d'Alalie et de Paralalie de mutisme et d'imperfection du parler que les Nosologistes ont mal connus. Journal de la Societe de Médecine-Pratique de Montpellier 7, 333–353, 417–433.
Luria, A.R., 1973. The Working Brain. An Introduction to Neuropsychology. Penguin Press, London.
Luria, A.R., 1976. Basic Problems of Neurolinguistics. De Gruyter, The Hague.
Luzzatti, C., 2002. Johann August Philipp Gesner 1738–1801. A review of his essay "The language amnesia" in the bicentennial anniversary of his death. Journal of the History of the Neurosciences 11, 29–34.
Luzzatti, C., Whitaker, H.A., 1995/1996. Johannes Schenck and Johannes Jakob Wepfner: clinical and anatomical observations on the prehistory of neurolinguistics and neuropsychology. Journal of Neurolinguistics 9, 157–164.
Marce, L.V., 1856. Memoire sur quelques observations de physiologie pathologique tendant a demontrer lexistence dun principe coordinateur de lecriture et ses rapports avec le principe coordinateur de la parole. Compte-rendu de la Socieetee de Biologie, Paris 3, 93–115.
Marshall, J.C., Newcombe, F., 1966. Syntactic and semantic errors in paralexia. Neuropsychologia 4, 169–176.

Marshall, J.C., Newcombe, F., 1973. Patterns of paralexia: a psycholinguistic approach. Journal of Psycholinguistic Research 2, 175–199.
McKeon, R., October 1946. Aristotle's conception of language and the arts of language. Classical Philology 41 (4), 193–206. January 1947, 42 (1), 21–50.
Monrad-Krohn, G.H., 1947. Dysprosody or altered 'melody of language'. Brain 70, 405–415.
Morgagni, G.B., 1762. The Seats and Causes of Diseases Investigated by Anatomy (Translation by B. Alexander, 1769 from: De sedibus et causis morborum per anatomen indagatis. Facsimile edition 1960. New York Academy of Medicine). Hafner Publishing, New York.
O'Neill, Y.V., 1980. Speech and Speech Disorders in Western Thought before 1600. Greenwood Press, London.
Ojemann, G., Whitaker, H.A., 1978. The bilingual brain. Archives of Neurology 35, 409–412.
Paradis, M., 1983. Readings on Aphasia in Bilinguals and Polyglots. Marcel Didier, Montreal.
Patterson, K.E., Marshall, J.C., Coltheart, M. (Eds.), 1985. Surface Dyslexia: Neuropsychological and Cognitive Studies of Phonological Reading. Lawrence Erlbaum Associates, London.
Penfield, W., Roberts, L., 1959. Speech and Brain Mechanisms. Princeton, New Jersey.
Pick, A., 1898. Beiträge zur Pathologie und pathologischen Anatomie des Centralnervensystems, mit Bemerkungen zur normalen Anatomie desselben. Karger, Berlin.
Pick, A., 1931. Aphasie. In: Bumke, O., Foerster, O. (Eds.), Handbuch der normalen un patholoqischen physioloqie, vol. 15. Springer, Berlin, pp. 1416–1524.
Pitres, A., 1884. Considérations sur l'Agraphie. Alcan, Paris.
Rommel, P., 1683. De Aphonia Rara. Miscellanea Curiosa Medico-Physica Academiae Naturae Curiosum 2 ser 2, 222–227.
Ross, E., Monnot, M., 2008. Neurology of affective prosody and its functionañ-anatomic organization in right hemisphere. Brain and Language 104, 51–74.
Roux, F.-E., Draper, L., Köpke, B., Démonet, J.-F., 2010. Who actually read exner? Returning to the source of the frontal "writing centre" hypothesis. Cortex 46 (9), 1204–1210.
Schmidt, J., 1676. Miscellanea curiosa medico-physica academiae naturae curiosum t. IV.
Snyder, P.J., Harris, L.J., 1997. The intracarotid amobarbital procedure: an historical perspective. Brain and Cognition 33, 18–32.
Sollier, P., 1892. Les troubles de la memoire. J. Rueff, Paris.
Steinthal, Ch, 1871. Einleitung in die Psychologie und Sprachwissenschaft. Dümmler's Verlagsbuchhandlung, Berlin.
Stemmer, B., Whitaker, H.A., 1998. Handbook of Neurolinguistics. Academic Press, San Diego.
Trager, E.C., 1960. The field of neurolinguistics. Studies in Linguistics 15, 70–71.
Trousseau, A., 1864. De l'aphasie, maladie decrite récemment sous le nom impropre d'aphémie. Gaz Hop civ milit 37, 13–14, 25–26, 37–39, 48–50.
Wada, J., 1997. Youthful season revisited. Brain and Cognition 33, 7–10.
Wepfer, J.J., 1727. Observationes medico-practicae, de affectionibus capitis internis et externis. Joh. Adam Ziegler, Schaffhausen.
Wernicke, C., 1874. Der Aphasische Symptomencomplex. Eine psychologische Studie auf anatomischer. Basis. Cohn & Weigert, Breslau.
Whitaker, H.A., 1982. Levels of impairment in disorders of speech. In: Malatesha, R.N., Hartlage, L.C. (Eds.), Neuropsychology and Cognition. Nijhoff, The Hague, pp. 168–207.
Whitaker, H.A., 1998a. Neurolinguistics from the middle ages to the pre-modern era: historical vignettes. In: Stemmer, B., Whitaker, H.A. (Eds.), Handbook of Neurolinguistics. Academic Press, San Diego, pp. 27–54.
Whitaker, H.A., 1998b. Electrical stimulation mapping of language cortex. In: Stemmer, B., Whitaker, H.A. (Eds.), The Handbook of Neurolinguistics. Academic Press, Orlando, pp. 125–130.
Wollock, J., 1997. The Noblest Animate Motion Speech, Physiology and Medicine in Pre-Cartesian Linguistic Thought. Benjamins, Amsterdam.
York, G., 2009. Localization of language function in the twentieth century. Journal of the History of the Neurosciences 18, 283–290.

Neurological Approaches to Agrammatism

Dimitrios S Kasselimis, University of Crete, Heraklion, Greece; and National and Kapodistrian University of Athens, Athens, Greece
Maria Varkanitsa, University College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The present article investigates agrammatism from a neurological/neuropsychological perspective. Agrammatism was originally defined as 'telegraphic' speech output characterizing nonfluent types of aphasia. During the last five decades, it has been well documented that agrammatic patients frequently present impaired comprehension of syntactically complex sentences, a phenomenon known as 'asyntactic comprehension.' Although agrammatism was traditionally linked to frontal lesion loci, such as Broca's area, contemporary lesion and brain-imaging studies depict a widely distributed neural network, consisting of frontal, temporal, and parietal cortical regions, associated with syntactic processing and agrammatic performance. Possible relationships between syntactic comprehension and working memory are also discussed.

Introduction

Agrammatism is a language deficit, usually accompanying post-stroke aphasic disturbances (Agrammatism is also observed in neurodegenerative diseases, like the nonfluent variant of primary progressive aphasia (Gorno-Tempini et al., 2004, 2011). However, such syntactic deficits are related to a different neuropathological substrate and are beyond the scope of this article.), which affects processing of sentences, while word-level processing is relatively well preserved. Sentence processing difficulties are common both in speech production and comprehension. Agrammatic patients have difficulties in understanding grammatically complex sentences and tend to omit function words (such as prepositions and determiners), thus using a 'telegraphic speech.' It is usually seen in patients with Broca's aphasia, in the context of nonfluent, dysprosodic speech.

The first case of a patient suffering from Broca's aphasia with agrammatism was published decades before the introduction of both terms in *The Phrenological Journal and Miscellany*. In 1824, Alexander Hood described the case of a gardener, presented with acute right hemiplegia and language disturbances after a stroke in the left hemisphere. The patient's profile is compatible with the contemporary clinical description of Broca's aphasic syndrome. Interestingly, Hood's patient is reported to speak telegraphically, using short, simple sentences, like 'something about it' (that was his answer when asked by the doctor to describe what happened to him). Even if Hood explained the case in the context of Phrenology, today – if tested in a modern neurological clinic – this patient could be diagnosed as a patient with Broca's aphasia and signs of agrammatism. Research on agrammatic speech per se was first conducted in the early 1900s on German-speaking patients, and subsequent studies with English-speaking patients with aphasia established the hallmark of the deficit to be omission of function words (Miceli, 2003). This is still considered to be one of the fundamental characteristics of agrammatism in contemporary clinical practice. Further support on this notion has been provided by analyses of agrammatic patients' connected speech (e.g., Rochon et al., 2000; Saffran et al., 1989) showing that the main symptom of agrammatic speech is the use of ungrammatical utterances due to omission (or substitution, depending on the language) of grammatical morphemes, both bound and free, such as verb inflection, determiners, and prepositions. Below is a speech sample of a chronic aphasic patient which was assessed in our clinic:

> […] much drink, eh? No, drink, much drink, eh, and then, ah, eh, stress, that's why I can't talk now, stress and I can't talk, after better now, but I stress and I can't talk […].
>
> translated from Greek

The patient (FS) is a male 49-year-old worker, with 9 years of formal schooling. He suffered a stroke 8 years ago, resulting in acute right hemiplegia and severe nonfluent aphasia. The brain lesion included the middle and inferior frontal gyrus, the precentral gyrus, the supplementary motor area, and the insula (Figure 1). During recovery, he regained his motor and every-day communicative abilities to an adequate degree, but the agrammatic pattern remained. As can be seen from the above sample, his speech completely

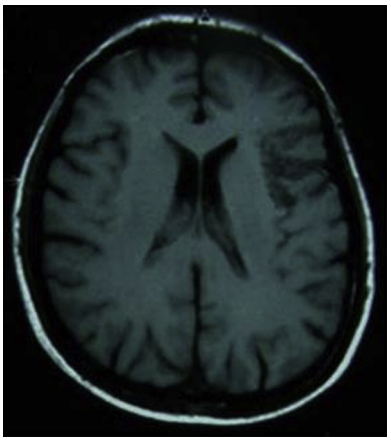

Figure 1 FS's lesion.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54022-7

lacks morphosyntactic structure. Verbs and pronouns were omitted and the meaning is almost exclusively conveyed by content words. A similar pattern was observed in his writing. FS could sufficiently write single words by dictation, but failed to form grammatically correct sentences when asked to describe an image within a short paragraph. Finally, he had difficulty in understanding syntactically complex sentences, such as *wh*-questions and passive sentences.

Based on the above clinical example, one could claim that agrammatism affects both production and comprehension. Since the seminal paper by Caramazza and Zurif (1976) it has been well documented that agrammatic patients have difficulties comprehending syntactically complex sentences (Such deficits are not generic, in the sense that comprehension is not globally compromised. When assessing a patient with impaired comprehension, a clinical neuropsychologist should be able to differentiate between impairment of syntactic abilities and other processing deficits. For example, patients with conduction aphasia often show impaired comprehension, but this can be attributed to a phonological WM deficit, while syntactic abilities remain intact (Gvion and Friedmann, 2012).), that is semantically reversible sentences that involve movement out of the object position, such as sentences presented in (1) below.

(1) a. [***Which girl***]$_1$ did the boy kiss t$_1$?

b. [***The girl***]$_1$ that the boy kissed t$_1$ was tall

c. It is [***the girl***]$_1$ that the boy kissed t$_1$

(2) a. [***Which boy***]$_1$ t$_1$ kissed the girl?

b. [***The boy***]$_1$ that t$_1$ kissed the girl was tall

c. It is [***the boy***]$_1$ that t$_1$ kissed the girl

Several cross linguistic studies, using sentence-picture matching, truth-value judgment, or/and grammaticality judgment tasks, have shown that patients' performance on these sentences is at chance level; patients are unable to correctly assign thematic roles to the arguments in half of the cases. When processing sentences like (1), for instance, patients are unable to understand who is doing the kissing. On the other hand, when they process semantically reversible sentences that involve movement out of the subject position, such as sentences in (2), their performance is above chance level. This dissociation is known as 'asyntactic comprehension' and has attracted most of the recent research in aphasia.

A number of linguistic theory-driven accounts have been put forward to explain agrammatic speech production and comprehension. Some researchers have suggested that language impairments in agrammatism stem from loss of core linguistic knowledge (representational accounts), whereas others have suggested that deficits in language processing is due to reduced working memory capacity in these patients (processing accounts). In this article the focus will be on agrammatism as a processing deficit within a neurological/neuropsychological framework. Representational accounts are discussed in Grodzinsky (1990, 2000), Friedmann and Grodzinsky (1997), and Wenzlaff and Clahsen (2004) among many others.

The Neurological Correlates of Agrammatism: Evidence from Lesion Studies

Despite the fact that detailed descriptions of lesion loci can be very informative, several papers on agrammatism do not report patients' lesion sites. Other studies give us a general idea of lesion sites, providing just a vague description of patients' lesion loci (e.g., Hanne et al., 2011). Especially for studies with small sample size, individual data such as lesion locus is considered essential in order for the reader to have an idea with regard to the homogeneity of the experimental group. Moreover, reporting brain lesion information in agrammatism research papers could provide useful insight on the possible brain-function relationship, when it comes to syntactic processing.

From studies that include lesion data it could be argued that it is usually patients with Broca's aphasia who demonstrate syntactic deficits (Ardila, 2012; Basso, 2003; Benson and Ardila, 1996; Caplan, 2003a; Goodglass, 1993; Kertesz, 1985; Marshall, 2011). This observation provided support for a localizationist view, according to which, different aspects of syntactic comprehension are localized in distinct areas of the left perisylvian cortex (see, for example, the Trace Deletion Hypothesis by Grodzinsky, 1990). It should be although noted that the lesion of such patients is never restricted to the foot of the third frontal convolution. Even the famous Monsieur Tan, the patient originally described by Pierre Paul Broca, had a rather large brain lesion that affected various cortical and subcortical areas (Broca himself acknowledged that the lesion was not restricted to the posterior part of the inferior frontal gyrus, although he claimed that Leborgne's aphemia was due to the destruction of the particular area (Broca, 1861).) (Dronkers et al., 2007; Signoret et al., 1984). Clinical research has clearly shown that a lesion restricted to BA44 is not sufficient to produce a Broca's aphasic syndrome (Mohr et al., 1978). Furthermore, there have been reported patients with Broca's aphasia, whose lesion spared the inferior frontal gyrus (Nespoulous et al., 1988; Basso et al., 1985; Kasselimis et al., 2012). It has also been claimed that syntactic abilities are not likely to be localized exclusively in Broca's area, even though higher level syntactic processes may be supported by the anterior perisylvian areas (Grodzinsky, 2000). In addition, it has been recently argued that lesion-analysis techniques used in studies supporting the localizationst framework are inadequate (Caplan et al., 2007). Apart from that, as mentioned above, there are several studies involving agrammatic patients that do not report lesion data; based on contemporary aphasia literature, one cannot simply assume that a nonfluent patient's lesion will include the inferior frontal gyrus. Moreover, syntactic deficits are present in other than Broca's aphasic syndromes (Grodzinsky and Finkel, 1998; Schwartz et al., 1987; Shapiro et al., 1993; Zurif and Caramazza, 1976; Berndt et al., 1996; Dick et al., 2001; Caplan, 1987a; Caplan et al., 1985). Finally, as it will be discussed below, agrammatic phenomena occur after various lesion sites throughout the perisylvian cortex.

Even if there isn't a distinct brain region specialized for syntax, several studies have pointed to specific areas of the left hemisphere which are involved in syntactic processing. Damage to Broca's area often result in agrammatic disturbances (Basso, 2003; Biran and Friedmann, 2012; Caplan, 2003b; Carreiras et al., 2012; Grodzinsky, 2000). However, lesions producing

syntactic deficits are not restricted to the inferior frontal gyrus. Several studies report agrammatic patients with lesions in the insular cortex, in temporoparietal and frontotemporal areas, as well as in subcortical tissue (Biran and Friedmann, 2012; Borovsky et al., 2007; Christiansen et al., 2010; Thompson et al., 2010; Ruiter et al., 2010; Peristeri et al., 2013; Fyndanis et al., 2010). It should be also noted, that if a post-stroke agrammatic patient's lesion does involve a specific area, such as BA44 or BA45, that does not mean that the syntactic deficit is due to damage in the particular area, nor that the lesioned area supported syntactic processing premorbidly. In a recent lesion study, Newhart et al. (2012) investigated the relationship between various brain areas and sentence comprehension. Their results indicated that asyntactic comprehension is associated with lesion at the angular gyrus, rather than BA44. Borovsky et al. (2007) recently published a paper on lesion correlates of speech production deficits. Although they concluded that lesions in the inferior frontal gyrus were related to impairment of both fluency and complexity, they could not provide robust evidence for a specific syntactic capacity of this particular brain region, since it was not specifically involved in complex speech production over fluency. Dronkers et al. (2004) failed to find significant correlations between lesion in Broca's area and measures of syntactic comprehension. Their results indicate that other brain areas, namely BAs 22, 39, 46, and 47, are crucial for processing highly demanding sentences, such as object-clefts, object- and subject-relative clauses, while the middle temporal gyrus may be involved in word-level comprehension. Research on treatment-induced aphasia recovery is in accordance with the notion that syntactic comprehension can be subserved by several cortical areas. Thompson et al. (2010) recently showed that rehabilitation of agrammatic patients is in parallel with a shift in their brain activation pattern toward angular gyrus and superior temporal lobule, thus indicating that these cortical regions can be recruited for processing complex syntactic stimuli. In their study, Tyler et al. (2011) assessed 14 aphasic patients with a syntactic ambiguity judgment task, and found that the integrity of the pars triangularis (in Brodmann's terms: BA45; the anterior part of the Broca's area) and the posterior part of the middle temporal gyrus is essential for syntactic abilities. The authors concluded that these two areas may be functionally related in the context of a neural network that supports syntactic computations. Caplan et al. (1996) studied 18 left-hemisphere damaged aphasic patients, in order to investigate possible relationships between lesion characteristics (locus and volume) and type or degree of syntactic deficits. Their findings showed that patients with anterior and posterior lesions could not be differentiated based on performance on measures of syntactic comprehension. Moreover, no correlation was found between the severity of syntactic impairment and lesion size. The authors conclude that "the complete picture is consistent with the models proposed by Mesulam (1990) and Damasio (1992), which involve both distributed and localized features" (Caplan et al., 1996, p. 946). More recently, Caplan et al. (2007) conducted a study using far more refined neuroimaging methods and a larger patient sample (42 post-stroke aphasic patients), in order to elucidate lesion-deficit correlations with regard to syntactic processing. According to their results, performance of measures of syntactic ability is dependent on perisylvian, as well as nonperisylvian (such as the anterior part of the inferior temporal cortex and the superior parietal lobule) lesion loci. Furthermore, patients with equivalent task performance had lesions of different sizes, and patients with similar lesion volume were characterized by great variability with regard to their syntactic processing abilities. Based on these data, the authors support a notion originally introduced as the "individual variability model" (1987a,b). According to this proposed framework, sentence comprehension and syntactic processing are not invariantly localized in one cortical area. Alternatively, it is suggested that these functions are related to multiple perisylvian and nonperisylvian areas that are not similar in terms of neither cytoarchitectonic features nor patterns of connectivity, and that such relations may differ among individuals (Caplan et al., 2007).

Overall, lesion studies point to several perisylvian areas, which are considered essential for syntactic processing. Even if agrammatism is considered to be the hallmark of Broca's aphasia, the crucial brain region for a possible syntax breakdown in case of brain damage extends beyond pars triangularis and pars opercularis of the inferior frontal gyrus, and includes several other cortical areas, such as the angular, the supramarginal, and the superior temporal gyri (Caplan, 2003b). The general image emerging from modern neuropsychology research depicts a possible neural network consisting of frontal, temporal, and parietal cortical regions that supports a rather complex linguistic function: syntactic processing.

The Neurological Correlates of Syntactic Processing: Evidence from Imaging Studies

The corpus of lesion studies' literature suggests that agrammatism is usually caused by damage in the left hemisphere (It should be although noted that brain-imaging studies also provide evidence of right-hemisphere activation during syntactic processing in healthy individuals (Just et al., 1996; Moro et al., 2001; Waters et al., 2003).), including but not limited to Broca's area. The fact that agrammatic patients have difficulties processing complex sentences, such as object-relative clauses, object-clefts, and object *wh*-questions, may suggest that Broca's area has a functional role in processing such sentences (see Swinney and Zurif, 1995; Swinney et al., 1996; Zurif et al., 1993). This has been actually supported by imaging studies with nonbrain-damaged adults. Functional brain-imaging studies using a variety of tasks have created a rich picture, which indicates that sentence processing (hence, syntactic processing) taxes Broca's region (Caplan et al., 2000; Just et al., 1996; Friederici et al., 2003; Hashimoto and Sakai, 2002; Heim, 2005; Meltzer et al., 2010; Thothathiri et al., 2012; Tyler et al., 2011).

Nevertheless, the specifics of the relationship between Broca's area and syntactic processing remain highly disputed. Some researchers have claimed that the renowned foot of the third frontal convolution has a general, linguistically nonselective role; it supports a domain-general verbal or syntactic Working Memory (henceforth WM), associated with assigning syntactic structure to a sentence – especially when that

sentence include intrasentential dependencies – and using that structure to determine sentence meaning. Caplan and his colleagues have conducted several PET studies that support this *Generalist* (The term *Generalist* view, as well as *Specifist* view that will appear later, has been introduced by Santi and Grodzinsky (2007).) view (see Caplan et al., 1998, 1999; Caplan and Waters, 1999; Stromswold et al., 1996). Specifically, they have tested normal adults while making plausibility judgments about simple subject-relative clauses (e.g., *The award thrilled the actress that praised the producer*) and more complex object-relative clauses (e.g., *The actress that the award thrilled praised the producer*). Their main finding was that regional cerebral flow (rCBF) was increased in Broca's area (particularly in the pars opercularis) when participants judged the semantic plausibility of the object-relative clauses, which are syntactically more complex, compares to subject-relative clauses, which are syntactically less complex (see also Just et al., 1996). These results have led to the view that increased activation in Broca's area is due to increased syntactic processing (i.e., processing syntactically complex sentences).

Recent fMRI studies have compared processing of complex movement-derived sentences with processing of sentences that are still complex, in terms of the length, but include no movement operation. This contrast has shown that greater activation within Broca's area is associated only with processing complex movement-derived sentences (Ben-Shachar et al., 2003, 2004; Friederici et al., 2006; Grodzinsky, 2000, 2005; Kluender and Kutas, 1993; Santi and Grodzinsky, 2007; Zurif et al., 1993). For instance, Ben-Shachar et al. (2003) designed an fMRI study in which 12 healthy native speakers of Hebrew judged the grammaticality of object-relative clauses (e.g., *I helped the girl that Mary saw in the park*) and sentences with embeddings (e.g., *I told Mary that the girl ran in the park*). Their findings demonstrated a movement-related activation within Broca's area. This has been replicated in a later study (Ben-Shachar et al., 2004), in which they contrasted processing of object *wh*-questions (e.g., *What have you forgotten?*) with processing of yes/no questions that in Hebrew do not require movement (e.g., *Have you forgotten?*). In another fMRI study, Santi and Grodzinsky (2007) compared processing of object-relative clauses with processing of sentences that included a reflexive (e.g., *John knows that Mary pinched herself*). In order to assign meaning to the reflexive (i.e., *herself*), a dependency between the reflexive and its antecedent (i.e., *Mary*) should be created. This dependency is very similar to the dependency between a moved constituent and the position from which it has been moved in movement-derived sentences. Having said that, Santi and Grodzinsky found increased activation within Broca's area only when processing movement-derived sentences. Taken together, these findings suggest that the functional role of Broca's area in WM and sentence processing is more specific (*Specifist* view); Broca's area is entrusted with handling movement.

However, Broca's area is not the only cortical part involved in syntactic processing. Several brain-imaging studies point to a widely distributed network associated with comprehension of syntactically complex sentences within the left-hemisphere cortex, which includes frontal, parietal, and temporal cortices (Ardila, 2012; Ben-Shachar et al., 2003; Caplan, 2001; Caplan et al., 1998, 2001; Cappa, 2012; Crinion et al., 2006; Dronkers et al., 1994; Friederici, 2011; Humphries et al., 2005; Iijima et al., 2009; Just et al., 1996; Kaan and Swaab, 2002; Mazoyer et al., 1993; Moro et al., 2001; Röder et al., 2002; Santi and Grodzinsky, 2007; Stowe et al., 1998; Stromswold et al., 1996). Thompson et al. (2010) recently confirmed previous studies via fMRI tasks in healthy individuals, by demonstrating increased activation of the inferior and middle frontal gyri, the anterior middle and superior temporal gyri, the angular gyrus, and the insula, during processing object-cleft structures. It is further suggested that such syntactic processes modulate the flow of information from the inferior frontal gyrus to the posterior superior temporal cortex (den Ouden et al., 2012). However, alterations in activation patterns of temporal, frontal, and parietal areas are found to be associated also with semantic abilities (Meltzer et al., 2010). To overcome this issue, Moro et al. (2001) conducted a PET experiment which required detection of anomalies in written sentences consisting of pseudowords, thus neutralizing possible confounding effects of the semantic component. Their results confirmed previous findings supporting the involvement of Broca's area in syntactic and morphosyntactic processing.

Friederici and his colleagues have conducted several brain-imaging studies showing an association between highly demanding complex syntactic tasks and activation of frontal and temporal regions, that constitute a left-lateralized frontotemporal language network (Fiebach et al., 2005; Friederici, 2012, 2011; Friederici and Kotz, 2003; Obleser et al., 2011). On the basis of combined brain-imaging and lesion data, they propose a model for processing and integrating semantic and syntactic information in a temporal/spatial framework (Friederici, 2012, 2011; Friederici and Kotz, 2003). According to this model, linguistic stimuli undergo three distinct stages of processing: (1) acoustic/phonological analysis and processes of local structure building, (2) lexical–semantic and thematic processes, and (3) processes of syntactic revision and final integration. Each phase requires engagement of specific cortical areas interconnected by white matter tracts, within a proposed dynamic frontotemporal neural net, which comprises distinct brain regions specialized for analyzing semantic and syntactic information (Friederici et al., 2003).

Working Memory and Agrammatism

Recent studies argue about a language network consisting of two processing streams, one ventral and one dorsal (Hickok and Poeppel, 2004, 2007). It has been shown that the ventral stream connecting temporal and prefrontal areas via the extreme capsule fasciculus (originally described by Petrides and Pandya, 1988; see also Frey et al., 2008) mediates higher-level language comprehension (Saur et al., 2008). Since the time of Wernicke, the left posterior temporoparietal region is traditionally considered to be crucial for comprehension. Modern lesion studies have confirmed the findings of the German physician, pointing out that posterior lesions, including the superior and middle temporal gyrus, and the inferior parietal lobule, are related to comprehension impairment in aphasia (Kreisler et al., 2000; Dronkers et al., 2004; Leff et al., 2009). There are also brain-imaging studies on healthy individuals that support the above clinical data.

Temporal and parietal cortices are shown to be involved in lexical-semantic analysis of spoken words (Price, 2000; Salmelin, 2007). However, contextual comprehension requires an exchange of information between the aforementioned posterior areas and the prefrontal cortex. This exchange is served by the extreme capsule fasciculus (Saur et al., 2008). In sum, with regard to auditory sentence comprehension and syntactic processing, the reviewed studies described in the two former sections of the present paper reveal a dynamic, left-lateralized frontotemporal network, consisting of several cortical regions that possibly serve distinct subprocesses of syntactic comprehension and integration, at different stages (Friederici, 2012, 2011; Friederici et al., 2003; Friederici and Kotz, 2003)

Research on neural substrate of WM points to similar cortical and subcortical regions. With regard to the role of posterior and anterior cortices in the brain circuits responsible for verbal WM, lesion studies suggest a relation between posterior regions – such as the superior temporal gyrus – and WM (Beeson et al., 1993; Leff et al., 2009). However, a *meta*-analysis carried out by Wager and Smith (2003) on brain-imaging studies reports consistent activation of prefrontal areas during verbal WM tasks. Gerton et al. (2004) recently showed that healthy participants' performance of verbal WM tasks, such as digits forward and backward, elicits hemodynamic responses in both frontal and parietal regions. There are also studies demonstrating the involvement of the left hemisphere in nonverbal stimulus processing (Baldo and Dronkers, 2006; Kessels et al., 2002). Ricciardi et al. (2006) claim that a frontoparietal network supports such nonverbal information processing. Overall, it could be the case that both anterior and posterior cortical regions are engaged in verbal and nonverbal WM tasks, as if the central executive required both prefrontal and parietal regions (Collette and Van der Linden, 2002). It is already stated that an analogous network, including similar cortical and subcortical areas of the left hemisphere has been proposed for language (Hickok and Poeppel, 2004, 2007), and more specifically for syntactic processing (Friederici, 2012, 2011). Based on the above, it could be hypothesized that linguistic processes, such as syntactic comprehension, and executive functions, such as WM, may be parallel processes in terms of anatomy and physiology.

Imaging studies with healthy adults have shown that Broca's area hosts WM mechanisms that support processing of syntactically complex sentences. Moreover, when this region is lesioned subjects have difficulties processing such sentences. These findings have led to the idea that language deficits in agrammatism could be explained by appealing to WM (Grodzinsky, 2005). In fact, WM abilities in patients with aphasia have received a great deal of attention in the literature in the last decades. Caplan (2006) interprets the underlying cause of comprehension deficits in aphasics, in terms of impaired processing due to resource reduction. Newhart et al. (2012) have shown that damage to BA44 is associated with deficits with regard to phonological short-term memory, a construct that is considered to be one of the slave systems of WM. There are also several studies demonstrating that patients with aphasia often present with WM deficits in conjunction with their language problems (Burgio and Basso, 1997; Erickson et al., 1996; Kasselimis et al., 2013; Martin and Ayala, 2004; Martin and Saffran, 1997; Ronnberg et al., 1996; Shallice and Warrington, 1970). Moreover, it is argued that agrammatic aphasia is associated with impaired nonlinguistic cognitive sequencing, thus indicating that there are domain-general brain mechanisms subserving both language and sequential learning (Christiansen et al., 2010; Dominey et al., 2003). Brain-imaging data support this notion, by demonstrating engagement of Broca's area in both sentence comprehension and abstract sequence processing tasks (Hoen et al., 2006).

Some researchers have further suggested that WM deficits not only co-occur with language impairments but they may also contribute to them (Christensen and Wright, 2010; Friedmann and Gvion, 2003; Laures-Gore et al., 2010; Martin, 2008; Potagas et al., 2011; Wright et al., 2007). Taking into consideration the above data, we could hypothesize that agrammatic patients may suffer from a general WM deficit. Nevertheless, functional imaging data indicate that agrammatism cannot be easily attributed to a domain-general WM impairment. Recent PET and fMRI studies have shown that increased regional cerebral blood flow in Broca's area is associated with processing syntactically complex structures per se, rather than short-term or WM (Caplan et al., 2000; Christensen et al., 2012; Waters et al., 2003). It could be thus the case, that agrammatic patients suffer from a WM deficit related to specific linguistic phenomena, and Broca's area is associated with a domain-specific WM function (Fiebach et al., 2005; Hashimoto and Sakai, 2002; Stowe et al., 1998). It should be although noted that Broca's area has been also found to be associated with domain-independent processes, such as visuospatial modality syntax and hierarchical structures in several functional domains (Tettamanti et al., 2009; Tettamanti and Weniger, 2006). From an evolutionary perspective, it is highly plausible that language functions in general and syntactic abilities in particular emerged via natural selection, based on physiological preadaptations, such as WM cortical networks, consisting of interconnecting frontal, temporal, and parietal areas (Aboitiz, 1995; Aboitiz and García, 1997; Aboitiz et al., 2006; Lieberman, 1985, 1992).

See also: Aphasia; Evolution and Language: Overview; Functional Brain Imaging of Language processes; Language Processing, Functional Magnetic Resonance Imaging of; Lateralization of Language as Demonstrated by Brain Imaging Procedures; Linguistic Approaches to Agrammatism.

Bibliography

Aboitiz, F., 1995. Working memory networks and the origin of language areas in the human brain. Medical Hypotheses 44 (6), 504–506.

Aboitiz, F., Garcõa, V.,R., 1997. The evolutionary origin of the language areas in the human brain. A neuroanatomical perspective. Brain Research Reviews 25 (3), 381–396.

Aboitiz, F., García, R.R., Bosman, C., Brunetti, E., 2006. Cortical memory mechanisms and language origins. Brain and Language 98 (1), 40–56.

Ardila, A., 2012. Interaction between lexical and grammatical language systems in the brain. Physics of Life Reviews 9 (2), 198–214.

Baldo, J.V., Dronkers, N.F., 2006. The role of inferior parietal and inferior frontal cortex in working memory. Neuropsychology 20, 529–538.

Basso, A., 2003. Aphasia and Its Therapy. Oxford University Press, New York.

Basso, A., Lecours, A.R., Moraschini, S., Vanier, M., 1985. Anatomoclinical correlations of the aphasias as defined through computerized tomography: exceptions. Brain and Language 26 (2), 201–229.

Beeson, P.M., Bayles, K.A., Rubens, A.W., Kaszniak, A.W., 1993. Memory impairment and executive control in individuals with stroke-induced aphasia. Brain and Language 45, 253–275.

Ben-Shachar, M., Hendler, T., Kahn, I., Ben-Bashat, D., Grodzinsky, Y., 2003. The neural reality of syntactic transformations: evidence from functional magnetic resonance imaging. Psychological Science 14 (5), 433–440.

Ben-Shachar, M., Palti, D., Grodzinsky, Y., 2004. Neural correlates of syntactic movement: converging evidence from two fMRI experiments. NeuroImage 21 (4), 1320–1336.

Benson, D.F., Ardila, A., 1996. Aphasia: A Clinical Perspective. Oxford University Press, Oxford.

Berndt, R., Mitchum, C., Haendiges, A., 1996. Comprehension of reversible sentences in "agrammatism": a meta-analysis. Cognition 58, 289–308.

Biran, M., Friedmann, N., 2012. The representation of lexical-syntactic information: evidence from syntactic and lexical retrieval impairments in aphasia. Cortex 48 (9), 1103–1127.

Borovsky, A., Saygin, A.P., Bates, E., Dronkers, N., 2007. Lesion correlates of conversational speech production deficits. Neuropsychologia 45 (11), 2525–2533.

Broca, P., 1861. Perte de la parole, ramollissement chronique de destruction partielle du lobe antrieur gauche du cerveau. Bulletins de la Societe d'Anthropologie de Paris 1861, 235–238.

Burgio, F., Basso, A., 1997. Memory and aphasia. Neuropsychologia 35 (6), 759–766.

Caplan, D., 2006. Aphasic deficits in syntactic processing. Cortex 42 (6), 797–804.

Caplan, D., 2003a. Aphasic syndromes. In: Heilman, K.M., Valenstein, E. (Eds.), Clinical Neuropsychology, fourth ed. Oxford University Press, New York, pp. 14–34.

Caplan, D., 2003b. Syntactic aspects of language disorders. In: Heilman, K.M., Valenstein, E. (Eds.), Clinical Neuropsychology, fourth ed. Oxford University Press, New York, pp. 61–91.

Caplan, D., 2001. Functional neuroimaging studies of syntactic processing. Journal of Psycholinguistic Research 30 (3), 297–320.

Caplan, D., 1987a. Discrimination of normal and aphasic subjects on a test of syntactic comprehension. Neuropsychologia 25, 173–184.

Caplan, D., 1987b. Neurolinguistics and Linguistic Aphasiology. Cambridge University Press, Cambridge.

Caplan, D., Alpert, N., Waters, G., 1998. Effects of syntactic structure and propositional number on patterns of regional cerebral blood flow. Journal of Cognitive Neuroscience 10 (4), 541–552.

Caplan, D., Alpert, N., Waters, G., 1999. PET studies of syntactic processing with auditory sentence presentation. NeuroImage 9 (3), 343–351.

Caplan, D., Alpert, N., Waters, G., Olivieri, A., 2000. Activation of Broca's area by syntactic processing under conditions of concurrent articulation. Human Brain Mapping 9 (2), 65–71.

Caplan, D., Baker, C., Dehaut, F., 1985. Syntactic determinants of sentence comprehension in aphasia. Cognition 21, 117–175.

Caplan, D., Hildebrandt, N., Makris, N., 1996. Location of lesions in stroke patients with deficits in syntactic processing in sentence comprehension. Brain 119 (3), 933–949.

Caplan, D., Vijayan, S., Kuperberg, G., West, C., Waters, G., Greve, D., Dale, A.M., 2001. Vascular responses to syntactic processing: event-related fMRI study of relative clauses. Human Brain Mapping 15, 26–38.

Caplan, D., Waters, G.S., 1999. Verbal working memory and sentence comprehension. Behavioral and Brain Sciences 22 (1), 77–94.

Caplan, D., Waters, G., Kennedy, D., Alpert, N., Makris, N., DeDe, G., Michaud, J., Reddy, A., 2007. A study of syntactic processing in aphasia II: neurological aspects. Brain and Language 101 (2), 151–177.

Caplan, D., 2003b. Syntactic aspects of language disorders. In: Heilman, K.M., Valenstein, E. (Eds.), Clinical Neuropsychology, fourth ed. Oxford University Press, New York, pp. 61–91.

Cappa, S.F., 2012. Imaging semantics and syntax. Neuroimage 61 (2), 427–431.

Caramazza, A., Zurif, E.B., 1976. Dissociation of algorithmic and heuristic processes in language comprehension: evidence from aphasia. Brain and Language 3 (4), 572–582.

Carreiras, M., Pattamadilok, C., Meseguer, E., Barber, H., Devlin, J.T., 2012. Broca's area plays a causal role in morphosyntactic processing. Neuropsychologia 50 (5), 816–820.

Christensen, K.R., Kizach, J., Nyvad, A.M., 2012. The processing of syntactic islands – an fMRI study. Journal of Neurolinguistics 26, 239–251.

Christensen, S.C., Wright, H.H., 2010. Verbal and non-verbal working memory in aphasia: what three n-back tasks reveal. Aphasiology 24 (6–8), 752–762.

Christiansen, M.H., Louise Kelly, M., Shillcock, R.C., Greenfield, K., 2010. Impaired artificial grammar learning in agrammatism. Cognition 116 (3), 382–393.

Collette, F., Van der Linden, M., 2002. Brain imaging of the central executive component of working memory. Neuroscience and Biobehavioral Reviews 26 (2), 105–125.

Crinion, J.T., Warburton, E.A., Lambon-Ralph, M.A., Howard, D., Wise, R.J., 2006. Listening to narrative speech after aphasic stroke: the role of the left anterior temporal lobe. Cerebral Cortex 16 (8), 1116–1125.

Damasio, A.R., 1992. Aphasia. New England Journal of Medicine 326 (8), 531–539.

Dick, F., Bates, E., Wulfeck, B., Utman, J., Dronkers, N., Gernsbacher, M., 2001. Language deficits, localization, and grammar: evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals. Psychological Review 108, 759–788.

Dominey, P.F., Hoen, M., Blanc, J.-M., Lelekov-Boissard, T., 2003. Neurological basis of language and sequential cognition: evidence from simulation, aphasia, and ERP studies. Brain and Language 86 (2), 207–225.

Dronkers, N.F., Plaisant, O., Iba-Zizen, M.T., Cabanis, E.A., 2007. Paul Broca's historic cases: high resolution MR imaging of the brains of Leborgne and Lelong. Brain 130 (5), 1432–1441.

Dronkers, N.F., Wilkins, D.P., Van Valin, R.D.J., Redfern, B.B., Jaeger, J.J., 2004. Lesion analysis of the brain areas involved in language comprehension. Cognition 92 (1), 145–177.

Dronkers, N.F., Wilkins, D.P., Van Valin, R.D.J., Redfern, B.B., Jaeger, J.J., 1994. A reconsideration of the brain areas involved in the disruption of morphosyntactic comprehension. Brain and Language 47, 461–463.

Erickson, R.J., Goldinger, S.D., LaPointe, L.L., 1996. Auditory vigilance in aphasic individuals: detecting nonlinguistic stimuli with full or divided attention. Brain and Cognition 30 (2), 244–253.

Fiebach, C.J., Schlesewsky, M., Lohmann, G., Von Cramon, D.Y., Friederici, A.D., 2005. Revisiting the role of Broca's area in sentence processing: syntactic integration versus syntactic working memory. Human Brain Mapping 24 (2), 79–91.

Frey, S., Campbell, J.S.W., Pike, G.B., Petrides, M., 2008. Dissociating the human language pathways with high angular resolution diffusion fiber tractography. Journal of Neuroscience 28 (45), 11435–11444.

Friederici, A.D., 2012. The cortical language circuit: from auditory perception to sentence comprehension. Trends in Cognitive Sciences 16 (5), 262–268.

Friederici, A.D., 2011. The brain basis of language processing: from structure to function. Physiological Reviews 91 (4), 1357–1392.

Friederici, A.D., Kotz, S.A., 2003. The brain basis of syntactic processes: functional imaging and lesion studies. Neuroimage 20, S8–S17.

Friederici, A.D., Rüschemeyer, S.A., Hahne, A., Fiebach, C.J., 2003. The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes. Cerebral Cortex 13 (2), 170–177.

Friederici, A.D., Fiebach, C.J., Schlesewsky, M., Bornkessel, I.D., Von Cramon, D.Y., 2006. Processing linguistic complexity and grammaticality in the left frontal cortex. Cerebral Cortex 16 (12), 1709–1717.

Friedmann, N., Grodzinsky, Y., 1997. Tense and agreement in agrammatic production: prunning the syntactic tree. Brain and Language 56, 397–425.

Friedmann, N., Gvion, A., 2003. Sentence comprehension and working memory limitation in aphasia: a dissociation between semantic-syntactic and phonological reactivation. Brain and Language 86 (1), 23–39.

Fyndanis, V., Varlokosta, S., Tsapkini, K., 2010. Exploring *wh*-questions in agrammatism: evidence from Greek. Journal of Neurolinguistics 23 (6), 644–662.

Gorno-Tempini, M.L., Dronkers, N.F., Rankin, K.P., Ogar, J.M., Phengrasamy, L., Rosen, H.J., Johnson, J.K., Weiner, M.W., Miller, B.L., 2004. Cognition and anatomy in three variants of primary progressive aphasia. Annals of Neurology 55 (3), 335–346.

Gerton, B.K., Brown, T.T., Meyer-Lindenberg, A., Kohn, P., Holt, J.L., Olsen, R.K., Berman, K.F., 2004. Shared and distinct neurophysiological components of the digits forward and backward tasks as revealed by functional neuroimaging. Neuropsychologia 42 (13), 1781–1787.

Goodglass, H., 1993. Understanding Aphasia. Academic Press, San Diego.

Gorno-Tempini, M.L., Hillis, A.E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S.F., Ogar, J.M., Rohrer, J.D., Black, S., Boeve, B.F., Manes, F., Dronkers, N.F., Vandenberghe, R., Rascovsky, K., Patterson, K., Miller, B.L., Knopman, D.S., Hodges, J.R., Mesulam, M.M., Grossman, M., 2011. Classification of primary progressive aphasia and its variants. Neurology 76 (11), 1006–1014.

Grodzinsky, Y., 2005. Syntacic dependencies as memorized sequences in the brain. Canadian Journal of Linguistics/Revue cancadienne de linguistique 50, 241–266.

Grodzinsky, Y., 2000. The neurology of syntax: language use without Broca's area. Behavioral and Brain Sciences 23 (01), 1–21.

Grodzinsky, Y., 1990. Theoretical Perspectives on Language Deficits. MIT Press, Cambridge, MA.

Grodzinsky, Y., Finkel, L., 1998. The neurology of empty categories: aphasics' failure to detect ungrammaticality. Journal of Cognitive Neuroscience 10 (2), 281–292.

Gvion, A., Friedmann, N., 2012. Does phonological working memory impairment affect sentence comprehension? A study of conduction aphasia. Aphasiology 26 (3–4), 494–535.

Hanne, S., Sekerina, I.A., Vasishth, S., Burchert, F., De Bleser, R., 2011. Chance in agrammatic sentence comprehension: what does it really mean? Evidence from eye movements of German agrammatic aphasic patients. Aphasiology 25 (2), 221–244.

Hashimoto, R., Sakai, K.L., 2002. Specialization in the left prefrontal cortex for sentence comprehension. Neuron 35 (3), 589–597.

Heim, S., 2005. The structure and dynamics of normal language processing: insights from neuroimaging. Acta Neurobiologiae Experimentalis (Wars) 65 (1), 95–116.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nature Reviews Neuroscience 8 (5), 393–402.

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92 (1–2), 67–99.

Hoen, M., Golembiowski, M., Guyot, E., Deprez, V., Caplan, D., Dominey, P.F., 2003. Training with cognitive sequences improves syntactic comprehension in agrammatic aphasics. Neuroreport 14 (3), 495–499.

Hood, A., 1824. Case 4th – July 28, 1824 (Mr. Hood's cases of injuries of the brain). The Phrenological Journal and Miscellany II, 82–94.

Humphries, C., Love, T., Swinney, D., Hickok, G., 2005. Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing. Human Brain Mapping 26 (2), 128–138.

Iijima, K., Fukui, N., Sakai, K.L., 2009. The cortical dynamics in building syntactic structures of sentences: an MEG study in a minimal-pair paradigm. NeuroImage 44 (4), 1387–1396.

Just, M.A., Carpenter, P.A., Keller, T.A., Eddy, W.F., Thulborn, K.R., 1996. Brain activation modulated by sentence comprehension. Science 274 (5284), 114–116.

Kaan, E., Swaab, T.Y., 2002. The brain circuitry of syntactic comprehension. Trends in Cognitive Sciences 6 (8), 350–356.

Kasselimis, D., Potagas, C., Kourtidou, E., Evdokimidis, I., 2012. Classification of the aphasias: maybe it's time for a change. Procedia – Social and Behavioral Sciences 61, 161–162.

Kasselimis, D.S., Simos, P.G., Economou, A., Peppas, C., Evdokimidis, I., Potagas, C., 2013. Are memory deficits dependent on the presence of aphasia in left brain damaged patients? Neuropsychologia 51 (9), 1773–1776.

Kertesz, A., 1985. Aphasia. In: Frederiks, J.A.M. (Ed.), Handbook of Clinical Neurology, vol. 45. Elsevier, New York.

Kessels, R.P.C., Kappelle, L.J., de Haan, E.H., Postma, A., 2002. Lateralization of spatial-memory processes: evidence on spatial span, maze learning, and memory for object locations. Neuropsychologia 40, 1465–1473.

Kluender, R., Kutas, M., 1993. Bridging the Gap: evidence from ERPs on the processing of unbounded dependencies. Journal of Cognitive Neuroscience 5 (2), 196–214.

Kreisler, A., Godefroy, O., Delmaire, C., Debachy, B., Leclercq, M., Pruvo, J., Leys, D., 2000. The anatomy of aphasia revisited. Neurology 54 (5), 1117–1123.

Laures-Gore, J., Marshall, R.S., Verner, E., 2010. Performance of individuals with left hemisphere stroke and aphasia and individuals with right brain damage on forward and backward digit span tasks. Aphasiology 25 (1), 43–56.

Leff, A.P., Schofield, T.M., Crinion, J.T., Seghier, M.L., Grogan, A., Green, D.W., Price, C.J., 2009. The left superior temporal gyrus is a shared substrate for auditory short-term memory and speech comprehension: evidence from 210 patients with stroke. Brain 132, 3401–3410.

Lieberman, P., 1992. Could an autonomous syntax module have evolved? Brain and language 43 (4), 768–774.

Lieberman, P., 1985. On the evolution of human syntactic ability. Its pre-adaptive bases—motor control and speech. Journal of Human Evolution 14 (7), 657–668.

Marshall, J., 2011. Disorders of sentence processing in aphasia. In: Papathanasiou, I., Coppens, P., Potagas, C. (Eds.), Aphasia and Related Communication Disorders. Jones & Bartlett Learning, Burlington, MA, pp. 197–216.

Martin, N., 2008. The role of semantic processing in short-term memory and learning: evidence from aphasia. In: Thorn, A., Page, M. (Eds.), Interactions between Short-term and Long-term Memory in the Verbal Domain. Psychology Press, Hove, UK, pp. 220–243.

Martin, N., Ayala, J., 2004. Measurements of auditory-verbal STM span in aphasia: effects of item, task, and lexical impairment. Brain and Language 89, 464–483.

Martin, N., Saffran, E., 1997. Language and auditory-verbal short-term memory impairments: evidence for common underlying processes. Cognitive Neuropsychology 14 (5), 641–682.

Mazoyer, B.M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., Salamon, G., Dehaene, S., Cohen, L., Mehler, J., 1993. The cortical representation of speech. Journal of Cognitive Neuroscience 5 (4), 467–479.

Meltzer, J.A., McArdle, J.J., Schafer, R.J., Braun, A.R., 2010. Neural aspects of sentence comprehension: syntactic complexity, reversibility, and reanalysis. Cerebral cortex 20 (8), 1853–1864.

Mesulam, M., 1990. Large-scale neurocognitive networks and distributed processing for attention, language, and memory. Annals of neurology 28 (5), 597–613.

Miceli, G., 2003. Agrammatism. In: Whitajer, H. (Ed.), Concise Encyclopedia of Brain and Language. Elsevier Ltd, Oxford, pp. 14–17.

Mohr, J., Pessin, M., Finkelstein, H., Duncan, G.W., Davis, K.R., 1978. Broca's aphasia: pathological and clinical. Neurology 28, 311–324.

Moro, A., Tettamanti, M., Perani, D., Donati, C., Cappa, S.F., Fazio, F., 2001. Syntax and the brain: disentangling grammar by selective anomalies. NeuroImage 13, 110–118.

Nespoulous, J.L., Dordain, M., Perron, C., Ska, B., Bub, D., Caplan, D., Mehler, J., Lecours, A.R., 1988. Agrammatism in sentence production without comprehension deficits: reduced availability of syntactic structures and/or of grammatical morphemes? A case study. Brain and Language 33 (2), 273–295.

Newhart, M., Trupe, L.A., Gomez, Y., Cloutman, L., Molitoris, J.J., Davis, C., Leigh, R., Gottesman, R.F., Race, D., Hillis, A.E., 2012. Asyntactic comprehension, working memory, and acute ischemia in Broca's area versus angular gyrus. Cortex 48 (10), 1288–1297.

den Ouden, D.B., Saur, D., Mader, W., Schelter, B., Lukic, S., Wali, E., Timmer, J., Thompson, C.K., 2012. Network modulation during complex syntactic processing. NeuroImage 59 (1), 815–823.

Obleser, J., Meyer, L., Friederici, A.D., 2011. Dynamic assignment of neural resources in auditory comprehension of complex sentences. NeuroImage 56 (4), 2310–2320.

Peristeri, E., Tsimpli, I., Tsapkini, K., 2013. The on-line processing of unaccusativity in Greek agrammatism. Applied Psycholinguistics 34 (2), 233–276.

Petrides, M., Pandya, D.N., 1988. Association fiber pathways to the frontal cortex from the superior temporal region in the rhesus monkey. Journal of Comparative Neurology 273 (1), 52–66.

Potagas, C., Kasselimis, D., Evdokimidis, I., 2011. Short-term and working memory impairments in aphasia. Neuropsychologia 49 (10), 2874–2878.

Price, C.J., 2000. The anatomy of language: contributions from functional neuro-imaging. Journal of Anatomy 197 (3), 335–359.

Ricciardi, E., Bonino, D., Gentili, C., Sani, L., Pietrini, P., Vecchi, T., 2006. Neural correlates of spatial working memory in humans: a functional magnetic resonance imaging study comparing visual and tactile processes. Neuroscience 139 (1), 339–349.

Rochon, E., Saffran, E.M., Berndt, R.S., Schwartz, M.F., 2000. Quantitative analysis of aphasic sentence production: further development and new data. Brain and Language 72 (3), 193–218.

Röder, B., Stock, O., Neville, H., Bien, S., Rösler, F., 2002. Brain activation modulated by the comprehension of normal and pseudo-word sentences of different processing demands: a functional magnetic resonance imaging study. NeuroImage 15, 1003–1014.

Ronnberg, J., Larsson, C., Fogelsjoo, A., Nilsson, L.G., Lindberg, M., Angquist, K.A., 1996. Memory dysfunction in mild aphasics. Scandinavian Journal of Psychology 37 (1), 46–61.

Ruiter, M.B., Kolk, H.H., Rietveld, T.C., 2010. Speaking in ellipses: the effect of a compensatory style of speech on functional communication in chronic agrammatism. Neuropsychological rehabilitation 20 (3), 423–458.

Saffran, E.M., Berndt, R.S., Schwartz, M.F., 1989. The quantitative analysis of agrammatic production: procedure and data. Brain and Language 37 (3), 440–479.

Salmelin, R., 2007. Clinical neurophysiology of language: the MEG approach. Clinical Neurophysiology 118 (2), 237–254.

Santi, A., Grodzinsky, Y., 2007. Working memory and syntax interact in Broca's area. NeuroImage 37 (1), 8–17.

Saur, D., Kreher, B.W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, W., Rijntjes, M., Hennig, J., Weiller, C., 2008. Ventral and dorsal pathways for language. Proceedings of the National Academy of Sciences of the United States of America 105 (46), 18035–18040.

Schwartz, M.F., Linebarger, M.C., Saffran, E.M., Pate, D.C., 1987. Syntactic transparency and sentence interpretation in aphasia. Language and Cognitive Processes 2, 85–113.

Shallice, T., Warrington, E.K., 1970. Independent functioning of verbal memory stores: a neuropsychological study. The Quarterly journal of experimental psychology 22 (2), 261–273.

Shapiro, L.P., Gordon, B., Hack, N., Killackey, J., 1993. Verb-argument structure processing in complex sentences in Broca's and Wernicke's aphasia. Brain and Language 45 (3), 423–447.

Signoret, J.L., Castaigne, P., Lhermitte, F., Abelanet, R., Lavorel, P., 1984. Rediscovery of Leborgne's brain: anatomical description with CT scan. Brain and Language 22, 303–319.

Stowe, L.A., Broere, C.A.J., Paans, A.M.J., Wijers, G.M., Vaalburg, W., Zwarts, F., 1998. Localizing components of a complex task: sentence processing and working memory. NeuroReport 9, 2995–2999.

Stromswold, K., Caplan, D., Alpert, N., Rauch, S., 1996. Localization of syntactic comprehension by positron emission tomography. Brain and Language 52 (3), 452–473.

Swinney, D., Zurif, E., 1995. Syntactic processing in aphasia. Brain and Language 50 (2), 225–239.

Swinney, D., Zurif, E., Prather, P., Love, T., 1996. Neurological distribution of processing resources underlying language comprehension. Journal of Cognitive Neuroscience 8 (2), 174–184.

Tettamanti, M., Rotondi, I., Perani, D., Scotti, G., Fazio, F., Cappa, S.F., Moro, A., 2009. Syntax without language: neurobiological evidence for cross-domain syntactic computations. Cortex 45 (7), 825–838.

Tettamanti, M., Weniger, D., 2006. Broca's area: a supramodal hierarchical processor? Cortex 42 (4), 491–494.

Thompson, C.K., den Ouden, D.B., Bonakdarpour, B., Garibaldi, K., Parrish, T.B., 2010. Neural plasticity and treatment-induced recovery of sentence processing in agrammatism. Neuropsychologia 48 (11), 3211–3227.

Thothathiri, M., Kim, A., Trueswell, J.C., Thompson-Schill, S.L., 2012. Parametric effects of syntactic–semantic conflict in Broca's area during sentence processing. Brain and language 120 (3), 259–264.

Tyler, L.K., Marslen-Wilson, W.D., Randall, B., Wright, P., Devereux, B.J., Zhuang, J., Papoutsi, M., Stamatakis, E.A., 2011. Left inferior frontal cortex and syntax: function, structure and behaviour in patients with left hemisphere damage. Brain 134 (2), 415–431.

Wager, T.D., Smith, E.E., 2003. Neuroimaging studies of working memory: a meta-analysis. Cognitive, Affective, & Behavioral Neuroscience 3 (4), 255–274.

Waters, G., Caplan, D., Alpert, N., Stanczak, L., 2003. Individual differences in rCBF correlates of syntactic processing in sentence comprehension: effects of working memory and speed of processing. NeuroImage 19 (1), 101–112.

Wenzlaff, M., Clahsen, H., 2004. Tense and agreement in German agrammatism. Brain and Language 89, 57–68.

Wright, H.H., Downey, R.A., Gravier, M., Love, T., Shapiro, L.P., 2007. Processing distinct linguistic information types in working memory in aphasia. Aphasiology 21 (6–8), 802–813.

Zurif, E.B., Caramazza, A., 1976. Linguistic structures in aphasia: studies in syntax and semantics. In: Whitaker, H., Whitaker, H.H. (Eds.), Studies in Neurolinguistics, vol. 2. Academic Press, New York.

Zurif, E., Swinney, D., Prather, P., Solomon, J., Bushell, C., 1993. An on-line analysis of syntactic processing in Broca's and Wernicke's aphasia. Brain and Language 45 (3), 448–464.

Neuromarketing

Fabio Babiloni, University of Rome Sapienza, Rome, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Neuromarketing is an area of applied neuroscience devoted to the analysis of brain responses to advertising stimuli. It is generally perceived by the public as promoted by commercial interest in the use of neurophysiological tools able to track the main neurophysiological signals and consumer's behavior. However, in the scientific literature, a precise definition of the word 'neuromarketing' has been attempted. Neuromarketing is defined as "the application of neuroscientific methods to analyze and understand human behavior in relation to markets and marketing exchanges."

Introduction

In the economically developed countries, each day a person is reached on average by more than 100 commercial advertisements, through different media such as television, newspapers, radio, and the Internet. Such crowding of the advertising makes the job of marketers really difficult. In fact, they have to design and generate advertising messages able to compete with the others already running in order to capture our limited attention and memory resources. It has been pointed out how our body and the related sensory receptors are targeted by 10^7 information bits per second from the external world, while the average amount of such sensory perceptions retained in our memory is just 10 bits per second (Wilson, 2002).

The capability of a commercial advertisement to attract our attention is also a function of how often it reaches us each day through the different media, such as the television, radio, and the Internet. The greater the 'pressure' exerted by the ad, the easier we will remember it. However, due to the current economic crisis in the Western countries, the use of such 'pressure' leverage to increase memorability of an ad has become limited. For the same economic reasons, the interest in producing efficient and attractive advertisements has greatly increased.

Traditionally, investigations of the consumer's needs and desires have been performed through methodologies involving interviews with single consumers or small meetings with them, called focus groups. Focus groups and surveys are usually quite simple to run and not expensive. However, they are not particularly accurate in reporting consumer intention, due to the fact that people usually do not report clearly and completely what they think about a product (Griffin and Hauser, 1993). In addition, interview subjects are often unable to correctly report their feelings about the product under consideration (Green and Srinivasan, 1990).

In fact, it is recognized that often the verbal advertising pretesting is flawed by the respondents' cognitive processes activated during the interview, with the implicit memory and subject's feelings often inaccessible to the interviewer who uses traditional techniques (Zaltman, 2003). In addition, it has also been suggested that the interviewer with typical pretesting interviews has a great influence on what the respondent recalls and on the subjective experiencing of it (McDonald, 2003). In fact, the mismatch between the verbal report of consumers and their true internal feelings about the product has been thought responsible for the flop of about 70% of the new products launched on the world market.

Keeping all of these considerations in mind, researchers have attempted to investigate the signs of the brain activity correlated with an increase of attention, memory, or emotional engagement during the observation of commercials. Researchers within the neuroscience community promote the view that findings and methods from neuroscience complement and illuminate existing knowledge in consumer research in order to better understand consumer behavior (e.g., Ambler and Burne, 1999).

In this scenario, in this last decade the use of scientific results deriving from the neuroscience area related to the cerebral processes of attention, memorization, and emotion during activities related to the economic transactions has been explored by an increasing number of researchers. In particular, researchers have begun to investigate the brain areas involved in the generation of economic transactions by using different neuroimaging devices, such as functional magnetic resonance imaging (fMRI), the electroencephalogram (EEG), or the magnetoencephalogram (MEG) (see Ariely and Berns, 2010 for a review). Researchers have recently begun to investigate the neuronal mechanisms underlying a more specific form of economically relevant behavior, that is, consumer choices concerning the selection and consumption of certain products. In this approach, one is interested in the relationship between marketing stimuli to which individuals are exposed and the related brain responses.

From an economic and/or marketing perspective, the aim of such investigations is a better understanding of how consumer advertising of (established) brands affects brain systems. From a neuroscience perspective, the broad goal is a better understanding of both the neural mechanisms underlying the impact of affect and cognition on memory and the neural correlates of choice and decision-making.

While neuroscience is a general term to indicate the study of the nervous system that seeks to understand the biological basis of behavior, the application of neuroscience to *consumer psychology* has gained popularity over the past decade in academic research and business practice. In particular, an adopted definition of the *consumer neuroscience* area of research is that it is a discipline that adapts methods and theories from neuroscience in conjunction with different behavioral theories,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.56029-2

models, and tested experimental designs from the area of consumer psychology. This is performed in order to develop a neuropsychologically sound theory to understand consumer behavior (Plassmann et al., 2012).

A final critical distinction is between *consumer neuroscience*, which refers to academic research performed in the area of consumer psychology and neuroscience, and the area related to the study of *neuromarketing*. The last area of study is generally perceived by the public as pushed forward by commercial interests in the use of neurophysiological tools able to track the main neurophysiologic signals and consumer's behavior to conduct company-specific market research.

However, in the scientific literature, a precise definition of neuromarketing has been attempted and proposed (Lee et al., 2007). In fact, the term 'neuromarketing' describes a field of study defined as "the application of neuroscientific methods to analyze and understand human behavior in relation to markets and marketing exchanges" (Lee et al., 2007). Such a definition has two main upshots: (1) it moves consideration of neuromarketing away from being solely the use of neuroimaging by commercial interests for their benefit; and (2) the scope of neuromarketing research is widened from solely consumer behavior to include many more avenues of interest, such as inter- and intra-organizational research, which are common in the marketing research literature. However, it must be stated that the term 'neuromarketing' is generally perceived as related to the study of the brain activity during its exposition to advertisements.

The incorporation of neuroimaging into the decision-making sciences has raised the interest of the marketing academia and professionals in this area of knowledge. It has been hoped that neuromarketing, with the associated neuroimaging technology, can aid marketers to generate a more efficient way to optimize the impact of advertisements on the human brain. In addition, it has also been thought that the use of neuroimaging measurements could return information about consumer preferences not immediately available through conventional methods (Ariely and Berns, 2010).

How Neuromarketing Methodologies Could Inform Marketing Processes

Nowadays, it has been clearly perceived that neuroscience imaging technology could be used to better understand the aspects of brain perception related to commercial advertisements. The use of such technology could be inserted in the cycle of the design, realization, launch, and support of a new product as described in **Figure 1**.

As described in **Figure 1**, neuroscience methodologies applied to the understanding of commercial advertising could have a role even before a single product could be realized. In fact, analysis of the consumer's brain responses due to the proposition of an 'idea' of a product could be performed. In this case, neuroimaging could be then employed to understand if consumers effectively agreed with the proposed idea or product concept, beyond their verbal opinion.

However, the most conventional use of neuroimaging in the analysis of marketing messages is generated in the area of the product cycle in which the product itself has been already realized and launched on the market. In this context, the analysis is of the impact on the consumer's brain of the particular advertising campaign supporting the product.

In this case, the dark box at the lower right of **Figure 1** describes the use of neuromarketing methodologies at the post-hoc evaluation of the advertising campaign produced. At this level, the corrections to be generated by the use of neuromarketing technologies could be applied not to the product itself but rather to the idea of the product conveyed to the consumers through the advertising messages. Neuroimaging studies could return information also useful for tailoring successive advertising campaigns related to the particular product launched.

Neurophysiologic Signals Used by Neuromarketing Methodologies

There are different kinds of neuroimaging technologies usually applied in the evaluation of the commercial products (or even ideas) in the neuromarketing area, and each specific technology involves the use of a particular neurophysiologic signal related to the brain activity.

The use of the variation of the hemodynamic blood flow in the brain during the observation of a particular ad is one of the most popular technologies employed in neuromarketing. In particular, fMRI has been largely validated in neuroscience during the last 20 years as a methodology able to reveal the activity of brain areas with a remarkable spatial resolution (on the order of mm) and an acceptable time resolution (seconds). Several companies offer neuromarketing services based on fMRI neuroimaging devices for the analysis of commercial advertising. The use of fMRI in this area of applied science is partially limited by the relatively high cost of the fMRI analysis per subject, in particular when compared to other neuroimaging technologies such as the EEG. However, the use of fMRI allows the analysis of the activity of particular subcortical structures known to be involved in the appreciation and avoidance of the presented stimuli, such as the nucleus accumbens. Such capability cannot be achieved by using EEG or MEG techniques. A partial limitation of fMRI for use in a marketing context is its limited temporal resolution (on the order of seconds), which represents an obstacle for the use of such a technique for the time-varying evaluation of commercial TV spots.

The measurement of electromagnetic activity related to the brain activity through an EEG or an MEG is another popular approach in the neuromarketing area (Ioannides et al., 2000; Vecchiato et al., 2010). The main limitations of EEG or MEG neuroimaging technologies are related to the moderate spatial resolution achievable (cm) and to the relative difficulty for the detection of the cerebral activity originating at the subcortical structures, such as the nucleus accumbens or the amygdala. However, a great advantage of the EEG and MEG technologies when compared to fMRI is related to their high temporal resolution, which is on the order of ms. Such temporal resolution is more than sufficient to track the time-varying changes of cortical brain activity during the observation of a TV commercial, making the EEG or MEG a good choice for the evaluation of cerebral impact of a TV

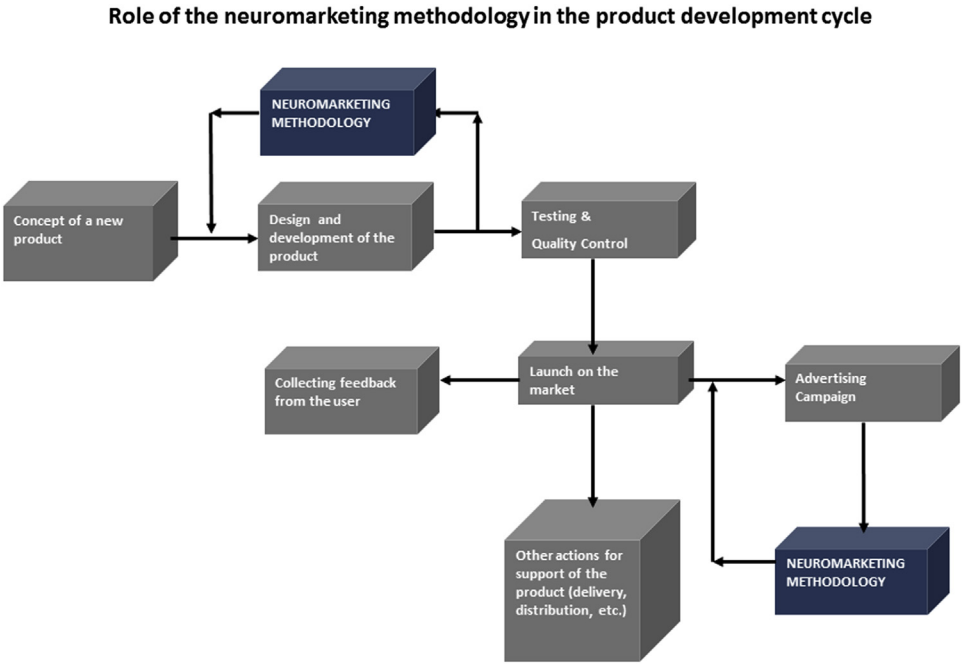

Figure 1 A possible scheme for the design and the launch of a new product on the market. The dark box indicates the possible use of neuromarketing methodologies, able to generate insights in the evaluation of the new concepts to be implemented in a product. Such use could take place in an early phase before the generation of the product itself (upper dark box) or in a post-hoc evaluation of the already-launched advertising campaign (the dark box at the bottom right of the figure).

ad. It is worth noting that MEG devices are relatively scarce in the world (due to their cost), while EEG devices are relatively cheap when compared to fMRI or MEG technologies. Such considerations explain why the use of EEG technology is largely adopted by many neuromarketing companies for the evaluation of marketing relevant stimuli.

There are other devices and technologies that could be used in the evaluation of the consumer's cerebral activity during the evaluation of marketing messages. These devices usually measure the indirect signs of the brain activity, mediated by the autonomous nervous system or through the facial muscles or eye movements.

Eye-tracking devices are usually employed in marketing research by tracking eye movements through different sensors such as glasses or even a camera placed on a monitor's screen. Some eye-tracking devices are also able to return information on the diameter of the pupil, which is thought to be linked to the general arousal level of the subject. Extensively used in neuromarketing practice, eye-tracking devices return information about the gaze direction and are used to evaluate the main attractive points of a static marketing message, such as an advertisement on a wall, rather than a TV commercial.

The time-varying signals provided by cutaneous impedance, usually measured at the finger level, is known as electrodermal activity (EDA; Critchley, 2002). EDA signals the variation of the autonomous sympathetic system, with a time resolution of less than 1 s. The kind of information returned by EDA measurements is related to the fluctuation of the arousal level, which is a valuable source of information about comfort or discomfort of the person at a point of sale or when watching a TV commercial. EDA evaluation is offered by several companies that link such arousal variation (Critchley, 2002) with the emotional involvement generated by the advertising.

Another signal employed in neuromarketing is the heart rate, that is, the instantaneous frequency of the heartbeats. Such a variable is again linked with the general arousal state of the person during the observation of advertising. Heartbeat is often collected in conjunction with the level of EDA responses, for a full characterization of the time evolution of the consumer's autonomic system in correlation with the presentation of the advertising.

In particular, by monitoring autonomic activity using devices able to record the variation of skin conductivity and heart rate, it is possible to assess the 'internal' emotional state of the subject. In fact, the EDA is actually viewed as a sensitive and convenient measure of indexing changes in sympathetic arousal associated with emotion, cognition, and attention (Critchley, 2002). Studies using functional imaging techniques have related the generation and maintenance of EDA level to specific brain areas. These specific regions are the ventromedial prefrontal cortex, orbitofrontal cortex, left primary motor cortex, and the anterior and posterior cingulate, which have been shown to be associated with emotional and motivational behaviors (Critchley, 2002). Such findings indicating the close association of peripheral and central measures of arousal reemphasize the close connections between EDA, arousal, attention, cognition, and emotion. In addition, the link between heart rate or heart rate variability and the sympatho/vagal balance has already been suggested (Malik, 1998).

Main Cerebral Structures Involved in Advertising and Brand Evaluation

It has been emphasized how a positive or negative emotional processing of advertising is an important factor for the formation of stable memory traces (Kato et al., 2009). Hence, it became relevant to infer the emotional engagement of the subject during the observation of an advertising message. In fact, indirect signs of emotional processing could be gathered by picking variations of the activity of the anatomical structures linked to such processing activity in humans. Indirect signs of emotional activity could be extracted by using galvanic skin response measurements, linked to the modulation of activity of the sweat glands on the hands or by using the heart rate measurements (Baumgartner et al., 2008). Direct signs of emotional activity could be obtained by gathering the variations of the cerebral activity on the prefrontal and frontal cortices (PFC and FC, respectively) by EEG (Davidson and Irwin, 1999).

In fact, the PFC region is structurally and functionally heterogeneous but its role in emotion is well recognized (Davidson, 2002). EEG spectral power studies indicate that the anterior cerebral hemispheres are differentially lateralized for approach and withdrawal motivational tendencies and emotional appreciation (Davidson and Irwin, 1999; Davidson, 2000, 2004). Specifically, several EEG findings suggest that the left PFC is an important brain area in a widespread circuit that mediates appetitive approach, while the right PFC appears to form a major component of a neural circuit that instantiates defensive withdrawal (Davidson, 2000, 2004).

It has also been proposed that the use of the unbalancing of the EEG spectral activity over the left and right prefrontal area in humans could be a reliable indicator of the prevalence of the 'approach' system when compared to the 'withdraw' one during the observation of different stimuli (Silberstein and Nield, 2012).

With such index, it has become possible to judge how much one likes a commercial during its temporal development, in order to determine the part of it that caused major appreciation from the observer and the part that did not.

However, the major part of the information about the brain structures involved in the appreciation of advertising in the literature has been obtained by the use of hemodynamic signals. In particular, an important body of research in advertising has been performed with the use of fMRI for the analysis of the 'brand' concept and for the localization of brain areas that respond to brands presented in different experimental conditions. Such research will now be briefly discussed.

It has been reported in a study by Deppe et al. (2005a) that the ventromedial prefrontal cortex (vmPFC) showed greater activity during the observation of a preferred brand as contrasted to the second or lower ranked brands in a healthy population. Of note, in a study by Koenigs and Tranel (2008), patients with damage in the same cortical structure (vmPFC) were not biased by brand information during blind versus open tasting of carbonated drinks (Coke and Pepsi). Taken together, such studies suggest an involvement of the vmPFC in favorite brand appreciation. However, the vmPFC is not the only brain structure that is active during such brand observation and appreciation. In fact, an fMRI study reports that observation of familiar brands by subjects activates their middle frontal gyrus when compared to the observation of unfamiliar brands (Schaefer et al., 2006). In addition, different studies suggested that brain activity at the anterior cingulate cortex could predict whether a person is biased by the brand name of a newspaper while evaluating the credibility of a headline or the attractiveness of print advertisements (Deppe et al., 2005b, 2007). Consistent activity was also observed in the ventral striatum during the imagined consumption of a product of preferred brands (Schaefer and Rotte, 2007), or during the observation of brands to which subjects are 'loyal' when compared to the observation of brands to which the subjects were 'unloyal' (Plassmann et al., 2007). In addition, the network of the cerebral areas active during the brand judgments are different than those elicited in the case of judgments related to humans traits (Yoon et al., 2006).

Imaging by fMRI has also been used to study how brand information could alter brand evaluation during product consumption. In a famous study, the stated preferences for Coke versus Pepsi did not correlate with the revealed preferences in blind tasting (McClure et al., 2004). However, revealed preferences correlated with the brain activity changes in the vmPFC and in the medial orbitofrontal cortex. In addition, awareness by the subjects that they were drinking Coke generated activity changes in memory and association areas (hippocampus, dorsolateral PFC), while no such difference could be found in the case of Pepsi (McClure et al., 2004).

In conclusion, the use of neuroscience information for the evaluation of commercial ads or brands has become popular nowadays and represents one of the first and most evident applications of neuroscience into an industrial area not related to healthcare. Although several difficulties are present in extending the above-mentioned neuroscience findings into practical and marketing-relevant information, namely due to the reverse inference problem, there is no doubt that the possibility of mapping brain activity related to the evaluation of marketing messages will be followed with a lot of attention by the marketing industry in the years to come.

See also: Attention, Neural Basis of; Cerebral Cortex; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Consumer Psychology; Electroencephalography: Basic Principles and Applications; Emotion in Cognition; Learning and Memory, Neural Basis of; Magnetic Resonance and Molecular Imaging in Psychiatry; Marketing Strategies; Prefrontal Cortex; Recognition.

Bibliography

Ambler, T., Burne, T., 1999. The impact of affect on memory of advertising. Journal of Advertising Research 39 (2), 25–34.

Ariely, D., Berns, G., 2010. Neuromarketing: the hope and hype of neuroimaging in business. Nature Reviews Neuroscience 11 (4), 284–292.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., Fehr, E., 2008. Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. Neuron 58, 639–650.

Critchley, H., 2002. Electrodermal responses: what Happens in the brain. Neuroscientist 8 (2), 132–142.

Davidson, R.J., 2000. Affective style, psychopathology, and resilience: brain mechanisms and plasticity. American Psychologist 55, 1196–11214.

Davidson, R.J., January 1 2002. Anxiety and affective style: role of prefrontal cortex and amygdala. Biological Psychiatry 51 (1), 68–80.

Davidson, R.J., 2004. What does the prefrontal cortex "do" in affect: perspectives on frontal EEG asymmetry research. Biological Psychology 67 (1–2), 219–233.

Davidson, R.J., Irwin, W., 1999. The functional neuroanatomy of emotion and affective style. Trends in Cognitive Sciences 3, 11–21.

Deppe, M., Schwindt, W., Kugel, H., Plassmann, H., Kenning, P., 2005a. Nonlinear responses within the medial prefrontal cortex reveal when specific implicit information influences economic decisionmaking. Journal of Neuroimaging 15, 171–182.

Deppe, M., Schwindt, W., Kramer, J., Kugel, H., Plassmann, H., Kenning, P., Ringelstein, E.B., 2005b. Evidence for a neural correlate of a framing effect: bias-specific activity in the ventromedial prefrontal cortex during credibility judgments. Brain Research Bulletin 67, 413–421.

Deppe, M., Schwindt, W., Pieper, A., Kugel, H., Plassmann, H., Kenning, P., Deppe, K., Ringelstein, E.B., 2007. Anterior cingulate reflects susceptibility to framing during attractiveness evaluation. Neuroreport 18, 1119–1123.

Green, P.E., Srinivasan, V., 1990. Conjoint analysis in marketing: new developments with implications for research and practice. The Journal of Marketing 54 (4), 3–19.

Griffin, A., Hauser, J.R., 1993. The voice of the customer. Marketing Science 12 (1), 1–27.

Ioannides, A.A., Liu, L., Theofilou, D., Dammers, J., Burne, T., Ambler, T., Rose, S., 2000. Real time processing of affective and cognitive stimuli in the human brain extracted from MEG signals. Brain Topography 13 (1), 11–19 (Fall).

Kato, J., et al., 2009. Neural correlates of attitude change following positive and negative advertisements. Frontiers in Behavioral Neuroscience 3, 6.

Koenigs, M., Tranel, D., 2008. Prefrontal cortex damage abolishes brand-cued changes in cola preference. Social Cognitive and Affective Neuroscience 3, 1–6.

Lee, N., Broderick, A.J., Chamberlain, L., 2007. What is "neuromarketing"? A discussion and agenda for future research. International Journal of Psychophysiology 63 (2), 199–204.

Malik, M., Jan 1998. Heart rate Variability. Current Opinion in Cardiology 13 (1), 36–44 (Review).

McClure, S.M., Li, J., Tomlin, D., Cypert, K.S., Montague, L.M., Montague, P.R., 2004. Neural correlates of behavioral preference for culturally familiar drinks. Neuron 44 (2), 379–387.

McDonald, 2003. Is Your Advertising Working? World Advertising Research Centre, Henley-on-Thames, U.K.

Plassmann, H., Kenning, P., Ahlert, D., 2007. Why companies should make their customers happy: the neural correlates of customer loyalty. Advances in Consumer Research—North American Conference Proceedings 34, 735–739.

Plassmann, H., Zoëga Ramsøy, T., Milosavljevic, M., 2012. Branding the brain: a critical review and outlook. Journal of Consumer Psychology 22, 18–36.

Schaefer, M., Rotte, M., 2007. Favorite brands as cultural objects modulate reward circuit. Neuroreport 18 (2), 141–145.

Schaefer, M., Berens, H., Heinze, H.J., Rotte, M., 2006. Neural correlates of culturally familiar brands of car manufacturers. Neuroimage 31, 861–865.

Silberstein, R.B., Nield, G.E., May–Jun 2012. Measuring emotion in advertising research: prefrontal brain activity. IEEE Pulse 3 (3), 24–27. http://dx.doi.org/10.1109/MPUL.2012.2189172.

Vecchiato, G., Astolfi, L., De Vico Fallani, F., Cincotti, F., Mattia, D., Salinari, S., Soranzo, R., Babiloni, F., 2010. Changes in brain activity during the observation of TV commercials by using EEG, GSR and HR measurements. Brain Topography 23 (2), 165–179.

Wilson, T.D., 2002. Strangers to Ourselves: Discovering the Adaptive Unconscious. Belknap Press of Harvard University Press, Cambridge, MA.

Yoon, C., Gutchess, A.H., Feinberg, F., Polk, T.A., 2006. A functional magnetic resonance imaging study of neural dissociations between brand and person judgments. Journal of Consumer Research 33 (1), 31–40.

Zaltman, G., 2003. How Customers Think. Harvard Business School Press, Boston, MA.

Neurons and Dendrites, Integration of Information in

Bartlett W Mel and Bryan Kolb, University of Southern California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 10600–10605, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Neurons are the building blocks of the brain. This article reviews the basic components and biophysical mechanisms that underlie nerve cell function, including a discussion of ion channels, the membrane potential, electrical signaling, synaptic function, and synaptic integration. The implications of these electrophysiological phenomena are then examined in relation to theories of information processing in single neurons.

Introduction and Scope

At the beginning of the twenty-first century, neural tissue remains arguably the most powerful substrate for information processing in existence. Unfortunately, the principles underlying brain function are still known only in rough outline. To unlock the many secrets of the brain and to help explain the panoply of neural functions ranging from simple motor reflexes to conscious thought and action, neuroscientists have amassed an extensive body of experimental data bearing on the properties of nerve cells (neurons) and their interconnections (synapses), which together form the building blocks of neural tissue. At the same time, the functional significance of the neurobiological data has been explored in computer modeling studies, which have often been used to test hypotheses inaccessible to direct experimental approaches. This article summarizes what is known about the intracellular physiology of individual nerve cells and interprets these data in relation to the information processing functions of single neurons. For comprehensive treatments of many of these same issues, see Stuart et al. (1999) and Koch (1999).

Given the morphological and functional diversity of neurons across the invertebrate and vertebrate phyla, the present article is focused on the principal neuron types of the vertebrate central nervous system (CNS), though much of what is discussed applies to neurons in general. This focus is justified, first, by the fact that neurons of the vertebrate CNS have been among the most frequent objects of correlated physiological, behavioral, and modeling studies, and second, by the fact that these cells are those which most directly underlie the cognitive and behavioral phenomena with which this volume is concerned.

Basic Construction of a Neuron

The typical neuron of the vertebrate CNS is composed of (1) a cell body, or soma, which contains the nucleus and other molecular machinery present in any living cell, (2) the dendritic tree, which emanates from the cell body and acts as the main input-receiving surface of the neuron, (3) the axon, which carries output signals generated by the neuron over long distances to other neurons, and (4) a large number of synapses, the specialized sites where axons make contacts with the dendrites and cell bodies of other neurons for subsequent processing (**Figure 1**).

The axon, a thin fiber emanating from the cell body, is the cell's final common output used to convey messages to other

Figure 1 (Left) Pyramidal cell from rat parietal cortex. Dendrites protruding from the cell body (dark pyramidal shape) have tiny protrusions (spines) covering the dendritic branches are the sites of most excitatory synapses. One pyramidal cell may contain tens of thousands of spines. Inhibitory synapses are less numerous, occur primarily on dendritic shafts and are concentrated near the cell body. Axon is the smooth protrusion from the bottom of the cell body. (Right). Model neuron with *m* identical branches containing *k* synapses each.

neurons. The axon carries trains of 'action potentials' (APs), all-or-none voltage spikes 60–100 mV in height and 1–2 ms in duration, which propagate undecremented along the length of the axon at speeds up to 100 m s^{-1}. The single axon of a neuron, which may travel over distances ranging from millimeters to meters depending on the neuron and which may branch profusely at one or more sites along its run, gives rise to many thousands of synaptic terminals impinging on the dendrites and cells bodies of hundreds to thousands of other neurons. Neurons can fire APs at rates up to 1000 impulses per second, though the maximum firing rate for most neurons in the CNS lies in the range of 100–200 Hz. APs are triggered when the membrane potential at the site of spike initiation reaches a firing threshold. In a typical neuron, this threshold is around −50 mV, that is, about 20 mV more positive than the cell's resting potential of −70 mV. While APs are thought to be triggered in the axon near the cell body, the electric currents that flow there represent the combined influences of many widely scattered synaptic inputs acting on the soma/dendritic membrane.

The dendritic tree is the main input-receiving domain of the neuron, generally accounting for more than 95% of the cell's total surface area, and receiving many thousands of synaptic inputs from other cells. With a few notable exceptions, the dendrites of CNS neurons are purely input structures (Shepherd, 1998). A synaptic input to a dendrite can be categorized roughly as either 'excitatory' or 'inhibitory,' meaning that when the synapse is activated, the membrane on the 'postsynaptic,' that is, dendritic side of the synaptic junction is driven transiently to a more positive (excitatory) or more negative (inhibitory) potential. Most of the contacts on a dendritic tree are excitatory (80–90%); inhibitory contacts can occur anywhere, but tend to be concentrated on or around the cell body. A synaptically evoked change in membrane potential can have a time course ranging from one to hundreds of milliseconds, depending on the type of synapse. In addition, synaptic inputs can trigger 'active' dendritic responses, including complex spike-like events involving both fast and slow components (e.g., Kamondi et al., 1998; Golding et al., 1999). Thus, the rules that govern how excitatory and inhibitory synaptic inputs are combined and transformed within the dendritic tree to produce outbound APs can be quite complex. This process commonly is referred to as synaptic or dendritic 'integration,' and is the focus of the remainder of this article.

The Membrane Potential

To understand how synaptic and other electrical events combine to generate APs, it is important to understand the basic electrical properties of 'neurites' (axons and dendrites).

The resting membrane potential of a neuron, as for any living cell, is established by molecular pumps, which actively maintain concentration gradients of sodium (Na^+) and potassium (K^+) ions across the cell membrane. Relative to the extracellular fluid, the inside of the cell contains a high concentration of K^+ ions and a low concentration of Na^+ ions. The constant pressure on these electrically charged ions to flow across the membrane down their respective concentration gradients (Na^+ in, K^+ out), is analogous to the pressure electrons feel that propels them from the negative to the positive terminal of a battery given a suitable path for current. The electrical pressure on each of these ions can thus be represented by a battery, whose voltage conveys the magnitude and direction of the ion's concentration gradient; the fact that the Na^+ and K^+ batteries point in opposite directions will turn out to be critical for electrical signaling.

Since the membrane–lipid bilayer is itself impervious to ions, currents can only flow through 'ion channels,' specially designed pores in the membrane formed by large, membrane-spanning protein complexes. Ion channels are generally quite selective as to which ions may flow through them, giving rise to the nomenclature 'sodium channels,' 'potassium channels,' and so on. Furthermore, certain ion channels are tonically open, while others can open and close under the control of various signals. For example, some channels open and close in response to changes in the membrane potential, such as occurs in the production of APs, others in response to chemical signals, such as occurs in the production of synaptic potentials.

The resting membrane potential of a neuron is a weighted average of ionic battery values, where the weight given to each ion's battery is determined by the membrane's total electrical conductance (g_I) to ion I through all open channels

$$V_m = \frac{g_{Na}E_{Na} + g_K E_K + \cdots}{g_{Na} + g_K + \cdots} \quad [1]$$

The usual resting potential around $V_m = -70$ mV is dominated by a K^+ conductance, whose associated battery value is $E_K = -90$ mV. An additional small resting Na^+ conductance, whose associated battery value $E_{Na} = +60$ mV or higher, pulls the resting potential to some 20 mV in the positive direction.

Electrical Signaling Results from Conductance Changes

Electrical signaling in a neuron is caused by transient changes in the conductance of various ion channels, which cause transient fluctuations in the membrane potential about its resting value. At every moment in time, the membrane potential moves toward the equilibrium value given by eqn [1] – not instantaneously – but at a rate limited by the need to charge or discharge the membrane capacitance. The classical signal generated by a neuron is the AP. When a synaptic or other injected current drives the membrane potential to a threshold of approximately −50 mV, a population of voltage-dependent Na^+ channels, normally closed at rest, opens abruptly to a large conductance value resulting in a large influx of current. As predicted by eqn [1], this leads to a rapid voltage upswing as the membrane potential spikes toward the Na^+ battery potential of +60 mV. Within a fraction of a millisecond, however, the Na^+ channels begin to close, or 'inactivate,' of their own accord. At the same time, voltage-dependent K^+ channels sense the voltage upswing on the leading edge of the AP and open to produce a large outward current. The combined effects of Na^+ channels closing and K^+ channels opening result in an abrupt downswing in the membrane potential to terminate the AP.

Active electrical responses of the neuronal membrane have historically been most closely associated with the long-range signal transmission function of the axon. It is now apparent, however, that the dendrites of many neuron types contain voltage-dependent channels capable of generating spikes, which are independent of the cell body (Amitai et al., 1993; Golding and Spruston, 1998; Kamondi et al., 1998). Unlike the stereotyped all-or-none form of the classical AP, however, synaptically evoked dendritic spikes are of variable magnitude and can consist of complex superimpositions of fast and slow events. As discussed below, this highly nonlinear electrical behavior of dendritic origin is likely to profoundly influence the integrative properties of nerve cells.

Synaptic Inputs

Synaptic potentials, the principal input signals to which a neuron responds, likewise, are evoked by changes in the conductance of ion channels in the postsynaptic membrane. In this case, the conductance changes are triggered by the binding of neurotransmitter molecules to receptors exposed on the extracellular surface of the membrane. The normal sequence of events in synaptic transmission is as follows. An AP invades the presynaptic axon terminal. The voltage upswing in the presynaptic terminal opens voltage-dependent calcium (Ca^{++}) channels leading to a rapid influx of Ca^{++} ions. A small number of vesicles (usually zero or one) containing neurotransmitters fuse with the presynaptic membrane at specialized release sites and disgorge their contents into the synaptic cleft. The neurotransmitter (e.g., glutamate or GABA) diffuses the several nanometers across the synaptic cleft in a fraction of a millisecond and binds with receptors on the postsynaptic membrane. Upon binding, the receptor–channel complex undergoes a conformational change, causing the channel to open. Alternatively, in the case of 'second messenger' transmission, the protein to which the neurotransmitter binds is not the channel itself, but instead triggers a cascade of biochemical reactions inside the postsynaptic cell, which lead indirectly to the opening or closing of ion channels. In either case, the resulting conductance change(s) cause the membrane potential in the postsynaptic compartment to fluctuate toward the characteristic battery value for that synapse, often called the synaptic 'reversal potential.' The reversal potential at an AMPA- or NMDA-type excitatory synapse is typically 0 mV, 70 mV more positive than the resting potential. This value results from an approximately equal conductance to Na^+ and K^+ ions. A synapse with a reversal potential around 0 mV is termed 'excitatory,' since under normal operating conditions with a negatively polarized membrane potential, activation of the synapse drives the membrane transiently to more positive potentials (i.e., toward 0 mV). This 'depolarization' increases the likelihood (or rate) of outbound AP production in the postsynaptic cell. Conversely, the reversal potential is −65 mV for $GABA_A$-type inhibitory synapses that conduct chloride ions and −90 mV for $GABA_B$-type inhibitory synapses that conduct K^+ ions. In either case, the activation of the synapse opposes synaptic excitation by transiently pulling the membrane potential to more negative values, thus reducing the likelihood (or rate) of AP generation.

Synaptic Integration

To this point, we have seen how transmitter-dependent and voltage-dependent gating of ion channels can produce the fluctuations in membrane potential that underlie both synaptic responses and APs. But what are the rules governing synaptic integration?

The simplest 'linear' model of dendritic integration holds that the neuron sums its excitatory and inhibitory influences across the dendritic arbor, followed by a global thresholding nonlinearity at the cell body (McCullough and Pitts, 1943; Rosenblatt, 1962; Rumelhart and McClelland, 1986). Despite its simplicity (**Figure 1**), this formulation has influenced thinking about neuronal function in both the experimental and computational communities for many years. Extensions to this idea that explicitly include the 'passive cable' properties of the neuron – accounting for attenuation of signals with time and distance (Rall, 1964; Koch et al., 1982; Agmon-Snir and Segev, 1993) – have also been highly influential and have led to a variety of productive interchanges between theory and experiment (Rall et al., 1992).

From the perspective of currently available neurophysiological data, however, much of it derived from direct intradendritic recordings, synaptic integration is likely to be more complex than can be captured by linear or passive models of dendritic function (Segev and Rall, 1998; Mel, 1999). The question is complicated for three reasons, principally relating to uncertainties as to how or to what extent synaptic inputs interact with each other through shared voltage or chemical signals within the dendrites.

First, the spatiotemporal patterns of synaptic activation driving neurons in the CNS *in vivo* are not well known; so that current best guesses as to the electrical and chemical environments in which synapses normally operate remain largely rooted in conjecture.

Second, synaptic inputs may trigger secondary events in the dendrites mediated by voltage-dependent NMDA, Na^+, K^+, and Ca^{++} channels, which could influence significantly how excitatory and inhibitory synaptic inputs are combined to produce overall neuronal responses. Synaptically evoked dendritic spikes have been observed *in vitro* (Golding and Spruston, 1998; Schiller et al., 2000) and *in vivo* (Kamondi et al., 1998), though it is not yet known whether such events occur commonly in awake, behaving animals, and if so whether they are localized to one or many dendritic branches, and so on. In addition, the spatial distributions of these channels on the cell surface, their kinetic properties, their dependence on voltage and other modulatory factors, and their efficacies under normal *in vivo* operating conditions are not sufficiently well known to make firm predictions regarding the contributions of these channels to the integrative processes of the cell.

Third, cable theory (Jack et al., 1975; Koch, 1999) informs us that the dendritic morphology of a typical CNS neuron – typically consisting of many thin-branch subtrees radiating

outward from the cell body and/or main dendritic trunks – is ideally suited to isolate voltage signals within specific subregions of the dendritic arbor. Thus, the electrical coupling between a thin branch and a main trunk (or soma) is expected to produce a significant attenuation of distally generated voltage signals when measured at the trunk, with the most profound suppression of signal transmission reserved for rapid voltage transients associated with fast synaptic potentials or dendritic spikes. Thus, if thin branches, which receive the vast bulk of the excitatory input in most neurons, routinely generate fast spike-like events in response to synaptic excitation, the cable structure of these cells appears designed to ensure that distally generated spikes are seen only in strongly attenuated form in other parts of the cell. Clearly, this type of functional compartmentalization could complicate the issue of synaptic integration, since the overall response of the cell would depend on whether synaptic inputs were delivered to the same or different dendrites, and whether dendritic spikes were generated or not.

Synaptic Plasticity

When a neuron fires an AP, it is propagated from the axon hillock down the axon. But there is often also a back-propagated action potential (bAP) going from the axon hillock back through the dendrites. The bAP produces a feedback signal that provides the entire cell body with information about the output state of the neuron (Magee and Johnston, 2005). The bAPs appear to act to enhance synaptic plasticity by tying together the timing of the Ca^{++} influx of the synaptic input and AP output, provided that they occur nearly coincidently within a tight window of <50 ms. This enables the production of a larger Ca^{++} signal than the input would produce alone and provides a candidate mechanism for aspects of synaptic plasticity. Such a mechanism lends more support to the view that dendrites are key integrative units playing computational roles such as in learning and memory, the mediation of attention or motor output (Behabadi and Mel, 2014; Shah et al., 2010).

Computational Implications

In spite of many uncertainties, currently available experimental data, combined with insights gained from theoretical and modeling studies, seem to support the hypothesis that dendritic trees are compartmentalized functionally and could consist of a moderately large number of, for example, up to 100, independent integrative 'subunits.' The idea that dendritic trees are compartmentalized functionally has been discussed in a variety of forms over the years (Koch et al., 1982; Shepherd et al., 1985; Rall and Segev, 1987; Mel, 1992a,b, 1993; Mel et al., 1998; Archie and Mel, 2000; Poirazi and Mel, 2001). In the continued absence of empirical evidence sufficient to confirm or deny this 'nonlinear subunit' hypothesis, several recent modeling studies have provided support for one simple version of the idea. According to this model, each dendritic subunit, corresponding to a single branch or small thin-branch subtree, acts as a separately thresholded neuron-like summing unit, and all the subunit responses are summed in the main trunks and the cell body to produce the cell's final output (**Figure 1**). This formulation ignores differences in the locations of synapses within a subunit, as justified by theoretical considerations and biophysical modeling studies, indicating that voltage signals are communicated far more effectively within a dendritic branch than between branches – especially for branches separated by a main trunk or the cell body. In mathematical terms, the neuron's overall output y is given by a sum of m independently thresholded subunit responses. Each subunit response is given by a weighted sum of k inputs followed by a fixed subunit nonlinearity, g, which is typically modeled as a sigmoid or power function

$$y = \sum_{j-1}^{m} g\left(\sum_{i-1}^{k} w_{ij}x_{ij}\right) \qquad [2]$$

This simple formula, which is similar to the expression describing a conventional two-layer neural network (Bishop, 1995), was found to predict the time-averaged firing rates produced by a biophysically detailed compartmental model of a pyramidal cell, a model which included passive membrane properties, a spatially extended dendritic morphology, synapses with realistic conductance models, and voltage-dependent ion channels in the dendrites (Archie and Mel, 2000).

From a functional perspective, it is interesting to note that the sum-of-thresholded-subunits model captures the underlying structure of a variety of classical and extraclassical receptive field properties found in cortical neurons, suggesting that dendritic subunits could contribute in important ways to the many quotidian signal-processing functions of cortical tissue (Mel, 1999). In a different vein, a recent modeling study has shown that subunit-containing neurons as described in eqn [2] could increase the trainable storage capacity of neural tissue by orders of magnitude (Poirazi and Mel, 2001).

While it is clear that many questions remain to be answered regarding the information processing functions of single neurons, one may be confident that with the steady advance of experimental methods available to study neuronal function at the required level of detail, coupled with the enormous flexibility of hypothesis testing provided by the computer modeling enterprise, the many fascinating secrets of neural information processing will gradually yield in the coming years.

See also: Computational Neuroscience; Ion Channels and Molecular Events in Neuronal Activity; Neural Plasticity.

Bibliography

Agmon-Snir, H., Segev, I., 1993. Signal delay and input synchronization in passive dendritic structures. Journal of Neurophysiology 70, 2066–2085.

Amitai, Y., Friedman, A., Connors, B., Gutnick, M., 1993. Regenerative electrical activity in apical dendrites of pyramidal cells in neocortex. Cerebral Cortex 3, 26–38.

Archie, K.A., Mel, B.W., 2000. An intradendritic model for computation of binocular disparity. Nature Neuroscience 3 (1), 54–63.

Behabadi, B.F., Mel, B.W., 2014. Mechanisms underlying subunit independence in pyramidal neuron dendrites. Proceedings of the National Academy of Sciences USA 111, 498–503.

Bishop, C., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford, UK.

Golding, N.L., Jung, H.-Y., Mickus, T., Spruston, N., 1999. Dendritic calcium spike initiation and repolarization are controlled by distinct potassium channel sybtypes in CA1 pyramidal neurons. Journal of Neuroscience 19, 8789–8798.

Golding, N.L., Spruston, N., 1998. Dendritic sodium spikes are variable triggers of axonal action potentials in hippocampal CA1 pyramidal neurons. Neuron 21, 1189–1200.

Jack, J., Noble, D., Tsien, R., 1975. Electric Current Flow in Excitable Cells. Clarendon Press, Oxford, UK.

Kamondi, A., Acsády, L., Buzsáki, G., 1998. Dendritic spikes are enhanced by cooperative network activity in the intact hippocampus. Journal of Neuroscience 18, 3919–3928.

Koch, C., 1999. Biophysics of Computation. Oxford University Press, New York.

Koch, C., Poggio, T., Torre, V., 1982. Retinal ganglion cells: a functional interpretation of dendritic morphology. Philosophical Transactions of the Royal Society of London B 298, 227–264.

Lübke, J., Markram, H., Frotscher, M., Sakmann, B., 1996. Frequency and dendritic distribution of autapses established by layer 5 pyramidal neurons in the developing rat neocortex: comparison with synaptic innervation of adjacent neurons of the same class. Journal of Neuroscience 16, 3209–3218.

Magee, J.C., Johnston, D., 2005. Plasticity of dendritic function. Current Opinion in Neurobiology 15, 334–342.

McCullough, W., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5, 115–133.

Mel, B.W., 1992a. The clusteron: toward a simple abstraction for a complex neuron. In: Moody, J., Hanson, S., Lippmann, R. (Eds.), Advances in Neural Information Processing Systems, vol. 4. Morgan Kaufmann, San Mateo, CA, pp. 35–42.

Mel, B.W., 1992b. NMDA-based pattern discrimination in a modeled cortical neuron. Neural Computation 4, 502–516.

Mel, B.W., 1993. Synaptic integration in an excitable dendritic tree. Journal of Neurophysiology 70 (3), 1086–1101.

Mel, B.W., 1999. Why have dendrites? A computational perspective. In: Stuart, G., Spruston, N., Haumlusser, M. (Eds.), Dendrites. Oxford University Press, Oxford, UK, pp. 271–289.

Mel, B.W., Ruderman, D.L., Archie, K.A., 1998. Translation-invariant orientation tuning in visual 'complex' cells could derive from intradendritic computations. Journal of Neuroscience 17, 4325–4334.

Poirazi, P., Mel, B.W., 2001. Impact of active dendrites and structural plasticity on the memory capacity of neural tissue. Neuron 29, 779–796.

Rall, W., 1964. Theoretical significance of dendritic trees for neuronal input-output relations. In: Reiss, R. (Ed.), Neural Theory and Modeling. Standard University Press, Stanford, CA, pp. 73–97.

Rall, W., Burke, R., Holmes, W., Jack, J., Redman, S., Segev, I., 1992. Matching dendritic neuron models to experimental data. Physiological Reviews 72 (4), S159–S186.

Rall, W., Segev, I., 1987. Functional possibilities for synapses on dendrites and on dendritic spines. In: Edelman, G., Gall, W., Cowan, W. (Eds.), Synaptic Function. Wiley, New York, pp. 605–636.

Rosenblatt, F., 1962. Principles of Neurodynamics. Spartan, Washington.

Rumelhart, D., McClelland, J., 1986. Parallel Distributed Processing. MIT Press, Cambridge, MA.

Schiller, J., Major, G., Koester, H.J., Schiller, Y., 2000. NMDA spikes in basal dendrites of cortical pyramidal neurons. Nature 404, 285–289.

Segev, I., Rall, W., 1998. Excitable dendrites and spines – earlier theoretical insights elucidate recent direct observations. Trends in Neuroscience 21, 453–460.

Shah, M.M., Hammond, R.S., Hoffman, D.A., 2010. Dendritic ion channel trafficking and plasticity. Trends in Neurosciences 33, 307–316.

Shepherd, G., 1998. The Synaptic Organization of the Brain. Oxford University Press, Oxford, UK.

Shepherd, G.M., Brayton, R.K., Miller, J.P., Segev, I., Rinzel, J., Rall, W., 1985. Signal enhancement in distal cortical dendrites by means of interactions between active dendritic spines. Proceedings of the National Academy of Science USA 82, 2192–2195.

Stuart, G., Spruston, N., Haumlusser, M., 1999. Dendrites. Oxford University Press, Oxford, UK.

Neurophysiological Markers of Language Development

Makiko Kaga, National Institute of Mental Health, National Center of Neurology and Psychiatry, Kodaira, Tokyo, Japan; and Tokyo Metropolitan Medical Center East for Children/Adults with Developmental Disabilities, Tokyo, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The development of language from neurophysiological aspects using event-related potentials (ERPs) are described. Mismatch negativities (MMN), P300, and N400 are the classification of ERPs based on the time course. MMN and P300 represent, respectively, passive perception and active discernment of the difference in several kinds of stimuli, especially pairs of tone bursts (TB) and verbal sounds (VS). N400 is related to the semantic processing of the language. The author will describe about developmental changes of MMN, P300, and N400 potentials as the neurophysiological markers in language development, which are highly related to the recognition of VS and the semantic meaning of the words.

Introduction

Language is the visible or invisible symbol of internal process of constructing idea, learning anything, and processing any perceptive or emotional stimulus. It is also an important medium for communicating with others. Markers of language development have historically been studied mainly from the viewpoint of developmental milestones by many psychologists, speech pathologists, and developmental pediatricians. Expressive and receptive abilities are usually the target of explorations of language evaluation in children, along with the development of intelligence, vision, hearing, social interactions, and so on. Especially, child's age at the utterance of the first word and sentence, the increment curve of vocabulary, and the capability of obeying oral commands are often the developmental markers of expressive or receptive language. Those topics are discussed precisely in other chapters.

Therefore, in this article, the development of language from neurophysiological aspects using event-related potentials (ERPs) will be described. Electroencephalographic procedures including evoked potentials (EPs) and ERPs can reveal the temporal resolution of the cognitive information processing. This kind of quantitative and objective measures of evaluation in language development will surely add some useful information about our understanding of the developing brain.

ERPs are recorded as computer processed electroencephalographic activities elicited by certain internal and external cognitive affairs. This cognition includes perception, attention, discrimination, judgment, expectation, and so on. ERPs are classified by time course, the sensory modalities, and passive/active attitude of subjects. Mismatch negativities (MMN), P300, and N400 are the classification based on time course. MMN usually appears around 200 ms after the stimuli presentation and represents passive perception of the difference in several kinds of stimuli. P300 is recorded about 300–400 ms after the beginning of tasks and represents active discernment of the same kinds of stimuli described above. N400 is elicited about 400–600 ms and is related to the semantic processing of the language. The origin of each ERP has been explored extensively and is also a subject of research. ERPs are applied to the typically developed subjects as well as patients with various neurological/psychiatric diseases and with developmental delay to reveal their cognitive function. In this article, the author will describe about developmental changes of MMN, P300, and N400 potentials, which are highly related to the recognition of verbal sounds (VS) and the semantic meaning of the words.

MMN

MMN was first reported by Naatanen in 1987. MMN is generated by automatic response to any change in outer sensory stimulation exceeding a certain limit of corresponding roughly to the behavioral discrimination threshold. MMN is a negative displacement at the fronto-central electrodes relative to reference electrodes in the difference wave elicited as the subtraction wave of frequent (standard) from rare (deviant) stimuli. Research in MMN has mainly been focused in the response elicited by auditory stimuli. Most studies have been done using tone bursts (TB) at the frequencies of 1000 and 1200 Hz with 60 to 70 dBHL in intensity. However, the difference wave induced by visual stimuli also disclosed the presence of MMN in a similar pattern (Horimoto et al., 2002).

Language development is highly influenced by the hearing acuity and understanding of spoken language. Consequently, researches using auditory stimuli are very important. The first gate of speech development is to differentiate VS unconsciously then consciously. Development of phonological perception has shown that even infants at the age of 1 month can differentiate (ba) (pa) sounds, as revealed by habituation methods (Eimas et al., 1971). Moreover, (da) (ta) sounds in infants (Novak et al., 1989) could be differentiated by auditory cortical response or (l) (y) sounds (Cheour-Luthtanen et al., 1995) in neonates of 1–4 days old could elicit mismatch negativity.

As to MMN to TB, Fuchigami (Fuchigami et al., 1995) examined the results of ERP of ignoring condition in 108 subjects aged 1 to 21 years old. They made clear that their latency was decreased as the subjects grew older. In the same year, Woods (1992) reported the middle aged and elderly adults showed the MMN using TB with the latency of 150–210 ms with reduced amplitude in the elderly subjects.

Kraus et al. (1992) reported that the MMN induced by speech sounds, (ba) and (da) in children aged 7 to 11 years old

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23135-8

in 1992. This potential was also confirmed by magnetoencephalography in 11 children aged 5–7 years old. MMN induced by TB stimuli and VS stimuli TB and VS develops independently but similarly (Pihko et al., 2005). Moreover, the latency of MMN to VS is longer than that of to TB. This suggests that the VS has more complex components than TB and need to be trained to acquire and recognize. We can only speculate the difficulty in distinguishing complex combination of words in the developing brain.

P300

P300 was first reported by Sutton et al. (1965) and has been widely applied to the patients with developmental disorders (Niwa et al., 1983; Courchesne et al., 1984; Kileny et al., 1997) and neurological/psychiatric diseases.

P300 is a response to represent the internal process of stimulus perception, discrimination, and judgment. Subjects must differentiate two or more different kinds of stimuli presented as the oddball paradigm and must choose the target stimuli. P300 appears clearly to the target stimuli after about 300 ms of stimulus presentation with centro-parietal dominancy. To make the best combination of the stimuli is very important to evaluate the ability of discrimination of the target stimulus.

Historically, most studies concerning the development of P300 were performed using nonverbal stimuli such as TB of 1000 and 2000 Hz, and rarely using verbal stimuli (Gordin et al., 1978; Polish et al., 1985; Martin et al., 1988). Concerning the cognitive development of the primary active discrimination of VS, P300 induced by TB stimuli should be compared with that of VS. Of course, to explore the cognition higher than that of verbal sound level, usage of words and sentences as the tasks could be ideal. However, P300 cannot be elicited by these tasks with such a long duration like sentences. As to the P300 induces by VS has been reported that the latency to VS was longer than to TB in adults and in children (Groenen et al., 2001; Hatori et al., 2004). This finding suggested that longer time was necessary to process multiple and complex frequency sounds (VS) rather than a single frequency sound (TB). Previous reports on the developmental change of P300 elicited by TB showed the inclination curve was calculated as the linear function (Gordin et al., 1978; Polish et al., 1985; Pearce et al., 1989). These approximations were elicited by the limited results in studies of children.

In 2004, a report compared results for P300 induced by TB and VS as an oddball paradigm in the same persons of a variety of age was published (Hatori et al., 2004; **Figure 1**; **Figure 2**). Their subjects, all right-hand-dominant native Japanese speakers, were divided into two groups; which one group consisted of 19 adults (24.3 ± 3.0 years old) and the other consisted of 13 children (10.0 ± 2.4 years old). Auditory oddball TB tasks were 1 kHz and 750 Hz in frequency, while VS tasks were [ae] and [a]. [ae] sound was the target and [a] sound was nontarget. [ae] is not present in Japanese pronunciation but [a] is present in Japanese pronunciation. None of the subjects made any mistakes in pressing key. Especially, the latency of P300 induced by VS was longer than that induced by TB in both groups. The dominant area of P300 was Pz in both children and adults (**Figure 3**). The peak latency of P300 (Y ms) and age (X years old) were approximated as $Y = 0.5071X^2 - 21.056X + 516.2$ for TB and $Y = 0.5683X^2 - 26.744X + 663.32$ for VS (**Figure 4**). Therefore, the age with the shortest latency of P300 to TB was 20.8 years old and was 23.6 years old to VS. Compared to TB, VS had a stronger inclination of the diminution speed of latency at 10 years old, and the latency of VS was longer than that of TB. This shows the learning process of differentiation of VS from TB progresses steadily and VS processing speed is faster than that of TB.

Thus the analysis of successive age subjects made clear the age of the shortest latency to TB and VS. Namely, the difference

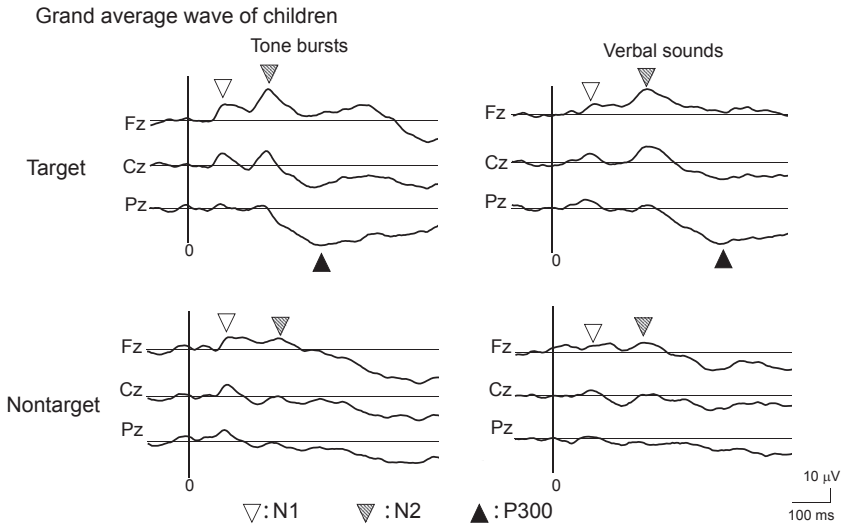

Figure 1 Grand average wave of children ($n = 13$), ▽: N1, ▼: N2, ▲: P300. Cited and modified from Hatori, T., Inagaki, M., Shirane, S., Kaga M., 2004. Clinical neurophysiological study of verbal and non verbal sound perception: normal development of P300 event related potential to different sound stimuli (in Japanese). No to Hattatsu 36, 232–239.

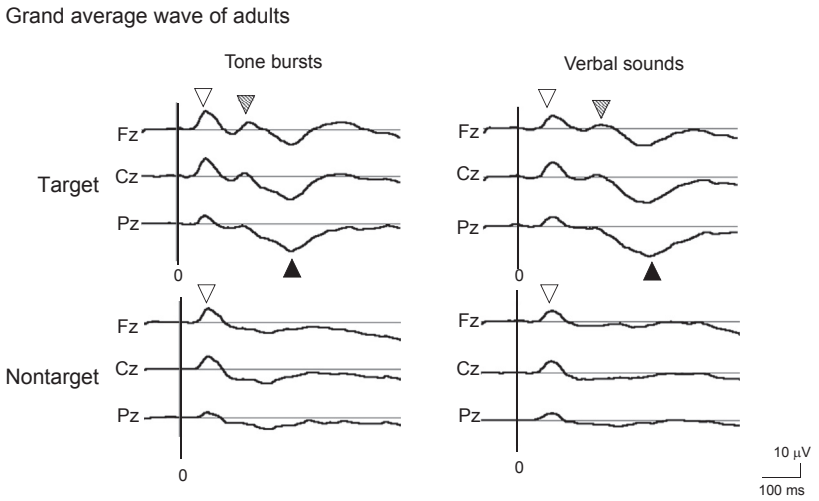

Figure 2 Grand average waves of P300 in adults ($n = 19$), ▽: N1, ▼: N2, ▲: P300. Cited and modified from Hatori, T., Inagaki, M., Shirane, S., Kaga M., 2004. Clinical neurophysiological study of verbal and non verbal sound perception: normal development of P300 event related potential to different sound stimuli (in Japanese). No to Hattatsu 36, 232–239.

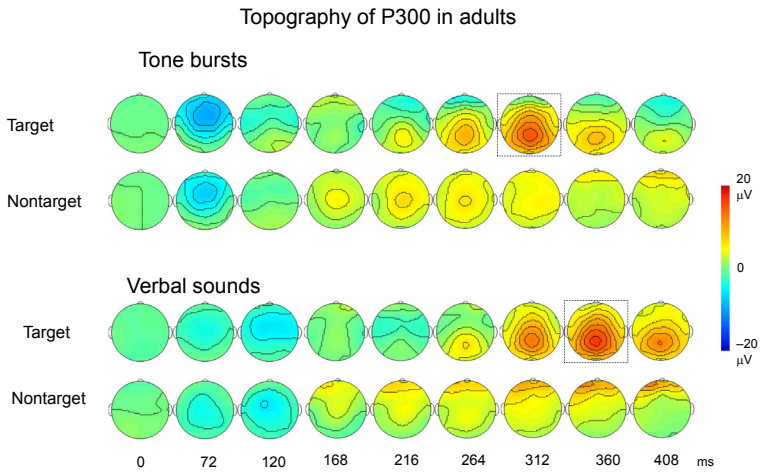

Figure 3 Topography of P300 in adults. (a) N1 peak to tone burst as target and nontarget stimuli appears at 72 ms at Fz and Cz dominancy. P300 appears only to the target stimuli with dominancy of Pz. (b) N1 peak to verbal sound as target and nontarget stimuli appear at 120 ms at Fz to Cz dominancy. P300 appears only to the target stimuli with dominancy of Pz. Cited and modified from Hatori, T., Inagaki, M., Shirane, S., Kaga M., 2004. Clinical neurophysiological study of verbal and non verbal sound perception: normal development of P300 event related potential to different sound stimuli (in Japanese). No to Hattatsu 36, 232–239.

was 3 years old. However, approximation curve revealed they were independent and never crossed over. Development of P300 latency reflects qualitatively different cognitive process. These results provide developmental process of perception and discrimination of TB and VS.

N400

N400 was first reported by Kutas and Hillyard (1980) and have been investigated for years by many researchers (Monetta, 2003; Kutas et al., 1994; Jouttonen et al., 1996; Yano et al., 1999; Yano and Kaga, 2000; Coch et al., 2002; Bonte and Blomert, 2004). N400 of ERP reflects linguistic process and has been studied with a semantic priming paradigm with lexical decision tasks (Holcomb, 1993; Kutas, 1993). For example, subjects were asked to judge whether a target word was a word or nonword, whether a presented sentence made sense or not, or whether a presented prime word was related to the target word or not. Many studies have been done applying to typical developed adults to solve these tasks in linguistics. So, it is often difficult to apply these tasks to children.

Despite many investigations on N400, studies using the same tasks in the different modality to the same subjects have

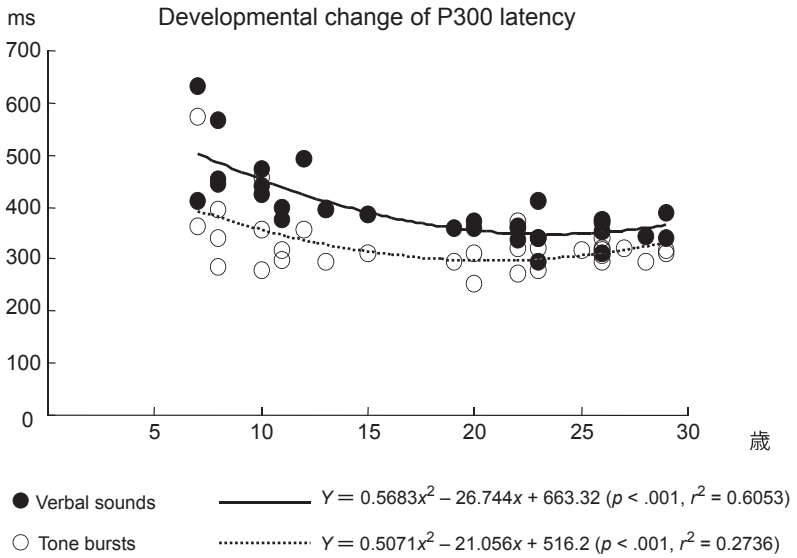

Figure 4 Developmental change of P300. The relation of ages (X years old) and P300 peak latency (Y ms) showed the below function. ○: tone bursts $Y = 0.5071X^2 - 21.056X + 516.2$ ($p < .001$, $r^2 = 0.2736$), ●: verbal sounds $Y = 0.5683X^2 - 26.744X + 663.32$ ($p < .001$, $r^2 = 0.6053$). Cited and modified from Hatori, T., Inagaki, M., Shirane, S., Kaga M., 2004. Clinical neurophysiological study of verbal and non verbal sound perception: normal development of P300 event related potential to different sound stimuli (in Japanese). No to Hattatsu 36, 232–239.

not been done. (Oana et al., 2006) made a new task called the semantic category decision task. This task is rather simple, so children who can read only kana letters (Japanese phonetic letters) can accomplish it. They provided three types of tasks; namely auditory, visual, and auditory–visual (bimodal), with which examined the developmental change of N400.

Their subjects were 38 typical developed children and adults aged 6–30 years old who were divided into three groups, 18 younger age group (6–9 years of age, mean standard deviation was 7.3 ± 1.1), 10 elder age group (10–14 years old, 11.6 ± 1.3), and 10 adults (22–30 years old, 26.6 ± 2.7).

In the semantic category decision task, a prime word was living thing or nonliving living. The target word was the name of concrete thing classified as living (for example, a cat, dog, or goldfish, and so on) or nonliving (for example, a pencil, desk, spoon, and so on). The target word was followed the prime word and the subject judged whether it was a living thing or nonliving thing. In this research, target words were based on the expressive vocabulary of younger children edited by National Institute for Japanese Language and Linguistics. Therefore, we can reasonably speculate about and discuss the development of semantic understanding of words in children of elementary school age.

The above results were shown in the figure (Figure 5). The most characteristic point was the different wave distribution and wave pattern among different age groups in each sensory modality. As for the distribution, N400 induced by auditory modality extends from frontal to parietal distribution in younger children while in adults it was limited to frontal distribution. We can speculate that in younger children group need a lot of energy from the broad area of brain to process phonetic and semantic processing (Figure 5(a)). As for the wave pattern (Figure 5(d)), it is monophasic in younger children but it is biphasic in the older children group, which is similar to that of the adult group. Moreover, the latency was almost the same as in older children and the adults. Thus semantic processing through phonetic route should be matured by the age of 10–14 years.

Contrary to the auditory modality, the wave form of visual N400 (Figure 5(b) and 5(e)) changed from biphasic in the younger children group to monophasic in the older children group. The second peak of negative deflection in younger children group can be thought to reflect the process, whether the word is congruent or incongruent to the target word. The peak latency of the second peak of younger age group was almost the same as the auditory stimulus-induced single peak latency. This finding suggests that the processing time course after recognition of semantic categories of words in the younger age group was almost the same between auditory modality and in visual modality. However, the simultaneous presentation of auditory and visual stimulus (Figure 5(c) and 5(f)) significantly shortened the peak latency of N400 (450 ms at Fz) compared to that with the single modality. Moreover, the response time in the bimodal condition also was shorter than that with single modality. Therefore, in younger age group, double-stimulus presentation facilitates the semantic processing of words. In the visual modality, the older age group and the adult group showed almost the same wave patterns and the same latencies as each other. This means that visual modality also develops after 10 years of age for the semantic processing. Moreover, the older age group showed the same wave patterns of visual and bimodal wave pattern. Thus in the older age group, visual dominancy is speculated than auditory modality in semantic processing. Robinson and Sloutsky (2004) and Napolitano and Sloutsky (2004) reported that infants under 4 years of age to whom simultaneous auditory and visual stimuli were presented, auditory route was dominant; vice versa, in the adults, visual route dominancy was shown. Thus, in the development of language acquisition, there seems to be a time of transient auditory dominant age.

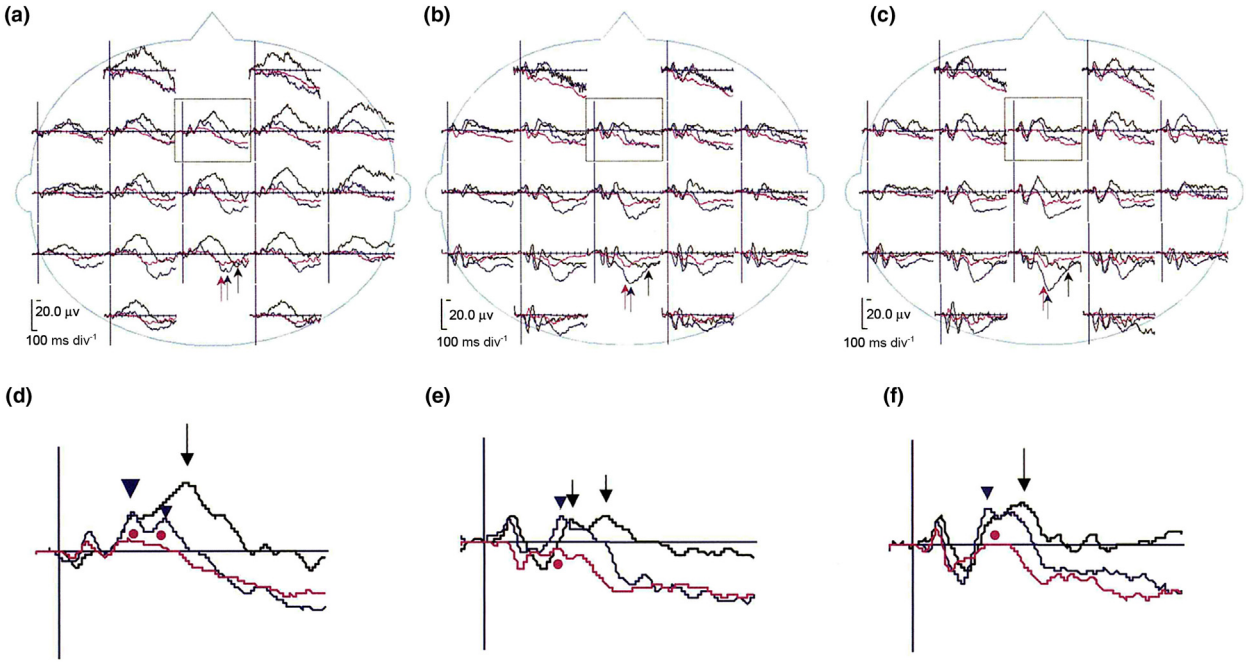

Figure 5 Developmental change of N400 in three modalities. Black line is the grand average pattern of the younger children group ($n = 18$) at Fz. Blue line is the grand average pattern of the elder children group at Fz. Red line is the grand average pattern of the adult group at Fz. (a) and (d) Auditory modality, Younger children group showed monophasic pattern (↓). Elder children group showed biphasic pattern (▼). Adult children group showed low amplitude pattern (●). (b) and (e) Visual modality. Younger children group showed biphasic pattern (↓), Elder children group (▼) and adult group (●) showed almost monophasic pattern. (c) and (f) Auditory–visual modality. The wave shape of younger children (↓) is similar to that of the auditory modality. The wave shapes of the elder children group (▼) and adult group (●) are similar to those of the visual modality. Cited and modified from Oana, S., Inagaki, M., Suzuki, S., Horimoto, R., Kaga, M., 2006. Developmental changes of N400 event-related potential of a semantic category decision task and modality specific findings in patients with developmental dyslexia. No to Hattatsu 38, 431–438.

Summary

In this article, the neurophysiological markers of language development were described explaining the developmental change of MMN, P300, and N400. These parameters can be compared with the abnormal or deviated development of language.

See also: Biological Preconditions for Language Development; Child-Directed Speech: Influence on Language Development; First Language Acquisition, Developmental Psychology of; Human Language, Evolution of; Language Development, Theories of; Language Development: Emergentist Theories; Language Development: Evolutionary Theories; Language Development: Influence of Socio-Economic Status; Multilingual Language Development.

Bibliography

Bonte, M., Blomert, L., 2004. Development changes in ERP correlate of spoken word recognition during early school years: a phonological priming study. Clin Neurophysiol 115, 409–423.

Cheour-Luthtanen, M., Alho, K., Kujala, T., Sainio, K., Reinikainen, K., Renlund, M., Aaltonen, O., Eerola, O., Näätänen, R., 1995. Mismatch negativity indicates vowel discrimination in newborns. Hearing Res 82, 53–58.

Coch, D., Maron, L., Wolf, M., Holcomb, W.P.J., 2002. Word and picture processing in children: an event-related potential study. Develop Neuropsychol 22, 373–406.

Courchesne, E., Kilman, B.A., Galambos, R., Lincoln, A.J., 1984. Autism: processing auditory information assessed by event-related brain potentials. Electroencephalogr Clin Neurophysiol 59, 238–248.

Eimas, P.D., Siqueland, E.R., Jusczyk, P., Vigorito, J., 1971. Speech perception in infants. Science 171, 303–306.

Fuchigami, T., Okubi, O., Fujita, Y., Kohira, R., Noguchi, Y., Fuchigami, S., Hiyoshi, K., Nishimura, A., Harada, K., 1995. Developmental changes in P300 wave elicited during two different experimental conditions. Pediatr Neurol 13, 25–28.

Gordin, D.S., Squires, K.C., Henderson, B.H., Starr, A., 1978. Age-related variations to auditory stimuli in normal human subjects. Electroencephalogr Clin Neurphysiol 44, 447–458.

Groenen, P.A.P., Beynon, A.J., Snik, A.D.F.M., van den Broek, P., 2001. Speech-evoked cortical potentials and speech recognition in cochlear implant users. Scandinav Audiol 30, 31–40.

Hatori, T., Inagaki, M., Shirane, S., Kaga, M., 2004. Clinical neurophysiological study of verbal and non verbal sound perception: normal development of P300 event related potential to different sound stimuli (in Japanese). No to Hattatsu 36, 232–239.

Holcomb, P.J., 1993. Semantic priming and stimulus degradation: implications for the role of the N400 in language processing. Psychophysiol 30, 47–61.

Horimoto, R., Inagaki, M., Yano, T., Sata, Y., Kaga, M., 2002. Mismatch negativity of color modality during a selective attention task to auditory stimuli in children with mental retardation. Brain Dev 24, 703–709.

Jouttonen, K., Revonsuo, A., Lang, H., 1996. Dissimilar age influences on two ERP waveforms (LPC and N400) reflecting semantic context effect. Brain Research. Cognitive Brain Res 4, 99–107.

Kileny, P.R., Boerst, A., Zwolan, T., 1997. Cognitive evoked potentials to speech and tonal stimuli in children with implant. Otolaryngol Head Neck Surg 117, 161–169.

Kraus, N., McGee, T., Sharma, A., Carrell, T., Nicol, T., 1992. Mismatch negativity event-related potential elicited by speech stimuli. Ear Hear 13, 158–164.

Kutas, M., Hillyard, S.A., 1980. Reading Senseless sentences: brain potentials reflect semantic incongruity. Science 207, 203–205.

Kutas, M., 1993. In the company of other words: electrophysiological evidence for single-word and sentence context effects. Lang Cogn Processes 8, 533–572.

Kutas, M., Steven, A., Hillyard, S.A., 1994. Brain potentials during reading reflect word expectancy and semantic association. Nature 307, 161–163.

Martin, L., Barajas, J.J., Fernandez, R., Torres, E., 1988. Auditory event-related potentials in well-characterized groups of children. Electroencephalogr Clin Neurophysiol 71, 375–381.

Monetta, L., 2003. Semantic processing of words, cognitive resources and N400: an event-related potentials study. Brain Cogn 53, 327–339.

Naatanen, R., Paavilainen, P., Alho, K., Reinikainen, K., Sams, M., 1987. The mismatch negativity to intensity changes in auditory stimulus sequence. Electroencephalogr Clin Neurophysiol (Suppl.) 40, S125–S131.

Napolitano, A.C., Sloutsky, V.M., 2004. Is a picture worth a thousand words? The flexible nature of modality dominance in young children. Child Develop 75, 1850–1870.

Niwa, S., Ohta, M., Yamazaki, K., 1983. P300 and stimulus evaluation process in autistic subjects. J Aut Develop Disrd 13, 33–42.

Novak, G.P., Kurtzberg, D., Kreuzer, J.A., Vaughan Jr., H.G., 1989. Cortical responses to speech sounds and their formants in normal infants: maturational sequence and spatiotemporal analysis. Electroencephalogr Clin Neurophysiol 73, 295–305.

Oana, S., Inagaki, M., Suzuki, S., Horimoto, R., Kaga, M., 2006. Developmental changes of N400 event related potential of semantic category decision task and modality specific findings in patients with developmental dyslexia (in Japanese). No to Hattatsu 38, 431–438.

Pearce, J.W., Crowell, D.H., Tokioka, A., Pacheo, G.P., 1989. Childhood developmental changes in the auditory P300. J child Neurol 4, 100–106.

Pihko, E., Kujala, T., Mickos, A., Antell, H., Alkuf, P., Byring, R., Korkman, M., 2005. Magnetic fields evoked by speech sounds in preschool children. Clin Neurophysiol 116, 112–119.

Polish, J., Howard, L., Starr, A., 1985. Effect of age on the P300 component of the event-related potential from auditory stimuli: peak definition, variation, and measurement. J Gerontol 40, 721–726.

Robinson, C.W., Sloutsky, V.M., 2004. Auditory dominance and its change in the course of development. Child Develop 75, 1387–1401.

Sutton, S., Braren, M., Zubin, J., John, E.R., 1965. Evoked-potential correlates of stimulus uncertainty. Science 150, 1187–1188.

Woods, D.L., 1992. Auditory selective attention in middle-aged and elderly subjects: an event-related brain potential study. Electroencephalogr Clin Neurophysiol 84, 456–468.

Yano, T., Kaga, M., Uno, A., 1999. Semantic categorical relations on N400. Tohoku Psychol Folia 58, 91–98.

Yano, T., Kaga, M., 2000. Semantic category discrimination and N400. Psychol Reports 87, 415–422.

Neuroprosthetics

Mounia Ziat, Northern Michigan University, Marquette, MI, USA
Sliman Bensmaia, University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, we present an overview of the current state of the art in neuroprosthetics, also known as brain–computer interfaces or brain–machine interfaces. First, we describe afferent neuroprosthetics, which are designed to convey sensory input through stimulation of neuronal populations that are implicated in perception. We then describe efferent interfaces that consist of using neuronal signals to control extra-corporeal devices.

Overview

Neuroprosthetics are devices that interface with the central nervous system (CNS) or the peripheral nervous system (PNS). While neuroprosthetics were first devised in the 1960s, recent technological advances have engendered important developments in this field, not only for neurorehabilitation, but also for gaming, virtual reality, and decision making. Neuroprosthetics can be classified into two main categories: afferent neuroprosthetics that write information into the brain, usually to provide sensory input and efferent neuroprosthetics that read information out of the brain to control an external device. These two categories can be further broken down into several subcategories that are usually defined by the site of stimulation for input brain–computer interface (BCI) and the type of response/command used for control in the case of output BCI.

Afferent Neuroprosthetics or Input BCI

Afferent neuroprosthetics consist of conveying sensory information by stimulating neuronal populations in the PNS or CNS. The oldest and most successful afferent neuroprosthesis to date is the cochlear implant (CI). Since the development of the CI, however, afferent neuroprostheses have expanded into other sensory modalities, and have recently begun to enter the cognitive domain as well.

Sensory Neuroprostheses

Auditory Neuroprostheses

Cochlear Implant

Children and adults with severe sensorineural hearing loss who cannot be helped by hearing aids are good candidates for CIs, as long as the auditory nerve is not damaged (**Figure 1**). Perceptive or sensorineural hearing loss is due to damage or loss of hair cells, which are the mechanoreceptors that mediate hearing, or damage to the cochlea (the structure in which these hair cells reside), and constitute a distinct condition from conductive hearing impairment loss, which affects the middle ear. Successful implementation of a CI depends on the age at onset, duration of deafness, the age at implantation, the etiology, and the implant model (NIH, 1995). Another important factor is whether the onset of hearing loss occurred before or after the development of a language. Indeed, while postlingually deaf adults experience an improvement in the perception of speech through their CI, prelingually deaf adults typically do not experience such a benefit. Nonetheless, noisy environments impair the ability of all CI patients to perceive speech, listen to music, or talk on the phone (Peterson et al., 2010). There is no age limit for a CI, but implantation before the age of 2 years tends to reduce the negative outcomes of auditory loss and allows for a more efficient acquisition of speech and language (NIH, 1995).

A CI comprises an internal component (the receiver), which consists of a module, placed beneath the skin, which contains a receiving antenna and an array of 8–22 electrodes. The electrode array is placed in the inner ear (cochlea) and stimulates the auditory nerve directly. The external part of the CI consists of a microphone that picks up surrounding sounds, a chip to process the information from the microphone, and a transmitter that sends signals to the receiver.

The surgery is performed under general anesthesia and typically lasts between 3 and 4 h. The patient lies supine with the head turned to the side. An incision is made behind the ear to allow the surgeon access to the layers of tissue between the

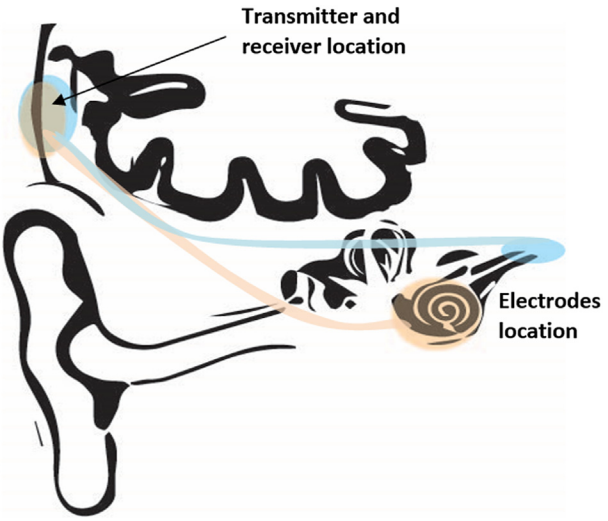

Figure 1 Location of cochlear implants (light orange) and brainstem implants (blue).

skin and the bone. A small cavity is drilled into the mastoid (the bone behind the ear) and the receiver is inserted into it. Another cavity is then drilled to access the inner ear and the auditory nerve, where the array of electrodes, shaped like a spiral to match the cochlea, is then inserted. About 3 weeks after the operation, the external component is installed and the device is calibrated. Calibration sessions consist of characterizing both perceptual thresholds and pain/discomfort levels. Then, different processing strategies are tested to determine which is best suited to the CI patient.

While single-channel CIs had been previously implanted, the first multichannel CI, known as the bionic ear, was implanted on 1 August 1978 by Professor Graeme Clark and his team at the University of Melbourne in Australia. The ten-channel implant recipient was Rod Saunders, a 48-year-old man who lost his hearing 18 months before the operation following an accident that damaged his cochlea. This implant led to several important discoveries about sound and speech processing (Clark, 2003).

Generally, the more electrodes or channels in a CI, the clearer is the perceived sound. While a single-channel implant can provide some conscious experience of sound, at least 6–8 channels are required for speech recognition. The number of implanted electrodes depends on the number of remaining hair cells in the cochlea. Commercial CIs have up to 22 electrodes. As might be expected, given the relatively small number of channels, CI-elicited sounds are far from natural.

Auditory Brainstem Implant

Patients who cannot benefit from a CI due to the absence of or damage to the auditory nerve can potentially benefit from auditory brainstem implants (ABIs), which bypass the auditory nerve and stimulate the brain directly. In most cases, patients suffer from neurofibromatosis type II (NF2), an inherited disease that requires the extraction of a nonbenign tumor in the vestibulocochlear nerve (cranial nerve VII), and thus of the auditory nerve itself. The design principle is similar to that of CIs, except that the electrode array is implanted in the cochlear nucleus in the brainstem (Rauschecker and Shannon, 2002).

The first ABI was implanted in an adult at the House Ear Institute, CA, in 1979 and consisted of two electrodes that stimulated the auditory brainstem. To date, about 900 adults worldwide have received an ABI. The Food and Drug Administration (FDA) approved clinical trials on children in January of 2013 and 3-year-old Grayson Clamp became the first child in the United States to receive the ABI in May of that year. The outcomes after the implantation are usually positive, as the patient is able to hear surrounding sounds and process speech during face-to-face communication. However, ABIs never reach the speech recognition level obtained with CIs for patients with NF2. Current ABIs consist of 10–12 electrodes that deliver electrical pulses to the surface of the cochlear nucleus. Colletti's team at the University of Verona developed an ABI with 8–10 penetrating electrodes (PABI) that provides better pitch sensations (Colletti et al., 2004; http://www.sciencedirect.com/science/article/pii/S019459980500344X) with reduced currents, although it does not produce a greater improvement in speech recognition than do standard ABIs (Otto et al., 2008; http://www.ncbi.nlm.nih.gov/pubmed/18931643). The level of speech recognition achieved with ABIs is equivalent to that of the first single-channel CIs.

Individuals who are not afflicted with NF2 were originally not eligible for an ABI because of the surgical risks. Indeed, in NF2 patients, the ABI is implanted just after the tumor removal and therefore does not require an additional surgery. With that proscription lifted, Colletti et al. showed that ABI patients without tumors exhibit better speech recognition than do ABI patients with tumors. In fact, their speech recognition is comparable to that of multichannel CI recipients. Importantly, as the surgical procedures, electrodes, and location are the same for ABI patients with and without tumor, it is likely that the removal of the NF2 tumor damages neural tissue that is essential to speech processing (Shannon, 2010). For children, the critical age for an ABI is the same as that for a CI.

Auditory Midbrain Implant

The fact that NF2 ABI patients' speech recognition does not improve significantly compared to that of non-NF2 ABI recipients suggests that the cochlear nucleus may be damaged by the tumor itself or during its removal. The inferior colliculus, an auditory structure in the midbrain, offers an alternative site of stimulation for NF2 patients (Lim et al., 2008, 2009). Although the results are encouraging, this technique is very recent, as less than 10 patients had been implanted with an AMI to date.

Visual Neuroprostheses

In contrast to their auditory counterparts, no visual prostheses are routinely used in the clinic. However, recent technological developments have opened new possibilities and some visual prostheses have been FDA and CE approved. To date, there are basically three main sites for interfacing with the visual system: the retina, the optic nerve, and the visual cortex. As is the case with CIs, an internal unit stimulates the cells after receiving the sensory information (in this case visual) from an external unit (**Figure 2**).

Retinal Implants

Retinal implants, whose scope is limited to retinal degradation or retinitis pigmentosa, are placed either on (epiretinal implant) or behind (subretinal implant) the retina, or above the vascular choroid (suprachoroidal implant).

1. Epiretinal implants

The external unit includes a miniature camera, an integrated circuit for signal processing, and a wireless connection

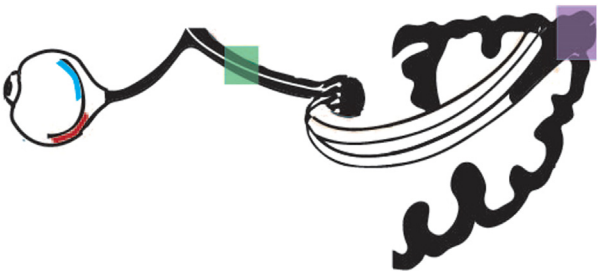

Figure 2 Implantation sites for epiretinal (blue), subretinal (red), optic nerve (green), and visual cortex (purple).

(radio frequency or infrared transmission), all mounted on a pair of glasses worn by the patient. The camera captures the images that are then processed by the integrated unit, which in turn sends stimulation commands to the internal implant. The implant consists of a receiver and an array of electrodes made that is mounted on a flexible silicon printed circuit and is attached to the inner layer of the retina. The necessary information and energy are received and sent to the electrodes that deliver electrical pulses to the retinal ganglion cells. Electrical stimulation of the retina typically results in the elicitation of phosphenes, spatially localized sensations of light. Several clinical trials of epiretinal implants have been successfully carried out on patients with retinitis pigmentosa (Humayun et al., 1996; Chader et al., 2009), and the first FDA-approved implant, ARGUS II, afforded patients spatial navigation and simple motor tasks (Ahuja et al., 2011).

2. Subretinal implants

In subretinal implants, an array of micro-photodiodes is implanted in the outer layer of the retina and replaces the damaged photoreceptors by capturing the light entering the eye, no external camera is required. Activation of individual photodiodes results in the delivery of electrical pulses to the other layers of the retina (biopolar and ganglion cells). However, since micro-photodiodes are situated in the outer layer of the retina, the path of light is sometimes impeded, so these types of implants require an additional energy source to compensate for this issue. While subretinal implants are not commercially available, patients in clinical trials have reported the perception of phosphenes (Benav et al., 2010; Zrenner et al., 2010). Some problems related to implant corrosion have led to explantation in some patients (Stingl et al., 2013) and need to be addressed before these devices become more widespread.

3. Suprachoroidal implant

In suprachoroidal–transretinal stimulation, an implant is placed in the suprachoroidal space within the sclera (white of the eye). While the implant is not in contact with the retina, this technique has the advantage over its epi- or subretinal counterparts in that the surgery is easier and safer for patients and results in more stable electrodes positioning (Fujikado et al., 2011). While implants in animals and a few human subjects have been promising, this approach is still experimental, as more efficient surgical techniques are under investigation. Moreover, the spatial resolution is still unknown and is expected to be lower than that of retinal implants.

4. Summary

The advantage of subretinal implants over epiretinal ones is that the former interface with the nervous system at an earlier stage, that is, when the light is converted into electrical impulses, which allows the image to move with the eyes in a natural way. Furthermore, the surgical procedure involved for subretinal implants is simple, as the implant is held in place by the space between the outer layer and the retinal pigment cells. In contrast, tacks are necessary to attach the epiretinal implant to the tissue. One of the major drawbacks of subretinal implants is that the light has to cross several layers to reach the micro-photodiodes. The quantity of light that reaches the implant affects the resolution, and often an external power source is added to the external module. Suprachoroidal–transretinal stimulation may offer more stable electrode attachment than do retinal implants and thus be less susceptible to variability in thresholds and to cause damage to retinal neurons.

Both retinal implants stimulate the first layer of cells in the retina if these are intact and functional. Both the implantation procedures are more straightforward than those required for optic nerve or cortical stimulation. However, retinal implants cannot be used in diseases that affect ganglion cells, such as glaucoma. To date, only one epiretinal implant, the Argus II, has been approved by the FDA and in Europe. A subretinal implant, the 'Retina' implant, manufactured in Germany, is also available commercially in CE. The surgical procedure for the suprachoroidal implant is the least risky, as the electrodes are inserted in the space between the sclera and the choroid and, therefore, are not likely to damage the retina. However, as the stimulation is applied to the choroid layers rather than to the retina itself, the resolution is potentially lower.

Optic Nerve Stimulation

An interface with the optic nerve was developed by Veraart et al. (2003) and successfully implemented in only one patient (Brelén et al., 2006). Indeed, the organization and structure of the optic nerve fibers limits the feasibility of optic nerve stimulation: removing the membrane that protects the 1 million optic nerve fibers of optic nerve that runs the risk of damaging the nervous tissue, which makes selective stimulation of optic fibers challenging. However, if deemed feasible, this method would offer the advantage of being applicable to any disease affecting the eye.

Visual Cortex Stimulation

1. History and overview

Foester first reported that stimulating the surface of the occipital cortex (epicortical stimulation) elicited the perception of spots of light called phosphenes. This discovery led to what can be considered the first cortical implant in 1968 (Brindley and Lewin, 1968), which consisted of an array of 80 surface electrodes implanted in the visual cortex of a 52-year-old patient with glaucoma (Weiland and Humayun, 2006). They estimated that 50 electrodes would be required to restore enough vision to read one letter at a time, and at least 600 electrodes to achieve normal reading.

Surface stimulation of visual cortex required high levels of electrical current to elicit visual percepts and led to very large phosphenes, thereby limiting the spatial resolution of the resulting signal. In the 1980s, invasive technology was developed that allowed for intracortical stimulation that consists of delivering electrical stimulation through electrodes that penetrate the cortical tissue. A commonly used intracortical interface is the Utah Electrode Array, which consists of a 10×10 grid of electrodes spanning 4.2×4.2 mm. Individual electrodes are 1- or 1.5-mm long and are separated from each other by 400 μm (Normann et al., 1999).

Experiments by Bartlett and Doty on macaques paved the way for the first human experiments in the early 1990s, in which they demonstrated that the minimum spacing between two electrodes to elicit distinct phosphenes is around 500 μm. Schmidt et al. (1996) implanted a 34-electrode array in the visual cortex of a 42-year-old patient who had been totally blind for over 20 years and showed that the patient was able to recognize simple patterns delivered through the intracortical microstimulation.

Dobelle and his colleagues (Dobelle, 2000) announced the first cortical implant that provides independent mobility to blind people. Initially, the 68-year-old subject could not recognize letters or numbers. However, after 3–4 h learning sessions 2–3 times a week, the subject was able to recognize a character inscribed in a 6 inch (15.24 cm) square at a distance of 5 feet, which corresponds to a visual acuity of 20/1200. In 2001, a zoom function was added to the camera improving the acuity to 20/400. The subject was able to take the subway by himself and navigate in a room with obstacles. That said, the success cannot be attributed to the implant alone, as an ultrasound telemeter was mounted on the camera to help with obstacle avoidance. More recently, wireless implants have been developed (Troyk et al., 2003; Hosseini-Khayat et al., 2012) that eliminate the transcranial cables and thus facilitate the surgical procedure.

2. Design concept

As is the case with other visual neuroprostheses, a camera captures images and sends the information to a processing unit, which enhances the contrast and reduces the resolution of the images. The unit then transmits a control command to the implant which then electrically stimulates the visual cortex. In theory, three parameters determine the quality of a pixelated image: the number of pixels, their density, and their range of achievable intensities. The number and density of pixels achievable with a cortical implant is related to the number of implanted electrodes and limited by the density of the electrodes and the spread of currents, which determines how distinct the phosphenes are elicited by two adjacent electrodes. There are many technological hurdles to producing arrays that are dense, robust, and biocompatible. Intracortical implants require complex and risky surgeries, which require not only creating a craniotomy, but also cutting a flap in the dura. Complications during surgery can lead to disastrous consequences on health, including death. However, the cortical implant is applicable to almost all types of blindness.

Somatosensory Neuroprostheses

Given the importance of somatosensory feedback in motor control, the restoration of touch and proprioception is critical for upper-limb neuroprostheses to be clinically viable. To date, most efforts to restore touch through neural stimulation have focused on interfacing with the nerve or with the brain. The underlying principles are similar to those underlying visual and auditory neuroprostheses: Signals from transducers, in this case mechanical transducers on the prosthetic limb, are converted into patterns of electrical stimulation of neural tissue to elicit meaningful somatosensory sensations (**Figure 3**).

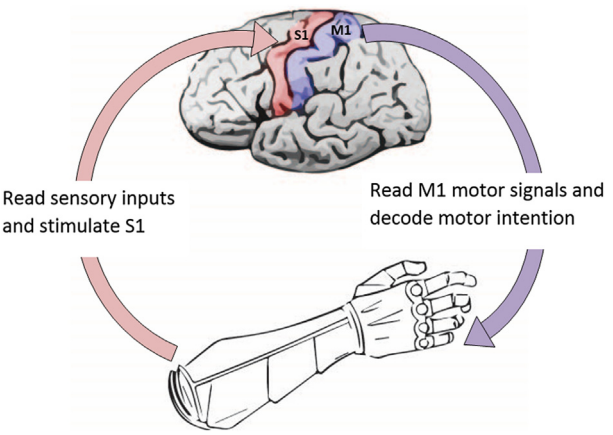

Figure 3 Pathway from M1 to S1 in a bidirectional BCI.

1. Peripheral interfaces

A few different interfaces have been developed for the nerve, some involving penetrating the nerve with electrode shanks (Branner and Normann, 2000), others flattening the nerve and stimulating it through its protective layers (Tyler and Durand, 2002), few others interfacing with the cell bodies in the dorsal root ganglia (Hokanson et al., 2011), and still others rerouting the nerve endings to intact skin and using tactors to stimulate them (Kuiken et al., 2007). The response properties of cutaneous mechanoreceptive afferents have been well characterized: indeed, models have been developed that can predict afferent responses to arbitrary spatial and temporal patterns of stimulation with millisecond accuracy (Dong et al., 2013; Kim et al., 2010, 2011). The relationship between afferent activation and perception has also been studied extensively. Indeed, the peripheral bases for form (Goodwin et al., 1997; Phillips and Johnson, 1981) and texture perception (Weber et al., 2013) have been characterized, as has that of perceived tactile intensity (Muniak et al., 2007) and pleasant touch (Löken et al., 2009). Furthermore, stimulation of individual afferents elicits very consistent and systematic tactile percepts that can readily be characterized by afferent type (Ochoa and Torebjörk, 1983). Thus, peripheral neural interfaces are supported by a strong empirical foundation. Naturalistic and well localized percepts have been achieved through stimulation of the peripheral nerve, both in acute preparations (Dhillon and Horch, 2005) and in chronic ones (Kuiken et al., 2007; Ledbetter et al., 2013; Raspopovic et al., 2014; Tyler and Durand, 2002).

2. Cortical interfaces

Wilder Penfield observed that stimulating the surface of primary somatosensory cortex elicits tactile percepts whose projected location on the body surface depended on where stimulation was applied (Rasmussen and Penfield, 1947). Romo and colleagues demonstrated that changing the parameters of intracortical microstimulation results in systematic changes in the evoked percepts (Romo et al., 1998). These early efforts laid the foundation for an approach to conveying somatosensory feedback, analogous to its visual counterpart, that consists in implanting arrays

of electrodes in primary somatosensory cortex (S1) and injecting small currents although these electrodes to elicit tactile and proprioceptive percepts (Berg et al., 2013; O'Doherty et al., 2011; O'Doherty et al., 2012; Tabot et al., 2013; Thomson et al., 2013; Venkatraman and Carmena, 2011). As is the case with their visual counterparts, the advantage of cortical over peripheral interfaces is that they are applicable to patients with upper spinal cord injuries. Stimulation through individual electrodes has been shown to elicit tactile percepts that are localized on the skin (Tabot et al., 2013). The perceptual magnitude of these percepts increases systematically and predictably with stimulation amplitude (Fridman et al., 2010; Tabot et al., 2013). While animals have been able to use this type of artificial sensory feedback to perform basic sensorimotor tasks (O'Doherty et al., 2011; Thomson et al., 2013), the intracortical approach is still in its infancy and has yet to be tested in humans.

Cognitive Neuroprostheses

Cognitive neuroprostheses aim to restore a cognitive function, such as attention, decision making, or memory. This approach typically consists of electrically stimulating the brain area that is thought to mediate the cognitive function in question. For instance, electrical stimulation of prefontal cortex in monkeys was shown to improve impaired decision making (Hampson et al., 2012), which might then pave the way to restore lost cognitive function in brain-damage patients. Among the prosthetics that have been approved by the FDA is deep-brain stimulation (DBS) that is described below.

Deep-Brain Stimulation

DBS is an FDA-approved procedure to treat movement disorders such as Parkinson's disease, dystonia, and essential tremor, but it may also be applicable to treat chronic depression and chronic pain (Young and Brechner, 1986; Anderson et al., 2012). In 1985, a patient suffering from essential tremor was to receive a thalamic ablation. Before the ablation, electrical stimulation was applied to determine the location of the tissue to be removed. When a specific portion of the thalamus was stimulated electrically, the tremor stopped immediately, which led the team, led by Benabid in Grenoble, France, to consider DBS as a treatment for motor disorders and paving the way for the first clinical trials of DBS in 1987 on a patient with Parkinson's disease (Benabid et al., 2009). More than 80 000 patients with Parkinson's disease have since been treated with DBS for pharmacologically intractable motor disorders. The surgical procedure for DBS is not without risk for the patient and can cause significant complications and side-effects such as mood changes or sensory/motor deficits. The site of stimulation depends on the type of disorder; the most common sites are subthalamic nucleus and globus pallidus for Parkinson's and dystonia, and the ventrointermediate nucleus for essential tremor. Stimulations of ventral posterolateral nucleus and of ventral posteromedial nucleus have been shown to alleviate pain due to cancer or due to phantom limb (Young and Brechner, 1986; Kringelbach et al., 2007). More recently, some evidence suggests that DBS could be used to treat chronic depression.

Vagus Nerve Stimulation

The vagus nerve, or 10th cranial nerve, regulates the heartbeat, controls some skeletal muscles, maintains breathing, mediates the transmission of a variety of chemicals through the body, maintains the digestive tract, contracts the muscles of the stomach and intestines to help process food, and conveys information about digestion and its metabolic consequences. Vagus nerve stimulation (VNS) is approved by the FDA to treat seizures (since 1997) and drug-resistant depression (since 2005). Similarly to DBS, VNS involves a pacemaker, implanted in the chest, which sends electrical impulses to the brainstem along the left vagus nerve. Because the right vagus nerve is involved in cardiac functions, stimulation of this structure could be fatal. The brainstem both relays and coordinates the information sent to several areas in the brain. Several groups (Sjögren et al., 2002; Clark et al., 1999; Groves and Brown, 2005) are investigating the possibility of extending VNS to treat Alzheimer's disease, chronic migraines, and tinnitus (a condition that affects hearing). VNS presents side-effects related to the laryngeal muscles. Several patients experience voice alteration, difficulty swallowing, throat pain, hoarseness, and coughing (Hsieh et al., 2008). Sleep apnea has also been reported in most patients (Marzec et al., 2003).

Efferent Neuroprostheses

In the first section, we discussed neuroprostheses that stimulate a tissue somewhere along the neuraxis. In another category of prostheses, information is read out of the nervous system to control an external effector. Measurements of neural responses can be split into two categories: invasive and noninvasive. The invasive techniques require a craniotomy and consist of measuring the activity of neurons by implanting electrodes at the surface of or into the brain. In these cases, either the responses of single neurons (single unit), of small groups of neurons (multiunit), or of large groups of neurons (electrocorticography or ECog) are recorded. Noninvasive techniques such as EEG or near-infrared spectroscopy (NIRS) can be used to measure the electrical activity at the surface of the scalp and thus do not require surgery. The choice between invasive or noninvasive technique depends on the application and on the targeted patient population. Noninvasive techniques may also be used in nonclinical applications but present the drawback of yielding a noisy signal.

Invasive Efferent Neuroprostheses

In an amputee or tetraplegic patient, the limb is either no longer there or no longer connected to the brain. However, the parts of the brain that sent signals to the muscles to move the limb are still functional. Signals from motor areas of the brain can be used to control an extracorporeal device, such as a prosthetic arm, or to stimulate populations of muscles to restore movement in the deafferented arm, an approach called functional electrical stimulation. The invasive version of this

approach consists in implanting multielectrode arrays in areas of the brain that are implicated in motor control, often primary motor cortex, as also premotor and parietal cortices. Then, the patient observes the movement of a prosthetic limb going through a series of motions, and is asked to imagine his or her own arm traversing those trajectories, while neuronal activity (in the form of single units, multiunits, or local field potentials) is recorded. Based on the recorded neuronal activity and the measured trajectories of the observed limb, a 'decoder' is built, which is a function that maps patterns of neuronal activity to kinematic states of the limb.

Algorithms to decode motor intention from neuronal signals have evolved rapidly over the past 15 years, from the one-dimensional control of a robotic arm used by rats to retrieve water droplets (Chapin et al., 1999), to two- (Serruya et al., 2002) and three- (Taylor et al., 2002) dimensional control of cursor movement by monkeys. Work with animal models paved the way for human experiments, where tetraplegic patients were able to control anthropomorphic robotic arms (Johannes et al., 2011) with many dimensions of control (Collinger et al., 2013; Hochberg et al., 2012). While these initial tests are remarkable and promising, many obstacles remain to be overcome to transfer this technology from the experimental laboratory to the home and to the clinic. For example, the chronically implanted electrode arrays have a relatively short life span (Judy, 2012) and thus would require periodic replacement, which is not an option given the dangerous surgery involved.

Noninvasive Efferent Neuroprostheses

In noninvasive efferent neuroprostheses, users modulate aspects of their brain activity, which are then used to control an external effector. Specifically, the signal is acquired and recorded in real time, features of interest are extracted (i.e., the aspects of brain activity that will drive the effector), and transmitted to an algorithm that converts these signals into a command to control a device such as a computer cursor or a robotic arm. Different types of brain features entail different types of control strategies (Zander et al., 2010), as detailed below.

Active BCI

In an active BCI, the signal is actively and explicitly controlled by the user. For example, the user might be explicitly instructed to imagine an action. Neuroimaging studies showed that imagining the movement activates similar areas in the motor and sensorimotor cortices as does performing the action (Jeannerod, 1997). Hence, this imagery-elicited signal can be used to control an external device. Signals that might be appropriate for an EEG-based BCI are sensorimotor μ and β rhythms. Indeed, when a user imagines a motor action, event-related desynchronization (ERD) decreases in the μ- (8–12 Hz) and β- (18–25 Hz) bands in the sensorimotor cortex; modulation in these bands can then be used to move a cursor (Pfurtscheller and Neuper, 2010; Brewster et al., 2014) or a prosthetic arm (Ramos-Murguialday et al., 2012; Gomez-Rodriguez et al., 2010).

Reactive BCI

In a reactive BCI, the neuronal activity evoked by repeated presentation of a stimulus is recorded. The subject is then asked to attend to one of the several stimuli presented, and algorithms are developed to identify the differential activity that is evoked by this target stimulus relative to that evoked by the other stimuli. Then, this differential activity can be used to decode a response from the subject, who then attends to one of the several stimuli to produce a response. In the decoding phase, the subject is presented with a set of interleaved stimuli. To produce a specific response, he or she must attend to its associated target. The cortical response to that target then shifts to an attended state, and the response associated with that target is recorded. In one approach, the P300, a positive deflection in the EEG signal that occurs at an average of 300 ms after stimulus onset, is used as a gauge of attention. The P300-based technique has been used to associate different targets with different letters, thereby allowing the subject to spell. This paradigm has been particularly useful in allowing patients with locked-in syndrome to communicate. An analogous approach consists in measuring the steady-state visually evoked potential (SSVEP). In this paradigm, the subject is presented with a visual stimulus that flickers at a constant frequency. The signal generated by the brain thus oscillates at the same frequency as the attended visual stimulus if the subject is attending to it. Thus, detecting the frequency present in the EEG can be used to determine to what stimulus a subject is attending to, and thus decoded as a response (in the same spirit as the P300 approach described above).

Passive BCI

Passive BCI extended the world of BCI to marketing, work environments, games, and human–computer interface. As indicated by its name, the user's brain activity is passive, which means that she is not asked to do anything. While she is going about some task that is unrelated to BCI control, like working on a computer, the BCI records and detects features from her brain activity that signal that she is stressed or that she may have made an error. If such an event occurs, the BCI then reduces the cognitive load imposed upon the user. A passive BCI record might then detect neural signatures of fatigue, arousal, working memory overload, or other cognitive or emotional states that might be triggered by the task. In one example of a passive BCI, participants were asked to navigate in virtual mazes with varying levels of difficulty (George et al., 2012). When the BCI detected a high mental load based on EEG signals, a haptic device was activated to guide the participant. Thus, the participant was doing the task unassisted unless his or her mental load reached a threshold, at which time assistance was provided. Using the same principle, some neuromarketing companies such as Neurofocus provide their clients with their consumers' physiological responses via a BCI while purchasing or being exposed to new products (Pradeep, 2010).

Conclusion

The development of neuroprosthetics has experienced exponential growth over the past two decades. Devices can be used to restore sensation, to control extracorporeal devices, to allow locked-in patients to communicate, and to monitor

physiological states to improve performance on a variety of tasks. Technologies range from highly invasive arrays of electrodes that are implanted in neural tissues to noninvasive devices that measure neuronal activity through the scalp. The promise of these technologies is only beginning to be realized, and it is difficult to imagine the state of the art of neuroprosthetics in 10 years.

See also: Cerebral Cortex; Electroencephalography: Basic Principles and Applications; Motor Cortex; Neuromarketing; Occipital Lobe (Visual Cortex): Functional Aspects; Prefrontal Cortex; Visual Perception, Neural Basis of.

Bibliography

Ahuja, A.K., Dorn, J.D., Caspi, A., McMahon, M.J., Dagnelie, G., Stanga, P., Greenberg, R.J., 2011. Blind subjects implanted with the Argus II retinal prosthesis are able to improve performance in a spatial-motor task. British Journal of Ophthalmology 95 (4), 539–543.

Anderson, R.J., Frye, M.A., Abulseoud, O.A., Lee, K.H., McGillivray, J.A., Berk, M., Tye, S.J., 2012. Deep brain stimulation for treatment-resistant depression: efficacy, safety and mechanisms of action. Neuroscience & Biobehavioral Reviews 36 (8), 1920–1933.

Benabid, A.L., Chabardes, S., Mitrofanis, J., Pollak, P., 2009. Deep brain stimulation of the subthalamic nucleus for the treatment of Parkinson's disease. The Lancet Neurology 8 (1), 67–81.

Benav, H., Bartz-Schmidt, K.U., Besch, D., Bruckmann, A., Gekeler, F., Greppmaier, U., Harscher, A., Kibbel, S., Kusnyerik, A., Peters, T., et al., 2010. Restoration of useful vision up to letter recognition capabilities using subretinal microphotodiodes. In: Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE, pp. 5919–5922.

Berg, J., Dammann, J., Tenore, F., Tabot, G., Boback, J., Manfredi, L., Bensmaia, S., 2013. Behavioral demonstration of a somatosensory neuroprosthesis. Neural Systems and Rehabilitation Engineering, IEEE Transactions 21 (3), 500–507.

Branner, A., Normann, R.A., 2000. A multielectrode array for intrafascicular recording and stimulation in sciatic nerve of cats. Brain Research Bulletin 51, 293–306.

Brelén, M.E., De Potter, P., Gersdorff, M., Cosnard, G., Veraart, C., Delbeke, J., 2006. Intraorbital implantation of a stimulating electrode for an optic nerve visual prosthesis: case report. Journal of Neurosurgery 104 (4), 593–597.

Brewster, P., Hock, C., Flint, L., Lanthier, A., Ziat, M., Dow, D., Hartman, M., 2014. BCI-Tactile cursor task. In Haptics Symposium (HAPTICS), 2014 IEEE, p. 1.

Brindley, G.S., Lewin, W.S., 1968. The sensations produced by electrical stimulation of the visual cortex. The Journal of Physiology 196 (2), 479–493.

Chader, G., Weiland, J., Humayun, M., 2009. Artificial vision: needs, functioning, and testing of a retinal electronic prosthesis. Progress in Brain Research 175, 317–332.

Chapin, J.K., Moxon, K.A., Markowitz, R.S., Nicolelis, M.A., 1999. Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex. Nature Neuroscience 2 (7), 664–670.

Clark, G., 2003. Cochlear Implants: Fundamentals and Applications. Springer, New York.

Clark, K.B., Naritoku, D.K., Smith, D.C., Browning, R.A., Jensen, R.A., 1999. Enhanced recognition memory following vagus nerve stimulation in human subjects. Nature Neuroscience 2 (1), 94–98.

Colletti, V., Carner, M., Miorelli, V., Guida, M., Colletti, L., Fiorino, F., 2004. Auditory brainstem implant (ABI): new frontiers in adults and children. Otolaryngology – Head and Neck Surgery 133 (1), 126–138.

Collinger, J.L., Wodlinger, B., Downey, J.E., Wang, W., Tyler-Kabara, E.C., Weber, D.J., Schwartz, A.B., 2013. High-performance neuroprosthetic control by an individual with tetraplegia. The Lancet 381 (9866), 557–564.

Dhillon, G.S., Horch, K.W., 2005. Direct neural sensory feedback and control of a prosthetic arm. IEEE Transactions on Neural Systems and Rehabilitation Engineering 13 (4), 468–472.

Dobelle, W.H., 2000. Artificial vision for the blind by connecting a television camera to the visual cortex. ASAIO Journal 46 (1), 3–9.

Dong, Y., Mihalas, S., Kim, S.S., Yoshioka, T., Bensmaia, S., Niebur, E., 2013. A simple model of mechanotransduction in primate glabrous skin. Journal of Neurophysiology 109 (5), 1350–1359.

Fridman, G.Y., Blair, H.T., Blaisdell, A.P., Judy, J.W., 2010. Perceived intensity of somatosensory cortical electrical stimulation. Experimental Brain Research 203 (3), 499–515.

Fujikado, T., Kamei, M., Sakaguchi, H., Kanda, H., Morimoto, T., Ikuno, Y., Nishida, K., 2011. Testing of semichronically implanted retinal prosthesis by suprachoroidal-transretinal stimulation in patients with retinitis pigmentosa. Investigative Ophthalmology & Visual Science 52 (7), 4726–4733.

George, L., Marchal, M., Glondu, L., Lécuyer, A., June 12–15, 2012. Combining brain-computer interfaces and haptics: detecting mental workload to adapt haptic assistance. In: Proceedings of Eurohaptics Conference. Tampere, Finland.

Gomez-Rodriguez, M., Peters, J., Hill, J., Schölkopf, B., Gharabaghi, A., Grosse-Wentrup, M., 2010. Closing the sensorimotor loop: haptic feedback facilitates decoding of arm movement imagery. In: Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on IEEE, pp. 121–126.

Goodwin, A.W., Macefield, V.G., Bisley, J.W., 1997. Encoding of object curvature by tactile afferents from human fingers. Journal of Neurophysiology 78 (6), 2881–2888.

Groves, D.A., Brown, V.J., 2005. Vagal nerve stimulation: a review of its applications and potential mechanisms that mediate its clinical effects. Neuroscience & Biobehavioral Reviews 29 (3), 493.

Hampson, R.E., Gerhardt, G.A., Marmarelis, V., Song, D., Opris, I., Santos, L., Deadwyler, S.A., 2012. Facilitation and restoration of cognitive function in primate prefrontal cortex by a neuroprosthesis that utilizes minicolumn-specific neural firing. Journal of Neural Engineering 9 (5), 056012.

Hochberg, L.R., Bacher, D., Jarosiewicz, B., Masse, N.Y., Simeral, J.D., Vogel, J., Donoghue, J.P., 2012. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. Nature 485 (7398), 372–375.

Hokanson, J.A., Ayers, C.A., Gaunt, R.A., Bruns, T.M., Weber, D.J., 2011. Effects of spatial and temporal parameters of primary afferent microstimulation on neural responses evoked in primary somatosensory cortex of an anesthetized cat. In: Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE. IEEE, pp. 7533–7536.

Hosseini-Khayat, S., Bahmanyar, P., Rahiminezhad, E., Sawan, M., 2012. Ultra-low power encryption engine of wireless implantable medical devices. In: 55th IEEE International Midwest Symposium on Circuits and Systems, MWSCAS 2012.

Hsieh, T., Chen, M., McAfee, A., Kifle, Y., 2008. Sleep-related breathing disorder in children with vagal nerve stimulators. Pediatric Neurology 38 (2), 99–103.

Humayun, M.S., de Juan Jr., E., Dagnelie, G., Greenberg, R.J., Propst, R.H., Phillips, D.H., 1996. Visual perception elicited by electrical stimulation of retina in blind humans. Archives of Ophthalmology 114 (1), 40.

Jeannerod, M., 1997. The Cognitive Neuroscience of Action. Wiley-Blackwell, Oxford.

Johannes, M.S., Bigelow, J.D., Burck, J.M., Harshbarger, S.D., Kozlowski, M.V., Van Doren, T., 2011. An overview of the developmental process for the modular prosthetic limb. Johns Hopkins APL Technical Digest 30 (3), 207–216.

Judy, J.W., 2012. Neural interfaces for upper-limb prosthesis control: opportunities to improve long-term reliability. Pulse, IEEE 3 (2), 57–60.

Kim, S.S., Sripati, A.P., Bensmaia, S.J., 2010. Predicting the timing of spikes evoked by tactile stimulation of the hand. Journal of Neurophysiology 104 (3), 1484.

Kim, S.S., Mihalas, S., Russell, A., Dong, Y., Bensmaia, S.J., 2011. Does afferent heterogeneity matter in conveying tactile feedback through peripheral nerve stimulation? IEEE Transactions on Neural Systems and Rehabilitation Engineering 19 (5), 514–520.

Kringelbach, M.L., Jenkinson, N., Owen, S.L., Aziz, T.Z., 2007. Translational principles of deep brain stimulation. Nature Reviews Neuroscience 8 (8), 623–635.

Kuiken, T.A., Marasco, P.D., Lock, B.A., Harden, R.N., Dewald, J.P., 2007. Redirection of cutaneous sensation from the hand to the chest skin of human amputees with targeted reinnervation. Proceedings of the National Academy of Sciences 104 (50), 20061–20066.

Ledbetter, N.M., Ethier, C., Oby, E.R., Hiatt, S.D., Wilder, A.M., Ko, J.H., Clark, G.A., 2013. Intrafascicular stimulation of monkey arm nerves evokes coordinated grasp and sensory responses. Journal of Neurophysiology 109 (2), 580.

Lim, H.H., Lenarz, T., Anderson, D.J., Lenarz, M., 2008. The auditory midbrain implant: effects of electrode location. Hearing Research 242, 74–85.

Lim, H.H., Lenarz, M., Lenarz, T., 2009. Auditory midbrain implant: a review. Trends in Amplification 13, 149.

Löken, L.S., Wessberg, J., McGlone, F., Olausson, H., 2009. Coding of pleasant touch by unmyelinated afferents in humans. Nature Neuroscience 12 (5), 547–548.

Marzec, M., Edwards, J., Sagher, O., Fromes, G., Malow, B.A., 2003. Effects of vagus nerve stimulation on sleep-related breathing in epilepsy patients. Epilepsia 44 (7), 930–935.

Muniak, M.A., Ray, S., Hsiao, S.S., Dammann, J.F., Bensmaia, S.J., 2007. The neural coding of stimulus intensity: linking the population response of mechanoreceptive afferents with psychophysical behavior. The Journal of Neuroscience 27 (43), 11687–11699.

NIH Consens Statement Online, May 15–17, 1995. Cochlear Implants in Adults and Children 13 (2), 1–30.

Normann, R.A., Maynard, E.M., Rousche, P.J., Warren, D.J., 1999. A neural interface for a cortical vision prosthesis. Vision Research 39 (15), 2577–2587.

Ochoa, J., Torebjörk, E., 1983. Sensations evoked by intraneural microstimulation of single mechanoreceptor units innervating the human hand. The Journal of Physiology 342 (1), 633–654.

O'Doherty, J.E., Lebedev, M.A., Ifft, P.J., Zhuang, K.Z., Shokur, S., Bleuler, H., Nicolelis, M.A., 2011. Active tactile exploration using a brain-machine-brain interface. Nature 479 (7372), 228–231.

O'Doherty, J.E., Lebedev, M.A., Li, Z., Nicolelis, M.A., 2012. Virtual active touch using randomly patterned intracortical microstimulation. IEEE Transactions on Neural Systems and Rehabilitation Engineering 20 (1), 85–93.

Otto, S.R., Shannon, R.V., Wilkinson, E.P., Hitselberger, W.E., McCreery, D.B., Moore, J.K., Brackmann, D.E., 2008. Audiologic outcomes with the penetrating electrode auditory brainstem implant. Otology and Neurotology 29 (8), 1147–1154.

Peterson, N.R., Pisoni, D.B., Miyamoto, R.T., 2010. Cochlear implants and spoken language processing abilities: review and assessment of the literature. Journal of Restorative Neurology and Neuroscience 28, 237–250.

Pfurtscheller, G., Neuper, C., 2010. Dynamics of sensorimotor oscillations in a motor task. In: Graimann, B., Pfurtscheller, G. (Eds.), Invasive and Non-invasive Brain-Computer Interfaces. Springer: The Frontiers Collection.

Phillips, J.R., Johnson, K.O., 1981. Tactile spatial resolution. II. Neural representation of bars, edges, and gratings in monkey primary afferents. Journal of Neurophysiology 46 (6), 1192–1203.

Pradeep, A.K., 2010. The Buying Brain. John Wiley & Sons.

Ramos-Murguialday, A., Schürholz, M., Caggiano, V., Wildgruber, M., Caria, A., Hammer, E.M., et al., 2012. Proprioceptive feedback and brain computer interface (BCI) based neuroprostheses. PloS one 7 (10) e47048.

Rasmussen, T., Penfield, W., 1947. The human sensorimotor cortex as studied by electrical stimulation. In: Federation Proceedings, vol. 6, no. 1, pt 2, p. 184.

Raspopovic, S., Capogrosso, M., Petrini, F.M., Bonizzato, M., Rigosa, J., Di Pino, G., Micera, S., 2014. Restoring natural sensory feedback in real-time bidirectional hand prostheses. Science Translational Medicine 6 (222), 222ra19.

Rauschecker, J.P., Shannon, R.V., 2002. Sending sound to the brain. Science 295, 1025–1029.

Romo, R., Hernández, A., Zainos, A., Salinas, E., 1998. Somatosensory discrimination based on cortical microstimulation. Nature 392 (6674), 387–390.

Schmidt, E.M., Bak, M.J., Hambrecht, F.T., Kufta, C.V., O'rourke, D.K., Vallabhanath, P., 1996. Feasibility of a visual prosthesis for the blind based on intracortical micro stimulation of the visual cortex. Brain 119 (2), 507–522.

Serruya, M.D., Hatsopoulos, N.G., Paninski, L., Fellows, M.R., Donoghue, J.P., 2002. Brain-machine interface: instant neural control of a movement signal. Nature 416 (6877), 141–142.

Shannon, R.V., 2010. Auditory prostheses for the brainstem and midbrain. In: Palmer, A.R., Rees, A. (Eds.), The Oxford Handbook of Auditory Science: The Auditory Brain. Oxford Library of Psychology.

Sjögren, M.J., Hellström, P.T., Jonsson, M.A., Runnerstam, M., Silander, H.C.S., Ben-Menachem, E., 2002. Cognition-enhancing effect of vagus nerve stimulation in patients with Alzheimer's disease: a pilot study. The Journal of Clinical Psychiatry 63 (11), 972–980.

Stingl, K., Bartz-Schmidt, K.U., Besch, D., Braun, A., Bruckmann, A., Gekeler, F., et al., 2013. Artificial vision with wirelessly powered subretinal electronic implant alpha-IMS. Proceedings of the Royal Society B: Biological Sciences 280 (1757), 20130077.

Tabot, G.A., Dammann, J.F., Berg, J.A., Tenore, F.V., Boback, J.L., Vogelstein, R.J., Bensmaia, S.J., 2013. Restoring the sense of touch with a prosthetic hand through a brain interface. Proceedings of the National Academy of Sciences 110 (45), 18279–18284.

Taylor, D.M., Tillery, S.I.H., Schwartz, A.B., 2002. Direct cortical control of 3D neuroprosthetic devices. Science 296 (5574), 1829–1832.

Thomson, E.E., Carra, R., Nicolelis, M.A., 2013. Perceiving invisible light through a somatosensory cortical prosthesis. Nature Communications 4, 1482.

Troyk, P., Bak, M., Berg, J., Bradley, D., Cogan, S., Erickson, R., Towle, V., 2003. A model for intracortical visual prosthesis research. Artificial Organs 27 (11), 1005–1015.

Tyler, D.J., Durand, D.M., 2002. Functionally selective peripheral nerve stimulation with a flat interface nerve electrode. IEEE Transactions on Neural Systems and Rehabilitation Engineering 10 (4), 294–303.

Venkatraman, S., Carmena, J.M., 2011. Active sensing of target location encoded by cortical microstimulation. IEEE Transactions on Neural Systems and Rehabilitation Engineering 19 (3), 317–324.

Veraart, C., Wanet-Defalque, M.C., Gérard, B., Vanlierde, A., Delbeke, J., 2003. Pattern recognition with the optic nerve visual prosthesis. Artificial Organs 27 (11), 996–1004.

Weber, A.I., Saal, H.P., Lieber, J.D., Cheng, J.W., Manfredi, L.R., Dammann, J.F., Bensmaia, S.J., 2013. Spatial and temporal codes mediate the tactile perception of natural textures. Proceedings of the National Academy of Sciences 110 (42), 17107–17112.

Weiland, J.D., Humayun, M.S., 2006. Intraocular retinal prosthesis. Engineering in Medicine and Biology Magazine, IEEE 25 (5), 60–66.

Wolpaw, J.R., Boulay, Ch B., 2010. Brain signals for brain–computer interfaces. In: Graimann, B., Pfurtscheller, G. (Eds.), Invasive and Non-invasive Brain-Computer Interfaces. Springer: The Frontiers Collection.

Young, R.F., Brechner, T., 1986. Electrical stimulation of the brain for relief of intractable pain due to cancer. Cancer 57 (6), 1266–1272.

Zander, T.O., Kothe, C., Jatzev, S., Gaertner, M., 2010. Enhancing human–computer interaction with input from active and passive brain-computer interfaces. In: Tan, D., Nijholt, A. (Eds.), Brain–Computer Interfaces: Applying Our Minds to Human–Computer Interaction. Springer, London, pp. 181–199.

Zrenner, E., Bartz-Schmidt, K.U., Benav, H., Besch, D., Bruckmann, A., Gabel, V.P., et al., 2010. Subretinal electronic chips allow blind patients to read letters and combine them to words. Proceedings of the Royal Society B: Biological Sciences 278, 1489–1497, rspb20101747.

Neuropsychological Testing

D Anne Winiarski, Emory University, Atlanta, GA, USA
Harry A Whitaker, Northern Michigan University, Marquette, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the applications of neuropsychological assessment, with particular emphasis on some of the most common standardized batteries available for testing a wide array of functions. Some of the most widely used and empirically validated cognitive, behavioral, academic, and personality measures are discussed. The article concludes with a brief overview of how to summarize findings from a clinical evaluation in a comprehensive neuropsychological report.

Neuropsychological assessment provides psychologists with a powerful tool for understanding the cognitive and behavioral manifestations of underlying brain functioning. Clinical neuropsychologists administer assessments to clients in order to determine whether they are suffering from conditions that may not be readily detectable using brain-imaging technology, such as a learning disability or Attention-Deficit Hyperactivity Disorder (ADHD). Neuropsychological evaluations also enable psychologists to determine what kinds of services, or accommodations, may be most beneficial to the client. This article briefly explores some of the applications of neuropsychological assessment across a variety of contexts, examines widely used reliable and valid measures of cognitive functioning, academic functioning, behavior, and personality, and concludes with a suggested approach for summarizing assessment findings in the form of a comprehensive report, as well as a discussion on the future of standardized neuropsychological assessment in an era of rapidly developing technological advancements. Some of the most widely used assessments are described below, but it should be noted that the selected assessments discussed in this article are not an exhaustive list, nor is it possible to expound on each assessment's psychometric properties at length. The interested reader is encouraged to review the manuals associated with each assessment, which are cited within the text. Furthermore, neuropsychological assessments related to speech deficits are not covered in this article, and the interested reader is encouraged to reference other articles (*see* Language Disorders, Treatment and Remediation of; Classical Tests for Speech and Language Disorders).

Goals of Neuropsychological Assessment

Neuropsychological assessment has many practical applications; six important ones stand out: diagnosis; patient care; treatment planning and remediation; treatment evaluation; research; and forensic applications (Lezak et al., 2004).

The first goal of neuropsychology, diagnosis, may seem quite obvious. Although superior imaging techniques exist for identifying the sites of brain damage, a neuropsychological evaluation illuminates the extent of deficits related to that injury, and how they manifest behaviorally and cognitively. Moreover, sensitive diagnostic neuroimaging techniques do not yet exist for identifying certain organic brain disorders (e.g., ADHD), further necessitating the utilization of neuropsychological assessment.

Neuropsychological testing also provides clinicians and treatment providers with pertinent information relating to a patient's current level of functioning. For example, an assessment can provide insight into whether a patient is able to care for him or herself or whether the patient is capable of successful academic or professional performance. Relatedly, information concerning treatment planning can typically be gleaned from 'accommodations' provided in a neuropsychological report. In some cases, neuropsychological assessment may be useful in evaluating pre- and postadministration of pharmacological, medical, or surgical procedures.

Neuropsychological assessments can also be applied in research and forensic contexts. From a research standpoint, assessments are included in a wide array of research protocols, including those reporting slower processing speed abilities among soldiers with comorbid traumatic brain injury and posttraumatic stress (Nelson et al., 2009). Another recent study comparing neurocognitive functions among preterm and full-term children found that preterm children performed worse on neuropsychological measures of processing speed, certain aspects of attention and executive functioning, visual-motor coordination, as well as face processing and emotion regulation (Potharst et al., 2013). Neuropsychologists are also increasingly called to the courtroom to address issues in the legal arena, such as the possible consequences of traumatic brain injury on employment (Bigler and Brooks, 2009). In criminal cases, a neuropsychologist may be called to assess a defendant's mental capacity to stand trial or to explain how neuropsychological factors may help explain criminal behavior (Raine, 2013: pp. 309–310).

Clearly, this cursory overview of the applications of neuropsychological assessment does not do justice to the myriad ways in which neuropsychological assessments can help individuals with cognitive, intellectual, and physical impairments. The interested reader is encouraged to look at more comprehensive reviews of neuropsychological assessment; some suggestions are provided below.

Malingering

Before delving into a discussion of specific assessments, it is essential to understand how malingering may influence a neuropsychological assessment. Malingering is defined as the intentional generation of false or exaggerated symptoms, and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54049-5

can manifest in two primary ways: (1) deliberate exaggeration of symptoms that are difficult to measure and (2) intentional poor performance on psychological and neuropsychological tests (Iverson and Binder, 2000). Although malingering generally occurs in 'high-stakes' situations, including those in which individuals are suing for worker's compensation, are seeking a declaration of 'incompetence' to avoid standing trial, or are trying to plead insanity in a criminal case, psychologists should be cautious of this phenomenon when working with any patient. Individuals who are malingering may exaggerate memory loss, poor concentration, personality changes, or other bodily or cognitive symptoms. Clinicians should always assess malingering in the context of a neuropsychological assessment.

Psychologists have proposed several strategies for identifying malingering. Larrabee (2003) identified clinically atypical patterns of performance on five standardized psychological and neuropsychological assessments: Benton Visual Form Discrimination, Finger Tapping, WAIS-R Reliable Digit Span, Wisconsin Card Sorting Failure-to-Maintain Set, and the Lees-Haley Fake Bad Scale from the Minnesota Multiphasic Personality Inventory-2 (MMPI-2). Using the cut-offs established from his analyses, Larrabee was able to correctly identify 15 of 17 subjects meeting criteria for probable malingering. These findings suggest that examiners may be able to utilize built-in empirical methods for assessing malingering without having to administer additional assessments.

Selecting an Assessment Battery

Before selecting an assessment battery, it is first necessary to determine the purpose of the assessment. Specifically, the clinician should have a clear picture of the patient's referral questions, as well as an understanding of what the patient can gain from the assessment (e.g., a request for specific accommodations, such as additional time on a test to accommodate a slower processing speed). A patient being evaluated for possible ADHD will likely require a different test battery from someone who needs an assessment following a traumatic brain injury. An important benefit of neuropsychological testing is that it allows for the selection of tests that tap into specific cognitive functions, and more importantly, a test battery can be tailored to the individual needs of the client.

Mini Mental Status Exam

The Mini Mental Status Exam (MMSE; Folstein et al., 1975) is not a standardized neuropsychological assessment per se, but it is a very important feature of assessment that can oftentimes be used as a screening tool, most commonly as a screening test for dementia. This screener requires an individual to respond to questions in eight categories. The first category addresses orientation to time (e.g., year, day of the week) and the second addresses orientation to place (e.g., can the patient state where he or she is). Next, the patient's registration is assessed, which is his or her ability to repeat three randomly selected words. In order to assess attention and calculation, a neuropsychologist may ask a patient to spell 'world' backward or to count backward from 100 by 7's, i.e., 100…93…86…19…etc.

After some delay, the examiner will ask the patient to recall the three words he or she had asked him to repeat verbatim earlier during registration assessment. The examiner may also wish to assess the patient's language by asking him to name everyday objects (e.g., the examiner's pen). The patient may also be asked to repeat a phrase back to the examiner exactly as he or she says it. The examiner may then conclude the MMSE by asking the patient to carry out more complex commands (e.g., follow three-sequence series of directions, e.g., 'pick up the pen, fold the piece of paper, and give me the magazine').

Cognitive Assessments

Intelligence

The Wechsler scales are among the most reliable and widely used neuropsychological evaluations of intelligence. There are essentially three classes of this assessment: the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV), which is administered to children aged 6 years, 0 months to 16 years, 11 months; Wechsler Adult Intelligence Scale, Fourth Edition (WAIS-IV), which is administered to individuals aged 16 years, 0 months to 90 years, 11 months; and an abbreviated version of these measures, the Wechsler Abbreviated Scale of Intelligence, now in its Second Edition (WASI-II), is available for individuals aged 6 years, 0 months to 90 years, 11 months.

The WAIS-IV and WISC-IV subtests are grouped into four indices or factors, and one global composite score, which is known as the Full Scale IQ (FSIQ). The four factors include Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed. The Verbal Comprehension Index (VCI) taps verbal reasoning and knowledge of word meanings; the Perceptual Reasoning Index (PRI) taps spatial visualization, visual relationships, and nonverbal reasoning; the Working Memory Index (WMI) taps short-term auditory information and the ability to hold and manipulate information in memory; and the Processing Speed Index (PSI) assesses overall speed of information processing and the ability to screen out (ignore) irrelevant stimuli.

Both the WISC-IV and WAIS-IV have 10 subtests (although the specific subtests on each assessment vary, as will be discussed below); three subtests comprise the VCI, three subtests comprise the PRI, and two subtests each comprise the WMI and PSI. Additionally, on both the WAIS-IV and WISC-IV, the three verbal comprehension subtests and three perceptual reasoning subtests can be utilized in the calculation of the General Ability Index (GAI), which is useful when the FSIQ is uninterpretable. The FSIQ may be uninterpretable in cases where the difference between the highest Index and the lowest Index exceeds 1.5 standard deviations, or 23 points (Flanagan and Kaufman, 2009).

The three subtests that tap into the VCI on the WAIS are the similarities subtest, vocabulary subtest, and the information subtest. There is also a supplemental comprehension subtest. The three subtests that tap into the PRI are the block design subtest, matrix reasoning subtest, and the visual puzzles subtest. The two supplemental subtests that tap into this Index are the picture completion and figure weights subtests. The WMI is composed of two core subtests, digit span and

arithmetic, as well as one supplemental subtest, letter–number sequencing. The two subtests making up the PSI are symbol search and coding, with an optional cancellation subtest.

The organization and administration of the WISC-IV are very similar, but some of the individual subtests comprising the indices differ from the WAIS-IV. Three subtests comprise the VCI, including vocabulary, similarities, and comprehension, with optional information and word reasoning subtests.

The WASI-II is an abbreviated measure of intelligence, with options of administering two or four subtests to generate the FSIQ (Wechsler, 2011). The two-subtest form requires the administration of the vocabulary and matrix reasoning subtests, and only yields the FSIQ. The four-subtest form requires the examiner to administer the vocabulary, similarities, block design, and matrix reasoning subtests, and generates an FSIQ-4, a VCI score, and a PRI score. Whereas the WISC and WAIS each take anywhere between 60 and 90 min to administer to core battery, the WASI-II can take about 15 (2 subtests) to 30 min (4 subtests), thereby increasing efficiency in the clinical or research setting.

The Wechsler Preschool and Primary Scale of Intelligence, Fourth Edition (WPPSI-IV) is also available to test younger populations (2 years, 6 months to 7 years, 7 months), and provides similar information as the other Wechsler instruments. The 14 core, supplemental, and optional subtests combine to generate five Primary Index scales. These scales include a VCI, Visual Spatial Index (VSI), WMI, Fluid Reasoning Index (FRI), and PSI. An FSIQ is also calculated. The WPPSI-IV has adequate reliability, and improved normative sampling from previous editions. Furthermore, the age range was extended from 7 years, 3 months (which was the maximum age in previous versions of the WPPSI) to 7 years, 7 months in the present standardized version.

The Differential Abilities Scale, Second Edition (DAS-II; Elliott, 2007) is another cognitive assessment that is available for testing of children. The DAS-II provides scores reflecting verbal, nonverbal, and spatial abilities, as well as a composite measure of general cognitive ability (General Conceptual Ability score, GCA). The GCA is defined as the "general ability of an individual to perform complex mental processing that involves conceptualization and the transformation of information" (Elliott, 2007: p.17). In addition to the composite measure, two or three lower-level composite or cluster scores can also be calculated. These cluster scores vary by age. The DAS-II is divided into two batteries, an Early Years battery (2 years, 6 months to 6 years, 11 months) and a School-Age battery (7 years, 0 months to 17 years, 11 months).

The DAS-II has shown very good internal reliability, with subtest coefficients ranging from 0.74 to 0.96 across both the Early Years and School-Age batteries. Inter-rater reliability among subtests requiring examiner judgment (e.g., Copying, Recall of Designs) was high, ranging from 0.95 to 0.99. There is also evidence for the high internal validity of the DAS-II (Davis et al., 2012).

Executive Functioning

One common assessment of executive functions is the Delis–Kaplan Executive Function System (D-KEFS), compromising nine tests designed to tap into various verbal and nonverbal executive functions. Each of the nine subtests generates information about an individual's performance, eliminating the need for an overall composite score. In fact, one of the many strengths of the D-KEFS is that it provides great flexibility in administration because the examiner can choose only those specific subtests that will answer the assessment question; he or she does not have to administer a predetermined number of tests before obtaining a clinically significant index score (Delis et al., 2001).

The D-KEFS has been shown to have impressive test-retest reliability, as well as high internal consistency on the Trail Making Test, Verbal Fluency Test, and Color-Word Interference Test. The authors of the manual also explain that factor analytic studies aimed at deriving index scores were not necessary given that each test is designed to be a stand-alone measure. The D-KEFS can be administered across several developmental periods, from ages 8 through 89.

Sometimes, clinicians must diagnose neuropsychological functioning in younger individuals, and when this becomes the case, tests like the D-KEFS are not normed for these younger populations. As such, it may be necessary to use a test like A Developmental Neuropsychological Assessment, Second Edition (NEPSY-II) in order to test these younger populations. The NEPSY-II is available in two forms: one for 3- to 4-year-old patients (Preschool Age), and one for 5- through 16-year-old patients (School Age). The six functional domains of the NEPSY-II were derived theoretically rather than empirically. These six domains cover attention and executive functioning, language, memory and learning, sensorimotor, social perception, and visuospatial processing. Like the D-KEFS, the NEPSY-II is a very versatile assessment, with selection of subtests being based on the referral question. Subtests can provide an overview of neuropsychological functioning, a diagnostic measure specific to the referral question, a selective assessment to determine information about cognitive abilities, or the examiner can provide a comprehensive neuropsychological evaluation (D'Amato and Hartiage, 2008).

Memory Assessment

One of the best-known memory assessments in the field is the Wechsler Memory Scale, Fourth Edition (WMS-IV; 2009). This assessment comprises seven subtests that tap into five Index Scores: Auditory Memory, Visual Memory, Visual Working Memory, Immediate Memory, and Delayed Memory. The WMS-IV can be administered to adolescent and adult patients (16 years, 0 months through 90 years, 11 months).

The California Verbal Learning Test, Second Edition (CVLT-II; Delis et al., 2000) is another measure used to specifically evaluate verbal memory, and is available for adolescents and adults (16–89 years), as well as for children (5 years, 0 months through 16 years, 11 months). In both assessments, individuals are asked to recall a list immediately, after a delay, with interference, and with or without cues.

Another memory assessment that was uniquely designed for children is the Children's Memory Scale (CMS), which is normed for children aged 5–16. Unlike the CVLT-II, which is specific to verbal memory, the CMS assesses a broader range of memory dimensions, including attention and working memory,

recall and recognition, short-delay and long-delay memory, as well as both verbal and visual memory (Cohen, 1997).

Finally, the Test of Memory and Learning, Second Edition (TOMAL-2; Reynolds and Voress, 2007) is a comprehensive memory assessment available for children and adults (5 years, 0 months to 59 years, 11 months). The assessment consists of eight core subtests, six supplemental subtests, and two delayed recall tasks. These various subtests generate three core indices (Verbal Memory Index, Nonverbal Memory Index, and Composite Memory Index) and six supplemental indices (Verbal Delayed Recall Index, Learning Index, Attention and Concentration Index, Sequential Memory Index, Free Recall Index, and Associative Recall Index).

Motor Functions

The Finger Tapping Test (FTT; Reitan and Wolfson, 1993) is one of the most widely recognized tests of manual dexterity. Previous studies have found that measures of simple motor speed like the FFT are able to demonstrate more reliable differences between individuals with head injuries and those without than specific measures of memory or attention (Dikmen et al., 1995).

The Peabody Developmental Motor Scales, Second Edition (PDMS-2) assesses fine and gross motor skills among children from birth to 5 months. The PDMS assesses both quantitative and qualitative aspects of motor development among children. For example, an examiner qualitatively evaluates a child's grasp to determine whether the grasp is mature for the child's age. The length of time it takes for the child to grasp the cube is also recorded (quantitative measure). The developers of the PDMS assert that the quantitative information gathered during the assessment may be especially useful in developing remediation strategies. The six subtests of the PDMS assess various milestones of fine and gross motor development, including grasping, locomotion, object manipulation, body control and equilibrium, and visual motor integration (Folio and Fewell, 2000).

Academic Functioning

While academic problems and specific learning disabilities are typically assessed using structured psychoeducational batteries, it is not uncommon to also find psychoeducational measures in some neuropsychological batteries, depending on the reason for referral. For example, a child with sickle cell anemia may present with recent academic difficulties, including problems with attention and reading comprehension. Upon further medical examination, it is determined that this child has experienced a series of silent strokes, one of the many typical complications of his or her disorder. He or she is referred for a neuropsychological evaluation to determine the cognitive consequences of his or her strokes, but in order to obtain special accommodations at his or her school, the neuropsychologist may need to provide evidence that the cognitive impairments influence the academic concerns that were reported. In this case, using standardized measures of academic functioning may prove to be very useful in identifying specific areas of academic concern.

One well-known and widely used measure of academic functioning is the Woodcock–Johnson Tests of Achievement, Third Edition (WJ-III ACH; Woodcock et al., 2001). The Standard Battery includes 12 subtests that assess performance in areas such as reading, oral language, math, written language, and general academic skills. The different start points within each subtest are determined based on the individual's level of schooling rather than age.

The Wechsler Individual Achievement Test, Third Edition (WIAT-III) assesses reading, math, writing, and oral language in a total of 16 subtests. It is a useful tool for identifying academic strengths and weaknesses, and can inform an examiner's decision about the presence of a learning disability (Miller, 2012).

Specialized assessments of various academic subjects are also available, but a thorough discussion of their use and psychometric properties is beyond the scope of this article. Examples of these assessments would be found in the Key-Math3 (Connolly, 2007), the Gray Oral Reading Tests - Fifth Edition (Wiederholt and Bryant, 2012), the Clinical Evaluation of Language Fundamentals - Fourth Edition (Semel et al., 2003), and the Test of Written Language -Fourth Edition (Hammill and Larsen, 2009).

Personality Assessment

Similar to academic functioning, personality assessment is not necessarily a core feature of neuropsychological batteries. Nevertheless, in some cases, it may be necessary to assess an individual's personality in order to obtain a complete picture of present psychosocial functioning. Furthermore, as was discussed above, individual scales of certain personality assessments can be used to detect possible malingering, especially in 'high-stakes' assessment (e.g., assessing someone who would like to use the insanity defense).

There are a wide variety of personality assessment measures available; one way to distinguish these measures is to classify them as 'projective' or 'objective.' Projective measures allow an individual to respond to ambiguous stimuli, and it is thought that these responses may reveal important characteristics about an individual's hidden emotions. Some of the better known projective measures include the Rorschach Inkblot Test, the Thematic Apperception Test, and the Draw-A-Person Test. For the purposes of this article, emphasis will be placed on the 'objective' measures of personality, such as the MMPI-2 (Hathaway et al., 2009) and the Personality Assessment Inventory (PAI). Unlike the projective tests listed above, objective tests require the individual being examined to answer specific questions typically presented in a multiple-choice format. These responses are then compared to responses generated from the normative samples utilized in the creation of the assessment.

The MMPI-2 is a 567-item questionnaire that is composed of over 120 scales, 10 of which are the well-known 'clinical scales.' One of the many strengths of the MMPI is its use of validity indicators, which can be useful to detect threats to a valid administration, such as malingering, response inconsistency, and superlative self-presentation. The 10 clinical scales on the MMPI-2 are (1) hypochondriasis; (2) depression; (3) hysteria;

(4) psychopathic deviate; (5) masculinity/femininity; (6) paranoia; (7) psychasthenia; (8) schizophrenia; (9) hypomania; and (10) social introversion. On the clinical scales of the MMPI-2, a T-score of 65 is considered clinically significant, but these elevations should be interpreted in light of other psychosocial variables, which can typically be gleaned from a clinical interview with the participant. An adolescent version (MMPI-A) was also developed, and is currently one of the most commonly used assessments of personality among adolescents (Merrell, 2008).

The PAI provides an assessment of psychopathology in individuals aged 18–89. There are 344 items on this assessment, and together they make up 22 scales, including four validity scales, 11 clinical scales, five treatment scales, and two interpersonal scales. This measure has been found to have high internal consistency, convergent validity, and discriminant validity (Morey, 1991). An adolescent version (PAI-A) is also available.

Behavioral Assessment

Behavioral ratings can also be used to supplement formal neuropsychological assessments, but are not necessarily core features of a neuropsychological battery. Since a neuropsychologist only evaluates an individual for a few hours, it is difficult to obtain a clear picture of this individual's behavioral context outside the testing session. As a result, administration of behavior rating scales to the individual, caregivers, parents, and other important individuals with whom the client interacts can provide a wealth of information about possible problematic behaviors the client may be experiencing. In certain contexts, such as an evaluation of a possible ADHD diagnosis, behavioral rating scales are critical in ascertaining that the characteristic ADHD behaviors of impulsivity and/or inattention occur across several contexts.

While there are certainly several empirically validated rating forms available (e.g., Child Behavior Checklist, Behavior Assessment System for Children, etc.), this article will limit its discussion to the Behavior Rating Inventory of Executive Function (BRIEF), which can be used to supplement neuropsychological evaluations of impairments in executive functioning (Gioia et al., 1996). This questionnaire is available for parents and teachers of children aged 5–18. There are also adolescent self-report forms (BRIEF-SR), preschool age forms for children between the ages of 2 and 5 (BRIEF-P), and self and informant rating forms for adults ranging in age from 18 to 90 (BRIEF-A).

While the number of items on each of the versions of the BRIEF differs, the overall clinical scales tap into the same constructs, including working memory, emotional control, inhibition, and planning/organization. Overall, the BRIEF's good reliability, validity, and clinical utility make this a useful addition to any neuropsychological battery, especially where questions of cross-situational impairment may arise.

The Report

After evaluating a client, a neuropsychologist generally writes a report summarizing the referral questions, pertinent background information, data gathered from the neuropsychological testing sessions, an interpretation of the data, and formal recommendations and/or accommodations for the patient. After summarizing the referral question and reasons for the assessment, the neuropsychologist will then highlight relevant background history information. This section may include information about developmental milestones, medical history, academic history, employment history, and any significant injuries, hospitalizations, etc. The behavioral observation section of the report outlines information about the client's mood, affective state, as well as any observations that the clinician made about the client's performance, effort, and/or behavior across the different tests. Based on these observations, a clinician should include a statement reflecting whether he or she believes the results of the assessment are an accurate reflection of the client's current level of functioning. Next, the clinician generally presents the data gathered from the assessment batteries that were administered, explaining why certain assessments were selected, and what conclusions can be drawn about the individual's functioning across the various domains of functioning that were assessed.

After the data are thoroughly presented, the report typically concludes with a brief summary of the referral question, relevant background information, and a condensed overview of the findings. This portion of the report should emphasize the most important findings clearly and concisely so that it is apparent how the diagnostic impressions (which are listed after the summary) were determined. Finally, the clinician concludes the report with important accommodations and recommendations that the individual, his or her family, other professionals, etc. may find to be useful in treatment planning, academic planning, or judiciary proceedings.

Concluding Comments and Future Directions

Neuropsychological reports can be used in a variety of settings, and for a variety of purposes. Parents can use these reports to help their children obtain 504 plans or Individualized Educational Plans (IEPs). Neuropsychological reports can also be used in forensic settings to determine whether someone is mentally fit to stand trial or to help an individual obtain worker's compensation in a lawsuit. Because individuals rely on these reports for sometimes life-changing purposes, it is essential to obtain and maintain competence in assessment before administering, scoring, and interpreting a neuropsychological battery (American Psychological Association, 2010). Furthermore, the neuropsychologist should make every effort to meet with the patient, his or her family, or both in order to communicate the findings of the assessment, explain the recommendations and/or accommodations, and answer any questions the patient and his or her family may have.

Neuropsychological assessment can provide valuable insight into the current state of functioning of individual patients, and can enable a clinician to make recommendations about treatment and services that an individual may need to thrive in spite of having neuropsychological deficits. It is the authors' hope that this very brief introduction has introduced the reader to some of the more widely used neuropsychological

assessments, to their utility, and to their applicability across various domains.

All of the assessments listed above are generally administered using pencil, paper, and stimulus materials. Test developers have recently begun taking advantage of newer technologies to create standardized assessments that can be administered entirely on computers and tablets. Not only does the portability of standardized assessments increase administration efficiency, it also enables examiners to administer assessments in less time, and it simplifies the scoring process by having all necessary materials in one place (the tablet/computer). The CVLT-II is one example of a standardized measure that can now be administered in its entirety using this tablet-based system.

See also: Classical Tests for Speech and Language Disorders; Intellectual Functioning, Assessment of; Personality Assessment.

Bibliography

American Psychological Association, 2010. Ethical Principles of Psychologists and Code of Conduct. Available at: www.apa.org/ethics/code/index.aspx.

Bigler, E.D., Brooks, M., 2009. Traumatic brain injury and forensic neuropsychology. The Journal of Head Trauma Rehabilitation 24 (2), 76–87.

Cohen, M., 1997. Children's Memory Scale. Pearson Education, Inc., Bloomington, MN.

Connolly, A.J., 2007. KeyMath-3. Pearson Education, Inc., Bloomington, MN.

D'Amato, R.C., Hartiage, L.C. (Eds.), 2008. Essentials of Neuropsychological Assessment: Treatment Planning for Rehabilitation, second ed. Springer, New York.

Davis, A.S., Finch, W.H., Tindal, G., 2012. Review of the differential ability scales. In: The Eighteenth Mental Measurements Yearbook, second ed. Available at: http://buros.unl.edu/buros (accessed 08.10.13.).

Delis, D.C., et al., 2000. California Verbal Learning Test, second ed. Pearson Education, Inc., San Antonio, TX.

Delis, D.C., Kaplan, E., Kramer, J.H., 2001. Delis-Kaplan Executive Function System. Pearson Education, Inc., San Antonio, TX.

Dikmen, S.S., et al., 1995. Neuropsychological outcome at 1-year post head injury. Neuropsychology 9 (1), 80–90.

Elliott, C.D., 2007. Differential Ability Scales-II (DAS-II). Harcourt Assessment, San Antonio, TX.

Flanagan, D.P., Kaufman, A.S., 2009. Essentials of WISC-IV Assessment, second ed. Wiley, Hoboken, NJ.

Folio, M.R., Fewell, R.R., 2000. Peabody Developmental Motor Scales, second ed. Pearson Education, Inc., San Antonio, TX.

Folstein, M., Folstein, S.E., McHugh, P.R., 1975. "Mini-mental state" a practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research 12 (3), 189–198.

Gioia, G.A., et al., 1996. Behavior Rating Inventory of Executive Function. Psychological Assessment Resources, Inc., Lutz, FL.

Hammill, D.D., Larsen, S.C., 2009. Test of Written Language, fourth ed. Pearson Education, Inc., Austin, TX.

Hathaway, S.R., McKinley, J.C., Butcher, J.N., 2009. Minnesota Multiphasic Personality Inventory-2. University of Minnesota Press, Minneapolis, MN.

Iverson, G.L., Binder, L.M., 2000. Detecting exaggeration and malingering in neuropsychological assessment. The Journal of Head Trauma Rehabilitation 15 (2), 829–858.

Larrabee, G.J., 2003. Detection of malingering using atypical performance patterns on standard neuropsychological tests detection of malingering using atypical performance patterns on standard neuropsychological tests. The Clinial Neuropsychologist 17 (3), 37–41.

Lezak, M.D., Howieson, D.B., Loring, D.W., 2004. Neuropsychological Assessment, fourth ed. Oxford University Press, Oxford.

Merrell, K.W., 2008. Behavioral, Social, and Emotional Assessment of Children. Routledge, New York.

Miller, M.D., 2012. Review of the Wechsler Individual Achievement Test.

Morey, L.C., 1991. Personality Assessment Inventory. Psychological Assessment Resources, Inc., Odessa, FL.

Nelson, L.A., et al., 2009. Relationship between processing speed and executive functioning performance among OEF/OIF veterans: implications for post-deployment rehabilitation. The Journal of Head Trauma Rehabilitation 24 (1), 32–40.

Potharst, E.S., et al., 2013. Perinatal risk factors for neurocognitive impairments in preschool children born very preterm. Developmental Medicine and Child Neurology 55 (2), 178–184.

Raine, A., 2013. The Anatomy of Violence: The Biologial Roots of Crime. Pantheon Books, New York.

Reitan, R.M., Wolfson, D., 1993. The Halstead-Reitan Neuropsychological Test Battery: Theory and Clinical Interpretation. Neuropsychology Press, Tucson.

Reynolds, C.R., Voress, J.K., 2007. Test of Memory and Learning, second ed. PRO-ED, Austin, TX.

Semel, E., Wiig, E.H., Secord, W.A., 2003. Clinical Evaluation of Language Fundamentals, fourth ed. Pearson Education, Inc., San Antonio, TX.

Wechsler, D., 2011. Wechsler Abbreviated Scale of Intelligence, second ed. Pearson Education, Inc., San Antonio, TX.

Wechsler, D., 2009. Wechsler Memory Scale, fourth ed. Pearson Education, Inc., San Antonio, TX.

Wiederholt, J.L., Bryant, B.R., 2012. Gray Oral Reading Tests, fifth ed. PRO-ED, Austin, TX.

Woodcock, R.W., et al., 2001. Woodcock Johnson III Normative Update (NU) Tests of Achievement. Riverside Publishing, Itasca, IL.

Neuroscience of Education

Andrew Tolmie, University of London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Educational neuroscience attempts to coordinate evidence from behavioral and neuroimaging studies in order to obtain a more complete understanding of learning that can then be used to specify the pedagogical approaches and educational systems that will support it most effectively. The translational nature of this objective poses challenges because of its requirement for collaboration between researchers and educational practitioners. However, educational neuroscience research in literacy, number development, and science learning illustrates the potential of the field to explain both typical and atypical learning in a coherent fashion and to identify novel pedagogical strategies that fully address individual variation in capability.

Introduction: A Brief Conceptual History of Educational Neuroscience

Interest in the potential applicability of neuroscience to education stemmed initially from the wider growth of cognitive neuroscience – research on the relationship between brain function and cognitive activity – during the late twentieth century. As technology that had been developed in medical contexts became more refined and more readily accessible, new techniques emerged to support *in vivo* studies of neural activity. These made it possible to examine what happens at the brain level during mental processing of information and coordination of behavioral responses, such as those involved in carrying out numerical calculations. The use of functional magnetic resonance imaging (fMRI) – a method of measuring areas of increased blood flow in the brain and therefore, by inference, areas of greater neural activity – proved particularly successful as a means of mapping the apparent organization of brain function. Identification of recurrent regularities in the way in which neural activity is coordinated and deployed in different contexts led in turn to important reconceptualizations of the nature of cognition. The most striking example was recognition of the inadequacies of information-processing models derived from computing, which had been the guiding framework for cognitive psychology and cognitive science since the 1960s. In particular, it became evident that the brain does not engage in narrow sequential processing of items of information, but operates instead through large-scale parallel processing and the extraction of stable patterns, as, for example, in the learning and use of letter–sound relationships in reading.

The success of core forms of cognitive neuroscience led naturally to its extension to other forms of cognitive activity, including the perception and management of social situations (social cognitive neuroscience) and the processes by which adult patterns of neural activity emerge (developmental neuroscience). Again, both led to new insights, particularly with regard to the relationship between more primitive brain structures in evolutionary terms (e.g., the amygdala, which plays a central role in processing of emotion) and more uniquely human structures, such as the prefrontal cortex, which is involved in the manipulation of information and the control of other systems.

Given the fundamental role of the brain in learning, and the central concern of education with the organization and promotion of learning, it was inevitable that this further extension – which has gradually come to be called educational neuroscience – would be mooted as a possibility. Interest in the potential connection was further fuelled by perceptions of the inadequate and failing nature of educational systems in much of the developed world, and the growth of arguments in favor of evidence-based approaches to schooling that take into account the role of the brain in learning (Chipman, 1986). It was helped too by interest on the part of many educators – echoing wider popular fashion – in the powerful but apparently unpredictable influence of the brain on what they observed in their classrooms. Summing up these perceptions, the Royal Society (2011) concluded that the application of a scientific approach to learning based on neuroscience had the potential to bring about "a future in which educational practice can be transformed … just as medical practice was transformed by science about a century ago" (p. v).

As the Royal Society quotation makes plain, though, educational neuroscience never was, and never could be, the same kind of endeavor as social cognitive or developmental neuroscience. The latter are fields of research as traditionally conceived, operating within relatively well-defined and preexisting realms of enquiry, and with the primary objective solely of increasing understanding. The resulting knowledge may well have potential for application, but it is not typically an immediate concern of researchers working within these fields to consider those applications in planning their work.

Educational neuroscience differs in two fundamental ways. First, because education is a social invention, not a naturally occurring process (although it may have its roots in evolutionarily advantaged processes of intergenerational transmission of learning), educational neuroscience lacks any similar fixed realm of enquiry. While it can (and often does) focus on an analysis of the relationship between neural activity and learning within some current educational setting, like engineering, the object of such analysis is rarely understanding in itself, but the design of as yet unrealized approaches that will work in better fashion. It is therefore inherently an area of research in which consideration of application is a central concern. This focus on application gives rise to the second key

difference, which is that educational neuroscientists are necessarily driven to work at least part of the time in collaboration with those responsible for the design and delivery of education, rather than primarily as a community of researchers. Such collaboration is crucial for the identification of research priorities – those areas where existing education systems appear to have greatest difficulty – and for the design and trialing of workable solutions.

If collaboration of this kind is a necessary part of the field, it also carries substantial challenges, especially with respect to building sufficient shared understanding to enable effective communication and coordination of activity between the different parties involved. The difficulties have been recognized more or less since the beginning. In an influential early critique, Bruer (1997) argued that the endeavor was 'a bridge too far,' because the perceived connections between education and neuroscience were too remote, and were based too often on misunderstandings or overgeneralizations of one kind or another. One example he cited as an illustration was the claim – common at the time among educators – that the proliferation of new synaptic connections between brain cells (synaptogenesis) that takes place during infancy and early childhood represents a critical period for promoting enhanced learning in a variety of areas. Such claims were based on mistaken beliefs about the general nature of synaptogenesis and the functions it supports, beliefs that were based on limited contact with the relevant literature. The process is in fact focused to a large extent on specific regions of the brain and linked primarily to the emergence of basic perceptual and control functions. The learning of explicit procedures and content that are focal to education continues to be refined long after the early synaptic spurt has given way to pruning and consolidation of synaptic connections, which suggests that synaptogenesis has no real bearing on it.

Other forms of neuromyth, as these mistaken conceptions are commonly termed, still persist (e.g., left- and right-brain learning – most cognitive functions rest, in fact, on neural systems distributed throughout the brain). Other ideas that have taken root have more credibility but little systematic evidence in their favor (e.g., the value of training key functions like working memory – while training produces improvement on standard measures, there seem in fact to be substantial problems with transfer of these gains to classroom performance). The misunderstandings are not solely on the part of educators, though. Hruby (2012) details instances of the failure of researchers to engage with – or even respect – expertise regarding practice, which is essential if translational goals are to be met. Common failures include glossing over the conclusions of practice-based research, or attributing perspectives to practitioners that they do not in fact hold.

It is evident therefore that there are problems of translation and of contact in both directions, and that it is only by addressing these problems that, as Hruby contends, the real potential of educational neuroscience will become clear. Bruer (1997) argued that progress was more likely to be achieved via a series of discrete linkages, between educators and instructional psychologists, on the one hand, and between psychologists and neuroscientists, on the other. In the intervening period, some of these boundaries have become blurred: many educational researchers have acquired expertise in neuroscience methodologies, and better lines of communication have been established between teachers and researchers of different types. Precise prescriptions as to productive linkages are therefore probably outdated. Nevertheless, it remains plain that if educational neuroscience has any real future, it will not be driven by any single disciplinary agenda, but by wide collaboration between researchers from different backgrounds and practitioners of different types. It will also need support from government and policy makers because of the implications it carries for the nature of educational systems. In this article, I outline current thinking on what the objectives of educational neuroscience might be, what models of collaboration might work best for achieving these, and what existing evidence reveals about how far the endeavor might be worth the effort.

What Is Educational Neuroscience?

Although it has become widely accepted, the term 'educational neuroscience' is problematic because of what it implies – that it is concerned solely with attempting to explain education through neuroscience principles – and what becomes obscured as a result. While there is no very clear consensual definition of educational neuroscience, there is in fact general agreement that it is not reductionist, that it involves a range of methodologies, and that it is fundamentally concerned with the coordination of evidence of many different types, with neuroscience research providing just one strand.

The overriding objective is to obtain a more complete understanding of processes of learning, which can then be used to specify the pedagogical approaches and educational systems that are likely to support these most effectively. The past 20 years have seen growing recognition that single-level accounts of learning based on cognition, behavior, social process, or biological systems are necessarily incomplete: since all have evidence to support them, they must each be capturing part of a substantially more complex set of processes. It seems plain, then, that the way forward lies in attempting where possible to integrate evidence and accounts from different levels into a more coherent whole. As neuroscience research has uncovered influences and constraints on learning, it has become increasingly apparent that any attempt at integration must include the neural level if it is to offer any prospect of a complete picture. The contribution of educational neuroscience is to draw attention to this dimension – but not to the exclusion of other levels.

The model of the learner that emerges (see, e.g., Mareschal et al., 2013) takes the form of a multilevel system operating simultaneously at neural, cognitive, and social–environmental levels, with complex and multidirectional interactions between these. These result in unexpected – and even, to the practitioner, puzzling – constraints on progress, and substantial individual variation both in presentation during schooling and in educational outcomes. This variation is masked to some extent by equifinality – the achievement of ostensibly the same outcome via different routes. This gives a mistaken impression of the extent to which there is uniformity across learners and how far they can be catered to by one-to-many delivery within classes

using uniform approaches to teaching. In fact, as learners' backgrounds become more diverse, especially within urban centers of population, and the criteria for achievement become more exacting, any assumption of uniformity becomes increasingly problematic. What is needed is coordinated effort to (1) map and understand the cross-level interactions, seeking regularities and typologies wherever possible; (2) derive implications for practice that are sensitive to individual – and topic-to-topic – variation in what is likely to be effective; and (3) consider methods of translating these into actual practice, taking pragmatic constraints into account.

If these are the objectives of educational neuroscience, the implication is that it is better thought of not as an area of research in the same way as developmental psychology or cognitive science, but as a wider endeavor more akin to public health. The history of public health as a discipline and a practice provides some indication of what in fact the development of educational neuroscience may involve, as well as some clues as to the structures that may need to be evolved in order to achieve the translational objectives. The origin of modern public health is generally traced to the application of germ theory by John Snow during the 1854 cholera outbreak in London. His approach suggested that specific microorganisms were responsible for the spread of the disease, which led in turn to his identification of a polluted public water well as its source. This notable success helped overturn miasma theory – the notion that the origin of such diseases lay in airborne emanations from rotting organic matter – and resulted in a rapid extension of germ theory to a range of infectious diseases, a marked growth in public sanitation works, and over the ensuing decades the development of programs relating to public education (e.g., on infant health), vaccination, road safety, occupational safety, and drug control. Public health is now an established part of daily life in developed countries, and an explicit objective of developing ones due to the growth of governmental and international agencies promoting good practice (e.g., the US Surgeon-General's Office and the World Health Organization).

If the translational goals of educational neuroscience parallel those of public health science, then this implies that it should (1) concentrate attention initially on a few key areas of educational need where good theory can make an obvious difference; (2) build outward from these via core teams of individuals representing the different contributing strands of activity (i.e., the equivalent of epidemiologists, biostatisticians, local and national government officials, and health service professionals), whose activity is focused on mutually identified areas of need or risk and methods of counteracting these; (3) promote public knowledge of effective practices (without necessarily worrying overmuch about grasp of why these are important); and (4) let governments take control ultimately, while continuing to feed them good, relevant evidence. The key step in this sequence is almost certainly the second one, which depends on building a consensus across key players in multidisciplinary teams within different professional backgrounds, based on (within bounds) shared knowledge of the relevant science – while as far as possible avoiding bias toward any one approach that may undermine that consensus.

This is a complex and difficult balancing act. Researchers have a critical role to play in providing reliable and systematic evidence that can steer effective action. However, educational neuroscience research to date is piecemeal and unevenly developed. There is much work on dyslexia that is in fact already beginning to inform both remedial and mainstream teaching of literacy (see, e.g., Hulme and Snowling, 2009), but few other areas approach this level of activity, and some (e.g., conceptual growth in science, and gifted and talented children) have so far been addressed by only a handful of researchers. Moreover, there is no unifying framework of the kind provided by germ theory; to the extent that there is a consensus across researchers, this is based on a shared belief that a full understanding of learning processes demands consideration of the neural level, but not what form the resulting models should take. A bigger picture is unlikely to emerge without a more coordinated program of research, covering typical and atypical learning in key areas, especially language and literacy, number and mathematics, and conceptual development and causal understanding in science, widely regarded as the three most important areas of the school curriculum and the principal focus of comparative international assessments.

Equally important is the need to progress as a community in a number of different senses. One aspect of this is researchers from different disciplinary backgrounds working as equal partners. However, researchers can only be one part of a wider community of engagement and exchange that helps set the research agenda and maintains a focus on the implications for practice, including delivery. This wider community also needs to encompass teachers, trainee teachers, teacher-training agencies, educational psychologists, speech and language therapists, pediatric neurologists, and members of other professions involved in implementing evidence-based support for learning and remediation of learning difficulties. If the public health analogy is taken seriously, the function of this wider community extends far beyond a consultative role: it needs to be an active partnership of researchers and professionals working together to identify issues, improve understanding through rigorous research, and develop solutions. Practitioners will therefore need to be involved at the heart of the research, and researchers will need to engage with issues of delivery. These roles are at present largely unfamiliar to all concerned and mostly lack any obvious institutional support. Any functioning team therefore demands an unusual commitment on the part of its members, based primarily on belief in the potential of the work rather than substantial concrete evidence of benefit.

This last point suggests that there also needs to be engagement with policy makers and policy shapers, in order to help ensure that educational neuroscience has socially perceived value, and that team members are positively sanctioned to contribute to the development and deployment of novel forms of provision. Given an environment within which politicians and policy specialists are subjected to constant streams of lobbying by organizations with competing vested interests, success on this front will depend on standing out in some way – hence the importance of good science and convincing applications based upon it. If this was true for public health science, it is even more so now.

Do We Really Need Educational Neuroscience?

Given the complexities involved, is educational neuroscience as an endeavor actually worth the effort? I will address this question by briefly considering work in the three key areas identified in Section What is Educational Neuroscience?: literacy development, number and mathematics, and science learning.

Literacy Development

Learning to decode written forms of language rests in fundamental ways on the child's earlier acquisition of oral language, especially where the orthography involved is alphabetic in character. Even where children grow up with a character-based orthography such as Chinese, initial language acquisition provides them with an implicit sense of the organization of the language. This includes the division of sounds into meaningful word units or morphemes, the basic vocabulary encoded in this way, and acceptable ways of putting this into sequences to create extended meaning – the grammar of the language. There is good evidence that children have an innate sensitivity to these characteristics as they manifest in their local language community, and that imitative observational learning plays an important role in acquiring use of them.

This prior knowledge provides the child with a valuable tool in the more deliberate – and instructed – process of learning to read, since it makes it possible to map the orthographic forms that they encounter onto known elements of language. Initially, at least, this mapping happens at the whole-word level and in rote fashion (what Hulme and Snowling, 2009, call the prealphabetic stage), with children simply learning the association between a given written symbol or set of symbols and a spoken word or morpheme that is significant to them. Rote learning of this kind is inefficient, however, and where the language uses alphabetic orthography, a further level of mapping is possible at the grapheme–phoneme (i.e., letter–sound) level. Guidance on these correspondences helps the child to begin to decode writing into known words without needing the assistance of others to tell them what each word is.

As soon as children have begun to acquire basic letter–sound and letter–name knowledge, they use this mapping to learn the relationship between letter strings and their pronunciation (the alphabetic stage). At first, they are better able to work with larger units than single letters and phonemes, with syllables ('crust') and onset–rime structures ('cr' and 'ust') marking important points of transition to the use of full phonological strategies, in which letters are sounded out and blended to arrive at the pronunciation of unknown words. Finally, as these strategies become more familiar, they become more automated, so that sounding out is no longer necessary. At this point, the child starts to extract the spelling rules that they encounter (the orthographic stage) and in a sense return to whole-word recognition – but on the basis of explicit knowledge about word composition and strategies for decoding unknown words.

Phonological awareness – the ability to identify sound values in speech and to decompose spoken words into these – is a key precursor of learning to read, since it underpins the shift to the alphabetic stage, and makes it possible to bind orthography into preexisting neurocognitive structures that link phonology and meaning. There has therefore been a gradual shift (in the United Kingdom, at least) toward the use of phonics teaching for reading in elementary schools. This involves working directly in the early years to promote awareness of phonemes. However, in English, this approach needs to be accompanied by work on whole-word recognition for high-frequency words because of the number of irregular spellings, which limit the value of grapheme–phoneme conversion as a method of learning to read. Children need to recognize where these mappings do not apply and words are not pronounced as they are written (e.g., 'yacht').

Neuroscience studies using fMRI have identified the neural pathways involved in reading and the anomalies found in children with dyslexia (Demonet et al., 2004). These studies reveal that the main problem is a deficit in the emergence of phonological awareness. Although there are competing theories about the nature of this deficit, the key step in learning to read efficiently appears to be the development of amodal processing structures in the temporoparietal and left ventral cortex, which integrate phonology, orthography, and semantics within what is known as the verbal word form area (VWFA). Metaphonological awareness – the capacity to deliberately utilize knowledge of word sounds – is strongly implicated in the building of these structures. This work has also demonstrated that the differences that separate nonimpaired from impaired readers also predict individual differences in typically developing children, implying that dyslexia is not a qualitatively different syndrome but a more pronounced degree of deficit. This appears to be marked by especially poor neural connectivity in the VWFA and anomalies in the profile of neurotransmitters, the chemical messengers that carry activation from neuron to neuron. It is important to note that how far this deficit has an impact is also affected by the nature of the letter–sound relationship in the child's native language. Reading disorders are less apparent where there is a consistent relationship between letters and phonemes, such as in Spanish or German, or in writing systems that do not require segmentation into phonemes, such as Chinese or Japanese. This suggests that the crucial dimension is sensitivity to complexity in orthographic mapping onto phonology.

Computer modeling of brainlike processing structures also highlights the importance of regularity in determining how difficult it is to build up stable representations of letter–sound relationships (Harm et al., 2003). Such modeling has been able to explore how the underlying deficit might be corrected in an irregular language such as English, as well as identify the stage at which remediation of different sorts is likely to be successful (Ziegler et al., 2007). This has led to the development of phoneme practice as a form of remediation, and evaluation of the brain activation changes in children who respond positively compared to those who do not. This work indicates that normalization is achievable in the majority of cases. Studies using electroencephalography (EEG) measures of electrical activity in the brain have also made it possible to identify children at risk of dyslexia before they begin to learn to read (Lyytinen et al., 2001), so that remediation can begin at an early age. Similarly, while it has long been known that reading difficulties tend to run in families, candidate genes that may affect brain development have now been found

(Schulte-Körne et al., 2007). Such understanding makes it possible to contemplate yet earlier forms of remediation.

Number and Mathematics

As with reading, it is the earlier stages of mathematics learning that are best understood. The focus of most research has been on the growth of children's understanding of number and arithmetic (the concrete use of number within the basic operations of addition, subtraction, multiplication, and division) in the preschool and elementary school years, since (as with word reading) this basic level underpins later achievement.

There is good evidence that infants (in common with other primates) possess two innate perceptual systems for recognizing number. One system enables them to distinguish, without counting, small exact quantities up to 3 – what is termed subitization. The other allows them to judge relative differences in large quantities (e.g., 16 vs 32) (Xu and Spelke, 2000). These systems have been argued to provide a foundation for the acquisition of verbal counting skills in the same way that phonological awareness provides a start point for reading. The linkage between perception of number and counting may be less exact, however, since children have to grasp new principles that apply only to the latter (Gelman and Gallistel, 1978). These include the one-to-one principle (each object to be counted only gets one count word), the stable-order principle (count words get used in fixed order), the order irrelevance principle (the order in which objects are counted makes no difference to the total), and the cardinality principle (the last count word used represents the total number of things in the set). Current evidence suggests that grasp of these principles is relatively hard-won on the basis of prolonged practice and that children move more easily from the innate systems to estimation (Gilmore et al., 2007).

Basic counting ability is usually in place by the time children go to school, and it forms the foundation for the growth of arithmetic skills (Hulme and Snowling, 2009). However, these skills are more disparate than those involved in reading, and they include single- and then multiple-digit addition, subtraction, multiplication, and division, along with varying strategies for carrying out these computations; understanding of tens and units, and the role of place value in distinguishing between these; translation between arithmetical problems presented in concrete, verbal, and numerical formats; knowledge of number facts (e.g., any number multiplied by 10 is equal to the original number shifted one place leftward with a 0 on the end); and derived fact strategies (e.g., use of the multiplication by 10 fact to decompose sums into workable subproblems).

There appears to be no consistent order for the emergence of these components, or any particularly strong relationship between them as they are acquired (Dowker and Sigley, 2010), so children may show surprising gaps and a wide range of individual variation. Dyscalculia, a basic deficit in the ability to learn arithmetic, was identified over 30 years ago, but the research on inherited number awareness has led to recent developments in understanding. The neural pathways involved in typical numerical capacities are now relatively well understood (Castelli et al., 2006), and dyscalculia appears to be due to a deficit in this number awareness associated with functional and structural abnormalities in a region of the parietal lobe known to be specialized for enumeration. This is reflected in poor performance on tasks such as saying how many dots there are in a visual display, or comparing two displays to say which has more (Landerl et al., 2004). Less is known as yet about the precise nature of this deficit than is the case for dyslexia, but research has led to interventions that are aimed at strengthening number sense via practice with the kinds of displays used to diagnose problems (Butterworth and Yeo, 2004). These appear to be more effective than more traditional interventions for children with number problems, which focus on simple repetition of number–object correspondences: it is not counting as such that is the source of the problem.

However, the relationship between arithmetical skills and innate abilities appears to be weaker than it is for reading. Achievement in arithmetic has been found to be more strongly and consistently related to knowledge of number facts and working memory ability (Cowan et al., 2009; Soltesz et al., 2010). Given the role of attention and executive control in counting and computation, and the way in which knowledge impacts the strategies used for solving problems, this is unsurprising. It does, however, make it harder to tell as yet what kinds of teaching strategies might be most effective, especially for learners with less distinct deficits.

Science Learning

As with reading and number, the cognitive capacities on which science understanding depends originate in the preschool period, and again key aspects appear to be innate. There are three elements of learning in science: factual knowledge (evidence), grasp of scientific procedure (generation of evidence), and conceptual grasp (understanding of phenomena). Despite the emphasis in much teaching on the first, and in research on the development of scientific thinking on the second, it is concepts that are central, especially during elementary school, since they provide unifying ideas about causal processes that help learners make sense of subsequent experience. These are essentially the ideas or theories that eventually provide the basis for meaningful scientific testing, and they therefore need to be put in place first.

There is good evidence that infants have an innate sensitivity to causal information, so that, for example, when one object strikes another and the second object moves, this is naturally perceived as the first object causing the second to move (Gopnik et al., 2004). Neuroscience evidence supports this claim, indicating a brain basis for the organization of causal perceptions (Satpute et al., 2005). By 2 years of age, children are capable of employing retrospective assessment of patterns of association between cause and outcome in order to anticipate the consequences of a specific event, and they do so with considerable statistical accuracy (Schulz et al., 2007).

However, this early understanding is not the resource for subsequent science learning that it has the potential to be.

While innately organized perceptions may provide a starting point, they are based on implicit sensitivity to covariation between events, and are not automatically verbally accessible. They are therefore not the same as explicit concepts, which can be broken down into elements that can be described in language and subjected to analysis and discussion. Explicit concepts appear typically to have a distinct point of initial origin in conversations around everyday experiences, and to relate to causal perceptions in at best a distorted and piecemeal fashion that is resistant to change (Howe et al., 2012). EEG studies with adults reveal separate neural activations associated with accurate implicit perceptions and inaccurate explicit concepts (Kallai and Reiner, 2010), indicating that failure to integrate the two is a commonplace outcome.

As with number, then, the crucial educational step is the accurate mapping of explicit concepts onto implicit perceptions and the consequent calibration between observable regularities and inferred causal relationships. However, this mapping appears to be constrained in a number of ways. In adults, fMRI studies show that coordination of incoming observational data with preexisting concepts is strongly associated with activity in the dorsolateral prefrontal cortex, a region linked with working memory and executive function (Fugelsang and Dunbar, 2005). Deliberate executive control is known to be relatively late maturing (Zelazo et al., 2008), suggesting that before the age of 10, sheer management of cognitive load may substantially limit implicit–explicit mapping. Although direct evidence is as yet scant, it also seems likely that individual variation in attentional and working memory abilities further impacts the pickup and coordination of perceptual information. There may even be yet-to-be-recognized dyscalculic-like deficits in sensitivity that impose additional constraints for some learners. At the other end of the mapping process, there is clear evidence of the importance for conceptual progression of learners being externally resourced with explicit vocabulary and explanatory constructs for encoding and manipulating concepts, and experiencing events within which language and observation are brought together (Philips and Tolmie, 2007). Both types of experience vary from individual to individual to a far greater extent than exposure to count words and counting experiences.

In terms of educational approaches, this analysis suggests that it is important that elementary school teachers provide children with concrete experiences that (1) involve manipulation of causal events, (2) draw attention to the exact sequence within these (e.g., what happens when an object that barely floats is dropped into a tank of water?), (3) provide descriptions to help make this sequence explicit, and (4) connect these descriptions to explanatory constructs. In fact, cooperative group work in science typically embodies all of these features and has been shown to produce robust improvements in conceptual understanding (Tolmie et al., 2010). Its success has generally been attributed to its capacity to generate explanatory dialogue, however. The educational neuroscience approach suggests that the combination of observation and dialogue is in fact more critical, and that the group context may further assist learning by establishing a setting in which there is good reason to explicitly refer to observations and where the informational load is made more manageable by distributing it between individuals.

Implications

Section Do We Really Need Educational Neuroscience? illustrates that educational neuroscience is indeed capable of making a distinctive contribution both to our understanding of key areas of learning and to the development of evidence-based pedagogical practices that will improve outcomes for typical and atypical learners. Although I have focused on literacy development and number and science learning, it is important to note that the range of relevant work is considerably wider. For instance, there is a growing body of research on socio-emotional development from infancy through adolescence, and how this is impacted by the organization and refinement of neural structures relating to processing of facial information, mentalizing (the attribution of mental states to others), and sensitivity to acceptance and rejection by others (Blakemore et al., 2013). The key nature of later refinements in particular argues that forms of social education have a potential importance in high school curricula that is presently largely unrecognized. Another strand of this research has helped identify abnormalities in the amygdala that are associated with impaired emotional awareness and callous-unemotional forms of antisocial behavior, suggesting alternative approaches to educational intervention with the most intransigent disruptive students (Jones et al., 2009). Research in the area of second-language learning has repeatedly demonstrated an important predictive role for low-level sensitivity to perceptual regularities in the language being learned (Frost et al., 2013), underscoring the importance of immersive pedagogies. Other pertinent work concerns shifts in adolescent sleep patterns and their implications for the timing of the school day (Lockley and Foster, 2012) and the development of attention and executive function (de Haan, 2013).

Across these different areas, the best work rests on a combination of behavioral and neuroscience methodologies, including both direct brain imaging and computational modeling. At present, effective interventions are based largely on behavioral work (much of this drawing on teacher expertise in its inception). However, by combining this with neuroscience, we are beginning to be able to explain better why these work and to build models of individual variation at a level of sensitivity that behavioral studies alone are not able to achieve (Thomas, 2013). In time, it is likely that wholly novel interventions that address this level of detail will begin to emerge.

There are other important implications, however. First, it is apparent that literacy, number, and science learning call upon a relatively unique constellation of functions and processes, with even the types of implicit–explicit mapping involved in each being distinctive in character. This may mean that there are likely to be few general principles underpinning successful learning across different areas. We need to keep looking for such principles as they may be beyond current conception, and more coordinated research activity is crucial for this reason. Even then, research in each area will probably remain at best only loosely interconnected for the immediate future.

The key objective at this stage is to map the skills necessary to progress in focal areas of the curriculum, the main steps involved in developing these, and the types of clinical and

nonclinical variations that can arise. These mappings can then be used to derive evidence-based strategies for promoting learning in each area and managing individual variability. There is a particular need for greater emphasis on mainstream development in order to extend the potential benefits of educational neuroscience to broad populations, rather than just those with specific deficits.

These points also mean that there are unlikely to be any effective overarching pedagogies – if different areas of education rest on different functions and skills, these will typically require different specific methods of support, and these may in turn differ according to individual need. The training of teachers and provision for support of their continuing development will therefore need to shift emphasis to promoting understanding of the range of evidence-based models that are applicable to their working context, placing them in a position somewhat akin to that of a medical general practitioner. This is a radical change, and given the scale of the new knowledge to be understood and deployed, it may entail a further need to strike a different balance between school-based and out-of-school education, with the former concentrating on provision of finely honed development of key skills, which are then elaborated according to personal choice in a diverse range of other settings. This pattern already exists to some extent, and it shows signs of accelerating. This in turn suggests a need for governments and wider society to come to a more explicit consensus on what the objectives of education should be, and how far to embrace (and resource) the evidence-based approach at the core of educational neuroscience.

See also: Bayesian Models in Neuroscience; Biological Preconditions for Language Development; Cognitive Development, Educational Theories of; Concept Formation in Childhood: Cross-Cultural Perspectives; Connectionist Models of Concept Learning; Developmental Behavioral Genetics and Education; Dyslexia, Developmental; Executive Functions During Childhood, Development of; Functional Brain Imaging of Language Processes; Mathematical Concepts During Early Childhood Across Cultures, Development of; Mathematics Education; Numbers and Number Sense; Reading Skills, Acquisition of: Cultural, Environmental, and Developmental Impediments; Science Education; Scientific Concepts During Childhood, Development of.

Bibliography

Blakemore, S.-J., Cohen Kadosh, K., Sebastian, C.L., Grossmann, T., Johnson, M.H., 2013. Social development. In: Mareschal, D., Butterworth, B., Tolmie, A. (Eds.), Educational Neuroscience. Wiley-Blackwell, Chichester, UK, pp. 268–296.

Bruer, J.T., 1997. Education and the brain: a bridge too far. Educational Researcher 26 (8), 4–16.

Butterworth, B., Yeo, D., 2004. Dyscalculia Guidance. NFER-Nelson, London, UK.

Castelli, F., Glaser, D.E., Butterworth, B., 2006. Discrete and analogue quantity processing in the parietal lobe: a functional MRI study. Proceedings of the National Academy of Sciences of the United States of America 103, 4693–4698.

Chipman, S., 1986. Integrating three perspectives on learning. In: Friedman, S.L., Klivington, K.A., Peterson, R.W. (Eds.), The Brain, Cognition, and Education. Academic Press, San Diego, CA, pp. 203–209.

Cowan, R., Donlan, C., Shepherd, D., Cole-Fletcher, R., 2009. Variation in number proficiency is not just a matter of fact. BPS Developmental Psychology Section Annual Conference, Nottingham, UK.

de Haan, M., 2013. Attention and executive control. In: Mareschal, D., Butterworth, B., Tolmie, A. (Eds.), Educational Neuroscience. Wiley-Blackwell, Chichester, UK, pp. 325–348.

Demonet, J.-F., Taylor, M.J., Chaix, Y., 2004. Developmental dyslexia. Lancet 363, 1451–1460.

Dowker, A., Sigley, G., 2010. Targeted interventions for children with arithmetical difficulties. In: Cowan, R., Saxton, M., Tolmie, A. (Eds.), British Journal of Educational Psychology Monograph Series II: Psychological Aspects of Education – Current Trends: Number Development and Difficulty. British Psychological Society, Leicester, UK, pp. 65–81.

Frost, R., Siegelman, N., Narkiss, A., Afek, L., 2013. What predicts successful literacy acquisition in a second language? Psychological Science 24, 1243–1252.

Fugelsang, J.A., Dunbar, K.N., 2005. Brain-based mechanisms underlying complex causal thinking. Neuropsychologia 43, 1204–1213.

Gelman, R., Gallistel, C.R., 1978. The Child's Understanding of Number. Harvard University Press, Cambridge, MA.

Gilmore, C.K., McCarthy, S.E., Spelke, E., 2007. Symbolic arithmetic knowledge without instruction. Nature 447, 589–591.

Gopnik, A., Glymour, C., Sobel, D.M., Schulz, L.E., Kushnir, T., Danks, D., 2004. A theory of causal learning in children: causal maps and Bayes nets. Psychological Review 111, 1–31.

Harm, M.W., McCandliss, B.D., Seidenberg, M.S., 2003. Modeling the successes and failures of interventions for disabled readers. Scientific Studies of Reading 7, 155–182.

Howe, C., Tavares, J.T., Devine, A., 2012. Everyday conceptions of object fall: explicit and tacit understanding during middle childhood. Journal of Experimental Child Psychology 111, 351–366.

Hruby, G.G., 2012. Three requirements for justifying an educational neuroscience. British Journal of Educational Psychology 82, 1–23.

Hulme, C., Snowling, M.J., 2009. Developmental Disorders of Language Learning and Cognition. Wiley-Blackwell, Chichester, UK.

Jones, A.P., Laurens, K.R., Herba, C., Barker, G., Viding, E., 2009. Amygdala hyporeactivity to fearful faces in boys with conduct problems and callous–unemotional traits. American Journal of Psychiatry 166, 95–102.

Kallai, A.Y., Reiner, M., 2010. The source of misconceptions in physics: when event-related potential components N400 and P600 disagree. European Association for Research on Learning and Instruction Neuroscience and Education Special Interest Group Biennial Meeting, Zurich, Switzerland.

Landerl, K., Bevan, A., Butterworth, B., 2004. Developmental dyscalculia and basic numerical capacities: a study of 8–9 year old students. Cognition 93, 99–125.

Lockley, S.W., Foster, R.G., 2012. Sleep: A Very Short Introduction. Oxford University Press, Oxford, UK.

Lyytinen, H., Ahonen, T., Eklund, K., Guttorm, T.K., Laakso, M.L., Leinonen, S., Leppanen, P.H.T., Lyytinen, P., Poikkeus, A.M., Puolakanaho, A., et al., 2001. Developmental pathways of children with and without familial risk for dyslexia during the first years of life. Developmental Neuropsychology 20, 535–554.

Mareschal, D., Butterworth, B., Tolmie, A. (Eds.), 2013. Educational Neuroscience. Wiley-Blackwell, Chichester, UK.

Philips, S., Tolmie, A., 2007. Children's performance on and understanding of the balance scale problem: the effects of parental support. Infant and Child Development 16, 95–117.

Royal Society, 2011. Neuroscience: Implications for Education and Lifelong Learning. Royal Society, London, UK.

Satpute, A.B., Fenker, D.B., Waldmann, M.R., Tabibnia, G., Holyoak, K.J., Lieberman, M.D., 2005. An fMRI study of causal judgments. European Journal of Neuroscience 22, 1233–1238.

Schulte-Körne, G., Ziegler, A., Deimel, W., Schumacher, J., Plume, E., Bachmann, C., Kleensang, A., Propping, P., Nöthen, M., Warnke, A., et al., 2007. Interrelationship and familiality of dyslexia related quantitative measures. Annals of Human Genetics 71, 160–175.

Schulz, L.E., Gopnik, A., Glymour, C., 2007. Preschool children learn about causal structure from conditional interventions. Developmental Science 10, 322–332.

Soltesz, F., Szucs, D., Szucs, L., 2010. Relationships between magnitude representation, counting and memory in 4- to 7-year-old children: a developmental study. Behavioral and Brain Functions 6 (13). http://www.behavioralandbrainfunctions.com/content/6/1/13 (accessed 03.06.13.).

Thomas, M.S.C., 2013. Educational neuroscience in the near and far future: predictions from the analogy with the history of medicine. Trends in Neuroscience and Education 2, 23–26.

Tolmie, A., Topping, K., Christie, D., Donaldson, C., Howe, C., Jessiman, E., Livingston, K., Thurston, A., 2010. Social effects of collaborative learning in primary schools. Learning and Instruction 20, 177–191.

Xu, F., Spelke, E., 2000. Large number discrimination in 6-month old infants. Cognition 74, B1–B11.

Zelazo, P.D., Carlson, S.M., Kesek, A., 2008. The development of executive function in childhood. In: Nelson, C.A., Luciana, M. (Eds.), Handbook of Developmental Cognitive Neuroscience, second ed. MIT Press, Cambridge, MA.

Ziegler, J., Castela, C., Pech-George, C., Georgel, F., Alarioa, F.-X., Perry, C., 2007. Developmental dyslexia and the dual route model of reading: simulating individual differences and subtypes. Cognition 107, 151–178.

Relevant Websites

http://www.aera-brain-education.org/ – Brain, Neurosciences and Education Special Interest Group, American Educational Research Association.

http://cen.squarevale.com/wordpress/ – Centre for Educational Neuroscience, London, UK.

http://www.bristol.ac.uk/education/people/group/centre/3125 – Centre for Mind and Brain in Educational and Social Contexts, Bristol, UK.

http://www.cne.psychol.cam.ac.uk/ – Centre for Neuroscience in Education, Cambridge, UK.

http://www.imbes.org/ – International Mind, Brain and Education Society.

http://www.earli.org/special_interest_groups/22._Neuroscience_and_Education – Neuroscience and Education Special Interest Group, European Association for Research on Learning and Instruction.

Neurosociology

David D Franks, Virginia Commonwealth University, Richmond, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article begins with a description of neurosociology, its basic unit of analysis, and how the name was coined. This is followed by generalizations about the brain – its complexity, its disadvantages, and how emotion organizes it. After placing the brain in evolutionary perspective, the concept of emergence is used to address the mind/body problem. The importance of the 'new' unconscious is discussed and Damasio's finding that emotion is necessary for rational decision making is detailed. Other subjects include consciousness and how the same things feel different to people. The mind and how it changes the brain is a current topic in neuroscience, and mirror neurons show how we are joined socially.

What Is Neurosociology?

Neurosociology is the study of the social nature of our brains as they developed through evolution because human intergroup cooperation had significant survival value. According to neuroscientist Leslie Brothers (1997), "We take the first step by acknowledging that the network of meanings we call culture arises from the joint activities of human brains…it (mind) cannot be derived from any one brain in isolation." What brings this further in line with sociological social psychology is her statement (1997: p. 39) that "While our individual brains are singular and self-contained, the processes on which they depend for functioning are social ones…there is no fully working brain without other brains."

The usefulness of neurosociology to its parent field of sociology can be seen partly as clarifying previous misunderstandings about biology. Sociology grew up in a time of clear divisions between nature and nurture. It defined itself as being opposed to natural instinct and insisted that there were as many human natures as there were cultures. Neurosociology shows clearly that we have a human nature and that it is social, however impressive the cultural variety of *Homo sapiens* may be. This means, as argued by scholars such as de Waal (2009), that the extreme individualism of our westernized culture, and more specifically our economy, is not tailored to our social natures and can be self-destructive, as events such as the Enron collapse (de Waal, 2009: p. 39) and our present economic predicaments demonstrate.

There was a time when sociologists were very critical of another area of biology – genetics. Rather than making either/or assertions about the effect of genes on human behavior, we now seek understanding in the interactions of genes and physical and social environments. A gene without experience in an environment is not a working gene. Even morality is seen to have a brain-based foundation that is part and parcel of social life.

For example, Gazzaniga argues that the human brain is by nature moralistic and that cross-cultural morals derive from the very social capacities that helped *H. sapiens* survive in the early days of our evolution. Churchland (2011: p. 59) also sees social and moral behaviors as parts of a single continuum. She notes that the same regions of the prefrontal cortex increase in activity whether the subject witnesses a purely social event or a clearly moral one.

Universal ethical issues focus on what types of killing are defined as murder and what forms of sexual intercourse are incestuous, what constitutes lies, and what it means to break promises.

Brothers and others point to the many social tendencies that infants have wired into their brains at birth (1997: pp. 12–13). She cites a study of 40 infants with a median age of a mere 9 min old who are more likely to turn their heads and eyes to an upright picture of a face rather than to an upside-down one, a scrambled one, or a blank one. She points to other studies showing that 7- to 11-week-old babies had a slight tendency to attend to the region of the eyes when a person was talking rather than to the constantly moving mouth. This should draw the most frequent attention, but it was the emotionally expressive eyes that drew interest.

> There is no question, Brothers argues, that when infants are simply fed by a bottle and are without other social and affective interactions, they do not develop emotionally. Because emotion is a critical organizer of the brain, they do not develop cognitively and become depressed and anxious. Such infants do not develop interest in, or even a capacity for, social activity. In its place is fear, anxiety, and depression and in many cases ordinary infant diseases lead to death. Tredway et al. (1999) have shown the specific brain processes that make this happen.

These examples provide important correctives to sociology while strengthening greatly the evidence of the profound degree of our social natures.

A History of the Name and Neurosociology's Basic Unit of Analysis

As early as 1972 Joseph Bogan, who was instrumental in the first split-brain research, and sociologist Warren TenHouten (2013) used the term neurosociology. The next year TenHouten and Charles Kaplan published the first sociologically oriented monograph dealing with neuroscience, *Science and its Mirror Image*. A significant number of sociologists began to think past E.O. Wilson's reductionist version of sociobiology and there was enough neurosociological work for Franks and Smith to edit *Mind, Brains and Society: Toward a Neurosociology*. In 1992, cognitive psychologists Cacioppo and Bernston used the term social

neuroscience to describe their entry into the social aspects of the brain.

The basic unit of analysis for neurosociology must be interactional, including at least two actors, whereas social neuroscience is not as constrained.

Theodore Kemper has referred to academia's 'fortress mentality,' wherein sociology and psychology are seen as being at odds with each other, but in brain science such a stance is counterproductive, and there is much we can learn from cognitive psychologists, especially those working in embodiment.

Important Things to Know about the Human Brain

Complexity

Edelman says that the cerebral cortex, covering only two-thirds of the brain's mass, contains at least 30 billion neurons, with 1 million billion associated connections or synapses. (A single neuron is merely a piece of meat; it requires other neurons to make a functioning neuron.) Thus, it is the synapses that matter. Counting one connection a second, it would take 32 My to count them all. This includes only the brain's neocortex, 80% of the brain. Eagleman (2012) writes that "there are as many synapses in a single centimeter of brain tissue as there are stars in the Milky Way Galaxy. Each of these cells is as complicated as a city."

According to Brothers, comprehending such complexity will have to be facilitated by a unifying theoretical framework specific to the complex workings of the brain. Examples of this kind of theory are "an atomic theory of matter, the germ theory of disease, a genetic theory of inheritance…a natural selection theory of evolution…" (Brothers, 2001: p. 67).

The Limitations of the Brain

The brain is not built to solve all the problems that experience brings to us. Perception is selective and memory is riddled with falsehoods. This lack of perfection – or even approximations of it – occurs because the brain is a 'tinkerer.' It can build only on brain structures that its past development have made possible. For example, the brain-circuits supporting consciousness were very small in our evolutionary past and had to borrow from and in some sense deplete circuits that existed next to them. We make a serious mistake when we see the 'higher' and newer prefrontal lobes as developing separately from the deeper limbic emotional system. The new works minimize and constrain the old, and the old works raise and influence the new.

The Brain as Relational

The brain is not sufficient unto itself. A functioning brain needs a social and physical environment. Persons placed in sensory deprivation tanks for extended periods of time have hallucinations and can have a number of distressing symptoms. Brain cells die if they are not used to act on the world in daily life and relations. "Use it or lose it" is not just a cliché. Children whose temper tantrums go unchecked simply do not develop the neuronal pathways from the prefrontal cortex going down to the emotional limbic system needed to control themselves. This problem is augmented by the fact that circuits going upward from the limbic system to the prefrontal cortex are much more robust than those going down.

Projection

Neuroscience has shown that any epistemology assuming that our brain-driven minds somehow passively copy what is 'out there' is misguided. This includes analytical philosophy, any kind of correspondence theory, and Locke's British empiricism, wherein experience 'writes itself' on the mind, which he conceived as a 'tabula rasa.' Our brains determine as much of what we observe as do the objects themselves. This is because our senses are transducers. Transducers take energy from one system and change it into a different form that will accommodate that system. The human eye takes light waves or amounts of electromagnetic energy and transforms these waves into what we experience as color. Without ears, there are compressions of air, but it takes an ear and a brain to make noise. True, stimuli are initiated independent of our will and in this sense there is an objective world, but how it is experienced depends on the brain/world interaction. According to Buckminster Fuller, what we can sense is less than a millionth of what is 'out there.' Various animals truly live in different worlds.

In sum, the brain selects, edits, and changes stimuli into what it can accommodate. It sees patterns where there are none, and much of the brain is constructed to receive only the gist of things.

The Brain Is Divided into Two Different Halves

Roger Sperry (1965) discovered that the two halves of the human brain were complementary to each other. The left side was logical, conscious, and speech driven. The right side was given to gestalts and 'getting the gist of things.' Sperry and his student Michael Gazzaniga treated epilepsy patients by severing the corpus callosum, which allows communication between the right- and left-brain hemispheres. The corpus callosum is a massive cable of 200 million fibers enabling the fully linguistic left-brain (in right-handed people) to know what the characteristically nonlinguistic right brain is doing. The mute right brain communicates by electrochemical means. Sensory information from the patient's left side is processed by the right side of the brain and vice versa.

After such severing, the right side can no longer tell the conscious left side what it is doing. Thus, a patient can be given a written message flashed to the right side of his face instructing him to draw a dog. But the severed corpus callosum prevents this communication to the conscious left hemisphere. This leaves one person with two brains, one of which is ignorant of the other. Patients had no clue about what they were drawing until its form becomes obvious during the sketching. Only then will they realize they are drawing a cat or dog, etc. When the right brain is asked why the patient drew it, the left, usually 'linguistic brain,' contrives an answer that makes some sense to the patient (who is the only one deceived). In another case, patients' right brains were asked to laugh and then they were

asked what was so funny. The patients never said they did not know why they were laughing. A reason was always forthcoming that only the patients themselves could believe because the researchers knew they were actually only following their directives for the research.

In other places I have argued that this provides evidence for notion of 'accounts.' What sociologists refer to as vocabularies, motive neuroscientists call confabulations.

Measures

If the human brain is imperfect, the measures we have for it are imperfect as well. There are a number of tests currently available and in wide use. The CAT scan uses X-rays taken from various angles, and they are often combined to produce three-dimensional (3D) images. The letters CAT stand for computerized axial tomography; PET stands for positron emission tomography. Small doses of radioactive traces are absorbed by the body part in question via the veins. The tracer substance releases energy, which is converted into 3D images. One advantage of PET scans is that only minimal doses of dangerous radiation are required, but the high cost of the short-lived radiation and the procedures to synthesize them before they decay mean that many hospitals cannot afford the equipment, making scheduling of the tests difficult.

Magnetic resonance image (MRI) measures give still pictures of brain structure. MRIs are used to correlate structural brain changes or differences in behavior.

Functional magnetic resonance images (fMRIs) give noninvasive 'movies,' or pictures of brain functioning, whereas MRI measures give only still pictures of brain structure. What is really measured is the oxygen levels in the brain. When neurons fire they take up oxygen, and this process is what is measured. fMRIs use a software program called the BOLD (blood-oxygen-level dependence) technique to measure the ratio of oxygenated to deoxygenated hemoglobin in the bloodstream; this simply correlates with neuronal processes in certain areas. It does not take direct single measures of neurons. Compared to other electrophysiological techniques, fMRIs have significant limitations, creating the need for extensive post-test processing. Published results are often averaged over time and smoothed across space using available software packages. Nevertheless, for better or worse, researchers have developed confidence in these measures because intrusive changes made purposely in the brains of rats have been correctly picked up by fMRI measures.

The New Unconscious

Since the 1970s, empirical approaches to the study of the unconscious have differed from Freud's notions, with exception of the defense mechanisms. Approximately 98% of our brain processes are unconscious. So 2% of the brain is conscious, but given the vastness of the human brain, this is more than enough. The evidence for the breadth of the unconscious is such that it demands attention. Most of what the brain does is 'procedural,' having to do with routine processes, such as breathing, perceiving 'unities' rather than separate attributes of objects, and learning balance, grammar, and feeling. The brain does these things automatically, which leaves us free for other critical tasks. The other dimension of the unconscious has to do with 'content,' i.e., unconscious fears and cognitions, defense mechanisms, emotions, and memories. When patients with amnesia were pinpricked or shocked by doctors who saw them soon after the onset of amnesia, the patients shied away from the doctors, even though they had no conscious memories of the abuse. In another study, patients with amnesia were placed with three groups of studiers' confederates: The first group was instructed to be pleasant to the patients; then the patients were placed with others who were instructed to be brusque and then with those who were neutral. The patients were shown photos of the three groups, and they easily picked out those they liked and disliked. Consciously, they had no memories of these experiences but unconsciously, they did.

The first undeniable evidence of unconscious emotion came from Zajonc, who showed that mere exposure to a familiar picture would create a positive feeling for it. Here, exposures of the pictures were presented too quickly for subjects to be aware of them.

Evolution of the Human Brain

The data of evolution are always open to doubt because the fossil record can never tell how many other examples existed before a particular finding. It gives only a minute portion of the actual history. Historically, brain size and its relation to intelligence have been given emphasis, but brain size alone does not produce intelligence, and intelligence can take many forms. Coordination with others and imitating others making stone tools were important to survival. Treating the brain, as it is isolated in the skull instead of functioning socially involving other brains, is an error according to Brothers (2001), who labels this 'neuroism.' The origins of human intelligence are, therefore, better sought in the environmental conditions that shape social life. The environmental changes moving early humans from the forest to the savannah put a survival premium on practical intelligence, emotional control, and social cooperation. Remains were found in the Kenyan savanna of the first of our forefathers, little Australopithecines with brain capacities of 450 cc, less than one-third the size of the brain capacities of *H. sapiens*. Their brains were rounder than those of chimpanzees, and their bone structures implies that they walked upright about 4.4 Ma.

Chimpanzee kinship ties are supported by mutual grooming, which produces opiates and increases solidarity. Group size is limited by the time needed for grooming. The larger the group, the more intergroup skills are needed. This puts survival pressure on these same skills, and it is critical for small animals whose size makes them dependent on forming alliances. Some experts suggest that Australopithecine communication was largely emotional, and emotion still organizes the hominid brain. Emotion also lays the basis for developing cognition and emotional control. The social emotions are, of course, central to group life. All these early characteristics formed the basis of our current social brains and impose limits on them.

Given the dangers of the times, it was probably healthy for the Australopithecine brain to react without much cognitive anticipation. These limits meant that they never wandered off to new environmental niches and were therefore mentally shaped by their environment. Nonetheless, they lasted more than 3.5 My.

Homo habilis arrived at least 2 Ma, and many think they were not that different from the numerous Australopithecines who used 'near' tools sufficient to break bones and eat their marrow. The stone choppers of *H. habilis* were useful only for scavenging. Nonetheless, their brains tripled in size during the one million years they lived. During that time, *H. habilis* never increased the efficiency of their tools. Nor did they have the cognitive capacity to imagine traveling to new places. Brain size does not drive innovation; innovation drives brain size.

Homo erectus arrived on the evolutionary scene 1.8–1.5 Ma. Although brainpower doubled to 1100 cc, they had no vocal organs to develop speech. The frontal and parietal lobes increased, flattening their faces. Females grew larger to accommodate the larger heads that had been developed, so the women could accommodate childbirth. Birth was still dangerous, but the large brain size increased brain development as well as infant dependence, which made things more dangerous for children and mothers. Breasts were enlarged and women were constantly receptive sexually. The brain was also lateralized. The right side allowed emotional communication and increasing sociality. The inventory of tools plus their variety increased. This allowed *H. habilis* to migrate to southern France in warm weather and return to northern Africa in the winter. This required the ability to make mental maps. They lasted one million years with their numerous tools and made no further advancements.

The Neanderthal and *Heidelbergensis* competed with *H. erectus*. The Neanderthal was more robust, more thickly boned, and more muscular. Their brains increased to 1400 cc – larger than the brains of *H. sapiens* brain.

H. sapiens developed in Africa 150 000 years ago. Between 200 000 and 40 000 years ago, they began to talk. Charles Li and Jean-Marien Hombert suggest there is a confluence of evidence that this date was as recent as 80 000–60 000 years ago. For Australopithecine, the amount of time spent grooming others determined group size, but language made grooming unnecessary, and group size was determined by food supplies. Domestication of animals allowed for fixed settlements, expansion of group size, and the advent of cities around 10 000–12 000 years ago.

Consciousness about one's self, was indicated 43 000 years ago when the oldest decorative beads were made. They were brought from great distances that gave them value and reflected concern about one's status and identity.

The Social Nature of the Brain

Leslie Brothers (1997) laid the verbal foundation for neurosociology. The term 'social editor' is Brothers' way of presenting the brain's social structuring. A critical structure here is the amygdala.

The amygdala is critical to socially related memories, as is the hippocampus. A lesson learned but forgotten is not really a lesson learned in the larger picture.

In cases of post-traumatic stress, the amygdala may be quick to learn, but it makes the situation hard to forget. The critical importance of the amygdala is that it enables so much. This includes fear, assessing facial expressions, attachment, unconscious learning, early emotional memory, and emotion across the life span. One should not, therefore, be surprised at its complexity, despite its small size. The amygdala exemplifies how the brain is not perfect. Known as a very fast warning system, it operates at speeds of less than 100 ms, allowing us to know whether an object is safe or not before we know what it is. At such speeds, it is often wrong. But when it is right, it can save our lives. It also assesses the emotions of others in terms of whether they are safe or dangerous. The speed of social interaction is so fast that we need such assessments, and the faster the better, according to many, including Brothers (1997: p. 12).

Mirror Neurons

A mirror neuron fires when we passively observe others perform certain actions. Next, it actively simulates the others' actions unconsciously in the observer's motor cortex. In an important sense, we act behaviorally according to what we watch. These neurons are inherently social and are directed toward the actions of others; they also actively combine motor activity and perception. Furthermore, and most important, mirror neurons fire when the action observed can be seen as reflecting intention on the part of the observed actor. Because it is unconscious, this response prepares for the lightening-quick responses necessary for social interaction. Some researchers have questioned whether mirror neurons are actually present in human beings because measures are so indirect and are limited to the amount of oxygen released in certain areas of the brain. Single-neuron measures are destructive of brain tissues surrounding the neurons in humans. The first direct recording of human mirror neurons used 21 patients with pharmacologically intractable epilepsy. Such research is obviously limited to what is possible in those kinds of cases. Even with that limitation, the researchers found that mirror neurons were more widely dispersed within the human brain than previously thought (Franks and Davis, 2012).

Mirror Neurons and Pragmatism

For researchers in the field of mirror neurons, the relationship between our motor intentions and the way we perceive the world is an 'affordance.' An affordance is something that the environment offers to make behavior possible, such as surfaces that provide support, objects to be manipulated, and substances that can be eaten. The Chicago pragmatists made the same point. They saw the epistemological link between the knower and the known as behavior. It is interesting that this is confirmed by current research into mirror neurons.

Researchers Rizzolatti and Sinigaglia (2008) argue that the visual perception of an object implies the immediate and automatic selection of those properties that facilitate our behavioral interaction with it. We perceive the world not so

much as it is in some final sense, but in terms of how it answers to or facilitates our actions.

Furthermore, our senses of and reasoning about the world are limited. The environment becomes constructed and objectified only in relation to our limited motor capacities, senses, and brains. The German term *Umwelt* captures this organism/environment relationship. It refers to the world carved out for our attention by our own capacities, sensitivities, and motor repertoires.

The research into mirror neurons implies the existence and relevance of just this social behaviorist view. The mirror neurons of dancers respond most strongly to observations of familiar dance steps in line with the style of dancing the observers routinely used. More pointedly, human subjects were shown videos of the motor behaviors of various animals communicating, such as monkeys lip-smacking, dogs barking, and humans talking; mirror neurons became engaged only during the observation of humans talking, even though visually, it was obvious that all behaviors were communications involving moving mouths. Because we have no spaces in our motor cortexes for barking and lip smacking, we cannot simulate them.

Mirror Neurons and Speech

For the social behaviorist George Herbert Mead (1938), human thought and speech were clearly intangible universals. Words represent general terms, subsuming all real particulars. But the process is also behavioral. As humans solve problems, they are really talking to themselves, making subliminal movements with those muscles associated with speech, such as the tongue, lips, and larynx. Deaf children privately thinking to themselves likewise make subliminal hand and body gestures that underlie their symbolic thoughts. The congenitally blind, who have never seen gestures, nonetheless make motor movements when they talk. Gestures are not ancillary to talk but are inherent parts of it, just as grammar is. We even gesture when talking on the phone when no one can see us.

According to Jonathan Turner, thinking is also accomplished in a gestalt-based mode. The Broca and Wernicke areas of the brain change speech into the brain's way of thinking. Because talk is behavioral, we simultaneously enact in our motor cortex the same speech movements others are making. Their talk literally becomes our talk, with the provision that it is automatic and unconscious. The theory of 'embodied semantics' holds that concepts are represented in the brain by the same sensory–motor circuitry on which the concept relies. The general concept of grasping would be represented by increased activity in the sensory motor areas devoted to grasping actions. Language is biologically linked to action through mirror neurons in a way that is congruent with the priority given to action by the pragmatists. Pragmatists would insist that language is not separate from motor behavior but, rather, is inherently grounded in actions. For example, kicking as a conceptual generality would be represented in the sensory–motor space given over to the leg and foot. Even metaphors such as "we will kick off the new year with a party" are represented in these foot and leg areas of the motor cortex. According to these researchers, the hypothesis of 'embodied semantics' has been confirmed by studies of patients with traumatized parts of the motor cortex and fMRI studies that take moving images of the activation of brain parts.

Emotion and Rationality in the Human Brain

Another critical part of the social editor is the often overappreciated prefrontal cortex. It is overappreciated because it is treated as if it stands alone to produce our intellectual capacities. Damasio as well as philosophers and even artificial-intelligence workers have convincingly shown that without certain emotions, we are unable to make the simplest choices. Alone, the prefrontal cortex gives us nothing. Patients traumatized in the ventral medial parts of the prefrontal lobes were capable of anger but incapable of feelings of apprehension, discomfort, or hunches, called 'somantic markers,' that lead us away from bad social associations, especially untrustworthy partners. On the other hand, they were not able to have normal emotions. While interviewers wept at their stories of their familial and business downfalls, they, themselves, felt no such remorse, even though many such patients were quite intellectually capable. Damasio and his wife, Hannah, developed a card game to measure this in which high-paying cards were also risky ones, with some demanding severe payments resulting in losing the game. Unlike normal subjects, these patients never learned to avoid high-paying but high-risk cards because they did not have somantic markers to make them feel apprehensive and reject them. A complete discussion of replications and their pros and cons is available in TenHouten.

Imitation

It is not apes or monkeys who do the aping. Imitation is an ability of humans and results in our capacity to learn from others and thus to pass culture on to other generations. Imitation is not literal or superficial because it must be generalized from situations different from the one where the behavior was first learned. We need to know the intentions of the persons we are imitating. We also simulate others' mouth movements in our motor cortex. Thus, Iacoboni (2008) argues that evolution leads from action recognition through imitation to language.

Infant imitation is hard-wired. Among newborns only 32 h old, the imitation of facial movements is common, even though they cannot see their own faces. The scope of imitation is broad. We imitate those we like, and they like us more commonly if we imitate them. Research using confederates shows how readily participants imitate the smallest movements. Participants primed with stereotypes of the elderly walked more slowly than unprimed ones. Imitation is seen as an unconscious social glue, but it is fragile and must be sustained by role taking and other processes.

The 'correspondence problem' and its solution are prerequisites for a full understanding of imitation. What we see is dependent on our spatial perspective, so the mapping of the motor cortex must make 'corrections.' When, as adults, we imitate expressions, how do we know we are correct when we

can't see our own faces? Solving the correspondence problem involves answers to these questions.

Determinacy and Empirical Evidence for Agency

In 1964 Roger Sperry won the Nobel Prize for his thesis that mind, seen as an emergent from neuronal processes in the brain, can in certain instances exert a causal force capable of changing the parts from which it emerged. Jeffery Schwartz and Sharon Begley (2002) applied Sperry's ideas to human patients suffering from obsessive compulsive disorder (OCD). In about 1969 Edward Traub, who worked with macaque monkeys in an NIH laboratory in Silver Springs, MD, established the experimental foundations for Schwartz's work. Traub clashed with animal-rights workers and was charged with animal cruelty because of his treatment of the monkeys, although his practices were not unusual for the time.

Traub tested the hypothesis that monkeys, whose ability to experience sensations in their arms was removed surgically, retained the capacity for purposeful behavior in those limbs but only if they were forced to use them and then were immediately rewarded. The monkeys were fed only minimally to the point of near starvation; their functioning arms were restrained; and electric shock was administered to them if they did not try to use the disabled arm. Finally, the monkeys regained the use of their once paralyzed arms. Autopsies showed that the somatosensory cortex, controlling feeling from the body to the cognitive prefrontal cortex, had been rezoned using neurons from the face.

In 1992, after he was reinstated into his profession, Traub applied his monkey experiments to humans who had suffered from strokes. In what he called 'constraint induced therapy,' he tied 21 patients' good arms in slings and had them do strenuous exercises with their weak arms. These patients had been paralyzed for an average of four and one half years. They now received therapy for 6 h a day for only 10 days. After only a 2-week session, the use of their once paralyzed arms was recovered and remained so after 2 years. For patients with OCD, Schwartz developed a successful four-step program that included relabeling, reattribution, refocusing, and revaluing. Two brain areas became overactive in OCD that were clearly visible in PET scans. Seeing these areas facilitated relabeling. For example, one woman witnessing this exclaimed, "That's not me, that's my OCD!" She was able to deconstruct her overpowering feelings as brain-driven illusions and separate them from her real self.

The theoretical implications of this for social psychology are significant indeed. First we have the notion of mind as a cognitive symbolic process instead of a reified noun. This allows the conception of mind as an emergent from but different from the deterministic brain. This leaves room for a sound concept of agency. Fuchs (2001), for example, sees agency as a residual category resulting from unexplained variance and then attributed to the actor. The patients who were cured, however, are direct empirical examples of will and the cognition reflecting agency. They demonstrated that mind and symbolic processes are causal forces over matter. This finding joins with those of mirror neurons in helping to explain our natures.

Current Selected Findings Relevant to Neurosociology

Jeff Davis on Persistent Inequality

Davis presents a compelling argument regarding the effects of persistent social inequality on the brain. He contends that structural explanations fall short of explaining the microfoundations of social inequality. Davis uses Tilly's model of inequality to identify four major institutions that weave inequality into the fabric of society: exploitation, hoarding of opportunities, emulation, and adaptation. Davis focuses on the last two processes.

The unpredictable nature of an impoverished environment impacts neural functioning as do other stress producers, creating preferences for short-term rewards and fostering little or no planning for the future as well as aggressive behaviors and higher risks for teenage parenthood. Brain supports for the processes involved in avoiding these behaviors are the prefrontal cortex, the anterior cingulate cortex, and the vertebrate social behavior network. The author links neurotransmitters to decision making, and deficits in the latter are linked to poverty. This means that once in such an environment, actors will have a hard time getting out because of their own behaviors.

Unconscious Voice Signals of Status

Thirteen years ago Gregory reported his findings concerning how to separate and thus analyze various dimensions of the spectral frequencies in human voices. He and his colleagues found that frequencies below 0.5 kHz signaled social status accommodation and social convergence. These original experiments were performed on conversation dyads. In 2002, Gregory and Gallagher demonstrated that analysis of low-frequency spectra from speech in presidential debates could be used to predict winners of eight of the past US elections, from 1960 to 2000. These remarkable findings were successfully replicated in 2008 for the debates between Senators Obama and McCain.

Conclusions

Notably, it was neuroscientists working with sociologists who first formulated the case for neurosociology. The two examples emphasized herein were Bogen, writing with TenHouten, and later Brothers, advised by Perinbanayagam, a symbolic interactionist. It is to be hoped that such interdisciplinary efforts will continue in brain studies.

Two tasks are still left undone for the final maturity of neurosociology. Sociology must be able to inform neuroscience more than it has in the past, and doctorate programs in neurosociology must be developed.

See also: Cognitive Development in Infancy, Neural Mechanisms of; Cognitive Neuroscience; Cortical Development: Transplantation and Rewiring Studies; Emergent Properties; Emotion, Neural Basis of; Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of; Language Development: Evolutionary Theories; Levels of Analysis in Social Psychology; Limbic System; Mind, Theories of; Motor Cortex; Neural Basis

of Language Development; Neural Plasticity; Pragmatism: Philosophical Aspects.

Bibliography

Brothers, Leslie, 1997. Friday's Footprint: How Society Shapes the Human Mind. Oxford University Press, New York.

Brothers, Leslie, 2001. Mistaken Identity: The Mind-Brain Problem Reconsidered. State University of New York Press, Albany, New York.

Churchland, Patricia, 2011. Braintrust: What Neuroscience Tells Us about Morality. Princeton University Press, Princeton, New Jersey.

Eagleman, David, 2012. Incognito: The Secret Lives of the Brain. Panthion Books, New York.

Franks, David, 2010. Neurosociology: The Nexus between Neuroscience and Social Psychology. Springer Press, New York.

Franks, David, Davis, Jeff, 2012. Critique and refinement of the neurosociology of mirror neurons. In: Kalkoff, W., Thye, S. (Eds.), Advances in Group Processes: Special Issue in Biosociology and Neurosociology, vol. 29. Emerald Group Publishers, New York, pp. 71–117.

Fuchs, Stephen, 2001. Beyond agency. Sociological Theory 1 (40), 24–40.

Gazzaniga, Michael, 1998. The Mind's Past. University of California Press, Berkeley, CA.

Gazzaniga, Michael, 2012. Who's in Charge? Free Will and the Science of the Brain. HarperCollins, New York.

Gregory, Stanley, Kalkholf, William, 2012. Comprehending the neurological substratum of paraverbal communications: the invention of SplitSpectechnology. In: Franks, D., Turner, J. (Eds.), The Handbook of Neurosociology. Springer Press, New York, pp. 369–383.

Iacoboni, Marco, 2008. Mirroring People: The New Science of How We Connect to Others. Farrar, Straus and Giroux, New York.

Mead, George Herbert, 1938. The Philosophy of the Act. Chicago Press, Chicago.

Ramachandran, Vilayanur Subramanian, 2011. The Tell-Tale Brain: A Neuroscientist's Quest for What Makes Us Human. W.W. Norton, New York.

Rizzolatti, Giacomo, Sinigaglia, Corrado, 2008. Mirror Neurons in the Mind: How Our Minds Share Actions and Emotions. Oxford University Press, New York.

Schwartz, Jeffrey, Begley, Sharon, 2002. The Mind and the Brain: Neuroplasticity and the Power of Mental Force. Regan Books, HarperCollins, New York.

Sperry, Roger, 1965. Mind, brain and humanistic values. In: Platt, J.R. (Ed.), New Views in the Nature of Man. University of Chicago Press, Chicago, pp. 71–92.

TenHouten, Warren, 2013. Emotion and Reason: Mind, Brain, and the Social Domains of Work and Love. Routledge, New York.

Traub, Edward, 2004. Harnessing brain plasticity through behavioral techniques to produce new treatments in neurorehabilitation. American Psychologist 59 (8), 692–704.

Neuroticism

Weiting Ng, School of Arts and Social Sciences, SIM University, Singapore

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Watson, volume 15, pp. 10609–10612, © 2001, Elsevier Ltd.

Abstract

The personality dimension of neuroticism, as defined by the trait perspective, refers to relatively enduring tendencies to think and react in anxious and maladaptive ways, and feel more negative emotions. Research has suggested that this personality dimension is universal, having been replicated across cultures and having a strong biological basis. Neuroticism is a robust and strong correlate of various constructs (e.g., emotional disorders, cognitive processes, emotion regulation, and coping) and is also an important predictor of diverse life outcomes (e.g., well-being, mental and physical health, social relationships). Although it demonstrates rank-order stability, extensive evidence indicates that there is mean-level decrease in neuroticism over the life span.

Definition of Neuroticism: A Trait Perspective

Traits refer to generalized and personalized tendencies – that is, consistent and stable modes of an individual's adjustment to his or her environment. In contrast to the nomothetic approach, which studies trait units that apply to all people, the idiographic approach emphasizes that every individual is unique and different, and general principles cannot be broadly applied to all people to understand and predict their behaviors. Although the concept of traits originated from Allport, the main emphasis of the modern-day trait approach has now diverged. The trait approach is now geared toward a descriptive and comparison orientation rather than being grounded in explanatory purposes. Though lacking in accuracy, as the trait approach does not provide a comprehensive theory of personality, theories based on this approach provide diverse perspectives, ranging from viewing traits as descriptive concepts to viewing them as biologically based causal concepts. The merits of this approach are that it is a convenient method for assessing personality characteristics and comparing people, and it provides a heuristic way of making predictions and judgments.

The most popular and commonly used trait theories are based on Costa and McCrae's five-factor model (FFM) and the Big Five (Costa and McCrae, 1992; John and Srivastava, 1999). Factor-analytic studies originating from Goldberg's lexical tradition and factor analyses of personality questionnaires consistently converged to yield five factors (Costa and McCrae, 1992; Goldberg, 1990), and the lexical and questionnaire traditions were merged to produce a five-factor structure. The Big Five dimensions are extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience. Extraversion assesses the quantity and intensity of interpersonal interaction, and one's activity level, need for stimulation, and capacity for joy. Agreeableness refers to a prosocial and communal orientation toward others and assesses the quality of one's interpersonal orientation along a continuum from compassion to antagonism. Conscientiousness describes the extent that an individual possesses socially prescribed impulse control such as delaying gratification, following norms and rules, organization, and persistence. Openness to experience describes the breadth, depth, originality, and complexity of an individual's mental and experiential life. It assesses an individual's proactive seeking and appreciation of experience and exploration of the unfamiliar. Finally, neuroticism assesses an individual's adjustment versus emotional instability. This personality dimension identifies individuals prone to psychological distress, unrealistic ideas, excessive cravings or urges, and maladaptive coping responses. Individuals who are low on neuroticism are calm, secure, and emotionally stable, whereas those high on neuroticism tend to be emotionally unstable and show characteristics of being anxious, depressed, moody, and irritable. Based on the NEO Personality Inventory – Revised (NEO-PI-R), the FFM further differentiates the broad domain of neuroticism into the facets of anxiety, angry hostility, depression, self-consciousness, vulnerability, and impulsiveness (Costa and McCrae, 1992).

The five-factor theory (FFT) posits that these five basic dimensions correspond to basic psychological tendencies in every individual and are biologically based. These personality dimensions interact reciprocally with external influences (e.g., cultural norms, life events) to influence characteristic adaptations (e.g., an individual's thoughts, feelings, and behaviors) (McCrae and Costa, 2008). The FFT assumes that these basic tendencies are uniform across cultures and are universal; this universality might be attributed to the species-wide biological bases of traits or might represent a psychological consequence of the shared human experiences of living in societies where similar life tasks might be promoted by most modern cultures (McCrae and Costa, 1997).

Universality of Neuroticism

Across the myriad of personality theories, various researchers have conceptualized neuroticism distinctively, assigning different labels to it (e.g., neuroticism, negative emotionality, negative affectivity, emotional instability). Yet, despite the disparate labels and definitions, the common underlying characteristics across the various personality models signal that they are referring to a fundamentally similar personality construct. For instance, although Eysenck's three-factor personality theory comprises only three basic dimensions – psychoticism, extraversion–introversion, and neuroticism – neuroticism as conceptualized in his theory is similar to that

in the FFM. It also subsumes lower-level traits such as worrying, anxious, depressed, moody, and emotional instability. Eysenck's three-factor theory emphasizes the biological determinants for the three superfactors, and he hypothesized that the biological basis for neuroticism stems from the autonomic nervous system; those higher in neuroticism have lower threshold activation in the autonomic nervous system (Eysenck, 1990). Equivalence can also be drawn between Tellegen's three- (or four-) factor model and the Big Five model. Assessed by the Multidimensional Personality Questionnaire (MPQ; Tellegen, 1985), Tellegen's model comprises positive emotionality, negative emotionality, and constraint (and absorption). Commonalities exist between the negative emotionality factor and neuroticism – individuals high in negative emotionality have a low threshold for the experience of negative emotions and tend to be involved in antagonistic relationships. Factor analyses indicated that the Big Five constructs can be subsumed under Tellegen's four dimensions because negative emotionality encompasses neuroticism and (the inverse of) agreeableness. Therefore, despite its varying labels, the conceptualization and inclusion of the neuroticism construct across different personality theories suggests that it is a universal personality dimension.

The replication of the neuroticism construct across cultures reinforces the notion that it is a universal personality dimension (McCrae and Costa, 2008). The FFM personality structure (including neuroticism) has been consistently replicated across cultures. For example, the same FFM based on American adults was replicated with German, Israeli, Korean, and Chinese adults (McCrae and Costa, 1997). Lexical studies also found cross-language agreement on the five factors – studies of German traits (using German adjectives) replicated the English-language five factors and clearly recovered the FFM, and Chinese trait terms also yielded five dimensions that overlap considerably with the Big Five dimensions. In sum, various cross-cultural studies, including a large-scale observer rating study across 50 different cultures, provide clear evidence for the universality of the FFM across different instruments and measurement methods (McCrae and Costa, 1997, 2008; McCrae and Terracciano, 2005).

Another line of evidence that supports the universality of neuroticism is its biological basis. Twin studies comparing monozygotic (MZ) and dizygotic (DZ) twins provide some evidence that neuroticism has a biological basis and is partly heritable. MZ twins correlated much higher (0.46) than DZ twins (0.2) on neuroticism. However, results from adoption studies undermine these findings that highlight the importance of genetic influence on neuroticism. Correlations between biological parents and their adopted-away offspring were almost zero, suggesting that the high correlations between MZ twins might have been partly caused by their more similar environments (Plomin and Caspi, 1999). To circumvent this problem and control for environmental effects, researchers compared reared-apart and reared-together MZ twins. Though various studies diverge on their estimates of the heritability of neuroticism (some found lower correlations for reared-apart MZ twins (0.25) than reared-together MZ twins (0.41), whereas others found similar correlations for reared-apart and reared-together MZ twins), the overall evidence indicates that neuroticism is partly biologically determined and heritable to some extent (Pedersen et al., 1988; Plomin and Caspi, 1999). In recapitulation, the replication of neuroticism across cultures and personality models and its biological basis substantiate the conclusion that it is a universal personality trait.

Links between Neuroticism and Other Constructs

Emotional Disorders

Before the *Diagnostic and Statistical Manual of Mental disorders* (*DSM-III*), early DSM-based categories of mood disorders included the clinical label of neurosis, which comprised individuals showing anxiety and depressive disorders. The personality trait neuroticism, however, is independent of the construct of neurosis. Subsequent editions of the *DSM* stopped employing the diagnosis of neurosis and instead classified patients into specific anxiety or mood disorders. Yet, the growing evidence highlighting the strong associations between neuroticism and various mental disorders suggests that neuroticism may represent a broad, underlying syndrome common across the disorders (Barlow et al., 2013).

In their review, Barlow et al. (2013) summarized theoretical and empirical evidence underscoring the role of neuroticism in the development and course of emotional disorders. First, there are high rates of comorbidity among anxiety and mood disorders. Second, treatments for a specific disorder often improve comorbid anxiety/mood disorders that are not targeted, demonstrating broad treatment responsiveness that may possibly be due to common features among different disorders. Finally, neuroscience evidence suggests that emotional disorders may share a common biological syndrome related to hyperexcitability of limbic structures (e.g., amygdala overactivation, limited cortical inhibition of amygdala responses). Likewise, individuals high in neuroticism exhibit similar amygdala overactivation. These three areas support the conception that there are commonalities among anxiety and mood disorders, which may arise from a broader, temperamental syndrome (i.e., neuroticism). These commonalities can be traced to the underlying structure of emotional disorders that are defined by the two core dimensions of neuroticism and extraversion (Barlow, 2002). The higher-order dimension of neuroticism explained most of the covariance among various anxiety and mood disorders (e.g., depression, social anxiety disorder, panic disorder) (Barlow et al., 2013).

The triple-vulnerability theory formulated by Barlow and colleagues serves as a framework to explain the origins of neuroticism and emotional disorders. The two generalized vulnerabilities in the theory – (1) a biological vulnerability that is heritable and (2) a psychological vulnerability that stems from adverse early experiences and predisposes one to a sense of unpredictability and uncontrollability – are hypothesized to be risk factors for the development and expression of neuroticism. When subjected to stressful environmental contexts, people with these two vulnerabilities develop a generalized, anxious syndrome. The last element of the theory – a specific psychological vulnerability (which is mainly learned) – accounts for why a particular form of emotional disorder is manifested. Given that trait neuroticism underlies the roots of anxiety and mood

disorders, Barlow and colleagues proposed that psychological interventions and treatments directed at addressing this temperamental vulnerability may yield greater success than traditional treatments targeting the symptoms of specific disorders.

Emotion

As reflected by its conceptualization and definition, neuroticism should be closely related to emotions. Indeed, numerous correlational, experimental, and longitudinal studies have converged on the conclusion that neuroticism is positively linked to negative emotions (e.g., Diener and Lucas, 1999; Tellegen, 1985; Watson and Clark, 1984). This robust relation has also been found across various cultures, where neuroticism predicted unpleasant affect over a period of 10 years. The temperament perspective posits that the biological basis of neuroticism leads to predilections to experience negative affect. Emerging research has begun to pinpoint some of these neurobiological bases; for example, high levels of neuroticism are linked to excessive sensitivity of the amygdala and low serotonin levels. This temperament perspective is also supported by heritability studies, demonstrating that genetics account for about 31% of the variance in neuroticism (Pedersen et al., 1988).

Temperamental differences in emotions can result from differences in reactivity or baseline affect (or both). According to the reactivity model, individuals high in neuroticism react more strongly to negative stimuli – this differential reactivity to emotional stimuli accounts for neuroticism differences in emotional experiences (Gross et al., 1998). In support of the reactivity model, experimental studies found that neuroticism was associated with higher negative emotions in unpleasant situations. Recent functional imaging studies also confirm that neuroticism is associated with brain reactivity to negative emotional stimuli (Canli et al., 2001). Ecologically valid studies examining naturalistic affective responses have also replicated the experimental findings. Diary studies, for example, documented that participants higher in neuroticism were more reactive to stress and showed a stronger association between daily stress and daily negative affect (e.g., Mroczek and Almeida, 2004). Another model, the affect-level model, posits that differences in neuroticism lead to differences in tonic levels of negative affect; hence, those high in neuroticism feel more negative regardless of the circumstances or stimuli (Diener and Lucas, 1999; Gross et al., 1998). In support of both models, studies by Gross et al. (1998) found that neuroticism correlated positively with baseline negative affect and with increases in negative emotions in response to unpleasant stimuli.

Cognitive Processing

According to the cognitive perspective, individual differences in cognitive processes or cognitive processing of emotional information can also account for neuroticism differences in emotional experiences (Rusting, 1998). Cognitive processes refer to ways of encoding and categorizing information, expectancies about outcomes and self-efficacy, self-regulatory processes, or coping styles. Personality theories like the FFT and the cognitive-affective processing system (CAPS) that incorporate additional elements (e.g., dynamic processes) and do not focus solely on traits may shed some light on how neuroticism differences in cognitive processing arise. The FFT postulates that the five factors (i.e., basic tendencies) interact with external influences through dynamic processes (such as cognitive styles) to shape characteristic adaptations. In turn, some dynamic processes can be differentially affected by personality traits (McCrae and Costa, 2008). CAPS theory conceives personality as a system characterized by interrelated cognitive and affective units that interact dynamically. Individual differences in behavioral and affective responses to the external environment are due to individual differences not only in the accessibility of particular cognitive-affective units but also in the distinctive organization and chronic activation levels of these units. Dynamic reciprocal transactions occur not only between this personality system and the external environment but also within the system, among the social-cognitive and affective variables (Mischel and Shoda, 1995).

Extant evidence does support the view of a more comprehensive system, where personality traits interact reciprocally with other processes (e.g., cognitive styles, affect). Other than affect (as discussed in the earlier section), neuroticism also influences cognitive processing. Studies have shown that individuals higher in neuroticism demonstrate a cognitive processing bias toward negative stimuli. For instance, high-neuroticism individuals recalled more negative words, exhibited a negative attentional bias, and responded faster to targets located with negative words. Not only do they process and recall negative stimuli faster and better on various cognitive tasks, they also interpret ambiguous stimuli negatively (Rusting and Larsen, 1998). Individuals high in neuroticism also tended to appraise the environment negatively (Watson and Clark, 1984). The superior processing of trait-congruent emotional information may be one possible explanation for why high-neuroticism individuals experience stronger negative emotions.

Emotion Regulation and Coping

The close associations between neuroticism and emotional experience also extend to emotion regulation. Neuroticism predicts efforts to control emotions, and correlational findings have revealed neuroticism differences in positive and negative emotion regulation (e.g., Gross, 1999). Tendencies to diminish or eliminate one's negative emotions or turn them in a more positive direction correlated inversely with neuroticism. This means that those higher in neuroticism are less likely to improve their negative mood or cope with negative outcomes and their emotional impact. They are also less likely to savor positive affect. Usage of the emotion regulation strategy, reappraisal, which leads to lower negative emotions and more positive emotions, is also correlated negatively with neuroticism (Gross and John, 2003). The link with emotion regulation extends beyond a concurrent association, as neuroticism also predicts future emotion regulation. Prior neuroticism predicted lower use of strategies to repair negative emotions at a later age (Kokkonen and Pulkkinen, 2001). Hence, the evidence

consistently documents that neuroticism is associated with use of maladaptive emotion regulation strategies.

Coping processes, which are considered as a type of cognitive process that people undergo when dealing with stressful or difficult events, have also been increasingly linked to neuroticism. Neuroticism relates to negative dispositional coping styles and cognitive appraisals and is consistently found to be associated with the use of ineffective coping mechanisms. High-neuroticism individuals also tended to use less adaptive coping strategies to cope with the daily stressors than low-neuroticism individuals. The links between neuroticism and coping were further confirmed in a meta-analysis, which found that neuroticism predicted specific problematic coping strategies like withdrawal, wishful thinking, and negative emotion-focused coping (Connor-Smith and Flachsbart, 2007).

Neuroticism not only exerts direct effects on constructs, such as emotion, emotion regulation, and cognitive processes, but also plays an important moderating role. For instance, not only does neuroticism affect people's emotional reactivity to stressors and their choice of coping strategies, it also moderates the efficacy of those coping strategies (Bolger and Zuckerman, 1995). There is also evidence that neuroticism moderates the effectiveness of cognitive strategies on emotion regulation. Reappraising an unpleasant real-life event or a hypothetical situation decreased negative emotions among low-neuroticism, but not among high-neuroticism, individuals (Ng and Diener, 2009). Conversely, negatively reappraising a positive event to dampen it decreased positive emotions among high-neuroticism, but not low-neuroticism, individuals. Hence, cognitive strategies have differential effects on positive and negative emotions for people varying in neuroticism. High-neuroticism individuals are less proficient at negative mood repair, but yet are more susceptible to the detrimental effects of maladaptive strategies.

Life Outcomes

In summary, the links between neuroticism and the various constructs highlight that neuroticism would predict outcomes in important life domains. A review of the extensive literature on the relationship between personality and subjective well-being confirmed that neuroticism is linked to lower well-being (Steel et al., 2008). People higher in neuroticism are lower in happiness and life satisfaction and experience greater negative affect. Neuroticism is also inversely related to the overall quality of life. Furthermore, neuroticism is closely associated with and predictive of mental and physical health. Extensive evidence indicates that it strongly correlates with mood, anxiety, somatoform, eating disorders, and schizophrenia, and correlates moderately with personality disorders. It is also associated with physical health problems such as cardiovascular disease and even predicts longevity in the general population (Lahey, 2009). However, a recent longitudinal study suggests that high neuroticism, when accompanied by high conscientiousness, can still be associated with some health benefits. Specifically, people who were high in both neuroticism and conscientiousness had lower levels of interleukin-6; that is, they had lower health risk for chronic diseases such as heart disease (Turiano et al., 2013).

In the work domain, negative emotionality (or neuroticism) predicted future work outcomes: Those higher in negative emotionality at age of 18 years subsequently had lower levels of occupational attainment, less job satisfaction, and less financial security at age of 26 years (Roberts et al., 2003). In the social domain, neuroticism is a consistent personality predictor of relationship outcomes (e.g., relationship dissatisfaction, conflict, abuse) and relationship quality (Caspi et al., 2005). It is also inversely related to marital satisfaction (Lahey, 2009). Therefore, work from different areas converges to underscore the pivotal role of neuroticism in life and how it serves as a robust and important correlate and predictor of key outcomes. Overall, the evidence underlines that higher neuroticism is related to more adverse outcomes – being less happy, less healthy, and less successful, although it may be associated with some health benefits when combined with other personality traits (e.g., conscientiousness).

Change in Neuroticism across the Life Span

Longitudinal studies on personality have shown that personality traits (such as neuroticism) are not fixed and can change over time. Despite its strong genetic basis and the belief that genes do not change, neuroticism is not entirely determined by genes and can also be influenced by environmental factors and the interactions between genes and environment. For instance, if one's environment does not fit with one's existing genetic predisposition, one can try to change the environment (manipulative process) or try to change one's personality traits to achieve a more congruent person–environment fit. Hence, as environments change and interact with genes via bidirectional and dynamic processes, these experiences can affect neuroticism change. Although people change as they age, there is still overall consistency in personality. A meta-analysis of longitudinal studies demonstrated high rank-order consistency (i.e., the relative ordering of individuals within a group over time) for the Big Five personality traits, as all five traits were fairly stable (with consistency estimates ranging from 0.5 to 0.55), with neuroticism showing a stability coefficient estimate of 0.50 (Roberts and DelVecchio, 2000).

Indeed, personality continuity is exhibited from childhood to adulthood. Childhood temperamental qualities are closely linked to adult personality traits. Findings indicated that children who exhibited the undercontrolled temperament type at age of 3 years showed high levels of negative emotionality (or neuroticism) at ages of 18 and 26 years (Caspi et al., 2003). Longitudinal and cross-sectional studies of personality trait change confirm that personality traits continue to change in adulthood, and even into old age (Roberts and Mroczek, 2008). Cross-sectional findings that middle-aged individuals are lower in neuroticism than young adults are corroborated by longitudinal studies that showed that there is increased emotional stability with age (Roberts et al., 2006). This mean-level change in neuroticism (i.e., the increases/decreases over time in a sample/population of individuals) occurs throughout the life span, starting from adolescence,

with larger decreases in young adulthood (age of 20–40 years), though people continue to decrease in neuroticism in middle and old age (Roberts et al., 2006).

This decline in neuroticism has been replicated across cultures, suggesting that its developmental trend is universal. Indeed, there is strong evidence of universal personality maturation, where neuroticism showed age-related decreases across 62 nations (Bleidorn et al., 2013). Consistent with the FFT perspective that personality traits are biologically determined, there should be no cultural differences in personality developmental trends. Yet, there is also evidence of significant cultural differences in age effects on personality traits across the 62 cultures; specifically, cultures with earlier job-role transitions showed more pronounced age-related decreases in neuroticism. Hence, these findings also support the social investment theory, which proposes that cultural norms regarding the timing of life transitions (e.g., family-role and job-role transitions) influence personality maturation (Bleidorn et al., 2013). That is, investment in social roles contributes to personality development.

Why do personality traits change over time? Although personality is partly determined by genes, personality change can occur via mechanisms such as life and work experiences, changes in cultures and social environments, and changes in expectancies and goals. Work and life experiences are a major factor for why people change as they age. Roberts et al. (2003) found that work experiences were related to changes in personality traits. For example, occupational attainment, work satisfaction, and financial security were related to decreases in negative emotionality (as assessed by the MPQ). Similarly, increases in work satisfaction and relationship satisfaction were associated with decreases in neuroticism (Scollon and Diener, 2006). In recapitulation, personality traits do change due to life experiences and through interactions with one's environment and culture.

Conclusion

It is evident that neuroticism is not merely a descriptive personality trait for comparing people. Neuroticism is also a strong predictor of emotion, mental and physical health, work outcomes, and social relationships. Its close links to adverse outcomes in important life domains highlight that it is germane to understand the mechanisms underlying these links. Development of models that would explain the causal relations between neuroticism and these life outcomes, as well as examine potential interventions or treatments, would be important. Such interventions could help to improve or even prevent the undesirable outcomes associated with high levels of neuroticism. For example, in the area of mental disorders, treatments that target temperamental vulnerabilities could be developed, as they would address commonalities underlying various disorders and would be more effective than treatments targeting symptoms. Additionally, future treatment approaches could examine ways of reducing high levels of neuroticism. This is pertinent, given that there is potential for change in neuroticism although it exhibits fairly large rank-order stability, as substantial evidence has noted mean-level decreases in neuroticism across the life span.

See also: Anxiety and Anxiety Disorders; Emotional Regulation; Personality, Trait Models of; Personality: Historical and Conceptual Perspectives.

Bibliography

Barlow, D.H., Sauer-Zavala, S., Carl, J.R., Bullis, J.R., Ellard, K.K., 2013. The nature, diagnosis, and treatment of neuroticism: back to the future. Clinical Psychological Science 1, 1–22.

Barlow, D.H., 2002. Anxiety and its Disorders: The Nature and Treatment of Anxiety and Panic, second ed. Guilford Press, New York, NY.

Bleidorn, W., Klimstra, T.A., Dennisen, J.J.A., Rentfrow, P.J., Potter, J., Gosling, S.D., 2013. Personality maturation around the world: a cross-cultural examination of social-investment theory. Psychological Science 24, 2530–2540.

Bolger, N., Zuckerman, A., 1995. A framework for studying personality in the stress process. Journal of Personality and Social Psychology 69, 890–902.

Canli, T., Zhao, Z., Desmond, J.E., Kang, E., Gross, J., Gabrieli, J., 2001. An fMRI study of personality influences on brain reactivity to emotional stimuli. Behavioral Neuroscience 115, 33–42.

Caspi, A., Harrington, H., Milne, B., Amell, J.W., Theodore, R.E., Moffitt, T.E., 2003. Children's behavioral styles at age 3 are linked to their adult personality traits at age 26. Journal of Personality 71, 495–513.

Caspi, A., Roberts, B.W., Shiner, R.L., 2005. Personality development: stability and change. Annual Review of Psychology 56, 453–484.

Connor-Smith, J.K., Flachsbart, C., 2007. Relations between personality and coping: a meta-analysis. Journal of Personality and Social Psychology 93, 1080–1107.

Costa, P.T., McCrae, R.R., 1992. NEO PI-R Professional Manual. Psychological Assessment Resources, Odessa, FL.

Diener, E., Lucas, R.E., 1999. Personality and subjective well-being. In: Kahneman, D., Diener, E., Schwarz, N. (Eds.), Well-being: The Foundations of Hedonic Psychology. Russell Sage Foundation, New York, pp. 213–229.

Eysenck, H.J., 1990. Biological dimensions of personality. In: Pervin, L.A. (Ed.), Handbook of Personality: Theory and Research. Guilford Press, New York, pp. 244–276.

Goldberg, L.R., 1990. An alternative "description of personality": the Big-Five factor structure. Journal of Personality and Social Psychology 59, 1216–1229.

Gross, J.J., 1999. Emotion and emotion regulation. In: Pervin, L.A., John, O.P. (Eds.), Handbook of Personality: Theory and Research. Guilford Press, New York, pp. 525–552.

Gross, J.J., John, O.P., 2003. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. Journal of Personality and Social Psychology 85, 348–362.

Gross, J.J., Sutton, S.K., Ketelaar, T., 1998. Relations between affect and personality: support for the affect-level and affective-reactivity views. Personality and Social Psychology Bulletin 24, 279–288.

John, O.P., Srivastava, S., 1999. The big five trait taxonomy: history, measurement, and theoretical perspectives. In: Pervin, L.A., John, O.P. (Eds.), Handbook of Personality: Theory and Research. Guilford Press, New York, pp. 102–138.

Kokkonen, M., Pulkkinen, L., 2001. Extraversion and neuroticism as antecedents of emotion regulation and dysregulation in adulthood. European Journal of Personality 15, 407–424.

Lahey, B.B., 2009. Public health significance of neuroticism. American Psychologist 64, 241–256.

McCrae, R.R., Costa, P.T., 1997. Personality trait structure as a human universal. American Psychologist 52, 509–516.

McCrae, R.R., Costa, P.T., 2008. The five-factor theory of personality. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research. Guilford Press, New York, pp. 159–181.

McCrae, R.R., Terracciano, A., 79 Members of the Personality Profiles of Cultures Project, 2005. Personality profiles of cultures: aggregate personality traits. Journal of Personality and Social Psychology 89, 407–425.

Mischel, W., Shoda, Y., 1995. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. Psychological Review 102, 246–268.

Mroczek, D.K., Almeida, D.M., 2004. The effect of daily stress, personality, and age on daily negative affect. Journal of Personality 72, 355–378.

Ng, W., Diener, E., 2009. Feeling bad? The "power" of positive thinking may not apply to everyone. Journal of Research in Personality 43, 455–463.

Pedersen, N.L., Plomin, R., McClearn, G.E., Friberg, L., 1988. Neuroticism, extraversion, and related traits in adult twins reared apart and reared together. Journal of Personality and Social Psychology 55, 950–957.

Plomin, R., Caspi, A., 1999. Behavioral genetics and personality. In: Pervin, L.A., John, O.P. (Eds.), Handbook of Personality: Theory and Research. Guilford Press, New York, pp. 251–276.

Roberts, B.W., Caspi, A., Moffitt, T., 2003. Work experiences and personality development in young adulthood. Journal of Personality and Social Psychology 84, 582–593.

Roberts, B.W., DelVecchio, W.F., 2000. The rank-order consistency of personality from childhood to old age: a quantitative review of longitudinal studies. Psychological Bulletin 126, 3–25.

Roberts, B.W., Mroczek, D., 2008. Personality trait change in adulthood. Current Directions in Psychological Science 17, 31–35.

Roberts, B.W., Walton, K., Viechtbauer, W., 2006. Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. Psychological Bulletin 132, 1–25.

Rusting, C.L., 1998. Personality, mood, and cognitive processing of emotional information: three conceptual frameworks. Psychological Bulletin 124, 165–196.

Rusting, C.L., Larsen, R.L., 1998. Personality and cognitive processing of affective information. Personality and Social Psychology Bulletin 24, 200–213.

Scollon, C.N., Diener, E., 2006. Love, work, and changes in extraversion and neuroticism over time. Journal of Personality and Social Psychology 91, 1152–1165.

Steel, P., Schmidt, J., Schultz, J., 2008. Refining the relationship between personality and subjective well-being. Psychological Bulletin 134, 138–161.

Tellegen, A., 1985. Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In: Tuma, A.H., Maser, J.D. (Eds.), Anxiety and the Anxiety Disorders. Erlbaum, Hillsdale, NJ, pp. 681–706.

Turiano, N.A., Mroczek, D.K., Moynihan, J., Chapman, B.P., 2013. Big 5 personality traits and interleukin-6: evidence for "healthy neuroticism" in a US population sample. Brain, Behavior, Immunity 28, 83–89.

Watson, D., Clark, L.A., 1984. Negative affectivity: the disposition to experience aversive emotional states. Psychological Bulletin 96, 465–490.

Neurotransmitters

Phillip Grant, Justus-Liebig-University, Giessen, Hesse, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.R. Cooper, volume 15, pp. 10612–10619, © 2001, Elsevier Ltd.

Abstract

Neurotransmitters act as signaling molecules between neurons and other neuronal or nonneuronal cells. They relay information across chemical synapses, whereby their effect, *in concreto*, depends upon the respective receptor they bind to. All known neurotransmitters have several different types of receptors they are capable of binding and, as such, are functionally highly versatile. They, additionally, share many commonalities with other signaling agents. Important neurotransmitters in mammals are acetylcholine, monoamines (like serotonin, dopamine, adrenalin, and noradrenalin), glutamate, and gamma-amino butyric acid. Changes in neurotransmission constitute, i.e., the first step in synaptic plasticity, the neural foundation of learning and memory.

Neurotransmitters are a group of signaling molecules that are synthesized and released by neurons. Prior to the discovery of enlargements in the gaps between axon terminals and other cells (the synaptic cleft) by Santiago Ramón y Cajal, signal transduction in the nervous systems was believed to be primarily electrical via so-called gap junctions. The first neurotransmitter was discovered in 1921 by Otto Loewi, who named it *Vagusstoff* due to its properties to mimic the inhibitory effects of the vagus nerve on the heart of the frog. *Vagusstoff* has since internationally come to be known as acetylcholine.

Definition

Due to the works of Loewi and other ensuing research, a number of criteria were laid down in the 1960s to classify neurotransmitters and distinguish them from other signaling molecules; e.g., hormones, neuromodulators, or neuropeptides. These criteria were briefly as follows:

1. Precursors and/or synthesizing enzymes must be present in the presynaptic neuron.
2. The molecule must be present in the presynaptic neuron and be released upon specific stimulation.
3. The available quantity must be sufficient to exert influences on postsynaptic cells.
4. Postsynaptic cells must have receptors capable of binding the molecule.
5. A deactivating biochemical mechanism must be present.
6. Antagonists or blockers must exist to inhibit the molecule's effect on postsynaptic cells.

Due to ongoing research, the development of new methods to identify potential neurotransmitters and the discovery of novel neurotransmitters, the aforementioned criteria have since been found to be mainly underinclusive and have, therefore, lost much of their importance. The most often cited example hereof is the gaseous molecule nitric oxide (NO), which – alongside a variety of other functions – also acts as a neurotransmitter. Due to its ability to diffuse through cell membranes it is, however, not necessarily bound to a single synapse, but instead may also retrogradely influence a number of cells adjacent to the releasing neuron (e.g., astrocytes). It is also not necessarily enzymatically degraded due to its radical nature and brief half-life, or it may react with intracellular superoxide anions to form peroxynitrite, although a number of NO-degrading mechanisms exist within the cellular redox pathways. Finally, due to its primary physiological mechanism of action – i.e., the activation of soluble NO-sensitive guanylyl cyclase – it does not have a receptor in the classic sense.

This issue is further complicated by the fact that a number of molecules – e.g., the monoamines serotonin (5-HT), dopamine (DA), noradrenalin (NA), and adrenalin (A) – act both as neurotransmitters as well as hormones in discrete bodily systems of communication.

It could, therefore, be concluded that the aspect of defining neurotransmitters has taken a back seat to the investigation of the actions of individual signaling molecules.

Classes of Neurotransmitters

Neurotransmitters belong to several distinct (bio)chemical groups. They are generally small nonpeptide molecules that are not encoded by genes but rather (mostly enzymatically) synthesized from precursors taken up with nutrition.

Monoamine transmitters are derivatives of essential amino acids, whereby serotonin (5-hydroxy tryptamine, 5-HT) is synthesized from tryptophane through the enzymes tryptophane hydroxylase and aromatic L-amino-acid decarboxylase. The catecholamines dopamine (DA), noradrenalin (NA), and adrenalin (A) are synthesized from each other starting from the amino acid tyrosine, which is converted to DOPA (3,4-dihydroxy phenylalanine) through tyrosine hydroxylase. DOPA is then metabolized to dopamine through aromatic L-amino-acid decarboxylase (aka DOPA-decarboxylase) and is then taken up into presynaptic vesicles via the vesicular monoamine transporter 2 (VMAT2). In (nor)adrenergic cells or chromaffine cells of the adrenal medulla, these vesicles possess membrane-bound dopamine-β-hydroxylase (DBH) producing noradrenalin. In adrenergic and chromaffine cells, noradrenalin is released back into the cytoplasm via the VMAT1 and here converted into adrenalin through phenylethenolamine-N-methyltransferase.

All monoamines are degraded through monoamine oxidases (MAOs), which exist as two isozymes, MAO-A and MAO-B. Serotonin may additionally be metabolized to melatonin in the pineal gland. Unlike 5-HT, catecholamines are additionally degraded through catechol-*O*-methyltransferase. An example of DA synthesis and degradation is given in Figure 1.

Acetylcholine (ACh) is metabolized from choline and actyl-CoA through choline-acetyltransferase and degraded through acetylcholine esterase forming choline and acetate, whereby the latter may be combined with coenzyme A to form new acetyl-CoA.

Glutamate (Glut) and gamma-aminobutyric acid (GABA) are both derivatives of the amino acid glutamic acid. Glut is the dissociation product of glutamic acid and can either be degraded through glutamate dehydrogenase to α-ketoglutarate or decarboxylased through glutamate/glutamic acid decarboxylase to form GABA, which in turn is degraded through GABA transaminase and succinate semialdehyde dehydrogenase to form succinate. Glut can additionally be synthesized from the amino acid glutamine through glutaminase.

An important fact to mention is that the commonly found classification of inhibitory versus excitatory neurotransmitters is strictly speaking a misnomer, since the action of any given neurotransmitter is not an inherent property of the molecule but rather determined through the individual receptor it binds to (see Section Neurotransmitter Receptors). Even though, e.g., glutamate usually has excitatory and GABA inhibitory effects on other cells, receptors with vice versa effects exist for both transmitters.

Neurotransmitter Systems

While glutamate and GABA as well as other related transmitters like aspartate are produced by a multitude of cells within the entire central nervous system (CNS), phylogenetically older transmitters like acetylcholine and the monoamines are usually only produced by distinct cell populations. These cells are mainly located within the brain stem (with the important exception of the cholinergic motoneurons of the spinal cord and the telencephalic Ncl. basalis of Meynert, which produces acetylcholine, but is only found in this form in *Homo sapiens*) and possess long axons transporting the transmitters to various regions of the CNS. These axon tracts are referred to as neurotransmitter systems, whereby some have additional individual names.

Apart from the aforementioned cholinergic neurons of the spinal cord and the Ncl. basalis of Meynert, acetylcholine is mainly produced in a diffuse network of nuclei within the

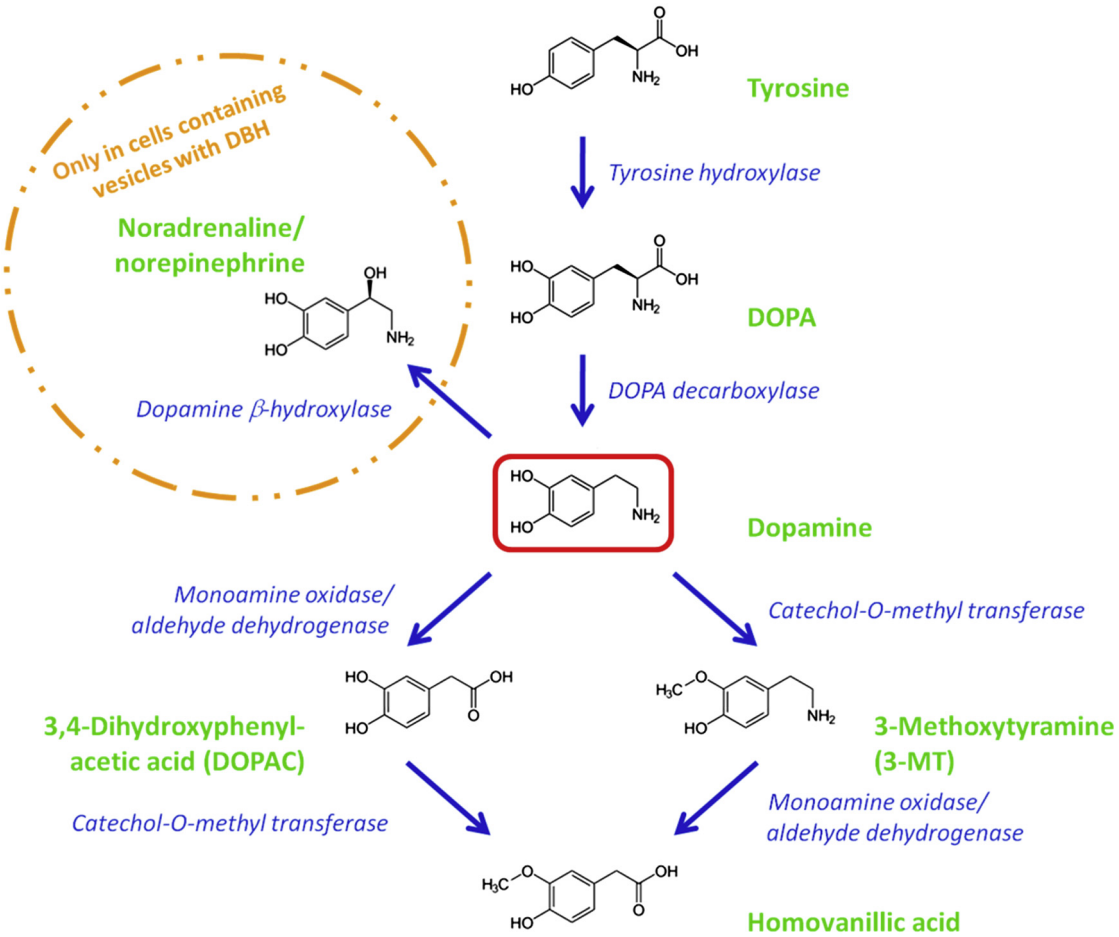

Figure 1 Enzymatic dopamine metabolism. Reproduced from Grant, P., 2011. Dopamine Neurotoxicity, Oxidative Stress and Schizophrenia – *In Vitro* and *In Vivo* Studies of Peroxisomal Reactions to Increased Dopamine. Giessen, University, Diss. URN: urn:nbn:de:hebis:26-opus-84954.

reticular formation of the brain stem. These neurons are believed to regulate arousal states both during sleep (induction of REM-phases) and waking hours, probably through projections to the Ncl. basalis of Meynert, and are therefore referred to as the ascending reticular activation system (ARAS). Many older sources link the ARAS to parts of the dorsal thalamus, called the unspecific thalamic nuclei or the truncothalamus, reporting a diffusely arousing function of these nuclei on the cerebral cortex. A vast number of publications of the early twenty-first century have, however, conclusively disproven this notion as well as the existence of the proposed unspecific thalamic nuclei in general.

The 5-HT-system originates from neurons of the midline group of the reticular formation, the raphe nuclei. The axons of the raphe nuclei project in descending (to the spinal cord), transverse (to the cerebellar cortex), and ascending fashion (to other parts of the brain stem, diencephalon, basal ganglia, and cerebral cortex). 5-HT mainly has a modulatory function on other discrete systems; e.g., ascending and descending pain-inhibition systems or modulation of cerebellar function through inhibition of Lugaro's neurons in the cerebellar cortex. According to Richard Depue, the cerebral 5-HT-projections modulate the stimulus-threshold of approach–avoidance reactions.

Noradrenalin and – to a lesser extent – adrenalin are mainly produced in the locus caeruleus located in the tegmentum of the pontomesensephalic border. The locus caeruleus projects, similar to the raphe nuclei, to a variety of regions within the brain stem, diencephalon, and forebrain. These projections have important functions regarding the regulation of central immunomodulatory processes, activation phases during sleep (induction of slow-wave sleep) and, according to the reinforcement sensitivity theory by Jeffery Gray, responses to potentially threatening external stimuli.

Dopaminergic projections, unlike those of other monoamines, form three discrete systems, whereof one originates in the hypothalamus and projects to the anterior lobe of the pituitary. Within this tubero-infundibular system, however, dopamine does not act as a neurotransmitter but as a hormone inhibiting the secretion of prolactin, and is therefore referred to as prolactin-inhibiting hormone. The two classical dopaminergic neurotransmitter systems both originate from the mesencephalon. The nigrostriatal system projects from the compact part of the substantia nigra (SNC) as well as the paranigral and retrorubral nuclei and projects via the median forebrain bundle to the dorsal striatum, where it modulates striatal gating. Due to the extremely high neurotoxicity potential of dopamine, the SNC cells are prone to neuronal death resulting in severe motor symptoms. The resulting clinical picture is called Parkinson's disease.

The second meso-telencephalic DA-system connects the ventral tegmental area with almost all components of the limbic system (e.g., the lateral hypothalamus, accumbens nucleus, hippocampus, amygdala, and the prefrontal and cingulate cortices) and is therefore referred to as the mesolimbic system. Due to the works of Olds and Milner (1954) and Crow (1972) on intracranial self-stimulation in rats, the mesolimbic system was for a long time believed to represent a reward and approach system. More recent findings have, however, shown that it is also involved in the processing of aversive stimuli and avoidance behavior (in the sense of negative reinforcement), wherefore it is today better referred to as conveying incentive salience of (mainly conditioned) stimuli. Dysregulation of this system is considered as one of the main factors in the ethiopathogenesis of psychosis (e.g., in schizophrenia).

Mechanisms of Neurotransmitter Action

Neurotransmitters usually act within chemical synapses by transducing information across the synaptic cleft. They are mostly taken up into vesicles within the synthesizing neuron by specific vesicular transporters and stored for action potential-dependent release. This uptake applies to both newly synthesized transmitters as well as to those that have been reuptaken from the synaptic cleft by so-called neurotransmitter transporters. This storage has been shown to be of great clinical relevance, since some transmitters – e.g., dopamine – have been shown to have neurotoxic effects, which are attenuated by higher pH values within the synaptic vesicles. The deleterious side effects of the drug methamphetamine (aka Chrystal Meth), e.g., are caused by the drug's inversion of the flow direction of both the vesicular transporter as well as the dopamine transporter, whereby methamphetamine acts as a potent dopamine releaser and rapidly induces dopamine neurotoxicity.

Upon an action potential reaching the presynaptic terminal, voltage-gated Ca^{2+} channels open, thereby allowing for the fusion of the synaptic vesicle with the presynaptic membrane and the exocytosis of a large amount of neurotransmitters into the synaptic cleft. The amount of transmitters stored within one vesicle is called a neurotransmitter quantum. Due to the continuous presence of Ca^{2+} ions in healthy cells, single quanta are released all the time, but do not suffice to exert significant influences on other cells. For this effect, the release of several quanta is necessary.

The released transmitters diffuse across the synaptic cleft and bind to receptors present on the membrane of postsynaptic cells. In most chemical synapses, neurotransmitters additionally bind to receptors within the membrane of the presynaptic neuron (autoreceptors), thereby contributing to the inhibition of synaptic transmission via a negative feedback mechanism. In other cases – e.g., within a majority of the synapses using the transmitter glutamate (Glut) – the negative feedback is achieved through more complicated mechanisms. Here, a specific postsynaptic receptor (the NMDA-receptor, named after its primary agonist *N*-methyl-D-aspartate) allows for the influx of Ca^{2+} ions into the postsynaptic cell, which in turn activates, i.a., the enzyme neuronal nitric oxide synthase (nNOS). As mentioned before, NO is then able to diffuse to neighboring astrocytes, where it activates enzymes that degrade important cofactors necessary for NMDA-receptor activation.

Apart from negative feedback and reuptake as well as similar uptake of neurotransmitters into astrocytes via neurotransmitter transporters, neurotransmitter action is also terminated through transmitter-degrading enzymes.

Finally, most synapses are complex, meaning that presynaptic neurons (regarding an individual synapse) are simultaneously the postsynaptic neuron of a second individual synapse and therefore possess receptors for the neurotransmitters used in the second synapse. From the point of view of the first synapse, these receptors are therefore called heteroceptors since they bind neurotransmitters other than the one

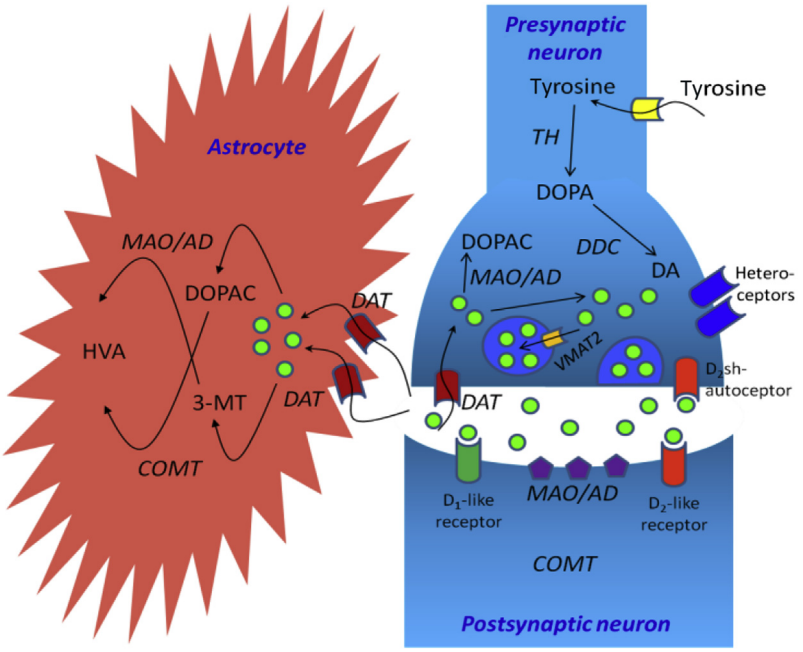

Figure 2 Dopaminergic synapse. From Grant, P., 2011. Dopamine Neurotoxicity, Oxidative Stress and Schizophrenia – *In Vitro* and *In Vivo* Studies of Peroxisomal Reactions to Increased Dopamine. Giessen, University, Diss. URN: urn:nbn:de:hebis:26-opus-84954.

produced by the presynaptic neuron. These heteroceptors may either facilitate or inhibit neurotransmission within the synapse, depending on their relative type (see Section Neurotransmitter Receptors).

An exemplary dopaminergic synapse is shown in Figure 2.

Additionally, a number of neurons – e.g., of peripheral axons belonging to the vegetative nervous system or those projecting from the hypothalamus to the posterior lobe of the pituitary – exhibit axonal tumescences, at which transmitters are released into the surrounding parenchyma. These synapses do not 'belong to' individual postsynaptic cells or correspond to the structure of classical chemical synapses and are therefore called *synapses en passant*.

A final important subtype of chemical synapses is the neuromuscular junction, a connection between axons of motoneurons of the spinal cord and skeletal muscle fibers. These synapses exclusively use the neurotransmitter acetylcholine and are, unlike other synapses, completely surrounded by Schwann cells (neurolemmocytes).

Neurotransmitter Receptors

Neurotransmitter receptors are transmembrane proteins consisting of various subunits. In general, they have two distinct features, namely the capacity of binding specific molecules (ligands) and the ligand-gated stabilization with a specific conformational state, which influences either the function or the structure of the receptor proteins. It has to be mentioned, however, that some ligands actually do not influence the receptor conformation, but only block the receptor's binding site for other ligands. These ligands are called competitive antagonists or blockers, while those that inhibit the actions of primary ligands by changing receptor conformation are noncompetitive antagonists/blockers. Ligands mimicking the effects of the primary ligand are referred to as receptor agonists.

Neurotransmitter receptors may be classified into four groups depending on their effect on the postsynaptic cell, which may be excitatory or inhibitory, and their biochemical mechanism of action, which may be ionotropic (i.e., the receptor is a ligand-gated ion channel, which will open upon binding of the primary ligand or an agonist) or metabotropic (i.e., the receptor influences an intracellular signaling cascade upon binding of the primary ligand or an agonist). The resulting four neurotransmitter receptor groups are ionotropic excitatory, ionotropic inhibitory, metabotropic excitatory, and metabotropic inhibitory.

Ionotropic receptors are made up of up to five subunits, which form a channel through the cell membrane (q.v., Figure 3) and open upon ligand binding. They are specific for certain ions, whereby Na^+- and Ca^{2+} channels have excitatory effects, due to a resulting cation influx and depolarization of the cell membrane. Inhibitory ionotropic receptors are K^+- or Cl^- channels, whereby the first group actively induces a hypopolarization of the cell membrane, whereas the latter changes the ion gradient over the cell membrane and, therefore, impedes upon the membrane's capability to depolarize.

Metabotropic receptors are bound to G-proteins (guanine nucleotide-binding proteins) on the intracellular side. Upon ligand binding, the conformational change leads to dissociation and thereby activation of the G-protein. The type of the G-protein thereby influences the actual mechanism of receptor

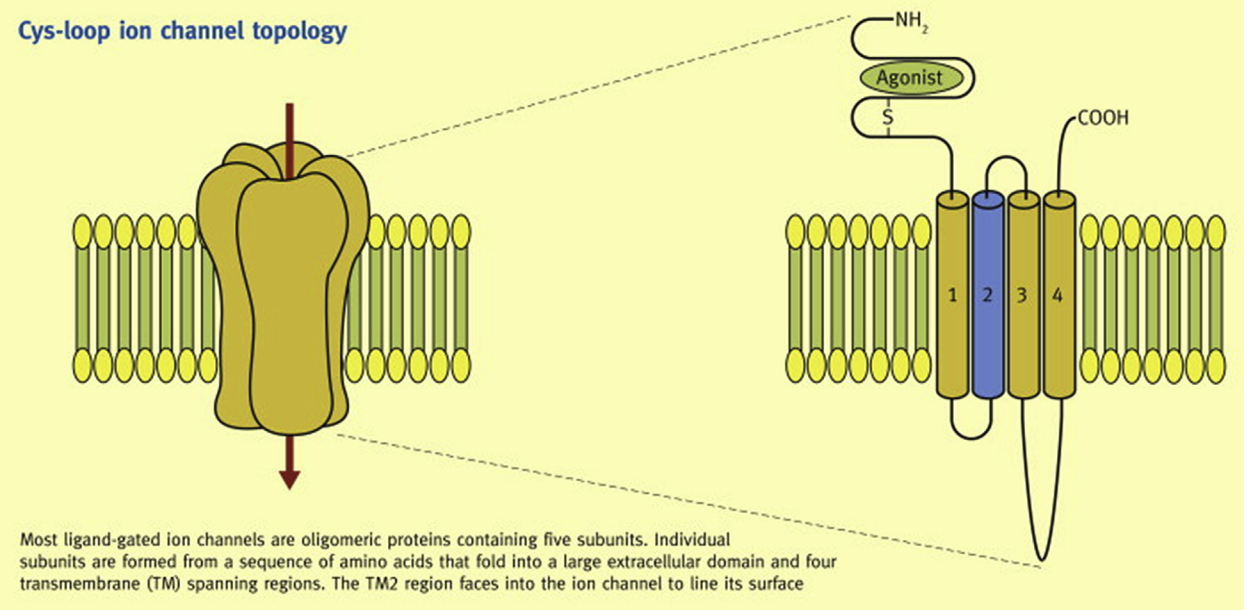

Figure 3 Schematic illustration of an ionotropic neurotransmitter receptor. From Cameron J. Weir. Ion channels, receptors, agonists and antagonists. Anaesthesia & Intensive Care Medicine 14 (9), 410–416. © Elsevier.

action, whereby G_i-proteins have inhibitory and G_s- and G_q-proteins excitatory effects on intracellular metabolism.

G-proteins influence the activity of enzymes, which in turn leads to a relative increase or decrease of so-called second messenger molecules. The most commonly found second messenger systems in the nervous system are the cyclic adenosine monophosphate (cAMP), the Ca^{2+}/Calmodulin (CaM), and the inositol triphosphate/diacyl glycerol (IP_3/DAG) systems.

cAMP is synthesized through adenyl cyclase (AC), whereby the G_s-protein activates and the G_i-protein inhibits AC-activity. An increase in intracellular cAMP concentration leads to the activation of a cAMP-dependent protein kinase (protein kinase A, PKA), which may either phosphorylate and thereby modify the function of ion channels within the cell membrane or alternatively phosphorylate the transcription factor CRE-binding protein (CREB) which in turn, when phosphorylated, dissociates into the nucleus and binds to the cAMP response element (CRE) within the promoter-regions of cAMP-dependent genes.

The IP_3/DAG-system is modulated through the G_q-protein, which increases the activity of the enzyme phospholipase C (PLC). PLC cleaves the membrane phospholipid phosphatidylinositol 4,5-bisphosphate (PIP_2) into IP_3 and DAG, whereby IP_3 binds and activates Ca^{2+} channels on the smooth endoplasmic reticulum, thereby increasing intracellular Ca^{2+} concentrations, and DAG activates the Ca^{2+}-dependent enzyme protein kinase C (PKC). PKC may, like PKA, phosphorylate ion channels, but also activates the transcription factors AP-1 (activator protein 1) and NF-κB (nuclear factor kappa-light-chain-enhancer of activated B cells; not only found in B cells).

Calmodulin (calcium-modulated protein) is activated by an increase in intracellular Ca^{2+} concentration, commonly either through Ca^{2+} influx via NMDA-receptors or as a secondary effect of the IP_3/DAG-system. The binding of Ca^{2+} to CaM (Ca^{2+}/Cam-complex) activates the enzyme Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII) as well as PKA. Both enzymes may phosphorylate ion channels and CREB, and CaMKII presumably also activates nNOS, thereby providing the aforementioned negative feedback for NMDA-receptor transmission.

All neurotransmitters bind to receptors specific to them. The relationship between transmitter and receptor is often compared to lock and key, whereby an analogy to a locking system is more appropriate, since first, all neurotransmitters bind to several specific receptors and, second, there are usually a number of ligands besides the neurotransmitter, which bind to specific receptors. As mentioned above, these other ligands are called receptor agonists or antagonists (see below).

The nomenclature of receptors is usually achieved through numbering or alphabetizing (e.g., dopamine receptors D_1 though D_5, α- and β-adrenoceptors, or $GABA_A$ and $GABA_B$ receptors), through numbering and alphabetizing (to classify subgroups; e.g., 5-HT_{2A}-receptor) or through their primary agonist or cofactor (e.g., NMDA-receptors or muscarinergic and nicotinergic ACh-receptors).

Receptor Agonists and Antagonists; Mechanisms of Drug Action

Apart from the primary ligand, receptors are also influenced by a number of other molecules. These are usually endogenous compounds, phytochemicals or artificial chemical compounds. This can be best illustrated using the model of ACh-receptors.

Two groups of ACh-receptors exist, named after their primary agonists: The muscarinic receptor binds the phytotoxins muscarin (agonist) – named after the fungus *Amanita muscaria*, from which it was first isolated – and atropin (antagonist) – derived from *Atropa belladonna*, commonly referred to as Deadly Nightshade. The second ACh-receptor group, the nicotinic

receptor, binds nicotine (agonist) – the alkaloid found in *Nicotiana* (tobacco) and other Nightshade-species – and curare (antagonist) – a group of alkaloid poisons extracted from South American vine-species like *Strychnos toxifera*.

A number of artificial chemical compounds exist, which act through activation or blockage of a specific neurotransmitter receptor, many of which are additionally used as recreational drugs. Common examples hereof are the $GABA_A$-agonist ethanol (drinking alcohol), the NMDA-receptor antagonists ketamine (Special K) or phencyclidine (PCP, Angel Dust), the first-generation antipsychotic haloperidol (blocks, i.e., the dopamine receptor D_2), β-blockers (antagonists of the β-adrenoceptors), the $5\text{-}HT_{2A}$-agonists mescaline and psilocybin (found in 'magic mushrooms') or lysergic acid diethylamide (LSD), which activates a variety of neurotransmitter receptors, including $5\text{-}HT_3$-, $5\text{-}HT_4$-, and β-adrenoceptors.

Apart from mimicking or preventing the actions of neurotransmitters through direct interaction with specific receptors, other drugs influence neurotransmission through either inhibition of reuptake into the presynaptic terminals or through actively releasing neurotransmission by reversing the flow directions of neurotransmitter transporters and vesicular transporters.

Commonly known examples for reuptake inhibitors are cocaine (dopamine reuptake inhibitor) or specific serotonin reuptake inhibitors. Other substances, like 3,4-methylene dioxy-*N*-methylamphetamine (MDMA, Ecstasy) or the aforementioned methylphenidate (Chrystal Meth) act as neurotransmitter releasers for (in this case) serotonin and dopamine, respectively.

Finally, drug influences on neurotransmission may act in a fashion not specific to any given neurotransmitter but rather on the more basic level of inhibition of action potentials (like *Tetraodontotoxin*; e.g., from puffer fish) or exocytosis of synaptic vesicles (like *Botulinum toxin*, *Botox*).

See also: Brain, Evolution of; Epigenetics of Brain Plasticity and Behavior; Hypothalamus; Ion Channels and Molecular Events in Neuronal Activity; Long-Term Potentiation and Long-Term Depression; Neural Plasticity; Neurons and Dendrites, Integration of Information in; Personality, Biological Models of; Personality, Trait Models of; Schizophrenia: Metabolic Consequences; Sleep: Neural Systems; Synaptic Mechanisms and Protein Synthesis in Memory.

Bibliography

Crow, T.J., 1972. A map of the rat mesencephalon for electrical self-stimulation. Brain Research 36 (2), 265–273.

Grant, P., 2011. Dopamine Neurotoxicity, Oxidative Stress and Schizophrenia – *In Vitro* and *In Vivo* Studies of Peroxisomal Reactions to Increased Dopamine. Giessen, University, Diss. URN: urn:nbn:de:hebis:26-opus-84954.

Kolb, B., Whishaw, I.Q., 2010. An Introduction to Brain and Behaviour, International ed. Worth Publishers.

Nieuwenhuys, R., Voogd, J., Van Huijzen, C., 2008. The Human Central Nervous System, fourth ed. Steinkopff.

Old, J., Milner, P., 1954. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. Journal of Comparative and Physiological Psychology 47 (6), 640–653.

Neutrality, Axiological

Nicholas Rescher, University of Pittsburgh, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Fillieule, volume 15, pp. 10619–10621, © 2001, Elsevier Ltd.

Abstract

There is no way to exclude axiological considerations from scientific inquiry. For one thing, a surgical fact/value separation is infeasible on theoretical grounds. And on practical grounds, science cannot operate without a recourse to cognitive values, in particular those that distinguish the trivial from the significant. The most and best that can be done is to adopt a position of axiological neutrality that minimizes any resort to tendentious and potentially contentious evaluations in science.

Is/Ought Separation and Value Exclusion from Science

The thesis that the sciences – and the social sciences in particular – should be value free has been astir since the early twentieth century, particularly under the influence of Max Weber (see, for example, Weber, 1917). The basic idea is that scientific claims should be formulated and validated impersonally, without involving the potentially variable value orientations of different individuals. Science should deal in observable facts, with issues of value – of deservability, preferability, normativity, objection, and the like – left to individuals to resolve independently on their own. Even as physicists look to numerically accessible phenomenon ('repeatability of experiments'), so the social sciences shall look to phenomena that are universally accessible, independently of potentially idiosyncratic orientations of particular individuals or groups. To be sure, social sciences can investigate the way in which people commit themselves to values, but any and all endorsement or espousal of values is to be proscribed in this domain.

The fact/value distinction was generally drawn on essentially terminological grounds. The idea is that fact statements simply maintain what is available to observation and inference therefrom. They report on reality's descriptive makeup. Value statements, by contrast, are not factual but axiological. They have a normative and axiological dimension owing to an involvement of some evaluative terms that give expression to placement along a positive/negative, favorable/unfavorable divide. So that cats chase mice is a factual claim. That it is a good (or bad) thing that cats chase mice is evaluative. The former may be germane to science, the latter must be excluded. But unfortunately, matters are not always so clear.

It is clear that any program of value exclusion from science must be predicated from the very outset on two indispensable presuppositions: (1) that it is possible to draw a clear line of distinction between statements of fact and judgment of value and (2) that inference across the boundary between these – and in particular the demonstration of judgment of value from statements of fact – is inherently impossible. But both these contentions are very questionable.

Why Fact/Value Separation Is Untenable: I. Logical Grounds

Considerations of general principle might seem to indicate that evaluative conclusions cannot be derived for purely factual premises. For by their very nature as such, evaluative conclusions must contain some sort of value terminology, and if the premises contain no such terms whatever, then it seems plausible to suppose that there is no ground in which a validly deducible conclusion can obtain them. But what seems plausible can turn out to be false – and the present case affords an instance of this.

However, a significant stumbling block for all theories of fact/value separation is the seemingly paradoxical circumstance that (1) $F \vee V$ follows deductively from the factual premise F (and thus is itself presumably factual), while nevertheless (2) $F \vee V$ in conjunction with a factual premise (viz. not-F) yields the evaluative conclusion V (and thus is itself presumably evaluative). (This difficulty was first noted by George Mavrodes. See his essay "On Deriving the Normative from the Non-Normative" in the *Papers of the Michigan Academy of Arts and Sciences*, vol. 53 (1968), pp. 353–365.)

The only plausible response here is to abandon the idea that the logical consequence of factual statements must always be factual, and proceed to class $F \vee V$ as hybrid rather than strictly or purely factual. This done, one would continue to hold to the principle that evaluative statements never follow from purely factual ones. But on this basis, one would then have to shift from a two-sided fact/value dichotomy to a tripartite division of statements as factual, evaluative, and hybrid, and this muddies the water. For the categories of fact and value are now seen as separated not by a sharp boundary line but by a broad corridor.

Why Fact/Value Separation Is Untenable: II. Practical Grounds

Consider the following inference:

All automobiles are man-made.
Therefore: All good automobiles are man-made.

The premise of this patently valid inference is clearly factual, and its conclusion is clearly evaluative. (Observe that the

statement "Dogs chase nice cats" combined with the factual thesis "Fido is a cat" and "No dog chases Fido" entails the clearly evaluative conclusion "Fido is not a nice cat"). Does such an example not refute our thesis that factual premise does not yield evaluative conclusions?

Not quite. For closer analysis shows that what is actually at issue in the preceding inference is an enthymematic syllogism that incorporates an unstated premise:

All automobiles are man-made.
(All good automobiles are automobiles.)
All good automobiles are man-made.

And here that enthymematic minor premise can (and should) be regarded as unproblematically acceptable on grounds of its effective triviality.

Again consider the following inference:

- *X* jabbed *Y* with a needle to cause *Y* pain because *X* did not like the way *Y* looked.
- *X* had no other reason for jabbing *Y* with a needle apart from wanting to cause *Y* pain because *X* did not like the way *Y* looked.

Therefore … *X* did something wrong/wicked.

Three things are clear here: (1) the premises all state matters of fact; (2) the premises entail the conclusion; and (3) the conclusion is evaluative. So it is clear that in this as in myriad other cases, factual premises can yield an evaluative conclusion.

To be sure, a stickler might object that this inference is not strictly speaking valid though also being enthymematic. And this is true. One needs to acknowledge that the reasoning at issue is enthymematic, and requires the addition of a further premise, viz.,

(P) When a person jabs another to cause that person pain simply and solely because he or she does not like his or her looks, then this person does something wrong/wicked.

But it is clear that this further premise is trivial, self-evident, and virtually tautological given the terms of reference in which it is formulated.

In cases of this sort, where the gap between factual premises and an evaluative conclusion is so narrow that it can be filled by an effectively trivial evaluative thesis, one has a situation where the fact/value gap is effectively bridgeable.

The price of a consistent theory involves taking the not implausible line that certain evaluative truisms (such as 'All nice cats are cats') cannot but be seen as unproblematically acceptable.

And of course there will be innumerable analogous cases, for example,

- *X* had food at his disposal.
- *X* realized that *Y* was very hungry and could not otherwise obtain food.
- *X* shared his food with *Y*.

Therefore: *X* did something morally credible.

Here again, one has an effectively trivial enthymematic bridge from the factual premises to the evaluative conclusion.

When factual premises yield evaluative conclusions in the face of premises that themselves are evaluatively trivial and self-evident, one not only can but also must cross the bridge over the fact/value divide. And here it does not matter whether one is a scientist speaking ex officio or a layperson enunciating common sense.

Viewed in this light, the idea that the conscientious social scientists cannot apply evaluative labels to certain deliberations and arguments – characterizing them as unsustainable, unworkable, unduly cumbersome, or counterproductive – is clearly problematic. The economist or sociologist no less than the historian or the physician can ex officio decry various matters in negative terms without being unprofessional about it. But of course he or she can only do this unproblematically when the issue is 'plain as a pikestaff' there for anyone to see – a matter of objective conditions rather than potentially idiosyncratic personal attitudes.

The fundamental contrast, of course, should be that between what can be observed and what cannot. The value of the item is not more perceptible than is the ownership of a piece of property. But that of course does not make it something mysterious and 'nonnatural' – the special object of a peculiar detection faculty, a value intuition. Value is not sense-perceptible but mind-judgmental: something to be determined not simply by observation of some sort but by reflective thought duly sustained by background information and suitably equipped with an awareness of principles.

What Is Left of Value Freedom

It is sometimes maintained that evaluative statements cannot figure in science because they are, by their very nature, neither true nor false. But this is simply nonsense. That it is foolish never to seek the advice of a knowledgeable person, and it is ill-advised to prefer a product one deems inferior to an equally available alternative that one sees as superior. And yet such evaluative claims are clearly every bit as true as the claim that grass is generally green. Such statements obviously state facts – only evaluative facts. (Yet another nail in the coffin of an insuperable fact/value divide.) One must avoid the confusion of values and tastes. "There's no disputing about tastes" may be true, but "There's no disputing about values" certainly is not. Values too can be altogether objective, in that value claims admit of rational substantiations – through impersonally cogent considerations.

Some value exclusionists hinge their position on an is/ought divide. Science, they say, informs us about what is, leaving matters of what ought to be to others. But this does science an injustice. For science tells us not only about what is, but also about what must be if – i.e., if certain conditions are to be brought to realization. And when one inquires into the conditions that must be met to make human life (be it at the individual of scientific level) possible – or indeed even pleasant – it is to science that one will have to look.

The biochemist who discovers that a particular production process poses a clear and present environmental risk is nowise stepping outside the range of scientific propriety in disrecommending it. And the same holds for the physician who detects available risks in a certain procedure. In expressing evaluative conclusions, such people are clearly acting within the bounds of scientific propriety rather than outside it.

The preceding deliberations indicate that an impassable fact/value barrier excluding values from science must be abandoned and a different basis provided for understanding the relationship between sciences and values. This revised basis will have to lie in the idea of universality of access. For science is and should be universally available, objective and impersonal in its findings and contentions. There is no place in the rational scheme of things for an Aryan physics or a Soviet biology. And so what is in order here is a transition from the (untenable) idea of a value-free science, one in whose operations a recourse to ideologically committed and potentially debatable and divisive values has no proper place. Any value commitment – any ideologically based perspectives or doctrines – has no place in science. It is not that values are to be excluded from science, but rather that whatever values are invoked are patently cogent, and nowise debatable or divisive. The oft-claimed axiological neutrality of science is not to be grounded in its Weberian Wertfreiheit but rather in its Kantian universality.

The conclusion that follows here is thus not that science is value free – that scientists acting ex officio can and should make no evaluative claims at all. It is, rather, that such claims can unhesitatingly be made whenever they happen to be uncontroversial and uncontestable.

So what is left of the Wertfreiheit program? What remains is the ideal of value neutrality: the idea that scientists should not qua scientists proclaim potentially controvertible evaluations – that their value claims should be either substantially trivial or at least such that reasonable people cannot reasonably avoid their acceptance.

Inquiry into the conditions that make individual human life possible (and even pleasant) is clearly part of the mandate of the life sciences. And inquiry into the conditions that make collective human life in society possible and congenial is part of the mission of the social sciences. Valuation is unavoidable in such contexts, and the exclusion of valuation in such matters makes no sense here. (Evaluation plays a legitimate and constructive role even in so theoretical a discipline as pure mathematics. Here *reductio ad absurdum* serves as a standard proof technique and is predicated in the thesis that absurdity is the focus of logically demonstrable self-contradiction in something that is obviously inappropriate and unacceptable.)

In the final analysis, science cannot possibly be value free. For the assessment of importance/significance/informativeness will unavoidably have to play a pivotal role in the formulation, promulgation, and publication of scientific work. If trivial details could not be distinguished from important facts – every one factual statement on a given topic was every bit as significant as any other – no meaningful discussion could ever be managed in science nor any meaningful instruction provided.

What It Means for the Working Scientists

As these deliberations indicate, it is deeply problematic to maintain that the scientist qua scientist must refrain from drawing such normative conclusions that follow from the facts at his or her disposal in an effectively unproblematic way.

It should, however, be granted that there is one important respect in which the advocates of scientific value neutrality have it right. For what is correct is not that social sciences are value free but that any policy recommendation about social arrangements should be clear and explicit about the values that speak for its acceptance. No hidden agenda, concealed interest, correct aim, etc.

The idea is not to keep value out, but make their role overt and explicit insofar as they enter in.

For while some self-evidently incontrovertible evaluative theses have to be seen as available anytime to anyone – scientist included – nevertheless contestable and potentially idiosyncratic evaluative claims are something else again.

So in the end, there are two distinct albeit related positions. One is based on the problematic but essentially factual idea of value exclusion from sciences predicated on the thesis that a Chinese wall of sorts exists to separate factual from evaluative claims. The other is based on the essentially normative idea that a scientific exposition should be candid and explicit about any operative value commitments in the interests enabling the neutrality of letting people form their own conclusions and averting a situation where advocacy wears the sheep's clothing of informativeness. Patently contestable value endorsements have no place in science. In this regard, science should be value neutral. There is – to reemphasize – no place for a Soviet biology or a Germanic physics.

The present article has argued that the first of these two positions is untenable and should be rejected but that the second is very much in order. On this perspective one cannot invoke an absolute fact/value divide to exclude evaluative claims from bearing any and all place in scientific deliberations. Rather one has to retreat to the position that what is questionable in science is significantly substantive and thereby potentially divisive evaluative claims.

Two decidedly different positions with respect to value freedom are at issue here. One is the problematic idea of value exclusion pivoting on that the sciences are altogether factual and have no evaluative or normative involvements. The other is the sensible and appropriate doctrine of value neutrality, viz., that the sciences properly should take sides on tendencies substantively controversial evaluative positions.

And this situation has important implications. For in many instances, the idea that a certain interaction is uncontestable and noncontroversial in itself is something that can be called into question. And this means that whenever there is even a remote chance of this, a scientist must be very clear and explicit about it so that his or her audience can judge matters for itself as regards the acceptability of the value claims at work. Just as the legislator should 'declare an interest' in an issue mode discussion, so the scientist should 'declare a normative commitment' in matters of deliberation.

The fallback from value exclusion to value neutrality is an important and well-advised step. For as Leo Strauss has rightly argued against the Weberian proscription of valuation in science, if there were no evaluative facts, then all evaluations would become arbitrary and unsustainable (see Strauss, 1953). And then if one scientific statement is no more important than another, one scientific claim no more reasonable than another, one scientific recommendation no more acceptable than another, and the ability and value of science itself

will be compromised beyond reason. If one claim is no more rationally appropriate than another, then where does that leave science itself? Valuation in science and valuation by science are essential aspects of overall functioning of our scientific endeavors.

In the final analysis, there is something rather self-inconsistent about the advocacy of value exclusion from science. For, of course, it is only possible justification of this contention would issue from a contention along such social lines as "Staking evaluative claims in the social sciences is inappropriate because doing so does disservice to the prospects of realizing the proper aims of the enterprise – viz. achieving a deepened explanatory understanding of social phenomena." But the fusion of factual and evaluative considerations in this claim itself is unmistakable the very surface of it.

See also: Interests in Sociological Analysis; Knowledge, Sociology of; Phenomenology in Sociology; Positivism, Sociological; Values, Sociology of.

Bibliography

Adorno, Theodor W., et al., 1975. der Positivismusstreit in Der Deutschen Soziologie. Luchterhand, Darmstadt. See especially Chapters 1 and 2.

Ayer, Alfred J., 1936. Language Trust and Logic. Gollancz, London. See especially Chapter 6.

Boudon, Raymond, 1995. Le juste et le vrai: études sur l'objectivité des valeurs et de la connaissance. Fayard, Paris. See especially Chapter 4.

Rokeach, Milton, 1973. The Nature of Human Values. Free Press, New York. See especially Chapters 1 and 2.

Strauss, Leo, 1953. Natural Right and History. University of Chicago Press, Chicago.

Weber, Max, 1917. Der Sinn der Wertfreiheit. Logos VII.

New Genetics and Race

Yulia Egorova, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores the way notions of race and ancestral origins are constructed in recent research in population genetics. The author focuses on studies used in health care, in the criminal justice system, and in genetic history. The article demonstrates how contemporary genetic research employs preexisting categories pertaining to race and descent, and discusses what implications it has had for processes of identity construction.

What is the relationship between genetic research and constructions of race? What contribution have geneticists made to the study of 'human genetic diversity'? What assumptions about the way humanity can be divided into distinct populations inform theoretical and methodological thinking, which in turn underpins research in the life sciences? Finally, in what ways has genetic research been used by the communities that it has researched, and by other stakeholders outside of the world of science? The author explores these issues by looking at literature within the social sciences and humanities addressing the sociocultural and political implications of genetic research. For the purpose of this analysis, the author divides the discussion of this field of literature along two main thematic lines. First, the author discusses studies that have focused on the role that genetics plays in categorizing humans into distinct populations either for medical purposes or for use in the criminal justice system. Second, the author explores genetic research initiated for the purposes of reconstructing the history of human migrations. In each case, the author focuses both on the genetic studies themselves and on their reception by the tested communities. In the end, the author discusses the way population genetics has been used to advance what has been described as 'genomic sovereignty' in countries and communities located outside Europe and North America.

Genetics and Race Revisited

In a seminal paper published in 1972, Richard Lewontin demonstrated that the major part of human genetic variation occurred among individuals within local geographic populations, which suggested that the concept of race was an invalid taxonomic construction (1972). The same idea was promoted by Luigi Luca Cavalli-Sforza, the initiator of the Human Genome Diversity Project (HGDP), which was supposed to map gene distributions of different populations around the world. The HGDP was launched almost in opposition to the Human Genome Project. Its initiators argued that the Human Genome Project did not take into consideration the diversity of human populations all over the world. They endeavored to fill this gap by collecting, preserving, and analyzing the DNA from about 500 populations around the globe that had been geographically isolated and had a distinct culture and language, as these would yield more information than genetically mixed urban ones. The collected materials were planned to be made available to scientists and scholars worldwide for research on the history of human migrations and biology (M'charek, 2005: pp. 6–7). The project initially received positive coverage in journals such as *Science* (e.g., Kahn, 1994: pp. 720–723) and was endorsed by the Human Genome Organisation. However, the HGDP soon ran into trouble, as its initiators were accused of exploitation and even racism (Haraway, 1997: pp. 248–253; M'charek, 2005: pp. 5–16; Marks, 2003: pp. 198–219). The accusation may seem surprising, given that one of the conclusions drawn by Cavalli-Sforza was that the notion of 'race,' which figured prominently in the scientific, historical, political, and popular discourses of the previous two centuries, had no genetic basis and that genetic differences that existed between any two individuals were vastly more significant than those between any two populations (Cavalli-Sforza, 2000: viii). Although such conclusions appear to be a powerful weapon in the struggle against racism, it has nevertheless been suggested that the HGDP and similar studies construct biological differences between various human populations: Cavalli-Sforza's endeavor may have thus helped to debunk the idea of dividing humanity into a limited number of major races, but it reinforced the notion that individuals could be divided into discrete populations on the basis of their genetic makeup. (For a critical discussion of the HGDP see M'charek, 2005; Reardon, 2005.)

It has been argued by commentators in the social sciences that this type of research in population genetics is permeated by 'racialist' thinking. (For a discussion of the distinction between racism and 'racialist' (or race) thinking, see (Skinner, 2006: p. 483). Indeed, some scientists have suggested that race is a useful operational category in genetic research and is of medical relevance, while the evolutionary biologist Armand Leroi (2005: p. A23) went as far as to declare that 'races clearly exist' and that 'the consensus about social constructs is unraveling.' For a very useful analysis see Hartigan (2008, 2010). For a discussion of internal debates among scientists about the usefulness of the term 'race' see Fausto-Sterling (2004).) Moreover, as Veronika Lipphardt has argued, drawing on Nancy Stepan (2003), any attempt to produce 'scientific facts' about human biological diversity is grounded in political agendas and social tensions (2012: S69). The way the categories of race and ethnicity are institutionalized and naturalized in population genetics has been discussed in a number of ethnographic studies of scientific practice (Reardon, 2005; M'charek, 2005; Fullwiley, 2007a,b). Duana

Fullwiley (2007a), for instance, has demonstrated in her analysis of two laboratories of pharmacogenetics that even though researchers themselves may struggle to define race and even have negative personal experiences of having to 'fit' their own identity into the predetermined categories of the US 'race and ethnicity' questionnaires, they still comply with the established practice of defining and comparing US groups in their work. It has therefore been demonstrated that the process of looking for genetic differences between populations is mediated from the outset by preexisting historical and sociocultural ideas about collectives of individuals constituting separate populations that are supposed to differ from each other on the level of biology. As sociologist of science, Troy Duster notes:

> It is possible to make arbitrary groupings of populations (geographic, linguistic, self-identified by faith, identified by others by physiognomy, etc.) and still find statistically significant allelic variations between those groupings. For example, we could examine all people in Chicago, and all those in Los Angeles, and find statistically significant differences in allele frequency at *some* loci (2003: 265).

Race, Genetics, Crime, and Medicine

What are the practical implications of population genetic research? In what ways has it been used 'on the ground'? First, it appears to be increasingly deployed in the name of improving health care provision in general and for 'minority' groups in particular. Some studies have aimed to include underrepresented communities in clinical trials to ensure that medical research does not rely entirely on people of European descent, or otherwise run the risk of jeopardizing 'minority' communities' chances to receive adequate pharmaceutical care. Social scientists have discussed the advantages and disadvantages of this approach pointing both to its practical, medical, and broader, ideational implications. Steven Epstein has suggested that this is "a victory worth savoring in a long struggle to bring medical attention to the excluded and underserved" (2007), but has also cautioned that this approach could reinforce racial categorization.

Other projects have attempted to seek a genetic basis for diseases and conditions with higher prevalence among particular populations. For instance, there has been a steady interest in connecting heart disease and hypertension in African-Americans to the genetics of 'race,' despite existing evidence that these complex conditions are associated with poverty and stress, among other factors, which could also explain why they are common in a community suffering from structural inequalities (Duster, 2003, 2005; Kahn, 2004). More specifically, in 2005, the Food and Drug Administration in the United States approved BiDil as a drug to treat heart failure specifically in African-Americans. As Jonathan Kahn has demonstrated, this was 'part of a much larger dynamic of reification in which the purported 'reality of race' as genetic may be used to obscure the social reality of 'racism,' including the health inequalities that map onto racial discrimination (Kahn, 2011: pp. 139–140).

Some genetic research has focused on groups that are at higher risk of recessive diseases, particularly Tay-Sachs disease, sickle-cell anemia, and cystic fibrosis (for a historical discussion see Wailoo and Pemberton, 2006). Such studies have undoubtedly contributed to a better understanding of these conditions and in some cases were even used by the 'affected' communities to improve health care provision for their members. For instance, an Orthodox Jewish group in the United States made use of research on Tay-Sachs disease to organize premarital genetic screening for their young men and women (Wailoo and Pemberton, 2006: pp. 17–20). However, it was pointed out that the danger of this type of research is that it opens the way for stigmatization of groups where a recessive condition is found at higher frequency. As Wailoo and Pemberton observe, the emergence of these conditions in the public consciousness in the United States transformed them into "vehicles for thinking about race and ethnicity" (Wailoo and Pemberton, 2006: p. 10). Not surprisingly, some communities protested against being associated with a genetic disorder. For instance, Ashkenazi Jewish women expressed concern about being identified as 'at risk' populations for breast cancer (Duster, 2003). Many African-Americans were wary of genetic screening and counseling for sickle-cell anemia, which for them had associations with the disturbing history of racial discrimination in the US health care (Wailoo and Pemberton, 2006).

Concern about racial discrimination has also been raised with respect to the use of DNA in the criminal justice system. It has been convincingly demonstrated that DNA databases used in the criminal justice systems in different countries around the Western world contain a disproportionate amount of DNA from people coming from 'minority communities' (Washington, 2010). Helen Wallace observes that in Britain, home to the largest DNA database in the world in terms of the proportion of the DNA samples of the population that it contains, "[a]pproximately 27 percent of the entire black population, 42 percent of the male black population, 77 percent of young black men, and 9 percent of all Asians have records on the National DNA Database, compared with just 6 percent of the white population" (2011: 76). Moreover, it has been reported that "an estimated 55 percent of those from black and minority ethnic communities who have their details on the database have never been charged or convicted of any offense," she notes (Wallace, 2011: p. 76). Similarly, Michael Risher has discussed how in California the retention of arrestee's DNA has led to the magnification of the racial disparities that had already been present in the criminal justice system of the state (2011: p. 47).

Genetic History

Studies in population genetics that concern themselves not with health care but with reconstructing the history of human migrations have also attracted a considerable amount of critical attention in social sciences. Such studies, which are sometimes referred to as genetic anthropology, anthropological genetics, or genetic history, constitute an important fraction of the growing body of population genetics. (For an excellent historical discussion of the emergence and early development of the field, see Sommer, 2008.)

This type of science is oftentimes described as 'genetic carbon dating,' as it is supposed to facilitate historical research, which otherwise would allegedly lack adequate tools for deciphering human past. As geneticists Mary-Claire King and Arno Motulsky put it, "the DNA of modern humans contains a record of the travels and encounters of our ancestors" and furthermore that

> The genotypes of people living today are the result of ancient human migrations, selection by climate and infection for genetic alleles that conferred a survival advantage, and mating patterns determined by cultural norms. By sampling genotypes from people across the globe, geneticists have reconstructed the major features of our history: our ancient African origin, migrations out of Africa, movements and settlements throughout Eurasia and Oceania, and peopling of the Americas. (quoted in Davis, 2004: p. 41)

To name just some such studies that specifically aimed to shed light on the processes of the formation of particular communities and thus to contribute to historical debates, one could refer to research on the ancestry of African-American and Native American groups and on different Jewish communities around the world.

One of the projects on African-American ancestry was launched by Rick Kittles then based at Howard University (Sabeti, 2002). The project was supposed to allow African-Americans to trace their ancestry to a particular region in Africa. Kittles argued that it could help this community to find out where exactly their ancestors were from. In addition, it was argued that such tests might turn out to be a very useful historical tool, as research on African-American roots was often limited due to lack of documentation. However, Kittles admitted that in some cases these tests could have a disruptive potential if their results contradicted family lore (see Sabeti, 2002). Other commentators have observed that this type of research has a danger of dividing African-Americans and distracting them from campaigns to improve the social and economic conditions of their community (Dula et al., 2003).

Similarly, there have been a substantial number of genetic studies addressing the origin of various Jewish groups. Perhaps the most widely publicized of these were the studies carried out on the *Cohanim* (Jewish priests) and on the origins of the Lemba Judaising group of southern Africa (Spurdle and Jenkins, 1996; Skorecki et al., 1997; Thomas et al., 1998, 2000).

The latter represent an example of a group whose communal narrative has been strengthened by genetic findings. (For a detailed discussion see Parfitt, 2002, papers by Azoulay, Johnston, Parfitt, and Zoloth in *Developing World Bioethics*, 2003, 3 (2), and Parfitt and Egorova, 2006.) In other cases, traditional creation stories have been threatened by genetic research and could even have negative practical outcomes. For instance, as Vine Deloria (1995: p. 82) observes, studies on Native American communities supporting the dominant story of the migration of their ancestors from the Bering Strait, promote the idea that Native Americans were not the original inhabitants of North America but 'late comers who had barely unpacked before Columbus came knocking at the door.' Problematically, the more Native Americans are seen as 'recent immigrants' the weaker their claims to land and sovereignty (Davis, 2004: p. 42). It has been pointed out that potential risks and benefits of this type of research go far beyond the individuals who actually agreed to take part in it and pertain to the group as a whole, which means a necessity for developing procedures for obtaining collective informed consent (Reardon, 2005; Widdows, 2007). However, the problem of seeking communal consent is complicated by the fact that the groups whose interests are concerned in these studies may be very diverse. As Davis (2004: p. 41) has put it, "a creation story that is a source of power for one subgroup may serve to marginalize another subgroup." (For further discussion see Brodwin, 2002; Elliott, 2003; Johnston 2003.)

Discussion

Scientists involved in such studies tend to portray their work as a neutral and objective contribution to historical research, a novel way of doing history by using 'superior' methods of genetics. Nevertheless, scholars coming from the perspective of social sciences and humanities disciplines have suggested that studies in genetic history indicate a worrying trend in DNA research, as they appear to naturalize social and cultural differences (e.g., Simpson, 2000; Abu El-Haj, 2007; Abu El-Haj, 2012; Palmie, 2007; Palsson, 2007; Reardon, 2005; Skinner, 2006; Smart et al., 2008).

A number of social scientists have paid particular attention to the way genetic anthropology has engaged with issues of personal and communal modes of self-identification, narratives of origin, and notions of relatedness. It has been lucidly demonstrated that such DNA studies are often informed by preexisting cultural and political discourses about the meaning of histories that they endeavor to reconstruct, and that in the imagination of the tested and their 'observers' the genetic markers 'assigned' to populations in the course of this kind of research are likely to be reinscribed as markers of social identification. To give an example related to genetic research on Jewish communities, Nurit Kirsh has argued that Israeli scientists involved in such studies in the 1950s almost unconsciously internalized Zionist ideology, which found expression in their work trying to prove the 'common origin' of various Jewish groups around the world (2003). Nadia Abu El-Haj cites the example of genetic research on the Lemba – a Judaising group in southern Africa, whose claims to Jewish origin have received a 'positive' response from geneticists, and suggests that the genetic study that established a 'biological' connection between the Lemba and the Jews has paved the way for their recognition by a number of Jewish organizations and educational charities (Abu El-Haj, 2012). (For a detailed analysis of genetic research on the Lemba see Parfitt and Egorova, 2006.)

At the same time, other commentators have pointed out that this 'new' genetic knowledge hardly superseded communal traditions or led to the emergence of new forms of belonging that would be completely at odds with those already in existence. Thus, Nikolas Rose suggests that "ideas about biological, biomedical, and genetic identity will certainly infuse, interact, combine and contest with other identity claims," but they can hardly be expected ever to supplant them (2007: p. 113).

Alondra Nelson in her study of the genetic ancestry tests offered to African-American and Black British citizens has convincingly argued that those who order these tests in an attempt to establish which part of Africa their ancestors may be from do not accept the results at face value but reinterpret them in light of their own 'genealogical aspirations.' Nelson therefore suggested that "while the geneticization of race and ethnicity may be the basic logic of genetic genealogy testing, it is not necessarily its inexorable outcome" (2008: p. 761). Reflecting on the role that genetic anthropology has played in such projects in the construction of collective and individual pasts, Marianne Sommer has observed that "[w]e have only just begun to understand the complex processes at work when DNA technologies enter into cultures of remembrance. Nonetheless, our current knowledge points towards the importance of the history and diversity of these cultures for the ways in which communities may or may not come to (re) imagine themselves in terms of new genetically based histories and in relation to biosocialities, which may or may not (re) form around genetic markers of ancestry" (2010: p. 387).

To return to the context of genetic research on the Jews, it has also been noted that it is too early to suggest that genetic constructions of a common Jewish origin are superseding other definitions of being Jewish. Thus, Barbara Prainsack and Yael Hashiloni-Dolev have argued that new 'discoveries' in 'Jewish genetics' have mostly remained without any political or practical consequences (2009). It has been demonstrated that in the case of the Lemba, as well as of some other 'emerging' Jewish communities, 'DNA evidence' did not play any role in defining their *halakhic* status (Halakhah is collective body of Jewish religious law) or their eligibility to the State of Israel (Parfitt and Egorova, 2006; Prainsack and Hashiloni-Dolev, 2009; Egorova, 2009a). Similarly, it has been demonstrated in relation to genetic research that engaged with the history of the caste system that, although 'DNA evidence' is widely used by 'lay commentators' as a rhetorical means for inscribing identities, it is often used selectively to support the more favored accounts about the origin and historical development of different groups. First, it was shown that genetic studies themselves offer different accounts of early Indian history: just like the explanations about the origin of the caste system provided by earlier commentators – from colonial scholars and administrators to Hindu reformers and nationalists were diverse – recent genetic studies have offered conflicting inferences on the nature of castes and tribes of the subcontinent. The models for the history of caste formation offered by geneticists were in their turn used by different social groups selectively in order to strengthen their own political agendas and to assert preferred identities (Egorova, 2009b). It could therefore be suggested that the interest that communities have expressed in embracing 'genetic history' indicates that while accepting its biological determinism, they also perceive it as imbued with liberatory potential (Egorova, 2009, 2010). Moreover, depending upon the specific social and political contexts in which genetic studies are commissioned, performed, and interpreted, their outcomes may have differing degrees of significance for the tested. As Sommer has argued in the context of her analysis of genetic ancestry testing in Switzerland, "there are people who perceive the genetic analysis as a new tool for self-(re)fashioning, and there are those whose (national) histories are controversial and for whom the genetic determination of origins and history becomes a serious arbiter in identity politics" (2012: p. 135).

Finally, the context that in recent years became particularly relevant to the discussion of the political significance of population genetics is that of biotechnology research in postcolonial world. It has been observed that genomics and biotechnology have become a measure of political and economic success for any country in the twenty-first century. (For an excellent discussion of the principle's applicability in 'non-Western' contexts see Ong, 2010; Sunder Rajan, 2006.) Therefore, it is not surprising that more and more governments outside Europe and North America are prepared to invest significantly in the development of the life sciences in an attempt to establish what has been described as 'genomic sovereignty' (Benjamin, 2009). Although these policies are designed to promote academic and economic independence for 'local' hubs of science and technology, they are also embedded in the global networks and processes of knowledge production. They 'naturalize' national populations in the name of postcolonial empowerment, but at the same time borrow practices and conceptual tropes from the wider context of 'genetic labeling' to reinforce already existing categories (Benjamin, 2009: p. 342).

To conclude, as Schramm, Skinner, and Rottenburg have put it, "The new genetics is part of changes that have a direct bearing on the notions of self and relatedness" (2012: p. 2). It is not at all surprising then that the new genetics was bound to introduce new dimensions into public, academic, and political discourses about the constructions of race and identity with the potential to reinstate old stereotypes and imbue them with new meanings, to continue the cultural work that was started by biological research of the past, and to add to it new complexity.

See also: Anthropology, Genomics, and Human Variation: National Roots; Genetic Ancestry Testing; Genetic Studies of Behavior: Methodology; Genetics and Indigenous Communities: Ethical Issues; Genetics and Society; Genetics: The New Genetics; Genomics, Ethical Issues In; Human Genomics in Asia; Race: Genetic Aspects.

Bibliography

Abu El-Haj, N., 2007. The genetic reinscription of race. Annual Review of Anthropology 36, 283–300.

Abu El-Haj, N., 2012. The Genealogical Science: the Search for Jewish Origins and the Politics of Epistemology. Chicago University Press, Chicago.

Benjamin, R., 2009. A lab of their own: genomic sovereignty as postcolonial science policy. Policy and Society 28, 341–355.

Brodwin, P., 2002. Genetics, identity, and the anthropology of essentialism. Anthropological Quarterly 75, 323–330.

Cavalli-Sforza, L.L., 2000. Genes, Peoples, and Languages. Pinguin Press, Allen Lane.

Davis, D., 2004. Genetic research and communal narratives. Hastings Center Report 34, 40–49.

Deloria Jr, V., 1995. Red Earth, White Lies: Native Americans and the Myth of Scientific Fact. Scribner, New York.

Dula, A., et al., 2003. The ethical and social implications of exploring African American genealogies. Developing World Bioethics 3, 133–141.

Duster, T., 2003. Buried alive. The concept of race in science. In: Goodman, A.H., Heath, D., Lindee, M.S. (Eds.), Genetic Nature/Culture. Anthropology and Science Beyond the Two-culture Divide. University of California Press, Berkeley, Los Angeles, London, pp. 258–278.

Duster, T., 2005. Race and reification in science. Science 307, 1050–1051.
Egorova, Y., 2009a. The truth is in the genes? Jewish responses to DNA research. Culture and Religion 10 (3), 159–177.
Egorova, Y., 2009b. De/geneticizing caste: population genetic research in South Asia. Science As Culture 18 (4), 417–434.
Egorova, Y., 2010. Castes of genes? Representing human genetic diversity in India, genomics. Society and Policy 6 (3), 18–35.
Elliott, C., 2003. Adventures in the gene pool. Wilson Quarterly 27, 12–22.
Epstein, S., 2007. Inclusion: the Politics of Difference in Medical Research. University of Chicago Press, Chicago.
Fausto-Sterling, A., 2004. Refashioning race: DNA and the politics of health care. Differences 15, 1–37.
Fullwiley, D., 2007a. The molecularization of race: institutionalizing human difference in pharmacogenetic practice. Science as Culture 16 (1), 1–30.
Fullwiley, D., 2007b. Race and genetics; attempts to define the relationship. Biosocieties 2, 221–237.
Haraway, D., 1997. Modest_Witness@Second_Millennium. FemaleMan_Meets_Oncomouse. Feminism and Technoscience. Routledge, New York.
Hartigan, J., 2008. Is race still socially constructed? The recent controversy over race and medical genetics. Science As Culture 17 (2), 163–193.
Hartigan, J., 2010. Race in the 21st Century: Ethnographic Approaches. Oxford University Press, New York, Oxford.
Johnston, J., 2003. Resisting a genetic identity: the black seminoles and genetic tests of ancestry. Journal of Law, Medicine and Ethics 31, 262–271.
Kahn, P., 1994. Genetic diversity project tries again. Science 266 (5186), 720–723.
Kahn, J., 2004. How a drug became "ethnic": law, commerce and the production of racial categories in medicine. Yale Journal of Health, Policy, Law and Ethics 4, 1–46.
Kahn, J., 2011. BiDiL and racialized medicine. In: Krimsky, S., Sloan, K. (Eds.), Race and the Genetic Revolution: Science, Myth, and Culture. Columbia University Press, New York, pp. 129–142.
Kirsh, N., 2003. Population genetics in Israel in the 1950s: the unconscious internalization of ideology. Isis 94, 631–655.
Leroi, A., 2005. A family tree in every gene. New York Times, 14 March.
Lewontin, R., 1972. The apportionment of human diversity. Evolutionary Biology 6, 391–398.
Lipphardt, V., 2012. Isolates and crosses in human population genetics; or, a contextualization of German race science. Current Anthropology 53 (Suppl. 5), S69–S82.
Marks, J., 2003. What Does it Mean to be 98% Chimpanzee? Apes, Peoples and the Genes. University of California Press, Berkeley and Los Angeles.
M'charek, A., 2005. The Human Genome Diversity Project. An Ethnography of Scientific Practice. Cambridge University Press, Cambridge.
Nelson, A., 2008. Bioscience: genetic genealogy testing and the pursuit of African ancestry. Social Studies of Science 38, 759–783.
Ong, A., 2010. Introduction. An analytics of biotechnology and ethics at multiple scales. In: Ong, A., Chen, N.C. (Eds.), Asian Biotech. Ethics and Communities of Fate. Duke University Press, Durham and London, pp. 1–55.
Palmie, S., 2007. Genomics, divination, "racecraft". American Ethnologist 34, 205–222.
Palsson, G., 2007. How deep is the skin? BioSocieties 2, 257--272.
Parfitt, T., 2002. Genes, religion, and history: the creation of a discourse on origin among a Judaizing African tribe. Jurimetrics 42, 209–219.
Parfitt, T., Egorova, Y., 2006. Genetics, Mass Media and Identity: a Case Study of the Genetic Research on the Lemba and Bene-Israel. Routledge, London.
Prainsack, B., Hashiloni-Dolev, Y., 2009. Religion and nationhood: collective identities and the new genetics. In: Glasner, P., Lock, M. (Eds.), Genetics and Society Handbook: Mapping the New Genomic Era. Routledge, London, pp. 404–421.
Reardon, J., 2005. Race to the Finish: Identity and Governance in an Age of Genomics. Princeton University Press, Princeton and Oxford.
Risher, M.T., 2011. Racial disparities in databanking of DNA profiles. In: Krimsky, S., Sloan, K. (Eds.), Race and the Genetic Revolution: Science, Myth, and Culture. Columbia University Press, New York, pp. 47–68.
Rose, N., 2007. The Politics of Life Itself: Biomedicine, Power, and Subjectivity in the Twenty-First Century. Princeton University Press, Princeton and Oxford.
Sabeti, P., 2002. DNA: an Alternative Record of African History. Available from: http://www.africana.com/Articles/tt_524.htm (accessed 19.06.02).
Schramm, K., Skinner, D., Rottenburg, R., 2012. Introduction. Ideas in motion: making sense of identity. In: Schramm, K., Skinner, D., Rottenburg, R. (Eds.), Identity Politics and the New Genetics. Berghahn Books, New York, Oxford, pp. 1–30.
Simpson, B., 2000. Imagined genetic communities: ethnicity and essentialism in the twenty-first century. Anthropology Today 16, 3–6.
Skinner, D., 2006. Racialized futures: biologism and the changing politics of identity. Social Studies of Science 36, 459–488.
Skorecki, K., et al., 1997. Y chromosomes of Jewish priests. Nature 385, 32.
Smart, A., Tutton, R., Martin, P., Ellison, G.T.H., Ashcroft, R., 2008. The standardization of race and ethnicity in biomedical science editorials and UK biobanks. Social Studies of Science 38, 407–423.
Sommer, M., 2008. History in the gene: negotiations between molecular and organismal anthropology. Journal of the History of Biology 41, 473–528.
Sommer, M., 2010. DNA and cultures of remembrance: anthropological genetics, biohistories and biosocialities. BioSocieties 5 (3), 366–390.
Sommer, M., 2012. Do you have Celtic, Jewish or Germanic roots? In: Schramm, K., Skinner, D., Rottenburg, R. (Eds.), Identity Politics and the New Genetics. Berghahn Books, New York, Oxford, pp. 116–141.
Spurdle, A., Jenkins, T., 1996. The origins in of the Lemba "Black Jews" of Southern Africa; evidence from p12F2 and other Y-chromosome markers. American Journal of Human Genetics 59 (5), 1126–1133.
Stepan, N., 2003. Science and race: before and after the human genome project, in fighting identities: race, religion and ethno-nationalism (Special issue). Socialist Register 39, 329–346.
Sunder Rajan, K., 2006. Biocapital. The Constitution of Postgenomic Life. Duke University Press, Durham and London.
Thomas, M., et al., 1998. Origins of old testament priests. Nature 394 (6689), 138–140.
Thomas, M., et al., 2000. Y chromosomes travelling south: the Cohen modal haplotype and the origins of the Lemba – the 'Black Jews' of Southern Africa. American Journal of Human Genetics 66, 674–686.
Wailoo, K., Pemberton, S., 2006. The Troubled Dream of Genetic Medicine: Ethnicity and Innovation in Tay-Sachs, Cystic Fibrosis, and Sickle Cell Disease. The Johns Hopkins University Press, Baltimore.
Wallace, H., 2011. Prejudice, stigma, and DNA databases. In: Krimsky, S., Sloan, K. (Eds.), Race and the Genetic Revolution: Science, Myth, and Culture. Columbia University Press, New York, pp. 68–99.
Washington, H., 2010. Base assumptions? Racial aspects of US DNA forensics. In: Hindmarsh, R., Prainsack, B. (Eds.), Genetic Suspects: Global Governance of Forensic DNA Profiling and Databasing. Cambridge University Press, Cambridge.
Widdows, H., 2007. Conceptualising the self in the genetic era. Health Care Analysis 15, 5–12.

New Institutionalism in the Analysis of Complex Organizations

Walter W Powell, Stanford University, Stanford, CA, USA
Patricia Bromley, University of Utah, Salt Lake, UT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The new institutionalism in sociology and organizational research is best represented as an extended family of scholars that share a broadly defined theoretical orientation. The multiple lines of thought in this tradition have a common interest in the relationship between social structure and organizations. Within this frame, researchers differ in their conceptualization of institutions (e.g., as specific legal or resource relationships versus a wider culture), the focus of causal arguments (e.g., top-down versus bottom-up), and the focal feature of organizations (e.g., formal structures versus practices and behavior). This robust theoretical tradition provides a valuable lens for understanding contemporary organizations and management.

Introduction

An institutional turn has taken place across the social sciences, with active strands arising in economics, political science, and sociology. In economics, rational choice institutionalism emphasizes coordination mechanisms that solve collective action problems by generating new forms of commitment and rule-following or norm-abiding behavior. Historical institutionalism, a cornerstone of comparative politics research, underscores how institutions emerge from and are embedded in temporal processes through path dependence and divergence at critical historical junctures. Sociological institutionalism shares this general emphasis on the influence of broader social structures, but takes a more cultural view of institutions, highlighting their role in providing models or scripts for action. We would not be the first to paraphrase former President Richard Nixon and argue: 'We are all institutionalists now.' In this article we focus on the sociological stream of new institutionalism, which has become particularly dominant in contemporary organization and management theory.

The Foundations of Sociological Institutionalism

In early analyses of the process of institutionalization – the 'old' institutionalism, if you will, scholars such as Selznick, Zald, and Clark explored how the policies and routines of a particular organization acquired local resonance and became self-sustaining. Selznick (1949) famously documented how the Tennessee Valley Authority (TVA)'s efforts to coopt grass-roots agrarian groups succeeded, and the resulting accommodations eventually took on a life of their own, supplanting the TVA's original purposes. Institutionalization, in Selznick's (1957: p. 17) eyes, refers to the process whereby practices become 'infused with value beyond the technical requirements of the task at hand.'

In contrast to this initial focus largely on internal organizational dynamics, a 'new' institutionalism shifted attention outward to examine how the external environment socially constructs organizations, providing them with templates for their formal structures and policies, and thereby increasing an organization's legitimacy in the wider world. New institutional studies of organizations in the 1970s and 1980s are largely characterized by emphases on diffusion, isomorphism, and decoupling. As put by DiMaggio and Powell (1991: p. 8): "The new institutionalism in organization theory and sociology comprises a rejection of rational–actor models, an interest in institutions as independent variables, a turn toward cognitive and cultural explanations, and an interest in properties of supraindividual units of analysis that cannot be reduced to aggregations or direct consequences of individuals' attributes or motives."

In a foundational paper, Meyer and Rowan (1977) set out a view of complex organizations in postindustrial society as reflecting wider myths in the institutional environment rather than the technical demands of production. In order to protect an organization's technical core, they posited a great deal of decoupling between espoused policies and actual practices, as well as internal buffering among subunits. Further, as the environment impinges on organization, considerable homogeneity among organizations develops in their formal structures, beyond any internal technical needs. Drawing on work in social psychology, Zucker (1977) emphasized that institutionalization is an organizational level process, and viewed diffusion as a consequence of institutionalization rather than a cause. Echoing this emphasis on mechanisms, DiMaggio and Powell (1983) brought causal force to the processes of diffusion and isomorphism by spelling out mimetic, normative, and coercive channels through which ideas and practices flowed. In so doing, they emphasized the important roles of the professions, higher education, and media in promulgating ideas and projects that are widely incorporated by modern organizations. In a subsequent article, DiMaggio and Powell (1991) drew on Bourdieu, Goffman, and ethnomethodology to fashion a theory of practical action, in an effort to build linkages between internal organizational routines and habits and macrosocial forces.

Even among these early studies, it is perhaps more appropriate to think of related lines of inquiry, rather than a unitary conceptual paradigm. Institutionalism in organization theory is a big tent, in which scholars share a general conception of institutions as having normative, cultural, and regulative elements (Scott, 2001). But under the broad canopy of

attentiveness to external institutional influences, scholars differ in both their conceptualization of institutional environments and in the extent to which individuals and organizations are seen as being fundamentally shaped *or* constituted by the external environment. At the most constructivist end of the continuum, John Meyer and colleagues emphasize the social creation of actors themselves (Meyer, 2010). In this view, individuals and organizations are constituted by their environment. At the other end of this spectrum are depictions of entrepreneurs pursuing self-determined interests within a set of institutional constraints and supports, and in so doing building or cocreating institutions (Battilana et al., 2009; Lawrence et al., 2009).

The conceptual breadth of the new institutional family, and the productive conflicts among its variants, contributes to the theory's ongoing intellectual vibrancy and growth. Older divisions over agency and power continue, but at least since the 1990s new debates have come to the forefront. First, recent scholarship has extended Zucker's (1977) emphasis on the microfoundations of institutional processes, bringing greater attention to the linkages between individuals and institutions. Second, others seek to move away from a focus on isomorphism and toward institutional sources of heterogeneity (Greenwood et al., 2011), most notably through ideas of institutional logics. A Scandinavian strand of work emphasizes translation and editing, recognizing that as institutionalized practices flow around the globe, the nature of reception and adoption often reflects strong local contingencies (e.g., Brunsson and Jacobsson, 2000; Czarniawska and Sevón, 1996). Others have reconceptualized theories of loose coupling (Bromley and Powell, 2012). Third, in macrolevel research, Meyer and Rowan's early insights about organizations have been extended globally, creating a stream of research known as world polity or world society theory. In what follows we discuss developments in new institutional theory in each of these areas since the 1990s.

Microfoundations: How Are Institutions Formed and Reproduced?

Overview

Initial work on microfoundations was prompted by a recognition that macrolines of research could profit from a micromotor. Such a motor would involve theories that attend to enaction, sensemaking, and meaning. Institutions are sustained, altered, and extinguished as they are enacted by individuals in concrete social situations. This line of work is guided by an effort to understand how individuals locate themselves in social relations and interpret and respond to their institutional context.

Institutional forces shape individual interests and desires, framing the possibilities for action and influencing whether behaviors result in persistence or change. Macroinstitutional effects create conventions that are the scripts for meaning making. This process is recursive and self-reinforcing. Powell and Colyvas (2008) lamented, however, that the individuals that populated much of institutional analysis were portrayed as either 'cultural dopes' or heroic 'change agents.' Moreover, too many analyses conflate macrofactors with structural forces and assume these factors only reinforce stability and homogeneity, while associating microfactors with entrepreneurship and agency. But individuals also play an important role in inhabiting institutions (Hallett and Ventresca, 2006), and organizations can serve as entrepreneurs. Moreover, macrotrends, such as globalization, can be profoundly destabilizing to local orders and individuals. Clearly, it is a mistake for institutional analysts to blindly equate change with the microlevel and persistence with the macrolevel.

Contributions

Institutions are reproduced through the everyday activities of individuals. Members of organizations engage in daily practices, discover puzzles or anomalies in their work, problematize these questions and develop answers to them by theorizing them. In turn, participants ascribe meaning to these theories and, in so doing, develop and reproduce taken-for-granted understandings. Jepperson (1991) was one of the first to emphasize microtranslation – an understanding of how macrocategories get inside the heads of individuals. Macroframings or values can be 'pulled down' to the everyday level of practice, as varied activities can be pursued under a common interpretation or account, or diverse practices can be pursued in the search of a common goal (Colyvas and Powell, 2006). Many microprocesses represent local instantiations of macrolevel trends. DiMaggio and Powell's (1991) overview of the elements of a theory of practical action similarly focused on microtranslation. Responding to readings of their 1983 article that contended that mimetic and normative isomorphism entailed 'mere' copying and replication, they emphasized that practical consciousness involves energy, effort, and reflection. DiMaggio and Powell argued that habitual action does not reflect passivity, but is a skilled means of directing attention. Everyday reasoning requires individuals to negotiate rules and procedures flexibly and reflexively to assure themselves and others around them that their behavior is sensible.

In a wide range of empirical ethnomethodological studies, categories and classifications become interpretive schema that members of organizations draw on. Over time, these schemas become a repository of organizational knowledge. As particular schemas become routinized through repeated application and use, they develop a habitual, taken-for-granted character. Berger and Luckmann (1967) emphasized that once joint activities are habitualized and reciprocally interpreted, patterns both harden and deepen as they are transmitted to others, particularly newcomers. When schemas become perceived as objective, exteriorized facts, their contingent origins are obscured.

Ethnomethodologists have demonstrated how classifications and categorizations are invoked on the fly by skilled actors to keep peace on the streets, in the courts, in hospital wards, and welfare agencies. Rather than struggling with or coping with uncertainty, the practical reasoning view emphasizes how situations are rendered comprehensible, and sees such efforts as an ongoing, contingent accomplishment. This is in contrast to a Weberian view of the typical bureaucracy more as a target or ideal. For the ethnomethodologists, bureaucracy

is neither a rarified nor lofty goal, but deeply embedded in commonsense routines of everyday life. Organization is a formula to which all sorts of problems can be brought for solution.

A related line of microresearch builds on ideas involving sensemaking, associated principally with Karl Weick (1995). Sensemaking is inherently retrospective and precedes action because situations are only understood upon completion. Meaning is shaped through attention to what has already occurred, and is therefore directed, not attached, to action. Because outcomes and subjective objects are implicit in interpretation, sensemaking entails a process that simultaneously enacts identity and institutions. Identity is central because individuals act based on who they are, not on what choices they have, and this feature is constituted out of the process of interaction. Mead's (1934) insights that each individual is a 'parliament of selves' and that 'social processes precede the individual mind' are critical. The environment is not viewed as a fixed and stable reality, but as a coconstruction of individuals' minds and their actions. Enactment represents the reciprocal interaction of the material and the cognitive world.

A sensemaking approach directs attention to the importance of language, routines, and communication for analyzing microprocesses. While emphasizing that various institutional materials are commonly 'pulled down' by individuals and translated within organizations, these processes may differ across circumstances. Multiple modes of meaning-making occur at the interface of identity and the enacted environment, and how such understandings are forged and enacted occur through retrospection. Sensemaking is thus a key micromechanism of institutionalization that allows consideration of both the 'cognitive complexities' that guide organizational behavior and recognition of the varied ways that institutionalized practices operate at the microlevel (Jennings and Greenwood, 2003).

Over the past decade, a new group of institutional scholars seeking to explain how extant institutional arrangements are altered have turned to the concept of institutional entrepreneurship, as highlighted by DiMaggio (1988: p. 14): "New institutions arise when organized actors with sufficient resources (institutional entrepreneurs) see in them an opportunity to realize interests that they value highly." The concept of institutional entrepreneurs was first developed by Eisenstadt (1980) in his work on Weber's concept of charisma, where he studied rare, but epochal, individuals who sparked institutional change. These are creative individuals whose social positions and skills allow them to recognize problems or opportunities and take advantage of enabling conditions to alter the social order (Battilana et al., 2009).

Work on institutional entrepreneurship has emphasized the microlevel processes through which intentional institutional change takes place, including acts of convening, the framing of projects, the process of mobilizing resources, and the binding of new practices and forms to extant institutions (Hardy and Maguire, 2008; provide a good overview of these studies). Such work has shed light on processes of organization-building and other kinds of institutional innovation that went largely unexamined in early studies.

Challenges

Discussions of microtheories that portray individuals and institutions as mutually constitutive are useful, but this work remains fairly conceptual and, thus far, lacks rich empirical grounding. Indeed, it is unclear at this stage how such work could be conducted, outside of detailed ethnographic study. Work on institutional entrepreneurship represents an effort to replace the imagery of oversocialized individuals, slavishly devoted to habits and fashions of the day. But the new celebration of entrepreneurs has possibly gone too far, as not all changes are led by purported entrepreneurs. A further limitation of work derives from the relatively short temporal frame covered by most studies of institutional entrepreneurship, which regularly slice historical processes that are unfolding on the scale of many decades or even centuries into very short time periods. Thus, critical macrolevel processes remain obscured from analytical view, especially when they are taking place beyond the immediate fields under investigation. Against this relatively static backdrop, the 'embedded agency' of individual institutional entrepreneurs can loom artificially large. Moreover, because the cases studied have almost exclusively involved successful projects, it has been tempting for researchers to seek out confirming signs of 'heroic' contributions to observed outcomes. This sampling on the dependent variable renders any possibility of generalization moot.

Crossing Levels: Institutional Logics

Overview

Scholarship on institutional logics focuses on how competing societal level belief systems shape both individuals and organizations. This stream of research focuses on the effects of distinct and diverse institutions in influencing behavior inside organizations. Thornton and Ocasio (2008: p. 99) suggest that: "By providing a link between institutions and action, the institutional logics approach provides a bridge between macro-/ structural perspectives and more micro-/ process approaches." The institutional logics perspective has proven valuable for spurring research on three fronts: change, complexity, and practice variation.

Thornton and Ocasio (2008: p. 101) developed a widely used definition of institutional logics as 'the socially constructed, historical patterns of material practices, assumptions, values, beliefs, and rules by which individuals produce and reproduce their material subsistence, organize time and space, and provide meaning to their social reality.' Heavily influenced by Friedland and Alford's (1991: p. 232) depiction of five key institutions – the capitalist market, bureaucratic state, democracy, nuclear family, and Christianity, each guided by a distinctive modus operandi, this approach sees change originating from contestation between rival logics. A more recent formulation of the interinstitutional system retains this macroemphasis, but melds the operational and ideological elements of institutions (e.g., the bureaucratic state versus democracy) and expands the application of the logics perspective beyond Christian and democratic countries (Thornton et al., 2012). The authors identify six distinctive societal institutions – the market, the corporation, the

professions, the state, the family, and religions. Within each of these metainstitutions, there are differences – in the sources of legitimacy, authority, and identity; basis of norms, attention, and strategy; informal control mechanisms; and economic rewards. Each societal sector is linked to a specific logic that shapes its material practices and organizational forms. There is much in common here with recent French social theory on different orders of worth, see particularly Boltanski and Thévenot (2006), though these parallels have not been pursued.

Contributions

The concept of logics has been valuable in depicting social transformations. Much research has focused on either a shift from one dominant logic to another, or the layering of multiple logics. Scott et al. (2000) detailed how the growth and expansion of new logics in healthcare led to the valorization of different actors, behaviors, and governance structures. Even as doctors and their patients were replaced by healthcare providers and consumers, they stressed that professional, bureaucratic and market logics intermingled, with one supplanting though not eliminating the other. The central idea is that institutional logics focus the attention of key actors on a particular set of problems and solutions. Changes in observed material practices and meanings can reflect movement between institutional logics.

Research on logics also provides a way for examining complexity, pluralism, and heterogeneity. Friedland and Alford's (1991) initial conception of logic was intended to describe competing and even at times contradictory practices and beliefs in modern Western societies. The tensions inherent in different logics provide the resources for individuals and organizations to transform social practices and structures. By focusing on how some fields are composed of multiple logics, and thus, varied forms of institutionally based rationality, institutional analysts offer new insight into practice variation and the dynamics of adoption. Multiple logics can create diversity in practices by enabling variety in cognitive orientations, prompting contestation over which practices are appropriate. These contradictions in multiple logics can also create ambiguity, leading to logic blending, the creation of new logics, or the continued emergence of new variants. This revision to a theoretically abstract and analytically distinct set of ideal types makes it useful for studying multiple logics in conflict and consensus, the hybridization of logics, and institutions in other parts of society and the world. This perspective emphasizes heterogeneity, whereas the earlier macroinstitutional perspective focused on similar changes across differentiated sectors that are attributed to one master form of rationality.

Challenges

Despite its benefits for understanding change, complexity, and practices, the institutional logics perspective is not without challenges. One issue lies in the source of logics. Where do they come from? Why these six? Do they operate with differential force, in different social settings? When and why is conflict among logics provoked? It is often not clear why actors begin to prefer one logic over another, thus much of the research is more descriptive than explanatory. (For a very different causal account of how distinctive roles and networks prompt conflicts and generate novelty, see Padgett and Powell, 2012.) A second criticism lies in ambiguity with the term 'institutional logic.' The definition put forth by Friedland and Alford (1991), as well as Thornton et al. (2012), is tied to megastructures in society, but at times the phrase is disconnected from the macrointerinstitutional system and used to describe nearly any moment of conflict or complexity, such as clashes between departments or individuals over 'love or money.' When microlevel observations are not linked back to the larger structures of society it is unclear what is particularly institutional about the arguments or what the larger logic involved may be.

Macrolevel: World Polity

Overview

Whereas the logics perspective emphasizes differences among societal level institutions, research in the world polity tradition illuminates how modernity itself is reshaping all sectors of society, from education to firms, and at all levels, from nation-states to individuals. This perspective is historically linked to John Meyer and his many collaborators at Stanford University. These scholars study the global arena and cross-national social changes in the period since World War II, and have been particularly prolific in the areas of management, education, science, international nongovernmental organizations, human rights, and the environment (Drori et al., 2006a). Drawing on Max Weber, these scholars emphasize rationalization, progress, and individualism as the foundational cultural assumptions underpinning modern organizations worldwide. These universalistic cultural foundations are assumed to shape complex organizations in similar ways, regardless of the substantive field or specific location. Churches, government agencies, universities, and private firms are thought to increasingly take on similarly rationalized elements. As with Berger and Luckmann's (1967) phenomenological view, actors and organizations are highly scripted. Thus, as with the microfoundations ideas, the social psychology involved is closest to that of Goffman and Mead, stressing the constructed nature of reality and emphasizing that much behavior reflects the enactment of socially appropriate frames in a given context.

Contributions

The earliest work in this tradition came from Meyer's observation of schools in the 1970s. In schools, and comparable types of complex organizations, there is greater structural similarity than seems plausible if these entities are thought to reflect only immediate technical needs or situated power dynamics. Instead, organizations are creations of their external environment. In the world society tradition, the cultural frame of reference is increasingly at the world level, rather than bounded by a particular industry or even national context (Meyer et al., 1997). The global cultural changes that underpin the expansion and internal elaboration of

organizations are characterized by increasing education (especially higher education), the empowerment of individuals through emphases on human rights, and the proliferation of scientific (and social scientific) thought and method (see, for example, the collection of writings in Krueken and Drori, 2009).

This intellectual tradition is one of the central sources of cross-national organizational research. For example, in a study of national governance practices, Drori et al. (2006b) find that increasingly rationalized administration is driven by linkages to world society (measured through trade openness, memberships in international organizations, and the penetration of science), net of measures of modernization. More recently, Schofer and Longhofer (2011) have documented how world society promotes the rapid emergence of nongovernmental organizations in countries worldwide. Thus, world society influences both the nature of organizational structures (highly rationalized) and the numbers of organizations found worldwide (increasing), over and above any local technical demands.

Challenges

Given that a central tenet of world polity research is that 'actors' and 'interests' are constituted by the wider institutional environment, it is no surprise that the strongest critiques revolve around the oversocialization of actors and a lack of attention to coercive power. To a certain degree, both the logics perspective and emphases on microfoundations emerged in reaction to phenomenological neoinstitutionalism. The ideas of institutional entrepreneurs and logics emphasize individuals pursuing interests within a social framework that is largely exogenous to their existence. In contrast, the world society view supposes that the larger framework is what gives entities their status as an actor and specifies the interests that are available to them. Further, a great deal of empirical work in this tradition documents the diffusion of all sorts of formal structures (e.g., policies, plans, or laws). In turn, such a focus generates two sorts of concerns. First, some contend that formal structures are superficial (relative to actual practices), and thus the research tackles a peripheral issue. Second, others interpret the emphasis on diffusion and homogeneity as having a teleological flavor; modernity and rationalization appear to be on an inevitably increasing trajectory toward an end where all entities look superficially similar.

Conclusion

We conclude with several intentionally provocative statements in the hope they might prompt theoretical progress and new research directions. First, existing conceptions of both embedded agency and structural constraints are rather simplistic. In lieu of descriptive accounts, future work might address: Under which conditions are organizations and their members likely to have autonomy from their environment, and when are they more constrained? How can we explain the variable nature of constraints and capacities of particular actors in different contexts? For example, in more centralized political or economic settings, there may be less room for robust action by the average individual or organization. Similarly, in settings where traditional forms of authority are strongest (e.g., religion, family), there may be less room for autonomy. Institutional environments with decentralized or fragmented sources of social authority may offer greater opportunity for microlevel strategic action. In short, more carefully specified, comparative conceptualizations of the relationship between institutions and actors would be valuable.

Second, the term 'institution' needs to be concretely specified, or else its purchase becomes weak. Hand-in-hand with the growth of the neoinstitutional family has come increasing diversity and vagueness in the use of this term. If the many implicit understandings of institutions found in the literature were made explicit, then clearer distinctions between the various strains of research would come into focus and more productive boundary-spanning discussions could ensue.

Presently, the institutional paradigm is quite highly developed. With such a rich and widely used research tradition, insularity is always a risk. Rather than solely conversing with each other, neoinstitutionalists should look to insights that can be gained from other approaches. Recent studies at the intersection of social movement theory and neoinstitutionalism are one such step. Efforts to link institutional research with European traditions, such as actor–network theory, are another. A capacious theoretical pluralism will help organizational research to remain firmly rooted in addressing key puzzles in the empirical world.

See also: Bureaucracy, Sociology of; Economic Sociology; Institutions; Organizations, Sociology of; Rational Choice Theory in Sociology; Social Networks.

Bibliography

Battilana, J., Leca, B., Boxenbaum, E., 2009. How actors change institutions: towards a theory of institutional entrepreneurship. The Academy of Management Annals 3 (1), 65–107.

Berger, P.L., Luckmann, T., 1967. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Anchor, New York.

Boltanski, L., Thévenot, L., 2006. On Justification: Economies of Worth. Princeton University Press, Princeton, NJ.

Bromley, P., Powell, W.W., 2012. From smoke and mirrors to walking the talk: decoupling in the contemporary world. The Academy of Management Annals 6 (1), 483–530.

Brunsson, N., Jacobsson, B., 2000. A World of Standards. Oxford University Press, Oxford.

Colyvas, J.A., Powell, W.W., 2006. Roads to institutionalization: the remaking of boundaries between public and private science. Research in Organizational Behavior 27, 305–353.

Czarniawska, B., Sevón, G. (Eds.), 1996. Translating Organizational Change, vol. 56. de Gruyter, Berlin.

DiMaggio, P.J., 1988. Interest and agency in institutional theory. In: Zucker, L. (Ed.), Institutional Patterns and Organizations: Culture and Environment. Ballinger Publishing Company, Cambridge, MA, pp. 3–22.

DiMaggio, P.J., Powell, W.W., 1983. The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. American Sociological Review 48 (2), 147–160.

DiMaggio, P.J., Powell, W.W. (Eds.), 1991. The New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago, IL.

Drori, G.S., Meyer, J.W., Hwang, H., 2006a. Globalization and Organization: World Society and Organizational Change. Oxford University Press, Oxford.

Drori, G.S., Jang, Y.-S., Meyer, J.W., 2006b. Sources of rationalized governance: cross-national longitudinal analyses, 1985–2002. Administrative Science Quarterly 51 (2), 205–229.

Eisenstadt, S.N., 1980. Cultural orientations, institutional entrepreneurs, and social change: comparative analysis of traditional civilizations. American Journal of Sociology 85 (4), 840–869.

Friedland, Roger, Alford, Robert R., 1991. Bringing society back in: symbols, practices, and institutional contradictions. In: DiMaggio, P., Powell, W.W. (Eds.), The New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago, IL, pp. 232–266.

Greenwood, R., Raynard, M., Kodeih, F., Micelotta, E.R., Lounsbury, M., 2011. Institutional complexity and organizational responses. The Academy of Management Annals 5 (1), 317–371.

Hallett, T., Ventresca, M.J., 2006. Inhabited institutions: social interactions and organizational forms in Gouldner's patterns of industrial bureaucracy. Theory and Society 35 (2), 213–236.

Hardy, C., Maguire, S., 2008. Institutional entrepreneurship. In: Greenwood, R., Oliver, C., Suddaby, R., Sahlin, K. (Eds.), The Sage Handbook of Organizational Institutionalism. Sage Publications, London, UK, pp. 198–217.

Jennings, P.D., Greenwood, R., 2003. Constructing the iron cage: institutional theory and enactment. In: Westwood, R., Clegg, S. (Eds.), Debating Organization: Point-counterpoint in Organization Studies. Blackwell Publishing Ltd, London, UK, pp. 195–207.

Jepperson, R.L., 1991. Institutions, institutional effects, and institutionalism. In: DiMaggio, P., Powell, W.W. (Eds.), The New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago, IL, pp. 143–163.

Krueken, Georg, Drori, Gili, 2009. The Writings of John W. Meyer. Oxford University Press, Oxford, UK.

Lawrence, T.B., Suddaby, R., Leca, B. (Eds.), 2009. Institutional Work: Actors and Agency in Institutional Studies of Organizations. Cambridge University Press, Cambridge, UK.

Mead, G.H., 1934. Mind, Self and Society. University of Chicago Press, Chicago, IL.

Meyer, J.W., 2010. World society, institutional theories, and the actor. Annual Review of Sociology 36, 1–20.

Meyer, J.W., Rowan, B., 1977. Institutionalized organizations: formal structure as myth and ceremony. American Journal of Sociology 83 (2), 340–363.

Meyer, J.W., Boli, J., Thomas, G.M., Ramirez, F.O., 1997. World society and the nation-state. American Journal of Sociology 103 (1), 144–181.

Padgett, J.F., Powell, W.W., 2012. The Emergence of Organizations and Markets. Princeton University Press, Princeton, NJ.

Powell, W.W., Colyvas, J.A., 2008. Microfoundations of institutional theory. In: Greenwood, R., Oliver, C., Suddaby, R., Sahlin, K. (Eds.), The Sage Handbook of Organizational Institutionalism. Sage Publications, London, UK, pp. 276–298.

Schofer, E., Longhofer, W., 2011. The structural sources of association. American Journal of Sociology 117 (2), 539–585.

Scott, W. Richard, 2001. Institutions and Organizations. Sage, Thousand Oaks, CA.

Scott, W Richard, Ruef, Martin, Mendel, Peter, Caronna, Carol, 2000. Institutional Change and Healthcare Organizations: from Professional Dominance to Managed Care. University of Chicago Press, Chicago, IL.

Selznick, P., 1949. TVA and the Grass Roots: A Study of Politics and Organization. University of California Press, Berkeley, CA.

Selznick, P., 1957. Leadership in Administration: A Sociological Interpretation. University of California Press, Berkeley, CA.

Thornton, P.H., Ocasio, W., 2008. Institutional logics. In: Greenwood, R., Oliver, C., Suddaby, R., Sahlin, K. (Eds.), The Sage Handbook of Organizational Institutionalism. Sage Publications, London, UK, pp. 99–129.

Thornton, P.H., Ocasio, W., Lounsbury, M., 2012. The Institutional Logics Perspective: A New Approach to Culture, Structure, and Process. Oxford University Press, Oxford.

Weick, K.E., 1995. Sensemaking in Organizations, vol. 3. SAGE Publications, Thousand Oaks, CA.

Zucker, L.G., 1977. The role of institutionalization in cultural persistence. American Sociological Review 42 (5), 726–743.

New Managerialism and New Public Sector Management

Tony Cutler, Centre for Research on Socio-Cultural Change, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article seeks to introduce the reader to 'new managerialism' (NM). It outlines the basic concepts and ideas underlying NM and considers the effects these have had in changing the way in which public sector services are organized. It also examines some of the principal criticisms of NM as a reform program designed to improve the performance and accountability of the public sector.

Introduction

In July 2006, the then United Kingdom Prime Minister, Tony Blair, told an interviewer from the (London) *Financial Times* that his administration was undertaking 'capability reviews' of central government departments. These reviews were designed to focus the government on 'delivery, project management and performance management' (Timmins, 2006). This statement reflected Blair's commitment to what has been termed the 'new public management' or 'new managerialism' (NM). This article is divided into four sections. The first considers the key concepts that define NM; the second examines the impact of such concepts on the organization of public sector services, particularly 'social' services such as education, health, and social care; the third section discusses the principal criticisms leveled at NM; and the concluding section seeks to bring the different strands of the argument together.

NM: Key Concepts

To approach the key concepts of NM, Hood's classic (1991) article provides a valuable starting point. Adapting his argument, three central strands of NM can be identified: the emphasis on improving the 'performance' of public sector services; the increased use of competition in the public sector; and related changes in the approach to 'managing' public sector organizations.

The emphasis on improving 'performance' involves collecting and publishing data on organizations providing public services such as hospitals, schools, and social services departments. Hood's (1991) concept also encompasses the use of incentives and sanctions (note Blair's reference to performance 'management'). The increased use of competition can take a variety of forms but a common feature is to render the provision of public services 'contestable.' The existing service provider is required to compete against alternative providers to retain the right to continue to provide the service (Hood, 1991). Finally, there are a set of changes in the approach to the management of the public sector. The public sector is 'disaggregated' (Hood, 1991) so that organizations responsible for 'strategic' decisions (not only policy making but also how performance will be measured and managed) are separated from organizations responsible for service delivery or implementing policy (see Duncan and Chapman, 2010 for this distinction in New Zealand, seen as the earliest example of a radical NM program). The management of organizations delivering services or implementing policy are to be set 'free' in that they are given increasing discretion over how performance objectives are to be achieved (Hood, 1991).

These changes are broadly justified with reference to two related objectives. The first is to focus on achieving 'results' in government (note Blair's reference to 'delivery'). The second is to create mechanisms that enhance 'accountability' for performance (Hood, 1991). Thus, NM can be presented as a coherent means of linking improved results and greater accountability. The emphasis on performance data is designed to make it clear whether the desired 'results' are being achieved and incentives and sanctions are means of 'managing' performance (Hood, 1991). Increased competition adds further pressure to 'perform' since incumbent service providers are threatened with losing their right to deliver the service. Performance data can also be used to assess how far a new provider is meeting the objectives set. Disaggregation is designed to make it easier to assess whether provider units are delivering 'results.' They are treated as 'stand-alone' organizations because the logic of disaggregation is that service managers are substantially 'free' of controls on how performance is achieved (Hood, 1991). Organizations designed to make strategic decisions are also seen as freed from having to consider detailed operational matters, which have now been 'devolved' downward.

Accountability for results is seen as having a plurality of dimensions but these dimensions are treated as complementary. The increased availability of performance data on service delivery provides 'upward' accountability to senior politicians and officials, occupying a 'strategic' role in the context of 'disaggregation' (Greener, 2009). However, such information (through the publication of performance data) is designed to increase accountability 'downward' to service users or 'consumers.' Increasing 'contestability' can also be seen as enhancing both 'upward' and 'downward' accountability. Alternatives to existing service providers can be used by senior politicians and officials to 'turn around' providers seen as consistently 'failing' to meet performance objectives (Cutler, 2007). Some forms of public sector markets, such as the use of individual budgets in social care, allow service users/consumers a role in purchasing services (Wilberforce et al., 2012). In other cases, service users are given a choice of service providers (see Considine et al., 2011 for the application of this principle in the Australian employment assistance

service). These practices can be seen as increasing accountability to users/consumers. Such a complementary view of forms of accountability conveys the impression that NM generates a 'win-win' situation where all forms of accountability for public sector service performance are improved.

The next section discusses the impact of key NM concepts on the organization of public sector services. However, before going on to address this question, three further issues are considered: the international scope of NM, the variant forms in which NM can appear, and the appropriateness of applying the adjective 'new' to NM.

NM has been embraced across a number of national jurisdictions but support for such programs has not been universal in the developed capitalist world. The principal comparative study by Pollitt and Bouckaert (2011) identified a group of 'leading' countries, Australia, New Zealand, the United Kingdom, and the United States, where support for NM-based concepts and practices were highly developed. It should also be noted with the respect to the United Kingdom that legislation on devolution of powers has meant that devolved governments in Scotland and Wales have considerable discretion over the organization of public sector social services and that this has resulted in NM organizational practices being much weaker in these devolved jurisdictions (Hood, 2007). In contrast to the 'leading' jurisdictions, NM has been much less significant in the major continental European economies such as France and Germany (Pollitt and Bouckaert, 2011 identify other countries where parts of the NM program have been adopted and others rejected). In the discussion in the second and third sections, these 'leading' NM countries will be used to give examples both of the application of NM concepts and the problems engendered by such applications.

As NM is composed of a set of elements, it is possible to find either variants in different 'leading' NM countries (Pollitt and Bouckaert, 2011) or changes in the emphasis on different aspects with changes of government, even though there is broad continuity in the adherence to NM concepts and practices (see, Duncan and Chapman, 2010 for such changes in the context of different administrations in New Zealand). An example is the extent and character of competition in public sector services. At one extreme, the role of the state can be limited to financing services and overseeing provider performance. In these, state service provision is ended and providers are drawn from the commercial ('for profit') or voluntary ('third') sectors. Such radical ventures into 'outsourcing' can also include commissioning roles for such case or care management as well as service provision for at-risk groups like children in the US state Kansas (Johnston and Romzek, 2008). In contrast, where competition is used in a 'mixed economy,' state provision is retained and incumbents are expected to compete against other public sector providers, for profit and/or third sector providers. Changes in the extent of public sector competition have varied with party control of central government. In Australia, both the Labor administration of Paul Keating and the Liberal–National administrations of John Howard exhibited a broad adherence to NM precepts. However, under Keating, the Australian employment assistance service operated with public, for profit and third sector providers, whereas Howard's administrations ended public provision in this area (Considine et al., 2011).

The general consensus in the academic literature is that NM emerged as a major public sector reform during the early 1980s (Cutler, 2007). The use of the adjective 'new' when referring to NM can broadly be treated in two ways. One is that NM is distinguished from earlier approaches to managing public sector services by the extent to which it has either invented or applied a 'new' set of techniques. This view is difficult to sustain. The logic of this position is that in the 'pre-NM' era techniques of performance measurement and management (like performance-related pay) were largely unknown in the public sector. However, there was a major performance-related pay scheme operating in elementary education in England and Scotland in the later half of the nineteenth century. This scheme linked student performance with the school budget and hence with teachers' pay (see Cutler, 2007 who gives other significant pre-NM cases of performance management in the public sector). Equally, as in the use of performance management techniques under NM, similar schemes were 'exported' to various Australian states. Arguably, the genuine claim for NM to be considered 'new' rests on the way the techniques have been applied as a set of 'doctrinal components' (Hood, 1991) across a range of services.

Thus, NM deploys a set of techniques designed to both improve the performance of public services and render providers of public services more 'accountable.' It is not a universal phenomenon. There are 'leading' countries where various forms of the broad program have been adopted, but in other major capitalist economies, its impact has been minimal. It is 'new' in the sense that, particularly in the 'leading' countries, a systematic attempt has been made to apply NM concepts across a broad range of public sector services, but individual techniques recommended by advocates of NM have a long history predating the past three decades when NM emerged as a major public sector reform program. In the next section, the argument turns to the impact of NM concepts on the organization of public sector services.

NM and the Organization of Public Sector Social Services

Measuring and Managing Performance

The emphasis in NM on measuring performance suggests that 'managing by numbers' in public sector organizations is highly desirable. 'Managing by numbers' can be classified under three principal headings: the use of 'targets'; the creation of 'ranking systems'; and 'intelligence systems' (Hood, 2007).

Targets are designed to measure performance against "one or more specified aspirational standards" (Hood, 2007: 95). An area in which targets have been frequently used is that of waiting lists and waiting times. In a number of public services, service users may be put on a 'waiting list' and the time they are waiting recorded. Instances include those waiting for an assessment by a social service department to see if they can receive a set of services, patients waiting for 'elective' (nonemergency) treatment in a hospital or waiting in a hospital accident and emergency department. The 'aspirational standard' could be to cut numbers on a waiting list to a given level or to set a maximum waiting time within which service users (or potential service users) should be assessed or receive a service. Instances of such practices include maximum

waiting time targets for elective hospital treatment, which were introduced in the English National Health Service (NHS) (see Bevan and Hood, 2006; Scotland and Wales did not have comparable targets); and times by which children at risk in Kansas would be expected to have permanent placements (Johnston and Romzek, 2008).

The logic of disaggregation means that the performance of organizations providing services can be assessed against such targets and where the target is not attained (e.g., maximum waiting times are exceeded) sanctions can be applied (Bevan and Hood, 2006). In the English NHS, for example, disaggregation has led to the creation of organizational units known as 'Trusts,' which are distinct accounting entities with accounts modeled on the form of the corporate sector and whose performance is measured. Some of these Trusts have what is termed 'Foundation' status, which means that they have more operational freedoms. Such freedoms are, however, conditional and 'Foundation Trusts' are currently regulated by a government appointed body called 'Monitor.' Like other Trusts, the performance against targets of 'Foundation Trusts' are monitored and where performance is seen as unsatisfactory one of the sanctions that Monitor can deploy is to demand an action plan from the relevant Trust management designed to correct the identified performance failure, which must be approved by Monitor.

'Ranking' systems are designed to measure the performance of 'comparable service units against one another' (Hood, 2007: 95) using a set of performance indicators. As the term suggests, performance data can be arranged in provider 'league tables' either constructed by a government agency or adapted from published government data. In the latter case, such league tables may be published in the media or incorporated in guides to the putative 'best' schools or universities among others. Publication is seen as a means to 'incentivize' improved performance, and (see below) published comparative performance data can also play a role in the extension of competition in the public sector.

The third principal form of 'managing by numbers' is 'intelligence systems.' However, in this form, the relevant 'numbers' are seen as providing 'background information,' which is subject to 'no fixed interpretation.' Such intelligence systems have a long history because the publication of data on numbers in public residential homes, patients treated in hospitals, or students attending schools or passing examinations have long been a staple feature of reports of government departments. While this approach is not absent in NM, it has played a relatively secondary role because intelligence systems provide no direct evidence as to whether 'results' are being delivered. In contrast both targets and ranking systems are designed to have a 'fixed interpretation.' Meeting the target or exceeding it (making assessments or treating patients well inside the target) is interpreted as 'success'; not meeting it, as 'failure.' The organization further up the rank order is regarded as more successful than the organization lower down and an organization consistently at the top of a rank order can be treated as a 'benchmark' for other organizations.

Increasing Competition in the Public Sector

The extension of competition in the public sector can take a variety of specific organizational forms. Insofar as such markets remain wholly or largely state financed, they have distinctive features from private sector markets (though it should be noted that where services such as railways are 'privatized' substantial public subsidies frequently operate). The 'market' aspect derives particularly from the separation of purchaser and provider roles (Duncan and Chapman, 2010). There are a number of variant forms of such markets. Various mixes of providers can operate in public sector markets. Purchasing structures can also vary considerably. Purchasing/commissioning can be devolved to professionals (case managers in social care, general practitioners in medicine). Education markets in contrast frequently operate by funding schools on age-weighted student numbers (Greener, 2009). The age weighting is designed to reflect the cost differences in providing education to students of different ages. In such a market, parents have a form of purchasing role, although supply constraints are likely to mean that while they can register a 'preference' for a given school, that school may be oversubscribed (Greener, 2009). In social care, there has been an increasing trend to 'individual budgets' where the service user is allocated a budget and becomes a direct purchaser of services (Wilberforce et al., 2012). Purchasing can thus vary from 'proxy' purchasing where the service is purchased on behalf of the service user to more direct roles in purchasing by service users.

Increased competition and performance management and measurement can operate in both complementary and competing ways. They are designed to be complementary insofar as performance data informs purchasing. Providers with a successful performance profile (shorter waiting times, 'better' examination results) could expect more 'business,' and vice versa. However, they can be competing in the sense that improved performance can be pursued either via performance management or through competition. Where the emphasis is on performance management, a key mechanism is for a central government department or a regulatory body to intervene where the performance of the provider unit is perceived as substandard. Where the emphasis is on competition, the stress is more on the pressure exerted on the provider unit by the actual or potential loss of revenue through desertion by purchasers. Even, however, where there is a preference for market mechanisms, this is not unproblematic since the obligation on government to provide locally accessible education, health, and social services can mean that allowing a public organization to 'exit' the market is limited (Greener, 2009).

Approaches to 'Managing' in the Public Sector under NM

The changing approach to 'managing' in the public sector has also led to an important set of structural changes. Disaggregation has led to the creation of new organizations. The most graphic example of this development has been the pattern of 'civil service reform' leading to creation of 'agencies' that are distinct from 'strategic' central government departments and have their own accounts and performance management regime (Duncan and Chapman, 2010). However, disaggregation does not necessarily involve the literal creation of new organizations. In education markets, for example, a given school might retain its name and be situated in the same place. Funding by age-weighted student numbers would, however, transform the

financial basis of the school by making the ability to attract students central to determining the budget. Where performance testing plays a salient role, the school may be judged on student performance in national or local examinations and tests. In both respects, the school can be seen as transformed into a quasi 'business' unit.

The logic of disaggregation has the important effect that provider units are subject to a performance management regime. The strategic/operational division means that schools, hospitals, and social services departments are to deliver services in line with performance objectives and their role in determining what these objectives will be is at best very limited (see Considine et al., 2011 for a discussion of some of the effects of such 'top-down' performance management).

Thus, NM concepts have important implications for the organization of public sector services. Provider units are usually subject to performance management regimes, which are designed to give incentives for 'good' and sanction 'poor' performance. Increased competition can take a variety of forms, some of which may result in reducing service provision by the public sector to a minimum. Disaggregation has led to the attempt to distinguish public sector organizations that play a strategic role from those that play a delivery role. The next section considers some of the principal criticisms leveled at NM.

The Critique of NM

Performance Measurement and Management

The principal criticisms of the NM approach to performance measurement and management can be broadly classified under three headings: the problem of 'proxy' measures, problems related to the attribution of responsibility for performance to particular organizational units, and the problem of 'gaming.'

The problem posed by 'proxy' measures arises because they may fail to reflect the objectives of a service. One broad objective of education, for example, could be the development of a capacity for critical thinking. However, educational performance measures usually focus on examination and test results. It is possible, however, to 'teach to the test,' i.e., coach students so they do well in particular tests and/or examinations. Such 'rote learning' can be contrary to a conception of critical thinking in which the student is able to transfer independent analytic skills to a number of contexts (for a discussion of both contemporary and 'pre-NM' examples of this problem, see Cutler, 2007).

Problems of allocating responsibility for performance come in a variety of forms. The notion that organizations designed to deal with service delivery/policy implementation can be regarded as 'stand-alone' bodies responsible for their performance is difficult to reconcile with interorganizational relationships in the public sector. An interesting example of the problems for performance management generated by such interactions is provided by Johnston and Romzek's (2008) study of services for children at risk in Kansas. Service providers are subject to a performance target for the time taken to place children whose cases they are managing. However, such placements are subject to the approval of a judge meaning that the service provider cannot control whether or not they achieve the target.

A different type of attribution problem arises with respect to ranking systems. These are often based on a composite set of indicators covering a wide range of services. Such rankings are subject, inter alia, to 'categorization' errors where "perplexity about how to fit cases into categories may result in faulty assignment of these cases" (Hood, 2007: 100). The nature of such errors can be understood if the mechanisms of such ranking systems are examined. A first stage in the process is the transformation of data on performance into a score for a given service. This in itself is a complex process because it is likely to require a large number of quantitative indicators (to reflect multiple dimensions of the services) and qualitative material such as inspection reports. A second stage is the generation of an organizational score that frequently will involve a service weighting. Such a weighting could be justified on the grounds of the relative importance of services (e.g., local government might regard social care as more important than leisure services) and/or on grounds of the relative expenditure on the services. Finally, the organization scores themselves can be classified into categories, e.g., a score in a certain range is 'poor,' another 'fair,' and another 'excellent.' The overall categorization might trigger actions like central government regarding a 'poor' local authority as requiring management changes.

Categorization error can occur at all stages. The first may involve implicit weightings of particular indicators and reports, but these may be difficult to justify. At the second stage, the principle of weighting might appear reasonable but specific weightings more difficult to justify; for example, should social care rate be twice as important as leisure services or three times? Finally, in terms of overall points scores, 30 points may be classed as 'fair' and 29 as 'poor' but there is no obvious justification for this, and in any case, these overall scores are themselves subject to categorization error.

The final key problem is posed by 'gaming.' As has been stressed, it is a characteristic of performance measurement and management under NM that there are powerful incentives and sanctions (Pollitt and Bouckaert, 2011). This engenders corresponding incentives to 'game' the system, i.e., adopt practices that can be seen as at variance with overall service objectives, while they enable the requisite performance objective to be met. A classic example of gaming occurs in educational performance measures that use a threshold. Such a performance measure would be the percentage of students achieving test or examination results within a particular range (e.g., grades A–C) (see Cutler, 2007 for an example from UK education performance indicators). A 'gaming' approach to such a threshold is for a school to concentrate resources on the group who are currently not expected to achieve the threshold but who, with sufficient additional help (e.g., extra tuition), might be expected to do so (the 'current Ds' but 'potential Cs'). This approach is designed to maximize the school's score on the relevant indicator within the resource constraints within which the school operates. The concentration of resources in this way means that two other groups are treated inequitably: those who are regarded as certain or highly likely to fall within the threshold because they do not need any additional help to get them over the threshold and those who are regarded as having virtually no chance of reaching the threshold. Arguably, this approach to maximizing the performance indicator score contradicts an implicit objective of the service, namely, to

provide, as far as possible, an appropriate education for individual pupils. The potential Cs are being favored not because they have any greater claim to educational resources but because favoring them is perceived as conducive to improving performance on the indicator (see Booher-Jennings, 2005 for such practices in the Texas education system).

Greater Competition

The principal problems with increasing the scope of competition in the public sector to be discussed in this subsection are the problems posed by managing public services through contracts, the impact of competition on individual service units and on overall service performance, and the impact of different conceptions of the service user/consumer of public services.

Insofar as maximum use is made of competition, this requires extending the range of providers used outside an internal market to encompass for profit and third sector organizations. However, this also requires drawing up contracts specifying not only expected costs of provision but also expected service 'outputs' and penalties for failure to achieve such outputs. The difficulty of doing this varies with the nature of the service provided. Where the service is relatively standardized (e.g., elective operations like hip replacements), such contractual terms may be relatively easy to specify. However, where services are more complex and expected outcomes less determinate (e.g., treatment for cancer), contracts are necessarily more problematic (Greener, 2009).

A related problem concerns the assumptions behind contracts in the context of contestability. If operations like hip replacements are treated as a 'quasi industrial' form of health care then it might be assumed that patients are relatively unconcerned regarding continuity of contract with the provider. The operation may be undertaken on a day surgery basis and continuity of subsequent care may be provided by a general practitioner not the surgical provider. However, in many instances of social care, such assumptions do not hold. Johnston and Romzek's (2008) study of services for at-risk children in Kansas found evidence that change or threat of a change in contract provider and case manager generated considerable unease in both children and their families. A state official (in a state that has completely outsourced this service) was quoted as saying that such contractual transitions were 'very tough on kids' and 'tough on families.'

An important example of the view that NM markets generate complementary benefits is the assumption that public sector markets extend 'choice' and that the exercise of choice drives up standards of service provision. Thus, Tony Blair argued that "choice puts the levers in the hands of parents and patients so that they as citizens and consumers can be a driving force for improvement in their public services" (cited in Clarke, 2006). What is assumed here is that there is a parallel gain at the individual and collective levels. However, it is incorrect to assume such a facile connection. An important illustration of attempts to link increased competition in the public sector with improvements in overall performance standards is the creation of new 'subsystems of schooling' (Morris, 2012), which compete for students with other schools in a national public school system. Examples of such subsystems are Charter Schools in the United States and 'free' schools in Sweden, which have been imported into English education by the current UK coalition government. The argument advanced by proponents of such increased competition is that it will improve both individual choice and overall school standards. However, such a view has been repudiated by the Organisation for Economic Cooperation and Development (OECD) (though the proxy problem should be noted here as the OECD discusses education standards predominantly in the context of test/examination data). The OECD concluded that while higher performance levels could be seen in many cases in these school subsystems, this was often accounted for by 'the higher socio-economic status' of students in such schools (cited in Morris, 2012: 102). The OECD went on to argue that 'cross-country analysis' suggested that school "systems as a whole do not benefit from higher rates of school competition" (cited in Morris, 2012: 102).

A final problem for public sector markets relates to the implications of changing and contradictory conceptions of the service user. In the passage quoted above, Tony Blair used the term 'consumer' but some supporters of public sector markets have wanted to replace public sector 'consumerism' with what they term 'personalization.' From a 'personalization' standpoint, the emphasis on offering 'choice' is too limited because it is choice between a 'menu' set by the service providers. Thus, a difficulty that can arise with choice is that it is between service providers who offer broadly the same service, so schools are often effectively competing to offer better measured academic performance, or hospitals, shorter waiting times (Cutler et al., 2007; see Considine et al., 2011 for increasing homogeneity of services in the Australian employment assistance service). In contrast, personalization requires 'coproduction' where the service user participates in the design of the service. There are clearly tensions between such a conception and the top-down approaches to performance management (Cutler et al., 2007). One of the conditions regarded as crucial by NM advocates of public sector markets is a plurality of providers as a condition of effective competition. However, there are potential tensions between personalization and the aim of having a plurality of potential service providers. For example, devolving budgets to individual users could throw up problems of providers having insufficient business to make them viable, having to incur higher service provision costs in the context of 'austerity'-driven reduced budgets, and having to bear higher transaction costs (Wilberforce et al., 2012).

The final set of critical issues regarding NM relates to the premises underlying the structural reorganization of public services driven by NM concepts. Here the distinction, underlying disaggregation of the public sector between organizations playing a strategic and those playing an operational/delivery role, will be considered. Disaggregation of the public sector was based on distinguishing strategic organizations like central government departments, which were responsible for setting policy and measuring the extent to which policy objectives were realized, and organizations concerned with the delivery of services and policy objectives. However, this presupposes that it is possible to distinguish clearly the effects of strategic and delivery-related decisions.

A major criticism of this view is that it is often unclear whether problems with the achievement of policy objectives were the responsibility of departments dealing with strategic

decisions or those charged with delivery/implementation. An instance of the problems of allocating responsibility in such a disaggregated structure is the issue of gaming. Gaming can be manifested in the form of shifting organizational resources from areas of organizational activity that are unmeasured or where a particular activity is not relevant to high-profile targets (for health, see Bevan and Hood, 2006). This, like the gaming education thresholds discussed above, can be regarded as generating inequitable practices. Users of, from a performance management point of view, low-profile health services or students regarded as either certain to get into the threshold range or having no chance of doing so suffer relative to service users who are 'relevant' to high-profile performance objectives (see Considine et al., 2011 for parallel practices in the Australian employment assistance service). In this sense, gaming represents a policy failure.

However, if this is a policy failure, it should be possible, in a disaggregated public sector organizational structure, to locate responsibility for this failure clearly. One immediate suggestion is that the 'fault' lies with the service delivery/policy implementation organizations. They are gaming the target/indicator and as a consequence responsible for the corresponding inequitable practices. Yet, a plausible case can be made that the organization responsible for strategy is at fault. The strategic organization was responsible for 'sending the message' that certain targets/indicators were much more significant. Equally, if such targets/indicators have the potential for gaming, why was this not controlled by the strategic organization (possible methods are discussed by Bevan and Hood, 2006). What is important here is that either account cannot be rejected unproblematically and that disaggregation does not locate responsibility for policy failures.

Conclusion

NM is based on a set of 'doctrinal components' (Hood, 1991) that have been applied to radically change the way in which government is organized. The effects of NM are diverse. It has greatly increased the measurement of the 'performance' of public sector organizations and spawned the application of incentives and sanctions to 'manage' performance. It has also encouraged much more extensive use of competition in the public sector. It has set in train a process of 'disaggregation' whereby organizations seen as specializing in 'strategic' decision making are separated from those involved in 'delivery' or policy implementation.

NM has operated on a contested terrain. Its advocates see it as a means of radically improving the performance of public sector organizations and bringing them into line with the demands of a more 'consumerist' society. The critics of NM perceive major problems in it. They argue that performance measurement and management has led to the use of crude performance measures and encouraged 'gaming,' which can compromise service standards. Increased competition in the public sector has been in tension with the desire of many service users for continuity of provision. Personalization, particularly in a putative 'age of austerity,' has been difficult to reconcile with encouraging a plurality of providers. Disaggregation, for critics of NM, is seen as offering a false hope of clear lines of accountability for policy success and failure. NM promised and promises a new era of a better managed public sector but 'delivering' on these pledges has proved problematic.

See also: Afrocentric Approaches to Social Work; Performance Appraisal and Evaluation; Public Choice Economics; Public Sector Organizations; Urban Neoliberalism.

Bibliography

Bevan, G., Hood, C., 2006. What's measured is what matters: targets and gaming in the English public health care system. Public Administration 84 (3), 517–538.

Booher-Jennings, J., 2005. Below the bubble, "educational triage" and the Texas accountability system. American Educational Research Journal 42 (2), 231–268.

Clarke, J., 2006. Consumers, clients or citizens? Politics, policy and practice in the reform of social care. European Societies 8 (3), 423–442.

Considine, M., Lewis, J., O'Sullivan, S., 2011. Quasi-markets and service delivery flexibility following a decade of employment assistance reform in Australia. Journal of Social Policy 40 (4), 811–833.

Cutler, T., 2007. A Necessary Complexity: History and Public Management Reform at: www.historyandpolicy.org (accessed 12.08.13.).

Cutler, T., Waine, B., Brehony, K., 2007. A new epoch of individualization? Problems with the 'personalization' of public sector services. Public Administration 85 (3), 847–855.

Duncan, G., Chapman, J., 2010. New millennium, new public management and the New Zealand model. Australian Journal of Public Administration 69 (3), 301–313.

Greener, I., 2009. Public Management: A Critical Text. Palgrave-Macmillan, Basingstoke, New York.

Hood, C., 1991. A public management for all seasons. Public Administration 69 (1), 3–19.

Hood, C., 2007. Public service management by numbers: why does it vary? Where has it come from? What are the gaps and the puzzles? Public Money and Management 27 (2), 95–102.

Johnston, J., Romzek, B., 2008. Social welfare contracts as networks: the impact of network stability on management and performance. Administration and Society 40 (2), 115–146.

Morris, P., 2012. Pick 'n' mix, select and project: policy borrowing and the quest for 'world class' schooling: an analysis of the 2010 schools white paper. Journal of Education Policy 27 (1), 89–107.

Pollitt, C., Bouckaert, G., 2011. Public Management Reform: A Comparative Analysis, third ed. Oxford University Press, Oxford.

Timmins, N., 2006. Back risk-taking civil servants, says PM. Financial Times, 19 July at: FT.Com (accessed 12.08.13.).

Wilberforce, M., Baxter, K., Glendinning, C., 2012. Efficiency, choice and control in social care commissioning. Public Money and Management 32 (4), 249–256.

New Media and Democracy in the Arab World

David M Faris, Roosevelt University, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

New media have had a particularly sharp impact on the Middle East, where states at the turn of the century remained dominated by authoritarian regimes and information monopolies. New media were used in the 2000s to circumvent censorship, to build communities of dissent and identity building, and to contribute in causally distinct ways to the wave of uprisings known as the Arab Spring. However, robust structures of authoritarian rule remain prevalent in the region, and regimes are increasingly savvy in leveraging new media tools to their own advantage.

What has been the impact of new media on political authoritarianism in the Middle East? Have emergent digital technologies overturned media hierarchies, challenged authoritarian rule, and transformed patterns of social mobilization in the region? Or have new media, as other scholars argue had little effect on entrenched patterns of authoritarian rule? The effort to theorize and analyze the impact of digital media on the Middle East has created a lively and sometimes contentious debate between scholars who study the region. New media in the Middle East have generated such sustained attention that there are now multiple academic journals devoted solely to regional mass media. In addition, articles about the impact of new media in the Middle East have appeared in dozens of communications, sociology, anthropology, and political science journals, as well as have been the focus of both single-author and edited volumes. This long-term scholarly effort has resulted in distinctive subfield whose signature feature is that it cuts across multiple disciplines, whose practitioners often do not read or cite one another's work.

First, a caveat: There is no longer anything 'new' about satellite television. Al-Jazeera is nearly 20 years old and the recipient of sustained scholarly attention that almost constitutes its own separate academic literature. Most major satellite broadcasters, including Al-Arabiya, the Lebanese Broadcasting Corporation and MBC, have been operating since the 1990s. This is not to suggest that satellite television is not an important subject of inquiry, but rather that patterns of consumption and use of television are so markedly different from those of the Internet and mobile phones as to necessitate a separate inquiry. It may not be long before digital media are also not 'new media' but for the time being the Internet and mobile phones can still be referred to as part of this general construct. This article will, therefore, confine itself to reviewing studies on the affects of the Internet and wireless telephony on society and politics in the Middle East. As these technologies 'took off' in the early 2000s for most societies of the region, the focus of this article will be on the period between 2000 and 2013. For the remainder of this article, new media can be assumed to refer both to the Internet, as well as to the mobile telephony platforms that are increasingly used to access the Internet.

In 2000, Jonathan Alterman argued that 'new media' might be received or used quite differently in the Middle East, suggesting that the Internet was unlikely to transform media consumption patterns in the region in the same way that they had done in other regions. He cautioned that "we need to think about the possibility that the Internet, which has transformed so many transactions and communications in North America and Western Europe, will not have as dynamic an effect on the Arab world" (Alterman, 2000: p. 356). The critique that new media's impact or growth curve in the Middle East would be markedly different from its development in the West has not been borne out by the unfolding of history since Alterman's article was published. If anything, the growth of Internet penetration in the region has been even more explosive than it was in the United States and Western Europe in the 1990s, and the region's citizens have used the new medium in more or less precisely the same way that it has been employed elsewhere. One can contest the relative importance of new media diffusion vis-à-vis other social and political variables in the region, but it would be difficult to argue that their impact has not been dynamic. This article focuses on a small number of new media impacts in the Middle East: first, on the practice of authoritarianism, second, on collective action, and third, on identity. The study is limited in nature, since the literature on new media in the Middle East has become expansive enough to defy even a lengthy cross-field summary.

New Media and Regional Authoritarianism

The academic literature on the Middle East is replete with unjustified exceptionalism. But one area where the Middle East was incontrovertibly different from other regions at the beginning of the new millennium was in the area of political rights and representation. At the turn of the new millennium, the Middle East was one of the last global bastions of entrenched authoritarian rule (Bellin, 2004), where there was one consolidated democracy (Israel excluding the Occupied Territories), one transitional democracy (Turkey), and a host of states that varied from hybrid regimes (Lebanon) to fully closed and brutal autocracies (Syria, Libya, and Iraq). Most of these regimes also exercised substantial or complete control over local media environments, to the extent that in many regional states, the constrained critiques emanating from Al-Jazeera represented the most radical dissent that most ordinary citizens had access to. In this constrained environment, the diffusion of self-publishing technologies like Blogger, as well as access to the largely unfettered (at least at first) global

 http://dx.doi.org/10.1016/B978-0-08-097086-8.95085-2

Internet, was genuinely disruptive in ways that helped lead to the Arab uprisings of 2010–11. Self-publishing technologies allowed individuals across the region the ability to create their own platforms of self-expression in the form of blogs and web sites, and gave others with Internet access the ability to access this new information and to engage in spirited debate about ideas. While satellite TV similarly crossed boundaries and carried unprecedented criticism of regime practices in the late 1990s and early 2000s (Lynch, 2006; Sakr, 2002), it was a medium that lacked the Internet's critical many-to-many communication properties.

At least in the early days of the Internet, authoritarian regimes in the region were generally incapable of censoring or filtering the Internet, and thus for a brief period of time in places like Tunisia, and for longer periods of time in countries like Egypt, the Internet was providing a site of free-wheeling discussion, dissent, and sharing, even if only for the few citizens with access to the Internet in the early 2000s. It was not long before citizens using these self-publishing platforms began to make organized calls for reforms in state public policy, particularly in the blogging hubs of Egypt and Bahrain (Al-Malaky, 2007). Unlike the stodgy state media outlets of the tolerated print media universe, the Internet tended to disproportionately feature the edgy voices of younger citizens, who felt left out of the power structure and who argued that the Internet gave "voice to the voiceless" (Faris, 2013: p. 119). Together these activists were able to create a set of "frames," or "conceptual tools for both media creators and audiences to convey, interpret, and evaluate information" (Mosemghvdlishvili and Jansz, 2012: p. 484). Finally, like many media commentators locally and in the West, authoritarian regimes for a time failed to take the threat of digital activism seriously. Together these factors led to several causal paths for the use of digital media and political change in the region. First, digital media platforms reduced communication and coordination costs for political actors operating under conditions of repression and information scarcity (Shirky, 2008). Second, early adopters of new media platforms and applications in certain regional states formed small, loosely-linked networks of dissent that proved critical in advancing long-term critiques of regime policies whose effects were difficult to see at the time (Faris, 2013: pp. 49–84), especially in Egypt and Bahrain (Schleusener, 2007). Crucially, new media offer the opportunity to publish opinion and join conversations on blogs or social media like Facebook and Twitter to any literate person with an Internet connection, leading to what Manuel Castells has termed 'the rise of mass self-communication' (Castells, 2007). Third, the Internet provided a forum for women, minorities, and other marginalized publics to shape their identity and participate in alternative public spheres (Wheeler, 2006). Fourth, new media made it easier for individuals in authoritarian societies to reveal their true feelings about authoritarian rule, and to avoid the public 'preference falsification' that deprives individuals of knowledge about the intentions and beliefs of those around them (Kuran, 1991). The default public nature of new media led to the creation of sometimes unintentional 'communality' (Bimber et al., 2005) and facilitated processes of 'frame-alignment' and in some cases, of 'informational cascades' – sudden, widespread changes in popular beliefs and preferences about public issues (Lustick and Miodownik, 2005; Bikchandani et al., 1992). Finally, new media 'allow communication to take place across borders, and forms of international solidarity to spring up relatively quickly' (Aday et al., 2010: p. 24). Collectively, these factors had the effect of posing new and unexpected challenges to authoritarian rule in the Middle East during the 2000–13 period, including undermining state media sovereignty and control (Lamer, 2012) culminating in the 2010–11 Arab Spring but followed by a period of political retrenchment and turmoil that highlighted the limitations of new media as an agent of social change.

New Media and Identity

The affordances of new digital technologies were not just seized upon by the secular activists who received so much press in the West. Scholars working on new media in the Middle East have largely assumed from the beginning that they would be particularly useful for transnational identity-building and linkage. Eickelman and Anderson argued in 1999 that "a proliferation of media and means of communication multiplied the possibilities for creating communities and networks among them, dissolving prior barriers of space and distance and opening new grounds for interaction and mutual recognition" (p. 3). Like satellite TV, the Internet by its very nature effortlessly crosses political boundaries and links "diffuse individuals with common interests" who may lack any other way to coordinate their actions (Shirky, 2008). In the Middle East, there are several plausible beneficiaries of digital media's transnational scope and reach. First, the Internet affords diaspora communities the opportunity to influence, and to be influenced by, political and social trends in home communities, perhaps deepening democratic engagement for citizens who would otherwise be marginalized (Brinkerhoff, 2009). During the Egyptian uprising, for instance, diaspora Egyptians in major cities across the world, including London, New York, and Chicago, used the Internet to organize rallies in support of the revolutionaries. The Syrian exile community has been similarly active both online and off-line since the civil war began in March 2011. Perhaps no community has seen a greater impact from digital media than the stateless Kurds, spread across multiple Middle East and North African (MENA) states including Syria, Turkey, Iraq, and Iran. The Kurds are subject to perhaps the strongest centripetal forces on national identity in the whole region – unlike, for instance, even the Egyptian Copts, the Kurds do not have a national home, and have been subject to the attempts of state leaders to suppress or eradicate expressions of Kurdish nationality, especially in Turkey during the course of the twentieth century. Kurdish web sites produced in diaspora have thus become an important source of contention, and have served to reinforce national identity in exile and in the homeland, with some scholars arguing that digital media have "assumed the role of print media nowadays' in the theory of 'imagined communities" advanced by Benedict Anderson (Candan and Hunger, 2008: p. 145). The Internet may also play a role in enhancing notions of pan-Arab nationalism, or *qowmiyya* nationalism, although this potentiality has yet to be the focus of sustained social–scientific inquiry. Certainly, the Internet has proven to be

a forum of empowerment for the region's women, many of whom played critical activist roles in the Arab Spring. As Courtney Radsch has argued, "Throughout the region, women have taken to the streets in unprecedented numbers, translating digital advocacy and organization into physical mobilization and occupation of public spaces in a dialectic of online and offline activism that is particular to this era" (2012: p. 4). Even in conservative Saudi Arabia, half of all bloggers are women, evidence that the new medium offers opportunities for participation and contestation that are not available offline (Open Net Initiative 2009).

Perhaps the transnational group that has received the most attention is global Islam. In fact, from the beginning, Islamist groups across the region used the proliferation of self-publishing tools and social media applications to network with one another, spread their message, and push for political reform of authoritarian states. In Egypt, young members of the Muslim Brotherhood began publishing blogs during 2006–07, in some cases challenging the organization's official positions (Lynch, 2007). The Egyptian Brotherhood was also at the forefront of its own kind of public diplomacy, using the group's English-language Web site, *Ikhwanweb*, to challenge both Egyptian and Western perceptions of the group, to advocate for imprisoned senior Brotherhood figures, and to infiltrate and disseminate information about military trials of the group's leaders (Faris, 2013: pp. 132–143). More broadly, the growth of digital media affords nonelites the opportunity to participate in conversations about Islam and Islamic practices, and to debate *al-maslaha* (the good life) together in digital forums (Bunt, 2009). One of the effects is a kind of broadening of religious authority, where less traditionally learned men may rule on religious matters through digital interaction in what Bunt terms Cyber Islamic Environments (CIEs) (Bunt, 2009).

New Media and the Arab Spring

The role of digital media in the series of uprisings that shook the Middle East beginning in Tunisia in December 2010 has also been the subject of sustained scholarly inquiry. Observers seeking a direct causal role for digital media were probably frustrated, but social media and mobile phones played at least a supporting role in nearly every revolt. The importance of new media in the Arab Spring is evidence of the multiple ways that these technologies have become embedded in everyday life, and how they have been transformed from the technologies of a wealthy elite. It was in Egypt where the impact of new media was the clearest. The Egyptian blogging community had over time been transformed by the arrival of social media platforms like Facebook and Twitter. Beginning in about 2008, young, digital-savvy Egyptians had used the group-forming capability of Facebook to organize increasingly daring challenges to the Egyptian state, beginning with the 6 April 2008 general strike organized online by a Facebook group (Faris, 2008). What became known as the April 6th Movement was one of the major organizers of the initial 25 January 2011 demonstrations that led to the ouster of Egyptian strongman Hosni Mubarak. The April 6th pioneers were joined by a group called We Are All Khaled Said, the namesake of a murdered Alexandrian whose plight came to stand in for experiences of torture, brutality, and impunity shared by many Egyptians. Founded in June 2012, the Khaled Said group soon featured hundreds of thousands of members, and whose Facebook wall was the site of open plotting about the date and form of the initial protests. What some have called 'cyberactivists' were also crucial in organizing opposition to the rigged December 2010 parliamentary elections, around which contestation spilled over into January 2011 (Khamis and Vaughn, 2011). While there is limited data about what led Egyptians to the protests on 25 January, there is some evidence to suggest that significant numbers of Egyptians learned about the protests through Facebook and Twitter (Tufekci and Wilson, 2012).

The role of digital media was much more varied across the other Arab Spring cases. In Syria, rebels leveraged the information capabilities to, at times, outsmart the regime. As Lindsey argues, "Amateur video showed rebels across the region the location of Syrian Army Republican Guard convoys, air assets, and checkpoints" (2013). Syrians also used Facebook to coordinate the dates and times of the initial Friday protests during March–April 2011. In Tunisia, unlike in Egypt, digital activists did not spark the initial revolt, but rather exploited it through publicizing subsequent protests (Faris, 2012), and some scholars have argued that the Wikileaks scandal was particularly damaging for the Tunisian regime, as it helped expose some of the more outrageous government corruption. And in Yemen, where rates of Internet and mobile penetration were particularly low, it is difficult to locate any unique causal role for new media.

However, the aftermath of the Arab Spring has certainly provided substantial evidence for scholars skeptical that digital media can permanently transform authoritarian regimes. In Egypt, the democratic transition instigated by digital activists collapsed in 2012–13 into a military coup and a seemingly strengthened authoritarian regime. Attempts to use existing digital platforms, including 'We Are All Khaled Said,' as well as new permutations of digitally mediated activists groups, like Third Square, were unsuccessful. Despite continued increases in access to the Internet in Egypt, Tunisia, Bahrain, and Iran, activists were unable or unwilling to follow their Arab Spring organizing methods with sustained political efforts. In Syria, while digital platforms were used to coordinate and facilitate early protests during March–April 2011, and while digital media have been crucial in keeping the global community apprised of what is happening inside the country (Aman and Jayroe 2013: p. 331), digital media have clearly been eclipsed by physical events on the ground and the overwhelming violence and brutality of the combatants. Digital dissidents were also comprehensively routed in Bahrain, as well as in Iran in the chaotic period following the reelection of Mahmoud Ahmedinejad in 2009.

The difficult aftermath of the Arab Spring, and the different trajectories of uprisings in Arab states, are both supportive of the notion that digital media's impact depends seriously on the political environment in each case. Scholars have thus far had a difficult time empirically linking the spread of digital technologies with democratization. Of the studies to undertake quantitative analyses of this question, there has been substantial disagreement. Among the first was Calfano and Sahliyeh's analysis of Internet penetration and political rights, which

found that an increase in Internet access was affiliated with a decline in political rights across the region (Calfano and Sahliyeh, 2008: p. 71). It is certainly true that the region's most wealthy states, and thus the places with the highest rates of Internet penetration, were among the least likely to witness widespread social unrest in 2010–11 (Wolfsfeld et al., 2013). Wolfsfeld et al. argue that protest tends to precede digital organizing, rather than the other way around. While there were stirrings of protest in Saudi Arabia, the United Arab Emirates, and Kuwait, there certainly was nothing like the broad-based, digitally-inspired street movement in Egypt, or even the digitally-coordinated protests that emerged in Tunisia following the initial spark.

On the other hand, there were modest protests in Saudi Arabia, Kuwait, and Oman (Howard and Hussain, 2013: pp. 6–9) and regimes responded with conciliatory policies including direct financial grants to citizens (Kuwait) or promises of reform (women gaining the right to vote in Saudi Arabia). And wealthy, highly-connected Bahrain was the site of a major uprising, albeit one with sectarian overtones, that required Saudi forces and extreme brutality to be put down. Other work has found an empirical link between increased access to digital media technologies and political reform over the long term (Howard, 2010). With very few exceptions, the most vociferous critics of the link between digital media and collective action have been nonacademics whose treatment of evidence and methodology are questionable at best. The crux of skepticism revolves around the continued importance of satellite television, particularly that of Al-Jazeera (Alterman, 2011: pp. 103, 104), and of the political context of collective actions. Very few scholars would likely be willing to sign onto the notion that digital media are either necessary or sufficient for widespread protest activity; at the same time, the reality of digital diffusion, the increasing density of digital networks, and the growing embedding of digital media in the everyday practices of ordinary life mean that, for the time being, we are unlikely to witness widespread protest activities in the Middle East without a leading digital component. And because of the failure of most Arab Spring uprisings to lead to genuine democratization, digital technologies are likely to remain the default mode of organizing and expression, even where states undertake sincere and sustained efforts to disrupt networks, erect information chokepoints, and imprison activists.

Regimes Respond

States, however, have not taken these challenges lightly, and have now responded in a variety of ways to the challenge of digital dissent. Generally, the more politically repressive states, including Iran, Syria, and Saudi Arabia, were most determined and successful in designing strategies to contain digital activism. While not technically part of the Arab world, Iran has been a key regional leader in disrupting the democratizing potential of new media. In 2006, the regime went as far as slowing the available download speed for Internet users, in an attempt to prevent the spread of viral videos challenging state authority. Tunisia, on the other hand, adopted a generic regime of censorship and 'filtering,' using off-the-shelf technology to block access to political sites on the Internet. Egypt was an outlier, in that the Mubarak regime made no serious attempt to interfere with the smooth functioning of the Internet – restricting itself largely to piecemeal harassment and detention of digital activists (Boas and Kalathil, 2002; Freedom House 2011). As Freedom House noted of Egyptian authorities in 2011, "Rather than relying on technical content filtering or monitoring, they typically employ 'low-tech' methods, such as intimidation, legal harassment, detentions, and real-world surveillance of online dissidence."

And by the standards of the region, even Egyptian harassment of bloggers and activists was tame. Generally, though, even deliberate and well-funded censorship regimes were vulnerable to circumvention by committed activists, who in Tunisia were able to use proxy servers to access blocked content (Faris, 2012). The main effect of the censorship regimes employed by Iran, Tunisia, Syria, and others was to drive activists into exile, and to heighten the importance of information produced in other countries and then made available to users via circumvention software. While the idea of an Internet 'kill switch' has received a great deal of press attention, most attempts to actually shut down Internet service had been relatively brief until the Syrian civil war.

By mid-decade, most regional states had recognized the threat posed by the diffusion of the Internet to the tight control over information traditionally exercised by authoritarian regimes. In responding to these threats, regimes have access to an increasing repertoire of global censorship, filtering, harassment, and innovation pioneered by likeminded authoritarians in China and Russia (both at the forefront of authoritarian responses to digital activism). As Youmans and York note "Repressive regimes are demonstrating learning curves when it comes to limiting the usefulness of social media for activists" (Youmans and York, 2012: p. 323). Jiang termed authoritarian public spheres that nevertheless contain technological or legal limits on digital speech 'authoritarian deliberation' (2010). Iran and Saudi Arabia remain at the forefront of regional attempts to construct tight censorship regimes that can prevent ordinary citizens from accessing sensitive or politically damaging information. While the crisis surrounding the reelection of Mahmoud Ahmedinejad to the presidency in 2009 provided proof that new media, particularly Twitter, played a substantial role in mobilizing external audiences to support the country's besieged protestors (Aday et al., 2010), those activists ultimately failed. Iran famously shut down its mobile networks during the 2009 postelection crisis (Rahimi, 2013: p. 40), but the state's attempt to build a longer-term and more durable censorship regime is of greater importance, highlighted by the creation of parallel social media Web sites, such as 'Velayatmadaran' (a Facebook copycat) that the state can suitably police and censor (Rahimi, 2013: p. 41). Saudi Arabia was among the first global regimes to construct an elaborate and expensive filtering regime, licensing private operators to provide Internet service but channeling all access through the same hub (Open Net Initiative 2010). Ultimately though, even determined attempts to shutter networks during moments of crisis in Libya and Egypt failed to dissuade knowledgeable and dedicated activists from accessing proxy servers and other workarounds.

However, it was also the embattled MENA autocrats in the post–Arab Spring period who deployed, under duress, new methodologies for shutting off access to the Internet. In Egypt,

the regime shut down access to the Internet altogether shortly after the 25 January protests began (Lynch, 2012), although by all accounts this tactic backfired as it forced those who were following along online to go outside and see what was happening in the streets. In Bahrain, regime loyalists used Facebook and blogs to attack leading dissidents, and in Syria, pro-Assad groups spammed Twitter with #Syria hashtag posts and created a popular Facebook group called 'Syria Revolution 2011' that hosted regime supporters (MacKinnon 63–64). Perhaps no effort was more comprehensive than the creation of the regime-supported Syrian Electronic Army (SEA) which was designed to harass, discredit, block, or dismantle the new media efforts of the opposition.

Conclusion

New media have had a clear and substantial impact on the Middle East. By offering alternative platforms of expression and dissent, the early Internet led to the creation of distinct blogging and collaborative communities in places like Egypt, Bahrain, and Iran (Sreberny, 2009). Later developments in what came to be known as 'social media' provided activists with new coordinating, information-sharing, and group-forming tools that led to escalating challenges against regional authoritarian regimes. New media are now deeply embedded in the everyday practices of the region's citizens, and while rates of access and use still vary widely across the region, the upward trend is clear even in the poorest states. These disparate forms of media are likely to play important roles in democratizing states like Tunisia and Iraq as well as in states where authoritarianism has weathered the Arab uprisings, such as in Egypt, Bahrain, Iran, and Syria. The causal impact of new media in the Middle East has been the subject of vigorous, popular, and scholarly dispute, but what is clear is that these technologies have served as important causal variables in social and political contestation even as local contexts serve as important intervening variables that may lead to divergent political trajectories. Above all, new media are now entrenched in the Middle East, and the study of their impact will remain a topic of sustained interest to scholars, journalists, and policy makers for the foreseeable future.

See also: Agenda Setting, Media Effects on; Anomie; Attitude Formation and Change; Attitudes, Political: Authoritarianism and Tolerance; Collective Action; Collective Behavior, Sociology of; Community and Media; Democratic Transitions; Freedom of the Press; Group Processes, Social Psychology of; Internet and Social Media: Anthropological Aspects; Media and Social Movements; Political Protest and Civil Disobedience; Representations, Social Psychology of; Social Media; Social Protest and New Media.

Bibliography

Aday, S., Farrell, H., Lynch, M., Sides, J., Kelly, J., Zuckerman, E., 2010. Blogs and Bullets: New Media in Contentious Politics. United States Institute For Peace, Washington, DC.

Alterman, J., Summer 2000. Counting nodes and counting noses: understanding new media in the middle east. Middle East Journal 54 (3), 355–361.

Alterman, J., Fall 2011. The revolution will not be tweeted. The Washington Quarterly 34 (4), 103–116.

Al-Malaky, R., Spring 2007. Blogging for reform: the case of Egypt. Arab Media & Society 1.

Aman, M.M., Jayroe, T.J., 2013. ICT, social media and the Arab transition to democracy: from venting to acting. Digest of Middle East Studies 22 (2), 317–347.

Bellin, E., January 2004. The robustness of authoritarianism in the middle east: exceptionalism in comparative perspective. Comparative Politics 36 (2), 139–157.

Bikchandani, S., Hirshleifer, D., Welch, I., 1992. A theory of fads, fashion, custom and cultural change as informational cascades. Journal of Political Economy 100 (5), 991–1026.

Bimber, B., Flanagin, A.J., Stohl, C., November 2005. Reconceptualizing collective action in the contemporary media environment. Communication Theory 15 (4), 365–388.

Boas, T.C., Kalathil, S., 2002. Open Networks, Closed Regimes: The Impact of the Internet on Authoritarian Rule. Carnegie Endowment For International Peace.

Brinkerhoff, J., 2009. Digital Diasporas: Identity and Transnational Engagement. Cambridge University Press, Cambridge, UK.

Bunt, G., 2009. iMuslims: Rewiring the House of Islam. Hurst & Co, London.

Calfano, B., Sahliyeh, E., Spring 2008. Transmitting reform? Assessing new media influence on political rights in the middle east. Critique: Critical Middle Eastern Studies 17 (1), 63–77.

Candan, M., Hunger, U., 2008. Nation building online: a case study of Kurdish migrants in Germany. German Policy Studies 4 (4), 125–153.

Castells, M., 2007. Communication, power and counter-power in the network society. International Journal of Communication 1, 238–266.

Eickelman, D.F., Anderson, J.W., 1999. New Media in the Muslim World: The Emerging Public Sphere. Indiana University Press, Indianapolis, IN.

Egypt, 2011. Freedom on the Net. Freedom House.

Faris, D., Fall 2008. Revolutions without revolutionaries? Network theory, facebook and the Egyptian blogosphere. Arab Media & Society 6, 1–11.

Faris, D., 2012. Beyond social media revolutions: the Arab spring and the networked revolt. Politique étrangèrel 77 (1), pp. 99–109.

Faris, D., 2013. Dissent and Revolution in a Digital Age: Social Media, Blogging and Activism in Egypt. I.B. Tauris and Co., London.

Howard, P.N., 2010. The Digital Origins of Dictatorship and Democracy: Information Technology and Political Islam. Oxford University Press, Oxford, UK.

Howard, P.N., Hussain, M.M., 2013. Democracy's Fourth Wave? Digital Media and the Arab Spring. Oxford University Press, Oxford, UK.

Jiang, M., 2010. Authoritarian deliberation on the Chinese Internet. Electronic Journal of Communication 20.

Khamis, S., Vaughn, K., Summer 2011. Cyberactivism in the egyptian revolution: how civic engagement and citizen journalism tilted the balance. Arab Media & Society 13.

Kuran, T., October 1991. Now out of never: the element of surprise in the east European revolution of 1989. World Politics 44 (1), 7–48.

Lamer, W., Fall 2012. Twitter and tyrants: new media and its effect on sovereignty in the middle east. Arab Media & Society 16.

Lindsey, R.A., July 2013. What the Arab spring tells us about the future of social media in revolutionary movements. Small Wars Journal 9 (7). http://smallwarsjournal.com/jrnl/art/what-the-arab-spring-tells-us-about-the-future-of-social-media-in-revolutionary-movements (accessed 17.01.14.)

Lustick, I.S., Miodownik, D., 1 September 2005. Neighborhoods and tips: implications of spatiality for political cascades. In: Paper Presented at the Annual Meeting of the American Political Science Association, Washington DC, pp. 29–30.

Lynch, M., 2006. Voices of the New Arab Public: Iraq, Al-Jazeera and Middle East Politics Today. Columbia University Press, New York.

Lynch, M., Winter 2007. Young brothers in cyberspace. Middle East Report 37, 26–33.

Lynch, M., 2012. The Arab Uprising: The Unfinished Revolutions of the New Middle East. PublicAffairs, New York.

MacKinnon, R., 2012. Consent of the Networked: The Worldwide Struggle for Internet Freedom. Basic Books, New York.

Mosemghvdlishvili, L., Jansz, J., 2012. Framing and praising Allah on Youtube: exploring user-create videos about Islam and the motivations for producing them. New Media and Society 15 (4), 482–500.

Radsch, C.C., May 2012. Unveiling the Revolutionaries: Cyberactivism and the Role of Women in the Arab Uprisings. Rice University James A. Baker III Institute for Public Policy Research Paper, 1–41.

Rahimi, B., 2013. The state of digital exception: censorship and dissent in post-revolutionary Iran. In: Hussain, M.M., Howard, P.N. (Eds.), State Power 2.0: Authoritarian Entrenchment and Political Engagement Worldwide. Asghate Publishing, Burlington, VT.

Sakr, N., 2002. Satellite Realms: Transnational Television, Globalization and the Middle East. I.B. Tauris and Co., London.

Internet Filtering in Saudi Arabia, 2009. Open Net Initiative. Available at: https://opennet.net/sites/opennet.net/files/ONI_SaudiArabia_2009.pdf (accessed 07.01.14.).

Schleusener, L., Spring 2007. From blog to street: the Bahraini public sphere in transition. Arab Media & Society 1.

Shirky, C., 2008. Here Comes Everybody: The Power of Organizing without Organizations. Penguin Books, New York.

Sreberny, A., 2009. Blogistan: The Internet and Politics in Iran. I.B. Tauris and Co., London.

Tufekci, Z., Wilson, C., 2012. Social media and the decision to participate in political protest: observations from Tahrir square. Journal of Communication 62, 363–379.

Wheeler, D.L., July 2006. Empowering Publics: Information Technology and Democratization in the Arab World: Lessons from Internet Cafes and Beyond. Oxford Internet Institute, Research Report No. 11, pp. 2–18.

Wolfsfeld, G., Segev, E., Sheafer, T., April 2013. Social media in the Arab spring: politics comes first. The International Journal of Press/Politics 18 (2), 115–137.

Youmans, W.L., York, J.C., 2012. Social media and the activist toolkit: user agreements, corporate interests, and the information infrastructure of modern social movements. Journal of Communication 62, 315–329.

Relevant Websites

http://www.arab-reform.net/ – Arab Reform Initiative.
http://cyber.law.harvard.edu/ – Berkman Center for Internet and Society.
http://www.freedomhouse.org – Freedom House.
https://opennet.net/ – The Open Net Initiative.
http://liberationtechnology.stanford.edu/ – Program on Liberation Technology.

New Media and Social Capital

Gustavo S Mesch, University of Haifa, Haifa, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The rapid increase in the adoption and use of information and communication technologies among the population of developed and less-developed countries has prompted researchers to investigate the social impact of these technologies on patterns of interpersonal communication; access to educational, housing, and job opportunities; and the attainment of social status. In the study of the social impact of information and communication technologies, the theory of social capital has served as an organizing framework. The purpose of this article is to expose the reader to the essential concepts and empirical findings that have guided this research. The focus of this article is on studies that have used social capital or some of its dimensions as the dependent variable of the study.

Social media have become a varied and pervasive platform for user-generated content. The use of such vehicles has implications for one's social capital, both on the individual level and on the collective level. This article reviews the early and more current literature about the relationship between the use of social media and the attainment of social capital for both the individual and the community at large. The findings also have implications for future studies in this area.

Social Media

Social media is a concept that involves the use of web and mobile-based technologies to facilitate the mass sharing of user-generated content and the maintenance of interactive communication with other users. In that sense, it is a concept that describes a wide variety of software enabling individuals, organizations, and communities to gather, communicate, create, and share multimedia content, collaborate with one another, and play together. Given its purpose, social media include six different types of tools: media for collaborative projects (e.g., Wikipedia), blogs and microblogs (e.g., Twitter), content communities (e.g., YouTube), social networking sites (e.g., Facebook), virtual game worlds (e.g., World of Warcraft), and virtual worlds (e.g., Second Life) (Kaplan and Haenlein, 2010). Social media is accessible to individuals, organizations, and communities, allowing them to easily produce and change content, and to reach large audiences in real time.

Social networks have played a critical role in the adoption and expansion of social media. Most of the applications require one to be invited to participate and/or to invite one's acquaintances, family, and friends to join and become linked to the user. Bringing along a network of acquaintances is a condition for having access to the content of others and for sharing one's own content with others. Thus, the worldwide adoption of these applications had led social media to play an important role in social and political mobilization, advertising, and brand building, and interpersonal and mass communication. The dissemination of content generated through these formats and their interactive characteristics have prompted researchers to try to understand the role of social media in the production and maintenance of social capital.

Social Capital

The idea behind social capital is that involvement and participation in groups can have positive consequences and benefits for individuals and communities. Examples of such benefits include access to information, emotional support, exposure to different ideas, and social and political deliberation, and engagement. Scholars regard social capital as being embedded in the structure of social relationships. Given that social applications on the Internet are used to access information, maintain existing ties and create new relationships, it is important to study the implications of the use of social media for the acquisition of social capital.

Social capital is complex, especially because researchers approach it from various disciplines and apply the concept to a wide variety of social phenomena. For these reasons, there are significant variations in the measurement, sources, and outcomes of the concept. The studies of Bourdieu (1986), Coleman (1990) and Putnam (1993) are the most cited sources in contemporary discussions on social capital. Bourdieu (1986) defined social capital as "the aggregate of the actual and potential resources which are linked to possession of a durable network of more or less institutionalized relationships of mutual acquaintance or recognition" (p. 248). For Coleman (1990), "social capital is defined by its function. It is not a single entity, but a variety of different entities having two characteristics in common: they all consist of some aspect of social structure, and they facilitate certain actions of individuals within the structure" (p. 302). According to Coleman, social capital can take on forms such as obligations and expectations, information potential, norms, and effective sanctions.

For Putnam (1993), social capital is a community level resource. He defines social capital as "… features of social organization such as networks, norms and social trust that facilitate certain actions of individuals who are within the structure" (p. 67). Putnam's argument is that social networks and social norms are critical for social cooperation, coordination, and collaboration. For this reason, social capital has important consequences for democracy, as civic engagement and participation are a measure and consequence of increasing or decreasing levels of social capital (Putnam, 1993).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.95087-6

From this review, we may infer that the most widely used dimensions of social capital are social networks, trust, and norms of reciprocity. Social networks concern the extent of individual participation in various types of organizations, communities, and informal networks. Trust is defined as the level of confidence that people will act or are expected to act as they say they will, or that what they say is reliable. Norms of reciprocity refers to shared understandings, and informal rules and conventions about continuing relationships of exchange at any given time. These norms involve mutual expectations that assistance given now will be repaid in the future (Putnam, 1993). This view of social capital stresses on aspects of the connections people create through civic engagements, community, and online-based activities.

Linking Social Media and Social Capital

Various disciplines have undertaken the study of the association between information and communication technologies and social capital. In general, these studies have focused on the effects on the individual level or on the collective, community level.

On the individual level, early studies were concerned with the extent to which frequency of Internet use and different types of use might have negative effects on social involvement with family and friends. Such studies focused on the well-being of Internet users. These studies were based on the social displacement hypothesis, which argues that time spent online displaces everyday social interaction with family and friends, and that social isolation has negative implications for users' psychological well-being. Early studies of Internet users' time online presented some evidence supporting this argument (Kraut et al., 1998; Nie, 2001). However, subsequent investigations showed that the effect appears to differ according to personality orientation and initial levels of social resources such as perceived social support, network size, and extraversion (Katz and Rice, 2002; Wellman et al., 2001; Bessière et al., 2008). A 'rich get richer' hypothesis explained these differential effects, indicating that very sociable individuals embedded in extensive networks of social support derive more social benefit from using the Internet. Very sociable people reach out to others on the Internet and use it particularly for communication. In contrast, individuals with lower levels of social support and more circumscribed social networks turn to the Internet to compensate for their lack of social resources. However, they do not reap the same level of benefits from online interpersonal communication and continue to suffer from social isolation and a poor sense of well-being (Bessière et al., 2008).

Later investigations failed to replicate the earlier findings. Indeed, studies have found no evidence that Internet use reduces social involvement (Rice and Katz, 2000; Wellman et al., 2001). Furthermore, some studies even suggested that Internet users tend to report more social involvement than non-Internet users do. Thus, several critiques have been leveled against the displacement argument (Gershuny, 2003). One criticism notes that Internet use does not necessarily come at the expense of other activities. This position rejects the idea that time is fixed. Second, the Internet facilitates information searching, communication, game playing, file downloading, and other activities, providing a technological infrastructure for a more efficient way of conducting these activities and thereby freeing up time for more social involvement. According to this view, individuals use these communication tools primarily to reinforce existing relationships with family, friends, and romantic partners. They are increasingly integrating these tools into their off-line worlds, for example, by using social networking sites to get more information about new entrants into their off-line world (Subrahmanyam et al., 2008).

Another criticism rejects the assumption that Internet use is a solitary activity, conducted alone and without the involvement of others. In fact, according to this view most Internet use is social. Rather than isolating the user, the Internet facilitates interpersonal contact, so that users may actually be more socially involved than others (Wellman et al., 2001). Finally, most of the early studies used only a single item to measure ICT use, namely, self-reported frequency of use. This single measure does not account for different uses of information and communication technologies. While some of these uses such as online gaming may be conducted in isolation, others allow one to connect and reconnect with acquaintances, family, and friends, expanding one's social circle. More recent studies have focused on the use of the Internet for information seeking, for entertainment, and for communicating with family and friends. These studies report an association between one's initial social resources, the type of Internet use, and the social outcomes (Shah et al., 2001). Individuals reporting high levels of social support and large social networks were more likely to use the Internet for communicating with family and friends. In contrast, individuals reporting lower levels of social support were more likely to use the Internet for entertainment purposes and for meeting with unknown others. All of the respondents used the Internet quite frequently to find information, and there was no association between this activity and previous levels of social support (Bessière et al., 2008). Studies also report an association between the types of Internet use and an individual's well-being. Using the Internet for entertainment was correlated with higher levels of perceived depression, whereas using the Internet for communicating with friends and family was correlated with lower levels of depression. Using the Internet for information search was uncorrelated with depression. These findings highlight the fact that the Internet is a multipurpose technology with a wide range of uses. Furthermore, one's initial social resources shape one's reasons for using the Internet. These various uses, in turn, may have different effects on one's well-being, social engagement, and involvement in social activities (Bessière et al., 2008).

Involvement in social networks is one of the central dimensions of social capital. Studies have established an association between social media use and network size and heterogeneity. Social media use is also associated with social capital because formal and informal conversations are important for exposure to nonredundant information, ideas, and opinions (Hampton et al., 2010; Hargittai and Hsie, 2010. Regarding the size of core networks, studies have found that using different social media applications such as instant messaging, social networking sites, and weblogs is not associated with a decrease in the size of one's core networks (Mazur and Richards, 2011). Furthermore, a study based on a large

representative sample of the US population found some indications that regardless of the application being used Internet users did not report larger networks, but were more likely to report greater heterogeneity in their core discussion networks (Hampton et al., 2010). Network heterogeneity is associated with the extent of overlap of the origins of one's social ties, whether online or off-line. The networks in applications such as instant messenger and social networking sites generally tend to be homogeneous and reflect off-line networks. In a study of preadolescents that compared the networks of 'friends' in social networking sites, instant messaging, and face-to-face communication, the results showed that there was a partial but substantial overlap between participants' online and off-line networks. Half of the closest instant messaging friends' names were also listed as the respondents' closest face-to-face friends, and 49% of the respondents' face-to-face friends were also the top social networking site friends. Thus, there is a large overlap between off-line and online worlds (Subramanyam et al., 2008).

Another study of adolescents (ages 16–19 years) used interactions in the public space of a social networking site ('the wall') as the unit of analysis and investigated whether the characteristics of online 'friends' were similar to or different from the characteristics of the young social networkers themselves. The study found that on average more than half of the interactions were with others who belonged to the same ethnic group, and 42% were with others of the same gender. In addition, the majority of the interactions reported were with similar others in terms of age category and place of residence (Mazur and Richards, 2011).

Other studies have investigated the extent to which social media is used to support existing ties or develop new social ties. In a study about the use of instant messaging, researchers found that it is generally used to support strong ties rather than weak ties (Mesch et al., 2012). In contrast, chat rooms and weblogs are often used for communication with those with whom one has weak ties rather than strong ties (Hargittai and Hsie, 2010).

Furthermore, more intense users of social media engage in more social activities on social networking sites than those who spend less time on such sites and use only one site. There are differences according to gender, as women pursue stronger tie activities than do men, such as interacting with existing friends. However, women engage in fewer weaker tie activities than men do, such as developing new relationships via these sites (Hargittai and Hsie, 2010).

In sum, the empirical evidence indicates that social media is a complementary communication channel that is used with existing social ties. Individuals tend to bring their existing connections to the social networking sites. The level of one's social resources is an antecedent of social media use. Those with high levels of social support and social involvement are the ones who gain the most from the use of social media. The use of social media may expand one's network size and increase its heterogeneity, but only modestly.

Collective Level Studies

On the collective level, research has focused on the use of social media use, civic engagement and political participation. Putnam (1993: p. 90) argued that "participation in civic organizations inculcates skills of cooperation as well as a sense of shared responsibility for collective endeavors." Indeed, early studies investigated the role of the frequency of use of social media on civic engagement and community involvement, with some of these studies reporting a negative relationship between the two (Kraut et al., 1998; Nie, 2001).

Nevertheless, there are grounds for assuming that different types of Internet use may have different effects on network diversity, interpersonal trust, and civic engagement. Information searching may promote civic engagement, because through such searches users gain knowledge about social and community issues, build connections, and coordinate their actions to address mutual concerns. Thus, searching for information online exposes the user to opportunities for participation in civic life, and may encourage citizens to participate in more and find more opportunities to contribute to society.

At the same time, using the Internet for entertainment does not provide the benefits of acquiring new information and making connections with new contacts for collective action. Online recreation consumes time and energy without providing opportunities to become involved in community or civic activities and may even weaken one's links to existing social ties. Various studies suggest that civic and social engagement is dependent on the type of Internet use. While informational use is positively associated with civic engagement, defined as participation in civic and community affairs, Internet use for social recreation has a small negative effect on interpersonal trust and does not affect civic engagement (Shah et al., 2001). These findings were also confirmed in longitudinal studies that compared the level of civic engagement of a 1965 high school class cohort before and after the adoption of the Internet. The study found that Internet access and its use to follow public affairs was positively associated with indicators of political involvement. It also established a strong relationship between Internet use and membership in voluntary organizations. It is very likely that early adopters of the Internet had an a priori tendency to be involved in civic and political affairs. Nevertheless, even controlling for previous civic involvement, the findings support the idea that Internet use for information access adds to the indicators of civic engagement and membership in voluntary associations (Jennings and Zeitner, 2003).

Using a large sample of the American population, a study that investigated the association between different social media applications and civic engagement in social and voluntary associations found that the effects differed according to the application. The use of weblogs increased participation in voluntary groups, participation in picture sharing social network sites, and increased participation in public spaces, and the use of social networking sites increased participation in locally based activities (Hampton et al., 2010).

One explanation for the variation in effects may be the fact that different online platforms fulfill different social functions. For example, weblogs are places of interaction in which users maintain their anonymity by using nicknames and are more likely to interact with others who might not be known to them and may reside in a different city, state, or country. One study found a positive association between writing a weblog and having a more diversified personal network (Marlow, 2005).

Given that the context of the weblog interaction is based on mutual, emotional, or intellectual interests, much of the interaction is with unknown individuals who share concerns, hobbies, or other interests (Subrahmanyam et al., 2008). In contrast, instant messaging provides text-based, near-synchronous communication between two or more users who are usually known to one another. In addition to supporting peer interaction, instant messaging is used to support communication with family and romantic partners. Indeed, the use of instant messaging for meeting new contacts is unusual (Mesch et al., 2012).

Social networking sites allow users to present information about themselves (such as age, gender, location, education, and interests), encourage users to link to known and likeminded users whose profile exists on the site, invite known and like-minded individuals to join the site, establish and maintain contact with other users, post content, create a personal blog, and participate in online groups (Boyd and Ellison, 2007). Social networking sites are used for communication between individuals who have some knowledge of one another and belong to the same social circle or to the social circle of their friends (Steinfeld et al., 2008).

Steinfield et al. (2008) found that the intensity of Facebook use was associated with changes in users' social capital. Among college students scoring higher on the scale was related to increases in one's social capital. In addition, social networking sites may act as mediators between dispersed groups. Social networking sites facilitate social awareness and social interaction, and allow the maintenance of ties that otherwise might be latent or dormant. Through intensive participation in social networking sites, individuals may have greater access to informational, emotional, and tangible support. Furthermore, participation in social networks enables individuals to develop norms of trust and reciprocity, which are necessary for successful engagement in collective activities. Studies on the effects of the use of the social network on trust and participation in civic and political activities among college students found that intense use of Facebook was associated with greater trust and participation in civic and political activities (Valenzuela et al., 2009).

Social networking sites can also be used for a variety of activities, such as one-to-one communication, broadcasting to a wider audience, and the passive consumption of news. Direct communication with individual friends consists of personal, one-to-one exchanges. On the other hand, social broadcasting refers to the updating of one's personal status and the posting of content for the consumption of unlimited others, while the passive consumption of news refers to browsing through the posts of other users. Thus, the use of social network sites for direct one-to-one communication and passive consumption of news appears to be associated with higher levels of social capital.

Social Media and Trust

Social capital is strongly interconnected with the concept of trust (Best and Krueger, 2006). Trust can be defined as the willingness to accept a vulnerable situation based on positive expectations about the actions of others. When social players do not trust others and do not feel that other players trust them, cooperation is unlikely. Trust is usually higher among individuals who participate in networks of social engagement (Blanchard and Horan, 1998).

Online trust is a difficult goal to achieve because it requires the establishment of trusting relationships in the online world. The Internet itself is an object of trust with uncertain conditions just because it is not based on dyadic or group relations alone (Wang and Emurian, 2005). Internet users perceive the online world as one in which they are uncertain about the risks and results of their online transactions. For this reason, the online environment is viewed as insecure. Studies of online self-disclosure have shown a direct relationship between trust in online providers and self-disclosure. Users tend to adjust how much information they disclose based on a consideration of perceived risks and trust (Joinson et al., 2010). Studies have also investigated the association between Internet use and trust. Early studies examined the extent to which Internet use promoted interpersonal trust and reported that Internet users were no more likely than nonusers to report interpersonal trust. However, trusting people are more likely to use the Internet for shopping and less likely to believe that their privacy will be violated online (Uslander, 2004).

As to the association between online and off-line trust and its effect on the disclosure of personal information, the Internet attribute approach implies that the attributes of the Internet (relative anonymity, lack of sufficient cues, and controllability) generate online behavior that is independent of off-line behavior. According to this approach, the structural attributes of the Internet encourage interactants to engage in more intimate exchanges in online settings than in face-to-face settings. Fewer nonverbal cues and greater controllability reduce people's inhibitions when interacting online, perhaps leading to the rapid development of online trust that in turn may lead to increased online self-disclosure. According to the Internet attribution approach, the ability of the Internet to free one from inhibitions will be positively associated with both the development of online trust and the disclosure of identifiable information. It is important to note that according to this approach, online trust is not associated with trust in individuals or trust in institutions, but develops as a result of the attributes of the online environment.

In sum, studies that have investigated the role of Internet use and social media have established that the acquisition of social capital differs according to the type and purpose of the social media one uses. Using social media to access information and enhance one's social ties improves one's level of social capital. At the same time, research has not established whether this association is a causal effect or the result of common antecedent variables. Some studies link one's level of initial social resources with the types of Internet and social media one uses, determining that when controlling for initial resources, the gain in social capital may be modest. This line of research requires more attention in the future. In addition, while the Internet is a global technology, the vast majority of the studies have been produced in only one country. As the use of social media spreads to a growing number of countries, comparative studies are required to be able to generalize the findings of existing research.

See also: Computer Mediated Communication; Information Society; Information and Communications Technology in Education; Media Effects; Public Sphere and the Media; Religion and New Media; Social Capital; Social Media.

Bibliography

Bessière, K., Kiesler, S., Kraut, R.E., Boneva, B., 2008. Effects of Internet use and social resources on changes in depression. Information, Communication & Society 11 (1), 45–70.

Best, S.J., Krueger, B.S., 2006. Online interactions and social capital: distinguishing between new and existing ties. Social Science Computer Review 24, 395–410.

Blanchard, A., Horan, T., 1998. Virtual communities and social capital. Social Science Computer Review 16, 293–307.

Bourdieu, P., 1986. The forms of capital. In: Richardson, J.G. (Ed.), Handbook of Theory and Research for the Sociology of Education. Greenwald Press, New York, pp. 241–255.

Boyd, d., Ellison, N., 2007. Social network sites: definition, history and scholarship. Journal of Computer Mediated Communication 13, 210–230.

Coleman, J., 1990. Foundations of Social Theory. Belknap Press of Harvard University, Cambridge, MA.

Gershuny, J., 2003. Web use and net nerds: a neofunctionalist analysis of the impact of information technology in the home. Social Forces 81 (1), 141–168.

Hampton, K., Sessions, L.F., Her, E.J., 2010. Core networks, social isolation and new Media. Information, Communication and Society 14 (1), 130–155.

Hargittai, E., Hsie, H.L., 2010. Predictors and consequences of differentiated practices on social networking sites. Information, Communication and Society 13 (4), 515–537.

Jennings, K.M., Zeitner, V., 2003. Internet use and civic engagement. Public Opinion Quarterly 67, 311–334.

Joinson, A., Reips, U., Buchanan, T., Paine, C.B., 2010. Privacy, trust and self disclosure online. Human Computer Interaction 25 (1), 24.

Kaplan, A., Haenlein, M., 2010. Users of the world, unite! the challenges and opportunities of social media. Business Horizons 53 (1), 59–68.

Katz, J.E., Rice, R., 2002. Social Consequences of Internet Use. MIT Press, Boston.

Kraut, R., Lundmark, V., Patterson, M., Kiesler, S., Mukipadhyay, T., Scheris, W., 1998. Internet paradox: a social technology that reduces social involvement and psychological well-being? American Psychologist 53 (9), 1017–1031.

Marlow, C.A., June 2005. Linking without thinking: Weblogs, readership, and online social capital formation. In International Communication Association Conference.

Mazur, E., Richards, L., 2011. Adolescents; and emerging adults' social networking online: homophilly or diversity? Journal of Applied Developmental Psychology 32 (4), 180–188.

Mesch, G.S., Talmud, I., Quan-Haase, A., 2012. Instant messaging social networks: individual, relational, and cultural characteristics. Journal of Personal and Social Relationships 26, 736–759.

Nie, N.H., 2001. Sociability, interpersonal relations, and the Internet: reconciling conflicting findings. American Behavioral Scientist 45 (3), 419–435.

Putnam, R.D., 1993. The prosperous community. The American Prospect 4, 13.

Putnam, R.D., 1995. Bowling alone: America's declining social capital. Journal of Democracy 6 (1), 65–78.

Rice, R.E., Katz, J.E. (Eds.), 2000. Internet and health communication: experience and expectations. Sage, Thousand Oaks, CA.

Shah, D.V., Kwak, N., Lance Holbert, R., 2001. "Connecting" and "Disconnecting" with civic life: patterns of internet use and the production of social capital. Political Communication 18, 141–162.

Steinfield, C., Ellison, N.B., Lampe, C., 2008. Social capital, self-esteem and use of online social network sites. Journal of Applied Developmental Psychology 29, 420–433.

Subrahmanyam, K., Reich, S.M., Waechter, N., Espinoza, G., 2008. Online and offline social networks: use of social networking sites by emerging adults. Journal of Applied Developmental Psychology 29 (6), 420–433.

Uslander, E.M., 2004. Trust, civic engagement and the internet. Political Communication 21, 223–242.

Valenzuela, S., Park, N., Kee, K.F., 2009. Is there social capital in a social network site: facebook use and college students life satisfaction, trust and participation. Journal of Computer Mediated Communication 14, 875–901.

Wang, Y.D., Emurian, H.H., 2005. An overview of online trust: concepts, elements and implications. Computers in Human Behavior 21, 105–115.

Wellman, B., Quan-Haase, A., Witte, J., Hampton, K., 2001. Does the Internet increase, decrease or supplement social capital? American Behavioral Scientist 45 (3), 436–455.

New Media and the Digital Divide

Alexander van Deursen and Jan van Dijk, University of Twente, Enschede, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

During the 1990s, researchers and policy makers began discussing the presence of a so-called 'digital divide,' a distinction of people who do and do not have access to information and communication technologies. The concept of the digital divide stems from a comparative perspective of social and information inequality and depends on the idea that there are benefits associated with ICT access. Originally, the digital divide was defined in terms of physical access to a technology. This is now considered superficial and the emphasis is shifting to new dimensions, that is, inequalities of skills and usage.

Introduction: The Digital Divide Concept

During the 1990s, researchers and policy makers began discussing the presence of a so-called 'digital divide,' a distinction of people who do and do not have access to information and communication technologies (ICTs). The concept of the digital divide stems from a comparative perspective of social and information inequality and depends on the idea that there are benefits associated with ICT access and usage and negative consequences attending non-access and usage. Originally, the term 'digital divide' mostly referred to gaps in access to computers. When the Internet became widely accessible in society and began to provide a primary means of computing, the term shifted to encompass gaps in Internet access. Defining the digital divide in terms of access to the Internet is now the most popular convention. However, other digital equipments such as mobile telephony and digital television are not ruled out by some users of the term digital divide.

The term digital divide probably has caused more confusion than clarification. According to Gunkel (2003), it is a deeply ambiguous term in the sharp dichotomy it refers to. van Dijk (2005) has warned against a number of pitfalls of this metaphor. First, the metaphor suggests a simple divide between two clearly divided groups with a yawning gap between them. In fact the divide is more like a spectrum with people who use computers and the Internet for about every daily task on the one side and the people not using them at all on the other. Secondly, it suggests that the gap is very difficult to bridge. A third misunderstanding might be the impression that the divide is about absolute inequalities, which is between those included and those excluded. In reality most inequalities of the access to digital technology observed are more of a relative kind (see below). A final wrong connotation might be the suggestion that the divide is a static condition while in fact the gaps observed are continually shifting.

An important theoretical distinction concerned is that between individualistic and relational conceptions of social inequality. The first conception departs from so-called methodological individualism (Wellman and Berkowitz, 1988). Differential access to information and computer technologies (ICTs) is related to individuals and their characteristics: level of income and education, employment, age, sex, and ethnicity, to mention the most important ones. This is the usual approach in survey research, which measures the properties of individual respondents. Making multivariate analyses of several individual properties and aggregating them to produce properties of collectivities, one hopes to find background explanations. An alternative notion of inequality uses a relational or network approach (Wellman and Berkowitz, 1988). Here, the prime units of analysis are not individuals but the positions of individuals and the relationships between them. Inequality is not primarily a matter of individual attributes but of categorical differences between groups of people. This is the point of departure of the pioneering work *Durable Inequality* by the American sociologist, Charles Tilly (1999). "The central argument runs like this: Large, significant inequalities in advantages among human beings correspond mainly to categorical differences such as black/white, male/female, citizen/foreigner, or Muslim/Jew rather than to individual differences in attributes, propensities, or performances" (Tilly, 1999: p. 7). In this conception gender inequality in using ICTs, for example, is not explained by the presumed characteristics of males and females (females having less technical interest etc.) but by the gender relations between them in which males first appropriate new technologies and exclude females in daily practice.

Despite these conceptual problems the term 'digital divide' drew attention to the important issue of information inequality in scholarly and political communities at the turn of the century. Countries increasingly realized that the digital divide reduces the potential of the labor force and innovation. Information and communication technologies were considered to be a growth sector in the economy that should be supported in global competition. Between 2000 and 2004, scientific and policy conferences concerning the digital divide were exceedingly popular, but attention to this matter began to decline in 2004 and 2005 (van Dijk, 2006). On political and policy-making fronts, many observers, particularly those in rich, developed countries, reached the conclusion that the problem was almost solved, as a rapidly increasing majority of their inhabitants obtained access to computers, the Internet, and other digital technologies.

The common current opinion among policy makers and the public at large is that the divide is closing between those who do and do not have access to computers, the Internet, and other digital media. In some countries, Internet connection rates in households have reached the figure of 90%. Computers, mobile telephony, digital televisions, and other digital media are becoming cheaper by the day, while their capacity to

perform complex tasks increases. These media are introduced on a massive scale and into all aspects of everyday life. Several applications appear so easy to use that basic literacy supposedly is the sole prerequisite for using them. However, simultaneously defining the digital divide in terms of physical access to a technology is considered superficial by digital divide researchers; physical access alone is no longer considered to be the most important factor explaining information superiority observed. The emphasis is shifting to new dimensions, which is inequalities of skills and usage.

Dimensions of the Digital Divide

First Dimension: Physical and Material Access

An important reason for the decreasing attention given to the digital divide in the first decade of the twenty-first century lies in the fact that divides in physical access to the Internet are closing in most Western countries. Concerns about acquiring physical access to digital media have completely dominated public opinion and policy perspectives in the last two decades. Indeed, these concerns are still paramount, as many people think the digital divide is closing because 90% of the population or more have access to computers and Internet. Such a number would put the Internet on par with television as a media source. One should note that the diffusion of the Internet in the past two decades has occurred even faster than that of television. However, on a global scale the situation is different; in 2010 Internet access was estimated between 20 and 25% of the world population, while in many developing countries, Internet access is still restricted to less than 10% of the population (UN/ITUstatistics).

Moreover, physical access is not equal to material access. Material access includes all costs related to the use of computers, connections, peripheral equipment, software, and services. These costs are diverging in many ways, and people with physical access have very different computer, Internet, and other digital media expenses. Considering the current economic crisis in the Western world, the problem of material access to computer and Internet resources for particular parts of the population might become more serious. In this regard, several scholars have pointed toward mobile phones and other portables such as tablet computers as technologies that have the potential to reduce the digital access divide. Mobile phones offer a more affordable means of access to the Internet than computers do when only simple applications that require small data capacity are used (Akiyoshi and Ono, 2008). They are supposed to offer a viable alternative for developing countries. However, one should also keep in mind that mobile phones are by no means a substitute for computers as they lack many advanced applications.

Whatever opportunities mobile phones offer, it is certainly a misconception to think that physical or material access to the Internet automatically brings all the benefits associated with Internet use. Indeed, it is rather schematic and superficial to conceive the digital divide in terms of a binary classification between those with and without physical access to computers or the Internet. Such a belief directly links rates of access to differences in material resources: one either does, or does not, have the resources to establish a connection to the Internet. Moreover, such a belief assumes that having a connection correlates with having access to all the advantages the Internet offers. Compaine (2001), for instance, relied on the Diffusion of Innovations (DI) theory of Everett Rogers (1963/1995), who theorized that innovations would spread through society in an S-curve. The S-curve measures the relative speed with which members of a social system adopt a particular innovation and focuses on the time required for that innovation to be adopted by a certain percentage of the system (Rogers, 1995). The adoption rate accelerates at a first tipping point, called critical mass; at this point, an innovation has been so widely adopted that its continued adoption is self-sustaining (Markus, 1987). At a second tipping point diffusion begins to slow down as the market starts to reach saturation. Here, the access divide starts to close. van Dijk (2012) has portrayed the current situation of the physical access divide on a world scale in terms of the S-curve with the locations of developed and developing countries mapped into it.

The two curved lines in **Figure 1** are splitting the main curved line that has the shape of an S to indicate that we have two sides or populations in the digital divide with different representations of the S-curve (that portrays the average). One curve is for those on the 'wrong' side of the digital divide, usually people with low education, low income, and higher age; and the other is for those on the 'right' side of the divide, mostly people with high education, high income, and lower age. When the two lines come together, the physical access divide is closing. However, at this point in time we still do not know whether it will close completely, called 'normalization' in the figure, or that a gap will remain because the social categories in the higher curve continue to have a lead because they first adopt every new innovation in the field of digital media (e.g., the transition from narrowband to broadband), which is called 'stratification' by Norris (2001).

If Compaine and van Dijk were correct in applying DI theory to the digital divide, increases in physical access to the Internet would no doubt correspond to the S-curve measuring

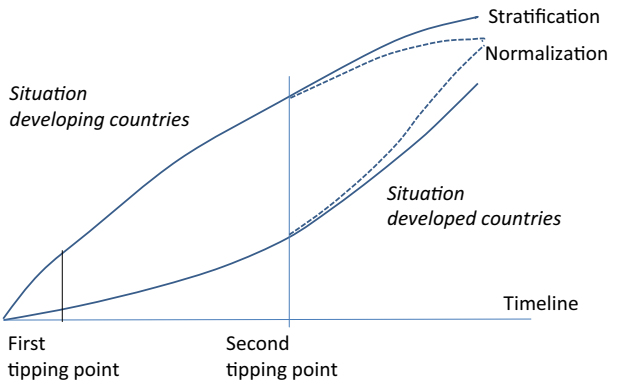

Figure 1 Evolution of the digital divide of physical access in time. Line below: access of categories of low education, low income, and higher age. Line above: access of categories of high education, high income, and lower age. Source: van Dijk, J.A.G.M., 2012. The evolution of the digital Divide. In: Brus, J., Crompton, M., Hildebrandt, M., Metakides, G. (Eds.), Digital Enlightment Yearbook 2012, IOS-Press, Amsterdam, pp. 57–75.

the adoption of innovations. From this point of view, the digital divide should steadily disappear as the diffusion rate reaches saturation, particularly when 'normalization' applies. The likelihood of the correspondence between the declining digital divide and the increasing rate of innovation diffusion would further be augmented as a result of the migration of the Internet to platforms such as digital television and mobile phones; indeed, the mistaken notion that the digital divide is a temporary problem of physical access has been reinforced by these migrations (Golding and Murdock, 2001). However, there are serious problems with applying the DI theory to the study of computer and Internet diffusion (Norris, 2001; van Dijk and Hacker, 2003). van Dijk (2005: pp. 62–65), which mentions several weak assumptions of this theory. First, why should the diffusion of each medium necessarily reach a 100%? Instead of this 'normalization' it might stop far before the stage of population-wide diffusion or become stratified (some categories keep adopting more and earlier than others). Secondly, while the DI theory assumes the existence of a single confined medium that does not change, in fact digital media devices often are combined (multimedia) and they continually change in functionality, quality, price, and appearance (think about the history of the PC). Finally, the determinism of DI theory is striking: why should there be innovators, early adopters, and early or late majority users with every medium?

Other Dimensions Come Forward: The Second-Level Digital Divide

Because it is wrong to assume that physical access to computers and the Internet automatically entails all benefits associated with their use, the digital divide should not be considered as a divide of physical access only. In the literature about the digital divide published after 2000, this conclusion comes forward stronger and stronger. Other dimensions have also come forward. Kling (2000), for instance, suggested a distinction between technical access (i.e., material availability) and social access (i.e., professional knowledge and technical skills necessary to benefit from information technologies). Attewell (2001) distinguished between a first digital divide and a second digital divide. Hargittai (2002), however, suggested what has become perhaps the most familiar distinction: that between first- and second-level digital divides. Besides these dimensions Warschauer (2003) also argued that factors such as content, language, literacy, educational level attained, and institutional structure must be considered. van Dijk (2005) proposed a causal model with four types of access to ICTs: motivational access (e.g., the lack of the elementary digital experience by people who have no interest or feel hostile toward ICTs), physical access (e.g., the availability of ICTs), digital skills (e.g., the ability to use ICTs), and usage access (the opportunity and practice of using ICTs). He calls the shift from physical access to skills and usage a 'deepening divide' as unequal skills and usage are deeply entrenched in existing social inequalities and because they will deepen these inequalities further.

The Dimension of Motivation

Each of these scholars shares the following view: while gaps in physical access might be closing in certain respects, other digital divides have begun to grow. In his discussion of motivation access, for instance, van Dijk (2005) argues that the wish to have a computer and a connection to the Internet precedes physical access. Thus, many of those who remain on the excluded side of the digital divide do so for motivational reasons. According to van Dijk, there are not only 'have-nots' but also 'want-nots.' At the start of the diffusion of new technologies, motivational concerns are the strongest forces to stimulate or prevent the acceptance of those technologies. In several European and American surveys conducted between 1999 and 2003, half of those individuals unconnected to the Internet explicitly stated that they would refuse to seek a connection for the following reasons: no need or significant usage opportunities, no time or liking, rejection of the medium (e.g., Internet and computer games viewed as 'dangerous' media), lack of money, or lack of skills (e.g., ARD-ZDF, 1999b and a Pew Internet and American Life survey: Lenhart et al., 2003). These observations lead us to one of the most confusing myths produced by popular ideas about the digital divide: that people are either in or out, included or excluded. To explain people's motivations for using digital technologies, mental and psychological conditions are often mentioned in literature about the digital divide. Here, the phenomena of computer anxiety and technophobia are still relevant and continue to create barriers to computer and Internet access in many countries, especially among seniors, people with low educational background, and segments of the female population (Brosnan, 1998; Chua et al., 1999; Rockwell and Singleton, 2002; UCLA, 2003).

The Skill Dimension

In contemporary literature, inequality of Internet skills is increasingly acknowledged as a key dimension of the digital divide. Several terms are used to frame Internet skills, e.g., digital or information literacy, computer skills, ICT literacy, or web fluency. To date, very little scientific research has focused on the actual level of digital skills possessed by various populations. Many large-scale surveys have revealed dramatic differences in skills among populations, including those populations in countries experiencing the extensive diffusion of new media (van Dijk, 2005; Warschauer, 2003). Nevertheless, these surveys measure actual levels of digital skills only by asking respondents to estimate their own proficiency.

A better way to obtain valid and complete measurements of digital skills is to implement performance tests; such tests would require participants to perform those computer and Internet tasks that are regularly performed in daily life. Hargittai (2002) has begun to implement performance tests in this field. Asking 54 demographically diverse Americans to perform different Internet search tasks, she discovered enormous differences in the levels of accomplishment and in the time needed to complete the tasks. In the Netherlands, van Deursen and van Dijk (2010, 2011) conducted performance tests in a university media laboratory on a cross-section of the Dutch population; more than 300 people were tested. Subjects who took the test showed a fairly high level of basic operational and formal skills, but they experienced much more difficulty in processing content-related information and in exercising strategic skills.

The results showed significant differences in performance between people of different ages and levels of education. Age primarily appears to be a significant contributor to the basic skills to use the Internet medium-related skills, as younger people perform better on these skills than that of older people. In contrast, older individuals performed better where content-related skills, including information and strategic skills, were needed; this occurred in all instances where the older individuals possessed an adequate level of medium-related skills. However, because many seniors tend to lack medium-related Internet skills, they are seriously limited in their content-related skills. Nevertheless, this observation provides another perspective on popular notions about the abilities of the so-called 'digital generation.' It also shows that the skills inequality problem will not automatically disappear in the future, and that life experience and substantial education of all kinds remain vital for acquiring digital skills.

The Usage Dimension

Aside from divergences in skill levels, the digital divide debate has increasingly drawn attention to the actual usage of the Internet. As a dependent factor, Internet use can be measured in several ways (e.g., usage time and frequency; number and diversity of usage applications; broadband or narrowband use; more or less active or creative use). Statistics regarding usage time and frequency are notoriously unreliable, as they rely on shifting and divergent operational definitions that are often determined by market research bureaus. These statistics give only some indication of the difference between actual use and physical access. It is certain, for example, that actual use diverges greatly from potential use. Furthermore, those who have a computer and/or Internet connection not always actually use them. Many assumed users actually use the computer or the Internet only once a week or a few times a month; some people never use them.

It is important to understand that when a physical access gap for a particular social category closes, usage of the medium concerned does not automatically equalize. Here the concept of 'usage gap' might apply. This concept is comparable to the concept 'knowledge gap' created in the 1970s by Tichenor et al. (1970). While the knowledge gap concerns the differential derivation of knowledge achieved through mass media and focuses, in particular, on mass media's influence on perception and cognition, the usage gap concept is much broader and potentially more effective in terms of social inequality because this gap concerns differential uses of and activities with computers and the Internet in all spheres of daily life, not just the derivation of knowledge.

The usage gap closure becomes most apparent when looking at types of usage. It is generally assumed that some Internet activities are more beneficial or advantageous for Internet users than others. Some activities offer users more chances and resources to move forward in their career, work, education, and societal position than others that are mainly consumptive or entertaining (e.g., Hargittai and Hinnant, 2008; Kim and Kim, 2001; Mossberger et al., 2003; van Dijk, 2005; Wasserman and Richmond-Abbott, 2005). In terms of the theories of capital inspired by Bourdieu (1984), one could also say that certain Internet activities allow users to accrue more economic, social, and cultural capital and resources than other activities. While some sections of the population will more frequently use those applications that have the greatest advantages for accruing capital and resources (such as work, career, study, and societal participation), other sections will choose to use those entertainment applications that have little or no advantage for accruing capital and resources (e.g., van Dijk, 1999; Bonfadelli, 2002; Park, 2002; Zillien and Hargittai, 2009).

These differences in types of Internet use call into question the belief that growing up in a digital world results in an intuitive and unproblematic use of digital technologies (see Prensky, 2001). A difference exists between the personal and purposeful uses of technologies such as the Internet. Recently, several scholars have addressed the digital divide by attempting to classify Internet usage types. Some of these classifications take the uses-and-gratifications approach (Katz et al., 1974) as a starting point, while others make use of the Technology Acceptance Model (Davis, 1989) or Social Cognitive Theory. Finally, there are scholars who account for differences in usage by grouping Internet users into use typologies (e.g., Ortega Egea et al., 2007).

The Digital Divide and Inequality

What is the stake or concern of the digital divide? Do people with no access or limited access actually experience disadvantages? Two contra-arguments are given frequently by people relativizing the importance of the digital divide. The first argument is that people still have the traditional media at their disposal, which also deliver information and provide the needed communication channels. For those who have no Internet, a multitude of broadcasting and print media is available, and for those who have no access to e-commerce, physical shops abound. People searching new social contacts or romantic encounters do not necessarily require a social-networking site or an online-dating service, as the choice of physical meeting places is immense. Those individuals who want to make a reservation can still pick up the phone. The second argument is that the nonhierarchical nature of the Internet, in conjunction with the declining cost of computing technologies and their increasing user friendliness, encourages social leveling and undermines existing patterns of class, race, and gender inequalities (see Tambini, 2000).

However, after the times of the Internet hype around the turn of the century most scholars now agree that differences in Internet access among segments of the population tend to reinforce social inequalities. Witte and Mannon (2009), for example, argue that the Internet is both intertwined with and consequential for inequality. Considering the Internet in a paradoxical way, they contend that it is not only an exemplar of a free and open society but also an active reproducer and potential accelerator of social inequality. To understand this paradox one has to realize that the digital divide is more characterized by relative than by absolute inequalities (see Introduction). The digital divide signifies a spectrum of positions of inclusion and exclusion, not a binary classification of absolute inclusion and exclusion. Ninety percent or more of a population might have access to computers and the Internet, while social and information inequality continue to rise on account of unequal material access, digital skills, and usage opportunities.

A classification of resources suggested by Bourdieu (1984) is often mentioned in discussions of the Internet's contribution to inequality. Bourdieu reimagined both Marx and Weber's ideas of social inequality in industrial society by defining economic, cultural, and social capital. Economic capital reflects Marx's ideas of property assets and other economic inequalities such as particular positions in the relations of production that might increase one's capacities (Hoffman, 2008). Meanwhile, social capital consists of resources based on group membership, relationships, networks, and support; according to Bourdieu, social capital "constitutes the totality of current and potential resources connected to a durable network of more or less institutionalized relations characterized by mutual knowing or acknowledgments; in other words, social capital is a resource based on the affiliation to a group" (Bourdieu, 1984: p. 190). Finally, cultural capital is formed by knowledge, skills, education, and social–cultural distinction, all of which give a person a higher status in society.

All forms of capital defined by Bourdieu are applicable to the Internet. In fact, Internet use and Bourdieuian forms of capital reinforce each other. On the one hand, all forms of capital affect Internet access: economic capital is a requisite to acquire physical and material access to computers and the Internet; social capital is needed to learn from others how to connect to the Internet, how to use it, and how to connect to those others; finally, cultural capital is required to cope with the diversity of content available to people of different cultural backgrounds. On the other hand, when these three requirements are met, the Internet in turn starts to affect the three forms of capital. For example, economic capital is increased by buying profitable resources online or by finding better jobs; social capital can be augmented by extending physical networks into virtual ones, perhaps increasing civic engagement and a sense of community (Katz and Rice, 2002); and cultural capital can be increased by using the Internet for learning purposes.

The result of the mutually determining relationship between forms of capital and the Internet is the usage gap discussed before which means that some individuals use the Internet in capital-enhancing ways, while others either do not use it or use it in less effective and less profitable ways; and the Internet may, of course, also be used simply for leisure (Hargittai and Shafer, 2006; Zillien and Hargittai, 2009). Furthermore, the intensive and extensive nature of Internet use among the well-off and well-educated correlates with an elite lifestyle, one from which those with less capital are excluded (van Dijk, 2005; Witte and Mannon, 2009). The result of this last observation is that groups with fewer forms of capital are likely to be affected in negative ways by the Internet: flights will be booked, concerts will be sold out, jobs will be given away, and dates will primarily be granted to those having access. This is the main counter-argument against the argument that traditional channels still offer all opportunities mentioned above. Rather than encouraging equality, the Internet tends to reinforce social inequality and can lead to the formation of more permanent disadvantaged and excluded groups, unless deliberate economic, cultural, and educational policies are installed by responsible policy makers (Golding, 1996; van Dijk, 2005). As Witte and Mannon (2009) contend, Internet access can be understood as a tool for the maintenance of class privilege and power as the inequalities upon which they rest are reproduced from one generation to the next.

Much theoretical work on the concept of network society also suggests that the Internet increases social inequalities. Castells (1999) considers the divide between those who are and are not networked as one of the major axes of social inequality (Castells, 1999). He contends that the particular framing of ICTs in the context of global, informational, and increasingly deregulated capitalism has been a major factor in the increase of inequality, in general, and of social exclusion, in particular. Because technological access is essential for both the improvement of living conditions and personal development, ICTs deepen discrimination and inequality in the absence of deliberate corrective policies (Castells, 1999).

van Dijk (1999, 2006, 2012) defines the network society as a form of society that organizes its relationships into media networks; these networks gradually replace or merge with the social networks of face-to-face communication. For van Dijk, media networks have become the nervous system of society and are shaping the primary means of social organization and the most important structures of modern society. In a network society, information should be considered a positional good, with some positions creating better opportunities than others for gathering, processing, and using valuable information (van Dijk, 2005). The positions people have in networks determine their individual potential power; this concept reflects the Weberian idea that human relationships should be considered alongside economic assets when discussing social inequality. When someone in a network society has only a marginal position or no position at all, he or she experiences social exclusion because the opportunities are reduced. Meanwhile, those who have a central position in society achieve social inclusion and thereby increase their power, capital, and resources (van Dijk, 2005). With unequal positions comes an unequal distribution of positional goods. According to the mechanism of opportunity hoarding, those individuals who have central positions in networks tend to appropriate more goods, to exert more control over particular facilities, and to wield greater power within networks; they then capture the returns of these resources and use them to reproduce the boundary between themselves and those who are socially excluded (Tilly, 1999).

In conclusion, most recent scientific literature on the digital divide suggests that the Internet has the potential to strengthen traditional kinds of inequality rather than to ameliorate them. Unequal access to the Internet has varying consequences in several areas of society: the economic (e.g., acquisition and maintenance of jobs), the social (e.g., development and maintenance of social contacts), the political (e.g., voting and other kinds of political participation), the cultural (e.g., participation in cyberculture), the spatial (e.g., the ability to lead a mobile life), and the institutional (e.g., recognition and attainment of citizens' rights).

Bibliography

Akiyoshi, M., Ono, H., 2008. The diffusion of mobile Internet in Japan. The Information Society 24, 292–306.

Attewell, P., 2001. The first and second digital divides. Sociology of Education 74 (3), 252–259.

Bonfadelli, H., 2002. The Internet and knowledge gaps: a theoretical and empirical investigation. European Journal of Communication 17 (1), 65–84.
Bourdieu, P., 1984. Distinction: A Social Critique of the Judgment of Taste. Routledge, London.
Brosnan, M.J., 1998. The impact of computer anxiety and self-efficacy upon performance. Journal of Computer Assisted Learning 14, 223–234.
Castells, M., 1999. Flows, networks and identity: a critical theory of the international society. In: Castells, M., et al. (Eds.), Critical Education in the New Information Age. Rowman and Littlefield, Lanham, pp. 37–64.
Castells, M., 2004. The Network Society: A Cross-Cultural Perspective. Edward Elgar, Northampton, MA.
Chua, S.L., Chen, D.T., Wong, A.F.L., 1999. Computer anxiety and its correlates: a meta-analysis. Computers in Human Behavior 15, 609–623.
Compaine, B.M., 2001. The Digital Divide: Facing a Crisis or Creating a Myth? MIT Press, London.
DiMaggio, P., Hargittai, E., Celeste, C., Shafer, S., 2004. From unequal access to differentiated use: a literature review and agenda for research on digital inequality. In: Neckerman, K. (Ed.), Social Inequality. Russell Sage Foundation, New York, pp. 355–400.
Davis, F.D., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quartely 13 (3), 319–340.
van Deursen, A.J.A.M., van Dijk, J.A.G.M., 2010. Measuring Internet skills. International Journal of Human-Computer Interaction 26 (10), 891–916.
van Deursen, A.J.A.M., van Dijk, J.A.G.M., 2011. Internet skills and the digital divide. New Media and Society 13 (6), 893–911.
van Dijk, J., Hacker, K., 2003. The digital divide as a complex and dynamic phenomenon. The Information Society 19 (4), 315–327.
van Dijk, J., 1999, 2006, 2012. The Network Society, First, Second and Third Edition. Sage Publicatons, London, Thousand Oaks CA, New Delhi.
van Dijk, J.A.G.M., 2005. The Deepening Divide, Inequality in the Information Society. Sage Publications, Thousand Oaks CA, London, New Delhi.
van Dijk, J., 2006. Digital divide research, achievements and shortcomings. Poetics 34 (4–5), 221–235.
van Dijk, J.A.G.M., 2012. The evolution of the digital divide. In: Brus, J., Crompton, M., Hildebrandt, M., Metakides, G. (Eds.), Digital Enlightment Yearbook 2012. IOS-Press, Amsterdam, pp. 57–75.
Golding, P., 1996. World-wide wedge: division and contradiction in the global information infrastructure. Monthly Review 48 (3), 70–86.
Golding, P., Murdock, G., 2001. Digital divides: communications policy and its contradictions. New Economy 8 (2), 110–115.
Gunkel, D., 2003. Second thoughts: toward a critique of the digital divide. New Media and Society 5 (4), 499–522.
Hargittai, E., Shafer, S., 2006. Differences in actual and perceived online skills: the role of gender. Social Science Quarterly 87 (2), 432–448.
Hargittai, E., 2002. Second-level digital divide: differences in people's online skills. First Monday 7 (4).
Hargittai, E., Hinnant, A., 2008. Digital inequality: differences in young adults' use of the Internet'. Communication Research 35 (5), 602–621.
Hoffman, R., 2008. Socioeconomic Differences in Old Age Mortality. Springer, Dordrecht.
Katz, E., Blumler, J., Gurevitch, M., 1974. Utilization of mass communication by the individual. In: Blumler, J., Katz, E. (Eds.), The Uses of Mass Communication: Current Perspectives on Gratifications Research. Sage, Beverly Hills, CA, pp. 19–34.
Katz, J.E., Rice, R., 2002. Social Consequences of Internet Use: Access, Involvement, and Interaction. MIT Press, Cambridge, MA.
Kling, R., 2000. Learning about information technologies and social change: the contribution of social informatics. The Information Society 16 (3), 217–232.
Kim, M., Kim, J., 2001. Digital divide: conceptual discussions and prospect. In: Kim, W., Ling, T.W., Lee, Y.L., Park, S.S. (Eds.), The Human Society and the Internet. Springer, New York, pp. 136–146.
Lenhart, A., Horrigan, J., Rainie, L., Allen, K., Boyce, A., Madden, M., et al., 2003. The Ever-Shifting Internet Population: A New Look at Internet Access and the Digital Divide. Pew Internet and American Life Project, Washington, DC.
Markus, L.M., 1987. Toward a "critical mass" theory of interactive media. Communication Research 14 (5), 491–511.
Mossberger, K., Tolbert, C.J., Stansbury, M., 2003. Virtual Inequality: Beyond the Digital Divide. Georgetown University Press, Washington, DC.
Norris, P., 2001. Digital Divide: Civic Engagement, Information Poverty and the Internet Worldwide. Cambridge University Press, Cambridge.
Ortega Egea, J.M., Menéndez, M.R., González, M.V.R., 2007. Diffusion and usage patterns of Internet services in the European Union, informations research. International Electronic Journal 12, 302.
Park, H.W., 2002. The digital divide in South Korea: closing and widening divides in the 1990s. Electronic Journal of Communication 12 (1–2).
Prensky, M., 2001. Digital natives, digital immigrants. On the Horizon 9 (5), 1–6.
Rockwell, S., Singleton, L., 2002. The effects of computer anxiety and communication apprehension on the adoption and utilization of the Internet. Electronic Journal of Communication/Revue de Communication Electronique 12 (1).
Rogers, E.M., 1963/1995. Diffusion of Innovations. Free Press, New York.
Steyaert, J., 2002. Inequality and the digital divide: myths and realities. In: Hick, S., McNutt, J. (Eds.), Advocacy, Activism and the Internet. Lyceum Press, Chicago, pp. 199–211.
Selwyn, N., 2006. Digital division or digital decision? A study of non-users and low-users of computers. Poetics 34 (4–5), 273–292.
Tambini, D., 2000. Universal Internet Access: A Realistic View. Retrieved from: http://www.csls.ox.ac.uk/.
Tichenor, P.J., Donohue, G.A., Olien, C.N., 1970. Mass media flow and differential growth in knowledge. Public Opinion Quarterly 34 (2), 159–170.
Tilly, C., 1999. Durable Inequality. University of California Press, Berkeley, CA.
UCLA, University of California, Los Angeles, Center for Communication Policy, 2003. The UCLA Internet Report: Surveying the Digital Future, Year Three. UCLA, Los Angeles.
Warschauer, M., 2003. Technology and Social Inclusion: Rethinking the Digital Divide. The MIT Press, Cambridge, MA.
Wasserman, I.M., Richmond-Abbott, M., 2005. Gender and the Internet: causes of variation in access, level, and scope of use. Social Science Quarterly 86, 252–270.
Wellman, B., Berkowitz, S.D. (Eds.), 1988. Social Structures, a Network Approach. Jai Press, Greenwich Conn., London.
Willis, S., Tranter, B., 2006. Beyond the digital divide. Internet diffusion and inequality in Australia. Journal of Sociology 42 (1), 43–59.
Witte, J.C., Mannon, S.E., 2009. The Internet and Social Inequalities. Routledge, New York.
Zillien, N., Hargittai, E., 2009. Digital distinction: status-specific types of Internet usage. Social Science Quarterly 90 (2), 274–291.

New Media, News Production and Consumption

Eugenia Mitchelstein and Pablo J Boczkowski, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the scholarship on online news consumption, addresses the limitations of this scholarship, and proposes avenues for future inquiries. It examines the scholarship in five areas: displacement and complementarity between traditional and online news consumption; audience fragmentation and homogenization; online news and political knowledge; Internet information as a resource for participation; and the user as a content producer. The assessment suggests that while audiences have straddled between tradition and transformation in their news consumption practices, most scholarly examinations have been characterized by stability rather than innovation.

Citizens in modern societies routinely engage in news media consumption. Since the mid-1990s, online news has been incorporated into the media diets of large segments of the population. In December 2011, four of ten Americans reported getting most of their national and international news from the Internet (Project for Excellence in Journalism, 2012). As online news audiences have expanded and become mainstream, research on online news consumption has increased and consolidated. However, but for a few exceptions (Boczkowski, 2002; Mitchelstein and Boczkowski, 2010), there have been few comprehensive analyses of the findings presented by these studies, and of the new directions research should take to contribute to a more thorough understanding of the manifold aspects of the phenomenon. This article aims to fill those two voids by providing first, a review of the literature on new media and news consumption published so far; and second, by addressing the main limitations in the existing scholarship and proposing new avenues for future inquiries. Building upon a prior article (Mitchelstein and Boczkowski, 2010), this article accomplishes these two goals by examining the research on five key areas of scholarship: the relationship of either displacement or complementarity between traditional media and online news consumption; audience fragmentation and homogenization; the link between online news access and political knowledge; Internet information as a resource for civic participation; and, the user as a content producer. Our assessment suggests that while audiences have straddled between tradition and transformation in their daily news consumption practices, most scholarly examinations have been characterized by stability rather than innovation, which may have obstructed a thorough assessment of the balance between steadiness and variance in online news consumption.

Does Online News Consumption Complement or Displace Traditional Media Consumption?

The widespread adoption of online news consumption has triggered studies on whether Internet news use complements or displaces traditional sources. One camp of scholarship proposes that online news consumption complements access to information in traditional media. For instance, Taneja and colleagues observe media use behaviors among more than 40 000 Americans and find that "that access to on-demand media (85% had home internet), does not displace linear television viewing" (Taneja et al., 2012: 964), which confirms the complementarity hypothesis. A different perspective draws on uses and gratification theory to propose that consumers' goals shape consumption more strongly than medium attributes, and thus, people who are interested in news tend to access this content in the multiple platforms available to them (Yuan, 2011).

Based on the principle of relative constancy, the opposite camp argues that audiences' adoption of Internet news displaces their use of traditional media (Gaskins and Jerit, 2012). The 2010 Pew Biennial News Consumption Survey indicates that the portion of Americans who read a newspaper 'yesterday' fell from 38 to 26% between 2006 and 2010, whereas the percentage who accessed their news online rose from 23 to 34% in the same period (2010). Temporal, demographic, and socioeconomic factors shape the interplay between complementarity and displacement. Some authors propose that audience behavior changes over time, as the appropriation of online technologies is normalized (Althaus and Tewksbury, 2000). Research indicates that displacement effects are greater among better educated and younger users (De Waal and Schoenbach, 2010). This effect is particularly salient among the digital natives, many of whom have tenuous links to traditional media. Gaskins and Jerit argue that young users are not likely to "substitute the Internet for print news since they would not [give up] a habit they had not acquired in the first place" (2012: 202).

Some authors have linked adoption of online news to the perceived credibility of Internet information (Bucy, 2003; Tsfati, 2010). Yuan conducted a survey of Chinese news consumers and proposes that "the more credible the news users believed a medium to be, the higher the odds were for them to include it in their news media repertoires" (2011: 1007). Regarding media factors, some studies find that online news is more credible than news on television and radio, but less than that in newspapers (Flanagin and Metzger, 2000; Schweiger, 2000). Other research studies look at differences concerning the institutional affiliation of online news sites and conclude that users rate traditional media sites as more credible than independent sites. For instance, in a survey of college students in the United States, Chung, Nam, and Stefanone found that

they rated mainstream media sites as more credible than independent sites, and propose that these "news providers benefit from their extant offline reputation" (2011: 182).

Fragmentation and Homogenization among Online Media Audiences

Scholarship on Internet news consumption has also examined whether the proliferation of information sources online is tied to either fragmentation or homogenization of the audience. The first stream of research draws on selective exposure theory to propose that diversification in online news supply may lead to polarization as people obtain their information increasingly from, and discuss public affairs with, like-minded individuals (Sunstein, 2001). For instance, Stroud examined online information seeking behavior during the 2005 presidential campaign in the United States and found that "partisans were more likely to select likeminded Web sites" (2011: 59).

An alternative stream of research contends that there is little evidence of ideological fragmentation in Internet news media consumption (Brundidge, 2010; Webster and Ksiazek, 2012). Researchers who study online consumption patterns propose that a small set of relatively moderate mainstream news outlets get the largest share of the traffic, while ideologically-oriented sites receive much less attention from the audience (Hindman, 2009; Meraz, 2009). Gentzkow and Shapiro examine browsing behavior in the United States from 2004 to 2008 and find that "online news consumption is highly concentrated … the top 20 sites account for nearly 80% [of the visits]" (2010: 18).

Other analysts have proposed that increased content choice could lead to lack of attention to public affairs news among online news consumers (Boczkowski and Peer, 2011; Prior, 2007). However, Boczkowski and colleagues find higher levels of interest in public affairs online news articles during the 2008 presidential campaign in the United States than exactly a year later, and suggest that "engagement with public affairs issues might vary not only from person to person but also across temporal and contextual circumstances" (Boczkowski et al., 2012: 361).

The Consequences of Online News Consumption for Political Knowledge

Segmentation in online news audiences according to their level and type of attention to public affairs content raises the issue of whether access to online news fosters or hinders political knowledge. Scholars disagree on how best to characterize the relationship between Internet information consumption and information acquisition. Some researchers draw on Downs' economic theory of democracy (1957) to propose that the increased accessibility of online news makes it easier for citizens to access political information (Johnson and Kaye, 2003; Tewksbury and Rittenberg, 2009: 197).

But other researchers contend that increased availability of news content online might only increase access to public affairs content for those already engaged in political issues (Margolis and Resnick, 2000). Some scholars have drawn upon the knowledge gap theory to propose that the infusions of information have an uneven effect on citizen knowledge, as the population with higher levels of education tends to acquire this information at a faster rate than those with fewer years of schooling (Jerit et al., 2006). For instance, Yang and Grabe compare knowledge acquisition among South Korean citizens from different educational backgrounds, and conclude that "less educated people comprehended significantly more public affairs news from reading a newspaper than using an online source while more educated people delivered similar comprehension performances across the two media" (2011: 1223).

A different stream of research proposes that use of the Internet might increase the likelihood of unplanned news consumption, as consumers may increase their awareness about public affairs subjects when they are exposed to online information as a consequence of general Web use (Lupia and Philpot, 2005). Rainie and colleagues analyze survey data from the 2004 presidential campaign in the United States and find that "fully half of internet users (50%) said they encountered political news by happenstance browsing" (Rainie et al., 2005: 9).

Scholars have also analyzed whether exposure to online news is more or less conducive to learning political knowledge than use of traditional news media. Some studies show that online information consumption reduces knowledge and recall of current events (Althaus and Tewksbury, 2002; Dalrymple and Scheufele, 2007). Schoenbach, de Waal, and Lauf conduct a survey of the Dutch population and find that "reading print newspapers contributes to awareness of more public events and issues than using online newspapers does" (2005: 253). However, other research indicates that access to online news does not lead to a decrease in information acquisition when compared to traditional news media use (Drew and Weaver, 2006). Kwak and colleagues conducted a survey in Ann Arbor, Michigan, and found that "the Internet, as compared to television, was more useful to younger respondents in their understanding of international affairs" (Kwak et al., 2006: 203).

Online News as a Resource for Civic Engagement

Scholarship on the impact of online news consumption on democratic participation has fallen into two camps. The first one proposes that there are no significant links between online news consumption and civic participation (Margolis and Resnick, 2000). Drew and Weaver examine media effects during the 2004 presidential campaign in the United States and find that "internet news exposure/attention were not significant predictors of intention to vote after demographics, campaign interest, and traditional media exposure and attention were controlled statistically" (2006: 32). However, other studies have shown a positive relationship between access to online information and political engagement. Boulianne conducts a meta-analysis of research on the relationship between online news use and political participation and concludes, "increased access to a large, diverse set of political information may help reinvigorate civic life" (2009: 205).

A cognate area of research suggests that political discussion, both online and off-line, mediates the relationship between

online news use and political participation, as news consumers who talk about public affairs are more likely to increase their engagement with public affairs (Shah et al., 2007). Hardy and Scheufele propose that "people who seek information on the internet are more likely to engage in participatory behavior, with this effect being greater for those individuals who talk about politics with others regardless of the medium in which the discussion takes place" (2005: 81).

An alternative stream of scholarship views online discussion of public affairs as a form of civic participation, and proposes that online news media have lowered the cost of debating opinions in public (Etzioni, 2003; Papacharissi, 2004). However, studies of online discussion suggest that dominance of loud minorities, lack of rationality and civility in the exchanges, and exclusion of dissenting voices, hampers understanding (Boczkowski, 2010; Ruiz et al., 2011). For instance, Mitchelstein examines readers' motivations to comment on various online spaces and reports that posting comments in newspapers is related to self-expression goals, such as making opinions public and blowing off steam, while posting comments on blogs is linked to discussion and socialization motives. She adds that "participation in blogs appears to lead to higher perceptions of political competence than does participation in newspapers" (2011: 2014–2034).

The User as a Content Producer

Other analysts focus on audience members as producers of news content. Some scholars propose that, as consumers are able to participate in content creation, online news may become the result of a collaborative process between reporters and audience members (Boczkowski, 2004; Deuze, 2003). However, research has identified at least three limitations to user-authored news as a form of participation. First, some scholars suggest that news organizations are not enthusiastic about allowing audience members to become co-authors of content (Robinson, 2007). Singer and Ashman interview journalists at a British newspaper and find that they "expressed concerns about the credibility of user contributions, as well as the effects of uncivil comments on personal and institutional credibility" (2009: 18).

Second, research shows that bloggers and citizen journalists rely heavily on mainstream media for information. Meraz (2009) examines linking patterns of independent blogs and concludes that more than half of the outgoing links from independent blogs go to mainstream media sites. However, other scholars have found that, while user-generated journalism participates in the agenda-setting process, it neither follows the mass media agenda nor does it revolutionize the mainstream news dominated hierarchy (Anderson, 2010).

Third, scholarship on online news content indicates that members of the audience show limited interest in contributing content to online news sites. Robinson and Deshano interview people who had received training to produce content as citizen journalists in Madison, Wisconsin, and yet failed to do so, and report that "they were not confident that they could do 'journalism' on any level" (2011: 648). However, "for those who did begin contributing regularly, a richer sense of community followed participation" (2011: 654). Furthermore, research indicates that most user-generated content does not cover public affairs topics, but rather topics related to the contributors' personal life and interests (Ornebring, 2008). Gans indicates that citizen journalists "cover the informally organized parts of society, the 'private' worlds of family, friends, compatible neighbors, [and] work colleagues" (2007: 163).

Limitations of Previous Research

The previous sections have critically assessed recent scholarship on online news consumption, addressing five key research topics: whether it displaces or complements traditional media consumption; audience fragmentation and homogenization; the link between online news consumption and political knowledge; online news as a resource for civic participation; and, the user as a content producer. This assessment indicates that online news consumers straddle between tradition and innovation in their daily practices. Audiences have increased their use of online news but have failed to abandon traditional media as sources of information. They also exhibit a mix of fragmentation and homogenization in their content choices, and take advantage of increased content choice online to deepen their preexisting preferences toward either public affairs content or sports and entertainment news. Consumers forsake the opportunities offered by the web for increased civic engagement frequently but not always; and, for the most part, have not taken up the opportunities offered to create their own news. Thus, this account belies both reports of radical change in news consumption in the networked environment, and portrayals of online news audiences that engage in wholesale repetition of their off-line behaviors.

In a previous article (Mitchelstein and Boczkowski, 2010) we argued that most scholarship in this area has been characterized by stability rather than change, which may have interfered with a thorough assessment of the balance between tradition and transformation in online news consumption. These approaches have made significant contributions, but have also exhibited five important limitations: the assumption of a division between print, broadcast, and online media; the tendency to focus on either ordinary or extraordinary patterns of phenomena but not on both at the same time; the inclination to concentrate exclusively on online news, which hinders comparisons with previous media developments; the notion that research should treat media features and social practices separately; and the dearth of ethnographic research on Internet news audiences.

The assumption of a division between print, broadcast, and online media has hindered the appraisal of whether Internet news complements or replaces traditional media in audiences' daily budget. With few exceptions (Taneja et al., 2012), most analysts have assumed that consumers devote their attention exclusively to newspapers, television, radio, or the Internet (Gaskins and Jerit, 2012), and have, therefore, failed to register the ways in which news consumers integrate use of different media, by, for instance, searching online for in-depth information about a topic featured in broadcast news, or commenting on a social networking site or on a personal webpage about a newspaper article.

The tendency to focus on either ordinary or extraordinary patterns of phenomena has limited understanding on whether online news audiences are becoming increasingly fragmented or homogenized. While some studies on politically interested information seekers show they tend to seek attitude-reinforcing content (Stroud, 2011), other researchers show that the most accessed online news sites tend to be mainstream news outlets with moderate political views and similar agendas (Hindman, 2009). Boczkowski proposes that while "homogenization might characterize the center of the media spectrum and segmentation its periphery" (2010: 250). He adds that if consumption "was arrayed in a single dimension, the more frequent activities would be located at the center and marked by rising homogenization, and the less frequent ones would be placed at both ends and be signaled by increasing segmentation" (2010: 250).

A third limitation to comprehensive understanding of Internet audiences is the dearth of historical comparisons. For instance, accounts of incidental information acquisition online echo Baum's research on the effects of talk shows on political knowledge and interest of otherwise inattentive audiences. The relationship between online news consumption and information acquisition would greatly benefit from analysis that considered the historical context in which audiences access information, making sense of the continuities and discontinuities (Boczkowski and Lievrouw, 2007) in motivations, behaviors, and interpretations.

Research on Internet news as a resource for civic participation may have also been hampered by a tendency to focus either on media features or on social practices, thus assuming that all news consumption online would either contribute to (Margolis and Resnick, 2000), or have no effect whatsoever on, political engagement (Esser and de Vreese, 2007). Lupia and Philpot contended that the Internet is not a monolith, and urge researchers to focus 'on how certain kinds of web sites affect certain kinds of people' (2005: 1125).

Finally, research on users as news content producers may have suffered from the paucity of ethnographic studies. Most research on citizen journalists and bloggers is based on either surveys or content analysis of their work. The few exceptions to this trend, based on ethnography and in-depth interviews of bloggers and citizens journalists (Anderson, 2010; Robinson and Deshano, 2011) provide a richer understanding of who, under what conditions, and with what consequences engages in news content production outside the realm of professional journalism.

Discussion

These limitations have not precluded previous research from making important contributions about the frequency and mode in which users engage with information online; the kind of news content they seek on the Internet; and the consequences of Internet information consumption for civic knowledge and participation. However, we would like to devote the remainder of this article to propose ways in which future research may overcome these limitations.

The reliance on methodological and theoretical frames inherited from the study of media throughout the better part of the twentieth century may explain, at least partly, the persistence of these limitations. Thus, research tends to be grounded on a unidirectional, one-to-many model of communication, in which there is little room for variance in the content accessed, the times and modes of attention to information, and the impact of news consumption on public life. Moreover, users-as-content producers are mostly examined with respect to their compliance with rules and routines of professional journalists, which may no longer hold true for all professional content producers online.

The online environment allows researchers to challenge traditional assumptions of scholarship on news consumption. For instance, studies of online news consumption could contest the notion that news consumption occurs at designated times and spaces, and instead seek to examine to what extent different users integrate their access to online news with other media use and their daily activities, such as work and commuting. Boczkowski's examination on news consumption at the workplace (2010) shows that access to online news is tied to the flow of the working day and serves as social currency among colleagues. Scholars may also question the stark division between consumption and production, and instead aim at analyzing under which conditions, and with what consequences, news media users engage in information production. Finally, researchers could also challenge the separation between mass and interpersonal communication, as the online environment allows users to access and disseminate information meant both for a mass audience and for a limited network of known acquaintances (such as via email or through social networking sites). This reassessment of the conceptual foundations of scholarship on media would help tackle unresolved dichotomies in the study of online news audiences in particular and in the broader field of media and society in general.

See also: Audiences, Media; Citizen Journalism; Journalism; Media Effects; New Media and Social Capital; Political Communication; Social Media.

Bibliography

Althaus, S.L., Tewksbury, D., 2000. Patterns of internet and traditional news media use in a networked community. Political Communication 17 (1), 21–45.

Althaus, S.L., Tewksbury, D., 2002. Agenda setting and the "new" news – patterns of issue importance among readers of the paper and online versions of the New York Times. Communication Research 29 (2), 180–207.

Anderson, C.W., 2010. Journalistic networks and the diffusion of local news: the brief, happy news life of the "Francisville four". Political Analysis 27 (3), 289.

Boczkowski, P.J., 2002. The development and use of online newspapers: what research tells us and what we might want to know. In: Lievrouw, L.A., Livingstone, S. (Eds.), Handbook of New Media: Social Shaping and Consequences of ICTs. Sage, London, pp. 270–286.

Boczkowski, P.J., 2004. Digitizing the News: Innovation in Online Newspapers. MIT Press, Cambridge.

Boczkowski, P.J., 2010. News at Work: Imitation in an Age of Information Abundance. University Of Chicago Press, Chicago.

Boczkowski, P.J., Lievrouw, L.A., 2007. Bridging STS and communication studies: scholarship on media and information technologies. The Handbook of Science and Technology Studies 949–977.

Boczkowski, P.J., Peer, L., 2011. The choice gap: the divergent online news preferences of journalists and consumers. Journal of Communication 61 (5), 857–876.

Boczkowski, P.J., Mitchelstein, E., Walter, M., 2012. When burglar alarms sound, do monitorial citizens pay attention to them? The online news choices of journalists and consumers during and after the 2008 U.S. election cycle. Political Communication 29 (4), 347–366.

Boulianne, S., 2009. Does internet use affect engagement? A meta-analysis of research. Political Communication 26 (2), 193–211.

Brundidge, J., 2010. Encountering "Difference" in the contemporary public sphere: the contribution of the internet to the heterogeneity of political discussion networks. Journal of Communication 60 (4), 680–700.

Bucy, E.P., 2003. Media credibility reconsidered: synergy effects between on-air and online news. Journalism & Mass Communication Quarterly 80 (2), 247–264.

Chung, C.J., Nam, Y., Stefanone, M.A., 2011. Exploring online news credibility: the relative influence of traditional and technological factors. Journal of Computer-Mediated Communication 17 (2), 171–186.

Dalrymple, K.E., Scheufele, D.A., 2007. Finally informing the electorate? How the internet got people thinking about presidential politics in 2004. Harvard International Journal of Press-Politics 12 (3), 96–111.

De Waal, E., Schoenbach, K., 2010. News sites' position in the mediascape: uses, evaluations and media displacement effects over time. New Media & Society 12 (3), 477–497.

Deuze, M., 2003. The web and its journalisms: considering the consequences of different types of newsmedia online. New Media & Society 5 (2), 203–230.

Downs, A., 1957. An Economic Theory of Democracy. Harper & Row, New York.

Drew, D., Weaver, D., 2006. Voter learning in the 2004 presidential election: did the media matter? Journalism & Mass Communication Quarterly 83 (1), 25–42.

Esser, F., de Vreese, C.H., 2007. Comparing young voters' political engagement in the United States and Europe. American Behavioral Scientist 50 (9), 1195–1213.

Etzioni, A., 2003. Are virtual and democratic communities feasible. In: Jenkins, H., Thorburn, D. (Eds.), Democracy and New Media. MIT Press, Cambridge, Mass, pp. 85–100.

Flanagin, A.J., Metzger, M.J., 2000. Perceptions of internet information credibility. Journalism & Mass Communication Quarterly 77 (3), 515–540.

Gans, H.J., 2007. Everyday news, newsworkers, and professional journalism. Political Communication 24 (2), 161–166.

Gaskins, B., Jerit, J., 2012. Internet news: is it a replacement for traditional media outlets? International Journal of Press-Politics 17 (2), 190–213.

Gentzkow, M., Shapiro, J.M., 2010. Ideological Segregation Online and Offline. Unpublished working paper. National Bureau of Economic Research.

Hardy, B.W., Scheufele, D.A., 2005. Examining differential gains from internet use: comparing the moderating role of talk and online interactions. Journal of Communication 55 (1), 71–84.

Hindman, M., 2009. The Myth of Digital Democracy. Princeton University Press, Princeton, NJ.

Jerit, J., Barabas, J., Bolsen, T., 2006. Citizens, knowledge, and the information environment. American Journal of Political Science 50 (2), 266–282.

Johnson, T.J., Kaye, B.K., 2003. Around the world wide web in 80 ways: how motives for going online are linked to Internet activities among politically interested Internet users. Social Science Computer Review 21 (3), 304–325. http://dx.doi.org/10.1177/0894439303253976.

Kwak, N., Poor, N., Skoric, M.M., 2006. Honey, i shrunk the world! The relation between internet use and international engagement. Mass Communication & Society 9 (2), 189–213.

Lupia, A., Philpot, T.S., 2005. Views from inside the net: how websites affect young adults' political interest. Journal of Politics 67 (4), 1122–1142.

Margolis, M., Resnick, D., 2000. Politics as Usual: The Cyberspace "revolution". Sage Publications, Thousand Oaks, CA.

Meraz, S., 2009. Is there an elite hold? Traditional media to social media agenda setting influence in blog networks. Journal of Computer-Mediated Communication 14 (3), 682–707.

Mitchelstein, E., 2011. Catharsis and community: divergent motivations for audience participation in online newspapers and blogs. International Journal of Communication 5, 2014–2034.

Mitchelstein, E., Boczkowski, P.J., 2010. Online news consumption research: an assessment of past work and an agenda for the future. New Media & Society 12 (7), 1085–1102.

Ornebring, H., 2008. The consumer as a producer of what? User-generated tabloid content in The Sun (UK) and Aftonbladet (Sweden). Journalism Studies 9, 771–785.

Papacharissi, Z., 2004. Democracy online: civility, politeness, and the democratic potential of online political discussion groups. New Media & Society 6 (2), 259–283.

Pew, 2010. Pew Research Center Biennial News Consumption Survey: Americans Spending More Time Following the News. Pew Research Center for the People and the Press, Washington, D.C.

Prior, M., 2007. Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections. Cambridge University Press, New York.

Rainie, L., Cornfield, M., Horrigan, J., 2005. The Internet and Campaign 2004. Pew Internet and American Life Project, Washington, D.C.

Robinson, S., 2007. Someone's gotta be in control here. Journalism Practice 1 (3), 305–321.

Robinson, S., Deshano, C., 2011. Citizen jourmalism and their third places: what makes people exchange information online (or not)? Journalism Studies 12 (5), 642–657.

Ruiz, C., Domingo, D., Mico, J.L., Diaz-Noci, J., Meso, K., Masip, P., 2011. Public sphere 2.0? The democratic qualities of citizen debates in online newspapers. International Journal of Press-Politics 16 (4), 463–487.

Schoenbach, K., de Waal, E., Lauf, E., 2005. Research note: online and print newspapers – their impact on the extent of the perceived public agenda. European Journal of Communication 20 (2), 245–258.

Schweiger, W., 2000. Media credibility–experience or image?: a survey on the credibility of the world wide web in Germany in comparison to other media. European Journal of Communication 15 (1), 37.

Shah, D.V., McLeod, D.M., Kim, E., Lee, S.Y., Gotlieb, M.R., Ho, S.S., et al., 2007. Political consumerism: how communication and consumption orientations drive "lifestyle politics". Annals of the American Academy of Political and Social Science 611, 217–235.

Singer, J.B., Ashman, I., 2009. "Comment is free, but facts are sacred": user-generated content and ethical constructs at the guardian. Journal of Mass Media Ethics: Exploring Questions of Media Morality 24 (1), 3–21.

Stroud, N.J., 2011. Niche News: The Politics of News Choice. Oxford Univ Press, New York.

Sunstein, C.R., 2001. Republic.com. Princeton University Press, Princeton, New Jersey.

Taneja, H., Webster, J.G., Malthouse, E.C., Ksiazek, T.B., 2012. Media consumption across platforms: identifying user-defined repertoires. New Media & Society 951–968.

Tewksbury, D., Rittenberg, J., 2009. Online news creation and consumption implications for modern democracies. In: Chadwick, A. (Ed.), Routledge Handbook of Internet Politics. Routledge, pp. 186–200.

Tsfati, Y., 2010. Online news exposure and trust in the mainstream media: exploring possible associations. American Behavioral Scientist 54 (1), 22–42.

Webster, J.G., Ksiazek, T.B., 2012. The dynamics of audience fragmentation: public attention in an age of digital media. Journal of Communication 62 (1), 39–56.

Yang, J., Grabe, M.E., 2011. Knowledge acquisition gaps: a comparison of print versus online news sources. New Media & Society 13 (8), 1211–1227.

Yuan, E., 2011. News consumption across multiple media platforms: a repertoire approach. Information Communication & Society 14 (7), 998–1016.

New Media, Political Mobilization, and the Arab Spring

Habibul H Khondker, Zayed University, Abu Dhabi, United Arab Emirates

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article seeks to review the impact of the new media on political mobilization in the Arab Spring of 2011. The article examines the relationship between the information communication technology, especially the new media and its consequences – both intended and unintended – on social movement in the Middle East and North Africa region. Tracing the various arguments in social science literature and taking into account the political developments in the region, the article concludes that on the one hand new media played a critical role in political mobilization culminating in the Arab Spring and on the other hand, the new media can also be used as repressive tools by the authoritarian rulers to their advantage.

Introduction

This article examines the popularity and impact of information communication technology (ICT) on all the spheres of society: the economic, social, and political. ICT was favored by the governments of the Middle East and North Africa (MENA) region, as in other parts of the developing world, as a means for achieving modernization, yet the unintended consequences were more unpredictable and led to political mobilization and played a key role in the Arab Springs of 2011.

The Tunis conference organized by World Summit for Information Society (WSIS) in 2005 was a project of the UNESCO hosted with the support of the Government of Tunisia to promote and institutionalize ICT as it was the mantra of the time to pave the way of a digital revolution. As the world was caught up with the new paradigm of the information society or knowledge-based economy, many developing countries too joined the caravan of information highway. In response to Tunisia's request, the United Nations General Assembly approved the creation of WSIS in late 2001. The WSIS Tunis Commitment recognized that "freedom of expression and the free-flow of information, ideas and knowledge are essential for the Information Society and beneficial for development" (UNESCO, n.d.).

Apart from highlighting freedom of expression as a "key component for the establishment of democratic processes" (UNESCO, n.d.), UNESCO's role was consistent with its promotion of continuing development of the Internet as a global public resource with open access to everyone. Little did the leadership of Tunisia envisage that by 2010, the new information technology would – in addition to playing a role in the creation of an information society and knowledge economy – contribute to the democratic opening of the society by removing authoritarian leaders.

Information Communication Technology and Development

Starting in the last decade of the twentieth century and reaching its peak in the early decade of the twenty-first century, information society, knowledge-based economy, knowledge society, and so on became the new mantra of development. The new development paradigm swept the world, and many leaders of developing countries embraced the ICT to hasten economic development.

Social scientists by and large expressed some sense of skepticism viewing the new technology with its penetrating reach, an invasion into the privacy. Some writers raised the fear of a surveillance society in the offing. Information technology, many thought, would grant government more control. Such fears have been rekindled with the popularization of "Big Data" projects. Some writers, however, saw a great deal of potential in opening up space for free, uncoerced discussions furthering the processes of democratization.

Social science research on the potential of the new media in shaping society began in earnest since Manuel Castells groundbreaking work, *The Rise of the Network Society* (1996) and more recently, *The Communication Power* (2009). A number of writers in the last decade of the twentieth century recognized the potential of Internet as a vehicle for disseminating democracy (Poster, 1995). More recent writers have identified the nuances as well as limits to the role of the new media in politics (Howard, 2011). The discussions on this subject are divided between those who emphasized the controlling role of the new media, a new tool of repression in the arsenal of the dictators and those who see it as tools for democratic openness. Even in democratic societies, the new technology, as some writers point out, pose a grave threat to the freedom and privacy of citizens. One writer traced the evolution of technology from the industrial revolution to the microchip revolution arguing that technology has historically increased control in society and the new microchip would not be an exception (Beniger, 1989). Other writers often get somewhat carried away with the potential role of the new media in shaping politics, opening up a new public sphere, especially in societies where a real public sphere is absent. There are however some writers who have presented a more balanced view on the pitfalls and potential, controlling as well as emancipatory role of the new media.

Academic discussion of the potential of Internet in opening up political space for democratization preceded the 'Arab Spring.' Studies were conducted elsewhere to examine the potential of Internet (Ho et al., 2002; Kalathil and Boas, 2003). The Berkman Center for Internet and Society at Harvard University published a study on Arabic blogosphere in June 2009 where it analyzed 35 000 blogs and systematically mapped the discussions around certain nodes covering such topics

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64135-1

as terrorism, human rights, minority issues, critique of USA, personal issues, and so on.

Their findings dispelled "two common threads of speculation about the role of the Internet in the Arab world. The first is a view of the Internet as primarily a vehicle for radicalization. We found very little support for terrorism or violent jihad in the Arabic blogosphere and quite a lot of criticism …. The second thread is the hope that the Internet will empower political movements that can move Arab societies toward democratic values and governance. This study supports some aspects of the view that the Internet can empower political movements in the region, since it provides an infrastructure for expressing minority points of view, breaking gatekeeper monopolies on public voice, lowering barriers to political mobilization (even if symbolic), and building capacity for bottom-up contributions to the public agenda. But we caution against the idea that these functionally 'democratic' characteristics necessarily travel in the company of classically liberal, historically Western, values regarding individual liberties, freedom of belief and expression, and, ultimately, secular foundations of political rights. The Internet lays a good foundation for a battle of ideas, but it does not necessarily favor a winner" (Etling et al., 2009: 10).

With hindsight, their second point touches the nub as the diverse developments in the postrevolutionary Tunisia and Egypt reveal: the former heading for a tolerant democracy, while the later portends a less-tolerant authoritarian rule. Political mobilization in the latter case seems to be top down.

ICT and the Arab World in Perspective

Since the 1990s Internet has revolutionized the world. The drift of this revolution reached the Middle East at the beginning of the twenty-first century. The rapid spread of Internet technology in the region was part of its high-paced economic transformation and was effectively pushed by the respective governments. Countries with new wealth sought to achieve modernization and bought into the paradigms of technology-driven development models. Especially the Gulf countries took the lead in the economic growth with the help of the wealth generated from the export of hydrocarbon and rapid building of infrastructure; developments of ICT were not lagging behind. Yet, those who share a technological deterministic explanation of social uprising fail to recognize that the countries with the highest penetration of ICT were not at the forefront of what has come to be known as the 'Arab Spring.'

It would be useful to remember that political protests and upsurge began in the MENA region following the controversial presidential elections of 2009 in Iran, which returned President Ahmedinajad to a second term. The elections were charged with widespread irregularities, and accusations of rigging were numerous. The so-called Green Revolution, since the protestors, who wore green failed in Iran but triggered a series of uprisings in the entire region that came to be known as the 'Arab springs.'

The Middle East as a region has a population of 231 million of which 90 million as of December 2012 are Internet users. If we look at the world, of course in terms of aggregate number of people using Internet we find that the largest users of the Internet are countries like the population giants such as China, India, Brazil, Indonesia, and Nigeria. However, when we look at the Internet penetration as percentage of the population, Europe and North America are the world leaders with the exception of South Korea.

Table 1 Internet penetration in the Arab world vis-à-vis leading countries of the world in 2014

Country	*Internet users (in millions)*	*Population (in millions)*	*% of Internet users*
China	641.6	1300	46.03
USA	279.8	322	86.75
India	243.1	1200	19.19
Japan	109.2	126	86.03
Brazil	107.8	202	53.37
Russia	84.4	142.4	59.27
Germany	71.1	82.6	86.78
Nigeria	67.1	178.5	37.59
UK	57.0	63.4	89.90
France	55.4	64.6	85.75
South Korea	45.3	49.5	91.52
Egypt	40.3	83.3	48.34
Iran	22.2	78.4	28.29
Morocco	20.2	33.4	60.33
Saudi Arabia	17.3	29.3	59.24
UAE	8.8	9.4	93.24
Algeria	6.6	39.9	16.70
Israel	5.9	7.8	75.80
Syria	5.8	21.9	26.66
Tunisia	5.0	11.11	45.46
Yemen	4.7	24.9	19.14
Jordan	3.3	7.5	44.98
Lebanon	3.3	4.9	67.19
Kuwait	3.0	3.4	86.86
Iraq	2.7	34.7	7.79
Oman	2.5	3.8	65.82
Qatar	2.1	2.2	96.65
Libya	1.3	6.2	21.79
Bahrain	1.29	1.34	96.53

Source: http://www.internetlivestats.com/internet-users-by-country/.

Given its sheer population size, China is a world leader in the use of Internet. As of 2014, China has 641 601 070 Internet users of its 1.3 billion population, which accounts for 46.03% penetration accounting for 21.9% of the Internet users of the world with 19.24% of the world share of population. However, in terms of penetration of Internet vis-à-vis the population, only four countries of the world have over 90% Internet penetration. Of the four countries, three are in the Gulf region of the Middle East. They are Qatar (96.65%), Bahrain (96.53), and UAE (93.24). South Korea has an Internet penetration rate of 91.52% (Table 1).

There are over 128 million Internet users and around 82 million users of social media in the Arab world as of 2014, which stands 22% of the population, and with around 41% Internet penetration rate and a whopping 110% mobile telephony penetration this region is ahead of other developing regions of the world. Within the region, the penetration rates vary nationally. For example, Qatar ranks number 1 in the use of Facebook followed closely by the UUAE social media.

According to the Arab Social Media Report of 2014,

> social media has been growing rapidly in the Arab World since 2011 and has continued to do so over the past year. Globally, Facebook boasted 1.28 billion monthly active users (MAUs) by the end of the first quarter of 2014, including companies, brands, and governments, with 1.01 billion of Facebook's monthly active users accessing it through smart mobile devices. Twitter has also seen similar strong growth globally, with 255 million monthly active users as of end of 1st quarter of 2014, with 198 million of them accessing the platform through mobile devices. LinkedIn, too, has seen healthy growth in users this year. As of January, LinkedIn had over 300 million members, with 39 million of them being students and college graduates.

> With these global highlights in mind, growth of social media usage in the Arab world has been equally strong, with the number of users increasing by 49% on Facebook, 54% on Twitter and 79% on LinkedIn since May 2013. (p. 6)

Compared to other prominent developing regions, the Internet penetration in the MENA is quite remarkable. For example, in India more people have access to mobile phone than to toilets. Just over half of India's population uses cell phones, yet only 13% of Indians use the Internet, with only 6% on social networking sites, according to the Pew Research Center. Other BRICS countries – Brazil, Russia, India, China, and South Africa – average 46.5% use, and worldwide, only 36% adults reportedly have access to the Internet (Reilly, 2013).

As social media may have played a role in facilitating social mobilization and social movement in some of the countries in the region, social movement or as in some cases it is called the 'Arab Spring' has also played a role in popularizing the Facebook as a social media in the region. The speed of penetration between 2011 and 2012, as **Table 2** shows, was nothing but spectacular in Jordan, Kuwait, and Tunisia.

Table 2 Facebook penetration in the MENA region

Country	*2010 (December)*	*2012 (June)*	*2014 (May)*
UAE	45.38	40.6	58.0
Israel	42.93	45.3	50.2[a]
Qatar	33.95	29.3	60.9
Bahrain	34.27	25.9	38.9
Lebanon	23.11	32.7	45.9
Kuwait	20.64	34.3	48.3
Tunisia	17.55	27.9	42.1
Jordan	17.06	34.1	47.9
Saudi Arabia	12.24	19.2	28.1
Palestine	10.76	21.2	31.4
Morocco	7.55	13.4	21.7
Oman	7.55	16.6	27.9
Egypt	5.49	13.1	22.4
Algeria	3.99	9.5	18.1
Libya	3.74	3.7	3.7
Iraq	1.26	5.0	19.5
Syria	1.07	15.3	29.3

[a]In December 2012.

The Arab Spring and Social Media

There is an inherent contradiction in the spread of information revolution since it is often the government, aided by corporate interests, that promotes the new media thus inadvertently creating a space for civic activism. As Roland Robertson argues in another context, the growing surveillance technology – an aspect of new, invasive information technology – threatens an open society leading to a syndrome of open society and closed mind (Robertson, 2007). Drawing on his metaphor, it can be said that in other contexts, it may create a tension between closed society and open mind. New information technology has clearly the transformative potential to open up spaces of freedom. The recent political transformations in the MENA region provide us an opportunity to examine both the limits and potential contributions of the new media in paving the way for freedom and openness.

The Arab Spring of 2011–12 brought to public attention the role of the social media in political mobilization and social movement in the Arab World. As long-standing authoritarian regimes in Tunisia and Egypt fell in the face of social uprisings, several commentators sought to analyze the role of the ICT in whipping up a veritable revolution pushing long-standing dictators out of office. Some commentators likened the new media as responsible for creating a virtual civil society while the real or off-line civil society was either nonexistent or anodyne at best.

Is Internet or Facebook, or Tweeter or their combined effect good enough to cause social revolution leading to freedom? What about the use of the same instruments of the new media by the political power as tools of repression? What are the ancillary factors that may come into tilting the contest one way or the other? Writers such as Rasha Abdulla (2011) and Eltantawy and Wiest (2011) argue that Facebook and Twitter were instrumental in both organizing and motivating social movements in the Middle East. Others favor a more integrated view in explaining the Arab Spring where new media and traditional media, that is, television and newspapers complemented each other (Khondker, 2011).

Tunisia

The Arab Spring began in Tunisia and then affected Egypt. This genealogy can be improved by adding the early, albeit much smaller-scale, ICT-mediated movements in Egypt since 2005, and the Iranian protests and social movement, known as green movement that rose and failed, yet was successful to draw international attention and support, modify the local political scene and possibly sent a message that the idea of a social movement is not off the agenda in the Gulf countries.

Tufekci and Wilson (2012) remind us: "From about 2005 on, there were persistent efforts by small but dedicated groups, starting with the Kefaya ("Enough") movement and later the April 6th Youth Coalition. The unrest was not, by any means, limited to the online sphere. Indeed, the April 6th movement was initially formed to support a labor strike. However, social media and satellite TV were the main public spaces where politically sensitive subjects were openly discussed and dissent could be expressed. The introduction of Facebook in Arabic in

2009 amplified this process. Political content online increased, with multiple campaigns by citizen journalists and online activists exposing poverty, corruption, and human rights abuses, including torture. However, until the Tunisian uprising, physical protests in Egypt remained fairly small and the regime appeared to be firmly in control" (Tufekci and Wilson, 2012: 364).

Why the 'Arab Spring' revolutions did not sweep the entire Gulf region felling all the authoritarian regimes paving the way of democracy across the region will be debated for some time. Failed or successful are the call of the historians for which some time has to be elapsed; however, what is clear from the uprisings in the region, is that information technology has been a crucial accessory in the opening up of the MENA region.

The authoritarian political regimes, ironically, brought the new information technology in their respective countries as a means for advancing or sustaining economic growth. It is the compulsion of modernization and keeping up with the rest of the economically (and thus technologically) advanced societies that made the political leaders to facilitate and implement the infrastructures of the advanced new technology that was meant to connect these societies to the global village. Little did the leaders know that technology, especially, the Internet technology is a double-edged sword. While it played a positive role in upgrading their economic systems, there were unintended consequences of the ICT, one of which was impact of the technology on the political processes.

In a study following the Egyptian social uprising that had Tahrir Square as the fulcrum of the movement it was found that the average age of the participants was 28.5 years with an average education of 5.3 years of schooling. More importantly, 80% of them had Internet access at home and 52% of them had mobile phone with Internet connectivity (Tufekci and Wilson, 2012: 369). Lest, this might give an impression that it was a movement of the techies, it was also interesting to observe that 34% of the respondents had previously attended political rallies (Tufekci and Wilson, 2012: 369).

The above-mentioned study shows that a combination of various factors, that is, presence of a revolutionary crisis, an aging authoritarian leader, a determined public spearheaded by an educated, technology savvy middle-class protestors, and the revolution became an inevitability. The introduction of Facebook in Arabic had a direct and penetrating impact as it indigenized Facebook to local population, taking it beyond the urban elite.

Many writers, however, fall in the trap of technological determinism giving excessive weightage to the forces of Internet and mobile technology as the causes of the Arab revolutions. All revolutions cannot be reduced to a singular cause. There are always multiple factors causing social revolutions.

Technology, whether it is old media or new media, plays a complementary role.

In the Arab Spring the new and the old media joined forces to unleash the revolutionary potential (Khondker, 2011).

In Tunisia, when the fruit-seller, Mohammed Bouaziz resorted to self-immolation to protest the price hike and political repression, the event became national and eventually international news thanks to the combined effects of conventional media and the new media. Both television networks (such as Al Jazeera) and Facebook played a significant role in disseminating information and mobilizing the masses of protestors in Tunisia. Both virtual and real revolutionaries came out in droves to protest.

Mr. Khaled Koubaa, the president of Internet Society in Tunisia reported that of the 2000 registered Twitters hardly 200 were active users but before the revolution there were two million users of Facebook. "Social media was absolutely crucial," says Koubaa. "Three months before Mohammed Bouazizi burned himself in Sidi Bouzid we had a similar case in Monastir. But no one knew about it because it was not filmed. What made a difference this time is that the images of Bouaziz were put on Facebook and everyone saw it." (Beaumont, 2011). Stressing the role of the new media, Zeynep Tufekci (2011) makes the point that in Tunisia protest movements in 2008 were crushed without significant backlash effects. Part of the reason was that at that time there were only 28 000 Facebook users in Tunisia. In other words, the new media penetration was low. In December 2010 the news of self-immolation of Bouazizi in a small town was transmitted by the new media triggering mass protests.

In 2010 at the onset of Tunisian revolution to oust Ben Ali, the government was not caught in surprise. It was doing all it could to subvert the movement by shutting down or hacking websites, and so on. Ammar was in the process of stealing an entire country's worth of passwords. Ammar is known as the Internet censorship in Tunisia which did all it could do to disrupt Internet, the very tool that people are using for their activism becomes the very means by which their identities could be compromised (Madrigall, 2011).

The Tunisian revolution spilled over into Egypt. In both cases, the new media played a key role. The revolution was labeled as Jasmine revolution in part because it was not as violent as it could have been – as it unfolded in Libya months later. Given the dominant role of the new media, Tunisian and Egyptian revolutions are called Facebook or Twitter revolutions. There is no question that the social networking applications played a vital role in organizing and publicizing social protests.

Egypt

On 6 April 2008 when the Egyptian authority locked up Kareem Al Beheiri, a worker-turned labor rights activist and a blogger for allegedly instigating riots, the news was picked up by the conventional media but did not either hog the headlines or stay in the media limelight for long. The news of his protest was subsumed under the broader protests and demonstrations against the rising food inflation in Egypt and a number of other developing countries. Beheiri was tortured in custody and later released. His cyber-activism was part of a growing movement, albeit gradual, where new media was used effectively to mobilize people against a regime in power for three decades. Cybe-ractivism in Egypt had emerged since 2004 as it began to spread in other parts of the Middle East, marking the ushering of cyber-civil society, and a virtual replacement for the muzzled media. Since 2008, in the face of a global economic crisis and enhanced political repression, Egypt has seen growing protests where various means of new media have played a crucial role. How significant is the role is, however, a matter of some

contention. When Wael Ghonim, a Google Marketing executive and a key cyber-activist in Egypt, stated in an interview with CNN days before the ouster of Hosni Mubarak, "If you want to free a society just given them internet access" (Khamis and Vaughn, 2011: 1) he was probably exaggerating. Shortly after the stepping down of President Mubarak, in an interview with CNN, Ghonim said:

"I want to meet Mark Zuckerberg 1 day and thank him [...] I'm talking on behalf of Egypt. [...]. This revolution started online. This revolution started on Facebook. This revolution started [...] in June 2010 when hundreds of thousands of Egyptians started collaborating content. We would post a video on Facebook that would be shared by 60 000 people on their walls within a few hours. I've always said that if you want to liberate a society just give them the Internet. [...] (Huffington Post, 11 February 2011).

Gamal Ibrahim, a young parent named his daughter "Facebook" in recognition of the role of the social media played in bringing about the revolution (CNN, 2011). The exuberance is understandable given the timing of the birth of the baby girl. But it was at best an example of exuberance.

The Egyptian revolution was well organized, coordinated and civil (Abaza, 2011; Bamyeh, 2011) and in every step, the new media played its part. On 6 June 2010 Khaled Said, an Egyptian blogger was dragged out of a cybercafé and beaten to death by policemen in Alexandria, Egypt. The café owner, Mr. Hassan Mosbah, gave the details of this murder in a filmed interview which was posted online and pictures of Mr. Said's shattered face appeared on social networking sites. On 14 June 2010 Issandr El Amrani posted the details on blog site Global Voices Advocacy (*Global Voices Advocacy*, accessed on 24 June 2011). A young Google executive Wael Ghonim created a Facebook page "We Are All Khaled Said" that enlisted 350 000 members before 14 January 2011 (Giglio, 2011: 15).

Digital media in Egypt was used to tell the stories of police brutality, violence, and blatant injustice. In the end the new media helped protestors mobilize specific political outcome such as the removal of President Hosni Mubarak.

Control of conventional media made the role of new media more relevant. During the anti-Mubarak protests, an Egyptian activist put it succinctly in a tweeter message: "we use Facebook to schedule the protests, Twitter to coordinate, and YouTube to tell the world" (Global Voice Advocacy, 2010). However, to overstate the role of the new media may not be helpful. Surely, social network sites and Internet were useful tools, but conventional media played a crucial role in presenting the uprisings to the larger global community who in turn supported the transformations.

Saudi Arabia

On 17 June 2011 when Maha al-Qahtani, a 39-year-old Saudi woman decided to drive a car in Riyadh, a conservative city with her husband on the passenger seat, she was arrested and later released. However, her defiant protest received wider media attention and got traction in the new media. Ms. Manal al-Sharif, a 32-year-old IT consultant who set up a Facebook campaign called women2drive was arrested on May 21 (Allam, 2011). Her Facebook site remains a space for cyber-activism. A woman posted a picture of a woman driving wearing an *abaya* (a black dress covering the entire body) on this site which now carries the name of Manal and Bertha. More and more women and men – Saudis and non-Saudis – continue to express solidarity posting various contents one of which includes a message of support for the freedom–loving Saudi women from the US Secretary of State, Ms. Hillary Clinton. The Facebook site has also hosted a YouTube video clipping made from a television coverage of an earlier protest of similar nature when 47 Saudi women of cross-sections broke the taboo by driving their cars in defiance of government ban on 6 November 1990. New comments and contents taken from televised news are being added to the women2drive site almost on a daily basis.

Whether these gestures mark the new beginning or not is a moot question. The incipient cyber-activism in Saudi Arabia may not herald an Arab Spring, at least, in the near future. The outcomes of the spread of the new media are likely to remain uneven in different parts of the MENA region. Despite the uneven outcomes, there seems to be no holding back of the marches of cyber-activism in this region. Cyber-activism has been defined as "the act of using the internet to advance a political cause that is difficult to advance offline" (Howard, 2011: 145).

As the cyber-protests signify a window of change in Saudi society thanks to the new media – assorted tools such as Facebook, Twitter, and Internet-based communication – there are sites hosted by conservatives who call these women names, create their own websites thus compete for capturing space in the new media. It is impossible not to pay attention to two points. First, the new media is a tool, a means for and not the ends of social movement. Hence, the role of new media is contingent on the movements on the street. And the new media has been a contesting site of resistance for all the interest groups in the twenty-first century (Ho et al., 2002). The debate over the competing role of the conventional media versus new media has been a false debate. Even before the uprisings in the MENA region, it was argued that "horizontal networks of communication, such as the Internet, to have significant impact on the majority of the population they need to be relayed by the mass media, as was the case in the diffusion of the pictures of torture in the Abu Ghraib" (Arsenault and Castells, 2006: 303).

There is no question that social media played a significant role in the political movements in Tunisia and Egypt but one should not overstate the role. Conventional media, especially television medium such as Al Jazeera's role was crucial. However, the most important underlying factor was the presence of revolutionary conditions and the inability of the state apparatus to contain the revolutionary upsurge. In this schema, social media was a vital tool – a necessary condition – especially in the face of a muzzled conventional local media but a tool nevertheless. It was not a sufficient condition. Stressing the role of Al Jazeera in Arabic as well as the BBC, France 24, Al Hiwar, and other channels in presenting the news of the Arab Spring, Manuel Castells states, "Al Jazeera has collected the information disseminated on the Internet by the people using them as sources and organized groups on Facebook, then retransmitting free news on mobile phones. Thus was born a new system of mass communication built like a mix between an interactive television, internet, radio and

mobile communication systems. The communication of the future is already used by the revolutions of the present. [emphasis in the original]."

Castells continues: "Obviously communication technologies did not give birth to the insurgency. The rebellion was born from the poverty and social exclusion that afflict much of the population in this fake democracy, …" (Castells interview posted on http://globalsociology.com/2011/02/07/the-sociology-elders-on-the-social-movements-in-tunisia-and-egypt/)

The new media's role can be likened with the historic role of print media in fostering nationalism through what Benedict Anderson (1991) calls 'print capitalism,' or the role of literacy in raising consciousness in the prerevolutionary eighteenth century France as the works of Robert Darnton (1982) and John Markoff (1986) illustrate. Markoff shows that in regions with higher literacy, revolutionary activities were more organized compared to those in low literacy regions. However, revolution did not wait for the even spread of literacy.

Social network or not, Tunisia, Egypt, Libya, Syria, Yemen, and Bahrain were ready for revolutionary movements due to an assortment of politico-economic conditions. Similarly, to what extent these revolutions will be successful or not depends on several factors – some known and some yet unknown, in which the social network may play a supportive role at best.

Like the 'revolution,' the discussion on social or new media or the role of ICT, generated more heat than the light. The revolution itself was not revolutionary in some cases as it replaced one military dictator with another with some diversionary electoral exercises to purchase some kind of short-lived political legitimacy. At least, that was the fate of the largest of the 'Arab Spring' countries, namely, Egypt. Ignoring the details of 2 years of political drama where in a popular vote, the Muslin Brotherhood supported leader, Dr. Mohamed Morsi, a US trained engineer, representing the Muslim Brotherhood was elected president of Egypt in June 2012, and as his regime played roughshod in implementing shariah-compliant laws that upset sections of the Egyptian middle class who were interested in seeing economic reforms rather than imposition of religious values. Many who supported the revolution quickly became disillusioned of Morsi's pro-Islamic regime.

Throughout these processes from the removal of the dictator Hosni Mubarak in January 2011, electing Mohamed Morsi in June 2012, and removing Mr. Morsi in July 2013, social media continued to play a role. And as Mohamed Morsi became president on the back of a new media generated social uprising, he turned against the new media and the bloggers reciprocated by turning against his pro-Muslim Brotherhood regime. A high-profile blogger Ahmed Doume was sentenced to 6 months prison term by Morsi regime. His was one of the two dozen cases of 'insulting the president' in the first 200 days of Morsi rule. Apparently, this was four times of the similar cases in the 30 years of Mubarak rule (VOA, 2013). As the new military leader General Sisi helped himself to the helms of affairs in July 2013 and consolidated his power through a show election in early 2014, which many Egyptians stomached despite its questionable legitimacy, because a large number of Egyptians, barring the Muslim Brotherhood, were so dismayed by the rule of President Morsi. However, Field Marshal Sisi too turned against the freedom of speech and matched the intolerance of the Muslim Brotherhood by his regime's intolerance of Muslim Brotherhood and the free use of media. One of the early casualties of his regime – as far as freedom of expression was concerned – the satirical television show of Dr. Bassem Youssef, the Jon Stewart of Egypt (Hellyer, 2014).

It is possible to see the Egyptian political drama as a fight on or off against the new media. When new media serves political power, it is acceptable and when it does not, it remains an instrument of social disruption and is suppressed. Political authorities by supporting or opposing recognize the potential power of the new media. An analysis of the blogosphere in Egypt and Saudi Arabia in 2013–14 showed a growing blogging fatigue especially of the left and the liberal voices. The Muslim Brotherhood blogs remain visible. There has also been a change of platforms from blogging to Twitters, especially, in Saudi Arabia (Othman, 2014).

Conclusions

From the very onset of the spread of ICT, many saw it as a tool of social and economic progress. Astute social scientists were skeptical seeing the new penetrating technology an invasion of privacy paving the emergence of surveillance society. Information technology, in the hands of the government, can be a source of control, while in the hands of the public it may be a tool for emancipation.

The spread of ICT in the emerging digital revolution, technology unwittingly played a role in Tunisia's democratization, spreading antiauthoritarian movements in Egypt and other MENA countries, which came to be known as the Arab Spring.

While some writers grant social media too much credit, for example, naming Iran's Green Movement the 'Twitter Revolution' and labeling the Arab Spring the 'Facebook Revolution,' others, however, were somewhat cautious and saw social media playing a complimentary role. Social media in tandem with other electronic media supported the social movement and political mobilization on the street, which had a deeper root in the social structures and political developments. Surely, in the political uprisings of 2011–12 new media's role in creating a virtual civil society in the absence of a real or off-line civil society needs to be recognized without falling in the trap of technological determinism. At the same time one has to consider the role of the new media in promoting right-wing movements and forces such as the Islamic State of Iraq and the Levant (ISIL) that threaten democracy and peace. The new media technology is Janus-faced.

See also: Information Society; Media and Social Movements; Mobile Communication; New Media and Democracy in the Arab World; Social Media; Social Protest and New Media; Social Protest.

Bibliography

Abaza, Mona, 2011. Revolutionary moments in Tahrir Square. Global Dialogue. Issue 3. May, International Sociological Association.

Abdulla, Rasha, 2011. The revolution will be tweeted. The Cairo Review of Global Affairs 3. http://www.aucegypt.edu/GAPP/CairoReview/Pages/articleDetails.aspx?aid=89.

Allam, Abeer, 2011. Saudi female drivers defy rules. Financial Times. June 17, 2011.
Anderson, B., 1991. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London.
Arab Social Media Report, vol. 1, no. 1, 2011. Dubai School of Government, Dubai.
Arab Social Media Report, vol. 2, no. 1, 2012. Dubai School of Government, Dubai.
Arab Social Media Report, sixth ed., 2014. Dubai School of Government, Dubai.
Arsenault, Amelia, Castells, Manuel, 2006. Conquering the minds, conquering Iraq: the social production of misinformation in the United States: a case study. Information, Communication and Society 9 (3), 284–308.
Bamyeh, Mohammed, 2011. The egyptian revolution: first impressions from the field. The Asia-Pacific Journal Japan Focus. http://japanfocus.org/-Mohammed-Bamyeh/3486.
Beaumont, Peter, Februart 2011. The Truth about Twitter, Facebook and the Uprisings in the Arab World. The Guardian.
Beniger, James, 1989. The Control Revolution: Technological and Economic Origins of the Information Society. Harvard University Press, Cambridge, MA.
Castells, Manuel, 1996. The Rise of the Network Society. Blackwell, Cambridge, MA.
Castells, Manuel, 2009. Communication Power. Oxford University Press, New York.
Castells Interview Posted On. http://globalsociology.com/2011/02/07/the-sociology-elders-on-the-social-movements-in-tunisia-and-egypt/ (accessed on 25.06.11.).
CNN, 2011. http://articles.cnn.com/2011-02-21/world/egypt.child.facebook_1_facebook-page-wael-ghonim-social-media?_s=PM:WORLD (accessed on 26.06.11.).
Darnton, Robert, 1982. The Literary Underground of the Old Regime. Harvard University Press, Cambridge, MA.
Eltantawy, Nahed, Wiest, Julie, 2011. Social media in the egyptian revolution: reconsidering resource mobilization theory. International Journal of Communications 5.
Etling, Bruce, Kelly, John, Faris, Robert, Palfray, John, 2009. Mapping of Arabic Blogosphere: Politics, Culture and Dissent. Berkman Center for the Study of Internet and Society, Harvard University, Cambridge, MA.
Giglio, M., February 21, 2011. The Facebook Freedom Fighters. Newsweek.
Global Voice Advocacy, 2010. Website: http://advocacy.globalvoicesonline.org/ (accessed on 23.06.11.).
Helleyer, H.A., June 3, 2014. The Silencing of Egypt's Jon Stewart. Foreign Policy.
Ho, Kong, Chong, Baber, Zaheer, Khondker, Habibul, 2002. 'Sites' of resistance: alternative websites and state – society relations. The British Journal of Sociology 53 (1), 127–148.
Howard, Philip N., 2011. The Digital Origins of Dictatorship and Democracy: Information Technology and Political Islam. Oxford University Press, Oxford.
Kalathil, Shanti, Boas, Taylor, 2003. Open Networks, Closed Regimes: The Impact of the Internet on Authoritarian Rule. Carnegie Endwment for International Peace, Washington, DC.
Khamis, Sahar, Vaughn, Katherine, 2011. Cyberactivism in the egyptian revolution: how civic engagement and citizen journalism tilted the balance. Arab Media and Society (Online Journal). http://www.arabmediasociety.com/?article=769 (accessed on 24.06.11.).
Khondker, Habibul H., 2011. Role of the new media in the Arab Spring. Globalizations 8 (5), 675–679.
Madrigall, Alexis, 2011. The inside Story of How Facebook Responded to Tunisian Hacks. The Atlantic, 24.01.11. http://www.theatlantic.com/technology/archive/2011/01/the-inside-story-of-how-facebook-responded-to-tunisian-hacks/70044.
Markoff, John, 1986. Some effects of literacy in eighteenth century France. Journal of Interdisciplinary History XVII (2), 311–333.
Othman, Dalia, 2014. "On Post Arab Revolution: What Social Media Is Telling Us" Presented at Berkman Center for Internet and Society. Harvard University. Available at: http://cyber.law.harvard.edu/interactive/events/luncheons/2014/05/othman.
Poster, Mark, 1995. Cyberdemocracy: Internet and the Public Sphere. http://www.humanities.uci.edu/mposter/writings/democ.html (accessed on 22.06.11.).
Reilly, Katie, 2013. India ends the Telegram, but Internet Access Is Not Close behind. Pew Research Center, 15.07.13. http://www.pewresearch.org/fact-tank/2013/07/15/india-ends-the-telegram-but-internet-access-is-not-close-behind/.
Robertson, Roland, 2007. Open societies, closed minds: exploring the ubiquity of suspicion and voyeurism. Globalizations 4 (3).
Tufekci, Zeynep, 2011. Technosociology. Website: http://technosociology.org/?page_id=2 (accessed on 22.06.11.).
Tufekci, Zeynap, Wilson, Christopher, 2012. Social media and the decision to participate in political protest: observations from tahrir square. Journal of Communication. 62, 363–379.
UNESCO, n.d. http://www.unesco.org/new/en/communication-and-information/flagship-project-activities/unesco-and-wsis/about/unesco-in-tunis-phase/ (accessed on 06.07.14.).
Voice of America news, 2013. http://www.voanews.com/content/egyptian-blogger-convicted-of-insulting-president-morsi/1674500.html.

New Religious Movements

Eileen Barker, London School of Economics, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'New religious movements' (NRMs) is a term used to describe minority religions that have recently become visible in a society. They have appeared throughout the world and throughout history, and they have differed from each other with respect to their beliefs, practices, and lifestyles, as well as the effects they have on their members and on society. However, they do tend to consist of converts; to have a membership that is atypical of the rest of the population; to have a charismatic leader; to change more rapidly and radically than older, more established religions; and to evoke fear, suspicion, and, not infrequently, antagonism from nonmembers.

Definitions

The concepts of 'newly arising religion,' then 'new religion,' entered Japanese vocabulary after the Second World War when restrictions imposed on religions other than state Shinto were lifted (Shimazono, 2004; Mullins et al., 1993). Around the late 1960s, Western scholars began to use the term 'new religious movement' (NRM) to describe (1) religions that were new to the West, although, like some forms of Tibetan Buddhism, they could have existed for hundreds of years in another part of the world; and (2) minority religious organizations and spiritual communities that might be referred to as a 'cult' or 'sect,' in order to distinguish them from a 'church' or 'denomination.' However, as the terms 'cult' and 'sect' were increasingly used in popular parlance to denote a negative evaluation, Western scholars had adopted the term NRM in an attempt to distance themselves from the blanket condemnations associated a priori with these labels and to present their analyses from a more objective perspective. This move led to accusations from sections of the media and anticultists (discussed in this article) that the scholars' use of the concept was an insult to 'genuine religion' and little more than an attempt to whitewash the movements – accusations that, by the turn of the twenty-first century, had resulted in some scholars and the movements themselves preferring to use terms such as 'alternative religions' or 'minority religions.'

There are, however, several other problems related to the term NRM. First, it has not been used to refer to all new religious groups (e.g., the United Reformed Church would not be called an NRM, although it came into being only in 1972 as the result of a merger between English Presbyterians and Congregationalists). Second, some groups classified as NRMs protest that they are not new but can claim a heritage of hundreds of years; the International Society for Krishna Consciousness (ISKCON) traces its lineage through the sixteenth-century monk Chaitanya Mahaprabhu. Interestingly, in Japan, movements such as Mahikari, Shinreikyō, and Agonshū that have appeared since 1970 are now referred to as 'new new religions' (*shinshinshūkyō*); indeed, some Japanese scholars are now calling movements that have emerged in the twenty-first century 'new new new religions.'

Third, several groups deny that they are religions, many preferring to classify themselves as spiritual, philosophical, or educational; and some, such as the Raelians, are avowedly atheistic. NRMs are, however, frequently defined rather loosely as new insofar as they have become visible in their present form since World War II, and religious insofar as they address at least some of the questions of ultimate concern that traditionally have been addressed by more established religions – questions such as 'What is the meaning of life?', 'Is there life after death?', and 'Is there a God?'.

Another way of defining a new religion is as a group or movement that consists predominantly of a first-generation membership. This has the advantage of allowing one to identify a relatively recognizable group that may share certain characteristics with other first-generation movements. It also helps to alert us to the fact that NRMs can change rapidly within a generation (discussed further in this article). There can, however, be difficulties when deciding how to classify movements that have existed for decades or even centuries as proselytizing movements; for example, 70% of the worldwide membership of the Church of Jesus Christ of Latter-day Saints (popularly known as the Mormon Church) are converts, although the Church itself has been around for nearly 200 years.

Numbers

The total number of NRMs is unknown and will, of course, differ according to the precise definition that is used. In the West, there are around 1500 groups, and more if all the 'self-religions' – that is, religious movements that (also) aim at self-improvement – are included (*see* Religiosity: Modern). Several thousand more may be found in Latin America, in the former Soviet Union, and throughout Asia, including the Indian subcontinent. It has been estimated that over 10 000 new religions are to be found in Africa, although many of these are referred to more frequently as African Independent Churches.

The number of members is even more difficult to assess, and estimates can vary by millions. Among the relatively few large international movements: in 2011, Sōka Gakkai was claiming around 12 million members worldwide; between 8 and 14 million persons are said to have taken one or more of the courses offered by Dianetics or the Church of

Scientology, but the vast majority of these people will no longer consider themselves to be Scientologists; it has been claimed that 6 million people have been trained as Transcendental Meditators; and Cao Đài also claims 6 million members, most of whom are ethnic Vietnamese. If nineteenth-century sects are defined as new religions, several of these have tens of thousands of members or more worldwide: Seventh-day Adventists claim over 15 million worldwide; Mormons claim over 14 million; Jehovah's Witnesses claim 7.5 million; the Salvation Army just over 1.5 million; and the Christian Community, popularly referred to as the 'Exclusive Brethren,' is thought to have around 43 000 members worldwide. There are, however, probably no more than a few hundred NRMs that have emerged in the West since the 1950s that have more than 1000 adherents; the majority are small, local groups with no more than 100 or so followers at any one time.

The difficulty of obtaining accurate figures is confounded by the fact that most movements have several levels of membership, ranging from the fully committed, who may live with and work for their movement, to 'sympathizers' who might belong to more than one group (especially within the New Age or Human Potential movements). Furthermore, both the NRMs and their opponents tend to exaggerate membership figures, the former because they wish to be seen as more successful than they really are. Opponents of NRMs often want to exaggerate the threat that the movements pose, and many opponents adhere to a brainwashing thesis (which claims members are recruited using irresistible and irreversible mind-control techniques), which leads them to ignore or deny the fact that large numbers of converts leave after a short period of membership (Barker, 1989).

Variety

Perhaps the most important generalization to be made about NRMs is that they are quite diverse. There have, of course, been new religions throughout history; early Buddhism, early Christianity, early Islam, and early Methodism were NRMs. But while most of the movements that arose in previous times were offshoots of the prevailing traditions in their society of origin, the range of traditions, permutations, and combinations to be found within the current wave of NRMs is without precedence. This may in part be explained by dramatically increased social and geographical mobility, the phenomenal expansion of the mass media, and various other processes subsumed under the rubric of postindustrial globalization. Thus, for example, one can find people brought up in the Judeo-Christian tradition deciding to join the Brahma Kumaris, Nichiren Shoshu, Mahikari, Falun Gong, or Hizb ut-Tahrir – NRMs drawn, respectively, from the Hindu, Buddhist, Shinto, Qi Gong, and Islamic traditions. There are, moreover, NRMs that have derived their inspiration from the beliefs and practices of Wicca, Paganism, Occultism, and Esotericism; others have turned to modern philosophies and ideologies, psychological and psychoanalytic theories, astrology, and, indeed, a variety of secularisms. Yet others, such as the Unarius Academy of Science, the Raelians, the Aetherius Society, and the Church of Scientology, owe much to ideas found in works of science fiction; and there are those that would seem to defy membership with any kind of tradition, or to be composed of a thoroughly syncretistic bricolage of traditions and innovative notions, two very different but colorful examples being L'Aumisme Religion Universelle de l'Unité des Visages de Dieu (also known as Mandarom) and the Nuwaubians (also known as the Holy Tabernacle Ministries) (Barker, 1989; Dawson, 1996; Melton, 1992; Miller, 1995). Finally, it should be added that there are now growing numbers of virtual new religions that exist purely or mainly on the Web (the Church of the Flying Spaghetti Monster with its membership of Pastafarians is an oft-cited example) and a series of movements that have been referred to as 'invented religions' or 'hyperreligions,' which include movements that have developed out of various forms of popular fiction, examples being Discordianism, The Church of All Worlds, Jediism, and Matrixism (Possamai, 2012).

Many of the better known NRMs (such as the Children of God/The Family International, the Church of Scientology, the Twelve Tribes, The Way, and the Church Universal and Triumphant) originated in the United States. Other movements developed in various parts of Western Europe (e.g., The Emin, Exegesis, Findhorn, and the Jesus Army in England; Damanhur in Italy; Smith's Friends in Norway; the Revelation of Arès in France; the Lou Movement in the Netherlands; and the Antoinistes in Belgium); and, particularly after the fall of the Berlin Wall in 1989, several emerged in Eastern Europe and the former Soviet Union (e.g., Vissarion's Church of the Last Testament in Siberia, the New Jerusalem in Romania, and Mariya Devi Khristos's White Brotherhood in Ukraine). Moreover, a large number of movements that are now familiar throughout the world originated in the Indian subcontinent (e.g., ISKCON, Sai Baba, the Rajneeshee/Osho sannyasins, the Brahma Kumaris, and Ananda Marga).

African new religions that have hit the headlines include the Lord's Resistance Army and the Movement for the Restoration of the Ten Commandments of God, but there are thousands of others, particularly in Sub-Saharan Africa, many of which, such as Cherubim and Seraphim as well as the Embassy of the Blessed Kingdom of God for all Nations, have migrated to the West and other continents around the world (Adogame et al., 2012; Clarke, 1998). The Rastafari movement arose in Jamaica, worshipping the emperor of Ethiopia, Haile Selassie. Latin American NRMs include various varieties of Candomblé and the Universal Church of the Kingdom of God from Brazil and Santa Muerte from Mexico (Torre Castellanos, 2012). From Indonesia, we have SUBUD; from Malaysia, there is the Sky Kingdom sect, sometimes known as the Teapot Cult because it built a giant teapot to symbolize its belief in the purity of water. Vietnam gave rise to Cao Đài, Australia to the Jesus Christians and the Magnificat Meal Movement, and New Zealand to Centrepoint.

As already intimated, Japan has been home to hundreds of movements, including Sōka Gakkai, Shinnyoen, Tenrikyo, Sūkyō Mahikari, Risshō Kōsei-kai, and the infamous Aum Shinrikyo, which now operates under the name of Aleph (Shimazono, 2004). The third most popular faith in Taiwan (after Buddhism and Taoism) is I-Kuan Tao (Yõguàn Dào or simply IKT), which originated in twentieth-century mainland

China, where it was outlawed (Lu, 2008). Other important Taiwanese NRMs include the True Buddha School and the Tzu Chi Foundation, known for its charitable work throughout the world. Throughout its history, mainland China has also spawned new religions of various forms, commonly derived from Confucianism; Taoism; Buddhism; Chinese 'folk religions'; and, more recently, the 'Qigong Fever' (Palmer, 2007). Following the partial relaxation of religious persecution in the People's Republic of China after the 'Cultural Revolution,' there has been a dramatic growth in unregistered religions that meet illegally for worship. So long as they are not seen as a threat by the state, these might be allowed to exist, but they are always in danger of being officially condemned as an 'evil cult,' with their members facing imprisonment and/or, most notably in the case of Falun Gong, 'reeducation.'

Numerous new religions have emerged from the Islamic traditions of various countries: the Nation of Islam began in the United States, Hitzb ut-Tahrir was founded by a Palestinian in Israel, the now-banned Al-Muhajiroun emerged in the United Kingdom, Tablighi Jamaat came out of India, the Gülen movement originated in Turkey, and the People's Mujahedin of Iran (MEK) operated out of Iraq. Some movements that are referred to as NRMs remain within the folds of their religious traditions. Thus, while the Apostles of Infinite Love and House of Prayer are rejected by the Vatican, Communione e Liberazione, Fokolare, the Neocatechumenal Way, and Opus Dei are accepted as part of the Church.

Some NRMs, such as the Unification Church, have relatively elaborate and systematic theologies; others consist of what might appear to be a jumble of uncoordinated pronouncements. Practices include chanting, meditation, ritual magic, dance, formal liturgy, and group therapy. Some members of NRMs live independently in high-rise apartments; others live together in rural communes (Miller, 1995). Lifestyles may be ascetic or hedonistic. Attitudes toward women, children, and the outside world vary. Some groups, such as the Rajneeshee movement, have indulged in promiscuity and sexual orgies; others, such as the Brahma Kumaris, practice celibacy. Some movements are rich; others are poor; sometimes the leaders become rich while the grassroots members become poor. The movement's organization may be democratic or authoritarian. The leader may be seen as a guru, teacher, prophet, messiah, or god. He (or sometimes she) might, like Sai Baba, perform miracles; or might, like Sun Myung Moon, claim to receive revelations directly from God – or, like Osho, claim to be a god. Alternatively, the leader might, like JZ Knight and Elizabeth Clare Prophet, channel messages from Ascended Masters, or, like Ruth Norman, Raël, and George King, they might be in touch with aliens from another planet. Some NRMs are destructive, and others benign. Most, like older, more traditional religions, can have positive effects some of the time for some of their members and negative effects at other times for other members.

Attempts to classify NRMs have been largely unsuccessful, the sheer variety having made well-nigh impossible any elegant typologies such as those that Ernst Troeltsch, Milton Yinger, and, perhaps most famously, Bryan Wilson (1970) elaborated for the largely Western, Christian-based sects of the nineteenth century. Possibly the most widely adopted attempt has been Roy Wallis's (1984) distinction between 'world-rejecting,' 'world-affirming,' and 'world-accommodating' movements.

Characteristics

Despite the enormous variety, there are, nevertheless, certain characteristics that are commonly found among NRMs merely because they are both new and religious. First, new religions, insofar as they are composed of first-generation members, are likely to display singular enthusiasms, because converts to any religion tend to be more zealous than those born into their religion. Second, the membership is usually atypical of the general population; for example, although in the past NRMs frequently appealed to the socially and/or economically oppressed, the NRMs that became visible in the West in the 1970s appealed disproportionately to young adults from middle-class backgrounds. The movements had, thus, a zealous, healthy, well-educated, but inexperienced membership with few dependents or responsibilities. Third, founders of new religions are frequently accorded a charismatic authority by their followers and are, thus, relatively unconstrained by rules or tradition and, consequently, both unpredictable and unaccountable to anyone.

Fourth, NRMs often embrace dichotomous worldviews that permit little qualification. The world is divided unambiguously into 'them' versus 'us' people, 'true' versus 'false' beliefs, and 'good' and 'godly' versus 'evil' and 'satanic' morals and practices, and time is split between 'before' versus 'after' significant events such as conversion or the coming apocalypse. Fifth, NRMs, given that they present an alternative to the status quo, are commonly regarded with suspicion and/or antagonism by the host society. Understandably, these fears have been reinforced by horrific incidents such as the mass suicides and murders associated with the People's Temple (Guyana, 1978), the Branch Davidians (Waco, TX, 1993), the Solar Temple (Quebec, Switzerland, and France, 1994–95), Aum Shinrikyo (Tokyo, 1995), Heaven's Gate (San Diego, CA, 1997), the Movement for the Restoration of the Ten Commandments of God (Uganda, 2000), and Al Qaeda (in the 9/11 attacks, New York and Washington, DC, 2001).

Changes

A sixth characteristic of NRMs is that they are likely to change far more rapidly and radically than older, more established religions. Taking demographic factors alone: youthful converts get older, and the arrival of a second generation will mean not only that resources such as time and money have to be devoted to their care and socialization, but also that they are likely to question many of the beliefs and practices their parents had adopted. As children cannot be expelled as easily as stroppy converts, the movement could have to adapt its beliefs and practices. Furthermore, charismatic leaders die, their authority usually being replaced by a more predictable and accountable leadership. Many of the sharp boundaries that distinguished NRMs from the rest of society can become permeated by qualification and accommodation, and the movements that survive tend, though not always, to merge less noticeably into

the pluralistic smorgasbord of contemporary democracies through a process commonly referred to by sociologists as 'denominationalization.'

Reactions to NRMs

Reactions to NRMs vary across both individuals and social groups. Such reactions can range from celebratory delight, through cautious curiosity, dismissive distaste, and outright antagonism to lethal persecution. Most people have their opinions formed by reports in the mass media, which, generally speaking, tend to select the unusual, exotic, and sensational, with 'bad stories' being viewed as 'good stories.' Frequently, the media obtain their information from worried relatives, disillusioned former members, and/or the so-called anticult movement (ACM), a number of organizations devoted to exposing the perceived dangers of NRMs in an attempt to warn others and/or have the movements controlled or banned (Bromley and Hadden, 1993). Originally, many of the members of the ACM were worried relatives of converts who believed that their loved ones were in danger, but they were soon joined by others, including former members and persons such as lawyers, private detectives, politicians, mental health professionals, and counselors who had developed a specialist interest in NRMs. Anticultist activities have ranged from persons who have engaged in forcible 'deprogramming' for thousands of dollars (a practice now rarely carried out in the West, although still prevalent in Japan) to those who merely attempt to educate the general public through literature, lectures, and counseling. By 2011, the International Cultic Studies Association (originally the American Family Foundation) had become the largest and most professional such organization, but numerous other organizations have played a role in opposing NRMs around the world, examples being the Dialogue Centre in Denmark; the French Interministerial Mission for Monitoring and Combatting Cultic Deviances (MIVILUDES); the Israeli Center for Victims of Cults; the Saint Ireneus of Lyons Informational Consultative Center, affiliated to the Russian Orthodox Church; and the Chinese Association for Fighting the Evil Cults. Other 'cult-watching groups' are more scholarly in their approach, and these include CESNUR (based in Turin); the International Society for the Study of New Religions (based in Sweden); INFORM (based at the London School of Economics); and Le Center Intercantonal d'Information sur les Croyances (in Geneva), with possibly the largest collection of material related to NRMs being housed at the University of California, Santa Barbara, under the auspices of the Institute for the Study of American Religion. Peer-reviewed journals specializing in the movements include *Nova Religio*, *The International Journal for the Study of New Religion*, and *The Cultic Studies Review*.

Several government enquiries have been commissioned to look into the activities of NRMs. Russian, Israeli, French, and, to a lesser extent, Belgian reports have tended to rely either solely or predominantly on anticult resources and to recommend that the movements should be restricted, while Swedish, Dutch, and Canadian reports have paid more attention to scholarly research. Countries that require religions to register with the government in order either to gain certain privileges, or to be allowed to practice at all, usually require the religion to have a minimal number of members (Slovakia has a requirement of 20 000 members) and to have been in the country for a minimum period of time, both of which criteria tend, obviously enough, to militate against NRMs.

Throughout history there have been, and in some cases there still are, societies (such as Saudi Arabia, North Korea, and Albania under the rule of Enver Hoxha) in which if NRMs were to operate at all, they would have to do so secretly underground. Elsewhere, state reactions to NRMs can be divided roughly into two approaches. On the one hand, in countries such as France, Russia, and China, the state contends that it has to protect its citizens in advance by introducing special legislation that will prevent 'cults,' 'sects,' and/or NRMs from inflicting harm. Alternatively, societies like the United States, Canada, Britain, and Scandinavia tend toward the position that there should not be special laws related to NRMs – although if any of their members are found guilty of criminal malpractice, they are subject to the law in exactly the same way as is any other citizen.

See also: Charisma and Charismatic; Charisma, Social Aspects of; Millennialism; Religion and Youth; Religion, History of; Religion, Sociology of; Religion: Evolution and Development; Religion: Morality and Social Control; Weberian Social Thought, History of.

Bibliography

Adogame, A., Chitando, E., Bateye, B. (Eds.), 2012. African Traditions in the Study of Religion in Africa: Emerging Trends, Indigenous Spirituality and the Interface with Other World Religions. Ashgate, Abingdon.

Barker, E., 1989. New Religious Movements: A Practical Introduction. HMSO, London.

Bromley, D., Hadden, J., 1993. The Handbook on Cults and Sects in America (Parts A & B). JAI Press, Greenwich, CT.

Clarke, P.B. (Ed.), 1998. New Trends and Developments in African Religions. Greenwood Press, Westport, CT.

Dawson, L. (Ed.), 1996. Cults in Context: Readings in the Study of New Religious Movements. Canadian Scholars' Press, Toronto.

Lu, Y., 2008. The Transformation of Yiguan Dao in Taiwan: Adapting to a Changing Religious Economy. Lexington, Lanham, MD and Plymouth, UK.

Melton, J.G., 1992. Encyclopedic Handbook of Cults in America, rev. ed. Garland, New York and London (original ed. 1986).

Miller, T. (Ed.), 1995. America's Alternative Religions. State University of New York, Albany, NY.

Mullins, M.R., Shimazono, S., Swanson, P.L. (Eds.), 1993. Religion and Society in Modern Japan. Asian Humanities Press, Berkeley, CA.

Palmer, D.A., 2007. Qigong Fever: Body, Science, and the Politics of Religion in China, 1949–1999. Hurst, London.

Possamai, A. (Ed.), 2012. Handbook of Hyper-Real Religions. Brill, Leiden.

Shimazono, S., 2004. From Salvation to Spirituality: Popular Religious Movements in Modern Japan. Trans Pacific Press, Melbourne.

Torre Castellanos, R., 2012. Religiosidades nómadas. Creencias y prácticas heterodoxas en Guadalajara. Consejo Estatal de Ciencia y Tecnologia de Jalisco (Coecytal), Jalisco, Mexico.

Wallis, R., 1984. The Elementary Forms of the New Religious Life. Routledge & Kegan Paul, London.

Wilson, B.R., 1970. Religious Sects: A Sociological Study. Weidenfeld and Nicholson, London.

New Urbanism

Jill L Grant, Dalhousie University, Halifax, NS, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

New Urbanism is a community design movement that originated in the United States in the 1980s. It has since spread across North America and also influenced urban development and planning in other parts of the world. It promotes mixed use, mixed housing types, pedestrian-oriented streets, and high design standards. While some criticize New Urbanism for practitioners' claims that good design can build social cohesion and social capital, supporters of the movement recognize its significant contribution to designing environments intended to enhance urban quality of life.

New Urbanism is a community design movement that began with localized responses to the sprawling monotony of American development practices in the 1980s (Talen, 2005); by the turn of the millennium, New Urbanism had become extremely influential in North American planning and urban design practices and had fused with ideas related to sustainable development, smart growth, and healthy communities to dominate planning principles for urban infill and suburban green-field development (Grant, 2009). Commonly associated with principles of compact form, mixed use, mixed housing types, higher urban densities, defined center and edge, walkability, and transportation alternatives, New Urbanism has influenced policy at all levels of planning and land development, especially in fast-growing regions.

The Origins of New Urbanism

New Urbanism draws on several historical trends in community planning. Two late nineteenth century/early twentieth century movements had significant impacts on the profession of planning, and ultimately on the new design approach. With its elegant public spaces, powerful urban geometries, and classical building forms, the City Beautiful Movement, led by Daniel Burnham, contributed to rethinking civic design. It proved especially influential for American cities, stimulating grand urban center plans for cities such as Chicago; Washington, DC; and Cleveland. The Garden City Movement linked to Ebenezer Howard and Raymond Unwin inspired visions of medium-density self-contained communities blending the best of town and country: it provided the impetus for places like Letchworth Garden City in the United Kingdom; Radburn, New Jersey (USA); and Don Mills, Ontario (Canada), and became the driving concept behind suburban development for much of the twentieth century. Most important for recent design models, however, were mid-twentieth century theorists highly critical of planning and of the products of garden suburb development. In her analysis of great American cities, Jane Jacobs (1961) argued that mixed use and community diversity stimulated urban vitality while conventional planning practices and processes sucked the vibrancy from cities. Léon Krier's (1978) focus on visual coherence and historical design antecedents, and his arguments for cloning traditional urban patterns for new growth, influenced two young American architects who would play key roles in the New Urbanism movement.

In 1982, developer Robert Davis hired Andrés Duany and Elizabeth Plater-Zyberk, who had studied briefly with Krier, to prepare a plan for Seaside, a beachfront community on Florida's Gulf Coast. The design for Seaside drew on regional architectural precedents to create a compact, walkable, mixed-use community with beautiful civic spaces (Katz, 1994). The design code book for Seaside established clear standards to guide urban form to create coherence and elegance in the spaces and buildings of the community. Seaside's site plan showed historic allegiance not only to the City Beautiful designs of the late nineteenth and early twentieth centuries but also to the Beaux Arts and Baroque plans of the seventeenth and eighteenth centuries. With its photogenic pastel homes, white picket fences, and stunning civic architecture, Seaside received considerable media attention. Originally intended to be a mixed-income community with a selection of affordable housing, Seaside's success turned it into a resort destination serving the affluent.

Duany and Plater-Zyberk followed up with another award-winning community design for The Kentlands in Gaithersburg, Maryland, this time drawing on New England architectural precedents (**Figure 1**). On a large site near Washington, DC, The Kentlands featured an ambitious plan with a town center, narrow streets and service lanes, and robust mix of housing types. The project proved highly successful in the market place, although the developer's plan for the regional shopping center undermined the functioning of the main street. Housing types and styles that appeared in The Kentlands soon influenced suburban developments in many other parts of the world. By the early 1990s, neotraditional town planning was drawing considerable media and professional attention as an alternative to the stereotypical blandness and placelessness of conventional suburban form. Images of The Kentlands and Seaside provided a template for small town urbanity and set the standard for the new approach.

On the west coast of the United States, Peter Calthorpe, Douglas Kelbaugh, and their colleagues began experimenting in the 1980s with pedestrian pockets and transit-oriented development (Calthorpe, 1993; Van der Ryn and Calthorpe, 1986). These planners sought to design sustainable communities that would coordinate land use and transportation options to reduce car use and energy demand. They experimented with urban infill projects, putting new housing into inner city areas made redundant by land use change. They proposed suburban developments that could be served by public transportation

Figure 1 The Kentlands established an early benchmark for New Urbanism.

nodes. Their work gained influence among planners in large cities where deindustrialization was increasingly opening opportunities for urban redevelopment where existing transportation infrastructure could support intensification. The publication of the World Commission on the Environment and Development's (1987) report, *Our common future*, accelerated international interest in the concept of sustainable development, aimed at achieving economic development while protecting environmental quality and resources for future generations. In concert with many planners for large urban centers in Canada and Europe, the transit-oriented designers argued that resolving some of the environmental and economic problems with (sub)urban development would require increased densities and a regional planning approach: a new kind of urbanism.

In 1991, the advocates of neotraditional town planning and of transit-oriented development agreed to collaborate on what they came to call the 'New Urbanism.' They issued a joint declaration – the Ahwahnee principles – and in 1993 formed the Congress for the New Urbanism (CNU). With annual conferences and regular publications, the CNU became an important mechanism for linking and educating practitioners and spreading the messages of the movement. Developers inclined to adopt innovative community design approaches increasingly turned to New Urbanism design strategies as a way of adding cachet to their projects (Ewing, 1996). House designs emulating historic models appeared with growing frequency in new developments. Skeptical at first, professional planning organizations gradually began to develop greater interest in the planning approach.

By the late 1990s, New Urbanism principles (Table 1) were influencing community planning theory and practice in many parts of North America and Europe. New Urbanism embraced discussions about intensification (that is, increasing densities for urban efficiency and vibrancy) and compact form (that is, reducing the area for urban development), which had become central themes in discussions of sustainable development and urbanism in Europe and Canada. Although the concept of New Urbanism remained fairly closely linked to practices in the United States, similar philosophies and practices underscored movements like the urban renaissance in Britain and European urbanism (Grant, 2006).

In the late 1990s, New Urbanism found a sympathetic partner in smart growth, a movement originating among state governors in the United States who were eager to respond to rising antigrowth sentiments. Smart growth advocated many of the same planning principles as New Urbanism while suggesting that appropriate government incentives and policies could facilitate urban growth in ways that would be efficient, fiscally responsible, and attractive. New Urbanism proponents also linked with the healthy cities movement to build support for their principles. By the first decade of the twenty-first century, plans and policies for many cities reflected values and premises drawing from these interlinked discourses. New Urbanism premises had been transformed into 'good planning principles.'

Table 1 Summary of key community principles

Ahwahnee principles	*New Urbanism principles*
Complete communities	Walkability
Walking distance for mixed use	Connectivity
Transit oriented	Mixed use and diversity
Mix of housing types to accommodate diversity	Mixed housing types
Jobs for residents	Quality design
Commercial and civic center	Traditional neighborhood structure
Open space network	Increased density
Public space 24/7 use	Green transportation
Defined community edge	Sustainability
Conserve resources, conserve water, and minimize wastes	Quality of life
Design for energy efficiency	

Examples of New Urbanism communities appeared outside the United States during the 1990s. Duany and Plater-Zyberk influenced Canadian projects in Calgary (Alberta) and Markham (Ontario). McKenzie Towne in Calgary began around 1995. It was planned as a 12-phase complete town project, but after 3 phases, the developers reverted to a conventional development model. Although the project was praised for its design qualities, its location and small lot sizes in a city dominated by the oil industry proved a tough sell. Calgary planners worked assiduously to promote sustainability, transit-oriented development, and intensification, but often encountered resistance from the development industry. By the late 1990s, however, the redevelopment of a former army base in Calgary provided an opportunity to make New Urbanism work. Garrison Woods, the first phase of base conversion into urban community, proved extremely successful, selling out two years ahead of schedule. Located relatively close to the urban core, it demonstrated the viability of urban infill according to New Urbanism principles.

Begun in the mid-1990s, Cornell, in Markham, reflected the ambitions of the Ontario provincial government to stimulate urban innovation and intensification in the Toronto area. Despite financial difficulties for its developers, Cornell remained relatively true to its original New Urbanism concept through various stages of development. The Town of Markham entrenched New Urbanism principles through its municipal planning policies in the late 1990s, creating a context where no other type of development became possible. Consequently, Markham developed a reputation as a show place for New Urbanism.

Although the label of New Urbanism has been used relatively rarely outside North America, the movement clearly influenced urban developments in other parts of the world. In the early 1990s, for instance, Krier prepared the plan for Poundbury, a new community developed by the Prince of Wales on the edge of Dorchester in Dorset in the United Kingdom. Reflecting street patterns of historic British villages and an eclectic mix of architectural styles, Poundbury included provisions for 20% affordable social housing. Like an earlier generation of experimental British new towns (such as Letchworth Garden City), Poundbury attracted urban innovators and artists to its homes. To celebrate the millennium, the UK government provided funding to redevelop former industrial lands and expand some green-field sites in sustainable urban forms. Projects such as Greenwich Millennium Village in London and Upton Village in Northampton showed the potential of the urban renaissance to reshape development practices evidently influenced by New Urbanism thinking.

Advocates of New Urbanism formed organizations to promote the vision in other regions of the world: the Council for European Urbanism, the Council for Canadian Urbanism, the Australian Council for New Urbanism, etc. By the turn of the millennium, planners increasingly accepted that many of the principles associated with New Urbanism were simply good planning. Earlier ideological separations between New Urbanism, smart growth, and sustainability had largely diminished in community planning practice. Urban design and physical planning grew in importance while regulatory and zoning approaches lost currency. An upstart movement initiated by architects had helped to redefine the philosophy of the urban planning profession and the understanding of the public of how to plan and design communities.

New Urbanism Practice

New Urbanism principles have been applied to the redevelopment of brown-field (former industrial lands), gray-field (former commercial or parking lands), and green-field (former forest, farm, or scrub land) sites. In some cases, such as the early stages of McKenzie Towne in Calgary (Figure 2), the principles shaped relatively small neighborhoods of a few hundred

Figure 2 McKenzie Towne brought New England urban design to the Prairies.

homes; in some instances, such as the redevelopment of Stapleton airport in Denver, New Urbanism guided the development of large master-planned communities; in a few cases, such as the town of Markham, New Urbanism provided the framework for regulating all new development. Conventional suburban developers have often selectively adopted elements of New Urbanism design, such as traditional architectural styles, making it challenging to differentiate comprehensively planned New Urbanism communities from their imitators.

While New Urbanism played a major role in reintroducing and popularizing traditional architectural styles, its principles of mixed use, walkability, transit orientation, and high density also influenced modernist, high-rise neighborhoods. In Vancouver, the best known example of the high-rise New Urbanism model, planners and designers promoted an urban form featuring ground-oriented podium buildings with commercial uses or town houses at street level and narrow residential towers set back from the street rising high above. While it avoided traditional architectural forms, Vancouver applied policies to ensure that buildings related well to the street, included a proportion of family-oriented units, and contributed to providing 20% affordable housing. Redevelopment of extensive former industrial lands provided the city with abundant opportunities to reshape the character of urban development to promote intensification, mixed use, and design quality.

To implement New Urbanism principles, its practitioners have developed a comprehensive toolkit of strategies (Duany et al., 2010; Steuteville, 2001). New Urbanism designers use design charettes to engage community decision makers and residents in working through local issues and fashioning options for development. In the wake of the damage done by Hurricane Katrina in 2005, for instance, a group of respected practitioners led Mississippi political leaders and residents through a series of charettes to identify strategies for rebuilding communities. The concept of the transect, which differentiates urban forms, densities, and patterns according to the proximity to the center or edge of a community, has been refined to inform codes and strategies planners can use to guide development (Duany and Talen, 2002). The *SmartCode* is an open source tool reflecting New Urbanism's interest in designing form-based codes to influence the character and arrangement of urban form. While conventional land use regulation involved zoning codes that specified allowable uses and densities, form-based codes regulate buildings and spaces. As New Urbanism and smart growth have grown in influence, many local governments have begun to consider form-based codes (especially for their urban cores). In recent years, the CNU worked with the US Green Building Council to develop LEED-ND (Leadership in Energy and Environmental Design-Neighborhood Development), a rating system to evaluate new residential developments on how they address energy and environmental concerns, and with the Institute of Transportation Engineers to try to develop new guidelines for building urban streets.

Extensive application of New Urbanism principles occurred in the United States under the program Housing Opportunities for People Everywhere (HOPE) VI, initiated by the Clinton administration. HOPE intended to renew dilapidated public housing areas by rebuilding them as mixed-use and mixed-income neighborhoods. Thousands of units of public housing were demolished. Proponents of HOPE VI saw signs of social integration and economic revitalization for highly disadvantaged neighborhoods (Duany et al., 2000). Critics described gentrification, displacement, and loss of affordable housing stock for the poorest of the poor (Smith, 2002). Major initiatives to implement New Urbanism principles revealed some of the significant challenges the movement and the theory faced in achieving some of its ambitions.

Challenges and Critiques

From early in its practice, New Urbanism faced criticism that its precepts promoted environmental or architectural determinism, the idea that building form shapes behavior. Practitioners building New Urbanism projects suggested that features such as front porches on houses could generate social interactions between neighbors; corner stores or local pubs would generate a sense of community. Critics challenged the notion that architectural or civic design could achieve social objectives such as community cohesion or neighborly interaction, suggesting that human behavior reflects a wide range of social and economic factors. Theorists developing New Urbanism arguments noted that ensuring place diversity (that is, a mix of uses and housing types) might help to produce social diversity (that is, a mix of classes, household types, and ethnicities), but cautioned that design should be seen as enabling rather than generative of particular social behaviors (Talen, 2008). New Urbanism aphorisms – such as 'eyes on the street enhance security,' 'front porches promote sociability,' and 'compact form reduces vehicle travel' – have come to influence public discourse despite evidentiary challenges.

New Urbanism seeks to produce complete communities with a mix of uses, incomes, household types, and ethnicities. In the context of contemporary community design practice, such ambitions are not easy to achieve. For instance, redeveloping public housing projects to integrate other uses and market housing is described by critics as generating gentrification and displacement. Some see efforts to reurbanize the city – making city living more attractive – as forcing poorer households to the suburban fringe. New Urbanism's success in infill development and suburban new towns has not been able to deflect forces that produce social homogeneity in neighborhoods. Even in places like Markham, Ontario, or Surrey, British Columbia, where New Urbanism communities show high levels of ethnic mixing, they reveal limited socioeconomic or household type diversity (Grant and Perrott, 2009). Strategies to produce social mix do not necessarily produce positive social interactions. Planners and designers have yet to experience great success in trying to achieve social objectives through community design.

Widespread borrowing of successful architectural styles and street designs soon led to concerns about copycat communities and cookie-cutter house forms. Although some large-scale developers built comprehensive communities according to New Urbanism principles, smaller projects selectively borrowed elements of the model and generated confusion over what New Urbanism required. New England-style brownstones and turreted Victorians were duplicated in town after town, while geometric civic center street plans proliferated. Historical

features of regional architecture were sometimes adopted with poor-quality building materials and were reproduced in contexts where they were locally inappropriate. For instance, critics of Celebration, Florida – a new town built by the Disney Corporation but housing relatively few African-Americans – noted that one of its civic buildings echoed the design of the slave market in St Augustine. To try to establish a level of quality control, the CNU established registration and awards programs and worked to differentiate between projects that met the principles of the movement and those that did not. Efforts to produce and disseminate the *SmartCode* – a system to encode urban form – facilitate local development of urban design codes to replace or augment zoning regulations but may threaten to create a new level of homogeneity in urban form.

Typically accompanied by new regulatory regimes to streamline development approvals, the new methods of community design reduce NIMBYism (not-in-my-backyard syndrome) by avoiding public processes for hearing applications. New Urbanism involves community members in setting a community vision. Designers produce elegant watercolor images of promised urban landscapes to assist the decision process. Once municipal officials adopt codes and regulations, implementation falls largely to qualified experts. Instead of planning processes intended to debate the merits of individual projects – and decide whether to approve them or not – New Urbanism encourages communities to identify the desired form of development and then provides mechanisms for planners and designers to implement it without further public consultation. This reliance on the role of experts to make decisions while community residents are excluded from discussions about particular development projects has led some community engagement advocates to call New Urbanism undemocratic.

While New Urbanism gives high priority to accessibility for pedestrians – often advocating grid or modified grid street layouts to ensure connectivity of pathways – the movement has come under scrutiny over accessibility for people with impaired mobility. Because New Urbanism practitioners place homes near the street to create an urban feel for the neighborhood, they characteristically call for the main level of homes to be elevated above street level to ensure visual privacy: front doors thus require stairs but small front yards prevent construction of ramps. Efforts by advocates for persons with disabilities to ensure 'visitability' in new homes – that is, to enable persons using wheelchairs to gain access to the main floor of the house and find an accessible washroom on that level – initially met resistance from New Urbanism. Similarly, the movement weathered debates about including high-rise buildings or street patterns using one-way couplets. Although some argued that only infill development should be promoted as New Urbanism, others believed that green-field projects were essential to the market and could also be sustainable. New Urbanists have forged considerable consensus on key principles while continuing debate on aspects of implementation.

The incorporation of the terminology of smart growth, sustainability, and energy efficiency in New Urbanism masks lingering debates about development priorities and strategies. The advocates of New Urbanism believe that community design strategies have the capacity to reduce energy demand, land consumption, and automobile use. Proponents of smart growth, sustainability, and energy efficiency agree with the aims but in some cases propose alternative strategies and policies, only some of which rely on design approaches.

New Urbanism communities often feature attractive open space networks and in some cases effectively safeguard environmentally sensitive areas. At the same time, however, New Urbanism has been criticized for taking an approach to urbanism that seeks to control nature, favoring urban edges and treatments rather than naturalized waterways or parks. New Urbanism projects have been proposed or built in fragile desert and prairie landscapes. In recent years, the movement has faced challenges from Landscape Urbanism, a competing philosophy developing from landscape architecture that seeks to put landscape processes and green infrastructure into the forefront of urban development.

As a community design movement, New Urbanism competes not only with conventional development approaches but also with niche marketing approaches such as gated communities. In the later decades of the twentieth century, enclosed, limited-access enclaves became a popular development form in many countries. Gated communities promised privacy, security, and exclusivity in urban environments often perceived as dangerous, crowded, and disorderly. While New Urbanists often contrasted their own projects with gated communities, some critics noted that both movements appealed to niche markets in rapidly growing areas, produced limited social diversity, and relied on private management (often condominium corporations or homeowners' associations). Gated communities have captured a larger share of new development but as a practice gating has been much less influential on community planners and designers than has New Urbanism.

Although New Urbanism has been highly influential in shaping planning and urban design philosophies and practices in North American cities, it has faced challenges in implementation. The urban landscape reflects a built environment legacy of decades of automobile culture. Governments have proved unwilling to invest sufficient resources in transportation alternatives or in low- and moderate-cost housing to ensure mix in market projects. Successful New Urbanism communities often contribute to inflationary housing markets that exclude the mix of classes intended in the original plans. Developers may cherry-pick the elements of New Urbanism communities that they find most readily salable, thus undermining the overall quality of the product generated. Local retail uses struggle to survive in a consumer culture where big-box stores prevail. Achieving the ideal of the small and relatively self-contained urban community of the philosophy proves difficult in the context of contemporary life.

The Future of New Urbanism

New Urbanism has become thoroughly embedded in community design philosophy and practice in North America. For many influential practitioners and theorists, its premises and principles have become articles of faith, no longer subject to empirical testing or challenge. What was originally called New Urbanism now is just as likely to be labeled simply 'smart' or 'sustainable' design and planning, embedded in policies, plans, and codes in communities large and small. Although

conventional suburban forms have not disappeared, the popularization of New Urbanism increasingly influences the shape and characteristics of new developments. Just as the Garden City model contributed to spreading serpentine street networks and detached houses on large lots designed for raising children in a blend of town and country living, the New Urbanism movement helped to bring back the grid of connected streets, shape the form of streetscapes, make the town house attractive, and popularize city living. New Urbanism helped design attractive and urbane new neighborhoods and in the process contributed to the contemporary interest in moving back to the city. If the success of the Garden City model is any indication, New Urbanism is likely to experience several more decades of popularity before another community design paradigm comes along to displace it.

See also: Cities and the Financial Crisis; Housing and Cities; Neighborhood Revitalization and Community Development; Regional Planning: China; Transportation Planning; Urban Policy in North America.

Bibliography

Calthorpe, P., 1993. The Next American Metropolis. Princeton Architectural Press, New York.

Duany, A., Plater-Zyberk, E., Speck, J., 2000. Suburban Nation: The Rise of Sprawl and the Decline of the American Dream. North Point Press, New York.

Duany, A., Speck, J., Lydon, M., 2010. The Smart Growth Manual. McGraw Hill, New York.

Duany, A., Talen, E., 2002. Transect planning. Journal of the American Planning Association 68, 245–266.

Ewing, R., 1996. Best Development Practices: Doing the Right Thing and Making Money at the Same Time. APA Planners Press, Chicago.

Grant, J., 2006. Planning the Good Community: New Urbanism in Theory and Practice. Routledge, London.

Grant, J.L., 2009. Theory and practice in planning the suburbs: challenges to implementing new urbanism, smart growth, and sustainability principles. Planning Theory and Practice 10, 11–33.

Grant, J.L., Perrott, K., 2009. Producing diversity in a new urbanism community: policy and practice. Town Planning Review 80, 267–289.

Jacobs, J., 1961. The Death and Life of Great American Cities. Vintage, New York.

Katz, P. (Ed.), 1994. The New Urbanism: Toward an Architecture of Community. McGraw Hill, New York.

Krier, L., 1978. The Reconstruction of the City. Rational architecture: La reconstruction de la ville Européanne. Archives d'Architecture Moderne, Bruxelles. pp. 38–42, 163–180.

Smith, J.L., 2002. HOPE VI and the new urbanism: eliminating low-income housing to make mixed-income communities. Planners Network 151, 22–25. Available at: http://www.plannersnetwork.org/publications/2002_152_spring/smith.htm.

Steuteville, R. (Ed.), 2001. New Urbanism: Comprehensive Report and Best Practices Guide, second ed. New Urban Publications, Ithaca.

Talen, E., 2005. New Urbanism and American Planning: The Conflict of Cultures. Routledge, New York.

Talen, E., 2008. Design for Diversity: Exploring Socially Mixed Neighborhoods. Architectural Press, Oxford.

Van der Ryn, S., Calthorpe, P., 1986. Sustainable Communities: A New Design Synthesis for Cities, Suburbs, and Towns. Sierra Club Books, San Francisco.

World Commission on the Environment and Development, 1987. Our Common Future, Report of the Brundtland Commission. Oxford University Press, New York.

Relevant Websites

http://www.lgc.org/ahwahnee/principles.html – Ahwahnee Principles. (Local Government Commission.)
http://acnu.org/ – Australian Council for New Urbanism.
http://bettercities.net/ – Better Cities and Towns [formerly New Urban News].
http://www.cnu.org/ – Congress for the New Urbanism.
http://www.canadianurbanism.ca/ – Council for Canadian Urbanism.
http://www.ceunet.de/whoweare.html – Council for European Urbanism.
http://www.charretteinstitute.org/ – National Charrette Institute.
http://www.newurbanism.org/newurbanism/principles.html – New Urbanism Principles.
http://www.smartcodecentral.org/ – Smart Code Central.
http://www.smartgrowth.org/ – Smart Growth Network.

News Interview

Steven E Clayman, University of California, Los Angeles (UCLA), Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The news interview is a locus for direct encounters between journalists and public figures, and as such provides a window into journalistic norms, press–state relations, and other aspects of the interface between media, politics, and culture. Its history is intertwined with the coevolution of journalistic and political institutions. In its contemporary broadcast form, the study of interview questioning practices has shed light on journalistic norms regarding neutralism and adversarialism, including their boundaries of relevance, manner of implementation, and varying intensity of observance. Quantitative research indicates that adversarial questioning of US presidents is more frequent on domestic than on foreign policy matters and during downturns in the business cycle. It has also grown more frequent since the 1960s, indicating a trend toward declining deference to political authority and more adversarial relations between press and state.

The news interview is a widespread method by which journalists gather information for incorporation into scripted news stories. It is also employed as a relatively unscripted form of news in its own right, a means by which current affairs information is simultaneously generated and presented for public consumption. The latter usage is particularly important in contemporary broadcasting, as a substantial proportion of news and public affairs programming now takes the form of journalist interviewing of one or more public figures before the media audience.

In its broadcast form, the news interview has attracted scholarly attention because of its manifest significance. It is a public arena for the enactment of journalistic professionalism and the norms that bear on it. It is a forum for political communication and expressions of opinion from diverse societal interests. It both reflects and constitutes relations between journalists on the one hand, and government officials, political candidates, certified experts, and other newsworthy individuals on the other. Correspondingly, the news interview offers a fresh source of data on numerous aspects of the institutional complex of media and politics.

Defining the Broadcast News Interview

The news interview is a familiar and readily recognizable genre of broadcast programming. Its basis in comparatively spontaneous interaction distinguishes it from the fully scripted narratives and stories that have long been a staple of news. The news interview also differs from other interaction-based genres of broadcast talk (e.g., talk shows, panel discussions, etc.) by its distinctive constellation of participants, subject matter, and interactional form. In a prototypical news interview, the interviewer acts as a professional journalist rather than a partisan advocate or celebrity entertainer. Interviewees are pubic officials, experts, or others whose actions or opinions are newsworthy. The discussion normally focuses on current events, is relatively 'formal' in character, and is managed primarily through questions and answers. Prototypical examples include Meet the Press and Nightline in the USA, and Newsnight and The World at One in the UK.

Beyond this prototype are a variety of closely related but less prototypical forms, including news conferences and briefings with large numbers of journalists alternating as questioners. Somewhat farther afield are overtly partisan interviews conducted by political commentators and celebrity talk shows featuring politicians as guests. These embody the fuzzy boundaries of the news interview as a broadcast talk genre, hybrid forms combining elements of the news interview and other 'infotainment' talk show formats (Ekstrom, 2011; Hutchby, 2011).

Origins and Institutionalization

Since the news interview often involves the interface between practitioners of journalism and politics, the history of the news interview is intertwined with the coevolution of these institutions.

Although it now seems quite natural for journalists to interview elected officials and other prominent public figures, it has not always been so. In the USA, interviewing was virtually nonexistent for the first half-century of the nation's existence. Institutions of national government only gradually became publicly accessible, and even as journalists were granted access first to the House of Representatives and later the Senate, verbatim quotations were normally prohibited (Leonard, 1986). The aloofness of government officials was matched by the disinterest of most journalists. Newspapers during this period were financed by political parties and were vehicles for editorial opinion rather than reportage.

The practice of interviewing for the record can be traced to the rise of the American penny press in the 1830s, the first papers to devote themselves primarily to 'news rather than views' and to employ reporters devoted to the task of news-gathering. This new form of journalism expanded rapidly in the USA, and then more slowly in England and other European countries, in part at the prompting of American journalists in Europe. Empirical frequency should not be equated with cultural normativity; throughout the nineteenth century, interviewing was recurrently attacked as an artificial and unduly intrusive journalistic practice (Schudson, 1994).

Although these criticisms would not disappear entirely, the news interview gained greater acceptance by the early decades of the twentieth century. This development roughly coincides with the growing stature and professionalization of journalism, and the rise of 'the bully pulpit' and public relations as tools of governance (Kernell, 1997; Tulis, 1987). Accordingly, the three US presidents most responsible for institutionalizing the presidential press conference – Theodore Roosevelt, Woodrow Wilson, and Franklin Roosevelt – were each progressive reformers who used such conferences to build public support for their ambitious policies. The advent of television further increased the prominence of news interviews and press conferences, as weekly and eventually nightly news programs were devoted to journalistic questioning of public figures.

In short, what used to be regarded as extraordinary has become standard practice across the political spectrum. Just as journalists were once criticized for questioning politicians, now both officials and candidates may be subject to criticism if they fail to make themselves sufficiently accessible to journalistic questioning.

Contemporary Norms and Practices

The news interview is not merely a reflection of journalistic and political institutions; it is also a social institution in its own right. Conduct within the news interview is organized around the roles of interviewer and interviewee, and is governed by a complex matrix of social norms and conventional practices.

Participation and Turn Taking

The most fundamental and pervasive characteristic of news interview interaction is that it unfolds as a series of questions and answers between interviewers and interviewees. This is both an empirical regularity that typifies news interview talk (Heritage and Roth, 1995), and a social norm that the participants are obliged to uphold (Greatbatch, 1988). The question-framework is constitutive of the news interview as a recognizable form of interaction, distinguishing it from other varieties of broadcast talk as well as from ordinary conversation. Underlying this seemingly obvious normative framework is a less obvious substrate of practices which are necessary to produce interaction in manifest compliance with the question–answer norm (Heritage and Greatbatch, 1991). These practices include the systematic avoidance of a wide range of acknowledgment tokens and other responsive behaviors (e.g., *uh huh*, *yeah*, *okay*, *right*, *oh*, *really*) which are pervasive in ordinary conversation, but which become incongruous in a context where the parties adhere to a formal question–answer framework. Both the question–answer norm and the practices underlying it are usually taken for granted by interview participants, but they may become more conscious of the ground rules at problematic or contentious moments, when those rules may be appealed to explicitly as a means or complaint or self-defense.

News interviews are generally conducted without a co-present studio audience, and with a media audience that cannot actively participate in the encounter. The audience is nonetheless tacitly involved, both as a direct addressee during the opening and closing phases, and as a ratified 'overhearer' during the interview proper. Interviewers maintain a tacit orientation to the overhearing audience by producing explicit displays of understanding (e.g., *So you're saying thatn…*) for the audience's benefit, and by avoiding other responsive behaviors (i.e., the aforementioned receipt tokens) that would 'get between' the interviewee and the audience (Heritage, 1985).

Question Design and Journalistic Norms

In building questions, interviewers are sensitive to two further journalistic norms. On the one hand, in accordance with the ideal of objectivity, they are supposed to remain neutral. While absolute neutrality is an unattainable ideal, interviewers do strive to maintain a formally neutral or 'neutralistic' posture by restricting themselves to actions accountable as 'asking questions,' avoiding all forms of acknowledgment, and avoiding flat assertions except as prefaces to a question (Clayman, 1988) or as attributed to a third party (Clayman, 1992). Quantitative studies indicate that the overwhelming majority (>90%) of interviewer contributions in both USA (Heritage and Roth, 1995) and UK (Tolson, 2012) news interviews are neutralistically framed. At the same time, it bears emphasis that adherence to a formally neutralistic posture is no guarantee of absolute neutrality, and efforts to test for partisan bias in the distribution of questioning practices have met with some success, with bias documented in Dutch (Huls and Varwijk, 2011) but not in Swedish (Ekström et al., 2012) national contexts.

While the norm of neutralism is very broadly relevant in news interviews, it is not without limits. One boundary condition for the applicability of this norm is the domain of mainstream politics and 'legitimate issues' about which it is believed that reasonable people can disagree. Here journalists are obliged to impartially represent the divergent views of established political factions (cf Hallin, 1984). Beyond this domain, however, in areas seen as consensually admirable or deviant, journalists appear less bound by neutralism and more willing to express consensus beliefs and values via unvarnished assertions of praise or condemnation (Liebes et al., 2008). A similar orientation to societal consensus may license interviewers to correct factual errors and obviously false statements by interviewees (cf Montgomery, 2011).

A second journalistic norm bearing on question design is that of adversarialism. Consistent with the ideal of journalistic independence and the watchdog role, interviewers should not permit the public figure to transform the interview into a purely self-serving public relations event. There are various avenues by which a critical or oppositional stance can be pursued within the confines of ostensibly neutralistic 'information-seeking' questions. Questions almost always set relevant topical and action agendas, encode presuppostions, and favor particular responses; interviewers often mobilize these aspects of question design in ways that are oppositional toward the interviewee (Clayman and Heritage, 2002a). An important resource for these various forms of adversarialism is the preliminary statements that interviewers often make when leading up to a question. Such statements are justifiable as providing 'background information' necessary to render the question intelligible to the audience, but they can be deployed in ways that

allow the interviewer to maximize control over the discussion agenda, introduce substantively oppositional information, and exert pressure on recalcitrant interviewees.

The adversarialism norm has boundary conditions that are narrower than the conditions for neutralism. The norm is activated when the independent watchdog ideal becomes salient; hence, in interviews with politicians and other elite newsmakers. Political candidates receive particularly vigorous interrogative scrutiny, and campaign interviews have played an important role in recent US elections (Clayman and Romaniuk, 2011). Conversely, the norm is less relevant in interviews with nonpartisan technical experts and ordinary persons (Montgomery, 2007). Even when it is situationally relevant, adversarialism is more variably enacted than neutralism, characterizing only a minority of questions even in the contemporary era (Clayman et al., 2006). Adversarial questioning tends to be relatively muted when the topic turns to foreign affairs and national security, and when the business cycle is on the upswing (Clayman et al., 2007). And in the broader context of journalistic work practices, adversarialism is less prominent in news interviews than in subsequent story-based news coverage that interviews may receive (Eriksson and Ostman, 2013; see also Eriksson, 2011a).

Answering and Resisting

Interviewees, in responding to questions, face a different set of cross-cutting interests and pressures. Since they generally prefer to stay 'on message,' adversarial questions create an incentive for resistant or evasive responses that shift the discussion agenda in a more desirable direction. However, the normative question–answer framework obliges interviewees to answer straightforwardly, so that failure to do so can have various costs. Audience members may infer that the interviewee has some ulterior motive for avoiding the question; interviewers often counter such maneuvers with probing follow-up questions and negative sanctions (Romaniuk, 2013b; cf Eriksson, 2011b); and the acts of evasion may be singled out in subsequent news coverage (Clayman, 1990).

Accordingly, interviewees typically design their resistant responses in ways geared to minimizing these undesirable consequences (Clayman and Heritage, 2002a). They may choose to sidestep the question in an overt or explicit manner, which allows for equally explicit forms of 'damage control,' such as justificatory accounts (Ekström, 2009), and requests for permission displaying deference to the interviewer. Alternatively, when evading the question covertly, they may take steps to obscure what is transpiring by giving it the surface form of an answer. There are also techniques of resistance that can be employed by interviewees earlier, as the question is unfolding (Romaniuk, 2013a).

Question Design and Press–State Relations

How do interviewing practices, in particular those associated with vigorous and adversarial questioning, vary over time? In exploring historical trends in interviews with government officials, the study of question design becomes a window into the evolving relationship between press and state.

Adversarial questioning has risen since the 1950s in both Britain and America, although this broad trend has developed somewhat differently in these countries (Clayman and Heritage, 2002a). In Britain, a robust tradition of government regulation of broadcasting, coupled the absence of competition prior to 1958, combined to foster a highly deferential style of questioning in BBC interviews of the 1950s. When the BBC monopoly was replaced by a duopoly in 1958, the resulting competition fueled a sudden and dramatic increase in adversarial questioning. In America, where government regulation of broadcasting has been comparatively minimal and where competitive pressures have been present from the outset, adversarial questioning has grown more steadily from a higher baseline.

To track this development in the US context, Clayman and Heritage (2002b), building on previous question design research, developed a system for quantitatively assessing the level of aggressiveness in journalistic questioning. This system, as subsequently refined (Clayman et al., 2006), involves practice-based indicators for five dimensions of aggressiveness: (1) initiative, (2) directness, (3) assertiveness, (4) adversarialness, and (5) accountability.

Initial applications of this system focused on US presidential news conferences and were primarily descriptive, documenting historical trends in the aggressiveness of the White House press corps over a half-century (Clayman et al., 2006). For nine Presidents from Eisenhower through Clinton, journalists grew significantly more vigorous on all five dimensions. To some extent this transformation affected the repertoire of practices actively in use, with certain extremely deferential practices (e.g., indirect questions such as *May I ask you to talk about X*) falling out of favor, and other extremely aggressive practices (e.g., accusatory *how could you* … questions) being added to the repertoire. For the most part, however, the transformation has affected only the relative frequency of practices. Journalists have tended to exercise greater initiative via more elaborate prefaced and compound questions; the content of their questions has grown increasingly oppositional and includes greater demands for accountability; and they exert greater pressure on the president to address such content via more direct and assertive forms.

Subsequent research on the same database has been explanatory in emphasis, using multivariate analysis to identify the social conditions associated with more vigorous questioning. Significant predictors include the question content (domestic affairs questions are more aggressive than foreign affairs questions), the timing of the conference within the president's tenure in office (second terms yield more aggressive questions than first terms), and the broader economic context (questions become more aggressive when unemployment and interest rates are rising) (Clayman et al., 2007). Further, returning to the explanation of historical trends, Clayman et al. (2010) document a punctuated equilibrium pattern with the late 1960s as a turning point. Controlling for the aforementioned social conditions, questions exhibited a stable level of deference toward the president from 1953 to 1968, and then jumped abruptly to a more aggressive level from 1969 to 2000. Thus the rise of more persistent adversarialism was not a transitory response to shifting circumstances, but reflected a more fundamental and

enduring transformation of the normative tenor of president–press relations in the USA.

Whatever its causes, the growth of vigorous and adversarial questioning has transformed the broadcast news interview into a more formidable instrument of public accountability. It is now more difficult for officials and other public figures to make purely self-serving statements in an interview context. However, this revolution has stimulated a counter-revolution by public figures: increasingly sophisticated strategies of resistance and evasion, aided by a burgeoning cottage industry of media advisors and consultants (Jones, 1993). At the same time, politicians have been exploiting what is perceived as the more congenial interactive environment of 'infotainment' programs (Baum, 2005), including celebrity, comedic, and satirical talk shows (Loeb, forthcoming).

See also: Journalism and Journalists; Journalism; Media Talk Shows; News: General; Political Communication.

Bibliography

Baum, M.A., 2005. Talking the vote: why presidential candidates hit the talk show circuit. American Journal of Political Science 49 (2), 213–234.

Clayman, S.E., 1988. Displaying neutrality in television news interviews. Social Problems 35 (4), 474–492.

Clayman, S.E., 1990. From talk to text: newspaper accounts of reporter-source interactions. Media, Culture, and Society 12 (1), 79–104.

Clayman, S.E., 1992. Footing in the achievement of neutrality: the case of news interview discourse. In: Drew, P., Heritage, J. (Eds.), Talk at Work. Cambridge University Press, Cambridge, pp. 163–198.

Clayman, S.E., Heritage, J., 2002a. The News Interview: Journalists and Public Figures on the Air. Cambridge University Press, Cambridge.

Clayman, S.E., Heritage, J., 2002b. Questioning presidents: journalistic deference and adversarianness in the press conferences of U.S. presidents Eisenhower and Reagan. Journal of Communication 52 (4), 749–775.

Clayman, S.E., Elliott, M.N., Heritage, J., McDonald, L., 2006. Historical trends in questioning presidents 1953–2000. Presidential Studies Quarterly 36, 561–583.

Clayman, S.E., Heritage, J., Elliott, M.N., McDonald, L., 2007. When does the watchdog bark?: conditions of aggressive questioning in presidential news conferences. American Sociological Review 72, 23–41.

Clayman, S.E., Elliott, M., Heritage, J., Beckett, M., 2010. A watershed in white house journalism: explaining the post-1968 rise of aggressive presidential news. Political Communication 27, 229–247.

Clayman, S.E., Romaniuk, T., 2011. Questioning candidates. In: Ekström, M., Patrona, M. (Eds.), Talking Politics in the Broadcast Media: Cross-Cultural Perspectives. John Benjamims, Amsterdam, pp. 15–32.

Ekström, M., 2009. Announced refusal to answer: a study of norms and accountability in broadcast political interviews. Discourse Studies 11 (6), 681–702.

Ekström, M., 2011. Hybridity as a resource and challenge in a talk show political interview. In: Ekström, M., Patrona, M. (Eds.), Talking Politics in Broadcast Media. John Benjamins, Amsterdam.

Ekström, M., Eriksson, G., Johansson, B., Wikström, P., 2012. Biased interrogations?: A multi-methodological approach on bias in election campaign interviews. Journalism Studies 14 (3), 423–439.

Eriksson, G., 2011a. Adversarial moments: a study of short-form interviews in the news. Journalism 12 (1), 51–69.

Eriksson, G., 2011b. Follow-up questions in political press conferences. Journal of Pragmatics 43 (14), 3331–3344.

Eriksson, G., Östman, J., 2013. Cooperative or adversarial?: journalists' enactment of the watchdog function in political news production. International Journal of Press/Politics 18 (3), 304–324.

Greatbatch, D., 1988. A turn-taking system for British news interviews. Language in Society 17 (3), 401–430.

Hallin, D.C., 1984. The media, the war in Vietnam, and political support: a critique of the thesis of an oppositional media. The Journal of Politics 46 (1), 1–24.

Heritage, J., 1985. Analyzing news interviews: aspects of the production of talk for an "overhearing" audience. In: van Dijk, T. (Ed.), Handbook of Discourse Analysis, Discourse and Dialogue, vol. 3. Academic Press, London, pp. 95–117.

Heritage, J., Greatbatch, D., 1991. On the institutional character of institutional talk: the case of news interviews. In: Boden, D., Zimmerman, D.H. (Eds.), Talk and Social Structure. University of California Press, Berkeley, pp. 93–137.

Heritage, J., Roth, A., 1995. Grammar and institution: questions and questioning in the broadcast news interview. Research on Language and Social Interaction 28 (1), 1–60.

Huls, E., Varwijk, J., 2011. Political bias in TV interviews. Discourse and Society 22 (1), 48–65.

Hutchby, I., 2011. Non-neutrality and argument in the hybrid political interview. Discourse Studies 13 (3), 349–367.

Jones, B., 1993. The pitiless probing eye: politicians and the broadcast political interview. Parliamentary Affairs 46 (1), 66–90.

Kernell, S., 1997. Going Public: New Strategies of Presidential Leadership. Congressional Quarterly Press, Washington.

Leonard, T.C., 1986. The Power of the Press. Oxford University Press, Oxford.

Liebes, T., Kampf, Z., Blum-Kulka, S., 2008. Saddam on CBS and Arafat on IBA: interviewing the enemy on television. Political Communication 25 (3), 311–329.

Loeb, L. The Celebrity Talk Show: Norms and Practices, forthcoming.

Montgomery, M., 2007. The Discourse of Broadcast News. Routledge, London.

Montgomery, M., 2011. The accountability interview, politics and change in UK public service braodcasting. In: Ekström, M., Patrona, M. (Eds.), Talking Politics in Broadcast Media. John Benjamins, Amsterdam.

Romaniuk, T., 2013a. Interviewee laughter and disaffiliation in broadcast news interviews. In: Glenn, P., Holt, E. (Eds.), Studies of Laughter in Interaction. Bloomsbury Press, London, England, pp. 201–220.

Romaniuk, T., 2013b. Pursuing answers to questions in broadcast journalism. Research on Language and Social Interaction 46 (2), 144–164.

Schudson, M., 1994. Question authority: a history of the news interview in American journalism, 1830s–1930s. Media, Culture, and Society 16, 565–587.

Tolson, A., 2012. You'll need a miracle to win this election" (J. Paxman 2005): interviewer assertiveness in UK general elections 1983–2010. Discourse, Context, and Media 1, 45–53.

Tulis, J., 1987. The Rhetorical Presidency. Princeton University Press, Princeton.

News: General

Svennik Høyer, University of Oslo, Oslo, Norway

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 10638–10642, © 2001, Elsevier Ltd.

Abstract

'News' alternately refers to how news events are evaluated, reported, edited, or presented as 'news' to the audience by journalists working within an organizational context. Thus news is part of the process of telling new stories about current affairs; it is intended as an integral part of an ongoing communication to keep a given community or society informed and functioning. As a single message 'news' may be defined as something current and important to tell by widely available media which is not known in advance by an anticipated audience, and which is likely to catch the attention of a sufficient number of recipients. This definition covers both a textual 'product' and the material and economic aspects of messages as well as the technological, social, and wider cultural 'process' aspects of news stories. The news media today include newspapers, news magazines, radio, television, and Internet services. Each medium has its own form, due to some means of presentation and format which is peculiar to their dominant technology of dissemination. However, news journalism also has some common features irrespective of medium, due to similar methods and norms of retrieving information and to the forms of an effective presentation of facts.

The communication of news has been part of a storytelling tradition since time immemorial. Storytelling to a recurring public is a way of making common sense of the environment. Stories and news about people, friends, and relatives have always passed through the grapevine. Both news and rumor give nourishment for social discourse; something that can be adapted to local experiences in the company of a small group of familiar faces (*see* Rhetorical Analysis).

Definition and Background

In its modern and more formal usage 'news' refers to certain types of texts produced by the personnel, institutions, and methods of news media. The news media today include news agencies, newspapers, news magazines, radio and television, and Internet services. Each medium has its own form linked to some means of presentation, which is peculiar to their dominant technology of dissemination. However, news journalism also has some common features of content, irrespective of medium, based on similar methods of choosing stories, retrieving information, and of presenting facts effectively.

A discussion over what are the typical contents of 'news' has been a part of news sociology for as long as the subfield has existed, i.e., since the 1960s. There is not as yet any widespread agreement as to the number and kind of criteria. But researchers emphasize some qualities of news texts more often than not. News is supposed to tell something that has really happened recently, to be important for the nation or for the community, and to interest a substantial part of the audience. Thus Mitchell Stephens (1997, p. 9) defines news as 'new information about a subject of some public interest that is shared with some portion of the public.' Brian McNair (1998, p. 4) stresses that 'news claims to be (i.e., is presented to its audience as) truthful statement about, or record of, some hitherto unknown (new) feature of the actual, social world.'

Even so, journalists thrive on rumor but do not necessarily publish all that they hear. Different media may differ in what they consider a threshold for what can be made public; sometimes depending on their means of presentation, the risk of litigation, or on the ideals of journalism media adhere to. News in popular tabloids differs in many respects from news in television and in the quality press.

News as a social habit existed before the first newspaper appeared in Antwerp during a siege of the city in 1605, and even before the printed word. Tidings, rumor, and bits of information from near and far away could be learned at certain places: in the corners of a village square, in postal offices, in coffee houses and taverns, in front of churches, at the harbor, in parks etc. *Gazetta* was the name of a small Venetian coin for the price of which you could purchase a newssheet carrying commercial and other news at the Rialto Bridge in seventeenth century Venice. From the sixteenth century, and long into the nineteenth century, flysheets provided stories comparable to the sensational tabloid news of today: romantic, adventurous, and melodramatic events; supernatural phenomena; the royal court; executions, wars, and witchcraft, etc.

Regularly published newspapers, usually weeklies, were established at the beginning of the seventeenth century in England, the American colonies, Holland, Denmark, etc. They were operated by jobbing printers who also doubled as postmasters and booksellers. Since few, if any, journalists were employed, printers took their news generally from other publications, mostly from the capital, from foreign newspapers, and from letters to the editor. With regularity of publication newspapers could report continuing stories, largely as a series of minor incidents successively published. Early newspapers almost exclusively used foreign news. Censorship or the threat of censorship made local news more dangerous to publish. Over time, as censorship was eased or lifted, national news stories became a more important part of the news content (Wilke, 1987).

Periods in national media histories are sometimes linked to major changes in media technologies, in economy, and to the size of the audience, which can be compared, cross-nationally to some extent. Periodization of the developing style and

content of journalism is often linked to changes in the political system where the causal links are sometimes evident, sometimes unclear. Commonly used labels in media histories are the early elite press, the party press, the mass press, national public service versus commercial broadcasting, etc.

The development of journalism was, until the mid-nineteenth century, closely bound to the book trade, in content as well as in authors. As the press expanded in terms of the number of fully employed journalists, journalism became more distinct. Divergent and seemingly contradictory trends have been observed cross-nationally. However, a general trend seems to be toward more variety in journalistic genres and methods of expression, especially after radio and television became the fastest news media. In the USA and in many countries of Europe two basic styles of journalism has been identified since the mid-nineteenth century, one discursive and another sensational. There is simultaneously a tendency toward a 'long journalism' with a slightly 'top-down' perspective, and another toward a more grassroots-oriented 'bottom-up' populist perspective.

The first trend has produced stories set within a wider frame of time and space. In a study of worklife journalism, criminal and accident reporting in *New York Times*, the *Chicago Tribune*, and the *Oregonian* between 1894 and 1994, Barnhurst and Mutz (1997) demonstrated that reporting developed toward longer stories, wider contexts, and more analysis. The "journalists identified individuals less often by name and more often by demographic group" (Barnhurst and Mutz, 1997, p. 41). Many events nowadays do not become newsworthy unless they fit into "a longer body of interpretations and themes" (Barnhurst and Mutz, 1997, p. 41). This development has been possible by the mobilization of substantial journalistic resources, which again is dependent upon advertising revenues and upon some ownership concentration in the audience market.

The other trend – mainly for apolitical tabloids – has been toward spot news as entertainment: centered both on ordinary individuals and on celebrities set within dramatized events where persons, actions, and feelings are foregrounded. Benjamin H. Day, pioneer of the tabloid genre from the 1830s, said that "News, properly so called … must generally tell of wars and fightings, of deeds of death, and blood, of wounds and heresies, of broken heads, broken hearts, of accidents by fire or flood, a field of possessions ravaged, property purloined, wrongs inflicted, rights unavenged, reputations assailed, feelings embittered, and oppressions exercised by nations, communities, or individuals' (quoted in Huntzicker, 1999, pp. 14–15). This tabloid formula seems to have been employed in urban boulevard newspapers in all kinds of political systems except the totalitarian ones.

The News Paradigm

Social scientists often use changes in technology as an index of wider social developments. Thus, the electric telegraph – invented in 1844 – and the practice of news agencies have served a wide range of explanations of why and how journalism has changed since the mid-nineteenth century. This infrastructure is said to produce alternately: the 24-h news cycle, the 'inverted pyramid form' of presentation, and 'objectivity' in journalism. Jürgen Wilke (1984) found e.g., that between 1856 and 1906 reported events that had occurred in the past 24 h increased from 11% to 95% in the German press. The international news agencies like Reuters have been described as pioneers of neutral reporting and of the 'lead' or the inverted pyramid form (Read, 1999). Likewise, Donald Shaw (1967) explained the alleged fall of the party press in the USA in the 1880s as a result of the increasing use and dependence on wired and politically neutral news by the newspapers. Many of these explanations have been challenged recently as technological determinism, ignoring journalism as a genuinely cultural phenomenon following its own trajectory.

Richard Kaplan (1997) found, for example, that the fall of the party press in Detroit did not coincide with the increased use of wired news. Michael Schudson (1978) interpreted the rise of 'objectivity' in journalism as influenced by a philosophical pragmatism in contemporary USA at the beginning of the twentieth century and by a detached professionalism in a growing number of occupations at the time. The canon of 'objectivity' became current in the late 1920s. But even by the turn of the twentieth century establishing facts was an ideal in journalism. 'Objectivity' in addition asked for balanced facts and opinions, which required new discursive practices (*see* Journalism).

The US newspaper industry had an unprecedented growth between 1880 and 1920 with widely expanding circulation, which included new layers of society in the informed public. In combination with previous breakthroughs in paper and print technology – which made cheap copy prices possible – it transformed the press into our first mass medium. The working environment and the job definition of reporters as different from writers changed rapidly (Schudson, 1995), and the audience easily became strangers as the circulation area expanded. "Journalists were thus becoming something of a separate class, a separateness indicated by increasing numbers of articles and lectures by journalists intended to explain how the press worked" (Dicken-Garcia, 1989, p. 230).

This was also a time ripe with experiments in journalism. The various ideals, values, criteria, definitions, etc., which were eventually assimilated into the modern idea of proper news, became a way of thinking in journalism from the beginning of the twentieth century. Høyer (1995) constructs a 'news paradigm' out of five distinct elements emerging in this period: (1) the event, (2) the idea of neutral news criteria, (3) the news interview, (4) the inverted pyramid method of presentation, and (5) objectivity. This combination of journalistic methods and genres appeared gradually as separate social and cultural inventions over a period of less than 50 years. The location of most innovations was New York and London, which were also centers of applied modern newspaper technology in the latter part of the nineteenth century. The five elements can be specified as follows:

1. The event may be defined as an interaction, an accident, a meeting, a crime, etc. which ideally shall begin and end between deadlines. With regular news services and the electric telegraph the time frame of a new event became the 24-h publication cycle.
2. News criteria are the use of nonsubjective criteria for the selection of news to be published.
3. The news interview has now become the main method of collecting information. It is a peculiar kind of conversation

where the information source feels obliged to respond to inquiries, without the journalist responding psychologically since the journalsist is talking on behalf of a third – silent but more important – part: the public. This form of interview became widespread in the US press during the 1870s and 1880s and was at first met with outright disdain from the elite, since it gave newspapers more control of what should become public news (Schudson, 1994)

4. The 'inverted pyramid' is a rhetorical form. It became a standard from the 1890s. The inverted pyramid emphasizes fast and short presentation of the event, in the headlines and especially in the opening paragraph of the story, known as the 'lead.' The lead contains the major facts and actors involved. The net result of this method was stories lacking an internal chronology with implied explanations. Stories were supplied with chronologies turned downside up, and written in a dry and down-to-earth style of language, hiding the journalist as the author responsible for the story. By the flexibility of computer technology the lead technique in presentation has recently been modified. Graphics, frames for facts and summaries of previous events, etc., serve as satellites to the main story, and allow the journalist to start closer to the chronological beginning of the reported event.
5. Objectivity – or more succinctly 'balance' of information from contending sources of information – is obtained by using many different authorities as sources. The number and variety of information sources attributed is sometimes given as the operationalized definition of 'objectivity' in journalism. This form of neutrality is different from the scientific or philosophical idea of objectivity, where the quality of data is methodically evaluated and intersubjectively controlled by repetitive tests.

The 'omnibus' newspaper became the arena for these diverse innovations; it became the answer to a growing heterogeneity of newspaper readers and to their diversity of interests. Increased advertising, cheap newsprint, etc. allowed for more space for journalism and for typographical experiments. The sum of stories is implicitly supposed to tell the most important of the latest events.

There are some apparent interconnections between the diverse elements of the news paradigm. The electric telegraph, which inspired the 24-h news cycle, also gave a time frame for news stories to be fresh and new, which subdued long-winded histories with complex cause and effects linkages. The event by definition became something that happened in a very short time. The emphasis on actuality, variety, brevity, and speed easily inspires the inverted pyramid form of presentation. When facts, events, and drama are foregrounded without any explicit interpretation by the journalist, the trustworthiness of facts and opinions becomes essential. The main method of gaining authority or 'objectivity' in journalism became the news interview of expert and elite information sources.

Diachronic and International Perspectives

The technology and economics of news production soon spread to many countries, while the diffusion of 'the news paradigm' lagged behind. The marked expansion in circulation between 1880 and 1920 observed in the USA has, e.g., been documented in Denmark, Finland, and Norway (Thomsen, 1972; Tomilla 1988; Høyer, 1995). Scattered data suggest that it took several more decades before the US-inspired news paradigm became widely adapted in most European countries, except for the UK. In 1905 and 1908 major newspapers in Copenhagen and Stockholm respectively adopted 'American journalism' (Marzolf, 1982), while Jean Chalaby (1996) suggests the 1930s for France. Countries like Spain and Russia lagged even further behind for obvious reasons. The local press seems to lag behind the major cities everywhere. The historical geography of changing journalism is different from the changing media technology, even if the diffusion process has not yet been mapped out in more detail.

The most common reason given for this cultural lag between the Anglo-American world, Scandinavia, and Continental Europe is the strong tradition of a party press following a sustained period of censorship. In France and the Germanies, for example, the original ideal of a journalist was the mature gentleman of the press – the *redakteur* – commenting on the affairs of the world and interpreting trends in politics, arts and science from a given philosophical position (Esser, 1998). With regard to Central and Eastern Europe, Owen V. Johnson (1999, p. 8) points to how journalism overlapped with politics and literature from the latter half of the nineteenth century to the Soviet period. Young writers made their living as newspaper journalists until they could make their mark as writers or politicians. In contrast, the Anglo-American ideal was the outgoing aggressive news hunter unbound by political allegiances (Köcher, 1986).

These differences have been explained by the longer periods of censorship in Continental Europe, which also produced a conception of 'state-professions' as regulated by the authorities, different from the Anglo-American system. By contrast to the US journalist the German *redakteur* demonstrated independence by giving personal and informed opinions which could deviate from more officially accepted views.

Jean Chalaby (1996) focuses more on cultural factors like the sharper hierarchy of literary genres in France as compared to Britain, fostering a literary and enigmatic journalism directed toward an elite public as the interpretive community. This tradition continued for decades after censorship was abandoned (Esser, 1998; Johnson, 1999). On the other hand, the introduction of an authoritarian regime in democratic Estonia 1933–40 did not prevent the breakthrough of more modern and US-inspired journalistic methods (Harro, 2000). In Scandinavia the party press subsided gradually somewhere in the 1970s and 1980s. But even before journalistic objectivity was given priority over alternative journalistic values, other elements of the news paradigm were adopted. Nonseid and Høyer (1998) found in a content analysis of four Oslo newspapers 1908–30 only some traces of evidence for Donald Shaw's (1967) thesis that news agencies influenced journalism more generally. They studied occurrences of 'the inverted pyramid,' which became a more common story telling method during the period, both for wired and self-produced stories. But, except for the wired news, a clear majority of news throughout the period was still cast in different and older forms of story presentation. The influence of news agencies could be traced, but was weak. Alternative cultural and political factors were apparently stronger.

The former USSR had different news criteria and different concepts of autonomy in media. The media should first of all be concerned with what the public ought to know according to the communist credo. The news should depict the party line as the guideline for correct behavior. The staple news diet contained all kinds of stories about economic and social progress within socialist countries. Soviet news was not sensational and rarely reported accidents of any kind. TASS – the news bureau – provided between a third and a half of the contents of published Soviet news. In spite of this, events were not necessarily defined by the 24-h news cycle. The timing of publication was mostly bound to political routines and necessities (*see* Mass Communication: Normative Frameworks). As a reflection of this conception of what is news Halliki Harro (2000) demonstrates in a recent content analysis of Estonian newspapers 1920–99 that the 'inverted pyramid' hardly existed during the Soviet period.

These international examples demonstrate a variety that denies the simple connections between technological and cultural innovations claimed by technological determinists. Even if there may be some links between media technology and the journalism, the connection is certainly indirect and not synchronous. Media technology seems to have diffused much faster and more widely than innovations in journalism, which is 'part of, and responding to, a deeper cultural movement' (Schudson, 1991, pp. 175–189).

News in Electronic Media

Radio news started gradually in the 1920s, first by some amateur ham operators reading directly from newspapers to an unknown public. Music, and to a lesser extent lectures and humor, were the great attractions of the new medium, not news. Until the mid-1930s few journalists were assigned to radio news. Only in 1934 did news become an independent department in the British BBC (Schlesinger, 1978). Typically, at the beginning, news stories consisted of short bulletins of wired news from major news agencies.

From the outset the press tried hard to repress, and later to control, radio news. The press in Europe most often controlled the main news agencies from where broadcasting corporations got most of their news. In return for access to the wired news the press in the 1920s and early 1930s demanded that broadcasters limit transmission to one news program at around 7.00 p.m., after listeners had got their evening newspapers.

Scarcity of frequencies regulated by the state, as a common good, made public service broadcasting more vulnerable to government control compared to the press. Authorities were determined to avoid controversial topics in radio. Political parties, trade unions, and special interest organizations were thus not welcome as topics or actors in the news. This made radio news rather unrealistic and broadcasters became the voice of some unidentified official establishment.

World War II broke most of these restrictions. Backgrounds and interpretations were needed to make sense of what happened and to bolster national morale. Day-round bulletins, mixed with reportage and interviews, now became the rule. Radio reporters became identifiable persons, sometimes celebrities, as they reported directly from the war theaters. Instrumental in this process was also the taperecorder, which came into use in 1939.

Television news started out as a copy of the cinema newsreel, with presenters heard but not seen. But the audience wanted to see the faces of announcers, who eventually developed into the role of the anchorperson and moderator. The anchorperson connected the steadily differentiating parts of a televised news story: a short announcement amounting to 'headlines,' a lead-type message followed by interviews, reports from eyewitnesses or reporters on location, and expert comments back in the studio to wind up the story. A study of 17 newscasts from eight countries identified different styles of anchorpersons in Western Europe: the 'Germanic' newsreader who briefly introduces the topic and leaves the substance of information to field reporters, as contrasted to the 'Romance' moderators who provide lengthy transitions between stories and subjective comments in the style of an editorial writer (Heinderyckx, 1993).

The Internet promises a lower threshold of access to the distribution of news and views for the regular public in the industrialized world. But to be effective as a news channel it seems that the Internet so far must rely on existing networks of people with special interests. The news sites serve as points of convergence both for a like-minded public internationally and for all sorts of changing information and images within specific topic areas (*see* Information Society). Within finance and some scientific disciplines, the Internet already serves as a major publishing channel or as a marketplace for news. But for the moment it lacks the ability to reach the broader audience during a regular news cycle. For the Internet to become a general news channel, its use must expand considerably above its present level of penetration.

See also: Celebrity; Entertainment; International Communication: History; Journalism; Mass Communication: Normative Frameworks; Mass Media, Political Economy of; Mass Media, Representations In; Mass Media: Introduction and Schools of Thought; News Interview; Printing as a Medium; Rhetorical Analysis.

Bibliography

Barnhurst, K.G., Mutz, D., 1997. American journalism and the decline in event-centered reporting. Journal of Communication 47 (4), 27–53.

Chalaby, J., 1996. Journalism as an Anglo-American invention. A comparison of the development of French and Anglo American journalism 1830s–1930s. European Journal of Communication 11 (3), 303–326.

Dicken-Garcia, H., 1989. Journalistic Standards in Nineteenth Century America. University of Wisconsin Press, Madison, WI.

Esser, F., 1998. Editorial structures and work principles in British and German newsrooms. European Journal of Communication 13 (4), 375–405.

Gans, H.J., 1979. Deciding What's News. A Study of CBS Evening News, NBC Nightly News, Newsweek & Time. Pantheon Books, New York.

Golding, P., Elliot, P., 1977. Making the News. Longman, London.

Harro, H., 2000. Changing journalistic conventions in the press. Empirical studies on daily newspapers under different political conditions in 20th century Estonia. Ph.D. dissertation, University of Oslo.

Heinderyckx, F., 1993. Television news programs in Western Europe: a comparative study. European Journal of Communication 8 (4), 425–450.

Høyer, S., 1995. Pressen mellom teknologi og samfunn (The press between technology and society). Universitetsforlaget, Oslo.
Huntzicker, W.E., 1999. The Popular Press, 1833–1865. Greenwood Press, Westport, CT.
Johnson, O.V., 1999. The roots of journalism in Central and Eastern Europe. In: Aumente, J., Gross, P., Hiebert, R., Johnson, O.V., Mills, D. (Eds.), Eastern European Journalism. Before, During and After Communism. Hampton Press, Cresskill, NJ, pp. 5–40.
Kaplan, R., 1997. The American press and political community: reporting in Detroit 1865–1920. Media Culture & Society 19 (3), 331–355.
Köcher, R., 1986. Bloodhounds or missionaries: role definitions of German and British journalists. European Journal of Communication 1 (1), 43–64.
Marzolf, M., 1982. Pioneers of "New Journalism" in early 20th Century Scandinavia. Pressens Årbog. 20.
McNair, B., 1998. The Sociology of Journalism. Arnold, London.
Michell, S., 1997. A History of News. Harcourt Brace, Fort Worth, TX.
Nonseid, J., Høyer, S., 1998. Diffusion of the news paradigm in Norway 1908–1930. Paper for International Association for Mass Communication Research, Glasgow. July.
Read, D., 1999. The Power of News. The History of Reuters, second ed. Oxford University Press, Oxford.
Schlesinger, P., 1978. Putting 'Reality' Together. BBC News. Constable, London.
Schudson, M., 1978. Discovering the News. A Social History of American Newspapers. Basic Books, New York.
Schudson, M., 1991. Historical approaches to communication studies. In: Bruhn Jensen, K., Janowski, J.W. (Eds.), A Handbook of Qualitative Methodologies for Mass Communication Research. Routledge, London, pp. 175–189.
Schudson, M., 1994. Question authority: a history of the news interview. Media, Culture & Society 16 (4), 565–589.
Schudson, M., 1995. The Power of News. Harvard University Press, Cambridge, MA.
Schudson, M., 2000. The sociology of news production revisited (Again). In: Curran, J., Gurevitsh, M. (Eds.), Mass Media and Society, third ed. Arnold, London.
Shaw, D.L., 1967. News bias and the telegraph: a study of historical change. Journalism Quarterly 44 (1), 3–16.
Stephens, M., 1997. A History of News. Harcourt Brace, Fort Worth, PA.
Thomsen, N., 1972. Dagbladskonkurrencen 1870–1970 (The Competition between Daily Newspapers 1870–1970). Gads, Copenhagen.
Tichenor, P.J., Donohue, G.A., Olien, C.N., 1980. Community Conflict & The Press. Sage, Beverly Hills, CA.
Tomilla, P., 1988. Suomen lehdistön historia (The history of the Finnish Press). Kustannuskiila, Kuopio, Finland.
Tuchman, G., 1978. Making News. A Study in the Construction of Reality. Free Press, New York.
Wilke, J., 1984. The changing world of media reality. Gazette 34 (3), 175–190.
Wilke, J., 1987. Foreign news coverage and international news flow over three centuries. Gazette 39 (2), 147–180.

Neyman, Jerzy (1894–1981)

Sandy L Zabell, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Born in 1894, Jerzy Neyman began his career as a promising Polish statistician working in Warsaw and Crakow. After meeting Egon Pearson in the 1920s during studies abroad, Neyman began a decade-long collaboration with Pearson that led to a systematic theory of statistical tests of hypotheses. When Egon Pearson succeeded his father Karl as professor of Statistics at University College London, Neyman joined him, continuing to make fundamental contributions to the theory of statistical estimation and testing. Initially on cordial terms with R. A. Fisher, the relationship between the two deteriorated, and Neyman elected to move to the University of California, Berkeley, leading to the formation of one of the most important centers of statistics in the United States. Although Neyman continued to make important contributions to mathematical statistics after his move to the United States, his efforts largely shifted to applied statistics and departmental administration.

Neyman is best remembered today for his development of the concept of confidence interval (which supplanted Fisher's earlier fiducial formulation), and his development, largely in collaboration with Egon Pearson, of the theory of statistical hypothesis testing, involving such concepts as size and power. The mathematical clarity of his approach attracted both mathematicians and practitioners. Neyman's legendary dispute with Fisher on such matters was both long and bitter. But equally important in assessing Neyman's impact on the statistical profession was his promotion of statistics at Berkeley, helping to create one of the most important academic centers of statistical research and practice in the United States.

Life

Jerzy Neyman was born in Russia, educated in Poland, did his most important work primarily in England, and was a major figure in the development of American statistics. Although a mathematician by training and inclination, his initial work and later life were primarily devoted to applied statistics. These apparent paradoxes only become understandable in the light of Neyman's formative years.

Russia and Poland

Jerzy Splawa Neyman's early years were spent in the politically uncertain environment of the final years of the Russian Empire. He was born on 16 April 1894 to Polish parents (Czeslaw Neyman, and Kazimiera Lutoslawska) in Bendery, a city in a region of repeatedly shifting allegiances to Romania, Russia, Transnistria, and Moldova. The family was descended from Polish nobility and the 'Splawa' refers to the family coat of arms. (Neyman himself preferred not to use the prefix Splawa, and it only appears in early papers.) Neyman's grandfather Hermogenes died in connection with an uprising against the Russian authorities in 1863, and 10 of his 11 sons were exiled to Siberia. Neyman's father Czeslaw, the youngest, was spared exile, but not permitted to live in Poland. He became a lawyer and later judge in Moldova, dying when Neyman was 12 years of age. The family then moved to Kharkov, Ukraine (then part of Russia).

Initially interested in experimental physics when he entered the University of Kharkov in 1912, ineptness in the laboratory led Neyman to mathematics instead. A happy suggestion by C. K. Russyan, then giving a course of lectures at the university on the theory of functions, led Neyman to Henri Lebesgue's *Leçons sur l'intégration et la recherche des functions primitives*, and a fascination with both the Lebesgue integral and measure theory. (A long paper by Neyman on the Lebesgue integral subsequently received a gold medal of the University.) Neyman also attended lectures by the famous Russian mathematician S.N. Bernstein on probability and statistics, including its applications in agricultural experimentation. Although of little interest to Neyman at the time, this exposure to applied statistics later played a key role in Neyman's career. It was also through Bernstein that Neyman became familiar with Karl Pearson's book *The Grammar of Science*, which was influential in shaping Neyman's view of science.

Unable to serve in the military during World War I because of his poor eyesight, Neyman pursued graduate studies in mathematics in Kharkov from 1917 to 1921, interrupted briefly by an arrest in 1920 by the Russian authorities as an enemy alien. In the fall of 1920, after passing his examinations for the master's degree, Neyman became a lecturer at the University; but in 1921, fearing rearrest at the time of the Treaty of Riga between Russia and Poland, he left Kharkov and Russia for Poland and Warsaw. On 4 May 1920, he had married Olga Solodovnikova, and he brought his wife with him.

After a brief time at the National Agricultural Institute in Bydgoszcz, Poland (where his brother lived), Neyman moved to Warsaw in 1922 to work at the State Meteorological Institute. In 1923 he obtained a position at the University of Warsaw permitting him to continue his studies there, and also became a special lecturer in mathematics and statistics at the SGGW (*Szkola Glowna Gospodarstwa Wiejkiego*), Poland's Central College of Agriculture. During his time in Warsaw, Neyman came in contact with other talented Polish mathematicians of his generation, including Alfred Tarski and Antoni Zygmund. In 1924 he received his doctorate, submitting a thesis on statistical problems in agricultural experimentation; his

examiners were Wacław Sierpinski, Stefan Mazurkiewicz, and his old teacher Antoni Przeborski from Kharkov. (He was not examined in statistics, because there was no one qualified to do so, in the sense of someone knowledgeable in both research-level mathematics and applied statistical theory; Neyman was, as he later put it, 'sui generis' – a unique blend of skills in the Poland of that era.)

In 1925 Neyman obtained a postdoctoral fellowship to study under Karl Pearson at University College London, where he first met some of the leading statisticians of the day, including W.S. Gosset ('Student'), R.A. Fisher, and Egon S. Pearson (son of Karl Pearson, and an important statistician in his own right). The next year Neyman obtained a fellowship from the Rockefeller Foundation to continue his studies in Paris, where he attended mathematical lectures by Borel, Lebesgue, and Hadamard. At the same time he began a collaboration with Egon Pearson lasting more than a decade, much of it by correspondence. During the summer of 1927, Neyman returned to Poland, and resumed teaching in Warsaw and Cracow.

Neyman remained in Poland from 1928 to 1934, working in both applied and mathematical statistics. He pursued a broad range of applied statistical interests in agriculture, biology, chemistry, and socioeconomics, ultimately leading to his appointment as head of the Statistical Laboratory of the Nencki Institute of Experimental Biology. But at the same time Neyman continued to work with Egon Pearson, resulting in some of their most important papers on the theory of statistical tests.

University College London

In 1933 Karl Pearson retired as Professor of Statistics from University College. Although Fisher was Pearson's obvious choice as successor, the two had been bitter enemies for more than a decade and, in Solomonic fashion, it was decided to divide the Department into two: a Department of Genetics to be headed by Fisher (as Professor of Genetics) and a Department of Applied Statistics to be headed by Egon Pearson. Not unnaturally, Egon Pearson immediately invited his collaborator and friend Neyman to join him at University College in the new Department of Statistics, initially as a Senior Lecturer and later as Reader. Neyman accepted Pearson's offer, a decision that was to have profound consequences for his career.

Despite the tensions that immediately arose between the two Departments, Neyman initially remained on cordial terms with Fisher (as evidenced both in their personal correspondence and their 1934 exchange at a meeting of the Royal Statistical Society during the discussion of Neyman's paper on confidence intervals). But their personalities and scientific philosophies were so different that conflict between the two was inevitable. Neyman's paper (1935) on the design of experiments (which criticized some of Fisher's most important work) enraged Fisher, and the clash between the two during the discussion of Neyman's paper led to a complete and permanent severing of relations between them. (See Sabbaghi and Rubin, 2014, who note there are serious errors in Neyman's paper.) Neyman's unfortunate 1961 retrospective polemic (Neyman, 1961) provides a useful illustration of just how bitter the dispute between the two became, although it exaggerates Fisher's role prior to the 1950s: after the first several years, Fisher largely ignored Neyman in print until returning to the attack in the mid-1950s.

Berkeley

In the 1930s statisticians in the United States followed the new developments in mathematical statistics in Europe with considerable interest, founding a new journal, the *Annals of Mathematical Statistics* in 1930, and a new professional organization, the *Institute of Mathematical Statistics* (IMS) in 1935. In 1938 the *Annals* became the official publication of the IMS, and acquired a genuinely international editorial board (including Cramér, Darmois, Fisher, von Mises, and Egon Pearson). Younger U.S. statisticians visited the United Kingdom for a year or more (such as Harold Hotelling and Churchill Eisenhart), and a number of statisticians from the United Kingdom were invited to the United States to lecture.

Neyman's move to Berkeley in 1938 reflected both the climate in the United States and Neyman's personal needs. In 1937 Neyman toured the United States, lecturing both at the U.S. Department of Agriculture (at the invitation of E. Deming) and a number of major universities, resulting in the offer of a position at the University of California, Berkeley. Attracted by the opportunity to build from the ground up a statistical group, Neyman accepted the offer and permanently left London for Berkeley in 1938.

Over the next several decades Neyman used his intellectual gifts, force of personality, and international standing in the statistical community to make the University of California, Berkeley, one of the leading centers of statistical research not just in the United States, but all over the world. In 1955 the continuing expansion of Neyman's Statistical Laboratory led to the formation at Berkeley of a new and independent Department of Statistics.

Throughout his career at Berkeley, Neyman continued to vigorously promote statistics at the international level as well in the United States; particularly noteworthy in this regard were the Berkeley Symposia on Probability and Mathematical Statistics, held every 5 years from 1945 until 1970. International visitors also came to Berkeley on a regular basis, such visits often either arranged by or due to Neyman.

During his long life, Neyman received numerous honors, many of these coming after his retirement in 1961. (The retirement was in name only. In an unusual exception to Berkeley's rules regarding Emeritus status, Neyman was permitted to continue as Director of the Statistical Laboratory.) In addition to being elected a member of the National Academy of Sciences in 1964, Neyman was also elected as a foreign member of the UK Royal Society, and the Polish and Swedish Academies of Science. Other honors included medals (the U.S. Medal of Science and the Guy Medal in gold of the Royal Statistical Society), and honorary doctorates (from Berkeley, the University of Chicago, the University of Warsaw, the University of Stockholm, and the Indian Statistical Institute).

Neyman remained active to the end, running a weekly seminar at which many of his visitors and guests spoke. He died in Berkeley, California, on 5 August 1981, after a brief illness.

Scientific Contributions

Hypothesis Testing

Although Neyman wrote papers both prior to his collaboration with Egon Pearson, and many after his departure for the United States, the decade from 1928 to 1938, when he collaborated with Pearson, marked the period of his most important and lasting contributions to statistics. In order to understand their work, some background is necessary.

From a Bayesian perspective, if H_0 and H_1 are competing hypotheses, E is evidence regarding them, $P(H_0)$ and $P(H_1)$ are their probabilities prior to observing E, and $P(H_0|E)$ and $P(H_1|E)$ are their probabilities after observing E, then one form of Bayes's theorem states that

$$\frac{P(H_0|E)}{P(H_1|E)} = \frac{P(E|H_0)}{P(E|H_1)} \cdot \frac{P(H_0)}{P(H_1)}$$

that is, the posterior odds in favor of H_0 relative to H_1 are the product of the likelihood ratio of the two times their prior odds.

Historically, controversy has surrounded the use of Bayes's theorem, due to differing views regarding the nature of probability; specifically, whether it is restricted to objective frequencies or can also include subjective beliefs. If the prior odds are known (or represent given objective frequencies), then the use of Bayes's theorem is not controversial; but some statisticians reject its use if the prior probabilities represent subjective beliefs. On the other hand, in many statistical applications, the hypotheses H_0 and H_1 are statistical in nature, $P(E|H_0)$ and $P(E|H_1)$ specify objective frequencies, and the likelihood ratio of the two is not in dispute.

A classical test of significance avoids the issue of the nature and value of the initial probabilities. A null hypothesis H_0 is specified, a test statistic T is computed whose statistical distribution is specified by H_0 and which has the property that (say) larger values of T call H_0 into question. Given an observed value t of T, a p-value, $p = P(T \geq t|H_0)$, is then computed. If the value of p is small, then one is faced with a dichotomy: either an unusual event has occurred (if the null is true), or the null is false.

The odds ratio form of Bayes's theorem makes clear the limits of this approach: the level of significance only takes into account the numerator in the likelihood ratio. In particular, as Egon Pearson realized in 1926 (due to a remark by Gosset), the only justification for rejecting a hypothesis on the basis of evidence unlikely under that hypothesis is another explanation which makes the evidence more likely. Pearson then began to formulate a mathematical theory (including the distinction between simple and composite hypotheses, rejection regions, and the two types of errors); but enlisted Neyman's help, recognizing his superior mathematical abilities. This resulted in a series of five fundamental papers between 1928 and 1938 (Neyman and Pearson, 1967).

These established a rigorous mathematical framework for the modern theory of hypothesis testing. The first of these, published in 1928, introduced the concepts of two types of error (type-one and type-two error), and the likelihood ratio, using the latter to derive as special cases some of the standard tests then in use. Ironically, in light of Neyman's later violent opposition to Bayesian methods, Pearson refused to be the coauthor of a second paper, drafted by Neyman (1929), because it appealed to the use of prior probabilities. Despite this disagreement, however, their collaboration continued, and in a short period of 6 years (1933–38) they proceeded to lay the foundations of the modern theory of hypothesis testing.

Their 1933 paper, on statistical tests, formalized the process of hypothesis testing by providing a clear mathematical framework. In the case of simple hypotheses, the Neyman–Pearson lemma was stated and proved; and in the case of composite hypotheses, the concepts of similar and best critical regions were introduced. Then, in a pair of papers that appeared in 1936 they further advanced the theory by introducing the concept of unbiased critical region and illustrated the use of power curves to compare tests (Neyman and Pearson, 1936a). They then discussed the role of sufficient statistics and the novel concept of uniformly powerful tests (Neyman and Pearson, 1936b). Finally, in a last paper, in 1938 they returned to the question of the existence of unbiased critical regions.

Statistical Estimation

Traditionally, estimates such as the sample mean attach some multiple of the estimated standard error to the sample statistic as an allowance for possible sampling error. In the wake of the exact small sample theory of 'Student' (Gosset), Fisher, and others, it was realized that it was possible under certain circumstances to make assertions about such intervals independent of the unknown parameters. For example, if $X_1, \ldots, X_n$ is a sample from a normally distributed population with mean μ and variance σ^2, then Student and Fisher were able to show that if $\overline{X}$ is the sample mean and S^2 is the sample variance, then

$$T := \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

that is, the statistic T has a distribution independent of the (in general) unknown parameters μ and σ (the 't-distribution on $n - 1$ degrees of freedom'). The importance of this is that this pivotal quantity can then be inverted to yield statements such as $(0 < \alpha < 1)$,

$$P\left(\overline{X} - c_\alpha \frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + c_\alpha \frac{S}{\sqrt{n}}\right) = 1 - \alpha,$$

where c_α is the appropriate percentile of the t-distribution. That is, the random interval $[L, R]$ (where $L = \overline{X} - c_\alpha \frac{S}{\sqrt{n}}, R = \overline{X} + c_\alpha \frac{S}{\sqrt{n}}$) has a specified coverage probability independent of the unknown parameters μ and σ^2, and therefore of any prior distribution on them.

This idea was introduced by Fisher in a paper in 1930, in which he gave a general procedure for constructing such intervals in the case of continuous distributions specified by a one-dimensional parameter. He described the probabilities generated this way as 'fiducial probabilities,' gave them the standard coverage interpretation, and noted that this could be done for every percentile θ_p, $0 < p < 1$. That is, for every p one could construct a statistic T_p such that $P(T_p > \theta_p) = p$. For any actual sample value of T_p, the function $p \to T_p$ is a distribution function in the classical sense, termed by Fisher a fiducial distribution, which he regarded as a substitute for the posterior distributions arising from the methods of inverse (i.e., Bayesian) probability. (It was this and only this point that later frequentists would reject.)

The extension of these ideas to the multiparameter case had been left open by Fisher in his 1930 paper. Neyman realized that if the essence of Fisher's method were regarded as the construction of a region which covered the unknown parameter with a specified probability, then one could easily extend the method to higher dimensions, which he proceeded to do in his 1934 Royal Statistical Society paper on sampling (Neyman, 1934), calling these 'confidence intervals' (in contrast to Fisher's terminology). At this stage, Fisher and Neyman were still on cordial terms, and the discussion after the paper was delivered was both polite and constructive, one of Fisher's main concerns being the lack of criteria for uniquely constructing such regions (which he thought of in terms of inductive inferences from data).

Originally Neyman deferred to Fisher's priority in discovering the basic underlying idea (that one could find sets having coverage probabilities independent of the parameter being estimated). But after his break with Fisher the next year, faced with both Fisher's criticism and indeed outright denial that Neyman's confidence intervals were fiducial intervals at all, Neyman proceeded on his own path, developing a complete theory of confidence intervals (Neyman, 1937). (For the evolution of Fisher's views on fiducial inference, and his dispute with Neyman, see Zabell, 1992.)

Sampling Theory

The 1934 RSS paper in which Neyman introduced the logic of confidence intervals was in fact primarily devoted to another subject, sampling theory. Even as late as the 1925 International Statistical Institute meeting in Rome, statisticians continued to debate the merits of 'purposive' versus random sampling. Neyman was a strong supporter of the use of randomization in sample selection, and his 1934 paper is commonly regarded as a turning point in the profession. Today, the best known result is his derivation of the fact that in stratified random sampling the optimal method of allocation (excluding cost) is to allocate the sample proportionately to the product of stratum size and stratum standard deviation (although the result was foreshadowed in earlier work of Chuprov). In brief, if j indexes the strata, N_j is the size of stratum j, S_j is the standard deviation of a quantity in stratum j, and n is the sample size, then one should draw.

$$n_j = n\frac{N_jS_j}{\sum_k N_kS_k}$$

elements from the jth stratum. But in fact Neyman's paper contains many other key elements of modern sampling theory, including the inclusion of cluster sampling within his framework, the importance of the sampling frame, and drawing upon whatever information is available (subjective or objective) in selecting strata, (as well as illustrating his ideas using the Galvani and Gini Italian census data).

Other Work

Neyman made many other important contributions to mathematical and applied statistics. On the mathematical side, these included his role in the Fisher–Neyman factorization criterion for suffciency, 'smooth' goodness of fit tests, best asymptotically normal estimation, and $C(\alpha)$ optimal tests of composite hypotheses; on the applied side, models for contagion, weather forecasting, and astronomy. Much of his later work was applied and done in collaboration with his long-time colleague and friend, Elizabeth Scott. He also wrote a textbook (Neyman, 1950).

Influence

It would be hard to overstate the impact that Neyman had on both the subject of modern statistics and the statistical profession. Although the English school of statistics, led by (among others) Galton, Karl Pearson, Weldon, Edgeworth, Gosset, Yule, and Fisher, both transformed the subject and vastly extended its scope, the process was undoubtedly hindered by concerns and doubts regarding both its conceptual underpinnings and rigorous mathematical structure. By advancing a purely frequentist setting for the subject, Neyman made the subject palatable to those suspicious of its Bayesian roots; and by presenting it as a natural branch of modern mathematics, using precise definitions and presenting many of its problems in purely mathematical guise, Neyman made it an attractive area of research for pure mathematicians.

This phenomenon was particularly striking in the theory of confidence intervals. Fisher's 1930 paper, 'Inverse Probability,' in which the term 'fiducial probability' first appears, had presented, in all essential respects, a general theory of confidence intervals for the one-parameter case. But as Fisher struggled to extend his approach to the multiparameter setting, many statisticians found his intuitive approach difficult to follow. Neyman, in contrast, abstracting out what he believed to be the mathematical essentials of the problem (the construction of sets that would contain a parameter with a long-term frequency independent of all unknown parameters), saw no difficulties in giving a mathematically precise extension of Fisher's approach, and did so in his 1934 paper. (It is indicative of the magnitude of the later break between the two that in his 1937 memoir on confidence intervals, Neyman not only did not cite Fisher's 1930 paper, but in a final footnote asserted that he had independently discovered the confidence interval for the t-statistic in 1930.) The conceptual and mathematical clarity of Neyman's approach eventually triumphed in statistical practice and textbook presentations (although even today the tension between the two approaches continues in the guise of the conflict between conditional and unconditional inference).

Similarly, the mathematical elegance and clarity of the Neyman–Pearson theory of tests made it the standard mode of presentation of the subject in many statistical textbooks despite Fisher's equally vehement opposition. (Fisher believed in, and indeed played a crucial role in the development of statistical tests of significance for a single null hypothesis, but thought that when more than one hypothesis was in question, parameter estimation rather than hypothesis testing was indicated.)

Closely tied to the acceptance of Neyman's approach was his impact on the statistical profession, especially in the United States. Under Neyman's leadership, the Berkeley Statistical Laboratory attracted and trained many of the leading figures in mathematical statistics in the decades immediately after the war. The Mathematical Genealogy Project lists Neyman as having 41 doctoral students and 2001 'descendants'

(students, students of students, and so on). His early students included Erich Lehmann, Joseph Hodges, and Lucien Le Cam, who themselves went on to have distinguished careers at Berkeley; their students in turn included such luminaries as Peter Bickel, Allan Birnbaum, Colin Blyth, Thomas Ferguson, Frank Hampel, and Stephen Stigler.

Literature

Many of Neyman's papers from 1923 to 1945 are conveniently collected in two volumes published in 1967 by Cambridge University and the University of California Press. Constance Reid's 1982 biography is an invaluable source of information on Neyman's life, based on extensive interviews with Neyman a few years before his death. Among the various biographical memoirs that appeared, notable are the ones by Kendall et al. (1982) for the Royal Society, and E.L. Lehmann (1994) for the U.S. National Academy of Sciences. Lehmann (2011) has also written a fair and balanced account of the contributions of both Fisher and Neyman to statistical science; see also Fienberg and Tanur (1996). For a statement of Neyman's philosophical perspective written towards the end of his life, see Neyman (1977). For his early work on agricultural experimentation and its connection with the later Rubin causal model, see Neyman (1923/1990). The history of the Department of Statistics at the University of California at Berkeley, including Neymans pivotal role, is summarized in Speed, Pitman, and Rice (2013). As they discuss, Neyman wrote a textbook (Neyman, 1950) that was used for several years there.

See also: Fisher, Ronald A (1890–1962); Hypothesis Testing in Statistics; Pearson, Karl (1857–1936); Statistics, History of.

Bibliography

Barnard, G.A., Plackett, R.L., Pearson, E.S., 1990. 'Student': A Statistical Biography of William Sealy Gosset. Clarendon Press, Oxford.

Fienberg, S.E., Tanur, J.M., December 1996. Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. International Statistical Review 64 (3), 237–253.

Fisher, R.A., 1930. Inverse probability. Proceedings of the Cambridge Philosophical Society 26, 528–535.

Kendall, D.G., Bartlett, M.S., Page, T.L., 1982. Jerzy Neyman: 1894–1981. Biographical Memoirs of Fellows of the Royal Society 28, 379–412.

Lehmann, E.L., 1994. Jerzy Neyman: April 16, 1894–August 5, 1981. In: Biographical Memoirs, vol. 63. US National Academy of Sciences, pp. 395–420.

Lehmann, E.L., 2011. Fisher, Neyman, and the Creation of Classical Statistics. Springer, New York.

Neyman, J., 1937. Outline of a theory of statistical estimation based on the classical theory of probability. Philosophical Transactions of the Royal Society of London, Series A. Mathematical and Physical Sciences 236, 333–380.

Neyman, J., 1938. Conferences on Mathematical Statistics and Probability. U.S. Department of Agriculture, first edition, second edition [revised and enlarged, 1952].

Neyman, J., 1950. First Course in Probability and Statistics. Henry Holt, NY.

Neyman, J., 1957. 'Inductive behavior' as a basic concept of philosophy of science. Review of the International Statistical Institute 25, 7–22.

Neyman, J., 1961. Silver jubilee of my dispute with Fisher. Journal of Operations Research Society of Japan 3, 145–154.

Neyman, J., 1967. A Selection of Early Statistical Papers of J. Neyman. University of California Press, Berkeley and Los Angeles.

Neyman, J., Pearson, E., 1936a. Contributions to the theory of testing statistical hypotheses 1: unbiased critical regions of Type A and Type AI. Statistical Research Memoirs, University College London 1, 1–37.

Neyman, J., Pearson, E., 1936b. Suffcient statistics and uniformly most powerful tests of statistical hypotheses. Statistical Research Memoirs, University College London 1, 113–137.

Neyman, J., Pearson, E., 1967. Joint Statistical Papers of J. Neyman and E.S. Pearson. Cambridge University Press, Cambridge, MA.

Reid, C., 1982. Neyman–from Life. Springer, New York.

Sabbaghi, A., Rubin, D.B., 2014. Comments on the Neyman-Fisher controversy and its consequences. Statistical Science 29 (2), 267–284.

Splawa-Neyman, J., 1990. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated and edited from the Polish original by D.M. Dabrowska and T.P. Speed. Statistical Science 5, 465–472 [with introductory remarks by T. P. Speed, pp. 463–464, and a comment by Donald B. Rubin, pp. 472–480].

Speed, T.P., Pitman, J., Rice, J., 2013. A brief history of the Statistics Department of the University of California at Berkeley. In: Agresti, A., Meng, Xiao-Li (Eds.), Strength in Numbers: The Rising of Academic Statistics Departments in the U.S. Springer, New York, pp. 317–338.

Zabell, S.L., 1992. R.A. Fisher and the fiducial argument. Statistical Science 7, 369–387.

Nietzsche, Friedrich (1844–1900)

Volker Gerhardt, Humboldt University, Berlin, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 10648–10650, © 2001, Elsevier Ltd.

Abstract

This contribution is based on the assumption that there is no philosopher whose biography is as indissolubly linked with the development of his work as Nietzsche. This is why the account of the several stages of Nietzsche's life, understood as important and molding factors for his thought, will include such elements as his descent from a family of parish priests, his links with the leading personalities of his time, and both his pathogenesis and its analysis by Nietzsche himself. It is against this background that the content of his philosophical works is outlined, the more important of which will be discussed more extensively. Nietzsche's fundamental theories are outlined, as well as his philosophical ideas such as the 'superman,' 'eternal recurrence,' 'the will to power,' 're-evaluation of values,' and 'nihilism.' This contribution provides an encompassing image of Nietzsche as a philosopher, classical philologist, psychologist, and man of letters. At the same time it is focused on his importance as the philosopher of the modern age. Finally it deals with Nietzsche's effects on art, science, and especially philosophy, from his lifetime to the present.

Nietzsche was born on 10 October 1844, in Röcken near Leipzig into a family with a long tradition of pastors. A defining traumatic experience of Nietzsche's childhood was the illness and death of his father, who died in 1849 of a brain disease; Nietzsche feared suffering a similar fate his entire life. One year later the pastor's widow moved to Naumburg on the river Saale (near Leipzig) with Nietzsche and his sister Elisabeth, born in 1846. Having attended the municipal boys' school and the Gymnasium of the diocese, the 14-year-old Nietzsche was awarded a vacancy at the renowned state school Zur Pforte. Here Nietzsche received thorough instruction and showed extraordinary talent in ancient languages and German literature. He composed, wrote poems and literary treatises. The content and problems in the philosophy of the later Nietzsche were prefigured in his writings at school, which is borne out by early evidence – even earlier than the remarkable essay on *Fate and History* (1862). Texts from his tenth year of life onward inform us of Nietzsche's intellectual development, which took the form of a preoccupation with Christian religion and morality, markedly influenced by a study of Greek thought, and was pursued, for social and cultural reasons, in a highly furtive and brooding manner.

In the fall of 1864, Nietzsche began a course of study in theology and philology in Bonn; in 1865 he continued his study of philology in Leipzig and finally decided against the theological career which had been designated for him. Although an investigation concerning Diogenes Laertius submitted for a university competition was awarded a prize, Nietzsche had plans for writing a philosophical dissertation concerning Kant and even considered changing his field of study to chemistry. Neither plan came to fruition. Instead, following completion of his military service (1867–68) with the mounted canoneers in Naumburg, the 24-year-old student was unexpectedly named Professor of Classical Philology at the University of Basel in February 1869 with a recommendation from the influential teacher Friedrich Wilhelm Ritschl.

After a brief tour in the Franco–German War, broken off after some weeks due to severe illness, Nietzsche applied in January 1871 for a vacant chair of philosophy at the University of Basel, but without success. Nietzsche had close contacts to his colleague Jacob Burckhardt, whose course of lectures – later to be published under the title *Observations of World History* – he attended in the summer of 1872. He also maintained a friendship with Richard and Cosima Wagner from the very beginning of his tenure in Basel.

In *The Birth of Tragedy*, his first publication, Nietzsche interpreted Greek culture in the spirit of Schopenhauer as the result of a reconciliation between two drives which mutually cultivate and challenge one another: the Apollinian, standing for those concepts created and formed by art, and the Dionysian, being the deeper lying, formless foundation of the existence of its counterpart. Nietzsche believed that the theoretical approach to science beginning with Socrates and continuing into the present was in crass opposition to life, and hoped this approach might be overcome through Wagner's cultural-political plans. His ambition to act as a political author and educator of the German people was formative for the composition of five lectures *On the Future of Our Educational Institutions*, as well as for a series of shorter essays which were written between 1871 and 1873 and first published in the posthumous works: *Philosophy in the Tragic Age of the Greeks*, which influenced our view of pre-Socratic philosophy decisively. *On the Pathos of Truth* (1872) and *Truth and Lie in an Extra-Moral Sense* (1873) contain Nietzsche's radical critique of metaphysics and knowledge, which seeks to prove both truth and knowledge to be measures set by humans. The *Warning to all Germans* (1873) was conceived as a programmatic essay to accompany Wagner's opera festival in Bayreuth and raised the demand for a kind of *Machtpolitik* under the aegis of culture.

Between 1873 and 1876, Nietzsche published four critical essays on culture and contemporary topics, the *Untimely Meditations*: the first work contains a sharp polemic against David Friedrich Strauss, who is represented as the very model of cultural philistinism in the academic establishment of the time. Nietzsche came out against the historicism of the nineteenth century in the second work and demanded a revision of the value of history for human life, such that historical inquiry might be placed at the service of the living moment. In the

third and fourth works, Arthur Schopenhauer and Richard Wagnerare portrayed as paradigmatic examples of the unique individual with a heroic life, to which Nietzsche attached hopes for a rebirth of culture as an aesthetic unity of all life experience.

The first symptoms of physical illness had become apparent during Nietzsche's time at Pforta. Beginning in 1873, Nietzsche was plagued by ever more frequent migraine-like attacks which forced him to interrupt his teaching. In addition, Nietzsche suffered from a chronically upset stomach and severe pain in the eyes. He took a leave of absence from the university for two semesters in the fall of 1876, after various therapies and several spa-visits showed no effect on his health. The intensity and frequency of his attacks increased so dramatically that in 1879, at only 34 years of age, Nietzsche submitted a request for his dismissal from the university. He obtained a pension settlement which was unusually generous for the time, but nevertheless extremely modest. Living on his pension from the University of Basel, Nietzsche led an unstable, lonely existence in constant search of favorable climates. He spent his winters mostly on the French and Italian Rivieras (Venice, Turin, Nice), his summers in Sils-Maria in Engadin. With the exception of occasional visits to Germany, his social contact was limited almost entirely to correspondence with Lou von Salome, Paul Ree, Franz Overbeck, Carl von Gersdorff, and Heinrich Köselitz. Preoccupation with the history of his own illness was to influence Nietzsche's concepts of 'health' and 'illness' as philosophical and existential categories, and played a significant role in his diagnosis of culture.

With the publication of *Human, All-too-Human* in 1878, and following the final break with Wagner, Nietzsche completed a transition to an independent manner of thinking in the style of an exposing critique. This committed him to a new literary style – a kind of essay writing *in nuce*. In the aphoristic works *Mixed Opinions and Sayings* (1879), *The Wanderer and his Shadow* (1880) (both published in a 2nd edition as Volume 2 of *Human, All-too-Human)* and *Dawn. With Thoughts Concerning Moral Prejudices* (1881), Nietzsche achieved fundamental insights into the origin, capacity, and importance of moral acts through his genealogical and psychological method. He thus arrived at a devastating result by showing the immorality of morals. With the loss of all transcendent and historical certainties there could only be for Nietzsche an 'experimental philosophy' founded on the basis of the perspectivism of human existence, a condition of the finite worldliness of human knowledge and action. It is this 'experimental philosophy' which Nietzsche draws from physical experience in *The Gay Science* (Books 1–4, 1882).

Nietzsche created his 'ideal' of a creative combination of all human contradictions through the literary rebirth of a Persian sage in *Thus Spoke Zarathustra* (Parts 1 and 2, 1883, Part 3, 1884, Part 4 as a private publication). Here we find the conception of the *Übermensch* or superman, as well as the doctrine of the eternal return of the same. We also find a theory of the will to power as the original unity of all mental and physical strength and potential – a theory intended to lend Nietzsche's practical expectations of a revaluation of all values a metaphysical foundation.

In the years 1886 and 1887, Nietzsche began a retrospective of his previous works, some of which he edited a new, added to, and prefaced. He thereby consciously sought a connection to his previous critical and disillusioning phase, begun with the first book of aphorisms in 1878. Simultaneously, the most important critiques of morality appeared: *Beyond Good and Evil* (1886) and *On the Genealogy of Morals* (1887). In this last work Nietzsche advances the thesis that Judeo-Christian religion has not only provoked a revaluation of all values with its life-hostile morality, but that – in an erstwhile established allegiance with modern science – it threatened the future of humanity. Nietzsche saw salvation from this threat in the merciless destruction of the old values, which would thereby both overcome nihilism and make a renewed revaluation of values possible.

In 1887 came yet another serious deterioration in Nietzsche's health. He nevertheless wrote his *Hymn for Life* at the end of the year, the only musical composition he ever submitted to press, and was extraordinarily productive in the months that followed. In the spring of 1888 Nietzsche wrote *The Case of Wagner* while continuing his main work, *Will to Power*. In the course of the year 1888 he slipped into a kind of ecstatic state which was hardly burdened by sickness. This enabled him to be immensely productive. Having broken off preparations for his main work, Nietzsche concentrated with increasing haste on other projects: a summary of the main themes of the planned work, *Twilight of the Idols* (1889) and *The Antichrist* (1894), a final reckoning with Richard Wagner in *Nietzsche contra Wagner* (1894), the autobiography *Ecce Homo* (1908), and a late collection of poems, *Dionysian Dithyrambs* (1894). Signs of a growing public interest in Nietzsche's work increased, and Georg Brandes held lectures on Nietzsche in Copenhagen.

An expression of Nietzsche's emotional state in this period are effusive letters in which he predicts immanent decisions concerning his fate. He portrays himself as playing the key role in these decisions. There is evidence of a progressive loss of a realistic view of himself by the end of 1888; Nietzsche seemed to be in a delirium of exaltation. Texts and fragments from the turn of the year document a gradual crossing over into insanity. In January 1889 Nietzsche experienced a mental breakdown, and he lived thereafter in a kind of mental derangement. His physicians' diagnosis – until this day uncertain – was progressive paralysis, that is, a continual decline of the activity of the brain, presumably as the result of a latent, long-term syphilis infection.

Following Nietzsche's brief stay in the psychiatric clinic in Basel and a year-long residence at the psychiatric clinic in Jena, his mother selflessly took over his care. Nietzsche himself was in a state of progressive dementia and lost nearly all physical and mental capacities in the course of the final 10 years of his life. Nietzsche's sister took over his care at her Villa Silberblick in 1897, following the death of their mother. Nietzsche's sister's relationship to the sick man and her connection to his work were rather ambivalent, however.

At the time of his death, on 25 August 1900, in Weimar, Nietzsche had long since become a symbolic figure in European literature. Today he is, as a philosopher, the modern classic *par excellence*.

See also: Historicism.

Bibliography

Behler, E., Montinari, M., Müller-Lauter, W. (Eds.), 1972. NietzscheStudien. Internationales Jahrbuch für die Nietzsche-Forschung. Walter de Gruyter, Berlin, Germany.

Danto, A.C., 1967. Nietzsche as Philosopher. Macmillan, New York.

Deleuze, G., 1967. Nietzsche et la philosophie. Presses Universitaires de France, Paris.

Gerhardt, V., 1999. Friedrich Nietzsche, third ed. Beck, Munich, Germany.

Janz, C.P., 1978. Friedrich Nietzsche. Biographie in drei Bänden. Carl Hanser Verlag, Munich, Germany.

Jaspers, K., 1936. Nietzsche. Einführung in das Verständnis seines Philosophierens. Walter de Gruyter, Berlin.

Kaufmann, W., 1974. Nietzsche. Philosopher, Psychologist, Antichrist. Princeton University Press, Princeton, NJ.

Kaulbach, F., 1980. Nietzsches Idee einer Experimentalphilosophie. Böhlau, Cologne, Germany.

Müler-Lauter, W., 1971. Nietzsche. Seine Philosophie der Gegensätze und die Gegensätze seiner Philosophie. Walter de Gruyter, Berlin.

Nehamas, A., 1985. Nietzsche: Life as Literature. Harvard University Press, Cambridge, MA.

Nietzsche, F., 1966. Werke in drei Bänden. In: Schlechta, K. (Ed.), Carl Hanser Verlag. Germany, Munich.

Nietzsche, F., 1967. Werke. Kritische Gesamtausgabe. In: Colli, G., Montinari, M. (Eds.). Walter de Gruyter, Berlin.

Nietzsche, F., 1975. Briefwechsel. Kritische Gesamtausgabe. In: Colli, G., Montinari, M. (Eds.). Walter de Gruyter, Berlin.

Reichert, H.W., Schlechta, K. (Eds.), 1968. International Nietzsche Bibliography, second ed. University of North Carolina Press, Chapel Hill, NC.

Ross, W., 1980. Derängstliche Adler. Friedrich Nietzsches Leben. Deutsche Verlags-Anstalt, Stuttgart, Germany.

Schmidt, H.J., 1991. Nietzsche absconditus oder Spurenlesen bei Nietzsche. Part I and II: Kindheit. IBDK, Berlin.

Schmidt, H.J., 1993–94. Nietzsche absconditus oder Spurenlesen bei Nietzsche. Part I und II: Jugend. IBDK, Berlin.

Volz, P.D., 1990. Nietzsche im Labyrinth seiner Krankheit. Eine medizinisch-biographische Untersuchung. Königshausen & Neumann, Würzburg, Germany.

Nisbet, Robert (1913–96)

Paul Gottfried, Elizabethtown College, Elizabethtown, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Among well-known sociologists, Robert Nisbet was almost unique in considering himself a 'conservative.' He defined this orientation not in terms of conventional American partisan affiliation but as a sociological and anthropological perspective. Nisbet viewed human relations as requiring organic relations, which he assumed to be hierarchical and communally determined. Such relations seemed to Nisbet, natural and even beneficial; and his lifelong interest in nineteenth-century social theory and his efforts to locate organicist positions among social thinkers of the left, especially Emile Durkheim, can be attributed to an attempt to substantiate his understanding of basic human relations. Nisbet also significantly viewed contemporary America as departing rapidly and perhaps irreversibly from what he considered to be a good society. This concern pervades his studies of social theory; and it is possible to interpret his emphasis on the 'conservative' origins of sociology and his attack on leveling public administration from the 1950s onward as related reference points. Against the social disintegration and social engineering he already criticizes in his early work The Quest for Community, Nisbet poses his vision of personal authority and communal hierarchy. This vision is present in Nisbet's writings, from his early studies of French counterrevolutionary thinkers through his final assaults on America in the present age.

The author of multiple works on social theory, the history of sociology, and social commentary, Robert Alexander Nisbet might be best remembered for The Quest for Community (1953), The Sociological Tradition (1966), and The Social Bond: lntroduction to the Study of Society (1970), all books that as late as 30 years ago were being assigned as college texts. Despite his fondness for hierarchical societies, Nisbet enjoyed the respect of those who stood on the opposite side of the political divide. His literary skills and vast theoretical knowledge account for this success, as did his unmistakably utilitarian approach to social institutions.

Though Nisbet did recognize religious mystery as an integral part of its foundation, he also defended social authority from a secular perspective, that is, as necessary for civilized life and human stability. Between Prejudices: A Philosophical Dictionary (1982) and The Present Age (1988), Nisbet moved from provisional agreement with neoconservatives on policy issues back toward an older tradition of American conservatism. Throughout his career, he underlined the need for a multiplicity of local, social groups as the natural realm of human action and as a bulwark against the centralized, unitary state of the twentieth century.

After a distinguished academic career at the University of California–Riverside in his native California, Nisbet went to Columbia University as Albert Schweitzer Professor of the Humanities. Among other numerous honors, he was a recipient of the Ingersoll Prize (1985) and in 1988 the Jefferson Day lectureship of the National Endowment for the Humanities. Nisbet was also a senior fellow of the American Enterprise Institute, and with his wife Caroline, he resided in Washington where he spent his time studying in 'self-imposed isolation from political intrigue.' His best-remembered achievement may have been his works on social theory, which continue to be widely read and which document the conservative origins of nineteenth-century sociology. Sociologists, observed Nisbet, once set out to demonstrate the truth of Aristotle's axiom that man is a social being. Ironically, he wryly noted, a discipline once identified with Aristotle's view about the value of custom and inherited order has been dragged into war against social authority.

Defining Nisbet's Identity

Although some of his interpreters have assigned to Nisbet, the label 'social theorist' or historian of social ideas, that is, someone who should be remembered for having brought forth such distinguished surveys of social theory as The Sociological Tradition (1966) and The Social Philosophers (1973), Nisbet would not have settled for this designation. His work, The Social Bond (1970) is a detailed study of the sociological method of investigation; and even his widely read defense of the 'sociological tradition,' as the creation of nineteenth-century opponents of the French Revolution and the modern 'unitary state,' is intended as a vindication of Nisbet's approach to social analysis. Nisbet would not have accepted for himself the title of being a mere chronicler of ideas – as opposed to an analyst of past and present societies, who appreciated the work of the founders of his discipline.

Clearly, Nisbet did not take a statistical approach to his subject and would have shunned the comparison of sociology, which he considered at least in part an 'art form,' as opposed to a physical science or a form of applied mathematics. What characterized his field of study were a tradition of inquiry and (although he never used the term) a social epistemology. Those whom Nisbet included as social theorists viewed individuals as being shaped by their communal contexts and as acquiring identities that were derived from their group association. Because of this understanding of his discipline, Nisbet felt no compunctions about working together with Marxist Tom Bottomore in putting together a provocative anthology, A History of Sociological Analysis (1978). The two sociologists also became close friends during this period of collaboration, on the basis of what seemed to be their shared views about the social origins of human behavior and their shared dislike for the liberal-left American academy.

A Sui Generis Conservative

Significantly, even nonrightist social thinkers who were congenial to Nisbet, like Durkheim and in some respects Marx, often become through his optic identified with 'conservative insights.' The awareness of collective identities arising from communal association became for Nisbet the point of division between good and bad social theorists. And this line of division sometimes elided in his work into judgments about sound and unsound thinkers. The author who in The Social Philosophers may have been treated the most negatively is Thomas Hobbes. In the case of this seventeenth-century English philosopher, we are told that "the only units Hobbes recognized as being metaphysically real are individuals. And the only forms of social relationship he was willing to accept are those that are geometrically rigorous logic can establish as arising from the very nature of man, individual man, with his instincts and reason. That history and convention can bring into existence social relationships which deserve legitimacy by the very virtue of history and convention was a proposition utterly foreign to Hobbes's way of thinking" (1973: 137–138).

Further, Hobbes's rejection of the validity of all intermediate associations between the individual and the state is intended to glorify centralized political power. It is with deep repugnance for this project that Nisbet observes that "the state became for Hobbes a community which does not permit within itself any lesser form of community that could conceivably challenge its unity, its indivisibility, and its absolute authority" (1973: 145).

Another influential social and political theorist whom Nisbet rejected on methodological and moral grounds was Jean-Jacques Rousseau. He insists that "as many popular movements of revolt have emanated from Rousseau's revolutionary philosophy of the general will as from Marx's idea of the revolutionary proletariat a century later" (1973: 145). "Rousseau's emphasis on the community has often been interpreted in a way that is foreign to his aim" (1973: 149). In contrast to "the social community as it existed in the thought of Thomas Aquinas or, later, in the theory of Althusius," as "a community of communities an assemblage of morally integrated minor groups," Rousseau's community "is a political one, which is indistinguishable from the state and which shares all the uniformitarian qualities of the state" (1973: 149). Rousseau avails himself of the medieval organic analogy in describing his ideal government but gives it a peculiarly modern twist: "The same centralization of control which exists in the human body must dominate the structure of the community; unity is conferred by the brain, which in Rousseau's analogy represents the sovereign power. The general will is the analogue of the human mind, and as such must remain as unified and undiversified as the mind itself. The general will, as he [Rousseau] is careful to indicate, is not synonymous with the 'will of all.' It is the will of the political organism, an entity with a life of its own quite apart from that of the individual members of which it is built" (1973: 149).

Although Nisbet had less sympathetic interest in the 'individual' than the community, he tried to deal with this concept as a historical reality that must be appreciated in its own right. When he discusses the Protestant Reformation, he deplores the thoroughness with which the Reformers attacked so many of the legacies of medieval Catholicism: "To Protestants of militant determination, everything associated with Catholic medieval society—literature, art, architecture, philosophy and scholarship, as well as outright catholic theology—was suspect" (1973: 215). Yet Nisbet also exhibited great respect for the transformative religious reform that took place in sixteenth-century Europe and which began in eastern Germany and Bohemia and eventually spread by way of Switzerland into English-speaking countries. "The clear and lasting contribution of Protestantism to political, economic and social modernity: its overriding emphasis on the individual, at the expense of all external ties, bonds, symbols, and works that are the attributes of community—that is, visible community. For, as we have seen, there was a profound conviction in the minds of Luther, Calvin, and all the major Protestant reformers, that to externalize religion is to run the risk of degrading or corrupting it" (1973: 215). Nisbet also had occasional high praise for other expressions of individual consciousness, particularly, the stress on individual reliance associated with the American work ethic, and traditional liberal constitutionalism, as revealed in the work of the American framers.

It may be relevant to distinguish between Nisbet's identification as a conservative and by now inflated use of the term to refer to Republicans or self-described members of the 'conservative movement.' Nisbet never sang the glories of 'American exceptionalism' or called for the spread of the American democratic model to the rest of the globe. From The Quest for Community (1953) onward, he targeted American crusades for democracy and American popular culture. In the epilogue to The Present Age, he depicts the United States as "a giant in military resources but not in the exercise of military power and responsibility. Befuddled by belief that God intended it to be the moral teacher to the world, our giant stumbles from people to people, ever eager to demonstrate that what American touches, it makes holy" (1988: 107). America is "a giant too in domestic bureaucracy," (1988: 107) which generates an "awful total of indebtedness" (1988: 107) and growing dependence on public administration.

Although Nisbet was a true product of the New World, born and raised in Southern California, and except for wartime service in Saipan, someone who spent most of his life in Southern California and Arizona, he never praised America, the Good. He was relentlessly critical of twentieth-century American life, and although by his own admission a man of the right, he scorned a moralistic foreign policy, consumer capitalism, and managed democracy. Nisbet frequently critically comments about "progress in anarchy in modern America" (1988) and the "twilight of authority" (2000) are produced with reference to his own society, and particularly the American academy.

In the conclusion of The Quest for Community, Nisbet sums up his unseasonable thoughts about modern centralized democracy. With the modern administrative state we are dealing with a "new classification" (1970: 281). "Traditional labels—democratic, republican, capitalist, socialist, etc.—have by now become nearly as archaic as older classifications of monarchy and aristocracy" (1970: 281). What is more, "government of, by, and for the people, for all its verity as an abstract proposition, becomes nearly as irrelevant in a world

in which all despotisms rest upon foundations of mass acquiescence and where the arts of political propaganda are employed to sink the roots of government deeply in popular consciousness and participation" (1970: 281).

Going beyond the prevalent distinction between democratic and totalitarian states, Nisbet arrives at a position in Quest for Community that may be unsettling for some of his readers. He treats the difference between democratic and nondemocratic regimes as being of only secondary importance relative to other changes that were taking place: "We must conclude that all States in the future will be able to demonstrate, and will have to demonstrate attributes of socialism. But, by themselves, these will promise nothing in the way of freedom" (1970: 281). It is consolidated public administration, promoting 'equality' that Nisbet has in mind when he speaks about 'the kind of state' that is always seeking "a higher degree of centralization in the conduct of its operation, always tending toward a wider measure of politicization of social, economic, and cultural life. [The state] does not do this in the name of freedom but of freedom from want, security, and minority tyranny" (1970: 282). Nisbet's attack on the 'type of state,' in which all human and social activities, "are made aspects of the administrative structure of political government" (1970: 282) is not an allusion to a historically remote event like the Terror of the French Revolution. Nisbet has in mind an imminent threat, namely, an American government that no longer protects a "pluralism of functions and loyalties in the lives of its people" (1970: 283).

Nisbet warned against the perils of 'democratic totalitarianism' quite early in the postwar years (1970: 283). His admonitions relate to the surge of centralized power that began with the Progressive Era and which received added impetus from American participation in two World Wars and the Cold War. Significantly, these admonitions came before the vast expansion of public administration, which started in the 1960s in the United States, Western Europe, and the British Commonwealth. A development of this magnitude was only incipiently present in most Western countries until after Nisbet had become an academic celebrity. Such governmental projects as fighting discrimination and punishing hate speech were not really issues when Nisbet produced The Quest for Community. Yet, he predicted the extension of public administration into forms of thought control that were still limited to hard totalitarian states in the 1950s. The author of The Quest for Community also presciently notes that the democratic administrative state creates its own "consensus for change" (1970: 87). It achieved this goal by molding 'popular consciousness' through its command of public education. Particularly, in his recognition that the expanding modern state always claims to be representing an expansion of freedom, Nisbet was looking well beyond the post-World War II era (1970: 86).

Against Constructivism

If there is any position that suffuses Nisbet's judgments about earlier social and political theorists, it is his pervasive rejection of the notion that civil society is composed of inherently autonomous individuals. This may explain why Nisbet had such an exceedingly broad taste in social theory, ranging from Bonald, Burke, Maistre, Alexis de Tocqueville, and Georg Wilhelm Friedrich Hegel down to the French Jewish socialist Emile Durkheim. These thinkers all rejected the notion of an atomistic society and warned against a power-hungry state that does not recognize the solidarity of family and communal relations.

Unlike his other heroes, Nisbet bestowed on Durkheim an entire intellectual biography; and anyone who reads Nisbet's Sociology of Emile Durkheim (1974) will recognize the impact of this French sociologist on Nisbet's thinking. His biographical work makes generous allowances for Durkeim's decidedly leftist politics, and Nisbet expresses no negative judgment concerning Durkheim's close friendship with the Marxist socialist Jean Jaurès. Nisbet valued his subject not for his partisan political positions, but for his emphasis on organic relations. Already Durkheim's earliest major work, The Division of Labor in Society (1892) stressed the bonds uniting real people, whether through 'mechanical solidarity' in tribal societies or through a more voluntaristic 'organic solidarity' in the modern West (Nisbet, 1974: 30). The concept of individuals living apart from each other and trying to achieve autonomy outside of complex social relations seemed to Durkheim as much as Nisbet to be a pure fiction: "Durkheim rejected individualism on every possible ground. He found it insupportable as a principle of social solidarity, as an ethic or moral value, as a cornerstone of the social order, and not least as the vantage point of social analysis" (1974: 16). Nisbet infers from this attitude and Durkheim's approach to his subject evidence of his "specific criticisms of the individualistically oriented titans of his day, such as Herbert Spencer and Gabriel Tarde" (1974: 16).

We learn that in his Rules of Sociological Method (1895) and in the journal he founded in 1898 L'Année Sociologique, the distinguished Sorbonne professor raised 'structural functionalism' to the prevalent form of social research. From this vantage point, according to both Durkheim and Nisbet, the researcher can approach his area of investigation 'holistically.' Neither an unadorned 'biologism,' that is, a preoccupation with individual drives and instincts, nor an emphasis on individual action, we are warned, should guide social research. Rather the researcher should be looking at how societies function collectively.

Despite Durkheim's secularist, anticlerical politics during the French Third Republic, Nisbet points to the fact that he published a ground-breaking work on religious sociology, Elementary Forms of Religious Life (1912). His biographer lavishes praise on this study: "Of the eternality of religion resting as it does on man's differentiation of the sacred from the profane, there is not the slightest doubt in Durkheim's mind. Durkheim may well have been atheist and disbeliever. The Elementary Forms of Religious Life is nevertheless one of the most powerful statements of the functional indispensability in society ever written" (*The Sociology of Emile Durkheim* 1974: 37).

A religious skeptic, albeit formally a High Church Episcopalian, with a German Jewish grandmother, Nisbet recognized in Durkheim his oft stated appreciation of religion as a social phenomenon. He and Durkheim were both aware of the value of the 'Sacred' in man's collective existence; and in the long

section on the 'Sacred' in The Sociological Tradition, one easily detects the legacy of the author of The Elementary Forms of Religious Life. From his glowing description of this work, it is clear that Nisbet had discovered certain conceptual overlaps between his subject and himself. He accentuates Durkheim's sound views about the continuity of the Sacred, his holistic approaches to the study of social phenomena, and his understanding of organic relations as necessary for fully developed social beings. Finally, as Nisbet mentions in the closing section of the biography, he and Durkheim shared forebodings about the effects of fragmented authority, a problem that had already emerged by the second half of the nineteenth century.

There is, however, another link between the biographer and his subject. Like Durkheim, Nisbet was searching for a modern basis for community that would lead away from the dangerous collectivism depicted in Quest for Community. The prescriptions he uncovered in European counterrevolutionaries, no longer had purchase in the modern world. There was no chance of going back to a system of 'personal authority' or to status relations of the kind envisaged by Bonald or Maistre. The social bond had to be preserved or renewed in some other way, particularly in the face of expanding public administration – and given the tendency of the modern state to push certain ideas about equality into ever more extreme forms.

Nisbet did not believe that the end of the Soviet Empire and the declining appeal of Marxist–Leninism for Western intellectuals would render the egalitarian idea less compelling. Rather, he assumed the quest for equality would become even more frenetic, as the 'democratic,' planning state triumphed as a vehicle for leveling down human differences. The disintegration of an old left based on Marxism was taking place in Western countries as public administration was taking over social and economic functions once reserved to families and communities. By the time Nisbet was writing The Present Age in the late 1980s and excoriating 'the new absolutism,' he had come to believe that the accumulation of centralized political power may have become the most frightening aspect of what he referred to as 'the crisis of the present age.'

This was developing apace with what Nisbet in the same work characterized as 'the loose individual,' and in a famous, slightly sarcastic definition he lets us know: "What sociologists are prone to call social disintegration is really nothing more than the spectacle of a rising number of individuals playing fast and loose with other individuals in relationships of trust and responsibility" (1988: 87). Nisbet's tract The Degradation of the Academic Dogma (1978) may be seen as another statement of his consternation at the collapse of responsibility in a society in which communal and ancestral bonds had been frayed. The young were free to act out; professors saw themselves as political revolutionaries rather than as custodians of an inherited civilization; the only real power broker was the Hobbesian sovereign state, which increased its power with the social breakdown that was taking place all around it. Intermediate institutions were under attack.

Nisbet's preoccupation with Tocqueville's discussion of communal and familial association as the substance of American democracy in the 1830s, his focus on Durkheim's notion of mechanical solidarity, and his references to seigniorial authority among classical conservatives, were all in different ways responses to the 'crisis of the present age.' Nisbet quotes Tocqueville in The Social Philosophers in the context of a consideration of the American democracy that Tocqueville observed in his travels through the newly formed United States. Unlike centralized states in Europe, the American federal government accepted "the principle of division of authority in society" (1973: 427). This "diversification of authority" helped to keep the state from engulfing civil society (1973: 427). There was also a 'federal system of political government,' in which different branches had assigned distinctive powers at the federal level and a division of authority existed between "the national government and the government of the component states" (1973: 427). These carefully devised arrangements worked to preserve a form of democracy that was "federalist, decentralist, and pluralist" (1973: 428).

But Tocqueville, says Nisbet, also saw a 'dark side of democracy,' the propensity of democratic governments to plunge into total wars that wreck destruction on constitutional liberties and the continued possibility of divided authority (1973: 426). War, as understood by Tocqueville, who pointed here toward his interpreter, "is more likely to leave its stamp of regimentation and despotism upon societies under democracies than other types of government" (1973: 429). Although, slower to engage in wars than traditional dynastic states, democracies "are more likely, once war is under way, to extend its duration as well as its social, moral, economic, and political scope" (1973: 426). "The effect of democracy is to give a mass basis to war and to moralize it-that is, to endow it with spiritual and moral objectives that serve to hide underlying material conditions. Moreover, the centralization of government to which democracies normally tend in their process of administration is greatly enhanced by war" (1973: 426).

Structural Defects of Late Modernity

After his comments on personal authority in The Social Bond, which "strongly emphasizes the corporate solidarity of the group," Nisbet goes on to broach a gloomier topic (1970:126). He perceives "a revolution in the nature of authority" that began centuries ago and which culminated in having "personal authority superseded by territorial authority" (1970: 126). Nisbet brings this into line with another tendency that is inherent in the modern state, namely its assault on intermediate institutions as standing in the way of centralized control. The centralized state, whether or not it called itself democratic, seized on military challenges to expand its administrative reach and to manipulate isolated individuals. And as time went on, there were fewer and fewer of those safeguards that the French political theorist Bertrand de Jouvenel called 'makeweights' as a counterforce to the existing power relations.

Given this unpromising historical situation, it was natural that Nisbet looked for 'conservative' antidotes wherever they seemed usable. His analysis of Durkheim and Alexis de Tocqueville and his respectful exposition of such European conservatives as Le Comte de Bonald, Edmund Burke, and the Hegel of The Philosophy of Right all had direct bearing for him on modern American government and society. A historian of social theory, he extracted from his journeys into the past whatever concepts could be applied to understanding his own age and to addressing its problems.

This mind-set was already operative in his dissertation The Social Group and French Thought submitted at Berkeley in 1939 and dealing with the French counterrevolutionary thinkers Bonald and Maistre. Nisbet's earliest publication, which was printed in the Journal of Sociology in November 1943, together with a related treatment of the same subject in the same journal in September 1952, provides a learned discussion of social critics of the right. Nisbet became interested in these thinkers while looking for a thesis theme that he might propose to his dissertation director, the famed intellectual historian Frederick Teggart. His topic was not picked at random but reflected Nisbet's social concerns as a graduate student.

The Search to Hold Things Together

At the time he was producing The Quest for Community, Nisbet still nurtured hope that the nascent 'conservative movement,' viewed as an intellectual force, could, in the words of his friend Russell Kirk, 'redeem the times' (1988: 64). His descriptions of this hope, as stated in The Present Age, should be quoted in full because it was recalled more than 30 years later as a disastrously lost opportunity:

> A veritable renascence of conservative ideology was under way by the end of the fifties; it was sufficient to carry with it an interest in both Edmund Burke and Tocqueville greater than perhaps in any prior decade. Overwhelmingly the new conservatism—in resolute opposition to liberals above all other groups—followed Burke and Tocqueville in espousing decrease in centralization, pluralism over monism in government, the free market in basic economic production and distribution, intermediate social groups like family and local community and voluntary associations—all calculated to take some part the load of responsibility from big government—and, inevitably, substantial decrease in bureaucracy. The new conservatism also emphasized some of the traditional moral values which, it plausibly argued, had gotten battered into passivity by the forces of modernism, political modernism most of all. In a word, the *autonomy* of social order and culture was the prized objective of the new conservatism.
>
> 1988: 64

Having stated this onetime cause for optimism, Nisbet proceeded to note: "At the present time, however, militant conservatism has as little to do with its historic substance as contemporary liberalism has to do with its birthright of devotion to individual liberty. What is most likely to be labeled 'conservative' by the media—and with considerable basis in reality—in militarism on the one hand and Christian Far Right evangelicism on the other, which is far more interested today in extending the power of the state—than in a free religion in a free political society" (1988: 64). Nisbet viewed the Reagan coalition as a loose alliance of sometimes discordant ideologies, which was trying to impose a unified face on "military hawkishness, evangelicism, libertarianism, supply-siders, the power-obsessed Right" (1988: 64). He was surprised 'the Reagan coalition lasted as long as it did' and he doubted 'it will be missed' after it falls apart (1988: 64).

There is a noticeable randomness in how Nisbet goes after his targets in The Present Age. Two of his favorite targets there are the far right and the Religious Right, although we may ask whether he himself had not tried to vindicate a more rightist European conservatism than any right he attacked in his later writings. Nisbet defended counterrevolutionaries who had upheld an alliance of throne and altar. Was the modern Religious Right, which he mocked in the 1980s, more extreme in its opposition to the modern social left than were those counterrevolutionaries whose teachings Nisbet consciously drew on? One of Nisbet's favorite authors, Tocqueville, was favorably impressed on the whole by the Protestant piety and morality that he encountered during his travels in the New World.

There is a palpable looseness in how Nisbet locates his enemies, particularly in the 1980s. Was the desire for military intervention a peculiarly right-wing tendency during most of Nisbet's life? In Nisbet's youth well into his adult years, such interventionism may have been more characteristic of Democratic liberal internationalists than of right-wing Republicans, who long included in their ranks a massive isolationist wing. The neoconservatives and neoliberals of the 1980s who favored a large American military presence overseas did not come out of the far right. They saw themselves as being faithful to the liberal internationalist foreign policy of someone who may have been Nisbet's least favorite American president Woodrow Wilson. Like Wilson, later liberal internationalists were eager to 'make the world safe for democracy,' a goal that required frequent overseas intervention, although it is hard to understand why this aggressive human rights idealism should be considered typical of the far right (1988: 9).

Curiously Nisbet did not place Wilson, who was one of his arch-villains, on the right. He viewed him as first and foremost an architect of the 'new absolutism,' a topic that receives extended treatment in The Present Age: "The blunt fact is that when under Wilson America was introduced to the War State in 1917, it was also introduced to what would be later known as the total, or totalitarian, state. There is this important point to add: the acts which transformed laissez-faire, entrepreneurial America into a total state for the duration were acts of Congress, not of a revolutionary minority as in Russia and Italy" (1988: 47). In Wilson's rule Nisbet discerns the coming together of democracy and the total state. Wilson used the powers of government to push the country into a devastating European war, then regulated the economy and suppressed opposition, all in the name of fighting for democracy. Although some of his wartime measures were later suspended or rescinded, Wilson's regime from Nisbet's perspective was "a brief forerunner to the kinds of states that would be 'managerial states' in the thirties, including those known as totalitarian" (1988: 47).

This was Nisbet's conclusive judgment about a towering managerial state, with an easily scared up democratic mandate for expanding its power. This judgment may explain why Nisbet disliked the Religious Right so profoundly. It had thrown in its lot with the centralized managerial state, which it expected quite unrealistically to do its bidding. Religionists had come to worship the 'new absolutism' as much as their opponents. But no clear terminological framework was available for what Nisbet saw as the political danger of the present age. What he described lay beyond clear right–left distinctions. He was talking about a development that could be labeled in different ways, and he assigned the forces he described in The Present Age to the right or the left almost indifferently. He noted how foreign war leaders during World War I acknowledged the

thoroughness of the American effort to mobilize its population around a crusading ideology. According to Nisbet, the centralized democratic character of Wilson's regime made it easy to root out opposition, with popular approval.

He also saw a natural progression from a certain type of 'national community' to the celebration of a world government based on ideas about human rights. Both represented stages in the expansion of centralized managerial government; and in his depiction of such activist presidents as FDR, Nisbet saw the beginning of politics in a new key. FDR's 'mesmerizing idea of national community' was the attempt to enact Rousseau's vision of a sense of solidarity that could be engineered by state planners. This project, according to Nisbet, went back to the Progressive Era, and it pointed toward a planned future for humanity, which Nisbet believed he already discerned.

Given his sense of social and political desperation about lost freedoms and community, it is easy to understand Nisbet's irritation with proposed political action intended to remedy a situation he considered to have been largely created by the modern democratic state. It may also account for why he often sounded like a libertarian, particularly in his later years. This may have seemed strange for someone who had a running feud with libertarians for many decades. In the 1960s, Nisbet and Russell Kirk were singled out by the conservative libertarian Frank S. Meyer as representatives of a 'new conservatism' that exalted state power. Although it is hard to find any predilection in Nisbet for big government, it is true that he disliked libertarians as constructivists and scorners of the social bond. Some of his harshest invective was reserved for Meyer's persuasion. It is, therefore, all the more remarkable that libertarians now quote Nisbet as a critic of centralized administration and as someone who spent his life attacking the modern state. This may be because if no other reason their shared enemy.

See also: Alienation, History of; Alienation, Sociology of; Culture, Sociology of; Elites: Sociological Aspects; Family and Kinship, History of; Family as Institution; Personality and Religion; Sociological Traditions; Solidarity, Sociology of; Tradition: Social.

Bibliography

Gottfried, Paul E., 2002. Introduction. In: Sociology as an Art Form. Transaction Publishers, New Brunswick, NJ, pp. IX–XVI.

Gottfried, Paul E., 2006. Entry on Robert Nisbet. In: Frohnen, Bruce, Beer, Jeremy, Nelson, Jeffrey O. (Eds.), American Conservatism: An Encyclopedia. ISI Books, Wilmington, DE.

Gottfried, Paul E., 2009. Encounters: My Life with Nixon, Marcuse and Other Friends and Teachers. ISI Books, Wilmington, DE.

Meyer, Frank S., 1962. In Defense of Freedom: A Conservative Credo. Henry Regnery Company, Chicago.

Nash, George H., 1996. The Conservative Intellectual Movement in America, second ed. ISI Books, Wilmington, DE.

Nisbet, Robert A., November 1943. The French revolution and the rise of sociology in France. The American Journal of Sociology 49, 156–164.

Nisbet, Robert A., September 1952. Conservatism and sociology. The American Journal of Sociology 48, 167–175.

Nisbet, Robert A., 1953/1970. The Quest for Community. Reprint. Original Copyright 1953. Oxford University Press, New York.

Nisbet, Robert A., 1966. The Sociological Tradition. Basic Books, New York.

Nisbet, Robert A., 1970. Tradition and Revolt. Random House (Vintage Books), New York.

Nisbet, Robert A., 1971. The Degradation of Academic Dogma: The University in America 1945–1970. Basic Books, New York.

Nisbet, Robert A., 1973. The Social Philosophers: Community and Conflict in Western Thought. Thomas Y. Crowell Company, New York.

Nisbet, Robert A., 1974. The Sociology of Emile Durkheim. Oxford University Press, New York.

Nisbet, Robert A., 1976. Sociology as an Art Form. Oxford University Press, New York.

Nisbet, Robert A., 1982. A Philosophical Dictionary. Harvard University Press, Cambridge, MA.

Nisbet, Robert A., 1988. The Present Age: Progress and Anarchy in Modern America. Harper& Row, New York.

Nisbet, Robert A., 1992. Teachers and Scholars: A Memoir of Berkeley in Depression and War. Transaction Publishers, New Brunswick, NJ.

Nisbet, Robert A., 2000. The Twilight of Authority. Liberty Fund, Indianapolis, IN.

Nisbet, Robert A., 2001. Conservatism: Dream and Reality. Transaction Publishers, New Brunswick, NJ.

Nisbet, Robert A., Perrin, Robert G., 1977. The Social Bond, second ed. Alfred Knopf, New York.

Perrin, Robert, 1999. Robert Alexander Nisbet. Proceedings of the American Philosophical Society 143, 695–710.

Stone, Brad L., 2001. Robert Nisbet: Communitarian Traditionalist. ISI Books, Wilmington, DE.

Nomads and Nomadism in History

Fred Scholz, Free University of Berlin, Berlin, Germany
Günther Schlee, Max Planck Institute for Social Anthropology, Halle (Saale), Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by F. Scholz, volume 16, pp. 10650–10655, © 2001, Elsevier Ltd.

Abstract

Pastoral nomadism has evolved out of mixed farming. It is a form of specialization in animal production at the expense of crop growing. It may have come into being when some herds became too large for all of them to stay around a village in a radius which allowed them to come back to the village daily or in the course of a year (transhumance), and mobile camps started to move independently, comprising entire families, without fixed abodes (nomadism). Another factor which periodically has contributed to the emergence of nomadism is a decline in crop production due to climatic change. This process has occurred in history repeatedly whenever these conditions were given. Nomadism thus is not an archaic way of life but a successful adaptation to changing ecological conditions and economic opportunities.

Nomads or nomadism has often been stereotyped as violent, threatening, or primitive, and backward. On the other hand, the transition to a settled life and a sedentary economy (sedentism) was considered to be an indicator of and a prerequisite for progress and civilization. However, according to current ethnoarcheological and sociogeographical research nomadism should be considered a rather sophisticated, economically successful, and sustainable way of life practiced by inhabitants of the Old World dry belt and the reindeer nomads of the Eurasian tundra (see Khazanov and Schlee, 2012 for a comprehensive treatment of both zones; in this article the focus will be on the dry belt, the much larger zone of pastoralism and historically the much more important one because of its influence on other civilizations). In response to ecological natural and social political conditions, pastoralists have developed an optimal form of safeguarding survival by using pasture animals and adopting a mobile way of life. Given the right conditions, nomadism was able to appear or disappear again and again in different places at different times. As the article shows, this 'appearance' and 'disappearance' of the nomads runs through the entire history of the Old World dry belt, proving that nomadism is as a constitutive element of society and a resilient form of economy within the Old World dry belt. As a politicomilitary factor, as builders or destroyers of states, the influence of nomads extends far beyond the dry belt to the fertile areas all around it.

The Term 'Nomadism'

Nomadism is a specific way of life practiced in the dry belt of the Old World (**Figure 1**), a socioecological mode of culture (sozioökologische Kulturweise) (Scholz, 1995), whose internal processes, governing factors, and external appearance essentially obey the elementary 'law' of safeguarding survival. Within the Old World dry belt, nomadism was able to reappear in new and original forms, at all times and places, fundamentally separate from the evolutionary stages of sedentary life. Nomadism is oriented toward living 'in' and 'with' nature, rather than dominating and exploiting it; as such, it represents a region-specific mode of subsistence that makes optimal use of ecological possibilities and sociopolitical conditions in order to survive. Thus it is an elementary, constitutive, and independent component of the cultural and social evolution of this dryland zone.

Debatable 'Stage Theories'?

Pohlhausen (1954: pp. 153, 167) still believed that pastoral nomadism started around AD 950 and that evidence existed of its monocentrist origin. But even the extensive ethnoarcheological studies of the past decades have failed to establish when and where nomadic peoples first appeared. The emergence of nomadism – a significant event in social history – may well have been due to "the herder of the steppe and oasis farmer having broken away from his agriculturalist brother" (Wissmann, 1961: p. 28). According to Lattimore (1940) this, or a similar process, may have occurred at the Inner Asian frontier, the marginal region between settled agriculture (practiced by the Chinese) and nonsedentary livestock-keeping (practiced by the Mongols). Indeed, it could have occurred anywhere within the Old World dry belt (**Figure 1**).

Plausible concepts of the origin of nomadism have sought to convey various (three-) stage theories of cultural and economic history (Vajda, 1968). While avoiding a lengthy discussion of such explanations – which have largely outlived their relevance – mention should at least be made of Adam Smith (1723–90) and Friedrich List (1789–1846). In his comments on the evolution of 'productive forces' in agriculture, List (1961: p. 57) expressed the opinion that, in genetic terms, the nomad represents a preliminary stage before the agriculturalist. This view prevailed from the nineteenth century up to late in the twentieth century and believed to be able to locate the center of origin of nomadism in the Altai region. This theory may seem logical at first sight, but in the end it does not fit the facts and has been considered refuted since Hahn (1891: p. 487) postulated that mixed farming was the prerequisite for the origin of nomadism.

More complex theories of origin were developed by, among others, Pohlhausen (1954) or Dittmer (1965), who believed the predecessors of pastoral nomadism (Hirtennomadismus) to have been alpine mountain nomadism arising from plow

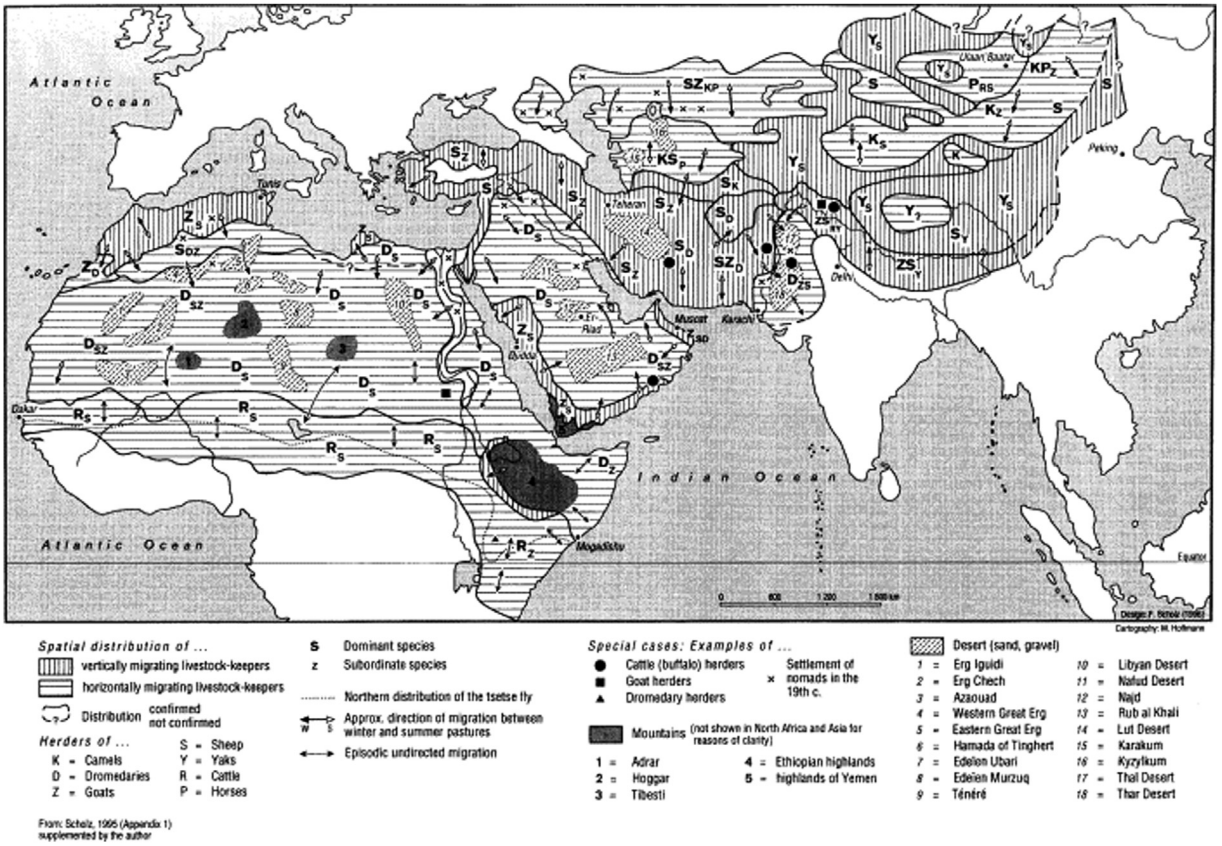

Figure 1 Nomadic livestock-keepers in the Old World dry belt (*c.*1900).

agriculture and subsequently, transhumance and cow herding (Küherwirtschaft). By contrast, Bobek (1959: p. 272) sees nomadism as a branch of clan farming (Sippenbauerntum) and relates this development 'on the one hand to the availability of extensive pasturelands that are only suitable for seasonal use and not really for permanent settlement and, on the other, to the dual economic interests of mixed farmers,' who are able 'to change their mode of subsistence relatively easily and quickly by shifting the emphasis from cultivation to animal husbandry and vice versa.' Basically similar ideas are put forward by Vajda (1968), special mention being made of the conception of a repeated alternation between nomadic and farming life as a result of natural catastrophes, armed conflicts, and economic setbacks. According to de Planhol (1975) the constant demographic surplus of the nomads substantially contributes to this situation.

These different views are plausible, but by no means imperative, in terms of time and space. The processes they describe could have happened anywhere and at any time, even long before the first written references (fourth millennium) to nomads.

Ethnoarcheological Findings

These cultural–historical views are, in the final analysis, conclusions drawn from archeological and ethnoarcheological research. Convincing work has recently been published on nomadic societies, which are generally not amenable to research methods based on permanent material evidence (Francfort, 1990). These studies have thrown light on prehistoric forms of dwellings and camps, their primary theme being the repeated alternation between sedentary agricultural and mobile livestock-keeping groups, i.e., they provide evidence of the recurring emergence of nomadic groups, of nomadism. These processes have been described for Baluchistan, for the Negev Desert, and for the Eastern Desert of Egypt. The view that the nomad did not solely evolve from the steppe farmer has been represented in various models. According to Lees and Bates (1974) the initial stage of nomadism was the transition from agriculture with animal husbandry to specialized irrigation-fed agriculture. On the one hand, this involved a population increase, specialized division of labor, and higher food requirements. On the other hand, animal husbandry was driven away from irrigable land, further and further into the steppes and mountains. Seasonal wanderings over increasing distances were the result, with nomads keeping their own separate herds. The different needs of the nomads and oasis dwellers to safeguard their existence, led to a symbiotic relationship of conflict and coexistence between the groups. The course of this relationship depended on the respective military power and was conductive to the nomads' organizational independence. With reference to Lattimore (1940), Nissen (1980: p. 287) developed a similar hypothesis regarding early

Babylonia whose history is marked by recurring invasions by 'so-called nomadic groups … in the fertile grounds.'

In fourth-millennium Sumerian texts, there are already many pointers to the repeated abandoning of agriculture in favor of mobile livestock-keeping or to the separation of livestock-keepers from agriculturalists; this process probably occurred at the start of the first millennium. Gilbert (1983) even puts forward the view that the developing urban centers the Zagros mountain region caused an increase in the demand for animal products. Simultaneously, however, more intensive agriculture was accompanied by a reduction in local animal husbandry, and niches opened up for mobile livestock-keeping and the nomadic way of life. This process is supposed to have first occurred in the seventh millennium and then again in the third and first millennia BC.

If the domestication of herd animals is accepted as the prerequisite for nomadism, then nomadism presumably did not exist in North Africa before about 5500 BC when the Sahara was a grassland niche which was increasingly occupied by pastoral people. Similarly, livestock-keeping communities would not have occurred in East Africa prior to about AD 500.

Recent archaeological evidence suggests that these dates have to be pushed back in time. While most domestic ungulates have been domesticated on the Eurasian continent and have penetrated into Africa in the care of humans, North Africa seems to have been a center of domestication of cattle of the *Bos taurus* subspecies independent of other centers of cattle domestication (The Fertile Crescent for *B. taurus* and the domestication of *Bos indicus* (the zebu) further east). Remainders of domesticated cattle in association with domesticated sorghum, other grass seeds, and pottery in the Western desert of Egypt point to cattle having been domesticated there already before 7400 BC (Wendorf and Schild, 1998). Most likely, "these cattle were first domesticated at an earlier, but unguessable, date in the Nile Valley" (Wendorf et al., 1987: p. 447). These findings push both plant domestication (agriculture) and cattle domestication back in time, without necessarily affecting their sequence. It is plausible to assume that universally, plant domestication preceded that of ungulates, as has been discussed above, in connection with the theories of different 'stages.' (For a discussion, see Schlee and Shongolo, 2012: pp. 6–10). From the time of these early agro-pastoralists onward, separations of people who specialize in livestock and become mobile may have occurred.

The transition to mobile forms of pastoralism goes along with changes in social organization. Generally, locality-based forms of organization recede and descent-based ones gain in importance. Some of the most elaborate genealogical systems (Somali and Bedouin Arabs) have been collected among pastoral nomads. Like their households, mobile pastoralists need to be able to take their group affiliations along wherever they go. A group affiliation which is primarily defined by residence in one place, however, becomes implausible when one has not been to that place for a long time.

Apart from the fact that their descent systems tend to be more elaborate, there is little commonality between pastoralists in this regard. The types of descent system they have differ widely. Patrilineal forms of descent reckoning prevail, but there are also the Tuareg where children of free-born mothers. In segmentary lineage systems, segmentation may occur whenever a man has more than one son, and smaller groups continuously become larger and genealogically deeper as new generations are added. These groups move up the scale to acquire the functions of the groups which were formerly the large ones and now have become very large and may fade away to become repositories of remote relations, while new small groups emerge at the bottom, in the growth zone, in the present where people marry and procreate. Apart from such segmentary lineage systems which are in continuous flux, there are social orders based on clans which are thought to be immutable and to have even fixed collective characters and biological peculiarities. These clans may also have higher time stability in reality, because otherwise such beliefs would have difficulties to persist.

Descent links someone to his ancestors (mostly, according to the form of descent reckoning, just one line of ancestors) and his progeny. If a descent group is narrowly defined, it severs rather than creates links to people of one's own generation. What mobile pastoralists need, however, is far reaching links to allow them to meet someone of their own (by one set of criteria or another) wherever they go. It is therefore useful to multiply the categories in which one can include someone as 'one's own.' If someone does not belong to ones patrilineal clan, he or she can belong to oneself in some other way. Links through women (affinal relations, uterine descent) are preserved by memory and appealed to when the need for that arises. Knowledge of history is also useful for pastoral strategists who seek military alliances or access to pastures in the vicinity of other groups. Descent units may have split in the past and relatives have ended up in different ethnic groups or tribes. The existence of such interethnic clan relationships may provide one with a far reaching net of relationships which one can fall back on in times of war or drought. The examples of social relations which are based on or complement and supplement shared unilineal descent could easily be multiplied.

Social institutions which systematically bridge the lines of fission between descent groups and create links between all people of one generation or similar age are the age set or generation set systems. Some of the most elaborate such systems have been recorded among pastoral groups (Boran Oromo, Maasai, and related groups). In addition to a ritual order for society which was believed to assure God's blessings like sufficient rains and protection against enemies, such systems did provide regular occasions to meet and engage in politics, and, on a smaller scale, solidarity and assistance on an individual basis, like hospitality for a traveler. Before colonial pacification it also was a means of military recruitment, and with the decline of statehood in many parts of Africa and the decentralization of violence, in many regions it has acquired that function again (Schlee, 2005).

Descent and age-grading systems, however do not appear to have satisfied the need for formalized social ties in pastoral societies. Even in setting where we find both systems, we also find a third, namely stock friendships. Giving and taking, including delayed reciprocity, are human universals. In the case of pastoral nomads, however, they are the basis of formalized institutions, namely various forms of animal loans and stock friendship. In stock friendships gifts may create obligations and loans may establish what has been called 'shared rights' in an animal. In fact, in the cases which have come to my notice, it is

not the same rights which are shared. One should rather say that different rights are held by different people in the same animal. Large stock (camels, cattle, and reindeer) may thus be the object of rather complex property regimes. The only other of objects to which similar systems of multiple rights by different sorts of holders or owners refer in preindustrial societies are land and ships. The topic 'Multiple and collective rights in animals' has been treated more fully elsewhere (Khazanov and Schlee, 2012).

It is unlikely that it will ever be known when this transition to nomadism and the forms of social organization associated with it first occurred. But what we do know is that nomadism could and did appear at different times and places, evidently unconnected with each other. It seems to have been the interplay between nature and society, together with the repeated long-term alternation of arid and less arid conditions and short dry periods since the early Holocene, that forced peoples to adopt a peripatetic way of life in order to survive. However, it is likely that social and political changes were no less important, such as the rise and fall of empires, population pressures, and famines and or tribal conflicts (Planhol, 1975). They triggered large- and small-scale migrations, caused nomadic hordes to overrun settled land, and forced settlers to move out into the steppes in times of emergency.

Historical Interpretation and Evidence

The first written reports about nomadic groups date to the fourth and third millennia. These and later scripts were almost all written by sedentary people and reflect their opinions and prejudices. Cuneiform texts, for example, always describe nomadic territory in negative terms, as the home of spirits and demons, the same word being used to denote both the desert and the underworld. In Asia the nomad is portrayed 'as a beast of hell' and 'incapable of civilized life' according to the Chinese Confucians (Khazanov, 1981: p. 142), or stereotyped as a 'barbarian,' a wolf or tiger. Little has changed since. However, now it is mainly Europeans who describe the historical role of nomads as destructive or nomadic existence as ahistorical.

The history of nomadism is voluminous, one-sided as regards its sources, doubtless complex, and difficult to unravel, and its inner contradictions have never been critically analyzed (Braudel, 1990). Its contradictory nature consists in the dichotomy between the ephemeral duration and recurrence of nomadic patterns. This will be kept in mind in the following discussion.

Historical descriptions are based on the generally accepted experience that the nomad raided settled land suddenly, like a cavalry attack, and after a while either disappeared or was assimilated. By contrast, the agriculturalist, natural conditions permitting, occupied nomadic country for permanent or at least long-term settlement and remained in logistic contact with the hinterland. Admittedly, nomads had a "considerable impact on the social and political development of many sedentary societies with more diversified economies and advanced technologies" (Khazanov, 1981: p. 142). His process gained importance as the mobility range of nomadic groups was enhanced by more powerful pack animals and the availability of lighter metal equipment. In addition, population pressure on the nomads increased, large (presumably wealthy) sedentary empires tempted nomads to trade or raid, and the scenario of conflict developed between central rule and peripheral tribes (Jettmar, 1966; Planhol, 1975). Yet these waves of nomadic raids on cultivated land are only part of the story, although sources mention them first of all. Much rarer are reports of sedentarists adopting a nomadic way of life. In this context Jettmar (1966: p. 5) even speaks of a 'victory of nomadism.' Even in Inner Asia, a region of almost classical 'antagonism' between nomads (Mongolia) and sedentarists (China), the situation is more one of overlapping and reciprocity (Lattimore, 1940: p. 328). Hence authors have spoken of an oscillation of nomadism's economic space or of a recurrent emergence of nomadism.

Indeed, empirical experience suggests that sedentism in conjunction with (oasis) agriculture offered safety and food in times of peace, but it was the mobile mode of life with herd-keeping in deserts and steppes that offered a means of survival in uncertain times. Considering the natural and socio-political settings and the short- and long-term changes occurring in the respective regions of the Old World dry belt, it may be assumed that a switch from sedentism to mobility or from agriculture to mobile livestock-keeping, and hence from agriculture to nomadism, occurred just as frequently as vice versa.

The historical view distinguishes two phases (Bobek, 1959):

1. The first phase differs only slightly from events in prehistoric times as described by archeologists. It is characterized by the keeping of small stock (sheep and goats), close links between agriculture and pastoralism, and probably also short, sometimes climate-induced switches from one to the other, leading to the recurrent separation of herders and cultivators within the same family or the same tribe. There is fairly reliable evidence of this from the Middle East. Similar events probably occurred in the Eurasian steppes although the actual 'nomadization' of the steppe agriculturalists presumably dates to the Scythian period (Iron Age, beginning around 800 BC) (Wissmann, 1961: p. 30).
2. This first phase ends during the second millennium BC. It was followed by first, domestication of the horse in the Eurasian steppes (in Turan about 2800 BC), by orienting entire peoples toward nomadism, by domestication of the camel in the Middle East (end of the second millennium), and resulting in greater mobility and the emergence of mounted warriors and hordes or 'empires' (e.g., the Cimmerians, Thracians, and Scythians, and later the Huns, Tatars, and (other) Turko-Mongolians). Second, the emergence of 'wealthy' centers of high culture, based on irrigation. They form the backdrop of the conflicts between nomadic and sedentary peoples that shape the subsequent history of this region. The clay tablets in the archives at Mari allowed scholars to infer these conflicts from the appearance of nomadic groups such as the Yaminites and Hanaeans (nineteenth century), the Sutaeans (eighteenth century), the Arameans (fourteenth century), or the Chaldeans (ninth century), in the irrigated parts of the drylands of Western Asia Mesopotamia. Lasting reminders of this are the fortified towns and villages of Iran, the famous 'walls' built as a defense against the nomadic threat in China (Great Wall, third century BC against the Huns, AD fifteenth century

against the Mongols), in phase with Turan or NE Mongolia at the mouth of the Gan, in the Middle East (Limes Syriae), North Africa (Limes Numidiae), Russia (Limes Daciae), or in central Europe (Limes Germaniae Raetiae).

In the seventeenth century BC, the Hyksos descended on Egypt, introducing the horse, and in the eleventh century the Midians overran Palestine and Canaan. Since the ninth century, the 'Arabs' (nomadizing Arab = Bedouins), mounted on camels, repeatedly entered fertile Mesopotamia and in the third century they came to Egypt. In the sixth century, the Persian Empire in the northern Middle East expanded northward (beyond the Jaxartes River) into the lands of the sedentarized steppe peoples. Like the later westward expansion of Hun rule (starting in the first century), this interrupted the sedentarization process that prevailed in the Eurasian steppes at the end of the Scythian period. Subsequently, the entire region experienced the upheavals of the migration of peoples. From then on, changes occurred more rapidly, and the nomad Bedouin continued to play a major role.

In the first two centuries AD Roman rule in the Middle East triggered a wave of nomadization; in this case the shift from a non-Bedouin to a Bedouin way of life. This process continued, as the decline of Marib Oasis (Yemen, sixth century) and the contemporaneous revival of Bedouin life show. Over the following centuries it received a new stimulus from the spread of Islam, particularly from the Islamic-Turkish nomadic armies of the Seljuqs (1048) in the mountains of Anatolia and Iran, Iranian tribes as far as Baluchistan, Sind, and Punjab, and the Beni Hilal (1051) and Beni Sulaym in North Africa. In the Khanate of the nomadic Khazars, north of the Caucasus, the sedentarization process that had begun in the eighth century was succeeded by a wave of renomadization as a result of conquests by the Pechenegs and Cumans in the eleventh and twelfth centuries. The same happened to the Polovtsy (Kipchaks) after the Mongol invasions in the thirteenth century. In the Iranian mountains, too, they triggered a real nomadization wave described as 'medieval Bedouinization' (Planhol, 1975: p. 238). The Mongol impact stretched as far as Western Asia, where for example, Baghdad was destroyed in 1258, the irrigation systems of Mesopotamia were devastated, arable land reverted to pasture, and the desert was the only safe refuge for any length of time (Wissmann, 1961: p. 47). Similar consequences followed the attacks by Timur in the fourteenth century or by the nomadic Turkmen tribes in Anatolia in the sixteenth century, where 'nomadism' disappeared in the West and began to flourish in the East (Braudel, 1990: p. 139). Until the nineteenth century Egypt experienced repeated phases of movement and settlement, and the shift between sedentarization and nomadization also continued into the nineteenth century in the Eurasian steppes, where it was quite common among the Turkmen of southern Kazakhstan for "a nomad who had lost his livestock to settle down on land, or for a rich farmer to buy a herd and to shift again to nomadism" (Khazanov, 1978: p. 121). This process has been impressively described for the Arabian Peninsula, for western Afghanistan, or for the western Sahara at the start of the twentieth century, where oasis farmers aspired to the 'higher' class of nomadic herders. With some exceptions, however, the trend since the late nineteenth century has been generally toward modern agriculture and sedentary life (Scholz, 1995).

In many instances, this trend is not caused by economic or ecological factors but by politics. Modern states, contrary to some of their medieval precursors like the empires of Chinggis Khan or Timur Leng, are not based on nomadism and do not have (former) nomads as their power elites. They are based on agriculture and on urban forms of life and share the age-old stereotypes of nomads some of which have been depicted above (see also Leder and Streck, 2005). Their main interest seems to be to control nomads, who are perceived as a security risk, and later, when the nomads no longer represent a threat, to appropriate pastoralist resources for other types of use. Also some of the historical theories described in this article affect these sedentist ideologies and policies. The conviction that nomadism represents an early development stage of humankind and stands for primitiveness and barbarism, although it has long been refuted by historians and anthropologists, lingers on in the minds of politicians and development planners. Rather than perceiving nomadism and other forms of mobile pastoralism (like long distance transhumance from a sedentary basis) as a comparatively recent development (old as they may be, they follow mixed agriculture rather than preceding it) and a sophisticated form of specialization which requires portable implements, great expertise in handling animals, flexible forms of organization (shrinking and widening forms of lineage alliance for joint defense, dispersal in times of scattered resource distribution), members of the state class tend to represent nomadism as archaic, traditionalist, unchanging, maladaptive. Anthropological studies of nomads as well as their resilience and reemergence throughout history, which has been the topic of the present article, suggest the opposite. Nomads are strategizing and opportunist users of resources in difficult environments. Recent studies (Babiker, 2013; Behnke and Kerven, 2013; Galaty, 2013) show that mobile pastoralism often compares favorably with alternative ('modern') forms of land use like irrigated agriculture, large-scale rain-fed mechanized agriculture or tourism (licensed hunting, photo-safaris) in terms of productivity per hectare, especially when one takes into account that the removal of key resources (access to riverbanks for watering, dry season pastures) often makes the use of much wider areas much less efficient or no longer feasible, so that in the interest of the overall economy pastoralist forms of production should be protected and, wherever possible, intelligently combined with other forms of land use rather than being abolished as 'archaic' or 'primitive.' Sadly, power elites often do not act in the interest of the overall economy but in their own interests and the interests of those with whom they identify with or with whom they are in alliance. This applies to the nomadic elites of Eurasian empires in antiquity and in medieval times, who were often perceived by their sedentary subjects as purely predatory, as well as to modern power elites who identify with the sedentary populations and have their own stakes in sedentary forms of economy.

Conclusions

This fragmentary outline of the 'origin' and history of nomadism may be summed up in the following four conclusions:

1. To a large extent, authors agree that stockkeeping (domestication of herd animals by farmers) forms the basis prerequisite for nomadism.
2. Historical sources 'report' in great detail, nomadic invasions of settled lands and nomadic switches to sedentism. This is probably why these events have enjoyed particular academic attention, wide and biased regard, and a historical interpretation of considerable influence on society.
3. Primarily ethnoarcheological studies have indicated that the shift from sedentary life to mobility, from agriculture to nomadism, occurred no less frequently. However, historical sources say little about such switches, and so there is some justification for assuming that insufficient importance has been attached to the historical and social relevance of these processes.
4. Nomadism is a human survival strategy in response to specific ecological and sociopolitical conditions within the Old World dry belt. As such, it represents a regionally immanent, logical 'idea' that could emerge recurrently, both locally and throughout the whole region.

The switch between sedentism and mobility, cultivator and nomad, occurred in both short- and long-term sequences, affecting both small groups and entire peoples. Its impact was local, but could also involve entire areas and regions.

In this context, it is no longer really necessary to consider Pohlhausen's (1954) hypothesis of a monocentric origin of pastoral nomadism and a multicentric origin of its various regional forms. The above archeological and historical findings, and the conclusions derived from them, suggest that it is more logical to assume that the 'idea' of a nomadic mode of culture was latent everywhere in the natural ecological and sociopolitical setting structures of the Old World dry belt. It only required a trigger, a change in prevailing conditions, for it to revive as a survival strategy or, indeed, to disappear.

See also: Middle East, North Africa: Sociocultural Overview; Migrations, Colonizations, and Diasporas in Archaeology; Pastoralism in Anthropology.

Bibliography

Babiker, M., 2013. Mobile pastoralism and land grabbing in Sudan: impacts and responses. In: Catley, A., Lind, J., Scoones, I. (Eds.), Pastoralism and Development in Africa: Dynamic Change at the Margins. Routledge, London and New York, pp. 177–185.

Behnke, R., Kerven, C., 2013. Counting the costs: replacing pastoralism with irrigated agriculture in the Awash Valley. In: Catley, A., Lind, J., Scoones, I. (Eds.), Pastoralism and Development in Africa: Dynamic Change at the Margins. Routledge, London and New York, pp. 57–70.

Bobek, H., 1959. Die Hauptstufen der Gesellschafts- und Wirtschaftsentfaltung in geographischer Sicht. Die Erde 90 (3), 259–298.

Braudel, F., 1990. Das Mittelmeer und die mediterrane Welt in der Epoche Philips II, vol. 1. Suhrkamp Verlag, Frankfurt am Main.

Dittmer, K., 1965. Zur Entstehung des Rinderhirtennomadismus. Paideuma. Mitteilungen zur Kulturkunde 11, 8–23.

Francfort, H.-P. (Ed.), 1990. Nomades et sédentaires en Asie centrale. Centre National de la Recherche Scientifique, Paris.

Galaty, J., 2013. Land grabbing in the Eastern African rangelands. In: Catley, A., Lind, J., Scoones, I. (Eds.), Pastoralism and Development in Africa: Dynamic Change at the Margins. Routledge, London and New York, pp. 143–153.

Gilbert, A.S., 1983. On the origins of specialized nomadic pastoralism in western Iran. World Archaeology 15, 105–119.

Hahn, E., 1891. Waren die Menschen der Urzeit zwischen der Jgerstufe und der Stufe des Ackerbaus Nomaden? Das Ausland. Wochenschrift für Erd- und Völkerkunde 64 (25), 481–487.

Jettmar, K., 1966. Die Entstehung der Reiternomaden. Saeculum 17 (1–2), 1–11.

Khazanov, A.M., 1978. Characteristic features of nomadic communities in the Eurasian Steppes. In: Weissleder, W. (Ed.), The Nomadic Alternative. Mouton, The Hague, pp. 119–126.

Khazanov, A.M., 1981. Myths and paradoxes of nomadism. Archives Europennes de Sociologie 22, 121–153.

Khazanov, A.M., Schlee, G. (Eds.), 2012. Who Owns the Stock? Collective and Multiple Property Rights in Animals. Berghahn Books, New York, Oxford.

Lattimore, O., 1940. Inner Asian Frontiers of China. Oxford University Press, New York.

Leder, S., Streck, B. (Eds.), 2005. Shifts and Drifts in Nomad-Sedentary Relations. Reichert, Wiesbaden.

Lees, S., Bates, D.G., 1974. The origins of specialized nomadic populations: a systematic model. American Antiquity 39, 187–193.

List, F., 1961. Das natürliche System der politischen Ökonomie Berlin Ökonomische Studientexte, 2. Akademie-Verlag. Lizenznummer 202, 100/27/61.

Nissen, H.J., 1980. The mobility between settled and non-settled in early Babylonia: theory and evidence. Colloques International du CNRS 580, 285–290.

Planhol, X., 1975. Kulturgeographische Grundlagen der islamischen Geschichte. Artemis, München.

Pohlhausen, H., 1954. Das Wanderhirtentum und seine Vorstufen. Eine ethnographisch-geographische Studie zur Entwicklung der Eingeborenenwirtschaft. Braunschweig.

Scholz, F., 1995. Nomadismus - Theorie und Wandel einer sozio-ökologischen Kulturweise, Erdkundliches Wissen 118. Steiner, Stuttgart.

Schlee, G., 2005. Forms of pastoralism. In: Leder, S., Streck, B. (Eds.), Shifts and Drifts in Nomad–Sedentary Relations. Reichert, Wiesbaden, pp. 17–53.

Schlee, G., Shongolo, A.A., 2012. Pastoralism and Politics in Northern Kenya and Southern Ethiopia. James Currey, Woodbridge.

Scholz, F., 1997. Gedanken zu einer These über die Entstehung des Nomadismus als sozioökologische Kulturweise. Petermanns Geographische Mitteilungen 141 (3), 187–196.

Vajda, L., 1968. Untersuchungen zur Geschichte der Hirtenkulturen (Study about the history of the pastoral culture, vol. 31). Harrassowitz, Wiesbaden.

Wendorf, F., Close, A.E., Schild, R., 1987. Early domestic cattle in the eastern Sahara. Paleoecology of Africa and the Surrounding Islands 18, 441–447.

Wendorf, F., Schild, R., 1998. Nabtsa Playa and its role in northeastern African prehistory. Journal of Anthropological Archaeology 17, 97–123.

Wissmann, H., 1961. Bauer, Nomade und Stadt im islamischen Orient. In: Paret, R. (Ed.), Die Welt des Islam und die Gegenwart. Kohlhammer, Stuttgart, pp. 22–63.

Non Declarative (Procedural) Memory

Leonora Wilkinson, Cognitive Motor Neuroscience Group, University College London, London, UK; Institute of Neurology, National Hospital for Neurology & Neurosurgery, University College London, London, UK; and Behavioral Neurology Unit, National Institute of Neurological Disorders and Stroke, Bethesda, MD, USA
Marjan Jahanshahi, Cognitive Motor Neuroscience Group, University College London, London, UK; and Institute of Neurology, National Hospital for Neurology & Neurosurgery, University College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abbreviations

CS Conditioned stimulus
DLPFC Dorsolateral prefrontal cortex
M1 Motor cortex
SMA Supplementary motor area
UCS Unconditioned stimulus
WPT Weather prediction task

Abstract

Declarative and procedural learning have been considered distinct learning systems, recruiting different brain networks. Procedural learning can be further subdivided into motor and perceptual/cognitive. A variety of tasks have been developed and used to assess specific types of procedural learning. Procedural learning involves different stages including cognitive, associative, and autonomous phases and is subject to off-line processes of consolidation. Dissociations between declarative and procedural memory deficits in patients with amnesic syndromes and neurodegenerative disorders including Parkinson's and Huntington's diseases together with evidence from brain imaging and animal studies indicated that the medial temporal lobe and basal ganglia mediate declarative and procedural learning, respectively. More recent evidence supports the notion of declarative and procedural learning as interacting rather than independent systems, depending on task demands.

Introduction

A wide spectrum of behavior is acquired through learning. The declarative memory system plays a role in the acquisition of knowledge in an intentional way including when learning facts such as the height of the Eiffel tower or a list of words. The declarative memory system is considered to operate consciously or explicitly. In contrast, the nondeclarative or procedural memory system is involved in learning motor skills such as riding a bicycle, playing tennis, or swimming, which are acquired through practice. Certain types of procedural learning are considered by some to be implicit or unconscious in nature (Butters et al., 1985; Squire, 2004). Procedural memory is the most primitive form of memory and refers to skills that are learned through acquisition of procedures, that is, associations between stimuli and responses. For example, when learning to swim, one learns the procedures, or some would say, motor programs for swimming. Once these procedures have been learned, then swimming can be performed automatically. Many of our essential motor skills in early life such as walking are acquired through procedural learning. Procedural learning is sometimes referred to as habit learning. A characteristic of procedural skill is that it can be demonstrated through action and performance, but explaining how the skill was learned is difficult, so knowledge cannot be verbally declared. For this reason, some forms of procedural learning are considered to be implicit and to occur without awareness.

As detailed below, Milner's (1962) and Corkin's (1968) seminal observations on the patient Henry Gustav Molaison, who despite dense amnesia showed intact performance on mirror tracing and pursuit manual tracking, were instrumental in fractionating the memory system. Subsequently, Cohen and Squire (1980) made a distinction between 'data-based' declarative memory and 'rule-based' procedural memory, while Graf and Schacter (1985) distinguished 'explicit' and 'implicit' memory. This led to the general view that the cortico-limbic-diencephalic structures were the substrates of declarative learning, while the striatum and its cortical projections supported procedural learning (Squire, 2004). However, in the intervening years, empirical evidence has failed to support a single procedural system and it has become clear that different types of procedural memory have distinct anatomical substrates, not solely mediated by the striatum (for review see Soliveri et al., 1992; Squire, 1987). Furthermore, recent evidence (e.g., Poldrack et al., 2001; Schendan et al., 2003) suggests that the declarative and procedural systems play an interactive rather than independent role in learning and development of skill.

Types of Procedural Memory and Tasks Used for Assessment

Studying lists of words followed by immediate or delayed word recall is the most common method for assessing short- and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.51052-6

long-term declarative memory in the laboratory. Skill learning is a relatively less intuitive phenomenon to measure. Nevertheless, over the years, numerous tasks have been developed to specifically assess procedural learning. Some of these procedural learning paradigms involve tasks that recruit the motor system directly, while others involve nonmotor tasks that have no direct motor component per se, but recruit the procedural system due to the nature of the complex information being learned during task performance and the fact that the knowledge becomes automatized and habitual.

Motor Procedural Skill Learning Tasks

1. *Mirror tracing*: This visual-motor task involves learning the new motor skill of drawing a figure (e.g., a star shape) by looking at its reflection in the mirror (Milner, 1962). The presentation of familiar but distorted visual information, which must be 'translated' before an accurate response can be made is the reason why this task is considered to involve procedural learning; once a participant has figured out how to draw the image in the mirror the first time, i.e., acquired the appropriate procedures, they will have little difficulty the second time they complete the task.
2. *Pursuit rotor*: The participants are instructed to keep a metal stylus in contact with a rotating disc. The amount of time on target typically increases with practice, and the acquired skill transfers to subsequent test occasions (Corkin, 1968).
3. *Tracking tasks*: The participants' task is to track the movements of a target (e.g., small rectangle) presented on a screen by moving the position of a pointer also presented on the screen. The difficulty of tracking can be manipulated by presenting various types of target. The percent time on target is measured which improves with practice.
4. *Serial reaction time task* (SRT task): This is a paradigm that has been employed extensively to study procedural/implicit motor sequence learning. The SRT task was developed by Nissen and Bullemer (1987). The SRT task involves presentation of stimuli in one of several target locations (e.g., in four horizontally presented boxes), to which participants respond as quickly as possible by pressing corresponding response buttons. Typically, participants perform several hundred trials of what is essentially a choice reaction time (RT) task; although unbeknownst to them, the targets actually appear in a predetermined, repeating sequence of locations (e.g., 3-4-2-3-1-2-1-4-3-2-4-1), the presentation of which is interspersed with random or pseudo-random trials. Following SRT training blocks, sequence learning is demonstrated by faster reaction times (RTs) to sequence compared to random blocks of trials. Learning on the SRT task is considered by some to be implicit and operating outside of awareness because it occurs incidentally and participants perform poorly on subsequent tests of awareness such as verbal report of their acquired sequence knowledge.

Perceptual or Cognitive Procedural Skill Learning Tasks

1. *Mirror reading*: This task assesses the ability to learn to read mirror-reversed material. Without the participants' knowledge, some of the stimuli are repeated. The rate of learning can therefore be assessed for both new material and repeated stimuli. As with mirror tracing, the task involves acquiring the procedures for translating familiar but distorted visual information before an accurate response can be made.
2. *Prism adaptation*: In this task, the participant wears prisms that produce a consistent, usually lateral, distortion of visual input. Initially, visually guided motor behavior is disturbed as the movement is based on the distorted visual information. With experience, the motor response adapts. That true perceptuo-motor adaptation has occurred (and not merely development of a conscious strategic change), becomes evident when the prisms are removed. When the prisms are removed, normal participants misreach in the opposite direction for a few trials before the visuo-motor system is 'retuned.' This 'negative after-effect' is the main criterion for determining that adaptation had taken place.
3. *Priming*: Priming refers to the facilitative effect of a pre-exposure to an item on subsequent processing of the same item (Schacter and Graf, 1986). Various tasks have been used to assess diverse types of priming, including perceptual priming, word stem completion, and semantic priming. In a typical perceptual priming task, for example, the stimuli employed are incomplete pictures (Gollin figures) or fragmented words. The participants try to identify the stimulus, starting with the most incomplete word/figure. If they fail, a more complete stimulus is presented. This process continues until identification occurs. Repeated testing on the same material in healthy participants leads to rapid learning, with participants being able to identify even the most degraded stimuli.
4. *Weather prediction task*: The weather prediction task (WPT) is a nonmotor probabilistic classification task involving incremental learning of a complex probabilistic rule (Knowlton et al., 1994). On each trial, participants are presented with a particular arrangement of one, two, or three of four possible tarot cards, each of which have different patterns on them (e.g., squares, diamonds, circles, or triangles). Participants are required to use the cards presented on each trial to predict a binary outcome; whether the weather will be rainy or fine. Each of the four cards is independently associated with the two possible outcomes with a fixed probability and overall, each outcome occurs equally often. For example, the squares, diamonds, circles, and triangles respectively predict the outcome 'fine' with a fixed probability of 0.2, 0.4, 0.6, and 0.8. Typically, participants perform around 200 trials of the WPT with corrective feedback on each trial. By learning the independent cue–outcome associations across trials, participants can improve their predictive accuracy on the WPT across training trials. Learning on the WPT task is considered by some to be implicit operating outside of awareness because participants perform poorly on subsequent tests of awareness.
5. *Classical conditioning*: Eyeblink conditioning is a form of classical conditioning that has been employed to study unconscious learning. This consists of pairing an auditory or visual stimulus (the conditioned stimulus – CS) with an

eyeblink-eliciting unconditioned stimulus (US) (e.g., a mild puff of air to the cornea). Participants initially produce a reflexive, unconditioned response (eyeblink) that follows US onset. After many CS–US pairings, an association is formed such that an eyeblink, or conditioned response, occurs and precedes US onset. The magnitude of learning is generally gauged by the percentage of all paired CS–US trials that result in a conditioned response.

Processes and Mechanisms of Procedural Learning

Stages of Learning

Most cognitive models of skill acquisition distinguish different phases of learning, involving distinct processes and modes of representation. Although the details differ, most of these models distinguish an initial phase requiring attention, conscious processing, working memory, and hypothesis testing with a later less demanding automatic phase. For example, Adams (1971) distinguished 'verbal-motor' and 'motor' stages, while Logan (1988) described 'algorithm-based' and 'memory-based' performance. Fitts (1954) hypothesized three main stages of skill learning. In the first 'cognitive phase' that requires attention, an individual develops an understanding of the demands of the task at hand, by breaking down the skill to be learned into its component parts and learning the basic procedures to be followed. The second 'associative phase' involves repeated practice of the task until successful patterns of responding emerge. Through reliance on feedback, in this associative phase, task components that lead to success are retained whereas components that are ineffective are dropped. The final 'autonomous phase' involves perfecting the acquired skill so that it can be performed quickly and automatically, without conscious attention.

Evidence from a study involving single-cell recordings from neurons in the striatum and motor cortex during motor skill learning in mice indicates there are different firing patterns in the motor system depending on whether skill learning is fast (learning that occurred during the first training session) relative to slow (learning that occurred across sessions). For example, during the first fast learning session, there was parallel activation of neurons in the striatum and motor cortex, whereas when the learning was slow the neuronal firing was differential in these structures rather than comparable (Costa et al., 2004). These findings could be considered as direct evidence for different neural signatures for the associative (fast) and associative/autonomous (slow) phases in Fitts' model.

Importance of Practice for the Development of Skill

A feature of procedural learning is that repetition and practice are important for the acquisition of skill, which is most evident in the case of motor skills. The importance of practice is reflected in Schmidt's (1982) definition of motor learning as "a set of processes associated with practice or experience leading to relatively permanent changes in the capability for responding." In this definition, "the permanent changes in the capability for responding" as indicative of learning distinguishes this from 'performance' reflected in transient changes in behavior. A recent primate study (Picard et al., 2013) examined changes in metabolic and single-neuronal activity in the motor cortex, with long-term training over 1–6 years of internally generated and visually guided sequences of reaching movements. A significant reduction of metabolic activity in the motor cortex was observed for the performance of internally generated compared to visually guided task, suggesting that extended practice results in reduction in the synaptic activity required to produce internally generated sequences of movement. It was concluded that practice leading to skilled performance is associated with more efficient generation of neuronal activity in the motor cortex.

Equally important for learning to occur are feedback and knowledge of results. Feedback can be intrinsic (produced by the response and could be visual, auditory, or proprioceptive) or external (provision of augmented information about performance by an external source) with the latter often referred to as knowledge of results. Less well known is the benefit that randomized (high contextual interference) versus grouped/constant (low contextual interference) training schedules can have on long-term skill acquired through practice. Termed the 'contextual interference effect,' this somewhat counter-intuitive benefit refers to the observation that higher levels of contextual interference, for example, when several tasks are being learned and practiced at the same time, result in poorer practice performance, but are associated with superior learning, retention, and transfer. This benefit has been demonstrated for a range of tasks, motor skills acquired explicitly and implicitly, and across age groups and in rehabilitation contexts (Song et al., 2012). Variability of practice is considered to accentuate task variations during learning such that, even though varied practice may produce poor performance during the acquisition phase, it is important for the development of the schemata, which are responsible for the assembly and improved retention and transfer of motor learning.

Consolidation of Procedural Knowledge and Sleep Benefit

While practice is essential for learning, further improvement of skill occurs after cessation of practice, which is referred to as consolidation. Such consolidation takes two forms. First, there is the 'off-line' improvement of skill, which occurs over time after training has ceased and following sleep. Notable improvements in perceptual or motor skill can occur following sleep, and such sleep-dependent benefit seems to be related to the percent or relative combinations of the different stages of sleep (Karni et al., 1995; Walker et al., 2002). Second, the skill and associated memory trace acquired through practice become stabilized and are less subject to decay or interference (Robertson, 2012). Experimental evidence suggests that consolidation of sequence knowledge and motor skill adaptation is more robust under conditions when initial skill learning is followed by a task that either engages 'e.g., word list learning' or inhibits (using inhibitory brain stimulation) the declarative system relative to when the declarative system is not disrupted or taken offline after learning (R.M. Brown and Robertson, 2007; Galea et al., 2010; Keisler and Shadmehr, 2010). These findings indicate off-line consolidation of procedural skill relies not only on the engagement of neuroplastic mechanisms,

but also on the disengagement of an interaction between declarative and procedural memory systems.

A study (Kantak et al., 2010) of the consolidation of knowledge following goal-directed skill learning of arm movements used inhibitory brain stimulation to interfere with the function of both the primary motor cortex (M1) and dorsolateral prefrontal cortex (DLPFC) following learning of the task with both randomized and constant practice schedules. Interestingly, disruption of the DLPFC-reduced consolidation of knowledge acquired following randomized but not constant practice, while interfering with M1-reduced consolidation of learning when the practice structure was constant but not when it was randomized. These findings demonstrate that the practice structure employed during learning has a direct impact on the networks recruited for subsequent consolidation. It appears that the goal-directed nature of the randomized schedule leads to the recruitment of more frontal cognitive networks, while constant practice leads to the engagement of motor associative regions with no requirement for the frontal lobes.

Procedural Memory and Retroactive and Proactive Interference

In contrast to interleaved practice of several tasks during training that can be associated with superior learning, performance of two procedural tasks in close temporal proximity can also result in various types of cross-task interference. Retroactive interference refers to the interference of a newly acquired task with performance of a previously learned task. There is evidence that similar to declarative memory, most forms of procedural learning are also susceptible to such retroactive interference effects, with these effects being reduced as the temporal interval between performance of the two tasks is increased (Robertson, 2012). For many years explicit memory was considered to be subject to 'proactive interference,' whereby the memory for most recently learned information can be compromised by previously learned material, whereas implicit memory was thought to be immune to this kind of proactive interference effect. However, more recently, evidence has emerged that proactive interference is not necessarily confined to explicit memory and occurrence of proactive interference effects has been documented for some procedural tasks such as word stem completion priming (Lustig and Hasher, 2001). In terms of long-term retention, well-learned motor skills, such as swimming, skiing or riding a bicycle, are maintained over time and are relatively resistant to forgetting over time.

Neural Substrates of Procedural Learning

The first evidence that procedural learning recruits different brain networks from those mediating declarative memory came from the results of some pioneering studies performed over several decades by Brenda Milner and Suzanne Corkin with an amnesic patient called Henry Molaison (HM) (Corkin, 2013). HM had a bicycle accident as a young child, and subsequently developed debilitating seizures. In 1953, at the age of 27, he underwent brain surgery for relief of his intractable epilepsy. The surgical procedure involved resection of the medial temporal lobe bilaterally. Following surgery HM was observed to have profound amnesia (i.e., impaired declarative memory) for events following his operation, as he could no longer encode newly learned declarative information into long-term memory. However, HM's memory for events before the surgery remained intact; furthermore, he was capable of learning new procedural skills. For example, in a study conducted by Milner (1962), HM acquired the new perceptual-motor skill of mirror tracing (drawing a figure, e.g., star shape by looking at its reflection in a mirror), despite having no conscious recollection of previous training episodes. In addition, HM was shown to have normal learning but little conscious recollection of the acquisition of other motor and cognitive and perceptual skill tasks such as pursuit rotor and bimanual tracking (Corkin, 1968).

Subsequently, it was shown that patients with Korsakoff syndrome with damage to diencephalic structures (mostly the mediodorsal nucleus of thalamus and the mamillary bodies) have severe anterograde amnesia, but also maintained a normal ability to learn cognitive skills, perceptual skills, and motor skills (Martone et al., 1984; Nissen et al., 1989; Saintcyr et al., 1988). Studies of amnesic patients with medial temporal lobe damage either due to lesions (Reber and Squire, 1994), Korsakoff syndrome (Nissen et al., 1989), or a combination of lesions and Korsakoff syndrome (Reber and Squire, 1998) showed normal learning on the SRT task, while verbal report of knowledge acquired during the training episode is impaired, indicating that explicit learning was dependent on the medial temporal lobe; while implicit motor sequence learning must involve regions of the brain other than the medial temporal lobe. It was further demonstrated that patients with dysfunction of the basal ganglia due to Parkinson's disease (PD) (Knowlton et al., 1996a; Saintcyr et al., 1988) or Huntington's disease (HD) (Knowlton et al., 1996a; Martone et al., 1984) showed the opposite pattern of results to amnesic patients as they had impaired procedural skill learning while being able to learn declaratively. This 'double dissociation' between declarative and procedural learning in amnesic patients and patients with PD and HD led to the general view that declarative and procedural learning are mediated by distinct and separable neural systems, and that while the cortico-limbic-diencephalic structures are the substrates for declarative learning, the striatum and its cortical projections support procedural learning nature (Squire, 2004). However, based on evidence from multiple sources, e.g., clinical patient studies and imaging, it has become clear that procedural learning is not mediated by a single unitary system and that the different forms of procedural learning are mediated by brain structures other than the striatum, for early review see Soliveri et al. (1992). In light of the heterogeneity of findings across groups with different neurodegenerative disorders across diverse tasks tapping procedural memory, Squire (1987) proposed a reformulation of procedural memory as "a collection of different abilities each dependent on its own specialised processing system … one therefore should not expect that a single lesion would affect all of procedural memory." Furthermore, in the intervening years, imaging has revealed that the hippocampus is engaged during procedural tasks such as motor sequence learning (e.g., Schendan et al., 2003), while under some circumstances, declarative tasks such as spatial navigation, considered to depend on the hippocampus, also recruit the striatum (e.g., Iaria et al., 2003).

Neural Substrates of Motor Skill Learning

Neuropsychological studies on patient groups with neurodegerative disorders and imaging have revealed that procedural learning does not depend on a single system but that different types of procedural learning are mediated by specific brain networks. Execution of the pursuit rotor task has been shown to be associated with a network of distributed regions including cortical, striatonigral, and cerebellar. However, learning on this task involves the primary motor cortex (M1), supplementary motor area (SMA), and thalamus (Grafton et al., 2001), suggesting engagement of the sensorimotor circuit between the putamen, premotor cortex, SMA, and M1. Not surprisingly, learning on a bimanual tracking task recruits similar regions to the pursuit rotor including the M1, posterior cingulate, putamen, and right anterior cerebellum (Puttemans et al., 2005). One study (Laforce and Doyon, 2001) showed that mirror tracing was not impaired in PD patients but was impaired in patients with cerebellar damage. These results indicate that in the mirror tracing task presentation of familiar but distorted visual information, which must be 'translated' before an accurate response can be made, necessitates cerebellar rather than striatal involvement. Furthermore, patients with PD (R.G. Brown et al., 2003; Deroost et al., 2006; Doyon et al., 1997; Jackson et al., 1995; Kelly et al., 2004; Muslimovic et al., 2007; Pascual-Leone et al., 1993; Stefanova et al., 2000; Werheid et al., 2003; Westwater et al., 1998; Wilkinson and Jahanshahi, 2007; Wilkinson et al., 2009) or HD (Knopman and Nissen, 1991) show impaired learning on the SRT, confirming that motor sequence learning is dependent on the basal ganglia and their cortical projections. The involvement of the fronto-striatal circuits, particularly the motor and associative circuits, has been confirmed by imaging studies, which have shown that motor sequence learning on the SRT is associated with activation of M1, SMA and premotor cortex, DLPFC, and the putamen and caudate (Grafton et al., 1995; Hazeltine et al., 1997; Poldrack et al., 2005; Schendan et al., 2003; Seidler et al., 2005).

With regard to motor skill learning, different fronto-striatal circuits have been implicated in different stages of learning (Graybiel, 2005; Hikosaka et al., 2002; Yin et al., 2009). Most recently, evidence from single-cell recording studies in mice indicates the dorsomendial striatum that is equivalent to the caudate in primates is involved in fast learning while the dorsolateral striatum – equivalent to the putamen in primates – is involved in slow automatized learning. This evidence suggests that in humans the associative circuit between the caudate and DPLFC and the sensorimotor circuit between the putamen and motor cortex are involved in fast and slow learning, respectively (Graybiel, 2005; Hikosaka et al., 2002; Yin et al., 2009).

Neural Substrates of Perceptual/Cognitive Procedural Skill Learning

Patients with PD (Knowlton et al., 1996a; Shohamy et al., 2004; Wilkinson et al., 2008; Witt et al., 2002) or HD (Knowlton et al., 1996a) are impaired at probabilistic classification learning on the WPT. However, imaging studies in healthy controls and PD patients have demonstrated involvement of both declarative and procedural networks in WPT learning, even though the time course of their recruitment during learning is different, and their activity is competitive (Aron et al., 2004; Foerde et al., 2006; Moody et al., 2004; Poldrack et al., 2001; Poldrack et al., 1999; Seger and Cincotta, 2006).

For other cognitive and perceptual procedural tasks, evidence indicates that they involve different networks. For example, a study of amnesic patients with medial temporal lobe damage showed they were impaired on an implicit conceptual memory priming task (Wang et al., 2010). Moreover, imaging evidence from healthy volunteers showed that implicit associative learning recruited the hippocampus rather than the basal ganglia (Degonda et al., 2005). So the role of the medial temporal lobe does not seem to be simply confined to declarative memory. Similarly, imaging evidence indicates that prism adaptation and mirror reading both recruit the superior temporal cortex and cerebellum (Luaute et al., 2009).

Other evidence indicates that the procedural-implicit/striatal versus declarative-explicit/medial temporal lobe distinction may be too simplistic. First, the idea these two memory systems are separable specifically on the basis of whether or not the learning is conscious in nature has been challenged (see Section The Question of Awareness below), as has the idea that nondeclarative procedural learning is the only form of skill learning dependent on the basal ganglia. For example, one study showed normal implicit SRT learning by patients with HD, while at the same time they were impaired at a modified SRT task, in which a sequence of target locations was explicitly learned by trial-and-error method (R.G. Brown et al., 2001). Another study found that both implicit and explicit sequence learning were impaired in PD (Wilkinson et al., 2009). Furthermore, similar to implicit sequence learning, the caudate, putamen, DLPFC, premotor cortex, and SMA are all activated during explicit sequence learning in healthy participants and are under activated in PD patients relative to controls (Carbon et al., 2003; Grafton et al., 1995; Jenkins et al., 1994; Jueptner et al., 1997; Lehericy et al., 2005; Nakamura et al., 2001; Schendan et al., 2003; Willingham et al., 2002). Second, it has been suggested that for some forms of procedural learning, the reason the basal ganglia are recruited is due to the fact that the task involves the processing of corrective feedback rather than because it is an implicit task per se (Holl et al., 2012; Shohamy et al., 2004; Wilkinson et al., 2008). Finally, and as mentioned above there is evidence that, rather than operating independently, the declarative and procedural systems interact in a dynamic manner (Foerde et al., 2006; Hartley and Burgess, 2005; Iaria et al., 2003; Myers et al., 2003; Sadeh et al., 2011).

The Question of Awareness

A core feature of some types of procedural learning is the concept that such learning is implicit and occurs without conscious awareness. For example, a great deal of research has been conducted on the question of whether knowledge acquired on the SRT is genuinely unconscious. Posttraining tests of awareness include asking participants to verbally report the sequence they learned. However, verbal report is not

sensitive to all of the conscious knowledge an individual may possess (Shanks and St John, 1994). Other more objective tests of awareness have been developed for the SRT such as free-generation and recognition tests. Some have proposed that the degree to which participants can control the generation of their sequence knowledge, either intentionally expressing or withholding it, is more relevant to understanding whether learning is genuinely unconscious. Destrebecqz and Cleeremans (2001) developed posttraining free generation tasks which participants performed under both 'inclusion' (express as much of the trained sequence as possible) and 'exclusion' (withhold production of the trained sequence) conditions. While Destrebecqz and Cleeremans' free generation under inclusion versus exclusion instructions data suggested that under some circumstances (e.g., speeded performance) learning on the SRT occurred with limited awareness, others (Wilkinson and Shanks, 2004) have not replicated their findings and questioned if learning on this task is ever truly implicit. Hence, the question of whether sequence learning can be implicit and whether genuinely implicit knowledge even exists is still debated.

Conclusions

The value of a procedural memory system is that once skill is acquired, performance can often proceed automatically without the need for allocation of significant attention. This liberates attention to be allocated to other urgent processes. For example, once the skill of driving a car has been acquired, a person can simultaneously drive and hold a conversation.

More recent evidence has highlighted the need for revision of the classic view of two independent memory systems: a declarative system mediated by the medial and diencephalic structures, and a procedural system based on the basal ganglia. First, instead of a single and unified procedural system based on the striatum, evidence across a range of different tasks suggests parallel and independent forms of procedural learning, only some of which may engage the striatum. Second, the view of two independent memory systems has been replaced with operation of two interactive memory systems that may become engaged at different stages of declarative and procedural learning.

See also: Amnesia: General; Amnesia: Psychogenic; Associative Learning in the Cerebellum; Declarative Memory, Neural Basis of; Hippocampal Amnesia; Hippocampus and Related Structures; Implicit Memory; Learning and Memory, Neural Basis of; Proactive Interference; Sleep and Memory; Transient Epileptic Amnesia; Transient Global Amnesia.

Bibliography

Adams, J.A., 1971. Closed-loop theory of motor learning. Journal of Motor Behavior 3 (2), 111–150.

Aron, A.R., Robbins, T.W., Poldrack, R.A., 2004. Inhibition and the right inferior frontal cortex (Review). Trends in Cognitive Sciences 8 (4), 170–177.

Brown, R.G., Jahanshahi, M., Limousin-Dowsey, P., Thomas, D., Quinn, N.P., Rothwell, J.C., 2003. Pallidotomy and incidental sequence learning in Parkinson's disease. Neuroreport 14 (1), 21–24.

Brown, R.G., Redondo-Verge, L., Chacon, J.R., Lucas, M.L., Channon, S., 2001. Dissociation between intentional and incidental sequence learning in Huntington's disease. Brain 124, 2188–2202.

Brown, R.M., Robertson, E.M., 2007. Inducing motor skill improvements with a declarative task. Nature Neuroscience 10 (2), 148–149.

Butters, N., Wolfe, J., Martone, M., Granholm, E., Cermak, L.S., 1985. Memory disorders associated with Huntingtons-disease – verbal recall, verbal recognition and procedural memory. Neuropsychologia 23 (6), 729–743.

Carbon, M., Ghilardi, M.F., Feigin, A., Fukuda, M., Silvestri, G., Mentis, M.J., et al., 2003. Learning networks in health and Parkinson's disease: reproducibility and treatment effects. Human Brain Mapping 19 (3), 197–211.

Cohen, N.J., Squire, L.R., 1980. Preserved learning and retention of pattern-analyzing skill in amnesia – dissociation of knowing how and knowing that. Science 210 (4466), 207–210.

Corkin, S., 1968. Acquisition of motor skill after bilateral medial temporal-lobe excision. Neuropsychologia 6 (3), 255.

Corkin, S. (Ed.), 2013. Permanent Present Tense: The Man with No Memory, and What He Taught the World. Penguin Books.

Costa, R.M., Cohen, D., Nicolelis, M.A.L., 2004. Differential corticostriatal plasticity during fast and slow motor skill learning in mice. Current Biology 14 (13), 1124–1134.

Degonda, N., Mondadori, C.R.A., Bosshardt, S., Schmidt, C.F., Boesiger, P., Nitsch, R.M., et al., 2005. Implicit associative learning engages the hippocampus and interacts with explicit associative learning. Neuron 46 (3), 505–520.

Deroost, N., Kerckhofs, E., Coene, M., Wijnants, G., Soetens, E., 2006. Learning sequence movements in a homogenous sample of patients with Parkinson's disease. Neuropsychologia 44 (10), 1653–1662.

Destrebecqz, A., Cleeremans, A., 2001. Can sequence learning be implicit? New evidence with the process dissociation procedure. Psychonomic Bulletin & Review 8, 343–350.

Doyon, J., Gaudreau, D., Laforce, R., Castonguay, M., Bedard, P.J., Bedard, F., et al., 1997. Role of the striatum, cerebellum, and frontal lobes in the learning of a visuomotor sequence. Brain and Cognition 34 (2), 218–245.

Fitts, P.M., 1954. The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental Psychology 47 (6), 381–391.

Foerde, K., Knowlton, B.J., Poldrack, R.A., 2006. Modulation of competing memory systems by distraction. Proceedings of the National Academy of Sciences of the United States of America 103 (31), 11778–11783.

Galea, J.M., Albert, N.B., Ditye, T., Miall, R.C., 2010. Disruption of the dorsolateral prefrontal cortex facilitates the consolidation of procedural skills. Journal of Cognitive Neuroscience 22 (6), 1158–1164.

Graf, P., Schacter, D.L., 1985. Implicit and explicit memory for new associations in normal and amnesic subjects. Journal of Experimental Psychology – Learning Memory and Cognition 11 (3), 501–518.

Grafton, S.T., Hazeltine, E., Ivry, R., 1995. Functional mapping of sequence learning in normal humans. Journal of Cognitive Neuroscience 7 (4), 497–510.

Grafton, S.T., Salidis, J., Willingham, D.B., 2001. Motor learning of compatible and incompatible visuomotor maps. Journal of Cognitive Neuroscience 13 (2), 217–231.

Graybiel, A.M., 2005. The basal ganglia: learning new tricks and loving it. Current Opinion in Neurobiology 15 (6), 638–644.

Hartley, T., Burgess, N., 2005. Complementary memory systems: competition, cooperation and compensation. Trends in Neurosciences 28 (4), 169–170.

Hazeltine, E., Grafton, S.T., Ivry, R., 1997. Attention and stimulus characteristics determine the locus of motor-sequence encoding – a PET study. Brain 120, 123–140.

Hikosaka, O., Nakamura, K., Sakai, K., Nakahara, H., 2002. Central mechanisms of motor skill learning. Current Opinion in Neurobiology 12 (2), 217–222.

Holl, A.K., Wilkinson, L., Tabrizi, S.J., Painold, A., Jahanshahi, M., 2012. Probabilistic classification learning with corrective feedback is selectively impaired in early Huntington's disease – evidence for the role of the striatum in learning with feedback. Neuropsychologia 50 (9), 2176–2186.

Iaria, G., Petrides, M., Dagher, A., Pike, B., Bohbot, V.D., 2003. Cognitive strategies dependent on the hippocampus and caudate nucleus in human navigation: variability and change with practice. Journal of Neuroscience 23 (13), 5945–5952.

Jackson, G.M., Jackson, S.R., Harrison, J., Henderson, L., Kennard, C., 1995. Serial reaction-time learning and Parkinsons-disease – evidence for a procedural learning deficit. Neuropsychologia 33 (5), 577–593.

Jenkins, I.H., Brooks, D.J., Nixon, P.D., Frackowiak, R.S.J., Passingham, R.E., 1994. Motor sequence learning – a study with positron emission tomography. Journal of Neuroscience 14 (6), 3775–3790.

Jueptner, M., Frith, C.D., Brooks, D.J., Frackowiak, R.S.J., Passingham, R.E., 1997. Anatomy of motor learning. 2. Subcortical structures and learning by trial and error. Journal of Neurophysiology 77 (3), 1325–1337.

Kantak, S.S., Sullivan, K.J., Fisher, B.E., Knowlton, B.J., Winstein, C.J., 2010. Neural substrates of motor memory consolidation depend on practice structure. Nature Neuroscience 13 (8), 923–925.

Karni, A., Meyer, G., Jezzard, P., Adams, M.M., Turner, R., Ungerleider, L.G., 1995. Functional MRI evidence for adult motor cortex plasticity during motor skill learning. Nature 377 (6545), 155–158.

Keisler, A., Shadmehr, R., 2010. A shared resource between declarative memory and motor memory. Journal of Neuroscience 30 (44), 14817–14823.

Kelly, S.W., Jahanshahi, M., Dirnberger, G., 2004. Learning of ambiguous versus hybrid sequences by patients with Parkinson's disease. Neuropsychologia 42 (10), 1350–1357.

Knopman, D., Nissen, M.J., 1991. Procedural learning is impaired in Huntingtons-disease – evidence from the serial reaction-time-task. Neuropsychologia 29 (3), 245–254.

Knowlton, B.J., Squire, L.R., Gluck, M.A., 1994. Probabilistic classification learning in amnesia. Learning and Memory 1 (2), 106–120.

Knowlton, B.J., Mangels, J.A., Squire, L.R., 1996a. A neostriatal habit learning system in humans. Science 273 (5280), 1399–1402.

Knowlton, B.J., Squire, L.R., Paulsen, J.S., Swerdlow, N.R., Swenson, M., Butters, N., 1996b. Dissociations within nondeclarative memory in Huntington's disease. Neuropsychology 10 (4), 538–548.

Laforce, R., Doyon, J., 2001. Distinct contribution of the striatum and cerebellum to motor learning. Brain and Cognition 45 (2), 189–211.

Lehericy, S., Benali, H., Van de Moortele, P.F., Pelegrini-Issac, M., Waechter, T., Ugurbil, K., et al., 2005. Distinct basal ganglia territories are engaged in early and advanced motor sequence learning. Proceedings of the National Academy of Sciences of the United States of America 102 (35), 12566–12571.

Logan, G.D., 1988. Automaticity, resources, and memory – theoretical controversies and practical implications. Human Factors 30 (5), 583–598.

Luaute, J., Schwartz, S., Rossetti, Y., Spiridon, M., Rode, G., Boisson, D., et al., 2009. Dynamic changes in brain activity during prism adaptation. Journal of Neuroscience 29 (1), 169–178.

Lustig, C., Hasher, L., 2001. Implicit memory is not immune to interference. Psychological Bulletin 127 (5), 618–628.

Martone, M., Butters, N., Payne, M., Becker, J.T., Sax, D.S., 1984. Dissociations between skill learning and verbal recognition in amnesia and dementia. Archives of Neurology 41 (9), 965–970.

Milner, B., 1962. Les troubles de la memoire accompagnant des 1esions hippocampiques bilaterales. Physiologic de l'Hippocampe. Centre National de la Recherche Scientifique, Paris, pp. 257–272.

Moody, T.D., Bookheimer, S.Y., Vanek, Z., Knowlton, B.J., 2004. An implicit learning task activates medial temporal lobe in patients with Parkinson's disease. Behavioral Neuroscience 118 (2), 438–442.

Muslimovic, D., Post, B., Speelman, J.D., Schmand, B., 2007. Motor procedural learning in Parkinson's disease. Brain 130, 2887–2897.

Myers, C.E., Shohamy, D., Gluck, M.A., Grossman, S., Kluger, A., Ferris, S., et al., 2003. Dissociating hippocampal versus basal ganglia contributions to learning and transfer. Journal of Cognitive Neuroscience 15 (2), 185–193.

Nakamura, T., Ghilardi, M.F., Mentis, M., Dhawan, V., Fukuda, M., Hacking, A., et al., 2001. Functional networks in motor sequence learning: abnormal topographies in Parkinson's disease. Human Brain Mapping 12 (1), 42–60.

Nissen, M.J., Bullemer, P., 1987. Attentional requirements of learning – evidence from performance-measures. Cognitive Psychology 19 (1), 1–32.

Nissen, M.J., Willingham, D., Hartman, M., 1989. Explicit and implicit remembering – when is learning preserved in amnesia. Neuropsychologia 27 (3), 341–352.

Pascual-Leone, A., Grafman, J., Clark, K., Stewart, M., Massaquoi, S., Lou, J.S., et al., 1993. Procedural learning in Parkinsons-disease and cerebellar degeneration. Annals of Neurology 34 (4), 594–602.

Picard, N., Matsuzaka, Y., Strick, P.L., 2013. Extended practice of a motor skill is associated with reduced metabolic activity in M1. Nature Neuroscience 16 (9), 1340–1347.

Poldrack, R.A., Clark, J., Pare-Blagoev, E.J., Shohamy, D., Moyano, J.C., Myers, C., et al., 2001. Interactive memory systems in the human brain. Nature 414 (6863), 546–550.

Poldrack, R.A., Prabhakaran, V., Seger, C.A., Gabrieli, J.D.E., 1999. Striatal activation during acquisition of a cognitive skill. Neuropsychology 13 (4), 564–574.

Poldrack, R.A., Sabb, F.W., Foerde, K., Tom, S.M., Asarnow, R.F., Bookheimer, S.Y., et al., 2005. The neural correlates of motor skill automaticity. Journal of Neuroscience 25 (22), 5356–5364.

Puttemans, V., Wenderoth, N., Swinnen, S.P., 2005. Changes in brain activation during the acquisition of a multifrequency bimanual coordination task: from the cognitive stage to advanced levels of automaticity. Journal of Neuroscience 25 (17), 4270–4278.

Reber, P.J., Squire, L.R., 1994. Parallel brain systems for learning with and without awareness. Learning and Memory 1 (4), 217–229.

Reber, P.J., Squire, L.R., 1998. Encapsulation of implicit and explicit memory in sequence learning. Journal of Cognitive Neuroscience 10 (2), 248–263.

Robertson, E.M., 2012. New insights in human memory interference and consolidation. Current Biology 22 (2), R66–R71.

Sadeh, T., Shohamy, D., Levy, D.R., Reggev, N., Maril, A., 2011. Cooperation between the Hippocampus and the striatum during episodic encoding. Journal of Cognitive Neuroscience 23 (7), 1597–1608.

Saintcyr, J.A., Taylor, A.E., Lang, A.E., 1988. Procedural learning and neostriatal dysfunction in man. Brain 111, 941–959.

Schacter, D.L., Graf, P., 1986. Preserved learning in amnesic patients – perspectives from research on direct priming. Journal of Clinical and Experimental Neuropsychology 8 (6), 727–743.

Schendan, H.E., Searl, M.M., Melrose, R.J., Stern, C.E., 2003. An fMRI study of the role of the medial temporal lobe in implicit and explicit sequence learning. Neuron 37 (6), 1013–1025.

Schmidt, R.A. (Ed.), 1982. Motor Learning: A Behavioral Emphasis. Champaign, IL.

Seger, C.A., Cincotta, C.M., 2006. Dynamics of frontal, striatal, and hippocampal systems during rule learning. Cerebral Cortex 16 (11), 1546–1555.

Seidler, R.D., Purushotham, A., Kim, S.G., Ugurbil, K., Willingham, D., Ashe, J., 2005. Neural correlates of encoding and expression in implicit sequence learning. Experimental Brain Research 165 (1), 114–124.

Shanks, D.R., St John, M.F., 1994. Characteristics of dissociable human learning systems. Behavioral and Brain Sciences 17 (3), 367–395.

Shohamy, D., Myers, C.E., Grossman, S., Sage, J., Gluck, M.A., Poldrack, R.A., 2004. Cortico-striatal contributions to feedback-based learning: converging data from neuroimaging and neuropsychology. Brain 127, 851–859.

Soliveri, P., Brown, R.G., Jahanshai, M., Marsden, C.D., 1992. Procedural memory and neurological disease. European Journal of Cognitive Psychology 4 (3), 161–193.

Song, S.B., Sharma, N., Buch, E.R., Cohen, L.G., 2012. White matter microstructural correlates of superior long-term skill gained implicitly under randomized practice. Cerebral Cortex 22 (7), 1671–1677.

Squire, L.R., 1987. Neural substrates of memory in humans and nonhuman-primates. Bulletin of the Psychonomic Society 25 (5), 345.

Squire, L.R., 2004. Memory systems of the brain: a brief history and current perspective. Neurobiology of Learning and Memory 82 (3), 171–177.

Stefanova, E.D., Kostic, V.S., Ziropadja, L., Markovic, M., Ocic, G.G., 2000. Visuomotor skill learning on serial reaction time task in patients with early Parkinson's disease. Movement Disorders 15 (6), 1095–1103.

Walker, M.P., Brakefield, T., Morgan, A., Hobson, J.A., Stickgold, R., 2002. Practice with sleep makes perfect: sleep-dependent motor skill learning. Neuron 35 (1), 205–211.

Wang, W.C., Lazzara, M.M., Ranganath, C., Knight, R.T., Yonelinas, A.P., 2010. The medial temporal lobe supports conceptual implicit memory. Neuron 68 (5), 835–842.

Werheid, K., Ziessler, M., Nattkemper, D., von Cramon, D.Y., 2003. Sequence learning in Parkinson's disease: the effect of spatial stimulus-response compatibility. Brain and Cognition 52 (2), 239–249.

Westwater, H., McDowall, J., Siegert, R., Mossman, S., Abernethy, D., 1998. Implicit learning in Parkinson's disease: evidence from a verbal version of the serial reaction time task. Journal of Clinical and Experimental Neuropsychology 20 (3), 413–418.

Wilkinson, L., Jahanshahi, M., 2007. The striatum and probabilistic implicit sequence learning. Brain Research 1137 (1), 117–130.

Wilkinson, L., Khan, Z., Jahanshahi, M., 2009. The role of the basal ganglia and its cortical connections in sequence learning: evidence from implicit and explicit sequence learning in Parkinson's disease. Neuropsychologia 47 (12), 2564–2573.

Wilkinson, L., Lagnado, D.A., Quallo, M., Jahanshahi, M., 2008. The effect of feedback on non-motor probabilistic classification learning in Parkinson's disease. Neuropsychologia 46 (11), 2683–2695.

Wilkinson, L., Shanks, D.R., 2004. Intentional control and implicit sequence learning. Journal of Experimental Psychology – Learning Memory and Cognition 30 (2), 354–369.

Willingham, D.B., Salidis, J., Gabrieli, J.D.E., 2002. Direct comparison of neural systems mediating conscious and unconscious skill learning. Journal of Neurophysiology 88 (3), 1451–1460.

Witt, K., Nuhsman, A., Deuschl, G., 2002. Dissociation of habit-learning in Parkinson's and cerebellar disease. Journal of Cognitive Neuroscience 14 (3), 493–499.

Yin, H.H., Mulcare, S.P., Hilario, M.R.F., Clouse, E., Holloway, T., Davis, M.I., et al., 2009. Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill. Nature Neuroscience 12 (3), 333–341.

Nonequivalent Group Designs

Charles S Reichardt, University of Denver, Colorado, CO, USA
Melvin M Mark, Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 10655–10660, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The nonequivalent group design is a quasiexperiment that draws comparisons among groups of participants who are assigned to different treatments in a nonrandom fashion. Selection differences are likely to bias the estimates of the treatment effects. Analysis strategies and design features for taking account of this bias are described. These analysis strategies and design features impose different assumptions about the nature of selection. Unfortunately, researchers will seldom have enough information to know which set of assumptions is correct in any given situation. In general, the recommended approach is to minimize the bias as much as possible by making the nonequivalent groups as similar as possible.

In nonequivalent group designs, comparisons are drawn between two or more groups of study participants with each group receiving a different treatment. The purpose, as in all quasiexperiments, is to assess the relative effectiveness of the different treatments. What distinguishes a nonequivalent group design from other designs is the manner in which the study participants are assigned to the different treatments. In a randomized experiment (*see* Random Assignment: Implementation in Complex Field Settings), participants are assigned to treatments at random. In a regression discontinuity design (*see* Regression Discontinuity Design), participants are assigned to treatments based on their value on an ordered quantitative variable. In a nonequivalent group design, participants are assigned to treatments in a nonrandom, nonquantitatively ordered manner. For example, a nonequivalent group design arises when participants self-select themselves into the different treatments, when an administrator assigns participants to different treatments, or when different treatments are implemented in different sites. In analyzing data from nonequivalent group designs, researchers must be concerned with biases that can arise because of the nonrandom assignment. The discussion describes the most widely used statistical analyses and design features for addressing these biases.

Threats to Internal Validity

Because the participants are not assigned to treatments at random, the treatment groups in a nonequivalent group design will differ, to some degree or other, even before the treatments are administered. Such differences among the treatment groups are called selection differences. In the nonequivalent group design, selection differences are a threat to internal validity (*see* Internal Validity) because they can bias the estimate of the treatment effect, that is, selection differences can cause posttreatment differences between the treatment groups, even if there are no treatment effects.

In addition to selection differences, other threats to internal validity can also arise in nonequivalent group designs. For example, Cook and Campbell (1979) list several additional threats to internal validity, including compensatory equalization of treatments, compensatory rivalry, resentful demoralization, diffusion or imitation of the treatment, and local history. Threats to validity that involve recognition that different participants receive different treatments, such as resentful demoralization, are likely to be more severe in randomized experiments than in nonequivalent group designs (Fetterman, 1982; Lam et al., 1994). On the other hand, threats such as local history will often be more severe in nonequivalent group designs than in randomized experiments. Although a variety of threats to internal validity can arise, the remainder of the present discussion will be concerned only with taking account of selection differences because it is that threat that most distinguishes nonequivalent group comparisons from other types of comparisons.

The Prototypic Nonequivalent Group Design

In the prototypic nonequivalent group design, only two treatments are compared: an experimental and a comparison treatment. In addition, the participants in each of these two treatment conditions are assessed at exactly two points in time, once before and once after the treatments are implemented. Differences between the two treatment groups on the posttreatment measure are then used to estimate the effects of the treatments and the pretreatment measure is used to take account of the effects of selection differences. If there were no pretreatment measure, the results would usually be uninterpretable because there would be no way to take account of the effects of selection differences.

Various statistical procedures have been proposed for taking account of the effects of selection differences, with the assistance of the pretreatment scores. The four most commonly used methods are described below. In general, all four of these procedures are most credible when the pretreatment measure is operationally identical to the posttreatment measure.

Each of the four analysis strategies imposes different assumptions about the nature of the selection differences. If the assumptions of a given procedure are correct, the resulting estimates of the treatment effects will be unbiased by selection differences. However, if the assumptions are incorrect the

estimates of treatment effects are likely to remain biased, perhaps severely so. Therefore, researchers must be diligent in choosing an analysis strategy whose underlying assumptions fit the circumstances at hand. Unfortunately, researchers will seldom, if ever, have enough information to know which set of assumptions is correct. As a consequence, researchers are advised to conduct multiple analyses that impose a range of plausible assumptions about the nature of selection differences. Even then, researchers will typically need to be cautious in interpreting the results from prototypic nonequivalent group designs.

Change-Score Analysis

In change-score analysis, the estimate of the treatment effect is the mean difference between the experimental and comparison groups in the amount of change that takes place from the pretreatment measure to the posttreatment measure. For this change from pretreatment to posttreatment to be meaningful, the pretreatment and posttreatment measures must be operationally identical. A change-score analysis can be implemented by calculating a pretreatment to posttreatment difference (i.e., a change score) for each individual and then performing a test of the mean difference between the treatment groups on the change scores (Donald and Lang, 2007). The same result can also be obtained using the pretreatment and posttreatment scores in a repeated-measures analysis of variance.

The underlying assumption in the change-score analysis is that the mean change from pretreatment to posttreatment will be the same in the two treatment groups, in the absence of a treatment effect. This assumption may or may not be correct. Sometimes treatment groups change at different rates so the gap between them either increases or decreases over time, even in the absence of a treatment effect. Such an outcome is called a selection-by-maturation interaction. For example, gaps in income often increase over time because the rich tend to get relatively richer. Patterns such as this can be accommodated in some cases by transforming the pretreatment and posttreatment measures to have equal variances (Judd and Kenny, 1981). In other circumstances, treatment groups change so that the gap between them decreases over time. For example, the gap would tend to decrease if individuals in one or both groups were selected for their extreme scores on the pretreatment measure and subsequent performances regressed back toward the mean. This is called a regression artifact (Campbell and Kenny, 1999).

Whether alternative explanations can account for the observed results sometimes depends on the pattern of the outcomes (Cook and Campbell, 1979). For example, when the experimental group starts out below the comparison group but ends up higher, the pattern of outcomes is called a crossover interaction. Such a pattern seldom can be plausibly explained as due to either a selection-by-maturation interaction or regression toward the mean. However, patterns that can be interpreted credibly are not always easy to produce in practice.

Matching or Blocking

In matching, each participant from the comparison group is paired with a participant from the experimental group so that they have the same or similar pretreatment scores (Rosenbaum and Rubin, 1985; Dehejia and Wahba, 2002). Participants who cannot be matched are discarded from the analysis. The effect of the treatment is then estimated by comparing the posttreatment scores in each matched pair (e.g., via a matched-pair *t* test). In blocking, the participants are matched in groups (or blocks) based on their pretreatment scores, and then comparisons are drawn between the posttreatment scores within each of the blocks (e.g., via a 'randomized' blocks analysis of variance).

The degree to which matching or blocking removes bias due to selection differences depends on several factors. First, bias reduction depends on how similar the participants are on the pretreatment measures within each pair or block (Cochran and Rubin, 1973). As the number of pretreatment measures increases, usually it is harder to form pairs or blocks in which the participants are similar on all the measures. This problem can be addressed by matching or blocking participants on the basis of propensity scores, which are single-variable aggregates of multiple pretreatment measures (Rosenbaum, 2002).

Second, bias will remain to the degree the pretreatment scores contain measurement error. Under some assumptions, measurement error can be taken into account by matching or blocking participants using pretreatment scores that are corrected for their degree of unreliability.

Third, bias reduction depends on the degree to which the pretreatment measures tap all the factors that influence the posttreatment scores and on which the treatment groups differ. For example, to the extent (1) the participants within each pair or block are not matched on, say, motivation, (2) the treatment groups differ in motivation, and (3) motivation influences the posttreatment scores, then matching or blocking will fail to control for the effects of all selection differences. Unfortunately, there is seldom any way to be sure that all the relevant selection differences have been assessed and included in the analysis. Often the best, though still fallible, way to assure that most of the relevant factors are taken into account is to use pretreatment measures that are operationally identical to the posttreatment measures.

Analysis of Covariance

Analysis of covariance (ANCOVA) is analogous to matching and blocking. The difference is that ANCOVA matches participants mathematically rather than by physically forming pairs or blocks of similar participants (Reichardt, 1979). By matching mathematically, ANCOVA avoids the problem of inexact matches that can arise when using pairs or blocks, but does so at the expense of having to impose assumptions about the shape of the regression surface between the pretreatment and posttreatment scores (Van Breukelen, 2006).

An analysis of covariance is accomplished by regressing the posttreatment scores on to both pretreatment measures and a dummy variable that indicates membership in the different treatment groups. The estimate of the treatment effect is the regression coefficient for the group-membership dummy variable. Although not as problematic for ANCOVA as for matching or blocking, having a large number of pretreatment measures can make the analysis difficult to implement, and this difficulty might again be reduced by using propensity scores.

Like matching or blocking, ANCOVA is susceptible to biases due to measurement error in the pretreatment scores. Analytic adjustments can be made in ANCOVA to address unreliability in the pretreatment scores (Sorbom, 1978). ANCOVA is also just as susceptible as matching or blocking to biases due to omitting pretreatment measures that influence the posttreatment scores and on which the treatment groups differ. Unless the pretreatment measures capture all of the selection differences that influence outcomes, the results of ANCOVA, like matching or blocking, are likely to remain biased.

Selection Modeling

In selection modeling (*see* Selection Bias, Statistics of), the posttreatment scores are regressed on to both pretreatment measures and a 'selection' variable that models the assignment of participants into the treatment groups. The estimate of the treatment effect is the regression coefficient for the selection variable in this regression equation (which will be called the 'outcome' regression equation).

The selection variable that appears in the outcome regression equation can be constructed in a variety of ways depending on the circumstances (Barnow et al., 1980). One way to construct the selection variable is to set it equal to the predicted scores from what will be called a 'selection' regression equation wherein a group-membership dummy variable is regressed linearly on to both the pretreatment measures that were used in the outcome regression equation and additional pretreatment measures that were not included in the outcome regression equation. The underlying assumption is that the unique pretreatment measures that are added to the selection regression equation are related to group assignment but do not themselves influence the posttreatment measures.

An alternative way to create the selection variable is to set it equal to the predicted scores when a group-membership dummy variable is regressed onto the pretreatment measures using probit or logistic regression. Using a probit or logistic, rather than a linear, regression produces a selection variable that is a nonlinear function of the pretreatment measures. The underlying assumption in this analysis is that the relationship between the pretreatment and posttreatment scores would be linear, in the absence of a treatment effect.

More Elaborate Nonequivalent Group Designs

A variety of design features are described below that can be added to the prototypic nonequivalent group design to improve the interpretability of the results (Shadish and Cook, 1999). The logic is to create a design where the pattern of results that would arise from a treatment effect is different from the pattern of results that would arise from selection differences (Reichardt, 2000).

Multiple Outcome Measures

The credibility of inferences from nonequivalent group comparisons can sometimes be increased by adding another outcome measure, often called a nonequivalent dependent variable (Cook and Campbell, 1979). Causal inference will be enhanced if selection differences have the same effects on both the original and additional outcome measures, but the effect of the treatment differs on the two measures (or vice versa). For example, in a study of a program to aid the homeless (Braucht et al., 1995), the primary intent of the treatment was to decrease substance abuse and only secondarily to improve other outcomes such as family relationships. In contrast, selection differences, such as differences in motivation to change one's lifestyle, should have had just as large an effect on family relationships as on substance abuse. That a large outcome difference appeared on measures of substance abuse but not on measures of family relationships increased the researchers' confidence that the positive results on substance abuse were due to the treatment, rather than to selection differences.

Multiple Comparison Groups

The credibility of inferences from nonequivalent group designs sometimes can be increased by using two comparison groups in the design rather than just one. For example, a design with two comparison groups allows the experimental group to be compared with each separately. This could strengthen confidence in a causal inference if the treatment were known, say, to have an effect in the same direction in both comparisons while the effects of selection differences were known to operate in opposite directions. In such a case, similar results in the two comparisons suggests the treatment, rather than selection differences, is responsible for the outcome.

Alternatively, additional comparison groups might be included in the design so that different comparison groups could be used to satisfy the assumptions of different statistical analyses or multiple comparison groups could be used to assess the plausibility of the assumptions of a given statistical analysis. For example, a change-score analysis could benefit greatly if, relative to the experimental group, one comparison group is initially inferior and the other initially superior on the pretreatment assessment. The most interpretable pattern of results would be where the mean changes from pretreatment to posttreatment were the same in the two comparison groups but different in the experimental group. Such an outcome would rule out both a selection-by-maturation interaction and regression toward the mean more credibly than would be possible with only a single comparison group.

Multiple Time Points

It can be useful to collect both pre- and posttreatment measures at additional time points. Pretreatment measures at additional time points can be used to improve the prediction of growth trajectories in the absence of the treatment. For example, suppose pretreatment measures are collected at both times 1 and 2, different treatment protocols are introduced at time 3, and posttreatment measures are collected at time 4. Then growth trajectories estimated using the time 1 and 2 measures can help distinguish between treatment effects and selection-by-maturation interactions. With even more pretreatment measures collected over time, researchers could also assess the plausibility of regression toward the mean. In either case, growth trajectories could be estimated using either mean level

analyses such as change-score models or individual level analyses such as latent growth curve models (McArdle, 1988; Muthén and Curran, 1997) (*see* Growth Curve Analysis) or hierarchical linear models (Bryk and Raudenbush, 1992).

Alternatively, pretreatment measures at times 1 and 2 could be analyzed as if they were a pretreatment and a hypothetical posttreatment measure. Because the different treatments were not introduced until later, such 'dry-run' analyses should produce null results except to the extent they are biased by selection differences or other threats to validity. In this way, the results of such analyses could be used to choose or modify statistical procedures so they are less likely to be biased by selection differences when used to analyze the actual pretreatment and posttreatment data (e.g., Wortman et al., 1978).

In a similar vein, posttreatment observations at additional time points sometimes can be used to distinguish between treatment effects and selection differences. For example, imagine that the effect of the treatment would increase over time while the effect of selection differences would remain constant or decrease over time. In such a case, a pattern of observed differences that increase over time would contribute to one's confidence that these differences are due to the treatment. For example, the conclusion that smoking causes cancer was strengthened when it was found that the rate of lung cancer in women smokers rose 20–30 years after smoking became prevalent in women, while the lung cancer rate in nonsmoking women remained relatively constant. The alternative explanation that genetic differences between smokers and nonsmokers was responsible for the observed association between cancer and smoking (because genes might produce both a liking for cigarettes and greater susceptibility to lung cancer) could not explain such an outcome, because genetic differences would be expected to be relatively stable over that 'short' a time period.

Variations in Amounts or Types of Treatments

Adding observations at additional time points or adding multiple comparison groups provides the opportunity to introduce repetitions of the same treatment, different amounts of the same treatment, or different types of treatments. For example, suppose the treatment could be introduced, removed, and then reintroduced to the experimental group and that observations could be collected at all the time points before and after these interventions for both the experimental and comparison groups. Further, suppose the effect of the treatment is short-lived (so the treatment's effects are not present when the treatment is removed) and that the effects of selection differences are constant over time. Then a pattern of outcome differences that varies over time in accord with the varying interventions could be strong evidence in favor a treatment effect.

Another possibility is the switching replication design (Cook and Campbell, 1979) where the treatment is administered to the experimental group at one point in time and then administered to the comparison group at a later point in time, with observations being collected on each group initially and after each of the treatment administrations. The premise of this design is that the difference between the groups in the effect of the treatment varies over the time periods because the administration of the treatment varies, but the effect of selection differences remains the same because the composition of the groups remains the same.

Another potential source of complex patterns with which to distinguish treatment from selection effects are designs in which groups of participants receive varying amounts of the treatment. For example, one set of the participants in the experimental group might receive the standard dose of a treatment while other sets of participants receive either half or twice the standard dose. The underlying premise of such designs is that different treatment doses produce systematically different outcomes (sometimes called the dose–response curve for the treatment), while the effects of selection differences would follow a different pattern.

The Benefits of Disaggregation

The preceding discussion describes potential benefits of additional outcome measures, comparison groups, time points, and treatments. Incorporating these design features often requires collecting additional data, but sometimes can be accomplished by disaggregating the available data. For example, sometimes it is possible to add a nonequivalent dependent variable by disaggregating the data from an outcome variable that consists of a battery of measures. Or a comparison group sometimes can be added by dividing what was initially a single comparison group into two groups. Or comparisons between different dosage levels of the treatment might be obtained by partitioning the experimental group according to the amount of treatment received. Even additional time points sometimes can be added to a design by partitioning data that initially were aggregated, say, at yearly time intervals, into data aggregated only at monthly intervals.

Minimizing Selection Differences

Researchers should try to implement designs and analyses that are most capable of taking account of selection differences in the specific situations at hand. Applying a complex statistical procedure to data from a prototypic nonequivalent group design may be appealing in theory, but a much simpler statistical analysis combined with a more elaborate design will often be more credible in practice.

However, regardless of the simplicity or elaborateness of the research design and statistical analysis, the credibility of a nonequivalent group comparison tends to increase when selection differences are made small and the treatment effects are made large. Random assignment tends to make selection difference small. In nonequivalent group designs, selection differences can often be minimized by using an assignment procedure that is as close to random as possible. For example, Langer and Rodin (1976) assessed the effects of a treatment by drawing comparisons across the clients on different floors of a nursing home. In this case, the assignment of clients to treatment conditions appeared to be quasirandom in that there were no obvious mechanisms by which different types of clients would be assigned systematically to different floors.

Often selection differences are smaller when the experimental and comparison groups are composed of individuals recruited from the same organization or locale (called internal

controls) than from different organizations or locales (called external controls). A special case of internal controls arises when the experimental and comparison groups are made up of earlier and later cohorts (Cook and Campbell, 1979). For example, the effect of a treatment that is given to college sophomores in a given year could be assessed by drawing a comparison with the performance of sophomores at that college from the prior year. Such a design is worth considering when the cohort of sophomores from the prior year is more similar to the treatment group than would be a contemporaneous comparison group of freshman or juniors from the same college or of sophomores from a different college.

Summary

Primarily because of biases due to selection differences and because of the difficulty in taking these biases into account, nonequivalent group designs generally produce less credible results than randomized experiments. Nonetheless, nonequivalent group designs are used widely because they can be far easier to implement than randomized experiments. Researchers should try to take account of the biasing effects of selection differences by the judicious combination of analysis strategies and design features. It is also advisable to minimize selection differences a priori so as to minimize one's reliance on either statistical procedures or design elaborations.

See also: Causal Inference: Overview; Comparative Studies: Method and Design; Experimental Design: Bayesian Designs; Experimental Design: Overview; Experimental Design: Randomization and Social Experiments; Field Observational Research in Anthropology and Sociology; Hierarchies and Markets; Inequality Measurement; Internal Validity; Latent Structure and Causal Variables; Mechanism-Based Causal Analysis; Observational Studies: Overview; Psychological Treatments: Randomized Controlled Trials; Quasi-Experimental Designs; Regression Discontinuity Design; Sample Surveys: Cognitive Aspects of Survey Design; Selection Bias, Statistics of; Social Experiments.

Bibliography

Barnow, B.S., Cain, G.C., Goldberger, A.S., 1980. Issues in the analysis of selectivity bias. In: Stromsdorfer, E.W., Farkas, G. (Eds.), Evaluation Studies Review Annual, vol. 5. Sage, Newbury Park, CA.

Braucht, G.N., Reichardt, C.S., Geissler, L.J., Bormann, C.A., Kwiatkowski, C.F., Kirby Jr., M.W., 1995. Effective services for homeless substance abusers. Journal of Addictive Diseases 14, 87–109.

Bryk, A.S., Raudenbush, S.W., 1992. Hierarchical Linear Models: Applications and Data Analysis Methods. Sage, Newbury Park, CA.

Campbell, D.T., Kenny, D.A., 1999. A Primer on Regression Artifacts. Guilford Press, New York.

Cochran, W.G., Rubin, D.B., 1973. Controlling bias in observational studies. Sankhya (Series A) 35, 417–446.

Cook, T.D., Campbell, D.T., 1979. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Rand McNally, Chicago.

Dehejia, R.H., Wahba, S., 2002. Propensity score-matching methods for nonexperimental causal studies. Review of Economics and Statistics 84, 151–161.

Donald, S.G., Lang, K., 2007. Inference with difference-in-differences and other panel data. The Review of Economics and Statistics 89, 221–233.

Fetterman, D.M., 1982. Ibsen's baths: reactivity and insensitivity. Educational Evaluation and Policy Analysis 4, 261–279.

Judd, C.M., Kenny, D.A., 1981. Estimating the Effects of Social Interventions. Cambridge University Press, New York.

Lam, J.A., Hartwell, S.W., Jekel, J.F., 1994. I prayed real hard, so i know i'll get in': living with randomization. In: Conrad, K.J. (Ed.), Critically Evaluating the Role of Experiments. New Directions for Program Evaluation, vol. 63. Jossey-Bass, San Francisco, pp. 55–66.

Langer, E.J., Rodin, J., 1976. The effects of choice and enhanced personal responsibility for the aged: a field experiment in an institutional setting. Journal of Personality and Social Psychology 34, 191–198.

McArdle, J.J., 1988. Dynamic but structural equation modeling of repeated measures data. In: Nesselroade, J.R., Cattell, R.B. (Eds.), The Handbook of Multivariate Experimental Psychology, second ed. Plenum, New York, pp. 561–614.

Muthén, B.O., Curran, P.J., 1997. General longitudinal modeling of individual differences in experimental designs: a latent variable framework for analysis and power estimation. Psychological Methods 2, 371–402.

Reichardt, C.S., 1979. The statistical analysis of data from nonequivalent group designs. In: Cook, T.D., Campbell, D.T. (Eds.), Quasi-Experimentation: Design and Analysis Issues for Field Settings. Rand McNally, Chicago, pp. 147–205.

Reichardt, C.S., 2000. A typology of strategies for ruling out threats to validity. In: Bickman, L. (Ed.), Research Design: Donald Campbell's Legacy, vol. 2. Sage, Thousand Oaks, CA, pp. 89–115.

Rosenbaum, P.R., Rubin, D.B., 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician 39, 33–38.

Rosenbaum, P.R., 2002. Observational Studies. Springer, New York.

Shadish, W.R., Cook, T.D., 1999. Comment–design rules: more steps toward a complete theory of quasi-experimentation. Statistical Science 14, 294–300.

Sorbom, D., 1978. An alternative methodology for analysis of covariance. Psychometrika 43, 381–396.

Van Breukelen, G.J., 2006. ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. Journal of Clinical Epidemiology 59, 920–925.

Wortman, P.M., Reichardt, C.S., St Pierre, R.G., 1978. The first year of the educational voucher demonstration: a secondary analysis of student achievement test scores. Evaluation Quarterly 2, 193–214.

Nongovernmental Organizations: Anthropological and Historical Aspects

Julia Elyachar, School of Social Sciences, UC of Irvine, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the brief history of the NGO, including its relation to development, neoliberalism, the United Nations, and mutual aid organizations. NGOs gather and render actionable forms of knowledge from diverse locations, at the interface of financial infrastructures and fields of organizational power. Like ethnography, NGOs render local forms of embodied knowledge explicit, quantifiable, and commensurate with knowledge in other locations of the globe. This has made them a popular if problematic location for fieldwork. As a flexible socio-technology, the NGO form interpolates global flows of knowledge, power, and finance.

Knowing and Working with NGOs

What are NGOs? This simple question has proven difficult to answer. Many social scientists write about NGOs. Many anthropologists in particular find professional employment in NGOs. NGOs are a common starting point for the signature research method of anthropology – fieldwork. Many things are confusing about NGOs. Perhaps the most confusing is how straight forward NGOs seem to be. The very name 'NGO' is a definition: An NGO is a nongovernmental organization. But that name/definition obscures as much as it explains. 'NGO' is neither a concept grounded in social or political theory, nor did it evolve out of critical engagements with classical social theory. NGO is not a translation of any particular 'native concept.' At the same time, something resembling NGOs exist everywhere. So do the local precedents for the NGO. Mutual aid societies across the (sometimes overlapping) Islamic majority world and the African continent have much in common with NGOs. So do the nineteenth-century Western charitable organizations aimed at reform and self-improvement of the poor in the West, and in the far reaches of empire at the turn of the twentieth century. NGOs look different in different places, depending on the nature of the state in which they are located and where they are active.

Social scientists encountered NGOs in the course of their research and began to write about them in the late 1980s. Influential reviews began to be published in the late 1990s (Fisher, 1997; Keck and Sikkink, 1998; Kearney, 1995; Mertz and Timmer, 2010). A next wave of research was presented in conference panels of the American Anthropological Association (AAA), organized by the Association of Political and Legal Anthropology in particular. By 2012, there was a robust body of published research on NGOs in social sciences, including anthropology. NGOs were being discussed within a wide range of social science disciplines, subfields, and an 'interest group' on NGOs in the AAA. Since NGOs are now everywhere, no area studies association specializes in their study.

NGOs are a kind of private voluntary organization (PVO). But not all PVOs are NGOs. Some NGOs are charitable organizations, but not all charitable organizations are NGOs. The organizing logic of NGOs is not profit, but NGOs are key intermediaries in global financial transfers. The organizing logic of NGOs is not formal politics, but NGOs are central to global political economy and the remaking of politics since the 1980s. NGOs appear in many shapes and forms. Some are linked to states, and some are linked to corporations. They can range from small local volunteer organizations to large international organizations with a professional staff. NGOs associated with states or corporations used to be seen as an aberration. The NGO was, in this first wave of research, assumed to be the organizational self-expression of 'the people.' But as the list of exceptional NGOs grew, it became harder to see them as exceptions. Overall, attempts to explain the NGO with recourse to another social science concept – be it civil society, neoliberalism, or antidevelopment – have proven insufficient. Recent writing on NGOs in anthropology has tended to be less normative in approach, and to focus more on what anthropology does best – analysis of tacit practices juxtaposed with analysis of programmatic statements.

Social scientists work with and on NGOs for different reasons. Some are pragmatic. Where it was once common for anthropologists to take 10 years to complete a degree that involved learning foreign languages and spending more than a year in the field, students are now under constraints to finish their PhD in 6 years. In the new climate, NGOs offer important benefits. English is the lingua franca of the NGO world. A student conducting fieldwork from an NGO need not invest heavily in language study. NGOs gather information about local conditions. That information is important to anthropologists and other social scientists for research. It is important to development agencies, international organizations, and corporations as well. Researchers who conduct fieldwork from NGOs can bypass immersion in the "imponderabilia of everyday life," to use the phrase made famous by anthropologist Branislow Malinowski, even while producing knowledge that is certifiable as 'local' and valuable for a number of ends. Furthermore, NGOs provide ethnographers with a ready-made set of introductions to informants. Entering the field via an NGO has implications for research outcomes that also need to be addressed. NGOs provide employment for an elite segment of educated youth throughout the global South. This leads to a growing congruence between the knowledge practices of social scientists studying NGOs and those working in NGOs (Elyachar, 2005; Riles, 2003). At the same time, to bypass the study of NGOs because they are not 'ethnographic' in the traditional sense would be absurd. The best ethnographic

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12214-7

research on NGOs situates them in a broader discursive and ethnographic field, and reveals their role in mediating other flows of meaning, value, power, and faith.

A (Very) Brief History of the NGO

The term 'nongovernmental organization' was formalized with the establishment of the United Nations (UN) in 1945. In the UN Charter, 'nongovernmental associations' were assigned 'consultative status' with the Economic and Social Council (ECOSOC). ECOSOC oversees about 70% of the resources of the entire UN system, including 14 specialized agencies, nine 'functional' commissions, and five regional commissions. The NGO branch services 3500 NGOs in consultative status with the ECOSOC, as well as "all other NGOs seeking to work with the UN" (http://csonet.org/).

And yet, few PVOs called themselves NGOs until the 1980s. NGOs were first studied in the field of development. Development is a discourse and a set of practices that organized relations between the former colonial powers, the US in its ascendency, and former colonies and territories of defeated empires. Before the 1980s, this went on via states and shifts in global financial flows are essential prehistory. From 1973 until the 1982 Latin American debt crisis, the US banking system channeled 'sovereign loans' to postcolonial states to carry out their development projects. This direct lending complemented lending through Bretton Woods institutions, in particular the World Bank. After the Latin American debt crisis in 1982, opinion about state-led development (as well as sovereign lending to Latin American countries) began to shift: postcolonial states began to be seen as 'corrupt.' These macro-level shifts in financial flows between global North and South are essential prehistory of the NGO. In this moment, 'communities' and 'social movements' appeared to be a better and nonoppressive framework through which to work for the improvement of life-chances of poor people. NGOs arose as an informal organizational vehicle to advance the aspirations of indigenous communities, social movements, or poor women.

By the 1990s, anthropologists began to critique NGOs as 'neoliberal' (Karim and Leve, 2001). The NGO form indeed overlaps with neoliberalism as a mode of governing in a number of ways. Neoliberalism is a constructivist project (Brown, 2003). The NGO has been one important agent of that construction, actively implanting new kinds of normative economic rationality in multiple spheres of life. Around the world, NGOs helped create new economic subjectivities of the dynamic 'entrepreneur' as opposed to class-based solidarity or citizenship (Elyachar, 2005; Brown, 2003). Indeed, neoliberalism provided the discursive terrain around which the organizational form of the NGO took shape. But as both neoliberalism and NGOs proliferated and transformed, it became untenable to explain one by way of the other.

The Rise of Civil Society

Three long-term (and related) transformations seem tied to the rise of NGOs in the 1980s: the appearance of civil society movements in eastern Europe and the Soviet Union, the collapse of the Soviet Union and the International Communist movement, and an interlinked set of phenomena glossed as globalization. 'Civil society' was the organizing concept of movements across eastern Europe in the 1980s that aimed at opening space for the 'social' outside the sway of the totalitarian state (Mastnak, 2005). Western scholars discovered, championed, and wrote about this movement, setting the stage for an uncritical view of 'civil society' as a space of freedom. The civil society movement was marked by the politics of antipolitics and a lack of theory of the state or economy. In contrast with the centrality of ideology and party rule under communism, the civil society movement stressed the rule of law and a legal rules-based approach (Mastnak, 2005).

The eastern Europe experience was a crucible in which development organizations and the US foreign policy establishment adopted civil society with religious zeal to champion in the 1990s across Latin America and Africa as well as the former communist world (Comaroff and Comaroff, 2000). The antipolitical and rules-based approach of the civil society movement was positioned in opposition to the 'totalitarian state.' The 'nonpolitical' nature of the civil society movement laid a perfect staging ground for the rise of the NGO. This pragmatic orientation took on different meanings after the fall of the Berlin Wall. In some places, the NGO became an excuse for the offloading of entire state institutions, as in some countries on the African continent (Ferguson and Gupta, 2007). At other places and times, the nonpolitical reputation of the NGO allowed political activists to carry out work that would have gotten them arrested were it conducted in the framework of (banned) political parties.

Critique of Development and Globalization

The rise of the NGO also coincided with the critique of development. A wide-ranging critique in the social sciences in the 1980s pointed to the oppressive power relations inherent to the very practice and notion of development. While development had once seemed to be a scientific approach to improving the quality of life for the populations of newly independent states, the critique of development argued that development was actually detrimental to local populations (Elyachar, 2012). This approach went much further than criticism of a particular development approach. Rather, it pointed to oppressive power relations inherent to development as a whole. The critique of development was implicitly a critique of the state as well, since the state was the primary vehicle through which development was effected and carried out together with interstate post-World War II institutions such as the World Bank. Social movements and NGOs came to seem like an alternative to development. NGOs could be a vehicle for 'globalization from below' (Appadurai, 2000) and a potential site for the generation of alternative forms of knowledge and power that could counter the weight of 'the state–capital nexus' (Appadurai, 2000). The notion of 'empowerment' that had begun in the women's movement shifted into development world, and NGOs became central to any plan aimed at empowerment of women and increasing their participation in development and economic growth.

Research burgeoned into apparent alternatives to state-led development: NGOs, social movements, and the informal economy. Traditional approaches to development were overthrown. 'The people' were no longer in need of development. Rather, the developed world needed to learn from backward peoples and their cultural practices. Poor peoples' social networks and their traditional cultural practices would be something to learn from instead. They were a gold mine for the 'social capital' essential for a well-functioning market society. The 'antiglobalization movement' of this period was also deeply critical of the state and the state-led development. NGOs found champions here as well. NGOs represented 'the people,' free of the state. This focus on 'the people' was part of a broader transformation in the nature of political life and public policy (Clarke, 2010). All of these efforts came together in the new microenterprise and microfinance movement, which was channeled through NGOs. Access to credit would set the poor and their entrepreneurial capacities free. Access to credit would tap new sources of economic growth for the economy as a whole. Over the next two decades, this approach to microlending moved out of the incubator of NGOs and into mainstream financial institutions (Roy, 2012).

The Left, Knowledge, and Communications Infrastructures

The end of the Soviet Union in 1991 is also linked to the rise of the NGO. Many leftist organizations, political movements, and solidarity movements around the world collapsed in the early 1990s. Solidarity movements in the West were essential to the mobilization of funding flows to liberation movements in the global South – in Latin America and Africa in particular. Those financial flows were not huge. But in retrospect we can see how they created channels for transmission of information in one direction and finance in the other. NGOs regularized this communicative and financial infrastructure that had been implicit in the original 'consultative role' of the NGO in the 1940s.

Across the third world, activists in local liberation movements shifted into NGO work, with a widespread 'NGO-ification' of politics. Those who might have been part of political movements in the 1970s found employment opportunities in NGOs instead. Solidarity workers in the global North moved into NGO work as well. Solidarity movements had relayed remarkably accurate and valuable information about conditions on the ground in battleground countries of the Cold War in Central and Latin America through the 1980s. In Latin America, NGOs built on former communications infrastructures developed by solidarity movements, and expanded into new communications technologies. Activists in remote locations in the 1980s had taught local NGO members how to access the Internet, and how to patch into telecommunications networks when infrastructure was not otherwise available. All of this helped create a new communications infrastructure across the global South of which NGOs were an essential part. This helped lay the building blocks for a number of trends to come, ranging from microenterprise, microlending, women's empowerment, mobile banking, and bottom of the pyramid (BOP) corporate planning.

Warfare, Humanitarianism, and Human Rights

The NGO began as a part of interstate warfare and its containment, attending to the suffering of ordinary soldiers on the battlefield, as told in histories of the Red Cross and humanitarianism (Bornstein and Redfield, 2010). It came into legal existence with the establishment of the UN at the end of World War II. The NGO came into its own toward the end of the Cold War with the rise of the civil society movement in eastern Europe. The end of the Cold War with the collapse of the Soviet Union set the stage for the proliferation of NGOs, as we saw in the wake of the US invasion of Afghanistan and Iraq. NGOs have been popping up like mushrooms in the wake of conflict situations ever since. NGOs poured into Afghanistan together with the US Army, bidding up the price of labor, supplies, and property, and distorting local economies as they carried out their missions for good. In the process, links were deepened between NGOs and the corporate world, via the military (Nordstrom, 2004). Recent work on NGOs in the aftermath of conflict also looks at different ways of doing good (Bornstein, 2012) and humanitarianism (Fassin, 2011). In the context of NGOs, actions carried out by individuals in NGOs become a form of moral action linked to visions of a greater good in the future (Bornstein and Redfield, 2010), or a modality of 'killing with kindness' (Schuller, 2012).

Human rights NGOs have been the most directly political of NGOs in many countries and have mediated transnational human rights movements and forms of local activism (Merry, 2006). They have gathered information about human rights violations in many countries of the global South since the 1990s (Abdulrahman, 2005; Willen, 2012). Human rights NGOs have often had a contentious relationship with the states in which they are located. In Egypt, for example, human rights NGOs were often accused by the Mubarak regime of being agents of the United States and the West. (Such accusations are made in many countries where the state has politics opposed to the United States.) Information gathered by Egyptian NGOs about police brutality and torture in prisons under Hosni Mubarak played an important role in creating international legitimacy for the overthrow of the Mubarak regime. Once in power, the new government led by Mohamed el-Mursi prosecuted some of those NGOs in turn. The link between anthropology and human rights has provoked research in all parts of the world where anthropologists work, and leads to theoretical reconsiderations as well (Willen, 2012).

The Religion of NGOs

Institutionalized forms of charity and giving can also be found in all religious traditions. The story of the transformation of indigenous forms of giving and regulating giving into something recognizable as an NGO is particularly striking in the case of religious NGOs. NGOs, like 'civil society,' were assumed in the 1980s to be 'secular.' The NGO as the building block of secular civil society was taken up with religious fervor in the Middle East. NGOs would counter 'Islamic fundamentalism' through the strengthening of civil society and its economic incarnation: the informal economy and microenterprise. At the same time, Islamist political parties, movements, and

associations were providing aid and succor of the kind associated with the NGO. Social development and charitable initiatives are central modalities of ethical being in the world. Such sustained practices of doing good and providing social and development services filled gaping holes in the provision of social services.

In the case of Israel/Palestine, Islamist parties began to fill some of the stark needs of an actively 'de-developed society' (Roy, 1999). There, too, with the decline of the left and its own solidarity movements and self-help initiatives, Islamist movements filled in the gap. These practices that in substance so resembled those of NGOs were adamantly kept out of the main narrative of 'development,' which was assumed to be 'secular' despite the roots of development in Christian theology (Elyachar, 2012). Islamist forms of 'doing good' were often called 'terrorist' after 11 September 2001 (de Goede, 2012).

Associational life is nothing new in today's Middle East (Abdulrahman, 2005; Atia, 2011). Systematized forms of giving, not to mention formal legal institutions such as *waqf*, are recorded in *hadith* of the Prophet Muhammad and are a central part of Islamic history (Atia, 2011). The same is true of Jewish communal institutions, which have always organized to support their religious brethren in far away places, sending help and succor across the stretches of empire and ocean. The establishment of the Jewish National Fund (JNF) in 1901 as an entity to gather and channel resources for Jews through and around a state defined as both Jewish and democratic has intriguing hallmarks of a vast NGO. At the same time, in a state defined as both democratic and Jewish, NGOs organized to provide development assistance to Palestinian citizens building on the earlier secular language associated with the NGO are targeted as a security threat. Research on NGOs self-identified as Christian has also been increasingly important, starting with Bornstein's book on Protestant NGOs in Africa (Bornstein, 2005).

Microcredit has been an interesting terrain bringing together the NGO with ethical subjectivities long associated in the research with 'neoliberalism' but now taking on explicitly Christian or Islamic orientations. Especially in the wake of the Arab Spring, Islamist political parties have adopted social technologies of microcredit and microlending that were once associated with the Mubarak government in Egypt. 'Islamic NGOs' are being recognized and cultivated as partners for global financial flows (Atia, 2011). They are becoming part of a global political and financial order. This could be noted, for example, in an article in *The Financial Times* in 2012 about the work of Hernando de Soto, champion of the microenterprise and entrepreneurial self-help capitalism, with the Muslim Brotherhood government in Egypt.

BOP, Social Business, and NGOs

In recent years, the NGO has also moved into the heart of the corporate world. The notion of a corporate NGO seems a misnomer. Nevertheless, as written in the World Association of NGOs' (WANGO) *NGO Handbook*, the NGO is nonprofit as well as nonstate. But local knowledge has become increasingly important in the corporate world, and as 'social goals' become increasingly integrated into mission statements, the import of the NGO – and ethnography – into the corporate world has become evident. Like ethnography, NGOs render explicit cultural practices that were once known only to insiders. As increasingly important mediators of global financial flows, NGOs are key nodes of translation of culture into economic value (Elyachar, 2005). They render explicit tacit practices in particular locations and render them commensurable in a global order of political and financial arrangements. Given the centrality of 'knowledge' as a category in the corporate world in the past 30 years, and the growing importance of poverty markets (Roy, 2012), it should not surprise us to see the pervasive spread of the NGO in the for-profit world.

This appearance of the NGO in the corporate world is linked to the rise of the movement for 'corporate social responsibility' (CSR) (Browne and Milgram, 2009). The rise of CSR in the corporate world drew on another concept of the 'stakeholder' in management theory. Others besides direct shareholders of a company are also affected indirectly by its actions. Their voices also need to be heard. The management world began to discuss unintended consequences or 'externalities' outside the range of economic theory or measurement. Here, the consultative role for the NGO outlined in the UN charter took on a new function, in an area unthinkable when the UN was founded: the world of the for-profit corporations. In later theories of the BOP, social value was moved to the forefront of corporate strategic planning, rather than being tacked on at the end. As argued in the seminal work of Prahalad, "Marrying the local knowledge of the NGO with the global reach of the multinational firm" could create new solutions for global problems even as it creates new sources of profit for MNCs (Prahalad, 2004).

In Prahalad's formulation, NGOs became incubators for business models, "lead experimenters in BOP markets" (Prahalad, 2004: 56). NGOs could help MNCs understand not only what poor people wanted as consumers, but also learn to take advantage of the incipient infrastructures poor people had built without compensation or help from the state. The NGO would be a translator and conveyor belt of knowledge from the BOP to the MNC.

Knowledge and Action Research: The Pragmatics of NGOs

Conceptual and epistemological overlaps among management theory, theories of knowledge in private firms, and NGOs are a newer area of interest. One framework to analyze those overlaps is through the prism of 'action research.' This overlap can be seen in how NGOs in all spheres of their work have implemented and participated in projects to institute best practices and create actionable research in their field of interest. Future research will need to trace the concepts used by NGOs and follow their shared genealogies in the fields of economic theory, organizational management, practice theory, and critical theory.

Research on the anthropological aspects of NGOs can draw on organizational theory, management theory, and theory of knowledge and information. It would also be fruitful to explore a number of other related topics: the role of pragmatism in relation to NGOs, how NGOs function as a mechanism of valuation, and how NGOs function as a devise for rending the tacit visible and for rendering apparently incommensurable

qualities commensurable as quantities. For this kind of work, we will need to take a classic anthropological take on such matters: study what NGOs say they do, study what they actually do, and try to get a grasp on the 'imponderabilia of everyday life' that make them tick, to return to Malinowski's term for the object of anthropological knowledge in fieldwork among natives. In the NGO as well, we need to pay more attention to the mundane, everyday worlds of NGOs in action and on the ground.

It will likely prove important to return to the original role of the NGO in its foundational movement of the Bretton Woods system and the UN: gathering and rendering actionable forms of knowledge from diverse locations, at the interface of complex systems regulating financial flows and structures of organizational power. NGOs are a central part of the financial and political order put into place at Bretton Woods together with the UN. That moment reorganized systems of finance, power, and international order. The origin of the NGO is located at this global nexus of violence, finance, and information flows. This peculiar location, I want to suggest, may just be what makes the NGO an NGO.

At the same time, there may be good reasons why we still lack a theory of NGOs. So many things constitute a referent to 'NGO' that any concept potentially delimiting them can end up being negative (not the state), vague (organizations to do good), or normative. The NGO is associated with many things, but is an expression of no one of them. NGOs might best be understood as a type of organization that arose concomitant with a discourse of 'neoliberalism,' and came to function well as a flexible socio-technology interpolating global flows of knowledge, power, and finance.

Social science research into NGOs has tended to move between two tendencies. One approach has attempted to pinpoint the true nature or essence of the NGO, in all of its variable locations. Another approach has documented the specificity of what particular types of NGOs do in one particular location. Neither approach can suffice. Like ethnography, NGOs render local forms of tacit embodied knowledge explicit, quantifiable, and commensurate with knowledge in other locations of the globe. The NGO now lies at the nexus of violence, finance, and information flows the world over. The myriad ways in which they do so could not have been envisioned by those who established the NGO together with the UN at the end of World War II.

See also: Conflict and War: Anthropological Aspects; Development: Social-Anthropological Aspects; Liberalism, Anthropology of; Organizational Culture, Anthropology of; Political Anthropology.

Bibliography

Abdulrahman, Maha, 2005. Civil Society Exposed: The Politics of NGOs in Egypt. I.B. Tauris, London.

Appadurai, Arjun, 2000. Grassroots globalization and the research imagination. Public Culture 12 (1), 1–19.

Atia, Mona, 2011. "A way to paradise": pious neoliberalism, Islam, and faith-based development. Annals of the Association of American Geographers 102 (4), 808–827.

Bornstein, Erica, 2012. Disquieting Gifts: Humanitarianism in New Delhi. Stanford University Press, Stanford.

Bornstein, Erica, 2005. The Spirit of Development: Protestant NGOs, Morality, and Economics in Zimbabwe. Stanford University Press, Stanford.

Bornstein, Erica, Redfield, Peter, 2010. Forces of Compassion: Humanitarianism between Ethics and Politics. School for Advanced Research Press, Santa Fe.

Brown, Wendy, 2003. Neoliberalism and the end of liberal democracy. Theory and Event 7 (1), 1–43. Accessed online: http://muse.jhu.edu/journals/theory_and_event/v007/7.1brown.html.

Browne, Katherine, Milgram, Lynne (Eds.), 2009. Economics and Morality: Anthropological Approaches. Altamira Press, Lanham, MD.

Clarke, John, 2010. Enrolling ordinary people: governmental strategies and the avoidance of politics? Citizenship Studies 14 (6), 637–650.

Comaroff, Jean, Comaroff, John L., 2000. First thoughts on a second coming. Public Culture 12 (2), 291–343.

Elyachar, Julia, 2005. Markets of Dispossession: NGOs, Economic Development, and the State. Duke University Press, Durham, NC.

Elyachar, Julia, 2012. Next practices: knowledge, infrastructure, and public goods at the bottom of the pyramid. Public Culture 24 (1), 109–129.

Encyclopaedia Britannica Online, September 16, 2012. Non-governmental Organization (NGO). Encyclopaedia Britannica Inc. Web edition.

Fassin, Didier, 2011. Humanitarian Reason: A Moral History of the Present. University of California Press, Berkeley.

Ferguson, James, Gupta, Akhil, 2007. Spatializing states: toward an ethnography of neoliberal governmentality. American Ethnologist 29 (4), 981–2002.

Fisher, William F., 1997. Doing good? The politics and antipolitics of NGO practices. Annual Review of Anthropology 26, 439–464.

de Goede, Marieke, 2012. Speculative Security: The Politics of Pursuing Terrorist Monies. University of Minnesota Press, Minneapolis.

Holmes, Douglas R., Marcus, George, 2008. "Para-Ethnography." The SAGE Encyclopedia of Qualitative Research Methods. SAGE Publications, New York. http://knowledge.sagepub.com/view/research/n307.xml.

Karim, Lamia, Leve, Lauren, 2001. Privatizing the state: ethnography of development, transnational capital, and NGOs. Political and Legal Anthropology Review 24 (1), 53–58.

Kearney, Richard, 1995. The local and the global: the anthropology of globalization and transnationalism. Annual Review of Anthropology 24, 547–565.

Keck, Margaret E., Sikkink, Kathryn, 1998. Activists beyond Borders: Advocacy Networks in International Politics. Cornell University Press, Cornell.

Mastnak, Tomaz, 2005. The reinvention of civil society: through the looking glass of democracy. European Journal of Sociology 46 (2), 323–355.

Merry, Sally E., 2006. Transnational human rights and local activism: mapping the middle. American Anthropologist 108 (1), 38–51.

Mertz, Elizabeth, Timmer, Andria, 2010. Introduction: getting it done: ethnographic perspectives on NGOs. Political and Legal Anthropology Review 33 (2), 171–177.

Nordstrom, Carolyn, 2004. Shadows of War: Violence, Power, and International Profiteering in the Twenty-First Century. University of California Press, Berkeley and Los Angeles.

Prahalad, C.K., 2004. The Fortune at the Bottom of the Pyramid: Eradicating Poverty through Profits. Harvard University Press, Cambridge.

Riles, Annelise, 2003. Anthropology in the realm of the pragmatic: studying pragmatism in law and politics. PoLAR: Political and Legal Anthropology Review 26 (2), 1–7.

Roy, Ananya, 2012. Ethical subjects: market rule in an age of poverty. Public Culture 24 (1).

Roy, Sara, Spring 1999. De-development revisited: palestinian economy and society since Oslo. Journal of Palestine Studies 28 (3), 64–82.

Schuller, Mark, 2012. Killing with Kindness: Haiti, International Aid, and NGOs. Rutgers University Press, New Brunswick, NJ.

Willen, Sarah S., 2012. Anthropology and human rights: theoretical reconsiderations and phenomenological explorations. Journal of Human Rights 11, 150–159.

World Association of Non-Governmental Associations (WANGO), 2010. NGO Handbook. http://www.ngohandbook.org/index.php?title=Main_Page.

Non-Heterosexual British Asian Men: Social Representations, Identity and Social Relations

Rusi Jaspal, De Montfort University, Leicester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores social representations, identity, and social relations among nonheterosexual British Asian men. After an overview of two relevant theories from social psychology – identity process theory and social representations theory, the article outlines dominant social representations of sexuality that inform individuals' meaning-making, the impact of these representations for identity processes, how individuals cope with potential threats to identity, and the development and management of social relations. It is argued that being Asian and gay can result in threatened identity, leading to low psychological well-being and negative interpersonal relations.

Introduction

In the Western world, the social, political, and legal positions of nonheterosexual individuals have greatly improved, leading to greater tolerance and acceptance at a societal level. Yet, nonheterosexual ethnic and religious minority individuals can continue to face discrimination from within their ethnic and religious communities (Yip, 2004). Stigma can result in decreased willingness to 'come out' and perceived conflict between sexual and ethnoreligious identities (Jaspal and Siraj, 2011). Identity conflict can lead to a variety of negative social and psychological outcomes – the experience of negative emotions, low psychological well-being, and social isolation. This article explores identity, social representations, and social relations among one such ethnoreligious minority group in the UK context – nonheterosexual British Indian and British Pakistani men. Here, the focus is on nonheterosexual men, rather than women, because most previous research has concentrated on the identities, experiences, and relationships of men.

British Asians: An Overview

The term 'British Asian' denotes a superordinate ethnoracial category, used typically to refer to individuals of Indian, Pakistani, or Bangladeshi descent. At the time of the 2011 UK Census, (http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/index.html) there were 2 984 670 individuals of South Asian descent resident in England and Wales. Within this superordinate category, Indians were the largest group (1 412 958); Pakistanis the second largest (1 124 511); Bangladeshis the third (447 201). British Asians constitute over 50% of the non-White population and are therefore the largest visible ethnic minority group in the country.

The sociodemographic makeup of British Asians is diverse and complex. The overwhelming majority of British Pakistanis are practicing Muslims, and most British Indians are, at least, nominally Hindu or Sikh (Ghuman, 2003). There have been various waves of immigration to the United Kingdom from the Indian subcontinent and therefore this is a demographically diverse group. Despite the sociodemographic variation, what most of the Asian immigrants shared was their socialization in a largely collectivist culture, which attached importance to kinship ('biraderi') and honor ('izzat'). These cultural tenets have remained powerful in the British Asian community (Ghuman, 2003), and in many cases, there remains a preoccupation among Asian parents that their children should be 'shielded' from the perceived immorality of elements of British/Western culture. Although there is little research into attitudes toward homosexuality among first-generation Asian immigrants, there is some evidence that this may be regarded in terms of a 'Western disease' that has 'infected' the Asian community (see Yip, 2007). Against this backdrop, this article examines and compares how men from the largest two nonheterosexual groups, British Indian and British Pakistani, construct identity and manage social representations and social relations.

Terminology

Various terms are used in academic work on sexuality. Terms such as 'gay,' 'lesbian,' and 'bisexual' are sometimes uncritically applied to research participants, despite the possibility that individuals themselves may deny that they are accurate descriptors of their sexual identity. Sexual self-definition depends on dominant social representations of sexual identity categories – what does it mean to be 'gay'? What are the cultural images of a 'typical' gay person? Where am I located in relation to these images? Constructivist approaches to sexuality emphasize fluidity in processes of self-categorization and self-definition (Seidman, 2003). Moreover, sexual self-definition can fluctuate throughout the life span – some individuals may initially view themselves as 'straight,' later concede that they may be 'bicurious,' and finally accept that they are actually 'gay.' As individuals encounter social representations of sexual identity categories, they may become more or less willing to embrace them. Thus, researchers should ensure that their terms capture potential fluidity in sexual self-identification.

In this article, the term 'nonheterosexual' is used. It has proven to be useful in research with ethnic and religious minority individuals, given that it embraces the plethora of sexual identity labels that individuals tend to use in contrast to 'heterosexual' (Yip, 2004). Moreover, it does not prioritize

any particular label (e.g., gay), which some individuals may find constraining and as inaccurately reflecting their sexual identity.

Theoretical Framework

In examining the construction and maintenance of identity, social representations, and social relations among nonheterosexual British Asian men, two theories from social psychology – identity process theory and social representations theory – have proven to be particularly insightful.

Identity Process Theory

Identity process theory (Breakwell, 1986; Jaspal and Breakwell, 2014) has provided a useful heuristic lens for understanding the construction and management of identity among nonheterosexual British Asian men. The theory integrates identity construction, threat, and coping, while synthesizing the social and psychological levels of analysis. Identity process theory proposes that identity construction is regulated by two universal processes, namely, (1) assimilation–accommodation and (2) evaluation. The assimilation–accommodation process refers to the absorption of new information in the identity structure (e.g., accepting that one is gay) and to the adjustment that takes places for it to become part of the structure (e.g., accepting that one is gay and therefore questioning one's membership in one's religious group). The evaluation process confers meaning and value upon the contents of identity (e.g., viewing one's sexuality positively).

These processes function to create specific desirable end-states for identity, namely, 'identity principles.' Identity processes are guided by the following principles: (1) continuity across time and situation (continuity), (2) uniqueness or distinctiveness from others (distinctiveness), (3) competence and control (self-efficacy), (4) personal and social worth (self-esteem), (5) inclusion and acceptance in groups (belonging), (6) significance and purpose (meaning), and (7) compatibility and coherence between interconnected identities (psychological coherence). Only the first four principles are consensually accepted in identity process theory – the latter three have been proposed by other researchers (Jaspal and Breakwell, 2014). The theory suggests that when identity processes cannot, for whatever reason, comply with psychologically salient principles, identity is threatened and the individual will engage in strategies to cope with the threat. A coping strategy is defined as "any activity, in thought or deed, which has as its goal the removal or modification of a threat to identity" (Breakwell, 1986: p. 78).

Coping functions at three levels: (1) intrapsychic, (2) interpersonal, and (3) intergroup. Intrapsychic strategies include inter alia denial and reconceptualization. Interpersonal strategies are those that rely on "changing relationships with others in order to cope with threat" such as self-isolation (Breakwell, 1986: p. 109). Intergroup strategies make use of intergroup dynamics in order to salvage identity and include making strategic use of multiple group memberships and actively seeking group support through social networks or consciousness raising/self-help groups.

Social Representations Theory

Social representations theory explains the development and dissemination of social knowledge. A social representation is defined as a system of values, ideas, and practices regarding a given social object, which facilitates communication and behavior (Moscovici, 1988). Social representations are context specific and create what one might call a shared social reality in which thinking and discussion regarding issues like sexuality can take place.

Moscovici (1988) outlines two processes that give rise to social representations, namely, (1) anchoring and (2) objectification. Anchoring refers to the process of making something unfamiliar understandable by linking it to something familiar. For instance, Jaspal and Siraj (2011) have found that nonheterosexual British Pakistani men may anchor homosexuality to 'decadent British culture,' which can provide an explanation for their sexual orientation. Objectification is the process whereby unfamiliar and abstract objects are transformed into concrete and 'objective' commonsense realities. The use of metaphor is one example of objectification.

Identity process theory and social representations theory share some common ground (Breakwell, 2014). Identity process theory acknowledges the importance of social representations in determining identity content, and the salience of identity principles in a given social context. Conversely, social representations may be embraced or resisted by individuals in order to provide feelings of self-esteem, continuity, and so on.

Identity and Sexuality: Managing Social Representations

Nonheterosexual British Asian men can experience identity threat as a result of exposure to negative and stigmatizing social representations of their sexuality, which in turn inform their own perceptions.

Sources of Social Representation

Nonheterosexual British Asian men constitute a diverse group. It has been argued that British Pakistanis tend to lack social capital, are less 'integrated' in British society and more ethnically and religiously 'insular' (Robinson, 2009). Conversely, British Indians are said to be relatively integrated in British society and to have access to a range of social group memberships – national and ethnic (Jaspal, 2011). Therefore, British Indians may be exposed to a wider range of social representations associated with these memberships. While the ethnoreligious in-group disseminates negative, stigmatizing representations of homosexuality, the sexual in-group (i.e., other nonheterosexual individuals) may be a source of positive representations. Conversely, nonheterosexual British Pakistani men may be more systematically exposed to negative and stigmatizing accounts of homosexuality, which are associated with their primary (ethnoreligious) group membership.

The availability of a variety of representations (associated with distinct group memberships) potentially provides individuals with opportunities to make strategic decisions about how they 'use' these representations. They may accept those representations that have favorable outcomes for identity processes while rejecting those leading to threat (Breakwell, 2014).

Sexuality: Behavior or Identity?

In general, being gay is defined in the West not only in terms of sexual behavior but also in terms of desire, lifestyle, culture, gender, politics, and identity. In short, being gay is frequently construed as an aspect of who one is. Conversely, in South Asian (and many other non-Western) cultures, men may indeed engage in homosexual acts but eschew self-categorization as gay – thus, it is not regarded as an aspect of one's identity (Murray and Roscoe, 1997).

Some nonheterosexual British Asian men may themselves eschew sexual self-categorization by describing their sexuality in behavioral terms, rather than as an identity (i.e., 'doing gay things rather than being gay'). This can have psychological advantages in the short term at least. The first concerns the perceived scope for changing one's sexual orientation. While the individual may view their identity as being static and resistant to change, they usually regard behavior as mutable and transient. Accordingly, some individuals may come to view their sexual orientation as a 'bad habit' rather than an identity (Jaspal and Cinnirella, 2010). The second advantage concerns the perceived 'origins' of homosexuality. Individuals who perceive their sexual orientation as a behavior may attribute their 'gay behavior' externally to their social surroundings. In one study, British Pakistani gay participants attributed their sexual orientation to the presence of 'gay culture' in 'British culture' (Jaspal and Cinnirella, 2010). The assertion was that something inherent to the sociocultural context of Britain had induced gay behavior.

Encountering Nonheterosexuality

Many nonheterosexual British Asian men believe that their ethnoreligious communities regard homosexuality as 'wrong' and, in some cases, 'sinful' (Yip, 2004). This perception seems to be reinforced by the negative, stigmatizing terms that are commonly used to denote homosexuality or gay people in individuals' heritage languages (Jaspal and Siraj, 2011). These terms are often derogatory and make reference to anal intercourse. Thus, sexualized terms, which are employed primarily as abusive remarks, can come to dominate thinking among nonheterosexual people, thereby encouraging a negative evaluation of homosexuality.

Research shows that individuals draw upon their previous encounters with the topic of homosexuality (involving their ethnoreligious in-group) in order to gauge potential reactions to their sexual orientation. In April 2009, the long-running popular British television soap opera *Eastenders* introduced a controversial storyline concerning a British Pakistani gay man, Syed Massoud. The storyline culminated in a dramatic and explosive public 'outing' of the character who was subsequently ostracized from both his family and ethnoreligious circles. Themes of family breakdown, depression, verbal and physical abuse, and suicide dominated Syed's coming out story. For some nonheterosexual British Asians, this storyline provided insight into what sexual self-disclosure might be like (Jaspal, 2010). Identification with the character and his plight appeared to dissuade individuals from ever coming out themselves. For some individuals, the storyline seemed to confirm and reinforce the perception that the ethnoreligious in-group is homophobic and will never accept homosexuality. This could create feelings of fear and anxiety about the future and prospect of coming out (Jaspal and Siraj, 2011).

There is an acute awareness of social representations of what it means to be a 'genuine' ethnoreligious in-group member, such as the importance of 'izzat' (honor) within family and ethnoreligious circles (Bradby, 2006). An essential tenet of maintaining family 'izzat' is the fulfillment of the ethnocultural expectation of marriage. However, it has been found that the expectation of marriage can be construed as a threat by nonheterosexual British Asians (Jaspal, 2012b). While the prospect of marriage may be avoided and suppressed in early adulthood, the completion of one's education seems to serve as a cultural cue for the prospect of marriage to reemerge. Conversely, a contravention of social representations concerning sexuality and marriage may be regarded as a threat to 'izzat' and lead to negative social and psychological outcomes. Individuals have reported fear of physical abuse from ethnic in-group members (and sometimes from family members). These fears are grounded in knowledge of 'honor crimes,' that is, psychological and/or physical abuse in the name of family honor.

Identity: Threat and Coping

Nonheterosexual British Asians may face 'hyperthreats,' that is, threats to several principles of identity. The following principles of identity seem particularly susceptible: (1) self-esteem, (2) belonging, (3) distinctiveness, (4) psychological coherence, and (5) continuity.

Self-Esteem

Homophobia and self-esteem are negatively correlated for gay men (Dupras, 1994). Individuals express acute awareness of the negative, stigmatizing social representations of homosexuality, which construct it as an abnormality or deliberate subversion, although the impact of this varies by cultural group. As observed above, many British Indian men seem to be able to overcome the social stigma associated with homosexuality through self-exposure to positive representations. In this case, self-esteem may be less susceptible to threat – individuals may in fact come to view their sexuality positively. However, in ethnoreligious contexts, individuals may be reminded that homosexuality is culturally unacceptable, due to the overt social stigma appended to this identity element. Furthermore, even among those men who have disclosed their sexuality to others, there may be a 'silencing' of homosexuality. Both overt condemnation of homosexuality and the more covert act of 'silencing' it can make it difficult for individuals to derive a positive self-conception on the basis of their sexuality – it is represented as something to be ashamed of. It threatens self-esteem only in those contexts in which social stigma acquires salience. Conversely, many nonheterosexual British Pakistani men, who tend to have less social capital, may themselves continue to evaluate their gay identity negatively. This negative evaluation of homosexuality at the psychological level can lead to a persistent threat to self-esteem.

Belonging

The perceived possibility of otherization and ostracization from their family and ethnoreligious circles – two valued group memberships – renders the belonging principle susceptible to threat. Exposure to negative representations can engender the belief that individuals will be excluded from relevant in-groups. Identity process theory argues that individuals will make use of their multiple social group memberships in order to safeguard identity. Accordingly, British Indians who decide to disclose their sexual identity to parents, family, and friends may attempt to derive feelings of acceptance and inclusion from alternative group memberships – by redefining themselves as 'primarily British' or as 'a gay man' (Jaspal, 2012b). In other words, their alternative group memberships (in this case, national and sexual, respectively) can acquire salience, becoming alternative sources of belonging when the ethnic/religious groups no longer satisfy the principle. However, this may be less of an option for British Pakistanis, many of whom have less social capital and may not have access to alternative group memberships.

Distinctiveness

Although a sexual minority identity is clearly one possible means of deriving feelings of distinctiveness from relevant others, it is likely that neither the level of distinctiveness nor the balance between distinctiveness and belonging will be satisfactory. Indeed, 'excessive' distinctiveness can be threatening for identity, since this can result in ostracization and exclusion. Moreover, distinctiveness that threatens self-esteem is unlikely to be beneficial for identity.

Distinctiveness can be threatened at another level. Nonheterosexual British Asian men generally express awareness of negative representations of homosexuality within their ethnoreligious in-group. In previous research, individuals have expressed fear of being 'homogenized' with 'typical' gay people with whom many perceive no affinity, solidarity, or 'oneness' (Jaspal and Cinnirella, 2010). There is evidence that many nonheterosexual British Asians have themselves internalized negative, stigmatizing representations of gay people – viewing them as 'camp,' 'flamboyant,' and 'too open' – from which they wish to distance themselves. The perceived possibility of being homogenized with 'typical' gay people may also result in threats to the distinctiveness principle. One means of protecting distinctiveness may be to avoid ever coming out (Jaspal and Siraj, 2011).

Psychological Coherence

In much recent research, the crux of the debate has been the extent to which sexual and ethnoreligious identities are subjectively perceived as compatible and coherent. Among British Pakistanis, the psychological coherence principle seems to be chronically threatened as a result of being gay and Muslim. Many individuals view the norms, values, and practices perceived to be associated with either identity as being in conflict. Among British Pakistani gay men, the perception that Muslim and gay identities are incompatible and incoherent can cause individuals to question their Muslimness, and to experience negative emotions such as shame, guilt, fear, and anger (Jaspal, 2012a). While British Pakistani gay men generally evaluate their Muslim identity positively and construct this identity as 'core,' they tend to evaluate their sexual identity negatively. Although the threat is often chronic, this asymmetry in identity evaluation can help to safeguard psychological coherence.

Conversely, there is little evidence that British Indian gay men experience threats to the psychological coherence principle. Unlike Islam, neither Sikhism nor Hinduism is particularly vocal on homosexuality. Thus, sexual and religious identities tend not to be 'interconnected' in the minds of individuals – they are simply not viewed as being relevant to one another (Amiot and Jaspal, 2014). For British Indians, psychological coherence may well be threatened during the early stages of sexual identity development, but does not seem to persist in the later stages of identity development. British Indians seem more able than British Pakistanis to reconstrue what it means to be gay and Asian – ethnic and sexual identities can cease to be 'interconnected' and, thus, psychological coherence becomes a nonissue.

Continuity

While psychological coherence is concerned with compatibility of identity elements at any given point in time, continuity reflects consistency in identity over time (Amiot and Jaspal, 2014). Maintaining a sense of temporal continuity is important for identity. The assimilation of gay identity in the identity structure (i.e., accepting that one is gay) reflects a change in self-perception and self-definition. Similarly, its accommodation in the identity structure (i.e., the changes that are made in order for it to become part of the structure) can necessitate a change in self-perception in that one may need to rethink the importance of other identity elements. For instance, upon coming out as gay, a British Asian man may attempt to accentuate his masculinity to counteract dominant representations concerning gender and sexuality. The introduction of this novel information in the identity structure (i.e., masculinity) can, however, jeopardize the psychological thread that connects past, present, and future.

Indeed, participants in previous research have explicated threats to continuity. In one case, a young British Indian participant described how his family and friends began to treat him in a more feminine manner when he came out as gay. He reported feeling the need to demonstrate his masculinity by engaging in 'masculine activities' such as boxing and playing football. These changes in self-definition and behavior were viewed as necessary for enhancing self-esteem and distinctiveness (from 'typical' gay people), but were threatening for continuity. He had never viewed himself in this way before but felt compelled to do so.

In addition to affecting identity processes, nonheterosexuality can impact the formation and management of relationships.

Forming and Managing Relationships

Individuals manage their family and friendship networks in ways that benefit identity.

Family Relations

One significant barrier to coming out is fear of rejection from family members. For many British Pakistanis, non-heterosexuality is threatening at a psychological level, since there is a perception that is 'wrong' and 'sinful,' and at a social level, given that individuals may anticipate a negative response from others. Although British Indians may cease to experience threats at a psychological level later in the life span, many decide not to come out due to the fear of disownment (Jaspal, 2012b).

Some British Pakistanis may regard negative responses from family and ethnoreligious in-group members as legitimate due to the perceived 'sinfulness' of homosexuality. There may be an aspiration to 'turn' straight and sometimes even an endorsement of homophobia. Individuals may not have had any exposure to positive representations that construct homosexuality as an acceptable alternative to heterosexuality and, thus, there is a reliance on heteronormative social representations. Clearly, this is dependent partly on the interpersonal networks to which an individual has access, since group memberships are sources of social representations. Conversely, some British Indians, and particularly those who have had exposure to more positive representations of homosexuality, manifest a more defensive response to perceived homophobia from others. For instance, individuals may find it unacceptable that their family members are unable to see 'beyond' their sexual orientation, leading to negative interpersonal relations with their family and ethnocultural in-group members.

Marriage has emerged as a particularly salient issue in research with nonheterosexual British Asians (Jaspal, 2012b). Many British Asian families continue to observe the norm of arranged marriage. Heterosexual marriage is a religiocultural rite that Asian families eagerly await and, despite the growing acceptance of gay marriage in the West, among South Asians this is not generally viewed as a legitimate alternative. In early adulthood, nonheterosexual men may completely avoid the topic of (heterosexual) marriage by attenuating thoughts about it and by evading it as a topic of conversation Jaspal (2013). They may defer marriage by attaching priority to education, career development, financial stability, and the purchase of property – these are 'legitimate' activities that can delay marriage. However, individuals may ultimately reach a point at which it becomes increasingly difficult to deflect the topic of marriage and they must take a stance on it.

In some cases, this can present a psychological dilemma – in the short term, one might acquiesce in their family members' expectation that they be married but, in the long term, one's nonheterosexual orientation could mean that the marriage is not successful. This can potentially lead to a breakdown in family relations. Conversely, one might reject the established norm of (arranged) marriage, which could similarly cause a breakdown in family relations. Indeed, individuals have reported the perceived necessity of 'distancing' themselves from their families in order to avoid marriage without having to disclose their sexual orientation. The idea is that one sacrifices family relations in order to avoid both marriage and the necessity of coming out. A breakdown in valued interpersonal relations can be threatening for belonging and continuity, in particular.

Friendship Networks

The development of friendship networks among non-heterosexual British Asians often depends upon their level of involvement in gay affirmative social contexts or 'gay space' (e.g., nightclubs, bars, social groups). These contexts can provide social capital and, thus, "an environment for the establishment of resources, networks, relationships, norms, meanings and trust that enable smooth engagement among members" (Yip, 2007: p. 86). Gay space is used to varying degrees. Jaspal and Cinnirella (2012) have argued that nonheterosexual British Pakistani men may avoid frequenting gay space in order to avoid having to acknowledge their sexuality and avoid being outed. Conversely, nonheterosexual British Indians seem to be more willing to frequent gay space and indeed some of the lesbian, gay, bisexual, and transgender-themed Asian nightclub events have drawn large numbers of nonheterosexual Indian men and women, in particular (Bassi, 2008).

Participation in gay space can potentially have favorable outcomes for the development of interpersonal relations and ultimately for identity processes. This can allow temporary or long-term sociopsychological 'shielding' from the negative and stigmatizing social representations of homosexuality. Through the dissemination and exchange of positive representations in these spaces, individuals may come to perceive their sexual orientation as a facilitator of a positive group membership, rather than as a personal trait (cf Jaspal and Cinnirella, 2010).

Yet, gay space can expose nonheterosexual British Asian men to stigma on the basis of their ethnic and/or religious background. In their study on gay Muslims in the United States, Minwalla et al. (2005: p. 122) highlight the potential importance of 'race' within 'social dynamics within white gay culture.' Emerging empirical research suggests that nonheterosexual British Asian men may be excluded from gay space, either implicitly or explicitly, due to racism, or forced to become 'culturally white,' that is, to abandon elements of their ethno-religious identities in order to 'fit in' (Bassi, 2008). This may decrease willingness to partake in gay space to develop interpersonal relations with other nonheterosexual individuals.

Despite these potential obstacles, many men do indeed participate in gay space. There has been some research into the interpersonal dynamics in these settings (Jaspal and Cinnirella, 2012). Although intergroup relations between Indians and Pakistanis can be negative in many social contexts, gay space is one context in which points of tension and conflict with non-heterosexual British Asians can become latent. There is evidence of a superordinate ethnosexual group identity around being a 'gay Asian,' which cohesively brings together non-heterosexual British Pakistanis and Indians. Key to this 'gay identity' is a common understanding of the social and familial constraints (e.g., honor and arranged marriage) faced by non-heterosexual British Asians.

Conversely, individuals may avoid interpersonal relations with white British gay men since there are perceptions that they (1) are 'excessively' open about their sexual orientation and expect nonheterosexual people to be open about their orientation, (2) attach priority to relationships, want to 'settle down,' and therefore categorize themselves as gay, and (3) fail to understand the social and familial constraints faced by nonheterosexual British Asians. Recent research indicates that

interpersonal encounters with white British gay men can result in threats to self-efficacy, since individuals feel that they have less control over their relationships and lives (e.g., possibility of being outed by a culturally unaware white person).

On the other hand, for some British Asians, contact with white British gay men can in fact serve as an 'identity enhancer' (Jaspal and Cinnirella, 2012). Sustained interpersonal relations with white British gay men can provide individuals with the necessary social capital to reject negative representations of homosexuality, which are associated with their heteronormative ethnoreligious cultures. The cultural necessity of heterosexual marriage in early adulthood is one such example.

Positive social representations from their new alternative interpersonal networks can provide a more psychologically satisfying 'replacement' of negative representations regarding their sexual orientation. Moreover, these networks can allow individuals to perceive their sexual orientation as 'normal' and, conversely, the negative reactions of their cultures and families as 'abnormal.' In short, self-alignment with white British gay men can have socially and psychologically empowering outcomes, resulting in the assimilation–accommodation and positive evaluation of their sexual orientation.

Conclusion

Nonheterosexual British Asian men can face social and psychological challenges to identity and well-being as a result of their sexual orientation. This article explores meaning-making vis-à-vis sexuality at the individual level, dominant social representations of homosexuality, how these representations might impinge upon identity, and how individuals, accordingly, manage social relations. This study focuses exclusively upon nonheterosexual men, and it is difficult to generalize these research findings to nonheterosexual British Asian women given the dearth of relevant research. It is, however, noteworthy that in most South Asian ethnoreligious communities, nonheterosexuality is viewed as contravening masculinity norms, in particular, and the act of 'sodomy' (anal intercourse between men) is viewed as particularly morally revulsive. Thus, there may be less acknowledgment of lesbianism and bisexualism among women and more scope for nonheterosexual British Asian women to remain socially 'hidden.' There is clearly a need for research into the identities and experiences of nonheterosexual British Asian women.

Existing research suggests that negative representations of homosexuality seem to be prevalent in family and ethnocultural/ethnoreligious contexts. In the long term, it is necessary to attempt to create a shift in social representations so that these communities adopt a more accepting stance on homosexuality. Drawing on social representations theory, this will necessitate positive forms of anchoring and objectification in relation to homosexuality. This may in turn encourage individuals to assimilate–accommodate and positively evaluate their nonheterosexuality. In the short term, it would be beneficial for nonheterosexual British Asians to gain some exposure to positive social representations of homosexuality, which might in turn facilitate the construction of a positive sexual identity. The task of facilitating positive identification among this population will need to be undertaken by a variety of actors – researchers, policy makers, educators, and community leaders alike.

See also: Ethnic Identity, Psychology of; Friendship Experiences of Lesbian, Gay, Bisexual, and Transgender Individuals; Gay, Lesbian, and Bisexual Youth; Heterosexism and Homophobia; Identity and Identification, Social Psychology of; Integration and Multiculturalism: Focus on Western Europe; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Same-Sex Sexualities in Turkey; Sexual Minorities and Sexual Citizenship; Sexual Minorities in the Maghreb; Sexual Orientation: Historical and Social Construction.

Bibliography

Amiot, C.E., Jaspal, R., 2013. Identity integration, psychological coherence and identity threat: linking identity process theory and notions of integration. In: Jaspal, R., Breakwell, G.M. (Eds.), Identity Process Theory: Identity, Social Action and Social Change. Cambridge University Press, Cambridge, pp. 155–174.

Bassi, C., 2008. The precarious and contradictory moments of existence for an emergent British Asian gay culture. In: Dwyer, C., Bressey, C. (Eds.), New Geographies of Race and Racism. Ashgate, Farnham, pp. 209–222.

Bradby, H., 2006. Understanding honour and religion as resource and constraint for young British Asians. In: Beckford, J.A., Walliss, J. (Eds.), Theorising Religion. Ashgate, Aldershot.

Breakwell, G.M., 1986. Coping with Threatened Identities. Methuen, London.

Breakwell, G.M., 2014. Identity and social representations. In: Jaspal, R., Breakwell, G.M. (Eds.), Identity Process Theory: Identity, Social Action and Social Change. Cambridge University Press, Cambridge, pp. 118–134.

Dupras, A., 1994. Internalized homophobia and psychosexual adjustment among gay men. Psychological Reports 75, 23–28.

Ghuman, P.A.S., 2003. Double Loyalties: South Asian Adolescents in the West. University of Wales Press, Cardiff.

Jaspal, R., 2010. Identity threat among British Muslim gay men. The Psychologist 23, 640–641.

Jaspal, R. 2011. The Construction and Management of National and Ethnic Identities among British South Asians: An Identity Process Theory Approach (Ph.D. dissertation). Royal Holloway, University of London, UK.

Jaspal, R., 2012a. Coping with religious and cultural homophobia: emotion and narratives of identity threat from British Muslim gay men. In: Nynäs, P., Yip, A.K.T. (Eds.), Religion, Gender and Sexuality in Everyday Life. Ashgate, Farnham, pp. 71–90.

Jaspal, R., 2012b. "I never faced up to being gay": sexual, religious and ethnic identities among British South Asian gay men. Culture, Health and Sexuality 14, 767–780.

Jaspal, R., 2014. Arranged marriage, identity and psychological wellbeing among British Asian gay men. Journal of GLBT Family Studies. http://dx.doi.org/10.1080/1550428X.2013.846105.

Jaspal, R., Breakwell, G.M., 2014. Identity Process Theory: Identity, Social Action and Social Change. Cambridge University Press, Cambridge.

Jaspal, R., Cinnirella, M., 2010. Coping with potentially incompatible identities: accounts of religious, ethnic and sexual identities from British Pakistani men who identify as Muslim and gay. British Journal of Social Psychology 49, 849–870.

Jaspal, R., Cinnirella, M., 2012. Identity processes, threat and interpersonal relations: accounts from British Muslim gay men. Journal of Homosexuality 59 (2), 215–240.

Jaspal, R., Siraj, A., 2011. Perceptions of 'coming out' among British Muslim gay men. Psychology and Sexuality 2, 183–197.

Minwalla, O., Rosser, B.R.S., Feldman, J., Varga, C., 2005. Identity experience among progressive gay Muslims in North America: a qualitative study. Culture, Health and Sexuality 7, 113–128.

Moscovici, S., 1988. Notes towards a description of social representations. European Journal of Social Psychology 18, 211–250.

Murray, S.O., Roscoe, W., 1997. Islamic Homosexualities: Culture, History, and Literature. New York University Press, New York.

Robinson, L., 2009. Cultural identity and acculturation preferences among South Asian adolescents in Britain: an exploratory study. Children and Society 23, 442–454.

Seidman, S., 2003. The Social Construction of Sexuality. Norton, New York.

Yip, A.K.T., 2004. Embracing Allah and sexuality? South Asian non-heterosexual Muslims in Britain. In: Jacobsen, K.A. (Ed.), South Asians in the Diaspora: Histories and Religious Traditions. Brill, Leiden, pp. 294–310.

Yip, A.K.T., 2007. Sexual orientation discrimination in religious communities. In: Badgett, L., Frank, J. (Eds.), Sexual Orientation Discrimination: An International Perspective. Routledge, London, pp. 209–244.

Nonparametric Regression

Martin L Hazelton, Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Nonparametric regression is a methodology for describing the trend between a response variable and one or more predictors. This approach differs from classical regression models in that it does not rely on strong assumptions regarding the shape of the relationship between the variables. Rather, the data are allowed to speak for themselves in determining the form of the fitted regression functions. Approaches include kernel smoothing and local polynomial regression, spline-based regression models, and regression trees and related techniques. Additive models are typically employed when there are several predictors to mitigate problems arising from the curse of dimensionality.

Introduction

Suppose that one is confronted with a data set comprising observations on multiple variables. An initial exploration of this data set would typically include visualization and computation of numerical summaries for individual variables. However, often it is the relationships between the variables that are of greater interest. Scatterplots of pairs of variables can provide some insight: it is often helpful to produce a matrix of all such possible plots, sometimes referred to as a 'pairs plot.'

As an illustration, we consider data from a UNICEF report from 2003 (UNICEF, 2003), describing childhood mortality and associated indicators by country. The variables recorded (where available) are listed in Table 1. A pairs plot of the data appears in Figure 1. There are some obvious pairwise relationships between variables. In some cases these appear approximately linear; for example, between **Under.5.MR** and **Infant.MR**. In other cases the forms of the relationships are more complicated. Consider the plots of **Infant.MR** against **GNI.per.cap**, or **Under.5.MR** against **Hi.20.pc** for instance.

Interpretation of scatterplots can be aided by adding smooth curves that describe the trend between the variables. Consider for example the plots in Figure 2. The left-hand panel contains data on **Infant.MR** against **GNI.per.cap**, while the right-hand panel shows the logarithm of **Under.5.MR** against **Hi.20.pc**. In both cases 'scatterplot smoothers' have been employed to help draw out the form of the relationships between the variables. In the left-hand panel the rapid rise in under-5 mortality as the share of income enjoyed by the richest 20% increases from 35 to 45% is quite striking, and is a feature of the data that arguably could have been missed without the scatterplot smoothing.

Table 1 Variables for the 2003 UNICEF data set

Variable	*Description*
Under.5.MR	Mortality rate for children under 5 years of age (per 100 000)
Infant.MR	Mortality rate for children under 1 year of age (per 100 000)
Total.pop	Total population (1000s)
Births	Annual number of births (1000s)
GNI.per.cap	Gross national income per capita (USD)
Life.exp	Life expectancy at birth (years)
Adult.lit	Adult literacy rate (%)
School	Attendance rate at primary school (%)
Low.40.pc	Percentage share of household income held by the poorest 40%
Hi.20.pc	Percentage share of household income held by the richest 20%

The trend curves displayed in Figure 2 were obtained using a type of nonparametric regression. Regression models in general seek to describe the variation in the mean (of expected value) of one variable, the response, as a function of one or more predictor (or explanatory) variables. In traditional parametric regression models, this relationship is assumed to be linear (or possibly to be quadratic or some other low order polynomial). However, in nonparametric regression we make almost no assumptions about the shape of the relationship, letting the data 'speak for themselves' in determining the estimated trend. While the plots in Figure 2 demonstrate the application of kernel regression to bivariate data (i.e., response and a single predictor variable), this type of methodology can also be employed with higher dimensional data, as discussed later.

An important use of nonparametric regression is as a tool for data exploration and visualization, where the aim is to gain a preliminary feel for the data before conducting any formal model fitting. Nonparametric regression can be employed for a number of other tasks, including prediction and hypothesis testing (e.g., Azzalini and Bowman, 1993). Nonparametric regression can also be combined with parametric models to form hybrid semiparametric models; see for example Ruppert et al. (2003).

Nonparametric Regression with a Single Predictor

We first look at the situation where we have bivariate observations (x_i, y_i) for $i = 1, \ldots, n$. A very general kind of regression model describing the relationship between these variables can be written as

$$y_i = m(x_i) + \varepsilon_i \quad (i = 1, \ldots, n) \qquad [1]$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are error terms that are assumed to have zero mean, so that $m(x_i) = \mathrm{E}[y_i|x_i]$. The regression function, m, therefore describes how the conditional mean of y depends on x.

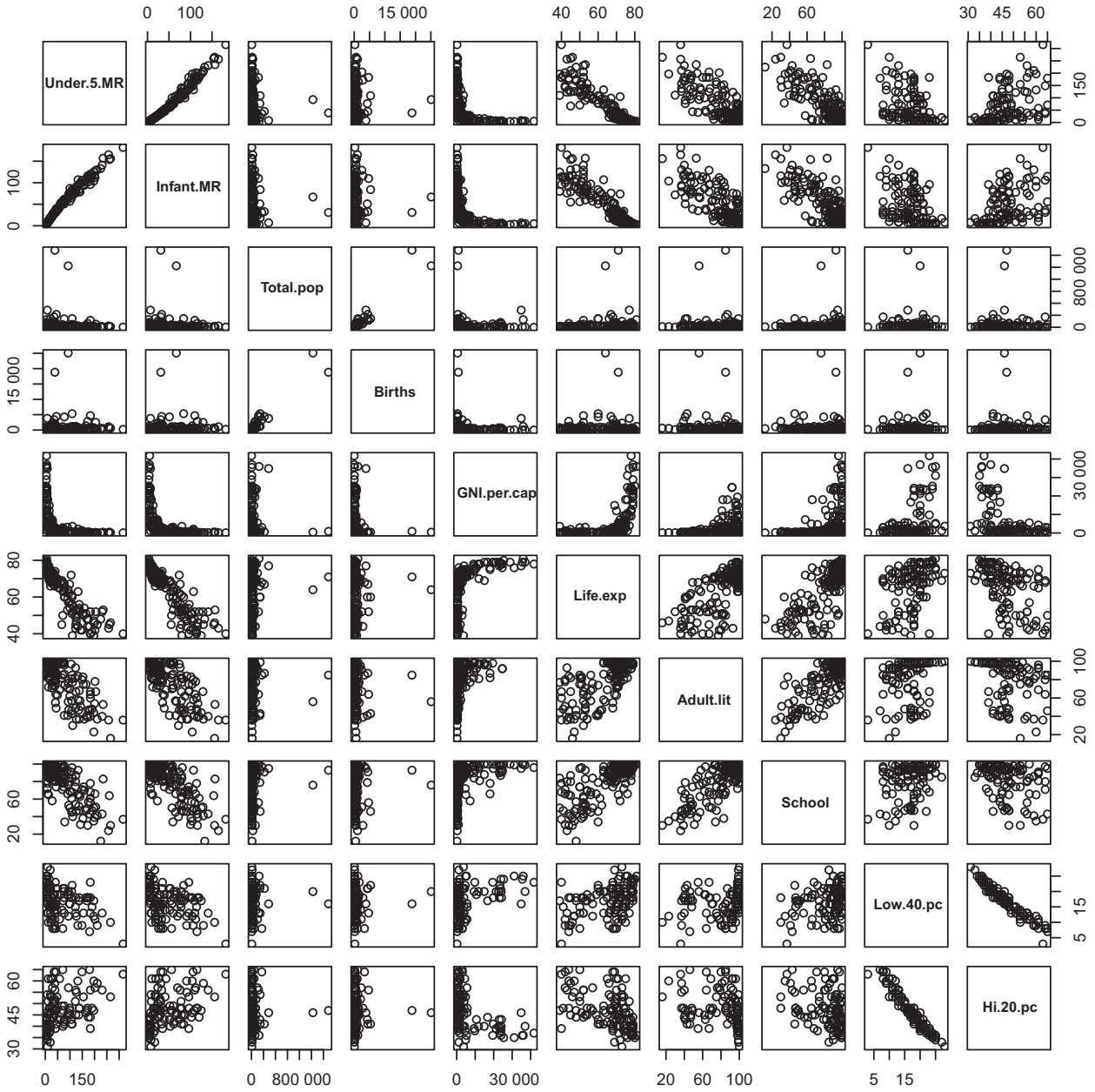

Figure 1 Pairs plot of the UNICEF data.

We can obtain the classical linear regression model from eqn [1] by setting $m(x) = \beta_0 + \beta_1 x$. In that case m is described in terms of just two parameters (the coefficients β_0, β_1) and so the regression model is parametric. In nonparametric regression, we do not specify any particular (simple) parametric form for m, we merely assume that it is some (typically smooth) function of x.

Nonparametric regression is characterized by 'local estimation' of the regression function. To understand what this means, consider estimation of m at a particular value x of the predictor variable. In linear regression we would form the estimate $\widehat{m}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$ where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are parameter estimates (usually obtained by the method of least squares in this context). Now, these parameter estimates are functions of all the observations. It follows that each observed data point (x_i, y_i) has an impact on the estimate $\widehat{m}(x)$, no matter how close or distant x_i is to the estimation point x. Classical linear regression is therefore based on 'global estimation.' In contrast, when estimating $m(x)$ locally, information is used only from data points with predictor values relatively close to x.

Kernel Smoothing

Perhaps the most straightforward method of nonparametric regression is to estimate $m(x)=\mathrm{E}[y|x]$ using a local average. Specifically, we can estimate the regression function at any

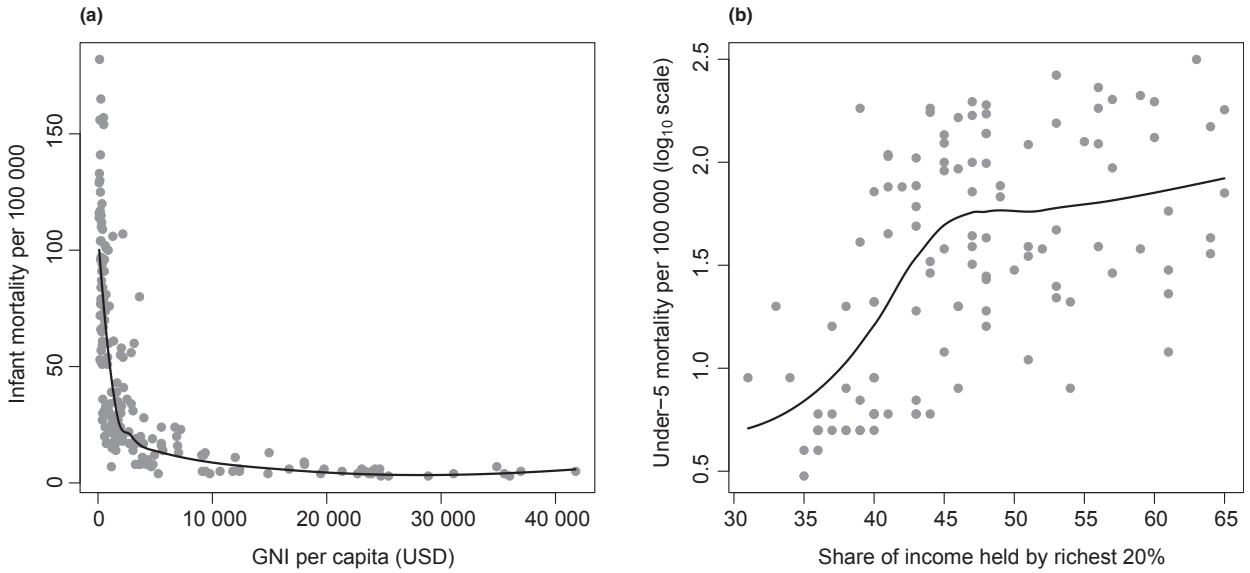

Figure 2 Plot of infant mortality against per capita GNI (left panel (a)) and under-5 mortality against share of household income held by the richest 20% (right panel (b)). In both cases nonparametric regressions have been added to help draw out the form of the relationships between the variables.

given point x by computing the (local) average of all data within some window centered on x. By sliding this window along the x-axis we will generate an estimate of the entire regression function (i.e., obtain estimates of $m(x)$ for all estimation points). We illustrate this idea for a regression of log (**Under.5.MR**) against **Hi.20.pc** in Figure 3, where the shaded regions in the left-hand panel represent instances of the estimation window. The resultant regression curve (shown right-hand panel) tracks the data quite well, but has an unattractively uneven appearance. Its blocklike form arises from sudden changes in the local mean that occur when a data point enters or leaves the window.

This problem can be addressed by replacing the local average by a weighted local average, where the weight for response y_i decreases smoothly according to how distant the predictor x_i is from the estimation point x. A convenient way of

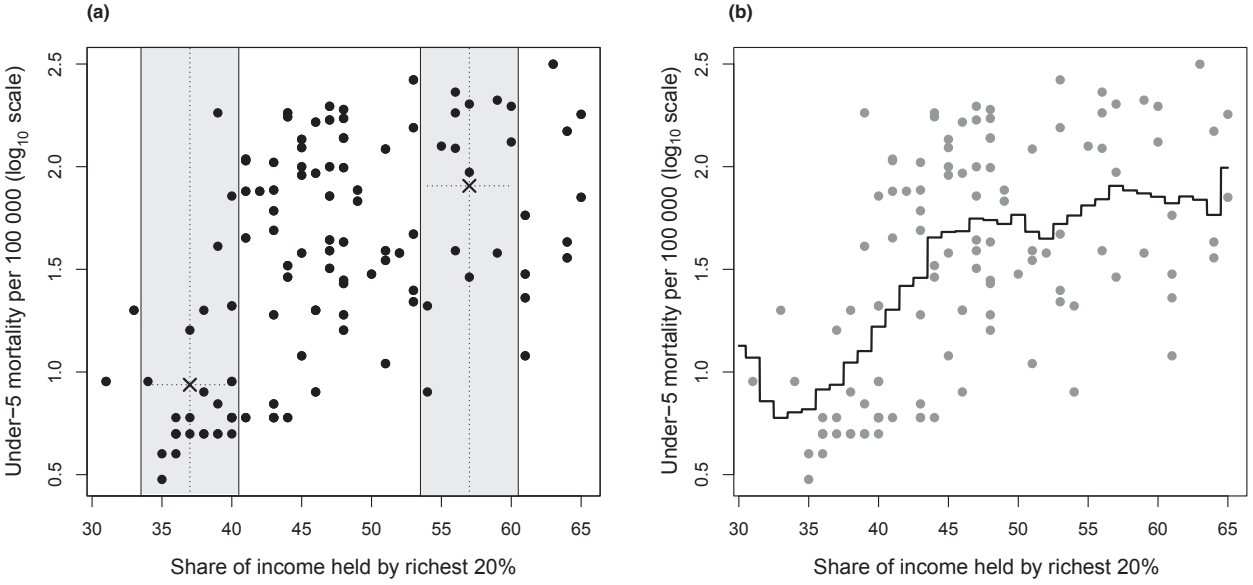

Figure 3 Illustration of local averaging to produce a regression of the logarithm of under-5 mortality against share of household income held by the richest 20%. In the left-hand panel (a) we show how estimates at $x = 37$ and $x = 57$ (located at the vertical dotted lines) are formed. In each case the estimated regression function, indicated by a cross, is obtained by computing the mean y-value for all data points falling in the surrounding shaded area. By computing such local averages for all values of x we obtained the estimated regression function depicted in the right-hand panel (b).

defining the weights is through a 'kernel function.' A kernel function K typically takes nonnegative values, and is unimodular and symmetric about zero. Examples include the Epanechnikov kernel defined by $K(x) = \frac{3}{4}(1-x^2)1_{|x|<1}$; the biweight kernel given by $K(x) = \frac{15}{16}(1-x^2)^2 1_{|x|<1}$; and the Gaussian kernel given by $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. These are plotted in Figure 4. In each case we can define the relative weight allocated to observation i for estimation at point x by $K(x-x_i)$.

Before we proceed, however, we need to address the question of scale. Specifically, how wide should we make the estimation window? In Figure 3, for example, we used a window of total width 7 for the local averaging. However, applying a window of the same width to the data on **Infant.MR** against **GNI.per.cap** from the left-hand panel of Figure 2 would clearly be inappropriate. For many x values, windows of that width would contain no data. A similar objection applies to using the kernels in Figure 4 without modification. To address this issue, we introduce a scaling parameter h called the 'bandwidth.' A scaled version of the basic kernel can then be defined by $K_h(x) = h^{-1}K(x/h)$.

Using the scaled kernel to defining weights, we can define a nonparametric regression estimator using local weighted averaging as follows:

$$\widehat{m}_{\mathrm{NW}}(x) = \frac{\sum_{i=1}^{n} y_i K_h(x-x_i)}{\sum_{j=1}^{n} K_h(x-x_j)} \qquad [2]$$

The denominator in eqn [2] means that the weight for the ith observation is $w_i = K_h(x-x_i)/\sum_{j=1}^{n} K_h(x-x_j)$, ensuring that $\widehat{m}_{\mathrm{NW}}(x)$ is a bona fide weighted average because $\sum_{i=1}^{n} w_i = 1$. The nonparametric regression defined by eqn [2] is often referred to as the Nadaraya–Watson estimator after its contemporaneous developers (Nadaraya, 1964; Watson, 1964), explaining the NW subscript to $\widehat{m}$. Its construction and application is illustrated in Figure 5.

The chosen value for the bandwidth h is of critical importance to the appearance of the fitted local regression. A very small bandwidth leads to weighted local averaging that involves only a very small number of data points. This results in an estimate with high variance, but low bias. The regression curve will appear undersmoothed, with multiple wiggles as it tracks noise as well as signal. A large bandwidth involves estimation over many more data points, reducing variability but introducing high bias because data points very distant to the estimation point play a significant role. The resultant regression curve will be oversmoothed, quite possibly failing to pick up interesting structure in the data. We illustrate these ideas in Figure 6.

A variety of practical methods of bandwidth selection have been proposed. The simplest is to choose h by eye, in order produce an aesthetically pleasing plot. Less subjective approaches include cross-validation and plug-in methods. The former seeks to choose the bandwidth that is best for predicting the response for observations independent from those used to estimate the regression. The latter uses asymptotic theory to derive a mathematical expression for the optimal bandwidth; that is, the bandwidth minimizing the mean squared error of $\widehat{m}_{\mathrm{NW}}$ integrated over the range of the predictor variable. Unknown quantities in the formula are replaced by pilot estimates thereof, producing a data-driven bandwidth. See Wand and Jones (1995) for details.

Local Polynomial Regression

The Nadaraya–Watson estimator can be derived as the solution to an optimization problem. Specifically, $\widehat{m}_{\mathrm{NW}}(x) = \widehat{\beta}_0$ where $\widehat{\beta}_0$ is the minimizer of the weighted sum of squares,

$$Q_0(\beta_0) = \sum_{i=1}^{n} K_h(x-x_i)(y_i-\beta_0)^2 \qquad [3]$$

A useful interpretation of this mathematics is that the Nadaraya–Watson estimate results when we fit a locally constant regression function $m(x) = \beta_0$ to the data by the method of least weighted sum of squares.

This idea can then be generalized. Suppose that we use a linear approximation of the regression function, so that at a point z close to the estimation point x we have $m(z) \approx \beta_0 + \beta_1(z-x)$. We can obtain parameter estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ as the minimizer of the weighted sum of squares,

$$Q_1(\beta_0,\beta_1) = \sum_{i=1}^{n} K_h(x-x_i)[y_i-\beta_0-\beta_1(x_i-x)]^2 \qquad [4]$$

These estimates provide the closest possible match of the approximate linear function to the true regression function. Moreover, since $m(x) \approx \beta_0 + \beta_1(x-x) = \beta_0$, it follows that

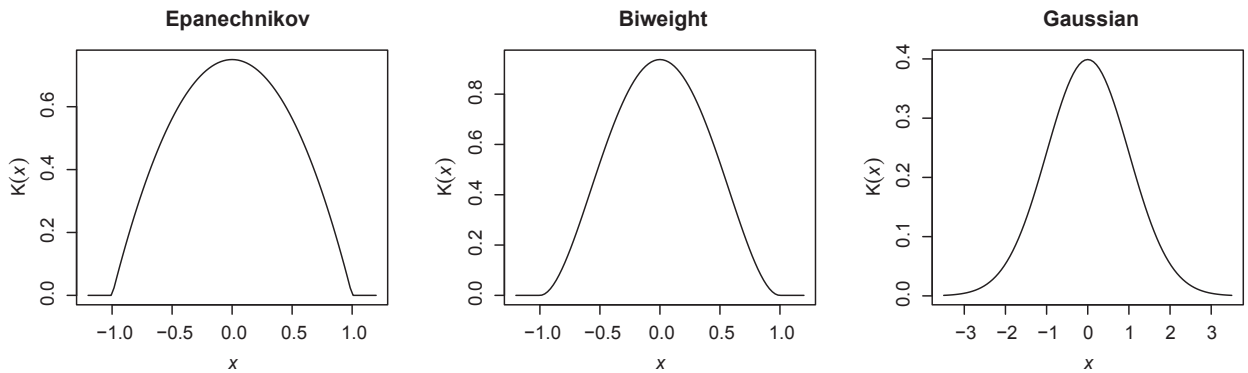

Figure 4 Three popular types of kernel function.

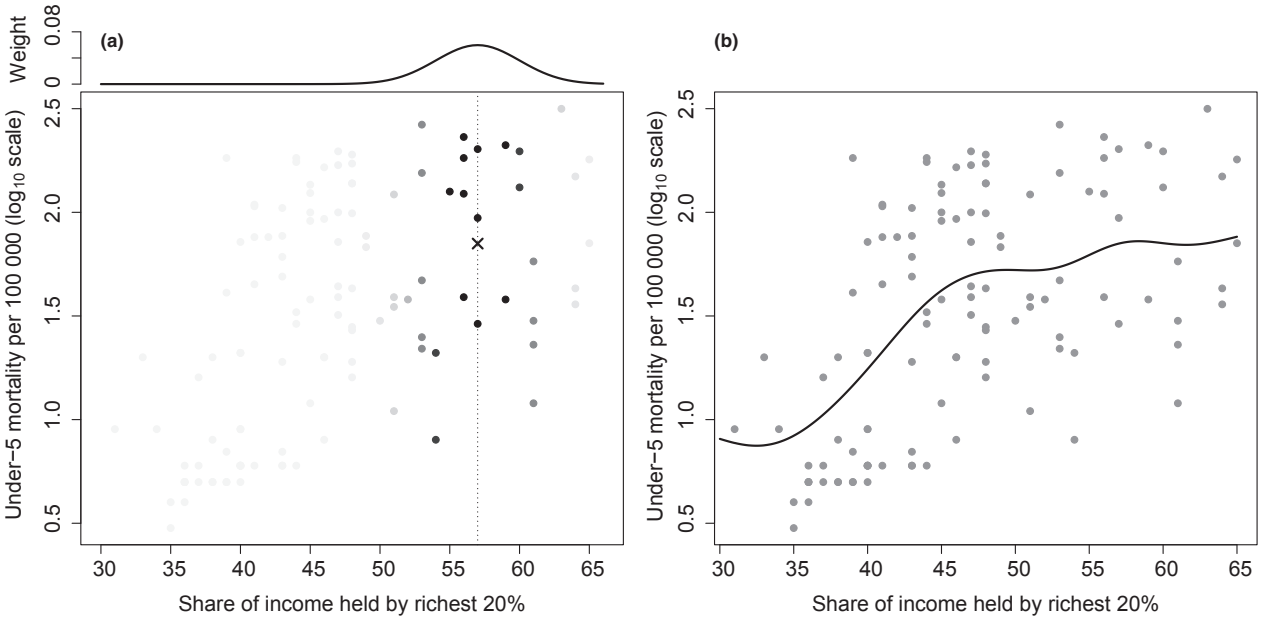

Figure 5 Nadaraya–Watson kernel regression for the logarithm of under-5 mortality against share of household income held by the richest 20%. The left-hand panel (a) shows how the estimate of the regression function is obtained at $x = 57$, with the weights derived from the Gaussian kernel function depicted at the top of the plot. The gray-scale intensity of each data point in the left-hand scatterplot reflects its weight for estimation at that point. The estimate of the regression function at $x = 57$ is marked by a cross. The right-hand panel (b) displays the complete nonparametric regression curve. A bandwidth of $h = 3$ was employed.

the value of the estimate $\widehat{\beta}_0$ the regression function at x. It is called the local linear estimator, which we will denote by $\widehat{m}_{\mathrm{LL}}(x)$. Its construction is illustrated in Figure 7. One of the advantages of local linear regression over the Nadaraya–Watson estimator is better performance near boundaries. This is illustrated in the middle- and right-hand panels of Figure 7, where the former methodology avoids the unnatural flattening near $x = 0$ that is observed in the curve using the latter approach.

We can generalize further, and seek to approximate $m(x)$ locally by a polynomial of degree p. If this is fitted by minimizing the kernel weighted sum of squares once again, then

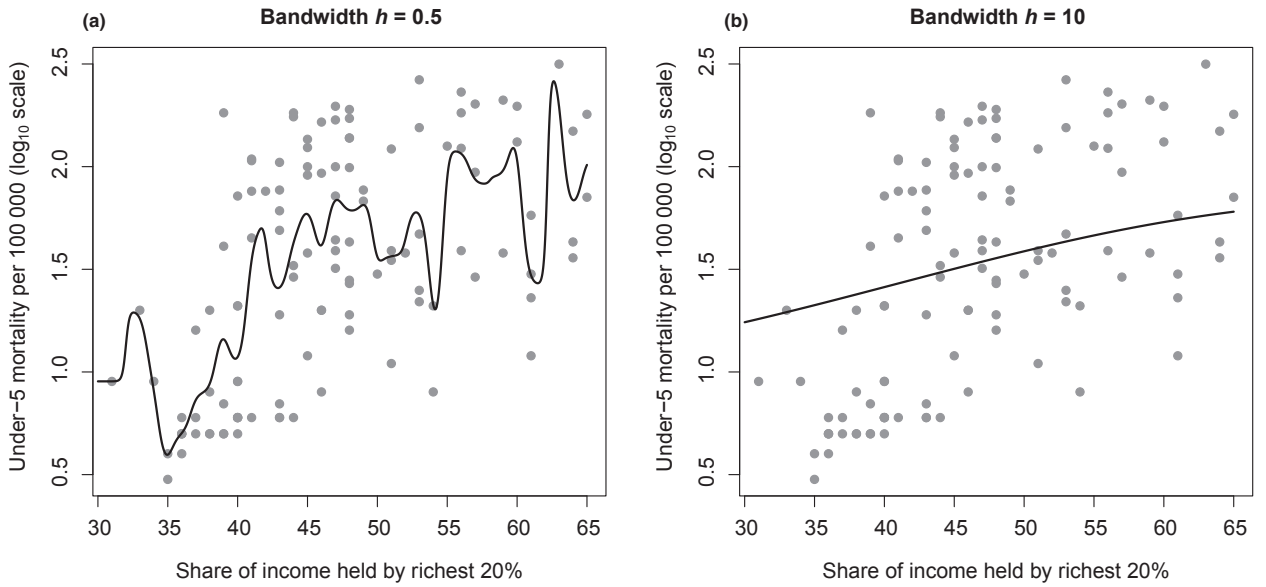

Figure 6 The effect of bandwidth on Nadaraya–Watson kernel regression. In the left-hand panel (a) a very small value $h = 0.5$ is used for the bandwidth, resulting in undersmoothing. In the right-hand panel (b) the fitted nonparametric regression is oversmoothed through the use of the large bandwidth $h = 10$.

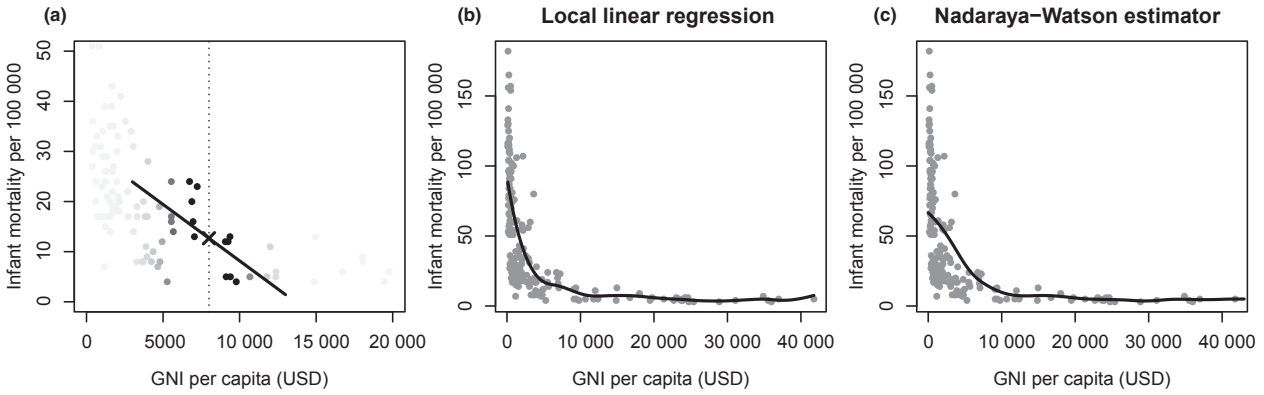

Figure 7 Local linear regression for the infant mortality against share of GNI per capita (USD). The left-hand panel (a) illustrates construction of the estimate at $x = 8000$. The gray-scale intensity of the points indicates the weights attached to each point when fitting the regression line. The cross indicates the local linear estimate at the point $x = 8000$. The middle panel (b) shows the complete local linear regression curve, while the right-hand panel (c) displays the Nadaraya–Watson estimate for comparison. Note how the local linear estimator provides a more plausible fit to the data near the boundary at $x = 0$. All estimates were computed using a bandwidth of $h = 8000$ and a Gaussian kernel.

we get the order p local polynomial estimator $\widehat{m}_p(x)$ by following the same kind of steps as above. The Nadaraya–Watson estimator is simply the order $p = 0$ local polynomial estimator, and hence is sometimes referred to as the local constant estimator.

For practical implementation of local polynomial regression we have a number of choices to make, concerning the order p of polynomial, the kernel function, and the value of the bandwidth. With regard to the first of these choices, local linear regression (corresponding to $p = 1$) is a popular and effective, in part, because of its good behavior near boundaries. Both theory and practice have shown that the choice of kernel function is of relatively minor importance; the biweight and Gaussian kernels are commonly used. The selection of the bandwidth h is far more critical. Major approaches to bandwidth selection for local polynomial regression mirror those discussed earlier in the context of the Nadaraya–Watson estimator. See Fan and Gijbels (1996) for further discussion of these matters.

LOWESS (an acronym of locally weighted scatterplot smoothing) and its close cousin LOESS are popular variants of local polynomial kernel regression (Cleveland, 1981). The principal difference between LOWESS (and LOESS) and the methods described above is that the local polynomial is fitted by weighed least squares to all data within a window the width of which is adjusted to contain a fixed proportion of the data. This is effectively local polynomial regression in which the bandwidth h for estimation at the point x adapts to the density of data close to x. The proportion of data determining the estimation window is called the 'span.' The fitted curves in Figure 2 were obtained using LOESS.

Regression with Splines

Spline smoothing provides an alternative approach to nonparametric regression. A spline is constructed by dividing the x-axis into a sequence of intervals. The spline is defined by different polynomials on each interval, but in such a way that these functions tie together smoothly at the 'knots,' i.e., the interval boundaries. The spline is said to be of pth order if its constituent polynomials are of that degree. A pth order spline must be $p - 1$ times differentiable at the knots.

Splines provide a very flexible class of functions for regression. Spline-based regression models have a nonparametric nature because the division of the x-axis into intervals means that estimation is essentially local. Nonetheless, splines are defined in terms of a large number of parameters – a number that is unbounded if the number of knots is allowed to increase indefinitely with the sample size n. This apparent paradox simply reflects the slightly misleading nature of the nomenclature 'nonparametric regression,' which nowadays is interpreted as a regression methodology that is local in nature.

The simplest variety is the linear (i.e., $p = 1$ order) spline, which is sometimes referred to as the broken stick model. A linear spline can be expressed as

$$s(x) = \beta_0 + \beta_1 x + \sum_{j=1}^{J} b_j (x - \kappa_j)_+ \qquad [5]$$

where $\kappa_1, \ldots, \kappa_J$ are the locations of J knots for the spline, and $(a)_+ = \max(0, a)$ for any number a. For a point $x < \kappa_1$, we have $s(x) = \beta_0 + \beta_1 x$. At the first knot the slope jumps by b_1, and hence the coefficient of x becomes $(\beta_1 + b_1)$ for $\kappa_1 \le x < \kappa_2$. The slope adjusts at each subsequent knot by the amounts specified by the parameters $b_2, \ldots, b_J$. This is illustrated in the left-hand panel of Figure 8.

A general pth order spline is given by

$$s(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{j=1}^{J} b_j (x - \kappa_j)_+^p \qquad [6]$$

where the change at knot j is to the coefficient of $(x - \kappa_j)^p$. As an example, consider the following quadratic ($p = 2$) spline with a single knot at $\kappa_1 = 1$:

$$s(x) = 5 + 4x + 3x^2 + 2(x - 1)_+^2$$

Figure 8 Some linear splines. The left-hand panel (a) illustrates the construction of a linear spline with knots at $x = 3$, 6, 9, marked by dotted vertical lines. To the left of the first knot the equation of the spline is $s(x) = 2 + x$. The slope then changes by $b_1 = -1.5$ at $\kappa_1 = 3$, and hence the coefficient of x is -0.5 for $3 \le x < 6$. The slope changes by $b_2 = 0.5$ and $b_3 = 0.9$ at the two subsequent knots. The central (b) and right-hand (c) panels depict linear splines fitted to the data on logarithm of under-5 mortality against share of household income held by the richest 20%. Locations of knots are indicated by the tick marks on the x-axis. The spline in the central panel is fitted using the method of least squares, and that in the right-hand panel is fitted using penalized least squares.

The quadratic equation defining the spline is different either side of the knot. We have

$$s(x) = 5 + 4x + 3x^2$$

when $x \le 1$, and

$$\begin{aligned} s(x) &= 5 + 4x + 3x^2 + 2(x-1)^2 \\ &= 7 + 5x^2 \end{aligned}$$

when $x \ge 1$.

We can use splines for nonparametric regression by replacing the general regression function m in eqn [1] by a spline s, so that $y_i = s(x_i) + \varepsilon_i$ (for $i = 1, \ldots, n$). Despite their flexibility, spline-based nonparametric regression is a linear model because the response y is a linear combination of the model parameters $(\beta_0, \ldots, \beta_p, b_1, \ldots, b_J)$. This means that splines can in principle be fitted to the data using the method of least squares. However, splines are so heavily parameterized that we will get severe over-fitting if this approach is employed without modification. This is illustrated in the middle panel of **Figure 8**, where a linear spline is fitted to the data on **Under.5.MR** against **Hi.20.pc**. The resultant regression curve is very jagged and noisy. This occurs because the changes to the slope of the spline, described by the parameter estimates $\widehat{b}_1, \ldots, \widehat{b}_J$, are too extreme. This is not a problem peculiar to linear splines. We would also get a noisy curve were we have to fit a higher order spline $(p \ge 2)$ using the method of least squares.

We can hope to get a more aesthetically pleasing, smooth fit if we reduce the magnitude of parameter changes at the knots. This motivates estimation of parameters based on minimization of a penalized sum of squares,

$$Q(\boldsymbol{\beta}, \mathbf{b}) = \sum_{i=1}^{n} \{y_i - s(x_i)\}^2 + \alpha \sum_{j=1}^{J} b_j^2 \qquad [7]$$

where α is a tuning parameter. In eqn [7] the standard sum of squares criterion (the first term of the right) is augmented by a penalty (the second term on the right) that is increasingly severe as the parameters $b_1, \ldots, b_J$ increase in magnitude. As a consequence, the estimates for this group of parameters obtained by minimizing eqn [7] will be shrunk toward zero in comparison to standard least squares estimates. This is illustrated in the right-hand panel of **Figure 8**, where a linear spline is fitted to the data on **Under.5.MR** against **Hi.20.pc** using the method of penalized least squares.

In order to implemented spline-based regression modeling in practice we need to make a number of choices, including selecting the degree of the spline and the number and position of the knots. With regard to the former, cubic splines $(p = 3)$ are most widely used because of their smoothness properties. Turning to the choice of knots, we could place a knot at each unique value of the predictor variable. This produces a class of models called 'smoothing splines' (Wegman and Wright, 1983; Green and Silverman, 1993). A very popular alternative is to select a much smaller number of knots, often following the rule of thumb $J = \min(35, \lfloor n/4 \rfloor)$ where $\lfloor a \rfloor$ denotes rounded down to the nearest integer. Models with this reduced number of knots, and which are fitted by minimization of eqn [7], are called 'penalized regression splines' (Eilers and Marx, 1996). The placement of the knots is not usually critical. Standard approaches are to space the knots evenly across the range of the predictor, or to place them at regular quantiles thereof (cf Ruppert, 2002).

A more significant issue is choice of the value of the tuning parameter α in eqn [7]. A relatively large value of α implies particularly heavy penalties for large estimates $\widehat{b}_1, \ldots, \widehat{b}_J$, and hence will tend to produce a smooth fitted spline with only minor changes to coefficients at the knots. Conversely a small value of α indicates only a weak effect for the penalty term, and hence a model fit little different to that obtained using standard least squares. A number of approaches to selecting α have been considered, including cross-validation, generalized cross-validation, and likelihood-based methods (cf Kauermann, 2005). The degree of smoothing in the linear spline in the right-hand panel of **Figure 8** was chosen using likelihood methods.

Nonparametric Regression with Multiple Predictor Variables

We have focused thus far on regression involving a single predictor. When there are multiple predictors, $x_1, \ldots, x_r$, the most general form of nonparametric regression model is

$$y_i = m(x_{i1}, \ldots, x_{ir}) + \varepsilon_i \quad (i = 1, \ldots, n) \qquad [8]$$

where x_{ij} is the value of the jth predictor for the ith observation. The regression function m now describes the variation in the conditional mean of y in terms of all the predictors, in an entirely general way.

Local polynomial kernel regression extends quite simply to this multiple predictor case. In the case of local constant regression, the estimate regression function still follows the form of eqn [2], but with the scalars x and x_i replaced respectively by vectors $\mathbf{x} = (x_1, \ldots, x_r)^T$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ir})^T$. The multivariate (scaled) kernel function $K_h(\mathbf{x} - \mathbf{x}_i)$ continues to define the weights used in the local averaging process. A popular choice is the multivariate Gaussian kernel, the bivariate version of which is plotted in **Figure 9(a)**. There are a number of types of scaling that can be applied to multivariate kernels, including different values for the bandwidth for each predictor (a necessity unless the predictors are pretransformed to a common scale). For pth order local polynomial regression the methodology continues to involve optimization of weighted sums of squares akin to eqn [4], although the polynomials will now be in terms of the multiple predictor variables. We illustrate local constant kernel regression of **Under.5.MR** on **Adult.lit** and **Hi.20.pc** in **Figure 9(b)**.

Fitting the nonparametric regression from eqn [8] is useful and viable only when the dimension r of the predictor variables is modest. An immediate practical problem is visualization of the fitted models. This is possible when $r = 2$ (for example, using perspective plots as in **Figure 9**, contour plots or heat maps), but not in higher dimensions. Of course, nonparametric regression models with $r \geq 3$ can be used for other purposes than visualization (e.g., prediction). However, estimation of such models rapidly becomes more difficult as r increases because of the 'curse of dimensionality,' which in essence requires the sample size to increase exponentially with r in order to retain any given level of precision.

To explain more fully the curse of dimensionality, suppose that we are interested in estimation of the regression function at a point $\mathbf{x}_0 = (x_{01}, x_{02}, \ldots, x_{0r})^T$. To start with, consider the case with a single predictor (i.e., $r = 1$) and suppose that 1/10 of the data have an x-value that falls within a window of length $2h$ centered on the estimation point (where h is the bandwidth). We can expect to compute a serviceable local average from this fraction of the data even for samples as small as $n = 100$. If we had $r = 2$ independent predictors with the same marginal property, then only $(1/10)^2 = 1/100$ of the data will fall in a square estimation window centered on the estimation point and with sides of length equal to $2h$. For r predictors of this type, only a fraction $(1/10)^r$ of the data will be within an r-dimensional hypercube window centered on $\mathbf{x}_0$ and with side lengths $2h$. Even with a sample size as large as $n = 10\,000$, we would probably see no data within such a window around the estimation point when there are $r = 6$ predictors.

A common method of addressing this problem is to assume that the effects of the predictors can be expressed in additive form, so that

$$y_i = \beta_0 + m_1(x_{i1}) + \cdots + m_r(x_{ir}) + \varepsilon_i \quad (i = 1, \ldots, n) \qquad [9]$$

where the regression functions $m_1, \ldots, m_r$ are each specific to a single predictor, and hence not subject to the curse of dimensionality. The term β_0 in eqn [9] models the 'grand mean' for the response, while each of the individual functions $m_1, \ldots, m_r$ is centered at zero on the response scale. It should be recognized that replacing the general model in eqn [8] by the additive model [9] requires the (strong) assumptions that there are no interactions between predictors. In other words, the relationship between y and predictor x_j is assumed to be the same regardless of the values taken by the other predictors.

The additive model defined by eqn [8] is often fitted using splines to represent the functions $m_1, \ldots, m_r$. See for example Wood (2006). The fitting algorithms employ penalized least squares. An advantage of using splines is that their linear model structure makes inference relatively straightforward, so that it is reasonably easy to derive quantities like standard errors and confidence intervals. The estimated functions $\widehat{m}_1, \ldots, \widehat{m}_r$ can be visualized to provide an impression of the relationship between the response and any given predictor, having taken account of the effects of the other predictors. This is illustrated in **Figure 10**, where we display results for the additive model

$$\begin{aligned} \mathsf{E}[\log(\mathbf{Under.5.MR})] = {} & \beta_0 + s_1(\mathbf{Adult.lit}) + s_2(\mathbf{Hi.20.pc}) \\ & + s_3(\mathbf{School}) + s_4(\mathbf{Low.40.pc}) \end{aligned}$$

where $s_1, \ldots, s_4$ are cubic spline representations of the component regression functions $m_1, \ldots, m_4$, respectively.

Other Models and Methods of Nonparametric Regression

Regression Trees and Related Methods

A wide variety of methods of nonparametric regression have been proposed. Many can be regarded as variations and extension of the kernel local polynomial regression and spline techniques discussed above.

One important class of models with somewhat different foundations comprises recursive partitioning methods (Breiman et al., 1984). These seek to divide the data into small groups of similar observations, and then estimate the mean response within the group by the group mean. In binary regression tree models, the partitions develop through a sequence of binary splits. In more detail, each partition involves splitting one of the current groups according to a division at some value of one of the predictor variables. The choice of variable and split value is made so as to minimize the residual sum of squares. This methodology is illustrated in **Figure 11**. Regression trees of this sort can be effective tools for prediction, but are less well suited to data visualization because they produce piecewise constant regression curves, as per the right-hand panel of **Figure 11**.

Multivariate adaptive regression splines (MARS) can be regarded as a generalization of regression trees that are better suited to handling continuous numerical predictors (Friedman,

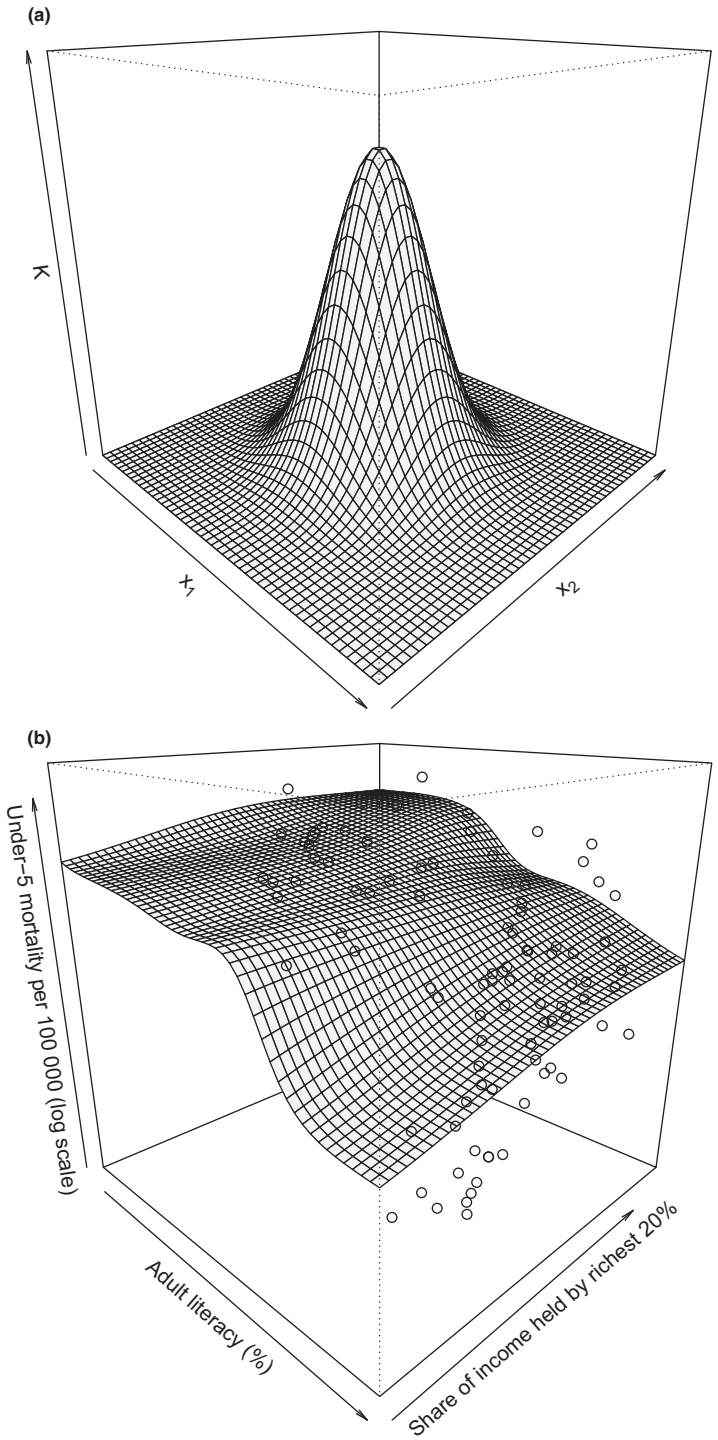

Figure 9 (a) a perspective plot of a bivariate Gaussian kernel. (b) a perspective plot of a local constant kernel regression of logarithm of under-5 mortality against adult literacy rate and share of household income held by the richest 20%, superimposed on a three-dimensional scatterplot of the data.

1991). While regression trees can be thought of as zero-order splines (since they are piecewise constant functions), MARS is based on linear splines. MARS is an adaptive method in the sense that the locations of the knots are not prespecified, but rather are selected in order to optimize fit to the data. The name 'MARS' is trademarked to Salford Systems, and as a consequence open-source implementations of this general methodology are often called 'Earth.'

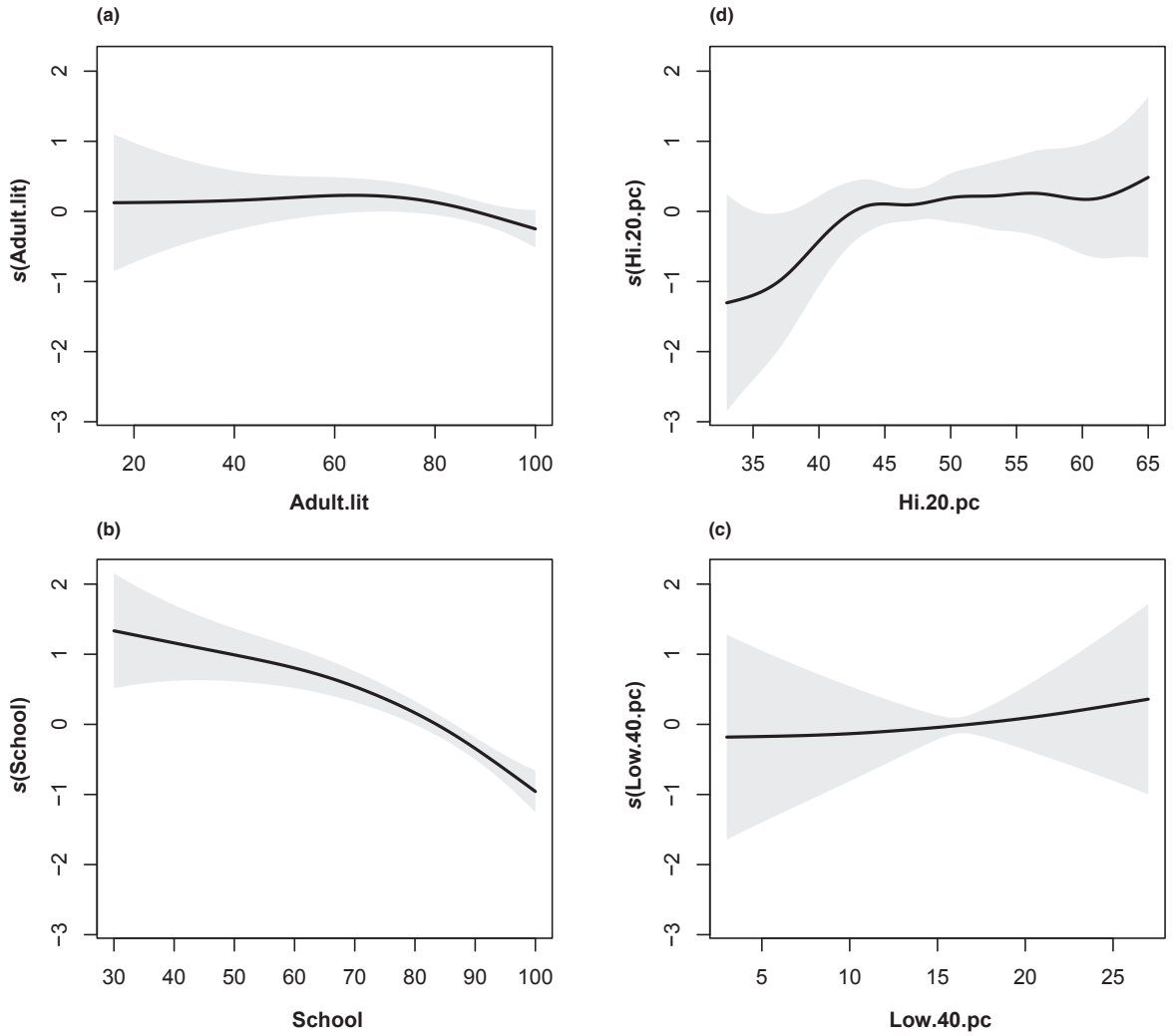

Figure 10 Estimates of component regression functions for an additive model relating logarithm of under-5 mortality to (a) adult literacy rate (%), (b) attendance rate at primary school (%), (c) share of household income held by the poorest 40%, and (d) share of household income held by the richest 20%. Each regression function is centered on zero on the response scale, and hence represents a change from the overall mean response. The shaded areas indicate 95% confidence intervals for the fitted curves.

Nonparametric Regression for Binary and Count Data

The methods discussed above are intended for use with a continuous numerical response variable y, typically assumed to follow a normal distribution conditional on x. However, nonparametric regression can also be employed to describe trends when the response y is binary (e.g., yes/no, died/survived) or a count (particularly a count sufficiently small that a continuous approximation is unappealing). A common and effective generalization of the additive model in eqn [9] is to model a transformed version of the mean response, so that

$$g(\mu_i) = \beta_0 + m_1(x_{i1}) + \cdots + m_r(x_{ir}) \qquad [10]$$

where $\mu_i = \mathrm{E}[y_i]$. The choice of 'link' function g depends on the nature of the response variable. For a binary response coded 0 or 1, the mean response is $\mu_i = \mathbb{P}(y_i = 1)$ and a natural choice of link is the logit transform $g(\mu) = \log(\mu/(1-\mu))$. For a Poisson distributed count response, $g(\mu) = \log(\mu)$ is natural. Equation [10] defines a class of models called generalized additive models, which can be regarded as nonparametric versions of classical generalized linear models. See Hastie and Tibshirani (1990).

Software Implementation

Most modern statistical software packages include an implementation of at least some basic types of nonparametric regression model, but the availability of more sophisticated techniques is patchy. Probably the most routinely available method is LOESS, which is implemented in Minitab, SPSS, Stata, SAS and R, amongst many others. Regression trees are also implemented in a wide variety of packages. Kernel local polynomial regression, penalized spline regression, and generalized additive models are less common. Nonetheless,

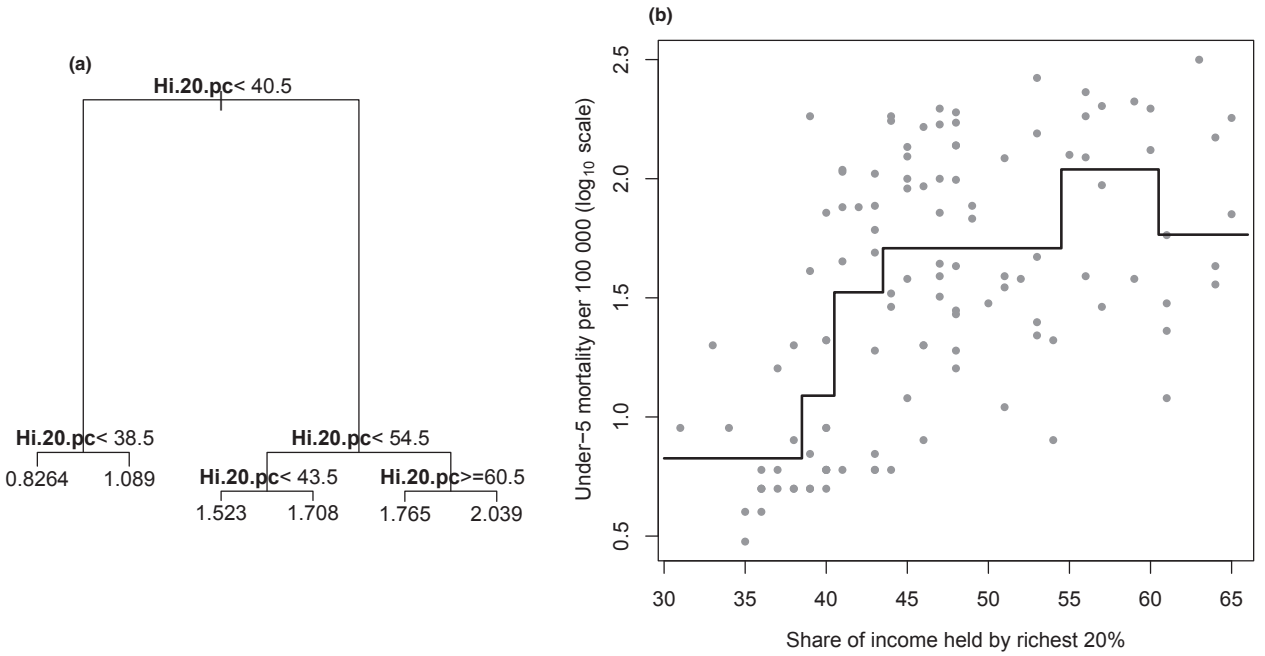

Figure 11 Regression tree for data on logarithm of under-5 mortality against share of household income held by the richest 20%. The left-hand panel (a) displays the tree defining the recursive partition of the data. Observations satisfying the inequalities at nodes proceed down left-hand branches; those failing to do so go down right-hand branches. Fitted values are given at terminal nodes (leaves) at the bottom of the plot. The right-hand panel (b) displays the resultant regression curve.

both SAS and R have algorithms for all of these techniques, and in the latter case a number of versions thereof.

Bibliography

Azzalini, A., Bowman, A., 1993. On the use of nonparametric regression for checking linear relationships. Journal of the Royal Statistical Society, Series B (Methodological), 549–557.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press, New York.

Cleveland, W.S., 1981. Lowess: a program for smoothing scatterplots by robust locally weighted regression. American Statistician, 35–54.

Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing with *B*-splines and penalties (with discussion). Statistical Science 11 (2), 89–121.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and Its Applications. CRC Press.

Friedman, J.H., 1991. Multivariate adaptive regression splines. The Annals of Statistics, 1–67.

Green, P.J., Silverman, B.W., 1993. Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. CRC Press.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall, New York.

Kauermann, G., 2005. A note on smoothing parameter selection for penalized spline smoothing. Journal of Statistical Planning and Inference 127 (1), 53–69.

Nadaraya, E.A., 1964. On estimating regression. Theory of Probability and Its Applications 9, 141–142.

Ruppert, D., 2002. Selecting the number of knots for penalized splines. Journal of Computational and Graphical Statistics 11 (4).

Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. Cambridge University Press.

UNICEF, 2003. The State of the World's Children. Oxford University Press.

Wand, M.P., Jones, M.C., 1995. Kernel Smoothing. Chapman & Hall, London.

Watson, G.S., 1964. Smooth regression analysis. Sankhyā, Series A 26, 359–372.

Wegman, E.J., Wright, I.W., 1983. Splines in statistics. Journal of the American Statistical Association 78 (382), 351–365.

Wood, S., 2006. Generalized Additive Models: an Introduction with R. CRC Press.

Nonparametric Statistics: Advanced Computational Methods

Myles Hollander and Jayaram Sethuraman, Florida State University, Tallahasse, FL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 10660–10667, © 2001, Elsevier Ltd.

Abstract

Statistical methods are useful in obtaining information about the unknown state of nature or the 'parameter' as it is usually referred to in the literature. A statistician collects suitable data whose distribution depends on the unknown parameter. A statistical inference procedure is then devised to produce information about the unknown parameter or a function of that parameter. The classical methods of inference assume that the probability distributions that govern the data depend only on a few unknown quantities (or parameters). Such procedures are called parametric procedures. When one is not able to make such strong assumptions about the probability models that govern the data, and/or when one can be more certain of the rankings among the data than the exact values, one uses robust inference procedures known as nonparametric methods. Tests and estimates based on the simpler nonparametric methods can be obtained from easy computations based on the data. Procedures that have validity under more general models will require more heavy computations. When there is some uncertainty about the probability models and/or when expert information concerning the problem at hand is available, one should use Bayesian methods. These methods can also tend to be computationally intensive. We describe briefly the main ingredients of some new Bayesian and computational methods in nonparametric inference. These include bootstrapping and Gibbs sampling.

Standard inference procedures are based solely on data whose probability distributions have been specified in terms of a few parameters. Bayesian methods incorporate existing knowledge (historical information, expert knowledge, previous data, etc.) about the parameter into the inference process. This is done by assigning a probability distribution, called the prior distribution, to the parameter governing the distribution of the data. This prior distribution represents the information available to us concerning the parameter before gathering data. After some data have been observed, we can compute the conditional distribution of the parameter given the observed data; this is called the posterior distribution and it represents our knowledge of the parameter at this stage. If one wants to estimate a function of the parameter, then one can just find the expectation of that function under this posterior distribution; it will be the best estimate under squared error loss. Similar methods of estimation under other loss functions are possible and they all depend solely on the posterior distribution. Thus Bayesian methods concentrate first on finding the posterior distribution or on simulating from the posterior distribution.

In a nonparametric problem, the common unknown distribution F of the data is the 'parameter.' A Dirichlet measure is a distribution for F and will serve as a prior distribution in nonparametric problems. The real problems here are how to define such nonparametric prior distributions for this and other wider classes of problems, and how to carry out the necessary computations.

In Section Bayesian Methods and Dirichlet Measures below we describe various ways to define the Dirichlet measure, which is a natural choice for a prior distribution in nonparametric problems. Closed form solutions for inference in Bayesian nonparametric problems are given in several cases together with some references for further study.

Since closed form solutions will in general not be available in interesting situations, advanced computational approaches are necessary to carry out inference in nonparametric Bayesian problems. Such computational methods have been used in regular nonparametric problems as well. In Sections Testing for Concordance and Testing Independence we give examples of computational intensive methods in two standard nonparametric problems, namely testing for independence and testing for concordance of two variables. In Section Gibbs Sampling, the estimation of a distribution function when observations have been censored by some censoring mechanism is treated as a Bayesian nonparametric problem. We describe the advanced computational approach, known as Gibbs sampling, which can be used for this problem. We present the exact Bayes estimate and the Bayes estimate obtained by Gibbs sampling to demonstrate that power of the computational method. For further comparison purposes, the estimate previously obtained by frequentist methods is also presented.

Bayesian Methods and Dirichlet Measures

Bayesian methods are quite general and apply to both parametric and nonparametric inference problems. It is easier in the parametric case since the unknown parameter usually varies in a finite dimensional space and we are familiar with probability distributions in finite dimensional spaces. In nonparametric problems the unknown 'parameter' can be as complex as the underlying distribution of the observed random variable. This is not only an infinite dimensional quantity, it also satisfies some inequality restrictions; for instance a distribution function $F(x)$ satisfies $0 \leq F(x) \leq 1$ for all $x \in R_1$, where R_1, is the real line, and is also a nondecreasing right continuous function with limits 0 and 1 as x tends to $-\infty$ and ∞, respectively. Thus the first problem faced by Bayesians in nonparametric inference problems is an appropriate definition of a probability distribution for $F(\cdot)$. Any such prior distribution will be useful only if it is tractable and yields closed form or computable solutions to the nonparametric inference problems on hand. These problems were daunting at first, but were elegantly solved by Ferguson (1973) who introduced Dirichlet measures as prior distributions.

We will look closely at probability measures P (which are equivalent to distribution functions) on R_1, the real line. The same can be done for probability measures on more general spaces. For any partition of $\mathscr{A} = (A_1, \ldots, A_k)$consisting of disjoint measurable sets whose union is R_1, the vector $P(\mathscr{A}) = [P(A_1), \ldots, P(A_k)]$ is a k-vector of nonnegative numbers adding to 1. Notice that P is completely determined if $P(\mathscr{A})$ is known for all or most partitions $\mathscr{A}$ and satisfies obvious additivity and consistency conditions. It can be shown that a consistent family of probability measures for $P(\mathscr{A})$ as $\mathscr{A}$ varies among all finite partitions will define a unique probability measure for P. This is the approach of Ferguson (1973).

To describe Ferguson's Dirichlet measure, we will first look at finite dimensional Dirichlet distributions. These are distributions in R_k concentrated on the subset

$$\mathscr{P}_k = \left\{\mathbf{p} = (p_1, \ldots, p_k); \quad p_1 \geq 0, \ldots, p_k \geq 0, \quad \sum_i p_i = 1\right\}$$

A random vector $(Y_1, \ldots, Y_k) \in \mathscr{P}_k$ is said to have a finite dimensional Dirichlet measure with parameters $(\beta_1, \ldots, \beta_k)$ with $\sum_j \beta_j > 0$, if $Y_i = Z_i/(\sum_j Z_j)$ where $Z_1, \ldots, Z_k$ are independent Gamma random variables with scale parameter 1 and shape parameters $\beta_1, \ldots, \beta_k$ respectively. When $\beta_i > 0$, $i = 1, \ldots, k$, this distribution is given by the more familiar density function of$(Y_1, \ldots, Y_{k-1})$:

$$\frac{\Gamma\left(\sum_i \beta_j\right)}{\Pi_j \Gamma(\beta_j)} \prod_j y_j^{\beta_j - 1}$$

where $y_k = 1 - \sum_1^{k-1} y_j$. This finite dimensional Dirichlet measure will be denoted in symbols by $\mathscr{D}(\beta_1, \ldots, \beta_k)$.

Let $\mathscr{P}$ be the space of all probability measures on R_1. Ferguson defined a Probability measure called a Dirichlet measure for a random Probability measure P taking values in $\mathscr{P}$. (To be more rigorous, we should also define the σ field in $\mathscr{P}$, but we shall avoid going into such details.) Ferguson defined the Dirichlet measure with parameters α and μ, where $\alpha > 0$ and μ is a probability measure on R_1, as follows: for every finite partition $\mathscr{A}$, $P(\mathscr{A})$ has the finite dimensional Dirichlet distribution with parameters $\mathscr{D}[\alpha\mu(A_1), \ldots, \alpha\mu(A_k)]$. In particular the distribution of $P(A)$ is a Beta distribution with parameters $[\alpha\mu(A), \alpha\mu(A^c)]$. It takes some more mathematical machinery to show that such a Dirichlet measure exists, but again we will not go into this here. The Dirichlet measure above will be denoted in symbols by $\mathscr{D}(\alpha, \mu)$.

Many interesting properties of Dirichlet measures are known. These are easy to derive from the representation of Dirichlet measures given by Sethuraman (1994) which is described in the next paragraph. The Dirichlet measure $\mathscr{D}(\alpha, \mu)$ gives probability one to the class of all discrete distributions in R_1 even if μ is a continuous distribution. The expected value and variance of $P(A)$ are $\mu(A)$ and $\mu(A)\mu(A^c)/(\alpha + 1)$. The expected value and variance of$\int f dP$ is $\int f d\mu$ and $(\int f^2 d\mu - (\int f d\mu)^2)/(\alpha + 1)$.

An easy constructive definition of a Dirichlet measure was announced in Sethuraman and Tiwari (1982) and proved in more detail in Sethuraman (1994). Consider independent random variables $Y_1, Y_2, \ldots, \theta_1, \theta_2, \ldots$ where the Y's are identically distributed (*iid*) with common distribution μ and the θ's are identically distributed with a common Beta distribution with parameters 1 and α. Define

$$p_1 = \theta_1, \quad p_k = \theta_k \prod_{j=1}^{k-1} (1 - \theta_j) \qquad [1]$$

$$P(A) = \sum_k p_k I(Y_k \in A) \qquad [2]$$

where for any event B, the indicator function $I(B) = 1$ if B occurs and $I(B) = 0$ otherwise. Clearly, P is a random discrete probability measure putting random probability mass p_k at Y_k. Its distribution is the Dirichlet measure $\mathscr{D}(\alpha, \mu)$.

Let $X_1, X_2, \ldots$ be random variables such that, given P, they are i.i.d. with common distribution P. Let F_n be the empirical distribution of $X_1, X_2, \ldots, X_n$, that is, $F_n(x) = \frac{1}{n}\sum_1^n I(X_i \leq x)$. The conditional distribution of P, given the sample $X_1, X_2, \ldots, X_n$, was shown by Ferguson (1973) to be the Dirichlet measure with parameters $(\alpha + n, \mu_n)$, where

$$\mu_n = \frac{\alpha}{\alpha + n}\mu + \frac{n}{\alpha + n}F_n \qquad [3]$$

Blackwell (1973) and Blackwell and MacQueen (1973) give alternative ways of defining Dirichlet measures and establishing the above result on the posterior distribution.

We will now briefly describe some popular uses of the Dirichlet prior in nonparametric problems. Suppose that we have data consisting of $X_1, \ldots, X_n$ which, given P, are independent and identically distributed with common probability distribution P. Suppose that we want to estimate P. The frequentist's usual estimator is $F_n(X) = \frac{1}{n}\sum_1^n I(X_i \leq x)$ the empirical distribution of the data $X_1, \ldots, X_n$.

A Bayesian would place a Dirichlet prior $\mathscr{D}(\alpha, \mu)$ as a prior distribution for P. The measure μ would represent his average guess of the unknown P and α would represent his measure of confidence in this guess, with large values of α denoting more confidence. Then the posterior distribution of P is the Dirichlet measure $\mathscr{D}(\alpha + n, \mu_n)$ where μ_n is given by eqn [3]. Under squared error loss, the Bayes estimate of any function $g(P)$ is simply its expectation of $g(P)$ under this posterior distribution. Thus the Bayes estimates of P, $P((-\infty, x])$ and $\int f(x)dP(dx)$ are μ_n, $\mu_n((-\infty, x])$ and $\int f(x)\mu_n(dx)$, respectively. Under absolute value of error loss function, the Bayes estimate of the median of P is the median of μ_n. The expected Bayes loss of all these estimates are obtained in Ferguson (1973). To summarize these results in simple terms, the Bayes estimate of P is μ_n which is a convex linear combination of the parameter μ of the prior distribution and the empirical distribution function F_n of the observed data, where the weight given to μ, the prior guess, goes to zero as the sample size of the data increases, a fact that agrees well with intuition. Susarla and Van Ryzin (1976) obtain Bayes estimates of P when the observations are censored by independent censoring variables, which we discuss later. For more examples of Bayes estimates of other functionals see Yamato (1975), Hollander and Korwar (1982), Zalkikar et al. (1986), Wells and Tiwari (1989), and Lu et al. (1994).

The influence of the parameter μ in the prior distribution goes to zero as n gets large and the Bayes estimate is close to F_n. In fact, if one were to formally put the parameter α of the prior distribution to be zero, then the Bayes estimate is F_n; this has

led some researchers to refer to α as the prior sample size. There is a difficulty with this notion. The prior distribution is not defined if $\alpha = 0$. However, the prior distribution converges to a proper prior distribution as $\alpha \to 0$ and μ is held fixed. This limiting prior corresponds to a random degenerate distribution, and as such will not be an acceptable prior distribution for Bayesian problems (see Sethuraman and Tiwari, 1982). In spite of this, many researchers check to see if they get familiar frequentist estimators when $\alpha \to 0$.

This summarizes the first results concerning applications of Dirichlet measures for Bayesian analysis in one-sample problems. Similar results which have wider applicability have been obtained for two-sample problems, Bayesian density estimation, censored samples, reliability, point processes, etc.

There has been work in finding prior distributions which are more general than Dirichlet measures, like measures that concentrate on absolutely continuous distributions, or distributions with increasing failure rate, and so on. The problem of constructing a random probability distribution requires a clear understanding of the structure of a probability measure.

Consider partitioning the whole space into two sets A_0, A_1, and then partitioning each of these as A_{00}, A_{01} and A_{10}, A_{11}, respectively. At the nth stage we will have 2^n sets A_{t_n}, where t_n is an n-vector of zeros and ones. The set A_{t_n} will be partitioned as $A_{t_n,0}$, $A_{t_n,1}$, and we will have 2^{n+1} sets at the $(n+1)$th stage, and so on. Suppose the partitioning sets from all stages, namely $\{A_0, A_1, \ldots, A_{t_n}, \ldots\}$, generate the Borel σ-field $\mathscr{B}$ in R_1. Any probability measure P on R_1 gives rise to the quantities $[u_e = p(A_1),\ u_0 = P(A_{01}|A_0),\ u_1 = P(A_{11}|A_1), \ldots,\ u_{t_n} = P(A_{t_n1}|A_{t_n}), \ldots]$ all of which lie between zero and one and conversely any sequence on numbers $(u_e, u_0, u_1, \ldots, u_{t_n}, \ldots)$ in [0, 1] will give rise to a unique probability measure P on R_1. Furthermore, the range of each u_{t_n} is unrestricted. Thus it is easy to assign probability distributions to (u_{t_n}). Any such distribution induces a probability measure on the space of probability measures on R_1 and can serve as a prior distribution in nonparametric problems. If we assume that the u's are independent with u_{t_n} having a Beta distribution with parameters $[\alpha\mu(A_{t_n1}),\ \alpha\mu(A_{t_n0})]$, then Blackwell (1973) showed that the resulting distribution for P is $\mathscr{D}(\alpha, \mu)$. By choosing other distributions for the u's one can obtain prior distributions that concentrate on absolutely continuous distributions, distributions which are symmetric, unimodal, or have increasing failure rate, etc.

Another way to introduce prior distributions for P arises from the fact that, when the distribution of $X_1, X_2, \ldots$ are i.i.d. P given P, the distribution of $X_1, X_2, \ldots$ is exchangeable. In fact, from De Finetti's theorem, if $X_1, X_2, \ldots$ is exchangeable, then there is a function P of $X_1, X_2, \ldots$, which is a random probability measure, and given this P, the $X_1, X_2, \ldots$ become i.i.d. P. The distribution of this P is a prior distribution on the space of all probability measures on R_1. Thus every sequence of exchangeable random variables corresponds to a prior distribution on the space of probability measures. Using this approach, Blackwell and MacQueen (1973) showed that the Pólyaurn scheme produces an exchangeable sequence of random variables leading to Dirichlet measures. Exchangeable distributions are usually derived as i.i.d. random variables with a common P having a prior distributions rather than vice versa, and so this method has not yielded a rich variety of prior distributions. There is, however, a mathematical beauty in this method.

Computational Methods in Nonparametrics: Bootstrap and Gibbs Sampling

Two methods, namely the bootstrap (*see* Resampling Methods of Estimation) and Gibbs sampling methods (*see* Monte Carlo Methods and Bayesian Computation: Overview), have proved very successful in handling a multitude of inference problems that could not be handled before since the various distributions needed for inference are generally mathematically intractable. These sampling methods replace the mathematical machinery with computational methods to obtain approximations for large classes of distributions. A lot of ingenuity still has to go into each situation to make these computational methods effective. Much theory has also been developed to prove that these computational methods work when the number of iterations and sample sizes become large, but we will not describe those developments here. We will give a general introduction and describe some examples.

Let $X_1, \ldots, X_n$ be iid with common distribution P. Let $F_n = \frac{1}{n}\sum_1^n I(X_i \le x)$ be the empirical distribution function which is an estimate of P. Suppose that we wish to estimate the parameter $\theta = g(P)$, where $g(P)$ is some function of P. The population mean $\theta = \int xP(\mathrm{d}x)$ is an example of such a parameter. An obvious estimate of this θ would be $g(F_n)$. One would also like to give the precision of this estimate, namely the standard error of this estimate. This can be done using the bootstrap. The bootstrap method says that if we draw a sample $X_1^*, \ldots, X_n^*$ of size n from F_n (i.e., a simple random sample of size n with replacement from the originally observed sample) and compute its empirical distribution function F_n^* then $g(F_n^*)$ bears the same relation to $g(F_n)$ as $g(F_n^*)$ bears to $g(P)$. If we repeat this procedure K times where K is large, then the standard deviation of the K values of $g(F_n^*)$ will be an estimate of the precision of the estimate $g(F_n^*)$. This, in a nutshell, is the bootstrap method in its simplest form, as introduced by Efron (1979). The asymptotic accuracy of the bootstrap has been established among others by Bickel and Freedman (1981) and Singh (1981). This method can apply to parametric as well as nonparametric problems. We will illustrate the bootstrap in two nonparametric problems below.

Testing for Concordance

Kendall's τ is a measure of concordance, namely the tendency of two random variables X and Y to increase or decrease together. It possesses the desirable property that it is invariant under monotone increasing transformations of X and Y and so depends only on the rankings and not the exact magnitudes of X and Y. Let (X, Y) have a joint distribution H. Let $I(\cdot)$ stand for the usual indicator function. Define the concordance between a pair of vectors (x_1, x_2) and (y_1, y_2) as

$$\begin{aligned} C[(x_1, x_2, (y_1, y_2)] &= I[(x_2 - x_1)(y_2 - y_1) > 0] \\ &\quad - I[(x_2 - x_1)(y_2 - y_1) < 0] \end{aligned}$$

which is equal to 1 if $(x_2 - x_1)$ and $(y_2 - y_1)$ are of the same sign, -1 if they are of opposite signs and 0 if either $x_1 = x_2$ or $y_1 = y_2$. The population value of τ is defined by

$$\tau = E\{C[(X_1, X_2), (Y_1, Y_2)]\} \quad [4]$$

Note that if X and Y are independent then $\tau = 0$. Consider a sample $(X_n, Y_n) = [(X_1, Y_1), \ldots, (X_n, Y_n)]$ which is iid with common distribution function $H(x, y)$. A natural sample estimate of τ_n of τ is given by

$$\tau_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} C\left[\left(X_i - X_j\right), \left(Y_i - Y_j\right)\right] \quad [5]$$

The above definition can be simplified if there are no ties in the data. The statistic τ_n can also be recognized to be a U-statistic and so one can derive its asymptotic distribution, see for instance Hoeffding (1948), under both the null and alternative hypotheses. However, the asymptotic distribution of τ_n depends on the joint distribution $H(x, y)$ and the variance should be estimated to produce confidence intervals for τ. We give below the details of a simple use of the bootstrap to solve this problem and illustrate it with a numerical example taken from Hollander and Wolfe (1999: Section 8.1).

Let $H_n(x, y)$ be a joint distribution that places mass $1/n$ at each of the observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. This H_n is the empirical distribution function of the bi-variate sample. Let $(X_n^*, Y_n^*) = [(X_1^*, Y_1^*), \ldots, X_n^*, Y_n^*]$ be a sample of size n from H_n and let τ_n^* be the τ-statistic computed from this sample. By repeating this K times, one gets K values of τ_n^*. Let a_n, b_n be the $B\alpha$-largest and $B(1 - \alpha)$-largest values of these K values of τ_n^* arranged in increasing order. The interval $[a_n, b_n]$ is an approximate $1 - 2\alpha$ confidence interval for τ. Standard theorems show that the confidence level of this confidence interval converges to $1 - 2\alpha$ as $K, n \to \infty$.

Rasekh et al. (1970) wish to evaluate the relative importance of factors contributing to tuna quality, in particular the importance of the Hunter L value factor. A small subset of the data is given in **Table 1** (see also Hollander and Wolfe, 1999).

The value of τ_n from this sample is $4/9 = 0.444\,4$ and this is an estimate of the population value of τ. This shows that there seems to be a positive dependence between the Hunter L value and tuna quality. A confidence interval for τ can now be obtained by the use of the bootstrap. One thousand samples were generated by the bootstrap sampling scheme from the given data using a program written in the statistical package S-Plus. The 90% confidence interval for τ was found to be $[-0.027\,8, 0.750\,0]$, the interval containing 90% of the 1000 values of τ_n generated by the bootstrap. Since zero falls in this interval, we cannot rule out that the Hunter L value has no influence on tuna quality, if we use a 90% confidence interval as a two-sided test of independence with type-I error probability approximately 10%.

Table 1 Hunter *L* values and consumer panel scores for nine lots of canned tuna

Lot	*Hunter* L *value (X)*	*Panel score (Y)*
1	44.4	2.6
2	45.9	3.1
3	41.9	2.5
4	53.3	5.0
5	44.7	3.6
6	44.1	4.0
7	50.7	5.2
8	45.2	2.8
9	60.1	3.8

Source: Rasekh, J.A., Kramer, A., Finch, R., 1970. Objective evaluation of canned tuna sensory quality. Journal of Food Science 35, 417–423.

Testing Independence

Suppose that we want to test that two random variables X and Y are independent. Consider a sample $(X_n, Y_n) = [X_1, Y_1, \ldots, (X_n, Y_n)]$ which is iid with common distribution function $H(x, y)$ with marginal distribution functions $F(x)$ and $G(y)$. We wish to test the hypothesis of independence, namely that

$$H(x, y) = F(x)\, G(y) \quad [6]$$

for all x, y. One may restate this problem as asking whether the joint distribution given by taking the marginals to be independent fits the actual distribution very closely; in this sense this problem can be considered as a goodness-of-fit problem. A standard test is to reject for large values of the statistic

$$T_n = T_n(X_n, Y_n) = \sup_{x,y} |H_n(x, y) - F_n(x) G_n(y)| \quad [7]$$

where $H_n(x, y)$, $F_n(x)$, $G_n(y)$ are the joint empirical and marginal empirical distributions. How should we choose the critical value for T_n?

The statistic T_n is distribution-free under the null hypothesis, if the distributions are continuous. The bootstrap implementation, which is valid even when the distributions are not continuous, proceeds as follows. Notice that $F_n(x)$ and $G_n(y)$ are the sample estimates of the marginal distribution functions. Under the null hypothesis, $F_n(x)G_n(y)$ will be an estimate of the joint distribution function. Thus one collects a sample $(X_n^*, Y_n^*) = [(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)]$ from the distribution $F_n(x)$ $G_n(y)$ and computes the statistic $T_n^* = T_n(X_n^*, Y_n^*)$. This is repeated K times. Let $c_n(\alpha)$ be the $K\alpha$-th largest value of these T_n^*s arranged in increasing order. This $c_n(\alpha)$ is taken to be the critical value at level α and the null hypothesis is rejected if T_n is larger than $c_n(\alpha)$. This implementation of the bootstrap is described and illustrated in two articles by Romano (1988, 1989), where it is shown that this bootstrap test is consistent, if $K, n \to \infty$.

Table 2 gives the data considered by Clark et al. (1961) in a study concerned with the relationship of scores on various psychological tests and certain physical characteristics of 13 nonidentical male twins (see Hollander and Wolfe, 1999: Section 8.1).

The value of the T_n given by eqn [7] for the data is 0.177 5. The bivariate distribution in which X and Y are independent and have the marginal distributions given by the above data was constructed. One thousand bivariate samples were generated from this distribution and the 5% upper cut-off point of the T_n statistic calculated from these 1000 samples were found to be 0.189 3. This shows that the scores of the two twins can be concluded to be independent at the 5% level.

Table 2 Psychological test scores of nonidentical male twins

Pair no	*First twin*	*Second twin*
1	277	256
2	169	118
3	157	137
4	139	144
5	108	146
6	213	221
7	232	184
8	229	188
9	114	97
10	232	231
11	161	114
12	149	187
13	128	230

Source: Clark, P.J., Vandenberg, S.G., Proctor, C.H., 1961. On the relationship of scores on certain psychological tests with a number of anthropomorphic characters and birth order twins. Human Biology 33, 163–180.

Gibbs Sampling

Gibbs sampling is a computational method to calculate distributions which are intractable by mathematical means. Suppose that we wish to obtain samples from a joint distribution H of a bivariate random variable (X, Y) or just want to evaluate $\int g(x, y)\mathrm{d}H(x, y)$for some function $g(x, y)$. This may be difficult because the joint distribution H is quite complicated or only indirectly defined. Suppose that the conditional distributions $Q(x|y) = P(X \leq x|Y = y)$ and $R(y|x) = P(Y \leq y|X = x)$ are known sufficiently well enough to generate observations from them. Then we can proceed as follows. Let (X_0, Y_0) be arbitrary. Generate X_n. from the distribution $Q(\bullet|Y_{n-1})$ and Y_n from the distribution $R(\cdot|X_n)$, $n = 1, \ldots, N$. One can check that $(X_0, Y_0), (X_1, Y_1), \ldots$, is a Markov chain with a transition function for which H is an invariant distribution. If this transition function satisfies some mild conditions, then this Markov chain converges in one of several senses to H, and so, for large N, one can take

$$\frac{1}{N}\sum I[(X_i, Y_i) \in A] \quad \text{and} \quad \frac{1}{N}\sum g(X_i, Y_i)$$

to be approximations to $H(A)$ and $\int g(x, y)\ \mathrm{d}H(x, y)$. If the convergence is stronger, then one can take (X_N, Y_N), as approximately having distribution H. This is the essence of the Gibbs sampler of which there are many other variations. See Gelfand and Smith (1990), Tierney (1994) (for more details *see* Monte Carlo Methods and Bayesian Computation: Overview).

We will now illustrate the Gibbs sampler in a nonparametric censoring problem. Suppose that, given F and G, $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ are independent with $X_1, \ldots, X_n$ iid with common distribution F and with $Y_1, \ldots, Y_n$ iid with common distribution G. We will assume that the distribution functions F and G are continuous functions. Suppose that we observe only $Z_i = \min(X_i, Y_i)$, $\delta_i = I(X_i \leq Y_i)$, $i = 1, \ldots, n$; Z_i is the censored value of X_i if $\delta_i = 0$ and the actual value of X_i if $\delta_i = 1$.That is, sometimes we observe an X_i and at other times we just know that $X_i \in (Y_i, \infty)$. The problem is to estimate $F(x)$ or the survival function $\overline{F}(x) = 1 - F(x)$. A Bayesian approach would be to put a prior on (F, G) and look at the conditional distribution of F given the data. Assume that under the prior distribution, F and G are independent and that F has a Dirichlet measure $\mathscr{D}(\alpha\mu)$. The distribution of G will be irrelevant and it can be shown in this Bayesian problem that we can consider the values of $Y_1, \ldots, Y_n$ as constants (see Lo, 1993 for instance). We can rearrange the sample so that $Z_1 = X_1, \ldots, Z_r = X_r, Z_{r+1} = Y_{r+1}, \ldots, Z_n = Y_n$ where r is the number of uncensored X's. The number of censored observations is $m = n - r$. This means the sample consists of $X_1, \ldots, X_r$ and the information $X_i \in A_i$, $i = r + 1, \ldots, n$, where $A_i = (Y_i, \infty)$, $i = r + 1, \ldots, n$. The Bayesian posterior distribution is the conditional distribution of F given the data. We will show now how Gibbs sampling can help us to simulate from this distribution after we describe earlier work on a closed form expression for the Bayes estimate of F.

The Bayes estimate of F is the expected value of F under this posterior distribution. This can be obtained without evaluating the complete posterior distribution as shown in Susarla and Van Ryzin (1976). Let $N_n(A) = \sum_1^n I(Z_i \in A)$ be the number of observations (censored or not) that fall in A. Let $Z_1^* < Z_2^* < \ldots < Z_m^*$ be the distinct values among the censored observations $Z_{r+1}, \ldots, Z_n$. Let λ_k be the number of censored observations equal to Z_k^*, $k = 1, \ldots, m$. Then the Bayes estimate of $\overline{F}(x)$ as obtained by Susarla and Van Ryzin (1976) is given by

$$\widehat{\overline{F}}_n(x) = \frac{\alpha\mu(x, \infty) + N_n(x, \infty)}{\alpha\mu(0, \infty) + n} \text{"times} \prod_{j=1}^{m}\left[\frac{\alpha\mu\left(\left[Z_j^*, \infty\right)\right) + N_n\left(\left[Z_i^*, \infty\right)\right)}{\alpha\mu\left(\left[Z_j^*, \infty\right)\right) + N_n\left(\left[Z_j^*, \infty\right)\right) - \lambda_j}\right] \quad [8]$$

for $Z_l^* \leq x < Z_{l+1}^*$, $l = 0, 1, \ldots, m$ where we set $Z_0^* = 0$, $Z_{m+1}^* = \infty$. This estimate converges to the standard Kaplan–Meier frequentist estimate (Kaplan and Meier, 1958) as $\alpha \to \infty$.

A full Bayesian analysis will require the posterior distribution of F. This can be accomplished in this problem through Gibbs sampling. This will also yield a way to obtain the posterior estimate of the survival function, by computational methods, which then can be compared with the exact estimate of Susarla and Van Ryzin given in eqn [8]. The conditional distribution of F given $X_1, \ldots, X_r$ (the observed X's) is $\mathscr{D}(\alpha + r, \mu_r)$ where μ_r is as defined in eqn [3] with $n = r$. Let F, $X_{r+1}, \ldots, X_n$ be such that F has distribution $\mathscr{D}(\alpha + r, \mu_r)$ and given F, let $X_{r+1}, \ldots, X_n$ be iid F. The posterior distribution is the conditional distribution of F given $X_i \in A_i$, $i = r + 1, \ldots, n$.

It is clear that we do not know the precise values of $X_{r+1}, \ldots, X_n$, but we know only that $X_i \in A_i$, $i = r + 1, \ldots, n$. The standard theory of Dirichlet distributions shows that the conditional distribution $Q(F|X_r, \ldots, X_n, \text{data})$ of F given $X_r, \ldots, X_n$ and the data is equal to $\mathscr{D}(\alpha + n, \mu_n)$ where μ_n is given by eqn [3]. We also know that the conditional distribution $R(X_{r+1}, \ldots, X_n|\ F, \text{data})$ of $X_{r+1}, \ldots, X_n$ given F and the data, is just that of independent random variables with distributions $F_{A_{r+1}}, \ldots, F_{A_n}$ where for any set A, F_A represents that distribution F restricted to the set A. Thus, in the joint distribution of F and $(X_{r+1}, \ldots, X_n)$ given the data, we know the two conditional distributions.

Table 3 Exact and computational Bayes estimates of $F(x)$ based on Kaplan and Meier Data ($\alpha = 8, \mu = exponential with scale parameter 0.12)

x	0.80	1.00	2.70	3.10	5.40	7.00	9.20	12.10
Exact	0.108 3	0.119 0	0.207 1	0.300 6	0.471 9	0.525 6	0.682 3	0.750 1
Gibbs	0.108 3	0.119 0	0.208 4	0.301 1	0.470 6	0.526 1	0.680 2	0.747 6
KM	0.125 0	0.125 0	0.125 0	0.300 0	0.475 0	0.475 0	0.737 5	0.737 5

Source: Doss, H., 1994. Bayesian nonparametric estimation for incomplete data via successive substitution sampling. Annals of Statistics 22, 1763–1786.

This provides the necessary parts to carry out Gibbs sampling. After simulating an observation $(F, X_{r+1}, \ldots, X_n)$ from the joint distribution of F and $(X_{r+1}, \ldots, X_n)$ given the data, we can retain just the F which then can be considered as an observation from the posterior distribution.

There is a special difficulty in this example of Gibbs sampling. How does one generate a distribution F from a Dirichlet measure? This F is not a simple as a real valued random variable. The constructive definition of a Dirichlet measure given in Sethuraman (1994) will be useful here. Suppose that at stage 0 of Gibbs sampling, we start with arbitrary or well chosen $F^0, X^0_{r+1}, \ldots, X^0_n$. Let $\mu^0_n = \frac{1}{n}[\alpha\mu + \sum_1^r \delta_{X_i} \sum_{r+1}^n \delta_{X^0_i}]$. Then F^1 has to be generated from $\mathscr{D}(\alpha + n, \mu^0_n)$. Thus we may take $F^1(x) \sum_k^\infty p_k I(W_i \le x)$, where $W_1, W_2, \ldots$ are iid μ^0_n and the p_ks are given by eqn [1] based on independent $\theta_1, \theta_2, \ldots$ which are iid Beta $(1, \alpha + n)$. This is an infinite sum and cannot be computed exactly. In fact, we will avoid generating this F^1 in a precise fashion and go to the next step of generating $(X^1_{r+1}, \ldots, X^1_n)$ as iid observations from this F^1 and satisfying the restrictions $X^1_i \in A_i,\ i = r+1, \ldots, n$. The constructive definition of Sethuraman (1994) states that $X^1_{r+1}, \ldots, X^1_n$ can be taken to be $W_{I_{r+1}}, \ldots, W_{I_n}$ where $I_{r+1}, \ldots, I_n$ are drawn iid from the random discrete distribution $\{p_k, k = 1, \ldots\}$. Thus one first generates a sufficiently large vector $W_1, \ldots, W_N$ which are iid μ^0_n. The random index I_{r+1} is then generated as min $\{j : \sum_1^j p_i \ge U\}$, where U is a uniform random variable on [0, 1]. This means that we will need to generate only a finite (though random) number of p_is to generate I_{r+1}. If the value $X^1_{r+1} = W_{I_{r+1}}$ is not in A_{r+1} we will repeat this step until $X^1_{r+1} \in A_{r+1}$. This will give us X^1_{r+1}. We continue until the remaining $X^1_{r+2}, \ldots, X^1_n$ are generated. In a similar fashion, we generate $(F^M, X^M_{r+1}, \ldots, X^M_n)$ for $M = 1, 2, \ldots, K$, noting that at each step we did not have to calculate the infinite series for F^M. The expected value of F^M, namely μ^M_n or the average of μ^M_n over several such runs, will be the computational Bayes estimate of μ. This computational method is described in Doss (1994). A result from Athreya et al. (1994) can be used to justify that the Gibbs sampler converges.

We illustrate the above with the following example taken from Doss (1994). Kaplan and Meier (1958) presented the following data on lifetimes on eight subjects in which four were not fully observed, that is, were censored (indicated with a+):

$$0.8,\ 1.0+,\ 2.7+,\ 3.1,\ 5.4,\ 7.0+,\ 9.2,\ 12.1+$$

The exact and computational Bayes estimates based on these data are very close to one another as can be seen from Table 3. For comparison purposes, the original standard frequentist estimate of Kaplan and Meier (KM) is also presented in the same table. As expected, the Bayesian estimate is smoother and varies over a wider range.

One can extend the above to the case where F is specified to come from a family $H(x - \theta)$, where H has a Dirichlet prior, and one can place an independent prior distribution on the real valued parameter θ. Doss (1994) extends the above Gibbs sampling method for this more general prior and obtains the posterior distribution and the Bayes estimate of θ. An illustrative numerical example can also be found in that paper.

The above constitute only a typical example of the use of Gibbs sampling in nonparametric Bayesian problems. There have been many interesting applications of Gibbs sampling; for some early references see Escobar (1994, 1995), and Escobar and West (1995).

Conclusion

Nonparametric methods have established themselves not just as competitors to parametric methods, but actually as being superior to them under certain circumstances, while still providing valid and robust procedures. Incorporating prior information and expert knowledge into nonparametric procedures becomes the next important advance in statistical methodology. In the first section we described various ways of defining Dirichlet measures which can serve as prior distributions in nonparametric problems. It turns out that the actual Bayesian solution to a particular nonparametric problem may not be in a simple closed form and therefore may require some intensive computing. The bootstrap method and Gibbs sampling represent two such computational procedures. We have presented two examples of the bootstrap: to test for concordance and to test for independence, and have illustrated the Gibbs sampler to estimate the survival distribution when censoring is present.

See also: Bayesian Statistics.

Bibliography

Athreya, K.B., Doss, H., Sethuraman, J., 1994. On the convergence of the Markov chain simulation method. Annals of Statistics 24, 69–100.

Bickel, P.J., Freedman, D.A., 1981. Some asymptotic theory for the bootstrap. Annals of Statistics 9, 1196–1217.

Blackwell, D., 1973. Discreteness of Ferguson selections. Annals of Statistics 1, 356–358.

Blackwell, D., MacQueen, J.B., 1973. Ferguson distributions via Pólya schemes. Annals of Statistics 1, 353–355.

Clark, P.J., Vandenberg, S.G., Proctor, C.H., 1961. On the relationship of scores on certain psychological tests with a number of anthropomorphic characters and birth order twins. Human Biology 33, 163–180.

Dalal, S.R., 1979. Dirichlet invariant processes and applications to nonparametric estimation of symmetric distributions functions. Stochastic Processes and Their Applications 9, 99–107.

Doss, H., 1994. Bayesian nonparametric estimation for incomplete data via successive substitution sampling. Annals of Statistics 22, 1763–1786.

Dykstra, R.L., Laud, P., 1981. A Bayesian nonparametric approach to reliability. Annals of Statistics 9, 356–367.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Annals of Statistics 7, 1–26.

Escobar, M.D., 1994. Estimating normal means with a Dirichlet process prior. Journal of the American Statistical Association 89, 268–277.

Escobar, M.D., 1995. Nonparametric Bayesian methods in hierarchical models. Journal of Statistical Planning and Inference 43, 97–106.

Escobar, M.D., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90, 577–588.

Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. Annals of Statistics 1, 209–230.

Gelfand, A.E., Smith, A.F.M., 1990. Sampling based approaches to calculating marginal densities. Journal of the American Statistical Association 85, 398–409.

Hoeffding, W., 1948. A class of statistics with asymptotically normal distribution. Annals of Mathematic Statistics 19, 293–325.

Hollander, M., Korwar, R.M., 1982. Nonparametric Bayesian estimation of the horizontal distance between two populations. In: Gnedenko, B.V., Puri, M.L., vincze, I. (Eds.), Nonparametric Statistical Inference. North Holland, New york, pp. 409–415.

Hollander, M., Wolfe, D.A., 1999. Nonparametric Statistical Methods, second ed. Wiley, New York.

Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53, 457–481.

Lo, A.Y., 1981. Bayesian nonparametric statistical inference for shock models and wear processes. Scandinavian Journal of Statistics 8, 237–242.

Lo, A.Y., 1993. A Bayesian bootstrap for censored data. Annals of Statistics 21, 100–123.

Lu, H.H.S., Wells, M.T., Tiwari, R.C., 1994. Inference for shift functions in the two-sample problem with right-censored data: with applications. Journal of the American Statistical Association 89, 1017–1026.

Mauldin, R.D., Sudderth, W.D., Williams, S.C., 1992. Pólya trees and random distributions. Annals of Statistics 20, 1203–1221.

Meyn, S.P., Tweedie, R.L., 1993. Markov Chains and Stochastic Stability. Springer-Verlag, New York.

Nummelin, E., 1984. General Irreducible Markov Chains and Non-negative Operators. Cambridge University Press, Cambridge, UK.

Rasekh, J.A., Kramer, A., Finch, R., 1970. Objective evaluation of canned tuna sensory quality. Journal of Food Science 35, 417–423.

Revuz, D., 1984. Markov Chains, second ed. North Holland, Amsterdam.

Romano, J.P., 1988. A bootstrap revival of some nonparametric distance tests. Journal of the American Statistical Association 83, 698–708.

Romano, J.P., 1989. Bootstrap and randomization tests of some nonparametric hypotheses. Annals of Statistics 17, 141–159.

Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Statistica Sinica 4, 639–650.

Sethuraman, J., Tiwari, R.C., 1982. Convergence of Dirichlet measures and the interpretation of their parameter. In: Gupta, S.S., Berger, J.O. (Eds.), Statistical Decision Theory and Related Topics III, vol. 2. Academic Press, New York, pp. 305–315.

Singh, K., 1981. On the asymptotic accuracy of Efron's bootstrap. Annals of Statistics 9, 1187–1195.

Susarla, V., Van Ryzin, J., 1976. Nonparametric Bayesian estimation of survival curves from incomplete observations. Journal of the American Statistical Association 71, 897–902.

Tierney, L., 1994. Markov chains for exploring posterior distributions. Annals of Statistics 22, 1701–1762.

Wells, M.T., Tiwari, R.C., 1989. Bayesian quantile plots and statistical inference for nonlinear models in the two sample case with incomplete data. Common Statistical Theory Methods 18, 2955–2964.

Yamato, H., 1975. A Bayesian estimation of a measure of the difference between two continuous distributions. Report of the Faculty of Science at Kagoshima University 8, 29–38.

Zalkikar, J.N., Tiwari, R.C., Jammalamadaka, S.R., 1986. Bayes and empirical Bayes estimation of the probability that $Z > X + Y$. Common Statistical Theory Methods 15 (10), 3079–3101.

Nonparametric Statistics: Asymptotics

Kei Takeuchi, Meiji Gakuin University, Yokohama, Japan

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 16, pp. 10667–10673, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Nonparametric methods are procedures of statistical inference for models without precise assumptions about the distribution of the sample. A typical case is the two-sample test problem, where samples from two populations are observed, and the hypothesis that two populations have the same distribution is tested. Various types of test procedures have been proposed, their distributions under the hypothesis are obtained from the permutation distribution, that is, the conditional distribution given the set of values of the two samples combined. In most cases, small sample exact distribution is complicated to compute, but asymptotic approximations based on the central limit theorem for the permutation distribution can be applied. Also, power of the tests can be calculated from the normal approximation of the test statistic under the contiguous alternatives, and the relative efficiency of the test is defined and calculated. There are classes of tests with high relative efficiency. Also optimal parametric test can be regarded as approximate nonparametric tests in this context. Similar theory obtains for other classes of nonparametric tests and for interval estimation procedures derived from them.

Nonparametric methods are methods of statistical inference without precise assumptions on the shape of probability distribution of the observations. More precisely, a statistical model where the joint probability distribution is determined by a finite set of real parameters, or a finite dimensional real vector is called parametric, and methods of statistical inference without assuming a parametric model, is called nonparametric. Nonparametric methods, and especially nonparametric tests, have a long history. They can be traced back to the nineteenth century. Before 1940, rank correlation, rank sum tests, and permutation tests were known. However, they were often considered as merely 'quick and simple methods, substituting for more exact' procedures like t-tests, analysis of variance, etc., which were thought to be tedious.

The true nature of nonparametric methods has come to be understood since the 1950s and some of the nonparametric procedures have been found to have not only validity irrespective of the shape of underlying distribution, but also to have rather high efficiency under standard cases such as the normal distribution. Hence they have a high degree of robustness. Although the computation of the distribution of nonparametric statistics is now not difficult with modern computers, small sample distributions are usually complicated. Though they may not be difficult to compute for specific cases, it is not easy to understand their general features. Hence asymptotic theory is called for, and gives the result not for only null cases but also for alternative cases, leading to the definition and calculation of the relative efficiency of the procedures.

The term nonparametric is sometimes used in a different sense. Thus nonparametric regression usually means not that the distribution of the error is unspecified, but that the form of the regression function is not specified by a finite number of parameters. However, such problems belong to a different area of theory and hence are not dealt with in this article. Basic aspects of the asymptotic nonparametric theory are expounded for some basic cases. There are two main nonparametric procedures: (a) testing hypotheses, and (b) interval estimation.

Two-Sample Tests

Among nonparametric procedures, the most common is the nonparametric hypothesis testing, where the hypothesis to be tested is expressed in a nonparametric framework. Simplest and most often appearing in practical applications is the so-called two-sample problem where it is required to test the equality of two probability distributions. More precisely, we have two sets of sample observations $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$ where X_is and Y_js are real random variables distributed according to continuous distributions, F and G respectively, and it is required to test the hypothesis that the two distributions are identical, that is

$$H : F \equiv G$$

A nonparametric test function φ is a measurable function of $\mathbf{X}$ and $\mathbf{Y}$, satisfying the conditions that

$$0 \leq \phi(\mathbf{X}, \mathbf{Y}) \leq 1,$$

and

$$E(\phi(\mathbf{X}, \mathbf{Y}) | F \equiv G) \leq \alpha \quad \text{for all possible } F$$

When $E(\varphi | F \equiv G)$, which is called the size of the test, is always equal to α, the test is called a similar, or exactly nonparametric test.

When F is assumed to be continuous, there is a general method to derive similar tests. We define the order statistics $\tilde{X} = (X_{(1)} < X_{(2)} < \ldots < X_{(n)})$ and $\tilde{Y} = (Y_{(1)} < Y_{(2)} < \ldots < Y_{(m)})$ where $X_{(i)}$s are values of X_is arranged according to their magnitudes, and $Y_{(j)}$'s are ordered values of Y_js. Given $\tilde{X}$ and$\tilde{Y}$, the conditional distributions of X and Y are independent of F and G. Hence the pair of $\tilde{X}$ and $\tilde{Y}$ is a sufficient statistic, and we may consider φ as the function of $\tilde{X}$ and $\tilde{Y}$ (Fraser, 1957, Lehmann, 1959). Now we also define the combined order statistics $Z=(Z_1<Z_2<\ldots<Z_{n+m})$ where Z_ks are either one of the X_is or

one of the Y_js, thus $\tilde{Z}$ is the set of values of $\tilde{X}$ and $\tilde{Y}$ combined and ordered. Then we have

$$P\{\tilde{X} = (Z_{i_1}, Z_{i_2}, \ldots, Z_{i_n}) | \tilde{Z} = (Z_1, Z_2, \ldots, Z_{n+m})\} = \frac{1}{{}_{n+m}C_n}$$

that is the conditional probability that the values of $\tilde{X}$ being equal to any combination of n values in $\tilde{Z}$ is always the same and equal to $(1/{}_{n+m}C_n)$, whatever the probability distribution $F \equiv G$ be, provided only that F is continuous so that no two values of Xs and Ys are equal.

Then if the test function satisfies the condition that

$$E(\phi(\tilde{X}, \tilde{Y}) | \tilde{Z}) \equiv \alpha \qquad [1]$$

that is, the conditional expectation of α given the combined order statistic $\tilde{Z}$ is always equal to α under the hypothesis, it follows that

$$E(\phi(\tilde{X}, \tilde{Y}) | F \equiv G) = \alpha \quad \text{for all } F = G \qquad [2]$$

hence φ is a similar test (Lehmann, 1959). Conversely if φ is similar, it is proved that (2) must hold for almost all values of $\tilde{Z}$.

Test Statistic

Various similar test functions can be constructed in the following manner. Let t be a function of n values depending on the combined order statistics $\tilde{Z}$. Then the conditional distribution given $\tilde{Z}$ of $T = t(\tilde{x})$ can be computed from (2). Choose a constant c and also $0 \le r \le 1$, and define

$$\phi(\tilde{x}) = \begin{cases} 1 & \text{if } t(\tilde{x}) > c \\ \mathrm{r} & \text{if } t(\tilde{x}) = c \end{cases}$$

so that $E(\phi(\hat{x}) | F \equiv G) = \alpha$. Thus we get a similar nonparametric test procedure.

There are two types of such functions. The first is a function whose marginal distribution is independent of the distribution F under the hypothesis. The rank of $x_i = x_i$ among $\tilde{Z} = (Z_1, Z_2, \ldots, Z_{n+m}) Z_1 < Z_2 < \cdots < Z_{n+m}$ is defined to be R_i if x_i is equal to Z_{Ri}. If t depends only on the ranks of X_is the distribution of $T = t(X) = t(R_1, R_2, \ldots, R_n)$ is clearly independent of F. By proper choice of t, most common nonparametric tests are obtained. The simplest one is the rank sum $T = \sum_{i+1}^{n} R_i$, and the hypothesis is rejected when T is either too large or too small.

The weighted rank or rank score sum test is a generalization of the rank sum test, defined by the criterion function $T = \sum_{i=1}^{n+m} c(R)_i$ where $c(R)$ is a function of the rank R. The score function $c(R)$ is often defined to be equal to the expectation of the R-th order statistic on the sample of size $n+m$ from some distribution. For example, one can take $c(R) = \xi_{R|n+m}$, the expectation of the R-th order statistics in the sample from the standerd normal distribution, or $c(R) = \frac{1}{R}$, the same from the exponental distribution. A more simple function is

$$c(R) = \begin{cases} +1 & \text{for } R \gtrless \dfrac{n+m+1}{2} \\ 0 & \text{for } R = \dfrac{n+m+1}{2} \end{cases}$$

which is called the median test. Sometimes a non-monotone function is used such as $c(R) = |\xi_{R(n+m)}|$.

Consider any statistic $T = t(X_1, X_2, \ldots, X_n)$. Its conditional distribution given $\tilde{Z}$ is called the permutation distribution of T. Then a nonparametric test can be obtained by using T as test criterion with the critical point calculated from the permutation distribution.

Using the idea of permutation tests every parametric test statistic can also be used as the second type of nonparametric test criterion. For example, if F and G are both normal distributions with possibly different means but with common variance, the best (in various senses) test statistic is the t-statistic defined as

$$T = \sqrt{\frac{nm}{n+m}} \frac{\tilde{X} - \tilde{Y}}{S}$$

where

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2 + \sum_{j=1}^{m} (Y_i - \overline{Y})}{(n+m-2)}$$

T is distributed according to the t-distribution with degrees of freedom $n+m-2$ under the normal hypothesis. But we can also consider the permutation distribution of T given $\tilde{Z}$. It can be shown that given $\tilde{Z}$, the statistic T is equivalent to $\tilde{X}$, the mean of the first sample.

Asymptotic Distribution Under the Hypothesis

The conditional distribution of any statistic $t(\tilde{X})$ given $\tilde{Z}$ is difficult to calculate in general, and so its asymptotic approximation plays an important role. A basic result under the hypothesis is the following (Fraser, 1957).

Theorem: Under some regularity conditions, the conditional distribution of $T = (1/\sqrt{nv})(\overline{X} - \mu)$ approaches the standard normal distribution given $\tilde{Z} = (Z_1, Z_2, \ldots, Z_{n+m})$ when n and m tends to be large where

$$\mu = \frac{1}{n+m} \sum_{i=1}^{n+m} Z_i, \quad v^2 = \frac{\sum_{i=1}^{n+m} (Z_t - m)^2}{n+m-1}$$

Here the regularity condition is troublesome to express rigorously, but it requires basically that all moments of Z_i relative to v are bounded.

The theorem is readily generalized to the case when instead of $\tilde{X}$

$$\overline{g} = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

is used. Here g is a function which may depend on $\tilde{Z}$. Then the asymptotic normality of the null distribution of rank sum statistics and various permutation statistics can be derived, and asymptotic critical values for test statistics are obtained.

Thus for example, for the rank score sum test statistic,

$$T = \sum_{i=1}^{n} c(R_i)$$

$$E(T) = \frac{n}{n+m} \sum_{r=1}^{n+m} c(r) = \overline{nc(r)}$$

$$V(T) = \frac{nm}{(n+m)(n+m+1)} \sum_{i=1}^{n+m} \left(c(r) - \overline{c(r)}^2 \right)$$

and the approximate critical points can be obtained from the normal distribution. The asymptotically conditional distribution of the usual t statistic is equivalent to its permutation distribution, and the percentage point obtained from the t distribution approximately corresponds to that of the permutation distribution (Fisher, 1935).

This fact may be interpreted to establish that the common t-test obtained under the assumption of normality is also asymptotically nonparametric, in the sense that it has approximately correct size when the sample size is large, if the two populations are the same, even if not normal.

Asymptotic Power

Although nonparametric tests are introduced to guarantee the level of the test under general conditions, it is also important to consider their power under the alternatives, consideration of which is the most important factor in choosing among many nonparametric tests. Before computers were developed, some nonparametric tests were preferred to the parametric tests because of computational simplicity, but this is no longer an issue.

In order to calculate the power we have to specify the alternative. Assume that the density function of **X** is expressed as $f_\theta(x)$ and that of **Y** as $f_0(x)$. Then the hypothesis to be tested is $H{:}\theta = 0$. Under the alternative, conditional distribution of $\tilde{X} = (X_1, X_2, \ldots, X_n)$ given $\tilde{Z}$ is expressed as

$$\begin{aligned} p(X_1 = Z_{i_1}, X_2 = Z_{i_2}, \ldots, X_n = Z_{i_n} | \tilde{Z}) \\ = c_\theta(\tilde{Z}) \prod_{j=1}^{n} f_\theta\left(Z_{i_j}\right) \end{aligned}$$

where $c_\theta(\tilde{Z})$ is a constant depending on $\tilde{Z}$. The conditional distribution of permutation test statistic can be derived from this formula.

When θ is small, that is the distribution of F_θ is close to F_0, the distribution under the alternative can also be approximated by the normal distribution. More precisely, consider a sequence of alternatives $\theta_N = \frac{\xi}{\sqrt{N}}$, and samples with

$$n = [\lambda N] \quad m = [(1-\lambda)N]$$

Where

$$0 < \lambda < 1$$

is a constant. Then we have

Theorem: (Fraser, 1957, Nikitin, 1995) Under a set of regularity conditions, the conditional distribution given $\tilde{Z}$ of

$$T = \sqrt{\frac{N\lambda}{1-\lambda}} \frac{\tilde{X} - \mu}{v}$$

approaches the normal distribution with mean ψ and variance 1 where ψ is given by

$$\psi = \xi\sqrt{\lambda(1-\lambda)} \int x \frac{\partial}{\partial\theta} f(x,\theta)|_{\theta=0} \mathrm{d}x$$

Normal approximation for other types of statistics can also be derived. Under the same set up, the asymptotically best test statistic (assuming f is known) is distributed asymptotically normal with variance 1 and mean

$$\psi^* = \xi\sqrt{\lambda(1-\lambda)}\sqrt{I(\theta)}$$

when $I(\theta)$ is the Fisher information, that is

$$I(\theta) = \int \left(\frac{\partial}{\partial\theta} \log f(x,\theta)\right)^2 f(x,\theta)\mathrm{d}x$$

The ratio $\psi^2/\psi^*2{=}e$ is called the relative efficiency of the nonparametric test, since the nonparametric test based on the sample of size N is nearly equal (in terms of power) to the best test based on a sample of size eN.

For the two-sample problem the most common model is that of shift parameter, that is, the two distributions are related by a real parameter θ so that

$$F(x) \equiv G(x-\theta)$$

Then the hypothesis corresponds to $H{:}\theta = 0$. For the general rank score test if the score function $c(R)$ can expressed as

$$c(R) = J\left(\frac{R}{n+m+1}\right)$$

where J is a continuous and smooth function with the conditions

$$\int_0^1 J(u)\mathrm{d}u = 0, \quad \int_0^1 J^2(u)\mathrm{d}u = 1$$

then the ψ above can be expressed as

$$\psi = \theta\sqrt{\lambda(1-\lambda)} \int f'(x)J(F(x))\mathrm{d}x$$

For example for the rank sum test,

$$J(u) = \frac{1}{\sqrt{12}}\left(u - \frac{1}{2}\right)$$

and

$$\begin{aligned} \psi &= \theta\sqrt{\frac{\lambda(1-\lambda)}{12}} \int f'(x)\left(F(x) - \frac{1}{2}\right)\mathrm{d}x \\ &= \theta\sqrt{\frac{\lambda(1-\lambda)}{12}} \int f^2(x)\mathrm{d}x \end{aligned}$$

from which its relative efficiency for the normal case is calculated to be equal to $3/\pi = 95$ percent.

If J is chosen to be equal to

$$J(u) = c\frac{d}{du}h_j(u), \quad \text{where } h_f(u) = f(F^{-1}(\mu))$$

the asymptotic efficiency of the rank score test is 100 percent. Thus for any specific distribution there corresponds a nonparametric rank score test which has 100 percent efficiency (Lehmann, 1959).

It is also easily proved that the permutation test derived from the best test statistic under a specified distribution is asymptotically equivalent to the best test. Since the best nonparametric test is dependent on the unknown shape of F, there is no uniformly best nonparametric test. But some information about the shape of the distribution is available from the sample, which may be utilized to choose the nonparametric test among a class of tests.

For example, the best score function $J(u)$ can be approximately constructed from the sample information. Such a method is called an adaptive procedure. There have been

proposals and theoretical results about adaptive nonparametric procedures, and there exist adaptive procedures which are 100 percent asymptotically efficient for all distributions within a rather large class (Nikitin, 1995). But these are usually very complicated, and some of them have distributions notoriously slow to converge asymptotically. Hence they are not necessarily recommendable in practical situations.

Common Nonparametric Tests

There are various problems of testing statistical hypotheses, where several types of nonparametric tests are derived in similar ways, as in the two-sample case.

Equality of Several Populations

Suppose that we have k sets of samples, each of size n_i from the population with distribution F_i. The hypothesis to be tested is $H{:}F_i \equiv F$. Let $X_i{=}(X_i, X_{i2}, \ldots, X_{in})$ be the set of the values in the sample from the i-th population, and $\tilde{Z} = (X_1, X_2, \ldots, X_k)$ conditional distribution given $\tilde{Z}$ is expressed as the total set of values of the k samples combined. Then under the hypothesis the

$$P(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_k | \tilde{Z}) = \frac{n_1! n_2! \ldots n_k!}{N}$$

And nonparametric tests can be derived from this permutation distribution.

One class of such tests can be obtained from permutation distribution of the usual test criteria such as

$$F = \frac{\sum_{i=1}^{k} n_i (\overline{X}_i - \overline{\overline{X}})^2}{(k-1)S^2}, \quad S^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \overline{X}_i)^2}{(N-k)}$$

where

$$N = \sum_{i=1}^{k_i} n_i, \quad x_i = \sum_{j=1}^{n_i} X_{i,j}/n_i, \quad \overline{\overline{X}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{i,j}/N$$

or in case $n_j = n$

$$R = \max_{i,i'} \sqrt{n}(\overline{X}_i - \overline{X}_{i'})S$$

When n_is are large, $(k-1)F$ is distributed asymptotically according to the chi-square distribution with k-1 degrees of freedom and R has the same asymptotic distribution as the same as the normal studentized sample range (Randles and Wolfe, 1979).

Another class of criteria is obtained by substituting the rank score $c(R_{i,j})$ for $X_{i,j}$, where $R_{i,j}$ is the rank of $X_{i,j}$ in $\tilde{Z}$. Test criteria corresponding to the F test can be expressed as

$$T = \sum_{i=1}^{k} n_i (\overline{c}_i - \overline{\overline{c}})^2 / v^2$$

where

$$\overline{\overline{c}} = \frac{1}{n} \sum_{r=1}^{N} c(r), \quad v^2 = \frac{1}{N-1} \sum_{j=1}^{N} (c(r) - \overline{c})^2,$$

$$\overline{c}_i = \frac{1}{n} \sum_{j=1}^{n_i} c(R_{i,j})$$

The distribution of T can be approximated by the chi-square distribution. Under the alternative close to the hypothesis, the asymptotic distribution of T is expressed as a non-central chi-square distribution. More precisely, when the distribution F_i is expressed as $F_i(x) = F_{\theta i}(x)$ with real parameter and known function $F_\theta(x)$, the hypothesis expressed as $H{:}\theta_i \equiv \theta_0$, and with the sequence of samples of size $n_i = \overline{\lambda}_i N$, $\sum_{i=1}^{k} \lambda_i = 1$ under the sequence of alternatives $\theta_i = \theta_0 + \xi_i/\sqrt{N}$, the statistic T is distributed asymptotically as the non-central chi-square distribution with degrees of freedom k-1, and non-centrality $\psi = \sum_{i=1}^{k} \lambda_i \xi_i^2 \times \delta$. The constant δ depends both on the shape of the distribution and the score function $c(R)$. The relative efficiency of such tests can be defined as in the two-sample case, and with the same score function, the relative efficiency of the rank score square sum test is equal to that of the rank score test in the two-sample case (Lehmann, 1975).

Location of a Symmetric Distribution

Consider the case when $X_1, X_2, \ldots, X_n$ is a sample from a symmetric distribution centered at θ, i.e., its probability density function $f(x\text{-}\theta)$ is an even function $f(\text{-}x) = f(x)$, but otherwise is not specified. It is required to test the hypothesis $H{:}\theta{=}\theta_0$. Non-parametric test procedures can be obtained in the following way.

Define $Z_i = |X_i\text{-}\theta_0|$ and $\varepsilon_i = sgn(X_i{-}\theta_0)$. Then it is easily shown that under the hypothesis, ε_is are independent and $P(\varepsilon_i = \pm 1){=}1/2$. Let $\tilde{Z} = (Z_1, Z_2, \ldots, Z_n)$ be the set of values of Z_i. Then given $\tilde{Z}$, the conditional distribution of the statistic

$$T = t(X_1, X_2, \ldots, X_n) = t(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n, Z_1, Z_2, \ldots, Z_n)$$

is obtained. Statistics of the form $T = \sum_{i=1}^{n} \varepsilon_i c(R_i)$ have the mean and variance. Its conditional distribution can be approximated by the normal distribution when n is large.

Let R_i be the rank of Z_i. Then the test based on $T = \sum_{i=1}^{n} \varepsilon_i R_i$ is called the signed rank sum test, and more generally is called a signed rank score test statistic. The relative efficiency of such a test is defined can calculated in a completely similar way, as in the two-sample case.

Independence of Two Variables

Let (X_i, Y_i), $i = 1, 2, \ldots, n$ be a sample from a bivariate distribution. Let $F(x, y)$ be the joint distribution function. Consider the hypothesis that X and Y are independent, i.e. $F(x, y) \equiv G(x)H(y)$ assuming G and H are absolutely continuous but without any further specification.

Let $\tilde{X} = (X_1, X_2, \ldots, X_n)$ and $\tilde{Y} = (Y_1, Y_2, \ldots, Y_n)$ be the set of X-values and Y-values. Then under the hypothesis the conditional distribution of (X_i, Y_i), $i = 1, 2, \ldots, n$ given $\tilde{X} = (x_1, x_2, \ldots, x_n)$ and $\tilde{Y} = (y_1, y_2, \ldots, y_n)$ is expressed as

$$P = (X_1 = x_{i_1}, Y_1 = y_{i_1}, X_2 = x_{i_2}, Y_2 = y_{i_2}, \ldots, X_n = x_{i_n}, Y_n = y_{i_n}) = \frac{1}{n!^2}$$

for any permutation $(i_1, i_2, \ldots, i_n)$ and $(j_1, j_2, \ldots, j_n)$. This distribution is also called the permutation distribution.

Then we may define the generalized correlation coefficient

$$R = \frac{\sum_{i=1}^{n} \phi(X_i)\phi(Y_i)}{\sqrt{\sum_{i=1}^{n} (X_i)^2 \sum \phi(Y_i)}}$$

with the condition

$$\sum_{i=1}^{n} \phi(X_i) = 0, \quad \sum_{i=1}^{n} \phi(Y_i) = 0$$

Under the hypothesis given $\tilde{X}$ and $\tilde{Z}$

$$E(R) = 0 \quad V(R) = \frac{1}{n-1}$$

As n tends to infinity the distribution of R approaches the standard normal distribution (Kendall, 1948).

When $\varphi(X_i) = R_i$, R is called the rank correlation coefficient (or more precisely Spearman's ρ). When $\varphi(X_i) = X_i$, R is equal to the usual (moment) correlation coefficient. Hence it can also be interpreted as a nonparametric correlation coefficient if its permutation distribution is taken into consideration.

Equality of Multivariate Distributions

Multivariate two-sample problems can be treated in the same way as in the univariate case. Suppose that we want to test the equality of two bivariate distributions.

Let $X = \{(X_{1,1}, X_{1,2}), (X_{2,1}, X_{2,2}), \ldots, (X_{n,1}, X_{n,2})\}$ be the bivariate sample of size n from the first distribution, and $Y = \{(Y_{1,1}, Y_{1,2}), (Y_{2,1}, Y_{2,2}), \ldots, (Y_{m,1}, Y_{m,2})\}$ be the sample of size m from the second distribution. The hypothesis to be tested is that the two distributions are continuous and identical, but not otherwise specified.

Let $\tilde{Z}$ be the totality of the $n+m$ pairs of values of $\tilde{X}$ and $\tilde{Y}$. Then given $\tilde{Z}$, the conditional probability that the pairs in X are equal to the specific n pairs in $\tilde{Z}$ is equal to $1/_{n+m}C_n$ as in the univariate case. Non-parametric tests can be derived from this fact.

Define $T_1 = \sum g_1(X_{i,1})$ and $T_2 = g_2(X_{i,2})$. Then under the hypothesis the conditional distribution given $\tilde{Z}$ of (T_1, T_2) approaches a bivariate normal distribution as n and m get large (under a set of regularity conditions). Hence we can define

$$\chi^2 = s^{11}(T_1 - m_1)^2 + s^{12}(T_1 - m_1)(T_2 - m_2) + s^{22}(T_2 - m_2)^2$$

where

$$m_1 = E(T_1|\tilde{Z}), \quad m_2 = E(T_2|\tilde{Z})$$

and s^{11}, s^{12}, s^{22} are the elements of inverse of conditional variance and covariance matrix of T_1 and T_2. Then under the hypothesis χ^2 is asymptotically distributed as chi-square distribution of 2 degrees of freedom. By various choices of the function g_1, g_2, we can get bivariate versions of rank sum, rank score, etc., tests (Puri and Sen, 1971).

Nonparametric Confidence Intervals

Nonparametric test procedures can be applied to construct nonparametric confidence intervals. For example, consider the two-sample location shift model i.e., the two distributions are related as $F(x) = G(x-\theta)$. Suppose we want to construct a confidence interval $(l(X, Y), u(X, Y))$, such that

$$P\{l(X,Y) < 0 < u(X,Y\} \geq 1 - \alpha$$

for all θ and for all continuous G.

Now consider the hypothesis $\theta = \theta_0$. It is equivalent to the hypothesis that $X'_i = X_i\text{-}\theta_0$, $i = 1, 2, \ldots, n$ has the same distribution as Y_i. Hence any two-sample nonparametric test procedure can be applied to $(X'_1, X'_2, \ldots, X'_n)$ and $(Y_1, \ldots, Y_m)$. Let $\phi(\tilde{X}, \tilde{Y})$ be the test function. Then $E_{\theta_0}(\phi(\tilde{X}, \tilde{Y})) \leq \alpha$.

Now consider the class of hypotheses for any specific value of θ, and let the test functions corresponding to θ be $\phi_\theta(\tilde{X}, \tilde{Y})$. Corresponding to $\phi_\theta(\tilde{X}, \tilde{Y})$, let $S \subset R^1$ (real line) be the set of values of θ for which $\phi_\theta(\tilde{X}, \tilde{Y}) = 0$. Then for any θ we have $P_\theta(\theta \in S) = E_\theta(1 - P_\theta(\tilde{X}, \tilde{Y})) \geq 1 - \alpha$. Therefore S is a confidence region of θ with level $1-\alpha$, and if S is always defined as an interval, it gives a nonparametric confidence interval with level $1\text{-}\alpha$.

With a monotone increasing score function we can always construct such a confidence interval as follows:

Let $R_{\theta,i}$ be the rank of $X_i - \theta$, $i = 1, 2, \ldots, n$ in the combined set of values $X_i - \theta$, $i = 1, 2, \ldots, n$ and Y_j, $j = 1, 2, \ldots, m$. Let $c(R)$ be the monotone increasing score function. Then we can determine two constants a and b such that

$$P\left\{a \leq \sum_{i=1}^{n} c(R_{\theta,i}) \leq b\right\} \geq 1 - \alpha$$

Then define

$$\underline{\theta} = \inf\left\{\sum_{i=1}^{n} c(R_{\theta,i}) \leq b\right\}$$

$$\overline{\theta} = \sup\left(\sum_{i=1}^{n} c(R_{\theta,i}) \geq a\right)$$

Since $\sum_{i=1}^{n} c(R_{\theta,i})$ is monotone non-increasing in θ, $\underline{\theta} < \theta < \overline{\theta}$ gives a confidence interval of level $1-\alpha$.

Also from the consideration of the permutation distribution of $T = \sum_{i=1}^{n} X'_i$, we can construct a nonparametric confidence interval of the form

$$\overline{X} - \overline{Y} - D < \overline{\theta} < \overline{X} - \overline{Y} + D$$

When n and m are large, asymptotic results can be applied, and the above interval becomes nearly equivalent to the t-interval derived from the assumption of normality.

See also: Nonparametric Statistics: Advanced Computational Methods; Nonparametric Statistics: Rank-Based Methods; Nonparametric Statistics: Records.

Bibliography

Fisher, R.A., 1935. The Design of Experiments. Oliver and Boyd.
Fraser, D.A.S., 1957. Nonparametric Methods in Statistics. Wiley.
Hajek, J., 1969. A Course in Nonparametric Statistics. Holden-Day.
Hettmansperger, T.P., 1984. Statistical Inference Based on Ranks. Wiley.
Hollander, M., Wolfe, D.A., 1973. Nonparametric Statistical Methods. John and Sons.

Hong, Yongmiao, White, Halbert, 2005. Asymptotic Distribution Theory for Nonparametric Entropy Measures of Serial Dependence. Econometrica 73 (3), 837–901.
Kendall, M.G., 1948. Rank Correlation Methods. Griffin.
Lehmann, E.L., 1959. Testing Statistical Hypotheses. Wiley.
Lehmann, E.L., 1975. Nonparametrics: Statistical Methods Based on Ranks. Holden-Day.
Masry, Elias, 2005. Nonparametric Regression Estimation for Dependent Functional Data: Asymptotic Normality. Stochastic Processes and their Applications 115 (1), 155–177.
Nikitin, Y., 1995. Asymptotic Efficiency of Nonparametric Tests. Cambridge University Press, Cambridge, UK.
Pratt, J.W., Gibbons, J.D., 1981. Concepts of Nonparametric Theory. Springer-Verlag.
Puri, M.L., Sen, P.K., 1971. Nonparametric Methods in Multivariate Analysis. John Wiley and Sons.
Randles, R.H., Wolfe, D.A., 1979. Introduction to the Theory of Nonparametric Statistics. Wiley and Sons.
Tsybakov, Alexandre B., 2009. Introduction to nonparametric estimation. Springer, New York.
Walsh, J.E., 1965. Results for Two and Several Sample Problems, Symmetry, and Extremes. In: Handbook of Nonparametric Statistics. Van Nostrand.
Wang, Qiying, Phillips, Peter CB., 2009. Asymptotic Theory for Local Time Density Estimation and Nonparametric Cointegrating Regression. Econometric Theory 25 (03), 710–738.

Nonparametric Statistics: Rank-Based Methods

Myles Hollander and Jayaram Sethuraman, Florida State University, Tallahasse, FL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 10673–10680, © 2001, Elsevier Ltd.

Abstract

From its early beginnings in the early 1900s, nonparametric statistics has been an active area that has stimulated elegant statistical theory and beautiful applications. Its development continues at a vigorous pace today. There are many reasons for the success of the nonparametric approach. Nonparametric methods have wide applicability and typically require only modest assumptions concerning the underlying populations from which the data are drawn. Under these mild assumptions, exact hypothesis tests, exact confidence intervals, exact multiple comparison procedures, and exact confidence bands can be obtained. Nonparametric methods have excellent efficiency properties with respect to their parametric competitors and are also robust in the sense that they are relatively insensitive to outlying observations and departures from the model. This article highlights some classical rank-based nonparametric methods in two- and κ-sample location problems. These problems and their generalizations were central to the early development of non-parametric methods and remain extremely useful for data analysis today.

A nonparametric statistical procedure is a statistical procedure, such as (1) a hypothesis test, (2) a confidence interval, (3) a multiple comparison procedure, (4) a confidence band, that has desirable properties that hold under mild assumptions concerning the underlying population from which the data are drawn. Typically, it suffices to assume the underlying population is continuous.

The term 'distribution-free' has a more precise meaning than the term 'nonparametric.' For example, in the context of hypothesis testing, a distribution-free statistic is one whose probability distribution, under the null hypothesis being tested, is independent of the underlying population distribution. In the context of interval estimates, a distribution-free $100(1-\alpha)\%$ confidence interval, say (Δ_L, Δ_U) for an unknown parameter Δ, is a random interval having the property that the probability that Δ is contained in the random interval is $1-\alpha$, independent of the underlying population distribution. There are related concepts for distribution-free multiple comparison procedures (see Section The κ-Sample Problem) and distribution-free confidence bands for distribution functions (see Section 11.5 of Hollander and Wolfe, 1999).

The success of nonparametric statistics over the last 70 years and its continued vigorous development today are due to the wide applicability of the nonparametric approach and the excellent efficiency properties of nonparametric procedures. Some of the main advantages can be briefly summarized as follows:

1. Nonparametric procedures provide exact P-values for tests, exact coverage probabilities for confidence intervals, exact experimentwise error rates for multiple comparison procedures, and exact coverage probabilities for confidence bands. These exact properties do not require an assumption that the underlying population have a specific form, such as a normal population.
2. Nonparametric procedures have excellent efficiency properties with respect to their parametric competitors.
3. Nonparametric procedures are relatively insensitive to outlying observations in the sense that a few wildly big (or small) observations that are far removed from their fellow observations have little influence on the procedure (test, confidence interval, multiple comparison procedure, etc.).

Section The k-sample Problem describes some basic aspects of two- and k-sample location problems. These problems and their generalizations drove the early development of nonparametric statistics, and the procedures developed remain extremely useful for data analysis today. More modern topics in nonparametrics include Bayesian and computationally intensive procedures. Section 1.7 of Hollander and Wolfe (1999) provides a brief history of nonparametric statistics and contains source references to work in other currently active areas including regression problems and survival analysis. Other recent books covering various aspects of nonparametric statistics include Sprent (1993), Gibbons (1997), and Hettmansperger and McKean (1998). Earlier applied books featuring nonparametric methods based on ranks include Noether (1967) and Lehmann (1975). Randles and Wolfe (1979) is an introduction to the theory of nonparametric statistics.

Rank-Based Procedures for Location Alternatives

The primary reasons for using ranks are to obtain exact procedures that enjoy a distribution-free property and to achieve robustness by diminishing the deleterious effects of outlying observations and departures from the model. One might think that the reduction to ranks from original data values would sacrifice too much information, but power and efficiency studies have shown that is not the case. Indeed the nonparametric rank-based procedures are only slightly less efficient than their normal theory competitors when the underlying populations are normal (the home court of normal theory methods), and can be much more efficient when the underlying populations are not normal. We begin our description of rank-based methods with procedures for the two-sample location problem, then consider the k-sample location problem and the

multivariate two-sample location problem. There is an analogous body of nonparametric techniques (which parallels the methods described in Section The k-sample Problem) for one-sample problems, paired-sample problems, randomized blocks, and multivariate one-sample problems. That body of procedures also enjoys excellent efficiency and robustness properties. See Hollander and Wolfe (1999) and Hettmansperger and McKean (1998) for coverage of these methods.

The Two-Sample Location Problem

A basic problem in statistics is the two-sample location problem. Assume that we have a random sample $X_1, \ldots, X_m$ from population 1 having distribution function $F(x) = P(X \le x)$ and an independent random sample $Y_1, \ldots, Y_n$ from population 2 having distribution function $G(x) = P(Y \le x)$. We wish to test the null hypothesis H_0 that the two populations are equivalent, that is,

$$H_0 : F(x) = G(x), \quad \text{for all } x \qquad [1]$$

If we are interested in location alternatives, these can be modeled using the location shift model

$$G(x) = F(x - \Delta), \quad \text{for all } x \qquad [2]$$

where the parameter Δ is commonly called the 'treatment' effect. Model [2] signifies that a randomly selected value Y from population 2 has the same distribution as $X + \Delta$ where X is a randomly selected value from population 1, that is $Y \overset{d}{=} X + \Delta$ where $\overset{d}{=}$ means 'has the same distribution as.' When $\Delta = 0$, H_0 is satisfied. In addition to testing H_0, it is useful to estimate Δ and give a confidence interval for Δ. An excellent way to test H_0, estimate Δ, and give a confidence interval for Δ is to use procedures based on (and derived from) the famous Wilcoxon (1945) two-sample rank sum statistic.

The Wilcoxon two-sample rank sum statistic is

$$W = \sum_{j=1}^{n} R(Y_j) \qquad [3]$$

where $R(Y_j)$ is the rank of Y_j in the joint ranking from least to greatest of the $N = m + nX$ and Y. Significantly large values of W indicate $\Delta > 0$; significantly small values of W indicate $\Delta < 0$. A one-sided upper-tail distribution-free test of H_0:$\Delta = 0$ versus H_1:$\Delta > 0$, at the α level of significance is,

$$\text{reject } H_0 \text{ if } W \ge w_\alpha; \quad \text{otherwise accept } H_0 \qquad [4]$$

where w_α is chosen to make the type-I error probability of the test equal to α and is the upper α^{th} percentile point of the null distribution of W.

Similarly, a one-sided lower-tail distribution-free test of H_0:$\Delta = 0$ versus H_2:$\Delta < 0$, at the $\acute{a}$ level of significance, is

$$\text{reject } H_0 \text{ if } W \le n(m+n+1) - w_\alpha; \quad \text{otherwise accept } H_0 \qquad [5]$$

A two-sided distribution-free test of H_0:$\Delta = 0$ versus H_3:$\Delta \ne 0$, at the α level of significance, is

$$\text{reject } H_0 \text{ if } W \ge w_{\alpha/2} \text{ or if } W \le n(m+n+1) - w_{\alpha/2}; \quad \text{otherwise accept } H_0 \qquad [6]$$

This is the two-sided symmetric test with $\alpha/2$ probability in each tail of the distribution.

The tests based on W are distribution-free under H_0 (eqn [1]) assuming the common value of F and G is a continuous distribution. For then, all $\binom{N}{n}$ possible assignments of the Y-ranks are equally likely, each having probability $1\backslash\binom{N}{n}$ when H_0 is true. Values of w_a can be obtained from Table A.6 of Hollander and Wolfe (1999). Exact p-values can also be obtained using the software StatXact (1999).

The large sample approximation to procedure [4] is

$$\text{reject } H_0 \text{ if } W^* \ge z_\alpha; \quad \text{otherwise accept } H_0 \qquad [7]$$

where

$$W^* = \frac{W - E(W)}{\{\text{var}_0(W)\}^{1/2}} = \frac{W - \{n(m+n+1)/2\}}{\{mn(m+n+1)/12\}^{1/2}} \qquad [8]$$

is the standardized version of W. It can be treated as an approximate $N(0, 1)$ random variable under H_0. In eqn [7], z_α is the upper α percentile point of an $N(0, 1)$ distribution.

If ties are present in the data, assign to tied observations the average of the ranks for which those observations are competing. Then procedure [4] remains applicable but it is now approximate rather than exact. Alternatively, one can use an exact conditional test based on the exact conditional randomization distribution of W, see Hollander and Wolfe (1999: Comment 4.5). When there are ties, the null mean $E_0(W)$ is unchanged but the null variance $\text{var}_0(W)$ is reduced to the right-hand-side of eqn [9]. Thus when applying the large sample approximation in the presence of ties, use average ranks and replace $\text{var}_0(W)$ in eqn [8] by

$$\text{var}_0(W) = \frac{mn}{12}\left(m + n + 1 - \frac{\sum_{j=1}^{g}(t_j - 1)t_j(t_j + 1)}{(m+n)(m+n-1)}\right) \qquad [9]$$

where g denotes the number of tied groups, t_j is the size of tied group j, and an untied observation is considered to be a tied group of size 1.

Thus far in this section we have considered testing H_0:$\Delta = 0$. To test $H_0'' : \Delta = \Delta_0$ where Δ_0 is some specified nonzero number, subtract Δ_0 from each Y to form $Y_1' = Y_1 - \Delta_0, \ldots, Y_n' = Y_n - \Delta_0$. Then compute W as the sum of the ranks of Y' in the joint ranking from least to greatest of the mX's and the nY''s. With W computed in this manner, the procedures described earlier in this section may be directly applied to test H_0''.

There are medical situations for comparing treatments when various restricted randomization designs are employed. In such situations, the test based on W will differ from that described here. See Hollander and Peña (1988) and Mehta et al. (1988) for details.

A point estimator of Δ, derived from the Wilcoxon rank sum statistic, is the Hodges–Lehmann (1963) estimator, namely

$$\hat{\Delta} = \text{median}\left\{Y_j - X_i, i = 1, \ldots, m, j = 1, \ldots, n\right\} \qquad [10]$$

That is, $\hat{\Delta}$ is the median of the $mn(Y - X)$ differences. Roughly speaking $\hat{\Delta}$ is the value $\hat{\Delta}$ that should be subtracted from each Y so that the value of W, when applied to the

aligned samples $X_1, \ldots, X_m, Y_1 - \hat{\Delta}, \ldots, Y_n - \hat{\Delta}$, is equal to $n(m+n+1)/2$. One can think of the method as estimating Δ by the amount $\hat{\Delta}$ the Y's should be shifted so that the sample $X_1, \ldots, X_m$ and the pseudo-sample $Y_1 - \hat{\Delta}, \ldots, Y_n - \hat{\Delta}$ look like two samples from the same distribution as viewed by W.

A symmetric two-sided $100(1-\alpha)\%$ distribution-free confidence interval for Δ, obtained by inverting the test statistic, is

$$\Delta_L = U^{(c_\alpha)}, \quad \Delta_U = U^{(mn+1-c_\alpha)} \qquad [11]$$

where

$$C_\alpha = \frac{n(2m+n+1)}{2} + 1 - w_{\alpha/2} \qquad [12]$$

$w_{\alpha/2}$ is the upper $\alpha/2$ percentile point of the null distribution of W, and $U^{(1)} \leq \cdots \leq U^{(mn)}$ denote the ordered values of the $mn(Y-X)$ differences. For large m and n, the integer C_α may be approximated by

$$C_\alpha \approx \frac{mn}{2} - z_{\alpha/2}\left(\frac{mn(m+n+1}{12}\right)^{1/2} \qquad [13]$$

The value C_α given by eqn [13] will typically not be an integer. To be conservative, when using the large sample approximation to the confidence interval, take C_α to be the largest integer that is less than or equal to the right-hand-side of eqn [13].

Example 1

We illustrate the two-sample Wilcoxon procedures with data from March et al. (1976) (see also Hollander and Wolfe, 1999) who examined differences between healthy and lead-poisoned Canadian geese. One of the measures they considered was plasma glucose (in mg/100 ml plasma). The data for eight healthy geese and seven lead-poisoned geese are given in **Table 1**. From **Table 1**, we find $W = 68$. Referring $W = 68$ to Table A.6 of Hollander and Wolfe (1999) with $n = 7$ and $m = 8$, we find $P_0(W \geq 68) = 0.095$. (In Table A.6, $n \leq m$, therefore designate the values corresponding to the smaller sample as Y.) That is, the probability, under the null hypothesis, of getting a value as larger or larger than 68, is 0.095. This is the p-value for the one-sided test of $\Delta = 0$ versus $\Delta > 0$. (StatXact software gives the p-value 0.094 6). Thus one would reject $\Delta = 0$ in favor of $\Delta > 0$ for all type-I error probabilities $\alpha \geq 0.095$ and accept $\Delta = 0$ for all type-I error probabilities $\alpha < 0.095$. Thus there is borderline evidence that the lead-poisoned geese have higher plasma glucose values.

Table 1 Plasma glucose values for healthy and lead-poisoned Canadian geese (ranks in parentheses)

Healthy geese	*Lead-poisoned geese*
297(8)	293(7)
340(12)	291(6)
325(10)	289(4)
227(1)	430(14)
277(3)	510(15)
337(11)	353(13)
250(2)	318(9)
290(5)	

Source: March, G.L., John, T.M., McKeown, B.A., Sileo, L., George, J.C. 1976. The effects of lead poisoning on various plasma constituents in the Canada goose. Journal of Wildlife Diseases 12, 14–19.

To compute the Hodges–Lehmann estimate $\hat{\Delta}$, we first obtain the ordered $Y - X$ differences $U^{(1)} \leq \cdots \leq U^{(56)}$. The ordered differences are: −51, −49, −48, −47, −46, −44, −36, −34, −32, −22, −19, −8, −7, −6, −4, −1, 1, 3, 12, 13, 14, 16, 16, 21, 28, 28, 39, 41, 41, 43, 56, 62, 63, 64, 66, 68, 76, 90, 91, 93, 103, 105, 126, 133, 140, 153, 170, 173, 180, 185, 203, 213, 220, 233, 260, 283. Then $\hat{\Delta}$ is found to be 41, the average of $U^{(28)}$ and $U^{(29)}$. To obtain a 96% confidence interval, set $\alpha/2 = 0.02$ and from Table A.6 of Hollander and Wolfe (1999) obtain $w_{0.02} = 74$. Then from eqn [12], $C_{0.04} = 11$, and from eqn [11] we obtain $\Delta_L = U^{(11)} = -19$, $\Delta_U = U^{(46)} = 153$.

The *k*-Sample Problem

The two-sample problem of Section The Two-Sample Location Problem is readily extended to $k \geq 2$ samples, one sample from each of k populations. The basic null hypothesis is that the k samples can be viewed as a single combined sample from a common population. Typically we are interested in alternatives that specify that the locations (medians, means) of the k populations are not equal. More formally, let $X_{ij}, X_{2j}, \ldots, X_{n_j j}$ denote the random sample of n_j observations from the jth population, which is assumed to have an unknown continuous distribution F_j, $j = 1, \ldots, k$ where $F_j(x) = P(X_{ij} \leq x)$ for $i = 1, \ldots, n_j$. We also assume that all $N = \sum_{j=1}^{k} n_j$ random variables are mutually independent. The null hypothesis of equivalent populations is

$$H_0 : F_1(x) = F_2(x) = \ldots = F_k(x), \quad \text{for all } x \qquad [14]$$

To study location alternatives it is natural to introduce the location shift model

$$F_j(x) = F(x - \tau_j), \quad j = 1, \ldots, k \qquad [15]$$

where F is an unknown continuous distribution function and τ_j is the unknown treatment effect corresponding to the jth population. In submodel [15], the null hypothesis H_0 of eqn [14] becomes

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_k \qquad [16]$$

An excellent way to test H_0 is to use the Kruskal–Wallis test (see Kruskal, 1952; Kruskal and Wallis, 1952). Let τ_{ij} denote the rank of X_{ij} in the joint ranking of all N observations from the k samples. Furthermore, let

$$R_j = \sum_{i=1}^{n_j} r_{ij}, \quad R.j = R_j/n_j, \quad j = 1, \ldots, k \qquad [17]$$

denote, respectively, the sum and average of the ranks received by the jth sample. The Kruskal–Wallis statistic is

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(R.j - \frac{N+1}{2}\right)^2$$
$$= \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j}\right) - 3(N+1) \qquad [18]$$

Significantly large values of H indicate the τ's are not all equal to a common value. A distribution-free test at the α level of significance, of H_0 in eqn [16], against the alternative that the τ's are not all equal, is

$$\text{reject } H_0 \text{ if } H \geq h_\alpha; \quad \text{otherwise accept } H_0 \qquad [19]$$

where h_α is chosen to make the type-I error probability of the test equal to α and is the upper α^{th} percentile point of the null distribution of H. (For $k = 2$, this test reduces to the two-sided Wilcoxon test of eqn [6]).

The test based on H is distribution-free under H_0 in eqn [14] assuming the common distribution of F is continuous. For then all $N!/(\prod_{j=1}^{k} n_j!)$ possible assignments of n_1 ranks to the sample 1 observations, n_2 ranks to the sample 2 observations, $\ldots n_k$ ranks to the sample k (Hollander and Wolfe, 1999). StatXact performs the exact test as well as the large sample approximation next described.

The large sample approximation to procedure eqn [19] is

$$\text{reject } H_0 \text{ if } H \geq \chi^2_{k-1,\alpha}; \quad \text{otherwise accept } H_0 \qquad [20]$$

where $\chi^2_{k-1,\alpha}$ is the upper αth percentile point of a chi-squared distribution with $k-1$ degrees of freedom.

For a ties correction to the large sample approximation, compute H using average ranks. That is, assign each of the observations in a tied group the average of the ranks those observations would receive if they were not tied. Then refer to H' to a chi-squared distribution with $k-1$ degrees of freedom where

$$H' = H \Bigg/ \left\{ 1 - \left(\sum_{j=1}^{g} \left(t_j^3 - t_j \right) / \left(N^3 - N \right) \right) \right\} \qquad [21]$$

In eqn [21], g denotes the number of tied groups and t_j is the size of tied group j (and an untied observation is considered to be a tied group of size 1). The ties correction of eqn [21] does not yield an exact α level test. To obtain an exact α-level test in the presence of ties, see Hollander and Wolfe (1999: Comment 6.8).

If the hypothesis H_0 in eqn [16] of equivalent treatments is rejected, it is natural to investigate which particular pairs of treatments differ from one another. This is the aim of multiple comparisons. There are various multiple comparison procedures available (cf Miller, 1981; Hettmansperger and McKean, 1998; Hochberg and Tamhane, 1987; Hollander and Wolfe, 1999). Here we will describe and illustrate the multiple comparison procedure proposed independently by Dwass (1960) and Steel (1960) for the equal sample size case and generalized by Critchlow and Fligner (1991) to allow unequal sample sizes. The procedure is based on $\binom{k}{2}$ separate pairwise rankings, corresponding to ranking jointly each pair of samples.

For each pair of treatments (u, v), let W_{uv} denote the Wilcoxon rank sum of the ranks attained by the n_v observations in sample v in the joint ranking of the n_u observations from sample u and the n_v observations from sample v.

Let

$$W^*_{uv} = \sqrt{2}\left(\frac{W_{uv} - E_0(W_{uv})}{\{\text{var}_0(W_{uv})\}^{1/2}} \right)$$
$$= \left(W_{uv} - \frac{n_v(n_u + n_v + 1)}{2} \right) \Bigg/ \left(\frac{n_u n_v (n_u + n_v + 1)}{24} \right)^{1/2},$$
$$1 \leq u < v \leq k \qquad [22]$$

A distribution-free two-sided all treatments multiple comparison procedure, at an experimentwise error rate of α, is

$$\text{decide } \tau_u \neq \tau_v \text{ if } \left| W^*_{uv} \right| \geq w^*_\alpha; \quad \text{otherwise decide } \tau_u = \tau_v \qquad [23]$$

The constant w^*_α is chosen to make the experimentwise error rate equal to α, that is, w^*_α is chosen so that

$$P_0\left(\left| W^*_{uv} \right| < w^*_\alpha,\ u = 1, \ldots, k-1;\ v = u+1, \ldots, k \right) = 1 - \alpha \qquad [24]$$

where the probability $P_0(\cdot)$ is computed under H_0. Selected values of w^*_α can be found in Table A.16 of Hollander and Wolfe (1999). Also see Steel (1960, 1961) and Critchlow and Fligner (1991). A large sample approximation to procedure [23] is

$$\text{decide } \tau_u \neq \tau_v \text{ if } \left| W^*_{uv} \right| \geq q_\alpha; \quad \text{otherwise decide } \tau_u = \tau_v, \qquad [25]$$

where q_α is the upper αth percentile point of the distribution of range of k $N(0, 1)$ random variables. Values of q_α can be found in Table A.17 of Hollander and Wolfe (1999). For the justification of the approximation, see, for example, Hollander and Wolfe (1999: pp. 246–247) and Hochberg and Tamhane (1987). There are also various procedures available to provide simultaneous confidence intervals for contrasts (cf Hollander and Wolfe (1999: Section 6.9) and Hettmansperger and McKean (1998: Section 4.3). Miller (1981) and Hochberg and Tamhane (1987) are books that specialize in multiple comparisons.

Example 2

We illustrate the k-sample procedures of this section with data from (Zelazo et al., 1972) (see also Brown and Hollander, 1977). The study concerned infants who were randomized to one of four treatments to determine if stimulation of the walking and placing reflexes in the newborn accelerates the time at which infants learn to walk by themselves. The subjects of **Table 2** were about 1-week-old male infants from middle-class and upper-middle-class families. Each infant was

Table 2 Ages (in months) of infants for walking alone (ranks in parentheses)

Active exercise	*Passive exercise*	*No exercise*	*8-week*
9.00(1.5)	11.00(10)	11.50(12.5)	13.25(20.5)
9.50(3.5)	10.00(7)	12.00(16.5)	11.50(12.5)
9.75(5)	10.00(7)	9.00(1.5)	12.00(16.5)
10.00(7)	11.75(15)	11.50(12.5)	13.50(22)
13.00(18.5)	10.50(9)	13.25(20.5)	11.50(12.5)
9.50(3.5)	15.00(23)	13.00(18.5)	
$R_1 = 39$	$R_2 = 71$	$R_3 = 82$	$R_4 = 84$

Source: Zelazo, P.R., Zelazo, N.A., Kolb, S., 1972. 'Walking' in the newborn. Science 176, 314–315.

randomly assigned to one of four groups: an experimental group (active exercise) and three control groups (passive exercise, no exercise, eight-week control). We apply the Kruskal–Wallis test to test if there are significant differences among the four groups. The reader will note from the rankings in **Table 2** that there are five tied groups of size 2, one tied group of size 3, and one tied group of size 4. Applying the large sample approximation with the correction for ties given by eqn [21], we find $H' = 6.88$. This agrees with the calculation obtained using StatXact, namely $H' = 6.88$ with an approximate p-value of 0.075 8 obtained by referring $H' = 6.88$ to a chi-squared distribution with three degrees of freedom. Thus there is evidence that the treatments are not equivalent. We now turn to multiple comparisons to see which treatments differ and which yield the earlier times to walking alone.

We will apply the approximate multiple comparison procedure given by eqn [25]. Performing the six two-sample rankings using the data of **Table 2**, we find from eqn [22], $W^*_{12} = 2.717$, $W^*_{13} = 2.038$, $W^*_{14} = 3.098$, $W^*_{23} = 1.132$, $W^*_{24} = 2.066$, and $W^*_{34} = 1.033$. From the distribution of the range of 4 $N(0, 1)$ random variables, we find $q_{0.10} = 3.240$ and $q_{0.20} = 2.784$. Thus at an approximate experimentwise error rate of 0.10, procedure [25] makes no significant decisions but at an experimentwise error rate of 0.20, we decide, since $|W^*{}_{14}| = 3.098 > 2.784$, that $\tau_1 \neq \tau_4$ and there is evidence that active exercise leads to earlier walking than is obtained by the 8-week control. We note that the conservative nature of this experimentwise error rate approach to multiple comparisons (as well as the two-sided procedure used here) makes it difficult to find significant differences at a low experimentwise error rate. If, however, one performs individual one-sided two-sample Wilcoxon rank sum tests of the form given by procedure [4], with sample sizes 6 and 5, as is appropriate for the (1, 4), (2, 4), and (3, 4) comparisons, the $\alpha = 0.063$ critical value for the upper-tail procedure [4] is 39 (see Table A.6, Hollander and Wolfe, 1999) and since W14 = 42, W24 = 38, and $W_{34} = 34$, we find that active exercise leads to earlier walking times than the 8-week control. Similarly, for the individual one-sided two-sample tests with sample sizes 6 and 6, as is appropriate for the (1, 2), (1, 3), and (2, 3) comparisons, the $\alpha = 0.066$ critical value for the upper-tail procedure (4) is 49 (see Table A.6, Hollander and Wolfe, 1999), and since $W_{12} = 51$, $W_{13} = 48$, and $W_{23} = 44$, we are able to judge active exercise as leading to earlier walking times than passive exercise.

The Multivariate Two-Sample Location Problem

The Wilcoxon statistic of Section The Two-Sample Location Problem can be generalized to the multivariate case (cf Puri and Sen, 1971: Chapter 5). Let

$$X^j_\alpha = \left(X^{(j)}_{1\alpha}, \ldots, X^{(j)}_{p\alpha}\right)', \quad \alpha = 1, \ldots, n_j, \quad j = 1, 2 \qquad [26]$$

be independent p-dimensional random vectors with X^j_α having distribution function $F_j(x)$. The hypothesis of eqn [1] is generalized to

$$H_0 : F_1(x) = F_2(x) \quad \text{for all } \boldsymbol{x} \qquad [27]$$

As with the procedures in Sections The Two-Sample Location Problem and The k-sample Problem, we will define a test intended to detect location differences of the form

$$F_j(\boldsymbol{x}) = F(\boldsymbol{x} + \boldsymbol{\delta}_j), \quad j = 1, 2 \qquad [28]$$

where for model [28], H_0 can be written

$$H_0 : \boldsymbol{\delta}_1 = \boldsymbol{\delta}_2 = \mathbf{0} \qquad [29]$$

that is, the random vectors can be viewed as coming from a common p-variate population.

To test H_0 versus the location alternatives $\boldsymbol{\delta}_1 \neq \boldsymbol{\delta}_2$, let W_i be the sum of the second sample's component i ranks in the joint ranking from least to greatest of the $N = n_1 + n_2$ values in the pth component of the first and second samples. We let $R_{i1}, \ldots, R_{in}$ denote the ranks obtained by the first sample and $R_{i,n_1+1}, \ldots, R_{i,n_1+n_2}$ denote the ranks obtained by the second sample in this joint ranking. To compute this multivariate two-sample Wilcoxon-type test, we thus need to perform p two-sample rankings, each of size $N = n_1 + n_2$. Let

$$V_i = W_i - n_2(N+1)/2, \quad i = 1, \ldots, p \qquad [30]$$

and let $\boldsymbol{V} = (V_1, V_2, \ldots, V_p)'$. The test statistic is

$$V^* = \mathbf{V}'\boldsymbol{A}^{-1}\mathbf{V} \qquad [31]$$

where the $p \times p$ matrix $\boldsymbol{A}$ has elements

$$a_{ii} = n_1 n_2 (N+1)/12$$
$$a_{ij} = [n_1 n_2/\{(N-1)N\}] \cdot \left[\sum_{s=1}^{N} R_{is}R_{js} - \left\{N(N+1)^2/4\right\}\right] \qquad [32]$$

The matrix $N^{-3}\mathbf{A}$ is a consistent estimator of the asymptotic covariance matrix of $N^{-(3/2)}\boldsymbol{V}$.

An asymptotically distribution-free test of H_0 defined by eqn [27] against the location alternatives $\boldsymbol{\delta}_1 \neq \boldsymbol{\delta}_2$, at the approximate α level, rejects H_0 if $V^* > \chi^2_{p,\alpha}$. This test is not distribution-free for finite samples, but a conditionally distribution-free test based on V^* can be obtained using a permutation principle (Section 5.3, Puri and Sen, 1971).

For simplicity we have chosen to describe (and illustrate in Example 3 below) the multivariate two-sample Wilcoxon test. For the k-sample multivariate Wilcoxon test, see Puri and Sen (1971: Chapter 5). For point estimation of $(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$ see Section 6.3 of Puri and Sen (1971) and for Bonferoni confidence bounds for $(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$, see Puri and Sen (1971: pp. 258–259).

Example 3

The taste aversion data in **Table 3** are from J.C. Smith and P. Smith (personal communication, 1999). Taste aversion learning is a situation in which an animal (or human) drinks a taste solution and then is subjected to some uncomfortable experience (e.g., radiation exposure, injection of some toxic drug such as lithium chloride, exposure to a high intensity magnet, etc.). After the 'pairing' of the taste and the drug, the subject is subsequently tested with the taste solution to see if an aversion has occurred. There has been much interest in how and where this association of taste and drug is made in the brain. It involves: (1) Where is the taste sensation mediated in the brain? (2) Where is the drug sensation mediated in the

Table 3 Total water intake (component ranks in parentheses)

	Day 1	*Day 2*	*Day 3*	*Day 4*	*Day 5*	*Day 6*
PBN	56.22(10)	53.64(9)	50.54(7)	50.66(7)	47.44(6)	45.68(5)
	13.76(1)	18.84(1)	18.94(1)	18.52(1)	20.28(1)	21.44(1)
	25.40(3)	27.44(3)	30.20(3)	31.00(4)	30.84(3)	33.52(3)
	28.22(4)	40.92(4)	42.04(5)	41.64(5)	41.92(4)	45.64(4)
	18.76(2)	25.18(2)	22.58(2)	22.40(3)	21.72(2)	26.24(2)
	57.92(11)	59.54(11.5)	58.88(12)	62.48(12)	58.62(12)	59.32(12)
	65.02(13)	61.28(13)	71.48(13)	65.14(13)	72.66(13)	60.06(13)
Control	51.12(7)	50.72(7)	53.20(8)	52.22(9)	55.74(10)	54.58(7)
	45.94(6)	44.34(5)	32.88(4)	20.22(2)	45.14(5)	55.04(8)
	53.22(8)	59.54(11.5)	55.82(10)	49.80(6)	52.32(7)	55.42(10)
	41.36(5)	50.44(6)	49.64(6)	52.08(8)	56.00(11)	53.56(6)
	53.34(9)	53.62(8)	57.24(11)	53.98(11)	54.24(8)	55.12(9)
	58.02(12)	55.28(10)	53.96(9)	53.12(10)	55.20(9)	56.16(11)
	$W_1 = 47$	$W_2 = 47.5$	$W_3 = 48$	$W_4 = 46$	$W_5 = 50$	$W_6 = 51$

Source: J.C. Smith and P. Smith (private communication, 1999).

brain? (3) Where is the association of the taste and drug mediated?

Taste is 'sensed' by taste nerves in the oral cavity and makes the first synapse in the nucleus of the solitary tract and the second synapse in the parabrachial nucleus (PBN). R. Norgren believes that the association (point (3) above) is made in the PBN. He has shown that when he destroys the PBN, a rat is unable to learn a conditioned taste aversion. In contrast, the rat can learn an aversion that does not involve taste, but does involve texture, temperature, etc. J.C. Smith and P. Smith can take a rat that normally prefers cold water over warm water, give it a pairing of cold water with lithium chloride and turn it into a warm water 'preferrer.' Norgren sent Smith and Smith some PBN-operated rats (along with some sham-operated controls) for testing with the warm/cold water procedure. Norgren predicted the PBN rat could learn the warm/cold discrimination since there is no reason to think this involves taste. The problem with making lesions in the brain is that most of the brain parts do more than one thing. Before Smith and Smith tested rats for Norgren, they felt it was a good idea to see if the PBN rats eat laboratory food and drink water like normal rats. The data in **Table 3**, with $n_1 = 7$ PBN-lesion rats and $n_2 = 6$ control rats, come from those observations. They represent a group of rats with PBN lesions and a control group that received surgery but the PBN was spared. Smith and Smith measured water intake over 6 days. They suspected a priori that there might be decrement in water drinking for the PBN rats. The analysis based on the two-sample multivariate Wilcoxon test leads to the conclusion that there are no significant differences between the two groups.

We used an S-plus program to calculate $V^*(31)$ and found $V^* = 5.368$. Referring $V^* = 5.368$ to a chi-squared distribution with four degrees of freedom yields a p-value of 0.498. Thus there is no evidence of differences in total water intake between the PBN and control groups. We note, however, that the test based on V^* is two-sided. Since Smith and Smith suspected a priori that the PBN rats might drink less, we should also check the one-sided p-values associated with the one-sided two-sample Wilcoxon test [4] to see if there is evidence that the PBN rats did less drinking. The one-sided p-values corresponding to $W_1 = 47$, $W_2 = 47.5$, $W_3 = 48$, $W_4 = 46$, $W_5 = 50$, and $W_6 = 51$ are (from Table A.6, Hollander and Wolfe, 1999), respectively 0.267, 0.267, 0.223, 0.314, 0.147, 0.117. Thus, again, we do not find significant differences between the two groups. In other words the individual tests are consistent with the overall test based on V^*. The latter, of course, accounts of the dependencies between days.

Efficiencies of Location Procedures

The location procedures of Sections The Two-Sample Location Problem, The k-Sample Problem, and The Multivariate Two-Sample Location Problem have desirable efficiency properties with respect to their parametric competitors derived under the distributional assumption of normality. For translation alternatives, the Pitman asymptotic relative efficiency (Puri and Sen (1971: Section 3.8.1)) of the test based on W with respect to the normal theory two-sample t-test is

$$E(W,t) = 12\sigma_F^2\left(\int f^2(x)\mathrm{d}x\right)^2 \qquad [33]$$

where σ_F^2 is the variance of the population with distribution F and f is the probability density corresponding to F. Expression [33] is also the asymptotic relative efficiency of the estimator $\hat{\Delta}$ in eqn [10] with respect to $\overline{\Delta} = \overline{Y} - \overline{X}$ (Hodges and Lehmann, 1963); the asymptotic relative efficiency of the confidence interval given by eqn [11] relative to the normal theory confidence interval centered at $\overline{\Delta}$ (Lehmann, 1963); the asymptotic relative efficiency of the Kruskal–Wallis test relative to the normal theory one-way layout $\mathscr{F}$-test (Andrews, 1954); the asymptotic relative efficiency of the multiple comparison procedure [23] with respect to the corresponding normal theory procedure based on sample means (Sherman, 1965). Hodges and Lehmann (1956) showed that for all populations $E(W, t)$ is at least 0.864. Values of this efficiency expression for various distributions are given in **Table 4**. The Pitman asymptotic relative efficiency of the test based on V^* in eqn [31] with respect to a likelihood ratio criterion test is quite complicated and depends on the univariate marginal distributions and bivariate marginal distributions of F. It is given by Puri and Sen (1971: p. 212).

Table 4 Values of $E(W, t)$ for various distributions

F	*Normal*	*Uniform*	*Logistic*	*Double-exponential*	*Cauchy*	*Exponential*
$E(W, t)$:	0.955	1.000	1.097	1.500	∞	3.00

Conclusion

We have highlighted some rank-based procedures in two- and *k*-sample location problems. These procedures and their generalizations played a key role in the early development of nonparametric statistics. They remain attractive to data analysts today because they enjoy excellent efficiency properties, they are relatively insensitive to outliers, and they are valid under very mild assumptions concerning the underlying populations from which the data are obtained. The subject of nonparametric statistics continues to be vigorously studied. Areas of interest at the time of this writing include Bayesian and computer intensive methods (*See* Nonparametric Statistics: Advanced Computational Methods), regression, and survival analysis. Recent books on nonparametric statistics that would be useful to the interested reader include Sprent (1993), Gibbons (1997), Hettmansperger and McKean (1998), and Hollander and Wolfe (1999).

See also: Nonparametric Statistics: Advanced Computational Methods; Nonparametric Statistics: Records.

Bibliography

Andrews, F.C., 1954. Asymptotic behavior of rank tests for analysis of variance. Annals of Mathematical Statistics 25, 724–736.

Brown Jr., B.W., Hollander, M., 1977. Statistics: A Biomedical Introduction. Wiley, New York.

Critchlow, D.E., Fligner, M.A., 1991. On distribution-free multiple comparisons in the one-way analysis of variance. Communications in Statistics – Theory and Methods 20, 127–139.

Dwass, M., 1960. Some *k*-sample rank-order tests. In: Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G., Mann, H.B. (Eds.), Contributions to Probability and Statistics. Stanford University Press, Stanford, CA, pp. 198–202.

Gibbons, J.D., 1997. Nonparametric Methods for Quantitative Analysis, third ed. American Sciences Press, Syracuse, New York.

Hettmansperger, T.P., 1984. Statistical Inference Based on Ranks. Wiley, New York.

Hettmansperger, T.P., 1985. Multivariate location tests. In: Kotz, S., Johnson, N.L., Read, C.B. (Eds.), Encyclopedia of Statistical Sciences, vol. 6. Wiley, New York, pp. 79–83.

Hettmansperger, T.P., McKean, J.W., 1998. Robust Nonparametric Statistical Methods. Arnold, London.

Hochberg, Y., Tamhane, A.C., 1987. Multiple Comparison Procedures. Wiley, New York.

Hodges Jr., J.L., 1967. Efficiency in normal samples and tolerance of extreme values for some estimates of location. In: Le Cam, L., Neyman, J. (Eds.), Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, pp. 163–186.

Hodges Jr., J.L., Lehmann, E.L., 1956. The efficiency of some nonparametric competitors of the *t*-test. Annals of Mathematical Statistics 27, 324–335.

Hodges Jr., J.L., Lehmann, E.L., 1963. Estimates of location based on rank tests. Annals of Mathematical Statistics 34, 598–611.

Hollander, M., Peña, E., 1988. Nonparametric tests under restricted treatment-assignment rules. Journal of the American Statistical Association 83, 1144–1151.

Hollander, M., Wolfe, D.A., 1999. Nonparametric Statistical Methods, second ed. Wiley, New York.

Kruskal, W.H., 1952. A nonparametric test for the several sample problem. Annals of Mathematical Statistics 23, 525–540.

Kruskal, W.H., Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association 47, 583–621.

Lehmann, E.L., 1963. Nonparametric confidence intervals for a shift parameter. Annals of Mathematical Statistics 34, 1507–1512.

Lehmann, E.L., 1975. Nonparametrics: Statistical Methods Based on Ranks. Holden Day, San Francisco, CA.

March, G.L., John, T.M., McKeown, B.A., Sileo, L., George, J.C., 1976. The effects of lead poisoning on various plasma constituents in the Canada goose. Journal of Wildlife Diseases 12, 14–19.

Mehta, C.R., Patel, N.R., Wei, L.J., 1988. Computing exact significance tests with restricted randomization rules. Biometrika 75, 295–302.

Miller Jr., R.G., 1981. Simultaneous Statistical Inference, second ed. Springer, New York.

Noether, G.E., 1967. Elements of Nonparametric Statistics. Wiley, New York.

Puri, M.L., Sen, P.K., 1971. Nonparametric Methods in Multivariate Analysis. Wiley, New York.

Randles, R.H., Wolfe, D.A., 1979. Introduction to the Theory of Nonparametric Statistics. Wiley, New York.

Sherman, E., 1965. A note on multiple comparisons using rank sums. Technometrics 7, 255–256.

Sprent, P., 1993. Applied Nonparametric Statistical Methods, second ed. Chapman and Hall, London.

StatXact, 1999. Cytel Corporation, Cambridge, MA.

Steel, R.G.D., 1960. A rank sum test for comparing all pairs of treatments. Technometrics 2, 197–207.

Steel, R.G.D., 1961. Some rank sum multiple comparisons tests. Biometrics 17, 539–552.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics 1, 80–83.

Zelazo, P.R., Zelazo, N.A., Kolb, S., 1972. 'Walking' in the newborn. Science 176, 314–315.

Nonparametric Statistics: Records

Mohammad Ahsanullah, Rider University, Lawrenceville, NJ, USA
Valery B Nevzorov, St. Petersburg State University, St. Petersburg, Russia

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by B.C. Arnold, volume 16, pp. 10681–10685, © 2001, Elsevier Ltd.

Abstract

In 1952, Chandler defined the so-called record times and record values. Records arise naturally in many fields of studies such as climatology, sports, medicine, traffic, and so on. The classical record schemes are based on sequences of independent identically distributed random variables. In the first part of the article, different types of the classical records are discussed. In the second part of the article, some modern generalizations of the classical record schemes are given. The so-called F^{α}-scheme, linear draft record model, δ-exceedance record scheme, and the models of records with restrictions and confirmations are presented here.

In 1952, Chandler defined the so-called record times and record values and laid the groundwork for a mathematical theory of records. For six decades, beginning with his pioneering work, there appeared about 500 papers and some monographs devoted to different aspects of the theory of records. This theory relies largely on the theory of order statistics and is especially closely connected to extreme order statistics. Records are very popular because they arise naturally in many fields of studies such as climatology, sports, medicine, traffic, industry, and so on. The records as such are memorials of their time. The annals of records reflect the progress in science and technology and enable us to study the evaluation of mankind on the basic of record achievements in various areas of its activity. A large number of record data saved for a long time inspirited the appearance of different mathematical models reflecting the corresponding record processes and forecasting the future record results.

Definitions of Records

Let X_1, X_2, … be a sequence of random variables and $X_{1,n} \le X_{2,n} \le \ldots \le X_{n,n}$, $n = 1,2,\ldots$, be the corresponding order statistics. For any $n = 1,2,\ldots$ denote also $M(n) = X_{n,n} = \max\{X_1, X_2,\ldots,X_n\}$ and $m(n) = X_{1,n} = \min\{X_1,X_2,\ldots,X_n\}$. Now one can define the classical upper record times $L(n)$ and upper record values $X(n)$ as follows: $L(1) = 1$, $X(1) = X_1$ and then,

$$L(n+1) = \min\left\{j : X_j > X(n)\right\}, \quad X(n+1) = X_{L(n+1)}, \quad n = 1,2,\ldots \tag{1}$$

One can use the following alternative definitions:

$$L(1) = 1, \quad L(n+1) = \min\left\{j : X_j > M(L(n))\right\}, \quad n = 1,2\ldots, \quad \text{and} \quad X(n) = M(L(n)), \quad n = 1,2,\ldots \tag{2}$$

If we replace the sign $>$ and $M(L(n))$ in [2] by $<$ and $m(L(n))$, respectively, we obtain the definitions of the lower record times and the lower record values. Indeed, the theories of upper and lower records coincide practically in all their details, since the lower records for the sequence $-X_1,-X_2,\ldots$ correspond to the upper records for the sequence $X_1,X_2,\ldots$. Using the sign $\ge$ in [1] instead of $>$ we introduce the so-called weak upper records, when any repetition of the previous record value is also considered as a new record. Analogically, the sign $\le$ in [1] gives the opportunity to define the weak lower record times and the weak lower record values. Note that the theory of weak records has practical meaning only for sequences of the initial X's, which have discrete distributions.

The so-called *k*th records are a natural extension of records. The *k*th record times $L(n,k)$ and the *k*th record values $X(n,k)$ for any $k = 1,2,\ldots$ are defined as follows:

$$L(1,k) = k, \quad L(n+1,k) = \min\left\{j > L(n,k) : X_j > X_{j-k,j-1}\right\}, \quad n = 1,2,\ldots, \tag{3}$$

and

$$X(n,k) = X_{L(n,k)-k+1,L(n,k)}, \quad n = 1,2,\ldots \tag{4}$$

To be precise, [3] and [4] define the *k*th upper record times and the *k*th upper record values, respectively. One can also introduce the *k*th lower record times and the *k*th lower record values changing the event $X_j > X_{j-k,j-1}$ in [3] by $X_j < X_{k,j-1}$ and replacing $X_{L(n,k)-k+1,L(n,k)}$ in [4] by $X_{k,L(n)}$. If $k = 1$, then definitions of *k*th record values $X(n,k)$ and *k*th record times $L(n,k)$ coincide with the definitions of $X(n)$ and $L(n)$ given in [1].

We will use $N(n)$ to denote the number of records among random variables $X_1,X_2,\ldots,X_n$, $n = 1,2,\ldots$, and $N(n,k)$ will denote the number of *k*th records in a sequence $X_1,X_2,\ldots,X_n$ respectively.

Sequential Ranks and Record Indicators

The essential role in the theory of records is played by the sequential ranks $R(n)$, the record indicators ξ_j, $j = 1,2,\ldots$, and the indicators of the *k*th records $\xi_j(k)$, $j = 1,2,\ldots$, $k = 1,2,\ldots$. The sequential rank $R(n)$ denotes the rank of X_n among $X_1,X_2,\ldots,X_n$, i.e.,

$$X_n = X_{R(n),n}, \quad n = 1,2,\ldots$$

The record indicator ξ_j is defined as follows: $\xi_j = 1$ if X_j is a record value and $\xi_j = 0$ otherwise. Analogically, the indicator $\xi_j(k)$ can be defined for any $k = 1,2,\ldots$ and $j \ge k$.

$\xi_j(k) = 1$, if X_j is the kth record value and $\xi_j(k) = 0$ otherwise. There are some useful relations between record indicators and sequential ranks. We will formulate some simple equalities for indicators of the upper records. Indeed, analogical results are also valid for the lower records. Note, that

$$\{\xi_j = 1\} = \{R(j) = j\} = \{X_j = M(j)\} = \{X_j = X_{j,j}\} = \{X_j \text{ is a record value}\}$$

and for any $k = 1,2,\ldots$ and $j \geq k$

$$\{\xi_j(k) = 1\} = \{R(j) \geq j-k+1\} = \{X_{j-k+1,j} \geq X_{j-k,j-1}\} = \{X_j > X_{j-k,j-1}\} = \{X_j \text{ is a } k\text{th record}\}$$

The record indicators allow us to give convenient relations for the numbers of records $N(n)$ and $N(n,k)$:

$$\{N(n) = m\} = \{\xi_1 + \ldots + \xi_n = m\}, \quad n = 1,2,\ldots, \quad m = 1,2,\ldots,n \qquad [5]$$

$$\{N(n,k) = m\} = \{\xi_k(k) + \ldots + \xi_n(k) = m\}, \quad n = k, k+1, \ldots, \quad m = 1,2,\ldots,n-k+1 \qquad [6]$$

The classical theory of records is connected with the situation, when the initial sequence of random variables $X_1,X_2,\ldots$ is a sequence of independent random variables with a common continuous distribution function (d.f.). In this case, sequential ranks and record indicators have a number of useful and rather convenient properties.

Theorem 1

For independent identically distributed random variables $X_1,X_2,\ldots$ with a continuous distribution function F the sequential ranks $R(1)$, $R(2),\ldots$ are independent and $P\{R(n) = m\} = 1/n$, $m = 1,2,\ldots,n$, $n = 1,2,\ldots$.

Theorem 2

Under conditions of theorem 1, for any fixed $k = 1,2,\ldots$, indicators $\xi_k(k)$, $\xi_{k+1}(k),\ldots$ are independent and $P\{\xi_j(k) = 1\} = k/j$, $j = k, k+1,\ldots$.

As a partial case of theorem 2, it follows that, under conditions of theorem 1, record indicators $\xi_1,\xi_2,\ldots$ are independent and

$$P\{\xi_j = 1\} = 1 - P\{\xi_j = 0\} = 1/j, \quad j = 1,2,\ldots$$

These results together with equalities [5] and [6] allow us to find that the distributions of $N(n)$ and $N(n,k)$ are expressed as distributions of sums of independent random variables. One can see also that under conditions of theorem 1 there are some convenient relations for the kth record times $L(n,k)$ and, in particular, for the record times $L(n)$:

$$P\{L(n,k) > m\} = P\{N(m,k) < n\} = P\{\xi_k(k) + \xi_{k+1}(k) + \ldots + \xi_m(k) < n\} \qquad [7]$$

and

$$P\{L(n) > m\} = P\{N(m) < n\} = P\{\xi_1 + \xi_2 + \ldots + \xi_m < n\} \qquad [8]$$

It follows from relations [7] and [8] that if $X_1,X_2,\ldots$ is a sequence of independent identically distributed (i.i.d.) random variables with any continuous d.f. F, then distributions of record times and numbers of records do not depend on F.

One more important situation in the classical record theory is connected with sequences of i.i.d. random variables having a discrete distribution. Without loss of generality, we can suppose that X's take nonnegative integer values. For discrete distributions, we introduce another type of record indicators. Let $\eta_n = 1$, if n is a record value in the sequence $X_1,X_2,\ldots$, that is there exists such $m = 1,2,\ldots$ that $X(m) = n$, and $\eta_n = 0$ otherwise (compare with indicators ξ_n!). Analogously, for any $k = 1,2,\ldots$ we can introduce indicators $\eta_n(k)$ for *kth* record values: $\eta_n(k) = 1$, if n is a kth record value in the sequence $X_1,X_2,\ldots$, and $\eta_n(k) = 0$ otherwise. The following results are valid for such type of indicators.

Theorem 3

Let $X,X_1,X_2,\ldots$ be a sequence of independent identically distributed random variables taking values $0,1,2,\ldots$ with probabilities $p_n = P\{X = n\} > 0$, $n = 0,1,2,\ldots$. Then for any fixed $k = 1,2,\ldots$ indicators $\eta_n(k)$, $n = 0,1,2,\ldots$, are independent and

$$P\{\eta_n(k) = 1\} = 1 - P\{\eta_n(k) = 0\} = (p_n/P\{X^3 \geq n\})^k, \quad n = 0,1,2,\ldots$$

Indeed, under $k = 1$, as a partial case of this theorem, one gets that record indicators $\eta_0,\eta_1,\eta_2,\ldots$ are independent and $P\{\eta_n = 1\} = 1-P\{\eta_n = 0\} = p_n/P\{X \geq n\}$, $n = 0,1,2,\ldots$.

It is easy to see that under conditions of theorem 3 we can express distributions of kth record values for discrete random variables via distributions of sums of independent indicators:

$$P\{X(n,k) > m\} = P\{\eta_0(k) + \ldots + \eta_m(k) < n\}, \quad m = 0,1,2,\ldots, \quad n = 1,2,\ldots, \qquad [9]$$

and, in particular, under $k = 1$ one has equality

$$P\{X(n) > m\} = P\{\eta_0 + \ldots + \eta_m < n\}, \quad m = 0,1,2,\ldots, \quad n = 1,2,\ldots \qquad [10]$$

As an example, we can consider the case, when X's have the geometric distribution with some parameter $0 < p < 1$, that is

$$P\{X_j = n\} = (1-p)p^n, \quad n = 0,1,2,\ldots, \quad \text{for} \quad j = 1,2,\ldots$$

In this situation

$$P\{\eta_n(k) = 1\} = (1-p)^k \text{ and}$$
$$P\{\eta_n(k) = 0\} = 1-(1-p)^k, \quad \text{for any } n = 0,1,2,\ldots$$

It means that the sum $\eta_0(k) + \ldots + \eta_m(k)$ has the binomial $B(m+1,q)$ distribution with a parameter $q = (1-p)^k$. Hence

$$P\{X(n,k) > m\} = \sum_{j=0}^{n-1}((m+1)!/j!(m+1-j)!)q^j(1-q)^{m+1-j}, \quad \text{if } 1 \leq n \leq m+1$$

and

$$P\{X(n,k) > m\} = 1, \quad \text{if } n > m+1$$

It was mentioned above that for discrete distributions, it is useful to introduce weak records together with classical (strong) record values. Weak records may arise, for example, in some sport competitions where any athlete who repeats the record achievement is also declared as a record holder. If we consider X's having a common discrete distribution, it is useful to introduce one more type of record statistics. Let the conditions of theorem 3 be valid. We define random variables $\mu_0,\mu_1,\mu_2,\ldots$, where μ_n denotes the number of those weak records in the sequence $X_1,X_2,\ldots$ that are equal to n. The following result is valid.

Theorem 4

Let $X,X_1,X_2,\ldots$ be a sequence of independent identically distributed random variables taking values $0,1,2,\ldots$ with probabilities $p_n = P\{X = n\} > 0$, $n = 0,1,2,\ldots$. Then for any fixed $k = 1,2,\ldots$, random variables $\mu_0,\mu_1,\mu_2,\ldots$ are independent and

$$P\{\mu_n = m\} = (1 - r(n))\,(r(n))^m,$$
$$n = 0,1,2,\ldots,\quad m = 0,1,2,\ldots$$

where

$$r(n) = p_n/P\{X \geq n\}$$

Let $X_\omega(1),X_\omega(2),\ldots$ denote the weak (upper) record values in the sequence $X_1,X_2,\ldots$. Then for any $n = 1,2,\ldots$ and $m = 0,1,2,\ldots$ the following relation is valid:

$$P\{X_\omega(n) > m\} = P\{\mu_0 + \mu_1 + \ldots + \mu_m < n\} \qquad [11]$$

Thus we see that there are some convenient representations of record values, record times, numbers of records ([5–11] among them), which allow us to impress these record statistics in terms of sums of independent random variables.

Distributions of Record Times

Let us consider the classical case, when $X_1,X_2,\ldots$ are independent and have a continuous d.f. F. Using the independence of the corresponding record indicators $\xi_1,\xi_2,\ldots$ one gets for any $n = 1,2,\ldots$ and $1 < j(1) < j(2) < \ldots < j(n)$ that

$$\begin{aligned}
&P\Big\{L(1) = 1, L(2) = j(2), \ldots, L(n) = j(n)\Big\}\\
&= P\Big\{\xi_1 = 1, \xi_2 = 0, \ldots, \xi_{j(2)-1} = 0, \xi_{j(2)} = 1, \xi_{j(2)+1} = 0, \ldots,\\
&\qquad \xi_{j(3)-1} = 0, \xi_{j(3)} = 1, \ldots, \xi_{j(n)-1} = 0, \xi_{j(n)} = 1\Big\}\\
&= P\Big\{\xi_1 = 1\Big\}P\Big\{\xi_2 = 0\Big\}\ldots P\Big\{\xi_{j(2)-1} = 0\Big\}P\Big\{\xi_{j(2)} = 1\Big\}\\
&\quad \times P\Big\{\xi_{j(2)+1} = 0\Big\}\ldots P\Big\{\xi_{j(3)-1} = 0\Big\}P\Big\{\xi_{j(3)} = 1\Big\}\ldots\\
&\quad \times P\Big\{\xi_{j(n)-1} = 0\Big\}P\Big\{\xi_{j(n)} = 1\Big\}\\
&= 1/(j(2) - 1)\,(j(3) - 1)\ldots(j(n) - 1)j(n)
\end{aligned} \qquad [12]$$

Note that here the joint distribution of record times does not depend on F.

Now one can see from [12] that

$$P\Big\{L(n) = m\Big\} = \sum 1/(j(2) - 1)(j(3) - 1)\ldots(j(n - 1) - 1)(m - 1)m$$

where the sum is taken over all $j(2),j(3),\ldots,j(n - 1)$, such that $1 < j(2) < j(3) < \ldots < j(n - 1) < m$.

It follows from [12] also that

$$P\{L(n) = j(n)|L(n - 1) = j(n - 1), L(n - 2) = j(n - 2), \ldots,$$
$$L(2) = j(2), L(1) = 1\} = j(n - 1)/j(n)\,(j(n) - 1)$$

and

$$P\{L(n) = j|L(n - 1) = i\} = i/j(j - 1),\quad n = 2,3,\ldots,\ j > i$$

Hence we see that the sequence of record times $L(1),L(2),\ldots$ in the announced situation forms a Markov chain.

It was mentioned above that record times are closely related to the random variables N(n), since for any $n = 1,2,\ldots$ and $m = 1,2,\ldots$ the following equalities are valid:

$$P\{L(n) > m\} = P\{N(m) < n\}$$

and

$$P\{L(n) = m\} = P\{N(m - 1) = n - 1, N(m) = n\} \qquad [13]$$

Equalities [5] and [13] allow us to express the distributions of L(n) in terms of independent record indicators:

$$\begin{aligned}
P\{L(n) = m\} &= P\{N(m - 1) = n - 1, \xi_m = 1\}\\
&= P\{N(m - 1) = n - 1\}/m \qquad [14]\\
&= P\{\xi_1 + \xi_2 + \ldots \xi_{m-1} = n - 1\}/m
\end{aligned}$$

Representations [5] and [14] of N(m) and L(n) give a possibility to find the distributions of these record statistics via the Stirling numbers of the first kind, $S_n(k)$, which are defined by equalities:

$$x(x - 1)\ldots(x - n + 1) = \sum_{k\geq 0} S_n(k)x^k$$

It appears that

$$P\{N(m) = k\} = (-1)^k S_m(k)/m! = |S_m(k)|/m!,$$
$$k = 1,2,\ldots,m$$

and

$$P\{L(n) = m\} = |S_{m-1}(n - 1)|/m!,\quad m = n, n + 1, \ldots$$

Representations [5] and [14] give the following formulas for the corresponding generating functions:

$$P_m(s) = Es^{N(m)} = s(1 + s)(2 + s)\ldots(m - 1 + s)/m!$$

and

$$\begin{aligned}
Q_n(s) &= Es^{L(n)} = 1 - (1 - s)\sum_{k=0}^{n-1}(-\log(1 - s))^k/k!\\
&= \int_0^{-\log(1-s)} v^{n-1}\exp(-v)dv/(n - 1)! \qquad [15]
\end{aligned}$$

Distributions of Record Values

Let us again consider the case when $X_1,X_2,\ldots$ are independent and have a continuous d.f. F. The record value X(n) can be presented as $X_{L(n)}$, where L(n) is the corresponding

record time with a generating function [15]. One can see that then

$$P\{X(n) < x\} = P\left\{X_{L(n)} < x\right\} = P\{M(L(n)) < x\} \qquad [16]$$

where $M(n) = \max\{X_1, X_2, \ldots, X_n\}$ and $P\{M(n) < x\} = F^n(x)$.

It follows from [15] and [16] that

$$\begin{aligned} P\{X(n) < x\} &= P\{M(L(n)) < x\} \\ &= \sum_{m \geq 0} P\{M(m) < x\} P\{L(n) = m\} \\ &= \sum_{m \geq 0} F^m(x) P\{L(n) = m\} \\ &= Q_n(F(x)) = \int_0^{-\log(1-F(x))} v^{n-1} \exp(-v) dv / (n-1)! \end{aligned}$$

The following result is valid for distributions of record values.

Theorem 5

Let $X(1) < X(2) < \ldots$ be the record values in a sequence of independent random variables having a common continuous distribution function F, and let $U(1) < U(2) < \ldots$ be the record values related to the uniform distribution on the interval [0,1]. Then for any $n = 1,2,\ldots$ the random vector $(F(X(1)),\ldots,F(X(n)))$ has the same distribution as $(U(1),\ldots,U(n))$.

Corollary

Let $X(1) < X(2) < \ldots$ and $Y(1) < Y(2) < \ldots$ be, respectively, the record values in a sequence of independent random variables $X_1, X_2, \ldots$ having a common continuous distribution function F_1 and in a sequence of independent identically distributed random variables $Y_1, Y_2, \ldots$ with a continuous distribution function F_2. Then for any $n = 1,2,\ldots$ the vector $(Y(1), Y(2), \ldots, Y(n))$ has the same distribution as the vector $(H_2(X(1)), H_2(X(2)), \ldots, H_2(X(n)))$, where $H_2(x) = G_2(F_1(x))$ and G_2 is the inverse function to F_2. Analogously, the vectors $(X(1), X(2), \ldots, X(n))$ and $(H_1(Y(1)), H_1(Y(2)), \ldots, H_1(Y(n)))$, where $H_1(x) = G_1(F_2(x))$ and G_1 is the inverse function to F_1, are identically distributed.

Let us consider the partial case of record values $Z(1) < Z(2) < \ldots$ related to the standard exponential E(1) distribution (the case, when $F(x) = 1 - \exp(-x)$, $x > 0$). We get

$$P\{Z(n) < x\} = \int_0^x v^{n-1} \exp(-v) dv / (n-1)!$$

that is, in this situation Z(n) has the gamma distribution with parameter n. It means that Z(n) has the same distribution as the sum $\nu_1 + \nu_2 + \ldots + \nu_n$ of independent E(1)-distributed random variables $\nu_1, \nu_2, \ldots$. Moreover, for any $n = 1,2,\ldots$ the vector $(Z(1), Z(2), \ldots, Z(n))$ has the same distribution as the vector $(\nu_1, \nu_1 + \nu_2, \ldots, \nu_1 + \nu_2 + \ldots + \nu_n)$. It means that the vector $(Z(1), Z(2) - Z(1), \ldots, Z(n) - Z(n-1))$ consists of independent elements and each of these elements has the standard exponential E(1) distribution.

Combining the previous results, one can get the following representation for record values $X(1) < X(2) < \ldots$ related to any continuous d.f. F. Let G below denote the inverse function to F.

Representation 1

For any $n = 1,2,\ldots$

$$\begin{aligned} &(X(1), X(2), \ldots, X(n))^d \\ &\quad = (H(\nu_1), H(\nu_1 + \nu_2), \ldots, H(\nu_1 + \nu_2 + \ldots + \nu_n)) \end{aligned}$$

where $\nu_1, \nu_2, \ldots$ are independent random variables having the exponential E(1) distribution and $H(x) = G(1 - \exp(-x))$.

Taking into account the property of the exponential records, it is not difficult to obtain the joint density function $f_n(x_1, x_2, \ldots, x_n)$ of the record values $Z(1), Z(2), \ldots, Z(n)$. It appears that

$$f_n(x_1, x_2, \ldots, x_n) = \exp(-x_n), \quad \text{if } 0 < x_1 < x_2 < \ldots < x_n$$

and $f_n(x_1, x_2, \ldots, x_n) = 0$ otherwise.

In the general case, when $X_1, X_2, \ldots$ have a d.f. F and a density function f, the joint density function of the record values $X(1), X(2), \ldots, X(n)$ is given by the formula:

$$\begin{aligned} &f_n(x_1, x_2, \ldots, x_n) = r(x_1) r(x_2) \ldots r(x_n)(1 - F(x_n)), \\ &\quad x_1 < x_2 < \ldots < x_n, \end{aligned}$$

where $r(x) = f(x)/(1 - F(x))$.

Now we consider the conditional distributions

$$\begin{aligned} &\varphi(x | x_1, x_2, \ldots, x_n) \\ &\quad = P\{X(n+1) > x | X(1) = x_1, X(2) = x_2, \ldots, X(n) = x_n\}, \\ &\quad x_1 < x_2 < \ldots < x_n < x \end{aligned}$$

for record values $X(1) < X(2) < \ldots < X(n) < X(n+1)$. It appears that

$$\begin{aligned} \varphi(x | x_1, x_2, \ldots, x_n) &= P\{X(n+1) > x | X(n) = x_n\} \\ &= (1 - F(x))/(1 - F(x_n)), \; x > x_n \qquad [17] \end{aligned}$$

It is interesting that equality [17] does not need the continuity of the d.f. F. It follows from [17] that record values $X(1), X(2), \ldots$ form a Markov chain.

If we now consider discrete X's taking values $0,1,2,\ldots$ then [17] can be rewritten in the form:

$$\begin{aligned} &P\{X(n+1) > j | X(n) = m\} = P\{X > j\}/P\{X \geq m+1\}, \\ &\quad j > m \geq n - 1 \end{aligned}$$

It follows from the latter equality that in this case

$$\begin{aligned} &P\{X(1) = j_1, X(2) = j_2, \ldots, X(n) = j_n\} \\ &\quad = P\{X = j_n\} \omega(j_1) \omega(j_2) \ldots \omega(j_{n-1}), 0 \leq j_1 < j_2 < \ldots \\ &\qquad < j_{n-1} < j_n, \end{aligned}$$

where $\omega(j) = P(X = j\}/P\{X > j\}$.

The simplest discrete case is presented by the geometric distribution. The following result is valid.

Theorem 6

Let $X, X_1, X_2, \ldots$ be independent identically distributed random variables such that

$$P\{X = j\} = (1 - p) p^{j-1}, \quad j = 1, 2, \ldots, \quad 0 < p < 1$$

and $X(1) < X(2) < \ldots$ be the record values in a sequence $X_1, X_2, \ldots$. Then the interrecord values $X(1), X(2) - X(1), X(3) - X(2), \ldots$ are independent and have the same geometric distribution as X.

Distributions of *k*th Record Values

The *k*th record values X(n,k) are a natural extension of records X(n). It is interesting that distributions of the *k*th records can be expressed via distributions of the classical record values. Really, together with a sequence of independent random variables $X_1, X_2, \ldots$ having a common d.f. F, let us consider one more sequence:

$$Y_1 = \min\{X_1, X_2, \ldots, X_k\},$$
$$Y_2 = \min\{X_{k+1}, X_{k+2}, \ldots, X_{2k}\}, \ldots$$

Let now X(n,k) be the *k*th record value based on X's and Y(n) be the classical records based on the sequence $Y_1, Y_2, \ldots$. It appears that for any fixed $k = 2,3,\ldots$ and any $n = 1,2,\ldots$ the vector $(X(1,k), X(2,k), \ldots, X(n,k))$ has the same distribution as the vector $(Y(1), Y(2), \ldots, Y(n))$.

Note that this result is valid for discrete X's as well. One can obtain immediately some important results for the *k*th records taking into account the analogous results for the classical record values. For example, if Z(n,k), $n = 1,2,\ldots$ denote the *k*th records for the standard exponential distribution, then the vector $(Z(1,k), Z(2,k), \ldots, Z(n,k))$ has the same distribution as the vector $(\nu_1/k, (\nu_1+\nu_2)/k, \ldots, (\nu_1+\nu_2+\ldots+\nu_n)/k)$, where $\nu_1, \nu_2, \ldots$ are the independent exponentially E(1) distributed random variables. Hence, the following relation is valid for *k*th records related to a sequence of $X_1, X_2, \ldots$ with a continuous d.f. F.

Representation 2

For any $n = 1,2,\ldots$

$$(X(1,k), X(2,k), \ldots, X(n,k))^d = (H(\nu_1/k),\ H((\nu_1+\nu_2)/k), \ldots, H((\nu_1+\nu_2+\ldots+\nu_n)/k)$$

where $\nu_1, \nu_2, \ldots$ are independent random variables having the exponential E(1) distribution, $H(x) = G(1-\exp(-x))$ and G is the inverse function to F.

Some useful results for the *k*th records follow immediately from representation 2 and analogous results for the classical records. Say, one obtain that

$$P\{X(n,k) < x\} = \int_0^{-k\log(1-F(x))} v^{n-1}\exp(-v)dv/(n-1)!$$

and this equality is valid for any $k = 1,2,\ldots$ and any continuous d.f. F.

Theorem 7

For any $k = 1,2,\ldots$ the sequence $X(1,k), X(2,k), \ldots$ forms a Markov chain and

$$P\{X(n+1,k) > x | X(n,k) = u\} = ((1-F(x))/(1-F(u)))^k, \quad x > u$$

Different Generalizations of the Classical Records

Indeed, for more than 60 years of its history, classical records were studied carefully. During this period, there appeared a number of 'nonclassical' models (see, for example, papers (Balakrishnan et al., 1996; Deheuvels and Nevzorov, 1993, 1994; Nevzorov and Saghatelyan, 2009; Sagateljan, 2008)). Descriptions of some of such schemes are discussed below.

F^α-Scheme

Yang (1975) considered records in the sequence

$$Y_k = \max\left\{X_{k,1}, \ldots, X_{k,n(k)}\right\}, \quad k = 1,2,\ldots$$

where $\{X_{k,j}\}$, $j = 1,2,\ldots,n(k)$, $k = 1,2,\ldots$, are i.i.d. random variables with a common continuous d.f. F. It is interesting that in his model, Yang proposed to take

$$n(k) = \lambda^{k-1}n(1), \quad k = 1,2,\ldots, \quad \lambda > 0.$$

Moreover, he preferred to take $\lambda = 2^{1/9}$ for his scheme analyzing the dynamics of the Olympic records and proposing that this dynamics really depends on the growth of population of our planet. Hence this value λ was chosen in view of the fact that the world population doubled during nine 4-year periods between Olympic Games, which were organized in 1900 and 1936. His partial model initiated the appearance of the more general, the so-called F^α-scheme (see, for details, papers (Deheuvels and Nevzorov, 1993, 1994) and books (Nevzorov, 2000, 2001)), where independent random variables $X_1, X_2, \ldots$ have different differential functions, but these are of the special form:

$$F_k(x) = F^{\alpha(k)}(x), \quad k = 1,2,\ldots, \quad \alpha(k) > 0.$$

That is, coefficients $\alpha(k)$ can take any positive values. The most important property of this scheme is that the record indicators ξ_n, $n = 1,2,\ldots$, defined above save the independence property and the following equalities are valid:

$$P\{\xi_n = 1\} = \alpha(n)/(\alpha(1)+\ldots+\alpha(n)), \quad n = 1,2,\ldots$$

These properties of record indicators provided the possibility to obtain for the F^α-scheme a lot of results for random record times and record values, which are analogous to the corresponding results for the 'classical' records.

Linear Draft Model

Let $\{X_n, n \geq 1\}$ be a sequence of i.i.d. random variables with d.f. F. Suppose that we deal with random variables $Y_n = X_n + cn$, $c > 0$, $n = 1,2,\ldots$. In this situation the distributions of record indicators ξ_n are given as

$$P\{\xi_n = 1\} = P(X_j < X_n + c(n-j),\ j = 1,2,\ldots,n-1) = \int_{-\infty}^{\infty}\prod_{j=1}^{n-1}F(x+cj)dF(x)$$

Let now F(x) in this model be the Gumbel distribution with

$$F(x) = e^{-e^{-x}}, \quad -\infty < x < \infty$$

and let us consider the corresponding Linear Draft model. In this case, we have

$$P\{Y_n < x\} = P\{X_n + cn < x\} = \exp\{-\exp(-x+cn)\} = (F(x))^{\alpha(n)} \quad [18]$$

where $\alpha(n) = \exp(cn)$, $n = 1,2,\ldots$, that is, we get the partial case of the F^{α}-scheme. It means that for this partial case of the Linear Draft model, record indicators ξ_n, $n = 1,2,\ldots$, are also independent.

δ-Exceedance Record Scheme

This scheme was suggested by Balakrishnan, Balasubramanian and Panchapakesan (Balakrishnan et al., 1996). A new observation X_j is called the *δ-exceedance upper record* if it is larger than the previous record by a prefixed positive quantity δ. In this model, $L(1) = 1$ and

$$L(n+1) = \min\left\{j | j > L(n), X_j > X_{L(n)} + \delta\right\}, \quad n = 1,2,\ldots$$

he sequence $\{X_{L(n)}, n = 1,2,\ldots\}$ forms the *δ-exceedance* upper record values. If we have the sequence of i.i.d. X's having d.f. F(x) and pdf f(x), then the joint pdf of the first m upper records X(1),X(2),…,X(m) is given in this scheme by equalities:

$$f_{X(1),X(2),\ldots,X(m)}(x_1,\ldots,x_m) = \frac{f(x_1)}{1-F(x_1+\delta)}\cdots\frac{f(x_{m-1})}{1-F(x_{m-1}+\delta)}f(x_m), \quad x_j > x_{j-1}, \quad j = 2,3,\ldots,m$$

Suppose that $F(x) = 1 - e^{-x}$, $x \geq 0$, ≥ 0,then

$$f_{X(1),X(2),\ldots,X(m)}(x_1,\ldots,x_m) = e^{(m-1)\delta}e^{-x_m}, \quad x_j > x_{j-1} + \delta, \quad j = 2,3,\ldots$$

The marginal pdf of X(m) is given as

$$f_{X(m)}(x) = \frac{1}{\Gamma(m)}(x - m\delta)^{m-1}e^{-(x-m\delta)}, \quad x > m\delta$$

Thus the random variable $X(m) - m\delta$ has the gamma distribution and one can see that for any *m* the vector {X(1),X(2),…, X(m)} has the same distribution as the vector $\{Y_1, Y_1 + Y_2,\ldots, Y_1 + Y_2 + \ldots + Y_m\}$, where Y_i's are independent and have the exponential distribution with $F(x) = 1 - e^{-(x-\delta)}$, $x \geq \delta$.

From this fact, it follows that interrecord values X(m)−X(m−1), m = 2,3,…, are independent.

Records with Restrictions I

Consider now the so-called *records with restrictions*, which are rather close in some sense to the δ-exceedance record scheme. Consider the sequence of i.i.d. random variables $X_1, X_2,\ldots$ and fixed some positive constant C. We take $X(1) = X_1$ and $L(1) = 1$ as the first record value and the first upper record time correspondingly. The next record times L(n) and record values X(n) are defined as follows:

$$L(n) = \min\{j > L(n-1): X(n-1) < X_j \leq X(n-1) + C\}, \quad X(n) = X_{L(n)}, \quad n = 2,3,\ldots$$

It means that we ignore as a record the new coming observation, if it exceeds essentially the previous record value.

Suppose that the initial random variables have a common pdf f(x). It is not difficult to get the expression for the conditional pdf $f_n(x_n|x_1,x_2,\ldots,x_{n-1})$, that is the expression for the pdf of X(n) under the condition that $X(1) = x_1, X(2) = x_2,\ldots, X_{n-1} = x_{n-1}$, where $0 < x_j - x_{j-1} \leq C$, $j = 2,3,\ldots,n-1$, are fixed. One obtains in this case that

$$f_n(x_n|x_1,x_2,\ldots,x_{n-1}) = f(x_n) - (F(x_{n-1}+C) - F(x_{n-1})) \quad [19]$$

if $x_{n-1} < x_n \leq x_{n-1} + C$, and $f_n(x_n|x_1,x_2,\ldots,x_{n-1}) = 0$ otherwise.

These equalities allow to find the corresponding relations for probability density functions $f_{n-1}(x)$ and $f_n(x)$ of random variables X(n−1) and X(n):

$$f_n(x) = f(x)\int_{x-C}^{x}(f_{n-1}(u)/(F(u+C) - F(u)))du, \quad n = 2,3,\ldots \quad [20]$$

From [20] one gets that the joint pdf $f_n(x_1,x_2,\ldots,x_n)$ of record values X(1),X(2),…,X(n) has the following form:

$$f_n(x_1,x_2,\ldots,x_n) = f(x_1)\frac{f(x_2)}{(F(x_1+C) - F(x_1))}\cdots\frac{f(x_n)}{(F(x_{n-1}+C) - F(x_{n-1}))} \quad [21]$$

if $x_1 < x_2 \leq x_1 + C$, $x_2 < x_3 \leq x_2 + C,\ldots, x_{n-1} < x_n \leq x_{n-1} + C$, and $f_n(x_1,x_2,\ldots,x_n) = 0$ otherwise.

These relations can be simplified essentially for the case of exponential distributions of X's.

Let us consider the sequence $Z_1, Z_2,\ldots$ of i.i.d. random variables having the standard E(1) d.f. $F(x) = \max\{0, 1-\exp(-x)\}$. Consider also $Z_0 = 0$. Denote the corresponding record (with the restriction $C > 0$) values and record times as

$$0 = Z(0) < Z(1) < Z(2) < \ldots \quad \text{and}$$
$$0 = T(0) < T(1) < T(2) < \ldots$$

One gets from [21] that the joint pdf of random variables Z(1),Z(2),…,Z(n) is given as

$$f_n(x_1,x_2,\ldots,x_n) = \exp(-x_n)/(1-\exp(-C))^n \quad [22]$$

if $0 < x_1 \leq C$, $x_1 < x_2 \leq x_1 + C,\ldots, x_{n-1} < x_n \leq x_{n-1} + C$, and $f_n(x_1,x_2,\ldots,x_n) = 0$ otherwise.

It follows from [22] that the joint pdf $g(v_1,v_2,\ldots,v_n)$ for differences

$$V(1) = Z(1) - Z(0), V(2) = Z(2) - Z(1), \ldots, V(n) = Z(n) - Z(n-1)$$

has the form

$$g(v_1,v_2,\ldots,v_n) = \exp\{-(v_1+v_2+\ldots+v_n)\}/(1-\exp(-C))^n, \quad 0 \leq v_1 \leq C, \ldots, 0 \leq v_n \leq C \quad [23]$$

This formula allows us to state that interrecord values V(1),V(2), …, V(n) are independent and identically distributed with the pdf

$$g(x) = \exp(-x)/(1-\exp(-C)), \quad \text{if } 0 \leq x \leq C, \text{ and } g(x) = 0, \quad \text{otherwise}$$

It means that vectors

$$\{Z(1), Z(2), \ldots, Z(n)\} \text{ and } \{W_1, W_1 + W_2, \ldots, W_1 + W_2 + \ldots + W_n\}$$

where independent random variables $W_1, W_2, \ldots, W_n$ have the pdf [23], are identically distributed. Thus the exponential random value Z(n) is presented as the sum of n i.i.d. random variables.

Note that the means and the variances of $W_1, W_2, \ldots$ are given as

$$a(c) = EW_k = \left(1 - (1+C)e^{-C}\right)/\left(1 - e^{-C}\right)$$

and

$$\sigma^2(C) = DV_k = (1 - (2 + C^2)e^{-C} + e^{-2C})/(1 - e^{-C})^2, \quad k = 1, 2, \ldots$$

It is possible based on these facts to obtain, in particular, different limit theorems for Z(n). For example, it follows immediately that random variables $(Z(n) - na(c))/\sigma(c)n^{1/2}$ tend as $n \to \infty$ to the standard normal distribution. Indeed, all results formulated above for the exponential E(1) random variables $Z_1, Z_2, \ldots$, can be rewritten analogously for the general case, when initial random variables have pdf's:

$$f(\alpha, \lambda) = \lambda \exp(-\lambda(x - \alpha)), \quad x \geq \alpha \quad \text{where} \quad \lambda > 0 \quad \text{and} \quad -\infty < \alpha < \infty$$

Records with Restrictions II

The next scheme is very close to the previous one. Now we also fix some constant $C > 0$, which determines the acceptable rate of exceeding of the previous record value. The discrepancy of these two schemes is the following. In the second case, we do not ignore the observations, which are essentially greater than the previous record value X(n). In this situation, we simply determine the new record value X(n + 1) as X(n) + C. In the case when the new coming observation belongs to the interval (X(n),X(n) + C), it is announced as the record value X(n + 1).

If we again consider the initial independent E(1)-distributed random variables (with pdf $f(x) = \exp(-x)$, $x \geq 0$) $Z_1, Z_2, \ldots$ and denote the records with such type of restrictions as $0 = Z(0) < Z(1) < Z(2) < \ldots$, then arguments analogous to those, which are given in the previous model, show that differences (interrecord values) $V_k = Z(k) - Z(k-1)$, $k = 1, 2, \ldots$, are also independent. In this case, we get that

$$P\{V_k < x\} = 1 - \exp(-x), \quad \text{if } 0 \leq x \leq C, \text{ and} \quad P\{V_k < x\} = 1, \quad \text{if } x > C. \qquad [24]$$

In particular

$$P\{V_k = C\} = \exp(-C), \quad k = 1, 2, \ldots$$

One can see that in this record scheme for any $n = 1, 2, \ldots$ the distribution of the vector $\{Z(1), Z(2), \ldots, Z(n)\}$ coincides with the distribution of the vector $\{V_1, V_1 + V_2, \ldots, V_1 + V_2 + \ldots + V_n\}$, where $V_1, V_2, \ldots$ are i.i.d. random variables having the d.f. [24].

Records with Confirmation

One more nonclassical record scheme – the so-called 'records with confirmation' (confirmed records), was considered in Sagateljan (2008) and Nevzorov and Saghatelyan (2009). In the simple option of this scheme to determine a new record value X(n + 1), one must wait for appearance of k (k is fixed number) observations, which will be greater than the previous record value X(n), and only after this event it is possible to determine X(n + 1) as the maximal value among these k observations. This scheme is useful in the situation when among X's, which have some fixed distribution, can appear an observation having another distribution (contamination of the original sample). The following new result generalizes this case. Let us consider the situation, when a new record will coincide with the mth ($1 \leq m \leq k$) in order of observations taken from k random variables, which are greater than the previous record value. The reason why we consider here the case when m can be less than k is rather simple. For example, imagine that we know that the given sample can be contaminated with some outliers and we expect that these outliers are presented in the sample by the 'top' observations. Hence, it is natural to delete these extraneous observations from the further consideration.

As above, the most interesting case here is connected again with the exponential distribution. The following result is valid.

Theorem 8

Let $F(x) = \max\{0, 1 - \exp(-x)\}$, and $1 \leq m \leq k$, defined above, be fixed. Then for any $n = 1, 2, \ldots$ the following equality holds:

$$\widehat{X}(n+1)^d = \widehat{X}(n) + X_1^{(n)}/k + \ldots + X_m^{(n)}/(k - m + 1) \qquad [25]$$

where $\widehat{X}(n), X_1^{(n)}, \ldots, X_m^{(n)}$ are independent and $P\{X_j^{(n)} < x\} = \max\{0, 1 - \exp(-x)\}$, $1 \leq j \leq k$.

Corollary

It follows from Theorem 8, that

$$\widehat{X}(n)^d = \mu_1/k + \mu_2/(k-1) + \ldots + \mu_m/(k - m + 1) \qquad [26]$$

where $\mu_1, \mu_2, \ldots$ are independent gamma-distributed random variables.

From this theorem, which is formulated for the exponential record values, it can be obtained (using the standard methods from the theory of records) that a set of all possible types of the limit distributions G(x) for the records with confirmations coincides with the corresponding set for the classical records:

$$G_i(x) = \Phi(-\ln(-\ln H_i(x))), \quad i = 1, 2, 3, \qquad [27]$$

where

$$H_1(x) = \exp\{-\exp(-x)\}, \quad -\infty < x < \infty$$

$$H_2(x) = 0, \text{ if } x < 0, \text{ and } H_2(x) = \exp\{-x^{-\delta}\}, \text{ if } x > 0, \ \delta > 0$$

$$H_3(x) = \exp\{-(-x)^{\delta}\}, \quad \text{if } x < 0, \ \delta > 0, \quad \text{and}$$

$$H_3(x) = 1, \quad \text{if } x > 0,$$

and $\Phi(x)$ is the normal N(0,1) distribution function.

Indeed, the centering and normalizing constants in this situation will differ from the corresponding constants for the classical records.

A more complete theory of records is given in monographs (Ahsanullah, 1988, 1995, 2004; Ahsanullah and Nevzorov, 2001; Arnold et al., 1998; Nevzorov, 2000, 2001). Different results for record values can be found in the Bibliography section.

Acknowledgments

The work of the first author was supported by Summer Research Grant of Rider University. The work of the second author was supported by grant NSh 13-02-00338 and by the grant of St. Petersburg State University N 6.38.672.2013.

See also: Nonparametric Regression; Nonparametric Statistics: Advanced Computational Methods; Nonparametric Statistics: Asymptotics; Nonparametric Statistics: Rank-Based Methods.

Bibliography

Adke, S.R., 1993. Records generated by Markov sequences. Statistics & Probability Letters 18, 257–263.

Ahsanullah, M., Nevzorov, V.B., 2001. Ordered Random Variables. Nova Science Publishers, New York, NY.

Ahsanullah, M., Nevzorov, V.B., 2004. Characterizations of distributions by regression properties of records. Journal of Applied Statistical Science 13, 33–39.

Ahsanullah, M., Nevzorov, V.B., 2005. Order Statistics. Examples and Exercises. Nova Science Publishers, New York, NY.

Ahsanullah, M., 1978. Record values and the exponential distribution. Annals of the Institute of Statistical Mathematics 30, A, 429–433.

Ahsanullah, M., 1979. Characterizations of the exponential distribution by record values. Sankhya B41, 116–121.

Ahsanullah, M., 1981. On a characterization of the exponential distribution by weak homoscedasticity of record values. Biometrical Journal 23, 715–717.

Ahsanullah, M., 1987. Two characterizations of the exponential distribution. Communications in Statistics: Theory and Methods 16, 375–381.

Ahsanullah, M., 1988. Introduction to Record Statistics. Ginn Press, Needham Heights, MA.

Ahsanullah, M., 1995. Record Statistics. Nova Science Publishers, Inc., Commack, NY.

Ahsanullah, M., 2004. Record Values Theory and Applications. University Press of America, Maryland, USA.

Akhundov, I., Nevzorov, V.B., 2008. Characterizations of distributions via bivariate regression on differences of records. In: Ahsanullah, M., Yanev, G.P. (Eds.), Records and Branching Processes. Nova Science Publishers, pp. 27–35.

Akhundov, I., Berred, A., Nevzorov, V.B., 2007. On the influence of record terms in the addition of independent random variables. Communications in Statistics: Theory and Methods 36, 1291–1303.

Andel, J., 1990. Records in an AR(1) process. Ricerche di Matematica 39, 327–332.

Arnold, B.C., Balakrishnan, N., Nagaraja, H.N., 1998. Records. John Wiley & Sons, NY.

Bairamov, I.G., 2000. On the characteristic properties of exponential distribution. Annals of the Institute of Statistical Mathematics 52, 448–452.

Balakrishnan, N., Nevzorov, V.B., 2006. Record values and record statistics. In: Encyclopedia of statistical sciences, the second ed., vol. 10. Wiley- Interscience, pp. 6995–7006.

Balakrishnan, N., Balasubramanian, K., Panchapakesan, S., 1996. δ-exceedance records. Journal of Applied Statistical Science 4 (2–3), 123–132.

Ballerini, R., Resnick, S., 1985. Records from improving populations. Journal of Applied Probability 22, 487–502.

Ballerini, R., Resnick, S., 1987. Records in the presence of a linear trend. Advances in Applied Probability 19, 801–828.

Berred, A., Nevzorov, V.B., Wey, S., 2005. Normalizing constants for record values in Archimedean copula processes. Journal of Statistical Planning and Inference 133, 159–172.

Biondini, R., Siddiqui, M.M., 1975. Record values in Markov sequences. In: Statistical Inferences and Related Topics –2. Academic Press, New York, pp. 291–352.

Chandler, K.N., 1952. The distribution and frequency of record values. Journal of the Royal Statistical Society. Series B 14, 220–228.

Deheuvels, P., Nevzorov, V.B., 1993. Records in F^{α}-scheme I. Martingale properties. Notes of Scientific Seminars of POMI 207, 19–36 (in Russian). English version: Journal of Mathematical Sciences 81 (1996), 368–2378.

Deheuvels, P., Nevzorov, V.B., 1994a. Limit laws for k-record times. Journal of Statistical Planning and Inference 38, 279–308.

Deheuvels, P., Nevzorov, V.B., 1994b. Records in F^{α}-scheme II. Limit theorems. Notes of Scientific Seminars of POMI 216, 42–51 (in Russian). English version: Journal of Mathematical Sciences 88 (1998), 29–35.

Deheuvels, P., 1984. The characterization of distributions by order statistics and record values – a unified approach. Journal of Applied Probability 21, 326–334 (Correction, 22, 997).

Dziubdziela, W., Kopocinsky, B., 1976. Limiting properties of the k-th record values. Zastos. Mat. 15, 187–190.

Foster, F.G., Stuart, A., 1954. Distribution free tests in time-series band on the breaking of records. Journal of the Royal Statistical Society B16, 1–22.

Gulati, S., Padgett, W.J., 2003. Parametric and Nonparametric Inference from Record Breaking Data. Springer.

Gupta, R.C., 1984. Relationships between order statistics and record values and some characterization results. Journal of Applied Probability 21, 425–430.

Haiman, G., 1987. Almost sure asymptotic behavior of the record and record time sequences of a stationary Gaussian process. In: Puri, M.L., Vilaplana, J.P., Wertz, W. (Eds.), Mathematical Statistics and Probability Theory, vol. A. D. Reidel, Dordrecht, The Netherlands, pp. 105–120.

Houchens, R.L., 1984. .Record Value Theory and Inference. Ph.D. Thesis. University of California, Riverside.

Nagaraja, H.N., 1978. On the expected values of record values. Australian Journal of Statistics 20, 176–182.

Nagaraja, H.N., 1982. Record values and extreme value distributions. Journal of Applied Probability 19, 33–239.

Nevzorov, V.B., Balakrishnan, N., 1998. Record of records. In: Balakrishnan, N., Rao, C.R. (Eds.), Handbook of Statistics, vol. 16. Elsevier, Amsterdam, pp. 515–570.

Nevzorov, V.B., Saghatelyan, V., 2009. On one new model of records. Proceedings of the Sixth St-Petersburg Workshop on Simulation 2, 981–984.

Nevzorov, V.B., Balakrishnan, N., Ahsanullah, M., 2003. Simple characterization of Student's t_2- distribution. The Statistician 52 (part 3), pp. 395–400.

Nevzorov, V.B., 1984. Record times in the case of nonidentically distributed random variables. Theory of Probability and Its Applications 29, 808–809.

Nevzorov, V.B., 1987. Records. Theory of Probability and Its Applications 32, 201–228.

Nevzorov, V.B., 1990. Generating functions for the kth record values – a martingale approach. Zap. Nauchn, Semin. LOMI 184, 208–214 (in Russian). Translated version in Journal of Soviet Mathematics 44, 510–515.

Nevzorov, V.B., 1992. A characterization of exponential distributions by correlation between records. Mathematical Methods of Statistics 1, 49–54.

Nevzorov, V.B., 1995. Asymptotic distributions of records in nonstationary schemes. Journal of Statistical Planning and Inference 44, 261–273.

Nevzorov, V.B., 2000. Records. Mathematical Theory. Phasis, Moscow (in Russian).

Nevzorov, V.B., 2001. Records: mathematical theory. Translations of Mathematical Monographs. American Mathematical Society 194.

Nevzorova, L.N., Nevzorov, V.B., Balakrishnan, N., 1997. Characterizations of distributions by extremes and records in Archimedean copula process. In: Johnson, N.L., Balakrishnan, N. (Eds.), Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz. Wiley, NY, pp. 469–478.

Pfeifer, D., 1982. Characterizations of exponential distributions by independent nonstationary record increments. Journal of Applied Probability 19, 127–135 (Correction, 19, 906).

Pfeifer, D., 1984. Limit laws for inter-record times from non-homogeneous record values. Journal of Organizational Behavior 1, 69–74.

Pfeifer, D., 1991. Some remarks on Nevzorov's record model. Advances in Applied Probability 23, 823–834.

Renyi, A., August 1–10, 1962. Theorie des elements saillants d'une suite d'observations. Colloquim on combinatorial methods in probability theory. Math. Inst., Aarhus University, Aarhus, Denmark. 104–117. See also: Selected papers of Alfred Renyi, vol. 3 (1976), Akademiai Kiado, Budapest, 50–65.

Resnick, S.I., 1973. Limit laws for record values. Stochastic Processes and their Applications 1, 67–82.

Sagateljan, V.K., 2008. On one new model of record values. Vestnik of St-Petersburg University 1 (3), 144–147 (in Russian).

Shorrock, R.W., 1972a. A limit theorem for inter- record times. Journal of Applied Probability 9, 219–223.

Shorrock, R.W., 1972b. On record values and record times. Journal of Applied Probability 9, 316–326.

Siddiqui, M.M., Biondini, R.W., 1975. The joint distribution of record values and inter-record times. Annals of Probability 3, 1012–1013.

Smith, R.L., Miller, J.E., 1986. A non-Gaussian state space model and application in prediction of records. Journal of the Royal Statistical Society: Series B 48, 79–88.

Smith, R.L., 1988. Forecasting records by maximum likelihood. Journal of the American Statistical Association 83, 331–338.

Stepanov, A.V., 1992. Limit theorems for weak records. Theory of Probability and Its Applications 37, 586–590.

Tata, M.N., 1969. On outstanding values in a sequence of random variables. Z. Wahrscheinlichkeitsth. verw. Geb 12, 9–20.

Vervaat, W., 1973. Limit theorems for records from discrete distributions. Stochastic Processes and their Applications 1, 317–334.

Williams, D., 1973. On Renyi's record problem and Engel's series. Bulletin of the London Mathematical Society 5, 235–237.

Yang, M.C.K., 1975. On the distribution of the inter-record times in an increasing population. Journal of Applied Probability 12, 148–154.

Nonphysician Health Professionals

Amy W Stimpfel, New York University, New York, NY, USA
Jill M Vanak and Linda H Aiken, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The organizational context of medical practice changed substantially over the past half century, resulting in new opportunities and a more favorable environment for the evolution of allied and nonphysician health professionals. This article chronicles the organizational developments leading to the growth in nonphysician health professions and the changing medical division of labor. Recent trends in nursing, the rise of advanced practice nurses and physician assistants (PAs), and the status of physical and occupational therapists are examined.

The Changing Context of Health Care

The organizational complexity of health care, including the rise of physician group practices and decline of the solo family physician, the entrance of for-profit hospitals and health systems, and the growing dominance of professional managers without a clinical background, has changed the nature of the patient–provider relationship (Starr, 1982). These changes coupled with societal pressures from better informed patients have resulted in the evolution of the physician role and relationship with patients, from one of dominance and authority to one of collaborator and shared-decision maker (Mechanic, 1986). Powerful stakeholders, including allied health professionals, nonphysician providers, and administrators have entered the system, often with a goal that differs from that of the physician. Hospital administrators, motivated to find cost-effective ways to deliver health care, began looking to nonphysician providers to expand their role in the provision of health services. This shift in provider configuration represents a trend occurring internationally.

With the global rise of health care costs, the emergence of national governments' involvement in the delivery and management of health care has grown extensively since the twentieth century. This has been evidenced in the United States through public policy with programs like Medicare and Medicaid, which provide insurance for the elderly, poor, and disabled. European nations have experienced government involvement in the form of austerity initiatives, designed to trim deficits and reduce services, which has led to public uncertainty about the future of health care in these nations.

Advances in biomedical research and technology, expansion of health insurance, increased incidence of chronic conditions, and expanding social definitions of the boundaries of medical care all contributed to increased utilization of health services and changing patterns of care. These changes fostered the development of an expanded division of medical labor in which a new group of health professionals emerged to complement and compete with physicians in the delivery of a broader array of health services. At the beginning of the twentieth century, physicians accounted for about one in three health care workers; by the 1980s, the comparable ratio was about 1 in 16 (Mechanic, 1983). This trend has continued into the twenty-first century with physicians accounting for a smaller and smaller share of the total health care workforce.

Changing Disease Patterns

Progress in medical science including antibiotics and vaccines as well as improved standards of living contributed to shifting the burden of illness from acute, infectious illnesses to complex, chronic conditions. While advances in modern medicine have produced important treatments for chronic illnesses, most have resulted in technologies that extend life but do not cure the underlying disease. Examples include long-term renal dialysis for the treatment of kidney failure, antiretroviral drugs that extend the lives of people with acquired immunodeficiency syndrome, and antineoplastic agents that extend survival in people with certain cancers.

Chronic conditions are accompanied by a myriad of functional limitations and social support needs that fall outside the conventional definition of medical practice. Nurses, social workers, occupational therapists, physical therapists, and other nonphysician health professionals have demonstrated interest and expertise in helping the chronically ill and frail elderly manage the sequelae of chronic conditions and medical treatments, thus creating new roles for themselves in the medical division of labor.

Accessibility of Health Care Services

Along with the technological and medical developments of the mid-twentieth and early twenty-first centuries came increasingly expensive, complex, inefficient, and unsustainable health care systems. Calls for comprehensive health reform ensued in many countries. After decades of public debate, the United States finally passed incremental health insurance reform, the Patient Protection and Affordable Care Act (PPACA), under the Obama administration, which was signed into law in early 2010. The PPACA required sweeping changes to the way health care is accessed and paid for by individuals and employers in the United States. Many of the components of the health reform directly impact nonphysician health care providers. Namely, a large cadre of newly insured citizens will be accessing

the health care system. Although the patterns of use and types of providers needed to accommodate an expanded patient population remain unknown, nonphysician providers are poised with the skills and knowledge to care for this newly insured group.

Organizational Changes Conducive to Expansion of Nonphysician Professions

As a result of the PPACA, approximately 30 million of the estimated 50 million uninsured people in the United States will have improved access to health care providers as of 2014. Many of the newly insured are likely to seek primary care services, which may be problematic because primary care physicians are in short supply due to a number of factors, including migration to specialty practices and changing practice patterns including less willingness of physicians to work long hours as in years past. Changes in reimbursement that favor specialty practice and heavy financial debt associated with medical school in the United States have contributed to a smaller share of US physicians electing careers in primary care. As more physicians gravitate toward specialized medicine, opportunities are increasing for nonphysician providers to enter the market for primary care services.

Globally, the health care environment has never been more receptive to nonphysician providers. Americans are not unique in their struggle with an increasingly costly, fragmented, and inequitable health care system. Rich and poor countries alike are facing the challenges associated with aging populations and increased demand for health care services without adequately trained, distributed, and diverse numbers of health care providers (World Health Organization, 2006). The Lancet Commission on the Future of Health Professions, written by a team of interdisciplinary experts, recommended a series of action items and reforms to redesign health provider's education, addressing the gaps that exist both in and across countries. The report highlights competency-based curricula, interprofessional education, use of information technology systems, and creation of global networks to share knowledge and resources as key pieces to a transformed education system for health professionals (Frenk et al., 2010).

Allied Health Professionals

The allied health workforce is diverse in term of roles, settings of practice, and educational requirements. Over 200 occupations fall under the 'allied health' umbrella, which contributes to the difficulty in distilling a singular definition or a complete list of allied health professionals. The Association of Schools of Allied Health Professions describes allied health professionals as being engaged in the delivery of health services such as evaluation and prevention of disease, nutritional services, and rehabilitation. Typically, physicians, dentists, nurses, pharmacists, chiropractors, and social workers are excluded from definitions of allied health because these disciplines are often educated in identifiable professional schools, whereas the other disciplines that comprise allied health professionals may be educated together. However, it is still common for the term allied health professionals to be used to refer to all nonphysician health professionals.

Allied health professionals fill a critical role in modern health care and are integral members of the health care team. Because many of the occupations have developed as a result of technological advances such as ultrasound technologists, the group of allied health professionals is constantly evolving and expanding. Two of the group's largest and oldest professions are occupational and physical therapists and are detailed in the following section.

Evolution of Occupational and Physical Therapy

Occupational therapy traces its origins back 200 years to a French physician's observation that mental patients given menial tasks to perform improved in functional capacity faster than those who were idle. At the end of the eighteenth century, American physician Benjamin Rush introduced work as a treatment for mentally ill patients at Philadelphia's Pennsylvania Hospital, and in 1906, the first training course for occupational therapists was established in Boston (Institute of Medicine, 1989). World War I created a demand for rehabilitation of large numbers of injured soldiers, which served as a major impetus for expanding the scope of practice of occupational therapy to physically as well as mentally disabled persons, and for establishing the new occupation of physical therapy.

Education and Characteristics

The two world wars stimulated the growth of physical and occupational therapists and the early recruits to both occupations were mainly women. Over time they have remained predominantly female occupations, although about a third of physical therapists are men compared with about 10% for nurses and occupational therapists. The first American training program in physical therapy began in 1917, as a short course for women aides. The success of these aides following World War I resulted in federal funding to increase the supply of therapists during World War II through accelerated programs to train college graduates in related fields such as physical education for roles in therapy. University-based degree-granting programs in physical therapy were established concurrently. Physical therapy has thus been a university-based educational program almost from its inception.

In 1979, the American Physical Therapy Association advocated for graduate level preparation for the profession by 1990. Although it took longer than perhaps desired, by 1 January 2016, the doctor of physical therapy (DPT) will be the required degree for all entry-level physical therapy programs. While therapists with a baccalaureate degree may still practice, the changes in educational standards have not been limited to future cohorts of students; many licensed practitioners have earned postprofessional DPT degrees or a transition DPT approximately 11 000 as of 2010. Like

registered nurses (RNs), physical therapists must be a graduate of an accredited physical therapy education program and pass the licensure exam prior to practicing. Physical therapists must be licensed by the state in which they practice and must renew their license on a regular basis. Physical therapy does not support career mobility from the physical therapy aide category to the professional category and has established nonoverlapping educational requirements for the two groups thus avoiding the problem nursing has faced. Physical therapy has also developed viable options for private practice, which allows the occupation to continue to evolve independently, to some extent, from institutional managerial control of salaried practice. Physical therapy is considered to be more visible and autonomous than occupational therapy for these reasons (Institute of Medicine, 1989).

The first formal 'professional' standards for education in occupational therapy were established in 1923, as 12-month training programs for high school graduates, in contrast to physical therapy, which moved directly to university-based education. As of January 2007, however, the entry-level degree for occupational therapists is postbaccalaureate, either a master's or doctoral degree. Licensure is state based; requirements include graduation from an accredited occupational therapy program, fieldwork, and passing the certification examination before obtaining a license as an occupational therapist. Technical training as an occupational therapy assistant occurs primarily in community colleges and in some 4-year university settings. Graduating from an accredited program and passing a licensing exam are also requirements from occupational therapy assistants.

Roles and Work Settings

Physical therapists plan and administer treatment to relieve pain, improve functional mobility, and limit the physical disability of people with injury or illness. Increasingly, physical therapists specialize within an area of practice such as children or orthopedics and sports medicine. The rapid growth in joint replacements overall and to younger patients as medical technology has advanced has significantly increased the demand for physical therapists in recent years. Most physical therapists work in private offices, clinics, or hospitals; other employers include home health agencies and nursing or residential care facilities.

Occupational therapists promote learning of skills necessary to perform daily tasks, diminish or correct pathology, and promote and maintain health. Occupational therapy focuses on both mental and physical disabilities, and some occupational therapists specialize in particular age groups such as the elderly. Occupational therapists working with the mentally ill or disabled in institutional and community-based settings typically provide activities that help their patients learn to cope with daily stresses and manage their work and leisure time more efficiently (Institute of Medicine, 1989). Like physical therapists, most occupational therapists work in hospitals and private offices, although nursing homes and home health care agencies are other employers.

Nursing

Nursing is the largest of the health professions with approximately 3 million RNs in the United States and roughly 19 million nurses and midwives worldwide. The American Nurses' Association Social Policy Statement (2010) defines the nursing profession as "… the protection, promotion, and optimization of health and abilities, prevention of illness and injury, alleviation of suffering through the diagnosis and treatment of human response, and advocacy in the care of individuals, families, communities, and populations" (American Nurses Association, 2010). This broad definition coupled with a valuable set of skills, knowledge, and public trust has allowed for role expansion of nurses both domestically and internationally.

Historical Perspectives

Modern nursing dates back to mid-nineteenth century, England, when hospitals were disorganized, unsanitary, staffed by unskilled workers, and were generally held in low esteem. Only the destitute sought help in hospitals, except in wartime. Florence Nightingale, a wealthy and educated English woman, accepted the challenge of bringing order to military hospitals during the Crimean War and in doing so began a social experiment that would change the nature of hospitals worldwide. Nightingale identified trained nurses as the vehicle for transforming hospitals but she recognized that nurses capable of skilled work and responsible ward management would have to be recruited from the better educated classes. To legitimize the work of nursing to middle to upper class women, she drew on the well-respected traditions of religious education and military-style authority. Young nurses motivated by a sense of dedication and service to mankind worked long hours for low wages under strict discipline (Abel-Smith, 1960).

Nightingale capitalized on the opportunity created by the war to not only become the founder of modern nursing, but also to become one of the first to use proportional statistics (Cohen, 1984). By tracking deaths by cause in military hospitals in Crimea, she demonstrated that more soldiers died of hospital environmental causes than of injuries received in battle. Nightingale's social position and statistical prowess forged a new future for hospitals and the nurses who worked in them.

Vestiges of History in Contemporary Nursing Roles

Nursing's history leaves a legacy that gives contemporary professional nurses special legitimacy in the medical division of labor and a high level of public trust. However, that same legacy creates barriers to the continued evolution of the profession of nursing in contemporary health care. Nursing developed as an ancillary occupation supportive to physicians. While nurses gained much of their professional status and public legitimacy from their close association with physicians, their subordinate status relative to medicine has hindered their quest for greater autonomy (Freidson, 1970). Medical dominance over nurses is linked to the organizational frame of the modern hospital where the majority of nurses practice

and where physicians exercise considerable professional authority. Indeed, nursing's quest for professional autonomy was not realized to any great extent until they began to practice in larger numbers and in a broader array of roles outside the hospital.

The association of early nurse training with religious orders has reinforced a continuing expectation of service and charity. These attributes are no longer valued as highly in Western culture, and no longer serve as sufficient motivating forces to recruit and retain professional nurses. Despite a shift from caring and nurturing functions to more technological responsibilities akin to those associated with male roles in other occupations, nursing has retained a feminine image and is thus subject to sex-role stereotyping and sex discrimination in professional relationships and labor practices. Although the nursing profession remains largely white and female dominated (U.S. Department of Health and Human Services, 2008), more gender, racial, and ethnic diversity has become a priority (Institute of Medicine, 2011).

Nursing Education

Professional nursing education has undergone a dramatic transformation since the 1970s, that is still underway globally (Lynaugh and Brush, 1996). Nursing education developed in hospital apprentice-like programs with a large service component. Until the Great Depression forced many graduate nurses into hospital employment, hospital nursing care was supplied largely by students in training. Leaders in American nursing advocated early in the twentieth century for the standardization of nursing education within institutions of higher education and a baccalaureate degree as minimum entry to practice, with the Goldmark Report being published in 1923. The Flexner Report (1910) had brought about similar reform in medical education in the United States. Subsequent reports and commissions supported the recommendation culminating in the American Nurses Association in the late 1970s setting 1985 as the target year for requiring the baccalaureate degree as minimum entry into practice. While progress has been made to this end, the new century dawned without this goal having been achieved in the United States, although other countries including Australia, New Zealand, Iceland, Chile, Norway, and Spain have successfully implemented minimum academic degree requirement for new entrants into nursing.

Nursing has been described as an incompletely closed profession because of its inability to set a minimum educational requirement (Freidson, 1970). Three generic types of education programs prepare professional nurses: 3-year hospital diploma programs, 2-year associate degree programs, and 4-year college or university programs. Following graduation from an accredited program, a state licensing exam must be passed in order to practice as an RN. Neither employers nor legal practice acts distinguish between professional nurses by type of education. The proportion of US nurses whose highest educational preparation was a diploma has dramatically decreased, while community college programs offering associate degrees increased. From 1980 to 2008, for example, the percentage of nurses educated at the diploma level decreased from 55 to 14% and associate degrees rose from 18 to 36%. In comparison, the rise of baccalaureate-prepared nurses was not as dramatic, only increasing from 22 to 37% in the same time frame. In 2008, only 13% of nurses held a master's degree, doctoral degree, or higher (Institute of Medicine, 2011). Multiple entry pathways into nursing have led to confusion for the public, as well as other health care professionals. The history of nursing is intertwined with the history of nursing education and nursing's quest for a professional identity (Allen et al., 2006). Although the goal of requiring a bachelor's degree as a minimum entry into practice has not been achieved in the United States, calls to standardize entry into practice continue.

In 2010, the Institute of Medicine released the landmark report, *The Future of Nursing*, which advocated for increasing the number of baccalaureate-prepared nurses in the workforce to 80% by 2020, also known as '80 by 20' (Institute of Medicine, 2011). Empirical research has identified a clear association between higher levels of nurse education and improved patient outcomes, including mortality (Aiken et al., 2011). Achieving '80 by 20' depends upon a variety of mechanisms including the creation of seamless pathways to Bachelor of Science in Nursing (BSN) education, funding BSN capacity expansion, increasing profiles in nursing leadership, establishing employer incentive programs, and enhancing diversity in the workforce.

Advanced Practice Nurses and Physician Assistants

Nurse practitioners (NPs), also known as advanced practice registered nurses (APRNs), and physician assistants (PAs) are two relatively new additions to the medical division of labor, both originating in the United States in the 1960s. NPs represent more of an evolution in the profession of nursing than a distinct new occupation while PAs constitute a new nonphysician profession. An APRN is an RN with a postgraduate degree and continued education in nursing. APRNs demonstrate effective integration of theory, clinical practice, and experience with degrees of autonomy in patient interventions. APRN education supports four recognized areas of specialization: nurse anesthetists (Certified Registered Nurse Anesthetists/CRNAs), nurse midwives (Certified Nurse Midwives), clinical nurse specialists, and NPs. In contrast, PAs are interdependent semiautonomous clinicians who practice in collaboration with physicians (Schneller, 1978). The impetus for the development of PAs and NPs was a shortage of primary care in the United States. However, over time PAs have followed the trends in physician specialization and now over half of PAs are in specialty practice. NPs are more likely to practice in primary care settings and care for underserved populations.

Evolution of NP and PA

While nursing's growth is strongly tied to the development of the modern hospital, nurses have a long and distinguished history in public health and home care, settings in which they exercised considerable autonomy and professional leadership. Lillian Wald invented the term 'public health nurse' in 1893 to describe a role for a nurse that was a combination of health promotion, illness care, and social reform (Buhler-Wilkerson, 1993). Public health nursing and public health in general

underwent substantial erosion in the United States with the rise of health insurance after World War II. However, the foundation for the emergence of NPs was the legacy of public health nursing.

In the case of NPs, however, these factors created a favorable environment for expanding the scope of nursing practice that had long been advocated by nursing's leaders and that was exemplified by the early public health nurses (Fairman, 2008). Thus, NPs are really not new at all but nurses with legally sanctioned expanded scope of practice including some domains formerly controlled solely by physicians, such as drug prescription authority (Lynaugh and Brush, 1996; Mezey et al., 2003).

The evolution of the development of PAs is often linked to a perceived shortage of generalist physicians in the United States in the 1960s and the market demand for physician substitutes. The PA profession was first proposed in 1961 when it was recommended to the American Medical Association to "create two new groups of assistants to doctors from nonmedical and nonnursing personnel." Most initial recruits to PA programs were experienced medical corps personnel and the initial training programs were short-term on-the-job training. While the profession was predominantly male in its inception, it is now almost equally male and female, and the most recent graduating classes are over 60% female (Hooker and Cawley, 1997). The early programs were located at medical schools but did not have academic standing or lead to degrees. Over time, degree-granting programs have been incorporated into schools of allied health.

Education and Characteristics

In the United States, a master's degree (or doctoral degree beginning in 2015 for CRNAs) is required for certification in advanced practice nursing. Education of the NP has been rapidly incorporated into the large existing university-based infrastructure of schools of nursing. Enrollments in NP programs have grown steadily and by 2009 there were about 150 000 NPs in the United States, with schools producing roughly 5000–6000 new NPs per year.

PA education is generally 24 months in length, following the acquisition of a bachelor's degree. A major difference between PAs and NPs is that PAs have accepted a legally dependent role working under the direction of a physician while NPs have fought for and obtained a legal scope of practice that does not require physician supervision (Sage and Aiken, 1997). In the United States, the profession is represented by the American Academy of Physician Assistants. Similar providers have different titles in other countries such as clinical officers in Africa, assistant medical officers in Malaysia, and assistant doctors in China. Upon graduation for accredited training programs, PAs must pass the national certifying examination of the National Commission on Certification of PAs, an independent accrediting agency, after which they must complete continuing medical education every 2 years and pass a recertification examination every 6 years in the United States.

Essentially, the role of the NP and the PA is quite similar in regard to scope of practice, function, and role as both professional groups can diagnose, treat, and prescribe. The training of each group varies in that PAs are trained under a generalist curriculum, modeled after medical school programs. All PAs learn primary care and then rotate through major specialty clinicals while in training. NPs are traditionally trained in a focused approach, specializing in a specific area of medicine or with focus on a specific population.

Both the NP and PA professions constitute an evolving health care workforce and the face of a new generation of health care workers. In most countries in which the role of an APRN exists, NPs provide a range of services that may otherwise be performed by physicians in primary care, including diagnostics, screenings, prescriptions of pharmaceuticals or medical tests, prevention and health education activities, the monitoring of patients with chronic illnesses, and care coordination (Delamaire and Lafortune, 2010). The education and training requirements to become an APRN vary both within and across countries; however, most require a graduate degree in nursing or a master's degree. The United States was the first country to introduce the right for nurses to prescribe drugs in the mid-1970s, followed by the United Kingdom, Australia, and Taiwan in the 1990s. The role of the NP has expanded dramatically around the world since its introduction in the 1960s, but a continued need for the development of the role to address health promotion, disease prevention, and to provide for evidence-based care to underserved global populations is warranted (Savrin, 2009). Worldwide, countries are exploring the concept of the PA as an efficient and quick means to train and employ highly autonomous health care workers to serve their populations. Professional organizations are proactively providing information about PAs within the global community and several countries are now utilizing US-trained PAs in pilot programs or in established training curriculums. A number of PA programs have affiliation agreements with institutions outside of the United States to host PAs for clinical training. With over 100 training programs globally, the PA model can be adapted to the needs of specific nations.

Roles and Work Settings

NPs and PAs provide generalist or specialty care including routine physical assessment, treatment of uncomplicated episodic illness, continuing care for persons with stable chronic conditions, and acute care in hospital settings. The role of the NP was developed in 1965 due to a need for a cost-effective avenue for increased access in rural communities in the state of Colorado. NP training programs developed in the 1970s with certification programs followed by master's level curriculum. The role of the NP has continued to expand in part due to legislative victories leading to an increasing independent role in practice. Such legislative initiatives include the NP being recognized for third party reimbursement, along with the addition of prescriptive authority. By the turn of the twenty-first century, NP programs had entered the mainstream of graduate nursing education.

The scope of practice that defines boundaries for NPs and PAs is determined by state licensing boards, as no federal regulation exists. Regulations vary widely and tend to focus on standards for physician oversight, prescribing authority and ratio of physicians to practitioners. Consequently, this makes

policy and regulation developments complex when distinguishing the role of physician providers versus the nonphysician health care workforce. As of 2013, 11 states allow NPs to practice independently of a physician, but the majority of states limit nonphysician providers' ability to prescribe medications. Although state boards of nursing have successfully lobbied to define a legal scope of autonomous practice for NPs, attempts by physicians to exert control over their practice and remuneration is a continuing source of conflict (Sage and Aiken, 1997). Licensing boards in the United States recognize PAs as practitioners authorized to perform medical tasks delegated to them by physicians. Legislation empowers PAs to perform any clinical task within the scope of practice and sanctioned by their collaborating and supervising physician (Mittman et al., 2002).

The passage of the PPACA marks a significant opportunity for the nursing profession to impact the delivery of health care. The designation of nurse-managed health clinics as eligible for federal funding will allow nurses to be trained in primary care and will increase access to citizens. The role of nursing will be highlighted by a number of programs directed at transitional and chronic care management, a role to which nurses are adeptly skilled. Nurses and nonphysician providers are tasked with the role of using the existing legislation to showcase the value of the professions in providing cost-effective, high-quality care to the population.

Global Perspectives

Maldistribution of health care providers is universal. Globally, few countries escape challenges in the number, quality, and skill mix of their health care workers (Chen et al., 2004). There is an estimated need for 6 million new health care workers, physicians, nurses, and midwives, based on the World Health Organization's goal of 2.3 health workers per 1000 people. Sub-Saharan Africa accounts for only 3% of the world's health care workers yet carries nearly a quarter of the disease burden compared to the Americas, which only account for 10% of the global disease burden and claim 37% of the health care workers (World Health Organization, 2006).

Allied health workers provide much of the medical care in poor countries but are in short supply and often poorly trained. The potential of nurses and APRNs to improve access to and quality of health care in developing countries is well recognized. However, many countries have insufficient resources to create employment opportunities for qualified nurses.

The inequitable distribution of health workers is seen in the nursing population, with nurses migrating from poor countries to wealthier ones. Low salaries, poor working conditions, lack of health care infrastructure, and social stigma associated with the profession are some reasons that nurse migration is widespread. Once nurses leave their native country, the working conditions become even worse for the nurses remaining, perpetuating a cycle of migration (Kaelin, 2011).

Thus, nonphysician providers and allied health occupations have good prospects for future growth in developing countries as well as in the sophisticated and highly technological health systems of industrialized countries.

See also: Healing Practices: Alternative and Complementary; Health Care Markets: Theory and Empirical Evidence; Health Insurance: Economic and Risk Aspects; Occupational Health.

Bibliography

Abel-Smith, B., 1960. A History of the Nursing Profession. Heineman, London.

Aiken, L.H., Cimiotti, J., Sloane, D.M., Smith, H.L., Flynn, L., Neff, D., 2011. The effects of nurse staffing and nurse education on patient deaths in hospitals with different nurse work environments. Medical Care 49 (12), 1047–1053.

Allen, M.P., Allison, M.M., Stevens, S., 2006. Mapping the literature of nursing education. Journal of the Medical Library Association: JMLA 94 (Suppl. 2), E122–E127.

American Nurses Association, 2010. Nursing's Social Policy Statement, third ed. American Nurses Association, Silver Spring, MD.

Buhler-Wilkerson, K., 1993. Bringing care to the people: Lillian Wald's legacy to public health nursing. American Journal of Public Health 83 (12), 1778–1786.

Chen, L., Evans, T., Anand, S., Boufford, J.I., Brown, H., Chowdhury, M., Cueto, M., Dare, L., Dussault, G., Elzinga, G., Fee, E., Habte, D., Hanvoravongchai, P., Jacobs, M., Kurowski, C., Michael, S., Pablos-Mendez, A., Sewankambo, N., Solimano, G., Stilwell, B., de Waal, A., Wibulpolprasert, S., 2004. Human resources for health: overcoming the crisis. Lancet 364 (9449), 1984–1990.

Cohen, I.B., 1984. Florence nightingale. Scientific American 250, 128–132.

Delamaire, M., Lafortune, G., 2010. Nurses in Advanced Roles: A Description and Evaluation of Experiences in 12 Developed Countries. OECD Health Working Papers, No. 54. OECD Publishing.

Fairman, J., 2008. Making Room in the Clinic: Nurse Practitioners and the Evolution of Modern Health Care. Rutgers University Press, New Brunswick, NJ.

Freidson, E., 1970. Professional Dominance: The Social Structure of Medical Care. Aldine, Chicago, IL.

Frenk, J., Chen, L., Bhutta, Z.A., Cohen, J., Crisp, N., Evans, T., Fineberg, H., Garcia, P., Ke, Y., Kelley, P., Kistnasamy, B., Meleis, A., Naylor, D., Pablos-Mendez, A., Reddy, S., Scrimshaw, S., Sepulveda, J., Serwadda, D., Zurayk, H., 2010. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. Lancet 376 (9756), 1923–1958.

Hooker, R.S., Cawley, J.F., 1997. Physician Assistants in American Medicine. Churchill Livingstone, New York, NY.

Institute of Medicine, 1989. Allied Health Services: Avoiding Crises. National Academies Press, Washington, DC.

Institute of Medicine, 2011. The Future of Nursing: Leading Change, Advancing Health. National Academies Press, Washington, DC.

Kaelin, L., 2011. A question of justice: assessing nurse migration from a philosophical perspective. Developing World Bioethics 11 (1), 30–39.

Lynaugh, J.E., Brush, B.L., 1996. American Nursing: From Hospitals to Health Systems. Blackwell, Cambridge, UK.

Mechanic, D., 1983. Handbook of Health, Health Care and the Health Professions. Free Press, New York.

Mechanic, D., 1986. From Advocacy to Allocation: The Evolving American Health Care System. Free Press, New York.

Mezey, M.D., McGivern, D.O., Sullivan-Marx, E. (Eds.), 2003. Nurses, Nurse Practitioners: Evolution to Advanced Practice, fourth ed. Springer Publication Company, New York, NY.

Mittman, D.E., Cawley, J.F., Fenn, W.H., 2002. Physician assistants in the United States. British Journal of Medicine 325 (7362), 485–487.

Sage, W., Aiken, L.H., 1997. Regulating interdisciplinary practice. In: Jost, T. (Ed.), Regulation of Health Professions. Health Administration Press, Chicago, IL, pp. 71–101.

Savrin, C., 2009. Growth and development of the nurse practitioner role around the globe. Journal of Pediatric Health Care 23 (5), 310–314.

Schneller, E.S., 1978. Physician's Assistants: Innovation in The Medical Division of Labor. Lexington Books, Lexington, p. 18.

Starr, P., 1982. The Social Transformation of American Medicine. Basic Books, New York.

U.S. Department of Health and Human Services, 2008. The Registered Nurse Population: Findings from the 2008 National Sample Survey of Registered Nurses. Health Resources and Services Administration, Bureau of Health Professions, Washington, DC.

World Health Organization, 2006. The World Health Report 2006: Working Together for Health. World Health Organization, Geneva, Switzerland.

Relevant Websites

www.aapa.org – American Academy of Physician Assistants.
http://nursingworld.org/ – American Nurses Association.
http://www.apta.org – American Physical Therapy Association, 2011. Today's Physical Therapist: A Comprehensive Review of a 21st Century Health Care Profession.
http://www.asahp.org/ – Association of Schools of Allied Health Professions, Allied Health Professionals.
http://www.bls.gov/ooh/healthcare/physical-therapists.htm – Bureau of Labor Statistics, 2012a. Physical Therapists.
http://www.bls.gov/ooh/healthcare/occupational-therapists.htm – Bureau of Labor Statistics, 2012b. Occupational Therapists.

Nonprofit Sector, Geography of

Vandana Desai, Royal Holloway, University of London, Egham, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W. Bielefeld, volume 16, pp. 10688–10691, © 2001, Elsevier Ltd.

Abstract

This article provides a critical overview of the changing role of the nonprofit sector, its definition, its evolving functions in the last few decades such as advocacy and empowerment, existing diversity within the sector, and various challenges that it faces in the twenty-first century.

Introduction: Scale and Diversity of the Nonprofit Sector

Nonprofit organizations exist worldwide and there is a huge range of diversity in activities that they take on and the scale of their operations. Much of the early research in the 1970s was done in the United States, but since the 1980s the study of nonprofit sectors has grown rapidly in other countries around the world especially in the developing world. Research has shown that the scope and dimensions of nonprofit organizations vary within and between countries, regions, cities, and neighborhoods. Nationally, too, nonprofit sectors vary in every country in terms of the numbers of nonprofit organizations, the services they provide, their revenues, and their employment (Salamon et al., 1999). Most of this research has involved case studies or comparisons among a small number of regions within each country, undoubtedly because the difficulty of acquiring data on these areas, the diversity of nongovernmental organizations (NGOs), and the diverse contextual factors that exist in their operations make generalizations difficult (see Desai and Preston, 2000).

While philanthropy and nonprofit organizations have long been considered in a number of social science disciplines, the nonprofit sector as a distinct field of study is a fairly recent phenomenon. Sparked mainly by a growing recognition of the importance of the nonprofit sector and debates about the future of the welfare state, interest in the nature of nonprofit organizations (also known as the third sector) has grown rapidly. At the global level in the 1980s significant growth took place in nonprofit organizations; this growth was stimulated in part by welfare state reorganization, which resulted in the substitution of private service provision for government services. Growth in nonprofit organizations in the 1980s was also due to increased activity in the areas of development and environmentalism, the rise in middle-class expectations, and improvements in communication technology. For example this rise is seen in the number of NGOs operating in the United States, estimated to be 1.5 million (http://www.humanrights.gov/2012/01/12/fact-sheet-non-governmental-organizations-ngos-in-the-united-states/ (visited on 26.10.13.)); on the other hand in India that number is estimated to have been around 3.3 million NGOs in 2009 (http://southasia.oneworld.net/news/india-more-ngos-than-schools-and-health-centres#.Une35xC8bX8 (visited on 26.10.13.)).

As a result NGOs are difficult to define (see the broad definition below) and classify and hence many different classifications are in use. The most commonly used framework of classification includes orientation (referring to the type of activities it takes on as its main focus, which might include human rights, combating child poverty, environmental, or development work) and level of operation (indicating the scale at which an organization works, such as local, regional, national, or international) (Vakil, 1997). Nonprofit research has shown that the diversity and scale can be due to broad societal patterns as well as factors that are more specific and contextual to that location or region. In the last two decades lots of research has taken place on NGOs in both the North and South to understand their functioning and explore patterns of their behavior. Broad patterns include historical development and social, economic, and political factors such as class relationships; state–society interactions; the nature of the legal environment; and the legitimacy of and trust in collective action through nongovernment. More specific factors include government policy about welfare funding and service delivery; the degree of social heterogeneity; the generosity of particular segments of the community; the needs and desires for public or semipublic goods, by either disadvantaged or middle/upper-class citizens; and the resources available for nonprofit organizations, including money, labor, and social entrepreneurs.

Relatively few systematic spatial studies of nonprofit organizations have been carried out at the neighborhood level. Some neighborhood studies have been descriptive and have used geographical information systems techniques to relate providers and needs spatially, within or across neighborhoods in metropolitan areas. In a typical example, a project by the Center on Nonprofits and Philanthropy at the Urban Institute used a spatial analysis to examine the relationships between the nonprofits that serve families and children and the socioeconomic characteristics of three low-income neighborhoods in the District of Columbia. The purpose was to show how neighborhood environments, including the presence of nonprofit organizations, affected the well-being of community residents (De Vita et al., 2000). A study of nonprofits across Dallas, Texas (Bielefeld et al., 1997) found that nonprofit organizations were more strongly influenced by demographic factors such as resources available locally and the specific need of the community in their immediate areas, than by the demographic characters of areas farther afield. These results indicate that the nonprofit sector is closely tied to the dynamics

of local communities (e.g., their levels of involvement and participation in programs and their own initiatives). Similar findings were reached by Desai (1999; Desai and Preston, 2000) in Mumbai, India. In Bangladesh for example NGOs have actively promoted neighborhood self-help savings groups, e.g., Grameen Bank or Kudumbashree in Kerala (India). Local formal savings or microfinance groups or microenterprises have emerged in communities as an approach to eradicate poverty supported by or linked to external organizations such as credit unions and banks. NGOs have been proactive in promoting such groups among less-educated, low-income women, because such women are the least likely to have access to formal means of saving and borrowing.

Definition of the Nonprofit Sector

Nonprofit researchers are interested in defining the nonprofit sector, delineating its scope, and understanding the factors determining its existence and activities. While definitions vary, it is generally agreed that the nonprofit sector is a set of social entities that are mainly formally organized, privately initiated, self-governing, involved in development issues on a nonprofit basis, and working to mobilize public spirit and altruism to create social and civic benefits for the needy. The term NGO is understood to refer to those autonomous, nonmembership, relatively permanent or institutionalized (but not always voluntary), intermediary organizations, staffed by professionals or the educated elite, which work with grassroots organizations in a supportive capacity. Grassroots organizations on the other hand are issue-based, often ephemeral, membership organizations; they may coalesce around particular goals and interests and then dissipate once their immediate concerns have been addressed. NGOs have become an important and vocal platform for the involvement of civil society in public affairs and may collaborate with grassroots/community-based organizations. There is also a distinction made between northern and southern NGOs, based on the organizations' location of operations. NGOs based in northern developed countries tend to fund programs and activities of NGOs in the southern developing countries. Northern NGOs seek funding from multilateral and government aid agencies and also develop their own financial base through private fund-raising activities. Recently a lot of southern NGOs have also started to aspire to financial independence and do their own private fund-raising activities locally.

NGOs are popular because they demonstrate unique characteristics and capabilities: they are perceived to be flexible, open to innovation, and able to access the poor through work at the grassroots level. NGOs play three main roles: service delivery, advocacy on behalf of the poor, and empowerment.

Functions of the Sector

Service Delivery

NGOs have come to play an increasingly important role in the formulation and implementation of policy, becoming key actors in the political economy of development. They have increasingly collaborated with governments and aid agencies in the developing world as the belief has grown that NGOs can offer an alternative model of development and play a key role in processes of democratization and decentralization of policies.

NGOs are seen as more administratively flexible, closer to the poor, innovative in problem solving, and more cost effective than corresponding state partners. Donor pressure toward structural reform and privatization in the Southern Hemisphere in the 1990s underlies the increased interest in NGOs as 'service deliverers,' part of a wider and explicit objective to facilitate productive NGO–state partnerships. There is a realization among donor countries in the global north that aid is becoming ever more complex as new instruments and players emerge and as criticisms over the effectiveness of aid increase. Private financial flows from corporate sector organizations in northern developed countries have increased rapidly in the wake of increased liberalization. These increases are in part the result of enhanced understanding of corporate social responsibilities and of the creation of organizations such as the Bill Gates Foundation, the Ford Foundation, and the Aga Khan Foundation. These developments have also led to an expansion in partnerships between northern NGOs and southern NGOs and have changed attitudes toward development interventions in the North. A view took hold that merely transferring resources in the form of tools or funds was not an adequate response to alleviating poverty, which was rooted in deeper structural problems. Indeed such transfers could just preserve the situation by creating financial dependency. Greater collaboration between corporations and NGOs can also create inherent risks of cooptation for the weaker partner, typically the nonprofit organization involved (Baur et al., 2012).

Former UN Secretary General Kofi Annan wrote in favor of international humanitarian intervention, arguing that the international community has a 'right to protect' citizens of the world against ethnic cleansing, genocide, and crimes against humanity. Northern NGOs such as Red Cross and Oxfam have a humanitarian function and respond to emergencies, providing short-term relief and long-term rehabilitation to victims of war and of natural or man-made disasters. These organizations raise money in the North from the general public, the private sector, and governments to pay for their work; they share funds as much as possible with their southern counterparts, help in building the capacity of southern NGOs, educate their own constituencies in the North about the underlying causes of poverty, and draw people into active lobbying and campaigning for change. Southern NGOs have the basic responsibility among NGOs for leading the development process in developing countries and the expertise to do so.

Policy Advocacy

The other frequent role for NGOs is policy advocacy, seeking social change by influencing attitudes, policy (see Pollard and Court, 2008), and practice; seeking to reform state services on the basis of NGO experiences and to lobby directly for the policy changes and reform. This is involvement in participatory, public-interest politics; NGOs engaging in such activity realize the increasing importance of information, which they gather through their own experience of working at the

grassroots level. They use the power of ideas and information to promote positive change in the wider structures of government and the official aid community. These NGOs often play a catalytic or seeding role by demonstrating the efficacy of a new idea, publicizing it, perhaps persuading those with access to greater power and budgets to take notice, and then encouraging the diffusion and adoption of the idea. In recent decades, though, NGOs have devoted their energies to advocacy campaigns. Drawing on new opportunities afforded by growth in their numbers and size and in the resources available to them, and facilitated by the revolution in communications technologies, NGOs have globalized, seeking to maximize their impact on global policy processes by working through transnational networks. In terms of environmental issues and sustainable development, the Earth Summit in Rio in 1992 was the first to show the power of international NGOs, when about 2400 representatives of NGOs came to play a central role in deliberations and highlighted that NGO transnational networking is now quite extensive.

The Jubilee 2000 Campaign was the first global campaign through which development NGOs united over a single issue – debt relief for the world's most heavily indebted nations. Jubilee 2000 mobilized over 24 million people from over 60 nations to sign a petition (Mayo, 2005: 174), drawing supporters of all ages and occupations from faith-based and secular organizations, trade unions, and businesses. With 110 member organizations, within 4 years of its launch in 1997, "Jubilee 2000's campaign for the remission of unpayable Third World debt made a remarkable impact, raising the issue's profile internationally as well as locally" (Mayo, 2005: 172). After a concerted lobbying campaign, in 1999 at the Cologne G8 Summit, world leaders agreed to cancel $100 billion of debts owed by the poorest nations (Bedell, 2005: 19). The campaign is credited with having contributed to poverty alleviation, enabling diversion of resources from debt repayment to increased spending on health care, housing provision, and education. Similar success is attributed to the 'Make Poverty History' campaign in London in 2005.

In 1995, Oxfam International established its Washington Advocacy Office to coordinate the development of joint policy and strategies for the nationally based Oxfams. Its location in Washington, DC, facilitated direct access to Washington-based multilateral agencies, the World Bank, the International Monetary Fund, and the United Nations, headquartered in New York. By 2012, Oxfam International supported advocacy and campaigning staff in Washington, DC, Geneva, New York, Brussels, Brasilia, and Addis Ababa (Oxfam International, 2012).

In the United Kingdom, Bond, the membership body for NGOs working in international development, established in 1993, had 400 members by October 2014 (Bond, 2014). Listed first among its activities, Bond "facilitates collective action by its members and wider UK civil society to influence the policies and practice of governments and institutions at UK, European and international levels" (Bond, 2012).

This trend has also reflected northern NGOs' reaction to calls from their southern partners to increase their engagement in advocacy. Because northern NGOs' greater access to power and resources enabled them to engage with global institutions and influence their policies, southern NGOs consistently called for northern NGOs to lobby and campaign about the global crises of poverty and environmental degradation. Concurrently, an evolving emphasis in development practice on the importance of empowering local communities to participate in all stages of the development process resulted in recognition that growing numbers of southern NGOs were best placed to initiate and manage development work in local communities. Northern NGOs thus responded by looking for new ways to contribute to poverty alleviation.

In recent decades, northern NGOs' financial commitment to advocacy has grown; this growth is evident in the budgets, staff time, and the departments dedicated to research and lobbying. In 1996, just 4.1% of total expenditure by international NGOs was dedicated to advocacy (Anderson, 2007). Anderson tracked the financial commitment of 14 international NGOs to policy research, lobbying, and public campaigning and found that between 1996 and 2005 this commitment increased dramatically. Only three increased their resource allocation to advocacy by less than 100%; over half increased the financial and human resources allocated to advocacy by three or six times (Anderson, 2007). Thus, in recent decades, NGOs have given greater priority to advocacy, prompted by the success of earlier campaigns.

Empowerment

The roles of service delivery and policy advocator include the process of empowerment. Some NGOs naturally progress from filling a gap in service delivery to recognizing the need to look outward to the wider context in which the need arises and find themselves drawn, possibly through involvement in NGO networks (Bebbington, 2003), into national or global policy advocacy (e.g., cancellation of debt of heavily in debt nations, placing it on the agenda of the G8 summit in Gleneagles in July 2005 – see Brown et al., 2012). This type of 'bottom-up democracy' has been successful in many instances in which it might eventually lead to 'top-down political change.'

NGOs have become key actors in a process of transformative development. They can effect norm changes that lead to regime change or restructuring of world politics. They do this (especially northern NGOs) through the communicative power of information, lobbying, research, campaigning, or media work, acting on the basis of their moral authority (for example, in the area of human rights (see Van Tuijl, 1999), free trade, debt relief, child labor, or fair trade). NGOs create alliances and networks to place pressure on the state (DeMars, 2005). Recently, NGOs have been drawn into networks of growing complexity. This has led for some to call for "power to be put back into empowerment" (Batliwala, 2007a), not only recognizing power as a vital component of the concept of empowerment, but also appreciating that power structures can hinder or affect an individual or group's ability to exercise power. Batliwala (2007a) suggests that notions of empowerment can be seen in the Protestant Reformation in Europe and, as a concept, can be traced through Quakerism, Jeffersonian democracy, early capitalism, and the black power movement. She also argues that empowerment as an idea is evident in the Veerashaiva movement in twelfth- and thirteenth-century Karnataka, India, which agitated against "caste and gender oppression, (calling) for the redistribution of power and access to spiritual knowledge through the destruction of these forms of social stratification" (2007a: 558). In

1970, Paulo Freire highlighted the importance of developing a critical consciousness, or to use his term 'conscientization' as a first step in processes of empowerment. He believed that everyone has the capability to participate in raising his or her own consciousness. This has been taken on by many in the NGO sector in various aspects of their work and has become part of many movements (e.g., feminism), and the importance of collective action has been adopted by those who understand the crucial elements of empowerment as power from *within* and power *with*.

Challenges Facing the Nonprofit Sector

Among the challenges facing the nonprofit sector are the following.

Sustainability and Long-Term Impact of NGO Operations

An important role for NGOs in the last few decades has been in mitigating the adverse costs of structural adjustment and promoting donor reform packages in offering insurance against a political backlash against harsh adjustment regimes. Such a role raises important questions. Patterns of service delivery through the voluntary sector may lack compatibility and coordination. Insofar as such efforts rely on government funding or private donations, the ultimate sustainability of NGOs is questioned. Will NGOs be able to fund their own operations and activities in the future? Will NGO funding create dependency? At a deeper level, there are worries about the long-term impact of NGO service provision on the sustainability of national health and education systems (rather than individual programs) and about access to quality services for all. NGOs also constantly face the challenge of the need to adopt approaches that 'best fit' in the context of a particular country, region, or community, and existing institutions rather than importing 'best practices,' which may not apply in certain local contexts (Dodman and Mitlin, 2013).

Accountability, Evaluation, and Performance

Funders generally require reporting and assessment; such information is not necessarily publicly available. Instead of relying on financial data or solely quantitative measures, qualitative evaluations might be more useful. Such evaluations could examine organizations' transparency, governance, and program effectiveness as well as the feedback mechanisms designed for donors and beneficiaries (Lowell et al., 2005). More generally, the popular discourse of nonprofit evaluation should move away from financial notions of organizational effectiveness and instead move toward assessing the impacts of programs that are implemented and focus on the beneficiaries who are the recipients of programs.

Legitimacy

As nonstate actors with considerable influence over governance in many areas, NGOs have generated concerns about the extent to which they represent the views of the public and the extent to which they allow the public to hold them to account. In recent decades NGOs have increased the number and range of their activities to a level where they have become increasingly dependent on a limited number of funders/donors (Edwards and Hulme, 1995). Consequently competition has increased for funding, as have the expectations of the donors themselves. This runs the risk of donors adding conditions that can threaten the independence of NGOs, and an overdependence on government aid has the potential to dilute "the willingness of NGOs to speak out on issues which are unpopular with governments" (Edwards and Hulme, 1996). In these situations NGOs are being held accountable by their donors, which can erode rather than enhance their legitimacy, a difficult challenge to overcome. Some commentators have also argued that changes in the sources of NGO funding have ultimately altered their functions.

The expansion of the nonprofit sector has also been accused of damaging the public sector in many developing countries. The mismanagement of NGOs has resulted in the breakdown of public health care systems in many Asian countries, such as India, Thailand, and Cambodia. Instead of promoting equity and alleviating poverty, NGOs have been under scrutiny for contributing to socioeconomic inequality and disempowering the services in the public sector of developing countries.

Administrative Cost

In recent years concerns have been raised regarding the level of administrative costs (e.g., office expenses, salaries) incurred by NGOs in managing projects and programs. The sector as a whole has developed various codes of ethics; for example, the World Bank typically allows 37%. High administrative costs can generate criticism that NGOs are being run simply to benefit the people working for them. Studies by the Urban Institute and the Center for Social Innovation at Stanford University have shown how rating agencies create incentives for nonprofits to lower and hide overhead costs, which may actually reduce organizational effectiveness by starving organizations of the infrastructure they need to deliver services effectively (Lowell et al., 2005).

Conclusion

Debates about the most appropriate forms of development intervention have persisted among development practitioners and theorists. Theorizing about development in the 1980s, notably by radical development theorists, contributed to understandings that 'underdevelopment' was a condition maintained by inequitable global structures, such as entrenched unfair terms of trade, poor commodity prices, oppressive debt obligations, and inequitable distribution of resources among different social groups and nations. Simultaneously, NGOs grappled with the uncertainties of development practice as they endeavored to find the most effective means to improve quality of life for those suffering the ravages of poverty and disadvantage. One common criticism of NGOs was that their focus on local development and empowerment meant they failed to consider the wider structural factors that perpetuated poverty (Edwards and Hulme, 1992). Decades of local development efforts also led to the realization that project work in communities is insufficient to bring about broader changes to address the ongoing causes of poverty. Responding to this insight and to

the new theoretical understandings, NGOs sought to tackle causal factors by increasing their commitment to advocacy, with the aim of maximizing the impacts and cost-effectiveness of their work.

Working collaboratively to broaden their impact, NGOs were drawn into networks of growing complexity. The primary impetus was growing awareness of the power of public action through experience of its success. Although arguably "little progress at the level of ideology and global systems" (Nyamagasira, 2002: 8) results from NGO action, through uniting in alliances NGOs "materially contributed to policy changes by Northern and Southern governments" (Anderson, 2002: 84). These included the success of campaigns relating to baby milk marketing codes, drafting of an international essential drugs lists, trade liberalization relating to clothing manufacture in southern nations, action against global warming and rain forest destruction, debt relief to African nations, fair trade pricing, and the imposition of sanctions to combat apartheid (Anderson, 2002).

See also: Development Geographies; Globalization: Geographical Aspects; Local Economic Development; Postcolonial Geography; Poverty, Geography of; Social Movements, Geography of.

Bibliography

Anderson, I., 2002. Northern NGO advocacy: perceptions, reality, and the challenge. In: Eade, D. (Ed.), Development and Advocacy. Oxfam Great Britain, Oxford.

Anderson, I., 2007. Global actions: international NGOs and advocacy. In: Rugendyke, B. (Ed.), NGOs as Advocates for Development in a Globalising World. Routledge, Abingdon and New York.

Atack, I., 1999. Four criteria of development NGO legitimacy. World Development 27 (5), 855–864.

Batliwala, S., 2007a. Taking the power out of empowerment-an experiential account. Development in Practice 17 (4–5), 557–565.

Batliwala, S., 2007b. Putting power back into empowerment. Open Deomcracy Available from. http://www.opendemocracy.net/node/34195/pdf (accessed on 06.04.14.).

Baur, Dorothea, Schmitz, Hans Peter, 2012. Corporations and NGOs: when accountability leads to co-optation. Journal of Business Ethics 106 (1), 9–21.

Bebbington, A., 2003. Global networks and local developments: agendas for development geography. Tijdschrift voor Economische en Sociale Geografie 94 (3), 297–309.

Bebbington, Anthony, Hickey, Samuel, Mitlin, Diana (Eds.), 2008. Can NGOs Make a Difference?: The Challenges of Development Alternatives. Zed Books, London.

Bedell, G., 2005. Make Poverty History: How You Can Help Defeat World Poverty in Seven Easy Steps. Penguin Books, London.

Bielefeld, W., Murdoch, J.C., Waddell, P., 1997. The influence of demographics and distance on nonprofit location. Nonprofit and Voluntary Sector Quarterly 26, 207–225.

Bond, 2014. Bond for International Development About us. Available online at www.bond.org.uk/pages/about-us.html (accessed 23.10.14.).

Brown, L.D., Ebrahim, A., Batliwala, S., 2012. Governing international advocacy NGOs. World Development 40 (6), 1098–1108. This article provides a comparative analysis of international advocacy NGOs, examining governance, organisational structure and the effectiveness of their advocacy.

DeMars, William E., 2005. NGOs and Transnational Networks: Wild Cards in World Politics. Pluto Press, London.

Desai, Vandana, Imrie, Rob, 1998. The new managerialism in local governance: north/south dimensions. Third World Quarterly 19 (4), 98.

Desai, Vandana, 1999. Anatomy of the Bombay NGO sector. Environment and Urbanization 11 (1), 247–265.

Desai, Vandana, Preston, Ian, 2000. Urban grassroots non-governmental organisations in Bombay: a suggested typology. Environment and Planning C: Government and Policy 18, 453–468.

De Vita, C.J., Manjarrez, C., Twombly, E.C., 2000. Organisations and Neighborhood Networks that Strengthen Families in the District of Columbia. Final Report to the Ammie E. Cassey Foundation. The Urban Institute, Washington, DC.

Dodman, David, Mitlin, Diana, 2013. Challenges for community-based adaptation: discovering the potential for transformation. Journal of International Development 25 (5), 640–659.

Edwards, Michael, Hulme, David (Eds.), 1992. Making a Difference: NGOs and Development in a Changing World. Earthscan Publications/SCF, London.

Edwards, Michael, Hulme, David (Eds.), 1995. Non-governmental Organisations – Performance and Accountability Beyond the Magic Bullet. Earthscan Publications/SCF, London.

Edwards, M., Hulme, D., 1996. Too close for comfort? The impact of official aid on non-governmental organisations. World Development 24 (6), 961–973.

Edwards, Michael, Fowler, Alan, 2002. The Earthscan Reader on NGO Management. Earthscan, London.

Freire, P., 1970. Pedagogy of the Oppressed. Penguin, London.

Hulme, David, Edwards, Michael (Eds.), 1997. NGOs, States and Donors: Too Close for Comfort?. Macmillan Press, London (in association with Save the Children). Probably the most comprehensive volume on NGDO relations with donors seen from an NGDO perspective. The papers describe the issue of co-optation and the danger of NGDO displacement from the poor and powerless to the more powerful.

Jude, Howell, Pearce, Jenny, 2001. Civil Society and Development: A Critical Exploration. Lynne Rienner Publishers, Boulder.

Kilby, Patrick, June 2006. Accountability for empowerment: dilemmas facing non-government organisation. World Development 34 (6), 951–963.

Lowell, Stephanie, Trelstad, Brian, Meehan, Bill, 2005. The ratings game: evaluating the three groups that rate the charities. Stanford Social Innovation Review, 39–45.

Mayo, M., 2005. Global Citizens: Social Movements and the Challenge of Globalization. Zed Books, London and New York. This accessible volume discusses the development of campaigning alliances by NGOs, community-based organisations and labour movements, case studies of social mobilisation and the challenges facing future attempts to transform global society.

Mitlin, D., Hickey, S., Bebbington, A., 2007. Reclaiming development? NGOs and the challenge of alternatives. World Development 35 (10), 1699–1720. This article explores the notion that NGOs would benefit from rethinking the notion of development alternatives, focussing on the political economy of social change, and reflects that the role of civil society has often been in conjunction with the political programs of social movements.

Najam, A., 1996. NGO accountability: a conceptual framework. Development Policy Review 14, 339–353.

Nyamagasira, W., 2002. NGOs and advocacy: how well are the poor represented. In: Eade, D. (Ed.), Development and Advocacy. Oxfam Great Britain, Oxford.

Pollard, Amy, Court, Julius, 2008. How civil society organizations use evidence to influence policy processes in Anthony Bebbington. In: Hickey, Samuel, Mitlin, Diana (Eds.), Can NGOs Make a Difference?: The Challenges of Development Alternatives. Zed Books, London, pp. 133–153.

Rugendyke, B. (Ed.), 2007. NGOs as Advocates for Development in a Globalising World. Routledge, Abingdon and New York Devoting increasing resources to advocacy campaigns directed at global actors, NGOs have globalised to maximize their impact. Reasons for these changes and the strengths, limitations and impacts of NGO advocacy at a range of scales are discussed.

Salamon, L.M., Anheier, H.K., et al., 1999. The Emerging Sector Revisited: A Summary. Institute for Policy Studies, Johns Hopkins University, Baltimore, MD.

Van Rooy, A., 2000. Good news! You may be out of a job: reflections on the past and future 50 years for northern NGOs. Development in Practice 10 (3/4) Oxford: Oxfam.

Van Tuijl, P., 1999. NGOs and human rights: sources of justice and democracy. Journal of International Affairs 52 (2), 493–512.

Vakil, Anna, December 1997. Confronting the classification problem: toward a taxonomy of NGOs. World Development 25 (12), 2057–2070.

Relevant Websites

www.bond.org.uk/ – Bond for International Development, a network of British NGOs.
http://www.globalaging.org/ – Global Action on Ageing.
www.jubileedebtcampaign.org.uk/ – Jubilee Debt Campaign.
www.makepovertyhistory.org/ – Make Poverty History.
http://www.oecd.dac – OECD Provides up-to date information about aid flows and analytical reports with a wealth of regularly updated statistical information.
http://www.dfid.gov.uk – Site of the British government's aid agency, the Department for International development.
http://www.usaid.gov – Site of the United States Agency for International Development.
http://www.worldbank.org – Site of the World Bank and including the International Development Association.

Nonsampling Errors

Elizabeth A Stasny, Ohio State University, Columbus, OH, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 10691–10696, © 2001, Elsevier Ltd.

Abstract

Data from sample surveys are subject to both sampling and nonsampling errors. The latter include errors from the sampling frame, the questionnaire, the interviewer, and the respondent. The article defines and sets out ways to deal with various errors. Sampling frame errors may be overcome by random digit dialing for telephone surveys or dual-frame sampling. Response error (incorrect answers) may be dealt with by randomized response techniques, while nonresponse errors may be minimized by weight adjustments and imputation methods. Panel rotation and survey mode errors are also briefly considered.

Data from sample surveys are subject to two types of errors: sampling and nonsampling errors. Sampling errors result from the fact that we only have information from a portion of the population of interest rather than from the whole population. Nonsampling errors include errors from all other sources: the sampling frame, the questionnaire, the interviewer, and the respondent. Data from both censuses and surveys may suffer from nonsampling errors, while only survey data are subject to sampling error.

Statisticians have traditionally been most concerned with controlling and measuring sampling errors, perhaps because they are easier to handle. Sampling errors are controlled by choice of the survey design; the size of the sampling error can be estimated using the survey data themselves. The emphasis of most classical survey sampling books such as Cochran (1977) and Kish (1965) is on measuring sampling error and designing surveys to reduce that error.

More recently, authors such as Lessler and Kalsbeek (1992) and Biemer et al. (1991) have focused on issues related to nonsampling error. Such errors must typically be assessed using information outside of the survey (or census) data or using a statistical model. As is noted by Mosteller (1978), sampling errors typically contribute to the variance of an estimate from survey data (and do not exist for census data), while nonsampling errors can contribute to both variance and bias of an estimate. In collecting survey or census data we try to control, reduce, or avoid nonsampling errors as much as possible because their size and effect on estimates are very hard to assess.

Frame errors, response errors, nonresponse errors, and other sources of errors are all nonsampling errors. The following describes these nonsampling errors and methods for dealing with them.

Frame Error

Definition of Frame Error

The sampling frame is the list from which units are drawn for the sample. The 'list' may be an actual listing of units, as in a phone book from which phone numbers will be sampled, or some other description of the population, such as a map from which areas will be sampled. Frame error results when the sampling frame is not an accurate and complete representation of the population of interest.

Frame errors arise from erroneous inclusions, erroneous exclusions, and multiple inclusions. Erroneous inclusions occur if the frame includes units not in the population of interest so that invalid units have a chance of being included in the sample. Erroneous exclusions occur when members of the population of interest are missing from the list so that they have no chance of being sampled. Multiple inclusions occur if some population members appear multiple times on the list so that they have a higher chance of being sampled.

For an easy example of these types of frame errors, consider sampling households within a city using the telephone white pages as the sampling frame. Some numbers listed in the phone book are no longer working numbers because the people in the household that had the number have moved out of the city; these are erroneous inclusions. Some households choose to have their phone numbers unlisted; these are erroneous exclusions. Some households have multiple listings in the phone book because they have additional telephone lines for a computer modem or for a fax machine or for the children; these are the multiple inclusions.

Frame error is a problem because a sample drawn from such a frame may not represent the population of interest, from which you think you are sampling, and some units may have a higher probability of being included in the sample than most units. These errors may result in sample estimates that are biased for the population quantity being estimated. For example, if wealthier households are more likely to have multiple phone lines, then those households are more likely to be included in the sample. This could lead to estimates of, say, amount spent on entertainment that are biased high.

Dealing with Frame Error

As is the case with all nonsampling errors, the best way to control frame error is to avoid or minimize it. Methods for doing this depend on the type of sampling frame that you are using. We will discuss two commonly used methods in this section: random digit dialing (RDD) and dual-frame sampling.

Random Digit Dialing

For telephone surveys of individuals in households, RDD is often used to avoid problems associated with using a telephone directory as the sampling frame. To create the sampling frame for an RDD sample, one must first list all working area codes in

the area to be sampled. Within each area code, one must list all working exchanges. If all possible four-digit numbers that could complete the phone number were used to make up the sampling frame of phone numbers, there would be numerous nonworking numbers and business numbers included. Because calling such numbers to determine that they are not valid household numbers is time-consuming and expensive, other methods may be used to reduce the number of invalid phone numbers in the frame. One method takes advantage of the fact that phone numbers are typically given out in banks of 100. Thus, the frame is taken to be the list of area codes with their exchanges followed by all possible first two digits of the remainder of the number. These eight-digit numbers are clusters of 100 possible phone numbers. A random sample of these clusters is taken and then two two-digit numbers are chosen at random to give two potential phone numbers within each cluster. The two sampled numbers within each cluster are called. If both are nonworking numbers or business numbers, then the cluster is discarded because it is likely that the remaining possible numbers in that cluster are also invalid numbers.

Once a valid household is reached, one still must determine who is to be interviewed within the household. If a single adult respondent is to be selected, a method proposed by Kish (1949) may be used to randomly select that adult. Because Kish's method involves creating a roster of all individuals within the household, which can be time-consuming and may be considered intrusive by potential respondents, other methods such as asking for the person who had the last birthday may be used as a quick method of selecting an individual within a household (see Lavrakas (1993: pp. 111–13) for a description of the last birthday and next birthday methods).

There are, of course, some problems with RDD samples. For example, persons who live in households with no telephone are not included in the sampling frame. Households with multiple telephone lines have a higher probability of being selected for the sample. One must account for this by asking respondents how many phone lines they have that are used for voice correspondence and use this information to adjust the selection probability for that household. If one is searching for a very specific population, such as households with disabled persons or early voters, then many of the households reached by telephone will be ineligible for the survey and must be screened out of the sample. The higher the proportion of households that must be screened out because of ineligibility, the higher the final cost of the survey will be.

Dual-Frame Sampling

In many sampling settings there are multiple lists or sources that can be used to develop a sampling frame. For example, the US Department of Agriculture uses lists of farms and area maps to select samples for their crop surveys. Samples from the lists of farms are easier to obtain but miss some farms; the area samples are difficult to handle but encompass all possible farms (see Bosecker and Ford, 1976 for details). A survey of patrons of the arts may use lists of season ticket holders to the symphony, ballet, and opera along with lists of persons who have made contributions to art museums. The challenge when multiple lists are used is to unduplicate the lists or to determine how many ways a person could be selected for the sample. Unduplicating lists to create a single sampling frame is complicated by problems such as incorrect spellings of names, use of complete first names versus initials, and spouses who may donate together or individually. Procedures for unduplicating lists must search for possible matches as well as for exact matches. These possible matches may need to be resolved by hand. Alternately, sampling may be done from the multiple lists and respondents may be asking how many of the lists they should appear on. This allows one to determine the probability that an individual was selected for the sample.

Response Error

Definition of Response Error

Response error occurs when a respondent intentionally or unintentionally gives an answer to a survey question that is not correct. An intentional response error occurs when the respondent purposely gives an incorrect answer, for example, to a question that asks for sensitive information about illegal activities (drug use, cheating on taxes), for personal information on sensitive topics (being a victim of rape), or about behavior that may not be perceived as socially acceptable (not voting, owning a gun). The respondent may give an intentionally incorrect answer for self-protection if he or she is concerned that the survey responses will not be held confidential and may be used against him or her. Or the respondent may not wish to appear in a bad light to the interviewer and may give the more socially acceptable answer.

Unintentional response errors can occur if questions are poorly worded and the respondent does not understand the question or terms in the question. The order in which questions and response choices are presented to a respondent can influence the responses that are given. In addition, response errors can occur if the respondent does not take time to think about a question that asks for an opinion and instead takes the easy option of agreeing with the statement or choosing the answer that is easiest to recall (called satisficing). (See Krosnick and Alwin (1987) for details on response-order effects and satisficing.)

Avoiding Response Error

Carefully designed and worded questionnaires, and, for in-person or telephone surveys, well-trained interviewers are the best tools for avoiding response errors (*see* Sample Surveys: Cognitive Aspects of Survey Design). Questionnaires must be pilot tested to ensure that respondents understand questions and that all possible responses are allowed for on closed-form questions.

Randomized Response Techniques for Sensitive Questions

Randomized response methods have been proposed by a number of researchers (see, for example, Warner, 1965; Nathan, 1988) to avoid the problem of having respondents intentionally give an incorrect answer to a question. The idea is to protect the respondent by ensuring that the interviewer is unaware of which question the respondent is answering.

There are several versions of the randomized response technique. In one, the respondent randomly selects one of two types of cards containing opposite versions of the question. For example,

Version A:
I have used cocaine at least once. YES/NO
Version B:
I have never used cocaine. YES/NO

Whichever the respondent answers, yes or no, the interviewer does not know whether the respondent has used cocaine or not. The data from such responses, however, can be used to estimate the proportion of the population that has used cocaine as long as there are unequal numbers of Version A and Version B cards.

Let p = population proportion of interest (here the proportion that has used cocaine). Let $\theta \neq 1/2$ be the proportion of Version A cards. Suppose that Y of the n respondents answered 'yes' to whichever question they were asked. One can use conditional probability arguments to show that an unbiased estimate of p, assuming simple random sampling and ignoring the finite population correction factor, is

$$\tilde{p} = (\theta - Y/n)/(2\theta - 1)$$

The variance of this estimator is

$$Var(\tilde{p}) = p(1-p)/n + \theta(1-\theta)/n(2\theta - 1)^2$$

which is larger than that obtained by asking the question directly, $p(1 - p)/n$, because we have introduced extra variability by asking a randomly chosen question. Notice that the variance will decrease as θ gets closer to 0 or 1, so that most of the cards are of one type (Version A or Version B). Thus, from a mathematical perspective, we prefer θ far from 1/2. Respondents, however, prefer θ close to 1/2, so that they feel confident that the interviewer does not know what question is being answered. The trade-off is an increase in variance in order to ensure that the respondent will provide a true, unbiased response.

In practice, randomized response techniques are not as successful as one would hope. The explanation of the technique may confuse the respondent, or the respondent may not really believe that the interviewer does not know what question is being answered. Thus, such techniques must be used with great care.

Nonresponse Error

Definition of Nonresponse Error

Almost every data collection endeavor results in missing data or nonresponse. In telephone or in-person surveys and censuses, for example, some sampled individuals cannot be located during the interview period, some are ill or unable to communicate with the interviewer, and some refuse to respond to the survey. The best that one can hope for when there is nonresponse is that the missing data are like a random sample of the units selected for the survey. If this is the case then the data are said to be missing completely at random and the effect on estimates is that they are more variable since they are based on a smaller sample size than planned. More often, however, those who do not respond to the survey differ from those who do respond on at least some survey variables of interest. In these cases the data are said to be missing at random if the missingness can be explained by other variables that are observed; nonresponse is nonignorable if the missingness depends on the value that is not observed (see Little and Rubin (1987) for a more complete discussion of types of missingness). For example, people who are not at home to respond to a telephone survey in the evening may watch less television than those who are at home. In this case, estimates of average hours of television viewing per week based only on data from respondents will be biased low. (Note that the above example also illustrates why it is not appropriate to replace nonresponding units in your sample with other units from the population: if there are differences between responders and nonresponders, choosing additional responders only repeats the bias on a larger scale.)

Nonresponse is a particular problem in longitudinal or multiple-wave surveys and studies where sampled individuals may drop out of the study after several interviews. If the reason for dropping out is related to the variables of interest, then nonresponse bias is introduced. For example, individuals in housing units sampled for the US National Crime Victimization Survey are interviewed seven times over 3.5 years. Individuals who move out of sampled housing units are not followed. If individuals who experience crime are more likely to move, perhaps to a safer neighborhood, then estimates of victimization rates will be biased low (see Lohr (1999) for a description of the National Crime Victimization Survey). A similar problem occurs in longitudinal medical studies if, for example, patients given the placebo see no change in their condition and thus are more likely to drop out of the study than patients receiving a treatment.

There are two types of nonresponse in surveys and censuses: unit nonresponse and item nonresponse. If no survey information is collected from a sampled unit, this is termed unit nonresponse; if a sampled unit responds to some, but not all, survey questions, this is termed item nonresponse. There are different methods for dealing with unit and item nonresponse.

Dealing with Nonresponse Error

The best way to deal with nonresponse error is to try to avoid it. Thus, good survey organizations use repeated callbacks on telephone and in-person surveys and reminder postcards or multiple mailings on mailed surveys to try to reach sampled individuals. Introductory letters or statements at the beginning of questionnaires should describe the purpose of the survey, provide information on who is funding and conducting the survey, stress the importance of the person's responses to the survey, and assure the respondent that all survey responses will be kept confidential. Some surveys provide a payment or gift to respondents or draw for prizes from among the respondents to encourage response.

In spite of the best efforts to encourage response, most surveys suffer from nonresponse. A researcher thus has the responsibility, at a minimum, to report the nonresponse rate for the survey along with any results from the survey. There

can be many definitions of the response rate for a survey, depending on whether one counts all units originally sampled or only those that led to eligible respondents or only those that resulted in some contact with a respondent, and so forth. See the American Association for Public Opinion Research (1998) for standard definitions of response rates.

The first step in dealing with nonresponse is to conduct exploratory data analysis on the survey data. For unit nonresponse, after determining the overall nonresponse rate one should then look at response rates by other variables that are available for both responding and nonresponding units. For example, in a mailed survey of a larger city, one could look at response rates by zip code. If there are big differences in response rates for various areas of the city, say that there are higher response rates from the more affluent areas, then survey estimates for variables that are related to affluence, such as support of the arts, are likely to be biased. In a survey of university alumni, one can compare respondents and nonrespondents on any of the variables available in the university alumni database such as GPA. If, for example, those with higher GPAs responded more often than those with lower GPAs, then survey estimates on variables related to GPA, say starting salary in the first job after graduation, are likely to be biased. If one does not find differences such as those suggested above, then at least one knows that there are not differences on the variables that could be explored. This does not, however, insure that there are no differences between respondents and nonrespondents on the survey variables of interest.

For item nonresponse, one should start by determining how much nonresponse there is for each variable with some missingness. For key study variables, one should look at summaries of other variables according to whether or not the key variable is missing. For example, if income is a key variable and subject to nonresponse, one could look at average age, average years in job, proportion married, proportion with high school diploma or less, and so forth for respondents who reported their income and those who did not. Obvious differences between the two groups on the other variables suggest that there are likely to be differences between the respondents and the nonrespondents on the income variable. It also suggests that the other variables might be useful in determining an adjustment for the nonrespondents, as described below.

Weight Adjustments

The sample-based weights assigned to cases in a sample survey can be interpreted as the number of people in the population of interest that the individual in the sample represents. The initial weight is usually the inverse of the selection probably for each sampled unit. One commonly used method to adjust for unit nonresponse in a sample is to increase the weights of respondents to account for nonrespondents. This adjustment is usually done within weight-adjustment cells within which it must be assumed that nonrespondents are 'just like' nonrespondents on variables of interest. That is, within weight-adjustment cells, the nonrespondents can be thought of as a random sample of the sampled units in that cell.

As a simple example, suppose you have a simple random sample of $n = 100$ individuals: 50 men, of whom 30 respond, and 50 women, of whom 40 reply. (Since we are assuming that a simple random sample of individuals was selected, all individuals in the sample will have equal selection probabilities and thus equal initial weights.) If you believe that within gender the nonrespondents are 'just like' the respondents (that is, the nonrespondents can be thought of as a random sample of those originally sampled) on survey variables of interest, then you may decide to use male/female as the weight-adjustment cells. The 30 responding males would represent themselves and the 20 nonresponding males. Thus, the original weights of the males must be multiplied by 50/30. Similarly, the original weights of the women would be multiplied by 50/40 to account for the nonresponding women.

In practice, there are usually more than the two nonresponse adjustment cells described in the simple example here. Little and Rubin (1987) describe methods for creating the nonresponse adjustment cells. Brick and Kalton (1996) provide details on weight adjustments for nonresponse.

Imputation Methods

Imputation methods are most often used to deal with item nonresponse. Imputation is defined as replacing a missing value by a plausible value. An advantage of imputation is that it creates a completed data set that can be analyzed using standard statistical methods. A disadvantage is that most subsequent analyses ignore the fact that imputation was done and, therefore, underestimate the variance by ignoring the uncertainty due to the imputation. There are many methods for creating plausible values to impute for a missing value. The better methods make use of information in the available data or in an auxiliary data source to estimate the missing values. The following describes some common methods of imputation. More details and additional methods may be found in Little and Rubin (1987) and Missing Data.

Mean imputation involves replacing a missing value of some variable by the sample mean computed from all individuals who reported a value for that variable. For example, mean income from those who reported income would be imputed for all those who refused to report income. The main disadvantages of mean imputation are that it distorts the distribution of the variable by creating a spike at the mean value, and it distorts relationships and correlations among variables. For example, the relationship between years experience on the job and income is not preserved when mean income is imputed for all those missing income regardless of years of experience. For this reason, mean imputation may be done within imputation cells that are determined by variables thought to be related to the missing values. Thus, years experience may be used to divide respondents into groups, and the mean income within each group is imputed for those missing income in the group. Mean imputation within group still created spikes in the data (at the mean for each group) and distorts the relationships among variables, but these effects should be less than with overall mean imputation.

Regression imputation involves using a regression model fit from complete cases to estimate missing values. For example, complete cases may be used to develop a model for income based on age, education, tenure in job, and other available data. This fitted model would then be used to estimate income when it is missing. Because this method may also create spikes

in the data when there are few possible combinations of the predictor variables used in the regression model, a variation involves adding a random error to the estimated value to be used for imputation.

Cold deck imputation involves using auxiliary data as a source for missing values. For example, data from administrative records might be used to obtain missing survey values, or, in a panel survey, values from a previous interview for the same respondent could be used to replace missing values. Hot deck imputation replaces missing values with values taken from other similar respondents' records in the same data set.

The imputation methods described above all result in underestimation of the variance because the imputed values are treated as 'truth' and subsequent analyses typically ignore the uncertainty in the imputations. Multiple imputation, first developed by Rubin (1987), allows one to account for this uncertainty. In multiple imputation, a stochastic imputation procedure is used to create, say, m imputations for each missing value. These m completed data sets are used in the standard analysis of the data. The variation among estimates from the m analyses is used to estimate the variance due to imputation. It is combined with the usual variance from the standard analyses to provide an overall variance of the estimates. For most larger surveys, m can be fairly modest: $m = 5$ is often adequate. For more details, *see* Multiple Imputation.

Other Sources of Error

There are other sources of nonsampling error that can affect survey data. Errors caused by interviewer bias, panel rotation, survey mode, coding, and data entry are all errors that should be avoided in conducting surveys. We will briefly consider panel rotation and survey mode errors.

Panel rotation error occurs when responses to a survey change depending on the length of time the panel has been included in the survey. Bailar (1975) described the effects of time-in-sample for the US Current Population Survey. In general, the longer a person has been in the sample, the less likely the person is to report being unemployed. A similar effect is found in the US National Crime Victimization Survey; persons in the survey for a longer time are less likely to report a victimization. Possible explanations for this phenomenon include that respondents learn to answer the survey so as to minimize the amount of time it takes to complete the survey or that differential dropout leads to a biased sample over time.

Survey modes include in-person, telephone, self-administered paper-and-pencil questionnaires, and, most recently, self-administered Web questionnaires. The same questions asked using different survey modes can elicit different responses (*see* Sample Surveys: Cognitive Aspects of Survey Design). For example, respondents are more concerned about their self-images in personal interviews than on paper-and-pencil surveys. Thus, they may be more likely to report personal problems on a self-administered questionnaire than in a face-to-face interview. On the other hand, security and confidentiality concerns may make respondents less likely to respond truthfully to surveys on the Web. The accuracy of responses to Web surveys is an area of ongoing research.

Additional information on these other sources of nonsampling error can be found in references such as Groves (1989) and Biemer et al. (1991). An excellent example of a survey for which all sources of error, both sampling and nonsampling, are explored in detail is the US Current Population Survey (see US Department of Commerce, 2000).

Bibliography

American Association for Public Opinion Research, 1998. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for RDD Telephone Surveys and In-person Household Surveys. AAPOR, Ann Arbor, MI.

Bailar, B.A., 1975. The effects of rotation group bias on estimates from panel surveys. Journal of the American Statistical Association 70, 23–30.

Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., Sudman, S., 1991. Measurement Errors in Surveys. Wiley, New York.

Bosecker, R.R., Ford, B.L., 1976. Multiple frame estimation with stratified overlap domain. Proceedings of the Social Statistics Section of the American Statistical Association, 219–224.

Brick, J.M., Kalton, G., 1996. Handling missing data in survey research. Statistical Methods in Medical Research 5, 215–238.

Cochran, W.G., 1977. Sampling Techniques, third ed. Wiley, New York.

Groves, R.M., 1989. Survey Errors and Survey Costs. Wiley, New York.

Kish, L., 1949. A procedure for objective respondent selection within the household. Journal of the American Statistical Association 44, 380–387.

Kish, L., 1965. Survey Sampling. Wiley, New York.

Krosnick, J.A., Alwin, D.F., 1987. An evaluation of a cognitive theory of response-order effects in survey measurement. Public Opinion Quarterly 51 (2), 201–219.

Lavrakas, P.J., 1993. Telephone Survey Methods, second ed. Sage, Newbury Park, CA.

Lessler, J.T., Kalsbeek, W.D., 1992. Nonsampling Error in Surveys. Wiley, New York.

Little, R.J.A., Rubin, D.B., 1987. Statistical Analysis with Missing Data. Wiley, New York.

Lohr, S.L., 1999. Sampling: Design and Analysis. Duxbury Press, Pacific Grove, CA.

Mosteller, F., 1978. Errors: nonsampling errors. In: Kruskal, W.H., Tanur, J.M. (Eds.), The International Encyclopedia of Statistics. Free Press, New York, pp. 208–229.

Nathan, G., 1988. A bibliography on randomized response: 1965–1987. Survey Methodology 14, 331–346.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

US Department of Commerce, Census Bureau, 2000. Current Population Survey: Design and Methodology. Government Printing Office, Washington, DC.

Warner, S.L., 1965. Randomized response: a survey technique for eliminating evasive answer bias. Journal of the American Statistical Association 60, 63–69.

Nonstandard Reasoning

Johan van Benthem, University of Amsterdam, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, nonstandard reasoning refers to the proliferation of reasoning styles investigated in modern logic beyond its traditional agenda. After a brief statement of standard logical approaches to consequence, we describe motivations for new systems. These include not only inference patterns with special vocabulary from mathematics, philosophy, and linguistics but also new styles of reasoning coming from computer science and artificial intelligence. The resulting landscape is diverse, but we discuss unifying themes such as structural rules, preferences, resources, information, and architecture of logical systems. Many of these reflect recent cognitive trends in modern logic putting 'social dynamics' at center stage: reasoning about one's own information and that of others, information update, acts of communication, processes of inquiry, and games.

The Emergence of Nonstandard Logics

Logic since antiquity has been concerned with 'the laws of thought,' supposedly unique and immutable principles of valid reasoning from premises to conclusions. Gradually, over the twentieth century, this monolithic view has changed to a pluralist one where logic tries to chart the different human reasoning styles and their matching logical systems. With the emergence of information, computation, and cognition as central scientific themes in recent decades, new items have kept appearing on this broader agenda. The resulting systems are often called 'nonstandard logics,' different from the established regime, but this label is relative. The background to current broader studies of reasoning are still the major approaches to valid inference in standard logic (semantic or proof-theoretic) developed by reflecting on philosophical argument and mathematical proof. Over time, however, these approaches have acquired richer vocabularies and additional notions of inference, whose motivation comes from philosophy, mathematics, linguistics, and many other disciplines. In particular, nonstandard logics often originate in computer science and artificial intelligence when thinking about key patterns in action, problem solving, or planning: arguably human core skills where reasoning may in fact have developed first in our evolutionary history. The resulting landscape of systems is diverse, but there are unifying themes such as structural rules of inference, information structure, preferences, resources, knowledge, or architectures for combining systems. These themes also reflect what might be called more realistic cognitive trends in modern logic: reasoning about one's own information and that of others, processes for information update and inquiry, communicative acts such as questions and answers, games and rational strategic interaction, and in all these, the use of heterogeneous information from symbolic, graphic, and other sources. In other words, the somewhat negative term 'nonstandard' will be understood in this article as natural agenda extensions of logic that go beyond the traditional core concerns.

Standard Logic: Valid Consequence, Proof Calculi, Completeness

The term 'standard logic' is mostly used to refer to the achievements of the renaissance of formal logic, starting with Boole, Frege, and Peirce in the nineteenth century, and culminating in the work of Gödel, Tarski, Turing, and others in the 1930s. The resulting core of the field consists of propositional logic (codifying reasoning about classifying situations using sentence combinations with the key operations 'not,' 'and,' and 'or') and predicate logic of quantifiers ('for all' and 'there exists') expressing statements about the internal structure of complex domains with objects and predicates. Together, these systems can describe a good deal of mathematical reasoning, and even quite a few core patterns in ordinary reasoning. Not surprisingly then, they have long been the standard core of most textbook presentations of modern logic. The two systems share what seems to be the most widely accepted semantic notion of valid consequence (due to Bolzano and Tarski) requiring 'transmission of truth' from premises P to a conclusion C:

Each situation in which all the premises P are true also makes the conclusion C true.

To describe and produce valid inferences, one can sometimes use simple mechanical methods, such as the truth tables of propositional logic. While truth tables are simple, and quite suggestive in assimilating reasoning to computation – a perennial theme in logic – they may promise more than can be delivered for logic in general. More general, and at the same time more relevant to capturing the undeniably stepwise structure of actual reasoning, are proof-theoretic approaches, deriving conclusions C from premises P via a system of rules:

Conclusion C follows from P if it can be derived by legitimate proof steps.

Simple steps staying within the reach of human understanding (not too easy, not too hard) are evident not only in mathematical proof but also in common sense reasoning. Proof systems of various sorts have turned out to have an additional benefit: they drive a wide variety of mechanized reasoning systems in artificial intelligence and computational logic.

In addition to system building, logic is also about system theory. Perhaps the first significant insight into logical systems

per se concerns the harmony between the preceding two perspectives. The semantic approach to valid consequence based on truth and the syntactic one based on proof coincide for quantificational logic: this is the content of Gödel's celebrated completeness theorem, which has spawned many similar results for other kinds of reasoning. Another key result is Gödel's even more famous 'incompleteness theorem' showing that formalized mathematical theories going beyond pure logic by adding arithmetical notions are nonaxiomatizable. One of the many consequences of the latter result is Church's theorem saying that no mechanical algorithm can decide validity in quantificational logic. Both positive and negative results have fueled progress in logic. For instance, one major source of variety in logical theories of inference is the search for decidable systems that escape Church's result by judiciously restricting the expressive power of the language of quantification.

Extending the Vocabulary: From Mathematics to Natural Language

Logical systems use 'logical forms' with fixed operators for key notions (Booleans, quantifiers), plus variable parts for expressions that can be interpreted freely (a typical valid form is 'from premises *A*-or-*B*, and not-*A*, infer conclusion *B*'). Propositional and predicate logic have been successful in analyzing mathematical proof, but gradually extensions had to be made to deal with the kind of argument found in philosophy, linguistics, and in the latter part of the twentieth century, computer science and cognitive science. For example, logical semantics of natural language deals with the much richer repertoire that we have for stating and communicating what the world is like according to us. This repertoire includes generalized quantifiers ('almost all,' 'most,' 'many,' and 'few') beyond the standard logical notions of 'all' and 'some,' plus a wide range of so-called modal expressions for our natural daily reasoning involving possibility, time, conditionality, obligation, knowledge, belief, action, etc. – often for both single and multiple agents. The extra vocabulary taken on board here may be considered 'nonstandard,' and determining the additional valid inferences supported by it may be subtle – but the canons of defining validity remain the earlier classical ones.

Alternative Views of Validity

There are also genuine deviations from the classical notion of valid consequence. A recurrent historical example is the view that conclusions only follow validly from consistent premises. Then, nothing follows from a contradiction *A*-and-not-*A*, while on the 'standard' account, everything does. However, this is just one of many hotspots. Alternative views of reasoning may also arise from studying the common vocabulary in human practice that often goes beyond the usual sentence connectives and quantifiers. Alternatives may also have conceptual motivations in philosophy and science.

Deviations in Natural Language, Cognition, and Philosophy

The idea that 'natural reasoning' in our ordinary language differs from standard logical calculi is a recurrent theme in the twentieth century. One hotspot for variations is the crucial logical notion of implication, represented by several expressions in natural language, with 'if…then…' as a ubiquitous example. It seems clear that natural implication does not behave like the truth-functional Boolean conditional – as shown by the 'paradoxes of material implication.' This has led to many deviant calculi not only for 'modal' or 'relevant' implication but also for counterfactual conditionals stating what would have been the case if certain things that are true now had not been the case. These are not just small variations. For instance, counterfactual reasoning keeps much of our thinking about what we do in the actual world 'in place,' and it represents a typical human skill. Some of these phenomena are also put into focus in cognitive psychology, where actual human reasoning behavior appears to be much richer than the repertoire that the standard logical canon has catered for. Finally, in addition to these empirical sources of diversity, there is also the influence of philosophical reflection on reasons, evidence, and knowledge or belief based on these. In contemporary epistemology, sophisticated new dynamic accounts of knowledge in terms of tracking the truth over time, or of constructing and maintaining spaces of relevant alternative worlds, suggest interesting new notions of reasoning that lack many of the basic features of classical systems.

Deviations in Scientific Reasoning

Other modulations on valid reasoning arise in the foundations of science. Starting in the heartland of mathematical logic, 'constructivism' in mathematics views proofs as constructions of mental entities, whose properties must be established by computational means. Accordingly, intuitionistic logic accepts only those inferences as valid that have a constructive proof. For instance, intuitionistically, one cannot prove an assertion 'there exists an x such that $F(x)$' without producing an explicit example, while classical logic would also accept an argument through the back door, namely, a refutation of the proposition that no object x has the property F. Intuitionistic logic is generalized in a wide family of constructive 'type theories' whose validities do not coincide with those of classical logical systems. However, logical proof theory continues to produce new perspectives on valid reasoning. A radical example from the 1980s is linear logic, which reinterprets the propositions involved in inference as information pieces or computational 'resources' that can be used only once. Alternative logics have also emerged in physics. Quantum mechanics has inspired quantum logics with failures of Boolean distributivity laws in reasoning about the location or momentum of particles. (Whether this has to be, or whether classical redescription is possible after all is one of the deep issues in the field.) Most of all, reflecting the close analogies between 'reasoning and reckoning' (as Hobbs put it in the seventeenth century when these views emerged), computer science has inspired many new systems of reasoning. Examples are dynamic logics describing the effects of computer programs and of actions generally, logic programming (using proof systems as computation procedures), description logics (well-behaved weak systems for data bases and knowledge representation), and a host of new systems discovered in artificial intelligence when engaging in the study of 'commonsense reasoning.'

The result of all these developments for logic today is a wide range of alternatives to classical systems, developed with equal rigor, but different in their valid principles, general philosophy, and mathematical properties.

Varieties of Inference

That 'natural reasoning' comes in different styles, with possibly different logical features, was already observed by Bolzano in the early nineteenth century. Scientific reasoning employs stricter standards than common sense (although Bolzano felt that philosophical reasoning was the most demanding skill of all), and common sense as encoded in natural language itself has many varieties, that may even be task dependent. The same variety also occurs in artificial intelligence with problem solving or planning (deductive or heuristic), where systems are more 'optimistic' than standard logic, representing a typical human ability of being able to jump to conclusions, a rich inferential power kept in check by our talent for self-correction once our beliefs turn out wrong. Indeed, more generally, reasoning cannot be seen in isolation from other cognitive skills. In the following subsections we give a few examples.

Nonmonotonic Logics

Practical reasoning uses defaults going beyond standard logical conclusions. Thus, we use some default rule ('Dutch trains run on time') in planning our escape home from a futuristic campus unless counterexamples are explicitly indicated, or, in judging our friends in a cold climate, we take an answer 'John, Mary' to the question 'Who were skating?' as conveying not just that John and Mary skated, but that no others did, assuming a maximally informative respondent. Crucially, such additional inferences may have to be withdrawn later, because new relevant information comes in: my train may be carrying soccer fans, the person answering my question may turn out to be a government employee. Reasoning in which a conclusion C may follow from premises P, but no longer from the richer set $P + Q$ is called 'nonmonotonic' – while classical consequence is monotonic: once drawn, conclusions persist. A general pattern behind much nonmonotonic reasoning is 'circumscription,' which, in its most abstract form, assumes that situations come ordered by some binary relation of preference, and then makes C follow from P if C holds in all most preferred situations where P holds. In the escapist example, the most preferred situations (for the time being) are those where trains run according to schedule, while those containing a disturbing factor are less preferred. This idea comes from artificial intelligence and philosophical logic, and it has a very wide sweep. Preferences may reflect genuine preference, but just as well notions of plausibility, utility, relevance, or simplicity, depending on circumstances.

Classical consequence is not in conflict with this view of reasoning. It is still present here as the special case when no models are preferred to others, making all models for the premises most preferred. Still, there is an intuitive difference, which comes to light in a richer picture of what is happening when these different reasoning styles are deployed. Classical reasoning is safe, and thus, it may be associated with knowledge, and ways of unpacking it. By contrast, nonmonotonic reasoning goes naturally with processes of forming and modifying beliefs, a more volatile attitude than knowledge, but one that is essential to practice. The tandem of inference and belief revision is a crucial practical skill, reflecting two aspects of what it means to 'learn.'

Directions and Purposes in Reasoning

Our presentation so far viewed reasoning as a forward-oriented process of accumulating conclusions from given premises. In addition to this forward mode, there are others. This is already clear with standard uses of classical reasoning for refuting, rather than confirming, opinions or even whole theories. Directions multiply when we consider the rich view of reasoning developed by Peirce around 1900, who distinguished deduction, induction, and abduction. Abduction is the common explanatory reasoning that we engage in when seeking plausible premises to support given observations, or in making sense of what we are told. It provides premises for conclusions, rather than the other way around. In contrast with this, induction looks for plausible generalizations beyond the immediate content of the premises, raising again the issue of learning scenarios, now for general facts going beyond specific observations. These different directions and purposes all matter when analyzing activities such as 'refutation,' 'explanation,' or 'confirmation' in the philosophy of science – but it is also easy to find counterparts for all of them in our daily practice. Moreover, like standard logic, these broader views of reasoning have found their way into computer science. Abduction has been studied in connection with logic programming; varieties of belief revision and induction occur in computational learning theories. In recent years, many of these systems have also been related to cognitive science. For instance, cognitive experiments with conditional reasoning point at nonmonotonic phenomena in actual behavior, and it has even been shown that neural networks in the brain can be represented in terms of nonmonotonic logics with inferences repeated over time.

Reasoning with Uncertainty and Probability

Induction leads naturally to the foundations of uncertainty and probability. Probabilistic reasoning seems a major mode of cognition in its own right. Its standard formalism is definitely not logic, but probability theory. However, starting from Carnap's 'inductive logic' in the 1950s, there is an increasing literature on interfaces between logical and probabilistic reasoning. These include more qualitative styles of probabilistic reasoning in 'fuzzy logic' and 'possibility theory,' and natural hybrids of standard logical and probabilistic systems. Moreover, statistical long-term bulk behavior is coming to the fore as a topic of study, and here 'inference statistics' in terms of percentages of truth and falsity across models, or records of proof attempts, meets logic proper. Finally, qualitative versions of fundamental probabilistic notions such as randomness and independence have been migrating into logic. In particular, new versions of standard predicate logic have attracted attention recently,

allowing for patterns of dependence and independence between variables that go far beyond what has been studied in existing logical systems.

Logical Dynamics: Reasoning as an Activity

Under the influence of ideas from philosophy, computer science, and cognitive science, many logicians have started viewing reasoning as a cognitive process whose dynamics must, and can, be brought out by logical means. This process view of logic is closer to actual argumentation, which is driven by local acts of inferential commitment, while longer term timing and procedural conventions are also crucial to the outcomes.

Dynamic Logics of Information Flow

In this dynamic perspective, reasoning is one of many cognitive tasks that can be studied together. As an early signal of this trend, already the Mohist logicians in ancient China thought that inference, observation, and communication are the three great natural sources of information that should work together. This interplay of different information sources can be observed in the empirical sciences, but just as well in daily practice. Modern logical systems merging inference with a broader information dynamics include dynamic epistemic logics that combine acts of inference, observation, and questions, and axiomatize the laws that describe which new information is generated in each step. The relevant information states can be those of individual agents, but given the social interactive character of much reasoning and related informational activities, these logics often work over multiagent models for different agents having different information.

A further source of diversity here, beyond the dynamic repertoire of actions and events, is the notion of information at work. Inquiry is tied to the usual semantic notion of information as a range of current options that shrinks when new information comes in. However, other powerful intuitions exist, such as the idea that information flows because of correlations or dependencies between different objects or situations that are brought together in reasoning. There is even a third notion. Valid inferences do not shrink a semantic range of options: they rather elucidate a more fine-grained syntactic notion of information, on whose nature there is no consensus at present. The precise sense of information in logic has itself become a topic of discussion and a source of variety.

Dynamic Semantics of Natural Language

Another take on information dynamics is in terms of the meaning of the very expressions used in reasoning. A modern program is dynamic semantics, a view of the linguistic meaning of a proposition as the update of a hearer's information state that is produced by it. Thus, propositions become cognitive procedures, and natural language now resembles a programming language modifying information states of agents.

Dynamic semantics does not stop at reinterpreting vocabulary. It has also been a source of new notions of inference and consequence. One natural dynamic notion of valid consequence from premises P to a conclusion C is the following 'fixed-point' view:

> *In final information states reached by processing the successive premises P, updating with the conclusion C will not have any further effect.*

Dynamic semantics exist for a broad variety of speech acts associated with reasoning and information flow, including questions, answers, suggestions, or commands. Again, multi-agent perspectives come to the fore, as speech acts make little sense in solitude.

Belief Revision, Learning, and Social Information Dynamics

A broader perspective on such cognitive actions arises in belief revision theory, which studies updates adding information to the current state, contractions that remove information, and revisions doing both – doing justice to the 'zigzag character' of much of our reasoning. We have already noted how acts of revision and self-correction form a natural complement to the somewhat risky nature of nonmonotonic reasoning. Forming hypotheses and learning from errors is also crucial to learning theory, which may be considered an extension of information dynamics to a longer term process horizon.

Belief revision and learning may also come in social flavors, since one of the most common causes for belief change is something said by someone else. However, groups of agents also bring their own dynamics, such as the phenomenon of 'belief merge' between different groups. Such phenomena have been studied for computational agents, and recently also in social philosophy and social choice theory, since votes are only stages in some larger setting of communication and deliberation, i.e., social reasoning.

Logic and Games

Perhaps the most 'activist' perspective on reasoning is one borrowing some major concepts from game theory. Many basic logical notions can be cast in the form of games. Argumentation is a key example, but the same is true for evaluating assertions between parties disagreeing on the truth value of a given statement, and many other logical tasks. By now, 'logic games' are a well-known technique, even in textbooks, for teaching logical basics. In particular, logical 'dialog games' suggest alternative views of valid reasoning in terms of winning and losing argumentation games, complementing the earlier views in terms of proof and semantics. The dialogical take on valid consequence employs the fundamental game-theoretic notion of a strategy, a response pattern for a player to any sequence of moves by one's opponents:

> *In fair rational debate, the proponent of the conclusion C has a winning strategy for upholding C against any play by an opponent granting the premises P.*

Logic games tie the study of reasoning to that of rational behavior in general. This is one of several places where logic interfaces with game theory, finding shared concerns in the

foundations of social action. The logic and games connection has led to new links of logic not only with philosophy but also with practical fields like argumentation theory.

Systematic Theory of Reasoning

With the rapid growth of nonstandard logics, the unity of classical logic may seem to dissolve. Even so, at a theoretical metalevel, unity may still prevail, be it in new more general guises – rather than the chanting of particular canonical laws of consequence. The usual mathematical notions and techniques of logic still work in today's broader realm, in charting possible systems, and exploring general trends. It is too early for a new synthesis of 'standard' and 'nonstandard,' but some unifying themes are noticeable.

Structural Rules

The above-discussed notion of nonmonotonicity is a hallmark of many nonstandard reasoning systems – although it is a symptom, not a diagnosis of underlying causes. However, there is a natural abstraction level here. In recent decades, it has turned out useful to classify styles of reasoning by their 'structural rules,' such as monotonicity and transitivity ('if P implies Q, and Q implies R, then P implies R'), and others, which move premises around, or duplicate them when needed. The structural rules of classical logic treat premises as an unordered set, whose presentation is irrelevant. By contrast, nonstandard reasoning tends to be more sensitive to the formulation of data, and the corresponding actions, which shows in failures of classical structural rules. Thus, in dynamic semantics, permuting two premises may change the update achieved – reflecting the typical order dependence of natural language. By now complete sets of structural rules have been found for many nonstandard reasoning practices, highlighting what makes them special. This variety has now also reached the philosophy of logic, where 'logical pluralism' is the position that this diversity is legitimate.

New Vocabulary

Reasoning and language are intimately connected. When an inferential practice changes, there may be a concomitant change in how we view the language where it all takes place. This language change can happen in two broad varieties. One is extension of classical languages. For instance, current dynamic logics of knowledge update or belief revision tend to add new logical operators of 'learning' connected to information changes, and then study their laws as an extension of existing logical systems. However, a more radical line is also noticeable in the current literature. A nonstandard notion of reasoning may lead to reinterpretation of the existing traditional logical operations. Thus, in logic games, a disjunction is a 'choice' of a move by the defending player, while a negation signals a 'role switch' between two players. Nonstandard reasoning may also involve new operations, outside the classical core. In particular, with games, there is not one conjunction, as in Boolean logic, but three: choosing which game to play at the start, playing one game after another, or playing two games concurrently. All these support logical laws for 'conjunction', that are related, but also part company in crucial aspects. Similar proliferation phenomena can be observed with the earlier mentioned resource-oriented proof-theoretic logics, from intuitionistic to linear systems. Thus modern calculi of reasoning can be much richer in vocabulary than classical logical systems: broader reasoning means richer language.

Again, theory is lagging behind agenda expansion here. The two strategies language extension and language reinterpretation seem to be connected in various ways. However, so far, we have only bits and pieces of translation results, and no general picture.

Mechanisms and Structures

More important than individual logical laws, deviant or not, are general mechanisms behind nonstandard reasoning systems. Striking unifying notions mentioned before included ordering relevant situations by various kinds of preference relations, fine-grained computational views of handling inferential resources, and natural repertoires of dynamic informational actions and processes. These notions provide a much richer picture of actual reasoning, and 'parameters' that can be set for different purposes.

Other important general trends go beyond the single-sentence habitat of standard logic. One of these is the temporal longer term. In addition to the usual local semantic 'meaning' of single inferences, acts of observation, or questions, there is the art of 'making sense' of these events. This art requires the setting of some longer term informational process, say a conversation, a process of inquiry, or a learning scenario. This was clear, for instance, with the game-theoretic view of validity that was discussed earlier: a strategy is a recipe for long-term behavior, guaranteeing certain outcomes in the long run. These temporal strategic aspects of reasoning are slowly coming to the fore in the modern literature, in the form of new research lines in between logic, process theories in computer science, and game theory.

Another large-scale aspect of reasoning that is ignored in a simple-minded premise-conclusion format is architecture of representation and processing. How are the data that reasoning has to work with usefully structured? How do different styles of reasoning cooperate to achieve useful effects? How do symbolic and nonsymbolic data (say visual, or just observing movement) cooperate in reasoning? One counterpart to the temporal long term here is spatial bulk. Reasoning takes place on the basis of experience, i.e., a vast reservoir of earlier reasoning tasks stored in some way in our memory (perhaps compressed into some sort of retrievable probabilistic version). How can we describe this, and what might be emergent statistical phenomena here going beyond our traditional idea of a logical system? Again, the earlier-mentioned interface between logic and probability becomes relevant here.

Finally, a pervasive theme in a broader understanding of reasoning is its social nature. Much reasoning is a many-mind activity involving agents that communicate, cogitate, and act – and the repercussions of this view of 'intelligent interaction' are slowly unfolding in current research on multiagent logics, computation, and games.

A Triangle of Disciplines

One way of looking at the nonstandard reasoning themes presented here is as a shift in character of logical theory, from purely normative to more descriptive. Now this may not be quite the right perspective: nonmonotonic logics can well be normative, and so can dynamic logics of observation, telling us what we should get out of what we see. However, what is true is that a natural triangle of disciplines is involved in most of the themes in this article. Some tend more toward theory (logic, philosophy, and mathematics), others are more driven by empirical data (linguistics or cognitive science), and interestingly, there was also a crucial third category of disciplines that are not easily classified as either normative or descriptive, such as computer science. Drawing sharp distinctions between the three kinds of discipline seems increasingly irrelevant in the study of reasoning.

Coda. Despite all this legitimate interdisciplinary variety at the level of cognitive activities, there may still be a privileged role for standard logic at another level. Significantly, in the metalevel system building about all these activities, logicians tend to follow classical methods from the field, since these have set standards of rigor and perspicuity that no one wants to give up. In this sense, there is much more continuity in the field than might appear from this article. The standard *modus operandi* is there to serve all.

Conclusion

Modern logic arose out of reflection on proof in mathematics and argumentation in philosophy, and created the standard systems of the twenty-first century. Modern studies of reasoning, while retaining the methods of this phase, deal with a much broader spectrum of laws, structures, and processes, than the repertoire of a standard logic textbook. Moreover, in doing so, they interface in new ways with disciplines such as philosophy, linguistics, computer science, cognitive science, game theory, argumentation theory, legal studies, and others. In this light, 'nonstandard reasoning' signals nothing marginal or subversive, except for insisting on this broader horizon.

See also: Axiomatic Theories; Deductive Reasoning Systems; Game Theory; Knowledge Representation; Logics for Knowledge Representation; Mathematical Models in Philosophy of Science; Scientific Discovery, Computational Models of.

Further Reading

Abramsky, S., Gabbay, D., Maibaum, T. (Eds.), 1992–2000. Handbook of Logic in Theoretical Computer Science. Oxford University Press, Oxford, UK.

Adriaans, P., van Benthem, J. (Eds.), 2008. Handbook of the Philosophy of Information. Elsevier, Amsterdam, The Netherlands.

Barwise, J. (Ed.), 1977. Handbook of Mathematical Logic. North-Holland, Amsterdam.

Gabbay, D., Guenthner, F. (Eds.), 1983–1988. Handbook of Philosophical Logic. Reidel, Dordrecht, The Netherlands; Revised and expanded, 2001. Kluwer, Dordrecht, The Netherlands.

Gabbay, D., Hogger, C., Robinson, J. (Eds.), 1993–1998. Handbook of Logic in Artificial Intelligence and Logic Programming. Oxford University Press, Oxford, UK.

Gabbay, D., Woods, J. (Eds.), 2002. Handbook of the Logic of Argument and Inference: The Turn towards the Practical. Elsevier, Amsterdam, The Netherlands.

Elsevier, Amsterdam; Revised and expanded edition 2010 Van Benthem, J., ter Meulen, A. (Eds.), Handbook of Logic and Language. Elsevier Insights, Amsterdam, The Netherlands.

Van Benthem, J., Gupta, A. (Eds.), 2011. Logic and Philosophy Today. College Publications, London.

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 17

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser–Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 17

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 17 *xlv*

VOLUME 17

Normality and Mental Health 1
Femi Oyebode

Norms 5
Karl-Dieter Opp

Norms in Science 11
Peter Weingart

North America, Archaeology of 15
Asa R Randall

North America: Sociocultural Aspects 22
Jessica Cattelino

Nosology in Psychiatry 27
Femi Oyebode

Nostalgia, Anthropology of 34
Millie Creighton

Notestein, Frank Wallace (1902–83) 39
John C Caldwell

Number, Grammatical 43
Guido G Corbett

Numbers and Number Sense 46
Ian M Lyons

Number Systems, Evolution of 57
Peter Damerow

Obedience (Research), History of 61
Stephen D Reicher and Alexander Haslam

Obedience: Social Psychological Perspectives 67
Thomas Blass

Obesity and Labor Market Performance 70
Susan L Averett

Obesity and Mortality 75
Neil K Mehta

Obesity: Cultural and Biological Factors 82
Alexandra A Brewis

Objectivity, History of 88
Stephen Gaukroger

Objectivity of Research: Ethical Aspects 92
David B Resnik

Objectivity: Philosophical Aspects 96
Paolo Savoia

Observational Studies: Outcome-Based Sampling 103
Daniel McFadden

Observational Studies: Overview 107
Paul R Rosenbaum

Obsessive-Compulsive Disorder 113
João M Castaldelli-Maia

Obsessive-Compulsive Disorder across the Life Span 119
Norbert Kathmann

Occipital Lobe (Visual Cortex): Functional Aspects 127
Oliver Braddick

Occupational Health 133
Töres Theorell

Occupational Health and Safety, Regulation of 141
Richard Johnstone

Occupational Stress Management 147
Peter Angerer and Andreas Müller

Oceania, Sociocultural Overviews: Melanesia 153
Robert J Foster

Oceania, Sociocultural Overviews: Polynesia and Micronesia 158
Niko Besnier

Oil Age: Energy in Transition 165
Colin J Campbell

Old Age and Centenarians: The Human 'Warranty Period' 172
Bruce A Carnes

Old Age: Definitions, Theory, and History of the Concept 178
Frédéric Balard

Old Age Work Participation 182
Arthur van Soest and Asghar Zaidi

Old and New Institutionalism in Economics 190
David Dequech

Olfactory System 196
Heather M Schellinck and Richard E Brown

Oligarchy, Iron Law of 201
Darcy K Leach

Olson, Mancur (1932–98) 207
Martin C McGuire

Online Dating 212
Monica T Whitty

Online Pornography 217
Susanna Paasonen

Onomastics 223
Wilhelm FH Nicolaisen

Open Education: Learning and Teaching 227
Eugene F Thibadeau

Openness to Experience 231
Sang E Woo, Rachel Saef, and Scott Parrigon

Opiate Use and Abuse, History of 236
Gene M Heyman

Opinion Formation, Theory of 243
Shawn W Rosenberg

Oppression 246
Katherine van Wormer

Optimal Control Theory 253
Ngo Van Long

Optimal Experimental Design 256
Anthony C Atkinson

Optimism and Pessimism 263
Charles S Carver

Optogenetics 268
Dominic Mudiayi, Scott Wong, and Aaron Gruber

Oral and Literate Expression 274
Ruth Finnegan

Oral Health, Psychology of 279
Renate Deinzer

Oral History 284
Andrea Hajek and Angela Davis

Order Statistics 291
HA David

Ordered Relational Structures 296
Reinhard Suck

Organization, Geography of 301
Martin Müller

Organization: Overview 307
W Richard Scott and Gerald F Davis

Organizational Behavior, Psychology of 313
Bernhard Wilpert

Organizational Citizenship Behavior 317
Dennis W Organ

Organizational Climate 322
Michael A West and Joanne Lyubovnikova

Organizational Climate in the Work Setting 327
Mark G Ehrhart and Maribeth Kuenzi

Organizational Commitment 334
Howard J Klein and Heeman Park

Organizational Control 341
Paul Thompson

Organizational Culture, Anthropology of 346
Jakob Krause-Jensen and Susan Wright

Organizational Decision Making 352
Anna Grandori

Organizational Ecology 358
Glenn R Carroll and Michael T Hannan

Organizational Emergence and Firm Formation 364
Elizabeth Garnsey, Sarah Lubik, and Paul Heffernan

Organizational Errors: Catastrophic 370
Karlene H Roberts

Organizational Fields 375
Melissa E Wooten

Organizational Justice 379
Russell Cropanzano and Agustin Molina

Organizational Path Dependence 385
Jörg Sydow and Georg Schreyögg

Organizations and Culture 390
Frank Dobbin and Shawna Vican

Organizations and the Law 397
Mark C Suchman

Organizations: Authority and Power 405
Cynthia Hardy

Organizations: Negotiated Orders 411
Tony J Watson

Organizations, Sociology of 415
Erhard Friedberg

Organizations: Unintended Consequences 420
Henk Flap

Orientalism 423
Daniel M Varisco

Orientalism, Oppression, and Contemporary Social Work 428
Elizabeth Beck and Stéphanie Wahab

Orthodox Churches 435
Sergei Hackel

Oversharing: The Eclipse of Privacy in the Internet Age 439
Ben Agger

Pacific Islands, Archaeology of 446
Patrick V Kirch

Pain, Health Psychology of 451
Herta Flor

Paleoethnobotany 456
Deborah M Pearsall

Panel Surveys: Uses and Applications 462
Greg J Duncan

Panic Disorder 468
Heather K Hood and Martin M Antony

Panic Disorder and Agoraphobia Across the Lifespan 474
Katharina Meyerbröker and Mark B Powers

Paralawyers: Other Legal Occupations 482
Herbert M Kritzer

Parallel Relationships 486
Catherine Hakim

Parenting Attitudes and Beliefs across Cultures 493
Çigdem Kagitcibasi and Bilge Yagmurlu

Parents and Teachers as Partners in Education 499
Janet Goodall

Pareto, Vilfredo (1848–1923) 503
Gottfried Eisermann

Parietal Lobe 506
Claudia LR Gonzalez and Jason W Flindall

Parliamentary Government 511
Klaus von Beyme

Parliaments, History of 515
Estevao C de Rezende Martins

Parsons, Talcott (1902–79) 520
David Sciulli

Partial Orders 525
Reinhard Suck

Participant Observation 530
Gary Alan Fine

Participatory Action Research in Environmental and Ecological Studies 535
Frances Fahy

Participatory Action Research in Social Research 540
María Elena Torre, Caitlin Cahill, and Madeline Fox

Partnership Formation and Dissolution in Western Societies 545
Brienna Perelli-Harris

Parts of Speech 553
Walter Bisang

Party Identification and Nonpartisanship 562
Russell J Dalton

Party Systems 567
Peter Mair

Passion and Motivation 570
Julia Moeller, Jacquelynne S Eccles, Katariina Salmela-Aro, Julia Dietrich, Barbara Schneider, and Robert Grassinger

Pastoralism in Anthropology 577
John G Galaty

Patenting in the Biological Sciences 584
Myles W Jackson

Pathological Gambling and Gambling Disorder 592
David M Ledgerwood and Nancy M Petry

Patient Adherence to Health Care Regimens 598
Tricia A Miller and M Robin DiMatteo

Patron–Client Relations, Social and Anthropological Study of 603
Luis Roniger

Pavlov, Ivan Petrovich (1849–1936) 607
Daniel A Powell

Pay, Compensation, and Performance, Psychology of 611
Joseph J Martocchio

Peace 618
Johan Galtung

Peace and Nonviolence: Anthropological Aspects 624
Leslie E Sponsel

Peace Keeping 631
Franz Kernic

Peace-Making in History 637
Peter J Lineham

Peace Movements 643
Marco Giugni

Peace Processes 648
John D Brewer

Peace Promotion, Psychology of 654
Michael G Wessells

Pearson, Karl (1857–1936) 660
John Aldrich

Peasantry in the Twenty-First Century 664
Jan Douwe van der Ploeg

Peasants and Rural Societies in History (Agricultural History) 669
Janken Myrdal

Peasants in Anthropology 676
Enrique Mayer

Peer Review and Quality Control 680
Silvio O Funtowicz and Jerome R Ravetz

Peer Review: Organized Skepticism 685
Katri Huutoniemi

Pension Policy 690
Karen M Anderson

People in Organizations 697
Jennifer A Chatman and Jack A Goncalo

Perception of Action Space: Using Multiple Frames of Reference 703
Carl Gabbard

Performance: Anthropological Aspects 708
Felicia Hughes-Freeland

Performance Appraisal and Evaluation 716
Dina Van Dijk and Michal M Schodl

Perinatal Psychiatry 722
Keiko Yoshida, Yumi Nishikii, and Hiroshi Yamashita

Period and Cohort Analysis in Demography 727
Yang C Yang and Ryan K Masters

Person and Self: Philosophical Aspects 735
Alessandro Ferrara

Person Perception, Accuracy of 739
David C Funder

Personal Identity: Philosophical Aspects 742
Gertrud Nunner-Winkler

Personal Intelligence and Competencies 750
John D Mayer, David R Caruso, and AT Panter

Personal Projects 757
Brian R Little and Simon Coulombe

Personality and Adaptive Behaviors 766
Gian V Caprara

Personality and Creativity 770
Jacob B Hirsh

Personality and Crime 774
David P Farrington and Darrick Jolliffe

Personality and Economics 780
Daniel D Santos

Personality and Educational Outcomes 787
Arthur E Poropat

Personality and Health 792
Christian Hakulinen and Liisa Keltikangas-Järvinen

Personality and Marriage 797
Benjamin R Karney

Personality and Religion 801
Vassilis Saroglou

Personality and Risk Taking 809
Bernd Figner and Elke U Weber

Personality and Self 814
Michael D Robinson and Clay Routledge

Personality and Values at Work 821
Ryan Klinger and Mark Mallon

Personality Assessment 827
Johannes Hewig, Uta Kraus, Patrick Mussel, Roman Osinsky, and Marko Paelecke

Personality Assessment, Faking and 834
Matthias Ziegler and Doreen Bensch

Personality Assessment, Forced-Choice 840
Anna Brown

Personality Assessment: Overview 849
Adrian Furnham

Personality, Biological Models of 857
Philip J Corr

Personality Changes During Adolescence Across Cultures 863
Boris Mlačić and Goran Milas

Personality, Cognitive Models of 870
Gerald Matthews

Personality Development and Temperament 876
Yiyuan Xu and Alexander W Krieg

Personality Development: Systems Theories 879
Katarína Millová and Marek Blatný

Personality Differences and Development: Genetic and Environmental Contributions 884
Christian Kandler and Wiebke Bleidorn

Personality Disorders 891
Peter Tyrer

Personality, Evolutionary Models of 899
Hannah J Haysom, Karin JH Verweij, and Brendan P Zietsch

Personality: Historical and Conceptual Perspectives 906
Frank Dumont

Personality Theory and Psychopathology 914
Kate E Walton and Stephanie R Pavlos

Personality, Trait Models of 920
Peter Borkenau and Nancy Tandler

Personalized Medicine 925
Robert A Phillips and Claus Vögele

Personhood, Anthropology of 932
Nigel Rapport

Personnel Selection, Psychology of 936
Dan Ispas and Walter C Borman

Persuasion Theories 941
Antonis Gardikiotis and William D Crano

Perversions, Sexual (Paraphilias) 948
Oyedeji Ayonrinde

Pestalozzi, Johann Heinrich (1746–1827) 955
Petra Korte

This page intentionally left blank

CONTRIBUTORS TO VOLUME 17

Ben Agger
University of Texas-Arlington, Arlington, TX, USA

John Aldrich
University of Southampton, Southampton, UK

Karen M. Anderson
Radboud University Nijmegen, Nijmegen, The Netherlands

Peter Angerer
Heinrich Heine Universität Düsseldorf, Düsseldorf, Germany

Martin M. Antony
Ryerson University, Toronto, ON, Canada

Anthony C. Atkinson
Department of Statistics, London School of Economics, London, UK

Susan L. Averett
Lafayette College, Easton, PA, USA

Oyedeji Ayonrinde
Bethlem Royal Hospital, South London and Maudsley NHS Foundation Trust, Beckenham, UK

Frédéric Balard
Université Lille 3, Villeneuve-d'Ascq, France

Elizabeth Beck
School of Social Work, Georgia State University, Atlanta, GA, USA

Doreen Bensch
Institut für Psychologie, Humboldt-Universität zu Berlin, Berlin, Germany

Niko Besnier
Universiteit van Amsterdam, Amsterdam, The Netherlands

Klaus von Beyme
University of Heidelberg, Heidelberg, Germany

Walter Bisang
Johannes Gutenberg-Universität Mainz, Mainz, Germany

Thomas Blass
University of Maryland, Baltimore, MD, USA

Marek Blatný
Institute of Psychology of Academy of Sciences of the Czech Republic, Brno, Czech Republic

Wiebke Bleidorn
Tilburg University, Tilburg, The Netherlands

Peter Borkenau
Martin-Luther-Universität Halle-Wittenberg, Halle, Germany

Walter C. Borman
Personnel Decisions Research Institutes, Inc. (PDRI), Tampa, FL, USA; and University of South Florida, Tampa, FL, USA

Oliver Braddick
University of Oxford, Oxford, UK

John D. Brewer
Institute for the Study of Conflict Transformation and Social Justice, Queen's University Belfast, Belfast, UK

Alexandra A. Brewis
Arizona State University, Tempe, AZ, USA

Anna Brown
School of Psychology, University of Kent, Canterbury, Kent, UK

Richard E. Brown
Dalhousie University, Halifax, NS, Canada

Caitlin Cahill
Pratt Institute, Brooklyn, NY, USA

John C. Caldwell
Australian National University, Canberra, ACT, Australia

Colin J. Campbell
Cork, Ireland

Gian V. Caprara
La Sapienza University, Rome, Italy

Bruce A. Carnes
College of Medicine, The University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

Glenn R. Carroll
Stanford University, Stanford, CA, USA

David R. Caruso
Yale University, New Haven, CT, USA

Charles S. Carver
University of Miami, Coral Gables, FL, USA

João M. Castaldelli-Maia
Medical School, Fundação do ABC, Santo André, São Paulo, Brazil; and Medical School, University of São Paulo, São Paulo, Brazil

Jessica Cattelino
University of California, Los Angeles, CA, USA

Jennifer A. Chatman
University of California, Berkeley, CA, USA

Guido G. Corbett
University of Surrey, Guildford, UK

Philip J. Corr
City University London, London, UK

Simon Coulombe
Université du Québec à Montréal, Montréal, QC, Canada

William D. Crano
Claremont Graduate University, Claremont, CA, USA

Millie Creighton
University of British Columbia, Vancouver, BC, Canada

Russell Cropanzano
Leeds School of Business, Boulder, CO, USA

Russell J. Dalton
Center for the Study of Democracy, University of California Irvine, Irvine, CA, USA

Peter Damerow[†]
Max Planck Institute for the History of Science, Berlin, Germany

H.A. David
Iowa State University, Ames, IA, USA

Angela Davis
University of Warwick, Coventry, UK

Gerald F. Davis
The University of Michigan, Ann Arbor, MI, USA

Renate Deinzer
Institute of Medical Psychology, Justus-Liebig-University Giessen, Gießen, Germany

David Dequech
Institute of Economics, University of Campinas, São Paulo, Brazil

Estevao C. de Rezende Martins
University of Brasilia, Brasilia, Brazil

Julia Dietrich
University of Jena, Jena, Germany

M. Robin DiMatteo
University of California Riverside, Riverside, CA, USA

Frank Dobbin
Harvard University, Cambridge, MA, USA

Frank Dumont
McGill University Montreal, Hudson, QC, Canada

Greg J. Duncan
Northwestern University, Evanston, IL, USA

Jacquelynne S. Eccles
University of California Irvine, Irvine, CA, USA

Mark G. Ehrhart
San Diego State University, San Diego, CA, USA

Gottfried Eisermann
Bonn, Germany

Frances Fahy
School of Geography and Archeology, National University of Ireland Galway, Galway, Ireland

David P. Farrington
Institute of Criminology, Cambridge, UK

Alessandro Ferrara
Università di Roma Tor Vergata, Roma, Italy

Bernd Figner
Behavioural Science Institute, Radboud University Nijmegen, Nijmegen, The Netherlands; and Center for the Decision Sciences, Columbia University, New York, NY, USA

Gary Alan Fine
Northwestern University, Chicago, IL, USA

Ruth Finnegan
Open University, Milton Keynes, UK

Henk Flap
Universiteit Utrecht, Utrecht, The Netherlands

Jason W. Flindall
University of Lethbridge, Lethbridge, AB, Canada

[†] Deceased.

Herta Flor
Medical Faculty Mannheim, Central Institute of Mental Health, Heidelberg University, Mannheim, Germany

Robert J. Foster
University of Rochester, New York, NY, USA

Madeline Fox
Brooklyn College, New York, NY, USA

Erhard Friedberg
Paris, France

David C. Funder
University of California Riverside, Riverside, CA, USA

Silvio O. Funtowicz
Centre for the Study of the Sciences and the Humanities (SVT), University of Bergen, Bergen, Norway

Adrian Furnham
University College London, London, UK

Carl Gabbard
Texas A&M University, College Station, TX, USA

John G. Galaty
McGill University, Montreal, QC, Canada

Johan Galtung
TRANSCEND Peace University

Antonis Gardikiotis
Aristotle University of Thessaloniki, Thessaloniki, Greece

Elizabeth Garnsey
Institute for Manufacturing, University of Cambridge, Cambridge, UK

Stephen Gaukroger
University of Sydney, Sydney, NSW, Australia

Marco Giugni
University of Geneva, Geneva, Switzerland

Jack A. Goncalo
University of California, Berkeley, CA, USA

Claudia L.R. Gonzalez
University of Lethbridge, Lethbridge, AB, Canada

Janet Goodall
University of Bath, Somerset, UK

Anna Grandori
Bocconi University, Milan, Italy

Robert Grassinger
University of Augsburg, Augsburg, Germany

Aaron Gruber
Canadian Centre for Behavioural Neuroscience, University of Lethbridge, Lethbridge, AB, Canada

Sergei Hackel[†]
Lewes, UK

Andrea Hajek
University of Glasgow, Glasgow, UK

Catherine Hakim
Civitas, London, UK

Christian Hakulinen
Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland

Michael T. Hannan
Stanford University, Stanford, CA, USA

Cynthia Hardy
University of Melbourne, Parkville, VIC, Australia

Alexander Haslam
School of Psychology, University of Queensland, St. Lucia, QLD, Australia

Hannah J. Haysom
University of Queensland, Brisbane, QLD, Australia

Paul Heffernan
Institute for Manufacturing, University of Cambridge, Cambridge, UK

Johannes Hewig
Julius Maximilians University Würzburg, Würzburg, Germany

Gene M. Heyman
Boston College, Chestnut Hill, MA, USA

Jacob B. Hirsh
Institute for Management & Innovation, University of Toronto, Mississauga, ON, Canada; and Rotman School of Management, University of Toronto, Toronto, ON, Canada

Heather K. Hood
Ryerson University, Toronto, ON, Canada

Felicia Hughes-Freeland
London University, London, UK

Katri Huutoniemi
Helsinki, Finland

Dan Ispas
Illinois State University, Normal, IL, USA

Myles W. Jackson
New York University, New York, NY, USA

[†] Deceased.

Richard Johnstone
Griffith University, Nathan, QLD, Australia

Darrick Jolliffe
University of Greenwich, London, UK

Çigdem Kagitcibasi
Koc University, Istanbul, Turkey

Christian Kandler
Bielefeld University, Bielefeld, Germany

Benjamin R. Karney
University of Florida, Gainesville, FL, USA

Norbert Kathmann
Humboldt-Universität zu Berlin, Berlin, Germany

Liisa Keltikangas-Järvinen
Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland

Franz Kernic
ETH Zurich / Swiss Federal Institute for Technology Zurich, Zurich, Switzerland

Patrick V. Kirch
University of California, Berkeley, CA, USA

Howard J. Klein
The Ohio State University, Columbus, OH, USA

Ryan Klinger
Old Dominion University, Norfolk, VA, USA

Petra Korte
Dülmen, Germany

Uta Kraus
Julius Maximilians University Würzburg, Würzburg, Germany

Jakob Krause-Jensen
Danish School of Education, Aarhus University, Copenhagen, Denmark

Alexander W. Krieg
University of Hawaii at Manoa, Honolulu, HI, USA

Herbert M. Kritzer
University of Minnesota Law School, Minneapolis, MN, USA

Maribeth Kuenzi
Cox School of Business, Southern Methodist University, Dallas, TX, USA

Darcy K. Leach
Bradley University, Peoria, IL, USA

David M. Ledgerwood
Wayne State University School of Medicine, Detroit, MI, USA

Peter J. Lineham
School of Humanities, Massey University, North Shore, Auckland, New Zealand

Brian R. Little
University of Cambridge, Cambridge, UK; and Carleton University, Ottawa, ON, Canada

Sarah Lubik
Simon Fraser University, BC, Canada

Ian M. Lyons
University of Western Ontario, London, ON, Canada

Joanne Lyubovnikova
Aston University, Birmingham, UK

Peter Mair
Leiden University, Leiden, The Netherlands

Mark Mallon
Old Dominion University, Norfolk, VA, USA

Joseph J. Martocchio
University of Illinois at Urbana-Champaign, Champaign, IL, USA

Ryan K. Masters
Department of Sociology and Institute of Behavioral Science, University of Colorado at Boulder, Boulder, CO, USA

Gerald Matthews
Institute for Simulation and Training, University of Central Florida, Orlando, FL, USA

Enrique Mayer
Yale University, New Haven, CT, USA

John D. Mayer
University of New Hampshire, Durham, NH, USA

Daniel McFadden
University of California, Berkeley, CA, USA

Martin C. McGuire
University of California-Irvine, Irvine, CA, USA

Neil K. Mehta
Rollins School of Public Health, Emory University, Atlanta, GA, USA

Katharina Meyerbröker
Department of Psychiatry, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

Goran Milas
Institute of Social Sciences Ivo Pilar, Zagreb, Croatia

Tricia A. Miller
University of California Riverside, Riverside, CA, USA

Katarína Millová
Institute of Psychology of Academy of Sciences of the Czech Republic, Brno, Czech Republic

Boris Mlačić
Institute of Social Sciences Ivo Pilar, Zagreb, Croatia

Julia Moeller
University of Helsinki, Helsinki, Finland

Agustin Molina
University of Valencia, Valencia, Spain

Dominic Mudiayi
Canadian Centre for Behavioural Neuroscience, University of Lethbridge, Lethbridge, AB, Canada

Andreas Müller
Heinrich Heine Universität Düsseldorf, Düsseldorf, Germany

Martin Müller
University of Zurich, Zürich, Switzerland

Patrick Mussel
Julius Maximilians University Würzburg, Würzburg, Germany

Janken Myrdal
Swedish University of Agricultural Sciences, Uppsala, Sweden

Wilhelm F.H. Nicolaisen
University of Aberdeen, Aberdeen, UK

Yumi Nishikii
Fukuoka, Japan

Gertrud Nunner-Winkler
University of Munich; Max Planck Institute for Human Cognitive and Brain Sciences, Munich, Germany

Karl-Dieter Opp
University of Leipzig, Leipzig, Germany; and University of Washington, Seattle, WA, USA

Dennis W. Organ
Bloomington, IN, USA

Roman Osinsky
Julius Maximilians University Würzburg, Würzburg, Germany

Femi Oyebode
National Centre for Mental Health, University of Birmingham, Birmingham, UK

Femi Oyebode
School of Clinical and Experimental Medicine, University of Birmingham, Birmingham, UK

Susanna Paasonen
University of Turku, Turku, Finland

Marko Paelecke
Julius Maximilians University Würzburg, Würzburg, Germany

A.T. Panter
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Heeman Park
The Ohio State University, Columbus, OH, USA

Scott Parrigon
Purdue University, West Lafayette, IN, USA

Stephanie R. Pavlos
St. John's University, Jamaica, NY, USA

Deborah M. Pearsall
University of Missouri, Columbia, MO, USA

Brienna Perelli-Harris
School of Social Sciences, University of Southampton, Southampton, UK

Nancy M. Petry
University of Connecticut Health Center, Farmington, CT, USA

Robert A. Phillips
University of Toronto, Toronto, ON, Canada

Arthur E. Poropat
School of Applied Psychology, Griffith University, Mount Gravatt, QLD, Australia

Daniel A. Powell
VA Medical Center, University of South Carolina, Columbia, SC, USA

Mark B. Powers
Behavioral & Integrative Treatment Development Lab, Institute for Mental Health Research, University of Texas at Austin, Austin, TX, USA

Asa R. Randall
University of Oklahoma, Norman, OK, USA

Nigel Rapport
University of St. Andrews, St. Andrews, UK

Jerome R. Ravetz
Institute for Science, Innovation and Society, University of Oxford, Oxford, UK

Stephen D. Reicher
School of Psychology, University of St. Andrews, Fife, UK

David B. Resnik
National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA

Karlene H. Roberts
Berkeley University of California, Berkeley, CA, USA

Michael D. Robinson
North Dakota State University, Fargo, ND, USA

Luis Roniger
Wake Forest University, Winston-Salem, NC, USA; Hebrew University of Jerusalem, Jerusalem, Israel

Paul R. Rosenbaum
Wharton School, University of Pennsylvania, Philadelphia, PA, USA

Shawn W. Rosenberg
University of California, Irvine, CA, USA

Clay Routledge
North Dakota State University, Fargo, ND, USA

Rachel Saef
Purdue University, West Lafayette, IN, USA

Katariina Salmela-Aro
University of Helsinki, Helsinki, Finland

Daniel D. Santos
Universidade de São Paulo, São Paulo, Brazil

Vassilis Saroglou
Université catholique de Louvain, Louvain-la-Neuve, Belgium

Paolo Savoia
Harvard University, Cambridge, MA, USA

Heather M. Schellinck
Dalhousie University, Halifax, NS, Canada

Barbara Schneider
Michigan State University, MI, USA

Michal M. Schodl
The Guilford Glazer Faculty of Management, Ben-Gurion University of the Negev, Israel; and Faculty of Health Sciences, Ben-Gurion University of the Negev, Israel

Georg Schreyögg
Freie Universität Berlin, Berlin, Germany

David Sciulli
Texas A&M University, College Station, TX, USA

W. Richard Scott
Stanford University, Stanford, CA, USA

Leslie E. Sponsel
University of Hawaii, Honolulu, HI, USA

Mark C. Suchman
Brown University, Providence, RI, USA

Reinhard Suck
Universitat Osnabrueck, Osnabrueck, Germany

Jörg Sydow
Freie Universität Berlin, Berlin, Germany

Nancy Tandler
Martin-Luther-Universität Halle-Wittenberg, Halle, Germany

Töres Theorell
National Institute for Psychosocial Factors and Health, Stockholm, Sweden; Karolinska Institutet, Sweden; and Stress Research Institute, Stockholm University, Sweden

Eugene F. Thibadeau
Indiana University of Pennsylvania, Indiana, PA, USA

Paul Thompson
The University of Strathclyde, Glasgow, UK

María Elena Torre
The Graduate Center, City University of New York, New York, NY, USA

Peter Tyrer
Imperial College, London, UK

Jan Douwe van der Ploeg
Wageningen University, Wageningen, The Netherlands

Dina Van Dijk
The Guilford Glazer Faculty of Management, Ben-Gurion University of the Negev, Israel; and Faculty of Health Sciences, Ben-Gurion University of the Negev, Israel

Ngo Van Long
McGill University, Montreal, QC, Canada

Arthur van Soest
Tilburg University, Tilburg, The Netherlands

Katherine van Wormer
University of Northern Iowa, Cedar Falls, IA, USA

Daniel M. Varisco
Qatar University, Doha, Qatar

Karin J.H. Verweij
Department of Developmental Psychology, EMGO Institute for Health and Care Research, VU University, Amsterdam, The Netherlands

Shawna Vican
Harvard University, Cambridge, MA, USA

Claus Vögele
Institute for Health and Behaviour, University of Luxembourg, Walferdange, Luxembourg

Stéphanie Wahab
University of Otago, Dunedin, New Zealand; and Portland State University, Portland, OR, USA

Kate E. Walton
St. John's University, Jamaica, NY, USA

Tony J. Watson
University of Nottingham, Nottingham, UK

Elke U. Weber
Center for the Decision Sciences, Columbia University, New York, NY, USA

Peter Weingart
University of Bielefeld, Bielefeld, Germany

Michael G. Wessells
Columbia University, New York, NY, USA

Michael A. West
Lancaster University, Lancaster, UK

Monica T. Whitty
University of Leicester, Leicester, UK

Bernhard Wilpert[†]

Scott Wong
Canadian Centre for Behavioural Neuroscience, University of Lethbridge, Lethbridge, AB, Canada

Sang E. Woo
Purdue University, West Lafayette, IN, USA

Melissa E. Wooten
University of Massachusetts, Amherst, MA, USA

Susan Wright
Danish School of Education, Aarhus University, Copenhagen, Denmark

Yiyuan Xu
University of Hawaii at Manoa, Honolulu, HI, USA

Bilge Yagmurlu
Koc University, Istanbul, Turkey

Hiroshi Yamashita
Fukuoka, Japan

Yang C. Yang
Department of Sociology and Lineberger Cancer Center, Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Keiko Yoshida
Fukuoka, Japan

Asghar Zaidi
University of Southampton, Southampton, UK

Matthias Ziegler
Institut für Psychologie, Humboldt-Universität zu Berlin, Berlin, Germany

Brendan P. Zietsch
University of Queensland, Brisbane, QLD, Australia

[†] Deceased.

This page intentionally left blank

Normality and Mental Health

Femi Oyebode, National Centre for Mental Health, University of Birmingham, Birmingham, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Offer, M. Sabshin, D. Albert, volume 14, pp. 9639–9645, © 2001, Elsevier Ltd.

Abstract

Concepts of normality are implicit in any description of health and disease. Although 'normality' is not itself defined, concepts of normality can be inferred from the concepts of health and disease. Normality as an ideal is described and then the varying concepts of disease are examined and reviewed. In psychiatry, these concepts are contested. The major and most influential theses of the fact, value, and concepts of mental disorder are examined and discussed. These include the contributions of Thomas Szasz, Christopher Boorse, Peter Sedgwick, Jerome Wakefield, and Robert Kendell.

Definitions of Normality

Definitions of normality go to the heart of medicine, its nature, and central concerns. It is how departures from normative values are defined and described that determines what the focus of interest of medicine ought to be. There is little doubt, too, that social context matters and that the definition of normality and by extension of disease and illness is constantly being negotiated between individuals, their doctors, and the wider society. It is also true that cultural beliefs and practices influence the manner in which medicine is practiced. In psychiatry, there is a further set of complications because the deviations from normative values are often, if not exclusively, behavioral deviations. This raises the profound problem of the boundaries of 'normal' behavior and whether aberrant behavior is evidence of disease or not. The history of psychiatry exemplifies the risk of identifying behavior that is socially disapproved of as evidence of disease. Behaviors such as masturbation, homosexuality, and slaves running away from their owners in the United States of America have all been designated as evidence of disease in recent times. Below is an examination of the various ways, in which disease is conceived in medicine starting with idealized concept of health promoted by the World Health Organization (WHO).

Normality as Idealized Norm

The WHO defines health as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" (WHO, 1948). This definition does not define disease itself but rather, equates health with the complete absence of disease. Thus disease is a departure from this *ideal* state. But the absence of disease is itself not a state of health. In this description, not having any identifiable disease is not synonymous with health. The emphasis is not upon the absence of disease but rather upon some yet to be determined ideal state of 'complete physical, mental, and social well-being.' It is also significant that 'normality' too is not defined.

Disease as Departure from Statistical Norm

The statistical approach to defining what is a disease is best summed up as a disease is a quantitative deviation from the norm (Cohen, 1981). The application of statistics to population values as a basis for determining whether disease is present or not has the advantage that the data appear objective and the diagnostic process has at least a superficial precision. Since the determination of disease status is the product of mathematical calculation, it is therefore assumed to be objective, value-free, and reliable. This means that a diagnostic process that relies on such data is thought of as independent of any particular observer's opinion and preferences. Measurements of height, weight, blood pressure, intelligence, and so on fall into this category and departures from mean values as a basis of making diagnostic decisions exemplify this paradigm. Hence the diagnosis of obesity, hypertension, learning disability, and so on is determined by objective scores that can be fitted to a chart and shown to deviate from the mean values or to fall without two standard deviations from the mean. Child development charts that help to track the height and weight of a child and readily identify anomalies in growth, for example, derive their power and influence from this way of conceiving disease.

Disease as Lesion (Typology)

Rudolf Virchow (1821–1902) located disease within the cell and argued that diseases are the physiological outcome of altered conditions. More significantly, he had the notion that the cellular changes were observable under light microscopy and demonstrated characteristic phenomena that were both describable and invariant. This notion of disease-as-lesion was occurring alongside the germ theory of disease propagated by Heinrich Koch (1843–1927). Koch's postulates established the conditions for identifying a particular microorganism as the causative agent of disease. The contributions of Virchow and Koch were followed by spectacular triumphs as pathological anatomy and identifiable causative agents advanced the understanding of diseases. The characteristic lesion of pulmonary tuberculosis visualized under light microscopy following staining is an example of the power of a visual image denoting the pathophysiology of an infective process, rendering visible and recognizable what was until then invisible. Thus, this approach to diagnosis embodied the modern approach to scientific medicine, what has come to be termed the 'medical model.' This can be the promise of science as a method to

unravel the mysteries of diseases and to bring the complexities of disease processes under scrutiny and control.

Disease as Loss of Homeostatic Equilibrium

The realization that disease-as-lesion paradigm failed to explain many of the already known diseases led to further exploration of other notions of disease. Terms such as 'disorder,' 'dysfunction,' and 'maladaptation' refer to a recognition that there might be failures of function without identifiable lesions. Diseases such as migraine and trigeminal neuralgia cause distress and disability but occur without readily identifiable pathological anatomy. The idea that health is a state of internal homeostatic equilibrium and that loss of this equilibrium can result in disorder or dysfunction at least went some way in potentially explaining some conditions. The regulation of hormones by feedback loops (positive and negative feedback) and the effects of hormones on target organs at a distance from the site of the hormonal glands gave additional impetus to this notion of homeostatic equilibrium as an aspect of health. Thus disruption of the fragile homeostatic balance could result in disease without apparent lesions. Diabetes mellitus is the primary example of a disease that results from a dysregulation of homeostatic balance. Once the notion of homeostasis of the internal physical environment was grasped, it was only a matter of time for the potential disruption of the intricate relationship of the individual to his social and cultural environment to be recognized as a basis for disease. Thus, the external homeostatic balance between an individual and his physical and social environments could be envisioned as the context of health or normality and disruption of this balance could be conceived of as potentially causative of disease.

Disease as Social Deviance

The foregoing discussions about disease-as-departure from statistical norm, disease-as-lesion, and disease-as-loss of homeostatic equilibrium are often described as the 'medical model' of disease. Talcott Parsons (1902–79) argued that health is included in the functional needs of the individual member of society so that from the point of view of functioning of the social system, too low a general level of health or too high an incident of illness is dysfunctional. Disease in this formulation incapacitates the effective performance of social roles and there is, therefore, social interest in the alleviation of disease. To put this in another way, disease is not purely or merely a natural phenomenon but a state of disturbance of the total human being including the state of the organism as a biological system and of his personal and social adjustments. This approach introduces the notion that behavioral deviance itself can be the source of disease. But, how is such deviance to be recognized and defined. Social deviance can be recognized by self-definition. The individual may come to hold the belief that he has a problem or there may be a societal reaction that an individual's behavior constitutes a problem. Societal reaction of this type might occur when a community comes to recognize a person's inability or reluctance to respond in a particular expected way. According to David Mechanic (1968), "the view taken of the deviant depends in large part on the frame of reference of the observer and the extent to which the deviant appears to be able or willing to control his responses. The evaluator views the act within the context of what he believes the actor's motivation to be. If the action appears reasonable in terms of the assumed motivation of the actor, there is a very good chance that deviant behavior will be defined in terms of the goodness–badness dimension. If the behavior appears to be peculiar and at odds with expectations of how a reasonable person might be motivated, such behavior is more likely to be characterized in terms of the sickness dimension." The problem with this is self-evident. Disease definition in this formulation seems significantly prone to error, to subjective judgment, and liable to be used as a tool of social control.

The degree to which the recognition of psychiatric disorders relies on societal responses makes psychiatric notions of disease problematic if not controversial. This fact has been used by various critics of psychiatry as the basis of their case for the illegitimacy of psychiatric diagnoses and for emphasizing the extent to which psychiatric practice is an extension of society's mechanism for social control of its members. Application of theories and concepts of disease in psychiatry have faced widespread criticism. At the heart of the debate is the distinction between objectivity of science and the subjective value judgment of individual practitioners. The most fundamental issue, and also the most contentious one, is whether disease and illness are normative concepts based on value judgments, or whether they are value-free scientific terms; in other words, whether they are biomedical terms or sociopolitical ones (Kendell, 1986). Below is a summary of the main critiques and their contribution to the philosophy of psychiatry.

Critique of Disease Model in Psychiatry

Thomas Szasz

Thomas Szasz (1920–2012) argued that mental diseases are purely normative concepts and that they exhibit only behavioral symptoms and behavioral abnormalities and furthermore that psychiatric disorders do not have demonstrable physical causes and hence should not be treated by physical intervention (1974). For Szasz, *real* diseases are by definition bodily diseases, in which there are disturbances of function or structure of an organ or part of the body. Szasz develops his argument by appealing to Anthony Ryle's view that the notion 'mind' is a theoretical concept and that the term does not denote an organ or part of the body. So, Szasz concludes that the mind cannot be diseased in the same sense as the body can. This controversial viewpoint extends well beyond commonsense notions of disease because it has implications for whether psychiatric disorders such as schizophrenia ought to receive medical attention or not and whether crimes committed in the context of abnormality of mind ought to excusable or not. Psychiatrists have responded to Szasz's charge by arguing either that many 'physical' disorders lack established characteristic pathology or that mental symptoms and abnormal mental states are epiphenomena of underlying physical disturbances yet to be revealed (Clare, 1979).

At the root of Szasz's contention is the age-old mind–body problem. There is little doubt that brain diseases such as Alzheimer's dementia and Parkinson's disease that have an

established morbid anatomy often present with mental symptoms such as depression, anxiety, persecutory beliefs, and hallucinatory experiences. In other words, coarse organic brain disease can manifest as mental symptoms. It is therefore reasonable to assume that in the absence of coarse organic brain change, mental symptoms of this kind can be attributed to yet to be discovered functional abnormalities. It is also apparent that given the complexity of the brain as a conglomeration of neural systems, it will be prone to error and malfunction. Arguing that the brain is exempt from abnormality or that mental symptoms are invariably due to social and cultural causes is merely to deny the role of the brain in mental life. Notwithstanding the problem of mind–body duality and how to reconcile facts about the brain and the apparent independence of mental phenomena, there is an undeniable link between mind and brain.

Christopher Boorse

Christopher Boorse's distinction between disease and illness is deservedly influential. He argued that an organism is *healthy* at any moment in proportion as it is not diseased; and a *disease* is a type of internal state of the organism that interferes with the performance of some natural function, that is some species typical contribution to survival and reproduction. And, that the disease state is not simply in the nature of the species, i.e., is either atypical of the species or, if typical, mainly due to environmental causes. Diseases become illnesses only when they satisfy certain further, and normative, conditions. A disease is an illness only if it is serious enough to be incapacitating, and therefore is regarded as undesirable, a title for special treatment and a valid excuse for normally criticizable behavior. For Boorse, mental functions such as perceptual processing, intelligence, and memory clearly serve to provide information about the world that can guide effective action. Drives serve to motivate it. Anxiety and pain function as signals of danger, language as a device for cultural cooperation and cognitive enrichment, and so on. He concludes, "it seems certain that a few of the recognized mental disorders are genuine diseases, whether mental or physical. Even without any knowledge of the relevant functional systems, one can sometimes infer internal malfunction immediately from biological incompetent behavior." Finally, Boorse thought that diseases are what doctors treat and illnesses are what people suffer from (Boorse, 1976).

Boorse's approach avoids the problem that Szasz posed, namely, that mental symptoms do not derive from bodily disease. Boorse locates certain mental functions firmly within the domain of brain function and concludes that impairment of these functions speaks directly to brain malfunction. However, there are objections to Boorse's approach. For example, Fulford (2001) in an open letter to Boorse argues that a value-free naturalist account such as that by Boorse fails when it is applied to psychiatric disorders because psychiatry is concerned with "desires, beliefs, emotions, motivations and so forth, areas of experience and behavior in which human values are characteristically and legitimately *diverse*" (italics in the original). For Fulford and Sedgwick (see below), value-free descriptions of diseases are impossible.

The distinction between disease and illness that Boorse draws is also the subject of Kleinman et al. (1978) application of anthropological concepts to medicine. They argue that a disease is "malfunctioning or maladaptation of biologic and psychophysiologic processes in the individual; whereas illness represents personal, interpersonal, and cultural reactions to disease or discomfort" (Kleinman et al., 1978). Hence, illness is not merely a disease that "satisfies certain, further and normative, conditions" (see above) but it is "culturally shaped in the sense that how we perceive, experience, and cope with disease is based on our explanations of sickness, explanations specific to the social position we occupy and systems of meaning we employ" (Kleinman et al., 1978). In this scheme, illness is dominant in the patient's mind, as it is constituted of the difficulties in living that the patient experiences, whereas disease is the central focus of attention of doctors. Kleinman et al. (1978) conclude that both illness and disease are explanatory concepts that help to construct the forms of clinical reality that is faced in clinical practice. In this formulation, they anticipate Peter Sedgwick's views as described below.

Peter Sedgwick

Peter Sedgwick argues that "It may prove possible to reduce the distance between psychiatry and the other streams of medicine…not by annexing psychopathology to the technical instrumentation of the natural sciences but by revealing the character of illness and disease, health and treatment, as social constructions." In other words, all medicine is underscored by values. He goes on "The blight that strikes at corn or at potatoes is a human intervention, for if man wished to cultivate parasites (rather than potatoes or corn) there would be no "blight," but simply the necessary foddering of the parasite-crop." Furthermore, "A tiger may experience pain or feebleness from a variety of causes. It may be infected by a germ, trodden by an elephant, scratched by another tiger, or subjected to the aging processes of its own cells. It does not present itself as being *ill* except in the eye of a human observer." It is therefore clear that "The medical enterprise is from its inception value loaded; it is not simply an applied biology, but a biology applied in accordance with the dictates of social interest." Sedgwick concludes "All illness is essentially deviancy. That is to say, no attribution of sickness to any being can be made without the expectation of some alternative state of affairs which is considered more desirable. All illness, whether conceived in localised bodily terms or within a larger view of human functioning, expresses both a social value-judgement…and an attempt at explanation. The physicalist psychiatrists are wrong in their belief that they can find objective-disease entities representing the psychopathological analogues to diabetes." Sedgwick predicts that "*The future belongs to illness*: we are just going to get more and more diseases, since our expectations of health are going to become more expansive and sophisticated" (Sedgwick, 1982).

Sedgwick's position is that heart failure, for example, or any disease for that matter is of interest to doctors primarily because the symptoms are negatively valued – breathlessness, swollen ankles, general malaise, and premature death. These complaints, what Boorse would call illness, draw attention to a devalued state that is then the focus of medical attention, thus revealing a disease process that informs treatment. For Sedgwick, this process is as true for a physical disease such as diabetes as for a mental disease such as schizophrenia. Hence in his terms, there are no illnesses or diseases in nature. All illness,

whether conceived in localized bodily terms or within a larger view of human functioning, express both a social value judgment and an attempt at explanation.

Jerome Wakefield and Teleology

Jerome Wakefield's account is sometimes referred to as a 'hybrid account.' Like Sedgwick, he accepts that social values determine what symptoms are of interest to medicine. But unlike Sedgwick, he argues for *dysfunction* as a failure of a biological system: "a disorder exists when the failure of a person's internal mechanisms to perform their function as designed by nature impinges harmfully on the person's wellbeing as defined by social values and meanings" (Wakefield, 1992). This view is a variant on the criterion of biological disadvantage as the defining characteristic of 'disorder' set out first by Scadding (1967, 1990) and later by Kendell (1975).

Scadding (1990) made the point that "The name of a disease refers to the sum of the abnormal phenomena displayed by a group of living organism in association with a specific common charcateristics, or set of characteristics, by which they differ from the norm for their species in such a way as to place them at biological disadvantage" and Kendell (1975) elaborated this view by linking biological disadvantage to survival and reproductive fitness, namely, lowered fertility and reduced longevity. This proposition allows for disease status to be conferred by an objective judgment that the identified phenomenon places an individual at biological disadvantage either by causing premature death or by reducing reproductive fitness.

Wakefield's fully developed concept of mental disorder runs as follows: "A disorder is a harmful dysfunction, where 'harmful' is a value term, referring to conditions judged negative by sociocultural standards, and 'dysfunctional' is a scientific factual term, referring to failure of biologically designed functioning. In modern science, 'dysfunction' is ultimately anchored in evolutionary biology and refers to failure of an internal mechanism to perform one of its naturally selected functions" (Wakefield, 1992). Here, Wakefield is distinguishing between 'biological disadvantage' as enunciated by Scadding and 'failure of function' as described in his own approach.

The problem with Wakefield's position is that its explanatory power relies on a teleological argument: dysfunction can only be understood in relation to a goal, the *telos* of the function. And, it is often hard to comprehend what the biological goal of the identified function is, be it emotion, motivation, drive, or whatever. And, to attempt to explain an event, by describing the goal of that event is not equivalent to describing its cause. One of the characteristics of causal explanation is the temporal ordering of events – effects cannot precede their causes.

Summary

Normality is an important concept in medicine but is hardly ever directly defined. It is implicit in the WHO definition of health and concepts of 'normality' inform the various ways in which disease is conceived and defined. It is clear that *values* are important in the recognition of all diseases as entities falling under the interest of medicine but that objective, value-free explanations, that is, science informs understanding of disease processes. This is as true for 'physical' diseases as it is for 'mental' diseases.

See also: Culture as a Determinant of Mental Health; Health Psychology; Health and Illness: Mental Representations in Different Cultures; Health: Anthropological Aspects; Personality Theory and Psychopathology.

Bibliography

Boorse, C., 1976. What a theory of mental health should be. Journal of the Theory of Social Behavior 6, 61–84.

Clare, A., 1979. The disease concept. In: Hill, P., Murray, R., Thorley, A. (Eds.), Essentials of Postgraduate Psychiatry. Academic Press, London, pp. 55–76.

Cohen, H., 1981. The evolution of the concept of disease. In: Caplan, A.L., Englehardt Jr., H.T., McCartney, J.J. (Eds.), Concepts of Health and Disease: Interdisciplinary Perspectives. Addison-Wesley, Reading MA, pp. 209–220.

Fulford, K.W.M., 2001. 'What is (mental) disease?': an open letter to Christopher Boorse. Journal of Medical Ethics 27, 80–85.

Kendell, R.E., 1975. The concept of disease and its implications for psychiatry. British Journal of Psychiatry 127, 305–315.

Kendell, R.E., 1986. What are mental disorders? In: Freedman, A.M., Brotman, R., Silverman, I., Hutson, D. (Eds.), Issues in Psychiatric Classification: Science, Practice and Social Policy. Human Science Press, New York, pp. 23–45.

Kleinman, A., Eisenberg, L., Good, B., 1978. Clinical lessons from anthropologic and cross-cultural research. Annals of Internal Medicine 88, 251–258.

Mechanic, D., 1968. Medical Sociology. The Free Press, New York.

Parsons, T., 1951. The Social System. The Free Press, New York.

Scadding, J.G., 1990. The semantic problem of psychiatry. Psychological Medicine 20, 243–248.

Scadding, J.G., 1967. Diagnosis: the clinician and the computer. The Lancet 290, 877–882.

Sedgwick, P., 1982. Psycho Politics. Harper and Row, New York.

Szasz, T., 1974. The Myth of Mental Illness: Foundations of a Theory of Personal Conduct. Harper and Row, New York.

Wakefield, J.C., 1992. The concept of mental disorder: on the boundaries between biological facts and social values. American Psychologist 47, 373–388.

World Health Organization, 1948. Preamble to the Constitution of the World Health Organization as Adopted by the International Health Conference, New York, 19 June–22 July 1946; Signed on 22 July 1946 by the Representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948.

Norms

Karl-Dieter Opp, University of Leipzig, Leipzig, Germany; and University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Norms are a major focus of attention of all social sciences, particularly sociology, political science, economics, anthropology, and social psychology. The social sciences are interested in the empirical study of norms, whereas jurisprudence and philosophy (moral philosophy and deontic logic) address the question of what 'good' norms are and how normative reasoning can be formalized. This article addresses the most important issues in the empirical study of norms.

We begin by outlining some basic facts about norms because explanations of norms should address these facts. Before proceeding further, it is useful to elaborate on the major dimensions of the norms concept that are found in the literature and to ask what norms definition is to be preferred. The next issue is how norms can be measured. The central theoretical questions are how norms originate and, if they exist, what effects they have. These questions are addressed in the last two sections of the article.

Norms are a major focus of attention of all social sciences, particularly sociology, political science, economics, anthropology, and social psychology. The social sciences are interested in the empirical study of norms, whereas jurisprudence and philosophy (moral philosophy and deontic logic) address the question of what 'good' norms are and how normative reasoning can be formalized. The literature on norms is so vast that a short article on this subject has to be restrictive. The following sections address the most important questions of the study of norms. We begin by outlining some basic facts about norms because explanations of norms should address these facts. Before proceeding further, it is useful to elaborate on the major dimensions of the norms concept that are found in the literature and to ask what norms definition is to be preferred. The next issue is how norms can be measured. The central theoretical questions are how norms originate and, if they exist, what effects they have. These questions are addressed in the last two sections of the article.

Some Basic Facts about Norms

When children grow up, they consider norms that are learned as universally valid. It is generally 'right' that one has to cover one's mouth when one yawns or that one must always tell the truth. But after some time, the child learns that there is a large heterogeneity of norms: different groups subscribe to different norms. This holds not only for immigrants but also for urban and rural populations or for younger and older people. The heterogeneity is still greater between countries. In India, for example, many people eat with their fingers. There is still female genital mutilation in several countries. In some countries like China tips are not given. Another basic fact about norms is that they change over time, in a given country as well as across countries. Some time ago it was scorned that women smoked in public. Foot-binding in China is no longer accepted. Another important feature of most norms is their conditionality. 'One must not lie' holds only under certain conditions. There may be white lies, that is, lies that avoid hurting others. In these situations, lies are allowed or even normatively expected. Another fact about norms is that they are related to each other. For example, a relatively general norm would be not to impose externalities on others. This implies or justifies the norm not to smoke in the presence of others, assuming that this hurts others. Norms vary in regard to their clarity. For example, under what conditions is what kind of 'self-defense' allowed? There are further many vague concepts in laws, especially in the constitutional laws.

So far the referents of norms were behavior in a wide sense, including thinking (one must not think of murdering somebody). There are further norms referring to beliefs (Christians must believe in God) or to goals, attitudes, and emotions ("Thou shalt not covet thy neighbour's wife," or "Love thy neighbour as thyself"). These facts are interesting in their own right, but also a challenge for explaining their origins and their effects.

The Definition of Norms

The literature abounds with definitions of norms. An overview and discussion of many norm definitions is provided by Gibbs (1965), Biddle and Thomas (1966), Rommetveit (1953), and Williams (1968). Crawford and Ostrom (1995) give a detailed analysis of the 'grammar' of norms. In this section, the dimensions of the norms concept that most definitions include are described, and it is discussed which definition of the norms concept is preferable. George C. Homans' definition of norms includes most of the dimensions that are ingredients of other definitions as well:

> A norm is a statement specifying how a person is, or persons of a particular sort are, expected to behave in given circumstances—expected, in the first instance, by the person that utters the norm. What I expect of you is what you ought to do (1974, p. 96).

Homans refers to four dimensions. One is oughtness: a norm is an expectation that some behavior ought to be (or ought not to be) performed. This is a basic characteristic that almost all norms definitions have in common. This is the oughtness definition of norms. A second dimension is that

people are expected to behave in a certain way 'in given circumstances'. Thus, norms may be conditional. A norm such as 'you shall not kill' holds only in certain circumstances or for certain kinds of actors (for example, soldiers are allowed to kill in times of war). Homans' definition further implies that norms refer to behavior, not to beliefs or attitudes. In contrast, Morris' definition (1956) also includes beliefs and attitudes (for examples see the first section). Fourth, the actors (i.e., the norm senders) who hold (or utter) a norm are individuals. In other definitions, the actors are collectives such as parliaments.

Two other criteria are used in many definitions of norms. Morris (1956) defines norms as "generally accepted, sanctioned prescriptions…" (p. 610). Thus, an expectation is, by definition, only a norm if the expectation is shared by the members of some group, and if conforming to or deviating from the expectation is sanctioned with some probability. It is important to note that a normative expectation is not identical to a preference. For example, I may not like it when you smoke in my company, but I may think that you are allowed to smoke.

A second type of definition – which may be called a behavioral definition of norms – states that a norm exists if there is a behavioral regularity and if sanctioning occurs with some positive probability in the case of deviating from this regularity. 'Sanctioning' means that others impose a cost on a person if a behavior is or is not performed. For example, a behavioral regularity may be that people keep promises. If someone does not keep a promise and if the person is sanctioned with some probability, then a norm exists. Refraining from a behavior may be a behavioral regularity as well. If people usually do not smoke (a 'behavioral' regularity) and if smoking triggers sanctions there is a nonsmoking norm. Thus, 'behavior' implies 'refraining from behavior'.

If a norm exists according to the behavioral definition, there need not be any element of oughtness. Assume that smoking is allowed (i.e., there may be no nonsmoking norm in the sense of the oughtness definition) and let nonsmoking be a behavioral regularity. Some nonsmokers do not like to be exposed to smokers. If this nonetheless happens smokers may be sanctioned: nonsmokers move to other tables in restaurants or express their anger about being disturbed. These actions may be costly to smokers, but they have nothing to do with a norm. Thus, although there is a behavioral regularity (most people do not smoke) and although there is some probability that deviations are 'sanctioned,' no norm is involved. Only if 'sanction' is defined as a reaction to the violation of a norm (in the sense of the oughtness definition) the behavioral definition includes oughtness. But most proponents of the behavioral definition do not like the oughtness definition because they prefer a pure 'behavioral' definition that allegedly allows easier measurement of norms (see below).

There are other definitions of norms in the literature. For example, Ensminger and Knight (1997, p. 2) define norms as "informal rules that structure behavior." Thus, only if a rule causes behavior can it be called a norm. If this definition is applied, it is no longer possible to explain a behavior by norms because a norm exists by definition only if it causes behavior. Since such definitions are relatively rare they will not be discussed here.

Laws (that will not be addressed in this article) are a subset of norms: they are norms that are promulgated by a public institution. There is a growing literature that explores differences between norms and laws (see, e.g., Posner and Rasmusen, 1999).

These definitions of norms are often used for defining institutions. The latter term often refers to single norms, as the following definition illustrates: "an institution is a regularity of behaviour or a rule that is generally accepted by members of a social group, that specifies behaviour in specific situations, and that is either self-policed or policed by external authority" (Rutherford, 1996, p. 182). 'Institution' often also denotes sets of norms as well as organizations (for a discussion see Calvert, 1995, pp. 216–218).

What definition is to be preferred? It is striking that despite the multitude of definitions in the literature, authors usually do not give reasons for preferring a certain definition over others. One important criterion for choosing a definition is the extent to which it refers to the explanatory problems social scientists want to solve. With regard to norms, social scientists are interested in explaining, among other things, when people accept normative expectations and under what conditions what kinds of sanctions are attached to normative expectations. Answering these questions is not possible if several dimensions of norms are included in a single complex definition. In those definitions, certain combinations of values are contained (such as existence of a behavioral regularity and a positive probability of sanctioning). Thus, only this combination of values can be explained. The upper part of Figure 1 depicts this situation.

However, it is also important to know how the dimensions are causally related (see the lower part of Figure 1). For example, if norms are shared and if strong sanctions go with the norm, it is likely that the behavior that the norm prescribes is performed. These complex interdependencies cannot be addressed if several dimensions are incorporated into a single

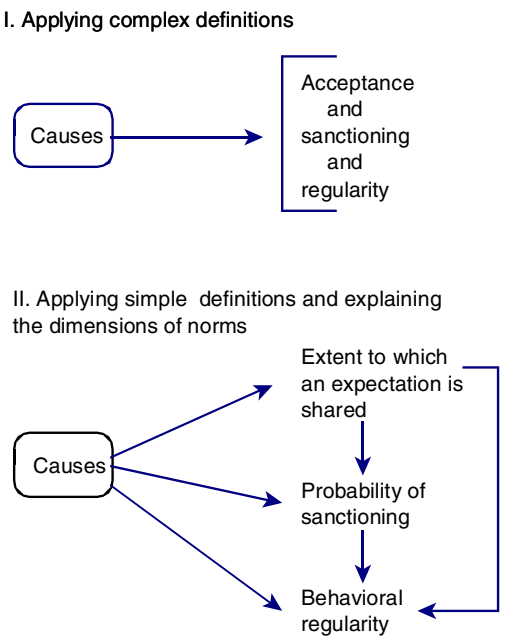

Figure 1 Different procedures for explaining the origins of norms.

definition. Treating the dimensions separately allows us to explain them and look at their interdependencies (for a similar argument see Hechter and Opp, 2001, p. 404).

It is further important that the definition used has theoretical import: it must permit the formulation of true and informative theories. The extent to which this criterion is fulfilled can only be judged if a theory with the respective concepts exists and has been tested. As long as such a theory does not exist, the definition is a preliminary working definition.

The concept of sanction has also different meanings in the literature. It may refer to a behavior that is intended to change a future behavior of a norm breaker. Sanctioning may also refer to a behavior that actually influences the future behavior of a person. Any reaction to a behavior – that need not be related to norms – is also sometimes denoted as a sanction. Researchers who use the sanction concept rarely specify its meaning. What is a theoretically fruitful sanction concept? This question is not addressed in the literature.

How Can Norms Be Measured?

If historians want to get information about norms they will look at documents such as books about manners (see, e.g., Elias, 1994). To learn about the norms of a cohort of young people, sociologists may analyze their diaries. The norms of voluntary associations may be taken from their statutes. One problem with these sources is to what extent the written norms are or were actually in effect. Diaries may only reflect the norms of a specific group of middle class youths. Books of manners may describe rather the ideals of the author than the manners that are effective. Another problem is the extent to which written documents state the conditions under which a norm holds. Members of a group may often not be aware of those conditions so that it is questionable whether documents state a given norm in all its complexity.

It seems that the method of observation is especially suitable for ascertaining norms according to the behavioral definition: behavior, in contrast to expectations, can be observed in principal, but practically, this is often difficult. It is not easy, for example, to measure the frequency of promise keeping (or of breaking promises), of certain sexual behavior, or of smoking. It is practically impossible to observe these behaviors for a large group: numerous observers and intrusion in the private sphere would be necessary. Even if these difficulties can be overcome, recording the behavior (even on a videotape) is not sufficient. A given behavior may have quite different meanings. For example, making a promise is a verbal behavior and the observer must know the language of a group in order to ascertain the extent to which promises are made or kept. Many kinds of sanctioning are verbal behaviors as well, and a given behavior may have different meanings. Shaking of the head may express disapproval but may also be a sign of astonishment. Thus, although behavioral regularities and sanctioning are in principle more accessible than expectations, practical measurement by observational techniques is extremely difficult. The conclusion is that the measurement argument that is often invoked for a behavioral definition is very questionable.

Surveys are probably used most frequently to measure norms. There are different procedures. For example, Jackson (1966) measures the extent of approval or disapproval of a behavioral dimension (such as the extent of approval of the number of times a person speaks in an hour's session of a discussion group). If 'frequency of behavior' is plotted on the x-axis and average approval/disapproval of the members of a group on the y-axis, a curve or line shows what behavior is more or less approved or disapproved. Based on such a curve various statistical measures can be computed. Another measurement procedure is to formulate various items that are designed to capture, among other things, the conditionality of a norm. For example, one of the items used to measure protest norms reads: "nobody can be expected to take part in political actions that might send her or him to jail" (Opp and Gern, 1993, p. 668). A third measurement device is the factorial survey that seems best suited to ascertain the conditionality of a norm. The first step is to specify the conditions under which a norm holds. For example, people may endorse an obligation to protest if the expected repression by police action is low or if the expected success of the collective action is high. Thus, the dimensions 'expected repression' (with values 'no' or 'yes') and group success (with the same values) may be important conditions for a felt obligation to protest. As a next step, descriptions of situations (the vignettes) are constructed from all possible combinations of the dimensions. Each of these situations is then judged by the respondents of the survey: they are asked to what extent a behavior (such as participating in a demonstration) is more or less obligatory in a given situation. The data enable the researcher to ascertain what effect each of the dimensions has on the overall normative judgment performing certain actions under a whole set of conditions (for details see Jasso and Opp, 1997; Opp, 2002).

Qualitative researchers may in general object that norms are not given but are largely 'constructed' in specific situations (see, e.g., Fine, 2001). This implies that the measurement of norms is impossible because they do not exist. Survey research does not confirm this view. If norms do not exist, the prediction is that factorial surveys yield a large number of missing values, which did not turn out to be the case (Jasso and Opp, 1997; Opp, 2002). It is surprising that despite the centrality of norms in the social sciences, there is rather little effort to develop standardized measurement procedures of norms. A detailed comparative discussion of the different procedures is also missing.

The Origins of Norms

The major difficulty of reviewing theories about the emergence of norms is the great variety of propositions, scattered in the literature, and the diversity of theoretical approaches. Recent reviews of the so-called new institutionalisms (see, e.g., Hall and Taylor, 1996) testify this. Nonetheless, a review of various propositions and approaches indicates that there seems to be one basic idea that underlies all or at least most explanations of norms. This is the instrumentality proposition. The idea is that norms emerge if their emergence is in the interest of a collective of people. A major proponent of this idea is Robert Ellickson who "advances the hypothesis that when people are situated in a close-knit group, they will tend to

develop for the ordinary run of problems norms that are wealth-maximizing" (1989, p. 84; see also 1991). Norms are thus instrumental (Opp, 2001). It is important to distinguish the instrumentality proposition from a functionalist hypothesis claim that norms emerge if they contribute to the equilibrium or survival of a social system. This hypothesis is flawed because it is not possible to explain the existence of some item by the effects that it would have if it existed. Such functionalist arguments are still widely used in the literature (see the critique by Hardin (1980) of Ullmann-Margalit (1977)). In contrast, the instrumentality proposition holds that situations or events that jeopardize the realization of group goals of actors provide incentives to the actors to create rules that make the group better off.

The instrumentality proposition is consistent with the widely held view that the emergence of norms is likely if there are externalities. These are actions that incur costs (negative externalities) or provide benefits (positive externalities) to other actors. Those who are affected by negative externalities have an interest in establishing norms that reduce the externalities. Those who cause positive externalities have an interest to establish a norm that compensates them. The basic reference for this theory is Demsetz (1967) (for a discussion see Eggertsson, 1990). This theory underlies Coleman's explanation of norms (Coleman, 1990, Chapters 10 and 11).

Norms are public goods, that is, goods that if provided are consumed (or enjoyed) by every member of a group even if the member has not contributed to their provision. For example, a nonsmoking norm holds for everybody. Public goods theory, which is consistent with the instrumentality proposition, argues that "groups exist in order to supply their members with some desired joint good. This good can be attained only if members comply with rules that are designed to assure its production" (Hechter, 1987, p. 41; see also Oliver, 1980; Heckathorn, 1990; Coleman, 1990). Thus, norms are instrumental in achieving group goals: norms are (second-order) public goods that are instrumental for providing (first-order) public goods.

There are different mechanisms that lead to the emergence of norms. Often, norms emerge by human design such as laws. Assuming that individuals have biased knowledge about the effects of their actions (a lawmaker may issue a law that is not accepted or has unintended negative side effects), it happens that norms are inefficient. This means that alternative norms would make the members of a group better off.

Inefficient norms may also originate due to path dependence (see, in particular, Arthur, 1994; Mahoney, 2000). In its most general sense, this means that the mere existence of norms or institutions heavily influences costs and benefits of change. Even if the Dvorak keyboard would allow much quicker typing than the QWERTY keyboard replacing the latter with the former would be very costly (see David, 1985; Lewin, 2002).

Kitts (2006, 2008) has conducted computer simulations showing that a regulatory interest (Heckathorn, 1988) as a condition for the emergence of norms may decline if selective incentives are rival. This means that the provision of a selective incentive to one member of a group diminishes its value to other members. These will then no longer be motivated to contribute to the collective good but enforce "antisocial norms that discourage peers' work" (Kitts, 2008, p. 143). Reference is here made to norms at the workplace. In this scenario, workers act 'instrumentally' as well, but their action will lead to norms that are likely to have negative effects for the workers themselves as well as for the firm.

We find the instrumentality proposition also in the work of qualitative sociologists. Fine (2001) explains why there is a norm that restrains mushrooming. He notes that mushrooming is competitive: if someone picks a mushroom another one cannot. Mushroomers are thus exposed to a negative externality. In this instance, the norm 'first come, first serve' might apply. However, this solution would breed strain. Mushroomers were interested in good personal relations that were inconsistent with an untamed competition. Thus, the emergence of a norm about constraining the collection of mushrooms is, among other things, due to the goals of the group of mushroom collectors.

Norms are often not the result of human design but the unintended and sometimes undesired consequences of human action. The emergence of such 'spontaneous' norms is consistent with the instrumentality proposition as well. Assume that all of a sudden being exposed to smokers becomes a strong externality for nonsmokers. They will increasingly punish smokers. Nonsmokers may be interested in the production of a 'private good' (being free of smoke) and not in establishing a general nonsmoking norm. Due to extensive sanctioning behavior, smokers will increasingly stop smoking and will get a bad conscience when they smoke in the presence of nonsmokers. A general nonsmoking norm has emerged. Various other mechanisms of the spontaneous emergence of norms are described in the literature. See particularly the literature on conventions (Schelling, 1960; Lewis, 1969; Sugden, 1995). Other examples of the spontaneous emergence of norms are explored by Mackie (1996), Opp (1982, 2002), and Horne (2001).

The instrumentality proposition assumes 'instrumental rationality' in the sense that individuals have goals and perceive that a norm x generates outcomes that attain their goals. Boudon (1997, pp. 19–22) argues that "many moral feelings are not the product of instrumental rationality" (p. 20). For example, the act of stealing cannot be shown to have negative consequences; stealing is bad because it contradicts basic moral principles. However, the instrumentality proposition does not rule out that the goal of a group or of individuals is to preserve general social principles (or values). The instrumentality proposition (and the theory of rational action in general) assumes 'instrumental rationality' or 'consequentialism' in a wide sense: a behavior (or a norm) may be instrumental in the sense that its aim is to influence phenomena outside the individual (such as to make other people better off) or to achieve goals related to internal states (such as following a norm). For details see Opp (2013).

Sociologists usually explain norms by the socialization proposition claiming that norms are transmitted (see, e.g., Boyd and Richerson, 1985). This may occur spontaneously (a norm is imitated) or in a planned way (parents socialize their children). One question that the socialization proposition does not address is under what conditions people transmit norms. Whatever these conditions are, sociologists seem to argue that it is not the overall instrumentality of a norm for a group that

is transmitted. This is suggested by the fact that norms may be dysfunctional (Merton, 1968, Chapter 3) and nonetheless spread. However, one would expect that parents who socialize children (or members of a group who socialize new members) are driven by incentives, that is, by their personal goals (Coleman, 1990, Chapter 11): parents teach their children good manners because they think that this is in the interest of the child or because they do not want to get a reputation of being bad parents who are not able to teach their children how to behave. Thus, parents aim at providing a private good ('good' children) that might lead to a norm on the macro level. The socialization proposition further does not specify how norms emerge before they are transmitted.

It might be argued that the previous propositions are inconsistent with a cultural approach such as symbolic interactionism. Hall and Taylor (1996) argue that a cultural approach does not reject that individuals are purposive, goal oriented, or rational (p. 949). But, they argue, the goals of individuals are conceptualized much broader: not only material well-being is at issue but individuals or organizations are also assumed "seeking to define and express their identity in socially appropriate ways" (p. 949). This critique holds for a narrow neoclassical version of rational actor theory but not for a wider version (Opp, 1999) that includes goals that go beyond material well-being. For example, preserving one's 'identity' may mean that members of a group subscribe to the goal of acting in accordance with important others. Furthermore, invoking 'world views' as conditions for norm emergence is also not against a rational choice approach because actors may realize that certain norms contradict values, and this may be a cost that discourages the acceptance of a norm.

The Effects of Norms

Assume that a norm has come into existence. What effects does the norm have on persons' actions? Role theory posits that norms are organized in roles that individuals take over. For example, the role of 'father' consists of a set of normative expectations about what a person with this role has to do and not to do. The message of the homo sociologicus is that individuals behave according to their roles (see, e.g., Biddle and Thomas, 1966). There are several problems with this proposition. One is that behavior that is not regulated cannot be explained. For example, there is no norm prescribing that one should not go to a cinema and instead watch television. A second problem is that the proposition is wrong – people often deviate from role expectations. 'Homo sociologicus' cannot explain such deviations. A third problem exists if there are conflicting role expectations: the theory does not tell us what expectation is heeded. The basic assumption of this theoretical orientation is that individuals are socialized into their roles and follow them blindly.

An alternative to this oversocialized conception of humans (Wrong, 1961) is to assume that in situations in which norms hold there is choice – as in any other situation. Individuals realize that they have at least two options: to follow or not to follow a norm. Each of these options has certain costs and benefits, depending on the situation. If norms are internalized, this means that breaking a norm ensues in a bad conscience, which is a cost. The severity and probability of external sanctions, including informal sanctions, are costly as well (see, e.g., Epstein, 1968; Hechter, 1984; Heckathorn, 1990; Posner and Rasmusen, 1999). But other nonnormative costs and benefits may instigate an individual to break a norm. What exactly the costs and benefits are that govern behavior if norms are given depends on the explanatory problem and cannot be specified in advance. For example, when people consider divorce, marriage norms play a more important role in Catholic than in Protestant countries. In committing murder norms may be more important than in tax evasions. In both situations, the kinds of nonnormative costs are very different. Thus, in explaining the effects of norms one has to include other nonnormative incentives and recognize that often norm following is a decision situation. This argument is based on the 'homo oeconomicus.' It suggests that homo sociologicus does not contradict homo oeconomicus.

To be sure, an individual has always a choice to follow a norm or break it. But often, the choice is made automatically, i.e., without thinking about the available options in a specific situation. To illustrate, let a person move to a new apartment and consider the possibilities of reaching the workplace. After weighing several alternatives, the person decides to take the subway, for the time being. It often happens that at some time a person makes a decision to act in a certain way and then performs the same action without deliberating. The action then becomes a habit – until there are certain changes in the situation (e.g., the subway schedule becomes unreliable). In regard to norms, it also happens that deviating from a norm is so costly that this action is never considered. This type of situation indeed resembles the situation that role theory seems to portray. But even if a behavior is normatively regulated, nonnormative incentives may be relatively strong so that the available options are weighed. When the effects of norms are considered, it is useful to distinguish between two types of effects. One concerns the direct effect of norms: compliance. If a norm (and the pertinent sanctions) has emerged, more people perform the behavior that is in accordance with the norm than before. If, for example, a government prohibits civil servants to smoke at their workplaces, smoking at the workplace will decrease.

This example illustrates a second type of indirect effect of norms. If fewer civil servants smoke, this may instigate others to stop smoking as well. This will reduce the number of people who get cancer and, thus, expenditures of health insurance companies will decrease. Indirect effects are often subtle and difficult to discover. For example, if companies have to pay increasing benefits to employed women (such as time off before and after pregnancy that the employer has to pay), the costs of employing women increase. The effect may be that in the long run unemployment of women increases and salaries decrease. This would certainly not have been intended by those who made the law. Increasing punishment for some crimes may amount to a higher relative benefit of committing other crimes and thus may lead to a displacement effect. In general, changing norms may change the costs or benefits of various types of actions and may thus have unanticipated and undesired indirect effects. What the indirect effects of norms are

cannot be said in advance. A detailed analysis of the incentives that prevail in the situation for which the effects of norms are to be explained is necessary.

Bibliography

Arthur, B.W., 1994. Increasing Returns and Path Dependence in the Economy. University of Michigan Press, Ann Arbor.

Axelrod, R., 1986. An evolutionary approach to norms. American Political Science Review 80, 1095–1111.

Bicchieri, C., Muldoon, R., 2011. Social Norms. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy (Spring Edition), http://plato.stanford.edu/archives/spr2011/entries/social-norms/.

Biddle, B.J., Thomas, E.J. (Eds.), 1966. Role Theory: Concepts and Research. Wiley, New York.

Boudon, R., 1997. The present relevance of Max Weber's Wertrationalität. In: Koslowski, P. (Ed.), Methodology of the Social Sciences, Ethics, and Economics in the Newer Historical School from Max Weber and Rickert to Sombart and Rothacker. Springer, Berlin.

Boyd, R., Richerson, P.J., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago.

Calvert, R.L., 1995. The rational choice theory of social institutions: cooperation, coordination, and communication. In: Banks, J.S., Hanushek, E.A. (Eds.), Modern Political Economy. Old Topics, New Directions. Cambridge University Press, Cambridge, UK.

Coleman, J.S., 1990. Foundations of Social Theory. Belknap Press of Harvard University Press, Cambridge, MA and London.

Crawford, S.E.S., Ostrom, E., 1995. A grammar of institutions. American Political Science Review 89, 582–600.

David, P., 1985. Clio and the economics of QWERTY. American Economic Review 75, 332–337.

Demsetz, H., 1967. Toward a theory of property rights. American Economic Review 57, 347–359.

Eggertsson, T., 1990. Economic Behavior and Institutions. Cambridge University Press, Cambridge, UK.

Elias, N., 1994. The Civilizing Process: The History of Manners and State Formation and Civilization. Blackwell, Cambridge, MA.

Ellickson, R.C., 1989. A hypothesis of wealth-maximizing norms: evidence from the whaling industry. Journal of Law, Economics, and Organization 5, 83–97.

Ellickson, R.C., 1991. Order without Law. How Neighbors Settle Disputes. Harvard University Press, Cambridge.

Ensminger, J., Knight, J., 1997. Changing social norms: common property, bride-wealth, and clan exogamy. Current Anthropology 38, 1–24.

Epstein, A.L., 1968. Sanctions. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences. Macmillan, New York.

Fine, G.A., 2001. Framing norms: mushrooming and the culture of expectations and explanations. In: Hechter, M., Opp, K.-D. (Eds.), Social Norms. Russell Sage Foundation, New York.

Gibbs, J.P., 1965. Norms: the problem of definition and classification. American Journal of Sociology 70, 586–594.

Hall, P.A., Taylor, R.C.R., 1996. Political science and the three new institutionalisms. Political Studies 44, 936–957.

Hardin, R., 1980. The emergence of norms. Ethics 90, 575–587.

Hechter, M., 1984. When actors comply: monitoring costs and the production of social order. Acta Sociologica 27, 161–183.

Hechter, M., 1987. Principles of Group Solidarity. University of California Press, Berkeley, CA.

Hechter, M., Opp, K-D., 2001. What have we learned about the emergence of norms? In: Hechter, M., Opp, K.-D. (Eds.), Social Norms. Russell Sage Foundation, New York.

Heckathorn, D.D., 1988. Collective sanctions and the creation of prisoner's dilemma norms. American Journal of Sociology 94, 535–562.

Heckathorn, D.D., 1990. Collective sanctions and compliance norms: a formal theory of group-mediated social control. American Sociological Review 55, 366–384.

Homans, G.C., 1974. Social Behavior. Its Elementary Forms. Harcourt, Brace & World, New York.

Horne, C., 2001. Sociological perspectives on the emergence of norms. In: Hechter, M., Opp, K.-D. (Eds.), Social Norms. Russell Sage Foundation, New York City.

Horne, Ch, 2009. The Rewards of Punishment. A Relational Theory of Norm Enforcement. Stanford University Press, Stanford.

Jackson, J.M., 1966. Structural characteristics of norms. In: Biddle, B.J., Thomas, E.J. (Eds.), Role Theory. Concepts and Research. Wiley, New York.

Jasso, G., Opp, K.-D., 1997. Probing the character of norms: a factorial survey analysis of the norms of political action. American Sociological Review 62, 947–964.

Kitts, J.A., 2006. Collective action, rival incentives and the emergence of antisocial norms. American Sociological Review 71, 235–259.

Kitts, J.A., 2008. Dynamics and stability of collective action norms. Journal of Mathematical Sociology 32, 142–163.

Lewin, P., 2002. The Economics of Qwerty. History, Theory and Policy. Essays by Stan J. Liebowitz and Stephen E. Margolis. Palgrave, Houndmills.

Lewis, D., 1969. Convention. A Philosophical Study. Harvard University Press, Cambridge, MA.

Mackie, G., 1996. Ending foot-binding and infibulation: a convention account. American Sociological Review 61, 999–1017.

Mahoney, J., 2000. Path dependence in historical sociology. Theory and Society 29, 507–548.

Merton, R.K., 1968. Social Theory and Social Structure. Free Press, New York.

Morris, R.T., 1956. A typology of norms. American Sociological Review 21, 610–613.

Oliver, P., 1980. Rewards and punishments as selective incentives for collective action. American Journal of Sociology 85, 1356–1375.

Opp, K.-D., 1982. The evolutionary emergence of norms. British Journal of Social Psychology 21, 139–149.

Opp, K.-D., 1999. Contending conceptions of the theory of rational action. Journal of Theoretical Politics 11, 171–202.

Opp, K.-D., 2001. How do norms emerge? An outline of a theory. Mind and Society 2, 101–128.

Opp, K.-D., 2002. When do norms emerge by human design and when by the unintended consequences of human action? The example of the no-smoking norm. Rationality and Society 14, 131–158.

Opp, K.-D., 2013. Norms and Rationality. Is Moral Behavior a Form of Rational Action? Theory & Decision 74, 383–410.

Opp, K.-D., Gern, C., 1993. Dissident groups, personal networks, and spontaneous cooperation: the East German revolution of 1989. American Sociological Review 58, 659–680.

Ostrom, E., 2005. Understanding Institutional Diversity. Princeton University Press, Princeton.

Posner, E.A., 2000. Law and Social Norms. Harvard University Press, Cambridge, Mass.

Posner, R.A., Rasmusen, E.B., 1999. Creating and enforcing norms, with special reference to sanctions. International Review of Law and Economics 19, 369–382.

Rommetveit, R., 1953. Social Norms and Roles. Universitetsforlaget, Oslo, Norway.

Rutherford, M., 1996. Institutions in Economics. The Old and the New Institutionalism. Cambridge University Press, Cambridge, UK.

Schelling, T.C., 1960. The Strategy of Conflict. Harvard University Press, Cambridge, MA.

Sugden, R., 1995. The coexistence of conventions. Journal of Economic Behavior and Organizations 28, 241–256.

Ullmann-Margalit, E., 1977. The Emergence of Norms. Clarendon Press, Oxford, UK.

Williams, R.M., 1968. The concept of norms. In: Sills, D. (Ed.), International Encyclopedia of the Social Sciences, vol. 2. Macmillan, New York.

Wrong, D.H., 1961. The oversocialized conception of man in modern sociology. American Journal of Sociology 26, 183–193.

Norms in Science

Peter Weingart, University of Bielefeld, Bielefeld, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

It is generally recognized that communication among scientists and scholars is governed by a particular scientific ethos, that is, by a set of rules that are supposed to establish trust in, and guarantee the reliability of, the knowledge created in the process. This ethos, although rarely made explicit in science, was given its most succinct and influential formulation by the American sociologist of science Robert K. Merton who defined it in terms of four basic norms and with that laid the groundwork for the *Sociology of Science*. This article presents the norms, their significance in scientific and scholarly communication, and some criticisms challenging their actual existence/validity motivated either theoretically or by factors related to the changing nature of science.

Introduction

It is generally recognized that communication among scientists and scholars is governed by a particular scientific ethos, that is, by a set of rules that are supposed to establish trust in, and guarantee the reliability of, the knowledge created in the process. This ethos, although rarely made explicit in science, was given its most succinct and influential formulation by the American sociologist of science Robert K. Merton (1957 [1942]), who defined it in terms of four basic norms, and with that laid the groundwork for the *Sociology of Science*. This article presents the norms, their significance in scientific and scholarly communication, and some criticisms challenging their actual existence.

The Norms of Scientific Communication

The scientific ethos refers to those patterns of behavior and implicit norms among scientists that can be traced back in history to the establishment of the academies in England and France in the seventeenth century. Regardless of many changes in detail there is a remarkable continuity in this ethos over a period of more than three centuries. Merton's formulation of the norms of science can be understood as an analytical condensation of the behavioral patterns that evolved over the duration of this period into a set of institutional imperatives or norms. The ethos, in his words,

> 'is that effectively toned complex of values and norms which is held to be binding on the man of science' and which, although not codified, can be inferred from, above all, the 'moral indignation directed toward contravention of the ethos.' (Merton, 1957: pp. 551–552)

The four key norms, according to Merton, were the following:

1. Universalism is the principle that truth claims are 'subjected to preestablished impersonal criteria' (Merton, 1957: p. 553) irrespective of the social attributes of their protagonists, for example, nationality, race, class, or religion.
2. Communism (later and preferably termed 'communalism') refers to the norm that the findings of science 'are a product of social collaboration and are assigned to the community' (Merton, 1957: p. 556). Property rights are kept to a minimum, and the scientist's claim to intellectual property may only be exchanged for recognition and esteem by the members of his community.
3. Disinterestedness demands that scientists should resist the temptation of using improper means (such as fraudulent manipulation of data) in the search for and communication of 'true' knowledge to their own advantage. It is a fundamental institutional element that has its basis in the public and testable nature of science, and it contributes to the integrity of scientists through accountability to their peers.
4. Organized skepticism is 'both a methodological and an institutional mandate.' It refers to the scrutiny of claims and beliefs by scientists' peers on the basis of empirical and logical criteria (Merton, 1957: p. 560).

These norms or values are interrelated and together shape a system of communication that is uniquely geared to produce knowledge that may be considered 'true' in the sense of being reliable, but by no means final. Universalism guarantees, at least formally, general social accessibility and at the same time it prevents the intrusion of any other criteria (political, religious, ethnic, etc.) in the communication of knowledge other than the certified knowledge accepted as belonging to science itself. Communalism subjects all knowledge to general and open communication, thereby subordinating the proprietary interests of the scientist to the less tangible rewards of gaining recognition by obtaining priority of discovery. It contributes to the social cohesion of the scientific community. Disinterestedness constitutes the self-reference of scientific communication. This means that science differs from other professions in that it has no clientele. The clients of science are scientists themselves; since all are engaged in a communally beneficial search for truth, and fraud or quackery may lead to temporary advantages at best. Organized skepticism is the reverse side of the norm of disinterestedness, as it stipulates the impersonal scrutiny of any claim to truth as a general principle of scientific communication. It is institutionalized in the peer review system of journals and funding agencies.

Merton's list of the norms of science was one of several descriptions of the 'scientific ethos' formulated during the late 1930s and early 1940s in reaction to the threats against science and democracy by fascist states in Europe (Hollinger, 1983). Merton's own account was written, originally in 1938, in reaction to the direct political interventions in science in National Socialist Germany (Merton, 1938; Mendelsohn, 1989). Later treatments of the scientific ethos have largely ignored this peculiar historical (and ideological) context of its formulation and have focused on its sociological relevance. Bernard Barber, working in the structural-functionalist tradition of Talcott Parsons and Robert Merton, added three values to the list of norms: faith in the virtue of rationality, 'emotional neutrality as an instrumental condition for the achievement of rationality,' and 'individualism' (Barber, 1953: pp. 86, 88). However, partly because of their overlap with the Mertonian norms and partly because of their different theoretical status, they never attracted the same attention.

The Theoretical Status and Significance of the Norms in Sociology of Science

In order to appreciate the significance of the norms one has first to realize that science in its modern form, as an institution that has the goal of the 'extension of certified knowledge' (Merton, 1957: p. 552), is an unlikely social invention. The improbability lies in the fact that science emerged from a historical situation in which the production of knowledge was immediately linked to the glorification of God or to the pursuit of economic utility. The unprecedented progress of science began when it became a 'self-validating enterprise' (Merton, 1957: p. 551), that is, a separate functional subsystem of society (Luhmann, 1990). On the institutional level, the remarkable social innovation was the separation of social status, religious affiliation, and, above all, prestige from the communication of truth claims. The rules and principles, which have been poignantly dubbed 'technologies of distance' (Porter, 1992), were instrumental in creating trust in scientific communication. They were gradually extended from individuals to institutions, allowing for a communication network that reached beyond generational limits in time, as well as beyond the social and geographical limits of personal acquaintances. Historically, the process of the formation of modern science spans from the seventeenth century to the early nineteenth century when the university emerged as the core institution and disciplines formed marking the establishment of self-validation and autonomy (Stichweh, 1984). Merton's norms codified the institutional patterns that emerged during this period into an analytical scheme that served to explain the unique status of science – as a set of methods to produce certified knowledge and to accumulate the knowledge that accrues from their application, as well as a set of cultural values and customs that governs these activities.

Thus, with the articulation of the norms, Merton also defined an aim for sociology of science: to be concerned primarily with the moral compulsion of methods and their impact on behavior, not with their technical expediency. Merton, in effect, conceived of the sociology of science as an analysis of the social prerequisites and institutional configurations of science. Sociologists would in no way be concerned with the substantive findings of science or their methodological underpinnings. Consequently his sociology never challenged the philosophical assumptions that shaped the concept of science at any particular historical moment, and which have changed over time; a limitation for which his views were subsequently criticized. Under the influence of Thomas Kuhn (1962), sociologists began to look into the normative nature of scientific paradigms and thereby crossed the line into a full-fledged 'sociology of scientific knowledge.' The theoretical status of the scientific ethos has been a persistent issue. It often has been argued that scientists do not behave in accordance with these norms and that the norms themselves are subject to historical change (Barnes and Dolby, 1970). Somewhat related to this is the critique that the norms are an edifying portrait of science that functions less as a description of actual behavior than as a professional ideology that has supported the post-World War II rise and establishment of 'basic science' (Mulkay, 1976). However, the existence of norms is not necessarily reflected in actual behavior. Merton's theoretical construction of the norms is a complex combination of different elements: (1) sociopsychological patterns of attitudes that are expressed in internalized (but not necessarily explicit) reactions to violations of the norms, in the awareness that one's own actions or those of others are breaching a code; and (2) sociostructural patterns of sanctions, that is, mechanisms institutionalized in science that sanction, positively or negatively, certain behaviors like plagiarism (negatively) or the open exchange of information (positively). Thus, to demonstrate whether the norms are operating or not, requires a more sophisticated approach than mere observation of behavior.

Merton himself responded to the challenges to the validity of the norms by pursuing two research strategies: (1) the analysis of apparent paradoxes in scientists' behavior and the resulting psychological and social conflicts (Merton, 1973b); and (2) the study of typical reasons for attacks against science or for conflicts between science and its societal environment (Merton, 1957 [1942]). On the one hand, Merton observed a peculiar ambivalence among scientists toward priority disputes and their own interest in priority of discovery. (His empirical evidence consists of biographical materials and correspondence of scientists dealing with priority disputes.) He attributed this ambivalence to the institutional mandate of disinterestedness and the primacy of the expansion of knowledge. Here again, sociological elements are coupled with sociopsychological ones. The need for recognition felt by the individual scientist is not an expression of idiosyncratic vanity but is initiated by the institution. 'Continued appraisal of work and recognition of work that was well done constitute one mechanism that unites the world of science' (Merton, 1973b: p. 401). Merton reacted to the changes in sociopsychological patterns affected by institutional changes. Observing the increasing number of multiauthored journal articles that indicate the rising importance of cooperative forms of research, he concluded that the 'interest in priority' – by no means historically invariant – may give way to an interest in the recognizability of one's work in a research team (Merton, 1973b: p. 409).

Merton's second approach to documenting the validity of the scientific ethos was to analyze conflicts between science and its societal environment. He observed two sources of hostility toward science: the conflict between methods or important results of science with important social values, and the conflict between the sentiments embodied in the scientific ethos and those of other institutions (Merton, 1957: p. 537). The most obvious conflict of the latter type arises in totalitarian political orders – the kind Merton had in mind at the time of his original writing. This was a conflict between the scientific ethos and the political code that routinely discredited science as 'cosmopolitan,' 'liberalistic,' or 'bourgeois.' (In his 1938 and 1942 essays Merton focused on Nazi Germany, after the war he included the Stalinist regime in the Soviet Union epitomized by the Lysenko affair.) A more complex effect may be seen to accrue from utilitarian demands on science. Originally in opposition to the demands of the church, the state, or the economy, the orientation to 'pure' science may undermine the social esteem of science. Although this purist orientation may be methodologically justified, as it minimizes bias and error by separating the production of knowledge from its application, its social effect may be resentment and revolt against the possible negative consequences of science (Merton, 1957: pp. 543–545). This diagnosis has proven to be farsighted. Since the 1960s, public resistance to and critique of technological projects have steadily increased and the social contract for 'pure science' has come under pressure.

Merton's formulation of the scientific ethos as a description of the special nature of science as an institution has been fruitful in two ways. On the one hand, it focused on the effects of structural changes in science, on the institutional mechanisms that guarantee the production of certified knowledge through channeling scientists' behavior. On the other hand, it focused on the conflicts that emerge between science and other institutions. The question then is what becomes of the scientific ethos as science and its institutional context change over time?

The Scientific Ethos and the Changing Nature of Science

Critics of Merton have pointed to the many changes science has undergone that are not accounted for in an ethos that is basically geared to an individualist concept of gentlemanly science. Merton himself addressed these changes and granted some plausibility to the view that the customs of scientists must have changed in view of the fact that all the "basic demographic, social, economic, political, and organizational parameters of science have acquired dramatically new values" (Merton, 1973a: pp. 328–329). He enumerated certain typical phenomena such as the premature publication of results before a theory is corroborated by more careful competitors, and an intensified rivalry occasioned by specialization as well as a new kind of cynicism, amorality, and disillusionment that goes along with it. However, in Merton's view, scientists' interest in priority, as such, has not changed. Rather than indicating a change of the ethos, public reactions to apparent or real breaches of the ethos reveal an 'instructive paradox,' namely, that the 'customs governing the public demeanor of scientists and the public evaluation of contemporaries have become more exacting rather than less' (Merton, 1973a: p. 338). Merton's observation is supported by heightened concerns in governments and the media about spectacular cases of fraud in science since at least the 1980s. Although it is by no means a proven fact that the incidence of fraudulent behavior among scientists has increased the structural pressures on scientists, especially in highly dynamic and capital-intensive research fields (e.g., biomedicine) and the concentration of reported cases in them suggest that it has increased. As a result, public awareness and vigilance have intensified. In several countries, institutions have been set up to monitor and sanction 'misconduct' in science (e.g., the Office of Research Integrity in the United States), and it is noteworthy that the 'codes of conduct' underlying their operation either explicitly refer to Merton's norms or are highly consistent with them.

Some theorists have suspected that the ethos of science is threatened by fundamental changes in social values. Yaron Ezrahi sees a challenge to the scientific ethos and to liberal democratic values by the counterculture that emerged from the radicalism of the 1960s, which favored values such as personality, authenticity, and creativity without standards. The belief that science is an enterprise governed by universalistic norms that produces objective knowledge by abstracting from its bearers is replaced, in his view, by an individualization of the perception of reality and the production of knowledge (Ezrahi, 1990: p. 54). In a similar vein, Paul Forman focuses on the effects of postmodernist values on science. He observes three main changes in the production of knowledge as causing the spread of these values: overproduction of knowledge and a resulting plurality of standards; instrumentalization of knowledge effecting an ethos of production as an end in itself; and an affirmation of knowledge as 'bound' and 'interested' resulting from an alleged incommensurability of different stocks of knowledge (Forman, 1997).

Another threat to the scientific ethos is seen to emerge from the 'commercialization' of science, mainly in the form of patenting and the imposition of licenses and other forms of intellectual property rights (IPR), together with their impact on the 'culture' of open academic science. It is plausible to assume – and in some cases has been shown to occur – that an undue orientation among scientists to monetary profit potentially violates the norms of communality (by way of delaying or interrupting the communication of research results through IPRs), organized skepticism, and disinterestedness (e.g., as a result of monetary conflicts of interest in review processes) (Mirowski and Sent, 2007). The underlying conflict is seen as occurring between 'public' and 'private' science.

Whether or not these diagnoses are correct will ultimately be verified by observable changes in the scientific ethos and by the emergence of radically different patterns of the production of certified knowledge. Merton's formulation of the scientific ethos has provided invaluable groundwork for an understanding of the interaction between normative, psychological, and institutional mechanisms in the production and communication of scientific knowledge.

See also: Afrocentric Approaches to Social Work; Conventions and Norms: Philosophical Aspects; Ethical Codes, Professional: Business Codes; Genomics, Ethical Issues In; Human Sciences,

History of; Norms; Research Ethics: Research; Research Funding: Ethical Aspects; Science, Sociology of; Scientific Knowledge, Sociology of; Technology, Social Construction of; Value Pluralism.

Bibliography

Barber, B., 1953. Science and the Social Order. George Allen and Unwin, London.

Barnes, S.B., Dolby, R.G.A., 1970. The scientific ethos: a deviant viewpoint. Archives Européennes de Sociologie 11, 3–25.

Ezrahi, Y., 1990. The Descent of Icarus: Science and the Transformation of Contemporary Society. Harvard University Press, Cambridge, MA.

Forman, P., 1997. Recent science: late-modern and post-modern. In: Soderquist, T. (Ed.), The Historiography of Contemporary Science and Technology. Harwood Academic Publishers, Amsterdam, pp. 179–214.

Hollinger, D.A., 1983. The defense of democracy and Robert K. Merton's formulation of the scientific ethos. Knowledge and Society: Studies in the Sociology of Culture Past and Present 4, 1–15.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. The University of Chicago Press, Chicago.

Luhmann, N., 1990. Die Wissenschaft der Gesellschaft. Suhrkamp, Frankfurt, Germany.

Mendelsohn, E., 1989. Robert K. Merton: the celebration and defense of science. Science in Context 3 (1), 269–289.

Merton, R.K., 1938. Science and the social order. Philosophy of Science 5, 321–337.

Merton, R.K., 1957. Science and democratic social structure. In: Merton, R.K. (Ed.), Social Theory and Social Structure, revised and expanded ed. The Free Press, Glencoe, IL (Originally published as: Merton, R.K., 1942. A note on science and democracy. Journal of Legal and Political Sociology 1, 115–126; reprinted as Merton, R.K., 1973. The normative structure of Science. In: Storer, N.W. (Ed.), The Sociology of Science. The University of Chicago Press, Chicago, pp. 267–278).

Merton, R.K., 1973a. Behavior patterns of scientists. In: Storer, N.W. (Ed.), The Sociology of Science: Theoretical and Empirical Investigations. The University of Chicago Press, Chicago, pp. 325–342.

Merton, R.K., 1973b. The ambivalence of scientists. In: Storer, N.W. (Ed.), The Sociology of Science: Theoretical and Empirical Investigations. The University of Chicago Press, Chicago, pp. 383–414.

Mirowski, P., Sent, E.-M., 2007. The commercialization of science and the response of STS. In: Hackett, E., Amsterdamska, O., Lynch, M., Wajcman, J. (Eds.), The Handbook of Science and Technology Studies. Sage, Thousand Oaks, CA and London, pp. 635–689.

Mulkay, M., 1976. Norms and ideology in science. Social Science Information 15, 637–656.

Porter, T., 1992. Quantification and the accounting ideal in science. Social Studies of Science 22, 633–652.

Stichweh, R., 1984. Zur Entstehung des modernen Systems wissenschaftlicher Disziplinen. Physik in Deutschland 1740–1890. Suhrkamp, Frankfurt, Germany.

North America, Archaeology of

Asa R Randall, University of Oklahoma, Norman, OK, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The archaeology of North America includes at least 16 000 years of occupation. The archaeological record is the product of communities of diverse cultural practices and worldviews who negotiated social and ecological environments. Although it contains evidence for many social and technical strategies, the archaeology of North America has much to contribute to contemporary and future concerns with climate change and the reconstitution of communities.

North America has an unbroken record of human occupation dating back at least 16 000 years. North American archaeology explores the ecological, political, and social dimensions of human experience recorded in material traces of past activities. Encapsulated in North America's history of occupation are processes of migration and interaction among culturally diverse communities. Many regions witnessed the emergence of complex societies whose political and economic institutions were materialized in objects and constructed landscapes. The development of traditional ecological knowledge and landscape management practices enabled communities to be resilient to environmental change, but also made some communities vulnerable to climatic or social perturbations. Finally, North American archaeology today is concerned with cultural heritage and the significance of the past, today, and in ancient times.

The archaeological record of North America is composed of the various objects, structures, and landscapes that humans created or interacted with. The most frequently occurring objects are tools made of stone and the debris generated during their manufacture and use. Vessels manufactured from ceramic, as well as items made of stone, marine shell, basketry, or textile were also produced in some regions. Many items of personal adornment are also known. Where conditions allow, the archaeological record contains objects manufactured from organic media (bone, wood), in addition to the remains of meals. The distribution of these various materials reflects accommodations of communities to particular social and ecological environments, as well as trade and interaction between groups. Archaeological sites in North America vary from the small scale of object scatters up to the large scale of ancient and modern cities, such as the thousand-year-old site of Cahokia near modern-day St. Louis, Missouri. Places of social poignancy and reverence are widespread through time and across space. Places of social significance include mortuaries, earthen or shell monuments, caves and rock shelters, and petroglyph panels. Included here as well should be 'natural places' that remain mostly unmodified but which were nonetheless significant to ancient communities and their descendants (Zedeño et al., 2009). A wide spectrum of societies has been documented in the region, whether one categorizes them by subsistence economies (hunter–gatherers, cultivators, and intensive agriculturalists), political economies (bands, tribes, chiefdoms, states, and empires), or historical processes (communities that are coalescent, diasporic, resistant, or colonial).

The Study of Archaeology in North America

The study of North America's past is structured by the interaction and power relationships between academic institutions, governmental agencies, legislation, descendant communities, landowners, and other stakeholders. Archaeology – broadly defined as the study of cultural variation through material remains – is a relatively recent social science which was formalized only in the late nineteenth century (Trigger, 2006). As an organized academic discipline in North America, archaeology is most frequently considered a subfield of anthropology. In this respect, archaeology is closely affiliated with other branches of anthropology including linguistics, sociocultural anthropology, and biological anthropology. The integration of these subdisciplines is particularly strong in the United States due to institutional and governmental interactions with indigenous communities dating back to the nineteenth century. Academic archaeologists who specialize in classical civilization, medieval history, or Middle Eastern studies are often housed in other disciplines. The vast majority of North American archaeologists are not employed by academic units. In Canada and the United States, legislation has been enacted to preserve cultural resources. Although varying in particulars, scale, and applicability, these laws mandate that the adverse effects of ground-disturbing activities on lands owned or managed by national, provincial, local municipal, and tribal organizations should be mitigated (Neumann et al., 2010). Legislative tools for preservation fostered the development of a field of archaeology in the 1970s referred to as cultural resource management. Where required by law, private firms and public agencies are contracted to conduct surveys and excavations to determine the significance of cultural resources prior to development and conduct full-scale data recovery when significant cultural resources will be otherwise destroyed.

Although closely aligned with anthropology, what tends to differentiate archaeology is a methodological emphasis on the variation of human societies across space and through time. The archaeology of North America is traditionally parsed out into culture areas which correspond to major physiographic divisions. These include the Arctic and sub-Arctic in Canada and Alaska; the northeast in New England; the Mid-Atlantic of the eastern seaboard; the Southeast associated with the Ohio River Valley, the southern Coastal Plain and Florida; the Midwest along the upper and middle Mississippi River valley; the Plains from the state of Texas to the Canadian provinces of

Alberta, Manitoba, and Saskatchewan; the Rocky Mountains, the Southwest from the four corners to the Sonora desert; the Great Basin; Southern California; and the Northwest coast extending from Oregon into coastal Alaska. The particular boundaries of these regions will vary depending on the analyst and the criteria used to delineate them. The logic for separating these regions was first espoused by Bureau of American Ethnology archaeologist William H. Holmes in 1914, who noted that, at the time of sustained European contact, living cultures that were coincident with major ecological and climate zones (Holmes, 1914). These boundaries continue to be used as heuristic and organizing devices. However, they can be criticized for representing a snapshot of cultural variation that likely obscures the complex interactions and movements of communities across these arbitrary boundaries over the past 16 000 years.

With respect to time, a fundamental distinction is traditionally made between historic and prehistoric archaeology. In this system, historic archaeology investigates social process with the aid of written documentary sources. There are no known pre-European contact systems of writing documented in North America, and thus many would restrict historic archaeology to the past 500 or so years, spanning early European colonialism and later national expansion. With regards to prehistoric archaeology of unlettered cultures, a sequence of chronological periods is recognized in many regions based on changes in subsistence, settlement organization, social organization, and technological innovations. The Paleoindian period is associated with the earliest inhabitants of North America during the Pleistocene geological epoch (Walker and Driskell, 2007). This period witnessed the initial colonization of the continent and the establishment of hunter–gatherer economies that brought communities into close contact with megafauna. Although well represented in the interior, the Paleoindian coastal record is poorly understood due to subsequent sea-level rise. The succeeding Archaic period corresponds to the demise of most megafauna at the onset of the Holocene geological epoch. The Archaic was characterized by the development of increasingly diverse hunter–gatherer traditions. It was during the Archaic that many social, political, and technological institutions that had currency later in time were developed (Sassaman, 2004). These included the invention of pottery, the domestication of endemic plants or adoption of tropical cultigens, the construction of mortuary monuments, investment in land management through the construction of permanent facilities, and the differentiation of communities and perhaps individuals based on socioeconomic status.

Archaeologists working in some regions of North America recognize a Woodland period, characterized by an increased emphasis on ritual and ceremony, and in many locations a shift to a mixed subsistence strategy of foraging and cultivation (Anderson and Mainfort, 2002). The Late Prehistoric era corresponds to the establishment of politically complex societies organized around either chiefly elites who capitalized on the labor of others, or around federations of communities (Bayman, 2001; Blitz, 2010; Bamann et al., 1992; Moss, 2011). Attendant to political shifts in some regions were agricultural intensification of tropical cultigens such as maize, beans, and squash, and the investment in large public works. Within each region of study, further chronological subdivisions are recognized based on shared material culture, subsistence, and political organization.

Chronological boundaries for each period will vary based on local processes. Ultimately these periods are used by archaeologists to consolidate research questions. Yet these divisions can also be criticized for essentializing diverse lifeways using select criteria, or for promoting a unilineal model of social change from simple to complex (Alt, 2010). Further, a close inspection reveals that whereas agriculture was adopted by certain communities in North America, hunting and gathering persisted in many regions (notably the higher latitudes, the Plains, and the Pacific, Atlantic, and Gulf of Mexico coasts), often alongside agricultural communities. Using temporal divisions to categorize societies, it is possible to have 'Archaic' societies coexisting alongside 'Woodland' societies, for example Desert Archaic hunter–gatherers and Freemont farmers in the Great Basin (1400–600 years ago), or hunter–gatherers and agriculturalists inhabiting northern Florida (AD 1000).

Like other social sciences, archaeological explanation in North America has vacillated between historical or particularizing frameworks on one hand, and generalizing or evolutionary perspectives on the other (Trigger, 2006). The shift between either pole was frequently informed by empirical contradictions and broader social concerns. During the eighteenth and nineteenth centuries, inquiry was antiquarian in principle: explanation took the form of speculation or interest in objects for their own sake. The defining controversy that organized interest in North American archaeology at this time was the so-called 'myth of the mound builders.' When European colonists and their descendants expanded throughout the Eastern Woodlands they encountered landscapes saturated with earthen and shell mounds that often contained evidence for elaborate mortuary ritual. These works were seen in an ethnocentric light as evidence for 'advanced civilizations' who possessed economic, political, and cultural capabilities exceeding those of indigenous Native Americans and their ancestors (Squier and Davis, 1848). The controversy was settled in the late nineteenth century by extensive ethnohistoric research and archaeological excavations that linked contemporary indigenous communities with the archaeological record. This research was sponsored by the United States federal government and did much to formalize the asymmetrical relationships between descendant communities, government, the archaeological record, and academic institutions (Thomas, 1894). At this same time, generalizing evolutionary schemes were developed by the likes of Herbert Spencer and Louis Henry Morgan that explained contemporary human variation as a unilineal sequence of development from cultural and moral simplicity to complexity. Descriptions of living communities were often used as analogues of ancient communities, often with racist and ethnocentric overtones (Lubbock, 1865).

Stadial schemes were mostly abandoned following developments in sociocultural anthropology led by Franz Boas at the Columbia University in the early twentieth century. North American archaeologists rejected unilineal evolution and instead investigated the distinct and particular sequences of change evidenced in the archaeological record. The culture-historical approach emphasized stratigraphic methods for recording archaeological finds. By closely documenting where

objects were found within the site, archaeologists could track changes through time (Kidder, 1924). Patterns of change at a local spatial scale could then be compared with regional finds, and organized into space–time taxonomies (Willey and Phillips, 1958). Culture-historians operated with a normative view of culture, such that objects and archaeological sites represented the ideals or norms of particular groups. Past archaeological cultures were defined principally by lists of traits (such lists would include the type of settlement pattern, subsistence, technological style, and mortuary tradition) which were thought to represent social norms. The result was a static and pessimistic view of change.

The assumptions of culture-historical, or particularist, archaeology were challenged by processualism (the 'new archaeology') in the middle of the twentieth century. Processualism was predicated on a particular view of the archaeological record: the distribution of objects and places reflected dynamic processes in which humans adapted to their social and ecological environments (Binford, 1962). Processual archaeologists adopted a logical-positivist epistemology that foregrounded objectivism, scientific method, and a systems theory approach (*see* Archaeology: Philosophy and Science). Explanation involved using the archaeological record to test hypotheses, often derived from ethnographic or modern analogs. Archaeologists used the distribution and function of archaeological resources and independent variables such as population density or ecological structure to test hypotheses (Watson et al., 1971). Underwriting processual thought was a model of social evolution in which societies became increasingly complex through time (e.g., Service, 1962). Archaeological cultures were often classified into types (such as bands, tribes, chiefdoms, and states) that were reflected in subsistence economy, demography, and political organization. Popular research concerned hunter–gatherer variation, the circumstances of plant domestication, and the origins of complex society. Lewis Binford, for example, examined the subsistence, settlement, and social patterns of historic and contemporary hunter–gatherers in North America and elsewhere (Binford, 1990). He argued that social complexity in ancient times may have emerged in situations where high population densities and spatially restricted resources promoted reduced mobility and technological innovations to efficiently capture food.

Beginning in the late 1970s, the processual approach was criticized for ignoring the ideology of power and authority, and for a lack of concern with meaning. The critique was applied equally to interpretations of the past and to relations between archaeologists and indigenous communities whose cultural resources were investigated by archaeologists. Post-processual approaches foregrounded the significance of genders, embodiment, cultural traditions, cosmology, and identity in structuring the archaeological record (Brumfiel, 1992). There has been a rapprochement with history. Archaeologists are increasingly concerned with how cultural traditions are created as a historical process in which different groups (such as classes, ethnicities, genders) negotiate systems of value. Research also investigates the ways in which ancient communities created their own histories through social memory (Pauketat, 2001). Tradition making can be investigated by examining the history of how people created objects, and the ways in which objects and landscape use changed in the context of social interaction at local and regional spatial scales. Independent of post-processualism, a community of archaeologists applies insights from evolutionary theory to understand ancient change. In particular, so-called evolutionary archaeologists have focused on Darwinian natural selection and its influence on changes in social organization and technology through time (Dunnell, 1989). In this school of thought, cultural practices (memes) are subject to Darwinian natural selection and replication. Another school of thought argues that anatomically modern humans have evolved to be behaviorally flexible and economically rational (Boone and Smith, 1998).

In the past quarter century, archaeology in North America has become increasingly self-reflexive. That means that many archaeologists are now aware of the colonial history of the discipline, and how the production of knowledge is not inherently objective. This recognition is largely due to Native American and First Nation civil rights movements. These actions resulted in eventual federal legislation (NAGPRA in the United States) that provided a mechanism for protecting indigenous graves and repatriating cultural patrimony (Preucel and Cipola, 2008). Some archaeologists practice public archaeology, which considers the practical consequences and ethical realities of incorporating stakeholder communities. There are indigenous archaeologies conducted by First Nations and Native American communities as well (Pyburn, 2011). In decolonizing archaeological interpretation, opportunities are made for self-representation of historically marginalized groups. New kinds of narratives relating communities to the past are now possible in a variety of media, and are supported by inclusive national museums in Canada and the United States (Colwell-Chanthaphonh et al., 2011).

Contemporary Research Themes

Research into North America's 16 000 years of history is influenced by academic, social, and pragmatic concerns. That said, contemporary research can be grouped according to themes that cross-cut particular chronological or spatial boundaries or theoretical divides. The themes considered here include: (1) population movements, encounters, and the creation of new identities; (2) documenting the circumstances attending the materialization of myth and history in the landscape; and (3) detailing human–environmental relations with an eye toward informing contemporary and future practice in an age of global climate change. These themes will be discussed with reference to particular examples.

Encounters and Movement

North American archaeology is defined by the movement of populations, encounters of culturally diverse communities, and subsequent social transformations. Among these events, it would be hard to overstate the impact of European colonialism. Indigenous populations were decimated by disease. Those persons who remained were faced with reconstituting communities by either being incorporated into, or resisting, a world system that promoted cash economies over traditional social relations. Equally significant was the inclusion of Africans, who were traumatically removed from their natal communities. Archaeologists working with written and

archaeological records have explored the dimensions of power and alienation through which material culture, places, and landscapes were reconstituted among plural communities. Importantly, these studies have exposed how multifaceted and varied colonial experiences were depending on the cultural dispositions and strategic intentions of colonial powers. So too, gender relations among diasporic communities and indigenous cultures influenced how new societies were created. For example, in Spanish *La Florida* in the southeastern United States, Timucuan, Guale, and other groups were valued by colonial powers in terms of the labor they could provide for agricultural production. Missions were established throughout the region to convert persons to Christianity and suppress resistance to European authority. The presence of Native American pottery within otherwise Spanish households, however, indicates that native women were incorporated into Euro-American communities (Deagan, 2003). This situation suggests that native women were an important influence in the process of ethnogenesis (in which new ethnic identities and cultural traditions emerged from Indigenous and Euro-American practices). Similar processes may have occurred among Russian trading outposts such as Fort Ross in northern California (Lightfoot et al., 1998). In the frontier of southern California, in contrast, the initial relations between Spanish men and indigenous persons were likely through forced conscription or wage labor. Not until imperial demands on morality were loosened during the eighteenth century would gender relations and ethnogenesis be materialized in domestic foodways and architecture (Voss, 2008).

Archaeologists working in the Southeast have, since the 1960s, explored the archaeology of enslaved, diasporic communities at Spanish, British, and American plantations (Singleton, 1995). Despite the suppression of African identity by many plantation owners, some enslaved persons managed to maintain African foodways and traditions, at least in secret, or otherwise exercised resistance to oppression (Leone et al., 2005; Ruppel et al., 2003). Detailed historic observations provide the framework for reconstructing in digital form the structure of plantations or other prominent landscapes (such as Thomas Jefferson's Monticello). These spatial models provide a basis for reflection, and often reveal the inherent contradictions between landowners, others, and nationalist narratives. Beyond the immediacy of direct contact, archaeologists and ethnohistorians are now beginning to understand the effects of colonialism at increasing spatial scales. Although slavery is frequently assumed to have been directed at those of African descent, slave-raiding of Native Americans communities was conducted by Native Americans along a frontier expanding outward from the zones of initial contact in the seventeenth and eighteenth centuries (Ethridge, 2006). The archaeological implications are only now being tested, but appear to be manifested in differing community size, material culture assemblages, and patterns of group dispersion.

While the archaeology of migration and displaced communities traditionally has been the purview of historic archaeology, more recent research recognizes that disruptive processes associated with the movements of persons also occurred in the pre-Columbian era (Cobb, 2005). This work represents an important correction by acknowledging that although the particular circumstances of the modern world system did not exist in North America prior to European contact, there was no inherent schism between modern and pre-modern people. For example, some of the more elaborate architectural and symbolic florescences in Late Prehistoric North America – Chaco Canyon in the northern Southwest (AD 1000–1150) and Cahokia in the Mississippi American Bottom in the Midwest (AD 1050–1150) – appear related to migration and the politics of placemaking. Expansive landscapes composed of civic buildings for ritual performances (often with sacred and celestial alignments) were rapidly constructed in both regions. In the case of Cahokia there was the elaboration of place-specific material symbols, such as Ramey Incised Pottery and chunkee gaming stones, that were distributed throughout Cahokia's sphere of influence and evoked the power and religious authority of that place (Pauketat and Emerson, 2008). At Cahokia and Chaco Canyon there is evidence for intercommunity conflict and environmental fluctuations coincident with the emergence of these important places. Such regional processes may have instigated the abandonment of some regions and the aggregation of culturally disparate communities in others. Historically, however, the production of each of these localities by diverse communities occurred under the direction (and likely promotion) of the few (e.g., Cameron and Duff, 2008). The elevation of central places was likely an attempt by emerging elites to attain political and ideological dominance. Elites used contrasting practices of sponsoring inclusionary public ritual on one hand, but excluding the majority of people from direct access to specific ritual knowledge or symbols on the other. The archaeology of migrant villages, in particular in the farming villages surrounding Cahokia, highlights how these communities did not accept Cahokian authority without challenge. New kinds of community organization emerged as migrants incorporated traditional and new values into existing lifeways. The collapse of both Chaco and Cahokia reverberated through their respective regions as power structures were renegotiated locally. Communities once again engaged in movement away from former power centers.

The example of North America's first-millennium societies highlights the political and symbolic aspects of migration in groups that were politically and economically ranked. However, migrations into and throughout North America occurred at larger spatial and temporal scales, among smaller-scale communities several times in the ancient past. Indeed, one of the defining questions in the study of North American archaeology is: who were the first Americans? Before 1924, it was assumed that Native Americans had entered the continent only a few thousand years prior to European contact, the so-called short chronology. All that changed with the discovery of a stone tool in association with a now-extinct species of bison at the Folsom site in New Mexico (Meltzer et al., 2002). In the decades since, a new model emerged based on the discovery of a characteristic projectile point found across much of North America south of the ice sheets, and often associated with mammoth and mastodon remains. Referred to as 'Clovis,' this archaeological culture was widely accepted as the first to inhabit North America. Recent discoveries of coeval, but distinctly non-Clovis lifeways, such as at the Paisley Caves in Oregon (Jenkins et al., 2012), or even chronologically earlier communities, have led to renewed discussion about who the

first Americans were, and how they got to North America (Meltzer, 2009). This discussion will no doubt continue. Regardless, non-Clovis or earlier traditions open up the real possibility of multiple migrations, by numerous pathways along the coasts and interiors. These multiple migrations imply that there were many moments of culture contact at the dawn of human occupation of the region. Archaeologists, often working with the aid of historical linguists, have identified at least two other pre-Columbian migrations into North America out of the western Subarctic and Arctic. These include the Athapascan migration, which resulted in the emplacement of ancestral Apache and Navajo groups in the Southwest during the fourteenth century AD, and the Thule migrations across the Subarctic, which resulted in the displacement of earlier Paleo-eskimo communities (Friesen, 2004; Seymour, 2012). The historical processes attendant to these migrations are poorly understood. In the case of the Subarctic, migrations appear to have involved competition and negotiation with between migrants and indigenous communities (Holly, 2005).

Materializing Social Complexities Through Ritual

One area in which the archaeology of North America is broadly relevant to global social scientific research is in documenting and problematizing diversity and apparent complexity of recent and ancient hunter–gatherer communities. Before the 1970s, it was thought that pre-Columbian hunter–gatherers were, by default, characterized by low population densities, high residential mobility, and limited differences in social status or rank. Because environmental processes were considered determinant of social variation, considerably little attention was paid to the ways in which ritual, religion, and social negotiations structured hunter–gatherer lifeways. This model contrasted with historic evidence in several portions of North America for large, semi-sedentary hunter–gatherer populations that engaged in widespread landscape modification and participated in elaborate exchange and ritual. Most notable are the communities of the Northwest Coast of the United States and Canada, the Ohlone of the San Francisco Bay, the Chumash of southern California, and the Calusa of southwest Florida (Arnold, 2001; Lightfoot and Parrish, 2009; Moss, 2011; Widmer, 1988). Although variable in customs and politics, one shared feature was subsistence economies grounded in naturally occurring abundant aquatic and terrestrial resources.

In the decades since these so-called complex hunter–gatherers were recognized historically, a number of apparently similar examples have been documented in antiquity. The majority lived in the Midwest and Midsouth where riverine or coastal resources were intensively exploited during the Archaic period. One materialization of these subsistence economies are mounds of earth or shell, in addition to wide-ranging exchange networks of raw materials and finished objects. The extent to which these traditions represent complex societies in an economic sense has been a source of significant debate (Gibson and Carr, 2004). Traditional models for social complexity, grounded in institutionalized status distinctions based on material wealth, have failed to explain archaeological patterning. The analysis of grave inclusions fails to discern status distinctions beyond gender or age. A further point of debate regards the social significance of mounds of earth and shell. Since the resolution of the moundbuilder debate, the construction of monuments has been coupled with other cultural traits such as agriculture or pottery production. It was once thought that North American communities lacked the labor, politics, and worldview to construct such places until the Woodland period. Archaic mounds are problematic in that they are chronologically early, and in many cases composed of apparently mundane materials. However, a consideration of the history of these places suggests that many were purposefully constructed as integrative ritual facilities that may have promoted the inclusion of diverse persons from disparate regions (Sassaman, 2010). For example, the Archaic freshwater shell mounds of the St. Johns River in Florida were constructed for a variety of purposes. Some were residential in nature, but others including the 6300-year-old Harris Creek site, were dedicated mortuary mounds that were constructed for communal or individual burial, and others were ceremonial mounds constructed during ritualized aggregations. When viewed in historical terms, the movement between and modification of shell sites generated a complex social landscape supported by riverine resource exploitation, but justified by ritual practice. Similar processes of ritual incorporation and situational leadership, but at a much larger scale, were likely behind the construction of North America's second largest mound center, Poverty Point in Louisiana (Ortmann and Kidder, 2013). In this case, elaborate earthworks were rapidly constructed. The largest, the 22-m high Mound A, may have been constructed in as few as 3 months by 12 000 or more individuals drawn from regional communities, perhaps from throughout the Midwest and Southeast.

Human–Environment Relations

North American archaeology is characterized by a deep-seated interest in the relations between the environment and human social and historical variation. This focus has a long history in prehistoric archaeological thought in general (dating back to at least the eighteenth century), but has been particularly prominent in North American theory and practice. Earlier approaches have been criticized for emphasizing the role of the environment in structuring the long-term histories and decisions of ancient North Americans. At best, such environmental approaches assume that the environment significantly influenced long-term trends in population density and social organization. At worst, they reflect a deep-seated pessimism regarding the capabilities of small-scale societies to innovate solutions to environmental concerns. Theory aside, the emphasis on environmental histories is uniquely suited to address timely questions regarding human actions in the world, cultural dispositions, and climate change (anthropogenic or not) (Dawdy, 2009). North American archaeologists have developed innovative methods for recovering and interpreting evidence of past resource exploitation and landscape modification. These methods include paleoethnobotany and zooarchaeology (the study of past uses of plants and animals, respectively), geoarchaeology (the application of geological techniques to define the initial state of an archaeological site and the effects of time on that deposit), and a host of specialized techniques for recovering paleoclimatic data that have been adopted from environmental sciences.

Environmental archaeology has helped reveal the scale at which pre-Columbian communities were involved in managing and transforming physical landscapes. In most of North America, localized fire was used to encourage mosaic environments that fostered niches for edge-browsing game and fowl. In California, hunter–gatherer communities created fire-managed landscapes of high diversity that provided subsistence, medicine, and raw material, and supported the high population densities noted by the Spanish (Lightfoot and Parrish, 2009). North America also provides multiple models of plant domestication and social transformation. Tropical cultigens such as maize were initially domesticated in Mesoamerica (*see* Food Production, Origins of and Mesoamerica, Archaeology of). The pace and mode of adoption was variable across North America. As documented at the Las Capas site in the Tucson basin of Arizona (3200 years ago), early agriculture was enabled by social transformations and technological innovations such as water control via canal construction (Mabry and Davis, 2008). By way of contrast, maize arrived relatively late in New England (after AD 1000), but when adopted appears to have been incorporated into managed landscapes (Chilton, 2008). In the Midwest and Southeast maize was present by perhaps 2000 years ago, but was not important until the rise of complex polities thousand years later. Part of the resistance to maize was likely the land use practices associated with plants domesticated in North America. The so-called eastern agricultural complex included several plant species, most notably goosefoot (*Chenopodium* sp.), sumpweed (*Iva annua*), little barley (*Hordeum pusillum*), erect knotweed (*Polygonum erectum*), maygrass (*Phalaris caroliniana*), and sunflower (*Helianthus annuus*). These plants were intensively collected during the Archaic and fully domesticated by the Woodland period (Smith, 1989).

The social circumstances attending the transition toward cultivation remain poorly understood, but may relate to changes in ritual politics and feasting, together with less predictable climate. Ancient communities managed landscapes in less obvious ways. For example, Plains bison hunters created extensive landscapes through the construction of drive lines and other features. These systems enabled repeated bison kills at jump sites such as Head-Smashed-In in Alberta, Canada (Brink, 2008). In other cases, rock art in regions created sacred ties with landscapes, and likely aided in long-term management decisions with respect to traditional ecological knowledge. Ancient management undermines claims that pre-Columbian North America was a wilderness and provides a baseline for reconsidering the vulnerability of contemporary land use practices. In a similar vein, large-scale climate process that will likely affect North America in the decades to come – such as droughts, floods, and sea-level change – have already been experienced by ancient North Americans. The long history of developing traditional ecological knowledge with respect to landscape-scale rhythms will no doubt be a crucial resource for future managers and planners (Reo and Parker, 2013).

North American Archaeology in Context

The archaeology of North America is the product of communities of diverse cultural dispositions and worldviews negotiating social and ecological environments. Of equal importance is the history of investigations of these cultural resources, which have framed the significance of social practices identified in the past as much as the representations for various stakeholders. Although it is the 'past,' the archaeology of North America has much to contribute to contemporary and future concerns with climate change and the reconstitution of communities.

See also: Archaeology, Politics of; Arctic Archaeology and Prehistory; Chiefdoms, Archaeology of; Environmental Archaeology; Ethnic Identity and Ethnicity in Archaeology; Food Production, Origins of; Geoarchaeology; Historical Archaeology; Horticultural Societies; Hunter-Gatherer Societies, Archaeology of; Hunting and Gathering Societies: Anthropology; Mesoamerica, Archaeology of; Migrations, Colonizations, and Diasporas in Archaeology; Paleoethnobotany; Zooarchaeology.

Bibliography

Alt, S., 2010. Considering complexity: confounding categories with practices. In: Alt, S. (Ed.), Ancient Complexities: New Perspectives in Precolumbian North America. University of Utah Press, Salt Lake City, pp. 1–7.

Anderson, D.G., Mainfort, R.C., 2002. An introduction to Woodland Archaeology in the Southeast. In: Anderson, D.G., Mainfort, R.C. (Eds.), The Woodland Southeast. University of Alabama Press, Tuscaloosa, pp. 1–19.

Arnold, J.E., 2001. The Origins of a Pacific Coast Chiefdom: The Chumash of the Channel Islands. University of Utah Press, Salt Lake City.

Bamann, S., Kuhn, R., Molnar, J., Snow, D., 1992. Iroquoian archaeology. Annual Review of Anthropology 21, 435–460.

Bayman, J.M., 2001. The Hohokam of southwest North America. Journal of World Prehistory 15, 257–311.

Binford, L.R., 1962. Archaeology as anthropology. American Antiquity 28, 217–225.

Binford, L.R., 1990. Mobility, housing, and environment: a comparative study. Journal of Anthropological Research 46, 119–152.

Blitz, J.H., 2010. New perspectives in Mississippian archaeology. Journal of Archaeological Research 18, 1–39.

Boone, J.L., Smith, E.A., 1998. Is it evolution yet? A critique of evolutionary archaeology. Current Anthropology 39, S141–S173.

Brink, J., 2008. Imagining Head-smashed-in: Aboriginal Buffalo Hunting on the Northern Plains. AU Press, Edmonton.

Brumfiel, E.M., 1992. Breaking and entering the ecosystem—gender, class, and faction steal the show. American Anthropologist 94, 551–567.

Cameron, C.M., Duff, A.I., 2008. History and process in village formation: context and contrasts from the northern southwest. American Antiquity 73, 29–57.

Chilton, E.S., 2008. So little maize, so much time: understanding maize adoption in New England. In: Hart, J.P. (Ed.), Current Northeast Paleoethnobotany II. University of the State of New York, Albany, pp. 53–59.

Cobb, C.R., 2005. Archaeology and the "Savage Slot": displacement and emplacement in the premodern world. American Anthropologist 107, 563–574.

Colwell-Chanthaphonh, C., Ferguson, T.J., Gann, D.W., 2011. Multivocality in multimedia: collaborative archaeology and the potential of cyberspace. In: Okamura, K., Matsuda, A. (Eds.), New Perspectives in Global Public Archaeology. Springer, New York, pp. 239–249.

Dawdy, S.L., 2009. Millennial archaeology. Locating the discipline in the age of insecurity. Archaeological Dialogues 16, 131–142.

Deagan, K., 2003. Colonial origins and colonial transformations in Spanish America. Historical Archaeology 37, 3–13.

Dunnell, R.C., 1989. Aspects of the application of evolutionary theory in archaeology. In: Lamberg-Karlovsky, C.C. (Ed.), Archaeological Thought in America. Cambridge University Press, Cambridge, pp. 35–49.

Ethridge, R., 2006. Creating the shatter zone: Indian slave traders and the collapse of the southeastern chiefdoms. In: Pluckhahn, T.J., Ethridge, R. (Eds.), Light on the Path: The Anthropology and History of the Southeastern Indians. The University of Alabama Press, Tuscaloosa, pp. 207–218.

Friesen, T.M., 2004. Contemporaneity of Dorset and Thule cultures in the North American arctic: new radiocarbon dates from Victoria Island, Nunavut. Current Anthropology 45, 685–691.
Gibson, J.L., Carr, P.J., 2004. Big mounds, big rings, big power. In: Gibson, J.L., Carr, P.J. (Eds.), Signs of Power: The Rise of Cultural Complexity in the Southeast. The University of Alabama Press, Tuscaloosa, pp. 1–9.
Holly, D.H., 2005. The place of "Others" in hunter-gatherer intensification. American Anthropologist 107, 207–220.
Holmes, W.H., 1914. Areas of American culture characterization tentatively outlined as an aid in the study of the antiquities. American Anthropologist 16, 413–446.
Jenkins, D.L., Davis, L.G., Stafford, T.W., Campos, P.F., Hockett, B., Jones, G.T., Cummings, L.S., Yost, C., Connolly, T.J., Yohe, R.M., Gibbons, S.C., Raghavan, M., Rasmussen, M., Paijmans, J.L.A., Hofreiter, M., Kemp, B.M., Barta, J.L., Monroe, C., Gilbert, M.T.P., Willerslev, E., 2012. Clovis age Western Stemmed projectile points and human coprolites at the paisley caves. Science 337, 223–228.
Kidder, A.V., 1924. An Introduction to the Study of Southwestern Archaeology with a Preliminary Account of the Excavations at Pecos. Yale University Press, New Haven, CT.
Leone, M.P., Laroche, C.J., Babiarz, J.J., 2005. The archaeology of black Americans in recent times. Annual Review of Anthropology 34, 575–598.
Lightfoot, K.G., Martinez, A., Schiff, A.M., 1998. Daily practice and material culture in pluralistic social settings: an archaeological study of culture change and persistence from Fort Ross, California. American Antiquity 63, 199–222.
Lightfoot, K.G., Parrish, O., 2009. California Indians and Their Environment: An Introduction. University of California Press, Berkeley, CA.
Lubbock, J., 1865. Pre-historic Times, as Illustrated by Ancient Remains and the Manners and Customs of Modern Savages. D. Appleton & Co., New York.
Mabry, J.B., Davis, O.K., 2008. Las Capas: Early Irrigation and Sedentism in a Southwestern Floodplain. Center for Desert Archaeology, Tucson, AZ.
Meltzer, D.J., Todd, L.C., Holliday, V.T., 2002. The Folsom (Paleoindian) type site: past investigations, current studies. American Antiquity 67, 5–36.
Meltzer, D.J., 2009. First Peoples in a New World: Colonizing Ice Age America. University of California Press, Berkeley, CA.
Moss, M., 2011. Northwest Coast: Archaeology as Deep History. Society for American Archaeology, Washington, DC.
Neumann, T.W., Sanford, R.M., Harry, K.G., 2010. Cultural Resources Archaeology: An Introduction. Alta Mira Press, Lanham, MD.
Ortmann, A.L., Kidder, T.R., 2013. Building Mound A at poverty point, Louisiana: monumental public architecture, ritual practice, and implications for hunter-gatherer complexity. Geoarchaeology 28, 66–86.
Pauketat, T.R., 2001. Practice and history in archaeology: an emerging paradigm. Anthropological Theory 1, 73–98.
Pauketat, T.R., Emerson, T.E., 2008. Star performances and cosmic clutter. Cambridge Archaeological Journal 18, 78–85.
Preucel, R.W., Cipola, C.N., 2008. Indigenous and postcolonial archaeologies. In: Liebmann, M., Rizvi, U.Z. (Eds.), Archaeology and the Postcolonial Critique. Alta-Mira Press, Lanham, pp. 129–140.
Pyburn, A., 2011. Engaged archaeology: whose community? Which public? In: Okamura, K. (Ed.), New Perspectives in Global Public Archaeology. Springer, New York, pp. 29–41.
Reo, N.J., Parker, A.K., 2013. Re-thinking colonialism to prepare for the impacts of rapid environmental change. Climatic Change 120, 671–682.
Ruppel, T., Neuwirth, J., Leone, M.P., Fry, G., 2003. Hidden in view: African spiritual spaces in North American landscapes. Antiquity 77, 321–335.
Sassaman, K.E., 2004. Complex hunter-gatherers in evolution and history: a North American perspective. Journal of Archaeological Research 12, 227–280.
Sassaman, K.E., 2010. The Eastern Archaic Historicized. AltaMira Press, Lanham, MD.
Service, E.R., 1962. Primitive Social Organization: An Evolutionary Perspective. Random House, New York.
Seymour, D.J., 2012. From the Land of Ever Winter to the American Southwest: Athapaskan Migrations, Mobility, and Ethnogenesis. University of Utah Press, Salt Lake City, UT.
Singleton, T.A., 1995. The archaeology of slavery in North America. Annual Review of Anthropology 24, 119–140.
Smith, B.D., 1989. Origins of agriculture in eastern North America. Science 246, 1566–1571.
Squier, E.G., Davis, E.H., 1848. Ancient Monuments of the Mississippi Valley. Smithsonian Institution, Washington, DC.
Thomas, C., 1894. Report on the Mound Explorations of the Bureau of Ethnology. Bureau of Ethnology, Washington, DC.
Trigger, B.G., 2006. A History of Archaeological Thought. Cambridge University Press, Cambridge.
Voss, B.L., 2008. Domesticating imperialism: sexual politics and the archaeology of empire. American Anthropologist 110, 191–203.
Walker, R.B., Driskell, B.N., 2007. Introduction: new developments in Paleoindian subsistence studies. In: Walker, R.B., Driskell, B.N. (Eds.), Foragers of the Terminal Pleistocene. University of Nebraska Press, Lincoln, pp. ix–xv.
Watson, P.J., Leblanc, S.A., Redman, C.L., 1971. Explanation in Archaeology: An Explicitly Scientific Approach. Columbia University Press, New York.
Widmer, R.J., 1988. The Evolution of the Calusa: A Nonagricultural Chiefdom on the Southwest Florida Coast. University of Alabama Press, Tuscaloosa, AL.
Willey, G.R., Phillips, P., 1958. Method and Theory in American Archaeology. University of Chicago Press, Chicago, IL.
Zedeño, M.N., Hollenback, K.L., Grinnel, C., 2009. From path to myth: journeys and the naturalization of territorial identity along the Missouri River. In: Snead, J.E., Erickson, C.L., Darling, J.A. (Eds.), Landscapes of Movement: Trails, Paths, and Roads in Anthropological Perspective. University of Pennsylvania Press, Philadelphia, PA, pp. 106–132.

Relevant Websites

www.cahokiamounds.org/ – Cahokia Archaeological Site.
canadianarchaeology.com/ – Canadian Archaeological Association.
www.civilization.ca/exhibitions/online-exhibitions/archaeology – Canadian Museum of Civilization Archaeology Exhibitions.
www.nps.gov/history/laws.htm – Federal Archaeological Legislation.
history.alberta.ca/headsmashedin/ – Head-smashed-in Archaeological Site.
www.monticello.org/site/research-and-collections/monticello-archaeology – Monticello Archaeological Site.
nmai.si.edu/home/ – National Museum of the American Indian.
www.saa.org – Society for American Archaeology.
www.archaeologysouthwest.org/ – Southwest Archaeology.

North America: Sociocultural Aspects

Jessica Cattelino, University of California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Sanjek, volume 16, pp. 10732–10737, © 2001, Elsevier Ltd.

Abstract

The United States and Canada have long been understood through analysis of the sociocultural aspects of both everyday life and structural forces. Historical transformations in North America have spurred the development of social scientific theory about cultural difference, its sources, and its implications. Several ways of accounting for sociocultural life in the United States and Canada have been salient: as nations of local communities, racially organized societies, nations of immigrants, capitalist economies, contradictory democracies, and, increasingly, as settler colonial societies.

Research on the sociocultural aspects of North America (here focused on the United States and, secondarily, Canada; Mexico is discussed in a separate article) has produced a wide-ranging ethnographic record that centers on a core question: What kinds of societies and polities are the United States and Canada, and to what extent does the answer lie with cultural difference? Answers have transformed along with historical conditions that include post-Cold War global political reorganization, changing structures of economic life and resulting shifts in immigration and gender patterns, welfare state shrinkage, and the persistent inequalities and diverse lived experiences along lines that include race and gender. Many of anthropology's core theoretical and empirical concerns have been developed through the ethnography and ethnology of American lands, peoples, and polities. Anthropologists and other social scientists usefully have analyzed American societies and polities as: collections of local communities; a racialized former slave state; nations of immigrants; capitalist dreams or nightmares; and contradictory democracies. By the late 1900s, they demonstrated that sociocultural life in the United States is organized by kinship, gender, and sexuality. Increasingly in the early 2000s, scholars analyzed the connections among indigenous and nonindigenous North Americans by pointing out the ways that Canada and the United States operate as settler colonial societies.

Theorizing Human Difference from North America

North American cultural lives and social organizations have inspired prominent theories of human difference in anthropology. This is in part because of the region's evident diversity and, relatedly, an outcome of its colonial projects. Three examples illustrate the importance of sociocultural conditions in North America to the development of anthropological theory and, more broadly, of concepts that inform public understandings of human difference. First, during the mid-1800s, at a time when the United States and Canada were consolidating nationhood and progress narratives, the influential early anthropologist Lewis Henry Morgan studied Haudenosaunee society and therefrom developed an influential general theory of the stages of social development. His theory informed Engels and Marx and contributed to the rise of social evolutionism. Second, in the early twentieth century, as scientific racism consolidated and immigration boomed, Franz Boas built from American research (with indigenous peoples and immigrants, in particular) to develop his discipline-defining theory of culture as a nonreductionist explanation for human difference. Third, in the mid-1900s, as Cold War in America debated the role of material conditions in shaping culture, Julian Steward built his theory of cultural ecology from research on the relationship between environment and sociocultural life among indigenous peoples in the Great Basin. Although the paradigmatic object of anthropological analysis has been a distant cultural other, anthropologies of – not simply in – North America long have structured the discipline's core theories.

As is often observed, much of the anthropology of North America that focuses on indigenous peoples has been undertaken in a salvage mode (i.e., as if indigenous peoples as such were about to disappear), with Native American and Aboriginal peoples understood to be culture bearers appropriate for anthropological study. Research has gone hand-in-hand with colonialism, within North America and abroad. Meanwhile, as Lee Baker (2010) argues, anthropologists often viewed other North American groups and formations – he draws a contrast with African-Americans – less as cultural groups and, therefore, as less appropriate for anthropological study. The settler colonial structure of American life, as discussed below, organizes knowledge. That said, there is an old and deep anthropological literature about non-Native communities: in fact, US-based anthropologists study the United States more than any other region.

Communities of Difference

Many Americans understand the United States to be a nation that is comprised of local 'communities.' An emphasis on community has facilitated ethnographic methods of close engagement with the everyday lives of Americans, and it has afforded insights into lived diversity. Critical community studies have reckoned with the powerful American ideology of community that merges communitarian with localist/federalist traditions. Especially in the mid-1900s, ethnographers studied communities that they took to represent a fundamentally

American way of life. Lynd and Lynd's 1929 study of 'Middletown' and Warner's of 1930s 'Yankee City' (condensed in Warner, 1963), for example, showed how people refracted paradigmatically American values of individualism and hard work through community dynamics that often produced social difference and stratification. Community studies of American culture, as such, most often have been situated in white neighborhoods and towns: whiteness is the unmarked racial category in North America. Similarly, as Di Leonardo (1998) argued, anthropological studies that attempt to show 'America' to be just as exotic as any other society have taken middle-class, white-majority communities as their starting point.

Community studies of poor neighborhoods revealed how inequality permeates everything from the experience of unemployment and youth in Boston (Whyte, 1943) to the everyday rhythms of drug use and commerce in New York City (Bourgois, 1995). 'Culture of poverty' theories, which asserted that cultural dynamics produced dysfunctional family forms and hindered individual and group mobility, gained policy traction amidst 1960s civil rights struggles. The most famous example was the then-Assistant Secretary of Labor Daniel Patrick Moynihan's report on 'The Negro Family: The Case For National Action,' which identified African-American family structure and gender as legacies of slavery and Jim Crow segregation that contributed to dysfunctional 'ghetto' life. An ethnographic rejoinder, Stack's 1974 study of kinship and friendship recast poor African-Americans' social organization as robust in the face of structural inequality. Late-twentieth-century policy debates and scholarship often deployed the term 'culture' as a power-evasive euphemism for race. This uptake of the quintessentially anthropological term *culture* contravened the Boasian legacy of dissociating the two concepts (see below).

Community is not self-evident in North American social life: instead, the very idea of community has afforded Americans a resource for organizing belonging, difference, and politics. As Gregory (1998: 11) showed in his study of political culture and activism in an African-American neighborhood of New York City, "community describes not a static, place-based social collective but the power-laden field of social relations whose meanings, structures, and frontiers are continually produced, contested, and reworked in relation to a complex range of sociopolitical attachments and antagonisms." Similar approaches have reinvigorated regional analysis. In her study of the poetics of place and region in Appalachia, which often symbolizes rural poverty in scholarship and public culture, Stewart (1996) interrogates space and its narration as process and possibility. Weston's 1991 ethnography showed how gay and lesbian family transformed American kinship and its theorization by highlighting how kinship is built from community, as it is variously understood and produced. Community studies have both contributed to and disrupted common understandings of the United States as a tapestry of localized sociocultural groups.

Racially Stratified Societies, a Former Slave State

North America is fundamentally shaped by the sociocultural parameters, experiences, and consequences of race. In the United States, the nation-shaping institution of slavery and the ongoing inequality of African-Americans call into question dominant American narratives of freedom and equality, and the distinctive forms of racialization experienced by this and other groups illuminates the complexity and salience of race in everyday life. Anthropology established – through efforts to map sociocultural difference onto racialized biological difference – and then struggled to dismantle Social Darwinist understandings of race. The influential anthropologist Franz Boas's early 1900s theories of race informed both American public culture and also the most important US civil rights judicial opinions and legislation of the twentieth century. Boas's research on the cranial size of racialized second-generation Americans showed that the physical differences that scientists often had attributed race (and then used to explain differences in culture and intelligence) were mutable.

Baker (2010) contends that twentieth-century scholars generally associated questions of race with African-Americans and those of culture with, especially, Native Americans. This, in turn, contributed to a broader devaluation of African-American cultural distinctiveness, despite Hurston's (1990(1935)) and others' portrayal of the complexity of African-American cultural life. Increasingly, as the disciplines of sociology and anthropology separated in the mid-1900s, sociologists addressed African-American life while anthropologists dominated the study of Native American life: the latter rarely was theorized in terms of political status or race, but instead as cultural. This cultural framing of Native America remains the case even though Blu (2001(1980)), among others, has shown how blood and race structure indigeneity as a political category in the United States. Other contrasts warrant remark. For example, Drake and Cayton's Works Progress Administration-organized 1945 landmark study of Chicago's 'black metropolis' emphasized that African-American neighborhoods, by contrast with models of urban immigrant ethnic succession, sustained community structures that had been developed in the face of ongoing barriers to economic and geographical mobility. With the rise of Ethnic Studies and post-Civil Rights debates about race, and as demographers predict that by the mid-2000s whites will comprise a minority of the population of the United States and that 'visible minorities' will comprise one-third of the population of Canada, Frankenberg (1993) and others examined the sociocultural contours of whiteness, including at its intersection with gender.

The sociocultural aspects of race in America are lived at the level of everyday experience. From conspiracy theories to DNA ancestry testing to hip-hop, Jackson (2005) shows how turn of the century African-Americans and their interlocutors mobilized race toward intersubjective evaluation and interconnection. As Greenhouse (2011) argued, many community studies in the late twentieth century debunked racial stereotypes but eschewed direct engagement with related policy debates (e.g., 1990s racially punitive federal welfare reform policies that dismantled public benefits and exemplified welfare state retrenchment). Overall, despite trenchant research, and even as anthropologists teach in classrooms that race is a social construction rather than a biological trait, anthropological research on race in North America has lagged by comparison to insights from sociology, history, and interdisciplinary ethnic and American studies.

Nations of Immigrants

That the United States and Canada are 'nations of immigrants' has been variously presumed and investigated. Following the (anthropologically engaged) early twentieth-century Chicago School of Sociology's ethnographic examination of 'ethnic succession' and acculturation in urban neighborhoods, anthropologists have attended to the ways that immigrants are slotted into American structures of difference and inequality. By the late twentieth century, research findings and debates over multiculturalism, along with the ongoing conversation about Quebec's status in Canada, challenged the assumption that becoming American meant the gradual adoption of (implicitly white Anglo) dominant American sociocultural forms of life. Instead of experiencing a universal path, immigrants differentially become racialized into unequal forms of citizenship. For example, Ong's complementary books on the 'flexible citizenship' of East Asian transnational elites in Canada and the United States (1999), on the one hand, and the stigmatized citizenship of low-income Cambodian refugees in California (2003), on the other hand, showed that becoming American in the late twentieth century came with distinct trajectories for different groups.

Research through the twentieth century, therefore, shifted from proving immigrants' equality by showing their capacity for assimilation to charting the ways that immigration structures subjectivity and inequality. Migration is a cultural process that often includes the everyday experience of (il)legality and deportability, as Coutin (2000) showed for Salvadoran migrants to California and De Genova (2005) for 'Mexican Chicago.' These and other ethnographies of the growing Latin American migration to the United States adopted a transnational perspective and thereby challenged the older assumption that migration was a one-way cultural journey. American cultural life changes and becomes all the more globally connected as a consequence of migration. Kinship, gender, and sexuality shape these patterns and experiences, as Manalansan (2003) demonstrated for Filipino gay men. If the United States and Canada are (in part) nations of immigrants, these are stratified and unruly migrations that continually prompt national debates about the meaning and qualities of citizenship.

Capitalist Economies

In scholarship and public debate, the United States has been viewed as an exemplar of capitalism, for better or worse, and both the United States and Canada are characterized by the contradictions of capitalism and inequality. From Goldschmidt's (1947) criticism of corporate agriculture's erosion of community to Nash's (1993) examination of the relationship between community and capitalist industry to Sacks's (1988) ethnography of health care worker rights struggles, research has shown economic organization – and, more specifically, the logics of capitalism – to permeate American social and cultural life. Studies of work by Chicana (Zavella, 1987) and immigrant women (Lamphere, 1987), among others, have revealed how gendered family forms and racial/ethnic subordination are produced and reinforced by market structures. These market structures are shaped by political processes, as poverty is produced and lived through neoliberal policy and economic restructuring (see, e.g., the approach to poverty studies in Goode and Maskovsky, 2001).

Capitalism affects sociocultural life through the direct effects of day-to-day work, the economic organization of communities, and the everyday experience of economic inequality, opportunity, and community in a class-structured society. Ortner (2003), for example, showed how class shaped the life course of her high school classmates from New Jersey. Chin's (2001) ethnography of how African-American children purchase and use dolls and other objects demonstrates that consumption structures not only class but also race and other social formations. Capitalism also structures American sociocultural transformation, the mundane desires and interactions of American people, and the circulation of cultural forms such as film and music (see, e.g., Ortner, 2013, on independent films).

Anthropological methods and ethics long have privileged research in which the scholar develops sympathy with – and sometimes advocacy for – research participants. Acknowledging that such an approach systematically produces research on power's effects rather than on its production, ethnographers have studied economic elites, power brokers, and others at the center of power. For example, ethnography of Wall Street (Ho, 2009) reveals the institutional practices by which inequality is perpetuated at local and global scales. 'The market,' such research shows, is neither an abstract force nor an acultural collection of individual and institutional actors: rather, markets are sites wherein people learn and propagate values through observable sociocultural processes.

Increasingly, American economic life and its sociocultural dimensions are studied as transnational processes: for example, the authors in Maskovsky and Susser (2009) view domestic policy as the home front of American empire, with empire always comingling political and economic projects. That is, global capitalism is not simply an extension of American power and values into far-flung corners of the globe. Instead, it is a force that is produced, extended, and transformed in part through transnational sociocultural processes that also transform life in North America.

Contradictory Democracies

US- and Canadian-style democracies are forms of political organization that emerge from – and, in turn, shape – social and cultural formations. Alexis de Tocqueville wrote about the ways that 1830s Americans' everyday values dovetailed with their democratic order, Jackson Turner advanced his 1893 'frontier thesis' that associated masculinist American spirit with colonial expansion, and American political discourse long has associated democracy with individual and group characteristics. Masco (2006) views the US as a Cold War state (exemplified by Los Alamos), with Cold War logics shot through science, human and nonhuman life forms, and Americans' psychological dispositions. The idea that the United States is 'exceptional' – as a nation founded in democracy and committed to the nonimperial spread of liberty – has been

roundly criticized by critics of American imperialism and settler colonialism.

As they do elsewhere, legal anthropologists of North America deploy the methods of ethnography and sociolinguistics to understand how legal processes take hold in everyday life. They show how court claimants understand justice, immigrants tackle the bureaucratic hurdles of authorization, or courtroom actors encode cultural values in speech and evidentiary standards. Law and policy changes at federal and state levels can have profound impacts on American lives, effects that too often go unseen in the social sciences. Ethnographic investigation of the consequences of 1990s welfare reform laws that transformed the role of work in access to state benefits, for example, called attention to the struggles of Americans to undertake social reproduction in a new policy context. In Canada, as Mackey (1999) has shown, state multiculturalist policy buttresses Canadian narratives of being a tolerant democracy while simultaneously ossifying the cultural criteria by which claims for tolerance can be made. Anthropologists have played a role in implementing federal policy, including as researchers and administrators at World War II-era Japanese internment camps.

Cultural commitments congeal and transform through social movements and the politics of knowledge. For example, Ginsburg (1989) deployed ethnographic methods to show how the two sides of an abortion debate in North Dakota drew different conclusions from their shared view that women are harmed by the systematic separation of wage labor from domestic life and reproduction. Rapp (1999) vividly presented the ethics and sociocultural impact of amniocentesis and thereby highlighted the everyday politics of knowledge production and dissemination.

Anthropologists have asked how the paradigmatic disciplinary concept of 'culture' itself becomes (or always already is) political in North America. Handler's study of cultural patrimony in Québec shows nationalism to be forged through the claim to 'having a culture.' More generally, cultural claims to political belonging are unevenly distributed across race, class, gender, and other forms of difference. Rosaldo (1997) and other theorists of 'cultural citizenship,' for example, observe that people of color build civic claims and experiences from positions of racialized and cultural difference. As with immigration (discussed above) and the contradictions of democratic life more generally, anthropologists have shown citizenship in the United States and Canada to be as much about everyday belonging and exclusion as about formal legal status.

Settler Colonial Societies

In this volume, as in many, research about the United States is categorized as largely distinct from that about American Indian nations. This does not generally result from a stated commitment to underlining indigenous sovereignty and the political distinctiveness of Native nations. To the contrary, Native peoples in the United States long have been studied in anthropology, but review articles and course offerings within the discipline more often than not treat Native America as a distinct 'culture area.' This scholarly tradition reproduces a broader US public culture in which indigenous peoples more commonly are associated with past tradition (or present-day socioeconomic failure and cultural decline) than with a vibrant and contradictory American present or future. In Canada, indigeneity joins Québec's status as a prominent topic in political theory, not only in anthropology. First Nations, Inuit, and Métis peoples more frequently are understood in scholarship and public culture to form a constitutive part of the Canadian political order, albeit not necessarily on the terms or with the projected future that Aboriginal peoples choose.

The United States and Canada are settler societies with ongoing settler colonial relations among indigenous and nonindigenous peoples. Settler colonialism describes myriad sociocultural and political-economic formations that typify liberal democratic white-majority societies with dispossessed indigenous minorities. These include legal systems that recognize multiple sovereigns, struggles over how to encode differentiated forms of citizenship alongside a liberal democratic commitment to equal citizenship, the appropriation of indigenous culture for settler-state national patrimony and identity, a public culture and scholarly tradition that relegates indigenous authenticity and power to the past, and the association of indigenous peoples with poverty and precapitalism as against settler modernity. Wolfe (1999) shows how anthropological research contributed to settler colonialism's eliminatory logics. Classic settler colonial societies stemming from the British common law tradition are the United States, Canada, Australia, and New Zealand; not coincidentally, those were the four major states that voted against (and subsequently became the last to support) the 2007 United Nations Declaration on the Rights of Indigenous Peoples.

Settler colonialism creates a set of structures, practices, ideological formations, and dilemmas that are open to social scientific analysis. Relative to analysts in other disciplines, for example, anthropologists are likely to study indigenous sovereignty's local manifestations in everyday life. Cattelino's (2008) study of tribal gaming showed Florida Seminole sovereignty to be enacted through the coproduction of economic power with cultural and political distinctiveness. Relative to other anthropologists of sovereignty who more commonly focus on the sovereign's power to harm, those who study indigenous sovereignties call attention to sovereignty's political capaciousness (see, e.g., Sturm, 2002, on Cherokee blood and citizenship politics and Kauanui, 2008, on Hawaiian genealogy and political belonging). Such a focus on indigenous sovereignties, in turn, illuminates settler-state sovereignty. For example, Simpson (2014) examines Mohawk narratives of, and embodied practices at, the US–Canadian border to show how indigenous nationalism unsettles settler-state sovereignty.

Settler colonialism, and particularly the related presumption that indigenous peoples are disappearing, has shaped research on the sociocultural characteristics of the United States and Canada. The specter of indigenous disappearance that built four-field American-style anthropology in the nineteenth century formed part of broader US debates about modernity, American exceptionalism, and democracy. Documentation of indigenous culture at the turn of the twentieth century, often funded by national institutions such as the federal Bureau of American Ethnology, went hand-in-hand with the establishment of national(ist) patrimony in a still-young settler nation

that remained unsure of its pedigree. Late-twentieth-century ethical and methodological debates over disciplinary anthropology's relationship with Native America went to the heart of how knowledge production has been complicit in the dispossession of indigenous political and economic power. Cruikshank's (2005) ethnography of glaciers and colonial encounters at the border of British Columbia, the Yukon Territory, and Alaska told how social and biophysical processes combine to produce knowledge. When research with (or often on) indigenous peoples is understood to be about the settler nation more generally, it becomes clearer that anthropological research in and about the United States and Canada, as settler nations, is nothing new.

To describe the United States and Canada as settler societies foregrounds indigeneity as a constitutively dispossessed condition of American life for all of the continent's residents. Indeed, emergent inquiry into settler colonialism and the politics of indigeneity has the potential to account for the ways that being a settler society structures all American lives. Doing so does not obviate the value of describing North America in other ways – through the study of community, race, immigration, capitalism, democracy, and more – but, rather, joins with these in the effort to capture both the structural forces and lived experiences of these diverse lands and peoples.

See also: Americas, Sociocultural Overviews: North America, Native Americans; Anthropology at Home; Borders, Anthropology of; Colonialism, Anthropology of; Family: The Anthropology of the Concept and Its History; Functionalism in Anthropology; Globalization, Anthropology of; Indigeneity; Industrial Societies; Liberalism: Political Doctrine and Impact on Social Science; Modernity: Anthropological Aspects; Nationalism: General; Poverty, Culture of; Urban Anthropology.

Bibliography

Baker, L.D., 2010. Anthropology and the Racial Politics of Culture. Duke University Press, Durham, NC.
Blu, K.I., 2001(1980). The Lumbee Problem: The Making of an American Indian People. With a New Afterword by the Author. University of Nebraska Press, Lincoln and London.
Bourgois, P.I., 1995. In Search of Respect: Selling Crack in El Barrio. Cambridge University Press, Cambridge and New York.
Cattelino, J.R., 2008. High Stakes: Florida Seminole Gaming and Sovereignty. Duke University Press, Durham.
Chin, E.M., 2001. Purchasing Power: Black Kids and American Consumer Culture. University of Minnesota Press, Minneapolis.
Coutin, S.B., 2000. Legalizing Moves: Salvadoran Immigrants' Struggle for U.S. Residency. University of Michigan Press, Ann Arbor.
Cruikshank, J., 2005. Do Glaciers Listen?: Local Knowledge, Colonial Encounters, and Social Imagination. UBC Press and University of Washington Press, Vancouver and Seattle.
De Genova, N., 2005. Working the Boundaries: Race, Space, and "Illegality" in Mexican Chicago. Duke University Press, Durham.
Di Leonardo, M., 1998. Exotics at Home: Anthropologies, Others, American Modernity. University of Chicago Press, Chicago.
Drake, S.C., Cayton, H.R., 1945. Black Metropolis: A Study of Negro Life in a Northern City. Harcourt, Brace and Co, New York.
Frankenberg, R., 1993. White Women, Race Matters: The Social Construction of Whiteness. University of Minnesota Press, Minneapolis.
Ginsburg, F.D., 1989. Contested Lives: The Abortion Debate in an American Community. University of California Press, Berkeley.
Goldschmidt, W., 1947. As You Sow, first ed. Harcourt, Brace, New York.
Goode, J., Maskovsky, J., 2001. New Poverty Studies: The Ethnography of Power, Politics, and Impoverished People in the United States. New York University Press, New York.
Greenhouse, C.J., 2011. The Paradox of Relevance: Ethnography and Citizenship in the United States. University of Pennsylvania Press, Philadelphia.
Gregory, S., 1998. Black Corona: Race and the Politics of Place in an Urban Community. Princeton University Press, Princeton.
Handler, R., 1988. Nationalism and the Politics of Culture in Quebec. University of Wisconsin Press, Madison.
Ho, K., 2009. Liquidated: An Ethnography of Wall Street. Duke University Press, Durham.
Hurston, Z.N., 1990(1935). Mules and Men. Harper Perennial, New York.
Jackson, J.L., 2005. Real Black: Adventures in Racial Sincerity. University of Chicago Press, Chicago.
Kauanui, J.K., 2008. Hawaiian Blood: Colonialism and the Politics of Indigeneity and Sovereignty. Duke University Press, Durham.
Lamphere, L., 1987. From Working Daughters to Working Mothers: Immigrant Women in a New England Community. Cornell University Press, Ithaca.
Lynd, R.S., Lynd, H.M., 1929. Middletown: A Study in Modern American Culture. Harvest Books, Harcourt Brace and Company, San Diego, New York, and London.
Mackey, E., 1999. The House of Difference: Cultural Politics and National Identity in Canada. Routledge, London and New York.
Manalansan, M.F.I., 2003. Global Divas: Filipino Gay Men in the Diaspora. Duke University Press, Durham.
Masco, J., 2006. The Nuclear Borderlands: The Manhattan Project in Post-Cold War New Mexico. Princeton University Press, Princeton.
Maskovsky, J., Susser, I., 2009. Rethinking America: The Imperial Homeland in the 21st Century. Paradigm Publishers, Boulder.
Nash, J., 1993. We Eat the Mines and the Mines Eat Us. Columbia University Press, New York.
Ong, A., 1999. Flexible Citizenship: The Cultural Logic of Transnationality. Duke University Press, Durham.
Ong, A., 2003. Buddha Is Hiding: Refugees, Citizenship, the New America. University of California Press, Berkeley.
Ortner, S.B., 2003. New Jersey Dreaming: Capital, Culture, and the Class of '58. Duke University Press, Durham.
Ortner, S.B., 2013. Not Hollywood: Independent Film at the Twilight of the American Dream. Duke University Press, Durham.
Rapp, R., 1999. Testing Women, Testing the Fetus: The Social Impact of Amniocentesis in America. Routledge, New York.
Rosaldo, R., 1997. Cultural citizenship, inequality, multiculturalism. In: Flores, W., Benmayor, R. (Eds.), Latino Cultural Citizenship: Claiming Identity, Space, and Rights. Beacon Press, Boston, pp. 27–38.
Sacks, K.B., 1988. Caring by the Hour: Women, Work, and Organizing at Duke Medical Center. University of Illinois Press, Urbana and Chicago.
Simpson, A., 2014. Mohawk Interruptus. Duke University Press, Durham.
Stack, C., 1974. All Our Kin: Strategies for Survival in a Black Community. Harper & Row, New York.
Stewart, K., 1996. A Space on the Side of the Road: Cultural Poetics in an "Other" America. Princeton University Press, Princeton.
Sturm, C., 2002. Blood Politics: Race, Culture, and Identity in the Cherokee Nation of Oklahoma. University of California Press, Berkeley.
Warner, W.L., 1963. Yankee City. Yale University Press, New Haven.
Weston, K., 1991. Families We Choose: Lesbians, Gays, Kinship. Columbia University Press, New York.
Whyte, W.F., 1943. Street Corner Society: The Social Structure of an Italian Slum. The University of Chicago Press, Chicago.
Wolfe, P., 1999. Settler Colonialism and the Transformation of Anthropology: The Politics and Poetics of an Ethnographic Event. Cassell, London and New York.
Zavella, P., 1987. Women's Work and Chicano Families: Cannery Workers of the Santa Clara Valley. Cornell University Press, Ithaca.

Nosology in Psychiatry

Femi Oyebode, School of Clinical and Experimental Medicine, University of Birmingham, Birmingham, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.E. Kendell, volume 16, pp. 10737–10744, © 2001, Elsevier Ltd.

Abstract

Nosology is a classification of disease in which disease categories are mutually exclusive and jointly exhaustive, and may also have other formal relationships to one another as well as being arranged either in tiers or in a hierarchy. The most influential classification systems in contemporary psychiatry are the World Health Organization's *International Classification of Diseases – 10th edition* (ICD-10) and the American Psychiatric Association's *Diagnostic and Statistical Manual – 5th edition* (DSM-5). These classification systems have converged over the years but are still distinct documents. This article reviews the historical aspects of the development of nosology and classification systems in psychiatry from the Classical Greek period to the modern day. It emphasizes the strength of contemporary nosologies, namely, the marked improvement in reliability that is due almost entirely to the development of formal diagnostic criteria and guidelines. The most obvious weakness of both contemporary nosologies is the limited evidence for the validity of most of their constituent categories. The combination of a limited understanding of the etiology of most mental disorders and the lack of evidence for natural boundaries or 'zones of rarity' separating their clinical syndromes means that most existing diagnostic concepts and definitions are not securely based.

Although mental disorders have figured in nosologies since Hippocrates, separate classifications of mental disorders did not appear until the nineteenth century. For a long time, confusion reigned and early attempts to secure agreement on a single international classification ended in failure. When it became clear to the World Health Organization (WHO) in the 1950s that few countries were using the International Classification of Diseases (*ICD-6*) which they had themselves formally approved, a concerted attempt was made both to improve the mental disorders section of the *ICD* and to persuade everyone to use it. Despite a widespread disenchantment with diagnoses, *ICD-8*, published in 1965, was a better and more widely used nosology than its predecessors, but the WHO lacked the political power to make major changes. The classification produced by the American Psychiatric Association in 1980 (*DSM-III*) provided these radical changes – multiple axes, operational definitions, a new grouping of syndromes, and rejection of several ancient terms. All subsequent nosologies, including the *ICD*, have incorporated most of these. At present, two nosologies are in widespread use – The American Psychiatric Association's *DSM-V* and the WHO's *ICD-10*. Although operational definitions of syndromes have greatly improved reliability, the validity of most diagnostic concepts is still uncertain and this is only likely to change with better understanding of etiology.

Nosology is the classification of diseases (from the Greek words *nosos*, disease, and *logos*, discourse), whereas taxonomy is the classification of animals or plants. A nomenclature is simply an approved list of categories or titles. In order to constitute a classification, its categories must be mutually exclusive and jointly exhaustive, and may have other formal relationships to one another as well, being arranged either in tiers or in a hierarchy.

History: From Hippocrates to the 1940s

For at least the last 2500 years, most generations of physicians and several philosophers have produced a nosology. Usually this was not because they were preoccupied with classification. It was because a nosology almost inevitably reflects the range and variety of diseases recognized by its authors, the extent of their clinical experience, their assumptions or knowledge about etiology, and their familiarity with or ignorance of the detailed manifestations of illness. Nearly all these nosologies have incorporated several categories of what we now regard as mental disorder, thereby demonstrating that their authors regarded these disorders as diseases to be treated by physicians, and usually with the same range of assumptions about causation and treatment as other diseases.

It is apparent from the corpus of writings attributed to him that Hippocrates recognized at least six different forms of mental illness:

1. phrenitis (acute mental disturbance with fever),
2. mania (acute mental disturbance without fever),
3. melancholia (chronic mental disturbance),
4. epilepsy,
5. hysteria (convulsions and other paroxysmal symptoms), and
6. scythian disease (cultural transvestism).

His contemporary, Plato, on the other hand, distinguished between Natural Madness originating in physical disease and several kinds of Divine Madness (Prophetic Madness given by Apollo, Religious Madness given by Dionysus, Poetic Madness inspired by the Muses, and Erotic Madness inspired by Aphrodite and Eros).

Many of the nosologies developed over the next 2000 years are listed by Menninger (1963). Most were fairly simple and incorporated Hippocrates' categories of mania, melancholia, and epilepsy. Others, like those of Boissier de Sauvages and William Cullen at the end of the eighteenth century, contained several hundred categories of disease arranged in tiers mimicking the orders, genera, and species of plants. Only in the nineteenth century, with the emergence of physicians specializing in the treatment of insanity (known as alienists), did separate classifications of mental illness start to appear. They

quickly proliferated, however, because every self-respecting alienist, and certainly every professor, produced his own nosology, some based on behavioral differences, others on presumed differences in etiology, others on psychological differences, and some on a mixture of all three. Indeed, as a medical historian subsequently observed, "to produce a well-ordered classification almost seemed to have become the unspoken ambition of every psychiatrist of industry and promise."

It was soon apparent that this multiplicity of classifications, with diverse meanings often attributed to the same terms, rendered communication between one hospital and another almost impossible, and with this came recognition of the need for a single nosology acceptable to everyone. At the Congress of Mental Medicine in Antwerp in 1885, a commission was appointed, chaired by a Belgian alienist, Morel, to produce this. The commission duly delivered a simple typology that was adopted formally by the International Congress of Mental Science in Paris in 1989, but as the English psychiatrist, Hack Tuke, had predicted, its nosology was generally ignored. However, as this was the first formal attempt to secure international agreement and the forerunner of the present International Classification, its 11 categories of insanity have some historical significance. They were

1. mania,
2. melancholia,
3. periodical insanity (Baillarger's *folie a double forme*),
4. progressive systematic insanity,
5. dementia,
6. organic and senile dementia,
7. general paralysis,
8. insane neuroses (hysteria, epilepsy, hypochondriasis, etc.),
9. toxic insanity,
10. moral and impulsive insanity,
11. idiocy, etc.

While Morel and his colleagues were engaged on this ultimately fruitless task, several national organizations were pursuing similar but more modest aims. In Britain, the Royal Medico-Psychological Association (the forerunner of the Royal College of Psychiatrists) produced a classification in 1882, which it recommended to its members for general use. Further drafts were produced in 1904, 1905, and 1906 before the Association finally accepted that there was no prospect of persuading its members to restrict themselves to the diagnoses listed in any official nomenclature, British or international. A similar sequence of events took place in the USA. In 1886, the Association of Medical Superintendents of American Institutions for the Insane (the forerunner of the American Psychiatric Association) adopted a modified version of the British classification of 1882. This was revised in 1913 to incorporate Kraepelin's concepts of dementia praecox, involutional melancholia, and manic depressive insanity but, even so, many States and asylums still preferred to use their own nosologies, or declined to be bound by any nomenclature, national, or local. Indeed, the New York State Commission in Lunacy continued to use its own private classification until 1968. In 1937, the need for an internationally agreed nosology was discussed once more at the Second International Congress for Mental Hygiene. Although it soon became clear that there was still little hope of any real consensus emerging, the simple eight-category classification proposed by the French psychiatrist Bersot was adopted by two countries, Portugal and Switzerland, and remained in use in both the countries for 20 years.

The Origins of the World Health Organization's International Classification

Largely at the instigation of two medical statisticians, William Farr in London and Jacques Bertillon in Paris, the International Statistical Congress had, as far back as 1853, recognized the need for '*une nomenclature uniforme des causes de décès applicable a tout les pays*' (a uniform nomenclature of causes of death applicable to all countries) and had not only produced such a nomenclature but revised it regularly for the next 50 years. Probably because of the importance of mortality statistics to public health and to governments, these successive editions of what came to be known as the Bertillon Classification of Causes of Death were used increasingly widely; and in 1900, the French Government assumed responsibility for the classification and convened a series of international meetings in Paris in 1900, 1920, 1929, and 1938, and thereby produced four successive editions of what was by then called the International List of Causes of Death.

When the World Health Organization (WHO) came into being in 1948, one of its first public actions was to produce a sixth revision of this International List. Renamed the International Statistical Classification of Diseases, Injuries and Causes of Death (*ICD*), this was for the first time a comprehensive nosology covering the whole range of disease, rather than merely causes of death, and so for the first time included a classification of mental illness. In fact, Section The Third Edition of the American Psychiatric Association's Classification (DSM-III) of this sixth revision (*ICD-6*), entitled 'Mental, Psychoneurotic and Personality Disorders,' contained 10 categories of psychosis, nine of 'psychoneurosis,' and seven of 'disorders of character, behavior and intelligence,' most of them subdivided further. Unfortunately, although the nomenclature of *ICD-6* was adopted unanimously by the 1948 revision conference, and was duly 'recommended for use' by all the member states of the WHO, its mental disorders section failed to gain acceptance, and 11 years later, it was found to be in official use only in Finland, New Zealand, Peru, Thailand, and the UK. History was repeating itself. Deeply concerned by this state of affairs, the WHO commissioned an English psychiatrist, Erwin Stengel, to investigate the situation and if possible to make recommendations.

The situation that Stengel encountered he described as one of 'almost general dissatisfaction with the state of psychiatric classification, national and international' and the attitude of many psychiatrists toward the classification of mental disorders seemed to have become 'one of ambivalence, if not of cynicism' (Stengel, 1959). Many countries had official classifications of their own, some of recent origin, others dating back to the 1930s, but none, new or old, was regarded as satisfactory by its users, or used conscientiously and consistently by them. In the UK, although the General Register Office published mental hospital admission rates and other data in the format of the *ICD*, the returns from individual hospitals from which these

data were derived frequently showed a blatant disregard for, or ignorance of, the requirements of the nomenclature. In the USA, Section The Third Edition of the American Psychiatric Association's Classification (DSM-III) of the *ICD* was ignored completely, despite the fact that American psychiatrists had taken a prominent part in drafting it. Instead, a nosology drawn up by the American Psychiatric Association and published in 1952 as a *Diagnostic and Statistical Manual* (*DSM-I*) was in widespread use. Even so, there were serious local problems. The New York State Department of Mental Hygiene still insisted on retaining its own classification, and influential psychoanalysts like Menninger (1963) argued forcefully that classifying patients into categories of illness did more harm than good and should be abandoned.

Reasons for Dissatisfaction with Psychiatric Classification

Many factors contributed to the unwillingness of psychiatrists to commit themselves to using a common nosology and to the general dissatisfaction with all classifications. Most fundamentally, there was no rational criterion available to enable the psychiatric community, faced with the need to choose between several alternative nosologies, to decide which was the best, or most valid. In other branches of medicine, the old diagnostic concepts and nosologies inherited from the eighteenth century or before had largely been replaced by new concepts and classifications based on at least a partial understanding of etiology. Phthisis had been replaced by pulmonary tuberculosis, myxoedema by hypothyroidism, and Mediterranean anemia by thalassemia, and no physician doubted the validity of these new concepts or the value of classification. Psychiatry, on the other hand, remained as ignorant as ever about both the pathology and the underlying causes of most of its common disorders, and so was compelled to continue defining these disorders by their syndromes, just as Sydenham had done in the seventeenth century, and went on using ancient Hippocratic terms like mania and melancholia.

A further problem was that many of these disorders appeared to merge into one another. A few conditions, like Down's syndrome and Huntington's chorea, did seem to be clearly demarcated but most were not. Patients with the characteristic symptoms of one disorder not infrequently presented with the typical symptoms of a quite different disorder on a subsequent occasion, and it was a commonplace to encounter patients with a mixture of the characteristic symptoms of what were supposed to be unrelated disorders. Such problems bedevilled the distinction between, for example, schizophrenic and manic depressive psychoses, and between depressive illnesses and anxiety states, and resulted in either persistent controversies about how best to distinguish the disorders in question or, even worse, doubts about the validity or usefulness of the diagnoses themselves.

The combination of these two problems – the lack of understanding of etiology or of any demonstrable pathology, and the absence of obvious boundaries between several of the most common syndromes – meant that, in many instances, there was no rational means of choosing between alternative diagnostic concepts or alternative nosologies. In this situation, most psychiatrists, understandably, opted to use whichever diagnostic terms and whichever classification were most familiar, particularly if they were their own.

Other considerations also contributed to the problem. Few of the diagnostic terms in common use held important therapeutic implications. Before the 1960s, there were few treatments that were only effective in a single diagnostic category. Electroconvulsive therapy was effective in most acute psychoses – schizophrenic, melancholic, or manic. Chlorpromazine and the other new neuroleptic drugs appeared to have a similarly broad range of efficacy. Most influential of all, particularly in North America, was the fact that a diagnosis had ceased to be a necessary precondition to treatment, because psychoanalytically oriented psychotherapy was widely regarded as an effective and appropriate treatment across almost the whole range of psychiatric disorders from mild character disorders to florid schizophrenia. The implications of diagnoses for prognosis were equally tenuous. There were, to be sure, well-established differences in outcome between populations of patients with different diagnoses, both in the short and the long terms. On average, the course and outcome of schizophrenic and affective psychoses, or of depressions and anxiety states, were significantly different. But the overlap was very substantial, and as a result, a diagnosis was only of limited value in predicting outcome in individual patients. By the early 1960s, it was also becoming apparent that the reliability of psychiatric diagnoses, even those of experienced psychiatrists, was often alarmingly low, and that even under optimal conditions, disagreements were commonplace – because the patient gave different information to different interviewers or because that information was interpreted in different ways or because there were covert differences in the interviewers' diagnostic concepts.

Finally, many psychiatrists and psychologists were concerned about the harmful effects that diagnostic labels often had on their patients. Labels like hysteria or schizophrenia sometimes encouraged hospital staff to treat the patients in their care as 'hysterics' or 'schizophrenics' rather than as distressed people. Several diagnostic terms also had pejorative overtones, lowering the patient's self-esteem even further than their illness itself had already done and encouraging relations, neighbors, colleagues, and employers to belittle them.

The Response to These Problems

The most important recommendations in Stengel's report to the WHO were that all diagnoses should be explicitly shorn of their etiological implications and regarded simply as 'operational definitions' for specified types of abnormal behavior, and that the next revision of the *ICD* should have an accompanying glossary 'available from the beginning in as many languages as possible' providing a brief description, or thumbnail sketch, of all the disorders in its nomenclature. Other, mainly American, psychiatrists advocated operational definitions as a means of improving reliability. The problem with the brief descriptions Stengel was advocating was that, like traditional textbook descriptions, they described the typical features of a clinical syndrome without making it clear which combinations of those features were adequate to establish the

diagnosis in question. A typical textbook description, or the thumbnail sketch of a glossary, lists the typical symptoms of disease X as A, B, C, and D, and often goes on to say that E and F are sometimes present as well. But it does not spell out which combinations of those symptoms are adequate and which are not to establish diagnosis X, and as a result, different diagnosticians make different assumptions, and so often end up making different diagnoses. An operational definition is simply a statement, or a set of statements, which makes it clear for all possible combinations of A, B, C, D, E, and F, which are and are not adequate to establish diagnosis X.

The 1960s also saw the introduction of multiaxial classifications. The Swedish psychiatrist Essen-Moller had long advocated a classification with separate axes for symptomatology (i.e., the syndrome) and etiology, primarily as a means of resolving the difficulties that inevitably occur in a classification in which some disorders are defined by their symptoms (e.g., mania) while others by their presumed etiology (e.g., puerperal psychosis). As he pointed out, having separate axes would make it possible to identify all manic illnesses, whether or not they occurred shortly after childbirth, and all puerperal illnesses regardless of their symptomatology. Shortly afterward, a WHO working party – set up to produce some much needed recommendations for an improved classification of childhood disorders in the *ICD* – made a firm recommendation that a triaxial classification with separate axes for the clinical syndrome, the child's intellectual level and 'etiological and associated factors' should be tried out and, if these pilot studies proved successful, incorporated into the next revision of the *ICD*.

The Eighth and Ninth Revisions of the International Classification (ICD-8 and ICD-9)

Stengel's report and the recommendations of its own expert committee were taken seriously by the WHO, and the next (eighth) revision of the International Classification, which was published in 1965 and came into use in 1969, was a considerable improvement on its predecessors. For the first time, all psychiatric disorders were brought together in the Mental Disorders section of the classification, thus making it self-sufficient and easier to use. The sections dealing with personality disorders, sexual deviations, mental retardation, alcoholism, and drug addiction were all expanded and recast and a new group of reactive psychoses was introduced at the request of Scandinavian psychiatrists. Even more important, the American Psychiatric Association was persuaded, despite considerable domestic opposition, to commit itself to using the nomenclature of *ICD-8* instead of that of its own *Diagnostic and Statistical Manual*. The German and Scandinavian psychiatric societies also agreed to use the new nomenclature so that, for the first time, all the major contributors to the psychiatric literature, with the notable exception of France, were committed officially to using the same classification. Unfortunately, the committee responsible for producing the companion glossary that Stengel had so strongly recommended did not start work until 1967 and the glossary was not published until 1974. The hiatus was partly filled, though, by the publication in 1968 of two national glossaries – American and British – to the nomenclature of *ICD-8*.

The ninth revision (*ICD-9*) was published in 1978, and this time the companion glossary was available from the beginning. But despite a series of eight international seminars held between 1965 and 1972, attended by psychiatrists from over 40 countries, there were no major changes. A few new categories, like nondependent abuse of drugs, were added in recognition of the broadening scope of psychiatric practice, a clearer distinction (reminiscent of Plato's distinction between natural and divine madness) was drawn between organic psychoses and other psychoses, and there was a substantial expansion of the number of categories allocated to disorders of childhood, but that was all. None of the fundamental problems – like the low reliability of diagnostic assignments, the tensions created by the incorporation of etiological assumptions into diagnostic terms, and the increasingly complex and chaotic classification of depressions – had been tackled effectively. It was also clear why this was so. The need for international consensus – the agreement of nearly 200 sovereign states to use the same nomenclature – was making it almost impossible to introduce radical changes; and an increasing number of sometimes incompatible terms were being added to the nomenclature with each new revision, because individual psychiatric societies always fought for the inclusion of their own favorite terms more persistently than they were prepared to oppose the attempts of others to do likewise. It was becoming apparent, in other words, that if fundamental improvements were ever to be achieved, it was unrealistic to expect the WHO to be the driving force.

The Third Edition of the American Psychiatric Association's Classification (DSM-III)

For the research community, the most important shortcomings of all classifications of mental disorders, national and international, were the low reliability and uncertain validity of most of their diagnostic terms. The Washington University Department of Psychiatry in St Louis in the USA reacted to this situation by publishing what came to be known as the Feighner criteria – a set of operational definitions, intended for research use, of the 15 major syndromes they regarded as having been adequately validated (Feighner et al., 1972). Other rival sets of definitions, like the Research Diagnostic Criteria, soon followed, and in this climate, the American Psychiatric Association committed itself to producing a new, third edition of its *Diagnostic and Statistical Manual*. When it was finally published in 1980, this *DSM-III* was dramatically different from any previous nosology of mental disorders. It was multiaxial with independent recording systems for five different kinds of information – the clinical syndrome or syndromes; associated personality or developmental disorders; associated physical disorders; the severity of recent psychosocial stresses; and the highest level of social functioning in the past year. The grouping of clinical syndromes was also novel, with all affective disorders brought together and the traditional distinction between psychotic and neurotic disorders abandoned. Even more provocatively, many of the psychiatry's most hallowed terms – like hysteria, manic-depressive illness, and even psychosis and neurosis – were discarded and replaced by unfamiliar and rather unattractive terms like somatoform

disorder, substance use disorder, and paraphilia. A brave but only partially successful attempt was also made to define the term 'mental disorder.' Finally, and most important of all, nearly all the 200 categories of disorder in the manual were provided with operational definitions, and it was demonstrated in extensive field trials that this made it possible to achieve higher reliability than had ever been attained previously.

None of the innovative features of *DSM-III* was entirely novel. The advantages of both multiple axes and operational definitions had been pointed out long before and both had already been introduced into clinical practice in some centers. The concept of hysteria had been under attack for over a generation, as had the distinction between psychotic and neurotic disorders. What was new was that a powerful professional body like the American Psychiatric Association was prepared to commit itself and its members to all these innovations simultaneously. For a time the fate of *DSM-III* hung in the balance, for in the eyes of many older and still influential psychiatrists, its authors had gone too far. But before long, its advantages were increasingly apparent and eventually it became a best seller, in not only the USA but also worldwide.

With hindsight, the adoption of *DSM-III* was a crucial step in the development of a scientific nosology of mental disorders. Certainly, all subsequent nosologies – national and international – have been deeply influenced by it. Of all its innovative features, the commitment to operational definitions was the most important. Not only did this result in a major improvement in the reliability of psychiatric diagnoses and eliminate the unsuspected differences in diagnostic criteria that had previously bedevilled international comparisons, they also made it explicit that diagnostic terms like schizophrenia were simply convenient designations for more or less arbitrary groupings of symptoms.

Nosologies (1992–2013)

The Fourth Edition of the American Psychiatric Association's Classification (*DSM-IV*)

The American Psychiatric Association published a revision of *DSM-III* in 1987 and a new fourth edition (*DSM-IV*) in 1994. Neither abandoned any of the fundamental features of *DSM-III*. For the most part, they simply corrected the anomalies that had come to light, reworded several operational definitions in order to improve their reliability or to conform with changes in expert opinion, and introduced a few additional categories of disorder. A new 'clinical significance' criterion was also added to the definition of many disorders in *DSM-IV* to counter the criticism that too many people met with the *DSM-III* or *DSM-IIIR* criteria for common conditions like major depression and phobic disorder and that many of them were not significantly handicapped.

The authors of *DSM-IV* also endorsed and expanded the rather cumbersome *DSM-III* definition of 'mental disorder' as "a clinically significant behavioral or psychological syndrome or pattern that occurs in an individual and that is associated with present distress or disability or with a significantly increased risk of suffering death, pain, disability, or an important loss of freedom." In addition, it was stipulated that "this syndrome or pattern must not be merely an expectable and culturally sanctioned response to a particular event, for example, the loss of a loved one," and that "whatever its original cause, it must currently be considered a manifestation of a behavioral, psychological, or biological dysfunction in the individual," thereby emphasizing that "neither deviant behavior nor conflicts that are primarily between the individual and society are mental disorders unless the deviance or conflict is a symptom of a dysfunction in the individual." The introduction to *DSM-IV* also emphasized, rather more succinctly, that "the term mental disorder unfortunately implies a distinction between mental disorders and physical disorders that is a reductionistic anachronism of mind/body dualism. A compelling literature documents that there is much physical in mental disorders and much mental in physical disorders. The problem raised by the term mental disorders has been much clearer than its solution, and, unfortunately, the term persists in the title of *DSM-IV* because we have not found an appropriate substitute."

The Tenth Revision of the International Classification (*ICD-10*)

The WHO published a tenth revision of the *International Classification* in 1992, after extensive consultations and field trials in several different parts of the world. Although the titles and definitions of many of the disorders listed in the Mental and Behavioural Disorders section of this *ICD-10* are different from those of either *DSM-III* or *DSM-IV*, its debt to the former is obvious. The general format of the two nosologies is very similar and, crucially, *ICD-10* provides 'diagnostic criteria for research' for every category (i.e., operational definitions) as well as 'clinical descriptions and diagnostic guidelines' for ordinary clinical use (WHO, 1992, 1993). This was only possible because, by the time the member states of the WHO were called upon to endorse the draft text of *ICD-10*, there had been time for the international psychiatric community to come to terms with the radical innovations of *DSM-III* and to appreciate their advantages. Significantly, though, no attempt was made to define the term 'mental or behavioral disorder.' Indeed, there is an apologetic note in the introduction explaining that "the term disorder is used…so as to avoid even greater problems inherent in the use of terms such as disease and illness. Disorder is not an exact term, but it is used here to imply the existence of a clinically recognizable set of symptoms or behavior associated in most cases with distress and with interference of personal functions."

At present, we are at cross-roads as *DSM-V* has just been published and *ICD-10* is undergoing revision. The two nosologies of so-called 'mental disorders' are in widespread use. In their essentials, they are very similar to one another. In points of detail – particularly, the titles and subdivisions of individual syndromes and the detailed definitions of these disorders – they are often different. The American classification is used mainly in North America and by research-oriented psychiatrists further afield. The *International Classification* is used elsewhere, particularly in Europe and developing countries, and by all governments for their formal statistical returns to the WHO. In some ways, it is regrettable that there are now, and will be for the foreseeable future, these two rival nosologies. However, the confusion generated by two rivals is much less than that

produced by a dozen, the nightmare of the 1950s. There are also advantages in having some choice. The coexistence of alternative titles, subdivisions, and definitions for several of the major syndromes helps to emphasize their arbitrary nature. It also invites formal comparisons, which may help to provide the evidential base for better categorization and definition in the future. It is likely, too, that for financial reasons, the American classification will be revised more frequently in the future than the *International Classification*, and some sections of the nosology may need to be revised quite quickly in the light of etiological research.

The Fifth Edition of the American Psychiatric Association's Classification (*DSM-V*)

DSM-V was published in May 2013. The most significant and perhaps controversial changes deal with neurodevelopmental disorders, depressive disorders, substance-related and addictive disorders, and personality disorders. Autism spectrum disorder is the new name for four previously separate conditions, namely, autistic disorder, Asperger's disorder, childhood disintegrative disorder, and pervasive developmental disorder. In addition, communication disorders now include social (pragmatic) communication disorder, the name for a new condition of persistent difficulties in the social uses of verbal and nonverbal communication. The previous bereavement exclusion criterion in Major Depressive disorder in *DSM-IV* has been removed. A new disorder, gambling disorder, has been added to the substance-related and addictive disorders. Finally, personality disorders now have a reduced (moderate) level of impairment in functioning as criterion with the implication of an increase in the numbers liable to receiving a diagnosis of personality disorder. These changes generated considerable debate in the lead up to publication of *DSM-V*. Allen Frances (2009), who had been the head of *DSM-IV* Task Force, argued that there was no need for a 'paradigm shift' in the structure and approach of *DSM-V* in comparison to previous editions. Particularly of concern was the idea that the threshold for diagnosing a condition might be lowered and that this was liable to result in "wholesale imperial medicalization of normality that will trivialize mental disorder and lead to a deluge of unneeded medication treatments – a bonanza for a pharmaceutical industry."

Shortcomings of Contemporary Nosologies

Uncertain Validity

The most obvious weakness of both contemporary nosologies is the limited evidence for the validity of most of their constituent categories. The combination of a limited understanding of the etiology of most mental disorders and the lack of evidence for the natural boundaries or 'zones of rarity' separating their clinical syndromes (*see* Differential Diagnosis in Psychiatry) means that most existing diagnostic concepts and definitions are not securely based. There are exceptions, of course. The etiologies of, for example, Huntington's chorea, Down's syndrome, the Fragile X syndrome, and Creutzfeldt Jacob disease are all well established, and enough is known of the neuropathology and etiology of Alzheimer's disease and vascular dementia, and of the various syndromes associated with the use of psychoactive substances, for them to be reasonably secure. But the same cannot be said for most of the other major syndromes. For many, there is cogent evidence that genetic factors are making a major contribution to their etiology, and that life experiences and other environmental factors are also involved. It seems increasingly likely, though, that many different genes, each on its own of relatively small effect, are responsible for the genetic basis of most syndromes, and that some of these genes are risk factors for more than one syndrome – for affective as well as schizophrenic psychoses, for example, or for anxiety states as well as depressions. Some adverse life experiences appear to predispose to a similarly broad range of apparently unrelated disorders. In essence, these two nosologies may have sacrificed validity for reliability. This point is underscored by the statement by the United States of America National Institute of Mental Health (NIMH) (Insel, 2013), on the eve of publication of *DSM-V* to the effect that "the strength of each of the edition of *DSM* has been 'reliability' – each edition has ensured that clinicians use the same terms in the same ways. The weakness is lack of validity. Unlike our definitions of ischemic heart disease, lymphoma, or AIDS, the *DSM* diagnoses are based on a consensus about clusters of clinical symptoms, not any objective laboratory measure." NIMH concluded that there was a need for a new approach, based on biology as well as symptoms and proposed a new Research Domain Criteria (RDoC), which is a hybrid of a dimensional approach mapped to cognitive, circuit, and genetic aspects of mental disorders. It is far too early for the full implications of this proposal to become clear.

The Dimensional and Hybrid Alternative

The statement by NIMH once again points to the fact that current genetic, or indeed, other findings do not support a categorical framework of classifying mental disorders. For this reason, at least, the possibility of representing variation in symptomatology by dimensions rather than categories needs to be considered. The idea is not, of course, new. Wittenborn developed an elaborate multidimensional representation of the phenomena of psychotic illness in the mid-twentieth century, and other psychologists and psychiatrists have subsequently advocated and used dimensional models to portray depressive or chronic schizophrenic illnesses, or even mental disorder as a whole. Dimensional classification solves at a stroke all the problems associated with boundaries and comorbidity, and may also be a powerful means of predicting outcome. Even so, there is no realistic prospect at present of any formal national or international classification adopting a dimensional format. Medicine as a whole is too firmly committed to categories of illness, and in any case, dimensions are better suited to the portrayal of variation in populations than to day-to-day decisions about the treatment of individual patients. This does not mean, though, that dimensional systems may not have an important role for research purposes, or as an experimental alternative to a traditional nosology.

Future Developments

In time, as understanding of the underlying determinants of most of the major syndromes slowly accumulates, mental disorders will cease being defined by their syndromes and be defined instead by their neuropathology, or perhaps by a characteristic combination of genetic risk factors, neuroimaging, circuitry abnormalities, and so on. This may not happen quickly, and when it does, the classification of mental disorders – if they are still so-called – may bear little resemblance to our present nosologies. What has to date been missing in the development of nosology and classification systems in psychiatry is the importance of social and cultural factors in shaping and determining the structure of the classification systems, the actual content and the formal criteria for determining class membership of the basic categories, and perhaps most importantly in clarifying the definition of boundaries between disorders. Social and cultural determinants undoubtedly influence the development, presentation, and behavioral response to emotional and psychological distress. Some would argue that language too has substantial influence not only in how distress is communicated but also in how it is experienced. Thus, future developments will have to take account of the ways in which social and cultural factors modify or mediate the experience and expression of psychiatric disorders. It is true that there are global, pan-cultural factors (etic) at play and implicitly constructing the ways in which human distress is manifested but it is equally true that local and parochial social and cultural factors (emic) also have significant influence. The so-called culture-bound syndromes may then have a place in the international classification systems that do not consign these conditions to the status of curios but to legitimate expressions of distress that are either unique or capable of being collapsed into existing nosology. In the immediate future, however, major change seems improbable.

See also: Differential Diagnosis in Psychiatry.

Bibliography

American Psychiatric Association, 1980. Diagnostic and Statistical Manual of Mental Disorders (DSM-III), third ed. American Psychiatric Association, Washington, DC.

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV), fourth ed. American Psychiatric Association, Washington, DC.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-V), fifth ed. American Psychiatric Association, Washington, DC.

Feighner, J.P., Robins, E., Guze, S.B., Woodruff, R.A., Winokur, G., Munoz, R., 1972. Diagnostic criteria for use in psychiatric research. Archives of General Psychiatry 26, 57–63.

Frances, A., 2009. A warning sign on the road to DSM-5: beware of its unintended consequences. Psychiatric Times 26 (8).

Insel, T., 2013. Transforming Diagnosis. www.nimh.nih.gov (accessed 16.05.13.).

Kendell, R.E., 1975. The Role of Diagnosis in Psychiatry. Blackwell, Oxford, UK.

Menninger, K.A., 1963. The Vital Balance: The Life Process in Mental Health and Illness. Viking Press, New York.

Stengel, E., 1959. Classification of mental disorders. Bulletin of the World Health Organization 21, 601–663.

World Health Organization, 1974. Glossary of Mental Disorders and Guide to Their Classification: For Use in Conjunction with the International Classification of Diseases, eighth revision. World Health Organization, Geneva, Switzerland.

World Health Organization, 1978. Mental Disorders: Glossary and Guide to Their Classification in Accordance with the Ninth Revision of the International Classification of Diseases. World Health Organization, Geneva, Switzerland.

World Health Organization, 1992. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines. World Health Organization, Geneva, Switzerland.

World Health Organization, 1993. The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research. World Health Organization, Geneva, Switzerland.

Nostalgia, Anthropology of

Millie Creighton, University of British Columbia, Vancouver, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews anthropological and sociological perspectives on nostalgia. It discusses collective nostalgia during times of rapid change as a response to insecurity and uncertainty. It addresses nostalgia for a past way of life due to a widespread sense that strong emotional bonds to other human beings and to community have diminished. It explores the romanticization of rural areas, small towns, and villages, of cultural calendars, and of childhood accompanying collective nostalgia in many parts of the world, as a reaction against urban and industrial anomie. It notes nostalgia's frequent collapse of time and place, with rural areas of the present nostalgically equated with the past, or desires to return to a geographical place of childhood really reflecting a desire for a past time. It situates collective nostalgia in theoretical discussions of community as a multivocal symbol in human life. It discusses 'nationalistic nostalgia,' showing how emotionally laden nostalgic sentiments are used in the nationalistic invention of states as imagined communities. It discusses ways where nostalgia for a lost past, rather than stagnating people in the past, can be a collective means to move forward into the future. It explores how anthropology itself has been viewed as the nostalgic conscious of the modern world for remote, small-scale, or preindustrial cultures and lifeways seemingly lost or sacrificed to modernity. It addresses earlier attempts at salvage anthropology, discussing these in relationship to both a positively espoused 'critical nostalgia,' and a negatively viewed 'imperialist nostalgia.' Finally, it addresses more recent trends in the anthropology of nostalgia that consider how nostalgic projections are used differentially by different elements within a culture in the on-going negotiation of culture, and how processes of world systems, and accelerating trends of globalization and transnationalism influence or induce cycles of collective nostalgia.

Nostalgia Isn't What It Used To Be, proclaims the title to actress Simone Signoret's autobiography (1978). Nostalgia, particularly for a romanticized version of the past as remembered, or imagined, accompanies feelings that life does not seem quite as vivid, or quite as good, as it did in an earlier, more golden time. Signoret's ironic title actually epitomizes this nostalgic sense by suggesting that even the present's poignant pangs of sentimental longing for the past do not seem either quite so romantic nor quite so melancholic as they once did, hence culminating in a nostalgia for a presumed more authentic nostalgia. The title also echoes a contemporary moment in anthropology, in which the anthropological exploration and study of nostalgia is not what it used to be. To explore these issues, this article will discuss nostalgia in individual and collective life, and also discuss transitions and trends in anthropological approaches to nostalgia, linked to the recognition of ongoing historical and global processes affecting human life and cultures.

At the individual level, nostalgia involves sentimentalized longings for the past. Most often, however, this is a combination of a remembered, an imagined, and a recreated past, which in the nostalgic mind-set seems much better – more benevolent, loving, and problem-free – than the past actually was. Individual nostalgia often involves recreated memories of childhood as a time of innocence, when life seemed to follow a meaningful, and patterned course, and one felt embedded in a nexus of human and humane relationships, rather than encumbered by bureaucratic and legalistic ones. Childhood is thus often focal to nostalgic memory and recalled as unproblematic, despite having been fraught with its own problems and tensions. At the collective level, the anthropology of nostalgia attempts to understand how and why such flushes of romanticized longing for the past that individuals periodically feel, become part of the collective imagination. The anthropology of nostalgia seeks to explain why nations, communities, or other collectivities of people seem to experience widespread cycles of nostalgic reminiscing for a presumed shared past.

Uncertainty, Change, and the Collective Imagination

The anthropology of nostalgia suggests that collective nostalgia waxes in times of transition or uncertainty. Increasing nostalgic sentiments frequently underlie a collective identity crisis indicating a sense of confusion or uncertainty surrounding a proclaimed cultural or national identity, or a sense that it is threatened. Increasing urbanization, technological changes, and globalization, all of which can cause swift changes in local lifestyles and cultures, can also give rise to collective nostalgia. Even among peoples or cultural groups seeking change, a seemingly contradictory combination of the rhetoric of progress, with an increase in rhetoric of nostalgia is frequently noted. The anthropology of nostalgia suggests this apparent contradiction is really a corollary response, indicating a fundamental ambivalence to sudden change and so-called modernization. Some anthropologists, such as Ivy (1988, 1995) contend that the contemporary condition of postmodernism, typified by rapidly changing lifestyles accompanied by the pronounced sense that life is shifting and uncertain, induces collective nostalgia.

The sense that life is changing and uncertain can also be associated with the feeling that one no longer has strong emotional bonds to other humans. In his treatise on the sociology of nostalgia, Davis (1979: p. 141) contends that collective nostalgia is often a response to cultural transitions that leave masses of people with feelings of loneliness and estrangement. According to Davis, such nostalgia is part of

a "collective search for identity," which "looks backward rather than forward, for the familiar rather than the novel, for certainty rather than discovery" (Davis, 1979: pp. 107–108). The past, since it has already occurred, emerges as a perceived anchor of stability. Smith (1982: p. 128) notes that the "past will always seem more stable than the present." Therefore, images of the past are invoked to offset the perceived threat of cultural change or identity loss. However, these tend to be highly romanticized images of the past. Tuan points out that "the cult of the past calls for illusion rather than authenticity" (1977: p. 194).

Nostalgic projections of a sentimentalized past involve what Hobsbawm and Ranger (1983) call 'invented traditions.' This term does not necessarily mean such traditions are false, but rather recognizes that the past is constantly undergoing recreation and reinvention by those in the present, in response to present needs.

Country, Community, Childlike, Calendric: The Time and Place of Nostalgia

Anthropologists working in different areas of the world have noted that collective nostalgia frequently romanticizes rural areas, a sense of community, childhood, and calendric rituals once more frequently or more intensely engaged in. Romanticization of the countryside, smaller communities, or rural life is common in collective nostalgia. Increasing urbanization often draws people away from small towns and villages where they belonged to networks of relationships, into the anonymity of city life. This influx of population into cities can also break down existing urban communities. Concomitantly, increasing industrialization, technological development, and globalizing economic systems change not only the methods humans use in productive work activities, but the relationships constructed via these activities. The *gemeinschaft* involvement thought to characterize preindustrial communities, in which life and work found meaning within a network of social relationships, is replaced by *gesellschaft* society, in which human interactions are more purely economic or business exchanges (Toennies, 1957). Romanticization of rural areas can result from urban anomie, or from frustrations with a perceived *gesellschaft* lifestyle, even for those still based in rural or smaller communities.

Romanticization of rural life frequently is noted in areas that have undergone rapid urbanization and industrialization. In Western Europe, industrialization created a romanticized image of the preindustrial countryside. Twentieth-century North America saw periodic flushes of nostalgia for small town life. Some Asian areas, such as Japan (Creighton, 1997, 1998) and Taiwan (Harrell, 1994) experienced nostalgia for agrarian village life after rapid industrial or economic growth.

Such romanticization of rural life suggests a desire to negate feelings of estrangement, by imagining a place and time of belonging. The particular images of rural life vary in each area noted, but all suggest a desire to return to a closer bond with nature, and to the nexus of reciprocal human relationships idealistically projected as typifying 'home town' or 'home village' life. Affirmations of a rural centrality, of the countryside, or of a 'back-to-nature' lifestyle, proffer the hope that a return to the presumed virtues of the past can offset the negative consequences of industrialization and technological development. The nostalgic emulation of deeply embedded human relationships thought to characterize rural areas, and more 'natural' lifestyles, involves a quest for community and the desire to affirm a collective identity.

'Community' is a powerful, multivocal symbol in many cultures. Despite variations in its particular ideal constructions, it generally suggests a desire for belongingness, to an essential need for human interaction in order to lead meaningful lives. Modern nostalgia, as a response to widespread feelings of homelessness and feared cultural loss, often projects symbolic images of a past rural existence to address the threat of estrangement by embracing an ideal of collective belongingness that represents 'community,' the 'good life,' 'wholesomeness,' and moral virtue.

Nostalgic projections of a home town, or home village past, for individuals who actually experienced these often represent what Bellah calls 'communities of memory.' According to Bellah (1985: p. 154), people who grew up in 'communities of memory' know the stories of a community's past, share its ideal values, and understand what it means to participate in the ritual, aesthetic, and seasonal practices that defined the community as a way of life. However, such imagery can invoke intense nostalgia even among those who did not grow up in such communities through the construction of the parallel 'communities of imagination.'

Nostalgia for a past way of life, which rural imagery suggests, is associated with nostalgia for the personal past, suggesting childhood. Sentimentalized projections of community and of childhood reflect the desire to find a 'place' where one belongs and feels secure. In mass nostalgic longings, there is a tendency to collapse time and place dimensions. Rural areas are geographical places that exist contemporaneously with urban areas. However, in the collective landscape of nostalgia, rural areas become symbols of an earlier existence. Childhood is a time in people's lives. However, in the nostalgic imagination, this time becomes a place. Childhood becomes that place where one felt sheltered and secure. Reflecting this emphasis on childhood, nostalgic imagery frequently projects maternal tropes. Both 'mother' and 'community' suggest intimate places of nurturance. For the young child, mother is not just a person who provides love and warmth, but also according to Tuan (1977: pp. 137–138) the child's primary 'place,' a 'haven of stability' where "fundamental needs are heeded and cared for without a fuss." The idealized past community is likewise that 'place' where one belonged, where one had obligations to others, and where one could depend upon others.

Calendric rituals and holidays are a focus of collective nostalgia. Cultural calendars punctuate experience with special days laden with symbolic meanings that have the capacity, in Zerubavel's (1981: p. 46) terms, to "synchronize the sentiments of a large number of people" by creating "emotional rhythms that affect large collectivities." Special calendric events integrate community life. They also play important symbolic and affective roles in both the nostalgic remembrance of community and the nostalgic imagination of community. For example, in an analysis of English Victorian life, Gillis (1989) contends that Christmas offered a nostalgic spiritual pilgrimage, a symbolic

journey of return and reunion to home, community, and 'mother.' In Japan, *Gyoji* are the annual cycle of calendric festival events, emotionally situated in seasonal associations that formerly orchestrated life in agrarian villages. Creighton (1998) describes how the nostalgia-driven 'retro-boom' prompted collective revivals of this traditional calendar of *gyoji* events, in both rural and urban areas, and in new ways in very modern settings such as department stores.

Nostalgia, Claimed Authenticity, Cultural Myths, and Negotiations of Culture

Whereas early anthropological inquiry tended to look for a shared similarity of understandings or world view among members of a culture, in more recent decades anthropology has increasingly explored culture as the organization of diverse elements, and with greater attunement to dynamics of power structures within societies. This orientation recognizes culture as something negotiated rather than fixed and something continually in the process of renegotiation, exemplified by Clifford's (1986: p. 19) declaration that; "Culture is contested, temporal, and emergent." Attunement to what is nostalgically espoused as vanishing from a culture is important to these considerations of culture as emergent and negotiated. Often, what is lamented as the past culture, and corresponding past values and virtues of the culture, or bemoaned as diminishing might actually be what certain members of the culture, whose interests such projections serve more, wish to assert as more 'culturally authentic' as a means of implementing or reimplementing such cultural orientations. What is nostalgically claimed to be 'our authentic culture,' which is eroding, may actually be a means of perpetuating a self-asserted myth about the culture or aspects of it, rather than represent the reality of past experience.

Examples of gender assertions help reveal this. Often reflections on 'traditional gender roles' mask desires to shape gender roles in a particular manner or constrain people within certain roles, based on nostalgic projections that things were different – and better – in the past. The commentary of the proverbial 'grumpy old man' that "back in my day, we did n't need this Women's lib stuff" denies the actual situation in which women (along with many men) were enmeshed in restrictive and oppressive gender expectations. Frequently expressed beliefs in traditional roles tend to reify them, and erase the actuality of past conditions. For example, in many industrial (or postindustrial) societies the image of the 'traditional woman' as an unemployed homemaker masks the reality that many women were – and had to be – in paid work in the past as well as in the present and potentially denies the validity of struggles for better employment conditions for women with the false suggestion that this was or is unnecessary. Such imagery can also be projected outward. For example, many in the West hold an image of the 'traditional Japanese woman' as unemployed, whereas statistics show women in Japan were in paid employment in relatively high and consistent percentages from the beginning of such statistics taking in the early 1900s (Brinton, 1993).

Nostalgic beliefs and representations of the past can also obviate against struggles of social groupings based on other criteria (ethnicity, age, abilities, lifestyle) to be as readily recognized as part of the cultural context. Nostalgic projections of a group's past culture, virtues, or values can also be used by nation-states to evoke nationalism or to prompt negative feelings toward other countries and groups of people in the fostering of positive feelings for one's own country or group.

Nostalgia and Nationalism

Anthropologists have pointed out that nostalgia is used to promote nationalism, through the projection of emotionally laden images, which are then linked to an asserted national way of life. This can involve a positive affirmation of a proud collective identity for a nation, or conversely, a negative and dangerous fostering of ultranationalism, or one that promotes negative feelings about those in other countries or as members of other ethnicities. In his famous treatise on nationalism and invented communities, Anderson says all national communities are imagined and stresses the importance of understanding "the style in which they are imagined" (1985, p. 15).

The nostalgic invocation of a presumed pristine past, an essentialized past culture, and sentimental reaffirmations of nostalgic holidays, are part of the stylistic imagination of national communities. Anderson claims that Christmas in modern Western nations functions as a calendric event, which helps perpetuate "that remarkable confidence of community in anonymity, which is the hallmark of modern times." Despite its appeal in different Western nations, its sentimentally loaded symbolic associations with growing up in a particular place, allow it to be associated with an imagined national community. Likewise in many Asian countries, New Year festivities are frequently used to reaffirm national identity, even though the holiday exists and is celebrated in different forms in other places.

Nostalgic associations with childhood experience are also used to promote nationalistic sentiment. 'Motherhood' frequently becomes a dominant, multivocal symbol used in the promotion of nationalism in diverse nation-states. Creighton (1991) explores the irony of motherhood as a nationalistic symbol. 'Mother' as a symbol could suggest a universal or shared humanity, since everyone presumably had a mother. However, mother symbolism is frequently used to emphasize a nationalistic identity and the exclusion of outsiders, by projections of a culturally specific mother prototype that is then linked to national identity.

In times of political, social, or economic strain, global or regional uncertainty, such uses of nostalgia in the promotion of nationalism can usurp grass roots attempts to reinvigorate bonds among people in local communities. Nationalistic uses of nostalgia can obviate against attempts to build associations that foster closer ties among human beings that go beyond national boundaries. Nationalistic nostalgia offers a sense of belonging to an imagined national community, but it is a form of belongingness, which frequently requires the exclusion and even repudiation of nonmembers. As countries vie for positioning on the world stage, rights over territory or resources, nostalgic longings of the imagined past can be actively utilized

by nation-states to prompt negative feelings toward current members of other nations based on reiterations of past wrongs, tensions, or glories lost, rather than solely focusing on solutions to help recognize and remedy previous wrongs.

There are sometimes reverse cases in the imagination of a (new) national identity in which nostalgia reflects a longing for a past not necessarily romanticized nor necessarily any longer desired. The term 'ostalgie' has been associated with postsocialist yearning in the former East Germany after reunification of Germany as one nation, to describe a yearning for what is lost even when one knows or believes the present is better than the past. Thus 'ostalgie' brings out a particular paradox of nostalgic possibilities.

Anthropology as Modernity's Nostalgia: Salvage Anthropology and Critical Nostalgia

It has been suggested that anthropology has served as the nostalgic consciousness of the so-called modern world. Anthropology's earlier focus on remote small-scale societies has been thought by some to reflect a nostalgia within the modern world for the type of societies and human relations possible in less industrialized or capitalistic systems. The movement for 'salvage anthropology,' referring to the drive to record presumably vanishing cultures before they disappeared, suggests this sort of nostalgia for possible past human lifeways. Associated with salvage anthropology was the concept of 'critical nostalgia' within anthropology suggested by Williams (1973) and discussed by Clifford (1986). Critical nostalgia as cultural analysis involved exploring what were frequently conceptualized as cultural forms anchored in the past as alternatives to the structures and organizational imperatives of present cultures otherwise seen as unavoidable and hence 'naturalized' (seen as 'natural' rather than as human constructions) by those living in modern societies. While some anthropologists considered 'critical nostalgia' as something potentially positive allowing for a critique of postmodernist societies, others have seen such tendencies to posit remote or seemingly exotic cultures nostalgically more negatively and have designated such practices as embedded in 'imperialist nostalgia' (Rosaldo, 1989).

Newer Trends in Anthropology and the Study of Nostalgia

As indicated above, since at least the mid-1980s anthropology tends to explore and analyze culture not as essentialized or involving a priori forms, but as enmeshed in on-going processes of reaffirmation, potential transition, and negotiation by inside as well as outside actors and elements. Cultures are no longer perceived as 'vanishing' in the sense that culture is viewed as processual. Culture is seen as emergent, and as something influenced by historical circumstances and larger world systems. In more recent years, the past one to two decades, there has been increasing emphasis within anthropology toward understanding how accelerating forces of globalization and transnational flows have affected cultures in both macro- and microcontexts and a greater emphasis on this in academic inquiry (Appadurai, 1996, 2001; Hannerz, 1996; Robertson and White, 2003; Tsing, 2000; Brooks and Normore, 2010). Thus greater attention has been paid to issues of so-called glocalization (Robertson, 1992, 1995) or understanding culture change in local contexts as entwined with globalizing processes. The anthropology of nostalgia has become increasingly important as a result. Anthropology still attempts to understand how and why sentiments of nostalgia spread through communities or collectivities or become at times more frequently or loudly voiced collectivity. As studies of globalization and transnationalism have become more important within anthropology, these trends are also looked at in addressing collective nostalgia, which is seen as potentially increasing or expanding due to the heightened effects of these trends, the spread of changes and the speed of changes resulting from them. The greater attunement to these larger and often seemingly external influences on cultures and individuals within them, does not replace an attempt to understand the construction of cultural idioms and meanings. The anthropological exploration of nostalgia brought on by such conditions is still about understanding the nature of self in relationship to society, but it also increasingly recognizes the need to understand the effects of larger world systems operating upon local cultures, and how cultures are embedded in and affected by historical processes, along with larger world economic systems, and increasing trends of transnationalism, translocalism, globalization, and glocalization.

Conclusions

The anthropology of nostalgia attempts to understand why there appear to be periods of mass nostalgia. Anthropologists have noted that collective nostalgia frequently accompanies or follows periods of change, uncertainty, or alienation. The collective landscape of nostalgia in industrialized or post-industrialized societies is typically rural in nature, reflecting an apparent collapse of time and place in the human imagination. Rural areas exist in present time, but are romantically envisioned as representing a past, more pristine way of life. In this collective nostalgic romanticization, place has been confused with time. Significant people of childhood, such as mother, and calendric holidays of a given culture that are emotively experienced during childhood, have strong nostalgia-inducing capacity. In this collective nostalgic romanticization time – the time of childhood – has become a place, a place where security and belongingness seemed assured.

Nostalgia can be channeled to promote nationalism through the projection of emotionally laden symbolism, which is then asserted to represent 'our way of life,' making the nation-state into an 'imagined community.' Anthropologists call this an 'imagined community' because in real communities the ties that bind people are created through actual, on-going, personal interactions. The use of nostalgia to promote nationalistic sentiment can have positive consequences in creating positive collective identity, or negative consequences in fostering ultranationalism, or hatred toward all members of another nation or grouping of people. Projections of a nostalgic

longing for a supposedly more authentic past, can also represent attempts of certain elements within cultures to inscribe or reinscribe forms of culture they find more desirable or which better suit their own interests relative to other elements of the culture, by making it appear that this was the once shared and understood culture.

Collective nostalgia can involve a melancholic longing for the past that prevents people from moving forward. However, many anthropologists suggest that collective nostalgia can also have positive effects, and help enable people to redirect themselves toward future goals. Nostalgic recreations of the past can inspire people to attempt to create deeper human relationships, and a stronger sense of community in their present lives. Nostalgia for the past can be part of a constructive process that allows people to deal with the uncertainty of the present in order to move forward. Collective romanticized sentiments help create a nostalgically imagined assertive linking of the present with the past, and thus with the future. Images of a presumed past in which life seems to have been more stable and loving can help people to offset the threats of insecurity and alienation in the present, allowing them to move forward into an uncertain future with renewed assurance that there is a continuity between past, present, and future, and that the human values they associate with the past still do and will continue to exist.

Anthropological explorations of remote, small-scale, or seemingly 'exotic' peoples at times spoke to the nostalgia of groups in postindustrialized societies as representing possible alternative modes of existence now lost to them. Some have suggested that anthropological studies at times reflected a nostalgic atonement for worlds destroyed in the processes of modernization (Fischer, 1986: p. 200). At one time attempts at so-called salvage anthropology sought to document and preserve the knowledge of cultures, with the possible suggestion that such preservation could possibly permit regenerative possibilities for future generations. Along with this a movement within anthropology designated as 'critical nostalgia' sought to explore such possibilities with an attempt to better the conditions of humanity by considering alternative lifeways shown by differing cultures, with the possibility of implementing aspects considered desirable into the organization of contemporary social systems. For some this is seen positively as potentially enabling a critique, reevaluation, and remaking of postmodern societies, while others view the tendency to contemplate other cultures in a nostalgic mood as more strongly constituting a form of 'imperialist nostalgia.'

The lament of modernity, represented in the title of Signoret's autobiography that "Nostalgia isn't what it used to be" suggests that even a nostalgia for nostalgia may be part of the modern condition. Whether or not nostalgia is what it used to be, the anthropological study of nostalgia is not what it used to be. In the past decade in particular, anthropology has taken a closer view of how accelerating processes of globalization, glocalization, translocalism, and transnationalism are reflected in collective surges of nostalgia. The anthropology of nostalgia still seeks to understand what nostalgic longings for the past reveal about the human condition, about collectivities and about individual social actors within cultural contexts. However, it adds to this a greater attunement to how self and society in localized cultural contexts is affected by historical processes, macroeconomic circumstances, and global world systems.

See also: Collective Beliefs, Sociological Explanation of; Collective Memory, Psychology of; Community Studies: Anthropological; Nationalism: General; Ritual; Sociological Traditions; Tradition, Anthropology of.

Bibliography

Anderson, B.R.O.'G., 1983. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London.
Appadurai, A., 1996. Modernity at Large: Cultural Dimensions of Globalization. University of Minnesota Press, Minneapolis, MN.
Appadurai, A., 2001. The production of locality. In: Beyer, Peter (Ed.), Religion in the Process of Globalization. Ergon Verlag, Wurzburg, pp. 99–123.
Bellah, R.N., Madsen, R., Sullivan, W.M., Swidler, A., Tipton, S.M. (Eds.), 1985. Habits of the Heart: Individualism and Commitment in American Life. University of California Press, Berkeley, CA.
Brinton, M., 1993. Women and the Economic Miracle: Gender and Work in Postwar Japan. University of California Press, Berkeley, CA.
Brooks, J., Normore, A., 2010. Educational leadership and globalization: literacy for a global perspective. Educational Policy 24 (1), 52–82.
Clifford, J., 1986. On ethnographic allegory. In: Clifford, J., Marcus, G.E. (Eds.), Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley, CA, pp. 98–121.
Creighton, M., 1991. Maintaining cultural boundaries in retailing: how Japanese department stores domesticate 'things foreign'. Modern Asian Studies 25 (4), 675–709.
Creighton, M., 1997. Consuming rural Japan: the marketing of tradition and nostalgia in the Japanese travel industry. Ethnology 36 (3), 254–269.
Creighton, M., 1998. Pre-industrial dreaming in post-industrial Japan: department stores and the commoditization of tradition. Japan Forum 10 (2), 1–23.
Davis, F., 1979. Yearning for Yesterday: A Sociology of Nostalgia. Free Press, New York.
Fischer, M.M.J., 1986. Ethnicity and the post modern arts of memory. In: Clifford, J., Marcus, G.E. (Eds.), Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley, pp. 194–233.
Gillis, J.R., 1989. Ritualization of middle-class family life in nineteenth century Britain. International Journal of Politics, Culture, and Society 3 (3), 213–235.
Hannerz, U., 1996. Transnational Connections: Culture, People, Places. Routledge, New York.
Harrell, S., 1994. Playing in the valley: a metonym of modernization in Taiwan. In: Harrell, S., Huang, C. (Eds.), Cultural Change in Postwar Taiwan. Westview Press, Boulder, CO.
Hobsbawm, E., Ranger, T. (Eds.), 1983. The Invention of Tradition. Cambridge University Press, Cambridge.
Ivy, M., 1988. Critical texts, mass artifacts: the consumption of knowledge in postmodern Japan. South Atlantic Quarterly 87 (3), 419–444.
Ivy, M., 1995. Discourses of the Vanishing: Modernity, Phantasm, Japan. University of Chicago Press, Chicago.
Robertson, R., 1992. Globalization: Social Theory and Global Culture. Sage, London.
Robertson, R., 1995. Glocalization: time-space and homogeneity-heterogeneity. In: Featherstone, M., Lash, S., Robertson, R. (Eds.), Global Modernities. Sage, London, pp. 25–44.
Robertson, R., White, K. El, 2003. Globalization: an overview. In: Robertson, R., White, K.E. (Eds.), Globalization: Critical Concepts in Sociology, vol. I. Routledge, London, pp. 1–44.
Rosaldo, R., 1989. Imperialist nostalgia. In: Culture and Truth: The Remaking of Social Analysis. Beacon Press, Boston, MA.
Signoret, S., 1978. Nostalgia Isn't What it Used to Be. Weidenfeld and Nicolson, London.
Smith, M.E., 1982. The process of sociocultural continuity. Current Anthropology 23 (2), 127–142.
Toennies, F., 1957. Community and Society [Gemeinschaft und Gesellschaft]. Michigan State University Press, Lansing, MI.
Tsing, A., 2000. The global situation. Cultural Anthropology 15 (3), 327–360.
Tuan, Y.-F., 1977. Space and Place: The Perspective of Experience. University of Minnesota Press, Minneapolis, MN.
Williams, R., 1973. The Country and the City. Oxford University Press, New York.
Zerubavel, E., 1981. Hidden Rhythms: Schedules and Calendars in Social Life. University of Chicago Press, Chicago.

Notestein, Frank Wallace (1902–83)

John C Caldwell, Australian National University, Canberra, ACT, Australia

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 15, pp. 10749–10753, © 2001, Elsevier Ltd.

Abstract

Frank Notestein did more than any other person in the twentieth century to develop demography as a field of study in the USA, and probably the world. He was a creator of institutions and a synthesizer of ideas. After playing an important role in developing a population program at the Milbank Memorial Fund 1928–36, he became successively the first head of Princeton University's Office of Population Research 1936–59, Consultant Director of the new United Nations Population Division 1946–48, and President of the Population Council, 1959–68. At Princeton he trained a high proportion of the next generation of US demographers. His name is indissolubly linked with the development of demographic transition theory, emphasizing the central role played by deliberate fertility control, the importance of massive social and economic change, and the global nature of both mortality and fertility declines. More controversially, he also argued from about 1945 that governmental intervention in developing countries, assisted by external technical aid, could and should hasten fertility transition and limit population growth. He forged the Population Council into an instrument for both helping to achieve these aims and measuring the achievement.

Life and Career

Frank Notestein did more than any other person in the twentieth century to develop demography as a field of study in the USA, and probably the world. He was born on 16 August, 1902, in Alma, Michigan to a family with academic connections. His undergraduate education was in the mid-West in Alma and Wooster colleges, but he earned his Ph.D. in Economics at Cornell University. There he came under the influence of Walter Willcox, which was important both because Willcox was interested in social and population statistics and because of his connections with the Milbank Memorial Fund (MMF), New York, which was to become an important early funder of population research. At Wooster, Notestein met Daphne Limbach whom he married in 1927.

Notestein worked with Edgar Sydenstricker from 1928 to 1936 helping to establish the MMF's population program, carrying out studies of both differential fertility and contraception, and arranging for MMF to fund a demographic survey in rural China. During this time the *Milbank Memorial Fund Quarterly* became, and remained for over a decade, the primary journal where demographic research was reported. In 1936, Frederick Osborn convinced Albert Milbank to fund Princeton University to establish the Office of Population Research (OPR). This was the first program in the USA to combine demographic research and graduate education. Notestein became its first head, 1936–59, directing and synthesizing its work with profound consequences for demographic thought. Because of Princeton's departmental structure OPR's faculty were assigned to appropriate university departments (Notestein to the Department of Economic and Social Institutions), as were its students for most of their course work. OPR provided a single course, given for the first years solely by Notestein, for which it was later to award its own certificate. This was where most of OPR's demographic discoveries and theories were first taught. OPR, because of its head start and its high standards, provided from its faculty and students an extraordinary proportion of the leading demographic teachers and researchers as the discipline expanded through the USA, and indeed the world, in the second half of the century. The two-year-old bibliographical journal *Population Index* was moved to OPR and Notestein became one of the editors.

Notestein took leave from OPR during 1946–48 to become the first Consultant Director of the new United Nations Population Division, which acted as the secretariat for the United Nations Population Commission. Ryder (1984, p. 10) believes that the decision that the United Nations should have such a commission was probably influenced by OPR's wartime research on demographic change in Europe, commissioned by the League of Nations. Notestein left Princeton in 1959 to become President of the Population Council, finally to retire from the Council back to Princeton in 1968 to the position of Visiting Research Demographer at OPR. The Population Council was established as a result of a meeting in Williamsburg, Virginia in 1952. It was held on the initiative of John D. Rockefeller III, but Notestein prepared the agenda and appears to have arrived with a clear concept of the form such an institution should take. For some years, the Council was the only organization which could provide expert advice to Third World governments on contraceptive technology and the establishment of family planning programs. During Notestein's presidency its annual budget rose from US$1 million to US$11 million and its professional staff from 10 to 90, with an increasing presence in developing countries. It also offered scholarships, predominantly to students from developing countries, to pursue graduate studies elsewhere in demography or reproductive biology.

To understand Notestein's achievements it is necessary to have some knowledge of the man himself. He was plain-speaking and straightforward, as was his prose. He had a great capacity for leadership, often in a fatherly way, perhaps because the Notesteins had no children. He had a gift for creating and building institutions, which was enhanced by his closeness to some of the rich and powerful, notably John D. Rockefeller III and Frederick Osborn. He moved easily between the foundation and the university worlds, partly because he believed that the aim of both should be to improve the human lot. He had great love for Princeton University but

was no believer in the ivory tower. He was strangely modest, saying that others had greater intellectual contributions to make than himself, but that his strength lay in his gift for recognizing ability and providing it with an opportunity. His successors at the various institutions he developed were men he had brought into the demographic field: Ansley Coale at OPR, John Durand in the United Nations Population Division, and Bernard Berelson at the Population Council.

Notestein's world was the institution he was building and a major source of information the seminars and papers of that institution. When first presenting demographic transition theory in its more fully developed form, he described it as 'some of the work which my colleagues and I, of Princeton's Office of Population Research, have undertaken ...' (Notestein, 1943, p. 165). Prominent among the colleagues contributing to demographic theory were Dudley Kirk and Kingsley Davis. The most employed of Notestein's outside sources were two works which significantly appeared in his first year of teaching: *World Population* (Carr-Saunders, 1936), which documented Third World mortality declines and population growth, and *Medical History of Contraception* (Himes, 1936), which showed how ancient were attempts to restrict fertility. In his two most famous papers on demographic transition, 'Population—the long view' (1945) and 'Economic problems of population change' (1953), there are only a couple of references which go beyond these two sources and work that either he or his colleagues had undertaken at MMF, OPR or the UN. All his papers drew quite early on ongoing research projects at MMF or OPR, which themselves were not often fully reported until years later. Many of his papers were given first to conferences of mixed disciplines and achieve their impact from their simplicity, common sense, and clarity. After 1936 surprisingly few of his publications were papers submitted without solicitation to demographic journals. All, as Coale (1979) notes of his writing and teaching, were marked by skepticism, respect for evidence, and insistence on rigor and technical competence.

Contribution to Demography

The greatest support for demographic studies and for their funding came between the World Wars from those in the eugenics movement. Their major concerns were that the lower socioeconomic classes were out-breeding the higher ones (and, they assumed, the genetically brighter) and, in the USA, that immigrants had higher fertility than the native-born. The contribution of Sydenstricker and Notestein (1930) was to present, virtually without comment on the implications for society, technically sophisticated analyses of a sample of previously untabulated 1910 US census returns to show that immigrants had higher fertility than the native-born in the first generation but lower in the second. They also showed there were marked rural–urban fertility differentials and somewhat smaller occupational ones. They tried to free the debate from ideology. After 1936 Notestein ceased publishing in the area of fertility differentials, although it was an important element in the 1941 Indianapolis Survey, which was recommended by a committee of which he was a member. That study was the world's first household fertility and family planning survey and was the direct ancestor of the global programs of such surveys in the later decades of the century.

During the 1930s Notestein, together with Regine Stix, analyzed information from interviews in 1932–34 of women who had attended a Margaret Sanger family planning clinic in 1931–32. Three conclusions were to be taught to all who passed through OPR. The first was that deliberate fertility control completely explained the fertility transition and that it did not have biological origins. The second was that 'New patterns of living and new values brought growing interest in family limitation that spread the use of known methods and stimulated the development of new ones. In a real sense, modern birth control is as much the result of new interest in family limitation as its cause' Stix and Notestein (1940, p. 150). The third conclusion, deriving from the discovery that most of these women (predominantly Jewish refugees from Europe) had earlier employed *coitus interruptus* (withdrawal) to limit fertility, was that all societies at all times could have reduced family size if the socioeconomic and cultural circumstances had made this desirable.

Notestein's name is most strongly linked with demographic transition theory, which is as important an intellectual framework in modern demography as evolution is in biology. In its simplest sense, namely, as societies evolve and become richer, mortality falls, to be followed later by a decline in fertility, the concept existed all through the twentieth century. Notestein taught demographic change within such a framework from the inception of his OPR course. His Cornell supervisor, Willcox, had published a paper drawing these conclusions in 1916, and Warren Thompson had divided the process into analytical stages in 1929. Notestein's and OPR's contributions were to identify the conditions of fertility decline, the role of contraception, the great spurt of population growth which occurred before the transition was over, and the spread of the phenomenon. This was greatly facilitated by the sponsorship of four major studies of demographic change in Europe at OPR by a division of the League of Nations transferred from Geneva to Princeton for the duration of World War II, and subsequent funding by the Carnegie Foundation through MMF. This work was published in the period 1944–46, but by February 1943 Notestein was reporting to a meeting of the American Philosophical Society the nature of the 'vital revolution,' the growth of population because of the 'lagging transition in growth rates,' the spread of this revolution to Southern and Eastern Europe, and the fact that mortality was already declining in agrarian Asia portending rapid population growth there.

The globalization of demographic theory was permitted by the funding in 1943 by the Geographer of the US Department of State of studies of population change in India (now India, Pakistan, and Bangladesh) and Japan. Later studies were to cover Formosa (now Taiwan), Malaya (now Malaysia), and Lebanon. Notestein's interest in Asia had been whetted by his part in funding and analyzing the population segment of John Lossing Buck's land utilization survey carried out in China in 1929–31. His conviction that the colonial world would face mounting population growth in the postwar period had been strengthened by the Annual Conference of the MMF in Princeton in early 1944 where OPR researchers gave papers on rapid population growth in India, Japan, and Egypt, and agreed

that global demographic transition was underway. This led to his paper, 'Population—the long view' (1945), presented to a 1944 conference of food specialists at the University of Chicago, and a 1952 address to a conference of agricultural economists, 'Economic problems of population change' (1953). With these two papers the theory was complete. Notestein maintained that fertility declined first in the urban populations of industrialized societies because of the development there of conditions that eroded the traditional family and ascribed status, favoring instead individualism and self-advancement. The decline was retarded because of the slow erosion of ancient institutions and beliefs which had been required to encourage the high fertility necessary to offset the high mortality of premodern times.

But Notestein was already entering a new stage in his life. He and other OPR researchers had recorded that European fertility had begun to fall centuries after mortality did so, with the result that population of European extraction increased sevenfold. This seemed to be an impossible scenario for agrarian Asia and suggested that mortality decline there might meet a Malthusian barrier strangling both the mortality transition and hopes for rising living standards. Already in 'Population—the long view' (1945, p. 52), he had raised the possibility that Asian population growth might be held to doubling by 'wise and vigorous leadership 'and' studied efforts ... to induce declining fertility.' By 1952, in 'Economic problems of population change' (1953, p. 28), his position was that 'It is within the bounds of possibility that the wise use of modern methods of communication and training to promote higher age at marriage and the practice of birth control would bring a considerable reduction of the birth rate even in peasant societies.' Birth control interventions were increasingly attempted, many with the assistance of the Population Council during Notestein's presidency. Fertility in most of the developing world has since declined, earlier than the OPR researchers of the 1940s thought likely, but neither Notestein nor others at OPR incorporated the new vital revolution into a revised demographic transition theory. Perhaps Notestein's stature and influence rested most on the extraordinary extent to which most of his colleagues and students retained the consensus which had been reached in OPR during the 1940s and which he expressed so well.

Revisionists

The movement of Notestein and others toward being more involved in policy formulation and less the detached intellectual observer disquieted parts of the academic demography community. Dennis Hodgson (1983, pp. 2–3) saw this change as being unfortunate, from the scientific point of view, arguing that 'There is a tendency ... for crisis research to be policy research rather than to be strictly academic in nature.' He contrasted this with an earlier, perhaps happier period, when 'Understanding, not change, was the goal.' Simon Szreter (1993) went much further and attributed the changes in Notestein's and his colleagues' interpretations and attitudes to demographic change to their appreciation of the USA's strategic interests during the Cold War period.

Hodgson's criticism probably would not have worried Notestein. He had always held that the overriding purpose of research was to help the human race, writing in his last published article that he was 'both a foundation and university man,' which he did not find incompatible, because most social scientists are 'reformers who are basically interested in the interplay of evolving technology, socioeconomic institutions, and population change as major determinants of the human condition' (Notestein, 1982, pp. 651–687). He had never ruled out policy interventions but regarded them as not only desirable but even to be expected, not only on the mortality side of demographic transition but also on its fertility side. He regarded them as social experiments worth undertaking and studying. His concluding chapter to Stix and Notestein (1940, pp. 150ff.) referred to the need to do something about the high fertility, poverty, and human suffering of half the world, and then discussed possible family planning interventions in distressed areas of the USA. His acceptance of assistance from the League of Nations, the Department of State, the World Bank, and others shows his determination to fund OPR adequately to investigate to its fullest the demographic transition as it had occurred in Western European populations and as it moved around the world. Indeed, if he had not done so, we would have heard little of him or of OPR. Even his later papers place less emphasis on saving the world than on giving priority to demographic research and its funding. The charge that Notestein gave priority to Asia because it was the prime area of American strategic interest, especially as China became communist (Szreter, 1993, pp. 675–680), overlooks the fact that Notestein regarded Asia as the main continent facing potential Malthusian pressures throughout his academic life.

Frank Notestein died at 80 years of age, of complications arising from emphysema, on 18 February 1983 in retirement at Newton, Pennsylvania, only a short drive from Princeton.

See also: Demographic Measurement: General Issues and Measures of Fertility; Demographic Models; Demography: History Since 1900; Eugenics as a Basis of Population Policy; Family Planning: Program Effects; Fertility Theory: Embodied-Capital Theory of Life History Evolution; Fertility Transition: Cultural Explanations; Fertility Transition: Economic Explanations; Human Ecology; Population Cycles, Formal Theory of; Population Dynamics: Mathematic Models of Population, Development, and Natural Resources.

Bibliography

Caldwell, J., Caldwell, P., 1986. Limiting Population Growth, and the Ford Foundation Contribution. F Pinter, London.

Carr-Saunders, A.M., 1936. World Population: Past Growth and Present Trends. Clarendon Press, Oxford, UK.

Coale, A.J., 1979. Notestein, Frank W. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences: Bibliographical Supplement. Free Press, New York.

Himes, N.E., 1936. Medical History of Contraception. Williams & Wilkins, Baltimore, MD.

Hodgson, D., 1983. Demography as a social science and policy science. Population and Development Review 9, 1–34.

Notestein, F.W., 1938. A demographic study of 38,256 rural families in China. Milbank Memorial Fund Quarterly 16, 57–79.

Notestein, F.W., 1943. Some implications of population change for post-war Europe. Proceedings of the American Philosophical Society 87, 165–174.

Notestein, F.W., 1945. Population—The long view. In: Schultz, T.W. (Ed.), Food for the World. University of Chicago Press, Chicago, IL, pp. 36–57.

Notestein, F.W., 1953. Economic problems of population change. In: Proceedings of the Eighth International Conference of Agricultural Economists. Oxford University Press, London, pp. 13–31.

Notestein, F.W., 1982. Demography in the United States: A partial account of the development of the field. Population and Development Review 8, 651–687.

Ryder, N.B., 1984. Frank Wallace Notestein (1902–1983). Population Studies 38, 5–20.

Stix, R.K., Notestein, F.W., 1940. Controlled Fertility: An Evaluation of Clinic Service. Williams & Wilkins, Baltimore, MD.

Sydenstricker, E., Notestein, F.W., 1930. Differential fertility according to social class. Journal of the American Statistical Association 25 (169), 9–32.

Szreter, S., 1993. The idea of demographic transition and the study of fertility change: A critical intellectual history. Population and Development Review 19, 659–701.

Thompson, W.S., 1929. Population. American Journal of Sociology 34, 959–975.

Willcox, W., 1916. The nature and significance of the changes in the birth and death rates in recent years. Journal of the American Statistical Association 15 (113), 1–15.

Number, Grammatical

Guido G Corbett, University of Surrey, Guildford, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 9, pp. 6340–6342, © 2001, Elsevier Ltd.

Abstract

Number is a much more challenging category than is generally believed. While in languages like English it is only a nominal category, for quantifying the number of entities, in many languages it is rather a verbal category, which indicates the number of events. Number is an obligatory category in English, and yet in many languages the use of the plural is not obligatory and it is used only when it is important to mark number. In English the majority of nominals can mark number, from the personal pronouns and nouns denoting persons, down to those denoting inanimate objects, even including some nouns denoting abstracts. Many languages restrict the number opposition to fewer nominals, namely those which are high in animacy. Animacy also helps determine the varying semantics of the category of number. While languages like English are restricted to singular and plural, many languages have a wider range of possibilities. We wish to establish how many number values natural languages may have, and which they are. Besides the well-known values, namely singular, plural and dual, there are also trials and paucals, and the largest systems have five members in all.

Number is often taken to be a relatively straightforward category. It is often used as the textbook illustration of an inflectional category, and the situation in languages like English is believed by many to be the normal if not the only one. English makes a semantic distinction, and a formal distinction between, for instance, *cat* and *cats*, *child* and *children*. In fact the number category reveals many surprises, and the system found in English is a rather extreme one, when seen within the space of possibilities established from a survey of the languages of the world. We shall consider characteristics of the English system in turn, showing some of the ways in which it is typical or atypical. We shall give greatest prominence to the question of number values.

Nominal and Verbal Number

In English number is a nominal category, it involves the distinction between instances such as *cat* and *cats*. When it appears on verbs, this is by agreement with a subject noun phrase, as in *the children were playing happily*. The plural verb still indicates the number of children, and not the number of playing events. There are many languages which have verbal number, that is, where the number of the verb indicates the number of events (see Durie, 1986 for examples).

Number as an Obligatory Category

Number is an obligatory category in English. If someone says, *there's a visitor waiting to see you*, when there are several, this is misleading and inappropriate. In many languages, however, the use of the plural is not obligatory, but is used when it is important to mark number. One example is the Cushitic language Bayso:

(1) Bayso (Dick Hayward, personal communications, Corbett and Hayward, 1987)

lúban	*foofe*
lion.GENERAL	watched.1SG

'I watched lion' (it could be one, or more than that)

Another language, from many examples, is the Arawak family language Bare of Venezuela and Brazil (Aikhenvald, 1995, pp. 18–19).

The Nominals Involved in the Number System

In English the majority of nominals can mark number, from the personal pronouns and nouns denoting persons (like *woman)* down to those denoting inanimate objects like *door*, and even to some denoting abstracts like *hour*. Many languages restrict the number opposition to fewer nominals, namely those which are high in animacy (Smith-Stark, 1974). Thus in Slave (an Athapaskan language spoken in parts of the Northwest Territories, British Columbia, and Alberta, Canada) number may be marked on the noun phrase provided the noun denotes a human or a dog (Rice, 1989, p. 247).

The Semantics of Number

As long as we stay with innocent examples like English *cat* and *cats*, the semantics of number look clear, since *cats* is appropriate for more than one instance of *cat*. In fact there are many concealed difficulties: a concise overview of work on number within formal semantics can be found in Link (1998), who shows how the study of plurals relates to different movements in linguistic semantics, and provides a substantial bibliography.

If we move from ordinary plurals to consider *I* versus *we*, it becomes clear that here we have an 'associative' use of the plural; *we* is most often used for 'I and other(s) associated with me.' At the other end of the Animacy Hierarchy, nouns like *coffee* have a plural, provided they are recategorized: if *coffee* denotes a type, or a typical quantity of coffee, then there is an available plural, as in: *she always keeps three different coffees*, and *we've ordered two teas and three coffees*.

Number Values

A striking way in which English is limited in its number system is in terms of the number values. In English we can distinguish

only between singular and plural. Many languages have more distinctions, and it is these on which we shall concentrate.

The Dual

A common system is one in which singular and plural are supplemented by a dual, to refer to two distinct real world entities. Examples are widespread. We will take Central Alaskan Yup'ik, as spoken in the southwestern Alaska (Corbett and Mithun, 1996); the number forms are: singular *qayaq* 'kayak,' dual *qayak* 'two kayaks,' and plural *qayat* 'three or more kayaks.' The addition of the dual has an effect on the plural, which is now used for three or more real world entities, and more generally a change in system gives the plural a different meaning.

The Trial

While the dual refers to two distinct real world entities, the trial is for three, and it occurs in systems with the number values: singular, dual, trial, and plural. This system is found in Larike, a Central Moluccan language spoken on the western tip of Ambon Island, Central Maluku, Indonesia. (Central Moluccan forms part of the Central Malayo-Polynesian subgroup of Austronesian.) This example is from Laidig and Laidig (1990, p. 7):

(2) *Kalu iridu-ta-peu* *au-na-wela*
if 2.TRIAL-NEG-go 1.SG-IRR-go.home
'If you three don't want to go, I'm going home.'

In this language the dual and trial forms can be traced back to the numerals 'two' and 'three,' and the plural comes historically from 'four.' Such developments are fairly common in Austronesian languages. Sometimes the term 'trial' is used according to the form of the inflections (derived historically from the numeral three), even when the forms are used currently for small groups including those greater than three (and so are paucals) and sometimes the term is used according to meaning (for genuine trials). The Larike trial is a genuine trial, used for exactly three (Laidig and Laidig, 1990, p. 92).

The Paucal

Paucals are used to refer to a small number of distinct real world entities, somewhat akin to the English quantifier 'a few' in meaning. There is no specific upper bound on the use of paucals (their lower bound will vary according to the system in which they are found). Let us return to the Cushitic language Bayso. Besides the general number forms (as in (1), where number is not specified), there are also singular, paucal, and plural. Here is the paucal:

(3) Bayso (Dick Hayward, personal communications, Corbett and Hayward, 1987)
luban-jaa *foofe*
lion-PAUCAL watched.1.SG
'I watched a few lions'

The paucal is used in Bayso for reference to a small number of individuals, from two to about six. (Bayso has this system in nouns but not in pronouns.) The Bayso system is rare; normally the paucal is found together with a dual, so that the number values are singular, dual, paucal, and plural. This widespread system is found, for example, in Yimas, a Lower Sepik language spoken in the Sepik Basin of Papua New Guinea (Foley, 1991, p. 216).

The Largest Systems

It is worth asking what is the largest system of number values. There are claims in the literature for languages with a quadral, for four distinct real world entities, thus giving a system of five number values. It is established that five number values can indeed be found, but the evidence specifically for a quadral is less clear. All the claims come from within the Austro-nesian family and the best documented case is Sursurunga (Hutchisson, 1986, and personal communications), spoken in southern New Ireland. The forms labeled quadral are restricted to the personal pronouns, but are found with all of them, the first person (inclusive and exclusive), the second, and third (**Table 1**).

Here is an example of a quadral form in use (we will retain the traditional label 'quadral,' until we have justified replacing the term for Sursurunga):

(4) Sursurunga (Hutchisson, 1986 and personal communications)
gim-hat *káwán*
1.EXCL-QUADRAL maternal.uncle:nephew/niece
'we four who are in an uncle-nephew/niece relationship'

The 'quadral' has two additional uses. First, in Sursurunga plural pronouns are never used with terms for dyads (kinship pairings like uncle-nephew/niece). Here the 'quadral' is used, for a minimum of four, and not just for exactly four (Hutchisson, 1986, p. 10). Second, in hortatory discourse, the speaker may use the first person inclusive 'quadral' in order to suggest joint action including the speaker, even though more than four persons are involved. These two special uses of the 'quadral' account for the majority of instances of its use. According to its meaning, the term 'quadral' is not appropriate. Rather the forms would be better designated 'paucal.'

Let us now consider the rest of the system of Sursurunga (data and judgments from Don Hutchisson, personal communications). The dual is used quite strictly for two people. It is also used for the singular when the referent is in a taboo relationship to the speaker. This is a special use; the dual's main use is as a regular dual. The trial will be used for three. However, it is also used for small groups, typically around three or four, and for nuclear families of any size. It is

Table 1 Emphatic pronouns is Sursurunga

	singular	*dual*	*'triai'*	*'quadrai'*	*plural*
1 exclusive	iau	giur	gimtul	gimhat	gim
1 inclusive	—	gitar	gittul	githat	git
2	iáu	gaur	gamtul	gamhat	gam
3	-i/on/ái	diar	ditul	dihat	di

Note: á is used to indicate schwa (ə)
Source: Hutchisson, D., 1986. Sursurunga pronouns and the special uses of quadral number. In: Wiesemann, U. (Ed.), Pronominal Systems (Continuum 5). Narr, Tübingen, Germany, pp. 217–255 and personal communications

therefore not strictly a trial, rather it could be labeled a paucal (an appropriate gloss would be 'a few'). We noted earlier that this is a frequent development for the trial. The 'quadral,' as we have just seen, is used primarily in different ways with larger groups, of four or more (an appropriate gloss would be 'several'). This too would qualify as a paucal; we therefore have two paucals in Sursurunga, a paucal (traditionally 'trial') and a greater paucal (traditionally 'quadral'). The plural, as we would expect, is for numbers of entities larger than what is covered by the greater paucal, however, there is no strict dividing line (certainly not at the number five). Using semantic labels, we might represent the system in Sursurunga as having: singular, dual, paucal, greater paucal, and plural. Sursurunga has a well-documented five-value number category. There are certainly other languages with five number values; we do not have such detailed information as for Sursurunga and there is no certain case of a quadral: it seems that in all cases the highest value other than plural in such systems can be used as a paucal. Besides the split in the paucal, we may also find a split in the plural, with 'greater plurals' of different types, as for instance in Mokilese (Harrison, 1976). Greater plurals may imply an excessive number or else all possible instances of the referent. They are as yet poorly researched, except in a few languages.

Conclusion

Number is indeed a much more complex category than is generally recognized, and the English system, though a common one, is at one extreme of the typological space. We noted that number may be a verbal category as well as a nominal one (but the possibilities are rather more restricted for verbal number). We also saw that number need not be obligatory, that the range of items which distinguish number varies dramatically from language to language, that the semantics of number are affected by the particular item involved, and that the values of the number category can be just two or, in various combinations, up to five. More complex issues, which we have not been able to address here, include various interactions, notably between the forms of marking of number, the particular nominals which distinguish number in a given language, and the systems of number values. For these issues see Corbett (2000).

See also: Gender: Grammatical; Number Systems, Evolution of.

Bibliography

Aikhenvald, A., 2000. 1995 *Bare* (Languages of the World/Materials 100). Cambridge University Press, Cambridge, UK. Lincom Europa, Munich, Germany.
Corbett, G.G., 2000. Number. Cambridge University Press, Cambridge, UK.
Corbett, G.G., Hayward, R.J., 1987. Gender and number in Bayso. Lingua 73, 1–28.
Corbett, G.G., Mithun, M., 1996. Associative forms in a typology of number systems: Evidence from Yup'ik. Journal of Linguistics 32, 1–17.
Durie, M., 1986. The grammaticization of number as a verbal category. In: Vassiliki, N., VanClay, M., Niepokuj, M., Feder, D. (Eds.), Proc. of the Twelfth Annual Meeting of the Berkeley Linguistics Society: February 15–17, 1986. University of California, Berkeley, CA, pp. 355–370.
Foley, W.A., 1991. The Yimas Language of New Guinea. Stanford University Press, Stanford, CA.
Harrison, S.P., Albert, S.Y., 1976. Mokilese Reference Grammar. University Press of Hawaii, Honolulu, HI.
Hutchisson, D., 1986. Sursurunga pronouns and the special uses of quadral number. In: Wiesemann, U. (Ed.), Pronominal Systems (Continuum 5). Narr, Tübingen, Germany, pp. 217–255.
Laidig, W.D., Laidig, C.J., 1990. Larike pronouns: Duals and trials in a Central Moluccan language. Oceanic Linguistics 29, 87–109.
Link, G., 1998. Ten years of research on plurals—where do we stand? In: Hamm, F., Hinrichs, E. (Eds.), Plurality and Quantification. Kluwer, Dordrecht, The Netherlands, pp. 19–54.
Rice, K., 1989. In: A Grammer of Slave. Mouton de Gruyter, Berlin.
Smith-Stark, T.C., 1974. The plurality split. In: La Galy, M.W., Fox, R.A., Bruck, A. (Eds.), Papers from the Tenth Regional Meeting, Chicago Linguistic Society, 19–21, April 1974. CLS, Chicago, pp. 657–671.

Numbers and Number Sense

Ian M Lyons, University of Western Ontario, London, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews several canonical signatures of number processing and the proposed theoretical interpretations of these signatures. It examines the biological origins of number sense in humans and nonhuman species alike, how to conceptualize the notion of numbers without symbolic form, and the role that number sense may play in human development. Symbolic numbers are discussed in terms of their history and the leading ideas about how the human brain assigns meaning to these crucial building blocks of modern human society.

Introduction

It is difficult to think of an aspect of modern society that has not been fundamentally shaped by the application of mathematics. Numbers, it may be argued, provide the basic cognitive scaffolding around which mathematical thinking is eventually constructed. The study of how numbers are represented and manipulated at both cognitive and neural levels thus provides a window into how the brain acquires and hones perhaps one of its most effective tools for shaping the world around us: mathematics. At the broadest level, the study of numerical cognition has divided numbers into two categories: symbolic and nonsymbolic. Symbolic numbers typically refer to exact quantities and are typically studied in written (e.g., Indo-Arabic numerals) or spoken (number words) form. Nonsymbolic magnitudes tend to be more approximate in nature – for instance, one may rapidly estimate which of two groups of dots contains more objects without counting or explicit reference to a symbolic numerical label. Number processing in both forms – symbolic and nonsymbolic – has been extensively studied in human children and adults, as well as many different nonhuman species. Methodological approaches tend to be behavioral (e.g., response times and/or error rates in numerical judgments) or neural (e.g., noninvasive neuroimaging in humans, as well as single-cell recording in nonhuman primates).

In the sections that follow, I review several canonical signatures of number processing and the proposed theoretical interpretations of these signatures. I then examine the biological origins of number sense in humans and nonhuman species alike, how to conceptualize the notion of numbers without symbolic form, and the role that number sense may play in human development. Finally, I discuss symbolic numbers in terms of their history and the leading ideas about how the human brain assigns meaning to these crucial building blocks of modern human society.

Signatures of Number Processing

There are several behavioral and neural signatures of number processing commonly found and discussed throughout the field of numerical cognition. Interpreting these effects has helped shape the ongoing debate about how numbers are represented and processed in the brain.

Numerical Distance Effect

Perhaps the most commonly cited numerical signature is the numerical distance effect. Note that the term 'distance' is something of a misnomer here, as it refers strictly to the absolute numerical difference between numbers, and not the physical distance between them. When an individual is asked to decide which of two numbers is numerically greater (commonly referred to as a numerical comparison task), performance depends systematically on the numerical difference between the two numbers. Performance tends to be worse (longer response times and higher error rates) when the numbers being compared are numerically close (e.g., performance on 5 vs 6 is worse than on 3 vs 8). This effect has been shown in both adults (Moyer and Landauer, 1967; Buckley and Gillman, 1974) and children (Sekuler and Mierkiewicz, 1977; Rubinsten et al., 2002). Moreover, neuroimaging studies have shown that brain responses – particularly in the intraparietal sulcus (IPS) – are modulated specifically by numerical distance during numerical comparison tasks (Pinel et al., 2001; Kaufmann et al., 2005; Ansari et al., 2006; Mussolin et al., 2013).

A related effect uses priming instead of comparison as a means of measuring the influence of numerical distance on brain and behavioral responses. In a priming paradigm, participants see a very brief stimulus (called a prime) immediately prior to completing some simple task. For example, when asked to name a number, numerically close primes lead to more facilitation (faster response times) than do numerically distant primes (Reynvoet et al., 2002). This effect is found in children as young as first grade, and is similar to that seen in adults (Reynvoet et al., 2009). The numerical distance between prime and target numbers also modulates neural responses in parietal cortex (Notebaert et al., 2010). In sum, numerical distance effects are present in both adults and children, influence both behavioral and neural responses, and are detectable whether measured via comparison or priming.

Numerical Ratio Effect

Another numerical signature is the numerical ratio effect, which refers to the ratio between two numbers, $(\frac{n_1}{n_2})$, typically where $n_1 < n_2$. (Notice that the ratio effect differs from the distance effect primarily in that the former takes into account the numerical size of the numbers in question. For example the distances between the pairs (2, 3) and (8, 9) are both 1;

 http://dx.doi.org/10.1016/B978-0-08-097086-8.57031-7

however, they differ in ratio $\left(\frac{2}{3} \neq \frac{8}{9}\right)$.) In number comparison tasks (subjects decide which of the two numbers is numerically greater), performance tends to be worse as ratio approaches 1 (e.g., Lyons and Beilock, 2009). Six-month-old infants' looking times are influenced by numerical ratio. For example, if an infant sees 16 items repeatedly displayed on a screen, the infant will look longer at the screen when it changes to 32 items (ratio = 0.5). However, the child will not increase its looking time when the screen changes from 16 to 24 items (ratio = 0.67) (Xu et al., 2005). The ratio at which children can distinguish between such perceptual magnitudes becomes more precise (i.e., closer to 1) as children get older (Halberda and Feigenson, 2008), asymptoting at around $\frac{10}{11}$ in adults (Halberda and Feigenson, 2008; Pica et al., 2004).

The method most commonly used for assessing neural responses as a function of numerical ratio is called adaptation. An individual whose brain is being imaged will passively view several repetitions of a given number (e.g., 6). This passive repetition is associated with a steady reduction in the neural response to the repeated (standard) number; that is, the relevant brain regions are thus 'adapted' (i.e., habituated) to the repeated stimulus (Grill-Spector et al., 2006). Researchers will then intermittently present different (deviant) numbers (e.g., one of 4, 5, 7, 8, etc.) and measure the resulting neural response. One can then plot the brain responses to the deviant numbers as a function of their numerical ratio relative to the standard number. Neural responses in the parietal lobe (especially the IPS) tend to be systematically greater as the ratio between the standard and the deviant diverges from 1 (Piazza et al., 2004; Holloway et al., 2013).

A close derivative of numerical ratio is sometimes referred to as the numerical Weber fraction, or w, which has become an especially popular measure of numerical comparison performance (e.g., Halberda et al., 2008, 2012). One reason for this is that w is an especially strong predictor of how accurately individuals compare perceptual magnitudes (such as arrays of dots). When fitting sample means, the R^2 fit between estimated w and actual error rates is often upward of 0.9 for children, monkeys, and adults (Pica et al., 2004; Cantlon and Brannon, 2007; Halberda and Feigenson, 2008; Lyons and Beilock, 2011).

A few points of caution with respect to w are worth noting, however. First, w is very closely related to numerical ratio. This is because the formula for estimating w when comparing two numbers n_1 and n_2 ($n_1 < n_2$) relies on the value $k = \frac{n_2 - n_1}{\sqrt{2(n_2^2 + n_1^2)}}$, which is very highly correlated with the simple ratio between n_1 and n_2 $\left(\frac{n_1}{n_2}\right)$ (for full details on the formula for estimating w, see Pica et al., 2004). For example, in a stimulus set containing all possible (nonequal) number pairs from the range 1–10, k and $\frac{n_1}{n_2}$ will be correlated at $r = -0.999$. Thus, an estimate of w should be nearly identical to a ratio effect. Second, w can only be assessed with respect to error rates and not response times. This may be especially problematic when examining symbolic number comparisons (i.e., using Indo-Arabic numerals) in adults, where error rates may be very close to zero, and hence the majority of meaningful performance variability is often found in response times. Finally, Inglis and Gilmore (2014) recently showed that test/retest reliability for both w and ratio effects is poor (though see also Price et al., 2012; for issues regarding reliability of symbolic distance effects, see Maloney et al., 2010). On the other hand, Inglis and Gilmore (2014) showed that reliability for overall average accuracy was quite high; hence, if one is concerned with individual differences (e.g., relating numerical performance with other measures, such as math achievement or IQ), then one may be better off forgoing measures related to numerical distance or ratio altogether.

In sum, like numerical distance, the influence of numerical ratio on both brain and behavioral responses has been identified across a wide range of ages, measures, and experimental paradigms. There has been strong recent interest in using w – a close cousin of numerical ratio. While w yields strong fit with performance in limited cases (e.g., error rates when comparing perceptual magnitudes), the usefulness of w is quickly exhausted – for instance, when considering individual differences or cases in which high accuracy rates are expected.

Interpretations

The most common interpretation given to both numerical distance and ratio effects (including w) is that of representational precision. In this view, distance/ratio effects arise because the representations of numerically close numbers are more likely to overlap than are those of numerically distant numbers (Dehaene and Changeux, 1993; Verguts and Fias, 2004; Nieder and Dehaene, 2009). Greater representational overlap makes it more difficult to distinguish between two numbers; for instance, when deciding which of the two numbers is numerically greater in a comparison task, or detecting numerical change in an adaptation or looking-time paradigm. This account explains the distance effect because the representations of 5 and 6 overlap to a greater degree than do those of 3 and 8. The ratio effect is accounted for by the assumption that the representations of larger numbers overlap to a greater degree than do smaller numbers; hence, while 8 and 9 overlap to a greater extent than do 7 and 9 (distance effect), 8 and 9 also overlap to a greater extent than do 2 and 3 (ratio effect).

Empirical justification for these assumptions has arisen primarily from work with perceptual magnitudes (e.g., estimating – without explicitly counting – the number of dots in an array, or which of the two arrays of dots contains more dots; Indow and Ida, 1977). Nieder and colleagues (for a review, see Nieder, 2005) recorded single-cell activity from individual prefrontal and parietal neurons in macaque monkeys while monkeys briefly held a quantity of dots in active memory. The authors found that some of the neurons thus sampled were tuned to specific quantities of dots. For instance, a given neuron might be most active (measured in terms of action-potential spike density) when the monkey temporarily held in memory three dots. Interestingly, the neuron would also fire when the monkey held in mind two or four dots; however, the net amount of activity was less than for the preferred quantity (three, in this example), and activity was still less for one or five dots. In this way, Nieder and colleagues showed that there are numerically tuned neurons in the brain, and that these representations are systematically imprecise (i.e., analog). Neuronal responses

to different numbers overlap with one another (implying a degree of representational imprecision), and this overlap varies systematically as a function of the numerical distance between them (see **Figure 1** for a simulated example based on human data in Merten and Nieder, 2009). Furthermore, both the neuronal evidence discussed above as well as behavioral evidence in humans (e.g., Nieder and Merten, 2007; Merten and Nieder, 2009) show that numerosity and tuning width are positively correlated. Specifically, as the preferred number increases, the width (i.e., systematic imprecision) of that preference also increases (note that this width is in fact constant on a log–log plot). This means that 8 and 9 will overlap more than 2 and 3, which, as discussed above, is consistent with the numerical ratio effect. Finally, neuroimaging evidence in humans also shows ratio-dependent tuning in parietal cortex – at least for nonsymbolic magnitudes (Piazza et al., 2004).

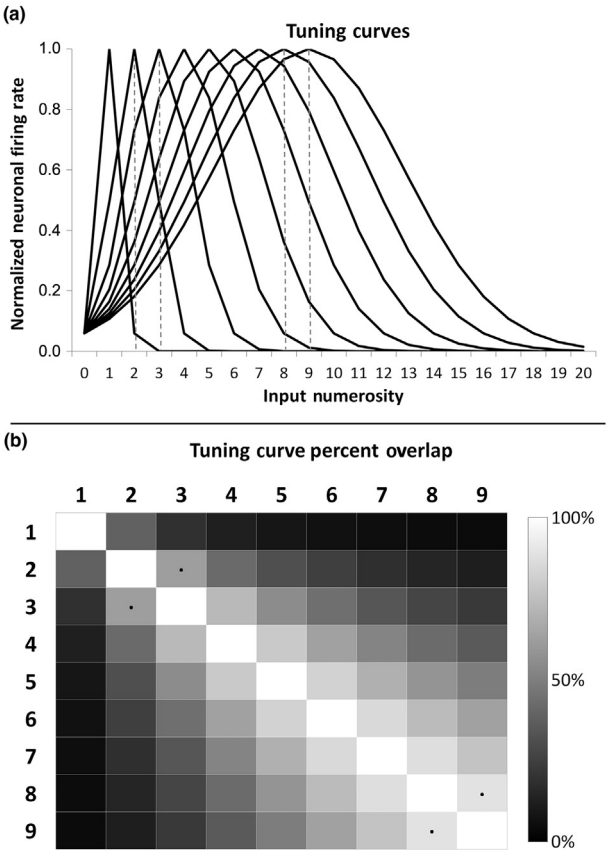

Figure 1 (a) Shows simulated tuning curves based on monkey neuronal data (for a review, see Nieder, 2005) and human and monkey behavioral data (Merten and Nieder, 2009). The *x*-axis is the number of dots in the stimulus, and the *y*-axis is the simulated firing rate. Each curve shows the simulated firing rate for a neuron that is tuned to the number over which it is centered (dashed lines are provided to help identify the curves for 2, 3, 8, and 9). That is, the relative firing rate for a neuron tuned to 9 fires maximally for 9 dots. Notice that this simulated '9-neuron' also fires for 8 dots, but a bit less so than for 9; indeed the firing rate falls off as the presented number of dots (*x*-axis) increasingly differs from 9. Finally, notice that these curves fall off more steeply for neurons tuned to smaller numbers. It is in this way that the width of tuning curves is positively correlated with the number being represented. (b) Plots the degree to which each pair of curves overlap with one another: $\frac{n_1 \cap n_2}{n_1 \cup n_2}$. Overlap proportions for 2 and 3 and 8 and 9 are marked with a dot, which demonstrates the ratio effect.

To briefly summarize, neural evidence is consistent with the interpretation that numerical ratio and distance effects are indicative of an underlying analog magnitude representation: numerically closer numbers show greater representational overlap (distance effect), and this overlap increases as the size of the numbers in question increase (ratio effect). However, the vast majority of this evidence comes from experiments that examined approximate, nonverbal, nonsymbolic, perceptual magnitudes (e.g., dot arrays; hereafter, simply 'nonsymbolic magnitudes'). When we look at symbolic numbers, such as Indo-Arabic numerals, the evidence is not as clear.

For symbolic numbers such as Indo-Arabic numerals, Verguts et al. (2005) proposed an alternative explanation, beginning with the observation that smaller numbers occur with far greater frequency than do larger numbers (both in speech and in written records; Dehaene and Mehler, 1992). Verguts et al. (2005) created a distributed model of number representation with constant representational precision (equally wide tuning curves) for all numbers, but matched number frequency during model training roughly to that reported by Dehaene and Mehler (1992). When performing a standard number comparison task, the model simulated a ratio effect. In other words, to explain the ratio effect, at least for symbolic numbers, one need not assume representational overlap that increases with the number being represented (see also Verguts and Van Opstal, 2005, for empirical verification of this model).

Sasanguie et al. (2011) have since shown that comparison and priming numerical distance effects are uncorrelated. These results essentially force one to choose between comparison and priming distance effects as the 'true' indicator of representational overlap, with the other indexing some other aspect of processing. To this end, Van Opstal et al. (2008) dissociated the notion of a comparison distance effect entirely from the notion of representational overlap. They did so by showing that comparison distance effects obtain for both number and letter comparisons (when asked 'which letter comes later in the alphabet,' participants were slower to compare, e.g., E vs F than C vs H), but the priming distance effect was found only for numbers. Hence, Van Opstal et al. argued that the priming distance effect is the superior index of number representation, and that the comparison distance effect is primarily due to more domain-general mechanisms, such as response mapping.

Another puzzling aspect of symbolic distance effects occurs when participants are asked to judge relative numerical order. Under such task demands, the effect of numerical distance is in fact reversed on trials where the numbers are in the correct order. For example, participants are faster and more accurate when confirming that 4 – 5 – 6 is in order than when confirming that 3 – 5 – 7 is in order (Turconi et al., 2006; Franklin, 2009; Franklin and Jonides, 2009; Lyons and Beilock, 2013). Crucially, no such reversal of the distance effect is found when participants judge the relative order of nonsymbolic perceptual magnitudes (arrays of dots; Lyons and Beilock, 2013). One interpretation is that

highly familiar associations between symbolic numbers (e.g., from the counting sequence) in essence trump the standard influence of numerical distance. Regardless of one's interpretation, the reversed ordinal distance effect demonstrates that simply changing task demands can qualitatively alter the influence of numerical distance on performance, which further calls into question the assumption that symbolic numerical distance effects are a straightforward index of symbolic number representation.

In sum, a common interpretation of numerical ratio and distance effects is that they index underlying representational overlap. In the case of nonsymbolic magnitudes (such as dot arrays), empirical evidence strongly supports this interpretation. However, with respect to symbolic number representation, several lines of evidence converge to render this assumption questionable at best.

Biological Origins – Number Sense

As noted above, adult and infant humans are capable of determining which of the two arrays of objects (e.g., dots) contains more items, even without counting (Buckley and Gillman, 1974). Adults' ability to do so depends on the ratio between the relative numbers of items in each of the two arrays – with a typical limit of about $\frac{10}{11}$ (Cantlon and Brannon, 2007). This pattern of results is robust across cultures and obtains even in individuals whom possess no words for numbers beyond three or four, or other kinds of formal mathematics (Pica et al., 2004). Individuals both with and without formal math education are capable of adding and subtracting such approximate magnitudes as well (Pica et al., 2004; Barth et al., 2006). The ability to process nonsymbolic magnitudes is commonly referred to as 'number sense' (Dehaene, 1997).

Number Sense across Species

When estimating changes in nonsymbolic magnitude, humans and (rhesus macaque) monkeys show a similar pattern of behavior (Merten and Nieder, 2009). A common experimental paradigm used to train and record neural and behavioral responses in monkeys is often referred to as 'delayed match-to-sample,' and it is similar to dot comparison tasks described previously. The difference is that arrays are presented sequentially. An individual sees a single array of dots, which is followed shortly thereafter by a second array. One's task is to determine if the second array matches the first in terms of the number of dots. The benefit of this approach is that one can more transparently record (and hence model) the performance function ('tuning curve') for a specific number – i.e., the first number presented and held in mind during the delay period. (In a standard number comparison task where both numbers are presented simultaneously, it is unclear whether performance and/or neural activity reflects the representation of one number, both numbers, the comparison process, or some combination of all three.) For a specific quantity of dots (e.g., 30), one can plot the probability of a 'same' judgment as a function of the quantity of dots in the second array. Even for quantities as high as 30, both monkeys and adult humans reliably report 'same' most often when indeed the second array contained 30 dots. This demonstrates that both species are capable of processing nonsymbolic magnitudes. Moreover, as the number of dots in the second array shifts away from 30, the probability of a 'same' response decreases systematically as a function of the numerical ratio of the two arrays. In other words, the closer the ratio between the two arrays is to 1, the more likely an individual is to perceive the two arrays as in fact containing the same number of items (thus demonstrating, in both species, the ratio effect, as discussed previously). Interestingly, the function that best describes the decrease in the probability of a 'same' response corresponds to a positively skewed distribution whose overall width increases as the target quantity (the first array) increases (somewhat akin to the curves plotted in **Figure 1**). That is, one can imagine a quasi-normal distribution centered on the target quantity, with a stretched out tail to the right, and a relatively compact tail to the left. Furthermore, the distribution (in both directions) centered on 30 will be wider than that for 25, and that for 25 will be wider for 20, and so on. Interestingly, the function that describes this tuning is quite similar in both humans and monkeys, which suggests a potential common mechanism across species, at least for detecting change in nonsymbolic magnitudes. (Plotted on a linear scale, the distribution widens at a rate of about 0.421 units per unit increase in the target number. On a log-scale, distribution widths are both symmetrical and constant with target number change (Merten and Nieder, 2009).)

The similarity between human and monkey processing of nonsymbolic magnitudes extends to other nonsymbolic magnitudes tasks, such as simultaneous nonsymbolic comparison (Brannon and Terrace, 2002), nonsymbolic ordering (Cantlon and Brannon, 2006), and nonsymbolic arithmetic (Cantlon and Brannon, 2007). Indeed, this ability is common to many species, including dolphins, macaques, capuchins, baboons, wolves, canids, mice, parrots, crows, robins, bees, fish, and even beetles (Agrillo et al., 2009; Armstrong et al., 2012; Baker et al., 2012; Barnard et al., 2013; Beran et al., 2012; Carazo et al., 2012; Nieder, 2005; Pepperberg, 2013; Piffer et al., 2013; Smirnova et al., 2000; Utrata et al., 2012; Yaman et al., 2012 – for reviews, see Agrillo and Beran, 2013; Pahl et al., 2013). In sum, it is clear that the ability to process nonsymbolic magnitudes is not unique to humans, and it is surmised that our ability to do so is part of an innate capacity with deep evolutionary roots (Hubbard et al., 2008).

Number Sense in Early Development

The ability to distinguish visual (Xu and Spelke, 2000) and auditory (Lipton and Spelke, 2003) nonsymbolic magnitudes is present in children as young as 6 months of age. The topography of neural signatures can distinguish between neural pathways related to object identification and nonsymbolic magnitudes; further, this topography is similar in adults, 4-year-olds, and 3-month-old infants, with sensitivity to magnitude centered on the IPS (Cantlon et al., 2006; Izard et al., 2008). Combined with the observation that something akin to a number sense is present across a range of

species, the results from human developmental studies suggest that the basis of humans' ability to process nonsymbolic magnitudes is present from a very early developmental stage, does not require formal math training, and may even be innate (Feigenson et al., 2004; Cordes and Brannon, 2008; Nieder and Dehaene, 2009).

This insight has led various researchers to hypothesize that children's number sense forms the foundation on which more sophisticated, culturally acquired mathematical skills rest (e.g., Dehaene, 1997). Indeed, over a dozen papers in the last several years have demonstrated a correlation between individual differences in adults', children's, and even infants' ability to process nonsymbolic magnitudes with performance on a wide range of formal math achievement tests (Bonny and Lourenco, 2013; Desoete et al., 2012; Gilmore et al., 2010; Halberda et al., 2008, 2012; Libertus et al., 2011, 2012, 2013; Lonnemanna et al., 2011; Lourenco et al., 2012; Lyons and Beilock, 2011; Mazzocco et al., 2011a,b; Piazza et al., 2010; Starr et al., 2013; for a review, see Feigenson et al., 2013). Training on approximate arithmetic (using dot arrays) has also been shown to preferentially improve symbolic math achievement scores (Park and Brannon, 2013, 2014; Hyde et al., 2014). These results point to the exciting prospect that numerical cognition may be a prime example of how evolutionarily ancient neural systems are co-opted by cultural inputs (number symbols) to serve a cognitive ability (mathematics) that is crucial to much of modern human life (Dehaene and Cohen, 2007).

On the other hand, it remains unclear precisely how this process occurs. With respect to training, other studies have shown that training on dot comparison tasks does not improve symbolic math scores (Wilson et al., 2006; Dewind and Brannon, 2012; Park and Brannon, 2014); thus, it is not so much improved processing of nonsymbolic magnitudes per se, but improved manipulation of those magnitudes in a mathematical context (arithmetic) that improves symbolic math performance (Park and Brannon, 2014). With respect to individual differences, the relation between nonsymbolic magnitude processing and math achievement has not replicated consistently (De Smedt et al., 2013), and a recent meta-analysis has shown that the relation between nonsymbolic magnitude processing and math achievement is relatively small ($r = 0.20$ for cross-sectional, and $r = 0.17$ for longitudinal studies; Chen and Li, 2014). Moreover, multiple studies have now shown that this correlation is often entirely eliminated once one controls for basic symbolic number processing abilities (such as number comparison, counting, or number ordering; Bartelet et al., 2014; Castronovo and Göbel, 2012; Fuhs and McNeil, 2013; Göbel et al., 2014; Holloway and Ansari, 2009; Kolkman et al., 2013; Lyons and Beilock, 2011, 2014; Sasanguie et al., 2013; Toll and Van Luit, 2014). In other words, the processing of number symbols appears to be more directly tied to more complex math processing (even after controlling for the fact that both are typically in the same visual format; e.g., Lyons and Beilock, 2011; Lyons et al., 2014).

The critical outstanding question, then, may be to understand precisely how nonsymbolic magnitudes aid young children as they first begin learning to comprehend and manipulate numerical symbols in mathematically meaningful ways. However, even in kindergarteners and preschoolers, the current literature remains mixed. For instance, contrast results confirming a relation between nonsymbolic and symbolic number processing in kindergarten or younger (Bonny and Lourenco, 2013; Jordan et al., 2009; Gilmore et al., 2010; Libertus et al., 2013; Mazzocco et al., 2011b; Starr et al., 2013) with results from studies showing largely the opposite (Bartelet et al., 2014; Fuhs and McNeil, 2013; Kolkman et al., 2013; Sasanguie et al., 2014; Toll and Van Luit, 2014). Indeed, there is even evidence to suggest that developmental influence runs in the opposite direction – that improvement in symbolic number abilities predict later accuracy in nonsymbolic magnitude comparison (Mussolin et al., 2014; see also Gelman and Gallistel, 2004). It may be that taking a broader view of how the key neural systems of young children change as a function of math education and math learning can provide crucial insight into this debate.

Number or Magnitude?

At this point, the reader may have noticed that I have used the term 'magnitude' when discussing the capacity to process approximate, nonverbal, perceptual, and nonsymbolic numbers or quantities. In recent years, the phrase 'approximate number system' (ANS) has appeared as an appellation for one's capacity to perform many of the nonsymbolic tasks described above (e.g., Halberda et al., 2006; Butterworth, 2010; Cantlon et al., 2009; McCrink and Spelke, 2010; Piazza, 2010). References to the ANS have become common in the literature; however, this has not occurred without some controversy. First, the scope of the term is not entirely clear, as it has been used refer to a range of tasks which are perhaps less consistently related to one another than might be expected if all relied upon a singular and unitary underlying construct (Gilmore et al., 2011; Inglis and Gilmore, 2014; though see also Price et al., 2012). Second, there remains some question as to whether what is being measured by these nonsymbolic tasks is indeed *numerical*.

This latter point of controversy is driven less by conflicting results than by a fundamental measurement limit: it is geometrically impossible to make all nonnumerical parameters strictly uncorrelated with number in a single trial. Take, for example, a nonsymbolic comparison task. On a given trial, one will see two dot arrays, usually arranged side-by-side. The task is to determine which array contains more items. Here, 'number' is conceptualized as the sum of the discrete items in a given array, and this is contrasted with continuous visual properties of the stimulus, such as individual dot size (area), the total area of all dots in a given array, the local spacing (Euclidean distance) between individual dots, global spacing (e.g., the average pairwise distance between all pairs of dots in the array – a property closely related to density), the outer contour length of the array (a semisubjective tracing of the outer envelope of the array), and so on. In contrast to a discrete conception of number, these properties vary continuously (akin to how the amount of water in a glass varies continuously, at least to the unaided human eye), and hence are typically labeled as 'magnitudes' as opposed to numbers. Notice that most of

these properties will tend to be correlated (or sometimes anticorrelated) with the relative number of dots in each array; that is, these continuous properties provide information about relative number. Hence, decisions may be made not on the basis of the number of items in each array, but may instead be driven entirely or in part by one or more of these covariate, continuous parameters. For instance, if the size of the individual dots is held constant, the array with a greater total area must contain more dots; if total area is equated across arrays, then the array with smaller dots on average must contain more dots. Analogous trade-offs present themselves for the other continuous properties as well. In sum, the very rules of geometry preclude the construction of a pair of arrays that differ in number of items and are exactly equal on all conceivable continuous (nonnumerical) parameters. The question, then, is, if we have no way to measure number independently of these other confounding variables, then can we ever be certain that what we are calling 'approximate number' is numerical at all? Might it be more appropriate to refer to an 'approximate magnitude system'? Indeed, one may question the logical soundness of an 'approximate number' to begin with, in that one may simply be confusing a property of a set (the number of items in the set) with the set's capacity for change (i.e., the possibility for the set to contain different numbers of objects with different probabilities). Indeed, this distinction is crucial for understanding and maintaining the difference between a number (even a number expressed as a ratio or proportion) and a probability. To be sure, debate is heated (e.g., Cordes and Brannon, 2009; De Hevia, 2011; Gebuis and Reynvoet, 2012a,b,c; Leibovich and Henik, 2013, 2014; Nys and Content, 2012); however, as has happened elsewhere when scientists have run up against what appears to be a fundamental limit on humans' capacity to measure a given phenomenon, one's ultimate conclusion may well be more a matter of philosophy (Heisenberg, 1979).

In sum, then, whether one chooses to refer to an 'ANS,' an 'approximate magnitude system,' 'number sense,' or simply 'nonsymbolic magnitude processing' may, at least at present, be largely a matter of opinion. Given the ongoing controversy, for present purposes, I have chosen what seems to me the most neutral term – nonsymbolic magnitude processing. However, the reader is encouraged to recognize these other terms in the literature, and, hopefully, consider the theoretical assumptions behind each.

Subitizing

As noted in the previous section, there is a certain conceptual difficulty (and perhaps methodological indeterminacy) with the notion of 'approximate number.' If one's requirement for nonsymbolic 'number' is that the numerical representation be exact, then one possible candidate is subitizing. Subitizing refers to the rapid, exact apprehension of the number of objects in a set without explicit counting (Mandler and Shebo, 1982; Dehaene and Cohen, 1994; Trick and Pylyshyn, 1994). Subitizing is subject to a limit of about four items that appear to be tied to a general processing capacity limit for visual short-term memory (Luck and Vogel, 1997; Piazza et al., 2011). When enumerating the exact number of objects in a set, when there are four or fewer, performance varies very little for sets with one versus four objects (both in terms of errors and response times; e.g., Trick and Pylyshyn; Revkin et al., 2008; Vuokko et al., 2013; Watson et al., 2007). This stands in contrast to enumerating larger sets of objects. One either begins to introduce increasingly large errors as the number of items increases (in the case of estimation; e.g., Revkin et al., 2008; Izard and Dehaene, 2008), or response times systematically increase as one counts the additional items (e.g., Trick and Pylyshyn, 1994; Vuokko et al., 2013). Furthermore, counting appears to rely disproportionately more on working memory resources than does subitizing (Shimomura and Kumada, 2011; Tuholski et al., 2001). Neuroimaging evidence also supports the notion that processing nonsymbolic sets >4 relies on qualitatively different neural systems than does processing subitizable set ≤ 4 (Ansari et al., 2007; Ester et al., 2012; Sathian et al., 1999; Vetter et al., 2011; Vuokko et al., 2013). There is also evidence to suggest there are two separate systems for processing small (≤ 4) and large (>4) numbers of items even in infants (Xu and Spelke, 2000). Taken together, there is evidence to suggest that subitizing operates in a qualitatively different manner than processing of larger nonsymbolic magnitudes. One important upshot (to which I will return in the next Section Symbolic Numbers) is that representation of exact number appears to be possible in the case of subitizing, even in young children (Le Corre and Carey, 2007).

Symbolic Numbers

Symbolic numbers are a relatively recent cultural invention when viewed from the broader context of human evolution. There is some evidence to suggest that tally sticks were used to record quantities perhaps as early as 40 000 years ago (Ifrah, 2000). However, evidence for more sophisticated systems for representing and manipulating abstract numerical symbols is absent from the archaeological record until the rise of Mesopotamian civilization around 3100 BC, which was followed shortly thereafter by the emergence of symbolic number systems in ancient Egypt and China only a few centuries later (Ifrah, 2000). Archaeological evidence for the use of spoken number words – another kind of number symbol – is more difficult to come by, though one may infer their use, for example, from root languages such as Proto-Indo-European (PIE). Many languages that have descended from PIE share similar forms for number words up to about 10 (Menninger, 1969/1992). As PIE is believed to have been originally spoken *c*.4200 BC (Blench and Spriggs, 1997), one may infer the use of similar number words perhaps as early as roughly 6000 years ago. On the other hand, present-day indigenous cultures without substantial contact with modern developed cultures have number words up to only three or four, and they lack more sophisticated means of computing exact quantities that exceed this limited range (Pica et al., 2004). In sum, while the ability to process nonsymbolic magnitudes appears to stem from a deep evolutionary history, the ability to manipulate exact numbers in symbolic

form is a much more recent cultural invention (see also Damerow, 2001). This places the acquisition and mastery of even basic processing of symbolic numbers (such as memorization and linking of simple arithmetic facts) at a fascinating junction between evolution and culture (Dehaene and Cohen, 2007; Ansari, 2008).

Symbol Grounding

A central question for research on symbolic numbers concerns what exactly number symbols represent – what do they refer? (Harnad, 1990). An early proposal is that they refer to the corresponding nonsymbolic magnitude: "When we learn number symbols, we simply attach their arbitrary shapes to the relevant nonsymbolic quantity representations" (Dehaene, 2008: p. 552; see also, Dehaene, 1997; Gallistel and Gelman, 2000). The appeal of this explanation is clear: the meaning of a symbolic number is grounded in what appears to be its evolutionary precursor in the (possibly innate) representation of the corresponding nonsymbolic magnitude, and support for this view has been echoed many times over (e.g., Verguts and Fias, 2004; Feigenson et al., 2004, 2013; Piazza et al., 2007; Hubbard et al., 2008; Libertus and Brannon, 2009; Nieder and Dehaene, 2009; Lyons and Ansari, 2009; Eger et al., 2009). More specifically, the argument is that symbolic numbers provide a more precise mapping (narrower tuning curve) onto the corresponding nonsymbolic magnitude (for a computational instantiation of this view, see Verguts and Fias, 2004).

Three different lines of evidence are often cited to support the proposal that symbolic numbers' meaning is derived from direct reference to their corresponding nonsymbolic magnitude. First, numerical signatures such as the distance and ratio effect show qualitatively similar patterns for symbolic numbers and nonsymbolic magnitudes (e.g., Buckley and Gillman, 1974; Dehaene, 2008). Second, neuroimaging evidence often points to similar neural substrates – specifically, the IPS – for symbolic and nonsymbolic number processing (Fias et al., 2003; Diester and Nieder, 2007; Piazza et al., 2007; Eger et al., 2009). Third, individual differences in nonsymbolic magnitude processing are related to more complex symbolic math abilities (as reviewed in the previous Section Number Sense in Early Development). In sum, one candidate for the meaning of a symbolic numbers – i.e., a potential solution to the symbol grounding problem – is that a given symbol simply refers to one's internal representation of the corresponding nonsymbolic magnitude.

Recently, several counterarguments to this proposal have arisen. First, one may note that distance effects are observed in virtually any discrimination task, whether one is distinguishing between items on the basis of perceptual dimensions (e.g., odor discrimination in drosophila; Parnas et al., 2013) or more abstract, categorical variables (e.g., distinguishing between species of animal figures; Gilbert et al., 2008). As discussed previously when interpreting distance and ratio effects, it is difficult to see how these effects in symbolic numbers are indicative of underlying representations in the same way as nonsymbolic magnitudes. Hence, it may be just as problematic to argue that qualitatively similar numerical distance/ratio effects are evidence of common representation across symbolic numbers and nonsymbolic magnitudes as it would be to argue that odors in drosophila and abstract animal categories in humans are underlain by a common representation.

Second, with respect to neural overlap of symbolic and nonsymbolic number processing, inferring a common neural mechanism from coactivation of brain regions (called reverse inference), though common, is logically problematic (Poldrack, 2006). Recent evidence suggests that such overlap in the case of numbers is dependent upon task demands in any case: neural activity for symbolic and nonsymbolic numerical ordering tasks shows a qualitatively different pattern of overlap relative to that for symbolic and nonsymbolic numerical comparison tasks (Lyons and Beilock, 2013). Further, of the three papers (Eger et al., 2009; Damarla and Just, 2013; Bulthé et al., 2014) that assessed whether distributed patterns of activity during processing of numbers in one format (e.g., symbolic) can be used to decode processing of numbers in the opposite format (e.g., nonsymbolic), only one (Eger et al., 2009) found evidence of successful cross-format decoding (and even in that case decoding was unidirectional and only slightly above chance: 57% accuracy, where chance was 50%). Cross-format fMRI adaptation has also yielded mixed results, with the responses to format change substantially larger than the responses to number change (Cohen Kadosh et al., 2011; though see also Piazza et al., 2007). In sum, while it is certainly possible that neural overlap of symbolic numbers and nonsymbolic magnitudes is more than a coincidence (Dehaene and Cohen, 2007), closer inspection suggests that there is substantial neural dissociation between the two formats. Behavioral evidence is consistent with this view: when asked to directly compare a symbolic number with a nonsymbolic magnitude (in an otherwise standard number comparison task), the cost of switching between formats is closer to what one would expect if the two formats referenced different as opposed to the same underlying representations (Lyons et al., 2012).

Third, with respect to the relationship between nonsymbolic magnitude processing and more complex math skills, as discussed in the previous Section Number Sense in Early Development, evidence for this result is quite mixed, even in preschoolers and kindergarteners who are just learning to use symbolic numbers. In sum, then, recent evidence has begun to erode a strong view of the notion that the meanings of number symbols are grounded in a direct reference to their nonsymbolic counterparts. An alternative explanation is that number symbols are initially linked exclusively via the exact, nonsymbolic quantities within the subitizing range (≤ 4; Carey, 2004). Indeed, Le Corre and Carey (2007) showed that 3- and 4-year-old children mapped number words onto corresponding magnitudes within the subitizing range prior to acquiring an understanding of the 'cardinality principle' (that counting to any number yields the number in the set, as indexed as the last number said). Children at this age failed to consistently map corresponding number words onto sets containing more than four items until several months after acquiring the cardinality principle. Further evidence consistent with this view has been found in adults: the cost of

mixing symbolic and nonsymbolic formats for small numbers (≤4) is substantially smaller than that found for large numbers (>4; Lyons et al., 2012), suggesting that subitizable symbolic and nonsymbolic numbers may retain a strong link even into adulthood.

Beyond Number Sense

How, then, do number symbols outside the subitizing range acquire meaning? One idea is that, having established a one-to-one mapping with nonsymbolic subitizable magnitudes, the meaning of (numerically) larger symbolic numbers are then bootstrapped from the smaller number symbols (Carey, 2004). Precisely how this process occurs, however, remains a point of debate (Ansari, 2008). One proposal is that the grammatical structure of language plays a critical role in inferring the existence and later the meaning of large exact numbers (Carey, 2004; Le Corre and Carey, 2007; Almoammer et al., 2013; Sarnecka et al., 2007; Sullivan and Barner, 2014). Another possibility is that mapping symbolic numbers to a visual-spatial mental number line is key (e.g., Gunderson et al., 2012). Specifically, understanding that the numerical distance between integers remains constant helps one to apprehend the notion of large exact numbers, which is evidenced by a linearization of the number line up to ever-increasing magnitudes (10, 100, 1000, etc.; Siegler and Ramani, 2008). Yet another view suggests that how we understand numbers as ordered sequences differs fundamentally between symbolic and nonsymbolic numbers (Lyons and Beilock, 2011, 2013). The ordinal associations between symbolic numbers may thus provide one distinct aspect of symbolic numbers' respective meaning (Nieder, 2009). These competing views need not be mutually exclusive, and indeed, they all share the general assumption that the meanings of larger number symbols do not remain directly tied to their nonsymbolic counterpart. (For a competing view that expressly disagrees with this assumption, see Feigenson et al. (2013).) Note that this assumption is consistent with a broader view of symbolic representation more generally (Deacon, 1997; for an early suggestion applying this perspective specifically to number representation, see Nieder, 2009). According to this view, it is the relations between symbols that ultimately come to define a symbolic system. Over time, these relations may even come to overshadow each symbol's link to an external referent. It is perhaps this that makes it possible to manipulate very large numbers such as 1 000 000 without having a concrete sense of what 1 000 000 is like. That said, even if such a view proves to be correct, a major unanswered question concerns the precise mechanism – whether that be linguistic, visual-spatial, ordinal, nonsymbolic magnitudes, some combination thereof, or something else entirely – that links number symbols to one another. Answering this question is thus a major contemporary topic of research in the field of numerical cognition. On a somewhat grander scale, such research can help us understand how cultural and evolutionary forces interact within the human brain in shaping our ability to manipulate the world of and with mathematics.

Acknowledgments

Thanks to Daniel Ansari and Elizabeth Necka for their helpful comments on earlier drafts of this article.

See also: Cognitive Development: Mathematics Learning and Instruction; Mathematical Concepts During Early Childhood Across Cultures, Development of; Mathematics Education; Number Systems, Evolution of; Number, Grammatical; Parietal Lobe; Serial Order, Representation of.

Bibliography

Agrillo, C., Beran, M.J., 2013. Number without language: comparative psychology and the evolution of numerical cognition. Frontiers in Psychology 4, 295.

Agrillo, C., Dadda, M., Serena, G., Bisazza, A., 2009. Use of number by fish. PLoS One 4 (3), e4786.

Almoammer, A., Sullivan, J., Donlan, C., Marušič, F., Žaucer, R., O'Donnell, T., Barner, D., 2013. Grammatical morphology as a source of early number word meanings. Proceedings of the National Academy of Sciences of the United States of America 110 (46), 18448–18453.

Ansari, D., Fugelsang, J.A., Dhital, B., Venkatraman, V., 2006. Dissociating response conflict from numerical magnitude processing in the brain: an event-related fMRI study. NeuroImage 32 (2), 799–805.

Ansari, D., Lyons, I.M., van Eimeren, L., Xu, F., 2007. Linking visual attention and number processing in the brain: the role of the temporo-parietal junction in small and large symbolic and nonsymbolic number comparison. Journal of Cognitive Neuroscience 19 (11), 1845–1853.

Ansari, D., 2008. Effects of development and enculturation on number representation in the brain. Nature Reviews Neuroscience 9 (4), 278–291.

Armstrong, N., Garland, A., Burns, K.C., 2012. Memory for multiple cache locations and prey quantities in a food-hoarding songbird. Frontiers in Psychology 3, 584.

Baker, J.M., Morath, J., Rodzon, K.S., Jordan, K.E., 2012. A shared system of representation governing quantity discrimination in canids. Frontiers in Psychology 3, 387.

Barnard, A.M., Hughes, K.D., Gerhardt, R.R., Divincenti Jr., L., Bovee, J.M., Cantlon, J.F., 2013. Inherently analog quantity representations in olive baboons (*Papio anubis*). Frontiers in Psychology 4, 253.

Bartelet, D., Vaessen, A., Blomert, L., Ansari, D., 2014. What basic number processing measures in kindergarten explain unique variability in first-grade arithmetic proficiency? Journal of Experimental Child Psychology 117, 12–28.

Barth, H., La Mont, K., Lipton, J., Dehaene, S., Kanwisher, N., Spelke, E., 2006. Non-symbolic arithmetic in adults and young children. Cognition 98 (3), 199–222.

Beran, M.J., Perdue, B.M., Parrish, A.E., Evans, T.A., 2012. Do social conditions affect capuchin monkeys' (*Cebus apella*) choices in a quantity judgment task? Frontiers in Psychology 3, 492.

Blench, R., Spriggs, M. (Eds.), 1997. Archaeology and Language III: Artefacts, Languages and Texts. Routledge, New York.

Bonny, J.W., Lourenco, S.F., 2013. The approximate number system and its relation to early math achievement: evidence from the preschool years. Journal of Experimental Child Psychology 114 (3), 375–388.

Brannon, E.M., Terrace, H.S., 2002. The evolution and ontogeny of ordinal numerical ability. In: Bekoff, M., Allen, C., Burghardt, G. (Eds.), The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition. MIT Press, Cambridge, MA, pp. 197–204.

Buckley, P.B., Gillman, C.B., 1974. Comparisons of digits and dot patterns. Journal of Experimental Psychology 103 (6), 1131–1136.

Bulthé, J., De Smedt, B., Op de Beeck, H.P., 2014. Format-dependent representations of symbolic and non-symbolic numbers in the human cortex as revealed by multi-voxel pattern analyses. NeuroImage 87, 311–322.

Butterworth, B., 2010. Foundational numerical capacities and the origins of dyscalculia. Trends in Cognitive Sciences 14 (12), 534–541.

Cantlon, J.F., Brannon, E.M., Carter, E.J., Pelphrey, K.A., 2006. Functional imaging of numerical processing in adults and 4-y-old children. PLoS Biology 4 (5), e125.

Cantlon, J.F., Brannon, E.M., 2006. Shared system for ordering small and large numbers in monkeys and humans. Psychological Science 17 (5), 401–406.

Cantlon, J.F., Brannon, E.M., 2007. Basic math in monkeys and college students. PLoS Biology 5 (12), e328.

Cantlon, J.F., Platt, M.L., Brannon, E.M., 2009. Beyond the number domain. Trends in Cognitive Sciences 13 (2), 83–91.

Carazo, P., FernÃ¡ndez-Perea, R., Font, E., 2012. Quantity estimation based on numerical cues in the mealworm beetle (*Tenebrio molitor*). Frontiers in Psychology 3, 502.

Carey, S., Winter 2004. Bootstrapping and the origins of concepts. Daedalus 59–68.

Castronovo, J., Göbel, S.M., 2012. Impact of high mathematics education on the number sense. PLoS One 7 (4), e33832.

Chen, Q., Li, J., 2014. Association between individual differences in non-symbolic number acuity and math performance: a meta-analysis. Acta Psychologica (Amst) 148, 163–172.

Cohen Kadosh, R., Bahrami, B., Walsh, V., Butterworth, B., Popescu, T., Price, C.J., 2011. Specialization in the human brain: the case of numbers. Frontiers in Human Neuroscience 5, 62.

Cordes, S., Brannon, E.M., 2008. Quantitative competencies in infancy. Developmental Science 11 (6), 803–808.

Cordes, S., Brannon, E.M., 2009. The relative salience of discrete and continuous quantity in young infants. Developmental Science 12 (3), 453–463.

Damarla, S.R., Just, M.A., 2013. Decoding the representation of numerical values from brain activation patterns. Human Brain Mapping 34 (10), 2624–2634.

Damerow, P., 2001. Evolution of number systems. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of Social and Behavioral Sciences, pp. 10753–10756.

Deacon, T., 1997. The Symbolic Species: The Co-evolution of Language and the Brain. W.W. Norton & Company, Inc., New York.

Dehaene, S., 1997. The Number Sense: How the Mind Creates Mathematics. Oxford University Press, New York.

Dehaene, S., Mehler, J., 1992. Cross-linguistic regularities in the frequency of number words. Cognition 43 (1), 1–29.

Dehaene, S., Changeux, J.P., 1993. Development of elementary numerical abilities: a neuronal model. Journal of Cognitive Neuroscience 5 (4), 390–407.

Dehaene, S., Cohen, L., 1994. Dissociable mechanisms of subitizing and counting: neuropsychological evidence from simultanagnosic patients. Journal of Experimental Psychology Human Perception and Performance 20 (5), 958–975.

Dehaene, S., Cohen, L., 2007. Cultural recycling of cortical maps. Neuron 56 (2), 384–398.

Dehaene, S., 2008. Symbols and quantities in parietal cortex: elements of a mathematical theory of number representation and manipulation. In: Haggard, P., Rossetti, Y. (Eds.), Sensorimotor Foundations of Higher Cognition (Attention and Performance). Oxford University Press, New York, pp. 527–574.

De Hevia, M.D., 2011. Sensitivity to number: reply to Gebuis and Gevers. Cognition 121 (2), 253–255.

De Smedt, B., Noël, M.P., Gilmore, C., Ansari, D., 2013. How do symbolic and non-symbolic numerical magnitude processing skills relate to individual differences in children's mathematical skills? A review of evidence from brain and behavior. Trends in Neuroscience and Education 2 (2), 48–55.

Desoete, A., Ceulemans, A., De Weerdt, F., Pieters, S., 2012. Can we predict mathematical learning disabilities from symbolic and non-symbolic comparison tasks in kindergarten? Findings from a longitudinal study. British Journal of Educational Psychology 82 (1), 64–81.

Dewind, N.K., Brannon, E.M., 2012. Malleability of the approximate number system: effects of feedback and training. Frontiers in Human Neuroscience 6, 68.

Diester, I., Nieder, A., 2007. Semantic associations between signs and numerical categories in the prefrontal cortex. PLoS Biology 5, e294.

Eger, E., Michel, V., Thirion, B., Amadon, A., Dehaene, S., Kleinschmidt, A., 2009. Deciphering cortical number coding from human brain activity patterns. Current Biology 19 (19), 1608–1615.

Ester, E.F., Drew, T., Klee, D., Vogel, E.K., Awh, E., 2012. Neural measures reveal a fixed item limit in subitizing. Journal of Neuroscience 32 (21), 7169–7177.

Feigenson, L., Dehaene, S., Spelke, E., 2004. Core systems of number. Trends in Cognitive Sciences 8 (7), 307–314.

Feigenson, L., Libertus, M.E., Halberda, J., 2013. Links between the intuitive sense of number and formal mathematics ability. Child Development Perspectives 7 (2), 74–79.

Fias, W., Lammertyn, J., Reynvoet, B., Dupont, P., Orban, G.A., 2003. Parietal representation of symbolic and nonsymbolic magnitude. Journal of Cognitive Neuroscience 15 (1), 47–56.

Franklin, M.S., Jonides, J., Smith, E.E., 2009. Processing of order information for numbers and months. Memory and Cognition 37 (5), 644–654.

Franklin, M.S., Jonides, J., 2009. Order and magnitude share a common representation in parietal cortex. Journal of Cognitive Neuroscience 21 (11), 2114–2120.

Fuhs, M.W., McNeil, N.M., 2013. ANS acuity and mathematics ability in preschoolers from low-income homes: contributions of inhibitory control. Developmental Science 16 (1), 136–148.

Gallistel, C.R., Gelman, R., 2000. Non-verbal numerical cognition: from reals to integers. Trends in Cognitive Sciences 4 (2), 59–65.

Gebuis, T., Reynvoet, B., 2012a. Continuous visual properties explain neural responses to nonsymbolic number. Psychophysiology 49 (11), 1481–1491.

Gebuis, T., Reynvoet, B., 2012b. The interplay between nonsymbolic number and its continuous visual properties. Journal of Experimental Psychology General 141 (4), 642–648.

Gebuis, T., Reynvoet, B., 2012c. The role of visual information in numerosity estimation. PLoS One 7 (5), e37426.

Gelman, R., Gallistel, C.R., 2004. Language and the origin of numerical concepts. Science 306 (5695), 441–443.

Gilbert, A.L., Regier, T., Kay, P., Ivry, R.B., 2008. Support for lateralization of the Whorf effect beyond the realm of color discrimination. Brain and Language 105 (2), 91–98.

Gilmore, C.K., McCarthy, S.E., Spelke, E.S., 2010. Non-symbolic arithmetic abilities and mathematics achievement in the first year of formal schooling. Cognition 115 (3), 394–406.

Gilmore, C., Attridge, N., Inglis, M., 2011. Measuring the approximate number system. Quarterly Journal of Experimental Psychology (Hove) 64 (11), 2099–2109.

Göbel, S.M., Watson, S.E., Lervåg, A., Hulme, C., 2014. Children's arithmetic development: it is number knowledge, not the approximate number sense, that counts. Psychological Science 25 (3), 789–798.

Grill-Spector, K., Henson, R., Martin, A., 2006. Repetition and the brain: neural models of stimulus-specific effects. Trends in Cognitive Sciences 10 (1), 14–23.

Gunderson, E.A., Ramirez, G., Beilock, S.L., Levine, S.C., 2012. The relation between spatial skill and early number knowledge: the role of the linear number line. Developmental Psychology 48 (5), 1229–1241.

Halberda, J., Sires, S.F., Feigenson, L., 2006. Multiple spatially overlapping sets can be enumerated in parallel. Psychological Science 17 (7), 572–576.

Halberda, J., Feigenson, L., 2008. Developmental change in the acuity of the number sense: the approximate number system in 3-, 4-, 5- and 6-year-olds and adults. Developmental Psychology 44 (5), 1457–1465.

Halberda, J., Ly, R., Wilmer, J.B., Naiman, D.Q., Germine, L., 2012. Number sense across the lifespan as revealed by a massive internet-based sample. Proceedings of the National Academy of Sciences of the United States of America 109 (28), 11116–11120.

Halberda, J., Mazzocco, M.M., Feigenson, L., 2008. Individual differences in non-verbal number acuity correlate with maths achievement. Nature 455 (7213), 665–668.

Harnad, S., 1990. The symbol grounding problem. Physica D Nonlinear Phenomena 42, 335–346.

Heisenberg, W., 1979. Philosophical Problems of Quantum Physics. Oxbow Press, Oxford.

Holloway, I.D., Ansari, D., 2009. Mapping numerical magnitudes onto symbols: the numerical distance effect and individual differences in children's mathematics achievement. Journal of Experimental Child Psychology 103 (1), 17–29.

Holloway, I.D., Battista, C., Vogel, S.E., Ansari, D., 2013. Semantic and perceptual processing of number symbols: evidence from a cross-linguistic fMRI adaptation study. Journal of Cognitive Neuroscience 25 (3), 388–400.

Hubbard, E.M., Diester, I., Cantlon, J.F., Ansari, D., Opstal, F.V., Troiani, V., 2008. The evolution of numerical cognition: from number neurons to linguistic quantifiers. Journal of Neuroscience 28 (46), 11819–11824.

Hyde, D.C., Khanum, S., Spelke, E.S., 2014. Brief non-symbolic, approximate number practice enhances subsequent exact symbolic arithmetic in children. Cognition 131 (1), 92–107.

Ifrah, G., 2000. The Universal History of Numbers: From Prehistory to the Invention of the Computer. John Wiley & Sons, Inc., New York.

Indow, T., Ida, M., 1977. Scaling of dot numerosity. Perception & Psychophysics 22 (3), 265–276.

Inglis, M., Gilmore, C., 2014. Indexing the approximate number system. Acta Psychologica (Amst) 145, 147–155.

Izard, V., Dehaene, S., 2008. Calibrating the mental number line. Cognition 106 (3), 1221–1247.

Izard, V., Dehaene-Lambertz, G., Dehaene, S., 2008. Distinct cerebral pathways for object identity and number in human infants. PLoS Biology 6 (2), e11.

Jordan, N.C., Kaplan, D., Ramineni, C., Locuniak, M.N., 2009. Early math matters: kindergarten number competence and later mathematics outcomes. Developmental Psychology 45 (3), 850–867.

Kaufmann, L., Koppelstaetter, F., Delazer, M., Siedentopf, C., Rhomberg, P., Golaszewski, S., Felber, S., Ischebeck, A., 2005. Neural correlates of distance and congruity effects in a numerical Stroop task: an event-related fMRI study. NeuroImage 25 (3), 888–898.

Kolkman, M.E., Kroesbergen, E.H., Leseman, P.P.M., 2013. Early numerical development and the role of non-symbolic and symbolic skills. Learning and Instruction 25, 95–103.

Le Corre, M., Carey, S., 2007. One, two, three, four, nothing more: an investigation of the conceptual sources of the verbal counting principles. Cognition 105 (2), 395–438.

Leibovich, T., Henik, A., 2014. Comparing performance in discrete and continuous comparison tasks. Quarterly Journal of Experimental Psychology (Hove) 67 (5), 899–917.

Leibovich, T., Henik, A., 2013. Magnitude processing in non-symbolic stimuli. Frontiers in Psychology 4, 375.

Libertus, M.E., Brannon, E.M., 2009. Behavioral and neural basis of number sense in infancy. Current Directions in Psychological Science 18 (6), 346–351.

Libertus, M.E., Feigenson, L., Halberda, J., 2011. Preschool acuity of the approximate number system correlates with school math ability. Developmental Science 14 (6), 1292–1300.

Libertus, M.E., Odic, D., Halberda, J., 2012. Intuitive sense of number correlates with math scores on college-entrance examination. Acta Psychologica 141 (3), 373–379.

Libertus, M.E., Feigenson, L., Halberda, J., 2013. Is approximate number precision a stable predictor of math ability? Learning and Individual Differences 25, 126–133.

Lipton, J.S., Spelke, E.S., 2003. Origins of number sense. Large-number discrimination in human infants. Psychological Science 14 (5), 396–401.

Lonnemanna, J., Linkersdörfera, J., Hasselhorna, M., Lindberg, S., 2011. Symbolic and non-symbolic distance effects in children and their connection with arithmetic skills. Journal of Neurolinguistics 24 (5), 582–591.

Lourenco, S.F., Bonny, J.W., Fernandez, E.P., Rao, S., 2012. Nonsymbolic number and cumulative area representations contribute shared and unique variance to symbolic math competence. Proceedings of the National Academy of Sciences of the United States of America 109 (46), 18737–18742.

Luck, S.J., Vogel, E.K., 1997. The capacity of visual working memory for features and conjunctions. Nature 390 (6657), 279–281.

Lyons, I.M., Ansari, D., 2009. The cerebral basis of mapping nonsymbolic numerical quantities onto abstract symbols: an fMRI training study. Journal of Cognitive Neuroscience 21 (9), 1720–1735.

Lyons, I.M., Beilock, S.L., 2009. Beyond quantity: individual differences in working memory and the ordinal understanding of numerical symbols. Cognition 113 (2), 189–204.

Lyons, I.M., Beilock, S.L., 2011. Numerical ordering ability mediates the relation between number-sense and arithmetic competence. Cognition 121 (2), 256–261.

Lyons, I.M., Ansari, D., Beilock, S.L., 2012. Symbolic estrangement: evidence against a strong association between numerical symbols and the quantities they represent. Journal of Experimental Psychology General 141 (4), 635–641.

Lyons, I.M., Beilock, S.L., 2013. Ordinality and the nature of symbolic numbers. Journal of Neuroscience 33 (43), 17052–17061.

Lyons, I.M., Price, G.R., Vaessen, A., Blomert, L., Ansari, D., 2014. Numerical predictors of arithmetic success in grades 1–6. Developmental Science 17 (5), 714–726.

Maloney, E.A., Risko, E.F., Preston, F., Ansari, D., Fugelsang, J., 2010. Challenging the reliability and validity of cognitive measures: the case of the numerical distance effect. Acta Psychologica (Amst) 134 (2), 154–161.

Mandler, G., Shebo, B.J., 1982. Subitizing: an analysis of its component processes. Journal of Experimental Psychology General 111 (1), 1–22.

Mazzocco, M.M., Feigenson, L., Halberda, J., 2011a. Impaired acuity of the approximate number system underlies mathematical learning disability (dyscalculia). Child Development 82 (4), 1224–1237.

Mazzocco, M.M., Feigenson, L., Halberda, J., 2011b. Preschoolers' precision of the approximate number system predicts later school mathematics performance. PLoS One 6 (9), e23749.

McCrink, K., Spelke, E.S., 2010. Core multiplication in childhood. Cognition 116 (2), 204–216.

Menninger, K., 1969/1992. Number Words and Number Symbols: A Cultural History of Numbers. Dover Publications, Inc., New York.

Merten, K., Nieder, A., 2009. Compressed scaling of abstract numerosity representations in adult humans and monkeys. Journal of Cognitive Neuroscience 21 (2), 333–346.

Moyer, R.S., Landauer, T.K., 1967. Time required for judgements of numerical inequality. Nature 215 (5109), 1519–1520.

Mussolin, C., Noël, M.P., Pesenti, M., Grandin, C., De Volder, A.G., 2013. Neural correlates of the numerical distance effect in children. Frontiers in Psychology 4, 663.

Mussolin, C., Nys, J., Content, A., Leybaert, J., 2014. Symbolic number abilities predict later approximate number system acuity in preschool children. PLoS One 9 (3), e91839.

Nieder, A., 2009. Prefrontal cortex and the evolution of symbolic reference. Current Opinion in Neurobiology 19 (1), 99–108.

Nieder, A., Dehaene, S., 2009. Representation of number in the brain. Annual Review of Neuroscience 32, 185–208.

Nieder, A., Merten, K., 2007. A labeled-line code for small and large numerosities in the monkey prefrontal cortex. Journal of Neuroscience 27 (22), 5986–5993.

Nieder, A., 2005. Counting on neurons: the neurobiology of numerical competence. Nature Reviews Neuroscience 6 (3), 177–190.

Notebaert, K., Pesenti, M., Reynvoet, B., 2010. The neural origin of the priming distance effect: distance-dependent recovery of parietal activation using symbolic magnitudes. Human Brain Mapping 31 (5), 669–677.

Nys, J., Content, A., 2012. Judgement of discrete and continuous quantity in adults: number counts! Quarterly Journal of Experimental Psychology (Hove) 65 (4), 675–690.

Pahl, M., Si, A., Zhang, S., 2013. Numerical cognition in bees and other insects. Frontiers in Psychology 4, 162.

Panteleeva, S., Reznikova, Z., Vygonyailova, O., 2013. Quantity judgments in the context of risk/reward decision making in striped field mice: first count then hunt. Frontiers in Psychology 4, 53.

Park, J., Brannon, E.M., 2013. Training the approximate number system improves math proficiency. Psychological Science 24 (10), 2013–2019.

Park, J., Brannon, E.M., 2014. Improving math with number sense training: an investigation of its underlying mechanism. Cognition 133 (1), 188–200.

Parnas, M., Lin, A.C., Huetteroth, W., Miesenböck, G., 2013. Odor discrimination in *Drosophila*: from neural population codes to behavior. Neuron 79 (5), 932–944.

Pepperberg, I.M., 2013. Abstract concepts: data from a grey parrot. Behavioural Processes 93, 82–90.

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., Dehaene, S., 2004. Tuning curves for approximate numerosity in the human intraparietal sulcus. Neuron 44 (3), 547–555.

Piazza, M., Pinel, P., Le Bihan, D., Dehaene, S., 2007. A magnitude code common to numerosities and number symbols in human intraparietal cortex. Neuron 53 (2), 293–305.

Piazza, M., 2010. Neurocognitive start-up tools for symbolic number representations. Trends in Cognitive Sciences 14 (12), 542–551.

Piazza, M., Facoetti, A., Trussardi, A.N., Berteletti, I., Conte, S., Lucangeli, D., Dehaene, S., Zorzi, M., 2010. Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. Cognition 116 (1), 33–41.

Piazza, M., Fumarola, A., Chinello, A., Melcher, D., 2011. Subitizing reflects visuo-spatial object individuation capacity. Cognition 121 (1), 147–153.

Pica, P., Lemer, C., Izard, V., Dehaene, S., 2004. Exact and approximate arithmetic in an Amazonian indigene group. Science 306 (5695), 499–503.

Piffer, L., Miletto Petrazzini, M.E., Agrillo, C., 2013. Large number discrimination in newborn fish. PLoS One 8 (4), e62466.

Pinel, P., Dehaene, S., Rivière, D., LeBihan, D., 2001. Modulation of parietal activation by semantic distance in a number comparison task. NeuroImage 14 (5), 1013–1026.

Poldrack, R.A., 2006. Can cognitive processes be inferred from neuroimaging data? Trends in Cognitive Sciences 10 (2), 59–63.

Price, G.R., Palmer, D., Battista, C., Ansari, D., 2012. Nonsymbolic numerical magnitude comparison: reliability and validity of different task variants and outcome measures, and their relationship to arithmetic achievement in adults. Acta Psychologica 140 (1), 50–57.

Revkin, S.K., Piazza, M., Izard, V., Cohen, L., Dehaene, S., 2008. Does subitizing reflect numerical estimation? Psychological Science 19 (6), 607–614.

Reynvoet, B., Brysbaert, M., Fias, W., 2002. Semantic priming in number naming. Quarterly Journal of Experimental Psychology A 55 (4), 1127–1139.

Reynvoet, B., De Smedt, B., Van den Bussche, E., 2009. Children's representation of symbolic magnitude: the development of the priming distance effect. Journal of Experimental Child Psychology 103 (4), 480–489.

Rubinsten, O., Henik, A., Berger, A., Shahar-Shalev, S., 2002. The development of internal representations of magnitude and their association with Arabic numerals. Journal of Experimental Child Psychology 81 (1), 74–92.

Sarnecka, B.W., Kamenskaya, V.G., Yamana, Y., Ogura, T., Yudovina, Y.B., 2007. From grammatical number to exact numbers: early meanings of 'one', 'two', and 'three' in English, Russian, and Japanese. Cognitive Psychology 55 (2), 136–168.

Sasanguie, D., Defever, E., Van den Bussche, E., Reynvoet, B., 2011. The reliability of and the relation between non-symbolic numerical distance effects in comparison, same-different judgments and priming. Acta Psychologica (Amst) 136 (1), 73–80.

Sasanguie, D., Göbel, S.M., Moll, K., Smets, K., Reynvoet, B., 2013. Approximate number sense, symbolic number processing, or number-space mappings: what underlies mathematics achievement? Journal of Experimental Child Psychology 114 (3), 418–431.

Sasanguie, D., Defever, E., Maertens, B., Reynvoet, B., 2014. The approximate number system is not predictive for symbolic number processing in kindergarteners. Quarterly Journal of Experimental Psychology 67 (2), 271–280.

Sathian, K., Simon, T.J., Peterson, S., Patel, G.A., Hoffman, J.M., Grafton, S.T., 1999. Neural evidence linking visual object enumeration and attention. Journal of Cognitive Neuroscience 11 (1), 36–51.

Sekuler, R., Mierkiewicz, D., 1977. Children's judgments of numerical inequality. Child Development 48, 630–633.

Shimomura, T., Kumada, T., 2011. Spatial working memory load affects counting but not subitizing in enumeration. Attention, Perception, & Psychophysics 73 (6), 1694–1709.

Siegler, R.S., Ramani, G.B., 2008. Playing linear numerical board games promotes low-income children's numerical development. Developmental Science 11 (5), 655–661.

Smirnova, A.A., Lazareva, O.F., Zorina, Z.A., 2000. Use of number by crows: investigation by matching and oddity learning. Journal of the Experimental Analysis of Behavior 73 (2), 163–176.

Starr, A., Libertus, M.E., Brannon, E.M., 2013. Number sense in infancy predicts mathematical abilities in childhood. Proceedings of the National Academy of Sciences of the United States of America 110 (45), 18116–18120.

Sullivan, J., Barner, D., 2014. Inference and association in children's early numerical estimation. Child Development 85 (4), 1740–1755.

Toll, S.W., Van Luit, J.E., 2014. Explaining numeracy development in weak performing kindergartners. Journal of Experimental Child Psychology 124, 97–111.

Trick, L.M., Pylyshyn, Z.W., 1994. Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. Psychological Review 101 (1), 80–102.

Tuholski, S.W., Engle, R.W., Baylis, G.C., 2001. Individual differences in working memory capacity and enumeration. Memory & Cognition 29 (3), 484–492.

Turconi, E., Campbell, J.I., Seron, X., 2006. Numerical order and quantity processing in number comparison. Cognition 98 (3), 273–285.

Utrata, E., Virányi, Z., Range, F., 2012. Quantity discrimination in wolves (*Canis lupus*). Frontiers in Psychology 3, 505.

Van Opstal, F., Gevers, W., De Moor, W., Verguts, T., 2008. Dissecting the symbolic distance effect: comparison and priming effects in numerical and nonnumerical orders. Psychonomic Bulletin and Review 15 (2), 419–425.

Verguts, T., Fias, W., 2004. Representation of number in animals and humans: a neural model. Journal of Cognitive Neuroscience 16 (9), 1493–1504.

Verguts, T., Fias, W., Stevens, M., 2005. A model of exact small-number representation. Psychonomic Bulletin & Review 12 (1), 66–80.

Verguts, T., Van Opstal, F., 2005. Dissociation of the distance effect and size effect in one-digit numbers. Psychonomic Bulletin & Review 12 (5), 925–930.

Vetter, P., Butterworth, B., Bahrami, B., 2011. A candidate for the attentional bottleneck: set-size specific modulation of the right TPJ during attentive enumeration. Journal of Cognitive Neuroscience 23 (3), 728–736.

Vuokko, E., Niemivirta, M., Helenius, P., 2013. Cortical activation patterns during subitizing and counting. Brain Research 1497, 40–52.

Watson, D.G., Maylor, E.A., Bruce, L.A., 2007. The role of eye movements in subitizing and counting. Journal of Experimental Psychology Human Perception and Performance 33 (6), 1389–1399.

Wilson, A.J., Revkin, S.K., Cohen, D., Cohen, L., Dehaene, S., 2006. An open trial assessment of the number race: an adaptive computer game for remediation of dyscalculia. Behavioral and Brain Functions 2, 20.

Xu, F., Spelke, E.S., 2000. Large number discrimination in 6-month-old infants. Cognition 74 (1), B1–B11.

Xu, F., Spelke, E.S., Goddard, S., 2005. Number sense in human infants. Developmental Science 8 (1), 88–101.

Yaman, S., Kilian, A., von Fersen, L., Güntürkün, O., 2012. Evidence for a numerosity category that is based on abstract qualities of "few" vs. "many" in the bottlenose dolphin (*Tursiops truncatus*). Frontiers in Psychology 3, 473.

Number Systems, Evolution of

Peter Damerow,[†] Max Planck Institute for the History of Science, Berlin, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 10753–10756, © 2001, Elsevier Ltd.

Abstract

A 'number system' consists of a set of mental entities and operations together with systems of symbols and symbolic activities. Until recently, cultures existed without any arithmetical techniques. The occurrence of the simplest genuine arithmetical activities dates back to the Late Neolithic and the Early Bronze Age. Advanced number systems evolved with the rise of early civilizations and the invention of writing. Context-dependent numerical sign systems were used to solve the administrative problems of early state organization. The later emergence of an abstract concept of number was a consequence of cultural exchange in the Hellenistic and Roman world. Limitations of this early number concept resulted from the introduction of algebraic notations, leading to the formalistic approach of the nineteenth century. Formalism finally eliminated in the twentieth century the conceptual distinction between number systems and other mathematical objects.

The Term 'Number System'

A 'number system' consists of a set of mental entities and operations together with systems of symbols such as number words, tallies, or numerical sign systems and of symbolic activities such as counting, addition, or multiplication representing them. Number systems evolved in the course of the development of most civilizations. This evolution changed not only numbers and number representations but also the meaning of the term 'number' itself.

Symbolic activities such as counting, the use of tallies for representing quantities, or performing calculations by means of tokens or systems of written symbols, are usually considered practices dealing with 'numbers.' However, such symbolic activities historically precede an explicit number concept. Explicit concepts of number are the outgrowth of reflection upon such activities.

When first explicit concepts of abstract 'numbers' emerged in history, they were restricted to the number system which is called today 'natural numbers,' that is, to the system '1, 2, 3, 4, …' The subsequent history of the term number is characterized by a series of extensions of the number concept to cover 'integers,' 'rational numbers,' 'real numbers,' 'complex numbers,' 'algebraic numbers,' and other mathematical objects designated, with lesser or greater justification, as a number system. These extensions usually included numbers of earlier stages as special kinds of numbers, but necessarily also dismissed certain properties which seemed to be essential in earlier stages. Consequently, any extension of number systems was accompanied by controversial discussions about the question whether the newly-considered mathematical objects were really numbers. These doubts are often reflected in the designation of such numbers as 'irrational,' 'negative,' 'imaginary,' etc. Finally, the variety of potential candidates for the term number became so great that the answer to the question of what is and what is not a 'number' turned into a merely conventional decision. The focus of attention shifted from the investigation of types of numbers as they were historically determined towards the study of their common structures.

Cultures without Number Systems

Until recently, 'prearithmetical' cultures existed without any arithmetical techniques at all such as finger counting, counting notches, counting knots, or tokens. The best known examples are the Australian aborigines (Blake, 1981; Dixon, 1980) and certain South American natives (Closs, 1986) whose languages did possess terms for quantities, yet of an exclusively qualitative nature, terms such as 'many,' 'high,' 'big,' or 'wide' (*see* Australian Aboriginal Society and Culture: An Overview; Justice: Philosophical Aspects; South America: Sociocultural Aspects). The quantitative aspects of an object of cognition were not yet distinguished from its specific physical appearance (*see* Indigenous Knowledge: Science and Technology Studies). Since from periods before the Late Neolithic no objects or signs have been identified that might have had some kind of arithmetical function, it is likely that up to that time human culture remained entirely prearithmetical. While artifacts with repeated signs and regular patterns from periods as early as the Late Paleolithic (*see* Modernity: Anthropological Aspects) have occasionally been interpreted as representations of numbers (Marshack, 1972), such an interpretation seems to be untenable since these sign repetitions lack the characteristic subdivision by counting levels which is present in all real counting systems (Damerow, 1998).

Proto-Arithmetic

The occurrence of the simplest genuine arithmetical activities known from recent nonliterate cultures (see, e.g., Saxe, 1982) date back to the Late Neolithic and the Early Bronze Age. These activities aiming at the identification and control of quantities are based on structured and standardized systems of symbols representing objects (*see* North America: Sociocultural

[†] Deceased.

Aspects; Ritual and Symbolism, Archaeology of). Their emergence as counting and tallying techniques may have been a consequence of sedentariness. Symbols are the most simple tools for the construction of one-to-one correspondences in counting and tallying that can be transmitted from generation to generation. The organization of agricultural cultivation, animal husbandry and household administration led to social conditions that apparently made symbolic techniques useful, and their systematic transmission and development possible. Such techniques are 'proto-arithmetical' insofar as the symbols represent objects and not 'numbers,' and consequently are not used for symbolic transformations which correspond to such arithmetical operations as addition and multiplication.

Early explorers and travelers encountering indigenous cultures using proto-arithmetical techniques often anachronistically interpreted their activities from a modern numerical perspective and believed the limitations of proto-arithmetic resulted from deficient mental abilities of such peoples. It was only in the first half of the twentieth century that anthropologists and psychologists challenged these beliefs (e.g., *see* Klein, Melanie (1882–1960); Boas, Franz (1858–1942)) and began seriously to study culturally specific mental constructions connected with proto-arithmetical activities (Levy-Bruhl, 1923; Wertheimer, 1925).

From Proto-Arithmetic to Systems of Numerical Symbols

A further step in the evolution of number systems resulted from the rise of early civilizations and the invention of writing. This transition is particularly well documented in the case of the Ancient Near East (*see* Writing, Evolution of). A system of clay tokens possessing proto-arithmetical functions has been identified which may have been widely used already during the period from the beginning of sedentariness in the areas surrounding the Mesopotamian lowland plain and in the Nile valley around 8000 BC until the emergence of cities around 4000 BC (Schmandt-Besserat, 1992).

In the fourth millennium BC this proto-arithmetic system of clay tokens became the central tool for the control of a locally-centralized economic administration. In order to serve this purpose, the proto-arithmetical capabilities of the system were exploited to the limits of their capacity. The safekeeping of tokens in closed and sealed spherical clay envelopes demonstrates that they were used for the encoding of important information. The tokens were complemented and later completely replaced by markings that were impressed on the surface of such clay envelopes or on the surface of sealed clay tablets. Thus, the system was transformed into a more suitable medium by substituting written signs for tokens. Around 3200 BC, the numerical notations were supplemented with pictograms, and the tablets achieved a more complex structure. They now displayed several quantitative notations arranged according to their function, as we find in an administrative form. The earliest attestation of addition as an operation with symbols is provided by the notation of totals of the entries on such tablets, usually inscribed on the reverse of the tablet containing these entries.

Based on their origin, the oldest written numerical notation systems exhibited, for a short period, very unusual characteristics. Since they initially still represented units of counting and measurements, and since the numerical relations between these units were dependent on the counted or measured objects, the numerical signs had no uniquely determined values. Their values resulted from their metrological context, and changed with the respective areas of application without any apparent attempt to attain unambiguous numerical values for the signs. From the viewpoint of modern arithmetic, the signs thus represented in different contexts different numbers (Damerow, 1995; Damerow and Englund, 1987). Some 500 years later, however, the system had developed into a sort of numerical sign system as is known from other early civilizations. In particular, the signs had now attained fixed numerical values.

Systems of Numerical Symbols in Early Civilizations

Most, if not all, advanced civilizations, in particular the Egyptian empire, the Mesopotamian city-states, the Mediterranean cultures, the Chinese empire, the Central American cultures, and the Inca culture, have, independent from each other or by adaptation from other cultures, developed systems of numerical symbols that exhibit similar semiotic characteristics. The basic symbols of these systems were the same as at the proto-arithmetical level, signs for units and not for numbers, but since they represented now complex metrologies they had necessarily also to deal with fractions. Furthermore, part of the systems were now complex symbol transformations, that is, arithmetical techniques such as Egyptian calculation using unit fractions (Chace, 1927), sexagesimal arithmetical techniques of the Babylonians (Neugebauer, 1934), transformations of rod numerals on the Chinese counting board (Lǐ and Dù, 1987), calendrical calculations in the pre-Columbian culture of the Maya (Closs, 1986; Thompson, 1960), or techniques of the use of knotted cords *(quipu)* as administrative tools by the Inca (Ascher and Ascher, 1971–1972; Locke, 1923).

The development of these numerical techniques was historically closely related with the administrative problems that had arisen through the concentration of economic goods and services in the governmental centers of early state organization (*see* State: Anthropological Aspects). The dramatic rise in the quantities of products had to be controlled, and the immense variety of decision-making implications had to be executed administratively (Høyrup, 1994; Nissen et al., 1993).

The rules dictating the use of numerical signs reflected this function; they corresponded to their meaning in the social context. The numerical sign systems thus exhibited a great variety of structures with no internal differentiation between rules representing a universal number concept and rules depending on the specific kind of symbolic representation and its function to control quantities in a specific social setting. It was only the cultural exchange between advanced civilizations with developed systems of numerical notations that created the preconditions for the differentiation between universal numbers and specific notations.

The Euclidean Conceptualization of Number

The emergence of an abstract concept of number as a consequence of cultural exchange is particularly well known from the ancient Greek culture which played a crucial role in the historical process of transforming the heritage of early Near Eastern civilizations into the intellectual achievements of the hegemonic cultures of the Hellenistic and Roman world. The reflective restructuring of divergent bodies of knowledge brought about new kinds of general concepts and, in particular, an abstract concept of number (*see* Reflexivity in Anthropology).

The oldest known examples of general propositions about abstract numbers, for instance, the statement that the number of prime numbers is infinite, are handed down to us through the definitions and theorems of Euclid's Elements (Heath, 1956). The relevance of these definitions and theorems was no longer based on their practical applicability, but only on their role within a closed system of mental operations. These operations reflected certain arithmetical activities in the medium of written language. In the case of the number concept these operations were the techniques of counting and tallying. Euclid defined a number as a 'multitude composed of units' and hence restricted the abstract concept of number to natural numbers. Platonism, in particular, dogmatically excluded from theoretical arithmetic all numerical structures that did not come under the Euclidian definition. Fractions of a number, for instance, and even more so irrational numbers such as the square roots of natural numbers that are not squares were not accepted themselves as numbers.

In order to circumvent the problems of incommensurability, a theory of proportions was developed, not for numbers, but for entities designated as 'magnitudes' (Heath, 1956). This theory served at the same time as a substitute for an expanded concept of number that could cover all the arithmetical activities and symbolic operations existing at that time.

At a practical level, however, operations with notations for other types of number systems had been developed and were used long before they were theoretically reflected in an extended number concept. Any exchange of goods by merchants or by administrators of a centralized economy implicitly involves fractions. The balancing of debits and credits in bookkeeping similarly involves negative numbers. Accordingly, arithmetical techniques for dealing implicitly or even explicitly with fractions or with negative and irrational numbers had thus been developed already in early civilizations long before they were reflected in an explicit number concept.

The Impact of Algebra

The problems resulting from the restrictive Greek number concept were aggravated with the renaissance of ancient mathematics in the Early Modern Era. In particular, the development of algebraic notation techniques and of analytical methods applied to continuous processes in mechanics made the distinction between numbers and magnitudes obsolete. The use of variables in order to solve equations was indifferent to this conceptual distinction. Fractions, negative numbers, irrational numbers, infinitesimals, and infinity became more and more accepted as numbers. Although infinitesimals were again expelled in the nineteenth century after their missing logical foundation resulted in serious contradictions in the conceptual system of mathematics, they were finally reintroduced in a new form by the nonstandard analysis of the twentieth century.

The introduction of algebraic notations thus contributed to the transformation of existing arithmetical techniques into conceptualized number systems, but it also initiated the creation of completely new ones. A new type of number system with no immediately obvious meaning emerged, reflecting no longer primary arithmetical techniques but rather abstract operations with variables and equations. A well-known example is provided by 'imaginary' square roots of negative numbers which occurred as solutions to higher degree equations. Such solutions, although calculated correctly according to rules for numbers, seemed to make no sense since squaring a number never results in a negative number. Nevertheless, it turned out to be possible to operate consistently with such imaginary numbers as if indeed they were, in fact, numbers. They were, thus, finally accepted as 'complex numbers.'

The justification of seemingly meaningless number systems was not the only problem posed by the development of mathematics in the nineteenth century. The term number became questionable for other reasons too. On the one hand, a productive new theory, even today called simply 'theory of numbers,' was created; it is, however, a discipline with a specific topic. The term number is in this theory restricted to (positive and negative) integers. On the other hand, a growing number of new types of mathematical systems emerged which, with greater or lesser emphasis, were called number systems, for example, 'algebraic numbers,' 'transcendental numbers,' 'higher order complex numbers,' 'quaternions,' 'biquaternions,' etc. In order to organize the hodgepodge of artificially-constructed number systems, a heuristic principle was formulated. The construction of number systems should follow the 'principle of the permanence of formal laws' of an *arithmetica universalis* (Hankel, 1867), that is, they should be constructed in such a way that some of the rules of ordinary numbers are preserved. It was hoped, in particular, that 'higher order complex numbers' of a certain type might serve as the foundation for an infinite number and virtually all possible number systems, an expectation which later turned out to be misguided.

Formalism and the Decline of Number

The formalistic approach, which in the nineteenth century led to the flourishing number systems and ascribed to them a key role in the foundation of mathematics, reduced their importance in the course of the structuralist rebuilding of mathematics in the twentieth century. In fact, the recombination of structures and operations of mathematical objects in order to create new ones with new properties made it more and more difficult to distinguish number systems from other systems of mathematical objects. The designation of an object as

a number was increasingly seen as merely an historical convention. Number systems were downgraded to instances of structures abstracted from them, in particular, of algebraic structures resembling addition and multiplication and topological structures determining distance and continuity. Not numbers, but 'sets' equipped with such structures, which were designated artificially as 'groups,' 'fields,' 'rings,' 'lattices,' 'compact spaces,' etc., became the building blocks of modern mathematics.

See also: Brain, Evolution of; Civilization, Concept and History of; Cultural Evolution: Overview; Human Cognition, Evolution of; Indigenous Knowledge: Science and Technology Studies; Knowledge, Anthropology of.

Bibliography

Ascher, M., Ascher, R., 1971–1972. Numbers and relations from ancient Andean Quipus. Archive for History of Exact Sciences 8, 288–320.

Blake, B.J., 1981. Australian Aboriginal Languages. Angus and Robertson, London.

Chace, A.B., 1927. The Rhind Mathematical Papyrus. Mathematical Association of America, Oberlin, OH.

Closs, M.P. (Ed.), 1986. Native American Mathematics. University of Texas Press, Austin, TX.

Damerow, P., 1995. Abstraction and Representation: Essays on the Cultural Evolution of Thinking. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Damerow, P., 1998. Prehistory and cognitive development. In: Langer, J., Killen, M. (Eds.), Piaget, Evolution, and Development. Erlbaum, Mahwah, NJ, pp. 247–270.

Damerow, P., Englund, R.K., 1987. Die Zahlzeichensysteme der Archaischen Texte aus Uruk. In: Green, M.W., Nissen, H.J. (Eds.), Zeichenliste der Archaischen Texte aus Uruk. Gebr. Mann, Berlin.

Dixon, R.M.W., 1980. The Languages of Australia. Cambridge University Press, Cambridge, UK.

Hankel, H., 1867. Theorie der complexen Zahlensysteme insbeson dere der gemeinen imaginären Zahlen und der Hamilton'schen Quaternionen nebst ihrer geometrischen Darstellung. Voss, Leipzig.

Heath, S.T.L. (Ed.), 1956. The Thirteen Books of Euclid's Elements. Dover, New York.

Høyrup, J., 1994. In Measure, Number, and Weight: Studies in Mathematics and Culture. State University of New York Press, Albany, NY.

Levy-Bruhl, L., 1923. Primitive Mentality. Macmillan, New York.

Lǐ, Y., Dù, S., 1987. Chinese Mathematics: A Concise History. Clarendon, Oxford.

Locke, L.L., 1923. The Ancient Quipu or Peruvian Knot Record. The American Museum of Natural History, New York.

Marshack, A., 1972. The Roots of Civilization: The Cognitive Beginnings of Man's First Art, Symbol and Notation. Weiden feld and Nicolson, London.

Neugebauer, O., 1934. Vorgriechische Mathematik. Springer, Berlin.

Nissen, H.J., Damerow, P., Englund, R.K., 1993. Archaic Bookkeeping: Early Writing and Techniques of Economic Administration in the Ancient Near East. Chicago University Press, Chicago, IL.

Saxe, G.B., 1982. Culture and the development of numerical cognition: Studies among the Oksapmin of Papua New Guinea. In: Brainerd, C.J. (Ed.), The Development of Logical and Mathematical Cognition. Springer, New York, pp. 157–176.

Schmandt-Besserat, D., 1992. Before Writing. University of Texas Press, Austin, TX.

Thompson, J.E.S., 1960. Maya Hieroglyphic Writing. University of Oklahoma Press, Norman, OK.

Wertheimer, M., 1925. Über das Denken der Naturvölker: Zahlen und Zahlgebilde. In: Wertheimer, M. (Ed.), Drei Abhandlungen zur Gestalttheorie. Palm and Enke, Erlangen, Germany.

O

Obedience (Research), History of

Stephen D Reicher, School of Psychology, University of St. Andrews, Fife, UK
Alexander Haslam, School of Psychology, University of Queensland, St. Lucia, QLD, Australia

Published by Elsevier Ltd.

Abstract

In this article we first trace the origins of Milgram's obedience studies in classic suggestion research and in the conformity studies of Solomon Asch. The basic Milgram paradigm is outlined, with particular emphasis on the so-called 'new baseline' condition. The findings in this and other conditions – which range from 0% to 100% obedience – are described. We then look at Milgram's rich speculations about the processes involved, which can be found in his experimental notebooks, archived at Yale University. We contrast this with the 'agentic state' explanation which Milgram later used to explain his findings. We also critique this explanation, showing that it cannot account for variations in obedience between conditions, nor the points at which people are most likely to disobey within conditions. We argue that its inadequacy lies in taking the relationship between experimenter and participant for granted, hence ignoring both when and how it is established. Drawing on some of Milgram's early speculations, we advance an 'engaged follower' perspective which explains obedience in terms of the identification of participants with the science of the study and with the experimenter as a representative of that science. We conclude by using this perspective to bring together both scientific and ethical dimensions of the Milgram's obedience studies.

More than any other topic in social psychology, perhaps, the study of obedience is dominated by one researcher – Stanley Milgram – and by one piece of research – his aptly named Yale Obedience Studies. To cite Muzafer Sherif "Milgram's obedience experiment is the single greatest contribution to human knowledge ever made by the field of social psychology, perhaps psychology in general" (cited in Takooshian, 2000, p. 10). But one underestimates Milgram's significance if one limits one's focus to the discipline of psychology. He has had considerable impact across a range of disciplines from sociology to theology, from history to law. And his influence goes far beyond the academy. It has marked popular culture. Milgram's work has been disseminated through television documentaries, theater plays, art shows, and even a film – *The Tenth Level* – in which Star Trek's Captain Kirk (William Shatner) plays a fictionalized Milgram. Moreover, as Peter Novick documents, since the 1960s Milgram's ideas have been "invoked in discussing everything from the Vietnam War to the tobacco industry, and, of course, reflecting back on discussions of the Holocaust" (2000, p. 137).

The link to the Holocaust is not coincidental. For Milgram's interest in, and research on, issues of obedience, was clearly rooted in his own experiences as the child of Jewish parents of East European origins, growing up in the United States during the World War II, acutely concerned as news of the unfolding tragedy filtered through. At his Bar-Mitzvah in 1946, he said "As I come of age and find happiness in joining the ranks of Israel, the knowledge of the tragic suffering of my fellow Jews throughout war-torn Europe makes this also a solemn event and an occasion to reflect upon the heritage of my people – which now becomes mine" (cited in Blass, 2004, p. 8).

Through his life, Milgram continued to reflect on this heritage and how it came to pass. His early research drew on the famous conformity studies of Solomon Asch, which in turn drew on a tradition of 'suggestion' research which goes back to the prehistory of social psychology (see Asch, 1955). In the late nineteenth century, specifically in France, researchers became increasingly interested in hypnotism, first as a means of diagnosing those 'hysterics' who were particularly prone to ideas implanted by others, later as a model of social functioning in general (see Barrows, 1981; Goldstein, 1982). This idea came to a head in the work of Gabriel Tarde who famously remarked that "la societe, c'est l'imitation, et l'imitation, c'est une espece de somnambulisme" (society is imitation and imitation is a form of somnambulism; Tarde, 1901, p. 95). (Translation by the authors. Note that Asch, in his 1955 paper, uses a variant of this quote: "social man is a somnambulist.") The idea, then, is that, especially when immersed in large groups with others, people cannot help but conform to the prevailing opinion.

Asch's aim was to investigate this idea of group pressure more closely and more systematically. Unlike previous studies, which had largely studied social opinions or aesthetic judgments, he addressed whether people would alter their judgments of physical stimuli (the length of a line) when confronted with the views of others. Asked to match one of several options with a target line, fully 76% of participants chose the wrong option at least once when it had previously been selected by other respondents (who were actually confederates of the experimenter). Overall, participants accepted the majority's false judgment on just over a third of the trials (36.8% to be precise; see Asch, 1955). As Jetten and Hornsey (2012) document, Asch himself provided a nuanced analysis

of his findings. On the one hand he bemoaned the fact that the tendency toward conformity is so strong that intelligent people can be led to call white, black. On the other hand, he stressed that we should not underestimate the human capacity for nonconformity. That is, however striking and surprising one-third of the responses where people went along with the confederates might be, we must not let them obscure the two-thirds of responses where they did not. To cite Asch himself "anyone inclined to draw too pessimistic conclusions from this report would do well to remind himself that the capacities for independence are not to be underestimated" (1955, p. 35).

Over time, however, Asch's caveats have been forgotten and the accounts of his studies have increasingly laid a one-sided stress on conformity over independence. Indeed, we shall see that this 'conformity bias' (Moscovici, 1976) is more general in social psychology and indeed leads to an equally one-sided account of Milgram's findings. But this is to get ahead of ourselves. For now, it is important to trace the passage from Asch's conformity studies to Milgram's obedience studies.

At first Milgram was interested in whether one could find national differences in the level of conformity. Would some conform more than others? But he remained dissatisfied with the basic paradigm itself. While the general phenomenon of conformity seemed important and relevant to societal phenomena, this was undermined by the fact that the topic of conformity – line judgments – seemed so trivial and so distant from events like the Holocaust (Blass, 2004). He began to ask whether people would conform when it came to more significant behaviors, like aggressing against another person. Moreover, would they do so if one substituted Asch's group of confederates for a single figure of authority? Putting these two things together, Milgram came up with the question: Just how far would a person go under the experimenter's order? That was the point where the obedience studies were born. Milgram later remarked "it was an incandescent moment" (cited in Blass, 2004, p. 62).

The Obedience Studies

If Milgram's obedience studies – or, at least, certain aspects of them – are exceedingly well known, the extensive efforts that Milgram spent in gradually shaping the paradigm in order to make it work smoothly is far less familiar. The basic outline of the studies was determined fairly early on. A participant would come along to the laboratory, ostensibly to take part in a learning experiment. Once there he (there was one study using women, but the great majority of respondents were male) would be greeted by the experimenter, allocated to the role of 'teacher' and instructed to administer an ascending series of electric shocks each time the learner (actually an actor who was a confederate of Milgram) made an error on the learning task. But beyond that, everything remained to be settled: how would the experimenter behave and how would he urge the participant on?; how would the participant be persuaded that real shocks were being administered?; how would the learner react as the shock level went up?

To start with, Milgram went to great lengths to choose an experimenter whom he considered to be sufficiently authoritative. He eventually settled on John Williams, "'a stern, intellectual looking man' who played his experimenter's 'role in a cold austere manner'" (from Milgram's notes, cited in Russell, 2011, p. 160). He carefully crafted an elaborate script in which the experimenter introduces the study, explains its scientific importance and its scientific originality. At one critical juncture the script reads "actually, we know *very little* about the effect of punishment on learning, because almost no truly scientific studies of it have been made in human beings" (Milgram, n.d., emphasis and capitals as in the original). He even crafted the things the experimenter should say in case participants seemed reluctant to continue giving shocks. There were four so-called 'prods' to be delivered in sequence: Prod 1 "Please continue"; Prod 2 "The experiment requires that you continue"; Prod 3 "It is absolutely essential that you continue"; Prod 4 "You have no other choice, you must go on" (Milgram, 1974, p. 21). Only if the participant continued to resist after the fourth prod would the session be terminated. (It should be noted, however, that the experimenter did not always stick strictly to the script and even employed impromptu devices – such as pretending to check up on the learner and confirming that he was alright – in order to persuade participants to continue (see Gibson, 2013).)

Milgram also devoted considerable attention to crafting his famous shock generator which now is kept at the Center for the History of Psychology at Akron University in Ohio. As Russell (2011) reveals, early prototypes had 12 switches increasing in 30-V steps from 0 V to 330 V, each with verbal labels attached, the final one initially reading 'LETHAL,' later changed to 'Extreme Shock: Danger.' The final version changed from 12 to 30 switches, increasing in 15-V steps up to a maximum of 450 V, the extreme now labeled with an ambiguous but ominous XXX (see Milgram, 1974, p. 28). Milgram was particularly concerned with making the machine seem impressive and professional. He was particularly delighted when "two electrical engineers examined the instrument and failed to realize it was a simulated device" (Russell, 2011, p. 157).

Finally, Milgram paid as much attention to choosing and scripting the 'learner' as he did the experimenter. His choice of who would play this role had less to do with experience – James McDonough, the successful candidate "probably could not act face-to-face," wrote Milgram (cited in Russell, 2011, p. 159) – than with his demeanor. Milgram thought McDonough to be "perfect as a victim" because he was "mild and submissive; not at all academic" (Russell, 2011, p. 159).

In early variants of the study, the learner said and did very little as the shocks were administered. Indeed, it appears that, in some of the initial pilot studies, nothing at all was heard from the learner as the shocks were delivered and here all the participants went up to the 450-V maximum – although it appears that Milgram thought this was atypical and reflected the nature of the Yale students ('Yalies') who had served as participants (A. Milgram, 2000). Henceforth he would use a broad cross section of the (mainly male) population in his studies. In the first of the published studies (Study 1 in Milgram's, 1974 summary of his work, called the 'Remote' condition), the learner, who was in a separate room to the participant, was restricted to banging on the wall at 300 V and then ceasing to respond after 315 V. All participants went up to the 300-V mark and 65% continued to 450 V.

The feedback became more elaborate in Study 2, the 'Voice-feedback' condition). Now the complaints of the learner could be heard clearly through the wall. Groaning would start at 75 V. At 150 V the learner would demand repeatedly "Let me out of here." The groans, screams, and demands would increase until, at 300 V the learner exclaimed "You can't hold me here," and, at 315 V, "I'm no longer part of this experiment." After the 330-V mark, he would fall silent and stop responding in any way.

This variant has sometimes been called the 'Baseline condition' and it elicited similar levels of total obedience to Study 1 (62.5% vs 65%), although here, people start to drop out much earlier (from 135 V). Further small amendments were made for Study 5, which Milgram terms the 'New baseline condition' in his 1974 book *Obedience to Authority*, but which (for reasons that will soon become clear) he originally dubbed the 'Coronary condition.' This is the best known of Milgram's studies, the one which generally appears in text books, and often the only condition that is mentioned when Milgram's studies are discussed. The full script is reproduced in *Obedience to Authority* (pp. 56–57) and it differs from Study 2 only in that, at various critical junctures (the 150, 195, and 330-V points), the learner refers to the fact that he has a heart condition. For instance, at 150 V the learner exclaims "Ugh!!! Experimenter! That's all. I told you I had heart trouble. My heart's starting to bother me now. Get me out of here please. My heart's starting to bother me. I refuse to go on. Let me out" (Milgram, 1974, p. 56). But still, 65% of participants carry on to the 450-V endpoint.

These results came as a major surprise. A large proportion of ordinary Americans, like Yale students it seemed, were willing to administer potentially lethal electric shocks to people simply because they had made errors on a learning task. This was completely at odds to how ordinary Americans thought that they would act. When he asked a sample of 110 people how they themselves would respond – a mixture of professionals, students, and middle-class adults – none thought they would go beyond 300 V and the mean point at which people thought they would break off was between 120 V and 135 V (whereas only 1 out of the 40 participants in the 'New baseline' condition stopped before 150 V). Moreover, the actual behavior was also completely at odds with the way that psychiatric professionals estimated that ordinary people would act. They predicted that most people would not go beyond 150 V and that only 0.125% – the pathological fringe – would go on to the end. Milgram appeared to have discovered a surprising and uncomfortable truth about human behavior. How could it be explained?

Explaining Obedience

In his early writings, Milgram wrestled long and hard with the significance of his findings. For instance, in his first published account of the research in 1963, Milgram lists 13 possible factors of significance: the reputation of Yale University where the research took place; the fact that the research is ostensibly about advancing a worthy purpose – how to help people learn; the way in which people voluntarily make a commitment to participate in the research; the ambiguity about the prerogatives of the experimenter and the rights of the learner; and so on.

Over time, however, Milgram's position changed. Even though his 1974 book mentions some of the possibilities that appear in the early writings, these are overshadowed by another explanation. This is the idea that, in the face of authority, people slip into an 'agentic state.' This term denotes the idea that people cease to act in terms of their own purposes and instead act as an agent for someone else's. In everyday terms "(t)he ordinary person who shocked the victim did so out of a sense of obligation – a conception of his duties as a subject – and not from any peculiarly aggressive tendencies" (Milgram, 1974, p. 6). Once in this state, people are so focused on how well they do the bidding of that authority that they lose concern for all else. In particular, they have little interest in the consequences of their actions and the plight of their victims. Essentially, they become bureaucratic zombies. They are able to commit harm largely because they do not think about what they are doing.

In formulating the concept of the agentic state, Milgram drew heavily on the ideas of Hannah Arendt. Arendt famously observed the trial of Adolf Eichmann in Jerusalem (which, coincidentally, took place at the exact same time as Milgram was conducting his obedience studies in 1961). Like all the others waiting in the courtroom, Arendt assumed that this man – the chief bureaucrat of the Holocaust – would appear monstrous. But what they saw was a seemingly insignificant character, middle-aged and balding, slightly hunched, fastidiously taking notes on the proceedings. Eichmann, she concluded, committed the most horrendous acts for the most banal of reasons – to please his superiors. "Eichmann," she says, "had no motives at all. He merely, to put the matter colloquially, never realized what he was doing" (Arendt, 1963/1994, p. 287).

Arendt summed the lesson that she drew from this in a ringing phrase "the lesson of the word-and-thought-defying banality of evil" (Arendt, 1963/1994, p. 252). Milgram concurred. Immediately before introducing his own definition of the 'agentic state' (cited above) he observed that "After witnessing hundreds of ordinary people submit to the authority in our own experiments, I must conclude that Arendt's conception of the banality of evil comes closer to the truth than one might dare imagine" (1974, p. 6).

The combination of Arendt's and Milgram's analysis has proved remarkably powerful and extremely resilient. The historical and psychological threads, wound together, provided a far more authoritative account than either could alone. They seemed to define something fundamental about human nature that could explain many, if not all, instances of toxic conformity.

Questioning the 'Agentic State'

In recent years, the 'banality of evil' account has come under increasing fire from historians of the Holocaust. They point out that Eichmann and his ilk, far from being passive 'desk murderers' who had little insight into the realities of mass murder, knew exactly what they were doing, believed in what they were doing, and worked independently and creatively to deport then massacre their victims. These were true Nazi believers who saw Jews as an enemy which had to be exterminated (e.g., Cesarani, 2004; Lozowick, 2002; Stangneth, 2014).

How did Arendt miss this? Vetlesen (2005) argues that, at his trial, Eichmann put on a deliberate act to undermine the prosecution case that he was a distinctively cruel individual. Vetlesen suggests that, in characterizing Eichmann as 'merely thoughtless,' Arendt "adopts the very self-presentation he cultivated" (p. 5).

Equally, Milgram's agentic state account has received mounting criticism from psychologists, even among his admirers (Blass, 2004). Indeed Darley goes so far as to say that "I was startled and appalled by… the odd and pseudoscientific/pseudophysiological concept of the agentic state" (1992, p. 207). In addition to the problem of positing a distinct mental state, which Darley alludes to here, there are two other fundamental problems with Milgram's account. The first is that it simply cannot account for the different levels of obedience that were observed in different variants of Milgram's paradigm. As we have already noted, in these, between 0% and 100% of participants were willing to go to the 450-V mark. But why would there be this variability if there is a general tendency for people to slip into an agentic state in the face of authority?

Giving Milgram the benefit of the doubt, one could treat the agentic state as a continuum and argue that obedience will be higher the more that authority imposes itself upon us. But there is no evidence for this. Thus, focusing on the perceptions of the participant, Mantell and Panzarella (1976) found no relationship between obedience and the extent to which people handed over responsibility for their actions. Equally, focusing on the actions of the experimenter, there is (if anything) a negative relationship between the amount of authority exerted by the experimenter and the degree of obedience. This is evident when one looks at the effects of the different prods that were used to urge on reluctant participants. Recall that only one of these, the fourth prod, was a direct assertion of authority ('You have no other choice, you must go on'), the others are more like requests or justifications. But examples used in Milgram's book suggest that Prod 4 was generally unsuccessful. In one oft-cited example (1974, p. 28), the participant, a Professor of Old Testament Liturgy responds by saying "If this were Russia maybe, but not in America." What is more, a recent replication of the Milgram studies (with the sole difference that participants only went as far as the 150-V point), participants refused to continue *every time* Prod 4 was used (Burger, 2009). Whatever the studies do show, then, it is emphatically not that people always follow orders, no matter how severe.

Moving on, the second fundamental problem with the agentic state account is that it reduces the richness of the obedience paradigm to the sole relationship between experimenter and participant. What is a multivocal reality in which people are torn between two voices invoking different obligations and placing incompatible demands on the participant (thereby creating the tension and the drama of the situation) is rendered univocal, as if the participant were in thrall to the sole voice of the experimenter and was unaffected by everything else. But, again, this cannot explain Milgram's own findings. When one looks closely at the points where people are most likely to disobey the experimenter, they are precisely the points where the learner makes his most powerful pleas – notably 150 V where (as we have described above) he first demands to be released from the study and, to a lesser degree, at 315 V where he declares that he is no longer part of the experiment (see Packer, 2008). The voice of the learner – as well as the experimenter – clearly does matter.

The deeper problem here is that, in this process of reduction, and by taking the influence of the experimenter for granted, all the important questions on which Milgram deliberates in his earlier notes – about which voice is listened to, about how the different conditions impact on this (and through what processes), and about how the experimenter (or the learner) achieves precedence – are instantly ruled out. These are questions of general relevance for psychology. For issues of social significance – say, how we should respond to immigrants – we do not come to an understanding through a process of silent contemplation. Rather we are assailed on all sides – in conversations with friends, headlines in newspapers, debates on the radio, television, and the internet – by various voices telling us what to think and do. The key question is how we allocate authority and credence among these voices. Psychological research, however, generally ignores this. Most of the time people are placed in a laboratory and asked to respond on their own. Occasionally, and at most, we look at how they respond to a single influence message. Milgram's paradigm is about the only one in which we are placed between several voices. It bears the promise of helping us understand the general nature of social understanding (and not only its toxic outcomes). The agentic state explanation kills that promise stone dead.

Forward to the Past: From 'Bureaucratic Sleepwalker' to 'Engaged Follower'

One obvious way to get beyond the impasse of the 'agentic state' explanation is to return to some of Milgram's earlier explanations. There are two elements in particular that can be drawn upon. First, Milgram points to the way in which the study is sold by the experimenter (and seen by the participant) as a worthy enterprise (Milgram, 1963, point 2 in the discussion). Second, he points to the way in which the participant comes to the study in order to form a relationship with the experimenter and hence this takes precedence over his relationship with the learner (point 3 in the discussion). In a later paper Milgram (1965) develops this last idea and talks about the way in which different experimental conditions (e.g., degree of spatial proximity) may impact on the formation of a group bond between the experimenter and participant, and hence affect the degree of obedience. This, then, begins to be a means of explaining the different levels of obedience in the different studies.

Putting the two elements together, one could argue that people are more likely to shock when they identify with a worthy scientific enterprise and with the experimenter as a representative of that enterprise. They are less likely to shock when they identify with victim as a fellow citizen. Put differently, one could characterize Milgram's compliant participants (like the Nazi bureaucrats) as 'engaged followers' who were far from unaware of the pain they were inflicting but who, ultimately, proceeded because they thought it was the right and noble thing to do.

There are a number of convergent pieces of evidence to support this approach. First, and most directly, using an

adapted version of the Milgram paradigm in which one is asked to commit increasingly difficult acts of derogation against members of other groups, participants proceed further when they are made to identify with science than when they are not (Birney et al., 2013).

Second, if one asks contemporary subjects to indicate how much they think they would identify with (1) the experimenter and (2) the learner in Milgram's different experimental variants (using the descriptions of these variants in his 1974 book), then there is a strong negative correlation between identification with the learner and the proportion of participants in the actual studies who went up to 450 V, a strong positive correlation between identification with the experimenter and the proportion of participants who went on to the endpoint, and an even stronger positive correlation between relative identification with the experimenter (i.e., identification with the experimenter minus identification with the learner) and this proportion (Reicher et al., 2012).

Third, when one examines the script used to introduce the study, Milgram devotes much time to explaining why the study is unique and important. That is, the participants do not come to the idea that they are involved in a noble enterprise of their own accord. Rather the experimenter actively construes the enterprise as noble in order to bind them in. Moreover, when the experiments were completed and Milgram wanted to reassure the participants, he uses almost identical language concerning the unique and important nature of the study (though this time the argument is that it provides new understandings of the obedience process rather than new understandings of the learning process). What is more, looking at participants reactions to this reassurance (which are contained in Box 44 of the Milgram archive) there is a clear relationship between belief in the scientific endeavor, on the one hand, and contentment with one's own participation, willingness to participate again, and willingness to persuade others to participate on the other (Haslam et al., 2013). Both during and after the study, then, what people did and how they felt was a function of the extent they could be signed up for the cause of science. This did not happen of itself. Compliance could not be taken for granted. Rather, Milgram and his experimenter succeeded in producing 'obedience' through their skills as recruiting sergeants for the cause.

Science, Ethics, and a Last Word

Any mention of how participants felt, and how Milgram reassured them, necessarily takes us to the issue of ethics. This is something that no discussion of the obedience research can or should ignore – for the studies are as famous (or rather notorious) for the cost to participants as they are for the gain to science. In general the two issues are treated separately leading to two camps, the one which condemns Milgram for his ethical misdemeanors, the other which praises him for his scientific contributions.

The ethical controversy was sparked immediately after Milgram's first publication in 1963 when Baumrind (1964) published a coruscating commentary in which she accused him of manipulating, embarrassing, and discomforting his participants while showing no concern for their suffering. She further dismissed his suggestions that any tensions were dissipated before people left the laboratory. But even this was mild compared to some other comments. The renowned child psychologist and concentration camp survivor, Bruno Bettelheim, declared the studies to be "vile" and "in line with the human experiments of the Nazis" (cited in Blass, 2004, p. 123). A newspaper editorial described them as "open eyed torture" (cited in Blass, 2004, p. 121).

Milgram confronted these criticisms head on. He collected systematic data after the experiment about the experience and reactions of participants. These largely bear out his contention that, whatever stresses were apparent during the studies, there is little evidence of distress after the studies. So, for instance, nearly 85% of participants were glad or very glad to have taken part and only just over 1% were sorry or very sorry. Similarly, only 7% of participants said that they had been bothered quite a bit by their experience. Overall, Milgram contends that most people were positive about the study and felt that it was time 'well-spent' (Milgram, 1974, p. 198).

One could obviously respond by arguing that even 7% of people feeling bothered, or 1% of people feeling sorry with what they had been led to do is unacceptable. But perhaps this misses the larger point. The controversy between Milgram, Baumrind, and others proceeds on the assumption that any problems have to do with whether participants were distressed – so if they were not distressed, there is no problem. But is there really no problem if there was no distress? Should not people have been distressed? Should not they worry about having apparently delivered lethal shocks to an innocent victim, and should not we worry that they are prepared to persuade others to participate in such studies in the future? That is, Milgram may have reassured his participants but at the cost of getting people to believe that it is fine to suffer for science, and fine to make others suffer for science. Arguably, this is the real ethical issue with the obedience studies, but it is one that has been almost totally overlooked. We are rightly concerned with whether individuals suffer in our studies, but, there is potentially a problem if this comes at the expense of consolidating ideologies which facilitate collective suffering in society.

To conclude, then, Milgram's obedience studies undoubtedly reveal, as he intended, a phenomenon of great consequence. However it is arguable that, when it comes to understanding this phenomenon, we have been led up a blind alley – or rather two blind alleys, one scientific and one ethical. On the scientific side, Milgram's agentic state account leads us away from asking how people are led to engage with a toxic cause and how authorities advance such causes. On the ethical side, the 'individual harm' debate leads us away from asking about the dangers of advancing causes that reconcile people to collective harm-doing.

From this perspective the scientific and ethical dimensions of the obedience studies are not two separate issues, but two sides of the same coin. Progress depends upon bringing them back together. It is precisely by understanding how people can be led to identify with the powerful and justify harm-doing against the powerless that we can disrupt the ease with which people abuse or else are complicit with abuse. This could be more uncomfortable than we imagine, for it may come very close to home. Indeed, we now look in horror at the past use of

racial supremacism as a justification for excluding and even exterminating the racial other. What we may be less aware of, and what Milgram's studies point to, is the continuing power of an ideology of scientific progress as a means of getting people to inflict harm and yet still feel noble.

See also: Compliance and Legal Authority; Obedience: Social Psychological Perspectives.

Bibliography

Arendt, H., 1963/1994. Eichmann in Jerusalem: A Report on the Banality of Evil. Penguin, New York.

Asch, S.E., 1955. Opinions and social pressure. Scientific American 193, 31–35.

Barrows, S., 1981. Distorting Mirrors: Visions of the Crowd in Late Nineteenth Century France. Yale University Press, New Haven, CT.

Banyard, P., Grayson, A., 2000. Introducing Psychological Research: Seventy Studies that Shape Psychology. Palgrave, London.

Baumrind, D., 1964. Some thoughts on ethics of research: after reading Milgram's "behavioral study of obedience". American Psychologist 19, 421–423.

Birney, M.E., Haslam, S.A., Reicher, S.D., 2013. In the name of scientific progress: evidence that identification with science increases obedience in an experimental analogue of the Milgram paradigm. Under review European Journal of Social Psychology.

Blass, T., 2004. The Man Who Shocked the World: The Life and Legacy of Stanley Milgram. Basic Books, New York.

Burger, J., 2009. Replicating Milgram: would people still obey today? American Psychologist 64, 1–11.

Cesarani, D., 2004. Eichmann: His Life and Crimes. Heinemann, London.

Darley, J.M., 1992. Social organization for the production of evil. Psychological Enquiry 3, 199–218.

Gibson, S., 2013. 'The last possible resort': a forgotten prod and the *in situ* standardization of Stanley Milgram's voice-feedback condition. History of Psychology 16, 177–194.

Goldstein, J., 1982. The hysteria diagnosis and the politics of anticlericalism in late nineteenth-century France. Journal of Modern History 54, 209–239.

Haslam, S.A., Reicher, S.D., Millard, K., McDonald, R., 2013. 'Happy to have been of service': box 44 in the Yale archive as a window into the engaged followership of participants in Milgram's 'obedience' experiments. Under review British Journal of Social Psychology.

Jetten, J., Hornsey, M., 2012. Conformity: beyond Asch's line judgment studies. In: Smith, J.R., Haslam, S.A. (Eds.), Social Psychology: Revisiting the Classic Studies. Sage, London, pp. 76–90.

Lozowick, Y., 2002. Hitler's Bureaucrats: The Nazi Security Police and the Banality of Evil (H. Watzman, Trans.). Continuum, London.

Mantell, D.M., Panzarella, R., 1976. Obedience and responsibility. British Journal of Social and Clinical Psychology 15, 239–245.

Milgram, A., 2000. My personal view of Stanley Milgram. In: Blass, T. (Ed.), Obedience to Authority: Current Perspectives on the Milgram Paradigm. Erlbaum, Mahwah, NJ, pp. 1–7.

Milgram, S. n.d. Preliminary Talk. Milgram Archive, Box 46. Sterling Memorial Library, Yale University.

Milgram, n.d. (f). Relative Identification with Authority or Victim. Milgram Archive, Box 46. Sterling Memorial Library, Yale University.

Milgram, S., 1963. Behavioral study of obedience. Journal of Abnormal and Social Psychology 67, 371–378.

Milgram, S., 1965. Liberating effects of group pressure. Journal of Personality and Social Psychology 1, 127–134.

Milgram, S., 1974. Obedience to Authority. Tavistock, London.

Moscovici, S., 1976. Social Influence and Social Change. Academic Press, London.

Novick, P., 2000. The Holocaust and Collective Memory. Bloomsbury, London.

Packer, D.J., 2008. Identifying systematic disobedience in Milgram's obedience experiments: a meta-analytic review. Perspectives on Psychological Science 3, 301–304.

Reicher, S.D., Haslam, S.A., Smith, J., 2012. Working toward the experimenter: reconceptualizing obedience within the Milgram paradigm as identification-based followership. Perspectives on Psychological Science 7, 315–324.

Russell, N., 2011. Milgram's obedience to authority experiments: origins and early evolution. British Journal of Social Psychology 50, 140–162.

Stangneth, B., 2014. Eichmann before Jerusalem. Knopf, New York.

Takooshian, H., 2000. How Stanley Milgram taught about obedience and social influence. In: Blass, T. (Ed.), Obedience to Authority: Current Perspectives on the Milgram Paradigm. Lawrence Erlbaum, Mahwah, NJ.

Tarde, G., 1901. L'opinion et la foule. Les Presses universitaires de France, Paris.

Vetlesen, A.J., 2005. Evil and Human Agency. Cambridge University Press, Cambridge.

Obedience: Social Psychological Perspectives

Thomas Blass, University of Maryland, Baltimore, MD, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 10765–10768, © 2001, Elsevier Ltd.

Abstract

Obedience is the act of carrying out the requests or commands of a person of higher status within a social hierarchy. The most distinctive feature of the social-psychological approach to obedience is the primary role it accords to situational determinants. The foremost social-psychological contribution to understanding obedience was that of Stanley Milgram, whose experiments demonstrated that a surprisingly high proportion of normal people would obey orders to hurt an innocent person. Milgram saw his experimenter as representing a legitimate authority – a person who is deemed to have the right to issue orders to others who, in turn, feel obligated to obey them. According to French and Raven, the power of authorities or other agents of influence can derive from five additional sources besides their legitimacy. These are: reward, coercive, expert, referent, and informational power. Kelman has pointed to three processes that help overcome the normal human reluctance to hurt another person: authorization, routinization, and dehumanization. Kelman and Hamilton distinguish among three broad individual orientations toward political authority: rule-oriented, role-oriented, and value-oriented. A current research emphasis is the study of harmful obedience in the workplace. Available evidence suggests that obedience to authority is one of the universals of social behavior, transcending both time and place.

Obedience refers to the act of carrying out requests or commands emanating from a person of higher status within a social hierarchy. It is to be distinguished from compliance – which also refers to a form of acquiescence, but does not necessarily involve a situation of status inequality. Obedience is also different from conformity (*see* Conformity: Sociological Aspects). In the latter, the source of influence comes from others of equal status to the target of influence; such as, a group of peers or other members of one's society or culture (*see* Social Influence, Sociology of). Additionally, while obedience – as well as compliance – is a result of direct influence attempts, conformity is usually the product of implicit, indirect group pressures felt by the individual to be impinging on him or her. Thus, while obedience represents the maintenance or creation of social inequality, conformity is a social leveler.

Distinctive Feature of the Social-Psychological Approach

Social-psychological perspectives on obedience to authority see it as a necessary feature of organized social life, contributing importantly to its efficient and orderly functioning. Their most distinctive feature is the primary role accorded to situational factors. Grounded in Kurt Lewin's ahistorical approach, which emphasized contemporaneous determinants of behavior, social psychologists tend to focus on situ- ational determinants to account for variations in obedience. This is in contrast to the personality psychologist who focuses on inner, stable dispositions to account for obedience and its variations. A prime example of the latter is the personality construct of authoritarianism, which consists of a number of interrelated elements, including a powerful propensity to obey authority (Adorno et al. 1950, Altemeyer 1988).

The Contribution of Stanley Milgram

The most important social-psychological contribution to understanding obedience was that of Milgram (1963, 1974). Although his aim was to study destructive obedience, his findings and insights are readily applicable to more mundane and benign forms of obedience. Milgram conducted a series of laboratory experiments at Yale University from July 1961 to May 1962 to answer the following question: If one person orders another to harm a third person, under what conditions would the second obey the first person's orders and under what conditions would he or she defy them? In his experiments, subjects who had been recruited for a study of memory and learning were instructed to give what appeared to be increasingly dangerous shocks to a learner every time he made a mistake on a word-matching task. The subjects delivered the shocks by pressing one of 30 switches arranged horizontally along the front panel of a box-shaped 'shock-generator.' Each subsequent switch was labeled with higher and higher voltages of shock, in 15-volt steps, beginning with 15 volts for the first switch and ending with 450 volts for the last one. The rule was that the subject teacher had to increase the shock level one switch at a time each time the learner made an error. In some of the conditions, as shock levels became more intense and painful, the learner responded with shrieks and complaints of increasing urgency and demanded that the experimenter terminate the experiment. After a certain point, the learner became silent, suggesting the possibility that he may have passed out.

The main finding was an unexpectedly high degree of obedience. As many as 65 percent of the subjects, ordinary residents of New Haven and Bridgeport, Connecticut, were fully obedient, that is, continued administering what they were led to believe were increasingly painful and dangerous shocks until the 450-volt maximum in response to the experimental authority's orders. In actuality, the shocks were not real and the learner was

a confederate who made errors according to a pre-arranged schedule. Some of Milgram's additional findings were: Obedience declined as the distance between teacher and learner was reduced and when the distance between the experimenter and subject was increased; obedience increased when the subject was only indirectly involved in punishing the learner, that is, he read the wordlist, but someone else pressed the switch; the subject's obedience rate diminished drastically after seeing two of his peers refuse to continue and thereby providing models of disobedience; and men and women were equally obedient, although the latter reported experiencing greater tension. Milgram believed that such extreme obedience to potentially destructive commands is made possible by the acceptance of the authority's definition of the situation and by the subject's entry into an 'agentic state,' a mental state whose central feature is the relinquishing of responsibility to the person in charge. According to Milgram, his experimenter represented a legitimate authority, a person who is seen as having the right to issue orders and as someone whom we feel an obligation to obey.

If the legitimacy of an authority is seen to be of prime importance in enabling him or her to be a powerful source of influence, a logical question is: How do authorities create and maintain their legitimacy? Recent research by Tyler (1997) has addressed this question. He distinguishes between two approaches that suggest why people see authorities as legitimate, and, hence, feel an obligation to obey them: an instrumental model and a relational model. According to the first approach, a prime determinant of peoples' acceptance of the legitimacy of a leader is whether they are seen as a source of benefits or desirable outcomes. According to the second approach, it is the nature of the treatment people receive from authorities – in particular the integrity and caring that mark their actions – that will determine their perceived legitimacy. Tyler finds that, although both instrumental and relational factors have an influence, the latter is the more dominant factor in judgments of an authority's legitimacy.

Other Social-Psychological Approaches

According to French and Raven (1959) and Raven (1999), a person's authority or power to influence others can derive from other factors besides from his or her legitimacy. Specifically, according to these authors, social power can derive from six different bases. Besides legitimate power, they are: reward, coercive, expert, referent, and informational power. When an authority or persuader has reward power, people obey or yield because they see the former as a potential source of rewards. Coercive power is operating when people yield out of fear of punishment. Persons are said to possess expert power when others go along with their wishes because of their expertise or special knowledge. An authority or influencer is said to have referent power when people acquiesce to his or her commands or wishes because they like or identify with him or her. And finally, an authority or other source is wielding informational power when people obey or are influenced because they find the information provided by that person to be intrinsically convincing or compelling.

According to Kelman (1973), there are three social processes that help weaken the normal human reticence against inflicting harm on another and thereby facilitate people's obedience to destructive orders: authorization, routinization, and dehumanization.

Through authorization, individuals are relieved of the necessity of assessing the morality of the actions they are directed to carry out. When an action is authorized, the usual moral considerations fade into the background, and the individual no longer feels responsible for his or her actions, nor any guilt, should those actions be harmful. Routinization – the absorption of the individual in the mechanical and routine details that are necessary for carrying out the authorized actions – minimizes the possibility that he will worry about its consequences. Dehumanization – viewing the target of one's harmful actions as less than human – further insures that the morality of the commands issued by an authority will not be questioned.

Going beyond obedience in authority–subordinate relationships between individuals, Kelman and Hamilton (1989) analyzed the effects of political authority – the nation–state and its various institutions – on the individual citizen. A central feature of their analysis is a threefold distinction among people's broad orientations to political authority – rule, role, and value orientations – that they see as potential mediators of the nature of citizens' responses to morally questionable policies or destructive orders, as the following quote indicates:

> For the *rule-oriented*, moral principles are largely irrelevant both to the actions of the citizen in an authority context and to the actions of the state in pursuit of its primary purposes: ensuring public order and national security. This view has a cynical overtone. The state cannot afford the luxury of concerning itself with higher principles, and individuals can be expected to do only what they must in order to stay out of trouble. For the *role-oriented*, special moral principles apply to the citizen in relation to the state as well as to the state itself. The morality of obedience—the obligation to obey and support the government—at the level of the individual citizen and the duty to pursue the national interest at the level of the state both override standard moral considerations. Finally, for the *value oriented*, principles of personal morality continue to govern the behavior of individuals as citizens, and adherence to moral principles is fundamental to the legitimacy of state action. Actions by and on behalf of the state may not always conform precisely to the expectations that govern personal relations, but they have to be chosen and evaluated with reference to standard moral principles (Kelman and Hamilton 1989, italics added).

A Contemporary Problem: Harmful Obedience in the Workplace

An important implication that Milgram (1974) had drawn from his research was about the potential dangers inherent in organizational behavior: When an individual 'merges... into an organizational structure, a new creature replaces autonomous man, unhindered by the limitations of individual morality, freed of humane inhibition, mindful only of the sanctions of authority.' This observation about the evaporation of individual responsibility in hierarchical organizations anticipated the kinds of tragedies that later resulted from car manufacturers' employees' readiness to

carry out their bosses' directives to produce unsafe automobiles, and the tobacco companies' employees' acquiescence to policies aimed at deceiving the public.

A current emphasis in research on obedience to authority is the systematic study of this 'darker side' to behavior within hierarchical organizations such as businesses and companies. Thus, for example, Hamilton and Sanders (1995), in a cross-cultural study, examined the factors that would affect judgments about a subordinate's responsibility when his company's actions could result in physical harm to the public. Brief et al. (2000) showed that, when a company president's orders not to hire minority group members was accompanied by a business justification – 'Our organization attempts to match the characteristics of our representatives with the characteristics of the population to which they will be assigned' – subordinates were more likely to discriminate against Black job candidates than when no such justification was provided.

The Pervasiveness of Obedience to Authority

Obedience to authority would seem to be one of the universals of social behavior, transcending both time and place, as suggested by two quantitative analyses. One analysis, using the findings of Milgram's standard conditions and those of all the replications conducted by others, correlated the amount of obedience found in each study with its year of publication and found absolutely no relationship between the two variables (Blass 2000). Another analysis (Blass 1998), also involving Milgram's standard conditions and their replications, reveals that the average obedience rate found in US studies (61 percent) was very similar to the average rate of obedience obtained in replications carried out in other countries (66 percent). As Milgram once wrote:

> We do not observe compliance to authority merely because it is a transient cultural or historical phenomenon, but because it flows from the logical necessities of social organization. If we are to have social life in any organized form—that is to say if we are to have society—then we must have members of society amenable to organizational imperatives (letter to Professor Alan Elms, Sept 25, 1973; The Stanley Milgram Papers, Yale University Archives).

Future Directions

Recently, there has been a discernible shift in emphasis away from obedience toward defiance – both as a focus of research (e.g., Modigliani and Rochat 1995, Rochat and Modigliani 2000) and in the implementation of changes in the workplace that facilitate employees' disobedience of illegitimate or destructive orders from their superiors, such as the creation of hot lines and the placement of ombudsmen. In the future, one can expect a substantive increase in these kinds of developments.

Bibliography

Adorno, T.W., Frenkel-Brunswik, E., Levinson, D.J., Sanford, R.N., 1950. The Authoritarian Personality. Harper, New York.

Altemeyer, B., 1988. Enemies of Freedom: Understanding Right-wing Authoritarianism. Jossey-Bass Publishers, San Francisco.

Blass, T., 1998. A cross-cultural comparison of studies of obedience using the Milgram paradigm. Paper presented at the annual convention of the American Psychological Association, San Francisco.

Blass, T., 2000. The Milgram paradigm after 35 years: Some things we now know about obedience to authority. In: Blass, T. (Ed.), Obedience to Authority: Current Perspectives on the Milgram Paradigm. Erlbaum Associates, Mahwah, NJ, pp. 35–59.

Brief, A.P., Dietz, J., Cohen, R.R., Pugh, S.D., Vaslow, J.B., 2000. Just doing business: Modern racism and obedience to authority as explanations for employment discrimination. Organizational Behavior and Human Decision Processes 81, 72–97.

French Jr., J.R.P., Raven, B.H., 1959. The bases of social power. In: Cartwright, D. (Ed.), Studies in Social Power. Institute for Social Research. University of Michigan, Ann Arbor, MI, pp. 150–167.

Hamilton, V.L., Sanders, J., 1995. Crimes of obedience and conformity in the workplace: Surveys of Americans, Russians, and Japanese. Journal of Social Issues 51 (3), 67–88.

Kelman, H.C., 1973. Violence without moral restraint: Reflections on the dehumanization of victims and victimizers. Journal of Social Issue 29 (4), 25–61.

Kelman, H.C., Hamilton, V.L., 1989. Crimes of Obedience: Toward a Social Psychology of Authority and Responsibility. Yale University Press, New Haven, CT.

Milgram, S., 1963. Behavioral study of obedience. Journal of Abnormal and Social Psycholog 67, 371–378.

Milgram, S., 1974. Obedience to Authority: An Experimental View. Harper & Row, New York.

Modigliani, A., Rochat, F., 1995. The role of interaction sequences and the timing of resistance in shaping obedience and defiance to authority. Journal of Social Issues 51 (3), 107–123.

Raven, B.H., 1999. Reflections on interpersonal influence and social power in experimental social psychology. In: Rodrigues, A., Levine, R.V. (Eds.), Reflections on 100 Years of Experimental Social Psychology. Westview Press, Boulder, CO.

Rochat, F., Modigliani, A., 2000. Captain Paul Grueninger: The chief of police who saved Jewish refugees by refusing to do his duty. In: Blass, T. (Ed.), Obedience to Authority: Current Perspectives on the Milgram Paradigm. Erlbaum, Mahwah, NJ.

Tyler, T.R., 1997. The psychology of legitimacy: A relational perspective on voluntary deference to authorities. Personality and Social Psychology Review 1, 323–345.

Obesity and Labor Market Performance

Susan L Averett, Lafayette College, Easton, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Obesity is the most pressing public health problem in developed countries and is rapidly becoming so in developing countries. Obesity is associated with a host of comorbidities and strains health care systems around the globe. Obesity is also associated with poor labor market outcomes such as lower wages and a lower probability of employment. While this negative correlation between obesity and wages/employment is well documented across numerous countries, sorting out causality from correlation has been challenging. Most studies that credibly disentangle correlation from causation find a negative and significant causal effect of obesity on wages and employment, particularly for women.

Introduction

We live in an increasingly obesogenic environment. A progressively more sedentary lifestyle coupled with low-cost, highly palatable calorically dense food has led to rising rates of obesity. Obesity is rapidly becoming the most urgent public health problem in the world. Obesity rates have been on the rise worldwide and even developing countries have not been spared the obesity epidemic. In fact, in many developing countries, the proportion of the population that is overweight or obese exceeds that which is underweight (Mendez et al., 2005). Recent estimates from the World Health Organization (WHO) indicate that 65% of the world's population lives in countries where overweight and obesity kill more people than does underweight and worldwide obesity rates have nearly doubled since 1980 (WHO, 2013).

Obesity is well known as a risk factor for a considerable number of diseases and medical conditions. These include diabetes, heart disease, stroke, hypertension, arthritis, sleep apnea, some cancers, and asthma. These conditions are responsible for a large and growing portion of medical care costs. Currently, obesity rivals, and according to some estimates surpasses, smoking as the leading preventable cause of death worldwide (Hennekens and Andreotti, 2013).

In addition to its negative impact on health and medical expenditures, there is a growing evidence that the obese, particularly women, pay a penalty in the labor market in that they are less likely to be employed and when employed they have lower wages (e.g., Averett, 2011, 2014; Averett et al., 2013). In this article, I first discuss the reasons why the obese may have lower wages and/or a lower probability of employment. I then outline the strategies and findings of the empirical studies that have attempted to determine if obesity is the cause of adverse labor market outcomes.

Obesity and Labor Market Outcomes: What Is the Relationship?

In the mid-1990s, economists and other social scientists began to explore the possibility of a causal relationship between obesity and labor market outcomes whereby they hypothesized that obesity might be the cause of lower wages and poor employment prospects (e.g., Averett and Korenman, 1996). Since that time, this literature has grown in both volume and sophistication of method.

Establishing a clear causal connection between obesity and labor market outcomes is not as straightforward as it might initially appear. Clearly, the aforementioned health conditions that the obese may suffer as a result of their obesity may limit their productivity in the labor market or may directly limit the type of work they can do and this may be a primary cause of lower wages for the obese. This may be particularly true for the morbidly obese and less of a concern for those whose obesity is not as severe. It likely also depends on the occupation – one can imagine that the obese are more incapacitated in occupations that require more physical exertion or agility but less impaired in an office setting, which is more sedentary. In addition, in some occupations cultural norms dictate that a thin person is appropriate for the job – fashion modeling is an oft cited example. In fact, in a column for Business Week magazine in 1998, Nobel Prize winning economist Robert Barro argued that an employee's physical appearance is a legitimate job qualification to the extent that it is valued by customers and coworkers. While he was speaking of beauty, in modern society beauty and obesity tend to be inversely correlated.

Obesity has also been directly linked to absenteeism. Given the health problems faced by the obese, they are more likely to be absent from work. Using US data, Cawley et al. (2007) reported that obese women were 61% more likely to miss work than were their nonobese counterparts independent of their occupation. Results for men were mixed and varied by occupation. He estimated the annual costs of obesity-related job absenteeism in the United States to be as high as 4.3 billion dollars in 2004. Using data from Europe, de Galdeano, 2007 confirms that for women, the costs of absenteeism associated with obesity are statistically significant, although for men obesity was not correlated with higher absenteeism.

Obesity may also negatively affect the acquisition of what economists call human capital – the skills that one brings to the work place that make them productive and hence enhance their employability and wages. Obesity could potentially affect educational attainment through several channels. The most obvious is a poor diet that fails to provide adequate nourishment. This idea that the obese are misnourished is gaining traction in the development economics literature where there is

 http://dx.doi.org/10.1016/B978-0-08-097086-8.94039-X

discussion of a 'nutrition transition' whereby poorer countries adopt the dietary habits of more affluent countries and thereby also the obesity problem of more affluent countries. Yet, this misnourishment has also been noted as a problem in more developed countries including the United States. A lack of important micronutrients may impair the academic performance and/or brain development of children directly reducing their human capital achievement.

Another way that obesity may reduce human capital acquisition is that teachers may discriminate against obese students making it difficult for obese children to perform well because they may feel stigmatized and even if they perform well, they may face discrimination in grading by teachers. There is a growing body of empirical evidence using data from the United States and Europe that indicates that obese children and adolescents have lower academic outcomes thus limiting their future productivity although this finding is not universal across all studies and more research is necessary on this topic (see Scholder et al., 2012, for further discussion). There is also growing support that the medical consequences of obesity may operate to cause physiologic brain changes that could impair general cognitive function or performance in some cognitive areas leading to lower educational attainment. All of the above suggests that obesity itself may be a cause of lower human capital attainment, which would translate into lower wages in the labor market.

Labor market discrimination is another possible explanation for the link between obesity and wages or employment. In this case, it is not that the obese are less productive than the nonobese but rather that they face discrimination in the labor market. Such discrimination could come directly from employers who have a subjective dislike of obese persons and thus do not want to hire them or if they hire them, they do so only if they are less expensive than nonobese employees. This 'taste-based' theory of discrimination was put forth by Nobel prize winning economist Gary Becker (1957) who theorized that this type of discrimination emanating from employers was not sustainable in the long run due to competitive pressures because employers who were unwilling to hire the obese would have higher costs and hence go out of business. It is also possible that customers may be less willing to purchase goods and services from the obese. That customers might have a subjective distaste toward the obese is consistent with the considerable evidence that the obese are stigmatized. In the United States there is evidence that discrimination against the obese is as prevalent as racial or gender discrimination.

Another source of labor market discrimination against the obese may be statistical discrimination. In contrast to taste-based discrimination, this is not based on any subjective dislike of the obese but rather refers to a situation where employers do not have perfect information about potential employees. This leads employers to make hiring decisions about individuals based on the characteristics of the group to which they belong. If employers believe obesity is associated with a lack of self-discipline and professionalism then they may choose a nonobese person to hire even if the obese person they are interviewing is highly productive.

Finally, the obese might earn lower wages because of their relatively greater health care costs. This is particularly salient in the United States where employers are responsible for a considerable portion of their employees' health care costs. Employers can compensate workers with a combination of wages and benefits. Economic theory posits that under certain conditions, employees pay for their own benefits in terms of higher wages. The employer is generally indifferent as to the exact wage/benefit combination as long as profits remain the same. Since the obese are likely to incur higher health care costs, their employer may pay higher premiums for employer-sponsored health insurance and therefore compensate the obese with lower wages to maintain the same cost of the wages/benefit mix (Bhattacharya and Bundorf, 2009).

All of the above explanations point to obesity as being a *cause* of lower wages. However, the causality may run the other direction such that those with lower wages become obese in part because they cannot afford healthy food and thus rely on calorie-dense fast foods. In most developed countries, there is an income–obesity gradient such that those with lower incomes are more likely to be obese (e.g., Averett, 2012; García Villar and Quintana-Domeque, 2009). It is also possible that an adverse labor market outcome leads to depression and/or low self-esteem and that this subsequent depression or low self-esteem leads to weight gain. Conversely, the obese, particularly women, may believe their marriage market prospects are low, and hence may invest more heavily in labor market–oriented human capital and actually have higher wages as a result of their obesity. The potential for reverse causality may be particularly relevant when employment is the outcome because obesity can be the cause of substantial work disabilities as noted earlier.

In contrast to the above explanations, it may be that there is no causal relationship in either direction between obesity and wages but that a third factor creates a correlation between obesity and labor market outcomes. For example, obesity could be correlated with undesirable personality traits that affect the productivity of the obese and hence their wages and their ability to find employment. In this case, the negative correlation between obesity and wages is not causal. These personality traits might include laziness, inability to delay gratification, lack of self-control, or poor social skills – all traits that the general public has often associated with obesity.

Another unobserved variable that might link obesity to poor labor market outcomes is that the obese may have a higher subjective discount rate; i.e., a higher degree of impatience, which implies a stronger preference for present consumption relative to future consumption. Thus, they may overeat today without thinking about the consequences that occur in the future. There is evidence that the marginal rate of time preference has increased leading to the obesity epidemic because individuals increasingly discount the future more heavily (Komlos et al., 2004). In addition, some people have high-discount rates for very short horizons but have considerably lower discount rates for longer horizons, thus indicating that some people are less patient in immediate future choices than in distant future choices. When people behave this way, economists say that their preferences are time-inconsistent. There is evidence that persons with a stronger tendency toward time-inconsistent preferences have higher BMIs (Ikeda et al., 2010).

These unobserved traits (e.g., laziness, high-discount rates, or time-inconsistent preferences) may make some individuals

both more likely to overeat and to invest less in wage-enhancing human capital leading to a negative correlation between wages and obesity. In other words, it may be this third factor that is the cause of their poor labor market outcomes; not their obesity directly. If this is the case then obesity and labor market outcomes would have what economists call a spurious correlation.

Obesity and Labor Market Outcomes: Empirical Methods and Findings

The fact that obesity is a growing concern around the world generating poor health and high costs associated with treating its comorbidities has led researchers to attempt to determine if obesity is the *cause* of adverse labor market outcomes. A major challenge in this line of research has been sorting out causation from correlation. Scholars have made use of complex econometric methods to identify there is any causal relationship at all between obesity and wages/employment.

Ideally, to measure the effect of obesity on labor market outcomes one would conduct a randomized controlled trial in which obesity was randomly assigned by the investigator and subsequent labor market outcomes were observed. Such an experiment would, of course, be unethical so the bulk of the research in this area relies on observational data collected through surveys. Most research in this area relies on the BMI as the measure of adiposity because it is easily calculated from height and weight. The BMI is calculated as weight (kg)/height^2 (meters). According to the WHO's international standards, a BMI of 25 or more signals overweight, and a BMI of 30 or more indicates obesity.

Early research linking obesity to labor market outcomes used ordinary least squares (OLS) regression – a statistical method that allows researchers to relate the outcome of interest (e.g., wages) to obesity while controlling for the influence of other factors that predict wages including education and work experience in order to isolate the effect of obesity on wages. The vast majority of these studies find a negative relationship between obesity and the labor market outcome of interest, particularly but not exclusively for women. Yet, although these models generally control for a host of observable socioeconomic and demographic factors, OLS cannot address the potential for reversed causality nor can it rule out that a third factor, unobserved to the researcher, is the cause of both the obesity and the labor market outcome. Thus, studies relying on OLS regression can only be taken as suggestive of a causal effect.

Given the limitations of OLS, researchers have used a variety of more sophisticated econometric techniques to determine if there is a causal effect of obesity on labor market outcomes. In what follows, I provide a brief description of these methods and the findings of the studies that employ them. Readers interested in a more technical discussion of each of the methods described below should consult Averett, 2011.

Many researchers relate an early measure of obesity to a later labor market outcome arguing that this temporal ordering of events cannot reflect reverse causality. Studies using this method report a negative association between obesity and wages, particularly for women. Results from these studies are most convincing if this early measure of BMI is taken prior to the labor market entrance of the individual so that it is not possible for the BMI to be affected by the labor market outcome.

Comparing siblings, one who is obese and the other who is not, has also been used to examine the relationship between obesity and wages on the assumption that differences between siblings removes the variation in weight attributable to a shared family environment. However, even after differencing across siblings there remains variation in weight ascribed to genetic makeup unshared by siblings and the variation in weight attributable to nongenetic factors. To the extent that these factors are not captured by the observable factors controlled for in the model such as education, or if parents treat children in the same family differently as a response to early academic potential that is related to future wages, this method will yield biased estimates of the effect of obesity on wages. Other researchers have used monozygotic twins as another way to remove unobservable genetic factors correlated with obesity. In this case, the researcher can control completely for genetic propensity to be obese since monozygotic twins share 100% of their genes, however, sample sizes of monozygotic twins tend to be small and thus the results of these studies may have limited external validity.

Longitudinal data (i.e., data where the same individual is followed over time) can be particularly useful because they enable the researcher to control for unobservable or hard to measure factors that might affect both obesity and the labor market outcome of interest as long as these unobservable factors are constant over time. As noted earlier, unobservable characteristics that might influence obesity and labor market outcomes could include an individual's rate of time preference, or a personality trait such as laziness. Essentially, with this type of data, each individual serves as their own control. However, if these unobservable factors affecting both weight and wages are time-varying they will not be accounted for with this method.

If the concern is that there is reversed causality or that there are important unobservable factors that vary over time then the method of instrumental variables (IV) is preferred. The drawback of this technique is that this technique involves finding a variable (the instrument) that is highly predictive of an individual's obesity but completely unrelated to that same individual's labor market outcomes except through its effect on obesity. As noted earlier, there are many potential factors that are related to both the probability that someone is obese and their labor market outcomes. Thus, finding plausible instruments is challenging.

Given that roughly half the variation in weight across individuals is genetic in origin; many studies have used the weight of a biological relative as an instrument. Yet, the validity of this instrument is questionable if both the respondent and the biological relative are affected by a common household environment that is also directly correlated with their labor market outcomes. In other words, it is possible that a substantial part of the genes responsible for obesity are also responsible for other factors that affect labor market outcomes, such as willingness to delay gratification (time discount rate) or other unobserved characteristics. Cawley (2004) argues for the validity of this instrument citing studies that have been unable to detect any effect of common household environment on body weight. Economists have also begun to use an individual's own genetic information as IV. Such information on specific genes

linked to obesity should be a strong instrument for obesity and is likely to be used more in the future, if it is available in the data.

Other instruments that have been used include the presence of other obese persons in the household, being an oldest child, and whether or not a parent has been prescribed medication for obesity. These are less likely to be valid instruments in that they are also likely correlated with an individual's labor market outcome independent of their association with obesity. In addition, a few studies have used the average BMI and the proportion obese who live in the same geographic area as the individual as an instruments. However, given that location of residence is a choice, this instrument could be related to occupational choices or other labor market characteristics rendering them invalid.

Using the methods described above, numerous studies based on US data have assessed the relationship between wages and obesity or wages and employment. One of the most robust findings from this research is that obese white women generally earn less than their nonobese counterparts even after using a wide array of statistical techniques as described above. That the effect is mainly found for women may indicate a form of cosmetic discrimination, which may exist in a culture that correlates thinness with attractiveness for women.

In addition to the United States, scholars around the world have conducted a number of studies relating obesity to labor market outcomes using data from Australia, China, Denmark, Finland, France, Germany, Iceland, Sweden, Taiwan, and the U.K. Similar to the case with US data, a negative and statistically significant effect of obesity on wages and employment is often, but not always found, particularly for women even using IV or other methods.

Even studies using data comparing several European countries have found that there is an obesity penalty in the labor market (e.g., García Villar and Quintana-Domeque, 2009). Assessing if the obese earn lower wages in Europe is particularly challenging because European labor markets are generally characterized by more compact wage distributions, a considerable share of the labor force in the public sector, and more rigid wage structures, which leaves little room for employer discretion with respect to wages.

The size of the effect found in this literature is economically meaningful. For example, the results in Cawley (2004) indicate that a difference in weight of 2 standard deviations among white females is associated with a difference in wages of 9%. This is equivalent to 1.5 years of education or 3 years of work experience.

The vast majority of the research in this area uses the BMI cutoff of 30 or over as the measure of obesity. But, cultural norms regarding social acceptable body size may play a role in whether or not there is a labor market penalty associated with obesity. In some cultures, thinness may be prized more than in others. Thus, perhaps the best measure is a relative measure of obesity – how much an individual's weight differs from the social norm may determine if they face lower wages. There is some empirical evidence for the relative threshold but more work remains to be done on this question. Other studies have noted that women's wages are highest at a body weight well below the BMI of 30 indicating a 'premium for thinness' that indicates that looks, rather than poor health, is driving the obesity penalty that we observe for women.

It is possible that cultural norms for thin body types are inversely related to the obesity prevalence in a society. If this is the case, in societies with relatively high obesity rates, the obese should have a lower wage penalty. Conversely, in countries which place a premium on frequent social interactions we should expect to observe a higher labor market penalty for obese persons; i.e., the obese will earn lower wages. There is some evidence in support of this hypothesis, particularly for women (e.g., García Villar and Quintana-Domeque, 2007).

The studies discussed above have all relied on BMI as the measure of whether an individual is obese or not. This is not surprising as social science data sets that contain detailed labor market outcomes were often not intended to collect health information and so it is fortunate that some collected information on height and weight. It has been established that BMI itself is not an accurate measure of obesity, particularly for males because it does not distinguish fat from muscle (Burkhauser and Cawley, 2008). In addition, in most data sets height and weight are self-reported and thus may be systematically misreported. Burkhauser and Cawley note that the "social science research on obesity would be enriched by greater consideration of more accurate measures of fatness" (Burkhauser and Cawley, 2008: p. 520).

Given the limitations of the BMI as a measure of adiposity, some studies have used fat-free mass (FFM) and body fat (BF) to examine the link between obesity and wages (e.g., Wada and Tekin, 2010). Similar to the studies using BMI, these studies often find that those who have higher BF and lower FFM to have lower wages. One study found that no association of BMI with wages or employment for men but did find that FFM and waist circumference were predictive of wages and employment. For women, all three measures were predictive of lower wages and a lower probability of employment (Johansson et al., 2009). This evidence is consistent with the concern that BMI is not an accurate measure of fatness for men and heightens awareness of the need for more data sets that include these more accurate measures of adiposity.

Another measure of obesity that is sometimes available is waist circumference. Waist circumference is a measure of central obesity, which occurs when excess fat is centrally distributed and some argue that it has at least two advantages over BMI. First of all, it is a stronger predictor of morbidity and mortality and second, it is a visible measure of fatness, which might be interpreted by employers, customers, or coworkers as an unattractive physical attribute. Some studies that have used waist circumference to measure obesity have also found evidence that the obese earn lower wages.

As noted earlier, the 'gold standard' for ascertaining causality is the randomized experiment. While such a study is ethically not feasible here, a related type of study is an audit study. In an audit study, fictitious individuals who are identical, except for some outwardly observable characteristic, apply for jobs and any group differences in outcomes are interpreted as reflecting discrimination. Rooth (2009) sent two equivalent applications to advertised job openings with the only difference being that one had a picture of a person who was not obese while the other had a picture of that same person, which had been digitally modified to make the same person appear obese. The advantage of this approach is that it helps ensure that other productivity-related characteristics of the job applicant are held constant – after all it is the same applicant just with a different picture. In order to isolate the effect of obesity and not confound it with a poor

job interview, the outcome in this study was the call-back rate (offer for an interview) rather than an actual job offer or wages. The results from this study indicated that the obese were less likely to be called back for an interview than the nonobese, although there were some differences in call-back rates across occupations and gender.

The variation by occupation in call-back rates across the obese and the nonobese found by Rooth (2009) could plausibly be interpreted as evidence of customer discrimination. If employer discrimination were responsible for these differences, we would not expect them to exist across occupations. If customer discrimination is the cause of the lower wages among the obese, then we would expect the effect to be stronger in occupations that involve significant customer contact. This hypothesis is bolstered by the findings of other researchers who have reported that the negative relationship between the BMI and wages is larger in occupations requiring interpersonal skills and that it is often obese women in sales and service occupations who suffer the greatest wage penalties.

As noted earlier, the Becker model states that discrimination should be incompatible with profit maximization. As a simple, though indirect test of this theory, Brunello et al., 2009, used data from nine European countries to calculate the correlation between wages and obesity rates, which they found to be negative in each country. They then compared this correlation to an OECD-created measure of product market regulation (as a proxy for the degree of competition in the economy) and found that the negative penalty for obesity is indeed higher in those countries where there is less competition as would be predicted by the theory of employer discrimination. In contrast, Lundborg et al. (2010) using a large sample of Swedish men, concluded that the 18% lower wages of obese men compared to nonobese men were more likely due to statistical discrimination as it appeared in their data that employers used obesity as a proxy for physical fitness since in their data, the obesity penalty could be explained almost entirely by differences in cognitive skills, noncognitive skills, and physical fitness.

There is a strong, negative correlation between obesity and employment and wages, particularly for women. Moreover, a host of careful econometric studies using data from many countries suggests that obesity is likely the *cause* of these adverse economic outcomes. The available evidence suggests that discrimination plays at least some role in these adverse outcomes although more research on this is necessary. While a discussion of policy alternatives is beyond the scope of this article (interested readers can consult Averett, 2014, for a discussion of policy alternatives), this social disparity across the obese and the nonobese has captured the attention of policy makers around the globe who have increasingly stepped up efforts to tackle the causes of obesity. The research summarized here makes it clear that in addition to improving people's health, reducing obesity will likely also improve the economic well-being of the obese.

See also: Beauty and the Labor Market; Discrimination, Economics of; Erotic Capital; Happiness and Work; Health Economics; Obesity and Mortality; Obesity: Cultural and Biological Factors; Value of Life, Economics of; Weight Regulation.

Bibliography

Averett, S.L., Korenman, S., 1996. The economic reality of the beauty myth. The Journal of Human Resources 31 (2), 304–330.

Averett, S.L., 2011. Labor market consequences: employment, wages, disability, and absenteeism. In: Cawley, J. (Ed.), The Oxford Handbook of the Social Science of Obesity. Oxford, New York, pp. 531–552.

Averett, S.L., 2012. Obesity, poverty and the ability to pay for calories. In: Jefferson, P. (Ed.), The Oxford Handbook of the Economics of Poverty. Oxford, New York.

Averett, S.L., 2014. Obesity and Labor Market Outcomes. IZA World of Labor. Available at: http://wol.iza.org/articles/obesity-and-labor-market-outcomes.pdf.

Averett, S., Argys, L., Kohn, J., 2013. Immigrants, wages and obesity: the weight of the evidence. In: Constant, Amelie F., Zimmermann, Klaus F. (Eds.), The International Handbook on the Economics of Migration.

Bhattacharya, J., Bundorf, K., 2009. The incidence of the healthcare costs of obesity. Journal of Health Economics 28 (3), 649–658.

Brunello, G., Michaud, P., Sanz-de-Galdeano, A., 2009. The rise of obesity in Europe: an economic perspective. Economic Policy 24 (59), 551–596.

Burkhauser, R.V., Cawley, J., 2008. Beyond BMI: the value of more accurate measures of fatness and obesity in social science research. Journal of Health Economics 2, 519–529.

Becker, G., 1957. The Economics of Discrimination. University of Chicago Press, Chicago.

Cawley, J., Rizzo, J., Haas, K., 2007. Occupation-specific absenteeism costs associated with obesity and morbid obesity. Journal of Occupational and Environmental Medicine 49 (12), 1317–1324.

Cawley, J., 2004. The impact of obesity on wages. Journal of Human Resources 39, 451–474.

de Galdeano, A.S., 2007. An economic analysis of obesity in Europe: health, medical care and absenteeism costs. Documentos de trabajo (FEDEA) 38, 1–31. Available at: http://www.fedea.es/pub/papers/2008/dt2007-38.pdf (accessed 24.06.13.).

García Villar, J., Quintana-Domeque, C., 2009. Income and body mass index in Europe. Economics & Human Biology 7 (1), 73–83.

García Villar, J., Quintana-Domeque, C., 2007. Obesity, employment and wages in Europe. In: Bolin, J.K., Cawley, J. (Eds.), Advances in Health Economics and Health Services Research, the Economics of Obesity, 17, 187–217.

Hennekens, C., Andreotti, F., 2013. Leading avoidable cause of premature deaths worldwide: case for obesity. The American Journal of Medicine 126 (2), 97.

Ikeda, S., Kang, M., Ohtake, F., 2010. Hyperbolic discounting, the sign effect, and the body mass index. Journal of Health Economics 29 (2), 268–284.

Johansson, E., Bokckerman, P., Kiiskinen, U., Heliovaarta, M., 2009. Obesity and labour market success in Finland: the difference between having a high BMI and being fat. Economics and Human Biology 7 (1), 36–45.

Komlos, J., Smith, P., Bogin, B., 2004. Obesity and the rate of time preference: is there a connection? Journal of Biosocial Science 36 (2), 209–219.

Lundborg, P., Nystedt, B., Rooth, D., 2010. No Country for Fat Men? Obesity, Earnings, Skills, and Health among 450,000 Swedish Men. IZA Discussion Paper No. 4775.

Mendez, M.A., Monteiro, C., Popkin, B., 2005. Overweight exceeds underweight among women in most developing countries. The American Journal of Clinical Nutrition 81 (3), 714–721.

Rooth, D., 2009. Obesity, attractiveness, and differential treatment in hiring a field experiment. Journal of Human Resources 44 (3), 710–735.

Scholder, S., Smith, G., Lawlor, D., Propper, C., Windmeijer, F., 2012. The effect of fat mass on educational attainment: examining the sensitivity to different identification strategies. Economics and Human Biology 10, 405–418.

Wada, R., Tekin, E., 2010. Body composition and wages. Economics & Human Biology 8 (2), 242–254.

World Health Organization, 2013. Obesity and Overweight. Fact Sheet No 311. Available at: http://www.who.int/mediacentre/factsheets/fs311/en/ (accessed 13.05.14.).

Relevant Websites

http://www.oecd.org/health/49716427.pdf – OECD Obesity Update, 2012.

http://www.who.int/mediacentre/factsheets/fs311/en/ – World Health Organization, 2013. Obesity and Overweight. Fact Sheet No 311.

Obesity and Mortality

Neil K Mehta, Rollins School of Public Health, Emory University, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Obesity is now considered to be a major contributor to premature mortality globally. Despite great interest in this topic, there remains much uncertainty over the most appropriate methods that should be applied to obtain accurate estimates of the obesity–mortality association. This article reviews the current evidence on the obesity–mortality association. I discuss pertinent methodological issues and recent advances in our understanding of this association. I then discuss the obesity–mortality association in an international context and potential secular changes in this association.

Introduction

Population scientists are often interested in how behavioral factors shape the risk of dying. One such behavioral factor that is receiving much attention is obesity. While the causes of obesity include both behavioral and nonbehavioral determinants, there is a general consensus that the rising level of obesity globally is a result of shifting patterns of diet and physical activity. Obesity is now often considered to be a major contributor of premature mortality in high-income countries and is of increasing concern in low- and middle-income countries.

The purpose of this article is to review our current understanding of the magnitude of the association between obesity and mortality. Despite the great interest in the mortality consequences of obesity, there remains much uncertainty in the most appropriate methods that should be applied to estimate this association. Estimates of the number of deaths attributable to obesity and obesity's influence on life expectancy, both of which are of primary interest to population scientists, nearly always rely on estimates of relative risks derived from survey data linked to prospective death records. This stands in contrast to approaches used to estimate cigarette smoking's effect on these aggregate mortality indicators. For cigarette smoking, researchers have developed indirect techniques to estimate the fraction of deaths attributable to this risk factor. These indirect techniques use deaths from lung cancer as an indicator of the overall mortality consequences of smoking (Peto et al., 1992; Preston et al., 2010). To date, no analogous indirect techniques have been developed for obesity. Therefore, much of our understanding of the aggregate mortality consequences of obesity is based on estimates of relative risks for obesity derived from survey data. I focus this article on the various methods used to estimate these relative risks and how the choice of methods influences our understanding of the aggregate mortality consequences of obesity.

The organization of this article is as follows. I begin with a brief overview of some salient methodological concerns in estimating the obesity–mortality association. I next turn to a review of recent evidence on the role of 'obesity histories' in helping to better characterize this association. I then discuss the obesity–mortality association in an international context and potential secular changes in the obesity–mortality association, respectively. I limit the discussion primarily to studies that use body mass index (BMI) as it is the most widely available indicator of weight status in survey data. In national populations, the correlation between BMI and other weight status indicators such as waist circumference and waist-to-hip ratio is generally high (Flegal et al., 2009; Vazquez et al., 2007). Current guidelines define obesity as a BMI ≥ 30.0 kg m^{-2} and overweight as a BMI of 25.0–29.9. In this article, I use the term 'obesity' in a general sense to refer to a BMI that is higher than optimal.

Methodological Concerns

Causal Identification

Because estimates of relative risk for obesity are based on observational studies, unobserved selection into the obese category is an important threat to causal identification. However, unlike other 'exposures' commonly examined in the social sciences, evidence supporting the causal effect of obesity on mortality can also be found in results from experiments conducted across species and from human clinical studies. Two large studies of bariatric surgery indicate that those who underwent surgery experienced 20–40% lower mortality compared to 'control' groups over a 7–15 year period (Adams et al., 2007; Sjöström et al., 2007). Neither study, however, was based on a random allocation of subjects into treatment/control groups nor was conclusive in isolating weight loss versus other factors as the direct cause of the mortality reduction. Experimental studies in nonhumans indicate that caloric restriction is associated with large reductions in mortality (Colman et al., 2009; Weindruch and Sohal, 1997; Weindruch and Walford, 1988). Vasselli et al. (2005), for example, show that obese rodents who were randomized to a weight loss diet experienced a 45% lower mortality compared to rats that remained on an obesity-sustaining diet over a 3 year period.

Instrumental variable approaches have also provided insights into the causal effects of obesity on health in observational settings (Cawley and Meyerhoefer, 2012; Nordestgaard et al., 2012; Richmond et al., 2014; Smith et al., 2009). Smith et al. (2009) use offspring BMI as an instrument for own BMI in a sample of Swedish parent–son pairs. They conclude that the magnitude of the obesity–mortality association has been underestimated in studies using conventional regression techniques. A similar conclusion was reached by

Nordestgaard et al. (2012) examining the association of BMI and ischemic heart disease in a sample of Danish adults and using genetic variation as an instrument.

Reverse Causation

Perhaps the most discussed challenge in estimating the obesity–mortality association in observational studies is that of reverse causation due to the presence of disease. Many diseases increase the risk of dying and can result in low weight. Diseases for which weight loss is common include many cancers, heart failure, renal failure, and chronic obstructive pulmonary disease (Hu, 2008). Reverse causation results in relative risks for obesity that would be downwardly biased, but the evidence supporting the extent of bias due to reverse causation is largely indirect. One approach has been to eliminate those with preexisting illnesses from the sample. Findings from large epidemiological cohorts in the United States suggest that when those with preexisting illnesses and smokers (who disproportionately experience disease) are eliminated, the relative risks for obesity increase by roughly 20–50% (Adams et al., 2006; Berrington et al., 2010; Calle et al., 1999). Such increases after similar eliminations, however, have not been observed universally (Flegal et al., 2007; Mehta and Chang, 2009). In addition, it is not clear if the removal of large segments of the sample is an effective method of assessing reverse causation as it may introduce additional selection effects that can change the obesity–mortality association independent of reverse causation (Banack and Kaufman, 2014; Flegal et al., 2011; Preston and Stokes, 2014).

Another indirect approach used to account for reverse causation is to eliminate the initial years of prospective mortality follow-up guided by the idea that the biasing effects of reverse causation are most pronounced in the early years after survey collection. A related approach is to allow the relative risks to vary by time in study, which allows for so-called 'time in study' effects. Evidence from studies that have addressed time in study effects is mixed with many studies reporting little to no change in the relative risks for obesity after eliminating the early years of follow-up (Allison et al., 1999; Flegal et al., 2007; Mehta and Chang, 2009). Some studies indicate that the relative risks for obesity grow stronger with increasing time in study (Zajacova and Burgard, 2012).

Alley et al. (2010) using the Baltimore Longitudinal Study of Aging, find that sustained weight loss begins 5–10 years before a cardiovascular death and 3 years before a cancer death in adults ages 60 and over. Thus, although the evidence on how large a role that reverse causation plays is not conclusive, reverse causation should be considered as an important biasing factor, particularly in older populations that suffer from a high burden of disease.

Variations in Relative Risks by Age and Smoking Behavior

Many studies highlight important variations in the relative risks for obesity across population subgroups. I focus here on variations by age and smoking status. The methods used to account for variations in the relative risks for obesity by these two variables have contributed to earlier and more recent controversies over the number of deaths attributable to obesity in the United States (Flegal et al., 2005; Masters et al., 2013b; Mehta and Chang, 2009; Mokdad et al., 2004).

Prior findings from large U.S. cohort studies and other international studies indicate that the relative risks for obesity decline with age (Prospective Studies Collaboration, 2009; Stevens et al., 1998). In fact, it is not uncommon to find that obesity is not associated with excess mortality in older populations (ages 60+) or is protective for mortality (the latter is often termed an 'obesity paradox' and is discussed more fully below) (Prospective Studies Collaboration, 2009). Various explanations have been proposed to explain the declining age pattern. From a biological perspective, body fat is thought to confer physiological advantages during periods of illness (of which older individuals suffer disproportionately from). Relatedly, the biasing effects of reverse causation may increase with age due to a higher prevalence of disease with increasing age.

Masters et al. (2013a,b) argue that the commonly reported declining age pattern is an artifact of cohort differences in mortality risk – related to the fact that the obese in earlier cohorts are likely to have lower durations of obesity compared to the obese in recent cohorts – and age-related survey selection. The latter refers to a specific bias whereby adults with obesity-related morbidities become increasingly less likely with age to participate in health surveys compared to healthier obese adults. I discuss the importance of duration effects in the obesity–mortality association under Obesity Histories.

Alternative explanations for the declining age pattern include competing risks and mortality selection. Competing risks becomes increasingly prominent with age as multiple body systems begin to fail. It is an oft cited explanation for the fact that many other risk factors including smoking also weaken with age. A second reason is mortality selection, a selection distinct from survey selection: among the obese at age x, there is a distribution of unobserved vulnerabilities to succumbing to obesity – the most vulnerable die first, leaving behind a successively less vulnerable group. To date, the role of these alternative explanations has not been investigated and further research is required to understand the relative contributions of all the various explanations put forth to understand why the relative risks for obesity vary by age.

Smoking status is considered an important confounder and modifier of the obesity–mortality association. Smoking reduces weight or otherwise prevents gains in weight in part because nicotine is thought to suppress appetite and increase metabolic demands (Audrain-McGovern and Benowitz, 2011). Smoking is also associated with a high burden of disease.

In the United States, smoking is negatively correlated with obesity (Chiolero et al., 2008). Failure to adequately control for smoking therefore can lead to underestimated relative risks for obesity as smoking operates as a confounder to the obesity–mortality association. In addition to its role as a confounder, smoking behaviors have also been reported to modify the obesity–mortality association. Numerous studies have reported that the relative risks for obesity are lower among current smokers compared to nonsmokers (Adams et al., 2006; Calle et al., 1999; Orpana et al., 2010). The reason for the differential relative risks is not fully explained. Reverse causation bias may play a larger role among smokers because of the high-disease burden in this population. Inadequate control for smoking behaviors is also a plausible explanation.

Often, data on smoking is based on self-reports of smoking behaviors and is amenable to misreporting. In addition to this potential misreporting bias, complete details about smoking intensity and duration are not always available or used in analyses (Hu, 2008). A number of studies report a moderate or otherwise imperfect correlation (0.40–0.70) between cotinine levels, a bioindicator of nicotine presence, and self-reports of smoking (see Stram et al., 2002). These factors would lead to 'residual confounding' of smoking on the obesity–mortality association contributing to the lower relative risks for obesity among smokers compared to nonsmokers. It is important to note that the confounding or modifying role of smoking has been observed to be small or nonexistent in some studies (Durazo-Arvizu and Cooper, 2008; Flegal et al., 2007; McGee, 2005; Mehta and Chang, 2009).

Obesity Paradoxes

The term obesity paradox has been used to describe circumstances in which obesity is found to be protective for mortality. The obesity paradox has been reported in individuals with a wide variety of diseases – kidney disease (Kalantar-Zadeh et al., 2005), heart failure and other heart diseases (Curtis et al., 2005), diabetes (Carnethon et al., 2007), hypertension (Carman et al., 1994), multiple cancers (Gonzalez et al., 2014) – and in older populations (especially among those 70 and over) (Al Snih et al., 2007; Kulminski et al., 2008). Physiological explanations emphasize the protective role of excess fat and are commonly employed in the biomedical literature (e.g., Oreopoulos et al., 2008). Recent evidence, however, suggests that in many cases the paradox is due to a statistical bias. The type of bias is known as 'collider bias' or 'selection bias.' This bias can arise when a sample is 'conditioned' on a common effect of exposure variables as in conditioning on a disease state (Banack and Kaufman, 2014; Glymour and Vittinghoff, 2014; Hernán et al., 2004; Preston and Stokes, 2014). Conditioning a sample on a common effect can change the correlation among exposure variables (Cole et al., 2010). Recently, Preston and Stokes (2014) developed the logic of collider bias in the context of the obesity paradox in a diseased population. They show that the relationship between smoking and obesity (both of which would be 'exposure' variables) is more negatively correlated among those with diabetes ('the common effect') than in the general population (e.g., those with and without diabetes). Using data from the National Health and Nutrition Examination Survey (NHANES), they show that the obesity paradox among those with diabetes is the result of the stronger negative correlation between smoking and obesity in this group compared to the source population that includes both smokers and nonsmokers. Banack and Kaufman (2014) demonstrate that collider bias is also a plausible explanation for the obesity paradox among those with cardiovascular disease (CVD) in the U.S.

The Role of Obesity Histories

Much of the early work on the obesity–mortality association has relied on a single observation of BMI. There has been increasing interest in understanding how weight dynamics over a lifetime influence the risk of mortality (Preston et al., 2013a). Mounting evidence now suggests that there are strong duration effects in the obesity–mortality association. In other words, the longer someone is obese, the higher the mortality risk. For example, Abdullah et al. (2011b) used multiple waves of the longitudinal U.S. Framingham Study and found a positive and dose–response relationship between obesity duration and mortality. In a related analysis, Abdullah et al. (2011a) show a similar relationship of obesity duration and diabetes suggesting that this disease may be an important pathway linking obesity duration and mortality.

Counting the number of years obese is only one method of incorporating weight dynamics in a statistical model. Preston et al. (2013a) provide a detailed typology of the various approaches that have been used when more than one observation of BMI is used in a regression equation. They outlined four specifications that have been most commonly implemented in prior research: additive (i.e., treating each BMI observation additively), duration (i.e., counting the time spent with a high BMI), weight change, and interactive (i.e., including interaction terms among BMI observations). Using baseline and retrospective information on height and weight contained in the NHANES, Preston et al. (2013a) estimated models based on each of the specifications and evaluated the models using two criteria: ease of interpretability and variance of mortality explained. They examined adults ages 50–74 at time of survey and used information on obesity at age 25, age 40, 10 years prior to survey, and at time of survey. Two key conclusions were made. The first is that obesity occurring at various adult ages was equally associated with mortality above age 50. The second is that duration of obesity models performed the best on the two evaluative criteria. The authors additionally found that weight change models, which are frequently used, are not straightforward to interpret. The details of the complexities inherent in interpreting weight change models go beyond the scope of this article and I refer the reader to Preston et al. (2013a) for a fuller discussion.

Mehta et al. (2014) examined the role of obesity histories to mortality at ages 50 and over using the nationally representative Health 2000 survey from Finland. Like NHANES, these data contain retrospectively reported information on weight at earlier adult ages in addition to weight and height collected at time of survey. When Mehta et al. (2014) compared their estimates from Finland with those reported by Preston et al. (Preston et al., 2013a) using the U.S. NHANES, they found that the association between obesity duration and mortality is remarkably similar in both countries (ages 50–74 at time of survey). Mehta et al. (2014) also examined the role of 'peak BMI' as an element of weight histories. Peak BMI refers to the highest BMI attained in life or is otherwise available in data regardless at what age it occurs. In Mehta et al. (2014), peak BMI was identified by the highest BMI recorded at ages 20, 30, 40, 50, and at time of survey. Mehta et al. (2014) found that peak BMI is significantly and positively associated with mortality even in the presence of BMI at time of survey. In fact, the excess risks of obesity using a measure of peak BMI appear double than that of using baseline BMI when each is considered in separate models. The hazard ratios for those

whose peak BMI was $\geq$30.0 kg m^{-2} was 1.42 (95% CI: 1.04, 1.95) relative to those whose peak BMI was less than a BMI of 25.0. The hazard ratio for obesity (BMI $\geq$ 30.0) at time of survey was 1.22 (95% CI: 0.91, 1.65) relative to a BMI < 25.0 at time of survey. The authors also found that peak BMI was moderately and positively correlated (rho = 0.62) with obesity duration.

A key strength of peak BMI is that it may be more robust to reverse causation compared to BMI at time of survey because peak records the highest achieved weight, which is not directly influenced by subsequent weight loss due to disease. Another advantage of peak BMI is that it can be efficiently collected in health surveys, even in instances when more complete histories cannot be collected. Mehta et al. (2014) concluded that peak body weight be regularly collected in health surveys and used to assess the obesity–mortality association.

The sharp period increases in obesity that many nations have experienced suggest that there are important birth cohort differences in obesity histories and obesity duration (among the obese and for the cohort as a whole). Compared to earlier cohorts, the most recent birth cohorts have been experiencing high levels of obesity at young ages. The finding by Preston et al. (2013a) and others that obesity in early adulthood is consequential to later life mortality indicates that obesity has already left an imprint on the future mortality experience of recent birth cohorts. Moreover, the use of peak BMI to estimate the number of deaths attributable to obesity will likely yield a much larger role for obesity than estimates based on BMI at time of survey. Using recent NHANES data, Stokes (2014) found that among those ages 50–84, the number of deaths attributable to a BMI $\geq$ 25.0 was 33% when it was calculated based on peak BMI compared to only 5% when calculated based on BMI at time of survey. Nearly all prior estimates on the number of deaths attributable to obesity have been based on BMI at time of survey.

The recent findings on the role of obesity histories has important relevancies to understanding the extent to which obesity is influencing aggregate levels of mortality in the future. Preston et al. (2013b) project both current obesity and obesity at age of 25 years by birth cohort to examine how rising obesity will affect life expectancy in the United States through 2040. The authors estimate that rising obesity will have a reducing effect of 0.73 years (males) and 0.82 years (females) on life expectancy at age of 40 years through 2040. This reducing effect of obesity represents approximately 29% (males) and 38% (females) of the projected growth in life expectancy during this period as forecasted by the Social Security Administration.

Obesity and Mortality in an International Context

There has been little work examining the role of obesity in contributing to international differences in mortality. Of particular interest has been the role that obesity is playing in contributing to the shortfall in U.S. longevity relative to other high-income countries (National Research Council, 2011). The United States has the highest levels of adult obesity among OECD nations (OECD, 2013). Using relative risks for obesity from NHANES, Preston and Stokes (2011) estimated that obesity is responsible for reducing U.S. life expectancy at age of 50 years by 1.54 years (men) and 1.85 years (women) in 2006. This reduction accounted for 42% (men) and 67% (women) of the U.S. shortfall in longevity at age of 50 years compared to other high-income countries with higher life expectancy at this age. To date, there has been little examination of whether the relative risks for obesity vary across countries. As described above, the findings by Mehta et al. (2014) indicate that the role of obesity duration in shaping mortality risks is remarkably similar in Finland and the United States. This cross-national comparability suggests that obesity is operating similarly on mortality in countries with relatively advanced obesity epidemics. Further cross-national comparisons of the relative risks for obesity will help identify the role obesity is playing in aggregate mortality patterns and in contributing to cross-national differences in longevity.

Another important consideration is the contribution of obesity to socioeconomic differences in mortality within countries. The strength and direction of the relationship between BMI and socioeconomic status varies considerably across countries with both positive and negative gradients observed (McLaren, 2011; Subramanian et al., 2013). In the United States, obesity is inversely associated with education (Sanchez-Vaznaugh et al., 2009) and income (Chang and Lauderdale, 2005) among women. Gradients tend to be flatter among U.S. men. The role that obesity plays in contributing to socioeconomic differences in mortality within countries has been an understudied area. In contrast, cigarette smoking alone has been shown to explain up to 50% of educational differences in male mortality in Europe (Jha et al., 2006). Using data from the Whitehall II study, Stringhini et al. (2010) report that four health behaviors (smoking, alcohol consumption, diet, and physical activity) explain approximately 70% of mortality differences between the highest (administrative) and lowest (clerical or support) occupational grades (ages 35–55) in this cohort. Of this, diet and physical activity jointly accounted for approximately 20% of the difference after accounting for the other behaviors. Using the nationally representative American Changing Lives Study, Lantz et al. (2010) report that physical inactivity and high BMI, along with smoking and alcohol consumption played a small role in explaining income and educational disparities in mortality among U.S. adults (ages 25+). Further work on the specific contribution of obesity to socioeconomic mortality differentials is required.

Secular and Cohort Changes in the Obesity–Mortality Association

Whether the relative risks of obesity have been changing over time also has important implications for the current and future aggregate mortality effects of obesity. To date, most studies examining secular or cohort changes have focused on the United States. Flegal et al. (2005) show that the relative risks for obesity appear to have declined among younger- and middle-aged adults (<70 years) over some successive NHANES surveys. Mehta and Chang (2011) show that there also appears to be a decline since the 1970s and 1980s

among middle-aged samples from NHANES, the U.S. National Health Interview Survey, and the U.S. Framingham Heart Study. One potential explanation or the declining relative risks is that the obese have disproportionately benefitted from improved treatments of cardiovascular risk factors, most notably hypertension and high cholesterol, relative to the nonobese (Gregg et al., 2005). Nonetheless, the decline in relative risks in the United States has not been observed universally. Calle et al., (2005) reported no decline in the relative risks of obesity among adult nonsmokers between 1982 and 1998 in the U.S. Cancer Prevention Study II. Yu (2012), also using NHANES, examined age, period, and cohort variations in the relative risks for obesity and concluded that the relative risks increased from earlier to more recent cohorts with no apparent secular decline among adults born between 1901 and 1957. The studies examining secular or cohort changes in the obesity–mortality association have not directly modeled the role of changes over time (or over cohorts) in obesity duration among the obese. The rapid increase in obesity prevalence in the U.S. and elsewhere may have led initially to a decline in the average duration obese for an obese person. However, such a change has not been confirmed empirically. In addition, many high-income nations are experiencing a 'behavioral replacement' away from smoking and toward higher obesity (Flegal et al., 2010; Giskes et al., 2005; Preston et al., 2013b). It is not fully understood how changes over time in the smoking composition of populations is influencing the relative risks for obesity.

Conclusion

My objective in this article was to highlight both advancements and continued areas of uncertainty in understanding the obesity and mortality association. Despite the large volume of research, many methodological concerns and controversies about its estimation remain. Recent work that has modeled weight dynamics over the life course have undoubtedly advanced our understanding of this complex association. So too has research highlighting the potential statistical biases that can arise in estimating the obesity–mortality association. Nonetheless, there are many avenues for continued research. These include a clearer understanding of the long-term effects of high levels of early life obesity in recent cohorts and the potential beneficial effects of weight loss and weight loss interventions among adults who have already experienced a long duration of obesity. We will also benefit from a clearer understanding of whether the obesity–mortality association is similar in different countries and contexts and whether secular or cohort changes are occurring. Moreover, there has been comparatively less research in developing countries whose populations also continue to experience concomitantly high levels of undernutrition (one exception is Zheng et al. (2011) who provided a comprehensive study of Asian populations).

See also: Adult Mortality in Developing Countries; Adult Mortality in Industrialized Societies; Mortality, Epidemiological, and Health Transitions; Obesity and Labor Market Performance; Obesity: Cultural and Biological Factors; Smoking and Health.

Bibliography

Abdullah, A., Stoelwinder, J., Shortreed, S., Wolfe, R., Stevenson, C., Walls, H., de Courten, M., Peeters, A., 2011a. The duration of obesity and risk of type 2 diabetes. Public Health Nutrition 14, 119–126.

Abdullah, A., Wolfe, R., Stoelwinder, J.U., de Courten, M., Stevenson, C., Walls, H.L., Peeters, A., 2011b. The number of years lived with obesity and the risk of all-cause and cause-specific mortality. International Journal of Epidemiology, 1–12.

Adams, K.F., Schatzkin, A., Harris, T.B., Kipnis, V., Mouw, T., Ballard-Barbash, R., Hollenbeck, A., Leitzmann, M.F., 2006. Overweight, obesity, and mortality in a large prospective cohort of persons 50 to 71 years old. New England Journal of Medicine 355 (8), 763–778.

Adams, T.D., Gress, R.E., Smith, S.C., Halverson, R.C., Simper, S.C., Rosamond, W.D., LaMonte, M.J., Stroup, A.M., Hunt, S.C., 2007. Long-term mortality after gastric bypass surgery. New England Journal of Medicine 357 (8), 753–761.

Al Snih, S., Ottenbacher, K.J., Markides, K.S., Kuo, Y.-F., Eschbach, K., Goodwin, J.S., 2007. The effect of obesity on disability vs mortality in older Americans. Archives of Internal Medicine 167 (8), 774–780.

Alley, D.E., Metter, E.J., Griswold, M.E., Harris, T.B., Simonsick, E.M., Longo, D.L., Ferrucci, L., 2010. Changes in weight at the end of life: characterizing weight loss by time to death in a cohort study of older men. American Journal of Epidemiology 172 (5), 558–565.

Allison, D.B., Faith, M.S., Heo, M., Townsend-Butterworth, D., Williamson, D.F., 1999. Meta-analysis of the effect of excluding early deaths on the estimated relationship between body mass index and mortality. Obesity Research 7 (4), 342–354.

Audrain-McGovern, J., Benowitz, N., 2011. Cigarette smoking, nicotine, and body weight. Clinical Pharmacology and Therapeutics 90 (1), 164.

Banack, H.R., Kaufman, J.S., 2014. The obesity paradox: understanding the effect of obesity on mortality among individuals with cardiovascular disease. Preventive Medicine 62, 96–102.

Berrington, d.G., Hartge, P., Cerhan, J.R., Flint, A.J., Hannan, L., MacInnis, R.J., Moore, S.C., Tobias, G.S., Anton-Culver, H., Freeman, L.B., Beeson, W.L., Clipp, S.L., English, D.R., Folsom, A.R., Freedman, D.M., Giles, G., Hakansson, N., Henderson, K.D., Hoffman-Bolton, J., Hoppin, J.A., Koenig, K.L., Lee, I.M., Linet, M.S., Park, Y., Pocobelli, G., Schatzkin, A., Sesso, H.D., Weiderpass, E., Willcox, B.J., Wolk, A., Zeleniuch-Jacquotte, A., Willett, W.C., Thun, M.J., 2010. Body-mass index and mortality among 1.46 million white adults. New England Journal of Medicine 363 (23), 2211–2219.

Calle, E.E., Teras, L.R., Thun, M.J., 2005. Obesity and mortality. New England Journal of Medicine 353 (20), 2197–2199.

Calle, E.E., Thun, M.J., Petrelli, J.M., Rodriguez, C., Heath Jr, C.W., 1999. Body-mass index and mortality in a prospective cohort of U.S. adults. The New England Journal of Medicine 341 (15), 1097–1105.

Carman, W.J., Barrett-Connor, E., Sowers, M., Khaw, K., 1994. Higher risk of cardiovascular mortality among lean hypertensive individuals in Tecumseh, Michigan. Circulation 89 (2), 703–711.

Carnethon, M.R., Biggs, M.L., Barzilay, J.I., Smith, N.L., Vaccarino, V., Bertoni, A.G., Arnold, A., Siscovick, D., 2007. Longitudinal association between depressive symptoms and incident type 2 diabetes mellitus in older adults: the cardiovascular health study. Archives of Internal Medicine 167 (8), 802–807.

Cawley, J., Meyerhoefer, C., 2012. The medical care costs of obesity: an instrumental variables approach. Journal of Health Economics 31 (1), 219–230.

Chang, V.W., Lauderdale, D.S., 2005. Income disparities in body mass index and obesity in the United States, 1971–2002. Archives of Internal Medicine 165 (18), 2122–2128.

Chiolero, A., Faeh, D., Paccaud, F., Cornuz, J., 2008. Consequences of smoking for body weight, body fat distribution, and insulin resistance. American Journal of Clinical Nutrition 87 (4), 801–809.

Cole, S.R., Platt, R.W., Schisterman, E.F., Chu, H., Westreich, D., Richardson, D., Poole, C., 2010. Illustrating bias due to conditioning on a collider. International Journal of Epidemiology 39 (2), 417–420.

Colman, R.J., Anderson, R.M., Johnson, S.C., Kastman, E.K., Kosmatka, K.J., Beasley, T.M., Allison, D.B., Cruzen, C., Simmons, H.A., Kemnitz, J.W., Weindruch, R., 2009. Caloric restriction delays disease onset and mortality in rhesus monkeys. Science 325 (5937), 201–204.

Curtis, J.P., Selter, J.G., Wang, Y., Rathore, S.S., Jovin, I.S., Jadbabaie, F., Kosiborod, M., Portnay, E.L., Sokol, S.I., Bader, F., 2005. The obesity paradox: body mass index and outcomes in patients with heart failure. Archives of Internal Medicine 165 (1), 55–61.

Durazo-Arvizu, R., Cooper, R., 2008. Issues related to modeling the body mass index–mortality association: the shape of the association and the effects of smoking status. International Journal of Obesity 32, S52–S55.

Flegal, K.M., Carroll, M.D., Ogden, C.L., Curtin, L.R., 2010. Prevalence and trends in obesity among US adults, 1999–2008. Journal of the American Medical Association 303 (3), 235–241.

Flegal, K.M., Graubard, B.I., Williamson, D.F., Cooper, R.S., 2011. Reverse causation and illness-related weight loss in observational studies of body weight and mortality. American Journal of Epidemiology 173 (1), 1–9.

Flegal, K.M., Graubard, B.I., Williamson, D.F., Gail, M.H., 2005. Excess deaths associated with underweight, overweight, and obesity. Journal of the American Medical Association 293 (15), 1861–1867.

Flegal, K.M., Graubard, B.I., Williamson, D.F., Gail, M.H., 2007. Impact of smoking and preexisting illness on estimates of the fractions of deaths associated with underweight, overweight, and obesity in the US population. American Journal of Epidemiology 166 (8), 975–982.

Flegal, K.M., Shepherd, J.A., Looker, A.C., Graubard, B.I., Borrud, L.G., Ogden, C.L., Harris, T.B., Everhart, J.E., Schenker, N., 2009. Comparisons of percentage body fat, body mass index, waist circumference, and waist-stature ratio in adults. The American Journal of Clinical Nutrition 89 (2), 500–508.

Giskes, K., Kunst, A.E., Benach, J., Borrell, C., Costa, G., Dahl, E., Dalstra, J.A.A., Federico, B., Helmert, U., Judge, K., Lahelma, E., Moussa, K., Ostergren, P.O., Platt, S., Prattala, R., Rasmussen, N.K., Mackenbach, J.P., 2005. Trends in smoking behaviour between 1985 and 2000 in nine European countries by education. Journal of Epidemiology and Community Health 59 (5), 395–401.

Glymour, M.M., Vittinghoff, E., 2014. Commentary: selection bias as an explanation for the obesity paradox: just because it's possible doesn't mean it's plausible. Epidemiology 25 (1), 4–6.

Gonzalez, M.C., Pastore, C.A., Orlandi, S.P., Heysmfield, S.B., 2014. Obesity paradox in cancer: new insights provided by body composition. The American Journal of Clinical Nutrition 99, 999–1005.

Gregg, E.W., Cheng, Y.J., Cadwell, B.L., Imperatore, G., Williams, D.E., Flegal, K.M., Narayan, K.M.V., Williamson, D.F., 2005. Secular trends in cardiovascular disease risk factors according to body mass index in US adults. Journal of the American Medical Association 293 (15), 1868–1874.

Hernán, M.A., Hernandez-Diaz, S., Robins, J.M., 2004. A structural approach to selection bias. Epidemiology 15 (5), 615–625.

Hu, F., 2008. Obesity and mortality. In: Hu, F. (Ed.), Obesity Epidemiology. Oxford University Press, New York.

Jha, P., Peto, R., Zatonski, W., Boreham, J., Jarvis, M.J., Lopez, A.D., 2006. Social inequalities in male mortality, and in male mortality from smoking: indirect estimation from national death rates in England and Wales, Poland, and North America. The Lancet 368 (9533), 367–370.

Kalantar-Zadeh, K., Abbott, K.C., Salahudeen, A.K., Kilpatrick, R.D., Horwich, T.B., 2005. Survival advantages of obesity in dialysis patients. American Journal of Clinical Nutrition 81 (3), 543–554.

Kulminski, A.M., Arbeev, K.G., Kulminskaya, I.V., Ukraintseva, S.V., Land, K., Akushevich, I., Yashin, A.I., 2008. Body mass index and nine-year mortality in disabled and nondisabled older U.S. individuals. Journal of the American Geriatrics Society 56 (1), 105–110.

Lantz, P.M., Golberstein, E., House, J.S., Morenoff, J., 2010. Socioeconomic and behavioral risk factors for mortality in a national 19-year prospective study of U.S. adults. Social Science & Medicine 70 (10), 1558–1566.

Masters, R.K., Powers, D.A., Link, B.G., 2013a. Obesity and US mortality risk over the adult life course. American Journal of Epidemiology 177 (5), 431–442.

Masters, R.K., Reither, E.N., Powers, D.A., Yang, Y.C., Burger, A.E., Link, B.G., 2013b. The impact of obesity on US mortality levels: the importance of age and cohort factors in population estimates. American Journal of Public Health 103 (10), 1895–1901.

McGee, D.L., 2005. Body mass index and mortality: a meta-analysis based on person-level data from twenty-six observational studies. Annals of Epidemiology 15 (2), 87–97.

McLaren, L., 2011. Socioeconomic status and obesity. In: Cawley, J. (Ed.), Handbook of the Social Science of Obesity. Oxford University Press, New York.

Mehta, N.K., Chang, V.W., 2009. Mortality attributable to obesity among middle-aged adults in the United States. Demography 46 (4), 851.

Mehta, N.K., Chang, V.W., 2011. Secular declines in the association between obesity and mortality in the United States. Population and Development Review 37 (3), 435–451.

Mehta, N.K., Stenholm, S., Elo, I., Aromaa, A., Heliövaara, M., Koskinen, S., 2014. Weight histories and mortality among Finnish adults: the role of duration and peak BMI. Epidemiology. http://dx.doi.org/10.1097/EDE.0000000000000147.

Mokdad, A.H., Marks, J.S., Stroup, D.F., Gerberding, J.L., 2004. Actual causes of death in the United States, 2000. Journal of the American Medical Association 291 (10), 1238–1245.

National Research Council, 2011. Explaining divergent levels of longevity in high income countries. In: Eileen Crimmins, Samuel Preston, and Barney Cohen for the Panel on Divergent Trends in Longevity. National Academy Press, Washington.

Nordestgaard, B.G., Palmer, T.M., Benn, M., Zacho, J., Tybjærg-Hansen, A., Smith, G.D., Timpson, N.J., 2012. The effect of elevated body mass index on ischemic heart disease risk: causal estimates from a Mendelian randomisation approach. PLoS Medicine 9 (5), e1001212.

OECD, 2013. OECD Health Statistics 2013.

Oreopoulos, A., Padwal, R., Kalantar-Zadeh, K., Fonarow, G.C., Norris, C.M., McAlister, F.A., 2008. Body mass index and mortality in heart failure: a meta-analysis. American Heart Journal 156 (1), 13–22.

Orpana, H.M., Berthelot, J.M., Kaplan, M.S., Feeny, D.H., McFarland, B., Ross, N.A., 2010. BMI and mortality: results from a national longitudinal study of Canadian adults. Obesity 18 (1), 214–218.

Peto, R., Lopez, A.D., Boreham, J., Thun, M., Heath Jr., C., 1992. Mortality from tobacco in developed countries: indirect estimation from national vital statistics. Lancet 339 (8804), 1268–1278.

Preston, S.H., Glei, D.A., Wilmoth, J.R., 2010. A new method for estimating smoking-attributable mortality in high-income countries. International Journal of Epidemiology 39 (2), 430–438.

Preston, S.H., Mehta, N.K., Stokes, A., 2013a. Modeling obesity histories in cohort analyses of health and mortality. Epidemiology 24 (1), 158–166.

Preston, S.H., Stokes, A., Mehta, N.K., Cao, B., 2013b. Projecting the effect of changes in smoking and obesity on future life expectancy in the United States. Demography 51 (1), 1–23.

Preston, S.H., Stokes, A., 2011. Contribution of obesity to international differences in life expectancy. American Journal of Public Health 101 (11), 2137–2143.

Preston, S.H., Stokes, A., 2014. Obesity paradox: conditioning on disease enhances biases in estimating the mortality risks of obesity. Epidemiology 25 (3), 454–461.

Prospective Studies Collaboration, 2009. Body-mass index and cause-specific mortality in 900000 adults: collaborative analyses of 57 prospective studies. Lancet 373 (9669), 1083–1096.

Richmond, R.C., Smith, G.D., Ness, A.R., den Hoed, M., McMahon, G., Timpson, N.J., 2014. Assessing causality in the association between child adiposity and physical activity levels: a mendelian randomization analysis. PLoS Medicine 11 (3), e1001618.

Sanchez-Vaznaugh, E.V., Kawachi, I., Subramanian, S.V., Sanchez, B.N., Acevedo-Garcia, D., 2009. Do socioeconomic gradients in body mass index vary by race/ethnicity, gender, and birthplace? American Journal of Epidemiology 169 (9), 1102–1112.

Sjöström, L., Narbro, K., Sjöström, C.D., Karason, K., Larsson, B., Wedel, H., Lystig, T., Sullivan, M., Bouchard, C., Carlsson, B., 2007. Effects of bariatric surgery on mortality in Swedish obese subjects. New England Journal of Medicine 357 (8), 741–752.

Smith, G.D., Sterne, J.A., Fraser, A., Tynelius, P., Lawlor, D.A., Rasmussen, F., 2009. The association between BMI and mortality using offspring BMI as an indicator of own BMI: large intergenerational mortality study. British Medical Journal 339.

Stevens, J., Cai, J., Pamuk, E.R., Williamson, D.F., Thun, M.J., Wood, J.L., 1998. The effect of age on the association between body-mass index and mortality. New England Journal of Medicine 338 (1), 1.

Stokes, A., 2014. Using maximum weight to redefine body mass index categories in studies of the mortality risks of obesity. Population Health Metrics 12 (1), 6.

Stram, D.O., Huberman, M., Wu, A.H., 2002. Is residual confounding a reasonable explanation for the apparent protective effects of beta-carotene found in epidemiologic studies of lung cancer in smokers? American Journal of Epidemiology 155 (7), 622–628.

Stringhini, S., Sabia, S., Shipley, M., Brunner, E., Nabi, H., Kivimaki, M., Singh-Manoux, A., 2010. Association of socioeconomic position with health behaviors and mortality. Journal of the American Medical Association 303 (12), 1159–1166.

Subramanian, S., Corsi, D.J., Subramanyam, M.A., Davey Smith, G., 2013. Jumping the gun: the problematic discourse on socioeconomic status and cardiovascular health in India. International Journal of Epidemiology 42 (5), 1410–1426.

Vasselli, J.R., Weindruch, R., Heymsfield, S.B., Pi-Sunyer, F.X., Boozer, C.N., Yi, N., Wang, C., Pietrobelli, A., Allison, D.B., 2005. Intentional weight loss reduces mortality rate in a rodent model of dietary obesity. Obesity Research 13 (4), 693–702.

Vazquez, G., Duval, S., Jacobs Jr., D.R., Silventoinen, K., 2007. Comparison of body mass index, waist circumference, and waist/hip ratio in predicting incident diabetes: a meta-analysis. Epidemiologic Reviews 29 (1), 115–128.

Weindruch, R., Sohal, R.S., 1997. Caloric intake and aging. The New England Journal of Medicine 337 (14), 986.

Weindruch, R., Walford, R.L., 1988. Retardation of Aging and Disease by Dietary Restriction. CC Thomas, Springfield.

Yu, Y., 2012. Reexamining the declining effect of age on mortality differentials associated with excess body mass: evidence of cohort distortions in the United States. American Journal of Public Health 102 (5), 915–922.

Zajacova, A., Burgard, S.A., 2012. Shape of the BMI-mortality association by cause of death, using generalized additive models NHIS 1986–2006. Journal of Aging and Health 24 (2), 191–211.

Zheng, W., McLerran, D.F., Rolland, B., Zhang, X., Inoue, M., Matsuo, K., He, J., Gupta, P.C., Ramadas, K., Tsugane, S., Irie, F., Tamakoshi, A., Gao, Y.-T., Wang, R., Shu, X.-O., Tsuji, I., Kuriyama, S., Tanaka, H., Satoh, H., Chen, C.-J., Yuan, J.-M., Yoo, K.-Y., Ahsan, H., Pan, W.-H., Gu, D., Pednekar, M.S., Sauvaget, C., Sasazuki, S., Sairenchi, T., Yang, G., Xiang, Y.-B., Nagai, M., Suzuki, T., Nishino, Y., You, S.-L., Koh, W.-P., Park, S.K., Chen, Y., Shen, C.-Y., Thornquist, M., Feng, Z., Kang, D., Boffetta, P., Potter, J.D., 2011. Association between body-mass index and risk of death in more than 1 million Asians. New England Journal of Medicine 364 (8), 719–729.

Obesity: Cultural and Biological Factors

Alexandra A Brewis, Arizona State University, Tempe, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Humans are evolved to gain weight easily under the conditions of plenty and maintain weight when food supplies are insecure. Globalization of low cost and readily available calories alongside changes in work and built environments that encourage sedentism has led to obesity as a common condition in many parts of the world. Obesity is also increasingly associated with poverty and social and structural disadvantage within countries. Despite obesity being identified as one of the most critical contemporary public health challenges, standard interventions to address it have often failed and both child and adult obesity rates continue to rise around the world. Cultural, biocultural, and ecological theories help explain both the rise of obesity and the resistance to any easy fixes. Social theories provide critical perspectives on the medicalization of obesity and challenge pervasive views of obesity that treat it as a sign of individual moral failing.

Introduction

Obesity technically refers to an excess amount of adipose tissue (body fat). While the way in which 'excess' is defined is often arbitrary (Romero-Corral et al., 2006), common clinical and scientific practice is to label an adult as obese, if their body mass index (BMI: weight/height^2) is 30 or higher. This cut point is considered to roughly match the level of body mass at which health starts to be noticeably compromised (Mascie-Taylor and Goto, 2007), placing people at heightened risk of such diseases as type II diabetes, hypertension, and heart disease. But several studies also show that many 'normal' weight people have metabolic profiles with high triglycerides, blood pressure, or insufficient insulin response that are considered a corollary of being obese, and that it is often fitness level rather than fatness level that counts (McAuley and Blair, 2011; Ruderman et al., 1998).

Using the standard definition of obesity represented by a BMI of 30 or more, currently over one-third of adults are obese in the United States, and proportion is even higher in some low-income nations. The only major world region that is yet to see the emergence of noticeable overweight and obesity in adulthood is north and east sub-Saharan Africa (**Figure 1**). Of the industrialized nations, the Nordic countries and Singapore stand out as having comparatively low rates. While rates in the industrialized North have been growing over the last several decades, adult and child weights in many low-income countries have begun accelerating more recently, and at a much faster pace (e.g., China, see Popkin, 2008). Recent projections suggest that if current trends continue, over half of US adults will be obese by 2030, and that low-income and high-income countries alike will be devastated by attendant disability and lost work productivity within a generation (e.g., Popkin, 2009).

Ecological Models of Obesity

Most fundamentally, the rise of obesity represents a growing and chronic imbalance between calories consumed and expended. With the globalization and commercialization of food systems since World War II, cost-effective foods that are high in fat, salt, and sugars have increasingly dominated diets throughout the world (Popkin, 2009). In many countries, growing household wealth has also allowed for greater food purchasing. With urbanization, changing transportation practices, new technologies, and increasing occupational sedentism, everyday exercise levels have at the same time decreased markedly for many (Dufour and Piperata, 2008).

Specific population case studies have demonstrated how these types of lifestyle changes have played out historically and in local ecologies to lead to rapid escalations in obesity risk. In comparing Samoa and American Samoa, where populations are genetically highly similar, McGarvey and

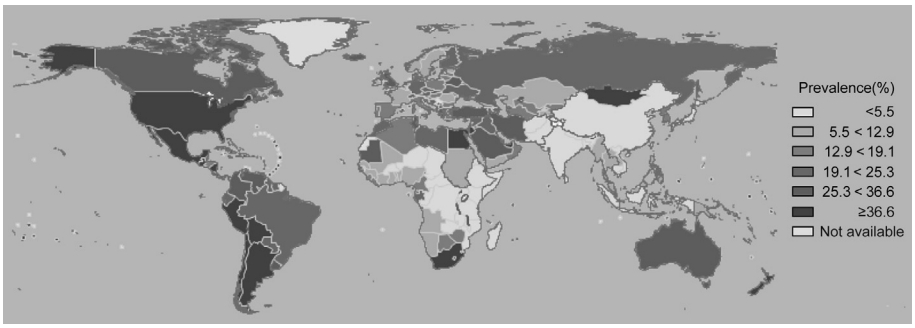

Figure 1 Prevalence (%) of obesity (BMI $\geq$ 30) in women over age 15, 2010 by country. Source: World Health Organization Global InfoBase.

colleagues (McGarvey, 1992; McGarvey et al., 1993, 1989) have shown how temporal changes in occupations (from farming and fishing to office work, for example) and diet (such as a move away from traditional diets) together promote very rapid growth in obesity risk at the population level within one generation. Carolyn Smith-Morris' (2004, 2006) historical study of the Gila River Akimel O'odham (Pima) of central Arizona shows how political, ecological, and cultural factors can force institutional changes (such as emergence of new dietary traditions, or health care practices) that spur rapid – and seemingly intractable – increases in obesity risk.

Many social science approaches thus recognize some social ecological environments as being 'obesogenic' and others as 'leptogenic' (Egger and Swinburn, 1997). This ecological approach to obesity is gaining traction within public health and emphasizes the normality of gaining weight when living in certain environments. It also suggests that neighborhood-level ecological change is the best way to reduce obesity risk in populations. A wide array of studies in geography and related fields have shown the effects of variations in built environment on obesity risk. For example, neighborhood walkability (influenced by such factors as road structure and repair, types of destinations available and proximity, and safety) makes a big difference to how much people living in cities walk (Rundle et al., 2009). Parks provide an important venue for both formal and informal exercise, and park access similarly is unevenly distributed; higher income areas tend to have better parks (more accessible, safer, better facilities) (Cutts et al., 2009; Saelens et al., 2003). People who live near parks or good physical fitness facilities in cities are shown to get significantly more exercise than those who do not (Giles-Corti et al., 2005; Gordon-Larsen et al., 2006) and child obesity rates are suggested to be significantly lower where access to green spaces is better (Sallis and Glanz, 2006). People's perceptions of environments, for example, as safe and pleasant, can be just as important in shaping neighborhood exercise patterns as actual amenities (Humpel et al., 2002; Timperio et al., 2005).

Cityscapes have uneven distributions of access to healthy, fresh, and affordable foods that make it easier or harder to eat an optimal diet (Gallagher, 2006; Horowitz et al., 2004; Morland et al., 2002; Powell et al., 2007; Sloane et al. 2003). Given the way cities develop, favoring the needs of the powerful, areas of low walkability and so-called food deserts are often clustered in lower income and minority inner city neighborhoods. These areas are often correspondingly filled with fast food advertising and restaurants.

The uneven distribution of worse exercise and food amenities in lower income and minority neighborhoods makes obesity risk an 'environmental justice' issue for some (Taylor et al., 2006; Cutts et al., 2009), and provides one explanation why minorities can be at higher risk of obesity even once socioeconomic factors are taken into account (Chang, 2006; Robert and Reither, 2004). Studies of built environment factors in low-income countries reveal patterns that are very different from that typically observed in high-income countries; for example, Mowafi et al.'s (2012) analysis of Cairo shows no significant association of access to green spaces with BMI variation.

Genetic and Epigenetic Explanations of Obesity

To add to the complexity of the historical and environmental picture, body mass and fat distribution patterns show heritability. A recent meta-analysis that pooled genetic data from 249 796 participants identified or confirmed 32 loci in different parts of the genome that helped explained susceptibility to high vs low body mass (Speliotes et al., 2011). Overall, the effect was noticeable but also modest – people with a large number of these variants were 15–20 pounds heavier than those with few of them (Speliotes et al., 2010). Epigenetic models, which focus on how genetic and environmental factors interact across generations to create variation in obesity risk, are increasingly seen as a compelling way to explain historical patterns of rapid escalations of population-level weight gain and obesity in the contexts of these genetic variants. Gluckman and Hanson (2005) coined the term "predictive adaptive responses" to describe important evolved mechanisms that allow humans to adapt to conditions in which food supplies are low or unpredictable. So, when children from low nutrient levels *in utero* suggestive of environmental insults, the body responds to this phenotypically by developing a lower metabolic rate and smaller body build – a so-called thrifty phenotype. This complex of features is advantageous under food-insecure or food-short conditions, but then predisposes via several different pathways to greater weight gain or reduced weight loss when people encounter calorie rich environments during childhood and adulthood (Kuzawa, 1998).

Demography, Socioeconomics, and Obesity

Within countries, states, and cities, there are major demographic variations in obesity, such as by gender, ethnicity, and immigrant status. In most national samples, there is a preponderance of female obesity compared to male (Mascie-Taylor and Goto, 2007) that is at least partially explained by social constructions of gender that shape daily exercise patterns (e.g., the roles of farmer vs housewife) (Brewis, 2011). Minority status, is often discussed in the United States as a major determinant of risk. For example, African-Americans and Mexican Americans are often reported to have the highest rates of obesity (Wang and Beydou, 2007). While relative wealth explains some variation in risk, the patterns are different across these groups; for example, Mexican American women living in poverty are the highest at risk within Mexican Americans, but among African-American men, those who are wealthier are more at risk (Walker and Kawachi, 2011). In response to these diverse findings, many interventions are targeted for subgroups and focus on different ecological and cultural explanatory models of how obesity risk accumulates.

Socioeconomic status (SES) is one of the best predictors of obesity risk within countries. Currently, the higher one's SES in a low-income country the higher BMI is likely to be, and the lower one's SES in a wealthy country then the higher BMI is likely to be (see Brewis, 2011). In 1989, Stunkard and Sobol published an oft-cited summary of 144 studies conducted in

1960s–80s. They described a basic pattern at that time whereby men, women, and children in lower- and middle-income countries were more likely to be overweight if they were socioeconomically advantaged. Newer transnational theories, considering how wealth differentials themselves affect obesity, propose that liberal market economic regimes (like the United States) act to proliferate obesity risk, especially when compared to more regulated welfare economies (like Sweden and Norway) where the wealth gap is narrower (Offer et al., 2010).

A different way to understand how socioeconomic context influences obesity comes from examining social networks – that is, people's everyday webs of social connections. The seminal study in this area by Christakis and Fowler (2007) used longitudinal data from the Framingham Heart Study to show that those whose associates (especially mutually close same-sex friends) gained weight also tended to gain weight themselves. Thus, over time, obesity tends to clump within social networks (as does slimness). Although the clumping itself has been shown in other studies, the underlying mechanisms are not well understood. While Christakis and Fowler posed that shared norms might be responsible, other studies have suggested that this 'social contagion' model is not the best explanation. For example, Cohen-Cole and Fletcher (2008) suggest shared environments are sufficient to explain the convergences in weight over time; Hruschka et al. (2011) show that – even though clustering could be confirmed – shared norms only accounted for 20% of the convergence observed (mainly through influences on ideal body image).

Cultural Meanings of Fat Bodies

Bodies are imbued with cultural meaning in all human societies and what people consider the ideal profoundly reflects and shapes identities and the broader social order (Brown and Konner, 1987; Brown and Sweeney, 2009; Douglas, 1970). A number of ethnographic studies conducted in the 1980s and 1990s have detailed social contexts in which women's fat bodies express beauty, marriageability, attractiveness, and an array of positive moral attributes such as control of selfish desires, closeness to God, generosity, fertility, familial responsibility, and social belonging (e.g., Brown and Konner, 1987; Brown and Sweeney, 2009; Popenoe, 2004; Sobo, 1994). An important example is Popenoe's (2004) study with Azawagh Arabs in Niger, in which she demonstrates how women use large bodies, rolled with fat to advance their social and economic goals, such as through better marriage. An ethnographic survey of several cultures by Brown and Konner (1987) suggested that the majority of human societies traditionally preferred plump women's bodies, in sharp contrast to the prevalent values placed on slimness in the West in the last several decades.

In industrialized contexts, best documented for the United States, women's bodies have long been the dominant, preferred, and inescapable symbol of their self and identity (Becker, 1995: p. 33; Degher and Hughes, 1999; Rubin et al., 1993). The writings of Becker (1995) and Martin (1989) explain how the body has emerged as 'something to be worked on.' Bordo's interpretive analysis discusses the importance of a slim – and for many unobtainable – female body as the cultural expression of an individual's mastery over their desires. Slimness is associated with self-discipline and control, but also health, beauty, intelligence, youth, wealth, attractiveness, grace, and goodness (Caputi, 1983; Moreno and Thelen, 1993). Fatness and obesity are by contrast associated with ugliness, sexlessness, and undesirability, but also with specifically moral failings such as a lack of self-control, social irresponsibility, ineptitude, and laziness (Becker, 1995; Cordell and Ronai, 1999; de Garine and Pollock, 1995; Martin, 1989; Stunkard and Sobal, 1995; Turner, 1984). For women in particular, a 'fat' body can trump *all* other aspects of identify formation (Carryer, 1997; Bordo, 1993), explaining in cultural and social terms why the emotional suffering around the experience of 'being fat' can be so devastating. Sociological studies of 'weight identities' have shown that people recognize and interpret a vast array of complex cues to determine if they are 'fat' (Degher and Hughes, 1999). After significant weight loss, while many describe social 'rebirths,' those of both genders with 'fat' identities can also find themselves struggling with their new 'thin' ones, often presaging weight regain (English, 1993; Granberg, 2006).

Weight Stigma and Discrimination

The social stigma that attends moral judgments about fat bodies explains the high rates of discrimination reported by the largest members of society in the United States (Puhl and Heuer, 2009; Sjoberg et al., 2005). The technically obese (those with BMI over 30) have less career and educational access, lower pay, worse health care service, and are significantly more likely to be fired, bullied, teased, and romantically rejected (Puhl and Heuer, 2009 provide a literature review). Given the predominant cultural model that individual moral failing underlies weight gain, public health and clinical approaches often place the 'blame' for failure on the individual. It seems that even the slim understand the very profound social costs of obesity: based on surveys of trade-offs people are willing to make to avoid obesity, many Americans say they would rather die younger or be blind than be obese (Schwartz et al., 2006).

Weight stigma has the dubious honor of underpinning what some have identified as the last socially acceptable form of discrimination (Puhl and Brownell, 2001). Anticipating or experiencing discrimination, because it challenges core social identities and often takes a chronic form, can prove a very powerful potential source of stress-induced pathophysiology/physiological responses (McEwen, 1998; Gravlee and Dressler, 2005), and this is a potentially important but as yet little studied aspect of obesity etiology (Rosmond and Bjorntop, 2000).

There appears to be a major global trend toward increasingly negative cultural and normative views of large/fat bodies. Using cultural survey data from 11 different locations globally, Brewis et al. (2011) reported that the highest levels of antifat attitudes were measured in middle income nations and territories, such as Paraguay and American Samoa. Anne Becker's work in Fiji shows that exposure to such new media,

such as television quickly changes how people view their large bodies in the context of self and society, moving them toward a slimmer ideal and away from the traditional notion that a sense of self was shaped by community membership rather than individual body habitus (i.e., as the embodiment of culture and ritual daily practices) (Becker, 1995, 2004). Based on recent work in Belize, Anderson-Fye proposes that participation in a transnational economy, drive for success in that new economy, and changing dietary options have introduced eating disturbance and awareness of body size and control among youth (both male and female, although most so for women): the slimmer women have more opportunity afforded to move into the higher paying tourism industry because they adhere to tourist body ideas, rather than local one (Anderson-Fye, 2004; Anderson-Fye and Lin, 2009). Based on a study of urban Paraguayan women, Brewis and Wutich (2012) pose an essentially reverse possible pathway: that changing body norms leads to fat stigmatization that then erodes both material and social capital and promotes weight gain.

Critical Perspectives

Obesity is a good example of a human condition that has become increasingly medicalized over the last several decades; meaning it is defined, discussed, and treated predominantly as medical problem (Sobal, 2011), such as through surgery or pharmacology (Sadler et al., 2009). Medicalization can be helpful in reducing stigma where it deemphasizes the role of individual responsibility, yet it also can promote stigma by labeling all obese bodies as a sick, regardless of the actual health of the individual.

Some social scientists have suggested that obesity in modern contexts is a form of 'moral panic' in which massive global alarm over obesity 'time bombs' and 'epidemics' speaks more to the ways in which obesity is thought to damage society, rather than good scientific evidence. Thus, the obesity epidemic is argued by some to be largely a fabrication that allows powerful groups (such as the commercial diet industry) to promote their own interests by selective use of evidence (Campos et al., 2006). The work of Marilyn Wann and others in the field of fat studies argues prevailing cultural ideas about fat as a form of oppression in itself, especially when the demography of obesity intersects so clearly with race/ethnicity, gender, and poverty (Rothblum et al., 2009). Scholarship in this area reveals alternative cultural models where negative ideas around fat are contested, and positive identities related to 'fat' are being promoted – such as the idea of a fat body as rebellious. This approach as has emphasized the importance of recognizing competing 'frames' – ways of understanding that provide and focus meaning, such as the competing explanatory models used by medically oriented obesity researchers vs fat acceptance activists (Saguy and Riley, 2006).

Obesity as a Complex Problem

After several generations of failed weight loss advice that focused on the importance of individual effort around dieting and exercising as the panacea, increasingly obesity is understood to be much more of a 'wicked problem' – tied in incomplete, complex, and often contradictory ways to changes in local and global ecological, economic, and political systems (not just around food), our urbanizing and increasingly built environments, our technological and social fixes, deeply imbedded in cultural context, and deeply imbedded in our psyches and psychologies, and all in ways that are difficult to recognize and certainly seemingly impossible to fix in any singular or straightforward manner. Systems science provides a means to bring these multiple aspects of obesity into a more comprehensive view that can move across different levels of analysis (Gortmaker et al., 2011). That is, systems models can incorporate the socioecological models already widely applied to include the complexity of myriad proximate and distal factors connected through tipping points and feedback loops and in nonlinear relationships (Finegood, 2011). In this way, the obese body is an emergent property of myriad interplaying factors. A further advantage of a systems approach is that the environmental and contextual causes of obesity tend to become the focus, rather than framing obesity as a product of individual behaviors or more particularly individual moral failings.

Conclusion

While the number of obese people rises globally, the topic of obesity remains ambiguously articulated in scholarly discussions in the social sciences. This includes ongoing and unresolved debates about when and how to define obesity as a 'problem.' Yet, as global public health campaigns gear up to take on the growing 'epidemic' of obesity throughout the world, basic and translatable knowledge about the social bases of obesity is critically needed. Particularly, social science studies to date suggest the better approach to curtailing population obesity is through comprehensive institutional transformations, such as in the built environment and social norms, rather than through individual behavioral or medical intervention.

See also: Anthropology: Overview; Body: Anthropological Aspects; Environmental Justice in the United States; Epigenetic Inheritance; Evolutionary Medicine; Globalization and Health; Health: Anthropological Aspects; Race: Ethnicity and Health; Sociology: Overview; Stigma, Social Psychology of.

Bibliography

Anderson-Fye, E., 2004. A coca-cola shape: cultural change, body image, and eating disorders in San Andrés, Belize. Culture, Medicine, and Psychiatry 28, 561–595.

Anderson-Fye, E., Lin, J., 2009. Belief and behavior aspects of the EAT-26: The case of schoolgirls in Belize. Culture, Medicine, and Psychiatry 33, 623–638.

Becker, A., 1995. Body, Self and Society: The View from Fiji. University of Pennsylvania Press, Philadelphia.

Becker, A., 2004. Television, disordered eating and young women in Fiji: negotiating body image and identity during rapid social change. Culture, Medicine and Psychiatry 28, 533–559.

Bordo, S., 1993. Unbearable Weight: Feminism, Western Culture and the Body. University of California Press, Berkeley, CA.
Brewis, A., Wutich, A., 2012. Implicit versus explicit fat-stigma. American Journal of Human Biology 24, 332–338.
Brewis, A., Wutich, A., Falletta-Cowden, A., Rodriguez-Soto, I., 2011. Body norms and fat stigma in global perspective. Current Anthropology 52, 269–276.
Brewis, A., 2011. Obesity: Cultural and Biocultural Perspectives. Rutgers University Press, New Brunswick, NJ.
Brown, P., Konner, M., 1987. An anthropological perspective of obesity. Annals of the New York Academy of Science 499, 29–46.
Brown, P.J., Sweeney, J., 2009. The anthropology of overweight, obesity, and the body. AnthroNotes 30 (1), 6–12.
Campos, P., Saguy, A., Ernsberger, P., Oliver, E., Gaesser, G., 2006. The epidemiology of overweight and obesity: public health crisis or moral panic? International Journal of Epidemiology 35, 55–60.
Caputi, J., 1983. One size does not fit all: being beautiful, thin, and female in America. In: Nachbar, J., Geist, C. (Eds.), The Popular Culture Reader, third ed. Popular Press, Bowling Green, OH, pp. 186–204.
Carryer, J., 1997. Embodied largeness: a feminist exploration. In: De Ras, M., Grace, V. (Eds.), Bodily Boundaries, Sexualized Genders, Medical Discourses. Dunmore Press, New Zealand.
Chang, V., 2006. Racial residential segregation and weight status among US adults. Social Science and Medicine 63, 1289–1303.
Christakis, N.A., Fowler, J.H., 2007. The spread of obesity in a large social network over 32 years. New England Journal of Medicine 357, 370–379.
Cohen-Cole, Ethan, Fletcher, Jason, 2008. Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic. Journal of Health Economics 27, 1382–1387.
Cordell, G., Ronai, C.R., 1999. Identity management among overweight women: narrative resistance to stigma. In: Sobal, J., Maurer, D. (Eds.), Interpreting Weight: The Social Management of Fatness and Thinness. Aldine de Gruyter, New York, pp. 29–48.
Cutts, B., Darby, K., Boone, C., Brewis, A., 2009. City structure, obesity, and environmental justice: an integrated analysis of physical and social barriers to walkable streets and park access. Social Science and Medicine 69, 1314–1322.
de Garine, I., Pollock, N.J., 1995. Social Aspects of Obesity. Gordon and Breach, Amsterdam.
Degher, D., Hughes, G., 1999. The adoption and management of a "fat" identity. In: Sobal, J., Maurer, D. (Eds.), Interpreting Weight: The Social Management of Fatness and Thinness. Aldine de Gruyter, New York, pp. 11–27.
Douglas, Mary, 1970. Natural Symbols: Explorations in Cosmology. Routledge, New York.
Dufour, Darna, Piperata, Barbara, 2008. Energy expenditure among farmers in developing countries: what do we know? American Journal of Human Biology 20, 249–258.
Egger, G., Swinburn, B., 1997. An 'ecological' approach to the obesity pandemic. British Medical Journal 315, 477–480.
English, Cliff, 1993. Gaining and losing weight: identity transformations. Deviant Behavior 14, 227–241.
Finegood, D.T., 2011. The complex systems science of obesity. In: Cawley, John (Ed.), The Oxford Handbook of the Social Science of Obesity. Oxford University Press, Oxford, pp. 208–236.
Gallagher, M., 2006. Examining the Impact of Food Deserts on Public Health in Chicago. Mari Gallagher Research and Consulting Group, Chicago.
Giles-Corti, B., Broomhall, M.H., Knuiman, M., Collins, C., Douglas, K., Lange, A., Donovan, R., 2005. Increasing walking: how important is distance to, attractiveness, and size of public open space? American Journal of Preventive Medicine 28, 169–176.
Gluckman, P., Hanson, M., 2005. The Fetal Matrix: Evolution, Development and Disease. Cambridge University Press, Cambridge.
Gordon-Larsen, P., Nelson, M., Page, P., Popkin, B., 2006. Inequality in the built environment underlies key health disparities in physical activity and obesity. Pediatrics 117, 417–424.
Gortmaker, S., Swinburn, B., Levy, D., Carter, R., Mabry, P., et al., 2011. Changing the future of obesity: science policy and action. Lancet 378, 838–847.
Granberg, E., 2006. 'Is that all there is?' Possible selves, self-change, and weight loss. Social Psychology Quarterly 69, 109–126.
Gravlee, Clarence C., Dressler, William W., 2005. Skin pigmentation, self-perceived color, and arterial blood pressure in Puerto Rico. American Journal of Human Biology 17, 195–206.
Horowitz, C.R., Colson, K., Hebert, P., Lancaster, K., 2004. Barriers to buying healthy foods for people with diabetes: evidence of environmental disparities. American Journal of Public Health 94, 1549–1554.
Hruschka, D., Hadley, C., Brewis, A., 2014. Disentangling basal and accumulated body mass for cross-population comparisons. American Journal of Physical Anthropology.
Hruschka, D., Brewis, A., Wutich, A., Morin, B., 2011. Shared norms provide limited explanation for the social clustering of obesity. American Journal of Public Health 101, S295–S300.
Humpel, N., Owen, N., Leslie, E., 2002. Environmental factors associated with adults' participation in physical activity: a review. American Journal of Preventive Medicine 22, 188–199.
Karelis, A.D., 2008. Metabolically healthy but obese individuals. Lancet 372, 1281–1283.
Kuzawa, C.W., 1998. Adipose tissue in human infancy and childhood: an evolutionary perspective. Yearbook of Physical Anthropology 41, 177–209.
Martin, E., 1989. The Woman in the Body: A Cultural Analysis of Reproduction. Beacon Press, Boston.
Mascie-Taylor, C.G.N., Goto, R., 2007. Human variation and body mass index: a review of the universality of BMI cut-offs, gender and urban-rural differences, and secular changes. Journal of Physical Anthropology 26, 109–112.
McAuley, P.A., Blair, S.N., 2011. Obesity paradoxes. Journal of Sports Sciences 2, 1–10.
McEwen, B.S., 1998. Protective and damaging effects of stress mediators. New England Journal of Medicine 338, 171–179.
McGarvey, S., Bindon, J., Crews, D., Schendel, D., 1989. Modernization and adiposity: causes and consequences. In: Little, M., Haas, J. (Eds.), Human Population Biology: A Transdisciplinary Science. Oxford University Press, Oxford, pp. 263–279.
McGarvey, S., Levinson, P., Bausserman, L., Galanis, D., Hornick, C., 1993. Population change in adult obesity and blood lipids in American Samoa from 1976-78 to 1990. American Journal of Human Biology 5, 17–30.
McGarvey, S., 1992. Economic modernization and human adaptability perspectives. In: Huss-Ashmore, R., Schall, J., Hediger, M. (Eds.), Health and Lifestyle Change, MASCA, vol. 9. University of Pennsylvania Museum of Archeology and Anthropology, pp. 105–112.
Moreno, A.B., Thelen, M.H., 1993. A preliminary prevention program for eating disorders in a junior high school population. Journal of Youth and Adolescence 22, 109–124.
Morland, K., Diez-Roux, A., Wing, S., 2002. The contextual effect of the local food environment on residents' diets: the atherosclerosis risk in communities study. American Journal of Public Health 92, 1761–1767.
Mowafi, M., Khadr, Z., Bennett, G., Hill, A., Kawachi, I., et al., 2012. Is access to neighborhood green space associated with BMI among Egyptians? A multilevel study of Cairo neighborhoods. Health and Place 18, 385–390.
Offer, A., Pechey, R., Ulijaszek, S., 2010. Obesity under Affluence Varies by Welfare Regimes: The Effect of Fast Food, Insecurity, and Inequality. University of Oxford Discussion Papers in Economic and Social History number 82.
Popenoe, R., 2004. Feeding Desire: Fatness, Beauty, and Sexuality among a Saharan People. Routledge, London, New York.
Popkin, B., 1994. The nutrition transition in low-income countries: an emerging crisis. Nutritional Reviews 52, 285–298.
Popkin, B., 2008. Will China's nutrition transition overwhelm its health care system and slow economic growth? Health Affairs 27, 1064–1076.
Popkin, B., 2009. The World is Fat: The Fads, Trends, Policies, and Products that are Fattening the Human Race. Avery, New York.
Powell, L., Auld, M., Chaloupka, F., O'Malley, P., Johnston, L., 2007. Associations between access to food stores and adolescent body mass index. American Journal of Preventive Medicine 33, S301–S307.
Primeau, V., Coderre, L., Karelis, A.D., Brochu, M., Lavoie, M.E., et al., 2011. Characterizing the profile of obese patients who are metabolically healthy. International Journal of Obesity 3, 971–981.
Puhl, R., Brownell, K.D., 2001. Bias, discrimination, and obesity. Obesity Research 9, 788–805.
Puhl, R.M., Heuer, C.A., 2009. The stigma of obesity: a review and update. Obesity Research 17, 941–964.
Robert, S.A., Reither, E.N., 2004. A multilevel analysis of race, community disadvantage, and body mass index among adults in the US. Social Science and Medicine 59, 2421–2434.
Romero-Corral, A., Montori, V.M., Somers, V.K., Korinek, J., Thomas, R.J., Allison, T.G., Mookadam, F., Lopez-Jimenez, F., 2006. Association of bodyweight with total mortality and with cardiovascular events in coronary artery disease: a systematic review of cohort studies. Lancet 368, 666–678.

Rosmond, R., Bjorntop, P., 2000. Occupational status, cortisol secretory pattern, and visceral obesity in middle-aged men. Obesity Research 8, 445–450.
Rothblum, E., Solovay, S., Wann, M., 2009. The Fat Studies Reader. NYU Press, New York.
Rubin, N., Shmilovitz, C., Weiss, M., 1993. From fat to thin: informal rites affirming identity change. Symbolic Interaction 16, 1–17.
Ruderman, N., Chisholm, D., Pi-Sunyer, X., Schneider, S., 1998. The metabolically obese, normal-weight individual revisited. Diabetes 47, 699–713.
Rundle, A., Neckerman, K., Freeman, L., Lovasi, G., Purciel, M., et al., 2009. Neighborhood food environment and walkability predict obesity in New York City. Environmental Health Perspectives 117, 442–447.
Sadler, J.Z., Jotterand, F., Lee, S.C., Inrig, S., 2009. Can medicalization be good? Situating medicalization within bioethics. Theoretical Medicine and Bioethics 30, 411–425.
Saelens, B.E., Sallis, J.F., Black, J.B., Chen, D., 2003. Neighborhood-based differences in physical activity: an environment scale evaluation. American Journal of Public Health 93, 1552–1558.
Saguy, A., Riley, K., 2006. Weighing both sides: morality, mortality, and framing contests over obesity. Journal of Health Politics 30, 869–921.
Sallis, J.F., Glanz, K., 2006. The role of built environments in physical activity, eating, and obesity in childhood. Future of Children 16, 89–108.
Schwartz, M., Vartanian, L.N.B., Nosek, A., Brownell, K., 2006. The influence of one's own body weight on implicit and explicit anti-fat bias. Obesity 14, 440–447.
Sjoberg, R.L., Nilsson, K.W., Leppert, J., 2005. Obesity, shame, and depression in school-aged children: a population-based study. Pediatrics 116, e389–e392.
Sloane, D.C., Diamant, A.L., Lewis, L.B., Yancey, A.K., Flynn, G., Nascimento, L.M., McCarthy, W.J., Guinyard, J.J., Cousineau, M.R., and the REACH Coalition of the African American Building a Legacy of Health Project, 2003. Improving the nutritional resource environment for healthy living through community-based participatory research. Journal of General Internal Medicine 18, 568–575.
Sobo, E.E., 1994. The sweetness of fat: Health, procreation, and sociability in rural Jamaica. In: Sault, N. (Ed.), Many Mirrors: Body Image and Social Meaning. Rutgers University Press, New Brunswick, NJ, pp. 132–154.
Smith-Morris, C., 2004. Reducing diabetes in Indian country: lessons from the three domains influencing Pima diabetes. Human Organization 63, 34–46.
Smith-Morris, C., 2006. Diabetes among the Pima: Stories of Survival. University of Arizona Press, Tucson.
Sobal, J., 2011. The sociology of obesity. In: Cawley, John (Ed.), The Oxford Handbook of the Social Science of Obesity. Oxford University Press, Oxford, pp. 105–119.
Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., et al., 2010. Association analyses of 249 796 individuals reveal 18 new loci associated with body mass index. Nature Genetics 42, 937–948.
Speliotes, E.K., Yerges-Armstrong, L.M., Wu, J., Hernaez, R., Kim, Li, et al., 2011. Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. PLoS Genetics 7, e1001324.
Stunkard, A.J., Sobal, J., 1995. Psychosocial consequences of obesity. In: Brownell, K.D., Fairburn., C.G. (Eds.), Eating Disorders and Obesity. Guilford Press, New York.
Taylor, W.C., Poston, W.S.C., Jones, L., Kraft, M.K., 2006. Environmental justice: obesity, physical activity, and healthy eating. Journal of Physical Activity and Health 3, S30–S54.
Timperio, A., Salmon, J., Telford, A., Crawford, D., 2005. Perceptions of local neighborhood environments and their relationship to childhood overweight and obesity. International Journal of Obesity 29, 170–175.
Turner, B., 1984. The Body in Society. Basil Blackwell, Oxford.
Wang, Y., Beydou, M., 2007. The obesity epidemic in the United States – gender, age, socioeconomic, racial/ethnic, and geographic characteristics: a systematic review and meta-regression analysis. Epidemiology Review 29, 6–28.
Walker, R., Kawachi, I., 2011. Race, ethnicity, and obesity. In: Cawley, J. (Ed.), The Oxford Handbook of the Social Science of Obesity. Oxford University Press, Oxford.

Relevant Websites

http://www.haescommunity.org/ – Health at Every Size.
http://www.healthyamericans.org/assets/files/TFAH2012FasInFat18.pdf – Trust for America's Health.
http://www.rwjf.org/ – Robert Wood Johnson Foundation.
http://www.yaleruddcenter.org/ – The Rudd Center.

Objectivity, History of

Stephen Gaukroger, University of Sydney, Sydney, NSW, Australia

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 10785–10789, © 2001, Elsevier Ltd.

Abstract

This article discusses what objectivity is, and the challenges that have been leveled against the idea that we can attain objectivity. (1) Possibility of objectivity: objection of ancient Pyrrhonists that everything is subjective, and how it might be met. (2) Relativity of objectivity: Why the fact that everything is relative to judgment or interpretation is not a problem. The question of observers affecting what they observe. (3) Objectivity in the natural and the social sciences: explanation vs interpretation. Could there be different kinds of objectivity, or just different ways in which we might establish objectivity. (4) Cases of conflict between theory and evidence: there is no hard rule on when the theory should be revised to accommodate the new evidence and when it should be abandoned.

Although the term 'objective' only came to have its modern sense in the eighteenth century, issues bound up with objectivity have been rife since seventeenth-century disputes over the respective merits of accounts based on facts and those relying on deductive argument (see Daston, 1992, 1994). Objectivity stands in contrast to subjectivity: an objective account is one which attempts to capture the nature of the object studied in a way that does not depend on any features of the particular subject who studies it. An objective account is, in this sense, impartial, one which could ideally be accepted by any subject, because it does not draw on any assumptions, prejudices, or values of particular subjects. This feature of objective accounts means that disputes can be confined to the object studied. But since there must be some input from the subject in terms of theories, assumptions, etc., this is possible under two conditions: neutral assumptions and shared assumptions. The first is obviously preferable to the second (because the shared assumptions might be mistaken), but there are questions as to whether the first is possible only in a limited range of cases, whether it is possible at all, and whether the second is possible in all, some, or no cases.

Much recent literature in the social studies of scientific knowledge effectively denies that either of these is possible, and argues that if we are to understand how scientific projects come to be accepted, we need to examine not their purported objectivity or truth but the rhetorical strategies employed by their proponents (see the discussions in Fuller, 1993; Golinski, 1998; and the essays in Megill, 1994; *see* Scientific Knowledge, Sociology of). Concern with the nature and possibility of objectivity is not new, however, and by looking at some traditional questions that have been raised about objectivity, we can attempt to identify what is genuinely problematic about objectivity and what is not.

The nature of objectivity can best be brought out by considering challenges to it. There are three traditional kinds of challenge that bring out some of the principal issues: the claim that objectivity is impossible; the claim that objectivity is always relative; and the idea that the form objectivity takes in the natural sciences might not be generalizable.

Is Everything Subjective?

The ancient Pyrrhonists denied that objectivity was possible (see Annas and Barnes, 1985). Everything, they argued, is relative to the observer, to the conditions under which the observation is made, and to the state the object is in when it is perceived. For example, what color we observe something to be will depend upon us – it will look different depending on whether we have jaundice or not – upon the medium in which it is observed – it will depend upon whether we observe the object through air or red-tinted glass – and it will depend upon the state of the object – such as whether the object is in a solid form or a powder (gold is yellow when solid and white when powdered, for example). Observation is always inescapably relative to a host of factors and not only can we not avoid this relativity, but, the Pyrrhonists argued, we cannot single out any factor as being objective. There is no objective fact of the matter as to what color something is, or even whether it is moving or stationary, large or small: it depends on the circumstances of viewing and we have no grounds for preferring observations made by a stationary observer to those of a moving one, those of a close observer to those of a distant one, or even those of an intoxicated observer to those of a sober one.

Now even if we were to accept that there are never any optimal viewing conditions (remembering that these kinds of argument often work by choosing examples where there are no optimal viewing conditions, and then moving illegitimately to the claim that there are *never* optimal viewing conditions), we would not thereby be forced to dispense with objectivity. An objective theory can predict how a powder will look when viewing through different colored glasses, how something will look if we have jaundice, and it may predict color changes when a solid is ground into a powder; and we can certainly correlate changes in perceived speed with our state of motion, even our state of motion relative to one another. In other words, an objective theory is not one that simply chooses one set of viewing conditions over others and then stipulates that these will be the objective ones: it is one that holds for all viewing conditions, and any choice of one set over others does not affect the objectivity of the claims of the theory.

Is Everything Relative?

Any perceptual judgment involves, as well as input from the object perceived, various inputs from our perceptual and cognitive faculties, and we might reasonably expect this latter type of input to take on an increasingly significant role as we move away from straightforward observation. What we see and what we judge is, therefore, dependent on various perceptual-cum-cognitive interpretations that we place on our observations: there is no such thing as a 'view from nowhere' (see Nagel, 1986). This does not rob these observations of objectivity. To think this is to assume that objectivity can be secured only in the absence of interpretation and judgment, but the question of objectivity does not even arise in the absence of these. Objectivity becomes an issue when we raise the question of the standing of our judgments and interpretations. We are not seeking a view from nowhere: we are seeking interpretations and judgments that hold up against all alternatives when judged by the same criteria. The fact that judgments involve conceptual structuring and interpretation (conceptual structuring and interpretation which might have been different) in itself has no bearing at all on the objectivity of those judgments.

The trouble is that when we move beyond this, to the question of theoretically motivated or theoretically guided interpretations of phenomena, differences in these interpretations may result in differences in what counts as evidence, differences in what counts as demonstration, and in what counts as a satisfactory explanation (see Kuhn, 1970; Gaukroger, 1978). Here we do face problems of objectivity. In such cases, questions may well arise of whether, in not treating a particular kind of evidence as revealing, for example, one is being objective. Now failure of objectivity may be very difficult to decide in such cases, but we must not confuse failure of objectivity with differing standards of objectivity. Kuhn's work, in particular, has sometimes been read in such a way that the move from one paradigm to another is akin to a religious conversion rather than something objective (Kuhn himself used this analogy, but arguably as a way of capturing the phenomenology of the experience of moving from one paradigm to another, not as something that was intended to throw light on how changes to the content of science are made). There may indeed be paradigm shifts that reflect no objectivity and are like religious conversions, but the ones that Kuhn described, which are very much the success stories of the physical sciences, are clearly not of this type: the move from Aristotelian to Newtonian to Einsteinian cosmologies is a progression with a clear rationale in terms of objectivity, which a move in the opposite direction, for example, would not be.

Genuine cases of different standards of objectivity are difficult, if not impossible, to find, but those who have tried to establish a relativity of objectivity to the investigators' beliefs or theories or values have sometimes invoked two cases in twentieth-century physics – relativity theory and quantum mechanics – which seem to provide examples of the relativity of observational results with respect to the observer. In fact, they do not. In the relativity case, what Einstein showed in his Special Theory was that we have to alter our assessment of what is relative and what is fixed, but no increase in relativity, as it were, results from this. In classical physics, it had been assumed that how quickly a body is moving will appear differently to different observers: if one is traveling in the same direction as the body, for example, it will appear to be moving more slowly than if one passes it traveling in the opposite direction. In other words, velocity is relative to the observer. But the speed of light in a vacuum turned out to be constant no matter in what direction, or at what speed, the measuring apparatus was moving. Einstein argued that the only explanation for this was that the speed of light was constant, but that space and time contracted or dilated depending on the direction and speed of the observer. Space and time, which were previously absolute, are now made relative to the observer (but not merely relative: they really do – objectively – dilate and contract!), but the speed of light, which was previously relative to the observer, is now made absolute.

The case of quantum mechanics is rather more complex, but again we cannot draw any general conclusions about the relativity of observations from it. The quantum mechanics case arises because we cannot measure both the position and momentum of a particle: measurement of one affects the other. One early interpretation of this phenomenon, the Copenhagen Interpretation, maintained that particles were not in a physical state when not being measured: it was the act of measurement, which involved an interaction between a measuring apparatus and a measured system, that put them in a physical state. Moreover, in the Copenhagen Interpretation, this interaction is very problematic and peculiar because the measuring system obeys classical laws, whereas the measured system obeys quantum laws. The lesson that some have drawn from this is that measurement or observation affects the system observed, so we can never know what the system would be like without the observation or measurement, which is what we seek to know.

The peculiarities of the quantum mechanics case make it difficult to understand even in its own right, but this has not prevented some from trying to take it as a general model. Making a very contentious interpretation of something which is poorly understood into a source of guidance is clearly foolish. Nevertheless, does there not seem to be an analogy between this case and one that may arise in the observation of human behavior, for example, where the fact of being observed might inevitably and perhaps unconsciously lead one to modify one's behavior? The analogy breaks down, however, because the case of modification of behavior is something which we can develop techniques for dealing with, whereas this is explicitly not possible in the Copenhagen Interpretation. Such techniques have been around since antiquity, and were developed in the sixteenth century in the context of legal and historical testimony in a particularly interesting way (see Franklin, 1963). Francesco Patrizzi, in his *Dialogues on History* of 1560, had attempted to show that the historian can either be impartial, or informed, but not both. He begins by rejecting secondary sources as virtual hearsay, and he divides primary sources into the partisan and the objective. Then, replying on a number of Machiavellian assumptions about the nature of rulers, he sets up a dichotomy between the partisan observer and the objective

observer. Partisan observers – in this case, those sympathetic to the ruler – in virtue of being partisan, have access to the relevant information, because the ruler can rely upon them; but because they are partisan they will not provide an objective rendering of this information. Objective observers, namely those who are prepared to be critical of the ruler if this is merited, on the other hand, will not, because they are objective, have the ruler's confidence, and so will not have access to the source of the relevant information. (If this seems farfetched, just think of the widespread practice of government briefings to selected journalists.) Patrizzi's conclusion is that it is "utterly and totally impossible for human actions to be known as they were actually done". The response from his contemporaries was that there are a number of ways in which we can establish credibility and plausibility, along the lines applied in law where, for example, we have doubts about a witness' credibility. In such cases we take into account such factors as the probative value of reinforcing testimonies: the problem is not one of a skeptical challenge to the objectivity of history per se but rather problems about evidence and reliability which practitioners of the discipline will usually be best placed to deal with. The same range of solutions is available to the anthropologist or psychologist, for example, who is concerned that the observation of his or her subjects causes them to modify their behavior or statements in some way.

Does Every Type of Explanation Call for Objectivity?

Objectivity seems to require us to stand back from the phenomena we are examining, as it were, so that we can investigate and assess them from a wholly impartial standpoint, which could be used for any phenomena of this kind. It is not always clear, however, that this approach is either the only one, or the most informative one. Consider the case of an anthropologist studying a rain dance. Assume that rain dances do not actually cause it to rain, and have never done so, yet are always performed at times of drought, and assume that there is evidence that, in times of drought, social strife and uncertainty about the nature of authority increases. In this case, it seems reasonable to answer the question why the natives perform the rain dance in functionalist terms, namely that the rain dance secures social cohesion. But there is also a sense in which this is a wholly inappropriate explanation: the natives perform the rain dance only when they want it to rain, and their reason for performing it is clearly because they believe that this will increase the chances of rain.

There are three kinds of response one can make to this issue. The first is to insist that only the first explanation is objective and impartial. It requires us to stand back from the phenomena and treat them as we would any other: after all, functional explanations can be given for phenomena in our own societies – one can imagine a functionalist explanation of regular presidential or parliamentary elections as securing social cohesion, for example – as well as in others. The theory it works with is something that has been developed in such a way as to eradicate partiality and the imposition of the values of its advocates. Generally speaking, one crucial thing that we seek in scientific explanations is algorithmic compression; that is, particularly economical statements of the principles underlying data that enable us to generate that data and predict outcomes. The functionalist criterion meets this requirement in a way that the account which sticks to what the participants themselves believe does not.

The second response is to argue that the issue hinges on the fact that the functionalist account fails to – indeed cannot but fail to – capture the thought that motivates the participants in the dance. One way in which this contrast is sometimes expressed is in terms of the distinction between reasons and causes: giving the reasons someone has for doing something (or interpreting the behavior) and giving the causes of their behavior are two different things. Some philosophers – most notably Vico and Dilthey – have argued that the natural sciences and the *human sciences* (*see* Human Sciences, History of) are quite distinct (see Berlin, 1976; Dilthey, 1976; Ermarth, 1978) in this respect. In particular, while it might be appropriate to 'stand back' from phenomena in the natural sciences to achieve objectivity, this is inappropriate in the case of the human sciences, where we are dealing not with an objectified realm, but with human beings who have intentional states, emotions, exercise judgment, etc., things they share with the investigator and which the investigator is therefore in a position to interpret and make sense of: something quite different from what we do in the natural sciences. In this view, to take the functionalist road would be to treat the rain dance as if it were on a par with a natural, nonintentional phenomenon and it would thus fail to appreciate the nature of the phenomenon under investigation. Proponents of such a view are not denying that objectivity in the sense of impartiality is appropriate in the human sciences, but rather that criteria of objectivity, or guides as to how objectivity might be achieved, can simply be imported from the natural sciences, because what is at stake is not the discovery of causes but the interpretation of behavior. Similarly with a third response, which denies that there is just one way of explaining the rain dance: different accounts bring to light different aspects of the phenomenon. The success of an explanation on this account is relative to what one wants explained, and there is no reason why the choices should be restricted to two mutually exclusive accounts.

These disputes raise a deep question about just how significantly the means by which objectivity is to be secured can vary. But they do not raise a question about the nature of objectivity as the dispute is over how objectivity in the human sciences is best achieved, not over whether objectivity consists of something different in the natural and human sciences. What it consists of can be spelled out in operational terms: whether one is seeking to discover the causes of behavior or to interpret it, what we seek is an impartial account, one which could ideally be accepted by any subject on the basis of evidence and/or arguments.

The Limits of Objectivity

It is a hallmark of objectivity that one compares one's theory with the evidence, but sometimes it is difficult to distinguish

between the evidence for one's theory and one's expectations about what a plausible theory should or will reveal. One always judges a theory against what one already knows or believes to be the case, and while in some cases the theory may change one's beliefs and expectations, in others it may lead one to challenge the theory by trying to develop a viable alternative, for example. There may simply be no objective matter of fact, in particular cases, about which direction to go in: one direction may be no more objective than another. In other words, there is nothing objective about accepting evidence at face value: sometimes it is the right thing to do, sometimes not. In the face of potentially falsifying evidence, for example, it is often better to revise one's theory, to accommodate the theory to the evidence, than to abandon it. Some philosophers have tried to capture the conditions under which it is no longer productive to continue to revise a theory (e.g., Lakatos, 1972), while others have denied that this is possible (e.g., Feyerabend, 1974).

What these disputes suggest is that guidelines marking out how we are to secure objectivity are much more difficult to formulate than in the heyday of theories of scientific method, when it was assumed that comparison of theories with the evidence was all there was to the matter. But this is, of course, quite different from the claim that the securing of scientific results is not a matter of objectivity.

See also: Content Validity; External Validity; Internal Validity; Laboratory Studies: Historical Perspectives; Measurement Theory: History and Philosophy; Objectivity of Research: Ethical Aspects; Objectivity: Philosophical Aspects; Quantification in the History of the Social Sciences; Reliability: Measurement; Science, History of; Science: Constructivist Perspectives, History of; Scientific Knowledge, Sociology of.

Bibliography

Annas, J., Barnes, J., 1985. The Modes of Scepticism. Cambridge University Press, Cambridge, UK.

Berlin, I., 1976. Vico and Herder. Chatto and Windus, London.

Daston, L., 1992. Objectivity and the escape from perspective. Social Studies of Science 22, 597–618.

Daston, L., 1994. Baconian facts, academic civility, and the prehistory of objectivity. In: Megill, A. (Ed.), Rethinking Objectivity. Duke University Press, Durham, NC.

Dilthey, W., 1976. Selected Writings (H.P. Rickman, Ed. and Trans.). Cambridge University Press, Cambridge, UK.

Ermarth, M., 1978. Wilhelm Dilthey: The Critique of Historical Reason. University of Chicago Press, Chicago.

Feyerabend, P., 1974. Zahar on Einstein. British Journal for the Philosophy of Science 25, 25–28.

Franklin, J.H., 1963. Jean Bodin and the Sixteenth Century Revolution in the Methodology of Law and History. Columbia University Press, New York.

Fuller, S., 1993. Philosophy, Rhetoric, and the End of Knowledge. University of Wisconsin Press, Madison, WI.

Gaukroger, S., 1978. Explanatory Structures. Harvester Press, Brighton, UK.

Golinski, J., 1998. Making Natural Knowledge. Cambridge University Press, Cambridge, UK.

Kuhn, T.S., 1970. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.

Lakatos, I., 1972. Falsification and the methodology of scientific research programmes. In: Lakatos, I., Musgrave, A. (Eds.), Criticism and the Growth of Knowledge. Cambridge University Press, Cambridge, UK.

Megill, A. (Ed.), 1994. Rethinking Objectivity. Duke University Press, Durham, NC.

Nagel, T., 1986. The View from Nowhere. Oxford University Press, New York.

Objectivity of Research: Ethical Aspects

David B Resnik, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA

Published by Elsevier Ltd.

Abstract

This article elucidates the generally accepted idea that scientists should strive to be objective. It provides a definition of objective research, describes the evolving debate concerning objectivity, examines the foundations of objectivity in science, and discusses some of the principles of conduct relating to objectivity, such as honesty and openness. In exploring this subject, this article addresses several topics related to objectivity, such as error and bias, peer review, cooperation and trust in science, conflicts of interest, social responsibility, and the moral limits of objectivity.

Introduction

This article elucidates the generally accepted idea that scientists should strive to be objective. It provides a definition of 'objective research,' describes the evolving debate concerning objectivity, examines the foundations of objectivity in science, and discusses some of the principles of conduct relating to objectivity, such as honesty and openness. In exploring this subject, this article addresses several topics related to objectivity, such as error and bias, peer review, cooperation and trust in science, conflicts of interest, social responsibility, and the moral limits of objectivity.

What Is Objective Research?

The phrase 'objective research' has several different meanings in the social sciences and humanities. We often think of objective research as research that is *unbiased or impartial*. Thus, an objective researcher (or project, report, or study) is like a judge who attempts to give a fair hearing to both sides of a legal dispute. The objective judge listens to both sides of a dispute without giving undue consideration to either. We also think of objective research as *value-free*, i.e., as research that is unaffected by moral, economic, social, political, or religious values. Thus, an objective researcher is like a judge who attempts to make decisions based on legal and empirical evidence without allowing his values to affect his judgment and reasoning. We also think of objective research as *reliable or trustworthy*. Thus, an objective researcher is like a thermometer that reliably reports the temperature. We can trust the thermometer's report. Finally, we also think of objective research as research that is *factual or real*. Objective theories (hypotheses or concepts) are theories that correctly depict facts or real phenomena, and they can be contrasted with fiction, propaganda, speculation, and fraud. All of these conceptions of objective research capture some important aspects of our use of this phrase; there are a variety of ways of understanding objectivity (Harre and Krausz, 1996). For the purposes of this article, objective research can be defined as research that is: (1) unbiased or impartial, (2) value-free, (3) reliable or trustworthy, or (4) factual or real (Resnik, 2007).

The Evolving Debate about Objectivity

The notion that scientists have ethical duties to be objective is hardly a modern invention. Scientists and philosophers, who laid the foundations for the modern scientific method, such as Francis Bacon, Rene Descartes, Galileo Galilei, and William Harvey, all viewed science as objective inquiry. Editors of the first scientific journals instituted peer review in order to publish articles based on objective evidence and reliable methods (LaFollette, 1992). Charles Babbage (1970) was perhaps the first scientist to emphasize the ethical component of scientific objectivity in print. He coined the terms 'fudging,' 'cooking,' and 'trimming' to describe some of the unscrupulous research practices he observed in England in the 1800s. In the last half of the 1900s, powerful discoveries and inventions, such as atomic power, the computer, and genetic engineering, have compelled researchers to recognize the social impacts of science and technology and the relationship between objective inquiry and moral responsibility (Shrader-Frechette, 1994).

In the last two decades, politicians, the press, and the public have raised questions about scientific integrity and ethics as a result of highly publicized cases of misconduct in science and questionable research practices (Broad and Wade, 1993; Shamoo and Resnik, 2009). Some of these episodes include, allegations of fraud, plagiarism, misappropriation of funds, and exploitation of subordinates, as well as controversies concerning secrecy, conflicts of interest, patents, and the funding of science. In response to these concerns, funding agencies, professional associations, and research institutions have adopted codes of ethical conduct in science that address issues relating to objectivity and other aspects of scientific integrity and ethics (Sigma, 1986; American Association for the Advancement of Science, 1980; Panel on Scientific Responsibility and the Conduct of Research, 1992; Committee on the Conduct of Science, 2009; Shamoo and Resnik, 2009).

While this debate about ethical duties concerning objectivity was taking place, many scholars and scientists were questioning the idea that science is or can be objective. This controversy about the objectivity of science is also not new – the ancient Greek philosophers developed accounts of objective knowledge in response to sophistry and relativism – but its twentieth-century incarnation has spurred an extensive inquiry

into the nature of scientific method and practice. Inspired by Thomas Kuhn's monumental work on the history of science, *The Structure of Scientific Revolutions* (1962), scholars from the social sciences and humanities have challenged the traditional view of science as objective knowledge. According to these scholars, otherwise known as 'constructivists,' scientific research is strongly influenced by moral, economic, historical, social, political, or religious interests or values (Shapin and Shaffer, 1985; Latour and Woolgar, 1986). Defenders of the traditional view have responded to these critiques with their own arguments (Kitcher, 1993; Gross and Levitt, 1994; Haack, 2003; Resnik, 2007, 2009). These debates about the objectivity of science – the so-called 'science wars' – have migrated from the halls of academia to the outside world, and they have received considerable attention from the press.

For the purposes of this article, this debate is relevant insofar as it has some bearing on the ethical aspects of objectivity in research. With this in mind, it is important to realize that strong versions of constructivism may be incompatible with the idea that researchers have ethical obligations to be objective, since it makes little sense to claim that a person has an ethical obligation that he cannot fulfill. If it is impossible to at least strive to be objective, then the claim that researchers have ethical duties to be objective is moot (Resnik, 1998). While it is important to acknowledge and understand factors that can undermine or compromise objectivity in research, this article will assume that it is at least possible for researchers to strive to be objective, even if they often fall short of this ideal (Resnik, 2007).

Foundations of Research Ethics

Ethical duties in science stem from two main sources: (1) the scientific profession itself, and (2) society at large (Resnik, 1998). Ethical standards (or norms) promote the aims and goals of science in several ways. Many standards promote the advancement of human knowledge. For example, honesty in science is important in obtaining truth and avoiding error and deception. Other ethical standards play a key role in promoting collaboration and trust among scientists by creating institutional structures that direct individual interests toward common goals (Merton, 1973). For example, authorship practices reward researchers for their efforts and ingenuity, insure that publications meet standards of quality and originality, and hold researchers accountable for their contributions (Shamoo and Resnik, 2009).

Scientists also should adhere to ethical standards in order to promote social values, uphold the public's trust in science, and secure the public's support for research. For example, data fabrication can lead to harmful social consequences, especially if the fabricated results pertain to applied sciences like medicine or engineering. Data fabrication can also be construed as a form of lying, which is widely regarded as immoral. Data fabrication violates the public's trust in science because the public trusts that scientists will not fake data. Finally, fabrication can undermine the public's support for science if people become concerned about unethical conduct among researchers and decide to shift government funds to other programs, such as education or health care.

Ethical Principles and Objectivity

There are a variety of ethical principles (or standards) that promote objectivity in research. The principles discussed below should be viewed as general guidelines instead of absolute rules. They must be elaborated, assessed, and interpreted before they can be applied to particular situations or cases.

Honesty

The principle of honesty implies a general prohibition against falsifying, fabricating, or misrepresenting data, results, or other types of information pertaining to scientific publication. Honesty applies to a variety of other aspects of research, such as grant proposals, peer review, personnel actions, accounting and finance, expert testimony, informed consent, media relations, and public education. As mentioned previously, honesty plays a key role in the search for knowledge and in promoting cooperation and trust among researchers. Few scientists or scholars dispute the importance of honesty and most people understand what it means to fabricate or falsify information pertaining to research. However, a few points of clarification will be helpful.

Error, Bias, and Disagreement

First, it is important to distinguish between dishonest research and error. Dishonesty implies the intent to deceive; error does not. Making an honest mistake in recording or analyzing data should not be equated to faking data. Although errors and dishonesty often have similar results; i.e., the deception of an audience, they arise from different motives. However, scientists also have an obligation to avoid making errors, since errors can also damage the quest for knowledge and undermine cooperation and trust among researchers. Although mistakes are, and always will be, a part of science – scientists are human beings, after all – excessive or egregious errors should be denounced as a form of negligence, since these errors have detrimental effects on research or its practical applications (Shamoo and Resnik, 2009).

Scientists have an obligation to prevent errors by carefully scrutinizing and reviewing their own work and the work of others. Although some errors result from haste or sloppiness, more subtle follies are due to systematic errors (or biases) and self-deception. For example, since the 1800s, racial and ethnic biases have plagued research on human intelligence. More recently, self-deception (or simple carelessness) was probably at work in the cold fusion research conducted by Stanley Pons and Martin Fleischmann, because they did not describe their methods in sufficient details for others to replicate their experiments (Shamoo and Resnik, 2009). Scientists also have a responsibility for reducing the effects of error by publishing errata, corrections, retractions, or apologies. Although the peer review system is not perfect, it can play an important role in reducing errors. Hence, reviewers and editors have ethical responsibilities pertaining to objectivity (LaFollette, 1992).

It is also important to distinguish between dishonesty and disagreement. Researchers often have fundamental disagreements about methods, concepts, and theories. Scientists may

disagree about research results (or their significance) without viewing their opponents as dishonest. Before an act can be judged to be dishonest, researchers must have some agreement on what constitutes the honest application of methods and concepts.

Misrepresentation of Data

The concept of 'misrepresentation,' unlike 'fabrication' and 'falsification,' is neither clear nor uncontroversial. Most scientists will agree that fabrication is making up data and falsification is changing data. But what does it mean to misrepresent data? As a minimal answer to this question, one can define 'misrepresentation of data' as 'communicating honestly reported data in a deceptive manner.' But what is deceptive communication? The use of statistics presents researchers with numerous opportunities to misrepresent data. For example, one might use a statistical technique, such as multiple regression or the analysis of variance, to make one's results appear more significant or convincing than they really are. Or one might eliminate (or trim) outliers when 'cleaning up' raw data. Other ways of misrepresenting data include: drawing unwarranted inference from data; creating deceptive graphs, figures, or images; not publishing all of the data; and using overly suggestive language for rhetorical effect.

However, since researchers often disagree about the proper use of statistical techniques and other means of representing data, the line between misrepresentation of data and 'disagreement about research methods' is often blurry. Since 'misrepresentation' is difficult to define, many organizations have refused to characterize misrepresenting data as a form of scientific misconduct. On the other hand, it is important to call attention to the problem of misrepresenting data, if one is concerned about promoting objectivity in research, since many of science's errors and biases result from the misrepresentation of data (Shamoo and Resnik, 2009).

Conflicts of Interest

In order to promote objectivity in research, it is important for scientists to minimize the effects of financial, personal, professional, and political interests that may potentially undermine their ability to conduct and evaluate research. Conflict of interest has emerged as an important ethical concern in science in this century as many researchers now have significant financial interests in their work (Macrina, 1995; Resnik, 2007). Private support for research and development (R&D) has risen steadily since 1970. Today, industry provides more than 60% of R&D funding in the United States alone, and it is expected that the trend toward the privatization of research will continue. Scientists working in the private sector receive remuneration in the form of salary, stock options, or royalties. Even researchers funded by governments or universities can have significant financial interests in the form of patents or copyrights. Although financial, and other, interests do not invalidate research – good research can still occur in troublesome circumstances – research that is conducted when conflicts exist is more likely to be biased or unreliable. For example, a scientist whose research is sponsored by a pharmaceutical company is more likely to distort results so that they favor the company than a scientist who is not working for the company. Even when research is not biased, the presence of a significant conflict can create the perception of a bias and distrust. For example, if people learn that a food company has sponsored research demonstrating the health benefits of its fat-free potato chip, they may believe that the research is biased, even if it is not. Conflicts of interest can arise in experimental design, publication, peer review, personnel decisions, expert testimony, media interviews, and other aspects of science (Resnik, 1998, 2007).

How should scientists deal with conflicts of interest? First, scientists should disclose conflicts of interest to those parties who may be affected by these conflicts. Most journals and funding organizations now require researchers to disclose sources of funding and conflicts of interest. Sometimes these disclosures are made available to the public when articles are published. Disclosure can help buffer the impact of bias by allowing reviewers, scientists, and the lay public to scrutinize and question research that may be affected by economic, social, or political interests. Second, although it is not possible to avoid all conflicts of interest, researchers have a duty to avoid those conflicts that cannot be corrected by peer review or may have a dramatic impact on science or society. For example, researchers should not review a grant application if the applicant is a close friend or bitter enemy. The decision to avoid a conflict of interest, also known as 'recusal,' can promote objectivity by removing a researcher from a situation that could create a bias or the perception of bias.

Openness

Since openness provides scientists with a reliable defense against error, bias, self-deception, dishonesty, and other phenomena that can threaten objectivity, it is an important ethical principle in research. Openness promotes objectivity in several ways. First, it enables researchers to criticize each other's work. Criticism can occur before, during, or after peer review. Frequently, the most important criticism takes place before an idea appears in print. Second, openness helps researchers to make effective use of human and technological resources. Problems can be solved more effectively when researchers work together than when they work in isolation (Merton, 1973). Third, openness allows the public to learn about research and take effective measures to protect public health and safety, society, and the environment.

Although openness is an important ethical principle in research, it often conflicts with personal or institutional interests. Sometimes researchers refuse to share data in order to protect scientific priority or intellectual property rights. For example, a researcher who is developing a new product may decide not to share preliminary data until he or she has applied for a patent. A researcher who discovers a new phenomenon may not disclose the discovery until the paper has been accepted for publication. Sometimes researchers refuse to share data in order to protect their reputations. For example, a scientist may decide not to share preliminary results from a study until they have had enough time to verify

and analyze those results. Researchers may refuse to share data because they have signed a contract with a corporate sponsor not to disclose data without prior authorization. Also, need to be taken to protect the privacy and confidentiality of human research participants. Since many different parties – government, industry, colleagues, the media – may seek to obtain scientific information, it is likely that researchers will often face difficult choices when deciding how and when to share information.

The Moral Limits of Objectivity

While objectivity is one of science's most important principles, it should be acknowledged that this norm sometimes conflicts with other ethical duties, such as the obligation to protect human subjects and social responsibility (Shrader-Frechette, 1994). For example, anthropologists may refuse to publish or disclose information in order to protect and promote the welfare of the human populations they are studying. In psychology, researchers have a duty to keep information about human subjects confidential in order to protect their privacy, dignity, and autonomy. An ecologist who is interviewed by the media must choose between their desire to advocate for the environment and their scientific duties to remain objective. Although scientists should strive to be objective, there are some situations where scientists should relax their commitment to objectivity in order to promote other moral and social values.

See also: Ethical Dilemmas, Research, and Access to Treatment; Ethical Guidelines and Codes of Conduct in Social and Behavioral Research; Objectivity, History of; Objectivity: Philosophical Aspects; Research Ethics, Cross-Cultural Dimensions of; Research Ethics: Research.

Bibliography

American Association for the Advancement of Science, 1980. Principles of Scientific Freedom and Responsibility. American Association for the Advancement of Science, Washington, DC.
Babbage, C., 1970. Reflections on the Decline of Science in England. Augustus Kelley, New York.
Broad, W., Wade, N., 1993. Betrayers of Truth, second ed. Simon and Schuster, New York.
Committee on the Conduct of Science, 2009. On Being a Scientist, third ed. National Academy Press, Washington, DC.
Gross, P., Levitt, N., 1994. Higher Superstition. Johns Hopkins University Press, Baltimore, MD.
Haack, S., 2003. Defending Science within Reason: Between Scientism and Cynicism. Prometheus Books, New York.
Harre, R., Krausz, M., 1996. Varieties of Relativism. Blackwell, Oxford.
Kitcher, P., 1993. The Advancement of Science. Oxford University Press, Oxford.
Kuhn, T., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago, IL.
LaFollette, M., 1992. Stealing into Print. University of California Press, Berkeley, CA.
Latour, B., Woolgar, S., 1986. Laboratory Life, second ed. Princeton University Press, Princeton, NJ.
Macrina, F., 1995. Scientific Integrity. American Society of Microbiology, Washington, DC.
Merton, R., 1973. The Sociology of Science. University of Chicago Press, Chicago, IL.
Panel on Scientific Responsibility and the Conduct of Research, 1992. Responsible Science, vol. 1. National Academy Press, Washington, DC.
Resnik, D., 1998. The Ethics of Science. Routledge, London.
Resnik, D., 2007. The Price of Truth: How Money Affects the Norms of Science. Oxford University Press, New York.
Resnik, D., 2009. Playing Politics with Science: Scientific Independence and Government Oversight. Oxford University Press, New York.
Shamoo, A., Resnik, D., 2009. Responsible Conduct of Research, second ed. Oxford University Press, New York.
Shapin, S., Shaffer, S., 1985. Leviathan and the Airpump. Princeton University Press, Princeton, NJ.
Shrader-Frechette, K., 1994. The Ethics of Scientific Research. Rowman and Littlefield, Boston, MA.
Sigma, Xi, 1986. Honor in Science. Research Triangle Park, NC.

Objectivity: Philosophical Aspects

Paolo Savoia, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by B. Leiter, volume 16, pp 10793–10797, © 2001, Elsevier Ltd.

Abstract

This article is an overview of philosophical conceptions of objectivity, and it is divided into three parts. First, it describes metaphysical objectivity, which concerns the extent to which the existence and character of some class of entities depends on the states of mind of persons. Second, it describes epistemological objectivity, which concerns the extent to which we are capable of achieving knowledge about those things that are objective. Finally, it shows that metaphysical and epistemological considerations of objectivity are also the focus of sharp critiques by post-positivist and feminist philosophers, who focus on the procedural nature of objectivity, and of an emergent historical–epistemological way of analyzing objectivity that claims to be of primary philosophical relevance.

Traditionally, there are two main kinds of philosophical questions about objectivity, mostly debated in Anglo-American philosophy, which in turn takes up some of the classical features of that notion: metaphysical and epistemological.

Metaphysical objectivity concerns the extent to which the existence and character of some class of entities depends on the states of mind of persons (i.e., their knowledge, judgment, belief, perception, or response). Epistemological objectivity concerns the extent to which we are capable of achieving knowledge about those things that are metaphysically objective. But metaphysical and epistemological considerations of objectivity are also the focus of sharp critiques by post-positivist and feminist philosophers, who focused on the procedural nature of objectivity; more recently, a historical–epistemological way of analyzing objectivity also emerged that claims to be of primary philosophical relevance. Various streams of contemporary thought thus put into question the metaphysical and traditional notion of objectivity following a variety of purposes: some of them want to substitute it with a more humble and 'communitarian' concept of objectivity; some want to destroy it; some want to understand how it historically came into being.

Metaphysical Objectivity

An entity (or a class of entities) is metaphysically objective if its existence and character are independent of the human mind. This 'independence requirement' is central to metaphysical objectivity (Brower, 1993; Sober, 1982), though its proper interpretation raises two important questions: first, in what way must a metaphysically objective thing be 'independent' of the human mind; and second, how much independence of the relevant kind is required?

What Kind of Independence?

The existence and character of some entity might be independent of the human mind in three senses: causally, constitutionally, and cognitively. Only the last two will matter for metaphysical objectivity. An entity is causally independent of the human mind as long as the causal trajectory producing it did not involve the human mind. Shoes, for example, are causally dependent on the human mind because the existence and character of any particular pair of shoes depends causally on a cobbler having had certain beliefs and desires (e.g., a desire to make a particular kind of shoe, and true beliefs about what needed to be done to produce such shoes). By contrast, the existence and character of the earth is causally independent of the human mind: no human intentions played a causal role in bringing about the existence of the earth or its specific character. Metaphysical objectivity, however, does not require causal independence. Even entities that are causally dependent on the human mind can be mind-independent in one of the other two senses (discussed later), and thus still be metaphysically objective. An entity is constitutionally independent of the human mind if its existence and character are not constituted by or identical with the mind. The claim that some entity is metaphysically objective almost always involves denying its constitutional dependence on the mind. The exception is for psychological entities (e.g., beliefs, desires, emotions): such things cannot be constitutionally independent of the mind because they just are facets of the mind. Yet surely psychological facts may also be metaphysically objective. If so, it must be in the final sense of 'independence' from the mind. An entity is cognitively independent of the human mind if its existence and character does not depend on any cognizing state of persons, for example, belief, sensory perception, judgment, response, and so forth. (A 'cognizing' state is one that is receptive to features of the world and thus is a potential source of knowledge of the world.) A metaphysically objective thing is, accordingly, what it is independent of what anyone believes or would be justified in believing about it (or what anyone perceives it to be or would perceive it to be under certain conditions, etc.). On this account, psychological facts about a person are metaphysically objective in virtue of not depending on what an observer of that person believes or would be justified in believing about that person's psychological state. Any kind of metaphysically objective fact (except for psychological facts) must necessarily be constitutionally independent of the mind. All metaphysically objective facts must also be cognitively independent. The common-sense picture of

the natural world presumes that its contents are metaphysically objective in this sense: ordinary people think that atoms and zebras and sulfur are not simply identical with the mind, and that they are what they are independent of what people may believe or be justified in believing about them. Science, then, aspires to epistemological objectivity by trying to accurately depict the way things (objectively) are.

How Much Independence?

There can be degrees of cognitive independence, and thus degrees of objectivity that may be distinguished; not everything that is objective may prove to be objective in the sense in which common sense understands the constituent elements of the natural world to be objective. The crucial notion for cognitive independence is independence from the cognizing states of persons: beliefs, sensory perceptions, judgments, responses, and the like. It is possible, then, to distinguish four kinds of claims about objectivity. According to *subjectivism*, what seems right to the cognizer determines what is right. According to *minimal objectivism*, what seems right to the community of cognizers determines what is right. According to *modest objectivism*, what seems right to cognizers under appropriate or ideal conditions determines what is right. According to *strong objectivism*, what seems right to cognizers never determines what is right. Subjectivism and strong objectivism represent the two classical and opposed philosophical positions of antiquity: Protagoras held that "man is the measure of all things" (subjectivism) (Plato, *Theaetetus* *152a, *166a–*8b), while Plato embraced a kind of strong objectivism (Plato, *Phaedo* *741–*75b, *Republic* *475–*80, *508d-e). The Protagorean position denies the objectivity of the world and everything in it: whatever each individual takes to be the case is the case (for that individual), and thus the existence and character of any particular thing depends (epistemically) on the (individual) human mind. By contrast, the Platonist affirms the complete and absolute objectivity of the world: what really is the case about the world is never fixed by what any person or all persons believe, has (or have) reasons to believe, or could have reasons to believe. Mistake, on a global scale, even under ideal epistemic conditions, is a possibility for the Platonist. Platonists are often called 'idealists,' but they believe in the absolute reality of Ideas, so this latter position can be described as 'realism,' or better, 'metaphysical realism.' Minimal objectivism and modest objectivism occupy conceptual space between these two familiar, historical positions. Minimal objectivism holds that whatever the community of cognizers takes to be the case is the case. This view, like its pure Protagorean cousin, issues in a kind of relativism (what is the case is relative to a particular community of cognizers), but by abstracting away from the subjectivity of the individual cognizer, it introduces a minimum amount of objectivity. It is also a kind of objectivity with some useful domains of application. What is and is not fashionable, for example, is probably minimally objective. What seems right to John about what is fashionable can be objectively wrong: John may be out of sync with the styles of his community, and thus it would be correct to say, "John is mistaken in thinking that a plaid shirt and striped pants go well together." But it does not seem that the entire community can be wrong about what is fashionable: in that sense, what seems right to the community determines what really is fashionable.

In most domains, however, minimal objectivity would be viewed as too close to subjectivism for comfort. Modest objectivity thus abstracts even further away from dependence on actual cognizers, individual or communal. Something is modestly objective if its existence and character depends only on what cognizers would believe under certain idealized conditions for cognition, conditions like full information and evidence, perfect rationality, and the like. (By hypothesis, under ideal conditions, all cognizers would come to the same belief about things.) Everyone on the planet can be wrong in his or her beliefs about a modestly objective entity; beliefs formed under ideal epistemic conditions, however, can never be wrong. This latter point is what differentiates modest from strong objectivity. Some philosophers have defended the idea that the truth about the world is at best modestly objective (e.g., the doctrine Putnam calls 'internal realism' in Putnam, 1981): what is true in any domain is simply whatever inquirers would agree upon under epistemically ideal conditions.

Alternative Metaphysical Conceptions

In the so-called continental tradition, the philosophical movement that had the greatest impact on the contemporary philosophy of objectivity, also by establishing a solid base for various forms of criticism of scientific objectivity, is phenomenology. Phenomenology gave birth to a variety of ways of thinking, but its founder is commonly recognized as Edmund Husserl.

The starting point of phenomenology was the debate on the status of the objects of thought and logic between those who thought them to be purely logic – objective – entities (Frege) and those who believed them to be psychological – subjective – entities (Brentano). Husserl criticized the psychology-oriented position, but maintained that concepts and objects have their origin in the 'intentional' activity of the mind: every object must correspond to the conscience of it. In other words, Husserl's problem was to conciliate the universal and objective nature of logical objects with the subjective experience of the cognizing mind. Husserl's phenomenology of the experience of the mind did not want to reduce objects to psychological states, nor to separate objective contents from the concrete activity of the mind, and he claimed that what had to be investigated was the 'phenomenon' itself, not conceived as opposed to a thing in itself (Kant), but as the way in which reality manifests itself in the mind. In order to do so, Husserl proposed a method he called *epoché*, consisting in freeing oneself from all the psychological and logical assumptions and describing the pure experience of objective reality as it appeared in the mind, thus embracing a strong idea of objectivity as a description of the original and essential correlation between the mind and the external world. In the view of its founder, phenomenology had to provide the foundation of all scientific thought (Husserl, 1970).

Husserl's philosophy later took the path of an extreme transcendentalism, embracing the absolute reality of the mind as a sort of renewed Cartesian *cogito*, while his early works

became the basis for the development of Heidegger's ontology. But by the 1930s almost all the members of the phenomenological school shared the view that objective knowledge, in the form of 'science' – often called 'objectivism' – became a mere technical activity more and more separated from the essence of philosophy and of authentic knowledge. Influenced by historicist thinkers such as Willhelm Dilthey, and by Heidegger's early existential analysis, Husserl and the phenomenologists elaborated the concept of the 'lifeworld,' a world of intersubjective concrete experience of reality that is the background, and the source, of all human activity (science and objective knowledge included). This preconceptual historical and social world of original and shared intersubjective experiences and activities is the basis of scientific objectivity, and as such it allowed phenomenologists to criticize science for detaching itself from the lifeworld and becoming a hollow and dangerous technological dominion over the natural and human world. Before becoming the explicit focus of natural science, the world is this set of shared horizons of meanings and experiences; objectivity was possible only insofar it remained connected to the intersubjective exchanges it came from (Husserl, 1970a). This does not mean that phenomenological thought has been antiscientific, but rather that it opened the way to procedural conceptions of scientific knowledge and objectivity, which could also be thought of in social, historical, and political terms.

Epistemological Objectivity

The demand for epistemological objectivity is the demand to be free of bias and other subjective factors that distort cognition, that prevent the things being cognized from presenting themselves as they really (metaphysically) are. More precisely, epistemological objectivity requires that the cognitive processes and mechanisms by which beliefs about the world are formed be constituted in such a way that they at least tend toward the production of accurate representations of how things are. Notice that epistemological objectivity does not require that cognitive processes always yield true representations: that would demand more than is attainable and more than is even expected. Epistemological objectivity obtains when either of the following is true: (1) the cognitive processes at issue reliably arrive at accurate representations, or (2) the cognitive processes are free of factors that are known to produce inaccurate representations. The obstacles to epistemological objectivity will vary with the domain under consideration. In the sciences generally, and the social sciences in particular, 'values' are often thought to present a special obstacle to epistemological objectivity insofar as they influence the choice of research topics and, most seriously, the selection and evaluation of evidence (Weber, 1949). As one author explains: "Values are thought to damage inquiry epistemically because they are held to be subjective – they come from us, not the world. Therefore, to allow values to influence inquiry into the nature of the world is to allow such inquiry to be subject to a control other than the world itself" (Railton, 1985: p. 818).

Epistemic values or norms – for example, norms about when evidence warrants belief – must of course play a role in all scientific inquiry; the worry is about nonepistemic values or norms, like the political ideology of the inquirer or the political climate in which inquiry takes place. Yet even here the effect on objectivity depends on the context and the precise role values play in inquiry. For example, where the political climate is hostile to the discovery of certain kinds of objective truths, only those inquirers motivated by a contrarian political ideology themselves are likely to pursue those objective truths (Miller, 1991). Here, of course, values supply a motive for allowing 'the nature of the world' to control the course of inquiry, rather than interceding and distorting cognition.

More generally, one "need not require freedom from all value and bias in order to have objective inquiry" since "there may yet exist mechanisms of belief formation that incorporate feedback from the object to the inquiring subject" (Railton, 1985: p. 818). The mechanism at issue is causal: metaphysically objective things make themselves felt causally, whatever our theoretical preconceptions and values. This, however, is merely an external criterion for objectivity, and does not yet show how inquirers could determine whether or not their inquiry is epistemologically objective.

Here, however, inquirers might look for certain familiar markers of epistemological objectivity, like the existence of intersubjective agreement in judgment, the publicity of evidence and standards of proof, and the reproducibility of the evidence for a judgment by different inquirers: "when these conditions are met, subjective contributions and biases will be excluded as far as possible" (Railton, 1985: p. 818). That physics constitutes a cross-cultural, global community of inquirers strongly suggests that it is epistemologically objective: if it were not, then one would expect local differences (in interests, ideology, and the like) to lead to markedly different discourses of physics.

Epistemological Critiques

The epistemology of objectivity has proved to be a fruitful field for critiques and different approaches aiming at reforming, correcting, or destroying the metaphysical notion of objectivity. In the first place, metaphysical thinking seems to imply as its epistemological correlative, a 'view from nowhere' (Nagel, 1997). Secondly, objectivity has been conceived as a point of attraction of a series of values and attitudes like detachment, impartiality, neutrality, lack of biases, and so forth, which should guide science toward truth via objectivity. Third, objectivity has been conceived in this way as coinciding with an ahistorical core of truth, rationality, and reality. Post-positivistic philosophy put into question all these assumptions: the idea of the dual structure of knowledge (scheme and reality, or language and empirical observation); the ahistorical nature of objective knowledge (knowledge is a historical process and as such it does not follow a linear line of progress); and the supposed moral neutrality of objective knowledge (objective knowledge is an intersubjective process, constitutively biased, dependent upon social and cultural assumptions).

Already in the 1930s the French philosopher and historian of science Gaston Bachelard put into question these three assumptions. In a series of very influential texts, he wrote a 'psychoanalysis of objective knowledge' that conceived objectivity in historical, dynamic, and procedural terms. The

word 'psychoanalysis' should not mislead us here, because it did not refer to the empirical process of the knowing individual mind, but rather to a collective psychology of knowledge itself. According to Bachelard, the epistemological problem of objective knowledge had to be raised in terms of obstacles. These 'epistemological obstacles' are not external obstacles that impose themselves upon knowledge (such as human fallibility) but are internal to the act of knowing itself, as its necessary inertia. The obstacles against objective knowledge are not simply errors that a progressively more perfect and objective knowledge rectifies, but are one of the conditions of the possibility of objective knowledge itself: indeed, there is no possible knowledge that is not in a dialectic relation with a previous obstacle. Influenced by phenomenology, Bachelard also believed that such a philosophy was insufficient because it was too abstract and it ended up underestimating the material and procedural aspects of objectivity, and thus he spoke of a 'phenomenotechnics' of objective knowledge. There are four main epistemological obstacles that scientific knowledge has to overcome in order to be objective. The first one is common opinion or common sense, with which objective knowledge has to break. The second is the notion of an original and first experience, which culminates in the notion, or better the myth, of the 'given' (equally put into question by postanalytic philosopher Wilfred Sellars) as the natural data upon which objectivity would be built; on the contrary, objective experimental science constructs its results upon nature and sometimes against it. The third major obstacle is the 'realist obstacle,' exemplified by the concept of substance, which is the naive way of conceiving the reality of the world. Bachelard did not advocate relativism, but tried to highlight the fact that reality was a process and that objectivity was a result of a complex interaction between subjects and objects. Finally, there was the 'animist obstacle,' which consisted in attributing magic and self-explanatory power to the concept of life, the valorization of the notion of life and 'the vital' in the pursuit of objective knowledge. Objective knowledge was in this picture the result of a historical process of intellectual catharsis and self-discipline (Bachelard, 2002). The so-called French tradition of epistemological history picked up this idea of procedural objectivity and, purified of its more psychological features, applied it mostly to the analysis of the life sciences and the human sciences.

In the Anglo-American context, the philosophy of Popper has been a transitional step toward more radical epistemological critiques. Popper believes that the problem of objectivity has to be approached from the point of view of that of realism, the simple idea that the world exists independently of the human mind and action. As such, science and knowledge presuppose realism. Popper's theory of the three worlds starts from this assumption. According to it there exist: (1) the world of physical objects and physical states; (2) the world of conscious or mind states; and (3) the world of 'objective contents of thought,' especially scientific and poetic thought. World (1) is the world of everyday objects, like tables and trees; it is objective because other people can experience it, and it is autonomous because their existence is not dependent upon us. World (2) is that of all our psychological experiences, like emotions, memories, and so forth; it is subjective because every such state is different from that of other people, and it is not autonomous, because it depends upon the mind of the humans. World (3) is the world of scientific and more generally intellectual human activity. Other kinds of objects do not fall within these two worlds: words, propositions, laws, symphonies, theories, numbers, and so forth. These are not material objects, but they are nonetheless objective; even if they are produced by the human mind, once created, they produce effects that were not predictable, nor included in the subjective intentions of their producers. Language plays a decisive role in this process, because it makes public, communicable (intersubjective) an object of the private world (2) (Popper, 1972: pp. 153–165).

From this theory derives that there are two kinds of knowledge: subjective and objective. Subjective knowledge is a state of mind or a disposition to behave in certain ways; objective knowledge consists in problem, theories, and argumentations as such. Traditional epistemology is mistaken because it takes scientific knowledge to be an object of world (2), thus focusing on the problem of justification, of how a private state of mind can be considered objective knowledge. Popperian epistemology wants, on the contrary, focus on problems and arguments as such, as a process of conjectures and confutations. Objective knowledge is an ever-evolving natural product of humans, and it grows through this continuous interactions of the world (3) with our subjective world (2) (Popper, 1972: pp. 106–115).

A central feature of this process is a sort of continuous public rational criticism that scientists apply to their production of objective knowledge. Popper rendered more dynamic the traditional problems of metaphysics and epistemology of objectivity, but this conception of rational criticism has been criticized by postanalytic philosophers and historians of science. Starting from the 1950s, the temporal dimension entered the realm of philosophical investigations of objectivity. Kuhn saw it as belonging only to the rare periods of revolution in the science, but inapplicable to the procedures of 'normal science,' while scientists are engaged in conforming previously shared assumptions (Kuhn, 1962). Feyerabend (1975) went beyond Kuhn in fighting every notion of scientific method that would guarantee a more objective knowledge. Imre Lakatos developed a more rationalistic and less relativistic notion of objectivity, embodied in his 'research programs' theory, which wanted to harmonize Popper's fallibilism and the critiques by Kuhn and Feyerabend. According to Lakatos, the development of the history of science is always characterized by the competition of two or more alternative research programs, built around a metaphysical core assumed as preliminary valid and not subject to criticism. Objective knowledge is not a simple process involving nature on the one hand, and scientists on the other, but a more complex one involving the historically changing research programs (Lakatos, 1978).

Some of these issues have been picked up by more recent thinkers (Dworkin, 1996; Nagel, 1997). In particular, these philosophers have raised two kinds of doubts about the preceding kinds of characterizations of objectivity. First, these philosophers question whether the conception of metaphysical objectivity (especially strong metaphysical objectivity) does not presuppose a vantage point on the way things 'really' are to which we can have no access. Second, these philosophers

wonder whether the conception of metaphysical objectivity as mind-independent is not a paradigm too closely tied to a picture of the objectivity of the natural world, and thus either does not make sense or should not be applied with respect to the objectivity of domains like ethics or aesthetics. In these evaluative domains, it makes no sense to ask about whether there are evaluative facts 'out there in the world.' The objectivity of evaluative discourse is simply a matter of its susceptibility to reasons, of our ability to subject ethical positions to rational scrutiny and discussion.

Picking up and putting together the various threads of postanalytic philosophy, someone like Richard Rorty has been able to consider objectivity as a useless, or worse a dangerous, concept, and proposed to substitute alternative notions that include social, or intersubjective, practices. The traditional ideal of objectivity as the true knowledge of what is real in the external world is put into question by denying that a distinction between human interests and practices and knowledge takes place. 'Objectivity' can only be a settlement on the best beliefs and reasons that human beings can give to each other and share, and thus it can only be based on the consensus of the scientific community. Thus Rorty conceives the sciences as aiming toward solidarity rather than metaphysical and epistemological objectivity: scientific knowledge, with its remarkable capacity of prediction and control, is to be praised because it fulfills in the best possible way human interests, not because it attains an objective representation of the mind-independent 'external world.' In other words, objectivity is not transcendent with respect to the human world, but it is given by a pragmatic research for shared and applicable knowledge. Objectivity does not need to be grounded in a transcendent conception of the world and in an epistemology of representation, because objective knowledge is not an enterprise different from other human activities, and as such it is a matter of beliefs, desires, debating, and of what works better. Significantly, Rorty praises the scientific enterprise in terms of the moral virtues of the scientists, aiming at an 'unforced agreement' based on persuasion rather than coercion, on respect for the opinions of others, and on curiosity. By putting into question the fundamental separation between subjective interests and objective representations, some of the traditional epistemological features of objectivity – a notion he ultimately wanted to destroy – became in Rorty's terms a moral and political model of virtues for the human community (Rorty, 1991: pp. 21–34).

Another deeply influential set of epistemological criticisms toward the traditional way of conceiving objectivity comes from feminist epistemology and philosophy of science. Feminist epistemologists deal with how gender influences the processes of the production and justification of knowledge and science, and how the knowing subject, individual and collective, is shaped by gender. Feminist critique of objectivity focuses in particular on three points: (1) the idea of objectivity as absence of perspective, (2) the idea of objectivity as tied to value neutrality, and (3) the idea of metaphysical objectivity as the picture of what is really real on the world independently from the human (gendered) mind. Feminist conceptions of objectivity are thus explicitly pluralistic.

Donna Haraway put into question the very fact that the scientific goal must be a single and unified vision of the world, which has historically been the vision of the dominant groups in a given society. Thus she proposed to substitute 'the god trick of seeing everything from nowhere' with what has been called 'situated knowledge.' In this view, objectivity remains a valuable goal, but it is conceived in different terms: 'objectivity turns out to be about particular and specific embodiment and definitely not about the false vision promising transcendence of all limits;' it follows that only partial – gendered, embodied, socially determined – perspectives promise objective vision. Since the knowing self is limited in every possible way, objectivity cannot be anything else than a communal experience where subjects put together their perspectives without claiming to abolish them (Haraway, 1988).

The emphasis on situated knowledge gave way to a general theory of knowledge called 'standpoint knowledge,' which argues that what happens in social relations both enables and limits, without rigidly determining, what we can know. Sandra Harding in particular has emphasized that a better vision of objectivity must not limit itself to criticize the naive individualism of traditional views, according to which objective knowledge can be attained by overcoming individual biases, but has to put into question the widely cultural assumptions held by scientific communities. Unlike 'weak objectivity,' relying only on an impossible neutrality, Harding proposed a 'strong objectivity' program focused on maximizing objectivity by detecting the values and interests that constitute the scientific projects, and by distinguishing those interests that enlarge from those that limit objectivity. Given the fact that the history of scientific practices and theories is dominated by androcentric values that benefited men, women are most likely to be critical of such values, because they have less to lose. The general conclusion is valid also for Eurocentric and bourgeois values of scientific communities: only by adopting a standpoint different from that of the dominant social groups can objectivity be maximized and more accurate empirical and theoretical accounts of the world provided. And this can in turn be obtained through 'reflexivity,' namely by making explicit all the partial standpoints of knowers and of the community of knowers, thus enhancing objectivity by avoiding a confusion of one's own partial perspective with a comprehensive view, and by highlighting contingencies of representation (Harding, 1995).

In a different but related vein, Helen Longino proposed a notion of objectivity as the result of the democratic life and discussion of a community of knowers, which has to have recognized places of criticism and shared standards, and must be responsive to criticism, with an equally distributed intellectual authority (Longino, 1990).

The Historical Epistemology of Objectivity

'Objectivity' tends to be conceptualized in terms of its opposite, 'subjectivity.' Subjectivity connotes variability and contingency, the perspective of an individual human subject, prey to local circumstances and molded by them into a distorting mirror. Objectivity is, by contrast, the 'view from nowhere,' with no local circumstances – and no perspective – to distort the mirror (Dear, 1992: pp. 619–620). In contemporary philosophy, objectivity has a 'hybrid' character. Lloyd (1995: p. 353) noted that 'objective' can mean detached, disinterested, unbiased,

impersonal, not having a point of view; it also means public, publicly available, observable, or accessible (at least in principle); it means existing independently or separately from us: that which is really existing, Really Real, the way things really are. We can say that the first two meanings are epistemological and the third metaphysical. Historians had already emphasized the fact that objectivity is not a single idea, but rather "a sprawling collection of assumptions, attitudes, aspirations and antipathies" (Novick, 1988: p. 1; Shapin and Schaffer, 1985). In recent years, another approach to objectivity has emerged, which can be placed under the label of historical epistemology, the historical study of the scientific categories, or meta-concepts, that structure our present thoughts and organize our practices, a study that regards the history of concepts, practices, instruments, and 'moral economies' of sciences. The historical epistemology of objectivity is aware and takes into full account the criticism coming from feminist and postanalytic philosophy, and it starts from the fact that "before we can decide whether objectivity exists, and whether it is a good or bad thing, we must first know what objectivity *is* – how it functions in the practices of science" (Daston and Galison, 2007: p. 51). According to this approach, (1) objectivity has to be treated historically; (2) the history of objectivity is the history of the merging of discourses, practices, epistemologies, and moral values; (3) the history of objectivity indicates the existence of a plurality of different epistemic virtues; and (4) contrary to a purely philosophical approach, objectivity must be considered as something different from semantic, epistemological, or metaphysical conceptions of truth.

Among historians of science, Lorraine Daston and Peter Galison gave us the most systematic account of the history of the concept of objectivity. Their core thesis is that the epistemic virtue called objectivity became the dominant scientific norm during the second half of the nineteenth century, through the accumulation and crystallization of a series of heterogeneous elements (such as techniques, instruments, scientific institutions, the birth itself of an international scientific community), and in correspondence with a general change of meaning of the meta-concepts – that is, concepts that organize other concepts – of subjectivity and objectivity. From Kant on, they argue, these two terms underwent a sort of 180-degree inversion and quite quickly took the meaning that we are familiar with: roughly speaking, objectivity designates the matter outside the mind, and subjectivity the mental and inner sphere of thought. For example, in the late Middle Ages, thinkers such as Duns Scotus and William of Ockham (fourteenth century) believed that the meaning of the couple of adjectives *objectivus/subjectivus* was, respectively: things that are present in the mind/things as they are in themselves; traces of this meaning can be found even in Descartes. Since about the second half of nineteenth century, all the principal European dictionaries began to register a meaning very close to ours. The only thing that remains the same in the history of these concepts is the fact that they always are in a relationship of mutual opposition and mutual definition: objective is the nonsubjective, and vice versa.

The history of the epistemic virtues related to objectivity is therefore linked to changing conceptions of subjectivity, which can be distinguished roughly as follows: the rational soul of the seventeenth century; the self as an aggregate of faculties of the eighteenth century; the post-Kantian self as a unity centered around the will of the nineteenth century; and finally, the vertical self, deep and unconscious, of the twentieth century (Daston and Galison, 2007: pp. 29–31). If one wishes to understand the emergence and the importance of objectivity as sets of epistemic virtues and as a scientific meta-concept, then one has to oppose it to the parallel notions of subjectivity. Daston and Galison start from two presuppositions: the self is historically determined; and it is shaped by a series of practices and exercises. In order to answer the question about the emergence of the meta-concept of objectivity one needs to look at the meta-concept of subjectivity, but starting 'from below:' who is the scientific self? The scientific self is anyone who becomes so as a result of the application of particular technologies of the self, deemed necessary to be a scientist. An epistemic virtue is then a melting between ethics and epistemology realized through instruments and technologies of the scientific self, regulated by an ethico-scientific ideal (Daston and Galison, 2007: pp. 198–199).

Daston and Galison take up from Foucault the concept of techniques of the self and they apply it to the formation of the scientific self with special reference to the second half of the nineteenth century and to what they call 'mechanical objectivity.' The techniques of the self corresponding to mechanical objectivity are practices such as the training of the senses in scientific observation, the keeping of laboratory notes, the monitoring of one's hypotheses and opinions, the canalization of one's attention, the control of one's beliefs and fantasies, and the suppression of the spontaneous forms of the will; the scientist applies these techniques to the scientific self and thus obtains, at the very same time, objective results, objective knowledge. In order to capture the peculiarity of nineteenth-century mechanical objectivity, Daston and Galison make a historical comparison with other epistemic virtues, insisting on the fact that every emergence of a new epistemic virtue is not a mere substitution of the existing ones, but rather a reorganization of the field of possible scientific practices and ideals. Among them, they call 'truth-to-nature' the epistemic virtue dominant in the eighteenth century, characterized by the scientist's ability of selecting and synthesizing, grasping and stylizing the essence of natural phenomena; and 'trained judgment,' which emerges in the first years of the twentieth century, a virtue marked by the scientist's use of unconscious intuitions and interpretative techniques about natural phenomena (Daston and Galison, 2007: pp. 42–50).

See also: Continental Philosophy of Language; Historical Explanation, Theories of: Philosophical Aspects; Phenomenology: Philosophical Aspects; Relativism: Philosophical Aspects; Science, History of; Self: Philosophical Aspects.

Bibliography

Bachelard, Gaston, 2002 [1938]. Formation of the Scientific Mind. Clinamen Press, Manchester.

Brower, Bruce W., 1993. Dispositional ethical realism. Ethics 103, 221–249.

Daston, Lorraine, Galison, Peter, 2007. Objectivity. Zone Books, New York.

Dear, Peter, 1992. From truth to disinterestedness in the seventeenth century. Social Studies of Science 22, 619–631.

Dworkin, Roland, 1996. Objectivity and truth: you'd better believe it. Philosophy and Public Affairs 25, 87–139.

Feyerabend, Paul K., 1975. Against Method. New Left Books, London and New York.

Haraway, Donna, 1988. Situated knowledges: the science question in feminism and the privilege of partial perspective. Feminist Studies 14, 575–599.

Harding, Sandra, 1995. 'Strong objectivity': a response to the new objectivity question. Synthese 104, 331–349.

Husserl, Edmund, 1970 [1900/1901]. Logical Investigations. Routledge, London.

Husserl, Edmund, 1970a [1936]. The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy. Northwestern University Press, Chicago.

Kuhn, Thomas, 1962. The Structure of Scientific Revolutions. The University of Chicago Press, Chicago.

Lakatos, Imre, 1978. The Methodology of Scientific Research Programmes. Cambridge University Press, Cambridge.

Lloyd, Elisabeth A., 1995. Objectivity and the double standard for feminist epistemologies. Synthese 104, 351–381.

Longino, Helen E., 1990. Science as Social Knowledge: Values and Objectivity in Scientific Inquiry. Princeton University Press, Princeton, NJ.

Miller, Richard W., 1991. Fact and method in the social sciences. In: Richard, Boyd, Philip, Gasper, Trout, J.D. (Eds.), The Philosophy of Science. MIT Press, Cambridge, MA.

Nagel, Thomas, 1997. The Last Word. Oxford University Press, Oxford.

Novick, Peter, 1988. That Noble Dream: The Objectivity Question and the American Historical Profession. Cambridge University Press, Cambridge.

Popper, Karl R., 1972. Objective Knowledge: An Evolutionary Approach. Clarendon Press, Oxford.

Putnam, Hilary, 1981. Reason, Truth, and History. Cambridge University Press, Cambridge.

Railton, Peter, 1985. Marx and the objectivity of science. In: Richard, Boyd, Philip, Gasper, Trout, J.D. (Eds.), The Philosophy of Science. MIT Press, Cambridge, MA.

Rorty, Richard, 1991. Solidarity or objectivity? In: Rorty, Richard (Ed.), Objectivity, Relativism, and Truth: Philosophical Papers. Cambridge University Press, Cambridge, pp. 21–34.

Shapin, Steven, Schaffer, Simon, 1985. Leviathan and the Air-Pump: Hobbes, Boyle and the Experimental Life. Princeton University Press, Princeton, NJ.

Sober, Elliott, 1982. Realism and independence. Nous 16, 369–385.

Weber, Max, 1949. The Methodology of the Social Sciences. Free Press, Glencoe, IL.

Observational Studies: Outcome-Based Sampling

Daniel McFadden, University of California, Berkeley, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 10804–10808, © 2001, Elsevier Ltd.

Abstract

Samples in observational studies are outcome-based if the probability of being sampled depends on responses that are the subject of study. This may occur inadvertently due to attrition, self-selection, or stratification that is correlated with response, or may result from deliberate design decisions to facilitate observation, as in case-control studies. Outcome-based samples require special attention to issues of robustness that arise because variations in sampling protocol may be confounded with response, and issues of inference that arise because procedures that ignore the selection induced by the sample frame may be statistically inconsistent. This article discusses these issues and outlines some practical methods for consistent estimation of population conditional response probabilities.

Introduction

An outcome-based sample in an observational study is one obtained by stratification on the basis of response behavior whose explanation is the sarget of study. Observations on response and explanatory variables (covariates) are collected within each stratum. These are then used for statistical inference on the conditional distribution of the response, given the covariates. For example, a study of occupational choice may draw a sample stratified by occupation, so the first stratum is a sample of engineers, the second stratum is a sample of educators, and so forth. Data are collected on covariates such as gender and utilization of training subsidies. The observations are then used to infer the impact of training subsidies on occupational choice.

Outcome-based sampling is also termed endogenous, intercept, on-site, experience-based, response-based, or choice-based sampling, and in biostatistics, retrospective or case-control sampling. To illustrate, case-control designs sample from one stratum that has a specified outcome (the cases) and a second stratum that does not (the controls). There may be further stratification of the controls so that they resemble the cases on some covariates such as age. Outcome-based samples can be contrasted with exogenous or prospective samples that sample randomly from the target population, or stratify solely on covariates. Outcome-based sampling methods have had wide application in social science, biostatistics, and engineering; examples are Akin et al. (1998), Domowitz and Sartain (1999), Laitila (1999), and van Etten et al. (1999).

Design and Robustness Issues in Endogenous Samples

Outcome-based samples are useful for studying rare responses that are difficult to observe in random samples of practical size, such as risk factors for a rare disease, and for studying outcomes that randomly sampled subjects are unable or unwilling to report. An example of the second type is the study of mental illness as a risk factor for homelessness, where the only practical sample methodology may be one that surveys cases who are homeless and controls who are not (Early, 1999; Lacy, 1997). Outcome-based sampling may also offer economies in recruiting and interviewing subjects. For example, consumers who purchase and register a product form an outcome-based stratum that the manufacturer can survey inexpensively.

Outcome-based samples may arise by deliberate design, as in case-control studies, but can also be unintentional. General purpose surveys often use cluster sampling to facilitate enumerating and contacting subjects. The cluster stratification is exogenous for most purposes, but has the properties of an outcome-based sample for behavior that is related to the stratification. For example, cluster sampling by census tract or area code becomes outcome based in studies of residential location of migration behavior. Further, screening, self-selection, and attrition create statistical problems like those in outcome-based samples. For example, a sampling protocol that selects subjects by random-digit dialing will over-sample individuals with more than one telephone line, resulting in a sample that is endogenous for study of the demand for telephone service. Differential attrition in which, for example, individuals who live alone are more difficult to contact or more frequently refuse to be interviewed, will produce a sample that is endogenous for the study of living arrangements.

Outcome-based samples are less robust than random samples. Because stratum boundaries coincide with the behavioral categories under study, variations in sampling protocol across strata are confounded with behavioral response. For example, in case-control studies of risk factors for disease, variations in question format between cases and controls, or even the context for questions induced by the disease status, can alter recall on exposure to hazards. Unlike prospective studies where randomization of treatments and double-blind designs can minimize the risks of contamination, retrospective studies are vulnerable to investigator bias and latent causal factors that are correlated with measured risk factors. As a result, differences in reported exposure to risk factors between cases and controls may arise from spurious sampling effects as well as from a true causal effect. The difficulty of avoiding contaminating bias in case-control settings is discussed further in Austin et al. (1994), Feinstein (1979), Lilienfeld and Lilienfeld (1979), and Taubes (1995).

Statistical Inference in Endogenous Samples

Statistical methods developed for random samples will often be inconsistent or inefficient when applied to endogenous samples. The essential problem is that the analysis is attempting to infer properties of the conditional distribution of outcomes given covariates, using observations that are drawn from conditional distribution of covariates given outcomes. The solution to the inference problem is to incorporate the mapping between the conditional distributions in the analysis, either by reweighting observations so that they behave *as if* they were drawn from a random sample, or by reweighting the probability model for a random sample so that it is consistent with the empirical sampling process. The statistical issues in analyzing outcome-based samples were treated in a seminal paper by Manski and Lerman (1977), with further results by Manski and McFadden (1981) and Cosslett (1981). A largely independent treatment of logistic regression in the case-control model appeared in biostatistics, notably in the papers of Anderson (1972) and Prentice and Pyke (1979). Recent papers by Hsieh et al. (1985) and Breslow (1996) sharpen the statistical analysis of case-control designs and show their relationship to the general problem of endogenous sampling.

Table 1 is a contingency table that depicts, schematically, the population probability law for an outcome variable y and a vector of covariates z. (This exposition treats y and z as discrete, but the discussion applies with minor modifications to the case where y and/or some components of z are continuous.) The joint probability of a (y, z) cell can be written as the product of the conditional probability of y given z times the marginal probability of z, $p(y, z) = P(y \mid z) \cdot p(z)$. The row sums give the marginal probability $p(z)$ of z, and the column sums give the marginal probability $q(y) = \sum_z P(y \mid z) \cdot p(z)$ of y. Dividing a cell probability by its column sum gives the conditional probability of z given y, $Q(z \mid y) = P(y \mid z)p(z) / q(y)$. This mapping between conditional probabilities is Bayes' Law. The target of statistical analysis is the conditional probability $P(y \mid z)$, sometimes termed the response probability. In applications, $P(y \mid z)$ is usually assumed to be invariant under treatments that alter the marginal probability of z; then knowledge of $P(y \mid z)$ permits the analysis to forecast y in new populations or under policy treatments where the z distribution is changed. (A conditional probability with this invariance property is sometimes said to define a causal model. It is true that a causal structure will imply this invariance property, but it is also possible for the invariance property to hold, making forecasting possible, without the presence of a deeper causal structure. Further, there are straightforward statistical tests for the invariance property, while detection of true causal structures is beyond the reach of statistics. For these reasons, it is best to avoid the language of causality and concentrate instead on invariance properties.)

Table 1

	y_1	y_2	…	y_J	
z_1	$P(y_1 \mid z_1, \beta_0)p(z_1)$	$P(y_2 \mid z_1, \beta_0)p(z_1)$	…	$P(y_J \mid z_1, \beta_0)p(z_1)$	$p(z_1)$
z_2	$P(y_1 \mid z_2, \beta_0)p(z_2)$	$P(y_2 \mid z_2, \beta_0)p(z_2)$	…	$P(y_J \mid z_2, \beta_0)p(z_2)$	$p(z_2)$
:	:	:		:	:
z_K	$P(y_1 \mid z_K, \beta_0)p(z_K)$	$P(y_2 \mid z_K, \beta_0)p(z_K)$	…	$P(y_J \mid z_K, \beta_0)p(z_K)$	$p(z_K)$
	$q(y_1)$	$q(y_2)$	…	$q(y_1)$	1

Random sampling draws from the table in proportion to the cell probabilities. Exogenous stratification draws rows, with probabilities that may differ from the population marginal probabilities $p(z)$, and then within a row draws columns in proportion to their population conditional probabilities $P(y \mid z)$. A simple outcome-based sampling design draws columns, with probabilities that may differ from the population marginal probabilities $q(y)$, then within a column draws rows in proportion to their conditional probabilities $Q(z \mid y) = P(y \mid z)p(z) / q(y)$.

More complex endogenous sampling designs are also possible. A general framework that permits a unified analysis of many sampling schemes characterizes the sampling protocol for a stratum s in terms of a probability $R(y, z, s)$ that a member of the population in cell (y, z) will qualify for the stratum. The joint probability that a member of the population is in cell (y, z) and will qualify for stratum s is then $R(z, y, s) \cdot P(y \mid z, \beta_0) \cdot p(z)$. Then the proportion of the population qualifying into the stratum, or qualification factor, is $r(s) = \sum_z \sum_y R(z, y, s) \cdot P(y \mid z) \cdot p(z)$, and the conditional distribution of (z, y) given qualification is $R(z, y, s) \cdot P(y \mid z) \cdot p(z) / r(s)$. When a fraction of the sample $f(s)$ is drawn from stratum s, the probability law for an observation from the pooled sample is $g(y, z) = \sum_s R(z, y, s) \cdot P(y \mid z) \cdot p(z) f(s) / r(s)$. The conditional distribution of y given z in this pooled sample is

$$\alpha_y^* + \gamma_{yv}^* = \alpha_y + \gamma_{yv} + \log\left(\sum_s R(v, y, s) \cdot f(s)/r(s)\right) g(y|z)$$

$$= \frac{P(y|z) \sum_\Theta R(z, y, s) \cdot f(s)/r(s)}{\sum_{\tau'} P(y'|z) \sum_\Theta R(z, y', s) \cdot f(s)/r(s)}$$

Note that this conditional probability depends on the marginal distribution of z only through the qualification factors.

When the sampling protocol is exogenous (i.e., $R(y, z, s)$ does not depend on y), the conditional probability $g(y \mid z)$ for the pooled sample equals the population conditional probability $P(y \mid z)$. Consequently, any statistical inference procedure designed to reveal features of the conditional probability $P(y \mid z)$ in random samples will apply to an exogenously stratified sample. In particular, if $P(y \mid z)$ is in a parametric family, then maximization of the random sample likelihood function in an exogenously stratified sample will have the same properties as in a random sample. However, in an endogenous sample in which the qualification probability $R(y, z, s)$ does depend on y, the conditional probability $g(y \mid z)$ for the pooled sample is *not* equal to $P(y \mid z)$. Consequently, statistical inference assuming that the data generation process is described by $P(y \mid z)$ is generally statistically inconsistent. Also, the distribution of covariates in an endogenous sample will differ from their population distribution, with

$$g(z) = p(z) \cdot \sum_s (f(s)/r(s)) \sum_y R(z, y, s) P(y|z)$$

and a corresponding correction factor must be applied to the sample empirical distribution of z to estimate population quantities consistently.

Manski and McFadden (1981) propose that statistical inference when $P(y \mid z)$ is parametric be based on the conditional likelihood $g(y \mid z)$, and term this the conditional maximum likelihood (CML) method. When the qualification factors $r(s)$ and sample frequencies $f(s)$ are known or can be estimated consistently from external samples, and the forms of $P(y \mid z)$ and $R(z, y, s)$ allow identification of any unknown parameters in $R(z, y, s)$, this approach is consistent. In general, the probability $g(y \mid z)$ is not in the same parametric family as $P(y \mid z)$. To illustrate, suppose a population has a binomial probit dose-response curve, $P(2 \mid z) = \Phi(\alpha + z\beta)$, and $P(1 \mid z) = 1 - \Phi(\alpha + z\beta)$. Suppose the sample consists of a randomly sampled stratum 1 with $R(z, y, 1) \equiv 1$, plus a stratum 2 drawn from the population with response $y = 2$, with $R(z, y, 2)$ equal to one if $y = 2$, and zero otherwise. This is called an enriched sample. The qualification factors are $r(1) = 1$ and $r(2) = q(2)$. If $q(2)$ is known, a consistent estimate of the slope parameter β in the dose-response curve can be obtained by the CML method with

$$g(1|z) = \frac{[1 - \Phi(\alpha + z\beta)]f(1)}{[1 - \Phi(\alpha + z\beta)f(1) + \Phi(\alpha + z\beta)[f(1) + f(2)/q(2)]}$$

By contrast, likelihood maximization using $P(y \mid z)$ is not consistent for β.

An important simplification of the CML method occurs for what in biostatistics is termed logistic regression in case-control designs. Suppose that the vector of covariates is partitioned into components $z = (v, x)$ with v discrete. (In biostatistics, v will often include variables such as age and gender whose distributions are matched between cases and controls.) Suppose that $P(y \mid v, x)$ has a multinomial logit form,

$$P(y|v,x) = \exp(\alpha_y + \gamma_{yv} + x\beta_y) / \sum_{y} \exp(\alpha_y + \gamma_{y'v} + x\beta'_y y)$$

In this model, the β_y are slope coefficients for the covariates x, and α_y and γ_{yv} are response-specific effects and interactions of response-specific and v-specific effects. Suppose that the qualification probability $R(v, x, y, s)$ does not depend on x. For identification, a normalization such as $\alpha_1 = \gamma_{1v} = \beta_1 = 0$ is imposed. The conditional probability $g(y \mid z)$ is again of multinomial logit form, with the same β_y parameters but with the remaining parameters shifted from their population values by sampling factors,

$$g(y|v,x) = \exp(\alpha^*_y + \gamma^*_{yv} + \beta_y) / \sum_{y'} \exp(\alpha'_{y'} + \gamma^*_{y'v} + \beta_{y'})$$

with

$$\alpha^*_y + \gamma^*_{yv} = \alpha_y + \gamma_{yv} + \log\left(\sum_s R(v,y,s) \cdot f(s)/r(s)\right)$$

Note that consistent estimation of this model requires the inclusion of all the alternative specific effects and interactions that are modified by sampling factors. However, if these variables are included, then the slope parameters β_y are estimated consistently without further adjustments for endogenous sampling. (If the raising factors are estimated rather than known, there is an additional contribution to the asymptotic covariance matrix (Hsieh et al., 1985). If the model had included interactions of v and x, then the coefficients on these interactions would contain sampling factors that must be removed to obtain consistent estimates of the corresponding population interactions). The simplification above was first noted by Anderson (1972).

Another solution to inference in choice-based samples is the weighted exogenous sample maximum likelihood (WESML) method proposed by Manski and Lerman (1977) in their seminal paper on this subject. This solution weights the observations, with a weight for (y, z) that is inversely proportional to $\sum_s R(z, y, s) \cdot f(s) / r(s)$. When these weights are finite and can be calculated or estimated from external data, consistent estimates can be obtained when the weighted observations are treated *as if* they were drawn from an exogenous sample. This is a pseudo-maximum likelihood method, and the associated covariance matrix will be of the sandwich form, with an outer product of the pseudo-score in the middle and the inverse of the pseudo-information matrix on the outside. For the case of pure outcome-based sampling, where the strata correspond to the possible values of y and $R(y, z, s)$ is one when $y = s$ and zero otherwise, the weights are $r(y) / f(y)$, the ratio of the population to the sample frequency of observing y.

A third approach to inference in outcome-based samples, developed by Cosslett (1981), achieves a semiparametric efficiency bound when $P(y \mid z)$ is parametric. This estimator is derived by maximum likelihood estimation that first formally concentrates out the marginal probability $p(z)$. The result is an objective function similar to $g(y \mid z)$, but with parametric sampling factors that satisfy side conditions. Hsieh et al. (1985) adapt Cosslett's analysis to improve the efficiency of the CML method in the presence of auxiliary information.

Extensions and Related Issues

Closely related to outcome-based sampling in the form described here is the problem of analyzing samples that are subject to self-selection. Several strands in the statistical and econometric literature have investigated estimators appropriate to such data: a seminal paper by Heckman (1974) on sample selection, further work on endogenous stratification by Hausman and Wise (1977), and related work on switching regression by Goldfeld and Quandt (1975).

Extensions of the basic framework for inference in outcome-based samples have been made for a variety of problems. Breslow (1996) provides methods for complex case-control sample designs. Imbens (1992) provides methods for combining outcome-based survey data with aggregate statistics. Imbens and Lancaster (1996) and McFadden (in press) have studied the problem of analysis of endogenously recruited panels.

Conclusions

Samples in observational studies often fail to qualify as random or exogenously stratified, and instead have the properties of endogenous or outcome-based samples. This may occur inadvertently due to attrition, self-selection, or stratification that is correlated with the response being studied, or may result from deliberate design decisions to facilitate

observation, as in case-control studies. This article has described the robustness issues and inference methods that are relevant for these samples.

Bibliography

Akin, J., et al., 1998. Price elasticities of demand for curative health care with control for sample selectivity on endogenous illness: An analysis for Sri Lanka. Health Economics 7, 509–531.

Anderson, J., 1972. Separate sample logistic discrimination. Biometrika 59, 19–35.

Austin, H., Hill, H., Flanders, W., Greenberg, R., 1994. Limitations in the application of case-control methodology. Epidemiologic Reviews 16, 65–76.

Breslow, N., 1996. Statistics in epidemiology: The case-control study. Journal of the American Statistical Association 91, 14–28.

Cosslett, S., 1981. Maximum likelihood estimation for choice-based samples. Econometrica 49, 1289–1316.

Domowitz, I., Sartain, R., 1999. Determinants of the consumer bankruptcy decision. Journal of Finance 54, 403–420.

Early, D., 1999. A microeconomic analysis of homelessness: An empirical investigation using choice-based sampling. Journal of Housing Economics 8, 312–327.

van Etten, M., Neumark, Y., Anthony, J., 1999. Male-female differences in the earliest stages of drug involvement. Addiction 94, 1413–1419.

Feinstein, A., 1979. Methodological problems and standards in case-control research. Journal of Chromic Diseases 32, 35–41.

Goldfeld, S., Quandt, R., 1975. Estimation in a disequilibrium model and the value of information. Journal of Econometrics 3 (4), 325–348.

Hausman, J., Wise, D., 1977. Social experimentation, truncated distributions, and efficient estimation. Econometrica 45 (4), 919–938.

Heckman, J., 1974. Shadow prices, market wages, and the labor supply. Econometrica 42, 679–694.

Hsieh, D., Manski, C., McFadden, D., 1985. Estimation of response probabilities from augmented retrospective observations. Journal of the American Statistical Association 80, 651–662.

Imbens, G., 1992. An efficient method of moments estimator for discrete choice models with choice-based sampling. Econometrica 60, 1187–1214.

Imbens, G., Lancaster, T., 1996. Efficient estimation and stratified sampling. Journal of Econometrics 74, 289–318.

Lacy, M., 1997. Efficiently studying rare events: Case-control methods for sociologists. Sociological Perspectives 40, 129–154.

Laitila, T., 1999. Estimation of combined site-choice and trip-frequency models of recreational demand using choice-based and on-site samples. Economics Letters 64, 17–23.

Lilienfeld, A., Lilienfeld, D., 1979. A century of case-control studies. Progress? Journal of Chronic Diseases 32, 5–13.

Manski, C., Lerman, S., 1977. The estimation of choice probabilities from choice-based samples. Econometrica 45, 1977–1988.

Manski, C., McFadden, D., 1981. Alternative estimators and sample designs for discrete data analysis. In: Manski, C., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications, pp. 51–111.

McFadden, D. On endogenously recruited panels. Journal of Applied Econometrics, in press.

Prentice, R., Pyke, R., 1979. Logistic disease incidence models and case-control studies. Biometrica 66, 403–411.

Taubes, G., 1995. Epidemiology faces its limits. Science 269, 164–169.

Observational Studies: Overview

Paul R Rosenbaum, Wharton School, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

An observational study is an empirical but nonexperimental investigation of the effects caused by a treatment. In an experiment, such as a clinical trial, the investigator assigns subjects at random to treatment groups ensuring that comparable subjects receive competing treatments. In an observational study, the investigator does not control the assignment of treatments, with the consequence that the individuals in different treatment groups may not have been comparable prior to treatment. Analytical adjustments, such as matching, are used to remove overt bias, that is, pretreatment differences that are accurately measured and recorded. In addition, there may be pretreatment differences that were not recorded, called hidden bias. An observational study is designed to permit detection of the most plausible hidden biases. In addition, the analysis of an observational study investigates the sensitivity of conclusions to hidden biases of plausible magnitudes. Strategies for making studies less sensitive to unmeasured biases are discussed.

Introduction

Definition

Cochran (1965) defined an *observational study* as a comparison of treated and control groups in which

> the objective is to elucidate cause-and-effect relationships [… in which it] is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects it is desired to discover, or to assign subjects at random to different procedures.

Treatments, policies, and interventions that are expected to benefit recipients can be studied in experiments that randomly allocate experimental subjects to treatment groups, so comparable subjects receive competing treatments (Boruch, 1997; Piantadosi, 2005). Observational studies are common, nonetheless, in most fields that study the effects of treatments on people, because harmful or unwanted treatments cannot be imposed on human subjects. When treatments are not randomized, subjects assigned to competing treatments may not be comparable, so differences in outcomes may or may not be effects caused by the treatments (Rubin, 1974).

Examples

Two examples of observational studies are described. Later sections refer to these examples.

Long-Term Psychological Effects of the Death of a Close Relative

Lehman et al. (1987) attempted to estimate the long-term psychological effects of the sudden death of a spouse or a child in a car crash. Starting with 80 bereaved spouses and parents, they matched 80 controls drawn from 7581 individuals who came to renew a driver's license, matching for gender, age, family income before the crash, education level, number and ages of children. Contrasting their findings with the views of Bowlby and Freud they concluded:

> Contrary to what some early writers have suggested about the duration of the major symptoms of bereavement … both spouses and parents in our study showed clear evidence of depression and lack of resolution at the time of the interview, which was five to seven years after the loss occurred.

Effects on Children of Occupational Exposures to Lead

Morton et al. (1982) asked whether children were harmed by lead brought home in the clothes and hair of parents who were exposed to lead at work. They matched children whose parents worked in a battery factory to unexposed control children of the same age and neighborhood, comparing the level of lead found in the children's blood, finding elevated levels of lead in exposed children.

In addition, they compared exposed children whose parents had varied levels of exposure to lead at the factory, finding that parents who had higher exposures on the job, in turn, had children with more lead in their blood. Finally, they compared exposed children whose parents had varied hygiene upon leaving the factory at the end of the day, finding that poor hygiene of the parent predicted higher levels of lead in the blood of the child.

Central Issues

Without random assignment of treatments, treated and control groups may not have been comparable prior to treatment, so differing outcomes may reflect either biases from these pretreatment differences or else effects actually caused by the treatment. The difficulty in distinguishing treatment effects from biases is the key consequence of the absence of randomization.

A variable measured prior to treatment is not affected by the treatment and is called a covariate. A variable measured after treatment may have been affected by the treatment and is called

an outcome. An analysis that does not carefully distinguish covariates and outcomes can introduce biases into the analysis where none existed previously.

A pretreatment difference between treated and control groups is called an *overt bias*, if it is accurately measured in the data at hand, and it called a *hidden bias*, if it is not measured. For instance, if treated subjects are observed to be somewhat older than controls, and if age is recorded, then this is an overt bias. If treated subjects consumed more illegal narcotics than controls, and if accurate records of this are not available, then this is a hidden bias. Typically, overt biases are immediately visible in the data at hand, while hidden biases are a matter of concerned speculation and investigation.

In most observational studies, an effort is made to remove overt biases using one or more analytical methods, such as matched sampling, stratification, or model-based adjustments such as covariance adjustment. Adjustments are discussed in Section Adjusting for Overt Biases. Once this is accomplished, attention turns to addressing possible hidden biases, including efforts to detect hidden biases and to study the sensitivity of conclusions to biases of plausible magnitude. Hidden biases are discussed in Sections Detecting Hidden Bias; Appraising Sensitivity to Hidden Bias; Reducing Sensitivity to Hidden Bias.

Adjusting for Overt Biases

Matched Sampling

Selecting from a Reservoir of Potential Controls

Matching is the most direct and intuitive method of adjustment, in which an effort is made to compare each treated individual to one or more comparable controls. Matched sampling is most common when a small treated group is available together with a large reservoir of potential controls (Rosenbaum, 2010: Part II; Stuart, 2010). For instance, in the study of bereavement by Lehman et al. (1987), there were 80 bereaved spouses and parents and 7581 potential controls, from whom 80 matched controls were selected.

The structure of the bereavement study is typical; computerized administrative records aid in identifying and matching treated and control subjects, but additional information is required from matched subjects for research purposes; in this case, psychiatric outcomes are required. Often it is not practical to obtain the needed research data from all controls in the reservoir and instead a sample of controls is used. Matched sampling selects controls who appear comparable to treated subjects.

Goals of Matching

Matching attempts to do three things: (1) produce matched pairs or sets that appear similar in terms of observed covariates, (2) produce treated and control groups with similar distributions of observed covariates, (3) remove that part of the bias in estimated treatment effects that is due to overt biases or imbalances in observed covariates. For instance, with pairs matched for age, goal (1) would be accomplished if each treated subject were matched to a control of the same age, whereas goal (2) would be accomplished if the entire treated group had an age distribution that closely resembled the age distribution in the matched control group, for instance, the same mean age, the same quartiles of the age distribution, and so on.

It is possible to demonstrate (Rosenbaum and Rubin, 1983) that goal (1) suffices but is not necessary for goal (2), and that goal (2) suffices for goal (3). Pairs that are individually matched for age will balance age, but it is also possible to balance age with matched pairs that are not individually matched for age, and any matching that balances the distribution of age removes the part of the bias in estimated treatment effects due to imbalances in age. This is fortunate, because it is difficult to match individually on many covariates at once, but it is not so difficult to balance many covariates at once. For instance, with 20 binary covariates, there are 2^{20} or about a million types of individuals, so even with thousands of potential controls, it will often be difficult to find a control who matches a treated subject on all 20 covariates. Nonetheless, balancing 20 covariates is often quite feasible. As a result, modern matching algorithms focus on balancing covariates, while comparability within pairs plays a secondary role. A tool in balanced matching is the propensity score.

Propensity Score

The *propensity score* is the conditional probability of receiving the treatment rather than the control given to the observed covariates (Rosenbaum and Rubin, 1983). Note carefully that the propensity score is defined in terms of the observed covariates, even if there may be hidden biases due to unobserved covariates. In the simplest randomized experiment, treatment or control is assigned by the flip of a fair coin, and the propensity score equals one-half for all subjects no matter what covariates are observed. In an observational study, subjects with certain observed characteristics may be more likely to receive either treatment or control, so the propensity score varies with these observed characteristics.

Matching on one variable, the propensity score, tends to balance all of the observed covariates. Matching on the propensity score together with a few other observed covariates – say, matching on the propensity score together with age and gender – also tends to balance all of the observed covariates. If it suffices to adjust for the observed covariates – that is, if there is no hidden bias due to unobserved covariates – then it also suffices to adjust for the propensity score alone. These results are theorems 1 through 4 of Rosenbaum and Rubin (1983).

In practice, the propensity score is unknown and must be estimated. For instance, one might estimate the propensity score using logistic regression of assigned treatment on observed covariates, perhaps with interactions and other transformations. Matching on a single variable, such as the estimated propensity score, is often practical whereas matching on many covariates is not; yet matching on the propensity score balances many covariates.

Matching Algorithms

Modern matching algorithms formulate matching as the solution to an optimization problem. Often, the optimal match minimizes the total covariate distance within disjoint matched

pairs or matched sets subject to constraints that force covariate balance (Rosenbaum, 2010: Part II; Stuart, 2010; Zubizarreta, 2012). A common covariate distance severely penalizes mismatches for the estimated propensity score and adds a robust Mahalanobis distance using key covariates (Rosenbaum, 2010: Chapter 8). Additional constraints such as 'fine balance' require perfect balance for that a nominal covariate, perhaps with thousands of levels, without constraining who is matched to whom (Rosenbaum, 2010: Chapter 10; Zubizarreta, 2012; Zubizarreta et al., 2013). In full matching, disjoint matched sets may have one treated subject and several controls or one control and several treated subjects; it is the form of an optimal stratification (Hansen, 2004; Stuart and Green, 2008; Rosenbaum, 2010: Section 8.6; Stuart, 2010: Section 3.2.2).

Several packages in the R statistical system construct optimal matched samples. The pairmatch and fullmatch functions in Hansen's (2007) optmatch package in R produce optimal treatment vs control matched pairs or sets. The nbpMatching package in R creates optimal matches without prespecified groups; see Lu et al.'s (2011) for a review of how this package may strengthen an instrumental variable or match with doses or match with longitudinal treatment assignment. Zubizarreta's (2012) mipmatch package in R implements a flexible form of optimal matching subject to balance constraints.

Model-Based Adjustments

Matched sampling compares treated subjects directly to actual controls who appear comparable in terms of observed covariates. For instance, as noted in Section Long-term Psychological Effects of the Death of a Close Relative, Lehman et al. (1987) compared bereaved individuals to matched controls who were comparable in terms of age, gender, income, education, and family structure. In contrast, model-based adjustments, such as covariance adjustment, use data on treated and control subjects without regard to their comparability, relying on a model, such as a linear regression model, to predict how subjects would have responded under treatments they did not receive.

If the model used for model-based adjustment is precisely correct, then model-based adjustments may be more efficient than matching, producing estimates with smaller standard errors. On the other hand, if the model is substantially incorrect, model-based adjustments may not only fail to remove overt biases, they may even increase them, whereas matching and stratification are fairly consistent at reducing overt biases (Rubin, 1979).

Alas, in practice, one can often detect substantial errors in a model, but rarely can have much confidence that the model is precisely correct, so there is rarely a clear choice between efficiency of model-based adjustments when the model is true and the robustness of matching and stratification when it is not. This has led to the use of combined techniques.

In simulation studies, Rubin (1979) compared matching, covariance adjustment, and combinations of these techniques, concluding that the use of covariance adjustment within matched pairs was superior to either method alone, being both robust and efficient.

Detecting Hidden Bias

Elaborate Theories

In an observational study, the investigator takes active steps to detect hidden biases – to collect data recording visible traces of unobserved pretreatment differences if they exist. In this, the investigator is aided by an 'elaborate theory,' defined by Sir Ronald Fisher in the following discussion from Cochran (1965: Section 6)

> About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: 'Make your theories elaborate.' The reply puzzled me at first, since by Occam's razor, the advice usually given is to make theories as simple as is consistent with known data. What Sir Ronald meant, as subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold.

Consider the study of lead exposures by Morton et al. (1982), discussed in Section Effects on children of occupational exposures to lead. Their elaborate theory made three predictions: (1) higher lead levels in the blood of exposed children than in matched control children, (2) higher lead levels in exposed children whose parents had higher exposure on the job, and (3) higher lead levels in exposed children whose parents practiced poorer hygiene upon leaving the factory. Since each of these predictions was consistent with observed data, to attribute the observed associations to hidden bias rather than an actual effect of lead exposure, one would need to postulate biases that could produce all three associations. The use of elaborate theories and further examples are discussed in Campbell and Stanley (1963), Campbell (1988), Rosenbaum (2002: Chapters 6–9; 2010: Chapters 5, 17, and 19), Shadish et al. (2002), and Sobel (2000).

Two Control Groups

The simplest and perhaps the most common study design to detect hidden bias uses two control groups. Since none of the control group received the treatment, systematic differences between the control groups cannot be effects of the treatment and must instead be some form of hidden bias. One seeks two control groups that would differ in their outcomes if a specific unobserved bias was present. In this way, hidden biases should produce a difference between the two control groups, whereas an actual treatment effect should produce relatively similar results in the two control groups and a different result in the treated group.

This is the principle of 'control by systematic variation,' which is due in general to Bitterman (1965), and is applied to observational studies by Campbell (1969) and Rosenbaum, (2002: Chapter 8, 2010: Chapters 5 and 19). If one cannot ensure comparability of treated and control groups with respect to a particular variable, say by matching, because that variable cannot be observed, then one seeks two control groups that are known to differ markedly on that unobserved variable, that is, to systematically vary the unobserved variable. Control by

systematic variation is an instance of 'strong inference' in the sense discussed by Platt (1964). Heller et al. (2010) use a computerized construction to build a second control group by not matching for certain observed covariates that are possibly innocuous, that is, possibly without influence on the outcome.

Appraising Sensitivity to Hidden Bias

The analytical adjustments discussed in Section Adjusting for Overt Biases can often remove overt biases accurately recorded in the data at hand, but there is typically a legitimate concern that treated and control groups differed prior to treatment in ways that were not recorded. A sensitivity analysis asks how such hidden biases might alter the conclusions of the study.

The first formal method of sensitivity analysis was developed to aid in appraising the effects of cigarette smoking on human health. Responding to objections that smoking might not cause lung cancer, but rather that there might be a genetic predisposition both to smoke and to develop lung cancer, Cornfield et al. (1959: p. 40) wrote:

> … if cigarette smokers have nine times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone X-producers among cigarette smokers must be at least nine times greater than among nonsmokers.

This statement was an important conceptual advance. It is, of course, commonly understood that association does not imply causation – that any observed association can be explained by a hypothetical argument, perhaps an implausible hypothetical argument, postulating an unobserved variable. Although this common understanding is always correct, it contributes little to our appraisal of an observational study, because it makes no reference to what was actually observed in this empirical investigation, and no reference to scientific knowledge about the phenomenon under study. A sensitivity analysis replaces the statement – 'association does not imply causation' – by a specific statement about the magnitude of hidden bias that would need to be present to explain the associations actually observed. Strong associations in large studies can only be explained by large, perhaps implausible, biases. Aspects of the method of Cornfield et al. (1959) are discussed by Greenhouse (1982). Although important as a conceptual advance, this method of sensitivity analysis is limited in that it ignores sampling error produced by having a finite sample rather than the population as a whole, and it is applicable only to binary outcomes.

A related method addresses sampling error and may be used with outcome measures of all kinds. Observing that in a randomized experiment, subjects in the same stratum or matched set have the same probability of receiving the treatment rather than the control, the sensitivity analysis introduces a sensitivity parameter Γ, which measures the degree of departure from random assignment of treatments. Specifically, two subjects who look the same in terms of observed covariates, and so are in the same matched set or stratum, may differ in their odds of receiving the treatment by at most a factor of Γ. In a randomized experiment, $\Gamma = 1$. If $\Gamma = 2$ in an observational study, then if two people look the same in terms of observed covariates, one might be twice as likely to receive the treatment as the other because they differ in terms of a covariate not observed. For each value of Γ, it is possible to place bounds on conventional statistical inferences – e.g., for $\Gamma = 2$, perhaps, the true p-value is unknown, but must be between 0.001 and 0.021. Similar bounds are available for point and interval estimates. One then asks, How large must Γ be before the conclusions of the study are qualitatively altered? If the conclusions change for $\Gamma = 1.05$, then the study is very sensitive to hidden bias – very small departures from a randomized experiment can explain the observed association between treatment and outcome – but if the conclusions do not change for $\Gamma = 5$, then only quite dramatic hidden biases can explain the observed association. A nontechnical introduction to this method of sensitivity analysis is given in Rosenbaum (1991), and a detailed development in Rosenbaum (2002: Section 4). In R, methods of this type are available in the senmv function of the sensitivitymv package, in the adaptive.noether.brown function of the SensitivityCaseControl package, and in Luke Keele's rbounds package.

Methods of sensitivity analysis are discussed by Copas and Li (1997), Cornfield et al. (1959), Diprete and Gangl (2004), Gastwirth (1992), Greenhouse (1982), Liu et al. (2013) and Rosenbaum (2002, 2010), among others.

Reducing Sensitivity to Hidden Bias

The Goals: Smaller Unmeasured Biases; Results that Are Insensitive to Larger Biases

Given the choice, one would prefer conclusions insensitive to hidden biases, so that there is less ambiguity about the effects caused by the treatment. In part, this means selecting circumstances for the study so that plausible unmeasured biases are not extremely large. In part, this means designing the study so that it is likely to be insensitive to larger biases. For a fixed study design, as the sample size increases without bound, there is limiting sensitivity to unmeasured biases called the design sensitivity (Rosenbaum, 2004, 2010: Part III; Zubizarreta et al., 2013; Stuart and Hanna, 2013). Eventually, in large enough samples, the study will be insensitive to biases smaller than the design sensitivity and sensitive to biases larger than the design sensitivity. We prefer a study setting and design in which anticipated biases are not large, but the design sensitivity is large. A diverse collection of strategies used in observational studies may be viewed as efforts to bring about this situation.

This section first discusses the choice of circumstances for the study so anticipated biases are not extremely large. Then attention turns to instrumental variables. Finally, aspects of study design that affect design sensitivity are discussed.

Choice of Circumstances

In the earliest stages of planning an observational study, the investigator chooses the circumstances in which the study will be conducted. These choices often affect both the size of the treatment effect and the magnitude of hidden bias that is

plausible. As a result, these choices affect the sensitivity of the study to hidden bias. Some illustrations of this use of 'choice' will be briefly mentioned. See Rosenbaum (1999, 2002: Chapter 11) for detailed discussion of these cases and others with examples, and see Meyer (1995), Angrist and Krueger (1999), Vandenbroucke (2004), Dunning (2012), and Sekhon and Titiunik (2012) for related discussion.

The investigator examines a broad research hypothesis, one that makes predictions about innumerable circumstances, but examines that hypothesis in particular circumstances where dramatic effects are anticipated, free of distortions from other factors that affect these same outcomes. The investigator seeks a situation in which there is a genuine control group completely shielded from the treatment, and a treated group that received the treatment at an intense dose. One prefers a treatment that is imposed suddenly, not gradually, and haphazardly, not in direct response to characteristics of the individuals under study.

For instance, in the study in Section Long-term Psychological Effects of the Death of a Close Relative by Lehman et al. (1987), the effects of bereavement were studied following sudden deaths of close relatives from car crashes. Moreover, employing standardized criteria, they used only those car crashes for which the victim's car was not responsible, reasoning that alcohol or drug abuse or certain forms of psychopathology might increase the risk of a car crash and also be associated with unfavorable psychological outcomes in the family.

As another example, Angrist and Lavy 1999 studied the effects of class size on academic achievement in carefully chosen circumstances. In Israel, a rule dating back to Maimonides requires class rooms with no more than 40 students, so a class with 41 students is divided in two classes of about 20 students. Whereas in the United States, a class of size 40 is likely to be found in a different economic environment from a class of size 20, in Israel a comparatively haphazard event – the enrollment of an additional child – often separates these two situations.

Instrumental Variables

In some circumstances, the treatment assignment itself is subject to hidden biases that are plausibly quite large, but there is another variable, called an instrument, which has no effect on the outcome, but is associated with the treatment. The hope, sometimes realistic, is that instrument is free of hidden bias, or at least that it is affected by hidden biases that are much smaller. In an instrumental variable analysis, the magnitude of the treatment effect is modeled in terms of the treatment received, but the assignment variable is the instrument, not the treatment.

A nice example with a clear theoretical development is given by Angrist et al. (1996). They consider aspects of the health effects of military service, but of course, people who serve in the military differ in many ways from those who do not. Their instrument is the draft lottery in use in the United States between 1970 and 1973. Many people who were drafted did not serve, and many who served were not drafted, so the draft is not at all the same as the treatment under study, namely military service. However, the draft was close to random, so it is not plausible that it is affected by large hidden biases. The instrumental variable analysis models the treatment effect in terms of actual service, but uses the draft as the 'assignment variable' in a sense made precise in Angrist et al. (1996) and in the discussion published with that paper.

An instrument is said to be weak, if changes in the instrument have only slight impact on exposure to the treatment. Bound et al. (1995) show that weak instruments create problems for conventional methods of analysis, such as two-stage least squares, so 95% confidence intervals have coverage that is far below 95%. More importantly, weak instruments are invariably sensitive to small unmeasured biases (Small and Rosenbaum, 2008). Baiocchi et al. (2010) use optimal nonbipartite matching to strengthen a weak instrument, thereby reducing both of the problems with weak instruments.

Tactics to Increase Design Sensitivity

The degree to which a study design is sensitive to unmeasured biases – more precisely, the design sensitivity – is affected by many factors over which the investigator may exert some control. To understand how design sensitivity behaves is to know how to make design decisions that will affect the degree to which the study is sensitive to unmeasured biases. This rather large topic is surveyed in Rosenbaum (2010: Part III), but a few aspects can be described briefly here. If the treatment under study comes in various doses or magnitudes of treatment, and if larger doses produce larger effects, then a study that eliminates negligible and small doses will have a larger design sensitivity (Rosenbaum, 2010: Section 17.3). The same treatment effect seen with less dispersion will have a larger design sensitivity, so a smaller study in which matched pairs are more homogeneous may exhibit less sensitivity to unmeasured biases than a much larger study of heterogeneous pairs, providing the typical effect is the same in both studies (Rosenbaum, 2010: Chapter 15). Some treatments affect only a fraction of the people who receive the treatment, and in this case studies that look at typical or average treatment effects will be more sensitive to unmeasured biases than studies that emphasize the affected portion of the distribution of treatment effects (Rosenbaum, 2010: Chapter 16). In general in an observational study, to know how to decide between two choices in design is to know how those choices affect the design sensitivity.

See also: Causal Inference: Overview; Experimental Design: Overview; Experimental Design: Randomization and Social Experiments; Panel Surveys: Uses and Applications; Quasi- and Field Experiment; Quasi-Experimental Designs.

Bibliography

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables (with discussion). Journal of the American Statistical Association 91, 444–469.

Angrist, J.D., Krueger, A.B., 1999. Empirical strategies in labor economics. Handbook of Labor Economics 3, 1277–1366.

Angrist, J.D., Lavy, V., 1999. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Quarterly Journal of Economics 114, 533–575.

Baiocchi, M., Small, D.S., Lorch, S., Rosenbaum, P.R., 2010. Building a stronger instrument in an observational study of perinatal care for premature infants. Journal of the American Statistical Association 105, 1285–1296.

Bitterman, M., 1965. Phyletic differences in learning. American Psychologist 20, 396–410.

Boruch, R.E., 1997. Randomized Experiments for Planning and Evaluation: A Practical Guide. Sage Publications, Thousand Oaks, CA.

Bound, J., Jaeger, D.A., Baker, R.M., 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. Journal of the American Statistical Association 90, 443–450.

Campbell, D.R., 1966. Pattern matching as an essential in distal knowing. In: Hammond, K.R. (Ed.), The Psychology of Egon Brunswik. Holt, Rinehart and Winston, New York, pp. 81–106.

Campbell, D.T., 1969. Prospective: artifact and control. In: Rosenthal, R.L., Rosnow, R.L. (Eds.), Artifact in Behavioral Research. Academic Press, New York.

Campbell, D.T., 1988. In: Methodology and Epistemology for Social Science: Selected Papers. University of Chicago Press, Chicago, IL, pp. 315–333.

Campbell, D.T., Stanley, J.C., 1963. Experimental and Quasi-experimental Designs for Research. Rand McNally, Chicago, IL.

Cochran, W.G., 1965. The planning of observational studies of human populations (with discussion). Journal of the Royal Statistical Society, Series A 128, 134–155.

Copas, J.B., Li, H.G., 1997. Inference for non-random samples (with discussion). Journal of the Royal Statistical Society, Series B 59, 55–96.

Cornfield, J., Haenszell, W., Hammond, E.C., Lilienfeld, A.M., Shimkin, M.B., Wynder, E.L., 1959. Smoking and lung cancer: recent evidence and a discussion of some questions. Journal of the National Cancer Institute 22, 173–203.

Diprete, T.A., Gangl, M., 2004. Assessing bias in the estimation of causal effects. Sociological Methodology 34, 271–310.

Dunning, T., 2012. Natural Experiments in the Social Sciences. Cambridge, New York.

Gastwirth, J.L., 1992. Methods for assessing the sensitivity of statistical comparisons used in title VII cases to omitted variables. Jurimetrics 33, 19–34.

Greenhouse, S.W., 1982. Jerome Cornfield's contributions to epidemiology. Biometrics Supplement 38, 33–45.

Hansen, B.B., 2004. Full matching in an observational study of coaching for the SAT. Journal of the American Statistical Association 99, 609–618.

Hansen, B.B., 2007. Optmatch: flexible, optimal matching for observational studies. R News 7, 18–24 (Package optmatch in R).

Heller, R., Rosenbaum, P.R., Small, D.S., 2010. Using the cross-match test to appraise covariate balance in matched pairs. American Statistician 64, 299–309.

Hill, A.B., 1965. Environment and disease: association or causation? Proceedings of the Royal Society of Medicine 58, 295–300.

Lehman, D., Wortman, C., Williams, A., 1987. Long-term effects of losing a spouse or a child in a motor vehicle crash. Journal of Personality and Social Psychology 52, 218–231.

Liu, W., Kuramoto, S.J., Stuart, E.A., 2013. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. Prevention Science 14, 570–580.

Lu, B., Greevy, R., Xu, X., Beck, C., 2011. Optimal nonbipartite matching and its statistical applications. Am. Statist 65, 21–30 (Package nbpmatching in R) Manski, C., 1995. Identification Problems in the Social Sciences. Harvard University Press, Cambridge, MA.

Meyer, B.D., 1995. Natural and quasi-experiments in economics. Journal of Business and Economic Statistics 13, 151–161.

Morton, D.E., Saah, A.J., Silberg, S.L., Owens, W., Roberts, M., Saah, M., 1982. Lead absorption in children of employees in a lead-related industry. American Journal of Epidemiology 115, 549–555.

Piantadosi, S., 2005. Clinical Trials, second ed. Wiley, New York.

Platt, J.R., 1964. Strong inference. Science 146, 347–353.

Rosenbaum, P.R., 1991. Discussing hidden bias in observational studies. Annals of Internal Medicine 115, 901–905.

Rosenbaum, P.R., 2002. Observational Studies, second ed. Springer, New York.

Rosenbaum, P.R., 1999. Choice as an alternative to control in observational studies (with discussion). Statistical Science 14, 259–278.

Rosenbaum, P.R., 2004. Design sensitivity in observational studies. Biometrika 91, 153–164.

Rosenbaum, P.R., 2010. Design of Observational Studies. Springer, New York.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688–701.

Rubin, D.B., 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. Journal of the American Statistical Association 74, 318–328.

Sekhon, J.S., Titiunik, R., 2012. When natural experiments are neither natural nor experiments. American Political Science Review 106, 35–57.

Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and Quasi-experimental Designs for Generalized Causal Inference. Houghton Mifflin, Boston.

Sobel, M.E., 2000. Causal inference in the social sciences. Journal of the American Statistical Association 95, 647–651.

Small, D.S., Rosenbaum, P.R., 2008. War and wages: the strength of instrumental variables and their sensitivity to unmeasured biases. Journal of the American Statistical Association 103, 924–933.

Stuart, E.A., 2010. Matching methods for causal inference. Statistical Science 25, 1–21.

Stuart, E.A., Green, K.M., 2008. Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. Developmental Psychology 44, 395.

Stuart, E.A., Hanna, D.B., 2013. Should epidemiologists be more sensitive to design densitivity? Epidemiology 24, 88–89.

Vandenbroucke, J.P., 2004. When are observational studies as credible as randomized trials? Lancet 363, 1728–1731.

Zubizarreta, J.R., 2012. Using mixed integer programming for matching in an observational study of kidney failure after surgery. Journal of the American Statistical Association 107, 1360–1371.

Zubizarreta, J.R., Cerda, M., Rosenbaum, P.R., 2013. Effect of the 2010 Chilean earthquake on posttraumatic stress: reducing sensitivity to unmeasured bias through study design. Epidemiology 24, 79–87.

Obsessive-Compulsive Disorder

João M Castaldelli-Maia, Medical School, Fundação do ABC, Santo André, São Paulo, Brazil; and Medical School, University of São Paulo, São Paulo, Brazil

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Broocks, F. Hohagen, volume 16, pp. 10815–10822, © 2001, Elsevier Ltd.

Abstract

Obsessive-Compulsive Disorder (OCD) is a chronic, disabling disorder that is characterized by the presence of obsessions and/or compulsions. OCD groups with other related disorders such as body dysmorphic disorder, hoarding disorder, trichotillomania (hair-pulling disorder), excoriation (skin-picking) disorder, and other disorders. However, it is important to note that this group of disorders guard close proximity with anxiety disorders. The prognosis of OCD improved dramatically with the introduction of special cognitive-behavioral techniques and the use of selective serotonin reuptake inhibitors. At present, researchers attempt to integrate psychological, behavioral, neurochemical, and neuroanatomical findings in a comprehensive psychobiological model of OCD.

Obsessive-Compulsive Disorder (OCD) is a chronic, disabling disorder that is characterized by the presence of obsessions and/or compulsions. Obsessions are recurrent and persistent thoughts, urges, or images that are experienced as intrusive and unwanted, whereas compulsions are repetitive behaviors or mental acts that an individual feels driven to perform in response to an obsession or according to rules that must be applied rigidly (APA, 2013). They cause marked distress and considerably interfere with daily functioning. For decades, OCD has been considered to be treatment-refractory. The first major advance was the observation that clomipramine, which differs from other tricyclic antidepressants by its potent inhibition of serotonin reuptake, proved to exert specific antiobsessional effects. The prognosis of OCD improved dramatically with the introduction of special cognitive-behavioral techniques (CBT) and the use of selective serotonin reuptake inhibitors (SSRIs). These specific therapeutic effects and other indices of altered serotonergic function were the background for a serotonergic hypothesis of OCD. The results of new neuroimaging techniques have produced evidence for basal ganglia involvement in the pathogenesis of OCD. In summary, it has been postulated that a disturbed striatal filter function causes a disinhibition of cerebral activities generated in orbitofrontal and cingulate structures. Positron emission tomography (PET) studies have also demonstrated that neurobiological changes associated with OCD are modulated by both pharmacological and psychological treatments. At present, researchers attempt to integrate psychological, behavioral, neurochemical, and neuroanatomical findings in a comprehensive psychobiological model of OCD.

The Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM-5; APA, 2013) proposes the OCD grouping with other related disorders such as body dysmorphic disorder, hoarding disorder, trichotillomania (hair-pulling disorder), excoriation (skin-picking) disorder, substance/medication-induced obsessive–compulsive (OC) and related disorder, OCD and related due to another medical condition, and other OCD and related specified and unspecified OCD. The inclusion of a specific chapter for OC and related disorders in DSM-5 reflects a progressive increase in scientific evidence that relate these disorders to each other with a series of diagnostic validators, and also the clinical utility of grouping these disorders (APA, 2013). However, it is important to note that this group of disorders guard close proximity with anxiety disorders.

Clinical Aspects

Clinical Symptomatology and Course

OCD is a chronic, disabling disorder that is characterized by recurrent obsessions and uncontrolled compulsions. Obsessions are defined as persistent ideas, thoughts, impulses, or images that are experienced as intrusive and inappropriate. Usually patients attempt to ignore or suppress the obsessions, which typically deal with contamination, doubts, or aggressive imaginations. Compulsions are repetitive behaviors (e.g., handwashing, checking locked doors) or mental acts (e.g., counting, repeating words) that the person feels driven to perform in response to an obsession or according to rigid rules. Characteristically, compulsions counteract the anxiety associated with an obsession. Diagnosis by DSM-5 also requires that the obsessions or compulsions are time-consuming (e.g., take more than 1 h per day) or cause clinically significant distress or impairment on social, occupational, or other important areas of functioning (APA, 2013).

Twenty-five percent of the cases of OCD start by 14 years. In the U.S., the mean age at onset is 19.5 years. Onset after 35 years is unusual but does occur. Males have an earlier onset than females (APA, 2013). The disorder tends to exhibit a waxing and waning course over the life span, with periods of relative good functioning and limited symptoms punctuated by periods of symptomatic exacerbation. A small minority of subjects exhibits either complete remission of their disorder or a progressive, deteriorating course. There is a high comorbidity with major depression, panic disorder, phobias, and eating disorders. Approximately half of all patients with Tourette's disorder meet criteria for OCD, although less than 11% of patients with OCD meet criteria for Tourette's disorder.

Differential Diagnosis

OC symptomatology is also found in various other psychiatric disorders. It can be difficult to differentiate OCD from major depression, which is sometimes associated with obsessive thoughts. In contrast to OCD, obsessional thinking in depression usually subsides in parallel to the remission of depression. Psychotic symptoms can also provoke obsessive thoughts and compulsive behaviors, but are accompanied by other characteristics of schizophrenia. Principally, OCD patients differ from schizophrenic patients by their knowledge that obsessions and compulsions are inappropriate, absurd, or irrational ('ego-dystonic'). However, when the condition is very severe the patient might experience his obsessions or compulsions as more ego-syntonic and is transiently not sure about the real risk of danger or the real necessity to perform a certain ritual in order to prevent a negative event. There is a subgroup of OCD patients who are frequently misdiagnosed as schizophrenic due to these 'psychosis-like symptoms.' OCD also displays some similarity to OC personality disorder, which is associated with an obsessive concern for details, perfectionism, and rigidity. However, a genuine syndrome of obsessions and compulsions is not present in OC personality disorder. Patients with OCD experience distress connected with their symptoms. By contrast, the traits of patients with OC personality disorder very often disturb others more than themselves. Finally, OCD is sometimes misdiagnosed in patients suffering from phobias (e.g., phobic avoidance of any assumed risks to be infected by the AIDS virus). However, phobic patients never develop the kind of rituals which are characteristic of OCD. Other disorders that should also be included in the differential diagnosis of OCD are hypochondria, eating disorders, generalized anxiety disorder, panic and agoraphobia disorder, and post-traumatic stress disorder (*see* Differential Diagnosis in Psychiatry).

Epidemiology

The prevalence of OCD in the various countries ranges from 1.1% to 1.8% (APA, 2013). Women are slightly more affected than men, but men are more affected than women during childhood (APA, 2013; *see* Mental Disorders, Epidemiology of).

Pathogenesis

There is strong evidence that neurobiological as well as psychological factors play an important role in the pathogenesis of OCD. Interestingly, OCD occurs across the world and there is a substantial similarity across cultures in the gender distribution, age at onset, and comorbidity of OCD (APA, 2013). Phenomenological similarities between focal dystonias and OCD are related to the persistent, perseverative, repetitive, and involuntary nature of the symptoms. The occurrence of OC symptoms as part of the postencephalitis Parkinsonian syndrome (Schilder, 1938) stimulated the search for a biological link between basal ganglia dysfunction and OC symptomatology. Compulsive behavior and obsessional thinking have been observed after basal ganglia lesions (Laplane et al., 1989; Swedo et al., 1989) and in patients with Gilles de la Tourette's syndrome (Pauls et al., 1986). Accordingly, symptomatic dystonias can be associated with structural brain lesions (Kostic et al., 1996). Patients suffering from spasmodic torticollis (Bihari et al., 1992) or from blepharospasm, conditions which are also related to basal ganglia dysfunction, showed significantly more OC symptoms than normal controls or patients with hemifacial spasms, a condition with similar symptomatology, but caused by an irritation of the facial nerve (Broocks et al., 1998b).

Neuroanatomical Hypothesis of OCD

Since around 1990 PET and other brain imaging studies have produced further evidence for basal ganglia involvement in the pathogenesis of OCD. Functional neuroimaging studies have shown an increase in metabolic activity in the orbitofrontal cortex, anterior cingulate gyrus, and caudate nucleus (Rauch, 2003). After receiving pharmacological and psychotherapy treatment, there was a reduction in the activity at these sites (Rauch, 2003). In addition, all surgical procedures for treatment of OCD refractory patients focus in the surgical disruption of these neural circuits (Lopes et al., 2009).

The pathological increase of metabolic activity in the orbitofrontal cortex (OFC) demonstrated in PET studies that most OCD patients experience the neurobiological expression of the sensations of anxiety and worry. The caudate nucleus and other striatal structures are stimulated by the OFC. Leading researchers in this field have postulated that in OCD a positive feedback mechanism between the OFC and the striatum is activated, which also involves thalamic structures. Physiologically, inhibitory pathways from the caudate nuclei to the thalamus (via the *globus pallidus*) have the task of keeping this neurosensory circuit in balance. One explanation of the proposed disbalance in OCD could be that an increased inhibition of the *globus pallidus* results in a disinhibition of the thalamus creating a self-sustaining circuit (**Figure 1**).

Neuropsychological testing has revealed that patients with OCD are impaired on measures of spatial working memory, spatial recognition, and motor initiation and execution (Purcell et al., 1998; Schmidtke et al., 1998). In contrast, performance of these tasks by patients with panic disorder or depression did not differ from that of controls.

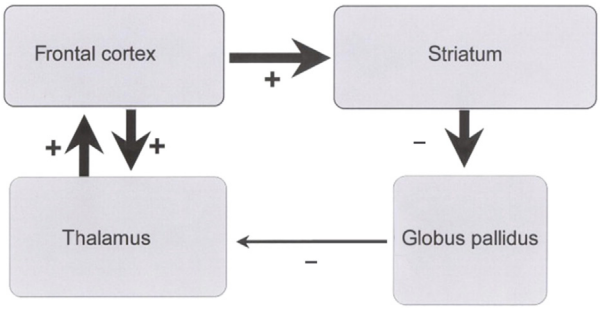

Figure 1 The proposed self-sustaining circuit in OCD.

Neurochemical Hypothesis of OCD

The most compelling findings suggesting a role of serotonergic pathways in the symptomatology of OCD come from well-replicated treatment studies. Most OCD patients improve during treatment with serotonin (5-HT) reuptake inhibitors including clomipramine, fluoxetine, fluvoxamine, paroxetine, and sertraline, but not with non-5-HT selective antidepressants (Murphy et al., 1996; Pigott and Seay, 1999). This observation was replicated in a multicenter trial enrolling 166 patients with concurrent OCD and major depression, in which the SSRI sertraline was more effective in reducing depressive and OC symptoms than the primarily norepinephrine reuptake inhibitor desipramine (Hoehn et al., 2000). Using a placebo-controlled design, discontinuation of clomipramine was followed by an increase of OCD symptoms (Pato et al., 1991). Interestingly, coadministration of the 5-HT2 receptor antagonist metergoline to OCD patients being treated with clomipramine was associated with higher OC symptoms in comparison to placebo coadministration (Benkelfat et al., 1989). An exacerbation of OC symptoms after administration of the nonselective serotonergic agonist, *m*-chlorophenylpiperazine (*m*-CPP) was observed in several, but not all studies (Broocks et al., 1998a; Charney et al., 1988; Hollander et al., 1992; Zohar et al., 1988).

When oral *m*-CPP was readministered during chronic treatment with clomipramine (Zohar et al., 1988) or fluoxetine (Hollander et al., 1992; Pigott et al., 1990), OCD patients did not show an exacerbation of OC symptoms although plasma prolactin and cortisol responses to *m*-CPP were not attenuated. This suggests that chronic clomipramine or fluoxetine treatment in OCD patients may be associated with the development of an adaptive subsensitivity in one or more serotonergic subsystems (Murphy et al., 1996).

Genetic Findings

There is a 3- to 12-fold increased risk in first-degree relatives of patients with OCD (Pauls, 2010). There are some genes that have been linked to OCD as SLC64A and SAPAP3 (Shavitt et al., 2012). The first gene is responsible for producing serotonin transporter. The second, when not functioning properly, leads to OC symptoms in animals and also seems to be related to trichotillomania (Zücher et al., 2010).

Immunological Findings

Recently, new evidence has emerged for possible immunological factors in the pathogenesis of OCD (Teixeira et al., 2014). It has long been known that about 10–30% of patients with rheumatic fever, caused by a streptococcal infection (Skoog and Skoog, 1999), develop Sydenham's chorea. It is believed that OCD happens for a cross-reactivity of antibodies produced against streptococci within the structures of the basal ganglia of the brain that are involved in the pathophysiology of OCD (Skoog and Skoog, 1999).

Treatment

For decades, OCD has been considered as a fascinating but treatment-refractory mental condition. Conventional pharmacological approaches as well as psychodynamic therapy did not lead to satisfactory improvements of OC symptomatology. The first major advances in the treatment of OCD came from two directions. Clomipramine, which differs from other tricyclic antidepressants by its potent inhibition of serotonin reuptake, proved to exert specific antiobsessional effects. At the same time, the introduction of special CBTs began to continuously improve the prognosis of OCD. In the following years, many studies have shown that behavior therapy, especially exposure with response prevention, and SSRIs significantly reduce OC symptoms.

Psychological Treatment

Exposure/response prevention (Williams et al., 2014) has been compared to various control treatments including relaxation, anxiety management training, and pill placebo (Kozak et al., 1999), leading to solid evidence for the efficacy of this procedure. Summarizing all controlled trials, average responder rates were 83% immediately after treatment and 76% after an average follow-up interval of 29 months.

Current exposure/response prevention programs usually combine *in vivo* exercises to feared stimuli with imaginative exposure. Repeated prolonged exposure leads to a psychophysiological habituation process, which also affects irrational beliefs held by the patient. It is crucial that the duration of the exposure exercise continues at least until the patient experiences a decrease of the induced distress. During this time therapists have to be alert to detect and restrain covert neutralizing thoughts (Salkovskis, 1999). If possible, a friend or a family member should be taught to support the patient between sessions in order to achieve adequate exposure. Irrespective of the emotional arousal, the patient has the task to describe precisely the experienced emotions and the accompanying cognitions. This description is the basis of the cognitive restructuring which most therapists view as an essential component of successful exposure.

Cognitive approaches have been proposed for the treatment of OCD, especially obsessional ruminations (Salkovskis, 1999). According to the cognitive hypothesis, obsessional problems occur when intrusive cognitions are interpreted as an indication that the person may be responsible for a certain harmful event, which has happened in the past or will happen in the future. Cognitive therapy is thought to work by altering both strongly held beliefs associated with OCD symptoms and interpretations of intrusive mental experiences (**Figure 2**). It is also important to help patients to identify and to modify underlying general assumptions, which might represent the roots of the more overt dysfunctional thoughts. As in other anxiety disorders, the general style of therapy is that of guided discovery. Further details of the cognitive treatment of obsessions including the use of loop tapes are described in Salkovskis (1999).

Schwartz has proposed a cognitive-biobehavioral self-treatment for OCD (Schwartz, 1996). It integrates the results of neuroimaging research into the cognitive approach based on the literature on the behavioral neurobiology of the basal ganglia. In a four-step process, the OCD patient learns to identify his or her obsessions and compulsions, to label them

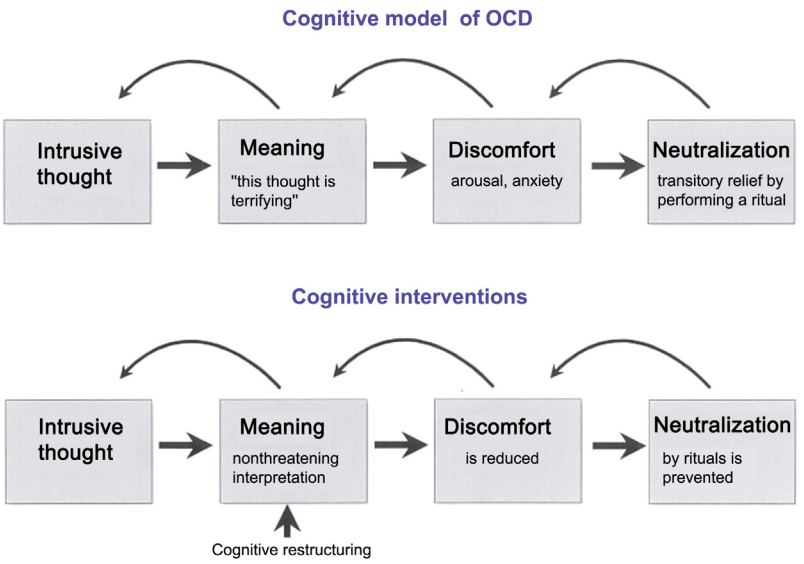

Figure 2 Cognitive therapies.

as such ('It is not me, it is my OCD'), and to attribute them to abnormal activity of the frontostriatal system. While 'working around OCD symptoms,' patients learn new adaptive responses to intrusive OCD thoughts and urges. They are tutored to become increasingly aware of avoidance behaviors and to use treatment techniques to perform previously avoided behaviors (Schwartz, 1996).

In addition to exposure, response prevention and cognitive techniques, the multimodal treatment of OCD also focuses on social deficits, reduced tolerance for intensive emotions, and reduced emotional awareness. In addition, it deals with deficits in problem-solving and self-initiated alternative behaviors as well as interpersonal problems. In severe cases, complex combinations of multimodal interventions are often needed, including marital or family interventions as well as individual or group treatments. Multimodal CBT has been shown to be effective with very low dropout rates in severely ill inpatients (Hohagen et al., 1998) and outpatients (Hand, 1998). Furthermore, its long-term efficacy has been demonstrated in follow-up studies in unselected patients (Hand, 1998). Whether multimodal behavioral therapy is superior to exposure therapy alone and whether it is appropriate to treat refusers, dropouts, and nonresponders to conventional exposure therapy with response prevention, are still open questions (*see* Behavior Therapy: Psychiatric Aspects).

Drug Treatment

In comparison to clomipramine, SSRIs have the advantage of a very low toxicity and a more favorable side effect profile. Consequently, these agents have been the focus of several large-scale, placebo-controlled studies of OCD. Interestingly, very low placebo response rates were observed in most studies. Based on meta-analyses of effect sizes in placebo-controlled trials clomipramine was considered more effective than the SSRIs. However, direct comparisons demonstrated similar efficacy between clomipramine and different SSRIs (Pigott and Seay, 1999). Pharmacological treatment of OCD usually involves relatively high daily dosages (clomipramine up to 250 mg; fluvoxamine up to 300 mg; fluoxetine, paroxetine, and citalopram up to 80 mg; and sertraline up to 200 mg a day), but some investigators believe that longer duration of treatment rather than higher dosage is essential for response. Ten weeks of treatment trial is currently the minimum recommendation, but it is not unusual to have treatment continue for several months before peak efficacy is reached. The algorithm in **Figure 3** can be recommended for the pharmacological treatment of OCD patients.

Combined Pharmacological and Cognitive-Behavioral Treatment

There seems to be superiority of combined treatment (pharmacological and psychotherapeutic) when compared to each treatment alone (Foa et al., 2005; Pediatric OCD Treatment Study, 2004; *see* Psychotherapy and Pharmacotherapy).

Prognosis

The presence of schizotypal, borderline, and avoidant personality disorders predicts a less favorable prognosis, whereas duration of illness, sex, age, and type of symptoms had no clear predictive value. In cognitive-behavioral therapy the best predictor of good outcome at the end of treatment and of follow-up was early compliance in doing exposure homework (De Araujo et al., 1996).

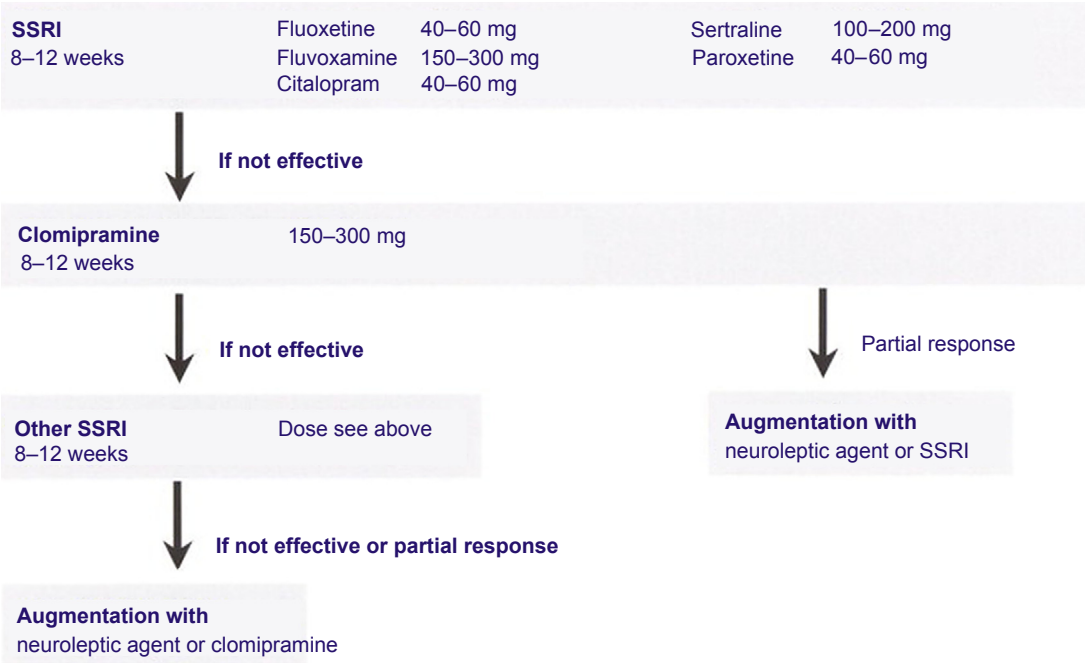

Figure 3 Further procedures in treatment of refractory cases.

Summary and Future Tasks

Despite all progress in treatment, a considerable percentage of patients experience only a partial unsatisfactory response (Dold et al., 2013). At least 25% of OCD patients do not comply with exposure therapy/response prevention (Schwartz, 1998). In addition to dropouts, 20–30% do not respond sufficiently. Since OCD is a chronic disorder, reoccurrence of the OCD symptomatology is likely when patients do not adhere to the newly adopted cognitive-behavioral habits due to motivational problems, acute life stress, or depression. With SSRI treatment alone, 40–60% of patients have not shown significant progress and have to be classified as nonresponders.

Resistant and refractory cases may benefit from an association of an SSRI and an atypical antipsychotic (Bloch et al., 2006), such as risperidone (up to 6 mg), quetiapine (up to 300 mg), or olanzapine (up to 20 mg). In addition, the association between Clomipramine and an SSRI is a good option for these cases (Shavitt et al., 2012). In severe and disabling cases with proven refractoriness, multiple drug interventions and CBT, surgical treatment should be considered (Lopes et al., 2004). The goals of surgical treatment of OCD are some points of the cortico–striatal–thalamic–cortical circuit (Lopes et al., 2004).

See also: Antidepressant Drugs; Anxiety and Anxiety Disorders; Behavior Therapy: Psychiatric Aspects; Child and Adolescent Psychiatry, Principles of; Differential Diagnosis in Psychiatry; Magnetic Resonance and Molecular Imaging in Psychiatry; Mental Illness, Genetics of; Personality Theory and Psychopathology; Post-Traumatic Stress Disorder; Psychotherapy and Pharmacotherapy; Women and Psychiatry.

Bibliography

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. Arlington, VA.

Benkelfat, C., Murphy, D.L., Zohar, J., Hill, J.L., Grover, G., Insel, T.R., 1989. Clomipramine in obsessive-compulsive disorder. Further evidence for a serotonergic mechanism of action. Archives of General Psychiatry 46, 23–28.

Bihari, K., Hill, J.L., Murphy, D.L., 1992. Obsessive-compulsive characteristics in patients with idiopathic spasmodic torticollis. Psychiatry Research 42, 267–272.

Bloch, M.H., Landeros-Weisenberger, A., Kelmendi, B., Coric, V., Bracken, M.B., Leckman, J.F., 2006. A systematic review: antipsychotic augmentation with treatment refractory obsessive-compulsive disorder. Molecular Psychiatry 11 (7), 622–632.

Broocks, A., Pigott, T.A., Hill, J.L., Canter, S., Grady, T.A., L'Heureux, F., Murphy, D.L., 1998a. Acute intravenous administration of ondansetron and m-CPP, alone and in combination, in patients with obsessive-compulsive disorder (OCD): behavioral and biological results. Psychiatry Research 79, 11–20.

Broocks, A., Thiel, A., Angerstein, D., Dressler, D., 1998b. Higher prevalence of obsessive-compulsive symptoms in patients with blepharospasm than in patients with hemifacial spasm. American Journal of Psychiatry 155, 555–557.

Charney, D.S., Goodman, W.K., Price, L.H., Woods, S.W., Rasmussen, S.A., Heninger, G.R., 1988. Serotonin function in obsessive-compulsive disorder. A comparison of the effects of tryptophan and m-chlorophenylpiperazine in patients and healthy subjects. Archives of General Psychiatry 45, 177–185.

De Araujo, L.A., Ito, L.M., Marks, I.M., 1996. Early compliance and other factors predicting outcome of exposure for obsessive-compulsive disorder. British Journal of Psychiatry 169, 747–752.

Dold, M., Aigner, M., Lanzenberger, R., Kasper, S., 2013. Antipsychotic augmentation of serotonin reuptake inhibitors in treatment-resistant obsessive-compulsive disorder: a meta-analysis of double-blind, randomized, placebo-controlled trials. International Journal of Neuropsychopharmacology 16 (3), 557–574.

Foa, E.B., Liebowitz, M.R., Kozak, M.J., Davies, S., Campeas, R., Franklin, M.E., Huppert, J.D., Kjernisted, K., Rowan, V., Schmidt, A.B., Simpson, H.B., Tu, X., 2005. Randomized, placebo-controlled trial of exposure and ritual prevention, clomipramine, and their combination in the treatment of obsessive-compulsive disorder. American Journal of Psychiatry 162 (1), 151–161.

Hand, I., 1998. Out-patient, multi-modal behaviour therapy for obsessive-compulsive disorder. British Journal of Psychiatry 45–52 (Suppl.).

Hoehn, S.R., Ninan, P., Black, D.W., Stahl, S., Greist, J.H., Lydiard, B., McElroy, S., Zajecka, J., Chapman, D., Clary, C., Harrison, W., 2000. Multicenter double-blind comparison of sertraline and desipramine for concurrent obsessive-compulsive and major depressive disorders. Archives of General Psychiatry 57, 76–82.

Hohagen, F., Winkelmann, G., Rasche, R.H., Hand, I., Konig, A., Munchau, N., Hiss, H., Geiger, K.C., Kappler, C., Schramm, P., Rey, E., Aldenhof, J., Berger, M., 1998. Combination of behaviour therapy with fluvoxamine in comparison with behaviour therapy and placebo. Results of a multicentre study. British Journal of Psychiatry 71–78 (Suppl.).

Hollander, E., DeCaria, C.M., Nitescu, A., Gully, R., Suckow, R.F., Cooper, T.B., Gorman, J.M., Klein, D.F., Liebowitz, M.R., 1992. Serotonergic function in obsessive-compulsive disorder. Behavioral and neuroendocrine responses to oral m-chlorophenylpiperazine and fenfluramine in patients and healthy volunteers. Archives of General Psychiatry 49, 21–28.

Kostic, V.S., Stojanovic-Svetel, M., Kacar, A., 1996. Symptomatic dystonias associated with structural brain lesions. Canadian Journal of Neurological Sciences 23, 53–56.

Kozak, M.J., Liebowitz, M.R., Foa, E.B., 1999. Cognitive behavior therapy and pharmacotherapy for OCD. In: Goodman, W.K. (Ed.), Obsessive-Compulsive Disorder: Contemporary Issues in Treatment. Harper, Washington, DC.

Laplane, D., Levasseur, M., Pillon, B., Dubois, B., Baulac, M., Mazoyer, B., Tran Dinh, S., Sette, G., Danze, F., Baron, J.C., 1989. Obsessive-compulsive and other behavioural changes with bilateral basal ganglia lesions. A neuropsychological, magnetic resonance imaging and positron tomography study. Brain 112, 699–725.

Lopes, A.C., de Mathis, M.E., Canteras, M.M., Salvajoli, J.V., Del Porto, J.A., Miguel, E.C., 2004. Update on neurosurgical treatment for obsessive compulsive disorder. Revista Brasileira de Psiquiatria 26 (1), 62–66.

Lopes, A.C., Greenberg, B.D., Norén, G., Canteras, M.M., Busatto, G.F., de Mathis, M.E., Taub, A., D'Alcante, C.C., Hoexter, M.Q., Gouvea, F.S., Cecconi, J.P., Gentil, A.F., Ferrão, Y.A., Fuentes, D., de Castro, C.C., Leite, C.C., Salvajoli, J.V., Duran, F.L., Rasmussen, S., Miguel, E.C., 2009. Treatment of resistant obsessive-compulsive disorder with ventral capsular/ventral striatal gamma capsulotomy: a pilot prospective study. Journal of Neuropsychiatry and Clinical Neurosciences 21 (4), 381–392.

Murphy, D.L., Greenberg, B., Altemus, M., Benjamin, J., Grady, T., Pigott, T., 1996. The neuropharmacology and neurobiology of obsessive-compulsive disorder. In: Westenberg, H.G.M. (Ed.), Advances in the Neurobiology of Anxiety Disorders. John Wiley & Sons Ltd, pp. 279–297.

Pato, M.T., Zohar-Kadouch, R., Zohar, J., Murphy, D.L., 1991. Return of symptoms after discontinuation of clomipramine in patients with obsessive-compulsive disorder. American Journal of Psychiatry 145, 1521–1525.

Pauls, D.L., Leckman, J.F., Towbin, K.E., Zahner, G.E., Cohen, D.J., 1986. A possible genetic relationship exists between Tourette's syndrome and obsessive-compulsive disorder. Psychopharmacology Bulletin 22, 730–733.

Pauls, D.L., 2010. The genetics of obsessive-compulsive disorder: a review. Dialogues in Clinical Neurosciences 12, 149–163.

Pediatric OCD Treatment Study (POTS) Team, 2004. Cognitive-behavior therapy, sertraline, and their combination for children and adolescents with obsessive-compulsive disorder: the Pediatric OCD Treatment Study (POTS) randomized controlled trial. Journal of the American Medical Association 292, 1969–1976.

Pigott, T.A., Pato, M.T., Bernstein, S.E., Grover, G.N., Hill, J.L., Tolliver, T.J., Murphy, D.L., 1990. Controlled comparisons of clomipramine and fluoxetine in the treatment of obsessive-compulsive disorder. Behavioral and biological results. Archives of General Psychiatry 47, 926–932.

Pigott, T.A., Seay, S.M., 1999. A review of the efficacy of selective serotonin reuptake inhibitors in obsessive-compulsive disorder. Journal of Clinical Psychiatry 60, 101–106.

Purcell, R., Maruff, P., Kyrios, M., Pantelis, C., 1998. Neuropsychological deficits in obsessive-compulsive disorder: a comparison with unipolar depression, panic disorder, and normal controls. Archives of General Psychiatry 55, 415–423.

Rauch, S.L., 2003. Neuroimaging and Neurocircuitry models pertaining to the neurosurgical treatment of psychiatric disorders. Neurosurgery Clinics of North America 14, 6–9.

Salkovskis, P.M., 1999. Understanding and treating obsessive-compulsive disorder. Behavioral Research and Therapy S29–52.

Schilder, P., 1938. The organic background of obsessions and compulsions. American Journal of Psychiatry 94, 1397–1414.

Schmidtke, K., Schorb, A., Winkelmann, G., Hohagen, F., 1998. Cognitive frontal lobe dysfunction in obsessive-compulsive disorder. Biological Psychiatry 43, 666–673.

Schwartz, J.M., 1996. Brain Lock. Harper Collins, New York.

Schwartz, J.M., 1998. Neuroanatomical aspects of cognitive-behavioural therapy response in obsessive-compulsive disorder. An evolving perspective on brain and behaviour. British Journal of Psychiatry 38–44 (Suppl.).

Shavitt, R.G., Diniz, J.B., Torres, A.R., Alvarenga, P.G., Cordás, T.A., 2012. Transtornos do espectro compulsivo ao longo da vida. In: Forlenza, O.V., Miguel, E.C. (Eds.), Compêndio de clínica psiquiátrica. Manole, Barueri, SP, Brazil.

Skoog, G., Skoog, I., 1999. A 40-year follow-up of patients with obsessive-compulsive disorder. Archives of General Psychiatry 56, 121–127.

Swedo, S.E., Rapoport, J.L., Cheslow, D.L., Leonard, H.L., Ayoub, E.M., Hosier, D.M., Wald, E.R., 1989. High prevalence of obsessive-compulsive symptoms in patients with Sydenham's chorea. American Journal of Psychiatry 146, 246–249.

Teixeira, A.L., Rodrigues, D.H., Marques, A.H., Miguel, E.C., Fontenelle, L.F., 2014. Searching for the immune basis of obsessive-compulsive disorder. Neuroimmunomodulation 21 (2–3), 152–158.

Williams, M.T., Farris, S.G., Turkheimer, E.N., Franklin, M.E., Simpson, H.B., Liebowitz, M., Foa, E.B., 2014. The impact of symptom dimensions on outcome for exposure and ritual prevention therapy in obsessive-compulsive disorder. Journal of Anxiety Disorders 28 (6), 553–558.

Zohar, J., Insel, T.R., Zohar Kadouch, R.C., Hill, J.L., Murphy, D.L., 1988. Serotonergic responsivity in obsessive-compulsive disorder. Effects of chronic clomipramine treatment. Archives of General Psychiatry 45, 167–172.

Zücher, S., Wendland, J.R., Ashley-Koch, A.E., Collins, A.L., Tran-Viet, K.N., Quinn, K., Timpano, K.C., Cuccaro, M.L., Pericak-Vance, M.A., Steffens, D.C., Krishnan, K.R., Feng, G., Murphy, D.L., 2010. Multiple rare SAPAP3 missense variants in trichotillomania and OCD. Molecular Psychiatry 14, 6–9.

Obsessive-Compulsive Disorder across the Life Span

Norbert Kathmann, Humboldt-Universität zu Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Obsessive-compulsive disorder (OCD) is a disabling disorder with heterogeneous symptoms. Diagnostic and Statistical Manual of Mental Disorders (Fifth edition) categorizes OCD apart from anxiety disorders, together with obsessive-compulsive spectrum disorders. Onset is frequent in childhood and adolescence but can also occur in late life. Several risk factors have been identified, with major genetic fractions. Traumatic or stressful events may play a role. Research identified alterations in brain structure and function, learning processes, action monitoring, feedback processing, and cognitive beliefs to contribute to OCD onset and development. Cognitive-behavioral psychotherapy and pharmacotherapy with serotonin reuptake inhibitors are the first-line treatments in both young and adult patients. More longitudinal research to explore the mechanisms of symptom development and maintenance as well as improved interventions together with efforts to disseminate empirically supported therapies is needed.

Phenomenology and Diagnosis

Symptoms

Obsessive-compulsive disorder (OCD) is a complex disorder that is primarily characterized by the interplay of obsessions and compulsions. Obsessions are unwanted intrusive ideas that suddenly enter consciousness and elicit distress, anxiety-like affect, or disgust. Typical obsessions are thoughts about contamination with germs, excrements, or cleansing agents, which occur in combination with the concern that these materials might infect, poison, and ultimately kill other individuals. Other contents circulate around attacking other people with a knife, hurting someone's baby, or behaving sexually inappropriate. There are many more topics that can turn into obsessive thoughts such as the lack of symmetry and order of objects, the possibility of deficient carefulness while closing a door or turning off a stove, or the possibility of performing or thinking blasphemously or immorally. Importantly, most of these ideas are contrary to the person's values and goals and impose on that person against his/her will. Compulsions are understood as responses to those obsessions that are initiated by the patient in an effort to overcome induced negative feelings, to neutralize aversive thoughts, and, ultimately, to prevent harm. Although patients usually acknowledge the irrationality of this logic, failing to perform the compulsion increases distress and leads to feelings of guilt. Examples for frequent compulsions are excessive hand washing; controlling switches and faucets for several minutes or even hours, thereby using a rigid ritual sometimes; thinking good thoughts to prevent bad events; repeating routine activities until one feels just right; ordering objects in a certain manner; counting until a specific number; or asking the same question until one feels certain. Patients feel more impaired by the obsessions whereas overt compulsions manifest as the leading symptom of OCD observable by other people. Only a minority of patients shows either obsessions or compulsions. In those cases, it is unclear whether obsessions cannot be reported or compulsions are too subtle to recognize them.

To achieve a more systematic description, attempts have been made to integrate the heterogeneous symptoms of OCD into subtypes and dimensions. The method most widely used to generate a dimensional system is factor analysis of identified symptoms. The Yale-Brown Obsessive Compulsive Scale Symptom Checklist (Y-BOCS-SC) measures more than 60 symptoms as present or not, either currently or throughout the whole lifetime. Thirteen symptom categories are determined by collapsing single items from the list on a rational basis. These categories were factor-analyzed in several studies converging in four-factor solutions. The factors were termed (1) 'taboos and forbidden thoughts' (aggressive, religious, sexual, and somatic obsessions, and checking behavior), (2) 'symmetry' (symmetry obsessions and repeating, ordering, and counting compulsions), (3) 'cleaning' (contamination obsessions and cleaning compulsions), and (4) 'hoarding' (hoarding obsessions and compulsions). An item-level factor analysis in a large sample showed the best fit for a five-factor solution consisting of (1) taboo, (2) contamination/cleaning, (3) doubts, (4) superstitions/rituals, and (5) symmetry/hoarding (Katerberg et al., 2010). Due to remaining inconsistencies, no consensus about the correct number and characterization of the symptom factors has been achieved so far. A limitation of the factor-analytic studies is that all of them are based on the Y-BOCS-SC, which was not constructed as a tool for measurement of the dimensional structure of OCD. New diagnostic instruments have been developed but require further validation before they can be used broadly. Nevertheless, dimensional diagnostics in OCD hold promise for future research because it allows for parsimonious but comprehensive description of cases, for covering heterogeneity of psychopathology, and, potentially, for refined etiological concepts.

Beyond the dimensional approach, several categorical subtypes for OCD have been suggested. *Early onset* (EO) in OCD is referred to the beginnings of the disease in childhood or adolescence. More than 50% of adults diagnosed with OCD are estimated to have EO. Latent class analysis has validated two onset groups (Taylor, 2011a). Patients with EO had a mean age of 11 years when OCD or OC symptoms were diagnosed the first time. They are more likely to be male, and report a higher global severity of symptoms in adulthood. Repeating compulsions, religious obsessions, symmetry obsessions,

checking compulsions, sexual obsessions, hoarding compulsions, and aggressive obsessions are more prevalent in adult OCD patients with EO compared to patients with late onset. Patients with EO are also exposed to a more than twofold risk of having comorbid tics. First-degree relatives of OCD patients with EO show a higher prevalence of OCD or OC symptoms than relatives of late-onset patients. These differences support the validity of the age of onset distinction. A second criterion used to categorize OCD patients in subgroups is *comorbidity with tics*. Tic-related OCD affects around one-fourth of the EO cases. Those cases are predominantly male and show more symmetry and ordering symptoms as well as taboo obsessions than other cases. Further, they report enhanced incidence of specific symptoms, like sensory phenomena, feelings of incompleteness, and higher rates of comorbidity with attention deficit hyperactivity disorder (ADHD), including disruptive behaviors. Recurrence risk for OCD is higher in first-degree relatives of tic-related OCD patients than in pure OCD cases. Response to cognitive-behavioral interventions, however, is not different for the tic-related subgroup. Although there is some overlap of the two subtyping schemes (EO and tic-related), a substantial proportion of EO patients has no tic symptoms. Therefore, and on the basis of the evidence reported above, both onset and tic comorbidity may be of value for refined etiological research and treatment.

Diagnosis

In the *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition (DSM-IV) as well as in the International Classification of Diseases, tenth revision (ICD-10), the presence of either obsessions or compulsions or both are core criteria for a diagnosis. ICD-10 requires symptoms to be present over a period of at least 2 weeks whereas DSM-IV defines that symptoms must be time consuming (e.g., taking more than 1 h per day). Also, according to DSM-IV and ICD-10, the person must recognize at some point that obsessions and/or compulsions are excessive or unreasonable. In 2013, the DSM-5 was introduced, including a new chapter on OCD and related disorders. This is a significant change because OCD is no longer classified as an anxiety disorder. Besides body dysmorphic disorder, hair pulling disorder, and skin picking disorder, hoarding disorder is now a diagnosis within the 'related disorders' spectrum (formerly a symptom category or a dimension of OCD). The DSM-IV specifier 'with poor insight' was changed in the DSM-5, and is now allowing for several degrees on a spectrum of insight: good/poor/absent or delusional OCD beliefs. Consequently, patients with delusional OCD symptoms can be diagnosed as OCD and do not warrant a diagnosis of psychosis or schizophrenia. Finally, a new specifier has been introduced for tic-related OCD.

Life Span Aspects

It is important to note that repetitive and ritualistic activities are frequent among children during some phases of development. Bedtime rituals, but also activities that are hardly to distinguish from compulsive symptoms like arranging objects, may occur in early childhood peaking around 2–3 years of age. Most of these behaviors disappear before school age, but little is known on the mechanisms underlying these behaviors and on possible links to OCD. In general, the phenomenology and the range of symptoms of OCD in childhood do not differ from those of adult patients. However, some specifics have been reported in small epidemiological samples. Content of harm and aggression obsessions often involved family members, and symmetry/ordering symptoms were more frequent in children, whereas contamination/cleaning symptoms appear to be less prevalent in children. Generally, insight with respect to the irrationality of the obsessions is low in young children. Therefore, the DSM-IV has omitted insight from the criteria list for an OCD diagnosis in children. As mentioned above, comorbidity with tics and ADHD is enhanced, and more boys than girls are diagnosed with OCD in childhood.

A critical phase for female individuals prone to develop OCD is pregnancy and postpartum. Based on a meta-analysis, prevalence in pregnant women is 2%, which is twice the prevalence of women in the general population. Postpartum women are suffering from OCD even more frequently than pregnant women (Russell et al., 2013). The clinical manifestation in these cases is dominated by obsessive thoughts with aggressive contents (e.g., to hurt or kill the baby) and/or contamination fears (e.g., that dirt or germs might be transmitted and may harm the baby). Excessive avoidance, compulsive cleaning, and excessively checking on the baby are observable compulsions, aiming to ensure life and health of the baby. In the case of clinically significant symptoms, OCD can interfere with the provision of adequate care and the development of bonding. However, the majority of new parents are experiencing some transient concerns of harming their newborn, and such fears are also part of postpartum depression. Thus, diagnosing OCD in the context of pregnancy or postpartum period requires careful assessment of the degree of functional impairment and distress.

OCD symptoms appear to decline in older adults, and patients of older age are less frequently seen in psychiatric and psychotherapeutic care. However, a recent epidemiological study in population aged 70 and above showed that the condition also represents a significant burden in older age (Klenfeldt et al., 2014). This may be partly due to a chronic course of OCD with onset earlier in lifetime but also to incidence in older adulthood. Cases with an onset beyond age 60 were also found. The doubting type of obsessions is frequently present in those patients, and compulsions typically involve checking but not cleaning. Generally, there is a need for more research on OCD in older age because data on this topic is scarce to date.

Epidemiology

OCD has been presumed to be a rare disorder until epidemiological field studies revealed a different picture. According to a recent study conducted in the USA, OCD has an estimated lifetime prevalence of 2.3% and a 12-month prevalence of 1.2% (Ruscio et al., 2010). Subthreshold OC symptoms were found in nearly 30% of the community. The majority of OCD cases met the lifetime criteria for one or multiple comorbid disorders. Anxiety disorders and mood disorders are the most prevalent of these comorbid conditions. Onset of mood

disorders precede or follow the onset of OCD with almost equal probability, whereas comorbid anxiety disorders like specific and social phobia mostly emerge before OCD onset. Other studies around the world using larger community samples revealed DSM-IV-based lifetime prevalence estimations between 0.5 and 3.8%. It is unclear whether regional and cultural differences in true prevalence or methodological aspects like definition of diagnostic threshold or the diagnostic instrument administered account for this variance.

Comorbid personality disorders (PD) are also found at high rates among OCD patients. A meta-analysis showed that OCD patients have the highest comorbidity with any PD (52%) among all anxiety disorders (Friborg et al., 2013). OCPD (20%) followed by avoidant PD are the most frequent comorbid PDs.

Although gender differences in prevalence is generally lower in OCD than in most anxiety disorders and affective disorders, there is a preponderance of female patients across studies. However, age substantially moderates gender differences. Patients with very early onset (<10 years) are predominantly male, females have more incidence in adolescence, and older age patients are mostly female.

Epidemiological studies specifically addressing childhood and adolescence populations report that OCD is rare (<0.1%) below age 13, but prevalence increases exponentially in adolescence (>0.5). Prevalence estimations for adolescents aged 14–18 do not substantially differ from those in adult populations in most studies. A recent population-based study found an even higher point prevalence of 3.3%. After age 30, prevalence appears to decrease monotonically. However, according to recent evidence, there may be a higher than expected prevalence in populations aged 70 and above.

Disability and Quality of Life

According to the World Health Organization (WHO), OCD ranges among the 10 most disabling diseases, being the fifth most disabling disease among women aged 15–44. Although many patients are successful in deceiving their fellow human beings about the real nature of their problems for a long time, about two-thirds of patients have severe disabilities and role impairments, mainly in the domains of work/school, relationships, and social functioning (Ruscio et al., 2010). Disability scales measure the degree to which OCD symptoms disrupt functions like managing work, school and household, and establishing partnership and social life. The degree of impairment depends not only on symptom severity but also on the number of comorbid conditions. Quality of life (QoL) is defined by the WHO as the subjective perception of one's position in life in the context of the personal goals and the value systems, the individual is living with. In addition, it represents satisfaction with health, and with occupational, social, and other activities. Patients suffering from OCD have seriously impaired QoL as reported in several studies. The deficits were more pronounced than those in panic disorder, social phobia, and kidney transplant patients, and at a similar level as in schizophrenia and depressive disorders (Macy et al., 2013). Improvements in role functioning and QoL are increasingly recognized as the ultimate goals for interventions in addition to symptom reduction. When measured as an outcome variable, impairments have been found to be reduced both after cognitive-behavioral psychotherapy (CBT) and pharmacotherapy, though remaining above the level in the general population. Less research has been conducted on QoL in pediatric and in elderly OCD patients. Notably, there seem to be serious impairments due to OCD in these age groups as well.

Risk Factors, Etiological Models, and Pathogenetic Mechanisms

Stressful Life Events and Traumatization

Several psychological factors that qualify to be etiologically relevant have been confirmed by empirical studies. Stressful life events are a general risk for the incidence of mental disorders and are also associated with OCD. They are assumed to trigger disease onset or to interact with biological factors and, thereby, contribute to the development of the disorder. Children with OCD were reported to have experienced more stressful events within 1 year before diagnosis than healthy children. Also, risk factors like pregnancy and the postpartum period typically implicate stressful events. In addition to everyday stress, traumatic events have been implicated in the etiology of OCD. Retrospectively, lifetime traumatic events were found in more than 50% of adult OCD patients. Present symptom severity, in particular checking and ordering, correlates with traumatic events in the past. Sexual as well as physical abuse before the age of 11, and a number of residence changes are prospectively associated with OCD. Furthermore, loss of a parent predicted checking symptoms (Grisham et al., 2011). However, some studies did not find differences in severe traumatization between OCD patients and controls. Overall, traumatization and comorbid trauma-related symptoms might be more relevant in severe and treatment-resistant cases. Parental rearing style is another candidate risk factor for OCD. In some studies, less paternal or parental warmth and more paternal or parental rejection and control were reported. However, diagnostic specificity is unlikely, and the retrospective nature of the studies limits their validity. A recent meta-analysis on twin studies assessing OC symptoms revealed that only a small proportion of variance (6%) is accounted for by shared environmental factors while more than 50% are related to nonshared environment (Taylor, 2011b). The latter proportion increases with age.

Genetics

Genetic factors contribute to mental disorders, and this applies for OCD as well. In fact, meta-analytic estimates of the variance explained by additive genetic factors amount to 38% (Taylor, 2011b). Family studies support the putative role of genetic effects for transmission of OCD. The risk to develop OCD for an individual with a family member already presenting with OCD is about four times higher than in the general population. Relatives of individuals with early age of onset have an even higher risk. Despite such consistent results in population genetic studies, there is little clarity about specific gene variants contributing to this risk so far. Many associations of OCD and functional candidate genes have been detected but successful

replications are rare. In a meta-analysis including 20 polymorphisms analyzed in at least five studies, variants of genes related to the serotonergic system (5-HTTLPR and HTR2A) and to catecholamine modulation (COMT and MAOA), with the latter in males only, were confirmed to be associated with OCD (Taylor, 2013). Less consistent were associations with dopamine-related polymorphisms (DAT1 and DRD3) and with the glutamate system (SLC1A1). Genome-wide association studies as well as further meta-analyses were not able to clarify the exact role of specific genes for the etiology of OCD, but the glutamate transporter gene appears to be involved. Even larger samples will be necessary to identify critical gene variants. Based on the present results, it can be concluded that many genes with small effects are likely to contribute to the genetic risk for OCD.

While a pure genetic explanation of OCD is incompatible with current findings, the fact that about 40% of the phenotypic variance in OC symptoms is accounted for by genetic variance clearly calls for incorporation of genetic factors in a comprehensive model of OCD. There is growing insight that future research must have a focus on gene-by-environment interactions, i.e., the effects of negative life events might come into play only or to a stronger degree in the presence of certain gene variants. One of the first studies of this type showed that emotional abuse and neglect during childhood increased the risk of having OCD, and this effect was enhanced in individuals with the BDNF Met-allele (Hemmings et al., 2013).

Immunological Factors

Infection with group A beta-hemolytic streptococcal bacteria has been suggested as a risk factor for OCD and tic disorder, since there are a number of childhood OCD cases with sudden symptom exacerbation after such an infection. The inflammation produces an autoimmune response involving antibasal ganglia antibodies, which are implicated in Sydenham chorea, a transient movement disorder frequently associated with OC symptoms, but OCD symptoms may also occur without movement disorder. The term PANDAS (pediatric autoimmune neuropsychiatric disorders associated with streptococcal infection) was introduced to indicate such cases. However, there is still some controversy about whether the presence of the antibodies is pathogenetically relevant for OCD or rather an epiphenomenon (da Rocha et al., 2008). Possibly, PANDAS concerns a relatively small subgroup of OCD patients with a different etiology and phenomenology. Antibiotic and immunomodulatory therapies have been proposed for those patients but require further evaluation.

Other Biological Risk Factors

Preterm birth and maternal health problems during pregnancy have been occasionally reported to be more prevalent among OCD cases. Maternal age at birth was associated with increased risk to develop OCD in a large-scale study, but it is known that this association also holds for other mental disorders. Based on a longitudinal study, perinatal complications were linked to obsessive shameful thoughts but not to OCD incidence in general (Grisham et al., 2011).

Pathogenetic and Etiological Models of OCD

Psychological Models

The first pure psychological model to account for OCD was developed by Sigmund Freud and later refined by psychodynamic theorists. Briefly, according to psychoanalytic theory, obsessional thoughts and compulsive acts are the results of a conflict between sexual and aggressive impulses, and the repression of those impulses. Due to fixation in the so-called anal phase of psychosexual development, the ego uses specific defense mechanisms to reconcile the forbidden impulses, thereby forming the symptoms as a compromise. However, only little empirical research was devoted to support this model. Therefore, it plays a minor role in modern concepts of understanding and treating OCD.

Behavioral models have received much more empirical attention. The most influential one is the two-factor model of fear development and maintenance by Mowrer (1960). It assumes that classical conditioning leads to the acquisition of fearful thoughts, and operant conditioning involving fear-neutralizing rituals makes the individual resistant to extinction. The latter component represents the target of the dominating intervention technique for OCD, namely exposure and response prevention (ERP). However, although a fraction of OCD patients actually report traumatic events before the onset of symptoms, empirical evidence for associative learning as a major mechanism for pathogenesis is limited. Extinction actually appears to be compromised both by excessive use of avoidant behaviors and putatively by a neurobiologically based impairment in extinction recall.

Cognitive models assume that symptoms may arise from dysfunctional cognitive contents and beliefs. They stimulated plenty of empirical studies, and are broadly accepted by practitioners, in particular by those applying CBT. In the case of OCD, inflated personal responsibility beliefs and the idea of thought–action fusion are presumed to induce negative appraisals of intrusive thoughts, and consequently negative affects like shame, guilt, and anxiety will arise. Compulsive rituals are adopted as a means to prevent both negative feelings and feared consequences. In fact, the rituals reinforce the negative thoughts and keep them activated. Finally, intrusions as well as negative feelings are kept alive (Salkovskis, 1985). Contemporary cognitive models have integrated additional OCD-typical beliefs like the overestimation of threat, the need to control thoughts, the intolerance of uncertainty, and perfectionism (OCCWG, 1997). Research has shown that (1) these beliefs are correlated with symptom expression, (2) experimental manipulations of the beliefs affect symptoms, and (3) symptom relief after cognitive-behavioral treatment also weakens OCD-typical beliefs. Factor-analytic studies confirmed three correlated belief domains, namely (1) perfectionism and intolerance of uncertainty, (2) overimportance of thoughts, and (3) inflated responsibility and overestimation of threat, together with a general factor explaining the largest amount of variance. It should be noted, however, that OCD-related dysfunctional beliefs are present only in a minority of patients and, therefore, are neither necessary nor sufficient conditions for OCD development. In sum, contemporary cognitive models consider an important part of psychological alterations implicated in OCD, but they

probably overestimate the contribution of cognition to symptom generation. Also, the origins of the biased beliefs remain open up to now.

Brain Circuitry

Progress in brain imaging techniques and findings in brain-lesioned patients have fostered concepts of OCD as a brain disorder with typical anatomical, functional, and connectivity alterations. Descriptions of patients with circumscribed lesions in the basal ganglia or neurological conditions like Huntington disease or Sydenham chorea revealed enhanced prevalences of OC symptoms. This observation gave rise to the conjecture that basal ganglia structures are the brain loci where OCD is mediated or even caused. Since the basal ganglia are part of brain circuits also encompassing the thalamus and subregions of the prefrontal cortex, the neuroanatomical model was extended to the cortico-striato-thalamo-cortical (CSTC) hypothesis of OCD generation (Saxena et al., 1998). In brief, imbalance of a direct and an indirect path within the circuit is assumed to set the course for enhanced activity in the orbitofrontal cortex, the ventral striatum, and the medio-dorsal thalamus. A wealth of brain imaging studies aimed at examining this hypothesis and its extensions. Results moderately support the model (de Wit et al., 2014) showing volume reductions in gray and white matter of the dorso-medial prefrontal cortex, the anterior cingulate cortex, the inferior frontal gyrus, and the anterior insula. Interestingly, altered neuroplasticity with increasing age of patients was detected in this large-scale analysis, with orbitofrontal-striatal regions being more preserved in higher-aged OCD patients, and a relative loss of limbic and paralimbic regions with increasing age. Functional brain imaging studies addressing hemodynamic patterns were also not fully consistent. A voxel-level meta-analysis of functional magnetic resonance imaging findings identified alterations in orbitofronto-striatal, lateral frontal, anterior cingulate, middle occipital and parietal cortices, and the cerebellum (Menzies et al., 2008). This highlights the need to implicate regions outside the classical CSTC areas to develop a comprehensive brain model of OCD. Recently, detection of abnormal interregional connectivity of CSTC regions as well as of cortical hubs, i.e., regions displaying a high degree of connectivity has gained more attention. OCD patients free of medication were reported to have greater distant connectivity in the orbitofrontal cortex and subthalamic nucleus, and greater local connectivity in the orbitofrontal cortex and the putamen (Beucke et al., 2013). The methods to analyze connectivity are under further development, but certainly promise more detailed insights into altered brain dynamics underlying psychopathological conditions. Thus far, studies were cross-sectional, and do not allow deciding whether the brain abnormalities are mere correlates or may qualify as vulnerability markers for OCD and other mental disorders.

Neuropsychological Deficits

As a consequence of brain abnormalities, impairments in cognitive and affective functions mediated by these altered brain areas can be expected. Those impairments are then assumed to contribute to the emergence of OCD symptoms. This neuropsychological deficit model has been widely studied. In a meta-analysis on 115 cross-sectional investigations, small to medium-size performance reductions of OCD patients compared to healthy control persons were found. Domains with effect sizes above 0.5 were nonverbal memory, set shifting/cognitive flexibility, and processing speed. Slightly smaller deficits were detected in response to inhibition and sustained attention (Abramovitch et al., 2013). The small effects suggest that moderators are likely, such as symptom expression on specific dimensions, which have been infrequently assessed in previous studies. Support for the role of neuropsychological deficits as vulnerability markers was provided by a longitudinal study showing that deficits in visuospatial, visuoconstructive, and visuomotor skills at age 13 were identified in individuals who had developed OCD at age 32.

Cognitive-Affective Processing

Several putative cognitive and affective mechanisms of OCD symptom development have been investigated using experimental approaches. For example, extinction learning is the core mechanism assumed to underlie the effects of behavior therapy (see Classical Conditioning Methods in Psychotherapy). Deficits in extinction learning, reflected by reduced activation of the ventromedial prefrontal cortex during extinction learning and skin conductance responses indicative of impaired extinction recall, have been demonstrated in OCD patients (Milad et al., 2013). Another promising area of research is processing of feedback. OCD patients are more sensitive to negative feedback in association learning tasks, and show inhibitory deficits after being punished (Endrass et al., 2013).

Endophenotypes

The search for the so-called endophenotypes or intermediate phenotypes is motivated by the goal to find the missing links between the clinical phenotype and its distal genetic or environmental causes. Endophenotypes are understood as more proximal products of the causes (e.g., genes) compared to clinical diagnoses. While diagnoses involve heterogeneous syndromes and have only limited biological validity, endophenotypes are thought to be biologically more homogeneous and closer to their causal underpinnings. Therefore, they might also help to successfully discover relevant susceptibility gene variants or environmental risk factors for disorders. They are also hoped to elucidate mechanisms of symptom generation and to identify individuals at risk for a disorder. Endophenotypes can be delineated at different stages of the pathogenetic pathway, including genetic, molecular, neurochemical, brain anatomical, large-scale neurofunctional, neuropsychological, information processing, cognitive-representational, and affective levels. To qualify as an endophenotype, a variable must be under genetic control and needs to be associated with the disorder or its symptoms. In addition, unaffected family members of a case proband must show higher rates of the endophenotype. Several candidates have been proposed as endophenotypes of OCD.

Probably, the most elaborate one is an EEG marker of error processing, termed the error-related negativity (ERN)

(Manoach and Agam, 2013). It is replicably enhanced in OCD patients, and twin studies confirmed heritability. In addition, its source lies in the anterior cingulate cortex, a region structurally and functionally altered in OCD; the amplitude is independent of symptom severity changes, and can be reliably measured; preliminary association studies revealed altered genes for the dopaminergic and the serotonergic system; and, importantly, in a family study, unaffected relatives of OCD cases showed enhanced ERN amplitudes (Riesel et al., 2011).

Other candidate endophenotypes are platelet serotonergic markers of central serotonergic functioning, white matter abnormalities in frontal brain regions, functional brain activation patterns during performance of reversal learning and response inhibition tasks, and performance deficits in motor inhibition and in set shifting tasks. These measures indicate executive function capacities that are relevant in the context of the CSTC model (Mataix-Cols and van den Heuvel, 2012). However, most studies were conducted in small samples, and replication of the central findings is necessary.

Need for Integrative Models

Obviously, most models take a selective view on the relevant factors underlying OCD. This is understandable in terms of feasibility issues, both for research and practical transfer. However, only relatively small amounts of variance can be explained by the factors within each model. Therefore, models taking a more integrative view are mandatory for further progress in the field. Using structural equation modeling of data from a twin sample, Taylor and Jang (2011) found evidence for a model, which assumed that genetic and environmental factors influence beliefs and OC symptoms and that beliefs also influence symptoms. Such biopsychosocial models hold promise to reconcile data from different research domains, will set the stage for a more realistic and comprehensive understanding of OCD, and may ultimately improve treatment of the disorder. Multilevel approaches are also needed in endophenotype research when trying to model the complex network of causes, influences, and moderators of processes involved in the generation of OC symptoms. A transdiagnostic, dimensional perspective, which detaches itself from fixed diagnostic categories may be helpful in this process.

Treatment

Natural Course and Placebo Response

OCD is a relatively chronic and persistent disorder, but waxing and waning of symptoms can be observed. Full or partial remission over at least eight consecutive weeks is achieved in only one-fourth of cases within an observation period of 2 years according to a prospective study conducted in a treatment-seeking sample with long illness duration. In a long-term follow-up study over 40 years, 20% of patients experienced full, and further 28% partial remission. That study was initiated before the advent of effective treatments (Skoog and Skoog, 1999). Evaluation trials of selective serotonin reuptake inhibitors (SSRIs) and clomipramine have used pill placebo to control for short-term symptom change due to treatment expectation and general attention effects. Placebo effects in terms of percent change of symptom severity over treatment periods of 4–12 weeks were in the 5–10% range on average, which is clearly less than in anxiety and depressive disorders. In conclusion, OCD is a disorder, which is quite resistant to spontaneous remission.

Empirically Supported Treatments

Cognitive-Behavioral Psychotherapy (CBT)

CBT for OCD is based on behavioral and cognitive models of this disorder (see above). ERP and cognitive techniques are the core elements of CBT. ERP teaches patients to expose themselves to feared stimuli and thoughts, and also to prevent passive and active avoidance strategies. Repeated exposures lead to habituation of anxiety and distress, and patients make corrective experiences regarding the actual degree of dangerousness of situations previously associated with aversiveness. Cognitive techniques promote modification of biased and dysfunctional beliefs by discussing, testing, and modifying them. The great majority of studies on the efficacy and effectiveness of psychological treatments have referred to this type of intervention. More than 15 randomized controlled trials (RCTs) have been conducted, and meta-analyses (e.g., Olatunji et al., 2013) provide strong evidence for reliable and high effect sizes (in the range of 1.0–1.5) for CBT compared to placebo, wait list, or active control treatments. Treatment effects appear to be stable in the majority of cases up to several years but sufficient follow-up data is currently lacking. CBT in OCD appears to show better efficacy compared to other anxiety disorders and depressive disorder. Nevertheless, about 40–60% of patients do not respond sufficiently, due to premature dropout, noncompliance with exposure work within and between sessions, or to residual symptoms and impairments despite adhering to the therapy protocol. In addition, a substantial proportion of patients are not willing to enter CBT due to its demanding and sometimes stress-evoking nature. Patients' adherence to ERP is one of the best predictors of outcome. Consequently, experts have proposed incorporation of specific interventions focusing on motivation and adherence in CBT programs. Preliminary studies suggest that motivational interventions improve outcome and reduce dropout rates. CBT is also successful in children and adolescents with similar effect sizes as in adults. Treatment techniques in young OCD patients are basically the same as in adults, but procedures have to be adapted to the developmental level. Both in youth and in adult OCD, effectiveness studies have confirmed the high usefulness of CBT, also in unselected patient samples and in routine clinical practice outside academic research centers. The format of interventions appears to be less important according to currently available data. CBT in groups has been shown to be similarly effective as in the individual setting, and both intense (daily sessions) and more widely spaced (weekly) formats can be effectively used (Franklin and Foa, 2011). Recently, the so-called third-wave psychotherapies like acceptance and commitment therapy and mindfulness-based cognitive therapy have been proposed as complementing treatment strategies. Preliminary evaluations revealed positive effects, but more controlled trials are necessary. Training protocols such as cognitive bias modification and meta-cognitive therapy have been developed and show some success in changing typical

interpretation biases of nonclinical individuals with OC symptoms. Attention training using a modified probe detection task designed to direct attention to the location of contamination-related words resulted in improved approach to feared objects. Further adaptation of these interventions to clinical needs is required.

Pharmacotherapy

Clomipramine, a nonselective serotonin reuptake inhibitor, was the first drug shown to be efficacious in reducing symptoms of OCD in controlled trials. Later, SSRIs fluvoxamine, fluoxetine, sertraline, paroxetine, and citalopram have been established as standard pharmacotherapy, mainly due to their slightly more favorable side effect profile. The effect size in RCTs amounts to 0.9 on average in adults, which is somewhat lower than that of CBT. In adolescents, treatment response is less impressive, with an average effect size of about 0.5. The most relevant disadvantage of pharmacotherapy with clomipramine or SSRIs is the high relapse rate after discontinuation of the medication, which necessitates maintenance of treatment or very slow tapering of the drugs. Pharmacological augmentation strategies aimed at addressing partial or even complete nonresponse to SSRIs or clomipramine have been proposed and evaluated. Augmentation with atypical neuroleptics, in particular risperidone, yielded more additional responders than adding a placebo but still only one-third of partial or nonresponders will sufficiently improve then. Potential adverse side effects have to be taken into consideration and may increase the risk to benefit ratio. Patients with comorbid tics or with schizotypal personality characteristics have been suggested to be the target population, which is most eligible for this augmentation strategy. There is no sufficient evidence for the efficacy of other pharmacological agents and for the use of second-generation antipsychotics as monotherapy. Antiglutamatergic agents like riluzole, memantine, or ketamine are in the early phase of evaluation and require further study (Fineberg et al., 2012).

Combination of Psychological and Pharmacological Treatments

While CBT can be considered as the first-line treatment of OCD in terms of its superior efficacy and safety, combined treatment with CBT and SSRIs or clomipramine is often recommended. There is little evidence for an additive effect of medication in excess of the CBT effect in OCD patients presenting with low comorbidity. However, highly depressed OCD patients might need antidepressant medication before being able to fully benefit from CBT, and pediatric OCD patients with ADHD comorbidity will also have better results when comorbid symptoms are targeted with adequate pharmacotherapy. Pharmacotherapy as a single treatment, conversely, is not fully adequate in the majority of cases since efficacy is limited, and relapse rates are high after discontinuation. CBT should be added if available and accepted by the patient. Thereby, relapse rates can be lowered and more remissions can be achieved (Franklin and Foa, 2011).

A pharmacological augmentation strategy with a different approach is the administration of D-cyloserine, a partial glutamatergic N-methyl-D-aspartate agonist, before each ERP session to enhance learning processes. D-cycloserine, originally developed as an antibiotic effective against *Mycobacterium tuberculosis*, has no psychotropic effect on its own, and is also not known to elicit adverse effects with the dosages used. Preliminary studies in small samples have been promising but further research is needed before implementation of this strategy in routine care.

Brain Stimulation and Psychosurgery

The efficacy of repetitive transcranial magnetic stimulation in OCD has been evaluated in several pilot trials with mixed success. Improvement of the protocols might result in more reliable effects.

Deep brain stimulation (DBS) of the nucleus accumbens via implanted electrodes has been developed as a last resort intervention for treatment-refractory patients with severe symptoms and impairments. The technique is invasive and therefore controversial, but complications and adverse events are relatively rare. Preliminary results from controlled studies are promising (Denys et al., 2010), showing about 50% improvement in symptom severity, and a 25% difference between active and sham stimulation. To make DBS useful for clinical routine, more evaluations of short- and long-term effects are needed. DBS obtains its convincibility from directly targeting brain structures assumed to be involved in the development of OC symptoms.

Care and Guidelines

Although efficacious therapies exist, many patients cannot benefit from them due to their limited availability. CBT, for example, is not provided in many regions, even not in highly developed industrial countries. If available, waiting time for beginning a CBT program may be several months. Sometimes, training and supervision of therapists are not optimal resulting in procedural shortcomings of provided CBT. For example, ERP is not conducted in many therapies described as cognitive-behavioral. This example makes clear why it is important to provide knowledge updates regarding empirically supported therapies to health care providers. To foster dissemination of this knowledge, practice guidelines have been developed (e.g., NICE, 2005). They give recommendations for therapy options that are graded according to their level of evidence. Global statements as well as such addressing specific subpopulations and problems are traceably derived from empirical studies. Nevertheless, the development of practice guidelines must be complemented by serious efforts to bring the guidelines to practitioners and to attain their acceptance.

Outlook

Theory and treatment of OCD have dramatically changed over the last decades, and the disorder is in the focus of contemporary research. Both the subjective and the socioeconomic burden of OCD have been acknowledged. We have efficacious psychological and pharmacological interventions at hand, but there is room for improvement in terms of success and dropout rates. To achieve such progress, more research into the mechanisms of symptom development and change must be conducted. Dimensional symptom description might help in identifying specific origins and interventions for different symptom expressions. Relatively less knowledge exists about

OCD in childhood and its role for the course of disorder in adulthood. Moreover, OCD in old age is underrecognized and deserves more attention. More refined etiological models involving endophenotypes have the potential to bridge the gap between pure psychological and pure biological approaches in future research. A major challenge will be the realization of prospective longitudinal studies to track developmental aspects of symptom generation and maintenance. Furthermore, efforts to establish international consortia for collection and analysis of large-scale data samples, for example from brain imaging and genetic research, have to be intensified. Such projects might overcome statistical power problems for testing factors with small effects and allow for the development of more complex and realistic models of the disorder. A translational research perspective and increased efforts to disseminate the knowledge gained from basic and intervention research will ensure that progress in the field will ultimately arrive at the patient.

See also: Evidence-Based Psychotherapy; Imaging Methods in Clinical Psychology; Mental Illness, Genetics of; Obsessive-Compulsive Disorder.

Bibliography

Abramovitch, A., Abramowitz, J.S., Mittelman, A., 2013. The neuropsychology of adult obsessive-compulsive disorder: a meta-analysis. Clinical Psychology Review 33 (8), 1163–1171.

Beucke, J.C., Sepulcre, J., Talukdar, T., Linnman, C., Zschenderlein, K., Endrass, T., Kaufmann, C., Kathmann, N., 2013. Abnormally high degree connectivity of the orbitofrontal cortex in obsessive-compulsive disorder. JAMA Psychiatry 70 (6), 619–629.

Denys, D., Mantione, M., Figee, M., van den Munckhof, P., Koerselman, F., Westenberg, H., Bosch, A., Schuurman, R., 2010. Deep brain stimulation of the nucleus accumbens for treatment-refractory obsessive-compulsive disorder. Archives of General Psychiatry 67 (10), 1061–1108.

Endrass, T., Koehne, S., Riesel, A., Kathmann, N., 2013. Neural correlates of feedback processing in obsessive-compulsive disorder. Journal of Abnormal Psychology 122 (2), 387–396.

Fineberg, N.A., Brown, A., Reghunandanan, S., Pampaloni, I., 2012. Evidence-based pharmacotherapy of obsessive-compulsive disorder. International Journal of Neuropsychopharmacology 15 (8), 1173–1191.

Franklin, M.E., Foa, E.B., 2011. Treatment of obsessive compulsive disorder. Annual Review of Clinical Psychology 7, 229–243.

Friborg, O., Martinussen, M., Kaiser, S., Overgård, K.T., Rosenvinge, J.H., 2013. Comorbidity of personality disorders in anxiety disorders: a meta-analysis of 30 years of research. Journal of Affective Disorders 145 (2), 143–155.

Grisham, J.R., Fullana, M.A., Mataix-Cols, D., Moffitt, T.E., Caspi, A., Poulton, R., 2011. Risk factors prospectively associated with adult obsessive-compulsive symptom dimensions and obsessive-compulsive disorder. Psychological Medicine 41 (12), 2495–2506.

Hemmings, S.M., Lochner, C., van der Merwe, L., Cath, D.C., Seedat, S., Stein, D.J., 2013. BDNF Val66Met modifies the risk of childhood trauma on obsessive-compulsive disorder. Journal of Psychiatric Research 47 (12), 1857–1863.

Katerberg, H., Delucchi, K.L., Stewart, S.E., Lochner, C., Denys, D.A., Stack, D.E., Andresen, J.M., Grant, J.E., Kim, S.W., Williams, K.A., den Boer, J.A., van Balkom, A.J., Smit, J.H., van Oppen, P., Polman, A., Jenike, M.A., Stein, D.J., Mathews, C.A., Cath, D.C., 2010. Symptom dimensions in OCD: item-level factor analysis and heritability estimates. Behavioral Genetics 40 (4), 505–517.

Klenfeldt, I.F., Karlsson, B., Sigström, R., Bäckman, K., Waern, M., Ostling, S., Gustafson, D., Skoog, I., 2014. Prevalence of obsessive-compulsive disorder in relation to depression and cognition in an elderly population. American Journal of Geriatric Psychiatry 22 (3), 301–308.

Macy, A.S., Theo, J.N., Kaufmann, S.C., Ghazzaoui, R.B., Pawlowski, P.A., Fakhry, H.I., Cassmassi, B.J., IsHak, W.W., 2013. Quality of life in obsessive compulsive disorder. CNS Spectrums 18 (1), 21–33.

Manoach, D.S., Agam, Y., 2013. Neural markers of errors as endophenotypes in neuropsychiatric disorders. Frontiers in Human Neurosciences 7, 350.

Mataix-Cols, D., van den Heuvel, O.A., 2012. Neuroanatomy of obsessive compulsive and related disorders. In: Steketee, G. (Ed.), The Oxford Handbook of Obsessive Compulsive and Spectrum Disorders. Oxford University Press, New York, pp. 126–144 (chapter 7).

Menzies, L., Chamberlain, S.R., Laird, A.R., Thelen, S.M., Sahakian, B.J., Bullmore, E.T., 2008. Integrating evidence from neuroimaging and neuropsychological studies of obsessive-compulsive disorder: the orbitofronto-striatal model revisited. Neuroscience and Biobehavioral Reviews 32 (3), 525–549.

Milad, M.R., Furtak, S.C., Greenberg, J.L., Keshaviah, A., Im, J.J., Falkenstein, M.J., Jenike, M., Rauch, S.L., Wilhelm, S., 2013. Deficits in conditioned fear extinction in obsessive-compulsive disorder and neurobiological changes in the fear circuit. JAMA Psychiatry 70 (6), 608–618.

Mowrer, O.H., 1960. Learning Theory and the Symbolic Processes. Wiley, New York.

National Institute of Health and Clinical Excellence [NICE], 2005. Obsessive-Compulsive Disorder. Core Interventions in the Treatment of Obsessive-Compulsive Disorder and Body Dysmorphic Disorder. NICE, London.

Obsessive Compulsive Cognitions Working Group [OCCWG], 1997. Cognitive assessment of obsessive-compulsive disorder. Behavior Research and Therapy 35 (7), 667–681.

Olatunji, B.O., Davis, M.L., Powers, M.B., Smits, J.A., 2013. Cognitive-behavioral therapy for obsessive-compulsive disorder: a meta-analysis of treatment outcome and moderators. Journal of Psychiatric Research 47 (1), 33–41.

Riesel, A., Endrass, T., Kaufmann, C., Kathmann, N., 2011. Overactive error-related brain activity as a candidate endophenotype for obsessive-compulsive disorder: evidence from unaffected first-degree relatives. American Journal of Psychiatry 168 (3), 317–324.

da Rocha, F.F., Correa, H., Teixeira, A.L., 2008. Obsessive-compulsive disorder and immunology: a review. Progress in Neuropsychopharmacology and Biological Psychiatry 32 (5), 1139–1146.

Ruscio, A.M., Stein, D.J., Chiu, W.T., Kessler, R.C., 2010. The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication. Molecular Psychiatry 15 (1), 53–63.

Russell, E.J., Fawcett, J.M., Mazmanian, D., 2013. Risk of obsessive-compulsive disorder in pregnant and postpartum women: a meta-analysis. Journal of Clinical Psychiatry 74 (4), 377–385.

Salkovskis, P.M., 1985. Obsessional-compulsive problems: a cognitive-behavioural analysis. Behavior Research and Therapy 23 (5), 571–583.

Saxena, S., Brody, A.L., Schwartz, J.M., Baxter, L.R., 1998. Neuroimaging and frontal subcortical circuitry in obsessive-compulsive disorder. British Journal of Psychiatry (Suppl. 35), 26–37.

Skoog, G., Skoog, I., 1999. A 40-year follow-up of patients with obsessive-compulsive disorder. Archives of General Psychiatry 56 (2), 121–127.

Taylor, S., Jang, K.L., 2011. Biopsychosocial etiology of obsessions and compulsions: an integrated behavioral-genetic and cognitive-behavioral analysis. Journal of Abnormal Psychology 120 (1), 174–186.

Taylor, S., 2011a. Early versus late onset obsessive-compulsive disorder: evidence for distinct subtypes. Clinical Psychology Review 31 (7), 1083–1100.

Taylor, S., 2011b. Etiology of obsessions and compulsions: a meta-analysis and narrative review of twin studies. Clinical Psychology Review 31 (8), 1361–1372.

Taylor, S., 2013. Molecular genetics of obsessive-compulsive disorder: a comprehensive meta-analysis of genetic association studies. Molecular Psychiatry 18 (7), 799–805.

de Wit, S.J., Alonso, P., Schweren, L., Mataix-Cols, D., Lochner, C., Menchón, J.M., Stein, D.J., Fouche, J.P., Soriano-Mas, C., Sato, J.R., Hoexter, M.Q., Denys, D., Nakamae, T., Nishida, S., Kwon, J.S., Jang, J.H., Busatto, G.F., Cardoner, N., Cath, D.C., Fukui, K., Jung, W.H., Kim, S.N., Miguel, E.C., Narumoto, J., Phillips, M.L., Pujol, J., Remijnse, P.L., Sakai, Y., Shin, N.Y., Yamada, K., Veltman, D.J., van den Heuvel, O.A., 2014. Multicenter voxel-based morphometry mega-analysis of structural brain scans in obsessive-compulsive disorder. American Journal of Psychiatry 171 (3), 340–349.

Occipital Lobe (Visual Cortex): Functional Aspects

Oliver Braddick, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Several aspects of structure and function of the occipital lobe provide insights into the mechanisms of visual perception. The occipital lobe includes the primary visual receiving area (V1), which is laid out as a retinotopic map of the visual field, allowing signals from neighboring parts of the visual field to interact, and the first functional convergence of signals from the two eyes takes place on neurons in V1. V1 neurons show responses, which are selective for various stimulus properties, including orientation, spatial frequency, direction of motion, binocular disparity, and color. Neurons that respond to similar stimuli are arranged in 'columns' running perpendicular to the cortical surface, which form a complex regular structure.

Areas V1 and V2 act to segregate three different kinds of visual information: color, spatial pattern, and motion in interleaved structures ('blobs' in V1 and 'stripes' in V2). This segregation of information serves as a starting point for the distribution of visual information to a hierarchy of multiple cortical areas, each with its own topographical map of the visual field and a degree of specialized function. The properties represented by neuronal activity become progressively more complex and global in higher areas, and progressively less localized. While these areas are hierarchical and divided between 'dorsal' and 'ventral' processing streams, there are rich interconnections between areas, as well as feedback pathways by which areas higher in the hierarchy can modulate the activity of lower areas. All these systems become functional early in postnatal life, and their early development is initially highly plastic.

The Occipital Lobe

The occipital lobe, at the back of the brain, is one of the four lobes into which each cerebral hemisphere is conventionally divided. The lobes are defined purely by anatomical landmarks, not functionally. However, the occipital lobe has a certain functional unity in that, so far as we know, the whole lobe is primarily concerned with processing visual information. How this processing works at the neural level, and how it is organized structurally, provides some insights about the mechanisms of visual perception (*see* Visual Perception, Neural Basis of). Much of this knowledge has come from studying the occipital lobe of the macaque monkey, but human neuroimaging studies have become increasingly detailed in recent years and have shown that the human and macaque occipital lobes are closely similar in many respects, containing multiple areas which represent visual information in the field of view. It is now clear that this array of areas also extends into the temporal and parietal lobes, with some areas spanning the notional boundaries between lobes, reinforcing the idea that these boundaries do not have detailed functional significance.

These areas maintain a basic spatial organization of visual information, but also extract and organize visual attributes such as local shape, motion, color, and depth. To understand this organization, it is important to appreciate that processing goes on in the cerebral cortex, a thin, folded sheet of tissue across which distinct and specialized visual areas are laid out, but also that the cortex is layered, with different layers containing a characteristic pattern of connections providing inputs and outputs of information to and from each area.

The Retinotopic Projection to Area V1

The fibers delivering visual information to the cerebral cortex terminate in a region around the occipital pole of each hemisphere, known as striate cortex, primary visual cortex, or area V1. Each incoming fiber, and each individual neuron in V1, carries information about a specific small region of the visual field (its 'receptive field'). These receptive fields are arranged in an orderly way across the cortex, so that neurons with neighboring receptive fields are close together. Overall, area V1 in each hemisphere is laid out like a spatial map of the opposite half of the field of view (*see* Topographic Maps). The field is split, with the right half-field represented in the left hemisphere, and vice versa. In this map, the vertical midline of the field is represented furthest back, near the occipital pole, and moving forward across the medial surface of the cortex corresponds to moving horizontally outward in the field of view. This orderly 'retinotopic' mapping must be functionally important, presumably because it allows neurons serving nearby field locations to interact readily. Several functions for such interactions can be suggested. First, to reliably represent local visual properties, such as the orientation of edges, requires information to be gathered and compared across cells representing a local area of the visual field. Second, local connections with a regular organization between cells can serve to link the features that form spatially connected objects, for example, adjacent parts of an object's outline. Third, connections linking nearby field regions may be critical in the development of cortical function, exploiting the statistical associations between the stimulation received from adjacent parts of the scene to build functional units that are well tuned to the structure of the visual world.

The exception to the rule that nearby locations in the field are represented close together in the cortical surface is provided by the division between hemispheres at the vertical midline. 'Local' connections across the midline have to be provided by long fibers that cross between hemispheres in the corpus callosum. Despite this, it is striking that our subjective impression of the visual field is quite seamless at the midline, and the

division between hemispheres has few if any functional consequences in normal individuals.

The scale of the retinotopic map is not uniform. The cells of the central retina are much more densely packed than in the peripheral field, and so provide much more information per unit area. This variation in information is reflected in a variation of the 'magnification factor' in V1, with a much larger area of cortex devoted to the central region than to the same area in a peripheral region. In this way, cortical organization provides the greater processing power needed for the fine detail and spatial precision of central vision. This principle extends to 'extrastriate' visual areas beyond V1, although the quantitative relationship between cortical 'magnification' and area of tissue varies between areas.

Binocular Organization

An important function of area V1 is that it provides the first point in the visual pathway where information from the two eyes can be integrated. The fibers from each eye are segregated into alternating bands ('ocular dominance stripes') as they arrive in layer 4 of cortical area V1 (see Section Columnar Organization). The first real interaction between left- and right-eye signals occurs when these fibers converge on cells in the layers above layer 4 (Hubel, 1988). This interaction allows both eyes to contribute to a unified visual representation of the world. It also allows neurons in V1 and beyond to register the disparity between the two eyes' views, which serves as the basis for stereoscopic depth perception (Parker, Chapter 49 in Chalupa and Werner, 2004).

Stimulus Selectivity of Cortical Neurons

The number of neurons in area V1 is at least a 100-fold greater than the number of input fibers. This enormous information-processing resource means that V1 does not simply map the visual input: it can also transform it. Each local region of the visual field is represented by an array of neurons, with each neuron responding selectively to a particular type of stimulus (Hubel, 1988). Most V1 neurons respond to contours, and are selective in several, if not all, of the following respects:

- strongest response to a preferred orientation of the contour, with response falling off as the angle is further from this optimum ('orientation tuning');
- a preferred type of contour: light bar, dark bar, or edge;
- a preferred spatial scale, so that the cell may respond either to the broad pattern of light and dark, or to fine detail ('spatial frequency selectivity');
- stronger response to one direction of motion than for the opposite direction;
- differences in binocular balance: some cells respond predominantly to the right eye, some to the left, and some only give their best response when both eyes are activated together. In the latter, binocular group, cells may be selective for stereoscopic depth, i.e., they respond best to objects that are either nearer, farther, or close to the distance on which the eyes are converged; and
- color, with one range of wavelengths preferentially activating the cell and another inhibiting it (sometimes combined with an enhanced response to color contrast).

This array of differently tuned neurons means that the cortex can represent all these diverse properties according to which particular pattern of neurons is active (a population-based 'place code'; *see* Visual Perception, Neural Basis of).

Columnar Organization

Retinotopic mapping (Section The Retinotopic Projection to Area V1) implies that within a small region of V1, all the cells deal with a local region of the field of view. The cortex also has a finer-grain spatial organization, reflecting the differences in tuning described in Section Stimulus Selectivity of Cortical Neurons. Connections between neurons are predominantly vertical (i.e., perpendicular to the cortical surface). Presumably as a result of these connections, neurons in a vertical column have similar response properties. In particular, there are columns of cells responding to the same orientation, and broader columns within which a particular eye predominates. The combination of these two forms of organization has led to the idea of the 'hypercolumn,' a module of visual cortex that includes a complete representation of both eyes' inputs and all possible orientations, for a particular local region of the visual field.

From column to column, there is also a spatial ordering; in a horizontal track across V1, the preferred orientation changes in an orderly sequence. As with the retinotopic organization, this order means that a regular, local pattern of connectivity between cells creates communication between cells responding to contours with the same or similar orientation, supporting the integrative processes discussed in Section The Retinotopic Projection to Area V1.

Thus, at the fine scale, cortical layout is not purely a retinotopic map. However, both retinotopic and orientation-based mapping mean that orderly spatial connectivity between cells can establish communication between cells responding to similar stimuli. The columnar organization has been described here for V1, but it is believed to be a general principle that applies also to the extrastriate areas described in Section Specialized Extrastriate Areas, and indeed to cortical organization more widely, including the representation of other sensory modalities besides vision.

Modern methods of optical imaging allow the array of neurons that respond to a particular stimulus to be highlighted, and so the spatial layout of the orientation columns can be visualized. It turns out that they are not so much columns as bands that converge at 'pinwheels,' so that the representation of contour orientations around the clock is arranged in a path around the pinwheel center. Motion responses also enter the organization: neurons responding to a particular contour orientation (e.g., vertical) may be selective for either of two opposite motions of the contour (e.g., leftward or rightward); there is a sharp discontinuity between regions of the two types, and these lines of discontinuity connect pinwheel centers. Another, related, organization is a periodic pattern of 'blobs' showing a high concentration of the metabolic enzyme cytochrome oxidase (CO), that can be revealed anatomically by staining cortical tissue. These blobs

contain many color-sensitive cells and little specificity for orientation. Overall, there appears to be an elaborate, repetitive, crystalline order to the functional selectivities of cortical neurons. Many details of this ordering are not yet understood, but it must be a key to understanding the circuits that define visual cortex function, and how these are constructed in development.

Interleaved Pathways through V1 and V2

The visual pathway leading to V1 in fact includes two major subpathways, and this division runs through processing in V1 and beyond (Daw, 2012). One group of cells, the magno cells, have large receptive fields, fast responses, and do not carry information about the color of light. The parvo cells, in contrast, respond less well to rapidly changing stimuli, but better to fine spatial detail, and are wavelength-selective (e.g., activated by one range of colors and inhibited by others). These two classes of cells appear to separate the information useful for detecting motion on the one hand, from that required for fine detail and for color vision on the other. Correspondingly, the perception of motion and of color seem to work largely independently of each other.

This division is maintained in the cortex. In V1, the fibers of magno cells terminate in a distinct layer (4B) containing many motion-selective cells. Within the parvo input, color-specific information is also anatomically separated into the CO 'blobs' described in Section Columnar Organization. Thus, in the fine structure of area V1, three different kinds of processing are segregated but interleaved: motion information in layer 4B, color information in the blobs, and spatial pattern (orientation) in the 'interblob' regions.

V1 sends information to a neighboring area, V2, which forms a second retinotopic map. The three types of visual information interleaved in V1 are kept segregated in V2, into three parallel sets of stripes that send their output to different extrastriate areas (Section Specialized Extrastriate Areas). Thus, apparently a major role of areas V1 and V2 is to marshal different kinds of information into separate cortical processing streams.

Specialized Extrastriate Areas

V1 and V2 are only two of what are now recognized to be a large number of distinct cortical visual areas forming an interconnected hierarchy (Zeki, 1993). Further forward in the occipital lobe, most of the lobe can now be divided into a mosaic of areas, referred to as 'extrastriate' since they lie outside the primary striate area. These areas can be distinguished by

- mapping: each area (at least at the lower levels of the hierarchy) forms a distinct, complete retinotopic map of the visual field;
- connectivity: there is an orderly set of projection fibers connecting one area to another and preserving the retinotopic mapping. Connections with 'higher' and 'lower' levels in the hierarchy are made by different cortical layers, which allow the hierarchical structure to be defined; and
- function: each area has a characteristic balance in the type of stimulus selectivity found in its neurons (e.g., color selectivity common and directional motion selectivity rare, or vice versa), leading to the idea that different areas have distinct functional roles in vision.

This functional division of visual areas was first suggested from recordings of the responses of single neurons in macaque cortex. More recently, it has been explored in the human brain by functional magnetic resonance imaging (fMRI; *see* Cognitive Neuroscience; Eye Movements in Reading) (Wandell and Winawer, 2011). Imaging allows areas to be defined by their retinotopic maps, and also by the specific visual properties that activate them (e.g., one area is activated best by color differences, a different area by coherent motion in the visual field). This idea of functional localization fits with findings in some patients with brain lesions, who show a deficit limited to one specific aspect of vision (e.g., a loss of just color or just motion processing).

The main areas that can be identified as retinotopic maps have been labeled as V3, V3A, V4, and V5 (**Figure 1**). Together they form a third tier (or possibly, third and fourth tiers) of a processing hierarchy beyond V1 and V2. Area V5, also known as MT ('middle temporal,' although this does not accurately describe its location in humans or macaques), lies on the boundaries of the occipital and temporal lobes in the human brain. V5 has been especially intensively investigated because it has a very well-defined specialization for processing motion information. At the same level in the macaque cortex, but complementary in its specialization, is area V4, which has the highest proportion of color-selective neurons, a type of selectivity almost absent in V5. Macaque V4 also shows processing of more complex aspects of spatial pattern than those found in V1 and V2 neurons.

Areas V4 and V5 represent a further stage in the segregated streams of processing found in V1 and V2 (see Section Interleaved Pathways through V1 and V2), with V5 as the destination of the motion information derived from layer 4B of area V1, and V4 receiving the color and form information derived from the blobs and interblob regions. Many V5 cells encode stereo disparity as well as motion, in a way that allows their combination to give information about the 3-D layout of the environment and the observer's motion through it. The functional distinctions between V4 and V5 are an aspect of the broader division between 'dorsal' and 'ventral' cortical streams (Daw, 2012) (*see* Visual Streams: Dorsal and Ventral). The dorsal stream consists of areas carrying motion and spatial information to the parietal lobe, where it provides the visual basis for controlling actions. The 'ventral' stream carries pattern and color information required for the identification of objects in the temporal lobe.

Functional anatomy, then, supports a view that the visual process is based on distinct modules with different functions; in particular, that there is a broad separation of the modules which tell us 'what' object we are looking at, from those that tell us 'where' or 'how' we can act upon it.

Evolution of Occipital Lobe Organization

It is clear that the same division between the streams exists in human as in monkey cortex, but some important details are

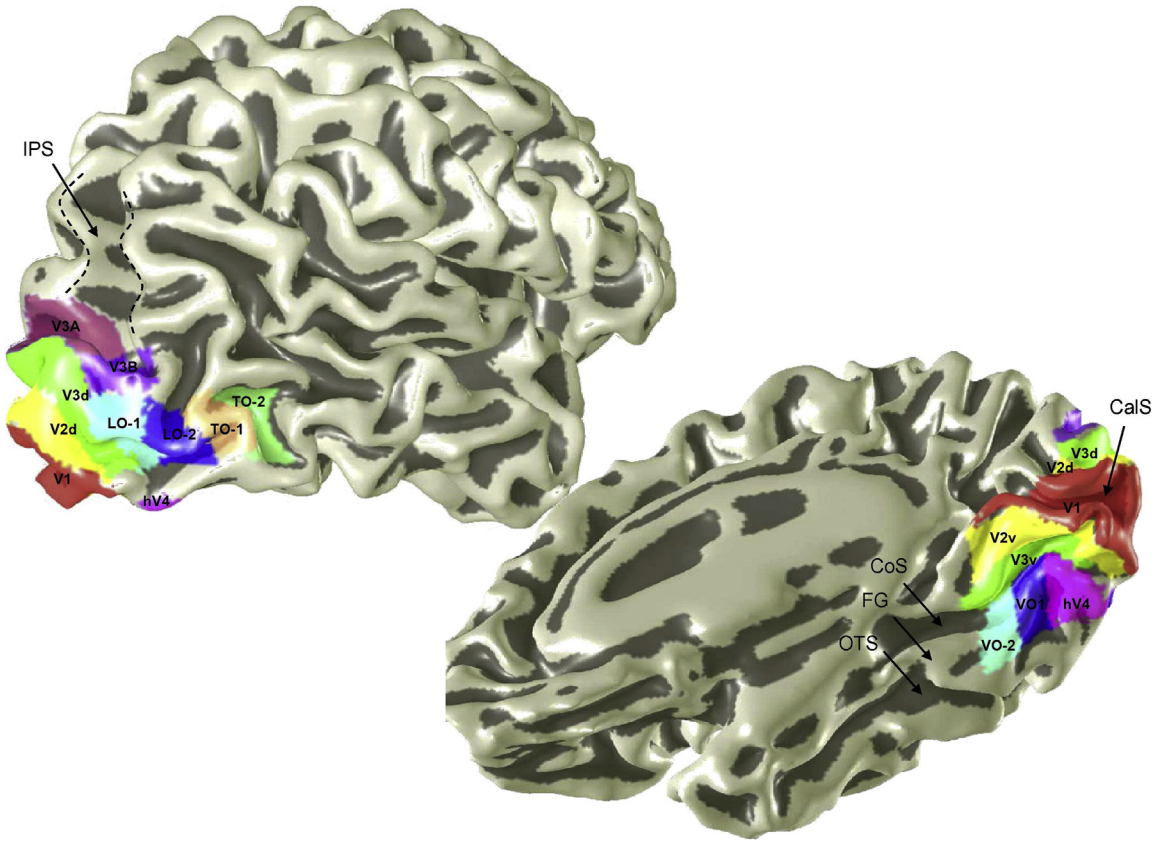

Figure 1 Distinct visual areas in the occipital lobe, revealed by fMRI mapping of visual field representations (from Wandell, Brian A., Winawer, Jonathan, 2011. Imaging retinotopic maps in the human brain. Vision Research 51, 718–737 by permission). The upper left figure illustrates the view of the outer (lateral) surface of the right cerebral hemisphere; the lower right shows the medial surface (which faces the other hemisphere across the midline). V1 (red) is the first visual area where the input fibers of the optic radiation arrive. Areas V2 and V3 have separate ventral (v) and dorsal (d) sections, mapping the upper and lower halves respectively of the visual field. V3A and V3B are distinct areas with their own maps, which despite this notation are not parts of V3. 'hV4' refers to human V4, 'LO-1 and LO-2' to maps of the field within the lateral occipital complex. 'TO-1, TO-2' denote maps in the temporal/occipital area identified by the authors, which are included within the motion-sensitive area MT+ (otherwise known as V5). The functional distinctions between these areas, and their interconnections, are discussed in the text. 'IPS,' 'CoS,' 'FG,' 'OTS,' and 'CalS' refer not to functional areas but to anatomical landmarks: intraparietal sulcus, collateral sulcus, fusiform gyrus, occipito-temporal sulcus, and calcarine sulcus, respectively.

different (Sereno, 1998; Van Essen, chapter 32 in Chalupa and Werner, 2004). Because of the great expansion of frontal cortex in humans relative to monkeys, the visual areas occupy a smaller percentage of the area of the human brain, but their absolute area is still several times greater than in the macaque monkey. Some areas in the human occipital lobe do not seem to have a direct correspondence with the monkey's organization. For example, human neuroimaging reveals a prominent lateral occipital area (LO), which responds much more strongly to the image of an object than to a scrambled version of the same object, and is presumed to be an important component of the ventral-stream network that enables us to recognize objects. Such an area has not yet been recognized in the monkey brain. The relation between color sensitivity and retinotopic mapping, found in monkey V4, seems to involve a more complex set of areas in the human brain. In the dorsal stream, the anatomy and connectivity of area V3A seems similar in both species, but in humans this area is highly motion-sensitive, whereas in the monkey it responds to stereoscopic depth but not particularly to movement. It is now possible to carry out closely analogous measurements using fMRI brain imaging in each species, and future studies with this method should reveal how the functional organization of the human occipital lobe has evolved.

Spatial Generalization, Functional Specialization, and Global Processing

Until recently, it was believed that retinotopic organization was limited to the areas V1–V5 described above, and that this feature was lost in further levels of the presumed hierarchy, such as LO area. Advances in imaging methods have led to revisions of this view, so that for instance LO has now been shown to consist of two areas each with its own retinotopic map. Nonetheless, it remains true that spatially precise mapping becomes a less dominant principle as we ascend the processing hierarchy. For example, within the dorsal stream, the directionally selective cells in area V1 have quite small receptive fields; near the center of the visual field these may span an area less than 1° across, while the receptive field size of

V5 cells in the same location is 4–5 times larger. A group of V1 cells are believed to provide subunits, which combine to yield the larger motion-sensitive area in V5. V5 cells in turn provide the inputs to cells in the nearby area MST ('medial superior temporal,' no specification of its location in humans or macaques), whose receptive fields are so large that they may extend into the opposite half of the visual field. Directional neurons in V1 represent motion in a very local region of the field; the increase in receptive field size in V5 and MST provides the basis for analyzing more global properties. Thus V5 neurons are highly sensitive to the coherence of motion direction over an extended region, and in turn MST cells combine V5 signals to detect the wide patterns of optic flow across the retina which occur as a result of the observer moving through the environment. Similarly global properties, but related to static pattern rather than large-scale motion, are represented in V4 and higher areas of the ventral stream.

There is a second potential reason why higher levels of the cortical hierarchy should show increasing receptive field size and coarser, or absent, retinotopic mapping. The recognition of an object should not depend on it being in a particular location but, rather, requires generalization across the visual field. Thus a neuron which plays a part in such recognition needs to gather information across a large spatial field.

In both dorsal and ventral streams, it seems to be a general principle that as visual information is progressively transformed and elaborated in the brain, the less relevant is retinal location for the way in which it is represented.

Development and Plasticity in the Occipital Lobe

In normal human development, the functions of occipital cortex are rudimentary at birth, and the visual behavior of newborn babies is largely controlled by neural networks in subcortical brain structures. However, in the first year of life very large numbers of synaptic connections are formed. This increasing connectivity allows the special functions of visual cortex to emerge: in the first 3–4 months orientation selectivity, directional selectivity, and binocular interaction develop in sequence, and global processing in extrastriate areas contributes to infants' visual abilities soon afterward (Braddick and Atkinson, 2011).

The emerging structural and functional organization of visual cortex do not follow from a rigid genetic wiring diagram, but result from the interplay of an intrinsic ground plan with the neural activity stimulated by visual experience. The most studied aspect of this is the interaction between signals from the two eyes (Section Binocular Organization) (chapters 4 and 12, Chalupa and Werner, 2004): if one eye is deprived of vision, that eye starts to lose its connections to area V1, the ocular dominance bands for that eye shrink and those for the other eye expand. This plasticity of cortical connections produces the phenomenon of amblyopia, or 'lazy eye,' in children who suffer an early imbalance between the eyes (for instance, if one eye is very long-sighted). Recent brain imaging studies have shown that even more radical reorganizations of the visual cortex are possible. For example, in an infant for whom one half of the brain is seriously damaged around birth, the map in the visual cortex can be completely rearranged so that both halves of the visual field are represented in the intact cerebral hemisphere. This plasticity is not restricted to the single sense of vision: functional imaging shows that in individuals who have been congenitally blind (but not in late-blind or sighted people), the visual occipital cortex is activated when they are reading Braille, implying that the sense of touch has in part taken over an area that is usually visual. All of these changes in the brain require the anomalous sensory input to take place in early childhood, supporting the idea of a 'critical period' in which brain networks can be adapted to the pattern of input, before neural connections become set. Research has found that dorsal stream functions are much more disrupted in developmental brain disorders than those of the ventral stream, but the neural basis for this 'dorsal stream vulnerability' is not yet understood.

Recurrent and Top-Down Influences

While it is clear that the visual areas can be ordered into a hierarchy, the flow of information is not simply upward from the lower levels. Almost every connection from a lower to a higher area has a corresponding connection in a reverse direction. Thus, in the neural structure, there exists the potential that low-level processing, e.g., in V1 and V2, does not run simply automatically, but is controlled and directed by top-down processes. There is also the potential for recurrent loops, by which information is exchanged between lower and higher areas and progressively refined. Such loops mean, for instance, that although the primary response of V1 neurons is to local features of the visual pattern such as an edge segment, these responses can be influenced by the global structure of the scene, such as whether the edge is on the left or right of an object (Lamme, chapter 45 in Chalupa and Werner, 2004). And although the different cortical modules have distinct functions, there are rich interconnections between them at the same level, including connections between dorsal and ventral streams. Occipital cortex is a modular system, but one in which the modules closely interact to produce integrated visual perception.

Top-down influences on occipital areas can also come from further afield in the brain. The decision to attend to one particular region of the field of view activates the corresponding part of the retinotopic maps in area V1 and elsewhere, acting to amplify the response to stimuli in the attended region (Kastner, chapter 101 in Chalupa and Werner, 2004). The effects of attention are not limited to sensitizing one area of space; for example, if a particular judgment is required, say of color or of motion speed, the visual brain areas responsible for the relevant kind of processing become differentially active. Perception is an active process, not just the passive processing of the visual input, and the occipital lobe participates in this goal-directed activity.

See also: Cognitive Development in Infancy, Neural Mechanisms of; Topographic Maps; Vision, High-Level Theory of; Vision, Low-Level Theory of; Visual Memory, Psychology of; Visual Perception, Neural Basis of; Visual Space, Geometry of.

Bibliography

Braddick, Oliver, Atkinson, Janette, 2011. Development of human visual function. Vision Research 51, 1588–1609.

Chalupa, Leo M., Werner, Jack S., 2004. The Visual Neurosciences. MIT Press, Cambridge, MA.

Daw, Nigel, 2012. How Vision Works. Oxford University Press, New York.

Hubel, David H., 1988. Eye, Brain and Vision. Scientific American Library, New York.

Sereno, Martin I., 1998. Brain mapping in animals and humans. Current Opinion in Neurobiology 8, 188–194.

Wandell, Brian A., Winawer, Jonathan, 2011. Imaging retinotopic maps in the human brain. Vision Research 51, 718–737.

Zeki, Semir, 1993. A Vision of the Brain. Blackwell, Oxford, UK.

Occupational Health

Töres Theorell, National Institute for Psychosocial Factors and Health, Stockholm, Sweden; Karolinska Institutet, Sweden; and Stress Research Institute, Stockholm University, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article a brief historical background of the development of social and psychological working conditions is followed by a description of the presently most widely utilized theoretical models for studying the psychosocial work environment in research and practice, the 'person–environment (PE) fit,' 'demand–control–support,' 'organizational justice' and 'effort–reward' models. Possibilities and difficulties in assessment are discussed. It is pointed out that both 'objective' and 'subjective' aspects are important and that these are not always strongly correlated. Physiological models linking the psychosocial models to health and disease are discussed and finally experiences of interventions are discussed. The main conclusion is that it is possible to define and assess psychosocial factors at work, that such factors are linked to physiological mechanisms, and that interventions based upon them are possible.

Introduction: Work and Health

During historical periods preceding industrialization, most of the work that humans were involved in was unpaid. The borders between work and leisure were not clear. It was with industrialism that paid work started to play such an important role in human life. For several generations, paid work has been the activity that has generated a family's needs for survival.

Health is not only the absence of illness. It is also a positive state in its own right. However, most of the research that has been published regarding work and health has been devoted to the establishment of relationships between certain adverse working conditions and illness. Accordingly, most of this article will be on work–illness relationships. But we shall also discuss possible pathways between good working conditions and the promotion of health.

An important basis for our understanding of the effects of working conditions on health has been the physiological stress research which was initiated by Hans Selye (1976). Initially stress researchers studied acute reactions to energy-demanding situations. Subsequently, they examined the biological consequences of long-lasting adverse life situations.

The Physiological Basis of the Effects of Work on Health

Genetic evolution and selection processes are slow. Therefore our biological reactions to stressful situations are likely to be approximately the same as they used to be several thousand years ago. It has been emphasized by many authors (Levi, 1972; Frankenhaeuser, 1980) that this may not be functional for man in modern societies and especially not at work where physical demands are becoming less and less important. It was functional for the primitive human being who needed to face dangerous animals to become strong, courageous, and insensitive to physical pain. However, most contemporary dangers have to do with complicated social interactions and in such situations the mobilization of physical energy is not functional. It is possible to regard some adverse aspects of lifestyle, such as smoking, the use of drugs stimulating the central nervous system, and the consumption of sweets and alcohol as 'external' efforts to cope with energy-demanding situations.

Energy Mobilization and Anabolism at Work

In Selye's (1976) general adaptation syndrome, 'stress' was seen as the general reaction to a nonspecific challenge or adverse condition. The situation that induced stress was labeled a stressor (factor which induces stress). Since it has central importance to the understanding of stress at work, a short description of energy mobilization will be given here. The most important biological process is the provision of energy – glucose and free fatty acids enter the blood and are used for the immediate production of energy. But there are several parallel phenomena, all of which aid the body in physical fight or flight responses. Examples are lowered excretion of water and salt, decreased sensitivity to pain, and decreased inflammatory responses to infections. Since energy mobilization (resulting in elevated blood concentration of glucose and free fatty acid) has the highest priority, anabolism (restorative and regenerative activities in the cells) is downregulated. Anabolism is central to the body's central defense of all the organ systems that need constant rebuilding and restoration. If this goes on for a long time (several months) increased sensitivity to physical and psychological stress in bodily organs is the ultimate result. There is a balance between regulation of energy mobilization and anabolism. There are two parallel regulatory systems ranging from the hypothalamus to the pituitary and to their respective endocrine organs, namely the hypothalamo–pituitary–adrenocortical (HPA axis) and the hypothalamo–pituitary–gonadal (HPG axis) systems. While the HPA axis facilitates energy mobilization, the HPG axis facilitates regeneration and reproduction. In many situations they counteract one another. For instance, during long periods of excessive arousal and effort the HPG axis is inhibited. But on the other hand, high HPG axis activity may protect the body against the adverse effects of long-lasting, excessive HPA axis activity

(Theorell, 2012). This has importance in the work environment in the sense that the psychosocial processes that stimulate the HPG axis may also protect against adverse effects of a stressful work environment.

Another consequence of long-lasting demands for energy mobilization is that the endocrine systems may change their regulatory patterns. This means that the ability of the body to stop energy mobilization when it is no longer needed may be disturbed or that the counter-regulation (inhibition of energy mobilization) is inhibited. Most processes in the body have a counter-regulatory mechanism that operates in order to inhibit a process that has been stimulated.

Physiological Short-Term and Long-Term Adaptation to Requirements at Work

The HPA axis is of central importance to the biological arousal mechanism. The steroid hormone cortisol which is released from the adrenal cortex is the most extensively studied 'agent' in this system. Cortisol facilitates biological arousal in most organ systems. Disturbed regulation of the HPA axis has been described as a possible consequence of long-lasting arousal. There are at least three kinds of such regulatory disturbances in the HPA axis, all of which are of relevance to health at work:

1. Lack of inhibition of arousal. All reactions in the body evoke counter-regulatory reactions. This is true also of energy mobilization, which should stop when the threat disappears. Normally cortisol inhibits higher centers in the HPA axis when it has been in the aroused state for some time. When this inhibition does not take place in the normal way, the levels tend to become constantly elevated.
2. Lack of response to stimulation of the HPA axis. This means that plasma cortisol levels are low and that normal activation does not take place when the HPA axis is stimulated artificially. This has often been labeled the 'flat cortisol curve.' This refers to the fact that the circadian variation is small and the 'awakening response' (the normal increase in cortisol levels from awakening until half an hour later, which could be regarded as a kind of stress response) is attenuated.
3. Unstable levels. In this kind of disturbance the plasma cortisol levels may be low at rest, but certain situations – that are specific to the individual and have to do with the symbolic significance of the event – trigger a strong HPA axis reaction that may seem to be out of proportion to the importance of the trigger. This occurs in subjects who have been exposed to extraordinarily stressful situations, such as torture, war, disasters, and incest.

The first kind of regulatory disorder (1) is typical of Cushing's syndrome (overproduction of corticosteroids including cortisol due to a disease in the adrenal cortex) and could be one alternative regulatory disturbance in clinically severe psychiatric depression. In patients hospitalized for severe psychiatric depression a pattern with constantly high cortisol levels was observed (Rubin et al., 1987). This pattern has become less prevalent with the frequent use of antidepressive medication in the population. Currently, in epidemiological studies of the relationship between cortisol excretion and depression, it is mostly not observed (see for instance Alderling et al., 2008).

The second kind (2) has been shown to exist in a subgroup of patients with the diagnosis of chronic fatigue syndrome or CFS (Demitrack et al., 1991). Conditions with long-lasting fatigue as a prominent feature, for instance the burnout syndrome, have been shown to be associated with poor awakening response, which is correlated with the 'flat cortisol curve' (Chida and Steptoe, 2009; Halford et al., 2012)

The third kind of regulatory disorder (3) is found in post-traumatic stress disorder (PTSD) (Charney et al., 1993). The key element in the etiology of PTSD seems to be that the long-lasting extreme arousal level increases the number of cortisol receptors on the cell surfaces and also increases their sensitivity. Since the sensitivity is high, only low concentrations of cortisol are needed to keep the system working at rest. This may be the reason why low serum cortisol levels are found at rest in patients with PTSD.

It is important to consider these different forms of regulatory disturbance in working subjects. Men and women with severe such regulatory disturbances are mostly unable to work and mild disturbance causes disturbed ability to cope with working tasks.

Unemployment and Stress

Unemployment is a common phenomenon in all industrialized countries. In most European countries it is presently of the order of 10%. Research has tried to establish the extent to which unemployment contributes to poor health. At present there is agreement that poor health increases the risk of becoming unemployed in periods of financial crisis but there is also agreement that unemployment per se contributes to poor health. Longitudinal studies of subjects who have gone through the different stages of unemployment (Brenner and Levi, 1987) or started activities again after a long period of unemployment have shown that several physiological systems are markedly affected by these processes. For instance, there is elevation of the secretion of cortisol (an index of energy mobilization) and prolactin (which mirrors passiveness in a crisis situation). It is accordingly reasonable to assume that health could be affected. There is published research showing that long-lasting unemployment increases the risk of mental disorder (Hallsten, 1998) and is also associated with a lifestyle that increases the risk of developing cardiovascular disease, particularly in young people (Janlert et al., 1991; Janlert and Hammarström, 1991). Economic hardship may add to the health consequences (Vinokur et al., 1996). Downsizing is associated with increased prevalence of poor mental health (emotional exhaustion and depression) not only for the dismissed workers but also for those remaining as workers in the 'downsized' workplace. A recent study of the situation in Hungary, Sweden, and France showed that the relationship between downsizing and poor mental health is particularly pronounced in Hungary where societal and financial support systems are weaker than in the other studied countries. In addition, studies show that subjects who remain employed after dismissal of colleagues in their workplace may also suffer from mental symptoms and that the way in which

management handles downsizing is crucial for the mental health of the employees (European Commission, 2011). Not unexpectedly, a well-organized downsizing process with good information and understandable goals does not result in as much health deterioration among employees as a chaotic, poorly organized downsizing process resulting in a less effective organization.

Jahoda (1979) formulated the basic reasons why work is such an important activity in modern life – the most central one being that work provides structure to life. There is extensive research showing that organized support can help the unemployed to regain employment (Vinokur et al., 1995). Evaluation of such programs, however, also shows that these programs may increase passiveness and feelings of low self-esteem if the initiative of the individual is taken away.

Combined Effects of Physical and Psychosocial Adversities at Work

Extensive research has been devoted to the disentanglement of physical stress from psychological stress in studies of health and disease. For instance, in the study of the importance of ergonomic load to the development of musculoskeletal disorders, it has been considered important to adjust for psychological stress. Extensive epidemiological studies of large samples of subjects have shown that ergonomic load contributes to the development of musculoskeletal disorder but that psychosocial factors at work may also contribute significantly (Kilbom et al., 1996). It has been shown that combinations of physical ergonomic and psychosocial factors may be of particular importance to low back pain (Vingård et al., 2000). The potential importance of the combination of noise exposure and job strain has also been documented in relation to morning cortisol excretion (Selander et al., 2012). Furthermore there seems to be an interaction between long-term symptoms of prolonged adverse stress (burnout syndrome, see above) and the ability to cope with a combination of acute stress and unpleasant sound, particularly among female working subjects. A study was based upon screening of working subjects with high scores on emotional exhaustion. Emotionally exhausted subjects showed lowered noise thresholds after acute experimental stress whereas normal subjects showed elevated thresholds. In other words, the normal way of coping with a combination of acute stress and noise is to elevate the noise tolerance whereas exhausted subjects become more sensitive to noise when there is acute stress (Hasson et al., 2013).

It has been important for the study of relationships between working conditions and health to have theoretical models that form the basis for hypothesis testing. The interplay between the individual and the environment has been an important point of departure in all theoretical modeling in this field. Some of the models have focused on environmental aspects – for instance, the demand–control and organizational justice models – and some on individual aspects, for instance, the eustress–distress model. The person–environment (PE) fit and the effort–reward imbalance models are other examples of models taking the interaction between the individual and the environment into account. These four models will be described briefly below.

Theoretical Models for the Study of Work and Health – Individual or Environment

The individual's response to demanding work situations has been the point of departure for Frankenhaeuser and her coworkers (Frankenhaeuser, 1980) who introduced one of the most influential models, the eustress–distress model. According to this model, the individual responds to the work situation with different degrees of effort. The bodily reactions to a high degree of effort are influenced by the concomitant degree of joy or distress in the situation. High effort with joy is labeled eustress while high effort with marked distress corresponds to a negative stress reaction.

The environmental counterpart of the effort–distress model is the demand–control model that has been used extensively. Accordingly this will be described in more detail. Karasek introduced the 'demand–control' model. It is a two-dimensional synthesis of the demand ('stress psychological') and the 'lack of control' (sociological) research traditions (Karasek and Theorell, 1990; Karasek, 1979). Lack of control has two components: 'Skill utilization' is related to the employee's control over the use and development of his/her skills, whereas 'authority over decisions' is related to the employee's control over decision-making relevant to his/her work tasks. The two factors are mostly added to one another to constitute 'decision latitude.'

The other dimension in the original demand–control model, psychological demands, includes both quantitative (how much work per time unit) and qualitative demands. Qualitative demands can be subdivided for instance into cognitive and emotional demands and demands to control one's emotions as in some kinds of health care work (Pejtersen et al., 2010). It has been claimed that different kinds of demands, for instance the ones related to intensity and to the number of working hours (extensiveness) should be treated separately for different groups of employees (Kristensen et al., 2004).

The high demand–low decision latitude combination, *job strain*, is regarded as the most dangerous in the development of illness. According to the theory, this kind of situation, if prolonged and repeated for a long time, increases sympatho-adrenal arousal and at the same time decreases anabolism, the body's ability to restore and repair tissues (Karasek and Theorell, 1990).

The combination of high psychological demands and high decision latitude is defined as the active situation. In this situation, the worker should be given more resources to cope with high psychological demands so that he/she can make relevant decisions, such as planning working hours according to his/her own biological rhythm. In addition he/she gets good possibilities to improve coping strategies – facilitating feelings of mastery and control in unforeseen situations. This situation corresponds to psychological growth.

The low demand–high decision latitude situation is the relaxed and theoretically ideal one, whereas the low demand–low decision latitude situation, which is labeled passive, may

be associated with risk of loss of skills and, to some extent, psychological capacity.

It soon became evident that a third component, social support, was needed in the demand–control model. This dimension had already been introduced previously in job stress research (House, 1981). Johnson has discussed this component extensively. He has pointed out that collective support may be difficult to separate theoretically from collective control. Workers who are exposed to adverse working conditions may join forces and increase their control over their working conditions. The worst hypothesized situation is 'iso-strain' – job strain and lack of social support (Johnson et al., 1989).

There is a striking difference between the effort–distress model and the demand–control model. Whereas the former is designed to explain differences in individual response to stressors, the latter is designed to assess the environmental conditions, – the stressors. There is of course a relationship between the models since eustress reactions are likely be more common in the 'active job' situation and negative stress reactions more common in the 'job strain' situation.

Interaction Between Individual and Environment

One of the most influential models has been that of the Michigan school, the PE fit model (Katz and Kahn, 1966). This theory postulates that adverse reactions at work arise when the PE fit is poor. In this model the interaction between a person's characteristics and the environment is the central theme. If the characteristics of the job do not fit the characteristics of the person there is a high likelihood that negative stress reactions may arise. There is considerable empirical support for this hypothesis (Kahn, 1981). Another important theme is the expectation of the person – those who have very positive expectations may be more disappointed than those who have negative expectations, if the conditions do not fit. The Michigan school also added social support at work as a central theme (House, 1981).

According to Siegrist (1996), a crucial job factor for health consequences is the degree to which the workers are rewarded for their efforts. When a high degree of effort does not meet a high degree of reward, emotional tensions arise and risk of illness increases. The 'effort–reward imbalance' model is described in another chapter by Siegrist in this book. It is a sociological theory that includes individual aspects. The effort component includes environmental (extrinsic effort, which is close to 'psychological demands' in the demand–control model) as well as individual aspects (intrinsic effort, overcommitment). A high degree of overcommitment could be the result of previous job experiences as well as of other experiences in the person's life. The reward component includes three different aspects of reward, namely monetary reward, promotion possibilities, and self-esteem. The construction of the questions specifically addresses whether the individual sees, for instance, a specific kind of extrinsic effort as a problem or not.

A more recent theoretical model related to the previous ones is the 'demand–resource' model (Demerouti et al., 2001). It is similar to the demand–control model, but it incorporates both individual and environmental resources. Whereas the demand–control model tries to focus on the organizational aspects of work, the demand–resource model has a mixed focus.

'Organizational justice' is another basic concept that has been used in job stress research. This concept builds upon the idea that job sites must have a system/organization for resolving conflicts and that this is important for employee health. It has also proved to be useful in illness prediction (Kivimäki et al., 2006).

Several multidimensional standardized questionnaires for the measurement of stressors and stress reactions at work have been published. The most extensively used one is the Occupational Stress Indicator (OSI), (Cooper and Payne, 1978; Cooper et al., 1988) that is based partly on the PE fit model and incorporates all of the dimensions in psychosocial work conditions.

During later years, considerable efforts have been made to explore the relationship between theoretical job stress models and 'hard endpoints' for health development. Previously scientific literature was criticized because most of the relationships had been observed in studies using self-reports both in descriptions of the stressors and the health outcomes (Kasl, 1982). There may be a risk in such studies that a pattern of complaining may affect both explaining variables and health outcomes. Therefore, studies have included prospective designs as well as objective outcomes (such as myocardial infarction or cardiac death) and objective recordings of stressors at work (through population means for each occupation regarding decision latitude, for instance; or by means of expert ratings of the crucial stressors). The demand–control model has been more extensively studied than the other models from this point of view.

Job Stress Models and Disease

The significance of the demand–control–support model to cardiovascular disease, in particular high blood pressure and myocardial infarction (Schnall et al, 1990; Theorell and Karasek, 1996; Melamed et al., 1998), has been tested in many epidemiological case-control as well as prospective studies. There is considerable support for the hypothesis that the combination of high demands and lack of control is associated with increased risk of developing myocardial infarction before retirement age. A recently published European study of a pooled cohort of almost 200 000 working men and women, showed a significant association between exposure to the combination of high psychological demands (above median for the demand scale) and low decision latitude (below median), and increased risk of developing a myocardial infarction during follow-up (Kivimäki et al., 2012). There were similar findings for men and women and they were robust when adjusted for age, other risk factors, and publication bias. The combination of high demands and low decision latitude was a better predictor than either demands or decision latitude separately. This study has been criticized because of the relatively weak exposure assessment with few questions assessing demand and decision latitude respectively. In addition, there was only one assessment of job strain before the follow-up started. It has been shown

that predictions gain precision if the exposure to job stress is assessed several times during the follow-up period, since participants experience changing work situations. It is therefore likely that the reported job strain-related myocardial infarction risk in the European pooled study (1.2 or 1.3) is lower than the true risk. In the early studies of demand and decision latitude in relation to myocardial infarction risk, the decision latitude was more important than the demand dimension but during recent years the opposite has been found to be true, possibly reflecting a changing working life (Eller et al., 2009). Loss of decision latitude has also been shown to be associated with increased risk of developing a first myocardial infarction in the near future, even after adjustment for other risk factors (Theorell et al., 1998). The support dimension has been less extensively studied than demand and control, but 'iso-strain,' – the combination of poor workplace support and job strain – has been shown to be associated with an increased development of the metabolic syndrome, which is a biological long-term consequence of adverse psychosocial stress and also a condition which increases the risk of developing cardiovascular disease (Chandola et al., 2010).

High blood pressure monitored during activities at work has shown a convincing relationship with job strain (Schnall et al., 1990; Laflamme et al., 1998; Riese et al., 2004). Serum lipids and smoking patterns have not shown consistent relationships with job strain (Theorell et al., 1998). Accelerated blood coagulation and activation of proinflammatory cytokines may be mechanisms that operate in the relationship between job strain and myocardial infarction risk, since studies have shown that high plasma fibrinogen (Tsutsumi et al., 1999; Brunner et al., 1996) and interleukin-6 (Theorell et al., 2000) are associated with low decision latitude and/or job strain.

Variations over time in job strain have been shown to be associated with changes in endocrine parameters and immune system function of relevance to stress theory. For instance, in one study of men, plasma testosterone (which is important to anabolism) decreased and immunoglobulin G increased (indicating increased immune system activity) when job strain increased and vice versa (Theorell et al., 1990a,b). In another study plasma testosterone increased when the total working life situation improved (Grossi et al., 1999).

Studies of the effort–reward model in relation to coronary heart disease (CHD) and its risk factors have also shown convincing relationships. A review by Kivimaki et al. (2006) showed that in men the CHD risk associated with effort–reward imbalance may even be more pronounced than the risk observed for job strain. High blood pressure and atherogenic lipids are associated with effort–reward imbalance. The patterns of association are different for men and women (Peter et al., 1998; Siegrist, 1996).

The PE fit model and related models have been used in studies of cardiovascular risk factors and positive findings have been reported (Cooper and Payne, 1978).

Musculoskeletal disorders such as low back pain and neck–shoulder pain have also been studied in relation to the demand–control model. The findings have been mixed. The relationships depend on the group and the kind of disorder studied. In some studies, social support at work is crucial whereas in others psychological demands or decision latitude may be important. When the associations between musculoskeletal disorder and psychosocial work environment are adjusted for physical load, the associations are often weakened. Among female workers, physical load and adverse psychosocial conditions are frequently correlated (Josephson et al., 1999). Vingård et al. (2000) showed that the prediction of the risk of developing an acute episode of low back pain is improved by combining psychosocial and physical working conditions. Similar findings have been made for neck–shoulder pain episodes (Wigaeus Tornqvist et al., 2001)

Integration of Effort–Reward Imbalance and Demand–Control

There is evidence (Bosma et al., 1998) that the decision latitude component of the demand–control model and the effort–reward imbalance model are contributing independently of one another to the prediction of episodes of CHD. Peter et al. (2002) found that a combination of the effort–reward and demand–control models provided better prediction of myocardial infarction risk than either one separately and also that the patterns were different for men and women. These findings confirm that the models are related to different psychosocial mechanisms linking psychosocial working conditions to health outcome. The models have psychological demands (extrinsic effort) in common but control (decision latitude) and reward are clearly different. A logical step would be to combine the models. Even when resources (control and support) are optimal for the development of good coping strategies in a highly demanding situation, the employees will need reward for high effort, and hence balance between the components is needed.

Most of the relationships that have been observed in this field would be without significance in society if it were not possible to prove that it is possible to improve health by means of changes in working conditions. Accordingly, the literature on job interventions and their possible effects on health is of considerable importance.

Other Job Stress Related Outcomes

Of particular importance to the link between job stress and cardiovascular disease is the development of atherosclerosis. During recent years it has become increasingly common to measure wall thickness in the carotid arteries – intima media thickness (IMT) – since this reveals early as well as late signs of atherosclerosis. In-depth studies of the early stages of atherosclerosis have shown that psychosocial stress is of significance. Sympathetic arousal increases the likelihood of soft plaques in the arterial wall through a complicated physiological process involving proinflammatory cytokines. Effective activation of the parasympathetic system and rapid slowing of the sympathetic system increases the likelihood that these reactions will be counteracted and the soft plaque will disappear without developing into more severe forms of atherosclerosis (De Couck et al., 2012). In consequence, there are studies which have shown that both the demand–control and

effort–reward models are useful in predictions of progression of IMT over time (Eller, 2012).

Another important step in the efforts to explore easily applicable biological stress markers that can be used in practice in the workplace has been the introduction of salivary cortisol measures. There has been great hope that salivary cortisol which mirrors the concentration of free cortisol in plasma would be a very useful indicator. It has been shown that the rise of the salivary cortisol from awakening to half an hour later, is a useful measure of the arousal associated with the day. The findings so far have been mixed (Kristenson et al., 2012) with regard to the usefulness of salivary cortisol variation in general as an index of stress states, with some significant findings for both models in some studies but null findings in several others (Karlson et al., 2012). Paid work is impossible if there is severe HPA regulation disturbance (see above). Most of the work stressors are of a mild nature and large samples are therefore needed for the study of these associations. The interpretations are complicated by the fact that the feeling of being stressed is dampened by a strong cortisol response (Het et al., 2012) and hence simple correlations between self-reported exposure intensity at work and cortisol level cannot be expected. Stressor peaks at work occur unexpectedly and when they occur, the subjects may not report feeling stressed (*because of* the cortisol peak). In a careful study – with a large number of observations – of the daily stressors for neonatal pediatricians and nurses in critical care, objectively assessed events gave rise to cortisol peaks; but almost three fourths of these peaks were not described as stressful by the subjects themselves (Fischer et al., 2000).

Another outcome that is relevant in linking job stress to illness is heart rate variability (HRV) that is recorded by means of continuous ECG recordings. As long as the subject does not have a large number of cardiac arrhythmias, a high variability in heart rate is an index of good health. Subjects with exhaustion and depression have low total HRV. There are also different components in HRV; 'high frequency' variations are reportedly related to the activity in the parasympathetic system whereas the ratio between 'low frequency' and 'high frequency' variations have been assumed to reflect sympatho-adrenal activity. There are many technical problems in interpreting the recordings and it also makes a difference whether recordings are made on a total day–night variability basis or whether short recordings are made during defined external conditions. Both the demand–control and the effort–reward models have been related to HRV. Thus job strain as well as lack of control (decision latitude) and effort–reward imbalance have been shown in several studies to be related to low total HRV, reduced high frequency power, and high ratio between low and high frequency power (Eller, 2012).

A great emphasis in the present survey has been on cardiovascular disease and related outcomes. The rationale behind this is that cardiovascular disease is an important health outcome that is partly determined by stress-related factors. However, many reports of relevance to job stress deal with emotional outcomes such as depression, depressive feelings, anxiety disorders, and anxious feelings as well as emotional exhaustion and burnout. A recent summary of prospective studies by Bonde (2011) concluded that perception of adverse psychosocial factors (predominantly the demand–control and effort–reward models or their components but also job insecurity and organizational justice) is related to an elevated risk of subsequent depressive symptoms or major depressive episodes.

Environmental and Individual Psychosocial Interventions at Work

Very few studies have actually documented the health consequences of psychosocial job interventions. One early study of cardiovascular risk factors (Orth-Gomér et al., 1994) showed that a program aiming at improved decision latitude and social support for the employees was followed by improved ratio between harmful and protective cholesterol (that is associated with reduced CHD risk) in the experimental but not in the comparison group. The findings could be interpreted to mean that job redesign aiming at improved social relationships and decision latitude could lead to decreased cardiovascular risk. Recently Brisson et al. (2006) have performed controlled evaluations of work organization intervention for health care workers. The intervention program was building upon the demand–control and effort–reward models. It was instituted in close collaboration with the participating worksites and adapted to the specific conditions of each site. Significant effects were reported on mental health outcomes. This group has later developed their intervention program and in new reports significant intervention effects in the form of reduced blood pressure have been reported (Trudel et al., 2011).

Experiences in all EU countries have been summarized (Kompier and Cooper, 1999). Evaluations of systematic efforts to improve working conditions by means of increased decision latitude, support, organizational justice, and reward in order to improve health have typically been more common in the Northern countries.

Recent evaluation studies of the effects of programs for improving managers' ability to decrease employee stress have been promising (Theorell et al., 2001; Romanowska et al., 2011).

Ongoing Trends

In an increasingly unstable labor market in the future, employment security will become less common and global competition will increase (Aronsson and Göransson, 1999; Johnson, 2008). In most European countries there has been a clear rise in work intensity since the beginning of the 1990s. There have been pronounced changes in the organization of the public sector. In the Nordic countries, for instance, there has also been a decrease in perceived decision authority since the late 1990s particularly among women working in health care and education. Problems with decision authority have arisen during a period when management philosophy has changed in the direction of more financial emphasis. Psychological demands will continue to rise, both because there is more cognitive demand (memorizing numbers, handling computers, etc.) and more emotional demands due to increased customer

demands. Numbers of collaborative partners are rising because of the increasing specialization of work, and thus there will be an increasing likelihood of complicated emotional communication for everybody.

Studies in several countries have already shown that workers in temporary employment have a bad position in decision processes. There are indications from several countries that differences in health between favored and less favored workers will increase. The importance of effort–reward balance and organizational justice increases in efforts to decrease job stress.

The basis for social support will change gradually since workers on all levels will rotate more frequently between worksites in different projects than previously. Employers will have a special responsibility for the organization of social support in worksites.

See also: Coping with Job Stress; Coronary Heart Disease: Psychosocial Aspects; Justice and Health; Social Epidemiology; Socioeconomic Status and Health; Stress at Work; Unemployment and Mental Health; Workplace Safety and Health.

Bibliography

Alderling, M., de la Torre, B., Forsell, Y., Lundberg, I., Söndergaard, H.P., Theorell, T., 2008. Psychiatric diagnoses and circadian saliva cortisol variations in a Swedish population-based sample (the PART study). Psychotherapy and Psychosomatics 77, 129–131.

Aronsson, G., Göransson, S., 1999. Permanent employment but not in a preferred occupation: psychological and medical aspects, research implications. Journal of Occupational Health Psychology 4, 152–163.

Bonde, J.P.E., 2011. Psychosocial factors at work and risk of depression: A systematic review of the epidemiological evidence. Occupational and Environmental Medicine 65, 438–445. http://dx.doi.org/10.1136/oem.2007.038430.

Bosma, H., Peter, R., Siegrist, J., Marmot, M., 1998. Two alternative job stress models and the risk of coronary heart disease. American Journal of Public Health 88, 68–74.

Brenner, S.O., Levi, L., 1987. Long term unemployment among women in Sweden. Social Science and Medicine 25, 153–161.

Brisson, C., Cantin, V., Larocque, B., Vezina, M., Vinet, A., Trudel, L., 2006. Intervention research on work organization factors and health: research design and preliminary results on mental health. Canadian Journal of Community Mental Health 25 (2), 241–259.

Brunner, E., Davey Smith, G., Marmot, M., Canner, R., Beksinska, M., O'Brien, J., 1996. Childhood social circumstances and psychosocial and behavioural factors as determinants of plasma fibrinogen. Lancet 13, 1008–1013.

Chandola, T., Heraclides, A., Kumari, M., 2010. Psychophysiological biomarkers of workplace stressors. Neuroscience and Biobehavioral Review 35, 51–57.

Charney, D.S., Deutch, A.Y., Krystal, J.H., Southwick, S.M., David, M., 1993. Psychobiological mechanisms of post-traumatic stress disorder. Archives of General Psychiatry 50, 294–330.

Chida, Y., Steptoe, A., 2009. Cortisol awakening response and psychosocial factors: a systematic review and meta-analysis. Biological Psychology 80, 265–278.

Cooper, C.L., Payne, R. (Eds.), 1978. Stress at Work. Wiley, Chichester, UK.

Cooper, C.L., Sloan, S., Williams, S., 1988. Occupational Stress Indicator. http://www.ase-solutions.co.uk.

De Couck, M., Mravec, B., Gidron, Y., 2012. You may need the vagus nerve to understand pathophysiology and to treat diseases. Clinical Science (London) 122, 323–328.

Demerouti, E., Bakker, A.B., de Jonge, J., Janssen, P.P., Schaufeli, W.B., 2001. Burnout and engagement at work as a function of demands and control. Scandinavian Journal of Work, Environment & Health 27, 279–286.

Demitrack, M.A., Dale, J.K., Straus, S.E., Laue, I., Listwak, S.H., Krusi, M.J.P., Chrousos, G.P., Gold, P.W., 1991. Evidence for impaired activation of the hypothalamic-pituitary-adrenal axis in patients with chronic fatigue syndrome. Journal of Clinical and Endocrinological Metabolism 73, 1224–1234.

Eller, N., 2012. Psychosocial Exposures at Work, Physiological Stress Response and Development of Atherosclerosis – Results from Two Cohort Studies (Doctoral thesis). Faculty of Medicine, Copenhagen University, Copenhagen.

Eller, N.H., Netterström, B., Gyntelberg, F., Kristensen, T.S., Nielsen, F., Steptoe, A., Phil, D., Theorell, T., 2009. Work-related psychosocial factors and the development of ischemic heart disease: a systematic review. Cardiology 17, 83–97.

European Commission, 2011. Working conditions and adaptation to change: Study of health issues associated with re-structuring, "Re-structuring survey". Service Contract No VC/2007/0482.

Fischer, J., Calame, A., Dettling, A., Zeier, H., Fanconi, S., 2000. Experience and endocrine stress responses in neonatal and pediatric critical care nurses and physicians. Pediatric Critical Care 28, 3281–3287.

Frankenhaeuser, M., 1980. Psychoneuroendocrine approaches to the study of stressful person–environment transactions. In: Selye, H. (Ed.), Selye's Guide to Stress Research, vol. 1. Van Nostrand Reinhold, New York.

Grossi, G., Theorell, T., Jurisoo, M., Setterlind, S., 1999. Psychophysiological correlates of organizational change and threat of unemployment among police inspectors. Integrated Physiology and Behavioral Science 34, 30–42.

Halford, C., Jonsdottir, I.H., Eek, F., 2012. Perceived stress, psychological resources and salivary cortisol. In: Kristenson, M., Garvin, P., Lundberg, U. (Eds.), The Role of Saliva Cortisol Measurement in Health and Disease. Bentham eBooks, pp. 67–87.

Hallsten, I., 1998. Psykiskt Välbefinnande Och Arbetslöshet (Psychological Health and Unemployment). Arbete och Hälsa 7. National Institute for Working Life Research, Solna, Sweden.

Hasson, D., Theorell, T., Bergquist, J., Canlon, B., 2013. Acute stress induces hyperacusis in women with high levels of emotional exhaustion. PLoS One 8 (1), e52945. http://dx.doi.org/10.1371/journal.pone.0052945. Epub 2 January 2013.

Het, S., Schoofs, D., Rohleder, N., Wolf, O.T., 2012. Stress-induced cortisol level elevations are associated with reduced negative affect after stress: indications for a mood-buffering cortisol effect. Psychosomatic Medicine 74, 23–32.

House, S.J., 1981. Work Stress and Social Support. Addison-Wesley, Reading, MA.

Jahoda, M., 1979. The impact of unemployment in the 1930's and the 70's. Bulletin of the British Psychological Society 32, 309–314.

Janlert, U., Asplund, K., Weinehall, L., 1991. Unemployment and cardiovascular risk indicators. Data from the Monica survey in northern Sweden. Scandinavian Journal of Social Medicine 1, 14–18.

Janlert, U., Hammarström, A., 1991. Alcohol consumption among unemployed youths: results from a prospective study. British Journal of Addiction 87, 703–714.

Johnson, J., Hall, E., Theorell, T., 1989. The combined effects of job strain and social isolation on the prevalence and mortality incidence of cardiovascular disease in a random sample of the Swedish male working population. Scandinavian Journal of Work Environment and Health 15, 271–279.

Johnson, J.V., 2008. Globalisation, workers' power and the psychosocial work environment – is the demand-control-support model still useful in a neoliberal era? Scandinavian Journal of Work Environment and Health (Suppl. 6), 15–21.

Josephson, M., Pernold, G., Ahlberg-Hulten, G., Härenstam, A., Theorell, T., Vingard, E., Waldenstrom, M., Hjelm, E.W., 1999. Differences in the association between psychosocial work conditions and physical work load in female- and male-dominated occupations. MUSIC-Norrtalje Study Group. American Industrial Hygiene Association Journal 60 (5), 673–678.

Kahn, R., 1981. Work and Health. Wiley, New York.

Karasek, R.A., 1979. Job demands, job decision latitude and mental strain: implications for jobs redesign. Administrative Science Quarterly 24, 285–307.

Karasek, R.A., Theorell, T., 1990. Healthy Work. Basic Books, New York.

Karlson, B., Lindors, P., Riva, R., Mellner, C., Lundberg, U., 2012. Psychosocial work stressors and salivary cortisol. In: Kristenson, M., Garvin, P., Lundberg, U. (Eds.), The Role of Saliva Cortisol Measurement in Health and Disease. Bentham eBooks.

Kasl, S.V., 1982. Chronic life stress and health. In: Steptoe, A., Mathews, A. (Eds.), Health Care and Human Behavior. Academic Press, London.

Katz, D., Kahn, R., 1966. Social Psychology of Organizations. Wiley, New York.

Kilbom, A., Armstrong, T., Buckle, P., Fine, L., Hagberg, M., Haring-Sweeney, M., Martin, B., Punnett, L., Silverstein, B., Sjogaard, G., Theorell, T., Viikari-Juntura, E., 1996. Musculoskeletal disorders: work-related risk factors and prevention. International Journal of Occupational and Environmental Health 2 (3), 239–246.

Kivimäki, M., Virtanen, M., Elovainio, M., Kouvonen, A., Väänänen, A., Vahtera, J., 2006. Work stress in the etiology of coronary heart disease – a meta-analysis. Scandinavian Jounral of Work Environment and Health 32 (6), 432–441.

Kivimäki, M., Nyberg, S.T., Batty, G.D., Fransson, E.I., Heikkilä, K., Alfredsson, L., et al., 2012. Job strain as a risk factor for future coronary heart disease: collaborative metaanalysis of 2358 events in 197 473 men and women. The Lancet 27 (380), 1491–1497.

Kompier, M., Cooper, C., 1999. Preventing Stress, Improving Productivity. European Case Studies in the Workplace. Routledge, London.

Kristensen, T.S., Bjorner, J.B., Christensen, K.B., Borg, W., 2004. The distinction between work pace and working hours in the measurement of quantitative demands at work. Work and Stress 18, 305–322.

Kristenson, M., Garvin, P., Lundberg, U. (Eds.), 2012. The Role of Saliva Cortisol Measurement in Health and Disease. Bentham eBooks.

Laflamme, N., Brisson, C., Moisan, J., Milot, A., Masse, B., Vezina, M., 1998. Job strain and ambulatory blood pressure among female white-collar workers. Scandinavian Journal of Work Environment and Health 24 (5), 334–343.

Levi, L., 1972. Stress and distress in response to psychosocial stimuli. Acta Medica Scandinavica 191 (Suppl.), 1–151.

Melamed, S., Kristal-Boneh, E., Harari, G., Froom, P., Ribak, J., 1998. Variation in the ambulatory blood pressure response to daily work load—the moderating role of job control. Scandinavian Journal of Work Environment and Health 24 (3), 190–196.

Orth-Gomér, K., Eriksson, I., Moser, V., Theorell, T., Fredlund, P., 1994. Lipid lowering through work stress reduction. International Journal of Behavioral Medicine 1, 204–214.

Pejtersen, J.H., Kristensen, T.S., Borg, V., Bjorner, J.B., 2010. The second version of the Copenhagen Psychosocial Questionnaire. Scandinavian Journal of Public Health. 38 (Suppl. 3), 8–24. http://dx.doi.org/10.1177/1403494809349858.

Peter, R., Alfredsson, L., Hammar, N., Siegrist, J., Theorell, T., Westerholm, P., 1998. High effort, low reward, and cardiovascular risk factors in employed Swedish men and women: Baseline results from the WOLF study. Journal of Epidemiology and Community Health 52 (9), 540–547.

Peter, R., Siegrist, J., Hallqvist, J., Reuterwall, C., Theorell, T., The SHEEP Study Group, 2002. Theory and methods. Psychosocial work environment and myocardial infarction: improving risk estimation by combining two complementary job stress models in the SHEEP Study. Journal of Epidemiology and Community Health 56, 294–300.

Riese, H., Van Doornen, L.J., Houtman, I.L., De Geus, E.J., 2004. Job strain in relation to ambulatory blood pressure, heart rate and heart rate variability among female nurses. Scandinavian Journal of Work Environment and Health 20 (6), 477–485.

Romanowska, J., Larsson, G., Eriksson, M., Wikström, B.-M., Westerlund, H., Theorell, T., 2011. Health effects on leaders and co-workers of an art-based leadership development program. Psychotherapy and Psychosomatics 80, 78–87.

Rubin, R.T., Poland, R.E., Lesser, I.M., Winston, R.A., Blodget, A.L., 1987. Neuroendocrine aspects of primary endogenous depression. I. Cortisol secretory dynamics in patients and matched controls. Archives of General Psychiatry 44, 328–336.

Schnall, P., Schwartz, J.E., Landsbergis, P.A., Warren, K., Pickering, T., 1990. The relationship between job strain, workplace, diastolic blood pressure and left ventricular mass index. Journal of the American Medical Association 263, 1929–1935.

Selander, J., Bluhm, G., Nilsson, M., Hallqvist, J., Theorell, T., Willix, P., Pershagen, G., 2 October 2012. Joint effects of job strain and road-traffic and occupational noise on myocardial infarction. Scandinavian Journal of Work Environment and Health. http://dx.doi.org/10.5271/sjweh.3324 pii: 3324.

Selye, H., 1976. The Stress of Life. McGraw-Hill, New York.

Siegrist, J., 1996. Adverse health effects of high effort/low reward conditions. Journal of Occupational and Health Psychology 1, 27–41.

Theorell, T., Hasselhorn, H.M., Vingård, E., Andersson, B., The MUSIC-Norrtälje study group, 2000. Interleukin 6 and cortisol in acute musculoskeletal disorders: results from a case-referent study in Sweden. Stress Medicine 16, 27–35.

Theorell, T., 2012. Stress reduction programs for the workplace. In: Gatchel, R., Schultz, I. (Eds.), Handbook of Occupational Health and Wellness. Springer Science, New York (chapter 18).

Theorell, T., Karasek, R.A., 1996. Current issues relating to psychosocial job strain and cardiovascular disease research. Journal of Occupational and Health Psychology 1, 9–26.

Theorell, T., Karasek, R.A., Eneroth, P., 1990a. Job strain variations in relation to plasma testosterone fluctuations in working men—a longitudinal study. Journal of Internal Medicine 227 (1), 31–36.

Theorell, T., Orth-Gomer, K., Eneroth, P., 1990b. Slow-reacting immunoglobulin in relation to social support and changes in job strain: a preliminary note. Psychosomatic Medicine 52 (5), 511–516.

Theorell, T., Tsutsumi, A., Hallquist, J., Reuterwall, C., Hogstedt, C., Fredlund, P., Emlund, N., 1998. Decision latitude, job strain and myocardial infarction: a study of working men in Stockholm. American Journal of Public Health 88, 382–388.

Theorell, T., Emdad, R., Arnetz, B., Weingarten, A.-M., 2001. Employee effects of an educational program for managers at an insurance company. Psychosomatic Medicine 63, 724–733.

Trudel, X., Milot, A., Brisson, C., Gilbert-Quimet, M., Vezina, M., Trudel, L., 2011. Blood Pressure Reduction Following Intervention on Psychosocial Work Factors: A Longitudinal Study. Abstract booklet of the European Society of Hypertension conference (oral presentation by Milot A). June. Milan, Italy.

Tsutsumi, A., Theorell, T., Hallqvist, J., Reuterwall, C., de Faire, U., 1999. Association between job characteristics and plasma fibrinogen in a normal working population: a cross sectional analysis in referents of the SHEEP Study. Stockholm Heart Epidemiology Program. Journal of Epidemiology and Community Health 53 (6), 348–354.

Vingård, E., Alfredsson, L., Hagberg, M., Kilbom, Å., Theorell, T., Waldenström, M., Wigaeus Hjelm, E., Wiktorin, C., Hogstedt, C., 2000. How much do current and past occupational physical and psychosocial factors explain of low back pain in a working population? Results from the MUSIC-Norrtälje study. Spine 25 (4), 493–500.

Vinokur, A.D., Price, R.H., Caplan, R.D., 1996. Hard times and hurtful partners: how financial strain affects depression and relationship satisfaction of unemployed persons and their spouses. Journal of Personality and Social Psychology 71, 166–179.

Vinokur, A.D., Price, R.H., Schul, Y., 1995. Impact of the JOBS intervention on unemployed workers varying in risk for depression. American Journal of Community Psychology 23, 39–74.

Wigaeus Tornqvist, E., Kilbom, Å., Alfredsson, L., Hagberg, M., Theorell, T., Waldenström, M., Wiktorin, C., Hogstedt, C., The MUSIC-Norrtälje Study Group, 2001. The influence on seeking care because of neck and shoulder disorders from work-related exposures. Epidemiology 12 (5), 537–545.

Occupational Health and Safety, Regulation of

Richard Johnstone, Griffith University, Nathan, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by R. Johnstone, W. G. Carson, volume 16, pp. 10835–10839, © 2001, Elsevier Ltd.

Abstract

This article traces the legal regulation of the health and safety of workers and others at work, from the early nineteenth century to the widespread reforms in occupational health and safety (OHS) regulation since the 1970s. The emphasis is on Europe, North America, and Australia. The article examines four themes of particular importance in the development of OHS regulatory regimes, wherever found. These are the different styles of OHS standard setting, the variable models of enforcement by state-sponsored inspectorates, and the legal institutions that promote worker participation in OHS. The article concludes with a discussion on the OHS regulation in developing countries.

This article outlines developments in the legal regulation of the health and safety of workers and others at work. It does not purport to offer a comprehensive survey of international occupational health and safety (OHS) practice and theoretical perspectives. Instead, it traces historically the main developments in the field of OHS regulation, with an emphasis on Europe, North America, and Australia. Within this framework, and with a special focus on the movement from command and control to self-regulatory models (or, more accurately, models of enforced self-regulation), the article examines four themes of particular importance in the development of OHS regulatory regimes, wherever found. These are the distinguishable features of OHS standard setting, requirements that duty holders seek expert assistance to improve compliance with OHS standards, the variable models of enforcement, and the relationship between OHS and industrial relations. The article concludes with a discussion on the difficulties of regulating OHS in developing countries.

The Emergence of the Traditional Approach to OHS Regulation

OHS regulation is generally traced back to the early nineteenth-century Factory Acts in the United Kingdom. The earliest enactment was the Health and Morals of Apprentices Act of 1802, essentially an extension of Elizabethan Poor Law. It regulated the working hours and conditions of apprentice pauper children in the water-based rural cotton factories of Northern England. The enforcement provisions relied on informers and visitors appointed by the magistracy from the clergy and justices of the peace, but were generally regarded as completely ineffective (Thomas, 1948; pp. 11–12), largely because of the limited powers given to visitors, magistrates' lack of interest in the factory legislation, and the fact that magistrates were too closely linked to mill owners to enforce the law (Gunningham, 1984; p. 38).

The ineffectiveness of this early factory legislation led to the enactment of the Factory Regulation Act of 1833, which regulated the employment of children in a wide range of textile trades. While the enactment of this law was partly a response to working-class pressure, and partly a response to agitation by middle- and upper-class philanthropists for shorter working hours for children and young people, it was undoubtedly also a result of calls for effective uniform regulation by larger urban manufacturers who had voluntarily introduced shorter working hours and who feared unscrupulous competition from smaller and rural-based manufacturers (Carson, 1974; Marvel, 1977).

The 1833 Act was the first to be enforced by an independent, salaried, factory inspectorate vested with broad powers, including the right of entry to factories and mills; the right to make rules, regulations, and orders; and the power to initiate and hear court proceedings to enforce the Act. Over time, however, these broad powers of inspectors were reduced to inspecting and taking enforcement action – a more 'legalist' regulatory form. This development was initiated by the inspectors themselves, to overcome allegations of class bias by presenting decision-making as a passive, neutral, nondiscretionary, and technical process of implementation of the rules set out in the Act (Field, 1990).

After initially making frequent use of prosecution to enforce the Act, the inspectors soon relied mainly on advice and persuasion to induce compliance, with prosecution a 'last resort.' Bartrip and Fenn (1983) suggest that, in doing so, inspectors were taking a cost-effective approach to enforcement. Prosecutions were time-consuming and costly; over time the ratio of inspectors to factory premises worsened significantly so that each inspector had an increasing number of factories to inspect; inspectors were ambivalent about the degree to which regulatory offenses committed by businesses creating wealth and employment were morally reprehensible; and inspectors were deterred by hostile attitudes of magistrates toward the provisions of the Act.

Taking a broader political economy approach, Carson (1979, 1980) has posed an alternative explanation of the origins of this approach to enforcement. He argues that by 1833 a number of large, urban manufacturers, for various reasons, supported the provisions of the 1833 Act, including strong enforcement by the inspectorate. The inspectors, however, found that their assumption that offenses against the Act were being committed by only a few employers was ill-founded, and that contraventions were widespread, and intrinsic to the prevailing production system (Carson, 1979). The heavy use of prosecution would have entailed "collective

criminalization" of employers "of considerable status, social respectability and … growing political influence" (Carson, 1979; p. 48). As a consequence, the early inspectors developed an approach to enforcement that relied principally on advice and persuasion, and resorted to formal prosecution only in the worst cases and where there was evidence of willful employer disobedience or obstinacy. Although officially criminal offenses, contraventions of the Factory Act were accepted as customary and those responsible were rarely subjected to criminal prosecution (the 'conventionalization' of factory crime; Carson, 1979). This 'ambiguity' of factory crime (Carson, 1980), although historically contingent, has proved to be deeply entrenched ideologically in most of the world's OHS regulatory systems.

The Factories Amendment Act of 1844 extended the coverage of the legislation to the working hours of women and included the first machinery safety provisions. In compliance with inspectors' wishes, it also deleted all references to willfulness in many of the employers' offenses (Carson, 1979; pp. 53–54). The 1833 and 1844 Acts marked the beginning of the traditional approach to the regulation of health and safety at work, in which the government set down highly detailed and technical requirements in strict liability 'specification' standards, and provided for the enforcement of these standards by a state inspectorate that relied principally on informal enforcement responses. This regulatory model was reproduced in other industries in Britain and was transported to the British colonies, including Australia, Canada, India, and New Zealand.

While these developments in the United Kingdom strongly influenced the OHS regulatory provisions of the rest of Europe (see Baldwin and Daintith, 1992), the regulatory approach in each European state was also shaped by national political considerations and the regulatory practices in the more powerful industrializing continental European countries, particularly Germany and France (Walters et al., 2011; p. 23). The German 'dual system' developed in the late nineteenth century. It comprised legislative standards enforced by a state inspectorate and complementary regulation by statutory mutual insurance companies (Schaapman, 2002). The legislative standards were not to be found in a specific OHS statute or code, but rather in paragraph 120(a) of the general Industrial Code (*Reichsgewerbeordnung*) for industry and paragraph 62 of the Business Code for commerce (Walters et al., 2011; p. 24). The first German Labor Inspectorate (established in 1853) was an optional service, and only in 1878 was an obligatory government inspectorate established to monitor and enforce the provisions of the Codes. In tandem with this state regulation was a second track of regulation and enforcement, initially enacted by the Industrial Accident Insurance Act of 1884, which created accident insurance funds (*Unfallversicherungsträger*) of which industrial employers' liability insurance funds (the *Berufsgenossenschaften*) were and are the most important. These bodies were responsible not only for compensation functions, but also for the prevention of illness and injury at work by making regulations to prevent injury and disease and enforcing those regulations through technical inspection services (Schaapman, 2002).

Subsequent development of the OHS regulatory provisions in Europe (and in the British colonies) in the first half of the twentieth century was piecemeal, in that detailed and technical standards ('prescriptive' or 'specification' standards) were introduced from time to time to address particular hazards, processes, or premises. One significant development during this period happened in Sweden, where in 1938 the employers and trade unions negotiated the Saltsjobaden Agreement to resolve a long period of industrial conflict. This agreement was a foundation stone in the development of Swedish corporatism, and in the development of participatory approaches to OHS at national, sectoral, and workplace levels (Walters et al., 2011; pp. 26, 122–123). Another important development was the introduction of occupational health services in some European countries, beginning with Norway in the late nineteenth century, and legislation to require firms to make use of occupational health services in addressing OHS issues. For example, France in 1946 passed legislation introducing occupational health services.

In the United States, similar models of OHS regulation were established at the state level from the late 1860s until the end of the nineteenth century. The state level statutes contained detailed, technical standards that focused on certain kinds of premises (such as factories), processes, or activities. The federal government relied on state regulation until 1970, when the Occupational Safety and Health Act (OSH Act) was enacted (Ashford, 1976; Noble, 1986).

Criticisms of the Traditional Command and Control Approach

The weaknesses of this traditional approach are well known (see Robens, 1972) and are widely regarded as having international relevance, because most OHS regulatory systems relied on detailed, technical prescriptive or specification standards enforced by state inspectorates. This traditional approach frequently resulted in a mass of detailed and technical rules, often difficult to understand and difficult to keep up to date. Standards were developed ad hoc to resolve problems as they arose, and concentrated mainly on factory-based physical hazards, resulting in uneven coverage across workplaces. Specification standards did not encourage or even enable employers to be innovative and to look for cheaper or more cost-efficient solutions. They also ignored the now well-accepted view that many hazards do not arise from the static features of the workplace, but from the way work is organized.

Critics have also pointed to the inadequacy of state approaches to enforcement, and to the ideological role of traditional factory legislation in encouraging dependency on state measures and in separating OHS from the conflictual terrain of industrial relations. This separation impedes the involvement of workers and their representatives in developing standards and measures for improving health and safety at the workplace. As noted previously, one exception to this trend has been some of the Scandinavian countries, most notably Sweden, which, from early in the twentieth century, has provided for trade union involvement in OHS (Navarro, 1983).

Since 1970, there have been major developments in the regulation of OHS. The general trend has been away from traditional 'command and control' models toward a more 'self-regulatory' system in which duty holders are given more

flexibility and choice of methods to achieve OHS standards. Under this approach there is greater opportunity for worker involvement in OHS, and approaches to state inspection focus as much on systems of work as on workplace hardware. In many countries, inspectors have been given a wider range of enforcement powers, including administrative sanctions intended to remedy hazards rather than punish offending duty holders.

The Modern Approach to OHS Regulation

Standard Setting

Since the mid-1970s, the trend in Europe, North America, and Australasia, and many countries in Asia, Africa, and South America, has been to replace or supplement specification standards with a combination of general duties, performance standards, process requirements, and documentation requirements. The trend has been most marked in Europe, Canada, and Australia, where the general duties increasingly have become process standards, and the subordinate legislation underpinning the general duties – regulations and guidance material – has increasingly adopted process, performance, and documentation standards.

The 'general duty' provisions require the duty holder (usually the employer, but in some countries including the self-employed, occupiers of workplaces, and manufacturers, suppliers, and designers of plant, substances and structures) to exercise reasonable care in the provision and maintenance of systems of work so as to ensure that workers are not exposed to health and safety risks. For example, the U.S. OSH Act of 1970 section 5 specifies that in all cases not covered by specific standards promulgated under the Act, the employer has a general duty to "furnish to each of his employees, employment and a place of employment which are free from recognized hazards that are causing or are likely to cause death or serious physical harm to his employees."

As this example shows, the post-1970s trend has been to base the general duties on the 'employment paradigm' – that is, the relationship between employers and employees. The dominance of this paradigm has been challenged since the 1980s by changes in work arrangements and, consequently, work relationships around the world, with employers increasingly outsourcing work to contractors and subcontractors, contracting with agencies in order to lease labor, or establishing franchise arrangements (see Walters et al., 2011; pp. 40–42). Research has documented the negative effect of these various work arrangements on the health, safety, and well-being of workers engaged in those arrangements, as measured by injury rates, disease, hazardous substance exposures, mental health, and work health and safety knowledge and compliance (see, for example, Quinlan et al., 2001a; b). This has caused some countries to move away from placing duties on 'employers' in relation to 'employees.' For example, in 2009, the Australian Commonwealth, state, and territory governments endorsed a Model Work Health and Safety Bill in which the primary duty holder is not the employer, but rather "a person conducting a business or undertaking," and this duty is owed to all types of workers 'engaged,' 'caused to be engaged,' 'influenced,' or 'directed' by the person conducting a business or undertaking. Further, a 'worker' is defined very broadly to include any person who carries out "*work in any capacity for* a person conducting a business or undertaking," and includes, for example, work as a contractor or subcontractor, work under a labor leasing agreement, or work as a volunteer.

The general duty provisions tend to be supplemented by a range of standards set out in subordinate legislation, which, increasingly in many countries, has moved away from specification standards to general duty approaches, and performance, process, and documentation standards.

Instead of telling duty holders exactly how they are to achieve compliance, performance standards define the duty holder's duty in terms of goals they must achieve, or problems they must solve, and leave it to the initiative of the duty holder to work out the best and most efficient method for achieving the specified standard.

Process requirements (sometimes known as 'management standards') prescribe a process, or series of steps, that must be followed by a duty holder in managing specific hazards, or OHS generally. They are often used when the regulator has difficulty specifying a goal or outcome, but has confidence that the risk of illness or injury will be significantly reduced if the specified process is followed. Typically, these standards require employers "to take an active, comprehensive, programmatic and enduring responsibility for OHS quality, through a systematic managerial process to detect, abate and prevent workplace hazards" (Frick et al., 2000; p. 1). For example, as a result of the EC Framework Directive of 1989, On the Introduction of Measures to Encourage Improvements in the Safety and Health of Workers at Work, all EC member states are required to introduce risk assessment processes into their OHS statutes. Consequently, the general duty provisions in most EC countries are couched as risk management standards (see Walters et al., 2011; pp. 32–40).

Process-based standards have spawned greater reliance on documentation requirements. Increasingly, OHS statutes are requiring duty holders to document measures they have taken to comply with process-based standards, performance standards, and general duty standards. An example is the requirement for health and safety plans and files in the 1992 EC Directive to Implement Minimum Health and Safety Requirements at Temporary or Mobile Construction Sites.

A current debate in OHS standard setting is whether OHS regulators can take measures to require employers and other duty holders to go beyond compliance with minimum standards, and to implement comprehensive OHS management systems that involve continuous improvement and a systems-approach to OHS management (see Frick et al., 2000; Gunningham and Johnstone, 1999). The debate is most advanced in provisions for the management of major hazardous facilities, which have largely been enacted in reaction to major disasters, and in the mining industry. Many jurisdictions now require facility operators to submit a 'safety case' for each of its installations, demonstrating that they have evaluated risks and set up a system of management and control to deal with the identified risks (Cullen, 1990; Gunningham, 2007; Kaasen, 1991; see the International Labour Organization's 1993 Convention on the Prevention of Major Industrial Accidents, and EU Directives 82/501/EEC and 96/82/EEC on the control of major hazard facilities).

The duties discussed thus far are duties imposed on the enterprise – usually a corporate entity, but occasionally a sole proprietor, partnership, voluntary association, or similar arrangement. A corporation is a fictional legal entity, and can operate only through its human directors, managers, and workers. Most contemporary OHS statutes impose duties of care on employees (and occasionally on other kinds of workers, who may not technically be 'employees') to ensure their own health and safety and the health and safety of others. Senior management leadership in OHS is critical to positive health and safety outcomes. Accordingly, some countries have also imposed duties on high-level corporate officers (who may also be employees and thus subject to the employee's duty as well).

For example, section 37 of the Health and Safety at Work etc Act of 1974 (UK) provides that where a body corporate's offense under the Act is proved to have been committed with the consent or connivance of, or to have been attributable to any neglect on the part of, any director, manager, secretary, or other officer of the body corporate, that officer is also guilty of the offense. This form of accessorial liability can also be found in Hong Kong, New Zealand, and the Australian state of Western Australia. Other statutes, such as those in Singapore or Malaysia, the Canadian province of British Columbia, and the Australian states of Victoria and South Australia, impute liability to an officer for a corporate OHS offense, and might give the officer a defense of due diligence, or only impose liability if the officer fails to exercise reasonable care. Some countries – for example, Germany, the Canadian province of Ontario, and the Australian jurisdictions adopting the Model Work Health and Safety Bill – are now imposing 'positive' duties on corporate officers to take reasonable care or to exercise due diligence. Because these latter duties are positive duties to ensure that the corporation complies with its OHS duties, inspectorates can take measures such as administrative sanctions if the officer fails to exercise due diligence or take reasonable care to ensure corporate compliance. In some countries, duties are imposed on individuals within the corporation rather than on the corporation itself. For example, in France duties are imposed on the Head of the Establishment (a director or a senior manager) and in Sweden, even though the duties are placed on the employer, the responsibility is borne primarily by the most senior manager. (For further discussion of these approaches, see Bergman et al., (2007)). In other countries – for example, the United States and the Netherlands – no provision is made for officer liability.

Provisions Requiring Duty Holders to Seek Expert Assistance in OHS

The modern OHS standards, with their reliance on performance and process (OHS management) standards, are demanding and often complex. While there is plenty of evidence to show that the promotion of systematic OHS management in OHS regulation is a positive development, improving management awareness of and involvement in OHS management (for a summary, see Quinlan et al., 2010; p. 358), it is also clear that duty holders often find it difficult to introduce effective measures to identify, assess, and eliminate or minimize OHS risks (see Walters et al., 2011; pp. 8–9). This is especially the case with small businesses (Hasle et al., 2009). Occupational health services can play an important role in supporting employers and improving their understanding of OHS matters. In the 1970s, Denmark, Sweden, and Norway, and in 1996 the Netherlands, legislated to require firms to engage occupational health services. Germany in 1974 required all firms to appoint occupational physicians and occupational safety specialists. The EC Framework Directive (1989) requires employers to enlist "competent services or persons" to assist in the implementation of the OHS duties in the Directive and "to organize protective and preventive measures." For example, consistent with provisions in other continental European nations, the Swedish Work Environment Act requires employers to provide an occupational health service. Most large companies have their own occupational health services, while joint occupational health service centers provide services to small- and medium-sized organizations. This continental tradition is not followed elsewhere, although UK regulations and some Australian and Canadian statutes require the appointment of 'competent persons' (and until 2012 Queensland required the appointment of workplace health and safety officers in large firms) to assist employers in implementing their OHS obligations.

Worker Participation

The nineteenth-century model of OHS regulation typically did not require worker involvement in OHS. There are, however, good reasons for involving workers in OHS. From an ethical perspective, workers have a direct interest in OHS because they run the major risks of injury if the law fails to protect them. Moreover, workers often know more about the hazards associated with their workplace than anyone else, because they regularly work with them. Where process standards require hazards at work to be identified and evaluated, workers' experience and knowledge is crucially important in successfully completing both of these tasks. Further, research from many countries demonstrates the positive benefits of worker participation in work health and safety, including a relationship between workplace structures enabling worker representation (union presence, joint safety committees, or worker/union safety representatives) and both objective indicators of OHS performance (such as reduced injury rates or hazard exposures) and better OHS management arrangements (see Walters and Nichols, 2007; Chapter 2).

European, Canadian, and Australian OHS statutes have made some statutory provision for OHS committees, and for worker health and safety representatives. Representatives generally have the right to inspect workplaces, to be consulted over OHS issues, and to be provided with OHS information. Some European, Australian, and Canadian statutes go further, and give elected health and safety representatives the right to require compliance with OHS regulatory provisions and to direct that dangerous work ceases. The EC Framework Directive requires employers to involve workers and their representatives in the implementation of the various duties set out in the directive. The U.S. OSH Act to date has made no provision for worker participation in OHS, although the Occupational Safety and Health Administration's (OSHA's) recent proposed safety and health program rule requires worker involvement in the development of a safety and health program.

In contrast, in Scandinavia, the worker participation elements of the OHS regulatory system derive from the general industrial relations system, and do not depend on the OHS statutory framework. For example, the Swedish Work Environment Act of 1977 enshrines the concept that worker control is a critical aspect of a healthy working environment, and provides health and safety representatives with extensive powers, including the right to halt dangerous processes and to participate in the planning of new work premises and processes. Industrial and local agreements regulate much of the administration of OHS. In Sweden, Spain, Italy, and Norway, the construction industry makes provision for coverage of small workplaces by enabling unions to nominate regional safety representatives for small workplaces (normally under 50 workers) to support the OHS initiatives by employers (see further Frick, 2009).

State Enforcement of OHS Obligations

The modern OHS inspectorates can call on a range of enforcement powers, usually culminating in prosecution. In most countries, inspectorates have a broad discretion to determine enforcement action, including negotiating compliance using informal processes such as education, advice, and persuasion. Since the 1970s, inspectorates in most countries have been given administrative sanctions such as improvement and prohibition notices and on-the-spot fines (usually imposing small fixed fines). In some countries inspectorates have the power to impose significant variable administrative penalties. For example, German Insurance Association inspectors can impose fines determined by reference to the amount of money that the firm has saved through noncompliance (Fooks et al., 2007; p. 43). In British Columbia, the basic penalty is assessed using multipliers relating to the size of the employer's assessable payroll, and can be adjusted up to 30 percent by reference to a range of aggravating and mitigating factors, and up to 24-fold if it is a repeat of a contravention committed within the previous three years; it can be even higher if willful or reckless contraventions result in death (see Fooks et al., 2007; pp. 43–44). In Australia since 2003 some OHS regulators have had the power to accept 'enforceable undertakings' in lieu of launching a prosecution for a contravention of the OHS statute. Enforceable undertakings are promises enforceable in court, and are usually offered by an individual or firm who has allegedly breached the law, and accepted by a regulator. If contravened, the undertaking is enforceable in court, often with additional penalties for the contravention of the undertaking (see Johnstone and King, 2008).

In the United States, enforcement is less discretionary, and principally involves administrative sanctions (citations imposing financial penalties and requiring the abatement of hazards) (see Kalestsky, 1997; Shapiro and Rabinowitz, 1997), with prosecution for a limited number of intention-based offenses. In other countries, the OHS inspectorates have been criticized for not adopting a sufficiently strong approach to enforcement, and for relying too heavily on advisory and persuasive approaches. Some commentators argue for a more vigorous approach to enforcement using the self-regulatory principles of the enforcement pyramid (a hierarchy of escalating sanctions) promoted by many regulatory theorists (see Ayres and Braithwaite, 1992; Chapter 2; Rees, 1988; see also Wright et al., 2004; p. vi). The challenge is to develop enforcement strategies that deter the worst offenders, while at the same time encourage and help most employers comply voluntarily. Such an approach to enforcement would require the availability of larger financial penalties and an array of new corporate sanctions (Gunningham and Johnstone, 1999; Chapter 4).

Inspection Approaches Aimed at Promoting OHS Management Systems

In some countries, OHS inspectorates have developed enforcement strategies that focus as much on OHS management as on workplace hardware. For example, in the United States, some OSHA inspection programs abbreviate inspections if employers can show that they have effective safety and health programs. Under the focused inspection program, piloted in the construction industry, if an OSHA inspector conducts an inspection and finds an effective safety program operating on site, then the remainder of the inspection will be limited to the top four hazards that kill workers in the construction industry. If these hazards are well controlled, the inspector concludes the inspection. If not, the inspector conducts a complete site inspection.

Such approaches to inspection are also increasingly being used in Europe and Scandinavia (see Walters et al., 2011). In Sweden, the Swedish Work Environment Authority inspectors routinely inspect 'Systematic Work Environment Management' (SWEM) when they visit workplaces – both by asking management about SWEM and by physical examination of OHS risks at the workplace (see Walters et al., 2011; Chapters 6 and 7).

Recent research suggests, however, that despite the regulatory shift in many countries to OHS management standards, and the increased focus of inspectors on inspecting systematic OHS management, "in most cases these changes have not resulted in clearly defined inspection strategies to achieve improved" OHS management, nor "have they given rise to sustained changes in regulatory inspection practice" (Walters et al., 2011; p. 308).

OHS Regulation in Developing Countries

The discussion so far has addressed the development of OHS regulation in developed countries. OHS and the regulation of OHS in African, South and Central American, and Asian and Pacific island countries face many obstacles not found in developed countries. In many developing countries these challenges include the rapid expansion of the informal economy, where economic activity is not recognized or protected by state authorities; the growth of new industries, some of which are difficult to regulate; and the outsourcing of production by the use of transnational supply chains enabling firms based in developed countries to exploit labor in developing countries from a distance (see Johnstone, 2012). At a broad level, OHS in developing countries is often competing with other priorities, for example, poverty alleviation, HIV/AIDS, and water and sanitation issues (Gutierrez, 2000). OHS regulatory measures can be undermined by inadequate levels

of education and deeply entrenched social and gender inequality, persistent poverty, threats to peace and security, and deregulation and reduced state intervention in the economy (World Health Organization, 2005; p. 20). Further, the contemporary approach to OHS regulation in developed countries demands significant resources from firms, unions, and regulators, resources that generally are not available to regulators, unions, and local small and medium-sized businesses in developing countries.

Many of these issues are a result of intensified globalization, and in particular the ability of capital to shift production to developing countries by direct investment in plant in developing countries, through supply chains or through purchasing contracts. The protection of workers' health is consequently a global issue, and needs global solutions built on cooperation between developed and developing countries (Johnstone, 2012). This requires consideration of matters such as the political, social, and economic dynamics leading to poor OHS in developing countries; how informal sectors might be effectively 'formalized' and properly regulated; how legally enforceable supply chain regulation in developed countries might be extended globally and greater cross jurisdictional cooperation be encouraged; and how the major international institutions, and international collaboration between developing and developed countries, can address the underlying factors leading to poor OHS in developing countries, and improve resources for OHS, and OHS regulation, in those countries (Johnstone, 2012).

See also: Employment and Labor Regulation; Job Design and Evaluation: Organizational Aspects; Occupational Health; Workplace Safety and Health.

Bibliography

Ashford, N.A., 1976. Crisis in the Workplace: Occupational Disease and Injury. MIT Press, Cambridge, MA.

Ayres, I., Braithwaite, J., 1992. Responsive Regulation: Transcending the Deregulation Debate. Oxford University Press, New York.

Baldwin, R., Daintith, T. (Eds.), 1992. Harmonization and Hazard: Regulating Health and Safety in the European Workplace. Graham & Trotman, London.

Bartrip, P.W.J., Fenn, P.T., 1983. The evolution of regulatory style in the nineteenth century British factory inspectorate. Journal of Law and Society 10, 201.

Bergman, D., Davis, C., Rigby, B., 2007. International Comparison of Health and Safety Responsibilities of Company Directors. HSE Books, London.

Carson, W.G., 1974. Symbolic and instrumental dimensions of early factory legislation. In: Hood, R.G. (Ed.), Crime, Criminology and Public Policy. Heinemann Educational, London.

Carson, W.G., 1979. The conventionalisation of early factory crime. International Journal of the Sociology of Law 7, 37–40.

Carson, W.G., 1980. The institutionalization of ambiguity: early British Factory Acts. In: Geis, G., Stotland, E. (Eds.), White Collar Crime: Theory and Research. Sage Publications, Beverly Hills, CA.

Cullen, W.D., 1990. Piper Alpha Inquiry. HMSO, London.

Field, S., 1990. Without the law? Arthurs and the early factory inspectorate. Journal of Law and Society 17, 445–468.

Fooks, G., Bergman, D., Rigby, B., 2007. International Comparison of (a) Techniques Used by State Bodies to Obtain Compliance with Health and Safety Law and Accountability for Administrative and Criminal Offences and (b) Sentences for Criminal Offences, RR607. Health and Safety Executive, London.

Frick, K., Jensen, P.L., Quinlan, M., Wilthagen, T. (Eds.), 2000. Systematic Occupational Health and Safety Management: Perspectives on an International Development. Elsevier, Amsterdam.

Frick, K., 2009. Health and safety representation in small firms: a Swedish success that is threatened by political and labour market changes. In: Walters, D., Nichols, T. (Eds.), Workplace Health and Safety: International Perspectives on Worker Representation. Palgrave MacMillan, Basingstoke, UK.

Gunningham, N., 1984. Safeguarding the Worker: Job Hazards and the Role of the Law. Law Book Company, Sydney, Australia.

Gunningham, N., 2007. Environmental regulation and non-state law: the future public policy agenda. In: Van Schooten, N., Vanschuren, J. (Eds.), International Governance and Law: State Regulation and Non-State Law. Edward Elgar, Cheltenham, UK.

Gunningham, N., Johnstone, R., 1999. Regulating Workplace Safety. Systems and Sanctions. Oxford University Press, Oxford, UK.

Gutierrez, E., 2000. Workers' health in Latin America and the Caribbean: looking to the future. Perspectives in Health 5 Available at: http://www.paho.org/English/DPI/Number10_article1.htm (accessed 27.05.10.).

Hasle, P., Kines, P., Anderson, L.P., 2009. Small enterprise owners' accident causation attribution and prevention. Safety Science 47 (1), 9–19.

Johnstone, R., 2012. Informal sectors and new industries: the complexities of regulating occupational health and safety in developing countries. In: Fudge, J., McCrystal, S., Sankaran, K. (Eds.), Challenging the Legal Boundaries of Work Regulation. Hart Publishing, Oxford, Oxford.

Johnstone, R., King, M., 2008. A responsive sanction to promote systematic compliance?: enforceable undertakings in occupational health and safety regulation. Australian Journal of Labour Law 21, 280–315.

Kaasen, K., 1991. Post Piper Alpha: some reflections on offshore safety regimes from a Norwegian perspective. Journal of Energy and Natural Resources Law 9, 281.

Kalestsky, R., 1997. OSHA Inspections. Preparation and Response. McGraw-Hill, New York.

Marvel, H.P., 1977. Factory regulation: a reinterpretation of the early English experience. Journal of Law and Economics 20, 379–402.

Navarro, V., 1983. The determinants of social policy, a case study: regulating health and safety at the workplace in Sweden. International Journal of Health Services 13, 517–561.

Noble, C., 1986. Liberalism at Work: The Rise and Fall of OSHA. Temple University Press, Philadelphia, PA.

Quinlan, M., Bohle, P., Lamm, F., 2010. Managing Occupational Health and Safety: A Multidisciplinary Approach, third ed. Palgrave MacMillan, Sydney.

Quinlan, M., Mayhew, C., Bohle, P., 2001a. The global expansion of precarious employment, work disorganisation, and consequences for occupational health: a review of recent research. International Journal of Health Services 31, 335.

Quinlan, M., Mayhew, C., Bohle, P., 2001b. The global expansion of precarious employment, work disorganisation, and occupational health: placing the debate in a comparative historical perspective. International Journal of Health Services 31, 507.

Rees, J.V., 1988. Reforming the Workplace: A Study of Self-Regulation in Occupational Safety. University of Pennsylvania Press, Philadelphia, PA.

Robens, Lord, 1972. Safety and Health at Work, Report of the Committee, 1970–72. HMSO, London. Cmnd 5034 (Robens Report).

Schaapman, M., 2002. Germany: occupational health and safety discourse and the implementation of the Framework Directive. In: Walters, D. (Ed.), Regulating Health and Safety Management in the European Union: A Study of the Dynamics of Change. P.I.E Peter Lang, Brussels.

Shapiro, S., Rabinowitz, R.S., 1997. Punishment versus cooperation in regulatory enforcement: a case study of OSHA. Administrative Law Review 49, 713.

Thomas, M.W., 1948. The Early Factory Legislation: A Study in Legislative and Administrative Evolution. Thames Bank Publishing, London.

Walters, D., Johnstone, R., Frick, K., Quinlan, M., Baril-Gingras, G., Thébaud-Mony, A., 2011. Regulating Workplace Risks: A Comparative Study of Inspection Regimes in Times of Change. Edward Elgar, Cheltenham, UK.

Walters, D., Nichols, T., 2007. Worker Representation and Workplace Health and Safety. Palgrave Macmillan, Basingstoke.

World Health Organization, 2005. Implementation of the Resolution of Occupational Health and Safety in the African Region, Meeting Report. World Health Organization, Regional Office for Africa, Brazzaville.

Wright, M., Marsden, S., Antonelli, A., 2004. Building an Evidence Base for the Health and Safety Commission Strategy to 2010 and Beyond: A Literature Review of Interventions to Improve Health and Safety Compliance. HSE Books, Norwich.

Occupational Stress Management

Peter Angerer and Andreas Müller, Heinrich Heine Universität Düsseldorf, Düsseldorf, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Work stress as measured by the work stress models of job strain, effort–reward imbalance, or injustice increases the risk for mental, psychosomatic, and cardiovascular disorders. On an individual level, interventions based on the principles of cognitive behavioral therapy are most effective and improve mental health outcomes immediately after the intervention. Little is known about long-term effects, and other outcomes. On an organizational level, interventions may effectively improve working conditions. However, results are heterogeneous as these interventions are highly sensitive to external disturbances, and require a systematic approach for sustainable implementation. The synergy between organizational- and individual-level interventions has not yet been fully used.

Introduction

Due to the changes within the working environment, especially in Europe and North America, some of the most hazardous work-related stressors nowadays are psychosocial stressors. A large body of scientific evidence based on cohort studies, experimental research, controlled randomized studies, and meta-analyses supports this hypothesis. Psychosocial work stressors have a significant impact on mental, psychosomatic, and cardiovascular health (Stansfeld and Candy, 2006; Kivimäki et al., 2007; Ndjaboué et al., 2012; Bambra et al., 2009). Musculoskeletal pain and cardiovascular disorders represent major causes for disabilities, absenteeism, and medical costs. Mental disorders, such as anxiety or depression, are increasing and account for a major burden of disease worldwide. Costs are high due to patients' impaired social functioning and sometimes poor integration into society, an above-average rate of unemployment and, more importantly for the working environment, a substantial reduction of productivity, an increase in absenteeism, or even permanent disability to work. The projection of the WHO Burden of Disease Project (Mathers and Loncar, 2006) implies a rising economic and social burden, which does not even sufficiently account for the burden on the social networks and families of affected patients. Health promotion and prevention of disease by improving work conditions and mitigating the impact of work stress on health must be one of the primary goals for health policy.

In the following article, we will briefly review the impact of work stress on health and outline the two major approaches to occupational stress management, the individual- and the organizational-level approach.

Occupational Stress Models

For the sake of clarity, some important terms related to occupational stress should be explained first.

In occupational contexts, *stress* can be defined as the psychological, physiological, or behavioral *response of a person* to a given work situation. Stress responses are determined by the quality and intensity of the working conditions, the duration of exposure, as well as on the enduring and momentary abilities and appraisals of the employee. Stress responses include not only physiological arousal like increased blood pressure and feelings such as to be challenged, but also anxiety, irritation, and rumination. Long-term consequences of stress on health and well-being mainly depend on the stress level and on the available internal and external resources to return to an equilibrium state. Stress mechanisms and their link to health are described in more detail elsewhere in this volume.

Although stress depends, to some extent, on individual conditions (e.g., Lazarus and Folkman, 1984), occupational stress models postulate that certain constellations of *external working conditions* (work tasks, work organization, social relationships at work, and environmental conditions) have a significant impact on health and well-being of the vast majority of employees. Generally, such occupational stress models distinguish between 'job demands' and 'job resources' considering the differential health-related impact of working conditions (e.g., Bakker and Demerouti, 2007).

Job demands, on the one hand, are all kinds of working conditions that require mental or physical energy. Within the job demands dimension, one can further distinguish between demands that constrain task accomplishment (i.e., *hindrance stressors* like interruptions or role conflicts) and demands that potentially provide the opportunity for personal development and growth (i.e., *challenge stressors* like task variety or complete tasks) (e.g., Podsakoff et al., 2007). This last-mentioned distinction is useful because challenge and hindrance stressors might be associated with health and well-being in a different way: Hindrance stressors might be a risk factor in any case, because they have to be compensated by additional efforts, and might be evaluated by employees as unnecessary or illegitimate (Semmer et al., 2010). Challenge stressors might only be a risk factor in case demands clearly fall below or exceed the capabilities of employees, but they are expected to contribute positively to health and well-being when there is a balanced relation to the capabilities of employees. *Job resources*, on the other hand, are regarded as working conditions that support people in accomplishing their work as well as safeguard

personal integrity and as such promote health and well-being (e.g., autonomy, social support, esteem).

Before describing basic intervention strategies, the most prominent occupational stress models are presented in the following section.

Job Demand-Control-Support Model

The Job Demand-Control (JDC) model established by Karasek (1979) is based on the idea that high job demands (e.g., high workload, time pressure) impair health particularly when employees possess only limited job control (e.g., narrow options to decide when and how to do the work). In contrast, work that includes high demands *and* high job control is expected to have a motivating function and to be conducive to learning. Accordingly, recent meta-analyses of epidemiological studies show that the combination of high job demands and low job control increased the risk of coronary heart disease (Kivimäki et al., 2012) and mental disorders (Stansfeld and Candy, 2006).

Beyond that, social support given through supervisors or colleagues is expected to have a direct stress-reducing and health-promoting effect and, above all, a buffering effect of job demands and job control on health (Johnson and Hall, 1988). Being short on experiencing social support comes with an increase in the risk of mental disorders (Stansfeld and Candy, 2006) and back pain (Hoogendoorn et al., 2000). Social issues should thus be considered when analyzing and designing work tasks (e.g., cooperation requirements), the work environment (e.g., common areas), and work organization (e.g., time buffers for mutual support).

Effort–Reward Imbalance Model

While the JDC model holds that individual control and decision-making opportunities are existential human needs and thus a prerequisite for health, the Effort–Reward Imbalance (ERI) model (Siegrist, 1996) is based on the idea that work can be further perceived as a social exchange process of individual efforts and received reward. High effort is fostered through high job demands (see above). Reward is particularly shown by money, esteem, career opportunities, and job security. An imbalance between high effort and low reward is expected to cause health impairments. There is good empirical evidence for this basic hypothesis of the ERI model, especially in regard to the association with coronary heart diseases (Kivimäki et al., 2006) and with mental disorders (Stansfeld and Candy, 2006).

Organizational Justice

Like the ERI model, organizational justice approaches are related to aspects of a fair distribution of rewards, but additionally encompassing a component of procedural and interpersonal of fairness (Greenberg, 1990). Procedural justice exists if employees are informed about important organizational matters and if decisions are made and implemented after participation of all parties. Relational justice particularly refers to the fair and respectful treatment of employees by their supervisors. Empirically, organizational justice is shown to be related to various health outcomes like cardiovascular disease and cardiovascular mortality as well as psychiatric disorders (Elovainio et al., 2010).

The Job Characteristic Model

The Job Characteristic model (Hackman and Oldham, 1980) specifically describes working conditions that contribute to intrinsically motivating and fulfilling work. The model initially postulated five 'core job dimensions' – skill variety (i.e., the use of different skills in a job), task identity (i.e., the possibility to complete a whole piece of work), task significance (i.e., the impact of the work on others' lives), feedback from the job (i.e., information about the own performance), and autonomy (i.e., the individual discretion to plan the work). These five core job dimensions operate through three psychological states – experienced meaningfulness, experienced responsibility, knowledge of results – on important work-related aspects of health like work motivation, job satisfaction, and low absenteeism. A recent meta-analysis (Humphrey et al., 2007) mainly supports these assumptions and adds further social and work context characteristics to the model.

Beyond work motivation, the Job Characteristic model also refers to basic processes of human action regulation, which especially benefit from autonomous goal setting and action planning and the accomplishment of complete tasks (e.g., Hacker, 2003). Thus from this model – just like the JDC model – it can be concluded that healthy work does not simply mean reducing job demands but rather setting the right demands.

Interventions to Reduce Occupational Stress

Individual-Level Interventions

Individual-level interventions target persons and change of individual behavior rather than work conditions. Thus, occupational stress according to the above mentioned stress models is usually not tackled, with few exceptions (e.g., Limm et al., 2011). Programs with the aim of better management of work stress are commonly called stress management interventions (SMIs). To date there is a large number of studies on the effects of SMI. A recent meta-review identified 27 meta-analyses and systematic reviews including 499 studies that examined the effects of SMI on common mental disorders (Bhui et al., 2012). Thus, there is a need to distinguish. SMIs can be classified in several ways:

Targets: In the context of work, individual stress management may be conceptualized as a program that focuses on change of behavior on different levels: A person may change work conditions, or personal appraisal of the conditions, or coping with emotional and physical reactions to the conditions (Ivancevich et al., 1990).

To give an example, a high effort and low reward situation can lead to stress reactions in the sense of emotional or cognitive irritation (Mohr et al., 2005), exhaustion, and sleep disturbance. To improve the situation, the affected person could try to decrease the effort or increase the reward by negotiating tasks and work conditions with managers

and colleagues. Alternatively, in a highly competitive profession, he or she could change appraisal and attitude by accepting the effort–reward imbalance situation as temporarily unavoidable and by perceiving the effort as a career strategy, which will lead to increased reward in the future (Siegrist and Marmot, 2004). Eventually, the person could use relaxation techniques or sports to counterbalance the stress reactions.

Intervention techniques: A common way to classify SMI is by the intervention technique. Interventions based on the principles of cognitive behavioral therapy (CBT) instruct participants about the role of their thoughts in eliciting stress reactions to stressful situations, and provide them with the skills to modify thoughts to better cope with stressful work conditions. Their main focus thus is the appraisal of the conditions, but CBT can as well support behavior change toward an active reducing of stressors or enhancing of resources. CBT-based SMI programs aim at empowering participants to actively manage stressful working condition. By this they differ from more passive relaxation or meditation techniques:

Interventions that focus on better coping with stress reactions are often based on the principles of muscle relaxation or meditation. They instruct participants how to reduce muscular tension, emotional arousal, or cognitive pondering due to work stress.

An alternative way is sports. Physical exercise may reduce stress reactions by distracting from the adverse conditions, enhance social support (if performed in groups), reduce muscle tension, stimulate endorphin production, and eventually strengthen emotional stability.

Other intervention techniques use a stress diary to write down, thereby verbalize and reflect stressful events, to identify causes of stress, to become aware of one's own reactions, and to set goals for managing stress.

Training practical skills as time management, problem solving, negotiating, or conflict resolution may reduce common sources of work stress. Interventions that teach more specific skill may serve to better handle stressful situations in special occupations, typically related to education, health services, or public security.

Level of prevention: Another way of conceptualizing interventions is by levels of prevention as primary, secondary, and tertiary prevention. Primary prevention interventions aim at influencing the sources of stress. Work design as an organizational approach of primary prevention is discussed below. However, also on the individual level the target of an intervention may be to alter work conditions. Secondary prevention interventions target stress reactions before they result in health problems. Tertiary prevention interventions target stress-related diseases as depression or anxiety; the approach could be early treatment or vocational rehabilitation (both are beyond the scope of this article). Most commonly SMIs are secondary prevention programs.

The duration of individual-level stress prevention programs usually ranges between several weeks and several months. Interventions may vary with respect to their intended outcomes, which can be psychological (well-being, mental health, etc.), physical (e.g., blood pressure, cortisol), or organizational (productivity, sick leave, presenteeism; work conditions, etc.).

Effectiveness of Individual-Level Interventions

Several recent systematic reviews and meta-analyses have attempted to synthesize the knowledge about the effectiveness of individual-level occupational SMIs (Bhui et al., 2012; Richardson and Rothstein 2008; Martin et al., 2009; LaMontagne et al., 2007; van Wyk and Pillay-van Wyk, 2010). Due to the large number and the considerable heterogeneity of studies, it is difficult to comprehensively summarize the evidence. As the cited reviews and meta-analyses have different aims and perspectives, their results are described separately before an overall conclusion is drawn.

Richardson and Rothstein (2008) included 36 experimental studies in a recent meta-analysis, analyzing 55 interventions with a total sample size of 2847 individuals. The average length of intervention was 7.4 weeks (range from 3 days to 7 months). Interventions were described as cognitive behavioral, relaxation, organizational, multimodal, or alternative. Most frequently, in 69% of interventions, relaxation interventions were tested. The effect of the SMI was measured by 60 different outcomes, mostly based on self-report. Typical outcomes included stress reactions, depression and anxiety, general mental health, and work satisfaction. Physiological outcomes were measured in a minority of studies, mostly by blood pressure registration. Organizational outcomes are even rarer, mostly measured as sick-leave days.

The overall weighted effect size (Cohen's d) for all studies was 0.526 (95% confidence interval 0.364, 0.687), which represents a significant medium to large effect. (The effect size is a statistical measure for the strength of an effect that facilitates comparison between different studies; as a convention, an effect size of 0.2 is regarded as small, 0.5 as medium, and 0.8 as big.)

Subgroup analyses showed that the largest effects were seen for interventions based on cognitive behavioral principles ($d = 1.164$; CI = 0.456, 1.871). If additional treatment components were added, the effect was reduced: The meta-analysis surprisingly finds that the more components were added to a cognitive behavioral intervention, the less effective was the outcome. Furthermore, multimodal in contrast to single component programs appear to lose effect with increasing treatment length.

Relaxation interventions were less effective; they constantly showed medium effect sizes ($d = 0.497$; CI = 0.309, 0.685). In the mixed group of 'alternative' interventions, studies that aimed at increasing participants' resources or management/job skills produced a combined significant large effect ($d = 1.414$, CI = 0.587, 2.241). These SMIs were designed to provide employees with individual tools to assist them to deal with the more stressful aspects of their work. With regard to specific outcomes, bias is introduced as the variables are likely to be chosen on the basis of the type of intervention: Thus the largest effects were found on psychological outcomes since CBT-based interventions only measured psychological, but no physiological outcomes, for which the overall effects were smaller. Productivity was the organizational outcome influenced most strongly.

Taken together, SMI that motivates the participants to actively manage their work stress by modifying or better handling it, or by changing their appraisal, seems to have stronger effects than passive coping with stress reactions by

relaxation, meditation, and alike. Targeted approaches in the sense of single component interventions have greater effects than multimodal multicomponent interventions. The effects are primarily seen in mental health (stress, depression, anxiety) and in productivity. Little is known about long-term effects. Furthermore, simultaneous assessment of various levels of outcome measures (psychological, organizational/process-orientated) in future studies can help to learn more about the mechanisms by which interventions reduce stress.

In a similar meta-analysis, Martin and colleagues (Martin et al., 2009) examined the impact of workplace health promotion programs that were evaluated by using a valid indicator or specific measure of depression or anxiety symptoms. They came to the conclusion that interventions that directly focused on mental health had positive but small effects on depression and anxiety, as did more unspecific, 'indirect' health promotion programs. However, they highlighted three studies with high above-average effects on mental health. These programs all used some kind of CBT approach.

Combining the results of 11 meta-analyses and 12 narrative reviews Bhui and colleagues (2012) draw a comprehensive picture of how SMIs at work influence anxiety, depression, and absenteeism. They confirm the findings outlined in the two meta-analyses cited above by showing that CBT-based SMI consistently produced the largest effect on these outcomes. The researchers could, however, not clearly state whether primary or secondary prevention is more effective and whether absenteeism can be influenced positively.

Systematic reviews for specific occupations were performed for health care workers and for policemen. In a Cochrane review, van Wyk and Pillay-van Wyk (2010) concluded that SMI seems to be effective to reduce job stress and prevent burnout in health care workers, mainly nurses, during the time of the intervention; however, there is insufficient evidence for the effectiveness beyond the intervention period. Thus, the overall effectiveness remains unclear. Similarly, Buchberger et al. (2011) conducted a health technology assessment evaluation studies on health promotion for nurses. They summarized that interventions for psychological health resulted in better stress management, more successful coping with workload, higher communication skills, and advanced training; however, the evidence is insufficient for a definite conclusion. For policemen, Peñalba et al. (2008) did not find studies that had sufficient quality to draw firm conclusions.

A systematic review addressing the socioeconomic results of preventive mental health promotion programs identified 47 studies worldwide. Especially for interventions at the workplace an economic return on investment was found (McDaid and Park, 2011). The cost–benefit ratio is of great importance, since SMI impose additional costs on employers or the health care system.

Implementation

It seems that only a small part of the recommended health care practices for prevention are implemented (Glasgow and Emmons, 2007). Several barriers have been identified as reasons, among them too demanding procedures for staff and participants, inflexibility, lacking alignment of organization and intervention, low participation or deficits in assessment or evaluation, just to name a few. Barriers can be manifold, but are most often found in the interaction of work setting and intervention in a lacking 'match.' An iterative closely intertwined process involving all partners in all stages of the process proposes a possible solution to this problem (Glasgow and Emmons, 2007). Hence, the integration of all stakeholders is essential for facilitating transfer and implementation as well as realization of measures (Driessen et al., 2010).

Organizational-Level Interventions

Organizational-level interventions (OLI) aim at the improvement of working conditions, i.e., the *structural* preconditions for health and well-being at the workplace. More specifically, according to the above introduced terminology, hindrance stressors shall be removed, a balance between challenge stressors and capabilities of employees shall be established, and job resources shall be maximized.

With that, OLI put specific emphasis on the primary prevention of disease (i.e., they eliminate stressful working conditions and establish job resources) but they also incorporate elements of secondary and tertiary prevention (i.e., they support more effective coping with stressful events, provide support, and help to minimize detrimental effects of existing health impairments). Due to its broad impact it is hypothesized that OLI are more effective than other approaches to maintain and improve the health and well-being of employees (LaMontagne et al., 2007).

Effectiveness of OLI

Although there is a strong emphasis on the implementation of OLI in occupational stress management systems, there are only few published studies on OLI with reasonable methodological quality, i.e., studies that used a control group to attribute observed effects on the intervention and not on confounding factors. Correspondingly, the evidence on the effectiveness of OLI is mixed: as cited above, Richardson and Rothstein (2008) conducted a systematic investigation of studies on controlled occupational SMIs. Out of the included 36 studies only 5 were OLI, which had a small and insignificant effect overall. LaMontagne et al. (2007) included 18 controlled interventions in their systematic review that exclusively targeted working conditions, or the interface between working conditions and the individual (see Appendix to LaMontagne et al., 2007). About two-third of these studies reported significant positive effects on favorable organizational level outcomes like autonomy or individual-level outcomes like stress reduction.

One reason for these inconsistent findings is that the effectiveness of OLI is highly sensitive to external disturbances, i.e., the same intervention can be highly effective in a supportive context but ineffective in an impedimental environment (i.e., Nielsen and Randall, 2012). This is because OLI often affect many organizational processes and structures and involves numerous stakeholders. Due to this complexity, there is the call for a more sophisticated evaluation of the intervention *process* beyond the mere evaluation of the intervention *effects* to reveal the mechanisms through which the intervention works (Nielsen et al., 2010).

The Implementation of OLI

Regardless of the specific target, the OLI process can be viewed as a cycle with the following critical phases: preparation, risk analysis, prioritization, development of measures, implementation of measures, evaluation, and improvement (e.g., healthy workplaces: WHO global model for action).

Intervention activities should be tailored to the specific needs and problems of the organization. This requires a *thorough risk assessment with established measures* (e.g., Nielsen and Randall, 2012). The adopted measures should capture the previously defined health-related working conditions with suitable scientific quality.

From the beginning *common goals and structured actions plans* need to be stipulated with all participants, as organizational stakeholders often have diverged interests and expectations in respect to OLI (e.g., work council vs management). As OLI frequently touch upon existing occupational areas of responsibility (e.g., occupational safety, quality management, health services, human resources management), implementation should be integrated into existing structures, as far as possible, to avoid redundancy and enable a sustainable implementation of organizational stress management measures.

Most importantly, OLI should only be introduced once *supervisor support* is assured. Of great importance is the structural integration of decision makers in a steering committee.

Participation of Employees

There are abundant approaches and procedures to conduct OLI. However, the indispensable core of all OLI should be the participation of employees: (1) Participation is a health-promoting factor in itself, because it accounts for a more fair organization of work processes and it strengthens employee competence. (2) Employees are experts in their work. Their participation increases the chance of adequate problem diagnosis as well as (3) the development of suitable and effective measures.

'Health circles' are one possibility of establishing participation during OLI (Aust and Ducki, 2004). Health circles are temporary work groups with employees that develop measures for the improvement of working conditions. There are first references for the effectiveness of participatory interventions like health circles (Aust and Ducki, 2004; Bourbonnais et al., 2011). However, there is substantial implementation effort of health circles that should be taken into account as a potential threat, especially so in cases of staff shortage and high workload.

Future Challenges

On the grounds of the knowledge summarized above open questions concern the sustainable implementation of effective occupational stress management into practice, the synergy between individual- and organizational-level approaches, and the long-term effects of occupational stress management on health, especially with respect to mental and cardiovascular diseases.

In the recent years an obvious change of work has occurred, which is, among others, marked by the global economic competition, rapid development of information technologies, blurred borders between time at work and leisure time, and a steadily increasing number of atypical work, like temporary contracts. Additional psychosocial risk factors for health and well-being may have emerged, which have not been considered sufficiently in the above-named classical organizational stress models. Hence, new conceptual and empirical models may be required to capture and systematize the ongoing changes of working and employment conditions and to allow forecasts of their effects on health and well-being.

The workforce in almost all industrialized countries is rapidly aging. Due to these demographic changes, it becomes more important to study occupational life-span models that enable healthy work careers and support employee's attempts to remain productive members of the workforce beyond the current retirement age. Moreover, due to the variety of aging processes, future work organizations have to create individually tailored solutions that enable older employees to compensate for specific age-related losses (e.g., health problems, decreased capacity of working memory) and to use age-related gains more efficiently (e.g., experience).

In conclusion, there is a clear imbalance between theoretical knowledge about the causes and processes of occupational stress on one hand and practical solutions to manage occupational stress on the other hand. To close this gap between theory and practice future research should therefore turn more intensively to the development and testing of effective occupational SMIs.

See also: Coping with Job Stress; Exposure Therapies and Stress Inoculation: A Brief Overview; Occupational Health; Stress and Cardiac Response; Stress at Work; Stress, Coping and Health; Stress: Psychological Perspectives; Teacher Stress and Burnout: Methodological Perspectives; Wellbeing and Burnout in the Workplace, Psychology of; Workplace Safety and Health; Workplace Stress.

Bibliography

Aust, B., Ducki, A., 2004. Comprehensive health promotion interventions at the workplace: experiences with health circles in Germany. Journal of Occupational Health Psychology 9 (3), 258–270.

Bakker, A.B., Demerouti, E., 2007. The job demands-resources model: state of the art. Journal of Managerial Psychology 22 (3), 09–328.

Bambra, C., Gibson, M., Sowden, A.J., Wright, K., Whitehead, M., Petticrew, M., 2009. Working for health? Evidence from systematic reviews on the effects on health and health inequalities of organisational changes to the psychosocial work environment. Preventive Medicine 48 (5), 454–461.

Bhui, K.S., Dinos, S., Stansfeld, S.A., White, P.D., 2012. A synthesis of the evidence for managing stress at work: a review of the reviews reporting on anxiety, depression, and absenteeism. Journal of Environmental and Public Health 2012, 515874.

Bourbonnais, R., Brisson, C., Vezina, M., 2011. Long-term effects of an intervention on psychosocial work factors among healthcare professionals in a hospital setting. Occupational and Environmental Medicine 68, 479–486.

Buchberger, B., Heymann, R., Huppertz, H., Friepörtner, K., Pomorin, N., Wasem, J., 2011. The effectiveness of interventions in workplace health promotion as to maintain the working capacity of health care personal. Health Technology Assessment 7 (Doc06). http://dx.doi.org/10.3205/hta000097. Epub 2011 Sep 28.

Driessen, M., Groenewoud, K., Proper, K.I., Anema, J.R., Bongers, P.M., van der Beek, A.J., 2010. What are possible barriers and facilitators to implementation of a participatory ergonomics programme. Implementation Science 5, 64.

Elovainio, M., Heponiemi, T., Sinervo, T., Magnavita, N., 2010. Organizational justice and health; review of evidence. Giornale Italiano di Medicina del Lavoro ed Ergonomia 32 (3), B5–B9.

Glasgow, R.E., Emmons, K.M., 2007. How can we increase translation of research into practice? Types of evidence needed. Annual Review of Public Health 28, 413–433.

Greenberg, J., 1990. Organizational justice: yesterday, today, and tomorrow. Journal of Management 16 (2), 399–432.

Hacker, W., 2003. Action regulation theory – a practical tool for the design of modern work processes? European Journal of Work and Organization Psychology 12, 105–130.

Hackman, J.R., Oldham, G.R., 1980. Work Redesign. Addison-Wesley, Reading, Massachusetts, MA.

Hoogendoorn, W.E., van Poppel, M.N.M., Bongers, P.M., Koes, B.W., Bouter, L.M., 2000. Systematic review of psychosocial factors at work and private life as risk factors for back pain. Spine 25 (16), 2114–2125.

Humphrey, S.E., Nahrgang, J.D., Morgeson, F.P., 2007. Integrating motivational, social, and contextual work design features: a meta-analytic summary and theoretical extension of the work design literature. Journal of Applied Psychology 92 (5), 1332–1356.

Ivancevich, J.M., Matteson, M.T., Freedman, S.M., Phillips, J.S., 1990. Worksite stress management interventions. American Psychologist 45, 252–261.

Johnson, J.V., Hall, E.M., 1988. Job strain, work place social support, and cardiovascular disease: a cross-sectional study of a random sample of the Swedish working population. American Journal of Public Health 78, 1336–1342.

Karasek, R.A., 1979. Job demands, job decision latitude and mental strain: implications for job redesign. Administrative Science Quarterly 24, 285–306.

Kivimäki, M., Nyberg, S.T., Batty, D.G., Fransson, E.I., Heikkilä, K., Alfredsson, L., et al., 2012. Job strain as a risk factor for coronary heart disease: a collaborative meta-analysis of individual participant data. The Lancet 380 (9852), 1491–1497.

Kivimäki, M., Virtanen, M., Elovainio, M., Kouvonen, A., Väänänen, A., Vahtera, J., 2006. Work stress in the coronary heart disease – a meta-analysis. Scandinavian Journal of Work, Environment & Health 32 (6), 431–442.

Kivimäki, M., Vahtera, J., Elovainio, M., Virtanen, M., Siegrist, J., 2007. Effort-reward imbalance, procedural injustice and relational injustice as psychosocial predictors of health: complementary or redundant models? Occupational and Environmental Medicine 64 (10), 659–665.

LaMontagne, A.D., Keegel, T., Louie, A.M., Ostry, A., Landsbergis, P.A., 2007. A systematic review of the job-stress intervention evaluation literature, 1990–2005. International Journal of Occupational and Environmental Health 13 (3), 268–280.

Lazarus, R.S., Folkman, S., 1984. Stress, Appraisal, and Coping. Springer, New York.

Limm, H., Gündel, H., Heinmüller, M., Marten-Mittag, B., Nater, U.M., Siegrist, J., Angerer, P., 2011. Stress management interventions in the workplace improve stress reactivity: a randomised controlled trial. Occupational and Environmental Medicine 68, 126–133.

Martin, A., Sanderson, K., Cocker, F., January 2009. Meta-analysis of the effects of health promotion intervention in the workplace on depression and anxiety symptoms. Scandinavian Journal of Work, Environment & Health 35 (1), 7–18.

Mathers, C.D., Loncar, D., 2006. Projections of global mortality and burden of disease from 2002 to 2030. PLOS Medicine 3 (11), e442.

McDaid, D., Park, A.L., December 2011. Investing in mental health and well-being: findings from the DataPrev project. Health Promotion International 26 (Suppl. 1), i108–i139.

Mohr, G., Rigotti, T., Müller, A., 2005. Irritation – ein Instrument zur Erfassung psychischer Beanspruchung im Arbeitskontext. Skalen- und Itemparameter aus 15 Studien (Irritation - an instrument assessing mental strain in working contexts. Scale and item parameters from 15 studies). Zeitschrift für Arbeits- und Organisationspsychologie 49 (1), 44–48.

Ndjaboué, R., Brisson, C., Vézina, M., 2012. Organisational justice and mental health: a systematic review of prospective studies. Occupational and Environmental Medicine 69 (10), 694–700.

Nielsen, K., Randall, R., 2012. Opening the black box: presenting a model for evaluating organizational-level interventions. European Journal of Work and Organizational Psychology 13 (5), 601–617.

Nielsen, K., Randall, R., Holten, A.-L., Gonzalez, E.R., 2010. Conducting organizational-level occupational health interventions: what works? Work & Stress 24 (3), 234–259.

Peñalba, V., McGuire, H., Leite, J.R., July 16, 2008. Psychosocial interventions for prevention of psychological disorders in law enforcement officers. Cochrane Database of Systematic Reviews (3), CD005601. http://dx.doi.org/10.1002/14651858.CD005601.pub2.

Podsakoff, N.P., LePine, J.A., LePine, M.A., 2007. Differential challenge stressor-hindrance stressor relationships with job attitudes, turnover intentions, turnover, and withdrawal behavior: a meta-analysis. Journal of Applied Psychology 92 (2), 438–454.

Richardson, K.M., Rothstein, H.R., 2008. Effects of occupational stress management intervention programs: a meta-analysis. Journal of Occupational Health Psychology 13 (1), 69–93.

Semmer, N.K., Tschan, F., Meier, L.L., Facchin, S., Jacobshagen, N., 2010. Illegitimate tasks and counterproductive work behavior. Applied Psychology: An International Review 59 (1), 70–96.

Siegrist, J., 1996. Soziale Krisen und Gesundheit. Hogrefe, Göttingen.

Siegrist, J., Marmot, M., 2004. Health inequalities and the psychosocial environment-two scientific challenges. Social Science & Medicine 58, 1463–1473.

Stansfeld, S., Candy, B., 2006. Psychosocial work environment and mental health – a meta-analytic review. Scandinavian Journal of Work, Environment & Health 32 (6), 443–462.

van Wyk, B.E., Pillay-van Wyk, V., 17 March 2010. Preventive staff-support interventions for health workers. Cochrane Database of Systematic Reviews (3), CD003541. http://dx.doi.org/10.1002/14651858.CD003541.pub2.

Oceania, Sociocultural Overviews: Melanesia

Robert J Foster, University of Rochester, New York, NY, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 14, pp. 9549–9554, © 2001, Elsevier Ltd.

Abstract

Anthropologists continue to use broadly etched contrasts – equality versus hierarchy, gift exchange versus commodity sale – as abbreviations for complex differences between Melanesian social forms and cultural practices, and those of the modern West. These contrasts, although problematic, effectively highlight the transactional nature of Melanesian economies and polities, kinship and social organization; they similarly underscore the malleability of gender identity, personhood, and the body in Melanesia. Recent Melanesianist anthropology has given fuller attention to processes of colonial history and postcolonial change, including Christian missionization, the spread of capitalist markets, and nation-state formation. This research foregrounds the variety of local modernities and creative adaptations that Melanesians have produced since the nineteenth century out of their uneven and contested encounters with agents and institutions of exogenous origin.

The convention of dividing Oceania into three distinct geographic regions or culture areas – Polynesia, Micronesia, and Melanesia – originated in 1832 with the French navigator J-C-S. Dumont d'Urvilles. By this convention, Melanesia today encompasses the independent nation-states of Papua New Guinea, Solomon Islands, Vanuatu, and Fiji; the French overseas territory of New Caledonia (Kanaky); the Indonesian province of Irian Jaya or Papua (West Papua); and the Torres Straits Islands, part of the Australian state of Queensland. These islands comprise about 960 000 km^2 with a total population of more than eight million people. The large island of New Guinea and its smaller offshore islands account for almost 90% of the land area and 80% of the population of Melanesia (see Figure 1).

Melanesian Diversities

The diversity of contemporary Melanesian political arrangements is exceeded by a linguistic, cultural, ecological, and biological diversity that impressed early European explorers and challenged later social and natural scientists. Although mostly pig-raising root-crop farmers, Melanesians occupy physical environments ranging from montane rainforests to lowland river plains to small coral islands; they live in longhouses, nucleated villages, dispersed hamlets, and urban settlements. While racially stereotyped as black-skinned and woolly-haired people, they display a wide range of phenotypic features. Similarly, Melanesian social organizations illustrate numerous anthropologically recognized possibilities with respect to rules of descent, residence, marriage, and kin terminology. Melanesians speak more than 1100 native languages, some with tens of thousands of speakers and others with less than a few hundred. English, French, and Bahasa Indonesia are recognized as official languages; several pidgin-creoles are widely spoken and also recognized as official languages in Papua New Guinea and Vanuatu. Melanesia is thus culturally heterogeneous to the point that its very status as a useful ethnological category is dubious; many anthropologists would assign Fiji to Polynesia and classify Torres Straits Islanders with Australian Aborigines. Ethnographic generalizations about the region often show a marked bias toward cases from some areas (e.g., the New Guinea Highlands) rather than others (e.g., south-coast New Guinea), and in any case remain rare.

Linguists, archeologists, and geneticists offer persuasive narratives of prehistoric migrations that account for Melanesian heterogeneity. These narratives, although not universally accepted, reveal how Dumont D'Urvilles' convention obscures historical and cultural connections between Melanesian and other Pacific populations as well as fundamental differences within Melanesian populations. At least 40 000 years ago, human populations of hunter – gatherers settled the supercontinent of Sahul (Australasia), reaching what is today the island of New Guinea, the Bismarck Archipelago and the Solomon chain by at least 30 000 years ago. Linguistic evidence suggests that these settlers spoke Non-Austronesian (NAN or Papuan) languages, which today number around 750 and occur mainly in the interior of New Guinea, with isolated instances through the islands to the east as far as the Santa Cruz Archipelago. By 3500 years ago, a second migration out of Southeast Asia had brought another population into sustained contact with the original settlers of the Bismarcks. This migration is associated by archeologists with the distinctive Lapita pottery complex, distribution of which suggests that 'Lapita peoples' had by 3000 years ago moved as far south and east as hitherto uninhabited Fiji and western Polynesia. Lapita peoples spoke an early language of the Austronesian (AN or Malayo-Polynesian) family labeled Proto-Oceanic. Austronesian languages are today distributed along coastal eastern New Guinea and through island Melanesia; all Polynesian languages and most Micronesian ones are also descended from Proto-Oceanic. Hence the extraordinary linguistic, cultural, and biological diversity of Melanesia by comparison with Polynesia and Micronesia derives from at least two prehistoric sources: (1) processes of divergence and differentiation among relatively separated populations over the very long period of human occupation in the region, especially on the island of New Guinea and (2) processes of mixing or creolization between the original NAN-speaking settlers and the later AN-speaking arrivals in island Melanesia.

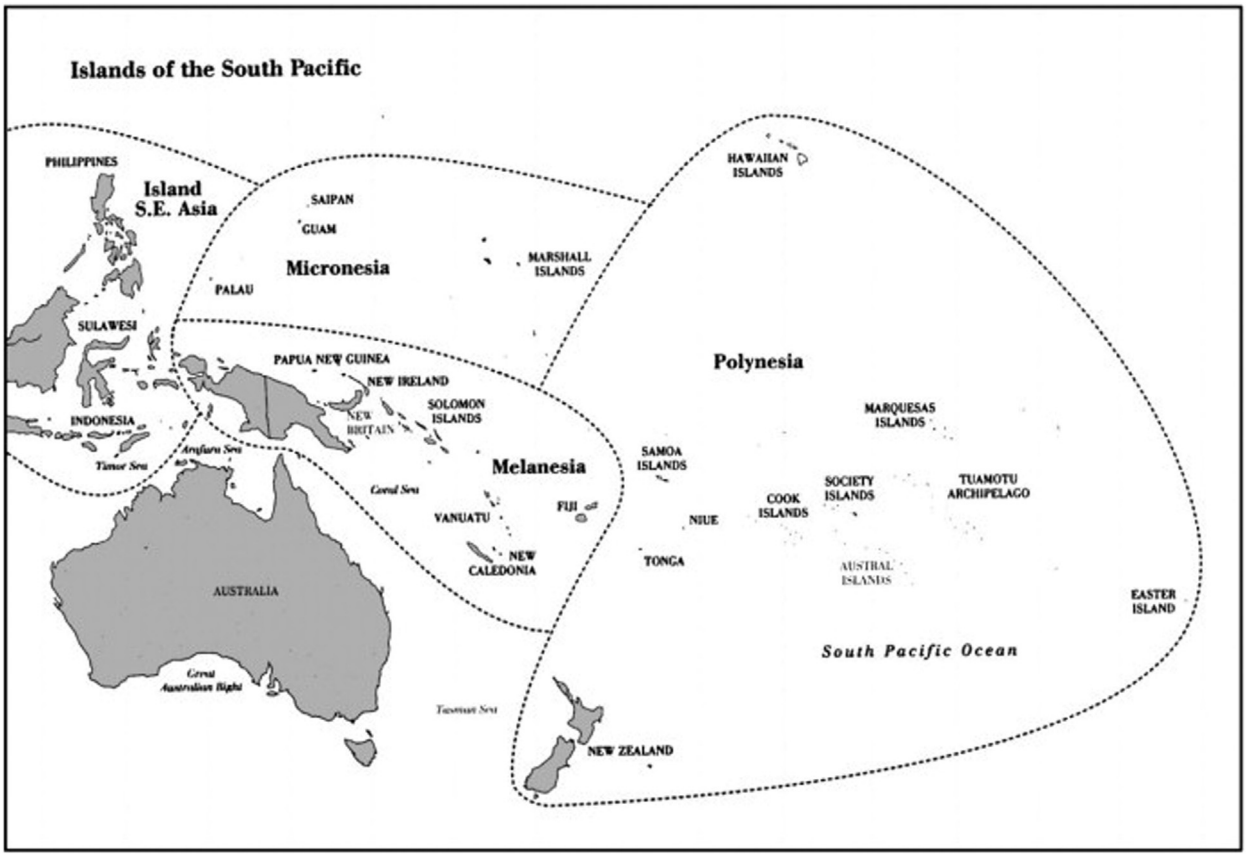

Figure 1 Islands of the South Pacific.

Social and Cultural Anthropology in/of Melanesia

Dumont D'Urvilles' classificatory legacy has been provocatively criticized on different grounds, namely that 'Melanesia' often designates one term of an invidious comparison with 'Polynesia.' Such comparisons, manifestly racist in many nineteenth-century versions, portray Melanesians as economically, politically, and culturally less developed than Polynesians and, by evolutionary extension, people of 'the West.' To Western political and economic interests, moreover, Melanesia has historically appeared marginal; Melanesianist anthropologists, for example, never found much support within the post World War II institutional apparatus of area studies. This marginality has rendered Melanesian studies as the exemplar of anthropology's authoritative discourse on 'primitives,' and contributed to the popular perception of Melanesian studies as the anthropological subfield most engaged in the contemporary study of radical (non-Western) Otherness. Such a perception not only turns Melanesian people into exotic vehicles for Western fantasy, but also suppresses the continuing history of colonialism in the region and misrepresents much Melanesianist anthropology, past and especially present.

Melanesian ethnography has played a central role in the history of the larger discipline's methods and theories. The work of W. H. R. Rivers on the 1898 Cambridge Expedition to the Torres Straits and later trips to Melanesia, and of Bronislaw Malinowski during his World War I era research in the Trobriand Islands, both mark important moments in the development of genealogical inquiry and extended participant observation as characteristic methods of social anthropology. Similarly, the Melanesian ethnography of M. Mead, G. Bateson, R. Fortune, and I. Hogbin, among others, guided anthropological approaches to questions of socialization and emotions, culture and personality, gender relations and sorcery, ritual process and social structure. From 1950 onwards, anthropology drew heavily on Melanesian ethnography – including the first waves of intensive fieldwork in the New Guinea Highlands – for diverse theoretical advances in the areas of cultural ecology, gender and sexuality, and in the study of metaphor, meaning, and knowledge transmission.

Gift/Commodity

Most prominent among the gate-keeping concepts that make Melanesia a distinctive anthropological place is exchange. Malinowski's classic description of *kula*, and Marcel Mauss's treatment of Malinowski's ethnography in *The Gift*, provided the grounds for understanding Melanesian sociality as fluid and contingent, performed through pervasive acts of give and take. In the 1960s, the critique of models of segmentary lineage structure in the New Guinea Highlands further enabled recognition of reciprocity and exchange as the Melanesian equivalents of the principle of descent in Africa. Exchange brings into

being the social categories that it perforce defines and mediates: affines and kin, men and women, clans and lineages, friends and trade partners, living descendants and ancestral spirits. The rhythms of Melanesian sociality thus develop out of infrequent large-scale prestations associated with collective identities and male prestige; occasional gifts of net bags from a sister to her brother or of magic from a maternal uncle to his nephew; and everyday exchanges of cooked food between households and betel nut between companions.

Before contact with Westerners, Melanesia was laced by extensive trade networks that connected coastal and inland areas, and linked numerous islands and archipelagoes. These regional networks crossed different ecological zones and organized the specialized production of food, pigs, crafts, and rare objects of wealth such as pearl shells and bird plumes. Wealth objects were in turn often displayed and circulated in sequential chained exchanges, of which the most complex and geographically far-reaching include *kula*, and the Melpa *moka* and Enga *tee* systems of the central New Guinea Highlands. These exchange systems often define a sphere of male-dominated activity that depends on the labor of women, but instances of women circulating women's wealth in public exchanges do occur throughout Melanesia. During the colonial period, many trade networks atrophied, while other forms of competitive exchange effloresced, incorporating new items of wealth. State-issued money now commonly circulates in exchange alongside indigenous currencies and wealth, especially as part of marriage payments and homicide compensations.

Melanesians sometimes self-consciously contrast their own reciprocal gift-giving with the market transactions of modern Western society; for example, the distinction between *kastam* ('custom') and *bisnis* ('business') occurs in the English-based pidgin-creoles of the region. In related fashion, anthropologists have used an ideal – typical contrast between a gift economy and a commodity economy to describe the social relations of reproduction characteristic of indigenous Melanesia. For Gregory (1982), 'gift reproduction' involves the creation of personal qualitative relationships between subjects transacting inalienable objects, including women in marriage; and 'commodity reproduction' involves the creation of objective quantitative relationships between the alienable objects that subjects transact. This dichotomy provides a way of talking about situations now common in Melanesia in which the status of an object as gift or commodity changes from one social context to another, and varies from one person's perspective to another. Such situations occur, for example, when kin-based gift relations become intertwined with the patronage system associated with elected political office or when marriage payments become inflated by cash contributions derived from wage labor and the sale of coffee and copra.

While Gregory's formulation has been criticized as obscuring the extent to which commodity-like transactions – barter, trade, localized markets – were always a feature of Melanesian social life (and gifts a feature of highly commoditized Western societies), it nonetheless highlights how Melanesians produce crucial social relations out of the movement of persons and things. Anthropologists have thus used the concept of personhood as a complement to the concept of the gift in understanding Melanesian moral relationships; for it is in the context of gift exchange that the agency of persons and the meaning of things appear most clearly as the effects of coordinated intentions. Marilyn Strathern contrasts a Melanesian relational person or dividual with a Western individual. Whereas Western individuals are discrete, distinct from and prior to the social relations that unite them, Melanesian dividuals are composite, the site of the social relations that define them. Similarly, whereas the behavior of individuals is interpreted as the external expression of internal states (e.g., anger, greed, etc.), the behavior of dividuals is interpreted in terms of other people's responses and reactions. The notion of dividuality makes sense of many Melanesian life-cycle rites that recognize the contributions of various kin to a person's bodily growth, as well as of mythological preoccupations with the management of ever-labile gender differences.

Equality/Hierarchy

Melanesian polities have often been characterized as egalitarian relative to the chiefly hierarchies of Polynesia; instances of ascribed leadership and hereditary office, numerous throughout insular Melanesia, have accordingly been regarded as somewhat anomalous, while the categorical distinction between 'chiefs' and 'people of the land' has served to mark Fiji's problematic inclusion within Melanesia. Traditional Melanesian political organization has been defined negatively, as not only stateless but also lacking any permanent form capable of defining large-scale territorial units. Precolonial warriors, such as those of south-coast New Guinea, distinguished themselves through vigorous practices of raiding, head-hunting, and cannibalism that ensured social and cosmological fertility. Local autonomy, armed conflict, and ongoing competition among men for leadership have thus been taken as the hallmarks of traditional Melanesian politics. This definition of precolonial Melanesian polities invokes a sharp contrast with the centralized, bureaucratic, violence-monopolizing states of the West. Yet, at the same time, Melanesian politics, epitomized in the figure of the self-made 'big man,' have sometimes been thought to recall values of individualism and entrepreneurship familiar from the West's own self-image.

Big men typically validate their personally achieved leadership through generous distributions of wealth. They manipulate networks of reciprocal exchange relationships with other men, thus successfully managing the paradox of generating hierarchy – even if only temporarily – out of equality. This paradox is variously expressed within Melanesia. In classic big-man exchange systems, known best from areas of high population density and intensive pig husbandry and horticulture in the New Guinea Highlands (Melpa, Enga, Simbu, Dani), a constant and imbalanced flow of shells, pigs, and money underwrites individual prestige. Such systems are not found everywhere in Melanesia, however, prompting Godelier (1986) to invent a new personification of power, the 'great man.' Great men emerge where public life turns on ritual initiations, where marriage involves the direct exchange of women ('sister exchange'), and where warfare once similarly prescribed the balanced exchange of homicides. The prestige and status of great men often depends upon their success in keeping especially valued 'inalienable possessions' out of general

circulation. While this typological innovation has been challenged, political life in the Sepik area and in eastern Melanesia frequently involves transmitting ritual resources (heirlooms, myths, names, and so forth) within a delimited kin grouping such as a clan. In the Mountain Ok area, ancestral relics (bones and skulls) provide the material focus of a secret men's cult that links contiguous communities and complements clan affairs; in insular Melanesia, forms of voluntary male political association detached from kin groupings are found, such as the public graded societies of northern Vanuatu. Nevertheless, the scope for establishing political relations in precolonial Melanesia was restricted by the instability of big men's exchange relations, or by the demands placed on high-ranking men to sponsor youthful aspirants, or by the limited extent of the communities or kin groups beyond which ritual resources have no currency.

Whereas Melanesian political relations have been deemed egalitarian by comparison with the West, gender relations have usually been characterized as manifestly inegalitarian. Ethnography of highlands New Guinea in particular suggests that women's productive labor in gardening and pig husbandry is eclipsed if not wholly appropriated by men who act exclusively as transactors in public exchanges. In areas with once secret and lengthy rites of male initiation, senior men socialized boys into ideologies of dangerous female power and pollution while enforcing strict physical separation of the sexes. Ethnography of insular Melanesia (AN speakers), by contrast, suggests that gender inequality is less severe, tempered by women's social and genealogical roles as sisters rather than wives, and as both producers and transactors of cosmologically valued wealth items such as finely woven mats. But Margaret Mead's pioneering work on sex and temperament in the Sepik area, and Knauft's (1993) review of south-coast New Guinea ethnography suggest the extreme variation possible in gender relations among even geographically proximate societies.

New forms of social inequality and hierarchy have taken shape since the mid-nineteenth century through intensified interactions with foreign traders, missionaries, and government officials. During the colonial period, 'chiefs' were often appointed in areas where they had not existed, while indigenously recognized leaders were often able to use their status to monopolize access to the expanding cash economy. Male labor migration put extra burdens on women who remained at home, their contacts with new ideas and values inhibited. In rural areas today, 'rich peasants' have emerged by manipulating customary land tenure – and sometimes using 'development' funds of the postcolonial state – to accumulate smallholdings of cash crops; conflicts have arisen over claims to cash crops that women raise on land owned by their husbands. In urban areas, political and economic elites attempt to reproduce their dominant class position through private education and overseas investments as well as cosmopolitan consumption practices.

Melanesian Modernities

The legacies of different colonial states (Great Britain, France, Germany, The Netherlands, Australia, and Indonesia) variously condition Melanesian political economies, especially with regard to control of land. In Papua New Guinea, limited land alienation by the crown, an indirect subsidy of contract labor, has preserved a vital resource in the defense against encroaching commodification; but French settlers in New Caledonia and, more recently, Javanese transmigrants in Irian Jaya violently displaced indigenous populations. In Fiji, complicated arrangements for leasing agricultural land, a consequence of British colonial policy, generate tension between Fijian landowners and their largely Indo-Fijian tenants, descendants of indentured plantation workers brought from India in the late nineteenth and early twentieth centuries. In addition, uneven capitalist markets and proliferating Christian denominations have added new dimensions to the diversity of Melanesia. Anthropologists readily acknowledge that even in those parts of Melanesia where the presence of foreign agents and institutions has been relatively weak (once a tacit justification for rendering colonial entanglements ethnographically invisible), received cultural orientations have become articulated with globalized elements of Western modernity. In other words, Melanesia is emerging as a key anthropological site for the comparative study of plural or local modernities.

Current Melanesianist research, historical as well as ethnographic, extends earlier studies of north-coast New Guinea cargo cults that sought to demonstrate how indigenous cosmologies shaped local interpretations of European material wealth and religious culture. This research highlights the syncretic products of intercultural encounters, noting how Melanesians have appropriated and reinterpreted such notable ideological imports as Christianity, millenarian and otherwise, and democracy. Even the impact of large-scale mining and timber projects has been shown to be mediated by highly local understandings of land tenure, kinship, and moral economy. In West Papua and Papua New Guinea, these projects define sites of struggle among transnational corporations, national states, and local stakeholders, often generating destructive environmental consequences (the Ok Tedi and Grasberg gold mines) and massive social dislocation (the civil war on Bougainville).

By the same token, current research emphasizes how the historical encounter with Western agents and institutions often transforms indigenous social practices, including language socialization, thus revaluing the meaning of tradition. For example, the flourishing gift economies of rural villagers are often a corollary of the remittances sent back by wage workers in town, while the ritual initiation of young boys might provide a spectacle for paying tourists from Europe and the USA. Since the 1970s, the emergence of independent nation-states has stimulated attempts to identify national traditions with habits such as betelnut chewing or artifacts such as net bags, both of which have become much more geographically widespread since colonization. In many of these states, 'chieftainship' furnishes the newly traditional terms in which elected political leaders compete for power and resources. In New Caledonia, the assertion of a unifying aboriginal (Kanak) identity and culture has been a feature of collective anticolonial protest.

Contemporary Melanesian societies and cultures are thus defined by a tense double process. On the one hand, they are increasingly drawn into a world system through the spread of commercial mass media; the force of economic reforms mandated by interstate regulatory agencies; and the sometimes tragic exigencies of creating political communities on the global model of the nation-state. On the other hand, they

continue to reproduce themselves in distinctive and local ways, organizing processes of historical change in terms of home-grown categories and interests.

See also: Big Man, Anthropology of; Colonialism, Anthropology of; Exchange in Anthropology; Kula Ring, Anthropology of; Malinowski, Bronislaw (1884–1942); Mauss, Marcel (1872–1950).

Bibliography

Douglas, B., 1998. Across the Great Divide: Journeys in History and Anthropology. Harwood Academic Publishers, Amsterdam.

Foster, R.J. (Ed.), 1995. Nation Making: Emergent Identities in Postcolonial Melanesia. University of Michigan Press, Ann Arbor, MI.

Gewertz, D.B., Errington, F.K., 1999. Emerging Class in Papua New Guinea: The Telling of Difference. Cambridge University Press, Cambridge, UK.

Godelier, M., 1986. The Making of Great Men: Male Domination and Power Among the New Guinea Baruya (R. Swyer, Trans.). Cambridge University Press, Cambridge, UK.

Gregory, C.A., 1982. Gifts and Commodities. Academic Press, New York.

Jolly, M., 1994. Women of the Place: Kastom, Colonialism and Gender in Vanuatu. Harwood Academic Publishers, Philadelphia.

Kirch, P.V., 1997. The Lapita Peoples: Ancestors of the Oceanic World. Blackwell, Cambridge, MA.

Knauft, B.M., 1993. South Coast New Guinea Cultures: History, Comparison, Dialectic. Cambridge University Press, Cambridge, UK.

Knauft, B.M., 1999. From Primitive to Postcolonial in Melanesia and Anthropology. University of Michigan Press, Ann Arbor, MI.

Kulick, D., 1992. Language Shift and Cultural Reproduction: Socialization, Self and Syncretism in a Papua New Guinea Village. Cambridge University Press, Cambridge, UK.

Lederman, R., 1998. Globalization and the future of culture areas: Melanesianist anthropology in transition. Annual Review of Anthropology 27, 427–449.

Pawley, A., 1981. Melanesian diversity and Polynesian homogeneity: A unified explanation for language. In: Hollyman, J., Pawley, A. (Eds.), Studies in Pacific Languages & Cultures in Honour of Bruce Biggs. Linguistic Society of New Zealand, Auckland, NZ, pp. 269–309.

Spriggs, M., 1997. The Island Melanesians. Blackwell, Oxford, UK.

Strathern, M., 1988. The Gender of the Gift: Problems with Women and Problems with Society in Melanesia. University of California Press, Berkeley, CA.

Weiner, A.B., 1992. Inalienable Possessions: The Paradox of Keeping-While-Giving. University of California Press, Berkeley, CA.

Oceania, Sociocultural Overviews: Polynesia and Micronesia

Niko Besnier, Universiteit van Amsterdam, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The terms 'Polynesia' and 'Micronesia,' nineteenth-century European inventions, have survived as convenient labels to refer to two large regions of the insular Pacific. These regions were peopled in the last 4000 years by waves of seafaring immigrants from Southeast Asia, who established patrilineal societies in Polynesia and matrilineal societies in Micronesia, which exhibit various degrees of stratification. The intensification of contacts with Westerners during the Enlightenment heralded an era of colonialism that brought about radical transformations. With few exploitable resources, contemporary societies of Polynesia and Micronesia tackle problems that range from deteriorating health to the possible effects of global warming.

History of Geographical Nomenclature

The region that has commonly been referred to as 'Oceania' since the early nineteenth century was peopled by successive waves of prehistoric settlers. The last of these were a seafaring people who settled in the islands of the Pacific, mingling with prior occupants in the case of New Guinea and neighboring areas and occupying uninhabited islands further west, all the way to Hawai'i in the Northern Pacific, New Zealand in the far south, and Easter Island in the southeast corner of the region. The social and cultural similarities among the islands east of and including Fiji reflect not only their relatively recent peopling by the same people but also the fact that islanders maintained active contact with one another throughout prehistory, even across vast distances, thanks to their sophisticated navigation technology. For them, the Pacific represented not a vast ocean dotted with minuscule islands, but rather a 'sea of islands,' in the memorable phrase of Tongan scholar Hau'ofa (1994), namely a vast network of human settlements connected to one another through kinship, exchange, alliance, history, and cosmology.

In 1756, French scholar and politician Charles de Brosses coined the term 'Polynesia' to refer to all islands of the Pacific east of Australia, hoping to persuade the French court to finance their exploration and eventual colonization, convinced as he was that they would prove to be 'rich in spices.' The label remained standard in European writings for several decades, during which Europeans began to take a keen interest in the Pacific Islands. Explorers, as well as their contemporaries who read and commented on the journals of their voyages that they published to great acclaim, were informed by the humanistic universalism of the Enlightenment and early Romantic eras that viewed humanity as one. They saw 'primitive man' as a model for humanity, living in harmony with nature and one another, and unencumbered by the trappings of 'civilization.' In particular, Tahiti and the Society Islands figured prominently in the elaboration of this romantic primitivism after first contact with British and French navigators in 1767 and 1768, respectively.

The early nineteenth century witnessed a radical shift in the European philosophical climate, particularly in France, with the emergence of a preoccupation with racial difference, which was now viewed as evidence of the polygenetic origin of humans and of the evolutionary inequality among them. It is in this racialist context that two French navigators with scholarly aspirations, Jules Dumont d'Urville and the lesser-known Grégoire-Louis Domeny De Rienzi, simultaneously devised a typology of Pacific Islanders based on their physical appearance, particularly skin color, and a division of the Pacific Islands into three large regions (in addition to a fourth one, 'Malaysia,' applied to insular southeast Asia): Melanesia (i.e., dark islands), Micronesia (i.e., small islands), and Polynesia (i.e., many islands), the referent of the latter being now much reduced in size. The different preoccupations of European observers of Pacific Islanders at different times are inscribed in the historical evolution of maps (Tcherkézoff, 2008) and iconographic representations (Smith, 1985/1969). At the time of contact, islanders themselves referred to their homes with names for individual islands if they were small enough or, in the case of larger islands, with the names of specific villages, regions, or language groups. Many of these toponyms were varied, complex, and contested.

Polynesia and Micronesia as Cultural Areas

Despite the morally problematic historical foundation of this nomenclature, 'Polynesia' and 'Micronesia' have remained to this day convenient labels for geographical areas whose social, cultural, archeological, historical, political, and linguistic significance is somewhat arbitrary. Polynesia and Micronesia are complementary to Melanesia, generally considered to comprise New Guinea and adjacent islands, the Solomon Islands, New Caledonia, and Vanuatu.

Micronesia encompasses Palau and outlying islands, Guam, the Mariana Islands, Caroline Islands, Marshall Islands, and Gilbert Islands, as well as two isolated islands, Nauru and Banaba (Ocean Island). Contemporary political considerations may include island groups of the Central Pacific like the Phoenix and Line Islands; although geographically located in Polynesia, these islands, which were not permanently inhabited until the modern age, are governed by Kiribati, the modern state based in the Gilbert Islands.

Polynesia includes all islands and groups falling within a large triangular area whose apexes are New Zealand to the south, Hawai'i to the north, and Rapanui (Easter Island) to the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12121-X

west. The largest islands and island groups of the region are Tonga, Samoa, Rarotonga and the Southern Cooks, Tahiti and the Society Islands, the Marquesas, and Hawai'i, in addition to the much larger New Zealand. Fiji and Rotuma are sometimes considered part of Melanesia and other times part of Polynesia. Politically or culturally notable smaller islands include Niue, Wallis and Futuna, Tuvalu, Tokelau, the Northern Cooks, Rapanui, and Pitcairn Island, as well as the Tuamotus, Austral, and Gambier Islands, which are politically subsumed under French Polynesia. In addition, a geographically heterogeneous group of about 18 islands and sections of islands are commonly identified as 'Polynesian Outliers' societies because they are located outside of this triangular area: in Melanesia to the west (e.g., Takuu off Bougainville Island, Sikaiana and Rennell in the Solomon Islands, Mae and Mele in Vanuatu, North and South Ouvea Atoll in the Loyalty Islands) and, in the case of two islands (Nukuoro and Kapingamarangi), in Micronesia.

Islands of many different types are found in the two regions: a few continental islands (e.g., Guam); volcanic structures, some of which can be substantial in size (e.g., Hawai'i); raised coral islands of different types (e.g., Tonga); atolls (e.g., Tuvalu); and combinations of these. These various morphological structures, combined with a vast array of climatic conditions (from subarctic to equatorial climates, from wet to dry), provided prehistoric settlers with very different problems of ecological adaptation.

The human groupings that inhabited each of the two regions at the dawn of contact with Europeans shared a number of social and cultural characteristics with one another, and continued to do so well into the modern age. Yet many societies of insular Melanesia display characteristics that are identical to those prevalent among Polynesian or Micronesian societies, reflecting the fact that all three regions were settled by the same people in prehistoric times, even though the newcomers found already inhabited land in the Western part of Melanesia. In addition, commonalities among the societies of both areas are balanced by important patterns of variation, to the extent that no feature of social organization of culture will be found in all societies in question.

Perhaps the least controversial way of identifying Polynesia is linguistically: all languages spoken natively in the islands are more closely related to one another than to any other language. Indeed, the reason for identifying Outlier societies as Polynesian is the fact that all their members speak languages that are most closely related to the languages spoken in Polynesia. In terms of social organization and culture, however, Outlier communities vary widely, from bearing considerable similarity to the rest of Polynesia (e.g., Tikopia) to having much more in common with their more immediate non-Polynesian neighbors (e.g., the Polynesian-speaking villages of Ouvea, Loyalty Islands). Yet even this criterion is not devoid of caveats: on the boundary between Fiji and Tonga, historical linguistic evidence suggests a continuum of gradual linguistic differentiation rather than a clean break. In contrast, linguistic factors are of little use as determiners of what Micronesia includes and what it does not: all languages spoken in the area are Austronesian, but some (e.g., Chamorro of the Marianas and Palauan) are closer to languages spoken outside of Micronesia (e.g., in the Philippines) than to the other languages of Micronesia.

Contemporary Polynesian and Micronesian societies can only be understood with reference to the fact that many became diasporic in the course of the second half of the twentieth century, a trend originally motivated by labor migration to the industrial centers of the Pacific Rim. Today, one finds in New Zealand significant communities of Samoans, Tongans, Niueans, and Cook Islanders; Samoans and Tongans are also settled in comparatively large numbers in Australia; both are also represented in Hawai'i and the Continental United States, where Caroline and Marshallese Islanders are also numerous, at least in comparison with the size of the populations in the island groups of origin. Emigrant communities in many cases are much more populous than the island-based communities. While the Cook Islands and Niue, two small Polynesian states partially under New Zealand's political jurisdiction, have 19 000 and 1400 inhabitants, respectively, Cook Islanders living in New Zealand number 58 000 and Niueans 22 000 (2006 figures). Micronesian and Polynesian people also migrate internally to urbanized or industrialized centers of the Pacific Islands. Thus Guam, New Caledonia, Fiji, and Hawai'i are home to many Pacific Island migrant communities originating from surrounding island groups.

Initial Human Settlement

Around 3500 BP, a wave of people traveled from East and Southeast Asia, settling in the islands of Melanesia, some of which were already inhabited while others were not, and occupying the entire insular region in a relatively short time (3500–1000 BP). While the process inevitably involved backtrackings, sidesteps, and false starts, its general West-to-East direction is reflected in the relative heterogeneity of cultures, social structures, and languages in the Western Pacific compared to the relative homogeneity of societies and languages further east.

The most salient archeological evidence for this conjectured prehistoric migration is a lowly style of decorated pottery referred to as 'Lapita,' fragments of which are found in insular Melanesia and Western Polynesia, carbon-dated to 3500–2000 BP. Lapita pottery makers and users appear to have been accomplished long-distance sailors who utilized large outrigger sailing canoes, skilled fisherfolk, and agriculturalists of staples that they brought with them, such as taro, believed to be among the first food crops ever to be domesticated (around 10 000 BP in tropical Asia). Among other sources of food that they may have transported on their long voyages figure pigs, dogs, and fowl. They organized their social life in hierarchical fashion. They spoke languages that belonged to the Austronesian language family, the largest in the world in terms of its geographical spread and number of different languages, with daughter languages spoken today from Madagascar to Easter Island, including Indonesia, Malaysia, and the Philippines. With few exceptions (such as the languages of Western Micronesia mentioned earlier), languages of these regions all belong to the Oceanic group of Austronesian.

Around 3000 BP, the settlers reached Fiji, Tonga, and Samoa, an area in which, according to some researchers, migrations may have 'paused' for a while. Such a pause would have allowed a social, cultural, and linguistic distinctiveness

to emerge that would eventually become what is now recognized as 'Polynesian.' From the Fiji–Western Polynesia area, early Polynesians settled the rest of Polynesia in the course of the following two millennia, finally reaching Hawai'i in about 650 AD and New Zealand around 1000 AD (Kirch, 2000). There is no convincing evidence of any subsequent prehistoric human settlement in the insular Pacific other than the Lapita potters and their descendants (particularly from south America), although they intermarried in Melanesia with their non-Lapita-making neighbors, forming as diverse a genetic pool as is found in any other part of the world.

Micronesia's prehistory is much more complex than that of Polynesia, and less well understood. Evidence of human settlement in the Mariana Islands dates back to approximately 4000 BP. The archeological record indicates a Southeast Asian connection for this early population, as does archeological evidence gathered on Palau and Yap. The rest of the Caroline Islands, the Marshall Islands, and the Gilbert Islands were settled by a northwestward back-migration from eastern Melanesia. Languages spoken in this region, commonly referred to as 'Nuclear Micronesian,' exhibit greater linguistic homogeneity than other Micronesian languages, suggesting a more focused history of prehistoric settlement.

Society and Culture

Many aspects of precontact society and culture have reproduced over time and continue to characterize contemporary circumstances. In precontact times, the societies of Polynesia and Micronesia displayed certain degree of sociocultural commonality, as witnessed, for example, in a tendency to have stratified political systems and in the elaboration of certain symbolic complexes such as *tapu* (roughly, 'religion-based prohibition') and *mana* (roughly, 'efficacy of divine origin') in Polynesia (Howard and Borofsky, 1989). However, the recognition of patterns of commonality must always be qualified in at least two ways. First, one must be attentive to the dynamics through which certain characteristics have come to be identified as 'typically' Polynesian or Micronesian in spite of the fact that these characteristics are absent in some societies of the region. Second, 'precontact times' is a vague characterization at best, and all island societies experienced radical changes over the centuries, some of which can be identified archeologically, in the course of which 'typical' Polynesian or Micronesian characteristics appeared in and disappeared from particular societies.

A significant number of societies in both regions were socially and politically stratified, the more so in Polynesia than in Micronesia. Particularly on large islands and island groups, chiefdoms were organized around a leadership that derived power through a combination of genealogical ties to both sacred and profane entities that confirmed claims to power (ascription) and personal performance that hopefully confirmed these claims (achievement). In some of more stratified societies, persons were also ranked into categories ranging from the highest ranks, which were commonly sacralized, to the lowest ranks, whose members were often not considered human. Other societies, in contrast, displayed little stratification and emphasized egalitarianism and consensual decision making. Political organization on some islands was centered on localized kinship units, particularly in Micronesia but also on the smaller islands of Polynesia. The rough correlation between the amount of resources to be produced and organized and the degree of social stratification, at least in Polynesia, led some early researchers to propose that one engenders the other: chiefs, as resource managers, increase in importance as resources increase in volume (Sahlins, 1958). Few anthropologists today would give such determinative power to resource management (and indeed the empirical evidence does not support the model), although it certainly plays a role in the development and elaboration of social stratification.

Prevalent patterns throughout the area centralized the identification of a common founding family ancestor, creating in some societies a pyramid-shaped kinship structure. In Polynesia, the most important branch was the senior patrilineal branch, and members of other branches reckoned their relative status according to their relative distance from the main descent branch, measured through relative patrilineality and seniority, although matrilineal descent could also play an important role. Micronesian societies are predominantly matrilineal. There, matrilineal clans have historic claims to particular pieces of land and are the basis for social organization, with rank being based on claims of founding settlement of particular pieces of land. Probably all Polynesian and Micronesian societies also offered the possibility of bilateral affiliation, whereby kinship could be based on either patrilineal or matrilineal principles, thus providing room for social and political maneuvering. Adoption in its various guises was and continues to be prevalent throughout Polynesia and Micronesia.

Throughout Polynesia, gender figures prominently in the reckoning of rank, although the specifics of this interaction differ across societies. In Samoa and Tonga, for example, the relationship between sisters and brothers is foundational to the cultural and social organization of society, with sisters being superior in rank to brothers, particularly if senior to them. However, the prevailing ideology throughout the region dictates that the husband is the head of the family, including his wife, although the extent to which this ideology is the result of nineteenth-century missionaries' reconfiguration of the Polynesian family is not known. At first contact with Europeans, Tahitian society included a category of transgender males that took some aspects of women's identity, which is now widespread throughout Polynesia, but less obviously so in Micronesia (Besnier and Alexeyeff, 2013).

Early Contacts with Europeans and Colonialism

While Pacific Islanders maintained active contacts between islands and island groups prior to establishing contact with Westerners in the region, Europeans' arrival and gradual colonial domination over the islands brought about fundamental changes. Spanish, Portuguese, and Dutch travelers had sailed through Polynesia and Micronesia since the very early days of the Age of Discovery, but their contacts with Islanders were sporadic, ephemeral, and often violent. The exception to the first two characteristics (but not the third) are Guam and the Marianas, which Spaniards saw as strategically situated on

the sea road between the colonial centers of the empire in the Philippines and the New World. Spain claimed these islands in the mid-sixteenth century (later extending its claim over the rest of Micronesia), a move which was followed by the virtual decimation of the Chamorro inhabitants of Guam and the Marianas through disease, slaughter, and forced resettlement.

Contacts between Westerners and Islanders in both Polynesia and Micronesia did not begin in earnest until the end of the eighteenth century, when the relatively peaceful period following the end of the Seven Year War in Europe (1763) enabled France and Britain to fund expeditions to the Pacific Islands and other parts of the world. These voyages were driven by complicated, overlapping, and sometimes contradictory motivations, including the desire to find lands hitherto unknown to Europeans and exploit the resources they might offer (particularly the conjectured *Terra Australis Incognita* or 'unknown land of the South'), the intellectual curiosity that reigned during the European Enlightenment, and scientific endeavors, the most celebrated being the mapping of the transit of Venus from Tahiti, which secured funding for Captain James Cook's first voyage around the world (1768–71).

Enlightenment-era contacts were considerably more peaceful than those of earlier centuries. Still, despite the ideals of the Enlightenment that suffused them on the European side, different perceptions of what these contacts were about, particularly different ideas about property, did occasionally lead to serious conflicts. Perhaps the most celebrated from the Westerners' perspective is Captain Cook's violent death on the Big Island of Hawai'i, the analysis of which became the topic of a spirited intellectual exchange in the 1990s between two prominent anthropologists about whether or not Hawaiians viewed Cook as a god (Sahlins, 1987; Obeyesekere, 1992). But even when islanders did not suffer casualties at the moment of contact, in many parts of Polynesian and Micronesia they subsequently suffered from the new diseases, economic upheavals, wars, land dispossessions, outright massacres, raids by slavers, and other calamities that were associated, directly or indirectly, with the increasing Western dominance over the region. Some scholars in the nineteenth century expressed concern over the possible complete disappearance of Pacific Islanders, a sentiment echoed in the twentieth century in the 'fatal impact' approach to early contacts advanced by some historians (Moorhead, 1987).

Early European travelers to the Pacific Islands were soon followed by a broad variety of Westerners, including adventurers, traders, whalers, and employees of colonial corporations seeking to establish plantations and trading empires. Those whose impact was most consequential were missionaries of various Christian denominations, particularly adherents of the nonconformist evangelical movements that were then sweeping Britain in the late eighteenth century. They later competed with Catholic missionaries (who had been firmly implanted in Guam since the sixteenth century) and Mormons, who first arrived in the Tuamotu Islands as early as 1844. Members of the London Missionary Society, founded in 1795, arrived in Tahiti in 1797 and, after some vicissitudes, managed to convert a chief whose power had been greatly increased by his direct contact with earlier European visitors. In the socially stratified societies of Polynesia, the strategy of first converting the powerful with the rest of the population following suit led to swift and effective conversions. Today, all islanders adhere to some form of Christianity, with different denominations often competing for adherents.

A notable consequence of Westerners' increasing visibility in nineteenth century Polynesia was the consolidation of hitherto politically fragmented chiefdoms. In Tonga, Fiji, Tahiti, and Hawai'i (and to a lesser extent Samoa and New Zealand), ambitious and astute chieftains utilized the new tools that Westerners introduced to the islands, trading natural resources for firearms and protection for religious conversion, to defeat their competitors and impose their rule over entire islands and island groups, establishing themselves as sovereigns. The emergence of these societies in the modern age was accompanied by radical reconfigurations of power structures, with the introduction of new forms of inequality but also emergent notions of citizenship protected by constitutions. Of the kingdoms established in the nineteenth century, only Tonga remains to this day, others having fallen to the imposition of colonial rule. The most egregious case was the illegal overthrow by the United States of the Hawaiian monarchy in 1893 with the pretext of protecting the interests of American plantation owners.

In the course of the nineteenth century, European powers and later the United States sought to establish colonies and spheres of influence, competing over the richer islands (e.g., Tahiti, New Zealand, Samoa), while annexing island groups with few or no exploitable resources in the context of international treaties designed to regulate their spheres of influence. In Polynesia, the key players in the nineteenth century were predominantly British, French, and German. Beginning in 1842, France established a protectorate, and later colony, over Tahiti and numerous adjacent islands. Māori chiefs in New Zealand were made to sign the infamous Treaty of Waitangi with Britain in 1840, the consequences of which are still argued over to this day. The Cook Islands became a British protectorate in 1888, the administration of which was transferred to New Zealand in 1901. In 1900, the British consul essentially forced King George II of Tonga to sign a treaty that made the kingdom a protected state of Britain. The Tripartite Convention of 1899, signed by Germany, the United States, and Britain, placed the larger islands of Samoa under German control, while the US Navy took over the administration of the smaller eastern islands of the group, which have been known as American Samoa to this day.

In Micronesia, Spain hung on to the last significant remnants of its colonial empire until its defeat in the Spanish–American War of 1898, when Guam came under US control (along with, outside the region, the Philippines and Puerto Rico, while Cuba became independent under the control of the US). A German–Spanish Treaty in 1899 added to the German empire the Caroline, Mariana, and Marshall Islands, as well as Palau and Nauru.

During these transformative decades, several pioneering scholars documented the societies and cultures of the two regions. Polish-born John Stanislaw Kubary (1846–96), for example, who eventually became an employee of the Hamburg-based plantation firm Godeffroy und Sohn (as well as an Australian citizen), wrote detailed accounts of the Micronesian societies among which he lived. German-educated Russian aristocrat Nikolai Mikloukho-Maclay (1846–88), who

also settled in Australia, documented Micronesian as well as New Guinea societies and championed Pacific Island people against colonial exploitation. Their important descriptive works have been largely forgotten because of their authors' short lives and the fact that they wrote in Polish and Russian, respectively. Slightly more enduring are the detailed works about Palau, Samoa, and Hawai'i of German ethnologist Augustin Krämer (1865–1941).

The end of World War I somewhat reshuffled the distribution of colonial powers, with Western Samoa passing into the hands of New Zealand after Germany's defeat. Japan, an ally of Britain and France at the time, emerged as a colonial power, being granted by the League of Nations in 1920 a South Pacific Mandate (*Nan'yō-chō*) over previously German-controlled Micronesia. During the two-and-a-half decades of Japanese rule, Micronesia experienced an influx of labor migrants, who came to work on plantations, in mines, and in fishing ventures, and eventually numbered more than twice the number of native Micronesians (approximately 100 000 immigrants for 40 000 Micronesians). The fact that most came from the former southern kingdom of Okinawa, which Japan had annexed 4 years after the 1868 Meiji Restoration, meant that they shared a colonial master with Micronesians, with whom many established cordial relations. Japanese rule and the Okinawan migration that it enabled left a strong cultural imprint on Micronesia.

World War II had a dramatic effect on all of Micronesia, which, with Melanesia, was the stage of some of the fiercest battles of the Pacific Theater. Micronesian lives were profoundly affected by the conflict and its aftermaths (Poyer et al., 2001). In contrast, Polynesia (other than Pearl Harbor in Hawai'i) remained largely in the background of the conflict. At the end of World War II, the United Nations created a Trust Territory of the Pacific Islands over all of Micronesia except the Gilberts, Nauru, and Guam, which the United States administered from 1947. The American administration took great pains to obliterate prior Japanese influence. Piggy-backing on the military administration, American anthropologists made Micronesian societies and cultures into the object of sustained research (Kiste and Marshall, 1999), studiously ignoring the substantial ethnographic corpus produced by Japanese anthropologists in prior decades, which to this date remains largely unknown outside of Japan.

Decolonization

The worldwide decolonizing trends of the second half of the twentieth century affected the entire Pacific Island region, as island nations began obtaining their independence from world powers. In Polynesia, Western Samoa (now Samoa) was granted independence in 1962, and Tonga ceased to be a British protected state in 1970. Decolonization in island groups held by the United States and France (particularly Micronesia and French Polynesia, respectively) was and still is a slower, more complex, and conflict-ridden process. A notable factor in both cases was the nuclear tests that both powers conducted in the islands until well into the 1990s, which created economic booms but also fundamental transformations in the lives of people, including ecological and health problems for the affected populations (Johnston and Barker, 2008).

Today, France continues to maintain a colonial presence in French Polynesia and Wallis and Futuna (both Polynesian groups), which is contested in the former but receives general local approbation in the latter. The United States signed a Compacts of Free Association with the Federated States of Micronesia and the Marshall Islands in 1986 and with Palau in 1994, and since then these entities have been recognized as sovereign nations despite continuing US security interests in the islands. American Samoa and Guam are unincorporated US territories in which the federal government maintains control of many functions, in spite of some local opposition in the case of Guam. The Northern Mariana Islands became a Commonwealth of the United States in 1986, whose citizens hold US citizenship but whose government controls internal affairs, although the extent of this control has been the subject of complex negotiations.

Rapanui (Easter Island) has been administered by Chile since 1888, despite political tension, and tiny Pitcairn Island (Polynesian from a geographical point of view) by Britain. New Zealand holds special political ties with three self-governing territories, namely Niue, Tokelau, and the Cook Islands. Hawai'i, which became the 50th state of the United States in 1959, and New Zealand are both postcolonial entities in which the politics of indigeneity are tense: indigenous Hawaiians and New Zealand Māori, numerical minorities in their own land, have claimed, with varying degrees of success, ownership of resources (primarily land) that were taken from their ancestors when Hawai'i and New Zealand became settler colonies. Fiji, which gained independence from Britain in 1970, has been rocked since 1987 by a series of coups triggered by ethnic conflicts between ethnic Fijians, who maintain ownership of key resources like land, and the descendants of immigrants from South Asia that the British colonial administration in 1879–1920 brought as indentured workers and later as free immigrants. By the late 1960s, Fiji Indians had become numerically dominant but, since the coups and the ethnic violence that accompanied them, many have emigrated to New Zealand, Australia, Canada, and the United States. These varied historical connections, as well as the different configurations in which domination, resistance, and everything in between are embedded, have contributed to the sociopolitical diversity of both regions. Standard reference works on the history of the Pacific Islands, including Polynesia and Micronesia, are Denoon (1997) and Lal and Fortune (2000).

Polynesia and Micronesia in the Global Age

The states and territories of Polynesia and Micronesia include some of the tiniest political entities of the world, such as Tuvalu (12 000 inhabitants, 26 km^2), with few viable economic resources. At the dawn of the twenty-first century, they have been particularly affected by economic problems. The dearth of natural resources in the two regions stands in sharp contrast with some island groups of Melanesia, where minerals and timber have generated substantial wealth for at least some segments of society. In post–World War II years, Nauru stood out as the exception, with one of the highest per capita incomes

in the world thanks to decades of phosphate mining, although extensive financial mismanagement since the end of mining in 2006 have radically altered this situation.

At the same time, capitalism and its institutions have become increasingly determinative, sidelining although not entirely obliterating traditional economies based on reciprocity. Similarly, global forces, in the form of increased population mobility and neoliberal policies enacted by such global entities as the World Trade Organization, are radically altering the social, cultural, and economic landscape in Polynesian and Micronesian societies. The region's economies (be it that of states or that of families) are experiencing increasing difficulties maintaining themselves simply with labor emigration, the remittances that emigrants generates, and a reliance on the continuation of foreign aid from wealthy donor nations, all of which are deeply vulnerable to the vagaries of global economies.

Some states of the region have experimented with what can be described as 'postmodern.' Some states of the region have experimented with what can be described as economic ventures. Polynesian and Micronesian microstates have been particularly proactive in these endeavors. Tuvalu, for example, has invested considerable energy in developing 'resources of jurisdiction,' that is, resources that only exist because of its status as a state. It began with the issuing of increasing quantities of postage stamps aimed at collectors. When philately waned (and the British company contracted to produce the postage stamps was found to be defrauding the government), Tuvalu persuaded Western donor nations to contribute to a trust fund instead of providing direct economic aid. Besides collecting license fees from foreign ships fishing within its 200-mile exclusive economic zone (totaling 600 000 square miles), the country also leased out its telephone country code to international corporate interests, although the Church of Tuvalu pressured the government to end the lease when it found out that the code was used mainly for phone-sex services. Most spectacular is the lease to a Canadian dot-com corporation of the Internet suffix, which fortuitously happens to be '.tv' (thus homophonous with 'television') and is now ubiquitous on the Internet.

Other ventures of the kind, which frequently come with serious social and political drawbacks, have included the sale of nonresidence passports to wealthy stateless foreigners (Tonga), offshore banking (Cook Islands), the sale of flags of convenience to shipping companies (Marshall Islands), the management of detention centers for asylum seekers to Australia (Nauru), and the establishment of offshore factories employing migrant workers (Mariana Islands). More traditional forms of economic development have included the attempt to transform the emigration of low-level unskilled labor into the export of professional labor, such as trained seamen, soldiers, and, most spectacularly, professional athletes. Rugby players from Fiji, Tonga, and Samoa have acquired visibility on the global circuit of athlete mobility, and Samoans who can legally migrate to the United States are overrepresented in American football. Significant numbers of (primarily indigenous) Fijians are recruited in the British armed forces.

These economic ventures, over which island states exert various levels of control, as well as the transnationalization of island populations, have had a considerable impact on the everyday lives of islanders, both materially and culturally. Today, many inhabitants of societies that are the dispersal centers of diasporas lead lives that refer at once to a local and a global context, generating a 'bifocality' of social life that aligns these societies with other diasporic societies of the world (Besnier, 2011). At the same time, island countries have become not only the departure point of labor migration but also its destination, particularly for labor migrants from the People's Republic of China. One manifestation of these complicated dynamics is the emergence, since the last decade of the twentieth century, of new globalized forms of charismatic and Pentecostal Christianity that challenge the hegemony of Christian denominations that have become definers of 'tradition' in the last two centuries and link disenfranchised citizens to global forms of belonging.

The profound changes that have accompanied the emergence of modernity have had a negative effect on the health of islanders. In addition to the surge in cancer rates among populations that have been affected by nuclear testing in the Marshall Islands and French Polynesia, noncommunicable diseases have increased dramatically. Nauru is reputed to have the highest rate of type-2 diabetes in the world, with Samoa and Tonga ranking very high as well. Contributing factors are a genetic propensity for obesity, increasingly sedentary lifestyles, and dramatic changes in dietary habits, including the consumption of low-quality fatty imported meats (Gewertz and Errington, 2010).

Finally, global warming, perhaps the most complex manifestation of the global condition, is poised to affect the region in a particularly dramatic fashion. Polynesia and Micronesia include several political entities that are in danger of complete disappearance within a few decades unless the industrial powers of the world implement radical policies. These include Tuvalu, Tokelau, Kiribati, and the Marshall Islands (as well as, elsewhere in the world, the Maldives and several Caribbean nations), all of which consist of low-lying atolls that are immediately threatened by rising sea levels. Of course, like all disasters and impending disasters, the predicaments faced by these countries and their peoples are as much an ecological phenomenon as a matter of discursive representation. Some politicians and citizens of these countries, for example, have argued that providing health care, adequate transportation, and unimpeded access to labor markets are more pressing priorities than global warming. Clearly, the future of the region is the object of multiple positions of a wide variety of global issues that will, one hopes, take into account the local dynamics of ordinary people's daily lives.

See also: Climate Change; Enlightenment; Oceania, Sociocultural Overviews: Melanesia.

Bibliography

Besnier, N., 2011. On the Edge of the Global: Modern Anxieties in a Pacific Island Nation. Stanford University Press, Stanford, CA.

Besnier, N., Alexeyeff, K. (Eds.), 2013. Gender on the Edge: Transgender, Gay, and Other Pacific Islanders. University of Hawai'i Press, Honolulu.

Denoon, D. (Ed.), 1997. The Cambridge History of the Pacific Islanders. Cambridge University Press, Cambridge, UK.

Gewertz, D., Errington, F., 2010. Cheap Meat: Flap Food Nations in the Pacific Islands. University of California Press, Berkeley, CA.
Hau'ofa, E., 1994. Our sea of islands. The Contemporary Pacific 6, 148–161.
Howard, A., Borofsky, R. (Eds.), 1989. Developments in Polynesian Ethnology. University of Hawai'i Press, Honolulu.
Johnston, B.R., Barker, H.M., 2008. The Consequential Damages of Nuclear War: The Rongelap Report. Left Coast Press, Walnut Creek, CA.
Kirch, P.V., 2000. On the Road of the Winds: An Archaeological History of the Pacific Islands Before European Contact. University of California Press, Berkeley, CA.
Kiste, R.C., Marshall, M. (Eds.), 1999. American Anthropology in Micronesia: An Assessment. University of Hawai'i Press, Honolulu.
Lal, B.V., Fortune, K. (Eds.), 2000. The Pacific Islands: An Encyclopedia. University of Hawai'i Press, Honolulu.
Moorhead, A., 1987. The Fatal Impact: The Invasion of the South Pacific, 1767–1840. Harper & Row, New York.
Obeyesekere, G., 1992. The Apotheosis of Captain Cook: European Mythmaking in the Pacific. Princeton University Press, Princeton, NJ.
Poyer, L., Falgout, S., Carucci, L. (Eds.), 2001. The Typhoon of War: Micronesian Experiences of the Pacific War. University of Hawai'i Press, Honolulu.
Sahlins, M., 1958. Social Stratification in Polynesia. University of Washington Press, Seattle.
Sahlins, M., 1987. Islands of History. University of Chicago Press, Chicago.
Smith, B., 1985/1969. European Vision and the South Pacific. Yale University Press, New Haven, CT.
Tcherkézoff, S., 2008. Polynésie/Mélanésie: L'invention française des 'races' et des régions de l'Océanie (XVIe–XIXe siècles). Aux Vents de Îles, Papeete, Tahiti.

Oil Age: Energy in Transition

Colin J Campbell, Cork, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Oil Age opened about 150 years ago and provided an abundant supply of cheap energy that fueled the world's economy, allowing the population to expand sevenfold. But oil and gas are finite resources that were formed in the geological past under processes that are now well understood. It means that they are subject to depletion: in other word's, for every gallon used, one less remains.

Analyzing the status of depletion is difficult because of confusion over the different categories of oil and gas, each having its own endowment, costs, and characteristics. Public information on discovery and reserves is also very unreliable. The skills of a detective are called upon to build a realistic model but a sound picture of the general situation begins to emerge. It shows that we are now about half way through the Oil Age, facing a very difficult transition from expansion to contraction with far-reaching social, economic, and political consequences.

Background

The planet, whose surface is heated by the sun, has had a long geological history spanning billions of years, and has seen many changes ranging from ice ages to epochs of global warming. The continents have broken apart and collided on the back of deep-seated convection currents in the Earth's crust. There have also been epochs of intense volcanic activity putting out massive lava flows and ash clouds, which have changed the composition of both the atmosphere and the oceans. Mountains were uplifted as continents collided, exposing rocks that were subject to erosion, the proceeds of which slipped downward under the forces of gravity to be deposited as gravels, sands, and clays in lowlands, lakes, and seas. Limestone and salt were also deposited under special conditions.

Chemical reactions over 500 million years ago led to early forms of life in lakes and seas, and the remains were preserved as fossils. Some, including, for example, the limpet (or *Patella* to give it its scientific name), have remained little changed, but other species evolved and proliferated when they found a niche that suited them only to die out when that closed from natural causes.

Homo sapiens arrived about 200 000 years ago followed by what could be described as *Modern Man* appearing about 12 000 years ago. People living in any particular chapter of subsequent history naturally assumed that it was the norm and expected it to continue, but history shows that there have often been radical changes, sometimes at short notice.

The Origin of Petroleum

Lough Hyne in SW Ireland is a lake with a narrow channel leading to the sea. It supports a rich fauna and flora, but in the summer, the surface waters warm up and circulation ceases with the result that the depths become stagnant, being deprived of oxygen. As a result its organic life dies. This is a present-day small example of the conditions responsible for the formation of petroleum. In fact, much of the world's oil was formed in two epochs of global warming 90 and 150 million years ago. Algae proliferated in the warm sunlit waters of lakes and restricted seaways in rifts that formed where the continents pulled apart. The organic remains sank to the stagnant depth to be duly buried by sands and clays washed in from the adjoining continents. The organic material was subject to chemical reactions becoming what are known as *kerogens*, some being oil prone and others gas prone. When buried to a depth of about 2000 m, they were heated enough to be converted into petroleum. If buried too deeply into zones of high temperature, the oil broke down into gas. In some cases, subsidence allowed the sea to break into the stagnant lakes and become subject to evaporation leading to the deposition of salt, which formed an excellent seal, preventing the escape of underlying petroleum.

Once formed in so-called source rocks, the oil and gas began to migrate upward to zones of lesser pressure, given that there was enough permeability, partly in the form of fractures, to allow movement. Some escaped at the surface to be weathered and degraded, with the tar sands of Canada being a well-known example. But, in some cases it was trapped in a reservoir of porous rocks at the top of domelike geological structures, known as anticlines. Other traps were formed by faults and where the reservoir pinched out. Such features form objectives for oil exploration. It did not take long for the early explorers to realize that finding an oil field depended on locating a place where the geological conditions of source, trap, reservoir, and seal came together, with timing being an additional important factor. There are many uncertainties, and the geological interpretation requires testing by an exploration borehole, known as a *New Field Wildcat*.

Petroleum Development

Oil and gas have been known from surface seepages since the earliest days, probably being responsible for the biblical burning bush. Commercial development began in the Nineteenth Century with an early mile stone being at Pichelbronn in eastern France where a heavy oil deposit was exploited with shallow wells in 1830. It was followed by other

developments in Romania and on the shores of the Caspian, with an important turning point being the discovery of oil in Pennsylvania in 1859.

At first, the explorers mapped the rocks exposed at the surface, using no more than a hammer, a hand lens, and a notebook, and succeeded in finding many of the major onshore productive provinces. Later came geophysical techniques whereby energy was released from an explosion or other source, and recorders measured the time taken for the echoes to return from rock surfaces far underground, which allowed the geology at depth to be mapped in detail. In recent years, computing power has added ever greater sophistication allowing even the smallest and most subtle prospects to be identified. Geochemistry has also provided a critical insight into where the oil and gas were formed.

As the onshore possibilities dwindled, the industry turned its attention offshore finding extensions of known provinces as well as entirely new ones. This involved the development of special geophysical and drilling equipment. The turning point came with the semisubmersible rig mounted on pontoons floating in stable conditions below the wave base. As the easier possibilities were exhausted, exploration extended into deeper waters, which called for ever more sophisticated technology but faced the occasional serious accident.

The drilling process is indeed complex and calls for a brief description. A drilling bit, even impregnated with diamonds to cut the rocks, is attached to a rotating pipe and drills an initial hole with a wide diameter of as much as 30 inches. The hole is progressively cased with steel pipe, which is cemented into the adjoining rocks, before drilling resumes in progressively narrower holes. A circulating heavy mud, impregnated with barytes, is pumped into the borehole to hold back the pressure from any oil or gas encountered, lubricate the bit, and bring the rock cuttings to the surface, where they are examined to reveal more of the geology. Such boreholes are normally drilled to depths of about 10 000 feet, although in some cases much deeper. They are duly logged by electronic devices to indicate the nature of the rocks penetrated and identify any oil- or gas-bearing zones. If such zones are found, the casing has to be perforated to allow the oil or gas to flow into the well. The borehole starts vertical but can be deviated at depth to run parallel with the strata and thereby put itself into contact with more productive rock. In some cases, an explosive charge was released in the well to fracture the reservoir making it more productive, and that has been followed in recent years by so-called fracking, whereby liquids and chemicals at high pressure are pumped into the rock to increase its permeability.

Once a find was made, attention turned to optimizing the development. This involved drilling appraisal wells to delineate the precise size and nature of the find before investing in the closely spaced development wells to produce it. The infrastructure, including pipelines, had to be planned and constructed at great expense, especially offshore. It was normal and very reasonable to make conservative estimates of the size of the field to minimize the risks to investment, which led to so-called *reserve growth* over the life of the field. The high oil prices of recent years have allowed ever smaller fields to be put on production and have also extended the lives of existing fields, especially offshore. But ironically, these additions leave less for the future.

Control of the Industry and Oil Price

The Standard Oil Company took a dominant early position in the US oil industry, starting primarily as a marketer and refiner before moving progressively into exploration and development, but its power was broken by antitrust legislation in 1911. Even so, three of its daughters, Exxon, Mobil, and Chevron, grew to become major oil companies in their own right, being joined by Texaco and Gulf, following the major discoveries in Texas in the 1930s, and Shell and British Petroleum (BP) in Europe. Together these companies, known as the *Seven Sisters*, controlled the world's supply. They even had a secret meeting in Achnacarry Castle in Scotland in 1928 to regulate production between themselves to support price.

Governments also began to take a role. The US Government moved in the 1930s to support price by limiting production to a certain number of days a month in a policy administered by the Texas Railroad Commission. Communist Russia nationalized its oil industry in 1928, and Mexico followed this example ten years later.

In most countries, mineral rights are owned by the Government that leases them out to oil companies as concessions lasting a set number of years and subject to drilling obligations. But the United States is an exception with the mineral rights belonging to the landowner. The oil fields became important financial assets, leading the Stock Market to impose strict rules on what could be reported as reserves, while smiling on under-reporting as commercial prudence. So-called *Proved Producing Reserves* represented the expected future production from producing wells, while *Proved Undeveloped Reserves* referred to that coming from low-risk infill wells between the existing ones before they had been drilled. *Probable* and *Possible Reserves* were also recognized but had little financial significance. A more sophisticated classification, known as the *Petroleum Resource Management System*, has been adopted in recent years.

In 1960 came the formation of OPEC (*The Organisation of Petroleum Exporting Countries*), which followed the example of the Texas Railroad Commission by limiting the production of its members to support price, the quota being based in part on their reported reserves. But it ran into difficulties when prices fell in the 1980s following the opening of the North Sea and other new offshore areas. In 1984, Kuwait was reporting reserves of 64 Gb (billion barrels) consistent with the long prior trend, but in the following year increased its report to 90 Gb although nothing particular had changed in the oil fields. Possibly, it started reporting *Original* rather than *Remaining* Reserves. Its motive presumably was to secure increased revenue from its resulting higher OPEC quota. Two years later, it announced a possibly genuine small increase to 92 Gb, but this proved too much for its rival OPEC members. Abu Dhabi decided to match Kuwait exactly at 92 Gb (up from 31 Gb); Iran went one better at 93 (up from 49 Gb); and Iraq capped both at a rounded 100 Gb (up from 47 Gb). Venezuela for its part jumped from 25 to 56 Gb but did so by including its heavy *nonconventional* oil that had not previously been counted for quota purposes. Saudi Arabia could not match Kuwait because it was already reporting more but in 1990 announced a massive increase from 179 to 258 Gb to hold its lead. The numbers suggest that Kuwait may have changed to report Original rather than Remaining Reserves, namely by not deducting past production.

The major companies naturally needed to track discovery and the size of reserves around the world in order to plan their investments. In earlier years they provided informal genuine information to Petroconsultants in Geneva, but it was sold on the death of its owner.

There are now several public sources of data on reserves including annual assessments by the Oil & Gas Journal and the BP Statistical Review, which carries a significant footnote to the effect that its published numbers do not reflect the company's own knowledge. The Energy Information Agency, an arm of the US Government, also publishes comprehensive data. There are in addition many web sites providing official assessments by country. These datasets commonly contain different information on reserves, and some show unchanged estimates year on year despite the utterly implausible implication that new discovery and/or *reserve growth* exactly matched the intervening production. There are in addition two industry datasets which are very costly and subject to strict confidentiality, such that they cannot be used for published analyses.

The challenge of determining accurate reserve data has become ever more difficult with the proliferation of small speculative companies and the increasing role of State entities, reluctant to release confidential information. Despite these difficulties, a reasonable approximation of future production by country and region can be made as summarized below.

The Status of Depletion

The first step in studying the depletion profile is to clearly identify the different categories of oil and gas of which there is no standard classification, a cause of further confusion. It is therefore useful to define and identify the following:

1. *Regular Conventional Oil and Gas*: a liquid, known as *condensate*, which naturally condenses from gas, may be treated together with *Regular Conventional Oil*.
2. *Nonconventional Oil and Gas*: made up of the following components:
   a. *heavy oils*: with a density greater than 17.5° API, including bitumen;
   b. *oil shale*: oil that can be produced by heating immature source rocks;
   c. *tight oil and gas* (*also called shale oil and gas*): as derived from rocks lacking adequate natural porosity and permeability that can yield production when artificially fractured;
   d. *deepwater oil and gas*: in water depths greater than 500 m.
   e. polar oil and gas;
   f. natural gas liquids from gas plants;
   g. *other nonconventional gases*: coalbed methane, hydrates, etc.

Regular Conventional, which was relatively easy and cheap to produce, has dominated all past production and is likely to continue to do so for many years to come, although *nonconventional* sources are being increasingly tapped.

It has also been found useful to treat everything in terms of production to the end of the Century to avoid having to worry about some insignificant tail end. Subtracting past production from the estimated total gives future production, a certain percentage being deemed to come from known fields, namely *reserves*, with the balance being yet-to-find.

There are 64 significant oil-producing countries to be evaluated. The study involves many iterative steps, starting with the input of past production, discovery, and the number of exploration boreholes drilled per year. The larger fields were usually found first for obvious reasons, and plotting cumulative past discovery against exploration boreholes, the so-called *creaming curve*, can be extrapolated to give an indication of the total, as illustrated by the example of Norway below. Another method is to analyze the distribution of field size against rank with the *parabolic fractal* (Laherrère, 1996). The indicated total is then evaluated and adjusted if necessary taking into account various relationships, including the status of depletion, with the peak of production normally coming approximately at the midpoint of depletion. Some 50 countries are now producing less than at some date in the past, many being long into decline.

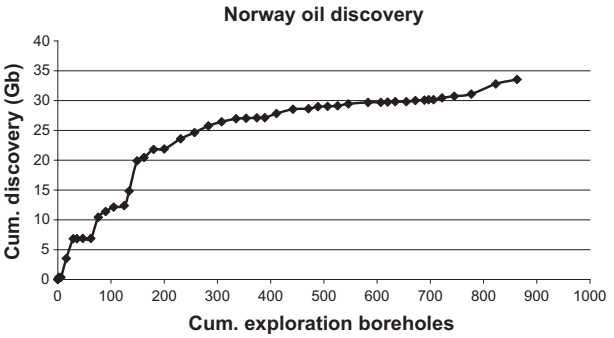

The following graphs summarize the conclusions. The first shows that discovery peaked in the 1960s, as confirmed by an Exxon executive (Longwell, 2002), such that by 1981 we were producing more than what was being found in new fields. The second plots the depletion profile of all categories of oil and gas. *Regular Conventional Oil* peaked in 2005, but all categories have followed a plateau over the past few years which may extend further. One year will obviously deliver a high point on that plateau, such being tentatively identified as coming in 2015. A debate rages as to the precise date but misses the point when what matters is the vision of the long decline on the other side of it. The longer the plateau, the steeper will be the subsequent decline. *Regular Conventional Gas* is expected to peak around 2015. More details are given in the following table and in the Atlas of Oil and Gas Depletion (Campbell, 2013).

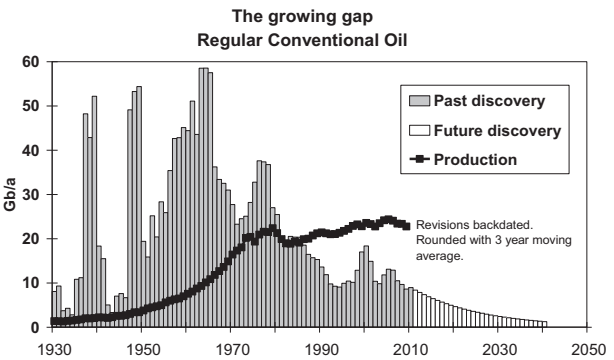

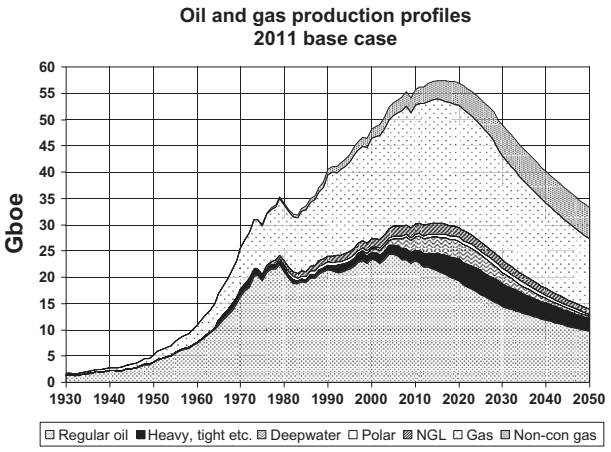

The Wider Significance in a Historical Context

The species *H. sapiens* arrived about 200 000 years ago, and settled agriculture followed about 12 000 years ago. Crops had to be stored between harvests and those controlling the storehouse had to keep accounts of how much was received and returned to the farmers. They were also able to give preference to the privileged members of their societies and charge interest for the loan of food after a bad harvest. In other words, they laid the foundations for banking.

We may imagine that at some point someone found a nugget of gold in a riverbed and showed it to a neighbor, who liked it even more, perhaps offering to exchange it for a sack of corn. It came to facilitate barter, laying the foundations for trade, and silver had a similar role.

Stone Age Man had used flints before people turned to bronze, iron, and steel for better tools and weapons. These

Resource based production forecast

Regular Conventional Oil by region to 2100

						Total (Gb)		*Percent*		*Peak dates*		
Mb day^{-1}	*2000*	*2010*	*2020*	*2030*	*DR (%)*	*Past*	*Future*	*Disc*	*Dep*	*Expl*	*Disc*	*Prod*
Africa	7.3	6.9	5.3	4.0	2.0	106	96	94	53	1981	1961	2005
Asia-Pacific	4.3	3.5	2.3	1.5	3.9	55	30	97	65	1990	1974	2000
Eurasia	11.2	15.8	12.1	8.5	3.4	231	166	93	58	1982	1960	2011
Europe	6.1	3.5	1.9	1.1	5.3	57	21	96	73	1986	1974	1999
L. America	8.1	5.7	3.9	2.5	3.5	127	60	96	68	1982	1914	1998
ME Gulf	18.8	19.8	22.3	18.3	1.6	300	416	94	39	1974	1948	2021
ME Minor	2.8	2.7	2.2	1.8	2.7	29	36	96	45	1991	1965	2000
N. America	5.1	4.2	2.4	1.3	5.5	201	25	99	89	1981	1930	1973
Other	0.5	I.0	0.7	0.4	4.6	9	8	81	54			2008
WORLD	64	63	53	39	2.5	1115	884	95	56	1981	1964	2005

Regular Conventional Gas region to 2100

						Total (Gb)		*Percent*		*Peak dates*		
Tcf/a	*2000*	*2010*	*2020*	*2030*	*DR (%)*	*Past*	*Future*	*Disc*	*Dep*	*Expl*	*Disc*	*Prod*
Africa	8.8	13.5	14.4	11.3	2.5	269	524	92	34	1963	1957	2022
Asia-Pacific	8.8	13.0	15.2	14.2	2.3	250	560	89	31	1990	1973	2021
Eurasia	25.3	31.2	42.1	34.3	2.1	1003	1560	85	39	1988	1966	2021
Europe	11.9	11.8	7.5	3.3	5.7	384	177	98	68	1986	1959	2004
L. America	7.2	10.5	9.2	6.4	2.7	245	380	93	39	1982	1941	2008
ME Gulf	8.1	15.1	17.6	18.1	0.9	310	1695	83	15	1992	1971	2030
ME Minor	3.2	7.2	9.5	13.3	0.8	95	1027	96	8	1992	1971	2030
N. America	31.4	21.7	9.8	5.1	6.9	1369	281	98	83	1981	1930	1973
Other	1.2	2.1	2.1	1.1	3.1	48	52	87	48			2007
WORLD	106	127	129	110	1.9	3975	6525	88	38	1981	1971	2015

Notes: DR refers to depletion rate (annual as a percent of future). Dep. stands for depleted. Eurasia is the former Communist bloc and China. N. America is USA & Canada. ME Gulf is Iran, Iraq, Kuwait, Neutral Zone, Saudi Arabia, and United Arab Emirates. Only major producing countries are included in each region, with the rest in Other The World Totals include a rounding of −19 Gb for oil and +271 Tcf for gas.

minerals were dug from surface outcrops before the pits were deepened into mines, but they flooded on hitting the water table. For fuel, people relied on wood and charcoal before turning to coal, also dug from pits that were deepened into mines.

Trade and industry developed over time, and the Eighteenth Century saw the birth of the so-called *Industrial Revolution* which led to the rapid increase in the demand for coal and minerals. Draining the mines led to a remarkable technological development: the hand pump evolved into the steam pump which in turn led to the steam engine. It greatly facilitated transport with the construction of railways and steam-driven ships that further encouraged trade and industry. The next important step came when a way was found to insert the fuel directly into the cylinder, giving the *Internal Combustion Engine* which was much more efficient. At first, it used benzene distilled from coal before turning to petroleum refined from crude oil. The first automobiles took to the road around 1880 and the first tractor ploughed its furrow in 1907 leading to the rapid expansion of agriculture to feed the expanding population, which then stood at about 1.5 billion.

This new prolific source of cheap energy changed the world in radical ways allowing the population to expand sevenfold in parallel. But it was not plain sailing as trade prompted the growth of empires which in turn led to major conflicts between rival powers. Whereas, in earlier years, short battles had been practical ways of settling disputes, the advances in weapons technology and the provision of railways to keep the front supplied with munitions and reinforcements transformed the First World War of 1914–18 into an appalling exercise of attrition on the killing fields of Flanders. Wars were expensive exercises largely funded by debt, which had to be settled after the war, leading in turn to social and political tensions. The Second World War of 1939–45 was essentially an extension of the First.

The British Empire was extinguished by the war, and the United States became the world's premier power with the dollar being adopted as a world trading currency. The so-called Cold War followed allowing the United States to increase its dominion around the world with military bases which indirectly strengthened its financial power and thereby trade. The dollar was at first backed by gold, but that practice was abandoned in 1971, allowing the financial institutions to adopt ever more complex and speculative manipulations, with debt itself even becoming a tradable commodity. The US industrial base, however, began to decline as its companies increasingly moved overseas for cheaper labor.

But whatever the conflicts, it was generally a successful chapter in history allowing the so-called standard of living for many people to grow as trade expanded. Women no longer saw their prime role as wives and mothers but developed careers and played an important part in the new consumeristic society. Radio, television, and, later, the Internet had a huge impact giving people a wider view of the world, subjecting them to the pressure of advertising.

The economic prosperity of this chapter of history was fueled by cheap oil-based energy allowing the roads to become choked with traffic while airliners carrying holiday-makers to distant beaches left trails across the skies. It marked the *First Half* of the *Oil Age* with production rising from about 4 million barrels a day in 1930 to pass 10 million in 1950 and 80 million in 2004 (a barrel contains 42 gallons). Oil prices were relatively stable over the last Century averaging $25 a barrel (quoted in terms of 2011 dollars), although there were surges to $53 in 1974, when OPEC countries restricted supply to counter US support for Israel, and to $101 in 1980 on fears of further interruptions following the fall of the Shah of Iran.

But the depletion of oil, which is a finite resource formed in the geological past, as described above, began to influence prices as the present Century opened. The major companies began to find it easier to acquire reserves than to find them by exploration, such that the *Seven Sisters* were reduced to four by merger. Most production came from *Regular Conventional* sources and prices reached to $63 in 2005 when its natural peak of production was passed. Traders spotted the rising trend and bought contracts on the Future's Market which led prices to soar to almost $150 by mid-2008. The scale of speculation is illustrated by the fact that contracts on the Future's Market exceeded actual production by factors of 10 to 30 (Tolub and Erb, 2010). The traders, realizing that the high prices would cause recession dampening demand, then started selling *short* on the market, while the industry for its part began draining its storage tanks to benefit from the high prices while they lasted. Prices accordingly fell back to $64 in 2009. Prices anomalously fell again in late 2014 to around $80, due, it is said to action by Saudi Arabia but it is too soon to be sure. The downturn is likely to be short-lived.

Governments, following outdated economic principles, reacted to the recession by printing yet more money out of thin air in the hope of stimulating consumerism to restore past prosperity. They met with a brief success such that the demand for oil again rose, and oil prices passed $111 in 2011, suggesting that recession is set to resume.

The industry reacted to the peak of *Regular Conventional Oil* in 2005 by turning more attention to *nonconventional* sources which are both more expensive and slower to extract. Wells were drilled to great depths in ever deeper waters, especially in the Gulf of Mexico and off Brazil and West Africa. In the United States, great attention has recently turned to so-called fracking, whereby expensive and short-lived highly deviated wells can extract limited production from rocks lacking adequate natural porosity and permeability to form a reservoir. The resource in the ground is enormous and unquantifiable, but extraction is slow and very costly, also facing environmental objections. No one would turn to this supply if there was adequate *Regular Conventional Oil* left.

The European Union was formed out of 27 countries, supporting a population of 550 million, to provide a common market and end the cause of conflicts that had led to the two world wars. Many of these countries adopted a common currency, the Euro, but it too has run into financial difficulties. In earlier years, Europe had led the world in terms of coal mining, which provided the energy for its rapid industrialization, but that production has long been in decline, Britain's production having peaked in 1914. France in particular turned to nuclear energy as a response. But the discovery of oil and gas in the North Sea led to a surge of new energy, although it too is subject to depletion with oil production passing a peak in 2000 and gas 4 years later. The high level of technical expertise and the stable operating conditions ironically accelerated depletion such that the *Depletion Rate* (Annual as a percentage of Future) stands at a relatively high level at 5% for oil and slightly more

for gas. It means that oil and gas production will have fallen to about half their present levels in about 10 years' time.

As countries begin to appreciate the depletion of their oil and gas, they are likely to restrict exports to conserve as much as possible for their own future needs. It makes eminent national sense although offending the principles of globalism whereby the resources of any country belong to the highest bidder. Argentina has already taken a lead in this direction, and King Abdullah of Saudi Arabia has perceptively said that he wishes to leave as much wealth as possible in the ground for his grandsons.

Every country and region faces comparable changes, such that the world production of *Regular Conventional Oil* and *Gas* will have fallen to half their present levels by about 2040 and 2065 respectively, suggesting that by the middle of the Century the world will be able to support no more than about half its present population in its present way of life.

The transition imposed by Peak Oil and Gas threatens to be a time of great tension as indeed has already been observed in the form of riots and revolutions around the world. Food prices have increased fivefold reflecting the rising cost of oil, agriculture having been described as a process which turns oil into food (Brown, 2012). The so-called *Arab Spring* affecting countries in North Africa and the Middle East was particularly severe. People became resentful and blamed their governments for their changed circumstances. Some of these countries are important oil producers and may be subject to foreign military intervention, as was the case in Libya.

The transition has also particularly affected the financial system. The banks had been lending more than they had on deposit confident that *Tomorrow's Economic Growth* was collateral for *To-day's Debt*. But growth is no longer possible without cheap oil-based energy, explaining why debt went bad. The United States Empire is essentially a financial one, and it faces a particularly difficult transition. It built up a substantial military presence around the world during the Cold War which helped ensure that countries used the dollar for trade and held their reserves in that currency. The North Atlantic Treaty Organization (NATO) was established as a defensive pact whereby it would go to the support of any of its members under attack, but the rules were progressively modified allowing intervention if its *vital interests* were at risk, oil supply being certainly such. Indeed, President Bush justified the invasion of Iraq with the words – *our energy supply was at risk*. The fall of the Soviets led to the opening of the oil-rich Caspian, which was once thought might rival the Middle East, and the United States attacked Afghanistan lying on a proposed export pipeline route. Iran, another major oil producer, has been under threat, having been subject to financial sanctions by the United States and some of its European supporters.

But it is not a hopeless situation as people and their governments can react in positive ways to meet the changed circumstances, once they come to understand that the depletion of the critical energy supply that supports their current style of life is imposed by natural depletion. It seems likely that there will be a reversion to localization as people again come to rely on whatever their particular region can support. Scotland is already considering seceding from the United Kingdom, as many as 45% of the voters agreeing in a recent referendum, as does the province of Catalonia from Spain. The various states making up the United States, which already have considerable political power, may also move to reduce the authority of the Federal Government, even if they do not formally secede from the Union.

It is not difficult to think of specific policies that would help reduce energy demand:

1. Adopt the *Oil Depletion Protocol*: Already the Portuguese Parliament has passed a resolution urging its Government to adopt a policy for cutting consumption to match world depletion rate (*Annual* as a percentage of estimated *Future* production), and a new political party in the Spanish province of Galicia has adopted it as part of its platform (Partido de la Terra). Depletion is currently running at no more than about 2.5% a year for *Regular Conventional Oil* and less than 1% for all categories. Cutting demand by such small amounts should be readily attainable. Such policies would also hold prices down helping the poorer nations and reduce the tensions.
2. Progressively disallow energy costs as a charge against corporate taxable income, which is a form of hidden subsidy. Management, facing these costs head-on, would likely react by paying more attention to energy efficiency, for which there is much scope.
3. Install smart meters so that households may be more aware of their electricity consumption, much of which is generated from gas, and impose new tariffs whereby the cost per unit rises with increasing consumption. France already has meters with a red light that comes on at periods of peak demand increasing the tariff at those times significantly.
4. Encourage car sharing whereby people come to give lifts, sharing the cost of the fuel, which itself might be subject to an affordable tradable ration, providing for minimal essential needs to people in different circumstances.
5. Encourage the development of energy from tides, waves, and wind, as well as that from solar and geothermal sources. Home-owned windmills and solar-power facilities can make households largely self-sufficient, even selling a surplus back to the grid. Anaerobic digestion, by which urban and agricultural organic waste is used to generate methane, can also contribute.
6. Encourage a new regionalism with local currencies, managed responsibly, whereby people may become less oil-dependent and again rely on what their particular region can support. This might prompt a new more positive democracy whereby the leaders are better linked to the people they represent, being freed from hidden financial and other pressures.
7. Above all, take measures to reduce the world population.

In conclusion, it may be said that the evolution of life on Earth has seen many changes. *H. sapiens* does not necessarily face an early extinction, but *Petroleum Man* will be about gone by the end of the Century. The challenges for survival are real but can be met once the issues at stake are better appreciated. Indeed, it is promising that the subject is now attracting much interest with academic research, notably by the UGES Research Group at Uppsala University in Sweden established by Prof. Aleklett, who also heads the Association for the Study of Peak Oil and Gas ('ASPO') with associates in more than 30 countries. But the challenges of adapting to the changes resulting from the radical fall in the principal energy source supplying the modern world are daunting.

Bibliography

Aleklett, K., 2012. Peeking at Peak Oil, ISBN 978-1-4614-3423-8.

Brown, L.R., 2012. Full Planet, Empty Plates, ISBN 978-0-393-34415-8.

Campbell, 2013, An Atlas of Oil and Gas Depletion, second ed. ISBN:978-1-4614-3575-4 Springer 2013. See also Campbell, C.J., 1997. The Coming Oil Crisis. ISBN:0-906522-11-0; Campbell, C.J., 2005. Oil Crisis. ISBN:0906522-39-0; Campbell, C.J. (Ed.), 2011. Peak Oil Personalities. ISBN:978-1-908378-06-4, and the many references contained therein.

Fleming, D., 2011. Lean Logic, ISBN 978-0-955-0849-6-6.

Heinberg, R., 2006. The Oil Depletion Protocol. ISBN:10-0-86571-563-7.

Laherrère, J.H., 1996. Distributions de type 'fractal parabolique' dans la nature. Comptes rendus de l'Académie des Sciences Paris 322 lla, 535–541.

Longwell, H., 2002. The future of the oil industry: past approaches, new challenges. World Energy 5 (3).

Partido de Terra, see http://wikipedia.org/wiki/Land Party.

Tolub, L., Erb, M.-A., 2010. Oil price band for the next decade: Utopia versus Reality. Swiss Derivatives Review 43.

Old Age and Centenarians: The Human 'Warranty Period'

Bruce A Carnes, College of Medicine, The University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

More humans are surviving to older ages than ever before. The age structure of our species is shifting from young to old, a phenomenon referred to as population aging. It is global in nature, uniquely human, and has become one of the greatest challenges ever encountered by our species. The basis for this assertion is that biological function (homeostasis) declines with advancing age and at the same time pathologies, frailty, and disability increase. Given these biological realities, this article examines whether humans have a warranty period, when do we grow old, and what are the odds of becoming a centenarian.

Humans are the only species who live most of their lives knowing they will eventually die. This knowledge has made duration of life a topic of discussion and conjecture throughout recorded history. Terminology related to duration of life has proliferated over the centuries. However, without a conceptual framework, this terminology has created confusion and misunderstanding across scientific disciplines as well as in the general public. In an age of population aging, scientific clarity is an essential precursor to the creation and implementation of effective social policy (*see* Population Aging: Economic and Social Consequences). An interdisciplinary framework for clarifying this terminology will be provided. The article will also examine commonly asked questions about age, aging, and longevity. For example: At what age do we become old? Can everyone become a centenarian? How long can humans live? We will examine whether these are even the right questions to ask. Identifying appropriate questions and answers is especially important now because the individual and societal consequences of extended longevity are threatening the social fabric of nations, communities, and families around the world.

Terminology and Premise

The terminology discussion begins with the most fundamental term, life span (Carnes, 2011). *Life span* is just what it sounds like; it is a term that describes the observed length of a person's life. It is an 'obituary' (birth to death) concept that replaces the older and more cumbersome *duration of life* terminology. *Longevity* is also defined as length of life, but dictionaries associate this term with a long duration of life. The latter distinction is important because genetic analyses of longevity suggest that it is heritable (i.e., influenced by genes). Life span lacks this genetic distinction (*see* Lifespan Development, Theory of).

Although seemingly straightforward, life span remains a fuzzy concept that creates considerable confusion in the field of gerontology. The basis for this confusion begins with the premise that nearly every death is premature. In other words, most deaths are avoidable, and human history can be viewed as a continuous search for ways to avoid avoidable mortality. By implication, everyone has a *potential life span* that under better circumstances (improved environment, better medical care) is within their grasp. This ability to pursue our own potential life span is personally empowering, further distinguishes humans from other species, and motivates the interdisciplinary pursuit of social policies aimed at increasing healthy life span.

The time between observed and potential life span is reclaimable without modifying our biology. Dentistry, medicine, public health, and other related pursuits are creations of human ingenuity intentionally designed to improve health and *manufacture survival time* (Olshansky et al., 1998). The clearest evidence of success for this remarkable Lamarckian (technological) endeavor is the rapid climb in life expectancy that has occurred over the last century. Our species has bypassed the rules of evolution through adaptations to environmental challenges that were achieved via culture and technology rather than biology.

The Great Unnatural Experiment

Peter Medawar (1952), Nobel laureate and one of the founding fathers of gerontology, declared that senility (or what we call senescence or aging today) is "an artifact of domestication; that is, something revealed and made manifest only by the most unnatural experiment of prolonging an animal's life by sheltering it from the hazards of its ordinary existence." Medawar saw the future long before the pace of manufacturing survival time had accelerated or the term population aging even existed or was an issue of societal concern. He understood that the future would be a world that bears almost no resemblance to the world from which our species and our biology emerged.

Human ingenuity not only created the opportunity for many to achieve or approximate their potential life span, it may have reached a level of technological sophistication that allows some people to exceed their potential life span as well. People who would have died at an earlier age in the past are routinely saved today. Historically, the older segments of human populations were populated by the robust survivors of an unforgiving gauntlet of death (demographic selection). Today, the older segments of our populations are far more diverse in a health sense than in the past and will be even more diverse as new ways are found to save lives in the future. Lowering the barriers to extended survival and the biomedical and social consequences that arise from it are already being felt around

the world (*see* Retirement and Encore Adulthood: The New Later Life Course).

Asking the Right Questions

For most of human history, life expectancy at birth for humans has been below 50 years. However, since 1900, life expectancy has climbed dramatically. This demographic trend has led both the public and the scientific community to ask: How long can humans live? The first reaction to this question is that it has no single answer because humans are not genetic clones. Madame Jeanne Calment, the longevity world record holder, died in 1997 at the age of 122 years and 164 days. As the cumulative number of humans ever born continues to increase, the odds will increase that her record will be broken some day. However, for every longevity record holder, there will be millions of fetuses that die before birth and billions of people who will die at double digit ages.

The rareness of people who achieve extreme longevity makes them a curiosity. Based on the 2006 U.S. period mortality table, the probability of surviving from birth to age 100 is 2.5% for females and 1.1% for males (NBER, 2013). However, demographers point out that life expectancy has increased at a pace of 1/4 year per year over the last two centuries (Burger et al., 2012). If this trend continues in high life expectancy countries, then demographic models predict that 50% of the babies born since the year 2000 will live to celebrate their 100th birthday (Christensen et al., 2009). If this prediction of extended survival becomes a demographic reality, then ways to extend healthy life span become essential in order to avoid trading off longer life for progressively worsening health (Olshansky et al., 1991). From a biodemographic perspective, biology trumps demography because it is more proximate to the underlying causes of both longevity and aging. As such, asking how long humans must live is a more relevant scientific question than asking how long can humans live. The former has a biological answer but the latter does not.

Warranties and Their Expiration Dates

We are all familiar with the expiration date of warranties; we look for them when we buy eggs or milk at the store and when we purchase cars. Everything created, including the human body, has an implied warranty period. As such, it is scientifically reasonable to search for ways to estimate an expiration date for the biological warranty period of humans. In order to avoid falling into conceptual traps, a few comments are required. The first of these is that a biological warranty period has nothing to do with life span limits. As mentioned earlier, life span limits just like Olympic records are invariably broken and exist only in a probabilistic sense; that is, the probability of surviving to the longevity record diminishes as the record is progressively surpassed. The second clarification about warranty periods is that they are not a death sentence. Eggs can be eaten beyond the expiration date, and cars do not fall apart on the day their warranty expires. Instead, the warranty period establishes a boundary between a period of normal operation characterized by health and vigor and a period of extended operation, where maintaining health and vigor becomes progressively harder to achieve. In a world of population aging, identifying the age of this transition is exceedingly relevant to the development of realistic policies associated with retirement, medical care, elder health, and social justice.

Medical and social science researchers partition the human life course into childhood, adulthood, and a postreproductive period (*see* Motivation: Life Course and Sociological Perspectives). This development-based framework enables them to organize their medical training and practice into pediatrics, internal medicine, and geriatric medicine and to identify the unique medical and social challenges that occur within these phases of the life course. For social scientists, it allows them to examine the unique life phases related to reproduction, mating, and employment. A warranty period is different; it provides an answer to the question: "How long must humans live?" The logic goes as follows. Every living thing eventually dies. Given this biological reality, life on the earth must have found a solution to the inevitability of death and it did – reproduction. Thus, the answer to the question is "humans must live long enough to replace themselves," or as a biologist would say, they must live long enough to achieve Darwinian success.

Multicellular organisms are not born mature; it takes time for them to reach a level of developmental maturity that makes reproduction biologically possible. As such, a highly orchestrated (i.e., genetically programmed) sequence of developmental milestones must be successfully navigated: conception, growth and development, maturation, reproduction, nurturing, parenting, postreproductive parenting, and in humans, grandparenting. The window of time needed to accomplish these biological milestones has been called adaptive life span (Gurven and Kaplan, 2007), essential life span (Rattan, 2000), or biological warranty period (Carnes and Witten, 2014). The heritability (attributable to genetics) of generic life span is approximately 15%, but the heritability of longevity (long life) increases as it is progressively lengthened (Gögele et al., 2011). All three concepts (life span, longevity, and warranty period) are measures of time created by varying degrees of biology/genetics (*see* Behavioral Genetics: Psychological Perspectives).

Quantifying the Warranty Period

Quantifying the human warranty period requires associating chronological ages with developmental milestones (*see* Age Structure; Age Stratification). Medawar's unnatural experiment complicates this task by distorting the natural timing of the milestones. The inevitability of death (mortality pressure) essentially defines many of the milestones because it creates a literal race between reproduction and death. Death before reproduction is the ultimate biological failure (*see* Fertility Theory: Theory of Life History Evolution). In nature, organisms must reproduce as early as is physiologically possible. Our distant ancestors as well as extant hunter-gatherers would have to do the same. Thus, on average, sexual maturity is achieved at 16.1 years of age and the first birth occurs at 19.5 years of age in hunter-gatherer societies (Walker et al., 2006).

In developed countries (countries experiencing Medawar's unnatural experiment), age of sexual maturity drops, on average, to 12.5 years, presumably because of easy access to

a surplus of food. Conversely, age of the first birth in developed countries is delayed to 24 years for all women or 26.5 years for educated women. Further, first births beyond age 30 are not unusual. So, women in developed countries are reaching sexual maturity earlier and reproducing later. This shift in reproductive behavior creates a large disparity between age of menarche and the first birth; it is 3.4 years in hunter-gatherer societies and 11.5 or 14 years in developed countries depending on whether you measure it for all women or women who are educated (Piñón, 2002). These disparities between ancestral and modern ages for important biological milestones raise an important question: Was it our biology, our behavior, or both that changed in response to changing environments?

Evolutionary biologists (e.g., Pritchard, 2010) suggest that the pace of global colonization by humans outpaced any credible rate of biological evolution (*see* Lifespan Development: Evolutionary Perspectives). To be sure, genetic change within our species did occur (e.g., lactose, hypoxia, melanin, and sickle cell gene), but recent genomic studies suggest the baseline rate of mutation (evolutionary change) is 50% slower than previously thought (Arsdale, 2012). This finding leads to the conclusion that cultural and technological adaptations are largely responsible for why humans were able to populate the globe so rapidly. In other words, our biology has remained largely unchanged, but our culture and technology has undergone dramatic transformations. To come full circle, this is exactly what Medawar meant with his colorful analogy to an unnatural experiment of prolonging life by creating artificial environments that shelter humans from the mortality risks of natural ecosystems.

If, as argued, human biology has remained largely the same, then hunter-gatherer 'ages' are likely to better reflect our ancestral biology than the same estimates derived from data for modern societies. Human societies sheltered from modernity avoid Medawar's unnatural experiment. Thus, age for maturity, first birth, and menopause are 16, 19, and 45 years, respectively (Piñón, 2002). This scenario also suggests grandparents would have emerged in the late 30s rather than the 50s and 60s observed in developed countries today. Although fertility patterns vary across many dimensions (e.g., geography and time), the vast majority of children ever born are born to mothers younger than 35 (Martinez et al., 2012). This age of cumulative reproduction can be referred to as the effective end of reproduction (EER) as opposed to menopause, the absolute end of reproduction.

Giving birth, however, is the beginning of a journey, not the end of one. Using 35 years as a conservative estimate for the EER, the mother would be 51 years old when her last child reaches the age of independence. Taken together, these milestones imply that the warranty period for humans is approximately 50–55 years (EER + age of independence). This age range may seem too young by today's standards, but it would not seem young 100 000 years ago or even in 1900 when life expectancy in the U.S. was 48.23 and 51.08 years for white males and females, respectively (HMD, 2013).

At What Age Do We Become Old?

Although the age-related physiological consequences of aging are fairly predictable (Shock, 1983), there is significant phenotypic and functional variation among individuals within and between chronological ages (*see* Age Stratification). As such, 'old' is a concept that cannot be uniquely measured or quantified from a biological perspective. Just as there is no specific age when a person is no longer young, there is also no specific age when a person becomes old. Further, identifying a specific chronological age as old is treating a continuous variable as though it is a discrete step-function variable. By general consensus, geriatricians view age 65 as the beginning of the geriatric years and by implication the gateway to old age. Of course, applying labels to segments of a continuum involves creating cut points that by their nature are arbitrary. The question, however, is whether the geriatric classification is reasonable.

The warranty period concept provides context for this question. Old, by definition (Merriam-Webster, 2013), means 'advanced in years or age' as well as 'showing the effects of time or use.' The latter implies that 'old' should be associated with elevated morbidity (frailty, disability) and their mortality risks (Fried et al., 2004). Thus, the chronological ages at which age-specific death rates begin to accelerate should serve as a harbinger of old age.

Parametric survival models provide a way to formally quantify the acceleration of mortality and its implied morbidity. Consider the Gompertz hazard rate (Elandt-Johnson and Johnson, 1980), which can be expressed as $\lambda(t) = \alpha e^{\beta t}$. In the language of physics, the hazard rate is a velocity function. As such, the first derivative of the hazard function measures acceleration and in the case of the Gompertz, it is nothing more than the slope parameter (β) times the hazard rate ($\lambda(t)$). The new function describes the 'acceleration of mortality' that a Gompertz population experiences as it travels along its dying out trajectory on the age axis.

Although demographers typically model all-cause mortality, there are many causes of death that compete for the lives of individuals. William Makeham (1867), an English actuary, recognized this issue and modified the Gompertz function to distinguish accidental deaths (age-independent, extrinsic) from biological deaths (age-dependent, intrinsic). In so doing, he improved the fit of the Gompertz function, created a new field of research now called competing risk theory, and became the father of biodemography. The latter moniker relates to his biologically motivated partition of mortality, which is a methodological characteristic that distinguishes biodemography from demography (Carnes and Olshansky, 1997). His binary cause-of-death partition allowed him to detect mortality dynamics that were otherwise hidden within the all-cause Gompertz pattern of age-specific mortality.

When applied to any mortality data for a developed country, the acceleration of intrinsic mortality will be minimal within the warranty period (Carnes and Witten, 2014). By age 50–55 years, mortality acceleration is approximately four times faster than during the decade of the 40s. However, a fourfold increase of a small acceleration is still a small acceleration. By age 65, mortality acceleration is approximately 12 times higher than what is observed within the warranty period. This level of acceleration is consistent with the geriatric notion of a gateway to old age – mortality risks are elevated, but not yet severe. The acceleration of mortality increases 20-fold by age 70, which would be consistent with 'old.' By age 80, mortality has accelerated by 40- to 50-fold, which would be consistent with Suzman et al. (1992) concept of the 'oldest old.' Although these

labels assigned to segments of a continuous variable may be conceptually useful and mathematically consistent, they must still be viewed as subjective.

There is also an acceleration of extrinsic mortality across the age spectrum, but it only achieves a fourfold increase (relative to within the warranty period) by age 80. In other words, the acceleration of extrinsic mortality is nearly constant. This finding is historically interesting because Makeham himself viewed his partition as distinguishing between age-dependent (intrinsic) and age-independent (extrinsic) causes of death. The minimal acceleration observed in the latter is therefore consistent with Makeham's predictions in 1867.

Can Most of Us Become Centenarians?

This is a frequently discussed issue in gerontology, and schools of thought fall into three general categories. Futurists suggest technology can scientifically engineer negligible senescence (SENS) such that life spans in excess of 1000 years and even immortality are already possible (de Grey, 2002; Kurzweil and Grossman, 2005). Forecasts made by respected demographers have predicted that 50% of every cohort from 2000 onward will survive to age 100 and beyond in developed countries (Christensen et al., 2009). Although biodemographers also forecast continued increases in life expectancies, their more modest longevity predictions (Carnes and Olshansky, 2007) have been described as an example of the pessimistic school of demographic thought (Bongaarts, 2006).

Despite the lack of evidence-based support, the SENS view was included here because this school of thought does have a visible presence within gerontology communities around the world. As mentioned, the demographic view of longevity not only has considerable support, but their methods of analysis and predictions influence agencies like the Social Security Administration when they create hypothetical scenarios in order to assess the impact of demographic stresses on the future solvency of this age-based entitlement program. By way of reminder, the biodemographic perspective is the one used throughout this article (Carnes and Olshansky, 1993).

As in the previous section, the warranty period concept provides a context for examining this extreme longevity question. Recall that ages 50–55 were declared the expiration of the human warranty period, and age 65 was identified as the gateway to old age. From this perspective, the difficulty of maintaining the biological integrity of an organism increases as survival beyond the warranty period increases. There are multiple reasons for this degradation, and they all fall within the rubric that while organisms are not designed to fail, neither are they designed for extended operation. One of the principle reasons for this phenomenon is aging itself, which is defined here as the functional degradation that arises from the "irreparable damage accumulation of biochemical side-reactions" (Yin and Chen, 2005).

An extension of this stochastic view of aging motivates the 'junk theory' of aging (Terman and Brunk, 2006), namely, "normal metabolism is associated with unavoidable mild oxidative stress" that results in "biomolecular damage that cannot be totally repaired or removed by cellular degradative systems." Thus, biochemical side-reactions and oxidative damage create an accumulation of 'irreversibly damaged and functionally defective structures (biological garbage)' that degrade homeodynamic processes and increase the age-specific probability of morbidity and eventual death. On a more evolutionary timescale, this degradation can be viewed as the cost of multicellularity. The mitochondria (cellular organelles), whose increased energy production made multicellular life possible also generate the unintended free radicals that contribute to the accumulating damage initially described in the free radical theory of aging (Harman, 1956).

These biological theories of aging suggest that survival to age 100 and beyond falls far beyond the human warranty period. Demographic evidence supports this conclusion (Carnes et al., 2012). Between 1935 and 2000, the probability of surviving to age 100 (based on U.S. period life tables) increased from 0.22% to 1.8% for females and from 0.09% to 0.49% in males (SSA, 2012). Even in Japan, where female life expectancy at birth is the highest in the world, the probability of surviving to age 100 in the year 2000 was 4.5% (HMD, 2013). These demographic results reveal an important biological reality: the probability of surviving to age 100 is exceedingly low – less than 2% for a female and only one-fourth of that for a male. It is equally important to recognize that biological reality establishes the boundary conditions within which demographic forecasts should or even must fall. Predictions of longevity that fall outside the boundaries of established biological plausibility cannot be justified mathematically; they require an evidence-based biological explanation.

What Makes Extreme Longevity Possible?

Despite poor longevity odds, there must be ways to achieve exceptional longevity, since an estimated 53 000 centenarians and supercentenarians were living in the U.S. in 2010 (Howden and Meyer, 2011). It is probable that every pathway to surviving 100 years requires winning the genetic lottery at birth (Gogele et al., 2011) while also successfully avoiding avoidable (premature) mortality (Perls and Terry, 2003; Willcox et al., 2006). The Forkhead Box O3A gene (FOXO3A) seems to be a genuine longevity gene (Willcox et al., 2008). As age increases, so does the prevalence of FOXO3A – an age pattern of demographic selection that is consistent with a survival advantage for individuals with this genotype.

As with FOXO3A and other longevity genes (Lombard and Miller, 2012) there is no real debate over whether these genes extend survival. In fact, there is good reason for expecting longevity genes to exist because longevity is a prerequisite for achieving the most fundamental task of any organism, reproduction; hence the warranty period and the genes that enable it. When environmental conditions challenge the survival capacity of organisms, the ability to survive one more day or week or month can make the difference between successful reproduction and death before reproduction. As such, the physiological plasticity to survive environmental challenges is an essential component of a broader genetic-based complex of adaptive responses that occur at all levels of biological organization within an organism. In other words, longevity or survival assurance is as much a part of an organism as its anatomy. In fact, biological complexity

theorists would say longevity and adaptation are emergent behaviors (the whole being greater than the sum of parts) of an organism (Carnes et al., 2010).

Living beyond 90 or 100 years, however, is survival that far exceeds expectation. It is an anomaly relative to the survival pattern observed for the vast majority of the population. Raymond Pearl (1946), one of the first biogerontologists, performed a detailed genealogical analysis of human pedigrees to reveal the conditions required for becoming longevous or what we would describe today as achieving extreme longevity. His analysis led him to conclude that in order to become longevous one must "select long-lived parents, and particularly long-lived mothers." However, to be fair to men, more recent analyses suggest that long-lived fathers may be just as important as long-lived mothers in determining the longevity of their offspring (Gavrilov et al., 1998). Of course, we do not pick our parents, but these genealogical findings are consistent with the notion that extreme longevity involves winning a genetic lottery at birth.

Although genetics plays a major role in establishing a warranty period and enabling extreme longevity, the genetic basis of these two phenomena probably differs. A vast majority of people possess the genetic capacity to survive 50–55 years in good health. Surviving 90 years or more in good health is far harder than surviving 50–55 years in good health. Do genes within either or both of these categories influence the biological rate of aging? The jury is out on this question and thus the scientific debates continue (Olshansky et al., 2002).

The disposable soma theory of aging (Kirkwood and Holliday, 1979) is one of the key theoretical foundations of gerontology. It posits that the inevitability of death (mortality pressures) gives reproduction a higher biological priority than maintenance of the soma (body). Hamilton (1966), one of our great evolutionary biologists, implied that the biology of a species is molded by the external mortality pressures acting upon it. This molding can be easily seen at a species level. Mice live short lives and produce large numbers of offspring because nearly everything eats mice. Mortality pressure on humans, however, is much lower, and this enables our species to mature more slowly, reproduce later, and have fewer children. In essence then, there is a trade-off between fertility and longevity. As mentioned, this trade-off is visible at a species level, but what about within species? Do women who become centenarians have fewer children than women who die at typical ages? Research on humans does support a trade-off between fertility and longevity (Westendorp and Kirkwood, 1998; Corbo et al., 2008), but it is also becoming clear that this presumed rule has exceptions (Partridge et al., 2005; Reznick et al., 2000).

Demographic evidence for extreme longevity is dubious for several reasons. The most important reason is that demography studies populations, not individuals. The actuarial rate of aging or the mortality rate doubling time has nothing to do with an individual; they are measures of the speed at which a population dies out. Between 1900 and 2000, life expectancy at birth (sexes combined) in Sweden increased from 58.03 to 79.83 years, a 27.56-year increase (HMD, 2013). Nearly all of that increase can be attributed to saving children from infectious and parasitic disease death and saving mothers from dying during childbirth. In other words, the collective extension of Swedish life spans has little or nothing to do with modifying or quantifying the rate at which Swedes age. This argument is not intended to demean demography; life expectancy is a valuable metric for quantifying population health trends over time. The primary point here is that life spans in the collective can be increased without modifying the biology of humans or the rate or rates of aging.

Closing Thoughts

When viewed through the prism of life on the earth, human biology is just a continuation of an ongoing theme. Out of that mundane status quo came something that could not have been predicted from the past. Virtually every species does something spectacular, but that ability has always had an explainable biological origin. However, no other species in the history of life on the earth has come even close to the ingenuity of hominids and the culture and technology that we have created. Much of that technological innovation has been directed toward progressively isolating ourselves from the mortality risks that have been imposed on us by the environment and microbes (Olshansky et al., 1997). This progressive sheltering from the hazards of ordinary existence is largely responsible for why every generation of humans is setting a new longevity record.

New technologies are being turned inward to reduce the health, disease, and mortality risks arising from bodies built from an imperfect design that uses imperfect building materials and is maintained by imperfect surveillance, maintenance, and repair processes. Despite these design challenges, our technologies have led to unprecedented population growth, unprecedented increases in life expectancy, and unprecedented survival to older ages. Our innovations have produced many blessings, but in doing so have created many new challenges. Fewer of us succumb to microbes, but one in every three seniors now dies with Alzheimer's or other dementia (Alzheimer's Association, 2013) (*see* Alzheimer's Disease, Neural Basis of). The biomedical sciences have found countless ways to manage the symptoms of human diseases, but managing a disease is not the same as curing a disease. Saving people who would otherwise have died has given loved ones more time to share together, but it has also produced a large number of older people struggling with multiple chronic health issues that put them at high risk for frailty and disability. Humans have exploited the natural world to meet our needs, but in so doing have degraded the ecosystems of the earth that are essential to maintain our own health.

Continually increasing life span is not biologically sustainable. The concept of a biological warranty period was used to identify a boundary between normal operation and extended operation. It is also a transition from an expectation of health and vigor to a period where continued life span extension meets resistance from the failures of an imperfect biology that eventually overwhelm even the most effective interventions. This is a new dilemma for our species. Technology has extended life span beyond its normal operating limits. End of life has become blurred, and philosophers and ethicists continue to debate whether death should be allowed to follow its natural course or whether life should be held onto at all costs. These are big questions and they do not have simple answers.

When our intellectual energies are focused on extending life span, we forget the important message that Bernice Neugarten (acclaimed sociologist) and Bob Butler (the first director of the

National Institute on Aging) constantly emphasized, namely, the goal of human technology should be to increase healthy life span even if duration of life remains unaltered. A secondary benefit of increasing healthy life span is that quality of life will also be improved. Finally, in an age when technology can defy both disease and death, we must constantly remind ourselves of the risk that we may be trading off longer life for worsening health.

See also: Age Stratification; Age Structure; Alzheimer's Disease, Neural Basis of; Behavioral Genetics: Psychological Perspectives; Fertility Theory: Theory of Life History Evolution; Lifespan Development, Theory of; Lifespan Development: Evolutionary Perspectives; Motivation: Life Course and Sociological Perspectives; Population Aging: Economic and Social Consequences; Retirement and Encore Adulthood: The New Later Life Course.

Bibliography

Alzheimer's Association, 2013. Report: Alzheimer's Disease Facts and Figures, 72 p.

Bongaarts, J., 2006. How long will we live? Population and Development Review 32 (4), 605–628.

Burger, O., Baudisch, A., Vaupel, J.W., 2012. Human mortality improvement in evolutionary context. Proceedings of the National Academy of Sciences 109 (44), 18210–18214.

Carnes, B.A., Witten, T.M., 2014. How long must humans live? Journal of Gerontology: Biological Sciences 69A (8), 965–970.

Carnes, B.A., Olshansky, S.J., Hayflick, L., 2012. Can human biology allow most of us to become centenarians? Journal of Gerontology: Biological Sciences 68 (2), 136–142.

Carnes, B.A., 2011. What is lifespan regulation and why does it exist? Biogerontology 12, 367–374.

Carnes, B.A., Staats, D.O., Vaughan, M., Witten, T.M., 2010. An organismal view of cellular senescence. Médecine & Longévité 2, 141–150.

Carnes, B.A., Olshansky, S.J., 2007. A realist view of aging, mortality, and future longevity. Population and Development Review 33 (2), 367–381.

Carnes, B.A., Olshansky, S.J., 1997. A biologically motivated partitioning of mortality. Experimental Gerontology 32 (6), 615–631.

Carnes, B.A., Olshansky, S.J., 1993. Evolutionary perspectives on human senescence. Population and Development Review 19 (4), 793–806.

Christensen, K., Doblhammer, G., Rau, R., Vaupel, J.W., 2009. Ageing populations: the challenges ahead. Lancet 374, 1196–1208.

Corbo, R.M., Ulizzi, L., Piombo, L., Scacchi, R., 2008. Study on a possible effect of four longevity candidate genes (ACE, PON1, PPAR-γ, and APOE) on human fertility. Biogerontology 9, 317–323.

Elandt-Johnson, R.C., Johnson, N.L., 1980. Survival Models and Data Analysis. John Wiley & Sons, New York.

Fried, L., et al., 2004. Untangling the concepts of disability, frailty, and comorbidity: implications for improved targeting and care. Journal of Gerontology: Biological Sciences 59, 255–263.

Gavrilov, L.A., Gavrilova, N.S., Semyonova, V.G., 1998. The regularities of inheritance of human life span. Proceedings of the Russian Academy of Sciences 360, 281–283.

Gögele, M., Pattaro, C., Fuchsberger, C., Minelli, C., Pramstaller, P.P., Wjst, M., 2011. Heritability analysis of life span in a semi-isolated population followed across four centuries reveals the presence of pleiotropy between life span and reproduction. Journal of Gerontology: Biological Sciences 66, 26–37.

Gompertz, B., 1825. On the nature of the function expressive of the law of human mortality and on a new mode of determining contingencies. Philosophical Transactions of the Royal Society of London 115, 513–585.

Gurven, M., Kaplan, H., 2007. Longevity among Hunter-Gatherers: a cross-cultural examination. Population and Development Review 33, 321–365.

de Grey, A.D.N.J., Ames, B.N., Andersen, J.K., Bartke, A., Campisi, J., Heward, C.B., McCarter, R.J.M., Stock, G., 2002. Time to talk SENS: critiquing the immutability of human aging. Annals of the New York Academy of Sciences 959, 452–462.

de Grey, A., Rae, M., 2007. Ending Aging: The Rejuvenation Breakthroughs that Could Reverse Human Aging in Our Lifetime, first ed. Martin's Press, New York.

Hamilton, W.D., 1966. The moulding of senescence by natural selection. Journal of Theoretical Biology 12, 12–45.

Harman, D., 1956. Aging: a theory based on free radical and radiation chemistry. The Journals of Gerontology 11, 298–300.

HMD, 2013. The Human Mortality Database. www.mortality.org.

Howden, L.M., Meyer, J.A., 2011. Age and Sex Composition: 2010. 2010 Census Briefs, pp. 1–15.

Kirkwood, T.B.L., Holliday, R., 1979. The evolution of ageing and longevity. Proceedings of the Royal Society of London. Series B 205, 531–546.

Kurzweil, R., Grossman, T., 2005. Fantastic Voyage: Live Long Enough to Live Forever. Plume, 464 p.

Lombard, D.G., Miller, R.A., 2012. Sorting out the sirtuins. Nature. http://dx.doi.org/10.1038/nature10950.

Makeham, W.M., 1867. On the law of mortality. Journal of the Institute of Actuaries 13, 325–358.

Martinez, G., Daniels, K., Chandra, A., 2012. Fertility of men and women aged 15–44 years in the United States: national survey of family growth 2006–2010. National Health Statistics Reports 51, 1–29.

Masoro, E.J., 2005. Overview of caloric restriction and ageing. Mechanisms of Ageing and Development 126, 913–922.

Mattison, J.A., et al., 2012. Impact of caloric restriction on health and survival in rhesus monkeys from the NIA study. Nature. http://dx.doi.org/10.1038/nature11432.

Merriam-Webster Dictionary, 2013. www.merriam-webster.com.

Medawar, P.B., 1952. An Unsolved Problem of Biology. Lewis, London.

NBER, 2013. The National Bureau of Economic Research. www.nber.org/data/multicause.html.

Olshansky, S.J., Hayflick, L., Carnes, B.A., 2002. Position statement on human aging. Journal of Gerontology: Biological Sciences 57A, B292–B297.

Olshansky, S.J., Carnes, B.A., Grahn, D., 1998. Confronting the boundaries of human longevity. American Scientist 86, 52–61.

Olshansky, S.J., Carnes, B.A., Rogers, R.G., Smith, L., 1997. Infectious diseases – new and ancient threats to world health. Population Bulletin 52 (2), 1–47.

Olshansky, S.J., Carnes, B.A., Cassel, C., 1993. The aging of the human species. Scientific American 268 (4), 46–52.

Olshansky, S.J., Rudberg, M.A., Carnes, B.A., Cassel, C.K., Brody, J.A., 1991. Trading off longer life for worsening health: the expansion of morbidity hypothesis. Journal of Aging and Health 3, 194–216.

Partridge, L., Gems, D., Withers, D.J., 2005. Sex and death: what is the connection? Cell 120, 461–472.

Pearl, R., 1946. Man the Animal. Principia Press Inc., Bloomington.

Perls, T., Terry, D., 2003. Understanding the determinants of exceptional longevity. Annals of Internal Medicine 139, 445–449.

Piñón, R., 2002. Biology of Human Reproduction. University of Science Books, Sausalito.

Pritchard, J.K., October 2010. How we are evolving. Scientific American, 41–47.

Rattan, S.I.S., 2000. Biogerontology: the next step. Annals of the New York Academy of Sciences 908, 282–290.

Reznick, D., Nunney, L., Tessier, A., 2000. Big houses, big cars, superfleas and the costs of reproduction. Trends in Ecology & Evolution 15, 421–425.

Shock, N., 1983. Aging of physiological systems. Journal of Chronic Diseases 36 (1), 137–142.

SSA, 2012. Table 6-Period Life Tables for the Social Security Area by Calendar Year. Office of the Actuary. http://www.ssa.gov/OACT/NOTES/as120/LifeTables_Tbl_6.html.

Suzman, R.M., Willis, D.P., Manton, K.G. (Eds.), 1992. The Oldest Old. Oxford University Press, New York.

Terman, A., Brunk, U.T., 2006. Oxidative stress, accumulation of biological 'garbage', and aging. Antioxidants & Redox Signaling 8 (1), 197–204.

Van Arsdale, A.P., 2012. A shifting framework for biological anthropology in 2012. American Anthropologist 115 (2), 262–272.

Vaupel, J.W., 2010. Biodemography of human ageing. Nature 464, 536–542.

Walker, R., Gurven, M., Hill, K., Migliano, A., Chagnon, N., De Sousa, R., Djurovic, G., Hames, R., Hurtado, A.M., Kaplan, H., Kramer, K., Oliver, W.J., Valeggia, C., Yamuchi, T., 2006. Growth rates and life histories in twenty-two small-scale societies. American Journal of Human Biology 18, 295–311.

Westendorp, R.G.J., Kirkwood, T.B.L., 1998. Human longevity at the cost of reproductive success. Nature 396, 743–746.

Willcox, B.J., Donlon, T.A., He, Q.M., Chen, R., Grove, J.S., Yano, K., Masaki, K.H., Willcox, D.C., Rodriguez, B., Curb, J.D., 2008. FOXO3A genotype is strongly associated with human longevity. Proceedings of the National Academy of Science 105, 13987–13992.

Willcox, B.J., He, Q.M., Chen, R., Yano, K., Masaki, K.H., Grove, J.S., Donlon, T.A., Willcox, D.C., Curb, J.D., 2006. Midlife risk factors and healthy survival in men. The Journal of the American Medical Association 296, 2343–2350.

Yin, D., Chen, K., 2005. The essential mechanisms of aging: irreparable damage accumulation of biochemical side-reactions. Experimental Gerontology 40, 455–465.

Old Age: Definitions, Theory, and History of the Concept

Frédéric Balard, Université Lille 3, Villeneuve-d'Ascq, France

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Smith, volume 16, pp. 10843–10847, © 2001, Elsevier Ltd.

Abstract

'Old age' and 'centenarian' refer to a social representation of an individual based on age. The assimilation of the retirement age to the beginning of old age is confusing because all pensioners cannot be assimilated to oldest old. Old age, as the last phase of human life, must be studied more specifically by questioning the factors that contribute to achieving this phase of life, the habits of this age group, and their possibilities of aging well. Moreover, in the framework of a social and behavioral paradigm, studying old age involves thinking of the role of the elderly in society.

The words 'old age' and 'centenarian' refer to a social representation of people based on age. Age is a chronological information (duration) that corresponds to the number of years lived since birth. Based on this definition, an individual could be considered as a centenarian if his or her life duration is equal to 100 years. In contrast, the term 'old' does not refer directly to any given chronological age. Basically, old age corresponds to the life phase lived by the old people. The issue is to define who these old people are, which involves answering the question when does old age begin or when and how an individual could be considered as old.

Before defining old age and what qualifies someone as old, it is questionable whether these concepts are found in all societies. Cross-cultural studies reveal that in some cultures, as for example, Cuiva of Colombia, the concept of old people does not exist. Once people become adults, they keep this status until their death. This is probably why Sokolovsky (2009), in the introduction of the book *The Cultural Context of Aging*, wrote "the conception of being old is *near* human universal".

However, a historical approach indicates that in most societies, human life is conceived of in three phases: childhood, adulthood, and old age. These three phases of life are mainly defined by the ability of individuals to perform the necessary survival tasks for themselves and their group. These tasks are culture specific. Thus, childhood would be the age at which the individual is 'too young' to accomplish the tasks for adults and 'old age', the age at which the individual would be too old. Modernity has substituted the notion of paid work to 'tasks necessary for survival'. Consequently, the official retirement age could be likened to the 'old age' beginning.

When Does Old Age Begin?

Janet Roebuck (1979) asked this question in 1979. Her historical reconstructions conducted mainly in England showed the vagueness of the concept of 'old age'. Under the influence of the Elizabethan Poor Law, it appears that a person is considered old if two conditions are met: reaching the age of 50 or 60 years and being unable to support oneself. In the years 1880–90, the need to match a chronological age to the concept of old age has been raised by the discussion on a state of old age pension scheme. As the combination of age and function would not lead to uniformity, a determination of chronological age was required. Considering the variety of ages selected from other European countries, 65 years was mentioned to be in correspondence with public opinion. However, the debate around the beginning of old age (and the corresponding pension) in the nineteenth century has been subjected to different types of arguments. Considering the age of 70 years as the beginning of old age was a way to save money for pensions. On the opposite side, social arguments were that some professions make people age and die more quickly than others. The age of 70 years was held in 1908, mainly for economic reasons. Subsequently, the pension age was lowered to 65 years for men and 60 years for women without even clearly defining 'old age'. Today, this chronological age is still associated with the beginning of old age in the collective representations. More than 30 years after the work of Roebuck, in France, B. Ennuyer (2011) asks nearly the same question "*A quel âge est-on vieux*". Citing several surveys on the representation of old age by different age groups, the author shows the subjectivity of the concept of 'old age'. For example, people aged between 15 and 25 years considered the beginning of old age around 60 years, while retired think that old age begins at 79 years. The author argues that the definition of old age is still very determined by official statistics and this contributes to a kind of reification.

Increase in Life Expectancy

During the twentieth century, life expectancy increased by 30 years or more in most developed countries. This general trend of longer life goes on with a reduction in mortality at older ages. In these countries, 60-year-old people benefit from more than 25 years of life. This increase in life expectancy contributes to a huge increase of oldest old people. In 2012, there were 17 087 centenarians and 561 646 nonagenarians in France, while these individuals were, respectively, 200 and 35 000 in 1950 (Balard et al., 2011). Old people and oldest old are becoming more numerous and reaching old age is becoming common. Some studies consider that for such countries, most babies born since the year 2000 could become centenarians (Christensen et al., 2009).

This demographic trend is the first upheaval in the way of conceiving the 'old age'. Indeed, the assimilation of the retirement to the beginning of old age, considered as the death-waiting life phase, is no longer valid. In this context of longer life, demographers, epidemiologists, and other biomedical researchers pay close attention to healthy life expectancy and life expectancy without disability (Robine et al., 2013). This research shows that the increase in life expectancy is accompanied by an increase of health expectancy. Despite the fact that this contributes to postponing the years lived in poor health and disabilities later in life, it is not possible to conclude to a generalization of the morbidity compression (Crimmins and Beltrán-Sánchez, 2011).

Fragmentation of the 'Old Age' Life Phase

The survival rate after 60 years, the increase of health expectancy beyond this age, and also the lifestyle disparities of old people have led to a fragmentation of the 'old age' phase into several phases. Thus, Laslett (1991) distinguished the Third Age and Fourth Age, while Neugarten (1974) referred to three categories of old people, the young old, the old old, and the oldest old. It is complex to match fixed chronological ages with these categories considering that the boundaries should fluctuate if we consider the increase of life and health expectancies. However, conventionally, people of the Third Age are those whose age is between 60 and 80 years and those of the Fourth Age are those over 80 years old, even if we find variations of 5 years of these boundaries. It is the same kind of chronological benchmark for the young old (65–75 years), old old (75–85 years), and oldest old (85+) people.

The Third Age and the young old categories are described by researchers as a new kind of old people. They are retired and they do not suffer from any incapacitating disease. These individuals are described as socially active and leisure consumers. Within the family, they are a pivotal generation. They can contribute to help their children financially and take care of their grandchildren and they can also provide a caregiving role with respect to their elderly parents.

Persons belonging to the Fourth Age category represent the 'real' old age because they suffer from comorbidities and activity limitations. From a demographic point of view, Baltes (1997) proposed to consider the Fourth Age as the chronological age where 50% of the birth cohort has died. Thus, the oldest old are part of the category of the Fourth Age and could be defined as those who exceeded the life expectancy of their birth cohort. The old old is an intermediate category between these two broad age categories, even though these categories are not perfectly homogeneous considering that it is possible to be 90 years old without being affected by any activity limitation. In addition, the concept of frailty (Morley et al., 2013) now appears to be crucial to establish distinctions among the elderly population.

From a social and behavioral point of view, it is not appropriate to give a chronological age to the beginning of old age because the social representation of old age depends on the population and the culture considered. The aging process is not only an individual and biological phenomenon but also a social phenomenon that permits to distinguish the oldest old from the other.

From social and behavioral perspectives, it is possible to distinguish three main issues related to old age and centenarians: the means to achieve old age, the old age habits, and the sociocultural meaning of old age.

Achieving Old Age

Despite the increase of lifespan and the oldest old population, reaching old age and becoming centenarian remains a mark of distinction. Consequently, one of the major research axes is the identification and the analysis of the longevity factors. All disciplines interested in human lifespan explored what allowed some people to reach old age with their own models and methods. Thus, the hypothesis of a genetic heritability of longevity has been largely explored as in the context of twin studies (Herskind et al., 1996). Given the relatively small proportion of the genetic heritability of longevity, these types of research are mainly focused on the biological factors that contribute to healthy aging (Skytthe et al., 2011). Social and behavioral sciences also present interesting results in this research field. Psychological researches show particularly strong links between personality traits and longevity (Chapman et al., 2011). These studies show such good evidence that higher level of conscientiousness and lower levels of hostility and 'distressed' personality are associated with greater longevity. Researches in public health, epidemiology, sociology, and anthropology have focused on the benefit of the sociocultural influence. Several studies have demonstrated the strong influence of the social status on life expectancy (Marmot, 2004).

Some researchers have attempted to analyze the lifestyles of populations in which the number of centenarians appears significantly higher than the national average, as in Okinawa and Sardinia. This allowed them to highlight, among other things, the influence of a low-calorie diet (Willcox et al., 2006) and physical activity (Pes et al., 2011) on the duration of human life.

Finally, the health gender paradox has to be mentioned as a key issue in this field. Even if women live longer and are more numerous in old age, men are in better physical and cognitive health, and are higher functioning. This observation implies to ask: if better health should lead to a longer life, why are men not living longer than women? Biologic research tries to answer this by focusing on the physiological differences between male and female, while social and behavioral research are mainly focused on the gender differences that could explain the stronger selection effect for men than women (Balard et al., 2011)

Aging Well

Research on longevity and aging are inseparable from the question on aging well, as evidenced by the myths or the first philosophical reflections such as those of Plato and Aristotle on the idea of positive aging for the first one and negative for the

second one. The fundamental question underlying this debate concerns the quality of life in old age, or to put it more directly if the years lived in old age are worth to be lived. Gerontology research debated on the aging well topic by developing concepts such as successful aging, active aging, productive aging, etc.

Despite the variety of definitions, the assessment of successful aging or quality of life in old age appears to be done mainly through the concepts of health, functional autonomy, and social productivity. Fernández-Ballesteros et al. (2013) indicate "Health maintenance and ADL preservation; high physical and cognitive functioning; positive affect and control, and social participation and engagement seem to be the most accepted domains." This aging well conception corresponds to the theoretical model of activity theory (Havighurst, 1961), which could be summarized in the preservation of the middle age lifestyles. The work of Rowe and Kahn (1987) is certainly the one that has most influenced this model by comparing usual aging to successful aging.

Some authors such as Tornstam (1992) criticized this paradigm based on productivity, efficiency, and independence as a kind of scientific realization of the middle-aged white Western men's values.

In contrast to the activity theory, the disengagement theory (Cumming and Henri, 1961) is mainly a psychological research paradigm. This model suggested that gradual disengagement from social life and focus on the inner self was a naturalistic process of human aging, because physical and mental decline is inevitable. Based on this paradigm, psychologists have developed several theories of successful aging. The Selection, Optimization, Compensation theory of Baltes (1997) has become a major model for this research field. It assumes that old people actively cope with losses by using selection (choosing their life goals), optimization (sophistication of ability, allocation, or adjustment of resources and regulation of motivation to accomplish goals), and compensation (drawing on a compensative resource, such as assistive devices or new strategies). The socioemotional selectivity theory (Carstensen, 1991) hypothesizes that individuals of any age focus on emotional euthymia rather than stressful situations, such as seeking new information or knowledge, when they realize that their life is limited. The gerotranscendence hypothesizes that gradual and natural changes in behavior and thought that come with aging provide an adaptive mechanism through which the elderly can maintain psychological well-being (Gondo et al., 2013).

Disengagement theory has also had some influence on the field of sociology of aging and the development of the concept of '*déprise*', which focuses on the arrangements and strategies made by old people when they are confronted with social and physiological trials (*épreuves*) (Caradec, 2007).

Old Age Habits

The analysis of the old age experience is not limited to the issue of aging well. It also aims to understand old people as different from younger people. In this context, particular attention should be paid to distinguish the biological age effects and the generation effects on life.

The aging process, whether we conceive of it in its biological and/or social dimension, is not sufficient to explain the specific behavior of the oldest old. In order to cope with the inherent fragility of senescence and the specific social events of this phase of life (isolation, loss of empowerment, etc.), the oldest old mobilize resources and tools that are often indicative of their generational affiliation. Generation constitutes a fairly narrow circle of individuals who, despite the diversity of other factors that have to be taken into account, are connected in a homogeneous whole in that they rely on the same main events and changes during their receptivity. In this context, all the research linked to the material and social environment of the oldest old must take into account the generational influence. Thus, the attitudes that they have with their health, their place of life, the technological equipment, the professional care as well as family relationships, and many other elements are strongly influenced by generational membership. This implies that precise attention should be paid to the particularities of the cohort studied and people's life course. A corollary of this is the almost permanent question on the results and theories developed regardless of the line of research considered. Indeed, it is likely that the baby boomers, i.e., the cohorts born after World War II, will report profoundly different behavior than the present-day centenarians.

The Sociocultural Meaning of Old Age

As indicated by Carstensen and Fried (2012), "population ageing presents a cultural problem". The increasing figures of the oldest old constitute an economic and health issue that is exacerbated in the present context of economic and social crisis in many societies. In addition, the changes in intergenerational relationships when they take the form of a decline in family support may lead some old people to a huge precariousness (Zhang, 2009). However, the social challenge related to old age is not limited to its economic and health dimensions. Indeed, it implies to question the meaning that our society is willing to pay to old age. In a social and behavioral perspective, one of the fundamental questions asked by old age is the social role of the oldest old people in our society. However, the place and the social role of oldest old is undermined by the movement of society medicalization that contributes to assimilate aging and senescence, making the old age inconsistent with the dominant values in modern societies such as productivity. Similarly, grouping presenile dementia and senile dementia under the term 'Alzheimer's disease', leads to a massive increase of 'illderly' (Ballenger et al., 2009), which contributes to the construction of a negative image of old age. Changing the perspective that we have of the oldest old now appears as a crucial issue as well as leads to research to improve their health and functioning. In this view, it seems necessary to remember that in some cultures, old age is not synonymous with senescence and obsolescence, but is rather a privileged status inherent to the experience and knowledge accumulated.

See also: Age: Anthropological Aspects; Aging and Health in Old Age; Differential Aging; Mortality, Biodemography of; Twin Studies in Demography.

Bibliography

Balard, F., Beluche, I., Romieu, I., Willcox, D.C., Robine, J.M., 2011. Are men aging as oaks and women as reeds? A behavioral hypothesis to explain the gender paradox of French centenarians. Journal of Aging Research. http://dx.doi.org/10.4061/2011/371039. http://www.hindawi.com/journals/jar/2011/371039/ (accessed 25.03.13.).

Ballenger, J.F., Whitehouse, P.J., Lyketsos, C.G., 2009. Treating Dementia: Do We Have a Pill for It? Johns Hopkins University Press, Baltimore, MD.

Baltes, P.B., 1997. On the incomplete architecture of human ontogeny: selection, optimization, and compensation as foundation of developmental theory. American Psychologist 52, 2366–2380.

Caradec, V., 2007. L'épreuve du grand âge. Retraite et société 3, 11–37.

Carstensen, L.L., 1991. Socioemotional selectivity theory: social activity in life-span context. Annual Review of Gerontology and Geriatrics 1, 195–217.

Carstensen, L.L., Fried, L.P., 2012. The meaning of old age. Global Population Ageing: Peril or Promise? In: Working Paper Series, vol. 15. http://diseaseriskindex.harvard.edu/pgda/WorkingPapers/2012/PGDA_WP_89.pdf#page=18 (accessed 25.03.13.).

Chapman, B.P., Roberts, B., Duberstein, P., 2011. Personality and longevity: knowns, unknowns, and implications for public health and personalized medicine. Journal of Aging Research. http://dx.doi.org/10.4061/2011/759170. http://www.hindawi.com/journals/jar/2011/759170/ (accessed 25.03.13.).

Christensen, K., Doblhammer, G., Rau, R., Vaupel, J.W., 2009. Ageing populations: the challenges ahead. The Lancet 374 (9696), 1196–1208.

Crimmins, E.M., Beltrán-Sánchez, H., 2011. Mortality and morbidity trends: is there compression of morbidity? The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 66 (1), 75–86.

Crimmins, E.M., Saito, Y., Ingegneri, D., 1989. Changes in life expectancy and disability-free life expectancy in the United States. Population and Development Review 15 (2), 235–267.

Cumming, E., Henri, W.E., 1961. Growing Old: The Process of Disengagement. Basic Books, New York.

Ennuyer, B., 2011. A quel âge est-on vieux ? La catégorisation des âges: ségrégation sociale et réification des individus. Gérontologie et société 138, 127–142.

Fernández-Ballesteros, R., Molina, M.A., Schettini, R., Santacreu, M., 2013. The semantic network of aging well. Annual Review of Gerontology and Geriatrics 33 (1), 79–107.

Gondo, Y., Nakagawa, T., Masui, Y., 2013. A new concept of successful aging in the oldest old development of gerotranscendence and its influence on the psychological well-being. Annual Review of Gerontology and Geriatrics 33 (1), 109–132.

Havighurst, R., 1961. Successful aging. The Gerontologist 1, 8–13.

Herskind, A.M., McGue, M., Holm, N.V., Sörensen, T.I., Harvald, B., Vaupel, J.W., 1996. The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870–1900. Human Genetics 97 (3), 319–323.

Laslett, P., 1991. Fresh Map of Life: The Emergence of the Third Age. Harvard University Press, Cambridge, MA.

Marmot, M., 2004. The Status Syndrome. Times Books, New York.

Morley, J.E., Vellas, B., Abellan van Kan, G., Anker, S.D., Bauer, J.M., Bernabei, R., Walston, J., 2013. Frailty consensus: a call to action. Journal of the American Medical Directors Association 14 (6), 392–397.

Neugarten, B.L., 1974. Age groups in American society and the rise of the young-old. Annals of the American Academy of Politics and Social Sciences 415, 187–198.

Pes, G.M., Tolu, F., Poulain, M., Errigo, A., Masala, S., Pietrobelli, A., Maioli, M., 2011. Lifestyle and nutrition related to male longevity in Sardinia: an ecological study. Nutrition, Metabolism and Cardiovascular Diseases 23 (3), 212–219.

Robine, J.M., Cambois, E., Nusselder, W., Jeune, B., Van Oyen, H., Jagger, C., 2013. The joint action on healthy life years (JA: EHLEIS). Archives of Public Health 71 (1), 2.

Roebuck, J., 1979. When does old age begin?: the evolution of the english definition. Journal of Social History 12 (3), 416–428.

Rowe, J., Kahn, R., 1987. Human aging: usual and successful. Science 237 (4811), 143–149.

Skytthe, A., Valensin, S., Jeune, B., Cevenini, E., Balard, F., Beekman, M., Slagboom, P.E., 2011. Design, recruitment, logistics, and data management of the GEHA (Genetics of Healthy Ageing) project. Experimental Gerontology 46 (11), 934–945.

Sokolovsky, J., 2009. A global vision of aging, cultural and context. In: Sokolovsky, J. (Ed.), The Cultural Context of Aging, Worldwide Perspective. Praeger, London, pp. 3–6.

Tornstam, L., 1992. The quo vadis of gerontology: on the scientific paradigm of gerontology. The Gerontologist 32 (3), 318–326.

Willcox, D.C., Willcox, B.J., Todoriki, H., Curb, J.D., Suzuki, M., 2006. Caloric restriction and human longevity: what can we learn from the Okinawans? Biogerontology 7 (3), 173–177.

Zhang, H., 2009. The new realities of aging in contemporary China: coping with the decline in family care. In: Sokolovsky, J. (Ed.), The Cultural Context of Aging, Worldwide Perspective. Praeger, London, pp. 196–207.

Old Age Work Participation

Arthur van Soest, Tilburg University, Tilburg, The Netherlands
Asghar Zaidi, University of Southampton, Southampton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

To understand how to raise labor force participation of older workers, it is necessary to understand demand, supply, and institutional factors. Labor supply and retirement decisions of older workers are at the core of microeconomic research and many countries have recently reformed their pension systems to make early retirement financially less attractive. Noneconomic factors like quality of work, health status, social networks, and peer group behavior are important as well. The demand side of the labor market is equally important, particularly now that recent pension reforms have raised labor supply.

Key questions are: How to maintain the productivity of older workers? How to eliminate existing prejudice toward older workers among employers and coworkers? How effective are laws against age discrimination or public campaigns to promote the image of older workers? Which demand side adjustments will be necessary to accommodate an increased supply of older workers, concerning, e.g., hiring policies, reducing wage costs, training of older workers, using alternative exit routes, accommodating workers with a health problem and facilitating gradual retirement?

In this article, we discuss the labor force participation of older age groups. We discuss definition and measurement of old age and work participation and present some data on OECD countries, provide an overview of the determinants of retirement and labor supply of older age groups, and summarize what we know about labor demand and productivity. We also discuss policy issues that may guide future research.

Introduction and Background

The increase in life expectancy in recent years poses new challenges for policymakers across the world. Policymakers faced with unprecedented rises in social expenditures on those in retirement and the need to ensure fiscal sustainability of social welfare and pension systems, have endeavored to extend working careers by attaining higher employment rates among older workers. In particular, many countries of the European Union face the challenge of reversing the trend toward early retirement by providing financial incentives to make people remain in work longer. This policy agenda requires a good understanding of demand and supply factors underlying the labor force participation of older age groups.

Since 1999, active aging has featured often in many of the EU policy frameworks. The goal of active aging has often been narrowed to extending working lives and discouraging early retirement. The focus on enhancing labor market participation of older workers is reflected in two EU-level targets: the 2001 Stockholm target to ensure that half of those in the age group of 55–64 years were in employment by 2010, and the 2002 Barcelona target to increase the average age of exit from the labor market (for retirement) in 2010 by 5 years. While none of the EU countries has managed to achieve the Barcelona target, Germany, Ireland, Cyprus, the Netherlands, and Finland did raise the employment rate for the age group of 55–64 years to over the 50% threshold by 2010 (Zaidi and Zolyomi, 2012). (In Sweden, Denmark, Portugal, and the UK the employment rate of this age group was already above 50% in 2001.)

Under the European Union's strategy for smart, sustainable, and inclusive growth, known as 'Europe 2020,' one of the five headline targets that EU Member States have agreed to implement to increase the employment rate of 20- to 64-year-old individuals to 75% by 2020. This target can only be achieved by increasing the employment rate of older workers and raising the effective retirement age. Another target is to lift at least 20 million people out of poverty and social exclusion in the next decade. This involves ensuring adequate pensions for older people, for which extended working careers are essential. In line with Europe 2020, the European Employment Strategy also promotes policies and measures targeted at supporting longer working lives. The Europe 2020 New Skills for New Jobs flagship initiative emphasizes better training and skills through occupational training and lifelong learning opportunities, among others for older workers (Zaidi, 2012).

The idea of the Silver Economy offers a solution to the current economic problems of lack of growth in the European economies exploiting the potential of a rising share of older people in society. It refers to the fact that the rising share of older consumers represents a potential for additional demand for new types of products and services, such as personalized care and technology that enables people to maintain healthy and independent lives as they age. Furthermore, the new cohorts of retirees are also expected to be wealthier than their predecessors, and thus generate more specific demand in the economy. The term 'Silver Economy' also sometimes encompasses the rising segment of older workers who ought to be considered as a resource for their employment potential and other social contributions (such as transfer of skills toward younger workers) in bringing economic growth.

To understand how to raise labor force participation of older workers, it is necessary to understand demand, supply, and institutional factors. Labor supply and retirement decisions of older workers are at the core of microeconomic research and many countries have recently reformed their pension systems to make early retirement financially less attractive. Noneconomic factors like quality of work, health status, social networks, and peer group behavior are important as well. Health plays a major role, in particular through limitations at work due to ill health and disability. The demand side of the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.94040-6

labor market is equally important, particularly, now that recent pension reforms have raised labor supply.

Key questions are: How to maintain the productivity of older workers? How to eliminate existing prejudice toward older workers among employers and coworkers? How effective are laws against age discrimination or public campaigns to promote the image of older workers? Which demand side adjustments will be necessary to accommodate an increased supply of older workers, concerning, e.g., hiring policies, reducing wage costs, training of older workers, using alternative exit routes, accommodating workers with a health problem and facilitating gradual retirement?

In this article we discuss the labor force participation of older age groups. First we discuss definition and measurement of old age and work participation and present some data on OECD countries. Then we provide an overview of the determinants of retirement and labor supply of older age groups, followed by a summary of what we know about labor demand and productivity. Finally, we discuss policy issues that may guide future research.

Old Age

At least three definitions of old age are in use in the social gerontology and economics of aging literature:

- Subjective definition – based on each individual's own assessment, usually derived from their labor market activity or health status.
- Functional definition – using objective information on how people function, considering restrictions due to health problems and frailty in different domains of life (e.g., in being employed, in physical and mental functioning in their day-to-day life).
- Chronological definition – using an objective age threshold beyond which the old age phase of life begins (e.g., the statutory pension age).

These definitions can provide an approximate entry into the old-age period of life around the time of retirement from the labor market or deterioration in health status. The chronological age cut-off can also be defined on another basis. For instance, the employment rate of older workers by EUROSTAT (the Statistical Office of the European Union) is calculated using the age group of 55–64 years. This age group is the primary focus of analysis in this article and in most of the literature on labor force participation of older age groups.

In addition, there is an increasing trend to work beyond the statutory retirement age (often age 65). Such workers are often referred to as 'silver workers.' Silver workers are increasingly visible in every sphere. In the UK, for instance, the most obvious examples include Queen Elizabeth (age 86) Sir Alex Ferguson (manager of Manchester United until age 71), and the actress Dame Judi Dench who competed for the Golden Globes at age 78. Employment after the statutory age is of special interest as it may provide a necessary complement to income and thus avoid poverty in old age. The financial aspect is not the only factor, however, as higher income workers are often overrepresented among post-retirement workers (Giarini, 2009). A formal job after retirement also serves as a possible means to remain active in other social and civic engagements.

One advantage of the chronological age definition is that age is by definition an exogenous attribute, whereas indicators of old age on the basis of labor market or health status are often affected by individual choices. An adoption of the definition on the basis of the chronological age will allow us to capture diversity within the old-age population in terms of their labor market participation (i.e., the endogenous variables). Moreover, the chronological definition is free from negative biases that arise from common myths and misconceptions about aging and older people, such as the perception that they are often frail, ill or disabled, and may thus be a burden on the community.

Work Participation

Individuals can be classified in three categories: employed, unemployed, or economically inactive. The most common definitions for these follow the Resolution of the 1982 International Conference of Labor Statisticians known as the 'ILO guidelines' (ILO, 1982). Two important elements are the selection of an age group (the reference group), and the time unit of the assessment (the reference week):

- Employed persons comprise those aged 15 years or above who either (1) worked for pay, profit, or family gain for at least 1 h during the reference week or (2) were not at work during the reference week but were temporarily absent from their job or business. This definition is applicable to employees, self-employed persons, and family workers. Pay includes cash payments or/and payments in kind (that is, in goods and services).
- Unemployed persons comprise those of ages 15–74 years who are: (1) not employed according to the definition mentioned above, (2) currently available for work (i.e., for paid employment or self-employment before the end of the 2 weeks following the reference week), and (3) actively seeking work (i.e., had taken specific steps in the 4 weeks ending with the reference week to seek paid or self-employment).

In some countries, the reference group is defined slightly differently. For instance, in Spain and Great Britain, the reference age group corresponds to those 16 years and more, while the upper age limit for the reference group is restricted to 74 years in Denmark, Estonia, Hungary, Latvia, Finland, Sweden, and Norway.

The residual third group is commonly referred to as 'inactive.' For older people, however, the distinction between unemployed and inactive is often blurred (Marin, 2013). For example, 'discouraged workers' who lost interest in searching for a job are categorized as 'inactive' rather than 'unemployed.' The most useful distinction for the analysis in this article is, therefore, employment versus nonemployment.

Employment Rates of the Age Group 50–64 Years

Figure 1 (males) and 2 (females) show large differences in the employment rates of workers aged 50–64 years across countries

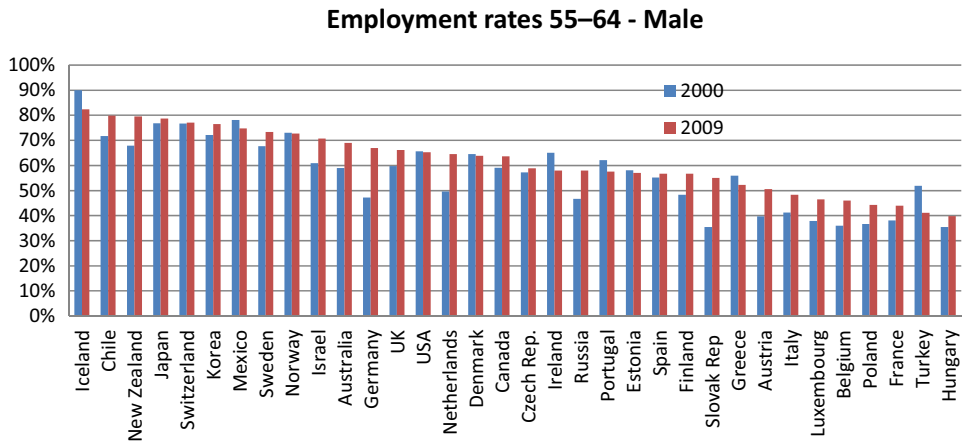

Figure 1 Employment rate for males, in the age group 55–64 years, during 2000 and 2009.

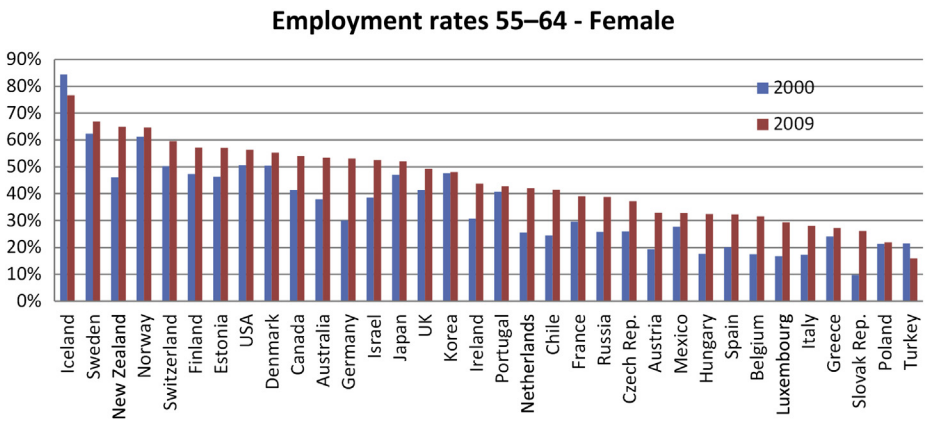

Figure 2 Employment rate for females, in the age group 55–64 years, during 2000 and 2009.

in 2000 and 2009. For males in 2009, some countries had employment rates of older workers close to 80%, whereas in other countries it was less than half as large (Hungary, Turkey, France, and Poland). Still, most countries improved in comparison to 2000. This positive trend seems to have stopped since 2009 (Zaidi and Zolyomi, 2012, Figure 6) as an effect of the global economic crisis since the fall of 2008. The trends for an earlier period (1998-2005) also show rising trends of employment among older workers (see Zaidi et al., 2007 for more discussion).

For older females, Iceland, Sweden, New Zealand, and Norway show employment rates in excess of 60%, and Switzerland, Finland, Estonia, and the United States lag only slightly behind (around 60%). At the other end of the spectrum, Turkey (16%) and Poland (22%) performed worst in 2009. In comparison to the trends for males, the improvement in the employment rate for female older workers is larger and observed in more countries. Still, the employment rate of older females remains quite low in many countries.

The Netherlands, Germany, and Finland are examples of countries where the employment rates among older workers have risen considerably because of specific policy measures. In the Netherlands, three public policy drivers are viewed to be responsible for these trends (Euwals et al., 2009): (1) Elimination of financial disincentives to delay retirement and make the pension system actuarially fair; (2) Measures to keep (older) people with reduced work capacity in the labor market instead of granting them permanent invalidity benefits; and (3) Other relevant policy measures such as stricter job search requirement for unemployed older workers.

In Germany, two partial retirement schemes (the partial retirement (Altersteilzeit) and the partial pension (Teilrente)) aim to provide a gradual transition from working life to retirement. Employees can claim a partial retirement if they have reached age 55 and wish to reduce their working time to half of the collectively agreed hours and if the employer agrees. Until 2010, the employer could receive subsidies to cover the costs if the part-time job is filled with a previously unemployed person. The partial pension can be 33, 60, or 67% of the full pension, but take-up of this pension is much lower than that of partial retirement.

In Finland, the 2005 reform of the earnings-related pension system introduced flexibility regarding the time of retirement between the ages 63 and 68 years. Within this age range, incentives to delay in retirement are given by higher accrual rates. From age 18 to 63 years a pension right of 1.5–1.9% of the pensionable wage is acquired each year. From age 63 until age 68, it is possible to receive an old age pension, but the

accumulation rate increases to 4.5% annually, creating a strong incentive to remain at work. If the old age pension is further deferred after age of 68 years then the amount is increased by 0.4% per month. These incentives had a strong impact, increasing the effective age of retirement.

Figure 3 (males) and 4 (females) show how the ratio of employment rates of older (aged 55–64 years) and prime-age (aged 35–44 years) workers have changed between 2000 and 2009. They show that in most countries the relative position of older workers compared to younger workers (both males and females) has improved. An exception is Turkey, where the employment situation among older female workers has deteriorated between 2000 and 2009. For Turkish males, there is also a decline in the relative position of older workers, but much less stark than for females.

The reforms across European countries are characterized by two broad types of policy measures:

1. Promotion of employment of older people, through financial incentives for work in the pension and old-age social security systems and through phasing out early retirement schemes, and by raising retirement ages or abolishing the mandatory retirement age; and
2. Improvement of conditions for the recruitment, retention, and productivity of older workers. This involves age management policies at the company level; public measures for the improvement and modification of qualifications and competences; reductions in labor costs of employment of older workers, and active labor market programs.

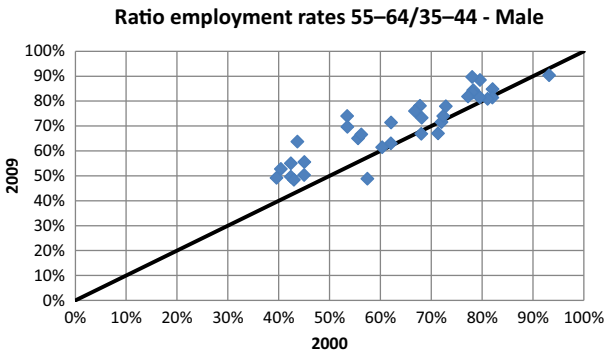

Figure 3 Ratio of employment rates for older workers (55–64) and younger workers (35–44), for males, during 2000 and 2009.

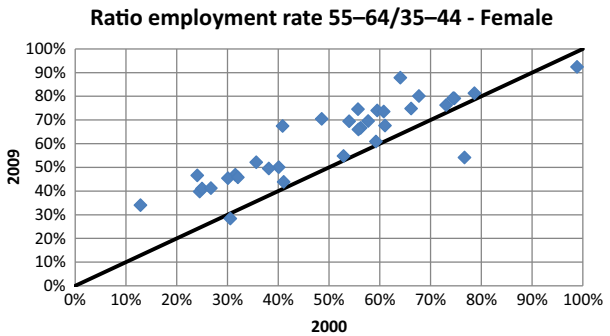

Figure 4 Ratio of employment rates for older workers (55–64) and younger workers (35–44), for females, during 2000 and 2009.

Many countries have successfully implemented the first type of policy measures during the late 1990s and early 2000s, and have now embarked on complementing them with the second type (European Commission, 2012). A good example is Poland, which after successful pension reforms, has now prepared a comprehensive program ('Solidarity across generations. Measures aiming at increasing the economic activity of people over 50') to achieve the Lisbon employment rate target of 50% for the age group of 55–64 years by 2020.

Labor Supply of Older Workers

At least three supply explanations for the large variation in retirement patterns across countries have been given (Kapteyn and Andreyeva, 2008). The first is financial incentives, emphasized by a group of researchers performing country-specific studies using a harmonized methodology (Gruber and Wise, 2004), evaluating the link between old-age social security programs and national retirement patterns. These studies consider the substitution effect as the dominant factor in shaping individual choices in the income–leisure trade-off: a higher reward for working longer raises the price of leisure around the retirement age, and therefore reduces the demand for leisure and increases labor supply. They emphasize that in many systems, retirement benefits did not increase with delaying retirement, so that the implicit tax on continued work was often 100%.

To illustrate: Gruber and Wise (2004) summarize their results of a hypothetical pension reforms in 12 countries, based upon a common model. The reform (named *Three-Year Delay*) shifts all pension entitlements by 3 years (if the actual entitlement at age X is Y, the simulation assigns entitlement Y to age X+3). They compute by how much the percentage out of labor force changes for a country-specific age group around retirement (the first age at which at least 25% of the workforce is out of the labor force plus the next 4 years). The results show that retirement behavior is sensitive to financial incentives: in 11 of the 12 countries, nonparticipation in the age group considered falls by more than 40% if pension entitlements are shifted by 3 years. Many other empirical studies have confirmed the sensitivity of retirement for financial incentives, with empirical strategies varying from reduced form difference-in-differences approaches to sophisticated structural dynamic programming models (e.g., French and Jones, 2011).

Another explanations for cross-country differences and particularly the gap between the United States and many European countries are based upon the larger extent to which home production can be substituted with services available in the labor market (Freeman and Schettkat, 2001), differences in personal preferences or culture (Blanchard, 2004), or the power of unions (Alesina et al., 2006).

A discussion of retirement patterns also has to consider the role of health. Failing health may lead to retirement (Kalwij

and Vermeulen, 2008). Retirement status may also have a (positive or negative) influence on health: it can remove mental stress and physical work effort, but may also be a life event that creates stress and reduces health. Existing studies often find a negative effect of retirement on, in particular, cognitive functioning, as a result of living less stimulating lives in retirement compared to working lives (Rohwedder and Willis 2010; Bonsang et al., 2012). An economic framework for studying the relation between health and retirement is the Grossman model (Grossman, 1972) in which individuals derive utility from consumption and health, and health also influences earnings. Galama et al. (2013) introduce variants of the basic Grossman model that include a retirement decision. Here individuals with lower human capital have fewer resources to invest in health so that their health deteriorates faster, implying that health will be positively associated with income and education. A higher earnings capacity also induces people to work longer, as observed in many studies. Yet, at the same time improvements in health over time have an income effect that reduces work effort and hence explains trends in early retirement.

The Demand for Older Workers: Productivity, Wages, and Substitution

Of the models that predict the demand for older workers from the firm's perspective, the best known is Lazear's 'delayed compensation contracts' model (Lazear, 1979). Lazear argues that a work contract where the worker is paid less than the value of the marginal product at younger ages, and more at later ages, has advantages for both workers and employers. For workers, this type of contract will increase lifetime wealth. Employers, although forced to bear the higher fixed costs associated with delayed compensation, gain from improvements in performance and stronger employee commitment induced by the workers' fear of losing delayed compensation. Mandatory retirement is required as a way of terminating the contractual work, as the worker would not voluntarily retire due to the high wage. In this model, firms avoid hiring older workers, as this will reduce the possible benefits of delayed compensation.

Skirbekk (2003) summarizes earlier studies relating age to productivity, finding evidence of a decline in several aspects of physical and mental functioning from around age of 50 years. On the other hand, other studies argue that older workers often rely on their professional experience to adapt and compensate for the decline in physical and mental ability and find no significant difference between the job performance of older and younger workers. Recent studies using matched employer–employee data suggest that individual productivity declines with age in some dimensions, but the decline can be partly compensated by experience, personal aids, and suitable workplace adjustments (Crépon and Aubert, 2003). In a study for Belgium, Lallemand and Rycx (2009) find that, particularly in ICT firms, young workers are significantly more productive than older workers, although age effects on productivity have substantially decreased over time. Empirically, higher wages for 55- to 64-year-old individuals relative to 25- to 29-year-old individuals are negatively correlated with activity rates at the age of 55–64 in the case of both men and women (OECD, 2006), in line with the firm's disincentive to employ older workers if productivity rises less with age than wage costs.

In view of the current context of high youth unemployment in many European countries, it is important to dispel the myth that a longer working career would imply less employment opportunities for younger people. The idea that the old should make place for the young is based on the incorrect notion that the total number of jobs is fixed (the 'lump-of-labor' fallacy). The same argument was used in the past to discourage women from joining the workforce and to motivate anti-immigrant policies. The lump-of-labor fallacy seems to pop up over and over again in the policy debate and becomes a serious concern for the intergenerational conflict. Evidence against it comes from the stylized facts that the high-employment rate among older workers (55–64) is typically positively associated with the high employment among youth (15–24). If the lump-of-labor arguments were correct, the opposite would be expected. Readers are referred to Kapteyn et al. (2004) for more rigorous empirical research supporting this line of argument.

The skills offered by older people are very often distinct and cannot always be substituted by those of younger workers. For instance, patients tend to prefer older physicians and older retail assistants often are more knowledgeable and patient. Earlier retirement, therefore, will probably not generate more jobs for young entrants. Instead, using the full potential of older workers and reaping the benefits of the demographic shifts in the demand for products and services will create more jobs both for the old and the young (Zaidi, 2012).

Employer Attitudes, Age Discrimination and Regulation

A special Eurobarometer survey (European Commission, 2007) shows that almost half of the population feels that age is, together with the onset of a disability, one of the most important criteria that might put a job candidate at a disadvantage when competing against someone with the same qualifications: 78% of respondents feel that a person aged 50-plus is less likely than a younger person to get a job, accepted for training, or promoted.

In a survey of 500 large employers in the UK, Taylor and Walker (1994) show that many had negative stereotypes of older workers, especially with regard to their openness to training and ability to adapt to new technologies. Laboratory and field experiments confirm that employers discriminate against older applicants (Riach and Rich, 2002). More recent studies show mixed results. For example, McNair et al. (2007) found that age stereotypes and attitudes tend to favor older workers in general (emphasizing more skills, life experience, reliability, loyalty, etc.) though negative views also exist (e.g., lack of willingness to adopt new work methods). Ilmarinen (2006) emphasizes the need to change toward a work culture that is better suited for workers of all ages, involving employers and employees as well as society. Van Dalen et al. (2009) analyze employer attitudes and actions in Greece, Spain, the Netherlands, and the UK and find that only in the UK, older workers are recognized as a valuable source of labor

supply. Vickerstaff (2010) summarizes the literature on this topic and concludes that most organizations have no specific strategies for age management and that the business case for a well diversified age composition of the workforce is extremely fragile, particularly during recession. Karpinska et al. (2013) find that Dutch managers generally do not find it important to keep older workers until the mandatory retirement age.

Cheron et al. (2011) extend the standard model of equilibrium unemployment with an explicit role of age. They argue that a standard retirement age explains the lower employment rate among the older population: it reduces the willingness of firms to recruit older workers, as expected returns will be lower. Moreover, the shorter expected job duration reduces the tendency of older workers to invest in job search. They conclude that existing policies in many countries present strong perverse effects on employment and social welfare.

There are several empirical studies on the impact of employment protection legislation. Bassanini and Duval (2006), using data from 21 OECD countries for 1982–2003, find a positive relation between employment protection legislation and the employment rate of workers aged 55 to 64. OECD (2004) finds that the negative relation between employment protection and the hiring rate of men aged 50 and over is compensated by a decrease in firings.

On-The-Job Training

Jacobs (2010) shows how to systematically analyze the interactions between human capital investments in on-the-job training (OJT), retirement choices and pension saving. The model adds an endogenous retirement decision to the standard life-cycle model of OJT investment and human capital formation. By extending the time-horizon over which investments in skills materialize, a higher retirement age promotes investments in OJT. Later retirement and OJT investment are therefore complementary and generous early retirement schemes therefore indirectly discourage investment in human capital. The intuition is that the opportunity return at which future labor earnings are discounted increases with the retirement age and human and financial capital are substitutes over the life cycle. This analysis implies that promoting lifelong learning or later retirement will not be effective as long as strong disincentives caused by labor market institutions, early retirement schemes and incentives for pension savings remain in place. Moreover, promoting private pension savings may inadvertently create implicit taxes on skill formation and indirectly stimulate early retirement, thereby worsening the aging problems.

Empirical work analyzing OJT of older workers is scarce. A major empirical problem is that investment in OJT is very difficult to measure. According to Heckman (2000) most training is informal rather than formal, limiting the applicability of commonly employed training measures, which are often based on subjective data (from firms or employees), on formal OJT investment. Moreover, firms and employees have different views on the participation intensity of training (Leuven, 2005). Not only the costs (i.e., the OJT investment), but also the returns (future wages) are difficult to measure, because earnings are not equal to labor productivity even if labor markets are perfectly competitive (since time investment in OJT drives a wedge between gross labor productivity and gross earnings).

Policy-Related Research Questions

Some major policy relevant research questions concerning old-age work participation in OECD countries are the following:

Which Economic and Noneconomic Factors Drive labor Supply of Older Workers?

Studies on the consequences of introducing and abolishing generous early retirement arrangements in many countries have shown that retirement decisions are sensitive to financial incentives. The quantitative importance of other factors (health, job characteristics, family considerations, peer group effects, etc.) is less clear.

What Are the Advantages and Disadvantages of Gradual Retirement?

Gradual retirement, in the form of either phased retirement reducing hours in the career job, a less-demanding bridge job, or self-employment, is more common in the United States than in Europe. It potentially offers an opportunity to keep people attached longer to the labor market.

What Are the Consequences of Working Longer for Well-Being in Old Age?

It seems clear that working longer may reduce old age poverty if it also leads to higher pensions, but studies on the consequences for health and cognitive skills seem inconclusive.

How can Negative Employer Attitudes toward Older Workers Be Eliminated?

Negative stereotypes and prejudice against older workers are diminishing but are still strongly embedded in society. Positive employer attitudes are often not translated into actions. How effective are legislation against age discrimination and public campaigns promoting a more positive view of older workers?

How Can Lifelong Learning be Promoted?

The effectiveness of investment in skills of older workers depends on how long these workers remain employed, leading to important relations between human capital investment, productivity, retirement saving, and early retirement arrangements.

What Is the Relation between Labor Market Flexibility and the Position of Older Workers?

Compared to the United States, European labor markets are characterized by larger employment protection, less job mobility, longer durations in jobs, and lower hiring rates of older workers. Would removing institutional restrictions and increasing flexibility improve their position?

Does the Economic Crisis Affect the Relevance of Existing Policy Recommendations?

The increasing unemployment rates (among young workers in particular) should not reduce the political pressure on increasing participation of older workers. It is important to dispel the myth of work sharing that early retirement would generate employment for younger workers.

Conclusions

All in all, we know more about labor supply of older workers than about their demand and productivity. Financial incentives matter for retirement decisions, though other, noneconomic factors like job satisfaction, health, working conditions, and attitudes of employers and coworkers also can have important effects. Still, more work is needed on, e.g., joint retirement decisions of husbands and wives, gradual retirement, substitution between exit routes, the interplay between economic and noneconomic factors, etc. Cross-country analysis exploiting the institutional variation seems particularly fruitful. Recent pension reforms in many countries give room for applying quasiexperimental approaches and improved identification of causal effects.

More research challenges remain concerning employer attitudes and labor demand. Good microdata are quite scarce here and the heterogeneous nature of jobs and firms increases the difficulty of this task. Qualitative research using cross-country comparisons suggests that a "coordinated and comprehensive package of age friendly employment measures and policies" leads to the best results (Sigg, 2007), but much more work is necessary to determine what characterizes optimal age management. In particular, the interplay between economic and noneconomic is poorly understood.

Older workers must be attractive to employers, because of their skills, productivity, and wage costs. Attitudes of employers toward older workers vary across countries and have to change in some countries. Subsidies for investing in older workers may help; the effects of employment protection and other features of labor market flexibility and opportunities for self-employment after a career job need to be better understood. Job characteristics, the organization of work, accommodation of workers with a disability and the recognition of the comparative advantages of older workers can contribute to increasing productivity and participation of older workers. In future research, the variation across countries in institutions and institutional reforms can be better exploited to improve our understanding of the mechanisms at work.

See also: Adult Education and Lifelong Learning: The US Experience and Beyond; Aging and the Labor Market; Family and Work; Happiness and Work; Welfare Reform; Welfare and Work.

Bibliography

Alesina, A., Glaeser, E., Sacerdote, B., 2006. Work and leisure in the U.S. and Europe: why so different? In: Gertler, M., Rogoff, K. (Eds.), NBER Macroeconomic Annual 2005. MIT Press, Cambridge, MA, pp. 1–64.

Bassanini, A., Duval, R., 2006. The Determinants of Unemployment across OECD Countries: reassessing the Role of Policies and Institutions. In: OECD Economic Studies, vol. 42. OECD, Paris.

Blanchard, O., 2004. The economic future of Europe. Journal of Economic Perspectives 18 (4), 3–26.

Bonsang, E., Adam, S., Perelman, S., 2012. Does retirement affect cognitive functioning? Journal of Health Economics 31, 490–501.

Cheron, A., Hairault, J.-O., Langot, F., 2011. Age-dependent employment protection. Economic Journal 121 (557), 1477–1504.

Crépon, B., Aubert, P., 2003. Productivité et salaire des travailleurs âgés. Economie et Statistique 368, 157–185.

European Commission, 2007. Discrimination in the European Union. In: Special Eurobarometer, 263. European Commission, Brussels.

European Commission, 2012. EEO Review: Employment Policies to Promote Active Ageing 2012. Directorate-General for Employment, Social Affairs and Inclusion, Brussels.

Euwals, R., de Mooij, R., van Vuuren, D., 2009. Rethinking Retirement from Participation towards Allocation. Special Publication No. 80, April 2009. CPB Netherlands Bureau for Economic Policy Analysis, The Hague, The Netherlands.

Freeman, R., Schettkat, R., 2001. Marketization of production and the US-Europe employment gap. Oxford Bulletin of Economics and Statistics 63, 647–670.

French, E., Jones, J.B., 2011. The effect of health insurance and self-insurance on retirement behavior. Econometrica 79 (3), 693–732.

Galama, T., Kapteyn, A., Fonseca, R., Michaud, P.-C., 2013. A health production model with endogenous retirement. Health Economics 22 (8), 883–902.

Giarini, O., 2009. The four Pillars, the financial crisis and demographics – challenges and opportunities. The Geneva Papers of Risk and Insurance 34, 507–511.

Grossman, M., 1972. On the concept of health capital and the demand for health. Journal of Political Economy 80 (2), 223–255.

Gruber, J., Wise, D., 2004. Social Security Programs and Retirement Around the World: Micro-estimation. University of Chicago Press, Chicago.

Heckman, J., 2000. Policies to foster human capital. Research in Economics 54, 3–56.

Ilmarinen, J., 2006. Towards a Longer Work Life! Ageing and Work Life Quality in the European Union. Finnish Institute of Occupational Health and Ministry of Social Affairs and Health, Helsinki.

International Labour Organisation (ILO), 1982. Statistics of Labour Force, Employment, Unemployment and Underemployment. Report Prepared for the Thirteenth International Conference of Labour Statisticians (Geneva, 18–29 October 1982).

Jacobs, B., 2010. Human capital, retirement and saving. In: Bovenberg, A.L., van Soest, A., Zaidi, A. (Eds.), Ageing, Health and Pensions in Europe: An Economic and Social Policy Perspective. Palgrave Macmillan, New York, pp. 283–317.

Kalwij, A., Vermeulen, F., 2008. Health and labour force participation of older people in Europe: what do objective health indicators add to the analysis? Health Economics 17, 619–638.

Kapteyn, A., Andreyeva, T., 2008. Retirement Patterns in Europe and the US. Panel paper No. 6. Tilburg University, Netspar.

Kapteyn, A., Kalwij, A., Zaidi, A., 2004. The myth of worksharing. Labour Economics 11 (3), 293–313.

Karpinska, K., Henkens, K., Schippers, J., 2013. Retention of older workers: impact of managers' age norms and stereotypes. European Sociological Review 29 (6), 1323–1335.

Lallemand, T., Rycx, F., 2009. Are older workers harmful for firm productivity? De Economist 157 (3), 273–292.

Lazear, E., 1979. Why is there mandatory retirement? Journal of Political Economy 87, 1261–1284.

Leuven, E., 2005. The economics of private-sector training: a review of the literature. Journal of Economic Surveys 19, 91–111.

Marin, B., 2013. Welfare in an Idle Society? Reinventing Retirement, Work, Wealth, Health and Welfare. Ashgate, Aldershot.

McNair, S., Flynn, M., Dutton, N., 2007. Employer Responses to an Ageing Workforce: A Qualitative Study. Research Report No. 455. Department for Work and Pensions, Leeds.

OECD, 2004. OECD Employment Outlook. OECD, Paris.

OECD, 2006. Live Longer, Work Longer. OECD, ELS, Paris.

Riach, P., Rich, J., 2002. Field experiments of discrimination in the market place. Economic Journal 112, 480–518.

Rohwedder, S., Willis, R., 2010. Mental retirement. Journal of Economic Perspectives 24 (1), 119–138.

Sigg, R., 2007. Extending working life: evidences, policy challenges and successful responses. In: Marin, B., Zaidi, A. (Eds.), Mainstreaming Ageing: Indicators to Monitor Sustainable Progress and Policies. Ashgate, Farnham, pp. 447–489.

Skirbekk, V., 2003. Age and Individual Productivity: A Literature Survey. Max Planck Institute for Demographic Research Working Paper 2003-028. Max-Planck Institute for Demographic Research, Rostock.

Taylor, P., Walker, A., 1994. The ageing workforce: employers' attitudes towards older people. Work, Employment and Society 8, 569–591.

Van Dalen, H.P., Henkens, K., Schippers, J., 2009. Dealing with older workers in Europe: a comparative survey of employers' attitudes and actions. Journal of European Social Policy 19 (1), 47–60.

Vickerstaff, S., 2010. Older workers: the unavoidable obligation of extending our working lives? Sociology Compass 4 (10), 869–879.

Zaidi, A., 2012. Mobilising the Potential of Active Ageing and Silver Economy: Opportunities and Challenges for Social and Economic Development. Discussion paper, Peer Review on age friendly goods and services, Poland 2012.

Zaidi, A., Markovec, M., Fuchs, M., 2007. Transition from work to retirement in EU countries. In: Marin, B., Zaidi, A. (Eds.), Mainstreaming Ageing: Indicators to Monitor Sustainable Progress and Policies. Ashgate, Farnham, pp. 395–419.

Zaidi, A., Zólyomi, E., 2012. Active Ageing: What Differential Experiences across EU Countries?. European Papers on the New Welfare, Paper No. 17: Counter-ageing Policies.

Old and New Institutionalism in Economics

David Dequech, Institute of Economics, University of Campinas, São Paulo, Brazil

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The original (or old) and the new institutionalisms are alive in economics. Both are heterogeneous, dynamic, and thus difficult to define. In addition, despite some important differences, there are also some similarities between them, as well as some complementarity. Consequently the intellectual demarcation between the old and the new institutional economics is not always clear-cut. Sociologically, the original institutionalism has been marginalized in the profession for several decades, whereas some subsets of the new institutionalism have managed to penetrate mainstream economics, partly, but not only, because of a greater degree of mathematical formalization. Furthermore, other approaches that emphasize institutions are also relevant.

Introduction

Although institutions – mostly by other names – had been discussed since the days of Adam Smith and other political economists, institutional economics as such was born in the United States in the early twentieth century, with Thorstein Veblen, Wesley Mitchell, and John Commons as its founding fathers.

According to Malcolm Rutherford (2001: pp. 173–174), this early institutionalism "became a significant element in American economics in the interwar period, only to decline rapidly in position and prestige after World War II." In other words, it became part of – and later was excluded from – what was mainstream economics in America, if one adopts the following definition: "mainstream economics is that which is taught in the most prestigious universities and colleges, gets published in the most prestigious journals, receives funds from the most important research foundations, and wins the most prestigious awards" (Dequech, 2007: p. 281). Also part of American mainstream economics at that time was neoclassical economics, which had emerged in the 1870s in Europe and is a school of economic thought defined by its reliance on the assumptions of (1) utility maximization, as a criterion for rationality, and (2) equilibrium, as the state in which the (micro or macro) economic system under study is or tends to be. It may seem obvious, especially in the eyes of other social scientists, that institutions matter greatly in economic issues, but, for several decades, this was not part of the consensus among economists. Neoclassical economics, in particular, emulated physics and often did not treat economics as an explicitly social science. After World War II, its prominence grew, together with its degree of mathematical formalization. Indeed, it can be argued that its influence grew – at the expense of the original institutionalism, in the case of the United States – largely because it became more mathematical. In the 1950s and 1960s, a specific subset of neoclassical economics, namely standard general equilibrium theory, represented by the Arrow–Debreu model, was very influential and prestigious. This model is substantially ahistorical and noninstitutional, at least in its explicit hypotheses (for a partly similar discussion by a new institutionalist, see Joskow (2008: pp. 1–4), on what he calls neoclassical economics (as distinct from a subset thereof), in the first three decades after 1945). In one of the possible interpretations, the Arrow–Debreu model is intended to be applicable to any economy, in any place and at any moment in time. In a less generalist reading, the model is mono-institutional, as it explicitly assumes only one institution: the market. Moreover, it conceives this institution in an extremely restrictive manner: this is not a market that resembles markets in the real world, but rather an imaginary market, reduced to the price mechanism (the interaction between supply and demand). Other institutions, including a few institutions without which real markets cannot exist, are ignored. Property rights are taken for granted. There are no transaction costs, no information asymmetries, and no binding contracts. The list of goods and services and their respective qualities are perfectly known by every potential buyer. Firms (or self-employed workers–producers) are reduced to a production function, and organizational issues are neglected.

This situation began to change in the 1970s, as economists became increasingly interested in institutions (again). In terms of impact in the profession, a crucial factor in this resurgence of interest was the appearance of the self-identified new institutional economics (NIE). The term 'new institutional economics' was coined by Oliver Williamson in a 1975 book. It was then that the approach originally developed by Veblen, Commons, Mitchell, and their descendants started to be called the 'old' institutional economics (OIE). Both because the adjective 'old' may sound derogatory and because that approach is still alive and has actually been revitalized and improved, the more neutral term 'original institutional economics' is employed here (with the same acronym, OIE).

Neither the OIE nor the NIE is easy to define or to characterize. As a matter of principle, the definitions should be based on ideas. In this sense, the economists who defend a certain set of ideas should be called original institutionalists or new institutionalists. At the same time, communication is not helped by definitions that are very distant from common usage. One must also consider the ideas that are defended by economists who see themselves, and/or are usually seen by others, as members of one of these two subsets of institutional economics. In this second sense, the OIE is that which the original institutionalists do; the same applies to the NIE. However, this does not necessarily amount, in each case, to a consistent group of ideas. In fact, both the OIE and the NIE are heterogenous. In addition, some of the ideas in the OIE may be similar to (and others may be compatible with) ideas in the NIE. As a result of all these complications, the demarcation between the OIE and the NIE is not always clear-cut. In any

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71070-1

case, the label that is usually attached to a work is not as important as the work's intellectual content.

The Original Institutional Economics

The American roots of the OIE can be traced back to the pragmatist philosophy of John Dewey, William James, and Charles Sanders Peirce. This was a strong and distinctive influence, which partly accounts for the originality of the approaches of Veblen, Commons, and other early institutionalists. For Veblen, in particular, Darwin was another major source of inspiration, leading him to propose an evolutionary economics. The term 'institutional economics' was coined in 1919 by Walton Hamilton in an article in the American Economic Review. In the following two decades, the OIE gained considerable prominence in American academia, as mentioned above. Although marginalized in the postwar period, and remaining so until today, the OIE has survived. At first, this happened mostly in the United States, even though European scholars like Karl Polanyi and William Kapp (who both taught at American universities for some time) became influential within the approach, as did Gunnar Myrdal, perhaps to a lesser extent. Authors like John Kenneth Galbraith, Clarence Ayres, Dudley Dillard, and Allan Gruchy were among the main exponents of the OIE in the first decades after World War II (Gruchy referred to this generation as the neoinstitutionalists, but the label is not used often). The OIE even gained an organizational basis, with the creation of the Association for Evolutionary Economics (AFEE), in 1965. AFEE has taken part in the Allied Social Sciences Association annual meetings and, since 1967, has published the Journal of Economic Issues (which is available through JSTOR, journal storage). Many of the scholars who can be called original institutionalists have some connection with AFEE and/or with its journal. Especially after the mid-1980s, the OIE has spread through Europe and other parts of the world.

There is no unified grand theory in the OIE. Hodgson (2004: p. vxii) states that "by 1945 American institutionalism was fractured and diverse, and lacking a consensus on its own methodological and theoretical foundations. Despite the contributions of postwar theorists, these fractured foundations largely account for the failure of institutionalism to re-establish anywhere the prominence that it enjoyed in interwar America." At least two strands of the OIE coexist, one more influenced by Veblen, the other by Commons (Rutherford, 1994: pp. 1–2). They are not necessarily in conflict, but a consensus synthesis does not seem to have emerged yet. Not even an outright rejection of neoclassical economics unites the OIE. According to Warren Samuels (1995: p. 569, n. 2), some institutionalists 'consider [original] institutionalism and neoclassicism to be supplementary.' This seems to be a minority position among current members of AFEE, but historically it has been mostly associated with the Commons' branch of the OIE. Samuels himself defended this view as an eclectic one, which for him did not require internal consistency.

Despite its internal diversity, the OIE can be said to embrace at least three core theoretical ideas (which are variously interconnected and developed within the approach): open-ended evolution, habit, and the strong influence of institutions on individuals.

Veblen was particularly influential in the adoption of an evolutionary perspective combined with the argument that there is no predetermined end to the evolutionary process. Relatedly, the notion of equilibrium, as distinct from process, is not often employed in the OIE. Institutions are the transmission mechanism in the economic evolutionary process. Other evolutionary approaches have been developed in economics, among which the neo-Schumpeterian one is the most compatible with the OIE. Without getting into details about the controversial relation between technology and institutions, one can more vaguely state that technological innovation plays a major role in any original institutionalist theory of economic change.

Habits have been conceptualized in different ways within the OIE. Some authors treat habits as synonymous with habitual behavior, while others see habits as dispositions (inclinations) to behave or to think in a certain way, so that habitual behavior and habitual thought are not habits, but manifestations of habits. In any case, habits imply that not all behavior and thought is completely deliberate and purposeful, in contrast to an influential view in economics. The role of habits has occasionally been overstated by some original institutionalists, which makes it difficult to explain endogenous change. In addition, habits imply some automaticity and inertia. On the other hand, the role of habits is more adequately understood when combined with the notion of bounded rationality, of which the early original institutionalists were precursors (Herbert Simon, who coined the term 'bounded rationality,' acknowledged Commons' influence). Given that the human mind is limited and decision environments are often complex, habits free mental resources from some tasks and thus allow us to make deliberate and sometimes innovative decisions about other issues. From this perspective, habits are not seen as usually opposed to purposeful deliberation nor to innovation, but rather are preconditions for them, although habits can in fact sometimes get in the way.

The OIE is against taking the individual as given in relation to institutions. On the contrary, it argues that individuals are strongly influenced by institutions. We propose below a way of specifying the meaning of this strong influence. In the past, some original institutionalists may have taken this argument too far and fallen into some sort of collectivist determinism. (See Rutherford (1994: pp. 38–42) and Hodgson (1998: p. 181). It must be noted, however, that not all self-identified methodological collectivists or holists have committed this mistake.) Currently there seems to be a consensus within the OIE that individuals and institutions are mutually dependent.

Since its inception, the OIE has been very open to influences from, and dialogue with, other social science disciplines. This may be a defining feature of the OIE (as argued by Gruchy), but it does not differentiate it from the NIE, which is also very interdisciplinary, as explicitly argued by some of its main exponents. What is distinctive of the OIE in comparison with the NIE in this regard is that the OIE is one of the schools of economic thought that are closest to the nonrational-choice approaches in the other social science disciplines.

In methodological terms, the OIE does not share the widespread belief among economists that rigor requires mathematical formalization. This is an important reason why the OIE is not part of contemporary mainstream economics, which is unified by that belief, if by any idea. Some OIE models are mathematical, but most are not.

In political-ideological terms, the OIE is predominantly in favor of an active role of government in planning and regulating the economy, although this need not be seen as a defining feature of the approach, as a few authors who have associated themselves with the OIE had a more laissez-faire orientation. In the interwar period in the United States, original institutional economists reached their peak of influence not only in academia, but also in government. Several of them held important positions during the New Deal.

The New Institutional Economics

Some of the earlier uses of the label 'new institutional economics' may have applied it to transaction cost economics, but the NIE became a much broader and more complex set of approaches. The NIE too has an organizational basis, namely the International Society for the New Institutional Economics (ISNIE), which was founded in 1997. ISNIE has an annual conference, but, unlike AFEE, it does not have its own journal. According to the ISNIE website, "[t]he New Institutional Economics is an interdisciplinary enterprise combining economics, law, organization theory, political science, sociology and anthropology to understand the institutions of social, political and commercial life. It borrows liberally from various social-science disciplines, but its primary language is economics. Its goal is to explain what institutions are, how they arise, what purposes they serve, how they change and how – if at all – they should be reformed." This characterization is quite open in terms of theoretical, methodological, and political ideas. It allows for the presence of different approaches within the NIE, with some important foundational differences among them. In turn, this makes it difficult to refer to the NIE as a school of thought – perhaps even more so than in the case of the OIE.

Transaction costs are an important concept in the NIE and play a central role in the work of Coase (who first developed the notion and its use in the theory of the firm), Williamson and North, among others. Given the importance of these authors within the NIE and the above-mentioned fact that Williamson coined the term 'new institutional economics,' no one would deny that transaction cost economics is a crucial subset of the NIE. It is based, according to Williamson (1985: p. 30), on two behavioral assumptions: bounded rationality and opportunism (self-seeking with guile). In addition to transaction cost economics, other subsets of the NIE also adopt some variant of the bounded rationality hypothesis. This leads us to a controversial issue: the relation between the NIE and neoclassical economics. Depending on how it is interpreted, the assumption of bounded rationality may be either contrasted with, or treated as a special case of, utility maximization. The latter alternative was rejected by Simon. In the former alternative, which opposes bounded and maximizing rationality, substantial parts of the NIE clash in this specific regard with neoclassical economics, as defined above. This seems to be implied by Claude Ménard and Mary Shirley (2005: p. 1), for example, when they state that NIE "abandons the standard neoclassical assumptions that individuals have perfect information and unbounded rationality and that transactions are costless and instantaneous. NIE assumes instead that individuals have incomplete information and limited mental capacity and because of this they face uncertainty about unforeseen events and outcomes and incur transaction costs to acquire information." This means rejecting at least one assumption that is essential for neoclassical economics in any of its variants (and not only the 'standard' one), namely unbounded rationality, which seems necessary for utility maximization. (We will not discuss here whether positive transactions costs and bounded rationality are inevitably linked to one another). Ménard and Shirley (2005: p. 2) insist that new institutionalists "accept orthodox assumptions of scarcity and competition," and thus "NIE does not [completely] abandon neoclassical economic theory" (for a similar statement, but without any references to utility maximization or bounded rationality, see Alston (2008)). Ménard and Shirley seem to define neoclassical economics more restrictively than we do here – perhaps as something similar to the Arrow–Debreu version of general equilibrium theory. For Ménard and Shirley, NIE "has a much broader reach than neoclassical economics, which has been largely concerned with prices and outcomes. … NIE asks questions that neoclassical economics does not address" (p. 2).

So far, we have mentioned the possibility of (partly) opposing the NIE and neoclassical economics and/or differentiating them in terms of scope. Other interpreters, however, identify a neoclassical wing or strand within the NIE. It is true that neoclassical economics did largely neglect institutions for several decades, but such neglect was not implied by the defining ideas of neoclassical economics, namely, utility maximization and equilibrium. On the contrary, since at least the 1970s the neoclassical approach has been extended in order to deal with institutions other than the market, as well as with the market as more than the price mechanism, moving beyond the Arrow–Debreu model in both aspects. In their book-length study of the NIE, Eirik Furubotn and Rudolf Richter (2005) accept that some members of the NIE do indeed adhere to core neoclassical assumptions, especially utility maximization, but Furubotn and Richter claim that this has persisted only at the cost of unsolved inconsistencies in what they call 'hybrid models.' The difficulties in formalizing bounded rationality seem to have played a part in the combination of the new institutionalism and the utility maximization hypothesis, but there exist mathematical models that assume some form of bounded rationality without treating it a special case of utility maximization.

Also controversial is the relation between the NIE and Austrian economics. One of the first books to refer to the NIE in the title or subtitle was a collection of essays edited by Richard Langlois (1986) and containing, among others, a few chapters written by Austrian economists, including Langlois. The NIE may not have evolved in the direction that Langlois and other Austrians had signaled, but some new institutionalists still have links with Austrian economics. Accordingly, some internal and external observers have referred to an Austrian strand within the NIE (perhaps even North could nowadays be classified

therein, as he has in some senses become closer to the Austrian approach over the years). In contrast, other important reviewers (e.g., Furubotn and Richter, 2005) seem to treat the NIE in such a way as to exclude an Austrian strand. Among the original proponents of the NIE, so did Williamson (1985).

In his characterization of the NIE and its scope, Williamson (2000) distinguishes among four levels of analysis. "The top level is the social embeddedness level. This is where the norms, customs, mores, traditions, etc. are located. … Level 1 is taken as given by most institutional economists." The supposed reason for this is that "[i]nstitutions at this level change very slowly – on the order of centuries or millennia" (p. 596). This would be the terrain of economic historians, sociologists, and other social scientists outside economics. Williamson seems unaware, however, that several economists have been studying informal institutions. This has long been true of the OIE, and it has become increasingly so within the NIE and within mainstream economics. "The second level is referred to as the institutional environment" (Williamson, 2000: p. 598). This includes formal rules such as constitutions, contract laws, and property rights. Only exceptionally will major events (such as civil wars, occupations, etc.) open a window of opportunity for broad reform. "Absent such a window, major changes in the [formal] rules of the game occur on the order of decades or centuries" (p. 598). The third level is that of the institutions of governance. This is what transaction cost economics investigates, focusing on the governance of contractual relations and comparing different governance structures (such as markets, hybrids, firms, bureaus, etc.). Here lies, for example, the issue of vertical integration or, in other words, the make-or-buy decision that was originally explored by Coase. "The possible reorganization of transactions among governance structures is re-examined periodically, on the order of a year to a decade, often at contract renewal or equipment renewal intervals" (Williamson, 2000: p. 599). A large part of the NIE works at levels two and three. When discussing level three, Williamson (2000: p. 599) himself notes that Commons was a precursor of this kind of analysis. Within the OIE, the strand more strongly influenced by Commons has continued this line of research. Finally, the fourth level refers to resource allocation, involving the determination of prices and quantities, and incentive alignment. This "is the level at which neoclassical economics works. Optimality apparatus, often marginal analysis, is employed" (Williamson, 2000: p. 600). In terms of time reference, 'adjustments to prices and output occur more or less continuously' (Williamson, 2000). Williamson's scheme may suggest that at this level institutions do not matter, except for incentives.

Methodologically, the NIE is, on average, much more inclined to mathematical formalization than the OIE. On the other hand, some important new institutionalists, such as Coase, North and Williamson, have not usually built mathematical models (although some of their ideas have been expressed in mathematical terms by others). Nor have the representatives of the Austrian wing of the NIE.

The greater degree of mathematical formalization certainly helps explain why a significant portion of the NIE has been admitted into mainstream economics, but the three authors mentioned in the previous paragraph have, even without mathematically formalizing their ideas, reached the pinnacle of prestige in the profession by winning the Nobel Prize in Economics. So did Elinor Ostrom, who shared the Prize with Williamson and was a member of ISNIE. In other cases, the fact that more traditional neoclassical economics already was part of mainstream economists has helped the extension of the neoclassical apparatus to a range of institutional issues join the mainstream club. In contrast, there are also parts of NIE – especially if we can include an Austrian wing therein – that have remained at the margins of mainstream economics.

The NIE is not defined by a unified political-ideological stance. Several of its proponents criticized the way earlier authors, using neoclassical theory and the notion of market failure, had defended state intervention on the basis of a comparison between a concrete economy and an ideal benchmark deprived of many institutions. The NIE argues instead for a more realist comparative institutional analysis, which may or may not be favorable to an institutional set up with a more active state. Some important authors within the NIE have had links with the Mont Pelerin Society, like Coase and property rights theorists Armen Alchian and Harold Demsetz, and several others are inclined toward a minimal state, but not all new institutionalists are neoliberals.

The Demarcation between the Original and the New Institutional Economics

There are numerous issues about which the OIE and the NIE can be compared and contrasted. In an attempt to differentiate and legitimate their approach, some authors, such as Coase, Langlois, and Williamson, have tried to distinguish the NIE from the OIE by arguing that the latter is nontheoretical or, even worse, antitheoretical. This was not true of Veblen and other early authors, nor is it true in recent decades. For example, the collection of articles originally published in the Journal of Economic Issues and edited by Marc Tool (1988) is intended to provide a comprehensive set of theoretical discussions of several issues that are relevant for the OIE (even if somewhat dated, this compilation is still useful, because of its organization, quality, and comprehensiveness). More recently, another major example is the work of Geoffrey Hodgson, the main and most prolific theorist in the original institutionalist tradition nowadays.

For Hodgson (1998), the crucial demarcation criterion is that the NIE takes the individual as given, for the purposes of economic analysis, while the OIE does not. Other important original institutionalists have defended similar arguments. Dequech (2002), however, maintained that recent developments in the NIE have made the demarcation more complicated. Some proponents of the NIE can be said to have acknowledged and even emphasized what we generically called above the strong influence of institutions on individuals. But what exactly does this strong influence mean?

This can be discussed with the help of the following distinction among types of influence that institutions exert on individual thought and behavior. The cognitive influence of institutions is threefold: informational, practical, and profound. They provide information (in particular about the likely behavior of others), they incorporate practical or tacit knowledge about how to do things, and they influence the way

in which individuals select, organize, and interpret information. The motivational influence of institutions is twofold: simple and profound. The simple variant consists in providing incentives, given the individuals' values, objectives, and obligations. The profound motivational influence is exerted on the values that individuals hold, the objectives they pursue in their instrumental actions, and the obligations they attribute to themselves in their noninstrumental actions. Institutions may also influence the individuals' emotional state, thereby having an emotional influence. Either directly (as rules that give power to individuals occupying certain positions) or through the above-mentioned roles they play, institutions allow individuals to do many things. They thus have an enabling influence. On the other hand, institutions also act as constraints; this is their restrictive influence.

What was called above the strong influence of institutions on individuals can now be argued to include, more specifically, the profound variant of (1) the cognitive and/or of (2) the motivational influences. While arguing that institutions constrain individuals and provide them with information and incentives, one could still see the individual as essentially given in relation to institutions. If, however, one accepts that institutions affect the way in which people select, organize, and interpret information, and/or that institutions influence people's values, objectives, and obligations, then one should acknowledge that institutions partly explain individuals, instead of being completely explained by them. In this case, the individual cannot be the single starting point of institutional theory.

This latter argument goes against what can be called extreme forms of methodological individualism. On the other hand, going beyond the argument that institutions strongly influence people's cognition and motivation, and maintaining that institutions determine them would correspond to an extreme form of methodological collectivism. However, terms like methodological individualism or methodological collectivism have been given a wide array of different meanings. Instead of using these labels, it is better to discuss how a certain approach or work allows or does not allow institutions to have those various types of influence on individuals, and individuals to transform institutions.

Whether the profound variety of the cognitive and motivational influences is highlighted or not is indeed a crucial question that distinguishes some approaches to institutions from others. It does work to distinguish the OIE from a large part of the NIE – but not all the NIE. Eminent authors who are associated with the NIE, such as North (1990, 2005), Greif (2006), and Aoki (2001), have, in their own way and with their own concepts, recognized these profound institutional influences, thus pushing the frontiers of the NIE and, more broadly, of mainstream economics.

This is one among other possible examples that labels like the OIE and the NIE have a limited usefulness.

Other Approaches that Emphasize Institutions

In addition to the OIE and the NIE, there are several other approaches within economics that emphasize institutions. Some of them, like the regulationist school and the economics of conventions (both originated in France), even have major exponents who describe themselves as institutionalists (of a different kind from both the OIE and the NIE). To a lesser extent, this also happens within Post-Keynesian economics, which since its inception in the 1970s has argued that institutions matter (Post-Keynesians who see themselves as institutionalists usually have closer ties to the OIE than to the NIE). The variety of behavioral economics put forward by Herbert Simon has placed special emphasis on organizations and, to a lesser degree, on other institutions. Partly under Simon's influence, so has neo-Schumpeterian economics. Austrian economics – which is not restricted to the Austrian wing of the NIE – has always devoted a great deal of attention to institutions, starting with Carl Menger. A more recent approach, evolutionary game theory, has focused on institutions assuming some sort of bounded rationality. In a sense, we are all institutionalists now.

In the last three or four decades, new institutionalisms have also appeared in other disciplines, like sociology and political science. As Paul DiMaggio (1998) convincingly argues, however, it is better to categorize different works by theoretical orientation than by discipline. DiMaggio distinguishes between three types of new (or neo) institutionalisms. Rational-action new institutionalism "views actors as coherent and strategic, and interests as stable and exogenous." Most of this work "explains the prevalence of economic and political institutions on the basis of problems that these institutions solve for actors who adopt them." Social-constructionist new institutionalism "argues that all the elements of rational-action models – actors, interests, and preferences – are 'socially constructed' and therefore endogenous and variable. … Its proponents emphasize that how pressures to appear legitimate, environmentally imposed uncertainty, and normative schemes embedded in professional training and practice" lead to organizational convergence, not necessarily to greater efficiency. Finally, mediated-conflict new institutionalism focuses "on how state institutions structure and mediate conflict among the interests of groups. … Cross-national policy differences and variation within polities over time are seen as reflecting changes both in the relative power of interests groups and in the openness of political institutions to different forms of political action" (DiMaggio, 1998: pp. 696–697).

A large part of the NIE can be seen as belonging to DiMaggio's rational-action new institutionalism. In contrast, the OIE has a greater degree of proximity to social-constructionist and mediated-conflict institutionalisms. However, classification is not so simple, both within economics, as already argued above, and across disciplines. DiMaggio himself points out that hybrids have emerged, and areas of potential collaboration exist among the three new institutionalisms he identifies. It should be added that outside economics, too, there are approaches that emphasize institutions without being labeled as institutionalist.

Much more dialogue is desirable among the various approaches that study institutions, both within economics and across disciplines. Their languages are varied, and communication is not always easy. Even after suitable translations there will be cases of conflicting ideas. However, there is a great deal of potential similarity or, more frequently and perhaps even

better, compatibility and complementarity between several approaches. Journals like the Journal of Institutional Economics, which is linked to the European Association for Evolutionary Political Economy (EAEPE), help promote an open-minded debate, by publishing contributions from mainstream and nonmainstream economists, as well as noneconomists. In the same vein, also promising is the recently formed World Interdisciplinary Network for Institutional Research (WINIR), which involves scholars from several disciplines and proponents of many different approaches.

See also: Ideologies, Institutions, and the New Institutionalism; Institutionalism and Neo-institutionalism: History of the Concepts; Institutionalism; Institutions; Law and Economics; New Institutionalism in the Analysis of Complex Organizations; Norms; Organizations, Sociology of; Regulation, Economic Theory of; Regulation, Political Economy of; Sociology and the New Institutionalism; Transaction Costs and Property Rights.

Bibliography

Alston, L., 2008. New institutional economics. In: Durlauf, S., Blume, L. (Eds.), The New Palgrave Dictionary of Economics, second ed. Palgrave, London. http://dx.doi.org/10.1057/9780230226203.1183 http://www.dictionaryofeconomics.com/article?id=pde2008_N000170 (accessed 02.02.14.).

Aoki, M., 2001. Toward a Comparative Institutional Analysis. MIT Press, Cambridge, MA.

Dequech, D., 2002. The demarcation between the "old" and the "new" institutional economics: recent complications. Journal of Economic Issues 36, 565–572.

Dequech, D., 2006. The new institutional economics and the theory of behaviour under uncertainty. Journal of Economic Behavior and Organization 59, 109–131.

Dequech, D., 2007. Neoclassical, mainstream, orthodox, and heterodox economics. Journal of Post Keynesian Economics 30, 279–302.

DiMaggio, P., 1998. The new institutionalisms: avenues of collaboration. Journal of Institutional and Theoretical Economics 154, 696–705.

Furubotn, E., Richter, R., 2005. Institutions and Economic Theory: The Contribution of New Institutional Economics, second ed. University of Michigan Press, Ann Arbor.

Greif, A., 2006. Institutions and the Path to the Modern Economy. Cambridge University Press, Cambridge.

Gruchy, A., 1987. The Reconstruction of Economics. Greenwood, New York.

Hodgson, G., 1998. The approach of institutional economics. Journal of Economic Literature 36, 166–192.

Hodgson, G., 2004. The Evolution of Institutional Economics. Routledge, London.

Hodgson, G., Samuels, W., Tool, M. (Eds.), 1994. Elgar Companion to Institutional and Evolutionary Economics. Edward Elgar, Aldershot.

Joskow, P., 2008. Introduction to new institutional economics: a report card. In: Brousseau, É., Glachant, J.-M. (Eds.), New Institutional Economics: A Guidebook. Cambridge University Press, Cambridge, pp. 1–19.

Langlois, R. (Ed.), 1986. Economics as a Process: Essays in the New Institutional Economics. Cambridge University Press, Cambridge.

North, D., 1990. Institutions, Institutional Change and Economic Performance. Cambridge University Press, Cambridge.

North, D., 2005. Understanding the Process of Economic Change. Princeton University Press, Princeton.

Rutherford, M., 1994. Institutions in Economics. Cambridge University Press, Cambridge.

Rutherford, M., 2001. Institutional economics: then and now. Journal of Economic Perspectives 15, 173–194.

Samuels, W., 1995. The present state of institutional economics. Cambridge Journal of Economics 19, 569–590.

Tool, M. (Ed.), 1988. Evolutionary Economics, vols. I and II. M.E. Sharpe, Armonk.

Williamson, O., 1985. The Economic Institutions of Capitalism. Free Press, New York.

Williamson, O., 2000. The new institutional economics: taking stock, looking ahead. Journal of Economic Literature 38, 595–613.

Olfactory System

Heather M Schellinck and Richard E Brown, Dalhousie University, Halifax, NS, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H. Scherer, A. Quast, volume 16, pp. 10847–10853, © 2001, Elsevier Ltd.

Abstract

This article describes the main olfactory system and the accessory olfactory system (AOS), their receptors and their neural projections to the brain. Both olfactory and vomeronasal systems use seven-transmembrane receptors, which respectively, send nerves to the main and accessory olfactory bulb (AOB). The main olfactory bulb has a number of cell types organized in layers, with the AOB situated in the dorsal posterior region of the olfactory bulb. The AOS sends signals to the amygdala and hypothalamus to stimulate pheromonal effects. Tests used to examine olfactory ability in animals and humans are described as are problems associated with the sense of smell in a number of neurological diseases.

The specific processes that result in the perception of an odorant, although crucial for the survival of most mammalian species other than humans, are yet to be fully understood. This article provides an overview of our current knowledge of the anatomical and functional pathways of the main olfactory system (MOS) as well as those of the accessory olfactory system (AOS) found in most mammals other than humans and some primates. We describe the measurement of olfactory function and discuss the factors, other than disease, that may influence the ability to process olfactory information. We conclude by describing diseases that disrupt our sense of smell; some of these disorders are particular to the olfactory system and others affect the central nervous system, in general.

The Main Olfactory System

The MOS consists of the nasal epithelium, the olfactory bulb, and its direct projections to the olfactory cortex on the base of the frontal lobe and the medial aspect of the temporal lobe. The olfactory epithelium, limited by dorsal aspects of the nasal vault, the septum and the superior and anterior middle turbinate bones, contains the bipolar olfactory receptor neurons (ORNs). The highly ciliated end of each ORN projects into the olfactory mucus where its receptors interact with airborne substances. The other end of the ORNs pass through the cribriform plate and project to networks of neuronal processes in the olfactory bulb called glomeruli; together the axons of the ORNs form the olfactory nerve (cranial nerve one). The olfactory epithelium contains supporting cells as well as basal cells that continuously replace ORNs, each of which only has a lifetime of several weeks (see **Figure 1**). Specialized glia, known as olfactory-ensheathing cells, assist in the process of regeneration of the axons of ORNs. Of clinical interest is the research involving the transplantation of these olfactory-ensheathing cells to stimulate regeneration following spinal cord injury (Franssen et al., 2007).

Linda Buck and Richard Axel received the Nobel Prize in 2004 for their discovery of genes in the mouse that code for G-protein-coupled olfactory receptors (Buck and Axel, 1991). The Nobel web site provides access to Linda Buck's address to the Nobel Society that describes their research and the workings of the olfactory system in some detail. Mice have approximately 1000 OR genes; humans have approximately 350 functional OR genes, each of which codes for a different OR protein (Malnic et al., 2004). Each OR neuron has one specific receptor type only.

As humans are able to identify 3000 odorants or more it is clear that each ORN cannot represent one odorant; rather, it is believed that a combinatorial code is in effect. The findings that a specific odorant may activate more than one ORN and similarly that different odorants may activate the same receptor provide evidence for this hypothesis. Exactly how an odorant activates a receptor is not known as of this writing. The most plausible theory suggests that a stereochemical relationship exists between an odorant and a receptor (or receptors) that initiates the G-protein-mediated cellular cascade resulting in the firing of the olfactory neuron (Buck and Bargmann, 2012).

As well as the G-protein-coupled olfactory sensory neurons, there is a second type of odor receptor neuron in the olfactory epithelium, the 'trace amine-associated receptors (TAARs),' which express a family of GPCR-related Taar genes. Odor receptors expressing Taar genes occur in small numbers throughout the olfactory epithelium in rodents and humans and are responsive to biogenic amines (Pacifico et al., 2012). There are seven different TAAR olfactory receptors (TAAR2–9).

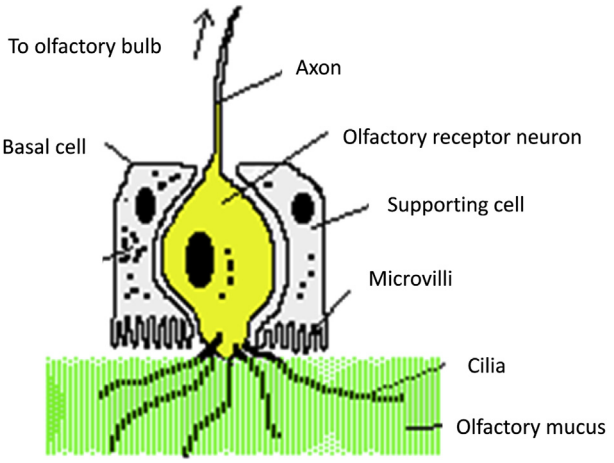

Figure 1 Schematic of the olfactory epithelium as described in the text. The function of the microvilli is not known. From Jacob, T.J.C., 2014. A Tutorial on the Sense of Smell. http://www.cardiff.ac.uk/biosi/staffinfo/jacob/teaching/sensory/olfact1.html (accessed 05.05.14.).

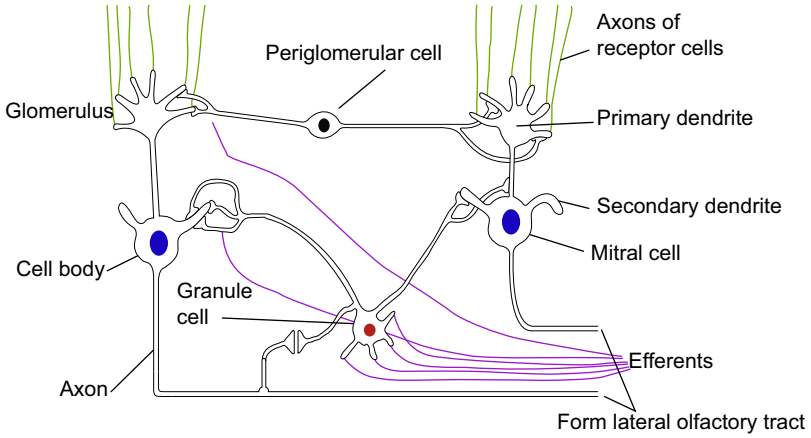

Figure 2 A schematic of the olfactory bulb with connections as described in the text. A tufted cell (not shown) would function similarly to a mitral cell in this diagram. From Jacob, T.J.C., 2014. A Tutorial on the Sense of Smell. http://www.cardiff.ac.uk/biosi/staffinfo/jacob/teaching/sensory/olfact1.html (accessed 05.05.14.).

Three murine TAARs respond to amines found in mouse urine, and one such receptor in mice (the mTAAR5) is responsive to the urine odors of adult male mice, but not to the urine odors of prepubertal males or females. A number of different volatile amine ligands have been found that bind to TAARs. The human TAAR5 is activated by trimethylamine (Wallrabenstein et al., 2013).

The TAAR olfactory sensory neurons project to discrete glomeruli in the dorsal olfactory bulb. Because the TAARs differ genetically from main olfactory receptors, and in their projection patterns to the olfactory bulb and in their axon guidance, they may constitute a distinct olfactory subsystem in the main olfactory epithelium (Johnson et al., 2012). These TAARs may be receptors within the olfactory epithelium that are hardwired to detect social odors such as pheromones or odors that are 'innately aversive' (Dewan et al., 2013)

The olfactory bulb consists of a variety of cell types organized in a number of layers, including the mitral cell, granule cell, and glomerular layers (see **Figure 2**). The glomerulus is spherical in nature and contains the axons of ORNs and dendrites of mitral, tufted, and periglomerular cells (Wachowiak and Shipley, 2006). Each glomerulus generally receives input only from ORNs that express the same receptor. These specific ORNs synapse on one medial and one lateral glomerulus in each olfactory bulb (Ressler et al., 1994). It is not yet clear how the axons of the ORNs arrive at the appropriate glomerulus. Mitral and tufted cells send output directly to the olfactory cortex, which includes such regions as the anterior and posterior cortical nuclei of the amygdala, the olfactory tubercle, and the piriform and entorhinal cortices. The entorhinal cortex sends fibers to the hippocampus, while the piriform cortex provides connections to the orbitofrontal cortex and the hypothalamus (see **Figure 3**). Information from the olfactory bulb output to the cortex does not follow the organizational principles found in the glomeruli. For example, an

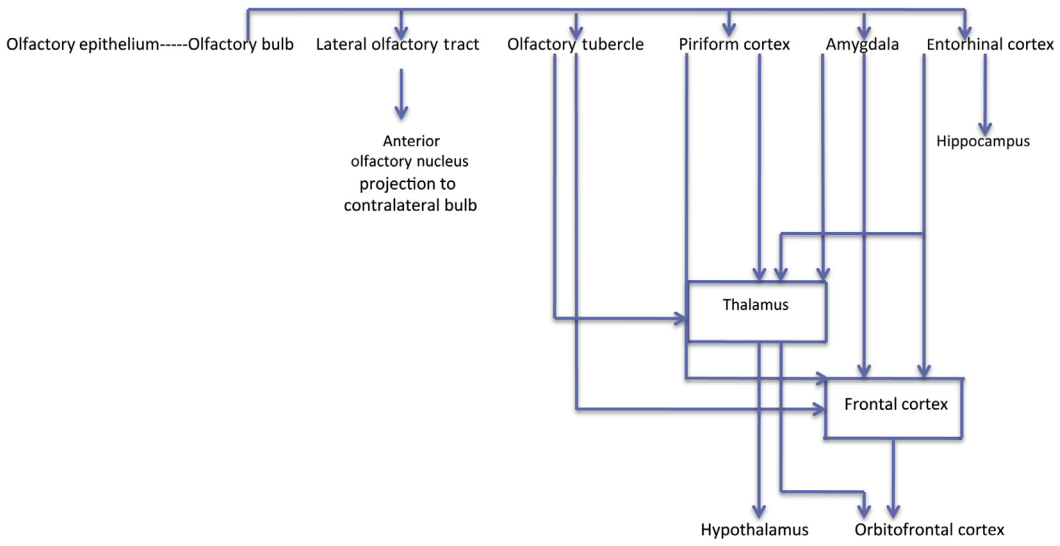

Figure 3 Projections of the olfactory bulb to the olfactory cortex. The amygdala also sends direct projections to the hypothalamus.

individual odor does not activate a particular area of the pyriform cortex, rather neurons appear to be activated in a random fashion. Moreover, mitral cell projections from different areas of the olfactory bulb may project to the same area of the piriform cortex.

Cells in the remaining layers of the olfactory bulb are primarily involved in regulating the output of mitral cells. For example, granule cells form dendrodendritic connections with mitral cells. As mitral cells release glutamate and granule cells release GABA, this can result in the unusual situation whereby mitral cells inhibit their own action or the action of other nearby mitral cells. While the function of this process is not clear, it has been suggested that inhibition may act to enhance the signal-to-noise ratio of incoming information by preventing the firing of other neurons (Arevian et al., 2008). The olfactory cortex through its retrograde projections to the bulb may also modulate responses to specific odors depending upon the motivational or physiological state of the individual.

The Accessory Olfactory System

The AOS consists of the vomeronasal organ (VNO), the accessory olfactory bulb (AOB), and its direct projections to the medial amygdala, which, in turn, sends signals to the hypothalamus. Although humans and other higher order primates have vestigial remnants of a VNO, evidence for any functional response remains elusive (Meredith, 2001). The VNO is known primarily for its role in the detection of pheromones. Pheromones are chemicals that produce a species-specific response, resulting in a change in the physiology or behavior of the animal receiving the signal. For example, a pheromone may signal the availability of a mate or result in a sexual or aggressive response. Although primarily a function of the AOS, it has been determined that, in some instances, the MOS also detects pheromones (Keller et al., 2009).

The VNO is a crescent-like tubular structure located bilaterally in the nasal septum at the base of the nasal cavity. Its duct opens into the nasal cavity where it receives nonvolatile chemical signals, frequently found in urine or in glandular excretions. These signals are detected by receptors on apical microvilli of the sensory epithelium that is located on the concave inner surface of the VNO. The stimuli are delivered to the epithelium through fluid produced by vomeronasal glands. Blood vessels associated with the VNO dilate or constrict and create a pumping action that facilitates the delivery process.

Similar to the ORNs, the VNO neurons have seven-transmembrane receptors but, unlike the olfactory epithelium, the VNO has several different types of receptors: V1R, V2R, and FPRs (Formyl peptide receptor-like proteins). They are structurally unlike ORN receptors are coupled with different G-proteins and use different signaling pathways. The V1R and V2R receptors are found in two distinct zones within the VNO. In the mouse, there are about 100 different receptors of both types (Herrada and Dulac, 1997). The role of FPRs in the VNO is not yet understood although it has been suggested that they may be involved in the sensing of pathogens. Like the relationship of ORNs to glomeruli, VNO receptors of the same type also converge upon mitral cells in specific glomeruli in the AOB.

The AOB itself is embedded in the dorsal posterior area of the main olfactory bulb. It sends signals directly to the medial amygdala, from there to the bed nucleus of the stria terminalis and on to the hypothalamus. It is the action of the hypothalamus that is responsible for some very specific responses to chemical stimuli. In mice, among the most well-known pheromonal effects other than the initiation of mating behaviors and aggression are the phenomena of pregnancy block (Bruce effect), puberty acceleration (Vandenbergh effect), and estrous cycle synchronicity (Whitten effect). Pregnancy block results when a female mouse aborts its pregnancy if it encounters the odor of the nonstud male prior to embryo implantation. The Vandenbergh effect refers to the acceleration of puberty in female mice following the exposure to the odors of an adult male conspecific. The Whitten effect occurs when female mice that are housed in unisexual groups begin to cycle in synchrony when an adult male is introduced to their cage. The chemical constituents of the latter pheromones have been identified and are found in urine. Debate continues regarding the role of other compounds found in urine such as major urinary proteins and major histocompatibility complex (MHC) fragments both of which it has been argued are involved in mate choice and kin and individual recognition.

The discovery that synchronization of the menstrual cycle occurs in women living in group housing (McClintock effect) gave rise to the suggestion that humans are also influenced by chemical stimuli. Follow-up studies by McClintock and others have demonstrated that humans undoubtedly react both consciously and unconsciously to the smells of other individuals. For example, females prefer (or at least find less aversive) the axillary odors of males with a MHC that is different than their own (Wedekind and Penn, 2000). Having a male with different MHC genes as a partner would be of considerable value as it would create greater immunocompetence in resulting offspring. Although it seems clear that humans do not have a functional AOS and that the term 'pheromone' may not be an appropriate designation for such a stimulus, this does not preclude our ability to detect and respond to the odors of another individual (Wyatt, 2014). The MOS includes projections to the limbic system that could mediate both physiological and psychological responses to an odorant.

Assessing Olfactory Function

A number of well-validated tests with normative data collected have been developed to assess olfactory ability in humans. The University of Pennsylvania Smell Identification Test or UPSIT, which can be self-administered and is available in multiple languages is perhaps the most well-known. It consists of 40 different 'scratch and sniff' squares each encapsulating a common odor such as cinnamon, peanut butter, cut grass, or motor oil. It may be used to assess detection, identification, and the ability to discriminate between or remember odors (Doty, 2006).

To measure absolute olfactory threshold, i.e., the lowest concentration of an odorant that can be detected, odors are usually presented sequentially from high to low concentrations (descending limits) or from low to high concentrations (ascending limits). Organic compounds such as phenyl-ethyl

alcohol or n-amyl acetate are generally used in such tests. For this procedure, a well-controlled odor delivery system is essential if accurate results are to be obtained. Electrophysiological techniques may also be used to assess olfactory function. Both the examination of olfactory event-related potentials and responses to an electroolfactogram, which detects electrical changes in the olfactory epithelium in response to an odor stimulus, are useful to assess individuals who are unable to provide reliable responses to less invasive tests. With the advent of structural and functional imaging techniques, the brain structures involved in olfactory information processing may also be examined. For example, an fMRI study has found that the orbital frontal cortex is activated during the judgment of an odor's hedonic value (Katata et al., 2009)

The ability to detect odors is influenced by a number of factors including age, gender, and hormonal status. It has been well established that females have a lower detection threshold than males. Moreover, olfactory acuity in females is more acute during the ovulatory phase of the menstrual cycle (Navarrette-Palacios et al., 2003). The ability of humans to detect odors also declines with normal aging and as a result of disease (see Section Olfactory Dysfunction and Disease). Specific anosmias also exist in a small fraction of the population. The most well-known example being the inability to detect 5-α-androst-16-en-3-one (androstenone) a boar pheromone and a compound found in urine and axillary sweat in humans (Wysocki et al., 1989). Mutations in specific odor receptors are likely to be the origin of specific anosmias.

A variety of paradigms are used to assess olfactory ability in nonhuman mammals. These include the buried food test and the habituation–dishabituation task. In the former paradigm, a food-restricted animal is required to find a food particle by digging in a substrate such as sand or bedding material. In the latter test, the animal's sniffing response to successive presentations of an odor is recorded. Successive presentation of the same odor will result in a decrease in investigation. When a second stimulus is presented, the previously habituated response, i.e., sniffing, will reoccur if the animal can discriminate between the different stimuli. Neither of these procedures is particularly valid or reliable as a negative result is impossible to interpret. If the animal does not detect a stimulus or does not discriminate between two stimuli, this failure could reflect a lack of interest or motivation rather than a lack of ability. Also, sniffing behavior, which is used to measure a response, is not often defined and this makes it difficult to compare results between different studies. Using a conditioned digging task or an operant olfactometer avoids the latter problems and has been demonstrated to provide more reliable results. Animals that have not dug for food or sniffed at a stimulus have been found to detect odors when assessed in these more rigorously designed tasks. The advantages and disadvantages of the various methods have been described by Schellinck and Slotnick (2013).

A comparison of olfactory ability in human and nonhuman mammals demonstrates the exceptional degree of odor sensitivity in some species. For example, although humans have been shown to detect an odorant in parts per billion, both dogs and rodents have been found to detect odors in parts per trillion. These differences may be accounted for in several ways: these animals may have longer cilia with more receptors; they possess a greater number of functional receptor genes and have a larger area of epithelium. Moreover, dogs unlike humans, have two separate airflows, one for respiration and one for olfaction, which provides more air for analysis (Craven et al., 2010). Also, they and other animals may make use of their AOS to process signals associated with bodily odors.

The extraordinary ability of dogs and rodents to detect odors has led to the use of these animals in a number of useful capacities. Tracker dogs have long been employed to search for lost or suspect individuals as well as to detect explosive devices or contraband substances. A species of rat is now replacing dogs in some instances to assist their trainers to find landmines in war-torn countries such as Mozambique and Cambodia and to detect explosives and drugs in Holland. Mice have also been trained to sniff out bombs in Israel. Currently, research is underway to determine if dogs can reliably detect specific volatile organic compounds or 'odors of disease' in the breath of individuals with Type 1 diabetes and cancers of the bladder, lung, and breast, in particular.

Olfactory Dysfunction and Disease

Problems associated with the sense of smell have been categorized according to the degree and type of dysfunction (**Table 1**). At least 70% of a person's sense of flavor is derived from olfactory cues such that anosmia results in food being experienced only as sweet, sour, salty, bitter, or umami. Despite the psychological and physiological deprivation that this situation created for its sufferers, clinicians showed little interest in olfactory dysfunction. With the discovery that the sense of

Table 1 Disorders of olfaction and common causes

Anosmia: Total absence of the sense of smell; this may be a result of the shearing of the olfactory nerve following head trauma; viral infections, toxins, or a genetic abnormality leading to lack of epithelium or receptor proteins.
Dysosmia: Partial impairment of the sense of smell; this is a general term and most disorders are more precisely described using the terms below.
Hyposmia: A Decrease in olfactory acuity; it may be associated with allergies, infection, head trauma, Alzheimer's disease, or Parkinson's disease.
Hyperosmia: An increase in olfactory acuity; this may be a result of a genetic polymorphism resulting in increased sensitivity to a particular odor (e.g., isovaleric acid) or from benzodiazepine withdrawal syndrome.
Parosmia: Smelling a substance other than the one that is present; the affected individual knows that their interpretation is incorrect. It may be caused by damage to ORNs as a result of infection but has also been found in patients with Parkinson's disease. Most commonly odors are identified as rotten or like feces. This is called cacosmia.
Phantosmia: A olfactory hallucination in which the individual believes a smell is present when it is not; it may be a result of psychoses but also experienced prior to an epileptic seizure or a migraine.
Pseudosmia: Misinterpretation of a real olfactory impression sometimes found in individuals with psychoses.

ORN, olfactory receptor neuron.

smell is altered in several neurological diseases, this consideration changed. Hyposmia or anosmia is one of the early signs of Parkinson's disease (PD) and occurs prior to motor symptoms; Individuals with Alzheimer's disease (AD) have deficits in both discrimination and identification of odors with detection becoming more of a problem in the later stages of the disease. Low scores on the UPSIT have also been found in patients with multiple sclerosis and motor neuron disease. For individuals with PD and AD, a greater understanding of olfactory dysfunction may create an early biomarker for these illnesses.

See also: Aging and Memory; Hypothalamus; Neuroendocrinology; Taste and Smell, Psychology of.

Bibliography

Arevian, A., Kapoor, V., Urban, N., 2008. Activity-dependent gating of lateral inhibition in the olfactory bulb. Nature Neuroscience 11, 80–87.

Buck, L., Axel, R., 1991. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell 65, 175–187.

Buck, L., Bargmann, C., 2012. In: Kandel, E., Schwartz, J., Thomas Jessell, T., Steven Siegelbaum, S., Hudspeth (Eds.), Smell and Taste: The Chemical Senses in Principles of Neural Science, fifth ed. McGraw-Hill, New York, pp. 713–735.

Bufe, B., Schumann, T., Zufall, F., 2012. Formyl peptide receptors from immune and vomeronasal system exhibit distinct agonist properties. Biological Chemistry 287, 33644–33655.

Craven, B., Paterson, E., Settles, G., 2010. The fluid dynamics of canine olfaction: unique nasal airflow patterns as an explanation of macrosmia. Journal of Royal Society Interface 6, 933–943.

Dewan, A., Pacifico, R., Zhan, R., Rinberg, D., Bozza, T., 2013. Non-redundant coding of aversive odours in the main olfactory pathway. Nature 497 (7450), 486–489. http://dx.doi.org/10.1038/nature12114. Epub 28 April 2013.

Doty, R.L., 2006. Olfactory dysfunction and its measurement in the clinic and the workplace. International Archives of Occupational and Environmental Health 79, 268–282.

Franssen, E.H., deBree, F.M., Verhaagen, J., 2007. Olfactory ensheathing glia: their contribution to primary olfactory nervous system regeneration and their regenerative potential following transplantation into the injured spinal cord. Brain Research Reviews 56, 236–258.

Herrada, G., Dulac, C., 1997. A novel family of putative pheromone receptors in mammals with a topographically organized and sexually dimorphic distribution. Cell 90, 763–773.

Jacob, T.J.C., 2014. A Tutorial on the Sense of Smell. http://www.cardiff.ac.uk/biosi/staffinfo/jacob/teaching/sensory/olfact1.html (accessed 05.05.14.).

Johnson, M.A., Tsai, L., Roy, D.S., Valenzuela, D.H., Mosley, C., Magklara, A., Lomvardas, S., Liberles, S.D., Barnea, G., 2012. Neurons expressing trace amine-associated receptors project to discrete glomeruli and constitute an olfactory subsystem. Proceedings of the National Academy of Sciences of the United States of America 109 (33), 13410–13415. http://dx.doi.org/10.1073/pnas.1206724109. Epub 26 July 2012.

Katata, K., Sakai, N., Doi, K., Kawamitsu, H., Fujii, M., Sugimura, K., Nibu, K., 2009. Functional MRI of regional brain responses to 'pleasant' and 'unpleasant' odors. Acta Otolaryngology Supplement 562, 85–90.

Keller, M., Baum, M.J., Brock, O., Brennan, P.A., Bakker, J., 2009. The main and the accessory olfactory systems interact in the control of mate recognition and sexual behavior. Behavioural Brain Research 200, 268–276.

Liberles, S.D., Buck, L.B., 10 August 2006. A second class of chemosensory receptors in the olfactory epithelium. Nature 442 (7103), 645–650. Epub 30 July 2006.

Malnic, B., Godfrey, Buck, L., 2004. The human olfactory receptor gene family. Proceedings of the National Academy of Sciences USA 101, 2584–2589.

Meredith, M., 2001. Human vomeronasal function: a critical review of best and worst cases. Chemical Senses 26, 433–445.

Navarrette-Palacios, E., Hudson, R., Reyes-Guerrero, G., Guevarapuzman, R., 2003. Lower olfactory threshold in the ovulatory phase of the menstrual cycle. Biological Psychology 63, 269–279.

Pacifico, R., Dewan, A., Cawley, D., Guo, C., Bozza, T., 2012. An olfactory subsystem that mediates high-sensitivity detection of volatile amines. Cell Reports 2 (1), 76–88. http://dx.doi.org/10.1016/j.celrep.2012.06.006. Epub 28 June 2012.

Ressler, K.J., Sullivan, S.L., Buck, L.B., 1994. Information coding in the olfactory system: evidence for a stereotyped and highly organized epitope map in the olfactory bulb. Cell 79, 1245–1255.

Schellinck, H., Slotnick, B., 2013. Olfactory acuity in Mus musculus. In: Crusio, W.E., Sluyter, F., Gerlai, R.T., Peitropaolo, S. (Eds.), Behavioural Genetics of the Mouse, Genetics of Behavioral Phenotypes, vol. 1. Cambridge University Press, Cambridge, UK, pp. 65–71.

Vassar, R., Chao, S., Sitcheran, R., Nuiiez, J., Vosshall, L.B., Axel, R., 1994. Topographic organization of sensory projection to the olfactory bulb activity-dependent gating of lateral inhibition in the mouse olfactory bulb. Cell 79, 981–991.

Wachowiak, M., Shipley, M., 2006. Coding and synaptic processing of sensory information in the glomerular layer of the olfactory bulb. Seminars in Cell and Developmental Biology 17, 411–423.

Walker, J.C., Hall, S.B., Walker, D.B., Kendal-Reed, M.S., Hood, A.F., Niu, X.F., 2003. Human odor detectability: new methodology used to determine threshold and variation. Chemical Senses 28, 817–826.

Walker, D., Walker, J., Peter James Cavnar, P.J., Taylor, P., Pickel, D., Hall, S., Suarez, J., 2006. Naturalistic quantification of canine olfactory sensitivity. Applied Animal Behaviour Science 97, 241–254.

Wallrabenstein, I., Kuklan, J., Weber, L., Zborala, S., Werner, M., Altmüller, J., Becker, C., Schmidt, A., Hatt, H., Hummel, T., Gisselmann, G., 2013. Human trace amine-associated receptor TAAR5 can be activated by trimethylamine. PLoS One 8 (2), e54950. http://dx.doi.org/10.1371/journal.pone.0054950. Epub 5 February 2013.

Wedekind, C., Penn, D., 2000. MHC genes, body odours, and odour preferences. Nephrology Dialysis Transplantation 15, 1269–1271.

Wyatt, T.D., 2014. Pheromones and Animal Behavior: Chemical Signals and Signature Mixtures. Cambridge University Press, Cambridge.

Wysocki, C.J., Dorries, K.M., Beauchamp, G.K., 1989. Ability to perceive androstenone can be acquired by ostensibly anosmic people. Proceedings of the National Academy of Sciences of the United States of America 86, 7976–7978.

Oligarchy, Iron Law of

Darcy K Leach, Bradley University, Peoria, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The iron law of oligarchy refers to a provocative and very influential theory posited by German social theorist, Robert Michels. In his seminal analysis of the German Social Democratic Party in 1911, Michels argued that all organizations – no matter how democratic their original intentions – eventually come to be ruled by a powerful entrenched minority that, when necessary, will act illegitimately to squelch internal opposition and divert the organization's goals in order to maintain its power. This article summarizes Michels' argument, reviews the theoretical debates it has provoked, outlines the findings of the empirical research evaluating his claims, and discusses some of the methodological difficulties that have plagued this inquiry.

In 1911, Robert Michels summarized his analysis of the German Social Democratic Party with the now-famous phrase "Whoever says organization, says oligarchy" (Michels, 1962: p. 365). Though a dedicated socialist at the time, Michels nevertheless concluded that in modern society, both socialism and democracy were structurally impossible – that the very principle of organization made oligarchy the inevitable result of any organized collective endeavor. This conclusion – known as the 'iron law of oligarchy' – has probably been the most influential and widely debated claim in the study of organizations. Despite over a century of empirical research in a range of subfields in political science and sociology, however, there is still no consensus about whether or under what conditions Michels' claim holds true. In what follows, I will first review Michels' original argument about the development of oligarchy and its consequences, then outline the impact of the iron law on the study of political parties, social movements and other voluntary associations, and workers' collectives.

Michels' original argument can be distilled into three basic claims:

1. *Bureaucracy happens*. Large-scale organizations in modern society will tend to develop a rational-bureaucratic structure (i.e., a system of procedural formality, jurisdictional specialization, and hierarchical authority). Bureaucracy arises due to a structural imperative that affects every organization, but increases as the organization grows: effectiveness requires administrative efficiency, which in turn requires a complex and hierarchical division of labor.
2. *If bureaucracy happens, power rises*. The hierarchical delegation of authority inherent in a rational-bureaucratic structure necessarily concentrates power in the hands of a minority. As this minority becomes intimately familiar with the day-to-day operations of the organization, it develops into a professional leadership with a monopoly of skills, knowledge, and resources. This state of affairs, aided by the supposed 'incompetence of the masses,' makes the leadership indispensable, as the rank and file come to rely on them for their expertise.
3. *If power rises, power corrupts*. Once power is concentrated in the hands of a minority, the leadership develops an interest in maintaining the status and perquisites of its position. Driven by a supposedly inherent 'desire to dominate,' the minority, having once attained this position, will do whatever is necessary to preserve its own power – even when that means going against the interests of the membership – and will use undemocratic means to stifle any internal opposition which it interprets as a threat to its authority. The organization becomes an end in itself, and as a result, the leadership becomes susceptible to co-optation and to the displacement of the membership's original goals for the sake of organizational maintenance.

Critics have debated and challenged Michels on each of these three claims. The earliest and least frequent line of attack came from an orthodox Marxist position that challenged the second step of Michels' argument, the claim that bureaucracy necessarily concentrates power. Scholars taking this line (e.g., Bukharin, 1925; Hook, 1933) argue that because power is based in the ownership of property, if bureaucrats exercised authority without owning the means of production (as they claimed is the case in a socialist state), the bureaucratic leadership would have no real power to either defend or abuse, and thus could not be considered oligarchic. This line of argument has been soundly critiqued by scholars of power who hold that people who *control* resources – i.e., who have administrative authority – can have at least as much impact on people's lives by monopolizing decision-making processes as those who *own* resources (see e.g., Dahrendorf, 1959). In other words, what makes a group oligarchic is their power to control the processes by which decisions are made and implemented, regardless of what allows them to wield that power.

By the 1950s and 1960s, a second set of arguments began to appear, this time focusing on the third step in Michels' argument. Taking a pluralist perspective, these critics accepted that bureaucracy happens and that in a bureaucracy, power rises, but they rejected the assertion that power must necessarily corrupt and posited various ways in which a powerful minority might be held accountable to the majority it is meant to represent (Edelstein, 1967; Gouldner, 1965; Lipset et al., 1956; Schumpeter, 1950). Others concerned with accountability have focused on the process of goal displacement, i.e., the tendency of entrenched leaderships to abandon the membership's original goals in favor of those that ensure organizational survival and growth (Barnes, 1987; Gusfield, 1955; Merton, 1957; Selznick, 1949; Sills, 1957).

In the 1980s, a new critique emerged that challenged the first step of Michels' argument – the underlying premise that organizations must necessarily develop a bureaucratic structure in order to be effective (Breines, 1980; Rothschild and Whitt, 1986). Even though his primary referent, the German Social Democratic Party, was in fact a large-scale bureaucratic organization, Michels explicitly stated that he was talking about *all* types of nominally democratic organizations: "The formation of oligarchies within *various forms of democracy* is the outcome of organic necessity, and consequently affects *every* organization …" (Michels, 1962: p. 366, emphasis added). And while Michels assumed that bureaucracy was the only sustainable form of organization, he nevertheless acknowledged that *if* bureaucracy could be avoided, then theoretically, so could oligarchy. Prior to the 1970s, however, empirical research had been restricted to large-scale, bureaucratically structured membership organizations. Critics felt that the wave of workers' collectives, communes, cooperatives, and political groups born of the New Left's commitment to participatory democracy could pose a serious challenge to the iron law, yet because nonbureaucratic organizations were being perceived as *dis*organized rather than *differently* organized, these potentially exceptional cases were not being acknowledged.

In any case, more attention is being paid to the applicability of the iron law in nonbureaucratic settings, in part because these groups clearly struggle with the problem of oligarchy, just as bureaucratic organizations do, but also because, contrary to Weber's prediction, nonbureaucratic organizations have become an increasingly prevalent feature of the modern institutional landscape. These structures are widely used in social movements, workers' cooperatives, protest/cultural encampments (e.g., Burning Man, Occupy, the Brazilian Landless Movement), communes, neighborhood associations, Internet campaigns, and forums for local self-governance. The impulse is even apparent in large corporations that have begun to adopt less hierarchical, more participatory team-based structures (Polletta, 2014; Rothschild, 2000).

Having outlined the most common avenues by which critics have evaluated the iron law, we now turn to a brief substantive review of their findings.

Conditions Affecting Oligarchy and/or Goal Displacement

Building on categorizations laid out by Zald and Ash (1966) and Lipset et al. (1956), I will organize the literature addressing Michels' thesis by the *types* of variables that have been found to contribute to the rise of oligarchy or its absence. I will incorporate findings from studies done in the whole range of organizational settings, including social movement organizations, unions, political parties, voluntary associations, local governing bodies, workers' cooperatives, and for-profit businesses. Reviews of the literature usually list findings separately for each of these settings, because they correspond to the disciplinary subfields that have been influenced the most by Michels' thesis (e.g., Ansell, 2001). While that approach is useful for those looking to focus on a particular subfield, however, one of the most important things we have learned from a century of research on oligarchy is that the *purpose* of the organization matters far less than characteristics related to how it is structured, the nature of its membership, and the environmental context in which it operates – factors that may be shared by a host of organizations with dissimilar goals and purposes. I highlight these factors in order to provide a better overall sense of where we stand in the study of oligarchy and what we have left to accomplish.

Forms of Authority: Bureaucracy, Collectivist Democracy, and Noncharismatic Personalism

Michels assumed, following his friend and mentor Max Weber, that traditional and charismatic forms of authority would increasingly be supplanted by rational-legal authority in modern society, and that rational-legal authority, expressed in bureaucracy, would give rise to oligarchy. However, two other forms of authority – neither of which were included in Weber's typology – have been hypothesized to affect the development of oligarchy. For example, in their study of three ideologically diverse political parties in Germany, France, and Russia, Ansell and Fish (1999) found that the basis of a leadership's indispensability depends not on their bureaucratic control of resources and information, but on the form of authority that prevails in the party, which in turn depends upon the types of cleavage structures that are most salient. In parties where both territorial and ideological cleavages are highly salient, they argued that a "non-charismatic personalist" form of authority develops, and the leadership's indispensability is based on its ability to resolve conflicts between factions. Such parties did develop an oligarchy, but via a different pathway than Michels laid out.

As mentioned above, another form of authority, collectivist democracy, may have the potential to prevent oligarchy altogether. Rothschild and Whitt examined this possibility in their study of five workers' cooperatives in the 1970s. One of their key findings was that having a "provisional orientation" which they define as the "sentiment that all operations and programs in an organization ought to be tentative" (1986: p. 83), militates against oligarchization. At the other extreme are those who hold that the lack of formal hierarchy makes collectivist organizations *more*, not less, susceptible to oligarchic control (Freeman, 1972; Staggenborg, 1988). Jo Freeman, for example, argued that the lack of formal structure and routinized procedures in collectivist organizations simply masks the rule of *informal* oligarchs, producing a 'tyranny of structurelessness.' Most scholars, however, have found that a collectivist structure is neither immune to oligarchy nor especially susceptible to it. In fact, research on organizations of all types suggests that while the vast majority do eventually succumb to oligarchy, whether, how, and how quickly it develops are contingent on a range of structural, environmental, and cultural factors.

Structural Factors

Certain characteristics of an organization's structure and practices have been found to affect the development of oligarchic rule. At least four such factors are likely to affect both formal and informal organizations. For one, scholars have generally supported Michels' claim that larger size makes oligarchy more likely (Johnson and Whyte, 1982; Rothschild and Whitt, 1986),

although some studies have found either no correlation or a negative correlation with size (Getz, 1998; Rosner, 1998). A second factor that could affect all organizations is professionalization. As with size, there is disagreement on the effect of this factor, with some supporting Michels' prediction that it facilitates oligarchization (Piven and Cloward, 1977; Zald and Ash, 1966) and others (e.g., Staggenborg, 1988) arguing that formal organizations with professional leaderships do not have to become oligarchic – that the routinized procedures make it harder for individuals to gain disproportionate power. Third, some kind of mechanism that assures the broad diffusion of knowledge, information, and technological skills among the membership is necessary for maximal participation and to keep elites from forming and/or becoming oligarchic (Katovich et al., 1981; Lipset et al., 1956; Rothschild and Whitt, 1986). Lastly, there needs to be some kind of structured mechanism to protect people from reprisals when they dissent or challenge the leadership (Bernstein, 1982; Lipset et al., 1956).

Other structural factors have been posited as specifically relevant in either bureaucratic or collectivist settings. In bureaucratic organizations, for example, the presence of factions and subgroups that vie for leadership helps to assure leadership accountability and/or turnover (Lipset et al., 1956; Zald and Ash, 1966). Similarly, Baiocchi (2005) found in his study of the participatory budgeting forums in Porto Alegre that in districts where popular councils and activist networks existed outside of the official forums, the process engendered more trust and participation and was less prone to debilitating conflict. Research on collectivist organizations has generated less consensus about which structured practices (stable norms) help groups to prevent/contain informal domination. Freeman (1972) specifically argues that a more structured division of labor (including a more fixed committee structure, the rotation of functionally important tasks, and a broad diffusion of information among members) as well as a more structured decision-making process are both necessary to keep informal elites from dominating the group. There is broad agreement about access to information and the rotation of tasks, but several scholars have found that a decision-making process that is too structured alienates the uninitiated and generates an in-group that is able to manipulate the process in its favor (Juris, 2008; Leach, 2013).

Organizational Environment

External factors having to do with the institutional context within which the organization operates can have an enormous effect on its ability to maintain a democratic distribution of power. A host of studies on social movement organizations have noted the impact of funding sources in this regard, as elite beneficiaries often make their contributions contingent upon the use of more conventional tactics, more formal organizational structures, and/or the ability to demonstrate high levels of operational efficiency, all of which push the organization in the direction of professionalization, oligarchization, and conservatism (Clemens, 1993; Ferree and Martin, 1995; Piven and Cloward, 1977; Rothschild and Whitt, 1986; Staggenborg, 1988; Voss and Sherman, 2000). The Mondragon complex in Spain, with over 160 workers' cooperatives and 23 000 members, successfully addressed this problem by creating its own banking system to fund new start-ups (Johnson and Whyte, 1982). Other scholars point to inter-organizational relationships and/or the organization's position in a broader organizational network as key, whereby the organization must either have a high degree of autonomy or be densely networked with other democratically inclined organizations to improve its chances of remaining democratic (Fisher, 1994; Lipset et al., 1956; Zald and Ash, 1966).

For political parties, electoral procedures, the conventions and regulations related to the governing process, and the demographics of potential members/voters have all been found to constrain a party's ability to remain internally democratic (Markovits and Gorski, 1993; Swerdlow, 1973; Zald and Ash, 1966). Similarly, research on workers' cooperatives has found that economic competition and the nature of markets can force organizations to sacrifice more democratic practices for the sake of efficiency, opening the door for oligarchy (Getz, 1998; Rosner, 1998). Breines' (1980) analysis of the collapse of Students for a Democratic Society – the leading organization of the student movement in the 1960s – argued that government repression was a major factor that fed the polarization, radicalization, and turn to militancy in the group, ultimately contributing to their rejection of participatory democracy and allowing an oligarchic minority to seize control of the organization.

Lastly, two recent studies of political parties present models combining external structural factors and internal cultural variables to explain the dynamics of power. In his study of left-libertarian parties, Kitschelt (1989) portrays the Michelsian dilemma as a conflict between two party logics – a logic of constituency representation, which works against oligarchy, and a logic of electoral competition, which hastens it. Which of these logics dominates depends on the relative strength of subgroups he labels ideologues, lobbyists, and pragmatists. This, in turn, is determined by the 'political opportunity structure,' which includes such external factors as the degree to which the libertarian left is mobilized, how existing elites respond to the entry of a new party into the political system, the competitive position of the party in relation to others, and the number of active left-libertarian movements. Rohrschneider (1994) similarly points to changing cleavage structures in the European party system – specifically, the emergence of the 'New Politics' orientation – to argue that New Politics proponents in the 52 established parties he studied were more likely to side with local constituencies than with leadership on contested party decisions. He argued that this was due to their socialist convictions, not party structure, and that the presence of this ideology undermines Michels' two classist psychological assumptions (that the masses are incompetent and that leaders will always have a will to dominate), thus potentially weakening Michels' argument.

Cultural Factors

Beyond these structural features, some scholars have pointed to aspects of organizational *culture* that may be necessary to avoid oligarchy. Such factors have been especially prevalent in research on collectivist organizations. For example, particular kinds of ideology have been found to enhance the effectiveness and longevity of collectivist groups, such as an ethic of

cooperation and egalitarianism (Iannello, 1992; Leviatan, 1998), an antinuclear ideology (Downey, 1986), religious ideologies such as Quakerism, Christian pacifism, or paganism (Epstein, 1991; Leach, 1998), and anarchist ideology (Day, 2005; Epstein, 1991, 2001; Graeber, 2009; Lyons, 1988). And, of course, feminism has been one of the driving forces behind experimentation with participatory forms of organization (e.g., Ferguson, 1984; Ferree and Martin, 1995; Harris and King, 1989). Other cultural factors found to contribute to the longevity of democratically inclined organizations more generally are an organizational culture that encourages dissent and conflict (Epstein, 1991; Rothschild and Leach, 2006), a homogeneous membership (Johnson and Whyte, 1982; Lipset et al., 1956; Rothschild and Whitt, 1986), status differences within the group (Lipset et al., 1956; Polletta, 2002), a culture of mutual self-criticism (Rothschild and Whitt, 1986), the social status of the members in the larger society (Clemens, 1993; Lipset et al., 1956), and the development of new forms of association within the larger decision-making units, such as affinity groups, consciousness-raising groups, pacifist 'disciplines,' or residential clusters (Polletta, 2002). Finally, some have suggested that groups with an instrumental logic of action are more susceptible to oligarchy than more expressive, value-rational groups (Koopmans, 1995; Kriesi, 1996), although oligarchy has certainly been found in expressive groups as well (Schmidt, 1973).

Methodological Issues and Emerging Directions

As is evident from the discussion so far, there is still much disputed terrain in the debate over the iron law. While it has provoked research for over a century now, questions about whether oligarchy is truly inevitable and the conditions under which it might be avoided remain unresolved. One major reason for this is that there has been no agreed-upon operational definition of what constitutes oligarchy across organizational forms (Leach, 2005). Different measures have been used from one context to the next – often without the choice being explicitly stated or defended. As Schmidt noted in 1973, "Since Plato and Aristotle, most writers who discuss oligarchy fail to define the concept, apparently because they assume the word is understood in the light of its Greek etymology (the rule of a few). Michels also fails to offer a formal definition …" (1973: p. 10).

Etymologically, oligarchy means 'the rule of the few.' The simplicity of this definition, however, belies the host of complex issues that arise in any attempt to empirically assess whether or not an oligarchy exists. For example, what, exactly, do we mean by 'rule'? What must 'the few' do in order to be considered oligarchic? Must they hold formal office or can they work behind the scenes? Is it only oligarchic when a minority is coercive and actively suppresses dissent or is a persuasive minority that gets people to go along with them out of respect for their competence or charismatic appeal also oligarchic? Is there a difference between 'ruling' and 'leading'? If so, is it the exercise of 'power' as opposed to 'influence' that differentiates the two? Or are power and influence both commodities that may be employed by democrats and oligarchs alike? And lastly, is it simply having a certain *amount* of power and/or influence that makes a minority oligarchic (in which case, how much is necessary and how can it be measured?), or does it depend on how the minority *uses* its power/influence that determines whether they have crossed the threshold into oligarchy?

Though definitions of oligarchy are often left implicit, two general tendencies are discernible in the literature. The most common approach has been to treat oligarchy as a feature of the formal authority structure (Lipset, 1952; Lipset et al., 1956; Michels, 1962; Schmidt, 1973; Stone, 1972). Lipset, Trow, and Coleman, for example, characterize oligarchy as a one-party political system where "one group, which controls the administration, usually retains power indefinitely, rarely faces organized opposition, and when faced with such opposition often resorts to undemocratic procedures to eliminate it" (1956: p. 3). This implies that the presence of organized opposition in a two-party system is a sufficient safeguard against oligarchic tendencies.

The problem with this definition is twofold: first, that it ignores the possibility of an oligarchy behind the scenes wielding *informal* power – that is, that a powerful economic or cultural elite could wield oligarchic influence without holding political office (Domhoff, 2014; Mills, 1956); and secondly, that it only works in bureaucratic organizations. If a minority has to have formal authority to be oligarchic, that would mean that collectivist organizations are *never* oligarchic, since no one has formal authority over anyone else. But such a claim flies in the face of a host of ethnographic research, not to mention common sense.

A second tendency has been to effectively equate oligarchy with goal displacement. While it is of course interesting to ask under what conditions organizations that start out with radical goals become more conservative over time – which often occurs as a consequence of oligarchization – asking why an organization changes its goals is not the same thing as asking how many people are ruling it. One cannot assume that just because an organization still has radical goals that it is therefore democratic; nor does a change in goals necessarily mean that an oligarchy imposed the decision. Oligarchies are not always conservative and, more importantly, radical organizations are not always democratic – which was precisely Michels' point about the Social Democratic Party. The direction of goal displacement, when it happens as a result of oligarchy, is generally dictated by the need for external support. Michels' argument was that where external support (in the form of votes or memberships or funding) is necessary for an oligarchic leadership to maintain its power, it will shift the organization's goals in whatever direction is most likely to attract that support.

Confusing bureaucratic conservatism with oligarchy and thereby equating radicalism with democracy is also problematic because it elides the fundamental concern underlying Michels' analysis: the question of how effective parliamentary strategies can be for realizing revolutionary goals. Michels' study of the German Social Democratic Party was largely a critique of the vanguardist strategy embraced by the parliamentary socialists of his day, who, in claiming to represent the emancipatory interests of 'the masses,' felt justified in repressing dissent within their own ranks. Debate over this question has been an especially prominent feature of leftist politics over the last half-century, underlying the New Left's embrace of participatory democracy and the extra-parliamentary stance of

many of the so-called New Social Movements in the postwar era (Breines, 1980, 1982; Epstein, 1991; Katsiaficas, 1997), and playing a central role in the factionalization of the German Green Party (Markovits and Gorski, 1993). If we imagine that all it takes to avoid oligarchy is to remain committed to a radical agenda, no matter how undemocratic the means used by 'revolutionary' leadership might be, we bypass the question at the very heart of Michels' work and indeed, at the very heart of much of activist history since then.

Instead of reducing oligarchy to a particular kind of formal authority structure or conflating it with a conservative policy orientation, it is more accurate to see it as a particular *distribution of power*, which may take the form of either formal authority or informal influence, and which may have a number of bases (e.g., economic, social, informational, or political). Elsewhere, I have proposed a formal definition of oligarchy as "a concentration of entrenched illegitimate authority and/or influence, such that *de facto* what that minority wants is generally what comes to pass, even when it goes against the wishes (whether actively or passively expressed) of the majority" (Leach, 2005: p. 329). This definition focuses on the decision-making process as the focal point for the contestation of power, and requires that, at least in decisions that it perceives could threaten its control, a minority must successfully impose its will through the illegitimate exercise of power, and against resistance, for the organization to qualify as oligarchic. It also notably leaves open the possibility that *any* organization – whether bureaucratic or nonbureaucratic – could potentially either succumb to oligarchy or resist it. In short, it is meant to provide a common measuring stick by which all three of Michels' claims (that bureaucracy is necessary, that it necessarily concentrates power, and that concentrated power necessarily corrupts) can be subjected to empirical disputation, regardless of the organizational setting. Recent studies have begun to use this definitional approach (Diefenbach, 2013; Toplak, 2011). It remains to be seen what progress it may facilitate.

Whatever approach we take, however, it is clear that Michels' groundbreaking work will continue to generate insights into the dynamics of power in organizations of all kinds. But the reach of his legacy goes far beyond the field of organizational studies. In fact, answering the questions Michels raises is key to unraveling perennial issues about larger processes of social change, such as whether grassroots and culturalist strategies for social change can be effective, whether working within the context of large-scale bureaucratic institutions can lead to lasting emancipatory change or if that would require the use of nonhierarchical processes, and whether we should continue to have faith in Tocqueville's argument about voluntary associations in civil society being vehicles for empowerment and democratization. One thing is clear: if, as advocates of participatory democracy have long held, grassroots democratic organizations are to be the breeding ground for an informed and active citizenry in a democratic society, then we will have to get better at breaking the iron law of oligarchy.

See also: Administration in Organizations; Authority, Social Theories of; Bureaucracy and Bureaucratization; Bureaucracy, Sociology of; Bureaucratization and Bureaucracy, History of; Collective Action; Collective Behavior, Social Psychology of; Collective Behavior, Sociology of; Cooperation: Sociological Aspects; Delegation of Power: Agency Theory; Deliberation and Democracy; Democracy, History of; Democratic Theory; Green Parties; Group Processes in Organizations; Group Processes, Social Psychology of; Labor Unions; Leadership in Organizations, Sociology of; Leadership; Movements, Social; Organizational Behavior, Psychology of; Organizational Decision Making; Organizations, Sociology of; Organizations: Authority and Power; Organizations: Unintended Consequences; Political Parties, History of; Political Parties: Communist; Political Sociology; Power; Social Influence, Sociology of; Voluntary Associations, Sociology of; Voluntary Organizations.

Bibliography

Ansell, Christopher K., 2001. Oligarchy (Iron law). In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of Social and Behavioral Sciences. Elsevier Press, Amsterdam.

Ansell, Christopher K., Fish, Steven, 1999. The art of being indispensable: noncharismatic personalism in contemporary political parties. Comparative Political Studies 32 (3), 283–312.

Baiocchi, Gianpaolo, 2005. Militants and Citizens: The Politics of Participatory Democracy in Porto Alegre. Stanford University Press, Stanford, CA.

Barnes, Donna A., 1987. Organization and radical protest: an antithesis. The Sociological Quarterly 28 (4), 575–594.

Bernstein, Paul, 1982. Necessary elements for effective worker participation in decision making. In: Lindenfeld, F., Rothschild-Whitt, J. (Eds.), Workplace Democracy and Social Change. Porter Sargent Publishers, Boston, pp. 51–81.

Breines, Wini, 1980. Community and organization: the new left and Michels' "iron law". Social Problems 27 (4), 419–429.

Breines, Wini, 1982. Community and Organization in the New Left: 1962–1968. Praeger Publishers, New York.

Bukharin, Nikolai, 1925. Historical Materialism. A System of Sociology. International Publishers, New York.

Clemens, Elisabeth S., 1993. Organizational repertoires and institutional change: women's groups and the transformation of U.S. Politics, 1890–1920. American Journal of Sociology 98 (4), 755–798.

Dahrendorf, Ralf, 1959. Class and Class Conflict in Industrial Society. Routledge & Kegan Paul, London.

Day, Richard J.F., 2005. Gramsci Is Dead: Anarchist Currents in the Newest Social Movements. Pluto Press, Ann Arbor, MI.

Diefenbach, Thomas, 2013. Hierarchy and Organisation: Toward a General Theory of Hierarchical Social Systems. Routledge, New York.

Domhoff, G. William, 2014. Who Rules America? The Triumph of the Corporate Rich. McGraw-Hill Education, New York.

Downey, Gary L., 1986. Ideology and the clamshell identity: oganizational dilemmas in the anti-nuclear power movement. Social Problems 33 (5), 357–373.

Edelstein, J. David, 1967. An organizational theory of union democracy. American Sociological Review 32, 19–31.

Epstein, Barbara, 2001. Anarchism and the anti-globalization movement. Monthly Review 53 (4), 1–14.

Epstein, Barbara, 1991. Political Protest and Cultural Revolution: Nonviolent Direct Action in the 1970s and 1980s. University of California Press, Berkeley.

Ferguson, Kathy, 1984. The Feminist Case against Bureaucracy. Temple University Press, Philadelphia.

Ferree, Myra M., Martin, Patricia Y., 1995. Feminist Organizations: Harvest of the New Women's Movement. Temple University Press, Philadelphia.

Fisher, Julie, 1994. Is the iron law of oligarchy rusting away in the Third World? World Development 22 (2), 129–143.

Freeman, Jo, 1972. The tyranny of structurelessness. Berkeley Journal of Sociology 17, 151–164.

Getz, Shlomo, 1998. The winds of change. In: Leviatan, U., Oliver, H., Quarter, J. (Eds.), Crisis in the Israeli Kibbutz: Meeting the Challenge of Changing Times. Praeger, Westport, CT, pp. 13–26.

Gouldner, Alvin W., 1965. Studies in Leadership and Democratic Action. Russell & Russell, New York.
Graeber, David, 2009. Direct Action: An Ethnography. AK Press, Oakland.
Gusfield, Joseph R., 1955. Social structure and moral reform: a study of women's Christian Temperance Union. American Journal of Sociology 61, 211–232.
Harris, Adrienne, King, Ynestra, 1989. Rocking the Ship of State: Toward a Feminist Peace Politics. Westview Press, Boulder, CO.
Hook, Sidney, 1933. Towards the Understanding of Karl Marx. John Day Co, New York.
Iannello, Kathleen P., 1992. Decisions without Hierarchy: Feminist Interventions in Organization Theory and Practice. Routledge, New York.
Johnson, Ana G., Whyte, William F., 1982. The Mondragon system of worker production cooperatives. In: Lindenfeld, F., Rothschild-Whitt, J. (Eds.), Workplace Democracy and Social Change. Porter Sargent Publishers, Boston.
Juris, Jeffrey S., 2008. Networking Futures: The Movements against Corporate Globalization. Duke University Press, Durham, NC.
Katovich, Michael, Weiland, Marion W., Couch, Carl J., 1981. Access to information and internal structures of partisan groups: some notes on the iron law of oligarchy. The Sociological Quarterly 22 (3), 431–445.
Katsiaficas, George N., 1997. The Subversion of Politics: European Autonomous Social Movements and the Decolonization of Everyday Life. Humanities Press, Atlantic Highlands, NJ.
Kitschelt, Herbert, 1989. The Logics of Party Formation: Ecological Politics in Belgium and West Germany. Cornell University Press, Ithaca, NY.
Koopmans, Ruud, 1995. Democracy from Below: New Social Movements and the Political System in West Germany. Westview Press, Boulder, CO.
Kriesi, Hanspeter, 1996. The organizational structure of new social movements in a political context. In: McAdam, D., McCarthy, J.D., Zald, M.N. (Eds.), Comparative Perspectives on Social Movements: Political Opportunities, Mobilizing Structures, and Cultural Framings. Cambridge University Press, Cambridge.
Leach, Darcy K., 1998. Why Just Go for 51%?: Organizational Structure in the Religious Society of Friends. CRSO Working Paper Series. Center for Research on Social Organization, University of Michigan, Ann Arbor.
Leach, Darcy K., 2005. The iron law of what again? Conceptualizing oligarchy across organizational forms. Sociological Theory 23 (3), 312–337.
Leach, Darcy K., 2013. Culture and the structure of tyrannylessness. The Sociological Quarterly 54 (2), 181–191.
Leviatan, Uriel, 1998. Second and third generations in kibbutzim – is the survival of kibbutz society threatened? In: Leviatan, U., Oliver, H., Quarter, J. (Eds.), Crisis in the Israeli Kibbutz: Meeting the Challenge of Changing Times. Praeger, Westport, CT, pp. 81–96.
Lipset, Seymour M., 1952. Democracy in private government (a case study of the International Typographical Union). British Journal of Sociology 3 (1), 47–63.
Lipset, Seymour Martin, Martin Trow, James Coleman, 1956. Union Democracy: The Internal Politics of the International Typographical Union. Free Press, Glencoe, IL.
Lyons, Matthew, 1988. The 'Grassroots' Network: Radical Nonviolence in the Federal Republic of Germany 1972–1985. Center for International Studies, Cornell University, Ithaca, NY.
Lyons, Matthew, 1999. The grassroots movement in Germany, 1972–1985. In: Zunes, S., Kurtz, L.R., Asher, S.B. (Eds.), Nonviolent Social Movements: A Geographical Perspective. Blackwell Publishers, Malden, pp. 81–95.
Markovits, Andrei S., Gorski, Philip S., 1993. The German Left: Red Green and Beyond. Oxford University Press, New York.
Merton, Robert, 1957. Social Theory and Social Structure. Free Press, Glencoe, IL.
Michels, Robert, 1962. Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracy. Free Press, New York.
Mills, C. Wright, 1956. The Power Elite. Oxford University Press, New York.
Piven, Frances F., Cloward, Richard A., 1977. Poor People's Movements: Why They Succeed and How They Fail. Pantheon Books, New York.
Polletta, Francesca, 2002. Freedom Is an Endless Meeting: Democracy in American Social Movements. University of Chicago Press, Chicago.
Polletta, Francesca, 2014. Participatory democracy in the new millenium. Contemporary Sociology 42 (1), 40–49.
Rohrschneider, Robert, 1994. How iron is the iron law of oligarchy? Robert Michels and national party delegates in eleven West European democracies. European Journal of Political Research 25 (2), 207–238.
Rosner, Menachem, 1998. Work in the Kibbutz. In: Leviatan, U., Oliver, H., Quarter, J. (Eds.), Crisis in the Israeli Kibbutz: Meeting the Challenge of Changing Times. Praeger, Westport, CT, pp. 27–40.
Rothschild, Joyce, 2000. Creating a just and democratic workplace: more engagement, less hierarchy. Contemporary Sociology 29 (1, Utopian Visions: Engaged Sociologies for the 21st Century), 195–213.
Rothschild, Joyce, Leach, Darcy K., 2006. Avoid, talk, or fight: cultural responses to oligarchic pressure in collectivist-democratic groups. In: Cnaan, R., Milofsky, C. (Eds.), Handbook on Community Movements and Local Organizations. Kluwer Books, New York.
Rothschild, Joyce, Whitt, J. Allen, 1986. The Cooperative Workplace: Potentials and Dilemmas of Organizational Democracy and Participation. Cambridge University Press, New York.
Schmidt, Alvin J., 1973. Oligarchy in Fraternal Organizations; a Study in Organizational Leadership. Gale Research Co, Detroit, MI.
Schumpeter, Joseph, 1950. Capitalism, Socialism, and Democracy. Harper & Bros, New York.
Selznick, Philip, 1949. TVA and the Grass Roots. University of California Press, Berkeley.
Sills, David, 1957. The Volunteers. Free Press, Glencoe, IL.
Staggenborg, Suzanne, 1988. The consequences of professionalization and formalization in the pro-choice movement. American Sociological Review 53 (4), 585–605.
Stone, Dale, 1972. SDS and the "iron law of oligarchy". The Kansas Journal of Sociology 8 (1), 59–64.
Swerdlow, Marian, 1973. Breaking the iron law: a critique of Michels' theory of political parties in the light of the history of the German Social Democratic Party. The Human Factor 12 (1), 1–14.
Toplak, Cirila, 2011. Hybridization of democracy in Central and Eastern Europe. Journal of Comparative Politics 4 (1), 76–90.
Voss, Kim, Sherman, Rachel, 2000. Breaking the iron law of oligarchy: union revitalization in the American labor movement. American Journal of Sociology 106 (2), 303–349.
Zald, Mayer N., Ash, Roberta, 1966. Social movement organizations: growth, decay and change. Social Forces 44 (3), 327–341.

Olson, Mancur (1932–98)

Martin C McGuire, University of California-Irvine, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

One of the founding figures in the economic discipline of 'Public Choice,' Mancur Olson developed the logic of how a group provides itself with nonrival, nonexcludable, public goods. Because individual rationality leads every member of a group to free riding behavior, the result of individual rationality is not the group rational outcome. His early writings established the nature of individually rational voluntary provision group equilibrium, described its characteristics, and identified its sweeping scope.

These early insights led to deepened understanding of the role of incentives in political organization and choice. Thus, a major impact of his work was to bring economic logic to bear on the analysis of political behavior, and this led to widespread utilization of economic methods in political science and other studies of nonmarket choice. Examples include interactions among countries in environmental, political, military, or economic alliances, and formulation of important concepts for understanding the success/failure of societies in governing themselves and in achieving economic prosperity.

Such concepts include (1) 'distributional coalitions' that fashion transfers to themselves from the society at large while degrading the economic performance of the whole; (2) a political group's 'encompassing interest' or stake in the overall economic health of its society, which limits the economic inefficiency it causes; or (3) 'demasclerosis,' the opportunity that social stability affords for the buildup of inefficiency causing special interests.

Biography

The son of Norwegian immigrants to the United States, Mancur Olson was born in Grand Forks, North Dakota, USA, on 22 January 1932, and raised near there on a farm. He died on 19 February 1998 at College Park, Maryland, USA. He studied economics at North Dakota State University, Oxford University, and Harvard University. He spent the greatest fraction of his career as an economics professor at the University of Maryland, USA, but before this, he undertook teaching assignments at the US Air Force Academy and Princeton University, and service in the US Department of Health Education and Welfare. Influential in professional associations, he had been President of the US Southern Economic Association, US Eastern Economic Association, Public Choice Society, and Vice President of the American Economic Association. He was a Fellow of the American Academy of Arts and Sciences, an Honorary Fellow of University College, Oxford, and recipient of many other honors and awards. At the time of his death, he was principal investigator and director of a US Government-sponsored development aid/research project with offices in 30 countries, and an ambitious agenda to bring the benefits of his life's work to the design and reform of the institutions of governance throughout the developing and former communist countries of the world.

Antecedents to Public/Collective Choice

One of the founders of Public or Collective Choice as a subdiscipline and movement within economics, Olson devoted his intellectual life to the analysis of collective action. Early, as a student he intuited that the then standard paradigm as to how people interacted economically with each other was seriously incomplete. In the traditional model, it was argued explicitly or implicitly that economic interaction while widespread and diverse, was normally mediated by prices and markets between otherwise isolated and independent agents. The first founding principle of 'Public Choice' and a concern that permeates Olson's life work is recognition that people exchange, produce, consume, and share many goods, in many environments that are not at all mediated by markets and prices, yet the behavior and analysis thereof is unquestionably 'economic.' Typically, in these environments, people belong to identifiable groups and their interaction partakes of their group identity in a manner not conveyed by Adam Smith's butcher or baker. This raises to paramount importance questions of how people and organizations that belong to a group do and should act to achieve their goals.

Oligopoly

Central to the Public Choice nexus is how the size of a group influences the behavior of its individual members and thus aggregate outcome. Other features of group membership such as homogeneity or diversity are also important but size was the first aspect to seize Olson's attention. With respect to industrial behavior, the question of how small numbers of firms would behave had occupied economists' attention for a generation by the time Olson became a student. Still earlier, the pricing, output, employment, and other profit-maximizing behaviors of a single, monopolistic, firm had been successfully analyzed as had those of a market consisting of a large number of atomistic competitive firms (Marshall, 1890). And by the time Olson arrived at Harvard in 1957, for the case of small numbers of firms in an industry, theories of oligopoly interaction (Cournot, 1897; Robinson, 1933; Chamberlin, 1933) had emerged, which demonstrated that in pursuing rational profit maximization for themselves individual suppliers would fail to achieve maximum profits for their industry – their group – as a whole. The reason for this is a lack of coordination: oligopolists fail to anticipate and include in their decision calculus the effects their narrow profit-making decisions have on the profits of other firms. In fact, it would be individually irrational for any single firm to include such external benefits in its

decision calculus since the individual firm could gain more by free riding off the efforts of other firms. A similar coordination failure motivated the fast developing field of Game Theory, which provided a rigorous framework for analysis of oligopolistic behavior. Actually Olson's mentor at Harvard (E. Chamberlin) had been one of the leaders in the evolving understanding of oligopoly with his introduction of the idea of 'monopolistic competition.' Here, individual firms in an industry resemble monopolists in that they recognize their supply decisions will influence the price they can obtain. At the same time, their industry is competitive, in a sense, because the price actually charged by any one firm influences the price–supply opportunities of every other firm.

Public Goods

Parallel with economists' study of producer oligopoly, discoveries preceding Olson showed that provision of some types of goods to consumers could lead to similar failures of coordination (Samuelson, 1954). 'Public goods' or 'collective goods' are such that once available to any member of a specified group, the same quantity and quality is equally available to every other member at no extra cost: this is because such goods are not used up when enjoyed and if available to any consumer cannot be excluded from any other group member. Early examples included the lighthouse beacon, which warns all ships (nonexcludable), and whose utilization by one ship in no way reduces its availability to other ships (nonrival). More contemporary examples include protection of pollution-free environments, provision of national defense, preservation of natural species, establishment of political freedoms, provision of property right guaranteed by an impartial and universal police/judiciary, and maintenance of government by rule law. These all should be understood to be economic 'goods' broadly speaking and all display the property that their enjoyment by one person in no way diminishes their availability to others to enjoy. But if every member of a group enjoys a good irrespective of who supplies it, no one acting voluntarily will provide the good – at least no one will provide a sufficient amount of it. Instead, all will have an interest in seeing others pay the supply costs while they enjoy the benefit so produced without making any cost contribution themselves. Every individual, that is, has good reason to free ride – in whole or in part – on the provisions of others. One, therefore, should anticipate that for the provision of such public goods, free markets based on individual rationality will fail. This means that the amount or quality of collective good provision to follow from individual rational behavior will fall short of what the aggregate of individuals in a group desire. In short, notwithstanding the fact that a group is nothing more than the collection of individuals who belong to it, individual rationality, even assuming it to be universal, will not secure the rational group outcome.

The Voluntary Provision Model and Olson's *Logic of Collective Action*

The initial reaction of economics to this understanding was to conclude that instead of individual action, society must rely on government power to enforce individual contribution via coercive taxation/expenditure. Olson's first insights came from taking the next step: because government itself is subject to the quandary that the free rider problem creates and, therefore, may not produce an optimum outcome, it cannot be assumed that government will necessarily or ordinarily close the gap between individual and group rationality. Olson, therefore, asked what to expect of individual voluntary behavior in the absence of governmental or coercive intervention. Such is the topic of *The Logic of Collective Action* (1965), Olson's Ph.D. thesis and first great contribution. Once the question was thus posed, it became clear that the range of circumstances where collective action obtained was far more vast, complex, and richer than the early oligopoly paradigm might lead one to expect. Moreover, contrary to simplistic theory, sometimes arrangements arise (in the form of intragroup monitoring or selective extra rewards to those who do not free ride) that allow the free rider failure to be reduced and possibly altogether circumvented. Olson's discovery and elaboration of the 'voluntary provision' model of public goods changed, in fact redefined, the topic for economists; it introduced public goods to other social sciences, especially political science and sociology and elevated it to a unifying principle throughout the social sciences. The result was an explosion of research into the structure of voluntary provision, and into the myriad of applications of the model, ranging from strictly economic to anthropological, legal, sociological, and political, from labor union behavior, to allocations within families, voluntary philanthropic giving for a group or social goal, and from public interest lobbying, special private interest favor seeking and professional associations to the behavior of nations as they formed alliances to combat common adversaries, or environmental challenges like ozone depletion.

In all of these instances, Olson discerned three factors paramount in explaining the degree of success a group has, absent government coercion, in providing itself with a collective good: these were (1) the size of the group, (2) the heterogeneity of its members, and (3) the ability of the group to offer selective incentives to individual members to induce contribution to the collective good. In fact, the international alliance example served as the first instance of voluntary provision that could be verified empirically. With coauthor Richard Zeckhauser (1966), Olson modeled the allocative behavior of the North Atlantic Treaty Organization (NATO) Alliance as it provided its constituent members deterrence against attack and defense should attack materialize. The Olson–Zeckhauser analysis comported empirically with the predictions from Olson's theory, namely, that when the membership of a group is heterogeneous, the large and/or rich members will voluntarily contribute disproportionately large shares of the total public good, an effect Olson had named 'exploitation of the great by the small.' Thus, crucially Olson–Zeckhauser demonstrated that Olson's insights into the applicability of economic structure to political phenomena could be formulated into rigorous logic and empirically tested against reality. Later extensions of the voluntary provision model have lead to surprising insights (Bergstrom et al., 1986) into equilibrium outcomes when individuals in a group interact but do not coordinate their interactions, including demonstrations that equilibrium group provision within a range is

independent of the distribution of a group's income or wealth among its members. This voluntary provision model and its Nash equilibria continue to be the paradigm for numerous behavioral patterns, from cooperation among countries in regulating the international economy to controlling pollution, to the economics of charitable giving.

Public Choice, Political Economy, and Olson's *Rise and Decline of Nations*

From the first principle of 'Public Choice' mentioned above, it follows that the extramarket behavior of the individuals who make up governments is itself governed by individual search for individual satisfaction. Always committed to individual rationality embodied in calculated decision, in the middle stages of his career, Olson's elaboration of the public choice theme develops several corollaries to this proposition.

First, rather than being a utopian instrument to correct social inefficiencies in the provision of public goods stemming from free rider behavior, government behavior itself is driven by free rider failures. The problem of correcting for free rider deficits is itself in other words beset by free rider effects; for example, who will voluntarily contribute the resources, attention, and energy necessary to form environmental policy to protect the public from degradation of the atmosphere? The answer is precisely not the general public, which benefits from environmental control but instead those interest groups sufficiently well organized to exact or induce participation by their membership. This leads to a second corollary.

The concept of collective goods and free riders applied to the larger social mosaic implies a hierarchy or cascade of nested interest groups. First of all, people will act in their own interests, not the interests of their group; but once or if they have successfully overcome intragroup free riding and harnessed their individual incentives to the group – if they ever do – then those groups will act in their own self interests, not in the interests of their larger society.

A third corollary applies these principles to redistribution, with introduction of the notion of a distributional coalition, a group that by collective or collusive action can gain an increased share of society's wealth not by working to create a larger societal total, but by imposing external costs on others in the society. Such groups of coordinated rent seekers work though various instruments, many effected and enforced by the government that they influence, instruments that include discriminatory taxes or tariffs, barriers to mobility, or to entry as well as downright corruption and theft.

It was Olson's focus on this structure that brought the free rider problem to the widest academic audience of sociologists, historians, and political scientists, and that showed the applicability of the idea to current world developments and practical policy. His formulation of the idea, propounded in *The Rise and Decline of Nations* (1982), assembles these corollaries into one grand principle: the notion that entire societies fail or succeed on the quality of their management of free riding. It is the success of interest groups in overcoming obstacles to their own collective action and the failure of any larger polity to overcome the obstacles to its collective need to leash protectionists that molds the destiny of nations. Moreover, there is an inherent tendency over time, Olson argued, for efficiency-destroying and innovation-strangling special interest groups to proliferate in stable societies, such as the United States and United Kingdom with institutions that survived World War II (WWII) more or less intact. By contrast, those societies that experience the social upheavals typical of war or revolution – like Japan or Germany at the end of WWII – find that special interest groups and distributional coalitions are dismantled, thereby giving the society as a whole another start at constructing social and economic institutions afresh, which will encompass the aggregate collective interest of all the people. The idea of social and institutional sclerosis sometimes called 'demosclerosis' was thus advanced as a crucial determinant of societal economic success. It was an audacious idea, but one has to be subject to empirical rejection. The idea was noticed by social scientists the world around and resonated in the popular press.

The Character of Governance: Olson's *Power and Prosperity*

The final focus of Olson's energy – summarizing papers from the last decade of his life in the posthumous *Power and Prosperity* (1999) – was on the greatest of the free rider puzzles of all: how does government even come into existence, given the immense obstacles to collective rationality presented by individually rational free riding? Reflecting on this question, he was well aware of the immense rewards a society reaps from establishing a peaceful social order, and submitting itself to the institutions of governance. So there is no doubt that good governance is in the collective interest of a society; by investing in certain public goods, governments can increase the productivity of the society dramatically, and this benefits everyone. Moreover, for groups larger than families or at most tribes, governments are the only institutions that can do this. These are public goods such as establishment of a rule of law, institution and maintenance of property rights, provision of common defense, stable currency, contract enforcement, elements of infrastructure, and public education and health, and so on. Such investments in public goods benefit everyone in the society by making its whole industrial, commercial, and economic base much more productive.

But, Olson was also deeply aware that provision of public goods is far from the only or even major function of government. A second function is redistribution. Historically, almost all systems of government have emerged as vehicles of redistribution from the people to an autocracy, monarchy, or oligarchy. Contemporaneously, our generation has seen massive increases in government-enforced redistributions throughout the world. These redistributions, which play so crucial a role in the political economy of the modern state, are undertaken by governments, which exact taxes from some groups and make transfers for the benefit of other groups. The beneficiaries in this process are in general the groups to whom governments are beholden, and the taxpayers those to whom they are not. Thus, (1) redistribution to favored interests and (2) allocation to enhance everyone's productivity are the two *de facto* primary economic functions of modern government. Public good, productivity-enhancing allocation, and group discriminatory redistribution are conflicting resource use

options between which any government must choose. Irrespective of its political makeup, this is a choice every government must make and a supremely important one for the welfare of its members.

But how is the conflict between the two objectives reconciled? In brief, Olson argues the redistribution that a society accepts is the price it must pay to achieve the gigantic collective benefits of peaceful governance. But a precondition for this bargain even to be struck in the first place is that free rider obstacles be overcome sufficiently for government to be enduring (Olson, 1991, 1993). Governance by transient 'roving bandits' is inefficient; only when the roving bandit settles down to become a 'stationary bandit' – that is, an established monarch or autocrat – can society enjoy the immense benefits of peaceful development. Assuming this much success in conquering free riding pitfalls, Olson then extrapolated a theme from earlier in his career: the idea of groups as differentiating themselves along a scale of their 'encompassing interests.' This is the notion that a group that reflects the interests of a large proportion of its society in its economic calculus will behave very differently from one with a narrow membership and support. The more encompassing group will internalize more of the cost it inflicts on the entire society in pursuit of its narrower self-serving distributional goals. This provides the clue to his discovery, with coauthor Martin McGuire (1996), that the political logic of these invest–redistribute decisions will rest on a spectrum between two polar extremes. At one extreme is an idealized, consensual democracy where all redistribution issues have been settled unanimously and taxation only serves public good investment objectives. At the other extreme is perfect autocracy ruled by a dictator who taxes and spends solely to satisfy his own selfish desires. This work then shows that realistic societies can be analyzed as a mixture, in fact a weighted average, of these two polar cases. Thus, in making the choice between investment in the entire society and redistributive taxation from the powerless to the powerful, every polity behaves somewhat like a pure democracy and somewhat like a selfish dictatorship. Informed by this logic, McGuire–Olson derive the allocation and distribution rules that various governments across the spectrum will follow, showing, for example, that even a purely selfish dictator (providing he controls an enduring government) will not only transfer resources to himself but also invest in his economy. Why? Because investment will make the economy prosper and add even further to the tax take. Also intriguing is the proof that all societies and forms of rule are limited by an automatic governor on their incentives to tax. This contradicts the pessimistic expectations of some political moralists that extension of the democratic franchise would necessarily induce endless voracious redistributive taxation, leading finally to the ultimate destruction of the economy itself. It demonstrates, to the contrary, that for every ruling group, incentives to tax reach a point well before confiscation where desired redistribution voluntarily ceases.

The thrust of Olson's thought toward the end of his life was to bring these insights to bear on the puzzling economic trajectory of the Soviet Union, which seemed so economically cumbersome succeeded by Russia whose fortunes so declined after its markets were unleashed.

Mancur Olson's Impact

Mancur Olson's legacy extends broadly across the overlapping frontiers of social science and philosophy. His life was marked by numerous conferences devoted to examination of his ideas and books devoted to his insights such as Sandler (1992). His impact derives first of all from his ability to develop a simple idea so that it becomes powerfully applied across a grand subject area. An enthusiast for basic ideas, and ever the economist, he illuminated political processes using economics ideas and showed how political processes could hobble economic productivity. The result was to gain far wider acceptance of economic method and theory in the work of political scientists. Many before Olson would take for granted that if some outcome should be beneficial to all members of a group, then the group would act in concert somehow to bring it about, thinking in terms of joint rather than individual action, without asking how the joint action would be enforced. It was no longer so after Olson. He had the rare ability to take highly abstract but ultimately simple ideas and to apply them to the reality of economic and political development. Because of his work, awareness of the difficulties and paradoxes of achieving collective rationality, no longer confined to a cadre of specialists, has entered into the discourse of the largest universe of social science and philosophy.

Concrete witness to his impact is the Institutional Reform and the Informal Sector project. This US Agency for International Development vehicle was created to promulgate Olson's ideas in more than 30 countries around the globe, many struggling with the reform and adjustment process in a transition from socialism. Although he explored and exposed the difficulties of achieving socially desirable outcomes through uncoordinated individual action, he was no naive advocate of central planning or coercion. For him, the operations of governments were to be subjected to the same analytic scrutiny as markets, and government failures had to be exposed just as fearlessly as market failures. Thus, it was within the practical context of practical reform of institutions that Mancur Olson's final contributions took effect as an advocate for the idea that good governance can manage free rider challenges and achieve prosperity in the developing world (Olson, 1996). His life was marked by innovative ideas, personal energy, and an unflinching conviction that institutional reform could yield prosperity over poverty. The now universal appreciation of the links between good governance and economic success is his living bequest.

See also: Capitalism; Dictatorship in History; Economics, History of; Interest Groups, History of; Political Economy, History of; Public Choice Economics; State, History of; Welfare State, History of.

Bibliography

Referenced Works of Mancur L. Olson Jr.

Olson, Mancur L., 1965. The Logic of Collective Action: Public Goods and the Theory of Groups. In: Harvard Economic Studies, vol. 124. Harvard University Press, Cambridge, MA.

Olson, Mancur L., Zeckhauser, Richard, 1966. An economic theory of alliances (with Richard Zeckhauser). Review of Economics and Statistics XLVII, 266–279.

Olson, Mancur L., 1982. The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities. Yale University Press, New Haven.

Olson, Mancur L., 1991. Autocracy, democracy, and Prosperity. In: Zeckhauser, Richard (Ed.), Strategy and Choice. MIT, Cambridge, MA, 131–157.

Olson, Mancur L., 1993. Dictatorship, democracy, and development. American Political Science Review 87 (3), 567–576.

Olson, Mancur L., Spring 1996. "Big bills left on the sidewalk: why some nations are rich, and others poor," distinguished lecture on economics in government in. Journal of Economic Perspectives, 3–24.

Olson, Mancur L., McGuire, Martin C., March 1996. The economics of autocracy and majority rule: the invisible hand and the use of force (with Martin C. McGuire). Journal of Economic Literature XXXIV, 72–96.

Olson, Mancur L., 2000. Power and Prosperity: Outgrowing Communist and Capitalist Dictatorships. Basic Books, New York.

Citations to Others

Bergstrom, Theodore C., Blume, L., Varian, Hal R., 1986. On the private provision of public goods. Journal of Public Economics 29, 25–49.

Chamberlin, Edward, H., 1933. The Theory of Monopolistic Competition. Harvard University Press, Cambridge, MA.

Cournot, Augustin, 1897. Researches into the Mathematical Principles of the Theory of Wealth. Macmillan, London.

Marshall, Alfred, 1890. Principles of Economics. Macmillan, London.

Robinson, Joan, 1933. The Economics of Imperfect Competition. MacMillan, London.

Samuelson, Paul A., 1954. The pure theory of public expenditure. Review of Economics and Statistics 36, 387–389.

Sandler, Todd, 1992. Collective Action: Theory and Applications. University of Michigan Press, Ann Arbor.

Online Dating

Monica T Whitty, University of Leicester, Leicester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ever since the early days of the Internet, men and women have been logging on to search for love. During this time there was much debate over whether 'real' relationships could be formed with 'strangers' online or whether instead these relationships were meaningless and potentially psychologically unhealthy. Theories such as the Social Presence and Social Context Cues Theory were developed to predict that online relationships were of little benefit to individuals. This article provides a summary of these early debates, and points out the importance of not drawing a distinction between 'physical world' versus 'virtual world,' but instead that there are many types of spaces online, with different features with different norms that govern these spaces. The article makes note of some of these spaces and draws distinctions between them. It also highlights the 'hyperpersonal' theory first noted by Walther (1996), which contends that under certain conditions individuals form closer relationships online than they would in the physical world. Contemporary examples of online dating examining online dating and social networking sites are examined to relook at early theories and point out how the different features and norms of the space lead to different trajectories when it comes to online dating. Finally, it examines a relatively new crime online, that exploits individuals hope to find 'true' love online – known as the romance scam. The theoretical reasons for its success are highlighted here.

From its early days the Internet has supported social interaction and the search for a soul mate. In early studies, there was much debate over whether 'real' relationships could be formed with 'strangers' online or whether instead these relationships were meaningless and potentially psychologically unhealthy.

As mentioned in the introduction to this article, in the early days of the Internet, there were two opposing theoretical views to online relationships: online relationships were shallow and meaningless versus the Internet provided new radical opportunities for forming relationships. When we consider these theories, we need to be mindful that the Internet was a very different space to how we know it today, and the types of people who inhabited the space were not represented of the general population. The virtual world was mostly constructed as text, where individuals' impression managed their online selves.

Social Presence Theory and Social Context Cues Theory

Advocates of Social Presence and Social Context Cues Theories took the view that online relationships had a noticeable deregulating effect on communication leading to unfulfilling relationships, that were low in intimacy (Hiltz et al., 1986; Kiesler et al., 1984; Sproull and Kiesler, 1986). Social presence was understood to mean the degree to which a person is perceived as 'real' and involved in mediated communication and scholars that devised these theories believed that given that online communication involved fewer nonverbal cues (such as facial expression, posture, dress, and so forth) and auditory cues in comparison to face-to-face communication, it was said to be extremely low in social presence and social context cues. According to these theorists, as social presence and social cues declines communication becomes more impersonal. In contrast, when more information is available about how one physically looks this leads to greater positive regard. Hence, given that there is less social presence online compared to other media online, communication was said to be less personal and intimate.

These views found some support by researchers. For example, Kraut et al. (1998) in a well-known study, referred to as the HomeNet Study, found that greater use of the Internet was associated with negative effects on an individual, such as decreases in the size of one's social circle and increases in depression and loneliness. In their longitudinal study, households who had never accessed the Internet were provided with a computer, a free telephone line, and free access to the Internet. During the course of the study, they tracked changes in psychological states over time. The study found a significant relationship between heavy Internet usage and loneliness. Kraut et al. argued that since initial loneliness failed to predict subsequent loneliness, the most likely explanation was that the increased use of the Internet was what caused the increase in loneliness. In other words, Internet usage was taking up time that could be better used for more psychologically beneficial interactions offline. Kraut et al. made the claim that *weak ties* were being established online, which were of poorer quality compared to the types of relationships, and strong ties already established offline.

Kraut et al.'s (1998) 'HomeNet Study' has been widely criticized. One of the major criticisms is that they used only three items from the UCLA loneliness scale to measure loneliness and their Cronbach's alpha of .54 was clearly poor (Grohol, 1998). Morahan-Martin (1999) points out that the sample size was too small and not randomly selected. Perhaps a more important criticism, however, is that Kraut et al.'s findings might only explain novice Internet users (LaRose et al., 2001). Moreover, as LaRose et al. (2001) have maintained, self-efficacy might be an important variable to consider. The individuals that spent more time online in Kraut et al.'s study might have been simply ineffective users of the Internet and the stress in trying to work out how to use this new technology might have caused them to become more depressed (Whitty and McLaughlin, 2007; Whitty and Carr, 2006). Interestingly, in the 3-year follow-up to the HomeNet Study the same

researchers found that almost all of the previously reported negative effects had dissipated (Kraut et al., 2002). Instead, higher levels of Internet use were positively correlated with measures of social involvement and psychological well-being. Perhaps such results might be explained by LaRose et al.'s (2001) claim that it is also important to consider self-efficacy. Whitty and Carr (2006) argue that participants in the HomeNet Study might have become more Internet savvy over time, which in turn altered the way they used the Internet.

Disinhibition Effect

Around the same time that theories such as Social Presence and Social Context Cues theories were being developed, other psychologists were observing a phenomenon fairly unique to online interactions, which they referred to as the 'disinhibition effect.' These individuals were observing that in some online spaces, are more likely to open up about aspects about themselves in cyberspace and act out behaviors they might not otherwise in traditional face-to-face settings (Joinson, 2001; Suler, 2004). Suler (2004) describes this as a double-edged sword as some people reveal secret emotions, fears, and wishes or show unusual acts of kindness and generosity, which he refers to as 'benign disinhibition.' Barak et al. (2008) found that because of the online disinhibition effect, individuals are able to open up more in online support groups. This, they argue, fosters personal empowerment, control, self-confidence, and improved feelings. In contrast, there are times when people are more rude, harsh critics, show anger, hatred, or even threats than they typically would be face to face. This Suler calls 'toxic disinhibition.' Cyber-bullying is an example of this type of toxic disinhibition. For example, when young people write cruel messages on social networking sites.

The disinhibition effect research demonstrated that there was a little more going on that those who subscribed to the social presence and social context cues theories. Although they too noticed that some online ties were 'weak' and found evidence for virulent relationships, at the same time they were finding the opposite results; that is, that individuals were self-disclosing more about themselves and being more kind than they would perhaps in the physical world.

Less Can Mean More

Others too were observing that traditional ideals and values important in the physical world were equally important online. Whitty and Gavin (2001), for instance, found in their study where they interviewed 60 Internet users, that ideals such as trust, honesty, and commitment were just as important in relationships formed in these spaces. Rather than finding evidence of less 'real' or less satisfying relationships online, some of the participants in their study reported that their relationships seemed to work better solely on the Internet. As this 18-year-old women reports:

> *It* [the relationship] developed through an interesting chat on IRC and a series of about 500 e-mails. The attraction was merely someone who cared and listened. He was very sensitive and caring, and his picture was hot! {laughs} … we exchanged addresses and he sent me presents on Valentines Day and Easter. We would write a two page e-mail every day, send sounds to each other, and eventually after 6 months we talked on the phone. Our phone conversation was very weak so we decided to stick to e-mail… We met after 8 months of exchanging e-mails. He was a great guy, and it would have worked but he lived in email. It was a good experience though and he was exactly like his photo.
>
> Whitty and Gavin, 2001, p. 628

In their study they also found that chat room users reported feeling less self-consciousness and less aware of being socially evaluated, which in turn allowed these individuals to reveal intimate details about themselves, while maintaining distance and personal space. Ironically, many of the male participants in this study believed that by disguising their identity they could be more emotionally honest and open.

Social Information Processing Theory

Given that researchers were finding that despite the lack of social presence and social context cues that 'real' relationships were forming online, theorists turned to explain the reason for these 'real' relationships. Of course, we need to be mindful that part of the explanation might be, as discussed earlier in this article, that individuals were getting better at communicating online and that different types of people were gravitating to the Internet. However, theorists noted that individuals might also be able to overcome some of the hurdles the Internet originally presented. Joe Walther is one such theorist who started to try to understand why 'real' relationships were emerging online.

While Walther (1995) subscribed to the reduced social cues view, he believed that many of these studies did not account adequately for time. Given the shortcomings of previous research, Walther and his colleagues sought to develop a more thorough understanding of the exchanges that take place in online environments. In doing so, Walther originally came up with the 'social information processing theory (SIP).' The SIP theory "refers to the way by which communicators process social identity and relational cues (i.e., social information) using different media" (Walther, 1995, p. 190). This theory proposes that the main difference between face-to-face communication and computer-mediated communication (CMC) is the pace that relationships develop in each space rather than capability to develop relationships. Walther argued that although CMC may be more aggressive at first, with time this dissipates. Walther (1992, 1995, 1996) stressed in his original thinking that many of the differences between online relationships and face-to-face relationships diminish over time, and although restricted bandwidth may limit the rate of information exchange, allowing longer and/or more frequent communication can alleviate this problem. However, Walther's theory did not end there – through extended empirical work, especially in lab environments, Walther started to notice that online relationships can be somewhat different to ones formed in the physical world. This led him to develop the theory of 'hyperpersonal communication.'

Hyperpersonal Communication

Walther extended his SIP theory to develop a hyperpersonal communication framework. This theory posits that "CMC users sometimes experience intimacy, affection, and interpersonal assessments of their partners that exceed those occurring in parallel FTF activities or alternative CMC contexts" (Walther et al., 2001, p. 109). Walther (1996) argued that people use the technical capacities of the Internet to assist in impression development. According to Walther (2007, p. 2539), the model "specifies several concurrent dynamics in sender, receiver, channel, and feedback systems that are affected by CMC attributes, which promote the development and potential exaggeration of impressions and relationships online …" According to this theory, receivers idealize partners because of the messages they receive, which they believe demonstrate the similarity in their online partner as well as a highly desirable character. In contrast, senders exploit the technology to selectively self-present aspects about themselves that the other would deem socially desirable. The CMC features allow one to be strategic in their presentation of self and the CMC environment creates a space where the outside world is filtered out and cognitive resources are instead employed to focus on the online communications. As Walther (2007) explains: "The CMC channel facilitates editing, discretion, and convenience, and the ability to tune out environmental distractions and re-allocate cognitive resources in order to further enhance one's message composition" (p. 2539). Once this cycle of communication has begun, Walther argues that "CMC may create dynamic feedback loops wherein the exaggerated expectancies are confirmed and reciprocated through mutual interaction via the bias-prone communication processes" (p. 2539).

Contemporary Theories about Online Relating

In modern times the Internet is becoming a more popular place for couples to meet. Dutton et al. (2008) sampled married couples in Australia, Britain, and Spain and found that 6% of married couples in the United Kingdom, 5% in Spain, and 9% in Australia had first met their partner online. The largest proportion of married couples that met online ranged from 26 to 55 years old.

As mentioned at the outset of this article, researchers need to be mindful that the Internet is not a homogenous space. More recent notions about the Internet examine the different spaces online, considering the different features of that space and the norms that govern that space. This article moves to focus on two such spaces to demonstrate by example, that one theory to one theory does not fit all: online dating and social networking sites.

Online Dating Sites

Online dating sites, in a fairly primitive form to the way we know them today, started appearing in the 1980s. In the early days, online dating was stigmatized and researchers were finding that shier people were gravitating to these sites to find a mate (Scharlott and Christ, 1995). In the 1980s these sites were largely text-based with limited space for individuals to create a dating profile in the hope to attract others. These days many online dating sites attempt to effectively match users of the service. Online daters are often expected to complete personality tests, as well as surveys on their interests and what aspects they are looking for in a partner. From there, matches are often given compatibility ratings. Other sites allow clients to find their own matches – where they need to work through a sea of profiles to make their choices.

Researchers have found that online daters often spend many hours contrasting an attractive profile (Whitty, 2008a,b,c; Whitty and Carr, 2006). Whitty (2007, 2008a) has argued that, as with newspaper ads, online daters are driven to 'commodify' themselves – ensuring they are presenting a self that others will feel compelled to 'buy' into. Some of the characteristics researchers have found that online daters consider important to present include: their looks, interests and activities, personality, humor, occupation, intelligence, uniqueness, and hopes and dreams (see Whitty and Carr, 2006; for an exhaustive list and percentages). Whitty (2007, 2008a) has found that online daters are savvy enough to know that simply presenting an attractive self on the site is not enough to attract the appropriate mate. Instead daters have to present a self that appears to have an honest and authentic presentation of themselves. Whitty (2007) has named this the BAR theory: that is, that to successfully attract a mate, online daters need to construct profiles that present a balance between an 'attractive' and a 'real' self.

Unlike what the hyperpersonal theory would predict, relationships developed between people that meet on dating sites are typically not intimate or close. This can be explained partly by the features of the space as well as the way relationships typically progress from these sites. As described above, online daters spend more time crafting a profile than they do getting to know one another. In fact, the processes involved in dating are quite different to the ones that Walther and others observed, where individuals spend much time communicating online, disclosing their private selves. Whitty (2008c) has drawn from Given's (1979) work on a traditional model of courtship to devise the five phases involved in online dating: *the attention phase* (which involves creating an online profile); *the recognition phase* (sending symbols such as 'winks' and 'kisses' to a profile that the person is interested getting to know the person whom it belongs to); *the interaction phase* (the daters begin communicating with each other through the site and then very quickly via Instant Messenger and e-mail); *face-to-face meeting* (this is a screening out process to determine if there is any chemistry and whether the person matches up to their profile); and finally *resolution* (starting to date). Given the way these relationships progress, there is not the opportunity, as the hyperpersonal theory predicts, for individuals to self-disclose in detail and become highly intimate and close with one another.

Social Networking Sites

Getting to know a person through a social networking site can again be potentially different to knowing someone via other online spaces or in our face-to-face interactions. Antheunis et al. (2010), for example, looked at the uncertainty

reduction process in initial online interactions on a Dutch social networking site. They found that participants used three types of uncertainty reduction strategies (URSs): passive, active, and interactive. Passive URS were the most commonly used, followed by interactive and active URS. Importantly they point out that:

> In contrast to text-based CMC environments, social network sites have more auditory and visual cues available and are relatively open systems. Apparently, this allows CMC participants not only to use interactive URS's (i.e., direct questioning and self-disclosure), but also to observe their target persons (passive strategy), or to ask other participants for information about the target persons (active strategy).
>
> p. 106

These researchers, however, also found that although the passive URS was the most frequently used strategy, the interactive URS was the only strategy that reduced the information seeker's level of uncertainty. As to be expected, they also found that a low level of uncertainty lead to social attraction. More importantly, however, they found that the relationship between level of uncertainty and social attraction is not direct, but rather moderated by perceived valence of the information (i.e., the extent to which an individual perceives the information of the target person as positive or negative, regardless of its credibility or quality).

Nosko et al. (2010) also ran a study to examine self-disclosure levels with Facebook users. They found that those seeking out a relationship on Facebook were more likely to disclose details about a number of topics, including views, relationship information, and personal information. In addition, they found that those seeking out relationships were more likely to disclose potentially stigmatizing information.

Social networking sites, such as Facebook, also play an important role once individuals have decided they want to date a person. Stern and Taylor (2007), for instance, found that individuals in romantic relationships used Facebook to check on the status of their relationship, commitment, and fidelity as well as to keep in touch with their partner.

Facebook provides a platform where individuals are able to monitor others, including loved ones. Researchers have found that this has lead to problems with individual's romantic relationships. Muise et al. (2009) found that increased Facebook use significantly predicted Facebook-related jealousy. Utz and Beukeboom (2011) examined SNS jealousy and SNS relationship happiness and the consequences of SNS use for romantic relationships. Overall, they found that participants experienced more happiness than jealousy in reaction to their partner's activities on social networking sites. However, individuals low in self-esteem were more SNS jealous than those high in self-esteem. Self-esteem also moderated the effects of SNS use and the need for popularity on SNS jealousy and SNS relationship happiness.

Dangers Involved in Online Dating

Although, as this article has demonstrated, the Internet provides new opportunities to meet romantic potentials and develop these relationships, researchers have found that criminals have found a new way to con individuals by setting up fake profiles and alluring victims into a fake relationship (Whitty, 2012; Whitty and Buchanan, 2012a,b). According to Whitty and Buchanan (2012a), the online dating romance scam emerged around 2007 and has its roots in paper-mail-based fraud. They have found that it is likely that at least 230 000 victims have fallen to this scam in Great Britain alone, and amounts lost range from £50 to £800 000. The online dating scam involves both identity fraud and mass marketing fraud. In this scam, criminals pretend to initiate a relationship through online dating or social networking sites with the intention to defraud their victims of large sums of money. Scammers create profiles with stolen photographs and at a very early stage the criminal declares their love for the victim and requests that the relationship move from the dating site to Instant Messenger and email, stating they want an exclusive relationship with the victim. Communication is frequent and intense, over periods of weeks, months, or even years. Phone calls might also be made. Fake documents are created to make the requests appear more genuine. Often small requests for money are made as a way of testing the waters and this is then followed by a crisis where the victims is asked for large sums of money that are urgently needed (crisis narratives include medical crisis, such as a car accident, business emergency, luggage needing to be released from customs containing gold bullion).

Relating this back to theories on online relationship development, Whitty and Buchanan (2012b) have argued that Walther's hyperpersonal theory can be applied to explain the success of this scam. Unlike typical relationships developed from an online dating site, where the individual physically meets the potential partner fairly quickly after meeting on the dating site, and spend little time getting to know them online; the victims of this scam spend copious amounts of time communicating with the criminal. Whitty and Buchanan (2012b) argue that the criminal uses Internet Communication Technologies (ICT) to groom the victim, sending poetry via email, and hours of time getting the victim to self-disclose highly intimate details about themselves and their lives via Instant Messenger. Victims often describe the relationship as the closest, most intimate relationship they have experienced in their life time, and admit that when they learn the relationship was fake, and even after losing large sums of money, they still want to maintain the relationship. This sometimes leads to a second wave of the scam, where the victim might have the criminal admit to deception, but during the scam they genuinely fell in love with the victim, and then the scam continues. The psychological impact from this scam is great, where the victim experiences a double hit of the loss of money and a close romantic relationship.

Conclusions

In the early days of the Internet a specific group of individuals gravitated to use it, in part to develop romantic relationships. In those days the Web was a space which mostly consisted of text, bandwidth was small, and download time very slow. Perhaps this accounts for the early theorists' views that the Internet is an impersonal space where it was believed that

relating online would lead to negative psychological well-being. The surprising result in the early days was that some online relationships were actually very intense and intimate and progressed healthy offline – a contrasting view that has been supported by empirical evidence. Despite these encouraging results, as the Internet has evolved we have learnt that we cannot develop one theory to explain all online relationships – individuals can be virulent and cold and calculating or intimate online, depending on the space and the norms of the space. Developing a relationship via an online dating site can be significantly different to meeting someone in a social networking site. The research on online dating has elucidated that there are certain skills involved in crafting a profile as well as making contact with others on the site. These relationships typically progress very quickly to offline settings, making it difficult to get to know one another online (a space where hyperpersonal theory is not supported). Facebook, however, is different again where individuals are often very strategic in the ways they impression-manage and the risks people are prepared to make to self-disclose. Facebook can lead to problems with already established romantic relationships, if individuals become too obsessive with monitoring their partner's activities. In addition to the new opportunities for online dating the Internet has opened up the floodgates for criminals to exploit spaces like online dating sites and social networking sites, to develop fake trusting relationships with unsuspecting victims and defraud them of large sums of money. The Internet will surely continue to grow and evolve over the forthcoming years, and no doubt will continue to play an important role in finding a romantic partner, and developing and maintaining a relationship with a loved one.

See also: Computer Mediated Communication; Love and Intimacy, Psychology of; New Media and Social Capital; Social Media.

Bibliography

Antheunis, M.L., Valkenburg, P.M., Peter, J., 2010. Getting acquainted through social network sites: testing a model of online uncertainty reduction and social attraction. Computers in Human Behavior 26, 100–109.

Barak, A., Boniel-Nissim, M., Suler, J., 2008. Fostering empowerment in online support groups. Computers in Human Behavior 24, 1867–1883.

Dutton, W.H., Helsper, E.J., Whitty, M.T., Buckwalter, G., Lee, E., 30 September 2008. Mate selection in the network society: the role of the internet in reconfiguring marriages in Australia, the United Kingdom and United States. Social Science Research Network. Found at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1275810.

Givens, D., 1979. The nonverbal basis of attraction: flirtation, courtship, and seduction. Psychiatry 41, 346–359.

Grohol, J., 1998. Response to the HomeNet study. PsychCentral. From: http://psychcentral.com/homenet.htm (retrieved 27.10.04.).

Hiltz, S., Johnson, M., Turoff, M., 1986. Experiments in group decision making: communication process and outcome in face-to-face versus computerized conferences. Human Communication Research 13, 225–252.

Joinson, A.N., 2001. Self-disclosure in computer-mediated communication: the role of self-awareness and visual anonymity. European Journal of Social Psychology 31, 177–192.

Kiesler, S.J., Siegel, J., McGuire, T.W., 1984. Social psychological aspects of computer-mediated communication. American Psychologist 39, 1123–1134.

Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukopadhyay, T., Scherlies, W., 1998. Internet paradox: a social technology that reduces social involvement and psychological well-being? American Psychologist 53, 1017–1031.

Kraut, R., Kiesler, S., Boneva, B., Cummings, J.N., Helgeson, V., Crawford, A.M., 2002. Internet paradox revisited. Journal of Social Issues 58, 49–74.

LaRose, R., Eastin, M.S., Gregg, J., 2001. Reformulating the Internet paradox: social cognitive explanations of Internet use and depression. Journal of Online Behavior 1 (2). From the World Wide Web: http://www.behavior.net/JOB/v1n2/paradox.html (retrieved 27.10.04.).

Morahan-Martin, J., 1999. The relationship between loneliness and Internet use and abuse. Cyberpsychology & Behavior 2, 431–439.

Muise, A., Christofides, E., Desmarais, S., 2009. More information that you ever wanted: does Facebook bring out the green-eyed monster of jealousy? CyberPsychology & Behavior 12 (4), 441–444.

Nosko, A., Wood, E., Molema, S., 2010. All about me: disclosure in online social networking profiles: the case of FACEBOOK. Computers in Human Behavior 26, 406–418.

Scharlott, B.W., Christ, W.G., 1995. Overcoming relationship-initiation barriers: the impact of a computer-dating system on sex role, shyness, and appearance inhibitions. Computers in Human Behavior 11 (2), 191–204.

Sproull, L., Kiesler, S., 1986. Reducing social context cues: electronic mail in organizational communication. Management Science 32 (11), 1492–1512.

Stern, L.A., Taylor, K., 2007. Social networking on Facebook. Journal of Communication, Speech & Theatre Association of North Dakota 20, 9–20.

Suler, J., 2004. The online disinhibition effect. CyberPsychology & Behavior 7, 321–326.

Utz, S., Beukeboom, C.J., 2011. The role of social network sites in romantic relationships: effects on jealousy and relationship happiness. Journal of Computer-mediated Communication 16 (4), 511–527.

Walther, J.B., 1992. Interpersonal effects in computer-mediated interaction: a relational perspective. Communication Research 19, 52–90.

Walther, J.B., 1995. Relational aspects of computer-mediated communication: experimental observations over time. Organizational Science 6, 186–203.

Walther, J.B., 1996. Computer-mediated communication: impersonal, interpersonal and hyperpersonal interaction. Communication Research 23, 3–43.

Walther, J.B., 2007. Selective self-presentation in computer-mediated communication: hyperpersonal dimensions of technology, language, and cognition. Computers in Human Behavior 23, 2538–2557.

Walther, J.B., Slovacek, C., Tidwell, L., 2001. Is a picture worth a thousand words? Photographic images in long-term and short-term computer-mediated communication. Communication Research 28, 105–134.

Whitty, M.T., 2012. True Romance. Police Professional.com. http://www.policeprofessional.co.uk/.

Whitty, M.T., 2007. The art of selling one's self on an online dating site: the BAR Approach. In: Whitty, M.T., Baker, A.J., Inman, J.A. (Eds.), Online Matchmaking. Palgrave Macmillan, Houndmills, pp. 57–69.

Whitty, M.T., 2008a. Revealing the 'real' me, searching for the 'actual' you: presentations of self on an Internet dating site. Computers in Human Behavior 24, 1707–1723.

Whitty, M.T., 2008b. The joys of online dating. In: Konjin, E., Martin, T., Utz, S., Linden, A. (Eds.), Mediated Interpersonal Communication: How Technology Affects Human Interaction. Taylor & Francis Group/Routledge, New York, pp. 234–251.

Whitty, M.T., 2008c. eDating: the five phases of online dating. In: Romm, C. (Ed.), Social Networking and eDating. Information Science Reference Hershey, PA, pp. 278–291.

Whitty, M.T., Buchanan, T., 2012a. The online dating romance scam: a serious crime. Cyberpsychology, Behavior, and Social Networking 15 (3), 181–183.

Whitty, M.T., Buchanan, T., 2012b. The Psychology of the Online Dating Romance Scam Report. University of Leicester. http://www2.le.ac.uk/departments/media/people/monica-whitty/Whitty_romance_scam_report.pdf.

Whitty, M.T., Carr, A.N., 2006. Cyberspace Romance: The Psychology of Online Relationships. Palgrave Macmillan, Basingstoke.

Whitty, M., Gavin, J., 2001. Age/sex/location: uncovering the social cues in the development of online relationships. CyberPsychology and Behaviour 4 (5), 623–630.

Whitty, M.T., McLaughlin, D., 2007. Online recreation: the relationship between loneliness, Internet self-efficacy and the use of the Internet for entertainment purposes. Computers in Human Behavior 23 (3), 1435–1446.

Online Pornography

Susanna Paasonen, University of Turku, Turku, Finland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Public debates on the cultural role and status of pornography have intensified in the course of the 2000s. The abundant and ubiquitous accessibility of online pornography has impacted the overall cultural visibility of sexually explicit media materials, while the diversification of production practices and available content has pushed the established definitions and assumptions concerning the genre. This article explores the developments in online pornography, its cultural and social implications, the transformations that have occurred the genre, as well as the challenges that all this poses for scholarly investigation and public debate.

Public debates on the cultural role and status of pornography have intensified in the course of the 2000s. The abundant and ubiquitous accessibility of online pornography has impacted the overall cultural visibility of sexually explicit media materials, while the diversification of production practices and available content has pushed the established definitions and assumptions concerning the genre. In a parallel development, scholarly investigations on the topic have given rise to something to an interdisciplinary field that can be loosely categorized as 'porn studies.' This article explores the developments in online pornography, its cultural and social implications, the transformations that have occurred the genre, as well as the challenges that all this poses for scholarly investigation and public debate.

Abundance

The Development of Online Pornography

Pornography (or porn for short) was, in both its amateur and professional variations, a crucial element in the exchanges in bulletin board systems, Usenet newsgroups, Internet-Relay Chat, and file transfer through FTP that preceded the World Wide Web. Internet users have, for decades now, shared pictures, text, and video files harvested from their personal collections, copied from other sources, or produced by the users themselves. The easy usability of graphical Web interfaces since the launch of the browsers Mosaic in 1993, Netscape Navigator in 1994, and Microsoft Internet Explorer in 1995, in combination with the spread of network connectivity at the home and the workplace, marked both the gradual rise in popularity of the Internet as communication and storage medium and a dramatic change in the ways of distributing and consuming pornography through online platforms. The consumption of online pornography necessitates no visits to a specialty shop or orders through a postal mail catalog: online porn can be consumed for free, in the privacy of one's home or office, anonymously, and with a bewilderingly diverse sets of options.

Online porn was from the beginning both visual and textual, although image size and resolution were rather low due to the limitations of dial-up modem connection speeds. With the gradual increase in high-speed Internet connections, online porn grew increasingly audiovisual and high in resolution while web camera services established their popularity among the consumers of pornography. By the early 2000s, online platforms had challenged magazines and DVDs as the main distribution formats for pornography while also supporting interaction opportunities and peer production among users.

Throughout the 1990s, pornography was considered one of the few profitable forms of online content production – and, before online shopping and gaming picked up, perhaps even the only profitable one. The needs of the porn industry, which is understood here as a broad umbrella term for divergent production practices, aesthetics, and economies, drove the development of Web technologies and business practices from hosting services to safe credit card processing, banner advertisement, pop-ups, Web promotion, and streaming video technology (Bennett, 2001; Johnson, 2010). Although pornography is often excluded from overviews of Web history or presented as its 'dark side' (Parikka and Sampson, 2009), it has been an organic element of the evolution of the Web, both technologically and economically. New media theorist Wendy Chun (2006: pp. 78–79) points out that the success of online porn in the mid-1990s set the example for more mainstream companies on the possibilities of e-commerce, safe credit card transactions, and potential profits that could be made with the then emergent mass medium of the Web. In this sense, pornography functioned as a prototype for other kinds of online content production and product marketing.

Large porn companies with established production and distribution in print and video – such as the North American Playboy or Vivid – branched online only after the mid-1990s. Up until then, online porn enterprises were primarily independent and semiamateur. In the early years of the Web, before the development of specialized professions and skillsets such as those of a Web designer (Kotamraju, 1999), it was easy to code one's own site with only the basic knowledge of HTML. Small porn ventures could gain considerable visibility with relatively limited resources when the total number of global Web sites still remained modest (see Mash, 2004). Independent and alternative pornographies catering to subcultural markets and making use of community features grew in popularity during the late 1990s and their principles of operation influenced those of more mainstream companies seeking out new audiences and uses for their online platforms (Attwood, 2007; Paasonen, 2010). Enterprises smaller and larger targeted both

mass audiences and those adding up to highly specialized sexual niches. Consequently, the landscape of online porn was, and remains, both heterogeneous and fragmented.

Social Media

Since the dot.com crash of the early 2000s, Web site concepts and business models have been increasingly built on user-generated content, social media, and community platforms. In the late 1990s, 'content producer' was a paid profession similar to the editor of a print publication. Content producers wrote short articles and provided visuals for Web sites and portals in order to attract paying customers or at least returning users that would translate as advertisement value. In the 2000s, content production became the preoccupation of the users themselves whereas sites aim to generate profit by collecting user data and selling targeted advertising based on these data. Terms such as 'social media' or 'Web 2.0' that are used to describe and pigeonhole such developments may nevertheless be ahistorical in that they turn attention away from the abundance of user-generated content that largely characterized the Web cultures of the 1990s (Bruns, 2008: pp. 3–4).

With the introduction of tube sites – video hosting sites modeled after YouTube (launched in 2005) – porn consumption has been largely centralized to platforms such as YouPorn, Xtube, RedTube, Tube8, xHamster, and PornHub, many of which remain in the top 100 of globally most visited sites. On tube sites, users can browse for content in multiple categories, do searches, click on related content without leaving the site, and, if they so desire, tag and rate videos and comment on them as registered users. This marks a clear departure from the link sites and click-through characteristics of porn sites well into the 2000s. According to estimates, the overwhelming majority of online porn users – well over 90% – do not pay for the content they consume. The effect of tube sites especially on the consumption of freely available content is therefore considerable while less pronounced on the operating principles of pay sites.

These developments have marked a major increase in the accessibility, visibility, as well as popularity of amateur pornography. While amateur content was broadly circulated already on preweb applications and platforms and its histories well predate the era of online communication, its current visibility and accessibility is unprecedented. Together with both legal and illegal file sharing in peer-to-peer (P2P) networks, tube sites have contributed to the decrease in the profits of the more traditional (print and DVD-based) porn production. Since online pornography is largely consumed for free, it seems that the profits currently made in the field are shifting from production to distribution. On tube sites, private users upload content from heterogeneous sources whereas companies distribute short clips, teasers, and trailers of their work through the same platforms in order to attract paying customers to their own sites (Morris and Paasonen, 2013: pp. 555–556).

Random Estimates

The percentage of Web traffic taken up by online porn or the total number of porn sites are difficult to pinpoint due to highly variable estimates. Search engines routinely exclude pornography from their freely published listings of the most popular search terms. The so-called adult sites are similarly not necessarily listed among the top global sites. Openly available statistics, many of which are published by sites promoting filtering software, argue that porn comprises between 12 and 34% of all Web sites whereas peer reviewed studies have, for almost a decade, set the estimate to around 4% (e.g., Spink et al., 2006). The number of pornographic sites is certainly much smaller than the most cited and inflated statistics claim, yet it should be noted that an individual site can be massive and feature tens of thousands of videos and image files.

On the one hand, there is currently more online pornography than ever in open circulation: videos, still images, and texts, circulate from one platform to another, materials produced in analogue formats are digitized for online consumption, new content is uploaded on a daily basis, pornography takes up considerable bandwidth in Web traffic due to the use of video, and the absolute volume of pornographic content accumulates. On the other hand, the proportional volume of pornography in terms of all Web content is smaller than it was in the 1990s. The number of porn sites has steadily increased yet other forms of content have done so even more rapidly and steeply (Coopersmith, 2006). It is therefore not entirely fair to say that the Web is full of porn but neither is there any shortage thereof.

Diversification

Articulated Distinctions

Pornography has, at least since the 1970s, been strongly divided into different subgenres that are marked apart from one other, from couples' videos to BDSM (bondage, domination/discipline, sadism, masochism), or gonzo (reality-style pornography that makes use of point of view camera and where performers double as cameramen). Online pornography has nevertheless involved an unprecedented visibility of diverse sexual subcultures, preferences, and niches, partly in novel forms and combination (Halavais, 2005; Paasonen, 2011a). This is partly due to the diversification of porn production: professional, semiprofessional, amateur, and professional-amateur (Pro-Am, i.e., productions that make use of paid actors performing as amateurs) productions of different sizes and foci abound online. Online distribution is relatively inexpensive in that it does not necessitate the manufacture of physical products such as magazines or DVDs, and bandwidth and server expenses have gradually grown more affordable. Online distribution has afforded global markets for products targeted at relatively marginal audiences of the kind that would not be commercially viable on DVD (e.g., Lane, 2001: pp. 223–227; Bennett, 2001: p. 384). Companies and individual people upload porn for others to watch for free, for pay, and in exchange for a site membership registration. When applied to online pornography, the notion of the 'porn industry' is by necessity a rather elastic marker of individual and collective agents, some of which have no direct part in the production of porn and which merely facilitate its distribution and consumption, as though hosting services.

The diversification of the pornographic content is also partly fueled by the information architecture of the Web.

Images and videos are rendered searchable through indexing and metadata, namely descriptive textual information as for the style and contents of the imageries in question that is either embedded in the source code or provided as tags, captions, and descriptions on the site. Working in tandem, indexing, tagging, metadata, and search engines have rendered all kinds of distinctions far more pronounced than they have been in pornography distributed through other media platforms. Due to this development, differences between diverse subgenres, performers, body styles, acts, locations, and areas of production have grown increasing articulated and recognizable.

Addressing the fragmentation of the mass market of pornography, Mark Dery (2007) argues sites aim to grab potential users 'by their eyeballs' by showing them images that are surprising or even amazing in their novelty or eccentricity. In this sense, the attention economy of the Web – the need to mark one's site apart from the competition and to attract returning users – also fuels the diversification and fragmentation of the pornographic content. For some scholars, the accumulation of alternative, queer, and experimental pornographies online, together with the ubiquity of amateur porn that challenges the body norms and production practices of mainstream commercial porn, marks a rupture between traditional pornographies and the sexual exchanges seen as native to online platforms (e.g., Jacobs, 2007).

Camera Sites and the Boundaries of Pornography

The increased popularity of webcam sites offering live video connection and interaction possibilities that seem to optimize the promises of immediacy and presence associated with the Internet as a communication medium is a telling, and crucial, development on a number of levels. As mediated encounters, webcam sites blur the previously established boundaries of pornography and live sex entertainment such as strip shows. In connection to this, they blur the markers of professionals and amateurs connected to both sex work and the porn industry. Webcam workers may offer their services through sites that host hundreds, perhaps even thousands of models and that set strict conditions to the performers' actions. Alternatively, models may operate independently for a fee paid by the customers, or in return for gift cards and products purchased from their gift lists at online marketplaces. Webcam workers can be professionals such as porn stars, semiprofessional, amateur, or any combination thereof; they can sport a range of body styles, be of different ages, ethnicities, and nationalities, and their performances may be highly particular and involve multiple interaction possibilities.

Users may post their own images and videos, or those of their partners, on platforms ranging from tube sites to subscription-based portals, and openly accessible forums. Users may exhibit themselves on webcam sites such as Chatroulette, engage in explicit textual exchanges, write their own porn stories, perform sex with the aid of avatars in an online role-playing community, and post suggestive personal profiles on dating sites or social networking sites with the aim of finding partners for sex (Attwood, 2006: pp. 79–81). The definitions of the pornographic content easily grow elastic in the context of user-generated content and personal exchanges where – like ultimately in most human activity – sexuality plays a central role. The practice of sharing an image with another user one-on-one via e-mail or a private message differs from distributing the same image in a newsgroup, uploading it to Web-based image gallery, or making use of it as one's personal profile in an adult dating service. The crucial question then remains, how such user practices are channeled on and by different platforms, how and if these practices are incorporated in the creation of value for the platform in question, and what kinds of control users may have over the content they produce and upload.

Multiple Concerns

Bad Effects?

As both a media genre and a cultural trope, pornography is generally identified with the less positive social aspects, implications, and effects. In addition to being categorized low in the hierarchies of culture as filth and scum with no redeeming artistic or social value, pornography is seen as having negative effects on gender equality, violence against women, child development, as well as sexual identities on an individual level. Online access is regularly connected to porn addiction and studies have, for decades, aimed to point out the harmful effects of pornography on attitudes and behavior, especially those of heterosexual men toward women. Causal effects have nevertheless been hard to prove (see Malamuth et al., 2012; Hald et al., 2013). This problem is partly due to the fact that one set of cultural and social influences – such as particular pornographic images – are difficult to isolate and separate from others sources through and with which individuals make sense of the world and their place within it. Separating pornography from advertising, gaming, television entertainment, film, or social networking sites is alone difficult, not to mention attempts to separate media effects from the impact of one's family, educational background, hobbies, religious and political views, regional practices, or social networks.

Online pornography is paradoxically both ubiquitous and effaced: it is extensively debated, mundane and popular, yet something that is constantly edited and filtered out and that has been relatively disregarded as a topic of scholarly investigation until the very recent years (Paasonen, 2011b). Existing research remains patchy especially when it comes to production practices and patterns of personal consumption, given that most projects have focused on content analysis of pornographic content. The shortage of research on the topic is largely due to pornography being conceived of as a sensitive, difficult, and problematic object of study that is associated with multiple ethical and political concerns.

Children and Online Pornography

The volume and easy availability of online pornography have given rise to recurrent public concerns and occasional moral panics especially related to child pornography and children's access to materials deemed adult (e.g., Patterson, 2004: pp. 104–105). As early as in 1995, the *Times* 'cyberporn' issue liberally categorized over 80% of all photographs online as pornographic. Since only an estimated 16 million people used

the Internet globally at that point, the issue, with its inflated and unsubstantiated estimates, was effective in shaping the national imagination concerning the new medium and its contents in the United States (Chun, 2006: pp. 77–80). Due to prosecution, child pornography has nevertheless not been a visible feature on the Web and P2P networks for file sharing among users have been heavily regulated (Greenfield, 2004; Coopersmith, 2006: pp. 14–15). Child pornography is mainly distributed in closed networks and in the so-called deep web that is inaccessible to search engines and therefore excluded from mass consumption.

Child protection remains a central strand in debates on online pornography, often in connection with acts of regulation. In the summer of 2013, Prime Minister David Cameron's government in the United Kingdom announced that, in the name of child protection, all Internet service providers in the country are to block access to online porn unless the customer specifically sets the filters off. In Australia, online content is classified for similar ends. In countries where such centralized filtering is not in place, local regulation can be bypassed by hosting the illegal content on a server located in another country.

Antipornography Debates

Generally speaking, discussions on online pornography remain difficult to detach from the dualistic lines drawn along the axis of the antipornography and prosex (or anti-antipornography) advocacies, moral panics, and debates concerning censorship and freedom of speech that are characteristic to the discourse in the United States. Antipornography feminists have, since the 1970s, associated pornography with the sexual exploitation and objectification of women (see Boyle, 2010). Given that hardcore audiovisual pornography involves actual bodies having sex, antipornography feminists have defined it as documentation of sexual acts that should therefore not be protected under the freedom of speech (Mason-Grant, 2004).

Other antipornography advocates see porn as a form and symbol of decay on moral grounds: as immoral filth and scum that needs to be warded off from public consumption. Advocates for the freedom of speech have, for their part, argued for pornography as a form of fantasy, and prosex advocates have emphasized the importance of sexual self-expression and the subversive potentiality of pornography in defying bourgeois mores of good taste (e.g., Califia, 1994; Kipnis, 1996). The lines of debate tend to be polarized as being categorically for or against pornography. Both stances are equally generalizing in that they make use of pornography as an assumedly unified entity, cultural symbol, and symptom.

Pornification

A more general debate on the cultural role and social impact of pornography has recently been waged under the rubrics of the 'pornification' and 'sexualization' of culture, namely the ubiquity of pornography and sexually suggestive imagery in popular media (Attwood, 2006; Paasonen et al., 2007). In this framework, no clear distinction is drawn between hardcore and softcore pornography since the attempt is to examine the depictions of sexuality and sexual desirability across media culture, the ways in which they may have become transformed, and the cultural resonances that they may have. Such lack of conceptual specificity may, however, lead to analytical shortcomings, given that differences between the distinctly disparate aesthetics and politics that (hardcore and softcore) porn productions may hold easily disappear and the actual topic under discussion may become obscured. Diagnoses of 'pornification' may not, for example, afford sufficient distinctions to be drawn between softcore and hardcore pornographies, the public circulation of sexually suggestive imageries, and the explicitness of action shots. The demarcation line between porn and more mainstream culture remains one of acceptability and unacceptability. While popular culture broadly flirts with partial nudity, sexually suggestive poses and the aesthetics of softcore, similar mainstream appropriation does not extend to the fleshy aesthetics, control scenarios, and minute displays of bodily fluids featured in hardcore pornography. Explicit hardcore imagery is not circulated in advertising, music videos, or popular cinema.

The term pornification can also be seen to imply a process where pornography seeps into culture from the outside in whereas the process arguably works in reverse: pornography is an element of culture that is constantly being filtered off and barred from the culture properly by drawing lines between the obscene and the mainstream, the unacceptable and the acceptable. Diagnoses of pornification may then work to uphold the perspective that marks porn as other to culture, rather than framing porn as the product of a culture. Furthermore, these diagnoses may not allow for a sufficient contextual specificity in terms of different cultural and social contexts. Despite familiar, repeating emphases, debates on the cultural role of pornography and its mainstreaming in North America differ in obvious ways from those waged in Europe or Asia. When it comes to defining pornography and the dynamics of public debate related to it, North American perspectives gain much too easy predominance.

Critical Contextual Analysis

As discussed above throughout this article, the shapes and forms of available pornography have undergone drastic transformations in the course of digital production and distribution: the pornographic content has grown fragmented indeed. Digital and network technologies have opened up new forms of social circulation and exchange while also incorporating and appropriating familiar aesthetics, commodity forms, and practices of usage. Pornography can be soft or hard, airbrushed or gritty, and tender or aggressive, it is catered to users of different genders, preferences, interests, and nationalities, and it features female, male, and transgender performers alike. One key problem is that pornography is nevertheless constantly treated as an assumedly uniform category with a known, albeit much understudied, social meaning, and impact.

Studying one set of pornographic materials and their production says very little of others. Hardcore gonzo pornography of the 1990s, which profiled itself through scenarios of male domination and female submission, such as the videos of Max Hardcore, have been much critiqued for their misogyny, which has been their very key selling point. Expanding such critiques categorically to all gonzo, video productions, or

online pornography is nevertheless untenable, given that one set of productions aims to profile itself apart from the competition, and that pornographic subgenres and niches operate centrally through distinction: a queer-feminist porn production is clearly profiled apart from the 'girl-on-girl' scenarios targeted at heterosexual consumers (Ryberg, 2012) just as pornography of specific national origins can be marketed through its specificity and difference to productions from the neighboring countries. As already pointed out, the categories and tags on tube sites facilitate and encourage similar forms of distinction.

In terms of research methodology, analysis of text and image provides limited analytical possibilities for exploring or mapping out the context of production or the labor conditions, political underpinnings, ethical principles, and economic considerations involved. Amateur productions, for example, may easily be seen to connote pornography that is more ethical in its principles of production and more real, raw, and innovative in its aesthetics than commercially produced materials. Such categorizations largely ignore questions concerning power in intimate relationships and family practices: an image of domestic sexual intimacy or intensity may be uploaded for online circulation without the permission of the performer and acts may be performed under duress. Similarly, a theatrical, commercially produced scenario of dominance of submission with dramatic displays of physical discomfort may result from a completely voluntary arrangement with regulated work conditions, and with no duress involved. Categorizations or judgments concerning pornographic images or videos based on content analysis alone can tell of the aesthetics and representational politics practices at play but they should not be extended to speculations concerning production without any empirical evidence.

Conclusions

There is little doubt as to the social, cultural, and economical significance of pornography in contemporary Western culture. Pornographic imageries depict forms of sexual desiring and desirability, visceral lust, domination and submission, and they often do this through repetitive, formulaic, and stylized templates. The perennial appeal of pornography among consumers from diverse background alone speaks of its active role in shaping and contributing to cultures of sexuality. At the same time, this contribution is hard indeed to map or define in terms of precise 'impact.' The shortage of empirical ethnographic work on the consumption – as well as the production – of pornography across different national contexts sets serious limits to what scholars can claim to know about the genre and its social implications. Debates on pornography are far too often grounded in assumptions, rather than knowledge, of its meanings and effects.

The boundaries of the pornographic content have grown elastic with peer production, the flirtation with the aesthetics of pornography across popular media, and the diversification of production practices. All this necessitates reflexivity in terms of what is meant with the term 'pornography' and how it is used as point of reference. Rather than a cultural symbol or symptom, pornography needs to be seen as an umbrella term for practices, many of which are mutually contradictory and even incompatible with one another. Given the plethora of online pornography, it is impossible to identify one set of ethical, political, or aesthetic principles that would apply to all of it, or to pornography as the general marker of genre. It therefore remains crucial to resist generalization, to remain specific in one's analysis and insist on drawing distinctions between different sets of materials, which are critiques of contemporary online pornography to have analytical leverage or explanatory power.

See also: Crime, Sociology of; Internet Research: Ethical Concerns; Internet and Culture; Internet and Social Media: Anthropological Aspects; Media Ethics; Oversharing: The Eclipse of Privacy in the Internet Age; Pornography; Sex Crimes; Sexuality and Social Work; Sexuality and the Internet; Sexuality: Cultural Aspects; Teens, Gender, and Self-Presentation in Social Media.

Bibliography

Attwood, Feona, 2006. Sexed up: theorizing the sexualization of culture. Sexualities 9, 77–94.
Attwood, Feona, 2007. No money shot? Commerce, pornography and new sex taste cultures. Sexualities 10, 441–456.
Bennett, David, 2001. Pornography-dot-com: eroticising privacy on the Internet. The Review of Education/Pedagogy/Cultural Studies 23, 381–391.
Boyle, Karen, 2010. Everyday Pornography. Routledge, London.
Bruns, Axel, 2008. Blogs, Wikipedia, Second Life, and beyond: From Production to Produsage. Peter Lang, New York.
Califia, Pat, 1994. Public Culture: The Culture of Radical Sex. Cleis Press, Pittsburgh.
Chun, Wendy Hui Kyong, 2006. Control and Freedom: Power and Paranoia in the Age of Fiber Optics. MIT Press, Cambridge, MA.
Coopersmith, Jonathan, 2006. Does your mother know what you really do? The changing nature and image of computer-based pornography. History and Technology 22, 1–25.
Dery, Mark, 2007. Paradise lust: pornotopia meets the culture wars. In: Jacobs, Katrien, Janssen, Marike, Pasquinelli, Matteo (Eds.), C'Lick Me: A Netporn Studies Reader. Institute of Network Cultures, Amsterdam, pp. 125–148.
Greenfield, Patricia M., 2004. Inadvertent exposure to pornography on the Internet: implications of peer-to-peer file-sharing networks for child development and families. Applied Developmental Psychology 25, 741–750.
Halavais, Alex, 2005. Small pornographies. ACM SIGGROUP Bulletin 25, 19–22.
Hald, Gert Martin, Malamuth, Neil N., Lange, Theis, 2013. Pornography and sexist attitudes among heterosexuals. Journal of Communication 63, 638–660.
Jacobs, Katrien, 2007. Netporn: DIY Web Culture and Sexual Politics. Rowman & Littlefield, Lanham.
Johnson, Jennifer A., 2010. To catch a curious clicker: a social network analysis of the online pornography industry. In: Boyle, Karen (Ed.), Everyday Pornography. Routledge, London, pp. 147–163.
Kipnis, Laura, 1996. Bound and Gagged: Pornography and the Politics of Fantasy in America. Duke University Press, Durham.
Kotamraju, Nalini P., 1999. The birth of web design skills: making the present history. American Behavioral Scientist 4, 464–474.
Lane III, F.S., 2001. Obscene Profits: The Entrepreneurs of Pornography in the Cyber Age. Routledge, New York.
Malamuth, Neil N., Hald, Gert Martin, Koss, Mary, 2012. Pornography, individual differences in risk and men's acceptance of violence against women in a representative sample. Sex Roles 66, 427–439.
Mash, Taylor, 2004. My year in smut: inside Danni's hard drive. In: Waskul, Dennis D. (Ed.), Net.Sexxx: Readings on Sex, Pornography, and the Internet. Peter Lang, New York, pp. 237–258.
Mason-Grant, Joan, 2004. Pornography Embodied: From Speech to Sexual Practice. Rowman & Littlefield, Lanham.
Morris, Paul, Paasonen, Susanna, 2013. Coming to mind: pornography and the mediation of intensity. In: Vernallis, Carol, Herzog, Amy, Richardson, John (Eds.), The Oxford Handbook of Sound and Image in Digital Media. Oxford University Press, Oxford, pp. 549–561.

Paasonen, Susanna, 2010. Labors of love: Netporn, Web 2.0, and the meanings of amateurism. New Media and Society 12, 1297–1312.

Paasonen, Susanna, 2011a. Carnal Resonance: Affect and Online Pornography. MIT Press, Cambridge, MA.

Paasonen, Susanna, 2011b. Online pornography: ubiquitous and effaced. In: Consalvo, Mia, Ess, Charles (Eds.), Blackwell Handbook of Internet Studies. Blackwell, Oxford, pp. 424–439.

Paasonen, Susanna, Nikunen, Kaarina, Saarenmaa, Laura, 2007. Pornification and the education of desire. In: Paasonen, Susanna, Nikunen, Kaarina, Saarenmaa, Laura (Eds.), Pornification: Sex and Sexuality in Media Culture. Berg, Oxford, pp. 1–20.

Parikka, Jussi, Sampson, Tony, 2009. On anomalous objects of digital culture. In: Parikka, Jussi, Sampson, Tony (Eds.), The Spam Book: On Viruses, Porn, and Other Anomalies from the Dark Side of Digital Culture. Hampton Press, Cresskill, pp. 1–18.

Patterson, Zabet, 2004. Going on-line: consuming pornography in the digital era. In: Williams, Linda (Ed.), Porn Studies. Duke University Press, Durham, pp. 104–123.

Ryberg, Ingrid, 2012. Imagining Safe Space: The Politics of Queer, Feminist and Lesbian Pornography. Acta Universitatis Stockholmiensis, Stockholm.

Spink, Amanda, Partridge, Helen, Jansen, Bernard J., 2006. Sexual and pornographic Web searching: trend analysis. First Monday 11. http://www.firstmonday.org/ojs/index.php/fm/article/view/1391 (accessed 27.11.13.).

Onomastics

Wilhelm FH Nicolaisen, University of Aberdeen, Aberdeen, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 10859–10863, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

'Onomastics' is the intellectual endeavor which studies names of all kinds, not simply as a subdiscipline of linguistics but involving several adjacent extralinguistic disciplines. In order to establish a sound basis for name studies, it is essential to clarify the distinction between names and words or, more generally, between lexicon and onomasticon. In the naming process, transfers occur not only from the former to the latter, but also from one name category to another. As name usage is the prerequisite for name survival, socio-onomastics has become a central concern of name studies. In a largely cross-disciplinary approach, literary onomastics, i.e., the study of names in literature, has recently become a strong focus, and an active and productive activity.

'Onomastics' (from Greek *ónoma*, 'name') is the intellectual endeavor which studies names of all kinds (geographical names, personal names, boat names, names in literature, etc.). As names are embedded in, and transmitted through, language their study is often regarded as a subdiscipline of linguistics, and more precisely of lexicology, the study of words. This limited approach, however, does not do them justice, for names are not simply words with additional properties, and therefore they transcend purely lexical considerations. In this broader sense, they are also the legitimate subject of investigation by scholars working in extralinguistic disciplines, such as history, geography, sociology, psychology, philosophy, law, religion, literary criticism, and so on. It is in this wider ambience that onomastics will be treated here. The most extensive and authoritative account of names and their study is to be found in *Name Studies: An International Handbook of Onomastics* (also containing articles in French and German), Eichler et al. (1995–96).

Names and Words

The study of names, whether pursued as an integral part of, parallel to, or complementary to the study of words, has a long and distinguished history, although its terminology and the concepts it represents have led to a certain amount of confusion, a state of affairs which has not been resolved adequately even in our time. It is therefore necessary to pay close attention to the criteria which are being applied in each approach. Furthermore, in English the situation is complicated by the fact that, in addition to its indigenous term *name* (Old English *nama*, *noma*), inherited ultimately from Indo-European and therefore shared with the other Indo-European languages, it has, among others, also adopted the term *noun*, which goes back to Latin *nomen*, via Old French *nun, num*, and is now usually applied to substantives. This contrast between *names* and *nouns* is, however, not carried through systematically, and one still comes across such largely synonymous pairs as *common noun* and *proper noun* on the one hand, and *common name* and *proper name*, on the other, in which *common* and *proper* reflect classical tradition. Though the term *name* and the concept behind it are consequently at least pan-Indo-European, it is advisable to treat its descendants with care within each individual language. Such caution is, naturally, also called for in connection with name studies in non-Indo-European languages. Otherwise, any attempt at a theory of names or, more correctly, a theory about names, is destined to founder from the outset. (For further reading see Algeo, 1973; Gardiner, 1954; Pulgram, 1954.)

Semantic Distinctions

Though names and words share many characteristics, they differ in one major respect which has fundamental consequences: their semantic properties. If words are to be used competently, they must have meaning on the lexical level; otherwise they fail to serve the purpose of linguistic communication. Such a condition is essential to words of all kinds, whether they apply to concrete (*table, stone, street)* or abstract (*kindness, politics, religion*) phenomena. Meaningless words cannot fulfill their connotative function which is inclusive in its potential reference to any kind of 'table' or any variety of 'kindness.' This basic prerequisite is not expected of names which function perfectly well without any lexical meaning in what has been called 'the speech act of identifying reference' (Searle, 1969, p. 174) which is both denotative and exclusive (*Aachen, Glasgow, Paris; Karin, Maggie, Robert*).

Although most names started out as words or combinations of words that, in their time, had been transparent semantically on the lexical level, it is their capacity to function while being opaque semantically that makes them such fascinating and useful evidence for scholarly investigation, i.e., for what has come to be known internationally as onomastics. What they do need instead of lexical meaning, however, is onymic contents – this city (*Toronto*), this hill (*Bennachie*), this person (*David Livingstone*), this ship (*Titanic*), this dog (*Rover*), etc. In this context, it is essential for effective communication that there is a minimum of shared knowledge about that content, in order to avoid misunderstanding or complete failure. We need to know names, not understand them. The investigation of brand names, which has become very popular in recent years, can be said to straddle onymic and extra-onymic fields of research (Eichler et al., 1995–96, Sect. XIX).

Name Clusters and Fields

Despite the basic role of names as identifiers of individual referents, to the exclusion of all others, the number of names which can justifiably be claimed to refer to just one place, person, boat, cat, etc. is very small in our day and age. For that reason, true individuation can only be expected in limited environments (one first name in one generation of a nuclear family, one street name in a small town, one fishing boat in one home port, one horse in one stable). It is, however, a truism that no name can function alone but can only be useful in relation to other names, either in loose name clusters or, particularly, in structured onomastic 'fields' (a metaphor borrowed from physics), such as all names in a family; all, or most, street names in a town; names of all fishing boats in a harbor (and perhaps even beyond); all names of horses in a stable. Research into 'onomastic fields' parallels the study of 'lexical fields' (Schmidt, 1973) but, because of their respective differences in semantic status, is subject to different criteria (Nicolaisen, 1998).

Lexicon and Onomasticon

It is customary to think of the inventory of words in any language as a lexicon of which dictionaries are the best-known repositories. The equivalent corpus of names may be regarded as an *onomasticon* which, like a lexicon, may range from all the names embedded in any culture or language to the personal inventory of names known to one individual. Their conventional home is the gazetteer. It is when lexical items cross the threshold from lexicon to onomasticon as part of the naming process that their semantic status changes drastically. Even if, as often happens at least in the initial stages, onymic items are still lexically transparent (*Neustadt, Westhill, Rose, Smith*), the recognition of that word meaning hardly ever interferes with the (newly) acquired onomastic contents, even if these two aspects apparently contradict each other, as in the case of a *Neustadt* ceasing to be a *neue Stadt* 'new city', or of a man named *Smith* becoming a baker. In fact, a *baker* called *Baker* is likely to be an amusing anomaly rather than the prevailing norm in cultures with heritable surnames. In this sense, the long-running controversy of whether a name is in essence more of a name when it is (lexically) meaningless than if it has preserved semantic transparency on the lexical level can be demonstrated to be based on false alternatives. This judgment is supported by the realization that not every name starts out as a word or word compound with lexical meaning, because any pronounceable sequence of sounds/letters can become a name, and in some instances there are intra-onomastic transfers in a personal name becoming a place name (*Jefferson*), a place name becoming a personal name (*Pittendriech*), or a place name or personal name becoming the name of a ship (*Deutschland, Queen Mary*). Sometimes this process can be repeated, as when the English place name *Washington* first turns into a surname and then into a place name again (*Washington*, DC *Washington* State).

Onomastic Methodology

It is clear that in such transfers from one onymic category to another, or in names with non-lexical origins, a search for the original word meaning would not only be fruitless but also inappropriate. Since they are, however, in the minority, and since much onomastic research has its origins in the investigation of linguistic history, it is not surprising that etymological approaches have been, and still are, very common, especially in the study of place names (*toponymy*). This is, in principle, not very different from the lexicographical search for the derivation of words insofar as its chief requirement is the utilization of the earliest recorded spellings or forms, but nevertheless differs from lexical etymology because its aim is to reduce the name to the word(s) it once was in the act of naming: nevertheless, etymological research must be regarded as a fundamental first step in most kinds of onomastic procedures, as long as its successful completion is not also taken to be their sole goal (*see* Etymology). The many volumes of the county survey by the English Place-Name Society, for example, speak of the necessity for, and efficacy of, this method (EPNS, 1924–).

Toponymic Research

The study of place names (Eichler et al., 1995–96, Sects. XV–XVII) has for a considerable time been one of the foremost occupations of the scholar of names. The main reason for this has been the very fact that names can function without the aid of lexical meaning, whereas words cannot. This characteristic has allowed names to survive not only when the words which have formed them have died out at an earlier stage of the language that coined them, but also to outlive the very death of that language itself and its replacement by another language. In Scotland, for instance, only a handful of words has been adopted by the English language from Gaelic, whereas the maps contain many thousands of place names of Gaelic origin; the same can be said of the Slavic ingredients in the vocabulary and toponymy, respectively, of the eastern parts of Germany. In some parts of Europe and in European-derived cultures, this process of toponymic survival at times of language replacement has sometimes been repeated more than once. In the north-east of Scotland, for instance, Pictish place names have survived from preGaelic times, and it is also possible to detect a preCeltic, i.e., preGaelic and prePictish layer of Indo-European river names ascribable to the same 'Old European' linguistic stratum which, on the European continent, is linked to the Bronze Age. Some of the major island names in the Northern and Western Isles reflect the sometime, presence of speakers of an as yet unidentified non-Indo-European language. In southern Europe, multiple layering in linguistic stratification is also demonstrated by toponymic evidence.

The value of place names as evidence for this kind of investigation increases with their relative antiquity. Research among living languages has shown that in periods of linguistic and cultural contact the main process by which place names pass sequentially from one language to another is chiefly by phonological adaptation. However, partial translation, full translation, and complete replacement also occur, albeit more rarely; it is also natural that many names do not survive language change. What has emerged from this type of study is the realization that lexical dialects and onomastic dialects are not always congruent (Nicolaisen, 1980). It has also become clear that the translations of maps showing the spatial

distribution of place names into maps demonstrating chronological sequences is fraught with difficulties.

Name Transfer

The kind of toponymic palimpsest which these processes create provides a key to a relative chronology in the settlement history of speakers of relevant languages. More often than not, individual strata, or the stratification as a whole, reflects the importation of languages from elsewhere. The settlement of the New World by Europeans since the sixteenth century is a particularly striking example, since many of the names designating new locations were imported unanalyzed from elsewhere, usually without regard to their semantic fitness (Zelinsky, 1967). It can also be shown that 'colonial' settlers take with them in their mental luggage not just a lexicon of words but also an onomasticon of name models which are applied to the new terrain in the creation of landscapes reminiscent of home (Nicolaisen, 1987). One instructive instance of this is the settlement of the Northern and Western Isles of Scotland from Norway, from the ninth century onward, for without names there are no landscapes and no habitats, but only wildernesses and disorientation.

Socio-Onomastics

Initially, name scholars concentrated largely on the naming process itself and on its product, the name, but in recent decades the focus of their research has shifted more to socio-onomastic concerns, and especially to the role of names in society (Eichler et al., 1995–96, Sect. XXI). Usage and process have therefore become more central to their discussions. There are various aspects to this new direction. A major facet of place names is, in this respect, their role in politics, especially in political change, of which the emergence of a 'new Africa' is an instructive example, having prompted the creation of a large number of country names in the post-colonial era (*French Soudan* becoming *Mali*, *Gold Coast* becoming *Ghana*, *Southern Rhodesia* becoming *Zimbabwe*). Of special interest is the Republic of South Africa, in which radical internal changes have necessitated the wholesale reconstruction of its place-nomenclature, a task in which scholars with extensive expertise in onomastics have become involved (Raper, 1998). More generally, in the 'global village,' the United Nations has created the 'United Nations Group of Experts on Geographical Names' (UNGEGN) to oversee particularly the standardization of names in official usage, a move which has, as a byproduct, created a greater awareness of exonyms (*Copenhagen* for *København*, *Cologne* for *Köln*, *Aix-la-Chapelle* for *Aachen*) and their legitimate but circumscribed usage.

A long-standing experience of professional involvement by name scholars in the shaping forces toward a politically, as well as linguistically acceptable toponymy is to be found in Québec and its aspirations regarding independence (Dorion, 1998); and the sociopolitical implications of the use of place names by speakers of minority languages in countries such as Italy, France, Switzerland, Belgium, the Netherlands, Spain, Scotland, etc. have attracted the attention of name scholars (Kühebacher, 1986), as have the alternative usage of place names in informal, unofficial onomastic registers and the minutiae of microtoponymic concerns.

Personal Names

It goes without saying that the study of personal names (*anthroponymy*) is central to socio-onomastic investigations, as names are an essential component of human beings as unique individuals, and of their role in society. Just as there are no landscapes without place names, there are no societies without personal names. Parallel to etymological investigations regarding other categories of names, a certain amount of basic research has gone into establishing the linguistic origins and meanings of personal names, but the main thrust has been toward a systematic investigation of personal names as markers of human identity, traditional family pressures, or fashionable models, in various contexts. These include the naming of children; the historical development of heritable surnames from their four main sources – patronymics, occupational names, nicknames, and place names – anthroponymic typologies; the morphology of personal names; the increasing use of middle names; the overt and covert use of bynames; continuity and change in favorite names, etc. (see Eichler et al., 1995–96, Sects. XI–XIII).

Names and Education

There has been a growing tendency to introduce the discussion of many aspects of personal names into the classroom, for didactic purposes (Franz and Greule, 1999). Their own names and other people's names are of interest to children from an early age, and psychologists as well as educationalists have studied their increasing awareness of names in the development of selfhood and in the process of socialization. Personal names also form an immediate basis for children's involvement in, and contribution to, such discussions, just as investigations into local place names provide convenient inroads into local history and spatial awareness.

Literary Onomastics

The study of names in literature (Eichler et al., 1995–96, pp. 533–536) attracted an increasing number of scholars in the last few decades of the twentieth century. Its double roots in onomastics and literary criticism have made it a truly interdisciplinary pursuit. Much of the research conducted in its early phases was directed toward the role of names in individual works by individual authors. This search for name inventories as texts within texts gave, and to a certain extent still does give, literary onomastics a strong bias toward intratextual questions. However, recently there has been a conscious movement toward intertextual approaches and names as devices in intertextual relationships, or as means by which the factual world can become fictionalized, and fictive persons and places can become factualized. A special branch of this approach is concerned with the translatability of names, on both the lexical and the onomastic levels, whereas other scholars are concerned with problems of name stylistics. The role of names as devices employed in autobiographical writings in attempts to create a credible past has also attracted special attention.

Conclusion

Having started out as a subdiscipline of linguistics and functioned for a long time as a pursuit which was, because of the special status of names, admirably suited to throw light on puzzling concerns of other intellectual endeavors, onomastics has gradually abandoned this handmaiden role and has developed into an independent, though amply connected, discipline which studies names as names, for their own sake, thus establishing a foundation for their effective investigations as interdisciplinary evidence. The introduction of socio-onomastic approaches has been especially productive in this respect, but onomastics has also widened its scope and diversified its methodology in other respects, in a concerted drive toward greater sophistication and a more systematic and systemic outlook. The growing scholarly interest in matters onomastic is reflected in the existence of several academic institutes devoted to name studies and usually attached to universities or academies (in Denmark, Norway, Sweden, Finland, Belgium, Scotland, England, Northern Ireland, etc.), as well as many international, national, and regional organizations in the field, and regular congresses and conferences, often dedicated to special facets of onomastics.

Editor's Note

The origins, history and processes of proper naming have continued to attract the attention of linguists, etymologists, sociologists, philosophers and others. Perhaps the leading journal in this field of inquiry is *Names: A Journal of Onomastics*, available online at: http://www.maneyonline.com/loi/nam. Readers may also be interested in *Onomastica Canadiana* (the official journal of the Canadian Society for the Study of Names): (http://www3.csj.ualberta.ca/sco/index.php/home/onomastica-canadiana/ Representative samples of current research in onomastics include Sofia Kotilainen, 2011; John Algeo, 2010; Lars Borin and Dimitrios Kokkinakis, 2010.

A great deal of onomastics research is essentially literary and studies the naming of people and places in the works, e.g., of Tolkein, Ursula LeGuin, Doris Lessing, or the Harry Potter novels (all subjects of frequently downloaded articles in *Names*), among many others. Another onomastics obsession, particularly among those of sociological persuasion, is trends in the naming of children. See, e.g., Jean M. Twenge et al., 2010; see also Stanley Lieberson, 2000.

See also: Etymology; Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects; Lexicology and Lexicography.

Bibliography

Algeo, J., 1973. On Defining the Proper Name. University of Florida Press, Gainesville, FL. University of Florida Humanities Monographs No. 41.

Algeo, John, 2010. Is a theory of names possible? Names 58 (2), 90–96.

Borin, Lars, Kokkinakis, Dimitrios, 2010. Literary Onomastics and Language Technology. In: van Peer, Willie, Zyngier, Sonia, Viana, Vander (Eds.), Literary Education and Digital Learning: Methods and Technologies for Humanities Studies. IGI Global.

Dorion, H., 1998. La toponymie, complice involontaire de la politique?. In: Nicolaisen, W.F.H. (Ed.), Proceedings of the XIXth International Congress of Onomastic Sciences, Aberdeen, August 4–11, 1996, 3 Vols, Vol. I. Department of English, University of Aberdeen, UK, pp. 5–19.

Eichler, E., Hilty, G., Löffler, H., Steger, H., Zgusta, L., 1995–96. Namenforschung/Name Studies/Les noms propres: Ein internationales Handbuch zur Onomastik/An International Handbook of Onomastics/Manuel international d'onomastique. de Gruyter, Berlin (in three parts).

EPNS, 1924–. The Survey of English Place-Names. English Place-Name Society, Nottingham, UK (75 volumes published so far).

Franz, K., Greule, A. (Eds.), 1999. Namenforschung und Didaktik. Schneider, Hohengehren, Germany.

Gardiner Sir, A.H., 1954. The Theory of Proper Names, second ed. Oxford University Press, London.

Kotilainen, Sofia, 2011. The Genealogy of Personal Names: towards a more productive method in historical onomastics. Scandinavian Journal of History 36 (1), 44–64.

Kühebacher, E., 1986. Amtlicher Gebrauch des geographischen Namengutes. Südtiroler Kulturinstitut, Bozen, Germany.

Lieberson, Stanley, 2000. A Matter of Taste: How Names, Fashions and Culture Change. Yale University Press, New Haven.

Nicolaisen, W.F.H., 1980. Onomastic dialects. American Speech 55, 36–45.

Nicolaisen, W.F.H., 1982. Lexical and onomastic fields. In: Kazimierz, R. (Ed.), Proceedings of the Thirteenth International Congress of Onomastic Sciences, Cracow, August 21–25, 1978. Pa'nstwowe Wydawn II, Warsaw, pp. 209–216.

Nicolaisen, W.F.H., 1987. Imitation and innovation in the Scandinavian place-names of the Northern Isles of Scotland. Nomina 11, 75–85.

Nicolaisen, W.F.H. (Ed.), 1998. Proceedings of the XIXth International Congress of Onomastic Sciences, Aberdeen, August 4–11, 1996, 3 vols. Department of English, University of Aberdeen, Aberdeen, UK.

Pulgram, E., 1954. Theory of Names. American Name Society, Berkeley, CA.

Raper, P., 1998. Standardizing geographical names in a changed society. In: Nicolaisen, W.F.H. (Ed.), Proceedings of the XIXth International Congress of Onomastic Sciences, Aberdeen, August 4–11, 1996, 3 Vols, Vol. I. Department of English, University of Aberdeen, UK, pp. 264–274.

Schmidt, L., 1973. Wortfeldforschung. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany.

Searle, J.R., 1969. Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press, Cambridge, UK.

Twenge, Jean M., Abebe, Emodish M., Campbell, W. Keith, 2010. Fitting in or standing out: Trends in American Parents' choices for children's names, 1880-2007. Social Psychological and Personality Science 1 (1), 19–25.

Zelinsky, W., 1967. Classical town names in the United States: The historical geography of an American idea. The Geographical Review 57, 463–495.

Relevant Websites

Modern nation-states also take pains to preserve the unique onomastics of distinctive national names of their people and places, often via dedicated websites. For examples, see:

http://www.gla.ac.uk/schools/humanities/research/historyresearch/historyresearchcentres/scottishcelticstudies/ – Centre for Scottish and Celtic Studies (University of Glasgow).

http://www.placenamesNI.org – Northern Ireland Place-Name Project (Queen's University Belfast).

http://www.gla.ac.uk/schools/humanities/research/celticgaelicresearch/currentresearchprojects/stit/ – Scottish Toponymy in Transition (University of Glasgow).

http://www.poms.ac.uk/ – People of Medieval Scotland: 1093–1314.

http://www.logainm.ie/ – Place names Database of Ireland.

http://names.darkgreener.com/ – England and Wales Baby Names.

http://www.gla.ac.uk/schools/humanities/research/celticgaelicresearch/currentresearchprojects/commemorationsofsaintsinscottishplace-names/ – Commemorations of Saints in Scottish Place-Names: (University of Glasgow).

https://www.facebook.com/EnglishPlaceNames – The Digital Exposure of English Place-Names.

Names in Denmark (in English: http://names.ku.dk/ and Danish: http://navn.ku.dk/).

http://www.stednavneudvalget.ku.dk/ – The Danish Place-Name Commission (in Danish) Stednavneudvalget.

http://www.sofi.se/ – The Swedish Institute for Language and Folklore (SOFI).

http://www.norsknamnelag.no/ – The Norwegian Name Association Norsk Namnelag.

http://www.kotus.fi/index.phtml?l=en&s=4 – The Names Archive of Finland.

http://publish.ucc.ie/doi/locus – Hogan's Onomasticon Goedelicum (University College Cork).

https://sites.google.com/site/ghanaplacenames/ – Ghana Place Names project.

Open Education: Learning and Teaching

Eugene F Thibadeau, Indiana University of Pennsylvania, Indiana, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 16, pp. 10863–10867, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

Learning and teaching in the open classroom is not a completed theory or system of education but, instead, a related set of ideas and methods. The four main operating principles that shape the open classroom (flexible use of space, multiple learning resources, students engaging in a variety of activities, teacher working with individual children or small groups) are explicated. The teacher's role as a facilitator of learning, which is not to judge but to reflect back to the student the consequences of the student's behavior by guiding the student through questions about activities and discoveries relative to the learning experience, is detailed. The effectiveness of the teacher in the diagnosis of learning events, which is to understand the student's thinking process by listening to and questioning the student so that the student can follow their own thinking, is presented. Five learning techniques commonly utilized in the open classroom are identified. An eight-dimensional taxonomy of the teacher's role, which is widely accepted, is introduced. The current negative status of the open classroom is noted.

Introduction

The open classroom is a highly structured and deliberately organized learning environment in which the teacher performs a significant and crucial role. This kind of schooling is known by several of its manifestations: open school, integrated day, open corridor, informal education, and British infant schooling, terms used to describe a learning approach, which discards the traditional classroom setting in favor of a less formal, highly individualized atmosphere in which direct experiences are central to the learning process. The adjective 'open' recurs because this form of education is not circumscribed in classroom space, time blocks, age grouping, or even standardized curriculums. Instead, each child proceeds at his or her own pace and according to his or her own interests within a rich and carefully patterned climate in which active investigation and enjoyment of varied materials lead to the learning of skills and concepts in an integrated, rather than an isolated, manner. This departure from traditional teaching methods affects the role of the teacher as it does the behavior of the students.

Although an agglomeration of concepts from a variety of sources have contributed to open-education theory, including Rousseau and Froebel, the theoretical basis of the open classroom is found in the work of Piaget who was frequently identified in the Plowden Report as being central to an understanding of the process of learning in the open classroom. This document recognized and authenticated the informal styles of teaching and learning that were developed in English primary schools during the previous four decades and investigated by Gardner at the University of London (Horwitz, 1979).

As a result of the impact of the Plowden Report upon American teacher advocates, 50% of all schools built in the United States during the late 1960s were of open design and a similar percentage of open schools were built in Canada, New Zealand, Australia, and the United Kingdom (Bennett et al., 1980). In addition, open classrooms have appeared in Japan, Israel, Thailand, and Norway (Bennett, 1981).

Learning in the Open Classroom

The Learning Environment

The underlying assumption of open classrooms is that in an enriched and carefully planned environment that supports the natural drive toward learning, children learn in encounters with the materials and people around them at their own irregular and individual pace. Children learn most intensely when they are interested and understand the relevance of what they are doing. The open classroom represents a way of thinking, a way of living with children in a learning environment that meets the individual needs of the students, and therefore its curriculum is individualized.

Four operating principles govern the learning environment. First, an open flexible space divided into functional areas rather than one fixed homogeneous unit. Second, the environment is rich in learning resources including concrete materials, books, and other media. Third, the students are engaged in a variety of activities and levels of operation utilizing the contract-commitment method in which each student 'contracts' to do a certain amount of work, or to complete a project within a specified time limit. In this dynamic atmosphere, where learning cuts across subject lines, students are free to talk to each other, move about the room, and they have a significant degree of responsibility for their own progress. Fourth, the teacher divides the time into small, concentrated periods, working with individual children or small groups, rather than addressing the whole class as one group all day.

Open-classroom teachers are firm in their assertion that the teacher's primary responsibility is to create an environment that stimulates learning and is altered and expanded as the needs of each child change and his or her interests become more complex. The child's response to the environment becomes the starting point for learning because children learn from the things they do and understanding comes through personal experiences. In this environment, the teacher develops a uniquely close relationship with each child, as learning is dependent upon the sensitive responding of the teacher to the needs and interests of the individual student. This

cultivated skill demands that the teacher become immersed in each child's activity, identifies the moment of readiness, and gently guides the child toward direct exploration and experimentation.

Learning Events

In the diagnosis of learning events, the role of the teacher is to understand the child's thinking process. By listening and questioning, the teacher draws the child out so that the child can be assisted in realizing their feelings, plans, and details. Teacher interviews illustrate the delicacy used to determine when to intervene, to push, or even to assist. "She stretches her empathy for potential, she thinks with the child as he discusses. The 'tuning in', the thinking 'with', the seeing with a 'child's eye' are part of the response to a child's own question. It is this question which a child must follow to follow the development of his thinking, that is prime to the teacher – not her own preconceived questions and answers" (Weber, 1971, p. 229). The teacher's intention as facilitator is not to judge but to reflect back to the child the consequences of the child's behavior by guiding the child through questions about activities and discoveries relative to the learning experience (Barth, 1972). Because diagnosis occurs constantly during teacher–child interactions, errors are acceptable as they provide necessary feedback for continued, guided learning.

Through frequent note-taking describing the child's feelings, attitudes, and experiences, the teacher assesses and evaluates growth and learning. From this perspective, formal evaluation procedures are ineffective because, rather than being preoccupied with standardized and normative expectations, the teacher is vitally concerned with knowing that each child is making the best possible use of his or her environment and that he or she is making progress. Careful observation of experiences, linked with accurate recording, constitutes the scientific study of the learning process.

This effectiveness in the diagnosis of learning events depends entirely on the teacher's knowledge of the child's ability, individual needs, stage of development, and learning style. The teacher's focus is on the child who, in turn, is expected to accept both responsibility and internal personal control. Barth (1972) extends this idea by identifying the child as 'an agent of his own learning' and teachers as 'facilitators of learning' who encourage, help, and assist rather than prescribe, mold, coerce, or manipulate. As a facilitator of learning, which is a highly individual process in which teaching styles are as numerous, complex, and individually effective as learning styles, there is no one specific approach to teaching. In addition, in the multiaged open classroom, students of proven ability, together with parent volunteers and aides, assist other students who are having difficulty in assimilating the knowledge, as research has shown that children benefit from differentiated staffing in which contacts with more than one teacher is more beneficial than having the same teacher all day long. Peer tutoring, where students must think through a topic carefully because they are responsible for teaching it to other students, and collaborative learning, where students share responsibility for learning something and must organize themselves and the topic to do it well, are necessary learning activities in the open classroom.

The belief that children can sit passively and absorb information is not one of the assumptions on which learning occurs in the open classroom but, instead, children must be actively engaged in the process of their own learning. Perkins (1995), in investigating the practice of ability grouping and tracking in K-12 education, viewed the open classroom (with heterogeneously mixed grades) as a superior learning environment because research shows that students understand and retain what they study when instruction engages them in thinking about, and with the content of, their curriculum. Perkins' thesis is that learning is a consequence of thinking, and if students are not constructively engaged and connected with their thinking processes and their curricular studies, they will neither fully understand nor retain them. From his perspective, the teachers' basic job in an open classroom is to ensure that the students do a majority of thinking about and with the content of what they are learning. This perspective implies a number of learning techniques: *problem-based learning* (where students study content by seeking out the information needed to solve the problems); *project-based learning* (where students engage in complex and often socially meaningful projects that enable them to gain content knowledge); *engagement in understanding performances* (that ask students to use what they know in order to demonstrate and build their understanding); *infusion of critical and creative thinking into subject-matter instruction* (where students learn to analyze, critique, defend, adopt alternative points of view, and ask 'what if' questions); and the use of *authentic problems* (that have real-world significance and a messy open-ended character). Although thinking-centered learning can be complex and challenging for teachers and administrators to sustain, the ways to configure thinking-centered learning are endless.

A Learning Concern

There does, however, appear to be a danger that students in the open classroom may display less involvement in their activities and, consequently, learn less. Large discrepancies in time on task appear to be widespread as the findings reported by Bennett (1981) on open education in Britain are very similar to those reported by Fisher et al. (1978) in the United States. "It is of interest to note that the curriculum areas which are allocated the most time show the poorest involvement levels" (Bennett, 1981, p. 23). In addition, transition time (from one activity to another), which was studied as part of a survey in the United Kingdom (Bennett et al., 1980), observed examples of wasting time and concluded that students in open schools spend too much time in transition from one activity to the next or from one space to another. Similarly, New Zealand elementary students who were taught at tables rather than seated in rows had increased peer interaction and transition time between activities.

Teaching in the Open Classroom

Important Teacher Functions

As teachers in the United Kingdom have frequently pointed out, the teacher's primary responsibility is to set the mood or climate of the learning environment as the needs of each

student change. In this stage-setting role, the teacher functions as a stimulator who encourages and guides but does not direct children. The teacher sets up opportunities for learning experiences and children's questions become curriculum-starting points. In fact, the teacher's role is to not impose on children, but to closely cooperate with their native interests and drives so that whatever they are led to do is viewed as something that comes out of themselves. Weber (1971), who calls this 'opening up a child's purposes,' values respect for the child's questions and purposes as serious forces in learning, while other educational theorists maintain that demonstrating respect for children and managing the learning environment are key teacher functions. Doyle (1986) characterized the open-classroom teacher as an effective manager who ensures that students are concentrating on appropriate work in an environment where they have a high degree of choice.

Common to all open-classroom facilitators is an enormous trust in learners and the ability to relinquish large amounts of control. Facilitators are more effective if they themselves are involved in learning and challenge their own beliefs and practices when confronted by students. Because, generally speaking, there is no standard curriculum imposed upon the student, three teacher functions have been identified as critical roles of the open-classroom teacher: individualizer of instruction, resource person and counselor, and provider of suitable materials and activities.

Walberg and Thomas (1972) identified a wide range of beliefs and behaviors about the teacher's role, which were grouped into eight themes that are multidimensional and interactive in nature, yet discrete enough to serve as an organizer for an understanding of the teacher's role in the open classroom. This eight-dimensional taxonomy, around which clusters of significant research have been conducted, includes *instruction* (guidance and extensions of learning), *provisioning* (provisioning the classroom for learning), *diagnosis* (diagnosis of learning events), *evaluation* (reflective evaluation of diagnostic information), *humanness* (respect, openness, and warmth), *seeking* (seeking opportunities to promote growth), *self-perception* (the teacher's view of herself and her role), and *assumptions* (ideas about children and the process of learning).

Planning, Relationships, and Management Skills

Research into open-classroom processes reveals that effective teaching requires a high degree of planning, positive personal relationships, and excellent classroom management skills. The teacher creates the environment by restructuring the dimensions of time, space, materials, grouping, and communication. The classroom reflects a diversity of homemade materials to interest the students, books geared to a wide range of reading and interest levels, and a variety of carefully planned activity areas that encourage the student to explore.

Personal relationships are positive as the teacher interacts easily with students who enjoy being around him or her. The teacher is a 'real' person, cognizant of and able to express feelings, needs, attitudes, and motives. Consequently, effective open-classroom teachers exhibit characteristics of self-acceptance, self-respect, security, and self-confidence; they are encouraged to be themselves and to reject adopting the role of a teacher at the expense of being oneself.

Several important aspects of teacher classroom management, which are more vital in open than in traditional classrooms, focus on clarity of explanations, vigilance and mobility, and pupil independence. Clarity of explanations is crucial in open-classroom teaching as students ought to have a clear idea of what the project entails and the procedure for accomplishing the task. Vigilance and mobility indicates that the effective open teacher has excellent eye contact, and is able to split attention between the individual or group the teacher is working with and the rest of the class. Periodically, the teacher scans the class and is mobile enough to intervene in any problem before it escalates. This vigilance is transmitted to the class when, at intervals, the teacher publicly reviews the progress of different students or groups of students. Another message regularly transmitted is that of pupil independence so that the teacher is not overwhelmed with six or eight students gathered around the teacher's desk or trailing the teacher around the room.

Teacher Characteristics, Attitudes, and Reactions

The redefinition of the teacher's role as that of a facilitator, which encourages short-term decisions directed toward individual students, increases cooperation and communication among faculty and staff members. This research conclusion was reinforced by Harrison and Glaubman (1990) when they studied a systematic random sample of 42 Israeli elementary schools of which 17 had been implementing open-education practices for at least 3 years. Teachers in the open-education schools were significantly younger, had less experience than teachers in traditional schools, and reported having more frequent discussions of school goals and school problems with administrators, supervisors, parents, and members of the community. They used more resources for curriculum development and, as might be expected, the Israeli teachers in open schools described their school decision-making process as being more decentralized than teachers in traditional schools. Finally, Israeli teachers, who taught in schools that adopted the open-classroom approach, reported significantly more skill development, as measured by the number of hours of school-base training, than did teachers in traditional schools.

Schubert and Baxter (1982) documented a close relationship between teachers' attitudes toward open education and noncognitive student outcomes. Teachers have almost uniformly rated themselves as more open than they are and the difference is especially pronounced in the traditional classroom. When open and traditional classrooms are compared, students in an open-space school or a closed-space school (where teachers have favorable attitudes toward open-education principles) consistently scored higher with regard to level of curiosity and self-esteem than students of teachers with unfavorable attitudes toward open education. A significant sample of public school teachers, who taught in both open and traditionally constructed schools, had significantly more positive attitudes toward open education than teachers who taught only in open-space schools. Consequently, there is a positive relationship between the number of years of teaching experience and attitudes toward open education because teachers who taught 10 years or more in the open classroom were significantly more positive in their attitudes toward open education than teachers who taught less than 10 years.

Because intrinsic rewards (i.e., to play a significant role in students' lives) are important for open teachers and because the classroom is the major arena for the receipt of the rewards (Sederberg and Clark, 1990), teachers in open schools, generally speaking, show no significant inclination to move back into the self-contained classroom. However, 10 Canadian studies detailing adverse teacher comments report increased workload and insufficient preparation time, as teaching in an open classroom requires continuous innovation, monitoring of students' performance, and well-developed skills in maintaining order without being authoritarian (Traub et al., 1976). New Zealand teachers stated that increased workload was their major problem and that demands on them were very high, citing the necessity for continuous student evaluation and careful curriculum planning. This teacher reaction to the open classroom was reinforced by Australian teachers who maintained that the open-classroom workload was having a negative effect on their private lives (Angus et al., 1975).

Conclusion

Research that indicated that cognitive achievement levels in open classrooms were not significantly different from those in traditional classrooms initiated the decline of open education in the United States and by the late 1970s the movement was in retreat. Today, the literature is clearly characterized by its critics but, because widespread research has indicated that students in open classrooms are more independent, less conforming, act more cooperatively, are more attentive and participate more actively in class discussions, and display more creativity than students in traditional classrooms (Rothenberg, 1989), open-classroom learning strategies continue to be implemented in Western Europe, North and South America, and Asia as many schools continue to offer students the option of learning in open classrooms.

See also: Classroom Climate; Environments for Education; Instructional Design; Instructional Psychology; International Organisations and Education; Rousseau, Jean-Jacques (1712–78); School (Alternative Models): Ideas and Institutions; Self-Regulated Learning: Theories, Measures, and Outcomes; Situated Learning (Learning In Situ); Teacher Behaviours and Student Outcomes.

Bibliography

Angus, M.J., Evans, K., Parkin, B., 1975. An Observation Study of Selected Pupil and Teacher Behaviours in Open Plan and Conventional Designed Classrooms. Educational Department of Western Australia Technical Report 4, Australian Open Area Schools Project, Perth.

Barth, R.S., 1972. Open Education and the American School. Agathon, New York.

Bennett, N., Andrae, J., Hegarty, P., Wade, B., 1980. Open Plan Schools: Teaching, Curriculum, Design. National Foundation for Educational Research, Windsor, Berkshire, UK.

Bennett, S.N., 1981. Time and space: curriculum allocation and pupil involvement in British open-space schools. Elementary School Journal 82, 18–26.

Doyle, W., 1986. Classroom organization and management. In: Wittrock, M.C. (Ed.), Handbook of Research on Teaching, third ed. Macmillan, New York.

Fisher, C., Filby, N., Marliave, R., Cahen, L., Dishaw, M., Moore, J., Berliner, D., 1978. Teaching Behaviors, Academic Learning Time and Student Achievement. Far West Lab, San Francisco.

Harrison, J., Glaubman, R., 1990. Innovations and school improvements. Journal of Research and Development in Education 24 (1), 21–33.

Horwitz, R.A., 1979. Psychological effects of the 'open classroom'. Review of Educational Research 49 (1), 71–78.

Perkins, D.N., 1995. Smart Schools: Better Thinking and Learning for Every Child. Free Press, New York.

Rothenberg, J., 1989. The open classroom reconsidered. Elementary School Journal 90 (1), 69–86.

Schubert, N.A., Baxter, M.B., 1982. The self-contained open classroom as a viable learning environment. Journal of Education 102 (4), 411–415.

Sederberg, C.H., Clark, S.M., 1990. Motivation and organizational incentives for high vitality teachers: a qualitative perspective. Journal of Research and Development in Education 24 (1), 6–13.

Traub, R., Weiss, J., Fisher, C., 1976. Openness in Schools: An Evaluation Study. Ontario Institute for Studies in Education, Toronto.

Walberg, H.J., Thomas, S.C., 1972. Open education: classroom validation in Great Britain and the United States. American Educational Research Journal 9, 197–208.

Weber, L., 1971. The English Infant School and Informal Education. Prentice-Hall, Englewood Cliffs, NJ.

Openness to Experience

Sang E Woo, Rachel Saef, and Scott Parrigon, Purdue University, West Lafayette, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Openness to experience is best understood as a multifaceted personality trait that covers a wide range of behavioral inclinations, attitudes, and interests associated with seeking for novelty and variety. Openness and its narrow traits (e.g., intellect, culture) are meaningfully linked to intelligence, academic achievement, creativity, social/political attitudes, religiosity/spirituality, subjective and psychological well-being, and work outcomes, which are discussed in detail.

Background, Definition, and Measurement

Openness to experience (in short, openness) is most commonly recognized as one of the Big Five personality dimensions and is associated with adjectives such as 'intelligent,' 'original,' 'curious,' 'broad-minded,' 'artistically sensitive,' and 'introspective.' Until recently, empirical research involving openness has been rather fragmented, mainly due to discord regarding its exact conceptualization and measurement. The divergence in definitions is perhaps best reflected in the often imbalanced contents of frequently used personality inventories that emphasize either the intellectual *or* nonintellectual aspects of openness, but rarely both. Recent personality literature addressed this imbalance by adopting a dual label for this construct, 'openness/intellect.' Most notably, DeYoung et al. (2007) factor analyzed the Abridged Big Five Circumplex measure from the International Personality Item Pool (AB5C-IPIP) and the Revised NEO Personality Inventory (NEO-PI-R), and identified two aspects (i.e., traits at the intermediate level of specificity), which were labeled as openness and intellect. However, in order to avoid any confusion around the fact that 'openness' can be used as both an aspect of openness to experience and as a shorter name for the general openness to experience construct, this article refers to these two aspects of openness to experience as *culture* and *intellect*, respectively.

Recently, Woo et al. (2014a) introduced a three-level structural model of openness to experience (http://www1.psych.purdue.edu/~sewoo/openness.php). This model was derived from a factor analysis of 36 existing measures of openness-related scales, which yielded six facets: *intellectual efficiency* (i.e., processing novel stimuli quickly, remembering information, being knowledgeable and intellectual), *ingenuity* (i.e., mental agility in manipulating ideas or concepts, refining existing information, creating something entirely new), *curiosity* (i.e., being inquisitive, perceptive, desiring to learn about scientific principals and related topics), *aesthetics* (i.e., appreciating various forms of art, open to aesthetic experiences), *tolerance* (i.e., enjoying learning about different cultures, attending cultural events, befriending people from other cultures, immersing oneself in a foreign culture when traveling), and *depth* (i.e., desiring to gain insight into self/world and to self-improve, discussing philosophy, self-reflecting, meditating).

According to Woo et al.'s (2014a) model, openness is a broad, multifaceted construct that embodies how one typically deals with novel stimuli. The narrower traits of openness encapsulate more specific behavioral patterns adopted by individuals that describe *how* people seek out, process, and manipulate novel stimuli. Novel stimuli can appear in either an 'encoded' form (e.g., symbols and equations used in scientific theories and models) or a 'raw,' preprocessed (e.g., shapes, colors, tastes, and people) form. Accordingly, Woo et al. categorized the six openness facets into two aspects, 'openness to intellectual stimulation' of encoded stimuli (or *intellect*, in short) and 'openness to cultural experience' or raw stimuli (or *culture*, in short). In further support of this, they presented a confirmatory factor analysis result that a model with the six facets and the two aspects of openness fit the data best. Woo et al. (2014a) also introduced a measure specifically designed to capture their theoretical model, which has been validated across three distinct samples from the United States and China.

Other Models of Lower-Level Openness Structure

In their renowned lexical study, Saucier and Ostendorf (1999) factor analyzed adjectives judged to be associated with openness, and identified three facets: imagination, intellect, and perceptiveness. The imagination facet had the highest loading on the openness factor and was marked by adjectives such as creative, inventive, clever, and innovative. The types of adjectives that loaded on their intellect facet described an individual who was intelligent, analytical, and knowledgeable. Lastly, the perceptiveness facet, which had the lowest loading on openness, was marked by adjectives such as perceptive, insightful, and foresighted. Based on the content of their facets, Saucier and Ostendorf's analysis provided a facet-level structure associated almost exclusively with intellectual functioning (e.g., inventive, analytical, insightful). While extremely illuminating, this methodology was unable to capture the vitally important experiential facets of openness that we now know compose the culture aspect of the construct.

Questionnaire-based studies, on the other hand, tend to reveal a much broader configuration of the openness construct. Costa and McCrae's (1992) NEO-PI-R is a good example. Its openness measure consists of six subscales (or facets): fantasy, aesthetics, feelings, actions, ideas, and values, which are based on their theoretical perspective of highly open people as being "imaginative, sensitive to art and beauty, emotionally

differentiated, behaviorally flexible, intellectually curious, and liberal in values" (McCrae and Sutin, 2009: p. 258). One of the strengths of this taxonomy is that it more accurately captures the domain of the multifaceted nature of openness, however, this taxonomy is limited because it was developed primarily based on the authors' intuitions rather than on a thorough, empirical and theoretical basis. As noted by Costa et al. (1991), they identified six facets for each dimension of the Big Five "not because each is naturally divisible into six parts, but because at least six distinctions were suggested by the literature, and more than six scales would tax the user's ability to learn and remember the facets," (p. 888).

On the other hand, openness facets of the Abridged Big Five Circumplex scales from the International Personality item Pool (AB5C-IPIP; Goldberg, 1999) had a clear theoretical rationale in their derivation. Each facet of the AB5C-IPIP was intended to represent a combination of two 'pure,' lexically derived Big Five factors (Hofstee et al., 1992). The assumption was that, by exhausting every possible 2-factor combination of the Big Five factors one would produce the most comprehensive range of Big Five facet-level traits. As a result of this pairing, the AB5C-IPIP contains nine openness subscales: intellect, ingenuity, reflection, competence, quickness, introspection, creativity, imagination, and depth (readers interested in the actual items of these scales may refer to http://ipip.ori.org/newAB5CKey.htm). Although a compelling methodology, we note that the AB5C approach was largely based on the notion that the core of 'Factor V,' a neutral label for openness within the lexical tradition, is 'creative mentality' (Johnson, 1994: p. 311). This inevitably biased these nine facets toward the construct's intellectual side, while essentially ignoring the cultural and artistic side of openness, as well as the vast range of stimuli associated with it.

Hough and Ones (2001) took yet another approach by consulting subject-matter experts to develop the openness of taxonomy. By aggregating opinions of experts, they proposed six openness 'working taxons' (complexity, culture/artistic, creativity/innovation, change/variety, curiosity/breadth, and intellect) and further listed some of the existing scales they believed to be associated with each taxon. The advantage of this approach is that it likely reflected a diverse range of researchers' experiences, and therefore a more balanced view of openness. The limitation of this approach may be that it is difficult to check for redundancies or to estimate the degree of relations between each proposed facet in a quantifiable, systematic manner.

Lastly, using a critical incidents sorting methodology, followed by theoretical discussions among subject-matter experts and subsequent meta-analytic investigation, Connelly et al. (2014a) identified four 'true' facets of openness: aestheticism, openness to sensation, nontraditionalism, and introspection. In addition, they also found four compound traits that were thought to be a result of both openness and other Big Five personality factors: thrill-seeking and variety-seeking were identified as reflective of high levels of openness and extraversion, and low conscientiousness; openness to emotions and innovation were identified as reflective of high levels of openness, extraversion, and emotional stability. Their final list of openness facets did not include certain constructs that represent curiosity and/or typical intellectual engagement, which was attributed to the fact that the subject-matter experts who participated in the theoretical sorting process did not see that they could be distinguished from the global openness trait. The authors noted, however, that including such constructs within the construct of openness is crucial, and as such, measures specifically designed to capture such facets are necessary.

Relationships with Other Constructs

Intelligence

It is widely recognized that individuals high in openness tend to seek intellectual stimulation and activities, behaviors which encourage further development of intelligence. Consistent with this, many empirical studies have found robust, positive correlations between measures of openness and cognitive abilities. Moving beyond the relationship between broad level openness and intelligence and taking into account their lower-level structures allows for a more nuanced understanding of how these two constructs are related. Past research that considered lower-level openness traits has suggested that the intellect aspect is primarily responsible for the relationship between openness and intelligence (e.g., DeYoung et al., 2014). This notion is directly supported by a recent discovery that measures of intellect were significantly associated with measures of *g*, verbal and nonverbal ability, whereas measures of culture (which DeYoung refers to as 'openness') were only associated with verbal ability measures (DeYoung et al., 2014).

Academic Achievement

In addition to general intelligence, openness has also been repeatedly linked to other various indicators of success in academic settings. Research shows that openness is positively related to factors such as deeper approaches to learning, intrinsic motivation to succeed in college, investigative and artistic interests, college GPA, primary school grades, SAT verbal section scores, as well as higher levels of educational attainment. However, recent research has suggested that the magnitude of correlation between openness (as a broad trait) and academic performance is often fairly modest, at best.

Further, researchers are not in agreement about the underlying psychological mechanisms by which openness influences academic success. One possible argument is that the positive predictive relationship between openness and academic success can simply be attributed to the conceptual overlap between the intellect aspect and intelligence. In line with this, Heaven and Ciarrochi (2012) found that openness to experience predicted increased academic achievement, but only for those who were high in intelligence. Some even argue that the intellect aspect of openness simply reflects intelligence, and that intelligence should be considered a part of personality (e.g., DeYoung, 2014).

Creativity (*see* Personality and Creativity)

Openness has shown consistent overlap with creativity-related constructs. Woo et al.'s (2014a) facet-level taxonomy of openness includes ingenuity as one of the openness facets, which was linked to high scores on Goldberg et al.'s (2006)

IPIP Creativity scale and other measures of creative personality. In line with this finding, previous empirical research has established a positive relationship between openness and creativity (e.g., Dollinger et al., 2004; Feist, 1998; Kaufman et al., 2009; McCrae, 1987; Nusbaum and Silvia, 2011). These studies considered creativity in contexts such as organizations, workgroups, academia, art, and science, as well as in various forms such as creative accomplishments, creative behaviors and/or performance, creative/divergent thinking, and self-ratings of creativity.

The culture and intellect aspects appear to have differential relationships with creativity-related constructs. For example, DeYoung (2013) posited that culture (which he called 'openness') would be more strongly associated with artistic creativity and intellect would be more strongly associated with scientific creativity. Similarly, Nusbaum and Silvia's (2011) study found that culture predicted creativity but intellect did not, when the assessment was geared toward artistic creativity. These findings echo previous research showing that the aesthetics facet of openness (a facet within the culture aspect) had the strongest relationship with artistic creativity whereas the ideas and values facets (facets within the intellect aspect) had the strongest relationships with scientific creativity (Perrine and Brodersen, 2005).

Political and Social Attitudes

Openness is the only Big Five personality factor that has consistently and uniquely predicted political orientation, with higher scores significantly associated with a stronger endorsement of liberalism. Liberals' "preference for social change and equality reflects and reinforces motivational needs for openness, creativity, novelty, and rebelliousness" (Carney et al., 2008: p. 817). Open individuals also tend to express less support for a government restricting civil liberties of prisoners and violating their human rights (Swami et al., 2012).

Beyond linking openness with specific social and political orientation, Mondak and Halperin (2008) found that people with high levels of openness also tend to possess more political interest and knowledge, engage in political conversations more actively, be more outspoken about their political beliefs, and have high levels of internal political efficacy (i.e., perceived understanding of government systems).

Religiosity and Spirituality

McCrae (1999) once noted that openness seems to be the most relevant personality trait to the topic of religion. In general, openness and its facets appear to be differentially associated with religiosity and spirituality.

With respect to the linkage to religiosity, some studies have found that liberal and fundamentalist Protestants were significantly different from each other on their level of openness (i.e., liberals higher on openness than fundamentalists; Streyffeler and McNally, 1998), and that individuals with high openness are less likely to endorse religious fundamentalism in general (Williamson et al., 2010). Also, Dollinger et al. (1996) found that those high in openness were less likely to value *salvation* and more likely to value *self-direction*. However, this research is somewhat mixed with a few studies not finding significant correlations between measures of openness and religious orientations (e.g., Robbins et al., 2010), while other studies have found significant relationships (e.g., Streyffeler and McNally, 1998), suggesting that more systematic research is needed.

While the exact relationship between openness and religiosity is mixed, research suggests that openness does tend to be positively correlated with spirituality. Duriez et al. (2004) found that openness was related to a tendency to deal with religion in a symbolic (as opposed to a literal) way, but was largely unrelated to a tendency to accept the existence of a transcendent reality (which is characteristic of the modern definition of spirituality). Further, based on a sample of 256 students studying education in a southern area of Spain, Saroglou and Munoz-Garcia (2008) found that openness facets were differentially related to religiosity and spirituality. While spirituality and religiosity were negatively correlated with openness to actions, spirituality was positively correlated with the values, feelings, aesthetics, and fantasy facets, whereas religiosity was negatively correlated with the ideas facet. A recent meta-analysis by Saroglou (2010) adds much needed clarity to this topic: he found that while openness was unrelated to subjective religiosity it was positively related to spirituality and faith maturity, and negatively to religious fundamentalism.

Subjective (and Psychological) Well-being

There seems to be lack of consensus on how openness relates to subjective well-being. Some researchers have argued that openness to experience is a 'double-edged sword' that predisposes individuals to feel both negative and positive emotions more intensely and to experience more positive and negative emotional states (McCrae and Costa, 1991; Gonzalez-Gutierrez et al., 2005). However, a meta-analysis by DeNeve and Cooper (1998) did not find strong support for this idea; it was revealed that openness was (positively) correlated with positive affect but not significantly related to negative affect. After a decade, during which numerous studies were published in the realm of positive psychology, Steel et al. (2008) sought to update and extend the findings from DeNeve and Cooper's (1998) meta-analysis. Steel et al. (2008) showed that openness generally had a smaller correlation with subjective well-being than the other Big Five personality factors, but was positively correlated with happiness (mean $r = 0.13$), positive affect (mean $r = 0.20$), and quality of life (mean $r = 0.16$), while being largely uncorrelated with negative affect.

Another line of psychological research of well-being proposed the construct of psychological well-being. Compared to subjective well-being, psychological well-being draws more heavily upon existential issues around basic life challenges and human development (Ryff, 1985). Schmutte and Ryff's (1997) study of midlife adults found that openness significantly predicted self-reported personal growth (defined as "a sense of continued growth and development as an individual" (p. 551)), which is conceptualized as one of the major domains of human functioning, thus contributing to one's psychological well-being. Further, Keyes et al. (2002) found that those with higher levels of *psychological* well-being than *subjective* well-being tended to have higher levels of openness. Based on these findings, Keyes et al. (2002) suggested that highly open individuals'

emphasis on personal growth may come with a cost regarding negative feelings and evaluations of life.

Work Outcomes

As described earlier, openness is positively associated with cognitive abilities, creativity and innovation, curiosity/willingness to learn, and adaptability to change; it seems intuitive that openness should predict individuals' effectiveness in the workplace. Surprisingly, however, relatively little evidence has been accumulated for the validity of openness in predicting job performance in general (e.g., Barrick and Mount, 1991).

Some have attributed the lack of evidence for openness' predictive validity to conceptual and measurement issues in previous research (see the Section Background, Definition, and Measurement for a detailed discussion of this issue). Given the multifaceted nature of openness, researchers proposed that a more detailed, nuanced investigation of the openness construct and its narrow traits may yield a better understanding of its relationship with job performance (Paunonen and Ashton, 2001; Woo et al., 2014b).

Woo et al. (2014b) suggest that as the measurement specificity of openness increases (i.e., from general, aspect, to facet), stronger correlations tend to exist between the components of intellect and the outcome variables. For example, at the facet level of the intellect aspect, the ingenuity facet had the strongest association with task performance, the intellectual efficiency facet predicted turnover, and the ingenuity facet predicted leadership effectiveness and adaptive performance. However, no such relationships were found for other facets, or at the broader levels of the openness construct, demonstrating the importance of the lower-level structure of the intellect aspect.

Others have proposed that openness may only predict certain types of performance/work outcomes, and that such prediction may be further moderated by situational factors (e.g., Tett and Burnett, 2003). In line with this, studies have found that open individuals perform better in jobs that encourage new behaviors and ideas (Bing and Lounsbury, 2000), as well as in jobs which value innovation and creativity (Pace and Brannick, 2010). Additionally, people higher in openness adjust more rapidly to cross-cultural settings (Lievens et al., 2003; Shaffer et al., 2006) and jobs that have changing and diverse work environments (LePine et al., 2000). Thus, supporting the possibility that a thorough understanding of the effects of openness may require an understanding of the situational factors that are present within an individual's work environment.

On the other hand, high levels of openness may not always be a good thing. For example, some research has suggested that highly open individuals may be more prone to quitting their job than others. Zimmerman's (2008) meta-analysis showed that openness was positively related to voluntary turnover, even after controlling for prior intention to quit. Similarly, Woo's (2011) study on 'hobo syndrome' suggested that those with high levels of openness are more likely to have an attitudinal and behavioral tendency to quit (and change) jobs more frequently compared to others (i.e., display frequent quitting behavior accompanied with positive attitudes toward quitting). All in all, these findings suggest that openness likely has complex relationships with workplace outcomes.

See also: Personality Assessment: Overview; Personality and Religion; Personality, Trait Models of; Political Psychology; Spirituality; Subjective Wellbeing, Psychology of.

Bibliography

Akrami, N., Ekehammar, B., 2006. Right-wing authoritarianism and social dominance orientation: their roots in Big-Five Personality Factors and facets. Journal of Individual Differences 27 (3), 117–126.

Barrick, M., Mount, M., 1991. The Big Five personality dimensions and job performance: a meta-analysis. Personnel Psychology 44 (1), 1–26.

Bing, M.N., Lounsbury, J.W., 2000. Openness and job performance in US-based Japanese manufacturing companies. Journal of Business and Psychology 14 (3), 515–522.

Carney, D., Jost, J., Gosling, S., Potter, J., 2008. The secret lives of liberals and conservatives: personality profiles, interaction styles, and the things they leave behind. Political Psychology 29 (6), 807–840.

Connelly, B., Ones, D., Davies, S., Birkland, A., 2014a. Opening up openness: a theoretical sort following critical incidents methodology and a meta-analytic investigation of the trait family measures. Journal of Personality Assessment 96 (1), 17–28.

Connelly, B., Ones, D., Chernyshenko, O., 2014b. Introducing the special section on openness to experience: review of openness taxonomies, measurement, and nomological net. Journal of Personality Assessment 96 (1), 1–16.

Costa Jr., P.T., McCrae, R.R., 1985. The NEO Personality Inventory Manual. Psychological Assessment Resources, Odessa, FL.

Costa Jr., P.T., McCrae, R.R., 1992. Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual. Psychological Assessment Resources, Odessa, FL.

Costa Jr., P.T., McCrae, R.R., Dye, D., 1991. Facet scales for agreeableness and conscientiousness: a revision of the NEO personality inventory. Personality and Individual Differences 12 (9), 887–898.

DeNeve, K.M., Cooper, H., 1998. The happy personality: a meta analysis of 137 personality traits and well-being. Psychological Bulletin 124 (2), 197–229.

DeYoung, C.G., 2014. Openness/intellect: a dimension of personality reflecting cognitive exploration. In: Larsen, R.J., Cooper, M.L. (Eds.), The APA Handbook of Personality and Social Psychology, Personality Processes and Individual Differences, vol. 3.

DeYoung, C.G., Quilty, L.C., Peterson, J.B., 2007. Between facets and domains: 10 aspects of the Big Five. Journal of Personality and Social Psychology 93 (5), 880–896.

DeYoung, C.G., Quilty, L.C., Peterson, J.B., Gray, J., 2014. Openness to experience, intellect, and cognitive ability. Journal of Personality Assessment 96 (1), 46–52.

Dollinger, S., Leong, F., Ulicni, S., 1996. On traits and values: with specific reference to openness to experience. Journal of Research in Personality 30 (1), 23–41.

Dollinger, S.J., Urban, K.K., James, T.J., 2004. Creativity and openness: further validation of two creative products measures. Creativity Research Journal 16 (1), 35–48.

Duriez, B., Soenens, B., Beyers, W., 2004. Personality, identity styles, and religiosity: an integrative study among late adolescents in Flanders (Belgium). Journal of Personality 72 (5), 877–908.

Ekehammar, B., Akrami, N., 2007. Personality and prejudice: from Big-Five personality factors to facets. Journal of Personality 75 (5), 899–925.

Feist, G., 1998. A meta-analysis of personality in scientific and artistic creativity. Personality and Social Psychology Review 2 (4), 290–309.

Gerber, A., Huber, G., Doherty, D., Dowling, C., Ha, S., 2010. Personality and political attitudes: relationships across issue domains and political contexts. American Political Science Review 104 (1), 111–133.

Goldberg, L.R., 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In: Mervielde, I., Deary, I., De Fruyt, F., Ostendorf, F. (Eds.), Personality Psychology in Europe, vol. 7. Tilburg University Press, Tilburg, The Netherlands, pp. 7–28.

Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R., Gough, H.G., 2006. The international personality item pool and the future of public-domain personality measures. Journal of Research in Personality 40 (1), 84–96.

Gonzalez-Gutierrez, J.L., Jimenez, M.B., Hernandez, E.G., Puente, C.P., 2005. Personality and subjective well-being: big five correlates and demographic variables. Personality and Individual Differences 38, 1561–1569.

Griffin, B., Hesketh, B., 2004. Why openness to experience is not a good predictor of job performance. International Journal of Selection and Assessment 12 (3), 243–251.

Heaven, P.C.L., Ciarrochi, J., 2012. When IQ is not everything: intelligence, personality and academic performance at school. Personality and Individual Differences 53 (4), 518–522.

Hofstee, W.K., de Raad, B., Goldberg, L.R., 1992. Integration of the Big Five and circumplex approaches to trait structure. Journal of Personality and Social Psychology 63 (1), 146–163.

Hough, L.M., Ones, D.S., 2001. The structure, measurement, validity, & use of personality variables in industrial, work, and organizational psychology. In: Anderson, N., Ones, D.S., Sinangil, H.K., Viswesvaran, C. (Eds.), Handbook of Industrial, Work, and Organizational Psychology, Personnel Psychology, vol. 1. Sage Publications Ltd, Thousand Oaks, CA, pp. 233–277.

Johnson, J.A., 1994. Clarification of factor five with the help of the AB5C model. European Journal of Personality 8 (4), 311–334.

Kaufman, J.C., Waterstreet, M.A., Ailabouni, H.S., Whitcomb, H.J., Roe, A.K., Riggs, M., 2009. Personality and self-perceptions of creativity across domains. Imagination, Cognition, and Personality 29, 193–209.

Keyes, C.M., Shmotkin, D., Ryff, C.D., 2002. Optimizing well-being: the empirical encounter of two traditions. Journal of Personality and Social Psychology 82 (6), 1007–1022.

Lievens, F., Harris, M.M., Van Keer, E., Bisqueret, C., 2003. Predicting cross-cultural training performance: the validity of personality, cognitive ability, and dimensions measured by an assessment center and a behavior description interview. Journal of Applied Psychology 88 (3), 476–489.

LePine, J.A., Colquitt, J.A., Erez, A., 2000. Adaptability of changing task contexts: effects of general cognitive ability, conscientiousness, and openness to experience. Personnel Psychology 53, 563–593.

McCrae, R.R., 1987. Creativity, divergent thinking, and openness to experience. Journal of Personality and Social Psychology 52 (6), 1258–1265.

McCrae, R.R., 1999. Mainstream personality psychology and the study of religion. Journal of Personality 67, 1208–1212.

McCrae, R.R., Costa Jr., P.T., 1991. Adding Liebe und Arbeit: the full Five-Factor Model and well-being. Personality and Social Psychology Bulletin 17, 227–232.

McCrae, R.R., Sutin, A.R., 2009. Openness to experience. In: Leary, M.R., Hoyle, R.H. (Eds.), Handbook of Individual Differences in Social Behavior. Guilford, New York, pp. 257–273.

Mondak, J., Halperin, K., 2008. A framework for the study of personality and political behaviour. British Journal of Political Science 38, 335–362.

Mussel, P., Winter, C., Gelléri, P., Schuler, H., 2011. Explicating the openness to experience construct and its subdimensions and facets in a work setting. International Journal of Selection and Assessment 19, 145–156.

Nusbaum, E., Silvia, P., 2011. Are openness and intellect distinct aspects of openness to experience? A test of the O/I model. Personality and Individual Differences 51 (5), 571–574.

Pace, V.L., Brannick, M.T., 2010. Improving prediction of work performance through frame-of-reference consistency: empirical evidence using openness to experience. International Journal of Selection and Assessment 18, 230–235.

Paunonen, S.V., Ashton, M.C., 2001. Big five factors and facets and the prediction of behavior. Journal of Personality and Social Psychology 81, 524–539.

Perrine, N.E., Brodersen, R.M., 2005. Artistic and scientific creative behavior: openness and the mediating role of interests. The Journal of Creative Behavior 39 (4), 217–236.

Robbins, M., Francis, L., McIlroy, D., Clarke, R., Pritchard, L., 2010. Three religious orientations and five personality factors: an exploratory study among adults in England. Mental Health, Religion & Culture 13, 771–775.

Ryff, C.D., 1985. Adult personality development and the motivation for personal growth. In: Kleiber, D., Maehr, M. (Eds.), Advances in Motivation and Achievement, Motivation and Adulthood, vol. 4. JAI Press, Greenwich, CT, pp. 55–92.

Saroglou, V., 2010. Religiousness as a cultural adaptation of basic traits: a Five-Factor Model perspective. Personality and Social Psychology Review 14 (1), 108–125.

Saroglou, V., Munoz-Garcia, A., 2008. Individual differences in religion and spirituality: an issue of personality traits and/or values. Journal for the Scientific Study of Religion 47, 83–101.

Saucier, G., Ostendorf, F., 1999. Hierarchical subcomponents of the Big Five personality factors: a cross-language replication. Journal of Personality and Social Psychology 76, 613–627.

Schmutte, P.S., Ryff, C.D., 1997. Personality and well-being: reexamining methods and meanings. Journal of Personality and Social Psychology 73, 549–559.

Shaffer, M., Harrison, D., Gregersen, H., Black, S., Ferzandi, L., 2006. You can take it with you: individual differences and expatriate effectiveness. Journal of Applied Psychology 91 (1), 109–125.

Sibley, C., Duckitt, J., 2008. Personality and prejudice: a meta-analysis and theoretical review. Personality and Social Psychology Review 12 (3), 248–279.

Steel, P., Schmidt, J., Shultz, J., 2008. Refining the relationship between personality and subjective well-being. Psychological Bulletin 134, 138–161.

Streyffeler, L.L., McNally, R.J., 1998. Fundamentalists and liberals: personality characteristics of protestant Christians. Personality and Individual Differences 24, 579–580.

Swami, V., Nader, I., Pietschnig, J., Stiegler, S., Tran, U., Voracek, M., 2012. Personality and individual difference correlates of attitudes toward human rights and civil liberties. Journal of Personality and Individual Differences 53, 443–447.

Tett, R.P., Burnett, D., 2003. A personality trait-based interactionist model of job performance. Journal of Applied Psychology 88, 500–517.

Von Stumm, S., Hell, B., Chamorro-Premuzic, T., 2011. The Hungry mind: intellectual curiosity is the third pillar of academic performance. Perspectives on Psychological Science 6 (6), 574–588.

Williamson, P., Hood Jr., R., Ahmada, A., Sadiqcand, M., Hill, P., 2010. The intratextual fundamentalism scale: cross-cultural application, validity evidence, and relationship with religious orientation and the Big 5 factor markers. Mental Health, Religion & Culture 13 (7–8), 721–747.

Woo, S.E., 2011. A study of Ghiselli's hobo syndrome. Journal of Vocational Behavior 79, 461–469.

Woo, S.E., Chernyshenko, O.S., Longley, A., Zhang, Z., Chiu, C-y., Stark, S.E., 2014a. Openness to experience: its lower-level structure, measurement, and cross-cultural equivalence. Journal of Personality Assessment 96 (1), 29–45.

Woo, S.E., Chernyshenko, O.S., Stark, S.E., Conz, G., 2014b. Validity of six openness facets in predicting work behaviors: a meta-analysis. Journal of Personality Assessment 96 (1), 76–86.

Zimmerman, R.D., 2008. Understanding the impact of personality traits on individuals' turnover decisions: a meta-analytic path model. Personnel Psychology 61, 309–348.

Opiate Use and Abuse, History of

Gene M Heyman, Boston College, Chestnut Hill, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

For millennia, opium was celebrated for its beneficial medical effects with no mention of the dangers of addiction. In the eighteenth century, the first documented cases of opium addiction appear along with prohibitions and illegal opium markets. This happens in China, and, then, in the early twentieth century roughly similar events take place in the United States. The transition from use to abuse is accompanied by changes in self-administration techniques (e.g., smoking), drug potency, economic conditions (e.g., an increase in disposable wealth and leisure time), and individual differences (e.g., genetic-based metabolic factors).

Archaeological evidence reveals that humans have been using mind-altering substances for thousands of years, and botanical and anthropological studies suggest that human drug use may be much older, perhaps stretching back to the Paleolithic Age. The inhabitants of early Neolithic settlements left behind fossilized psychoactive plant materials, wine-stained pottery, and containers that once held cannabis and ephedra (e.g., McGovern et al., 2004; Merlin, 2003). However, fossils and pottery may not capture the true age of psychotropic drugs. The opium poppy is a cultigen that has not-known wild variants; this suggests an association with humans that predates the development of pottery. In support of the botanical evidence, anthropologists (e.g., La Barre, 1970) report that hunting societies placed more importance on drug-induced experiences than did agriculturists, suggesting that mind-altering substances were known prior to the advent of agriculture. Thus, our hominid ancestors may have been using intoxicants well before the Neolithic period. In contrast, the historical record provides no consistent evidence of drug abuse until relatively recent times. Descriptions of what we now call 'alcoholism' did not appear until the early seventeenth century (e.g., Warner, 1994), and accounts of opiate addiction first appeared in China in the eighteenth century (Spence, 1975). Thus, any mention of drug abuse comes quite late in the history of mind-altering substances. Indeed, commentary prior to the seventeenth century routinely praised opiates for their restorative powers, and early discussions of the ill effects of alcohol focused on drunkenness, with no mention of alcoholism. Other drugs have similar histories. For example, cocaine and marijuana were also once perceived as largely beneficial agents that promoted well-being. However, the story of the transition from use to abuse is best told by opiates. Opiates have long been the prototypical addictive drug, and we know much more about the history of opiates than the histories of cocaine and cannabis.

The Opium Poppy

The opium poppy is a widely cultivated annual herb that thrives in many temperate and subtropical regions of the world. Its scientific name, *Papaver somniferum*, translates as 'sleep-inducing poppy.' Although best known as a source for narcotic drugs, the plant also has nonintoxicating uses. Its seeds are tasty and can be turned into oils that are used in cooking, lighting, and painting. Its flowers are prized by gardeners and florists, ranging from white to pale blue to fire-engine red. Even its stalks serve practical ends. Some dairy farmers claim that poppy-fed cows produce the best milk for making yogurt (Booth, 1996). But, the opium poppy is most famous as the source of morphine. Morphine is nature's most potent pain killer, antiquity's most celebrated palliative, and the basis for heroin, the most notorious of all addictive drugs.

Opium

Opium is derived from the sap of the poppy's seed pod. This is a globular structure that begins to ripen once the poppy's petals have dropped off. Starting at about the size of a pea, it reaches the size of a small egg in about 10 or so days. As it ripens, the pod fills with morphine-rich sap, which can then be extracted by hand or machine. The non-mechanized method, which was universal until recently, requires only a sharp knife. The farmer scores the bulbous pod so that the sap oozes out and coagulates on the plant's surface. The sap is then scraped off, dried, and shaped into an 'opium cake.' The cake contains about 10% morphine by weight, plus a number of other psychoactive alkaloids, of which codeine is the best known.

Opium History: From the Neolithic to the Seventeenth Century

Sixth millennium BC European communities provide the earliest unambiguous evidence that the opium poppy was harvested for human use. Litter from these early settlements includes fossilized domesticated grains, the skeletons of domesticated animals, and fossilized opium poppy seeds and poppy seed cakes. Poppy seeds also show up in a structure, dating from 5700 BC, which archaeologists speculate was used for religious rites (Merlin, 2003). However, given the poppy plant's various uses, it is not possible to say definitively that Neolithic Europeans took advantage of the plant's mind-altering properties.

Representational artifacts provide the first direct evidence that opium was valued for its subjective effects. Mid-fourth

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45023-3

millennium Sumerians inscribed clay tablets with an ideogram that refers to the opium poppy and has been translated as 'joy plant.' The Ebers papyrus, dating from approximately 1550 BC, lists some 700 medicinal potions, including one for colicky babies. It instructs concerned parents to mix opium with fly droppings and administer for four days (Booth, 1996). Bronze Age Greek artifacts (1600–1200 BC) depict the opium seed capsule in association with the gods of the night, sleep, dreams, and medicine (Nyx, Hypnos, Morpheus, and Asclepius, respectively). Greek myths and Homer's epics tell of opium's narcotic properties. Theseus uses opium to put Cerberus to sleep, the many-headed dog that guards Hades. In the Iliad, which is believed to date from the eighth century BC, Helen mixes 'nepenthe' with wine to soothe the distraught survivors of the Trojan War. Scholars believe that nepenthe was a mixture that contained opium.

The early written accounts of opiates were universally positive. In classical Greek and Roman texts and tales, opium soothes pain, quiets coughs, calms fractious babies, and relieves grieving warriors of their painful memories. There is no mention of abuse or even recreational use. The positive slant could of course be a matter of selective reporting. However, classical Greek and Roman essayists and philosophers valued moderation and did not hesitate to chastise their follow citizens for drunkenness and gluttony. (In *The Republic*, Plato describes drunkenness as a selfish myopic state and likens it to the psychology of lunatics and tyrants.) Thus, it is reasonable to suppose that if even a small number of people routinely abused opium, the philosophers and orators who stressed a balanced life would have held them up for criticism.

In the following centuries, medical texts become more detailed and comprehensive, thereby opening the way for negative as well as positive comment. But in regards to opium, the narrative turns even more salutary. Paracelsus (1493–1541), the leading medical figure of the sixteenth century, refers to tinctures of opium and alcohol by the made-up word 'laudanum,' a neologism derived from *laberle*, the Latin verb for 'to praise.' Elsewhere he refers to opium as the 'stone of immortality,' suggesting that opium extends life. Several generations later, Thomas Sydenham (1624–89), who is sometimes referred to as the father of English medicine, is even more laudatory. He writes, "among the remedies which it has pleased Almighty God to give to man to relieve his sufferings, none is so universal and so efficacious as opium."

Opium Becomes a Drug of Abuse

Opium's reputation is soon to change. In the first quarter of the eighteenth century, negative comments regarding opium and opium users begin to appear. In 1729 a government prohibits opium for the first time. These events take place in China.

Opium Smoking in China

In the early seventeenth century, shops in Jakarta and South China seaports began selling mixtures of opium and smoking tobacco. Soon after, Chinese living in Taiwan and along the Fujian coast began smoking pipes filled just with opium. This was new. For the first time in recorded history, significant numbers of people were using opium for nonmedical purposes (Zheng, 2005; Spence, 1975). Also for the first time, opium became the focus of criticism. The diary entry of a soldier stationed in Taiwan in 1724 notes that opium smoking was:

> … a harmful trap, set by the barbarians in Taiwan to ensnare Han Chinese: neophytes were given free meals and free opium at first, but once they were hooked they were made to pay. Addiction was common in Taiwan, and smoking had been widespread … . (Spence, 1975, pp. 147–148)

At about the same time, a local government official adds that opium smokers 'were either the criminal or the gullible.'

These passages are remarkable. For more than 5000 years, the historical record contains nothing but acclaim for the benefits of opium – "a gift from God that extended life" – but with the advent of smoking and recreational use, it turns into a quagmire for deviants and fools.

Opium Addiction in China

The disparaging remarks were followed by the first nationwide prohibition on opium. In 1729, Yongzheng, the emperor of China, banned opium smoking. However, the ban did not stop the growth of opium smoking. Opium smoking became widespread, affecting all social strata. Historians claim that the increase was directional, starting first with the educated and wealthy (e.g., Spence, 1975; Trocki, 2002; Zheng, 2005), but Chinese government officials of the day described opium's dissemination the other way around – with the 'poor and disreputable infecting' the upper classes, particularly their children. However, there is no doubt that over the course of the eighteenth and nineteenth centuries, opium smoking in China dramatically increased. Foreign interests, particularly the British East Indian Company, exploited the demand for opium. They joined forces with local warlords to set up illegal opium markets. This led to a series of wars with Britain over the importation of opium. Britain easily vanquished the Chinese forces. The resulting peace agreements called for the emperor to accept the opium trade, cede Hong Kong and other territories to the British, and submit to political and economic concessions that together initiated 'China's century of humiliation.'

On the basis of Dutch and British trading company records, historians have estimated the size of the opium market and the prevalence and frequency of opium smoking in China over the nineteenth and early twentieth centuries. Between 1729 and 1906, tons of imported opium increased from approximately 15 000 to 26 000. Domestic production increased dramatically as well so that at the beginning of the twentieth century, the Chinese were consuming about 54 000 tons of opium a year. In many towns, opium shops greatly outnumbered tea and wine shops, and in some regions, opium was the most widely cultivated plant (e.g., Fields and Tararin, 1970; Newman, 1995; Spence, 1975; Trocki, 2002). On the basis of these figures and other historical documents, Newman estimates that in 1906 between 50% and 60% of the population smoked opium two to four times a year, about 12% smoked once or twice

a day, and about 2% smoked four or more times a day. These numbers suggest that the percentage of Chinese who would have met today's criteria for addiction in 1906 was somewhere between 2% and 12%. Other experts come up with similar estimates of overall opiate use (e.g., Spence, 1975; Trocki, 2002). To help put these numbers into perspective, there is wide agreement that the prevalence of lifetime opiate addiction in the United States has remained well below 1% for years (e.g., Anthony and Helzer, 1991; Stinson et al., 2005), whereas lifetime alcoholism rates have been steady at about 13% for some time (e.g., Hasin et al., 2007). That is, in the late nineteenth century, the prevalence of opiate addiction in China may have been equal to the prevalence of alcoholism in the United States in the early twenty-first century.

In 1799, Emperor Kia King issued a more far-reaching ban on opiates, which was accompanied by a frequently quoted rationale for the new prohibitions. A version in Latimer and Goldberg's (1981) history of opium reads:

> The use of opium originally prevailed only among vagrants and disreputable persons … but has since extended itself among the members and descendants of respectable families, students, and officers of the government. When this habit becomes established by frequent repetition, it gains an entire ascendance, and the consumer of opium is not only unable to forbear from its daily use, but on passing the accustomed hour, cannot refrain from tears or command himself in any degree. The extraordinary expense of this article is likewise to be noticed … which the fortunes of the bulk of the community are unable to satisfy, and are therefore wholly dilapidated and wasted away. (p. 106)

The emperor's account, although now over 200 years old, describes many of the key elements of addiction as it is understood by researchers and clinicians today. Frequent use leads to withdrawal symptoms, compulsive use, and a general loss of will. To make matters worse, the opium market was an economic drain on China. The middlemen traded away valuable goods (e.g., silver and porcelain) to foreign traders – which they could then sell for a profit – for a commodity that literally went up in smoke.

Why China?

For thousands of years, opium was the physician and herbalist's most revered and powerful potion. But as soon as large numbers of people began using opium recreationally, it became a vector for personal and social problems. The most obvious explanation for this turn of events is the change in the mode of self-administration. Inhalation bypasses first-pass drug metabolism, allowing the circulatory system to deliver morphine molecules to the brain in a matter of seconds. As a result a given amount of morphine is much more concentrated when it reaches the brain, thereby producing a much stronger effect. The stronger effect attracted many new users.

But smoking cannot be the sole reason that opium addiction first emerged in China. By the eighteenth century, tobacco smoking was familiar to much of the world and opium was available throughout Europe, North Africa, the Near East, and Asia. The British were the major opium dealers and were ever eager to create new markets for their vast Indian opium fields. Opium smoking could have taken hold virtually anywhere. Yet only in China did it take place on a scale that prompted government prohibitions. Similarly, merchants from many countries encountered opium smoking in Southeast Asia in the seventeenth century, yet only those from China found an eager market for this new practice when they returned home.

Many factors contribute to drug abuse, and accordingly there are a number of different accounts of why opiate abuse first emerged in China. Spence (1975) suggests that opium smoking offered relief from the boredom and frustration that was endemic to the rigid social structure of eighteenth and nineteenth century China. Zheng (2005) makes the case that Chinese cultural traditions paved the way for opium smoking. Wealthy households and businessmen offered their guests and clients an after-dinner smoke of opium along with fine teas and tobacco. Moreover, in China opium already had a link with pleasure. According to some texts it was an aphrodisiac – a belief that seems to be unique to China. Thus, an aesthetic sensibility that extolled private pleasures prepared the way for opium smoking. Trocki (2002) emphasizes economics, particularly China's advanced consumer culture. He contends that China had a highly developed market system (relative to Europe), and that Chinese consumers embraced opium as the next, new, 'had-to-have' luxury.

Opium and the Genetics of Alcohol Metabolism

These three theories are reasonable and not mutually exclusive. It is easy to imagine that individuals trapped in a rigid social structure would turn to opium, and, similarly, it is easy to imagine that a culture that emphasized sensual pleasures, particularly those that could be pursued in quiet repose, would welcome opium, and given how avidly Chinese took to tobacco smoking, it is easy to imagine that they were now looking for something new to spend money on. However, these traits are not unique to China. People everywhere seek sensual pleasures, wish they could advance to a higher social status faster, and like new things. Since opium addiction occurs first in China, there should be one or more etiological factors that are unique to China.

The genetics of alcohol metabolism differentiates China from most other countries. The key fact involves the second step in the metabolic chain. First, alcohol dehydrogenase transforms alcohol into acetaldehyde, which is toxic. Then acetaldehyde enzymes detoxify the acetaldehyde. However, some people inherit an allele that results in a slow-functioning acetaldehyde enzyme. When they drink, the acetaldehyde toxins increase, resulting in dizziness, flushing, and increases in heart rate. These effects are correlated with low rates of binge drinking and low rates of alcoholism (Luczak et al., 2001). The allele that codes for the less functional acetaldehyde enzyme is much more common in Asians than in Europeans. For instance, in a study conducted at the University of California at San Diego about 50% of the Chinese students were positive for the slow-metabolizing allele, whereas for students of European ancestry there was not one positive case (Luczak et al., 2001).

Assuming that the distribution of the acetaldehyde enzyme alleles in Chinese Americans has not greatly changed, many

nineteenth century Chinese had no easy way to get intoxicated. Alcohol made them sick. Consequently for those who found pleasure in intoxication, opium filled a niche that heretofore had no occupants.

The genetic explanation is perfectly compatible with the psychological, cultural, and economic accounts, but is different from them in that it identifies a unique feature of China, thereby making China's unique history with opium more understandable. Finally, it should also be pointed out that each of these accounts presumes a sizable population with leisure time on their hands and disposable wealth. Opium was a luxury and smoking opium to the point of intoxication typically takes at least two to three hours of time.

Summary

For millennia, opium was the herbalist and physician's most effective and celebrated treatment for medical disorders. In the seventeenth century this changed. Chinese abroad and then at home began smoking opium. The practice spread across all social strata; use turned to abuse; and government officials placed restrictions on the sale and use of opium, labeling opium a 'trap' and its users 'criminal' and 'gullible.' Smoking, a new form of self-administration, helped trigger the transformation of opium from boon to bane. However, inhalation was not a sufficient causal factor. During the eighteenth century, opium smoking could have become widespread in many places, yet only in China did opium smoking reach proportions that resulted in a thriving black market and government prohibitions. This was due to a unique combination of factors, including relatively large amounts of disposable wealth, relatively generous amounts of leisure time, cultural values that stressed private, sensory pursuits, a genetic predisposition that discouraged alcohol consumption, and weakly enforced prohibitions. European and British economic and political exploitation of China must have played a role also, but Western colonialism was worldwide, whereas the 'opium epidemic' was unique to China.

When Addictive Drugs Were Legal in the United States

Prior to the 1914 Harrison Narcotics Tax Act, addictive drugs, including heroin and cocaine, were legal in the United States. This proved fertile grounds for the patent medicine industry. Scores of companies marketed mixtures of alcohol, opiates, and cocaine, with colorful labels and grandiose claims. *Mrs. Winslow's Soothing Syrup* "soothed any animal or human," particularly colicky infants, and Dr *McMunn's Elixir of Opium* relieved "morbid irritability of body and mind." Those who had no local pharmacy to go to could send away to Sears Roebuck and other mail order companies for the drug of their choice, including morphine (which came with the promise that it would keep husbands home at night, away from the saloon). There are no data on actual sales, but on the basis of import records and historical documents, Courtwright (1982) has estimated the amounts of opium and morphine (which chemists had synthesized in the early nineteenth century) consumed by Americans over the years 1827–96. Consumption peaked in the 1890s. In this decade, Americans consumed about 3685 tons of opium each year. Thus, over the second half of the nineteenth century and for almost two decades of the twentieth, opiates were legal and not difficult to find in the United States.

Americans appear to have had as easy access to opium as did the Chinese. Indeed in China, opium was officially illegal, whereas in America patent medicines laced with opium were legal and popular. Nevertheless, opiate use in the United States remained well below the rates in China. Assuming widely accepted population estimates (400 million and 70 million people, respectively), opium consumption in China was approximately 0.2700 pounds per person, and in the United Sates it was approximately 0.0033 pounds per person. However, the per capita calculations overlook important differences in the two countries. By the second half of the nineteenth century, opium smoking in China was most common among laborers and rickshaw pullers. In the United States, smoking was but one form of opiate self-administration, and the demographic correlates were diverse, varying with the mode of consumption.

Laudanum Drinkers ('Opium Eaters')

Laudanum was the most widely used opiate. For millennia, it was the healer's most reliable antidote for coughs, pain, and distress. But in the nineteenth century, laudanum became a drug of abuse. Thomas De Quincey's famous memoir of addiction, *Confessions of an Opium Eater* (1821), featured laudanum, and physicians began to specialize in the treatment of laudanum 'inebriates.' According to reports from druggists, physicians, and newspaper reporters, laudanum users included, men, women, the aged, the young, suburbanites, urbanites, and even the very well-to-do (Brecher, 1972; Courtwright, 1982; Musto, 1973). Although it is often stated that most laudanum users were women from well-to-do families (e.g., Brecher, 1972), this claim may reflect a sampling bias. Women were more likely to go to a physician or clinic for help in dealing with laudanum addiction than were men, so that the physician reports may be biased (Ahn, 1996). However, that some laudanum users were wealthy and female was intriguing. Consequently, newspapers began portraying laudanum as a refined, sophisticated indulgence. For example, an 1881 editorial in the *Catholic World* labeled laudanum drinking an 'aristocratic vice' that is more common among the educated and wealthy, although the writer goes on to say that it spares no one (quoted in Brecher, 1972).

Although laudanum attracted all sorts, there was a common factor in how it was used. Laudanum was consumed in private, behind closed doors. As a result, laudanum drinkers were not perceived as a threat to public safety. Instead, they were perceived as ill or neurotic, and among physicians, they inspired a new medical specialty, the study and treatment of 'inebriation.'

Opium Smokers

In the third quarter of the nineteenth century, more than 250 000 Chinese men immigrated to the American West to work as laborers. They laid railroad track, dug tunnels through the Sierras, and scoured the earth for gold and silver. They were

poorly paid, ostracized by non-Chinese, and terrorized by vigilante groups and organized labor. In the labor camps and Chinatowns, opium smoking flourished. In contrast to laudanum drinking, it was a group activity that took place in discreet, hidden-away rooms ('opium dens'). A few non-Chinese Americans joined in. In contrast to laudanum drinkers, they had a distinct demographic profile. In newspapers of the day, they were described as 'evil' men, 'fallen' women – gamblers and prostitutes (e.g., Courtwright, 1982). No one would confuse laudanum drinkers with opium smokers.

Heroin Sniffers

In 1895 the Bayer Pharmaceutical Company began marketing a semisynthetic opium-based drug they called 'Heroin.' This was a made-up word that suggests the German word for heroic (*heroisch*). The ads bragged that when it came to suppressing respiratory symptoms, 'heroin had no equal.' This claim rested on fact. Heroin is much more lipid soluble than morphine, which allows it to enter the circulatory system and brain more rapidly. As a result, a given dose of heroin is about 10 times more potent than the same amount of morphine. As the morphine content of opium is about 10%, heroin is about 100 times more potent than opium. The Bayer managers also claimed that heroin was not addictive. They were not being purposely misleading. A few Bayer employees had been persuaded to try heroin. As is often the case with initial exposure to heroin, the experience was highly unpleasant. Bayer officials concluded that this ruled out addiction.

The intoxicating effects of heroin soon became known and sought out. However, little is known about the first generation of heroin users. The most frequently cited reference is an 'eyewitness' account by Pearce Bailey (1916), a physician and New York State public health official who specialized in 'defectives.' According to Bailey, 'heroin sniffers' were young men who hung out in gangs in East Coast urban centers. They had quit school, rarely worked, and were often in trouble with the law. For these young men, heroin functioned as a badge of identity, signifying rebelliousness, and disdain for conventional employment. At the end of his article, Bailey proposed that state officials relocate the 'heroin boys' to a rural setting and allow the restorative powers of agricultural work to do its magic.

Did Addiction Increase in Epidemic Proportions?

Given America's history of prosecuting drug use, it is reasonable to suppose that drug use grew at epidemic proportions prior to the passage of the Harrison Narcotics Tax Act. However, historians estimate that there were never more than 300 000 opiate addicts prior to 1914 (Courtwright, 1982; Musto, 1973). This implies an overall prevalence of less than half of 1% (0.0043). This is markedly less than even the lowest estimates for China in the same period (e.g., 2%). The pre-prohibition rate is also within today's rates. In the most recent and largest national survey (Conway et al., 2006), lifetime frequency of opiate dependence was 0.3% and in the Epidemiological Catchment Area survey of the early 1980s (Anthony and Helzer, 1991), lifetime frequency of opiate abuse and/or dependence was 0.7%.

There is of course much uncertainty about these estimates, particularly those from a hundred years ago. Historians have to rely on indirect measures, such as pharmacy sales and taxes on imported drugs. Today's researchers have the advantage of a scientific research tradition and well-funded projects that enlist the help of trained interviewers. Nevertheless, the old and new survey results are in rather close agreement. The simplest explanation is that the lifetime prevalence of opiate addiction in the United States at the end of the nineteenth century and the end of the twentieth century were not that different, somewhere between about 0.3% to perhaps 0.7%, and likely closer to the 0.3% mark. The broader implication is that legal restrictions were not the only curb on opiate addiction in the United States. In support of this point, opiate use and addiction peaked in the mid-1890s, well before the Harrison Narcotics Act.

The Response to Opiate Use in the United States

Although there is no evidence of an addiction epidemic prior to the Harrison Act, opiate use did elicit interest and concern among physicians, social reformers, and politicians. A number of physicians began to specialize in the study and treatment of addiction, the Temperance movement added opium addiction to their list of concerns, and legislators began passing laws to limit the use of opiates.

Physicians

In the late 1860s, a handful of physicians founded a new specialty, the treatment and study of 'alcohol and drug inebriety' or what we would now call alcoholism and addiction. They built treatment facilities, founded the American Association for the Cure of Inebriety in 1870, and in 1876 began publishing the *Journal of Inebriety*. The editor and contributors were physicians, and the one common theme of the published papers was that alcohol and opiate inebriety were diseases (Weiner and White, 2007). However, these early addiction specialists met stiff and eventually overwhelming opposition. As they put it, the claim that addiction was a disease was met with 'hysterical denunciations.' The prevailing view was that 'inebriates' got drunk or high because they wanted to not because they had to (Levine, 1978).

The Legislative Response to Opium Addiction

The Association for the Cure of Inebriety was at loggerheads with a powerful grassroots temperance movement. Concerned citizens, clergy, politicians, and many physicians joined forces to end drunkenness, addiction to patent medicines, opium smoking, heroin sniffing, and saloons. Their efforts won wide support. In 1906 President Theodore Roosevelt signed into law the Pure Food and Drug Act, which called for patent medicine manufacturers to label their bottles with the warning 'May Be Habit Forming' and a list of the amounts of opium, cocaine, and alcohol. According to Musto (1973), the labeling act reduced patent medicine purchases by as much as one-third. Then in 1914 came the legislation that transformed American drug use and set the terms for American drug policy ever since.

The Harrison Narcotics Tax Act limited the use of opiates and cocaine to the practice of medicine. The legislation did not mention addiction nor define medical practice, but focused on the regulation of trade. However, the justice department targeted all use except for the relief of pain in medical settings (Courtwright, 1982; Musto, 1973). Doctors who had been prescribing opiates to addicts stopped doing so, and with few exceptions (e.g., cough syrups containing small amounts of codeine) Americans could no longer buy opiates or cocaine at their local pharmacy. Indeed using opiates and cocaine without a prescription was now a criminal act. Drug users were arrested and sent to jail. In a sign of the times, the Association for the Cure of Inebriety stopped publishing their journal in 1914, and in a few years the medically staffed sanatoriums for the treatment of addiction were a relic of the past (Weiner and White, 2007).

The Consequences of the Harrison Narcotics Tax Act

Following the Harrison Act, laudanum drinkers and opium smokers all but disappeared. Heroin use persisted, but because it was illegal, it became even more closely tied to criminal activity. Criminal gangs took over heroin's distribution, adulterated it with inert substances, and raised prices. Users switched to injecting heroin in order to get the same kick that snorting had provided. Apparently, laudanum drinkers were unwilling to inject themselves with a substance that by weight was about a hundred times more powerful than opium. Street addicts now included repeat offenders who had committed serious crimes. Among the first to document the transformation of opiate use in America were Lawrence Kolb and A.G. Du Mez, physicians who worked for the Public Health Service and specialized in addiction. They characterized the demographic consequences of the Harrison Act in the following words (1924):

> addiction is becoming more and more a vicious practice of unstable people, who, by their nature, have abnormal cravings which impel them to take much larger doses that those which were taken by the average person who so often innocently fell victim to narcotics some years ago. Normal people now do not become addicted or are, as a rule, quickly cured, leaving as addicts an abnormal type with a large appetite and little means of satisfying it. (p. 1191)

Put another way, the middle-class laudanum drinkers disappeared, so that the remaining addicts were hardened versions of Bailey's 'heroin boys.'

Lessons

The American Experience

Although opiates and cocaine were legal and widely available in the last half of the nineteenth century and early twentieth century, there was no opium epidemic, as in China, and the prevalence of opiate use and addiction may have been about what it is today. The simplest explanation is that for most Americans intoxication was usually not a particularly desirable state. For instance, intoxication interferes with the business of life, particularly conventional activities, so that most of the time it would be impractical to take heroin or other intoxicating drugs. Second, opium smokers and heroin sniffers included a disproportionate number of individuals who were unemployed, undereducated, and involved in illegal activities. Thus, the claim that drug prohibitions turned addicts into criminals is at best a simplification of the facts. Third, antidrug legislation reduced drug use in those who by today's standards would have been considered addicts. However, the legislative effects were selective. The Harrison Act virtually eliminated opium smoking and laudanum drinking, whereas heroin use persisted. Events much later in the century reveal that these differences were not strictly a matter of heroin's powerful pharmacological actions. Quit rates for heroin addicts are about the same as for other drugs (Anthony and Helzer, 1991; Stinson et al., 2005), and the correlates of quitting include the ordinary pressures of adult life, such as economic well-being, familial responsibilities, and the opinions of others (Waldorf, 1983), which is also the case for other drugs (e.g., Heyman, 2009).

The Transition from Use to Abuse

Laboratory psychopharmacology research shows that the behavioral effects of psychotropic drugs invariably reflect the setting and individual differences. For example, in rat studies the same dose of amphetamine can either increase or decrease reinforced lever pressing as a function of differences in the animals' reinforcement history. The history of the transition from opiate use to opiate abuse involves the same sort of contextual influences. Americans had access to opiates that were 10 (morphine) and 100 (heroin) times more potent than the opium available in China, yet the rates of opiate addiction in China were 10–20 times higher than in the United States. Conversely, the prevalence of alcoholism is substantially lower in China than in the United States, yet the United States has put more road blocks in the way of drinking (e.g., Xiang et al., 2009). Thus, to make sense of addiction it is necessary to take into account characteristics of the setting (e.g., legal prohibitions and cultural traditions), characteristics of the individuals (e.g., the genetics of drug metabolism), and characteristics of the pharmacological experience (e.g., the transition from drinking tinctures of opium to smoking opium).

See also: Drugs, Decriminalization of: Cultural Concerns; Drugs: Illicit Use and Prevention.

Bibliography

Ahn, C.H., 1996. Nineteenth-Century Morphine Addiction: A Socially Constructed Phenomenon (thesis). Harvard University, Cambridge, MA.

Anthony, J.C., Helzer, J.E., 1991. Syndromes of drug abuse and dependence. In: Robins, L.N., Regier, D.A. (Eds.), Psychiatric Disorders in America: The Epidemiologic Catchment Area Study. Free Press, New York, pp. 116–154.

Bailey, P., 1916. The heroin habit. New Republic 6, 314–316.

Booth, M., 1996. Opium: A History. Simon & Schuster, London, New York.

Brecher, E.M., 1972. Licit and Illicit Drugs; the Consumers Union Report on Narcotics, Stimulants, Depressants, Inhalants, Hallucinogens, and Marijuana – Including Caffeine, Nicotine, and Alcohol, first ed. Little, Brown, Boston.

Conway, K.P., Compton, W., Stinson, F.S., Grant, B.F., 2006. Lifetime comorbidity of DSM-IV mood and anxiety disorders and specific drug use disorders: results from the National Epidemiologic Survey on Alcohol and Related Conditions. Journal of Clinical Psychiatry 67, 247–257.

Courtwright, D.T., 1982. Dark Paradise: Opiate Addiction in America before 1940. Harvard University Press, Cambridge, MA.

Fields, A., Tararin, P.A., 1970. British Journal of Addiction 64, 371–382.

Hasin, D.S., Stinson, F.S., Ogburn, E., Grant, B.F., 2007. Prevalence, correlates, disability, and comorbidity of DSM-IV alcohol abuse and dependence in the United States: results from the National Epidemiologic Survey on Alcohol and Related Conditions. Archives of General Psychiatry 64, 830–842.

Heyman, G.M., 2009. Addiction: A Disorder of Choice. Harvard University Press, Cambridge, MA.

Kolb, L., Du Mez, A.G., 1924. The prevalence and trend of drug addiction in the United States and factors influencing it. Public Health Reports 39, 1179–1204.

La Barre, W., 1970. Old and new world narcotics: a statistical question and an ethnological reply. Economic Botany 24, 73–80.

Latimer, D., Goldberg, J., 1981. Flowers in the Blood: The Story of Opium. F. Watts, New York.

Levine, H.G., 1978. The discovery of addiction. Journal of Studies on Alcohol 39, 143–174.

Luczak, S.E., Wall, T.L., Shea, S.H., Bun, S.M., Carr, L.G., 2001. Binge drinking in Chinese, Korean, and white college students: genetic and ethnic group differences. Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors 15, 306–309.

McGovern, P.E., Zhang, J., Tang, J., Zhang, Z., Hall, G.R., Moreau, R.A., Nuñez, A., Butrym, E.D., Richards, M.P., Wang, C., Cheng, G., Zhao, Z., Wang, C., Bar-Yosef, O., 2004. Fermented beverages of pre- and proto-historic China. Proceedings of the National Academy of Sciences of the United States of America 101, 17593–17598.

Merlin, M.D., 2003. Archaeological evidence for the tradition of psychoactive plant use in the old world. Economic Botany 57, 295–323.

Musto, D.F., 1973. The American Disease: Origins of Narcotic Control. Oxford University Press, New York.

Newman, R.K., 1995. Opium smoking in late imperial China: a reconsideration. Modern Asian Studies 29, 765–794.

Spence, J., 1975. Opium smoking in Ch'ing China. In: Waksman, F.E., Grant, C. (Eds.), Conflict and Control in Late Imperial China. University of California Press, Berkeley, CA.

Stinson, F.S., Grant, B.F., Dawson, D., Ruan, W.J., Huang, B., Sasha, T., 2005. Comorbidity between DSM-IV alcohol and specific drug use disorders in the United States: results from the National Epidemiological Survey on Alcohol and Related Conditions. Drug and Alcohol Dependence 80, 105–116.

Trocki, C., 2002. Opium as a commodity and the Chinese drug plague. In: Hanna, B., Woodley, E., Buys, E., Summerville, J. (Eds.), Social Change in the 21st Century Conference Proceedings: Refereed Conference Papers Brisbane, Queensland.

Waldorf, D., 1983. Natural recovery from opiate addiction: some social-psychological processes of untreated recovery. Journal of Drug Issues 13, 237–279.

Warner, J., 1994. 'Resolv'd to drink no more': addiction as a preindustrial construct. Journal of Studies on Alcohol 55, 685–691.

Weiner, B., White, W., 2007. The Journal of Inebriety (1876–1914): history, topical analysis, and photographic images. Addiction 102, 15–23.

Xiang, Y., Ma, X., Lu, J., Cai, Z., Li, S., Xiang, Y., Guo, H., Hou, Y., Li, Z.B., Li, Z.J., Tao, Y., Dang, W., Wu, X., Deng, J., Lai, K.Y.C., Ingvar, G.S., 2009. Alcohol-related disorders in Beijing, China: prevalence, socio-demographic correlates, and unmet need for treatment. Alcoholism: Clinical and Experimental Research 33, 1111–1118.

Zheng, Y., 2005. The Social Life of Opium in China. Cambridge University Press, New York, Cambridge, UK.

Opinion Formation, Theory of

Shawn W Rosenberg, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 10876–10879, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Public opinion is a contested concept. The nature of the 'public,' the political concepts and values that constitute 'opinion,' and the process whereby public opinion is formed are all matters of debate. In the dominant view, the public is a composite of individual citizens and opinion consists of the beliefs that are subjectively associated with an attitude object. The process of public opinion formation therefore revolves around individuals. Empirical research adopting this view focuses on information processing and thus examines the role of perception, reasoning, and memory on attitude formation. The conclusion is that people generally do not attend to, understand, or care about politics. What attitudes they express are typically unstable, inconsistent with one another, and do not reflect any overarching ideological point of view. Public opinion formation is also understood from three other perspectives. The first retains a psychological focus but takes a more holistic view of political meaning and examines qualitative differences in how individuals represent, understand, and judge experience and information. The second shifts the focus from individuals to discourse and considers how political concepts and values are constructed through democratic deliberation. The third perspective views publics in holistic terms and argues that public opinion is an irreducibly collective phenomenon, one that is shaped by commensurately large-scale social forces.

The concept of 'public opinion' has been debated across disciplines and centuries. At issue is the nature of the 'public,' the political concepts and values that constitute 'opinion,' and the process whereby the public opinion is formed. One key point of division is epistemological: whether to view political meaning as an aggregate or a structured whole. According to the aggregative view, political meaning consists of a set of elements (representations of experience), which are related to one another either through passively learned association or through active mental calculation. The alternative view suggests that political meaning is constructed. How it is constructed delimits the nature of the representations that may be formed and the quality of the associations or calculations that can be made.

A second divide is more explicitly political. The issue here is where to locate public opinion and, by implication, where to look in order to explain it. One view assumes the public is a composite phenomenon, an aggregate function of its individual citizens. In this view, political meaning and value are attributes of individuals. Analysis therefore focuses on the qualities of cognition and, in this light, how individuals process the experiences and social influences to which they are exposed. The opposing view presumes that the public is a collective entity, a corpus of its own. Here political meaning and value are emergent properties of the collective. They are viewed as essentially cultural phenomena and are explained with reference to large-scale social forces.

The Mainstream Perspective: Publics and Meaning as Aggregates

Although these debates remain central in contemporary political theory, a consensus emerged in the empirical study of public opinion formation. Its basic contours were defined by Allport's (1935) seminal discussion of attitudes, the Yale persuasion studies (Hovland et al., 1953), and Converse's (1964) analysis of political belief systems. Although pursuing different topics from different disciplinary perspectives, the research is oriented by the same political and epistemological vision: the public is an aggregate of its individual citizens, and political meaning is a matter of calculated or learned associations among representations and preferences. In this light, the formation of public opinion is understood to be a process that revolves around individuals. It begins with their exposure to politically relevant experiences and information. Each individual processes this information, thereby coming to a judgment that yields an attitude. The attitudes of different individuals are then aggregated, either through informal interactions or more formal mechanisms, such as elections or opinion polls.

For the most part, this research effort has focused on the second step in the opinion formation process, the social psychology of how individuals process information (either from direct or indirect experience) when forming an attitude. The critical outcome or dependent variable is an attitude. Typically, this is defined as an individual's overall evaluation of an object with some degree of favor or disfavor. In this view, an attitude consists of three types of elements: cognitive (beliefs); affective (emotions or feelings); and behavioral (intent to act). These elements can affect one another. While they are usually consistent, they may not be. Attitudes are also related to one another.

In a way that has important consequences for attitude strength and stability, this interattitudinal structure may vary with regard to its elaboration, its coherence, and the degree to which it has a hierarchical structure. A central aim of the research is to examine how different conditions (both situational and dispositional) affect the processing of information and thus the resulting attitude. The research is designed on the presumption that all individuals operating under similar conditions will process information in the same way. Consequently subjects' responses are aggregated, and inferences are made regarding how the 'average' person will form, maintain, or change their opinions under the conditions examined.

Adopting this general framework, a wide variety of competing theories or models of public opinion formation were presented from the 1950s to the 1980s. Some posit a rational, almost logical, integration of information. A prominent example is the value-expectancy model (e.g., Fishbein and Ajzen, 1975). According to this model, an individual forms or changes an attitude by reviewing their beliefs about the attitude object. A weight is assigned to each belief by multiplying its assessed value by the probability that the belief correctly pertains to the object in question. The expected values of all the beliefs are then averaged, leading to an overall attitude or judgment being formed.

Other models accept the basic features of such rational computational models but focus more on considerations of memory and its effect on what beliefs will be considered. The suggestion here is that an individual may have a diverse set of contradictory beliefs or affects associated with an attitude object. Situational and internal cues affect which of these beliefs is likely to be remembered on any given occasion. Consequently, although individuals may process beliefs rationally, the ways in which memory functions will bias which beliefs are processed and therefore what judgment is finally made. How mass media prime candidate evaluations and the apparent instability of political opinions are readily explained in these terms (e.g., Zaller, 1992). On-line processing models of political candidate evaluation de-emphasize the importance of recall (e.g., Lodge and Steenbergen, 1995). In this view, individuals do not judge an object by recalling all relevant beliefs. Rather, they simply remember their prior evaluation of the object and adjust this in light of the new information provided. While this information may then be forgotten, the overall assessment is carried forward when further judgment is required.

Other models of information processing suggest that judgment is more irrational. One example is the family of cognitive consistency models which suggest that individuals relate objects to one another, not only on categorical or logical grounds, but also on the basis of evaluation or feeling. A particularly influential extension of this is Tajfel's (1981) analysis of the consequences of group identification. His research suggests that a person's desire to maintain self-esteem (positive evaluation) and their negative identification with an out-group (categorical link) will result in negatively biased evaluations and behaviors toward that group. Following a learning theory tradition, other research demonstrates how regularly pairing an unconditioned stimulus (an object already valued positively or negatively) with a conditioned stimulus (an object not valued) will shape the subsequent evaluation of the condition stimulus.

In recent years, there has been a shift to more eclectic dual processing models of attitude formation. Self-consciously integrating earlier conceptions, these models suggest that different processes will operate under different conditions. Two types of processes are distinguished. Central processes, such as the value expectancy model, are relatively demanding and involve the evaluation of relevant arguments and information. Peripheral processes are less demanding than central processes and include classical conditioned responses, reliance on peripheral situational cues (e.g., physical attractiveness of the source), or the use of simple heuristics (e.g., experts know what they are talking about). The focus of this research is to discover (1) what conditions foster what type of processing and (2) how different kinds of processes affect attitude formation under different conditions. This has led to a reconsideration of the early research on the persuasive effect of source variables (e.g., credibility, expertise), message variables (e.g., argument quality, personal relevance), recipient variables (e.g., prior attitude, personality), and context variables (e.g., distractions, audience reaction). The net result is a more complex and contingent understanding of how attitudes are formed, sustained, and changed (Petty and Wegener, 1998).

Political scientists have drawn upon these models in their attempt to understand opinion formation in mass publics. Thus they have focused on who is presenting what information (the agents of socialization – family, workplace, school, social groups, and the mass media), to whom, and with what effect on individuals' judgment and knowledge of political leaders, issues, and events. The research on agent effects on opinion formation suggest that whereas the family is not as influential as initially believed, it does provide individuals with one of their most enduring characteristics: their political identity. Supporting early survey research, contemporary work indicates that group identification can influence powerfully individuals' attitudes toward issues, especially those emphasized by the group. Research also indicates that the mass media, while not the powerful persuader it was initially feared to be, nonetheless is able to affect people's sense of political priorities and to prime the specific considerations they may bring to bear when judging a particular issue or candidate.

In the case of knowledge and sophistication, the evidence suggests overwhelmingly that the vast majority of people have little accurate information about political leaders, current legislation, or foreign affairs. Nor do their judgments reflect much sophistication or consideration. What attitudes they express do not cohere in any apparent ideological fashion, nor are they particularly reliable over time. Even slight changes in how questions are posed can produce different attitudinal responses. Some studies do show correlations between apparently abstract and more concrete attitude items. This result is interpreted to suggest people do base particular judgments on more abstract core values. However evidence from open-ended interviews indicates that whereas a few people do think about politics in this deductive way, most do not. Their responses are generally ill-considered, unjustified, and readily changed by new considerations that pop to mind. Overall, the evidence suggests that citizens lack the information, understanding, or considered judgment typically called for in democratic political theory.

Alternative Direction: Constructivism, Discourse, and Collective Publics

Questions have been raised regarding the basic assumptions orienting this dominant approach to the study of public opinion. The results of open-ended interviewing have sparked an increasing interest in taking a more constructive, holistic view of political thinking. This has lead to a renewed interest in constructivist psychologies that consider basic differences in how individuals reason (e.g., Rosenberg, 2002). In this view,

how an individual thinks about social or political events will determine the kinds of interrelationships and the kinds of representations (and interests) they can construct. Moreover, different individuals may think in qualitatively different ways.

For example, some people tend to focus on a simple causal or categorical relationship cued by the situation when formulating an understanding, and they draw primarily on social convention or group interests to make their judgments. Other people focus on only the concrete contours of the immediate situation and draw primarily on present feelings to make a judgment. Still others infer quite abstract issues that are only implicit in the situation and judge them with reference to some notion of the general interest or abstract principle. Contrary to the common assumption of the 'average citizen,' this work suggests that people may differ significantly, both in how they are affected by the same social influences and in their capacity to respond to different kinds of institutional regulation.

This interest in reasoning dovetails with the dissatisfaction expressed by some political theorists regarding representative democracy and its emphasis on electoral mechanisms of preference aggregation. Instead, they advocate more deliberative forms of democracy that require individuals to cooperate in the generation and resolution of political issues (e.g., Habermas, 1996). In this light, considerations of the requirements of citizenship shift. Rather than personal judgment, the focus is on cooperation with others in order to come to common understanding and judgment. This involves taking the perspective of another, relating it to one's own, and then critically considering both. The latter not only entails compromise, but the self-conscious construction of new positions that incorporate and go beyond the original ones. Research motivated by these concerns seeks to determine how differing institutional arrangements impact on deliberation and with what effect on the collective judgments that ensue. It does so by eschewing a focus on individuals in favor of a direct examination of the quality of deliberations themselves and the social representations and group opinions they produce.

Finally, there is renewed interest in a more holistic, constructivist view of the public. Rejecting what is referred to as methodological individualism, social theorists have argued that consciousness and meaning making are essentially collectively structured phenomena that are best thought of as a culture or mode of discourse. Some view culture as singly structured and coherent, and others argue that it is essentially an arena of conflict. However, all agree that the individual's understanding is a generally poor reflection of a larger social process and an inappropriate object of inquiry. In this more sociological vision, changes in public opinion reflect changes in the underlying structure of collective meaning. Examples include the shift in public discourse from appeals to the sacred to reasoned argument regarding the profane, or a population coming to imagine itself as a nation (e.g., Anderson, 1991). Such changes are explained with reference to collective forces such as the form of the relations of production, the mode of information, or a new geographical positioning of a population. With such changes, alternative forms of institutional regulation become possible (e.g., rule by personal power vs rule by law) and new kinds of political values are constructed (social tradition vs principled critique). The challenge to this line of inquiry is to develop an empirical methodology as sophisticated as the theorizing.

Editor's Note

Recent works of interest on the processes of opinion formation in mass publics include: Druckman et al., 2013; Leeper and Slothuus, 2014; Gelpi, 2010; Galam, 2013; Fatas-Villafranca et al., 2011.

See also: Attitude Formation and Change; Attitude Measurement; Attitudes and Behavior.

Bibliography

Allport, G.W., 1935. Attitudes. In: Murchison, C.A. (Ed.), Handbook of Social Psychology. Clark University Press, Worcester, MA.

Anderson, B.R.O.'G., 1991. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London.

Converse, P.E., 1964. The nature of belief systems in mass publics. In: Apter, D.E. (Ed.), Ideology and Discontent. Free Press, New York.

Druckman, James N., Peterson, Erik, Slothuus, Rune, 2013. How elite partisan polarization affects public opinion formation. American Political Science Review 107 (1), 57–79.

Fatas-Villafranca, Francisco, Saura, Dulce, Francisco, J., Vazquez, 2011. A dynamic model of public opinion formation. Journal of Public Economic Theory 13 (3), 417–441.

Fishbein, M., Ajzen, I., 1975. Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research. Addison-Wesley, Reading, MA.

Galam, Serge, January–April 2013. Modeling the forming of public opinion: an approach from sociophysics. Global Economics and Management Review 18 (1), 2–11.

Gelpi, Christopher, 2010. Performing on cue? The formation of public opinion toward war. Journal of Conflict Resolution 54 (1), 88–116.

Habermas, J., 1996. Between Norms and Facts. MIT Press, Cambridge, MA.

Hovland, C.I., Janis, I.L., Kelley, H.H., 1953. Communication and Persuasion: Psychological Studies of Opinion Change. Yale University Press, New Haven, CT.

Leeper, Thomas J., Slothuus, Rune, 2014. Political parties, motivated reasoning and public opinion formation. Politial Psychology 35 (1), 129–156.

Lodge, M., Steenbergen, M., 1995. The responsive voter: Campaign information and the dynamics of candidate evaluation. American Political Science Review 83, 309–326.

Petty, R.E., Wegener, D.T., 1998. Attitude change: mutiple roles for persuasion variables. In: Gilbert, D.T., Fiske, S.T., Lindzey, G. (Eds.), Handbook of Social Psychology, fourth ed. vol. 1. McGraw-Hill, New York.

Rosenberg, S.W., 2002. The Not So Common Sense: Different Ways People Judge Social and Political Events. Yale University Press, New Haven, CT, pp. 323–390, Chapter 8.

Tajfel, H., 1981. Human Groups and Social Categories: Studies in Social Psychology. Cambridge University Press, Cambridge, UK.

Zaller, J.R., 1992. The Nature and Origins of Mass Opinion. Cambridge University Press, Cambridge, UK.

Oppression

Katherine van Wormer, University of Northern Iowa, Cedar Falls, IA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article considers the root causes of oppression from social work's biopsychosocial framework, looking at biological, psychological, and social theories of oppression and explores types of oppression as delineated in the literature. Macro- and microforms of oppression are considered. From a macro- and Marxist perspective, oppressive aspects of economic globalization are discussed. From a micro-perspective, attacks on individuals on the grounds of personal or cultural difference are briefly considered. The second part of the article focuses on antioppressive theory and antioppressive practice as concepts associated most closely with the social work profession. Finally, examples of antioppressive practice are presented.

Introduction

Why is there oppression in the world? Why is there a tendency for one group to oppress another? What do scientific theories say about the basis of this social phenomenon? Although no definitive answers to these age-old questions are possible, and scholars disagree about the causes if not consequences of oppression, these are important issues to explore. In this article, I consider definitions of oppression, explanations about its prevalence, and implications for practice.

Defining Oppression

There are various definitions of oppression (Dominelli, 2002). However, because social workers are concerned (at individual and policy levels) with the victims of oppression and its prevention, define oppression as in the *Social Work Dictionary*:

> Oppression [is] the social act of placing severe restrictions on an individual group, or institution. Typically, a government or political organization in power places restrictions formally or covertly on oppressed groups so they may be exploited and less able to compete with other social groups. The oppressed individual or group is devalued, exploited, and deprived of privileges by the individual or group who has more power.
> (Barker, 2003: p. 307)

Devaluation, exploitation, and privilege are key concepts in it, which are inextricably linked to the phenomenon of oppression, which exists in all societies, and has probably done so throughout history. Attempts to eradicate oppression in society are relatively new, and the focus may be on one group to the neglect of another. For example, in the United States and Northern Ireland during the 1960s, while the civil rights movements brought a focus on discrimination and denial of human rights to minority groups on the basis of race and religion, respectively, the denial of women's rights were largely ignored. In the US, the forces that brought about freedoms for the groups in question (blacks and other minority groups) ultimately led to the feminist movement and social changes that extended across the Western world in subsequent decades.

The Nature of Oppression: What Are the Root Causes?

Iris Young's (1990) formulation of five categories of oppression is helpful to social workers seeking to understand the oppressions faced by the people with whom they work. The five forms of oppression included in her model are exploitation, marginalization, powerlessness, cultural imperialism, and violence. Drawing on Young's work, Dorothy Van Soest (2008) delineates the following key characteristics of oppression: bestowing power and advantage on those regarded as the norm and denying power and advantage to others who are seen as different; and using ideology and violence or the threat of violence to retain their privileges. Society's institutionalization of the norms of oppression and the invisibility of members in oppressed group are additional characteristics. Privilege is the flip side of oppression, oppression cannot exist in the absence of privilege, and those who have privilege may not even be aware of it (Robbins, 2011). With in-groups and out-groups in a society, what is oppression in one context may be privilege in another (Samuels and Ross-Sheriff, 2008).

These basic details about the nature of oppression challenge social workers, social scientists, and other students of human behavior to take a holistic approach to understanding privilege and oppression at the macro-, structural level, as well as in seeing how they are played out at the micro-, personal level. Samuels and Ross-Sheriff illustrate this through adopted biracial children brought up in middle-class homes. Although members of families with access to material and social resources and privilege, they may still encounter racial discrimination as blacks in external social settings.

Taking a biopsychosocial perspective in *Confronting Oppression, Restoring Justice*, van Wormer et al. (2012) consider theories of oppression provided by biologists, psychologists, and sociologists. These perspectives differ on the particular forms of oppression emphasized and the disciplines represented. Evolutionary biologists, e.g., stress genetic and instinctual components and focus on aggression in nonhuman and human behavior (Green, 2001; Wilson et al., 2012). Within each species, strong animals dominate and suppress the others; these survive and have most of the offspring. Similarly, Kurzban and Leary (2001) argue from a survival-of-the-fittest rationale that the social exclusion of people with disabilities among humans today is a throwback to an evolutionary past

when intergroup competition was fierce, exclusion and even elimination of 'unfit' members of the group would have enhanced the group's survival. A very different biologically based theory is presented by Cunningham et al. (2004). These researchers conducted laboratory experiments that suggested racism was an unconscious negative fear response to images of people who look different from one's social circle.

In the psychological realm, there are seemingly natural tendencies to blame the victim (Ryan, 1976; van Wormer et al., 2012). Such victim blaming can be considered a form of oppression. People who have lost their homes, persons who are chronically ill; those who have been raped and beaten – all are likely to be held at least partially responsible for their fate by the general public and their own families. Ryan (1976) described the phenomenon of blaming the victims of misfortune as a generic process applied to almost every social problem that exists. Victim blaming occurs when the downtrodden or 'underdogs' of society are held responsible for creating their own hardships. This can be understood as a psychological defense mechanism, as a way for people to cope with the suffering of others – by turning away from it. A related theoretical explanation is the theory of displaced aggression, which describes situations in which a person transfers his or her feelings of anger from one source to another, weaker source (see Allport, 1981/1954), e.g., an entire people may redirect their anger about a bad economy onto a marginalized group such as a racial or ethnic minority. This introduces the social realm. Before expanding the unit of analysis, I consider the psychological impact on the person who is victimized in these ways. Internalized oppression, or aggression turned inward, is often the end result. As African-American bell hooks (1993) states:

> Black people are indeed wounded by forces of domination. Irrespective of our access to material privilege we are all wounded by white supremacy, racism, sexism, and a capitalist economic system that dooms us collectively to an underclass position. Such wounds do not manifest themselves only in material ways; they affect our psychological well-being.
> (p. 11)

Social scientists focusing on conflict consider oppression as structural, with causes to be found in society as a whole (Corrigan and Leonard, 1978; Lundy, 2011; Mullaly, 2007; Vickers, 2012). This perspective, described and categorized by Malcolm Payne (2005) as radical and socialist-collectivist, is also consistent with social work values and ethics. In their change efforts, many social workers come into conflict with powerful forces in society that resist progressive change. This perception of society as torn by conflict rather than united in consensus had its origins in the teachings of Karl Marx. Today, radical sociologists and social workers highlight the powers of corporations and big banks in preserving the status quo and maintaining their own economic interests at the expense of people excluded from this power structure (Corrigan and Leonard, 1978; Lundy, 2011; Mullaly, 2007; Reisch, 2013; Vickers, 2012). Corporate control of the media portrays images of 'the just society' with great effectiveness and can spread conservative ideology consistent with the vested interests of rich and superrich people.

Other representations of this school of thought include David Gil (1998) who revealed how to confront injustice and oppression, Anne Bishop (2002) who focused on becoming allies of various oppressed groups in effecting social change, and Bob Mullaly (2007, 2010) who produced influential theories about structural social work in Canada.

Biology, psychology, and sociology provide insights into the nature of oppression with the most relevant from an anti-oppressive perspective being the radical sociological or structural approach. Oppression theory, as Susan Robbins (2011) suggests, is derived from several disciplines and shares conceptual frameworks with sociological conflict theory, critical social theory, feminist theory, and empowerment theory, radical economics, and the interdisciplinary approaches of radical social policy and critical theory. All of these have contributed significantly to understanding the nature of oppression. Marxist concepts about the interlocking systems of privilege and domination, exploitation of powerless people by rich ones who get richer, and alienation of workers, are infused in the writings of oppression theorists. Conventional social work seeks to preserve the prevailing social order (Mullaly, 2010).

Intersectionality

On the American continent, evidence of yesterday's oppression appears in the jolting words of bell hooks (2009), who recently returning to her roots in the hills and hollows of Kentucky, said:

> I was a little girl in a segregated world when I first learned that there were white people who saw black people as less than animals. Sitting on the porch, my siblings and I would watch white folks bring home their servants, the maids and cooks who toiled to make their lives comfortable. These black servants were always relegated to the back seat. Next to the white drivers in the front would be the dog and in the back seat the black worker. Just seeing this taught me much about the interconnectedness of race and class.... I was the girl they [the workers] did not see, sitting in the swing, who felt their pain and wanted to make it better.
> (pp. 148–149)

Much of the discourse about oppression addresses a dichotomous relationship between specific dominant and subordinate groups that assumes individuals must fit neatly into one group or the other. By not considering the intersections of multiple roles and identities, arguments of comparison about whose oppression is worst can prevail. The end result is that oppression can then be blamed on just one group (e.g., white working-class men) rather than considering power within its social and political contexts or analyzing how oppression is maintained. Patricia Hill Collins (1999) locates the solution to problems of 'difference' in what she, and others before her, have termed *intersectionality*. In oppression theory, this term is used to address the experiences of people who are subjected to multiple forms of oppression and domination. This multidimensional lens, facilitates understanding of how membership in a marginalized group like the female gender, combines with other key social characteristics

and statuses related to oppression, such as race or ethnicity, to produce a synergistic or multiplying effect that is stronger than the sum of the separate categories of difference. National origin, culture, religion, caste or class, age, gender identity and sexual expression, (dis)ability, and/or having the status of exconvict among others are categories that in combination can subject people to heightened levels of discrimination. Devah Pager's experimental research on responses to employment applications beautifully illustrates the point. Pager (2003) and Pager et al. (2009) tested the concept of intersectionality in their research on the low-wage labor market. Their research provides evidence that the effect of race can be magnified by additional factors, e.g., criminal history. Within social work, intersectionality is increasingly used in journal articles on minority issues and social work practice textbooks. Hulko (2009) discusses the theoretical paradigm of intersectionality relating to interlocked oppressions involving women. Writing from a feminist standpoint, Mehrotra (2010) similarly discusses the impact of intersectionality on South Asian women who identify or are identified as lesbians. In their text, *Multidimensional Contextual Practice*, Guadalupe and Lum (2005) emphasize how membership in more than one marginalized group at a time has a compounding effect on one's vulnerability to discrimination.

Antioppressive Theory

Antioppressive theory is widely used in all the English-speaking countries other than the United States. American social work professionals tend to draw upon an empowerment perspective, which also is concerned with power arrangements in society when working through the system to effect social change (Payne, 2005). Antioppression can be likened to an empowerment approach through its attention to power differentials in worker/client relationships and in helping clients gain control of their lives (see Humphries, 1996). Practitioners, as Payne (2005) suggests, can avoid oppressing (thereby empowering) clients through partnership, client choice, and seeking changes in the agency and wider systems that adversely affect clients. Both orientations are about minimizing power differences in society and maximizing the rights to which all people are entitled (Dalrymple and Burke, 2007; Dominelli, 2002). Both approaches seek the eradication of oppression at multiple levels of social life: the personal, interpersonal, structural, and cultural (Mullaly, 2010; Robbins, 2011). Empowerment social workers operate more at the individual and small-group levels from a strengths-based perspective, while antioppressive social workers are more critical of the status quo.

Dominelli's (2002) articulation of *Anti-Oppressive Social Work Theory and Practice* stresses the holistic role of social work within a rapidly globalizing economy, and has been influential in the profession, especially in the UK, Australia, and Canada. Central to her theoretical approach is an understanding of the deleterious effects that macro-level economic forces can have on peoples' daily lives. Dominelli's approach is multidimensional – socioeconomic, political, cultural, spiritual, and environmental, embedded in social justice and human rights, and argues for oppression being tackled at both the individual and/or interpersonal level and the structural one simultaneously.

Another major contribution to the antioppressive literature is the structural model introduced in Mullaly's (2010) *Challenging Oppression and Confronting Privilege*, where he describes antioppression as a starting point from which social workers can tackle oppression and discrimination. Structuralism and antioppressive perspectives are the dominant ones taught to Canadian social workers. The Canadian Association of Social Work Education, like its British counterpart, requires that Canadian schools of social work demonstrate their commitment to social justice and antioppressive social work in their curricula (Mullaly, 2010).

Lundy (2004) has built on Mullaly's ideas to develop the antioppressive conceptual framework and lay out concrete social work practice skills for individual, group, and community interventions. Using a structural approach, Lundy emphasizes human rights and views capitalism as the fundamental cause of the social and economic injustice that poor people experience. Within the context of the privatization of social services and erosion of social welfare programs, social workers are in precarious positions as they try to abide by principles of social justice.

Antioppressive Practice

Antioppressive theory, developed through the ranks of progressive practice and social action in the feminist and black activist movements, provides "an important explanatory base for action" (Payne, 2005: p. 13). Practicing from this perspective, social workers avoid blaming the victims of social problems, address the systemic nature of oppression, and rely on strategies such as consciousness raising, advocacy on behalf of marginalized groups, and community organizing for direct action. Antioppressive practitioners are critical of conventional social work practice and organizational failure in confronting unjust social systems. Van Soest (2008) challenges social work "to uncover, make visible, expose, and oppose oppressive relationships and systemic power arrangements that give privilege and advantage to one group over another" (p. 323). Ferguson and Lavalette (2006) urge that social workers move toward 'a social work of resistance' (title page).

The belief that personal troubles often stem from political sources is integral to the belief system of radical and antioppressive social work. The contribution of highlighting awareness that the 'personal is political' was originally made by feminist social workers who focused on services for women that were empowering and based on understanding that their individual problems were often systemic rather than personal (Dalrymple and Burke, 2007; Dominelli, 2002). Antioppressive social work practice has a rich history, as Dalrymple and Burke suggest. The teachings did not just emerge from the writings of academics but were a product of the real struggles of social workers in daily practice. Their struggles were not only against oppression of marginalized groups within society, but also against those elements of social control engrained within mainstream social work practice and reflected in the organizations where practitioners worked (Dominelli and McLeod,

1989). Contemporary antioppressive practice, as Dalrymple and Burke (2007) and Dominelli (2002) note, occurs within the context of globalization and the widening gap between rich and poor people linked to the neoliberal trends toward privatization of the social welfare state. Carniol (2000) articulates this position well:

> For social workers who engage in anti-oppression practice, there is a strong connection between, on the one hand, providing individual assistance to people belonging to disempowered groups, and, on the other hand, working with social movements connected to these disempowered groups. By linking these two ways of working, social service providers are challenging social services from the ground up. We are reframing 'private' problems as public issues.
> (p. 115)

To effectively confront oppression and injustice, social workers need an awareness of the forces, including psychological and structural factors, involved in their perpetuation. Such awareness in conjunction with a transnational perspective is essential for citizens to create a countervailing force to corporate power and unregulated capitalism. 'Think globally, act locally', as the environmentalists say (see Oberst, 2013; Winograd and Hais, 2009).

Global Forces of Oppression

How is it possible, Michael Dover (2008) asks, to overcome racism and other forms of oppression without fundamental changes to the system of capitalism? His discussion of worker exploitation and dehumanization is timely in considering the impact of global market capitalism at every level of the social structure because it hits hardest economically disadvantaged individuals and devastates most, low-income nations of the Global South with economies intertwined with those of the Global North. A study of the principles of global economics and free trade is thus an aspect of social work education that can no longer be neglected. For contemporary writings by social work academics on the impact of economic globalization, see Dominelli (2010), Ferguson and Lavalette (2006), Karger and Stoesz (2013), Link and Ramanathan (2011), and Reisch (2013).

Economic globalization encourages a skewed distribution of wealth, and wealthy people can access capital to invest in profitable enterprises. Access to money held by major US corporations brings with it political clout and control of elections as politicians depend on corporate finance for their campaigns. This is economic privilege to an extreme degree, and legally sanctioned recently by the US Supreme Court in *Citizens United v. Federal Election Commission* (2010). This decision allowed billionaires, corporations, and labor unions to spend unlimited sums of money on advertising that demands the defeat of individual candidates whose policies they oppose. At the same time, current US policies create tax breaks for rich people while increasing restrictions for welfare recipients and people facing unemployment. Privatized health care policies ensure continued high profits for pharmaceutical and health insurance companies; and the political clout of these corporate companies is enormous (Karger and Stoesz, 2013). In competition for government funding, between social services and corporate interests, the corporations generally win (Karger and Stoesz, 2013; Reisch, 2013). The implications for social work are significant.

At the global level, unrestricted capitalism has failed to raise the well-being of those in poverty, as Link and Ramanathan (2011) suggest, because the market has been manipulated by corporations. Competition, therefore, does not flow freely. Companies like Enron and Monsanto manipulate, pollute, and exploit for the sake of immediate profits at the expense of long-term gains. In the absence of tight regulation, corporate monopolies such as Monsanto, which produces genetically engineered seeds, can enforce policies to eliminate competition and name their price (van Wormer et al., 2012). Additionally, corporations, by moving their factories to the Global South engage in worker exploitation as far as is politically feasible. Social workers have important roles to play in terms of advocacy on behalf of individual workers (Link and Ramanathan, 2011). Collectively, social workers can organize for legislative protection of these workers' health and safety (see www.statepolicy.org to review social workers current efforts to influence US health and safety policies).

Social Work Contributions

Working globally, social work practitioners and academics produced a landmark document, the *Global Agenda*, following 3 years of collaborative initiatives undertaken by three international organizations, the International Federation of Social Workers, International Association of Schools of Social Work, and International Council on Social Welfare. The *Global Agenda* highlights the appalling conditions shaping human lives in a world that fails to hold to account market forces and enforce international standards for labor. These "have damaged the health and well-being of peoples and communities, causing poverty and growing inequality" (IASSW, 2012: p. 1; ICSW, 2012: p. 1; IFSW, 2012: p. 1). According to the *Global Agenda*, human rights are in danger worldwide (p. 1), and some aspects of economic globalization have had especially negative consequences for indigenous and First Nations peoples (p. 3). Recognizing these challenges, the *Global Agenda* pledges a commitment to focus on the following areas: "promoting social and economic equalities; promoting the dignity and worth of peoples; and working toward environmental sustainability" (p. 1).

An encouraging development took place in the United States, when the Council on Social Work Education (CSWE) (2008) issued its revised Educational Policy and Assessment Standards. The new standards placed a major stress on the teaching of concepts related not only to diversity, but also, human rights, oppression, and intersectionality. The social work curriculum was also expected to emphasize the global interconnections of oppression and cover "issues of power, privilege, and economic dominance" (p. 5).

Through the incorporation of theories of oppression, this educational revision implies a pronounced change in the focus of the standardized social work curriculum. The shift is from race to racism, sex to sexism, ethnicity to ethnocentrism, and

oppressed to oppressor. This shift is far more radical than that a quick reading of CSWE's (2008) revised curriculum standards would indicate. The new focus is structural, and on societal practices that perpetuate oppression rather than on teaching about diversity and cultural differences. This shift is consistent with social work values of social justice and nondiscrimination, and preparing students to become change agents. It is too early to tell what actual effects this curriculum change will have on social work in the USA long-term but it is an educationally significant development.

Besides preparing social work students for roles that many will play as therapists in social services agencies 'to help people help themselves', social work educators also prepare future generations to participate in policy making, as earlier generations have done. The social work curriculum emphasizes knowledge of the law in many countries, and in some such as the United States, social welfare policies are strongly emphasized. These enable students to face, "the daunting task of translating that understanding into actions designed to facilitate social change for social justice" (Van Soest, 2008: p. 347).

In the belief that policy and practice are inextricably linked, van Wormer et al. (2012) have introduced an antioppressive policy analysis model to guide social work class projects. The model leads students to discover solutions to problems that have been successfully applied elsewhere and to appreciate the many obstacles that they will encounter in introducing policy changes. A secondary goal of the exercise is to show that social workers and other helping professionals have vested interests in maintaining control and power, that for those employed in the field of social work, the 'personal is political.' Because research has shown that social work students have problems in recognizing the strategies of dominance and control are used to defeat progressive change, this exercise can be highly instructive. When tackled from an antioppressive perspective, policy analysis links the theory of oppression to policy practice.

Antioppressive Practice: Call to Action

Organizational constraints can turn antioppressive practice into a theoretical construct, or way of thinking, rather than actual practice. As social workers' belief-systems influence their professional behavior, familiarity with basic theories of oppression is a starting point for progressive policy practice. Social work practitioners and academics often occupy subordinate roles within their own agencies that require them to carry out policies with which they disagree. Agency and government policies can be oppressive when implemented in practice within a conservative setting and run according to a business model. What do social workers do then? They can quit their job or transfer to a different department; they can work within the system to change the system; they can go outside for influence, perhaps even become a whistleblower; or they can find ways to get around or even defy the most harmful aspects of the policies they are supposed to enforce. In the UK, e.g., the forced removal of unaccompanied minor asylum seekers when they reach adulthood at age 18 poses irreconcilable moral dilemmas for social workers (Wright, 2012). Social work values and ethics inform social workers in conflicts involving government legislation that their first duty is to the client. American social workers face similar situations because immigration is becoming increasingly criminalized and forcing agency employees to discriminate against undocumented clients (Furman et al., 2012).

Nzira and Williams (2009) tackle this common ethical dilemma. "Even if we feel we cannot have great influence over cultural or structural forces" Nzira and Williams (2009) advise, "we can strive to be personally antioppressive toward those we meet" (p. 38). Antioppression does not mean treating everyone the same but allowing for individual differences (Dalrymple and Burke, 2007; van Wormer et al., 2012). At the interpersonal level, social workers can engage in an antioppressive practice that is active and thoughtful. In working with people of different cultures, acquiring some knowledge of culture, language, beliefs, and likely experiences of the person can help put them at ease, show respect, and, if necessary, adopt an advocacy or support role based on partnership with service users and mutual learning.

To 'act locally', policy makers, helping professionals, and advocates who care about ending poverty and oppression need to develop skills and abilities to enable them to step forward as change agents. Activists can learn techniques of organizing through the Internet mobilization, regional teach-ins, and use of the alternative press to produce tangible results (see Haynes and Mickelson, 2009). Social workers engage in grassroots organizing on behalf of deprived groups including women living in countries where their rights are unprotected, indigenous populations, and preserving the physical environment from which all life flows. A new consciousness is promoting the protection of environmental and human resources, from Southeast Asia to the Americas. This consciousness is related to antioppressive approaches in several ways. Dominelli's (2012) *Green Social Work* reveals the connections between radical social work's values of solidarity and equality, and the questioning today's dominant model of economic development. Green social work is explicitly political, and the "values of antioppressive social work can become useful tools in mobilizing people to care for the environment" (Dominelli, 2012: p. 131). (See Gray et al. (2008), for indigenous approaches to the environment and a perspective that is consistent with antioppressive principles; and Besthorn and McMillen's (2002) article drawing parallels between the oppression of women and the oppression of nature.)

The Occupy movement, according to the *Huffington Post*, is appropriate for social workers who witness firsthand the connection between economic stability and physical and mental well-being (Jeltsen, 2011; Mason, 2012). As Mason (2012) states, social workers can learn from the Occupy Wall Street movement to shift their attention from concerns with individual minority groups to economic injustice and the need to alleviate gross economic inequalities.

Members of the Social Welfare Action Alliance (SWAA), to which many social workers belong, are actively involved in the national convergence of training, education, and action. Their goal is to train 100 000 people across the country in sustained, nonviolent direct action to expose the injustices of the moment

and demand change from those responsible (Johnson, 2012). SWAA members regularly organize rallies and teach-ins that reveal how the United States violates the United Nations Universal Declaration of Human Rights. The Kensington Welfare Rights Union (KWRU), affiliated with SWAA, develops leadership from the ranks of poor people to provide housing for homeless people. This organization uniquely links the civil rights movement with international human rights and uses civil disobedience is a tactic to house homeless people in abandoned dwellings in the Kensington neighborhood of Philadelphia. An internship program at the University of Iowa School of Social Work offers an immersion experience for social work students to learn about macro-level work from the KWRU in North Philadelphia (van Wormer et al., 2012). The documentary film, *Living Broke in Boom Times*, available at New Day Films, provides commentary by key activists in the movement.

Another movement that draws on human rights to combat oppression internationally is a form of restorative justice, e.g., Truth and Reconciliation Commissions (Androff, 2013; Stauffer, 2012) that promote healing through public testimony by the survivors of systemic oppression of minority groups through persecutions, war, and genocide (Androff, 2013). Such Commissions have been organized in South Africa, following apartheid; Timor Leste (formerly East Timor) at the end of a brutal occupation by Indonesia; and Canada to repair damage done to First Nations children, who having been forcibly removed from their homes and families, were placed in boarding schools and abused. The social work values of self-determination, truth telling, and peace, and principles of antioppression are central to this healing format. When oppressed peoples tell their stories, 'not primarily to entertain but to make a point – to demonstrate, describe, explain, or justify something to others', they move from a position of oppression, a state in which they were silenced, to a public expression of injustice (Young, 2002: p. 72).

Conclusion

This article has moved from a review of biological, psychological, and sociological theories about the nature of oppression to a theoretical/social action framework – the antioppressive model – which directly addresses the causes and consequences of oppression in society, and which has a major following among social workers and social work academics internationally. Insights from antioppressive perspectives can help social workers better understand the sources of injustice and oppression and productive ways of confronting them.

See also: Critical Social Work Practice; Global Agenda for Social Work and Social Development; Human Rights and Social Work; Intersectionality; Marxist Approaches to Social Work; Risk and Uncertainty in the Social Sciences: Implications for Social Work Theory and Practice; Social Work with Racialized Groups: Frameworks for Practice; Whiteness and White Supremacy.

Bibliography

Allport, G., 1981/1954. The Nature of Prejudice. Addison-Wesley, Reading, MA.
Androff, D., 2013. Truth and reconciliation commissions and transitional justice in restorative justice context. In: van Wormer, K., Walker, L. (Eds.), Restorative Justice Today: Practical Applications. SAGE, London, pp. 205–213.
Barker, R.,L., 2003. The Social Work Dictionary, fifth ed. NASW Press, Washington, DC.
Bishop, A., 2002. Becoming an Ally, second ed. Fernwood Press, Halifax, NS.
Carniol, B., 2000. Case Critical: Challenging Social Services in Canada, fourth ed. Between the Lines Press, Toronto.
Citizens United v. Federal Election Commission. Docket Number 08-205. 558 U.S. 310. Washington, DC: U.S. Supreme Court. www.scotusblog.com (accessed 21.08.13.).
Collins, P.H., 1999. Black Feminist Thought. Routledge, New York.
Corrigan, P., Leonard, P., 1978. Social Work under Capitalism. Macmillan, London.
Cunningham, W., Johnson, M., Raye, C., Gatenby, J.C., Gore, J.C., Banaji, M., 2004. Separable neural components in the processing of black and white faces. Psychological Science 15 (12), 806–813.
Dalrymple, J., Burke, B., 2007. Anti-Oppressive Practice, second ed. Open University Press, Buckingham.
Dominelli, L., 2002. Anti-Oppressive Social Work Theory and Practice. Palgrave, New York.
Dominelli, L., 2010. Social Work in a Globalizing World. Polity Press, Cambridge.
Dominelli, L., 2012. Green Social Work. Polity Press, Cambridge.
Dominelli, L., McLeod, E., 1989. Feminist Social Work. Macmillan, London.
Dover, M., 2008. Oppression, dehumanization and exploitation: connecting theory to experience. In: Van Soest, D., Garcia, B. (Eds.), Diversity Education for Social Justice, second ed. Council on Social Work Education, Washington, DC, pp. 366–393.
Ferguson, I., Lavalette, M., 2006. Globalization and global justice. International Social Work 49 (3), 309–318.
Furman, R., Ackerman, H., Loya, M., Jones, S., Negi, N., 2012. The criminalization of immigration. Journal of Sociology and Social Welfare 39 (1), 169–185.
Gil, D., 1998. Confronting Injustice and Oppression. Columbia University Press, New York.
Gray, M., Coates, J., Yellow Bird, M., 2008. Indigenous Social Work around the World. Ashgate, Farnham.
Green, R., 2001. Human Aggression, third ed. Open University Press, Buckingham.
Guadalupe, K.L., Lum, D., 2005. Multidimensional Contextual Practice. Cengage, Belmont, CA.
Haynes, K., Mickelson, J., 2009. Affecting Change, seventh ed. Pearson, Boston, MA.
hooks, bell, 1993. Sisters of the Yam. South End Press, Boston, MA.
hooks, bell, 2009. Belonging: A Culture of Place. Routledge, New York.
Hulko, W., 2009. The time and contingent nature of intersectionality and interlocking oppressions. Affilia 24 (1), 44–55.
Humphries, B., 1996. Critical Perspectives on Empowerment. Venture Press, New York.
International Association of Schools of Social Work (IASSW), 2012. Global Agenda for Social Work and Social Development to Action. www.iassw-aiets.org (accessed 21.08.13.).
International Council on Social Welfare (ICSW), 2012. Global Agenda for Social Work and Social Development to Action. www.icsw.org (accessed 21.08.13.).
International Federation of Social Workers (IFSW), 2012. Global Agenda for Social Work and Social Development Commitment to Action. www.ifsw.org (accessed 21.08.13.).
Jeltsen, M., 27 October 2011. Occupy Boston: Social Workers Staff Camp, Help Defuse Situations. Huffington Post. www.huffingtonpost.com (accessed 21.08.13.).
Johnson, D., 2012. You Should Know about the "99%". Institute for America's Future. http://institute.ourfuture.org (accessed 21.08.13.).
Karger, H.J., Stoesz, D., 2013. American Social Welfare Policy. Pearson, Boston, MA.
Link, R., Ramanathan, C., 2011. Human Behavior in a Just World. Rowman and Littlefield, Lantham, MD.
Lundy, C., 2004. Social Work and Social Justice. Broadview, Petersborough, OD.
Lundy, C., 2011. Social Work, Social Justice, and Human Rights. University of Toronto Press, Toronto, ON.
Mason, S.E., 2012. The Occupy movement and social justice economics. Families in Society 93 (1), 3–4.
Mehrotra, G., 2010. Toward a continuum of intersectionality theorizing for feminist social work scholarship. Affilia 25 (4), 417–430.

Mullaly, B., 2007. The New Structural Social Work, third ed. Oxford University Press, Oxford.
Mullaly, B., 2010. Challenging Oppression and Confronting Privilege, second ed. Oxford University Press, Oxford.
Nzira, V., Williams, P., 2009. Anti-Oppressive Practice in Health and Social Care. SAGE, London.
Oberst, J.R., 2013. Think globally, act locally: science careers. Science. www.sciencecareers.sciencemag.org (accessed 21.08.13.).
Pager, D., 2003. The mark of a criminal record. American Journal of Sociology 108, 937–975.
Pager, D., Western, B., Bonikowski, B., 2009. Discrimination in a low-wage labor market. American Sociological Review 74 (5), 777–799.
Reisch, M., 2013. Radical community organizing and community practice. In: Weil, M., Reisch, M., Ohmer, M. (Eds.), Handbook of Community Practice, second ed. SAGE, London, pp. 361–382.
Robbins, S.P., 2011. Oppression theory and social work treatment. In: Turner, F. (Ed.), Social Work Treatment: Interlocking Theoretical Approaches, fifth ed. Oxford University Press, Oxford, pp. 343–353.
Ryan, W., 1976. Blaming the Victim. Random House, New York.
Samuels, G., Ross-Sheriff, F., 2008. Identity, oppression, and power: feminisms and intersectionality theory. Affilia 23 (1), 5–9.
Stauffer, C., 2012. Restorative interventions for postwar nations. In: van Wormer, K., Walker, L. (Eds.), Restorative Justice Today. SAGE, London, pp. 195–204.
Van Soest, D., 2008. Oppression. In: Mizrah, T., Davis, L.E. (Eds.), Encyclopedia of Social Work, twentieth ed., vol. 3. Oxford University Press, Oxford, pp. 322–324.
Vickers, T., 2012. Refugees, Capitalism and the British State. Ashgate, Aldershot.
Wilson, M., Kahlenberg, S., Wells, M., Wrangham, R., 2012. Ecological and social factors affect the occurrence and outcomes of intergroup encounters in chimpanzees. Animal Behaviour 83 (1), 277–291.
Winograd, M., Hais, M., 2009. Millenials think globally, act locally. New Geography. www.newgeography.com (accessed 21.08.13.).
van Wormer, K., Kaplan, L., Juby, C., 2012. Confronting Oppression, Restoring Justice. Council on Social Work Education, Alexandria, VA.
Wright, F., 2012. Social work practice with unaccompanied asylum-seeking young people facing removal. British Journal of Social Work. http://dx.doi.org/10.1093/bjsw/bcs175.
Young, I.M., 1990. Justice and the Politics of Difference. Princeton University Press, Princeton, NJ.
Young, I.M., 2002. Inclusion and Democracy. Oxford University Press, Oxford.

Optimal Control Theory

Ngo Van Long, McGill University, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explains the basic structure of an optimal control problem, and shows how the optimal control theory can be usefully applied to the analysis of dynamic problems in behavioral sciences. Examples of applications in economics, finance, and management science are provided. Extensions to stochastic control theory and differential games are explained. The article concludes with notes on future directions for research.

The optimal control theory is a mathematical theory that deals with optimizing an objective function in a dynamic environment. In this theory, there are two basic types of variables: state variables and control variables. The state variables describe the current state of the system, for example, a state variable may represent the stock of equipment present in the economy, the size of a forest, the amount of carbon dioxide in the atmosphere, or a biomass of a certain fish species in a given lake. The control variables are what the decision makers can directly control, such as the rate of investment or the harvesting efforts. A system of differential equations, often called the transition equations, describes how the rate of change of each state variable at a given point of time depends on the state variables and the control variables, and possibly on time itself. Given the initial values of the state variables (often called the initial stocks), the decision maker chooses the time path of the control variables. This choice determines the rates of change of the state variables, and hence their future values. The objective of the decision maker is to maximize some performance index, such as the value of the stream of discounted profits, or some index of the level of well-being of a community, which may depend on current and future values of the control variables and the state variables. Solving an optimal control problem means determining the time paths of all the control variables so as to maximize the objective function.

It turns out that in solving an optimal control problem, it is useful to add a new set of variables, called costate variables; for each state variable, one assigns a corresponding costate variable. In economics, the costate variables are often called shadow prices, because each costate variable is an indicator of the worth of the associated state variable to the decision maker. With this additional set of variables, one can formulate an elegant set of necessary conditions, called the maximum principle, which is extremely useful in constructing the solution for the optimal control problem, and which, as a bonus, provides deep insight into the nature of the solution.

The Maximum Principle

We now define a simple optimal control problem and state the maximum principle for it. Let t denote time. Let $s(t) = (s_1(t), s_2(t), \ldots, s_n(t))$ the vector of n state variables, and $c(t) = (c_1(t), c_2(t), \ldots, c_m(t))$ be the vector of m control variables (m can be greater than, equal to, or less than n). The transition equations are given by

$$ds_i(t)/dt = f_i(s(t), c(t), t) \quad i = 1, 2, \ldots, n$$

The initial values of the state variables are $s_i(0) = s_{i0}$, $i = 1, 2, \ldots, n$. The terminal time T is fixed in this simple problem. The decision maker must find the time path of $c(t)$ to maximize the sum of (1) the value of the integral of the flow of benefit, with the benefit at time t being denoted by $u(s(t), c(t), t)$, and (2) the decision maker's valuation of the terminal stock, denoted by $\phi(s_T, T)$.

$$\int_0^T u(s(t), c(t), t)dt + \phi(s_T, T)$$

(The functions $u(.)$, $\phi(.)$ and $f_i(.)$ are exogenously specified.) At each time t, the values that $c(t)$ can take on are restricted to lie in a certain set A.

For each state variable s_i, one introduces a corresponding costate variable π_i. Let $\pi(t) = (\pi_1(t), \pi_2(t), \ldots, \pi_n(t))$. The costate variables must satisfy certain conditions, as set out below, and their optimal values must be determined as part of the solution of the problem. A key step in solving an optimal control problem is the construction of a function called the Hamiltonian, denoted as $H(s(t), c(t), \pi(t), t)$, and defined as

$$H(s(t), c(t), \pi(t), t) = u(s(t), c(t), t) + \sum_{i=1}^{n} \pi_i(t) f_i(s(t), c(t), t)$$

The maximum principle states that an optimal solution to the above problem is a triplet $(s(t), c(t), \pi(t))$ that satisfies the following conditions.

1. $c(t)$ maximizes $H(s(t), c(t), \pi(t), t)$ for given $s(t)$, $c(t)$, and t, over the set A. (This is a very attractive property, because the Hamiltonian can be interpreted as the sum of the current benefit and an indicator of future benefits arising from current actions, as reflected in the value of the rates of changes of the state variables that are influenced by the control variables.)
2. The costate variables and the state variables are related by the following differential equations

$$d\pi_i(t)/dt = -\partial H/\partial s_i(t), \quad i = 1, 2, \ldots, n$$

$$ds_i(t)/dt = \partial H/\partial \pi_i(t), \quad i = 1, 2, \ldots, n$$

Note that the first differential equation has the following appealing interpretation. The costate variable $\pi_i(t)$ being the shadow price of the state variable $s_i(t)$, the left-hand side of that equation may be thought of as the capital gain (the increase in the value of that asset). The optimal level of asset holding must be such that the sum of (1) the marginal contribution of that asset, $\partial H/\partial s_i(t)$, and (2) the capital gain term is equal to zero, (3) $2n$ boundary conditions must be met, namely, $s_i(0) = s_{i0}$, $i = 1,2,\ldots,n$ and

$$\pi_i(T) = \partial\phi/\partial s_i(T), \quad i = 1,2,\ldots,n$$

These last n conditions given are also called the transversality conditions. They say that at the terminal time T there must be an equality between $\pi_i(T)$ (the 'internal worth' of the stock i) and its 'external worth' at the margin (recall that the function $\phi(.)$ is externally specified).

The optimal time path of the control variables can be represented in the open-loop form, i.e., each $c_i(t)$ is expressed as a function of time, having as parameters the initial conditions $s_i(0)$, $i = 1,2,\ldots,n$; closed-loop form; or feedback form, i.e., an optimal rule that determines $c(t)$ using only the knowledge of the concurrent values of the state variables, and of t.

There are more complicated optimal control problems, which may involve several initial or terminal constraints on the state variables, the optimal choice of the terminal time T (no longer fixed) or of the initial time (e.g., at what time a mining firm should abandon a mine, a researcher should drop a research project, or a city should start building an airport). These problems give rise to a complex set of transversality conditions; see Léonard and Long (1992) for a useful tabulation.

Connections with the Calculus of Variations and Dynamic Programming

The optimal control theory has its origin in the calculus of variations, which concerns essentially the same types of optimization problems over time. However, with the maximum principle, developed by Pontryagin and his associates (see Pontryagin et al., 1962), the class of problems that the optimal control theory can deal with is much more general than the class solvable by the calculus of variations.

The optimal control theory is formulated in continuous time, although there is also a discrete time version of the maximum principle; see Léonard and Long (1992). In discrete time, the most commonly used optimization technique is dynamic programming, developed by Bellman (1957), using a recursive method based on the principle of optimality. Bellman's principle of optimality can also be formulated in continuous time, leading to the Hamilton–Jacobi–Bellman equation (or HJB equation for short), which is closely related to Pontryagin's maximum principle, and is often used to provide a heuristic proof of that principle (see Arrow and Kurz, 1970). Such a heuristic proof reveals that the shadow price (i.e., the costate variable) of a given state variable is in fact the derivative of the value function (obtained by solving the HJB equation) with respect to that state variable, thus justifying the name shadow price. This approach yields equations that are in agreement with (2) in the preceding section.

Applications of Optimal Control Theory to Economics, Finance, and Management Sciences

Economists initially used the optimal control theory primarily to address the problem of optimal capital accumulation and optimal management of natural resources and the environment (Arrow and Kurz, 1970; Pitchford and Turnovsky, 1977; Kemp and Long, 1980). The optimal control theory has been used to study a variety of topics, such as macroeconomic impacts of fiscal policies (e.g., the effect of a temporary increase in government spending on private consumption and investment) and optimal consumption planning of an individual. (For example, Becker and Murphy (1988) addressed the question of rational addiction: if you know that your current level of consumption of a drug will determine your degree of addiction, which in turn will increase your future demand for it, how should you determine your time path of consumption of the drug, knowing the consequence of addiction?) Recently, the optimal control theory has been used to explore conditions under which a Green Paradox might occur: given that oil firms optimize their time path of production, would a rising time path of carbon tax (which is meant to discourage the use of fossil fuels) make them extract their resource stocks faster, bringing global warming damages closer to the present? (see Sinn, 2008.)

In management sciences, optimal control has been the management of problems such as the optimal holding of inventories, the optimal time path of advertising (e.g., under what conditions would it be optimal to vary the intensity of advertising over the product cycle), the optimal maintenance and sale age of a machine subject to failure, and so on. In finance, theories of optimal portfolio adjustment over time, and theories of pricing of financial options, have been developed using optimal control methods, in a stochastic setting (see Malliaris and Brock, 1982).

It is important to note that while the variable t often stands for time, and the objective function is often an integral over a time interval, it is not necessary to restrict the theory to that interpretation. In fact, in some important applications of the optimal control theory to economics, t has been interpreted as an index of the type of an individual in a model where there is a continuum of types.

Extensions of Optimal Control Theory

There are two important extensions of the optimal control theory. The first extension consists of allowing for uncertainty in the transition equations, leading to the development of stochastic control theory. The second extension is to allow for strategic interactions among decision makers who pursue different objectives, or who have diametrically different interests. The body of literature that deals with this extension is called the theory of differential games.

Stochastic Optimal Control Theory

A typical transition equation in the stochastic optimal control theory is

$$ds(t) = f(s(t), c(t)) + g(s(t), c(t))$$

where dz represents an increment in a stochastic process z called the Wiener process. In this equation, $f(s(t),c(t))$ is the expected rate of change in the state variable s, the term dz represents a random disturbance in the external environment, and $g(s(t),c(t))$ indicates that the effect of the disturbance on the evolution of the state variable may depend on the values taken by the control variable and the state variable.

In some contexts, it is more appropriate to consider random disturbances that occur only rarely, but that may have great impacts when they occur, for example, an earthquake, a currency crisis, or a discovery of an entirely new production technique. To deal with such discrete events, the Poisson process (also called jump process) is used instead of the Wiener process, which is suitable for continuous random variations.

In finding a solution of a stochastic control problem, one makes use of a stochastic version of the HJB equation. Sometimes it is possible to provide an explicit solution in the feedback form. For example, some adjustment rules for portfolios of financial assets can be derived. Another interesting application of the stochastic optimal control theory is the problem of determining the optimal time to switch from one regime to another. This is often referred to as the optimal stopping problem. One asks, for example, when one should exercise a call option (stop waiting for a further increase in price), or when to start setting up a factory in a foreign country (stop waiting for a further favorable movement in the exchange rate).

Differential Games

A differential game is a game involving N players, each solving an optimal control problem, knowing that the other players are also solving their own optimal control problems, and that these problems are related because the controls chosen by each player have an impact on the evolution of the state variables, and possibly on the benefit functions u_i. An example of a differential game is the transboundary fishery problem faced by two adjacent coastal nations. Each nation knows that the other nation's harvesting strategy will have an impact on the growth rate of the fish stock. If the two nations do not cooperate, they are said to be playing a noncooperative differential game. In trying to predict the outcome of such games, theorists have formulated various concepts of equilibrium. A Nash equilibrium of a differential game is a pair of strategies, one for each player, such that each player's strategy is the best course of action (i.e., it solves the player's optimal control problem), given the strategy of the other players.

The nature of the solution depends on the set of strategies available to the players. If one assumes that, given the initial conditions, the players can only choose control variables as functions of time (and not as a function of the concurrent values of the state variables), then the strategies are called open-loop strategies, and the resulting Nash equilibrium is an open-loop Nash equilibrium. On the other hand, if players can observe values of the state variables and choose closed-loop strategies (i.e., strategy that specify the values to be taken by their control variables at any time as functions of the concurrent values of the state variables, and possibly of time as well) then the resulting equilibrium is a closed-loop Nash equilibrium. A refinement of the closed-loop Nash equilibrium requires that at any point of time, and given any situation the system may reach, the strategy chosen by each player must remain optimal for that player, given that other players also stick to their chosen strategies. This property is called subgame perfectness. A closed-loop Nash equilibrium that satisfies the subgame perfectness property is called a Markov-perfect Nash equilibrium.

Both deterministic and stochastic differential games have been applied to problems in economics and management science. Examples include games of transboundary fishery, advertising games, innovation games, global warming games, and so on (see Dockner et al., 2000 and Long, 2010).

Conclusions and Future Directions

The optimal control theory has become a standard tool in economics and management science. It has facilitated the solution of many dynamic economic models, and helped to provide a great deal of insight into rational behavior and interaction in a temporal context. Several unsolved problems are currently attracting research interests. Among these are (1) finding Markov-perfect equilibrium in more general games (such as games with leaders and followers) and (2) allowing coalitions of players to be formed or dissolved as the games proceed.

See also: Dynamic Decision Making; Game Theory; Self-Organizing Dynamical Systems.

Bibliography

Arrow, K.J., Kurz, M., 1970. Public Investment, the Rate of Return, and Optimal Fiscal Policy. Johns Hopkins University Press, Baltimore.

Becker, G.S., Murphy, K.M., 1988. A theory of rational addiction. Journal of Political Economy 96 (4), 675–700.

Bellman, R., 1957. Dynamic Programming. Princeton University Press, Princeton, New Jersey.

Dockner, E., Jorgensen, S., Long, N.V., Sorger, G., 2000. Differential Games in Economics and Management Science. Cambridge University Press, Cambridge, UK.

Kemp, M.C., Long, N.V., 1980. Exhaustible Resources, Optimality, and Trade. North-Holland, Amsterdam.

Léonard, D., Long, N.V., 1992. Optimal Control Theory and Static Optimization in Economics. Cambridge University Press, Cambridge, UK.

Long, N.V., 2010. A Survey of Dynamic Games in Economics. World Scientific, Singapore.

Malliaris, A.G., Brock, W.A., 1982. Stochastic Methods in Economics and Finance. North-Holland, Amsterdam.

Pitchford, J.D., Turnovsky, S.J., 1977. Applications of Control Theory to Economic Analysis. North-Holland, Amsterdam.

Pontryagin, L.S., Boltyanski, V.G., Gamkrelidze, R.V., Mishchenko, E.F., 1962. The Mathematical Theory of Optimal Processes. Wiley, New York.

Sinn, H.-W., 2008. Public policies against global warming: a supply-side approach. International Tax and Public Finance 15 (4), 360–394.

Optimal Experimental Design

Anthony C Atkinson, Department of Statistics, London School of Economics, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The methods of optimal experimental design allow cogent comparisons of designs and provide algorithms for the construction of designs in nonstandard cases. This article covers optimal design for simple regression and then develops the general theory. A sequence of designs for specific models of increasing complexity include the 2^m factorials and their fractions and generalized linear models, particularly the logistic linear model for binary data. Results on parameter uncertainty are applied to designs for conjoint choice experiments and nonlinear regression. Reference is made to several recent books and to sources of numerical algorithms.

Introduction

Theoretical results from optimal experimental design allow the comparison of proposed designs and the selection of the design that is most efficient for the specific purpose of the experiment. For example, power of important tests of hypotheses may be maximized. Methods derived from the theory provide algorithms for the construction of designs in nonstandard cases, such as those when certain treatment combinations are impossible or when responses are nonnormal.

Many experiments in behavioral science have a simple treatment structure. Study 3 of Correll et al. (2007) on the decision to shoot by police officers uses a four factor design with all factors at two levels, that is, a full 2^4 factorial design. In contrast, the choice experiments described in Section Conjoint Choice Designs of this article have a complicated structure, the selection of items to be compared depending on prior beliefs about parameter values. Optimal design algorithms are essential for the efficient design of such experiments.

Although the factorial design of Correll et al. (2007) has a simple structure, Sherman and Strang (2004) stress the close relationship between experimental designs in behavioral science and those in agriculture and medicine, where there may be important heterogeneity between experimental subjects. The data analysis of Correll et al. (2007) is for the 58 participants who completed the full set of factorial tasks, for whom some information is given on gender and ethnicity. However, there is no analysis of the heterogeneity in responses using these covariates, perhaps because of serious imbalance – only one participant was black. Atkinson and Biswas (2014: Chapter6) describe optimal design methods for balancing treatment allocation over covariates in clinical trials.

This article starts in Section Simple Regression with optimal design for simple regression. The next three sections develop the general theory of optimal design, before attention turns to designs for specific models of increasing complexity. Section 2^m Factorial Designs and Fractions covers the 2^m factorials and their fractions. Generalized linear models are in Section Generalized Linear Models with a discussion of the effects of parameter uncertainty on designs for nonlinear models in Section Optimal Design and Parameter Uncertainty. Section Locally Optimal Designs for Logistic Regression in One Variable illustrates these topics with designs for the linear logistic model. Designs for the related conjoint choice experiments are in Section Conjoint Choice Designs with designs for nonlinear regression models in Section Nonlinear Regression Models. The article concludes with a brief mention of the algorithms for design construction that are such an important consequence of the theory of optimal design.

Simple Regression

Many optimal designs minimize the variances of estimated parameters, or the variances of functions of the parameters, such as predictions. We therefore need to specify a model. The simplest interesting case is linear regression in one variable, which serves to introduce some important ideas. Section Locally Optimal Designs for Logistic Regression in One Variable describes designs for models with binary responses.

There is a single response y, the expected value of which depends linearly on the value of the variable x through the relationship $E(y) = \beta_0 + \beta_1 x$. The design problem is to choose N values of x at which measurements are to be taken. Although there are N observations, the number of distinct values of x used in the experiment, n, may be much less than N. These n values of x must lie in the experimental region χ, determined by experimental constraints. For a single variable we can scale the values of x such that $-1 \le x \le 1$ where ± 1 correspond to the minimum and maximum values of the variable.

When the variance of the observations is constant, that is $\mathrm{var}(y) = \sigma^2$, least squares is the appropriate method of parameter estimation. The least squares estimators of the parameters are $\widehat{\beta}_0$ and $\widehat{\beta}_1$ with

$$\mathrm{var}\left(\widehat{\beta}_1\right) = \sigma^2 \Big/ \sum (x_i - \bar{x})^2 \qquad [1]$$

where the sample average $\bar{x} = \sum x_i / N$. Here and in eqn [1], the summations are over all N values x_i. From eqn [1], the variance of $\widehat{\beta}_1$ is minimized when $\sum (x_i - \bar{x})^2$ is maximized. For a fixed number of trials N this occurs when all trials are at $+1$ or -1. If N is even, exactly half the trials will be at $x = +1$ and the other half at $x = -1$. This design has $\bar{x} = 0$, so that $\mathrm{var}(\widehat{\beta}_0) = \sigma^2/N$, which does not depend any further on the design. This design is therefore optimal for both parameters.

This design exhibits several features of optimal designs. The number of support points of the design n, here two, may be

appreciably less than the number of trials N. The fine structure of the design depends on the value of N, here whether it is even or odd, in which case one more trial is put at one end of the design region than at the other; it does not matter which. A third feature is that the design covers the whole of the design region, using the most extreme values of x available. In any application of the design, the order in which the two treatments $x = -1$ and $x = 1$ are applied should be randomized to avoid any confounding with omitted variables.

The optimal design found here for parameter estimation also has a minimax property for prediction: it minimizes the maximum variance of the prediction $\widehat{y}(x)$ over χ.

Optimal designs are tailored to the model that is assumed; particularly when there is one factor, the designs often have the same number of design points as there are parameters. An example is the design at two points for the two-parameter regression model. An objection to such a design is that it does not provide for checking the model, for which trials at three or more values of x are needed. This suggests that the design should also be efficient in case the relationship is a second-order model

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 \qquad [2]$$

It is always possible that an even higher-order model is required, but third- or higher-order models are rarely needed. If such models seem to be required, data transformations, such as the power transformations analyzed by Box and Cox (1964), are often called for.

Design for the three-parameter second-order model, eqn [2], introduces further aspects of the dependence of the optimal design on the precise purpose of experimentation. Unlike the two-point design for the first-order model, there is now no overall optimal design for all parameters. Testing whether the first-order model is adequate is equivalent to testing whether $\beta_2 = 0$. The variance of $\widehat{\beta}_2$ is minimized by a design that puts half the trials at $x = 0$ and divides the remaining half between $x = -1$ and $x = 1$. But the design that provides the best estimates of all three parameters, in the sense of D-optimality described in the next section, puts one-third of the trials at each of these three points. In a satisfactory manner, the optimal design depends on the precise purpose of the experiment.

Criteria of Optimality

This section mainly describes the criterion of D-optimality, which provides designs minimizing the generalized variance of the estimated parameters.

We consider a general experiment in which there are m factors and write the linear model as

$$E(y) = F\beta \qquad [3]$$

Here y is the $N \times 1$ vector of responses, β is a vector of p unknown parameters, and F is the $N \times p$ *extended* design matrix. The ith row of F is $f^T(x_i)$, a known function of the m explanatory variables. For the quadratic model eqn [2], for which $m = 1$ and $p = 3$, $f^T(x_i) = \begin{pmatrix} 1 & x_i & x_i^2 \end{pmatrix}$. The design is determined by the experimental values of x. The *extended* design matrix reflects not only the design but in addition the model to be fitted.

For the moment we continue to assume that the observational errors are independent with constant variance, σ^2. The least squares estimator of the parameters can be written as

$$\widehat{\beta} = \left(F^T F\right)^{-1} F^T y \qquad [4]$$

where the $p \times p$ matrix $F^T F$ is the information matrix for β. The larger $F^T F$, the greater is the information in the experiment.

With σ^2 constant, the covariance matrix of the least squares estimator is

$$\operatorname{var} \widehat{\beta} = \sigma^2 \left(F^T F\right)^{-1} \qquad [5]$$

When interest is in the comparison of experimental designs, the value of σ^2 is not relevant, since it is the same for all proposed designs for a particular experiment.

The volume of the confidence ellipsoid for all p elements of β is inversely proportional to the square root of the determinant $\left|F^T F\right|$. Designs which maximize $\left|F^T F\right|$ minimize the generalized variance of $\widehat{\beta}$ are called D-optimal (for Determinant).

Interest in the fitted model may be not only in the parameters, but also in the quality of the predictions. The predicted value at x of the fitted response for the p-parameter model eqn [3] is

$$\widehat{y}(x) = \widehat{\beta}^T f(x) \qquad [6]$$

with variance

$$\operatorname{var} \widehat{y}(x) = \sigma^2 f^T(x) \left(F^T F\right)^{-1} f(x) \qquad [7]$$

Designs which minimize the maximum over χ of $\operatorname{var} \widehat{y}(x)$ are called G-optimal. As the theorem of Section The General Equivalence Theorem and Design Efficiency indicates, as N increases, D-optimality and G-optimality give increasingly similar designs.

An alternative to this minimax approach to $\operatorname{var} \widehat{y}(x)$ is to find designs that minimize the average value of the variance over χ. Such designs are variously called I-optimal or V-optimal from Integrated Variance. Another criterion of importance in applications, particularly for nonlinear models, is that of c-optimality in which the variance of the linear combination $c^T \widehat{\beta}$ is minimized, where c is a $p \times 1$ vector of constants. When c is a vector of zeroes and ones, the criterion selects a subset of parameters and reduces to **D_S-optimality**. We do not exhibit such designs in this article, but focus on D- and G-optimality. Details of these and other design criteria are in Chapter 10 of Atkinson et al. (2007). Goos and Jones (2011: Section 4.3.5) stress the advantages of I-optimality.

Exact and Continuous Designs

Instead of finding *exact* optimal designs for specific values of N, there are mathematical and computational advantages in finding the limiting *continuous* design independent of N. Such designs can both easily be checked for optimality and readily be approximated to provide good exact designs for many N.

The theory was introduced in Kiefer (1959) and developed in a series of further papers. Kiefer's work on optimal design is collected in Brown et al. (1985). A basic idea is to write designs as probability measures on the n support points of the design

$$\xi = \begin{Bmatrix} x_1 & x_2 \ldots x_n \\ w_1 & w_2 \ldots w_n \end{Bmatrix} \quad [8]$$

where the weights $w_i > 0$ and $\sum w_i = 1$. The measure ξ therefore specifies the experimental conditions and the proportion of experimental effort at each condition.

The information matrix can be written in terms of this measure as

$$M(\xi) = \sum w_i f(x_i) f^T(x_i) \quad [9]$$

The variance of the predicted response can be written in the standardized form

$$\mathrm{d}(x, \xi) = f^T(x) M^{-1}(\xi) f(x) \quad [10]$$

At the end of Section Simple Regression it was stated that the D-optimal design for the quadratic model in one variable with $\chi = [-1, 1]$ put weights one-third at each of the three levels of x: -1,0, and 1. For any N that is a multiple of three, the exact optimal design puts $N/3$ trials at each of these sets of experimental conditions. For other values of N a good design is to distribute the trials as evenly as possible over these three points.

The General Equivalence Theorem and Design Efficiency

D-optimal continuous designs maximize $|M(\xi)|$. Kiefer and Wolfowitz (1959) provide an equivalence theorem, relating D- and G-optimal designs, that provides a method of checking the D-optimality of a continuous design.

If ξ^* is a D-optimal design measure, then ξ^* also minimizes the maximum over χ of the standardized variance of prediction $d(x, \xi)$ given by eqn [10]; that is ξ^* is also G-optimal. If we write this maximum as

$$\overline{d}(\xi) = \max_{x \in \chi} d(x, \xi) \quad [11]$$

then $\overline{d}(\xi^*) = p$, the number of parameters of the linear model. For any nonoptimal design ξ the maximum over χ of $d(x, \xi)$ is greater than p, which provides a method for checking for D-optimality.

The theorem holds for continuous optimal designs and, of course, for exact designs with replications that are an integer multiple of the weights of the continuous designs. But, for example, the exact D-optimal design for the quadratic model when N is not a multiple of three will not satisfy the theorem.

To compare a continuous design ξ with the optimal design ξ^* we define the D-efficiency as

$$D_{\text{eff}} = \left\{ \frac{|M(\xi)|}{|M(\xi^*)|} \right\}^{1/p} \quad [12]$$

The comparison of information matrices for designs that are measures removes the effect of the number of observations. Taking the pth root of the ratio of the determinants in eqn [12] results in an efficiency measure which has the dimensions of variance, irrespective of the dimension of the model. Since the variance of estimated regression coefficients is inversely proportional to the number of observations, two replicates of a design measure for which $D_{\text{eff}} = 50\%$ would be as efficient as one replicate of the optimal measure.

Quadratic Model in One Variable

The claim that the final design of Section Simple Regression with weights 1/3 is D-optimal for the quadratic model when $\chi = [-1, 1]$ can be checked using the General Equivalence Theorem through calculation of the variance of the predicted response.

For this design $\mathrm{d}(x, \xi^*)$ is the quartic given by

$$\mathrm{d}(x, \xi^*) = 3 - 9x^2/2 + 9x^4/2 \quad [13]$$

which has a maximum over χ of 3 at $x = -1$,0, or 1, the three design points. Further, this maximum value is equal to the number of parameters p. The plot in **Figure 1** exhibits the optimality of the design since the values of three at the design points are indeed the optima over χ.

Equivalence theorems for other criteria are given in Chapter 10 of Atkinson et al. (2007).

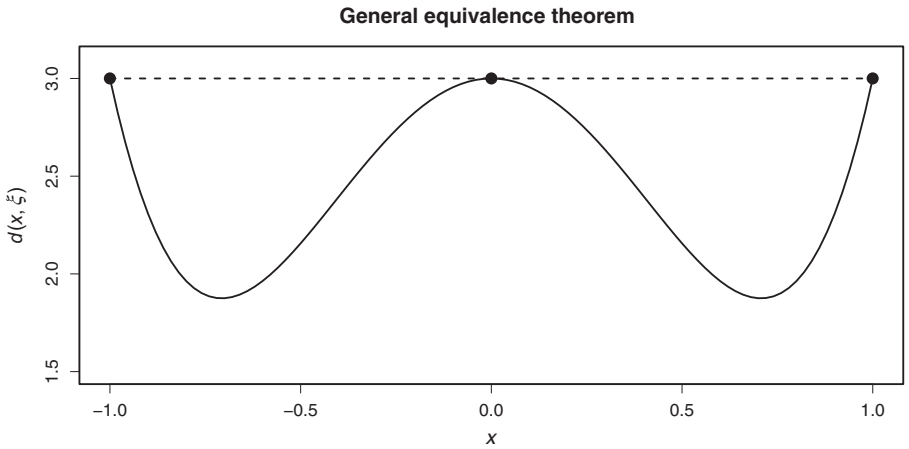

Figure 1 Quadratic model in one variable; $d(x, \xi^*)$ over χ for the design putting equal weights at $x = -1$, 0, and 1. Since the maximum value is three, which is the number of parameters in the model, the design is D-optimal.

2^m Factorial Designs and Fractions

The first-order model in m factors with main effects and all possible interactions may be written as

$$E(\gamma) = \beta_0 + \sum_{i=1}^{m} \beta_i x_i + \sum_{i=1}^{m-1} \sum_{j=1+1}^{m} \beta_{ij} x_i x_j + \cdots + \beta_{ijk\ldots} x_i x_j x_k \cdots \quad [14]$$

When χ is cuboidal, the D-optimal continuous design is the equireplicated 2^m factorial, which has $n = 2^m$ points of support at all combinations of values of ± 1 for each factor. If the full model eqn [14] is fitted $p = N$ and $M(\xi^*)$ is the identity matrix, then $\mathrm{d}(x, \xi^*) = p$ at the design points and the equivalence theorem for continuous D-optimality is satisfied. Because of the orthogonality of the design, reflected in the diagonal nature of $M(\xi^*)$, the same design is optimal if terms are omitted from eqn [14] or if the design is projected by the omission of one or more factors. Provided all the terms in the model can be estimated, the regular fractions of the design, that is the series of 2^{m-f} fractional factorials, are also the continuous D-optimal designs.

The 2^{m-f} fractional factorials are often used when it is known that some interactions are not important and the full factorial design will lead to unnecessarily precise estimation of the parameters, with a consequent waste of resources. However, these orthogonal fractions are restricted to values of N that are powers of 2. Goos and Jones (2011: Section 8.2.2) demonstrate the use of algorithmic methods for the construction of optimal designs that find fractional factorials for a specified value of N. Although the designs are not quite orthogonal, highly efficient parameter estimates are obtained.

Generalized Linear Models

The Family of Models

The family of generalized linear models (McCullagh and Nelder, 1989) extends normal theory regression to several useful distributions, including the gamma, Poisson, and binomial. The models have three components:

- A distribution for the univariate response γ with mean μ.
- A linear predictor, $\eta = \theta^T f(x)$. As for regression, $f(x)$ is a vector of p known functions of the explanatory variables.
- A link function, $g(\mu) = \eta$, relating x to the mean μ.

The distribution of γ determines the relationship between the mean and the variance of the observations. The information matrix for a single observation at x is

$$M(x;\theta) = u(x) f(x) f^T(x) \quad [15]$$

with the weights $u(x)$ for individual observations depending on both the distribution of γ and the link function. Since eqn [15] is the information matrix for weighted least squares, there are important relationships between optimal designs for regression models and those for other generalized linear models.

Binomial Data

In models for binomial data with R successes in n observations, the response γ is defined to be R/n. The link function should be such that, however the values of x and θ vary, the mean μ satisfies the physically meaningful constraint $0 \le \mu \le 1$. The most commonly used link function is the logistic link in which

$$\eta = \log\{\mu/(1-\mu)\} \quad [16]$$

In this model the log odds is therefore equal to the linear predictor. The weight $u(x) = \mu(1-\mu)$. The weights for other links are more complicated.

Poisson Data

For Poisson data we require that $\mu > 0$. The log link, $\log \mu = \eta$, is standard for the analysis of Poisson data in areas such as medicine, social science, and economics; see, for example, Agresti (2002: Chapter 9). This link leads to models with $\mu = \exp \eta$, which satisfy the constraint on values of μ, and weights $u(x) = \mu$.

Gamma Data

The gamma family is one in which the correct link is often in doubt. The physical requirement is again that $\mu > 0$. A useful, flexible family of links that obeys this constraint is the Box and Cox family. (See Box and Cox, 1964 for the use of this function in the transformation of data.) For the special case of the log link, the weights $u(x)$ equal one. Therefore optimal designs for gamma models with this link are identical to optimal designs for regression models with the same linear predictors.

Optimal Design and Parameter Uncertainty

The optimal designs for regression in the earlier sections of this article maximized a function of the information matrix, $M(\xi)$, which did not depend on the values of the parameters of the model. However, for generalized linear models, the information matrix depends on θ through the weight function $u(x)$. We accordingly write the information matrix as $M(\xi;\theta)$ and consider maximization of some function, $\Psi\{M(\xi;\theta)\}$. The numerical values of θ required for this optimization can be specified in several ways.

Locally Optimal Designs

The optimal design is found for θ_0, a point prior estimate of θ, that is we maximize $\Psi\{M(\xi;\theta_0)\}$. In the conjoint choice designs of Section Conjoint Choice Designs these values serve to indicate the signs of effects, the designs being relatively insensitive to the precise values of θ_0.

Bayesian Designs

Let the prior distribution of θ be $p(\theta)$. Bayesian designs are found to maximize

$$\Psi\{M(\xi)\} = E_\theta \Psi\{M(\xi;\theta)\} = \int \Psi\{M(\xi;\theta)\} p(\theta) \mathrm{d}\theta \quad [17]$$

The ease of calculation depends on the form of the prior $p(\theta)$ and of $\Psi(.)$ as well as, often crucially, on the region of integration. The easiest case is when θ has a multivariate

normal distribution over $\Re^p$, although numerical approximation is still needed. For less amenable cases, a standard solution is to sample from the prior distribution and to use an average objective function. An example for a nonlinear model is in Section 18.5 of Atkinson et al. (2007).

Designs maximizing expectations such as eqn [17] ignore the additional effect of the prior information about θ on the information matrix and make no assumption of a Bayesian analysis. The designs are accordingly sometimes called pseudo-Bayesian. A discussion of Bayesian experimental design is given by Chaloner and Verdinelli (1995).

Sequential Designs

Where it is possible, sequential experimentation can provide an efficient strategy in the absence of knowledge of plausible parameter values. A locally optimal, or other initial, design is followed by alternating phases of parameter reestimation and locally optimal design based on the revised parameter estimates. The procedure stops when sufficient accuracy is obtained or the experimental resources are exhausted. An early example, for nonlinear regression, is Box and Hunter (1965), extended by Atkinson et al. (2007: Section 17.7). Dror and Steinberg (2008) developed a Bayesian sequential design methodology for generalized linear models.

Minimax and Maximin Designs

This approach replaces dependence of designs on the unknown value of θ by finding the best design for the worst case when the parameter θ belongs to a set Θ. A design, ξ^* is found for which

$$\Psi\{M(\xi^*)\} = \max_{\xi} \min_{\theta\in\Theta} \Psi\{M(\xi;\theta)\} \qquad [18]$$

A potential objection to these designs is that the minimax or maximin design is often close to a combination of locally optimal designs for values of θ at the edges of the parameter space. If a prior distribution is available, such points may have a very low probability; their importance in the design criterion may therefore be overemphasized by the minimax criterion. Providing adequate performance in these unlikely worst case scenarios may greatly affect overall design performance. These designs can also be hard to find. Numerical procedures are described in detail by Pronzato and Pázman (2013: Section 9.3). Minimax designs for generalized linear models have been found by Sitter (1992) and by King and Wong (2000).

Small Effects

If the effects of the factors are slight, the mean, μ_i of each observation will be similar, and so will the corresponding model weights, $u(x_i)$. The information matrix will then, apart from a scaling factor, be close to the unweighted information matrix, $F^{\mathrm{T}}F$ and the optimal designs for the weighted and unweighted matrices will be close. Optimal designs for generalized linear models with small effects will therefore be close to optimal designs for regression models with the same linear predictor (Cox, 1988).

Locally Optimal Designs for Logistic Regression in One Variable

With a single explanatory variable and the logistic link

$$\log\{\mu/(1-\mu)\} = \eta = \theta_0 + \theta_1 x \qquad [19]$$

it is clear that observations at values of x for which the responses are virtually all 0 or 1 will be uninformative about the values of the parameters. This intuitive result indeed follows from the weight $u(x) = \mu(1-\mu)$.

We start by finding the locally D-optimal design for the *canonical* case of $\theta = (0,1)^T$. With a response symmetrical about zero, the design can also be assumed symmetrical about zero. It is thus simple to find the D-optimal design as

$$\xi^* = \left\{ \begin{matrix} -1.5434 & 1.5434 \\ 0.5 & 0.5 \end{matrix} \right\} \qquad [20]$$

which puts equal weights at two points symmetrical about $x = 0$, provided χ is sufficiently large. At these support points, $\mu = 0.176$ and $1 - 0.176 = 0.824$.

Although designs for other values of the parameters can likewise be found numerically, design problems for a single x can often be solved in such a canonical form, yielding a structure for the designs independent of the particular parameter values.

For the upper design point in eqn [20] the linear predictor, $\eta = 0 + 1 \times x$ has the value 1.5434, which is the value needed for the optimal design whatever the parameterization. If we solve eqn [19] for the η giving this value, the upper support point of the design is given by $x^* = (1.5434 - \theta_0)/\theta_1$.

For linear regression the D-optimal design puts half of the design weight at the two extreme values of χ, whereas, for logistic regression, the design does not span χ, provided the region is sufficiently large. Note that as $\theta_1 \to 0$, the value of x^* increases without limit. This is an example of the result of Cox (1988) mentioned above. In practice, the design region will not be unlimited and, depending on θ, the optimal design may put weight on one boundary point and an internal point, or on the two boundary points of χ.

Ford et al. (1992) provide similar results on the structure of designs for other single-variable generalized linear models. Optimal designs for models with more than one explanatory variable usually have to be found numerically. The optimality of continuous designs can again be checked using an equivalence theorem. For D-optimality, eqn [10] is replaced by the weighted form $\mathrm{d}(x,\xi) = u(x)f^T(x)M^{-1}(\xi)f(x)$.

Chapter 22 of Atkinson et al. (2007) provides examples of the numerical calculation of locally optimal designs for logistic and gamma models with two explanatory variables. Atkinson and Woods (to appear) collect results on optimal designs for models with multifactor first-order linear predictors. They also calculate Bayesian and other designs when there is uncertainty about the values of the parameters, and discuss designs for dependent data arising from random effects in the model.

Conjoint Choice Designs

Conjoint choice experiments are frequently performed in marketing and other fields of behavioral research in order to

establish individual preferences between items. The factors are typically discrete and at a few levels. As an example, factors in the choice between laptop computers might include discretized values of weight, RAM, screen size, computational speed, and price. The design region, χ consists of all feasible combinations of the factors (and so may not form a full factorial array, since such extremes as cheap fast computers with large screens are not a practical reality). A single observation comes from comparing the desirability of J items formed by choosing J rows from χ to give a choice set, s. The utility attached to x_j, the jth row of χ, is customarily modeled as $\propto x_j^{\mathrm{T}}\beta$ plus an error term. Since only comparisons between items are meaningful, the first-order model excludes a constant.

The response to the experiment might be a scoring or ranking of the J items. However, the literature summarized by Kessels et al. (2006) indicates that results closest to purchasing behavior are obtained by asking the subject to choose one of the J items. Then the multinomial logit probability model that profile j in choice set s is chosen as

$$p_{js} = \exp\left(x_{js}^{\mathrm{T}}\beta\right) \Big/ \left\{\sum_{\iota=1}^{J} \exp\left(x_{\iota s}^{\mathrm{T}}\beta\right)\right\} \qquad [21]$$

The design matrix for choice set s is the $J \times p$ matrix, X_s, with rows $x_{\iota s}^{\mathrm{T}}$, $(\iota = 1, \ldots, J)$. In the simplest case, each subject performs evaluations of N choice sets. The information matrix for the exact design for a single subject is then

$$NM(\xi_N) = \sum_{s=1}^{N} X_s^{\mathrm{T}}\left(P_s - p_s p_s^{\mathrm{T}}\right)X_s \qquad [22]$$

where $p_s = (p_{1s}, \ldots, p_{Js})^T$, $P_s = \mathrm{diag}(p_{1s}, \ldots, p_{Js})$, and p_{js} is given by eqn [21]. The number N of choice sets seen by each subject is determined by how quickly subjects are expected to tire. The number of subjects is determined by N, the available resources, and the desired accuracy. The remaining optimal design problem is then to determine the N choice sets of size J that maximize the exact information matrix eqn [22]. Often J is taken as two.

Clearly, maximization of a function of eqn [22] is related to the optimal design problems for generalized linear models of Section Generalized Linear Models. In particular, the design depends on the unknown parameters, β through the probabilities p_{js} in eqn [21]. The problems caused by parameter uncertainty described in Section Optimal Design and Parameter Uncertainty, and the methods for their solution, therefore also apply to choice experiments.

The locally optimal design with all elements of $\beta = 0$ is called utility neutral, since there is no prior expression of any preference between elements of χ. The optimal designs are then related to the factorials and their fractions of Section 2^m Factorial Designs and Fractions. Kuhfeld and Tobias (2005) describe algorithms for the construction of such designs. However, many of the sets selected by utility neutral designs may result in noninformative choices. For example, it is clear in comparing computers that higher speed and lower weight are both desirable. Accordingly, Kessels et al. (2006) provide algorithms for D-optimal and other designs for these models.

Two further points are important. The model of utility used in eqn [21] assumes a fixed effect, that is, that all subjects have the same value for the parameter β. However, parameters are likely to vary between subjects, leading to random effects models (Yu et al., 2009). Further, the experiments are sequential, so that it should be possible to learn about a subject's parameter value in real time. However, at present, the algorithms are not fast enough to allow updating without severely delaying the subject's flow of answers.

Nonlinear Regression Models

The dependence of designs for generalized linear models on the parameters θ entered only through the weight function, $u(x)$. In nonlinear models, such as those derived from chemical kinetics, the response function may depend nonlinearly on all the parameters in a more complicated way.

For example, in the compartmental model presented by Box and Lucas (1959) the mean of the regression model is

$$E(Y) = \eta(t, \theta) = \theta_3\{\exp(-\theta_2 t) - \exp(-\theta_1 t)\} \quad (t \geq 0) \qquad [23]$$

with all three parameters positive and $\theta_1 > \theta_2$. The exponential structure of such models reflects an underlying scheme of first-order chemical reactions.

Locally optimal designs are found by Taylor series expansion. Let θ be of dimension $p \times 1$, with θ_0 the vector of p prior values of θ. Then the vector of partial derivatives of the model, often called parameter sensitivities, is

$$f^T(t, \theta_0) = \left\{\frac{\partial\eta(t,\theta)}{\partial\theta_1}\frac{\partial\eta(t,\theta)}{\partial\theta_2}\cdots\frac{\partial\eta(t,\theta)}{\partial\theta_p}\right\}\Bigg|_{\theta=\theta_0}$$

which forms the terms of the p variable linearized multiple regression model. The vector of sensitivities for observation i is written, $f^T(t_i, \theta)$ and forms row i of the extended design matrix F for this linearized model; the information matrix is $F^{\mathrm{T}}F$. Note that θ_3 occurs linearly in eqn [23], so that optimal designs will not depend in the value of this parameter.

Chapter 17 of Atkinson et al. (2007) describes a variety of approaches to design for nonlinear models. Because of the dependence of the designs on unknown parameter values, the results of Section Optimal Design and Parameter Uncertainty are again pertinent.

Discussion and Literature

The principles and theory of optimal experimental design can yield insight into the form of efficient experiments. However, application of the methods to the more complicated, parameter-dependent, designs in Sections Generalized Linear Models; Optimal Design and Parameter Uncertainty; Locally Optimal Designs for Logistic Regression in One Variable; Conjoint Choice Designs and Nonlinear Regression Models may require powerful numerical algorithms.

Atkinson et al. (2007) relies on SAS software for the numerical calculation of designs. A general discussion of some algorithms for continuous designs harnessing the power of standard optimization methods is in Sections 9.4 and 9.5. Chapter 12 covers the rather different algorithms needed for the construction of exact D-optimal designs. The book of Goos

and Jones (2011) proceeds by numerical examples using JMP software to find exact designs and illustrate the methods of optimal design. The basic method is the coordinate-exchange algorithm introduced by Meyer and Nachtsheim (1995).

The first exposition of optimal design in its modern form is Kiefer (1959), although the subject goes back to Smith (1918). In addition to the books already cited in this article, recent books on optimal design include Fedorov and Leonov (2014), who provide a solid introduction to the theory of optimal design illustrated with pharmaceutical applications, particularly dose finding. Pronzato and Pázman (2013) present the theory of optimal design for nonlinear models. Berger and Wong (2009) are more concerned with applications.

See also: Binary Response Models and Logistic Regression; Crossover Designs; Experimental Design: Large-Scale Social Experimentation; Experimental Design: Overview; Experimental Design: Randomization and Social Experiments; Generalized Linear Mixed Models; Linear Mixed Models; Model Selection and Model Averaging.

Bibliography

Agresti, A., 2002. Categorical Data Analysis, second ed. Wiley, New York.

Atkinson, A.C., Biswas, A., 2014. Randomised Response-adaptive Designs in Clinical Trials. Chapman and Hall/CRC Press, Boca Raton, FL.

Atkinson, A.C., Donev, A.N., Tobias, R.D., 2007. Optimum Experimental Designs, with SAS. Oxford University Press, Oxford.

Atkinson, A.C., Woods, D. Designs for generalized linear models. In: Bingham,D., Dean, A., Morris, M., Stufken, J. (Eds.), Handbook of Experimental Design. Chapman and Hall/CRC Press, Boca Raton, FL, to appear.

Berger, M., Wong, W.K., 2009. An Introduction to Optimal Designs for Social and Biomedical Research. Wiley, New York.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations (with discussion). Journal of the Royal Statistical Society, Series B 26, 211–246.

Box, G.E.P., Hunter, W.G., 1965. Sequential design of experiments for nonlinear models. In: Proceedings IBM Scientific Computing Symposium: Statistics. IBM, New York, pp. 113–137.

Box, G.E.P., Lucas, H.L., 1959. Design of experiments in nonlinear situations. Biometrika 46, 77–90.

Brown, L.D., Olkin, I., Sacks, J., Wynn, H.P. (Eds.), 1985. Jack Carl Kiefer Collected Papers III. Wiley, New York.

Chaloner, K., Verdinelli, I., 1995. Bayesian experimental design: a review. Statistical Science 10, 273–304.

Correll, J., Park, B., Judd, C.M., Writtenbrink, B., Sadler, M.S., Keesee, T., 2007. Across the thin blue line: police officers and racial bias in the decision to shoot. Journal of Personality and Social Psychology 92, 1006–1023.

Cox, D.R., 1988. A note on design when response has an exponential family distribution. Biometrika 75, 161–164.

Dror, H.A., Steinberg, D.M., 2008. Sequential experimental designs for generalized linear models. Journal of the American Statistical Association 103, 288–298.

Fedorov, V.V., Leonov, S.L., 2014. Optimal Design for Nonlinear Response Models. Chapman and Hall/CRC Press, Boca Raton, FL.

Ford, I., Torsney, B., Wu, C.F.J., 1992. The use of a canonical form in the construction of locally optimal designs for non-linear problems. Journal of the Royal Statistical Society, Series B 54, 569–583.

Goos, P., Jones, B., 2011. Optimal Design of Experiments: A Case Study Approach. Wiley, New York.

Kessels, R., Goos, P., Vandebroek, M., 2006. A comparison of criteria to design efficient choice experiments. Journal of Marketing Research 43, 409–419.

Kiefer, J., 1959. Optimum experimental designs (with discussion). Journal of the Royal Statistical Society. Series B 21, 272–319.

Kiefer, J., Wolfowitz, J., 1959. Optimum designs in regression problems. Annals of Mathematical Statistics 30, 271–294.

King, J., Wong, W.K., 2000. Minimax D-optimal designs for the logistic model. Biometrics 56, 1263–1267.

Kuhfeld, W., Tobias, R., 2005. Large factorial designs for product engineering and marketing research applications. Technometrics 47, 132–141.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman and Hall, London.

Meyer, R.K., Nachtsheim, C.J., 1995. The coordinate exchange algorithm for constructing exact optimal experimental designs. Technometrics 37, 60–69.

Pronzato, L., Pázman, A., 2013. Design of Experiments in Nonlinear Models. Springer, New York.

Sherman, L., Strang, H., 2004. Verdicts or inventions? Interpreting results from randomized controlled experiments in criminology. American Behavioral Scientist 47, 575–607.

Sitter, R.R., 1992. Robust designs for binary data. Biometrics 48, 1145–1155.

Smith, K., 1918. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. Biometrika 12, 1–85.

Yu, J., Goos, P., Vandebroek, M., 2009. Efficient conjoint choice designs in the presence of respondent heterogeneity. Marketing Science 28, 122–135.

Optimism and Pessimism

Charles S Carver, University of Miami, Coral Gables, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Optimism is the extent to which people hold favorable expectancies for their future. Higher optimism has been linked to higher engagement coping and lower avoidance coping. Optimism is also associated with taking proactive steps to protect health, whereas pessimism is associated with health-damaging behaviors. The energetic, task-focused approach that optimists take to goals also predicts benefits in the socioeconomic world. Optimism predicts more persistence in educational effort and higher later income. Optimists also appear to fare better than pessimists in relationships. Although optimism sometimes fails to convey an advantage, and occasionally may even convey a disadvantage, those instances are rare.

The concepts of optimism and pessimism are rooted in centuries of folk wisdom about core differences among people. In this case, the difference among people is in their expectations for the future. This difference has reverberations throughout people's lives and relationships with other people. Optimists, expecting to see the best, tend to find the best in whatever circumstances they confront. Pessimists, expecting trouble at every turn, tend to experience all of life as less than what it ought to be.

Contemporary research on optimism and pessimism also has its roots in people's expectations for the future (Carver and Scheier, 1998, 2014; Carver et al., 2010; Segerstrom, 2006). In taking that starting point, this body of research connects to more than a century of psychological theory. That is, the expectancy construct links research on optimism to a collection of theories of motivation and action that collectively are called expectancy-value theories (Eccles and Wigfield, 2002). This link, in turn, provides a clear rationale for examining the way in which the traits of optimism and pessimism are expressed in motivation, behavior, and feelings.

Expectancy-value models begin with the idea that behavior is aimed at attaining desired goals. These goals reflect the 'values' in the label expectancy value. Without a valued goal, there is no reason to act. The other element is expectancy: confidence or doubt about being able to attain the goal. If a person lacks confidence, again there will be no action. Only with sufficient confidence do people exert efforts and remain engaged in those efforts. These principles pertain to specific values and focused confidence, and they also apply to the more general traits of optimism and pessimism. In optimism and pessimism the 'confidence' involved is simply broader in scope.

These motivational principles provide the basis for many predictions about optimists and pessimists: When confronting a challenge, optimists should take a posture of confidence and persistence, even if progress is difficult or slow. Pessimists should be more doubtful and hesitant, quicker to give up their efforts. This divergence should be amplified under conditions of adversity. Optimists are likely to believe the adversity can be handled successfully; pessimists are likely to anticipate disaster. These differences in orientation can lead to differences in active efforts to take precautions, differences in persistence, differences in many decisions and actions that pertain to health and other risks, and differences in how people approach relationships.

People's overt behavioral responses to adversity are important, but overt behavior is not the only response that occurs when people confront adversity. People also experience emotions in such situations. Encountering difficulties results in a mixture of feelings and optimism influences the balance among the feelings. Again, predictions are easily derived from theory. Optimists expect to have good outcomes, even when things are difficult. This confidence should yield a more positive mix of feelings. Pessimists expect bad outcomes. This should be related to more intense negative feelings – anxiety, sadness, or despair.

Relations between optimism and distress have been examined in a wide range of contexts (Carver et al., 2010). The studies that have been done vary in complexity and what they are therefore able to show. Some of the work is cross-sectional, showing only that lower optimism relates to reports of more distress in some difficult situation. Those studies cannot show whether pessimistic people had more distress even prior to the adversity. Other studies assess people at multiple time points. These studies give a better picture of how distress shifts over time and circumstances and allow researchers to control for initial levels of distress.

A good deal of the work on optimism and subjective well-being has been done in the context of medical settings. Several projects have studied people having coronary artery bypass surgery. Several projects have also studied optimism in the context of treatment for breast cancer. Other studies have examined optimism in the context of *in vitro* fertilization, a procedure that helps people overcome fertility problems, and in the context of treatment for ischemic heart disease. Another project studied a group of cancer patients and their family caregivers. Family caregiving for a serious illness is highly stressful. In this study caregivers' optimism predicted less depression over time and less adverse impact of caregiving on their physical health.

The experiences of intense distress, hopelessness, and disengagement from effort have physical concomitants; so do positive emotions and engagement in positive efforts. A number of psychologists and others believe that these physiological responses can play a role in health outcomes, thus resulting in health outcomes that differ between

optimists and pessimists (Boehm and Kubzansky, 2012; Rasmussen et al., 2009). Such processes may influence a person's probability of getting a disease in the first place, they may influence the progression of diseases, and they may even influence mortality.

Studying Optimism and Pessimism

The easiest way to think about and measure optimism and pessimism is in terms of expectancies, asking people directly whether they expect outcomes in their lives to be good or bad. This approach is reflected in the life orientation test (LOT) and its successor the life orientation test-revised (LOT-R; Scheier et al., 1994). The LOT-R consists of a set of statements (e.g., 'I am always optimistic about my future,' 'I rarely count on good things happening to me' (reversed)) to which people indicate their agreement or disagreement on a multipoint scale. This measure yields a continuous distribution of scores. It is common to refer to optimists and pessimists as though they were distinct categories of people, but this is simply a verbal convenience. Almost never is a line drawn and people placed into two groups. People range from very optimistic to very pessimistic, with most being somewhere between.

There has, however, been some controversy about whether the optimism construct should be seen as one bipolar dimension or whether there are two separable dimensions, one pertaining to affirmation versus denial of optimism, the other pertaining to affirmation versus denial of pessimism. This controversy stems mostly from the fact that factor analyses of the LOT-R usually produce two factors, one marked by the optimistically worded items, the other by the pessimistically worded items. The core question in the controversy is whether or not this separation (a pattern that is quite common in measures that have both positively and negatively worded items) reflects a substantive difference. Some people believe the two sets of items should be kept separate, others disagree. Sometimes separating the item sets from each other has led to better prediction of outcomes, but not always. A number of projects trying to settle the issue have come to different conclusions.

For the sake of simplicity, in this article optimism–pessimism is treated as one dimension. Keep in mind, however, that in some studies what mattered most was the extent to which people endorse versus reject a pessimistic outlook; in other studies what matters most was the extent to which people endorse versus reject an optimistic outlook. In yet other studies, this issue did not matter at all.

Optimism and Coping

As indicated at the outset, the definition of optimism and pessimism in terms of people's expectations for the future links the concepts closely to a family of motivation theories. The most obvious reflection of motivation is how people behave. One place where behavioral differences between optimists and pessimists have been investigated closely is how they cope with adversity. The ways in which optimists and pessimists differ in coping resemble the differences in broad behavioral tendencies discussed earlier in the article. That is, people who are confident about their eventual success continue trying to move forward, even when the going is hard. People who are doubtful about being able to succeed try to escape the adversity by wishful thinking; they are drawn into temporary distractions that do not help solve the problem; and they sometimes even stop trying to do anything.

Differences in coping between optimists and pessimists that correspond to this picture have been found in a number of studies (for detailed review and *meta*-analysis see Solberg Nes and Segerstrom, 2006). Early studies examined student reports of both situational coping responses and general coping styles, and found that optimists appear generally to be approach copers and pessimists appear to be avoidant copers. Conceptually similar results have come from many subsequent studies.

Other projects have studied what coping strategies were used spontaneously by people in specific difficult contexts (for detail, see Carver et al., 2010). For example, a study of coronary artery bypass surgery assessed people's attentional cognitive strategies as ways of dealing with the experience. Before surgery, optimists more than pessimists reported already making plans for their future and setting goals for recovery. Optimists also focused less than did pessimists on negative aspects of the experience – distress and symptoms. Once surgery was past, optimists more than pessimists reported seeking out information about what the physician would require of them in the months ahead. There was also evidence that the positive impact of optimism on quality of life 6 months later occurred through the indirect effect of these differences in coping.

Another study looked at how women coped with treatment for breast cancer during the first year after diagnosis and treatment. Both before and after surgery, optimism related to coping that involved accepting the reality of the situation as one that must be dealt with, placing as positive a light on it as possible, and trying to relieve the situation with humor. Pessimism related to overt denial (reports of trying to push the reality of the situation away) and to giving up tendencies at each time point. The coping responses that were related to optimism and pessimism were also related to distress. Further analyses revealed that the effect of optimism on distress was largely indirect through coping, particularly at postsurgery.

There are many distinct ways to cope, and many ways to categorize these various responses (for review see Carver and Connor-Smith, 2010). A particularly important distinction is between engagement or approach coping – aimed at dealing with the stressor or emotions stemming from it – and disengagement or avoidance coping – aimed at escaping the stressor or emotions stemming from it. In their *meta*-analysis of optimism and coping, Solberg Nes and Segerstrom (2006) fit a range of specific coping responses from various studies into four categories. Optimism was positively associated with broad measures of engagement coping, and was about equivalently related to two subsets of engagement coping responses: those that were problem focused (e.g., planning, seeking instrumental support) and those that were emotion focused (e.g., cognitive restructuring, acceptance). Optimists also varied their coping responses according to what sort of

stressor was being confronted. They reported more problem-focused coping with controllable stressors (such as academic demands) and more emotion-focused coping with uncontrollable stressors (such as trauma). Thus, optimism predicted active attempts to both change and accommodate to stressful circumstances, in ways that reflect flexible engagement.

The pattern for disengagement coping was opposite that of engagement coping. Optimism related inversely to disengagement coping, and to both problem-focused disengagement (e.g., behavioral disengagement) and emotion-focused disengagement (e.g., denial, wishful thinking). Thus, as would be expected from the expectancy-value motivational viewpoint, the relationship of optimism to coping differed substantially between engagement and disengagement.

In sum, optimists differ from pessimists in stable coping tendencies and in ways of coping when confronting specific stressful situations (Solberg Nes and Segerstrom, 2006). Of particular note maybe the contrast between acceptance and active denial. Denial (refusing to accept the reality of the situation) means trying to maintain a worldview that is no longer valid. Acceptance implies restructuring one's perceptions to come to grips with the situation. Optimists accept that reality, whereas pessimists try to deny it as long as possible. It should be stressed that acceptance here does not mean giving up. Acceptance of the reality of a challenge, even something as severe as a fatal illness, has different consequences. By accepting that one's life is compromised (in the case of a fatal illness), optimists shift the parameters within which they will live the time they have left. Acceptance may actually serve the purpose of keeping the person goal-engaged and indeed 'life-engaged.'

Fostering and Interfering with Well-Being

The concept of coping with adversity broadens easily to include a variety of related content areas. A simple extension is to what has been called preventive or proactive coping, processes that proactively promote good health and well-being rather than just reacting to adversity. An interesting question is whether optimists take active steps to ensure positive outcomes in their future, as opposed to just waiting for those outcomes to occur. This would resemble problem-focused coping, except that it is intended to prevent a stressor from arising in the first place.

An example of how this might occur is seeking knowledge pertaining to areas of potential risk. For example, middle age is a time when a variety of health risks begin to become relevant to people's lives. Would you expect adults who are optimistic about life to make the effort to learn about such risks, or to just ignore the question? Perhaps surprisingly, people high in dispositional optimism have been found to actually know more about risk factors for heart attacks than people who are less optimistic. They seem to be taking steps to be prepared for potential challenges (for review of studies see Carver et al., 2010).

Proactive efforts in health promotion have also been studied among patients recovering from heart attacks. In this research optimism was associated with success in lowering saturated fat, body fat, and an index of overall coronary risk. Optimism was also related to an increase in exercise. In another study of coronary artery bypass patients, 5 years after surgery, optimists were more likely to be taking vitamins, eating low-fat foods, and be enrolled in a cardiac rehabilitation program. Thus, optimists do not seem to take for granted that everything will turn out fine without their own personal involvement. Instead, they seem to believe that they have opportunities to help produce those outcomes, which they should take advantage of.

It might seem paradoxical that people who expect good things to happen take active steps to make sure good things *do* happen. But experience presumably teaches people that their own efforts play an important part in many kinds of life outcomes. Optimists maybe more confident about their outcomes because they are more confident that their efforts will be successful. For that reason, they are quicker to engage those efforts when there is a need for them.

If optimists are persistent in trying to reach goals, then pessimists should therefore be less persistent and more likely to give up. There is, in fact, a variety of evidence of giving up tendencies among pessimists. These tendencies often have adverse consequences. For example, giving up may underlie excessive alcohol use, which is often seen as an escape from problems. Pessimists are more vulnerable than optimists to such maladaptive behavior.

Another study examined a different indicator of giving up the disruption of normal social activities. In that study (Carver et al., 2003), breast cancer patients rated illness-related disruption of social activities at several points after treatment. At each assessment, pessimism predicted more disruption, along with emotional distress and fatigue. When confronted with a health threat, then, pessimism can lead to a withdrawal from the social activities that are important to a normal life.

Optimism as an Engine of Resources

Health psychology has probably been the main arena for studying effects of optimism, but not all research on this trait comes from health psychology. The tendency among optimists toward persistent goal pursuit, their active coping with stressors, and even their better health, may make it possible for them to translate their short-term approach tendencies into the longer-term development of resources.

One kind of resource is socioeconomic. Although this topic has not been much studied with respect to optimism, the available evidence points toward correlations between optimism and two indicators of socioeconomic status (SES): education and income. In one large sample of first-year undergraduates, dispositional optimism before starting school predicted a significantly higher probability of returning for the second year (Solberg Nes et al., 2009). The dropout rate for pessimists was roughly twice that for the very optimistic. In a smaller study of law students, optimism before starting school predicted higher income 10 years later.

Interestingly, socioeconomic resources also predict the development of optimism over time. One study assessed parents' SES (by education level, occupational class, and employment status) among children who were either 3 or

6 years of age. The children were assessed 21 years later, at 24 and 27 years old. There was a significant association between initial parental indicators of SES and adult optimism 21 years later, even when adult SES was controlled. Thus, poor childhood socioeconomic circumstances lead to pessimism later in life.

Another important kind of resources is social capital. Optimists also benefit in this domain. For example, a study of how students coped with the challenge of starting college found that optimists experienced greater increases in their social networks across the first semester than did pessimists. Other research has also found associations between expecting positive outcomes in the future and having broader social networks.

Why do optimists have better social connections than pessimists? One reason is that optimists are easier to like than pessimists. Studies confirm that people are more accepting of someone who expresses positive expectations for the future and more rejecting of someone who expresses negative expectations. Another study found that social interactions with optimistic people are more positive than those with less optimistic people. In yet another study, pessimism among men about to undergo coronary artery bypass surgery predicted greater caregiver burden from their wives 18 months later.

Another contributor may derive from the fact that optimists tend to see things in the best light. It seems likely that this includes things pertaining to their relationships. This might make optimists more satisfied in the relationship even when things are not perfect. Consistent with this idea, one recent study found that optimists had greater relationship satisfaction than pessimists, a difference that was mediated by their perceptions that partners were very supportive of them. Of course, it maybe that partners of optimists really *are* more supportive than partners of pessimists, because optimists are easier to like and thus support. This study controlled for that possibility, however. Even with that control, optimists perceived more supportiveness in their partners than pessimists.

Yet another reason why optimism represents a resource for relationships maybe that optimists work harder (or work more effectively) at their relationships. This would be consistent with their greater engagement with high priority tasks. In a study bearing on this question (Srivastava et al., 2006), relationship partners first had a conversation in the laboratory about the area of their greatest disagreement. Then the couples rated both their own behavior and their partner's behavior during that interaction. From this was created an index of engagement (being a good listener, not criticizing, trying to understand the other's point of view). Optimists were rated as being more engaged than those who were less optimistic. A week later, the couples were asked how well the conflict they had discussed had been resolved. More resolution came from those who were more optimistic and that effect was partly mediated by positive engagement in the previous discussion.

Another recent project tested the idea that optimists would have an orientation to relationships that fosters effective problem solving, just as optimists engage in problem-focused coping when under stress. This project (Assad et al., 2007) studied married couples across a 2-year period. Participants completed measures of cooperative problem solving, both for themselves and for their spouses. They were also recorded while discussing their relationship. Raters coded the recordings for relationship quality and negative interactions. Optimism was associated with better relationship quality, with less negative interactions, and with higher levels of cooperative problem solving.

In sum, although there are relatively few studies of the role of optimism in relationships, the evidence that does exist is consistent in indicating that pessimists have a harder road than do optimists. Given the importance of close relationships to life, this represents yet one more area in which the optimist appears to have the advantage.

Is There a Downside to Optimism?

Thus far, it seems that optimists have somehow found the keys to a rich and fulfilling life. Compared to people who are more pessimistic, they experience less distress when they encounter adversity. They cope with difficult situations by remaining engaged in the goals and activities that the stressor is threatening. They engage in problem-focused coping when there is something to be done, and they engage in accommodative coping when the adversity simply has to be endured. Perhaps as a result of these differences in coping, they also have better health-related outcomes and better social connections, both broadly and in their close relationships. These properties sound quite adaptive. Optimists seem to have the kind of life most of us want.

Some have asked, though, whether optimism can potentially lead to problems in certain kinds of contexts. Consider problem gambling. Gambling can cost large amounts of money and create serious problems in work and relationships. Gambling is a context in which having expectancies of good outcomes and being highly persistent might be counterproductive. Indeed, research has found some worrisome tendencies regarding gambling among optimists. Optimists have more positive expectations for gambling than did pessimists and they were less likely to reduce their bets after poor outcomes. The people in that research were not people with actual gambling problems. But the evidence suggests the possibility that optimists maybe more likely to develop such problems than pessimists.

Another set of studies has asked whether the persistence of optimists can create problems because they fail to recognize what they cannot accomplish. More simply, maybe optimists do not know when they should quit. Certainly circumstances do exist in which people have to recognize that their goals are lost, and that the adaptive course is to turn away from them. Does the dogged persistence of the optimist prevent that from occurring?

One project on this question examined people's willingness to disengage from tasks on which they were failing (because the task actually was impossible). In one condition, there was no alternative task to turn to; in other conditions there was an alternative. When there was no alternative task to switch to, everyone persisted at the impossible task. When there was another task, optimism related to faster switching from the impossible task. In effect, optimists gave up on a task they

could not master in order to turn to a similar task that they *could* perhaps master.

Another question is whether optimism causes people to see only what they want to see and ignore threats. The first evidence suggested the opposite that optimists pay closer attention to information about health threats than pessimists, provided the threat is serious and is relevant to them. More recently, however, other researchers found the reverse effect. Other studies have tied optimism to an attentional bias toward positive over negative stimuli. Exactly, how to view the information available on this question is not clear. There do appear, however, to be at least some cases in which optimism has drawbacks or blind spots. It is not clear how circumscribed these cases are, but this doubtless will remain a topic for future work.

Cultural Issues

The description throughout this article has been relative coherent and internally consistent. It should be noted, however, that most of what is known about optimism comes from studies of North Americans, mostly of European descent. An important question is the extent to which the knowledge derived from these studies generalizes to other groups. The information on that question is limited, but so far there have been both differences and similarities.

One difference is that Asians seem to distinguish more sharply than do European Americans between affirmation of an optimistic view of the future and rejection of a pessimistic view of the future. The cultures do not consistently report overall mean differences in optimism (Chang, 2002). Thus far, the evidence suggests that optimism and pessimism relate to quality of life outcomes in the same general way across cultures. This line of inquiry is likely to become more important over time.

Concluding Comment and Future Directions

A large literature indicates that people who hold positive expectations for the future respond to adversity in more adaptive ways than people who hold negative expectations. Furthermore, optimism appears to confer benefits in both intrapersonal and interpersonal domains, even in the absence of stress. Expectancies influence how people approach both stressors and opportunities, and they influence the success with which people deal with them. There are some ways in which the focused efforts and persistence of the optimist can go awry; however, these cases are few in number compared to the benefits that optimism seems to confer.

Research has linked optimism to better emotional well-being, more effective coping strategies, and even to better outcomes in several aspects of physical health. The advantages of optimism also appear to translate into the domains of interpersonal relationships: optimists are better liked than pessimists, they benefit from their natural tendency to see things in the best light, and they appear to engage more productive effort in the sorts of problem solving that keep relationships alive.

It is clear that optimism is an individual difference variable that plays a key role in human experience. From a personality psychologist's view, this variable displays the important role of individual differences across many domains of life. From a clinical psychologist's point of view, the evidence reviewed here represents an opportunity to learn from optimists, so that the processes that characterize their approach to the world can be taught to pessimists.

See also: Coping across the Lifespan; Emotion and Expression; Motivation and Actions, Psychology of; Personality and Adaptive Behaviors; Personality and Risk Taking; Temperament and Motivation.

Bibliography

Assad, K.K., Donnellan, M.B., Conger, R.D., 2007. Optimism: an enduring resource for romantic relationships. Journal of Personality and Social Psychology 93, 285–297.

Boehm, J.K., Kubzansky, L.D., 2012. The heart's content: the association between positive psychological well-being and cardiovascular health. Psychological Bulletin 138, 655–691.

Carver, C.S., Connor-Smith, J., 2010. Personality and coping. Annual Review of Psychology 61, 679–704.

Carver, C.S., Lehman, J.M., Antoni, M.H., 2003. Dispositional pessimism predicts illness-related disruption of social and recreational activities among breast cancer patients. Journal of Personality and Social Psychology 84, 813–821.

Carver, C.S., Scheier, M.F., 1998. On the Self-regulation of Behavior. Cambridge University Press, New York.

Carver, C.S., Scheier, M.F., 2009. Optimism. In: Leary, M.R., Hoyle, R.H. (Eds.), Handbook of Individual Differences in Social Behavior. Guilford Press, New York, pp. 330–342.

Carver, C.S., Scheier, M.F., 2014. Dispositional optimism. Trends in Cognitive Sciences 18, 293–299.

Carver, C.S., Scheier, M.F., Segerstrom, S.C., 2010. Optimism. Clinical Psychology Review 30, 879–889.

Chang, E.C., 2002. Cultural influences on optimism and pessimism: differences in Western and Eastern construals of the self. In: Chang, E.C. (Ed.), Optimism and Pessimism: Implications for Theory, Research, and Practice. American Psychological Association, Washington, DC, pp. 257–280.

Eccles, J.S., Wigfield, A., 2002. Motivational beliefs, values and goals. Annual Review of Psychology 53, 109–132.

Rasmussen, H.N., Scheier, M.F., Greenhouse, J.B., 2009. Optimism and physical health: a meta-analytic review. Annals of Behavioral Medicine 37, 239–256.

Scheier, M.F., Carver, C.S., Bridges, M.W., 1994. Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the life orientation test. Journal of Personality and Social Psychology 67, 1063–1078.

Segerstrom, S.C., 2006. Breaking Murphy's Law. Guilford, New York.

Solberg Nes, L., Evans, D.R., Segerstrom, S.C., 2009. Optimism and college retention: mediation by motivation, performance, and adjustment. Journal of Applied Social Psychology 39, 1887–1912.

Solberg Nes, L., Segerstrom, S.C., 2006. Dispositional optimism and coping: a meta-analytic review. Personality and Social Psychology Review 10, 235–251.

Srivastava, S., McGonigal, K.M., Richards, J.M., Butler, E.A., Gross, J.J., 2006. Optimism in close relationships: how seeing things in a positive light makes them so. Journal of Personality and Social Psychology 91, 143–153.

Optogenetics

Dominic Mudiayi, Canadian Centre for Behavioural Neuroscience, University of Lethbridge, Lethbridge, AB, Canada
Scott Wong, Canadian Centre for Behavioural Neuroscience, University of Lethbridge, Lethbridge, AB, Canada
Aaron Gruber, Canadian Centre for Behavioural Neuroscience, University of Lethbridge, Lethbridge, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Optogenetics is a term encompassing a wide array of tools and techniques for manipulating cell function with light. The effector molecules are genetically encoded, which allow manipulations to be targeted to select cell types. This selectivity, along with the high temporal resolution afforded by light actuation has several important advantages over traditional tools for altering brain function, and it enables novel types of experiments. Optogenetic tools presently enable scientists to drive changes in cellular excitability, intracellular signaling, gene expression, and receptor function. Here, we introduce the foundations of this important new technology, and discuss its rapidly expanding capabilities for enabling neurobiological research.

Studying the Brain and Behavior

Traditional Tools for Manipulating Brain Function

Scientists have long sought to understand the physiological basis of normal and abnormal brain function. A common approach has been to assess changes in behavior and/or neural activity following manipulation of targeted brain regions or systems. The ideal experimental tool would allow for reversible functional manipulations of well-defined neuronal populations on timescales ranging from the action potential to behavioral epochs. Historically, three classic approaches have been used to selectively manipulate brain regions in order to dissect the neural components underlying complex behaviors: lesions, pharmacological manipulations, and electrical stimulation. Each of these has unique strengths and weaknesses.

Brain lesions usually cause permanent loss of neural tissue in a target region, and therefore have poor cell selectivity and temporal resolution. Lesions are often generated with electrical current, mechanical damage, or injection of toxic substances into specific brain regions. This technique is well suited for discerning whether a given brain structure is necessary for particular behaviors or information processing. The long-term and irreversible nature of the brain damage can foster adaptation in other brain structures and secondary effects on behavior that can present confounds (Traversa et al., 1997).

Pharmacological approaches have led to tremendous advances in neuroscientific discovery. Although pharmacological approaches are usually reversible and often highly selective for particular mechanisms, such as specific receptors, they are usually not selective to particular cell types. The temporal precision of this technique is typically limited by the rate of diffusion or metabolization of the drug/ligand used. Therefore, many drugs used to study important brain mechanisms impact the brain over longer timescales lasting from dozens of minutes to hours (Kramer et al., 2013).

Intracranial electrical stimulation of neural tissue enables manipulation with high temporal resolution, and has provided a powerful tool for investigating behavior, plasticity, and functional brain connectivity. Despite its continued fundamental importance in brain research, electrical stimulation has several limitations. Stimulation current is not selective for cell types, it generates electrical artifacts that obscure neural recordings, and evokes both orthodromic and antidromic activity propagation by directly stimulating afferent fibers as well as efferent axonal projections.

These classic tools for manipulating brain function continue to contribute tremendously to neurobiological research. They are, however, limited by an inability to reversibly activate or inactivate a predefined cell type with high temporal resolution. In the late twentieth century, neuroscientists began to increasingly use genetic tools to overcome the problem of cell selectivity. These tools allow for removing ('knocking out'), reducing ('knock down'), or inserting ('knocking in') genes encoding for proteins of interest in well-defined neuronal populations in freely moving animals. This approach, however, still lacked high temporal precision. An augmented strategy was needed to actuate changes more rapidly. Actuation refers to the control of cellular function by biological or nonbiological components that transduce input stimuli into physiologically relevant outputs. An actuator is the physical substrate required to achieve such control. Scientists realized that light-driven actuators derived from photoreceptors and light-driven protein effectors (e.g., protein pumps, channels, and G-protein-coupled receptors (GPCRs)) would have several useful properties. In principle, they could be genetically encoded so as to be expressed in select cells and could provide highly precise, and reversible, cellular actuation on a timescale similar to the processes that govern neuronal excitability (Miesenbock and Kevrekidis, 2005).

The use of genetically encoded, light-driven protein effectors to trigger temporally precise cellular actuation in neurons defined the origin of the field of optogenetics (Zemelman et al., 2002; Lima and Miesenbock, 2005; Boyden et al., 2005). The versatility of the optogenetic technique has since expanded neuroscientists' ability to investigate the functional impacts of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.55060-0

specific cell types and mechanisms on downstream and upstream neural microcircuitry, their functional connectivity, and the collective neural dynamics that underlie behavior (Fenno et al., 2011).

Opsins and Optogenetic Technology

Historical Development

Opsins form a large and diverse class of light-sensitive proteins that are ubiquitous in all kingdoms of life. There are two classes of opsins: type I opsins encapsulate both light sensing and ion conductance properties within a single protein unit and are only found in microbes; type II (animal) opsins are GPCRs found only in eukaryotes where they are involved in vision and pigmentation.

Initial investigations of light-driven protein effectors for use in the field of optogenetics were focused on the metabotropic photoreceptors of vertebrate and invertebrate eyes, which possess innate light sensitivity and indirectly drive changes in cell excitability (Zemelman et al., 2002). However, the focus shifted to ionotropic microbial opsins because these exhibit fast, direct light-dependent ion conduction across the cell membrane. Genomic studies have continued to uncover more members of the microbial opsin family, each with different conductance properties that allow for either depolarization or hyperpolarization of the membrane potential, thus leading to excitation or inhibition of neural action potential generation. The predominant opsins currently in use include halorhodopsin (chloride pump), channelrhodopsin (nonselective cation channel), Mac (proton pump), and archaerhodopsin (proton pump).

The light-sensing component (chromophore) in microbial opsins is a vitamin A derivative termed retinal. In type I microbial opsins, incident photons cause the all-trans form of retinal to isomerize, which facilitates conformational changes in the opsin protein. This leads to increased ion transport across the cellular membrane by either opening an ion channel or by actively pumping ions. Once isomerized, the chromophore in the opsin must be replaced with a new molecule that is in the all-trans isomer form. A source of all-trans retinal is therefore required for opsins to function. Intriguingly, many scientists were initially hesitant to investigate the use of microbial opsins to elicit membrane potential changes in mammalian cells based on the false premise that mammalian cells lacked sufficient sources of all-trans retinal. This was exacerbated by misconceptions about the feasibility of microbial opsin expression, toxicity, and functionality in mammalian cells (Yizhar et al., 2011).

Important early breakthroughs in optogenetic research came from Miesenbock and colleagues (Zemelman et al., 2002, 2003; Lima and Miesenbock, 2005). This group targeted a ligand-gated ion channel to dopaminergic neurons in *Drosophila* (fruit fly) and injected them with an inactive, photosensitive ligand. Light stimulation then activated the ligand, which initiated action potentials in the dopaminergic neurons and evoked the fly's escape behaviors. Although an important landmark in optogenetic research, this method lacked the very high temporal resolution needed for rapid, reversible control of neurons on the timescale of individual action potentials (Boyden et al., 2005). Therefore, the search continued for direct light-driven control of the protein channels and pumps.

The discovery of the microbial channelrhodopsin-2 (ChR2) by Nagel et al. (2003) led to subsequent breakthroughs by Boyden, Deisseroth, and colleagues, who showed that ChR2 could be expressed in cultured mammalian neurons using a delivery vector based on a virus to insert DNA into these cells (Boyden et al., 2005). They also showed that these cells could be activated by light to reliably evoke action potentials. This work demonstrated an efficient technique for cell type-specific targeting and optical control with high temporal precision on a timescale roughly 1000 faster than all previous approaches, allowing for manipulations on a millisecond timescale, rather than seconds or minutes (Boyden et al., 2005). This work also countered prior misconceptions that microbial membrane proteins would be excessively toxic to mammalian cells, and showed that there were sufficient levels of all-trans retinal in mammalian cells for ChR2 function. These advances set the stage for the subsequent rapid development and adoption of optogenetic approaches for neuroscience research in mammals.

Methodology

The concept of the optogenetic technique consists of two general steps. First, light-sensitive proteins are selectively expressed in the cell type of interest using genetic techniques. Second, these opsins are actuated with light.

Methods for Targeting Specific Cell Types

The first step to obtain selective optical control is to identify a gene that is specific to the target cell type. For example, dopaminergic and noradrenergic neurons are the only cells in the brain that express the tyrosine hydroxylase gene. The expression of genes is controlled by regulatory DNA sequences called promoters; these are typically positioned upstream to the DNA sequence of the gene that is to be expressed. Promoters enable enzymes to convert downstream genes into messenger ribonucleic acid, which is then translated into protein. Promoter activity can be widespread among many different types of neurons (e.g., Elongation Factor-1 alpha) or it can be confined to a specific class of neurons (e.g., tyrosine hydroxylase). Restricting the expression of opsins to select cells can be achieved by restricting the expression of opsin genes by promoters that are unique to the cells of interest.

The predominant method for introducing DNA into neurons is the use of a 'vector' derived from an inactive form of a virus. The vector is loaded with a DNA 'construct' that contains several DNA segments with distinct functions. The DNA construct usually encodes (1) a promoter region to allow expression of the other DNA segments, (2) the opsin, (3) membrane trafficking sequences, and (4) a fluorescent reporter protein that is fused to the opsin to facilitate the visualization of opsin expression. The encoded opsin protein, tagged with a fluorescent reporter, is subsequently localized in the neuronal cell membrane of select cells. The fluorescent reporter remains in the cytosol and does not interfere with

opsin function. Robust optogenetic response requires sufficient time for opsin densities to reach optimal levels in the membrane, which is typically between 3 and 5 weeks post vector injection (Yizhar et al., 2011). This incubation period varies depending on the type of viral vector used and the nature of the preparation (whether *in vitro* or *in vivo*). The appropriate promoter and membrane trafficking sequences must be selected to ensure that the genetically encoded microbial opsins are expressed at adequate levels and localized in the membrane of the specific class of neurons that are targeted.

Viral vector DNA delivery methods are a versatile means of transfecting cells and result in greater genetic and anatomical specificity than other common cell transfection techniques (Fenno et al., 2011). Another important advantage is that viral-based vectors can introduce multiple gene copies into the host cell, which increases protein density and photocurrents, leading to more robust optical control. Such vectors do not elicit an adverse immune response, and are widely used in experiments with rodents and primates. The tissue type targeted and extent of gene expression within that tissue can be modified by varying the viral coat glycoproteins (thereby changing the serotypes) or by using different viruses. However, viral-based vectors are limited in the size of the DNA construct that can be delivered, which can restrict the use of large promoter sequences. Furthermore, some vectors integrate the DNA construct into the host genome, whereas others do not and remain in plasmid form; this impacts the duration of expression. Finally, opsins can be selectively targeted to cell bodies, surrounding axons, or both by selecting the appropriate type of viral-based vector.

Two commonly used viral vector types in optogenetic experiments are lentiviral (LV) vectors and adeno-associated viral (AAV) vectors. While both are able to transfect both dividing and nondividing cells, they have several important differences. LVs are able to integrate randomly into the host genome, thus providing the potential for more long-term gene expression than nonintegrative vectors such as AAVs. Although LVs are easily produced with simple in-lab tissue culture techniques, they are typically designated for work under Biosafety level-2 (BL-2) laboratory conditions. Consequently, the majority of optogenetic assays make use of different serotypes of AAVs, which can normally be used under BL-1 laboratory conditions. Both AAV and LV are derived from live viruses, but they have been engineered with numerous safety measures. These vectors cannot replicate and are not known to cause disease (Yizhar et al., 2011).

Attempting to use virus-based vectors to express a DNA construct encoding an endogenous promoter and the microbial opsins in a specific class of neuron is possible, but faces several challenges. First, it is impossible to package some large cell-specific promoters into current vectors. Further, several small cell-specific promoters have been reported to be too weak to drive the expression of microbial opsins (Tsai et al., 2009). Second, the target cell must have membrane receptors that are recognized by the viral vector for vector insertion to take place. An alternate methodology that overcomes these problems has been adopted from genetic engineering techniques. Animals have been engineered with a permanent genetic modification to introduce a gene encoding a protein that is not present in wild-type animals (Arenkiel et al., 2007). These 'exogenous' genes can be expressed robustly in select cell types using genome modification approaches that do not depend on viral-based vectors, and therefore avoid their limitations. The predominant strategy is to use an exogenous enzymatic protein called cre recombinase as a 'driver' for functional opsin expression in the animal brain. Many rodent lines expressing cre recombinase in specific neuronal cell populations have been generated (Gong et al., 2007; Witten et al., 2011). Vectors carrying novel DNA constructs whose expression is dependent on cre recombinase can then be injected into a 'cre driver' animal's brain. The novel DNA construct is then only expressed in neuronal cells containing cre recombinase. Consequently, the cre driver lines are useful for functional opsin expression for several reasons. First, they allow the experimenter to select animals with cre recombinase in select cell types, and restrict the targeting specificity of the opsin by choosing an appropriate location for the vector injection. Second, they circumvent the problem of identifying regulatory elements (e.g., promoters, enhancers) of adequate size and sufficient functionality to fit in the viral envelope along with the opsin. These benefits allow researchers to optogenetically control many types of neurons using off-the-shelf vectors, which allow these techniques to be utilized by scientists with little molecular biology experience.

An alternative approach is to generate a transgenic line of animals that permanently express opsins in select cells (Arenkiel et al., 2007). This method has some drawbacks including low levels of expression by the endogenous promoters, and an inability to confine opsin expression to specific anatomical loci. The cre-driver approach is currently better in these aspects (Tsai et al., 2009; Witten et al., 2011).

Functionality

Biomolecular engineering and genomic studies continue to generate new opsins with a variety of characteristics including variation in the specificity of ion conductance, kinetics, and sensitivity to varying wavelengths of light. The functionality of opsins can be broadly divided in terms of their functional effect on the cell: direct activation, direct inhibition, bistable modulation, modulation of intracellular signaling, and altered gene expression.

Direct Activation

ChR2 was the first opsin to be used for direct activation of mammalian neurons (Boyden et al., 2005). When illuminated, it allows nonselective conductance of positively charged ions (cations), resulting in depolarization of the membrane and increased neuronal excitation. It is activated by blue light, although other related variants show sensitivity to a range of wavelengths corresponding to blue-green visible light (450–505 nm). Several variants of ChR2 have been produced in effort to tailor the functional characteristics of ChR2 to a variety of experimental uses, such as slower inactivation and changes in peak activation wavelength. For example, the H143R variant has an amino acid residue substitution that elicits larger photocurrents relative to the wild-type ChR2, although at the expense of having slower inactivation. A channelrhodopsin discovered in *Volvox carteri* called VChR1 is

another optically actuated cation channel used for excitation. Unlike ChR2, it has a peak activation wavelength that is shifted toward longer wavelengths (red-shifted), thus allowing use in parallel with ChR2 for separate optogenetic control of different cell types (Zhang et al., 2011).

Direct Inhibition

The most commonly used tools for direct inhibition are the outward proton pump archaerhodopsin (Arch) and the inward chloride pump halorhodopsin (HR). These are actuated by yellow-green light and yellow-red light, respectively. Both of these pumps have undergone mutagenesis to alter and/or improve their functional performance in inhibiting the generation of action potentials (Gradinaru et al., 2007). Whereas ChR2 is an ion channel conducting a depolarizing current, Arch and HR are ion pumps that generate polarizing currents and therefore inhibit action potential generation. Arch and HR variants are unable to generate photocurrents that are equivalent in magnitude to those generated by channels (e.g., ChR2). For this reason, functional inhibition during behavior therefore requires intense illumination (Zhang et al., 2011).

Bistable Modulation

Point mutation variants of ChR2 have been generated that enable bistable (on/off) control of the neuronal membrane potential. These ChR2 variants are termed step-function opsins (Yizhar et al., 2011). Brief illumination with blue light causes them to reconfigure to an active state in which they pass constant depolarizing current. Illumination with yellow light turns them to the inactive, nonconducting state. This approach allows neurons to remain responsive to native synaptic inputs rather than biasing neural activity by directly generating action potentials entrained with photo stimulation (Fenno et al., 2011).

Modulation of Intracellular Signaling

Rather than directly influencing membrane excitability, other types of opsins can instead interact with intracellular signaling cascades. For instance, rhodopsin is a eukaryotic type II opsin that is both a light-sensing protein and a GPCR. Such opsin-receptor (termed optoXR) chimeras based on the rhodopsin architecture have been generated to effect temporally precise, light-driven modulation of a specific intracellular cascade *in vitro* and *in vivo* (Airan et al., 2009). Generating an entire library of optoXR chimeras will undoubtedly help to unravel biochemical mechanisms essential for plasticity, behavior, and neurological disease. Variants of the optoXR family can also been applied to study nonexcitable cells, extending the reach of optogenetics outside of the brain (Fenno et al., 2011).

Gene Expression

Recently, the optogenetic toolbox has been extended to include the optical control of the expression of endogenous genes in mammals (Konermann et al., 2013). This system can easily be customized to target many endogenous genomic loci, and no exogenous factors are required. The rapidly expanding toolkit for driving gene expression is described in more detail in a subsequent section of this article.

Light Delivery and Neural Interfaces

The opsins expressed in the brain must be activated with light. The requirements for light delivery differ significantly depending on whether an *in vitro* (brain slice or cultured neuron) or an *in vivo* (anesthetized or freely behaving animal) preparation is used. For *in vivo* preparations, two solutions are presently used: light emitting diodes are implanted in the brain, or optical fibers are used to direct light from sources external to the brain to the desired targets. The fibers can either be inserted through permanently implanted guide cannulae for the duration of each experiment, or a small piece of fiber attached to a connector (called a 'ferrule') is permanently implanted in the brain so that the ferrule is accessible on the skull. The permanent implant approach is more widely used because it involves less risk of tissue damage or infection that arises from repeatedly inserting materials into the brain after surgery.

Delivering sufficient light to activate opsins to deep brain structures faces several challenges. First, the optical fibers are typically at least 100–200 μm in diameter, which damage tissue around the region of interest and may increase immune response. Second, light intensity rapidly decays away from the fiber tip due to diffraction and absorption by neural tissue (Yizhar et al., 2011). This heightens the importance of proper surgical placement of the optical fiber, and prevents uniform activation of cells within a structure. An online tool is available for estimating light power density at a given distance from the fiber tip in neural tissue (www.optogenetics.org/calc). Third, light in the brain produces heat. The effects of light-induced heating on neural tissue have not been studied enough to draw any definitive conclusions regarding the significance of thermal changes in neural tissue. This potential concern, however, limits the use of optogenetic approaches to illumination with brief light pulses. This is particularly limiting for inactivation studies during behaviors that require more than a few seconds. These challenges necessitate control experiments in animals lacking opsins to account for effects of illumination on neural tissue (Yizhar et al., 2011).

Emerging Directions

Restricted Transgene Expression in Homogenous Cell Populations

With the precise temporal control and ability to target specific cell populations offered by optogenetics, researchers have been able to investigate the function of the brain in ways that were previously impossible. This is demonstrated in a recent study by Liu et al. (2012), who investigated fear memory recall and its effects on behavior in mice. Animals first underwent a fear conditioning paradigm using an auditory tone followed by a mild electrical foot shock. Once conditioned, the mice demonstrated a freezing motor response to the tone, which was indicative of the retrieval of fear memory. Liu and colleagues then showed that the fear memory could also be retrieved by direct activation of neurons in the dentate gyrus of the hippocampus by selectively activating a subset of neurons in this structure with ChR2. This study illustrates that the

stimulation of a population of neurons is capable of eliciting macroscopic behavioral changes. Additionally, the research showcases a novel modification of established transfection techniques to enhance cell specificity by coupling ChR2 expression only to neurons that are actively depolarizing. This allows the optical stimulation to only activate neurons that were active during the formation of the initial memory engram, despite being in a larger population of the same cell type. To accomplish this, Liu and colleagues used molecular methods to restrict expression of ChR2 to cells that expressed *c-fos*, a gene that is rapidly expressed following increased neural activation. This novel method allows for even greater specificity in optogenetic experiments by restricting ChR2 expression to active neurons within a larger homogenous population.

Targeting Neuron Projections

Another example of further refinements in optogenetic techniques for elucidating brain circuits and their direct effect on behavior is work from Tye et al. (2011). These researchers demonstrated that the optical stimulation of terminals in the central nucleus of the amygdala (CeA) originating from the basolateral amygdala (BLA) exerts an anxiolytic effect in mice. Two classic methods to assess behaviors associated with anxiety are the open field test and the elevated-plus maze. Anxious animals spend less time in the center of the open field and the exposed arms of the maze. When ChR2 was optically activated in the terminals from the BLA, mice spent more time in these areas, suggesting they were less anxious. Conversely, using the inhibitory HR to silence these same terminals led to a decrease in the open areas, which is indicative of increased, anxiety.

Perhaps the most interesting discovery in these experiments was that while the stimulation of ChR2-expressing projections from the BLA to the CeA reduced anxiety, this was not true when the cell soma located in the BLA that were the source of these projections were preferentially stimulated with light. Tye and colleagues tested this by beveling the light-delivering fiber cannula so that light was directed toward the cell bodies rather than the terminals. This simple modification permits an even more focused delivery of light so that only select parts of a neuron can be bathed in illumination.

Modifications to Receptor Function

Another recent approach has been to modulate the functionality of specific receptors. Here, protein channels or receptors are covalently linked to an exogenous photo-switchable compound. This photo-switchable compound is fused to a ligand such that illumination with light will trigger the ligand to act on the protein channel/receptor (Kramer et al., 2013). Similarly to cell type-specific targeting in the optogenetic technique, a viral-based vector is used to carry and deliver a DNA construct encoding the modified protein to select cells. For example, any of the numerous subtypes of the voltage-gated K^+ channel can be modified with a covalently attached photo-switchable 'blocker.' Illumination with light will cause the blocker to inhibit the channel (Kramer et al., 2013). Although this approach does not use opsins, it provides tools for functional studies of endogenous channels or receptors with high temporal resolution rivaling that seen in opsin based optogenetic experiments.

Other Emerging Directions

In addition to using light to alter the function of cells, new optogenetic tools are being developed for reading out the activity from large populations of defined cell types. These are based on genetically encoded fluorescent proteins that report intracellular or extracellular physiological changes (membrane potential, Ca^{2+} concentrations, synaptic transmission, and metabolite fluctuations) in genetically targeted cell populations (Carlson and Campbell, 2013; Akerboom et al., 2012; Dreosti et al., 2009). For instance, viral vectors can be used to express the Ca^{2+} indicating protein, $GCaMP_3$, to targeted cells in mouse hippocampus. Then, a miniature fluorescence camera can be used to image Ca^{2+} dynamics in hundreds of cells in a freely moving mouse, providing a measure of neural activity (Ziv et al., 2013).

Although light can rapidly actuate protein function, it is limited to small regions of tissue and to brief illumination times due to diffraction and heat. Light is therefore not ideal for actuating proteins that are distributed widely in the brain, or actuating over long behavioral epochs. An alternate approach is to use a ligand as the actuator. The primary technique for this type of actuation is called Designer Receptor Exclusively Activated by Designer Drugs (DREADDs). DREADDs use chemical ligands in lieu of light to activate genetically encoded GPCRs. This enables targeting of specific G-protein signaling pathways with effects on membrane excitability or initiation of gene expression widely in the brain of freely behaving animals over the course of tens of minutes (Rogan and Roth, 2011).

Conclusion

Optogenetics is a revolutionary new technology for manipulating and monitoring brain function that overcomes several key problems of traditional methods, namely low cell type selectivity and low temporal precision. This rapidly expanding set of tools allows for control not only of the membrane potential, but also of gene expression, intracellular signaling cascades, and receptor function. Undoubtedly, the optogenetic toolbox will feature prominently in future neurobiological research and lead to breakthroughs in our understanding of the neural underpinnings of animal behavior and disease states.

See also: Anxiety and Fear Conditioning, Neural Basis of; Brain Stimulation Reward; Ion Channels and Molecular Events in Neuronal Activity.

Bibliography

Airan, R.D., Thompson, K.R., Fenno, L.E., Bernstein, H., Deisseroth, K., 2009. Temporally precise, *in vivo* control of intracellular signaling. Nature 458, 1025–1029.

Aravanis, A.M., Wang, L.P., Zhang, F., Meltzer, L.A., Mogri, A.Z., Schneider, M.B., Deisseroth, K., 2007. An optical neural interface: in vivo control of rodent motor cortex with integrated, fiberoptic and optogenetic technology. Journal of Neural Engineering 4, S143–S156.

Arenkiel, B.R., Peca, J., Davison, I.G., Feliciano, C., Deisseroth, K., Augustine, G.J., Ehlers, M.D., Feng, G., 2007. *In vivo* light-induced activation of neural circuitry in transgenic mice, expressing channelrhodopsin-2. Neuron 54, 205–218.

Akerboom, J., Chen, T.W., Wardill, T.J., Tian, L., Marvin, J.S., Mutlu, S., Calderon, N.C., Esposti, F., Borhguis, B.G., Sun, X.R., Gordus, A., Orger, M.B., Portugues, R., Engert, F., Macklin, J.J., Filosa, A., Aggarwal, A., Kerr, R.A., Takagi, R., Kracun, S., Shigetomi, E., Khakh, B.S., Baier, H., Lagnado, L., Wang, S.S., Bargmann, C.I., Kimmel, B.E., Jayaraman, V., Svoboda, K., Kim, D.S., Schreiter, E.R., Looger, L.L., 2012. Optimization of a GCaMP calcium indicator for neural activity imaging. Journal of Neuroscience 32, 13819–13840.

Boyden, E.S., Zhang, F., Bamberg, E., Nagel, G., Deisseroth, K., 2005. Millisecond-timescale, genetically targeted optical control of neural activity. Nature Neuroscience 8, 1263–1268.

Carlson, H.J., Campbell, R.E., 2013. Mutational analysis of a red fluorescent protein-based calcium ion indicator. Sensors 13, 11507–11521.

Chow, B.Y., Han, X., Dobry, A.S., Qian, X., Chuong, A.S., Li, D., Henninger, M.A., Belfort, G.M., Lin, Y., Monahan, P.E., Boyden, E.S., 2010. High-performance genetically targetable optical neural, silencing by light-driven proton pumps. Nature 463, 98–102.

Dreosti, E., Odermatt, B., Dorostkar, M.M., Lagnado, L., 2009. A genetically encoded reporter of, synaptic activity *in vivo*. Nature Methods 6, 883–889.

Elwassif, M.M., Kong, Q., Vazquez, M., Bikson, M., 2006. Bio-heat transfer model of deep brain, stimulation induced temperature changes. Conference Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society 1, 3580–3583.

Fenno, L., Yizhar, O., Deisseroth, K., 2011. The development and application of optogenetics. Annual Review of Neuroscience 34, 389–412.

Gong, S., Doughty, M., Harbaugh, C.R., Cummins, A., Hatten, M.E., Heintz, N., Gerfen, C.R., 2007. Targeting Cre recombinase to specific neuron populations with bacterial artificial chromosome constructs. Journal of Neuroscience 27, 9817–9823.

Gradinaru, V., Thompson, K.R., Zhang, F., Mogri, M., Kay, K., Schneider, M.B., Deisseroth, K., 2007. Targeting and readout strategies for fast optical neural control *in vitro* and *in vivo*. Journal of Neuroscience 27, 14231–14238.

Konermann, S., Brigham, M.D., Trevino, A.E., Hsu, P.D., Heidenreich, M., Cong, L., Platt, R.J., Scott, D.A., Church, G.M., Zhang, F., 2013. Optical control of mammalian endogenous transcription and epigenetic states. Nature 500, 472–476.

Kramer, R.H., Mourot, A., Adesnik, H., 2013. Optogenetic pharmacology for control of native, neuronal signaling proteins. Nature Neuroscience 16, 816–823.

Lima, S.Q., Miesenbock, G., 2005. Remote control of behavior through genetically targeted, photostimulation of neurons. Cell 121, 141–152.

Liu, X., Ramirez, S., Pang, P.T., Puryear, C.B., Govindarajan, A., Deisseroth, K., Tonegawa, S., 2012. Optogenetic stimulation of a hippocampal engram activates memory recall. Nature 484, 381–385.

Miesenbock, G., Kevrekidis, I.G., 2005. Optical imaging and control of genetically designated, neurons in functioning circuits. Annual Review of Neuroscience 28, 533–563.

Nagel, G., Szellas, T., Huhn, W., Kateriya, S., Adeishvili, N., Berthold, P., Ollig, D., Hegemann, P., Bamberg, E., 2003. Channelrhodopsin-2, a directly light-gated cation-selective membrane channel. Proceedings of the National Academy of Sciences of the United States of America 100, 13940–13945.

Rogan, S.C., Roth, B.L., 2011. Remote control of neuronal signaling. Pharmacological Reviews 63, 291–315.

Traversa, R., Cicinelli, P., Bassi, A., Rossini, P.M., Bernardi, G., 1997. Mapping of motor cortical, reorganization after stroke: a brain stimulation study with focal magnetic pulses. Stroke 28, 110–117.

Tsai, H.C., Zhang, F., Adamantidis, A., Stuber, G., Bonci, A., de Lecea, L., Deisseroth, K., 2009. Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. Science 324, 1080–1084.

Tye, K.M., Prakash, R., Kim, S., Fenno, L.E., Grosenick, L., Zarabi, H., Thompson, K.R., Gradinaru, V., Ramakrishnan, C., Deisseroth, K., 2011. Amygdala circuitry mediating reversible and bidirectional control of anxiety. Nature 471, 358–362.

Witten, I.B., Steinberg, E.E., Lee, S.Y., Davidson, T.J., Zalocusky, K.A., Brodsky, M., Yizhar, O., Cho, S.L., Gong, S., Ramakrishnan, C., Stuber, G., Tye, K.M., Janak, P.H., Deisseroth, K., 2011. Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. Neuron 72, 721–733.

Yizhar, O., Fenno, L.E., Davidson, T.J., Mogri, M., Deisseroth, K., 2011. Optogenetics in neural systems. Neuron 71, 9–34.

Zemelman, B.V., Nesnas, N., Lee, G.A., Miesenbock, G., 2003. Photochemical gating of, heterologous ion channels: remote control over genetically designated populations of neurons. Proceedings of the National Academy of Sciences of the United States of America 100, 1352–1357.

Zemelman, B.V., Lee, G.A., Ng, M., Miesenbock, G., 2002. Selective photostimulation of genetically chARGed neurons. Neuron 33, 15–22.

Zhang, F., Gradinaru, V., Adamantidis, A.R., Durand, R., Airan, R.D., de Lecea, L., Deisseroth, K., 2010. Optogenetic interrogation of neural circuits: technology for probing mammalian brain structures. Nature Protocols 5, 439–456.

Zhang, F., Vierock, J., Yizhar, O., Fenno, L.F., Tsunoda, S., Kianianmomeni, A., Prigge, M., Berndt, A., Cushman, J., Polle, J., Magnuson, J., Hegemann, P., Deisseroth, K., 2011. The microbial opsin family of optogenetic tools. Cell 147, 1446–1457.

Ziv, Y., Burns, L.D., Cocker, E.D., Hamel, E.O., Ghosh, K.K., Kitch, L.J., El Gamal, A., Schnitzer, M.J., 2013. Long term dynamics of CA1 hippocampal place codes. Nature Neuroscience 16, 264–266.

Relevant Websites

Information on Optogenetic Hardware and Tools

http://www.openoptogenetics.org/index.php?title=Main_Page – OpenOptogenetics.
http://www.stanford.edu/group/dlab/optogenetics/ – Optogenetics.
http://syntheticneurobiology.org/protocols – Synthetic Neurobiology Group.

Information on Transgenic Animals

http://research.jax.org/grs/optogenetics.html – The Jackson Laboratory.
http://www.rrrc.us/search/?strainStatus=available – Rat Resource & Research Center.

Information on Vector Supplies

http://www.med.unc.edu/genetherapy/vectorcore – UNC Vector Core.
http://www.med.upenn.edu/gtp/vectorcore/ – Perelman School of Medicine Gene Therapy Program.

Oral and Literate Expression

Ruth Finnegan, Open University, Milton Keynes, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The phrase 'oral and literate expression' refers to verbalizd formulations communicated through speech and writing respectively. 'Oral' and 'literate' are often seen as contrasted or complementary (hence their joint treatment in this article), but the relations between them, even their juxtaposition, are matters of debate. Recent research focuses on their overlap, taking them as dimensions of a multiplex and socially mediated continuum that continues to evolve. This article provides an overview of their main manifestations, then considers issues of definition and approach to their study, and lastly, discusses recent research trends.

The phrase 'oral and literate expression' refers to verbalized formulations communicated through speech and writing respectively. 'Oral' and 'literate' are often seen as contrasted or complementary (hence their joint treatment in this article), but the relations between them, even their juxtaposition, are matters of debate. Recent research tends to focus on their overlap, taking them as dimensions of a multiplex and socially mediated continuum. This article starts with a short overview of their main manifestations, then considers issues of definition and approach, the application of these concepts in the current situation of changing technologies, and recent research trends.

Spoken and Written Expression in Human Culture

Oral Expression

The rich complexities of oral forms, long ignored in the scholarly focus on written expression, are now being increasingly appreciated and studied (for examples and discussion, see Amodi, 2006; Feld, 1990; Finnegan, 2012, 2013; Foley, 2002; Okpewho, 1992; Perks and Thomson, 2006). Their nature and functioning are not necessarily identical to those of writing. In many disciplines, from philosophy and sociolinguistics to anthropology and education, the focus has shifted away from treating oral forms as if they are written texts toward active speech processes, conversations, and discourse-in-context. As with written genres, so too there are oral (or speech) genres; for example, different styles are expected for political speeches, spoken prayers, or personal greetings. Oral expression uses varying permutations of human voices (speaking, singing, chanting, shouting), singly or combined, with generic conventions based on, for example, appropriate verbal and performance style for speakers, audiences, context, or purpose.

These forms of expression are deployed in manifold ways in innumerable formal and informal situations, from intimate interchanges to public rituals. Many oral genres typically involve speakers present at the occasion and transient performances (unlike the greater permanence of writing), but some are now also enacted over far distances through electrical and electronic technologies or stored and reproduced in audio-recordings. Like writing, they are both interwoven with human organization and creatively used by individuals for an infinite number of purposes (Duranti, 1997; Finnegan, 2013a).

In most cultures people single out certain oral genres for heightened attention (Duranti, 1997; Feld, 1990; Finnegan, 2012; Okpewho, 1992). Far from being crude, communal, or uncreative, as once supposed, 'nonliterate' formulations can be extremely sophisticated in both verbal exposition and performance. Examples include the magnificent oral artistries of the twentieth-century South Slavic epics, the lofty diction of Zulu panegyrics, the brilliant imagery of Somali 'miniature' lyrics, the deeply felt 'wept thoughts' of some South Pacific laments, or contemporary performance poetry. Oral genres such as narratives, myths, epics, eulogies, protest poetry, laments, love poems, oratory, and song-texts (themselves a form of lyric poetry) are variously found throughout the world, and also include condensed but sometimes highly artful proverbs, jokes, anecdotes, or after-dinner tales. These are often described as 'oral literature' and although the term is not wholly satisfactory (it underplays performance dimensions, for example) it usefully highlights their literary qualities; a few scholars prefer the neologism 'orature'. Their study has enabled a radical reassessment among scholars of the achievements of oral expression, once regarded as being without creativity or individually shaped verbal artistry, notably in the lyrics of popular music and in the cultures of Africa, Native America, or Oceania.

Expression in Writing

Through the ages humans have developed many forms of writing: systems for communicating relatively specific ideas through standardized and relatively permanent visible marks, conventionally ordered and interpreted. Many systems are based on close relations between written marks and speech sounds, conveyed through syllabic, consonantal, or, most familiar to Western readers, alphabetic symbols (phonographic systems). Others, such as the Chinese and Aztek systems, use visual signs, sometimes pictorially based, to represent meanings rather than sounds (variously termed logographic, ideographic, and/or semasiographic systems); these allow communication across language boundaries, notably in the far-spread Chinese script. Some scholars include within writing pictographic systems such as American hobo signs or Native American pictorial records of events; others prefer narrower definitions like 'visible speech' (DeFrancis, 1989; Mallery

1893). The classifications and boundaries of writing are thus not fully agreed, partly because most systems combine several principles and the relations between speech (and other) sounds and written signs are variable (Boone and Mignolo, 1994; Daniels and Bright, 1996; Ellingson, 1986; Harris, 2000; Van Leeuwen, 1999).

Writing appears in many different materials and forms, versatile enough to be turned to almost any of the multifarious purposes of human interaction. These include encapsulating, storing, and transmitting information; counting and accounting (sometimes seen as the origin of writing); formulating and extending memory; bridging space and time; and facilitating personal and artistic expression. Writing is also used in the formulation of recognized literary and religious works and their communication within and across human cultures. Western Europe, China, and the Islamic world are famous examples; but throughout history and, nowadays, in every area of the globe, writing has played powerful roles – for good or for ill – in large-scale organizations; in personal and group expression at every level; and in the human endeavor to transcend space and time. It is inextricably embedded in human social institutions, for example, in the educational, artistic, and organizational hierarchies of the modern West – but not only there (Barber, 2006; Barton and Hamilton 1998; Cavallo and Chartier 1995; Coleman, 1996).

The Concepts 'Oral' and 'Literate': Problems and Approaches

The distinction between oral and literate expression seems at first self-evident, familiar from phrases like 'the spoken and written word'. However 'oral' and 'literate' are complex and multiform concepts, both in their varying cultural manifestations and in the theoretical controversies surrounding them. They are all the more 'loaded' because they invoke emotive and far-reaching issues about the nature and history of human culture.

The terms themselves are awkward in English. 'Oral' mostly refers to what is spoken or uttered by the mouth, but it also sometimes refers to anything that is not written: the first sense excludes deaf signing and nonverbalized material culture, the second includes them; phrases like 'oral tradition' can thus be ambiguous. The term 'literate' partly complements 'oral' but is not precisely parallel and is sometimes replaced by the term 'written'. While it can readily be assumed that all members of a society normally gain some capacity to speak and hear, the skills of reading and (not necessarily the same) of writing are less widely distributed – though contrary to some presuppositions, many so-called oral societies also recognize training and expertise in particular forms of spoken art and forms that are proscribed to certain sections of society (Bauman and Sherzer, 1989; Finnegan, 2012, 2013a).

The referents and connotations of each term have also been channeled through a series of theoretical frameworks and questions. Many of these broadly match those elsewhere in the social sciences, successively dominated by, for example, evolutionist, diffusionist, and functionalist analyses; ideas of reflection; structuralism, poststructuralism, and post-modernism; and a variety of Marxist-inspired and cognitivist approaches. Each has it own 'fix' on the terms and their interrelationships. I have selected a few below to illustrate the contrasting and changing perspectives.

Romantic and Evolutionary Paradigms

Some interest in contrasts between spoken and written expression dates back at least to classical Greece but in more recent centuries the most influential model has posited a teleological progression in human history, from oral to written. Spoken and written forms are viewed as themselves distinct and as characterizing distinctive evolutionary stages, with writing destined to supersede the oral phase in the upward path to enlightened modernity. Complementing and partly overlapping this is a nostalgia about what was supposedly sacrificed in the process: a romantic glorification of that lost 'other'. Thus so-called oral cultures are visualized both as the valued repository of deep-seated, communal, and emotionally sensitive qualities, as well as a 'primitive stage' that modern peoples have left behind.

Consequences of Orality and Literacy

The *consequences* of oral and literate forms of expression attracted much scholarly and popular interest through the mid- and late twentieth century. Somewhat counter-intuitively this seems in practice to have been started off by the work on literacy, this in turn drawing attention to what before had just been presupposed as the 'natural' – and thus uninteresting – order of things. The British anthropologist Jack Goody argued from the 1960s that the functions and effects of literacy should be studied comparatively (Goody, 1968). Parallel questions were raised in history, sociology, and comparative literary studies, further reinforced by interest in development, economic 'modernization', educational applications, and the major comparative features of societal organization. Marshall McLuhan's writings on the media (1964) stirred popular imagination and intermixed with continuing evolutionary and romantic perspectives. The result was intensive discussion about the social and cognitive consequences of literacy. Among them were rationality; objectivity; the emergence of history, science and/or philosophy; secularism (but in other accounts religious obscurantism); political and religious control (or alternatively, freedom); modernization – the 'take-off' for economic development and/or imperial rule both in early history and the present; innovation; the transition from auditory to visual perception; bureaucracy; and individualism.

The suggested consequences between them thus spanned almost every feature associated either ideally or in practice with Post-Enlightenment Western society; and chimed well with narratives of the print-based Protestant Reformation and of the expansionist West disseminating literacy through its civilizing mission.

Most of this presupposed some independent variable from which these consequences flowed. Usually this was literacy, with orality being implicitly its more shadowy opposite. However just what this literacy comprised was less clear. Some of the posited effects occurred in some eras but not in others or were even mutually contradictory (e.g., literacy leading both to freedom *and* to oppression), so they could scarcely be

automatic results from some uniform entity. As in other areas of the social sciences, many analysts thus now speak more of 'implications' rather than 'unicausal consequences', and of 'possibilism' rather than 'determinism', recognizing that literacy takes many forms and is used differently in different settings.

The Orality–Literacy Debate

Linked to the interest in consequences was the so-called orality–literacy debate. This revolved round the question of whether there is some generalizable divide between oral and literate cultures, and/or between oral and literate minds.

On the one side were writers like Marshal McLuhan (1964, 1982), Goody (2000), and Olson and Torrance (1991), who not only strongly contrasted 'literacy' and 'orality' but often envisaged human history as a series of major changes – from orality, to writing, to print. Each stage was seen as having distinctive properties, with the 'Great Divide' between 'primitive' and 'civilized' marked by literacy. This resonated with evolutionary and romantic notions, carrying connotations of orality as primeval, emotional, and homogeneous, characterized by unchanging and communal tradition rather than individual creativity, as against the rationality, individualism, science, urbanism, modernity, and economic development associated with literacy. This in turn recalled the binary division between types of society envisaged by classic nineteenth and early twentieth century social theorists.

Other scholars argued that the basic question was 'over general', giving a spurious identity to multiplex notions (Boyarin, 1993; Finnegan, 2013b; Schousboe and Larsen, 1989; Street, 1993). They challenged the ethnocentric Western bias of the contrast and its implicit technological determinism. In their view literacy and orality are neither autonomous nor unitary. The practices of speaking and writing coexist, they are socially as much as technologically shaped, and cultural variations undermine sweeping generalizations.

A further twist came in McLuhan's (1964) and Ong's (1982) vision of electrical and electronic technologies as ushering in a new age of 'secondary orality'. This further linked to suggestions that the apparent bias toward vision in Western cultures, often associated particularly with Gutenberg's printing press, was now being redressed by new opportunities for audio-communication, and hence providing a greater role, once again, for oral expression. The position is complicated, not least because others propound a new 'visual turn', but computer technologies certainly open new opportunities in many directions. They also challenge older assumptions about oral forms being fleeting, evanescent, and essentially variable in contrast to the fixed, permanent, and authoritative properties of writing. Computer 'soft' text transcends that postulated divide and erodes yet further the once-confident distinctions (Danet, 1997).

The trend is now to take more account of diversity and change rather than generalizing about 'orality' or 'literacy' as supposedly uniform entities. But the question is constantly being recycled, sometimes by writers unacquainted with recent – or, indeed earlier – work, and still stirs both academic and popular interest.

Ethnographic and Historical Specificities

Many scholars now look instead toward empirical investigation of specific cases (Amodio, 2005; Barber, 2002; Bauman and Sherzer, 1989; Cope and Kalantzis, 2000; Danet, 1997). Here 'oral' and 'literate' turn out to be less hard-edged and autonomous than once assumed and to come in varied, culturally specific, forms. Speaking and writing intermingle in manifold genres rather than exist in monolithic and clearly separated form. The famous Greek epics of Homer, for example, were probably first composed and performed orally, but then transmitted in writing. Oral and written traditions are not independent streams, as once supposed; the same works sometimes appear in both modes and move between them (Coleman, 1996; Finnegan, 1992, 2013a; Henigan, 2012). Examples are national anthems, carols, well-known religious texts, 'urban myths', Hausa political poetry, and English-American ballads.

In classical and medieval Europe, literary works were commonly composed for the voice and 'published' by being read or performed aloud to an audience; comparable interplays of oral and written processes occur in sermons, plays, or conference papers today (Amodio, 2005; Coleman, 1996). Song lyricists have often used mixes of written and oral media to compose sung words and guide their performances. Similarly written formulations are related to oral practices in a huge variety of ways, from dictated transcriptions and performance scores to crib sheets, prompt books, or jotted notes. Radio and television broadcasts and the icons, graphics, and audio-messages of modern computers further blur the once-clear divisions (for discussion and illustration see Boone and Mignolo, 1994; Boyarin, 1993; Cavallo and Chartier, 1995; Fardon and Furniss, 2000; Jorholt, 2001; Van Leeuwen, 1999).

The notions of oral versus literate thus no longer appear definitive and clearly opposed because they so readily coexist at any given time (Finnegan, 2012; Hofmeyr, 2004; Julien, 1992; Peek and Yankah, 2004), nor are they seen as carrying already known attributes when applied to cultures, genres, or individuals whether now or in the past. Rather they emerge as broad and somewhat ill-defined terms covering a spectrum of varying practices and artefacts, resources which human beings draw on in a multitude of overlapping, merging, and intermingled ways.

Recent Insights and Future Directions

While the earlier approaches described above still carry influence, current research could be said to be characterized by four interrelated trends.

Multiplicity and Differentiation

Recent analyses continue and extend the interest in historically and culturally specific contexts. Scholars now seldom speak of purely 'oral cultures', recognizing not only the lengthy and geographically widespread interconnections between oral and graphic systems, but also the complex and shifting varieties of current spoken/written/electronic interactions. Multiple versions of writing have existed and coexisted both in the past and in more recently developed systems, like the little-known

but ingenious indigenous scripts of West Africa (Finnegan, 2012). A definition based on Western history and alphabetic writing is no longer accepted as universally valid (Cavallo and Chartier, 1995; Cope and Kalantzis, 2000; Street, 1993).

Unlike earlier models envisaging homogeneous patterns of *either* orality *or* literacy throughout a culture as a whole, recent studies explore how different groups or individuals use and/or conceptualize speech and writing in diverse directions, have differential access to them, and deploy them differently in different situations. Even within the same locality, different categories of people use writing (or, indeed, speaking) in different ways. Some can read but not write; others may be multiply literate in different languages or different genres, using them for different purposes; others again may specialize in particular forms. There are accepted social controls on who can speak, how, and where, as well as on the learning, practice, and purposes of writing (Duranti, 1997; Finnegan, 2013a).

The focus is thus now on the multiplex, localized, and differentiated ways that both speaking and writing are learned, valued, or practised according to, for example, gender, rank, religious or educational role, age, geographical area, ethnicity, or situation. Critical and postmodernist analyses emphasize that both ideologies and practices may be related to the exercise of power, and can divide as well as unite. These spill over into studies of popular culture and the contemporary media, and of the role of literacy in the ideologies and practices of imperialism (Cope and Kalantzis, 2000; Street, 1993)

Practices and Processes Rather Than Texts and Products

A further trend goes beyond models of oral and literate expression as forms of text. This partly dates back to work on performance by linguistic anthropologists, folklorists, and others from around but has intensified since the 1960s through more recent social science interests in process and practice (Bauman and Sherzer, 1989; Gerstle et al., 2005; Schechner, 2002). Many scholars have turned from studying fixed products, formal typologies, or even cognitive meanings, toward investigating everyday situations, intertextual processes, emotive experiences, and the diversity of socially mediated expectations about how people do or should communicate.

Such studies throw up the creative and diverse involvements of individuals in oral and written interaction, revealing something of the infinite inventiveness of human beings in the composition, enactment, and reproduction of verbal expression. Even in a single communicative situation, different individuals may be participating differentially, actively constructing the communicative process rather than merely forming a passive or homogeneous 'audience' for an already defined message. This can be applied to written forms too. Readers approach and interpret their material in diverse ways, with marked differences between cultures, historical periods, sectional groups, and even individuals (Boyarin, 1993; Cavallo and Chartier, 1993; Manguel, 1996). The reality lies there as much as in the ostensibly objective and enduring text.

The investigation of 'text' continues (notably Barber, 2007; Honko, 2000) but the term is now less often taken as a given. Rather it is approached as a multifaceted issue for exploration, opening up questions about, for example, the socially shaped ways in which actions and communications are, or are not, crystallized in verbalized forms; how these relate to one another and to other aspects of social life; and the diverse processes through which they are actively put together and experienced.

The Senses and Multimodality

This article started by defining the key concepts in terms of 'verbalized' expression. But current research increasingly emphasizes that verbal expression consists of more than words. Speaking uses uttered words, certainly, but also regularly relies on concurrent 'visual communication', through participants' bodily gestures, movements, and positioning. In writing there are images, significant layouts, and visual displays over and above signs representing speech sounds, and sometimes tangible, acoustic, and olfactory dimensions. The materiality of artefacts such as books or letters now attracts interest, both in new transdisciplinary approaches to material culture and because computer developments challenge older and established print forms; and in so doing raises awareness of properties previously taken for granted. The remarkable surge of interest in the senses has also valuably added to the analysis of both oral and written media by drawing attention not only to the visual and auditory dimensions but also the olfactory, tactile, musical, and dynamic features which so enhance – or, rather, constitute – both written and spoken expression (Finnegan, 2014; Howes, 2005).

Earlier ideologies positing the neutrality, universality, and transparency of print now themselves turn out to be culturally moulded and implicated in material artefacts, not just mental representations. The multimodality of most, perhaps all, oral and written expression is likely to form one revealing area of research in the future.

The Virtues of Ethnography

There is currently great interest not just in the importance of accurate ethnographies, but also their comprehensive grasp. This has valuable implications for the study of oral and literate expression for it means that rather than positing some divide or, at best, interaction or transition between the two (not that this is always irrelevant; see, for example, Amodio, 2005; Ricard and Veit-Wild, 2005) the emphasis is now on a full account of the many media, senses, and genres, with all their complexities and elusiveness, in a given cultural context (Barber, 2002; Barton and Hamilton, 1998; Fox, 2000; Henigan, 2012; Hofmeyr, 2004). In this context, verbal expression in all its forms seems to shade into the many other art forms of culture, an amazing spectrum from the pictorial, and static to the dynamics of the fleeting dynamic art forms (Finnegan, 2013b; Howes, 2005).

There are other perspectives on speech and writing but a virtue of focusing on the interrelation of oral and written expression, even on partially outmoded debates, is that it brings to the surface and challenges, some powerful but often-unacknowledged presuppositions. In doing so it opens doors to a greater appreciation of both the continuities and the multiplex powers of these primarily (but not exclusively) verbal resources of human expression.

See also: Anthropological Writing; First Language Acquisition, Developmental Psychology of; Language Acquisition; Language and Social Class; Language and Society; Linguistic Anthropology; Popular Culture; Speech Perception; Speech Production, Psychology of; Writing Systems, Psychology of; Writing Systems.

Bibliography

Amodio, M.C. (Ed.), 2005. New Directions in Oral Theory. Arizona Center for Medieval and Renaissance Studies, Tempe, AZ.

Barber, K. (Ed.), 2006. Africa's Hidden Histories: Everyday Literacy and Making the Self. Indiana University Press, Bloomington, IN.

Barber, K., 2007. The Anthropology of Texts, Persons and Publics. Oral and Written Culture in Africa and beyond. Cambridge University Press, Cambridge.

Barton, D., Hamilton, M., 1998. Local Literacies. Reading and Writing in One Community. Routledge, London.

Bauman, R., Sherzer, J. (Eds.), 1989. Explorations in the Ethnography of Speaking, second ed. Cambridge University Press, London.

Boone, E.H., Mignolo, W.D. (Eds.), 1994. Writing without Words. Alternative Literacies in Mesoamerica and the Andes. Duke University Press, Durham, CA.

Boyarin, J. (Ed.), 1993. The Ethnography of Reading. University of California Press, Berkeley, CA.

Carpenter, E., McLuhan, M. (Eds.), 1960. Explorations in Communication. Beacon Press, Boston, MA.

Cavallo, G., Chartier, R. (Eds.), 1995. Storia Della Lettura Nel Mondo Occidentale. Laterza, Roma [1999. A History of Reading in the West. Polity, Cambridge].

Coleman, J., 1996. Public Reading and the Reading Public in Late Medieval England and France. Cambridge University Press, Cambridge.

Cope, B., Kalantzis, M., 2000. Multiliteracies. Literacy Learning and the Design of Social Futures. Routledge, London.

Danet, B., 1997. Books, letters, documents. The changing aesthetics of texts in late print culture. Journal of Material Culture 2, 5–38.

Daniels, P.T., Bright, W., 1996. The World's Writing Systems. Oxford University Press, New York.

DeFrancis, J., 1989. Visible Speech. The Diverse Oneness of Writing Systems. University of Hawaii Press, Honolulu, HI.

Duranti, A., 1997. Linguistic Anthropology. Cambridge University Press, Cambridge.

Ellingson, T., 1986. Buddhist Musical Notations. In: Tokumaru, Y., Yamaguti, O

Fardon, R., Furniss, G. (Eds.), 2000. African Broadcast Cultures. Radio in Transition. James Currey, Oxford.

Feld, S., 1990. Sound and Sentiment: Birds, Weeping, Poetics, and Song in Kaluli Expression, second ed. University of Pennsylvania Press, Philadelphia, PA.

Finnegan, R., 1992. Oral Poetry: Its Nature, Significance and Social Context, second ed. Indiana University Press, Bloomington, IN.

Finnegan, R., 2012. Oral Literature in Africa. second ed. (illustrated). Open Book, Cambridge.

Finnegan, R., 2013a. Studying the Oral, a Handbook for 21st Century Researchers. Callender Press, Milton Keynes.

Finnegan, R., 2013b. Reflections on Language, Literature and Performancel. Callender Press, Mioton Keynes.

Finnegan, R., 2014. Communicating. The Multiple Modes of Human Communication, second ed. Routledge, London.

Foley, J.M., 2002. How to Read an Oral Poem. Illinois University Press, Champaign, IL.

Fox, A., 2000. Oral and Literate Culture in England 1500–1700. Oxford University Press, Oxford.

Gerstle, A., Jones, S., Thomas, R. (Eds.), 2005. Performance literature. Oral Tradition 20, 1–2(special issue).

Goody, J. (Ed.), 1968. Literacy in Traditional Societies. Cambridge University Press, London.

Goody, J., 2000. The Power of the Written Tradition. Smithsonian Institution Press, Washington, DC.

Harris, Roy, 2000. Rethinking Writing. Athlone Press, London.

Henigan, J., 2012. Literacy and Orality in Eighteenth-Century Irish Song. Pickering & Chatto, London.

Hinde, R.A., 1975. The comparative study of non-verbal communication. In: Benthall, J., Polhemus, T. (Eds.), The Body as a Medium of Expression. Allen Lane, London.

Hofmeyr, I., 2004. The Portable Bunyan. A Transnational History of the Pilgrim's Progress. Princeton University Press, Princeton, NJ.

Honko, L. (Ed.), 2000. Textualization of Oral Epics. de Gruyter, Berlin.

Howes, D. (Ed.), 2005. Empire of the Senses. The Sensual Culture Reader. Berg, Oxford.

Jorholt, E., 2001. Africa's modern cinematic griots – oral tradition and West African cinema. In: Baaz, M.E., Plamberg, M. (Eds.), Same and Other: Negotiating African Identity in Cultural Production. Nordic Africa Institute, Uppsala.

Julien, E., 1992. African Novels and the Question of Orality. Indiana University Press, Bloomington, IN.

Mallery, G., 1893. Picture-Writing of the American Indians. Smithsonian Institution, Washington, DC.

Manguel, A., 1996. A History of Reading. HarperCollins, London.

McLuhan, M., 1964. Understanding Media: The Extensions of Man. Routledge and Kegan Paul, London.

Okpewho, I., 1992. African Oral Literature. Background, Character, and Continuity. Indiana University Press, Bloomington, IN.

Olson, D.R., Torrance, N. (Eds.), 1991. Literacy and Orality. Cambridge University Press, Cambridge.

Ong, W., 1982. Orality and Literacy. The Technologizing of the Word. Methuen, London.

Peek, P.M., Yankah, K. (Eds.), 2004. African Folklore. An Encyclopedia. Routledge, New York.

Perks, R., Thomson, A. (Eds.), 2006. Handbook of Oral History, second ed. [third edition in preparation]. Routledge, London.

Ricard, A., Veit-Wild, F. (Eds.), 2005. Interfaces between the Oral and the Written: Versions and Subversions in African Literature. Rodopi, Amsterdam and New York.

Schechner, R., 2002. Performance Studies. An Introduction. Routledge, London.

Schousboe, K., Larsen, M.T. (Eds.), 1989. Literacy and Society. Akademisk Forlag, Copenhagen.

Segerstråle, U., Molnár, P. (Eds.), 1997. Nonverbal Communication. Where Nature Meets Culture. Erlbaum, Mahwah, NJ.

Street, B. (Ed.), 1993. Cross-cultural Approaches to Literacy. Cambridge University Press, Cambridge.

Tedlock, D., 1977. Toward an oral poetics. New Literary History 8, 507–519.

Tokumaru, Yosihiko, Yamaguti, Osamu (Eds.), 1986. The Oral and the Literate in Music. Academia Music, Tokyo.

Van Leeuwen, T., 1999. Speech, Music, Sound. Macmillan, Basingstoke.

Wagner, D.A., Venezky, R.L., Street, B.V. (Eds.), 1999. Literacy: An International Handbook. Westview Press, Boulder, CO.

Oral Health, Psychology of

Renate Deinzer, Institute of Medical Psychology, Justus-Liebig-University Giessen, Gießen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The present article provides an overview about major aspects of the psychology of oral health. These comprise oral health-related quality of life, oral health behavior, dental anxiety, psychological stress and oral health, psychological factors in chronic facial pain, and psychological factors in denture satisfaction and adaptation. It is concluded that psychosocial factors play a major role in nearly every aspect of oral health. Still, further research is needed, especially with respect to psychosocial interventions. Additionally, psychological expertise and modeling is often missing in studies assessing psychosocial aspects of oral health. A claim is thus made for more interdisciplinary approaches including the expertise of both dentists and psychologists.

Introduction and Overview

The human face and therein the mouth is one of the most significant areas of the body for a variety of major psychological functions. It serves as a region of mutual recognition of persons and thereby also defines identity not only for the social counterpart but also for the individual itself. It is most important in social communication both with respect to the ability to communicate emotions by facial expression and with respect to the articulation of speech and other vocal utterances. The human smile is a most important part of social communication and is characterized beneath other factors by the visibility of the teeth. The teeth may form in this context a major aspect of one's attractiveness. The lips and the oral cavity are also regions of intimate social contacts. They represent the most sensitive parts of the human body provided with the highest density of sensible nerves and a disproportionally high representation in the sensory cortex exceeding that of any other body region.

According to all these factors, oral health comprises an important part of human well-being. At the same time, oral ill health and its treatment may become a psychological challenge. The present article provides an overview over these and other biopsychosocial interactions in the area of oral health and disease. In particular, it deals with the psychological impact of oral health, psychological aspects of the prevention of oral disease, psychological obstacles to adequate oral care, manifestations of psychological problems in oral health, and psychological aspects of treatment failure. It concludes with a statement concerning future research and practice of oral health care under a biopsychosocial perspective.

Psychological Impact of Oral Health: Oral Health-Related Quality of Life

The research about and assessment of health-related Quality of Life (QoL) denotes a rethinking in the consideration of health and health-related issues, which found its most prominent manifestation in the definition of health by the WHO from 1946: "Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" (WHO, 1948). According to this definition, well-being is an important outcome measure of health assessment and health-related QoL is considered an important aspect of this well-being. Health-related QoL has been found to be associated not only with somatic factors and severity of disease but also with environmental factors and factors of the person like coping, self-efficacy, anxiety, depression, and so on.

Soon after the introduction of the concept of health-related QoL, it became obvious that generic measures of QoL often miss disease-specific impairments. As a consequence, disease-specific instruments were developed, e.g., instruments to assess oral health-related QoL. The assessment usually comprises impairments with respect to oral health itself and treatment of oral disease (like gingival bleeding, facial appearance), oral functioning (like chewing, talking), environmental, social, and emotional outcomes (Sischo and Broder, 2011). They have been used to assess the impact of various oral diseases (e.g., caries) and the outcome of treatments aiming to reduce impairments due to chronic conditions like periodontal disease or tooth loss (Sischo and Broder, 2011). Research, for instance, found oral health-related QoL to be related to the extent and severity of periodontal disease and demonstrated nonsurgical treatments of periodontitis to increase oral health-related QoL while surgical treatment had a lower impact (Shanbhag et al., 2012).

Oral health-related QoL is also assessed to define treatment needs and priorities at a population level. It has been found to be reduced under conditions of dental anxiety, periodontitis, caries, HIV, and so on. Oral health-related QoL is also associated with age, ethnicity, and socioeconomic status, and social disparities with respect to oral health-related QoL have been unveiled (Sischo and Broder, 2011). In this context, it is important to note that these relationships are not always simple. There is, for instance, some research indicating that oral health-related QoL is sometimes worse in younger generations as compared to older ones despite the differences in oral care between these groups, which point into the other direction (Slade, 2012). Other studies found first-generation migrants to report better oral health-related QoL than second-generation migrants (Sischo and Broder, 2011).

Oral health-related QoL is thus a meaningful outcome measure for oral treatments and an important parameter for the

assessment of health-related impairments, which might even affect health policy decisions.

Prevention of Oral Disease: Health Psychology in Oral Care

Maintenance and rehabilitation of oral health depend to a great degree on patient behavior. The two most prevalent oral diseases, caries and periodontal disease, are considered to be avoidable in most cases if nutritional factors are considered, professional care (i.e., regular visits at the dentist) is provided and oral hygiene is maintained (Baehni, 2012). All three factors depend mainly on patients' behavior. Thereby, oral health becomes a matter of health psychology (Renz et al., 2008). Several investigations have linked oral health behavior to behavioral skills, dental knowledge, health beliefs, self-esteem, self-efficacy, reasoned action, and so on (Deinzer et al., 2005).

The majority of interventional studies in preventive oral health behavior have focused on oral hygiene. Nevertheless, research in this area is hampered by a lack of scientific evidence for both, the aims of oral hygiene behavior (i.e., what is proper oral hygiene; Baehni, 2012) and the ways how to establish it (e.g., how often teeth should be cleaned and which brushing technique would be the best; Muller-Bolla et al., 2011). If the aims of oral hygiene were perfect oral cleanliness; a recent German study demonstrated that students did not have sufficient skills to even come close to that aim: more than 60% of the tooth surfaces showed persisting plaque after thorough oral hygiene (Harnacke et al., 2012). Little research is, however, available both regarding the best oral hygiene techniques (Muller-Bolla et al., 2011) and regarding the best ways to teach these techniques. From a psychological point of view, the latter is a matter not only of health psychology but also of kinesiology as tooth brushing is a complex motoric task (Harnacke et al., 2012).

Irrespective of these yet unresolved research questions there is a large number of education programs in dentistry, which mostly address patients' knowledge about dental diseases and teach patients oral hygiene techniques. Still, the effectiveness of these programs tends to be low and short-termed, if at all (Renz et al., 2008; Baehni, 2012). Recent reviews have, therefore, demanded more theoretically driven research, incorporating insights from health psychology. Renz et al. (2008) reviewed the available literature on interventions based on psychological models aiming to improve adherence to oral hygiene instructions in periodontitis patients. They only found four papers meeting their research criteria. From these studies they conclude: "There is tentative evidence that psychological approaches to behaviour management such as the use of reinforcement, goal setting and the provision of feedback can improve oral hygiene and oral hygiene related behaviours" (Renz et al., 2008: p. 11). They conclude that more research of high methodological quality is needed, based on psychological models and theoretical frameworks. Similarly, Baehni (2012) state that empowerment of patients is needed in a sense of a comprehensive education and competence training with realistic aims and well-defined action plans. He underlines the social and psychological competences necessary for oral care professionals to meet these requirements.

Psychological Obstacles to Adequate Oral Care: Dental Fear and Phobia

Psychological competences of oral health professional are of special significance when it comes to the prevention of and dealing with dental fear and anxiety (DFA), dental phobia (DP), and dental behavior management problems (DBMP). DFA is defined as strong aversive feelings against a dental visit and/or treatment while DP is a severe type of anxiety fulfilling clinical criteria of a phobic state. DBMP comprise uncooperative and disruptive behavior during the dental treatment. The prevalence of both DFA and DBMP in children and adolescents is estimated at approximately 9% (Klingberg and Broberg, 2007; Zhou et al., 2011). It is important to realize that both conditions do not necessarily coincide. In adults, the prevalence rate for DFA is estimated at 20%, as opposed to DP with 5% (Wide Boman et al., 2013). Several instruments have been developed to assess DFA in children, adolescents, and adults (Porritt et al., 2013; Wide Boman et al., 2013).

Considering the psychological significance of the mouth and the oral cavity mentioned in the introduction, reluctance to dental treatment is a natural response to a condition where intimate barriers are violated. Dental patients are exposed to a situation, which induces unpleasant feelings in several aspects. They are positioned in a supine position in a chair with usually no way to escape. This results in a feeling of loss of control. Furthermore, the dental professionals approach one of the most intimate regions of the body. Finally, dental treatment may be associated with pain or at least uncomfortable sensations and noises either due to the treatment itself or, e.g., the injection of anesthetics. Additionally, the dental practice is an environment of outstanding salience as it differs from most environments of daily exposure. These factors, the salient environment and the uncomfortable experiences, are best suited for classical conditioning. Thereby, while DFA may be a natural response in the beginning it easily transforms into a classically conditioned response, which emerges even in the absence of the unconditioned stimuli. The dental setting is, however, not only a situation easily being associated with physical threat. It is also a place where embarrassing situations are likely to occur, if, for instance, the dentist comments on oral hygiene.

Dental anxiety often results in avoidance of dental visits and treatments and thereby in reduced dental care, which again enhances dental problems and thereby DFA (Armfield, 2013; Wide Boman et al., 2013). It is thus pivotal for maintenance of oral health to reduce DFA in its very early stages. Zhou et al. (2011) reviewed studies analyzing the effects of dental staff behavior on child dental anxiety and behavioral problems. The few studies available do not allow for firm conclusions. Still, they give some hints that empathic communication, clear-cut and specific instructions, and appropriate physical contact in combination with explanations and verbal reassurance help to reduce both DFA and DBMP. Applied on their own, explanations, reassurance, or physical contact appear to be of low efficiency.

Several groups have also analyzed the effects of treatment programs aiming to overcome DFA and DP in adults. A recent review found interventions based on behavior therapy or cognitive behavior therapy to be effective in reducing dental anxiety and increasing patients' acceptance of dental treatment though the quality of the studies was moderate to low (Wide Boman et al., 2013). Interestingly, this psychological approach appeared to be more effective than a mere physiological approach like sedation during treatment.

Manifestation of Psychological Problems in Oral Disease: Periodontitis, Bruxism, and Orofacial Pain

While dental anxiety is a psychological problem immediately related to dental treatment and its setting, other less specific psychological problems might affect oral health as well. Among them is psychological stress. Gingivitis and periodontitis are the most prevalent dental disturbances in adults. Persisting dental plaque adjacent to the gingival margin induces an inflammatory response of the gingiva called gingivitis. Plaque removal results in a restitutio ad integrum in that stage of the disease. However, complete plaque removal is hardly ever achieved by the dental patient (Harnacke et al., 2012). Gingivitis may persist over long times without any further complications. But in many cases it progresses from a localized inflammation, affecting the gingiva alone to a more widespread infection of the gingiva and the adjacent periodontium, a periodontitis. The result is an irreversible inflammatory destruction of desmodontal fibers and the alveolar bone, both supporting the tooth to remain at its place. Thereby, tooth loss is the final consequence of untreated periodontitis.

Psychological stress has been considered to be an etiological factor in the development of periodontal disease (Rosania et al., 2009). Both, animal experiments and correlative human studies show relations between psychological stress and alveolar bone loss (Rosania et al., 2009; Akcali et al., 2013) and stress has been considered to explain at least some of the socioeconomic disparities with respect to periodontal health (Thomson et al., 2012). Several mechanisms have been discussed as potential explanation for these relationships (Preeja et al., 2013), among them immune-regulation and oral health behavior, which both have indeed been found to be altered under stress by our and other groups (e.g., Weik et al., 2008; Deinzer et al., 2005; Thrombelli and Farina, 2013). Though, it is now widely accepted to consider psychological stress as a risk factor for periodontal disease, intervention studies aiming to ameliorate adverse stress effects as an adjunctive approach to periodontal treatment are missing so far.

Chronic orofacial pain is another condition often believed to depend on more general psychological functions than to be a mere representation of a physical disorder. Orofacial pain comprises four symptomatic groups: temporomandibular disorder, burning mouth syndrome, atypical facial pain, and atypical odontalgia, i.e., tooth ache. All disturbances have in common that they cannot be explained merely by physiological findings. Often, patients complain about other yet unexplained physical symptoms as well (Aggarwal et al., 2011). Major symptoms include difficulties to open the mouth and sensations as if the teeth would not fit together. Medical interventions are often invasive in that they induce irreversible alterations of the stomathognathic system, e.g., by altering the occlusion in excavating disturbing occlusal contacts. Still, these and other medical interventions appear to have little effect, if any (Aggarwal et al., 2011). Several studies have investigated psychological factors potentially involved in these conditions. Perceived pain control, self-efficacy, and somatization were found to be psychological correlates. Behavioral factors may also be important, e.g., overactivity due to stress and anxiety, which result in enhanced muscular activity and teeth clenching, which in turn induce pain. Pain may then lead to fear-avoidance behavior impairing the functionality of the whole system (Aggarwal et al., 2011). Indeed, bruxism is considered a major risk factor for temporomandibular disorder. Bruxism is a motor disorder of the stomathognathic system resulting in clenching or grinding at day and/or during sleep. It results in disturbances of the stomathognathic system including tooth wear, muscular tensions, and disturbances of the temporomandibular joints. As potential etiological or maintaining factors for this disturbance, overactivity of dopaminergic systems is currently discussed, as are psychosocial factors. With respect to the latter, it appears mandatory to differentiate between sleep-bruxism and bruxism during daytime. While the latter clearly appears to depend on psychosocial factors, results concerning the former are less definitive (Manfredini and Lobbezoo, 2009). This may be, however, also due to the fact that it is much more difficult to diagnose sleep-bruxism, which would require polysomnographic recordings. Awake-bruxism appears to depend on emotional tension (e.g., stress, anxiety) and some personality factors and behavioral styles related to emotional tension (e.g., trait anxiety, anger-in-behavior, emotional coping), and its prevalence is enhanced in some mood disorders (Manfredini and Lobbezoo, 2009). Though bruxism not always results in orofacial pain, it is considered to be a major risk factor for temporomandibular disorder. The latter has been addressed in a variety of psychosocial randomized controlled trials (RCTs). Though the majority of these studies had a moderate to high risk of bias they provide weak evidence that cognitive behavioral interventions might be helpful in the treatment of temporomandibular disorder. This is remarkable considering that mere physiological treatment strategies tend to fail in these patients (Aggarwal et al., 2011). Interestingly, RCTs on the effects of psychosocial interventions for chronic orofacial pain appear to have focused on temporomandibular disorders and neglected the other conditions belonging to this field of oral health disorders (Aggarwal et al., 2011).

Psychological Factors in Treatment Failure: Dentures and Denture Adaptation

As mentioned in the introduction, the face and the mouth form an important part of human identity. Losing one's teeth can thereby become a psychological challenge and dentures are thus provided not only to overcome functional but also psychological and social problems related to edentulism. Indeed, prevalence of edentulism remains high irrespective of improving dental care because increasing live expectancy. Endentulsim is a chronic condition resulting in a variety of

impairments and in functional loss, for instance, with respect to eating, speaking, and social interaction in general. It is accompanied by a loss of alveolar bone, reduced support of soft tissues and facial musculature, and reduced masticatory force, which results in a functionality of mastication of 1/4 to 1/7 of that of a dentate adult (Roumanas, 2009). While the majority of patients are satisfied with their dentures, fewer than 10% (Roumanas, 2009) or between 10 and 20% (Carlsson, 2009) are dissatisfied. Though there is some correlation between patients' complaints and the objective quality of the denture or anatomical characteristics of the patient, these associations are far too small to explain patients' satisfaction or dissatisfaction (Carlsson, 2009; Roumanas, 2009). Typical complaints comprise instability of the denture, pain, difficulties in mastication, and aesthetics. With respect to the last, there is often a mismatch between patients' and dentists' ideas of what would be a good aesthetic result (e.g., when dentists chose a tooth color corresponding to patients' age while patients' prefer a brighter teeth color, which the dentist would judge as unnatural) (Roumanas, 2009). In an attempt to better understand factors predicting patients' satisfaction with their denture, several studies have been conducted. These indicated that patients' previous experiences and expectations are important predictors (Roumanas, 2009). Personality has also been found to be an important predictor, and a recent longitudinal study found neuroticism to predict patients' dissatisfaction even 2 years after they had received their denture (Roumanas, 2009; Fenlon et al., 2007). Another factor considered of high significance is the dentist–patient interaction and it has been suggested by several reviews that this factor is even more important than the 'technically perfect denture construction' (Carlsson, 2009). Again, however, intervention studies aiming to facilitate patients' satisfaction by psychosocial interventions appear to be rare.

Concluding Remarks

This short overview of psychological aspects of oral health is not comprehensive in that not all relevant areas of oral health and dental treatment related to psychological factors are covered, due to space limitations. For example, special problems of orthodontic treatment of adolescents, oral cancer and its psychosocial consequences, implant failure, and stress as a predictor, could not be discussed here. Still, this overview leads to some conclusions, which hold true for these and other areas of psychodental relationships as well.

- First and foremost, the dentist–patient relationship plays a pivotal role for nearly every area of oral health. A qualitatively good and empathic relationship serves not only as an ice-breaker and makes the interaction between the dentist and patient more comfortable. It also has a significant effect on oral health outcome.
- Second, psychological factors affect nearly every aspect of oral health. Patients' personality, their learning history, previous experiences, expectations, and current stress and strain may in some areas explain even more variance of oral health conditions than physical factors.
- Third, it is for these reasons that psychosocial knowledge and competences of the dentist are considered invaluable for qualitatively high dental care.
- Finally, though there exists some research on psychodental relationships, studies of high methodological quality and with good psychological foundation are sparse. Further research providing better scientific evidence is needed. More interdisciplinary approaches including the expertise of both dentists and psychologists would be helpful to fulfill these requirements.

See also: Anxiety and Fear Conditioning, Neural Basis of; Behavior Therapy: Background, Basic Principles, and Early History; Behavioral Medicine; Chronic Pain: Models and Treatment Approaches; Doctor–Patient Interaction in the West: Psychosocial Aspects; Early Emotional Development and Cultural Variability; Emotion, Neural Basis of; Evidence-Based Medicine; Fear Conditioning: Overview; Fear: Psychological and Neural Aspects; Health Behavior, Psychosocial Theories of; Health Behaviors; Health Education and Health Promotion; Health Literacy; Health Psychology; Pain, Health Psychology of; Patient Adherence to Health Care Regimens; Personality and Health; Phobias Across the Lifespan; Psychological Treatment, Effectiveness of; Psychoneuroimmunology; Social Phobia across the Lifespan; Stress, Coping and Health; Stress: Psychological Perspectives; Subjective Wellbeing, Psychology of.

Bibliography

Aggarwal, V.R., Lovell, K., Peters, S., Javidi, H., Joughin, A., Goldthorpe, J., 2011. Psychosocial interventions for the management of chronic orofacial pain. Cochrane Database of Systematic Reviews 9 (11). Art. No.: CD008456.

Akcali, A., Huck, O., Tenenbaum, H., Davideau, L., Buduneli, N., 2013. Periodontal diseases and stress: a brief review. Journal of Oral Rehabilitation 40, 60–68.

Armfield, J.M., 2013. Predicting dental avoidance among dentally fearful Australian adults. European Journal of Oral Sciences 121, 240–246.

Baehni, P.C., 2012. Translating science into action – prevention of periodontal disease at patient level. Periodontology 2000 60, 162–172.

Carlsson, G.E., 2009. Critical review of some dogmas in prosthodontics. Journal of Prosthodontic Research 53, 3–10.

Deinzer, R., Granrath, N., Spahl, M., Linz, S., Waschul, B., Herforth, A., 2005. Stress, oral health behaviour and clinical outcome. British Journal of Health Psychology 10, 269–283.

Fenlon, M.R., Sherriff, M., Newton, J.T., 2007. The influence of personality on patients' satisfaction with existing and new complete dentures. Journal of Dentistry 35, 744–748.

Harnacke, D., Mitter, S., Lehner, M., Munzert, J., Deinzer, R., 2012. Improving oral hygiene skills by computer-based training: a randomized controlled comparison of the modified bass and the fones techniques. PLoS ONE 7 (5), e37072. http://dx.doi.org/10.1371/journal.pone.0037072.

Klingberg, G., Broberg, A.G., 2007. Dental fear/anxiety and dental behaviour management problems in children and adolescents: a review of prevalence and concomitant psychological factors. International Journal of Paediatric Dentistry 17, 391–406.

Manfredini, D., Lobbezoo, F., 2009. Role of psychosocial factors in the etiology of bruxism. Journal of Orofacial Pain 23, 153–166.

Muller-Bolla, M., Courson, F., Maniére-Ezvan, A., Viargues, P., 2011. Toothbrushing: which methods to use? Revue d'Odonto Stomatologie 40, 239–260.

Porritt, J., Buchanan, H., Hall, M., Gilchrist, F., Marshman, Z., 2013. Assessing children's dental anxiety: a systematic review of current measures. Community Dentistry and Oral Epidemiology 41, 130–142.

Preeja, C., Ambili, R., Nisha, K.J., Seba, A., Archana, V., 2013. Unveiling the role of stress in periodontal etiopathogenesis: an evidence-based review. Journal of Investigative and Clinical Dentistry 4, 78–83.

Renz, A., Ide, M., Newton, T., Robinson, P., Smith, D., 2008. Psychological interventions to improve adherence to oral hygiene instructions in adults with periodontal diseases. Cochrane Database of Systematic Reviews (2). Art Nr.CD005097.

Rosania, A.E., Low, K.G., McCormick, C.M., Rosania, D.A., 2009. Stress, depression, cortisol, and periodontal disease. Journal of Periodontology 80, 260–266.

Roumanas, E.D., 2009. The social solution-denture esthetics, phonetics and function. Journal of Prosthodontics 18, 112–115.

Shanbhag, S., Dahiya, M., Croucher, R., 2012. The impact of periodontal therapy on oral health-related quality of life in adults: a systematic review. Journal of Clinical Periodontology 39, 725–735.

Sischo, L., Broder, H.L., 2011. Oral health-related quality of life: what, why, how, and future implications. Journal of Dental Research 90, 1264–1270.

Slade, G.D., 2012. Oral health-related quality of life is important for patients, but what about populations? Community Dentistry and Oral Epidemiology 40, 39–43.

Thomson, W.M., Sheiham, A., Spencer, A.J., 2012. Sociobehavioral aspects of periodontal disease. Periodontology 2000 60, 54–63.

Thrombelli, L., Farina, R., 2013. A review of factors influencing the incidence and severity of plaque-induced gingivitis. Minerva Stomatologica 62, 207–234.

Weik, U., Herforth, A., Kolb-Bachofen, V., Deinzer, R., 2008. Acute stress induces proinflammatory signalling at sites of chronic inflammation. Psychosomatic Medicine 70, 906–912.

WHO, 1948. Preamble to the Constitution of the World Health Organization as Adopted by the International Health Conference. New York, 19–22 June 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948.

Wide Boman, U., Carlsson, V., Westin, M., Hakeberg, M., 2013. Psychological treatment of dental anxiety among adults: a systematic review. European Journal of Oral Sciences 121, 225–234.

Zhou, Y., Cameron, E., Forbes, G., Humphris, G., 2011. Systematic review of the effect of dental staff behaviour on child dental patient anxiety and behavior. Patient Education and Counseling 85, 4–13.

Oral History

Andrea Hajek, University of Glasgow, Glasgow, UK
Angela Davis, University of Warwick, Coventry, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Oral history is a versatile research method that can be used in different ways and for different purposes. In this article we examine, first, the definition and origins of oral history and its relation to memory and processes of remembering; second, we offer a theoretical background to the study of oral data and a methodological exploration of oral history research; third, we provide practical information and advice on how to prepare for and conduct an interview. The article concludes with a brief discussion of recording equipment, ethical issues and details on the archivization, and dissemination of oral data.

Introduction

Attempting to define oral history, Donald Ritchie (2003) once said that oral history "collects memories and personal commentaries of historical significance through recorded interviews". However, oral history is 'too dynamic and creative a field' to be entirely captured by a single definition (p. 19). Similarly, Bruno Bonomo (2013) identified two ways of defining oral history: on the one hand it refers to oral traditions, for example in non-Western societies and indigenous cultures where tradition is transmitted orally from one generation to another, but which are also used to share family stories; on the other hand, 'oral history' refers to a method of historical reconstruction of the past which is based on oral sources (Bonomo, 2013: pp. 13–14). These have made important contributions to our understanding of the nature and meaning of oral history accounts (Thomson, 2006: p. 51). It may thus take the form of eyewitness evidence about the past, but can also include myths, songs, and stories passed down over the years by word of mouth. Furthermore, while oral history is an exceptionable way of preserving the knowledge and understanding of older people, it can also be used to engage with younger generations. Hence it contributes to a better understanding of both the past and the present. Nor should we ignore the fact that oral history is a highly interdisciplinary research area which links to other areas of research, such as sociology, folklore studies and anthropology, although the relationship between the disciplines has varied in different parts of the world (Thomson, 2006: p. 51; Evans, 1972: p. 56). Finally, oral history interviews can be used for a variety of purposes including for academic research, publication in a radio or video documentary, in museum exhibitions or posted on the Internet (Ritchie, 2003: p. 19).

In short, oral history is a versatile research method that can be used in different ways and for different purposes. In this article we will examine the definition of oral data and the relation of oral history to memory and historiography, before moving to an historical, theoretical, and methodological exploration of oral history research.

What Is Oral History?

Historian Lynn Abrams (2010) describes oral history in terms of a 'catch-all term' which reflects, on the one hand, the process of conducting and recording interviews; on the other hand, it is the very product of that interview, and as such it is both a research methodology and the result of that research process (p. 2). Indeed, the process of interviewing cannot be disconnected from the outcome of the interview (p. 3), as oral data are not simply 'produced' during the interview: this happens in the context of a communicative event, which strongly determines what and how it is remembered (Bonomo, 2013: pp. 23, 25). Hence the oral history interview reflects a performance and an act of narration, where the interviewee constructs a sense of self through the telling of stories (Abrams, 2010: p. 35). However, oral history is not a monolog but a dialog, where data is produced through a 'discursive weaving' between two or more individuals (Bonomo, 2013: pp. 25–26), as we shall see.

A History of Oral History

Historians have, since ancient times, relied upon eyewitness accounts in the creation of historical data. During the Renaissance, however, humanist scholars began considering written documents in the reconstruction of the past (Bonomo, 2013: p. 19). With the advent – in the nineteenth-century – of Positivism and the development of a discipline of academic history, archives and documentary material became the primary sources of research, and oral evidence was marginalized (Thompson, 2000: pp. 25–81; Ritchie, 2003: pp. 19–21; Sharpless, 2006: pp. 19–20; Bonomo, 2013: p. 19). It was not until the 1940s that the modern practice of oral history developed. Ronald Grele (2006) states that the attitudes and traditions within which it did reflected the tension between those who saw oral history as an archival practice and those who envisioned it as social history, and these were geographically specific: in the United States, oral history originated in archival projects while the European introduction to oral history was through the work of social historians (p. 44). Consequently,

historians in England led the way in documenting lives of ordinary people, as opposed to Americans who tended to focus on elites.

Oral history in the modern form of audio recordings has its origins in the work of Allan Nevins at the University of Columbia (Thomson, 1998: p. 581). Nevins began to record the memories of 'persons significant in American life' in 1948. He was the first to initiate a systematic and disciplined effort to record on tape, preserve, and make available for future research recollections deemed of historical significance. While working on a biography of President Grover Cleveland, he found that Cleveland's associates left few of the kinds of personal records that biographers generally rely upon. Nevins thus came up with the idea of conducting interviews with participants in recent history to supplement the written record.

In contrast to Nevin's 'great men' approach, George Ewart Evans (1972) – the pioneer of oral history in England – collected memories of life and work in Suffolk villages, where "the old survivors were walking books" (p. 57). These were published in *Ask the Fellows Who Cut the Hay* (1956). Alistair Thomson (2006) states that oral history was an essential source for the 'history from below' approach fostered by politically committed social historians in Britain, such as Evans, from the 1960s onward (p. 52). Paul Thompson, for example, was a social historian at the University of Essex who played a leading role in the creation of the British Oral History Society in the early 1970s and the subsequent development of an international oral history movement from the end of that decade.

Within this developing international movement, oral historians approached the collection, analysis, and dissemination of oral history in different ways. Graham Smith (2010) notes that in broad terms, while oral historians in Western Europe and North America have often focused on issues of identity and cultural difference, oral historians in Latin America and Eastern Europe have tended to pursue more overtly political projects. Ritchie (2003) states that worldwide political and social changes during the last few decades of the twentieth century confronted historians with the inadequacy of archival documentation. He argues, for example, that newly emerging nations in Asia and Africa found that the written documents reflected the views of former colonial masters, and used oral history to revive buried national identities (p. 23). Indeed, decolonialization processes in the 1950s led to the development of an interest in African studies where oral history research and anthropology merged, as in the influential works of anthropologist and historian Jan Vansina (Abrams, 2010: pp. 25–26; Bonomo, 2013: p. 48). Similarly, when the Soviet Union dissolved, a wave of autobiographical works were published as Russian and European oral historians sought to reexamine and rewrite their history by collecting personal testimony suppressed under Communist regimes (Bonomo, 2013: p. 67). Last, in recent years Brazil and Argentina have also witnessed a boom in oral history projects, to record the experiences of victims of military dictatorship and state terrorism (Ritchie, 2003: p. 23; Bonomo, 2013: p. 68).

Despite these valuable contributions to the historical record, oral history has nevertheless come under criticism: the speaker's personal knowledge has been held to be less reliable than written documents. Oral history is indeed reliant on memories, usually of older people, which are criticized for being sentimental, vague, inaccurate, and biased. Yet, as Elizabeth Tonkin (1992) argues, oral history is not either intrinsically more or less likely to be accurate than a written document: the latter are often no more than orality recorded (pp. 113–114). In addition, Alistair Thomson (2006) has demonstrated that oral historians have developed their own standards to assess the reliability of oral memory, for example, by learning from social psychology and anthropology how to determine the bias and fabulation of memory, the significance of retrospection, and the effects of the interviewer upon remembering (p. 54). Furthermore, from the late 1970s imaginative oral historians such as Luisa Passerini, Alessandro Portelli, and Michael Frisch have turned these criticisms on their head and argued that the so-called unreliability of memory was also its strength, and that the subjectivity of memory provided clues not only about the meanings of historical experience, but also about the relationships between past and present, between memory and personal identity, and between individual and collective memory (Thomson, 2006: p. 54).

This 'rehabilitation' of oral history continued in the 1980s, with the end of the 'grand narratives' as theorized by Jean-François Lyotard (1984), the revival of the narrative element in history writing, and the so-called linguistic turn (Bonomo, 2013: pp. 64–65). Connected to this is the fact that memory production in the 1980s became more focused on the analysis of collective traumatic events such as the Holocaust and on the survivor's authority as a witness (Bonomo, 2013: p. 37), symbolized in Annette Wievorka's definition of *The Era of the Witness* (1999), or what Thomson (2006) has defined a "biographical era" (p. 57).

Over the past two decades, finally, oral history has largely been accepted as a research methodology: it is practiced both in academia and in the public sphere, and is represented by a variety of associations and societies across the world. It has gained a particular value in processes of community building and reconstruction. As Lynn Abrams (2010) has justly observed, oral history "is one of the few ways by which those who have traditionally been silenced in History may be heard" (p. 24). Similarly, Alistair Thomson (1999) argues that oral history gives a voice to the voiceless, such as women, workers, ethnic minorities, indigenous peoples, and marginalized groups, and enables historians to investigate underresearched areas of history such as the family and private issues. Hence it is able to provide information about everyday life and give insights into the mentalities of what are sometimes termed 'ordinary people', offering data that simply cannot be drawn from more traditional sources (p. 291). Oral histories also eloquently make the case for the active agency of individuals whose lives have been lived within deeply constraining circumstances. In other words, oral history can open up new views of the past. For in an interview the voice of the narrator literally contends with that of the historian for control of the story. Recounting the experiences of everyday life and making sense of that experience, narrators turn history inside out, demanding to be understood as purposeful actors in the past, and talk about their lives in ways that do not easily fit into preexisting categories of analysis.

Memory and Remembering

There are several difficulties attributed to the use of oral data, due to the way in which people remember. Memory can, for example, be distorted by external constraints. It is also selective, as people often recall things that correspond with the image of themselves that they are trying to present. In his interviews with Australian First World War veterans, Alistair Thomson (1994) found that his respondents remembered events in such a way as would prove acceptable accounts of their war service to their present selves. When asked about their enlistment, for example, the veterans described reasons that were not untrue but which were angled to reflect the image they wanted to present (Thomson, 1990: pp. 27–31).

Memories are also subject to self-censorship, as is particularly evident in memories of Nazi Germany. People cannot remember events without benefit of hindsight as memory is embedded in present experience. Thus Michelle Mouton and Helena Pohlandt-McCormick (1999) argue that Germans do not suffer from collective amnesia, but from a collective revision of memory (p. 43). In order to deal with the horrors of war and the guilt associated with the genocide committed during the Nazi regime, they try to create an acceptable memory. Many of the German women interviewed by Mouton and Pohlandt-McCormick initially referred to the impossibility of acting against the Nazis, as this was a sanctioned way to remember the past (pp. 43–44). However, during the course of the interviews it became clear that several of the women had supported National Socialism at the time. The women seemed to contradict themselves, but this was not to be regarded as their trying to present misleading pictures of their pasts: they felt compelled to construct an interpretation of their past acceptable to the present climate. When using oral data the complexity of people's memories must therefore be observed.

Another good example of the complexities of remembering is the death of Italian steel worker Luigi Trastulli, shot by police during an anti-North Atlantic Treaty Organization (NATO) workers' rally in 1949. In what is perhaps the most cited article in the literature of oral history, Alessandro Portelli (1991) brilliantly demonstrates why oral accounts of this incident routinely get the date, place, and reason for Trastulli's death wrong. Narrators manipulated the facts surrounding this tragic incident to render it less senseless and more comprehensible to them. Although oral historians do attempt to get the story straight through careful background research and informed questioning, they are ultimately less concerned with the vagaries of individual memories than with the larger context within which individual acts of remembering occur, or with what might be termed social memory. As Portelli argues, "errors, inventions, and myths lead us through and beyond facts to their meanings" (p. 2).

Elizabeth Tonkin's statement about people speaking in genres brings us to another characteristic of oral history. 'Doing an interview' is an oral genre, which conditions how people tell their story as they have a model of what is expected of them in their head. They therefore use a mode of storytelling they think is suitable in this context (Tonkin, 1992: pp. 50–65). People also use narrative genres as a way of structuring their accounts. In her study of British women's lives during the Second World War, Penny Summerfield (1998) found that women adopted certain roles in which to present their war service. She showed how the women she interviewed about their experiences of the war principally adopted two narrative models: the stoic and the heroic (pp. 87–99, 284–285). In other words, some remembered their experiences as war heroes or as active participants in the shaping of events, with their narratives constructed to exemplify this, while others presented themselves as passive victims of the war. Seen in the context of a range of other sources, oral history can, then, reveal the complex ways in which women and men compose their narratives in order to reconcile ideals of femininity and masculinity with the reality of their own lives (Davies, 1992: p. 55).

Finally, what respondents do not say can be as important sources of information as what they do, and even when the facts they give us are wrong, the reasons why they do so can tell us something about the society at the time they are describing, and that of the present. Luisa Passerini (1987), for example, analyzed the silences and inconsistencies in Italian working-class memories of Benito Mussolini's interwar Fascist regime, demonstrating how Fascism had infiltrated everyday life and the difficulties of remembering personal involvement in a Fascist regime. When using oral sources it is important to be aware that silences do not just mean something has been forgotten about. Françoise Zonabend (1984) studied personal memories in a Burgundian village in 1970s. She found that her interviewees did not discuss the Resistance during the Second World War with her. Their silence about the Resistance did not mean it was forgotten, though: it was regularly commemorated at elections. It shows, instead, how informants may not tell observers everything they know or think. The Resistance was a divisive issue in the village, which was not the image of unity that the villagers wanted to create (pp. 196–203).

Subjectivity, Intersubjectivity, and Composure

The fact that oral history builds on people's memories of the past raises the problem of objectivity. According to James Fentress and Chris Wickham (1992), memories are subjective by nature, although they can contain objective facts (p. 7). However, the very subjectivity of oral history is also its advantage. In the words of Portelli (1991), oral sources "tell us not just what people did, but what they wanted to do, what they believed they were doing, what they now think they did…Subjectivity is as much the business of history as the more visible 'facts'" (p. 50). We can use official records, for example, to tell us the date of a marriage or the birth of a child, but not how a bride felt on her wedding day or a mother on the birth of her child. Like any other source, oral history data simply need careful analysis. It is not the case that written documents are reliable and oral sources are not, but there are additional factors when analyzing oral data that need to be taken into account, such as the nature of memory and the relationship between interviewer and interviewee. If done so, oral data can provide an invaluable source of information to the historian.

Indeed, an important issue to bear in mind is the relationship between interviewer and interviewee. While the historian always engages in a relationship with source material, in oral sources they tell the story together. To quote Portelli (1997)

again: "Oral history…refers [to] what the source [the narrator] and the historian [the interviewer] do together at the moment of their encounter in the interview" (p. 3). The historian can influence how and what someone remembers, people can say what they think the interviewer wants to hear and are influenced by the character of the interviewer, and the reverse is also true. People are not just living documents, and there is a special relationship between interviewer and interviewee. This is not necessarily a bad thing: a great advantage of a living source is that you can ask people questions, which you cannot ask documents.

Finally, the narrator often 'performs' his or her identity, and in doing so may engage in a dialogic process where his or her consciousness engages with existing discourses or cultural frameworks: "the self is unstable, performative and draws heavily upon culture" (Abrams, 2010: p. 59). In her aforementioned research on Britain during the Second World War, Penny Summerfield (1998) provided examples of how people draw on public versions of the past as these have been transmitted through a wider, cultural memory (films, photographs, etc.) when constructing their own personal accounts. Drawing on Graham Dawson's concept of 'composure', i.e., the process of "creating accounts of experiences and achieving personal composure, or equilibrium, through constituting oneself as the subject of those stories", she demonstrates how oral history helped her uncover and 'expose' legends for the 'fabrications' that they are (p. 17).

Oral History in Practice: Preparing for the Interview

Before starting an oral history project, it is important to consider costs for traveling; recording equipment and transcription services; potential funding possibilities; and research outputs. Interviewees can be found through local newspapers, residential homes, local history centers and archives, community projects and religious and political organizations (Smith, 2010: p. 22). Snowball sampling, where each respondent gives the name of another person to participate, is a tried and tested means of identifying potential interviewees. According to Katherine Field (2001) snowballing is a particularly appropriate method for finding elderly respondents because it helps secure the trust of interviewees through being 'recommended' to them by their friends (p. 103). However, some awareness of who might be missing from recommendations to interview is useful, as it is likely that snowball sampling will reproduce preexisting networks (Smith, 2010: p. 23).

There are four principal interviewing techniques. First of all, unstructured or in-depth interviewing allows the researcher to gain an understanding of the interviewee's life history. It consists of open-ended questions which allow the interviewee to talk about his or her life in a free-form manner. The life-story approach, as Marie-Françoise Chanfrault-Duchet (1991) has argued, "makes it possible to go beyond the preconstituted discourses and 'surface assertions' collected through survey research. It highlights the complexity, the ambiguities, and even the contradictions of the relations between the subject and the ideological image of woman" (p. 89).

Second, semistructured interviewing combines organized questions with more open-ended questions. It allows for the collection of more specific information on a subject all the while giving the interviewee space to talk freely. Structured interviewing, on the other hand, makes use of a standard set of questions which are ordered in a specific manner, and is most useful when doing quantitative analysis. A focus group or witness seminar, finally, is a collective interviewing technique where a group of people is interviewed all at once. It is a highly suitable technique for gathering a large quantity of information, for example, when doing research on a particular geographical area or a city. It is also useful when investigating issues of collective memory or trauma. At the same time, though, it may pose problems when, for example, interviewees refuse to reveal certain information to one another.

Credibility and Reliability

As with all historical sources, researchers must exercise critical judgment when using interviews – just because someone says something is true does not mean it is true, however convincing they are. Similarly, being a witness to events does not mean the narrator understands the significance of the events they describe. Linda Shopes has therefore advised that the first step in assessing an interview is to consider the credibility of the interviewee and their account. The narrator's relationship to the events under discussion, his or her personal stake in presenting a particular version of events, physical and mental state at the time of the events under discussion and of the interview, as well as the internal consistency of the account all figure into the narrator's reliability. Comparing what the interviewee says both with other respondents and different types of evidence can help establish the accuracy of their account. Does it complement or supplement the other sources, or does it contradict them? If the latter, it is necessary to analyze why this difference occurred. Were interviewees differently situated in relationship to the events under discussion? Might they have different agendas, leading them to tell different versions of the same story? Could the written sources be biased or limited in a particular way? Might intervening events – for example, ideological shifts between the time of the events under discussion and the time of the interview, or subsequent popular cultural accounts of these events – have influenced later memories? Put simply, we need to ask: who is saying what, to whom, for what purpose, and under what circumstances?

The Interviewee: Who Is Talking?

Who a narrator is becomes a cognitive filter for their experiences. In Shopes' words, "What a narrator says, as well as how the narrator says it, is related to his or her social identity (or identities)". Identities are neither singular nor fixed and are defined by both the speaker's relationship to the events under discussion and their temporal distance from them. Feminist historians have shown that women often articulate their life stories around major events in the family life cycle, such as dating events in relation to when their children were born. Women also tend toward "understatement, avoidance of the first person point of view, rare mention of personal accomplishments, and disguised statements of personal power"

(Gwen Etter-Lewis cited in Sangster, 1998: p. 89). In contrast, men are more likely to connect their personal chronologies to public events like wars, elections, and strikes. Ethnicity is also influential. Scott Ellsworth (1982) used the phrase 'segregation of memory' to describe the varying ways blacks and whites remembered the 1921 race riot in Tulsa, Oklahoma (p. 98). Whites recalled either 'very little at all' about members of minority groups or that 'we all got along'. In contrast, blacks offered a less positive and more complex account (pp. 104–107).

The Interviewer: Who Is Asking?

While the interviewee is the central focus of the interview, it is the questions posed by the interviewer that shape the interview's form and content. Smith (2010) recommends a combination of broad questions (i.e., 'Tell me something about…'), precise enquiries (i.e., 'Could you please tell me your date of birth?'), and questions that bring a topic into focus by asking who, what, when questions (p. 27). The interviewer's influence reaches beyond their questioning, though. Linda Shopes explains that the social identities of both the narrator and interviewer are played out in the dynamic of the interview with narrators assessing what they should and should not say to the interviewer. For example, a grandparent being interviewed by a grandchild for a family history project may suppress less savory aspects of the past in an effort to shield the child, serve as a responsible role model, and preserve family myths. Consequently, the interviewer must be aware of what is happening in the interpersonal relationship during the interview situation (Yow, 1994: p. x).

The Interview: What Is Said?

As we have seen, oral history interviews are broad in scope and approach. Shopes reminds us that what is analytically important is the way narrators structure their accounts and the way they select and arrange the elements of what they are saying. As important as what is said is, again, what is not said, i.e., what a narrator misconstrues, ignores, or avoids. Silences can signify many things, though: it can be a simple misunderstanding, a discomfort with a difficult or taboo subject, but it may also reflect mistrust of the interviewer and a cognitive disconnection between interviewer and narrator.

The purposes and outputs of an interview also influence and shape the narrative. For example, community-based oral history projects, which seek to enhance feelings of local identity and pride, often sidestep more difficult and controversial aspects of a community's history. What is said can also be shaped by the intended outcomes of the interview. Narrators whose interviews are intended for web publication, with a potential audience of millions, are perhaps more likely to exercise a greater degree of self-censorship than if their interviews were intended to be placed in an archive.

Finally, the circumstances of the interview also affect what is recalled. Generally, interviews for which both interviewer and interviewee have prepared themselves are likely to be fuller and more detailed accounts than more spontaneous exchanges. Similarly, physical comfort and adequate time help create the expansive mood and unhurried pace that enhances recall. Shopes suggests that the location of the interview subtly influences what a narrator talks about and how they talk about it. Interviews in a person's office, for example, tend to be more formal, less intimate, with the narrator emphasizing public rather than private life. Likewise, an interview with more than one person simultaneously or the presence of a third person in the room where an interview is taking place can constrain a narrator, turning a private exchange into something more akin to a public performance.

Recording and Transcribing Interviews

A good example of how the environment shapes what is said during the interview is when video recording is used instead of audio recording. The advantage of video is that it allows the interviewer (and subsequent users of the recordings) to 'hear' more by seeing body language, facial expressions, and emotions. However, it may prove difficult to manage both interviewing and filming the interviewee, and it is advisable to have someone else handle the camera. As discussed above, the presence of a third person (and indeed the presence of the video camera itself) makes the interview a less private matter. In consequence the interviewee may feel uncomfortable and hold back information. Another side-effect of video interviews is that the narrator might become too attentive to his or her physical appearance or eloquence, to the detriment of the interview itself.

In addition to the choice between audio and video recorder, there are other decisions to be made in respect of recording equipment, with a large range of different types of equipment available in the market. For many researchers, the most convenient audio recorders will be 'solid-state' digital recorders: these usually have a Universal Serial Bus (USB) connection so that the recordings can be easily uploaded to a computer. It is advisable to buy an external microphone – for example a simple tie clip microphone – and headphone for the best sound quality. A unidirectional hand-held microphone or a rifle microphone can be a useful addition if the interview is taking place outdoors.

A good quality recording is important when making transcriptions of interviews, or, at the very least, a summary of the interview detailing the main themes that were discussed, the date, location, and duration of the interview, and some biographical information about the interviewee. A full transcription is a time-consuming matter. While transcription software programs may help, these never transcribe more than 80% of the recording, and it is advisable to hire a professional transcriptionist.

Ethics and Copyright

The interviewee is the copyright holder of the content – i.e., the words – of the interview, which cannot be used unless the interviewee grants his or her 'informed consent'. This will usually be obtained through the interviewee signing a written consent form, although it can be obtained orally on the

recording (which may be more appropriate if the interviewee has low literacy). A consent or clearance form also helps avoid issues with data protection. Whether obtained in writing or orally, the statement of consent will affirm that the interviewee consents to take part in the recording; that they permit you to use it in an agreed way or ways (which need to be specified), for example, as an online publication or to be deposited in an archive; and that they have had the opportunity to specify any restrictions they wish to make. While consent can be obtained after the interview the best time to acquire it is immediately at the end of the interview when both interviewer and interviewee are present to discuss the implications of the statement.

All oral history interviewing requires the researcher to consider a complex range of ethical issues requiring them to balance the need to respect the wishes of the informant whilst also maintaining professional integrity in writing the history which they believe does best justice to all the sources. Some of the most difficult ethics issues relate to oral history interviews involving children, detainees or medical patients, but there are legal and ethical limitations which are applicable to all oral history sources. Some interviewees may have personal issues, for example, a traumatic experience or a disabled child they may not want the researcher to write or talk about in public, or indeed to discuss at all. In such cases it is important to state in advance what kind of information is being collected and how it will be used, so as to enable potential interviewees to decide whether the research project is something they want to participate in.

Archiving and Disseminating Oral History

Oral history that is to be archived needs to comply with a number of requisites: quality and good audio; documentation (i.e., information about interviewee(s), recording dates/venues, etc.); and the presence of signed consent forms. In our current, digital times, many archives also choose to digitize interviews and publish them online. This allows archives to save on transcription costs and make the oral data more easily accessible for researchers located elsewhere. Interviews are also used ever more frequently in museum exhibitions and on museums web sites. As such they manage to enliven the exhibition, increase public support and attract new users or visitors. They also allow visitors to imagine what it was like to live in the past. More importantly, they offer a variety of perspectives on events, thus challenging dominant narratives while giving voice to marginalized groups such as immigrants or women. However, this approach to oral data poses problems, not only in terms of ethics. Making a selection from an interview may not only lead to misinterpretations and wrong impressions, but there is also the risk of error if the digitized interviews are transcribed without the original participants in the interview being able to review the transcript (Ritchie, 2003: p. 176). Finally, there is a danger involved when visitors are exposed to interpretations of ambiguous events, such as genocide and dictatorial regimes, as not all visitors may have the capacity to critically interpret certain perspectives. It is important, then, that the oral data used reflect a wide array of positions and are backed up by historical information or commentary.

Conclusion

If oral history has been growing – as a research method – for several decades now, it was only since the 1980s that it has gained repute and an established position in historical research. Although criticism has not entirely been refuted, for example with regard to the subjective nature of oral history data, nowadays oral history is widely applied both in and outside academia. Researchers intending to partake oral history research should understand the dynamic nature of oral history and its relation to processes of remembering, be aware of (inter)subjective elements involved in interviewing, engage in supplementary and comparative research of written documents, and study the various interviewing techniques and ethical challenges.

See also: Communication and Transportation, History of; Contemporary History; Cultural History; First World War, The; Intercultural Communication: An Anthropological Perspective; Second World War, The.

Bibliography

Abrams, L., 2010. Oral History Theory. Routledge, London and New York.
Bonomo, B., 2013. Voci della memoria. L'uso delle fonti orali nella ricerca storica [Voices of Memory. The Use of Oral Sources in Historical Research]. Carocci, Rome.
Chanfrault-Duchet, M.-F., 1991. Narrative structures, social models and symbolic representation in the life story. In: Berger Gluck, S., Patai, D. (Eds.), Women's Words: The Feminist Practice of Oral History. Routledge/Chapman and Hall, London and New York, pp. 77–92.
Davies, B., 1992. Women's subjectivity and feminist stories. In: Ellis, C., Flaherty, M.G. (Eds.), Investigating Subjectivity: Researching on Lived Experience. Sage Publications, Newbury Park, pp. 55–76.
Ellsworth, S., 1982. Death in a Promised Land: The Tulsa Race Riot of 1921. Louisiana State University Press, Baton Rouge.
Evans, G.E., 1972. Approaches to interviewing. Oral History 1 (4), 56–71.
Fentress, J., Wickham, C., 1992. Social Memory. Blackwell, Oxford.
Field, K., 2001. Children of the Nation? A Study of the Health and Well-being of Oxfordshire Children, 1891–1939 (Unpublished D.Phil. thesis). Oxford University, Oxford.
Grele, R.J., 2006. Oral history as evidence. In: Charlton, T.L., Myers, L.E., Sharpless, R. (Eds.), Handbook of Oral History: Theories and Applications. Altamira, Lanham, MD, pp. 43–101.
Lyotard, J.-F., 1984. The Postmodern Condition. A Report on Knowledge. Manchester University Press, Manchester.
Mouton, M., Pohlandt-McCormick, H., 1999. Boundary crossings: oral history of Nazi Germany and Apartheid South Africa: a comparative perspective. History Workshop Journal 48, 41–63.
Passerini, L., 1987. Fascism in Popular Memory: The Cultural Experience of the Turin Working Class. Cambridge University Press, Cambridge.
Portelli, A., 1991. The Death of Luigi Trastulli and Other Stories: Form and Meaning in Oral History. State University of New York Press, Albany.
Portelli, A., 1997. The Battle of Valle Giulia: Oral History and the Art of Dialogue. The University of Wisconsin Press, Madison, WI.
Ritchie, D.A., 2003. Doing Oral History. A Practical Guide, second ed. Oxford University Press, Oxford.
Sangster, J., 1998. Telling our stories: feminist debates and the use of oral history. In: Perks, R., Thomson, A. (Eds.), The Oral History Reader. Routledge, London, pp. 87–100.
Sharpless, R., 2006. The history of oral history. In: Charlton, T.L., Myers, L.E., Sharpless, R. (Eds.), Handbook of Oral History: Theories and Applications. Altamira, Lanham, MD, pp. 19–42.
Smith, G., 2010. Historical Insights: Focus on Research. Oral History. The Higher Education Academy in collaboration with the Institute of Historical Research, Coventry.
Summerfield, P., 1998. Reconstructing Women's Wartime Lives. Manchester University Press, Manchester.

Thomson, A., 1990. Anzac memories: putting popular memory theory into practice in Australia. Oral History 18 (1), 25–31.
Thomson, A., 1994. Anzac Memories: Living with the Legend. Oxford University Press, Oxford.
Thomson, A., 1998. Fifty years on: an international perspective on oral history. The Journal of American History 85 (2), 581–595.
Thomson, A., 1999. Making the most of memories: the empirical and subjective value of oral history. Transactions of the Royal Historical Society, Sixth Series 9, 291–301.
Thomson, A., 2006. Four paradigm transformations in oral history. The Oral History Review 34 (1), 49–70.
Thompson, P., 2000. The Voice of the Past: Oral History, third ed. Oxford University Press, Oxford.
Tonkin, E., 1992. Narrating Our Pasts: The Social Construction of Oral History. Cambridge University Press, Cambridge.
Yow, V.R., 1994. Recording Oral History: A Practical Guide for Social Scientists. Sage Publications, Thousand Oaks, CA.
Zonabend, Z., 1984. The Enduring Memory: Time and History in a French Village. Manchester University Press, Manchester.

Relevant Websites

http://www.canoha.ca/ – Canadian Oral History Association.
http://www.oralhistory.org/ – Oral History Association (US).
http://www.bl.uk/reshelp/findhelprestype/sound/ohist/oralhistory.html – Oral History at the British Library.
http://www2.warwick.ac.uk/fac/cross_fac/heahistory/elibrary/internal/rg_smith_oralhistory_20111015 – Oral History guide by Graham Smith.
http://ohda.matrix.msu.edu/ – Oral History in the Digital Age.
http://www2.warwick.ac.uk/fac/cross_fac/ias/current/networks/oralhistory – Oral History Network at the University of Warwick.
http://oralhistorynoticeboard.wordpress.com/ – Oral History Noticeboard.
http://www.oralhistory.org.uk/ – Oral History Society (UK).

Order Statistics

HA David, Iowa State University, Ames, IA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 16, pp. 10897–10901, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

The subject of *order statistics* deals with the properties and applications of ordered random variables and functions involving them. Let the variates $X_1, \ldots, X_n$ be arranged in ascending order and written $X_{1:n} \leq \cdots \leq X_{r:n} \leq \cdots \leq X_{n:n}$ Then $X_{r:n}$ is the rth order statistic, $X_{1:n}$ and $X_{n:n}$ are the extremes, and $X_{(n+1)/2:n}$ is the median for n odd. Important functions of order statistics include the range $W_n = X_{n:n} - X_{1:n}$, the *trimmed means* $(X_{r+1:n} + X_{r+2:n} + \cdots + X_{n-r:n})/(n-2r)$, $1 \leq r < n/2$, and the L-statistics $c_{1n}X_{1:n} + \cdots + c_{nn}X_{n:n}$. Both finite-sample and asymptotic theory are outlined. Major applications of order statistics include: (1) the estimation of location and scale parameters; (2) the analysis of data where the smallest r_1 and/or the largest r_2 observations have been 'censored'; (3) the treatment of outliers; (4) robust estimation; (5) selection procedures; (6) data compression; and (7) probability plotting. If the bivariate sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ is ordered by the Xs, then the Y-variate paired with $X_{r:n}$ is called the concomitant of the rth order statistic. Some uses of these concomitants are also noted.

The field of *order statistics* deals with the properties and applications of ordered random variables and of functions involving them. Simple examples, useful as descriptive statistics, are the sample maximum and minimum, median, range, and trimmed means. Major applications include the estimation of location and scale parameters in complete or censored samples, selection procedures, the treatment of outliers, robust estimation, simultaneous inference, data compression, and probability plotting. Although important in the theory of nonparametric statistics, the subject of order statistics is very largely parametric, that is, its results depend generally on the form of the population(s) sampled.

Introduction

Let the variates $X_1, \ldots, X_n$ be arranged in ascending order and written $X_{1:n} \leq \cdots \leq X_{r:n} \leq \cdots \leq X_{n:n}$. Then $X_{r:n}$ is the rth order statistic ($r = 1, \ldots, n$), a term also applied to the corresponding observed value, $x_{r:n}$. The *median* is $X_{(n+1)/2:n}$ for n odd and $\frac{1}{2}(X_{n/2:n} + X_{n/2+1:n})$ for n even. Helped by their simplicity, the median and the *range*, $X_{n:n} - X_{1:n}$, have long been valuable as measures of location and scale, respectively.

Consider the following illustrative data, with $i = 1, \ldots, 8$

x_i :	16	40	23	12	240	33	17	27
$x_{i:8}$:	12	16	17	23	27	33	40	240

These figures may represent the annual incomes (in multiples of \$1000) at age 30 of eight high-school dropouts. Because of the *outlier* $x_5 = 240$, the median, $\frac{1}{2}(23 + 27) = 25$, clearly gives a better picture of the group than the mean, 51. The median would in fact be unchanged as long as $x_5 \geq 27$, which shows just how *robust* the median is.

Next, suppose that the x_i are test scores out of a total of 40. If we are unable to correct the impossible score of 240, we eliminate or *censor* it. The median is then $x_{4:7} = 23$, no longer very different from the mean, $\bar{x} = 24$. The range is not robust, dropping from 228 to 28, and is therefore useful in detecting outliers, which are seldom as obvious as in the above data (see Section Outliers and Selection of the Best). For 'well-behaved' data, the range provides a simple measure of variability or scale.

In time series, $X_1, X_2, \ldots$, these descriptive statistics are particularly helpful when based on small moving subsamples $X_j, X_{j+1}, \ldots, X_{j+n-1}$ ($j = 1, 2, \ldots$). The resulting *moving* (or *running*) *medians* provide a current measure of location that, unlike the moving mean, is not too sensitive to outliers. The sequence of moving medians represents a smoothing of the data, with important applications in signal and image processing. Moving ranges are useful in quality control charts, where their sensitivity to outliers is a desirable feature. If more robust current estimators of dispersion are required, one can turn to the difference of the second largest and the second smallest in each moving sample, etc.

Again, the above x_i may represent response times to a stimulus. The *extremes*, 12 and 240, are then of particular interest (see Section Extreme Order Statistics). Of course, we would want more extensive data for which we may also wish to study the *upper* and *lower records*, here (16, 40, 240) and (16, 12) (*see* Nonparametric Statistics: Records).

Basic Distribution Theory

If $X_1, \ldots, X_n$ is a random sample from a population with cumulative distribution function (cdf) $F(x)$, then the cdf of $X_{r:n}$ is given by

$$F_{X_{r:n}}(x) = P(X_{r:n} \leq x) = P$$

(at least r of $X_1, \ldots, X_n$ are less than or equal to x)

$$= \sum_{i=r}^{n} \binom{n}{i} F^i(x)[1 - F(x)]^{n-i} \qquad [1]$$

since the term in the summand is the binomial probability that exactly i of $X_1, \ldots, X_n$ are less than or equal to x.

In handling order statistics, usually it is assumed that the underlying variate X is continuous. Note that this assumption is not needed in the derivation of eqn [1]. Also, if X can take only integral values, then $P(X_{r:n}=x)=F_{xr:n}(x)-F_{xr:n}(x-1)$. However, if X is discrete, the probability of ties among the observations is positive, which tends to complicate the distribution theory (see, e.g., David, 1981: p. 13, Arnold et al., 1992: p. 41).

If the probability density function (pdf) $f(x)$ exists, as we will assume from here on, then differentiation of eqn [1] yields

$$f_{X_{r:n}}(x) = \frac{n!}{(r-1)!(n-r)!}F^{r-1}(x)f(x)[1-F(x)]^{n-r} \qquad [2]$$

Also, the joint pdf of $X_{r:n}$ and $X_{s:n}$ $(1 \le r < s \le n)$ is, for $x \le y$ (e.g., David, 1981: p. 10; Arnold et al., 1992: p. 16)

$$f_{X_{r:n}X_{s:n}}(x,y) = \frac{n!}{(r-1)!(s-r-1)!(r-s)!} \times F^{r-1}(x)f(x)[F(y)-F(x)]^{s-r-1}f(y)[1-F(y)]^{n-s} \qquad [3]$$

Equations [2] and [3] are fundamental and permit the calculation, usually by numerical integration, of the means, variances, and covariances. Also, eqn [3] makes possible the derivation, by standard methods, of the pdf of functions of two order statistics, such as the range $W_n = X_{n:n} - X_{1:n}$. To facilitate the use of order statistics, extensive tables have been developed, especially when $F(x)=\Phi(x)$, the standard normal cdf (see Pearson and Hartley, 1970/1972; Harter and Balakrishnan, 1996, 1997).

Distribution-Free Interval Estimation

Three important types of distribution-free intervals involving order statistics may be distinguished. For simplicity we assume that $F(x)$ is strictly increasing for $0 < F(x) < 1$.

Confidence Intervals for Quantiles

The (population) quantile of order p, ξ_p, is defined by $F(\xi_p)=p$, $0<p<1$. In particular, $\xi_{1/2}$ is the median. Since, for $r<s$ $P(X_{r:n}<\xi_p<X_{s:n}) = P(X_{r:n}<\xi_p) - P(X_{s:n}<\xi_p)$, it follows from eqn [1] that the interval $(X_{r:n}, X_{s:n})$ covers ξ_p with probability

$$P_1 = \sum_{i=r}^{s-1}\binom{n}{i}p^i(1-p)^{n-i}$$

Tolerance Intervals

The focus here is on the probability, P_2, with which $(X_{r:n}, X_{s:n})$ covers a proportion γ of the underlying population. Thus $P_2 = P\{F(X_{s:n}) - F(X_{r:n}) \ge \gamma\}$ Since $F(x)$ is a uniform (0, 1) variate, U, $P_2 = P\ (U_{s:n} - U_{r:n} \ge \gamma) = 1 - I_\gamma(s-r,\ n-s\ +r+1)$, with the help of eqn [3], where $I_\gamma\ (a, b)$ is the incomplete beta function.

Prediction Intervals

Let $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ be independent variates with cdf $F(x)$. The Xs may represent current observations and the Ys future observations. What is the probability, P_3, that $Y_{t:n}$ $(t=1, \ldots, n)$ falls between $X_{r:m}$ and $X_{s:m}$ $(1 \le r < s \le m)$? It can be shown that $P_3 = P(Y_{t:n} > X_{r:m}) - P(Y_{t:n} > X_{s:m})$, where

$$P(Y_{t:n} > X_{r:n}) = \sum_{i \ge r}\binom{m}{i}\binom{n}{r+t-1-i}\Big/\binom{m+n}{r+t-1}$$

A good account from the user's point of view of all these intervals is given in Chapter 5 of Hahn and Meeker (1991).

Estimation of Parameters

An important simple example is the use of range in quality control to estimate the standard deviation (sd) σ in a normal sample. If $w_{n;\alpha}$ denotes the upper α significance point of the range in a standard normal population, then from $P(w_{n;1-a/2} < W_n/\sigma < w_{n;a/2}) = 1-\alpha$ it follows that a $(1-\alpha)$ confidence interval (CI) for σ is

$$\frac{w_n}{w_{n;\alpha/2}} < \sigma < \frac{w_n}{w_{n:1-\alpha/2}}$$

where w_n is the observed value of the range.

An unbiased point estimator of σ is $\widehat{\sigma}_W = W_n/d_{n,}$, where $d_n = E(W_n/\sigma)$ is widely tabulated. Let $\widehat{\sigma}_s$ denote the unbiased root-mean-square estimator of σ. In small samples the efficiency of $\widehat{\sigma}_W$, i.e., var $\widehat{\sigma}_s$/var $\widehat{\sigma}_W$, is quite high (0.955 for $n=5$).

More systematically, linear functions of the order statistics (L-statistics) can be used to estimate both location and scale parameters, μ and σ (not necessarily mean and sd), for distributions with pdf of the form

$$f(x;\mu,\sigma) = \frac{1}{\sigma}g\left(\frac{x-\mu}{\sigma}\right), \quad \sigma > 0 \qquad [4]$$

With $Y=(X-\mu)/\sigma$, let $\alpha_{r:n} = E\ (Y_{r:n})$ and $\beta_{rs:n} = \text{cov}\ (Y_{r:n},\ Y_{s:n})$ $(r,\ s=1,\ \ldots,\ n)$. It follows that $X_{r:n} = \mu + \sigma Y_{r:n}$, so that

$$E(X_{r:n}) = \mu + \sigma\alpha_{r:n} \qquad [5]$$

$$\text{cov}(X_{r:n}, X_{s:n}) = \sigma^2\beta_{rs:n} \qquad [6]$$

A major advance was made by Lloyd (1952), who recognized that the generalized Gauss–Markov theorem can be applied to the estimation of μ and σ, in view of eqns [5] and [6]. This yields the best linear unbiased estimators (BLUEs).

$$\mu^* = \sum_{i=1}^{n} a_{in}X_{i:n} \text{ and } \sigma^* = \sum_{i-1}^{n} b_{in}X_{i:n} \qquad [7]$$

where the a_{in} and b_{in} are functions of the $\alpha_{i:n}$ and $\beta_{ij:n}$ and can be evaluated once and for all. For tables in the normal case, see Sarhan and Greenberg (1962).

The statistics in eqn in [7] provide estimators that are unbiased and of minimum variance within the class of linear unbiased estimators. They are particularly convenient for data in which some of the smallest and/or largest observations have been censored, since maximum likelihood (ML) estimation tends to be difficult in this situation. Also, for complete and censored samples, ML estimation may produce biased estimators. However, μ^* and σ^* may not have the smallest mean

squared error, even among L-statistics. Simplified approximate estimation procedures are also possible (see, e.g., David, 1981: p. 128 or Arnold et al., 1992: p. 171).

Asymptotic Theory

Except for the distribution-free results of Section Basic Distribution Theory, the practical use of order statistics in finite samples depends heavily on the availability of tables or computer programs appropriate for the underlying population. In large samples, asymptotic theory can provide approximate results that are both simpler and more widely applicable.

It is useful to distinguish three situations as $n \to \infty$: (1) *Central order statistics* $X_{(np):n}$, where (np) denotes here the largest integer $\leq np+1$, with $0<p<1$; (2) *Extreme order statistics* $X_{r:n}$, where $r=1, \ldots, m$ or $r=n-m+1, \ldots n$, and m is a fixed positive integer; (3) *Intermediate order statistics* $X_{r:n}$ or $X_{n-r+1:n}$, where $r \to \infty$ but $r/n \to 0$.

Central Order Statistics

Corresponding to the population quantile ξ_p of Section 2.1, it is natural to call $X_{(np):n}$ a *sample quantile*. The following important result then holds (e.g., David, 1981: p. 254). Let $X_1, \ldots, X_n$ be a random sample from a population with pdf $f(x)$ and put $n_j=(np_j)$, $j=1, \ldots, k$, where k is fixed and $0<p_1<p_2<\cdots<p_k<1$. If $0<p(\xi_{pj})<\infty$, $j=1, \ldots, k$, then the asymptotic joint distribution of $\sqrt{n}(X_{n_1:n}-\xi_{p_1}), \ldots, \sqrt{n}(X_{n_k:n}-\xi_{p_k})$ is k-dimensional normal, with zero mean vector and covariance matrix $(\sigma_{jj'})$, where

$$(\sigma_{jj'}) = \frac{p_j(1-p_j)}{f\left(\xi_{p_j}\right)f\left(\xi_{p_j}\right)} j \leq j'$$

As a very special case, it follows that for the median $M_n=X_{(n+1)/2:n}$ (n odd), as $n \to \infty$, $\sqrt{n}(M_n-\mu) \to N(0, \pi\sigma^2/2)$ showing incidentally that, as an estimator of μ, M_n has asymptotic efficiency (relative to $\overline{X}$) of $2/\pi=0.637$. The same result holds for n even.

Extreme Order Statistics

The situation is more complicated when $p_1=0$ or $p_k=1$, corresponding to the *extremes* $X_{1:n}$, $X_{n:n}$ and more generally to the mth extremes $X_{m:n}$, $X_{n+1-m:n}$, where m is a fixed positive integer. For $X_{n:n}$, suitably normalized, there are three limiting distributions, the simplest of which is the Type III extreme-value distribution with cdf $\Lambda_3(x)=\exp(-e^{-x})$ $-\infty<x<\infty$. This distribution has been found very useful in describing the behavior of floods and other extreme meteorological phenomena. A convenient sufficient condition for $\Lambda_3(x)$ to be the appropriate limiting distribution is that the underlying distribution satisfy

$$\lim_{x\to\infty} \frac{d}{dx}\left[\frac{1-F(x)}{f(x)}\right] = 0 \qquad [8]$$

Then

$$\lim_{x\to\infty} P\{(X_{n:n}-a_n)nf(a_n) \leq x = \Lambda_3(x)$$

where a_n is given by $F(a_n)=1-1/n$. Distributions for which eqn [8] holds include the normal, gamma, and logistic families. For a full treatment of the extreme-value case, see Galambos (1987).

Intermediate Order Statistics

An example of an intermediate order statistic is $X_{\sqrt{n}:n}$. A suitably normalized form of $X_{r:n}$ may have either a normal or nonnormal asymptotic distribution (e.g., Arnold et al., 1992: p. 226).

Asymptotic Independence Results

A useful result is that, after appropriate normalization, upper extremes are asymptotically independent of lower extremes. Moreover, all extremes are asymptotically independent of the sample mean (e.g., David, 1981: pp. 267–270).

L-Statistics

It is clear from the above that a linear function of sample quantiles, $\sum_{j=1}^{k} c_j X_{nj:n}$, tends to normality as $n \to \infty$ (k fixed). On the other hand, the asymptotic distribution of linear functions of the extremes, such as the range, will be more complicated. An L-statistic, $\sum_{j=1}^{k} c_{in} X_{i:n}$, involving extreme, intermediate, and central order statistics, may be asymptotically normal or nonnormal, depending on the underlying distribution and on how much weight is placed on the extreme order statistics.

Applications

Probability Plotting

This very useful graphical method typically is applied to data assumed to follow a distribution depending, as in eqn [3], only on location and scale parameters μ and σ. The ordered observations $x_{i:n}(i=1, \ldots, n)$ are plotted against $\nu_i=G^{-1}(p_i)$, where G is the cdf of $Y=(X-\mu)/\sigma$ and the p_i are probability levels, such as $p_i=(i-1/2)/n$. The plotting is immediate if probability paper corresponding to G is available, as it is for, for example, the normal, lognormal, and Weibull distributions. If a straight-line fit through the points $(\nu_i, x_{i:n})$ seems unreasonable, doubt is cast on the appropriateness of G. Often this simple procedure is adequate, but it may be followed up by a more formal test of goodness-of-fit.

If a straight line fits satisfactorily, a second use may be made of the graph: σ may be estimated by the slope of the line and μ by its intercept on the vertical axis. A detailed account of probability plotting and the related *hazard plotting* is given in Nelson (1982).

Selection Differentials and Concomitants of Order Statistics

If the top k ($<n$) individuals are chosen in a sample from a population with mean μ and sd σ, then $D_{k,n}=(X_{n-k+1:n}+\cdots+X_{n:n}-k\mu)/k\sigma$ measures the superiority, in sd units, of the selected group. The mean and variance of this *selection differential* is given immediately by eqns [5] and

[6] as $E\ (D_{k,n}) = \sum \alpha_{i:n}/k$, var $(D_{k,n}) = \sum\sum \beta_{ij:n}/k^2$, where all summations run from $n-k+1$ to n.

Often one is interested in how this group performs on some related measurement Y. Thus, X may refer to a screening test and Y to a later test or X to a parent animal and Y to an offspring. The Y-variate paired with $X_{r:n}$ may be denoted by $Y_{[r:n]}$ and called the *concomitant of the rth order statistic*. Corresponding to $D_{k,n}$, the superiority of the selected group on the Y-measurement is, with obvious notation, $D_{[k,n]} = (\sum Y_{[i:n]} - k\mu_Y)/k\sigma_Y$. If Y is linearly related to X, such that $Y_i = \mu_Y + \rho\sigma_Y(X_i - \mu_x)/\sigma_x + Z_i$ where ρ is the correlation coefficient between X and Y, and Z_i is independent of X_i, with mean 0 and variance $\sigma_Y^2(1-p^2)$, then $Y_{[r:n]} = \mu_Y + \rho\sigma_Y(X_{r:n} - \mu_x)/\sigma_x + Z_{[r]}$ where $Z_{[r]}$ is the Z-variate associated with $X_{r:n}$ (e.g., David, 1981: p. 110). It follows that $EY_{r:n} = \mu_Y + \rho\sigma_Y\alpha_{r:n}$, $\mathrm{var}Y_{[r:n]} = \sigma_Y^2(p^2\beta_{rr:n} + 1 - p^2)$, $\mathrm{cov}(Y_{[r:n]}, Y_{[s:n]}) = p^2\sigma_Y^2\beta_{rs:n} r \neq s$ so that the mean and variance of $D_{[k,n]}$, the *induced selection differential*, are $ED_{[k,n]} = \rho ED_{k:n}$, var $D_{[k,n]} = \rho^2 \mathrm{var}\ D_{k:n} + (1-\rho^2)/k$. The asymptotic distribution of $D_{[k:n]}$ as well as many other properties and applications of concomitants of order statistics are treated in David and Nagaraja (1998).

Outliers and Selection of the Best

Order statistics, especially $X_{n:n}$ and $X_{1:n}$, enter naturally into many outlier tests. Thus, if $X_1, \ldots, X_n$ are supposedly a random sample from a normal distribution with unknown mean and known variance σ^2, then large values of $B_n = (X_{n:n} - \overline{X})/\sigma$ make this null hypothesis, H_0, suspect. The corresponding two-sided statistic is $B'_n = \max_{i=1,\ldots,n} \left| X_i - \overline{X}\right)/\sigma$. If σ is unknown, studentized versions of B_n or B'_n can be used. The subject, with its many ramifications, is treated extensively in Barnett and Lewis (1994), where tables needed to carry out such tests are given. If the observed value of the chosen statistic is large enough, H_0 is rejected and $X_{n:n}$ (or $X_{1:n}$) becomes a candidate for further examination or rejection.

The foregoing approach is related closely to the following multiple decision procedure, illustrated on the balanced one-way classification with n 'treatment' means $\overline{X}_i$, $i = 1, \ldots, n$, each based on m replications. To test the null hypothesis H_0 of equal population means $\mu_1, \ldots, \mu_n$ against the 'slippage' alternative $\mu_1 = \cdots = \mu_{i-1} = \mu_{i+1} = \cdots \mu_n$ and $\mu_i = \mu_1 + \Delta$, where i and $\Delta < 0$ are unknown, $n+1$ decisions may be taken, that is, D_0:H_0 holds or D_i: the ith treatment mean has slipped to the right. Under the usual normal-theory assumptions, an optimal procedure at significance level α is to choose D_M if (e.g., David, 1981: p. 226) $C_{n,\nu} = m(\overline{X}_M - \overline{X})/(\mathrm{TSS})^{1/2} > c_n, \nu : \alpha$, where M is the subscript corresponding to the largest $\overline{X}_i$, TSS is the analysis of variance total sum of squares, and $c_{n,\nu:\alpha}$, with $\nu = n(m-1)$, is given in Table 26a of Pearson and Hartley (1970/1972). If $C_{n,\upsilon} < c_{n,\upsilon:\alpha}$ the decision D_0 is taken. Slippage in either direction can be handled similarly with Table 26b.

In the above, the focus is on the detection of a single outlier or a single superior (or inferior) treatment. The handling of multiple outliers is an important part of Barnett and Lewis (1994) (*see* Statistical Analysis, Special Problems of: Outliers).

Simultaneous Inference

Order statistics play an important supporting role in simultaneous inference procedures. A well-known example is Tukey's use of the studentized range in setting simultaneous confidence intervals for the treatment means in fixed-effects balanced designs (e.g., *see* Hypothesis Tests, Multiplicity of).

Robust Estimation

Here one gives up the classical idea of using an estimator that is optimal under tightly specified conditions in favor of an estimator that retains reasonably good properties under a variety of departures from the hypothesized situation. Since the largest and smallest observations tend to be more influenced by such departures, which include the presence of outliers, many robust procedures are based on the more central order statistics. A simple example is provided by the *trimmed means*

$$T_n(j) = \sum_{i=j+1}^{n-j} \frac{X_{i:n}}{n-2j} \quad 0 < j < \frac{1}{2}n$$

which include the sample median as a special case. For a very readable treatment of robust methods see Hoaglin et al. (1983) (see also Robustness in Statistics).

Data Compression and Optimal Spacing

Given a large random sample of size n from a population with pdf, eqn [4], it is possible to estimate μ or σ or both from a fixed small number k of well chosen order statistics. The loss in efficiency of estimation may be more than compensated by the storage space gained, especially as a test of the assumed underlying distribution can still be made by probability plotting. For example, in the case of a normal population, the optimal estimator of μ is, for $k = 4$,

$$\mu^* = 0.1918\left(X_{(0.1068n):n} + X_{(0.8932n):n}\right) + 0.3082\left(X_{(0.3512nn):n} + X_{(0.6488n):n}\right)$$

where $(0.1068n)$ is the largest integer $\leq 0.1068n + 1$, etc. The estimator μ^* has asymptotic efficiency (AE) 0.920 and, since it does not involve the more extreme order statistics, is more robust than$\overline{X}$. Also, μ^* is much more efficient than the median (AE = 0.637) (see, e.g., David, 1981: Section 7.6).

Concluding Remarks

The two large multiauthored volumes (Balakrishnan and Rao, 1998a,b) cover many more topics in both theory and applications of order statistics than can even be mentioned here. Additional applications include *ranked set sampling* and especially the use of *median, order-statistic*, and other *filters* in signal and image processing.

Bibliography

Arnold, B.C., Balakrishnan, N., Nagaraja, H.N., 1992. A First Course in Order Statistics. Wiley, New York.

Avérous, J., Genest, C., Kochar, S.C., 2005. On the dependence structure of order statistics. Journal of Multivariate Analysis 94 (1), 159–171.

Balakrishnan, N., Rao, C.R. (Eds.), 1998a. Handbook of Statistics 16, Order Statistics: Theory & Methods. Elsevier, Amsterdam.
Balakrishnan, N., Rao, C.R. (Eds.), 1998b. Handbook of Statistics 17, Order Statistics: Applications. Elsevier, Amsterdam.
Barnett, V., Lewis, T., 1994. Outliers in Statistical Data, third ed. Wiley, New York.
Cramer, Erhard, Kamps, Udo, 2003. Marginal distributions of sequential and generalized order statistics. Metrika 58 (3), 293–310.
David, H.A., 1981. Order Statistics, second ed. Wiley, New York.
David, H.A., Nagaraja, H.N., 1998. Concomitants of order statistics. In: Balakrishnan, N., Rao, C.R. (Eds.), Handbook of Statistics 16, Order Statistics: Theory & Methods. Elsevier, Amsterdam, pp. 487–513.
Galambos, J., 1987. The Asymptotic Theory of Extreme Order Statistics, second ed. Krieger, Malabar, FL.
Hahn, G.J., Meeker, W.Q., 1991. Statistical Intervals: A Guide for Practitioners. Wiley, New York.
Harter, H.L., Balakrishnan, N., 1996. CRC Handbook of Tables for the Use of Order Statistics in Estimation. CRC Press, Boca Raton, FL.
Harter, H.L., Balakrishnan, N., 1997. Tables for the Use of Range and Studentized Range in Tests of Hypotheses. CRC Press, Boca Raton, FL.
Hoaglin, D.C., Mosteller, F., Tukey, J.W. (Eds.), 1983. Understanding Robust and Exploratory Data Analysis. Wiley, New York.
Jones, M.C., 2004. Families of distributions arising from distributions of order statistics. Test 13 (1), 1–43.
Kamps, U., Cramer, E., 2001. On distributions of generalized order statistics. Statistics 35 (3), 269–280.
Lloyd, E.H., 1952. Least-squares estimation of location and scale parameters using order statistics. Biometrika 38, 88–95.
Nelson, W., 1982. Applied Life Data Analysis. Wiley, New York.
Pearson, E.S., Hartley, H.O. (Eds.), 1970/1972.Biometrika Tables for Statisticians, third ed., vols. 1 and 2. Cambridge University Press, Cambridge, UK.
Sarhan, A.E., Greenberg, B.G. (Eds.), 1962. Contributions to Order Statistics. Wiley, New York.

Ordered Relational Structures

Reinhard Suck, Universitat Osnabrueck, Osnabrueck, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 16, pp. 10901–10906, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

The concepts of a relation, a mapping, and an operation are introduced. An ordered relational structure consists of a set and a few relations defined on it. The formal definition is presented and illustrated by numerous examples from social sciences and from mathematics. Homomorphisms and isomorphisms between structures are defined. These are mappings of one structure into another which, in a way, keep its structural properties, while possibly changing the meaning of its primitives. Automorphisms are mappings of the structure onto itself. They reveal most of its important properties. The group of automorphisms of a structure is described leading to the concept of homogeneity and uniqueness of a structure. The Alper/Narens theorem characterizes structures on the reals with specified homogeneity and uniqueness properties of their automorphism group. This famous theorem is discussed along with a recent generalization. The order topology and some results in connection with it are presented. Finally, the problems which arise when a structure is embedded into a Dedekind complete ordered relational structure, are briefly mentioned. This article is closely related to the article on representational measurement theory. Furthermore, the section on the completion of structures draws on Section Order Topology of the article Partial Orders.

Ordered relational structures are the working place for representational measurement theory. If one thinks of measurement theory as a kind of interface between empirically observed phenomena and abstract structures which one can work on mathematically, a common language for both is needed. In this context the objects under scrutiny are just elements of a set and the observed interrelations (at least those which are deemed relevant by some scientific question) are modeled as abstract relations between these elements. In the investigation of such structures nobody is interested in the actual meaning attached to the object and their relations. This approach is to some extent similar to the development of geometry, where one works on points, lines, planes, etc., without ever defining these concepts. The only thing provided is a set of axioms which these objects satisfy. (*see* Axiomatic Theories for a deeper description of this approach.) Furthermore, the theory must take care of the process of mapping one structure into another while keeping its essential features intact.

Many of the phenomena for which one needs measurements exhibit some kind of dominance, preference, or ordering. Therefore, most of the relational structures which are used in measurement theory carry some kind of abstract ordering relation (linear, weak, or partial orders).

In short, an 'ordered relational structure' consists of a baseset, here always denoted by A, an order relation on A, and possibly further relations on A. It is the foundation on which representational measurement theory is built. Sometimes A is the Cartesian product of other sets, sometimes it is the union of disjoint sets and some of the relations are between these sets (see Section Relations, Functions, Operations for definitions). While the investigation of one such structure may be a part of mathematics, in applications in the behavioral and social sciences the common frame consists of at least two relational structures, one to describe in a qualitative way the investigated part of reality and the other a formal, mathematical, or numerical structure that is well understood or that can be investigated by means of mathematics, logic, or computation. Such a development can yield results relevant to the empirical subject under scrutiny if the investigator succeeds in faithfully mapping the qualitative structure into the formal structure.

The requirements for this process, i.e., building useful structures and mapping them 'faithfully' into one another are discussed in the sequel and in the article Measurement, Representational Theory of.

Relations, Functions, Operations

Let A and B be two sets. The 'Cartesian product' of these sets is denoted by $A \times B$ and consists of all ordered pairs (a, b) with $a \in A$ and $b \in B$. Any subset $R \subseteq A \times B$ is called a 'binary relation' between A and B. The case $A = B$ is of particular interest; in this situation R is called a 'relation' on A. Clearly, this construction can easily be generalized to a k-ary relation, denoting a subset of the Cartesian product of k sets, i.e., $R \subseteq A_1 \times A_2 \times \ldots \times A_k$. If in this case all sets A_i equal A, then the Cartesian product is denoted as A^k and R is a k-ary relation on A. This broad concept is often used to describe parts of mathematics, mostly with further restrictions and properties on R, but it is also possible to use it in everyday situations. For example, let A be a set of persons and let $(a, b) \in R$ denote the fact that a and b have the same birthday. Or if A is the set of participants in a tournament and $(a, b) \in R$ denotes the result 'a beats b,' then R is an encoding of the outcome of this tournament.

We now classify the possible relations. Among the 'binary relations' there is the prominent class of 'order relations' (*see* Partial Orders for definitions). In this article we only consider weak and linear orders, denoted by $\precsim$. A 'weak order' satisfies by definition 'transitivity,' i.e., $a \precsim b$ and $b \precsim c$ implies $a \precsim c$, and 'completeness,' i.e., for all $a, b \in A$ holds $a \precsim b$ or $b \precsim a$. A

'linear order' is a weak order satisfying 'antisymmetry,' i.e., for all $a, b \in A$ holds, if $a \precsim b$ and $b \precsim a$ then $a = b$.

Now we are ready to define an 'ordered relational structure.' It consists of a nonempty set A, a weak order $\precsim$ on A, and possibly more relations R_i with $i \in I$ for some index set I. In the sequel an ordered relational structure is denoted by

$$\mathscr{A} = (A, \precsim, R_i, i \in I)$$

If R_i is a k_i-ary relation, then $(k_i, i \in I)$ is the *type* of the structure.

Closely related to order relations are 'equivalence relations.' Such a relation $\sim$ satisfies by definition 'reflexivity,' i.e., $a \sim a$, 'symmetry' i.e., $a \sim b$ implies $b \sim a$, and transitivity. A set on which an equivalence relation is defined can be partitioned into disjoint subsets of equivalent elements. These subsets are called 'equivalence classes.' Often one can reduce a structure by replacing the base set A with the set of its equivalence classes.

There are a few operations on relations that can be employed to generate new relations. We define them only for binary relations on a set A. Thus, in what follows $R, S \subseteq A \times A$.

$$R^*: = \{(b, a); (a, b) \in R\}$$

$$R \circ S: = \{(a, b); \exists c \in A (a, c) \in R, (c, b) \in S\}$$

$$R^2: = R \circ R$$

$$R^n: = R \circ R^{n-1}$$

Thus, if $(a, b) \in R$ means 'a beats b in a tournament' then $(a, b) \in R^*$ means 'a is beaten by b,' and aR^2b means 'a has beaten someone who in turn has beaten b.'

Mappings between two sets A, B are a particular kind of binary relation. A relation f between A and B is a 'mapping' (*or* 'function),' when, for all $a \in A$ there is exactly one $b \in B$ such that $(a, b) \in f$. In this case one writes $f(a) = b$ rather then $(a, b) \in f$. The comprehensive notation is

$$f: A \to B, \quad a \mapsto b = f(a)$$

A is called the 'domain' and B the 'range' of f. If $b = f(a)$ then b is the image of a (under f) and a is a preimage of b (there may be more than one). For a subset A' of A the set of images of all $a \in A'$ is denoted by $f(A')$. In particular $f(A) = \{b \in B;$ there is an $a \in A$ such that $f(a) = b\}$. A mapping is 'surjective' if $f(A) = B$ and 'injective' if $f(a_1) = f(a_2)$ implies $a_1 = a_2$. Thus, an injective function allows at most one preimage of each element in its domain. An injective and surjective function is called 'bijective' or a 'bijection.' Bijective functions can be 'inverted;' this concept denotes the function $f^{-1}: B \to A$ such that $f^{-1}(b) = a$ if $f(a) = b$.

Sometimes bijections $f: A \to A$ are called 'permutations.' For instance, when dealing with finite sets this terminology is quite usual. Under these bijections there is one which is particularly simple, the identity mapping, denoted by id_A. It is defined by $\mathrm{id}_A: A \to A$ and $\mathrm{id}_A(a) = a$ for all $a \in A$.

Functions can be 'concatenated,' i.e., if $f: A \to B$ and $g: f(A) \to C$ then $g \circ f: A \to C$ by setting $g \circ f(a) := g(f(a))$. Note that the order of f, g in $g \circ f$ is important – $f \circ g$ is not always defined and if it is, e.g., if $A = f(A) = C$, then in general $f \circ g \neq g \circ f$. One has $f \circ \mathrm{id}_A = f$ and $\mathrm{id}_{f(A)} \circ f = f$. If f is bijective then $f^{-1} \circ f = \mathrm{id}_A$ and $f \circ f^{-1} = \mathrm{id}_B$.

If the respective conditions on the domains and ranges are met then the $\circ$-operation of functions is 'associative,' i.e., $h \circ (g \circ f) = (h \circ g) \circ f$. Thus, the brackets can be omitted, and we simply write $h \circ g \circ f$.

Mathematical operations such as addition of real numbers are special functions. A '(binary) operation' $\bullet$ on A is a function.

$$\bullet: A \times A \to A \quad (a, b) \mapsto a \bullet b$$

Thus, it is a special kind of ternary (i.e., 3-ary) relation. It is a 'partial' operation if its domain is a proper subset of $A \times A$. Examples of structures with operations frequently occur in measurement theory under the name 'extensive' or 'positive concatenation' structures (*see* Measurement, Representational Theory of).

An operation is 'associative' if $a \bullet (b \bullet c) = (a \bullet b) \bullet c$. Sometimes the notion of 'weak associativity' is needed. If $a \backsim b$ denotes $a \precsim b$ and $b \precsim a$ then the operation $\bullet$ is weakly associative if $a \bullet (b \bullet c) \sim (a \bullet b) \bullet c$. As already mentioned, weak association might arise in a structure $(A, \precsim, \bullet)$, where $\precsim$ is a weak order.

It should be noted in passing that the relation $\sim$ defined in the previous paragraph is an equivalence relation. It is often used if one wants to replace a given weak order by a linear order (*see* Partial Orders).

Examples of Ordered Relational Structures

Weights

Consider a collection of weights and let $a \precsim b$ stand for: If a is put in one pan of a balance and b in the other, then the one containing b remains in place or deflects the pointer to b's side. Let $a \bullet b$ denote putting a and b in one pan. Let A denote the set of all weights and of all combinations that can be formed by putting two or more of the weights in one pan. It can be assumed, with reasonable accuracy that $(A, \precsim, \bullet)$ is an ordered relational structure. Furthermore, the operation $\bullet$ will be approximately weakly associative and positive, i.e., $a \prec a \bullet b$ for all $b \in A$ ($x \prec y$ is short for $x \precsim y$ and not $y \precsim x$). Still more, it will satisfy 'commutativity' $a \bullet b = b \bullet a$.

Guttman Scaling

Let A be a set of testees and B a set of items to be solved by each $a \in A$. For $a \in A$ and $b \in B$ let aRb denote that person a fails to solve item b correctly. Now assume that the relation R (which is between A and B) satisfies

$$aRb, \ a'Rb, \ \text{not } a'Rb' \quad \text{implies} \quad aRb'$$

The system (A, B, R) is not an ordered relational structure in the sense of the previous section because R, being a biorder (*see* Partial Orders), is not a weak order on one of the sets A, B. However, we can easily construct an ordered relational structure from it. Let $C := A \cup B$ and on C define: for $a, a' \in A$ $a \precsim a'$ if person a' solves at least as many items as does person a, and for $b, b' \in B$ define $b \precsim b'$ if b' is solved by at most as many

persons as b. Finally define $a \precsim b$ if aRb, and $b \precsim a$ if not aRb. Then it can be shown that $\precsim$ is a weak order on C; thus, $(C, \precsim)$ is an ordered relational structure.

Several Orders on One Set

Let S be a collection of different societies. Assume they can be weakly ordered with respect to social mobilization (relation K) and social frustration (relation L) of its members, and also with respect to economic development (relation M).

This is a case of a structure $\mathscr{A} := (A, K, L, M)$ where K, L, M are all weak orders on A. Let the weak orders satisfy the following compatibility relations: For all $a, b \in A$

$$\text{if } aKb, bLa \quad \text{then } aMb \qquad [1]$$

$$\text{if } aKb, bMa \quad \text{then } aLb \qquad [2]$$

$$\text{if } aLb, aMb \quad \text{then } aKb \qquad [3]$$

Suck (1990) has shown that such a structure can be transformed to an ordered relational structure on a Cartesian product which turns out to be a conjoint measurement structure satisfying independence (*see* Measurement Theory: Conjoint). As an application of such a structure Suck and Getta (1994) gave an axiomatization and representation of Coombs's portfolio theory in which A is a set of gambles (wins and losses with certain probabilities), K is an ordering of these gambles according to preference to play such gambles by subjects, L is an ordering by expected value, and M is an ordering by perceived riskiness of the gambles. The possibility to associate this situation with conjoint measurement proved very helpful in analyzing the structure and data.

Wille and Wille (1993) considered similar relational structures. They investigated the possibility of homomorphically mapping a structure satisfying eqns. [1–3] into the reals without the detour via conjoint measurement. They related their investigation to the 'Huntington controversy' arising from Huntington (1968) who presented 'equations' such as:

$$\frac{\text{Social mobilization}}{\text{Economic development}} = \text{Social frustration} \qquad [4]$$

This use of 'equations' was severely criticized, mainly with the argument that the variables of the equation are not measurable entities (*see* Mathematical Psychology). Formulating the structure in the above way can provide the means to solve this controversy. Either applying conjoint measurement in the sense of Suck (1990) or the results of Wille and Wille (1993) would yield a framework in which a decision on the validity of eqn. [4] is possible.

Homomorphisms, Isomorphisms

It is often useful and even necessary to consider several relational structures at the same time. In this case it is important to describe how similar these structures are, and how one can go from one structure to the next or – more intuitively – how one can transfer knowledge from one structure to another. Examples of this technique occur in measurement theory where usually one structure has as its base set (a subset of) the reals, or more generally, when there is one structure which is thoroughly investigated and all or many of its properties are well known. In such cases it is without doubt helpful to compare such a structure to the one under empirical investigation and to point out the manner in which the two structures resemble each other.

Let $\mathscr{A} = (A, R_i, i \in I)$ and $\mathscr{B} = (B, S_j, j \in J)$ be two relational structures. Then a mapping $\varphi: A \to B$ is called a 'homomorphism' between $\mathscr{A}$ and $\mathscr{B}$ if the structures are of the same type $(k_i, i \in I)$ and for all $i \in I$ and for all $(a_1, \ldots, a_k k_i) \in A^{ki}$:

$$(a_1, \ldots, a_{k_i}) \in R_i \quad \text{iff} \quad (\varphi(a_1), \ldots, \varphi(a_{k_i})) \in S_i \qquad [5]$$

$\mathscr{A}$ and $\mathscr{B}$ are said to be 'homomorphic' if a homomorphism exists between A and B, If the homomorphism is a bijection, then it is an 'isomorphism.'

Isomorphic structures are, in a sense, identical, although their base sets A and B can be quite different, but each property of $\mathscr{A}$ expressible in terms of the relations R_i corresponds to a property of $\mathscr{B}$ expressible in terms of the relations S_j.

If the structures $\mathscr{A}$ and $\mathscr{B}$ are both ordered, i.e., $\mathscr{A} = (A, \precsim; R_i, i \in I)$ and $\mathscr{B} = (B, \precsim'; S_i, i \in I)$ where $\precsim$ and $\precsim$ are weak orders, then the homomorphism is sometimes called 'a monotonic' function. It satisfies

$$a_1 \leq a_2 \quad \text{iff} \quad \varphi(a_1) \precsim' \varphi(a_2)$$

If $\mathscr{A}$ is mapped homomorphically into a structure on the reals then the relation on $\mathbb{R}$ corresponding to $\precsim$ is usually, but not necessarily, the $\leq$ relation on $\mathbb{R}$. This situation is typical of representational measurement theory (*see* Measurement, Representational Theory of).

As an example, let $\mathbb{R}^+$ denote the positive reals and let $\mathscr{A} := (\mathbb{R}^+, \leq, \cdot)$ and $\mathscr{B} := (\mathbb{R}, \leq, +)$. Then $\varphi(x) := \log x$ is an isomorphism between these numerical structures.

As another example, let A be a set of players in a tournament and let $\precsim$ denote a ranking of these players with ties allowed. Here $\precsim$ is a weak order but in general it is not linear because of the possibility of ties. If $\mathscr{B} := (\mathbb{N}, \leq)$ with $\mathbb{N}$ the natural numbers, then the mapping φ which associates his or her place in the ranking with each player, is a homomorphism but not an isomorphism.

Homomorphisms can be concatenated provided the range of the first and the domain of the second coincide. More precisely, let $\mathscr{A} = (A, \precsim, R_i, i \in I)$, $\mathscr{B} = (B, \precsim, S_i, i \in I)$, and $\mathscr{C} = (C, \precsim, T, i \in I)$ be homomorphic structures with the homomorphisms

$$\varphi: A \to B \quad \text{and} \quad \psi: B \to C$$

where φ is surjective. Then $\mathscr{A}$ and $\mathscr{C}$ are homomorphic by the homomorphism $\xi := \psi \circ \varphi$. This fact is of particular importance when the structures are all identical; this case is considered in Section The Automorphism Group of an Ordered Relational Structure.

The Automorphism Group of an Ordered Relational Structure

An isomorphism which maps its domain structure onto itself is called an 'automorphism.' Let Aut $\mathscr{A}$ denote the set of automorphisms of a given structure $\mathscr{A}$. Clearly, Aut $\mathscr{A} \neq \varnothing$

because $\mathrm{id}_A \in \mathrm{Aut}\ \mathscr{A}$, i.e., the identity map is always an automorphism. If $\alpha, \beta \in \mathrm{Aut}\ \mathscr{A}$ then $\alpha \circ \beta \in \mathrm{Aut}\ \mathscr{A}$; furthermore, $\alpha^{-1} \in \mathrm{Aut}\ \mathscr{A}$. Thus $(\mathrm{Aut}\ \mathscr{A}, \circ)$ has the properties of a 'group,' i.e., a set with an operation on it which is associative, has an identity element (id_A in this case), and a right and left inverse (α^{-1}) Therefore $(\mathrm{Aut}\ \mathscr{A}, \circ)$ is called the 'automorphism group' of $\mathscr{A}$.

Modern measurement theory exploits the properties of the automorphism group very intensively. The main ideas on how this is done are now described.

The structure $\mathscr{A}$ is called 'homogeneous' if for each $a, b \in A$ there is an $\alpha \in \mathrm{Aut}\ \mathscr{A}$ such that $\alpha(a) = b$. The group and the structure are called 'n-homogeneous' if for any two n-tuples $a_1 \prec \ldots \prec a_n$ and $b_1 \prec \ldots \prec b_n$ there is at least one automorphism α such that $\alpha(a_i) = b_i$. Similarly, they are called 'n-unique' if for any two n-tuples $a_1 \prec \ldots \prec a_n$ and $b_1 \prec \ldots \prec b_n$ there is at most one automorphism α such that $\alpha(a_i) = b_i$. n-uniqueness can be reformulated as: if two automorphisms agree on n points then they are identical.

If M is the maximal homogeneity and N the minimal uniqueness of a structure, then the 'scale type' of $\mathscr{A}$ (or of Aut $\mathscr{A}$) is denoted by the pair (M, N). Clearly, one has $M \leq N$. $N = \infty$ means that there is no minimal degree of uniqueness. A structure with $N < \infty$ is called 'finitely unique.'

One of the most important results of measurement theory in the 1980s is the following theorem developed by Narens (1981) and Alper (1985, 1987).

Theorem 1 (Alper/Narens Theorem)
Let $A = \mathbb{R}$ *with* $\mathscr{A}$ *homogeneous and finitely unique. Then it is of scale type* (1, 1), (1, 2) or (2, 2). Type (1, 1) is a ratio scale and type (2, 2) an interval scale (see Section Examples of Ordered Relational Structures of Measurement, Representational Theory of).

The importance of this theorem for measurement theory is discussed in the articles on measurement theory (*see* Measurement, Representational Theory of). Suck (2000) proved a partial generalization for ordered structures with an arbitrary base set. Essentially, his result shows that for scale type (N, N), in many, though not all cases, the Alper/Narens Theorem remains true if the base set A of the structure is different from $\mathbb{R}$.

Order Topology

If one wants to exploit the power of the continuity concept for the investigation of ordered relational structures, then one must construct a 'topology' on A which is compatible with the order relation and, if possible, with the other relations as well.

A topology of A is a collection $\mathscr{T}$ of subsets of A, called the 'open sets,' such that the following three conditions are satisfied:

(T1) $\emptyset \in \mathscr{T}$ and $A \in \mathscr{T}$.

(T2) If $\mathscr{S} \subseteq \mathscr{T}$ then $\bigcup_{T \in \mathscr{S}} T \in \mathscr{T}$.

(T3) if $\mathscr{S}$ is a finite subset of $\mathscr{T}$ then also $\bigcap_{T \in \mathscr{S}} T \in \mathscr{T}$

The structure $(A, \mathscr{T})$ is a 'topological space.' For a development of the theory of topological space see Kelley (1955) or Dugundji (1966). The most important topological space is constructed on $\mathbb{R}$ by defining as open all subsets of $\mathbb{R}$ which are arbitrary unions of open intervals. This construction can be carried out in any ordered structure and is called the 'order topology' (see below). Thus, any $\mathscr{A}$ of the kind considered in this article can be regarded as a topological space endowed with the order topology.

In topological spaces the well-known concepts of convergence, limits, and continuity can be defined, and in this way, far reaching generalizations of familiar theorems of analysis are available in the theory of ordered relational structures.

We give the topological definition of 'continuity.' Let $(A, \mathscr{T}_A)$ and $(B, \mathscr{T}_B)$ be topological spaces. A function $f: A \to B$ is 'continuous' if for all $U \in \mathscr{T}_B$ the preimage under f, i.e., the set $f^{-1}(U) := \{a \in A; f(a) \in U\}$ is in $\mathscr{T}_A$.

Next, we are going to define the 'order topology' of an ordered structure $\mathscr{A}$. We assume that A does not contain a maximal or a minimal element. We define: for all $a, b \in A$

$$(a, b) := \{x \in A; a \prec x \prec b\}$$

$$[a, b] := \{x \in A; a \precsim x \precsim b\}$$

$$(0, a) := \{x \in A; x \prec a\}$$

$$(a, \infty) := \{x \in A; a \prec x\}$$

and call these sets intervals (with respect to $\precsim$). Note that the symbols 0, ∞ do not denote elements of A.

The intervals $[a, b)$, $(a, b]$, $(0, a]$, and $[a, \infty)$ are defined analogously. (a, b), $(0, a)$, and (a, ∞) are referred to as 'open intervals.' The open intervals are used as a 'basis' for the order topology denoted by $\mathscr{T}_{\precsim}$, i.e., $\mathscr{T}_{\precsim}$ is the collection of those subsets of A which can be generated by union from an arbitrary set of open intervals.

The construction of the order topology for an ordered relational structure is useful when the other relations of $\mathscr{A}$ are in some sense compatible with $\mathscr{T}_{\precsim}$; e.g., when there is an operation among the R_i then it should be continuous.

If $\precsim$ is a linear order, $\mathscr{T}_{\precsim}$ is easily seen to separate points, i.e., for all distinct $a, b \in A$ there are open and disjoint sets U, V such that $a \in U$ and $b \in V$. In particular, consider $a \prec b$ then, if $z \in (a, b)$ exists, $(0, z)$ and (z, ∞) are such sets; if $(a, b) = \emptyset$, then $(0, b) \cap (a, \infty) = \emptyset$ and $a \in (0, b)$ and $b \in (a, \infty)$. Technically, such a space is called a 'Hausdorff space.'

A bijective and continuous map $f: A \to B$ between topological spaces $(A, \mathscr{T}_A)$ and $(B, \mathscr{T}_B)$ is a *homeomorphism* if its inverse is also continuous. Topology is mainly concerned with properties of spaces which are shared by all spaces that are homeomorphic to a given one. These are called 'invariants' of the space.

Specializing this concept on order topologies, we observe: Isomorphic linear orderings are homeomorphic with respect to the order topologies and an isomorphism is a homeomorphism. Formulating this result as a theorem one gets:

Theorem 2
Let $(A, \precsim_{\mathscr{A}})$ *and* $(B, \precsim_{\mathscr{B}})$ *be two linear orderings and* $f: A \to B$ *strictly monotonic and surjective. Then* f *is*

a homeomorphism with respect to the order topologies induced by $\precsim_{\mathscr{A}}$ *on A and* $\precsim_{\mathscr{B}}$ *on B.*

It follows from Theorem 2 that automorphisms of a structure $\mathscr{A} = (A, \leq, R_i, i \in I)$ are homeomorphisms of $(A, \mathscr{T}_{\precsim})$ onto itself.

Summarizing, we note that the order topology carries a lot of information about an ordered relational structure. Despite this fact, it has not been used very much in measurement theory until very recently. Wakker (1988) pointed out that in many cases the algebraic conditions used by Krantz et al. (1971) and others are more general than topological ones. As long as the use of topological concepts is restricted to the order topology no extra conditions need be imposed. In some modern investigations the order topology could be put to use as a fruitful tool. For instance, even the topological properties of the automorphism group can play an important role (Suck, 1998, 2000).

Finally, it should be mentioned that the power of topological methods shows up in infinite structures (although the definitions are not restricted to this case). Some approaches to measurement try to avoid such structures, especially the continuum. As an example, we refer to Sommer and Suppes (1997).

Completion of Structures

The article on partial orders introduces the 'McNeille completion' of a partial order (*see* Partial Orders). This construction is also useful for ordered structures in the sense of the present article. The property of existence of infimum and supremum is equivalent to the existence of a cut element for each 'Dedekind cut.' It is called 'Dedekind completeness' (cf Narens, 1985: p. 16).

If $\mathscr{A}$ is an ordered relational structure then the McNeille completion of $(A, \precsim)$ is 'Dedekind complete.' However, a major problem arises if one wants to extend the other relations of $\mathscr{A}$, i.e., the R_i for $i \in I$, to the complete superset of A. There may exist no extension at all, or too many – more than one is already too many; Narens (1985) and Cohen (1988) investigated this problem for so-called positive concatenation structures. The results are discussed in Luce et al. (1990) in Chap. 19.4.

In the context of the order topology Dedekind completeness is equivalent to 'connectedness' in the topological sense, i.e., there are no two disjoint open sets U, V such that $A = U \cup V$. Ovchinnikov (1996) proved the surprising result that a homogeneous structure $(A, \precsim)$ with a linear order $\precsim$ has either a connected order topology or each of its maximal connected subsets consist of one element only. This result has far reaching consequences for ordered relational structures and their automorphism groups. It is at the heart of Suck's (2000) generalization of the Alper/Narens result. Thus, if a completion is feasible then many properties of the structure follow from this theory.

See also: Axiomatic Theories; Decision and Choice: Luce's Choice Axiom; Factor Analysis and Latent Variable Models in Personality Psychology; Mathematical Models in Philosophy of Science; Measurement Theory: Conjoint; Measurement Theory: History and Philosophy; Partial Orders.

Bibliography

Alper, T.M., 1985. A note on real measurement structures of scale type (m, m+1). Journal of Mathematical Psychology 29, 73–81.

Alper, T.M., 1987. A classification of all order-preserving homeomorphism groups of the reals that satisfy finite uniqueness. Journal of Mathematical Psychology 31, 135–154.

Cameron, P.J., Nešetril, J., 2006. Homomorphism-homogeneous relational structures. Combinatorics. Probability and Computing 15, 91–103.

Cohen, M., 1988. Dedekind completion of positive concatenation structures: necessary and sufficient conditions. Journal of Mathematical Psychology 32, 64–71.

Dugundji, J., 1966. Topology. Allyn and Bacon, Boston.

Foniok, J., Nešetřil, J., Tardif, C., May 2008. Generalised dualities and maximal finite antichains in the homomorphism order of relational structures. European Journal of Combinatorics 29 (4), 881–899.

Huntington, S.P., 1968. Political Order in Changing Societies. Yale University Press, New Haven, CT.

Kelley, J.L., 1955. General Topology. Van Nostrand, New York.

Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A., 1971. Foundations of Measurement, vol. 1. Academic Press, New York.

Luce, R.D., Krantz, D.H., Suppes, P., Tversky, A., 1990. Foundations of Measurement, vol. 3. Academic Press, New York.

Narens, L., 1981. On the scales of measurement. Journal of Mathematical Psychology 24, 249–275.

Narens, L., 1985. Abstract Measurement Theory. MIT Press, Cambridge, MA.

Ovchinnikov, S., 1996. Means on ordered sets. Mathematical Social Sciences 32, 39–56.

Sommer, R., Suppes, P., 1997. Dispensing with the continuum. Journal of Mathematical Psychology 41, 3–10.

Suck, R., 1990. Conjointness as a derived property. Journal of Mathematical Psychology 34, 57–80.

Suck, R., 1998. A qualitative characterization of the exponential distribution. Journal of Mathematical Psychology 42, 418–431.

Suck, R., 2000. Scale type (N, N) and an order based topology induced on the automorphism group. Journal of Mathematical Psychology 44, 582–599.

Suck, R., Getta, A., 1994. A reaxiomatization of portfolio theory. Journal of Mathematical Psychology 38, 115–127.

Wakker, P., 1988. The algebraic versus the topological approach to additive representations. Journal Of Mathematical Psychology 32, 421–435.

Wille, R., Wille, U., 1993. On the controversy over Huntington's equations: when are such equations meaningful. Mathematical Social Sciences 26, 173–180.

Organization, Geography of

Martin Müller, University of Zurich, Zürich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The geography of organization denotes the interdisciplinary study of the spatial aspects of organizations and processes of organizing. Geography as well as organization and management studies are the major fields to contribute to this area. Three major strands of research can be distinguished. One focuses on 'organizations in space' and how organizational action transforms space and vice versa. The second one looks inside organizations and is interested in how space shapes organizational life. The third one moves away from the organization as an entity but rather centers on how processes of organizing create emergent sociomaterial orders.

The geography of organization refers to the interdisciplinary study of the spatial aspects of organizations and processes of organizing. It analyzes how space and organization interact and shape each other. 'Organization' in the singular is the term that encompasses both 'organizations' – associations of human and nonhuman elements that are meant to work toward specific goals – as well as 'organizing' – the practices of establishing these associations. The term 'institution,' by contrast, which can sometimes be found in the geographical literature as a synonym for organization, is reserved for the dominant understanding in political science that sees institutions as the formal and informal rules that structure social and political behavior.

The intersection of space and organizations has grown in importance due to the proliferation of organizations worldwide and their rising significance in all aspects of contemporary societies. At the same time, processes of globalization and attendant time-space compression have also altered the role of space for organizational activities. Rather than a discernible field *sui generis*, the geography of organization is more a shared interest in some common themes. Major contributions to studying the geography of organization have originated from human geography as well as from organization and management studies. Both fields have experienced a surging interest in space and organization, respectively, in the 2000s (in organization and management studies, for example, Clegg and Kornberger, 2006; Dale and Burrell, 2008; Hernes, 2004; in geography Müller, 2012; Philo and Parr, 2000).

For a long time, the implicit division of labor between geography on the one hand and organization and management studies on the other was such that geography focused on 'organizations in space,' treating the organization as a black box, whereas organization and management studies were interested in 'space in organizations,' however, often ignoring what was going on outside the organization. Organization and management scholars studied the role of space for activities in organizations, whereas geographers tended to be interested in the external implications of the activities of organizations for space. This division has become blurred with the embrace of poststructuralist approaches that are concerned with organizing in a broader sense than with organizations as discernible units. This article brings together the literature from the two fields, charting the major lines of flight and exploring commonalities and differences. In so doing, it also seeks to tap the potential for closer dialog for two literatures that have hitherto often followed rather separate trajectories.

Theoretical Perspectives

Three major different social theoretical perspectives have characterized work on the geography of organizations: positivism, critical realism, and poststructuralism. **Table 1** provides a schematic overview for a rough first orientation, although it should be acknowledged that the boundaries often blur.

Table 1 Major theoretical perspectives on the geography of organization

	Positivism	*Critical realism*	*Poststructuralism*
Approaches	Spatial science Functionalism Rational choice	Institutionalism Structuration Critical theory	Discourse theory Practice theory Actor-network theory Embodiment
Organization	Discrete, bounded entity	Embedded entity	Temporarily fixed, constructed entity
Ambition	Mapping the organization	Contextualizing the organization	Deconstructing organization
Key question	What are the spatial patterns of organizational relationships?	How is the organization related to its socioeconomic context?	How is organization as a precarious order maintained and subverted across space?
Methods	Descriptive and inferential statistics GIS Survey research	Structured and open interviews Participant observation Descriptive statistics	Participant observation Open interviews Text analysis Visual analysis

Adapted from Del Casino Jr., V., Grimes, A.J., Hanna, S.P., Jones III, J.P., 2000. Methodological frameworks for the geography of organizations. Geoforum 31, 523–538.

The positivist understanding sees organizations as bounded, discrete units with a fixed membership and a defined set of rules. Going back to Weber (1922), it stresses the organization as an ordered and stable entity with predictable routines and rules. It is most interested in the spatial patterns of organizational relationships, either in the organization itself or with its environment, and is associated with approaches such as spatial science and rational choice. This perspective can be descriptive, mapping spatial patterns, but it can also be explanatory, trying to predict spatial patterns through organizational variables or vice versa. While dominant until the 1970s, not least fueled by the quantitative revolution of the 1960s, the positivist perspective has lost ground since then with turns toward social and critical theory in both geography and organization and management studies.

The critical realist perspective dissolves the assumption of the boundedness of organizations in favor of viewing them as linked up with wider social, economic, and political structures. The actions of organizations thus cannot be studied without taking into account the relations with their surrounding. This understanding owes much to Granovetter's (1985) notion of embeddedness, but also to Giddens' (1984) structuration theory that privileges neither structural nor agential explanations of organizational action. Veering away from the quantitative description and explanation of spatial patterns, the critical realist perspective increasingly employs qualitative methods such as semistructured interviews in addition to quantitative methods. It has been particularly useful for conceptualizing the relationship between the organization as an actor and space as the variable context in which it is embedded and enjoys particular popularity in economic geography.

The poststructuralist understanding moves from a focus on 'the organization' as entity to 'organization' as a temporarily fixed but in general ephemeral effect of the process of organizing. It is thus more interested in the becoming than in the being of organization and in the processes that bring about organization in the first place. It was Weick (1979) who pioneered this shift in attention in the studies of organization with his focus on the 'enactment' of organization. This perspective sees organization as unbounded in time and space: it is an emergent phenomenon, multiple, and amorphous. Deconstructing how this precarious order is temporarily stabilized or subverted is the central ambition of such a poststructuralist perspective that draws on a range of thinkers from Foucault to Deleuze, Latour, and Lefebvre. It has been the driving force behind the spatial turn in organization and management studies since the 2000s and has also enjoyed increasing popularity in geography with the reception of more-than-representational forms of theorizing and the concomitant focus on embodiment. However, it also encompasses variants of discourse analysis that concentrate on the meaning-making in and of organizations and thus on the semiotic side of social life.

Types of Organizations

The types of organizations that are studied in the geography of organization vary greatly. Organization and management studies as well as economic geography are most interested in firms as pivotal actors in the modern economy. In organization and management studies, this empirical focus is related to the common affiliation of the field with business schools. Transnational Corporations (TNCs) have enjoyed particular attention in economic geography, because their far-flung, intricate global networks make them ideal cases to study the geography of organization. However, empirical interest has branched out in recent years to other organizations that partake in economic activity, such as trade unions, not-for-profit organizations or public administration, and service organizations.

Political geography devotes more attention to organizations involved in politics, whether that is institutionalized big P Politics or the small p politics of social agonism and contestation. While the state and its multiple organizational manifestations have been at the center of interest for a long time (see State, Geography of), other organizations such as international and transnational organizations, NGOs, grassroots movements, schools, or the church have broadened the array of actors that are studied. In social and cultural as well as in urban geography, the range of organizations encompasses the various aspects of public administration and service as well as social movements, neighborhood and community organizations, and NGOs. Here, some organizations also figure under the moniker of 'institutions,' in particular hospitals, asylums, schools, and so on.

Organizations in Space

Research on organizations in space asks how organizational action is mediated through and across space as well as how it transforms space and vice versa. This research area has attracted increasing attention not least due to the growing need of organizational processes to be coordinated across space in the wake of globalization, but also because of the major influence organizations have on spatial patterns of the economy. It is no surprise then that the role of organizations in space is a key ingredient in many geographical theories, in particular in economic geography. Storper's (1997) 'holy trinity' of technologies, organizations, and territories equates organizations with firms and production systems. He follows Coase in seeing the major rationale of organizations in reducing transaction costs. The main goal of organizations then is to internalize and coordinate action in the economy. To this end, organizations establish a system of rules to adhere to, thus creating conventions and relations over and above mere traded transactions. Bathelt and Glückler (2003), too, have organizations as one of their four ions of a relational economic geography, alongside interaction, innovation, and evolution. For them, organizations exist to coordinate economic inputs across an increasingly differentiated spatial division of economic activity in the world, while at the same time being embedded in local social, cultural, and institutional structures. It is this tension between the recursive relationship of organization and space that characterizes work on organizations in space.

Organizations Forming Space

One strand of research seeks to explain spatial patterns, whether economic, social, or political, through the study of

organizations and organizational structures. Organizations here figure as an alternative *explanans* to complement those of individual choice and capitalist market forces. Among others, scholars have discussed the influence of government organizations on such things as the patterns of agricultural activity, migration, political boundaries, natural resource exploitation, or industrial location (e.g., Flowerdew, 1982; Scott, 1998). A groundbreaking study of the way that organizations impact on geography is Selznick's (1949) *TVA and the Grass Roots*. It is an in-depth analysis of the organizational rationalities of the Tennessee Valley Authority (TVA), a government agency that was set up in 1933 to facilitate resource development and regional planning in that area and today is the largest regional planning association in the United States. Selznick considers how the TVA coopted local residents and organizations into its land use policy, which were to fundamentally transform the spatial pattern of the Tennessee Valley with the construction of multiple dams and navigation channels and the introduction of new agricultural and forest policies.

In economic geography, the geography of enterprises, or corporate geography, sought to relate the spatial patterns of economic activity to decision making in organizations. Corporate strategies, such as cost leadership or differentiation, or vertical and horizontal integration versus disintegration have an effect on the location and clustering of production activities and thus on regional economies. The large corporation with a centralized, hierarchical structure became the archetypal organization to be studied under this rubric. Yet, corporate geography came under fire for lacking proper theoretical foundation, having no additional explanatory power over the forces of market capitalism, and being too preoccupied with the single large corporation (Walker, 1989). As a response, scholarship moved to embrace the concept of the network, which recognized the increasing differentiation of organizational forms and the fraying boundaries of organizations (Dicken and Thrift, 1992). Enjoying considerable prominence in economic geography, a set of approaches has sought to conceptualize the spatially dispersed nature of the production of commodities and services in the global economy. Global production networks and global value chains seek to capture the flows that bind headquarters and subsidiaries or customers and suppliers together for the sourcing and distribution of products (Henderson et al., 2002). Yet again, however, the prominent concept of the lead firm, which coordinates and controls operations across several countries, focuses on one large corporation that has the power to manipulate global space.

Space forming Organizations

The relation between *explanans* and *explanandum* is inverted in the second strand of research which considers not the effects of organizations on space but that of space on organizations. It acknowledges that organizations do not exist as quasi-independent entities that are isolated from the space surrounding them (Dicken and Thrift, 1992). The central idea is that of the embeddedness of organizations, i.e., the influence of institutions outside the organization on organizational action (Hess, 2004). Here, the term institutional thickness has been used to describe an integrated web of institutions such as enterprise support systems, political institutions, and social citizenship that create a shared sense of purpose and synergies of interaction between organizations.

In what has become known as the 'new regionalism' the institutional thickness of regions as well as the close proximity of organizations to each other is thought to grant a competitive edge to firms located there. This assumption is captured in a range of concepts such as the cluster, learning region, creative milieu, or industrial district, which are assumed to enhance interorganizational trust and boost learning and innovation in organizations located there. This spatial proximity between organizations is considered of particular importance for the exchange of tacit knowledge, which is at the heart of many innovation processes. Saxenian's (1996) comparison of the Silicon Valley with the Boston Route 128 cluster, for example, attributes the comparative success of Silicon Valley precisely to the presence of advantageous regional effects on organizational action. Geographers have celebrated this new regionalism as signifying the continued importance of space for organizational success, despite predictions of the death of distance in the age of globalization.

Producing Spatial Representations

A third sense in which organizations are involved in space relates to the production of representations of space. This strand of research starts from the assumption that organizations created shared symbolic meanings of spaces. Studies in this vein tend to operate with the concept of discourse or other semiotic approaches and are concerned with how language attaches differential meaning to space, which, in turn, might influence social action. Some pinpoint single organizations, such as government agencies, the EU, or the media, while others examine the representations produced by a multitude of different organizations but with respect to the same spatial category, such as 'Eastern Europe' or 'Africa.' An illustrative example of this category is the analysis of organizational discourses in the wake of 9/11 (Dahlman and Brunn, 2003). The study makes the case for recognizing NGOs as "important agents in the production of ideational content and the circulation of contemporary global visions and political (re)orderings" (p. 256). On the basis of press releases and public statements immediately after 9/11 from a broad variety of 23 organizations, ranging from the WWF to the World Bank, the IOC, and the World Council of Churches, the authors perform a discourse analysis of the spatial framing of the terrorist events.

Space in Organizations

For a long time, geography has been interested in the effects and rules that organizations produce, which is reflected through the frequent use of the term 'institution.' The mechanisms and processes of rule production within organizations, however, have tended to escape scrutiny. This predicament of treating the organization as a black box has drawn increasing critique in different geographical subdisciplines, provoking calls to examine and theorize what is going on inside organizations (Müller, 2012; Taylor and Asheim, 2001). This kind of

research has been the staple of organization and management studies. Space, however, did not explicitly appear on its agenda until the 2000s. When it made its entry, it happened with a flourish, prompting talk about a 'spatial turn' in the discipline and even attempts at rethinking organization through space, since "any act of organizing is about creating a space for human action and interaction" (Hernes, 2004: 11).

Ordering Space

Much of the concern with space in organizations has revolved around how the ordering of space can become a source of power. Work in this tradition often connects to critical management studies in highlighting how management can produce unequal and exploitative relations. Foucault and Lefebvre act as the major sources of theoretical inspiration. For Foucault, "discipline proceeds from the distribution of individuals in space … [it] organizes an analytical space" (Foucault, 1979[1975]: 141, 145), a dictum that he exemplified in his work on the prison as a panopticon and on schools as forming docile bodies. What has attracted most attention in Lefebvre's work is the dialectic of the spatial triad of conceived space (*le conçu*), perceived space (*le perçu*), and lived space (*le vécu*) (Lefebvre, 1991[1974]). Conceived space corresponds more or less to Foucault's idea of space: the spatial order that relations of production impose, expressed in plans, calculations, and models. Perceived space refers to spatial practices and the everyday appropriation of space through routines and rituals that reproduce space. Lived space, finally, represents the dynamic element, in which opportunities for difference and subversion arise that overflow the rigid spatial models of conceived space.

At the center of attention is the question of how space shapes organizational life and vice versa. The stress falls on the material aspects of space, or physical space, as it manifests itself in office layouts or the architecture of buildings (Dale and Burrell, 2008; van Marrewijk and Yanow, 2010). This (re-) discovery of material space in organization and management studies is often phrased as engaging in a dialog with the discipline of architecture rather than that of geography (Clegg and Kornberger, 2006; Kornberger and Clegg, 2004). This might be because geography, until recently, has rather sought to overcome the notion of physical space, emphasizing the constructedness and malleability of space. Claims about how space shapes organizational functions stir unpleasant memories of geographical determinism with geographers.

This strand of research examines how the material spaces in an organization are often enlisted to increase control over staff, enhance productivity, or save costs under the disguise of a rhetoric of making office space more livable and worker-friendly. An example is the enchantment of modern management with open plan offices, which have done away with walls in office buildings. The intention of open plan offices is to buttress information exchange and communication and increase the identification with the organization. At the same time, it also facilitates surveillance of staff through placing people and screens in plain view of the panoptic gaze. The original intentions of this conceived space – to follow Lefebvre's terminology – are often subverted, however, in the everyday practices of the social appropriation of it. Workers sometimes improvise makeshift partitions between workplaces with files and books or use earplugs to escape the constant background noise of the office. The continuous circulation of people, fueled through policies of hot-desking, often contributes to social withdrawal rather than to the desired social integration and exchange. In this sense, conceived space does not determine organizational life, but there is a recursive relationship in which its social appropriation leads to initially unintended effects.

Folding Space

While the research on ordering space makes much of distances and spatial partitioning, research on folding space takes the opposite tack: it is interested in how large physical distances are reduced to become small. Such an approach owes much to actor-network theory (ANT) and Serres and Latour's (1995) notion of topological space, where distance is a function of the length of a relation between two elements and as such not pre-given. Latour and Serres use this metaphor of 'folding space' to liken space to a handkerchief which can be folded and crumpled so that two points that were distant in the beginning end up close together. For example, I can be further away from the colleague who sits in the next cubicle but works for a different department than from a collaborator who works thousands of kilometers away but who is on my Skype list. The possibilities of folding space in organizations have increased with new channels of communication such as e-mail and social media, which allow people to feel connected and develop a form of sociality across vast physical distances, as well as with cheaper long-distance travel (see Globalization, Geographical Aspects). We thus increasingly find teams that are not co-located but rather work on the same project from different locations across the world.

The notion of folding space is of particular relevance for organizational arrangements where elements from different locations need to come together to produce organizational action. This empirical phenomenon has been captured in concepts such as that of the 'community of practice' and others that revolve around a distanciated sociology of learning (Amin and Cohendet, 2004: Chapter 5). In communities of practice, knowledge is created and passed on in relationships between people, and learning occurs through repeated interaction. "What allows members [of communities of practice] to share knowledge is not the choice of a specific form of communication … but the existence of a shared practice – a common set of situations, problems and perspectives" (Wenger et al., 2002: 25). Communities of practice therefore do not have to consist of localized individuals, but can be made up of a network of individuals dispersed around the globe that are nevertheless tied together through common relations. This assumption of knowledge relations stretched over long distances complements but does not supplant that of the territorial embeddedness of learning and innovation, which stresses the importance of spatial proximity.

Organizations have an active part to play in this folding of space, for the stretching and maintenance of relations over long distances requires much effort. Organizations provide the informational infrastructure, be it e-mail, wikis, intranets, project management software, or video conferencing

equipment, but also facilitate the moving of human bodies between places through business travel. Forms and documents of all manners act as intermediaries between different locations, since they can be moved easily, and help align organizational action. But organizations also create a shared practice through orienting staff toward a common goal, establishing formal and informal codes of conduct and social practices – often referred to as organizational culture. Examining how those relations across distance are established and maintained, what kind of organizational action they result in, and where they break down is the key interest of research in this mold. For that purpose, researchers often adopt a research strategy of 'following the thing,' reconstructing the networks people and things weave as they move from place to place and enroll new elements for organizational action.

Organizing Space

Emerging with a turn toward poststructuralist theorizing, the *process* of organizing has attracted increasing attention. This has implied a move away from the organization as a self-enclosed entity and distinctive focus of research, broadening the empirical scope of interest and also the possibilities for convergence between geography and organization and management studies. The central question here is how organization as a precarious order is maintained and subverted across space. 'Organizing' thus is often used as a synonym for 'ordering,' i.e., arranging things and people for a certain amount of time. This shift foregrounds the spontaneous, ad hoc character of organization and how it is actualized in social practices. For Thrift and Olds (1996: 319–320), four implications result from this shift. First, organization is improvised and tentative and not as goal-oriented and purposive as much of the literature has tended to assume. Organizational action has a haphazard character and affords a rather partial view of the world. Second, organization is constituted in action and talked into being and thus needs continued maintenance, rather than existing as an unproblematic a priori. Third, organization is always open and has fuzzy boundaries, bringing in new elements, while discarding others at the same time. Fourth, certain interactional structures are linked to organizational forms that cannot be adequately described with the existing notion of the organization. This applies, for example, to the precarious and dispersed nature of projects and social movements.

For the theoretical underpinning of this understanding of organizing, research has turned to inspiration from a heterogeneous range of authors such as Derrida, Deleuze, and Guattari, actor-network theorists such as Latour, and phenomenologists such as Spinoza. Three ideas can be highlighted as central. One is that of the *performativity* of organizing, whereby organization is enacted through repetitive practice. This calls for a shift of attention to the mundane practices through which organization is brought into being at the level of the individual, which is in tune with the budding interest in theories of practice. It reverses the idea that organizations give rise to certain actions, claiming that instead actions constitute organization. This assumption is encapsulated in Kafka's often-cited fragmentary novel *Der Prozeß*, where the very actions of the protagonist K. give birth to the trial and the court at which it is pursued in the first place.

The second idea is that of taking *materiality* seriously, both in the sense of giving more weight to things and to the corporeality of the human body and its senses. The claim that 'things matter' recognizes that organization is a sociomaterial accomplishment, in which things – whether mundane such as partition walls or complex such as software – often provide the cohesive glue to make organizational arrangements durable at least for some time. For Latour, for example, technology is society made durable and in his concept of agencements (assemblages) Deleuze equally insists that the combination of social *and* material elements creates action. This insight breaks with the tradition of seeing organizations as associations of individuals and skirting over the role of the material world, which often served as a mere backdrop. Grappling with the specific materiality of the human body involves treating it as a living organism with all the senses and forces that work through it. Giving due attention to affects such as fear or desire calls the concept of the rational, calculating individual into question, but at the same time opens up new perspectives on what makes organization durable across space or fall apart. However, it also throws up challenges about the research methods with which to approach such phenomena as affect, which often seem to be too elusive to be captured in interviews, for example.

Multiplicity, a notion at the heart of Deleuze's thought, is the third central idea of research on the processes of organizing space and refers to the precarious, contingent nature of organization, in which arrangements could always be otherwise. It thus has an avowedly political component in that it affirms the openness of the current state of affairs to political contestation, since there is no natural, generative order of things. A concern with organizing thus goes lockstep with an interest in how orders are always temporary and subverted, how power is necessarily limited and much escapes its structuring effect and how creativity is at work in unexpected places. But it also implicates an ethical call for the researcher to abandon a position of purportedly detached observation to engage in what has been called ontological politics, i.e., to think about the worlds we want to help make through our research.

Such efforts to push the boundaries of research on the geography of organization have gelled into studies that sweep across a broad range of empirical territory. They have encompassed a reliving of the riveting experience of slow-motion video as a corporeal sensitization for the minutiae of affect in everyday life (Beyes and Steyaert, 2012) or the development of new sensitivities for the experiential spaces of a voluntary welfare agency, with their fleeting and affective encounters (Conradson, 2003). The established boundaries between geographers and scholars of organization and management studies have increasingly become erased and transgressed in the course of this and there is increasing recognition of each other's work. Yet much remains to be done in the years to come to bring the conceptual promises and ambitions of this research direction to full empirical fruition.

See also: Administration in Organizations; Networks, Geography of; Organization: Overview; Organizations, Sociology of; People in Organizations.

Bibliography

Amin, A., Cohendet, P., 2004. Architectures of Knowledge: Firms, Capabilities, and Communities. Oxford University Press, Oxford.

Bathelt, H., Glückler, J., 2003. Toward a relational economic geography. Journal of Economic Geography 3, 117–144.

Beyes, T., Steyaert, C., 2012. Spacing organization: non-representational theory and performing organizational space. Organization 19, 45–61.

Clegg, S.R., Kornberger, M., 2006. Space, Organizations and Management Theory. Liber, Malmö.

Conradson, D., 2003. Doing organisational space: practices of voluntary welfare in the city. Environment and Planning A 35, 1975–1992.

Del Casino Jr., V., Grimes, A.J., Hanna, S.P., Jones III, J.P., 2000. Methodological frameworks for the geography of organizations. Geoforum 31, 523–538.

Dahlman, C.T., Brunn, S.D., 2003. Reading geopolitics beyond the state: organisational discourse in response to 11 September. Geopolitics 8, 253–280.

Dale, K., Burrell, G., 2008. The Spaces of Organisation and the Organisation of Space: Power, Identity and Materiality at Work. Palgrave Macmillan, Basingstoke.

Dicken, P., Thrift, N., 1992. The organization of production and the production of organization: why business enterprises matter in the study of geographical industrialization. Transactions of the Institute of British Geographers 17, 279–291.

Flowerdew, R. (Ed.), 1982. Institutions and Geographical Patterns. Croom Helm, London.

Foucault, M., 1979[1975]. Discipline and Punish: The Birth of the Prison. Vintage, New York.

Giddens, A., 1984. The Constitution of Society: Outline of the Theory of Structuration. University of California Press, Berkeley and Los Angeles.

Granovetter, M., 1985. Economic action and economic structure: the problem of embeddedness. American Journal of Sociology 91, 481–510.

Hess, Martin, 2004. 'Spatial' relationships? Towards a reconceptualization of embeddedness. Progress in Human Geography 28, 165–186.

Henderson, J., Dicken, P., Hess, M., Coe, N., Yeung, H.W.-C., 2002. Global production networks and the analysis of economic development. Review of International Political Economy 9, 436–464.

Hernes, T., 2004. The Spatial Construction of Organization. John Benjamins, Amsterdam.

Kornberger, M., Clegg, S.R., 2004. Bringing space back in: organizing the generative building. Organization Studies 25, 1095–1114.

Lefebvre, H., 1991[1974]. The Production of Space. Blackwell, Oxford.

Müller, M., 2012. Opening the black box of the organization: socio-material practices of geopolitical ordering. Political Geography 31, 379–388.

van Marrewijk, A., Yanow, D. (Eds.), 2010. Organizational Spaces: Rematerializing the Workaday World. Elgar, Cheltenham.

Philo, C., Parr, H., 2000. Special issue: institutional geographies. Geoforum 30, 513–599.

Saxenian, A., 1996. Regional advantage: culture and competition in Silicon Valley and Route 128. Harvard University Press, Cambridge, MA.

Scott, J.C., 1998. Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed. Yale University Press, New Haven.

Selznick, P., 1949. TVA and the Grass Roots: A Study in the Sociology of Formal Organization. University of California Press, Berkeley.

Serres, M., Latour, B., 1995. Conversations on Science, Culture, and Time. University of Michigan Press, Ann Arbor.

Storper, M., 1997. The Regional World: Territorial Development in a Global Economy. Guilford Press, New York.

Taylor, M., Asheim, B., 2001. The concept of the firm in economic geography. Economic Geography 77, 315–328.

Thrift, N., Olds, K., 1996. Refiguring the economic in economic geography. Progress in Human Geography 20, 311–337.

Walker, R., 1989. A requiem for corporate geography: new directions in industrial organization, the production of place and uneven development. Geografiska Annaler B 71, 43–68.

Weber, M., 1922. Wirtschaft und Gesellschaft. Mohr, Tübingen.

Weick, K., 1979. The Social Psychology of Organizing. Random House, New York.

Wenger, E., McDermott, R., Snyder, W.M., 2002. Cultivating Communities of Practice. Harvard University Press, Cambridge, MA.

Organization: Overview

W Richard Scott, Stanford University, Stanford, CA, USA
Gerald F Davis, The University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by W.R. Scott, Volume 16, pp. 10910–10917, © 2001, Elsevier Ltd.

Abstract

Organizations have been studied from many disciplinary perspectives and at many levels of analysis. This article focuses on organizations as distinct units, covering the twentieth-century evolution of the study of organizations and briefly summarizing the major theoretical approaches. We emphasize the shift from a focus on organizations as bounded units to organizing as a process and conclude with contemporary challenges for organization theory.

Organizations

Organizations are a defining feature of modern societies. Modern societies are distinguished from traditional ones in part by the numbers and variety of organizations in existence, and their utilization to support collective action in every sector. Organizations are both contexts for the actions of individuals and groups, and collective actors in their own right. Some, such as corporations, are given legal recognition and constituted so as to own and dispose off the property, enter into binding contracts, and mobilize effort in the pursuit of specified goals. Organizations can be formatted to increase the reliability of behavior, but in doing so, they often become rigid and constraining. They also are able to provide greater accountability, maintaining records of how things are done and by whom, but in doing so, may become overly procedural and legalistic. Today's organizations, operating in a highly rationalized and dynamic environment, face different challenges than early organizations confronting a more traditional and stable context. Conceptions of organizations, theoretical perspectives, and favored levels of analysis undergo change, in part because organizations and organizing processes themselves have changed.

Organizations have been studied from many disciplinary perspectives and at many levels of analysis. Psychologists, sociologists, political scientists, and economists have all made central contributions to the understanding of organizations. Psychologists typically study individuals, dyads, and groups within an organizational context to uncover distinctive features of organizational behavior, such as how the design of jobs or pay systems influences individual motivation, or how groups of managers make decisions. Political scientists have extended the study of decision making to the design of organizations and the pathologies of bureaucracy. Sociologists examine both the structures of individual organizations, their relations with other organizations, and aggregate features of organizational populations and fields. Economists study how markets influence the shape of organizations – where they place their boundaries, how they are owned and governed – and how firms interact in industries. Thus, organizational scholars span a range of units of analysis, from individuals and groups, to departments and divisions, to organizations and industries, to networks and fields, and even to the level of the nation-state and the global system itself. In each case, the vocabulary and theoretical toolkit varies according to the level, as do the questions considered 'interesting.'

This article focuses on organizations as distinct units (i.e., on 'organization theory') but it also acknowledges that there are alternative perspectives and trade-offs involved in choosing one focus rather than another. It first describes how dominant theoretical views of organizations evolved over the twentieth century. It then analyzes the distinguishing features of organization from a substantivist and a relational perspective. Next it discusses different views on the origins of organizations and how different levels of analysis channel different approaches to their object of study. Finally, it concludes with a discussion of new directions in the study of organizations.

Evolving Perspectives on Organizations

Three definitional perspectives on organizations, each embracing a number of schools, have been pervasive during the twentieth century in organizational studies, and each has been animated by its own dominant metaphor for organizations (Scott and Davis, 2007). The 'rational system' model of organizations was the dominant perspective from early in this century to the 1930s. Organizations are viewed as rationally designed instruments for the attainment of specific goals. A mechanical model provides the guiding image and emphasis is placed on those features that differentiate organizations from other social groups, such as communities or families. Exemplars include the 'scientific management' approach of F.W. Taylor, Weber's theory of bureaucracy (1968), and Simon's model of (boundedly) rational decision-making hierarchies (March and Simon, 1958). The twin attributes of goal specificity and formalization – the deliberate design of roles and social structures – are privileged.

The 'natural system' model of organizations emerged during the 1930s and became dominant in the scholarly community in the 1940s and 1950s. Organizations were viewed as social groups whose participants are joined by common purpose and informal ties, but whose primary goal is the survival of the system. Rational, formal attributes were de-emphasized as attention was focused on interpersonal ties and leadership. An

organic image supplanted that of the well-oiled machine; not distinguishing elements, but rather those common to all social groups were emphasized.

During the 1960s, general systems theory swept over the field of organization studies giving rise to the 'open system' model of organization. This model emphasized the extent to which organizations were dependent on and penetrated by facets and forces in the wider environment. By contrast, previous models were exposed as giving insufficient attention to actors and forces external to system boundaries. But as open system models surged into the arena, they quickly were coupled with, first, rational, and then natural system concerns (Scott and Davis, 2007). Thus, rational-open theorists like Thompson (1967) and Lawrence and Lorsch (1967) sought to explain how organizations could attempt to behave rationally while also open to their environments; March and Simon (1958) considered how decision rules can be adjusted to accommodate greater amounts of uncertainty; and Williamson (1975) theorized that organizations arise and develop more complex governance systems to deal with increasing amounts of uncertainly and complexity. Contingency theory, decision-making theory, and transaction cost theory represent attempts to combine rational and open systems conceptions.

These developments dominated well into the 1970s, but they then were joined by a resurgence of natural system theorists attempting to incorporate and take into account open system insights. Principal approaches included *socio-technical systems* theory, emphasizing the need to design technologies to take into account human and social factors, and *resource-dependence* theory, as developed by Pfeffer and Salancik (1978), which stressed the ways in which economic (resource-based) dependencies generated power problems that, in turn, were addressed by political solutions. In general, organizations confronted by increasing levels of interdependence respond by creating more encompassing coordinative and governance systems. *Population ecology* models (Hannan and Freeman, 1989) examine the ways in which similar types of organizations compete for scarce resources within the same environmental niche. And *institutional theorists* (Meyer and Rowan, 1977; DiMaggio and Powell, 1983) consider the effects of wider cultural rules and normative structures on the emergence and *legitimation* of organizations.

Thus, conceptions of organizations have been influenced by the theoretical stance of scholars, as debates revolved around two axes: (1) the extent to which organizations are means – disposable, deliberately designed instruments for goal attainment – or value-impregnated, ends in themselves and (2) whether organizations are self-sufficient, relatively insulated forms or substantially constituted, influenced, and penetrated by the wider environment.

Why Do Organizations Exist, and What Distinguishes Them?

Max Weber, who provided a theoretical foundation for much subsequent work, was among the first social theorists to recognize and attempt to characterize the features that distinguished organizations from other social groupings. First, a 'legitimate order' is a social relation oriented to a normative framework or set of maxims viewed as obligatory (Weber, 1922/1968: pp. 26–53). Social relations vary in the extent to which they are open or closed, that is, limiting access to outsiders. 'Organizations' (Verband) are social relationships that are closed, with enforced boundaries between the inside and outside. 'Enterprises' (Betreib) are organizations that have developed a system of continuous purposive activity directed to specified goals; and a 'formal organization' (Betriebsverband) is an enterprise based on associative relations resting on rationally motivated adjustments of interests rather than on communal ties, which are based on kinship or the subjective feelings of the parties. The two latter definitions have been employed widely by subsequent scholars.

Chester Barnard, perhaps the most influential early American organizational theorist, was a telephone executive who drew on his own practical managerial experiences to develop his conception of the distinctive properties of organizations. Barnard (1938) emphasized the physical, biological, and psychological limitations of individual actors, limitations that greatly constrain their ability to make intelligent choices. Cooperative systems arise to enable individuals to collectively pursue a given purpose, the purpose acting to both support and constrain individual choice and decision making. Such systems are 'effective' to the extent that they achieve their purpose; they are 'efficient' if they satisfy the motives of individual participants. To be viable, systems must induce participants to make contributions; thus, only efficient systems can survive. Organizations represent a distinctive type of cooperative system: a 'formal organization' is 'that kind of cooperation among men that is conscious, deliberate, purposeful' (p. 4), although all organizations incorporate both formal and informal systems.

Weber's intellectual descendants include a host of sociological scholars, from Parsons to Etzioni, Blau, and Albrow. All of these theorists follow Weber in emphasizing goal specificity and formalization as distinguishing features of organizations and in recognizing the importance of power and hierarchy both in defining goals and controlling access and specifying the terms of participation. Barnard's followers encompass a diverse collection of researchers, including political scientists Simon and March, economists Arrow and Williamson, and sociologists Selznick and Perrow. Devotees from political science and economics embrace Barnard's view of organizations as arising from the limitations of individuals – physical, but more important, cognitive constraints – organizational structures and processes acting to both inform and channel their decision making. These same theorists also stress the rationality of individuals, insisting that their participation and willingness to make contributions is contingent on the continuing receipt of appropriate inducements. A modified version of Barnard's approach, developed by Cyert and March (1963), views organizations as coalitions of interests, with diverse parties vying for dominance. Sociologists, by contrast, focus on Barnard's recognition of the motivating power of purpose and the ways in which organizational structures and procedures themselves become infused with value, so that, as both instruments and ends in themselves, survival of the organization becomes an overriding goal for participants (Selznick, 1949). Organizations are observed to dispense with

specific objectives when their pursuit undermines organizational viability.

Relational Views of Organizations

Although Weber and his followers have viewed organizations as substantive entities, emphasizing boundaries and goals, other theorists have taken a more relational perspective focused on 'organizing' rather than 'organization.' Karl Weick was a major early proponent of a more relational conception of organizations:

> We observe either an ongoing process that appears "frozen" and steady because it is glimpsed only briefly, or we observe that the process is continuously changing if we watch for a longer span of time. The point is that the crucial events to be explained are processes, their structuring, modification, and dissolving. (Weick, 1969: p. 16)

This stance has been embraced and elaborated by theorists such as Giddens (1984), who proposed that 'structure' be reconceptualized as 'structuration' – the continual production and reproduction of resources and rules or schemas – as well as by interpretive and postmodernist theorists who emphasize the social construction of reality and the reflexive and contextualized nature of knowledge. Among organizational scholars, Silverman (1970) employed an interpretive framework to argue that the dominant systems-based view of organizations needed to be replaced by an action-based account. He asserted that action analysis is particularly well suited to addressing neglected issues, such as the origins of organizations and organizational change. In a more programmatic statement, Burrell and Morgan (1979) distinguished between theoretical approaches stressing stability and order versus those emphasizing radical change. Their work stimulated efforts to advance interpretive and radical structuralist paradigms. Among theorists developing a more interpretive or cultural approach to organizations, a number emphasize the value of appropriating and adapting the tools developed by literary and humanistic scholars and suggest that society and its component units can be viewed and analyzed as 'texts' – that is, as a structure or grammar of rules and communication acts. Some propose the use of 'dramaturgic' approaches, focusing on the expressive dimension of social relations and the role of rituals and ceremonies in expressing commitments and dealing with social uncertainties. Others assert the value of 'narrative' analysis, which views organizations as ongoing conversations among interacting individuals involved in formulating and reformulating intentions, interpretations, and accounts to themselves and each other (Czarniawska, 1997).

Other relational scholars have emphasized not interpretive construction but rather network connections. *Network theorists* emphasize the value of viewing organizations as shifting networks of social relations embedded in fields of wider relational systems. Location or position in the network may be of greater significance in shaping an organization's behavior than its internal attributes. Both the organization's capacities and constraints on its action are determined by the ever-changing configuration of ties (and the lack thereof) among an organization and counteractors.

Network scholars have examined both networks within organizations and networks across organizations. Interorganizational networks can be created through a variety of connections: alliances; buyer–supplier relations; personnel flows; shared directors on the board; and even mere physical proximity, which can create informal connections (e.g., in Silicon Valley, where engineers from different companies and universities famously swap stories and ideas at shared watering holes). Network theorists have evolved an armory of methods and measures to examine the structure of social relations, and their insights have been absorbed into other areas of organizational scholarship For instance, institutional scholars have examined how the structure of social networks can channel the diffusion of innovations, often leading to field-level conformity in practices and structures. Other scholars have generated distinct insights on networks in themselves. Burt (1992) provided extensive theory and evidence showing that actors holding a position rich in structural holes (standing at the nexus of otherwise-unconnected clusters) often are presented with opportunities to act as a broker, innovator, or entrepreneur, bringing together ideas, people, and resources that otherwise might not have been connected. Podolny (2005) noted that networks are not just 'pipes' for channeling information and resources but also 'prisms' that reflect one's status: who you know matters, but who you are seen with sometimes matters even more.

The relational perspective has found allies in surprising places. Economists recognize that one of the most important features of an organization is its status as an independent legal entity. It is capable of entering into binding contracts in its own name, separate from those involving its individual participants. Parallel to the relational view of organizations in psychology and sociology, economists developed a view of the organization as a 'nexus of contracts': the firm as a 'focus for a complex process in which the conflicting objectives of individuals (some of whom may 'represent' other organizations) are brought into equilibrium within a framework of contractual relations' (Jensen and Meckling, 1976: p. 311). The stability of the equilibrium reached is a function of many factors, including changes in legal systems and contracting norms as well as the composition of and changing interests of participants.

Relational approaches celebrate process over structure – that is, becoming over being (cf Emirbayer, 1997). What is being processed varies greatly. In some versions, it is symbols and words; in others, relationships or contracts; in still others, assets. But in relational approaches, if structures exist it is because they continually are being created and recreated, and if the world has meaning, it is because actors are constructing and reconstructing intentions and accounts and, thereby, their own and others' identities.

Emergence of Organizations

Organizations are viewed as a construction of the modernist world, but there is much debate as to whether they represent accompaniments to and products of *industrialization* or, more

generally, vehicles of *rationalization*. From Adam Smith to Clark Kerr and Alfred Chandler, economists have viewed organizations as handmaidens to industrialization. The subdivision of tasks enables specialization and mechanization, but it usually is accompanied by agglomeration – the gathering of workers together in a common location – and requires coordination to integrate discrete tasks, typically through a command and control structure. A more recent but compatible conception advanced by *institutional economists* (Williamson, 1975), views organizations as operating to reduce transaction costs occasioned by the cognitive constraints (bounded rationality) of individual actors and the need to regulate opportunism among exchange partners. In either case, organizations are viewed as emerging to enhance efficiency: to reduce costs, whether of production or transactions.

Many of the most prevalent features of organizations – rules, formalized roles, dedicated machines, and established work routines – are intended to contribute to the reliability of performance. As often noted, however, these same mechanisms often prevent the organization from changing quickly enough to keep pace with environmental demands (Hannan and Freeman, 1989).

Sociological institutionalists argue that organizations should be viewed as a cultural construction: A product of increasing rationalization as generalized means-ends rules are promulgated and become valued independent of their contributions to heightened effectiveness and efficiency. The proliferation of organizations, and their infusion into virtually every arena of contemporary society, is viewed as resulting not only from their contribution to the coordination of differentiated, specialized tasks but also to their capacity to symbolize rational (rule-based, means-ends) procedures (Meyer and Rowan, 1977). The formal structure of organization, hierarchical controls, and the compilation of records all contribute to ensuring the accountability of organization performance and thereby enhance their legitimacy. These same structures, however, also are associated with red tape and ritualistic practices.

Viewed historically, organizations represent new types of social forms in that they are collective or corporate actors, accorded rights, capacities, and resources that are independent of those of any of its individual participants. Coleman (1974) employed changes in the law as important indicators of the growing independence and powers of organizations as they developed during the early modern epoch. Institutionalists such as Meyer, Dobbin, and Shenhav viewed the same process as being fueled primarily by the growth of a scientific ethos emerging during the Enlightenment that created abstract and general categories to classify and enumerate first the biological and physical universe and subsequently the social world. Local knowledge and culture, over time, succumbed to a new inductive epistemology viewed as universal. In the realm of purposive social organization, engineers and managers envisioned and then gradually helped to bring about the development of a single type of social entity (the 'organization') and a generic set of skills ('managerial') required for their maintenance. It is a vision that has been embraced and, hence, reinforced by the emergence of an academic discipline and profession devoted to the pursuit of organizational studies.

Levels of Analysis

The concept of organization is employed at various levels of analysis. More micro approaches treat organizations as distinctive types of social contexts that, in turn, affect the behavior and attitudes of their participants, whether considered as individuals or as groups or teams. Thus, the first level concentrates on organizational behavior. Viewed as social contexts, organizations generally are thought to be 'strong situations': Much effort and various techniques and devices are employed in an attempt to influence participant behavior. Positions typically are predefined, roles are prescribed, rules are formulated, authority relations are specified, information is routed selectively, specialized tools and equipment are provided, training is offered, incentives are created and sanctions are employed, and beliefs and values are inculcated. As Pfeffer (1997: p. 100) observed: "control is at once the essential problem of management and organization and the implicit focus of much of organization studies." Theorists differ in the relative priority they assign to different modes of control, whether economic, social, or cultural; but organizations are apt to combine many varieties of incentives and constraints in such a manner as to maximize the likelihood of compliance.

Moving up one level from a focus on organizational behavior, many analysts focus on the structural and other features of the organization qua organization. Here the organizational arrangements – extent of formalization, bureaucratization, centralization, mechanization, professionalization – instead of serving primarily as hallmarks of the setting become attributes of the corporate actor to be described and explained. Extensive comparative research has been conducted attempting to account for the widely varying features of organizational structure. Among the causal factors examined are the characteristics of the task and institutional (including cultural and political) environments, the demography of participants, and the life cycle of the organization.

When a particular organization is singled out and its context is described as it relates to this focal unit, an organization set level is being employed. This level was developed in association with the early stages of the open systems perspective and facilitates the examination of the particular information and resource flows and types of relations affecting the organization of interest. It is instructive to replace general descriptions of environmental variables, such as complexity and uncertainty, with specific information on who is providing what information and resources under what terms to the focal organization. This level of analysis has been particularly supportive of analysts employing *resource dependence*, *transaction cost*, and *strategic management approaches* (Pfeffer and Salancik, 1978; Thompson, 1967; Williamson, 1975).

The organizational population level is defined to identity and encompass all instances of a geographically bounded aggregation of all organizations of the same type. This level roughly parallels that of an *industry*, except that attention is limited to organizations of the same general type. The concern here is not the individual organization but rather the *organizational form*: variously defined as similar structures, a common blueprint or archetype, or distinctive constitutive information. Researchers working at this level examine the conditions under which new ways of organizing and new types of routines arise

and coalesce to constitute a recognizable, accepted form for the pursuit of specified goals. Analysts of populations examine the conditions under which populations of organizations emerge, grow, decline, and fail (Hannan and Freeman, 1989).

At the macro end of the scale, theorists have identified the organizational field as an important level linking individual organizations and organizational systems to wider, societal structures and processes. Organizational fields are defined as 'those organizations that, in the aggregate, constitute a recognized area of institutional life: key suppliers, resource and product consumers, regulatory agencies, and other organizations that produce similar services and products' (DiMaggio and Powell, 1983: p. 143). Fields are made up of multiple populations of organizations and their respective organization sets. Organizational fields vary greatly in terms of the extent of their organization and the degree of their structuration; and the emergence, development, and decline of organizational fields has become an important new focus of study.

Changing Theories; Changing Organizations

It would be odd indeed if there were no relation between our changing definitions and conceptions of organizations and changes in the phenomena itself. Over the past few centuries, organizations have evolved from relatively embedded and dependent components of traditional society to more independent systems that attempted to insulate themselves from the surrounding society, to systems that are much more open, more flexible, and more penetrated by their context. Since the beginning of the industrial period, organizations have developed from highly centralized to more decentralized structures; and they have tended to move from more tightly to more loosely coupled systems. Rather than having their component units be fixed jobs and departments, they increasingly are made up of more diffusely defined skill categories whose participants are grouped into changing configurations of teams and project groups. Also, jobs formerly linked densely in *internal labor markets* providing clear career lines to employees now more often are held by temporary workers or independent contractors whose mobility routes cross organizational boundaries. Rather than attempting to incorporate all functions within the same hierarchical ownership structure, organizations increasingly find ways to contract work out to relationally linked independent firms or to develop alliances with a shifting set of independent partners. New developments in information processing allow the collaboration of geographic remote partners and various combinations of off-site, part-time, and loosely linked participants. Ownership is no longer viewed as necessary – and in some situations is seen to undermine – control of essential resources and services.

Two master trends have shaped new understandings of organizations. First, globalization – the increasing interconnections in information, culture, and trade among the different countries of the world – make it clear that 'organizations' and our understanding of them are somewhat like breakfast. Although the first meal of the day might be labeled 'breakfast' everywhere, the actual content and size of that meal, and what is considered appropriate and sensible, varies wildly. In the United States, breakfast might consist of dry cereal and milk; in Switzerland, yogurt and muesli, in the United Kingdom, eggs, baked beans, and sausage; in Korea, kimchee, soup, and rice; in France, a croissant; in Argentina, steak. Appropriate beverages range from milk to tea to coffee to hard liquor. Similarly, organizations may share a designation, but surprisingly little else. Documenting diversity around the world has been a growth industry in the study of organizations.

Second, information and communication technologies – computers, telephony, and particularly the advent of the Internet – drastically have changed the feasible forms of organization around the world by vastly reducing the costs of coordination. The transaction cost profiles of different forms of collective action have changed radically, such that it now is possible to accomplish through dispersed horizontal groups feats that would have required formal hierarchies in prior times. Linux is now the world's most pervasive operating system, used on literally billions of electronic devices, and Wikipedia is the world's most widely used encyclopedia (and an indispensible tool for students everywhere); both were created via open-source communities consisting of tens of thousands of individuals who have never met – a format that fits our definitions of 'organization' only problematically. 'Organizing' has surpassed 'organization' in field after field (Scott and Davis, 2007).

We thus are left with a paradox: On the one hand, we seem to live in a society of organizations in which nearly every aspect of our lives occurs in and through formal organizations, including birth and other vital events, education, politics, economic mobility, and even family life itself, as more of what we do on a daily basis, from cooking food to changing the oil to tutoring our children can be contracted out to suppliers (cf Perrow, 1991). On the other hand, societies around the world have become so rationalized that the order formerly created within organizations increasingly is embedded in society itself (Meyer, 1983). The largest corporations increasingly are evanescent, relying on networks of contractors and fictional legal devices around the world to design, produce, and distribute their products. The functions of organization increasingly are available off the shelf to entrepreneurs who snap them together and take them apart like interlocking plastic toys that increasingly resemble the 'nexus of contracts' described by economists (Davis, 2009). The twenty-first century already has shown that organizing will continue to demand new strategies for understanding.

See also: Bureaucracy and Bureaucratization; Bureaucracy, Sociology of; Bureaucratization and Bureaucracy, History of; Hierarchies and Markets; Narrative Approaches to Organizations; Organizational Behavior, Psychology of; Organizational Climate; Organizational Culture, Anthropology of; Organizational Decision Making; Organizational Ecology; Organizations and the Law; Organizations, Sociology of; Rational Choice and Organization Theory; Social Networks; Weber, Max (1864–1920).

Bibliography

Barnard, C.I., 1938. Functions of the Executive. Harvard University Press, Cambridge, MA.

Burrell, G., Morgan, G., 1979. Sociological Paradigms and Organisational Analysis. Heinemann, London.
Burt, R.S., 1992. Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge, MA.
Coleman, J.S., 1974. Power and the Structure of Society. Norton, New York.
Cyert, R.M., March, J.G., 1963. A Behavioral Theory of the Firm. Prentice-Hall, Englewood Cliffs, NJ.
Czarniawska, B., 1997. Narrating the Organization: Dramas of Institutional Identity. University of Chicago Press, Chicago.
Davis, G.F., 2009. Managed by the Markets: How Finance Reshaped America. Oxford University Press, Oxford, UK.
DiMaggio, P.J., Powell, W.W., 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. American Sociological Review 48, 147–160.
Emirbayer, M., 1997. Manifesto for a relational sociology. American Journal of Sociology 103, 281–317.
Giddens, A., 1984. The Constitution of Society: Outline of the Theory of Structuration. University of California Press, Berkeley, CA.
Hannan, M.T., Freeman, J., 1989. Organizational Ecology. Harvard University Press, Cambridge, MA.
Jensen, M., Meckling, W., 1976. Theory of the firm: managerial behavior, agency costs, and ownership structure. Journal of Financial Economics 3, 305–360.
Lawrence, P.R., Lorsch, J.W., 1967. Organization and Environment: Managing Differentiation and Integration. Graduate School of Business Administration, Harvard University, Boston.
March, J.G., Simon, H.A., 1958. Organizations. Wiley, New York.
Meyer, J.W., 1983. Institutionalization and the rationality of formal organizational structure. In: Meyer, J.W., Scott, W.R. (Eds.), Organizational Environments: Ritual and Rationality. Sage, Beverly Hills, CA, p. 261.
Meyer, J.W., Rowan, B., 1977. Institutionalized organizations: formal structure as myth and ceremony. American Journal of Sociology 83, 340–363.
Perrow, C., 1991. A society of organizations. Theory and Society 10, 725–762.
Pfeffer, J., 1997. New Directions for Organization Theory: Problems and Prospects. Oxford University Press, New York.
Pfeffer, J., Salancik, G., 1978. The External Control of Organizations. Harper, New York.
Podolny, J.M., 2005. Status Signals: A Sociological Study of Market Competition. Princeton University Press, Princeton, NJ.
Scott, W.R., Davis, G.F., 2007. Organizations and Organizing: Rational, Natural and Open System Perspectives. Pearson Prentice-Hall, Upper Saddle, NJ.
Selznick, P., 1949. TVA and the Grass Roots. University of California Press, Berkeley, CA.
Silverman, D., 1970. The Theory of Organisations: A Sociological Framework. Heinemann, London.
Thompson, J.A., 1967. Organizations in Action. McGraw-Hill, New York.
Weber, M., 1922/1968. In: Roth, G., Wittich, C. (Eds.), Wirtschaft und Gesellschaft: Grundriss der verstehenden Soziologie. Economy and Society: An Outline of Interpretive Sociology. J. C. B. Mohr, Tubingen.
Weick, K.E., 1969. The Social Psychology of Organizing. Addison-Wesley, Reading, MA.
Williamson, O.E., 1975. Markets and Hierarchies: Analysis and Antitrust Implications. Free Press, New York.

Organizational Behavior, Psychology of

Bernhard Wilpert[†]

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 16, pp. 10917–10921, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editors.

Abstract

Organizational behavior (OB) denotes a relatively new area of research and practice of psychology. It refers to the study of both the psychological and social antecedents and consequences of behavior of individuals and groups in organizations, as well as the psychological and social antecedents and consequences of the behavior of organizations. Stressing a psychological point of view, this article defines the field of OB and describes the main characteristics of its dominant domains, their methodological implications, and the likely future directions of OB.

Organizational behavior (OB) denotes a relatively new area of research and practice of psychology. It refers to the study of both the psychological and social antecedents and consequences of behavior of individuals and groups *in* organizations, as well as the psychological and social antecedents and consequences of the behavior *of* organizations. Stressing a psychological point of view, this article defines the field of OB and describes the main characteristics of its dominant domains, their methodological implications, and the likely future directions of OB.

The Nature of the Field

As in any new scientific field, various attempts have been made by different authors to define the field of OB. In the very first review entitled 'Organizational Behavior' in the *Annual Review of Psychology*, Mitchell (1979: p. 244) considers OB as a "distinct discipline with a focus on individual and group behavior in the organizational context," thus making it virtually synonymous with traditional work and organizational psychology. Staw (1984: p. 628) underlines the interdisciplinary character and the two main orientations of the field, the stress on micro- and macroissues: "organizational behavior is an interdisciplinary field that examines the behavior of individuals within organizational settings as well as the structure and behavior of organizations themselves." According to this view, microaspects are dealt with by psychology whereas macroaspects are treated by sociology, political science, and economics, making OB appear as two distinct subdisciplines, an impression that is further fostered by different professional identifications of the students of OB. Others simply define OB operationally in terms of the organizational research topics and issues published over time in scientific psychological periodicals (Rousseau, 1997) covering, for instance, leadership, motivation, organizational technology, job analysis, legal issues, and organizational performance.

A more integrative stance is taken by Schneider (1985: p. 574) in focusing on the types of studies conducted in OB: "OB is the confluence of individual, group, and organizational studies flowing from industrial–organizational (I/O) psychology and organizational and management theory (OMT) with headwaters in psychology (social, psychometrics), sociology (organizational, work, and occupational), and management (scientific, human relations)." Given this mixed heritage of OB, its scope is extremely broad, covering both micro- and macrophenomena of organizations. McGrath (1976), in an early, even more comprehensive manner, conceptualizes OB as the intersecting area of three systems: the physical–technical/structural, the social–interpersonal, and the person system of organization members.

From a psychological point of view, OB may thus be defined as an emergent field of scientific inquiry, which is constituted by the converging contributions of various social and behavioral sciences focusing on multilevel and multidimensional antecedents and consequences of psychological phenomena in organizations. The multilevel perspective of OB implies that literally *all relevant levels and dimensions* that either condition or are conditioned by psychological phenomena in organized settings fall into the scope of OB studies: the levels of the individual–personal, the group, the managerial, the technological, the organizational–structural, the interorganizational, and environmental. The notion of multiple dimensions refers here to organizational practices, processes, and outcomes of any of the relevant system levels as proper problems of OB studies.

Although the study (Greve, 2013) of organizational issues has been part and parcel throughout the history of work and organizational psychology (Münsterberg, 1912), it remains a moot question why it was only during the last three decades of the twentieth century that OB entered the scene as a distinct subfield of scientific inquiry. One may hypothesize that factors seem to have interacted here that are partly due to change in society and partly due to intrinsic developments of science, both of which made their effects particularly felt in the decades after World War II. Epochal shifts in the work force such as the growth of the service sector of economy, the demographic changes in terms of age and gender composition implied drastic restructuring of vocational and professional profiles. The cost-effective wholesale introduction of new information and communication technologies in work settings created

[†] Deceased.

dramatic challenges to a novel division of labor. Social value changes such as the changing subjective meaning of working (MOW, 1987) and the increase of international market competition may have further been contributing factors of societal change to highlight the need for the reform of work organizations and social institutions that had to go far beyond a restructuring of individual work places. Required were more holistic approaches to the study and practice of organizational change and performance that eschewed monodisciplinary psychological treatments. This is where progress in science offered itself as a promising solution. The emergence of systems theory as a metatheoretical intellectual vehicle to facilitate multilevel studies and to link different scientific disciplines to each other in the analysis of complex scientific and practical problems appears to have been a further necessary condition for OB to acquire scientific credibility as an interdisciplinary field.

Despite its comparatively short history, OB is far from displaying a unified common theoretical framework and a corresponding set of issues under study. The most important theoretical underpinnings and the substantive problems studied in their perspective will be addressed next. These theoretical perspectives represent umbrella orientations usually encompassing diverse theoretical positions of varying scope, which depending on one's meta-theoretical penchant, may also be classified differently.

Theoretical Paradigms and Topical Foci

Psychology of Individual Differences

The basic tenet of differential psychology within OB is the assumption that personal characteristics constitute the core of variables to be studied by psychology either as causal (at least moderating) factors of consequent psychological/social phenomena or as the most important outcomes of individual experience in organizational settings. The micro-perspective dominates here. Within this theory class might also be counted the important approaches of expectancy and of goal setting theory in as much as individual characteristics (e.g., preferences, competencies) need to be considered by them. The former surmises that subjectively expected outcome values (subjective expected utilities (SEU)) guide individual decision-making behavior (Naylor et al., 1980), the latter postulates that goals, self-set individually or interactively by subordinates and superordinates, constitute potent predictors of subsequent behavior in organizations.

Major topical concerns within this paradigm concentrate on issues such as the role of personality and leadership style as conditioning work group and organizational performance (including absenteeism from work, turnover), work attitudes (causes and consequences of job satisfaction, meaning of working, and work commitment/motivation).

Bureaucratic Theory

An important contribution to the advancement of OB was made by the interdisciplinary team of Aston University (Birmingham, UK) under the leadership of a psychologist, Derek Pugh (Pugh and Hickson, 1976). Taking Max Weber's theory of bureaucracy as the starting point, the Aston group deviated from the received single-level analytic procedure and proposed to overcome this by integrating the analysis of three levels: situational factors (size of organization, technology, environment), organizational structure (decision centralization, specialization, formalization, configuration), and member experience and behavior. A basic assumption was that cross-level effects had to be taken into account in the sense that an organization's technology would impact the division of labor, i.e., the role structure of the organization which, in turn, would influence autonomy and individual outcomes such as the evaluation of one's work setting (motivation, commitment, satisfaction). Thus the Aston approach covers both micro- and macro-dimensions of OB.

As a major achievement of the Aston group, apart from introducing from its outset a theoretical multilevel and interdisciplinary perspective must be considered its success in setting current psychometric standards for the development of instruments for an adequate analysis of organization structure and its interaction with variables on other system levels. A burgeoning body of studies followed from their lead.

Cognitive Psychology

The generally noted 'cognitive shift' in psychology during the latter part of the twentieth century soon was also to reach OB and to influence paradigmatic thinking in the field. Cognitive processes are in this perspective considered to mediate or govern behavior. Various positions fall under the umbrella category of cognitive psychology, some of which achieved prominence in OB; the three most important ones are social cognition, expectancy theory, and attribution theory (Hodgkinson, 2003).

Social cognition refers to processes among relevant actors, which produce socially shared representations, views, and perceptions of organizational realities. Thus some 'social constructivist' authors claim that organizational phenomena are mainly socially constructed and articulated through negotiated goals, structures, artifacts, through myths, rituals, stories, and symbols, which result in organization-specific, intersubjectively shared meanings and values (Turner, 1990). Social constructivism favors an ethno-methodological and anthropological, holistic approach to organizations, which lends itself to the analysis of molar organizational features such as organizational culture and climate. Under the heading of social cognition must also be mentioned the significant contribution of equity theory, which focuses on an actor's input–output ratio in comparison with the respective ratio of similar others with the resulting energy expenditure and work satisfaction. *Expectancy theory*, as described above, addresses the cognitive processes of organization members' choices according to perceived SEU. It has shown considerable fruitfulness in the study of leadership, self-efficacy, and work motivation. *Attribution theory*, owing to its focus on perceptual processes concerning perceived causes of success and failure, has to a large extent influenced theorizing and empirical research on

achievement motivation and respective performance in organizations.

Action Theory

Many different theoretical social and behavioral science approaches go under the guise of 'action theory.' Some of them play a major role in OB. A large body of German theoretical and empirical literature under the name of action control/regulation theory needs to be subsumed here (Frese and Zapf, 1994). It combines elements of cybernetics and cognitive aspects of goal-oriented behavior (actions) that is assumed to be regulated by feedback loops concerning the deviations from goal achievement, which necessarily includes the consideration of task and technical components of task and organization structures. The approach has proven its usefulness in a wide variety of domains, e.g., human error analysis, personality and competence development, individual task and organization design, linking socially shared meanings, and values to collective action control.

A notable French approach of sociological provenance (Crozier and Friedberg, 1977) puts social actors (individuals, groups, organizations) and their various strategies of playing their power games into the center of its theoretical and empirical analysis. In this context, games are not understood in a metaphoric sense, but as concrete mechanisms to structure organizational power relations among interdependent actors negotiating degrees of freedom within given organizational and environmental constraints and opportunities. Games are here considered as the very instruments that regulate and sustain stability and integration of organized behavior based on accepted rules of the games. Thus the authors succeed in encompassing both micro- and macroaspects of OB.

The notion of autopoiesis, derived from biology (Maturana et al., 1980), has attracted growing interest also in OB. The concept refers to the continuous and structurally coupled adjustment of living systems to their extant environment and their continuous systemic self-reproduction. It has been applied to, among others, self-organizing capacities of institutions of public administration and their sustained regeneration of the bureaucracy and the autonomy and self-governance of organizations acting in the public area. The link of micro- and macroaspects of organizations in their given environment becomes clearly dominant in this perspective.

Systems Thinking

A lasting paradigmatic contribution to OB was made by the so-called socio-technical systems approach (STSA), developed soon after World War II by a group of social scientists of the London-based Tavistock Institute of Human Relations (Emery and Trist, 1960). The approach grew out of general systems theory and is predicated on important assumptions concerning the interaction of an organization's *social* subsystem (its members, role structure) and its *technical* subsystem (task structure, technology). In linking task and technological elements to the social patterns of organization, STSA paved the ground for systematic analysis and design of work places and work organizations. Work and organizational design was postulated to proceed always through joint optimization of both subsystems in order to guarantee efficiency. The increased participation of employees in organizational change programs became a leading principle. Thus, it was possible to take into account both the characteristics of organization members (their values, aspirations, expectations, social and hierarchical relations, role ascriptions, etc.) and the organization's technical dimensions.

The theoretical innovations of the STSA proved to be of great practical usefulness in providing guidelines for the design of work and work organizations which met, for instance, the demands of differential levels of human competencies and also of technical and task demands. The large majority of work humanization programs of the 1970s and 1980s profited from them in many industrialized countries. It facilitated improvements in the adequate introduction of new information and communication technologies, the creation of flexible production units, the formation of semiautonomous work groups, and the creation of large-scale industrial democracy schemes.

Unresolved Issues and Future Directions of OB

The multilevel perspective of OB poses an intriguing *theoretical problem*: How can and should the relationships of several system levels be conceptualized? One case in point is problems such as the structural interrelations of organization-specific features (organization strategy and structure, membership resources, organizational practices, output) and characteristics of their institutional and societal environment (dominant culture, education and legal system, economic and technological status). While there is growing evidence that these features are interrelated and influence each other, theoretical integration remains lacking. A possibly promising theoretical solution might be found in Giddens' theory of structuration. Giddens (1984) postulates a dual nature of institutions recursively reconstituting and maintaining each other. On the one hand, institutions, by virtue of their existence, facilitate action of individuals and of organizations, and on the other hand, by virtue of the acting actors, the institutions are recursively constituted and sustained. The theory still being on a rather abstract level requires further development in order to make it fruitful in OB.

A *methodological problem* of considerable difficulty is the need to account for effects of time. Change and continuity of organizations may be likened to quasistationary equilibria (Lewin, 1952). Like rivers they are in continual flux while preserving their identity and basic characteristics for some time. One methodological response of OB to study change and continuity was to repeat several cross-sectional studies at various points in time, thus approaching some sort of longitudinal evaluation. Another response was to trace organizational developments historically through document analysis and interviewing. Yet another approach seems to be to employ more 'dense' descriptions of organizational processes over time by way of ethnographic participant observation. This is a choice that is increasingly advocated and implemented by many researchers in spite of its cost, simply because of its superior information yield.

Finally, it may be that at the turn of the millennium, *societal changes* as alluded to above will gather further momentum to

induce also a new organizational era impacting on OB. According to Rousseau (1997), organizations seem to be faced with the challenge of increasing upheaval and transition of depth and magnitude that were hitherto unknown. She refers to governmental measures of deregulation, global competition, the explosion of information technology in manufacturing and service sectors, the uncoupling of traditional education from work demands, the emergence of small-firm employment, new and more differentiated employment relations, outsourcing among firms and interorganizational networking, shifts from managerial prerogatives to self-management, transition from rule-regulated work to self-reliance, and continuous adaptation of workers. In short, traditional internal organizational structures are becoming more fluid and the boundaries of organizations blurred. The perceived meaning of organization as a clear-cut entity seems in need of change to adaptive *organizing*. A reorientation of OB with reference to thematic foci in addition to theoretical underpinnings would have to be the consequence of such developments. Theoretically, a renewed emphasis on the conceptualization of the dynamics of organizational and interorganizational processes is called for. New topical issues are likely and desirable to emerge as concern for OB, such as the changing nature of employment and equitable forms of remuneration (Scott and Davis, 2007), emergence of multiple loyalties, individual learning for self-management and career development, organizational learning, and intervention methods to introduce and to sustain change. OB seems to be facing new and wide areas to be examined, a task for which its intrinsic multilevel and interdisciplinary outlook will be a definite asset.

Comprehensive treatments of the field can be found in the following publications: *Handbook of Industrial and Organizational Psychology*, vols. 1–4 (see Frese and Zapf, 1994); *Annual Review of Psychology; International Reviews of I/O Psychology; Administrative Science Quarterly; Organizational Studies; Human Relations;* and *Journal of Occupational and Organizational Psychology*.

See also: Behavioral Theories of Organization; Corporate Culture; Group Processes in Organizations; Human Resource Management, Psychology of; Organization: Overview; Organizational Climate; Organizational Decision Making; Organizations: Authority and Power; Wellbeing and Burnout in the Workplace: Organizational Causes and Consequences.

Bibliography

Crozier, M., Friedberg, E., 1977. l'Acteur et le Système. Les Contraintes de l'Action Collective. Editions du Seuil, Paris.

Emery, F.E., Trist, E.C., 1960. Socio-technical systems. In: Churchman, C.W., Verhulst, M. (Eds.), Management Science, Models and Techniques. Pergamon, New York, pp. 83–97.

Frese, M., Zapf, D., 1994. Action as the core of work psychology. A German approach. In: Dunnette, M.D., Hough, L.M., Triandis, H. (Eds.), Handbook of Industrial and Organizational Psychology, second ed., vol. 4. Consulting Psychologists Press, Palo Alto, CA, pp. 271–340.

Giddens, A., 1984. The Constitution of Society: Outline of the Theory of Structuration. University of California Press, Berkeley, CA.

Greve, H.R., 2013. Microfoundations of management: Behavioral strategies and levels of rationality in organizational action. The Academy of Management Perspectives 27 (2), 103–119.

Hodgkinson, G.P., 2003. The interface of cognitive and industrial, work and organizational psychology. Journal of Occupational & Organizational Psychology 76 (1), 1–24.

Lewin, K., 1952. Group decision and social change. In: Swanson, G.E., Newcomb, T.M., Hartley, E.L. (Eds.), Readings in Social Psychology. Holt, New York, pp. 459–473.

Maturana, H.R., Varela, F.J., Cohen, R.S., 1980. Autopoiesis and Cognition: The Realization of the Living. Reidel, Boston.

McGrath, J., 1976. Stress and behavior in organizations. In: Dunette, M. (Ed.), Handbook of Industrial and Organizational Psychology. Rand McNally, Chicago, pp. 1351–1396.

Mitchell, T.R., 1979. Organizational behavior. Annual Review of Psychology 30, 243–281.

MOW—Meaning of Working International Research Team, 1987. The Meaning of Working. Academic Press, London.

Münsterberg, H., 1912. Psychologie und Wirtschaftsleben. Barth, Leipzig, Germany.

Naylor, J.C., Pritchard, R.D., Ilgen, D., 1980. A Theory of Behavior in Organizations. Academic Press, New York.

Pugh, D.S., Hickson, D.J., 1976. Organizational Structure and its Context. The Aston Programme I. Saxon House, Farnborough, UK.

Rousseau, D.M., 1997. Organizational behavior in the new organizational era. Annual Review of Psychology 48, 515–546.

Schneider, B., 1985. Organizational behavior. Annual Review of Psychology 36, 573–611.

Scott, W.R., Davis, G.F., 2007. Organizations and Organizing: Rational, Natural and Open System Perspectives. Pearson Education, Upper Saddle River.

Staw, B.M., 1984. Organizational behavior: a review and reformulation of the field's outcome variables. Annual Review of Psychology 35, 627–666.

Turner, B.R. (Ed.), 1990. Organizational Symbolism. de Gruyter, Berlin.

Organizational Citizenship Behavior

Dennis W Organ, Bloomington, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Organizational citizenship behavior (OCB) refers to discretionary, nonrequired contributions by members to the organizations that employ them. Evidence indicates that job satisfaction is more closely related to such contributions than to productivity in core job tasks. Other data suggest that personality also is more likely to predict such discretionary behaviors rather than task productivity. Various typologies and dimensions of OCB have been posited, but the most parsimonious structure consists of two major dimensions, roughly analogous to discretionary help and support to particular persons and discretionary levels of conformity to organizational rules. Compelling findings support the hypothesis that unit level OCB indeed predicts various criteria of organizational effectiveness. Cross-cultural research in OCB has begun and is proceeding at an accelerating pace, with indications that the structure of what is perceived as OCB varies somewhat across cultures. Moreover, some forms or levels of intended OCB might prove dysfunctional for either the organization or the individual or both.

The Construct and Its Boundaries

Organizational citizenship behavior (OCB) pertains to discretionary behavioral contributions that members (e.g., employees) render to their organizations. The qualifier 'discretionary' conveys the notion that such behavior is either not part of the job description or is a dimension of the job requirements that cannot be easily quantified (e.g., 'cooperates with colleagues') or enforced beyond some minimal level (such as attendance or punctuality). Some examples of specific OCB are helping a new hire 'learn the ropes', exemplary attendance, making timely suggestions to the appropriate parties and in an appropriate forum for more efficient operations, acting to prevent or resolve conflicts between or among colleagues, or spontaneously helping a coworker with some problem or urgent task. In sum, OCB subsumes those actions generally desired by organization leaders but often not easy to define operationally with respect to time, place, and amount.

Furthermore, because OCB is seldom easily quantified or monitored, it does not lend itself to incentive systems in the manner of, for example, physical output or sales revenues. This is not to say that abundant OCB on the part of an employee could not or would not be recognized and rewarded in some fashion, such as a later increase in wages or salary, recommendation for promotion, or informal favored treatment by a supervisor. But such rewards would more typically arise from reputational effects that foster holistic impressions among colleagues, and possibly superiors, based on numerous brief, specific episodes in which OCB was exhibited and later recalled by various other parties. In other words, OCB is rewarded on the basis of the extent to which specific behaviors lead to a general impression, which in turn might or might not lead to recompense.

Theoretical Roots

Perhaps the earliest articulation of something evocative of OCB came from Chester Barnard (1938). His basic notion of organization was what he called a 'cooperative system'. He described as 'indispensable' the 'willingness of persons to contribute efforts' to such a system. For him, this 'willingness' captured the more common terms of solidarity, *esprit de corps*, or loyalty. He added that such "willingness" was "…something different from effectiveness, ability, or value of personal contributions…." (Barnard, 1938: p. 84)

Roethlisberger and Dickson (1939), in their pioneering account of the famed Hawthorne studies of the 1920s and 1930s, noted that an 'informal organization' is part of every organization and described it as a 'necessary condition for effective collaboration' (Roethlisberger and Dickson, 1939: p. 562). In other words, much of what makes an organization effective comes from spontaneous and informal exchanges of help, information, and support.

From Katz and Kahn (1966, 1978) came a more precisely articulated version of such collaboration, close to what is now often referred to as OCB. Their analysis posited three forms of behavior that organizations must evoke in order to function: (1) attract and hold people in the system; (2) ensure that members exhibit reliable in-role performance, i.e., do the assigned jobs for which they were recruited and trained and which largely determines base compensation; and (3) encourage spontaneous behavior beyond role or job prescriptions when such behavior is needed. Katz and Kahn elaborated upon this third requirement to say that it includes cooperative interactions with coworkers and colleagues, suggestions for improvement and solving problems, actions protective of the system, and promoting a favorable image of the organization in the external environment.

Current Conceptions

Organ (1977), in an essay on the sparse evidence supporting the 'satisfaction causes productivity' hypothesis, suggested that job satisfaction is likely to be more closely related to OCB – that is, the ongoing stream of what Barnard described as 'willingness to contribute', Roethlisberger and Dickson referred to as 'collaboration', and Katz and Kahn called 'spontaneous' gestures of cooperation, supportiveness, and protection. Organ

reasoned that objective, quantifiable individual productivity is more typically a function of skill, training, technology, and in some instances the explicit contractual incentives attached to objective individual performance. Job satisfaction – as a prototype attitude about the organization, its supportiveness and fairness of treatment – is more likely to affect those behaviors that are not constrained (e.g., by the technology of the job functions in which the individual is involved or the individual's level of relevant technical expertise), nor governed by compelling extrinsic incentives or penalties. Thus, Organ's (1977) interest in OCB arose from the effort to identify the behavioral manifestations of positive attitudes about the job environment. Initially, the focus was on general job satisfaction, but eventually included beliefs about fairness, particularly perceived fairness of pay, benefits and supervisory treatment.

A meta-analysis (Organ and Ryan, 1995) of 28 studies found a weighted average raw correlation of 0.24 between satisfaction and OCB, with a 95% confidence interval of 0.20–0.27. An earlier review of satisfaction and overall job performance (Iaffaldano and Muchinsky, 1985) had reported a 95% confidence interval of 0.13–0.17. Thus, there seemed to be some empirical support for the belief that job satisfaction and related positive attitudes are more likely to have their effect on OCB rather than productivity or overall performance.

What Constitutes OCB?

The original conception of OCB was holistic and not broken down into components. That is, early studies cobbled together a list of behaviors that seemed to share the underlying idea of 'doing something good that is not part of strict job definition and does not qualify for guaranteed reward'. Thus, items were designed that referred to helping others, living up to the spirit as well as the letter of the rules, above average attendance and punctuality, not complaining about temporary inconveniences, interpersonal courtesies that help prevent problems for others, acting to defuse interpersonal conflicts, and taking part in discussions in a constructive, proactive fashion. However, factor analysis of the OCB measure in an early study (Smith et al., 1983) revealed two distinct underlying factors: interpersonal OCB (OCB-I) and organizational OCB (OCB-O). The distinction between the two factors had to do with the *target* of the behavior, that is, whether the focus was a specific individual or individuals vs the more general or collective formal organization. Thus, helping, courtesy, and peacemaking benefit one or more specific individuals, typically part of the work group, while exemplary rule compliance and speaking up in meetings usually are directed toward the larger organizational collective. Across various studies using different measures of OCB, this two-factor structure remains the most reliable and robust, and probably most parsimonious, representation of the larger construct.

Contextual Performance

Borman (1983, 1987) offered a construct that has much in common with OCB on its most substantive or operational level. However, he defined it in terms of the contributions pertaining to the *context* within which core task performance of the group occurs. In particular, his emphasis was on the psychological and social context that supports and facilitates performance by the group as a whole. The specific items in the measure of contextual performance (CP) overlap considerably with OCB measures in referring to such contributions as helping others, making constructive suggestions, providing a model of self-discipline, and loyalty to the organization (e.g., representing it positively to outsiders). Moreover, measures of CP typically break down into two factors – interpersonal facilitation and job dedication – that bear much resemblance to the OCB-I and OCB-O factors in OCB measures.

However, Borman's scheme does not rule out some expectation of reward, nor does it require that the behavior in question lie outside the scope of the job or role definition.

Furthermore, whereas OCB was formulated as a means of defining the types of positive contributions that were more likely to be a function of job satisfaction, Borman was interested in the important work behaviors more likely than individual task performance to reflect the importance of personality dimensions. However, his rationale was much like that of Organ (1988). Personality, like attitude, predicts behavior much better in 'soft' than 'hard' situations. Soft situations are those in which no particular behavior is forced (by law, custom, compelling rewards or punishments), i.e., where there is discretion. Borman, not discouraged by the disappointing results in trying to link personality to job performance or productivity, reasoned that productivity is constrained by skill differences, technologies, and the salience of compelling rewards or sanctions. He viewed CP as the domain in which differences in personality would have their major effects.

Some evidence does support the proposition that certain personality dimensions more consistently predict CP than hard measures of productivity. The evidence comes mainly from studies using the 'Big Five' conceptions of personality (McCrae and Costa, 1987): agreeableness, conscientiousness, extraversion, neuroticism, and openness to experience. In particular, the dimension 'agreeableness' has some empirical connection to interpersonal facilitation, while 'conscientiousness' more reliably predicts task dedication.

On the other hand, Organ and Ryan's (1995) meta-analysis found less support for a relationship between personality dimensions and either holistic OCB or particular dimensions of it, particularly when looking only at studies that used something other than a self-report of OCB (i.e., a supervisory or coworker rating rather than the respondent's own self-rating). The only personality–OCB connection that stood up to conventional standards of statistical reliability was the correlation between the Big Five's measure of conscientiousness and OCB-O (more impersonal OCB).

Still, it seems intuitively compelling that some dispositional factors would logically pertain to most forms of OCB. Perhaps a better test of this belief would take the form of observer ratings of subject personality and a supervisory rating of OCB. Mount et al. (1994) used such a design, with customer, coworker, and supervisor as procedurally independent ratings of personality, CP, and in-role performance. When different variables were matched up with different sources, the correlations between personality and CP were noticeably stronger than were typically found in prior research.

OCB and Organizational Effectiveness

The presumption at the outset of work on OCB was that the sorts of behaviors defined by the concept operationalizing it would have something to do with how well organizations function. The actual items that comprise measures of OCB sound much like what we would assume to be the profile of a healthy, efficient enterprise. Indeed, even to suggest that such behaviors make organizations function better could be considered a truism.

However, the tautology resemblance is less troubling if we at least surmise *why or how* such behaviors benefit the larger collective. Organ et al. (2006) have articulated some reasoning along such lines. First, some forms of OCB represent *added resources*. For example, someone who takes the time to help a new hire learn some operations is providing additional inputs to those not otherwise forthcoming in the mix of organizational resources. Second, OCB serves to economize on the use of resources already at the organization's disposal. This might be illustrated by the staff member who steps into the breach to resolve conflict or prevent conflict from taking its toll on the stamina of colleagues or draining the time and energy of managers who have to become referees. Similarly, timely suggestions in the appropriate tone and forum often provide the means of reducing waste of materials, money, and time. Finally, as we will see below, a transactional analysis of organizations (Williamson, 1975) argues that the very reason for the existence of organizations is that they minimize the extent and cost of transactions needed to negotiate or contract for the inputs they require. Absent OCB, much effort would be diverted into ascertaining what benefits would derive from each and every efficacious input.

Ultimately, though, the tautology problem can be dealt with only by empirical research that takes OCB to the macro level. We must show that in organizations or portions of organizations where more OCB occurs, the more effective they are judged to be in terms of some specified criteria.

However, doing this is not an easy task. Such research requires a group of organizations, a group large enough to have at least minimally adequate statistical power. Furthermore, each organizational unit within that group must lend itself to at least one objective criterion of effectiveness, preferably more such criteria.

One approach to dealing with this problem takes the form of studying small but semiautonomous units of chain-type organizations, as are found in retail firms and restaurants. Walz and Niehoff (2000) used this strategy with a chain of 30 limited-menu restaurants and found significant positive relationships between the OCB dimension of helping behavior and the measured criteria of operating efficiency, customer satisfaction, and an index of quality of performance. Koys (2001) followed a similar research strategy, but with a longitudinal design, in order to determine if earlier OCB predicted later effectiveness better than the converse (prior effectiveness predicting subsequent OCB). The results favored an inference in which OCB was the prior predictor or cause, as opposed to representing the effect of prior effectiveness.

Subsequently, Podsakoff et al. (2009) conducted a meta-analysis of 38 studies that investigated OCB and organization-level outcomes, with varied criteria, such as productivity, reduced costs, and customer satisfaction. The estimated population correlations between unit level OCB and 'positive' criteria ranged from 0.15 (for unit profitability) to 0.47 for subjective or reputational measures of effectiveness. For 'negative' criteria (staff turnover and operating costs), the correlations were negative, i.e., higher unit level OCB was associated with lower turnover and costs. The authors also noted that time-lagged correlations between earlier OCB and later outcomes were generally stronger than those associated with same-time measures.

OCB and Individual Outcomes

We noted earlier that OCB was conceived as discretionary behavioral contributions. 'Discretionary' implies the absence of any guaranteed or contractual reward for such behaviors, as well as the absence of any expected sanction for not displaying those behaviors. Moreover, while some such behaviors might fit within a general statement of certain expectations in a job description – such as 'cooperative', 'courteous', 'self-disciplined', or 'reliable' – seldom are such expectations defined in operational terms. On the other hand, an organization often does cite behaviors, such as deliberately falsifying documents, insubordination, physical or verbal abuse of others, that are subject to punishment or discharge. Positive actions that might be specified as qualifying for remuneration could include productivity (especially in sales jobs but also in others that lend themselves to objectively quantifiable individual output), attendance, and remaining in the organization in good standing. In between those extremes are many small positive or negative behaviors that can easily escape notice, are not quantifiable, and ultimately are matters of reputation within the work group, from customer feedback, or sometimes noted by officials in other departments. Some people practice more, others less, of these behaviors. Reputations based on such 'microcontributions' usually lag behind their rendering by days, weeks, months, or more.

What we do find, however, is that reputations for doing these things do often meet with positive outcomes. The Podsakoff et al. (2009) meta-analysis cited above found abundant and conclusive evidence that OCB does often redound to the individual's benefit. Overall, characteristic OCB is associated with more positive performance evaluations (by peers as well as supervisors and higher level management), recommendations for promotion, and in general with more favorable treatment in terms of the social exchange process between individuals and their supervisors.

OCB and the Cultural Matrix

Early research on OCB was almost wholly confined to USA settings. Therefore, the theory and research were predicated on the thinking about how OCB would be manifested in American organizations.

In recent years, empirical research in industrial and organizational psychology in other countries has increased significantly, and by the 1990s scholars had raised questions about the degree to which OCB research in the USA, or even the very

idea of what OCB meant, could be generalized to other cultures. For example, the US is generally agreed to exhibit a quite individualistic culture and, not surprisingly, OCB is regarded as something an individual chooses to contribute or not contribute. In more collectivist cultures, this concept might not fit as well; rather, there are likely to be strong and shared norms about work behavior, such that members should do certain things and not do others, regardless of what some individuals might prefer to do or not do. In cultures characterized by well-defined levels of power, doing what is 'discretionary', i.e., beyond one's instructions, carries some risk, as it might be resented by the boss.

A study by Farh et al. (2004) gave Chinese respondents a general definition of OCB and asked them to describe examples from their experiences and observations at work, producing 756 distinct examples that they thought exemplified OCB. Content analysis of these produced 10 distinct dimensions. One of those was 'helping', at first glance similar to what obtains in US samples. However, the Chinese interpretation of 'helping' as OCB went well beyond the workplace. It was considered OCB if someone helped the family member of a coworker, or helped with work on a colleague's house. 'Harmony' was a separate brand of OCB, indicating a major ongoing concern with avoiding conflict in the workplace. In state-owned firms, OCB included a factor that pertained to volunteer work in the community. Another factor that emerged was self-training – on one's own, learning more skills that could be applied at work.

If the conceptual structure of OCB varies across national or regional cultures, one might well ask whether different industries and occupations – which plausibly generate differences in organizational culture – might also have different conceptions about OCB. Health care organizations, research laboratories, advertising, correctional facilities, arts organizations are just few that come readily to mind, with varied ideas as to what best fits important categories of OCB. Thus, one daunting challenge in the community of OCB theorists and researchers is to figure out what level of breadth vs specificity will prove most efficacious in future work in this area.

Is There a 'Dark Side' to OCB?

While we have reported here a fair amount of evidence that OCB at the macro or collective level does predict reasonably well a number of criteria of organizational effectiveness and often generates positive outcomes for members who exhibit more of it, some concern has also been expressed about the 'down side' of OCB. One such concern is what it does to the health, home life, and 'burnout' of those who expend enormous amounts of time and energy doing things that go beyond what the job requires. Others worry about the toll that OCB might take on some individuals' performance of their major job-related tasks. Some have suggested that a potentially large portion of OCB is performed as part of a calculated strategy of ingratiating supervisors and higher levels of management (Bolino, 1999). Others have suggested that some forms of OCB might be unethical or even illegal – i.e., doing something regarded as 'good' for the organization but 'wrong' in other, broader respects, possibly unethical or even illegal.

No doubt all of us could summon from memory some people who seriously overloaded themselves with OCB, possibly leading to health breakdowns; individuals who involved themselves so much in doing extraneous tasks that they neglected their core formal responsibilities; quite a few that seemed to curry favor with important people by appearing always to go the extra mile.

In answer to some of these concerns, some would point out that meta-analyses indicate that OCB not only does not negatively correlate with of core job performance, but is either essentially unrelated to in-role productivity (when different sources are used for the two types of contributions) or substantially positive (when the same source rates both factors; Organ et al., 2006). Furthermore, a means of discerning whether and to what extent OCB arises from an attempt to manage impressions of others, is to use multiple sources of rating. Peers and lower-level members would not likely be misled by apparent OCB that is exhibited only for the benefit of the boss.

Furthermore, to have concern about OCB because it might represent a tool of ingratiation seems to suggest that OCB should have the right motive, otherwise it is a negative form of behavior. Such a stance would denigrate a wealthy donor's large contribution to a charitable organization if the motivation – or even part of the motivation – were to involve building a good reputation in the community. One is hard-pressed to argue that the dollars in the donation do any less good for the recipient. In any event, OCB is not defined by its underlying motive. Arguably a complex of varied motives underlies much behavior at work. To date, the most straightforward conjecture about a dominant motive for OCB comes from social exchange theory (Blau, 1964), i.e., people will try to reciprocate good treatment by someone with comparably treatment to that person – or organization. If the individual has reason to perceive the organization as a microcosm of 'a just world', one in which the long-run dynamic is toward fairness, then that individual will not stint in discretionary contributions.

As for the concern that excessive OCB, piled upon diligent care to perform one's core job well, might cause health breakdowns or domestic problems, we have no data at hand to document the frequency or severity of this potential problem. We find it plausible that such cases exist. But we find equally plausible that some people have overextended themselves within the core functions of the job, regardless of how much or how little OCB added to it.

Current Trends and Future Tasks

A major trend in recent years is the status of OCB as a criterion or dependent variable in organizational research. Research designs that once studied such factors as pay systems, leader behavior, organization design, and personality, as they related to job satisfaction and productivity, now often include a measure of OCB or CP. The range of industries and organizational types in which is studied continues to increase. Also, there is anecdotal evidence that some management practitioners now attempt to give more weight to OCB in overall evaluations of employees and that they now search actively for some means of predicting it in the personnel selection system.

However, one would hope that organizational psychologists will continue to work toward more conceptual clarity in the OCB concept, the optimum level of generality vs specificity, and measurement tools that yield the best return in value-added decision-making for the cost of including it in the management information system. There must also be some reckoning with what might the limits of application of OCB concepts, i.e., contexts in which it informs neither theory nor practice.

See also: Cross-Cultural Training in Industrial and Organizational Psychology; Indigenous Management Styles; Job Satisfaction; Leader–Member Exchange Theory; Organizational Climate in the Work Setting; Organizational Commitment; Organizational Justice; Pay, Compensation, and Performance, Psychology of; Psychological Contracts; Teamwork and Team Performance Measurement; Work Motivation.

Bibliography

Barnard, C.I., 1938. The Functions of the Executive. Harvard University Press, Cambridge, MA.

Blau, P., 1964. Exchange and Power in Social Life. Wiley, New York.

Bolino, M.C., 1999. Citizenship and impression management: good soldiers or good actors? Academy of Management Review 24, 82–98.

Borman, W.C., 1983. Implications of personality theory and research for the rating of work performance in organizations. In: Landy, F., Zedeck, S., Cleveland, J. (Eds.), Performance Measurement and Theory. Lawrence Erlbaum, Hillsdale, NJ.

Borman, W.C., 1987. Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: explorations in an Army officer sample. Organizational Behavior and Human Decision Processes 40, 307–322.

Farh, J.L., Zhong, C.B., Organ, D.W., 2004. Organizational citizenship in the people's republic of China. Organization Science 15, 241–253.

Iaffaldano, M.T., Muchinsky, P.M., 1985. Job satisfaction and job performance: a meta-analysis. Psychological Bulletin 97, 251–273.

Katz, D., Kahn, R.L., 1966,1978. The Social Psychology of Organizations. Wiley, New York.

Koys, D.J., 2001. The effects of employee satisfaction, organizational citizenship behavior, and turnover on organizational effectiveness: a unit-level, longitudinal study. Personnel Psychology 54 (1), 101–114.

McCrae, R.R., Costa Jr., P.T., 1987. Validation of the five-factor model of personality across instruments and observers. Journal of Personality and Social Psychology 52, 81–90.

Mount, M.K., Barrick, M.R., Strauss, J.P., 1994. Validity of observer ratings of the big five personality factors. Journal of Applied Psychology 79 (4), 272–280.

Organ, D.W., 1977. A reappraisal and reinterpretation of the satisfaction-causes-performance hypothesis. Academy of Management Review 2, 46–53.

Organ, D.W., 1988. Organizational Citizenship Behavior: The Good Soldier Syndrome. Lexington Books, Lexington, MA.

Organ, D.W., Ryan, K., 1995. A meta-analytic review of attitudinal and dispositional predictors of organizational citizenship behavior. Personnel Psychology 48 (4), 775–802.

Organ, D.W., Podsakoff, P.M., MacKenzie, S.B., 2006. Organizational Citizenship Behavior: Its Nature, Antecedents, and Consequences. Sage Publications, Thousand Oaks, CA.

Podsakoff, N.P., Whiting, S.W., Podsakoff, P.M., Blume, B.D., 2009. Individual and organizational-level consequences of organizational citizenship behaviors: a meta-analysis. Journal of Applied Psychology 94, 122–141.

Roethlisberger, F.J., Dickson, W.J., 1939. Management and the Worker. Harvard University Press, Cambridge, MA.

Smith, C.A., Organ, D.W., Near, J.P., 1983. Organizational citizenship behavior: its nature and antecedents. Journal of Applied Psychology 68, 653–663.

Walz, S.M., Niehoff, B.P., 2000. Organizational citizenship behaviors: their relationship to organizational effectiveness. Journal of Hospitality and Tourism Research 24, 301–319.

Williamson, O.E., 1975. Markets and Hierarchies. Free Press, New York.

Organizational Climate

Michael A West, Lancaster University, Lancaster, UK
Joanne Lyubovnikova, Aston University, Birmingham, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M.A. West, volume 16, pp. 10923–10926, © 2001, Elsevier Ltd.

Abstract

This article examines the concept of organizational climate and its metaphorical derivation. It begins by defining what is meant by the construct and the specific conceptual challenges that researchers face in trying to agree upon a single definition. Different statistical functions of organizational climate scores which have recently emerged in the literature are outlined using examples from relevant research findings. Empirical research into global and facet-specific organizational climate is then described. Given the plethora of facet-specific climate studies which have emerged in the literature over the past decade, the article concludes with a call for a more unified theoretical and methodological approach when conducting future research on organizational climate.

Defining Organizational Climate

For many years, management scholars have grappled with the challenge of understanding how organizational context impacts and shapes the attitudes and behavior of organizational members. Much of this endeavor has centered on the construct of organizational climate. Climate has been variously described as "the way things are around here" (Schneider, 1990) to "a molar concept reflecting the content and strengths of the prevalent norms, attitudes, behaviors, and feelings of the members of the social system which can be operationally measured through the perceptions of system members, or observational means" (Payne, 1990: p. 156). The *Oxford English Dictionary* defines meteorological climate as "the prevailing atmospheric phenomena and conditions of temperature, humidity, wind, etc., of a country or region." When applied to communities, climate is "the mental, moral, etc., environment prevailing in a body of people in respect of opinion, some aspect of life, etc." Thus, just as we describe climates in terms of sunshine, prevailing winds, levels of rainfall, etc., organizations are described in terms of dimensions such as leadership style (autocratic versus democratic), concern with the well-being of employees, emphasis on productivity, formalization or bureaucracy, innovation versus traditionalism, and commitment to quality in service or production.

Although there is a general consensus around what organizational climate means, trying to agree upon a specific definition in the literature remains a challenging task, once described as like trying to "nail Jell-O to the wall" (Schneider, 1990: p. 1). Questions which create difficulties for researchers include: Are climate phenomenon perceptual in nature, or do their capture objective attributes of an organization? How does the concept organizational climate differ from organization culture? And at which level of analysis should organization climate be measured? This has led copious climate definitions in the literature. However, for the purposes of this article, we adopt the widely accepted conceptualization proposed by Schneider and Reichers (1983), which defines organization climate as a set of shared perceptions about the practices, procedures, and policies that an organization expects, supports, and rewards. Organizational climate captures the aggregation of individual psychological meanings of their work environment and is therefore a collective perceptual phenomenon (Kuenzi and Schminke, 2009). Research on organizational climate explores how individual perceptions regarding their workplace impact both individual outcomes (such as job satisfaction, individual performance, and organizational citizenship behavior (OCB)) both group and organizational outcomes (such as team performance and customer satisfaction). As shall be examined later in this article, empirical studies are therefore increasingly conducted at multiple levels of analysis to understand more about how individuals, groups, and organizations interact.

Another continuing debate in the literature concerns the demarcation of organizational climate from organizational culture. Schneider and Reichers's (1983) definition establishes that climate, although related, is distinct from culture. Indeed, both of these concepts are root metaphors, which are very different in orientation. Organizational culture is taken to refer to the underlying patterns of interpretations of formal practices, such as pay levels, hierarchy, and job descriptions, and informal practices such as norms, espoused values, rituals, organizational stories, jargon, humor, and physical environment. Conversely, researchers generally agree that climate is a measure of the surface manifestations of culture and is not entirely distinct from culture. Most argue that culture, due to its deep and often hidden embeddedness in organizations, can only be measured by qualitative methodologies, whereas climate as a more superficial characteristic of organizations can be assessed using quantitative questionnaire measures. Furthermore, whereas culture researchers are concerned with the manifestation of underlying assumptions which create shared values, climate researchers are more interested in the ways in which organizational members attend to these shared values (Kuenzi and Schminke, 2009). However, others would argue that these distinctions are largely artificial and it is true that both concepts of climate and culture are the subject of

continuing debate because of their lack of theoretical and conceptual underpinnings.

In defining the construct of organizational climate, it is also important to distinguish between global climate and facet-specific climate. Indeed, over the years, there has been much uncertainty and debate about whether climate should be conceived of as a multidimensional concept or a unitary concept. Initially, climate researchers focused on establishing a global conceptualization of organizational climate which could account for the total influence of situational variables in organizations, and how these impact individual attitudes and behavior. For example, Campbell et al. (1970) argued for a small number of common dimensions that constitute climate, including individual autonomy at work, the degree of structure imposed upon people's work roles, reward orientation (positive or negative), and warmth and support. Similarly, Jones and James (1979) proposed four dimensions of climate perceptions: job challenge and autonomy; role stress and lack of harmony; leadership facilitation and support; and work-group cooperation, friendliness, and warmth. However, the focus on climate as a global phenomenon has been riddled with conceptual and methodological problems. First, given the abundance of global climate definitions in the literature, researchers have been unable to reach a consensus on a single definition, leading to conceptual imprecision and ambiguity surrounding measurement. Existing research focused on global climate has also been criticized for lacking rigorous theoretical rationale.

These conceptual and methodological challenges have fueled a growing interest in facet-specific organizational climates over the past decade. For example, Clarke (2006) looked at a climate for *safety*, Anderson and West (1998) at a climate for *innovation*, and Martin and Cullen (2006) at a climate for *ethics*. Indeed, Schneider and Reichers (1983) initially suggested that work settings actually have many climates and argued for the study of facet-specific climates rather than global organizational climate. To this way of thinking, the choice of climate dimensions should be closely linked to the criteria of interests or climates 'for something.' This should reflect the specific strategic outcomes of the organization at hand, for example, a climate for justice, ethics, diversity, innovation, customer service, or safety. Because these each represent distinct aspects of an organizational environment, facet-specific climates can be seen as independent and may therefore be present simultaneously at any given time in an organization.

Although facet-specific climate research has helped to develop our understanding about how work climates impact individual and organizational performances, this approach has also led to substantial fragmentation in the field, with research on particular facets becoming splintered and published under different topical areas. Some suggest that this has led to poor integration and unification of the most recent research on organizational climate (Kuenzi and Schminke, 2009). Others also continue to argue that organizational climate is most usefully applied as a multidimensional construct and that some dimensions are generic and applicable to most organizations. Of course, there will be some which are unique or at least applicable to only a small range of organizations (e.g., climates for secrecy in the defense industry or for caring in health care organizations). One orientation to the mapping of climate that has proved powerful is the competing values model of organizations (Quinn and Rohrbaugh, 1981), which distinguishes among human relations, rational goal, open systems, and internal process constellations of organizational values.

The Competing Values Model

The competing values approach incorporates a range of fundamental dimensions of values into a single model. It calls attention to how opposing values exist in organizations and how organizations have different mixtures of values that influence their goals and how they go about achieving those goals. These values will most fundamentally manifest themselves in whether the organizational orientation varies more toward flexibility or control, and whether the focus of the organization varies more toward the external or internal environment. A major theoretical strength of this model is its derivation from four orientations to the study of organizational effectiveness reflecting long traditions in work and organizational psychology. Thus the rational goal approach (external focus, but with tight control within the organization) reflects a rational economic model of organizational functioning, in which the emphasis is upon productivity and goal achievement. The open systems model (external focus and flexible relationships with the environment) emphasizes the interaction and adaptation of the organization in its environment, with managers seeking resources and innovating in response to environmental (or market) demands. The internal process approach reflects a Tayloristic concern with formalization and internal control of the system in order that resources are efficiently used. Finally, the human relations approach reflects the tradition derived from socio-technical theory, emphasizing the well-being, growth, and commitment of the community of workers within the organization. By combining these orientations into one model, Quinn and Rohrbaugh aimed to provide a broad conceptual map of the domains of organizational theory over the twentieth century. Such a map is useful in identifying the required topography of a climate measure applicable to a wide range of organizations. The model is also useful in reflecting the means for implementing those values in terms of managerial practices and the ends or outcomes which are emphasized or which compete in each domain. The model does not propose that organizations will locate predominantly in one quadrant, but, reflecting the rich mix of competing views and perspectives in organizations, proposes that organizations will be active and give emphasis to each domain but with differing strengths.

Measuring Organizational Climate

Having established what we mean by 'organizational climate,' in this section, we shall explore the different statistical functions of organizational climate scores. Indeed, research on organizational climate continues to be hampered by a particular methodological debate. It concerns the practice of

aggregating individual employees' scores on climate dimensions and taking the mean as a representation of organizational climate. This has been referred to as a composition theory for climate (James et al., 2008) which proposes that the aggregation of individual climate perceptions (i.e., psychological climate) provides a powerful measure of organizational climate. Given that organizational climate is a collective phenomenon representing shared psychological meanings, statistical aggregation to this higher level construct requires perceptual agreement between individual climate scores. Therefore, the validity of aggregate climate used to describe organizations depends upon the demonstration of agreement between individual members and their perceptions. In such circumstances, the use of aggregated individual data to measure organizational climate is appropriate.

Nevertheless, the main methodological problem with this approach, a demonstration of shared conceptions for aggregate climates, has provoked much debate in the research literature. Indeed, writers have not specified what is an appropriate level of agreement. Some suggest that dichotomous questionnaire items should be used to measure climate and that the only items to be treated as descriptive of organizational climate should be those in which the frequency of endorsement is not significantly different from 0 to 100%. Others have suggested that 66% was reasonable, but few researchers have followed either prescription. The validity of aggregating climate perceptions to the organizational level is further called into question by the demonstration that there are significant differences between members of different organizational subgroups in their perceptions of climate. Such differences have been found between groups within the same organization, based on factors such as hierarchical level, departments, divisions, regions, and work groups.

One approach to overcoming the aggregation issue is to examine the level of inter-rater agreement between individuals within an organization, in relation to their perceptions of climate. A common way of achieving this is through the use of intraclass correlation coefficients. Such coefficients are based on a one-way analysis of variance and assesses the ratio of variation within organizations to variation among organizations. However, this coefficient has been criticized as inadequate which led to the development of the $R_{wg(j)}$ index of multiple item scales developed by James et al. (1993). This is a technique for assessing agreement among the judgments made on multiple item scales. Another salient approach is the average deviation (AD) index (Burke et al., 1999). Although there is continuing debate about the advantages and disadvantages of such orientations, these measures for justifying aggregation continue to be widely used. Indeed, it remains likely that organizational climate measures will continue to be used, despite there being no simple or widely agreed criteria for aggregation. Part of the problem is that the level of agreement within organizations will depend partly on the measures employed, the context, and the domains tapped by the measures. However, simply using the mean may mask profound differences between organizations and the variation of responses given by employees. Indeed, research within organizations has shown clear differences between hierarchical levels and departments in ratings of climate, suggesting a number of subclimates within each organization. Overall, when the level of analysis changes from the individual to the organization, the validity of the data is uncertain.

More recently, researchers have begun to explore new relationships between organizational climate and its antecedents and outcomes by using alterative statistical functions of climate scores. For example, Schneider et al. (2002) proposed the related construct of '*climate strength*' which captures within-group variability of perceptions of organizational climate. The researchers found that climate strength moderated the relationship between managerial practices (one of four dimensions of climate) and customer experience. In another study of 197 work units, Gonzalez-Romá et al. (2002) also examined the moderating influence of three climate strength facets (support, goals orientation, and innovation).

Alternatively, in a study of emergency planning committees, Lindell and Brandt (2000) used the constructs of '*climate quality*' (mean climate perceptions) and '*climate consensus*' (level of perceptual agreement on climate) as mediating variables in their analysis. Their results showed that in comparison to climate consensus, climate quality was more strongly related to both antecedent and outcomes variables. More recent research is therefore moving beyond the composition theory for climate, exploring the influence of more sophisticated statistical functions of climate scores.

Research on Antecedents and Outcomes

So what has research told us to date about the importance of organizational climate in predicting individual, group, and organizational outcomes? What factors influence the development of organizational climate in organizations? In this section, we shall begin by summarizing research evidence on the consequences of organizational climate for organizations and their members, before considering several important antecedents.

One of the earliest uses of the concept of climate was in research conducted by Lewin et al. (1939), who examined the effect of the leadership style of boys upon their work groups. Leadership style was categorized as autocratic, democratic, or laissez-faire and the work-group climates were described, respectively, as hostile, unproductive, and unenjoyable; busy, cooperative, productive, and enjoyable; and anarchic, unproductive, not cohesive, and unsatisfying. More recent research has concentrated on climate at work group, department, or organizational levels. As had been discussed, these supra-individual climates have been operationally constructed by aggregating individual perceptions of psychological climate to the appropriate level and using the mean to represent the climate at that level of aggregation. These organizational climate measures are now widely used in public and private sector organizations to determine the prevailing climate, often being called employee attitude surveys or employee opinion surveys.

One of the reasons why organizational climate surveys have become so widely used is that climate has been shown to predict organizational productivity and profitability, customer attitudes, innovation, team performance, OCBs, safety, and job

attitudes. For example, Patterson et al. (1997) found that an organizational climate oriented toward good human relations predicted 29% of the variation between companies in productivity and profitability in the UK manufacturing sector. In a study of 137 clusters of bank branches, Gelade and Ivery (2003) also found that global organizational climate was positively related with sales against targets, clerical accuracy, customer satisfaction, staff retention, and overall performance. Furthermore, Glisson and James (2002) found that positive climates predicted more favorable work attitudes in a study of 238 managers. This study also found evidence that organizational climate is distinct from culture. Overall, because of these strong associations between climate and performance, both researchers and practitioners remain intensely interested in the construct (Kuenzi and Schminke, 2009).

There has also been an increasing amount of empirical attention paid to facet-specific climates over the past decade. For example, Hofmann et al. (2003) explored the moderating effect of safety climate on the association between leadership (specifically leader–member exchange (LMX) and safety citizenship roles. Using multilevel analysis, group-level safety climate was found to moderate the LMX–safety citizenship roles relationship. In another facet-specific study, King et al. (2007) analyzed data from over 22 000 members of 131 health care organizations in the United Kingdom and found that climate for innovation was positively related to organizational performance and also moderated the relationship between work demands and organizational performance. Furthermore, Chen et al. (2007) adopted a group-level approach to explore the effects of team empowerment climate on an individual's sense of empowerment, whereas Maynard et al. (2007) collected data from 637 customer service engineers to examine how resistance to empowerment climate was related to team performance and customer satisfaction. These are just a few of many published studies which have examined outcomes of facet-specific climates over recent years.

But where do climates come from? What influences their emergence and development? Researchers have now started to pay more attention to the individual-, group-, and organization-level predictors of climate, mainly through quantitative cross-sectional field studies. However, relative to the amount of research examining the consequences of global and facet-specific climates, far fewer studies have so far examined climate antecedents. One antecedent that has received the most interest is that of leadership. For example, transformational leadership was shown to be a significant antecedent of climate for involvement in a study by Richardson and Vandenberg (2005). Similarly, in a study of over 2000 military soldiers, Zohar and Luria (2004) found transformational leadership to positively predict safety climate. In another leadership study, Ehrhart (2004) found that servant leadership was positively related to procedural justice climate, which itself was a mediator between servant leadership and OCB. Leader personality characteristics have also been explored as antecedents of facet-specific climate. In a study of 3445 employees in 383 US department stores, Mayer et al. (2007) found that leader agreeableness and conscientiousness were positively related to procedural justice climate, whereas neuroticism was negatively related, and extraversion unrelated. Finally, in a study of 212 bank employees, Walumbwa et al. (2008) found that contingent reward transactional leadership positively predicted procedural justice climate and procedural justice climate strength, which in turn, were related to satisfaction with supervisor, OCB, and organizational commitment. Research to date therefore suggests that leadership is a key mechanism through which organizational climate can be developed and nurtured to have a positive impact on outcomes.

However, an increasing amount of research has also looked beyond leadership to consider other important antecedents of facet-specific climates. In keeping with the theme of procedural justice, Tangirala and Ramanujam (2008) found that employee silence was a negative antecedent of procedural justice climate, which itself moderated the relationship between employee silence and work group identification, as well as professional commitment. Naumann and Bennett (2000) examined whether demographic similarity, group cohesion, and supervisor viability were antecedents of climate strength in a study of 323 bank employees. They found that procedural justice climate was indeed related to cohesion and supervisor viability, but not to demographic similarity, suggesting that shared climates can emerge even in highly diverse organizations. Finally, Pirola-Merlo et al. (2002) conducted research with 53 research and development teams and found that obstacles had a negative impact on team climate, which in turn, was positively related to team performance. Furthermore, team climate mediated the relationship between leadership and team performance. This is by no means an exhaustive list of research on climate antecedents, but serves to demonstrate the growing level of interest and diversity in the type of research currently being conducted in the field. We anticipate that this will only continue to grow as organizations are required to do more and more with less and less, and therefore require further understanding about how exactly contextual factors impact and shape the working lives of organizational members.

Conclusion

Theory and research on the phenomenon of organizational climate is continuing to develop and gain momentum in the organizational sciences. In a recent qualitative review, Kuenzi and Schminke (2009) identified 241 published empirical studies in the area. Indeed, because organizational climate is a phenomenon that researchers experience in practice, in working in different organizations, and because the measures of organizational climate relate to important outcome measures, it is likely that the concept will continue to be used. Progress has been made on how to best capture and operationalize the climate construct. However, the increasing focus on facet-specific climates means that the literature has become splintered and somewhat fuzzy in recent years. Future research endeavor must therefore be designed and conducted through a more unified theoretical and methodological lens to enable organizational climate scholars to better synthesize their research findings and move the field forward in a coherent and coalesced way.

See also: Organizational Climate in the Work Setting; Organizations: Authority and Power.

Bibliography

Anderson, N., West, M., 1998. Measuring climate for work group innovation: development and validation of the team climate inventory. Journal of Organizational Behavior 19, 235–258.

Burke, M.J., Finkelstein, L.M., Dusig, M.S., 1999. On average deviation indices for estimating interrater agreement. Organizational Research Methods 2, 49–68.

Campbell, J.P., Dunnette, M.D., Lawler, E.F., Weick, K.E., 1970. Managerial Behavior, Performance and Effectiveness. McGraw-Hill, New York.

Chen, Z., Lam, W., Zhong, J.A., 2007. Leader–member exchange and member performance: a new look at individual-level negative feedback-seeking behavior and team-level empowerment climate. Journal of Applied Psychology 92, 202–212.

Clarke, S., 2006. The relationship between safety climate and safety performance: a meta-analytic review. Journal of Occupational Health Psychology 11, 315–327.

Ehrhart, M., 2004. Leadership and procedural justice climate as antecedents of unit-level organizational citizenship behavior. Personnel Psychology 57, 61–94.

Gelade, G.A., Ivery, M., 2003. The impact of human resource management and work climate on organizational performance. Personnel Psychology 56, 383–404.

Glisson, C., James, L., 2002. The cross-level effects of culture and climate in human service teams. Journal of Organizational Behavior 23, 767–794.

Gonzalez-Romá, V., Peiró, J.M., Tordera, N., 2002. An examination of the antecedents and moderator influences of climate strength. Journal of Applied Psychology 87 (3), 465–473.

Hofmann, D.A., Morgeson, F.P., Gerras, S.J., 2003. Climate as a moderator of the relationship between leader–member exchange and content specific citizenship: safety climate as an exemplar. Journal of Applied Psychology 88, 170–178.

James, L.R., Demarée, R.G., Wolf, G., 1993. r_{wg}: an assessment of within-group interrater agreement. Journal of Applied Psychology 78, 306–309.

James, L.R., Choi, C.C., Ko, C.E., McNeil, P.K., Minton, M.K., Wright, M.A., Kim, K., 2008. Organizational and psychological climate: a review of theory and research. European Journal of Work and Organizational Psychology 17 (1), 5–32.

Jones, A.P., James, L.R., 1979. Psychological climate: dimensions and relationships of individual and aggregated work environment perceptions. Organizational Behavior and Human Performance 23, 201–250.

King, E.B., de Chermont, K., West, M., Dawson, J.F., Hebl, M., 2007. How innovation can alleviate negative consequences of demanding work contexts: the influence of climate for innovation on organizational outcomes. Journal of Occupational and Organizational Psychology 80, 631–645.

Kuenzi, M., Schminke, M., 2009. Assembling fragments into a lens: a review, critique, and proposed research agenda for organizational work climate literature. Journal of Management 35 (3), 634–717.

Lewin, K., Lippitt, R., White, R.K., 1939. Patterns of aggressive behavior in experimentally created social climates. Journal of Social Psychology 10, 271–299.

Lindell, M.K., Brandt, C.J., 2000. Climate quality and climate consensus as mediators of the relationship between organizational antecedents and outcomes. Journal of Applied Psychology 85 (3), 331–348.

Martin, K., Cullen, J., 2006. Continuities and extensions of ethical climate theory: a meta-analytic review. Journal of Business Ethics 69, 175–194.

Mayer, D., Nishii, L., Schneider, B., Goldstein, H., 2007. The precursors and products of justice climates: group leader antecedents and employee attitudinal consequences. Personnel Psychology 60, 929–963.

Maynard, M.T., Mathieu, J., Marsh, W., Ruddy, T., 2007. A multilevel investigation of the influences of employees' resistance to empowerment. Human Performance 20, 147–171.

Naumann, S., Bennett, N., 2000. A case for procedural justice climate: development and test of a multilevel model. Academy of Management Journal 43, 881–889.

Patterson, M.G., West, M.A., Lawthom, R.L., Nickell, S., 1997. People Management and Business Performance. Institute for Personnel and Development, London.

Payne, R., 1990. Madness in our method. Journal of Organizational Behavior 11, 77–80.

Pirola-Merlo, A., Härtel, C., Mann, L., Hirst, G., 2002. How leaders influence the impact of affective events on team climate and performance in R&D teams. The Leadership Quarterly 13, 561–581.

Quinn, R.E., Rohrbaugh, J., 1981. A competing values approach to organisational effectiveness. Public Productivity Review 5, 122–140.

Richardson, H.A., Vandenberg, R.J., 2005. Integrating managerial perceptions and transformational leadership into a work-unit level model of employee involvement. Journal of Organizational Behavior 26, 561–589.

Schneider, B., 1990. Organizational Climate and Culture. Jossey-Bass, San Francisco.

Schneider, B., Reichers, A.E., 1983. On the etiology of climates. Personnel Psychology 36, 19–39.

Schneider, B., Salvaggio, A.N., Subirats, M., 2002. Climate strength: a new direction for climate research. Journal of Applied Psychology 87 (2), 220–229.

Tangirala, S., Ramanujam, R., 2008. Employee silence on critical work issues: the cross level effects of procedural justice climate. Personnel Psychology 61 (1), 37–68.

Walumbwa, F.O., Wu, C., Orwa, B., 2008. Contingent reward transactional leadership, work attitudes, and organizational citizenship behavior: the role of procedural justice climate perceptions and strength. The Leadership Quarterly 19, 251–265.

Zohar, D., Luria, G., 2004. Climate as a socio-cognitive construction of supervisory safety practices: scripts as proxy of behavior patterns. Journal of Applied Psychology 89, 322–334.

Organizational Climate in the Work Setting

Mark G Ehrhart, San Diego State University, San Diego, CA, USA
Maribeth Kuenzi, Cox School of Business, Southern Methodist University, Dallas, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Organizational climate research represents one of the primary approaches to understanding work environment in organizations. Employee perceptions of organizational climate affect nearly all aspects of organizational life including employee attitudes and behavior, group processes, and organizational performance. Research on organizational climate gained momentum in the late 1960s and early 1970s, but waned in the 1980s as a number of conceptual and methodological challenges were raised by critics. The resolution of these challenges formed the foundation for contemporary research on climate, which has thrived over the past decade due to its usefulness in understanding workers' experiences and predicting organizational effectiveness.

Researchers and practitioners alike have been increasingly interested in the impact of the organizational context on important outcomes such as employee attitudes, group processes, and firm performance. One of the primary and most fruitful lines of research on these contextual factors has focused on the topic of organizational climate. Organizational climate is defined as "the shared meaning organizational members attach to the events, policies, practices, and procedures they experience and the behaviors they see being rewarded, supported, and expected" (Ehrhart et al., 2014: p. 2). The study of organizational climates is important in that they impact nearly all aspects of organizational life. To provide an overview of this important topic, below is summary of the history of organizational climate, its overlap with and distinctions from organizational culture, and how it has been conceptualized, measured, and studied.

History of Organizational Climate

The history of the organizational climate construct can be traced to research done by Lewin et al. (1939) on the effects of leadership on the 'social climate' that formed in groups of 10-year-old boys. They examined how democratic, autocratic, and laissez-faire leadership influenced the nature of the interactions among the boys. The critical aspects of this research were its focus on the group environment or climate as a whole and the importance of that climate for the group's functioning and performance. Interest in the environment of organizations as experienced by its employees continued to grow in subsequent decades, particularly the study of climate. The interest in organizational climate was aligned with the broader movement among organizational researchers to move beyond the focus on individual differences that had been the typical focus of industrial psychology. Thus, as the fields of organizational psychology (later merged with industrial psychology as industrial/organizational psychology) and organizational behavior began to grow through the 1960s, the concept of organizational climate attracted increasing attention by researchers.

By the late 1960s and early 1970s, research on organizational climate had proliferated, but that growth was quickly followed by a number of critical reviews of the literature. The concerns were many, including definitional issues such as whether climate was an attribute of individuals or organizations, the overlap of climate with other variables including job satisfaction and structure, the mixed findings between climate and performance, and methodological issues such as whether there was adequate consensus in climate perceptions to justify aggregation (Ehrhart et al., 2014). The resolution of these issues formed the foundation for contemporary research on organizational climate to be described in what follows. Before doing so, we first address another relevant issue surrounding the definition of climate: the relationship between organizational climate and organizational culture.

Organizational Climate versus Organizational Culture

The history of organizational climate is intertwined with the related concept of organizational culture. Although organizational climate and organizational culture have similarities, they are distinct constructs. Organizational culture has been defined as "a pattern of shared basic assumptions learned by [an organization] as it solved its problems of external adaptation and internal integration, which has worked well enough to be considered valid and, therefore, to be taught to new members as the correct way to perceive, think, and feel in relation to those problems" (Schein, 2010: p. 18). In early climate research in the 1950s and 1960s, the terms climate and culture were often used interchangeably as the domains of either had not been clearly articulated. A distinct literature on organizational culture emerged in the 1980s, and for many years the two concepts coexisted with little discussion of their relationships or overlap. More recently, however, this issue has received growing attention, including by Denison (1996), Ehrhart et al. (2014), Ostroff et al. (2012), Schneider et al. (2011b), and Zohar and Hofmann (2012).

There are many similarities between the concepts of organizational climate and organizational culture. Ehrhart et al.

(2014) identified several, including their common focus on the macro environment of the organization, their emphasis on the shared experiences of workers, the centrality of meaning that employees' experiences hold for them, the important role of leaders in influencing the climate or culture, and the constructs' influence on organizational effectiveness. Nevertheless, they and others (particularly Denison, 1996) have also identified several critical differences between climate and culture. One is the theoretical tradition of each variable, as organizational climate was primarily developed within the field of psychology and organizational culture was drawn from the field of anthropology. An implication of this difference has been that climate has been more typically studied using quantitative methods whereas culture research has more traditionally used qualitative methods (although there has been growing use of organizational culture surveys over the years, which has served to confound the confusion between the constructs). Climate and culture also differ with regard to their breadth, as culture typically includes a much broader range of variables (such as language, stories, rituals, traditions, behavioral norms, dress, and physical arrangements) and climate is more focused on how employees perceive the organization's policies, practices, and procedures. Similarly, organizational culture includes deep assumptions that may not even be in the conscious awareness of employees; climate clearly exists within the conscious awareness of employees. Because of organizational culture's breadth and depth, it is notably difficult to change. Organizational climate is not easy to change either, but is likely more susceptible to organizational change interventions. Finally, research on organizational climate has developed such that it has a strategic focus. In other words (and as we describe in more detail in later sections), there are many climates within organizations which relate to the strategic outcomes or processes in organizations such as justice, safety, or service (Schneider, 1975). Because culture tends to include the totality of employees' experiences in their organizations, it is more commonly studied as a general concept without a focus.

The complexities surrounding the constructs of organizational climate and organizational culture raise questions about how the two constructs are related to each other. Schneider et al. (2011b) emphasized culture as both forming a foundation for strategic climates (which then predict strategic outcomes) and making the organization an attractive place to work (which results in a strong organizational reputation and high employee retention). Ostroff et al. (2012) similarly placed culture as a foundation for climate, noting that climate focuses on what happens in the organization as perceived by employees, and culture explains why those things happen. Furthermore, they describe the policies, practices, and procedures in the organization as the linking mechanism between culture and climate, such that those practices are a manifestation of cultural assumptions and subsequently become the focus of employees' climate perceptions. Zohar and Hofmann (2012) emphasized organizational climate as an overall gestalt perception based on the pattern of cultural artifacts employees experience in the organization. These perceptions inform understandings of the enacted values in the organization, which when contrasted with the espoused values, provide insight into the deeper layers of culture (i.e., the organization's core values and assumptions). Although each of these perspectives has its own emphasis, they share in common the idea that both organizational climate and organizational culture offer unique insights into organizational functioning and approaches integrating both concepts offer much potential.

Molar Climate versus Focused Climate

As mentioned previously, there were a number of concerns voiced in the 1970s as the climate construct grew in prominence, and two of the criticisms of climate were its overlap with other constructs and its inconsistent relationship with outcomes. Some early efforts to develop measures of climate attempted to capture the full breadth of work environment attributes as experienced by workers, or what has been referred to as the molar climate approach. The problem is that such measures captured many other organizational constructs with their own independent literature, like structure, leadership, stress, and cooperation. In addition, attempts to show that molar climate measures related to effectiveness were inconsistent because the measures lacked focus or a direct link to the criteria of interest. Schneider (1975) argued that to address these problems, climate should be studied with a particular focus or a climate for 'something.' Thus, instead of trying to measure the totality of the work environment, he asserted that the goal should be to measure those aspects of the environment that are most relevant for predicting a specific criterion. In the years following, there was a divergence in the field between research on molar climate and focused climate.

Across research on molar climate, a variety of dimensions within various taxonomies have been proposed. One such taxonomy that sought to organize the molar climate literature up to that point was proposed by Ostroff (1993). She organized climate dimensions into three general categories: affective, cognitive, and instrumental. The affective climate category addresses relationships among employees and includes the dimensions participation, cooperation, warmth, and social rewards. The cognitive category captures how individuals relate to their work, and includes the dimensions of growth, innovation, autonomy, and intrinsic rewards. Finally, the instrumental category includes climate dimensions related to how tasks are performed, such as achievement, hierarchy, structure, and extrinsic rewards. Another approach by Patterson et al. (2005) was developed by linking climate dimensions to the competing values framework (CVF). The CVF results from crossing the dimensions of flexibility versus control and internal versus external orientation, resulting in four approaches to organizational effectiveness: the human relations approach, the internal process approach, the open systems approach, and the rational goal approach (Quinn and Rohrbaugh, 1983). In the development of their Organizational Climate Measure (OCM), Patterson et al. (2005) organized 17 climate dimensions within each of these approaches, such that the human relations model included the dimensions of autonomy, integration, involvement, supervisory support, training, and welfare; the internal process model included the dimensions of formalization and tradition; the open systems model included the dimensions of innovation and flexibility, outward focus, and reflexivity; and the rational goal model

included the dimensions of clarity of organizational goals, efficiency, effort, performance feedback, pressure to produce, and quality. Because of the strong correlations across a number of molar climate dimensions, some have proposed that molar climate approaches capture the extent to which an organization is a positive place to work, or a general climate for well-being (Schneider et al., 2011a).

Although molar climate research has continued to this day, the dominant approach to studying organizational climate has been the focused climate approach. The earliest efforts to study focused climates were on the climate for customer service and climate for safety, and both of those research streams have produced a large body of research. Beyond service and safety, the focused climate approach has been applied to a number of additional areas. A partial listing of these climates would include climate for innovation, justice climate, ethical climate, diversity climate, leadership climate, climate for sexual harassment, and climate for initiative, among many, many others (Kuenzi and Schminke, 2009). The large number of applications of climate has resulted in some researchers proposing ways of categorizing these various focused climates. For instance, Schneider et al. (2011a) distinguished between two types of focused climate. Strategic climates focus on the achievement of the organization's strategic goals that can typically be measured by external criteria, such as service, safety, or innovation. In contrast, process climates focus on the organization's internal processes, and include justice climate or ethical climate. Another organizing scheme for focused climates was proposed by Kuenzi and Schminke (2009). They proposed four general categories of focused climates: climates focused on behavioral guidance (e.g., ethics and justice), climates focused on involvement (e.g., participation, support, and empowerment), climates focused on development (e.g., learning and creativity), and climates focused on core operations (e.g., service and safety).

In a similar way that organizational culture can serve as a foundation for climate, some have proposed that molar climate and process climates may best be conceptualized as serving as a foundation for strategic climates (Ehrhart et al., 2014). Because strategic climates involve those aspects of the work environment that are most relevant for the accomplishment of the relevant strategic goals, they are the more proximal predictor of outcomes. Nevertheless, workers are unlikely to be motivated to pursue the organization's strategic goals unless they are treated well and have the necessary resources and training to be successful. In subsequent sections describing different approaches to studying climate, we will highlight research that supports this conceptualization and that also provides examples of various approaches to studying organizational climate.

Measuring Organizational Climate

In addition to the early concerns voiced by critics over the conceptualization of climate, there were also concerns about its measurement. One of the core problems was confusion over the proper level of analysis for the construct. Although historically the writing and research on climate had emphasized its role as a macro construct functioning at the organizational level, it was typically measured based on individuals' reports. With most of the interest in climate coming from psychological researchers, the tendency early on was to subsequently study it at the individual level rather than aggregating to a higher level, raising questions about the proper level of analysis for the construct. One important resolution of this issue came from James and Jones (1974), who distinguished between climate studied at the unit level (which was labeled as organizational climate) and at the individual level (which was labeled as psychological climate). Although research on both psychological climate and organizational climate continues to this day, recent literature has focused more on organizational climate most likely because of the historical emphasis of climate as a macro construct and the usefulness of organizational climate for understanding important unit-level outcomes. For a recent review of the literature on psychological climate, see James et al. (2008).

Related to the level of analysis issue was another criticism that climate seemed to overlap with job attitudes, particularly job satisfaction. The shift to the unit level of analysis addressed this concern in part, but another aspect of it had to do with the content of climate items. Because some climate measures emphasized individuals' affective responses to their work responses, the unique contribution of climate beyond work attitudes was questioned. The resolution of this concern, as advocated for by Schneider (1975) and Glick (1985), was for climate measures to focus on description rather than evaluation and to use a unit referent for climate items. In other words, if employees are asked to describe the shared experiences of all employees in the workplace, then the overlap with job attitudes like satisfaction is minimized. Furthermore, the unit referent reinforces that climate is a property of organizations or their units, not individuals. Recent research suggests that using a unit referent also increases the validity of the climate construct for predicting performance outcomes (Klein et al., 2001; Wallace et al., in press; Whitman et al., 2012).

Once researchers have decided their focus is on organizational climate as a unit-level construct measured by descriptive items with a unit referent, they still must demonstrate justification for aggregating individual reports to a higher level of analysis. The amount of agreement that is needed, how to statistically test for that agreement, and the number of employees needed from the unit to test for agreement has been the subject of much discussion for several decades, going beyond organizational climate research to include any unit-level constructs that researchers wish to study based on individual-level reports. Although many alternatives exist, the most common within-group agreement statistics found in research in industrial/organizational psychology and management is $r_{WG(j)}$ (James, 1982; James et al., 1984). Within-group agreement statistics are often coupled with intraclass correlations, such as ICC(1) and ICC(2), that combine information on agreement and reliability. More in-depth information on handling these statistical issues can be found in Bliese (2000) and LeBreton and Senter (2008). The take-home message is that researchers should demonstrate that individual-level climate data can be meaningfully aggregated to represent the climate of the unit as a whole.

One final point of clarification on the topic of measurement is that the climate literature suggests that it can operate at

multiple levels of analysis. Thus, there may be an overall climate in an organization as a whole, but there also may be subclimates in departments, stores, or work groups. In some cases the climate at one level may be dominant and studying climate at other levels may not be appropriate. More likely is the case that climate operates at multiple levels simultaneously and that different climate issues are more relevant depending on what level is the focus of the research. An excellent example of such multilevel climate research is Zohar and Luria's (2005) research on safety climate. Among other findings, they demonstrated that the safety climate in a manufacturing plant predicts the safety climate of the work groups in that plant, which subsequently predicts the group's safety behavior.

Research on Organizational Climate

The early popularity of organizational climate waned with the concerns raised by many critics and the growing interest in organizational culture. Nevertheless, since the late 1990s interest in organizational climate has risen dramatically (Kuenzi and Schminke, 2009), and research on the topic has expanded and in numerous directions. Much of the early interest in organizational climate was based on its hypothesized importance for organizational effectiveness; therefore, it is not surprising that one of the major lines of research on the topic has been in demonstrating its relationship with critical organizational outcomes. Although molar climate has been demonstrated to relate to organizational performance (e.g., Patterson et al., 2004), the relationships have generally not been as strong as those found for focused climates as demonstrated by several meta-analyses in recent years. For instance, Hong et al. (2013) showed that service climate was significantly related to service performance and customer satisfaction, in addition to financial performance. In the safety domain, meta-analyses by Christian et al. (2009) and Beus et al. (2010) have demonstrated significant results across studies for a relationship between safety climate and safety outcomes like accidents and injuries. As a final example, Whitman et al. (2012) showed that justice climate was significantly related to unit performance.

In light of the strong evidence for focused climate's relationships with organizational effectiveness, researchers have begun to incorporate more complexity in their research to better understand climate's effects. One issue addressed by this research is how climate has its effects on outcomes (i.e., what are the mediators of this relationship?). For instance, researchers have investigated that service climate's relationship with customer outcomes is mediated through employee behaviors toward customers (Schneider et al., 2005). Research from the safety domain has suggested that motivation also plays a role in the process, such that the climate increases employee motivation, which then is related to their behavior and the subsequent outcomes (Neal and Griffin, 2006). Such research has been helpful in detailing the processes through which organizational climate influences outcomes.

Mediation models have not been limited to studying the mediators of climate's relationship with outcomes. Climate itself has been studied as a mediator, explaining how other influences on the organization and work units influence employees and unit outcomes. From the beginning (e.g., Lewin et al., 1939), climate has served as an explanatory variable to understand how leaders have their influence. In other words, as leaders create policies, procedures, and reward systems in their units, employees experience those systems and form shared gestalt perceptions of what the priorities are in the unit (i.e., climate), which subsequently guides employee behavior and ultimately the unit's outcomes. In the same way that climate research has distinguished between molar and focused approaches, research on leadership's relationship with climate has begun to distinguish between general and focused leadership. Meta-analytic results from Hong et al. (2013) on service climate demonstrated that both general leadership and service leadership had significant relationships with service climate, but the strength of the relationship was stronger for the focused (i.e., service) leadership approach. Ehrhart et al. (2014) explain these findings in terms of the effects of leadership on employee motivation. When leadership is generally positive, employees respond favorably and are more motivated, but also need more information on where to target their motivation, and that is what the focused leadership communicates. In addition to leadership, other variables have been studied as potentially foundations for the development of focused climates. Examples include work facilitation (Schneider et al., 1998) and organizational resources (Salanova et al., 2005) as foundations for service climate, and organizational rewards as a foundation for safety climate (Wallace et al., 2006). This research is in line with the idea that a molar climate for employee well-being (and possibly related process climates) is needed in order for more focused climates to be effective.

In addition to climate studies focusing on mediation, researchers have also studied moderated relationship involving climate. Perhaps most common is to study moderators of climate's relationship with outcomes. The assumption of this research is that climate will have stronger effects in some situations or contexts than in others, so the research is focused on identifying the critical variables that influence this variation. The relevant moderators often will be specific to the focused climate. For instance, service climate has been shown to have stronger effects when levels of customer contact are higher and the product is less tangible (Dietz et al., 2004; Mayer et al., 2009a), and safety climate has been shown to have stronger effects on injuries and errors when complexity was higher (Hofmann and Mark, 2006). There are some moderators, however, that can affect all climates. For instance, the moderator that has perhaps received the most attention in the organizational climate literature is climate strength. As opposed to traditional climate research that focuses on the level of the climate in the unit (typically the unit's mean climate score), the climate strength approach focuses on the variability within the unit in climate perceptions (e.g., the standard deviation). Researchers studying climate strength as a moderator of the relationships between climate level and outcomes typically hypothesize that the strength of those relationships will be stronger when the climate is stronger, such that a positive climate will have a more positive impact when unit members agree it is positive and a negative climate will have a more negative impact when unit members agree it is negative. This approach has been applied to research on service climate (e.g., Schneider et al., 2002), justice climate (e.g., Colquitt et al., 2002), and molar climate

(e.g., Lindell and Brandt, 2000), among many others. The research findings on this hypothesis have been somewhat mixed perhaps as a result of the lack of sufficient variability in climate strength to reveal significant moderated relationships (Ehrhart et al., 2014). Nevertheless, the climate strength approach has continued to grow, and researchers have begun to examine what causes climates to be strong or weak, as well as possible direct outcomes of climate strength.

A final general direction of climate research has been to study climate as a moderator of other individual- or unit-level relationships. One theme in early climate research was that certain personal characteristics of workers would matter more or less depending on the climate. More specifically, if the characteristics were aligned with the climate, then positive outcomes (e.g., performance) would result, and if they were not aligned, then negative outcomes (e.g., stress) would result. Other variations of this thinking were similar to the idea of situational strength (Mischel, 1968), such that certain climates were thought to encourage and others discourage the display of individual differences (Schneider, 1975). Recent research along these lines has demonstrated that employees' service orientation is a significant predictor of customer-oriented behaviors when service climate is high but not when it is low because a high service climate allows employees with high customer orientation the opportunity to display customer-oriented behaviors (Grizzle et al., 2009). Climate has also been studied as a moderator of unit-level relationships with outcomes, such that certain unit-level characteristics or processes are less likely to be successful unless the context (i.e., the climate) is right. Examples can be found in research on innovation climate with regard to outcomes of process innovations (Baer and Frese, 2003) and in the domain of safety climate with regard to safety training (Burke et al., 2008).

Future Directions

Many of the concerns of early climate research have been addressed, but as climate researchers continue to become more sophisticated in their thinking, it will be important to revisit some of these concerns in new ways as well as to examine new issues that have surfaced as the field has advanced. One area to revisit is the relationship between climate and culture. Although the distinction between the two constructs has been made, we are left with how they can be integrated together. As described above, there has been increasing attention in recent years in the theoretical connections between the constructs, but little empirical research has resulted as of yet. It is likely that it will change in coming years, especially with regard to how organizational culture across levels (assumptions, espoused values, and artifacts; Schein, 2010) serves to support specific strategic climates in organizations.

Another area that is likely to see increased attention is the development of climate over time. There is some theoretical work hypothesizing how climates are formed, but little empirical work. Schneider and Reichers (1983) suggested that climates develop in three different ways. First, as individuals interact with each other, they will converge on shared meanings (symbolic interaction approach). Second, homogeneity can occur due to the way employees are attracted and selected to the organizations and eventually how individuals that do not fit well will leave through attrition. Finally, climates can develop as individuals are exposed to the same practices, policies, and procedures (i.e., a structuralist approach). More recently, Roberson and Colquitt (2005) integrated network theory with climate research to explain how justice climates can emerge and to identify some of the barriers to that emergence in teams. Nevertheless, neither of these theories addresses how much influence leaders have on climate development or if climates can change over time as feedback is received by the managers from organizational outcomes or employees. Again, little longitudinal research on climate has been conducted, so our understanding of how climates are created and how they change over time is mainly theoretical at this point.

A third area for future research is on climate across levels of analysis. As there has been an increased recognition that individuals and organizational units are interdependent, there has been a steady increase in multilevel research across many domains. Zohar and Luria (2005) provided an excellent demonstration of how both climate level and climate strength operate across manufacturing plants and the work groups within those plants. In other areas, there has been interest in how strategies and values 'trickle down' from top management to managers to employees (Mayer et al., 2009). More research is needed clarifying the issues that are most relevant for climate across levels and how climates at different levels can work synergistically or at cross-purposes.

Finally, in coming years we are likely to see more research on how multiple types of climate operate simultaneously. There are many climates that have been linked to similar outcomes such as performance and customer satisfaction. However, we do not know much as to how multiple climates can have a synergistic effect, neutralizing effect, or possibly even compete against each other to affect employee attitudes and organizational outcomes. Some very preliminary work has already been done in this area; for instance, McKay et al. (2011) studied both diversity climate and service climate and found that diversity climate had the strongest relationship with customer satisfaction when both service climate and minority representation levels were high. However, more theoretical and empirical work is needed in this area to shed light on such issues as how the relationship between process and strategic climates operates and the implications for units when there are competing priorities from multiple strategic climates.

Conclusion and Implications

Research on organizational climate is growing and expanding in numerous directions. Not only is it being studied in new and exciting ways, but it is being applied in an increasing number of domains. This growth is testament to the usefulness of the construct for understanding employee experiences in organizations and how they contribute to organizational effectiveness. In that sense, organizational climate is a very practical construct; indeed, Johns (2006) has suggested that it is impossible to understand behaviors in organizations without understanding the situation. Thus, we would be remiss if we did not briefly highlight the practical implications of research

on organizational climate to the functioning of organizations. Perhaps the most important lesson from the literature on organizational climate is that of alignment. If managers have strategic goals they wish to accomplish, then the climate literature suggests the importance of aligning their prioritization of those goals through the messages that are sent in the organization's policies, practices, procedures, and reward systems. If those strategies are simply espoused by management but not enacted through their behavior and the systems they create, then employees will be unlikely to be motivated to work toward achieving those goals. Climate surveys can assist managers in diagnosing these issues within their organization. For instance, climate surveys can help assess if there is alignment between the organization's strategic goals and its policies, practices, and procedures as perceived by employees. It is not uncommon for top-level managers to have different perceptions of organizational climates than their employees, so if employee perceptions are not assessed, managers may not have an accurate picture of what is happening in their organization. By examining specific dimensions and items in climate surveys, managers in organizations can derive specific actionable steps to help make the desired climates become more salient to employees. In addition, surveys addressing molar or process climates can help leaders understand the foundation they must establish in order to build the strategic climate they desire. Thus, addressing organizational climate in all of its manifestations can provide an evidence-based framework for organizational change and, ultimately, organizational effectiveness.

See also: Group Processes, Social Psychology of; Industrial–Organizational Psychology: Science and Practice; Job Satisfaction; Teamwork and Team Performance Measurement.

Bibliography

Baer, M., Frese, M., 2003. Innovation is not enough: climates for initiative and psychological safety, process innovations, and firm performance. Journal of Organizational Behavior 24, 45–68.

Beus, J.M., Payne, S.C., Bergman, M.E., Arthur Jr., W., 2010. Safety climate and injuries: an examination of theoretical and empirical relationships. Journal of Applied Psychology 95, 713–727.

Bliese, P.D., 2000. Within-group agreement, non-independence, and reliability: implications for data aggregation and analyses. In: Klein, K.J., Kozlowski, S.W.J. (Eds.), Multilevel Theory, Research and Methods in Organizations: Foundations, Extensions, and New Directions. Jossey-Bass, San Francisco, CA, pp. 349–381.

Burke, M.J., Chan-Serafin, S., Salvador, R., Smith, A., Sarpy, S.A., 2008. The role of national culture and organizational climate in safety training effectiveness. European Journal of Work and Organizational Psychology 17, 133–152.

Christian, M.S., Bradley, J.C., Wallace, J.C., Burke, M.J., 2009. Workplace safety: a meta-analysis of the roles of person and situation factors. Journal of Applied Psychology 94, 1103–1127.

Colquitt, J.A., Noe, R.A., Jackson, C.L., 2002. Justice in teams: antecedents and consequences of procedural justice climate. Personnel Psychology 58, 83–109.

Denison, D.R., 1996. What is the difference between organizational culture and organizational climate? A native's point of view on a decade of paradigm wars. Academy of Management Review 21, 619–654.

Dietz, J., Pugh, S.D., Wiley, J.W., 2004. Service climate effects on customer attitudes: an examination of boundary conditions. Academy of Management Journal 47, 81–92.

Ehrhart, M.G., Schneider, B., Macey, W.H., 2014. Organizational Climate and Culture: An Introduction to Theory, Research, and Practice. Routledge, New York.

Glick, W.H., 1985. Conceptualizing and measuring organizational and psychological climate: pitfalls of multilevel research. Academy of Management Review 10, 601–610.

Grizzle, J.W., Zablah, A.R., Brown, T.J., Mowen, J.C., Lee, J.M., 2009. Employee customer orientation in context: how the environment moderates the influence of customer orientation on performance outcomes. Journal of Applied Psychology 94 (5), 1227–1242.

Hofmann, D.A., Mark, B., 2006. An investigation of the relationship between safety climate and medication errors as well as other nurse and patient outcomes. Personnel Psychology 59, 847–869.

Hong, Y., Liao, H., Hu, J., Jiang, K., 2013. Missing link in the service profit chain: a meta-analytic review of the antecedents, consequences, and moderators of service climate. Journal of Applied Psychology 98, 237–267.

James, L.R., 1982. Aggregation bias in estimates of perceptual agreement. Journal of Applied Psychology 67, 219–229.

James, L.R., Choi, C.C., Ko, C.H.E., McNeil, P.K., Minton, M.K., Wright, M.A., Kim, K., 2008. Organizational and psychological climate: a review of theory and research. European Journal of Work and Organizational Psychology 17, 5–32.

James, L.R., Demaree, R.G., Wolf, G., 1984. Estimating within-group interrater reliability with and without response bias. Journal of Applied Psychology 69, 85–98.

James, L.R., Jones, A.P., 1974. Organizational climate: a review of theory and research. Psychological Bulletin 81, 1096–1112.

Johns, G., 2006. The essential impact of context on organizational behavior. Academy of Management Review 31, 386–408.

Klein, K.J., Conn, A.B., Smith, D.B., Sorra, J.S., 2001. Is everyone in agreement? an exploration of within-group agreement in employee perceptions of the work environment. Journal of Applied Psychology 86, 3–16.

Kuenzi, M., Schminke, M., 2009. Assembling fragments into a lens: a review, critique, and proposed research agenda for the organizational work climate literature. Journal of Management 35, 634–717.

LeBreton, J.M., Senter, J.L., 2008. Answers to twenty questions about interrater reliability and interrater agreement. Organizational Research Methods 11, 815–852.

Lewin, K., Lippitt, R., White, R.K., 1939. Patterns of aggressive behavior in experimentally created "social climates". Journal of Social Psychology 10, 271–299.

Lindell, M., Brandt, C., 2000. Climate quality and climate consensus as mediators of the relationship between organizational antecedents and outcomes. Journal of Applied Psychology 85, 331–348.

Mayer, D.M., Ehrhart, M.G., Schneider, B., 2009a. Service attribute boundary conditions of the service climate – customer satisfaction link. Academy of Management Journal 52, 1034–1050.

Mayer, D.M., Kuenzi, M., Greenbaum, R., Bardes, M., Salvador, R., 2009b. How low does ethical leadership flow? Test of a trickle-down model. Organizational Behavior and Human Decision Processes 108, 1–13.

McKay, P.F., Avery, D.R., Liao, H., Morris, M.A., 2011. Does diversity climate lead to customer satisfaction? it depends on the service climate and business unit demography. Organization Science 22, 788–803.

Mischel, W., 1968. Personality and Assessment. Wiley, New York.

Neal, A., Griffin, M.A., 2006. A study of the lagged relationships among safety climate, safety motivation, safety behavior, and accidents at the individual and group levels. Journal of Applied Psychology 91, 946–953.

Ostroff, C., 1993. The effects of climate and personal influences on individual behavior and attitudes in organizations. Organizational Behavior and Human Decision Processes 56, 56–90.

Ostroff, C., Kinicki, A.J., Muhammad, R.S., 2012. Organizational culture and climate. In: Schmitt, N.W., Highhouse, S. (Eds.), Handbook of Psychology, Industrial and Organizational Psychology, second ed., vol. 12. Wiley, Hoboken, NJ, pp. 643–676.

Patterson, M.G., Warr, P., West, M.A., 2004. Organizational climate and company productivity: the role of employee affect and employee level. Journal of Occupational and Organizational Psychology 77, 193–216.

Patterson, M.G., West, M.A., Shackleton, V.J., Dawson, J.F., Lawthom, R., Maitlis, S., Robinson, D.L., Wallace, A.M., 2005. Validating the organizational climate measure: links to managerial practices, productivity, and innovation. Journal of Organizational Behavior 26, 379–408.

Quinn, R.E., Rohrbaugh, J., 1983. A special model of effectiveness criteria: toward a competing values approach to organizational analysis. Management Science 29, 363–377.

Roberson, Q.M., Colquitt, J.A., 2005. Shared and configural justice: a social network model of justice in teams. Academy of Management Review 30, 595–607.

Salanova, M., Agut, S., Peiró, J.M., 2005. Linking organizational resources and work engagement to employee performance and customer loyalty: the mediation of service climate. Journal of Applied Psychology 90, 1217–1227.

Schein, E.H., 2010. Organizational Culture and Leadership, fourth ed. Jossey-Bass, San Francisco, CA.

Schneider, B., 1975. Organizational climates: an essay. Personnel Psychology 28, 447–479.

Schneider, B., Ehrhart, M.G., Macey, W.H., 2011a. Organizational climate research: achievements and the road ahead. In: Ashkanasy, N.M., Wilderom, C.P.M., Peterson, M.F. (Eds.), Handbook of Organizational Culture and Climate, second ed. Sage, Thousand Oaks, CA, pp. 29–49.

Schneider, B., Ehrhart, M.G., Macey, W.H., 2011b. Perspectives on organizational climate and culture. In: Zedeck, S. (Ed.), APA Handbook of Industrial and Organizational Psychology, Building and Developing the Organization, vol. 1. American Psychological Association, Washington, DC, pp. 373–414,.

Schneider, B., Ehrhart, M.G., Mayer, D.M., Saltz, J.L., Niles-Jolly, K., 2005. Understanding organization-customer links in service settings. Academy of Management Journal 48, 1017–1032.

Schneider, B., Reichers, A.E., 1983. On the etiology of climates. Personnel Psychology 36, 19–39.

Schneider, B., Salvaggio, A.N., Subirats, M., 2002. Climate strength: a new direction for climate research. Journal of Applied Psychology 87, 220–229.

Schneider, B., White, S.S., Paul, M.C., 1998. Linking service climate and customer perceptions of service quality: test of a causal model. Journal of Applied Psychology 83, 150–163.

Wallace, J.C., Edwards, B.D., Paul, J., Burke, M., Christian, M., Eissa, G. Change the referent? A meta-analytic investigation of direct and referent-shift consensus models for organizational climate. Journal of Management, in press.

Wallace, J.C., Popp, E., Mondore, S., 2006. Safety climate as a mediator between foundation climates and occupational accidents: a group-level investigation. Journal of Applied Psychology 91, 681–688.

Whitman, D.S., Caleo, S., Carpenter, N.C., Horner, M.T., Bernerth, J.B., 2012. Fairness at the collective level: a meta-analytic examination of the consequences and boundary conditions of organizational justice climate. Journal of Applied Psychology 97, 776–791.

Zohar, D., Hofmann, D.H., 2012. Organizational culture and climate. In: Kozlowski, S.W.J. (Ed.), The Oxford Handbook of Industrial and Organizational Psychology. Oxford University Press, Oxford, UK, pp. 643–666,.

Zohar, D., Luria, G., 2005. A multi-level model of safety climate: cross-level relationships between organization and group-level climates. Journal of Applied Psychology 90, 616–628.

Organizational Commitment

Howard J Klein and Heeman Park, The Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Organizational commitment is the psychological bond an employee has to their employing organization. Commitment has been one of the most frequently examined constructs in organizational research because of the demonstrated impact it has on outcomes important to individuals and organizations. In this article, we define commitment, discuss the evolution of the construct, overview the outcomes and antecedents of commitment, contrast commitment to other related constructs, and outline current issues in the study of commitment requiring attention in future research.

Commitment is a psychological bond that an individual forms with a particular foci or target. Our focus is on organizational commitment and we will henceforth use the term commitment to refer to the commitment an employee has to their employing organization, even though bonds can be formed between employees and other organizations (e.g., client organizations, professional associations, unions) as well as to numerous other workplace targets (e.g., team, supervisor, goal; Becker, 1992; Reichers, 1985). Commitment is a frequently examined construct, because of the demonstrated impact it has on outcomes important to individuals and organizations such as absenteeism, turnover, motivation, performance, prosocial behaviors, and well-being (Becker et al., 2009; Mathieu and Zajac, 1990; Meyer et al., 2002).

Commitment has been studied for decades and many definitions and conceptualizations of commitment have emerged over time. Commitment has often been referred to as a work attitude, but the nature of the construct does not fit current definitions of an attitude. Instead, commitment is more accurately conceptualized as a psychological state (Meyer and Allen, 1991). Klein et al. (2012) proposed a definition that we adopt here that is applicable to any target, excludes confounds present in prior definitions, and highlights the distinctiveness of commitment bonds. In the past, the term commitment has been used to encompass a wide range of different types of bonds or attachments (e.g., Kanter, 1968). Klein et al. (2012) define commitment as a particular type of bond and differentiate it from other bond types (e.g., instrumental or identification). They specifically define commitment as "*a volitional psychological bond reflecting dedication to and responsibility for a particular target*" (p. 137). Klein et al. (2012) also present the process model of commitment shown in **Figure 1**. We overview the outcomes and antecedents of commitment identified in this model after briefly summarizing prior treatments of the construct.

Evolution of the Construct

The earliest research on commitment was conducted by sociologists who examined group attachments as a mechanism for generating and sustaining social order (e.g., Parsons, 1951) or in understanding social roles (e.g., Gouldner, 1960). Scholars in psychology and management began to pay more attention to commitment in the 1960s, with that early commitment research taking two very different perspectives, reflecting either a behavioral or psychological view. The behavioral view considered commitment to be the result of behavior, with prior acts 'committing' people to a future course of action due to a desire for consistency, sense of obligation, and the 'sunk costs' of previously invested time and effort (e.g., Becker, 1960; Kiesler, 1971; Salancik, 1977). Work from the behavioral perspective provided important insights into the phenomena of commitment, particularly in terms of decisions to stay with and contribute to organizations based on behavioral consistency, investment, and exchange. This focus on behavior and exchange was, however, viewed by many as providing an incomplete explanation.

The alternative psychological or attitudinal view considered commitment as a driver of behavior rather than resulting from behavior. This perspective focused on how individuals identify or relate to their organization (e.g., Buchanan, 1974; Porter et al., 1974). From this perspective, commitment was presented as an important determinant of turnover and other outcomes, and as an alternative to job satisfaction that could be measured and influenced. Mowday et al. (1982) summarized the early work conducted from these perspectives and attempted to reconcile the two views, with primacy given to the psychological view.

Within and across these two views, there were numerous variations in focus and definition (see Klein et al., 2009), leading to efforts at developing integrative frameworks. For the past 20 years, the most popular of those frameworks has been the three-component model (TCM) presented by Meyer and Allen (1991). In this model, commitment is defined as "a force that binds an individual to a course of action of relevance to one or more targets" (Meyer and Herscovitch, 2001: p. 301), with that force experienced as one or more mind-sets – affective ('want to' or desire), normative ('ought to' or obligation), and continuance ('have to' or cost). This model, presented in **Figure 2**, has generated a tremendous amount of research regarding the antecedents and consequences of commitment. The TCM has, however, been criticized for a number of reasons including a lack of theoretical justification for the three mind-sets and being overly broad, with definitional overlaps with closely related constructs, leading to the Klein et al. (2012) reconceptualization.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.22032-1

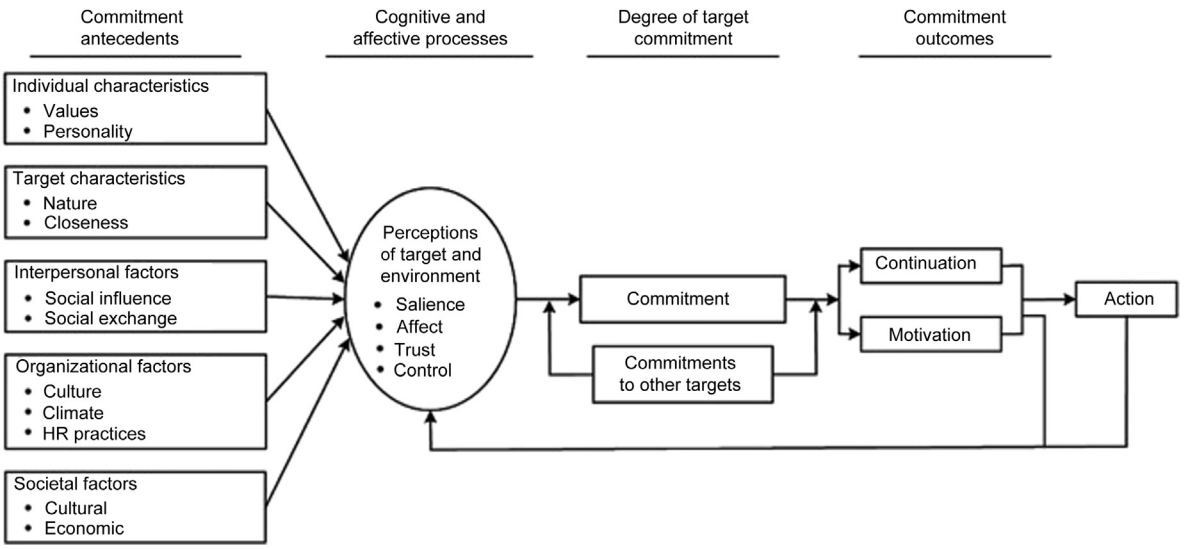

Figure 1 Klein et al.'s (2012) process model of commitment. Reprinted from, Klein, H.J., Molloy, J.C., Brinsfield, C.T., Copyright 2012. Reconceptualizing workplace commitment to redress a stretched construct: revisiting assumptions and removing confounds. Academy of Management Review 37 (1), 139, with permission from Academy of Management.

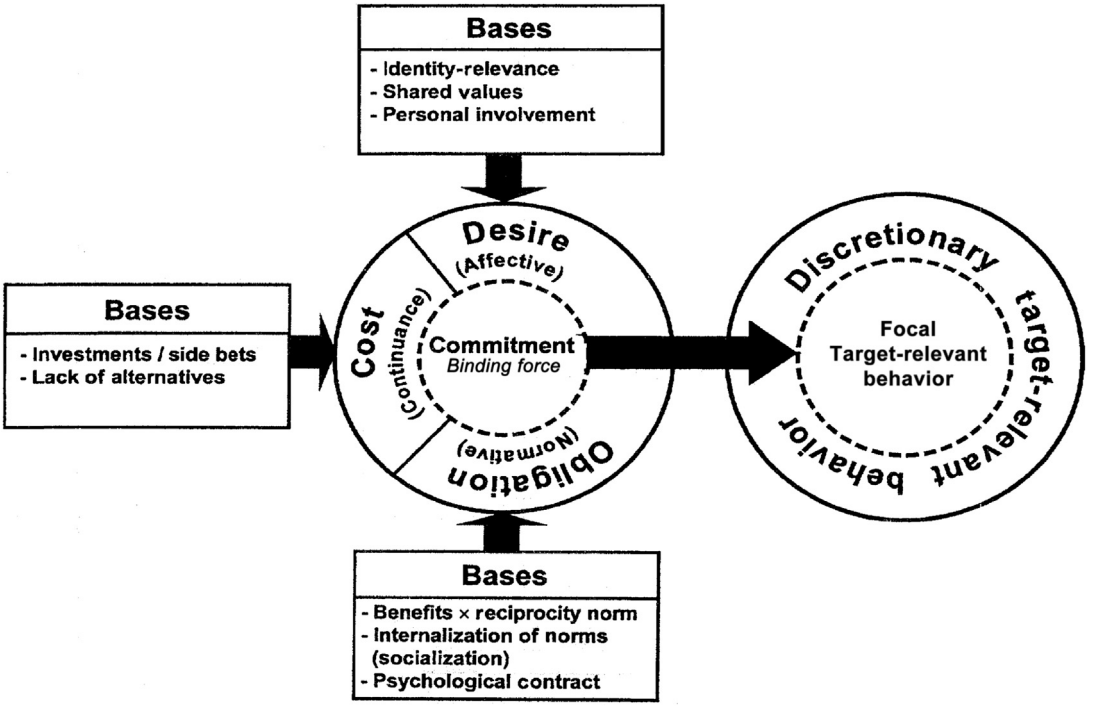

Figure 2 Meyer and Herscovitch's (2001) general model of workplace commitment. Reprinted from Meyer, J.P., Herscovitch, L., Copyright 2001. Commitment in the workplace toward a general model. Human Resource Management Review 11 (3), 317, with permission from Elsevier.

Commitment Outcomes

As noted previously, Commitment is associated with a number of desirable consequences. The Klein et al. (2012) model, shown in **Figure 1**, proposes two immediate outcomes (i.e., continuation and motivation) with action and the results of those actions (e.g., performance) being more distal outcomes. Continuation includes both the intention to continue with the organization and continuation behaviors. In short, committed individuals are less likely to withdraw from the organization. Meta-analyses (Mathieu and Zajac, 1990; Meyer et al., 2002) have consistently demonstrated negative relationships between commitment and turnover intentions, actual turnover, and other withdrawal behaviors (e.g., tardiness, absenteeism).

In terms of motivation, high commitment results in individuals' allocating more effort and resources in support of the organization and being more willing to make trade-offs in favor of the organization when allocating constrained resources such as time and attention (Klein et al., 2012). The relationship between commitment and motivation has been supported in prior research, summarized in the above referenced meta-analyses. Higher performance, both inrole and extrarole (e.g., organizational citizenship behaviors), can be expected to result from higher commitment, but indirectly, through the effects of continuation and motivation (Klein et al., 2012). Prior research supports these relationships, with stronger effects generally observed for extrarole than inrole performances.

Antecedents of Commitment

A vast number of factors have been identified as antecedents of commitment. The process model shown in **Figure 1** suggests that four proximal states are the immediate determinants of commitment with an inclusive set of more distal antecedents, organized by level that can influence the development of commitment through the four proximal states (see Klein et al., 2012). The four proximal states, which based on cognitive and emotional evaluations of the target and environment, are salience, positive affect, trust, and perceived control. First, based on field theory, the more salient the organization, relative to other potential commitment targets, the more likely that commitment will develop. Stated differently, people are more likely to develop a commitment bond to when they feel closer to the organization.

The second factor is positive affect. Individuals are more likely to dedicate themselves to and care about positively evaluated organizations than negatively evaluated organizations. This is based on approach motivation (i.e., individuals approach pleasurable situations and avoid painful situations) among other theories and is supported by the positive relationship demonstrated between commitment and satisfaction. Trust, the acceptance of some vulnerability based on positive expectations that others will reciprocate, is the third proximal state. Without trust in the organization, the dedication that is at the core of commitment is unlikely. The final proximal state, perceived control, is based on the theory of planned behavior, social cognitive theory, and other research. Essentially, individuals are more willing to commit to an organization when they feel they have a sense of control over the situation and confidence in their ability to meet the demands of their work role.

A wide range of factors influence those proximal states, which are grouped, as illustrated in **Figure 1**, from individual characteristics to target, relational, organizational, and societal factors. Individual level factors are attributes of the employee such as values and personality traits. Personal values such as work ethic and work centrality have been shown to lead to organizational commitment, as has the congruence between personal and organizational values. Personality traits found to relate to commitment include locus of control, conscientiousness, extraversion, and regulatory focus. Other individual differences shown to influence perceptual evaluations, such as trait affect, also influence commitment via the above proximal states.

A second set of antecedent factors arises from the social relationships that individuals have, both work and nonwork. It has been shown, for example, that having a dense friendship network is positively associated with commitment (Morrison, 2002). In addition, both social influences and exchanges resulting from these relationships can impact the identified perceptual evaluations and subsequently commitment. Examples of these social influences include what others think and say about the organization and the commitment of others. Social exchanges include the effects of perceived organizational support, supervisor supportiveness, and leader–member exchange.

With the organization as the target, target and organizational characteristics become a single category. Target characteristics include factors such as the psychological proximity of the organization, which impacts salience. Similarly, characteristics such as the legitimacy and reputation of the organization can influence commitment through evaluations of trust and positive affect. Other organizational factors that would impact the more proximal states include culture (and subcultures) and climate relative to numerous foci (e.g., safety, justice productivity, burnout). Finally, an organization's human resource practices can facilitate commitment bonds between employees and the organization. For example, Wayne et al. (2013) found that family supportive organizational practices are positively related to an employees' organizational commitment via the family's positive affect toward the organization.

The final antecedent category reflects societal-level factors – the broader socioeconomic environment in which the individual and organization are embedded. Because national culture shapes perceptions and ascribed meaning, cultural factors can influence the proximal states that determine commitment. Uncertainty avoidance, power distance, and collectivism, for example, can impact perceptions of control and trust, and have been shown to influence commitment. Macroeconomic factors can similarly impact perceptions and, in turn, commitment. Factors relevant to target salience and trust toward an organization, for example, include a country or region's economic conditions, business system characteristics, and labor market features (e.g., segmentation, strength, and role of unions).

Related Constructs

The distinctions between commitment and several related constructs have often been unclear because of the various, frequently confounded, prior definitions and measures of commitment. Below we address the distinctions and relationships between commitment and satisfaction, identification, motivation, engagement, and burnout.

Satisfaction

Prior definitions of commitment have blurred the boundary between commitment and job satisfaction by emphasizing affective attachment. As a result, some authors have suggested

that satisfaction and commitment can be combined into an overall job attitudes composite (Harrison et al., 2006) or even that commitment is a redundant and unneeded construct given the empirical overlap (Le et al., 2010). Although meta-analytic investigations have shown the two constructs to be highly related, there are conceptual distinctions that can be made between satisfaction and commitment. For example, attitudes such as satisfaction involve evaluative judgments (i.e., one's summative evaluation of the organization), whereas commitment, as defined here, does not involve an affect-based evaluative aspect. The causal ordering of the relationship between satisfaction and commitment has also been of interest and a debated issue. Some scholars have asserted that satisfaction causes commitment (e.g., Mowday et al., 1982) and others the opposite (e.g., Bateman and Strasser, 1984). Conceptually, we conclude that a reciprocal causal relationship (e.g., Farkas and Tetrick, 1989) is the most appropriate view, with satisfaction being a stronger antecedent than consequence. Specifically, satisfaction likely impacts commitment through its primary antecedents. That is, job satisfaction should facilitate commitment, as individuals are more likely to commit to positively evaluated targets. In turn, being committed to an organization likely makes the positive aspects of the organization more salient, biasing subsequent affect. Additional empirical research is needed to tease apart and verify the casual ordering of these relationships, over time and within person, using more recent and less confounded measures.

Identification

Prior definitions of commitment have also blurred the boundary between commitment and identification. For example, Porter et al. (1974) in part defined commitment as the strength of an individual's identification with a particular organization. Identification is a type of bond, like commitment, but there are unique aspects of commitment not captured by identification and vice versa. For example, the 'forming of oneness' with the target or the notion that the target becomes part of the individual's self-concept or image is central to identification (Riketta, 2005). Others have further detailed the conceptual differences and the need to better differentiate these constructs (see Mael and Ashforth, 1992; Meyer et al., 2006). Meta-analyses (e.g., Riketta, 2005) have demonstrated very strong correlations between identification and commitment. This is partially due to prior confounds in definition and measurement, with some prior measures of commitment largely indistinguishable from measures of identification (Klein et al., 2014). Yet studies have demonstrated empirical distinctions (e.g., Mael and Ashforth, 1992; Riketta, 2005). Thus, both conceptual and empirical evidence supports the distinctiveness of the two constructs (Riketta, 2005) and the existing evidence suggests that social identities are more likely to be antecedents than a consequence of commitment (Meyer et al., 2006).

Motivation

Some authors have defined commitment as a force (e.g., Meyer and Herscovitch, 2001), and motivation is typically defined as a set of internal and external forces that initiate behavior and determine its form, direction, intensity, and duration (Pinder, 1998). Meyer et al. (2004) proposed a model that presents motivation and commitment as related yet distinguishable concepts, with commitment being one of several energizing forces that constitute motivation. Klein et al. (2012) position commitment as distinct from motivation, with motivation being a primary commitment outcome as described above. In addition, since commitment is one of several determinants of motivation, other factors (e.g., needs, incentives) may be more salient in determining motivation at a given point in time (Meyer et al., 2004). Klein et al. (2009) summarize additional arguments for the distinctiveness of these two constructs.

Engagement

A concept sometimes defined in motivational terms (e.g., Kahn, 1990; Rich et al., 2010) is engagement. Other authors define engagement as a trait, state, presence/enthusiasm, and the opposite of burnout (e.g., Macey and Schneider, 2008; Schaufeli et al., 2002). Because of this definitional diversity, it can be challenging to differentiate engagement from commitment. Across all these definitions, engagement would be expected to relate positively to commitment. When defined in motivation terms (i.e., the willingness to expend personal energies on behalf of a target, or the harnessing of oneself to a role), engagement is an expected, motivational outcome of commitment. Of these varied definitions, those that discuss engagement as a state are most similar to commitment; yet state engagement is frequently conceptualized as a broad, higher order concept incorporating commitment and other constructs. For instance, Macey and Schneider (2008) discuss state engagement as including energy, absorption, satisfaction, involvement, commitment, and empowerment. Similarly, Schaufeli et al. (2002) conceptualize engagement as being involved, dedicated, enthused, inspired, committed, engrossed, and attached to one's work. Note that in defining commitment in terms of dedication, Klein et al. (2012) were not equating commitment with the dedication dimension in the Schaufeli et al. typology.

Burnout, Stress, and Well-being

Given that engagement is sometimes defined as the opposite of burnout, and the above-noted positive relationships between commitment and engagement, it is not surprising that commitment has been found to relate negatively to burnout (e.g., Cropanzano et al., 2003). In addition to burnout, research has consistently demonstrated associations between commitment and other employee physical and psychological health outcomes. Specifically, positive outcomes have been demonstrated with employee well-being (including physical well-being and physical general health, and mental health) and negative associations with physical and mental health complaints (including anxiety and depression), burnout and stress (see Meyer and Maltin, 2010). In addition to these main effects, several other studies have examined the role of commitment in buffering or mitigating the negative effects of perceived stressors on tension, anxiety, burnout, and depression. Commitment is thought to provide a sense of purpose and resilience, serving as a resource to attenuate the harmful

effects of stressors. Results are generally supportive of this moderating effect, though there are some inconsistencies in the findings (Meyer and Maltin, 2010).

Despite this evidence that commitment is generally positively associated with well-being, there may be situations where commitment may threaten a person's sense of well-being and lead to negative outcomes, resulting in what Reilly (1994) termed the "paradox of commitment". First, commitment may actually increase stressor–strain relationships because highly committed employees are likely to experience greater stress stemming from the increased salience of job demands relative to less committed employees who may not care. Second, commitment may negatively impact well-being when there are competing commitments – commitments that are in direct conflict or simply in competition for an individual's limited time or attention (Mathieu and Zajac, 1990). Similarly, over commitment – when the strength of a commitment results in an individual expending levels of time and effort that are not sustainable – can also result in high levels of stress and burnout (Schaufeli et al., 2002). A final potentially negative consequence of commitment on well-being pertains to the cessation or loss of a commitment, which has not been adequately considered in prior research. In addition, much of this research has been cross sectional, and longitudinal studies are needed to better understand how commitment relates to these employee-relevant outcomes. The evidence to date, however, does suggest that commitment has important consequences for employees themselves.

Issues Needing Research and Resolution

Competing Conceptual Definitions

As reviewed by Klein et al. (2009), numerous conceptualizations of commitment can be found in the literature. Inherent in this definitional debate, which has implications for the relationship between commitment and other constructs, is the location of the construct boundary. For example, investments that create a bond (e.g., sunk costs, side bets; Becker, 1960) as well as goal or value congruence (e.g., Porter et al., 1974) are included in some definitions of commitment, but excluded from others and instead viewed as antecedents. Similarly, some definitions include aspects of motivation, behavioral intentions, or even behaviors within the definition of commitment, whereas other definitions exclude these concepts, instead viewing them as distinct constructs that result from commitment. The location of the construct boundary matters as, depending on the definition used, a measure is either valid or confounded, constructs redundant or distinct, and observed empirical relationships accurate or inflated due to construct and measurement overlap (Klein et al., 2014). Klein et al. (2009) argue that five of the eight categories of definitions they identified are problematic because of conceptual overlap with antecedents or consequences. The remaining three conceptualizations view commitment as an attitude, a binding force, or as a bond or attachment. Klein et al. (2009) further note that commitment is not an attitude as attitudes are currently defined, and make the case for commitment as a bond over a binding force. There is not as yet, however, agreement on this in the literature.

A second aspect of the conceptual debate is whether commitment is multidimensional. There are several frameworks that view commitment as multidimensional, but there appears to be consensus that commitment, at its core, is one dimensional (Becker et al., 2009), with the dimensionality debates more around the multiple targets of commitment (e.g., organization, supervisor, career) and whether/how they are structured, the multiple bases that can lead to the formation of commitment (e.g., identification, compliance), and the multiple mind-sets that can characterize a commitment (e.g., normative, affective). With regard to commitment mind-sets, the empirical evidence does not always support the TCM, with arguments made for there being anywhere from two to eight distinct mind-sets (Klein et al., 2014). A related issue is whether all of these mind-sets should be considered commitment. The Klein et al. (2012) reconceptualization, for example, in differentiating among types of bonds and defining commitment as a particular type of bond, eliminates the need for ancillary concepts such as bases and mind-sets. Given the above conceptual disagreements, it is not surprising that there are also questions regarding the proper measurement of commitment. One set of issues centers around whether an uni- or multidimensional scale is needed and the extent to which those measures are construct valid. Another issue is whether the same measure can be used across targets or if distinct measures are needed for each commitment target.

Differences in Commitment across Targets

Organizational commitment is just one of a multitude of commitments individuals hold within and outside of work contexts. Those work-related commitments could be to other organizations (e.g., unions, professional associations, client organizations), individuals within or outside the organization (e.g., coworkers, supervisors, suppliers, customers), groups (e.g., work teams), and various ideas and initiatives (e.g., values, goals, decisions, policies, change programs). Commitment to different targets has, in the past largely, been studied separately, with different researchers examining different targets in different literature, and commitment to these different targets viewed as distinct (Klein et al., 2012). There have been efforts to integrate the research on commitment across targets and both the TCM and the more recent Klein et al. (2012) reconceptualization view the construct of commitment to be the same regardless of target. Meyer and Herscovitch (2001) extended the original TCM to be applicable to other targets; Klein et al. (2012) redefined commitment in a manner that is equally applicable to any target and concluded that although commitments to different targets may serve different purposes, there is a generalizable and distinctive meaning common to all commitments regardless of context or target. That conclusion was based on observed similarities across targets in terms of (1) how it is defined, (2) how it develops, (3) its antecedents, (4) the processes by which it operates, and (5) its consequences.

Defining commitment in a target-free manner does not mean that the target does not matter or that targets are interchangeable – only that distinct constructs are not needed to reflect commitment to distinct targets. There are likely phenomenological differences between commitments to different categories of targets (e.g., entities vs individuals), and

variation across targets can be expected in terms of the exact manifestation and relative importance of the identified causes and consequences of commitment (Klein et al., 2012). The need exists to systematically examine the differences and interplay among targets, but such research has been difficult up until now due to the absence of a brief measure that is applicable and comparable across any target. The need also exists, for a logically meaningful and parsimonious typology for categorizing commitment targets. Examples of needed research across commitment targets include identifying how the relative importance of the antecedents and outcomes might vary across targets and whether there are systematic differences in the development, maintenance, or dissipation of commitment over time across targets.

Temporality

Commitment has often been viewed as static construct, yet as a psychological state, levels of commitment can ebb and flow. Commitment theorists have recognized that commitment is a process that develops over time, yet there has been surprisingly little research on that development process, variation in commitment within individuals over time, or the dissipation of commitment. In terms of the formation of commitment, there is a large body of cross-sectional research examining correlates of commitment that are likely antecedents, but relatively little longitudinal research examining the development of commitment as a process that unfolds over time. There is theory to guide such research, but the presumed mechanisms (e.g., attributions, trust, salience; Klein et al., 2012; Meyer et al., 2004), and other possible factors warrant further investigation. In addition to better understanding the formation of commitment, research on factors that may strengthen, weaken, or even eliminate commitment is needed as commitments do not simply develop and then exist forever and key factors in altering commitment may not be the same as those that initially facilitate the development of commitment. There are only a few studies that have examined within-person changes in commitment over time and even fewer looking at how and why employees once committed to their organizations became uncommitted. As a result, we know little about how the consequences of such dissipation of commitment on work attitudes, behavior, or well-being; or how those effects differ depending on the reasons for the loss of commitment.

Downsides of Commitment

Finally, greater research attention needs to be given to the potential 'dark side' of commitment. The literature has largely assumed that more commitment is better without sufficiently recognizing that the consequences of high commitment can be either positive or negative for the individual, the organization, or both. This issue has been previously identified (e.g., Randall, 1988), but unlike the work on escalation of commitment, research on organizational commitment has almost exclusively focused on positive outcomes. High commitment can potentially be detrimental when the target is problematic (e.g., an unethical organization, abusive supervisor, flawed strategy), or when commitment becomes so strong as to bias judgment and decision making. Negative outcomes may also occur when commitment results in an individual expending levels of time and effort on behalf of a target that are not sustainable without sacrificing other commitments and/or well-being. This relates to another understudied downside of commitment, conflicting commitments. Research suggests that individuals can simultaneously commit to multiple targets and that such commitments need not be in conflict (Randall, 1988). There are, however, limits on an individual's resources (e.g., time, attention) and conflict is likely to result when multiple commitments create incompatible demands. Additional research is thus needed on both over commitment and conflicting commitment.

Cultural Differences and Influences

Commitment research has been conducted in numerous countries and it appears that commitment is relevant across cultures, but many questions remain concerning whether there are cultural differences in how commitment develops, is experienced, and influences behavior. There is some research suggesting that commitment models and measures developed in the Western world transfer reasonably well when applied in other cultures. Others (e.g., Wasti and Önder, 2009), however, question those results and the largely used imposed etic approach. Qualitative research is probably needed to better understand what commitment means to individuals within different cultures rather than simply evaluating the transportability of existing models. At that point, and after revising our models if needed, other issues including the appropriate translation of items and measurement equivalence, and the effects of cultural and values and norms on the development and experience of commitment can be examined.

Conclusion

Commitment is studied across a wide variety of disciplines, suggesting that it is a fundamental concept for understanding human behavior. Within the study of organizational phenomena, commitment has been one of the most frequently examined constructs. As a result of that attention, great strides have been made in our understanding of the antecedents and consequences of commitment, but many questions warrant further investigation and clarification. A better understanding of commitment is of both scientific and practical importance to organizations and individuals alike, as changes in the nature of organizations and work have not mitigated the need for a committed workforce.

See also: Industrial–Organizational Psychology: Science and Practice; Job Satisfaction; Organizational Citizenship Behavior; Personality and Values at Work; Psychological Contracts; Work Motivation; Workplace Stress.

Bibliography

Bandura, A., 1997. Self-Efficacy: The Exercise of Control. Freeman, New York.

Bateman, T.S., Strasser, S., 1984. A longitudinal analysis of the antecedents of organizational commitment. Academy of Management Journal 27 (1), 95–112.

Becker, H.S., 1960. Notes on the concept of commitment. American Journal of Sociology 66, 32–40.

Becker, T.E., 1992. Foci and bases of commitment: are they distinctions worth making? Academy of Management Journal 35, 232–244.

Becker, T.E., Klein, H.J., Meyer, J.P., 2009. Accumulated wisdom and new directions for workplace commitments. In: Klein, H.J., Becker, T.E., Meyer, J.P. (Eds.), Commitment in Organizations: Accumulated Wisdom and New Directions. Routledge/Taylor and Francis, New York, pp. 419–452.

Buchanan, B., 1974. Building organizational commitment: the socialization of managers in work organizations. Administrative Science Quarterly 19, 533–546.

Cropanzano, R., Rupp, D.E., Byrne, Z.S., 2003. The relationship of emotional exhaustion to work attitudes, job performance, and organizational citizenship behaviours. Journal of Applied Psychology 88, 160–169.

Farkas, A.J., Tetrick, L.E., 1989. A 3-wave longitudinal analysis of the causal ordering of satisfaction and commitment on turnover decision. Journal of Applied Psychology 74, 855–868.

Gouldner, H.P., 1960. Dimensions of organizational commitment. Administrative Science Quarterly 4, 468–490.

Harrison, D.A., Newman, D.A., Roth, P.L., 2006. How important are job attitudes? Meta analytic comparisons of integrative behavioral outcomes and time sequences. Academy of Management Journal 49, 305–325.

Kahn, W., 1990. Psychological conditions of personal engagement and disengagement at work. Academy of Management Journal 33, 692–724.

Kanter, R.M., 1968. Commitment and social organization: a study of commitment mechanisms in utopian communities. American Sociological Review 33, 499–517.

Kiesler, C.A., 1971. The Psychology of Commitment: Experiments Linking Behavior to Belief. Academic Press, New York.

Klein, H.J., Cooper, J.T., Molloy, J.C., Swanson, J.A., 2014. The assessment of commitment: advantages of a unidimensional, target-free approach. Journal of Applied Psychology 99, 222–238.

Klein, H.J., Molloy, J.C., Brinsfield, C.T., 2012. Reconceptualizing workplace commitment to redress a stretched construct: revisiting assumptions and removing confounds. Academy of Management Review 37 (1), 130–151.

Klein, H.J., Molloy, J.C., Cooper, J.T., 2009. Conceptual foundations: construct definitions and theoretical representations of workplace commitments. In: Klein, H.J., Becker, T.E., Meyer, J.P. (Eds.), Commitment in Organizations: Accumulated Wisdom and New Directions. Routledge/Taylor and Francis, New York, pp. 3–36.

Le, H., Schmidt, F.L., Harter, J.K., Lauver, K.J., 2010. The problem of empirical redundancy of constructs in organizational research: an empirical investigation. Organizational Behavior and Human Decision Processes 112 (2), 112–125.

Macey, W.H., Schneider, B., 2008. The meaning of employee engagement. Industrial and Organizational Psychology: Perspectives on Science and Practice 1, 3–30.

Mael, F.A., Ashforth, B.E., 1992. Alumni and their alma mater: a partial test of the reformulated model of organizational identification. Journal of Organizational Behavior 13, 103–123.

Mathieu, J.E., Zajac, D.M., 1990. A review and meta-analysis of the antecedents, correlates, and consequences of organizational commitment. Psychological Bulletin 108, 171–194.

Meyer, J.P., Allen, N.J., 1991. A three-component conceptualization of organizational commitment. Human Resource Management Review 1, 61–89.

Meyer, J.P., Becker, T.E., Vandenberghe, C., 2004. Employee commitment and motivation: conceptual analysis and integrative model. Journal of Applied Psychology 89, 991–1007.

Meyer, J.P., Becker, T.E., Van Dick, R., 2006. Social identities and commitments at work: toward an integrative model. Journal of Organizational Behavior 27, 665–683.

Meyer, J.P., Herscovitch, L., 2001. Commitment in the workplace: toward a general model. Human Resource Management Review 11, 299–326.

Meyer, J.P., Maltin, E.R., 2010. Employee commitment and well-being: a critical review, theoretical framework and research agenda. Journal of Vocational Behavior 77, 323–337.

Meyer, J.P., Stanley, D.J., Herscovitch, L., Topolnytsky, L., 2002. Affective, continuance, and normative commitment to the organization: a meta-analysis of antecedents, correlates, and consequences. Journal of Vocational Behavior 61, 20–52.

Morrison, E.W., 2002. Newcomers' relationships: the role of social network ties during socialization. Academy of Management Journal 45 (6), 1149–1160.

Mowday, R.T., Porter, L.W., Steers, R.M., 1982. Employee-Organization Linkages: The Psychology of Commitment, Absenteeism, and Turnover. Academic Press, New York.

Parsons, T., 1951. The Social System. Taylor & Francis, England.

Pinder, C.C., 1998. Motivation in Work Organizations. Prentice Hall, Upper Saddle River, NJ.

Porter, L.W., Steers, R.M., Mowday, R.T., Boulian, P.V., 1974. Organizational commitment, job satisfaction, and turnover among psychiatric technicians. Journal of Applied Psychology 59 (5), 603–609.

Randall, D.M., 1988. Multiple roles and organizational commitment. Journal of Organizational Behavior 9, 309–317.

Reichers, A.E., 1985. A review and reconceptualization of organizational commitment. Academy of Management Review 10, 465–476.

Reilly, N.P., 1994. Exploring the paradox: commitment as a moderator of the stressor–burnout relationship. Journal of Applied Social Psychology 24, 397–414.

Rich, B., Lepine, J.A., Crawford, E.R., 2010. Job engagement: antecedents and effects on job performance. Academy of Management Journal 53, 617–635.

Riketta, M., 2005. Organizational identification: a meta-analysis. Journal of Vocational Behavior 66, 358–384.

Salancik, G.R., 1977. Commitment and the control of organizational behavior. In: Staw, B.M., Salancik, G.R. (Eds.), New Directions in Organizational Behaviour. St. Clair Press, Chicago, pp. 1–54.

Schaufeli, W.B., Salanova, M., Gonzalez-Romá, V., Bakker, A.B., 2002. The measurement of engagement and burnout: a two sample confirmatory factor analytic approach. Journal of Happiness Studies 3, 71–92.

Wasti, S.A., Önder, C., 2009. Commitment across cultures: progress, pitfalls and propositions. In: Klein, H.J., Becker, T.E., Meyer, J.P. (Eds.), Commitment in Organizations: Accumulated Wisdom and New Directions. Routledge/Taylor and Francis, New York, pp. 309–343.

Wayne, J.H., Casper, W.J., Matthews, R.A., Allen, T.D., 2013. Family-supportive organization perceptions and organizational commitment: the mediating role of work–family conflict and enrichment and partner attitudes. Journal of Applied Psychology 98, 606–622.

Organizational Control

Paul Thompson, The University of Strathclyde, Glasgow, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although control can be seen as the essential problem of management, its treatment in mainstream organizational literature is problematic. It is frequently asserted that new competitive pressures and the enhanced demands of a service-oriented, knowledge-based economy meant that 'command and control' is no longer an option for successful business. Yet, while changes in types and levels of control have taken place, the weight of evidence is that most large organizations remain traditionally managed. In addition, it is often claimed that the contemporary locus of control has become internal rather than external, through the spread of normative controls or panoptic surveillance. Considerable problems remain, however, with the evidence on the extent of such practices and their effectiveness. Changes in the structure of the economy and corporations, however, have led to a widening of the scope of and greater hybridity in controls.

Mainstream Approaches: Ambiguities and Absences

As Pfeffer noted, "control is at once the essential problem of management and organization and the implicit focus of much of organization studies" (1997: p. 100). Yet there are wide disagreements as to its significance and substance. In mainstream models of organization behavior, control is treated, at best, ambiguously. Some standard textbooks simply do not have chapters on the issue. Systems theory is based on equilibrium and interdependence, and therefore it tends to assume goal consensus and downplays systematic conflicts of interest or identity in the employment relationship.

When control is discussed it is often alongside coordination. Any complex division of labor requires mechanisms to set goals, allocate responsibilities, and evaluate the effectiveness of performance. Coordination is a more neutral term than control and more compatible with an assumption of goal consensus. In this context, control tends to be treated as a *monitoring* device, with management's role to check progress, and ensure that actions occur as planned and to correct any deviation. The emphasis is on technical inputs and outputs in a self-adjusting system, performance standards, and feedback mechanisms.

Such emphases, however, can be misleading. Control often *is* present in mainstream approaches, it just is hiding in the shadows of something else. The key term used by Pfeffer was *implicit*. Texts that do have chapters with control in the title often pass over quickly onto other issues, such as job design, organizational structure, or leadership. In the first instance, debate is focused on designing structures that facilitate levels of control and coordination appropriate to types of work that require different levels of discretion and standardization. In the second, discussion of leadership styles, such as the classic polarity of authoritarian versus democratic approaches, constitute ways of discussing control, but do so through the language of influence and motivation.

A Break from Command and Control?

Reinforcing this ambiguity about whether control is necessary, the more practical management theorists continually have tried to abolish it or proclaim its imminent disappearance. From the human relations writers of the 1950s, with their distinctions between (bad) theory X and (good) theory Y, to the advocates of 1970s-style job enrichment and contemporary empowerment; control has been presented as destructive and out of date. The idea was that a high degree of autonomy over one's own work was a precondition for job satisfaction, and high productivity has been a cornerstone of organizational psychology (as captured in notions like self-efficacy). From the mid-1980s onward *necessity* was added to desirability. Influential articles such as by Walton (1985) declared that new competitive pressures and the enhanced demands of a service-oriented, knowledge-based economy meant that 'command and control' was no longer an option for successful business. We were moving from control to commitment and from rules to values as means of coordination. The future was empowered employees in self-directing teams, participating in and sharing responsibility for a wider range of decisions – more voice and choice at work.

A problem with such approaches is their aspirational nature and the subsequent tendency to conflate prescription and description. It is accepted across different perspectives that changes have taken place with respect to controls. In particular, there has been, through teamwork, increased involvement of employees in microlevel decision making about how to do their work – a form of democratized Taylorism (Adler, 1993). Much of this has been influenced by the importation of Japanese management techniques, repackaged in the 1990s as lean production. Although this sometimes has achieved high levels of productivity, continuous improvement, and quality assurance, most studies conclude that such methods are accompanied by intense surveillance, extensive standardization, and new forms of cultural or normative controls (see Graham, 2005). At a broader level, studies are consistent in showing that most organizations, at least in Britain and North America, remain traditionally managed and continue to take a low-trust, low-skill path to competitiveness. In the context of generally tougher times, marked by downsizing and delayering, many employees experience empowerment largely as a form of work

intensification. Firms are finding it harder to deliver on the kind of 'package' envisaged Walton, in which increased commitment and responsibilities would be offset by greater job security and other enhancements to the employment relationship.

Although companies are indeed seeking to enhance service quality and efficiency, they are also investing in new control systems to monitor the consistency and characteristics of those outputs. In services, generally there is ample evidence that organizations are scripting employee–customer interactions that previously might have been left to employee discretion (Ritzer, 1993), or specifying feelings rules for the mobilization of emotional labor (Hochschild, 1985). This emphasis on the body has been expanded in recent research on aesthetic labor, showing how some organizations are imposing standards of 'looking good and sounding right' in some service settings. Indeed, customers can be incorporated into control processes through mechanisms such as mystery shoppers and forms of feedback and evaluation. The redefinition of hospital patients or students as 'customers' can be used to initiate coercive evaluations of public sector professionals. Such employees increasingly are faced with audits and other forms of accountability to government bodies. Contrary to claims about inevitably enhanced autonomy in knowledge-intensive settings, the work of scientific and technical staff increasingly is subject to measurement and codification through the metrics and tools of knowledge management. Such trends are consistent with surveys showing that in many sectors levels of discretion are falling, especially among professional and expert labor.

What this discussion suggests is that mainstream conceptual frameworks are not only overoptimistic, but they are insufficiently robust to deal with changes in the *types* and *levels* of control. For example, it is perfectly possible to have increased operational autonomy for 'empowered' managers, existing alongside much tighter financial controls through targets and budgets.

Control: the Basic Building Block of Radical Organizational Theory

As Pfeffer (1997: p. 135) noted, those ambiguities toward the idea of control are in part responsible for the development of more critical perspectives on organizations. For a variety of radical theories, control is the core concept. Whereas mainstream perspectives treat control and coordination together, radical theorists argue that management performs a *dual* function in the enterprise. Managerial practices are a necessary means of *coordinating* diverse activities; however, they also bear the imprint of conflicting interests in the labor process; a conflict that reflects the unique nature of labor as a commodity. When capital purchases labor, it has only a potential or capacity to work. To ensure profitable production, capital must organize the conditions under which labor operates to its own advantage. But workers pursue their own interests for job security, higher rewards and satisfying work, developing their own counterorganization through informal job controls and output norms.

To resolve this problem and because they are under competitive pressure from other firms to cut costs and raise productivity, employers seek to control the conditions under which work takes place. Control is not an end in itself, but rather it is a means to transform the capacity to work established by the wage relation, into profitable production. This idea of the transformation of labor power into labor is taken from Marx and is the starting point of what has become known as *labor process theory* (LPT). Its adherents describe control as a set of mechanisms and practices that regulate the labor process. Edwards (1979) distinguished three elements in any system of control:

- direction and specification of work tasks;
- evaluation, monitoring, and assessment of performance; and
- the apparatus of discipline and reward to elicit cooperation and compliance

Such elements, however, may be best described as *detailed* control, in that they normally are connected to immediate work processes, whereas *general* control refers to management's capacity to subordinate labor to their direction of the production process as a whole. This allows for the recognition of tremendous variations in how detailed control is exercised. Such a model can even allow for employers giving workers significant discretion over tasks, as in semiautonomous work groups, *if* it maintains their overall control. Control also is not absolute, but, at least at the micro level, it is a contested relationship.

Strategies of Control: Debates and Disagreements

Radical perspectives have been conditioned by Braverman's (1974) argument that the twentieth century has seen the tightening of managerial control, primarily through the application of Taylorist and scientific management strategies. He provided an important corrective measure to the widespread view that Taylorism was a failed system, superseded by more sophisticated behavioral theories to be used for motivational and job design tools. But it is recognized widely that Braverman overestimated the dominance of Taylorist strategies and practices, and underestimated the varied and uneven implementation.

Nevertheless, because Braverman confused a particular system of control with management control in general, the question of *alternative strategies* was put firmly on the agenda. The best known is Richard Edward's (1979) model that is based on historically successive dominant modes of control, which reflect worker resistance and changing socioeconomic conditions. A nineteenth-century system of *simple* or *personal* control by employers exercising direct authority gave way to more complex *structural* forms with the transition from small business, competitive capitalism to corporate monopolies. The first of these forms was *technical* control, typified by the use of the assembly line, which can pace and direct the labor process. The contradiction for management is that it created a common work experience and basis for unified shop floor opposition. In contrast, a system of *bureaucratic* control embedded in the social and organizational structure of the

firm rather than in personal authority, offers management a means of redividing the workforce and tying it to impersonal rules and regulations.

In contrast, the other most influential model from Friedman (1977) eschews the notion of stages, preferring to set out ideal types or strategic poles of responsible autonomy and direct control, which run parallel throughout the history of capitalism. Paralleling mainstream notions of mechanistic and organic structures, or low- and high-trust employment relations, each strategy is said to generate its own inflexibilities in such areas as hiring and firing and task specification. The choice of strategy is governed by variations in the stability of labor and product markets, mediated by the interplay of worker resistance and managerial pressure.

Although the emphasis on alternative strategies was welcomed, such conceptual contributions were criticized by some commentators for the 'panacea fallacy' – that is, the idea that capital always seeks and finds definitive and comprehensive modes of control as the solution to its problems. Another strand of criticism drew from long-standing issues about the relations between ownership and management. Capital's interests are not given and management practices cannot be 'read-off' from them. Assumptions of a 'tight-coupling' underestimate the diversity and complexity of such practices and the significance for decision-making processes within the enterprise. In addition, there are conflicts within management, reflecting contending interests groups and the difficulty of carrying out integrative functions.

What has happened in practice is that researchers influenced by LPT have largely ignored such overarching frameworks and have used particular concepts, such as technical or bureaucratic control, to explain particular trends in new technology, recruitment practices, or internal labor markets. Alongside such usage, other researchers have opened up the empirical and theoretical terrain by demonstrating how control processes inside work or are shaped by the broader social divisions of labor connected to gender and family relations. LPT also has attempted to develop a more complex framework that can deal with the tensions between different management levels, functions, and ideologies in framing and exercising control. The most innovative is associated with Armstrong's (1989) work on *interprofessional competition*. Management functions for capital are mediated by competition among occupational groups to become the trusted agency to carry out control functions. Armstrong charted the historical and contemporary forms of competition between engineers, personnel, accounting, and other professions within the business enterprise. Who comes out on top is a function not just of their power resources, but the societally specific constitution of managerial expertise (for an overview of such research, see Thompson and van den Broek, 2010).

New Challenges, Critiques, and Responses

Although a degree of common ground emerged both about the varied influences on control strategies and the variations in combination and context through which control is exercised, new challenges have arisen from critical readings of contemporary social and organizational changes.

The Locus of Control: From External to Internal?

Rather than denying the salience of control, new challenges were made concerning where and how it was taking place. What was discussed earlier as a move from control *to* commitment was instead seen as a form of *internal control* that does not rely on external rewards and sanctions, or rule-following. Outside the largely consensual corporate culture literature, theorists within a labor process tradition already had gone someway down this route. Burawoy (1979) had shown how employee participation in production games and workplace rituals generated consent to existing social relations, constituting workers as individuals rather than as members of a class. In other words they 'helped' to control themselves. Later, Ray (1986) argued that control by corporate culture was the last frontier, in that it had enabled organizations to generate sentiment and emotion, simultaneously internalizing control and linking personal with corporate identity.

A further twist to such arguments has come through more recent influences from Foucault and poststructuralism. At the height of the attention paid to corporate culture in the early 1990s, theorists accepted the intent and outcomes of managerial action but redescribed it in critical terms, referring to the closing down of rival cultural spaces and the production of 'designer' employees (Casey, 1995). The specifically Foucauldian influence can best be seen in some influential accounts of changes in factory work. Sewell (1998) claimed to have identified a new model of labor process control that counters the optimistic gloss of the empowerment and teamwork literature, while moving beyond the confines of traditional LPT. 'Panoptic' forms of control augment existing practices through the interaction of 'vertical' electronic surveillance and the 'horizontal' dimension of peer group scrutiny within teams. Sewell made much of the enhanced capacity of management to collect, display, and attribute performance data. With respect to teams, he relied on other cases, including McKinlay and Taylor (1996) to make a case for horizontal control. Here, team members create the intersubjective conditions for compliance by creating disciplinary rules within delegated authority, thus collaborating with management to identify and reward the 'good worker.'

The implication of Sewell's argument is that such a combination solves the 'control-engage' dilemma identified by LPT and suffered in different ways by generations of managers. The new techniques allow management to rationalize knowledge and work, but not through obtrusive Taylorist and bureaucratic means, while mobilizing and manipulating the subjectivity of employees. The rapid growth of call centers and automated call distribution systems also provided plenty of scope for Foucault-influenced commentators to reinforce a focus on the increased surveillance potential of the electronic panopticon. Some commentators became convinced that the call center represented the ultimate and total control system. The capacity of the computer monitoring

screen to not only direct tasks, but to get inside the heads of Customer Service Representatives (CSRs) rendered supervisory power both 'perfect' and its actual use unnecessary.

Although the focus on culture change has faded in tougher times, poststructuralist writers have continued to highlight the significance of identity in exploring the continuance of normative controls through managerial discourses; human resource management practices, such as induction, training, and promotion (Alvesson and Willmott, 2002); and the more recent emphasis on employees 'living the brand.' It is argued that such approaches are particularly pertinent in knowledge-intensive firms where employees are not amenable to direct or task-based controls. McKinlay has argued that "[t]he primary means of managerial control of knowledge work is the regulation of employees self rather than work flows or task" (2005: p. 245).

Changes in the terrain and frontier of control take place continually, and it would be difficult to deny that *some* shift toward electronic and self-surveillance has taken place in *some* industries. Similarly, while explicit culture change programs may have gone out of fashion, increased attention does seem to have been placed in *indirect* forms of control that would include the use of intensive recruitment methods. In putting more effort into recruiting the 'right people' with compatible attitudes and orientations, there would be less need for direction of work tasks.

A range of objections can and have been made, however, to these twists on accounts of organizational control (Thompson and Ackroyd, 1995). Few of its enthusiasts adequately consider issues of time, place, and people – in other words, the generalizability of such controls outside of the high-tech, strong culture firms from which the evidence largely comes. The central problem, however, is the confusion of the formal capacities of technological and managerial systems with their actual usage and effectiveness. This is compounded by a tendency to draw evidence primarily from managerial sources, underestimating the capacity of labor to find individual and collective spaces to dissent or resist. Discourse and identity writings are still light on actual evidence of high commitment and internalized corporate norms (Leidner, 2006). In cases in which these problems are considered more fully, as in the McKinlay and Taylor case study of an electronics plant in Scotland, a more rounded and complex picture of the gap between intent and outcome is revealed. Finally, it is not only debatable how internal*ized* the new controls are, but also how *internal*. Normative rules in strong culture companies may be different from traditional task-based control, but they still are rules. Management's sources and use of information may have increased, but there is little evidence that is unobtrusive or regarded by employees as objective and unconnected to visible authority relations. This is strongly confirmed in most research on call centers. Most studies show that employees are able and willing to question the purposes and accuracy of 'the stats' and to find ways to both evade some of the pressures of call monitoring and locate blind spots in the system.

One of the problems with shifting locus arguments is the implication that shifts are substantive, if not permanent. Given the highly dynamic nature of the frontier of control, it is equally possible that traditional controls may reemerge, albeit in some new forms. This is what appears to be happening as large organizations in the private and public sectors are subject to heightened competitive and financial pressures. Performance management appears to be an increasingly central focus of managerial attention. It is unusual for most employees not to be faced with a plethora of targets, key performance indicators, and related controls in the context of external and internal restructuring and cost pressures. In increasingly 'lean' environments, tighter and more pervasive systems of monitoring and surveillance, supported by more visible accountability and stringent disciplinary procedures are more common.

The previous era partial shift to consent making may have been eroded by the principles and practices of 'market rationalism' in the context of shareholder-driven economic pressures. Normative interventions promoting commitment and cultural change are being marginalized (Kunda and Ailon-Souday, 2005). Compliance rather than commitment may suffice in circumstances in which organizations and their workers are subject to enhanced market discipline and insecurity (McGovern et al., 2007).

Questioning the Centrality and Scope of Workplace Control

Traditional controls may persist, but another kind of locus argument has been raised by recent research. Models such as those associated with Richard Edwards conceptualized developments largely in terms of managerial strategies within large companies pursued at the workplace or plant level.

Friendly critics of LPT always raised questions as to whether the centrality given to control of labor is actually reflective of managerial behavior. It was argued that we cannot view management strategies and tactics primarily from the vantage point of the labor process. If anything, managers are dominated by problems of the outcomes of that process, including sales, marketing, supply, and cash flow. The concept of *the full circuit of capital* is used to argue that we must be concerned not only with the *extraction* of surplus value through controlling the labor process, but its realization through the sale of commodities, as well as the prior purchase of labor.

It was widely accepted within LPT that some of these objections were a necessary corrective measure to a debate that often had lost sight of the role the broader political economy because of the emphasis on *management* control. Such concerns have been amplified by more recent developments signaling a move away from managerial capitalism and the emergence of a more financialized economy with more complex, fluid interorganizational relations that raise issues about the changing nature of the workplace. Contemporary organizational and employment structures often are dispersed and disaggregated. Controls thus operate at an interorganizational level, across conglomerates, production networks, or supply chains. Some traditional analytical tools are ill-equipped for understanding this. A good illustration of these complexities is the retail supply chain. Prominent examples such as Wal-Mart illustrate retailers' growing ability to transfer risks and costs to suppliers, and to exercise controls indirectly or directly over aspects of their operations. This is facilitated by investment in retail technology and information technology systems that enable retailers to manage and monitor information and goods within the overall supply chain.

Supply chains are just one form of more complex interorganizational relations. As corporations get larger, their activities become more diverse, and their goals become more shareholder driven, such 'directed' networks exercise significant power at the center that constrains unit autonomy (Ackroyd, 2002). Studies report a transfer of powers from local, unit-level managers to corporate executives who are more directly accountable to shareholders. Managers are under continual pressure to cut labor costs and restructure operations to meet capital market requirements (Appelbaum et al., 2013). This adds up to a strengthening of financial control systems across a variety of organizational boundaries. Information and Communications Technology (ICT) systems facilitate the capacity of corporate management to operate sophisticated surveillance mechanisms and resource allocation models. These frequently lead to a cascading down of monitoring and target setting for performance measures that tighten controls for both managers and workers alike, albeit in different ways.

Some researchers argue that new loci for value creation and decision making challenges work-based control perspectives traditionally favored by LPT and employment relations scholars. Reedhas made a similar point: "The major focus of control has moved away from the design and monitoring of standardized work tasks toward the management of highly complex informational, human and material 'flows'" (2010: p. 42). These are important points, although LPT already has been moving away from workplace-centric control models. It also is the case that the centrality of labor costs within restructuring and value creation means that broader changes are highly likely to have consequences for workplace controls.

Conclusion

Management always may be looking for the final frontier, but like the crew of Star Trek, they never really find it. This is hardly surprising. Conflicts of interest and constraints of circumstance run too deep. A close examination of contemporary organizational trends and their analysis reveals considerable continuity in theory and practice. Just as newer normative controls do not necessarily replace traditional task-based ones, extraworkplace mechanisms add to rather than displace intraorganizational forms. Sometimes theorists are describing the same practices in a different conceptual language. Electronic surveillance can look remarkably like Edward's concept of technical control; the exaggerated claims for empowerment strongly resemble those made for work humanization in the 1970s. Willing internalization in one framework can impose normative controls in another. It would be foolish to claim that nothing changes. Emphases within the repertoire of controls change, particularly as the wider structure of economy and society is transformed. Broadening the scope of controls in the new interorganizational landscape is a prime example. Reed (2010) has attempted to sum up these trends by referring to new control logics – complex, multilevel hybrids that combine the hard and soft dimensions from electronic surveillance to cultural engineering and performance management. Nevertheless, while work remains a contested terrain, there will be a frontier of control. For example, changes in the labor force and broader cultural norms can bring new problems, such as the rising tide of sexual misbehavior at work. In turn, this creates the conditions for new controls, such as codes of conduct that specify appropriate boundaries of behavior and what sanctions might follow if they are transgressed. It is the task of organizational analysis to explain the character, combination, and circumstances of that changing frontier.

See also: Control, Social; Corporate Finance: Financial Control; Foucault, Michel (1926–84); Interorganizational Relationships and Networks; Technology and Social Control.

Bibliography

Ackroyd, S., 2002. The Organization of Business. Oxford University Press, Oxford.
Adler, P.S., 1993. Time-and-motion regained. Harvard Business Review, January–February, 97–107.
Alvesson, M., Willmott, H., 2002. Identity regulation as organizational control: producing the appropriate individual. Journal of Management Studies 39 (5), 619–644.
Appelbaum, E., Batt, R., Clark, I., 2013. Implications of financial capitalism for employment relations research: evidence from breach of trust and implicit contracts in private equity buyouts. British Journal of Industrial Relations 51 (3), 498–518.
Armstrong, P., 1989. Management, labour process and agency. Work, Employment and Society 3 (3), 307–322.
Braverman, H., 1974. Labor and Monopoly Capital: the Degradation of Work in the Twentieth Century. Monthly Review Press, New York.
Burawoy, M., 1979. Manufacturing Consent: Changes in the Labor Process Under Monopoly Capitalism. University of Chicago Press, Chicago.
Casey, C., 1995. Work, Self and Society: After Industrialism. Routledge, London.
Edwards, R., 1979. Contested Terrain: The Transformation of the Workplace in the Twentieth Century. Heinemann, London.
Friedman, A., 1977. Industry and Labor: Class Struggle at Work Monopoly Capitalism. Macmillan, London.
Graham, L., 2005. Manual workers: conflict and control. In: Ackroyd, S., Batt, R., Thompson, P., Tolbert, P. (Eds.), Oxford Handbook of Work and Organization. Oxford University Press, Oxford, pp. 20338–20355.
Hochschild, A.R., 1983. The Managed Heart: Commercialization of Human Feeling. University of California Press, London.
Kunda, G., Ailon-Souday, G., 2005. Managers, markets and ideologies: design and devotion revisited. In: Ackroyd, S., Batt, R., Thompson, P., Tolbert, P. (Eds.), Oxford Handbook of Work and Organization. Oxford University Press, Oxford, pp. 200–219.
Leidner, R., 2006. Identity and work. In: Korczynski, M., Hodson, R., Edwards, P. (Eds.), Social Theory at Work. Oxford University Press, Oxford.
McGovern, P., Hill, S., Mills, C., White, M., 2007. Market, Class and Employment. Oxford University Press, Oxford.
McKinlay, A., Taylor, P., 1996. Power, surveillance and resistance: inside the factory of the future. In: Ackers, P., Smith, C., Smith, P. (Eds.), The New Workplace and Trade Unionism. Routledge, London.
Pfeffer, J., 1997. New Directions for Organizational Theory: Problems and Practices. Oxford University Press, Oxford.
Ray, C.A., 1986. Corporate culture: the last frontier of control? Journal of Management Studies 23 (3), 287–297.
Reed, M., 2010. Control in contemporary work organizations. In: Blyton, P., et al. (Eds.), Re-assessing the Employment Relationship. Palgrave Press, Basingstoke, pp. 41–70.
Ritzer, G., 1993. The McDonaldization of Society. Pine Forge Press, London.
Sewell, G., 1998. The discipline of teams: the control of team-based industrial work through electronic and peer surveillance. Administrative Science Quarterly 43 (2), 406–469.
Thompson, P., Ackroyd, S., 1995. All quiet on the workplace front? Sociology 29 (4), 19–33.
Thompson, P., van den Broek, D., 2010. Managerial control and workplace regimes: an introduction. Work Employment and Society E-special 1.
Walton, R.E., 1985. Towards a strategy of eliciting employee commitment based on policies of mutuality. In: Walton, R.E., Lawrence, P.R. (Eds.), Human Resource Management, Trends and Challenges. Harvard University School Press, Boston, Mass.

Organizational Culture, Anthropology of

Jakob Krause-Jensen and Susan Wright, Danish School of Education, Aarhus University, Copenhagen, Denmark

Published by Elsevier Ltd.
This article is a replacement of the previous edition article by W. Koot, volume 16, pp. 10934–10938, © 2001, Elsevier Ltd.

Abstract

Anthropologists have studied organizations since the 1930s. In the 1980s, anthropologists' concepts of culture were instrumentalized by nonanthropologists to promote 'organizational culture' as a management tool. In subsequent decades, concern shifted to welding employees from different 'national cultures' into transnational corporations and organizations concerned with international governance. In such organizations, anthropology graduates are increasingly employed as 'cultural experts.' We track the anthropological research on organizational culture and argue that the sensibilities and analytical skills acquired and the concepts developed through the ethnographic encounter gives anthropology a unique voice in the study of cultural matters in organizations.

Introduction

'Organizational culture' came into vogue as a management tool in the late 1970s. Anthropologists had been engaged in research on industrial and other organizations since the 1930s, but these studies took a new turn when managers, consultants, and organizational researchers drew on anthropological concepts of culture in the process of devising and managing the culture of their organizations. Anthropologists found their analytical concept of culture was implicated in the phenomenon they were trying to study empirically. Anthropologists try to maintain a distinction between their own analytical (etic) concepts and the (emic) concepts used in the everyday life of the people they study, whereas now they found the emic and etic were entwined.

This article will focus on anthropological studies of 'organizational culture' as a tool of management and a *raison d'etre* for organizations. The first section gives a brief history of anthropological approaches to organizational culture. The second section outlines the rise of 'organizational culture' as a key management concern. It considers the 'Japanese connection,' the advocates of 'organizational culture' in the United States, notably Peters and Waterman, and Schein, and the further transition into 'values-based management.' We explore how exponents of these approaches to management drew on concepts and understandings of culture that were often at variance with contemporaneous anthropology. The third section addresses managing culture in transnational corporations and the fourth section concerns working with culture in organizations that are engaged with international governance, notably the European Union, which tried to develop a 'European culture,' and the United Nations Educational, Scientific and Cultural Organization (UNESCO), which engages with multiple concepts of culture in preserving cultural heritage and promoting a 'culture of peace' in the world. The final section sums up the argument by emphasizing the critical use of anthropological concepts and strategies in organizations. This is especially important given that employment in organizations as 'cultural experts' is a growing field of work for anthropologists.

Anthropological Approaches to Organizational Culture

It is common to date the beginning of anthropologists' involvement with industrial organizations to the Hawthorne Project (1927–32). This project was led by Harvard psychologist Elton Mayo and challenged Taylorist management strategies by arguing that managers should take into account the emotional needs of employees, including their need to belong to a group or a team. He knew two prominent anthropologists, Radcliffe-Brown and Malinowski, and one of Malinowski's students, Lloyd Warner, was involved in designing and carrying out the observations. Those studies were taken further by British anthropologists who employed participant observation in a series of shop-floor studies in Manchester in the 1950s and 1960s (Wright, 1994: 6–13). They explored in great detail the strikingly different ways that workers organized themselves in comparable factories. Drawing on the Manchester School's sophisticated analyses of conflict, notably in sub-Saharan Africa, these studies generated a range of analytical concepts for a much more nuanced understanding of the class conflicts that marked British industrial relations at the time.

In the Hawthorne factory studies, anthropologists tended to take the managers' rationality as a given and 'study down.' Perhaps even more than in the colonies (Loizos, 1977), they investigated the everyday life and ways of thinking of the workers in order to explain these to managers (Wright, 1994: 8). The Manchester shop-floor studies adopted some of the radical promise of modern anthropology, by presenting workers' ways of organizing themselves as rational and worthy of respect, even if not in accord with those who managed and governed them. But it was only in the 1970s that anthropologists turned their attention to studying the elites themselves, their ways of governing or managing, and their claims to define 'culture' for those they governed. A surge in studies of nationalism explored the ways that elites invented national 'traditions' (Hobsbawm and Ranger, 1983) and as Gellner (1983) put it, constructed a political and symbolic formation that gathered one people, one ethnicity, with a common language, shared attributes, and sense of belonging under the one

'political roof' of a modern nation-state. This focus on the *constructed* nature of national culture coincided, in England at least, with an emerging middle class among descendants of postwar migrants and new immigration from a worldwide movement of people from the global south to the global north, including a diasporic intelligentsia. This indigenous and imported black and Asian middle class unsettled the English elite's assumption that they had the right to define national culture and the age-old canons of "the best which has been thought and said in the world" (Arnold, 1869: viii). Just at the moment when this emergent force claimed that everyone had the right to contest and contribute to the core concepts of Englishness, management studies developed the idea that a managerial elite could devise a company's culture, as a set of values and behaviors that they would expect all workers to share and demonstrate in their daily work culture as a top-down and uncontested tool of management.

By the 1990s and into the new century, there were four major meanings of 'culture' in play. First, national culture was projected as a shared history, common adherence to key institutions and symbols, similar ways of behaving and communicating, and a sense of belonging. Although moments of crisis (for example, the Rushdie Affair, Asad, 1993) exposed the work of elites in maintaining this construction, their success was evident in the popular assumption that it was natural – even God-given – to see the world as made up of such nation-states. A second notion of culture focused on 'communities' in smaller localities or 'subcultures' of ethnic minorities, working class, or youth. Largely dismissed by anthropologists, but still used widely by politicians, in the media and in everyday discourse, this idea of culture portrayed members of a 'community,' 'people,' or 'subculture' as homogeneous, as if any member could be 'known' by a checklist of characteristics and as if they uniformly and unchangingly shared the same values and ways of behaving. Third, and in contrast to both of the meanings above, in a new approach in anthropology and cultural studies, culture was no longer seen as consensual but as a site of contestation. Within any nation or community, there would be a political struggle between different individuals and interest groups over who had the power to define the core characteristics of a culture, how other individuals or interest came to be dominated or subsumed, and with what material consequences (Wright, 1998a). Fourth, as we will see below in the discussion about UNESCO, a pluralist view of everyday life conceived of individuals or small groups migrating and living in a variety of 'mixed communities' in the course of their lives. In the process, they collected ideas, values, and ways of living from multiple sources, from which each assembled an idiosyncratic bricolage. In this view, culture is not shared and predictable, but negotiated and contested in a continual process of creation.

It was especially the first two concepts of culture that appealed to management and came to provide a fragile and ambiguous connection between anthropology and the world of business. Notably, a single quotation from the anthropologist, Clifford Geertz, was to be found in numerous management publications and consultants' power points:

> Believing, with Max Weber, that man (sic) is an animal suspended in webs of significance he himself has spun, I take culture to be those webs, and the analysis of it to be therefore not an experimental science in search of law but an interpretive one in search of meaning.
> Geertz, 1973: 5

The article from which this is drawn analyzed how three people involved in an incident had very different understandings of a sequence of actions. Geertz showed that each was weaving a different web from the same threads and each had different structural power and personal ability to impose their meaning on the others, with the dominant one accruing substantial material outcomes. This view of culture as a political and contested process is akin to the third meaning of culture above. Yet in management studies, this quote was used to argue for managers' generating a single vision for the organization, implemented from the top, so as to align all the informal structures and practices of all the departments into one 'web of meaning' or corporate culture. This smacks of the second meaning of culture, and was precisely the meaning of culture that Geertz was writing against. Anthropologists found their concepts used in management in such disconcerting ways. As George Marcus put it: "For anthropologists especially there will be a particular irony in this migration of their signature concept appropriated and transformed in a sphere of their own society quite foreign to them — that of business — and without any consultation or credit either" (Marcus, 1998: 8). In the 1980s, when the culture concept (in the second sense of shared values and behavior among a group of homogenous people) was just becoming the greatest conceptual export success in the history of anthropology, anthropologists had already turned their backs on it. The problems and dangers of the politicization of the culture concept have received so much attention within anthropological discourse (Wright, 1998a) that some discarded it as an analytical tool altogether (Abu-Lughod, 1991).

The Rise of 'Organizational Culture'

The Japanese Connection

It was the staggering Western economy compared to impressive Japanese growth rates that initially sparked the interest of Western firms in establishing 'corporate cultures.' In his business bestseller from 1981, *Theory Z. How American Business Can Meet the Japanese Challenge*, the Japanese-American organization researcher William G. Ouchi (1981) argued that the deepest reason for the Japanese *Wirtschaftswunder* was not technological inventions or tight financial policy. The root cause was the Japanese culture: what characterized Japanese companies was the loyalty of their employees and the close social relations between colleagues. Contrary to Weber's (1947) idea that bureaucracy was built on the absolute separation of the personal from the professional, according to Ouchi, in Japan, the personal and the professional met in a sublime and productive synthesis. Such depictions of synthesis between national and individual 'character traits' in Japan have a long legacy, starting with Benedict's (1946) treatment of Japanese wartime ideology as the depoliticized and ahistorical cultural characteristics of the population. For Ouchi, the Japanese tradition to be personally involved in the workplace was founded in an unselfish discipline, which is

rooted in traditional Japanese virtues of 'intimacy' and 'trust' (Benedict, 1946: 9). The implication of Ouchi's study for the West was, of course, that the desired corporate culture (in the second meaning of culture, above) was not isomorphous with national culture (in the first meaning, above) and thus did not grow organically from existing social forms. If the Japanese growth depended on an essential 'Japaneseness,' then how could it be transmitted? How could Western business leaders learn from it? In Europe and the United States, corporate culture had to be a 'project'; organizational culture had to be engineered.

US Corporate Culture

In the beginning of the 1980s, a number of influential US business bestsellers appeared. They all argued that organizations were not driven simply by concerns for hard-line efficiency, but should rather be seen as collective worlds governed by values and meanings (Peters and Waterman, 1982: 56, 61). Peters and Waterman's bestseller *In Search of Excellence* argued that the recipe for company success was a corporate culture. Managers and consultants started to pay attention to their organizations' collective identities: they developed symbols, enacted rituals, and identified corporate cultures to motivate and inspire employees (Deal and Kennedy, 1982; Morgan et al., 1983). Culture also became an important part of the curriculum in management studies, be it human resource management or cross-cultural management.

In organizational studies, Edgar Schein is the doyen of organizational culture and his influence can hardly be exaggerated. His book *Organisational Culture and Leadership* (1986) is probably the single most important book in the research field and has inspired both managers and other researchers working with 'organizational culture.' Schein's relationship to anthropology is ambivalent. On the one hand he draws a contrast between his own 'clinical-consultant' approach to organizational culture and the ethnographer's participant observation (Schein, 1992(1985): 47–48). On the other hand he draws on the "complex notions of culture of anthropologists" (Schein, 1992(1985): 3) and their ability to 'find' and 'uncover' culture in people's everyday lives to mark himself off from 'culture-vulture'-consultants and their account of organizational values as something that can be readily invented by managers and implemented in organizations. Instead, he develops a three-tiered model of culture as 'artifacts,' 'espoused values,' and 'tacit assumptions,' putting his greatest emphasis on this last level of unconscious 'basic underlying assumptions.' It is this last level that defines his celebrated conception of culture as

> a pattern of shared basic assumptions that the group learned as it solved its problems of external adaptation and internal integration that has worked well enough to be considered valid and therefore, to be taught to new members as the correct way to perceive, think, and feel in relation to those problems.
> Schein, 1992(1985): 12

Schein's insistence on using 'complex anthropological notions' rings a bit odd, though, as this conception resonates with the homeostatic models of culture (the second idea of culture, above), which most anthropologists had left behind before Schein wrote his book. Schein's ideas of organizational culture as a set of fundamental assumptions that guide behavior and are reproduced between generations of workers seem empirically implausible in large organizations, where people occupy different hierarchical positions. They are even less credible in today's turbocharged labor markets, where people change jobs and organizations with increasing frequency (Krause-Jensen, 2010: 56–68).

It is quite curious that one of the most widely read and classic examples of an ethnographically grounded analysis of organizational culture that chimes with the third anthropological approach to culture (above) was authored by one of Schein's students, Gideon Kunda (1992). In the 1980s, he did a detailed study of the corporate culture project of a US-based Hi-Tech firm, 'Tech.' Kunda analyzed the company's efforts to instill a corporate culture as an example of soft normative control, and he drew on a range of theories, including about ritual, to understand how corporate culture was communicated, what interests and ideas lay behind it, and how members responded to the efforts. The main ethnographic site of his fieldwork was the engineering division of the company, which pointed to the double entendre of the book's title, *Engineering Culture*. What people came to share was not a corporate culture, but an ambivalence about their involvement: the harder the company tried to engineer its culture and propagate organizational ideology through seminars and corporate newsletters, the more the members responded with resentment and cynicism. Their sense of self as engineers – their own 'engineering (sub)culture' – grew from their collective if discreet dissociation from the espoused corporate culture.

Value-Based Management

As described above, the notion of culture has been central to human resource management as a relay between the individual and the organization. The central aim behind establishing an organizational culture is to make a bridge between employee concerns and aspirations, and the strategic aims of the organization. Kunda and Van Maanen (1999: 65) argue that there used to be a cultural contract or social understanding behind the 'corporate culture project': "Loyalty in return for a career, commitment in turn for an identity, dedicated performance in return for meaningful work." However, since the 1980s, continuous waves of reengineering, downsizing, and outsourcing have seriously challenged the premises on which this contract rested. With an increasingly mobile labor force, the project of training every employee in the corporate culture through away days and training camps became too expensive.

In his book on the Danish company Bang & Olufsen, anthropologist Jakob Krause-Jensen traced how the word 'culture' largely vanished from the organizational strategies and was supplanted by 'value-based management.' This new concept continued the company's attempts to align individuals with corporate strategies. Krause-Jensen explored the semantic shifts involved. 'Culture' fell out of favor because it carried cumbersome connotations of tacit understandings and norms slowly accumulated through stable communities living and working together. This idea was out of synch with the conditions of flexible capitalism. 'Values' did not imply socialization in the same way as culture. Belonging to the realm of ideas,

values were easier to disseminate. High-ranking managers and consultants identified the fundamental values of Bang & Olufsen as 'poetry,' 'synthesis,' and 'excellence.' They claimed to have 'discovered' the organization's inherent values, rather than inventing and imposing them. They therefore expected these values to percolate through every aspect of the employees' working lives and to resonate with the consumers. Previously, 'corporate culture' had been the domain of the human resources department, whereas the marketing department joined in the project to identify and promote the company's values as part of establishing the corporate brand. Within the organization, however, this project did not achieve the aim of creating a common platform of shared values. No sooner had the values been excavated and highlighted than employees in different departments began to contest them or dismiss them as too remote from their working reality.

Managing Culture in Transnational Organizations

Global capitalism has spurred another strategic interest in culture in organizations. The growth of multinational organizations has been accompanied by attempts to understand and manage national cultural differences within them. A particular and very influential way of conceptualizing and working with national cultures in organizational studies can be seen in Geert Hofstede's *Culture's Consequences* (1984). Hofstede used questionnaires distributed in the sales organizations of IBM in the 1970s to classify different nationalities along five value dimensions that he calls "power distance, uncertainty avoidance, individualism, masculinity, pragmatism and indulgence" (http://geert-hofstede.com). According to Hofstede, culture should be understood as the 'collective programming of the mind' (1984: 25). Although the book challenged the idea that American management practices could be implemented unproblematically across the globe, anthropologists will find many reasons to question its methods, and challenge its concepts and conclusions. Hofstede's approach and view of national culture as an immutable essence that can be isolated, measured, and condensed into a limited number of values clashes with some basic convictions of most fieldworking anthropologists for whom thoughts and practices cannot be properly understood outside the context in which they are situated.

Anthropologists have developed understandings of national differences and how they are articulated in transnational organizations that are very different from Hofstede's. In the 1990s, Swedish anthropologist Christina Garsten did a multisited ethnography of Apple Inc. Through fieldwork in three different sites – Stockholm, Paris, and Cupertino, where Apple's headquarters are located – she analyzed how the company attempted to mobilize employees through defining itself against the titans of the industry like IBM. They projected an image of themselves as a transnational corporation with a coherent identity marked by a countercultural zest, lightness, and entrepreneurialism. This study is an example of how a very strong corporate culture is modulated, ironically embraced, or differently interpreted in national subsidiaries at various distances from the corporate center.

Anthropologist Tomoko Hamada also developed an approach quite different from Hofstede's when she analyzed how national cultural differences are played out in a US subsidiary of a Japanese company. Literally an 'extended case study,' she described events surrounding a sexual harassment lawsuit from the standpoints of the various people involved. The point of her analysis is that national differences were important, but certainly not the only relevant distinctions in the organization. Class, gender, and race were equally important and the participants' perceptions of their national differences underwent decisive changes as the case developed.

A third example is Galit Ailon's (2007) study of a successful Israeli-American Hi-tech merger. She directly challenges Hofstede's ideas that every nation is regulated by an essential set of values. Through detailed descriptions of everyday work life, the book shows how the Israelis' sense of identity is not *a priori* but is sparked by encounters with the American merger partners. She treats national identity as "a symbolic resource that assumed a potent role in the social production of a *sense* of difference," rather than an immutable essence (Ailon, 2007: 138).

Working with Culture in Transnational Organizations

The notion of culture has also played a crucial role in organizations concerned with inter- and supranational governance. Anthropologist Cris Shore (2000) has studied the European Commission's attempt to create 'ever closer union among the peoples of Europe.' Crucially, this has involved trying to create a public or political community with a common sense of belonging to 'Europe.' To overcome the 'democratic deficit' and build the cultural bedrock of such a democratic regime, the Commission engaged in a range of cultural policies. These aimed at transferring popular loyalty from the nation-state and engineering a love of Europe as a *patria* and a European consciousness in people's everyday lives. They tried to generate an idea of a shared European heritage and common contemporary culture through the creation of a flag, an anthem, European statistics, EU citizenship, and a common currency, with banknotes and coins designed as icons and symbols of new Europe. Simultaneously, they tried to forge a sense of Europeanness among the national civil servants appointed to the Commission. It was expected that, by working together, there would be an 'enmeshing' (*engrenage*) of the European elites: they would shift their ideals and loyalties to build an integrated system of public administration and a supranational institution that would become the custodian of the European ideal. However, in practice, a system of informal networks, patronage, and backdoor recruitment developed, until an investigation found an 'organizational culture' marked by widespread fraud, mismanagement, and endemic corruption.

UNESCO has no such problems with legitimacy: it is known as the intellectual arm of the United Nations system and its Conventions and Declarations wield considerable moral authority among member states. With its overordinate aim to foster world peace, UNESCO uses the same key concepts of culture to order both its internal operations and to mobilize its participants throughout the world. But the meaning of 'culture' has gone through several transformations (Nielsen, 2011; Wright, 1998a,b). Initially, the conception of culture was like Arnold's, as the canon of great art and architectural and archaeological monuments with universal value, which would be disseminated by an intellectual and political elite through

UNESCO. Instead of being linked to national identity (as in version 1 above) in the founding charter, culture was seen as a universal route to human perfection. Culture would "alter … the minds of men," create "intellectual and moral solidarity" and shore up peace. In the 1990s, there was finally a shift to a demotic idea of culture, but organized in 'archipelagos' (Eriksen, 2001), with each 'people' having its own culture (as in version two above). Now the promotion of cultural diversity, respect for difference, and the rights of minorities would be the route to peace. However, this relativism was tempered by an implicitly Western normative judgment that some cultural practices, like genocide or honor killings, should be condemned (Wright, 1998b: 179–180; Nielsen, 2011). Cultural diversity soon morphed into the fourth, pluralist, idea of culture. With the global flows of people, images, ideas, and goods, everyone should be able to enrich their lives by choosing and mixing elements of their own and others' cultures. Culture comes to be seen as a creative capacity in the hands of people themselves, a dynamo for change, a source of freedom, through which individuals can respond with imagination and vision to changes in the world. Successful development would be evidenced by a 'flowering of culture' (UNESCO, 1996) and a world 'culture of peace.'

Each of these transitions is the culmination of long processes of contestation over the meaning of culture (as in version three, above), but new meanings are never closed and final; the older meanings are always available and kept in play. Where in UNESCO do these contests over the organization's meanings of culture take place? Nielsen's (2011) study of the UNESCO bureaucracy shows that this is not the site of such contestation. A successful UNESCO bureaucrat is one who writes documents by sustaining the ambiguity of keywords like diversity, creativity, tolerance, freedom, and culture (Nielsen, 2011: 282). Bureaucrats take care not to define or dislodge the precarious meanings of 'resource-generating key words' like 'culture of peace,' which are at the center of projects reliant on member states for funding (Nielsen, 2011: 284). Instead, bureaucrats assemble meetings of academic 'experts' and nongovernmental organizations to present and discuss contested ideas of culture. It is in the regional and world conferences to develop new instruments of governance, for example, on Cultural Diversity or Intangible Cultural Heritage, that these contestations over the meanings of key words come to a head, not least in the all-night sessions where government delegates draft an Action Plan or a new Convention (Wright, 1998b). The major project led by Bendix et al. (2013) shows that as soon as a Convention, such as that on Intangible Cultural Heritage is ratified by a state, a whole new assemblage of actors – governmental, nongovernmental, commercial, professional, and community – begin to contest the ways that particular concepts and instruments will be used in a specific context. Their study traces who gets to define UNESCO's meaning of culture, for whom, and with material consequences, in places far distant from UNESCO's Paris headquarters.

Conclusion

These analyses of organizational culture do not result in a checklist of the key characteristics of an anthropological approach (we are not trying to turn anthropology itself into a culture *à la* version two above). However, it is possible to identify some family resemblances, which have grown out of anthropology's disciplinary legacy of cultural critique and its ethnographic methodologies.

As demonstrated above – with the risk of exaggerating a bit – it is possible to distinguish anthropology (with its concomitant ethnographic research strategies) from mainstream perspectives on organizational culture (Gellner and Hirsch, 2001; Ybema et al., 2009). A major difference is that much nonanthropological description of culture shares an appetite for 'how to,' which often leads to treating culture and values as 'variables' that cause or regulate behavior in uniform ways. There is a huge difference, however, between the lived experience of working in organizations, with their ambiguity and confusion, and the tidy management texts and reports on 'organizational behavior.' A major tenet and ambition of anthropology is to capture those experiences and understand them in the context in which they are situated.

The anthropologists who study organizations ethnographically try to get behind the common sense and official, managerially sanctioned worldviews often implied in the ideas of 'an organizational culture.' Such notions would often emphasize convergence, sharedness, similarity, and normative consensus. Anthropologists, on the other hand, do not assume consensus and regard sharedness as an empirical question. Researchers in organizational behavior tend to see organizational culture and values as the description of a deeply sedimented, communal identity, whereas anthropologists regard it as an often ambiguous and contested social accomplishment tied to particular (managerial) interests. As a consequence, anthropological/ethnographic analysis of organizational culture will often be in the form of a 'thick description' rather than distilling 'factors' and isolating 'variables' to predict outcomes. An important undertaking for anthropologists and ethnographers who work with organizational culture has been to look into the ideas and the practices surrounding culture projects and to do a cultural critique (in the enlightenment sense) of the culture concept and the way it has been understood and put to use in organizations and among management researchers.

A growing number of anthropologists work in organizations outside academia. They are sometimes employed in a consultant role as 'culture experts' able to grasp the otherwise intangible workings of the organization. Currently, many anthropologists are also employed in organizations as 'cultural experts' in another sense. Through their ethnographic methods they are expected to be able to uncover consumer culture: to identify customers' social aspirations and the contexts within which they use products. In particular, anthropologists are finding employment in user-based design from General Motors to Xerox's copying machines (Cefkin, 2009: 4–5). Previously, the concept of culture was largely used by consultants from a nonanthropological background, whereas now, anthropology graduates are employed to use the concept themselves in organizations. Such anthropologists face on a daily basis the challenge of using and developing 'culture' as an analytical concept to gain a distanced and critical purchase on the way ideas of 'culture' are used ethnographically in the management and everyday life of organizations.

See also: Anthropology at Home; Business Anthropology; Corporate Culture; Cultural Critique, Anthropological; Culture, Sociology of; Ethnography; Expertise, Anthropological Approaches to; Intercultural Communication: An Anthropological Perspective; Myths and Symbols: Organizational; National Character; Nongovernmental Organizations: Anthropological and Historical Aspects; Organizational Behavior, Psychology of; Organizational Climate; Organizations and Culture; Work, Anthropology of.

Bibliography

Abu-Lughod, L., 1991. Writing against culture. In: Fox, R. (Ed.), Recapturing Anthropology: Working in the Present. University of Washington, New Mexico, pp. 137–162.

Ailon, G., 2007. Global Ambitions and Local Identities: An Israeli-American High Tech Merger. Berghahn Books, New York.

Arnold, M., 1993[1869]. 'Culture and Anarchy' and Other Writings. Cambridge University Press.

Asad, T., 1993. Multiculturalism and British identity in the wake of the Rushdie affair. In: Genealogies of Religion. Discipline and Reasons of Power in Christianity and Islam. John Hopkins University Press, Baltimore, pp. 239–268.

Bendix, R., Eggert, A., Peselmann, A., 2013. Heritage Regimes and the State. Universitätsverlag Göttingen, Göttingen.

Benedict, R., 1989[1946]. The Chrysanthemum and the Sword: Patterns of Japanese Culture. Houghton Mifflin, Boston.

Cefkin, M. (Ed.), 2009. Ethnography and the Corporate Encounter: Reflections on Research in and of Corporations. Berghahn Books, New York; Oxford.

Deal, T.E., Kennedy, A.A., 1982. Corporate Cultures: The Rites and Rituals of Corporate Life. Addison-Wesley, Reading, MA.

Eriksen, T.H., 2001. Between universalism and relativism: a critique of the UNESCO concept of culture. In: Cowan, J., Dembour, M.-B., Wilson, R. (Eds.), Culture and Rights: Anthropological Perspectives. Cambridge University Press, Cambridge.

Fox, R.G., King, B.J., 2002. Anthropology beyond Culture. New York University Press, New York.

Garsten, C., 1994. Apple World. Core and Periphery in a Transnational Organization. Stockholm University Press, Stockholm.

Geertz, C., 1973. The Interpretation of Cultures. Harper Collins, United States of America.

Gellner, D.N., Hirsch, E., 2001. Inside Organizations: Anthropologists at Work. Berg, Oxford.

Gellner, E., 1983. Nations and Nationalism. Cornell University Press, Ithaca.

Hobsbawm, E., Ranger, T., 1983. The Invention of Tradition. Cambridge University Press.

Hofstede, G., 1984. Culture's Consequences: International Differences in Work-Related Values. Sage, London.

Hofstede, G., 2013. The Hofstede Centre [Online]. Available: http://geert-hofstede.com (accessed 10.04.13.).

Krause-Jensen, J., 2010. Flexible Firm: The Design of Culture at Bang and Olufsen. Berghahn Books, New York, Oxford.

Kunda, G., 1992. Engineering Culture: Control and Commitment in a High-Tech Corporation. Temple University Press, Philadelphia.

Kunda, G., Maanen, J.V., 1999. Changing scripts at work: managers and professionals. The Annals of the American Academy of Political and Social Science 561, 64–79.

Loizos, P., 1977. Anthropological research in British colonies: some personal accounts. Anthropological Forum, 137–144.

Marcus, G. (Ed.), 1998. Corporate Futures: The Diffusion of the Culturally Sensitive Corporate Form. University of Chicago Press, Chicago.

Morgan, G., Frost, P.J., Pondy, L.R., 1983. Organizational symbolism. Organizational Symbolism 3, 35.

Nielsen, B., 2011. UNESCO and the 'right' kind of culture: bureaucratic production and articulation. Critique of Anthropology 31, 273–292.

Ouchi, W.G., 1981. Theory Z: How American Business Can Meet the Japanese Challenge. Addison-Wesley, Reading, MA; London.

Peters, T.J., Waterman, R.,H., 1982. In Search of Excellence: Lessons from America's Best-Run Companies. London, Harper & Row, New York.

Schein, E.H., 1992. Organizational Culture and Leadership. Jossey-Bas, San Francisco.

Shore, C., 2000. Building Europe: The Culture Politics in European Integration. Routledge, London.

UNESCO, 1996. Our Creative Diversity: Report of the World Commission on Culture and Development. UNESCO, Paris.

Weber, M., 1947. The Theory of Economic and Social Organization (A.M. Henderson, T. Parsons, Trans.). Oxford University Press, New York.

Wright, S. (Ed.), 1994. Anthropology of Organizations. Routledge, London.

Wright, S., 1998a. The politicization of 'culture'. Anthropology Today 14, 7–15.

Wright, S., 1998b. Encaging the wind. International Journal of Cultural Policy 5, 173–182.

Ybema, S., Yanow, D., Wels, H., Kamstee, F.H., 2009. Organizational Ethnography: Studying the Complexity of Everyday Life. Sage Publications, Thousand Oaks, CA.

Organizational Decision Making

Anna Grandori, Bocconi University, Milan, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Theory and research of organizational decision making (ODM) is organized in two parts: grouping contributions mainly concerned with the 'logic' and the 'politics' of ODM. Within each, the analysis develops in three movements. First, it points out how some classic 'divides' marked the field's origin, in relation to how uncertainty and conflict of interests may be dealt with. Second, it traces the key subsequent developments showing how the various models can be positioned on an ideal 'continuum.' Next, more recent works and research in relevant and related fields are considered, highlighting how they would allow to go beyond those classic continua and to indicate responses to increasing uncertainty and conflict implying 'stronger' rather than 'weaker' forms of thought.

Introduction

The field of organizational decision making (ODM) is not easy to define. As to the 'organizational' qualification, those who engaged in a conceptual definition have usually contrasted it with individual decision making, as ODM entails multiple individuals – whereby issues of communication and conflict may arise – and multiple occasions over time and matters – thereby excluding one-shot decisions among unrelated players – and have usually conceived it as decision making occurring 'in organizations,' intended as entities. However, those definitions are in many ways too narrow, with respect to both conceptual consistency and correspondence to the ODM actual empirical research tradition. In fact, to start with, any decision-making entity can be conceived as 'unitary' or not depending on the purpose of analysis. Individuals may not be such a unitary unit of analysis as often taken for granted: There can be conflict between objectives and 'multiple selves' also within an individual (Elster, 1985). Conversely, a multiperson system may be conceived as a unitary actor, if on a problem at hand, its behavior can be modeled 'as if' it were a homogeneous node of knowledge and preference – for example, a firm in the context of defining an alliance with other firms, or a functional unit in the context of defining a new product with other functions. Second, another important sense in which decision making can be qualified as 'organizational' is not related to the multiplicity of actors, but to whether an actor is deciding out of an organizational role or on a private basis (Barnard, 1938). Third, although earlier ODM research has been focused on processes going on within the institutional boundaries of organizations as entities (firms, public agencies, etc.), the main propositions and models are actually applicable to any ongoing cooperation, including interorganizational relations.

Hence, we shall adopt here a broad definition of ODM as both single actor and multiple actors, provided that decisions are taken in an organized context: ongoing relations among actors (not instantaneous transactions), but not necessarily internal to an entity, acting for some purpose of effectiveness (not just for leisure and pleasure). (In spite of its breadth, this definition leads to exclude from our analysis of ODM those 'phenomenological' traditions that have described all kinds of decision behaviors that may be observed in organization, without an interest in modeling their properties or conditions for effective or ineffective occurrence.)

As to the meaning of 'decision making,' the minimum common denominator of all notions can be found in the etymology of the term: 'de cedere' means cutting out, partitioning, discriminating, separating, and opting for some thing(s) rather than others. This activity may be performed in a deliberate and conscious way or not. In an organized context, deliberate and conscious action is of course particularly important, as effectiveness matters. Utility theory models coming from economics, as well as most models generated by the 'cognitive revolution' in interpreting human behavior are in fact of that type. Nevertheless, all deliberate decision-making processes are unavoidably embedded in other processes and factors that drive, constrain, and reduce the variety of possible actions that go beyond actor control and deliberation. On one side, those factors are embodied in the actors themselves, as 'taken for granted' and inherited background knowledge (Simon, 1990), or as 'normative' and 'affective' factors (Zey, 1992). On the other side, decision making is embedded in ecological processes, whereby it is the context/environment that in the end selects proper or viable actions, 'as if' actors' actions were random or blind variations. (To avoid confusion, it should be noticed that the term 'ecological rationality' has been used with various different meanings. In its strongest meaning, it has been used by organizational ecologists, in the sense that 'it is the environment that optimizes,' by selecting actions by means of selecting the subjects carrying them. In a second meaning, it has been contrasted with constructivist rationality (see Vernon Smith's Nobel Lecture) to refer to the emergence of 'patterned action without design' out of the combination of myriads of actions, in turn largely 'determined by cultural and biological inheritage.' Third, it has been taken to refer (e.g., by Gigerenzer and some related authors) to what in decision-making research (and in the present article as well) has always (and I think more properly) called 'contingent rationality': the deliberate choice by a subject, of a decision strategy deemed appropriate to (his model of) the environment/problem at hand (actually more a 'metaconstructivist' rather than ecological notion of rationality).) Actually, before the upsurge of cognitive sciences, decision behavior was predominantly interpreted in terms of instinct and affect or in terms of context-driven selection of actions.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.73076-5

Herbert Simon (1955) provided the most prominent foundations to the entire ODM field and put cognition center stage with respect to both 'ecological' rationality and to unconscious and emotional impulses. In his founding works, Simon, as well as some other important social scientists concerned with the foundations of human action (as Kelly and Bandura) criticized behaviorism and psychologism and invited to develop a variety of cognitive models of decision making, suspending a view of human behavior as driven by imperscrutable intuition and emotion. This invitation did not go unheard; actually it had a paramount influence on the evolution of the field. Simon kept a door open to the possibility that once the cognitive modeling of decision making would have been mature, eventually 'the pendulum would have swung again' toward an analysis of their relations with affective factors – something that in fact is an emerging trend nowadays (Gavetti et al., 2007; Hodgkinson and Starbuck, 2012). (Less fortunately, also the use of the notion of 'intuition,' as contrasted with analytic decision making, is reemerging in ODM literature. However, as Simon repeatedly noticed (e.g., 1989), distinguishing (no matter whether for opposing or blending) 'analytic' from 'intuitive' thinking is logically incorrect, because no one knows what intuition is; it is like distinguishing processes which have been analyzed from processes that simply have not been. Actually a core aim in Simon's research program was precisely to discover which logics, which types of analyses, methods, heuristics lay behind what we call intuition or even genius. Hence, here we do not make any use of the term intuition.) Anyhow, models of decision always imply a mix of constraint and freedom (Loasby, 1976): that very mix is what makes choice an interesting activity, with respect to being completely determined (if there is no choice, decision is impossible) or completely unconstrained (if everything works, decision is irrelevant).

The 'Logic' of ODM

The Rationality Divide

The history of thought on ODM did have a 'big bang,' with the concept of 'bounded rationality' (BR) (Simon, 1955). The theory was developed in contrast with utility theory, the dominant model of conscious, purposeful, effectiveness-oriented decision making available at the time, developed especially in economics, and applicable to the behavior of either individual actors or collective organized actors. Since then, most expositions of ODM models are 'anchored' to that model, in the sense of elaborating new models by relaxing some of assumptions of that approach, as to both knowledge and interest, considered to be very restrictive. Those assumptions are deemed to be narrow as they include knowledge of all the relevant alternatives and possible states of the world on a matter or problem, as well as the consequences of their combinations and their possible value (utility) for the decision maker (Simon, 1955). The economic model of rational ODM is still summarized in this way in most accounts (e.g., Hodgkinson and Starbuck, 2012) and 'therefore' considered to be practically 'inapplicable.' A minority of scholars have in turn criticized this characterization of 'value maximization' as a straw man, and pointed out that a complete knowledge assumption is different from a value maximizing strategy, which is nothing else than a set of procedures that can be applied with claim of efficiency *if* the problem at hand is 'closed' (Nutt, 1976), the alternatives in the relevant set are countable (Hatchuel, 2001), the task at hand is not too complex (Beach and Mitchell, 1978; Payne, 1982; Peters, 2003), and if knowledge on cause–effect relations as well as on preferences is available (Thompson and Tuden, 1959) or can be constructed in a valid and reliable way (Grandori, 1984). In those works, accordingly, the ODM models 'alternative' to value maximization are considered as alternative, *contingently effective decision strategies*, applicable in other circumstances, not as rival theories of ODM. Therefore, the adoption of alternative decision strategies is seen as metarational behavior rather than as a form of incapacity or of practically useful but inferior, or even biased thought (as it became in the 'heuristic and biases' behavioral decision theory). For one reason or another, anyway, most scholars agree on the point that, on most problems, actors (either individual or collective) cannot or even should not model problems so that utility maximizing calculations are applicable, either because it is too costly or cognitively unfeasible to do so, or because it would require to simplify problems to an extent at which they no longer represent reality well. In addition, economists themselves have highlighted that the existence and uniqueness of optimal solutions faces limitations and paradoxes in many multiple actor, multiple interest, interactive decision-making situations. The main alternative models, with a claim of being applicable and predictive under conditions of strong uncertainty and conflict of interests, are illustrated next.

The Classic Continuum of ODM Models

Simon outlined the most important of these models, the '*satisficing model*' of search and choice. The simplest basic version of that model states that actors will accept the first encountered alternative superior to a given aspiration level, truncating search at that point. If it is difficult to find acceptable alternatives, aspiration levels fall, and if it is easy, they rise. In that version, satisficing would describe primarily a decision behavior capable of generating 'good' solutions while reducing the costs of search.

Various other decision models, with claims of being able to manage even stronger conditions of uncertainty and conflict, have also been formulated in ODM research.

An '*incremental model*' has been identified in research on ODM especially in the context of public administration (Lindblom, 1959). A starting provocative empirical finding has been that the best predictor of a year's budget for a program or activity in local government is the budget of the previous year (Wildavsky, 1964). In an incremental process, the decision system should be able at least to discern where it wants to go: increase rather than reduce production capacity, improve levels of education, and increase well-being. However, it can generate action without making use of theories on the cause–effect relationships that regulate how the action system works and changes. Rather it employs a 'linear' choice rule – try 'incremental' solutions that differ marginally from those in use and learn afterward.

A *cybernetic model*, built in conceptual frames such as of control theory and stimulus/response behaviorism, relies even more heavily on *ex post* learning, and more precisely on learning by reinforcement (Steinbrunner, 1974). A cybernetic decision strategy implies only the following kinds of judgement: the capacity to recognize situations (e.g., a configuration of costs); the capacity to recognize performance gaps with respect to a standard (works/does not work; positive/negative); and possessing a repertory of possible actions that are applicable to eliminate the gap or respond to the situation, adjourned by a principle of reinforcement (repeat successful action, avoid previously failing actions). Processes that can be described and predicted fairly well as cybernetic processes range from some real-time unreflexive stimulus–response motivation processes to some innovation processes (e.g., the diffusion of codified technologies).

A *garbage can model* of ODM, elaborated in the early 1970s and enjoying a period of significant popularity afterward, can be seen as a limiting case, in which actions are tried at random, guided by chance encounters between actors, with an opening of attention at a certain time, and the emergence of choice occasions (Cohen et al., 1972). It is supposed to be relevant especially under ambiguity: Research using that model has typically studied institutions or processes with 'ambiguous' goals or highly unpredictable consequences (e.g., universities and cultural institutions). Although it has been seen as a provocative criticism of the means-ends 'determinism' of rationalist models of ODM, and as opposed to rational choice, the actual finding of the 'garbage can model' is that if knowledge and preferences are unclear, and alternatives tend to be 'undistinguishable,' then random choice (with eventual *ex post* learning) is the only (and arguably rational) option.

In sum, the main ODM models specified in the second part of the 1900s have been developed by relaxing and extending some of the conditions of uncertainty and conflict which characterized Simon's initial satisficing model, similarly to how the latter model was developed by relaxing and extending some conditions characterizing economic value maximizing models.

The general pattern envisaged and the implicit or explicit conjecture laying behind all these models is that the higher the level of uncertainty and conflict among interests becomes, the less 'rational' the sense of comprehensive and deductively ordered the logic of ODM becomes; the less investment in *ex ante* knowledge acquisition and the more reliance on *ex post* learning is made; and the less preferences are integrated and limited agreement is sought on action.

This orderly range of discrete behavioral alternatives on a 'continuum,' though, does not tell the whole story, nor makes justice to more recent developments in ODM research.

First of all, the decision procedures and heuristics that 'usually characterize' a decision pattern or model may actually be *combined* to generate hybrid decision behavior that better suits circumstances or captures the advantages of different decision modes simultaneously. Actually, it can be noticed that decision rules traditionally belonging to different models, being respectively focused on the different subprocesses of search, choice and learning, may well 'need each other' and be complementary rather than alternative (before applying any choice rule, a problem needs be defined and alternative generated through search and research; and any choice, even if taken according to a correct procedure, may result to be wrong upon test, hence learning is always worthwhile). In addition, it has been noticed early that *all* these traditional models, including both classic maximizing models and classic BR models, do not seem to encompass all feasible decision behavior or to specify the 'continuum' of decision strategies (if it is a continuum) completely. What is neglected is that in decision making, as in any knowledge-based processes, there are 'epistemic' issues, which perhaps are the most important issues, such as interpretation and mindful 'sense-making' (Weick, 1979), the reliability of information and of action based on it (Weick et al., 1999), and the validity of the knowledge generated and used in the process (Grandori, 2010). In one word, 'epistemic rationality' is of paramount importance in good judgment under uncertainty, but it has not been well represented and analyzed in the classic models of ODM.

More precisely and provocatively, traditional ODM research falls short of modeling those decision behaviors which respond to uncertainty not by avoiding it, but by investing in research and knowledge construction, and more precisely in reliable and valid knowledge generation, thereby leading to effective discovery and innovative decision behaviors (Grandori, 1984). At the time this observation was made, though, innovation was not so central as it is nowadays. In the next section, more recent research on effective ODM for innovation, which can fill this gap, is reviewed.

ODM for Innovation and Discovery

A fair amount of research is available on ODM in areas such as entrepreneurship, technological innovation, and strategic innovation. Studies on decision processes in the more innovative and knowledge-intensive parts of the economy show a rich array of decision procedures and heuristics that interestingly differ from those contemplated in the classic approaches so far considered. Rather than postulating a reduced content or a weaker form of thought in decision making as uncertainty increases, research on innovative ODM overall highlights that the opposite strategy – invest more effort and use stronger forms of thought – is conceptually and empirically possible, and even justified as a 'superior' strategy according to the canons of logics and epistemology, as a set of 'rational heuristics' for discovery. (See Grandori (2010, 2013), for a wider review and assessment of those heuristics, as well as for indications of the various studies from which they can be extracted, which cannot be quoted here for reasons of space.) For example, research on the entrepreneurial discovery of new products and services has consistently shown that effective logics include 'disciplined' and 'systematic search'; the reasoned formulation and controlled test of hypotheses; and processes of 'theoretical' and 'causal' modeling, rather than only of experiential learning. Research on the logic of technological innovation has highlighted how it can exploit the multifunctionality of resources and the multipurposedness of action, whereby logics of 'resources/alternatives in search of uses/problems to be solved,' rather than the classic reverse, can increase the rate of success in discovery and reduce downside risk in implementation. Studies on processes of strategic innovation are rich with instances of what Bandura called a 'modeling' approach to the discovering and learning of successful action, as opposed to

experiential direct or vicarious learning: constructing causal models of a situation using theory and relevant empirical evidence rather than 'experience,' and then 'test' them through action.

Considered together, these ODM processes can be considered as offering a further broad family of 'epistemic' models of ODM, complementing the two classic families of 'economic' and 'behavioral' models.

The 'Politics' of ODM

The Power/Efficiency Controversy

A different track of development of ODM studies has been concerned more with the problem of diverging interests and the relative 'weight' or power of actors holding them than on the problem of knowledge. A starting point has been a criticism of Simon's 'administrative man' for 'neglecting,' or not being particularly concerned with, the political and strategic dimensions (Pettigrew, 1973).

American ODM research on the politics of ODM has often contrasted organizational actors' power with organization system's efficiency, showing empirically that often the particularistic objectives and interests of subunits have more weight on the allocation of resources than the overall effectiveness and efficiency objectives held by the central agents of the system (Pfeffer and Salancik, 1974) and that 'subgoal pursuit' typically leads to 'fatter' organization (e.g., larger slack, larger organizational boundaries) (Williamson, 1964).

European studies have considered actors' power in a more positive light, casting the ODM problem as 'the integration of the many,' as contrasted with the 'one actor theorem.' (See Grandori (1995) for a review of the European tradition in ODM.) Research conducted in that perspective has been especially concerned with the strategic use of uncertainty and the determinants of the decision weight of organizational units and groups (Crozier and Friedberg, 1977; Hickson et al., 1971). It has been argued and found that the relative influence of organizational actors in ODM is a function of 'the control of critical uncertainties,' of the degree of substitutability, of the centrality in work and information flows, and of the perception and use of these *atouts* as bases of negotiation power. In that perspective, then, organizational power is a dependent rather than independent variable, and its allocation as a function of critical uncertainty may even be considered efficient, rather than an alternative to efficiency (Grandori, 1993).

A critical observation, on this way of 'extending' ODM models for accommodating conditions of different interests, is parallel to that raised toward the type of 'extension' of economic models of decision making for accommodating strong uncertainty provided by BR models. In the face of strong conflict, as in the face of strong uncertainty, and especially in the face of both, a loss of rationality and fairness is expected. An objection to this position is that this is not the sole possible behavior, nor it seems to be the best response that social and economic actors may be capable of. Also on the terrain of conflict resolution under uncertainty, a 'third' approach is possible and indeed followed both in theory and practice. Streams of studies highly relevant for ODM, combining the acknowledgment of both conflict and uncertainty, and a concern with efficiency and equity in conflict resolution do exist, although they are seldom included in standard reviews of ODM.

Conflict Resolution for Efficiency and Equity

Organizations are typically systems composed by parts. Actually one basic definition of the organization of a system is the mode in which it is partitioned and in which parts are related. Hence, it is more the rule than the exception that a system cannot be modeled properly as 'one single actor,' It might be so, for some classes of decisions (as clarified in the Introduction). But for most decisions, the 'locus' of knowledge and interests is more elementary than the entire organization; actually one basic raisons d'etre of organization is precisely to integrate and coordinate a constellation of interdependent nodes of knowledge and preference. Therefore, rather than asking whether power or efficiency are more important in actually shaping organizational structures and actions, in this perspective the relevant question is to devise structures and actions that are 'superior' for all the relevant actors having a claim to decide on a matter.

A core mechanism which may sustain this type of search is negotiation. There is plenty of evidence, in the negotiation literature, that many if not most organizational actions, and even organization structures themselves, are defined by negotiation. A negotiation perspective is quite different from a 'power perspective' as, while not neglecting the bargaining power of the actors, negotiation analysis provides tools for assessing solutions in terms of efficiency and fairness.

Negotiation is not the only mechanism for solving conflicts in an efficient and fair way though. Other coordination mechanisms are effective alternatives and/or complements to negotiation and negotiated orders in the governance of pluralistic systems and have also been studied in strands of organization studies connected or contiguous with ODM research: for example, voting and organizational democracy constitute another important alternative; alternatively, gaming through both unilateral strategies and coalition formation are possible and indeed likely and can be assessed in terms of Pareto efficiency and fairness and equilibrium of solutions. If these 'voting and gaming' conflict resolution mechanisms have to work effectively, though, problems need be structured, alternatives well defined, and payoffs clear; hence they are applicable to narrower conditions with respect to negotiation.

Hence, the models concerned with the 'politics of ODM' could be ordered on a sort of continuum, as the models concerned with the 'logic of ODM.' Again, the selection of a mode can be conceived in contingency terms, as a function of the clarity and differentiation of knowledge and interests. If at one extreme we locate systems for the aggregation of well-defined preferences over well-defined alternatives, as vote-based organizational democracies and gaming-based organizational poliarchies, at the other extreme we may find 'organized anarchies' (Cohen et al., 1972). In fact, the latter type of decision systems and processes has been identified and connected to conditions of extreme ambiguity, in which preferences and participants themselves are unclear/shifting; there is no predefined pattern or mechanism according to which to take decisions, and everyone is free to address any problem and coalesce with any other actor.

Although connected poliarchies, democracies, negotiated orders, and organized anarchies provide alternatives to markets and hierarchies for governing pluralistic and interdependent decision making in complex systems, those distributed decision-making systems also have costs and limits and differ in terms of the types of performance outcomes they are more likely to produce.

Economists have typically emphasized that the wider and more internally differentiated the set of actors holding decision rights, the better the representation/fairness of the decision system is, but the costlier decision processes are. On the other side, differentiated and integrated ODM, either through simple aggregation of preferences and judgments, or, and even more, through voice, discussion, and negotiation, is likely to produce more innovation than either a disconnected poliarchy or a completely connected and homogeneous community (Grandori, 2009).

The costs and difficulties of all these pluralistic governance decision systems can be shifted downward, though. One way is the institutionalization of framing norms and rules, of 'constitutional orders' which reduces the 'degrees of freedom,' and the variance in actors' preferences; that is, the embedding or blending of conscious, deliberate, ad hoc analysis and choice with the suspension of critical judgment in certain zones of action. Another way is to increase the capacity of information processing through information systems: A set of supports that in the current context of massive development of information and communication technologies (ICTs) are obviously gaining importance and radically enhancing the feasibility of distributed decision making.

Finally, as for decision strategies, recent contributions have highlighted that solutions may be improved and trade-offs simplified by *combining* coordination mechanisms, rather than considering them as mutually exclusive. In fact, in a design perspective on distributed ODM, there is an increasing awareness of the imperfections and paradoxes of any single mode and model and a combination of them recommended to sustain effective and innovative ODM (Grandori 2013).

Concluding Remarks

A Twofold Trend

It is common to recognize that the economic and social world is becoming increasingly complex, more uncertain, differentiated, and conflict laden. The evolution of ODM models in fact is marked by attempts to address increasingly complex conditions. However, this trend has been twofold: Two different types of response and direction of study can be identified in the ODM models for a complex world.

On one side, there are models and studies driven by the hypothesis that in the face of an increasing complexity of the world, thought becomes weaker and rationality more and more limited, with respect to both problem solving and conflict resolution.

On the other side, there are models and studies building on the hypothesis that in the face of the increasing complexity of the world, actors *can* stretch and strengthen their minds, invest in research rather than reduce search, and try to solve conflicts by devising fair higher-order rules of the game and constitutions rather than indulge in power games.

These two trends may well reflect a polarization of ODM system in reality. The first type of reaction is likely to be more common, to describe well average behavior, but to 'leave resource on the table.' The second type of reaction is likely to be less frequent but 'superior,' likely to be found in 'outliers' more than in mainstream practice, but inspired to the 'best patterns of thought and action' that are logically and practically possible. As an agenda for future research, the second perspective may deserve more attention, not only because the majority of mainstream ODM can be located in the first approach, but also for its higher potential for improving ODM under complexity.

Implications for Structures

ODM is influenced by organizational architectures, and vice versa structures are influenced by ODM processes and approaches. Information processing-based organization theory stated a wide set of propositions on how to design organization structures so as to set 'decision premises' stimulating more optimizing and computational decision processes in units dealing with structured tasks and devoted to the exploitation of existing resources, leaving space for more heuristic and experiential approaches in unstructured task environments and exploration-oriented parts of the organization.

The enrichments that occurred in decision models for knowledge-intensive and innovative decision making highlighted that in those conditions there are further implications for structures. Organizational systems should become, almost by definition, less 'decomposable' if problems are not known enough for being decomposed and new knowledge is to be generated in ODM processes. Hence, somewhat contrary to some current organizational fashions, effective structures for the growth of knowledge should be expected to become not only less hierarchical, but also more connected, that is, less 'loosely coupled' and less 'modularized' (in fact, hierarchy is quite modular a system). The general principle that, in managing uncertainty and innovation, both differentiation and integration are required, seems to remain valid; however, the forms of specialization and coordination that are likely to be most innovative are expanded by the examined recent trends in ODM. The classic notion was that some units should be flexible and exploration oriented and some structured and devoted to exploitation. The emerging model is more networked: If the growth of knowledge is central, the 'parts' or units should be characterized by differentiated knowledge that should be all integrated to contribute to organizational learning. As illustrated in the decision logic section, heuristics such as 'resources in search of uses' and multipurposedness imply that the very search for better exploitation of existing resources implies exploring new uses, activities, and services. Hence, in terms of future research, some of the currently popular distinctions in ODM, in particular that between 'exploration' and 'exploitation,' may deserve some revision in the light of the logics of ODM for innovation and discovery specified in recent research.

More horizontal, paritarian, pluralistic, and representative organization forms should also be favored by the multiplication of the relevant actors and constituencies, as

illustrated in the politics of decision-making section. To put the study of pluralistic governance arrangements high on research agendas seems to be overdue in economic and management sciences, still seeing governance as limited to command-based, incentive-driven, or conflict-free communitarian alternatives. The development of new information and communication technologies should provide extensive and yet far from fully exploited support to forms of ODM that are both pluralistic and connected.

See also: Decision and Choice: Bounded Rationality; Decision and Choice: Heuristics; Organizations: Negotiated Orders; Rational Choice and Organization Theory.

Bibliography

Barnard, C.I., 1938. The Functions of the Executive. Harvard University Press, Cambridge, MA.

Beach, L.R., Mitchell, T.R., 1978. A contingency model for the selection of decision strategies. Academy of Management Review 3, 439–449.

Cohen, M.D., March, J.J., Olsen, J.P., 1972. A garbage can model of organizational choice. Administrative Science Quarterly 17, 1–25.

Crozier, M., Friedberg, E., 1977. L'Acteur et le systéme. Editions du Seuil, Paris.

Elster, J. (Ed.), 1985. The Multiple Self. Cambridge University Press, Cambridge, UK.

Gavetti, G., Levinthal, D., Ocasio, W., 2007. Neo-Carnegie: The Carnegie School's past, present, and reconstructing for the future. Organization Science 18 (3), 523–536.

Grandori, A., 1984. A prescriptive contingency view of organizational decision making. Administrative Science Quarterly 29, 192–208.

Grandori, A., 1993. Notes on the use of power and efficiency constructs in the economics and sociology of organization. In: Lindenberg, S., Schreuder, H. (Eds.), Interdisciplinary Perspectives on Organization Studies. Pergamon Press, Oxford.

Grandori, A., 1995. Models of pluralistic organization: the contribution of European decision making research. In: Bacharach, S., Gagliardi, P., Mundell, B. (Eds.), Research in the Sociology of Organizations. The European Tradition, vol. 13. Jay Press, San Francisco, CA.

Grandori, A., 2009. Poliarchic governance and the growth of knowledge. In: Foss, N., Michailova, S. (Eds.), Knowledge Governance. Oxford University Press.

Grandori, A., 2010. A rational heuristic model of economic decision making. Rationality and Society 22, 477–504.

Grandori, A., 2013. Models of rationality in economic organization: economic, experiential, epistemic. In: Grandori, A. (Ed.), Handbook of Economic Organization. Integrating Economic and Organization Theory. Edward Elgar.

Hatchuel, A., 2001. Towards design theory and expandable rationality: the unfinished program of Herbert Simon. Journal of Management and Governance 5 (3–4), 260–273.

Hickson, D., Hinings, C.R., Lee, C.A., Schnek, R.E., Pennings, J.M., 1971. A strategic contingencies theory of intraorganizational power. Administrative Science Quarterly 16, 216–229.

Hodgkinson, G.P., Starbuck, W.H., 2012. The Oxford Handbook of Organizational Decision Making. Oxford University Press.

Lindblom, C.E., 1959. The science of muddling through. Public Administration Review 19, 78–88.

Loasby, B., 1976. Choice, Complexity and Ignorance. Cambridge University Press.

Nutt, P.C., 1976. Models for decision making in organizations and some contextual variables which stipulate optimal use. Academy of Management Review 1, 84–98.

Payne, J.W., 1982. Contingent decision behavior. Psychological Bulletin 92, 382–402.

Peters, E., 2003. Simple and complex market inefficiencies: integrating efficient markets, behavioral finance, and complexity. The Journal of Behavioral Finance 4 (4), 225–233.

Pettigrew, A.M., 1973. The Politics of Organizational Decision Making. Harper and Row, New York.

Pfeffer, J., Salancik, G.R., 1974. Organizational decision making as a political process: the case of a university. Administrative Science Quarterly 19, 135–151.

Simon, H.A., 1955. A behavioral model of rational choice. Quarterly Journal of Economics 69, 99–118.

Simon, H.A., 1989. Making management decisions: the role of intuition and emotion. Academy of Management Executive 1 (1), 57–64.

Simon, H.A., 1990. A mechanism for social selection and successful altruism. Science 250, 1665–1668.

Steinbrunner, J.D., 1974. The Cybernetic Theory of Decision. Princeton University Press, Princeton, NJ.

Thompson, J.D., Tuden, A., 1959. Strategies, structures, and processes of organizational decision. In: Thompson, J.D., Hammond, P.B., Hawkes, R.W., Junker, B.H., Tuden, A. (Eds.), Comparative Studies in Administration. University of Pittsburgh Press, Pittsburgh, PA.

Weick, K.E., 1979. Cognitive processes in organizations. Research in Organizational Behavior 1, 41–74.

Weick, K.E., Sutcliffe, K.M., Obstfeld, D., 1999. Organizing for high reliability: processes of collective mindfulness. In: Sutton, R.S., Staw, B.M. (Eds.), Research in Organizational Behavior, vol. 1. Jai Press, pp. 81–123.

Wildavsky, A., 1964. Politics of the Budgetary Process. Little, Brown, Boston, MA.

Williamson, O.E., 1964. The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm. Prentice Hall, Englewood Cliffs, NJ.

Zey, M. (Ed.), 1992. Decision Making. Alternatives to Rational Choice Models. Sage, Beverly Hills, London.

Organizational Ecology

Glenn R Carroll and Michael T Hannan, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.T. Hannan, volume 17, pp. 11780–11784, © 2001, Elsevier Ltd.

Abstract

Organizational ecology refers to a sociologically oriented research program on organizations. It involves the empirical study of populations of organizations and a theoretical emphasis on processes of selective replacement of relatively inert organizations. Much organizational ecology research uses common methodological presumptions and practices, including the analysis of entry, exit, and growth hazard rates in large historical populations from their dates of origin. Development and testing of ecological theory has taken place within cumulative theory fragments, such as niche width, density dependence, and resource partitioning. Each fragment addresses focused research problems in a coherent way using middle-range theory. Combining the theory fragments remains an unrealized goal, but one which has spurred a great amount of theory and research on how institutionalized social categories shape organizations and guide organizational action.

Organizational ecology (OE) is a major theoretical perspective that attempts to explain organizational diversity, famously framed by the orienting question of Hannan and Freeman (1977), why are there so many different kinds of organizations? OE uses a selection model of organizational evolution to explain organizational diversity in a population. In the imagery of this model, organizational change over time occurs through the selective replacement of some kinds of organizations in the population with other kinds as environmental circumstances change. In selection models, at any point in time the environment operates as a screen, letting ill-fitted organizations fall through or fail and catching or supporting better-fitted organizations. Selection models assume individual organizations face constraints that limit their abilities to make major transformations, which meet the changing demands of the environment (Hannan and Freeman, 1984). Accordingly, research in OE focuses on the vital rates of populations – rates of founding and mortality in particular, although growth and transformation remain of interest. Selection-driven evolution is sometimes mistakenly thought to lead to a narrowing of organizational diversity, a winnowing. Although winnowing may be more likely when constraints such as regulation limit new entry, various selection processes can drive population diversity up or down (Carroll et al., 2012).

One foundational pillar of the OE research program examines the demography of organizational populations. It finds that organizational mortality processes depend upon the age and size of the organization, as well as on characteristics of populations and environments. Moreover, research confirms an imprinting process, meaning that environmental conditions at certain early phases in an organization's development exert long-term consequences. In particular, organizations subject to intense competition show elevated mortality hazards at all ages.

OE theory invokes a central assumption of structural inertia, the tendency for organizations to respond slowly relative to the speed of environmental change. A core argument holds that inertia derives from the very characteristics that make organizations favored kinds of actors in modern society: reliability and (formal) accountability. It follows logically that change in an organization's core features disrupts activities and increases the hazard of mortality, at least in the short-run, a prediction supported empirically (Barnett and Carroll, 1995).

Within OE, the niche concept organizes theoretical ideas relating environmental variations and competition to population vital rates. The niche specifies the resource space within which a population (or even an individual organization) exists. Much empirical work in OE examines the niches of organizational populations along dimensions of social, political, and economic environments, a tradition dating back to the human ecology of Hawley (1968). Recent research conceptualizes niches in terms of identities within institutionalized categories, an approach that has proven very insightful (Ruef, 2000; McKendrick et al., 2003).

Development and testing of ecological theory has taken place within cumulative theory fragments. Each fragment addresses focused research problems in a coherent way using middle-range theory (Merton, 1968). Within basic demography, fertile fragments include aging, inertia, and niche width. Major fragments dealing with population dynamics include density dependence and resource partitioning.

The theory of density-dependent organizational evolution integrates ecological and institutional processes. It posits that growth in the number of organizations in a population (density) drives processes of social legitimation and competition that, in turn, shape the vital rates. This argument implies certain nonmonotonic relationships between density and rates of founding and mortality.

Resource-partitioning theory concerns the relationship between increasing market concentration and increasing proliferation of specialists in mature industries. The key implication of this theory concerns the effects of concentration on the viability of specialist organizations (those that seek to exploit a narrow range of resources).

The Basic Demography of Organizations

With its concentration on the vital rates of founding and mortality, OE has a strong demographic character. Key

structuring processes are specified at the population level. Mortality histories can readily be associated with individual organizations, but histories of foundings cannot because the absence of foundings in some period or place must be considered. Populations must have unit character to serve as useful objects of analysis (Hawley, 1968). Researchers generally regard populations as spatial–temporal instantiations of organizational forms, templates for organizing. Several conceptual approaches to form have guided these efforts: common resource dependence, structural equivalence in network position, and more recently identities coded in cultural rules, especially institutionalized categories.

Constructing organizational demography and ecology requires attention to how organizations differ from organisms. In particular, organizations experience more varied 'birth' and 'death' events, typically lack clear 'parents,' consist of multilayered, partly decomposable structures, and can transform themselves into very different kinds of entities. Methodological issues in OE often involve dealing with these complexities (Carroll and Hannan, 2000; Phillips, 2005).

The prototypical OE research design examines the full histories of organizational populations, because experience shows that early events can exert lasting consequences for population dynamics. The design gathers life-history data on all organizations in the population(s), including the large and famous as well as the small and insignificant; such inclusiveness is crucial for avoiding problems of selectivity bias. The design records detailed information about type of entry (new founding, entry from another industry, merger, etc.) and exit (disbanding, acquisition, transformation, etc.). Finally, it uses event-history methods to estimate the effects on vital rates of characteristics of organization, population, and environment. Used earliest by OE analysts, this focused population design diffused widely to other kinds of organizational research; and now seems primed to affect the study of products (Khessina and Carroll, 2008; Carroll et al., 2010).

The prototypical OE design typically uncovers many more organizations operating in well-studied industries than is usually assumed. Careful archival searches usually find information on hundreds of organizations that get overlooked in published counts, some because of their small size, and some because of their ephemerality. Analysts using traditional approaches counting only bigger and more stable organizations defend the exclusion of these entities based on their apparent minimal impact on major economic and social outcomes. However, ecologists point out that minimal impact at any given point does not preclude subsequent impact, or potential impact, an insight of evolutionary approaches. Classic examples of companies that might be excluded by such rules would include Microsoft, Apple, and Google at the time of their origins, leading the analyst to either include them retrospectively based on later performance (creating selection bias) or including them only when they reach a certain size (creating an unnatural 'founding' date).

An unexpected insight generated by the prototypical OE design was that the foundational periods of industrial populations often show activity typically seen in social movements, apparently noncommercial pursuits. Carroll and Hannan (2000: p. 231) made the general observation that: "To say that the origin periods of industries resemble social movements means that (1) they represent challenges to existing companies and industries, and (2) they are populated with individuals and organizations devoted to causes, lifestyles, and visions of a better future for all (rather than with profit-maximizing entrepreneurs engaged in competitive battles based primarily on self-interest)." Other insights from social movement theory have subsequently been added to OE analysis (Wade et al., 1998; Rao et al., 2003).

The OE design also reveals that organizational life expectancies are generally an order of magnitude smaller than for humans, prompting great attention to organizational aging. Stinchcombe (1965) observed that organizations (and organizational populations) experience a liability of newness. Efficient organization requires trust among members, which takes time to build. Creating roles and routines, learning about the environment, and developing relationships with other organizations also take time. Once established, these patterns of relationships enhance survival chances. Much early research found this pattern. Some later research found a variation, a liability of adolescence: the hazard rises during the early lifespan – while initial stocks of endowments get exhausted – before declining. Recent research, using designs that measure (age-varying) organizational size, has produced (mixed) evidence in favor of positive age dependence (Carroll and Hannan, 2000).

Two interpretations for this pattern have been proposed (Barron et al., 1994; Sørensen and Stuart, 2000). If inertial forces are strong, then the possibilities of adapting to changing environments are limited and older cohorts of organizations have lower fitness – there is a liability of obsolescence. Alternatively, the accumulation of rules and routines impedes adjustment to environmental change – the liability is one of senescence.

Subsequent research sought to clarify the conditions under which the various patterns rise. Hannan (1998) proposed a partial integration using first-order logic. The next step required use of nonmonotonic logic, which allows premises to be overridden by more specific ones so that the conclusions of an argument can change as more knowledge about specifics becomes available (Hannan et al., 2007). In recent work, Le Mens et al. (2011) revised the theory in terms of the stock of organizational capital and show that the various patterns of age dependence can be integrated using weaker assumptions.

Inertia and Selection

Stinchcombe's (1965) insight that new organizations get imprinted by their environments provided inspiration for organizational ecologists. According to Stinchcombe, new organizations get tested against assumptions that are taken for granted but change over time. Hence, the characteristics of organizations which pass such tests and emerge will reflect the social structure of the time. Imprinting requires both a mapping of environmental conditions onto the organization and inertia in the imprinted characteristics. Absent the second condition, later adjustments will erode the association between founding conditions and those features.

OE analysts built on this image of inertia (relative to the speed of environmental change). Hannan and Freeman (1984)

claimed inertia prevails as an inadvertent by-product of a selection process favoring the properties of reliability (low variance in the quality of performance) and accountability (the ability to construct rational accounts for actions). Achieving these properties depends on structures being reproduced faithfully over time. Yet, high reproducibility means that structures resist transformation. Therefore, this selection process inadvertently favors corporate actors with strong inertial tendencies.

Fundamental change presumably diminishes reliability and accountability. Old learning loses value, and new processes must be learned. Vestiges of the old system conflict with the new emerging system and disrupt activities and operations. Thus, even if the intended change has long-term benefits (the new form might be better aligned with environments and internal processes), the change process might increase mortality hazards in the short-run. If so, then reorganization-prone organizations are less likely to be present in future populations.

Ample research examines the implications of this argument. Most well-designed studies reveal that changing core structures does indeed increase the hazard of mortality in the short-run, often substantially (Carroll and Hannan, 2000). Moreover, the magnitude of this effect increases with organizational age.

A different way of viewing imprinting yielded one of OE's most robust empirical findings: density delay (Carroll and Hannan, 1989). Organizations subject to intense competition at founding (as evidenced by a high density of organizations in the population – see below) have elevated mortality hazards at all ages. In other words, density at founding has a (partially) delayed effect on mortality processes. This relationship has been interpreted as reflecting enduring effects either of building structures under scarcity or of initial organizational positioning at the margins of resource distributions.

More recently, theorists suggest that identity enforcement processes might generate inertial mechanisms as well as punishments for fundamental changes (those that violate codes – see Zuckerman, 1999; Hannan et al., 2007; Hsu et al., 2009).

Organizational Niches and Institutionalized Categories

Within OE theory, the niche concept organizes ideas for relating environmental variations and competition to organizational demography and population dynamics. An organization's or a population's fundamental niche consists of the ranges of environmental parameters for which its growth rate (fitness) is nonnegative in the absence of competition. Much empirical work in OE examines the niches of organizations and populations in terms of dimensions of social, political, and economic environments (e.g., Dobrev, 1999; Boone et al., 2004; Goldstein and Haveman, 2013). Even when such issues do not stand in the forefront theoretically, ecologists routinely control statistically for the effects of such dimensions in examining dynamic processes.

Recent efforts to rebuild the foundations of the OE theory depict organizational forms as based on culturally coded and enforced identities (Hannan et al., 2007). These identities operate within institutionalized social categories, agreed upon classification systems that shape the organizational world with labels and associated schemata. For instance, 'private liberal arts college' represents a type of higher educational organization which likely contains many specific features about its personnel, employee contracts, technology, funding base, impact, and the like. Organizations may be more or less members of specific categories, defining their 'grades of membership' in a category, and organizational may span across multiple categories, thereby increasing their niche width in the space of categories. Boundaries between categories may be strongly or weakly enforced by institutional gatekeepers and others. And, categories vary in their apparent distinctiveness based on category permeability and organizational distributions; ecologists describe these variations as category 'contrast' (Negro et al., 2010; Kovaçs and Hannan, 2010).

Niche theory has been central to research on competition (Hannan and Freeman, 1977, 1989; McPherson, 1983). Populations compete if the presence of each reduces the spaces in which the others can sustain themselves (their realized niches). By this reasoning, populations compete if and only if their fundamental niches intersect. The assumed equivalence of niche overlap and competition plays a crucial role in allowing organizational ecologists to relate empirical observations on realized niches to models of population dynamics.

Niche width concerns the variance of resource utilization. Does evolution favor populations of Jacks-of-all-trades and masters of none (generalists) or specialists at a few tasks? Ecologists call this the Principle of Allocation (POA) problem. Research shows that the organizational world tends toward segmentation into specialist populations (each highly isomorphic) when relevant environmental variations are uncertain, large in magnitude, and long in duration (Hannan and Freeman, 1989; Péli, 1997).

Recent research examines the POA problem in the context of form identities and social categories. This research asks, does category spanning lead to deleterious consequences for organizations? Findings to date suggest strongly that when an organization's niche width crosses category boundaries, the organization will pay a price in terms of recognition and valuation (Hsu, 2006; Hannan et al., 2007; Negro et al., 2010; Pontikes, 2012). The penalty seems to be more severe when the categories involved are defined in opposition to each other, for example, mass producer and artisanal craft producer.

Density Dependence

The theory of density-dependent organizational evolution integrates ecological and institutional processes. It posits that growth in the number of organizations in a population (density) drives processes of social legitimation and competition that, in turn, shape the vital rates (Hannan and Carroll, 1992). An organizational form gains (constitutive) legitimation when it becomes taken for granted (Meyer and Rowan, 1977). New organizational forms generally lack such legitimation, which makes organizing difficult. If the form gets promulgated, legitimation increases. Initially, when there are few organizations in a population and the form lacks legitimacy, the returns to legitimation of adding another organization are great. However, when the population

contains many organizations, constitutive legitimation increases little – if at all – as density rises further. In other words, a population's legitimation increases with density at a decreasing rate and approaches a ceiling.

Competition refers to the consequences of efforts by multiple agents to acquire the same limited resources. Density-dependence theorists focus on diffuse competition because it has a clear relationship with density as the intensity of diffuse competition within an organizational population rises as a function of the number of potential bilateral competitors. As density increases linearly, the number of possible competitive relations increases geometrically. Thus diffuse competition rises with density at an increasing rate. The theory posits that founding rates are directly proportional to legitimation and inversely proportional to the intensity of diffuse competition and that these relationships are inverted for mortality hazards.

This theory (formalized appropriately) yields two theorems. First, as density increases, founding rates initially rise, reach a peak, and then decline. Second, the relationship for mortality is the mirror image. These theorems, expressed in parametric forms, can be falsified empirically. They have been tested for diverse populations, including automobile manufacturers in Europe and the United States, brewers in Germany and the United States, banks in Denmark, Japan, Singapore, and the United States, and American labor unions, life insurers, newspapers, and social movements of women and ethnic groups. These studies have supported the theory in most cases (full citations can be found in Carroll and Hannan, 2000).

In recent formulations, density dependence plays a role in the cultural coding process of forms. Ecologists use this approach to weight density using grades of membership (Kuilman and Li, 2009; Boone et al., 2011).

Resource Partitioning

Resource-partitioning theory addresses market segmentation; it concerns the relationship between increasing market concentration and increasing proliferation of specialists in mature industries (Carroll, 1985). The theoretical imagery involves crowding in a heterogeneous market. Organizations initially target resource segments. Specialists choose narrow, homogeneous targets; generalists choose broad, heterogeneous ones.

The theory assumes an advantage of scale (in production, marketing, or distribution). Because organizations targeting narrow segments cannot grow very large, they have higher (per unit) costs. Organizations with scale advantages expand product offerings to broaden their appeal without losing established domains. Small organizations then face formidable invaders, and many fail. Competition is most intense in resource-dense centers, resulting in dominance by large generalists. Failures of smaller generalists cause some resources to become free. Because larger generalists can best secure these resources, the surviving generalists become larger and more general. However, generalists can rarely secure all of the newly freed resources. As the struggle among generalists proceeds to its monopoly equilibrium, the surviving generalists grow and broaden, but the set of resources they hold collectively actually declines. In other words, more resources become available for specialists as the market concentrates and their viability increases as well: specialist founding rates rise and mortality rates fall. Accordingly, the specialist advantage in concentrated markets comes as a consequence of strategic positioning by generalists in a market conceived of spatially (Péli and Nooteboom, 1999).

Subsequent formulations of the theory highlight the market attractions of specialist identities, clarifying certain issues in resource partitioning. Specialization based on claims of authentic identity (as in the case of the rise of specialist producers of microbrewed beer in the United States and 'real ale' in the United Kingdom) can prevent large generalists from gaining complete control of concentrated markets; this development makes it very hard, if not impossible, for generalists to compete, especially within certain categories (Carroll and Swaminathan, 2000; Koçak and Özcan, 2013). Generalizations of this insight have focused broadly on the appeal of products, services, and performances perceived to be authentic, because of aspects of the producer's identity (Carroll and Wheaton, 2009; Phillips, 2013; McKendrick and Hannan, 2013). Empirical support for the theory has been found in a variety of contexts including airlines, automakers, commercial and investment, brewers, whiskey distillers, microprocessor manufacturers, newspaper publishers, telephone companies, television broadcasters, and wineries (full citations can be found in Carroll and Hannan, 2000).

Emerging Trends

Current research seeks to extend many of these and other OE theory fragments, to connect to other social science research programs and to expand the range of applications to other phenomena.

As for extensions, the most active fragment by far is that concerning identities and categories. Work proceeds on fundamental issues such as how categories emerge, evolve, and impact (Negro et al., 2010b). It concerns organizations but also labels, boundaries, regulators, and other gatekeepers (Negro et al., 2010a; Hsu et al., 2012). The theory invoked in this program relies heavily on cognitive psychology and cultural sociology (Hannan, 2010; Negro and Leung, 2012).

OE analysts also keep integrating ideas about social networks into the analysis. Cattani et al. (2008) demonstrated how density-dependent legitimation grows hand in hand with development of the interorganizational network. Intriguing recent research on the relational structure of schemas within an audience shows promise for advancing understanding of audience dynamics (Goldberg, 2011; Koçak et al., in press).

Organizational status has been shown to affect organizational life chances, with the impact increasing with the uncertainty in evaluations of quality (Podolny et al., 1996). Kovacs and Sharkey (2014) show that status shocks sometimes produce unexpected audience dynamics. Social movement research relies increasingly on OE as a guide to modeling dynamics (Minkoff, 1997; Soule and King, 2008).

The empirical scope of OE continues to expand. Current research expands the diversity of organizational populations studied. In addition to those already mentioned, these include:

accountancies, bicycle manufacturers, computer manufacturers, disk-drive producers, footwear manufacturers, orchestras, hotels, law firms, magazine publishers, music recording firms, paper manufacturers, restaurants, sex shops, Islamic banks, and software producers. A major thrust of current work is on connections among the organizations (including flows of personnel among them). These studies highlight again the value of social network analysis for understanding the ecology of organizations.

Other important social and economic phenomena are also being brought into the picture. Studies of careers find that organizational demography induces considerable job mobility (Haveman and Cohen, 1994); at the same time, careers of managers shape strategies and thereby influence life events for organizations (Sørensen, 1999). The density and diversity of organizational forms in a community also apparently affect social and legal change (Koçak and Carroll, 2008; Boone et al., 2011; Negro et al., 2013) as well as the degree of civic engagement (Sampson et al., 2005), fulfilling the promise of Hannan and Freeman's (1977) orienting question about organizational populations.

See also: Darwinism: Social; Ecological Economics; Ecology: Organizations; Geography: Overview; Social Categorization.

Bibliography

Barnett, W.P., Carroll, G.R., 1995. Modeling internal organizational change. Annual Review of Sociology 21, 217–236.

Barron, D.N., West, E., Hannan, M.T., 1994. A time to grow a time to die: growth and mortality of credit unions in New York City, 1914–1990. American Journal of Sociology 100, 381–421.

Boone, C., Brouwer, A., Jacobs, J., van Witteloostuijn, A., 2011. Religious pluralism and organizational diversity: an empirical test in the city of Zwolle, the Netherlands, 1851–1914. Sociology of Religion 73, 150–173.

Boone, C., Carroll, G.R., van Witteloostuijn, A., 2002. Resource distributions and market partitioning: Dutch daily newspaper organizations from 1968 to 1994. American Sociological Review 67, 408–431.

Boone, C., Carroll, G.R., van Witteloostuijn, A., 2004. Size, differentiation and the performance of Dutch daily newspapers. Industrial and Corporate Change 13, 117–148.

Carroll, G.R., 1985. Concentration and specialization: dynamics of niche width in populations of organizations. American Journal of Sociology 90, 1262–1283.

Carroll, G.R., Hannan, M.T., 1989. Density delay in the evolution of organizational populations: a model and five empirical tests. Administrative Science Quarterly 54, 524–541.

Carroll, G.R., Hannan, M.T., 2000. The Demography of Corporations and Industries. Princeton University Press, Princeton, NJ.

Carroll, G.R., Harrison, J.R., McKendrick, D.G., 2012. Selection and variation in organizational evolution. Industrial and Corporate Change 21, 217–244.

Carroll, G.R., Khessina, O.M., McKendrick, D.G., 2010. The social lives of products: analyzing product demography for management theory and practice. Academy of Management Annals 4, 157–205.

Carroll, G.R., Swaminathan, A., 2000. Why the microbrewery movement? Organizational dynamics of resource partitioning in the U.S. brewing industry. American Journal of Sociology 106, 715–762.

Carroll, G.R., Wheaton, D.R., 2009. The organizational construction of authenticity: an examination of contemporary food and dining in the U.S. Research in Organizational Behavior 29, 255–282.

Cattani, G., Ferriani, S., Negro, G., Perretti, F., 2008. The structure of consensus: network ties, legitimation and exit rates of US feature film producer organizations. Administrative Science Quarterly 53, 145–182.

Dobrev, S.D., 1999. The dynamics of the Bulgarian newspaper industry in a period of transition. Industrial and Corporate Change 8, 573–605.

Goldberg, A., 2011. Mapping shared understandings using relational class analysis: the case of the cultural omnivore reconsidered. American Journal of Sociology 116, 1397–1436.

Goldstein, A., Haveman, H.A., 2013. Pulpit and press: denominational dynamics and the growth of religious magazines in antebellum America. American Sociological Review 78, 797–827.

Hannan, M.T., 1998. Rethinking age dependence in organizational mortality. American Journal of Sociology 104, 85–123.

Hannan, M.T., 2010. Partiality of memberships in categories and audiences. Annual Review of Sociology 36, 159–181.

Hannan, M.T., Carroll, G.R., 1992. Dynamics of Organizational Populations. Oxford University Press, New York.

Hannan, M.T., Freeman, J., 1977. The population ecology of organizations. American Journal of Sociology 82, 929–964.

Hannan, M.T., Freeman, J., 1984. Structural inertia and organizational change. American Sociological Review 49, 149–165.

Hannan, M.T., Freeman, J., 1989. Organizational Ecology. Harvard University Press, Cambridge, MA.

Hannan, M.T., Pólos, L., Carroll, G.R., 2007. The Logics of Organization Theory: Audiences, Codes and Ecologies. Princeton University Press, Princeton.

Haveman, H.A., Cohen, L.E., 1994. The ecological dynamics of careers. American Journal of Sociology 100, 104–152.

Hawley, A.H., 1968. Human ecology. In: Sills, D. (Ed.), International Encyclopedia of Social Sciences. Macmillan, New York, pp. 328–337.

Hsu, G., 2006. Jacks of all trades and masters of none: audiences' reactions to spanning genres in feature film production. Administrative Science Quarterly 51, 420–450.

Hsu, G., Hannan, M.T., Koçak, Ö., 2009. Multiple category memberships in markets: a formal theory and two empirical tests. American Sociological Review 74, 150–169.

Hsu, G., Negro, G., Perretti, F., 2012a. Hybrids in Hollywood: a study of the production and performance of genre-spanning films. Industrial and Corporate Change 21, 1427–1450.

Hsu, G., Roberts, P., Swaminathan, A., 2012b. Evaluative schemas and the mediating role of critics. Organization Science 23, 83–97.

Khessina, O., Carroll, G.R., 2008. Product demography of de novo and de alio firms in the optical disk drive industry, 1983–1999. Organization Science 19, 25–38.

Koçak, Ö., Carroll, G.R., 2008. Growing church organizations in diverse U.S. communities 1890–1926. American Journal of Sociology 113, 1272–1315.

Koçak, Ö., Hannan, M.T., Hsu, G. Emergence of market orders: audience interaction and vanguard influence. Organization Studies, in press.

KoçakÖ., Özcan, S., 2013. How does rivals' presence affect firms' decision to enter new markets? Economic and sociological explanations. Management Science 59, 2586–2603.

Kovaçs, B., Hannan, M.T., 2010. The consequences of category spanning depend on contrast. Research in the Sociology of Organizations 35, 175–201.

Kovaçs, B., Sharkey, A., 2014. Paradoxical publicity: how status can lead to lower perceived quality. Administrative Science Quarterly 59, 1–33.

Kuilman, J., Li, J.T., 2009. Grades of membership and legitimacy spillover: foreign banks in Shanghai, 1847–1935. Academy of Management Journal 52, 229–245.

Le Mens, G., Hannan, M.T., Pólos, L., 2011. Founding conditions, learning, and organizational life chances: age dependence revisited. Administrative Science Quarterly 56, 95–126.

McKendrick, D.G., Jaffee, J., Carroll, G.R., 2003. In the bud? Analysis of disk array producers as a (possibly) emergent organizational form. Administrative Science Quartely 48, 60–93.

McKendrick, D.G., Hannan, M.T., 2013. Oppositional identities and resource partitioning: distillery ownership in scotch whisky, 1826–2009. Organization Science 25 (4), 1272–1286.

McPherson, J.M., 1983. An ecology of affiliation. American Sociological Review 48, 519–535.

Merton, R.K., 1968. Social Theory and Social Structure. Free Press, Glencoe, IL.

Meyer, J.W., Rowan, B., 1977. Institutionalized organizations: formal structure as myth and ceremony. American Journal of Sociology 83, 340–363.

Minkoff, D., 1997. The sequencing of social movements. American Sociological Review 62, 779–799.

Negro, G., Hannan, M.T., Rao, H., 2010a. Categorical contrast and audience appeal: niche width and critical success in winemaking. Industrial and Corporate Change 19, 1397–1425.

Negro, G., Koçak, Ö., Hsu, G., 2010b. Research on categories in the sociology of organizations. Research in the Sociology of Organizations 31, 3–35.

Negro, G., Leung, M.D., 2012. "Actual" and perceptual effects of category spanning. Organization Science 24, 684–696.

Negro, G., Perretti, F., Carroll, G.R., 2013. Challenger groups, commercial organizations, and policy enactment: an empirical study of local lesbian/gay rights ordinances. American Journal of Sociology 119, 790–832.

Péli, G., 1997. The niche hikers guide to population ecology. In: Raftery, A. (Ed.), Sociological Methodology. Blackwell, Oxford, UK, pp. 1–46.

Péli, G., Nooteboom, B., 1999. Market partitioning and the geometry of resource space. American Journal of Sociology 104, 1132–1153.

Phillips, D.J., 2005. Organizational genealogies and the persistence of gender inequality: the case of Silicon Valley law firms. Administrative Science Quarterly 50, 440–472.

Phillips, D.J., 2013. Shaping Jazz: Cities, Labels, and the Global Emergence of an Art Form. Princeton University Press, Princeton, NJ.

Podolny, J.P., Stuart, T.E., Hannan, M.T., 1996. Networks, knowledge and niches. American Journal of Sociology 102, 659–689.

Pontikes, E.G., 2012. Two sides of the same coin: how ambiguous classification affects multiple audience evaluations. Administrative Science Quarterly 57, 81–118.

Rao, H., Monin, P., Durand, R., 2003. Institutional change in Toqueville: Nouvelle Cuisine as an identity movement in French gastronomy. American Sociological Review 108, 795–843.

Reis, S., Negro, G., Sorenson, O., Perretti, F., Lomi, A., 2012. Resource partitioning revisited: evidence from Italian broadcasting. Industrial and Corporate Change 22, 459–487.

Ruef, M., 2000. The emergence of organizational forms: a community ecology approach. American Journal of Sociology 106, 658–714.

Sampson, R.J., McAdam, D., MacIndoe, H., Weffler-Elizondo, S., 2005. Civil society reconsidered: the durable nature and community structure of collection action. American Journal of Sociology 111, 673–714.

Sørensen, J.B., 1999. The ecology of organizational demography. Industrial and Corporate Change 8, 713–744.

Sørensen, J.B., Stuart, T.E., 2000. Aging, obsolescence, and organizational innovation. Administrative Science Quarterly 45, 81–112.

Soule, S.A., King, B.G., 2008. Competition and resource partitioning in three social movement industries. American Journal of Sociology 113, 1568–1610.

Stinchcombe, A.L., 1965. Social structure and organizations. In: March, J.G. (Ed.), Handbook of Organizations. Rand-McNally, Chicago, pp. 143–193.

Wade, J.B., Swaminathan, A., Saxon, M.S., 1998. Normative and resource flow consequences of local regulations in the American brewing industry. Administrative Science Quarterly 43, 905–935.

Zuckerman, E.W., 1999. The categorical imperative: security analysts and the legitimacy discount. American Journal of Sociology 104, 1398–1438.

Organizational Emergence and Firm Formation

Elizabeth Garnsey, Institute for Manufacturing, University of Cambridge, Cambridge, UK
Sarah Lubik, Simon Fraser University, BC, Canada
Paul Heffernan, Institute for Manufacturing, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents the emergence and evolution of an organization as a response to the demands faced by the newly created firm, shaped by a process of problem solving in response to change emerging both from the environment and from the firm itself. Successful firms are those that successfully adapt as demands evolve, in this case, specifically with respect to their organization. Such adaptation requires flexibility and diversity within the firm, attributes that must be shaped by the firm's managers.

Introduction

The main focus of this article is the creation and development of organization within new and young firms. In this article, organizations are viewed as more or less enduring groups of people in a collective with a common identity, engaged in some common activity that is structured through a division of labor (differentiation) and distribution of authority. Self-sustaining organizations can be viewed as input–output systems, obtaining resources from their ecosystem and transforming them into outputs that bring in the returns that sustain the organization. A commercial organization must appropriate and retain some of the value it generates through market transactions (revenue) if it is to continue (Davis, 1973; Casson, 1982). Not-for-profit organizations that rely on donors or public revenues also must generate value of some kind if external resource providers are to continue supporting them. Private sector firms, the basic units of a market economy (Edquist, 2005), are the focus of this article, with special reference to new firms, which widely are acknowledged to contribute to economic growth and industrial revitalization.

Despite increasing media and academic attention to organizations that are new entrants to industries, there are still many gaps in our understanding of how new organizations emerge and develop. This can be attributed in part to the fragmented analysis to which this topic has been subjected and partly to misapprehensions as to what can be explained. Penrose (1959) showed that the specific determinants and behavior of a particular unit, such as a firm, is inherently unpredictable, but nevertheless, we can aim to gain a better understanding of *processes* of development.

Building on the work of Penrose (1995), we focus on the processes of change in new firms (Garnsey, 1998; Brush et al., 2001). Applying the ideas of Penrose to the development of new firms shows how emergence is driven by both internal and external dynamics. We first look at new firm formation and value creation, from the perspective of firm growth and organizational evolution. We then examine firm fitness in relation to its business environment, considering both the firm's ability to influence its environment and the firm's evolution to fit its ecosystem. We then discuss dominant approaches to the study of new firm formation. We conclude with recommendations for further study and policy.

New Firm Formation and Value Creation

For a firm to survive in a market economy, opportunities must be detected and input resources obtained to produce offerings that can be exchanged for income over and above the cost of production (Druilhe and Garnsey, 2006). Thus, the firm must create value for customers equal or superior to market alternatives and capture enough value to keep the firm in business, Opportunity recognition, the building of a productive base and creating and sustaining returns, will be discussed in the following sections along with the organizational evolution that follows each step.

Opportunity Recognition

Perception of an opportunity to create value, either one that currently exists or could be created, is often what triggers the process of new firm formation by the founder or entrepreneur (Penrose, 1959). Firms are sometimes founded, however, before the business idea has fully taken form, as might be the case if a firm is founded to exploit a new technology without a clear choice of application. The availability of a new resource may raise awareness of an opportunity; for example, advances in genetics drew the attention of microbiologists to new opportunities in pharmaceuticals. The perception of a new market need may trigger awareness of an opportunity; for example, the need for reliable energy in developing countries is increasingly inspiring technologists and entre-preneurs to find low-cost energy solutions. Once organi-zations start to develop, that opportunity detection process is unlikely to be the preserve of a single individual, and the extent to which the firm is able to develop effective learning processes becomes important (Lumpkin and Lichenstein, 2005).

Early Organizational Development

The early development of the firm involves both maturation and growth. Development refers to a process of maturation involving changes in structure and activity, while growth in a firm refers to an expansion in size or other indicators. Emergence of the firm may be preceded by a period of preparation; as opportunity scanning, consultation, and discussion often take place before start-up, the date marking the formal

 http://dx.doi.org/10.1016/B978-0-08-097086-8.73099-6

founding of the firm is to some extent arbitrary. If this process takes place, the maturity and market readiness of the business concept on start-up is variable. Serial entrepreneurs or those spinning their firm out of an existing organization (such as a large firm or university lab) may have developed and even tried out a business concept in a previous organization; alternatively the initiative may be pioneering. Some firms are started by first-time entrepreneurs or with technology still in need of further development; for example science-based university spinouts. Such firms may grow significantly in terms of resource endowment and employees while still being far from market, as occurs in the case of science-based firms in which the technology has a long gestation period, as in biopharm and new materials.

The development of the firm from original concept to operational entity involves cumulative problem solving; the original ideas for the business are modified frequently, either because of resource availability or because the entrepreneurs' assessment of the opportunity shifts. In contrast to this experimental mode, the corporate strategic planning that occurs in established firms assumes much greater consistency of direction as required for hierarchical control and accountability.

Preliminary scanning for business opportunities gives way to management tasks as the new firm matures. But this does not imply that problems are solved once and for all as 'stages of growth' perspectives imply. Entrepreneurs frequently revert to opportunity scanning when their goals come into question or if they start another line of business. Scanning opportunities for new product introduction is a continuous process when product life cycles are short.

The opportunities open to the new firm and how fast it is able to grow initially depends on resource endowments the founders bring to the start-up. When founders bring unequal contributions to the ventures' initial resource endowment, this should be reflected in the founders' agreement that records the new firm's ownership structure at the time of incorporation. Among the benefits of founding a company is that company law can be used to define ownership and avoid some of the complications of multiple ownership. For example, lack of clarity over intellectual property (IP) and other ownership issues can result in conflict later on. Limited liability and company law offer further benefits to a firm registered as a company, beyond those available to an unincorporated form of organization.

Since it takes time to produce outputs that generate revenues, a critical initial issue is how to cover the costs of start-up. For entrepreneurs, convincing funders of the prospects for their venture is the key to openings beyond their own immediate means. They are more likely to attract investors on terms acceptable to them if they can develop and add value to their own resources. A working prototype is more valuable than a design and a customer base makes a firm more valuable still. The ability to attract finance depends on the entrepreneur's ability to 'create confidence' in investors (Penrose, 1995: p. 38). If confidence is lost, investors can withdraw support or exert other sanctions. Investors with a large stake in a venture sit on the new firm's board of directors. An entrepreneurial chief executive can be replaced by a dissatisfied board. Entrepreneurs who want to retain autonomy may choose to make do without recourse to external investment, at the risk of having too few resources to realize extensive opportunities.

Building a Productive Base

A key challenge is to build a productive base from which to pursue opportunities (Garnsey et al., 2006). This productive base is made up of all resources, including human, financial, physical, technological, and organizational resources, that are required to seize the identified opportunity (Clarysee et al., 2005). With regard to human capital, entrepreneurs have to find partners, assistants, champions, and other potential resource providers and convince them that they too can benefit from the proposed business experiment (Lubik et al., 2011). Many studies show that prospects are better for ventures founded by a team of entrepreneurs (Storey, 1994; Reynolds and White, 1997). Depending on the sector and business model of the firm, more or less financial capital must be raised. Physical and technological resources can be accessed through a number of often-creative arrangements (Garnsey et al., 2007). The resources required will differ according to the activity undertaken; for example, firms that chose to manufacture a complete product are likely to require a much larger resource base, in terms of financial and physical resources, than a firm competing in the market for technology (licensing).

The revenue-generating activities a firm plans to undertake, such as licensing, service, manufacturing, or franchising, have implications for both requisite investment and competition. The production of new IP (as opposed to IP inherited or bought from the new firm's initial endowment) requires a distinctive base for productive activity, as does software production (Arora et al., 2001). Scientific research often requires a laboratory base in a research environment, which is very costly to duplicate in a new firm. For this reason, young technology firms are likely to engage in partnership arrangements. But entrepreneurs' specialist skills can constitute a large part of the productive base if they use their knowledge base for consulting, which allows them to continue the development of their core technology while generating early revenue streams. At the other extreme, if the productive base is highly capital intensive (i.e., requires significant external finance and investment in physical assets), this constitutes a barrier to entry for most entrepreneurs. Activities that have low barriers to entry, such as catering, are inevitably subject to greater competition. To address common problems, some start-ups replicate management procedures used elsewhere. For example, franchising arrangements provide detailed specifications for operational procedures. Franchising arrangements guide the creation of a productive base in a new unit and accelerate start-up through imitation, but they prevent further innovation.

Because many organizations are concerned with using knowledge to solve problems for producers and users, a productive base that relies more on knowledge rather than physical output is increasingly common and distinctive (Teece, 1998). Knowledge cannot remain exclusive once it has been shared, and hence there is a need to find ways to create IP that can be protected from competition and bring in license fees and royalties through patents and copyright. It is

unusual for licensing revenues to be sufficient to sustain a company, except with valuable new medical entities or well-protected generic technologies. Protection usually is afforded by extensive patenting portfolios, but patents are difficult for a new firm to defend. Corporate partners licensing the technology from a new firm will have an interest in protecting its IP (to prevent competitors from obtaining it illegally), deterring potential patent infringers, as was the case for the chip design company ARM.

Organizing Productive Activity

In a firm that is set up anew, with no precedents to follow, the firm's productive activity involves a division of labor and methods of coordination that start out fluid and experimental (Utterback, 1994). As experience is accumulated, a problem-solving repertoire develops and ways of carrying out tasks and requirements are embodied in procedure (Lumpkin and Lichenstein, 2005). Roles are assigned and rules are adopted to save time and improve coordination (Mintzberg, 1979: p. 2). By drawing on previous learning, the emergence of routines can improve the organization of production and revenue generation. Over time, routines affect outlook as well as procedure. Solutions that have been found to work come to be taken for granted and part of the unconscious assumptions of the organization. Tacit knowledge is conveyed to new members as the way to do things. Underlying assumptions about how things should be done make up the emerging culture of the firm, which is influenced by the outlook and preconceptions of the founding entrepreneurs.

At first, tasks are shared but as soon as a division of labor sets in, specialist functions emerge: finance, production, sales, marketing, and human resources. In this way, the firm begins to develop competence, which is the product of problem-solving activities. Competences enable the firm to respond to changing opportunities and threats. The new firm develops problem-solving skills and embeds these in procedures and routines. Routines can be viewed as "patterns of interactions that represent successful solutions to particular problems" (Teece et al., 1997) and become regular features of the organization's way of doing things. As the firm grows more complex, standardization of procedures is needed so that decision making can deal mainly with exceptions and the need for supervision can be reduced (Cyert and March, 1963). Official standard procedures are available for purchase and can help firms learn from practices used elsewhere.

Creating and Sustaining Returns

Problems of coordination and direction become more acute as the firm grows more complex and may require skills beyond those of the founders. Thus, the rate of return can be improved by speeding up throughput in more rapid production cycles, resulting in shorter lead times from input to invoice. But even if professional managers improve efficiency, further expansion is by no means ensured. The original offering of a service, product, or product range will have to be updated or extended. Many firms are unable to generate or mobilize sufficient additional resources to support the growth they need to make them less vulnerable (Garnsey and Heffernan, 2005). Sales growth is needed to achieve the minimum efficient scale (MES) of operations in an activity or in cases in which only a higher level of market share can secure the firm against competition. To achieve MES, the firm requires additional resources and must retain talent through opportunities for promotion. This requires development capital, which is never easy to obtain from external sources. Retained earnings provide the most autonomous source of investment capital but are unlikely to be sufficient for new companies with few reserves. Servicing the interest charges on bank loans raises costs. These challenges make it difficult for young firms to show an increase in profits even where their output and revenues are increasing.

Whether or not a primary aim of the entrepreneurs is profit, opportunity realization depends on returns that cover costs and provide reserves to see the firm through contingencies that endanger cash flow. The need to be profitable was brought home after the dot.com crash during which investors lost patience with ventures burning through initial investment without advancing toward profitability. Failure to show profits can result in a vicious cycle of investor and customer withdrawal, outcomes that distorted accounting practices attempt to conceal but instead may precipitate.

Not all new firms achieve the competitive advantage or secure a rent-generating position that enables them to accumulate resources in a growing product and asset base (Tidd et al., 1998). The advantage of the minority who do so is reinforced through reinvestment of retained earnings. New opportunities may be identified – positively affected by the increased knowledge base of the firm. This resource accumulation process provides necessary reserves for dealing with the potential resource shortfalls (Cyert and March, 1963). Organizational slack, the amount of resources the firm has in excess of the minimum required to remain operational, provides a buffer against external shocks and offers the means to explore new opportunities without endangering the current resource-generation process.

The firm itself becomes a useful asset once it has a market value as an entity (Lubik et al., 2012). If the firm initially was funded with venture capital (this applies to only a small minority of start-ups), their investors may insist on an early exit route. Realizing the value embodied in the firm is a challenge, as the value is only made manifest when on exit (via a buy out, a takeover, or bankruptcy) or if the firm launches on the stock market in an initial public offering. The advantage of going public is that the firms' shares become a liquid asset that can be bought and sold on the market. In return for funding, however, the founder managers may have to cede control of the firm to a board of directors that includes external investors. For the founders and managers of some young firms it is a problem, indeed a dilemma, that resources for growth come at the cost of loss of autonomy.

The Firm in Its Selection Environment – the Issue of Fitness

The firm is not passive in relation to its environment as the founding entrepreneurs choose their business environment in selecting their business opportunity (Penrose, 1959). Evolutionary theory represents the firm's environment as the arena in which selection forces are exerted; these reflect the fitness criteria that are required of firms in that environment

(Nelson and Winter, 1982). Selection processes are experienced as the firm interacts with resource providers (including investors, the knowledge environment, and labor sources), coproducers, customers, and competitors and regulators. If they cannot find others prepared to do business with them, they are eliminated from that environment. There is scope for agency by the entrepreneurial firm, however, which can be proactive in a number of ways, including selecting a business mode and choosing activities and business partners best suited to a particular environment. For example, a venture producing solar cells may become involved in lobbying and standards setting, thus influencing the selection criteria in its environment. The following section examines the firm's coevolution with its environment as its ecosystem and business evolves.

Ecosystem Evolution

The firm's environment is likely to be dynamic; with continual changes taking place in the firm's input markets and its output markets. Important aspects of input markets include the allocation of resources by resource providers, and the prospects for coproduction with other organizations or of outsourcing. The growth and decline of product markets, and the changing level of competition, affect opportunities and firms' prospects (Utterback, 1994). Important features of the new firm's environment include the operations of the capital market that affect conditions of exit, merger, and acquisition. A key question in this respect is how firms coevolve with others in the markets and networks in which they operate, and this depends on how the firm is organized to create and capture value in its business environment. For example, in the early days of the fuel cell market, new ventures often were forced to develop the full spectrum of innovations and complements required for application, from fuel cell-specific components to hydrogen generators. As the industry matured and more companies entered the market, however, some firms began to specialize in hydrogen production, others in fuel cells for specific applications, and others in distribution.

Organizational Evolution

The business model provides a concept that bridges the internal organization of the firm and the way the firm organizes relations and transactions with external players in its business environment (Amit and Zott, 2010), which also can be viewed as a business ecosystem (Adner and Kapoor, 2010). As the organization emerges, it relies on an implicit or explicit business model to enable it to create value. The extent to which a firm can implement the business model depends also on what Teece (2007) has described as the 'microfoundations' of enterprise performance "… the distinct skills, processes, procedures, organizational structures, decision rules, and disciplines …" (p. 1319).

The firm's business model represents the way the firm is organized to create and capture value (Teece, 2010) and specifies the firm's role in the wider value chain or web. In developing a business model, decision makers are 'modeling' the conditions in which the firm will operate, its competitors, and it complementors in the innovation ecosystem, and on this basis, they are deciding how to create and capture value. A business model that allows the firm to create and capture value is a good fit with its firm's ecosystem.

Dynamic value creation and capture in a firm is developmental and occurs through economies of scope, resource building, and harnessing complementarities through benefits from partnerships. These are anticipated in a predictive business plan and are operationalized in the business model, which may be deliberate or emergent. In young firms operating under high uncertainty, business models are likely to evolve in unanticipated ways. Similar business models reflect similar responses to the business environment. For example, many ventures in biotech conduct early research and testing with close ties to university research, leaving marketing and distribution to the incumbent pharmaceutical firms that are likely to acquire them if their new medical entities are successful in clinical trials.

The selection environment may select differentially for value creation and value capture over the course of business cycles. For example, there was plentiful seed capital and venture capital for start-ups that were far from realizing profit during the Internet boom, but resources were withdrawn when the dot.com ventures failed to create and capture value by producing products and services and reaching profitability. Surviving Internet firms learned from these selection force effects. The business models of successful firms provide proxy evidence of selection requirements in that environment. If the firm is to succeed, its business model must meet the fitness requirements of the business environment. If a market economy is to succeed, market forces must exert selection criteria that favor organizations creating value for users and investors – not only over the short term but also over the longer term. If selection forces exerted by the financial system favor short-term value capture over value creation, this erodes the ability of emerging firms to create wealth in the longer term.

Alternative Perspectives on Early Firm Growth

There are two dominant approaches for looking at the emergence of new firms: the stages of growth process and the developmental process approach. These two approaches (e.g., Greiner, 1972; Kazamjian, 1988) both focus on dominant problems in the life of new firms. When considering stages of growth, Kazamjian (1988) distinguished six categories of dominant problems to growth in technology-based new ventures: those related to organizational systems, sales and marketing, people, production, strategic positioning, and external relations. When critical problems are solved successfully, the firm is able to continue to grow. This type of study focuses on taxonomies of problems in distinctive stages and assumes that firms progress sequentially across a specific sequence of stages.

In contrast, the developmental approach focuses on processes of change and feedback effects that influence those changes. If similar developmental problems are addressed in sequence through similar phases of activity, regularities will be

observed, but firms operating a different business model or engaged in different modes of activity will evolve differently (Garnsey, 1998). Regularities can be observed in companies producing similar products on the basis of similar business models (e.g., in early US computer hardware ventures). But the building of different kinds of resource base involves different sequences of activity – for example, developmental sequences in a research consultancy differ from those in a company producing medical instruments in house. Moreover, ventures may or may not inherit a resource base from another organization through demerger or spinout, and this will affect the sequence of activity required. On the basis of these considerations, we subscribe to a developmental approach, holding that there are no invariant phases of activity for all new firms as they emerge; instead, there are common requirements for development into an economically viable unit, but this can be achieved in a variety of ways.

Just as there are common developmental processes at work as firms solve common problems, there are shared constraints to growth. Metcalfe (1998: p. 45) identified five specific kinds of limitation to growth. Limitations can be "in relation to the ability to purchase inputs and sell output as determined by the growth of relevant market environments; in relation to the availability of internal and external finance to expand capacity; in relation to the managerial implications of growth for the ability to control costs; in relation to the growth of rival firms and thus the specific market of the firm; and in relation to the ability to imagine and articulate growth opportunities" (Penrose, 1959). According to Metcalfe (1998), all of these elements come together to determine the economic fitness of the firm.

Conclusion

This article has discussed how new firms move from recognizing or creating a business opportunity to becoming a viable market actor, and it considered how they evolve as organizations. This leads us to consider avenues for further study and practical policy recommendations.

Economists examine the growth performance of new firms in the aggregate but do not enter the 'black box' of development processes within the firm. Entrepreneurship studies examine case histories, events, and situations, but they fail to set the firm in a wider context, which makes it possible to identify what features are specific to that case and what is an instance of more generic developments found in other cases. Strategy studies have only recently begun to address the issue of new firms and their growth; this issue increasingly is being addressed in terms of the business models of new organizations, the final topic of this article. Although the orthodox study of industry structure is relevant to understanding the business environment of new firms, the approach is too static to deal with the dynamic nature of the environment facing many new entrant organizations. The study of how organizations emerge shows the need for enterprises to adopt a dynamic approach to strategy (and the business models that operationalize strategy) in new organizations.

Policy interventions should consider the coevolution of the firm and its environment. Policy makers who aim to promote thriving new firms should ensure that markets and regulators are applying selection criteria that promote sustainable value creation by new firms. These include the availability of finance to cover the costs of start-up and early development before the firm can create value. Beyond competitive selection forces, partnerships are needed to enable newcomers to do business with established players in clusters and strategic alliances that promote open innovation. In the absence of such conditions, firms are unlikely to develop sustainable business models or to grow to a size at which a significant number of new jobs can be created.

Bibliography

Adner, R., Kapoor, R., 2010. Value creation in innovation ecosystems: how the structure of technological interdependence affects firm performance in new technology generations. Strategic Management Journal 31, 306–333.

Amit, R., Zott, C., 2010. Business model design: an activity system perspective. Long Range Planning 43, 216–226.

Arora, A., Fosfuri, A., Gambardella, A., 2001. Markets for technology and their implications for corporate strategy. Industrial and Corporate Change 11, 419–451.

Casson, M., 1982. Entrepreneur: an economic theory. Barnes & Noble Books, Totowa, NJ.

Clarysee, B., Wright, M., Lockett, A., Velde, E., Vohora, A., 2005. Spinning out new ventures: a typology of incubation strategies from European research institutions. Journal of Business Venturing 20, 183–216.

Cyert, R., March, J., 1963. A Behavioral Theory of the Firm. Prentice-Hall, Englewood Cliffs, NJ.

Davis, K., 1973. The case for and against business assumption of social responsibilities. Academy of Management Journal 16, 312–322.

Druilhe, C., Garnsey, E., 2006. Diversity in the emergence and development of academic spin-out companies: a resource-opportunity approach. In: Bernasconi, M., Harris, S., Moensted, M. (Eds.), High-Tech Entrepreneurship. Routledge, London.

Edquist, C., 2005. Systems of innovation. In: Fagerberg, J., Mowery, D., Nelson, R. (Eds.), The Oxford Handbook of Innovation. Oxford University Press, Oxford, UK.

Garnsey, E., 1998. A theory of the early growth of the firm. Industrial and Corporate Change 7, 523–556.

Garnsey, E., Heffernan, P., 2005. Growth setbacks in new firms. Futures 37, 675–697.

Garnsey, E., Dee, N., Ford, S., 2006. Clean Technology Ventures and Innovation. CTM Working Papers.

Garnsey, E., Heffernan, P., Stam, E., 2007. Firm, sector and cluster dynamics: the Cambridge case Presentation. European Network on the Economics of the Firm Conference. Cambridge, UK.

Kazamjian, R.K., 1988. Relation of dominant problems to stages of growth in technology-based new ventures. Academy of Management Journal 31, 257–279.

Lubik, S., Maine, E., Garnsey, E., 2012. Value creation in science-based ventures: a taxonomy of value indicators. R&D Management Conference. Grenoble, France.

Lubik, S., Garnsey, E., Minshall, T., Platts, K., 2011. Maximizing value creation from the innovation environment: partnership strategies from university spin-outs. R&D Management Conference. Norrköping, Sweden.

Lumpkin, G., Lichenstein, B., 2005. The role of organizational learning in the opportunity-recognition process. Entrepreneurship: Theory and Practice 29, 451–472.

Metcalfe, J.S., 1998. Evolutionary Economics and Creative Destruction. The Graz Schumpeter Lectures 1. Routledge, London.

Mintzberg, H., 1979. The Structuring of Organizations. Prentice-Hall, Englewood Cliffs, NJ.

Nelson, R., Winter, S., 1982. An Evolutionary Theory of Economic Change. Harvard University Press, Cambridge, MA.

Penrose, E., 1959. The Theory of the Growth of the Firm. Oxford University Press, Oxford, UK.

Penrose, E., 1995. The Theory of the Growth of the Firm. Resources, Firms and Strategies, third ed. Oxford University Press, Oxford, UK.

Reynolds, P., White, S., 1997. The Entrepreneurial Process. Quorum Books, Westport, CT.

Storey, D., 1994. Understanding the Small Business Sector. Routledge, London.

Teece, D.J., 1998. Capturing value from knowledge assets. California Management Review 40, 55–79.

Teece, D.J., 2010. Business models, business strategy and innovation. Long Range Planning 43, 172–194.

Teece, D.J., 2007. Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance. Strategic Management Journal 28, 1319–1350.
Teece, D.J., Pisano, G., Shuen, A., 1997. Dynamic capabilities and strategic management. Strategic Management Journal 18, 509–533.
Tidd, J., Pavitt, K., Bessant, J., 1998. Managing Innovation: Integrating Technological, Market and Organization Change. Wiley.
Utterback, J., 1994. Mastering the Dynamics of Innovation. Harvard Business School Press, Boston, MA.

Organizational Errors: Catastrophic

Karlene H Roberts, Berkeley University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses a relative newcomer to the social and behavioral sciences, the study of organizations in which errors can have catastrophic consequences. As we design ever more complex and technologically advanced organizations we create more opportunities for catastrophic error. The piece discusses precursors to present day social science research beginning with the sociological examination of the devastation of most of Halifax, Nova Scotia by an explosion in 1917. According to psychologist, Barry Turner, until 1978, the only interest in disasters was in responses to them. Today interest has broadened to include precursors, preventive strategies and mitigation.

The Past

Researchers and practitioners alike usually view accidents as slips, trips or falls. They then proceed to find the perpetrator and name, blame, train, or fire that person. Slips, trips, and falls, while taken seriously by their victims, are not usually harmful to the organizations in which they occur. Some accidents are seriously harmful to organizations. Take, for example, the 2010 BP oil spill in the Gulf of Mexico which flowed unabated for three months and may continue to seep. That accident killed 11 men, injured 17 others, and spilled 4.9 million barrels of crude oil (National Commission, 2011).

While catastrophic errors, risk, and reliability improvement have long been subjects of interest to practitioners and engineers, interest in these phenomena is newer in the social sciences. The first sociological approach to understanding accidents was Prince's (1920) dissertation on Canada's worst catastrophe, the 1917 munitions ship explosion in Halifax (Turner, 1978). This was the largest man-made disaster in the world to that date. Prince concentrates only on its aftermath. Prior to dropping atom bombs on Hiroshima and Nagasaki, the U.S. government used the Halifax accident to try to predict the sociological outcomes of these events.

In the United States just after World War (WW) II the sociological approach to the study of risk and catastrophes is primarily associated with E.L. Quarantelli and his colleagues. The many consistent themes in their publications include the myths of disaster behavior, the social nature of disaster, adaptation of community structure in the emergency period, dimensions of emergency planning, and differences among social situations that are conventionally considered as disasters (Dynes et al., 1994). The work primarily contributes to the sociological literature on collective behavior and social movement theory, and its focus is on how rents in the social fabric caused by a disaster can be repaired. In his seminal book, *Man-Made Disasters*, the British psychologist, Barry Turner, notes that until 1978, the only interest in disasters was in responses to them, and this was certainly true of the research mentioned above.

A corollary development that began during WW I was a focus on human factors which is a merger of engineering and experimental psychology. At first this field focused on finding men to fit systems. Later the field moved to examine both system and people characteristics which would allow the maximal match between men (and presumably later women) and systems. Such interests contributed to the psychological literature on occupational safety, which had a 'slips, trips, and falls' character.

Process safety emerged as a separate field from occupational safety in the 1980s (U.S. Department of Labor, 1994). It initially focused on understanding and explaining the nature and causes of major accidents and then on estimating their impacts. An emphasis was placed on plant and equipment design, which led to quantitative risk assessments that focused particularly on equipment failures. A focus on safety auditing systems resulted from the recognition of the importance of humans and was promoted internationally in the 1990s. Many industrial communities employed probabilistic risk assessment processes with increasing intensity until the early 1990s (Stamatelatos, 2000). Then these communities began to understand the limitations of such processes and search for improved ways to assess and manage the risks of their enterprises. These limitations were centered on the unpredictability and uncontrollability of the multitude of complex factors and interactions that determined the real risks these systems faced and on organizational factors.

Today such accidents are increasingly labeled 'process accidents.' They seriously disrupt or destroy the organizations in which they occur and ensuing clean-ups are hugely expensive. Such accidents are almost always rooted in the organization's fabric, and most often they have long periods of latency. An example is United Airlines DC-10 Flight 232 from Denver to Chicago on 19 July 1989. The plane suffered an engine failure and hydraulic system destruction and crash landed in Sioux City, Iowa, killing 115 people. 185 people survived. The engine failure was later traced back to a manufacturing defect in the fan disk, a defect that should have been discovered many times during regular aircraft maintenance (National Transportation Safety Board, 1990).

Recent Approaches to Error

There is a foreboding that these kinds of catastrophic accidents are on the rise. Man- and woman- kind build and try to manage ever more large and complex organizations. Greater size and complexity are associated with more errors that can have catastrophic consequences. Behavioral science researchers are

 http://dx.doi.org/10.1016/B978-0-08-097086-8.73126-6

increasingly concerned with finding ways to avoid such accidents or to mitigate their effects. Doing this requires the development of new conceptual and theoretical bases on which to understand the causes and consequences of such accidents. While this work has been ongoing since 1985 (see for example, Rochlin et al., 1987) it is still nascent.

There are two dominant approaches to understanding and avoiding catastrophic accidents. One is called high reliability theory (HRT) and the other normal accidents theory (NAT). HRT recognizes that some organizations in which catastrophic errors should occur, seem to avoid them and tries to understand how they do so. The underlying NAT logic is that if an accident can occur it will. Research from the first perspective focuses on such organizational processes as decision making, communications, how the organization is structured, etc. Research from the second perspective examines underlying potential causes of failure.

Both approaches suffer from the non-falsifiability problem. In neither case can you support the approach by showing how many accidents did or did not happen because no one has decided on the number of accidents or successes that would prove or disprove the theory. That NAT and HRT should be seen as competing theories seems silly to high reliability researchers. The competitive view began with the publication of Scott Sagan's book, *The Limits of Safety*, in 1993. He says:

> The first is the optimistic view of… "high reliability theory" whose proponents argue that extremely safe operations are possible, even with extremely hazardous technologies, if appropriate organizational design and management techniques are followed. The second school … "normal accidents theory" presents a much more pessimistic prediction: serious accidents with complex high technology systems are inevitable. (p. 13)

Clarke compares the two approaches in his review of *Limits of Safety* (Sagan) and *New Challenges to Understanding Organizations* (Roberts, 1993). He says:

> Sagan poses his central concern thus: "There has never been an accidental or unauthorized detonation of a nuclear weapon, much less escalation to an accidental nuclear war. Why?" The short answer: mainly luck (Clarke, 1993: p. 676).

About high reliability organizations(HRO) Clarke says:

> How do HROs do it? In my view, the force of the theory, as presented by HRO authors, is in a cultural factor and two structural ones. The cultural factor is represented by the metaphor of organizational learning; the structural factors are functional decentralization and redundancy (Clarke, p. 679)

Clarke then goes on to criticize both the HROs and the authors who discuss them by stating that both are filled with deception. Clarke concludes by pointing out that we need new ideas more than new data, and feels that Sagan provided those new ideas.

In 1994, the *Journal of Contingencies and Crisis Management* (vol. 2) published a four paper symposium airing the differences between NAT and HRO proponents. In a back and forth between Perrow (1994) and La Porte (1994) and then La Porte and Rochlin (1994) each author made his case. The basis of Perrow's argument is that technologically advanced organizations that can do harm to people and their environments are tightly coupled and complex. This combination of tight coupling and complexity leads to their failure. He also argues there are ways to minimize risk but that in the end complexity and coupling will doom the organization. Elite group interests and power, Perrow continues, allow the few to impose risk on the many for the benefit of the few. Under these circumstances, the elite may not do what is in the best public interest. Perrow concludes that the optimism of HRT is evidence that high reliability researchers do not consider power.

La Porte and then La Porte and Rochlin state that HRT and NAT are not competitive but complementary. La Porte and Rochlin say:

> …the HRO group takes as its starting point the perception by the organization that it is inherently susceptible to accidents, including very serious ones. We then attempted to find out what strategies have been developed by organizations that have demonstrated high levels of performance in the face of demanding conditions…. (p. 221).

Sagan (1994) is the bookend to this symposium. He agrees with Perrow and then offers suggestions for future research. He says that the pull of power and interests in organizational politics might complicate matters and should be included in research. He adds that we need much more study of redundancy. When does it backfire? When does adding more redundancy create common mode failures? And when does it add unreliability because individuals or subunits shirk their duties knowing that redundancy backs them up. Sagan claims that organizational learning is difficult in these advanced technological systems and suggests that perhaps introducing a political approach may uncover barriers to learning.

Since that time there have been a number of comparisons of NAT and HRO (e.g., and Leveson et al., 2009), and a number of attempts to merge the two approaches (e.g., Brady and Davies, 2010; Rosa, 2005).

In the 2000s more research has been devoted to teasing out high reliability processes and constructs than to unpacking NAT. This may be, in part, because results of early research on high reliability organizing were filtering their way into major accident investigations and major policy recommendations. For example, the *Columbia Accident Investigation Board's* (2003) discussion of organizational contributions to the accident is quite consistent with early high reliability findings. In 2008, the *Agency for Healthcare Research and Quality* (AHRQ) published a lengthy report entitled 'Becoming a High Reliability Organization: Operational Advice for Hospital Leaders'. AHRQ's mission is to support research to improve quality, safety, efficiency, and effectiveness in healthcare for all U.S. citizens. The report cites Weick and Sutcliffe's (2001) discussion of challenges organizations face that need to operate highly reliably:

Hypercomplexity. HROs exist in complex environments that depend on multi team systems that must coordinate for safety.

Tight coupling. HROs consist of tightly coupled teams in which the members depend on tasks performed across their team.

Extreme hierarchical differentiation. In HROs, roles are clearly differentiated and defined. Intensive coordination efforts are needed to keep members of the teams working cohesively. During times of crisis, however, decision making is deferred to the most knowledgeable person on the team, regardless of their position in the organization.

Multiple decision makers in a complex communication network. HROs consist of many decision makers working to make important, interconnected decisions…. HROs must develop processes that allow these decision makers to communicate effectively with each other.

High degree of accountability. HROs have a high degree of accountability when an error occurs that has severe consequences.

Need for frequent, immediate feedback. HROs exist in industries where team members must receive frequent feedback. This feedback and the opportunity to make continuous adjustments based on it are essential to anticipate and avert problems before they become crises.

Compressed time constraints. Time constraints are common to many industries… In HROs, the systems and culture allow people to identify when they lack time to reliably complete all needed tasks and obtain additional assistance (AHRQ, 2008).

The AHRQ report then goes on to mention the five reliability enhancing characteristics discussed by Weick and Sutcliffe:

Sensitivity to operations. HROs recognize that manuals and policies constantly change, and are mindful of the complexity of the systems in which they work. HROs work quickly to identify anomalies and problems in their system to eliminate potential errors. Maintaining 'situational awareness' is important for staff at all levels because it is the only way anomalies, potential errors, and actual errors can be quickly identified and addressed.

Reluctance to simplify. HROs refuse to simplify or ignore the explanations for difficulties and problems that they face. These organizations accept that their work is complex and do not accept simplistic solutions for challenges confronting complex and adaptive systems. They understand that their systems can fail in ways that have never happened before and that they cannot identify all the ways in which their systems could fail in the future.

Preoccupation with failure. HROs focus on predicting and eliminating catastrophes rather than reacting to them. These organizations constantly entertain the thought that they may have missed something that places people and their organizations at risk. 'Near misses' are viewed as opportunities to improve current systems by examining strengths, determining weaknesses, and devoting resources to improve and address them.

Deference to expertise. HROs cultivate a culture in which team members and organizational leaders defer to the person with the most knowledge relevant to the issue they confront. The most experienced person or the person highest in the organizational hierarchy does not necessarily have the information most critical to responding to a crisis.

Resilience. HROs pay close attention to their ability to quickly contain errors and improvise when difficulties occur. Thus, systems can function despite setbacks. An HRO assumes that, despite considerable safeguards, the system may fail in unanticipated ways. They prepare for these failures by training staff to perform quick situational assessments, working effectively as a team that defers to expertise, and practicing responses to system failures (Weick and Sutcliffe, 2001).

While HRO research began by in depth studies of three organizations that were identified as reaching or exceeding their reliability goals today both the contexts and issues in this research area have broadened. The first investigations were of U.S. Navy aircraft carrier operations, operating a commercial nuclear power plant, and the Federal Aviation Administration's air traffic control system. Today this kind of research is done in healthcare facilities (e.g., Vogus et al., in press), banks, mining operations, software firms (e.g., Vogus and Welbourne, 2003), the construction of Heathrow Airport's terminal 5 (Brady and Davies, 2010), etc. While early studies focused almost exclusively on organizational structuring, culture, and decision making, more recent studies examine such issues as conceptual under specification, measurement, counter factuals, organizational learning, team dynamics, etc.

W.R. Scott (1994) noted "one of the puzzling aspects of the growing literature on organizations prone to accidents…is the lack of connection to mainstream organization theory… the study of HRO needs to be better integrated into the study of organizations in general (p. 25)". As the research broadens out to include new issues it necessarily examines more mainstream organizational issues such as those mentioned above. Another criticism of the work is that it relies on case study methodology. That, too, has broadened out. We see a growing number of methodological approaches (e.g., Morgeson et al., 2011).

What's Missing

A major missing piece is action research that would provide managers and scholars with information about reliability processes that can be implemented with favorable outcomes. Most attempts at implementation of organizational research fail. In this area there is no scientifically based information on what to implement and how.

Today the research takes both micro and macro perspectives. It fails to take a systemic perspective. That is, it fails to examine the interdependencies of units or organizations working with one another, thus making the error of the third kind (Mitroff and Silvers, 2010), or solving the wrong problem precisely.

Returning to the BP example which opened this article, as night fell on 20 April 2010 a series of explosions rocked the semisubmersible drilling platform leased by BP from Transocean Ltd. The accident resulted in a sea floor oil gusher and is the largest accidental marine oil spill in the history of the petroleum industry. To date BP and Transocean have paid the U.S. Government approximately USD 9.2 billion in settlement fees (Reuters, 2013). A number of books have been written about the accident, numerous radio and television reports

were filed about it, Congress took testimony by parties to the case and a plethora of law suits were filed.

No one has examined BP in the context of its interactions with its regulatory agency, the Minerals Management Service (MMS), now the Bureau of Safety and Environmental Enforcement (BSEE), with its subcontractors, or with other companies in the same industry. Many large organizations are deeply interlinked with other organizations and understanding one organization in the set provides an incomplete picture. Some macro sociologically based organizational research does examine sets of organizations (e.g., network research, population ecology research) but those perspectives have not found their way into HROs research. It is unlikely to do so because of its perceived cost and the amount of effort required that would probably prevent any academically based researcher from further promotion.

What's Needed?

A number of things are needed to move this body of research along. Moving from the easiest to the more difficult approaches, work needs to continue to delineate what is meant by high reliability processes and to think about whether current labeling draws attention away from what we are really doing. Some writers have recently begun to differentiate between HROs (HROs, those that have reached this state) with reliability seeking organizations (RSOs) (Vogus and Welbourne, 2003). Unlike HROs, which operate high-hazard technologies, RSOs operate in high-hazard environments. RSOs are tightly coupled to their unpredictable and complex environments in such a manner that although the human mortality

> rate is low, the risk of small failures amplifying into organizational mortality is high. To cope with these environments, reliability-seeking organizations organize to remain open and flexible to emerging information and achieve the reliability demanded by their environments – intensity of innovation (Vogus and Welbourne, p. 877).

These differentiations focus attention on different associated processes. Research emanating from the HRO perspective is still in search of those processes associated with reliable performance. Research based on the RSO perspective focuses on processes involved with what we have to do to get there.

This leads us to think about taking a broader perspective on error by considering the language we adopt doing this kind of research. What do we call 'errors' and what do we call 'incidents'? These two words have different emotional pulls as do many other words we often use interchangeably in this line of research. This suggests a line of experimental research in which subjects evaluate the same incidents using different descriptive scenarios.

The role of emotions in high reliability organizing and in reliability seeking processes has, to date, been avoided. Yet, there is no doubt that different emotions are provoked in wildfire fighting, piloting aircraft in war time, the precision landing of aircraft, medical emergencies, operating complex technologically advanced equipment near their levels of maximum capacity, etc., than in product sales, housekeeping, and many other tasks done in organizations every day. We are long overdue in examining the roles of such emotions as guilt and shame in relation to reporting errors.

As yet there has been little effort to systematically examine the environmental requisites for highly reliable or reliability seeking organizations. The three organizations in the initial research are strongly regulated. When one moves to more market oriented organizations there may be no (e.g., healthcare) or weak (e.g., petroleum exploration) regulation. This undoubtedly influences an organization's attention to safety processes. Other aspects of the environment certainly influence how organizations think about reliability, such as the stock market, other external incentives, industry practices, etc.

Finally, research on organizational reliability is inherently interdisciplinary. Even within the social sciences both the journals and academic institutional arrangements are stove piped.

The problems requiring research solutions are not stove piped. Legal issues interact with organizational, personality, motivational, training, cultural, engineering design, and other issues. Looking again at the BP Gulf accident, all of these issues in addition to the systemic issues discussed would have to be examined to fully understand the accident.

See also: Corporate Culture; Corporate Governance; Organization: Overview; Organizations: Unintended Consequences; People in Organizations; Wellbeing and Burnout in the Workplace: Organizational Causes and Consequences; Workplace Safety and Health.

Bibliography

Agency for Healthcare Research and Quality, 2008. Becoming a High Reliability Organization: Operational Advice for Hospital Leaders (AHRQ 08-0022).

Brady, T., Davies, A., 2010. From hero to hubris – reconsidering the project management of Heathrow's terminal 5. International Journal of Project Management 28, 151–157.

Clarke, L., 1993. Drs. Pangloss and Strangelove meet organizational theory: high reliability organizations and nuclear weapons accidents. Sociological Forum 8, 675–689.

Columbia Accident Investigation Board Report, 2003. U.S. Government Printing Office, Washington, D.C.

Dynes, R.R., Tierney, K.J., Fritz, C.E., 1994. Forward: the emergence and important of social organization: the contributions of E.L. Quarantelli. In: Dynes, R.R., Tierney, K.J. (Eds.), Disasters, Collective Behavior, and Social Organization. University of Delaware Press, Newark, Delaware, pp. 9–17.

La Porte, T., 1994. A straw man speaks up. Journal of Contingencies and Crisis Management 2, 207–211.

La Porte, T., 1996. High reliability organizations: unlikely, demanding, and at risk. Journal of Contingencies and Crisis Management 4, 60–72.

La Porte, T., Rochlin, G., 1994. A rejoinder to Perrow. Journal of Contingencies and Crisis Management 2, 221–227.

Leveson, N., Dulac, N., Marris, K., Carroll, J., 2009. Moving beyond normal accidents and high reliability organizations. Organization Studies 30, 205–226.

Mitroff, I., Silvers, S., 2010. Dirty Rotten Strategies: How We Trick Ourselves and Others into Solving the Wrong Problems Precisely. Stanford University Press, Stanford, California.

Morgeson, F.P., Narhgang, J., Hofmann, D.A., 2011. Safety at work: a meta analytic investigation of the link between job demands, job resources, burnout, engagement, and safety outcomes. Journal of Applied Psychology 96, 71–94.

National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011. Deep Water: the Gulf Oil Disaster and the Future of Offshore Drilling. U.S. Government Printing Office, Washington, D.C.

National Transportation Safety Board, 1 November 1990. Aircraft Accident Report: UAL F 232 McDonnell Douglas DC 10–10 Sioux Gateway Airport, Sioux City, Iowa, January 19, 1989.

Perrow, C., 1994. The limits of safety: the enhancement of a theory of accidents. Journal of Contingencies and Crisis Management 2, 212–220.

Prince, S.H., 1920. Catastrophe and Social Change: Based upon a Sociological Study of the Halifax Disaster. Columbia University Press, New York.

Reuters, 3 January 2013. Transocean to pay 1.4 billion dollars for role in BP oil spill. Chicago Tribune.

Roberts, K.H. (Ed.), 1993. New Challenges to Understanding Organizations. Macmillan, New York.

Rochlin, G.I., La Porte, T.R., Roberts, K.H., 1987. The self-designing high-reliability organization: aircraft carrier flight operations at sea. Naval War College Review 40, 76–90.

Rosa, E.A., 2005. Celebrating a citation classic and more. Organization and Environment 18, 229–234.

Sagan, S., 1993. The Limits of Safety. Princeton University Press, Princeton, New Jersey.

Sagan, S., 1994. Toward a political theory of organizational reliability. Journal of Contingencies and Crisis Management 2, 228–243.

Scott, W.R., 1994. Open peer communication on "accidents in high risk systems,". Technology Studies 1, 23–25.

Stamatelatos, M., 2000. Probabilistic Risk Assessment: What Is It and Why Is It Worth Performing It. NASA, Washington, D.C.

Turner, B.M., 1978. Man-Made Disasters. Wykeham, London.

U.S. Department of Labor: Occupational Safety and Health Administration, 1994. Process Safety Management Guidelines for Compliance (OSHA 3132).

Vogus, T., Welbourne, T., 2003. Structuring for high reliability: HR practices and mindful processes in reliability-seeking organizations. Journal of Organizational Behavior 74, 877–903.

Vogus, T., Sutcliffe, K., Weick, K.E., in press. Searching for safety culture: an integration and research agenda. Academy of Management Annals.

Weick, K.E., Sutcliffe, K., 2001. Managing the Unexpected: Assuring High Performance in an Age of Complexity. Jossey Bass, San Francisco.

Organizational Fields

Melissa E Wooten, University of Massachusetts, Amherst, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Organizational field scholarship illuminates the process by which globally accessible templates of behavior emerge and influence organizational action. An interest in understanding the sources of constraint for organizational adaptation lie at the core of this literature. Early scholarship focused on how the environment or institution shaped organizations. The institution determined whether fields were marked by stability or tumult. Alternatively, recent scholarship depicts a reciprocal relationship between fields and institutions placing greater emphasis on the ways in which interactions at the field level shape institutions. The organizational field remains a vital unit of analysis for detecting institutional effects and change.

Introduction

Is an organization free to adapt in any manner of its choosing? This question lay at the heart of any investigation situated within the organizational field literature. The empirical setting may vary from Fortune 500 companies (e.g., Davis, 1991) to city municipalities (e.g., Tolbert and Zucker, 1983). The topic of interest may range from defensive managerial practices (e.g., Davis and Greve, 1997) to maternity leave policies (e.g., Dobbin and Kelly, 1999). Yet, underneath these differences remains the same quest to understand organizational adaptation. By and large, organizational field theorists have concluded that even if they must and can do so, organizations are constrained in their ability to adapt. Customer preferences in the automotive industry may shift toward smaller more fuel-efficient vehicles. As US manufacturers can attest, however, years of investment in equipment, facilities, and relationships tailored toward large vehicle production present serious adaptive challenges. No matter how inclined an organization is toward change, the environment can erect insurmountable barriers.

Foundations

Perhaps the most oft-cited definition of an organizational field stems from Paul DiMaggio and Walter Powell's 1983 *American Sociological Review* article "The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields." Drawing on insights from Max Weber's *Economy and Society*, DiMaggio and Powell (1983) sought to understand the modern-day sources of bureaucracy and rationalization among organizations. At the time, organization theorists primarily were concerned with explaining variation in organizational structure and behavior (e.g., Lawrence and Lorsch, 1967; Hannan and Freeman, 1977). Instead, DiMaggio and Powell (1983) proposed to explain the lack of diversity among organizational forms and practices, that is, they asked, "Why had organizations become so similar?"

In their quest to identify the forces driving homogeneity, focus went to the organizational field or "those organizations that, in the aggregate, constitute an area of institutional life: key suppliers, resource and product consumers, regulatory agencies, and other organizations that produce similar services or products" (DiMaggio and Powell, 1983: 148). Because it placed emphasis on the "totality of relevant actors," the term organizational field was ideal for understanding the sources of homogeneity as lying outside the boundaries of any one organization. Various actors, including the state, professional associations, and competitors exert pressures on organizations that over time lead them to adopt similar procedures, programs, and policies. Despite the existence of more efficient or rational ones, organizations feel compelled to incorporate practices solely because they have received the endorsement of important field members (Meyer and Rowan, 1977).

In his 1994 essay "Conceptualizing Organizational Fields," W. Richard Scott defined organizational fields as "a community of organizations that partakes of a common meaning system and whose participants interact more frequently and fatefully with one another than with actors outside of the field" (1994, pp. 207–208). Less concerned with explaining the sources of bureaucracy, Scott's definition highlights the relational nature of the interactions within fields and how shared understandings of a situation bind organizations to one another.

The conceptual definitions of organizational fields developed by DiMaggio and Powell and Scott have empirical implications. Both definitions highlight an aggregate of organizations, yet, otherwise offer few restrictions on what aggregates can and cannot be considered organizational fields. This lack of empirical specificity has resulted in a diverse array of organizational aggregates that have served as the empirical settings for field studies. Scholars have taken organizations listed on the Fortune 500 (Davis, 1991), the largest firms (Fligstein, 1991), organizations that share an interest (Hoffman, 1999), and geographically proximate organizations (Lounsbury, 2001) as members of a field.

Although scholarship on organizational fields experienced a renaissance during the past decades of the twentieth century, the concept had existed in one form or another for far longer within the organization studies tradition. Scott's imagery of a field harkened back to Ronald L. Warren's 1967 essay the "Interorganizational Field as a Focus for Investigation," published in *Administrative Science Quarterly*. Writing about community organizations, such as local banks, churches, and schools, Warren drew attention to situations in which organizations had dissimilar goals, yet found it advantageous to interact with one another to accomplish a given task like revitalizing a defunct business district. A shared understanding of the value of a thriving economic zone can provide disparate

organizations with a reason to work together. Warren's work in turn drew from that of Selznick.

Philip Selznick's (1949) study of the Tennessee Valley Authority exemplified not only the durability but also the evolution of the construct. Writing in the 1940s, like DiMaggio and Powell, Selznick concerned himself with the environment's effect on organizations, particularly its ability to constrain choice and agency. Yet, reflective of the time in which he wrote, Selznick's conception of the environment was locally rooted and understood as the interests, values, and attitudes of organizational constituents. For Sezlnick, organizations did not gain legitimacy by adopting practices deemed legitimate by some omnipresent forces within the field. Instead, organizations gained legitimacy by ensuring that their constituents' values and beliefs were reflected in the organizational goals. Here, the environment did not provide a generic template of action. No, the environment encapsulated people and groups with the power to change the agenda of an organization.

Within the more recent instantiations of organizational field scholarship, the environment has expanded. Technological advancements have made it such that local or even national borders no longer bind an organization's activities. Products made in one country are easily accessible to consumers in another. Employees seamlessly move between firms and across industries over the course of a career. Organizations exist within a globally defined and generated culture from which to draw templates of action (Weber, 2005).

Empirical Interests

Key to any investigation of organizational fields is the relationship between fields and the institutional literature more broadly. Although often thought of as interchangeable, organizational fields and institutions are in fact distinct. Defined as actions (Zucker, 1977), as facts that must be taken into account by social actors (Meyer and Rowan, 1977), and as the structures that provide stability and meaning to social behaviors (Scott, 1995, 2001), institutions are enduring features of social life (Giddens, 1979). At its core, institutional theory seeks to explain the process of institutionalization – that is, how actions, facts, or structures come to take on a rulelike status within society. Institutions structure an organization's environment (Meyer and Rowan, 1977), and within this paradigm, the organizational field serves as a level of analyses for investigating an institution's effects. Methodologically speaking, the organizational field becomes the unit of observation for analyses in which scholars ultimately hope to learn something about the institution.

Although subtle, the difference between institutions and organizational fields is important for it is through observations at the field level that we are able to learn about the effects of institutions. Many erroneously conclude that the primary takeaway from the organizational field literature is that fields are stable. Yet, organizational fields constantly shift. If fields appear at all stable, it is only because the effect of the institution in which these organizations exist is constant. Marquis' (2003) study of intercorporate networks demonstrated this quite well. Firms located in communities that began before the era of auto and air travel had more locally based director connections than firms located in communities that began after auto and air travel became prevalent. This result did not occur because there had been no change at the level of the field – many firms had entered and exited these particular communities. This result occurred because the cultural–cognitive templates or institutions that guided the construction of intercorporate networks remained constant within these communities.

Identifying the Effect of Institutions on Fields

Of particular interest to scholars concerned with investigating institutional effects has been exploring the mechanism proposed to account for homogeneity in structure and outcomes among field members. DiMaggio and Powell (1983) argued that coercive, normative, and mimetic forces led organizations to become institutionally isomorphic. A series of essays by Edelman (1990, 1992) provided a classic example of this investigative approach. Edelman sought to explain the proliferation of affirmative action and equal employment offices following the implementation of the employment-related sections of the 1964 Civil Rights Act. Through their professional associations, personnel officers touted the benefits of using bureaucratic policies to guide the recruitment and retention of women and minorities in the workforce. Nowhere in the act had Congress mandated that employers establish such policies. Yet, government authorities, including the Office of Federal Contract Compliance, the Justice Department, and the Equal Employment Opportunity Commission, came to view these policies as a necessary component in the fight to end employment discrimination. As a result, these policies and the offices overseeing them multiplied among organizations subject to the legislation.

Each of the three mechanisms thought to produce homogeneity played an essential role in the organizational response to equal employment legislation. Professional norms dictated that personnel officers respond to the uncertainty of the legislation in a particular manner. Once a critical number of companies had instituted bureaucratic policies and structures to oversee the hiring of women and minorities, other companies mimicked this behavior. As the bodies responsible for monitoring compliance with the employment regulations, agencies like the Office of Federal Contract Compliance and the court systems' endorsement of these bureaucratic policies and structures further exacerbated their use. The institution, primarily understood as regulatory agencies and courts overseeing these companies and personnel professionals working within them, encouraged the use of formal procedures to regulate the hiring and retention of women and minorities.

The perspective that institutions had a homogeneous effect on organizations dominated the literature for many years until scholarship on nuance and change emerged. Scholars such as Czarniawska and Jorges (1996) and Zilber (2006) challenged the theoretical understanding of diffusion within fields by arguing that organizational practices are not adopted intact, but instead are translated. Just as literal translations from one language to another often have no meaning, adopting a prevailing practice 'as is' will not work. Field members must

determine how to bend and shape a prevailing practice such that it will work given the particulars of their organizational context. Although some focused on complicating our understanding of the processes by which practices spread through fields, others looked to challenge our understanding of institutions as stable, inert entities.

Investigations illustrated that institutions vary, and support for particular organizational configurations are contingent. To illustrate, Kraatz and Zajac (1996) argued that for much of their existence, the institution – conceptualized primarily as student interest or consumer demand – called for liberal art colleges to prepare students for humanistic inquiry. As student career preferences shifted away from valuing liberal education toward a more occupationally rooted education experience, the institution no longer supported a focus on humanistic inquiry. Instead, the institution made it such that, to survive, liberal arts colleges had to institute vocationally and professionally focused courses (Kraatz and Zajac, 1996).

In a similar vein, building on the work of Friedland and Alford (1991), Thornton and Ocasio (1999) argued that institutions have historically variant logics. Economic and social structural shifts create a space for new logics to arise. As these new logics take hold, at the level of the field, organizational behavior changes. For instance, Thornton (2001) found that acquisition patterns within the higher education publishing industry varied according to the dominant institutional logic. When a market logic dominated the industry, publishers following an imprint or distribution strategy faced a higher risk of acquisition than others. When an editorial logic dominated the industry, imprint and distribution strategies had no significant effect on a publisher's likelihood of acquisition. As the logic changed, the pattern of acquisitions changed within the field.

In another instance, Scott et al. (2000) argued that between 1945 and 1995, institutional logics privileged particular organizational configurations for health care delivery over others. The authors linked the development of new organizational forms in the San Francisco Bay area health care field to the changing institutional environment. In the 1940s, hospitals represented the major organizational form within the field and with the exception of Kaiser Permanente no other health maintenance organization (HMO) existed. As the 1970s approached, however, HMOs became prominent within the field. The increased presence of HMOs corresponded to a shift away from a logic that privileged the professional dominance of doctors to one that privileged markets and the efficiency of service provision.

Identifying the Effect of Fields on Institutions

Scholarship pointing to nuance and change illustrated a more reciprocal process in which fields influence institutions and institutions influence fields. Field members have their own political and organizational interests (DiMaggio, 1988) that lead them to respond strategically to their environments (Oliver, 1991). As field members pursue their interests, the institution itself changes. This perspective has coalesced more recently under the rubric of institutional work. Key to this area of inquiry is the desire to challenge the conceptualization of field members as relatively powerless to resisting the isomorphic tendencies produced by institutions (Lawrence et al., 2009). Focusing on fields in this manner makes it clear that identifying the mechanisms responsible for producing homogeneity is not the only means by which to understand the relationship between fields and institutions. It is at the level of the field that institutions are created, maintained, and disrupted (Lawrence and Suddaby, 2006).

Focusing on the dynamics and consequences of the structuration process among organizations further illuminates the reciprocal relationship between fields and institutions. Structuration studies aim to understand how fields come together. Moreover, investigating structuration gives insight into how the process of field formation gives rise to particular institutional definitions. Hoffman's (1999) study focused on the development of the corporate environmentalism field. The actors within the field included petrochemical companies, environmental activists, regulatory agencies, investors, and insurers. As each actor entered the field, they brought their own understanding of corporate environmentalism to it. Thus, the process of field structuration became seen as not just a means to establish connections among the various actors – that is, creating the field – but also as a process that had implications for the definition of corporate environmentalism as an issue.

Likewise, Wedlin (2006) argued that the structuration process is about more than establishing connections among similar organizations. The structuration process acts as a mechanism that enables potential field members to debate and ultimately classify the organizations that should and should not belong to a given community. As the importance and desirability of certain organizational characteristics and attributes is weighed, the field's boundaries meld. To the extent that an organization is understood to possess the favored attributes, it can gain membership in this community. Consequently, the process of structuration leads to institutional definitions of organizational and field identity.

Conclusion

Initially developed to explain how external parties altered an organization's goal-setting apparatus, organizational field scholarship now illuminates the process by which globally accessible templates of behavior emerge as well as their influence on organizational action. Despite this shift in the literature, the interest in understanding the sources of and constraints on organizational adaptation remains at its core. Common across any investigation of organizational fields is the desire to identify the organizations, collectives, and individuals that interact with one another to produce some outcome – be it homogeneity or the development of a shared understanding.

Essential to this literature is the proposition that organizational action is contingent on the environment or the institution in which it resides. The institution dictates whether the organizational field is marked by the dynamics of rapid change or constant stability. Although historically dependent on itself, the institution shapes and molds the possibilities of organizations at the level of the field. Once conceptualized

as a unidirectional process, scholars now understand the relationship between fields and institutions as reciprocal. Institutions themselves are created, maintained, and disrupted through field members' interactions with one another. Accordingly, the organizational field remains a vital unit of analysis for detecting institutional effects and change.

See also: Bureaucracy and Bureaucratization; Bureaucracy, Sociology of; Corporate Culture; Ecology: Organizations; Narrative Approaches to Organizations; Network Paradigm: Applications in Organizational Science; Organization: Overview; Organizations, Sociology of.

Bibliography

Czarniawska, Barbara, Jorges, Bernward, 1996. Travels of ideas. In: Czarniawska, Barbara, Sevon, Guje (Eds.), Translating Organizational Change. Walter de Gruyter, Berlin, pp. 13–48.

Davis, Gerald F., 1991. Agents without principles? The spread of the poison pill through the intercorporate network. Administrative Science Quarterly 36, 583–613.

Davis, Gerald F., Greve, Henrich R., 1997. Corporate elite networks and governance changes in the 1980s. American Journal of Sociology 103, 1–37.

DiMaggio, Paul J., 1988. Interest and agency in institutional theory. In: Zucker, Lynn (Ed.), Institutional Patterns and Organizations. Ballinger, Cambridge, MA, pp. 3–21.

DiMaggio, Paul J., Powell, Walter W., 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. American Sociological Review 48, 147–160.

Dobbin, Frank, Kelly, Erin, 1999. Civil rights law at work: sex discrimination and the rise of maternity leave policies. American Journal of Sociology 105, 455–492.

Edelman, Lauren, 1990. Legal environments and organizational governance: the expansion of due process in the American workplace. American Journal of Sociology 95, 1401–1440.

Edelman, Lauren, 1992. Legal ambiguity and symbolic structures: organizational mediation of civil rights law. American Journal of Sociology 97, 1531–1576.

Fligstein, Neil, 1991. The structural transformation of American industry: an institutional account of the causes of diversification in the largest firms, 1919–1979. In: Walter W. Powell and Paul J. DiMaggio (Eds.), The New Institutionalism in Organizational Analysis. The University of Chicago Press, Chicago, pp. 311–336.

Friedland, Roger, Alford, Robert R., 1991. Bringing society back in: symbols, practices, and institutional contradictions. In: DiMaggio, Paul J., Powell, Walter W. (Eds.), The New Institutionalism in Organizational Analysis. The University of Chicago Press, Chicago, pp. 232–266.

Giddens, Anthony, 1979. Central Problems in Social Theory: Action, Structure, and Contradiction in Social Analysis. University of California Press, Berkeley, CA.

Hannan, Michael T., Freeman, John, 1977. The population ecology of organizations. American Journal of Sociology 82, 929–964.

Hoffman, Andrew J., 1999. Institutional evolution and change: environmentalism and the U.S. chemical industry. Academy of Management Journal 42, 351–371.

Kraatz, Matthew S., Zajac, Edward J., 1996. Exploring the limits of the new institutionalism: the causes and consequences of illegitimate organizational change. American Sociological Review 61, 812–836.

Lawrence, Paul R., Lorsch, Jay W., 1967. Differentiation and integration in complex organizations. Administrative Science Quarterly 12, 1–47.

Lawrence, Thomas B., Suddaby, Roy, 2006. Institutions and institutional work. In: Clegg, Stuart, Hardy, Cynthia, Lawrence, Thomas, Nord, W.R. (Eds.), Handbook of Organization Studies, second ed. Sage, Thousand Oaks, CA, pp. 215–254.

Lawrence, Thomas B., Suddaby, Roy, Leca, Bernard, 2009. Introduction: theorizing and studying institutional work. In: Lawrence, Thomas, Suddaby, Roy, Leca, Beranard (Eds.), Institutional Work: Actors and Agency in Institutional Studies of Organization. Cambridge University Press, Cambridge, UK, pp. 1–27.

Lounsbury, Michael, 2001. Institutional sources of practice variation: staffing college and university recycling programs. Administrative Science Quarterly 46, 29.

Marquis, Christopher, 2003. The pressure of the past: network imprinting in intercorporate communities. Administrative Science Quarterly 48, 655–689.

Meyer, John W., Rowan, Brian, 1977. Institutionalized organizations: formal structure as myth and ceremony. American Journal of Sociology 83, 41–62.

Oliver, Christine, 1991. Strategic responses to institutional processes. Academy of Management Review 16, 145–179.

Scott, W. Richard, 1994. Conceptualizing organizational fields: linking organizations and societal systems. In: Derlien, Hans-Ulrich, Gerhardt, Uta, Scharpf, Fritz W. (Eds.), Systems Rationality and Partial Interests. Nomos Verlagsgesselschaft, Baden-Baden, Germany, pp. 203–221.

Scott, W. Richard, 1995. Institutions and Organizations. Sage Publications, Thousand Oaks, CA.

Scott, W. Richard, 2001. Institutions and Organizations. Sage Publications, Thousand Oaks, CA.

Scott, W. Richard, Ruef, Martin, Mendel, Peter, Caronna, Carol, 2000. Institutional Change and Healthcare Organizations: From Professional Dominance to Managed Care. University of Chicago Press, Chicago.

Selznick, Philip, 1949. TVA and the Grass Roots. University of California Press, Berkeley, CA.

Thornton, Patricia H., 2001. Personal versus market logics of control: a historically contingent theory of the risk of acquisition. Organization Science 12, 294–311.

Thornton, Patricia H., Ocasio, William, 1999. Institutional logics and the historical contingency of power in organizations: executive succession in the higher education publishing industry, 1958–1990. American Journal of Sociology 105, 801–843.

Tolbert, Pamela S., Zucker, Lynne G., 1983. Institutional sources of change in the formal structure of organizations: the diffusion of civil service reform, 1880–1935. Administrative Science Quarterly 28, 22–39.

Warren, Ronald L., 1967. The interorganizational field as a focus for investigation. Administrative Science Quarterly 12, 396–419.

Weber, Max, 1922. 1978. Economy and Society. University of California Press, Berkeley.

Weber, Klaus, 2005. A toolkit for analyzing corporate cultural toolkits. Poetics 33, 227–252.

Wedlin, Linda, 2006. Ranking Business Schools: Forming Fields, Identities, and Boundaries in International Management Education. Edward Elger Publishing, Inc, Northampton, MA.

Zilber, Tamar, 2006. The work of the symbolic in institutional processes: translation of rational myths in Israeli high tech. Academy of Management Journal 49, 281–303.

Zucker, Lynne G., 1977. The role of institutionalization in cultural persistence. American Sociological Review 42, 726–743.

Relevant Websites

http://www.princeton.edu/csso – Center for the Study of Social Organization.

http://www.casbs.org – Center for Advanced Study in the Behavioral Sciences at Stanford University.

http://www.icos.umich.edu/lectures – Interdisciplinary Committee on Organizational Studies.

Organizational Justice

Russell Cropanzano, Leeds School of Business, Boulder, CO, USA
Agustin Molina, University of Valencia, Valencia, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Research on organizational justice explores the psychological mechanisms by which people render judgments of fairness, as well as their responses to these perceptions. When workers believe that they have been treated justly, they tend to show higher job performance, better work attitudes, and lower levels of stress. With these benefits in mind, this article reviews the literature on organizational justice, placing emphasis on the different types of fairness, as well as interactions among these types. Individual-level and group-level approaches to justice are also contrasted, with special attention to recent theoretical and empirical developments.

Within the organizational sciences, the term 'organizational justice' can be understood as having two meanings, one narrow and one general. The narrow meaning refers to particular judgments concerning various types of events. When deciding whether an occurrence is just or unjust, individuals (or groups) ascertain whether or not it is morally appropriate. More generally, the term 'organizational justice' can also refer to scholarly literature. That is, it refers to a body of knowledge about these fairness evaluations, including such topics as how they are formulated, their dimensional structure, and their individual- and group-level effects.

In general, workers respond more favorably – both for themselves and for their employers – when they perceive that they have been treated justly. They respond less well when they perceive an injustice. Research has found that justice perceptions are positively correlated with organizational commitment, job performance, leader–member relations, and helpful organizational citizenship behaviors (OCB). They are negatively correlated with work stress, counterproductive work behaviors, and turnover intentions (Cohen-Charash and Spector, 2001; Rupp et al., 2014).

Descriptive vs Normative Justice

Within the organizational sciences, justice is generally studied descriptively. That is, it focuses on the perceptions or evaluations of fairness that are formulated by people at work. Put more starkly, an event is treated as 'just' or 'unjust' because some person or persons believe it to be. This approach is subjective and allows disagreement as to what is and is not appropriate. The descriptive paradigm taken by organizational justice researchers can be distinguished from the more familiar normative perspective that is often employed by philosophers. Speaking very generally, ethical philosophers use logic and observation to determine what is 'prescriptively' or 'normatively' fair (Cugueró-Escofet and Fortin, 2014). That is, philosophers seek to understand what the objective events or attributes are that make something just (Cropanzano et al., 2011b).

In practice, organizational justice researchers tend to borrow ideas from ethical philosophy. For example, seminal thinking on procedural justice was taken from John Rawls, and work on distributive justice traces back (at least) to Aristotle. Regardless, even when borrowing philosophical concepts, organizational justice researchers study them descriptively. For example, they are unlikely to emphasize whether, say, a voice procedure is truly fair, but are more likely to investigate whether people evaluate it as fair. Organizational justice scholars also study the consequences of having made a fairness judgment.

Forming Perceptions of Justice

This descriptive perspective favored by organizational justice researchers places a great deal of emphasis on individual perceptions of fairness. There are a number of different theories that attempt to account for how justice evaluations are formulated. However, most of these conceptual models share certain common features (Cropanzano et al., 2011b), which are reviewed here.

In most cases, individuals make fairness judgments when an event, and often one that is personally disappointing, occurs. To understand whether or not the event was unfair, rather than simply unfortunate or surprising, individuals make use of a comparative referent, which acts as a sort of moral yardstick. These points of comparison go by a number of names, such as 'norms', 'standards', 'rules', or even 'justice criteria'. When events are negative and inconsistent with these referent points, then people are apt to decide that an injustice has occurred. In short, individuals compare an actual occurrence to a normative standard and then evaluate the discrepancies (Hollensbe et al., 2008). As described so far, the formation of justice evaluations may appear to be somewhat simple and logical. However, this omits a number of complicating factors.

First, before people begin to appraise the fairness of an event, it needs to be salient. If it does not catch sufficient attention, then individuals will be less likely to do the cognitive work involved in formulating a fairness judgment. This is one reason why negative events that are surprising and personally relevant are especially likely to be viewed as unfair.

Second, in real-world settings people do not always go through the careful steps outlined above. Often, employees

simply model the behavior of their coworkers, thereby adopting the views of their peers rather than working out their own personal assessments. This is an important finding, which as will be seen is relevant to the formation of group-level justice perceptions.

Third, people differ as to who is and who is not included within their *scope of justice*. When other people are within this scope, such as in-group members, evaluators are more solicitous and scrupulous about applying fairness rules. Unfortunately, when others are outside the scope of justice, such as out-group members, then evaluators becomes less circumspect and more disinterested in how potential victims are treated. Findings for scope of justice may go far in explaining how observers can be nonchalant about the mistreatment of others. Fortunately, everyone does not show this effect to the same degree. The scope of justice effect is stronger for those with weak moral identities. Conversely, for individuals who see morality as important to their identity, this effect is weaker (Reed and Aquino, 2003).

Fourth, while people seem to care about fairness, they show considerable egocentric biases in their judgment. These biases can be automatic and unconscious. For instance, when individuals are assigned a side in a legal case, either defendant or plaintiff, and then asked to read about the alleged events, they show biased recall of facts and more extreme evaluative judgments. However, when individuals are told to read about the alleged events and only later assigned a role in the case, they show more accurate recall of acts and less extreme judgments (Thompson and Loewenstein, 1992). This suggests that personal relevance affects the processing of factual information.

Dimensions of Justice

Thus far we have described the precipitating stimuli for justice judgments broadly, using some terms as 'events' or 'occurrences'. In fact, justice scholars have devoted considerable attention to the types of things that people evaluate. Broadly speaking, these can be broken into three classes of event – outcomes (distributive justice), decision-making processes (procedural justice), and interpersonal treatment (interactional justice). Each of these is reviewed each, introducing alternative formulations where these are relevant.

Distributive Justice

Distributive justice refers to the perceived fairness of an allocation or, more broadly, to how people judge what they receive. It was probably the first type of justice to gain the attention of organizational justice scholars and continues to receive wide attention. In general, there are three rules that people can use to decide whether or not their outcomes are fair – equity, equality, and need (Deutsch, 1985).

Equity allocations treat a fair allocation as one that pays individuals in accordance with their contributions. There are different models of equity – Aristotle even had one – but by far the most influential is Adams' (1965) *equity theory*. According to Adams, individuals calculate equity comparing two outcome/input ratios. The first ratio is their own outcomes divided by their own inputs, while the second is a referent's outcomes divided by the referent's inputs. In order for equity to exist, these two ratios should match. A counterintuitive prediction of equity theory is that people care about overreward, as well as underreward. If an individual's ratio is larger than that of a referent person, this means that the individual has greater outcomes given his or her inputs. In such a circumstance, Adams found that the individual will attempt to restore equity by either raising performance (if pay is fixed) or lowering performance (if pay is piece rate). In general, equity theory has received support, including predictions for overreward. But it is not the whole story.

An equality allocation provides everyone with the same amount, without regard to contributions. Equality allocations are most likely to be used when the goal is to maximize group harmony; equity allocations are most common when the goal is to maximize productivity. As such, equality is relatively common among close, communal groups. There is some evidence that equality, as opposed to equity, is also preferred when decision making is allocating something negative, such as budget cuts. If we must share suffering, it seems that individuals often prefer to suffer by the same amount or percentage. Interestingly, equality may also have an information processing advantage, as it is a very simple allocation rule that requires less effortful thought than others. In this regard, researchers have documented the existence of an *equality heuristic*, whereby people simply divide things evenly unless they have the cognitive resources to consider a more complex norm (such as equity).

A third allocation rule is need, which provides outcomes on the basis of a perceived deficit. Perhaps surprisingly, need allocations have been less widely studied in the organizational sciences, although even profit-making firms make at least some use of them. For instance, family leave policies are allocated to those who need them – only people with particular needs (e.g., child care or elder care) receive time off.

In practice, when people are asked to make allocation decision, they often mix the rules together, such that they are not always used in a pure form (Cugueró-Escofet and Rosanas, 2013). For example, one strategy might be to provide 'necessities' or a minimum income to all. This would be an equality allocation. However, above this equal minimum the equity rule could be employed. In this regard, one might use merit to allocate additional payments. Another mixed approach might be to allocate different goods in different ways. Among American employers, for example, socioemotional goods indicative of group standing are often allocated equally, whereas economic goods are often allocated through equity (Chen, 1995). These and other sorts of mixed approaches seek to balance the strengths and weaknesses of the different allocation rules.

Procedural Justice

Outcomes need to arrive from somewhere. Procedural justice refers to the decision-making process or the set of policies that are used to make allocation decisions. For example, trials are one way that disputes are adjudicated in many societies. Research suggests that, to the extent that people perceive control within the legal process (e.g., they have 'voice', or the

opportunity to present their case), they are more willing to accept an unfavorable decision (Tyler, 2006). Interestingly, there seems to be marginal 'fairness returns' on at least some procedural safeguards. If a valued process is absent, then even small gains in its amount will boost perceptions of justice. However, at a certain point, some amount of the process appears to be satisfactory and additional gains will produce fewer benefits.

Leventhal et al. (1980) provided a list of six well-known criteria that are used to decide whether a procedure is just. In general, the process should treat all parties consistently, be free from bias, use accurate information in rendering decisions, take into account the views of all (something like voice), be correctable in the event of an error, and remain consistent with prevailing ethical norms. These six criteria have remained influential, although some additional rules might also be important in certain settings. For example, in some settings, people prefer advance notice before something potentially negative occurs.

Research reviewed by Tyler (2006) suggests that procedural justice place an important role in promoting citizen compliance with the law. When people believe that the political system is essentially fair, such as when citizens are adequately represented by their elected representatives, they are more likely to accept legal rules, even when these might not be viewed as personally desirable.

Interactional Justice

Bies (2001) has argued that in addition to formal outcomes and procedures, individuals evaluate fairness via social or communication criteria. That is, they look to how they were treated by others. Broadly defined, these social criteria have been found to fall into at least two categories (although there may be more than two) – interpersonal justice and informational justice.

Interpersonal justice pertains to the dignity and respect with which people are treated. An interpersonally fair transaction would avoid personal attaches, refrain from unnecessary harshness, eschew bigotry, and so forth. Some scholars have argued that this interpersonal variety of interactional justice has an especially unambiguous relationship to judgments of immorality.

Informational justice, as its names applies, refers to the provision of relevant evidence and explanations, and this is especially important when things go wrong. In particular, research has found that organizations should provide social accounts, indicating to the employee why things happen as they do. Such accounts are important (and surprisingly effective) in mitigating feelings of unfairness (Shaw et al., 2003).

There is some debate over interactional justice's structure. Many scholars combine interpersonal and informational justice together. This results in a three-factor model for fairness – distributive, procedural, and interactional. Others separate out the interpersonal and informational components, producing a four-factor solution (Colquitt, 2001). Both approaches seem to work psychometrically. Choosing between the three- and four-factor models may partially have to do with the goals of the researcher and the underlying theory.

Justice Rules in Real-World Settings

In general, fairness researchers argue for three or four families of events – outcomes, processes, interpersonal transactions, and information – that are appraised in accordance with a set of normative standards. While these standards certainly cover a broad class of situations, it is not clear that actual employees apply them as neatly as some research might imply. Some research suggests that decision makers supplement the better researcher approaches to justice with so-called emergent rules (Hollensbe et al., 2008). For instance, individuals may notice high turnover and work backwards to assume that this was caused by injustice. Alternatively, individuals might assess the level of supportiveness and use this to infer that a supervisor is fair. These emergent rules are important and should be the subject of future research.

Interactions between and among the Types of Justice

So far we have only considered the individual main effects of the different types of justice. Research has long supported the idea that the various types of fairness interact. For instance, individuals dislike receiving an unfair, or even unfavorable, outcome. However, the ill effects of an unjust allocation can be partially addressed if the allocation process is viewed as fair. For example, people who lose their court cases are unhappy with the outcome, but they are less likely to derogate the legal system if they perceive that it treated them fairly. Similar effects have been found for workplace decisions (for a review, see Brockner, 2002).

In general, the outcome by process interaction suggests that ill effects of fairness are maximized when both types of justice are low (e.g., low distributive justice + low procedural justice). Either type of fairness can buffer the effects of the other. Thus, it could also be said that individuals prefer to be provided with, say, voice, which is a characteristic of fair procedures. However, the absence of voice will not be problematic unless an outcome is unfair or at least unfavorable.

Despite these supportive findings, there are other times when a fair process is unlikely to ally the anger from an unfavorable outcome. A good example of this comes from research on 'moral mandates' (Mullen and Skitka, 2006). A moral mandate refers to a strongly held belief or opinion, which is grounded in an ethical judgment. A moral mandate would exist for an issue like abortion, when the debate is framed as an issue of human rights (e.g., the right to choose vs the right to life). For those possessing strong moral mandates about an issue, an unfavorable outcome is more difficult to mitigate by means of the procedure. Rather, the process itself is likely to be derogated as a result of the moral mandate violation.

Interesting, the two-way interaction works the same for interactional justice as it does for procedural justice. That is, interactional justice has the same buffering effects on unfortunate outcomes, and this has been observed in both the laboratory and the field (Brockner et al., 2003). This raises the possibility that any of the three types of fairness could allay the negative impact of the other two. By extension, this suggests that there should be a three-way interaction, such that the effects of unfairness are disproportionately present when all three are low. There is some evidence for this. Studies have

found that this three-way interaction predicts workplace retaliation and organizational attractiveness. At the present time, research on the three-way interaction is limited but promising, as the effect has been replicated (Cropanzano et al., 2011). However, less work has examined the four-factor model of justice in this way, although presumably this could be done as well.

Overall Justice

A recent trend within justice research, called *overall justice*, suggests that individuals form global judgments of how they are treated. Instead of focusing on the different sets of justice perceptions as three (or four) pieces of a puzzle, this approach allows researchers and practitioners to treat justice as a holistic assessment. The overall justice approach does not reduce the importance of the specific types of fairness. On the contrary, evidence suggests that overall justice judgments can be made in a 'bottom-up' fashion, as the specific justice perceptions are cognitively combined into an overall opinion. In other words, the types of justice act as the antecedents of overall justice judgments (Ambrose and Schminke, 2009).

Multifoci Justice

In the research viewed thus far, justice judgments have been treated as if they were directed toward a class of events, such as outcomes or processes. However, it is also important to consider the entity, such as an organization or a supervisor, who is the target of these appraisals. In this way, multifoci justice theory adds to this body of literature by evoking the importance of the *foci* or sources of justice. In the workplace, employees generally engage in social exchange relationships with multiple parties (Lavelle et al., 2007). Based on the nature of those exchanges, employees can form different perceptions for each of the sources of justice with whom they interact (e.g., customers, supervisor, coworkers, and the organization as a whole). Multifoci theory then highlights the importance of holding others accountable for the just or unjust acts that individuals experience in the workplace.

In short, research on this area has argued that employees can distinguish among different sources (Rupp and Spencer, 2006) and meta-analytic research has supported this contention (Rupp et al., 2014). For example, Rupp and Cropanzano (2002) found that employees form supervisory – procedural and interactional – justice perceptions. These authors further observed that those supervisory-focused perceptions could be differentiated from organizational-focused perceptions – organizational procedural and interactional justice. Another interesting finding from multifoci theory is that the relationship between a justice perception and a variable of interest will be stronger when the source of justice matches that of the outcome (Lavelle et al., 2007). If an organization that values its employees' commitment as core strength, for example, was worried about the impact of a new set of policies that have been implemented, the best approach to examine the antecedents on organizational commitment would be to measure its employees' perceptions of organizational procedural justice.

Group-Level Justice

The discussion so far has focused on individual-level perceptions of justice. In the workplace, though, individuals are constantly submerged within collective social structures. These may take such forms as teams, branches, and overall organizations. Justice scholars are no strangers to these structures or to their subjacent social dynamics. Over the last two decades, organizational justice has extended the scope of interest to include the appraisal of justice events and occurrences at the group level. Even though there are several arguments to describe the emergence of group-level justice phenomena, social information processing (SIP) theory and the attraction–selection–attrition (ASA) model represent the two approaches most used.

Social Information Processing Theory

According to the SIP theory (Salancik and Pfeffer, 1978), individuals do not work in isolation from one another. On the contrary, employees generally engage in social interactions to make sense of the events that take place during their daily activities. As explained in a previous section, employees may form justice assessments by simply modeling the behavior of their coworkers (Hollensbe et al., 2008). Consequently, as time goes by, the justice perceptions of each member of the group are said to converge. In other words, if group members receive similar outcomes, are subjected to similar procedures, and are treated in a similar fashion by a source of fairness (e.g., supervisor), over time, a shared judgment of justice emerges. In direct support of the SIP theory, Roberson (2006) demonstrated that the more time spent in sense-making activities, the higher the convergence of group members' procedural and distributive perceptions.

Attraction–Selection–Attrition Model

The ASA model (Schneider et al., 1995) posits that individuals with similar characteristics are attracted to, selected into, and retained by the same group. The first step, the *attraction*, suggests that individuals find organizations differentially attractive based on the congruence of the organizations' goals and their own personalities. The next step, the *selection*, concerns the recruitment and hiring of individuals who possess the desired attributives looked for by the organization. The final step, the *attrition*, argues that the employees who do not fit the organization will leave their work, whereas employees who do fit the organization will remain in their jobs.

Colquitt et al. (2002) were among the first scholars to provide evidence for the ASA model. These scholars observed that a characteristic such as the demographic – age, ethnicity, and gender – diversity within a team predicted the level of convergence among team members' perceptions on procedural justice.

Justice Climate and Peer Justice

Thanks to the advancements on multifoci theory, research on group-level justice has been able to distinguish between two

sources of justice from which shared perceptions can emerge. These are justice climate and peer justice. The former refers to how the unit believes that it is treated by an authority figure, while the latter refers to how teammates behave toward one another.

Justice Climate

Naumann and Bennett (2000) were the first to use the term *justice climate*. They referred to this construct as the shared perception of how fairly a group is treated by an authority figure – such as the organization or the supervisor. In their study, Naumann and Bennet measured the group's shared perceptions on the procedures enacted by the organization (i.e., procedural justice climate). Their results revealed that procedural justice climate predicted helping behaviors above and beyond individual-level perceptions of procedural justice. That is, the perception of the group explained a part of the helping behavior performed by employees that could not be explained by its individual counterpart (individual-level procedural justice). Several studies have been conducted since then attesting the importance of justice climate as a predictor of workplace attitudes (e.g., satisfaction, commitment), processes (e.g., citizenship behaviors, cooperation), withdrawal behaviors (e.g., absenteeism, turnover intentions), and unit performance (e.g., customer satisfaction, financial performance; Whitman et al., 2012).

Peer Justice

This construct refers to the shared perception regarding how individuals judge the fairness with which they treat one another (Li and Cropanzano, 2009). Peer justice thus examines how people who do not have formal authority over each other behave, or believe that they behave, internally to the group. Despite the novelty of this construct, peer justice has already demonstrated a great potential for the organizational behavior domain. Cropanzano et al. (2011a), for instance, showed that peer justice predicts task and interpersonal teamwork processes which, in turn, predict performance and citizenship behaviors. Similar to justice climate, the effects of peer justice were above and beyond the effects of its individual-level counterpart.

Justice climate and peer justice provide two key instruments for industries and organizations that keep increasing the importance of team settings. By measuring both constructs, researchers and practitioners can assess the full spectrum of group-level justice phenomenology. Whereas justice climate reflects the treatment the group receives from an outside person or entity, peer justice provides a picture from inside the group, shedding light on the internal social dynamics.

Concluding Thoughts

Organizational justice is an area of study that investigates how and why people come to form judgments of fairness in regard to their work settings. Justice scholars also explore the consequential behavior and attitudes that result from these evaluations. In general, justice perceptions can be organized into at least three broad families – distributive, procedural, and interactional – and also examined holistically. Justice perceptions can be directed at specific targets (multifoci justice) or formulated at the individual or group level of analysis. In all these ways, justice has been shown to be an important aspect of work life.

See also: Emotions and Work; Happiness and Work; Organizational Citizenship Behavior; Organizational Commitment; Supervision, Abusive; Work Motivation; Workplace Deviance; Workplace Stress.

Bibliography

Adams, J.S., 1965. Inequity in social exchange. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 2. Academic Press, New York, pp. 267–299.

Ambrose, M.L., Schminke, M., 2009. The role of overall justice judgments in organizational justice research: a test of mediation. Journal of Applied Psychology 94 (2), 491–500.

Bies, R.J., 2001. Interactional (in)justice: the sacred and the profane. In: Greenberg, J., Cropanzano, R. (Eds.), Advances in Organizational Justice. Stanford University Press, Stanford, pp. 89–118.

Brockner, J., 2002. Making sense of procedural fairness: how high procedural fairness can reduce or heighten the influence of outcome favorability. Academy of Management Review 27 (1), 58–76.

Brockner, J., Heuer, L.B., Magner, N., Folger, R., Umphress, E., Van den Bos, K., Vermunt, R., Magner, M., Siegel, P.A., 2003. High procedural fairness heightens the effect of outcome favorability on self-evaluations: an attributional analysis. Organizational Behavior and Human Decision Processes 91 (1), 51–68.

Chen, C.C., 1995. New trends in rewards allocation preferences: a Sino-U.S. comparison. Academy of Management Journal 56 (2), 408–428.

Cohen-Charash, Y., Spector, P.E., 2001. The role of justice in organizations: a meta-analysis. Organizational Behavior and Human Decision Processes 86 (2), 278–321.

Colquitt, J.A., 2001. On the dimensionality of organizational justice: a construct validation of a measure. Journal of Applied Psychology 86 (3), 386–400.

Colquitt, J.A., Noe, R.A., Jackson, C.L., 2002. Justice in teams: antecedents and consequences of procedural justice climate. Personnel Psychology 55 (1), 83–109.

Cropanzano, R., Li, A., Benson III, L., 2011a. Peer justice and teamwork processes. Group and Organization Management 36 (5), 567–576.

Cropanzano, R., Stein, J.H., Nadisic, T., 2011b. Social Justice and the Experience of Emotion. Routledge, New York.

Cugueró-Escofet, N., Fortin, M., 2014. One justice or two? A model of reconciliation of normative justice theories and empirical research on organizational justice. Journal of Business Ethics 124, 435–451.

Cugueró-Escofet, N., Rosanas, J.M., 2013. The justice design and use of management control systems as requirements for goal congruence. Management Accounting Research 24 (1), 23–40.

Deutsch, M., 1985. Distributive Justice. Yale University Press, New Haven.

Hollensbe, E.C., Khazanchi, S., Masterson, S.S., 2008. How do I assess if my supervisor and organization are fair? Identifying the rules underlying the entity-based justice perceptions. Academy of Management Journal 51 (6), 1099–1116.

Lavelle, J., Rupp, D.E., Brockner, J., 2007. Taking a multifoci approach to the study of justice, social exchange, and citizenship behavior: the target similarity model. Journal of Management 33 (6), 841–866.

Leventhal, G.S., Karuza, J., Fry, W.R., 1980. Beyond fairness: a theory of allocation preferences. In: Mikula, G. (Ed.), Justice and Social Interaction. Springer-Verlag, New York, pp. 167–218.

Li, A., Cropanzano, R., 2009. Fairness at the group level: interunit and intraunit justice climate. Journal of Management 35 (3), 564–599.

Mullen, E., Skitka, L.J., 2006. Exploring the psychological underpinnings of the moral mandate effect: motivated reasoning, identification, or affect? Journal of Personality and Social Psychology 90 (4), 629–643.

Naumann, S.E., Bennett, N., 2000. A case for procedural justice climate: development and test of a multilevel model. Academy of Management Journal 43 (5), 881–889.

Reed II, A., Aquino, K., 2003. Moral identity and the expanding circle of moral regard toward out-groups. Journal of Personality and Social Psychology 84 (6), 1270–1286.

Roberson, Q.M., 2006. Justice in teams: the activation and role of sensemaking in the emergence of justice climates. Organizational Behavior and Human Decision Processes 100 (2), 177–192.

Rupp, D.E., Cropanzano, C., 2002. The mediating effects of social exchange relationships in predicting workplace outcomes from multifoci organizational justice. Organizational Behavior and Human Decision Processes 89 (1), 925–946.

Rupp, D.E., Shao, R., Jones, K.S., Liao, H., 2014. The utility of a multifoci approach to the study of organizational justice: a metaanalytic investigation into the consideration of normative rules, moral accountability, bandwidth-fidelity, and social exchange. Organizational Behavior and Human Decision Processes 123, 159–185.

Rupp, D.E., Spencer, S., 2006. When customers lash out: the effect of customer interactional injustice on emotional labor and the mediating role of discrete emotions. Journal of Applied Psychology 91 (4), 971–978.

Salancik, G.R., Pfeffer, J., 1978. A social information processing approach to job attitudes and task design. Administrative Science Quarterly 23 (2), 224–253.

Schneider, B., Goldstein, H., Smith, D.B., 1995. The ASA framework: an update. Personnel Psychology 48, 747–773.

Shaw, J.C., Wild, E., Colquitt, J.A., 2003. To justify or excuse?: a meta-analytic review of the effects of explanations. Journal of Applied Psychology 88 (2), 444–458.

Thompson, L., Loewenstein, G., 1992. Egocentric interpretations of fairness and interpersonal conflict. Organizational Behavior and Human Decision Processes 51 (2), 176–197.

Tyler, T.R., 2006. Why People Obey the Law. Princeton University Press, Princeton.

Whitman, D.S., Caleo, S., Carpenter, N.C., Horner, M.T., Bernerth, J.B., 2012. Fairness at the collective level: a meta-analytic examination of the consequences and boundary conditions of organizational justice climate. Journal of Applied Psychology 97 (4), 776–791.

Organizational Path Dependence

Jörg Sydow and Georg Schreyögg, Freie Universität Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Organizations are increasingly required to change rapidly and radically, i.e., to innovate or adapt to new circumstances in the technical and/or institutional environment. However, sometimes such organizational changes turn out to be extremely difficult, if not impossible. The theory of organizational path dependence features prominently in explaining this somewhat puzzling inertia. The antecedents, mechanisms, and outcomes of path dependence for organizations and interorganizational arrangements like alliances and networks are outlined.

Introduction

Why is organizational change so difficult, sometimes even impossible? We may almost expect large and bureaucratic organizations to be difficult to change, even in face of significant pressure from the technical and/or institutional environment (Amburgey et al., 1993). But this is much less obvious regarding smaller organizations or allegedly more flexible interorganizational arrangements like strategic alliances or regional networks. Nevertheless, under certain circumstances these may also become structurally inert or institutionally persistent. This is demonstrated by an increasing number of empirical studies (e.g., Maurer and Ebers, 2006; Manning and Sydow, 2011). The theory of organizational path dependence has been developed to explain these puzzling persistencies by highlighting the antecedents and mechanisms driving organizations and even networks into hyperstability.

The Theory of Organizational Path Dependence

The theory of organizational path dependence builds on earlier efforts in economics (David, 1985; Arthur, 1994) and political science (Mahoney, 2000; Pierson, 2000), and sets out to clarify the processes that lead institutions into a lock-in. Up until now, the explanatory power of the theory of organizational path dependence has been demonstrated in several empirical studies (e.g., Koch, 2011; Schreyögg et al., 2011; Burger and Sydow, 2014).

The basic point of any path-dependence argument lays stress on the importance of past events for current and future actions. This very basic insight has certainly advanced the building of a time-sensitive understanding of organizational phenomena, as well as helping to overcome the ahistorical, unbounded perspective of rational choice thought. Highlighting the fact that 'history matters,' however, is too vague a theoretical framework. In an attempt to develop a more specific understanding of path dependence, the drivers of the process have to be unearthed. The focus is on dynamics triggered by a critical event (or a sequence of events) that narrows the scope of action, eventually resulting in a state of persistence or lock-in. An important point of nearly all path-dependence thought is that this development need not be conceived as a negative spiral in terms of stabilizing an early bad decision or action pattern (as it is, for example, in the case of escalating commitment; cf Staw, 1984). In contrast, most path-dependence processes start quite promisingly; it is only later that the rationale changes and the development becomes disturbed by a dysfunctional flip. Path dependence, therefore, is inherently ambivalent in character.

With regard to antecedents, mechanisms, and outcomes of path-dependence processes in and among organizations, we have identified three different phases governed by three different regimes (Sydow et al., 2009). *Phase I* – the Preformation Phase – can be characterized by a largely unrestricted scope of action, where the choices taken cannot be predicted by prior events or initial conditions (Mahoney, 2000: 511). Up to a certain degree, however, the first phase is also influenced by the past: all choices are historically framed and therefore, in a way, 'imprinted.' Hence, we should refrain from conceiving this phase as a completely unrestricted choice in terms of rational choice models (as do David, 1985; Arthur, 1994; in their seminal works). The shadow in Phase I in **Figure 1** is intended to indicate this idea. Among the decisions made – on whatever basis – one choice may turn out to be a 'small event' that triggers further developments. From a path dependence point of view, this rather innocent or even random decision gains importance if, and only if it sets self-reinforcing processes into motion. At this point, nobody can predict that this initial action will end up in path dependence; this is only possible at a later stage of the process. At this point in time choices are still reversible, as well.

In *Phase II* – we suggest calling it the Formation Phase – the range of options narrows increasingly and it becomes progressively difficult to reverse the initial choice. In this – at least at the beginning, still quite open and possibly contested – phase of path-building, a new regime takes the lead: the dynamics of self-reinforcing processes. Arthur (1994) elaborates on these driving forces, conceiving them from an economic perspective as increasing returns (thereby excluding constant or diminishing returns as explanatory forces). On a general level, the notion of increasing returns highlights positive-feedback processes in which the increase of a particular variable leads to a further increase of this very variable. The notions of increasing returns or positive feedback indicate self-reinforcing processes with growing benefits when using a specific pattern of action or routine. There is a dynamic of doing more of the same. Eventually, a dominant organizational

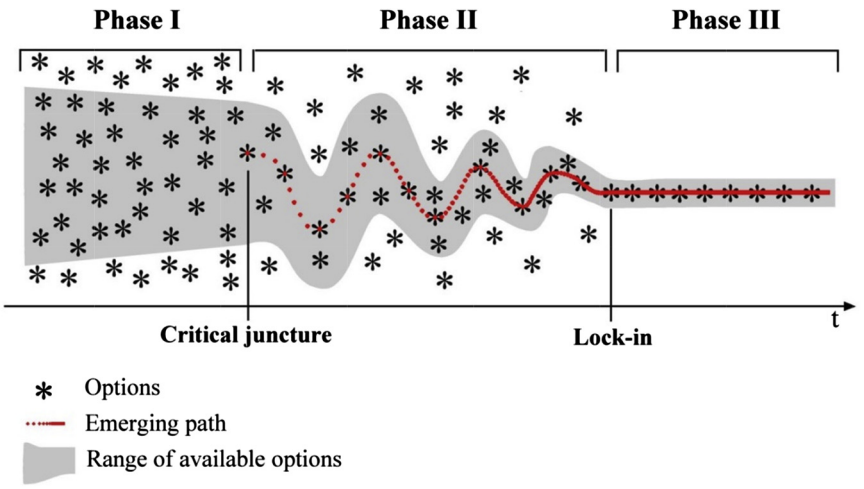

Figure 1 The constitution of an organizational path. Sydow, J., Schreyögg, G., Koch, J., 2009. Organizational path dependence: opening the black box. Academy of Management Review 34 (4), 692.

solution emerges. The other side of the coin is that the whole process is becoming more and more irreversible: all alternatives are excluded more and more cognitively, and also materially in terms of investments. Decision processes or action patterns in Phase II are still contingent, however; i.e., while essentially constrained, choices are still possible. As David (1985) put it, the processes are 'nonergodic,' i.e., they are not accidental, but do not yet fully converge to a fixed point of distribution.

In the economic and institutional studies mentioned earlier, several types of self-reinforcing dynamics have been identified to date. Among them are (1) coordination effects, (2) complementarities, (3) network externalities, (4) learning effects, and (5) adaptive expectations. As empirical studies show, they are relevant in the organizational context as well (e.g., Schreyögg et al., 2011; Burger and Sydow, 2014). However, they may be of different importance there and act in a slightly modified manner.

1. When explaining organizational path dependence, the *coordination* effect features most prominently. Building on the benefits of rule-guided behavior, this effect relates to the heart of organizational theory: the more actors adopt and apply a specific institution (a law, a norm, a routine, etc.), the more efficient the interaction is among these actors, as their behavior is guided by a rule and can therefore be anticipated, and so on. In consequence, the more returns there are from sharing a rule, the more attractive (beneficial) it becomes for other people to adopt and so follow this very same institution.
2. Similarly important when explaining organizational path dependence are *complementarity* effects (see also Dobusch and Schüßler, 2013). These result from synergies from the interaction of two or more separate but interrelated resources, rules, or practices. In the case of complementarities, the advantages of repeatedly combining interrelated activities do not simply add up, they produce an additional surplus. In complementary settings, self-reinforcing processes occur when routines and/or practices are interconnected in such a way that it becomes more and more attractive to exploit the synergies or – when referring to the reverse aspect – to save misfit costs caused by solutions deviating from the established organizational capability. As a result, distinct sets of activity patterns become progressively dominant and deeply embedded in an organization ('deep structure'), i.e., they become organizationally path dependent.
3. Self-reinforcing effects are also brought about by *direct and indirect network externalities*. A user's benefits from a purchased good or service increase overproportionally, the more other actors use the same good or service too. For example, the purchase of a telephone was all the more rewarding the more others decided to use this device as well. Or in the case of video technology, replicating the initial choice for a system (e.g., VHS) was more rewarding the more people were already using this technology (Cusumano et al., 1992). This effect becomes even more rewarding and dynamic if indirect externalities exist too, i.e., the demand becomes even more focused because complementary products and services accompany the focal product – as was, for instance, the case with video stores, video recorders, and so on, in the case of VHS. These returns eventually led to a dominant solution irrespective of the fact that the first users had chosen that solution accidentally and a technically more advanced solution was available.

   This effect works in an intra- or interorganizational environment as well. Take, for instance, knowledge management systems. They work better the more employees (possibly also including participants from other organizations) subscribe to them, and it is more attractive to join the network the more employees use it already and the better it fits together with other information systems (e-mail, portals, etc.).
4. The *learning* effect holds that the more often an operation is performed, the more efficiency will be gained when operating subsequent iterations (Argote, 1999). The operations are performed more skillfully (faster, more reliable, and with less errors), which, in turn, means a decrease in average costs per unit of output. And the more attractive the chosen

solution becomes due to accumulated skills and decreasing costs, the less attractive it will be to switch to new learning sites (where the actors would have to start from scratch). Only sticking to the once chosen solution promises continued returns – although, as is well known, the resulting cost curve flattens after a while.

Self-reinforcing learning effects can be found on various organizational levels. One well-known example from organization-level learning indicates that a focus on the advantages of exploitative learning may increasingly drive out explorative learning (March, 1991). For various reasons (e.g., prevailing organizational culture and reward system), the motivation to improve everyday practices is likely to gain further acceptance or legitimacy (and thus more rewards) from the organization, whereas the motivation to look for fresh alternatives and to examine well-established organizational practices critically will shrink progressively.

5. The self-reinforcing effect of *adaptive expectations* originally relates to the interaction of demand and preference: i.e., preferences are not considered to be individually fixed, but rather vary along with expectations of what others do. Often-quoted examples highlight the need for social belonging and the desire to end up on the side of the winners. The first is best known as the bandwagon effect, stating that people trust in the wisdom of the majority. The more people purchase a particular good (and not another one), the more attractive it becomes for new buyers (Leibenstein, 1950). Users are uncertain about the right choice and feel rewarded by the fact that others have bought or will buy the same good or service – and the more the better. Because of these increasing returns, a dominant solution is likely to emerge (in most cases on the basis of more or less accidental first choices). The diffusion of a best practice in organizations often follows the same logic; organizational members are willing to adopt a practice because they expect others to do the same – a tendency which is also driven by legitimacy seeking. Subsystems not subscribing to the best practice run the danger of losing legitimacy.

A similar expectation effect stems from the desire to avoid getting into a failing position (e.g., Haller and Norpoth, 1994). Prospective users choose those options which they expect to win (in internal resource allocations, for instance). The more actors exchange and share this expectation, the more others will subscribe to the solution which is generally assumed to be the winning one. The more individuals adopt this expectation and base their choices on it, the more probable it is that the favored solution will gain the winning position – and the more rewarding it will be to join the winning solution. In other words, the emerging dominant solution stems from a self-fulfilling prophecy and its increasing returns for those who always want to join the winner's side (Pierson, 2000). In organizations, this effect can often be observed in political processes and coalition building. Once a suggestion by a group turns out to be successful, other organizational members are likely to join this group in order to avoid getting into a failing position.

As already indicated, most of these effects have originally been modeled in a restricted economic way, i.e., they develop only if the decision to reproduce a particular option is suggested by a utility maximizing calculus by an individual in terms of increasing returns. Without excluding this case, it seems too restrictive a starting point for an explanation of organizational paths. As is well known from a long-standing tradition in organizational research, there are also other self-reinforcing patterns in organizations, for instance, based on emotional reactions, cognitive biases, or political processes. These dimensions have to be included when setting out to understand the specific dynamics of organizational paths. Our suggestion, therefore, is to broaden the scope of self-reinforcing effects by including all kinds of positive self-reinforcing feedback cycles. In addition, a broader perspective beyond individual decision-making is needed to account for the social setting in which the positive feedback processes are embedded. It is the broader organizational context that informs actors and presses for self-reinforcing processes (see also Koch, 2011).

Finally, it should be stressed that self-reinforcing effects are not fully separate; rather, they often occur jointly and overlap (see also Dobusch and Schüßler, 2013). Take, for instance, information and communication technology, where the value of a specific terminal and the whole network depends to a large extent upon the number of users subscribed to this system. These direct network externalities are complemented by indirect network effects when the system is compatible with other already available technologies (printers, software, etc.), which are likely to generate synergies (complementarities). These effects can, in turn, be reinforced through increasing returns resulting from learning in the form of economies of scale: the increasing number of system users allows for a higher volume and, hence, declining costs per unit. These savings may be forwarded (partially) to the buyer – via lower prices – which will increase the attractiveness of the system even further. Adaptive expectations can speed up the spiraling process: prospective buyers observe the trend toward a particular solution and are therefore likely to choose the same system, thereby encouraging other potential buyers.

The transition from this second phase to *Phase III* – the Lock-in Phase – can be characterized by a further constriction which is likely to (but will not necessarily) lead into a lock-in, i.e., the dominant pattern gets fixed and even gains a quasi-deterministic character. Further decisions or actions are bound to the organizational path that has emerged; even new entrants have to adopt it. A lock-in means that the path has led the organization into a situation of irreversibility; actually, alternative organizational solutions cannot be realized. A lock-in is not necessarily inefficient in the beginning, but it is always problematic from a strategic perspective that requires organizational flexibility in increasingly 'turbulent fields' (Emery and Trist, 1965). However, the problem of a lock-in, even from a short-term perspective, becomes obvious in cases where a more efficient alternative appears but a switch is no longer possible. The result is actual or potential inefficiency. Therefore a major element of the third phase is (potential) inefficiency.

Considering specifically organizational paths (particularly in contrast to technological paths in a market), it may however be inadequate to conceive the third phase as a situation of total rigidity (or monopoly) in terms of one solution; a somewhat modified conception of lock-in seems to be required here.

Due to their social character, organizational processes are more complex and ambiguous in nature than market standardization (Cyert and March, 1963). They are not likely to amount to a concrete monopolistic solution that excludes any further choices. Rather, they are prone to bring about a specific action pattern that gets deeply embedded in practice and replicated across various situations. Fixed recursive patterns of action, however, cannot be equated with determinism, as practices have to be reproduced, and all reproduction includes interpretation and at least some minor deviation (Giddens, 1984). For these reasons, we should, in particular with regard to organizations, refrain from reifying paths and attributing an objective quality to them. Instead of a fully determined lock-in, no matter whether cognitive, normative, or resource-based, a theory of organizational paths is well advised to conceptualize the final stage of a path-dependent process in a less restrictive way, leaving some scope for variation (similarly: Pierson, 2000; Martin and Sunley, 2006). By implication, it seems more adequate to conceive of the lock-in state not in terms of total rigidity, but rather as a matter of degree accounting for some variance in the actual practicing of the path. A corridor may serve best to illustrate this reasoning; the shadow in Phase III in **Figure 1** is designed to indicate this corridor.

In sum, the proposed theory – as shown in **Figure 1** – conceptualizes an organizational path as a historically imprinted social process that is initially, in Phase I, shaped by a contingent choice, possibly leading to a critical juncture triggered by one or several events. Phase II, the Path Formation Phase, is governed by a regime of positive, self-reinforcing feedback loops, constituting and setting up a specific pattern of actions (social practices), which gains progressive predominance against alternative solutions to such an extent that, at least potentially, it leads in Phase III into a lock-in, in terms of a pattern underlying actions and/or practices. Thus all three stages pictured in **Figure 1** are governed by different regimes allowing us to differentiate between the phases not only in the theoretical model but also in the empirical data. Though this model focuses on the organizational level of analysis, it may also be helpful to analyze path dependence on the lower levels of organizational subsystems or higher levels such as interorganizational networks or fields, and to explore the interrelationships among various paths (cf Sydow et al., 2012).

Organizational Path Dependence and Related Concepts

The theory of organizational path dependence only provides explanations for very specific kinds of institutional persistence and structural inertia. Therefore, it may be helpful to distinguish this particular explanation from institutionalization and inertia. Because of the important role of imprinting, in particular in Phase I of a path-dependent process, it is useful to elaborate on its distinctive features as well.

The concept of *imprinting* figures prominently among approaches that seem to address a process very similar to organizational path dependence (Johnson, 2007). Basically, this concept postulates that either the initial cognitive schemes or the competences of an actor, a founding entrepreneur or a team, for instance, or specific contextual conditions (organizational structure, financial crisis, etc.) at the time of founding imprint upon organizational processes at later stages and, eventually, amount to a replicated pattern. While there are doubtless striking similarities that lead one to either simply equate imprinting and path dependence or to conceive the former as a specific variant of the latter, the process of becoming path dependent is, however, governed by a different logic. First, the replicated pattern in the imprinting approach is ready-made at the beginning; it is a specific scheme that persists and continues to influence future processes. In contrast, the gestalt of an organizational path is not clear at all in the early stage; it is an unforeseeable product of later processes. Second, a theory of organizational paths – as opposed to the imprinting approach – has to explain the unfolding of the process of path formation, not only the reproduction of structural properties because of either efficiency or a lack of competition, or the presence of institutionalization processes (Johnson, 2007). Nevertheless, imprints doubtlessly play an important role in many organizational processes. In path-dependent organizational processes, they can, for instance, explain the restrictions in the Preformation Phase.

Another concept that shares striking similarities with organizational path dependence is *structural inertia*, which results mainly from the increased routinization of organizational processes over time (Hannan and Freeman, 1984). Like organizational path dependence, structural inertia implies the hyperstability of organizational arrangements in spite of environmental change. Routinizing of organizational activities is seen as imperative in order to guarantee efficiency, reliability, accountability, and, finally, survival in competitive environments. Inertia is, however, like organizational path dependence, a double-edged sword: it is considered as a precondition to organizational efficiency and legitimacy but it may at the same time threaten the organization's survival when it leads to a misfit with changing environmental circumstances. However, in sharp contrast to organizational path dependence, inertia occurs via the intended establishment of reliable organizational structures; there are no structural dynamics. It is a universal requirement that all organizations have to fulfill. And all organizations, especially when growing and aging, are also expected to become hyperstable and face difficulties in meeting new environmental challenges. Opposed to that, the suggested framework of path dependence does not apply to all organizations, highlighting only special cases.

Contextual shaping forces play a major role in *institutionalization*, which also seems to come close to path dependence (e.g., Scott, 2008). Not unlike structural inertia, the institutionalization of (inter-) organizational structures and processes is demanded by the environment of the system, in particular the symbolic-normative environment. In the institutionalization process, specific organizational elements become sedimented and taken for granted over time. Despite similarities, institutionalization is a much broader concept and does not address the systemic logic of the escalating reinforcement of an action pattern leading into an organizational path (see Berthod and Sydow, 2013; for a further elaboration of the similarities and differences).

Implications for Organizational Research and Managerial Practice

This article has suggested a theory of organizational paths that builds, on the one hand, upon the seminal works of David and Arthur in order to transcend a purely metaphorical use of the notion of path dependence. On the other hand, this theory goes beyond these works, first by specifying the regimes under which the process of becoming path dependent unfolds in and among organizations, and second by modifying important assumptions in the light of organizational research. While an uncountable number of studies have investigated the persistence of institutions and/or the rigidity of organizations and even interorganizational networks (e.g., Amburgey et al. 1993; Maurer and Ebers, 2006), only relatively few have so far adopted an analytical approach based upon (strict concepts from) the theory of path dependence as presented above (e.g., Koch, 2011; Schreyögg et al., 2011; Manning and Sydow, 2011; Burger and Sydow, 2014).

The implications for organizational research are straightforward: more processual and detailed studies, possibly even using ethnographic methods (Yanow, 2009), are needed to understand better how events trigger self-reinforcing processes that – at least potentially – lead organizations or even interorganizational networks into a lock-in; a situation that is always problematic, at least strategically. Such studies should also be used to develop the theory of organizational path dependence into a more general approach to path constitution that informs us about possibilities of path breaking, path forking, and even path creation (Sydow et al., 2012).

While management may be particularly interested in these last alternatives, we suggest – given the present state of theorizing – a more modest approach and a focus on monitoring and understanding processes of organizational path dependence, along with their more often than not hidden dynamics within and among organizations (Schreyögg and Sydow, 2010). Although it is tempting to wonder about possibilities of path breaking or even path creation, the present state of theorizing provides firmer ground for understanding organizational path dependence.

See also: Boundaries and New Organization Forms; Design and Form: Organizational; Development: Organizational; Ecology: Organizations; Modularity and Organizations; Narrative Approaches to Organizations; Network Paradigm: Applications in Organizational Science; Organization: Overview; Organizations, Sociology of.

Bibliography

Amburgey, T.L., Kelly, D., Barnett, W.P., 1993. Resetting the clock: the dynamics of organizational change and failure. Administrative Science Quarterly 38 (1), 51–73.

Argote, L., 1999. Organizational Learning: Creating, Retaining and Transferring Knowledge. Kluwer Academic, Boston.

Arthur, W.B., 1994. Increasing Returns and Path Dependency in the Economy. University of Michigan Press, Ann Arbor.

Berthod, O., Sydow, J., 2013. Locked in the iron cage? when institutionalization is (not) a path-dependent process. In: Sydow, J., Schreyögg, G. (Eds.), Self-reinforcing Processes in and Among Organizations. Palgrave Macmillan, London, pp. 204–229.

Burger, M., Sydow, J., 2014. How inter-organizational networks can become path-dependent: bargaining practices in the photonics industry. Schmalenbach Business Review 66 (1), 73–99.

Cusumano, M.A., Mylonadis, Y., Rosenbloom, R.S., 1992. Strategic maneuvering and mass-market dynamics: the triumph of VHS over Beta. Business History Review 66, 51–94.

Cyert, R.M., March, J.G., 1963. A Behavioral Theory of the Firm. Prentice-Hall, Englewood Cliffs, NJ.

David, P.A., 1985. Clio and the economics of QWERTY. American Economic Review 75 (2), 332–337.

Dobusch, L., Schüßler, E., 2013. Theorizing path dependence: a review of positive feedback mechanisms in prominent cases. Industrial and Corporate Change 22 (3), 617–647.

Emery, F.E., Trist, E.L., 1965. The causal texture of organizational environments. Human Relations 18, 21–32.

Giddens, A., 1984. The Constitution of Society. Polity Press, Cambridge.

Haller, B.H., Norpoth, H., 1994. Let the good times roll: the economic expectations of US voters. American Journal of Political Science 38, 625–650.

Hannan, M.T., Freeman, J., 1984. Structural inertia and organizational change. American Sociological Review 49, 149–164.

Johnson, V., 2007. What is organizational imprinting? Cultural entrepreneurship in the founding of the Paris Opera. American Journal of Sociology 113 (1), 97–127.

Koch, J., 2011. Inscribed strategies: exploring the organizational nature of strategic lock-ins. Organization Studies 32 (3), 337–363.

Leibenstein, H., 1950. Bandwagon, snob, and Veblen effects in the theory of consumer's demand. Quarterly Journal of Economics 64, 183–207.

Mahoney, J., 2000. Path dependence in historical sociology. Theory and Society 29, 507–548.

Manning, S., Sydow, J., 2011. Projects, paths, practices: sustaining and leveraging project-based relationships. Industrial & Corporate Change 20 (5), 1369–1402.

March, J.G., 1991. Exploration and exploitation in organizational learning. Organization Science 2, 71–87.

Martin, R., Sunley, P., 2006. Path dependence and regional economic evolution. Journal of Economic Geography 6, 395–437.

Maurer, I., Ebers, M., 2006. Dynamics of social capital and their performance implications: lessons from biotechnology start-ups. Administrative Science Quarterly 51, 262–292.

Pierson, P., 2000. Increasing returns, path dependence, and the study of politics. American Political Science Review 94, 251–267.

Schreyögg, G., Sydow, J. (Eds.), 2010. The Hidden Dynamics of Path Dependence. Palgrave-Macmillan, London.

Schreyögg, G., Sydow, J., Holtmann, P., 2011. How history matters in organizations – the case of path dependence. Management & Organizational History 6 (1), 81–100.

Scott, W.R., 2008. Institutions and Organizations, third ed. Sage, Thousand Oaks.

Staw, B.M., 1984. The escalation of commitment to a course of action. Academy of Management Review 6 (4), 577–587.

Sydow, J., Schreyögg, G., Koch, J., 2009. Organizational path dependence: opening the black box. Academy of Management Review 34 (4), 689–709.

Sydow, J., Windeler, A., Müller-Seitz, G., Lange, K., 2012. Path constitution analysis – a methodology for understanding path dependence and path creation. Business Research 5 (2), 155–176.

Yanow, D., 2009. Organizational ethnography and methodological angst: myths and challenges in the field. Qualitative Research in Organizations and Management 4 (2), 186–199.

Relevant Websites

http://economics.stanford.edu/node/105 – Stanford University Department of Economics.

http://www.wiwiss.fu-berlin.de/en/forschung/pfadkolleg/index.html – Freie Universitat Berlin.

Organizations and Culture

Frank Dobbin and Shawna Vican, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cultural approaches to understanding the modern organization date back to the Human Relations studies of the 1920s. To understand the phenomenology of the modern firm, cultural theorists have explored the roles of power and meaning in the social construction of rational, and just, organizational practices. Neoinstitutionalists have recently turned their attention to how organizational practices travel across time and space, the role of agency in the construction and sustenance of the cultural practices of the firm, and the microfoundations of meaning in organizational practice.

In organizational sociology, the tradition of examining the role of informal, cultural patterns dates to the 1930s and the Hawthorne plant studies that spawned the Human Relations School (Mayo, 1933; Roethlisberger and Dickson, 1939). As Barley and Kunda (1992) argue, there have been waves of rationalist and culturalist theories of management, and of organizations, since then. The social constructionist approach to culture that has come to dominate the field has its roots in single-organization studies exploring culture, and in psychological theories of cognition. First, the single-organization studies were pioneered by institutionalist Philip Selznick and his students (Selznick, 1957; Zald and Denton, 1963; Clark, 1960). They explored how members of an organization become irrationally committed to the organization's routines and practices. Their method was to chart natural histories of individual organizations, often nonprofits. They found that members develop a collective understanding of the utility, efficacy, and rationality of organizational practices, and overestimate the importance of particular practices to the organization's success. Experiencing an organization's routines, and playing a role in the replication of those routines, produces this commitment. Culture is at the heart of the process.

Second, James March and Herbert Simon (1958) pioneered the cognitive approach to organizational practice in *Cognitive Limits on Rationality*, drawing conclusions similar to those of the institutionalists, but from a base in psychology rather than from historical studies of individual organizations. They observe that managers seldom consider all available solutions to a particular problem, and describe how consciousness is shaped by life within an organization – how an organizational worldview emerges. When faced with a problem, people typically think of an analogous problem from the past and apply the solution used in that case. The menu of solutions that comes to mind, then, is a consequence of organizational experience. The culture of a firm is embedded in its routines and practices, and the collective understandings that emerge to support those routines and practices. Organizational culture theorists would later describe the same phenomena in the terms of cultural anthropology (Schein, 1996; Martin, 2002).

In organizational psychology, Karl Weick's theory of sensemaking suggests that there are shared rationales of action that we pick up through experience, and that we choose from among these cultural constructs to understand behavior in organizations. For Weick (1993, 1995), the interpretive frames offered by society, and by the focal organization, shape how the individual understands her own actions. People may make sense of their own actions retrospectively, describing their behavior as part of a plan (Weick, 1979: p. 194). Sensemaking is like cartography in Weick's view, in that there is no one best map of the world but many useful maps that people can use to make sense of their actions. In sensemaking, people assemble causal maps from bits and pieces of practice and theory: "What the world is made of is itself a question which must be answered in terms of the available conceptual resources of science at a particular time" (Fay, 1990: p. 36). Cognitive maps are derived from experience and institutions. These maps are shared. They are cultural.

The Phenomenology of the Firm

In suggesting that organizational theory undergoes swings from cultural to rational theories, Barley and Kunda (1992) challenge the view that knowledge about management is cumulative. They also challenge the prevailing view of the firm *c*.1970, which was functionalist. The dominant thinking was that economic laws determined 'best practices' and that those 'best practices' would come to the surface everywhere. If organizations look alike, it is because they are subject to the same laws of efficiency. In *Institutionalized Organizations: Formal Structure as Myth and Ceremony*, the seminal article in the new institutional tradition, John Meyer and Brian Rowan (1977) attacked this view, proposing a constructionist approach to understanding organizational practices and routines. Meyer and Rowan describe modern organizations as adopting structures that symbolize rationality and fairness, and describe those structures as 'myth and ceremony.'

Organizations adopt practices that embody myths of rationality to both trumpet their commitment to efficiency and achieve it. We socially construct reality generally by classifying behavior patterns, norms, and rules and linking them to myths of rationality. Meaningful social conventions thus shape behavior in organizations: "Institutionalized rules are classifications built into society as reciprocated typifications or interpretations" (Meyer and Rowan, 1977: p. 341). Firms come to look alike because they follow the same rationalized myths of organizing rather than because each divines the same laws of efficiency. New myths of organizational rationality spread like

fads, causing organizations to adopt, and reject, matrix management and quality circles en masse.

The intellectual foundations of this approach are found in phenomenology and symbolic interactionism. Selznick, March and Simon, and Weick suggested that experience with practices in organizations shapes cognition and understanding, but the phenomenologists and symbolic interactionists focused more on the collective construction of cultural meaning in society generally, and less on the cognitive responses of individuals. In so doing they prepared the ground for a truly cultural theory of the organization.

Symbolic interactionist Erving Goffman (1974) built on Durkheim's (1961[1912]) ideas about classification and meaning in premodern societies, found in *The Elementary Forms of the Religious Life*, to argue that in the modern world, the individual "is allotted a kind of sacredness that is displayed and confirmed by symbolic acts" and arguing that "a version of Durkheim's social psychology can be effective in modern dress" (Goffman 2005[1967]: p. 47). Meyer and Rowan (1977) made very similar arguments about the modern organization: that it has become a sacred realm, and a locus of myth and ceremony in modernity. They built on Peter Berger and Thomas Luckmann's *The Social Construction of Reality* (1967) in the lineage of Alfred Schutz's *The Phenomenology of the Social World* (1967 [1932]). For Berger and Luckmann (1967: p.20), the task of sociology is to grasp "the objectivations of subjective processes (and meanings) by which the *inter*subjective commonsense world is constructed." The inclination to assign objective status to intersubjective reality is characteristic of human society, they contend. In mystified, religious, philosophical, and rationalized social systems alike, individuals make causal connections on the basis of the wider system of meaning institutionalized in concrete customs. People participate in sustaining ideas about reality, and thus, "Knowledge about society is … a realization in the double sense of the word, in the sense of apprehending the objectivated reality, and in the sense of ongoingly producing this reality …. The sociology of knowledge understands human reality as socially constructed reality" (Berger and Luckmann, 1967: pp. 210–211). Like tribal societies, modern societies trace social conventions to forces in the wider universe. But where Durkheim described totemic societies as tracing social conventions to the will of the animistic totems, today we trace conventions to self-interest and natural laws governing markets and bureaucracies. Modern culture is highly rationalized, but in modern societies, cultural meaning is organized around science and artificial laws of nature.

In 1983, Paul DiMaggio and Walter Powell built on Meyer and Rowan's ideas, sketching the ways in which new rational customs spread across organizations. Sometimes governments, or other extraorganizational forces, encourage organizations to adopt new conventions ('coercive isomorphism'), as when federal regulators dictate school standards (Meyer and Scott, 1983). Sometimes professional networks that span organizations promote new conventions with theories of their efficacy ('normative isomorphism'), as when finance managers promote the portfolio approach to corporate diversification (Fligstein, 1990). Sometimes managers copy practices of successful organizations without an idea of what makes them work ('mimetic isomorphism'), as when the U.S. automakers copied Toyota willy-nilly in the 1980s.

Diffusion of management myths, or cultures, made organizations within sectors quite homogenous in the postwar period. In recent years, diffusion across sectors has made organizations in very different sectors quite homogenous (Meyer, 1994). The corporate model, in particular, has spread to nonprofits and government agencies, such that they all have CEOs and mission statements.

How Power Shapes Culture

Power theorists have incorporated cultural insights from social constructionists to explain how groups promote policies and practices that serve their own interests by depicting them as in the interest of all. In Neil Fligstein's (1996) political–cultural approach to the firm, belief in the efficiency of institutions comes from rational theorization by consultants, economists, and management theorists. When an external shock – a recession, a policy shift – destabilizes one broad management regime, entrepreneurial consultants and managers promote innovations that are in their interest, or in the interest of their patrons. Thus the powerful are most likely to be able to put their preferred alternatives into place. They hawk those innovations by theorizing them as Pareto optimal, rather than as promoting the interests of a particular group. Thus they actively engage in the social construction of rationality, but of a politically infused rationality. Once in place, the theory holds the innovation in place. The conglomerate model of the firm, then, was shaken by policy and economic shifts *c.*1980, and institutional entrepreneurs successfully promoted a new model rooted in ideas about 'core competence' in management (Davis et al., 1994; Fligstein and Markowitz, 1993). Fligstein's power theory, then, builds directly on the insights that Meyer and Rowan developed in their institutional approach to explain how culture and legitimacy reinforce particular organizational strategies, across the organizational field.

William Roy, in *Socializing Capital*: *The Rise of the Large Industrial Corporation in America* (1997), sketches a somewhat different theory of the role of power in the modern firm, but one that also relies heavily on culture. For Roy, the social construction of efficiency keeps particular corporate strategies in place. Public and organizational policies that are adopted to support the interests of particular groups become legitimated through rhetorical strategies and economic theories, and group power differentials thus become institutionalized and obscured.

For instance, Roy (1997) explains the wave of mergers at the start of the twentieth century not as a consequence of economies of scale, as leading capitalists came to explain it. The cause was the enforcement of antitrust, which led large firms to force their smaller competitors to sell out because they could no longer fix prices together. The dominant players that emerged jumped on the idea of economies of scale to explain the merger wave, as a way to justify the rise of oligopolies and refute the notion that big industrialists had destroyed their smaller competitors. Thus the theory of economies of scale got a boost, and helped to justify concentrated ownership even in industries that enjoyed no scale economies. For Roy, once this pattern was established, it became self-reinforcing, largely because people make sense of the world around them by attributing rationality to practices, explaining surviving economic

solutions teleologically, as a consequence of evolution and natural selection. The successful theorization of new organizational practices means that they no longer have to be actively supported by the powerful. They become taken for granted, and advantages to certain groups (in the United States, owners of large corporations) become institutionalized.

Comparative students of organizations have explored differences in organizational cultures across nations. Hofstede (1980), for instance, shows that across countries, orientations to authority, innovation, and cooperation vary widely. The national business systems approach traces broad differences in organizational patterns not only to the organization-level construction of rationality, but to the wider constructions of economic actors and institutions that the organization engages with (e.g., Dore, 2000; Whitley and Kristensen, 1996).

New Directions in Cultural Approaches to Organizations

In what follows, we examine contemporary trends in the study of culture and organizations. Scholars have been examining how institutions travel across organizations and countries. They have been exploring the role of agency in institutional change. And they have been delving into the microfoundations of institutional persistence and change. Cultural approaches to organizations have been flourishing in recent years. Studies have shown the utility of a cultural and institutional approach for understanding organizational practices, their origins, and their sustainability.

How Institutions Travel

Studies by organizational institutionalists in the 1980s were dominated by research into processes of diffusion, or how organizational practices spread across fields (e.g., Tolbert and Zucker, 1983; Baron et al., 1986; Hirsch, 1986). This literature was later critiqued for an overly deterministic approach to diffusion, spurring a number of theoretical refinements and reconceptualizations of how institutions emerge and spread (Sahlin-Andersson, 1996). Building on the standard diffusion model, Strang and Meyer (1993) highlight the importance of expert theorization of new practices in the diffusion process, and reassert the importance of culture in the spread of organizational practices. Moving even further afield, Czarniawska and Joerges (1996) explicitly reject the idea of diffusion, drawing from Latour (1986) to suggest instead a model of translation and adaptation. Organizations do not simply adopt innovations, they translate and adapt them to local cultures. What goes by the name of matrix management may be entirely different in two different firms. While offering distinct theoretical contributions, both of these research streams complicate the institutional model of diffusion, offering greater attention to the symbolic nature of institutions through a nuanced account of how ideas, meanings, and practices move across an institutional field.

Theorization

A growing body of work has developed to challenge the traditional model of diffusion (e.g., Kennedy and Fiss, 2009). This line of research suggests several ways in which culture shapes the spread of organizational practices, first arguing that emergent practices diffuse not only through relational connections (or patterns of interaction) but also through cultural connections, and second arguing that emergent practices will diffuse more rapidly when they are theorized (Strang and Meyer, 1993; Greenwood et al., 2002). While previous diffusion studies focused on direct relationships, Strang and Meyer (1993) highlight the importance of 'cultural linkage,' meaning that two entities belong to the same constructed social category. A number of studies illustrate the importance of these shared social categories. In tracing the history of student activism in the United States, Sarah Soule (1997) finds a particular protest tactic spread quickly between similar universities, or those that shared institutional characteristics such as prestige, endowment size, and institutional type.

Theorization provides an account of why some innovations diffuse rapidly while others do not. Diffusion is aided by theorization: "the self-conscious development and specification of abstract categories and the formulation of patterned relationships such as chains of cause and effect" (Strang and Meyer, 1993: p. 492). Theorization includes two parts, the specification of an 'organizational failing' or a problem that the new practice will solve, and a story about how the innovation bests existing solutions (Tolbert and Zucker, 1996). Theorization allows new practices to be more readily adopted by organizations, as elements or properties of the practices are simplified and their outcomes are clearly defined (Greenwood et al., 2002). Thus theorization links the preinstitutionalization phase of institutional change, where local innovations or new ideas emerge, with the later stages of diffusion and institutionalization.

Theorization has been shown to aid diffusion processes across a number of empirical settings (Greenwood et al., 2002; Strang, 1997). For instance, Kelly and Dobbin (1998) trace the emergence of corporate diversity management, arguing that personnel professionals retheorized politically contentious affirmative action practices as diversity management practices. By retheorizing and renaming existing practices, to break the tie to federal regulations and create a tie to a looming change in the demography of the workforce and customers, personnel managers created a new rationale for practices such as race relations training (now 'diversity training') and boosted their diffusion (Kelly and Dobbin, 1998). This study points to the role of professionals in theorizing organizational practices.

Translation, Editing, and Adaptation

Sharing this focus on the spread of organizational practices, Scandinavian institutional theorists offer another perspective on the role of culture in institutional processes, with great attention to meaning, interpretation, and local context. Organizations, like individuals, are culturally embedded and thus the ideational and material aspects of organizational institutions (practices and their meanings) are translated or altered as they move from one place to another (Sahlin and Wedlin, 2008). As institutions travel they are modified to fit the new context; translation allows for variation to emerge based on local culture (Boxenbaum, 2006). From this perspective, change is both emergent and dynamic, neither fully strategic nor simply

the result of environmental pressures, but rather "a blend of intentions, random events and institutional norms" (Czarniawska and Sevon, 1996: p. 8).

In developing an account of 'how ideas travel,' Kerstin Sahlin-Andersson (1996) proposes an editing process. In part a reaction to the 'automatic process' of diffusion sketched by DiMaggio and Powell (1991), Sahlin-Andersson's (1996: p. 92) approach aligns more closely with that of Friedland and Alford (1991) in their original conception of institutional logics. While Friedland and Alford (1991) argue that multiple, sometimes contradictory, institutional logics allow for organizational choice in selecting particular practices, Sahlin-Andersson (1996) argues that variation in local practices can be explained by the editing process whereby new models are not simply imported wholesale (or selected due to the logic they instantiate) but are modified through a process of adaptation. This process entails shifts in both the material and ideational aspects of institutions to align with local models and practices.

Studies of translation highlight the importance of viewing institutions as embedded and best understood in context. In tracing the spread of corporate diversity management from the U.S. to Denmark, Eva Boxenbaum (2006) finds that the process of translation involves: (1) individual preference, as some elements of the frame appeal to particular translators, (2) strategic reframing of the practice, to best appeal to interests of powerful actors whose support is necessary for implementation, and (3) local grounding, to match the framing of the practice to existing practices in the local context.

Agency and Institutional Change

Institutional theory has long been critiqued for neglecting or undertheorizing agency (Battilana et al., 2009; Powell and Colyvas, 2008; Barley and Tolbert, 1997). As DiMaggio (1988) argued, institutional theory must account for "the reality of purposive, interest-driven, and conflictual behavior" (1988: p. 5). Recent work on institutional entrepreneurship has demonstrated a range of field-, organization-, and individual-level factors that enable individual actors to change institutions (Battilana, 2006; Greenwood and Hinings, 1996; Greenwood and Suddaby, 2006).

Institutional Logics

The institutional logics approach, while clearly building from the neoinstitutional framework, takes a sort of middle ground in theorizing the link between agency and structure in institutional studies. Proponents eschew rational actor models by characterizing individual interests, identities, and understandings as embedded within logics, while at the same time describing actors as capable of overcoming institutional constraints to innovate (Greenwood et al., 2008). Critical of the institutional focus on isomorphism and the inherent stability of institutions, the institutional logics camp developed a metatheory and analytical framework that allows an analysis of the interrelations between institutions, organizations, and individual actors (Thornton et al., 2012). There are various institutional orders – family, religion, state, market, profession, and corporation – and there are elemental categories of institutions aligned to each, including sources of legitimacy, authority, and identity, as well as bases of norms, attention, and strategy (see also Boltanski and Thevenot, 1991). The logics approach views cultural content, the symbols and practices of each logic, as potentially complementary or contradictory. It is through transposing symbols and practices across institutional orders within an institutional field that change occurs (Thornton, 2004; see also Padgett and Powell, 2012).

While logics can be seen as constraining, in as much as an actor's social position makes particular logics more accessible, the fact that actors are exposed to multiple logics allows for agency. Individuals and organizations are part of multiple institutional orders, which provide both symbolic and material resources. Thus actors can combine logics, highlight contradictions, and otherwise reinterpret and mobilize meanings and practices to drive institutional change (Thornton and Ocasio, 2008). A range of empirical work supports these theoretical claims. Scholars have shown that more than one logic may be present within an organization (Lounsbury, 2007; Reay and Hinings, 2009). Individuals within organizations may mobilize elements of disparate logics. They can also make use of elements of old institutions. Dacin and Dacin (2008) study deinstitutionalization, finding that even while institutions decline, their remnants remain potent and serve to allow for the reemergence or reinvention of institutions. Actors play a key role here, serving as 'custodians' of institutionalized practices.

Microfoundations of Institutions

In concert with calls for revisiting agency in organizational institutionalism, there have been calls for increased attention to the microfoundations of institutionalization and the process by which the institutional environment shapes organizational practices (Zucker, 1977; Tolbert and Zucker, 1983, 1996). Studies of the microfoundations of institutions can be traced back to the origins of the field. Lynne Zucker drew from both the phenomenology of Berger and Luckmann (1967) and the ethnomethodology of Garfinkel (1967) in developing an argument for how institutions consist of taken-for-granted meanings (Zucker, 1977; Tolbert and Zucker, 1996). These same traditions shaped the thinking of Meyer and Rowan (1977) and DiMaggio and Powell (1991). As DiMaggio and Powell argue, "ethnomethodology and phenomenology together provide the new institutionalism with a microsociology of considerable power" (1991: p. 21). However, as Barley (2008) and others have contended, neoinstitutional theory has not provided adequate attention to issues of agency or practice. Recently however, there has been increased attention within organizational institutionalism to the importance of the interactionist approach (Fligstein, 2001; Hallett and Ventresca, 2006), the microprocesses of institutionalization, and organizational change at the microlevel (Kellogg 2009; Johnson et al., 2000). While Thornton et al. (2012) provide a theoretical account of microfoundations in developing the institutional logics approach, empirical work at the microlevel is scant.

Bricolage

Campbell (1997) draws the term bricolage from the anthropologist Claude Levi-Strauss (1963) to describe knowledge

creation that is put together piecemeal from a variety of sources. Baker and Nelson (2005) use the idea of entrepreneurial bricolage to explain that, rather than being constrained by their resource environments, certain firms are able to make use of specific physical (materials, skills, labor), social (customers and markets), and institutional (standards and regulations) resources that other firms lack access to. This approach suggests that resource environments, or fields, are socially constructed. Baker and Nelson problematize the view that resources are objective or clearly definable, defining bricolage as "making do by applying combinations of resources at hand to new problems and opportunities" (2005: p. 333). A growing body of work employs bricolage in explaining the emergence of hybrid practices (Campbell, 1997) and the creation of new political organizations (Clemens, 1993).

Institutional Work

Institutional work is "the purposive action of individuals and organizations aimed at creating, maintaining, and disrupting institutions" (Lawrence and Suddaby, 2006: p. 215). This bottom–up approach to how actions shape institutions seeks a middle ground between 'cultural dopes' in the neoinstitutional literature and the overly agentic institutional entrepreneur literature (Powell and Colyvas, 2008). The approach also bridges institutional theory and practice theory. The overarching claim of institutional work theorists is that institutions should not be viewed as taken for granted and stable. Rather, institutions require maintenance, not simply via their regulative and normative pillars but also via the cultural–cognitive pillars (Lawrence and Suddaby, 2006).

Zilber (2009) considers how broad, societal metanarratives are used to support and give legitimacy to local narratives. Here we see how local institutions are shaped, in part, by broader context. As Zilber demonstrates, institutional maintenance can also include elements of translation, as institutional metanarratives are translated into organization-specific stories (Zilber, 2009). Zilber highlights the importance of power, as organizational elites shape the organizational 'reality' through their storytelling and translation of societal metanarratives.

Marti and Mair (2009) adopt an institutional work perspective to study rural health care organizations, finding that actors borrow logics from multiple domains, such as cultural and religious domains, in order to drive change. They argue that this is not a case of bricolage, where actors borrow institutional elements from another field. Rather, actors must 'navigate subtly' between several distinct logics that operate in multiple domains. For instance, efforts to provide microcredit to women in the developing world must fit within the financial and business logic, but also link to distinct cultural and religious norms.

Institutions are composed of both material and ideational elements. Berger and Luckmann (1967) conceive of institutions as made of practices and meanings that became taken for granted through enactment. Meyer and Rowan's 'rational myths' are simply taken-for-granted parts of modern institutions. While the symbolic has been a central part of neoinstitutional theory, empirical work has focused more prominently on organizational practices or material elements than on meanings and the ideational elements (Zilber, 2008). Heeding calls for more attention to meanings, studies of institutional work may inform understandings of the cognitive underpinnings of institutions. However, the institutional work literature is quite new. Recent studies test particular types of institutional work sketched by Lawrence and Suddaby (2006), suggesting there is much to be done in fleshing out the theoretical model.

Conclusion

The study of organizations and culture is one of the most active fields in organizational sociology today. Institutional theory is based in the social constructionism of Peter Berger and Thomas Luckmann (1967). The perspective first set out in Meyer and Rowan's (1977) *Institutionalized Organizations: Formal Structure as Myth and Ceremony* and developed in discussions of theory (Meyer et al., 1987; Meyer and Jepperson, 2000), describes the modern worldview as a historical construct. So are organizations and their various routines and practices. Meyer and colleagues offer a radical challenge to the taken-for-granted categories of modernity and the taken-for-granted causal relationships. In Berger and Luckmann's (1967: p. 20) brand of social phenomenology, which Meyer and colleagues draw on, the task is to grasp how collective understandings of the social and physical worlds around us become objectified, such that we do not see them as collective understandings. They ask how our subjective 'knowledge' of the world comes to have the feel of objective reality. Berger and Luckmann focus on the different intersubjective realities found in mystified, religious, and rational settings. These realities emerge through experience, and at the level of cognition, the individual makes causal connections on the basis of the wider system of meaning institutionalized in concrete customs. Students of organizations and culture first built on these insights by tracing the spread of new ideas about management and the practices associated with them. They went on to explore the role of expert theorization, as modern-day witch doctors who call themselves sociologists, economists, and management consultants divine laws of the organizational universe and promote practices that conform to those laws. In recent years they have turned their attention to the cognitive processes, mental maps, institutional logics, and microlevel processes that undergird, and ultimately that produce and change, organizational practices and their meanings.

See also: Collective Identity; Culture and Economy; Culture and Institutional Logics; Culture and Networks; Culture, Cognition and Embodiment; Ideologies, Institutions, and the New Institutionalism; Networks and Meaning; New Institutionalism in the Analysis of Complex Organizations; Old and New Institutionalism in Economics; Organizational Culture, Anthropology of; Organizations, Sociology of; Sociology and the New Institutionalism.

Bibliography

Baker, Ted, Nelson, Reed E., 2005. Creating something from nothing: resource construction through entrepreneurial bricolage. Administrative Science Quarterly 50 (3), 329–366.

Barley, Stephen, 2008. Coalface institutionalism. In: Greenwood, Royston, Oliver, Christine, Suddaby, Roy, Sahlin, Kerstin (Eds.), The SAGE Handbook of

Organizational Institutionalism. Retrieved 06.03.14.: http://www.barnesandnoble.com/w/sage-handbook-of-organizational-institutionalism-kerstin-sahlin-andersson/1100613929.
Barley, Stephen R., Kunda, Gideon, 1992. Design and devotion: surges of rational and normative ideologies of control in managerial discourse. Administrative Science Quarterly 37, 363–400.
Barley, Stephen R., Tolbert, Pamela S., 1997. Institutionalization and structuration: studying the links between action and institution. Organization Studies 18 (1), 93–117.
Baron, James N., Dobbin, Frank, Devereaux Jennings, P., 1986. War and peace: the evolution of modern personnel administration in U.S. Industry. American Journal of Sociology 92, 350–383.
Battilana, Julie, 2006. Agency and institutions: the enabling role of individuals' social position. Organization 13 (5), 653–676.
Battilana, Julie, Leca, Bernard, Boxenbaum, Eva, 2009. How actors change institutions: towards a theory of institutional entrepreneurship. The Academy of Management Annals 3 (1), 65–107.
Berger, Peter L., Luckmann, Thomas, 1967. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday & Company Inc., New York.
Boltanski, Luc, Thevenot, Laurent, 1991. De La Justification: Les Economies de La Grandeur. Gallimard, Paris.
Boxenbaum, Eva, 2006. Lost in translation the making of Danish diversity management. American Behavioral Scientist 49 (7), 939–948.
Campbell, John L., 1997. Mechanisms of evolutionary change in economic governance: interaction, interpretation and bricolage. In: Magnusson, Lars, Ottosson, Jan (Eds.), Evolutionary Economics and Path Dependence. Edward Elgar, Cheltenham, United Kingdom, pp. 10–32.
Clark, Burton R., 1960. The Open-door Colleges: A Case Study. McGraw-Hill, New York.
Clemens, Elisabeth S., 1993. Organizational repertoires and institutional change: women's groups and the transformation of US politics, 1890–1920. American Journal of Sociology 98 (4), 755–798.
Czarniawska, Barbara, Joerges, Bernward, 1996. Travels of ideas. In: Sevón, Guje, Czarniawska, Barbara (Eds.), Translating Organizational Change. de Gruyter, Berlin, pp. 13–48.
Czarniawska, Barbara, Sevon, Guje, 1996. Translating Organizational Change. Walter de Gruyer & Co., Berlin.
Dacin, M. Tina, Dacin, Peter A., 2008. Traditions as institutionalized practice: implications for deinstitutionalization. In: Greenwood, Royston, Christine, Oliver, Roy, Suddaby, Kerstin, Sahlin (Eds.), The SAGE Handbook of Organizational Institutionalism, pp. 327–351.
Davis, Gerald F., Diekmann, Kristina A., Tinsley, Catherine H., 1994. The decline and fall of the conglomerate firm in the 1980s: the deinstitutionalization of an organizational form. American Sociological Review 59, 547–570.
DiMaggio, Paul J., 1988. Interest and agency in institutional theory. In: Zucker, Lynne G. (Ed.), Institutional Patterns and Organizations. Ballinger, Cambridge, MA, pp. 3–21.
DiMaggio, Paul J., Powell, Walter W., 1991. The New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago.
Dore, Ronald, 2000. Stock Market Capitalism: Welfare Capitalism – Japan and Germany versus the Anglo-Saxons. Oxford University Press, New York.
Durkheim, Emile, 1961. The Elementary Forms of Religious Life. Collier, New York.
Fay, Brian, 1990. Critical realism? Journal for the Theory of Social Behaviour 20, 33–41.
Fligstein, Neil, 1990. The Transformation of Corporate Control. Harvard University Press, Cambridge, MA.
Fligstein, Neil, 1996. Markets as politics: a political-cultural approach to market institutions. American Sociological Review 61, 656–674.
Fligstein, Neil, 2001. The Architecture of Markets: An Economic Sociology of Twenty-first Century Capitalist Societies. Princeton University Press, Princeton, NJ.
Fligstein, Neil, Markowitz, Linda, 1993. Financial reorganization of American corporations in the 1980s. In: Wilson, William Julius (Ed.), Sociology and the Public Agenda. Sage Publications, Beverly Hills, pp. 185–206.
Friedland, Roger, Alford, Robert R., 1991. Bringing society back in: symbols, practices, and institutional contradictions. In: Powell Walter, W., DiMaggio, Paul J. (Eds.), The New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago, pp. 232–263.
Garfinkel, Howard, 1987. Studies of the routine grounds of everyday activities. Studies in Ethnomethodology, 35–75.
Goffman, Erving, 1974. Frame Analysis. Harvard University Press, Cambridge, MA.
Goffman, Erving, 2005. Interaction Ritual: Essays in Face to Face Behavior. Aldine Transaction, New Brunswick, NJ.
Greenwood, Royston, Hinings, Christopher R., 1996. Understanding radical organizational change: bringing together the old and the new institutionalism. Academy of Management Review 21 (4), 1022–1054.
Greenwood, Royston, Oliver, Christine, Suddaby, Roy, Sahlin-Andersson, Kerstin (Eds.), 2008. The SAGE Handbook of Organizational Institutionalism. SAGE, Los Angeles, London.
Greenwood, Royston, Suddaby, Roy, 2006. Institutional entrepreneurship in mature fields: the big five accounting firms. Academy of Management Journal 49 (1), 27–48.
Greenwood, Royston, Suddaby, Roy, Hinings, C.R., 2002. Theorizing change: the role of professional associations in the transformation of institutionalized fields. Academy of Management Journal 45, 58–80.
Hallett, Tim, Ventresca, Marc J., 2006. How institutions form: loose coupling as mechanism in Gouldner's patterns of industrial bureaucracy. American Behavioral Scientist 49, 908–924.
Hirsch, Paul M., 1986. From ambushes to parachutes: corporate takeovers as an instance of cultural framing and institutional integration. American Journal of Sociology 91, 800–837.
Hofstede, Geert, 1980. Culture's Consequences: International Differences in Work Values. Sage, Beverly Hills, CA.
Johnson, Gerry, Smith, Stuart, Codling, Brian, 2000. Microprocesses of institutional change in the context of privatization. Academy of Management Review 25 (3), 572–580.
Kellogg, Katherine, 2009. Operating room: relational spaces and microinstitutional change in surgery. The American Journal of Sociology 115, 657–711.
Kelly, Erin, Dobbin, Frank, 1998. How affirmative action became diversity management: employer response to antidiscrimination law, 1961–1996. American Behavioral Scientist 41, 960–984.
Kennedy, Michael T., Fiss, Peer C., 2009. Institutionalization, framing, and diffusion: the logic of TQM adoption and implementation decisions among U.S. hospitals. Academy of Management Journal 52, 897–918.
Latour, Bruno, 1986. The powers of association. In: Law, John (Ed.), Power, Action and Belief. A New Sociology of Knowledge?, pp. 264–280. Sociological Review Monograph, Keele.
Lawrence, Thomas B., Suddaby, Roy, 2006. Institutions and institutional work. In: Clegg, S.R., Hardy, C., Lawrence, T.B., Nord, W.R. (Eds.), The Sage Handbook of Organization Studies, pp. 215–254, p. 215.
Lévi-Strauss, Claude, 1963. The structural study of myth. In: Structural Anthropology. Basic, New York, pp. 206–231.
Lounsbury, Michael, 2007. A tale of two cities: competing logics and practice variation in the professionalizing of mutual funds. Academy of Management Journal 50, 289–307.
March, James G., Simon, Herbert A., 1958. Organizations. Wiley, New York.
Marti, Ignasi, Mair, Johanna, 2009. Bringing change into the lives of the poor: entrepreneurship outside traditional boundaries. In: Institutional Work. Cambridge University Press. Retrieved: http://dx.doi.org/10.1017/CBO9780511596605.004.
Martin, Joanne, 2002. Organizational Culture – Mapping the Terrain. Sage Publications, Thousand Oaks.
Mayo, Elton, 1933. The Human Problems of an Industrial Civilization. Graduate School of Business Administration, Harvard University, Boston.
Meyer, John W., 1994. Rationalized Environments. In: Scott, W.R., Meyer, John W. (Eds.), Institutional Environments and Organizations. R. Scott and J. Meyer. Sage, Newbury Park, pp. 28–54.
Meyer, John, Scott, W.R., 1983. Organizational Environments: Ritual and Rationality. Sage, Beverly Hills, CA.
Meyer, John W., Boli, John, Thomas, George M., 1987. Ontology and rationalization in the western cultural account. In: Thomas, George M., Meyer, John W., Ramirez, Francisco, Boli, John (Eds.), Institutional Structure: Constituting State, Society, and the Individual. Sage, Newbury Park, pp. 12–37.
Meyer, John W., Jepperson, Ronald, 2000. The 'actors' of modern society: cultural rationalization and the ongoing expansion of social agency. Sociological Theory 18, 100–120.
Meyer, John W., Rowan, Brian, 1977. Institutionalized organizations: formal structure as myth and ceremony. American Journal of Sociology 83, 340–363.
Padgett, John F., Powell, Walter W., 2012. The Emergence of Organizations and Markets. Princeton University Press, Princeton.
Powell, Walter W., Colyvas, Jeanette A., 2008. Microfoundations of institutional theory. In: Greenwood, Royston, Oliver, Christine, Sahlin, Kerstin, Suddaby, Roy (Eds.), The Sage Handbook of Organizational Institutionalism. Sage Publications, London, pp. 276–298.
Reay, Trish, Hinings, C.R., 2009. Managing the rivalry of competing institutional logics. Organization Studies 30 (6), 629–652.
Roethlisberger, Fritz J., Dickson, William, 1939. Management and the Worker. Harvard University Press, Cambridge, MA.
Roy, William, 1997. Socializing Capital: The Rise of the Large Industrial Corporation in America. Princeton University Press, Princeton, NJ.

Sahlin-Andersson, K., 1996. Imitating by editing success: the construction of organization fields. In: Czarniawska, B., Sevón, G. (Eds.), Translating Organizational Change. Walter de Gruyter, Berlin, New York, pp. 69–92.

Sahlin, K., Wedlin, Linda, 2008. Circulating ideas: imitation, translation, and editing. In: Greenwood, Royston, Oliver, Christine, Suddaby, Roy, Sahlin, Kerstin (Eds.), The SAGE Handbook of Organizational Institutionalism, pp. 218–242.

Schein, Edgar H., 1996. Culture: the missing concept in organization studies. Administrative Science Quarterly 41, 229–240.

Schutz, Alfred, 1967. The Phenomenology of the Social World. Northwestern University Press, Evanston, IL.

Selznick, Philip, 1957. Leadership in Administration: A Sociological Interpretation. Harper and Row, New York.

Soule, Sarah A., 1997. The student divestment movement in the United States and tactical diffusion: the Shantytown protest. Social Forces 75 (3), 855–882.

Strang, David, 1997. Cheap Talk: Managerial Discourse on Quality Circles as an Organizational Innovation. Presented at the annual meetings of the American Sociological Association, Toronto.

Strang, David, Meyer, John W., 1993. Institutional conditions for diffusion. Theory and Society 22, 487–511.

Thornton, Patricia H., 2004. Markets from Culture: Institutional Logics and Organizational Decisions in Higher Education Publishing. Stanford Business Books. Stanford University Press, Stanford, CA.

Thornton, P.H., Ocasio, W., 2008. Institutional logics. In: Royston, Greenwood, Oliver, Christine, Sahlin, Kerstin, Suddaby, Roy (Eds.), The Sage Handbook of Organizational Institutionalism, pp. 99–120.

Thornton, Patricia, H., Ocasio, William, Lounsbury, Michael, 2012. The Institutional Logics Perspective: A New Approach to Culture, Structure, and Process. Oxford University Press, Oxford, UK.

Tolbert, Pamela S., Zucker, Lynne G., 1983. Institutional sources of change in the formal structure of organizations: the diffusion of civil service reform, 1880–1935. Administrative Science Quarterly, 22–39.

Tolbert, Pamela S., Zucker, Lynne G., 1996. The institutionalization of institutional theory. In: Clegg, S., Hardy, C., Nord, W. (Eds.), Handbook of Organizational Studies. Sage, London, pp. 175–190.

Weick, Karl, 1979. The Social Psychology of Organizing. McGraw-Hill.

Weick, Karl E., 1993. The collapse of sensemaking in organizations: the Mann Gulch disaster. Administrative Science Quarterly 38, 628–652.

Weick, Karl E., 1995. Sensemaking in Organizations. Sage, Thousand Oaks, CA.

Whitley, Richard, Kristensen, Peer Hull, 1996. The Changing European Firm: Limits to Convergence. Routledge, London, New York.

Zald, Mayer N., Denton, Patricia, 1963. From evangelism to general service: the transformation of the YMCA. Administrative Science Quarterly 8, 214–234.

Zilber, Tammar B., 2008. The work of meanings in institutional processes and thinking. In: Greenwood, Royston, Oliver, Christine, Suddaby, Roy, Sahlin, Kerstin (Eds.), The SAGE Handbook of Organizational Institutionalism. SAGE, Los Angeles; London.

Zilber, Tammar B., 2009. Institutional maintenance as narrative acts. In: Institutional Work. Cambridge University Press. Retrieved: http://dx.doi.org/10.1017/CBO9780511596605.008.

Zucker, Lynne G., 1977. The role of institutionalization in cultural persistence. American Sociological Review 42, 726–743.

Organizations and the Law

Mark C Suchman, Brown University, Providence, RI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The relationship between organizations and the law is rich and complex. Organizations encounter law in at least three guises: Law as a *facilitative arena* for interaction; law as a *regulative code* for conduct; and law as a *constitutive schema* for experience. In each of these guises, law is simultaneously an independent variable affecting organizational activity and a dependent variable reflecting organizational activity. Moreover, whether as cause or as consequence, law operates simultaneously through material incentives and through cultural beliefs. This article explores the diverse body of scholarship on organizations and law that has emerged to make sense of this varied terrain.

Scholarship on organizations and the law addresses the relationship between formal organizations – such as private businesses, public agencies, and nonprofit associations – and the legal institutions that govern them. Since both organizations and law receive separate treatment elsewhere in this encyclopedia, this article concentrates solely on the connections between the two. These connections are elaborate and diverse, and different research traditions have explored different territories along three cardinal axes: First, various researchers have emphasized the relationship between organizations and at least three distinct aspects of law: (1) legal venues as *facilitative arenas* for organizational interaction (especially disputing), (2) legal precepts as *regulative codes* for organizational conduct, and (3) legal definitions as *constitutive schemas* for organizational experience. Second, the relationship between organizations and each of these aspects of law can involve multiple causal orderings: Law not only shapes organizational activity, but organizational activity simultaneously shapes law – and at times both take shape together. Third, for each aspect and each causal order, the relevant mechanisms may operate either through material rewards and punishments, or through cultural morals and meanings, or through complex mixtures of the two. The following pages examine the diverse images of organizations and the law that lie at the intersections of these cardinal dimensions.

Intellectual History

Classical Roots

Classical social theorists saw organizations and law as closely linked. To Marx (e.g., 1842, 1844), capitalist organizational structures provided the economic premises for 'bourgeois law,' with its characteristic coupling of formal individual rights – especially rights of property and contract – and repressive police interventions against the working class. To Durkheim (1893), society's evolution from a unified to a specialized division of labor (and the corresponding shift from communal to limited-purpose organizations) generated a parallel evolution in law, from punitive demonstrations of shared outrage to restitutive compensations for individualized damages. To Weber, 1978/1925: pp. 654–658, 956–1005), both modern organizations and modern law drew legitimacy from a shared 'legal-rational' reliance on purposeful, systematic, impersonal rules; and as a result, both embodied similar tensions between responsiveness and predictability, and between autonomy and domination.

Mid-Century Revival

Despite these early linkages, subsequent scholarship on organizations and on law developed for the most part independently. A few theorists (e.g., Hurst, 1964) attempted to bridge the gap, but most research in each tradition remained oblivious to the other. Since the mid-twentieth century, however, the separation between organizational and legal studies has become increasingly untenable. Organizations form through the legal act of incorporation, and they dissolve through the legal act of bankruptcy. In between, they raise capital under securities law, hire employees under labor law, and exchange goods and services under contract law. Law, for its part, arises from the demands of organizations, and it topples under pressure from organizations as well. In between, it develops meaning through organizational compliance, receives enforcement through organizational litigation, and encounters limits through organizational recalcitrance. Moreover, the organizational and legal worlds have become increasingly intertwined over time, as organizations experience rising levels of litigation, regulation, and certification, and as law grapples with an ever more bureaucratized society.

In light of these affinities, a reconciliation between organizational and legal studies was perhaps inevitable, and by the mid-1960s, the early signs of convergence were beginning to emerge. At least four distinct lines of scholarship contributed to this rapprochement: a political science literature on regulation, a sociology literature on legalization, an economics literature on transaction costs, and a legal literature on commercial contracting.

Studies of Regulation

From political science (and to a lesser extent economics and sociology), a number of researchers have explored the relationship between regulatory agencies and the industries those agencies supposedly oversee (e.g., Bernstein, 1955; Stigler, 1971; Posner, 1974; Peltzman, 1976; Hawkins, 1984; Kagan and Scholz, 1984). Although many such investigations have

started with the assumption that well-chosen penalties can deter organizational wrongdoing, the empirical evidence suggests a gloomier story: In what economists have labeled the 'market for regulation' (Ulen, 1980), industries often appear to 'capture' their regulators (Laffont and Tirole, 1991; Dal Bó, 2006), resulting in legal rules that limit competition and preserve prevailing practices, rather than imposing reforms. Moreover, even when agencies do attempt to regulate in the public interest, regulated organizations are less likely to be deterred from wrongdoing than to be impelled toward evasion. In recent years, the rhetoric of regulation has shifted away from top-down 'command and control,' toward a more collaborative philosophy of 'industry self-governance' (Lobel, 2004); however, even where regulation on-the-books conforms to the command-and-control paradigm, regulation in-action rarely looks so simple (Short and Toffel, 2008; *see* Regulation).

Studies of Legalization

The pessimism of the regulation literature has been counterbalanced by a second body of research, originating from sociology. Examining the emergence of lawlike organizational structures such as formal work rules and disciplinary hearings, Selznick and his followers (Selznick, 1969; Edelman, 1990; Dobbin et al., 1993) have suggested that administrative and political pressures can lead organizations to develop internal practices that draw upon the public legal order for models of fairness and objectivity. Over time, these introjections of legality tend to become 'institutionalized' – both within organizations as formal policies and procedures, and outside organizations as legal doctrines and statutes. The resulting 'legalization of the workplace' (Sitkin and Bies, 1993; Sutton et al., 1994), holds the potential to transform organizations from authoritarian hierarchies into normatively constrained polities, thereby limiting official power and securing substantial 'citizenship' rights for lower-level participants (Dobbin and Sutton, 1998; *see* Legalization).

Studies of Transaction Costs

In contrast to the literatures on regulation and legalization, a third body of scholarship, emerging primarily from economics, downplays concerns about making organizations more socially responsible and focused instead on making them more efficient. Drawing heavily on the work of Coase (1937) and Williamson (1975), this tradition links law and organizations through a set of generalizable arguments about the costs and benefits of 'hierarchical governance,' whether in the form of legal rules or organizational policies. Because human beings are only 'boundedly rational' (i.e., unable to anticipate all future contingencies) and 'opportunistic' (i.e., prone to self-interested dishonesty), certain recurring exchanges will exhibit sizable 'transaction costs' in the market: Detailed long-term contracts will be difficult to write and enforce, and hierarchical authority structures may consequently prove more efficient than market trading. Since organizations and laws represent alternative forms of hierarchical governance, this analysis speaks equally to the question of when the legal system should intervene in market activity and to the question of when organizations should internalize market exchanges (*see* Transaction Costs and Property Rights).

Studies of Commercial Contracting

Somewhat ironically, at about the same time that transaction cost economists began calling attention to the trade-offs between market contracts and hierarchical authority relations, a small group of legal scholars began calling attention to the remarkably minor role of formal contracting even in the market itself. Sparked by the empirical fieldwork of Stewart Macaualy (1963), and systematized in the 'relational contract' writings of Ian Macneil (1977), this tradition highlights the extent to which ongoing business relations give rise to informal norms of reciprocity, forbearance, and mutualistic decision-making that largely displace the right-versus-wrong, winner-take-all dichotomies of the formal law. Subsequent economic accounts have debated whether such 'incomplete contracts' (Hart and Moore, 1988) and 'private ordering' regimes (Ellickson, 1991; Bernstein, 1992) serve to minimize transaction costs. However, the original insight of the relational contract literature was less about economic efficiency than about social embeddedness: Most business commitments arise within preexisting organizational communities, where formally alienable legal rights often become inextricably entangled with identity-defining moral claims (Suchman, 2003). This proposition offers a sociological counterpoint to economic theories of 'transaction cost engineering' (Gilson, 1984), in much the same way that the legalization literature offers a sociological counterpoint to economic theories of the 'political market' for regulation (Peltzman, 1976).

Current Scholarship

In recent years, these four bodies of research have spawned a resurgence of scholarship on organizations and the law. This resurgence spans the disciplines of political science, sociology, and economics, with offshoots extending into traditional legal and managerial studies as well. The present section reviews the state of this emerging literature, with emphasis on the three major aspects of the legal environment introduced in the opening paragraph above. Following Edelman and Suchman (1997), these aspects are designated the 'facilitative,' 'regulative,' and 'constitutive' legal environments, respectively.

Law as a 'Facilitative' Arena for Interaction

At its most passive, the legal system simply provides places, personnel, and procedures to accommodate organizational interactions that might otherwise occur elsewhere. The archetypal manifestation of this 'facilitative' aspect of law is the lawsuit: Organizations routinely employ litigation to open or defend markets, to attack or protect reputations, to reveal or conceal information, and so on. Courthouses are not the only relevant legal facilities, however; for example, organizations also use standard-setting bodies as vehicles for market dominance, patent filings as signals to potential investors, and Freedom of Information Act requests as instruments of corporate espionage. Across these and other contexts, the unifying theme of the facilitative legal environment is that, for better or worse, legal arenas offer congenial public venues for organizational activity, much as do public marketplaces and municipal convention centers.

As the preceding examples suggest, contemporary legal systems provide facilities for a wide array of organizational endeavors. To date, however, research on facilitative law has focused almost exclusively on one: law as an arena for disputing. Legal disputing takes place not only through traditional courtroom litigation, but also 'in the shadow of the law' (Mnookin and Kornhauser, 1979), through appended fora offering alternative dispute resolution (ADR) techniques such as mediation and arbitration (*see* Litigation; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Dispute Resolution in Economics). The best available evidence suggests that, in the United States at least, litigation against and among organizations has increased significantly since the early 1970s, at least in part because of a rise in the rate at which businesses sue one another (Dunworth and Rogers, 1996; Cheit and Gersen, 2000). Alongside this rise in interorganizational litigation, however, recent years have also brought increased use of ADR, which proponents claim can curtail the cost, delay, and adverse publicity of conventional adjudication (Lande, 2000). The resulting regime has been described as "process pluralism" (Bush, 1984), with an elevated level of overall disputing accompanying a reduced level of courtroom centrality.

Facilitative Law as Independent Variable

Although most studies of facilitative law portray organizations as the active players and the legal system as merely a passive arena, most also emphasize that the contours of this arena may determine the course of the game. In this sense, facilitative law tends to be depicted as an exogenous contextual variable in the disputing process (Hendley et al., 1999; Djankov et al., 2003; Ginsburg and Hoetker, 2006). Because much of this literature emerges from policy debates over litigation reform, investigation generally centers on whether particular procedural rules elicit or suppress lawsuits, and on whether changing rates of litigation exert positive or negative influences on nonlegal organizational activities (Dreyfuss, 1995; Choi et al., 2009). Standing-to-sue doctrines, class-action requirements, and contingency fee arrangements have all attracted attention as determinants of litigation behavior, and rising litigation has, in turn, received blame for a wide variety of adverse side effects, including increased insurance use, elevated bankruptcy rates, and a societal undersupply of high-risk innovation (Huber and Litan, 1991; Viscusi and Moore, 1993; Kagan, 2001). At the same time, a more upbeat literature on ADR asserts that the growing burdens of full-dress litigation have sparked managerial interest in nonjudicial alternatives, and that the resulting shift in dispute processing may have eased workplace tensions and strengthened ongoing business relations (Lande, 2000; Stipanowich, 2004). Other commentators, however, have noted that litigation may serve the salutary purpose of promoting corporate social responsibility, and that ADR, in contrast, may cool out genuine grievances and undercut public debate (Edwards, 1986; Edelman et al., 1993; Hensler, 2002).

Research on the link between forum structure and disputing behavior generally adopts a materialist orientation, examining how legal procedures affect the concrete costs and benefits of various organizational actions. If litigants enter court strategically, then the 'house rules' should determine both the attractiveness of the arena and the moves of the players. In addition to such obvious factors as expense, delay, and the magnitude of awards, litigation rates should be particularly sensitive to the predictability of trial outcomes, since if both sides have similar expectations about the results of a trial, rational parties will generally prefer to settle their disputes out of court (Priest and Klein, 1984). The most efficient judicial system, in this view, would be one in which the predictability of legal procedures encouraged disputants to resolve most matters through private negotiation, leaving the courts free to decide genuinely contested points of law with as little expense and delay as possible.

Recently, however, a number of culturalist researchers have borrowed from the literature on individual-level disputing to suggest that such rational-actor analyses may provide only a partial account, at best. Even among sophisticated business organizations, legal disagreements often involve substantial expressive and symbolic elements, and these cultural factors may play as great a role as rational profit maximization in determining organizational disputing behavior (Cheit, 1991; Lind et al., 1993). Thus, where materialists see litigation as 'economic competition by other means,' culturalists emphasize the relationship between law and 'softer' organizational phenomena such as alliance formation, community building, and reputation management (Blegvad, 1990). This perspective is particularly common in studies of ADR, which highlight the ability of dispute processing institutions to preserve or disrupt relationships, to harmonize or polarize viewpoints, and to privatize or publicize disagreements (Fuller et al., 2000; Hoffmann, 2005).

Facilitative Law as Dependent Variable

Alongside these analyses of how legal arenas shape organizational behavior, a handful of studies have reversed the causal arrow, to explore how organizational behavior shapes legal arenas. As the transaction cost tradition suggests, law is only one of many governance mechanisms, and if litigation arises in part from the failure of other controls, then extralegal organizational conditions may play a crucial role in shaping the work of the courts.

For the most part, the existing literature assumes that the relevant extralegal conditions are material rather than cultural, with factors such as economic globalization, industrial turbulence, and a shift from 'relational' to 'transactional' commerce topping the list (Cheit and Gersen, 2000; Kenworthy et al., 1996). Clearly, however, cultural conditions may also exert a significant influence. Preliminary findings suggest, for example, that professional ideologies strongly affect whether an organization's culture favors ADR or litigation; and organizational cultures, in turn, affect whether employees seek to resolve workplace grievances in-house or through the courts (Bies and Tyler, 1993; Fuller et al., 2000). Further, in a mirror image of Selznick's legalization dynamic, the spread of ADR techniques in the workplace tends to inject private organizational values into public legal discourses: Several major employment laws now explicitly encourage the use of ADR to resolve legal claims, and courts themselves increasingly employ 'appended' ADR to expedite settlement negotiations (Edelman et al., 2011; Hensler, 2002). Thus, the changes that began with rising litigation rates now appear to be coming full circle: Where organizations originally embraced ADR to escape the perils of formal legal arenas, the legal system is now

remodeling its own facilities to replicate these once-alien organizational proceedings. Even when legal institutions are at their most passive, the relationship between law and organizations proves to be both reciprocal and complex.

Law as a Code of Conduct

In addition to providing a procedural forum for organizational interaction, law can also act as a system of substantive edicts, imposing societal authority over various domains of organizational activity – from antitrust to labor relations to environmental protection (*see* Antitrust Policy: Lessons from the US Labor Unions). Formal regulatory oversight represents the archetypal manifestation of this 'regulative' aspect of law. But similar applications of sovereign authority can also occur through common-law precedents, street-level policing, and privately-enforced policy initiatives, such as tort claims against harassment, nuisance, or fraud. In all its guises, regulative law operates not merely as a neutral arena, but as a structure of precepts and enforcement mechanisms purporting to channel organizational behavior in particular directions. While regulative law may not always accomplish (or even sincerely intend to accomplish) its stated goals, the hallmarks of this aspect of the legal environment are a pervasive rhetoric of active social control and a pervasive concern with organizational compliance and evasion.

Regulation is almost certainly the most widely studied aspect of the relationship between organizations and the law. Unfortunately, most investigations to date have focused narrowly on particular substantive domains, and systematic efforts to explore underlying social processes remain relatively rare. Nonetheless, this literature displays a number of recurring themes. From a materialist perspective, regulative law serves to counterbalance market imperfections, imposing supplemental incentives and deterrents that recalibrate the implicit prices of various organizational behaviors. In this view, organizations respond to regulation by seeking to maximize total returns in light of the legal incentive structure, whether through petition and compliance or through manipulation and evasion. Alongside such economic dynamics, however, legal regulations also implicate deep-seated normative and cognitive beliefs (*see* Norms), and tactical maneuvers often cannot fully extricate organizations from these broader cultural processes: As statements of public values, laws shape and reflect what organizational stakeholders want, and as repositories of public meaning, laws frame and recapitulate what organizational stakeholders expect. Consequently, the underlying cultural logics of the organizational and legal realms often move in tandem, even as particularly recalcitrant firms may seek to amend or evade particularly burdensome strictures.

Regulative Law as Independent Variable

Tracing, first, the pathway from regulative edicts to organizational responses, research suggests that such legal pronouncements can operate both as material threats and as cultural sermons. As material threats, laws exert their greatest influence when standards are clear, sovereignty uncontested, and sanctions substantial (*see* Compliance and Legal Authority). In practice, however, such ideal conditions rarely pertain, and in their absence, self-interested organizations react not only to the incentives enunciated by law-on-the-books, but also to the gaps provided by law-in-action – ambiguous statutes, fragmented jurisdictions, manipulable enforcement agencies, and the like. These gaps, substantial in themselves, grow still wider in the face of organizational characteristics such as specialization and decentralization, which often marginalize legal voices and insulate managers from responsibility for corporate transgressions (Stone, 1975; Ashforth and Anand, 2003). As a result, the history of regulation is replete with unintended consequences: disclosure requirements that foster uninformative disclaimers (Samuels, 1996); tort liability that shifts hazardous activities into undercapitalized subsidiaries (Barney et al., 1992); air-quality standards that export pollutants to downwind communities (Ackerman, 1981). Although few researchers would rule out the possibility of successful deterrence, the common theme behind much of the materialist literature is that regulative law can easily mobilize evasion and resistance rather than obedience.

The pessimism of this literature stems in large part from the assumption that rational organizations lack the moral sentiments that promote individual-level respect for law (*see* Compliance and Legal Authority). Recent culturalist studies, however, call this assumption into question (Gunningham et al., 2005). Organizational behavior, after all, reflects the activities of culture-bearing individuals operating within decision structures that incorporate cultural values and cultural templates at every turn; it would be a bit odd, then, if organizational behavior responded only to law's material threats, while remaining impervious to law's cultural sermons (Howard-Grenville et al., 2008).

The leading attempt to examine regulative law through such a culturalist lens comes from 'neo-institutional' sociologists, working broadly in the Selznickian tradition. To these researchers, legal compliance simply represents a special case of 'institutional isomorphism' – the process by which organizations facing similar external rule systems become both increasingly alike and increasingly congruent with the myths of their shared environment (DiMaggio and Powell, 1983). In addition to applying material coercion, legal institutions also homogenize organizational behavior by shaping cultural beliefs: law enunciates normative values and ethics, which organizations then elaborate and internalize; and law provides cognitive models and scripts, which organizations then imitate and enact (Fuller et al., 2000; Stryker, 2003; Albiston, 2005). While the resulting isomorphism often amounts to little more than superficial ceremony, the neo-institutional perspective suggests that even 'empty' symbolism reenacts and thereby substantiates the law's cultural claims in ways that outright defiance would not.

Thus, investigations into the impact of regulative law on organizational activity suggest an ironic result: As a material deterrent, law is rarely sufficiently explicit, authoritative, and coercive to accomplish much beyond mere symbolism. As a cultural pronouncement, however, 'mere symbolism' can prove quite consequential – often extending a law's practical reach (for better or worse) well beyond the law's professed targets.

Regulative Law as Dependent Variable

Investigations into the reciprocal impact of organizational activity on regulative law reveal similarly complex patterns. At a

basic level, research confirms conventional wisdom: Organizations make substantial investments in electoral politics and backstage agenda-setting, and organizational actors occupy central positions on all sides of most policy networks (*see* Policy Networks). Contrary to many popular accounts, however, organizations often seem less interested in averting government 'interference' than in using regulation proactively to stabilize markets or manage competition (Fligstein, 1996). Legal incentives are easier to control than market prices, and as materialist accounts would predict, organizations invest heavily in securing this resource pool. These efforts extend well beyond formal lawmaking to encompass subsequent implementation activities as well. Legislative compromises can leave regulatory agencies with little substantive guidance and even less monetary support, and as a result, administrators often find themselves dependent on outside organizations for information, expertise, and even political cover (Baldwin and Cave, 1999; Pedriana and Stryker, 2004). Far from imposing top-down social control, the resulting process of regulatory capture – in which regulators become subservient to those they supposedly regulate – tends merely to institutionalize the indigenous practices of the regulated industry itself (*see* Regulation, Political Economy of).

Culturalist accounts supplement this Machiavellian power-dependence analysis in two ways. First, organizational participation in the lawmaking process may reflect ideological concerns, as well as material interests. Professional and social movement organizations, in particular, often blend moral entrepreneurship with institutional entrepreneurship, to align the regulative environment with their own ethics and worldviews (Hoffman, 1999; Greenwood et al., 2002; Dobbin, 2009; *see* Entrepreneurship and Organizations). Second, organizational and industrial culture may act as a 'filtering agent' between law-on-the-books and law-in-action (Kidder, 1983). Since popular understandings of legality come largely from mundane experiences in concrete settings, living law ultimately emerges not from legislatures or regulatory agencies but from real-world communities, where organizational beliefs and practices give form and content to otherwise vague legal principles. Indeed, many legal enactments are initially so ambiguous and contradictory as to do little more than signal that something in the regulative environment has changed – leaving the actual meaning of compliance to be constructed by subject organizations themselves, through an iterative process of interpretation, imitation, litigation, and reinterpretation (Edelman, 1992).

This culturalist perspective suggests a somewhat less sinister view of regulatory capture. Both regulator and regulated operate within the same institutional field, and shared exposure to a common set of filters and experiences can produce congruent outlooks even in the absence of *quid pro quo* power relations. Moreover, this dynamic can sometimes amplify, rather than dampen law's regulative impact: The literature reveals several instances in which savvy lobbyists, lawyers, and other professions have actually exaggerated legal demands in order to increase their own symbolic and practical importance – and have thereby constructed a regulatory environment that is more constraining than the original statutory framework may have envisioned (Edelman et al., 1992; Dobbin, 2009).

Law as a Typology of Experience

As well as providing an arena for interaction and a lever for social control, law can also enter organizational life more subtly, as a framework of categories and definitions that identifies the legally recognized components of the social world and that explains their natures and attributes. The archetype of this 'constitutive' aspect of law is the doctrine of corporate personhood, which defines particular social groups as unitary collective actors with coherent legal identities, interests, and capacities (*see* Multinational Corporations). Beyond incorporation, however, constitutive law also underpins virtually every other feature of the organizational world, from financial securities to intellectual property to employment relations (*see* Intellectual Property: Law in Context; Employment and Labor Regulation). Unlike facilitative and regulative law, constitutive law rarely attracts public attention, but its invisibility belies its agenda-setting power: By establishing fundamental background assumptions, legal definitions help to determine which types of organizations come into existence and which types of organizational activity come to be taken-for-granted. As a collection of basic templates and mechanisms, constitutive law often seems more placid and routine than its facilitative and regulative counterparts; in reality, though, this aspect of the legal environment provides the cognitive building blocks that underpin the other two.

Constitutive Law as Independent Variable

Because the materialist perspective emphasizes concrete conditions, it tends to downplay law's constitutive aspects: The physical realities of organizational life, one might argue, do not depend on the presence or absence of legal definitions – especially given that economic actors can usually specify, by mutual agreement, anything not already defined by law. Nonetheless, hints of a materialist approach to constitutive law appear, at least embryonically, in certain writings from the transaction cost school. Masten (1990), for example, suggests that the fundamental distinction between markets and hierarchies lies in the different 'default rules' that govern these two types of economic activity. Although a carefully crafted network of market contracts could theoretically achieve the same results as a corporate charter, the ready-made template of incorporation reduces the information and coordination costs of constructing such an architecture (Klausner, 1995). In a world of bounded rationality, many organizations would never come into existence without this handy shortcut. Law, then, can exert a substantial constitutive influence on organizational life simply by creating a basic tool kit of familiar categories and prebundled rights (Campbell and Lindberg, 1990).

Culturalist analyses go even further, seeing constitutive law not merely as a convenient starting point for strategic negotiation, but more fundamentally as a preconscious framework for making sense of the social world. In this view, law plays a central role in the "social construction of reality" (Berger and Luckmann, 1966), with legal definitions often becoming reified and institutionalized as taken-for-granted components of 'the way the world works' (*see* Law as Constitutive; Reification: History of the Concept). Culturalist researchers have explored such dynamics in several aspects of organizational life, including interorganizational relations, organizational structures, and the general cultural logic of 'legal rationality.' In the

constitution of interorganizational relations, contract law establishes rituals for forming binding commitments, property law determines the appropriability of resources and ideas, tort law assigns responsibility for social misfortunes, and so on (Suchman, 2003; Kapczynski, 2008; Clarke and Short, 1993; *see* Contracts: Legal Perspectives; Property: Legal Aspects; Torts in Law). In the constitution of organizational structures, law defines the characteristics of particular organizational features, such as affirmative action policies and 'poison pill' takeover defenses, and law also formulates the ground rules for entire organizational forms, such as private companies, public agencies, and nonprofit associations (Edelman, 1992; Dobbin et al., 1993; Davis, 1991; Salamon and Anheier, 1997). Finally, in the constitution of legal rationality, law generates sympathetic vibrations in other formalized and rationalized social structures, including organizational bureaucracies, drawing organizational embodiments of 'fairness,' 'citizenship,' and the like into synchrony with legal exemplars (Edelman, 1992; Sutton et al., 1994).

Constitutive Law as Dependent Variable

In depicting legal definitions as basic atoms of social structure, researchers from both the materialist and the culturalist camps tend to give short shrift to the study of constitutive law as a dependent variable. Nonetheless, legal categories, no matter how fundamental, are never facts of nature, and investigations into the relationship between organizations and the law have much to gain from examining how organizational activity may shape various constitutive constructs.

The existing literature offers several thought-provoking suggestions. First, in Gilson's (1984) memorable phrase, corporate lawyers may act as 'transaction cost engineers,' responding to unresolved governance challenges by recombining preexisting constitutive building blocks into novel configurations. Over time, the resulting 'legal devices' can serve as prototypes for increasingly routine reproductions – and the resulting standardized and typified artifacts can in turn provide the basis for revised legal categories (Powell, 1993; Suchman, 2003). Second, even without intentional innovation, organizations may inadvertently destabilize established legal definitions through imperfect implementation or awkward elaboration, generating legal pluralism and contradiction as an unintended consequence of everyday activity (cf Smets et al., 2012). Third, since legitimacy often stems from sheer numerical prevalence, organizations may skew the development of constitutive law simply by disproportionately employing some legal alternatives rather than others (Suchman, 2000; Edelman et al., 2011). Finally, as described above, organizations may participate directly in creating, interpreting, and implementing new legal categories and definitions, shaping the constitutive legal environment as advocates, as filtering agents, and as the concrete 'data points' that give substance to the law's abstract labels.

Future Directions

The study of organizations and the law has accomplished much in recent years; however, much still remains to be done. At a technical level, although both empirical and conceptual analyses abound, the former tend to focus on narrow policy concerns, while the latter tend to rest on meager evidentiary foundations. The challenge for future researchers will be to build firmer linkages between theory and data, and from those linkages, to develop a systematic body of intercomparable results. Cross-temporal and cross-cultural studies, in particular, seem essential to resolving the field's persistent debates over causal sequences and material/cultural mechanisms.

As such research takes shape, investigators are likely to encounter many phenomena that bridge the three major aspects of law and that comingle material and cultural processes. Ambiguous regulative statutes, for example, acquire meaning largely by motivating organizations to invent experimental legal devices that recombine old and new constitutive definitions in pursuit of plausible and advantageous models of 'compliance.' These recombination efforts inevitably generate test cases in facilitative legal arenas, which cases in turn further clarify and institutionalize the underlying constitutive definitions, the prevailing regulative standards, and the legitimate compliance options. In such dynamics, cultural and material considerations presumably overlap, since most laws set forth norms and penalties simultaneously and most legal devices blend technical efficacy and symbolic meaning. Indeed, under certain circumstances, the material/cultural distinction can virtually disappear, as when material punishment becomes a symbol of societal commitment, or when cultural legitimacy becomes an asset in market competition (Suchman, 1997).

In exploring such intersections, future research may accelerate a trend, already emerging among neo-institutionalists (e.g., Edelman et al., 2011), toward conceptualizing the relationship between organizations and law less as a matter of unilateral causation than as a matter of endogenous coevolution: Given the inherent ambiguity of law-on-the-books, the practical meaning of law-in-action develops only through a highly interactive process of social construction – involving both symbolic sense-making and concrete dealmaking. This 'structuration' dynamic (Barley and Tolbert, 1997) encompasses not only agents of the official legal system but also members of the focal organizational field, all tacitly collaborating to create simultaneously the forums and procedures of facilitative law, the demands and sanctions of regulative law, and the schemas and devices of constitutive law. Eventually this collaboration may produce a working agreement on what the law 'is' and on what it 'requires,' and that agreement may then become institutionalized in formal conventions and taken-for-granted beliefs. But saying that legal rules 'cause' organizational practices (or vice versa) would be a gross oversimplification. Rather, rules and practices arise in tandem, as part of the consolidation of a complexly intersectional organizational/legal field.

See also: Legal Personality; Regulation and Administration; Science and Law.

Bibliography

Ackerman, B., 1981. Clean Coal/Dirty Air: Or How the Clean Air Act Became a Multibillion-Dollar Bail-Out for High-Sulfur Coal Producers. Yale University Press, New Haven.

Albiston, C.R., 2005. Bargaining in the shadow of social institutions: competing discourses and social change in workplace mobilization of civil rights. Law and Society Review 39, 11–50.
Ashforth, B.E., Anand, V., 2003. The normalization of corruption in organizations. Research in Organizational Behavior 25, 1–52.
Baldwin, R., Cave, M., 1999. Understanding Regulation: Theory, Strategy, and Practice. Oxford University Press, New York.
Barley, S.R., Tolbert, P.S., 1997. Institutionalization and structuration: studying the links between action and institution. Organization Studies 18, 93–117.
Barney, J.B., Edwards, F.L., Ringleb, A.H., 1992. Organizational responses to legal liability: employee exposure to hazardous materials, vertical integration, and small firm production. Academy of Management Journal 35, 328–349.
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, Garden City, NY.
Bernstein, M., 1955. Regulating Business by Independent Commission. Princeton University Press, Princeton, NJ.
Bernstein, L., 1992. Opting out of the legal system: extralegal contractual relations in the diamond industry. Journal of Legal Studies 21, 115–157.
Bies, R.J., Tyler, T.R., 1993. The "litigation mentality" in organizations: a test of alternative psychological explanations. Organization Science 4, 352–366.
Blegvad, B.-M., 1990. Commercial relations, contract, and litigation in Denmark: a discussion of Macaulay's theories. Law and Society Review 24, 397–411.
Bush, R., 1984. Dispute resolution alternatives and the goals of civil justice: jurisdictional principles for process choice. Wisconsin Law Review 893–1034.
Campbell, J., Lindberg, L., 1990. Property rights and the organization of economic activity by the state. American Sociological Review 55, 634–647.
Cheit, R.E., 1991. Corporate ambulance chasers: the charmed life of business litigation. Studies in Law, Politics, and Society 11, 119–140.
Cheit, R.E., Gersen, J.E., 2000. When businesses sue each other: an empirical study of state court litigation. Law and Social Inquiry 25, 789–816.
Choi, S.J., Nelson, K.K., Pritchard, A.C., 2009. The screening effect of the private securities litigation reform act. Journal of Empirical Legal Studies 6, 35–68.
Clarke, L., Short Jr., J.F., 1993. Social organization and risk: some current controversies. Annual Review of Sociology 19, 375–399.
Coase, R., 1937. The nature of the firm. Economica 4, 386–405.
Dal Bó, E., 2006. Regulatory capture: a review. Oxford Review of Economic Policy 22, 203–225.
Davis, G.F., 1991. Agents without principles? The spread of the poison pill through the intercorporate network. Administrative Science Quarterly 36, 583–613.
DiMaggio, P.J., Powell, W.W., 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. American Sociological Review 48, 147–160.
Djankov, S., Porta, R.L., Lopez-de-Silanes, F., Shleifer, A., 2003. Courts. Quarterly Journal of Economics 118, 453–517.
Dobbin, F., 2009. Inventing Equal Opportunity. Princeton University Press, Princeton, NJ.
Dobbin, Frank, John R. Sutton, 1998. The strength of a weak state: the rights revolution and the rise of human resources management divisions. American Journal of Sociology 104 (2), 441–476.
Dobbin, F., Sutton, J.R., Meyer, J.W., Scott, R., 1993. Equal opportunity law and the construction of internal labor markets. American Journal of Sociology 99, 396–427.
Dreyfuss, R.C., 1995. Forums of the future: the role of specialized courts in resolving business disputes. Brooklyn Law Review 61, 1–44.
Dunworth, T., Rogers, J., 1996. Corporations in court: big business litigation in U.S. federal courts, 1971–1991. Law and Social Inquiry 21, 497–592.
Durkheim, E., 1964[1893]. The Division of Labor in Society. Free Press, New York.
Edelman, L., 1992. Legal ambiguity and symbolic structures: organizational mediation of law. American Journal of Sociology 97, 1531–1576.
Edelman, L.B., 1990. Legal environments and organizational governance: the expansion of due process in the american workplace. American Journal of Sociology 95, 1401–1440.
Edelman, L., Krieger, L., Eliason, S., Albiston, C., Mellema, V., 2011. When organizations rule: judicial deference to institutionalized employment structures. American Journal of Sociology 117, 888–954.
Edelman, L.B., Erlanger, H.S., Lande, J., 1993. Internal dispute resolution: the transformation of civil rights in the workplace. Law and Society Review 27, 497–534.
Edelman, L.B., Abraham, S.E., Erlanger, H.S., 1992. Professional construction of law: the inflated threat of wrongful discharge. Law and Society Review 26, 47–83.
Edelman, L.B., Suchman, M., 1997. The legal environments of organizations. Annual Review of Sociology 23, 479–515.
Edwards, H.T., 1986. Alternative dispute resolution: panacea or anathema? Harvard Law Review 99, 668–684.
Ellickson, R., 1991. Order without Law: How Neighbors Settle Disputes. Harvard University Press, Cambridge, MA.
Fligstein, N., 1996. Markets as politics: a political-cultural approach to market institutions. American Sociological Review 61, 656–673.
Fuller, S.R., Edelman, L.B., Matusik, S.F., 2000. Legal readings: employee interpretation and mobilization of law. Academy of Management Review 25, 200–216.
Gilson, R., 1984. Value creation by business lawyers: legal skills and asset pricing. Yale Law Journal 94, 239–313.
Ginsburg, T., Hoetker, G., 2006. The unreluctant litigant? An empirical analysis of Japan's turn to litigation. Journal of Legal Studies 35, 31–59.
Greenwood, R., Suddaby, R., Hinings, C.R., 2002. Theorizing change: the role of professional associations in the transformation of institutionalized fields. Academy of Management Journal 45, 58–80.
Gunningham, N.A., Thornton, D., Kagan, R.A., 2005. Motivating management: corporate compliance in environmental protection. Law and Policy 27, 289–316.
Hart, O., Moore, J., 1988. Incomplete contracts and renegotiation. Econometrica 56, 755–785.
Hawkins, K., 1984. Environment and Enforcement: Regulation and the Social Definition of Pollution. Clarendon Press, Oxford, UK.
Hendley, K., Murrell, P., Ryterman, R., 1999. Do repeat players behave differently in Russia? Contractual and litigation behavior of Russian enterprises. Law and Society Review 33, 833–867.
Hensler, D.R., 2002. Suppose it's not true: challenging mediation ideology. Journal of Dispute Resolution 2002 (1), 81–99.
Hoffman, A.J., 1999. Institutional evolution and change: environmentalism and the US chemical industry. Academy of Management Journal 42, 351–371.
Hoffmann, E.A., 2005. Dispute resolution in a worker cooperative: formal procedures and procedural justice. Law and Society Review 39, 51–82.
Howard-Grenville, J., Nash, J., Coglianese, C., 2008. Constructing the license to operate: internal factors and their influence on corporate environmental decisions. Law and Policy 30, 73–107.
Huber, P.W., Litan, R.E., 1991. The Liability Maze: The Impact of Liability Law on Safety and Innovation. Brookings Institution, Washington, DC.
Hurst, W., 1964. Law and Economic Growth: The Legal History of the Lumber Industry in Wisconsin 1836–1915. Belknap Press, Cambridge, MA.
Kagan, R.A., 2001. Adversarial Legalism the American Way of Law. Harvard University Press, Cambridge, MA.
Kagan, R., Scholz, J., 1984. The criminology of the corporation and regulatory enforcement styles. In: Hawkins, K., Thomas, J. (Eds.), Enforcing Regulation. Kluwer-Nijhoff, Boston, MA, pp. 67–95.
Kapczynski, A., 2008. The access to knowledge mobilization and the new politics of intellectual property. Yale Law Journal 117, 804–885.
Kenworthy, L., Macaulay, S., Rogers, J., 1996. "The more things change…": business litigation and governance in the American automobile industry. Law and Social Inquiry 21, 631–678.
Kidder, R., 1983. Connecting Law and Society. Prentice-Hall, Englewood Cliffs, NJ.
Klausner, M., 1995. Corporations, corporate law, and networks of contracts. Virginia Law Review 81, 757–852.
Laffont, J.-J., Tirole, J., 1991. The politics of government decision-making: a theory of regulatory capture. Quarterly Journal of Economics 106, 1089–1127.
Lande, J., 2000. Getting the faith: why business lawyers and executives believe in mediation. Harvard Negotiation Law Review 5, 137–231.
Lind, E.A., Kulik, C.T., Ambrose, M., de Vera Park, M.V., 1993. Individual and corporate dispute resolution: using procedural fairness as a decision heuristic. Administrative Science Quarterly 38, 224–251.
Lobel, O., 2004. Renew deal: the fall of regulation and the rise of governance in contemporary legal thought. Minnesota Law Review 89, 262–390.
Macaulay, S., 1963. Non-contractual relations in business: a preliminary study. American Sociological Review 28 (1), 55–67.
Macneil, I., 1977. Contracts: adjustment of long-term economic relations under classical, neoclassical, and relational contract law. Northwestern University Law Review 72, 854–902.
Marx, K., 1971a[1842]. Proceedings of the sixth Rhenish parliament: debates on the law on thefts of wood. In: McLellan, D. (Ed.), Early Texts. Blackwell, Oxford, UK, pp. 49–51.
Marx, K., 1971b[1844]. On the Jewish question. In: McLellan, D. (Ed.), Early Texts. Blackwell, Oxford, UK, pp. 85–114.
Masten, S., 1990. A legal basis for the firm. In: Williamson, O., Winter, S. (Eds.), The Nature of the Firm: Origins, Evolution, and Development. Oxford University Press, New York, pp. 196–212.
Mnookin, R., Kornhauser, L., 1979. Bargaining in the shadow of the law: the case of divorce. Yale Law Journal 88, 950–997.

Pedriana, N., Stryker, R., 2004. The strength of a weak agency: enforcement of title VII of the 1964 civil rights act and the expansion of state capacity, 1965–1971. American Journal of Sociology 110, 709–760.
Peltzman, S., 1976. Toward a more general theory of regulation. Journal of Law and Economics 19, 211–240.
Posner, Richard A., 1974. Theories of economic regulation. Bell Journal of Economics 5 (2), 335–358.
Powell, M.J., 1993. Professional innovation: corporate lawyers and private lawmaking. Law and Social Inquiry 18, 423–452.
Priest, G., Klein, B., 1984. The selection of disputes for litigation. Journal of Legal Studies 13, 1–55.
Salamon, L.M., Anheier, H.K., 1997. Defining the Nonprofit Sector: A Cross-national Analysis. Manchester University Press, Manchester, UK.
Samuels, P.D., 14 April 1996. Litigation Law Creates Work for Disclaimer Writers. New York Times section 3, p. 3.
Selznick, P., 1969. Law, Society, and Industrial Justice. Sage, New York.
Short, J.L., Toffel, M.W., 2008. Coerced confessions: self-policing in the shadow of the regulator. Journal of Law, Economics and Organization 24, 45–71.
Sitkin, S.B., Bies, R.J., 1993. The legalistic organization: definitions, dimensions, and dilemmas. Organization Science 4, 345–351.
Smets, M., Morris, T., Greenwood, R., 2012. From practice to field: a multi-level model of practice-driven institutional change. Academy of Management Journal 55 (4), 877–904.
Stigler, G., Spring 1971. The theory of economic regulation. Bell Journal of Economics 2, 3–21.
Stipanowich, T.J., 2004. ADR and the "vanishing trial": the growth and impact of "alternative dispute resolution." Journal of Empirical Legal Studies 1, 843–912.
Stone, C.D., 1975. Where the Law Ends: The Social Control of Corporate Behavior. Harper and Row, New York.
Stryker, R., 2003. Mind the gap: law, institutional analysis and socioeconomics. Socioeconomic Review 1, 335–367.
Suchman, M.C., 2000. Dealmakers and Counselors: law firms as intermediaries in the development of Silicon Valley. In: Kenney, M. (Ed.), Understanding Silicon Valley: The Anatomy of an Entrepreneurial Region. Stanford University Press, Palo Alto, CA, pp. 71–97.
Suchman, M., 1997. On beyond interest: rational, normative and cognitive perspectives in the social scientific study of law. Wisconsin Law Review 475–501.
Suchman, M., 2003. The contract as social artifact. Law and Society Review 37, 91–142.
Sutton, J.R., Dobbin, F., Meyer, J.W., Scott, W.R., 1994. The legalization of the workplace. American Journal of Sociology 99, 944–971.
Ulen, T.S., 1980. The market for regulation: the ICC from 1887 to 1920. American Economic Review 70, 306–310.
Viscusi, W.K., Moore, M.J., 1993. Product liability, research and development, and innovation. Journal of Political Economy 101, 161–184.
Weber, Max, 1978 [1925]. Economy and Society: An Outline of Interpretive Sociology. In: Roth, G., Wittich, C. (Eds.), University of California Press, Berkeley, CA.
Williamson, O., 1975. Markets and Hierarchies. Free Press, New York.

Organizations: Authority and Power

Cynthia Hardy, University of Melbourne, Parkville, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

There are countless ways of defining power in organizations and explaining its relationship to authority. In this article, some of these differences are explored in terms of three broad perspectives found in the literature. First, *managerialist* perspectives differentiate legitimate, formal authority from illegitimate, informal power. Authority is assumed to be rational and functional; while power is typically seen as 'political' and dysfunctional. Second, *critical* approaches directly challenge managerialist approaches. They see organizations, not as rational, neutral, and objective, but as tools designed to fulfill the interests of one group at the expense of another. Power is hidden behind the mask of authority and embedded in the fabric of organizational life, operating to privilege employers over employees and managers over workers. Third, *discursive* views of power examine how communicative practices create organizational realities and identities that both constrain and enable individual action and thought. This article emphasizes the limits of trying to exercise power while, at the same time, helping us to understand how some actors are constructed as authoritative, albeit within the limits of the broader discursive context. Each approach is discussed in depth, and their assumptions and insights compared.

Managerialist Approaches: Power vs Authority

Managerialist research on power and authority is predicated on a rational orientation that seeks to make organizations more efficient and effective (or to understand why they are not). Authority is defined as the recognized, legitimate right to make a decision. In contrast, power is typically employed by those who oppose management, resist change, or pursue self-interest – those who seek to subvert the 'legitimate' power embodied in the organizational hierarchy (see Thomas and Hardy, 2011). Even when these researchers acknowledge that managers do use power, as opposed to authority, they suggest that it is only in self-defense – to pursue organizational goals in the face of resistance. Since authority is considered a 'normal' and 'inevitable' phenomenon that results from the formal design of the organization, it is largely excluded from analysis as research concentrates on power that is exercised outside formal hierarchical structures. Accordingly, managerialist work on power stems from an interest in why and how decisions, which 'ought' to be rational, instead become political as groups use their power to influence decision outcomes in their favor.

The managerialist approach to power in organization and management theory (OMT) is similar in its underlying assumptions to the early studies of community power carried out by American political scientists. In this section, an overview of this article is provided, followed by a discussion of the OMT literature in more detail.

Community Studies of Power

Some of the most influential early work on power was conducted, not in management, but in community studies conducted by political scientists. These researchers (e.g., Dahl, 1957) analyzed the decision-making process in American towns and communities to explore the prevailing power relations. The object was to determine who made these decisions: if the same groups were responsible, the community could be said to be ruled by an élite group as had been proposed by sociologists like C. Wright Mills (1956). If, on the other hand, a variety of groups had an impact on decision outcomes, the community could be described as pluralist. Dahl and other researchers found evidence of the latter, and hypothesized that America as a whole was a pluralist society insofar as, rather than being dominated by a small number of élites, it was populated by diverse groups with equal opportunities to exercise power and influence decisions.

Several assumptions underlay this research. First, researchers assumed that individuals were aware of any grievances they had regarding the political system and, further, they acted upon them by participating in decision making and attempting to influence key decisions. Second, the decision-making arena was open to anyone with a grievance and, therefore, if individuals did not participate, it indicated that they agreed with decision outcomes and the political system more broadly. Third, power was exercised only in decisions where conflict was clearly observable: if conflict was not observed, it was because consensus prevailed. Moreover, such conflict was resolved primarily through political decision-making arenas (see Lukes, 1974).

Lukes (1974) referred to such views of power as 'one dimensional'. He goes on to identify a second dimension of power that emerged as researchers challenged these assumptions. Specifically, Bachrach and Baratz (1962) questioned whether decision-making processes were open to everyone. They argued that nonparticipation did not necessarily reflect genuine satisfaction but, instead, meant that individuals were unable to secure access to decision-making processes. Researchers began to focus on how participation in decision making was constrained – by excluding individuals from the decision-making arena or limiting the participation of those who did secure access to it. Bachrach and Baratz (1962) showed how dominant actors were able to confine the agenda to 'safe' questions and the arena to 'safe' participants, while more difficult matters were dealt with outside formal decision making. They called this process 'nondecision making' because it allowed the more powerful actors to determine outcomes from behind the scenes.

The work on nondecision making broadened the research agenda, but it continued to operate on similar assumptions as the one-dimensional view of power. Specifically, it assumed that some form of conflict was necessary for power to be exercised, i.e., groups had to have some form of grievance, and desire different decision outcomes if their political opponents were to initiate nondecision making to try and ensure that this did not happen (Lukes, 1974). Thus, from a theoretical perspective, the first- and second-dimensional views of power converge: both focus on the exercise of power in or around the decision-making arena as part of a deliberate strategy to achieve intended outcomes in the face of conflict. Power mobilized through nondecision making may be less visible than decision-making power – since opponents may be sidelined rather than directly confronted – but it is still premised on the existence of conflict and opposition (Hardy, 1985).

Managerialist Studies of Power in OMT

Although conducted separately and with a different rationale, managerialist studies of power in OMT adopted a similar view of power as the literature described above insofar as researchers explored how individuals in organizations exercised power as part of a deliberate strategy to achieve intended outcomes in the face of opposition. In particular, researchers focused on uncovering the sources of power that allowed actors who did not possess legitimate authority to influence decision outcomes. In this way, researchers began to differentiate between formal, legitimate power – authority – and informal, illegitimate power.

Crozier (1964) studied how maintenance workers in a French state-owned tobacco monopoly had a high degree of power in relation to their position in the hierarchy because they controlled the uncertainty associated with breakdowns of equipment. This and similar work contributed to the strategic contingencies theory of intraorganizational power (Hickson et al., 1971), which argued that subunits within an organization were powerful when they were able to cope with greater amounts of uncertainty, were central, and were not easily substitutable – in other words, when other subunits depended upon them. Similar assumptions underpinned the resource dependency view of power.

> This theory states that power resides among a set of interdependent subunits or organizations that exchange resources with each other. The value of the resources that a subunit/organization controls and the extent to which those resources can be obtained elsewhere (i.e., the subunit/organization's substitutability in the exchange relationship) determine the terms of exchange, and thus the power in relation to other subunits/organizations. If the value that a subunit/organization provides can be replaced (i.e., substituted), then there is little dependence on that subunit/organization, which consequently has little power in that social relationship.
> (Magee et al., 2008: p. 362)

Researchers began to identify the resources that constituted the bases or sources of power, such as information, expertise, credibility, prestige, access to higher echelon members, and the control of money, rewards, and sanctions (Pettigrew, 1973). They also questioned whether simply possessing scarce resources was enough to confer power, arguing that actors also had to be aware of their contextual pertinence and sufficiently skilled to deploy them. This process of mobilizing power was typically referred to as 'politics', which further reinforced the idea that the use of power was illegitimate, dysfunctional, and self-serving.

Researchers using this approach to power, concentrated on the link between power and the control of scarce resources – or dependencies – that helps actors to influence decision-making processes in the face of opposition. Those individuals with the greatest access to power sources and the greatest skill in putting them to use are the most likely to prevail in decision making. Decision outcomes produced in this way are described as political, rather than rational, and are typically assumed to be at odds with the interests and objectives of those in formal authority. Power is only used in the first place because of the existence of conflict.

> [It] is clear that political activity is activity which is undertaken to overcome some resistance or opposition. Without opposition or contest within the organization, there is neither the need nor the expectation that one would observe political activity.
> (Pfeffer, 1981: p. 7)

In this way, the use of power is inextricably linked to the situations of overt conflict associated with struggles among actors seeking to protect vested interests.

This managerialist work on power thus converges theoretically with the study of community power. First, both community and OMT researchers challenged the existing models of decision making. Studies of community power refuted the élitist views of decision making suggested by some sociologists to propose a pluralist perspective; while management studies of power challenged the existence of the rational model in organizations by saying that decision making was, in fact, political (although, whereas political scientists tended to promote pluralism – and the attendant politics – over élitism, management researchers clung to rationality as the desirable mode of functioning). Second, both bodies of literature adopted a behavioral approach that focused on the overt exercise of power in the decision-making arena, and both argued that decision outcomes were influenced by actors' access to and expertise in using power (although OMT researchers seemed uninterested in the concept of nondecision making, which did not gain much traction in this literature). Third, both bodies of work employed similar definitions of power, i.e., the ability to get others to do what you want them to, if necessary against their will (Weber, 1947) or to do something they otherwise would not do (Dahl, 1957).

While managerialist approaches in OMT have produced insight into some aspects of power, existing organizational arrangements – particularly in the guise of legitimate authority – are largely excluded from analysis. Power represents dysfunctional interruptions in the far more desirable flow of rational decisions by self-interested actors who engage in its political mobilization, a term whose negative connotations helped to reinforce the view that power is illegitimate and dysfunctional (except, perhaps, when used by managers in defense of organizational goals). As power and authority were pulled further apart, authority was increasingly taken for

granted; with researchers rarely stopping to consider how it got there or whose interests it serves. Power was what counted – and what fascinated – researchers (Hardy and Clegg, 2013).

> Power, related to one's control over valued resources, transforms individual psychology such that the powerful think and act in ways that lead to the retention and acquisition of power.
> (Magee et al., 2008: p. 351)

As the managerialist literature has grown, it has increasingly conceptualized power in highly individualized terms: as individual control over valued resources that are asymmetrically distributed in social relations. Confidence in the predictive ability of power mechanisms such as hierarchical position, perceptions of competence, self-enhancement, hedonic sense-making, and the desire to win has led researchers like Jeffrey Pfeffer (2013) to claim their universality as determinants of organizational outcomes.

> Not only does power induce different motivations in individuals, but the powerful soon learn that they can get away with things that others cannot.
> (Pfeffer, 2013: p. 277)

In contrast, Willmott (2013a: p. 281) argues, "power and influence reach well beyond the surface mechanisms identified and naturalized by Professor Pfeffer." He argues that managerialist work ignores and obscures the ways in which power is institutionalized and reproduced – to the extent that this approach can be said to constitute *uncritical* management studies.

> [Power] tends to be conceived shallowly as an attribute of individuals and groups who possess it, and to be manifest in zero-sum games involving acquisition, exercise, and potential loss of power [which] reflects and reinforces a highly individualized and individualizing representation of an episodic manifestation of power that largely displaces consideration of its relational formation and systemic operation.
> (Willmott, 2013a: p. 285)

So, let us now turn to *critical* management studies to see how work in this area has treated power and authority.

Critical Approaches: Power, Authority, and Domination

Research in critical management studies, a term used here to refer to a diverse body of research (see Willmott, 2013b), adopts a very different approach to power and authority than the work described above. Critical management researchers argue that power is embedded in formal organizational structures and that decision-making processes, which may appear to be neutral, rational, or logical are, in fact, designed to serve managerial interests. Hence, hierarchy is not authority, but power and domination; while actions taken to challenge it constitute resistance to domination, rather than the illegitimate use of power. Critical management researchers examine how power is embedded in societal and organizational structures, ideologies, and narratives. They assert that much of what appears to be apolitical is, in fact, both a product and a medium of power. This view has helped to expose how power, concealed in the legitimate trappings of structures, rules, class mechanisms, and cultures, produces quiescence and prevents conflict from arising. In the following section, Lukes' (1974) third dimension of power is examined and the adoption of a similar radical or critical view in OMT is explored.

The Third Dimension of Power

Having identified the first and second dimensions of power described above, Lukes (1974: p. 24) presented a third dimension of power. He argued that power can be used to shape perceptions, cognitions, and preferences to the extent that individuals "accept their role in the existing order of things, either because they can see or imagine no alternative to it, or because they view it as natural and unchangeable, or because they value it as divinely ordained and beneficial." Consequently, the study of power cannot be confined to observable conflict, the outcomes of decisions, or even suppressed issues. It must also consider the question of political *quiescence*: why grievances do not exist; why demands are not made; and why conflict does not arise in situations of inequality? In other words, researchers were interested in how power is used to maintain the dominance of élite groups by reducing the inclination of subordinates to employ their discretionary power.

Lukes' interest in the societal and class mechanisms, which perpetuate the status quo, relates to the ability of dominant classes to define reality to support and justify their material domination, thus preventing challenges to their position, and avoiding the necessity of using more coercive, visible forms of power. According to this view, stable organizing power is based on the third dimension – the ability to avoid conflict and opposition by making power appear legitimate, thereby rendering its use by dominant actors less visible.

Critical Views in OMT

The fact that organizations tend to serve managerial interests had not been lost on all organizational researchers. Nor had the political advantages of legitimacy gone completely unnoticed as Pettigrew (1979) showed how interest groups legitimized their demands to prevent conflict through symbolic activity. Critical management researchers went further to examine the way in which power is embedded at a 'deeper' level in organizational structures, cultures, and ideologies, and as political actors try to section off spheres of influence where domination is perceived as legitimate and thus unchallenged (e.g., Clegg, 1975). This use of power is largely invisible and designed to avoid conflict altogether rather than trying defeat it once it has arisen (Hardy, 1985).

Another critical perspective on power was provided by labor process theorists (e.g., Burawoy, 1979) who focused on the dynamics of power and resistance in relation to phenomena such as gender, technology, ethnicity, managerial work, and other aspects of the structuration of work and its organizational context. More recently, such work has critically examined the empowerment and teamwork practices that characterize

contemporary management practice. Far from providing employees with more power, researchers argue that these practices embody subtle, invisible controls. The terminology of empowerment helps to avoid conflict and resistance by emphasizing consensus and cooperation, while managerial control is reinforced through the language of the team effort and peer pressure. A critical perspective thus sees empowerment as an exercise in the management of meaning, designed to enhance the legitimacy of organizational goals and to influence behavior unobtrusively.

In summary, critical approaches to power are substantively different in their assumptions compared to managerialist views of power. Power is domination; authority is one form of domination; and resistance is a repudiation of that domination. Power and resistance are differentiated in order to indicate "qualitatively different contributions to the outcome of power relations made by those who exercise power over others, on the one hand, and those subject to that power, on the other" (Barbalet, 1985: p. 545). According to this approach, power is used not simply to defeat conflict, but to prevent it from arising in the first place. Critical management research focuses on this subtle, repressive use of power, attempting to reveal the hidden aspects of domination, helping to facilitate resistance, and enabling groups to free themselves from the effects of power.

Discursive Power: Reconstructing Authority

Critical research has itself come in for criticism, particularly for the idea that revealing and resisting the hidden way in which power is embedded in the fabric of organizations can produce a situation 'free' of power, and that critical theorists have a privileged access to reality that permits them to stand outside the power relations that they are exploring. As a result, many researchers – including many already inside the critical camp – turned to the work of poststructuralist researchers to adopt a discursive view of power. In this section, work that links power to discourse is explored, particularly the work of Michel Foucault (e.g., 1980).

Foucault: The End of Authority?

The work of Foucault challenges the very idea of authority by downplaying the relevance of 'sovereign' power in contemporary society. Gone is the idea of a supreme authority that determines organizational outcomes. Power is no longer a convenient, manageable resource controlled by autonomous individuals, whether by exploiting or circumventing hierarchy. Instead, a network of power relations – or discourses – captures the advantaged and disadvantaged alike.

> [I]n any society there are manifold relations of power which permeate, characterize and constitute the social body, and these relations of power cannot themselves be established, consolidated nor implemented without the production, accumulation, circulation and functioning of a discourse.
> (Foucault, 1980: p. 93)

Discourses are collections of interrelated texts and practices that "systematically form the object of which they speak" (Foucault, 1972: p. 49). They produce particular meanings, as a result of which particular kinds of objects and subjects become 'known', through which and upon which particular relations of power are realized. In influencing the way in which a topic or issue can be talked about – or acted upon – power circulates through discourse, constraining and enabling what individuals can think, say, and do.

Actors may try to manipulate resources, structures, or meanings with the intention of achieving particular outcomes but, according to Foucault, no matter how hard they pull on the 'strings' of power, they are unlikely to attain their desired ends.

> People know what they do; they frequently know why they do what they do; but what they don't know is what what they do does.
> (Foucault quoted in Dreyfus and Rabinow, 1982: p. 187)

These limits to power and authority can be found in the complexity of power relations: the way in which they are productive, and not only repressive; their link to knowledge; and their pervasiveness.

This network of power produces the subject (Knights and Morgan, 1991). According to this view, the individual is not a stable constellation of essential characteristics, but is associated with multiple, socially constituted categories. Identities are salient only insofar as they are socially recognized. Thus the subject is not a 'given' but produced historically through power–knowledge relations. This effect of power is not necessarily alienating or repressive. It is also productive and positive insofar as it transforms individuals "into subjects who secure their sense of what it is to be worthy and competent human beings" (Knights and Morgan, 1991: p. 269). Foucault also challenges the idea that one can escape the regimes of power through knowledge, arguing that knowledge can never be separated from the effects of power. Knowledge does not strip away the effects of power to reveal the 'truth'; all it accomplishes is the embodiment of new and different forms and relations of power. What is accepted as 'true' are those mechanisms that "enable one to distinguish true and false statements … the techniques and procedures accorded value in the acquisition of truth; [and] the status of those who are charged with saying what counts as true" (Foucault, 1980: p. 131). Finally, the embedding of power relations in every aspect of one's life – including identity – makes them difficult to repudiate. They are experienced as reality, making alternatives difficult to conceive of, let alone enact. Resistance can usually only be conceived of in terms of the prevailing power relations and is subsequently at a risk of being 'colonized' by them; it is as likely to reinforce existing power relations as to overthrow them (Clegg, 1989). In this way, the emancipation for which critical approaches strive becomes illusory: power relations may change as a result of resistance but they do not disappear.

Foucault's work thus posits a very different conceptualization of power to both managerialist and critical views. Not only is there no autonomous subject waiting – or even willing – to be liberated through critical awareness, it is also erroneous to assume that emancipation can be achieved by the knowledge gained through critical reflection. Even if resistance is

successful, the resulting situation can never be free of power – it simply represents a *different* power regime. Moreover, these power relations affect all actors – capital and labor, employers and employees, researchers and research subjects. No one has the privilege of standing outside power.

> Power must be analyzed as something which circulates, or rather something which only functions in the form of a chain. It is never localized here or there, never in anybody's hands, never appropriated as a commodity or piece of wealth. Power is employed and exercised through a net-like organization…[Individuals] are always in the position of simultaneously undergoing and exercising this power.
> (Foucault, 1980: p. 98)

This power is impossible to escape. It encompasses, surrounds, and infiltrates in ways of which individuals are often unaware. It is productive yet constraining; it makes them what they are and, yet, makes it difficult to become anything else: authority becomes an illusion and politics, an indeterminate enterprise.

Organizational Discourse: Reconstructing Authority

Foucault's work focuses attention on how the institutionalization of practices and structures embody sets of power relations that are deeply embedded in the taken-for-granted rules and practices that constrain action (Clegg, 1989). In linking power and discourse, he shows how actors are locked into positions by way of the linguistic practices available to them to make sense of a situation. This form of power lies neither in the possession of individuals nor in the dependency relationships among them. It is embedded in the fabric of organizational life, reproduced by the day-to-day communicative practices and struggles of its members. Organizational reality is characterized less by the domination of one group over another, but by complex discursive practices that "define what it means to be an organizational member and … allow for the privileging of managerial interests over others" (Deetz and Mumby, 1990: pp. 32, 39).

Some writers have criticized Foucault for a supposedly agentless and overly pessimistic conceptualization of power, partly stemming from the belief that in linking power to discourse, Foucault was reducing it to language. However, as Foucault (1980: p. 57) himself has pointed out, discourse is not composed only of linguistic components, it is also embedded in material practices: "nothing is more material, physical, corporeal than the exercise of power." Accordingly, his work allows researchers to ask more far-reaching questions about power (Hardy and Thomas, 2014). Rather than simply asking who has power, researchers can ask how does power circulate through discourse, with a view to problematizing its institutionalization and reproduction as well as identifying its limits. Rather than seeing resistance only as an adversarial response to authority, researchers can ask how power and resistance operate together as part of the complex web of relations linking actors in everyday organizational life. Rather than asking how is power exercised repressively 'over' others, researchers can ask what does power produce and how does it shape our everyday life.

> [Power] produces reality; it produces domains of objects and rituals of truth. The individual and the knowledge that may be gained of him belong to this production.
> (Foucault, 1979: p. 194)

In this way, researchers can focus on how a particular discourse positions actors: which actors secure rights to speak, and which actors do not. They can examine the way in which actors draw upon discourses, and the language, genres, and linguistic strategies they use. They can explore the meaning attached to different categories of object in the discourse, the procedures, and practices that are invoked, and the way in which particular actors stand to gain or lose from the deployment of discourse. This article, implicitly or explicitly, resurrects the notion of authority by acknowledging that certain actors and institutions are constructed as authorities and authoritative.

Conclusions

This article has examined a number of approaches and explored their insights into the relationship between power and authority. Managerialist views of power differentiate legitimate, formal authority from illegitimate, informal power. The former is assumed to be rational and functional, the latter is typically seen as political and dysfunctional. This view of power is not contained within a particular body of literature on power; rather it permeates much of the mainstream work in OMT that sees organizations as rational decision-making systems, albeit flawed ones. Critical approaches directly challenge this managerialist work. They see organizations, not as rational, neutral, and objective but as managerial tools designed to fulfill the interests of one group at the expense of another. Power is embedded in the fabric of organizational life, hidden behind the mask of authority, and designed for domination. Discursive views of power wrestle with the notion of authority. They dispel the idea that either authority or power can be a pliable means of domination but, at the same time, show how authority emerges from contested, productive, and transgressive power–knowledge relations.

See also: Critical Theory; Dahl, Robert A (1915–2014); Foucault, Michel (1926–84); Mills, Charles Wright (1916–62); Organizational Decision Making; Postmodernism in Sociology; Power in Society; Power; Social Constructionism.

Bibliography

Bachrach, P., Baratz, M.S., 1962. The two faces of power. American Political Science Review 56, 947–952.
Barbalet, J.M., 1985. Power and resistance. British Journal of Sociology 36 (4), 531–548.
Burawoy, M., 1979. Manufacturing Consent. Chicago University Press, Chicago.
Clegg, S.R., 1975. Power, Rule and Domination. Routledge, London.
Clegg, S.R., 1989. Frameworks of Power. Sage, London.
Crozier, M., 1964. The Bureaucratic Phenomenon. University of Chicago Press, Chicago.

Dahl, R., 1957. The concept of power. Behavioral Science 20, 201–215.
Deetz, S., Mumby, D., 1990. Power, discourse and the workplace: reclaiming the critical traditions. Communication Yearbook 13, 18–47.
Dreyfus, H.L., Rabinow, P., 1982. Michel Foucault: Beyond Structuralism and Hermeneutics. Harvester Press, Brighton, UK.
Foucault, M., 1972. The Archaeology of Knowledge. Harper & Row, New York.
Foucault, M., 1979. Discipline and Punish: The Birth of the Prison. Penguin, Harmondsworth.
Foucault, M., 1980. Power/Knowledge: Selected Interviews and Other Writings 1972–1977. Harvester Press, Brighton, UK.
Hardy, C., 1985. The nature of unobtrusive power. Journal of Management Studies 22 (4), 384–399.
Hardy, C., 1994. Managing Strategic Action: Mobilizing Change. Sage, London.
Hardy, C., Clegg, S., 2013. Some dare call it power. In: Clegg, S.R., Hardy, C., Lawrence, T.B., Nord, W. (Eds.), Handbook of Organization Studies. Sage, London.
Hardy, C., Thomas, R., 2014. Strategy, discourse and practice: the intensification of power. Journal of Management Studies 51 (2), 320–348.
Hickson, D.J., Hinings, C.R.A., Lee, C.A., Schneck, R.E., Pennings, J.M., 1971. A strategic contingencies theory of intraorganizational power. Administrative Science Quarterly 16 (2), 216–229.
Knights, D., Morgan, G., 1991. Corporate strategy, organizations, and subjectivity. Organizational Studies 12 (3), 251–273.
Lukes, S., 1974. Power: A Radical View. Macmillan, London.
Magee, J.C., Galinsky, A.D., Wagner, R.F., 2008. Social hierarchy: the self-reinforcing nature of power and status. Annals of the Academy of Management 2 (1), 351–398.
Pettigrew, A.M., 1973. The Politics of Organizational Decision Making. Tavistock, London.
Pettigrew, A.M., 1979. On studying organizational cultures. Administrative Science Quarterly 24, 570–581.
Pfeffer, J., 1981. Power in Organizations. Pitman, Marshfield, MA.
Pfeffer, J., 2013. You're still the same: why theories of power hold over time and across contexts. Academy of Management Perspectives 27 (4), 269–280.
Thomas, R., Hardy, C., 2011. Reframing resistance to organizational change. Scandinavian Journal of Management 27 (3), 322–331.
Thomas, R., Sargent, L.D., Hardy, C., 2011. Managing organizational change: negotiating meaning and power-resistance relations. Organization Science 22 (1), 22–41.
Weber, M., 1947. The Theory of Social and Economic Organization. Oxford University Press, New York.
Willmott, H., 2013a. Reflections on the darker side of conventional power analytics. Academy of Management Perspectives 27 (4), 281–286.
Willmott, H., 2013b. OT and power: the significance of value-orientations and a plea for pluralism. Tamara – Journal for Critical Organization Inquiry 11 (2), 53–66.
Wright Mills, C., 1956. The Power Elite. Oxford University Press, New York.

Organizations: Negotiated Orders

Tony J Watson, University of Nottingham, Nottingham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The negotiated order of an organization is the pattern of activities that has emerged over time as an outcome of the interplay of the variety of interests, understandings, reactions, and initiatives of the individuals and groups involved in the organization. To examine negotiated orders in any given organization is to turn away from the more traditional way of looking at organizations that give primacy of attention to the pattern or ordering of activities chosen (or 'designed') by those officially in charge of the organization. The concept started its life in the study of the hospital and was originally used within a broadly interactionist frame of reference. However, it has close affinities with a series of other developments in organization theory, and it provides a way of dealing with the issues that were once considered in terms of the theoretically inadequate distinction between formal and informal organizations. This is especially the case if the distinction between official and unofficial aspects of organizing is incorporated into a negotiated order perspective. More and more, the negotiated order notion is being linked to that of the social institution, and this makes it an invaluable conceptual resource for the type of multilevel analysis to which researchers are increasingly turning.

Origins of the Concept in a Hospital Study

The concept of 'negotiated order' first appeared in a study of a psychiatric hospital (Strauss et al., 1963). The study focuses on the interactions of the variety of individuals and groups who are involved in its activities. These include the doctors, nurses, patients, patients' families, social workers, nutritionists, administrative workers, and so on. Although the rules, policies, official procedures, and hierarchical statuses of the hospital play a part in the patterns that emerge from the interactions that take place, the patterns are shown to be emergent. They emerge from the processes whereby different parties make use of rules, procedures, and information and negotiate between them what actually happens. These negotiations are characterized in a variety of ways, such as agreements, contracts, understandings, pacts, and so on. But two things are emphasized throughout. First, it is stressed that these negotiated outcomes are always temporary and have to be made and remade over time. A head nurse, for example, may gain agreement to move a patient to another ward, after it is agreed that the first ward has 'tried for a little longer' with that patient. In the light of how that patient fares, however, the patient may or may not subsequently be moved. The 'deal' will have been remade, implicitly or explicitly. Second, it is stressed that there is a patterning to what might, at first sight, appear to be a series of unique encounters. The deal between the head nurse and the doctor over a particular patient in specific circumstances, for example, takes place in the light of how interactions between members of these two occupational groups generally occur across the hospital.

Out of the ongoing multiplicity of interactions and negotiations, a pattern of hospital life, a negotiated order, emerges. Order is thus not imposed on social actors by the hospital but the 'order,' which, in effect, is the hospital is socially constructed by the interactions of the various individuals and groups. And each of these brings to the negotiation of order their own interests, understandings, loyalties, knowledge, occupational standing, and social origins. All these elements themselves are liable to change as they are taken into processes of reality negotiation.

Negotiated Order: Processes, Rules, and Rules

When Strauss et al. introduced the concept of negotiated order in their hospital study, they did not claim a generalized relevance for the concept. However, the leading researcher, Strauss (1978), later showed how it could be used more generally to understand organizational processes, including ones in industrial organizations. Its insights have subsequently been incorporated into a broad stream of thinking about organizations that focuses on *processual* aspects of organizational life. The approach emerged from the symbolic interactionist tradition of social psychology, a perspective that stresses the two-way relationship between human beings and their social context. People construct their realities through processes of interaction with others, and their identities emerge as part of this process of interacting and reality making. Such a theoretical approach, with its strongly antideterminist orientation, inevitably leads to a questioning of views of organizations that treat the formal arrangement of roles, rules, and procedures as the major determinants of the activities of organization participants. Indeed, Strauss et al. explicitly identify their study with such a critical stance on orthodoxy: "Students of formal organization tend to underplay the processes of internal change as well as overestimate the more stable features of organizations – including its rules and its hierarchical statuses" (1963, p. 147).

The emphasis of the Strauss et al. study on interactive and processual aspects of organizations was not peculiar to this group of social scientists and their study of a mental hospital. In particular, Goffman's similarly interactionist classic study, *Asylums* (1961), has clear parallels with the Strauss et al. study (which was partly influenced by early accounts of Goffman's research). And the same argument can be made with regard to Dalton's participant observation study of managerial work in an industrial setting. In fact, it was this study that Strauss (1978) used to illustrate the position he later came to take: that the negotiated order concept had a wider currency than had earlier been implied. He notes that Dalton's detailing of what Strauss calls "sub-processes of negotiation" is especially rich. Many researchers in managerial behavior would probably agree that few studies have surpassed Dalton's analysis of how

managers engage in; as Strauss puts it, "trading off, the paying off of accumulated obligations, the covert making of deals and other kinds of secret bargains, the use of additional covering negotiations for keeping hidden the results of previous covert arrangements, the bypassing of negotiations, and the mediating of negotiations; also, a very complex balancing of accumulated favors and obligations, along with the juggling of commitments within the negotiations itself" (1978, p. 139).

One particular area where Strauss et al. (1963) follow Dalton (1959) is in attending to the frequency with which formal rules are broken, stretched, or ignored, not only to suit sectional purposes but also to 'get things done.' A similar insight is central to ethnomethodological writing on organizations. This contributes to an understanding of how 'orders' are produced in organizations by focusing closely on how individuals, in their organizational interactions, 'work,' especially through talk, to present what they are doing as 'sensible' because it accords with rules (Bittner, 1965; Silverman and Jones, 1976). The rule (or even the organization itself) is not something preexisting or existing outside ongoing organizational interactions. It is, rather, a resource that people draw upon to legitimize their actions and further whatever projects they are pursuing in the work context.

This perspective, combining ethnomethodological insights with the notion of negotiated order, can be illustrated with a small case study. A van driver was overheard explaining to a factory security officer who had stopped him why he and his colleagues regularly took an illegal shortcut across a large industrial plant. He said, "The company says that all employees should use their initiative, always thinking of the customer first. By driving this way we are getting the goods out to the customer as quickly as we can. The other security men know this, and don't stop us." Ethnomethodological analysis would point to the way this driver uses a notion of the organization (the company) as one resource and utilizes a rule (use initiative and put customers first) as another. This is the negotiating of order in practice. Unofficial negotiations between van drivers and security officers that allowed driving along an illegal route through the plant to become part of the order of the organization had occurred. It suited both parties (a quiet life for the security people and a saving of time and effort for the drivers), and it got the job done. However, we must remember that negotiated order thinking sees such deals as continually liable to change. In fact, such a change was happening on the day these events were observed. A new security man had started work. And he believed that the one-way rule (which made the short-cut illegal) was necessary for protecting the safety of employees working in the plant. He, too, utilized a notion of 'the organization' and drew on a rule when he said, "You should realize that the company puts employee safety before even customers. Don't drive down here again." He got his way – for the time being.

Ambiguity, Interpretation, and Enactment

It might be argued that the above case is not looking at rules in a proper sense and that the individuals in the case study were talking simply in terms of inferred rules. But organizational thinking that follows a broad negotiated order approach would accept that all rules, or any other organizational procedure for that matter, are ambiguous. An important work in this tradition by March and Olsen (1976) argues that all organizational choices occur in a context where there is ambiguity about everything, from organizational goals and rules, through to the technologies used and people's knowledge of the past. Weick (1979), in another important contribution to this broad style of thinking, similarly points to equivocality in every aspect of the work of organizing. Members of organizations create the organization (or, rather, achieve organizing) through continuous processes of interpretation and action based on interpretations: They enact both the organization and the environment to which it relates.

Out of a range of partly linked theoretical origins – symbolic interactionism, phenomenology, and Weberian interpretative sociology – a general style of organizational analysis has developed which, in one way or another, incorporates the negotiated order insight. It is a processual style of analysis. It treats organizational patterns as emergent rather than as the creations of their managerial designers. And such an emphasis has been reinforced by the influence on organization theorizing of postmodern thinkers. Chia, for example, adopting an 'ontology of becoming,' encourages the analysis of the "discursive micro-practices which generate [the] stabilized effects" which constitute 'organizing' (Chia, 1996, p. 149).

The Negotiation of Order: Building upon and Going beyond the Classic Formal/Informal Organization Distinction

The concept of negotiated order provides a valuable analytical handle for dealing with the interactive and processual aspects of organizations, and one of the most important ways in which the broad idea of organizations as negotiated orders has influenced the study of organizations has been in its displacing of a powerful earlier distinction between formal and informal organization. This distinction was once widely used in organization theory, industrial sociology, and managerial writing to draw attention to aspects of patterned activity in work organizations that are not part of the formally or officially designed rules, relationships, and procedures that constitute the formal dimension of the organization. Informal organization is seen variously as something that works, or might be made to work, in conjunction with the formal organization, something that can exist with a degree of independence from it or something that can work against the formal aspects of the organization – especially when it takes the form of deviant behavior that resists or defies managerial authority. The formal part of an organization was identified in *Management and the Worker* (1939), a book that is famous as the key account by Roethlisberger and Dickson of studies carried out by academics from Harvard University and members of the Western Electrical Company at that company's Hawthorne plant. 'In short,' say Roethlisberger and Dickson, "the patterns of human interactions, as defined by the systems, rules, policies, and regulations of the company, constitute the formal organization." It is commonly assumed that the "organization of a company corresponds to a blueprint plan or organization chart" (Roethlisberger and Dickson, 1939, p. 559). 'Actually,' they say, "it never does." The study revealed

what Roethlisberger later referred to as "the emergent patterns of behavior that evolved within this formal organization" (Roethlisberger, 1968, p. 262). This informal organization was made up of "the practices, values, norms, beliefs, unofficial rules, as well as the complex network of social relations, membership patterns and centers of influence and communication that developed within and between the constituent groups of the organization under the formal arrangements but that were not specified by them" (Roethlisberger, 1968, p. 262).

These insights from the Harvard-based factory studies were taken up and developed at the managerial level of organization by Barnard in another classic Harvard study *The Function of the Executive* (1938). Barnard noted the paradox that, although executives tend to speak of their work primarily with reference to the "formal scheme of work tasks – and, indeed, deny the importance of the informal side – they nevertheless intuitively recognize its vital significance." "Learning the organizational ropes" in most organizations, he points out, "is chiefly learning who's who, what's what, why's why, of its informal society" (Barnard, 1938, p. 121). Consequently, it is on this aspect of the organization that executives need to focus their attention.

It is impossible to overestimate the importance of these insights into the development of organization and management studies, and it is clear that the empirical knowledge produced by these enormously important writers is vital to any consideration of the negotiation of order in organizations. Also, as ethnographic and, especially, participant observation studies of organizational activities began to appear, the more attention was drawn to facets of organization life that could not be accounted for in terms of formal administrative systems. Roy's (1960) participation observation of shopfloor work drew attention to how his colleagues used games and rituals to shape their day and overcome the monotony that the straightforward compliance with formal procedures would create. And later research by Burawoy (1979) in the same plant showed how workplace practices which might, at first sight, seem to challenge the dominant interests that the formal arrangements of the factory were designed to serve, in practice accommodated workers into the dominant patterns of interest. At the managerial level, Dalton's (1959) participative observation study showed the considerable extent to which senior organizational members organized themselves and behaved in ways that would be quite unexpected from a reading of their formal roles and duties. And Mangham and Pye's (1991) interview-based study of senior executives shows how close much of the thinking of these individuals was to Barnard's conception of the executive role and its focus on informal activities. This includes their relationships with other senior managers: "Many stressed that informal organizing, the structuring of the top team, was as important as, if not more than, the articulation of a formal organization" (1991, p. 8).

Given all of this, one is bound to ask why so little use is made of the formal/informal distinction in contemporary literature. This question can be answered by pointing to the movement in organization theory away from systems thinking and toward more processual thinking, the spirit of which is better captured by the idea of the negotiation of order than by the notion of there existing formal and informal elements of organizational systems. Roethlisberger was explicit, for example, in saying that the "notion of a natural social system" should be understood as 'underlying' the way of thinking which distinguishes between formal and informal organization (Roethlisberger, 1968, p. 261). And theorists, in part influenced by the way in which ethnographic work has tended to demonstrate the greater emphasize on matters of process rather than the existence of systems, have increasingly found this notion unhelpful. Further, systems thinking is seen as representing work enterprises in too unitary a manner, as failing to recognize the multiplicity of meanings and interests existing within the organization and as underplaying the relationships between organizations and wider societal patterns and processes. It can also be argued that with the managers of organizations now setting out to manipulate corporate cultures, publish mission statements, facilitate team working, and empower workers, there is such a blurring of the formal and the formal aspects of managerial initiative that the distinction is difficult to use when setting out to analyze the organizations in which this occurs. Managers have incorporated into the formal aspect of their organizations' activities that were once regarded as informal. An alternative distinction to the formal/informal organization one is that between official and unofficial aspects of organizations. This pair of concepts enables fully processual and negotiation-of-order analyses to be made as the researcher examines the interplay between "the managerial aspirations expressed in official management statements, policies, and claims and the observable patterns of belief and behavior that prevail across organizations in practice" (Watson, 2012, p. 119).

The Continuing and Future Power of the Negotiation of Order Perspective

The original way in which the negotiated order concept was used was criticized, even by broadly sympathetic critics like Day and Day (1977), for insufficiently grounding its analysis of interactional processes in their wider political, structural, and historical context. Strauss's later work responded positively to some of these points and, in particular, demonstrated how the approach could handle such allegedly neglected matters as conflicts that are "endemic, or essential to relations between or among negotiating parties" (1978,p. 290). Although, as has been demonstrated above, the negotiated order insight runs through a great deal of organization analysis, its terminology has not been used in an explicit manner as widely as might be expected. This may reflect reservations about the extent to which the analysis of negotiated orders within organizations can add to organizational theorizing generally if it focuses primarily on interactional patterns within those organizations. Perhaps the perspective, at least its explicit adoption, has had less impact than it deserves because it has been deemed inadequate as a general perspective for the study of organizations and their management. This problem can be overcome if the significant insights that it offers are incorporated into broader perspectives that link activities at the intraorganizational interactional level to wider societal patterns and processes.

Harking back to the origins of the negotiated order concept in the study of health provision, Sambrook (2001) examines the emergent patterns of human resource development in the British National Health Service by setting personal interests and

actions of individuals in hospitals to both the level of organizational strategy and state power, showing how roles and practices are "negotiated to meet both personal and organization strategies at the same time, often, as needing to comply with government legislation … meet the demands of the health service … comply with requirements of government bodies" (2001, p. 190). Perhaps surprisingly, the notion of institution is not used in this study, but institutional theory is brought together with the negotiated order perspective to cover similar issues of strategic innovation in public organizations by Grenier to show how actors negotiate and elaborate local and situated agreements to strategically implement new organizational arrangements in institutional contexts introduced by public policies (Grenier, 2011, p. 8). This is a valuable move in that it helps overcome the criticism frequently made of neo-institutional theory that it pays insufficient attention to the part played by action and interaction in processes of institutional change. This is noted by Bechly who shows how combining institutional thinking with the negotiated order concept allows our research fully to acknowledge that people do not directly respond to social structures, but rather to the situations they face and their interpretations of them (Bechly, 2011, p. 1157).

Watson and Watson (2012) echo Bechly's call for multilevel approaches to theorizing. In their study of the interlinking of 'real' beer and public-house-related narratives at the individual level, the organizational/family/group level/community level, and the societal and social change level, they produce an extended version of the concept of negotiated order. The concept of identity-work provides the conceptual focus at the individual level; the notion of the social construction of reality provides the conceptual focus at the societal level; and the extended negotiation of order concept gives such a focus at the middle level. The negotiation of order here is "the pattern of activities and understandings which emerge over time as an outcome of the interplay of the various interests and initiatives of the individuals and groups involved in a group, community, family or organization" (2012, p. 687). Such an extended concept of the negotiation of order, or something like it, has considerable potential for enabling us to appreciate in our research that one cannot understand meanings, actions, and processes within an organization and the way in which order and patterning emerge without observing how the negotiations that occur between all the social actors involved with an organization are intimately related to the negotiated orders of other human grouping in which they are involved, as well as to the structures, cultures, discourses, and narratives that operate at the level of individual lives and at the level of societal structures and processes.

See also: Corporate Culture; Ethnography; Ethnomethodology, General; Institutions; Interactionism, Symbolic; Participant Observation; Symbolic Interaction: Methodology; Work and Industry, Sociology of.

Bibliography

Barnard, C., 1938. The Functions of the Executive. Harvard University Press.
Bechly, B.A., 2011. Making organizational theory work: institutions, occupations, and negotiated orders. Organization Science 22 (5), 1157–1167.
Bittner, E., 1965. The concept of organization. Social Research 32, 239–255.
Burawoy, M., 1979. Manufacturing Consent. University of Chicago Press, Chicago.
Chia, R., 1996. Organizational Analysis as Deconstructive Practice. de Gruyter, Berlin.
Dalton, M., 1959. Men Who Manage. Wiley, New York.
Day, R.A., Day, J.V., 1977. A review of the current state of negotiated order theory: an appreciation and a critique. The Sociological Quarterly 18, 126–142.
Goffman, E., 1961. Asylums. Doubleday, New York.
Grenier, C., 2011. The appropriation of new arrangements of public organizations: locally negotiate to strategically act. Administraţie şi Management Public 17, 6–30.
Mangham, I.L., Pye, A., 1991. The Doing of Management. Blackwell, Oxford.
March, J.G., Olsen, J.P., 1976. Ambiguity and Choice in Organizations. Universitetsforlagtt, Oslo.
Roethlisberger, F.J., 1968. Man-in-Organization: Essays of F. J. Roethlisberger. Harvard University Press, Cambridge, MA.
Roethlisberger, F.J., Dickson, W.J., 1939. Management and the Worker. Harvard University Press, Cambridge, MA.
Roy, D.F., 1960. Banana time: job satisfaction and informal interaction. Human Organization 18, 156–168.
Sambrook, S., 2001. HRD as an emergent and negotiated evolution: an ethnographic case study in the British National Health Service. Human Resource Development Quarterly 12 (2), 169–193.
Strauss, A., 1978. Negotiations. Wiley, New York.
Strauss, A., Schatzman, L., Ehrlich, D., Bucher, R., Sabsin, M., 1963. The hospital and its negotiated order. In: Friedson, E. (Ed.), The Hospital in Modern Society. Macmillan, New York.
Silverman, D., Jones, J., 1976. Organizational Work: The Language of Grading and the Grading of Language. Macmillan, London.
Watson, T.J., 2012. Sociology, Work and Organisation. Routledge, London.
Watson, T.J., Watson, D.H., 2012. Narratives in society, organisations and individual identities: pubs, identity work and the pursuit of the real. Human Relations 65 (6), 683–704.
Weick, K.E., 1979. The Social Psychology of Organizing. Addison-Wesley, Reading, MA.

Organizations, Sociology of

Erhard Friedberg, Paris, France

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 10968–10973, © 2001, Elsevier Ltd.

Abstract

The sociology of organizations, like its larger sister the theory of organizations, has seen many different approaches which have left their imprint on our way of thinking about organizations. However, one can distinguish two basic sociological templates in understanding organizations. Within certain limits, the *explanandum* is of course similar. For both, it is the structure and the functioning of organizations but they have different starting points, take different routes, and put the emphasis differently. Thus, the first starts with organizations and will focus on the variation of their forms: as a consequence, organizations are its basic units of analysis and it will analyze the social dynamics on the inter-organizational, sectoral, or societal level in order to explain organizational form. The second, on the other hand, will start with social action and will consider the sheer existence and maintenance of organizations as the basic problem to be addressed: as a consequence, it will concentrate its analysis on interaction processes within organizations, will have individuals (their decisions and their actions) as its basic unit of analysis, and will stress the differences between individual organizations even if they belong to the same organizational field.

The sociology of organizations, like its larger sister the theory of organizations, is not a unified field of research. In its already long history, it has seen many different approaches which have left a more or less profound imprint on our way of thinking about organizations. It is beyond the scope of this article to attempt a complete account of this diversity.

However, taking a distant view, one can distinguish two sociological templates which structure around two understandings of the word organization. The first sees organizations as structural forms, the nature, characteristics, and dynamics of which have to be explained, whereas the second insists on the processes of structuring, i.e., organizing collective human action. Within certain limits, the *explanandum* is of course similar. For both, it is the structure and the functioning of organizations but they have different starting points, take different routes, and put the emphasis differently. Thus, the first starts with organizations and will focus on the variation of their forms: as a consequence, organizations are its basic units of analysis and it will analyze the social dynamics on the inter-organizational, sectoral, or societal level in order to explain organizational form. The second, on the other hand, will start with social action and will consider the sheer existence and maintenance of organizations as the basic problem to be addressed: as a consequence, it will concentrate its analysis on interaction processes within organizations, will take individuals (their decisions and their actions) as its basic unit of analysis, and will stress the differences between individual organizations, even if they belong to the same organizational field.

Both strands have their intellectual roots in the seminal work of Max Weber. In his *Wirtschaft und Gesellschaft* (1964), Weber develops a theory of the forms of legitimate domination in society (traditional, charismatic, and rational–legal forms of domination) which roughly define the bases of the legitimacy of power and influence in a given society and condition the types of organizations to be found within each of these regimes of domination. He then argues that bureaucracy is the organizational form developing within the framework of the rational–legal form of domination and proceeds to give an ideal–typical description of bureaucracy, i.e., a highly formalized and hierarchical form of organization characterized by the predominance of written rules and procedures, formalized tasks, precise definitions of competence, clear lines of subordination, the explicit separation of ownership and management responsibility, and the merit principle as the only legitimate route of access to the different functions in the organization. Weber sees unifying dynamics at work, pushing toward the development and generalization of bureaucratic organizations as they are both legitimate and more efficient. He sees the roots of this greater efficiency essentially in the fact that the arbitrary imposition of power and the resulting interpersonal conflicts are limited by written rules and procedures which allow for more rational, foreseeable, and standardized execution of tasks.

This analysis includes all the themes which organizational sociology has developed and dwelled on through all these years. It also contains the roots for the divide which still structures two traditions in organizational sociology today. One is concerned with the study of organizational forms as they are shaped by the constraints of efficiency (this is the main concern of what has become known as 'structural contingency theory') or legitimacy (this is the theme of the 'neo-institutional school' in organizational sociology). The other tradition emphasizes the study of interaction and decision-making processes which 'produce' as well as 'reproduce' regularities and structure, i.e., organization, and which first developed out of the empirical analysis of public and private bureaucracies.

From the Dysfunctions of Bureaucracy to Organizations as Constructs of Collective Action

In the 1940s and 1950s, bureaucracy was a central theme of empirical study of organizations, especially in the United

States. This has certainly to do with the proliferation of huge administrative bureaucracies and the seemingly unlimited growth of the large corporation which was looked upon with both fascination, and fear. Following Weber's analyses, this movement was interpreted as the proof of the greater efficiency made possible by the standardization, formalization, and depersonalisation characteristic of this form of organization. However, following R. Michels' (1914) iron law of oligarchy, this secular movement raised fears because the conditioning powers of these huge organizations and the oligarchic and technocratic tendencies inherent in their functioning were seen as a threat to democracy and the ideals of reform: organizations as means for collective action in favor of reform set free forces which were in contradiction with these ideals.

The main interest of the early studies of bureaucracy (Merton, 1940; Selznick, 1949; Blau, 1955; Gouldner, 1954, 1955; Crozier, 1961, 1964) was to have considered bureaucracies as a complex social system which had to be studied empirically in order to reconstruct and understand its informal structures and dynamics. Applying the lessons and methods from industrial sociology which developed out of the Hawthorne experiments and the human relations movement sparked by Roethlisberger's and Mayo's account of them, these studies produced results which were in contradiction to both the extrapolations and fears mentioned above. They showed that bureaucracies not only were not always efficient, but also produced informal behavior and dynamics which were often dysfunctional, i.e., detrimental to the attainment of their goals and which in any case created a greater diversity in the ways organizations function than would have been expected on theoretical grounds. Thus, Merton shows that the impersonal rules on which bureaucracies are built induce rigid and ritualistic behavior on the part of their members which are dysfunctional for their capacity to respond to the needs of their clients. Blau demonstrates that the efficiency of bureaucracies is not the product of their formal characteristics but bases his argument on the fact that their members take it upon themselves to break the rules in order to enhance the quality and efficiency of their work.

These studies, however, go beyond merely showing the inefficiencies or deviations of bureaucracies. Following Robert Merton's analysis (1936) of the unanticipated consequences of purposive social action and of the latent functions of social structures, there are two complementary explanations not only for the functioning, but also for the emergence of bureaucracy.

On the one hand, these studies show that bureaucracy is the place (and also the product) of 'vicious circles of bureaucratization,' as the attempts at countering unanticipated and dysfunctional consequences of bureaucratic structures lead to an accentuation of the very characteristics which have generated these dysfunctions. This is, for instance, the mechanism analyzed by Gouldner in his study of a gypsum mine: the formal rules instituted to strengthen control over workers and to fight against their low work morale instead increase the latter's apathy, which management tries to counter by strengthening the surveillance of the workers and an increase in the rules on which this surveillance is based. This attempt, in turn, further increases workers' apathy and passivity, and so on. Crozier's analysis (1964) of the vicious circle generated by impersonal rules and the resulting parallel power relations follows a similar pattern. He shows how the proliferation of impersonal rules in view of eliminating the uncertainties which condition the satisfactory functioning of an organization increases the power of those whose contribution is needed in order to cope with the remaining sources of uncertainty. The interpersonal power and dependence relations which are generated around these uncertainties and the (unofficial and illegitimate) privileges they produce, will justify fears of personal power and dependence relations and therefore increase the pressure for further eliminating remaining uncertainties through centralization and new impersonal rules, which in turn will generate new parallel power relations, and so forth.

On the other hand, these studies explain the bureaucratization of organizations by the latent functions it fulfills for their members. Gouldner (1954) distinguishes five latent functions of bureaucracy: impersonal rules allow for distant control, thus creating a filter and a protection by reducing interpersonal relations; they legitimize sanctions but they also restrict the freedom of behavior of the members of the hierarchy by codifying the possibilities of sanction; they make possible apathy, i.e., behavior which restricts itself to the strict application of the rules; and therefore they are a resource for bargaining with members of the hierarchy if they need some extra commitment, which they always do. Moreover, Gouldner shows that bureaucratization is an answer for the problem of succession in organizations, in so far as impersonal rules can at least partially be a substitute for the personal legitimacy which the outgoing managers enjoyed, but which the incoming managers lack. Crozier's analysis (1964) follows a similar pattern but completes Gouldner's interpretations with a broader theory: the bureaucratic mode of organization based on the proliferation of written rules and procedures is a way to avoid face-to-face relations and to escape the arbitrariness and uncertainties of direct power and dependence relations. Its latent function is a general one and relates to the necessity of taming and structuring the power and dependence relations which are at the heart of collective action and without which no cooperative endeavor is feasible.

The essential merit of these seminal studies is to enable us to go beyond the sterile opposition inherited from the human relations tradition between, on the one hand, formal structure understood to be the incarnation of rationality as well as efficiency and, on the other, the actual behavior of its members which result from their irrationality, their affectivity, and their conditioning by prior processes of socialization. In their theorizing, the rationality of formal structure is no more opposed to the affectivity of human behavior. On the contrary, they interpret these structures as an answer to the cognitive limits and to the constraints stemming from human affectivity. The structure and the goals of an organization can thus no more be understood as the expression of a logic which could be independent from, and superior to, the relations between the members of an organization. They are created by these relations and draw their significance and their justification from them. In short, they are no longer an exogenous variable. They have been put back into the dealings between the members of an organization and become the endogenous result of them: they

cannot be understood independently from the interaction and bargaining processes for which they constitute only the framework. They therefore cannot escape the limits of rationality characterizing the human behavior which produces them, and have to be analyzed together.

One can clearly see the thrust of this line of reasoning. Drawing on these studies of bureaucracy as well as on the seminal work of the Carnegie group around H. Simon on a behavioral theory of decision making in organizations (March and Simon, 1958; Cyert and March, 1963) and on the analyses of collective action and public or private decision making in the United States and in Europe (see Allison, 1971; Crozier and Friedberg, 1977; Grémion, 1976; Hirschmann, 1967, 1970; Lindblom, 1959; Olson, 1965; Schelling, 1961, 1978), it points toward a behavioral conceptualization of organizations which links organizational behavior to the cognitive and relational capacities of their members. An organization, in this view, can be understood as an arena (Cyert and March, 1963) or as game structures (Crozier and Friedberg, 1977, 1995), where participants are free to choose their behavior, but within limits. The range of their choices is restricted, because if they do not want to lose in the transactions with the other participants, they have to take into account the rules of the game which prevail and which determine the value of the resources of each participant and the appropriate ways of using them in transactions with other participants. But within these limits, which may be variable and leave more or less leeway, they can and will actually choose their behavior which can thus be understood as the way in which they adjust to the constraints of their situation as they see it, while simultaneously trying to further their interests (in whatever way they may define them) by using the resources and opportunities they perceive. Their behavior therefore cannot be considered as completely hazardous: it is the expression of a decision which makes sense to the person who is choosing and therefore is rational in the sense of H. Simon's bounded rationality (Simon, 1957; March and Simon, 1958). It is a reasonable adjustment by any participant to his or her situation; i.e., the network of interdependencies within which each participant is located, an adjustment which is reasonable within the limits of his or her perception of the opportunities and constraints contained in this situation and of his or her capacity to make use of these opportunities and constraints.

Such an approach to organizations, which emphasizes the constraints on organization resulting from the limited (but extendable) cognitive and relational capacities of human beings, naturally transcends the usual distinctions between different forms of organizations (hospitals, firms, administration, etc.), and its heuristic value clearly is not limited to the analysis of formal organizations. Its target is, in fact, a much broader issue which all kinds of organizations and all forms of collective action have to solve; i.e., the problem of cooperation and coordination between actors pursuing and continuing to pursue divergent interests. It is, in a way, centered around a theory of the organizational phenomenon which aims to understand how participants who continue to pursue divergent interests can nonetheless organize or accept to be organized in the pursuit of collective goals. Such an approach radically banalizes formal organizations which, in this view, are only one of many possible forms of contexts of action the characteristics of which constrain the collective action of the various participants (Friedberg, 1993). Organizations thus become the artificial device which help analyze and understand the general problem of human cooperation and coordination. Organizational sociology becomes a way of theorizing about collective social action.

Efficiency and Legitimacy as Unifying Forces Shaping Organizational Forms

The other tradition, to which we will now turn, has a different starting point. It sees organizations as structural forms, the nature, characteristics, and dynamics of which have to be explained. Thus, it starts with organizations and will focus on the variation of their forms: organizations are its basic units of analysis and it will try and analyze the social dynamics on the inter-organizational, sectoral, or societal level in order to explain organizational form. These dynamics are traced to two distinct constraints: efficiency and legitimacy, each of which is stressed by a different strand of analysis.

Efficiency is the constraint stressed by structural contingency theory. This paradigm emerged in the middle of the 1960s and developed as a critical reaction to the theoretical and methodological perspective characteristic of organizational sociology of the 1950s. With regard to methodology, the then dominant qualitative case-study method was criticized because it provided merely a thick description but no grounds for generalizations or for the construction of a general theory of organizations. With regard to theory, it was argued that the over-emphasis on motivations and human relations characteristic of organizational thinking so far had had two consequences detrimental to our understanding of organizations: the role of structure and its influence on these relations had been downplayed, while, and this was seen as even more important, the context of an organization and the way its characteristics condition an organization's structure and functioning had been largely ignored.

As a consequence, structural contingency theory set out on a different program. Its focus was not on action or behavior within organizations, but on organizations as structured entities whose characteristics and change over time have to be explained using quantitative methods for the statistical study of samples of organizations in order to list, describe, and, if possible, measure the influence which the main dimensions of an organization's context exert on its structures, its functioning, and its performance. In other words, this paradigm was concerned with two main questions: which dimensions of context affect an organization's (mainly structural) characteristics and to what extent? what is the influence of each of these characteristics on the performance of an organization?

Structural contingency theory has been the dominant paradigm in the field of organization studies from the middle of the 1960s up to the first half of the 1980s, especially in the Anglo-Saxon world. It has generated an immense effort to determine and measure the impact of the various dimensions of context. Let us mention a few particularly significant research programs. The influence of technology on the structure of

organizations has been explored by J. Woodward (1958, 1965, 1970) and C. Perrow (1967, 1970); the Aston group around D. Pugh (1963) and D. Hickson et al. (1969) as well as P. Blau in the United States (Blau and Schoenherr, 1971) has explored the link in particular between size (among other, less central variables) and organizational structure. The impact of the technical, economic, and social characteristics of an organization's environment on its structure and mode of functioning have been independently studied in the seminal work of Burns and Stalker in England (1961) and Lawrence and Lorsch in the Unites States (1967). Last, but not least, the more conceptual work of J. Thompson (1967) has also been very influential, especially in regard to his conceptualization of what he called the 'task environment' of an organization. This approach is developed further by the population ecology of organizations, which builds on the seminal work of Aldrich (1979) as well as Hannan and Freeman (1977) and which aims at studying the contextual conditions which explain the emergence, the diffusion, and the disappearance of populations of organizations which share the same characteristics and which fit certain contextual conditions or ecological niches.

The main contribution of this quantitative and apparently more scientific approach to the study of organizations has been to demonstrate empirically the impossibility to define a single best way for structuring an organization. The good, i.e., the efficient structure, cannot be defined in general and beforehand. It is a function of the context and can only be defined after the different dimensions of this context have been recognized and taken into account in the organizational design. However, this very important contribution should not cover up the theoretical and empirical shortcomings of the approach. Indeed, according to the reasoning on which it is based, context becomes a constraint because organizations are viewed as driven by the constraints of efficiency. Indeed, so this reasoning goes, organizations have to adjust to their contexts because their performance depends on this fit: in order to survive, they have to be efficient and, in order to be efficient, they have to adjust to the demands of their context. Although there is certainly some truth in this hypothesis, the empirical diversity of organizations with similar contexts has shown that structural contingency theory has vastly overestimated the unifying power of the constraint of efficiency. It has enlarged our understanding of the forces which shape organizations but in the process has overstated its case and has been proven wrong by empirical analysis.

Against this reductionism, which pretends to analyze organizations from a purely technical or economic viewpoint (the pressure of the constraint of efficiency), the neo-institutional school in organizational analysis has promoted a more sociological perspective. It emphasizes the symbolic and normative dimensions of action in and between organizations, and stresses the role of their culture, i.e., a set of cognitive and normative frames, to explain their mode of functioning. In other words, it puts forward a less intentional and rational perspective on organizations. In this view, organizations are neither the simple tools of their masters nor machines to maximize efficiency. They are also institutions, i.e., social worlds with their specific identity and culture, which, once created, take on a life of their own and develop their own ends which can never be reduced to mere considerations of efficiency.

Sociological neo-institutionalism builds on, and tries to integrate, several theoretical perspectives: the work of P. Selznick (1943, 1949) on organizations as institutions, the work of H. Simon and his group at Carnegie on bounded rationality and cognitive frames, and the work of Berger and Luckmann on the processes of institutionalization understood as processes of the social construction of reality. In organizational analysis, it brings together the work of American and European sociologists, such as N. Brunsson, P. DiMaggio, N. Fligstein, J. G. March, J. Meyer, J. P. Olsen, W. W. Powell, B. Rowan, W. R. Scott, and L. G. Zucker, whose analysis of organizations is based on roughly three main premises.

First, an organization is an institution because it is structured by a set of cognitive, normative, and symbolic frames, which shape the behavior of its members by providing them with the tools necessary to observe and perceive the world around them, to interpret and understand their counterparts' behavior, and to construct their own interests as well as possible ways to further them. Through their structures – formal (organizational forms, procedures, institutional symbols) as well as informal (myths, rituals, social norms) – organizations shape the perceptions, calculations, reasoning, interpretations, and actions of their members by defining acceptable and legitimate behavior, i.e., behavior which is appropriate in the context of its culture.

Second, no organization exists independently of other organizations which share the same characteristics and which together form an organizational field: e.g., the organizational field of universities, hospitals, schools, airlines, museums, etc. Such organizational fields have their producers and consumers, cognitive and normative frames, their power structures, control mechanisms. In short, they have their own institutional structure and their own dynamics which are brought about by competition as well as interdependence between the constitutive organizations, processes of professionalization, i.e., the establishment of cognitive and normative frames for the field, by government intervention and the like. The dynamics of these organizational fields exert unifying pressures on the individual organizations which, in order to enhance their legitimacy, tend to adopt similar, if not identical, institutional forms and procedures.

Third, while recognizing the importance of the technical and economic environment which was stressed by the structural contingency theory, the neo-institutionalist perspective is interested mainly in the influence ofthe societal and institutional environment. The institutional environment concerns the characteristics of the organizational field of which an organization is a part, and of the rules it has to follow if it wants to obtain resources from the field and strengthen its legitimacy in it. The societal environment designates the norms and values of modern societies which, according to DiMaggio and Powell (1983) or Meyer (Meyer and Scott, 1994), are characterized by processes of rationalization (to a certain extent, this can be understood as the reedition of Max Weber's process of the disenchantment of the world) and the diffusion of standardized norms as a consequence of increased intervention by states, professions, science, and organizational fields. Both institutional and societal environments constitute a sort of unifying matrix for

organizations which have to conform to their pressures if they want to be accepted and thus able to draw resources for their functioning.

In short, the neo-institutional perspective stresses the constraint of legitimacy, as opposed to the constraint of efficiency. Rational structures (formal organizations) do not dominate the modern world because they are efficient. They adopt rationalized institutional norms because these will enable them to obtain the resources necessary for their success and their survival as they increase their legitimacy in a wider cultural environment (rationalist Western society and culture).

Taken together, the two strands of reasoning mentioned in the opening remarks provide a complete panorama of the forces shaping organizations: efficiency and legitimacy are the constraints, within which the games being played in and between organizations are embedded, but the games that depend upon the cognitive and relational capacities of the individuals playing them are in turn mediating these constraints. The unifying forces stressed by contingency theory and the neo-institutionalist perspective should therefore never be overestimated: they are themselves subject to differentiating pressures stemming from the cognitive and relational capacities of the humans who play games in contexts structured by constraints of efficiency and legitimacy. These relational and cognitive capacities are never determined and are never final: they are the motor for the infinite variance which is observable in organizational life, they are the motor of innovation which succeeds in destabilizing even the best established technical or institutional environments.

See also: Bureaucracy and Bureaucratization; Bureaucracy, Sociology of; Bureaucratization and Bureaucracy, History of; Collective Action; Institutionalism; Oligarchy, Iron Law of; Organization: Overview; Organizational Behavior, Psychology of; Organizational Climate; Organizational Culture, Anthropology of; Structural Contingency Theory; Weber, Max (1864–1920).

Bibliography

Aldrich, H., 1979. Organizations and Environments. Prentice-Hall, Englewood Cliffs, NJ.
Allison, G.T., 1971. The Essence of Decision: Explaining the Cuban Missile Crisis. Little Brown, Boston.
Blau, P.M., 1955. The Dynamics of Bureaucracy. Chicago University Press, Chicago.
Blau, P.M., Schoenherr, R., 1971. The Structure of Organizations. Basic Books, New York.
Brunsson, N., Olsen, J.P., 1993. The Reforming Organization. Routledge, London.
Burns, T., Stalker, G.M., 1961. The Management of Innovation. Tavistock, London.
Crozier, M., 1961. De la bureaucratie comme système d'organisation. Archives Européennes de Sociologie 2, 18–52.
Crozier, M., 1964. The Bureaucratic Phenomenon. Chicago University Press, Chicago.
Crozier, M., Friedberg, E., 1977. L'Acteur et le système. Editions du Seuil, Paris [trans. 1981 *Actors and Systems*. Chicago University Press, Chicago].
Crozier, M., Friedberg, E., 1995. Organizations and collective action: our contribution to organizational analysis. In: Bacharach, S.B., Gagliardi, P., Mundell, B. (Eds.), Studies of Organizations in the European Tradition, Research in the Sociology of Organizations, vol. 13. Jay Press, pp. 71–93.
Cyert, R., March, J.G., 1963. A Behavioral Theory of the Firm. Prentice Hall, Englewood Cliffs, NJ.
DiMaggio, P.J., Powell, W.W., 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. American Sociological Review 48, 147–160.
Friedberg, E., 1993. Le Pouvoir et la règle. Editions du Seuil, Paris [trans. 1997 *Local Orders*, Jay Press].
Gouldner, A.W., 1954. Patterns of Industrial Bureaucracy. The Free Press, Glencoe, IL.
Gouldner, A.W., 1955. Wild Cat Strike. Free Press, Glencoe, IL.
Grémion, P., 1976. Le Pouvoir périphérique. Editions du Seuil, Paris.
Hannan, M.T., Freeman, J.H., 1977. The population ecology of organizations. American Journal of Sociology 82, 929–964.
Hickson, D., Pugh, D.S., Pheysey, D.C., 1969. Operations technology and organization structure: An empirical reappraisal. Administrative Science Quarterly 16, 378–397.
Hirschman, A.O., 1967. Development Projects Observed. The Brookings Institution, Washington, DC.
Hirschman, A.O., 1970. Exit, Voice and Loyalty: Response to Decline in Firms, Organizations and States. Harvard University Press, Cambridge, MA.
Lawrence, P.R., Lorsch, J.W., 1967. Organization and Environment: Managing Differentiation and Integration. Harvard Business School, Cambridge, MA.
Lindblom, C., 1959. The science of muddling through. Public Administration Review 19, 79–88.
March, J.G., 1988. Decisions and Organizations. Blackwell, New York.
March, J.G., Olsen, J.P., 1984. The new institutionalism: organizational factors in political life. American Political Science Review 78, 734–749.
March, J.G., Olsen, J.P., 1989. Rediscovering Institutions: The Organizational Basis of Politics. Free Press, New York.
March, J.G., Simon, H.A., 1958. Organizations. Wiley, New York.
Merton, R.K., 1936. The unanticipated consequences of purposive social action. American Sociological Review 1, 894–904.
Merton, R.K., 1940. Bureaucratic structure and personality. Social Forces 18, 560–568.
Meyer, J.W., Rowan, B., 1977. Institutionalized organizations: Formal structures as myth and ceremony. American Journal of Sociology 83, 340–363.
Meyer, J.W., Scott, R.W. (Eds.), 1994. Institutional Environments and Organizations: Structural Complexity and Individualism. Sage, Thousand Oaks, CA.
Michels, R., 1913. Les Partis politiques. Giard et Brière. Flammarion, Paris, 1971 et 1978.
Olson, M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.
Perrow, C., 1967. A framework for the comparative analysis of organizations. American Sociological Review 32, 194–208.
Perrow, C., 1970. Organizational Analysis: A Sociological View. Tavistock, London.
Powell, W.W., DiMaggio, P.J. (Eds.), 1991. The New Institutionalism in Organizational Analysis. Chicago University Press, Chicago.
Pugh, D.S., 1963. A conceptual scheme for organizational analysis. Administrative Science Quarterly 8, 289–315.
Schelling, T.C., 1961. The Strategy of Conflict. Harvard University Press, Cambridge, MA.
Schelling, T.C., 1978. Micromotives and Macrobehavior. Norton, New York.
Scott, R.W., 1987. The adolescence of institutional theory. Administrative Science Quarterly 32, 93–111.
Selznick, P., 1943. An approach to a theory of bureaucracy. American Sociological Review 8, 47–54.
Selznick, P., 1949. TVA and the Grass-Roots. University of California Press, Berkeley, CA.
Simon, H.A., 1957. Administrative Behavior, second ed. Macmillan, New York.
Simon, H.A., 1957. Models of Man, Social and Rational. Wiley, New York.
Thompson, J.D., 1967. Organizations in Action. McGraw-Hill, New York.
Weber, M., 1964. Wirtschaft und Gesellschaft, second ed. Kiepenheuer & Witsch, Cologne, Germany.
Woodward, J., 1958. Management and Technology. H.M.S.O, London.
Woodward, J., 1965. Industrial Organization: Theory and Practice. Oxford University Press, London.
Woodward, J. (Ed.), 1970. Industrial Organization: Behavior and Control. Oxford University Press, New York and London.

Organizations: Unintended Consequences

Henk Flap, Universiteit Utrecht, Utrecht, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 16, pp. 10973–10976, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editors.

Abstract

Organizations are often constructed to make up for the unintended, sometimes perverse consequences of individual actors being unable on their own to act in their common interest; yet organizations produce *de novo* similar and new unintended consequences among and within themselves, as well as for society at large. It is not the problem as such but its manifold manifestations that is of interest, as well as their explanation. The multifarious forms of unintended consequences can all be subsumed under the problem of rationalization. A useful typology containing the basic configurations of unintended consequences has six dimensions: (1) the number of individual actors that realize their goals. (2) whether they are positive and/or negative; (3) number of other actors affected; (4) whether they are predictable; (5) the time it takes for them to develop; and (6) whether the focal actor is the corporate actor itself, or a set of individual members. Sociology provides empirical research on concrete instances, for example, oligarchic tendencies, rate-busting, and the performance paradox, whereas economics produces theoretical accounts, for example, external effects, transaction cost theory, and agency theory. Sociology focuses more on unanticipated consequences and explains these by referring to the lack of or faulty information. Economics concentrates on unintended yet frequently anticipated consequences that stem from devious incentive structures.

Unintended consequences are the grand theme of sociology. Like all social sciences sociology assumes that people act intentionally; however, in sociology these actions are seen to be socially conditioned, and while people act upon their intentions they produce unanticipated, unintended, even perverse consequences. While the notion of social conditioning has become part of the western worldview, this has not been the case for the notion of unintended consequences.

Organizations are often constructed by people to make up, through the pooling of resources, for the unintended consequence of individuals being unable to act in their common interest. Yet organizations produce new second-order unintended consequences in and among themselves and for society at large (Scott, 2014).

History of the Problem within Sociology

Interest in unintended consequences started in the eighteenth century with the individualistic tradition of the Scottish moral philosophers. In accord with Bernard de Mandeville's words, "private vices, public benefits," Adam Smith analyzed the unintended benign effects of the free market: as if moved by a hidden hand free trade transforms the self-centered motives of individual actors into a greater wealth for all. The French contract tradition provided other insights. Condorcet showed that when parties, each with consistent but different preferences on various issues, come together to collectively decide upon them, preferences on the collective level might be inconsistent, resulting in voting cycles. Rousseau described *avant la lettre* what is now known as a prisoner's dilemma, a situation in which individual and collective rationality conflict. Two persons while hunting deer fail in their objective because instead of cooperating – what they agreed upon and what is necessary to catch such a big game – one of them is tempted by the immediate gain of catching a hare. The unintended outcome is that both are collectively worse off than they could be. Thereafter, collectivist traditions prevailed. Hegel hinted at unintended consequences in his work on the long-term historical development when he spoke about the 'cunning of reason.' However, he did not analyze why they emerge. Marx adopted these ideas of laws in history and the notion that people make history without being aware, with one exception: under communism people can decide upon their common plan and escape from unintended consequences because they are unfettered by unjust power differences. The Durkheimian program of collectivist sociology did not encourage students to look for or analyze unintended consequences; although many of the phenomena Durkheim described can be reconstructed as being the unanticipated, unintended outcome of individual actions, for example, the solidarity that flows from ceremonial gatherings, or a rise in suicide rate after an economic boom, a so-called *crise heureuse*.

Merton (1949) placed the theme back on the research agenda with his analysis of manifest and latent functions; that is, consequences of individual actions that are either intended and recognized or unintended and unanticipated. He provided case studies of self-fulfilling prophecies, such as customers beginning a run on a bank they wrongly believe to be insolvent, causing it to go bankrupt, or the Mattheus effect in science, that is, how fame begets fame.

About the same time Elias presented his general ideas on unintended consequences in a historical study on state-formation and the refinement of manners. According to him unintended consequences, for example, refined manners, will be more common in larger groups, with more complex social networks, and smaller power differentials.

Lately, with the emergence of a rational choice sociology and organizational economics, unintended consequences become more central to research, although often under different headings (*see* Rational Choice and Organization Theory; Long et al., 2014).

Unintended Consequences in Organizations

Even a cursory review of the research literature shows that unintended consequences in organizations come in many forms and with multiple origins.

In the wake of Weber's discussion as to why Marxism did not succeed in its revolutionary goals, Michels (1911) formulated the iron law of oligarchy, probably the most famous instance of an unintended effect. This 'law' states that with time and growth of the organization there will be oligarchic tendencies in every organization, also in socialist organizations with a democratic ideology and constitution. Why is it rational not to be interested in one's own interests and not use one's democratic rights? Over time and with growing size it becomes less practical that all participate, and even leaders have to specialize and delegate. As the social distance between members and leaders grows, the cost for members to stay informed and to coordinate their actions increases. This gives discretionary power to the leaders also, if they only have eyes for the interests of the members. There are, however, exceptions to the iron law, like the ITU, the union of typographers that was studied by Lipset et al. (1956).

Just before World War II, industrial sociology discovered the informal group in formal organizations. Informal relations sometimes go against the objectives of the firm as they lead to rate busting. Yet, other research underlined their productive contribution. The work-leader's centrality within this informal network seems critical: employees will cooperate more readily and management can keep track of ties between employees (*see* Network Analysis).

Since informal networks can have negative consequences, a work leader has to keep track of and control all the possible ties between employees and, therefore, his span of control should be limited. Research showed an intricate unintended relation between span of control, levels of hierarchy within an organization, and level of pay: a smaller span of control leads to a taller hierarchy and greater wage differentials in an organization. Promotion schemes are another known source of unintended consequences. Stouffer (1949) gave the classic example of soldiers who were less satisfied and had lower morale in companies with a higher rate of promotion. To evaluate their own status as well as to estimate what actions would be in their own interest people compare themselves to similar others. They care for relative, and not for absolute status. Recent research, however, could not replicate Stouffer's findings. Even those who are not promoted interpret high promotion rates as a sign of trustworthiness of the employer.

Lately, an important instance of unintended consequences in organizations has been discovered, that is, the performance paradox (Meyer and Gupta, 1994). Where performance of a department or organization can be established, somehow measurements run down over a period of time for a number of reasons: (1) positive learning, that is, as performance improves, indicators become less useful, (2) perverse learning, that is, performance seems to improve without actually doing so because employees or managers adapt to the measurement system and learn how to fix the indicators, and (3) selection or replacement of poor performers for better performers. Often these processes occur simultaneously (Harbour, 2009).

A Typology of Unintended Effects

It is not the problem as such but its manifold manifestations, their explanation, and how they can be amended that is of interest. Yet the various forms of unintended consequences have a common denominator in that they can all be subsumed under the problem of rationalization, since they affect efficient goal attainment.

A useful typology that covers the basic forms should at least have six dimensions (Boudon, 1977):

1. number of individual actors that realize their goals,
2. whether the effects are positive, negative, or both,
3. number of actors affected by the outcomes,
4. whether they were anticipated or not,
5. the time it takes for consequences to develop,
6. whether a corporate actor or individual persons are the focal actor(s).

Explanation Sketches in Sociology and Economics

Existing theoretical knowledge on unintended consequences is not well integrated. The accounts given refer to a number of recurring elements, though: lack of or faulty information and incentives to act opportunistically. Sociologists traditionally focus on the former, and recently economists have emphasized the latter.

Until Coase explained why organizations exist, economists were silent on what happened in organizations. They exist because they are more efficient in certain situations as they save not only on production costs but also on transaction costs stemming from recurrent transactions and transaction-specific investments (*see* Hierarchies and Markets).

Economists made clear that there are dilemma situations in which consequences emerge that are unintended but foreseen. The employment relation, for example, is fraught with anticipated, yet often inevitable unintended effects. Employer and employee are collectively better off if they trust each other and cooperate, yet each is tempted by opportunism.

Firm-specific investments in human capital made by the employees can, depending on the alternatives both have, be misused by both parties against the other. An employer may cheat upon his promises just as the employees may sometimes shirk work or reverse who controls whom.

What situational characteristics lead to these deficits in information and incentives to act opportunistically? The examples of unintended consequences discussed either stem from functional, cognitive, or structural interdependencies between various actors (Lindenberg, 1997). People need each other to cooperate in the attainment of a common goal, to acquire or reduce information, and to get advice, gain approval, or emotional support. Moreover, persons who are functionally interdependent, for example, in their tasks, will interact and it will also affect their mind.

Another relevant distinction is whether the unintended effects of a transaction affect only those involved in the transaction or also other third parties. Economists call the latter 'externalities.' Sociologists would consider them third-party effects. Actors do not want to be responsible for negative

externalities but they would like to internalize positive ones (Opp, 2002).

If groups grow in size, Olson (1965) showed that this logic will keep people from contributing to the common good. Many unintended effects in organizations come from the outside as external effects; for example, growing time pressure and growing wealth in society at large make people want to trade time for money resulting in many voluntary organizations nowadays being 'check book writing organizations.'

Many unintended effects come about through path dependency of social processes; that is, effects of earlier choices feed back into the social situation under which individuals decide on their next move, augmenting original differences, as with, for example, oligarchic tendencies.

Why are there exceptions to unintended effects? From the relevant research literature one might take the following advice. To come up with explanations one could stick to a behavioral theory of action and think about who are relevant actors, what are their goals, what actions would further these goals at what costs given the situation, and what actions would constitute the unintended effect. Oligarchy, for example, is reflected by nonparticipation by members and leaders not acting upon members' interests. In the case of the ITU, for example, the three kinds of interdependencies came together such that the difference in knowledge and political experience was far less than usual and that leaders were also part of the informal groups. All these are conditions that promote a choice for the democratic alternative. The absence of voting chains that might complicate originally democratic choices also helps (see Wippler, 1986).

What can be done about negative unintended effects? It makes a difference whether there are central agencies that try to stabilize, correct, or foreclose the emergence of unintended negative outcomes. Functional interdependencies can be changed through social technology, for example, through redefinition of rights of actors (*see* Design and Form: Organizational).

Cognitive and especially structural interdependencies are harder to manipulate. As can be learned from principal agent and transaction cost theory, many unintended effects in organizations come from the split of ownership and usage rights, or more generally from the split between authority and responsibility. Organizational rules can redefine property rights and thereby align interests such that those who have control right over a productive capacity also have an interest in employing this capacity in a productive way. A simple application of this advice is to bind salaries of managers to performance of their department or firm. Another one is to make employees not responsible for mistakes made by others; for example, in a sequential production process a system of quality control through backward policing allows employees to refuse products delivered by others (see Milgrom and Roberts, 1992).

See also: Conflict: Organizational; Corporate Culture; Information and Knowledge, Organizational; Intelligence, Prior Knowledge, and Learning; Leadership in Organizations, Sociology of; Leadership; Learning: Organizational; Organizational Behavior, Psychology of; Organizational Decision Making; Organizations, Sociology of; Rational Choice Explanation: Philosophical Aspects; Risk, Sociology and Politics of; Technology and Organization.

Bibliography

Boudon, R., 1977. Effets pervers et ordre social. Presses Universitaires, Paris.
Harbour, J.L., 2009. The Performance Paradox: Understanding the Real Drivers that Critically Affect Outcomes. CRC Press, New York.
Long, W., Chen-Bo, Z., Murnighan, J.K., 2014. The social and ethical consequences of a calculative mindset. Organizational Behavior & Human Decision Processes 125 (1), 39–49 sert refer.
Lindenberg, S., 1997. Grounding groups in theory: functional, cognitive and structural interdependencies. Advances in Group Processes 14, 281–331.
Lipset, S.M., Trow, M.A., Coleman, J.S., 1956. Union Democracy. Free Press, New York.
Merton, R.K., 1949. Social Theory and Social Structure. Free Press, Glencoe, IL.
Meyer, M.W., Gupta, V., 1994. The performance paradox. Research in Organizational Behavior 16, 309–369.
Michels, R., 1911–1925. Zur Soziologie des Parteiwesens in dermodernen Demokratie. Untersuchungen uv ber die oligarchischenTendenzen des Gruppenlebens. Alfred Kroener Verlag, Stuttgart, Germany.
Milgrom, P., Roberts, J., 1992. Economics, Organization and Management. Prentice-Hall, Englewood Cliffs, NJ.
Olson, M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.
Opp, K.-D., 2002. When do norms emerge by human design and when by the unintended consequences of human action? The example of the no-smoking norm. Rationality and Society 14 (2), 131–158.
Stouffer, S., 1949. The American Soldier. Free Press, New York.
Scott, W.R., 2014. Institutions and Organizations: Ideas, Interests, and Identities, fourth ed. Sage, London.
Wippler, R., 1986. Oligarchic tendencies in democratic organizations. Netherlands Journal of Sociology 22, 1–17.

Orientalism

Daniel M Varisco, Qatar University, Doha, Qatar

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R.G. Fox, volume 16, pp. 10976–10978, © 2001, Elsevier Ltd.

Abstract

Interest in Europe in the study of the 'Orient' as both a geographical and an imaginary space has a long and contested history. Orientalism, or Oriental Studies, evolved as an interest in the languages and cultures of the region. By the eighteenth century, with the invasion of Egypt by Napoleon, European imperial expansion had created the opportunity for travel and a new need for analysis. Edward Said's seminal book *Orientalism* (1978) sparked a critique of academic Orientalism and its political implications. The ongoing debate, despite flaws in Said's presentation, has had a positive impact on academic study of the regions once assumed to be an unchanging Orient.

Orientalism is a contested term, primarily because of its imaginative geographical reference to a variety of major civilizations in the east of Europe. By the eighteenth century, the term referred primarily to scholarship that studied and used so-called Oriental languages. The reputed founders of Western Orientalism, such as Guillaume Postel (1510–81), Abraham-Hyacinthe Anquetil-Duperron (1731–1805), and William Jones (1746–97), were all philologists. By the nineteenth century, several societies devoted to the study of Oriental languages and cultures had been established in Europe, including the Société Asiatique in Paris (1821), the Royal Asiatic Society of Great Britain and Ireland (1823), and the Deutsche Morgenländische Gesellschaft in Leipzig (1845), as well as the American Oriental Society, founded in 1842. In 1873, the first International Congress of Orientalists was held in Paris. The papers covered ancient and contemporary Near Eastern civilizations, India, Southeast Asia, China, Japan, modern Greece, and an array of ethnographic examples including *sauvages océaniens*. As a term, *Orientalism* also has several specific usages for periods, including the early phase of British involvement in India from 1784 to 1835, before the Anglicanism exemplified by the ethnocentric Lord Thomas Babington Macaulay (1800–59). The genre of nineteenth century European art characterized by exotic harem beauties and Bedouin sheikhs on horseback is styled Orientalist, especially the paintings by Jean Auguste Dominique Ingres (1780–1867).

The primary interaction between Europe and the so-called Orient focused on the Near East. Christian pilgrims routinely visited the sites of the Holy Land in Palestine. Trade flourished in the Mediterranean after the spread of Islam into Spain in the first part of the eighth century CE. Much of the early scholarly interest came from Christians opposed to Islam, including the 1143 CE Latin translation of the Quran commissioned by Peter the Venerable (Kritzeck, 1964). The formal university study of Arabic in Europe, for example, did not begin until the Church Council of Vienna in 1312; the first Sanskrit text translated into a European language appeared in the seventeenth century. By the time of Napoleon's invasion of Egypt in 1798, a defining moment in European colonial involvement in the Orient, Arabic was taught at most of the major European universities. As Samir Amin (1989: 100–101) argues, the term *Orientalism* should not be conflated to include all study by Western authors of people encountered to their east, but rather to "the ideological construction of a mythical 'Orient,' whose characteristics are treated as immutable traits defined in simple opposition to the characteristics of the 'Occidental' world."

Oriental studies in an academic sense evolved primarily as philology, closely related to the long-standing interest of both Christians and Jews in Hebrew and Aramaic. The expansion of European states into the Orient created a need to translate and understand major languages, such as Arabic, Persian, Sanskrit, and Chinese. In the nineteenth century, archaeological discoveries led to decipherment of several ancient Near Eastern languages, including Egyptian hieroglyphics and Akkadian. Fascination with the customs of the 'Bible Lands' created a popular genre of writing by travelers and missionaries such as William Thomson's (1859) *The Land and the Book*. Several travel accounts of the Near East were widely read in Europe and America, including Richard Burton's (1855–56) account of a secret visit to Mecca in 1853, Mark Twain's (1869) satirical account of a steamship tour to the Holy Land, and Charles Doughty's (1888) eloquent account of travels in the interior of Arabia.

A wide range of academic texts have been published in the broad field of Oriental studies, which many universities have renamed Asian, Near Eastern, or Middle Eastern studies in the past four decades. For Islam and the Middle East, the most authoritative source is *The Encyclopedia of Islam*, originally published between 1913 and 1938 in English, German, and French, a second edition published between 1954 and 2005, and a third edition began in 2007. Brill's *Handbuch der Orientalistik* series has published, since 1952, monographs concerning the civilizations of the Near East and Asia. Brill also published the *Index Islamicus*, a bibliographic database that extended back to 1906 and currently covers more than 3000 journals. Although dated, the guide to resources by Jean Sauvaget (1965) is still useful for earlier references to Islam in the Middle East and North Africa.

Edward Said on Orientalism

In 1978, the literary scholar Edward Said (1935–2003) published his seminal *Orientalism*, a book that redefined and reshaped the study of Orientalism as an intellectual field of study. Said (1979: p. 1) broadened the preexisting sense of Orientalism beyond philological and area studies and suggested that the Orient 'was almost a European invention.' The core of his argument is that the word *Orient*, as used in European and American discourse,

should be labeled *Orientalism* and should not be confused with the 'real Orient.' Said situates his critique of the European invention of the Orient as someone who is an Oriental in two senses. First, he himself is stereotyped according to the ethnocentric and racist bias that is often found both in popular media and in some academic forums. In response to a later critical review, Said (1982: p. 45) sarcastically refers to critics like Bernard Lewis and Clifford Geertz as treating him like a 'wog,' a racial epithet used by the British against Arabs and North Africans. The second sense is what authorizes Said, a literary critic with little academic knowledge of the history of Oriental studies, to be taken seriously; the author is an Arab Palestinian, whose repulsion at the stereotyping is thus not "an exclusively academic matter" (Said, 1979: p. 27). Largely because of his public stand for Palestinian rights, Said has been labeled as Professor of Terror and has received death threats. The Orient, as invented, may not exist, but flesh-and-blood Arabs, Palestinians, and Muslims certainly do.

At the start of his text, Said (1979: pp. 2–3) presents three working definitions for *Orientalism*. The first, by virtue of its being 'most readily accepted,' is seemingly academic: "Anyone who teaches, writes about, or researches the Orient – and this applies whether the person is an anthropologist, sociologist, historian, or philologist – either in its specific or its general aspects, is an Orientalist, and what he or she does is Orientalism." The gist here is that Orientalism is what Orientalists do, and that whatever Orientalists do is Orientalism. Such a tautology provides the critic virtually unlimited ground for maneuvering among a wide variety of texts no one previously would have grouped together. Thus an Arabist like Edward Lane (1801–76), whose Arabic-English lexicon is still definitive after a century and a half, can be lumped together with a libertine novelist such as Gustave Flaubert (1821–80). Also new was the idea that anthropologists and sociologists, who rarely looked at Oriental texts, were doing essentially the same thing as those who did. Orientalism is further assumed to permeate all corners of academe and live on 'through its doctrines and theses about the Orient and the Oriental.'

The second sense of *Orientalism* proposed by Said is a general but more straightforward reduction of Orientalism to "an ontological and epistemological distinction made between 'the Orient' and (most of the time) 'the Occident.'" The epistemological dimension is self-evident; such a division had been widely used for centuries as an imagined cultural overlay on geographical space. Said (1979: p. 57) focused on the fifth century BCE play *The Persians* to argue that the East–West divide goes back to classical Greek culture. But the term *Oriental* was not used to refer to the inferior other by the Greeks; their term was *barbaros* (precursor of the modern term *barbarian*) and was not reserved for the Persians and other peoples to the east of Greece (Romm, 1998: pp. 95–96). Said adopts what historians call a Whiggish perspective, looking at the history of relations between peoples in Europe and the Near East through the filter of a post-Enlightenment, imperialist age of European empire. This distinction between East and West is necessarily relational but not always essential. In the realm of representations, which Said addressed in his book, the philosophical debate about ontology is not discussed. Even so, as Johannes Fabian (1999: p. 38) comments, on a different author in a different context, "ontology does not get us through the day." The point is that distinguishing Western self from Eastern other is an obvious epistemological rather than a dubious ontological issue.

It is the third meaning of *Orientalism* that fully engages the critical acumen of Said and most excites and incites his readers. The two generalized senses of *Orientalism* as whatever Orientalists do and what they always do to an Orient are situated 'historically and materially' in the colonial expansion of Britain and France over the past two centuries. Orientalism is now fully exposed "as a Western style for dominating, restructuring, and having authority over the Orient" (Said, 1979: p. 3). Style is, of course, too light a term to denote the politically charged sense of hegemonic ideology that Said's literary imagination conjures out of Michel Foucault's notion of discourse and Antonio Gramsci's elaboration of hegemony. Said employs 'discourse' to argue that Orientalism is an institutionalized force present in every Western author who writes about the Orient. Although differing in principle with Foucault, Said's (1979: p. 23) stated belief in the "determining imprint of individual authors" does not extend to even the most sympathetic of scholars such as Louis Massignon (1883–1962). Indeed, Said (1979: p. 319) argues that Orientalism is "a dogma that not only degrades its subject matter but also blinds its practitioners." In the same vein, Said (1979: p. 307) writes, "Of itself, in itself, as a set of beliefs, as a method of analysis, Orientalism cannot develop. Indeed it is the doctrinal antithesis of development." The primary reason for this is the cultural hegemony of Europe as an imperialist and colonizing power in the real Orient, which is assumed to be a passive victim.

Akbar Ahmed (1992: p. 185) observes that Said suggests mainly what this style is not, so we are told it is not what colonialist ideology says it is. It is not objective and not concerned with the real Orient. "Indeed," admits Said (1979: p. 12) "my real argument is that Orientalism is – and does not simply represent – a considerable dimension of modern political-intellectual culture, and as such has less to do with the Orient than it does with 'our' world." The fundamental corollary of Orientalism's stylistic production of the Orient is, despite Said's ontological tease, its representational function as 'a sort of surrogate and underground self' to imperial Europe. As the Syrian scholar Sadiq Jalal al-'Azm (1984: p. 351) observes, "Certainly Homer, Euripides, Dante, St. Thomas, and all other authorities that one may care to mention, held the more or less standard distorted views prevalent in their milieus about other cultures and peoples." As several critics have countered, there was an equal disdain of Europeans by peoples living in the so-called Orient.

Said noted that his threefold definition "will become clearer" later in his text. The focal point of this clarification, in Said's own words (1979: p. 206), occurs in his distinction between latent and manifest Orientalism:

> The distinction I am making is really between an almost unconscious (and certainly an untouchable) positivity, which I shall call *latent* Orientalism, and the various stated views about Oriental society, languages, history, sociology, and so forth, which I shall call *manifest* Orientalism. Whatever change occurs in knowledge of the Orient is found almost exclusively in manifest Orientalism; the unanimity, stability, and durability of latent Orientalism are more or less constant.

By *manifest*, Said refers to those parts of Orientalist discourse that can be negotiated, viewed differently, or even debated. The manifest comprises the material elements of Orientalism: texts, authors, guilds, the apparatus of academic disciplines, professional societies, the publishing industry, colonial administrative structure, and the like. Said draws examples from these material elements, especially texts, but his whole point is that nothing manifest makes sense without grounding the whole complex in the latent 'positivity' that ultimately defines it as hegemonic discourse. Unlike his use of Foucault's term *discourse* and Gramsci's term *hegemony*, the borrowing of the terms *manifest* and *latent* is not itself made manifest in *Orientalism*. These are, of course, terms laden with meaning in the psychoanalytic interpretation of dreams, especially in Lacan's reworking of Freudian theory (Yegenoglu, 1998: p. 23).

Critiques of Said's Orientalism Thesis

Said's Orientalism quickly became required reading in a number of fields and is credited as a founding text of postcolonial studies. The text, which was never revised but did receive from Said an afterword in 1994 and a preface in 2003, is available in at least 36 languages. Aijaz Ahmad (1992: p. 14) refers to Said as the central figure influencing the study of postcolonial literature. The visibility of Said's text was heightened by the increased attention paid to political events in the Middle East during the 1970s, including the 1973 Ramadan war and Arab oil embargo, the civil war in Lebanon, the jihad against the Soviets in Afghanistan, and the Iranian Revolution of 1979. The initial reviews, of which there were more than 50, were mixed. Most praised the effort of Said to expose the blatant ethnocentrism and racism in much of the previous literature on the Orient, but many pointed out historical errors and Said's lack of familiarity with the academic genre. "The basic problem," wrote historian Richard Bayly Winder (1981: p. 618), "is that *Orientalism* is a polemic, in a good cause, but a polemic."

The most caustic critic, historian Bernard Lewis, was a target of Said's disdain. Lewis is dismissed by Said (1979: p. 316) as having "a project to debunk, to whittle down, and to discredit the Arabs and Islam." As an example, Lewis is criticized for saying in a minor conference article that one of the Arabic terms for *revolution* in a political sense has an etymology that refers to the awkward rising of a camel. "Instead of revolution there is sedition," writes Said (1979: pp. 315–316), "setting up a petty sovereignty, and more excitement, which is as much as saying that instead of copulation the Arab can only achieve foreplay, masturbation, coitus interruptus." Said attributes a sexual slant to the editor of the conference's proceedings as well, ignoring the editor's observation that the concept of revolution in European usage "attempts to create substitutes for reality: to control the known world by bringing it under its history-moulding will" (Vatikiotis, 1972: p. 8). Not only does this view resonate with Said's own approach to Orientalism as discourse, but in fact the 'bad kind of sexuality' that Said teases out of the passage would target revolutions in Europe and not in the Middle East. Lewis (1982) waited three years before responding to Said in print, fueling the personal antagonism that dominates their exchanges.

Anthropologists, while respecting Said's critique of Western ethnocentrism, have for the most part faulted Said for his lack of knowledge about ethnography and anthropological theories of culture. The 'rise of ethnography' is traced by Said (1993: pp. 44, 108) to Gobineau, Maine, Renan, and von Humboldt – none of whom actually did ethnographic fieldwork or provided methodological models embraced by modern anthropology as a formal discipline. Said's view of culture is based on the elitist literary definition by Matthew Arnold, with no concern for the evolution of cultural concepts within anthropology since Edward Tylor (Varisco, 2004). George Marcus and Michael Fischer (1986) begin their book *Anthropology as Cultural Critique* by citing the 'broad and indiscriminate' brush used in Orientalism to attack Western representation of the Orient. Michael Richardson (1990: p. 17) accuses Said of "as much mis-representation of Orientalists as he accuses them of making in their representations of the Orient."

One of the few anthropologists to speak positively about Said's methods as a literary critic useful for anthropologists is Lila Abu-Lughod. After Said's 1987 lecture at the annual meeting of the American Anthropological Association, Abu-Lughod (1993: p. 10) engaged, at Said's suggestion, to write against essentializing discourse in anthropological theory. The specific notion of culture that Lila Abu-Lughod is writing against is one that most anthropologists would concur needs to be delegitimized. Unlike Said, she notes that cultural anthropologists have long questioned the idea of scientific objectivity, have consistently opposed racism, and have developed a reflexive critique of their own discipline. Said, who is not alone, is right to question the damaging binary of West vs East as an imaginative geography that necessarily instates a difference based on power. Yet Abu-Lughod is writing primarily against an assumed tendency to essentialize or freeze cultures rather than the common anthropological position that human beings all share, in their collective evolution, a qualitative cultural distinction. Andrew Shryock (1997: p. 314) suggests that Abu-Lughod is essentially writing against 'androcentric, agonistic discourse,' adding that this "slights a world of experience and concern; moreover, it renders much of intellectual and political life in the Middle East immediately unintelligible."

Numerous critics have noted that the central problem with Said's Orientalism thesis is not that there was no prejudicial writing about the imagined Orient, but that he essentializes Western Orientalists as part of a homogeneous discourse, a kind of Orientalism in reverse. Said has been praised for applying Foucault's notion of discourse and Gramsci's sense of hegemony. Unlike Foucault, however, Said derives Orientalist discourse from master texts and canonical authors and thus misunderstands the Foucauldian approach to power (Bhabha, 1994: p. 72; Clifford, 1988: p. 265). Said also ignores the Marxist underpinnings of Gramsci's writing and fails to show how Orientalist discourse manipulates consent by real people in the Orient. Dennis Porter (1994: p. 152) notes that Said's Orientalism lacks "a sense of hegemony as process in concrete historical conjunctures, as an evolving sphere of superstructural conflict in which power relations are continually reasserted, challenged, modified."

Book-length summaries of the historical errors, rhetorical devices, and distortions in Said's *Orientalism* have been provided by Irwin (2006) and Varisco (2007). One of the most

common criticisms is Said's failure to include German-speaking Orientalists, who are among the most important scholars studying Islam and the Near East. Notably absent is the Hungarian-born Ignaz Goldziher (1850–1921), who is considered to be one of the major influences on the academic study of Islam in Europe (Patai, 1987). Said also ignores real Oriental scholars who responded to Western prejudice, including 'Abd a-Rahman al-Jabarti (1756–1825)' (Keddie, 1968). Arab critics generally dismissed Said's work, which ignores any critique written in Arabic. The Lebanese scholar Mahdi 'Amil (1985) provides a detailed rebuttal to Said's dismissal of Karl Marx as an Orientalist and rejection of all Marxist theory. Feminist scholars have noted Said's lack of engagement with female authors and the idea of the female racial other (Holmlund, 1993: p. 1).

Two egregious copying errors have drawn the attention of several critics. The first is a mistranslation of lines from Goethe's famous 1819 *Westöstlicher Diwan*. The phrase "Gottes ist der Orient!" is garbled by Said as "God is the Orient," but the German indicates the genitive. This is given by Said in a passage that blames Goethe for romantic pantheistic tendencies, but in this case Goethe was translating directly from the German translation of a Qur'anic verse (Surat al-Baqara 2:142), which reads "The East and the West are God's." A second example is a misquote of a passage Said uses to illustrate the anti-Oriental bias of Gustave Flaubert. Said (1979: p. 184) cites Flaubert as writing "Inscriptions and bird-droppings are the only two things in Egypt that give any indication of life" as an example of Flaubert's 'grotesque nastiness.' The actual text, both in the English translation cited by Said and in the original French, includes the words 'in the ruins of Egypt,' but Said does not provide an ellipsis for the missing words. Flaubert wrote this while visiting the monumental Colossus of Memnon, and it is clear from the French that the reference is not to hieroglyphic inscriptions but to travelers' graffiti. Such careless errors do not invalidate Said's overall thesis, but they indicate that his polemical interests outweighed his attention to sound scholarship.

Critiques of Orientalism before Edward Said

Although Said's *Orientalism* is the major English-language book to draw attention across disciplines and fields to bias in Orientalist writing, it was not the first demythologizing of the Orient concept. The most thorough study of the intellectual roots of the 'myth' of East vs West is Kurt Goldammer's (1962) *Der Mythus von Ost und West: Eine kultur und religionsgeschichtliche Betrachtung*. Targeted in this earlier study, not known by Said, is the tendency in the literature to ignore factual differences and to posit a false uniformity so as to reduce the East to a racial question and portray the East falsely as a mirror of the West. 'Das ganze Phänomen,' concludes Goldammer (1962: p. xx) "gehört zu der Unzahl von vorgefassten schematischen Meinungen und Begriffs-fixierungen, mit denen wir leben, unter denen wir leiden, die unseren Alltag und unsere Bildung erfüllen, und deren wir selbst nach generationslanger Abnutzung nicht überdrüssig werden" ("The entire phenomenon belongs to a large number of preconceived, schematic ideas and fixed notions with which we live and under which we suffer, that imbue our daily life and education, and which do not disgust us even after generations of use" (my translation)). Other pertinent criticisms of such an East–West binary were available, even in English. "Many a writer on Asia treats the Orient as though it were a single entity (which it is not) – and thus postulates a unity that has no real existence outside his own imagination," wrote John Steadman (1969: p. 14) a decade before *Orientalism* appeared. Arguing that there was "no objective basis for the dichotomy of East and West," Steadman proposed, in his study of the 'myth' of Asia, to provide "the anatomy of an illusion, the dissection of an eidolon." At almost the same time, Stephan Hay (1970: p. 30) went so far as to say that among scholars the "idea of 'the East' is now all but extinct." In 1973 the term Orientalism was abandoned at the 29th International Congress of Orientalists in 1973. The binary that Said rightly targeted for deconstruction was already under indictment by scholars within the very guild Said thought incapable of such self-critique.

Resources for Orientalism

There are several intellectual histories of the rise of Oriental studies, most now overshadowed by the polemical work of Edward Said. The major study through the middle of the twentieth century was Johann Fück's (1955) *Die Arabischen Studien in Europa* and Maxime Rodinson's (1991) *Europe and the Mystique of Islam*. These are now supplemented by Zachary Lockman's (2004) *Contending Visions of the Middle East*; Daniel Martin Varisco's (2007) *Reading Islam: Said and the Unsaid*; Ursula Wokoeck's (2009) *German Orientalism: The Study of the Middle East and Islam from 1800 to 1945*; and Ivan Kalmar's (2012) *Early Orientalism: Imagined Islam and the Notion of the Sublime Power*. Edward Said drew much of his historical information on Orientalists from Raymond Schwab's (1950) *La renaissance orientale*, published in English in 1984. For American Orientalism, two useful sources are Douglas Little's (2003) *American Orientalism: The United States and the Middle East since 1945* and Mari Yoshihara's (2003) *Embracing the East: White Women and American Orientalism*. A concise overview of Orientalism is given by Jacques Waardenburg (1992) in the second edition of the *Encyclopedia of Islam*. Oxford Bibliographies Online has an updatable bibliography devoted to Orientalism and Islam (Varisco, 2010).

Bibliography

Abu-Lughod, Lila, 1993. Writing Women's Worlds: Bedouin Stories. University of California Press, Berkeley.

Ahmad, Aijaz, 1992. In Theory: Classes, Nations, Literatures. Verso, London.

Ahmed, Akbar, 1992. Postmodernism and Islam. Routledge, London.

'Amil, Mahdi, 1985. Marks fi istishraq Adward Sa'id. Dar al-Farabi, Beirut.

Amin, Samir, 1989. Eurocentrism (Russell Moore, Trans.). Monthly Review Press, New York.

Al-'Azm, S.J., 1984. Orientalism and orientalism in reverse. In: Rothschild, Jon (Ed.), Forbidden Agendas: Intolerance and Defiance in the Middle East. Al-Saqi, London, pp. 349–376.

Bhabha, Homi, 1994. The Location of Culture. Routledge, London.
Burton, Richard, 1855–56. Personal Narrative of a Pilgrimage to Al-Madinah & Meccah. Longman, Brown, Green, and Longmans, London.
Clifford, James, 1988. The Predicament of Culture: Twentieth-Century Ethnography, Literature, and Art. Harvard University Press, Cambridge.
Doughty, Charles, 1888. Travels in Arabia Deserta. Cambridge University Press, Cambridge.
Fabian, Johannes, 1999. Comments. Current Anthropology 40, 38. (Suppl.).
Fück, Johann, 1955. Die Arabischen Studien in Europa bis in den Anfang des 20. Jahrhunderts. Harrassowitz, Leipzig.
Goldammer, Kurt, 1962. Der Mythus von Ost und West: Eine kultur- und religionsgeschichtliche Betrachtung. Ernst Reinhardt, Basel.
Hay, Stephen N., 1970. Asian Ideas of East and West: Tagore and His Critics in Japan, China, and India. Harvard University Press, Cambridge.
Holmlund, Christine Anne, 1993. Displacing limits of difference: gender, race, and colonialism in Edward Said and Homi Habha's theoretical models and Marguerite Duras's experimental films. In: Naficy, Hamid, Gabriel, Teshome H. (Eds.), Otherness and the Media: The Ethnography of the Imagined and the Imaged. Harwood Academic Publishers, Chur, Switzerland, pp. 1–22.
Irwin, Robert, 2006. For Lust of Knowing: The Orientalists and Their Enemies. Allen Lane, London.
Kalmar, Ivan, 2012. Early Orientalism: Imagined Islam and the Notion of Sublime Power. Routledge, London.
Keddie, Nikki R., 1968. An Islamic Response to Imperialism: Political and Religious Writings of Sayyid Jamal Ad-Din "al-Afghani". University of California Press, Berkeley.
Kritzeck, James, 1964. Peter the Venerable and Islam. Princeton University Press, Princeton.
Lewis, Bernard, 24 June 1982. The question of orientalism. The New York Review of Books 29 (11), 49–56.
Little, Douglas, 2003. American Orientalism: The United States and the Middle East since 1945. University of North Carolina Press, Charlotte.
Lockman, Zachary, 2004. Contending Visions of the Middle East: The History and Politics of Orientalism. Cambridge University Press, Cambridge.
Marcus, George E., Michael, M.J. Fischer., 1986. Anthropology as Cultural Critique: An Experimental Moment in the Human Sciences. University of Chicago Press, Chicago.
Patai, Raphael, 1987. Ignaz Goldziher and His Oriental Diary. A Translation and Psychological Portrait. Wayne State University Press, Detroit.
Porter, Dennis, 1994. Orientalism and its problems. In: Williams, P., Chrisman, L. (Eds.), Colonial Discourse and Post-colonial Theory: A Reader. Columbia University Press, New York, pp. 150–161.
Richardson, Michael, 1990. Enough said: reflections on Orientalism. Anthropology Today 6 (4), 16–19.
Rodinson, Maxime, 1991. Europe and the Mystique of Islam (Roger Veinus, Trans.). University of Washington Press, Seattle.
Romm, James, 1998. Herodotus. Yale University Press, New Haven.
Said, Edward, 1979. Orientalism. Vintage Books, New York.
Said, Edward, 12 August 1982. Orientalism: an exchange. The New York Review of Books 29 (13), 44–46.
Said, Edward, 1993. Culture and Imperialism. Knopf, New York.
Sauvaget, Jean, 1965. Introduction to the History of the Muslim East: A Bibliographical Guide. University of California Press, Berkeley.
Schwab, Raymond, 1950. La Renaissance orientale. Payot, Paris.
Shryock, Andrew, 1997. Nationalism and the Genealogical Imagination: Oral History and Textual Authority in Tribal Jordan. University of California Press, Berkeley.
Steadman, John M., 1969. The Myth of Asia. Simon and Schuster, New York.
Thomson, William, 1859. The Land and the Book. Harper and Brothers, New York.
Twain, Mark, 1869. The Innocents Abroad. American Publishing Company, Hartford.
Varisco, Daniel Martin, 2004. Reading against culture in Edward Said's Culture and Imperialism. Culture, Theory and Critique 45 (2), 93–112.
Varisco, Daniel Martin, 2007. Reading Orientalism: Said and the Unsaid. University of Washington Press, Seattle.
Daniel Martin Varisco, 2010. Orientalism and Islam. Oxford Bibliographies Online. Electronic document, http://oxfordbibliographiesonline.com/display/id/obo-9780195390155-0058.
Vatikiotis, P.J. (Ed.), 1972. Revolution in the Middle East and Other Case Studies. Allen and Unwin, London.
Waardenburg, Jacques, 1992. Mustashrikun. In: The Encyclopaedia of Islam, second ed., vol. 7. Brill, Leiden. pp. 735–753.
Winder, Richard Bayly, 1981. Orientalism. Middle East Journal 35 (4), 615–619.
Wokoeck, Ursula, 2009. German Orientalism: The Study of the Middle East and Islam from 1800 to 1945. Routledge, London.
Yegenoglu, Meyda, 1998. Colonial Fantasies: Towards a Feminist Reading of Orientalism. Cambridge University Press, Cambridge.
Yoshihara, Mari, 2003. Embracing the East. White Women and American Orientalism. Oxford University Press, Oxford.

Orientalism, Oppression, and Contemporary Social Work

Elizabeth Beck, School of Social Work, Georgia State University, Atlanta, GA, USA
Stéphanie Wahab, University of Otago, Dunedin, New Zealand; and Portland State University, Portland, OR, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Edward Said's work on Orientalism gives meaning to social workers, particularly those interested in social work, which attends to issues of representation, power, privilege, and oppression. Both authors place Orientalism in context by examining social work discourse and practices associated with gender, anti-Arab racism and Islamophobia, Zionism, cultural competency, and international social work.

Orientalism in Context for Social Work

The word Orientalism predates Edward Said, but it was his analysis of the term combined with his determination to 'speak truth to power,' that has rendered the word transformative, contested, and revered. It is also his exploration of Orientalism that gives meaning to social workers. Attention to Orientalism can support social workers to understand common practices associated with helping 'others,' work across differences, and unpack normative professional narratives that inadvertently create knowers (i.e., social work professionals) and objectify 'others' (i.e., 'clients,' service users). For social workers addressing issues of social justice, Orientalism can provide insight into the dangers of objectification, mechanisms that create and maintain oppression and dominance – including ideas associated with neoimperialism – as well as the enduring nature of anti-Arab and anti-Muslim prejudice. Additionally, Orientalism provides social workers with a postcolonial lens that forces us to ask whose definition of social justice is being advanced and what is its relationship to dominance? As authors, one of us is a Palestinian-French Canadian with American citizenship, and the other a secular Jew who identifies as an ally to Palestinians, a framing of ourselves that constitutes the 'we' in this article.

Key Concepts in Orientalism

Edward Said's 'Orientalism'

As a Professor of English and Comparative Literature, Said's Orientalism was derived from his close readings of Anglo, French, and British post-Enlightenment literature as well as texts ranging from travelogues to political speeches. Although the geographical area explored by Said over the course of his 20 books was expansive, Arabs and Islam constituted his primary focus. The decision to focus on Arabs and Islam was based on both the historic position of the Orient and Said's own experience as an 'Oriental.' East and West are used to bifurcate the globe between Western parts of the globe, i.e., America, England, and Europe, and the exotic and less civilized East. Other terms like Global North and Global South are used to delineate industrialized nations from the nonindustrialized ones. In this article, East is used to represent both the geographic East and countries outside of the Western and industrialized hegemony.

Said was born in Palestine, schooled in Egypt, and worked in New York City. As an Oriental living in the West, Said both studied and experienced the themes raised in Orientalism, themes that included "racism, cultural stereotypes, political imperialism and dehumanizing ideology" (Said, 1992: p. 27).

What Is Meant by Orientalism

Prior to Said, Orientalism was an academic discipline dominated by Westerners who studied the near and far East. With the 1978 publication of his book *Orientalism*, Said shifted this subject of study. No longer was Orientalism a study of the Orient and its culture, rather it became the study of how the Orient is viewed, stereotyped, and thereby dominated by its opposite, the Occident, or the West.

The domination that Said spoke about included traditional colonization to what is presently considered neoimperialism. In the latter, former colonies are economically indebted to and controlled by the West through such entities as multinational corporations, international monetary structures, and even some forms of foreign aid in which the emphasis is on the security and economic interest of the donor agency rather than local need (Bayoumi and Rubin, 2000; Burney, 2012).

For a basic understanding of Orientalism, it is instructive to explore both the theories that underpinned Said's thinking as well as some of the stereotypes identified by Said. Said's Orientalism was influenced by the scholarship of French Psychoanalyst Jacques Lacan, French Philosopher Michel Foucault, and Italian Marxist Antonio Gramsci. Building on Lacan's notion of 'the gaze,' in which the objectification of another diminishes the object's autonomy, Said shifts the subject from the gaze on the Orient to the West's construction of the Orient (Bayoumi and Rubin, 2000). Said then locates Orientalism within Foucault's framework of discourse by arguing that the Orient is a creation of its binary, the Occident. In this binary, Occident is 'normative' and Orient is 'other.' Harkening to Foucault's work on knowledge production and its ability to manufacture 'truthful statements' about the subjects studied, Said argued that Western studies and its use of authoritative discourse created and then exploited a problematic view of the Orient. Gramsci's work on hegemony, as conceptualized by ideological leadership, allows Said to

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28066-5

explain the enduring nature of Orientalism, even when faced with evidence to the contrary (Bayoumi and Rubin, 2000; Burney, 2012; Said, 1978).

Orientalist Stereotypes

Oriental stereotypes, according to Said (1978), are used to rob the Orient of its truths while both silencing and diminishing the Orient to the Occident. For example, Westerners celebrate Sir Edmund Hillary's achievement as the first man to conquer Mount Everest, despite evidence suggesting that Hillary's Nepalese Sherpa, who clearly carried the larger portion of the heavy lifting, most likely reached the top before Hillary. The example of Hillary and the Sherpa are repeated in many variations including Western geographic or archeological 'discoveries' of inhabited land.

In addition to stereotypes involving the Orient as a whole, Said spoke about the significance of several pervasive and self-reproducing stereotypes. These include Oriental depravity, Oriental splendor, and Oriental sensuality. Oriental depravity assumes many forms and a number of them are centered on the Prophet Mohammed, who is often vilified as an 'imposter.' Said explains that the label of imposter was the result of the incorrect assumption that Mohammed was to Islam what Christ was to Christianity, thus forcing Mohammed into a binary in which Christ was legitimate and Mohammed an imposter who needed to be destroyed or punished. Consider, for example, Dante's *Inferno* in which Mohammed's sin of deception places him with his entrails exposed in the eighth of the nine circles of hell.

As the Nepalese Sherpa is silenced, so are Oriental women who also do not speak for themselves. For example, consider Said's description of Flaubert's courtesan. She "never spoke for herself, she never represented her emotions, presence, or history. He spoke for and represented her" (Said, 1978: p. 72).

Oriental splendor and sensuality are iconic in figures such as Cleopatra, Salome, and Isis. While women such as Cleopatra conjure their own images, she is used in multiple other works (see for example, Charolette Bronte's *Villette*) to express the desires of nineteenth-century women. Additionally, images such as the Harem, as often depicted in nineteenth- and twentieth-century Orientalist paintings, shows both the exoticism and the depravity of the Orient. The Harem is often used in books and movies to illustrate splendor and depravity, and storylines tend to include a subtext of Western intervention. For example, it is often seen that the Western hero is often portrayed as seeking to save an Eastern woman or wandering Western women from the depravity and exoticism of the harem (Abdulhadi et al., 2011).

Orientalism and Representations of Arab and Muslim Women Post-9/11

There is relatively new literature exploring the post-9/11 depictions of Arab and Muslim women as found in the Western media and in policy debates. Given that Orientalized women do not speak for themselves, one of the major themes of the post-9/11 representations is the hyperoppressed, shrouded woman in need of Western intervention. This theme is beautifully expressed by Lila Abu-Lughod (2002) when she asked *Do Muslim Women Really Need Saving?* Abu-Lughod argues that in the post-9/11 world, depictions of Muslim women are largely essentialized as victims of 'the veil,' 'stoning,' and 'honor killing.'

While violence against women must never be condoned, images of hyperoppressed women have been used to justify war in Afghanistan, and distract from a broader analysis of global injustice. For example, in the case of Afghanistan, American constructions of Middle Eastern women as unable to speak for themselves reinforced Orientalist views, while 'women saving' supplied a liberal tilt to an unpopular war (Jarmakani, 2010: p. 228). For social workers committed to human rights, one can see how public support for Saudi women drivers or Afghani women's rights not to be fully covered serves as a distraction from systemic concerns associated with war, occupation, and colonization (Erakat, 2011). As different groups of Arab and Muslim women are seen as terrorized by Arab and Muslim men, their multiple and complex realities are simplified, essentialized, and appropriated by Western feminists (Abdulhadi et al., 2011) and their needs associated with national liberation, decolonization, freedom from racism, and poverty are erased and ignored.

For many Muslim and Arab women, wearing the veil is a choice that is associated with concepts of identity, personhood, religious and cultural beliefs, and esthetic preferences (Abdulhadi et al., 2011). Unfortunately, the relevance and reverence for traditions associated with wearing the veil for Arab and Muslim women are often obscured by stereotypes that lead to the pitying of them as the significance of the veil (to Westerners) is often misunderstood, and the agency of the women wearing it diminished.

Additionally, the Orientalized view of gender and gender oppression can distract social workers from engaging in any form of critical analysis of gender, religion, nationality, and ethnicity necessary for antioppressive practice. The Orientalized view of gender can lead to single and essentializing narratives such as Arab women are oppressed by their men, Muslim women have little agency in their day to day lives, and Arab women who wear hijab are oppressed. Antioppressive practice requires a critical analysis of power that includes attention to resistance, agency, and diversity across identities. There is no single story about Arab or Muslim women. Women, Arab and Muslim social workers, scholars, and activists find themselves responding to what Jarmakani (2010) calls false binaries such as whether the veil is oppressive or not, rather than the complex and lived experiences of patriarchy (because the question of gender oppression is not denied). In other words, efforts to support equality in gender and sexual orientation are interrupted as scholars' and activists' attention is turned from these issues to refuting Orientalist constructions of Muslim and Arab women by non-Muslims or non-Arabs.

Orientalism and the Persistence of Islamaphobia and Anti-Arab Racism

Written in 1530, and currently available in multiple stores, libraries, and on the Internet, the book entitled *Here after Foloweth a Lytell Treatyse Agaynst Mahumet and his Cursed Secte*

(Anon, 1530) speaks to the enduring nature of Orientalized depictions of depravity and the ways in which violence against Muslims has been justified. In 1978, Said wrote, that "practically the *only* ethnic group about whom in the west racial slurs are tolerated, even encouraged is the Arabs" (p. 26). Said's sentiment was affirmed 20 years later in 1999 in a widely cited report issued by the Runnymead Trust in England, indicating that anti-Muslim and anti-Arab discourse had become normative.

Today, the term Islamophobia is used to describe the racialized bigotry, and discriminatory policies and practices geared toward Muslims across the world, as well as to Arabs (and non-Arabs mistaken as Arabs) from a variety of religious backgrounds. Islamophobia and anti-Arab prejudice is particularly strong in Europe and the United States where Arabs and Muslims have been the target of discriminatory public policies, rising levels of intolerance, and hate crimes (Sayyid and Vakil, 2011). The vitriolic comments made by the North American woman who confessed to murdering Indian immigrant Sunando Sen in December 2011 are illustrative. She is quoted as saying "I pushed a Muslim off the train tracks because I hate Hindus and Muslims …. Ever since 2001 when they put down the Twin Towers, I've been beating them up" (Santora, 2012: p. A15).

For Muslims, Arabs, and individuals who are of Muslim or Arab decent or are perceived as such, living in the West includes many of the issues associated with being marked as racially and ethnically different. Difficulties include marginalization, racism, violence, isolation, and cultural imperialism. The resulting harms involve exclusion, stigma, hate speech and crimes, as well as stress including traumatic stress. Stereotypes of violence and terrorism accompany Muslim and Arab identities. In her exploration of Arab and Arab American youth living on the West coast of the United States, Naber (2012) interviewed youth in San Francisco who shared their disappointment and fear that the only portrayals of Arabs in the media are ones of Arabs as killers and radical terrorists.

While Orientalism helps place Islamophobia in context, social workers are faced with two important questions. What are the implications of Islamophobia and anti-Arab racism for social work practice? This question has particular significance given that social workers are often in the role of providing services to and working alongside Arab and Muslim individuals, communities, and colleagues. This question also has particular implications for social workers and social work students who identify as or are perceived to be Arab or Muslim. The second question, given our profession's commitment to social justice is what should be our role in addressing Islamophobia and anti-Arab or anti-Muslim prejudice?

Implications of Islamophobia and anti-Arab racism on social work practice include the need to incorporate Orientalism theory and content in social work education to attend to ethical standards that guide our profession, as well as to support antioppressive practice. This is suggested by the International Federation of Social Workers (2012), in Ethical Principle 4.2 – Social Justice – which identifies social workers' responsibility to challenge negative discrimination. Social workers cannot challenge Islamophobia or anti-Arab racism at micro-, mezzo-, or macrolevels if they cannot recognize either the phenomenon or their participation in it. Instead students, practitioners, researchers, and academics must understand the macroforces that have infused what individuals know about Arabs and Muslims and how they know it. Social workers must be able to unpack their own social identities, as well as how these inform their working relationships with diverse groups of Arabs and Muslims so that they can more effectively interrupt discrimination and oppression while working toward social justice.

Palestine

For Said (1992), the apex of Orientalism in the modern world and in his own biography was Palestine. Understanding the question of Palestine, as constructed in the United States, is instructive for social workers as it illustrates modern day experiences of Orientalism. Social workers, particularly in the United States, have largely been exposed to a single story about Palestine and Palestinians. This story paints an 'us–them' binary with a benign 'us' and 'them' as radical terrorists who undermine the security of Israel and stability of the entire Middle East, thus providing a contemporary instance of Oriental depravity. As social workers find themselves in positions to address the harm done to Palestinians at the micro-, mezzo- and macrolevels, they need to have a greater understanding of the experiences of Palestinians from diverse Palestinian perspectives. Whether engaged in international social work, or microlevel practice with Palestinians living abroad, especially during times of war, social workers must be able to think critically about the ways in which Orientalism informs their theories, practices, and biases. By recognizing, for example, the existence of multiple Palestinian narratives associated with struggle, survival, gender, religion, place, family, and nationality among other issues, social workers, particularly in the United States, become exposed to the diversity of Palestinian experiences and realities, as well as counternarratives that challenge Orientalist and Zionist knowledge about Palestinians.

In his book, *The Question of Palestine*, Said (1992: p. 15) writes that support of Israel or Zionism has "rationalized the denial of present reality in Palestine." The website, *If Americans Knew*, offers a quick overview of some of the hardships and humiliations that comprise the present realities of Palestinians. The Web site makes clear that since the year 2000, there have been at least six Palestinians killed for every Israeli and more than 10 times more children. We do not use these numbers to suggest that any death is tolerable, but rather to show the scale between them. Political imprisonment and detention as well as restrictions on movement and water are common elements of life on the West Bank (*If America Knew*, n.d.).

Although the United Nations has issued 65 separate resolutions indicating human right abuses by Israel regarding the Palestinians (none have been issued against Palestine), Israel is portrayed as the victim in much of the North American public discourse, and efforts to suggest otherwise can carry heavy costs (Carter, 2006b). For example, accusations of being "negative, anti-Semitic, or an apologist for Islam and the Arabs" (Said, 1992: p. 25) are leveled when one deviates from the narrative of Israel as holding a higher moral ground, as evidenced by the vitriolic criticism from some US Democratic leaders and American Jewish organizations when American President

Jimmy Carter (2006) used the word apartheid in his book *Palestine: Peace Not Apartheid* in describing the Palestinian–Israeli conflict (see, for example Northam, 2006). This limited tolerance for a full and critical discourse places Palestinians and Arabs at risk for further marginalization and isolation, in the United States in particular.

Therefore, we know that in addition to marginalization and isolation, individuals of Palestinian descent know our pain, that of our family and community members, and the pain in our country's recent history. For example, Thabet et al. (2002) studied the prevalence of posttraumatic stress disorder (PTSD) among 180 Palestinian children who either experienced bombardment or demolition of their home by the Israeli government when 18 000 Palestinian homes had been demolished. This study found that 54% of the children exposed to this horror experienced clinical levels of PTSD. This prevalence of PTSD affects social workers working with these youth and their families, personally and professionally. For social workers in the West with family and friends living in the occupied territories, the impact can be particularly acute. Such high levels of PTSD also carry implications for social work advocacy concerned with social justice and human rights, and carry implications for the public and mental health effects of occupation and war.

In many parts of the West (especially in the United States), the pain of the Palestinian people is largely unrecognized. For people with Palestinian ties living in the West, the set of feelings regarding the homeland, the fears for loved ones who remain, and the news of traumatic experiences engender grief that is largely unacknowledged and misunderstood by Orientalist societies.

Ken Doka (1989) coined the term 'disenfranchised grief' to write about the pain and difficulty of healing associated with grief that is not supported by society or of which there is stigma associated with the loss. To support Palestinian healing and restore self-determination lost as a result of war and occupation, social workers in the West must attain appropriate information about the effects of Orientalism on Palestinians. They can begin by unpacking the ways by which they have internalized Orientalist constructions of Arabs and Muslims before attempting to help. Given the dominance of Zionist narratives about Palestinians in the United States, accessing this type of information is neither easy nor self-evident. Social workers wanting to educate themselves, prior to working directly with Palestinians in or out of their own countries, may begin by reading some of the references provided in the body of this text, or consider some of the texts listed on *The Faculty Lounge: Conversations about law, culture and academia* (2012). When ready, social workers might engage in antioppressive practice by becoming allies to Palestinians (in or out of their own countries).

The issues affecting Arab and Palestinian youth in particular are potentially important to social workers. Studies speak to the ill effects of racial profiling on youth, but studies informing social workers what it means to go through life profiled as a terrorist, are lacking. Consequently, there is a need for social workers to learn about and research the harm caused by Orientalist stereotypes of Palestinians and Arabs. Beyond educating themselves and gaining the ability to unpack their own Orientalism, social workers can seek out youth groups and organizations that work with young Arabs to become allies that support their self-identified needs. Working with those most affected by Islamophobia and anti-Arab racism, however, is not the only (or even most desirable) way to engage in anti-oppressive practice with these issues. Social workers can also focus their advocacy on changing systems (i.e., racial profiling), institutions (i.e., schools in the West that do not allow young Arab women to wear the hijab), and norms (i.e., constant depictions in the media of Arabs as violent and depraved) that perpetuate these harms.

Critiques

Said's work and his construction of Orientalism is contested. Critiques of Orientalism include that it does just what Said warned about, that is, take a reductionist, essentialized, and global view of the West and the West's description of the Orient. To Said, a critical component of Orientalism is that the West speaks for the Orient. Interestingly, Muslim scholars accused Said, who is Christian, of speaking for Muslims (Varisco, 2007).

The Western Gaze Creating Others in Social Work

Cultural Competence

Through the vantage point of Orientalism, cultural competence in social work takes on additional dimensions and requires further critique. Cultural competence generally refers to developing cultural knowledge and awareness of 'different' cultures, so that one is able to work across these differences.

An important set of critiques found in the cultural competency literature highlights the lack of attention to power associated with specific cultures because culture is typically constructed as neutral (Dominelli, 2002; Pon, 2009). Viewing culture as neutral leaves little room for addressing and critiquing the role that culture plays in perpetuating systems of privilege and oppression. Thus, culture is seen as dichotomized between the poles of normative and 'othered' perspectives. That is, normative is defined as from the West, white, middle class, heterosexual, Christian, male, able bodied, and so on; white and the 'othered,' includes those who identify or are perceived as identifying as people of color; indigenous; noncitizen; gay, lesbian, bisexual, transgender, queer, asexual, intersex; Muslim; poor; disabled; and female, among others. Without a discourse of power associated with culture within cultural competence, an analysis of domination cannot occur (Dominelli, 2002; Park, 2005; Pon, 2009; Sakamoto, 2007).

Rossiter (2007) states, "Domination takes place when the dominator has the power to define the Other from within the dominator's frame of reference and interests" (p. 24). Using Said's (1978) concept of positional superiority, one can argue that by controlling the production of knowledge about culture in the East, the Occident, or West has not only asserted its viewpoint but also enshrined hegemonic power. Consequently, Orientalism is embedded within the notion and practice of cultural competence to the extent that social workers, particularly in the West, study, create knowledge about, and then intervene with 'cultured,' marginalized, oppressed, and

objectified groups. Cultural competence as often practiced in the West can be problematic for a number of reasons including that it encourages social workers to think about and engage in (cultural) differences at the individual and family level, much to the exclusion of macroforces. For example, a cultural competence approach for a Pākehā (Māori term for being of European decent) social worker working with a Māori family focuses on learning about te reo (Māori language) or Tikanga (general behavior guidelines for daily life and interactions in Māori culture) rather than learning about how racism and white supremacy operate in Aotearoa/New Zealand. Without a critical discussion of power, some of the issues associated with the act of being different from the norm – for example, marginalization, silencing, isolation and poverty – are also excluded from the discussion (Dominelli, 2004).

Finally, without a structural analysis of culture and, difference, there is little room to consider the positional privilege associated with social work, that is, the privilege bestowed on certain professions (including social work, mental health, law, medicine, academia) based on their gatekeeping roles and exercise of power. As Margolin (1997) suggests, lack of attention to positional privilege can lead social workers to view themselves as innocent helpers, and not consider the ways in which structural inequalities inform their practices and policies. Dominelli (2002), among others, has specifically urged social workers to focus on how knowledge of 'others' is constructed in and through practice to become proficient practitioners providing locally specific and culturally relevant services rather than learning about reified 'culture' when working with 'others.'

International Social Work

Whether social workers define international social work as 'global awareness' (Asamoah et al., 1997) or as more practice focused, social workers from the West or Global North interested in joining international human rights organizations and nongovernmental organizations (NGOs) operating in the East need to explore Orientalism and its impact on the construction of social problems, service delivery, and policy. Below, we discuss anti–sex trafficking approaches to highlight concrete examples of Orientalism surfacing within and across international social work issues and practices. Anti–sex trafficking approaches offer a vivid example.

Social works' involvement with individuals perceived as sexual slaves/fallen women dates back to the Progressive Era when social workers advocated against what was then called 'white slavery' (Chapkis, 1997; Wahab, 2002). While addressing slavery of any type can be seen as significant for social workers, historians have argued that the fear of 'white slavery' was based not on a large number of documented cases, but rather, societies' fears of cultural contamination due to changing immigration patterns, social anxieties about changing gender roles, sex, class, and race relations at the turn of the twenty-first century (Doezema, 2010; Hobson, 1987). Hobson (1987) points out that the presence of Asian women in brothels drew particular attention as Orientalist notions of Asian women as hypersubmissive fed assumptions that they were imported as sexual slaves. This view removed attention from the race- and class-based exploitation that the women may have experienced upon their arrival in the United States – exploitation that may have led them to the brothel (Hobson, 1987).

Similar Orientalist notions that regard certain Asian, indigenous, and Middle Eastern women as passive, exotic, and victims of their own cultures exist today in social work (Mukherjee and Chowdury, 2012). Evidence shows that sex trafficking exists in Canada and the United States. However, American social work efforts on sex trafficking often emphasize cultural and global 'others.' For example, a piece on human trafficking in the National Association of Social Workers' (Salett, 2006) *Human Rights and International Affairs Practice Update* presents seven examples of what human trafficking and slavery may look like. None of these include North Americans, focusing instead on individuals from India, Thailand, Sudan, Mexico, Moldovia, West Africa (*sic*), and Russia brought into the United States (Salett, 2006).

Orientalist stereotypes have empowered women from the West to define and attempt to solve the issue of sex trafficking for women in the East. The Coalition Against Trafficking in Women is an international NGO with strong local affiliates throughout the world. Its American founder, Kathleen Barry, once stated "In Thailand, religious ideology and patriarchal feudalism reduce the value of women's lives to that of sexual and economic property, which in turn validate prostitution" (1995: p. 182). Instead of recognizing and framing forced prostitution as a consequence of globalization, and complicated socioeconomic, and political forces, many anti–sex trafficking organizations in the West focus on the cultures of the East as the source of oppression for trafficked people. Ditmore and Thukral (2012) provide useful information for service providers who wish to help survivors of sex trafficking without causing additional harm. Their paper outlines the dangers associated with raids to save 'trafficked victims' and they argue for greater law enforcement accountability through human rights and survivor-centered approaches. Social workers can facilitate these approaches by building empowering relationships with immigrant communities and individuals involved in informal economies as well as advocating for greater accountability of labor laws to prevent debt bondage and other violations (Ditmore and Thukral, 2012: p. 147).

Teaching Orientalism in Social Work

A Foundational Critical Theory

Orientalism is taught in some social work curricula to support students' understanding of theory and social work practice. For instance, faculty members (Anaru Eketone and Shayne Walker) at the University of Otago (Dunedin, New Zealand) teach Orientalism to undergraduate social work students to help them understand the ways in which Māori have been constructed, studied, and represented by Pākehā. They also teach Orientalism as a means of supporting the ethical *use of self* in practice. Said's concepts of latent and manifest Orientalism are introduced in a Kaupapa Māori course to unpack the ways in which historical and contemporary constructions of Māori by Pākehā have marginalized Māori culture, tikanga, and te reo through government legislation (Walker, 1990). To gain a deeper understanding of Orientalism as a concept, students

are asked to select a piece of media that depicts Māori or Pacific Islanders in New Zealand. Their task is to critically analyze the selected piece through an Orientalist lens as a means of making 'othering' explicit and tangible. The goal of the assignment is to use the us/them binary of Māori/Pākehā as a method to explore multiple ways in which 'others' are constructed for the benefit of those who are doing the othering, albeit unconsciously at times. In this particular course, faculty members also teach Orientalism as a means of helping students consider their constructions and the gazes of 'others' more generally (not only Māori and Pākehā identities), better to understand privilege and oppression in social work practice across the micro-, mezzo-, and macrolevels.

By understanding Orientalism, students can become more aware of and confront the ways in which objectification, hegemony, and knowledge production are used as policy tools to reduce services and vilify oppressed peoples. Even the term 'welfare queen,' (meaning a woman who cheated the welfare system in such a way that she was able to amass wealth, while being lazy) was utilized to produce major restrictions to welfare policies during the 1990s in countries such as the United States, Canada, and New Zealand. These policy changes focused on controlling the behaviors of poor people rather than addressing poverty through local, regional, national and/or global policy changes, and social and political action. To be an effective policy advocate for individuals who are marginalized, social workers have to see how stereotypes influence the formation of policy and practice.

Conclusion

Orientalism provides a critical theoretical framework through which social workers can explore numerous issues including social workers: participating in systems of inequality, affirming commitments to social justice, and supporting those harmed by stereotypes and oppression, particularly those viewed as coming from the Orient. Conceptually, Orientalism helps social workers make sense of certain types of knowledge construction associated with cultural competence, negative perceptions of Arab and Muslim women, aspects of international social work, Islamophobia, and anti-Arab sentiments, speech, and practices. It can also support social workers in considering certain types of liberatory praxis for developing important counternarratives.

See also: Indigenous Child Welfare Practice: The Canadian Case; Indigenous Psychology; Indigenous Rights; Indigenous Social Work; Islamic Social Work: Hope and the Commitment to Change; Neoliberalism; Oppression; Orientalism.

Bibliography

Abdulhadi, R., Alsultany, E., Naber, N., 2011. Arab and Arab American Feminisms: Gender, Violence, & Belonging (Gender, Culture, and Politics in the Middle East). Syracuse University Press, Syracuse.

Abu-Lughod, L., 2002. Do Muslim women really need saving? Anthropological reflections on cultural relativism and its others. American Anthropologist 104, 783–790.

Anon, 1530. Foloweth a Lytell Treatyse Agaynst Mahumet and His Cursed Secte. EEBO Editions, ProQuest. Nabu Press, Charleston, SC.

Asamoah, Y., Healy, L.M., Mayadas, N., 1997. Ending the international-domestic dichotomy: new approaches to a global curriculum for the millennium. Journal of Social Work Education 33, 389–401.

Barry, K., 1995. The Prostitution of Sexuality. New York University Press, New York.

Bayoumi, M., Rubin, A., 2000. The Edward Said Reader. Vintage Books, New York.

Burney, S., 2012. Pedagogy of the Other: Edward Said, Postcolonial Theories and Strategies for Critique. Peter Lang Publishing Inc., New York.

Carter, J., 2006a. Palestine Peace Not Apartheid. Simon and Schuster, New York.

Carter, J., 12 December 2006b. Israel, Palestine, Peace and Apartheid. The Guardian. http://www.theguardian.com/commentisfree/2006/dec/12/israel.politicsphilosophyandsociety (retrieved 04.06.13.).

Chapkis, W., 1997. Live Sex Acts: Women Performing Erotic Labor. Routledge, New York.

Ditmore, M., Thukral, J., 2012. Accountability and the use of raids to fight trafficking. Anti-trafficking Review 1, 134–148.

Doezema, J., 2010. Sex Slaves and Discourse Matters. The Construction of Trafficking. Zed Books, London; New York.

Doka, K., 1989. Disenfranchised Grief: Recognizing Hidden Sorrow. Lexington Books, Lexington.

Dominelli, L., 2004. Culturally competent social work: a way towards international anti-racist social work? In: Guttierez, L., Zuniga, M., Lum, D. (Eds.), Education for Multicultural Social Work Practice. Council on Social Work Education, Alexandria, VA, pp. 281–294.

Dominelli, L., 2002. Anti-oppressive Social Work: Theory and Practice. Palgrave Macmillan, Hampshire, UK.

Erakat, N., 2011. Arabiya made invisible: between marginalization of agency and silencing of dissent. In: Abdulhadi, R., Alsultany, E., Naber, N. (Eds.), Arab and Arab American Feminisms: Gender, Violence, and Belonging. Syracuse University Press, New York, pp. 174–183.

Hobson, B.M., 1987. Uneasy Virtue: The Politics of Prostitution and the American Reform Tradition. The University of Chicago Press, Chicago.

If America Knew, n.d. Water in Palestine. [Online]. Available from: http://www.ifamericansknew.org (accessed 01.03.13.).

International Federation of Social Work, 2012. Statement of ethical principles [Online]. Available from: http://ifsw.org/policies/statement-of-ethical-principles/ (accessed 31.05.13.).

Jarmakani, A., 2010. Arab American feminisms: mobilizing the politics of invisibility. In: Abdulhadi, R., Alsultany, E., Naber, N. (Eds.), Arab and Arab American Feminisms: Gender, Violence, and Belonging. Syracuse University Press, New York, pp. 227–241.

Margolin, L., 1997. Under the Cover of Kindness: The Intention of Social Work. University Press of Virginia, Charlottesville, VA.

Mukherjee, D., Chowdury, D., 2012. What do the flyers say? Embedded "orientalist" constructions in social work study abroad programs in the United States. International Social Work. http://dx.doi.org/10.1177/0020872812441644 [First published online 13 June 2012].

Naber, N., 2012. Arab America. New York University Press, New York; London.

Northam, J., 2006. Jimmy Carter's Book Stirs Criticism, Complaint [Online]. Available from: http://www.npr.org/templates/story/story.php?storyId=6605934 (accessed 01.07.13.).

Park, Y., 2005. Culture as deficit: a critical discourse analysis of the concept of culture in contemporary social work discourse. Journal of Sociology and Social Welfare 32, 11–33.

Pon, G., 2009. Cultural competence as new racism: an ontology of forgetting. Journal of Progressive Human Services 20, 59–71.

Rossiter, A., 2007. Self as subjectivity: toward a use of self as respectful relations of recognition. In: Mandell, D. (Ed.), Revisiting the Use of Self: Questioning Professional Identities. Canadian Scholars' Press Inc., Toronto.

Runnymede Trust, n.d. Islamophobia a challenge for us all. [Online]. http://www.runnymedetrust.org/uploads/publications/pdfs/islamophobia (accessed 15.12.12.).

Salett, E.P., 2006. Human Trafficking and Modern Day-slavery. Human Rights and International Affairs Practice Update. National Association of Social Workers [Online]. Available from: http://www.naswdc.org/diversity/affirmative_action/humanTraffic1206.PDF (accessed 30.11.12.).

Said, E., 1978. Orientalism. Vintage Books, New York.

Said, E., 1992. The Question of Palestine. Vintage Books, New York.

Sakamoto, I., 2007. An anti-oppressive approach to cultural competence. Canadian Social Work Review 24, 105–118.

Santora, M., 2012. Women Is Charged with Murder as a Hate Crime in a Fatal Subway Push. New York Times [Online]. Available from: http://www.nytimes.com/2012/12/30/nyregion/woman-is-held-in-death-of-man-pushed-onto-subway-tracks-in-queens.html?_r=0.> (accessed 01.04.13.:A15).

Sayyid, S., Vakil, A., 2011. Thinking through Islamophobia: Global Perspectives. Columbia University Press, New York.

Smelser, N.J., Baltes, P.B., 2001. International Encyclopedia of Social and Behavioral Sciences. Elsivier Science Ltd., Oxford.

Thabet, A.A.M., Abed, Y., Vostanis, P., 2002. Emotional problems in Palestinian children living in a war zone: a cross sectional study. The Lancet 359, 1801–1804.

The Faculty Lounge, 2012. The Israeli-Palestinian Conflict: A Reading List [Online]. Available from: http://www.thefacultylounge.org/2012/11/the-israeli-palestinian-conflict-a-reading-list.html.> (accessed 05.04.13.).

Varisco, D., 2007. Reading Orientalism: Said and Unsaid. University of Washington Press, Seattle, WA.

Wahab, S., 2002. For their own good? Sex work, social control and social workers, a historical perspective. Journal of Sociology and Social Welfare 29, 39–57.

Walker, R., 1990. Ka Whawhai Tonu Matou: Struggle without End. Penguin, Auckland, New Zealand.

Relevant Website

http://www.ifamericansknew.org/ (accessed 01.03.13.) – If Americans Knew.

Orthodox Churches

Sergei Hackel,[†] Lewes, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, Orthodoxy, volume 16, pp. 10981–10985, © 2001, Elsevier Ltd.

Abstract

The Orthodox Church emerged with the Roman Catholic Church from the apostolic successions in early Christianity, and separated from the Roman tradition sometime between the 11th and 13th Centuries. It is the central Christian denomination in Eastern Europe, Africa, and the Middle East, with many national branches ranging from Russia to Ethiopia. Most were connected for long periods of time to the central Patriarch in Constantinople during the Ottoman Empire. The breakdown of the Ottoman Empire also led to the decentralization of the Orthodox Churches and migration has also created diaspora Orthodox communities.

The term 'Orthodox Church' is the normal term to describe a body that like the Roman Catholic Church, traces its origins to apostolic times. However, it ceased to share in the life of the Roman Catholic Church from the early Middle Ages. It would be misleading to give a single date for the separation of the two traditions. Certainly, 1054 should no longer be treated as the decisive moment of division. More important was the year 1204, when the fourth Crusade sacked the city of Constantinople and thereby demonstrated the 'otherness' of Western as compared to Eastern Christians. The very use of such terms suggests that there were cultural, not merely dogmatic, differences between these areas of the Christian world.

Meanwhile, the Orthodox have spread far beyond their one-time homelands in the 'East.' This was largely the Near East, as defined from a West European point of view. From the ninth century, Central and Eastern Europe were added to this notional 'East.' The modern move of Orthodox communities to other parts of the world tends to make 'Eastern Orthodox' an imprecise if not misleading term.

Some centuries before the separation of the Orthodox and Roman Catholic worlds occurred a different separation in the 'East.' The church council of Chalcedon (451) was the occasion on which a number of church bodies failed to accept the teaching, therefore also the fellowship of the established Orthodox Church in the Byzantine empire. It was (mistakenly) asserted that these were Monophysites. In other words it was presumed that they believed in Christ's single nature as God at the expense of his humanity. The Churches that were supposed to maintain this teaching belonged to the Near and Middle East, as well as to Africa. They are often described as non-Chalcedonian. Less precisely, they are defined as 'Oriental' Orthodox. In the course of the twentieth century, they also have moved far beyond their traditional locations. Unlike the 'Eastern' Orthodox, all of whom follow the Byzantine rite, the 'Oriental' Churches have their separate liturgical traditions.

[†] Deceased.

The 'Eastern' Orthodox at Home

The Former Ottoman Empire

World War I saw the disintegration of the Ottoman empire in which the Orthodox had been always granted a recognized if subordinate role as a people (*millet*).

Under Ottoman rule this *millet* had constituted a clearly segregated society within the Muslim realm. The Orthodox patriarch of Constantinople was treated as its principal administrator or ethnarch. It was possible to escape from the restraints imposed on the *millet* by conversion to Islam. Individuals who chose this way to advancement in the Muslim administration naturally lost the esteem of their former coreligionists. Yet those Orthodox who were executed after subsequent repudiation of Islam came to be venerated as neomartyrs.

The Orthodox Church fostered ethnic or national aspirations in Ottoman times. In effect it was a people's Church. All the more was it ready for new prominence in public life once Ottoman power had come to an end.

Thus the Church of Greece came into its own after the Greek uprising against the Turks (1821). In due course it was proclaimed the established Church of the new Greek kingdom (1852). Likewise the Serbian Church was able to proclaim its independence in 1879. This was in the aftermath of Serbian emergence from the Turkish yoke. Hence the country's Orthodox were eager to assert the unity of Church and state.

The Orthodox of Bulgaria aspired to such unity as well. But the Bulgarian Church claimed its independent status (1870) some years in advance of national independence (1878). Its aspirations were therefore condemned by the patriarchate of Constantinople as untoward, if not heretical (1872–1953). By contrast, the newly established Rumanian Church was to gain the patriarchate's approval in 1885. It had already been recognized by the Rumanian constitution as the country's national ('dominant') Church in 1866. Even after all the assaults of communism the Church could claim the allegiance of 87 percent of the population. The small Albanian Church was not to gain its independence until 1937. Unlike its Balkan partners, it could not claim to represent a majority of the population (no more than 25 percent).

These Orthodox populations could now swim with the tide and use their new opportunities to consolidate their dignity and self-esteem. But in the process they tended to despise their Muslim masters of the recent (even medieval) past. This was to motivate (politically motivated) attempts at ethnic 'cleansing' in Bosnia and Kosovo.

Memories of subjugation could also inspire a revanchist ideology among the Greeks. An aspiration to restore Constantinople as the center of an Orthodox empire animated religious and secular thinkers alike. The latter saw this as their 'Great Idea.' It created no problems until the disarray of the postwar years enticed the Greek authorities to put it into practice. Their ill-fated attempt to restore the one-time Byzantine territories of Asia Minor to the Greeks (1919–22) led to a disaster for the Greek-speaking Orthodox of the region, who were subsequently hounded thence by the victorious armies of Kemal Ataturk. The convention of Lausanne (1923) seemed to legitimize an earlier form of 'ethnic cleansing.' Significantly, this treaty determined that religious criteria should decide how peoples should be moved in the compulsory exchange of populations between Turkey and Greece. Only a small minority of Orthodox Greeks remained in Asia Minor, with the status of their former ethnarch correspondingly diminished. The patriarch of Constantinople was thought to retain merely a symbolic role. Certainly no ethnarch he, since the *ethnos* was displaced.

Only in the Church of Cyprus was the role of ethnarch to retain, even to regain, its former role. Cyprus had been taken from the Turks in 1878 and ceded to the British crown. But the ruling bishop continued to be the leader of the people. This was clearly revealed when Cyprus was seeking independence from the British. The Cypriot archbishop of the day (Makarios) was not only leader in the struggle to achieve this. In the immediate aftermath he was also to gain election as president of the new state (1960–77).

The Church of Cyprus received its independence in 431. The patriarchates of Africa (Alexandria) and the Middle East (Antioch and Jerusalem) are also rooted in the early medieval past.

Despite its once modest numbers, Alexandria claims pastoral responsibility for the whole of Africa. It has consequently acquired a sizeable flock in 60 different countries. This has modified its previously well-established character as a factor of Greek mercantile life.

Greek influence is strong among the leaders of the Jerusalem patriarchate. But it is an influence that is critically assessed by a predominantly Arab flock (perhaps 100 000 strong). By contrast, Antioch (in effect Syria and Lebanon) insists that its Arabic culture is central to its concerns. It is therefore well placed to promote dialogue with its Muslim neighbors.

The Former Russian Empire

Orthodoxy was well entrenched in Russian society by the end of the Middle Ages, and it had few competitors on native soil. But expansion of the Muscovite state from the mid-sixteenth century incorporated peoples with allegiance to other faiths. Muslim peoples were to remain proof against conversion. By contrast, the various pagan peoples of Siberia were often persuaded to join the state Church. Here, as elsewhere, conversion served as a function of colonization. More easily incorporated were the Orthodox of neighboring nations, who joined the Russian empire as the result of one treaty or another. Thus did Ukrainians merge with the Russian Orthodox Church (1686), while the ancient Church of Georgia was accommodated merely as a province (1811). After the church reforms of Peter I (1722), the Russian Orthodox Church was overtly a state church, with the emperor's lay nominee at the head of its subservient synod. It was taken for granted that the population would normally belong to this body. To this end, Orthodoxy was taught in the schools. An annual confession and communion were seen as civic duties. Until an imperial edict of toleration was published in 1905, non-Orthodox religious rivals were restricted in their public life, if not actively restrained; some kind of pluralism might have been envisaged.

But it was not to be. The Bolshevik party, which seized power in 1917, proclaimed separation of Church and state on January 23, 1918. It was soon evident that this involved control, if not suppression of the Church. Most of the Church's many institutions were abolished, and although Stalin's constitution proclaimed 'freedom of worship' (1936), hardly a hundred churches were left open by 1939. Mere membership of the Church involved risk of liberty and life. Among the millions who vanished into concentration camps were almost all the bishops. Meanwhile, anti-religious propaganda was deployed in the formation of *homo sovieticus*. 'Scientific atheism' was taught throughout the schools and universities of the USSR. Membership of the communist party required commitment to the atheist cause.

During World War II a degree of toleration of the Russian Orthodox Church was introduced by Stalin in an attempt to raise national morale (1943) in those parts of the USSR that had escaped German occupation. It was toleration that extended also to the Georgian Church, which had reclaimed independence in 1917. However, such toleration signified no lessening of controls by the party and the state. Rather were these to be extended to Orthodox institutions in the larger Soviet Union of the postwar years.

With equal rigor were controls imposed on Eastern Europe. This meant that the Orthodox church establishment of Rumania and Bulgaria were used to further Soviet interests, while atheism was promoted in society at large. *Mutatis mutandis* the Serbian Church was similarly affected in Tito's Yugoslavia. Yet only in Albania was official atheism to be taken to its logical conclusion: by 1967 religion was prohibited in any of its forms.

By contrast, the Greek Orthodox Church was left free of communist rule at the conclusion of a bitter civil war (1948–49). It was therefore able to continue its existence as established Church. Only one other matched it in this role, the Orthodox Church of Finland. But the two are curiously contrasted. Whereas the Church of Greece represents 98 percent of the population (hence problems of nominal membership), the Orthodox of Finland form hardly more than 1 percent of a predominantly Lutheran population.

The end of communist regimes in Eastern Europe and the former USSR has allowed the local Orthodox Church to reassert its place in secular society. But its potency and numbers are less than might have been expected in the aftermath of large-scale persecution. However, it is no longer barred from education or

from social work, or from public propagation of its faith. Lack of an official ideology has encouraged the substitution of popular (often nationalistic) Orthodoxy by spokesmen of the Church. The Russian form of this nationalism involves an anti-Western stance, with antisemitism one of its ingredients.

The 'Eastern' Orthodox Abroad

Missions

The missionary activities of St Cyril and St Methodios in Moravia (ninth century), and the associated translation of sacred texts from Greek into Slavonic, had legitimized the use of languages other than Hebrew, Greek, and Latin in church life. Later missionaries could therefore work unconstrained in a variety of cultures. Orthodox Christianity was able to bring literacy as well as other social tools to its clients. This was the case with the Siberian and Central Asian missions of the Russian Church. These had their extension to China (1695) and Alaska (1794). In due course there were Orthodox parishes founded in Japan (1868) and Korea (1897). These were modest-sized monocultural outposts of the Orthodox world and thus free of the conflicting loyalties that were later to burden the emigres of Western Europe, Australia, and North America.

The Orthodox Diaspora

Economic pressures and political oppression both led to the migration of Orthodox populations in the course of the twentieth century. Communist rule in Russia and Eastern Europe was among the most potent of the forces that propelled communities into exile. The number of North American immigrants from Russia was swelled by the Russian revolution and civil war of the post-1917 period. Communist rule in Eastern Europe after 1945 also brought Ukrainians, Rumanians, Serbs, Albanians, and Bulgarians to the new world. Meanwhile, Greeks had arrived in considerable numbers after their expulsion from Asia Minor in 1922–23. There were also Palestinians after 1948. In due course Orthodoxy came to be recognized as one of the 'major' religions of the USA.

However, there was never a united church administration. The attempt to form a largely English speaking 'Orthodox Church in America' (1970) did not meet with universal approval. Apart from a Greek archdiocese, which retained its links with the patriarchate of Constantinople, there was a bewildering variety of national jurisdictions, and not infrequently more than one of each. Many immigrants preferred to retain their languages and customs. Even (or especially) in exile their church life was a function of national identity. Similar fragmentation was to be observed elsewhere (Western Europe or Australia, for example).

Diaspora and Mission

However nation-oriented the diaspora, it could still have an effect on its non-Orthodox host environment. The Orthodox may not have sought to proselytize their neighbours. Nevertheless, conversions took place. A common cause might be intermarriage between partners of different faiths. But other less specific causes might also prompt the conversion of individuals or groups.

Some of these have gravitated towards Orthodoxy of their own accord: here is hardly 'mission' in the usual sense of the term. Thus the self-established Orthodox of Uganda and Kenya were assimilated into the patriarchate of Alexandria (1946), and without becoming honorary Greeks in the process. Rather the black Orthodox of Africa (one million strong) have come to outnumber the Greek Orthodox founders of the patriarchate (200 000).

The 'Oriental' Orthodox at Home

However, there were already large numbers of indigenous African Orthodox outside the boundaries of the Alexandrian patriarchate, members of the non-Chalcedonian ('Oriental' Orthodox) Churches of Egypt and Ethiopia.

The Coptic Orthodox Church

The Coptic (= Egyptian) Orthodox Church encompasses much of the non-Arabic part of the population, something like 10 percent of the total. This Church emerged as a distinct entity in the fifth and sixth centuries. But its distinctness was confirmed when the Arabs conquered Egypt in 639–642, since this freed it from Byzantine (= Chalcedonian) control. It has retained its character as the church of its nation. Members of the Coptic Church have experienced constraints from the Muslim establishment. Some of the discrimination exercised by extremists is not far removed from persecution. All the more noteworthy is the vigor of church life among the Copts. In a low-income community this finds its expression in service of the poor.

The Ethiopian Orthodox Church

The Coptic Church supervised the life of the Ethiopian Orthodox Church from the fourth to the twentieth centuries. Not until 1957 was the formal autonomy of the Ethiopian Church proclaimed, and this despite the number of its members (20 million), which far exceeds that of any of its 'Oriental' sister Churches. This accounts for almost half the population. The state encouraged loyalty to the Church when it was an established institution. But much loyalty survived the rigors of a Marxist revolution, not only anti-clerical but anti-religious though it was (1974).

The Syrian Orthodox Church

The 'Oriental' Churches are also represented in the Middle East. Members of the Syrian Orthodox ('Jacobite') Church are distributed between Syria, Lebanon, and Iraq. In terms of culture and membership, this Church was at its strongest in the Middle Ages, but its numbers have declined to little more than 100 000. In the search of anonymity and self-protection (many have fled post-Ottoman Turkey since the 1920s for lack of adequate safeguards) its members tend to emulate their Arab neighbors' way of life.

The Syrian Orthodox Church of South India

However, the Syrian Orthodox gain extra prominence from the existence of their co-religionists in South India, who have been linked with them for much of the period since 1665 through acceptance of a common head, the ('Oriental') Syrian Orthodox patriarch of Antioch. However, the Church of South India, with its 1 500 000 adherents, has its independent history since early Christian times, and its autonomy is fiercely defended by many of its members to this day. As an indigenous church it is well integrated into its milieu (Kerala).

The Armenian Apostolic Church

Armenian home territories are difficult to define. The Armenian republic (formerly a constituent republic of the USSR) is central to them, not least since the spiritual center of all Armenians (Etchmiadzin) is located there. It was founded in 301. However, only a third of the Armenian people reside in that republic. Many find themselves in a diaspora. But others are distributed within the confines of an adjacent administration, the catholicosate of Cilicia (Lebanon) (founded 1441).

Armenia adopted Christianity as its state religion in 301. The Armenian Apostolic Church (its name in later centuries) has remained the Church of the nation since, irrespective of changing frontiers. Armenians in the Ottoman empire gained protection by virtue of their own *millet* (1461). But they could still be subject to pressures and persecution. The last decades of the nineteenth century and the first of the twentieth witnessed massacres of the Armenian population by the Turks that reached their climax in the early years of World War I. Here was a genocide that claimed something like one and a half million lives, a third of the nation. Hence the dispersal of Armenians throughout the world. The persecutions confirmed Armenians in their commitment to their nation, their culture, and their Church.

The 'Oriental' Orthodox Abroad

Missions

The 'Oriental' Orthodox have been little involved in such political imperialism as might have prompted missionary outreach to nations other than their own.

Diaspora

Yet each of the 'Oriental' Churches has branches abroad. Thus Armenians have their communities in North and South America, in Russia, Western Europe, and the Middle East. As many are in the diaspora as in Armenia itself.

See also: Catholicism; Christianity, Origins of; Historical Thought and Historiography: Christian Tradition; Religion, History of; Religion, Sociology of; Religion: Evolution and Development.

Bibliography

Atamian, S., 1955. The Armenian Community. Philosophical Library, New York.
Beeson, T., 1982. Discretion and Valour: Religious Conditions in Russia and Eastern Europe. Collins, London.
Betts, R.B., 1979. Christians in the Arab East: A Political Study. SPCK, London.
Davis, N., 1995. A Long Walk to Church: A Contemporary History of Russian Orthodoxy. Westview Press, Boulder, CO.
Horner, N., 1989. A Guide to the Christian Churches in the Middle East. Mission Focus, Eikhart, IN.
Hourani, A.H., 1947. Minorities in the Arab World. Oxford University Press, London.
Lane, C., 1978. Christian Religion in the Soviet Union; A Sociological Study. State University of New York Press, Albany.
Pospielovsky, D., 1984. The Russian Church Under the Soviet Regime 1917–1982. St Vladimir's Seminary Press, Crestwood, NY.
Rinvolucri, M., 1966. Anatomy of a Church: Greek Orthodoxy Today. Burns and Oates, London.
Saloutos, T., 1964. The Greeks in America. Harvard University Press, Cambridge, MA.
Sarkissian, K., 1968. The Witness of the Oriental Orthodox Churches. Beirut.
Ware, T., 1993. The Orthodox Church. Penguin Books, Harmondsworth, UK.

Oversharing: The Eclipse of Privacy in the Internet Age

Ben Agger, University of Texas-Arlington, Arlington, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Oversharing refers to disclosing intimate details of one's personal life, especially via electronic means of communication. The causes and consequences of oversharing are examined as a sociology of self-disclosure (or presence) is developed. Of central issue is the fading boundary between private and public life, which is a distinguishing feature of postmodern capitalism. Lacking close friends, people use social media, texting and tweeting, to stay connected. Children and adolescents inhabit a nighttime cyberworld of their own, protesting their overburdening by parents and teachers by disclosing their needy selves. Oversharing might be protest against performativity – the expectation that 'everything counts' and must be graded.

Oversharing (see Gould, 2008; Agger, 2012) is Emily Gould's (2008) term for disclosing too much about oneself. By now, merely half a decade later, oversharing has entered the vocabulary, as both a verb and noun ('that's an overshare…'). It is also called TMI – too much information. It is characteristic of communication in the Internet age (Turkle, 2011), when people are spared social embarrassment by engaging in electronically mediated interaction. Oversharing helps constitute what I call a 'pornographic public sphere', in which intimacy goes public. In this article, I explore social and psychological dimensions of the erosion of boundaries between personal and public lives, a story that will necessarily involve the Internet and cell phones. I am writing on the 10-year anniversary of the inception of Facebook, which began as a Harvard dating site and now boasts a billion users worldwide, with over 30% of them being 35 or over.

Genealogy and Definitions of 'Oversharing'

Oversharing occurs through social media such as Facebook. As of 2014, people may overshare by texting, tweeting, Skyping, and posting porn on the Internet. They may also use Internet dating sites and craigslist to overshare, spin, and engage in self-inflation.

The prefix 'over' implies that people are doing more than sharing. They are sharing too much. This degree of sharing violates Goffman's (1959) distinction between front stage and back stage behavior – the former in role, the latter after one has removed one's costume and makeup and returns to an authentic selfhood. Oversharing violates the perceived boundary between private and public selves, diminishing the realm of privacy, secrecy, and intimacy. While different cultures define the private domain differently, Goffman's insight that all cultures have some sense of privacy is relevant here.

Lasch (1979) wrote of a 'culture of narcissism' in Western countries, characteristic of a stage of capitalism after WWII. His analysis derives from the Frankfurt School, which postulated an 'eclipse' (Horkheimer, 1974) of the self, the autonomous subject capable of mediating between himself and the world. The world saturates and overwhelms us, whittling us down. This makes us needy and self-absorbed, which, clinically, might be described as narcissism. The so-called narcissistic personality disorder involves low self-esteem, which might seem counter-intuitive. I argue that one of the features of alienation in the early twenty-first century is the loss of self-esteem, which leads people to seek solace in largely electronic relationships with others.

Oversharing is a baring of the soul that solicits intimacy and reciprocity. For example: 'I did 10 shots at the bar last night', or 'I slept with a stranger I met there', and even 'I got arrested for DUI'. It involves making the private public to nonintimate 'friends', in Facebook parlance. Of course, what counts as oversharing depends on context. Disclosing intimacies with one's therapist is different from posting them on Facebook, for one's hundreds of 'friends'. Discussing sex with one's partner is different from discussing sex with one's work colleagues. At issue is the context-bound nature of boundaries, which shift and evolve.

One notices a widespread deboundarying in the post-modern stage of capitalism (Agger, 2002; Jameson, 1994). Perhaps the postmodern moment is defined by the suspension of boundaries, especially between the private and the public. People were more discreet in an earlier modernity, and not only because they were Puritanical. There was general agreement that self-disclosure should be boundaried by a sense of propriety, but it was also true that people tended to have a few close friends and intimates with whom to share. The early social theorists, Durkheim, Weber, and Marx, all identified a tendency for modern life to leave people with few strong ties. Durkheim's term anomie and Marx's term alienation were basically synonymous. The decline of face-to-face intimacy in our Internet age corresponds to deboundarying, which leads people to seek intimacy on social-media sites, message board, dating sites, chat rooms, and via text messaging. In other words, the decline of intimate community ('Gemeinschaft') leads to making the private public, dissolving boundaries. Perhaps anomie involves undersharing, where, in a pre-Internet moment, a person lacks intimacy in everyday life and also cannot find it online.

Adults who use Facebook may wonder whether a 1000 Facebook friends are equivalent to a few 'real' friends. That is a tempting question, but not the appropriate one. It already decides in favor of offline versus online relationships as a source of meaningful intimacy. Kids raised in the Internet and Facebook age 'do' social relationships by moving back and

forth between the offline and the online, especially where they can text, which in a way breaches offline and online communication and community by making one's interlocutor present. For the young, texting is more immediate and intimate than voice-to-voice conversation, allowing for greater self-disclosure.

Age and generation are slippery. Divorced parents may seem child-like to their worldly children who watch them return to adolescent dating rituals. In this article, I draw a boundary between people raised before the Internet and people raised since the Internet – pre-digital and digital generations. As with all boundaries, this one blurs as elders learn to negotiate life on the screen and are positioned by it.

Young people and even their parents use FB to arrange movie dates, which they pursue offline. They also use FB to sustain more distant friendships with kids in another county, or with whom they have occasional meetings such as at a quarterly district tennis tournament. Kids use many, and often simultaneous, portals, posting, texting, hanging out. Perhaps the only medium they do not use (for oversharing) is the old-fashioned telephone, which asks of them too much instantaneity and thus exposes them to risk as they cannot stall, edit, mediate, filter. Every parent is exasperated when their kids do not answer their phones, but instead engage in lengthy text exchanges, which they probably conduct while driving or doing their homework.

But Daniel Miller (Miller, 2013) has recently found that 16–18 year olds are leaving Facebook in favor of Twitter, regarding Facebook as a stodgy domain for their parents and other elders. Hill (2009) offers a different perspective on why he left Facebook, largely involving time use and inauthenticity. Miller's work is part of the Global Social Media Impact Study (http://gsmis.org/) which tracks trends in Internet and social media use.

The Needy Self

The needy postmodern self who overshares in order to seek intimacy and reciprocity is characterized by a certain plasticity, or displacement of the ego. Some address personality disorders such as borderline and narcissistic in terms of the loss of self, or of its decentering. One tries to be everything to everyone, manipulating impressions in order to solicit care and love. This is eminently possible using the Internet, which allows one to create one's profile in fanciful and falsifying ways. Everyone has heard stories of dating site profiles that transform the chubby self into svelte and take years off of one's age. This sort of spinning is characteristic of Internet self-assemblage, which also encourages a kind of oversharing as one fills in the broad brush strokes of self-description with the fine details of one's preferences – food, leisure, sex. The Internet calls forth an overindulgence in self-description. Before she bemoaned the privatizing tendencies of the Internet, Turkle (1995) noted that the Internet offered people a 'life on the screen' that allows them to experiment with different identities, something that is important for young people. Indeed, one could define adolescence as the period of growing a self, by trial and error.

Oversharing includes not only these intimate self-descriptions, not to mention self-made amateur pornography, but an opinionated announcement of one's views of the world, especially posted on message boards. As a runner, I have consulted a popular running message board, letsrun.com. This is a board for serious and competitive runners, not for weekend warriors who run a Thanksgiving turkey trot. The letsrun message board is described as 'world famous,' which is perhaps just Internet-era grandiosity.

On the message board, one will find lengthy threads about training theory and the recent performances of elite athletes. The Oregon virtuoso, Galen Rupp, who won silver in the London 2012 Olympic games and recently ran a 3:50 mile indoors, is often discussed, as is the European marathon guru, Canova. But there are other threads on politics, Obama, taxation, sexuality, sexual orientation, and everything under the sun. These posts are sometimes vitriolic, parochial, and often poorly crafted, probably written by young men who run and who also participate in this Age of Opinion. Nevertheless, there is an addictive quality about the message board for serious runners, as we seek mystical wisdom about getting faster, even as we cannot turn away from the train wreck of unsolicited opinion about current events. The boundary between sharing and oversharing seems tenuous here.

Oversharers take license to opine about everything, whether on Facebook or the letsrun message board. An overtly political message board, redstate.com, offers commentary on the Kenyan socialist in the White House as well as on the Newtown tragedy. Crackpotism abounds, which is manifest as a message-board oversharing about all matter of ephemera, from Kim Kardashian to a possible assault-rifle ban. Of course, one's person reasonableness may be another person's lunacy – again the problem of boundaries.

One might distinguish between populism, which involves a certain leveling, and a public sphere in which people produce informed and learned discourse. This is not an argument for academic credentialism, which, as Russell Jacoby (1987) has pointed out, can lead to self-satisfied obscurantism. I applaud his concept of the public intellectual, who writes big books about big issues in accessible prose. Jacoby might agree with me that our Age of Opinion, manifested in untutored oversharing on message boards and in blogs, is inimical to public discourse, and hence democracy. Academization is not an adequate alternative to manic message-board postings which risk regression to the premodern – a kind of flat-earth Babbitism represented by Sarah Palin and other magical thinkers on the right who probably cost Romney the 2012 Presidential election with their bigotry and bad reproductive science.

Gould and I use 'oversharing' with tongue in cheek; whether one discloses too much depends on the standards in use. The term captures the trend toward greater self-disclosure in public, especially in a cybersphere in which people are 'alone together' (Turkle, 2011). What I call a pornographic public culture, arose with (1) the Internet, (2) a larger media culture in which 'reality' becomes entertainment, and (3) sexualization of public discourse and imagery. I hasten to add that pornography is effect, not cause; it is an example of oversharing, of rendering the private public. One can defend filmic sexuality as an overcoming of Puritanism, just as one can defend 'sharing' intimacies with others.

It is not the prurient nature of oversharing that makes it potentially unhealthy. It is the way that oversharing

deboundaries the self by baring all that is at issue. The Frankfurt School's thesis of the eclipse of subjectivity is confirmed by this self-deboundarying, even though they aptly provide no clear guidelines for where boundaries should be drawn. Postmodernists might respond that the Enlightenment concept of a stable self or subject was always a fiction; hence, they prefer to rename the self a 'subject position'. But this endorses alienation and refuses utopia, a good society in which healthy people keep some of themselves hidden while entering public discourse.

Undersharing

Just as one can disclose too much on the first date – sexual preferences, drinking habits, salary – one can disclose too little, which I term undersharing. Undersharing may shade into deception, for example when one does not acknowledge one's marital status or alcoholism. Between undersharing and deception is spinning, when one embellishes one's exercise routine, educational background, love of pets, and sunset walks on the beach. We embellish, spin, deceive, and conceal because people, especially daters, are awash in a mating market that, in the Internet era, can be cruel, as Vitzthum (2007) explains in her *I Love You, Let's Meet*.

Dating sites are a proving ground for my concerns about oversharing. We find self-constructed personal profiles which overshare as well as contextualize, spin, and deceive. Perusing date sites also reveals appropriate sharing, especially by women who, in general, seek relationships whereas Internet-dating men seek sex. There is blurring, especially as people live fast lives and just want to hook up. Divorced parents who work may use dating sites and craigslist to find casual sexual partners, even as we find gender differences in self-presentation, with women likelier than men to tell the truth, but appropriately, without oversharing.

Oversharing may be a relational strategy designed to earn trust and thus advance the relationship. But in telling too much too soon, mystery declines and people move quickly from date to date, relationship to relationship. Vitzthum (2007), using her personal Internet-dating experience, describes coming home from a date and going online in order to search the dating site or sites for a newer, better option. The restlessness in Internet dating mimics and reproduces restlessness in all contemporary life, as people are perpetually unfulfilled in the relational marketplace.

Oversharing is done efficiently as people learn not to waste time on probably fruitless relationships. They are upfront about 'deal breakers' – smoking, children, pets, a certain age range, height and weight, religion. People who engage online seem picky, because they want to move on quickly and because they are accustomed to deception and duplicity. Alienation begets alienation as people are rejected and learn to reject, without allowing relationships to develop serendipitously and in the fullness of time. The reciprocal revelation packed into the first drive-by date at Starbucks, after dating-site picture sharing has been accomplished, is not viewed as oversharing but as the efficient use of time. Here, undersharing would be inefficient, requiring tedious future dates to discover that being a vegan Buddhist is a deal breaker.

Undersharing may respond to structural changes in American society. It may simply be that people have fewer close friends with whom to share, as Brashears (2011) has argued. He tracks the decline of close friends over the past 25 years from three to two. He doubts that Internet friendships, which constitute weak ties in most cases, compensate for the loss of close friends. But others (Hampton et al., 2009) paint a less bleak picture and argue that Internet-based relationships are ample and diverse. Both approaches tell part of the story. I think that the reality is complex because offline and online relationships blur and intermingle, especially for young people, who are adept at moving between the two.

Finally, some people undershare because they are shy, as Cain (2012) has argued, Jung allows for introversion as the anchor of a common personality type (or archetype), as reflected in results on the Myers-Briggs personality test. Many of us bookish types have archetypes that begin with the letter 'I', for introverted. Introversion may be a reasonable response to a world in which the boundary between personal and public is fading; people draw into their shells because they seek to protect their psychic space under assault by electronic invasion.

Boundaries and Personality Disorders

There is something fundamentally different about trading intimacies face to face than online. It is less dehumanizing when staring into the other's eyes, which have a mind and voice of their own. Too little is risked when saying 'I love you' via texting, and risk is essential if the reward of reciprocation is to be gained. Of course, for diminished selves in our fast, postmodern capitalism, little can be ventured emotionally because people are emotionally embattled. Terms such as 'borderline' and 'narcissistic' psychiatrize what is really a social condition: not liking oneself sufficiently. Low self-esteem leads to online intimacy, which involves quick revelation and equally quick rejection or disconnection. Being 'unfriended' or having one's friend request ignored on Facebook both sting.

And so is being ignored, which is the cruelest cut in Internet interaction. Attractive women are overwhelmed with interest by male suitors (or bedmates) on Internet dating sites. There is an erotic economy which governs how many, if any, responses an attractive woman will be able to offer the dozens who message her everyday. This reproduces the pre-Internet dating minuet, where men do the chasing. Women can pick and choose, which gives them a certain upper hand. They learn to be disdainful, and ignored or rejected men lick their wounds, perhaps learning misogyny in the meantime. Why is there an Internet-era erotic economy? Because people work a lot and cannot find time to date, unless they date people from work. And because many are unlucky in love, with divorce and separation littering the landscape. In these contexts, people have less time to 'do' relationships than they did in slower times, when relationships emerged gradually from shared activities and did not move too quickly.

Here, oversharing is an expedient, a quick means of testing the water. But it bears the cost of rejection and being ignored, which only deepens alienation. Whence alienation? From the loss of community, which sociologists have been tracking since

the nineteenth century through our era of suburbs, divorce, workaholism, etc.

Alienation can be traced to Marx's 1843 formulation. The concept has had a Hegelian-Marxist career, becoming reification (Lukacs, 1974) and domination (Frankfurt School). Weber talked about a bureaucratic iron cage, but without Marx's utopian optimism about overcoming it. Being 'alone together', captured in the image of a silent family all texting during dinner, is the latest version of alienation. Oversharing is a desperate attempt to overcome alienation and disconnection, restoring friendship, intimacy, and self-esteem. The Frankfurt School deepened Marx by examining the psychic roots and manifestations of economic alienation. They realized that the self in late capitalism was losing the ability to stand apart from the world and appraise it critically. Instead, people were merely vessels into which commodities and meaning are poured. Before people could engage in the critique of ideology, understanding capitalism and formulating plans to change it, they had to work on themselves by reboundarying their psyches and everyday lives. Retreat from the world would afford the opportunity to read, write, reflect, and imagine utopia.

This is not to indulge in a technological determinism whereby the Internet, cell phones, and the media culture are to be blamed for everything from porn to school shootings. I have argued in *Texting toward Utopia* (2013) that kids are using these prosthetic communication technologies to write their ways out of their misery. Only we elders fail to recognize that this is the most literary of ages, even as youngsters do not turn in their term papers on Shakespeare on time. They text and post up a storm, sharing secret code with other alienated youngsters off the grid of adult expectations and discipline. Seen this way, oversharing has the dialectical potential to undo itself as people seeking intimacy actually find it – love, relationships, affinity groups, social movements. The Internet and texting can be used to organize Occupy Wall Street as well as to post photos of last night's drunken binge at the bar. Adult conceptions of privacy and propriety may need to be rethought at a time when young people 'share' more globally, quickly, and explicitly than we did, especially if young people's prison talk about their oppression by school and parents leads to the mobilization of the young as a new and powerful minority group. Jacobs (2008a,b, 2009) joins me in valuing kids' writing and reading it for tea leaves of critique, resistance, and imagination.

I am, of course, describing the black and white children's movements of the 1960s. The 1962 Port Huron Statement of the Students for a Democratic Society SDS was Xeroxed, stapled, and passed around. It went viral, but only slowly, in bits and pieces. One wonders whether the Port Huron Statement or Malcolm's autobiography, if posted on the Internet, would have gone viral or sunk into the vast Sargasso Sea of the Web. It seems that only viral YouTube videos suggest post-textual manifestos that constitute a cultural politics of resistance and reconstruction. Rock, marijuana, hippies were pre-political, Woodstock making way for marches on Washington and perhaps also to Days of Rage. Yesterday's hippies are today's slackers, kids so turned off by school, teachers, parents, career that they spend their time online, gaming, texting, blogging. This is a kind of Luddite protest against kids' regimentation and more-than-40 hour work week, including homework and college-bound resume building.

Overburdened kids overshare. They are insomniac, too wired to sleep. They claim the night as their own, becoming vampires. They send text messages in a bottle out to sea, hoping for connection and then perhaps community, or at least a date. According to Pew data, the American youngster sends an average of 109 texts a day.

Prison code has always united isolated prisoners. Kids are imprisoned in schools that they hate, and set adrift as latchkey children in families in which everyone has too much to do and is in a separate orbit. Fast families breed virtual children (Agger and Shelton, 2007). Who is to say that sharing is excessive when there is so much pain in the land? School shooters intend suicide as well as others' death. Oversharers produce what I call disturbia – signs and signals of their angst and possibly of their self-destructive intentions. The Columbine shooters, Harris and Klebold, let everyone know: Their friends bought them ammo and videoed them shooting in the woods. One of the fathers found a homemade pipe bomb. The police arrested them for robbing a van. Similarly, Cho at Virginia Tech acted out in his English class, provoking one professor to avoid him, in fear, and another to try to reach him. Gun control would have done little in either case, but love and care would have helped buffer mortal aloneness. Harris and Klebold were not bullied; that is urban legend (Cullen, 2009). They were relatively popular, but it seemed that few adults spent time with them, giving them boundaries and inspiration. The adults are always too busy.

The whole subject of online bullying is an intriguing one. Hinduja and Patchin (2008, 2012) have found that there is relatively little online bullying. Everyone hates a bully, but their findings, like Cullen's, above, about Columbine, suggest that kids use the Internet more for consensus building and community formation than for bullying, although, certainly, gossiping and oversharing abound.

Oversharing, thus, may be a cry for help, especially when kids can use social media to announce their angst. Some of this might be posturing, a way of profiling themselves as cool and diffident. Or it might be real, a reflection of their aloneness and anxiety. School weighs heavily on many kids, as do parental expectations. Adults, especially teachers, are experienced as authoritarian, and oversharing is a reaction, a way of venting and building solidarity.

There may be a blurry boundary between narcissism – 'look at me' – and political mobilization. Cries for help may become calls to arms as kids take back their lives, and their schools. It may be difficult to imagine Facebook, Twitter, and texting as platforms of social change, but they could be, as we see in the Occupy Wall Street and Arab Spring movements. I am hesitant about this because information, communication, and entertainment technologies impose themselves on users and turn self-revelation and compulsive surfing into a nearly inevitable fate. My students check their phones before, during, and after class. They check for messages, mail, alerts. They may be conducting multiple conversations, with complicated threads that never seem to quit. It is difficult for many to be off the grid, especially when one's interlocutors may be employers and parents. Time has expanded into 'iTime', elastic and never-ending (Agger, 2011). People fill it

with their electronic devices, which, unlike radio and television, allow them to interact and write.

This most literary of ages sees billions of words texted and tweeted every day. This should tell us that people have a protean desire to author their own lives, to be writers and thus citizens. The young text and tweet toward utopia, even if they do not spell out their utopias, their ideal worlds, having been focused on fact-based education and not thinking outside the box, let alone thinking about a new world. Lasch's argument about a culture of narcissism is important here. He is noticing that we teach and train people to be inward-looking, focusing on 'self' and not also focusing on others, let alone the whole society. We pour out our hearts in blogs, not in political manifestos, even though the diminished, needy self is a casualty of a postmodern capitalism in which the psyche is invaded and recruited as an insatiable consumer.

So critics like Turkle can easily interpret oversharing as a symptom of narcissism. I could not disagree. However, I think texting and tweeting are dialectical phenomena, perhaps having the potential to become liberating means of Socratic dialogue, public discourse, and political action. Of course, some already use social media to change the world, as the Arab Spring, Occupy Wall Street, and Obama's campaigns indicate. This is a mixed picture: The Internet permits the self to be diverted into consumerism and entertained harmlessly, while the Internet also makes cyberdemocracy possible.

Performativity Provokes Slacking

I have argued that the inward-looking self has become a political factor as well as a casualty of an invasive society that breaks down all boundaries. So what is narcissism? Although the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 2000) is removing narcissistic personality disorder as an official psychiatric condition, everyone understands narcissism to be a personal problem, if not also a social problem. It signifies lack of empathy and perhaps hubris as self becomes central. Clinicians rightly understand that narcissistic and borderline personality disorders involve not self-inflation but low self-esteem. On this view, self-orientedness signals not a Napoleonic impulse, a will to power (although narcissistic leaders can seem that way) but deep self-doubt that is diminished by grandiosity. That does not work and self-esteem remains low. Narcissists who run organizations can quickly seem sociopathological. But there is a narcissist in all of us as most of us are insufficiently satisfied by a world which we experience as harsh and threatening. It is no wonder that kids hide their phones in their hoodies and text each during the interminable school day, when they face performativist expectations and too much boring school work and homework (Kralovec and Buell, 2001).

'Performativity' is the tendency to evaluate and measure all human behaviors. Kids from K through 12, and even beyond, into college, are graded numerous times a day, on quizzes, homework, tests, classroom comportment, hallway and lunchroom behavior. They are then evaluated in their extra-curricular activities, from orchestra and cheerleading to sports. Kids reasonably resist performativism by retreating to a cyberworld in which kids form solidarity by using prison codes in their texting, tweeting, and posting: 'PIR (parent in room)'. People resist performativity, without fully understanding that it is bad for them.

Young and older people have resisted performativism since the Luddite reaction to industrialization starting in the early 1800s. Fordism describes Henry Ford's era of mass production, beginning a century later, made possible by the assembly line. Technology makes for Frederick Taylor's time and motion studies that measure the worker's output and efficiency. Capitalism drives performativity because it wants to maximize profit, which flows from human labor power. The only limit to production is the need for sleep. Sleep deprivation boomerangs as children and their parents awaken bleary-eyed, unable to do on Thursday what they did on Wednesday.

Young people, especially in schooling, are members of a pre-labor force that acquires knowledge, skills, and work habits suited to later marketplace work performance and engages in pre-market work itself, especially homework. This is both preparation for adult work and work itself, which diverts kids from utopia. Whereas Ford-era scientific management evaluated and expedited work performance of adults, today childhood and adolescence have been attenuated as schooling has fallen under sway of performativity, beginning even before elementary school. I think childhood performativity serves capital by preparing a future labor force (Bowles and Gintis' classic 1976 argument) and also diverts kids from non-utilitarian activities in the moment. Adults, themselves alienated, want children to share their pain and hence they pile on the homework and other evaluated activities.

Kids, then, work a more-than-40 hour work (pre-labor) week. They experience insomnia, consume caffeine and speed, learn to hate school, and resent adult authority. Slacking is a response to compulsive performance evaluation and it involves sharing and oversharing among slackers who are, in effect, Luddites, albeit Luddites armed with 3G and 4G cellular technologies.

Oversharing as resistance? Perhaps, inasmuch as saying too much reflects a frenetic desire to connect and seek comfort. How else are we to explain the fascinations of Facebook and message boards, which suck up many hours of attention and diversion? Television is privatizing, whereas sharing is not. Even playing online video games connects, as gamers know. For we elders, that these are not (yet) face-to-face connections matters, but for our progeny avatars stand in for the real thing. And presence is conjured out of reciprocal messaging, composed quickly and in computer quick, deploying slang, acronyms, and emoticons. 'I luv u' is met with 'LOL' and a knowing wink. This is real. Whether it is 'as' real as an in-person exchange depends on one's generation. For the young and Internet adept, 'reality' is not a continuum that comes in less and more. It is the presence of the other, which manifests offline, online, and somewhere in between, perhaps in memory and imagination. Kids who text during school meet after, but do not necessarily discuss what was said. There is a sliding into role that is nearly effortless, much more for the Internet generation than for mine, the boomer generation. Much more can be left unsaid. People who did not grow up with the Internet and texting may find it difficult to cross the boundary beween online and offline relationships, whereas their children

effortlessly relate to each other using multiple portals, including face-to-face interaction. This is why authors of my age risk denigrating online communication as inauthentic, when, for young people, online and offline blur to the point of near identity.

Self-disclosure, then, is a tricky business, requiring us to rethink some standard sociological notions of the stable self, interaction, communication. We need e-sociologies addressing e-life and iTime. We sociological elders are not innocent in these ways. We already know email and cell phones, and many of us text. Many Facebook users are between 18 and 40, although there has been rapid growth of social media participation among people over 40. We rationalize this as a matter of convenience, a way to stay in touch with far-flung family and friends. But all this is imprinting, as we gradually replace face–face and voice-to-voice communication with virtual presence. This asks of us the ability to conjure presence out of absence, taking the role of the other, as symbolic interactionists such as Mead (1962) term it, in order to get on with a new, virtual construction of reality.

That people used to have three close friends, now diminished to two, is bad for intimacy and exacerbates anomie. But the missing friend has been replaced, or augmented, with one's virtual friends who might become 'real' friends and even partners and spouses. E-sociology is sufficiently postmodern that we will not distinguish between the real and the virtual, as I just did. Everything is real, with presence emerging from the screen as well as the gaze and touch. Perhaps that is the twenty-first century predicament: Spilling our gut to a screen, behind which hide many insatiable others.

A Sociology of Presence

A sociology of self-disclosure, then, is a sociology of presence and absence. The Internet and cellular communication call forth these new intellectual approaches. We are virtual selves (Agger, 2004) and we need new frameworks to understand a pixelated, global existence in which downtime nearly disappears. Oversharing might be defined as what you say when you cannot conjure the presence of an other, but only, perhaps, of a generalized other – people like me who might be listening. One cannot imagine blushing, judging, condemning. This is perhaps the difference between writing in a diary when you think your mother would not find and read it, and anticipating that she will read your confession about a first kiss or cutting class. Oversharing occurs when the person does not anticipate that she will be heard and acknowledged; it is a cry for attention. Ignoring and being ignored constitute alienation in the Internet age. This is not to say that telling too much is a viable strategy for being noticed. But people do not always think clearly when they are drunk, awake too late, driving, heartbroken. How many of us have read emails fired off in the heat of the moment? We want to look away.

Presence, then, buffers the self's mortal aloneness, which may be simply to say that we all need friends and intimates. Freud knew this, as did Durkheim. Marx noticed that capitalism pits us against each other, making sharing difficult. Friendship is portable; it can cross the online/offline boundary. One texts one's wife: "Your hair looks good today." Later, one might caress it. Or, "Plz bring home some oranges if you go to the store." Or, even, "here's my latest essay; fyi."

Jurgen Habermas (1989) argues that one of the positive features of early-modern, industrializing Europe was a vibrant public sphere, with coffee houses, newspapers, and libraries. Drawing on images of public life in the Athenian city-state, he applauds a 'public sphere' in which people come together to trade ideas and conduct democracy. Today, the public sphere, rendered digital, invades people's private hours, homes, leisure, even their psyches. Habermas never anticipated oversharing; he linked sharing and democracy. A Habermasian agenda today would include means of protecting the private sphere, in which people turn off, and turn away from, lives on the screen that exert a certain disciplining force (Foucault, 1977).

See also: Computer Mediated Communication; Information Society; Internet and Culture; Interpersonal Trust across the Lifespan; Online Dating; Social Media; Youth Culture, Sociology of.

Bibliography

Agger, Ben, 2002. Postponing the Postmodern: Sociological Practices, Selves, and Theories. Rowman & Littlefield, New York.
Agger, Ben, 2004. The Virtual Self: A Contemporary Sociology. Blackwell, Malden, MA.
Agger, Ben, 2011. iTime: labor and life in the smartphone era. Time & Society 20 (1), 119–136.
Agger, Ben, 2012. Oversharing: Presentations of Self in the Internet Age. Routledge, New York.
Agger, Ben, 2013. Texting Toward Utopia: Kids, Writing, and Resistance. Paradigm, Boulder.
Agger, Ben, Shelton, Beth Anne, 2007. Fast Families, Virtual Chidren: A Critical Sociology of Families and Schooling. Paradigm, Boulder, CO.
American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders, fourth ed, Washington, DC.
Bowles, Samuel, Gintis, Herbert, 1976. Schooling in a Capitalist America: Educational Reform and the Contradictions of Economic Life. Basic, New York.
Brashears, Matthew E., 2011. Small networks and high isolation?: a reexamination of American discussion networks. Social Networks 33 (4), 331–341.
Cain, Susan, 2012. Quiet: The Power of Introverts in a World that Can't Stop Talking. Crown, New York.
Cullen, Dave, 2009. Columbine. Twelve, New York.
Foucault, Michel, 1977. Discipline and Punish: The Birth of the Prison. Verso, New York.
Goffman, Erving, 1959. The Presentation of Self in Everyday Life. Doubleday, New York.
Gould, Emily, 25 May 2008. Exposed. New York Times.
Gould, Emily, 2010. And the Heart Said Whatever. Free Press, New York.
Habermas, Jurgen, 1989. The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society. MIT Press, Cambridge, MA.
Hampton, Keith, Goulet, Lauren Sessions, Ja Her, Eun, Rainie, Lee, 2009. Social Isolation and New Technology. Pew Internet and American Life Project. November 4. http://www.pewinternet.org/Reports/2009/18–Social-Isolation-and-New-Technology.aspx.
Hill, David W., 2009. Reflections on leaving facebook. Fast Capitalism 5.2. www.fastcapitalism.com.
Hinduja, Sameer, Patchin, Justin W., 2008. Cyberbullying: an exploratory analysis of factors related to offending and vicitimization. Deviant Behavior 29 (2), 129–156.
Hinduja, Sameer, Patchin, Justin W., 2012. School Climate 2.0: Cyberbullying and Sexting One Classroom at a Time. Sage Publications, Thousand Oaks, CA.
Horkheimer, Max, 1974. Eclipse of Reason. Seabury, New York.
Jacobs, Gloria, 2008a. We learn what we do: developing a repertoire of writing practices in an instant messaging world. Journal of Adolescent and Adult Literacy 52 (3), 203–211.

Jacobs, Gloria, 2008b. Saying something or having something to say: attention seeking, the breakdown of privacy, and the promise of discourse in the blogosphere. Fast Capitalism 4.1. www.fastcapitalism.com.

Jacobs, Gloria, 2009. Adolescents and Instant Messaging: Literacy, Language, and Identity Development in the 21st Century. VDM Verlag, Saarbucken.

Jacoby, Russell, 1987. Last Intellectuals: American Culture in the Age of Academe. Basic, New York.

Jameson, Fredric, 1994. Postmodernism, or, the Cultural Logic of Late Capitalism. Duke University Press, Durham, NC.

Kralovec, Etta, Buell, John, 2001. The End of Homework: How Homework Disrupts Families, Overburdens Children, and Limits Learning. Beacon, Boston, MA.

Lasch, Christopher, 1979. The Culture of Narcissism. Norton, New York.

Lukacs, Georg, 1974. History and Class Consciousness. Merlin, London.

Mead, George H., 1962. Mind, Self and Society. University of Chicago Press, Chicago.

Miller, Daniel, 20 December 2013. Facebook's so Uncool, but Its Morphing into a Different Beast. http://theconversation.com/facebooks-so-uncool-but-its-morphing-into-a-different-beast-21548.

Turkle, Sherry, 1995. Life on the Screen: Identity in the Age of the Internet. Simon & Schuster, New York.

Turkle, Sherry, 2011. Alone Together: Why We Expect More from Technology and Less from Each Other. Basic, New York.

Vitzthum, Virginia, 2007. I Love You, Let's Meet: Adventures in Online Dating. Little, Brown, Boston, MA.

P

Pacific Islands, Archaeology of

Patrick V Kirch, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Pacific Islands, comprising Polynesia, Micronesia, and Melanesia, were explored and settled by humans over a long period beginning in Near Oceania during the late Pleistocene and ending in eastern Remote Oceania *c.* AD 1000–1200. This article reviews recent archaeological, human biological, and linguistic evidence for these complex migrations. Two other major topics of interest to Pacific archaeologists – the evolution of complex societies and human impacts on island ecosystems – are also considered.

The Pacific Islands, also known as Oceania, have been subdivided traditionally into three main geographic regions: Melanesia, Micronesia, and Polynesia. This tripartite division, based largely on crude racial categories, was first proposed by the early-nineteenth-century French explorer Dumont d'Urville. Nonetheless, these geographic categories continue to be used by many anthropologists. Following Green (1991), archaeologists now stress a distinction between *Near Oceania* in the west (comprising New Guinea, the Bismarck Archipelago, and the Solomon Islands) and *Remote Oceania* (which includes all of island Melanesia southeast of the Solomons, along with Polynesia and Micronesia). This latter division recognizes that Near Oceania was colonized by humans in the Pleistocene, whereas the widely dispersed islands of Remote Oceania were not discovered and settled until after *c.* 2000 BC.

Archaeological research in Oceania has a long history, but modern efforts emphasizing stratigraphic excavations did not begin until after World War II (Kirch, 2000), when Edward Gifford of the University of California undertook pioneering excavations in Fiji and New Caledonia, revealing stratified sequences of culture change. Almost all of the main archipelagoes and islands of Oceania have now been subject to archaeological investigations, so that the broad outlines of the region's prehistory are known.

Early Human Settlement of Near Oceania

The oldest known occupation sites in Near Oceania are radiocarbon dated to *c.* 36 000 years ago (the late Pleistocene) on the large island of New Guinea and on New Britain and New Ireland islands in the adjacent Bismarck Archipelago (Allen, 1996). At several times during the Pleistocene, New Guinea was joined to Australia as a single land mass (known as Sahul); it seems probable that human entry into and expansion throughout this vast Australasian region occurred rapidly. Late Pleistocene sites in the Admiralty Islands, New Ireland, and Buka (Solomons), all would have required open ocean transport, indicating the presence of some form of watercraft (possibly rafts, bark boats, or dugouts) (Irwin, 1992).

Early human colonists in Near Oceania were hunters-and-gatherers, who exploited tropical rainforests as well as inshore marine resources. Long-distance communication and exchange are indicated by the movement of obsidian between islands, as well as by the translocation of animal species such as the spotted cuscus (*Phalanger maculatus*). The material culture of these Pleistocene inhabitants of Near Oceania consisted largely of informal flake and core lithics although distinctive 'waisted blades' resembling large axes, as well as tanged core tools, are known from several localities. Although there is no direct evidence of the linguistic affiliation of these early Near Oceanic populations, we can infer that they spoke a variety of proto-languages ancestral to the later Papuan or Non-Austronesian languages of the region.

By the early Holocene period (after 8000 BC), there is archaeobotanical evidence for domestication of tree, root, and tuber crops (such as the *Canarium almond* and various aroids) within Near Oceania. Archaeological evidence suggests that cultivation of swamplands at Kuk in the Highlands of New Guinea commenced as early as 7000 BC (Golson, 1988). These archaeological indications confirm long-held ethnobotanical claims that Near Oceania was one of the several independent centers for the origins of tropical horticulture, including such important crops as taro (*Colocasia esculenta*), sugar cane (*Saccharum officinarum*), and the *Australimusa* bananas.

Austronesian Expansion and Lapita

During the early Holocene, speakers of Papuan or Non-Austronesian languages had expanded as far east as the southeastern Solomon Islands, which marked the eastern limit of human expansion into Oceania. Beginning around 2000 BC, a major expansion or diaspora of people speaking

languages belonging to the Austronesian language family commenced (Blust, 1995). The Austronesian family includes about 1200 extant languages spread from Madagascar to Easter Island. The immediate homeland of the Austronesians has generally been regarded as consisting of the island of Taiwan and, quite likely, adjacent coastal areas of mainland China. The ability of early Austronesians to disperse rapidly has been attributed to their invention of the outrigger sailing canoe (Pawley and Ross, 1993). The Austronesians were horticulturalists who transported root, tuber, and tree crops via their canoes, along with breeding stocks of domestic pigs, dog, and chickens.

The Austronesian diaspora rapidly encompassed the major archipelagoes of island Southeast Asia. One branch of Austronesian speakers expanded along the north coast of New Guinea into the Bismarck Archipelago. This branch is known to linguists as Oceanic, and the Oceanic languages (numbering about 450 modern languages) include most of those spoken throughout the Pacific. The great exception is New Guinea (along with nearby New Britain, New Ireland, and Bougainville), where roughly 750 *Non-Austronesian* languages are spoken. The Non-Austronesian or Papuan languages do not constitute a single unified language family and are highly diverse, reflecting the great antiquity of human settlement in Near Oceania.

Archaeological evidence for the initial Austronesian dispersal into the Pacific comes from both western Micronesia (the Marianas and Palau archipelagoes), and from the Bismarck Archipelago. In western Micronesia, early sites contain red-slipped pottery, some of which is decorated with lime-filled, impressed designs (Carson et al., 2013). These sites, along with radiocarbon-dated sediment cores exhibiting signals of human presence (e.g., high influxes of microscopic charcoal resulting from anthropogenic burning), suggest that humans settled the Marianas and Palau no later than 1500 BC, and possibly as early as 2000 BC.

In the Bismarck Archipelago, the arrival of Austronesian speakers has been correlated with the appearance of a distinctive suite of sites, also containing pottery with lime-infilled decorations, but with motifs made largely by a technique of dentate stamping, in which several kinds of toothed combs were impressed into the damp clay. These sites and the associated artifacts (such as *Tridacna* shell adzes and *Trochus* shell fishhooks, as well as ornaments) represent the earliest known phase of the Lapita cultural complex, dating to *c.* 1500–1300 BC (Gosden et al., 1989; Kirch, 1997, 2000; Spriggs, 1997). Early Lapita sites were frequently hamlets or villages consisting of houses elevated on posts or stilts, situated over tidal reef flats or along shorelines, such as the Talepakemalai site on Eloaua Island in the Mussau group, first occupied around 1500–1300 BC. Excavated plant and animal remains indicate a mixed economy with horticulture and marine exploitation. Substantial quantities of obsidian, chert, pottery, shell artifacts, and other materials were exchanged between Lapita communities (Kirch, 1997).

Correlations among archaeological assemblages, language groups, and human biological populations are often complex, and need not be isomorphic. However, a strong consensus has emerged among scholars in several disciplines that the initial phase of the Lapita cultural complex can be correlated with the *Proto-Oceanic* interstage of the Austronesian language family. Moreover, genetic evidence (such as mitochondrial DNA (mtDNA) and nonrecombining Y-chromosome (NRY) DNA) supports the view that the appearance of Lapita sites reflects a substantial population intrusion into the Bismarck Archipelago, deriving out of island Southeast Asia (Hill and Serjeantson, 1989; Friedlaender, 2007). At the same time, these Proto-Oceanic speakers undoubtedly had considerable interaction (cultural, linguistic, and genetic) with the indigenous Non-Austronesian speaking populations who already occupied the Bismarck region in the mid-Holocene. For this reason the Lapita cultural complex is seen as an outcome of a complex set of cultural interactions summarized by Green's 'Triple-I' model of *intrusion*, *integration*, and *innovation* (Green, 1991).

Human Colonization of Remote Oceania

Beginning *c.*1200 BC, Lapita pottery-makers expanded rapidly beyond the Solomons and – for the first time in human history – into the southwestern archipelagoes of Remote Oceania: Vanuatu, the Loyalty Islands, New Caledonia, Fiji, Tonga, and Samoa. Numerous radiocarbon-dated archaeological sites document that Lapita sites appear in all of these archipelagoes no later than *c.*900 BC. Recent high-precision uranium series dating on coral abraders at the Nukuleka site on Tongatapu pinpoints the arrival of Lapita pottery-makers in this Western Polynesian archipelago at 2838 ± 8 BP (Burley et al., 2012).

We have already noted that the westernmost islands of Micronesia were colonized directly out of island Southeast Asia by Austronesian speakers *c.* 2000–1500 BC. Around 2000 years ago, Oceanic speakers, who made plainware pottery (a late form of Lapita) and who used shell adzes, fishhooks, and other implements, founded settlements on several volcanic islands of central Micronesia (Chuuk, Pohnpei, and Kosrae). The atolls of the Marshall Islands were also colonized at this time.

A distinctive Ancestral Polynesian culture emerged in the Western Polynesian archipelagoes of Tonga and Samoa (along with adjacent Futuna and 'Uvea) out of its Lapita roots during the course of the first millennium BC, through processes of local adaptation and gradual cultural change. Archaeologically, this transition is marked first by changes in the pottery complex, with the loss of the earlier dentate-stamped decoration and development of a distinctive Polynesian Plainware ceramic tradition, as well as by changes in the adz kit, now manufactured from dense basalt. Historical linguists have also traced the development of a distinct Proto-Polynesian language interstage (ancestral to all contemporary Polynesian languages) marked by at least 1300 lexical innovations (Marck, 2000). Applying a 'triangulation approach' that integrates evidence from archaeology, historical linguistics, and comparative ethnology, Kirch and Green (2001) extensively reconstructed this Ancestral Polynesian culture, including aspects of subsistence economy, social organization, and ritual and calendrics.

The final stage in the human settlement of the Pacific Islands began in the first millennium AD, with the Polynesian dispersals eastward out of Western Polynesia (the Tonga and Samoa regions). The timing of Polynesian dispersals into the

central and eastern Pacific (or what is known as Eastern Polynesia) has been a matter of considerable debate among archaeologists (Spriggs and Anderson, 1993; Wilmshurst et al., 2011). Archaeological work, in the 1950s and 1960s, in several Eastern Polynesian islands led to the view that this region had been settled as early as AD 300, but the validity of most of the earlier radiocarbon dates has now been called into question. Recent redating of samples from older excavations combined with new finds in the Marquesas and Mangareva Islands strongly suggests initial human incursion into central Eastern Polynesia between *c*.AD 850 and 1100. The earliest dates for the Hawaiian archipelago are in the range of AD 940–1130 (Athens et al., 2014). The twin large, temperate islands of New Zealand appear to have been the last to have been colonized by Polynesians, around AD 1250 (Anderson, 1989; Davidson, 1984). Critical to the success of this unprecedented diaspora was the double-hulled sailing canoe, capable of carrying 40 to 60 people on voyages lasting one month or longer (Irwin, 1992).

The possibility of Polynesian contact with the Americas has long intrigued scholars. That the Polynesians reached South America and returned is indicated by preserved remains of the sweet potato (*Ipomoea batatas*), a South American domesticate, in several prehistoric Polynesian sites (Hather and Kirch, 1991). Radiocarbon-dated bones of domestic chicken (*Gallus gallus*) from the El Arenal site in Chile have also been claimed to reflect pre-Columbian Polynesian visits to South America (Storey et al., 2007, 2008), although this remains controversial.

Because it was the last sector of Remote Oceania to be settled, and because its populations represent a single radiation or diaspora, Polynesia constitutes a *monophyletic* cultural and linguistic group. Thus, Polynesia has often been regarded as an ideal region for testing models of cultural differentiation from a common ancestor (e.g., Kirch, 1984; Kirch and Green, 2001).

Development of Complex Societies

When the eighteenth-to-nineteenth-century European voyages of discovery inspired by the Enlightenment encountered Pacific Island societies, they frequently found large, dense populations organized into complex, hierarchical sociopolitical formations. With populations often numbering into the tens or hundreds of thousands, such societies had two to three decision-making levels, and hereditary leaders who enjoyed elite privileges and status markers. Anthropologists classify such sociopolitical formations as chiefdoms, and indeed, the Polynesian chiefdoms are often considered the archetypal model (*see* Chiefdoms, Archaeology of).

The origins, development, and elaboration of Pacific Island chiefdoms have been a major topic of archaeological research (e.g., Davidson, 1984; Kirch, 1984; Sand, 1995; Spriggs, 1997). Based on linguistic and archaeological evidence, early Austronesian societies were characterized by some degree of internal ranking (especially between senior and junior branches of a descent line), but were probably heterarchical rather than *hierarchical* in structure. However, heterarchic competition (in such social arenas as marriage and exchange, as well as competition for land) between social groups provided the basis for true hierarchy (and eventually, in the largest societies, class stratification) to emerge.

Archaeologists have identified several factors and processes that were significant in the rise of Oceanic chiefdom societies. Many of these were closely linked, and should not be considered unicausal variables. For example, population growth leading to large, high-density populations can be identified as a necessary, but probably not sufficient, cause underlying sociopolitical complexity (Kirch, 2000; Sand, 1995). The human populations of the volcanic islands typically reached densities of between 100 and 250 persons per km^{-2} prior to European contact, resulting in intense competition for arable land and other resources. Such conditions encouraged centralized, hierarchic control, as well as providing incentives for militaristic aggression.

A second process linked to population growth and to increased hierarchy was intensification of production, including agricultural systems and other forms of production or resource extraction, as well as economic specialization (e.g., in pottery production and trade). On many islands, large-scale irrigation works or dryland field systems were developed during late prehistory. Although population increases may have initially spurred intensification, once in place such intensive food production systems provided a means for surplus extraction by chiefs and other leaders, thus encouraging hierarchy. Often culturally marked as tribute, such surpluses were the economic foundation of an emergent elite, including not only hereditary chiefs, but priests, warriors, craft specialists, and others.

Ideology likewise played a key role in Pacific Island societies, with the elite cadres of the larger and most complex societies actively employing ideological control as a means of legitimation. The origins of Oceanic ritual systems can be traced back to common Austronesian concepts of the sacredness of ancestors; these concepts later became elaborated as cults in which the highest chiefs were seen as directly descended from powerful gods, and hence essential to the continued well-being of the society at large. Archaeologically, the rise of elite-dominated ideological systems is particularly reflected in monumental architecture, of which the most impressive examples are the giant statue-bearing temples of Easter Island (Van Tilburg, 1994), and the site of Nan Madol on Pohnpei (Kirch, 2000). Other forms of monumental architecture, however, are ubiquitous throughout Pacific Islands chiefdoms. Even when monumental architecture is absent, material signs of ideological control can be quite striking, as in the multiple sacrificial interments associated with the burial of Roy Mata, a chief of Vanuatu (Garanger, 1972).

Finally, competition, conflict, and warfare also characterized many of the complex societies of the Pacific, especially following the rise of large and dense populations (*see* Conflict and War, Archaeology of: Behavior and Social Organization). Archaeologically, warfare is marked by a diversity of kinds of fortifications, such as the pallisaded *pa* volcanic cones and headlands of New Zealand, or the ring-ditch fortified villages of Fiji. Another, more gruesome signal of the levels that inter-societal aggression reached on some islands is cannibalism (or para-cannibalistic treatment of enemies,

such as dismembering, roasting, and non-funerary discard of skeletal remains). Although some anthropologists have expressed skepticism regarding European voyagers' accounts of cannibalism in the Pacific, there is now direct archaeological evidence for cannibalistic or para-cannibalistic practices in late prehistory on Easter Island, the Marquesas, New Zealand, Mangaia, and Fiji (Kirch, 2000).

While most of the complex, stratified societies of Oceania would be regarded as chiefdoms, it has recently been argued that the society of Hawai'i, and possibly that of Tonga as well, constituted archaic states prior to European contact (Kirch, 2010; Hommon, 2013). In Hawai'i, a political system with divine kings (*ali'i akua*) who ruled over a society marked by endogamous elite and commoner classes emerged on the large islands of Hawai'i and Maui in the early seventeenth century. This shift from complex chiefdom to archaic state entailed a number of essential changes including the alienation of land from commoner lineages and the imposition of a system of ritualized tribute collection. Similar transformations may have occurred in late pre-contact Tonga, although the Tongan case requires further study.

Human Impacts to Island Ecosystems

The islands of Remote Oceania, due to isolation and related factors, provide model conditions for studying the effects of human colonization and land use on pristine ecosystems. Interdisciplinary research among archaeologists and natural scientists over the past three decades has amplified our understanding of such human–ecosystem interactions (Kirch and Hunt, 1997).

Because of the substantial open-ocean distances isolating them from continents as well as other islands, and the difficulty of dispersal to islands, prior to human arrival, oceanic ecosystems were typically characterized by high species-level endemicity, but lower diversity in higher-level (generic and family) taxa; lowered competition; survival of archaic forms; and *vulnerability* to disturbance from outside agents. Larger vertebrates such as marsupials (wombats and cuscus) and rats, snakes, frogs, and most lizards were restricted primarily to Near Oceania, with only a handful of species declining in numbers eastwards to Fiji and Samoa. The reef and marine resources of Pacific Islands also display a west-to-east decline in species diversity. Throughout most of Remote Oceania, pre-human vertebrate faunas were dominated by birds (including many flightless forms which had evolved in situ from flighted ancestors). Prior to human arrival, these bird populations lacked large vertebrate predators, and presumably also a typical predator avoidance strategy. They must have been extremely easy prey for the first humans to step foot on islands.

When humans first arrived in Remote Oceania, they typically found the islands to be forested, and inhabited by a range of largely endemic species, dominated by birds, along with invertebrates such as landsnails and insects. Oceanic peoples possessed a successful colonization strategy that allowed them to exist on isolated islands by (1) transporting in their sailing canoes stocks of horticultural crop plants, along with domestic pigs, dog, and chickens (rats came along, presumably as 'stowaways'); (2) clearing areas of rainforest for garden land; and (3) intensively exploiting the abundant natural avifaunal and marine resources.

This colonization strategy had several consequences for island ecosystems, all of which are increasingly well documented through both archaeological and paleoenvironmental indicators. Forest clearance on many islands is signaled in changing pollen spectra from sediment cores, with tree taxa rapidly giving away to ferns and grasses; also characteristics are sharp increases in microscopic charcoal influxes, indicating human-induced burning, in most cases probably associated with shifting cultivation. On some islands, forest clearance led to increased erosion rates, along with alluviation of valley bottoms or along coastal plains. The exploitation of natural resources is particularly evident in the zooarchaeological assemblages from early settlement sites, which are characterized by high numbers of land and seabirds, many of them representing now extinct or extirpated species (Steadman, 2006). A dramatic case of avifaunal extinctions on Pacific Islands is that of the *moa*, a group of 13 species of large, flightless birds which became totally extinct in New Zealand during the brief period of Polynesian occupation (Anderson, 1989).

Cumulative effects of human actions on islands led to irreversible changes, such as dramatic declines in biodiversity, and the conversion of natural rainforests to intensively managed, anthropogenic landscapes. The consequences for human populations themselves were undoubtedly mixed. The replacement of natural ecosystems with intensive food production systems enabled the growth of large and dense human populations. At the same time, reduction or depletion of natural resources, coupled with the necessity for intensive land use, encouraged highly complex sociopolitical systems which at times competed fiercely for control of land and the means of production.

See also: Australia and New Guinea, Archaeology of; Chiefdoms, Archaeology of; Food Production, Origins of; Horticultural Societies; Migrations, Colonizations, and Diasporas in Archaeology; Oceania, Sociocultural Overviews: Melanesia; Oceania, Sociocultural Overviews: Polynesia and Micronesia; States and Civilizations, Archaeology of.

Bibliography

Allen, J., 1996. The pre-Austronesian settlement of island Melanesia: implications for Lapita archaeology. Transactions of the American Philosophical Society 86 (5), 11–27.

Anderson, A., 1989. Prodigious birds: moas and moa hunting in prehistoric New Zealand. Cambridge University Press, Cambridge.

Athens, J.S., Reith, T.M., Dye, T.S., 2014. A paleoenvironmental and archaeological model-based estimate for the colonization of Hawai'i. American Antiquity 79, 144–155.

Blust, R., 1995. The prehistory of the Austronesian-speaking peoples: the view from language. Journal of World Prehistory 9, 453–510.

Burley, D., Weisler, M.I., Zhao, J.-X., 2012. High precision U/Th dating of first Polynesian settlement. PLoS 7 (11), e48769.

Carson, M., Hung, H., Summerhayes, G., Bellwood, P., 2013. The pottery trail from Southeast Asia to remote Oceania. Journal of Island and Coastal Archaeology 8, 17–36.

Davidson, J.M., 1984. The Prehistory of New Zealand. Longman Paul, Auckland.

Friedlaender, J.S. (Ed.), 2007. Genes, Language, and Culture History in the Southwest Pacific. Oxford University Press, Oxford.

Garanger, J., 1972. Archéologie des Nouvelles-Hébrides. Publication de la Société des Océanistes No. 30, Paris.

Golson, J., 1988. The origins and development of New Guinea agriculture. In: Harris, D., Hillman, G. (Eds.), Foraging and Farming: The Evolution of Plant Exploitation. Unwin Hyman, London.

Gosden, C., Allen, J., Ambrose, W., Anson, D., Golson, J., Green, R., Kirch, P.V., Lilley, I., Specht, J., Spriggs, M., 1989. Lapita sites of the Bismarck Archipelago. Antiquity 63, 561–586.

Green, R.C., 1991. Near and remote oceania: disestablishing "Melanesia" in culture history. In: Pawley, A. (Ed.), Man and a Half: Essays in Pacific Anthropology and Ethnobiology in Honour of Ralph Bulmer. The Polynesian Society, Auckland.

Hather, J., Kirch, P.V., 1991. Prehistoric sweet potato (*Ipomoea batatas*) from Mangaia Island, Central Polynesia. Antiquity 65, 887–893.

Hill, A.V.S., Serjeantson, S.W. (Eds.), 1989. The Colonization of the Pacific: A Genetic Trail. Clarendon Press, Oxford.

Hommon, R.J., 2013. The Ancient Hawaiian State. Oxford University Press, Oxford.

Irwin, G., 1992. The Prehistoric Exploration and Colonisation of the Pacific. Cambridge University Press, Cambridge.

Kirch, P.V., 1984. The Evolution of the Polynesian Chiefdoms. Cambridge University Press, Cambridge.

Kirch, P.V., 1997. The Lapita Peoples: Ancestors of the Oceanic World. Blackwell Publishers, Oxford.

Kirch, P.V., 2000. On the Road of the Winds: An Archaeological History of the Pacific Islands before European Contact. University of California Press, Berkeley.

Kirch, P.V., 2010. How Chiefs Became Kings: Divine Kingship and the Rise of Archaic States in Ancient Hawai'i. University of California Press, Berkeley.

Kirch, P.V., Green, R.C., 2001. Hawaiki, Ancestral Polynesia: An Essay in Historical Anthropology. Cambridge University Press, Cambridge.

Kirch, P.V., Hunt, T.L. (Eds.), 1997. Historical Ecology in the Pacific Islands: Prehistoric Environmental and Landscape Change. Yale University Press, New Haven.

Marck, J., 2000. Topics in Polynesian Language and Culture History. Pacific Linguistics, Australian National University, Canberra.

Pawley, A.K., Ross, M., 1993. Austronesian historical linguistics and culture history. Annual Review of Anthropology 22, 425–459.

Sand, C., 1995. "Le Temps d'Avant": La Préhistoire de la Nouvelle-Calédonie. L'Harmattan, Paris.

Spriggs, M.J.T., 1997. The Island Melanesians. Blackwell Publishers, Oxford.

Spriggs, M.J.T., Anderson, A., 1993. Late colonisation of East Polynesia. Antiquity 67, 200–217.

Steadman, D.W., 1995. Prehistoric extinctions of Pacific Island birds: biodiversity meets zooarchaeology. Science 267, 1123–1130.

Steadman, D.W., 2006. Extinction and Biogeography of Tropical Pacific Birds. University of Chicago Press, Chicago.

Storey, A.A., Ramirez, J.M., Quiroz, D., Burley, D.V., Addison, D.J., Walter, R., Anderson, A.J., Hunt, T.L., Athens, J.S., Huynen, L., Matisoo-Smith, E.A., 2007. Radiocarbon and DNA evidence for a pre-Columbian introduction of Polynesian chickens to Chile. Proceedings of the National Academy of Sciences 104, 10335–10339.

Storey, A.A., Quiroz, D., Ramirez, J.M., Beavan-Athfield, N., Addison, D.J., Walter, R., Hunt, T.L., Athens, J.S., Huynen, L., Matisoo-Smith, E.A., 2008. Pre-Columbian chickens, dates, isotopes, and mtDNA. Proceedings of the National Academy of Sciences 105, E99.

Van Tilburg, J.A., 1994. Easter Island: Archaeology, Ecology and Culture. British Museum Press, London.

Wilmshurst, J.M., Hunt, T.L., Lipo, C.P., Anderson, A.J., 2011. High precision radiocarbon dating shows recent and rapid initial human colonization of East Polynesia. Proceedings of the National Academy of Sciences USA 108, 1815–1820.

Pain, Health Psychology of

Herta Flor, Medical Faculty Mannheim, Central Institute of Mental Health, Heidelberg University, Mannheim, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Acute pain is an adaptive phenomenon that signals impending danger. It is a multidimensional, sensory, and emotional experience and must be differentiated from the physiological process of nociception. When pain becomes chronic, it develops into a disorder in its own right where psychological factors dominate. These include nonassociative (sensitization) and associative (respondent, operant conditioning) learning and memory processes, as well as cognitive and emotional factors. A biobehavioral perspective of chronic pain is based on the view that biological, psychological, and social factors contribute to the experience of chronic pain and they must be addressed in assessment and treatment.

Basic Terms

Pain is an adaptive phenomenon that signals impending danger to the body. Pain can, however, also be maladaptive, for example, in states of chronic pain, and thus can itself become a significant disorder. Epidemiological studies show that 80% of the population experience recurrent pain, with about 19% of the population being significantly disabled by it (Reid et al., 2011). Although pain has for a long time been viewed a purely sensory phenomenon or an epiphenomenon of a medical disorder, this view changed during the last 40 years of the twentieth century. It has been recognized that pain is a psychobiological experience, with emotional aspects being as important as sensory pain characteristics. Thus nociception, the physiological process of a noxious signal being transmitted from the periphery to the brain, has been differentiated from the experience of pain, which encompasses all aspects including psychological, social, and cultural factors. The International Association for the Study of Pain (IASP) definition takes this shift from a biomedical to a biobehavioral or psychobiological perspective into account by characterizing pain as an "unpleasant sensory and emotional experience associated with actual or potential tissue damage or described in terms of such damage" (Merskey and Bogduk, 1994). This definition was prompted by the fact that pain may often be experienced in the absence of identifiable objective pathology. A major change in the traditional view of pain was the gate control theory of pain proposed in Melzack and Wall (1965). More important than its physiological aspects, which have not all been confirmed, was the conceptual model behind it. The gate control theory stated not only that pain could be modulated by both afferent and efferent factors at the level of the spinal cord, but also emphasized that pain has sensory–discriminative, motivational–emotional, and cognitive–evaluative components, and thus it assigns an important role to psychological factors. A core feature of this model is the integration of peripheral stimuli with psychological variables such as mood, attention, or cultural experience in the perception of pain. Thus, the gate control theory has been instrumental in abolishing the dichotomy of psychogenic versus somatogenic pain – psychological and physiological factors always interact in the production of the experience of pain and are not mutually exclusive causes of pain.

Classification of Pain

The common distinction between acute and chronic pain – referring to pain of at least 3–6 months duration and/or pain that exceeds the normal time for healing in the case of an acute injury – is useful, because chronic pain is often maladaptive and needs special attention. While, acute pain such as pain related to medical procedures or childbirth usually leads to anxiety and apprehension, chronic pain tends to be associated with helplessness, depression, and irritability, as well as interference with family function, work, or social interaction (Flor and Turk, 2011). In addition to the large number of workdays that are lost as a consequence of it, chronic pain is the leading cause of invalidity in the age-group of under-50s and causes enormous costs for the health care system (Gatchel et al., 2007).

The IASP classification of pain is organized in five general categories: (1) the bodily system affected, (2) time characteristics, (3) intensity of pain, (4) duration of pain, and (5) its presumed etiology. This classification lacks reliability especially in respect of the last characteristic. In addition, it neglects psychological factors that are only coded with respect to the presence of a psychiatric disorder or within the category 'psychophysiological dysfunction.' The International Classification of Diseases of the World Health Organization as well as the Diagnostic and Statistical Manual of the American Psychiatric Association used to adhere to the distinction between psychogenic–somatogenic or physical–somatoform pain, which must be viewed as obsolete, given the large number of research findings emphasizing the importance of psychological factors in any type of pain, both acute and chronic (Gatchel et al., 2007). Thus, in its more recent versions, both classification systems focus on the symptoms and adverse consequences of chronic pain rather than its putative psychogenic origin and thus favor a biobehavioral or biopsychosocial view.

The biopsychosocial model of pain defines chronic pain as a complex response that can be described on the verbal–subjective, the motor–behavioral, and physiological levels. Pain can be based on nociceptive input, but nociception is not a prerequisite of the experience of pain, which can be an exclusively central phenomenon, but always has physiological antecedents and consequences (Flor and Turk, 2011). The multiaxial classification of pain as proposed by Turk and his

colleagues (cf Turk and Rudy, 1988) categorizes pain patients along both the somatic and the psychosocial dimensions. This classification has yielded three subgroups of patients in the psychosocial domain characterized as adaptive copers, interpersonally distressed, and dysfunctional, with significant differences in the course of the illness and response to treatment. A new multidimensional taxonomy proposed by the American Pain Society (Fillingim et al., 2014) focuses on core diagnostic criteria, common features, common medical comorbidities, neurobiological, psychosocial, and functional consequences, as well as putative neurobiological and psychosocial mechanisms, risk factors, and protective factors.

The Role of Learning in Chronic Pain

Both associative and nonassociative learning processes as well as social learning have been found to be of fundamental significance for the development of chronic pain (Apkarian, 2008; Sandkühler, 2000). The repeated application of painful stimuli usually leads to reduced responsivity, i.e., habituation to the painful stimulation. In many states of chronic pain, sensitization rather than habituation occurs due to changes both at the level of the receptor and the central nervous system. Sensory information accelerates habituation and reduces activation caused by surprise, insecurity, and threat. This mechanism may underlie the effects reported in a large number of studies that support the positive results of preparatory information prior to acutely painful procedures such as surgery or bone marrow aspiration.

The most influential model of psychological factors in chronic pain was Fordyce's assumption that chronic pain can develop and be maintained due to operant conditioning of pain behaviors, i.e., overt expressions of pain. Fordyce (1976) postulated that acute pain behaviors such as limping or moaning may come under the control of external contingencies of reinforcement, and thus develop into a chronic pain problem. Positive reinforcement of pain behaviors (e.g., by attention and the expression of sympathy), negative reinforcement of pain behaviors (e.g., the reduction of pain through the intake of medication or by the cessation of activity) as well as a lack of reinforcement of healthy behaviors could provoke chronicity in the absence of nociceptive input. Thus, pain behaviors, originally elicited by nociceptive input, may, over time, occur in response to environmental contingencies. This model has generated much research, which has not only confirmed the original assumptions made by Fordyce; it has also been shown that in addition to pain behaviors the subjective experience of pain, as well as physiological responses related to pain, are subject to reinforcement learning.

A special role has been assigned to the 'significant others' of chronic pain patients, who have high reinforcement potential. When solicitous, i.e., pain-reinforcing, spouses were present, several studies found that the patients were more pain-sensitive during acute pain tests than when the spouses were absent. The patients with nonsolicitous spouses did not differ whether the spouse was present or absent. These studies suggest that spouses can serve as discriminative stimuli for the display of pain behaviors by chronic pain patients, including their reports of pain intensity. Health care providers may also serve as discriminative cues influencing patients' responses. Solicitous responses by significant others can also lead to increased physiological pain responses in the patients, whereas the focusing on healthy behaviors by significant others can have positive effects on the pain experience (cf Flor and Turk, 2011).

Of equal importance is the operant conditioning related to the intake of pain medication. Patients are often told by their physicians or by well-meaning family members that they should not continue to take analgesic medication unless the pain increases to a point where it becomes intolerable (referred to as *prn* from the Latin, 'take as needed'). When pain medication is taken at that stage, both pharmacological and behavioral factors can contribute to the development of the misuse of medication and, in severe cases, even drug dependence. If analgesic medication is taken only at peak pain levels, the effect of the medication is less potent, and patients cycle between high and low levels of medication that facilitate the development of dependence. In addition, medication intake is negatively reinforcing, since the intake of medication ends an aversive state (pain). Subsequently, the pain-reducing behavior (use of analgesics) increases in frequency. Thus, both pharmacotherapists and behavioral psychologists recommend that analgesic medication should be taken not in a pain-contingent but rather in a time-contingent fashion, adapted to the specific pain level of the patient and to the half-life of the drug.

The negative reinforcement of activity levels is an important process in the development of disability. A specific activity, for example walking, is performed until pain sets in, at which point the activity is interrupted and replaced by rest. Subsequently, the pain will be reduced. The reduction of an aversive state – pain – negatively reinforces the cessation of activity. As was the case with the intake of analgesic medication, the cessation of activity has to be made dependent on the amount of activity achieved – quota-based (e.g., number of stairs climbed, distance walked) – rather than on the amount of pain. Thus, the pain-reinforcing quality of rest is eliminated. This formulation supports the strategy of encouraging patients to perform activities to meet a specific quota, and not until pain is perceived as overwhelming. More recently, the operant formulation of pain has been extended to reward learning, and it has been postulated that pain relief may be processed as rewarding and thus contribute to the maintenance of chronic pain (Becker et al., 2012; Leknes and Tracey, 2008).

The respondent conditioning model postulates that numerous, formerly neutral, cues can be associated with the experience of pain and can – over time – themselves elicit responses that lead to an increased pain response, and can create the experience of pain in the absence of nociceptive input. To illustrate this process, the patient may have learned to associate increases in muscle tension with all kinds of stimuli that were originally associated with nociceptive stimulation. Thus, sitting, walking, bending, or even thoughts about these movements may elicit anticipatory anxiety and an increase in muscle tension. This fear of movement or 'kinesophobia' has been discussed as an important factor in the maintenance and exacerbation of chronic pain (Leeuw et al., 2007). Subsequently, patients may display maladaptive responses to any number of stimuli and reduce the frequency of performance of many activities other than those that initially reduced pain. Thus, although the original association between injury and

pain results in anxiety regarding movement, over time, the anxiety may lead to increased muscle tension and pain even if the nociceptive stimuli are no longer present. In addition, stress situations can increase muscle tension levels and cause sympathetic activation and may, thus, reinforce this process.

Many patients report that an acute pain problem evolved into chronic pain at a time where personal stressors co-occurred with the pain. Stress situations may serve as additional unconditioned stimuli and also as conscious stimulus for muscle tension increases, increased sympathetic activation, and subsequently pain. Nonoccurrence of pain is a powerful reinforcer for reduction of movement. Thus, the original respondent conditioning may be complemented by an operant process, whereby the nociceptive stimuli need no longer be present for the avoidance behavior to occur. People who suffer from acute back pain, regardless of the cause, may adopt specific behaviors (e.g., limping) to avoid pain and they may never obtain 'corrective feedback' because they fail to perform more natural movements. Reduction in physical activity may subsequently result in muscle atrophy and increased disability. In this manner, the physical abnormalities proposed by biomechanical models of pain may actually be secondary to changes in behavior initiated through learning. Chronic pain patients tend to focus their attention on impending pain and subsequently avoid many types of activity, thus fostering the development of disability and depression. The release of endogenous opioids – the body's own analgesic system – may also be influenced by respondent conditioning as well as brain responses related to the experience of pain. Moreover, fear of pain and subsequent avoidance of activity is one of the best predictors of subsequent chronicity (Turk and Wilson, 2010; Wideman et al., 2013).

Response acquisition through the observation of others and modeling is an essential mechanism of learning new patterns of behavior (Martin et al., 2014). Children acquire attitudes about health and health care, and the perception and interpretation of symptoms and physiological processes from their parents and their social environment. They also learn appropriate responses to injury and disease, and thus may be more or less likely to ignore or overrespond to normal bodily sensations they experience. The culturally acquired perception and interpretation of symptoms determines how people deal with illness; the sight of others in pain is an event that captures attention and may have survival value, may help to avoid experiencing more pain, and help to learn what to do about acute pain. Modeling probably plays a part in the phenomenon of 'pain-prone families,' families with a significantly elevated occurrence of pain problems. It was, for example, reported that children show the same pain syndromes as their parents currently have, rather than the pain problems their parents had in their own childhood. The large cultural variations in pain expression are also important. In common clinical practice, the acquisition or extinction of pain-related behavior by means of modeling has received little attention. However, there are occasional indications for the role of modeling in treating pain problems in children in burn units and in the treatment of postoperative pain.

Despite the great deal of data available on the modification of experimentally induced pain behavior by means of modeling in normal (healthy) subjects, there are few experimental results concerning chronic pain patients. Nor are there any longitudinal studies of the development of pain syndromes in 'pain-prone families.' Further investigation is necessary for the evidence of modeling as a factor in the development of chronic pain disorders.

Cognitive Factors and Pain

Cognitive–behavioral models of chronic pain emphasize that the evaluation of the pain experience by the patient greatly determines the amount of pain that is experienced as well as its negative consequences (Flor and Turk, 2011; Skinner et al., 2012). General assumptions that characterize the cognitive–behavioral perspective are: (1) people are active processors of information, not passive reactors; (2) thoughts (e.g., appraisals, expectancies) can elicit or modify mood, affect physiological processes, influence the environment, and serve as impetus for behavior. Conversely, mood, physiology, environmental factors, and behavior can influence thought processes; (3) behavior is reciprocally determined by the person and by environmental factors; (4) people can learn more adaptive ways of thinking, feeling, and behaving; and (5) people are capable and should be involved as active agents in change of maladaptive thoughts, feelings, and behaviors.

From the cognitive–behavioral perspective, people suffering from chronic pain are viewed as having negative expectations about their own ability to control certain motor skills such as performing specific physical activities (e.g., climbing stairs, lifting objects) that are attributed to one overwhelming factor, namely, a chronic pain syndrome. Moreover, chronic pain patients tend to believe that they have a limited ability to exert any control over their pain. Such negative, maladaptive appraisals about their situation and personal efficacy may reinforce the experience of demoralization, inactivity, and overreaction to nociceptive stimulation.

A great deal of research has been directed toward identifying cognitive factors that contribute to pain and disability. These have consistently demonstrated that patients' attitudes, beliefs, and expectancies about their plight, themselves, their coping resources, and the health care system affect their reports of pain, activity, disability, and response to treatment.

Pain, when interpreted as signifying ongoing tissue damage or a progressive disease, seems to produce considerably more suffering and behavioral dysfunction than if it is viewed as being the result of a stable problem that is expected to improve. A number of studies using experimental pain stimuli demonstrated that the belief in personal control can ameliorate the experience of experimentally induced nociception. Moreover, the type of thoughts employed during exposure to painful stimulation has been related to pain tolerance and pain intensity ratings. Catastrophizing thoughts have been associated with lower pain tolerance and higher ratings of pain intensity. In contrast, coping thoughts have been related to higher pain tolerance and lower pain intensity ratings.

Certain beliefs may lead to maladaptive coping, increased suffering, and greater disability. Patients who believe their pain is likely to persist may be passive in their coping efforts and fail to make use of available strategies, even if in their repertoire, to cope with pain. Patients who consider their pain to be an

'unexplainable mystery' may negatively evaluate their own ability to control or decrease pain, and are less likely to rate their coping strategies as effective in controlling and decreasing pain.

Once beliefs and expectancies (cognitive schemata) about a disease are formed, they become stable and very difficult to modify. Patients tend to avoid experiences that could invalidate their beliefs, and they adapt their behavior in accordance with these beliefs, even in situations where the belief is no longer valid (no corrective feedback is received to discredit this belief). For example, feeling muscular pain following activity may be caused by lack of muscle strength and general deconditioning, and not by additional tissue damage.

Self-regulation of pain and its impact depends upon a person's specific ways of dealing with pain, adjusting to pain, and reducing or minimizing pain and distress caused by pain – their coping strategies. Coping is assumed to be manifested by spontaneously employed purposeful and intentional acts, and it can be assessed in terms of overt and covert behaviors. Overt, behavioral coping strategies include rest, medication, and use of relaxation. Covert coping strategies include various means of distracting oneself from pain, reassuring oneself that the pain will diminish, seeking information, and problem solving. Studies have found active coping strategies (efforts to function in spite of pain, or to distract oneself from pain such as activity, or ignoring it) to be associated with adaptive functioning; and passive coping strategies (depending on others for help in pain control and restricted activities) to be related to greater pain and depression. However, beyond this there is no evidence supporting the greater effectiveness of one active coping strategy compared to any other. Specific coping strategies need not always be adaptive or maladaptive. It seems more likely that different strategies will be more effective than others for some people at some times, but not necessarily for all people all of the time, or even the same person at different times and coping should be understood in the context of self-regulation (Van Damme and Kindermans, 2014).

Pain and Emotional Factors

The affective factors associated with pain include many different emotions, but they are primarily negative in quality. Anxiety and depression have received the greatest amount of attention in chronic pain patients; however, anger has also received considerable interest as an important emotion in chronic pain patients. Research suggests that from 40 to 50% of chronic pain patients suffer from depression. There have been extensive and fruitless debates concerning the causal relationship between depression and pain. In the majority of cases, depression appears to be the patients' reaction to their plight. The presence of depression is closely related to the feelings of loss of control and helplessness often associated with pain (Goesling et al., 2013).

Several investigators have also found a close association between fear of pain and dysfunctional coping. In addition, high comorbidity between anxiety disorders and pain seems to be present. Muscular hyperreactivity to stress seems to be closely associated with fear of pain.

Anger has been widely observed in individuals with chronic pain. The internalization of angry feelings seems to be strongly related to measures of pain intensity, perceived interference, and reported frequency of pain behaviors. Anger and hostility are closely associated with pain in persons with spinal cord injuries. Frustrations related to persistence of symptoms, limited information on etiology, and repeated treatment failures along with anger toward employers, the insurance, the health care system, family members, and themselves, also contributed to the general dysphoric mood of these patients. The impact of anger and frustration on exacerbation of pain and treatment acceptance has not received adequate attention. It would be reasonable to expect that the presence of anger may serve as an aggravating factor, associated with increasing autonomic arousal, and having the effect of blocking motivation and acceptance of treatment oriented toward rehabilitation and disability management rather than cure, which are often the only treatments available for chronic pain.

Biobehavioral Perspective

A biobehavioral perspective of chronic pain needs to consider the factors discussed above and their mutual interrelationships in the explanation of chronic pain (Flor and Turk, 2011). The existence of a physiological disposition or diathesis is one important component. This predisposition is related to a reduced threshold for nociceptive stimulation and can be determined by genetic factors or acquired through early learning experiences. It is known that large genetic variations in individual pain sensitivity exist (Mogil, 2012). Furthermore, early stress experience greatly impacts later pain processing (Chapman et al., 2008). A further component of the biobehavioral model is a response stereotype of a particular bodily system, such as exaggerated muscular responses of the lower back muscle to stress and pain that is based both on the diathesis and on aversive experiences present at the time of the development of the response. These aversive stimuli may include personal or work stress or problematic occupational conditions and will lead not only to painful responses, but also to avoidance behaviors and associated maladaptive cognitive and affective processes. The cognitive evaluation of these external or internal stimuli is of great importance in the pain response, as discussed above.

Memory for Pain

An important maintaining factor in this chronicity process is the development of pain memories (Erskine et al., 1990). For example, pain patients have a tendency to remember preferentially negative and pain-related life events and show a deficit in the retrieval of positive memories. The experience of chronic pain also leads to the development of somatosensory pain memories, for example, an expanded representation of the affected body part in the primary somatosensory cortex. This expanded cortical representation is accompanied by increased sensitivity to both painful and nonpainful stimuli, and may be further enhanced by learning processes or attention to painful stimulation. A dramatic example of a learned memory for pain has been found in phantom limb pain patients (Flor et al.,

2006). In upper extremity amputees, the magnitude of the phantom limb pain was found to be proportional to the amount of reorganization in the primary somatosensory cortex, namely, the shift of the cortical mouth representation into the area where the amputated limb was formerly represented. The brain obviously maintains a memory of the former input to the deafferented area. Subsequently, stimulation stemming from areas adjacent to the deafferented zone elicits sensations and pain in the now absent limb. Phantom sensations and cortical reorganization are absent in congenital amputees. The focus of the biobehavioral perspective is thus on the patient and not just on the symptoms or the underlying pathology. This focus also requires that the treatment of the patients be tailored not only to medical factors, but that it incorporates psychosocial variables that may often be predominant in states of chronic pain (*see* Chronic Pain: Models and Treatment Approaches).

Future Perspectives

At the end of the twentieth century, there has been an enormous increase of research in the neurosciences, where the mutual interactions of molecular and genetic as well as neuronal and behavioral variables have begun to be examined. This integration of the psychosocial and biomedical perspective is much needed and will bring substantial advances in our understanding of the phenomenon of pain. A further important development is the finding in neuroscience that the adult brain is enormously plastic, and that learning processes as well as training measures have great impact on these plastic changes. This opens the door for new approaches to treatment (Flor, 2008; Sandkühler and Lee, 2013). It will also aid in delineating and eliminating the factors that lead to resistance to the treatment of chronic pain syndromes.

See also: Chronic Illness, Psychosocial Coping with; Chronic Pain: Models and Treatment Approaches; Coping with Job Stress; Health Behavior, Psychosocial Theories of; Occupational Stress Management; Stress, Coping and Health.

Bibliography

Apkarian, A.V., 2008. Pain perception in relation to emotional learning. Current Opinion in Neurobiology 18, 464–468.

Becker, S., Gandhi, W., Schweinhardt, P., 2012. Cerebral interactions of pain and reward and their relevance for chronic pain. Neuroscience Letters 520, 182–187.

Chapman, C.R., Tuckett, R.P., Song, C.W., 2008. Pain and stress in a systems perspective: reciprocal neural, endocrine, and immune interactions. Journal of Pain 9, 122–145.

Erskine, A., Morley, S., Pearce, S., 1990. Memory for pain: a review. Pain 41, 255–265.

Fillingim, R.B., Bruehl, S., Dworkin, R.H., Dworkin, S.F., Loeser, J.D., Turk, D.C., Widerstrom-Noga, E., Arnold, L., Bennett, R., Edwards, R.R., Freeman, R., Gewandter, J., Hertz, S., Hochberg, M., Krane, E., Mantyh, P.W., Markman, J., Neogi, T., Ohrbach, R., Paice, J.A., Porreca, F., Rappaport, B.A., Smith, S.M., Smith, T.J., Sullivan, M.D., Verne, G.N., Wasan, A.D., Wesselmann, U., 2014. The ACTTION-American Pain Society Pain Taxonomy (AAPT): an evidence-based and multidimensional approach to classifying chronic pain conditions. Journal of Pain 15, 241–249.

Flor, H., 2008. Maladaptive plasticity, memory for pain and phantom limb pain: review and suggestions for new therapies. Expert Review of Neurotherapy 8, 809–818.

Flor, H., Nikolajsen, L., Jensen, T.S., 2006. Phantom limb pain: a case of maladaptive CNS plasticity? Nature Reviews Neuroscience 7, 873–881.

Flor, H., Turk, D.C., 2011. Chronic Pain: An Integrated Biobehavioral Approach. IASP Press, Seattle, WA.

Fordyce, W.E., 1976. Behavioral Methods for Chronic Pain and Illness. Mosby, St. Louis.

Gatchel, R.J., Peng, Y.B., Peters, M.L., Fuchs, P.N., Turk, D.C., 2007. The biopsychosocial approach to chronic pain: scientific advances and future directions. Psychological Bulletin 133, 581–624.

Goesling, J., Clauw, D.J., Hassett, A.L., 2013. Pain and depression: an integrative review of neurobiological and psychological factors. Current Psychiatry Reports 15, 421.

Leeuw, M., Goossens, M.E., Linton, S.J., Crombez, G., Boersma, K., Vlaeyen, J.W., 2007. The fear-avoidance model of musculoskeletal pain: current state of scientific evidence. Journal of Behavioral Medicine 30, 77–94.

Leknes, S., Tracey, I., 2008. A common neurobiology for pain and pleasure. Nature Reviews Neuroscience 9, 314–320.

Martin, L.J., Tuttle, A.H., Mogil, J.S., 2014. The interaction between pain and social behavior in humans and rodents. Current Topics in Behavioral Neurosciences [published online ahead of print 21 February 2014]. http://dx.doi.org/10.1007/7854_2014_287. http://link.springer.com/chapter/10.1007/7854_2014_287.

Melzack, R., Wall, P.D., 1965. Pain mechanisms: a new theory. Science 150, 971–979.

Merskey, H., Bogduk, N., 1994. Classification of Chronic Pain. Descriptions of Chronic Pain Syndromes and Definitions of Pain Terms. IASP Press, Seattle, WA.

Mogil, J.S., 2012. Pain genetics: past, present and future. Trends in Genetics 28, 258–266.

Reid, K.J., Harker, J., Bala, M.M., Truyers, C., Kellen, E., Bekkering, G.E., Kleijnen, J., 2011. Epidemiology of chronic non-cancer pain in Europe: narrative review of prevalence, pain treatments and pain impact. Current Medical Research and Opinion 27, 449–462.

Sandkühler, J., 2000. Learning and memory in pain pathways. Pain 88, 113–118.

Sandkühler, J., Lee, J., 2013. How to erase memory traces of pain and fear. Trends in Neurosciences 36, 343–352.

Skinner, M., Wilson, H.D., Turk, D.C., 2012. Cognitive-behavioral perspective and cognitive-behavioral therapy for people with chronic pain: distinctions, outcomes, and innovations. Journal of Cognitive Psychotherapy 26, 93–113.

Turk, D.C., Rudy, T.E., 1988. Toward an empirically derived taxonomy of chronic pain patients: integration of psychological assessment data. Journal of Consulting and Clinical Psychology 56, 233–238.

Turk, D.C., Wilson, H.D., 2010. Fear of pain as a prognostic factor in chronic pain: conceptual models, assessment, and treatment implications. Current Pain and Headache Reports 14, 88–95.

Van Damme, S., Kindermans, H., 2014. A self-regulation perspective on avoidance and persistence behaviour in chronic pain: new theories, new challenges? Clinical Journal of Pain [published online ahead of print 21 March 2014]. http://dx.doi.org/10.1097/AJP.0000000000000096. http://journals.lww.com/clinicalpain/pages/results.aspx?txtKeywords=Kindermans%2c+Hanne.

Wideman, T.H., Asmundson, G.G., Smeets, R.J., Zautra, A.J., Simmonds, M.J., Sullivan, M.J., Haythornthwaite, J.A., Edwards, R.R., 2013. Rethinking the fear avoidance model: toward a multidimensional framework of pain-related disability. Pain 154, 2262–2265.

Relevant Websites

http://apkarianlab.northwestern.edu/ – Apkarian Lab (Pain and Passions).
http://www.bodyinmind.org/ – Body in Mind (Australia).
http://depts.washington.edu/anesth/ – Department of Anesthesiology & Pain Medicine.
http://www.efic.org/ – EFIC (European Pain Federation EFIC).
http://www.iasp-pain.org/ – IASP (International Association for the Study of Pain).
http://iannettilab.webnode.com/ – Iannetti Lab (Human Sensory Neuroscience and Pain Perception).

Paleoethnobotany

Deborah M Pearsall, University of Missouri, Columbia, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Paleoethnobotany is the study of human–plant interrelationships through the archaeological record. Paleoethnobotanists recover, identify, and interpret evidence of past plant use documented through macroremains, pollen, starch, and phytoliths. Not all human activities involving plants result in the deposition and preservation of plant remains in archaeological sites, and natural processes may also introduce plant remains into sites. Paleoethnobotanical data and methods are reviewed, and examples of paleoethnobotanical research presented, including the topics of food and human evolution, the origins of agriculture, uses of food and drink in ritual and feasting, and the case of 'Ötzi,' the Tyrolean Iceman, and his environment.

Paleoethnobotanical Data

Paleoethnobotany, the study of human–plant interrelationships through the archaeological record, provides direct evidence on how past populations met subsistence and other daily needs and contributes to understanding human interactions with their environments, and with each other. Human activities involving plants often leave traces, such as charred seeds, fruit rind fragments, wood charcoal, tuber or root peels, pollen and starch grains, and phytoliths (plant opal silica bodies), which become incorporated into archaeological and geological records. Recovery, identification, and interpretation of such plant remains are the tasks of paleoethnobotanists.

Macroremains

Seeds, fruits, nuts, roots/tubers, and wood recovered from archaeological sites are called macroremains – plant tissues visible to the naked eye or under low magnification (Pearsall, 2000; Hastorf and Popper, 1988). Since such materials deteriorate rapidly under most conditions, survival of macroremains depends on slowing the agents of decay by charring, waterlogging, mineralizing, or desiccation. Incomplete accidental or deliberate burning, resulting in charring of plant tissues, is the most common mode of preservation.

Macroremains are recovered by collection of material *in situ*, through screening, and by a water recovery technique called flotation (Pearsall, 2000). *In situ* recovery preserves the association of macroremains and nonbotanical artifacts, but limits recovery to materials that can be easily seen. Screening allows for more systematic recovery of macroremains larger than the screen mesh, but can damage fragile remains. Water flotation, a technique that uses differences in density of organic and inorganic materials to achieve separation of organic remains from soil, greatly enhances the quantity and size range of macroremains that can be recovered (**Figure 1**). Manual and machine-assisted flotation has become routine in archaeology. Fine-sieving is often substituted for flotation in waterlogged or desiccated deposits, since waterlogged materials are not buoyant, and desiccated remains may be damaged by immersion in water.

Macroremains are identified by comparing unknown archaeological materials to known modern comparative specimens, augmented by published and online sources (Pearsall, 2000). Characteristics important for identification vary by type of remains. For example, seeds are identified by distinctive features of the seed coat, the amount of endosperm and its position in the seed relative to the embryo, and overall size and shape, while wood is identified by its internal structure (vessels and their arrangement, size

Figure 1 Students learning to operate a water flotation system that uses water pressure to separate organic materials from soil matrix.

and arrangement of rays, and abundance and nature of parenchyma). Common challenges to the identification include fragmentation, damage and loss of key characteristics, redundancy of characteristics among related plants, and lack of comparative materials.

Not all human activities involving plants result in the deposition and preservation of macroremains, creating interpretive challenges for paleoethnobotanists. For example, foods consumed or processed away from an occupation site may never become deposited in a garbage midden, while foods eaten raw or prepared as beverages may never be exposed to fire and preserved. This makes interpretation of absence especially difficult. Natural processes, rather than human behavior, may bring plants into what will become archaeological deposits, creating other interpretative challenges. For example, weedy plants may contribute substantial numbers of seeds to midden areas or abandoned dwellings, and wind or water may bring botanical materials into sites. If macroremain preservation is through charring, then material deposited through natural processes, being rarely charred, may have minimal impact on assemblages. Natural processes may also remove macroremains, for example, though erosion of site sediments.

Phytoliths

Phytoliths are produced by the deposition of silica from groundwater into plant tissues. There is a strong genetic component to phytolith formation; orders and families show tendencies for deposition or lack of deposition of silica (e.g., Pearsall, 2000; Piperno, 2006; Runge, 2000; among others). While silicification patterns are redundant in some groups, distinctively shaped phytoliths are formed in many, providing a means of identifying plants at family, genus, or species level. Most phytoliths are the size of silt soil particles (2–250 μm).

Phytoliths are recovered in a specialized laboratory setup for chemical processing from small samples taken for that purpose (Pearsall, 2000; Piperno, 2006). Basically, phytoliths are floated from the matrix using a heavy liquid of specific gravity 2.3 after samples have been treated with a series of chemicals that break down chemical bonds between phytoliths and the matrix. Phytolith extracts are slide mounted and examined under 200–400 power magnification or higher. Phytoliths are identified by comparison of unknowns to phytoliths extracted from identified plant specimens, augmented by online and published sources. Characteristics useful for identification vary widely among plant groups but include three-dimensional morphology, size, surface characteristics, and attachments. Typically, slides are examined until a predetermined count or sum is reached.

Since phytoliths are inorganic, they survive in circumstances under which organic tissues may be lost (Pearsall, 2000; Piperno, 2006). Phytolith representation in archaeological and geological sediments is thus impacted chiefly by phytolith production patterns, i.e., whether or not plants and plant tissues targeted by humans are silica accumulators and produce diagnostic phytoliths. Phytoliths are not indestructible, however. Lightly silicified tissues may break up in soil, and phytoliths may dissolve in highly alkaline environments (pH 9 and higher). Unlike pollen and seeds, phytoliths are not involved in plant reproduction, and so do not have a dispersal mechanism. Phytoliths are released back into the soil by processes that break down plant tissues: organic decay, burning, and digestion. Because phytoliths are released into soil, a large proportion of phytoliths deposited in sediments or soils represent *in situ* deposition. Research indicates, however, that if soil, especially silt and fine sand fractions, moves, phytoliths move. Phytoliths are part of wind-blown dust and soil washed into lakes and swamps, for example. Because both natural processes and human behavior result in deposition of plant remains in archaeological sites, phytoliths become deposited through both these pathways.

Pollen

Pollen is formed in the anthers of seed plants. Most are regular rotation ellipsoids (i.e., symmetrical around an axis) that usually have three concentric layers. The outermost, the exine, contains sporopollenin. This is one of the most resistant natural organic compounds, and is preserved in fossil pollen (Faegri et al., 1989). Although all seed plants produce pollen, pollen abundance and morphology are affected by pollination mechanism. For example, wind-pollinated plants produce many more pollen grains than do animal-pollinated plants, whose pollen has features that ensure that grains stick to the pollinator. Pollen varies in size from 5 to 200 μm.

Like phytoliths, pollen is extracted from small samples in a specialized laboratory. The basic aim is to concentrate pollen and render grains as visible as possible. Concentration is the process of removing everything else from a sample with a series of chemicals, leaving only pollen and spores; rendering the grains visible involves removing clays and other soil components, mounting residue in a suitable medium, and staining (Faegri et al., 1989; Moore et al., 1991). Identification follows the approach already described for phytolith analysis. Features of the exine, as well as grain shape and size, are important for identifying pollen (Kapp, 1969).

The likelihood that pollen will become deposited in archaeological sediments, on artifacts, or in a lake and swamp sediments is affected by pollination mechanism. For example, a major pathway by which pollen is deposited on the landscape is through pollen rain, the mix of pollen transported by air currents (Faegri et al., 1989). Thus, pollen of wind-pollinated plants is highly likely to become deposited in open-air contexts, less likely in closed contexts (e.g., inside structures), while pollen of animal-pollinated plants may only be introduced through the deposition of flowers or fruits with pollen adhering to them. Once pollen is deposited, sporopollenin is tough, but not indestructible. Pollen damage or destruction is reduced in arid or waterlogged soils and in deposits covered by peat formation (acid conditions) because of lower microbial activity (Dimbleby, 1985). Burial of archaeological deposits (by volcanic ash, sand, under-earth mounds, pavements, and within burials) and contact of pollen with artifact surfaces (grinding stones, ceramic vessels, and metal objects) may enhance preservation. In many cases, however, pollen representation in archaeological sediments is a matter of preservation; conditions encountered in many sites are not ideal for pollen preservation (Faegri et al., 1989).

Starch

Starch serves as a plant energy reserve (Gott et al., 2006). It forms in amyloplasts, beginning at a point called the hilum, and grows by successive layers (lamellae), which may remain visible on the granule. Starch is semicrystalline and exhibits strong birefringence, i.e., under polarized light it appears white against the black background, and an extinction cross, a dark cross centered on the hilum, is visible. Starch designed for long-term energy storage is what humans target for food. It tends to be concentrated in seeds and underground storage organs (USOs), but it may also be found in fruits. Storage starch morphology is largely under genetic control, and many plants can be identified by their starch.

Like phytoliths and pollen, starch is extracted in a specialized laboratory from small samples taken for that purpose. The general methodology follows that of phytoliths and pollen, to remove other particles and concentrate the starch (Torrence, 2006). Identification follows the approach already described. Among the characteristics used for identification of starch are granule shape and size, hilum location, extinction cross characteristics, fissure presence and shape, surface and edge characteristics, and whether granules are simple or compound (Gott et al., 2006).

Like phytoliths, starch is not involved in plant reproduction, and does not have a dispersal mechanism. A number of different pathways have been proposed by which starch enters the archaeological record: directly from starchy plant part to soil (e.g., through decay of tubers in an agricultural field); by being consumed or altered by people, then discarded (e.g., decay of tuber peelings in a midden); and by deposition on artifacts used to process starch-rich tissues (Beck and Torrence, 2006). The last named has been the focus of much archaeological research, as well as experiments to investigate how starch is modified by common food preparation techniques. But starch granules have been recovered from a wide range of archaeological contexts, including soils, coprolites, and human dental calculus (Barton and Torrence, 2006). Factors that influence starch survival in archaeological contexts are not yet well understood.

Paleoethnobotany in Archaeology

In the limited space remaining, it is impossible to do justice to the diverse and significant insights into human–plant interrelationships gained through the study of ancient plant remains. A 'taste' will have to serve to illustrate what paleoethnobotany contributes to archaeology.

Human Evolution and the Emergence of Modern Behavior

Understanding early hominin diet is an important component of studying human evolution. Dental calculus traps food particles, including phytoliths and starch grains, and provides direct evidence of diet. Recovery of phytoliths from dental calculus of two 2 million-year-old *Australopithecus sediba* individuals from Malapa, South Africa, revealed consumption of dicotyledonous leaf, fruit, and wood/bark as well as sedge, grass, and palm (Henry et al., 2012). In combination with dental microwear analysis and carbon isotope data, these results revealed consumption of some hard foods and nearly exclusive feeding on a wide variety of C_3 resources, in preference to more widely available C_4 grasses, providing insight into the nature of the environmental relationships of *A. sediba*.

Paleoethnobotany has provided considerable insight into the nature of diet of early *Homo sapiens*. A study of phytoliths from Amud Cave, Israel, suggested that a broadening of diet, including utilization of wild grasses and other seeds, began with Neanderthals (Madella et al., 2002). Cemented sediment patches and intact hearths were sampled, and abundant phytoliths were recovered. Among the plants identified were dicot wood and leaves, grasses, including phytoliths from mature grass seed heads, palm, and fig-tree family (Moraceae), reflecting in large part plant-related human activities, including seed gathering. The evidence indicated broad-spectrum plant foraging during the Middle Paleolithic. Research at the Ohalo II site (23 000 BP), Israel, demonstrated through the study of plant macroremains (Weiss et al., 2004) and starch grains (Piperno et al., 2004) that by the Upper Paleolithic the plant component of human diet in southwest Asia was dominated by an array of wild grasses, and augmented by nuts, olives, fruits, and berries, among other plants. Grass seeds were processed by grinding.

Additional insights into plant-based subsistence strategies in the Middle Paleolithic of southwest Asia come from paleoethnobotanical studies at Kebara cave, Israel. Lev et al. (2005) analyzed charred macroremains from the site. Charred remains were found primarily near hearths. The vast majority of the 3956 identified seeds were legumes (Papilionaceae), but a wide variety of seeds were found in small numbers, including grass, goosefoot, safflower, seeds of several plants producing USOs, including wild radish and nut grass, and several species known for medicinal uses. Acorn and pistachio nut fragments were also recovered. The presence of charring suggested cooking of vegetal foods. Albert et al. (2012) studied phytoliths and thin sections taken from hearths and combustion features at Kebara. The generally high number of phytoliths present in samples suggested that plants were abundant in the cave. Those most directly related to anthropogenic deposition were from wood used as fuel for fires and grasses used to help start fires (or adhering to wood). Abundance of phytoliths from grasses and seed coats of the Boraginaceae in samples near the entrance of the cave was likely due to their growth at the entrance.

Study of microfossils in dental calculus of Neanderthal individuals provides further support for breadth of Neanderthal diet (Henry et al., 2011). Henry studied dental calculus from three teeth from Shanidar III, Iraq, and two each from Spy I and II, Belgium. Numerous starch grains were recovered and some phytoliths. In the warm Mediterranean region, Neanderthal individuals consumed grass seeds of the Triticeae tribe, including some that appeared to have been cooked, legume seeds, an unknown USO, and date palm. In cold northwestern Europe, consumption included an unknown USO and grass seeds (Andropogoneae tribe). These results showed that two major hominin dietary adaptations, cooking and incorporation of a diversity of relatively costly plant foods, had taken place by the Middle Paleolithic.

Beyond diet, paleoethnobotany contributes to understanding the emergence of other modern behaviors in the Middle Paleolithic. For example, research at Tor Faraj, a rock shelter site in Jordan, provided insight concerning the extent to which Neanderthal behavior resembled that of modern humans (Henry, 2003). Rosen (2003) conducted a study of plant microfossils, focused on reconstructing local vegetation, paleoclimate, and pattern of plant use by the shelter's inhabitants. Samples were taken horizontally across the living floor. Abundant phytoliths were recovered, as well as some spores, starch grains, and calcium oxalate crystals. Among the plants identified were pooid (festucoid) grasses, reed, sedge, date palm, and abundant phytoliths from trees or shrubs, suggesting that much moister conditions existed in the area at the time of site occupation than at present. Date palm fruits and/or leaves were likely used. Several kinds of starch grains were found, likely indicative of nuts, such as pistachio, roots, and tubers brought into the shelter. The distribution of microbotanical remains across the living floor indicated that grasses were spread on the floor near the back wall and in between the hearths, perhaps indicating prepared sleeping areas. Phytoliths from grass seeds and palm were concentrated near a hearth, suggesting a food preparation area, while high concentrations of dicot phytoliths beyond the drip line of the shelter suggested a brush windbreak. The microbotanical results contributed significantly to the overall finding that the inhabitants organized their activities essentially along the lines of modern hunter-gatherers (Henry, 2003).

A phytolith study conducted at Esquilleu Cave, Spain, by Cabanes et al. (2010) also documented the use of grass as bedding during the Late Mousterian. Most of the phytoliths recovered from samples were of grasses (leaf and inflorescence) but several hearth samples had abundant wood and bark phytoliths. Phytolith dissolution occurred in some areas. In areas of good preservation, accumulation was related to use of grass in bedding (in repetitive layers) and wood/bark for combustion. Interpretation of such layers of phytoliths as bedding is supported by an experimental study by Miller and Sievers (2012). They studied the micromorphological signatures of burned bedding, in order to gain insight into how laminated layers of carbonized material and phytoliths were produced, in this case, at the Sibudu site, South Africa. The experiment revealed that large volumes of vegetal material could have produced thin archaeological deposits of burnt bedding, and that if carbonized material was present, burning was most likely deliberate.

Origins of Food Production

Paleoethnobotany is at the heart of research into the origins of food production. The American tropics provide an example that illustrates the power of combining multiple indicators of plant use to document early agriculture. The early history of American plant domestication begins in lower Central America and northwestern South America, and is known in large part from phytolith, starch grain, and pollen evidence. Pollen and phytolith sequences from lakes indicated that populations began modifying landscapes in the early Holocene. Burning of forests and small-scale land clearance was documented then at Lake La Yeguada in Panama, for example (Piperno and Pearsall, 1998). Phytoliths also fill in the record at early sites with poor preservation of macroremains. For example, arrowroot, llerén (two root/tuber crops), gourd, squash, and maize (corn), all identified by phytoliths, were found at two early sites in Panama, Cueva de los Vampiros and Aguadulce.

Research at Xihuatoxtla shelter in the Central Balsas River valley, west Mexico, has now documented early maize in the dry tropical forest setting of its wild ancestor (Piperno et al., 2009). Maize phytoliths were recovered from site sediments, and maize starch and phytoliths from grinding stones, dating to 6700 BC. The spread of maize through the lowland tropics from its area of origin has been documented in part by pollen and phytoliths in sediment cores. Evidence for maize was found, for example, at 5100–5000 BC in a sediment core on the Gulf coast, and maize pollen and/or phytoliths documented the crop in south Pacific coastal Mexico, Pacific coastal Guatemala, northern Belize, and Honduras by 3500 BC (Piperno and Pearsall, 1998). Maize was carried south through the tropical lowlands prior to this time, however, as it is documented earlier in Panama, Colombia, and Ecuador. Perhaps the best known maize remains come from two caves, Coxcatlán and Guilá Naquitz, each located in the semiarid highlands of central Mexico. Maize macroremains were well preserved in these dry sites, but when directly dated, not all were as ancient as expected from site stratigraphy. Maize has been direct-dated to 3600 BC at Coxcatlán and 4250 BC at Guilá Naquitz (Long et al., 1990; Piperno and Flannery, 2001).

The emergence of agriculture in the dry tropical forests of southwest coastal Ecuador has also been documented largely through microfossil data (Piperno and Pearsall, 1998). Squash phytoliths were recovered from terminal Pleistocene and early Holocene strata at preceramic Vegas sites; other Vegas crops included gourd and llerén, and eventually maize. Maize, identified by phytoliths and starch grains, continued to be grown at Real Alto and Loma Alta, two later, Valdivia tradition farming villages, along with cotton and jack bean (preserved as macroremains), achira, manioc, chili pepper, llerén, and arrowroot (identified by phytoliths and/or starch grains) (Chandler-Ezell et al., 2006; Zarillo et al., 2008). Agriculture in coastal Ecuador remained broad-based for many millennia, incorporating wild/managed tree fruits as well as annual crops (Pearsall, 2004).

Research at early preceramic sites in the Ñanchoc valley, located in the lower western slopes of the Andes in northern Peru, has provided the first detailed look at early farming communities in the Andes. Recovery of macroremains from sealed house floors and hearths has documented cultivation of squash, peanut, quinoa (all by 7500 BP or earlier), and cotton (by 5500 BP) (Dillehay et al., 2007). Manioc and other unidentified tubers and fruits were also recovered. The early presence west of the Andes of peanut and manioc, originating in eastern South America, and quinoa, from the southern highlands, has documented the extent of networks of communication and exchange among early agriculturalists. Study of starch recovered from the dental calculus of 39 teeth from the same contexts as the dated macroremains led to the identification of starch from bean and pacae seeds (tree fruit), squash flesh, and peanut (Piperno and Dillehay,

2008). Starch from these four taxa was commonly recovered; rarer types identified included possible manioc and tree fruits. Irrigation technology, vital to the later development and expansion of agriculture in desert coast Peru, was also documented early in the Ñanchoc valley, in the form of small-scale gravity canals (Dillehay et al., 2005). Paleoethnobotanical and archaeological evidences from this small valley combine to indicate how early effective farming emerged in western Peru, and took on a regional identity (Dillehay, 2011).

The Iceman and His Environment

In September 1991, a well-preserved corpse of a man was discovered frozen in the Alps at 3214 m elevation. Eventually dated to more than 5000 years ago, studies of 'Ötzi,' the Tyrolean Iceman, and his tools and clothing have provided a snapshot of life in the Alps during the Neolithic. Detailed paleoethnobotanical studies played a central role in understanding Ötzi's life and death and his natural environment. For example, study of pollen sequences from bogs and peat deposits demonstrated that the alpine grasslands of the region were used as pastures since the beginning of the Neolithic (Bortenschlager, 2000). Forests were also cleared to increase pasturage. Ötzi was likely involved in an early form of transhumance, moving his flocks seasonally to higher elevation pastures. Wooden objects found with Ötzi were identified through dendrological analysis. Ötzi's bow and the handle of his ax, for example, were made of yew, while arrow shafts were made of wayfaring tree (Oeggl and Schoch, 2000). His pack frame was made of hazel and larch. A birch-bark container was filled with maple leaves and charcoal fragments, and probably served as a storage container for cold embers. The majority of wood species belonged to the montane region, and technological analysis suggested that Ötzi selected species based on the requirements of each object.

Study by Oeggl et al. (2007) of a sequence of gut samples revealed what Ötzi ate during the last hours of his life, and what kinds of environments he traveled through. Pollen, plant macroremains, whipworm eggs, and muscle fibers were recovered. Pollen can become incorporated into food residue in the gut in a variety of ways, for example, through food consumption, drinking water, and breathing. Only frequently occurring pollen was interpreted as associated with food consumption. Ötzi's last meals included meat, wheat, bracken, chenopod, and plants in the primrose and legume families. At least three different meals were identified. Abundant charcoal particles suggested that Ötzi's meals were prepared over an open fire. Background pollen from the gut series indicated that Ötzi traveled from the subalpine region down to the valley bottom, and then back up to the alpine region, where he died from an arrow wound.

Study of macroremains recovered from the location of Ötzi's discovery, a rocky hollow, provide further insights into the Iceman, his life, and his environment (Heiss and Oeggl, 2009). Materials studied included Ötzi's clothing and equipment, some of which was made of plant material, such as a grass cape/mat and grass shoe insulating material. Plant remains were also recovered from Ötzi's clothing and equipment and from wet-sieving or floating sediments from the bottom of the hollow. In all, over 40 000 plant macroremains have been studied, from at least 66 species, representing environments from the valley bottom to the snow line. However, only Ötzi's equipment, clothing, and the cultivated plants found in association with him, hulled barley, einkorn, millet, opium poppy, and flax, can be unambiguously considered the result of human rather than natural deposition through long-distance transport.

Ritual and Feasting

Beyond their roles in diet and economy, foods (and drinks) play important roles in social relationships, both public and private. In the Andes, for example, maize beer *(chicha)* is an important part of cultural identity, and plays a central role in ceremonies. Documenting when this role emerged was one goal of paleoethnobotanical research at Formative sites on the Taraco Peninsula, Bolivia (Logan et al., 2012). Phytolith and starch grain analyses were used to investigate maize processing and its consumption in both ceremonial and domestic contexts. Charred maize remains were lacking at sites, and no maize microfossils were recovered from cooking pot residues. The only evidence for maize processing came from phytolith or starch residues on grinding stones from a ceremonial space, on teeth of a sacrificed individual, and in other ceremonial contexts, such as burials and sunken courts. The evidence supported a ceremonial role for maize, probably in liquid rather than cooked form.

Study of macroremains, starch, and phytoliths from a buried ceremonial room at the Buena Vista site provides insights into ritual feasting during the preceramic period in coastal Peru (Duncan et al., 2009). The ceremonial room, a small temple of the Kotosh tradition, was intentionally buried, and food remains were part of the ritual entombment. Macroremains recovered from the room, well preserved by the desert conditions of the coast, were dominated by nonedible parts of food plants and broken gourd and squash serving vessels. These were not food offerings, but the refuse from feasting associated with the ritual burial of the temple (Duncan et al., 2009). Ritual entombment of buildings has been documented in the region, but not the association of feasting with temple burial. Study of starch residues from the gourd and squash artifacts recovered starch from manioc, potato, chili pepper, arrowroot, and algaroba. Starch and phytoliths recovered from artifacts and sediments led to identification of foods (and drinks) not present as macroremains.

There is a long tradition of studying plant food offerings from cremation cemeteries of the Iron Age and Roman Empire (Rottoli and Castiglioni, 2011). Since all offerings were subject to burning during the funerary rites, chances of preservation are high, and preservation is less selective, in comparison to macroremains preservation at contemporary habitation sites, for example. In their review of plant food offerings from Roman cremation sites in northern Italy, Rottoli and Castiglioni (2011) found that while cereals, legumes, cultivated and gathered fruits, vegetables, and prepared foods, such as bread, were often all present, fruit remains were especially common. This abundance contrasted markedly to abundance of fruits preserved at habitation sites. Presence of

grapes and hazelnuts in Roman cremations indicated continuity of local, Iron Age, traditions, to which offerings of other fruits, such as fig, date palm, olive, and pine nut became increasingly common. Another legacy of more ancient customs was indicated by isolated offerings of wild fruits such as acorn, strawberry, and rose. No evidence of consumption of foods was found, i.e., no remains of meals or preparation waste. Fruits, seeds, and prepared foods were offered in their entirety. Studies of plant food offerings provide insights into the Romanization of new territories and document the growth of trade, including new and exotic fruits, in Roman times.

See also: Environmental Archaeology; Zooarchaeology.

Bibliography

Albert, R.M., Berna, F., Goldberg, P., 2012. Insights on Neanderthal fire use at Kebara Cave (Israel) through high resolution study of prehistoric combustion features: evidence from phytoliths and thin sections. Quaternary International 247, 278–293.

Barton, H., Torrence, R., 2006. Looking ahead. In: Torrence, R., Barton, H. (Eds.), Ancient Starch Research. Left Coast Press, Walnut Creek, CA, pp. 217–223.

Beck, W., Torrence, R., 2006. Starch pathways. In: Torrence, R., Barton, H. (Eds.), Ancient Starch Research. Left Coast Press, Walnut Creek, CA, pp. 53–74.

Bortenschlager, S., 2000. The Iceman's environment. In: Bortenschlager, S., Oeggl, K. (Eds.), The Iceman and His Natural Environment. Palaeobotanical Results. Springer-Verlag Wien, New York, pp. 11–24.

Cabanes, D., Mollol, C., Expósito, I., Baena, J., 2010. Phytolith evidence for hearths and beds in the late Mousterian occupations of Esquilleu cave (Cantabria, Spain). Journal of Archaeological Science 37, 2847–2957.

Chandler-Ezell, K., Pearsall, D.M., Zeidler, J.A., 2006. Root and tuber phytoliths and starch grains document manioc (*Manihot esculenta*), arrowroot (*Maranta arundinacea*), and Ilerén (*Calathea* sp.) at the Real Alto site, Ecuador. Economic Botany 60 (2), 103–120.

Dillehay, T.D. (Ed.), 2011. From Foraging to Farming in the Andes. New Perspectives on Food Production and Social Organization. Cambridge University Press, New York.

Dillehay, T.D., Eling Jr., H.H., Rossen, J., 2005. Preceramic irrigation canals in the Peruvian Andes. Proceedings of the National Academy of Sciences 102 (47), 17241–17244.

Dillehay, T.D., Rossen, J., Andres, T.C., Williams, D.E., 2007. Preceramic adoption of peanut, squash, and cotton in Northern Peru. Science 316, 1890–1893.

Dimbleby, G.W., 1985. The Palynology of Archaeological Sites. Academic Press, London.

Duncan, N.A., Pearsall, D.M., Benfer, R.A., 2009. Gourd and squash artifacts yield starch grains of feasting foods from Preceramic Peru. Proceedings of the National Academy of Sciences 106 (32), 13202–13206.

Faegri, K., Kaland, P.E., Krzywinski, K., 1989. Textbook of Pollen Analysis by Knut Faegri and Johs. Iversen, fourth ed. John Wiley and Sons, Chichester, England.

Gott, B., Barton, H., Samuel, D., Torrence, R., 2006. Biology of starch. In: Torrence, R., Barton, H. (Eds.), Ancient Starch Research. Left Coast Press, Walnut Creek, CA, pp. 35–45.

Hastorf, C.A., Popper, V.S. (Eds.), 1988. Current Paleoethnobotany: Analytical Methods and Cultural Interpretations of Archaeological Plant Remains. University of Chicago Press, Chicago.

Heiss, A.G., Oeggl, K., 2009. The plant macro-remains from the Iceman site (Tisenjoch, Italian–Austrian border, eastern Alps): new results on the glacier mummy's environment. Vegetation History and Archaeobotany 18, 23–35.

Henry, A.G., Brooks, A.S., Piperno, D.R., 2011. Microfossils in calculus demonstrate consumption of plants and cooked foods in Neanderthal diets (Shanidar III, Iraq; Spy I and II, Belgium). Proceedings of the National Academy of Sciences 108 (2), 486–491.

Henry, A.G., Ungar, P.S., Passey, B.H., Sponheimer, M., Rossouw, L., Bamford, M., Sandberg, P., de Ruiter, D.J., Berger, L., 2012. The diet of *Australopithecus sediba*. Nature 487, 90–93.

Henry, D.O., 2003. Behavioral organization at Tor Faraj. In: Henry, D.O. (Ed.), Neanderthals in the Levant. Behavioral Organization and the Beginnings of Human Modernity. Continuum, London, pp. 237–269.

Kapp, R.O., 1969. How to Know Pollen and Spores. William C. Brown, Dubuque, Iowa.

Lev, E., Kislev, M.E., Bar-Yosef, O., 2005. Mousterian vegetal food in Kebara cave, Mt. Carmel. Journal of Archaeological Science 32, 475–484.

Logan, A.L., Hastorf, C.A., Pearsall, D.M., 2012. "Let's drink together": early ceremonial use of maize in the Titicaca Basin. Latin American Antiquity 23 (3), 235–258.

Long, A., Benz, B.F., Donahue, D.J., Jull, A.J.T., Toolin, L.J., 1990. First direct AMS dates on early maize from Tehuacan, Mexico. Radiocarbon 31 (3), 1035–1040.

Madella, M., Jones, M.K., Goldberg, P., Goren, Y., Hovers, E., 2002. The exploitation of plant resources by Neanderthals in Amud Cave (Israel): The evidence from phytolith studies. Journal of Archaeological Science 29, 703–719.

Miller, C.E., Sievers, C., 2012. An experimental micromorphological investigation of bedding construction in the Middle Stone Age of Sibudu, South Africa. Journal of Archaeological Science 39, 3039–3051.

Moore, P.D., Webb, J.A., Collinson, M.E., 1991. Pollen Analysis, second ed. Blackwell Scientific Publications, Oxford.

Oeggl, K., Kofler, W., Schmidl, A., Dickson, J.H., Egarter-Vigl, E., Gaber, O., 2007. The reconstruction of the last itinerary of "Ötzi", the Neolithic Iceman, by pollen analyses from sequentially sampled gut extracts. Quaternary Science Reviews 26, 853–861.

Oeggl, K., Schoch, W., 2000. Dendrological analyses of artefacts and other remains. In: Bortenschlager, S., Oeggl, K. (Eds.), The Iceman and His Natural Environment. Palaeobotanical Results. Springer-Verlag Wien, New York, pp. 29–61.

Pearsall, D.M., 2000. Paleoethnobotany. A Handbook of Procedures, second ed. Left Coast Press, Walnut Creek, CA.

Pearsall, D.M., 2004. Plants and People in Ancient Ecuador: The Ethnobotany of the Jama River Valley. Case Studies in Archaeology. Wadsworth/Thomson Learning, Belmont, CA.

Piperno, D.R., 2006. Phytoliths. A Comprehensive Guide for Archaeologists and Paleoecologists. AltaMira Press, Lanham, MD.

Piperno, D.R., Dillehay, T.D., 2008. Starch grains on human teeth reveal early broad crop diet in northern Peru. Proceedings of the National Academy of Sciences 105 (50), 19622–19627.

Piperno, D.R., Flannery, K.V., 2001. The earliest archaeological maize (*Zea mays* L.) from highland Mexico: new accelerator mass spectrometry dates and their implications. Proceedings of the National Academy of Sciences 98 (4), 2101–2103.

Piperno, D.R., Pearsall, D.M., 1998. The Origins of Agriculture in the Lowland Neotropics. Academic Press, San Diego.

Piperno, D.R., Ranere, A.J., Holst, I., Iriarte, J., Dickau, R., 2009. Starch grain and phytolith evidence for early ninth millennium B.P. maize from the central Balsas River valley, Mexico. Proceedings of the National Academy of Sciences 106 (13), 5020–5024.

Piperno, D.R., Weiss, E., Holst, I., Nadel, D., 2004. Processing of wild cereal grains in the Upper Palaeolithic revealed by starch grain analysis. Nature 430, 670–673.

Rosen, A.M., 2003. Middle Paleolithic plant exploitation: the microbotanical evidence. In: Henry, D.O. (Ed.), Neanderthals in the Levant. Behavioral Organization and the Beginnings of Human Modernity. Continuum, London, pp. 156–171.

Rottoli, M., Castiglioni, E., 2011. Plant offerings from Roman cremations in northern Italy: a review. Vegetation History and Archaeobotany 20, 495–506.

Runge, F., 2000. Opal-Phytolithe in den Tropen Afrikas. Books on Demand GmbH, Paderborn.

Torrence, R., 2006. Starch in sediments. In: Torrence, R., Barton, H. (Eds.), Ancient Starch Research. Left Coast Press, Walnut Creek, CA, pp. 145–176.

Weiss, E., Wetterstrom, W., Nadel, D., Bar-Yosef, O., 2004. The broad spectrum revisited: evidence from plant remains. Proceedings of the National Academy of Sciences 101 (26), 9551–9555.

Zarillo, S., Pearsall, D.M., Raymond, J.S., Tisdale, M.A., Dugane, J.Q., 2008. Directly dated starch residues document early formative maize (*Zea mays* L.) in tropical Ecuador. Proceedings of the National Academy of Sciences 105 (13), 5006–5011.

Panel Surveys: Uses and Applications

Greg J Duncan, Northwestern University, Evanston, IL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 11009–11015, © 2001, Elsevier Ltd.

Abstract

A panel survey makes similar measurements on the same units (e.g., individuals, firms) at different points in time. Analytic advantages of panel as opposed to cross-sectional survey data include the ability to measure components of individual change, the ability to collect a wider range of data on each sampled element, and the provision of high-quality intergenerational data and measurements of the frequency, timing, and duration of events. Problems with panel but not cross-sectional surveys are panel nonresponse and, in some cases, panel conditioning, where responses in a given interviewing round may be conditioned by participation in prior rounds of interviews. Examples of panel data analysis include poverty dynamics, the intergenerational transmission of economic status, unemployment and labor turnover, and antecedents of life events.

A panel survey makes similar measurements on the same sample units (e.g., individuals, firms) at different points in time (Duncan and Kalton, 1987). Panel surveys can be distinguished from: (1) cross-sectional surveys, which take measurements on sample units from a population of interest at a single point in time; and (2) repeated surveys, such as the US General Social Survey (Davis and Smith, 1994), in which similar measurements are made on samples from an equivalent population (e.g., defined by identical geographical boundaries or birth cohorts) at different points of time, but without attempting to ensure that any elements are included in more than one round of data collection.

A widely used alternative term for panel survey is longitudinal survey. However, it is useful to use 'longitudinal' only in the context of longitudinal *data*; such data may be collected by a panel survey, but they could also come from retrospective reports in a cross-sectional survey.

Most panel surveys follow their sample elements for the duration of the survey. Other panel designs include a rotating panel survey, in which sample elements have a restricted panel life; as they leave the panel, new elements are added. Examples of this are the monthly US Current Population Survey (US Bureau of Labor Statistics, 2000a) and the Canadian Labour Force Survey (Statistics Canada, 2000), in which the sample at each month is composed of a number of rotation groups, each of which is a self-contained sample from the population of interest. One rotation group drops out and one rotation group enters each month. A split panel survey is a combination of a panel and a repeated survey (Kish, 1986).

Panel surveys can differ markedly in the interval between rounds of data collection and the overall length of the survey. On the one hand, a consumer panel in market research may involve weekly contacts with respondents to collect their diary records of purchases. On the other hand, there are long-term panel surveys that may collect data only annually or even less frequently, and these surveys may continue for many years. Examples are the Panel Study of Income Dynamics (PSID), which had been collecting data on US households annually (biannually since 1997) since 1968 (Hill, 1992), and the British National Child Development Study, which started in 1958 and has collected data on the sample persons on five occasions in the 33 years to 1991 (Institute of Education, 2000).

Some panel surveys are concerned with population subgroups that have experienced the same event during the same time period, such as attending eighth grade in a given year, having been born in a particular week, or having been married during the same period. Since these sample universes are defined as cohorts, such studies are often called cohort studies. Examples include the various cohorts defined by age and sex in the US National Longitudinal Surveys (US Bureau of Labor Statistics, 2000b). A limitation of cohort studies is that their results relate formally only to the particular cohorts from which the sample was selected. An extension of the approach is the multiple cohort design, in which several cohorts are followed in similar panel surveys. Thus, the National Longitudinal Surveys have begun three youth cohort panel surveys, in 1968 (males) and 1968 (females), 1979, and 1997, while the Monitoring the Future studies have selected samples of US high school seniors each year, and follow each sample in a panel survey (Monitoring the Future, 2000).

Despite their lack of formal population representation, still other panel surveys are valuable because of their detailed interviews and often very long durations. Examples include Elder's (1999) 30-year follow-up of individuals who had attended Oakland, California elementary schools and were enrolled in the Oakland Growth Study when in fifth and sixth grades and Furstenberg's (Furstenberg et al., 1987) 30-year panel study of Baltimore teen mothers who had enrolled in a teen parenting program in 1965.

Uses of Panel Surveys

To appreciate the analytic advantages of panel surveys, it is important to realize that the time dimension complicates analyses because: (1) characteristics of population elements may change over time, for instance, an individual may be single one month and married the next; and (2) the population itself may change in composition, with both sample units entering into and exiting from the population.

These complications, especially the first, give rise to a variety of objectives for analysis (Kish, 1986; Duncan and Kalton, 1987), including:

1. Estimates of population parameters (e.g., monthly rates of unemployment, public opinion regarding gun control) at

 http://dx.doi.org/10.1016/B978-0-08-097086-8.44040-7

distinct time points. The resulting time series of such estimates may be the input for further analysis; see, for example, Smith (1994).

2. Measurement of net change, defined as change at the aggregate level. An example is the change in the UK unemployment rate between two consecutive months or change in US public opinion regarding gun control before and after a gun-related event.
3. Measurement of various components of individual change, including: (i) gross change – change at the individual level between two time points (e.g., relating the poverty status of sample members in a given year to their poverty status ten years later); (ii) average change, or trend, for each individual over a period of time, with the average serving to smooth out the effects of isolated perturbations and measurement errors; or (iii) instability for each individual (e.g., the stability of locus of control over 25 consecutive weeks from a sample of elderly in Eizenman et al., 1997).
4. Aggregation of data for individuals over time. An example is the collection of incomes on an annual basis to relate childhood-long economic well-being to children's developmental outcomes (Blau, 1999).
5. Measuring the frequency, timing, and duration of events occurring within a given time period. Examples are the proportion of persons who were ill in the past two weeks and the average duration of their illnesses, and the complete fertility histories of postmenopausal women.

If the composition of a population is sufficiently stable to be treated as static for the time period of interest, it may be possible to collect the data required to meet all of these objectives from a single cross-sectional survey conducted at the end of the period. Crucial here is that errors of measurement are acceptably small. Unfortunately, respondents' memories are notoriously inadequate when a substantial time is involved or when the characteristics or events have a low degree of salience for the respondents (Bound et al., 2000). For example, respondents are often unable to recall important events such as unemployment (Mathiowetz and Duncan, 1984) or hospitalizations (Cannell et al., 1981) within periods as short as several months, and they cannot reliably recall subjective experiences (Moss and Goldstein, 1979). Thus, event data often need to be collected at several points throughout the period.

Repeated surveys can meet some of these analytic goals. Their particular strength is that at each round of data collection they routinely select a sample of the population existing at that time. They thus provide a series of cross-sectional estimates that can be used for objective (1) above. They can also be used to measure overall net change (2). The major limitation of a repeated survey is that it does not yield data to satisfy objectives (3) and (4): since elements are not explicitly included in several rounds, none of the components of individual change can be measured and individual data cannot be aggregated across rounds. Data on the frequency and timing of events in a specified period, as in objective (5), can be collected in a repeated survey, but with the measurement error problems noted above.

The major advantage of a panel survey over a repeated survey is its much greater analytical potential. It enables components of individual change to be measured, objective (3), and also the summation of a variable across time, objective (4). By providing high-quality measurements of the frequency, timing, and duration of events, panel surveys provide data useful for the estimating of event-history models (Tuma and Hannan, 1984; Allison, 1982; Yamaguchi, 1991).

A panel survey can be much more efficient than a repeated survey for measuring net change, objective (b). Let the net change be $y_2 - y_1$, where $y_1$and y_2 are the means of the variable of interest at times 1 and 2. Then the variance of the net change is $V(y_2 - y_1) = V(y_1) + V(y_2) - 2\rho[V(y_1)V(y_2)]^{1/2}$ where ρ is the correlation between y_1 and y_2. In a repeated survey with two independent samples, $\rho = 0$, but in a panel survey ρ is the correlation between an individual's y-values on the two occasions, which is often quite high. In this case, a panel survey will yield a much more precise estimate of net change, objective (b), than will a repeated survey of the same size. On the other hand, as discussed below, panel nonresponse between times 1 and 2 may result in more bias in the estimation of y_2 in a panel than repeated survey.

A panel survey also permits the collection of a wider range of data on each sampled element than is possible with a repeated survey. The amount of data collected from any one respondent on one occasion is limited by the burden imposed; however, different data can be collected on different rounds of a panel survey, thus providing the possibility of accumulating an extremely wide range of variables for each sampled element. In the US Survey of Income and Program Participation (SIPP), for instance, core data on income and program participation are collected on each round, and additional data on different topics (e.g., health care utilization and financing, pension and retirement issues, housing conditions) are collected on specific rounds only (US Bureau of the Census, 2000).

Some long-term panels have important intergenerational components, with data on children and their family contexts taken in early waves and measurements of the adult attainments of the children taken in later waves. Examples include population panels such as the National Longitudinal Survey youth panels, the PSID, the British Household Panel Survey (Institute for Economic and Social Research, 2000), the German Socio-Economic Panel, and more specialized data collections such as the Oakland Growth Study (Elder, 1999).

Problems with Panel Surveys

By attempting repeated interviews with the same sample, panel surveys have problems not found in single or repeated cross-sectional designs, the most important of which is panel nonresponse (initial-wave respondents may not respond in later waves). An additional potential problem with panel surveys is panel conditioning, where responses in a given interviewing round may be conditioned by participation in prior rounds of interviews.

Methods developed to cope with nonresponse bias include minimizing nonresponse in panel surveys and developing statistical adjustments for existing nonresponse. Existing panel surveys typically devote sizeable resources to maintaining high response rates, and sometimes are quite successful. For example, the National Longitudinal Survey of Youth conducted interviews in 1991 with 89% of the respondents in its initial,

1979, interview (MaCurdy et al., 1998). Losses in the British National Survey of Health and Development amounted to only 12% after 26 years (Atkins et al., 1981).

Incentive payments, respondent reports, persuasion letters, using administrative data for tracing, and collecting extensive contact information (e.g., on friends and relatives not living in the household who would know of address and telephone number changes) help minimize these problems (Freedman et al., 1980; Clarridge et al., 1978; Call et al., 1982).

As with any survey, sizeable nonresponse in a panel survey gives rise to concerns about nonresponse bias. The situation with the first wave of a panel survey corresponds to that with a cross-sectional survey in that very limited information is available on the nonrespondents. The situation with later wave nonresponse in a panel survey is, however, different: in this case a good deal of information is available about later wave nonrespondents from their responses on earlier waves. The earlier wave information can be used to investigate the possibility of nonresponse bias and to develop imputation and weighting nonresponse adjustments that attempt to reduce the bias (Kalton, 1986; Lepkowski, 1989).

With regard to conditioning, there is ample evidence from several surveys that initial responses in a panel survey differ substantially from those given in subsequent waves (Bailar, 1975, 1979; Ghangurde, 1982). In the case of the US Current Population Survey, estimates of unemployment from households entering the sample for the first time are almost 10% larger than the average over all eight monthly reporting periods. It is not clear whether there is more response bias in the initial or subsequent waves, because the repeated contact with respondents has ambiguous effects on the quality of the data. The crucial question, as yet unanswered for most phenomena reported in surveys, is whether it is merely the reporting of behavior or the behavior itself that is affected by panel membership.

It may be that data collected in subsequent panel waves is less biased, because repeated contact increases the probability that respondents understand the purposes of the study and are thus increasingly motivated to make the effort necessary to give more accurate answers. On the other hand, there is evidence from a validation study (Traugott and Katosh, 1979) that extended participation in a panel study on election behavior not only increased the accuracy of responses on voting behavior but may indeed have increased the amount of voting, so that the behavior of the panel was no longer representative of the behavior of the population at large.

It seems unlikely that panel participation has pervasive behavioral effects, especially when changes in the behavior under investigation require more effort than making a trip to the polls. For example, economic behavior such as work effort, saving, commuting, and home ownership are all unlikely to be affected by responses to occasional interviews. Responses to attitudinal questions may be affected by panel membership if participation stimulates interest in the subject matter of the survey.

The limited membership in a rotating panel acts to reduce the problems of panel conditioning and panel loss in comparison with a nonrotating panel survey, and the continual introduction of new samples helps to maintain an up-to-date sample of a changing population. Rotating panels are used primarily for the estimation of cross-sectional parameters, objective (1), for the estimation of average values of population parameters across a period of time, objective (2), and for measuring net changes, objective (3). A rotating panel survey will generally provide more precise estimates of point of time and, especially, of change parameters than a repeated survey of the same size. Moreover, a rotating panel survey will sometimes have a cost advantage over a repeated survey. This will occur when it is cheaper to conduct a reinterview than an initial interview, as for instance is the case in the US Current Population Survey where initial interviews must be conducted by personal visit whereas reinterviews on some waves may be conducted by telephone (US Bureau of the Census, 1978).

The ability of a rotating panel survey to measure components of individual change, objective (3), and to aggregate data for individuals across time, objective (4), is clearly restricted. Since rotating panels are not intended to serve these objectives, they can be designed to avoid the heavy expense of following movers that occurs with nonrotating panel surveys. Thus, for instance, the Current Population Survey employs dwellings, not households or persons, as the sampled units, so that there is no need to follow households or persons moving between panel waves.

In a split panel survey, the panel survey component can be used to measure components of individual change, objective (3), and to aggregate data for individuals over time, objective (4). Its permanent overlap aids in the estimation of net change, objective (2), between *any* two waves whereas the overlap in a rotating panel survey aids only in the estimation of net change between certain prespecified waves.

Both rotating and split panel survey designs provide samples of new entrants to the population and the capacity to use their panel survey components to check on biases from panel conditioning and respondent losses.

The Problem of Changing Populations

The composition of almost every population of interest, whether of individuals, families, firms, or housing units, changes over time. Individuals enter the population when they are born, immigrate, or attain the age or other status that is used to define the population of interest, and depart when they die, emigrate, move into institutions such as jails, nursing homes, or the military, or in some other way lose that status. Families are 'born' when children leave their parents and set up their own independent households or when a divorce or separation breaks up one family into two; they 'die' when all members of the original household die or when two households are merged into a single one through marriage or less formal living arrangements. Over a survey's time span an element may both enter and leave the population, and may do so more than once, as for instance in the case of a person who has several stays in a nursing home during the course of a survey of the noninstitutionalized population. Comparable changes through construction and demolition, or incorporation, merger, and bankruptcy apply to populations of dwellings and firms.

All surveys over time must address the issues raised by these changes in population composition (Duncan and Kalton, 1987). In a repeated survey the cross-sectional parameters of

objective (1) relate to the population as it exists at the time the sample is drawn for the particular round of the survey. This procedure reflects both the changing values of the variables under study and the changing population over the time period involved. The estimates are thus average values for a form of average population over time.

Panel studies have adopted a variety of strategies for coping with changing populations. The simplest are panel studies of birth cohorts of individuals in the population, where as many members as possible of the original sample are followed throughout the panel period. In effect, these panels ignore all problems of population composition change except death. Disregarding nonresponse and immigration, these studies represent their respective cohorts as they age, while they gradually cease to represent individuals in the original age range. The National Longitudinal Survey cohorts of Older Men and Mature Women are examples. As discussed above, multiple-cohort panel designs avoid this problem by including in subsequent interviewing waves a sample of new entrants into the population.

Panel designs such as that of the PSID, the British Household Panel Survey, and the European Community Household Panel surveys, contain a mechanism for adding to their samples individuals and families that are 'born' into the population, so they have the potential for maintaining representative samples of nonimmigrant individuals and families throughout their duration. For example, the PSID began with a probability sample of dwellings that, in turn, provides a representative sample of subunits within those dwellings: households, families, subfamilies, transfer program recipiency units, and individuals. The selection probability of each of these units is identical to the selection probability of the dwelling itself. Other panels composed these subsamples with known selection probabilities through other means. With a properly specified set of rules regarding the definition of units, newly formed subunits of interest (including individuals) enter into the sample with known selection probabilities and reflect corresponding changes that are taking place in the population at large, for example, Hill (1992).

Because they originate from existing panel households, newly formed families and individuals are more 'clustered' by family lineage in these designs than in repeated cross-sectional samples; however, the statistical inefficiency of this clustering is heavily outweighed by the analytic advantages of being able to link these newly formed units with information on their families of origin. Thus, information on children setting up independent households can be matched to reliably reported information on their parents, information on ex-spouses can be linked together if the marital split occurred during the panel period, etc. Immigration during the panel period is usually ignored although can be a serious problem for a long-term panel survey confined to a city or a local area, but it is a less serious problem for a national panel survey.

Examples of What Panel Surveys Have Found

To help convey the analytic advantages of panel surveys, we list in this section some prominent examples of panel-survey based analyses.

Poverty and Welfare Dynamics

Trends in annual US poverty rates (i.e., the fraction of the population with family cash incomes below a set of 'poverty thresholds' that vary with family size) are tracked through cross-sectional estimates based on the Current Population Survey. Net change from one year to the next in these rates rarely amount to as much as a single percentage point (US Bureau of the Census, 1999), producing perceptions that 'the poor' are an ever-present and little-changing group.

By applying event history methods to the time series of annual poverty measurements for PSID sample members, Bane and Ellwood (1986) characterized the nature and determinants of poverty experiences by the length of their 'spells' (i.e., continuous periods of poverty). They find that a clear majority (60%) of families who first begin a poverty experience do so for short (i.e., spells of 1 or 2 years) periods of time. On the other hand, a subset of poor families (14%) have longer-run (8 or more years) spells, and the remaining one-quarter (26%) have medium-length (3–7 years) spells. Thus, there is no single 'correct' characterization of poverty – transitory or persistent – since poverty experiences are a mixture of transitory and long-term.

By combining individual spells into simulated lifetime histories, Bane and Ellwood (1994) provide estimates of the likely total number of years of receipt for families just starting to receive cash assistance from the US Aid to Families With Dependent Children program. They find a roughly even distribution of first-time welfare recipients across the three time intervals; roughly one-third (36%) have very short welfare experiences, a third (35%) medium-length experiences, and the final third (29%) long-term receipt. With welfare, as with poverty, heterogeneity is a key feature, with many families using AFDC as short-term insurance, but a substantial fraction of recipients receiving long-term assistance.

Intergenerational Transmission of Economic Status

Correlations in earning between fathers and sons is sometimes treated as a useful measure of intergenerational (im)mobility. Solon (1992) argues that there are two important sources of bias in attempts to estimate such correlations: single year proxies for long-run economic status contain transitory fluctuations that, under plausible assumptions, impart a downward bias to the correlation estimates, and samples used in previous studies (e.g., Mormon brothers, white twins who served in the armed forces) tend to be peculiarly homogeneous, also resulting under plausible assumptions in a downward bias to estimates based on homogeneous than population-based samples.

Data from both the NLSY (in Zimmerman, 1992) and PSID (in Solon, 1992) have been used to gauge the likely biases. For example, Solon et al. (1992) find that increasing the measurement interval for economic status from one to several years causes their estimate of the father-son earning correlation to jump from 0.25, a number consistent with past studies based on single-year earnings measurement and homogeneous samples, to 0.45. Thus, intergenerational economic mobility appears to be substantially lower than had been depicted in

studies based on homogeneous samples and single-year measures of income or earnings.

The more specialized panel studies of Elder and Furstenberg provide much more detail on the process of intergenerational mobility. Using 30-year follow-up data on the Oakland Growth Study cohorts, Elder (1999) found in these children of the Great Depression strong associations among economic hardship, parental psychological well-being, and adolescent well-being in intact families.

Fathers who experienced job loss and economic deprivation were more distressed psychologically and prone to explosive, rejecting, and punitive parenting. Preschool-age children in these families, especially boys, were more likely to exhibit problem behaviors, while adolescent girls were more likely to have lower feelings of self-adequacy and to be less goal-oriented. Elder (1999) speculated that the gender and age differences reflected different experiences in families during the deprived times. Adolescent boys reacted to family economic hardship by looking for economic opportunities outside of the home. This time spent taking on responsibility in extra-family contexts reduced the time they spent with their families and may have provided a buffer to the effects of negative family interactions. Younger children and girls were unable to seek such extra-familial opportunities as readily and therefore did not have access to buffering opportunities.

Furstenberg et al. (1987) take a detailed look at intra- and intergenerational mobility in a panel study of a poor teen mothers participating in a program in a Baltimore hospital. Their 20-year follow-up revealed considerably diversity in the mothers' economic and demographic statuses, with most self-supporting but very few stably married. Most (60%) of the daughters of the teen mothers had not themselves become teen mothers and three-quarters enjoyed at least a modest degree of economic success. In contrast, fewer than half of the sons of the teen mothers had attained successful economic trajectories.

Unemployment and Labor Turnover

Both unemployment and job turnover are most naturally thought about in a 'spells' framework: how long does a spell of unemployment last before employment recommences? How long does a spell of employment with one firm last before the worker leaves to take (or look for) alternative employment? Given the inherently longitudinal nature of these processes, it is not surprising that panel surveys have played an important role in research on these topics. While much research on this topic has used administrative records, a panel survey has the advantages of including those who do not receive unemployment benefits, for comparison to recipients, and it continues observing spells after benefits have been exhausted (i.e., it measures total spells, not just insured spells), so one can observe changes in job-finding rates after benefits are exhausted.

An example is Katz and Meyer's (1990) study of the impact of unemployment insurance on unemployment duration. Katz and Meyer find that probability of recall declines as the duration of a spell increases, for both recipients and nonrecipients. New-job rates, however, are upward-sloping for recipients but relatively flat for nonrecipients. Moreover, there are spikes in both the recall and new-job hazard rates around 26 and 39 weeks (when benefits were typically exhausted in their sample period) for recipients but not for nonrecipients. Thus, not only is there evidence that the approaching end of benefits conditions recipients' job finding, but it also conditions firms' recall policies.

Antecedents of Life Events

There is great interest in understanding the consequences of important life events such as divorce, widowhood, unemployment, and retirement. Cross-sectional data provide comparisons of, for example, the living standards and labor supply of divorced women and 'otherwise similar' married women, or the health of unemployed or widowed men with 'otherwise similar' employed or married men. But panel surveys provide valuable post- and, interestingly, pre-event data on the same individuals.

For example, Johnson and Skinner (1986) use data from the PSID to examine work hours of women between seven years before and four years after divorce or separation. Consistent with other research, they find that the average work hours of these women rose by half (from 1024 to 1551 hours) between one year before and four years after the divorce or separation. But Johnson and Skinner (1986) also found a big jump (from 744 to 1024 hours) over the six-year period *prior* to the divorce or separation.

Zick and Smith (1991) analyze the economic antecedents and consequences of widowhood by comparing family-income trajectories of a control group of intact couples with widowed women and men between five years before and five years after the spouse's death. They too find differences prior to the event, with income changes over the five years *prior* to the death accounting for about one-quarter of the control-widow difference at $t+5$ and for fully half of the control-widower difference. A clear message from both of these analyses is that impending life events such as divorce and death produce measurable economic and social changes long before the events actually occur. An implication is that fairly long panels are needed to capture the complete dynamic.

See also: Large-Scale Social Surveys, Analysis of; Longitudinal Research: Panel Retention; Sample Surveys: Cognitive Aspects of Survey Design; Survey Research Centers and Companies; Surveys and Polling: Ethical Aspects.

Bibliography

Allison, P., 1982. Discrete-time methods for the analysis of event histories. In: Leinhardt, S. (Ed.), Sociological Methodology 1982. Jossey-Bass, San Francisco, pp. 61–98.

Atkins, E., Cherry, N., Douglas, J.W.B., Kieman, K.E., Wadsworth, M.E.J., 1981. The 1946 British birth cohort: an account of the origins, progress, and results of the National Survey of Health and Development. In: Mednick, S.A., Baert, A.E. (Eds.), Prospective Longitudinal Research. Oxford University Press, Oxford, UK, pp. 5–30.

Bailar, B.A., 1975. The effects of rotation group bias on estimates from panel surveys. Journal of the American Statistical Association 70, 23–30.

Bailar, B.A., 1979. Rotation sample biases and their effects on estimates of change. Bulletin of the Statistics Institution 48 (2), 385–407.

Bane, M.J., Ellwood, D.T., 1986. Slipping in and out of poverty: the dynamics of spells. Journal of Human Resources 21, 1–23.

Bane, M.J., Ellwood, D.T., 1994. Welfare Realities. Harvard University Press, Cambridge, MA.
Blau, D., 1999. The effect of income on child development. Review of Economics and Statistics 81 (2), 261–276.
Bound, J., Brown, C., Mathiowetz, N., 2000. Measurement error in survey data. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 5. North Holland, Amsterdam.
Call, V.R.A., Otto, L.B., Spenner, K.I., 1982. Tracking Respondents: A Multi-Method Approach. Lexington Books, Lexington, MA.
Cannell, C.F., Miller, P.V., Oksenberg, L., 1981. Research on interviewing techniques. In: Leinhardt, S. (Ed.), Sociological Methodology, 1981. Jossey-Bass, San Francisco, pp. 389–437.
Clarridge, B.R., Sheehy, L.L., Hauser, T.S., 1978. Tracing members of a panel: A 17-year follow-up. In: Schuessler, K.F. (Ed.), Sociological Methodology, 1978. Jossey-Bass, San Francisco, pp. 185–203.
Davis, J., Smith, T.W., 1994. General Social Surveys, 1972–1994: Cumulative Codebook. National Opinion Research Center, Chicago.
Deutsches Institut für Wirtschaftsforchung, 2000. German Socio-Economic Panel Home Page. http://www.diw-berlin.de/soep/.
Duncan, G., Kalton, G., 1987. Issues of design and analysis of surveys across time. International Statistical Review 55, 97–117.
Eizenman, D.R., Nesselroade, J.R., Featherman, D.L., Rowe, J.W., 1997. Intra-individual variability in perceived control in an older sample: The MacArthur successful aging studies. Psychology and Aging 12, 489–502.
Elder, G., 1999. Children of the Great Depression: Social Change in Life Experience, 25th Anniversary Edition. Westview, Boulder, CO.
Freedman, D., Thornton, A., Camburn, D., 1980. Maintaining response rates in longitudinal studies. Sociological Methods and Research 9, 87–89.
Furstenberg Jr, F.F., Brooks-Gunn, J., Morgan, S.P., 1987. Adolescent Mothers in Later Life. Cambridge University Press, New York.
Ghangurde, P.D., 1982. Rotation group bias in the LFS estimates. Survey Methodology 8, 86–101.
Hill, M.S., 1992. The Panel Study of Income Dynamics. Sage, Newbury Park, CA.
Institute for Economic and Social Research, 2000. British Household Panel Survey Home Page. University of Essex. http://www.irc.essex.ac.uk/bhps.
Institute of Education, 2000. British National Child Development Study Home Page. Center for Longitudinal Studies. http://www.cls.ioe.ac.uk/Ncds/nhome.htm.
Johnson, W., Skinner, J., 1986. Labor supply and marital separation. American Economic Review 76 (3), 455–469.
Kalton, G., 1986. Handling wave nonresponse in panel surveys. Journal of Official Statistics 2, 303–314.
Katz, L., Meyer, B., 1990. The impact of potential duration of unemployment benefits on the duration of unemployment. Journal of Public Economics 41 (1), 45–72.
Kish, L., 1986. Timing of surveys for public policy. Australian Journal of Statistics 28, 1–12.
Lepkowski, J., 1989. The treatment of wave nonresponse in panel surveys. In: Kasprzyk, D., Duncan, G., Kalton, G., Singh, M. (Eds.), Panel Survey Design and Analysis. Wiley, New York, pp. 348–374.
MaCurdy, T., Mroz, T., Gritz, R.M., 1998. An evaluation of the National Longitudinal Survey on Youth. Journal of Human Resources 33 (2), 345–436.
Mathiowetz, N., 1984. Temporal patterns of response errors in retrospective reports of unemployment and occupation. In: Proc. Survey Res. Meth. Sect. American Statistical Association, Washington, DC, pp. 652–657.
Monitoring the Future, 2000. Monitoring the Future Home Page. http://monitoringthefuture.org/.
Moss, L., Goldstein, H., 1979. The Recall Method in Social Surveys. University of London Institute of Education, London.
Smith, T.W., 1994. Is there real opinion change? International Journal of Public Opinion Research 6, 187–203.
Solon, G., 1992. Intergenerational income mobility in the United States. American Economic Review 82 (3), 393–408.
Statistics Canada, 2000. Guide to the Labour Force Survey Statistics Canada web site. http://www.statcan.ca/.
Traugott, M., Katosh, K., 1979. Response validity in surveys of voting behavior. Public Opinion Quarterly 79, 359–377.
Tuma, N.B., Hannan, M.T., 1984. Social Dynamics: Models and Methods. Academic Press, Orlando, FL.
US Bureau of Labor Statistics, 2000a. Current Population Survey Home Page. http://www.bls.census.gov/cps/cpsmain.htm.
US Bureau of Labor Statistics, 2000b. National Longitudinal Surveys Home Page. http://stats.bls.gov/nlshome.htm.
US Bureau of the Census, 2000. Survey of Income and Program Participation Home Page. http://www.sipp.census.gov/sipp/sipphome.htm.
US Bureau of the Cenus, 1978. The Current Population Survey: Design and Methodology. US Government Printing Office, Washington, DC. Technical Paper 40.
US Bureau of the Cenus, 1999. Poverty in the United States: 1998. US Government Printing Office, Washington, DC, P60–207.
Yamaguchi, K., 1991. Event History Analysis, Applied Social Research Methods, Vol. 28. Sage, Newbury Park, CA.
Zick, C., Smith, K., 1991. Patterns of economic change surrounding the death of a spouse. Journal of Gerontology 46 (6), S310–S320.
Zimmerman, D.J., 1992. Regression toward mediocrity in economic stature. American Economic Review 82 (3), 409–429.

Panic Disorder

Heather K Hood and Martin M Antony, Ryerson University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Panic disorder (PD) is a condition associated with recurrent, unexpected panic attacks, characterized by sudden surges of physical arousal symptoms and fear. PD occurs frequently in the general population and is often associated with other psychological problems, including other anxiety disorders, depression, and substance-use problems. PD appears to stem from an interaction of biological, psychological, and social factors. This article begins with a review of diagnostic criteria, and then provides a brief summary of the biological and psychological factors involved in the onset and maintenance of PD. This article concludes with a review of effective medications and psychological treatments.

Panic disorder (PD) is an anxiety disorder associated with recurrent, unexpected panic attacks, characterized by sudden surges of physical arousal and fear. In addition, many individuals with PD develop agoraphobia to varying degrees. Agoraphobia is a fear of situations in which escape might be difficult or embarrassing, or in which help might be unavailable in the event of a panic attack.

The terms 'PD' and 'panic attack' first appeared in the diagnostic nomenclature with the publication of the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-III, American Psychiatric Association, 1980). Initially, panic attacks were seen primarily as a feature of agoraphobia. In the *Diagnostic and Statistical Manual of Mental Disorders-III* (DSM-III), agoraphobia could be diagnosed with or without panic attacks, and individuals who experienced unexpected panic attacks in the absence of agoraphobia received a diagnosis of PD without agoraphobia.

With the publication of the next edition of the DSM (e.g., DSM-IV, American Psychiatric Association, 1994), the relationship between PD and agoraphobia was reversed, such that PD was seen as the primary problem, and agoraphobia was viewed as a complication of PD. The disorder 'agoraphobia with panic attacks' was essentially replaced with 'PD with agoraphobia' to reflect the observation that, in clinical groups, agoraphobia can develop as a response to having unexpected panic attacks, rather than the other way around (Craske and Barlow, 1988).

More recently, with the release of DSM-5 (American Psychiatric Association, 2013), agoraphobia has been restored as a distinct diagnostic category. This change is supported by evidence indicating that 46–85% of people with agoraphobia do not have panic attacks (Faravelli et al., 2009). Other studies have not found a consistent temporal relationship between the onset of panic attacks and agoraphobic avoidance (e.g., Wittchen et al., 2008), calling into question the validity of agoraphobia as a secondary or subordinate diagnostic entity. In addition, agoraphobia without panic symptoms is associated with significant impairment and disability, and has a distinctive incidence, course, temporal progression, and treatment response compared to PD with or without agoraphobia (for a review, see Wittchen et al., 2010).

Definitions and Description

Panic Attack

DSM-5 (American Psychiatric Association, 2013) defines a panic attack as an abrupt surge of intense fear of discomfort that peaks within minutes. A panic attack must also be accompanied by at least 4 out of a list of 13 symptoms, which includes physical sensations (i.e., racing or pounding heart, sweating, trembling, shortness of breath, choking feelings, chest discomfort, nausea, dizziness, feelings of unreality or depersonalization, numbness or tingling sensations, hot flashes or chills) and cognitive symptoms (i.e., fear of losing control or going crazy, fear of dying).

Panic attacks are common in the general population and are also associated with a broad range of psychological problems. For example, people who experience intense worry may have a panic attack in response to a particular worrisome thought. Individuals who are fearful of a specific situation (e.g., seeing a spider, being in a high place) may experience panic attacks when confronted with these situations. In addition, the presence of cooccurring panic attacks in the context of other psychological disorders have been associated with increased severity, distress, and poor response to treatment of the principal diagnosis (Craske et al., 2010). Panic attacks may also occur in the absence of any obvious cue or trigger. In fact, the hallmark symptom of PD is the presence of panic attacks that occur out of the blue.

Panic Disorder

People who suffer from PD experience recurrent unexpected or uncued panic attacks. In addition, they must experience, for at least 1 month, one or more of the following: (1) persistent worry about having more panic attacks or worry about the implications or consequences of the attacks (e.g., that an attack will lead to a heart attack, death, fainting, diarrhea, vomiting, embarrassment, or some other catastrophe); or (2) a significant maladaptive change in behavior related to the attacks (e.g., avoidance of certain situations or activities). Finally, before assigning a diagnosis of PD, the clinician must establish that the panic attacks are not directly related to another psychological problem (e.g., a specific phobia), a medical condition

(e.g., hyperthyroidism), or use of a substance (e.g., cocaine, caffeine). PD may be diagnosed with or without agoraphobia.

Agoraphobia

A significant proportion of individuals suffering from PD develop agoraphobia, a fear of situations in which escape might be difficult, or in which help might not be available in the event of a panic attack (American Psychiatric Association, 2013). DSM-5 requires that individuals with agoraphobia report fear or anxiety in at least two of the following five situations: (1) using public transportation (e.g., driving, flying, or traveling on buses or trains), (2) being in open spaces (e.g., bridges, parking lots), (3) being in enclosed places (e.g., concerts, movies, shopping malls), (4) standing in a line or being in a crowd, or (5) being outside the home alone (American Psychiatric Association, 2013). For some individuals with agoraphobia, it is the distance from home, rather than the situation itself, that determines the level of fear and avoidance. For example, driving may be difficult, but only if the car ventures more than a few kilometers from home. For many people who suffer from agoraphobia, entering feared situations is easier if accompanied by a safe person, such as a spouse, family member, or close friend. In extreme cases, agoraphobia may be associated with a complete inability to leave the home, even when accompanied.

In addition to the obvious avoidance associated with agoraphobia, individuals with PD often engage in more subtle forms of avoidance. For example, a person might see a movie in a theater, but only if sitting in an aisle seat, near the exit. Or an individual might agree to ride a bus only while engaging in distracting behaviors such as reading or listening to a portable radio. Individuals with PD may also carry objects with them to feel safer, including medication, extra money in case of an emergency, or a portable telephone. Finally, they may engage in overprotective behaviors such as frequently checking their heart rate or blood pressure.

Epidemiology, Onset, and Demographics

PD begins on average when a person is in his or her mid- to late-20s, although onset can occur any time between childhood and late in life (Kessler et al., 2005). Risk factors include a period of life stress preceding the first attack, being female, and a history of either substance abuse or dependence (Antony and Swinson, 2000).

In the National Comorbidity Survey Replication (a study of over 9000 Americans in the general population), 4.7% of individuals reported symptoms meeting criteria for PD with or without agoraphobia at some time in their life (Kessler et al., 2006). In the same study, the lifetime prevalence of agoraphobia without history of PD was 1.4%. Although some studies report substantially higher prevalence rates of agoraphobia among community samples, agoraphobia is rarely diagnosed separately from PD in clinical settings (Wittchen et al., 2010).

People with PD report impairment across a broad range of life domains including work, recreation, and social functioning (Olatunji et al., 2007). Furthermore, PD is associated with enormous costs to sufferers and to society, in the form of increased health care utilization, lost wages, and time off work (e.g., Greenberg et al., 1999). In fact, PD is associated with higher rates of medical utilization at general and specialist settings compared to all other anxiety disorders (Deacon et al., 2008). In addition, PD is associated with a higher risk of developing other psychological problems, including other anxiety disorders, depression, certain personality disorders (e.g., dependent, avoidant), and substance-use disorders (for a review, see Antony and Swinson, 2000).

Biological Approaches to Understanding PD

Biological Models of PD

Most biological approaches to understanding PD have assumed a dysfunction in one or more neurotransmitter systems, including norepinephrine (NE), serotonin (5-HT), and cholecystokinin (CCK). In addition, theorists are increasingly seeking to understand the contribution of genetic and neuroanatomical factors involved in the pathogenesis of PD (e.g., Maron et al., 2010).

Perhaps the most influential biological model of PD in recent years has been Donald Klein's suffocation alarm model (Klein, 1993), which proposes that panic attacks are caused by a pathological misfiring of the suffocation alarm system. According to Klein, humans are equipped with two types of alarm reactions. The first is controlled by the hypothalamic–pituitary–adrenal axis and is responsible for dealing with real or perceived emergency situations. It is this alarm system that is activated when an individual is confronted with a feared object or situation.

The second alarm system is the suffocation alarm system, which is mediated by an area of the brain known as the *locus ceruleus*. This alarm system is triggered by increased levels of carbon dioxide in the blood, a sign that individuals might be in a small enclosed area, forced to breathe their own exhaled air, which is rich in carbon dioxide. Klein hypothesized that unexpected panic attacks occur when an individual has a pathologically lowered threshold for the activation of the suffocation alarm system. Presumably, the brains of such individuals are prone to mistakenly conclude that there is not enough air to breathe. Evidence regarding Klein's model has been mixed (for a review, see Antony and Swinson, 2000), particularly, with regard to the hypothesized causal role of respiratory symptoms (Vickers and McNally, 2005).

Evidence Regarding Biological Approaches to Panic

Despite inconsistent evidence for Klein's model, there are many studies supporting the view that biological processes are involved in the pathophysiology of PD. PD tends to run in families and evidence from twin studies consistently showing that genetics plays a significant role in the transmission of PD across generations (Finn and Smoller, 2001). Findings from brain imaging studies suggest that the temporal areas of the brain, including the hippocampus and amygdala, as well as the medial prefrontal cortex, are involved in the experience of panic attacks and fear (de Carvalho et al., 2010).

Evidence regarding the role of neurotransmitters in PD has included studies of altered receptor functioning, levels of particular neurotransmitters and related metabolites, and responses to substances that are known to cause changes in neurotransmitter levels. Among neurotransmitters, NE is the one most often implicated in PD and agoraphobia (Kalk et al., 2011), although other neurotransmitter systems, such as 5-HT (Maron and Shlik, 2006) and CCK (Zwanzger et al., 2012), appear to also play a role. There is evidence that levels of NE metabolites are elevated in the peripheral, but not in the central nervous system of individuals with PD (Kalk et al., 2011) and of elevated NE receptor sensitivity in people suffering from PD (Nutt, 1989). In addition, drugs that block the reuptake of NE, including the class of drugs known as serotonin–norepinephrine reuptake inhibitors (SNRIs) such as venlafaxine and duloxetine, appear to be effective for treating PD (Pollack et al., 2007).

Finally, a large number of studies have shown that certain substances that affect the NE system are more likely to trigger panic attacks in people with PD than in various comparison groups. Infusions of sodium lactate, carbon dioxide inhalation, and yohimbine injections have all been shown to trigger panic attacks. However, it should be noted that psychological and contextual factors, such as perceived control over the procedure, the presence of safety cues, and the expectations of physical arousal, can influence whether an individual will experience a panic attack following exposure to these substances (Sanderson et al., 1989; Telch et al., 2011).

Psychological Approaches to Understanding PD

Psychological Models of PD

Attempts to explain PD using a psychological framework have come from a broad range of theoretical perspectives, including cognitive theory, learning theory, and psychodynamic theory. However, the most influential of these models have been those developed by cognitive and behavioral theorists such as David M. Clark (1986) and David H. Barlow (1988, 2002).

From a cognitive behavioral perspective, PD is viewed as being maintained by a fear of benign physical sensations. For example, according to David Clark's cognitive model of PD, panic attacks stem from a tendency to misinterpret normal physical sensations (e.g., racing heart, dizziness) as indicating some immediate danger. As a result of interpreting these sensations as being threatening, the symptoms escalate, leading to further catastrophic interpretations and predictions. Very quickly, the symptoms spiral into a full-blown panic attack, seemingly out of the blue. Clark and other cognitive theorists acknowledge that biological factors may also contribute to the experience of panic. The sensations that initially start the panic cycle may be caused by caffeine, intense emotions, exercise, hormonal changes, fatigue, hunger, hyperventilation, attention to bodily feelings, or other factors. Consistent with the cognitive behavioral view, 70% of individuals with PD report that their panic attacks are triggered by a physical sensation (Breitholtz et al., 1998).

Evidence Regarding Psychological Approaches to Panic

A number of studies have confirmed that people with PD are fearful of the physical sensations associated with panic attacks, compared to people without PD and people with other anxiety disorders (McNally, 2002). In addition, researchers have demonstrated that people with PD are more likely than others to interpret ambiguous physical sensations as a sign of some impending physical or mental disaster and to believe these thoughts quite strongly (Clark et al., 1997).

Numerous studies have also found that people with PD have information-processing biases that are consistent with their fears. For example, a number of studies have found that PD is associated with a heightened ability mentally to track physical sensations such as heart rate (e.g., van der Does et al., 2000). In addition, results from studies of attention and memory have also been consistent with psychological models of PD. Compared to people without PD, individuals with PD attend more to threatening faces and words (Reinecke et al., 2011), and are more likely to show threat-related biases in explicit (i.e., conscious, intentional recollection), but not implicit (i.e., unintentional, automatic processing), memory (Mitte, 2008).

In summary, individuals with PD are typically more fearful of and more vigilant for panic-related sensations, compared to people without PD. PD is also associated with a tendency to hold fearful beliefs about panic-related sensations, and to show biases in attention and memory for threat-related information.

Treatment of PD

The recently updated American Psychiatric Association (2009) practice guidelines for the treatment of PD discuss important issues in the assessment and treatment of this condition, and reviews the main approaches that have been shown to be useful for treating PD. Essentially, these approaches include treatment with medications, particularly selective serotonin reuptake inhibitors (SSRIs) and SNRIs medications, and treatment with cognitive behavior therapy (CBT). A review of studies on the relative and combined effects of these interventions is available elsewhere (e.g., Koerner et al., 2012). A brief summary of the major findings is provided below.

Medication Treatments

A large number of controlled clinical trials have established that certain medications are useful for treating PD with and without agoraphobia. Controlled trials have established that SSRIs (e.g., fluoxetine, fluvoxamine, sertraline, paroxetine, citalopram) and SNRIs (e.g., venlafaxine) are also useful for treating PD; in fact, the SSRIs and SNRIs are now often considered to be the pharmacological treatments of choice for this condition because of the lower rates of adverse effects and withdrawal symptoms with SSRIs and SNRIs compared to some other medication classes (Mitte, 2005). Although tricyclic antidepressants, and imipramine in particular, have

also been shown to be an effective treatment for PD (e.g., Barlow et al., 2000), they are often associated with pronounced adverse side effects, limiting their clinical utility (Mavissakalian et al., 2002). In addition, monoamine oxidase inhibitors (MAOIs), such as phenelzine, have been demonstrated in a limited number of controlled trials to be useful in treating PD (Buigues and Vallejo, 1987). However, MAOIs are associated with an increased risk of drug interactions, necessary dietary restrictions when taking the medications, and only modest response rates; as such, MAOIs are generally not recommended as a first line treatment for PD (American Psychiatric Association, 2009). Studies with high-potency short-acting (e.g., alprazolam) and long-acting (e.g., clonazepam) benzodiazepine anxiolytics have consistently shown these medications to be effective as well (Moylan et al., 2011), although there remains a significant risk of withdrawal and rebound anxiety upon discontinuation (Chouinard, 2004).

There is now consistent evidence that SSRIs and SNRIs are preferred medication options for PD. However, other medication classes may be preferable for some patients. Choosing among medications for a particular individual usually depends on factors such as potential side effects, possible problems during discontinuation (e.g., rebound anxiety), the patient's previous history of treatment response, the cost of the drug, and potential interactions with other medications, substances, or medical illnesses.

Cognitive Behavior Therapy

CBT for PD includes five main types of strategies: (1) psychoeducation about the nature of anxiety and panic; (2) exposure to feared situations such as crowds, public transportation, and being away from home alone (i.e., *in vivo* exposure); (3) exposure to exercises that trigger feared sensations (e.g., overbreathing or spinning to induce feelings of dizziness); (4) cognitive restructuring (i.e., systematically considering evidence that contradicts fearful thinking); and (e) relaxation-based strategies (e.g., learning to slow down one's breathing). Most empirically-supported CBT protocols (e.g., Barlow and Craske, 2006; Clark et al., 1999), include various combinations of these strategies.

A number of comprehensive reviews on CBT for PD have been published (e.g., Koerner et al., 2012; Taylor, 2000). Generally, there is little evidence that it is necessary to include all five treatment components to achieve maximum results. *In vivo* exposure to feared situations appears to be particularly important for treating individuals who are agoraphobic. Recently, the inclusion of relaxation-based strategies, including breathing retraining, have been called into question as studies have failed to demonstrate an additive effect of including breathing retraining into CBT treatment protocols (Schmidt et al., 2000).

Although CBT for PD typically consists of 10–15 weekly sessions, a number of recent studies suggest that gains can be made relatively quickly, with a minimal investment of therapist time. For example, therapist-administered CBT via telephone (Swinson et al., 1995) and video-conferencing (Bouchard et al., 2004) have been successfully used to treat people who do not have easy access to a therapist. In addition, internet-based (Carlbring et al., 2005) and bibliotherapy (Febbraro, 2005) self-help have also been reported to be an effective alternative to therapist-administered CBT requiring minimal therapist contact (for a review, see Lewis et al., 2012). Further, a single session of CBT has been shown to prevent the onset of PD in a significant percentage of individuals who present to hospital emergency rooms with a panic attack (Swinson et al., 1992). Finally, a number of studies have found that brief CBT lasting only five sessions (Otto et al., 2012) can produce clinically significant improvements in PD symptoms comparable in magnitude to those obtained with standard treatment protocols (*see* Cognitive Behavioral Therapy).

Combination Treatments

Several meta-analyses of randomized controlled trials have consistently demonstrated that medications, CBT, and their combinations are effective for treating PD (e.g., Furukawa et al., 2006; Mitte, 2005). Although some studies have demonstrated a slight advantage for combination treatments at the end of the acute treatment phase, it should be noted that most studies have focused primarily on the short-term effects of treatment over a period of only 2 to 3 months, and combined treatments are no more effective than standard monotherapy at long-term follow-up (Furukawa et al., 2006). In fact, the studies that have examined the relative and combined effectiveness of CBT with either imipramine (Barlow et al., 2000), alprazolam (Marks et al., 1993), or fluvoxamine (Sharp et al., 1996) have consistently found that individuals who received CBT (alone or in combination with medication) were more likely to maintain their gains over the long term than individuals who received only medication or placebo treatment. In addition, CBT has been used successfully to prevent relapse and facilitate discontinuation of benzodiazepines among individuals with PD (Otto et al., 2010). As such, it has been suggested that combined approaches may be most effective when CBT is introduced during discontinuation of pharmacological treatment to protect against relapse (Otto et al., 2009).

Taken together, the literature on treating PD suggests that CBT and pharmacotherapy are both effective treatments in the short term. However, there is evidence that CBT is a more effective treatment for ensuring that gains are maintained over the long term.

See also: Anxiety and Anxiety Disorders; Anxiety and Fear Conditioning, Neural Basis of; Fear Conditioning: Overview; Fear-Potentiated Startle; Fear: Psychological and Neural Aspects.

Bibliography

American Psychiatric Association, 1980. Diagnostic and Statistical Manual of Mental Disorders, third ed. (DSM-III). American Psychiatric Association, Washington, DC.

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders, fourth ed. (DSM-IV). American Psychiatric Association, Washington, DC.

American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders, fourth ed., text revision. (DSM-IV-TR). American Psychiatric Association, Washington, DC.

American Psychiatric Association, 2009. Practice Guideline for the Treatment of Patients with Panic Disorder. http://psychiatryonline.org/guidelines.aspx (accessed 17.04.13.).

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. (DSM-IV-TR). American Psychiatric Association, Washington, DC.

Antony, M.M., Swinson, R.P., 2000. Phobic Disorders and Panic in Adults: A Guide to Assessment and Treatment. American Psychological Association, Washington, DC.

Barlow, D.H., 1988. Anxiety and its Disorders: The Nature and Treatment of Anxiety and Panic. Guilford Press, New York.

Barlow, D.H., 2002. Anxiety and its Disorders: The Nature and Treatment of Anxiety and Panic, second ed. Guilford Press, New York.

Barlow, D.H., Gorman, J.M., Shear, M.K., Woods, S.W., 2000. A randomized controlled trial of cognitive-behavioral treatment vs imipramine and their combination for panic disorder: primary outcome results. Journal of the American Medical Association 283, 2529–2536.

Barlow, D.H., Craske, M.G., 2006. Mastery of Your Anxiety and Panic. Workbook, fourth ed. Oxford University Press, New York.

Bouchard, S., Paquin, B., Payeur, R., Allard, M., Rivard, V., Fournier, T., Renaud, Lapierre, J., 2004. Delivering cognitive-behavior therapy for panic disorder with agoraphobia in videoconference. Telemedicine Journal and e-Health 10, 13–25.

Breitholtz, E., Westling, B.E., Öst, L.G., 1998. Cognitions in generalized anxiety disorder and panic disorder. Journal of Anxiety Disorders 12, 567–577.

Buigues, J., Vallejo, J., 1987. Therapeutic response to phenelzine in patients with panic disorder and agoraphobia with panic attacks. Journal of Clinical Psychiatry 48, 55–59.

Carlbring, P., Nilsson-Ihrfelt, E., Waara, J., Kollenstam, C., Buhrman, M., Kaldo, V., Söderberg, Ekselius, Andersson, G., 2005. Treatment of panic disorder: live therapy vs self-help via the internet. Behaviour Research and Therapy 43, 1321–1333.

de Carvalho, M.R., Rozenthal, M., Nardi, A.E., 2010. The fear circuitry in panic disorder and its modulation by cognitive-behaviour therapy interventions. World Journal of Biological Psychiatry 11, 188–198.

Chouinard, G., 2004. Issues in the clinical use of benzodiazepines: potency, withdrawal, and rebound. Journal of Clinical Psychiatry 65, 7–12.

Clark, D.M., 1986. A cognitive approach to panic. Behaviour Research and Therapy 24, 461–470.

Clark, D.M., Salkovskis, P.M., Hackmann, A., Wells, A., Ludgate, J., Gelder, M., 1999. Brief cognitive therapy for panic disorder: a randomized controlled trial. Journal of Consulting and Clinical Psychology 67, 583–589.

Clark, D.M., Salkovskis, P.M., Öst, L.G., Breitholtz, E., Koehler, K.A., Westling, B.E., Jeavons, A., Gelder, M., 1997. Misinterpretation of body sensations in panic disorder. Journal of Consulting and Clinical Psychology 65, 203–213.

Craske, M.G., Barlow, D.H., 1988. A review of the relationship between panic and avoidance. Clinical Psychology Review 8, 667–685.

Craske, M.G., Kircanski, K., Epstein, A., Wittchen, H., Pine, D.S., Lewis-Fernández, R., Hinton, D., 2010. Panic disorder: a review of DSM-IV panic disorder and proposals for DSM-V. Depression and Anxiety 27, 93–112.

Deacon, B., Lickel, J., Abramowitz, J.S., 2008. Medical utilization across the anxiety disorders. Journal of Anxiety Disorders 22, 344–350.

van der Does, A.J.W., Antony, M.M., Barsky, A.J., Ehlers, A., 2000. Heartbeat perception in panic disorder: a re-analysis. Behaviour Research and Therapy 38, 47–62.

Faravelli, C., Furukawa, T.A., Truglia, E., 2009. Panic disorder. In: Andrews, G., Charney, D.S., Sirovatka, P.J., Regier, D.A. (Eds.), Stress-Induced and Fear Circuitry Disorders. American Psychiatric Association, Arlington, VA, pp. 31–58.

Febbraro, G.A.R., 2005. An investigation into the effectiveness of bibliotherapy and minimal contact interventions in the treatment of panic attacks. Journal of Clinical Psychology 61, 763–779.

Finn, C.T., Smoller, J.W., 2001. The genetics of panic disorder. Current Psychiatry Reports 3, 131–137.

Furukawa, T.A., Watanabe, N., Churchill, R., 2006. Psychotherapy plus antidepressant for panic disorder with or without agoraphobia: systematic review. British Journal of Psychiatry 188, 305–312.

Greenberg, P.E., Sisitsky, T., Kessler, R.C., Finkelstein, S.N., Berndt, E.R., Davidson, J.R.T., Ballenger, J.C., Fyer, A.J., 1999. The economic burden of anxiety disorders in the 1990s. Journal of Clinical Psychiatry 60, 427–435.

Kalk, N.J., Nutt, D.J., Lingford-Hughes, A.R., 2011. The role of central noradrenergic dysregulation in anxiety disorders: evidence from clinical studies. Journal of Psychopharmacology 25, 3–16.

Kessler, R.C., Berglund, P., Demler, O., Jin, R., Merikangas, K.R., Walters, E.E., 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. Archives of General Psychiatry 62, 593–602.

Kessler, R.C., Chiu, W.T., Jin, R., Ruscio, A.M., Shear, K., Walters, E.E., 2006. The epidemiology of panic attacks, panic disorder, and agoraphobia in the national comorbidity survey replication. Archives of General Psychiatry 63, 415–424.

Klein, D.F., 1993. False suffocation alarms, spontaneous panics, and related conditions: an integrative hypothesis. Archives of General Psychiatry 50, 306–317.

Koerner, N., Vorstenbosch, V., Antony, M.M., 2012. Panic disorder. In: Sturmey, P., Hersen, M. (Eds.), Handbook of Evidence-Based Practice in Clinical Psychology, Adult disorders, vol. II. John Wiley and Sons, Hoboken, NJ, pp. 285–311.

Lewis, C., Pearce, J., Bisson, J.I., 2012. Efficacy, cost-effectiveness and acceptability of self-help interventions for anxiety disorders: systematic review. British Journal of Psychiatry 200, 15–21.

Marks, I.M., Swinson, R.P., Basoglu, M., Kuch, K., Noshirvani, H., O'Sullivan, G., Lelliott, P.T., Kirby, M., McNamee, G., Sengun, S., Wickwire, K., 1993. Alprazolam and exposure alone and combined in panic disorder with agoraphobia: a controlled study in London and Toronto. British Journal of Psychiatry 162, 776–787.

Maron, E., Hettema, J.M., Shlik, J., 2010. Advances in molecular genetics of panic disorder. Molecular Psychiatry 15, 681–701.

Maron, E., Shlik, J., 2006. Serotonin function in panic disorder: important, but why? Neuropsychopharmacology 31, 1–11.

Mavissakalian, M., Perel, J., Guo, S., 2002. Specific side effects of long-term imipramine management of panic disorder. Journal of Clinical Psychopharmacology 22, 155–161.

McNally, R.J., 2002. Anxiety sensitivity and panic disorder. Biological Psychiatry 52, 938–946.

Mitte, K., 2005. A meta-analysis of the efficacy of psycho-and pharmacotherapy in panic disorder with and without agoraphobia. Journal of Affective Disorders 88, 27–45.

Mitte, K., 2008. Memory bias for threatening information in anxiety and anxiety disorders: a meta-analytic review. Psychological Bulletin 134, 886–911.

Moylan, S., Staples, J., Ward, S.A., Rogerson, J., Stein, D.J., Berk, M., 2011. The efficacy and safety of alprazolam versus other benzodiazepines in the treatment of panic disorder. Journal of Clinical Psychopharmacology 31, 647–652.

Nutt, D.J., 1989. Altered central α2-adrenoceptor sensitivity in panic disorder. Archives of General Psychiatry 46, 165–169.

Olatunji, B.O., Cisler, J.M., Tolin, D.F., 2007. Quality of life in the anxiety disorders: a meta-analytic review. Clinical Psychology Review 27, 572–581.

Otto, M.W., Behar, E., Smits, J.A.J., Hofmann, S.G., 2009. Combining Pharmacological and Cognitive Behavioral Therapy in the Treatment of Anxiety Disorders. Oxford University Press, New York, pp. 429–440.

Otto, M.W., Tolin, D.F., Nations, K.R., Utschig, A.C., Rothbaum, B.O., Hofmann, S.G., Smits, J.A.J., 2012. Five sessions and counting: considering ultra-brief treatment for panic disorder. Depression and Anxiety 29, 465–470.

Otto, M.W., McHugh, R.K., Simon, N.M., Farach, F.J., Worthington, J.J., Pollack, M.H., 2010. Efficacy of CBT for benzodiazepine discontinuation in patients with panic disorder: further evaluation. Behaviour Research and Therapy 48, 720–727.

Pollack, M.H., Lepola, U., Koponen, H., Simon, N.M., Worthington, J.J., Emilien, G., Tzanis, Salinas, Whitaker, Gao, B., 2007. A double-blind study of the efficacy of venlafaxine extended-release, paroxetine, and placebo in the treatment of panic disorder. Depression and Anxiety 24, 1–14.

Reinecke, A., Cooper, M., Favaron, E., Massey-Chase, R., Harmer, C., 2011. Attentional bias in untreated panic disorder. Psychiatry Research 185, 387–393.

Sanderson, W.C., Rapee, R.M., Barlow, D.H., 1989. The influence of an illusion of control on panic attacks induced via inhalation of 5.5% carbon dioxide-enriched air. Archives of General Psychiatry 46, 157–162.

Schmidt, N.B., Woolaway-Bickel, K., Trakowski, J., Santiago, H., Storey, J., Koselka, M., Cook, J., 2000. Dismantling cognitive-behavioral treatment for panic disorder: questioning the utility of breathing retraining. Journal of Consulting and Clinical Psychology 68, 417–424.

Sharp, D.M., Power, K.G., Simpson, R.J., Swanson, V., Moodie, E., Anstee, J.A., Ashford, J.J., 1996. Fluvoxamine, placebo, and cognitive behaviour therapy used alone and in combination in the treatment of panic disorder and agoraphobia. Journal of Anxiety Disorders 10, 219–242.

Swinson, R.P., Fergus, K.D., Cox, B.J., Wickwire, K., 1995. Efficacy of telephone-administered behavioral therapy for panic disorder with agoraphobia. Behaviour Research and Therapy 33, 465–469.

Swinson, R.P., Soulios, C., Cox, B.J., Kuch, K., 1992. Brief treatment of emergency room patients with panic attacks. American Journal of Psychiatry 149, 944–946.

Taylor, S., 2000. Understanding and Treating Panic Disorder: Cognitive-Behavioural Approaches. John Wiley and Sons, New York.

Telch, M.J., Harrington, P.J., Smits, J.A., Powers, M.B., 2011. Unexpected arousal, anxiety sensitivity, and their interaction on CO_2-induced panic: further evidence for the context-sensitivity vulnerability model. Journal of Anxiety Disorders 25, 645–653.

Vickers, K., McNally, R.J., 2005. Respiratory symptoms and panic in the national comorbidity survey: a test of Klein's suffocation false alarm theory. Behaviour Research and Therapy 43, 1011–1018.

Wittchen, H., Gloster, A.T., Beesdo-Baum, K., Fava, G.A., Craske, M.G., 2010. Agoraphobia: a review of the diagnostic classificatory position and criteria. Depression and Anxiety 27, 113–133.

Wittchen, H., Nocon, A., Beesdo, K., Pine, D.S., Höffer, M., Lieb, R., Gloster, A.T., 2008. Agoraphobia and panic: prospective longitudinal relations suggest a rethinking of diagnostic concepts. Psychotherapy and Psychosomatics 77, 147–157.

Zwanzger, P., Domschke, K., Bradwejn, J., 2012. Neuronal network of panic disorder: the role of the neuropeptide cholecystokinin. Depression and Anxiety 29, 762–774.

Panic Disorder and Agoraphobia Across the Lifespan

Katharina Meyerbröker, Department of Psychiatry, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands
Mark B Powers, Behavioral & Integrative Treatment Development Lab, Institute for Mental Health Research, University of Texas at Austin, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article is aimed to give a short overview over the most important features and research findings about panic disorder and agoraphobia. The diagnostic criteria will be discussed reviewing the changes in the history of the Diagnostic Statistical Manual for Mental Disorders. Epidemiology will be reviewed within development and maintenance factors. Different theoretical models will be discussed and the evidence-based treatment approach will be explained. Possible mechanisms of change are explained and the article ends with a short overview of clinical implications and future directions of research.

Definition and Diagnostic Criteria

Panic Disorder

Panic disorder and agoraphobia can be disabling conditions, which are associated with significant interference in occupational and social functioning (Schmidt et al., 2013). Certain degrees of anxiety are adaptive in daily life and can protect us when confronted with dangerous situations. Thus, we can think of the continuum from worry to anxiety to panic as an alarm system designed to protect us from danger. However, when this alarm system goes off even in relatively safe situations and we fear the alarm system itself, this can have a major negative impact on the individual's daily functioning.

According to the Diagnostic and Statistical Manual of Mental Disorders fifth edition (DSM-5), panic disorder refers to experiencing recurrent and unexpected panic attacks (Criterion A). A panic attack is an abrupt surge of intense fear or intense discomfort that reaches a peak within minutes, and during which time four or more of a list of 13 physical and cognitive symptoms occur (American Psychiatric Association (APA), 2013). This includes palpitations, pounding heart, or accelerated heart rate; sweating; trembling or shaking; sensations of shortness of breath or smothering; feelings of choking; chest pain or discomfort; nausea or abdominal distress; feeling dizzy, unsteady, lightheaded, or faint; chills or heat sensations; paresthesia; derealization or depersonalization; fear of losing control; and fear of dying. The term 'recurrent' refers to have experienced more than one panic attack and 'unexpected' refers to the fact that no obvious trigger has elicited the panic attack. Panic attacks can even occur when the individual is relaxing or during nighttime. The distinction between expected and/or unexpected panic attacks has to be made carefully, based on open and structured questioning about the attacks and circumstances. In the United States and Europe, approximately one-half of individuals with panic disorder have expected panic attacks as well as unexpected panic attacks (Craske et al., 2010). Thus, according to the DSM-5, the presence of expected panic attacks does not rule out the diagnosis of panic disorder (APA, 2013).

However, most patients with panic disorder experience a mixture of different types of panic attacks (Schmidt et al., 2014). During the onset of panic disorder, panic attacks occur spontaneously and unexpectedly. However, as time progresses, patients are often able to establish a link between panic and a variety of situations, which trigger their panic attacks at an earlier stage, so that panic attacks appear to be more situationally bound over the course of panic disorder.

The frequency and severity of panic attacks can vary widely and depend largely on the individual coping style and avoidance behavior. In panic disorder, experiencing panic attacks is accompanied by frequent worries about the next panic attack and the possible consequences. These consequences include the presence of a life-threatening illnesses (e.g., cardiac disease, seizure disorder) or social concerns, such as embarrassment or fear of being judged negatively by others because of visible panic symptoms and concerns about mental functioning such as 'going crazy' or losing control (Criterion B) (APA, 2013). Often maladaptive behaviors occur to minimize the occurrence of panic attacks. Examples include avoiding physical exertion, reorganizing daily life to ensure that help is available in the event of a panic attack, restricting usual daily activities, and avoiding situations, such as using public transportation or shopping. If intense fear of multiple situations is present, a separate diagnosis of agoraphobia is given.

Agoraphobia

The essential feature of agoraphobia is marked or intense fear or anxiety triggered by real or anticipated exposure to a wide range of situations (Criterion A; APA, 2013). The diagnosis requires the occurrence of symptoms in at least two of the following five situations: (1) using public transportation, such as automobiles, buses, trains, ships, or planes; (2) being in open spaces, such as parking lots, marketplaces, or bridges; (3) being in enclosed spaces, such as shops, theaters, or cinemas; (4) standing in line or being in a crowd; or (5) being outside of the home alone (APA, 2013). When experiencing fear and anxiety cued by such situations, individuals typically experience thoughts that something terrible might happen (Criterion B). Individuals frequently believe that escape from such situations might be difficult (e.g., 'cannot get out of here') or that help might be unavailable (e.g., 'there is nobody to help me') when paniclike symptoms or other incapacitating or

 http://dx.doi.org/10.1016/B978-0-08-097086-8.21014-3

embarrassing symptoms occur. 'Paniclike symptoms' refer to any of the 13 symptoms included in the criteria for panic attack, such as dizziness, faintness, and fear of dying. 'Other incapacitating or embarrassing symptoms' include symptoms such as vomiting and inflammatory bowel symptoms, as well as, in older adults, a fear of falling or, in children, a sense of disorientation and getting lost.

Fear or anxiety is evoked nearly every time the individual comes into contact with the feared situation (Criterion C). Thus, an individual who becomes anxious only occasionally in an agoraphobic situation (e.g., when standing in line on only one out of every five occasions) would not be diagnosed with agoraphobia (Craske et al., 2010; Wittchen et al., 2010).

From DSM-IV to DSM-5

The definition and diagnostic criteria of panic disorder and agoraphobia have undergone significant changes over time in the DSM (Schmidt et al., 2014). In particular, the description of the relationship between panic disorder and agoraphobia has undergone interesting changes within the DSM. While Marks (1970) originally suggested to classify agoraphobia as a phobic disorder arising from fears of public places that may or may not occur with panic attacks, as was done in the DSM-III (American Psychiatric Association (APA), 1980), over time researchers increasingly recognized that agoraphobia can occur as a consequence of panic attacks and in the DSM-III-R and DSM-IV the relationship between these conditions was reversed. Agoraphobia was considered to be a consequence of panic attacks and thus, agoraphobic behaviors were classified more as panic-related sequelae (Frances et al., 1993; Goldstein and Chambless, 1978; McNally, 1994).

In the DSM-5 (APA, 2013), agoraphobia is classified as a distinct disorder from panic disorder. This means that patients can receive either a diagnosis of panic disorder or a diagnosis of agoraphobia, or a comorbid diagnosis of panic disorder and agoraphobia. The separation of panic disorder and agoraphobia is based on large cohort and prospective studies wherein it is shown that agoraphobia exists in individuals who never experienced any panic attacks or paniclike symptoms (Faravelli et al., 2008; Fava and Morton, 2009; Wittchen et al., 1998). For a recent review see Emmelkamp and Powers (2009). A detailed analysis of the prospective reports as well as analysis of the incidence patterns of agoraphobia indicate that a substantial percentage of individuals with this diagnosis show no history of panic or paniclike symptoms (Wittchen et al., 2008). However, fairly often, panic disorder and agoraphobia are comorbid conditions. For a more complete discussion on the diagnostic changes between DSM-IV and DSM-5 with regard to the diagnosis of panic attacks, panic disorder, and agoraphobia see Schmidt et al. (2013).

Prevalence

The 12-month prevalence of panic disorder ranges between 1% and 3% across the United States (Kessler et al., 2012, 2005) and Europe (Goodwin et al., 2005; De Graaf et al., 2012). The 12-month prevalence of agoraphobia is approximately 1.7% in adolescents and adults (Kessler et al., 2012; Wittchen et al., 2011). Onset of agoraphobia can occur during childhood but the most vulnerable periods to develop the disorder are late adolescence and early adulthood (Beesdo et al., 2009; Bittner et al., 2007). The gender ratio is 2:1 with females experiencing twice as often agoraphobia as men (Wittchen et al., 2010). This may be for several reasons: (1) males (especially young) tend to be less willing to openly discuss agoraphobic fears and avoidance behavior, (2) because of this desire to not appear 'weak,' they might expose themselves to the feared situations with fear reduction over time as an unintended consequence, or (3) they use other methods to enter the feared situations (alcohol/substances).

Development and Maintenance Factors

Development and Avoidance Behavior

Epidemiological research has investigated potential risk factors for panic disorder. As in most psychiatric disorders, a diathesis-stress model is commonly used to explain the development and maintenance of the disorder (Roy-Byrne et al., 2006).

While panic attacks are quite common in the general population, only a small percentage of people develops panic disorder (Emmelkamp and Powers, 2010). While most people who experience a panic attack during their lifetime just go on with their life, a small percentage starts worrying about their next panic attack and their perceived consequences. There is some evidence that this panic-related apprehension is one of the key factors in developing panic disorder (Telch et al., 1989). Panic appraisal is usually considered to consist of three different aspects: (1) perceived likelihood of panic occurrence, (2) the perceived negative consequences, and (3) the perceived efficacy in panic coping (Telch et al., 1989). There are some indications that these aspects of panic appraisal are the best predictor for agoraphobic avoidance (Smits et al., 2004; Telch et al., 1989; Cho et al., 2007).

Other researchers emphasize the importance of cognitive factors in the development of panic disorder (Beck and Emery, 1985; Clark, 1986), for example, anxiety sensitivity (McNally, 2002). Anxiety sensitivity is referred to as the fear of benign bodily sensations, which is caused by anxiety – also referred to as fear of fear (Goldstein and Chambless, 1978; Reiss et al., 1986) and has long been considered a specific vulnerability of panic disorder (White et al., 2006). Anxiety sensitivity is mostly conceptualized as a trait that remains relatively stable over time; although there are more studies showing that anxiety sensitivity decreases following treatment (Naragon-Gainey, 2010; Smits et al., 2008).

Currently, it remains unclear why some individuals with panic attacks start avoiding situations while others do not (Emmelkamp and Powers, 2010). Most individuals with agoraphobia experience some panic attacks or show some signs of anxiety before the onset of panic disorder and/or agoraphobia (Fava et al., 1992). The initial onset of agoraphobia is typically before the age of 35 years. There is a latent incidence risk in late adolescence and early adulthood. The mean age for onset of agoraphobia without preceding panic attacks is between 25 and 29 years (Nocon et al., 2008; Wittchen et al., 2010). Usually the course of agoraphobia is chronic and if not treated the remission rate is low (10%) (Emmelkamp and Wittchen, 2008). With increasing severity of agoraphobic symptoms, chronicity

increases and full remission rates decrease. In the long-term, a range of comorbidity may complicate the course of agoraphobia; depressive disorders, substance use disorder and traits of personality disorders. The clinical feature of agoraphobia is relatively consistent across the life span, although situations, which trigger fear and avoidance as well as the type of cognitions, may change. For example, for children being outside their home alone is the most frequently feared situation, whereas in adults standing in line in a shop and being in open spaces are most often feared (Wittchen et al., 2010). Common cognitions in children include becoming lost while in older adults cognitions focus more on falling.

Although panic severity does not predict the development of agoraphobia, the patients' exaggerated likelihood estimation of panic occurrence in situations is the best predictor of agoraphobic avoidance accounting for almost 80% of variance in avoidance behavior (Rachman, 1994; Telch et al., 1989). In a recent trial, it was shown that initial changes during therapy in agoraphobic cognitions predict later changes in avoidance behavior (Meyerbroeker et al., 2013).

Twin studies show 40% heritability with contributions from common familial environmental effects (<10%) and unique environmental effects (>50%) (Roth et al., 1984). It has further been suggested that an anxious temperament, which is characterized by neuroticism functions as an important risk factor (Watanabe et al., 2005).

Life Events and Stress

There is increasing evidence that life events often precede the onset of panic disorder. In a number of studies, it was found that up to two-thirds of patients with panic disorder experienced a significant personal loss in the year preceding the onset of panic disorder (Faravelli and Pallanti, 1989; Milrod et al., 2004; Pollard et al., 1989; Scocco et al., 2007). Most individuals report stressors in the months before their first panic attack, which can be interpersonal stressors, stressors related to physical well-being, or work-related stressors. Another important aspect is that childhood experiences of sexual, physical, and/or emotional abuse. These experiences are commonly reported in individuals with panic disorder (Roy-Byrne et al., 2006; Schade et al., 2004).

Comorbidity

High comorbidity rates have been found in individuals with anxiety disorders for substance abuse, especially alcohol abuse (Grant et al., 2004). In patients with an alcohol-related substance use disorder, an overall rate of more than 42% was found to have a comorbid anxiety disorder (Schneider et al., 2001). Most prevalent were social phobia (14%), agoraphobia (13%), and panic disorder (5%). The relationship between substance abuse and panic disorder and agoraphobia seems to be reciprocal (Emmelkamp and Powers, 2010). Individuals are often convinced that the use of alcohol reduces their anxiety and substance abuse might be attributed to anxiety as a form of self-medication (Zimmerman and Chelminski, 2003). On the other hand, symptoms of panic and anxiety are common symptoms during withdrawal and should be distinguished from primary anxiety disorders (Emmelkamp and Vedel, 2006).

Medical Conditions

Panic disorder is commonly associated with numerous medical conditions and symptoms such as dizziness, cardiac arrhythmias, apnea, hyperthyroidism, asthma, COPD, and irritable bowl syndrome (APA, 2013). However, it remains unclear, what the exact nature of this relation is (cause or effect) and it might be that some physical symptoms just overlap and are due to the medical condition.

Theoretical Background

Depending on theoretical background, different models have been developed to explain the development and maintenance of panic disorder and agoraphobia. Although, different theoretical background often lead to different treatment models. In panic disorder and agoraphobia in general the treatment consists of the same basic components.

Cognitive Model

As in most anxiety disorders, the appraisal of threat is central to the cognitive theory of panic disorder and agoraphobia (Beck and Emery, 1985; Salkovskis et al., 2007). One possible explanation is that panic occurs because individuals are appraising normal somatic symptoms of anxiety as excessively threatening (Clark, 1986). According to the cognitive model of panic disorder, panic can be classified as an anxiety, which focuses on somatic sensations, which are usually increased while experiencing anxiety (Salkovskis, 2007). According to this view, the cognitive hypothesis of panic is based on the premise that individuals who are experiencing recurrent panic attacks do so because they have a tendency to misinterpret bodily sensations as dangerous or life threatening. Thus, in long-term individuals fear their fear.

Behavioral Model

In panic disorder, the pairing of intense fear and experiencing bodily sensations at the same time is central to develop the disorder in the classical conditioning paradigm. According to a classical conditioning model, the occurrence of one of the symptoms (e.g., palpitations) leads 'automatically' to the other (anxiety: 'I might have a heart attack') thus provoking a full panic attack. According to the conditioning paradigm, the same holds true for the development of panic disorder and agoraphobia. In the comorbid condition, panic attacks are correlated with certain situations, which in the long run lead to more paniclike symptoms leading to a situational-bound panic attack. The maintenance of anxiety disorders in general can be explained by operant conditioning when leaving the feared stimulus leads to reward (Emmelkamp et al., 1995).

Different theories have been discussed to explain the major effects of exposure therapy in anxiety disorders (Telch et al., 2014). It has been argued that exposure can be explained by means of habituation (Marks, 1978). According to Wolpe (1958), the effects of exposure therapy can be explained by means of reciprocal conditioning: the central idea is that fear could be eliminated by pairing a patient's fear response with

a physiological response, which was incompatible with the fear response (e.g., relaxation), and which would lead to inhibition of the original fear response. In experiments all over the world, it was shown that relaxation was not a necessary component to achieve meaningful fear reduction (Emmelkamp, 1982). In more recent views, treatment success is explained in terms of a learning paradigm (Bouton, 2000; Craske et al., 2008).

Biological Model

As in most other psychiatric disorders, multiple gene interactions are central to the biological models (Craske and Barlow, 2008). It is suggested that relatively broad or nonspecific genetic factors influence the vulnerability to panic and anxiety. This gene interaction probably explains why panic disorder and agoraphobia tend to run in families. For a recent review, concerning the debate about gene–environment interaction in panic disorder see Battaglia (2013).

Mechanisms of Change

Panic Appraisal

There seems to be preliminary evidence for cognitive mediation during cognitive behavioral therapy (CBT) in panic disorder (Hofmann et al., 2007). In a randomized controlled trial, 91 subjects were allocated to receive CBT alone, imipramine alone, CBT plus imipramine, or CBT plus placebo. Individuals who received one of these interventions were assessed before and after acute treatment, and after a 6-month maintenance period. It was shown that changes in panic-related cognitions-mediated changes in panic severity only in treatments that included CBT. This is in line with the findings of Fernandez-Arias et al. (2013) who argued not to combine psychological treatment with pharmacological treatment. First results are promising, but more research is needed to investigate the assumed mechanisms of change in exposure-based treatments in panic disorder. Panic appraisal has been found to play an important role in the development and maintenance of panic disorder and agoraphobia, but more research is needed to replicate and further extend these findings and investigate whether these findings are specific for panic disorder and agoraphobia or whether anxiety appraisal might be a more general factor across anxiety disorders.

Anxiety Sensitivity

From a cognitive perspective, anxiety sensitivity can be considered a significant factor in the development and maintenance of panic disorder (Barlow, 2002). Several other studies suggest that anxiety sensitivity mediates improvement in the treatment of panic disorder (Cho et al., 2007; Hofmann et al., 2007; Smits et al., 2004). For example, one study suggested that changes in anxiety sensitivity fully mediated improvement in patients with panic disorder after CBT when compared to a wait list control condition and partially mediated improvement in patients with agoraphobia (Smits et al., 2004). These results were stable at 6-months follow-up (Cho et al., 2007). In another study, it was shown that reductions in anxiety sensitivity preceded reductions in panic symptoms (Smits et al., 2008), thus meeting the temporal precedence criteria as a mediator of change in panic disorder. Anxiety sensitivity is associated with most anxiety disorders, but a particularly strong relationship between panic attacks and anxiety sensitivity has been demonstrated (Bentley et al., 2013). Anxiety sensitivity has also been shown to predict spontaneous panic attacks (Maller and Reiss, 1992; Plehn and Peterson, 2002).

In a large recent treatment study with 361 individuals with panic disorder, effect sizes showed that the greatest changes in anxiety sensitivity occurred at early treatment stages (Gallagher et al., 2013). Additionally, it was shown that anxiety sensitivity predicted changes in panic disorder symptoms, thus indicating that anxiety sensitivity might be a mechanism of change, which is particularly important during the early stages of treatment.

In terms of transdiagnostic approaches, anxiety sensitivity might be a marker across anxiety disorders. There is more research needed to investigate whether the above-mentioned findings can be replicated in other anxiety disorders. These findings are interesting in times of unified protocols and more dimensional perspectives of mental disorders. The above-mentioned findings are important for understanding the mechanisms of change in specific anxiety disorders, but research has also shown that anxiety sensitivity plays an important role across other anxiety disorders (Vanden Bogaerde and De Raedt, 2011).

Evidence-Based Treatment Approaches

Guidelines

According to the NICE guidelines (National Institute for Health and Clinical Excellence, 2011), there is a consensus that empirically supported psychological treatments are the treatment of choice for panic disorder and agoraphobia. This can be combined with pharmacological treatments, which can produce better results in reducing relapse rates and improving maintenance of gains (Barlow et al., 2000; Otto et al., 2005), although this has been questioned recently (Fernandez-Arias et al., 2013; Hofmann et al., 2007). Please see also the section about pharmacological treatment and the paragraph about mechanisms of change.

Although different theoretical background often leads to different treatment models, in panic disorder and agoraphobia in general the treatment consists of the same components: (1) psychoeducation about the nature and causes of panic disorder and agoraphobia, (2) cognitive restructuring, which targets the negative cognitions associated with panic, and (3) exposure to interoceptive bodily sensations and agoraphobic situations. Sometimes, this program is completed with breathing skills training and relaxation but in clinical practice these have been largely abandoned given findings that they do not enhance outcome (Schmidt et al., 2000).

Dismantling studies of the various components of CBT for panic disorder and agoraphobia have had inconsistent findings (Craske and Barlow, 2008). On one hand, it was found that cognitive therapy as a component may be effective (Williams and Falbo, 1996) even when it is conducted without exposure and other behavioral components (Salkovskis et al., 1991). On the other hand, it was found that cognitive therapy does not have an additional value above exposure in

patients with agoraphobia (Craske et al., 1997; Ost et al., 2004). In more recent research, it was found that exposure-based interventions are usually superior to cognitive interventions in agoraphobia (Moscovitch, 2009; Sanchez-Meca et al., 2010).

Psychoeducation

The first phase of treatment for panic disorder and/or agoraphobia usually includes psychoeducation about the nature and causes of panic disorder and agoraphobia. This information usually consists of making a more detailed and clear distinction between anxiety and panic. Further, the relationship between experiencing anxiety, panic-related bodily symptoms, and cognitions is explained (e.g., the vicious circle of panic, (Craske and Barlow, 2008)). Detailed analyses of recent panic attacks often help the patient to become aware of these relationships. If appropriate, the relationship with agoraphobic avoidance behavior is further explored in collaboration with the patient. To obtain objective information in treatment sessions, it is important to let the patient record their panic attacks and the correlating symptoms, feelings, and thoughts.

Exposure *in vivo*

Marks (1978) has argued that exposure therapy works through emotional habituation. Habituation is defined as a decrease in fear reactions in response to repeated exposure to the feared stimulus. In panic disorder and agoraphobia, exposure targets fear of bodily sensations (interoceptive exposure) and agoraphobic situations (Emmelkamp and Powers, 2010; Meuret et al., 2012).

Meta-analyses and reviews of exposure for panic disorder and agoraphobia (Mitte, 2005; Sanchez-Meca et al., 2010) show that exposure *in vivo* has a substantial effect size for reducing agoraphobic symptoms. Additionally, exposure-based interventions (including virtual reality exposure, Meyerbröker, 2014) show strong effect sizes in reducing agoraphobic avoidance behavior (Meyerbroeker et al., 2013; Moscovitch, 2009). Interestingly, therapist guidance during exposures appears important (Gloster et al., 2011). This large Randomized Control Trial showed that guided exposure *in vivo* was more effective than exposure without therapist guidance in reducing the number of panic attacks, avoidance behavior, and improvement in global functioning.

Interoceptive Exposure

Because individuals with panic-related symptoms usually misinterpret their bodily sensations and therefore panic even worse, one important aspect in the treatment of panic disorder is interoceptive exposure. The aim of the exposure is twofold: first, one can tolerate uncomfortable bodily sensations and second that these symptoms will decrease over time without any intervention. Thus, the patient will learn to tolerate the symptoms without misinterpreting the symptoms as resulting in the feared consequences. Additionally, the patient can compare the provoked paniclike symptoms and compare them to the recently, nonprovoked symptoms, and panic attacks.

Cognitive Restructuring

Cognitive restructuring is often introduced in combination with the recording of the panic attacks over the early therapy sessions. The purpose of cognitive restructuring is to investigate together with the patient what thoughts and possible misinterpretations they experience along with panic attacks. Often, these thoughts or misinterpretations come along with negative emotions, which in turn can strengthen the symptoms (e.g., vicious circle of panic). By having a closer look together with the therapist at these thoughts and emotions, the patient might become more aware of certain thoughts. These thoughts – automatic thoughts – are important in panic disorder because they often generate negative behavior due to the misinterpretation of the symptoms. When patients have identified these automatic thoughts, they can be challenged with more rational alternatives.

Relaxation and Breathing Training

Often, relaxation and/or breathing training is added to the treatment of panic disorder. The purpose is to help the patient to relax more easily when a panic attack is occurring. However, research shows that the above-mentioned treatment components are usually more effective than relaxation training (Barlow et al., 1989) and results are maintained at 2-year follow-up (Craske et al., 1991). The breathing training is often offered but not actively practiced in therapy sessions.

Pharmacological Treatment

Generally, selective serotonin reuptake inhibitors are the medication of choice for panic disorder and agoraphobia as indicated by placebo-controlled randomized clinical trials (Roy-Byrne and Cowley, 2002). In meta-analyses, medium to large effect sizes are reported when compared to Placebo (Mitte, 2005; Bakker et al., 2002). The use of benzodiazepines is not recommended because of their physiological dependence and withdrawal (Roy-Byrne and Cowley, 2002) and because of only temporary symptom remedy (for a review about the use of benzodiazepines in combination with psychotherapy see Watanabe et al., 2007).

Recently, it has been questioned whether the combination of pharmacological treatment with psychological treatment has an additional value (Fernandez-Arias et al., 2013). It was found that in the long run, patients profited more from empirically supported psychological treatment than from psychological treatment in combination with pharmacological therapy.

Future Directions

Although we discussed some findings on potential mediators and moderators of treatment outcome, future studies should increase our confidence in more specific targets. Further, new analysis strategies offer more powerful methods to test such mediation targets (Smits et al., 2012). For example, Kazdin (2007) outlined seven conditions to establish mediation including (1) the selection of mediators must be guided by theory; (2) treatment studies must include measures of potential mediators; (3) the timeline of the proposed mediator

and outcome must be established; (4) studies must assess more than one mediator; (5) studies must use designs that can evaluate mediators; (6) different types of studies must provide converging evidence; and (7) treatment studies must be complemented by experiments that manipulate the mediator to provide converging evidence. As more researchers use these more stringent criteria, our understanding of mediators and moderators should improve.

Future research will also likely include transdiagnostic models of psychology focusing on cognitive endophenotypes of psychopathology from an integrative cognitive psychology perspective (Emmelkamp et al., 2014). For example, emotion dysregulation (Hofmann et al., 2012) and rumination or repetitive thinking (Ehring and Watkins, 2008) are implicated across various mood and anxiety disorders. Such findings have led to transdiagnostic manualized protocols. Initial results are promising. Two open clinical trials of a transdiagnostic treatment suggest that this approach can distil common strategies and enhance the effectiveness by targeting core affective 'higher order' components, which results in substantial clinical improvement in primary and secondary disorders (Ellard et al., 2010). These findings were extended by Farchione and colleagues (Farchione et al., 2012), who treated 38 patients with a unified protocol consisting of 5 modules, which targeted key aspects of emotional processing and regulation of emotional experiences: (1) increasing present-focused emotion awareness, (2) increasing cognitive flexibility, (3) identifying and preventing patterns of emotion avoidance, (4) increasing awareness and tolerance of emotion-related physical sensations, and (5) interoceptive and situation-based emotion focused exposure (Farchione et al., 2012).

Clinical Implications

It is important to translate research into clinically meaningful working protocols. As mentioned above, recent approaches to more transdiagnostic models of psychopathology are a promising first step into clinically meaningful results. If, for example, multiple diagnosis and specific treatment protocols can be combined, this can result in more fruitful and effective treatment planning. If this approach proves to be successful in emotional disorders, it might promote dissemination of evidence-based treatments. Another important aspect is the need for more effective studies wherein not all comorbidities are excluded. This makes studies more representative for clinical practice and might help to clarify why some treatments work for some individuals but not for others.

See also: Behavior Therapy: Background, Basic Principles, and Early History; Classical Conditioning Methods in Psychotherapy; Concepts and Methods of Cognitive Therapies; Epidemiology, Concepts and Methods of; Exposure Therapies and Stress Inoculation: A Brief Overview; Manual Based Treatments.

Bibliography

American Psychiatric Association (APA), 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5), fifth ed. American Psychiatric Association (APA), Washington, DC.

American Psychiatric Association (APA), 1980. Diagnostic and Statistical Manual of Mental Disorders (DSM-III), third ed. American Psychiatric Association (APA), Washington, DC.

Bakker, A., Van Balkom, A., Spinhoven, P., 2002. SSRIS vs. TCAs in the treatment of panic disorder: a meta-analysis. Acta Psychiatrica Scandinavica 106 (3), 163–167.

Barlow, D.H., 2002. Anxiety and Its Disorders: The Nature and Treatment of Anxiety and Panic, second ed. Guilford Press, New York.

Barlow, D., Craske, M., Cerny, J., Klosko, J., 1989. Behavioral treatment of panic disorder. Behavior Therapy 20 (2), 261–282.

Barlow, D., Gorman, J., Shear, M., Woods, S., 2000. Cognitive-behavioral therapy, imipramine, or their combination for panic disorder (vol. 283, pp. 2529, 2000). JAMA-Journal of the American Medical Association 284 (19), 2450.

Battaglia, M., 2013. Gene-environment interaction in panic disorder and posttraumatic stress disorder. Canadian Journal of Psychiatry-Revue Canadienne De Psychiatrie 58 (2), 69–75.

Beck, A.T., Emery, G., 1985. Anxiety Disorders and Phobias: A Cognitive Perspective. Basic Books, New York.

Beesdo, K., Knappe, S., Pine, D.S., 2009. Anxiety and anxiety disorders in children and adolescents: developmental issues and implications for DSM-V. Psychiatric Clinics of North America 32 (3), 483.

Bentley, K.H., Gallagher, M.W., Boswell, J.F., Gorman, J.M., Shear, M.K., Woods, S.W., Barlow, D.H., 2013. The interactive contributions of perceived control and anxiety sensitivity in panic disorder: a triple vulnerabilities perspective. Journal of Psychopathology and Behavioral Assessment 35 (1), 57–64.

Bittner, A., Egger, H.L., Erkanli, A., Costello, E.J., Foley, D.L., Angold, A., 2007. What do childhood anxiety disorders predict? Journal of Child Psychology and Psychiatry 48 (12), 1174–1183.

Bouton, M., 2000. A learning theory perspective on lapse, relapse, and the maintenance of behavior change. Health Psychology 19 (1), 57–63.

Cho, Y., Smits, J.A.J., Powers, M.B., Telch, M.J., 2007. Do changes in panic appraisal predict improvement in clinical status following cognitive-behavioral treatment of panic disorder? Cognitive Therapy and Research 31 (5), 695–707.

Clark, D.M., 1986. A cognitive approach to panic. Behaviour Research and Therapy 24 (4), 461–470.

Craske, M.G., Barlow, D.H., 2008. Panic disorder and agoraphobia. In: Barlow, D.H. (Ed.), Clinical Handbook of Psychological Disorders. Guilford, New York, pp. 1–64.

Craske, M., Brown, T., Barlow, D., 1991. Behavioral treatment of panic disorder – a 2-year follow-up. Behavior Therapy 22 (3), 289–304.

Craske, M., Rowe, M., Lewin, M., Noriegadimitri, R., 1997. Interoceptive exposure versus breathing retraining within cognitive-behavioural therapy for panic disorder with agoraphobia. British Journal of Clinical Psychology 36, 85–99.

Craske, M.G., Kircanski, K., Epstein, A., Wittchen, H., Pine, D.S., Lewis-Fernandez, R., Hinton, D., DSM V Anxiety OC Spectrum Posttrau, 2010. Panic disorder: a review of DSM-IV panic disorder and proposals for DSM-V. Depression and anxiety 27 (2), 93–112.

Craske, M.G., Kircanski, K., Zelikowsky, M., Mystkowski, J., Chowdhury, N., Baker, A., 2008. Optimizing inhibitory learning during exposure therapy. Behaviour research and therapy 46 (1), 5–27.

De Graaf, M.R., Ten Have, M., Van Gool, C., Van Dorsselaer, S., 2012. Prevalentie van psychische aandoeningen en trends van 1996 tot 2009; resultaten van NEMESIS-2. Tijdschrift voor Psychiatrie 54, 27–38.

Ehring, T., Watkins, E.R., 2008. Repetitive negative thinking as a transdiagnostic process. International Journal of Cognitive Therapy 1 (3), 192–205.

Ellard, K.K., Fairholme, C.P., Boisseau, C.L., Farchione, T.J., Barlow, D.H., 2010. Unified protocol for the transdiagnostic treatment of emotional disorders: protocol development and initial outcome data. Cognitive and Behavioral Practice 17 (1), 88–101.

Emmelkamp, P.M.G., 1982. Anxiety and fear. In: Bellack, A., Hersen, M., Kazdin, A. (Eds.), International Handbook of Behavior Modification and Therapy. Plenum Press, New York, pp. 349–395.

Emmelkamp, P.M.G., Bouman, T.K., Scholing, H.A., 1995. Angst, Fobieën en Dwang: Diagnostiek en behandeling. Van Loghum Slaterus, Deventer.

Emmelkamp, P.M.G., David, D., Beckers, T., Muris, P., Cuijpers, P., Lutz, W., Andersson, G., Araya, R., Banos Rivera, R.M., Barkham, M., Berking, M., Berger, T., Botella, C., Carlbring, P., Colom, F., Essau, C., Hermans, D., Hofmann, S.G., Knappe, S., Ollendick, T.H., Raes, F., Rief, W., Riper, H., Van Der Oord, S., Vervliet, B., 2014. Advancing psychotherapy and evidence-based psychological interventions. International Journal for Methods in Psychiatric Research 23 (S1), 58–91. http://dx.doi.org/10.1002/mpr.1411.

Emmelkamp, P.M.G., Powers, M.B., 2010. Agoraphobia. In: Thomas, J.C., Hersen, M. (Eds.), Handbook of Clinical Psychology Competencies. Springer, New York, pp. 723–758.

Emmelkamp, P.M.G., Powers, M.B., 2009. Neo-Kraepelinian diagnosis: adequacy for phobias and panic. In: Mckay, D., Abramowitz, J.S., Taylor, S., Admundson, G. (Eds.), Current Perspectives on the Anxiety Disorders: Implications for DSM-V and Beyond. Springer Press, New York, pp. 41–76.

Emmelkamp, P.M.G., Vedel, E., 2006. Evidence-Based Treatment in Alcohol and Drug Abuse: A Practitioners Guide to Concepts, Methods and Research. Brunner-Routledge/Taylor & Francis, New York.

Emmelkamp, P.M.G., Wittchen, H.U., 2008. Specific phobias. In: Andrews, G., Regier, D.A., Charney, D., Sirovatka, P.J. (Eds.), Stress-Induced and Fear Circuitry Disorders: Defining the Research Agenda for DSM V. American Psychiatric Publishing, Arlington, USA, pp. 77–104.

Faravelli, C., Cosci, F., Rotella, F., Faravelli, L., Catena Dell'osso, M., 2008. Agoraphobia between panic and phobias: clinical epidemiology from the Sesto Fiorentino Study. Comprehensive Psychiatry 49 (3), 283–287.

Faravelli, C., Pallanti, S., 1989. Recent life events and panic disorder. American Journal of Psychiatry 146 (5), 622–626.

Farchione, T.J., Fairholme, C.P., Ellard, K.K., Boisseau, C.L., Thompson-Hollands, J., Carl, J.R., Gallagher, M.W., Barlow, D.H., 2012. Unified protocol for transdiagnostic treatment of emotional disorders: a randomized controlled trial. Behavior Therapy 43 (3), 666–678.

Fava, L., Morton, J., 2009. Causal modeling of panic disorder theories. Clinical Psychology Review 29 (7), 623–637.

Fava, G., Grandi, S., Rafanelli, C., Canestrari, R., 1992. Prodromal symptoms in panic disorder with agoraphobia – a replication study. Journal of Affective Disorders 26 (2), 85–88.

Fernandez-Arias, I., Labrador, F.J., Estupina, F.J., Bernaldo De Quiros, M., Alonso, P., Blanco, C., Gomez, L., 2013. Does adding pharmaceutical medication contribute to empirically supported psychological treatment for anxiety disorders? Psicothema 25 (3), 313–318.

Frances, A., Miele, G., Widiger, T., Pincus, H., Manning, D., Davis, W., 1993. The classification of panic disorders – from Freud to DSM-IV. Journal of Psychiatric Research 27, 3–10.

Gallagher, M.W., Payne, L.A., White, K.S., Shear, K.M., Woods, S.W., Gorman, J.M., Barlow, D.H., 2013. Mechanisms of change in cognitive behavioral therapy for panic disorder: the unique effects of self-efficacy and anxiety sensitivity. Behaviour Research and Therapy 51, 767–777.

Gloster, A.T., Wittchen, H., Einsle, F., Lang, T., Helbig-Lang, S., Fydrich, T., Fehm, L., Hamm, A.O., Richter, J., Alpers, G.W., Gerlach, A.L., Stroehle, A., Kircher, T., Deckert, J., Zwanzger, P., Hoefler, M., Arolt, V., 2011. Psychological treatment for panic disorder with agoraphobia: a randomized controlled trial to examine the role of therapist-guided exposure in situ in CBT. Journal of Consulting and Clinical Psychology 79 (3), 406–420.

Goldstein, A.J., Chambless, D.L., 1978. A reanalysis of agoraphobia. Behavior Therapy 9 (1), 47–59.

Goodwin, R., Faravelli, C., Rosi, S., Cosci, F., Truglia, E., De Graaf, R., Wittchen, H., 2005. The epidemiology of panic disorder and agoraphobia in Europe. European Neuropsychopharmacology 15 (4), 435–443.

Grant, B., Hasin, D., Stinson, F., Dawson, D., Chou, S., Ruan, W., Pickering, R., 2004. Prevalence, correlates, and disability of personality disorders in the United States: results from the National Epidemiologic Survey on alcohol and related conditions. Journal of Clinical Psychiatry 65 (7), 948–958.

Hofmann, S.G., Meuret, A.E., Rosenfield, D., Suvak, M.K., Barlow, D.H., Gorman, J.M., Shear, M.K., Woods, S.W., 2007. Preliminary evidence for cognitive mediation during cognitive-behavioral therapy of panic disorder. Journal of Consulting and Clinical Psychology 75 (3), 374–379.

Hofmann, S.G., Sawyer, A.T., Fang, A., Asnaani, A., 2012. Emotion dysregulation model of mood and anxiety disorders. Depression and Anxiety 29 (5), 409–416.

Kazdin, A.E., 2007. Mediators and mechanisms of change in psychotherapy research. Annual Review of Clinical Psychology 3, 1–27.

Kessler, R.C., Chiu, W.T., Demler, O., Walters, E.E., 2005. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey replication. Archives of General Psychiatry 62 (6), 617–627.

Kessler, R.C., Petukhova, M., Sampson, N.A., Zaslavsky, A.M., Wittchen, H., 2012. Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. International Journal of Methods in Psychiatric Research 21 (3), 169–184.

Maller, R., Reiss, S., 1992. Anxiety sensitivity in 1984 and panic attacks in 1987. Journal of Anxiety Disorders 6 (3), 241–247.

Marks, I.M., 1978. Behavioral psychotherapy of adult neurosis. In: Garfield, S.L., Bergin, A.E. (Eds.), Handbook of Psychotherapy and Behavior Change. John Wiley, New York, pp. 493–589.

Marks, I.M., 1970. The classification of phobic disorders. British Journal of Psychiatry 116 (533), 377–386.

McNally, R.J., 1994. Panic Disorder: A Critical Analysis. Guilford Press, New York.

McNally, R., 2002. Anxiety sensitivity and panic disorder. Biological Psychiatry 52 (10), 938–946.

Meuret, A.E., Wolitzky-Taylor, K.B., Twohig, M.P., Craske, M.G., 2012. Coping skills and exposure therapy in panic disorder and agoraphobia: latest advances and future directions. Behavior Therapy 43 (2), 271–284.

Meyerbroeker, K., Morina, N., Kerkhof, G.A., Emmelkamp, P.M.G., 2013. Virtual reality exposure therapy does not provide any additional value in agoraphobic patients: a randomized controlled trial. Psychotherapy and Psychosomatics 82 (3), 170–176.

Meyerbröker, K., 2014. Virtual reality exposure. In: Emmelkamp, P.M.G., Ehring, T. (Eds.), The Wiley Handbook of Anxiety Disorders, Blackwell-Wiley, vol. II. Oxford, pp. 1310–1324.

Milrod, B., Leon, A., Shear, M., 2004. Can interpersonal loss precipitate panic disorder? American Journal of Psychiatry 161 (4), 758–759.

Mitte, K., 2005. A meta-analysis of the efficacy of psycho- and pharmacotherapy in panic disorder with and without agoraphobia. Journal of Affective Disorders 88 (1), 27–45.

Moscovitch, D.A., 2009. What is the core fear in social phobia? A new model to facilitate individualized case conceptualization and treatment. Cognitive and Behavioral Practice 16 (2), 123–134.

Naragon-Gainey, K., 2010. Meta-analysis of the relations of anxiety sensitivity to the depressive and anxiety disorders. Psychological Bulletin 136 (1), 128–150.

National Institute for Health and Clinical Excellence, 2011. Generalised anxiety disorder and panic disorder (with or without agoraphobia) in adults: management in primary, secondary and community care (Clinical guideline). http://www.nice.org.uk/guidance/cg113.

Nocon, A., Wittchen, H., Beesdo, K., Brueckl, T., Hofler, M., Pfister, H., Zimmermann, P., Lieb, R., 2008. Differential familial liability of panic disorder and agoraphobia. Depression and Anxiety 25 (5), 422–434.

Ost, L., Thulin, U., Ramnero, J., 2004. Cognitive behavior therapy vs exposure in vivo in the treatment of panic disorder with agrophobia. Behaviour Research and Therapy 42 (10), 1105–1127.

Otto, M., Powers, M., Smits, J., 2005. Adding cognitive-behavioral therapy to pharmacotherapy for panic disorder: issues and strategies. CNS Spectrums 10 (9), 32–39.

Plehn, K., Peterson, R., 2002. Anxiety sensitivity as a predictor of the development of panic, symptoms, panic attacks, and panic disorder: a prospective study. Journal of Anxiety Disorders 16 (4), 455–474.

Pollard, C., Pollard, H., Corn, K., 1989. Panic onset and major events in the lives of agoraphobics – a test of contiguity. Journal of Abnormal Psychology 98 (3), 318–321.

Rachman, S., 1994. The overprediction of fear – a review. Behaviour Research and Therapy 32 (7), 683–690.

Reiss, S., Peterson, R., Gursky, D., McNally, R., 1986. Anxiety sensitivity, anxiety frequency and the prediction of fearfulness. Behaviour Research and Therapy 24 (1), 1–8.

Roth, W., Telch, M., Taylor, C., Sachitano, J., Gallen, C., Pfefferbaum, A., Agras, W., Kopell, B., 1984. Autonomic activation in agoraphobia with panic attacks. Psychophysiology 21 (5), 596.

Roy-Byrne, P.P., Cowley, D.S., 2002. Pharmacological treatments for panic disorder, generalized anxiety disorder, specific phobia, and social anxiety disorder. In: Nathan, P.E., Gorman, J.M. (Eds.), A Guide to Treatments That Work, second ed. Oxford University Press, New York, pp. 337–365.

Roy-Byrne, P.P., Craske, M.G., Stein, M.B., 2006. Panic disorder. Lancet 368 (9540), 1023–1032.

Salkovskis, P.M., 2007. Cognitive–behavioural treatment for panic. Psychiatry 6 (5), 193–196.

Salkovskis, P.M., Hackmann, A., Wells, A., Gelder, M.G., Clark, D.M., 2007. Belief disconfirmation versus habituation approaches to situational exposure in panic disorder with agoraphobia: a pilot study. Behaviour Research and Therapy 45 (5), 877–885.

Salkovskis, P., Clark, D., Hackmann, A., 1991. Treatment of panic attacks using cognitive therapy without exposure or breathing retraining. Behaviour Research and Therapy 29 (2), 161–166.

Sanchez-Meca, J., Rosa-Alcazar, A.I., Marin-Martinez, F., Gomez-Conesa, A., 2010. Psychological treatment of panic disorder with or without agoraphobia: a meta-analysis. Clinical Psychology Review 30 (1), 37–50.

Schade, A., Marquenie, L., Van Balkom, A., Koeter, M., De Beurs, E., Van Den Brink, W., Van Dyck, R., 2004. Alcohol-dependent patients with comorbid phobic disorders: a comparison between comorbid patients, pure alcohol-dependent and pure phobic patients. Alcohol and Alcoholism 39 (3), 241–246.

Schmidt, N.B., Korte, K.J., Norr, A.M., 2014. Panic disorder and agoraphobia. In: Emmelkamp, P.M.G., Ehring, T. (Eds.), The Wiley Handbook of Anxiety Disorders, Blackwell-Wiley, vol. I. Oxford, pp. 321–356.

Schmidt, N.B., Norr, A.M., Korte, K.J., 2013. Panic disorder and agoraphobia: considerations for DSM-V. Research on Social Work and Practice. online before print.

Schmidt, N., Woolaway-Bickel, K., Trakowski, J., Santiago, H., Storey, J., Koselka, M., Cook, J., 2000. Dismantling cognitive-behavioral treatment for panic disorder: questioning the utility of breathing retraining. Journal of Consulting and Clinical Psychology 68 (3), 417–424.

Schneider, U., Altmann, A., Baumann, M., Bernzen, J., Bertz, B., Bimber, U., Broese, T., Broocks, A., Burtscheidt, W., Cimander, K., Degkwitz, P., Driessen, M., Ehrenreich, H., Fischbach, E., Folkerts, H., Frank, H., Gurth, D., Havemann-Reinecke, U., Heber, W., Heuer, J., Hingsammer, A., Jacobs, S., Krampe, H., Lange, W., Lay, T., Leimbach, M., Lemke, M., Leweke, M., Mangholz, A., Massing, W., Meyenberg, R., Porzig, J., Quattert, T., Redner, C., Ritzel, G., Rollnik, J., Sauvageoll, R., Schlafke, D., Schmid, G., Schroder, H., Schwichtenberg, U., Schwoon, D., Seifert, J., Sickelmann, I., Sieveking, C., Spiess, C., Stiegemann, H., Stracke, R., Straetgen, H., Subkowski, P., Thomasius, R., Tretzel, H., Verner, L., Vitens, J., Wagner, T., Weirich, S., Weiss, I., Wendorff, T., Wetterling, T., Wiese, B., Wittfoot, J., 2001. Comorbid anxiety and affective disorder in alcohol-dependent patients seeking treatment: the first multicentre study in Germany. Alcohol and Alcoholism 36 (3), 219–223.

Scocco, P., Barbieri, I., Frank, E., 2007. Interpersonal problem areas and onset of panic disorder. Psychopathology 40 (1), 8–13.

Smits, J.A.J., Julian, K., Rosenfield, D., Powers, M.B., 2012. Threat reappraisal as a mediator of symptom change in cognitive-behavioral treatment of anxiety disorders: a systematic review. Journal of Consulting & Clinical Psychology 80, 624–635.

Smits, J., Powers, M., Cho, Y., Telch, M., 2004. Mechanism of change in cognitive-behavioral treatment of panic disorder: evidence for the fear of fear mediational hypothesis. Journal of Consulting and Clinical Psychology 72 (4), 646–652.

Smits, J.A.J., Berry, A.C., Rosenfield, D., Powers, M.B., Behar, E., Otto, M.W., 2008. Reducing anxiety sensitivity with exercise. Depression and Anxiety 25 (8), 689–699.

Telch, M.J., Cobb, M.R., Lancaster, C.L., 2014. Exposure therapy for anxiety disorders: procedural variations, clinical efficacy, and change mechanisms. In: Emmelkamp, P.M.G., Ehring, T. (Eds.), The Wiley Handbook of Anxiety Disorders.

Telch, M., Brouillard, M., Telch, C., Agras, W., Taylor, C., 1989. Role of cognitive appraisal in panic-related avoidance. Behaviour Research and Therapy 27 (4), 373–383.

Vanden Bogaerde, A., De Raedt, R., 2011. The moderational role of anxiety sensitivity in flight phobia. Journal of Anxiety Disorders 25 (3), 422–426.

Watanabe, N.A., Churchill, R., Furukawa, T.A., 2007. Combination of psychotherapy and benzodiazepines versus either therapy alone for panic disorder: a systematic review. BMC Psychiatry 7 (18).

Watanabe, A., Nakao, K., Tokuyama, M., Takeda, M., 2005. Prediction of first episode of panic attack among white-collar workers. Psychiatry and Clinical Neurosciences 59 (2), 119–126.

White, K., Brown, T., Somers, T., Barlow, D., 2006. Avoidance behavior in panic disorder: the moderating influence of perceived control. Behaviour Research and Therapy 44 (1), 147–157.

Williams, S., Falbo, J., 1996. Cognitive and performance-based treatments for panic attacks in people with varying degrees of agoraphobic disability. Behaviour Research and Therapy 34 (3), 253–264.

Wittchen, H., Jacobi, F., Rehm, J., Gustavsson, A., Svensson, M., Jönsson, B., Steinhausen, H.C., 2011. The size and burden of mental disorders and other disorders of the brain in Europe 2010. European Neuropsychopharmacology 21 (9), 655–679.

Wittchen, H., Gloster, A.T., Beesdo-Baum, K., Fava, G.A., Craske, M.G., 2010. Agoraphobia: a review of the diagnostic classificatory position and criteria. Depression and Anxiety 27 (2), 113–133.

Wittchen, H., Nocon, A., Beesdo, K., Pine, D.S., Hoefler, M., Lieb, R., Gloster, A.T., 2008. Agoraphobia and panic. Psychotherapy and Psychosomatics 77 (3), 147–157.

Wittchen, H., Reed, V., Kessler, R., 1998. The relationship of agoraphobia and panic in a community sample of adolescents and young adults. Archives of General Psychiatry 55 (11), 1017–1024.

Wolpe, J., 1958. Psychotherapy by Reciprocal Inhibition. Stanford University Press, Standford, CA.

Zimmerman, M., Chelminski, I., 2003. Clinician recognition of anxiety disorders in depressed outpatients. Journal of Psychiatric Research 37 (4), 325–333.

Paralawyers: Other Legal Occupations

Herbert M Kritzer, University of Minnesota Law School, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although legal services are thought to be the domain of lawyers, those employed in many other occupations do law-related work, some incidental to other work and some, here called para-lawyers, as the sole focus of what they do. Para-lawyers have long existed, but they are becoming more visible as providers of legal services with the development of systems for licensing and regulating para-lawyers. Research shows that para-lawyers can be effective providers of a range of legal services, and that under some circumstances they can be more effective than fully qualified lawyers. Understanding the phenomenon of para-lawyers raises issues related to closure, deprofessionalization, postprofessionalism, and gender.

Definition

Para-lawyers are persons who perform 'law work' although they are not themselves licensed as lawyers. 'Law work' refers to those activities requiring at least some of the specialized legal knowledge and skill of the type associated with the work of lawyers. Thus, para-lawyers are 'law workers' who lack the full credentials and/or status associated with the legal profession but who provide services that draw on expertise in law and/or legal processes. This negative definition reflects the complexity of defining the practice of law and what constitutes membership in a legal *profession*.

Drawing the line between lawyers and para-lawyers varies from country to country because of variations in the nature of legal tasks within different legal systems and how those tasks are assigned to particular occupations. In many countries, the exact line is not clear. In some countries, the term most equivalent to 'lawyer' is customarily applied to an occupation engaged in a very narrow range of activities, while there are other occupations requiring substantial formal legal training that perform work that elsewhere would be the exclusive domain of the legal profession. Deciding whether to include such occupations under the label of 'lawyers' or 'para-lawyers' requires drawing arbitrary lines. One example is the legal occupation of notary as it exists in many civil law countries (Malavet, 1996). Historically, the occupation of notary arose to provide a neutral person who could be trusted to draw up fair contracts to be executed by one or more persons unable to read or write; today, notaries continue to play a role in the drawing up of and execution of contracts, and they apply specialized legal knowledge related to those activities. If the notary's tasks involve anything more than ministerial functions such as witnessing the execution of contracts, as is the role of the notary in the United States, it is arbitrary whether the notary is deemed to be part of the legal profession or is instead categorized as a para-lawyer, although some distinctions might be made based on the education and training requirements to become a notary.

In addition to problems of drawing a line between lawyers and para-lawyers, it is difficult to draw a line between para-lawyers and some 'nonlaw' occupations because many nonlaw occupations require the use of some very specific legal knowledge, typically applied in a very routine way. Examples include drawing up contracts for routine property transactions (in the United States typically done by real estate sales agents), drawing up routine loan contracts (done by bank officials), settling tort-based injury claims (done by nonlawyer representatives of insurance companies), or advising on estate or tax issues (done by financial advisors and accountants). In one sense, all of these occupations could be labeled para-lawyers; for some of the occupations the para-lawyer work is only a part of their activities (e.g., real estate sales agents) while for others (e.g., insurance claims representatives) it is the vast bulk of their work.

In some countries (e.g., Japan), rights of audience in court might serve as a clear demarcation between lawyers and nonlawyers. However, in other countries there are branches of the legal profession that have no rights of audience (e.g., notaries in civil law countries) and in others some para-lawyers have been granted rights of audience in court (e.g., England for legal executives and Ontario for licensed paralegals). In fact, in England it is now possible for 'chartered legal executives' to become District Judges (the lowest level of professional judges in England, once called 'stipendiary magistrates'). Note here the reference to 'rights of audience in courts'; in many countries nonlawyers may represent parties before administrative tribunals and undertake various types of work not involving litigation. In recent years, at least two American states (California and Washington) have established procedures for licensing nonlawyers to assist clients in preparing a range of legal documents. In many jurisdictions there is no licensing procedure, and a range of legal work can be done legally by anyone who chooses to offer the service.

Finally, there is no recognized standard of education or specialized training to become a para-lawyer. Although there are training courses for occupations variously labeled 'paralegal,' 'legal assistant,' or 'legal executive,' these courses do not provide formal qualifications, and most people enter such occupations without any formalized training (Johnstone and Wenglinsky, 1985: pp. 121–123). Other kinds of para-lawyer occupations involve the application of knowledge learned in some other role; examples include former police officers representing traffic offenders in Ontario (including those charged with serious offenses such as driving while intoxicated), former Internal Revenue Service (IRS) auditors representing or assisting taxpayers in disputes with the IRS in the United States, former insurance company claims inspectors or adjusters

representing injury claimants in England, and legal secretaries who take on increasing responsibilities to the point where they are engaged more in law work than in secretarial work. Thus there is no clear way to decide whether a person offering a particular set of services, or engaged in a particular set of tasks, should be classified as a para-lawyer; nor is the term 'para-lawyer' in common use in a way that would allow for self-identification.

Empirical and Theoretical Issues

Even with these definitional and boundary problems, the concept of para-lawyers, and para-professionals more generally, is useful theoretically and important for practical purposes. It raises the issues of how law work is and should be structured, and what qualifications are needed to be effective in law work. In addition, there are three other lines of theoretical inquiry that are important vis-à-vis para-lawyers: achieving profession or 'closure,' 'deprofessionalization,' and the gendering of occupations.

Structure of Law Work

Law work readily divides along at least three dimensions: institutional setting; legal advice vs legal assistance (preparing documents, advocacy before a neutral third-party, negotiating with another party); and routine vs nonroutine tasks. The degree to which para-lawyers have in the past, and are today, providing legal services varies along these dimensions. As one moves among institutional settings – from office, through tribunal, to lower courts, and finally to the higher courts – the role of para-lawyers tends to decrease. Para-lawyers are most often excluded from law work in upper-level trial and appellate courts, and least often from tasks that can be completed in the privacy of an office. For example, in England, lawyers enjoy a monopoly on representation in the higher courts; in contrast, office-based work generally is not limited to lawyers, nor is representation before administrative tribunals. As noted previously, legal executives in England can now appear in some court proceedings to handle preliminary matters in criminal cases, and licensed paralegals in Ontario can handle small claims litigation, traffic offenses, and minor criminal offenses in lower level courts.

Continuing with the example of England, para-lawyers direct a larger proportion of their effort toward providing advice on legal matters, broadly defined, while lawyers are more likely to engage in active assistance. In part this is because accounting firms and advice bureaus staffed by nonlawyers are major providers of legal advice in England (Baldwin, 1989; Citron, 1989). Similar advice bureaus also exist in New Zealand and in at least some states of Australia. In other countries, it is likely to be the case that para-lawyers are more involved in advice-giving simply because it is less visible, and hence difficult for regulators to limit or otherwise control.

The more that routines can be established to handle a legal task, the more possible it is for someone with very specific training and/or expertise to handle the task, even if it is highly technical. For example, in the United States, appeals of certain types of government benefit claims involve highly technical issues of both regulations and medical assessment. Such cases tend to be handled by specialists, some of whom are lawyers and some of whom are para-lawyers. On the less technical side, there are many legal processes that are routine but require the completion of various kinds of forms; para-lawyers can both identify the forms needed for various processes and assist consumers in the completion of those forms, although increasingly there are online services (e.g., LegalZoom, Nolo, Rocket Lawyer) that can assist consumers with such tasks.

One final issue concerning the structure of legal work and para-lawyers is whether they work as independent service providers or under the supervision of lawyers. Many lawyers employ para-lawyers to complete routine and/or specialized tasks within the lawyers' practices (Johnstone and Wenglinsky, 1985); many legal departments inside organizations employ a mix of lawyers and para-lawyers all under the supervision of senior lawyers (Kelly, 1994: pp. 96–99). Typically, para-lawyers working under the supervision of lawyers is an accepted delegation of responsibility, although there is substantial variation in drawing the line between what a para-lawyer may do and what requires a lawyer. As mentioned earlier, one frequent boundary is the courtroom door: representation in court tends to be limited to lawyers, although that may be less so in the lower courts (the example of Ontario noted previously) and it is changing under both political and economic pressures (the example of para-lawyers handling certain criminal matters in England, and the role of specialized advocates in certain types of cases such as domestic violence).

Effectiveness in Delivery of Legal Services

A key question with regard to the work of para-lawyers is how it compares in quality to the same activities carried out by lawyers. Although lawyers claim that use of para-lawyers raises concerns about poor-quality work, systematic comparisons of the work of lawyers and para-lawyers fail to support this contention (Kritzer, 1998; Moorhead et al., 2003). The extant research demonstrates that para-lawyers can be as effective as lawyers in providing specific legal services. Achieving effectiveness involves a combination of specific knowledge or expertise vis-à-vis substantive law, institutional process, and other actors in the system. Para-lawyers who obtain this combination can achieve levels of effectiveness comparable to that of lawyers, and specialized para-lawyers may be more effective than lawyers lacking one or more of the core elements of knowledge.

The relationship between the service provider and the client can also influence the relative effectiveness of the provider. Specialist para-lawyers who have ongoing relationships with clients may be more effective than even specialist lawyers brought in on an ad hoc basis to handle a specific matter; this reflects the knowledge base created by the ongoing relationship. The payment arrangement between service provider and client may also make a difference. One study (Kritzer, 1998: pp. 111–149) that found specialist lawyers to be more effective than specialist para-lawyers in a particular area attributed the difference in significant part to the lawyers being paid on a no-win, no-fee basis while the para-lawyers' compensation was unaffected by the outcome of specific cases. Another study found little or no difference between lawyers and para-lawyers, and in that study the lack of difference may have been due to different

compensation models for lawyers and para-lawyers that effectively allowed the para-lawyers to devote more time to an individual matter than could a solicitor (Moorhead et al., 2003).

Achieving Profession, Deprofessionalization, and Postprofessionalism

One issue for para-lawyers has been achieving recognition as a distinct profession. Typically this is reflected in efforts to establish a system of licensing that is somewhat under the control of the profession itself. Such a system imbues the occupation with status it previously lacked while also creating what sociologists of the professions have typically labeled 'closure' (Abel, 1989: pp. 20–24; Witz, 1992). Para-lawyers have had mixed success in their efforts to achieve closure. Most successful have been the English legal executives who are now regulated by their own body, the Institute of Legal Executives. Ontario paralegals may have achieved closure, but they are subject to regulation by the Law Society of Upper Canada, which has long been the regulatory authority for lawyers in Ontario. The legal document assistants in California who register with and are regulated by the state Department of Consumer Affairs have achieved an element of closure but not in a form associated with a professional status. Importantly, both the Ontario licensing scheme and the California system for what amounts to licensing legal document assistants (and which includes a second group called unlawful detainer assistants who handle eviction proceedings) was initiated by members of the legal profession rather than by members of the occupational group subject to the regulation.

'Deprofessionalization' refers to the shift from professional to nonprofessional status. Professional status is conferred on those occupations that maintain a monopoly over a theoretical base of knowledge, have expectation of autonomy in regulation and day-to-day work, and promote a belief in the professional having a service ethos. Deprofessionalization refers to the loss of these supposedly unique occupational qualities (Haug, 1973; Rothman, 1984). Nonprofessionals such as para-lawyers increasingly do work once the province of professionals. Deprofessionalization reflects several specific developments.

First is the combination of the rationalization of knowledge, the rationalization of tasks employing knowledge, and the growth of tools to assist in applying knowledge. The more that tasks can be compartmentalized and the more that knowledge-based tools can be applied, the easier it is to assign the task to a person with limited, specific expertise. The result is the de-skilling of knowledge-based occupations.

A second development is the loss of governmental support for restrictions limiting the role of para-lawyers. In England, solicitors held a monopoly on handling land transfers for many years, but these tasks are now open to licensed nonlawyer specialists. In the United States, limitations have been under attack politically at both the national and state level (Commission on Nonlawyer Practice, 1995).

The third development is the increasing bureaucratization of legal service delivery. Once organizational structures become established, there is a drive to rationalize, compartmentalize, and reduce costs. Bureaucratic structures also provide means for supervision and control, which both creates tensions for professionals expecting autonomy and provides a setting to monitor the work of para-lawyers.

A frequent claim by lawyers seeking to secure and maintain restriction on the delivery of legal services is that opening such work to para-lawyers and other nonlawyers jeopardizes the interest of clients both because of lack of competence and the absence of the stringent ethical norms of the professional. As noted in the previous section, systematic research in several countries has failed to support this claim (Kritzer, 1998; Moorhead et al., 2003). The loss of this argument combined with drives to contain costs through broadening options in service-oriented markets pushes toward less and less control of 'professional' services.

The concept of deprofessionalization suggests the loss of something. An alternative concept, 'postprofessionalism' (Kritzer, 1999), reflects most of the same elements but sees the opening of legal and other professional services to para-professionals, such as para-lawyers, as part of a natural development. Postprofessionalism builds on the growth of technological tools to access knowledge previously the exclusive realm of persons with professional-level training (Susskind, 1996, 2008). These developments parallel the industrial revolution, when the rationalization of tasks and the development of machines led to the shift from crafts-based production to factory-based production. Postprofessionalism captures the image of a similar shift for people who work with their heads rather than their hands: rationalization of knowledge combined with information technology tools make it possible to produce services using workers with lower levels of training and knowledge.

Gendering of Occupations

A final issue in the analysis of para-lawyers relates to the gendered nature of occupations. Historically, professions have been male-dominated while many para-professions have been female-dominated. Within law-work, women were largely excluded from the professional ranks until about 1970. In the early twentieth century, some of the American states formally excluded women from the bar; even where they were not excluded, women were often confined to roles such as legal secretary even if they had the formal credentials of a law degree. Since 1970, most countries have witnessed the opening of the legal profession to women, and the number of women who practice as lawyers has increased sharply.

Nonetheless, para-lawyer roles tend to be dominated by women; this is particularly true of para-lawyer positions that are specifically subordinate to lawyers (Johnstone and Wenglinsky, 1985: p. 69). Even for those para-lawyer positions that function independently of lawyers, the gender composition may limit the likelihood of achieving some sort of full professional status as long as they are seen as in competition with the existing legal profession (Hearn, 1982). At the same time, efforts to increase occupational status may lead to certain kinds of 'closure' strategies that bear a resemblance to those employed by lawyers: formal licensing, testing requirements, educational qualifications, and so forth (Witz, 1992; Johnstone and Wenglinsky, 1985: pp. 165–166). Whether such efforts ultimately improve status or lock in gendered distinctions is an important theoretical and empirical issue.

A final gender-related issue concerns the professional roles assumed by many of the increasing number of women in the legal profession. A common observation in the literature on the struggles of women to achieve status and economic parity with men in the legal profession is that women lawyers are disproportionately to be found in law-work positions that are themselves in some sense gendered (e.g., family law practice, as employees rather than as principals, etc.). In some ways, the professional roles occupied by many women lawyers are closer to those of para-lawyers than to the traditional image of the work of male lawyers (Sommerlad and Sanderson, 1998). This development may lead to a blurring of lines between lawyers and para-lawyers while preserving patriarchical status structures in the workplace.

Future Research and Theory

As theoretical and empirical concepts, para-lawyers (and para-professionals more generally) are underdeveloped. Current dynamics in the changing occupational structures, such as postprofessionalism, suggest that the para-lawyers should become increasingly important in the study of law work and in the delivery of legal services. The relationship of para-lawyers to the increasing role of information technology is one avenue of potential theory development and research; a second is the economics of service delivery; and a third potential area is in the relationship of para-lawyer occupations to broader patterns of changing social structures, particularly the changing impact of gender on social structure.

See also: Empirical Legal Studies; Gender and the Law; In-House Counsel; Law Firms; Law, Sociology of; Lawyers, Regulation of; Lawyers: Social Organization of the Profession; Professions and Professionalization, History of; Professions, Sociology of.

Bibliography

Abel, R.L., 1989. American Lawyers. Oxford University Press, New York.
Baldwin, J., 1989. The role of citizens advice bureaux and law centres in the provision of legal advice and assistance. Civil Justice Quarterly 8, 24–44.
Citron, J., 1989. The Citizens' Advice Bureaux: For the Community by the Community. Pluto Press, London, UK.
Commission on Nonlawyer Practice, 1995. Nonlawyer Activity in Law-Related Situations: A Report with Recommendations. American Bar Association, Chicago.
Haug, M.R., 1973. Deprofessionalization: an alternative hypothesis for the future. Sociological Review Monograph 20, 195–217.
Hearn, J., 1982. Notes on patriarchy, professionalization and the semi-professions. Sociology 16, 184–198.
Johnstone, Q., Wenglinsky, M., 1985. Paralegals: Progress and Prospects of a Satellite Occupation. Greenwood Press, Westport, CT.
Kelly, M., 1994. Lives of Lawyers: Journeys in the Organization of Practice. University of Michigan Press, Ann Arbor, MI.
Kritzer, H.M., 1998. Legal Advocacy: Lawyers and Nonlawyers at Work. University of Michigan Press, Ann Arbor, MI.
Kritzer, H.M., 1999. The professions are dead, long live the professions: legal practice in a post-professional world. Law & Society Review 33, 713–759.
Malavet, P.A., 1996. Counsel for the situation: the Latin American notary, a historical and comparative model. Hastings International and Comparative Law Review 19, 389–488.
Moorhead, R., Sherr, A., Paterson, A., 2003. Contesting professionalism: legal aid and nonlawyers in England and Wales. Law & Society Review 37, 765–808.
Rothman, R.A., 1984. Deprofessionalization: the case of law in America. Work and Occupations 11, 183–206.
Sommerlad, H., Sanderson, P., 1998. Gender, Choice and Commitment: Women Solicitors in England and Wales and the Struggle for Equal Status. Ashgate Publishing, Aldershot, UK.
Susskind, R.E., 1996. The Future of Law: Facing the Challenges of Information Technology. Clarendon Press, Oxford, UK.
Susskind, R.E., 2008. The End of Lawyers? Rethinking the Nature of Legal Services. Oxford University Press, New York.
Witz, A., 1992. Professions and Patriarchy. Routledge, London.

Relevant Websites

http://www.citizensadvice.org.uk/ – Citizen Advice Bureaux (UK).
http://www.cab.org.nz/Pages/home.aspx – Citizen Advice Bureaux (New Zealand).
http://www.cilex.org.uk – Legal Executives (England).
http://www.lsuc.on.ca/licensingprocessparalegal/ – Paralegals (Ontario).
http://www.calda.org/Visitors.asp – Legal Document Assistants (California).
http://wspaonline.org – Washington State Paralegal Association.

Parallel Relationships

Catherine Hakim, Civitas, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Parallel relationships encompass many types of extramarital sexual encounters and relationships, including flings, hookups, casual sex, mistresses, courtesans, concubines, swinging, sex parties, and long-term romantic or sexual relationships. Sex surveys suggest a rising trend in such recreational sexual activity, prompted by the contraceptive revolution of the 1960s. The incidence of extramarital affairs also rises with duration of the primary relationship, as boredom sets in. Parallel relationships may overlap with use of commercial sexual services, and are more common within highly educated and high-status groups.

Sex researchers in Scandinavia adopted the term parallel relationships to cover sexual or romantic liaisons, whether brief or long lasting, concurrent with any primary sexual or romantic relationship (Kontula, 2009). The term thus encompasses sexual relationships that are variously called hookups, flings, one-night encounters, swinging, casual sex, and extramarital affairs of any duration. Strictly speaking, hookups with men and women who sell sexual services should also be included, especially as these may become ongoing relationships with regular clients. The dividing line between professional and nonprofessional sexual liaisons that was so important in previous centuries in the Western world is fading rapidly in the twenty-first century. Just one example is the relationship between the 'sugar daddy' (or mommy) and the young, attractive girlfriend (or boyfriend) who they support financially (Croydon, 2011). However, the formal sale and purchase of sexual services, whether gay or heterosexual, have distinctive features, are criminalized in some countries (such as the USA), not in others (such as most of Europe and New Zealand), or is tolerated even when strictly unlawful (as in Thailand), so it is covered separately in this article.

It is essential to note that paid sexual services and unpaid casual sexual relationships (including affairs) are functional alternatives that substitute for each other and cannot always be differentiated (Hakim, 2011: 150–152). People who report two or more sexual partners in sex surveys may be referring to liaisons in the commercial sex industry as well as to ongoing lovers. By eliminating the risk of unwanted pregnancies, the contraceptive revolution of the mid-1960s made premarital and nonmarital sexual relationships less problematic. This opened the door to an expansion of recreational sex for people of all ages, whether married or not, an option that had previously been restricted to royalty and the aristocracy, people wealthy enough to cope with the financial costs of any consequential offspring. The defining characteristic of parallel relationships, like commercial sexual services, is a primary focus on recreational sex, or fantasies of recreational sex, and related entertainments, in contrast to the very different long-term goals of marriage and cohabitation, which usually include raising children and a common lifestyle.

The most immediate effect of the introduction of reliable contraception (the pill, intrauterine device, and sterilization) in the 1960s was a rapid expansion of premarital sex among young single people, often considered shocking at the time. The second, less visible, and much slower consequence was an increase in extramarital relationships, swinging, flings, and casual sex among people of all ages – what the French call *aventures* and *petites aventures*. The third effect was a decline in male demand for commercial sexual services (and a consequential fall in the prices for such services), as men switched to using nonprofessional sexual liaisons as a substitute for commercial services (Levitt and Dubner, 2009: 29–31). It is notable that the official report on the 1966 Swedish national sex survey, in line with comments by Swedish social scientists, takes it for granted that men can get as much sex as they want from nonprofessionals on the 'normal sex market' of nonmarital sex (Lewin, 2000; Hakim, 2011: 148–149). It thus makes sense to think of a wider sex market that includes amateurs as well as professionals, marital and nonmarital liaisons, whether heterosexual or homosexual – even if notional 'prices' and values vary between sections of the market due to scarcity and demand (Laumann et al., 1994: 11–12; Baumeister and Vohs, 2004; Hakim, 2011: 194–200).

The focus here is on extramarital sexual and/or romantic liaisons among people who are married or in committed relationships. The article focuses on modern societies where monogamy, or serial monogamy, is the norm, and on 'vanilla' sex, excluding the special features of minority sexual entertainments, clubs, and communities such as Bondage Domination Sadism Masochism (BDSM) and other sexual variations. For simplicity, longer-term relationships are labeled as marriages.

The statistical and survey evidence quoted is taken from the nationally representative sex surveys of the 1980–2010 period around the world that are reviewed by Hakim (2011: 31–61, 2012: 33–42). Case study evidence from around the globe is provided by Druckerman (2007).

Incidence and Trends

The incidence of affairs varies between countries and cultures, men and women, and high-libido and ordinary people. But there is no doubt at all that the incidence of parallel relationships is increasing in many modern societies, even if comprehensive and reliable statistics are not always available. Statistics for the various types of parallel relationship generally show an upward trend in the proportions reporting *ever* having an affair or fling, or reporting such activity in *the last year*, or some other

specified period. However, case studies and qualitative research indicate that depth interviews can double admissions of affairs, from 30 to 60%, in countries where they are considered deviant (Hakim, 2012: 18).

Between 1990 and 2000, the proportion of men and women admitting to extramarital affairs doubled in Britain, and then rose again in the next decade. The upward trend was unmistakable, even if there was some increase in willingness to report such activity, which is usually carefully concealed. Nonetheless, the incidence remains low compared to other European countries. In Britain, among people married for over 5 years, 10% of men and 5% of women admit to ever having an affair. In the USA and Britain, across all age groups from 18 to 60 years, 2% of wives and 5% of husbands admit to an affair in the last year (Hakim, 2012: 18).

The duration of a relationship is relevant: percentages rise with increasing durations of marriage (and hence with age). American, French, and English surveys all confirm that the incidence of extramarital affairs and flings rises the longer people have been married, from close to zero in the first 3 years of a relationship, the 'honeymoon' phase, to over 20% (among French men) and around 14% (among French women) after 5 years of marriage. This pattern is observed in all age groups, in both first and second marriages (Hakim, 2011: 126).

In some countries, surveys report the incidence of two or more sexual partners – in the last year, the last 5 years, or ever – without trying to specify the nature of the relationships. These statistics inevitably *under*state the incidence of parallel relationships, because a relatively high proportion of marriages is celibate, or virtually sexless, especially when spouses are over 40 years of age (Hakim, 2011: 41–44, 2012: 65–70). Men who report a single sexual partner may be referring to a mistress or casual hookup rather than their wife. This is clear from the fact that wives are generally twice as likely as husbands to report their marriage has been celibate – for the last month, the last year, last 5 years, or whatever period is used. Many husbands resolve the problem of a sexless marriage through affairs, the sex industry, or casual hookups.

The American 1992 sex survey (Laumann et al., 1994) did identify spouses in celibate marriages who had lovers: 15% of husbands and 8% of wives. Reports of affairs are broadly consistent across surveys in America: one-quarter of husbands and 15% of wives report ever having an affair. The proportions rise with age: one-fifth of men with 30–40 years of age vs two-fifths of men with 50–60 years of age ever had another sex partner. These figures suggest that about half of all affairs occur within celibate marriages. The survey also reports people who had two or more partners in the last year after controlling for age, marital status, race, education, and other factors, to arrive at a standardized average for people admitting to more than two partners (19% for men and 13% for women), or more than five partners (4% for men and 2% for women). Everything points to men being twice as likely as women to have several sexual partners during marriage, and even more so before or after marriage (Laumann et al., 1994: 210–217).

A common reaction to these figures is to say they must be unrealistic, as male and female sexual activity must balance out, overall. No, it does not – for three reasons. First, men and women who sell sexual services (as a full-time or part-time activity) do not participate in sex surveys, as these are clearly not addressed to their situation, and participation is voluntary. Since the vast majority of sex workers are females, this removes women with high numbers of partners from survey results. Furthermore, the vast majority of male sex workers are selling services to gay men, not women, so they are reported as partners by other men, not by female clients. Second, sexual markets are not restricted by national borders, unlike sex surveys. Men who have affairs, or purchase sexual services, may do so outside their own country, while away on business trips abroad or during sex tourism holidays in places with more tolerant policies such as Thailand or Spain. For example, Swedish surveys show that 80% of men who buy sexual services do so outside Sweden (Lewin, 2000; Hakim, 2011: 149). The policy of criminalizing such activities in Sweden leads to the industry being outsourced to foreign countries. Third, heterosexuals invariably overlook the impact of the gay community on overall sex statistics. All studies show that gay men report the highest numbers of partners, far above the maximum numbers reported by heterosexual men. In gay sex, both partners are male; no females are involved. These three factors together far outweigh any impact of men's tendency to exaggerate and women's tendency to minimize reports of sex partners.

Parallel relationships are more common in Europe and other countries (Druckerman, 2007; Hakim, 2012: 13–21). In Italy, one-quarter of men and one in eight women admit at least one affair at some time, with the highest incidence among graduates. Less than half of Italians (two in five) regard affairs as unacceptable. In Spain, half of all husbands and one-quarter of all wives admit to affairs. Affairs are not always successfully concealed: one in ten married people in Spain are aware that *both* spouses have had an affair. Attitudes are consistent: one-quarter of Spaniards do not regard sexual fidelity as fundamentally important, and fear of discovery is admitted to be the main barrier to engaging in such dalliances. Clearly, the values promoted by the Catholic Church are not accepted unquestioningly, and many Spaniards must be using effective contraception. In France, it is estimated that one-quarter of adults are having affairs at any one time. Attitudes in France are most permissive, as affairs are taken for granted as something that happens throughout life, if you are lucky. Two-thirds of French men and half of French women accept that sexual attraction inevitably leads to intimacy. Two-fifths of the French do not equate love with sexual fidelity, which is another way of saying that they continue to love their spouse even when seeing a lover. The incidence of parallel relationships seems to be highest in Finland, where the surveys do not bother to count flings and one-night encounters. Here, half of all men and one-third of all women have had at least one serious affair while married. A minority of men report over 10 such relationships. Sweden's reputation as the land of 'free love' must refer to premarital sexual activity rather than affairs, as only one-fifth of women and two-fifths of men have ever had an affair while living with a primary partner. In Japan, the 'love hotels' of big cities and country spa inns are used by married people and their lovers as much as by courting singles seeking privacy. Even in straight-laced China, one in ten husbands and one in twenty wives admit to at least one affair (Hakim, 2012).

Concurrent sexual partners are generally far more common among young people who are not yet married. However, most

sex surveys notice that there is a small minority of sexual 'superactives', who have an early sexual debut, numerous partners, and wide sexual experience. This group is dominated by men – gay, heterosexual, or both – who espouse a distinctive ethos of sexual exploration and adventure. For example, the Swedish survey found men were three times more likely to define the ideal sexual lifestyle as having a variety of concurrent lovers: 20% vs 6% of women (Lewin, 2000).

Several features of modern societies are conducive to the growth of affairs and flings, even in cultures that are firmly opposed to the idea, such as Britain and the USA, where 80 and 90%, respectively condemn infidelity (Hakim, 2012). With life expectancy increased to around 80 years in Western Europe, marriages can potentially last for 60 years or more today, whereas they lasted around 20 years on average in 1900. Boredom with the regular partner is an important factor in affairs; the posthoneymoon phase is now very long indeed. Another factor is the increasing secularization of cultures and values. The steady rise in civil (nonreligious) marriages is relevant: no 'vows' or promises are made, unless the couple actively chooses to include some, and they can specify their own norms and rules to suit themselves. A factor that influences some people is awareness of the hedonistic values of sexual adventure and exploration that apply in many parts of the gay community. Many gay men in long-term relationships still enjoy flings and casual sex with other men, sometimes in explicitly 'open' relationships. Other people look enviously at the sexual freedoms enjoyed by young singles, including their own children. A sexualized visual culture may have a diffuse impact on attitudes and behavior: advertising, pop videos, films, and TV display nudity and sexuality more frequently today than in the twentieth century, awakening or reinforcing libidos. The expansion of foreign travel and holidays is another factor, increasing awareness of the more tolerant attitudes toward affairs and flings in Southern Europe, for example. People who travel a lot also find it easier to have affairs in other locations, away from home and the constraints of normal life. Rising levels of education and greater affluence in a society are contributing factors, as all surveys show the incidence of nonmarital sexual liaisons to be highest in highly educated and high-income groups – because they have more relaxed and cosmopolitan attitudes, or because they can afford the inevitable costs.

Urbanization is another important macrolevel factor, as it increases the feasibility of affairs. Couples can meet more discreetly in cities and metropolitan areas than in small closed rural communities where everyone knows everyone else's activities, and gossip is a major form of social control. Urban areas increase the diversity of people encountered in workplaces, leisure centers, and other public places, and offer greater anonymity. Finally, internet dating and the rising number of specialized web sites catering for married dating facilitate meetings between like-minded people and fuels the expansion of what Hakim (2012) terms *playfairs*.

Another invisible but important factor may be the degree of multiculturalism in the local environment, as well as its size. For example, some 300 languages are spoken in London, England, making it the most multiethnic and multicultural city in the world. This type of cosmopolitan environment reinforces the idea of cultural relativism, and promotes the idea that you can step outside the norms of your own community by engaging with someone in a different community – a kind of 'virtual' cultural holiday. However, there is also a high incidence of affairs in culturally homogenous societies where extramarital liaisons are widely tolerated, so long as they remain discreet – such as France, Italy, or Japan.

Since 2000, there has been a proliferation of dating Web sites dedicated to married people seeking casual sex and affairs – in the USA, Britain, Scandinavia, and elsewhere. These Web sites offer a potentially global service. For example, a manager in Hong Kong who will be visiting London or New York on business can search Web sites for women interested in meeting him for a dinner date, and perhaps more. A professional in New York traveling to Paris for a meeting can search for potential partners in that city.

Typically, such Web sites boast that they are among the largest and most successful in their field, with more than 500 000 members often claimed. Many imply that they have broadly equal numbers of male and female subscribers – a claim that is completely implausible according to one study (Hakim, 2012). Typically, men pay subscription fees to use the service (to send and receive messages) whereas women pay nothing at all – in contrast to singles dating Web sites where both parties pay fees. This is a strong indicator that men vastly outnumber women. Hakim (2012) estimates that men outnumber women by at least 10 to 1, possibly 15 to 1, on these specialized Web sites. This illustrates the sex differential in interest in parallel relationships and recreational sex, and is consistent with the results of sex surveys.

The proliferation of married dating Web sites since 2000 reflects and facilitates increasing interest and activity in parallel relationships. Although such dating Web sites do not cost much to run, and rely on standard singles dating software, advertising can be costly, especially as normal advertising channels are often closed to such 'deviant' enterprises. Web sites that do not attract enough customers to become profitable close down fairly quickly. At least one dating Web site switched from supporting singles dating to focus on married dating instead because the niche market was more profitable.

Parallel relationships take three distinct forms: the traditional kept mistress; playfairs and flings; and the more 'democratic' modern swinging and sex parties. The social element declines sharply from the first to the last type.

The Kept Mistress

Throughout history, the mistress and concubine has been an indicator of status and wealth, just as a large number of wives remains a status symbol in countries where polygamy is still practiced – such as Nigeria. Chinese emperors had hundreds of concubines. French kings often had an official mistress (*maitresse en titre*), usually French and freely chosen, who was generally more popular than the official wife, who was inevitably a foreign princess chosen for political reasons. The most famous was Madame de Pompadour, the long-lasting mistress of King Louis XV. Such mistresses were invariably married women, hence respectable ladies. Old traditions persist in new forms. President Francois Mitterand had a long-lasting mistress who was almost a second wife, with a daughter raised in secret.

French law protects the privacy of everyone, including public figures, so the media did not dare discuss the matter until after Mitterand's death.

No statistics are available on the numbers of long-term mistresses, whether financially independent, or in receipt of financial support in one form or another. In some countries, such as Brazil, discreet 'sugar-daddy' arrangements between affluent (and possibly older) men and attractive young women are not stigmatized (Hakim, 2011: 198). There are now specialist dating Web sites in the USA and Britain for 'sugar-daddy' men and attractive young women, many of them students in higher education with debts to pay off. Typically, these relationships are short-lived, lasting for 6 months or so. The global reach of Web sites means that such arrangements are often cross-national in scope. A first-person account of such relationships shows that most if not all the women in question are highly intelligent, well-educated, autonomous, ambitious as well as attractive people (Croydon, 2011). The average age of the successful sugar daddy is around 40–50 years. Many are single men who are dedicated to their careers or business; they want female company at certain times, but choose to avoid the permanent encumbrance of a wife and (potentially) children. Others are older men seeking an attractive younger mistress.

Relationships that develop within a single country have the potential to be longer-lived, especially when they involve a married man and an unmarried woman, who effectively provides a second home as well as sexual entertainment. This type of relationship is unbalanced and potentially exploitative, in contrast to affairs where both partners benefit from the security of marriage (Hakim, 2012).

First-person accounts by modern mistresses in Europe range from those that take a perverse pride in an emotionally dependent secret relationship that confers no financial or social benefits whatsoever on the woman and restricts her social life (Griffin, 1999) at one extreme, to those that rage against the injustice of an exploitative relationship, even when the mistress has received substantial financial and other benefits from her wealthy lover (Symonds, 2007) – with every shade in between.

In the past, the mistress and the courtesan might have similar social status and problems of stigmatization, as illustrated by the Russian novel *Anna Karenina*, and the *grandes horizontales* (courtesans) of the French *Belle Epoque*. These arrangements have become rarer in the Western world, being replaced by playfairs and flings. They are sometimes revived by rising affluence. For example, Hong Kong businessmen who travel regularly to mainland China often maintain a mistress in the Chinese cities they visit, in order to enjoy home comforts while traveling (Druckerman, 2007: 251–270). In China, corruption scandals in big cities occasionally reveal officials with dozens of mistresses.

Playfairs and Flings

What Hakim (2012) terms playfairs and flings are nonmarital, or extramarital relationships of relatively short duration, normally shorter than the relationship with a mistress, and with less financial support. Playfairs may last weeks, months, or (rarely) years.

Historically, nonmarital sex was condemned because of the risk of pregnancy. For an unmarried woman, an illegitimate child represented a major financial burden and ruined her chances of marriage and respectability. Married women could avoid such problems (unless the lover was of a different race, which might be reflected in the baby's appearance). But they would be condemned for duping their husband into supporting another man's child. Men avoided these problems, and could engage in affairs with a lighter heart and even irresponsibility. After modern contraception broke the link between sexuality and fertility, affairs and flings became technically unproblematic. However, modern morality has yet to adapt to the new situation, especially in Puritan Christian cultures. In some countries, the old values became institutionalized in concepts of love, trust, and betrayal that insist on absolute sexual fidelity. Most surveys suggest that the AIDS crisis of the 1980s temporarily reinforced attitudes of disapproval and intolerance everywhere, due to the new health risks.

Two distinct cultures can be identified as regards the theory and practice of playfairs and flings. In Puritan Anglo-Saxon countries, such as the USA and Britain, affairs are widely condemned and are less common. In continental European countries such as Italy, Spain, and France, attitudes are more relaxed and tolerant, albeit never unanimously in favor, and affairs are far more common. Cultures where playfairs are accepted also have well-developed conventions regarding their conduct. For example, affairs should be conducted with elegance, courtesy, and great style in order to guarantee that the liaison offers pleasure, entertainment, and enjoyment on all levels. In France, for example, playfairs are almost a work of art, with elegant dinners, regular gifts, well-planned excursions, and weekends away in romantic places such as Venice or Marakesh. In Japan, when a man wishes to terminate an affair, it is customary for him to make a substantial financial parting gift to his lover, as compensation (Druckerman, 2007).

The general pattern is for Puritan countries to have low tolerance for affairs, high divorce rates, and serial monogamy. Countries with high tolerance for affairs tend to have much lower divorce rates and generally long-lasting marriages. In the USA, around half of all marriages end in divorce. In Britain, about one-third of first marriages end in divorce. Although divorce rates rose in Catholic countries after divorce became legalized through the civil courts, they remain much lower, as also in Japan. In effect, there seems to be a choice between serial monogamy and playfairs as alternative routes to achieving variety and novelty within long-term intimate relationships – for those who need it (Hakim, 2012). Either way, sexual fidelity has become a personal choice, not a necessity or an obligation, in secular modern societies.

In Britain, a large study of men and women engaged in extramarital affairs through the new specialized Web sites (Hakim, 2012) found that a wide variety of triggers and catalysts prompted an affair, such as: a partner's illness or unavailability for other reasons; the early death of a friend reminding them of their own mortality; a desire to enjoy (or regain) youthful vitality and sexual excitement; or a desire to experience the youthful hedonism that eluded them due to early marriage. However, the root cause of the majority of affairs was a celibate or sexually starved marriage. This common explanation is corroborated by evidence on celibacy

from national sex surveys and other research showing that men's sex drive is stronger than women's (Baumeister et al., 2001; Hakim, 2011: 31–61, 2012: 65–70).

In Europe and the USA, in the absence of any distorting social constraints, there are no sex differences in sexual interest among young people, up to the age of about 30 years. From then on, women start to lose interest in sex, while men's libido does not diminish. Half of all men over the age of 30 wish they had sex more often in their current relationship, compared to less than one in ten women. From the age of 35 years onward, celibacy increases rapidly among women, but rises very slowly among men (Hakim, 2011: 39–50). In the USA, one-fifth of marriages are sexually inactive. In Britain, among people with 45–59 years of age, one-fifth of married couples have been celibate for the last year, and one in ten have been celibate for 5 years or more.

The celibate character of many marriages is reflected in attitudes. In Britain, two-thirds of people agree that sex is not the most important element in a marriage. Curiously, this is in line with the continental European view that marriage is primarily about investments in children, property, a common lifestyle, values, and friends.

Inevitably, married people who have recourse to sex therapists and marriage counselors are a self-selected unhappy group who identify serious problems with their relationship that potentially lead to separation if not resolved. So the picture of affairs that emerges from books by counselors is typically negative, and differs fundamentally from the picture from wider population samples (Hakim, 2012: 35–42). Men and women using the dating Web sites for married people are typically happily married, and certain that they do not want to change their situation. They are not seeking an excuse to terminate their primary relationship, by 'falling in love' with someone else, or by provoking their spouse's rage at discovering an affair, to prompt a rapid divorce. These successful affairs are labeled *playfairs* by Hakim (2012) to differentiate them from the emotional–romantic liaisons that are both cause and consequence of an unsatisfactory marital relationship.

Paradoxically, men and women using the dating Web sites for married people often express strong family values, especially as regards their responsibilities for their children's well-being and long-term care. Some insist that they have a 'perfect marriage' except for the sharp imbalance between spouses in sexual interest. This leads them into strict procedures to maintain the secrecy of their affairs. In effect, they observe the two fundamental rules of the affair identified in French culture. First, an affair must never be obvious, should always remain discreet, so that the spouse is never embarrassed in any way. Second, affairs should never be conducted close to home, within your neighborhood, workplace, or friendship circle, as this raises the likelihood of accidental exposure and embarrassment.

Some users of the married dating Web sites treat their affairs primarily as sexual adventure. The emphasis is on sexual exploration and sensuality. This group veers toward casual sex and one-off hookups rather than continuing liaisons (Hakim, 2012), although some approach the continuity displayed in the film *Last Tango in Paris*. This type of heterosexual affair is a close parallel to affairs in the gay community, where casual and anonymous sex is far more common, among both partnered and unpartnered men (Weinberg and Williams, 1975). The key difference is that heterosexual affairs are almost invariably concealed; they are not part of an 'open' relationship, even when there is implicit tolerance of occasional flings.

Web sites for playfairs are popular because they facilitate secrecy. They allow people to avoid meeting anyone they might potentially know socially or through their occupation, by excluding people within the same neighborhood or type of business. The Web sites thus operate in the *opposite* way to singles dating, where people mostly choose to meet others who live in the same area, with similar occupations, or similar leisure interests. People on the married dating Web sites *actively exclude* anyone who might be too close to them, for example by avoiding people in their own ethnic community or religious group.

Given the statistics on celibate marriages, it is no surprise that men outnumber women on the Web sites for affairs. Some married men also use singles dating Web sites and singles bars with a view to achieving sexual encounters with young single women. Estimates suggest that around one-third of men on singles dating Web sites are married men 'on the prowl' for extra sex.

Physical and social attractiveness, or what Hakim (2011) terms *erotic capital*, is an important determinant of men's and women's success in attracting and keeping partners, and this is especially true of affairs (Hakim, 2011: 194–198, 2012). Long-term relationships often include investments in children, property, joint friends, common interests, and lifestyle. This sort of glue is missing in short-term liaisons, so erotic capital and entertainment are highlighted. Attractive people have an earlier sexual debut, and acquire more sexual experience. In Europe, they attract more sexual partners. On the specialized Web sites, people with high erotic capital attract more interest, have greater choice of partner, and can impose their own conditions for any liaison, including expectations of entertainment, gifts, and even financial support. Here again, there is an overlap between the affair, the 'sugar-daddy' relationship, the kept mistress, and the commercial sex industry, with fuzzy boundaries. Just as marriages sometimes involve an *implicit* exchange of beauty (hers) and money (his), affairs are sometimes based on an *explicit* exchange of erotic capital and economic capital.

Swinging, Sex Parties, and 'Open' Relationships

Swinging is a modern egalitarian form of sex orgy. The 'free for all' and competitive nature of the sex orgy means that some participants may have less choice of partner and less luck than others, especially when there are unequal numbers of men and women. Swinging is normally organized around couples (although singles sometimes join in to boost numbers). This ensures balanced sex ratios and equal opportunities for all participants (Brandon, 2008).

Swinging parties are normally arranged in private homes, nowadays through specialist Web sites or magazines. They seem to be more popular with men than women; wives report being pressurized into going along with their partner. They are seen to offer 'fair play', because nonmarital sex is available equally to both spouses, there are no secrets about what happens, everyone is clear that sexual adventure is the main entertainment, there are no emotional attachments, and the

semipublic nature of events eliminates any security risks. However, laws on pimping and brothels can make such sex parties technically illegal – once again underlining the lack of any sharp distinction between amateur and professional sex parties.

Participants in swinging parties sometimes engage in other sexual variations, such as threesomes of one type or another, and voyeurism. This category of nonmarital liaison is strongly focused on sexuality and sexual performance, quite independently of any social or emotional relationships. Swinging thus represents one extreme end of the parallel relationships continuum, along with traditional sex orgies. Catherine Millet's sexual memoirs offer a rare first-person account of involvement in these activities, from the perspective of a woman, one who is also a respected member of the French intelligentsia.

In 2011, Dominique Strauss-Kahn, then head of the International Monetary Fund (IMF) in Washington, became world famous for all the wrong reasons. Arrested in New York for imposing sexual demands on a hotel chambermaid, he was subsequently released without charge (because the woman was an unreliable witness). But he lost his job as head of the IMF, was unable to run for the French presidency in 2012 as he intended, was divorced by his wealthy journalist French wife, and was forced to pay millions in compensation to the hotel maid in order to avoid a civil suit in the USA. In 2013, a former lover published *Belle et Bete*, denouncing Strauss-Kahn's libidinous lifestyle. As details of his lively sexual life flooded into the media, he was shown to lack respect for women, and he became discredited as an academic, intellectual, and politician in France. One of the many media stories concerned his participation in regular sex parties (without his wife) in which prostitutes made up the shortfall of women. Prostitution is legal in France. Even so, Strauss-Kahn insisted he did not know that the women attending his sex parties were professionals – despite the fact that someone paid them. This example illustrates how commercial sexual services and other nonmarital sexual affairs and flings are substitutes for each other, and are enjoyed interchangeably by wealthy libidinous men.

A contrasting case is the Italian business magnate and politician Silvio Berlusconi. While he was President, media stories circulated about the sex parties he attended, to which beautiful call girls were invited. Prostitution is legal in Italy, and Berlusconi, a married man, was never embarrassed to admit that he enjoyed the company of party girls, and paid generously for their time. Despite numerous legal cases against him, including one for his liaison with an underage nightclub entertainer (which is illegal everywhere), he retained substantial public and political support, as evidenced by his party's strong result in the Italian 2013 hung election.

The situation in Puritan Anglo-Saxon countries is quite different. Swinging and sex parties may possibly occur, but the Puritan culture of the USA and Britain ensures they remain hidden and rarely acknowledged.

Sex parties and swinging are probably the dominant form of parallel relationship in the gay community. They generally take the form of gay men visiting gay bars, bathhouses, and parks late at night with a view to achieving one or more sexual encounters. These liaisons can be anonymous, impersonal, and brief. Men with regular partners use such venues as well as the unattached, both heterosexual and openly gay. Heterosexual men also organize and attend sex parties to engage discreetly in homosexual activities while leading a 'normal' heterosexual life. For example in New York, the classic French '5 to 7' *rendez-vous* for affairs is used for private sex parties in hotels, attended by men exclusively. Most then go home to their wives and children and classic heterosexual family lifestyle.

Conclusions

Parallel relationships may be kept hidden, not denied, or admitted frankly in an 'open' primary relationship. They may be publicly accepted or condemned, common or rare, involve amateurs or professionals. But whether they are flings, playfairs, mainly sexual, or romantic–emotional liaisons, parallel relationships remain separate and different from long-term marital relationships. A study of American executives concluded that only 3% of men eventually marry their mistress or lover (Halper, 1988; Hakim, 2012: 22, 277). It appears that partners for long-term and short-term relationships are usually chosen using quite different criteria.

There is no evidence that the expansion in parallel relationships of all kinds in developed economies threatens the institution of marriage. There is good evidence that parallel relationships are a functional alternative to patronage of the sex industry and to serial monogamy in the context of life expectancies of 80–100 years in modern societies of the twenty-first century.

However, for wealthy people who intend to remain childfree, or have already raised children in one marriage and are starting a new solo life as an older nonmarried hedonist, a succession of (multiple) playfairs can become a lifestyle choice. This option seems to be restricted to a tiny fraction of people who are both healthy and wealthy, a group that is coalescing and creating enclaves in locations where the wealthy congregate, such as the south of France and in cosmopolitan cities. For some men, the ideal sexual lifestyle of multiple lovers and playfairs can be realized (Hakim, 2012).

See also: Erotic Capital; Hooking up and Casual Sex; Lesbian, Gay, Bisexual and Trans-sexual: Commercial Sex; Popular Sexual Cultures in Western Europe (1930s to the Present); Sexual Motivation; Sexuality and Social Work; Sexuality and Space; Sexuality, Theories of; Sexuality: Cultural Aspects.

Bibliography

Baumeister, R.F., Vohs, K.D., 2004. Sexual economics. Personality and Social Psychology Review 8, 339–363.

Baumeister, R.F., Catanese, K.R., Vohs, K.D., 2001. Is there a gender difference in strength of sex drive? Theoretical views, conceptual distinctions, and a review of relevant evidence. Personality and Social Psychology Review 5, 242–273.

Brandon, M., 2008. Swinging: Games Your Neighbours Play. HarperCollins Friday Books, London.

Croydon, H., 2011. Sugar Daddy Diaries. Mainstream, London.

Druckerman, P., 2007. Lust in Translation: The Rules of Infidelity from Tokyo to Tennessee. Penguin Press, New York.

Griffin, V., 1999. The Mistress. Bloomsbury, London.

Hakim, C., 2011. Erotic Capital: The Power of Attraction in the Boardroom and the Bedroom. Basic Books, New York. Published in Britain as Honey Money: Why Attractiveness is the Key to Success. Penguin, London.

Hakim, C., 2012. The New Rules: Internet Dating, Playfairs, and Erotic Power. Gibson Square, London.

Halper, J., 1988. Quiet Desperation: The Truth about Successful Men. Warner Books, New York.

Kontula, O., 2009. Between Sexual Desire and Reality: The Evolution of Sex in Finland (M. Makinen, Trans.). Population Research Institute, Helsinki, Finland.

Laumann, E.O., Gagnon, J.H., Michael, R.T., Michaels, S., 1994. The Social Organisation of Sexuality: Sexual Practices in the United States. University of Chicago, Chicago, IL.

Levitt, S.D., Dubner, S.J., 2009. Super Freakonomics. Allen Lane, London.

Lewin, B. (Ed.), 2000. Sex in Sweden 1996. National Institute of Public Health, Stockholm.

Millet, C., 2002. The Sexual Life of Catherine M (A. Hunter, Trans.). Serpent's Tail, London.

Symonds, S.J., 2007. Having an Affair? A Handbook for the Other Woman. Red Brick Press, New York.

Weinberg, M., Williams, C., 1975. Gay baths and the social organisation of impersonal sex. Social Problems 23, 124–136.

Parenting Attitudes and Beliefs across Cultures

Çigdem Kagitcibasi and Bilge Yagmurlu, Koc University, Istanbul, Turkey

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article addresses parental attitudes, beliefs, and more generally parental orientations in terms of some topical categories, that is, parenting orientations and culture, parenting orientations and socioeconomic context, parenting orientations and family change, and parenting orientations and culture contact. Although these areas of study overlap to some extent, they provide a systematic framework to address the important and complex issues involved in parenting. Both contextual and universal approaches are pursued to examine differences and commonalities from a global perspective. A great deal of research is reviewed, and theoretical perspectives are put forward to provide insights into parenting attitudes and beliefs across cultures.

For some time, parenting styles and practices have been the main focus of research on parenting. However, it is now generally recognized that an appropriate model of parenting must also involve underlying factors such as parental cognitions. Although the association between parental beliefs and behaviors is usually found to be modest (belief–behavior dilemma) (e.g., Hofer et al., 2012), researchers have shown that their link is actually stronger when the focus is on implicit behaviors (e.g., nonverbal behavior) or beliefs (implicit cognitions of warmth and power), rather than explicit ones (e.g., verbal behavior or socialization goals) (Keller et al., 2007). This growing confidence regarding the correspondence of parental cognitions and practices has provided an additional impetus for studying parental beliefs. Accordingly, in the last decades, interest in parents' ideas (Cheah and Chirkov, 2008), values (Harwood, 1992) and goals (Durgel et al., 2009; Yagmurlu and Sanson, 2009) has increased. Accordingly, in this article, the authors will use the term 'parenting orientations' with an understanding that it includes parenting attitudes and beliefs, as well as behavior.

Most of the knowledge on parental socialization comes from studies of Western families. However, socialization is culture bound (Kagitcibasi, 2007). Culture influences all processes that affect developmental outcomes; it does not simply affect parental values or behaviors, but it is at the core of them. Children learn values and rules in a cultural context. Cultural values are transmitted to them as they are socialized by parents, family, and other agents in the larger society. Hence, human development must be understood in the context of family and family in the sociocultural context. This is the basic assumption of contextualism, which claims that behavior cannot be understood outside of the context in which it occurs (Kagitcibasi, 2000). The recognition that parental beliefs and attitudes should be studied in sociocultural context is of particular importance to develop an understanding of why certain beliefs show systematic variation among different groups. The answers are often found in underlying functional adaptive mechanisms. Thus, parents' values and beliefs about child rearing reflect parents' conceptualization of the functional attributes of an 'adaptive adult' in that culture (Rosenthal and Roer-Strier, 2001). In-depth studies within context, as, for example, carried out by anthropologists, have stressed the importance of 'lay' theories or beliefs about children, parents, and development, more recently called 'ethnotheories' (Greenfield et al., 2003; Harkness et al., 2011). These belief systems are key aspects of the 'psychology of the caretaker,' which in turn constitutes one of the three components of the 'developmental niche' (Harkness and Super, 1993).

While understanding variations in parental beliefs and attitudes in different contexts is important, going beyond them, one can also search for insights regarding commonalities, even possibly universals, mostly based on shared needs of protecting and nurturing the young and biologically anchored mechanisms of survival. Indeed, a concern for context need not and should not lead to relativism. Rather, one should endeavor to rise to the challenge of integrating the contextual and the universal approaches – "Achieving cultural contextualism without complete relativism" (Kagitcibasi, 2000). Such an approach promises to provide an understanding of the functional mechanisms underlying observed commonalities and differences. Nevertheless, this article is selective in focusing more on diversity, given its cultural emphasis. Notwithstanding this focus, one needs to be careful concerning the meaning attributed to similarities and differences. When a difference is found across cultures, a contextual (environmental) interpretation is used. However, when a similarity emerges, there is an inherent ambiguity here. Because, while shared biology is likely to be the reason, there can be other causes, such as shared psychological, ecological, social, or cultural structures (e.g., all known societies appear to have incest taboo). Or there may be the possibility of an interaction of biology and culture. Research involving evolutionary approaches to the study of culture and biology provides us with new insights into the study of human development in culture, which have implications for variations and commonalities in parental orientations (Keller et al., 2002).

Parenting Orientations and Culture

Given that culture is a powerful context that shapes societal and familial values, one faces the dilemma of coming up with

countless indigenous societal and familial values accompanying countless cultures. Such an approach would yield an unwieldy and basically incomparable body of knowledge. Thus, there is a need for systematic grouping of cultural contexts on the basis of a few underlying constructs that can serve as a paradigm.

Since 1980s, individualism–collectivism is the most widespread framework to explain the role of culture in parenting and indeed in most psychological phenomena (Kagitcibasi, 1997; Triandis, 1995). Although there are numerous definitions of the 'individualism' and 'collectivism' constructs, they refer mainly to cultures' different worldviews, value systems, and approaches to socialization (for an extensive coverage see, Kagitcibasi, 1997). A person in a collectivistic culture defines himself/herself in terms of membership of an in-group (Triandis, 1995). Strong family ties, interdependence, and obedience are valued in such a culture, and an obedience-oriented parenting style is often observed, aimed at achieving a compliant child who gives priority to group needs and close-knit ties. On the other hand, an individualistic culture values personal needs and aims to achieve autonomy and assertiveness in the child (Triandis, 1995). Individualistic parenting style tends to reinforce self-reliance and autonomy in the child, characteristics that are highly valued in individualistic societies. After a decade of extensive research and writing on individualism–collectivism, Markus and Kitayama's (1991) seminal paper has shifted the attention to independence and interdependence. Although different terms are used, there is a correspondence between independence and individualism and interdependence and collectivism. Research has shown that goals related to 'self-enhancement/autonomy' and 'compliance/relatedness to others' highlight a robust difference in maternal socialization across cultural groups.

Much research in Europe and the United States (e.g., Leyendecker et al., 2002; Ragavan et al., 2010) showed that Anglo-American and European mothers placed greater emphasis on individualistic goals (e.g., self-maximization such as self-confidence and self-reliance), whereas Latin-American and immigrant mothers emphasized goals that are representative of a more collectivist pattern (e.g., proper demeanor such as fulfillment of the role obligations and the ability to get along with others). In the same vein, German (Lamm and Keller, 2007) and Dutch mothers (Willemsen and van de Vijver, 1997) are found to strongly endorse independence and assertiveness goals, and even infants are expected to be able to play alone. Similar contrasting results are indicated for Chinese and English mothers. Pearson and Rao (2003) found that English mothers in the United Kingdom underlined the importance of letting the child develop his/her abilities as an individual, and Chinese mothers in Hong Kong were more likely to value filial piety, respectfulness, proper demeanor, and obedience. A study in Australia (Rosenthal et al., 1989) compared Greek-Australians, Anglo-Australians, and Greeks in Greece and their adolescent children. While Greek and Greek-Australian mothers emphasized being good and respectful as desirable child features, Anglo-Australians valued happiness, independence, and assertiveness in the child, which were indicative of individualism.

Further cross-cultural differences are observed in different parenting orientations, including verbal (discourse style) and nonverbal strategies (body contact and face-to-face contact) (Keller et al., 2011). This difference in parenting is stable at least across early childhood (Keller et al., 2010) suggesting that mothers display parenting strategies (e.g., distal vs proximal) that are most conducive to the child's adaptive development in that cultural context (Keller et al., 2011).

Parental beliefs do not concern only children, but also other aspects of family dynamics, including their cognition and construal of their own roles as parents. Parental role definitions have been shown to vary a great deal. For example, while Chinese and Japanese mothers are found to define their maternal roles to include coaching and supporting their children's school-related success (Chao, 2000; Jose et al., 2000), most other mothers do not (Dekovic et al., 2006; Nunes, 1993).

Parenting Orientations and Socioeconomic Context

Parenting attitudes and beliefs are closely related to socioeconomic status (SES) as well as culture. A great deal of research focusing on parental verbal behavior, in particular, has pointed to the variations in early home literacy environment with profound effects on child outcomes along social class differences (e.g., Aksu-Koç, 2005; Hart and Risley, 2003; Hoff, 2003; For an extensive review see Kagitcibasi, 2007: Chapter 3).

From early on, the general conclusion is that there are significant differences among cultural groups but they are lessened when SES (mainly maternal education) is taken into account (Cashmore and Goodnow, 1986). One of the earliest studies that examined the interaction between socioeconomic context and parental orientations is the classical work of Kohn (1963). Taking a sociological perspective, Kohn suggested that social class shapes a person's life perspective. Life perspective has an impact on which personal characteristics are perceived as desirable, influencing parenting goals and disciplinary behaviors. For example, working-class occupations require the person to obey directives of those in authority. Accordingly, working-class parents want their children to be obedient and respectful and they employ power assertion to achieve these functional behaviors. On the other hand, middle-class jobs emphasize self-direction rather than direct supervision; this affects middle-class parents' life perspectives and approaches to child rearing. They employ more autonomy granting goals (e.g., assertiveness and creativity).

Subsequent studies provided support to Kohn's theory (e.g., Ellis et al., 1978). For example, Tugde et al. (1999) showed that American, Korean, Russian, and Estonian parents did not differ in their self-direction goals, but this value was endorsed more highly by middle-class parents. Furthermore, control and discipline were valued more by working-class parents. In accord with this, middle-class children displayed more self-initiation in their activities, which was considered to be related to their mothers' goals for self-direction. Eisenberg (1996) further showed that working-class mothers in both Hispanic-American and Anglo-American families displayed less verbalization and more aversion (e.g., negative affect) in conflict situations than middle-class mothers, but no significant cultural difference was reported. Luster et al. (1989)

investigated American mothers with children 9–23 months old and also found results consistent with Kohn's hypothesis even at this young age. Higher social class standing was positively associated with parents' valuing self-direction (e.g., to think for himself/herself and to be curious about many things) and negatively associated with valuing conformity (e.g., to keep himself/herself and his/her clothes clean and to be polite to adults).

However, other findings contradict Kohn's (1959, 1963) assumption of linearity between parenting and social class (Abell et al., 1996). Harwood et al. (1996) found that although European American mothers coming from a lower socioeconomic background were more concerned with proper demeanor, as opposed to their high SES counterparts who emphasized self-actualization, Puerto Rican mothers coming from both middle and lower socioeconomic backgrounds endorsed goals for proper demeanor highly. Thus, the relative contributions of culture and social class can show variation, and this is an empirical issue.

An early large-scale study on the 'Value of Children' (VOC) (Kagitcibasi, 1982) provided insights into the place of the child in family and society and the changes thereof. The VOC study showed that in contrast with the United States and Germany, obedience was the most valued child characteristic, and independence and self-reliance were the least valued ones in the less-developed Indonesia, the Philippines, Turkey, and Thailand. 'Being close, loyal, and faithful to parents' was an attribute valued in adult children. Old-age security value of children (adult offspring seen as a source of economic security in old age) was also highly valued in less-developed contexts. This value was not endorsed in the United States and Germany, showing that children were raised to become independently functioning adults and were not expected to support their elderly parents. Similarly, another early study investigating parental values in Turkish families showed that parents from lower SES expected their children to be grateful to them when they grow up, whereas high-SES parents valued autonomy and did not expect gratitude, yet they wished for their children to be close to them (Imamoğlu, 1987).

A replication of the 'VOC' study across three generations in Turkey (Kagitcibasi and Ataca, 2005) revealed that material expectations from children declined with socioeconomic development, whereas no such decline was observed for emotional (psychological) interdependence. Other research (Chang, 2006) also suggests that, in line with changing cultural and economic circumstances, Chinese parents are coming to value attributes such as assertiveness, competitiveness and independence more than those traditionally valued such as control and inhibition. Such findings suggest the malleability of parenting in response to changing socioeconomic and cultural demands.

Parenting Orientations and Family Change

Up to this point, this article has revealed a great deal of diversity along cross-cultural, ethnic, and social class differences. At the same time, one sees that family orientations and parenting can change in response to changing sociocultural environmental demands. One is faced with challenges in making sense out of the complexity involved. In particular, one can aim to understand the systematic causal relationships that link culture/society, parenting, and child outcomes. For example, beyond asking 'What are the outcomes of a certain type of socialization?,' one can ask the question 'Why does a certain type of socialization occur in a certain type of context?' Kagitcibasi attempted to address these questions in her 'Family Change Theory' deriving originally from the VOC study results (Kagitcibasi, 1990, 2007; Kagitcibasi and Ataca, 2005). The authors will present this theory briefly as an example of a framework dealing with changes in parenting in response to changes in family and society.

Three prototypical family models are distinguished (Kagitcibasi, 1990, 2007). The first one is the family model of interdependence, which is mostly seen in the collectivistic, traditional, and rural contexts. The family pattern in this model is characterized by overall material and emotional intergenerational interdependence. The child has economic and utilitarian value and contributes to family well-being both as a child and especially as a grown-up offspring, thus the obedience of the child is functional and valued, while autonomy of the child is dysfunctional and is not fostered. This is because an autonomous child may grow up to be an independent young adult who attends to his/her own needs more than those of the family. Thus, the parenting style in this model is obedience oriented, and the self-developing in this family is the 'connected/relational self.' The contrasting model, the family model of independence, is prototypical of the individualistic Western middle-class nuclear family where intergenerational independence is valued. In this family model, relatively permissive and self-reliance-oriented parenting is functional since autonomy and separation/individuation are considered to be prerequisites of healthy development. The 'separate self' develops in this family. The third family model is prototypical of the urbanizing, more developed sectors of collectivistic societies with higher education. It entails material independence and psychological interdependence between generations. With increased affluence and changing life styles, dependence on the offspring for old age security decreases. Therefore, obedience-orientation decreases, and autonomy is granted in child rearing. Also, autonomy of the growing child becomes adaptive in higher education and specialized employment contexts that require individual decision making. However, given the collectivistic culture, the family still values close familial ties. Hence, the self-developing here is the 'autonomous-related self.'

Kagitcibasi's (1990, 2007) family change theory explains the relative salience of autonomy and/or relatedness in child-rearing orientations as a function of the family patterns that are impacted by both socioeconomic and also cultural factors and changes in these. It points to the possibility of being both autonomous and related. Furthermore, as autonomy and relatedness are generally considered to be basic human needs, the autonomous-related self is proposed as a healthy model of self (Kagitcibasi, 2007, 2012). The above discussion pointed to the importance of living conditions and environmental demands and the changes in these for parenting beliefs and behaviors. It is important to analyze these links among contextual factors and parenting to

understand why parents behave as they do. As mentioned earlier, parents have ethnotheories; they are aware of the adaptive value of different child characteristics and adjust their child-rearing orientations accordingly. The different family models discussed above render different child outcomes more or less functional.

Several studies provide supportive evidence for the family change theory (see Kagitcibasi, 2007 for an extensive review). For example, using in-depth semistructured interviews, Yagmurlu and colleagues (Yagmurlu, 2009; Yagmurlu et al., 2009) found that socialization goals of the less-educated Turkish mothers typically exemplify the family model of interdependence. These mothers mostly mentioned being kind and considerate toward others, and being compliant to societal norms as desirable child characteristics; and they emphasized an anticipation to be looked after by their children when they get older. However, mothers with high education tended to set socialization goals that enhance self-reliance and autonomy more than dependency. Nevertheless, maintaining warm and close relationships with family was a goal that these mothers shared with the less-educated mothers, in line with an earlier study by Imamoğlu (1987), mentioned above. Thus, socialization goals of the mothers with high education exemplify the family model of psychological interdependence. This implies that psychological interdependence is needed and preserved in the face of rapid urbanization, modernization, and economic development (Kagitcibasi, 2007).

Comparative work across countries concur. Mayer et al. (2012) in a study using the recent VOC study results from Germany, Turkey, and India found three different clusters corresponding to the independence, psychological interdependence, and the interdependence family models. As expected, the highest prevalence of the independent family model was found in Germany, that of psychological interdependent model in Turkey, and that of interdependent model in India. The relative significance of the utilitarian VOC (material interdependence) in the family emerges as the key factor differentiating the three family models. This factor is directly influenced by socioeconomic characteristics such as affluence. In the same vein, in a 27-country study with university students, Georgas et al. (2006) found evidence for the prevalence of the psychologically interdependent family.

Parenting Orientations and Culture Contact (Immigration and Acculturation)

The discussion up to now, especially in the last section, has drawn attention to the interactions between context (cultural and socioeconomic environment) and family/parenting orientations. The home culture (e.g., the values, practices, and cultural background of the child's immediate family) and the societal culture, together form the 'dual climate' of culture (Greenfield and Suzuki, 1998). In general, the 'cultural climate' in these two contexts is not very different; cultural/societal values and parental values tend to be congruent. However, at times there can be mismatches between changing environmental requirements and parenting conventions, which can cause problems for all concerned. When changes are profound, such as in immigration or sudden social change, the likelihood of mismatches may increase.

Thus, studying parenting in acculturation contexts gives us further insights into the dynamics of stability and change in parenting orientations. Indeed, the distinction between the home culture and the societal culture is usually larger for immigrant or minority children in multicultural societies. Immigrant children learn the norms and practices of the dominant culture through their peers and teachers at school and through other channels such as the media. Immigrants' parenting values and practices may also change as they are exposed to the host society (Bornstein and Cote, 2006).

When people move to a new culture, they take their cultural values and beliefs with them, and the reciprocal relations with the dominant society influence these original cultural values and beliefs. Berry (1997) defined this process of acculturation in terms of the immigrant's attitudes toward maintaining his/her own cultural identity and/or becoming a part of the host society, and described four types of possible response to the host culture – assimilation, separation, integration, and marginalization. These 'acculturation strategies' have significant implications for parenting orientations in culture contact contexts.

Reviews of research on ethnic and minority parenting (e.g., Garcia Coll et al., 1995) reveal that acculturation is an important predictor of mothers' developmental expectations, values, parenting style, and interaction with the child. Most studies that pertain to acculturation and parenting have been conducted in North America, Australia, and Europe. Research with Mexican-American groups showed that first-generation Mexican immigrants promoted greater interdependence than later generations (Delgato-Gaitan, 1994), and those mothers who were more integrated to the US culture (as indicated by their language preference) displayed less hostile control and inconsistency in their disciplinary strategies (Hill et al., 2003). Similarly, Puerto Rican mothers who embraced US values were found to display more warmth and involvement in their parenting (Calzada and Eyberg, 2002), and living in the United States was associated with a reduction in their use of corporal punishment (Jambunathan and Counselman, 2002). Aycan and Kanungo (1998) showed that Indo-Canadian immigrants who endorsed an assimilation attitude were more likely to support autonomy in their children, particularly in social life. Those who were integrated emphasized the role of parents in both guiding children toward schooling and career choices and teaching responsibilities toward the family. Indo-Canadian parents who tended to separate from the dominant culture favored control over children and attached more significance to moral life.

A study conducted in Australia with Greek immigrant mothers' parenting values and goals showed a shift toward the values of Australian mothers (Rosenthal et al., 1989). They combined an emphasis on autonomy goals, as the Australian mainstream mothers did, and on being respectful, like the Greek mainstream mothers. Yagmurlu and Sanson (2009) also found that Turkish immigrant mothers who had a tendency to integrate with Australian society reported higher levels of self-direction goals and inductive reasoning,

and lower levels of compliance goals and obedience-demanding behavior. Similar results were found in research conducted on parenting goals of immigrants in Europe. Leyendecker and colleagues have shown that second-generation Turkish mothers in Germany valued autonomy and self-maximization of their children significantly more than first-generation mothers (Leyendecker et al., 2006). However, even second-generation Turkish immigrant mothers expected their children to be respectful, starting from the early ages, like first-generation mothers. In a study with Dutch, German, Turkish-Dutch, and Turkish-German mothers, Durgel et al. (2009) found that Dutch and German mothers displayed a family model of independence, whereas Turkish immigrant mothers favored an interdependence model. Those who were more adjusted to the mainstream culture endorsed autonomy goals more, such as professional and educational success, and demanded less obedience from their children, thus moving to a 'psychological interdependence family model.' This study showed that family change theory (Kagitcibasi, 1990, 2007) was useful in understanding changing parenting orientations in immigration contexts.

In general, these studies point to the greater adoption of host society parenting orientations through generations. However, there are complex issues involved in acculturation research. For example, the SES standing, especially the education of the parent needs to be considered. Often, low-educated working class immigrants reflect interdependent traditional family values, while more-educated immigrants from the same original country may have psychologically interdependent families. Their parental orientations and relative adoption of host culture norms tend to be different. Another issue concerns the Western individualistic theoretical assumptions of the researchers. Because often there is not much recognition of the psychological interdependence family model and the autonomous-related self (Kagitcibasi, 2007), close-knit family relations and parental control tend to be interpreted in negative terms. For example, when Dekovic et al. (2006) in their study of several ethnic minorities in the Netherlands found that parents use strict discipline, but they are also warm toward their children, they called this an 'unlikely combination.' Indeed, Gonzales et al. (1996) noted earlier that often ethnic minority parents are too readily labeled authoritarian because they appear too controlling to the Western mainstream culture researcher. Some recent volumes on acculturating families and parenting orientations in culture contact aim to provide new insights into these complexities and address problems and their solutions (e.g., Bornstein and Cote, 2006; Masten et al., 2012).

In conclusion, one can say that examining family in culture contact contexts leads to a better understanding of how parenting orientations change. This article has attempted to address parental attitudes, beliefs, and more generally parental orientations in terms of some topical categories, that is, parenting orientations and culture, parenting orientations and socioeconomic context, parenting orientations and family change, and parenting orientations and culture contact. Although these topical areas of study tend to overlap to some extent, they provide a systematic framework to address the important and complex issues involved in parenting orientations across cultures.

Bibliography

Abell, E., Clawson, M., Washington, W.N., Bost, K.K., Vaughn, B.E., 1996. Parenting values, attitudes, behaviors, and the goals of African American mothers from a low-income population in relation to social and societal contexts. Journal of Family Issues 17 (5), 593–613.

Aksu-Koç, A., 2005. Role of the home-context in the relations between narrative abilities and literacy practices. In: Ravid, D., Bat-Zeev Shyldkrot, H. (Eds.), Perspectives on Language and Language Development. Kluwer, Dordrecht, pp. 257–274.

Aycan, Z., Kanungo, R.N., 1998. Impact of acculturation on socialization beliefs and behavioral occurrences among Indo-Canadian immigrants. Journal of Comparative Family Studies 29, 451–467.

Berry, J.W., 1997. Immigration, acculturation and adaptation. Applied Psychology: An International Review 46, 5–68.

Bornstein, M.H., Cote, L.R., 2006. Parenting cognitions and practices in the acculturative process. In: Bornstein, M., Cote, L. (Eds.), Acculturation and Parent–Child Relationships: Measurement and Development. Lawrence Erlbaum, New Jersey, pp. 173–196.

Calzada, E.J., Eyberg, S.M., 2002. Self reported parenting practices in Dominican and Puerto Rican mothers of young children. Journal of Clinical Child and Adolescent Psychology 31, 354–363.

Cashmore, J.A., Goodnow, J.J., 1986. Influences on Australian parents' values: ethnicity versus socioeconomic status. Journal of Cross-Cultural Psychology 17, 441–454.

Chang, L., July 2006. Confucianism or Confusion: Parenting Only Children in Urban China. Keynote Address Presented at the 19th Biannual Conference of the International Society for the Study of Behavioral Development. Melbourne, Australia.

Chao, R., 2000. The parenting of immigrant Chinese and European American mothers: relations between parenting styles, socialization goals, and parental practices. Journal of Applied Developmental Psychology 21, 233–248.

Cheah, C.S.L., Chirkov, V., 2008. Parents' personal and cultural beliefs regarding young children: a cross-cultural study of Aboriginal and Euro-Canadian mothers. Journal of Cross-Cultural Psychology 39, 402–423.

Dekovic, M., Pels, T., Model, S., 2006. Unity and diversity in child rearing: looking back, looking forward. In: Dekovic, M., Pels, T., Model, S. (Eds.), Child Rearing in Six Ethnic Families: The Multi-cultural Dutch Experience. Edwin Mellen Press, Lewiston, pp. 279–303.

Delgato-Gaitan, C., 1994. Socializing young children in Mexican American families: an intergenerational perspective. In: Greenfield, P.M., Cockings, R.R. (Eds.), Cross-Cultural Roots of Minority Child Development. Erlbaum, Hillsdale, NJ, pp. 55–86.

Durgel, E., Leyendecker, B., Yagmurlu, B., Harwood, R., 2009. Sociocultural influences on German and Turkish immigrant mothers' long-term socialization goals. Journal of Cross-Cultural Psychology 40, 834–852. http://dx.doi.org/10.1177/0022022109339210.

Eisenberg, N., 1996. The conflict talk of mothers and children: patterns related to culture, SES, and gender of child. Merrill-Palmer Quarterly 42 (3), 438–452.

Ellis, G.J., Lee, G.R., Petersen, L.R., 1978. Supervision and conformity: a cross-cultural analysis of parental socialization values. American Journal of Sociology 84 (2), 386–403.

Garcia Coll, C.T., Meyer, E.C., Brillon, L., 1995. Ethnic and minority parenting. In: Bornstein, M.H. (Ed.), Handbook of Parenting, Vol. 2. Biology and Ecology of Parenting. Lawrence Erlbaum, Hillsdale, NJ, pp. 189–209.

Georgas, J., Berry, J.W., van de Vijver, F., Kagitcibasi, C., Poortinga, Y.H. (Eds.), 2006. Families Across Cultures: A 30 Nation Psychological Study. Cambridge University Press, Cambridge.

Gonzales, N.A., Cauce, A.M., Mason, C.A., 1996. Interobserver agreement in the assessment of parental behavior and parent-adolescent conflict: African American mothers, daughters, and independent observers. Child Development 67, 1483–1498.

Greenfield, P.M., Maynard, A.E., Childs, C.P., 2003. Historical change, cultural learning, and cognitive representation in Zinacantec Maya children. Cognitive Development 18, 455–487.

Greenfield, P.M., Suzuki, L.K., 1998. Culture and human development: implications for parenting, education, pediatrics, and mental health. In: Damon, W. (Ed.), Handbook of Child Psychology, fifth ed., vol. 4. John Wiley & Sons, Inc., New York, pp. 1059–1109.

Harkness, S., Super, C.M., 1993. The developmental niche: implications for children's literacy development. In: Eldering, L., Lesemen, P. (Eds.), Early Intervention and Culture: Preparation for Literacy. UNESCO, Paris, pp. 115–132.

Harkness, S., Super, C.M., Mavridis, C.J., 2011. Parental ethnotheories about children's socioemotional development. In: Chen, X., Rubin, K.H. (Eds.), Socioemotional Development in Cultural Context. Guilford Press, New York, NY, pp. 73–98.

Hart, B., Risley, T.R., 2003. The early catastrophe: the 30 million word gap by age 3. American Educator 27, 4–9.

Harwood, R.L., Schölmerich, A., Ventura-Cook, E., Schulze, P.A., Wilson, S.P., 1996. Culture and class influences on Anglo and Puerto Rican mothers' beliefs regarding long term socialization goals and child behavior. Child Development 67, 2446–2461.

Hill, N.E., Bush, K., Roosa, M., 2003. Parenting and family socialization strategies and children's mental health: low-income Mexican-American and Euro-American mothers and children. Child Development 74, 189–204.

Hofer, J., Schröder, L., Keller, H., 2012. Motivational basis of body contact: a multicultural study of mothers with their three-month-old infants. Journal of Cross-Cultural Psychology 43 (5), 858–876. http://dx.doi.org/10.1177/0022022111414418.

Hoff, E., 2003. The specificity of environmental influence: socioeconomic status affects early vocabulary development via maternal speech. Child Development 74, 1368–1378.

Imamoğlu, O., 1987. An interdependence model of human development. In: Kagitcibasi, C. (Ed.), Growth and Progress in Cross Cultural Psychology. Swets and Zeitlinger, Lisse, The Netherlands, pp. 138–145.

Jambunathan, S., Counselman, K.P., 2002. Parenting attitudes of Asian Indian mothers living in the United States and in India. Early Child Development and Care 172, 657–662.

Jose, P.E., Huntsinger, C.S., Huntsinger, P.R., Liaw, F.-R., 2000. Parental values and practices relevant to young children's social development in Taiwan and the United States. Journal of Cross-Cultural Psychology 31, 677–702.

Kagitcibasi, C., 1982. Old-age security value of children: cross-national socioeconomic evidence. Journal of Cross-Cultural Psychology 13, 29–42.

Kagitçibasi, C., 1990. Family and socialization in cross-cultural perspective: a model of change. In: Berman, J. (Ed.), Nebraska Symposium on Motivation: Cross-Cultural Perspectives. University of Nebraska Press, Lincoln, pp. 135–200.

Kagitcibasi, C., 1997. Individualism and collectivism. In: Berry, J.W., Segall, M.H., Kagitcibasi, C. (Eds.), Handbook of Cross-Cultural Psychology, vol. 3. Allyn & Bacon, Boston.

Kagitcibasi, C., 2000. Cultural contextualism without complete relativism in the study of human development. In: Comunian, A.L., Gielen, U. (Eds.), International Perspectives on Human Development. Pabst, Rome, pp. 97–115.

Kagitcibasi, C., 2007. Family, Self, and Human Development Across Cultures: Theory and Applications, second ed. Lawrence Erlbaum (Taylor & Francis), Mahwah, NJ.

Kagitcibasi, C., Ataca, B., 2005. Value of children and family change: a three-decade portrait from Turkey. Applied Psychology: An International Review 54, 317–337.

Keller, H., Abels, M., Borke, J., Lamm, B., Wingshan, L., Yanjie, S., Yifang, W., 2007. Socialization environments of Chinese and Euro-American middle-class babies: parenting behaviors, verbal discourses and ethnotheories. International Journal of Behavioral Development 31 (3), 210–217.

Keller, H., Borke, J., Chaudhary, N., Keller, H., Lamm, B., Kleis, A., 2010. Continuity in parenting strategies: a cross-cultural comparison. Journal of Cross-Cultural Psychology 41 (3), 391–409. http://dx.doi.org/10.1177/0022022109359690.

Keller, H., Borke, J., Lamm, B., Lohaus, A., Dzeaye, Y.R., 2011. Developing patterns of parenting in two cultural communities. International Journal of Behavioral Development 35 (3), 233–239. http://dx.doi.org/10.1177/0165025410380652.

Keller, H., Poortinga, Y.H., Schölmerich, A., 2002. Between Culture and Biology: Perspectives on Ontogenetic Development. Cambridge University Press, Cambridge, UK.

Kohn, M.L., 1963. Social class and parent child relationships: an interpretation. American Journal of Sociology 68, 471–480.

Lamm, B., Keller, H., 2007. Understanding cultural models of parenting: the role of intracultural variation and response style. Journal of Cross-Cultural Psychology 38, 50–57.

Leyendecker, B., Harwood, R.L., Lamb, M.E., Schölmerich, A., 2002. Mothers' socialization goals and evaluations of desirable and undesirable everyday situations in two diverse cultural groups. International Journal of Behavioral Development 26, 248–258.

Leyendecker, B., Schölmerich, A., Citlak, B., 2006. Differences and similarities among first- and second-generation Turkish immigrants in Germany. In: Bornstein, M.H., Cote, L.R. (Eds.), Acculturation and Parent–Child Relationships: Measurement and Development. Erlbaum, Mahwah, NJ, pp. 297–315.

Luster, T., Rhoades, K., Haas, B., 1989. The relation between parental values and parenting behavior: a test of the Kohn hypothesis. Journal of Marriage and the Family 51, 139–147.

Markus, H.R., Kitayama, S., 1991. Culture and the self: implications for cognition, emotion and motivation. Psychological Review 98, 224–253.

Masten, A.S., Liebkind, K., Hernandez, D.J. (Eds.), 2012. Realizing the Potential of Immigrant Youth. Cambridge University Press, New York.

Mayer, B., Trommsdorff, G., Kagitcibasi, C., Mishra, R.C., 2012. Family models of independence/interdependence and their intergenerational similarity in Germany, Turkey, and India. Family Science 3, 64–74. http://dx.doi.org/10.1080/19424620.2011.671503.

Nunes, T., 1993. Learning mathematics: perspectives from everyday life. In: Davis, R., Maher, C. (Eds.), Schools, Mathematics, and the World of Reality. Allyn and Bacon, Needham Heights, MA, pp. 61–78.

Pearson, E., Rao, N., 2003. Socialization goals, parenting practices and peer competence in Chinese and English preschoolers. Early Child Development and Care 173, 131–146. http://dx.doi.org/10.1080/0300443022000022486.

Ragavan, C.S., Harkness, S., Super, C.M., 2010. Parental ethnotheories in the context of immigration: Asian-Indian immigrant and Euro-American mothers and daughters in an American town. Journal of Cross-Cultural Psychology 4, 617–632. http://dx.doi.org/10.1177/0022022110362629.

Rosenthal, D.A., Bell, R., Demetriou, A., Efklides, A., 1989. From collectivism to individualism? The acculturation of Greek immigrants in Australia. International Journal of Psychology 24, 57–72.

Rosenthal, M., Roer-Strier, D., 2001. Cultural differences in mothers' developmental goals and ethnotheories. International Journal of Psychology 36 (1), 20–31. http://dx.doi.org/10.1080/00207590042000029.

Triandis, H.C., 1995. Individualism and Collectivism. Westview, Boulder, CO.

Tugde, J.R.H., Hogan, D.M., Lee, S., Meltsas, M., Tammeveski, P., Kulakova, N.N., Snezhkova, I.A., Putnam, S.A., 1999. Cultural heterogeneity: parental values and beliefs and their preschoolers' activities in the United States, South Korea, Russia, and Estonia. In: Göncü, A. (Ed.), Children's Engagement in the World. Cambridge University Press, New York, pp. 62–96.

Willemsen, M.E., van de Vijver, F.J.R., 1997. Developmental expectations of Dutch, Turkish-Dutch, and Zambian mothers: towards an explanation of cross-cultural differences. International Journal of Behavioral Development 21, 837–854.

Yagmurlu, B., 2009. Culture and migration: a comparison of Turkish migrant and non-migrant mothers' long-term socialization goals. In: Icduygu, A., Kirişçi, K. (Eds.), Land of Diverse Migrations. Challenges of Emigration and Immigration in Turkey. Bilgi University Press, Istanbul, pp. 299–363.

Yagmurlu, B., Citlak, B., Dost, A., Leyendecker, B., 2009. Child socialization goals of Turkish mothers: an investigation of education related within-culture variation. Turkish Journal of Psychology 24 (63), 1–15.

Yagmurlu, B., Sanson, A., 2009. Acculturation attitudes and parenting in Turkish-Australian mothers. Journal of Cross-Cultural Psychology 40, 361–380. http://dx.doi.org/10.1177/0022022109332671.

Parents and Teachers as Partners in Education

Janet Goodall, University of Bath, Somerset, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.A. Payne, volume 16, pp. 11045–11048, © 2001, Elsevier Ltd.

Abstract

This article examines the concept of parental engagement with children's learning. First it looks at what this concept entails, then at the value that this concept has been shown to have. It then moves on to examine how support can be offered for parental engagement, under the headings of home–school links, and support and training for parents; both of these sections include examples. The article then moves on to examine the barriers to parental engagement, then to critique the field of research around this topic, and to suggest future areas for research.

Parental Engagement with Their Children's Learning

What Is Parental Engagement in Children's Learning?

Parental involvement with the learning of their children may be rightly called ubiquitous – from a child's first moments of life, it is interaction with their parents and carers. (In this article, the term 'parents' is used to signify those adults who have the primary care of minors.)

Children learn the basics of social interaction from their parents and within the family unit. This learning link to parents – and the wider group of the family and community – continues throughout the child's life. Although schools and school staff have long acknowledged the importance of this learning link (with early parent-education classes being offered in the mid-1800s in the USA; Turney and Kao, 2009), it is only with the rise of compulsory schooling, and the view that education takes place within the boundaries of the school, that there has been a need to investigate the role that parents play in relation to their children's learning. What has come to be acknowledged with increasing force over the past 30 years is the value of parental engagement with children's learning. Conceptions of this process have also changed, and we are now able to state that, when properly supported, parental engagement with children's learning can be "one of the most powerful levers for school improvement" available (Harris and Goodall, 2008). This article approaches the issue in relation to parental engagement in children's learning aligned to the work of schools, as both have the same ultimate aim, that of the best possible outcome for the child.

Originally, all learning took place within the family unit. Early interventions were aimed at supporting parents in relation to child health, and in creating healthy, well-disciplined homes. However, as early as in 1977, there was a call for training to support parents to negotiate educational and social systems (Gordon, 1977). By the late 1980s, there was research highlighting the value of supporting parents of preschool children (Powell, 1989). Governmental agencies took up the banner, mandating parental involvement and participation in schools and other agencies. In fact, there was support even in the mid-1980s for parents teaching their children to read and write (Berger, 1991). In spite of these calls and evidence, however, parental engagement with their children's learning still lacks adequate support in many schools.

Recently, there has been a move from conceptions of parental involvement (see, e.g., Baumrind, 1989; Catsambis, 2001; Brooks, 2004; Berthelsen and Walker, 2008), to a discussion of parental engagement (Barton et al., 2004; Carréon et al., 2005; Harris and Goodall, 2008; Epstein, 2009; Hollingworth et al., 2009; Goodall and Vorhaus, 2011; Goodall, 2012, 2013; Goodall and Montgomery, 2013). Encompassed within this change is a more fundamental movement, from seeing parents as secondary partners in the schooling of their children, to seeing them as fundamental to the wider concept of children's learning. We may now define parental engagement as "parents' engagement in their children's lives to influence the children's overall actions" (Kim, 2009: p.89).

This article addresses parental engagement in the following sections, namely the value of parental engagement, support for parental engagement and the forms it can take, barriers to that engagement, critique of the research in the field, and areas for further research.

What Is the Value of Parental Engagement?

Literature in the field, while not unanimous and while certainly not complete (Goodall and Vorhaus, 2011), gives strong indications of the value of parental engagement with children's learning. The value of this engagement is not necessarily affected by the background of the family (Catsambis, 2001; Sylva et al., 2003). Parental engagement can be shown to be positively associated with children's achievement (Smith and Hausafus, 1998; Harris and Goodall, 2007; Hill and Tyson, 2009; Jeynes, 2012) and with children's self-regulation in learning (Xu et al., 2010); supporting parental engagement can impact positively on young people's motivation, self-efficacy, engagement in learning, and achievement (Fan and Williams, 2010).

There is a need for a broad understanding of learning and education. While it is clear that children learn in venues other than school, 'education' is often considered to take place within schools alone. However, children's achievement can be supported by blurring these lines; training parents to teach their children to read (rather than merely listening to them doing so) can have significant positive impact (up to 6 months of reading

age; Erion, 2006; Sénéchal and Young, 2008; Sylva et al., 2008b).

Parenting programs, aimed at supporting parenting skills, appear to have more modest but still positive effects, leading to reductions in tension between parents and teenage children (Ralph and Sanders, 2003), better relationships between parents and school staff, and increased bonding between mothers and young children (Erion, 2006; Reid et al., 2007; Lexmond et al., 2010), and increased parental confidence (Norwood et al., 1997). The positive effects of even short-term interventions to support parenting have been recognizable over many years (Brooks et al., 2008).

Afterschool, family engagement programs have been shown to have a positive effect on students (Lexmond et al., 2010) and attainment (McDonald et al., 2006; Brooks et al., 2008), in both literacy and numeracy.

Support for Parental Engagement

Parental engagement may usefully be discussed under two headings: home–school links and support and training for parents. It should be stressed, however, that these are not separate strands, which can stand alone. The greatest benefit from support for parental engagement will come from holistic, coordinated efforts.

Home–School Links

This need for 'joined-up thinking' is nowhere more clear than in the area of home–school links. To adequately support parent's engagement, there is a need for a whole-school approach, underpinned by a broad understanding of that engagement, and supported by senior leaders.

Good results have been found when schools are proactive in working with parents, and where the process is supported by senior leaders (Redding et al., 2004; Etsyn, 2009). The links between home and school need to be based on an understanding of cooperation and partnership (Goodall and Montgomery, 2013). Examples of successful programs which have forced such links and had a positive impact on children's achievement include the Manchester Transition Project (Dyson et al., 2007), Solid Foundation Schools (Redding et al., 2004), School Home Practitioners (Rogers et al., 2010), Homes School Knowledge Exchange (Greenhough et al., 2007), and the Parent Support Advisor Trial (Lindsay et al., 2009).

Part of an effective exchange of knowledge between home and school relies on the information passed between them. Studies have found that not all information sent to homes by schools is fit for this purpose (Crozier and Davies, 2007; Harris and Goodall, 2007). Information needs to be in language(s) and formats which are accessible to parents, and which contain information parents can use to support the learning of their children. Parents given clear, easily followed suggestions are more able to support their children's learning than those given more general information (Toomey, 1993).

This exchange of knowledge works best when it is based on a broad and accurate understanding of parental engagement. Research has shown that parents, school staff, and children conceive of parental engagement differently (Martinez and Velazquez, 2000; Harris and Goodall, 2007). The engagement which is most valuable is support for learning in the home (Desforges and Abouchaar, 2003; Sylva et al., 2008a), which aligns with what older pupils reported seeing as the greatest value of parental engagement – moral support (Harris and Goodall, 2007). Schools may not recognize all the forms of support parents are giving to students (Lopez et al., 2001), leading to an undervaluing of parents by schools, and contributing to feelings of alienation by parents.

Examples

In this area, as in all others, it is very important that interventions are underpinned by a coherent, whole school approach. Examples under this heading would include out-of-hours activities (clubs, the FAST program, and outreach programs). Links can also be forged by careful planning of school work to include parents, and by activities designed to help pupils link home and school (homework about interviewing family members and bringing photos of home to school).

Support and Training for Parents

Interventions to support parental engagement are more effective if they are based on clear knowledge and understanding of the needs and desires of parents (Lopez et al., 2001; Carpentier and Lall, 2005; Feiler et al., 2006; Brooks et al., 2008; Statham et al., 2010). Parents may need support in areas such as literacy and numeracy to support their children (Statham et al., 2010). Programs to support parents also need to show cultural sensitivity to parents and communities (Kumpfer and Alvarado, 2003).

Research has shown that parental engagement is enhanced when parents receive clear, targeted information about supporting their children, rather than more general information. Parents also benefit from a clear understanding of what any school-based interventions will entail.

Examples

Parent-to-parent tutoring has been shown to be effective, allowing parents to feel more relaxed and confident. Training parents to teach their children to read, as mentioned above, has also been shown to have significant benefits, as can more general programs aimed at supporting parents' learning, such as Supporting Parents on Kids' Education (SPOKES).

What Are the Barriers to Parental Engagement in Children's Learning?

In various studies, parents have reported a number of barriers to their engagement with their children's learning, and to engaging with schools about that learning. The two most often reported barriers are time (Peters et al., 2007) and their own experience of education (Harris and Goodall, 2008; Harris et al., 2009).

Parents have reported a number of ways that lack of time poses issues for engagement. In relation to engaging with schools, parents are often unable to come into school during the day due to other commitments (work, caring for other children or relatives). These same commitments may also make it difficult for parents to find time to interact with their children at home.

Secondly, parents have reported that their own, negative experiences of education can present a significant barrier to anything which requires their presence at school or interaction with school staff. To overcome this, schools must seek to build up trust with groups of parents, to allow them to form new concepts of schools and schooling.

There are other issues which may also prevent parents from engaging with their children's learning or with schools. Numerous authors have discussed a cultural and social gap between groups of parents and schools (see the subsequent section), which renders some groups of parents less able to negotiate the school system or make their needs known. Many parents also experience difficulties with English, particularly if schools do not make information available in a range of languages.

Critique of Research in the Field

The foregoing must not be seen as offering parental engagement with children's learning as a panacea or uncontested good. Particularly, in the last few years, legitimate critiques of the concepts have risen. These tend to center on two questions. The first question, raised above, centers around the exact definition of parental engagement. Above, we examined the issue of differing definitions of parental engagement between different stakeholder groups; here, the issue relates to differences in perception of parental engagement between cultural groups. This is most obvious when there is a cultural gap between parents/families and school staff.

Research has critiqued the concept of parental engagement as given in the literature as one that reflects the practice of white, middle-class families. The practices of other groups are, therefore, sometimes perceived as lacking or in deficit and parents have reported feeling disenfranchised, with their support of their children being undervalued or ignored. Further, parents and families are conceived to be homogeneous groups, when in fact they exist in great variety. Many parental engagement interventions are seen to support what the school sees as good for students (and parents), and the appropriate roles for both, rather than being based on a cooperative stance between parents and teachers. Some parental engagement programs can actually be detrimental, in that they have been found to reinforce existing power differentials between parents and schools (Crozier et al., 2000; Henderson and Mapp, 2002).

Lack of educational experience and engagement can reduce the value that parents are able to take from interventions to support them (Manz et al., 2010), just as the lack of this experience can be a barrier to engagement. It is also possible that programs do not seem to have cultural validity for some groups of parents, where the values and norms of parents are not taken into account (Manz et al., 2010). However, when programs are appropriate and sensitive to the needs and norms of parents, significant benefit can be derived for children from disadvantaged backgrounds, outlasting the effects for more advantageous peers (Brooks et al., 2008).

The second question is raised around the claims made for parental engagement (Hartas, 2008). It is notoriously difficult to make robust claims of causality in educational research, and the area of parental engagement is no exception. It is very difficult to isolate the effects of any given intervention among the myriad of influences on the achievement of young people.

Future Research

As noted, while the field has benefitted from a considerable array of research, there are still significant gaps in the literature base. Further work is needed, particularly around issues such as access and equity, as well as wide-scale testing and trial of interventions to determine good and best practices across a wide range of families.

Conclusion

As can be seen, parental engagement with children's learning is an important area of school and educational life. It is not, however, an area which has received enough research and there are a number of significant questions still outstanding. We do not yet have enough information about the effects of various interventions, nor do we know enough about how to support different groups of parents. We do know that engaging parents in their children's learning is a significant lever for improvement.

See also: Behavioral Problems, Effects of Parenting and Family Structure on; Family and Schooling; Prosocial Behavior, Effects of Parenting and Family Structure on; School Achievement as Moderated by Sociodemographic Status: The US Case; School Achievement: Motivational Determinants and Processes.

Bibliography

Barton, A.C., Drake, C., Perez, J.G., Louis, K.S., George, M., 2004. Ecologies of parental engagement in urban education. Educational Researcher 33 (4), 3–12.

Baumrind, D., 1989. Rearing competent children. In: Damon, W. (Ed.), Child Development Today and Tomorrow. Jossey-Bass, San Francisco, CA, USA, pp. 349–378.

Berger, E.H., 1991. Parent involvement: yesterday and today. The Elementary School Journal 209–219.

Berthelsen, D., Walker, S., 2008. Parents' involvement in their children's education. Family Matters 79, 34–41.

Brooks, G., Pahl, K., Pollard, A., Rees, F., 2008. Effective and Inclusive Practices in Family Literacy, Language and Numeracy: A Review of Programmes and Practice in the UK and Internationally. CfBT Education Trust.

Brooks, R., 2004. My mum would be as pleased as punch if I actually went, but my dad seems a bit more particular about it: paternal involvement in young people's higher education choices. British Educational Research Journal 30 (4), 495–519.

Carpentier, V., Lall, M., 2005. Review of Successful Parental Involvement Practice for 'Hard to Reach' Parents. Institution of Education, University of London, London.

Carréon, G.P., Drake, C., Barton, A.C., 2005. The importance of presence: immigrant parents' school engagement experiences. American Educational Research Journal 42 (3), 465–498.

Catsambis, S., 2001. Expanding knowledge of parental involvement in children's secondary education: connections with high school seniors' academic success. Social Psychology of Education 5, 149–177.

Crozier, G., Davies, J., 2007. Hard to reach parents or hard to reach schools? A discussion of home-school relations, with particular reference to Bangladeshi and Pakistani parents. British Educational Research Journal 33 (3), 295–313.

Crozier, G., Dewey, J., Husted, T., Kenny, L., 2000. Community, society and the school. Educational Administration Abstracts 35 (3), 359–370.

Desforges, C., Abouchaar, A., 2003. The Impact of Parental Involvement, Parental Support and Family Education on Pupil Achievement and Adjustment: A Literature Review. Department of Education and Skills.

Dyson, A., Beresford, E., Splawnyk, E., Britain, G., 2007. The Manchester Transition Project: Implications for the Development of Parental Involvement in Primary Schools. DfES Publications.

Epstein, C., 2009. Rules of engagement: increasing students' engagement and parental involvement in maths. Learning Matters 14 (1), 37–41.

Erion, J., 2006. Parent tutoring: a meta-analysis. Education and Treatment of Children 29 (1), 28.

Etsyn, 2009. Good Practice in Parental Involvement in Primary Schools. In: Her Majesty's Inspectorate for Education and Training in Wales, vol. 27.

Fan, W., Williams, C.M., 2010. The effects of parental involvement on students' academic self-efficacy, engagement and intrinsic motivation. Educational Psychology: An International Journal of Experimental Educational Psychology 30 (1), 53–74.

Feiler, A., Greenhough, P., Winter, J., Salway, L., Scanlan, M., 2006. Getting engaged: possibilities and problems for home-school knowledge exchange. Educational Review 58 (4), 451–469.

Goodall, J., 2012. Parental engagement to support children's learning: a six point model. School Leadership & Management 1–18.

Goodall, J., 2013. Parental belief and parental engagement: how do they interact? Journal of Beliefs & Values 34 (1), 87–99.

Goodall, J., Montgomery, C., 2013. Parental involvement to parental engagement: a continuum. Educational Review 66 (4), 399–410.

Goodall, J., Vorhaus, J., 2011. Review of Best Practice in Parental Engagement. Department of Education, London.

Gordon, P., 1977. Commitments and developments in the elementary school curriculum 1870–1907. History of Education 6 (1), 43–52.

Greenhough, P., Hughes, M., Andrews, J., Goldstein, H., McNess, E., Osborn, M., Pollard, A., Stinchcombe, V., Yee, W., 2007. What Effect Does Involving Parents in Knowledge Exchange Activities during Transfer from Key Stage 2 to Key Stage 3 Have on Children's Attainment and Learning Dispositions?.

Harris, A., Andrew-Power, K., Goodall, J., 2009. Do Parents Know They Matter? Raising Achievement through Parental Engagement. Continuum International Publishing Group, London.

Harris, A., Goodall, J., 2007. Engaging Parents in Raising Achievement. Do Parents Know They Matter? Department for Children, Schools and Families.

Harris, A., Goodall, J., 2008. Do parents know they matter? Engaging all parents in learning. Educational Research 50 (3), 277–289.

Hartas, D., 2008. Practices of parental participation: a case study. Educational Psychology in Practice: Theory, Research and Practice in Educational Psychology 24 (2), 139–153.

Henderson, A.T., Mapp, K.L., 2002. A New Wave of Evidence: The Impact of School, Family and Community Connections on Student Achievement. National Center for Family and Community Connections with Schools, Austin, TX.

Hill, N.B., Tyson, D.F., 2009. Parental involvement in middle school: a meta-analytic assessment of the strategies that promote achievement. Developmental Psychology 45 (3), 740–763.

Hollingworth, S., Allen, K., Kuyok, K.A., Mansaray, A., Rose, A., 2009. Parental Engagement, Learning in Families and Technology: Groundwork Literature Review. BECTA.

Jeynes, W., 2012. A meta-analysis of the efficacy of different types of parental involvement programs for urban students. Urban Education 47 (4), 706–742.

Kim, Y., 2009. Minority parental involvement and school barriers: moving the focus away from deficiencies of parents. Educational Research Review 4 (2), 80–102.

Kumpfer, K.L., Alvarado, R., 2003. Family-strengthening approaches for the prevention of youth problem behaviors. American Psychologist 457–465.

Lexmond, J., Bazalgette, L., Margo, J., 2010. It Is Time to Be Honest about What Good Parenting Involves…. The Home Front, Demos.

Lindsay, G., Davis, H., Strand, S., Cullen, M.A., Band, S., Cullen, S., Davis, L., Hasluck, C., Evans, R., Stewart-Brown, S., 2009. Parent Support Advisor Pilot Evaluation Final Report. DCSF, London.

Lopez, G.R., Scribner, J.D., Mahitivanichcha, K., 2001. Redefining parental involvement: lessons from high-performing migrant-impacted schools. American Educational Research Journal 38 (2), 253–288.

Manz, P., Hughes, C., Barnabas, E., Bracaliello, C., Ginsburg-Block, M., 2010. A descriptive review and meta-analysis of family-based emergent literacy interventions: to what extent is the research applicable to low-income, ethnic-minority or linguistically-diverse young children? Early Childhood Research Quarterly 25 (4), 409–431.

Martinez, Y.G., Velazquez, J.A., 2000. Involving Migrant Families in Education. In: Eric Digest, vol. 4. ERIC Clearinghouse on Rural Education and Small Schools, Charleston, WV.

McDonald, L., Moberg, D., Brown, R., Rodriguez, I., 2006. After-school multifamily groups: a randomized controlled trial involving low-income, urban, Latino children. Children & Schools 28 (1), 25–34.

Norwood, P.M., Atkinson, S.E., Tellez, K., Saldana, D.C., 1997. Contextualizing parent education programs in urban schools: the impact on minority parents and students. Urban Education 32 (3), 411–432.

Peters, M., Seeds, K., Goldstein, A., Coleman, N., 2007. Parental Involvement in Children's Education 2007. Department for Children, Schools and Families, London.

Powell, D.R., 1989. Families and Early Childhood Programs. In: Research Monographs of the National Association for the Education of Young Children, vol. 3. ERIC.

Ralph, A., Sanders, M.R., 2003. Preliminary evaluation of the group teen triple P program for parents of teenagers making the transition to high school. Australian e-Journal for the Advancement of Mental Health 2 (3), 1–10.

Redding, S., Langdon, J., Meyer, J., 2004. The Effects of Comprehensive Parent Engagement on Student Learning Outcomes. American Educational Research Association, San Diego.

Reid, M., Webster-Stratton, C., Hammond, M., 2007. Enhancing a classroom social competence and problem-solving curriculum by offering parent training to families of moderate-to high-risk elementary school children. Journal of Clinical Child & Adolescent Psychology 36 (4), 605–620.

Rogers, L., Hallam, S., Long, M., 2010. Evaluation of the Impact of 'Unlocking Potential' Pilot on Children, Young People and Families. Institute of Education, University of London, London.

Sénéchal, M., Young, L., 2008. The effect of family literacy interventions on children's acquisition of reading from kindergarten to grade 3: a meta-analytic review. Review of Educational Research 78 (4), 880.

Smith, F.M., Hausafus, C.O., 1998. Relationship of family support and ethnic minority students' achievement in science and mathematics. Science Education 82, 111–125.

Statham, J., Harris, A., Glenn, M., Morris, M., Marshall, H., Bergeron, C., White, K., 2010. Strengthening family wellbeing and community cohesion through the role of schools and extended services. Centre for Excellence and Outcomes in Children and Young People's Services (C4EO).

Sylva, K., Melhuish, E., Sammons, P., Siraj-Blatchford, I., Taggart, B., Elliot, K., 2003. The Effective Provision of Pre-school Education (Eppe) Project: Findings from the Pre-school Period. Institute of Education, University of London, London.

Sylva, K., Melhuish, E.C., Sammons, P., Siraj, B.I., Taggart, B., 2008a. Final Report from the Primary Phase: Pre-school, School and Family Influences on Children's Development during Key Stage 2 (Age 7–11), p. 165.

Sylva, K., Scott, S., Totsika, V., Ereky, S.K., Crook, C., 2008b. Training parents to help their children read: a randomized controlled trial. British Journal of Educational Psychology 78 (3), 435–455.

Toomey, D., 1993. Parents hearing their children read: a review. Rethinking the lessons of the Haringey Project. Educational Research 35, 223–236.

Turney, K., Kao, G., 2009. Barriers to school involvement: are immigrant parents disadvantaged? The Journal of Educational Research 102 (4), 257–271.

Xu, M., Benson, S.N.K., Mudrey-Camino, R., Steiner, R.P., 2010. The relationship between parental involvement, self-regulated learning, and reading achievement of fifth graders: a path analysis using the ECLS-K database. Social Psychology of Education 13 (2), 237–269.

Pareto, Vilfredo (1848–1923)

Gottfried Eisermann, Bonn, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 11048–11051, © 2001, Elsevier Ltd.

Abstract

Fritz Wilfried Pareto was born on 15 July 1848 in Paris, France, the son of an Italian political émigré. Due to an amnesty granted by the House of Savoy, the Pareto family was able to move back to Italy. Fritz Wilfried's given name was italianized and shortened to Vilfredo. He enrolled at Torino's *Scuola di Applicazione per Ingeneri*, where, in 1870, he received his doctorate. In 1882, Pareto, retired from his career as a manager with the *Societá dell'Industria di Ferro*, resolved to dedicate his remaining life to his scientific pursuits. Italy's leading economist at the time, Maffeo Pantaleoni, lent his support when Pareto was named to the Lausanne University chair held previously by Léon Walras. As an academic teacher, Pareto had considerable success. His busy teaching schedule did not keep him, however, from publishing his *Cours d'économie politique*, a masterpiece of neoclassical economic theory. In *Les systèmes socialistes*, Pareto concluded that, theoretically, a socialist economy might be able to solve its problems as well as any nonsocialist economy. In his *Manuale dieconomia politica*, Pareto presented a reworked version of his theory of economics with special emphasis on the various states of economic equilibrium. In the two volumes of his *Trattato di sociologia generale*, Pareto undertook the immense task of integrating his economic and social analyses: 'Economics is but a small part of sociology, and pure economics is but a small part of economics'. Proud of his anticipatory powers, he foresaw the catastrophic repercussions of the Treaty of Versailles – for purely economic reasons. As early as 1917, he had asked, 'whether the Pax americana will turn out to be as productive as the onetime Pax romana?' Pareto died on 23 August 1923. He was buried in Céligny, his last place of residence, in a modest grave listing only the years of his birth and death, and his *nom de guerre*.

Life

Fritz Wilfried Pareto was born on 15 July 1848 in Paris, France, the son of an Italian political èmigrè, Raffaele Pareto (1812–82). Raffaele's father, Raffaele Benedetto Pareto (1768–1831), had been made a marchese di Parigi – or, as some have claimed, even a pair de France – by Napoleon, in recognition of his collaboration in erecting the Cisalpine Republic. The Pareto family was rooted in the trade bourgeoisie of Genoa, Italy, where it had been registered in the *Libro d'oro*.

In addition to her son, whose given name was Italianized and shortened to Vilfredo, Pareto's mother, Marie Mètenier (1816–89), also had two daughters. Due to a political amnesty granted by the House of Savoy, probably as early as 1858, the Pareto family was eventually able to move back to Italy. Pareto's father started to work as a French instructor for Genoa's *Scuola di Marina*. Vilfredo attended the same school, the beginning of his Italian education. When, later, Pareto's father was offered a better position in Turin, Vilfredo followed him to the capital of the Kingdom of Savoy. Here, in 1867, Vilfredo received his *Licenza in scienze matematiche e fisiche*.

He then enrolled at the *Scuola di Applicazione per Ingeneri*, where, in 1870, he received his doctorate. Much later, Pareto still considered his dissertation – on 'Principi fondamentali della teoria di elasticità dei corpi solidi' – important enough to include it in his *Collected Essays on Economic Theory*. He began his professional career in Florence, working for the *Centrale della S. A. delle Strade Ferrate Romane*. From there, he went on to a management position with the *Societá dell'Industria di Ferro*, traveling widely in Germany, Austria, France, England, and Switzerland.

His career left him with a sound knowledge of practical business matters. In 1882, the year his father died, Pareto chose professional retirement and settled in Florence. At around this time, he made an unsuccessful attempt at running for political office on a platform promoting his ideal of a liberal market economy. Pareto became resolved to dedicate his remaining life to his scientific pursuits. He joined Italy's oldest learned society, the highly respected *R. Accademia Economico-Agraria*. He also became a member of a second society, *Adamo Smith*, where he gave several lectures on scholarly topics and economic policy, which he also published. Moving in the upper-class circles of Florence – which then exerted a leading cultural influence throughout Italy – Pareto made a number of influential friends: the Lord Mayor Peruzzi and his wife; the politician S. Sonnino; P. Villari, who wrote a famous biography of Machiavelli; Colli, the author of *Pinocchio;* the anthropologist Mantegazza; such renowned philologists as D. Comparetti, A. Franchetti, and A. Linaker, as well as several others. It was at this time that he married Alessandra (Dina) Bakounine (1860–1940), a Russian (no relation to the anarchist of the same name).

When Pareto met Italy's leading economist, Maffeo Pantaleoni, they soon became friends. Pantaleoni supported Pareto, as a successor had to be appointed for the Lausanne University chair held by Lèon Walras, who had had to vacate his position due to his worsening mental illness. As an academic teacher, Pareto had considerable success. However, he soon asked to be relieved of his teaching duties. An unexpected inheritance set him free to concentrate on his preference for research and scholarship. Pareto's *Feste Giubilari* was published in 1917 to great critical acclaim. While this represented a personal triumph for him, Pareto continued to view himself as a victim of Italy's politicians and the *Camorra universitaria italiana*. When the Fascists took over Italy's government, various honors were offered to

him; he either declined them explicitly or did not respond to them at all.

Only once did Pareto return to Italy and then only because he was forced to visit the Rome family record registration office in person. To get a divorce and be able to remarry, he needed his birth certificate. Dina, his first wife, had left him. In the summer of 1923, he married Jeanne (Jane) Régis (1879–1948), his second wife. He had met her through a newspaper advertisement in Paris. When Pareto died two months later, on 23 August 1923, he no longer possessed an Italian passport. To be able to marry Jane, he had been forced to turn in his passport at the Republic of Fiume. Conforming to his last wishes, he was buried in Céligny, his last place of residence, in a modest grave listing only the years of his birth and death and his *nom de guerre*.

Works

From his beginnings as a researcher, Pareto placed immense importance on the question of method. He had named his approach the logical-empirical method. In his *Manuale d'economia politica* (1994), first published in 1906, he devoted no less than 55 paragraphs to it. In his *Trattato di sociologia generale* (1983), that number rose to 144 paragraphs. 'My only intention,' Pareto wrote, 'is to describe facts and to draw logical conclusions from them' (Pareto, 1983: p. 55). Facts were to be weighted according to relevance, not quantity. Facts based on experience were considered valid only 'as far as experience does not contradict observation' (Pareto, 1983: p. 6).

Pareto considered this method applicable not only to the field of economics but to the social sciences as well and – implicitly or explicitly – to every field of scholarship (Pareto, 1994: Vol. 1, 14). However, every inductive approach will always contain an implicit assumption. In the sciences, this would be the supposedly homogeneous character underlying all of nature. Thus, Pareto's 'survey' is strongly similar to J.A. Schumpeter's concept of 'vision.' According to Pareto, it is quite feasible to differentiate between economic and social phenomena initially. To regard them separately does not violate the dictates of empiricism. But 'both parts then have to be integrated with each other. Only a full view of the whole thing possesses any empirical reality' (Pareto, 1994).

At the heart of Pareto's *Cours d' économie politique* (1896–97, 2 Vols.) lies his theory of economic equilibrium. He refers to two sets of mathematical equations to describe two 'economic systems.' One system applies to barter and exchange, the other to production. If combined, both sets of equations attain validity only in a state of equilibrium (Pareto, 1994, App. 131–133). 'The most interesting case,' according to Pareto, is described as a situation where the number of equations is less than the unknowns contained in the equations: 'The system is then flexible and can balance itself' (Pareto, 1994, App. 131–133). 'Mathematics,' Pareto insisted, 'should be used only if it allows for an expression of certain relations between facts which otherwise could not have been expressed or which could not have been formulated using standard language' (Pareto, 559, n. 4, cf Pareto, 1983, 1732).

To describe different states of economic equilibrium, Pareto coined two terms: ophelemité and utilité. When Irving Fisher wrote a review of Pareto's work, Pareto tried in vain to discourage Fisher from confusing his terms with the terminology of Walras and Edgeworth. Fisher, Pareto argued, did not understand enough French and even misunderstood some of Pareto's mathematics (Pareto, 1960: Vol. 3, p. 408).

In 1902–03, Pareto published the two volumes of *Les systèmes socialistes*, an incisive analysis of historical and present-day socialist theories, including a discussion of Werner Sombart (he had treated Karl Blücher in a similar way in his earlier *Cours*). He paid due reverence to Karl Marx as a sociologist. Economically, however, he did not give Marx much credit. Theoretically, a socialist economy might be able to solve its problems as well as any nonsocialist economy. Earlier, Pareto had already integrated socialist society and collective capital with his overall theory of economics (Pareto, Vol. 2, pp. 714–724). Indeed, his followers, E. Barone and O. Lange, later based their own socialist theories of economics on this insight. Pareto argued against socialism not on economical, but on social and anthropological grounds. He concluded that all socialist systems were eventually based on untenable social and anthropological contradictions.

In his *Manuale d'economia politica* (1994), Pareto presented a reworked version of his theory of economics with special emphasis on the various states of economic equilibrium. Typically, he differentiated between them in his famous dictum:

> We can say that the members of a community, in a certain situation, enjoy the optimum of ophelemit if it has become impossible to distance oneself from this situation even to a small degree without benefit or damage to all members of this community; even the smallest step then necessary leads to profit for some members of the community or loss for others. (Pareto, 1994: p. 33)

This concept has been named the Pareto optimum. Later, it served as the foundation for welfare economics. It led to several attempts at an improved theory of income distribution.

Pareto always insisted that to understand his *Manuale*, one needed to know his *Cours*. In a typical misreading, J.R. Hicks complained that in the *Manuale* – by now a classic on the theory of consumption – the problems of capital and interest were hardly noted (Hicks, 1950), ignoring their extensive treatment in the *Cours*.

It was Pareto's intention to show 'by examples rather than declarations to prove the relation between economic phenomena and other phenomena of society' (Pareto, 1994, Foreword). Thus, it becomes clear how deplorable was the omission, in the French edition of the *Manuale* (Pareto, 1981) of Pareto's Foreword or *Proemio*. Still, improvements had been made to other parts of the French edition, including Pareto's mathematics. He was of the opinion that 'pure' theory was applicable only to the flow of goods and trade within a 'purely' competitive economy. Yet, the reality of monopolies, cartels, and trusts asked for a different description of modern markets and their mechanisms. Eventually, he discarded his entire theory of economic equilibrium as a 'dead end,' in spite of his groundbreaking anticipation of what later essentially came to be called the

theory of monopolistic competition (cf the French edition of the *Manuale*, Vol. 3, 40–48, 160–164, Vol. 5, 8–9, 71; App. 141). With reference to the effects of economical cycles, Pareto already differentiated between their objective and subjective sides. In this, he could claim no less a predecessor than Montesquieu.

Pareto's final great work, the *Trattato*, was preceded by the brief *Mythe vertuiste et la littérature immorale* (Pareto, 1911), an unforgiving settling of scores with any kind of literary censorship. Taken to its logical conclusion, Pareto argued, censorship would have to call for all of world-class literature to be censored – starting with the Bible.

In the two volumes of his *Trattato di sociologia generale* (1983), Pareto undertook the immense task of integrating his economic and social analyses. It had long been clear to Pareto that for any model of economic action fully to grasp reality, it had to encompass a theory of society. Next to those actions described as logical, recognizably illogical actions had to be equally accounted for. Such actions called illogical by Pareto are regularly cloaked by pseudo-logical explanations. On this insight, Pareto erected an entire sociology of knowledge (cf Eisermann, 1987, pp. 170ff.). Logical actions, to Pareto, were not synonymous with rational actions, however. For any act to be described as rational, it would have been necessary actually to look into the mind of the social actor. Pareto did not think such introspection was feasible. Rather, he conceptualized such actions as residual social facts. Thus, Pareto arrived at an inventory of six residues, in which instinct of combinations and persistence of aggregates were the most important ones.

He distinguished between two groups of economic actors: 'rentiers' and 'speculators.' Rentiers in general are secretive, cautious, timid souls, mistrustful of all adventure, not only of dangerous ventures but also of any venture to have even the remotest semblance of not being altogether safe (Pareto, 1983: p. 2313). Pareto's rentiers are ruled by aggregates of persistence and his speculators by their instinct of combination. Speculators

> are usually expansive personalities, ready to take up with anything new, eager for economic activity. They rejoice in dangerous economic ventures and are in the watch for them, but they work underground and know how to win and hold the substance of power, leaving the outward forms to others. (Pareto, 1983: p. 2313)

Accordingly, Pareto expanded his analysis of the economic equilibrium into an analysis of the social equilibrium (Pareto, 1983: pp. 2067–2087). The crucial movement inside the system of society is construed as *circolazione delle classe scelte*, class circulation. Society is then ruled by fractions, or minorities, which are subject to a continuous turnover, or exchange (2025–2046 and throughout).

Pareto always insisted that his 'pure' theory simply was an integral part of his overall economics, and that his economics necessarily formed part of his sociology. He claimed to have introduced the principle of relativity to the social sciences. He took equal pride in his anticipatory powers of foresight. Like his contemporary J.M. Keynes, he foresaw the catastrophic repercussions of the Treaty of Versailles – for purely economic reasons. As early as 1917, he asked, 'whether the Pax americana will turn out to be as productive as the onetime Pax romana?' (Pareto, 1960: Vol. 3, p. 240).

See also: Economic Sociology; Economics, History of; Elites: Sociological Aspects; Political Sociology; Rationality in Society; Sociology, History of; Sociology: Overview; Welfare.

Bibliography

Bobbio, N., 1969. Saggi sulla scienza politica in Italia. Laterza, Bari, Italy.

Busino, G., 1989. L'Italia di Vilfredo Pareto, vol. 2. Banca commerciale italiana, Milan.

Convegno Internazionale, V.P., 1975. Accademia Nazionale dei Lincei. Rome.

Demaria, G., 1949. L'oeuvre économique de Vilfredo Pareto. Revue de l'économique politique 59, 517–544.

Eisermann, G., 1961. Vilfredo Pareto als Nationalkonom und Soziologe. Mohr, Tübingen, Germany.

Eisermann, G., 1987. Vilfredo Pareto. Mohr, Tübingen, Germany.

Eisermann, G., 1989. Max Weber und Vilfredo Pareto, Dialog und Konfrontation. Mohr, Tübingen, Germany.

Eisermann, G., Malinvaud, E., 1992. Vilfredo Pareto und sein 'Manuale.' Verlag Wiltschaft und Finanzen, Düsseldorf, Germany.

Freund, J., 1974. Pareto. Seghers, Paris.

Hicks, 1950. Value and Capital, third ed. Oxford University Press.

Kirman, A.P., 1987. Pareto as economist. In: Eatwell, J., Milgate, M., Newman, P. (Eds.), 1987. The New Palgrave, vol. 3. Macmillan Press, London, pp. 804–809.

Pareto, V., 1911. Le mythe vertuiste et la littrature immorale. Riviére, Paris.

Pareto, V., 1920. I sistemi socialistici. Italy, Turin.

Pareto, V., 1952. Scritti teorici. Edizioni di storia e letteratura, Milan.

Pareto, V., 1960. Lettere a Maffeo Pantaleoni, vol. 3. Edizioni di storia e letteratura, Rome.

Pareto, V., 1963. The Mind and Society, second ed. Dover, New York.

Pareto, V., 1964–89. Oeuvres compltes, vol. 30. Librarie Droz, Geneva, Switzerland.

Pareto, V., 1965. Croncache italiane. Morcelliana, Brescia, Italy.

Pareto, V., 1966. Scritti sociologici. UTET Turin, Italy.

Pareto, V., 1981. Manuel d'économie politique. Librarie Droz, Geneva.

Pareto, V., 1920. I sistemi socialistici. Turin, Italy.

Pareto, V., 1983. Trattato di sociologia generale, fourth ed. AMS Press, New York.

Pareto, V., 1994. Manuale dieconomia politica, fourth edn. Tesi, Pordenone, Italy.

Parsons, T., 1968. Vilfredo Pareto. In: Sills, D.I. (Ed.), 1968. International Encyclopedia of the Social Sciences, vol. 11. Macmillan, New York, pp. 411–416.

Schneider, E., 1961. Vilfredo Pareto in the light of his letters to M. Pantaleoni, Banca Naz. Lavoro Quarterly Review 58.

Schumpeer, J., 1951. Vilfredo Pareto. In: Ten Great Economists. Oxford University Press, New York.

Valade, B., 1990. Pareto. Presses Universitaires de France, Paris.

Parietal Lobe

Claudia LR Gonzalez and Jason W Flindall, University of Lethbridge, Lethbridge, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.E. Goldberg, volume 16, pp. 11051–11054, © 2001, Elsevier Ltd.

Abstract

The parietal lobe is one of the four major divisions of the cerebral cortex. The function of the parietal lobe is to transform sensory input into motor output. This transformation is influenced by a host of cognitive computations processed by the parietal lobe including attention, spatial representation, number processing, and memory. Ultimately, the parietal lobe integrates sensory, motor, and cognitive information for the production of purposeful guided behavior.

Introduction

The parietal lobe constitutes approximately one-fifth of the cerebral cortex. Situated in a pivotal location, the parietal lobe receives projections from the frontal, occipital, and temporal lobes. Broadly speaking the parietal lobe has two main functions: somatosensation and sensorimotor integration. The parietal lobe can be considered the 'association cortex' as it integrates visual, auditory, and somatosensory information for the purpose of guiding behavior. An example that illustrates the integrative functions of the parietal lobe is a behavior that we perform multiple times on a daily basis: reaching and grasping for objects. Imagine, for example, reaching for and picking up a glass from the dinner table. This seemingly simple action requires the visual system to extract information about the location and the intrinsic properties (e.g., size, shape) of the glass, and then process this information in order to plan and execute the movement. Once grasped, sensory receptors in the hand assess properties of the glass such as weight, texture, and temperature. This information will in turn be used to modify the details of the grasp in order to successfully transport the glass to the mouth. In addition, the parietal lobe plays a key role in processing information about the spatial arrangement of the body with respect to the glass, and the location of the glass with respect to the other objects on the table. The parietal lobe is also essential in a number of cognitive functions such as attention, numerical processing, and working memory. Together with the frontal lobe, the parietal lobe demonstrates the greatest evolutionary expansion in terms of cortical size. Relative to body size, the human parietal cortex is 20 times larger than it is in other primates. This expansion has been attributed to the increased complexity of visuospatial integration, multimodal processing, and cognitive abilities supported by the human parietal lobe.

The bulk of our knowledge regarding the functions of the parietal lobe has come from the study of patients with lesions to this area of the brain. The neuropsychological lessons have highlighted how important the integrity of the parietal lobe is to the processing of sensory information, sensorimotor transformation, and spatial computations. More recently and with the advent of modern functional neuroimaging, mainly functional magnetic resonance imaging (fMRI), much more has been learned about functions of the parietal lobe particularly with respect to cognitive abilities such as attention, memory, and number processing.

The following sections will describe the anatomy and main functions of the parietal lobe. Particular emphasis is placed on the major deficits associated with damage to this part of the brain.

Gross Anatomy and Functional Divisions

The parietal lobe is located under the parietal bone, which forms the lateral wall of the skull just above the ear. The parietal lobe is superior and anterior to the occipital cortex, posterior to the frontal cortex, and superior with respect to the temporal lobe. The parietal lobe is bound by four prominent sulci (see Figure 1): anteriorly, the parietal lobe is separated from the frontal lobe by the central sulcus (cs); posteriorly, separated from the occipital lobe by the parieto-occipital sulcus; laterally, separated from the temporal lobe by the lateral sulcus (ls; with the supramarginal gyrus (SMG) extending upward into the parietal lobe from the dorsal edge of that sulcus); and medially, the longitudinal fissure divides the left and right parietal cortices. The parietal lobe can be subdivided into two functionally distinct zones: the primary and secondary somatosensory cortices (S1 and S2 respectively) in the anterior zone, which are the main areas responsible for processing the sense of touch, and the posterior parietal cortex (PPC) which is, broadly speaking, the area responsible for sensorimotor integration and control.

Primary and Secondary Somatosensory Cortex (S1, S2)

Anatomy

S1 and S2 are located on the anterior edge of the parietal lobe, bordering the primary motor cortex (M1; Brodmann's area (BA) 4) on the dorsal boundary of the frontal cortex. This region (anatomically referred to as the postcentral gyrus, PoG) is delimited anteriorly by the cs and posteriorly by the postcentral sulcus. It extends laterally from the longitudinal fissure to the ls (Sylvian fissure). S1 is located superficially, and is composed of BAs 3, 1, and 2. S2 is located on the superior bank on the ls. In addition to all three subdivisions of S1 being interconnected, S1 has connections to the frontal lobe, specifically both the M1 (BA 4) from BAs 3 and 2, and the

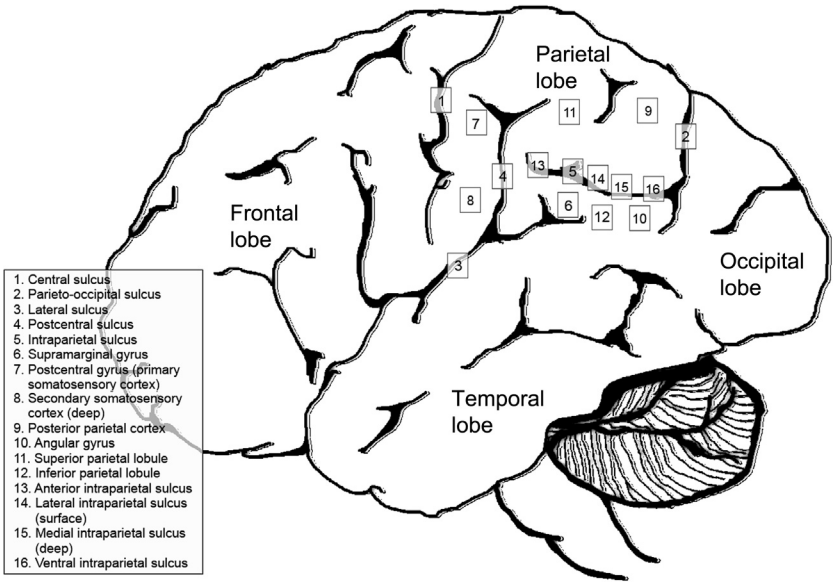

Figure 1 Schematic representation of human parietal lobe. Numbers represent main divisions.

premotor cortex (BA 6) from BA 2. S1 has projections to the PPC (BA 5) and S2 from all three BA subdivisions. Area S2 is densely interconnected with the contralateral S2, and also projects to area parietal ventral (adjacent anteriorly to S2), the PPC (BA 7b), the insular cortex, the amygdala, and the hippocampus.

Functions

S1 and S2 receive information from receptors in the skin (touch, pressure, vibration, temperature, and pain), and proprioceptive information from receptors in the muscles and joints. This information is utilized for detecting the presence and magnitude of sensory stimuli as well as its location on the body surface. S1 contains a tactile *homunculus* (Latin for 'little person') – a representation of the entire human body. This representation in the somatosensory cortex is not a true copy of the body; instead it is distorted. The largest proportion of S1 is dedicated to receiving input from areas on the body with highest receptor density and thus greatest sensitivity (i.e., the hands and face). Patients with damage to S1 or S2 have difficulty with tactile perception on the side of the body opposite to the lesion (i.e., the contralateral side). This deficit is more noticeable in the hands and face and could present as a loss of touch, temperature, pain, and/or proprioception. In addition, anosognosia, hemianesthesia, and astereognosis are three well-characterized disorders that arise from damage to S1 and/or S2.

Deficits Associated with S1/S2 Damage (Most Common)

Anosognosia

Perhaps the most extreme deficit that results from lesions to S1 and S2 is anosognosia (denial of motor deficits). This neurological condition is characterized by the patient's inability, or refusal, to acknowledge motor deficits on the contralesional side of the body. In extreme cases patients will deny the presence of a limb altogether, or report it as belonging to someone else. Anosognosic behavior may be exceptionally illogical in this regard, as the patient may admit the affected limb is attached to his/her shoulder, while simultaneously insisting it belongs to someone else. Anosognosia most frequently occurs following unilateral or bilateral lesions of S1 and S2, though it may be associated with frontal and/or temporal lesions, especially if the lesion-affected area encompasses both parietal and frontal structures. Though many regions have been reported to have a critical role in the cause of the disorder, damage to the thalamoparietal radiations within the right optic thalamus and/or damage to the right parietal cortex appears necessary to produce symptoms of anosognosia. Though anosognosia is most often diagnosed following right parietal lesions, this may be confounded by the fact that patients with left parietal lesions are often aphasic, and thus unable to communicate the symptoms of the disorder.

Hemianesthesia

Hemianesthesia describes a condition wherein patients lose complete pain, temperature, and/or tactile sensibility on one side of the body, while retaining motor ability. Hemianesthesia most often affects the left side, following damage to the right parietal cortex (primarily S1), though left parietal lesions may also result in right-side symptoms. As diagnosis usually follows a patient's failure to report tactile sensation on one side of the body, hemianesthesia diagnoses may be confounded in cases of unilateral neglect or hemi-inattention.

Patients with hemianesthesia may retain (or, more commonly, regain) motor ability in the affected limb despite persistent loss of somesthetic sensitivity. Interestingly, patients

may be able to localize tactile stimuli, despite being unable to accurately report whether or not a stimulus even occurred. For example, a patient may be unable to differentiate between hot or cold, rough or smooth, or heavy or light, but may be able to correctly report whether an 'unfelt' stimulus occurred on his/her fingers or palm, or his/her hand or forearm. This 'blind touch' phenomenon suggests an unconscious network processing somesthetic input – one that is spared following lesions to S1.

Astereognosis

Astereognosis is a disorder described by a loss of ability to discriminate between familiar objects based on touch alone. Despite tactile sensation being otherwise intact, patients with this disorder show a distinct loss of size and shape discrimination ability on the contralesional side. Astereognosis results after damage to the somatosensory hand area, located on the anterior middle third of the PoG (S1).

Posterior Parietal Cortex

Anatomy

The PPC is split longitudinally by the intraparietal sulcus (IPS), which divides the PPC into the superior parietal lobule (SPL; BAs 5, 7a and 7b) and the inferior parietal lobule (IPL; BAs 40 and 39). *Superior Parietal Lobule*: In the SPL, BA 5 is a sensorimotor region that receives input from M1, S1, and BA 7b ipsilaterally, and BAs 2, 5, and 7b contralaterally. It projects to the premotor, motor, and sensorimotor cortices, as well as the arcuate and principal sulci, cingulate, basal ganglia, superior temporal gyrus (STG), and claustrum. BA 7a is an area important for spatial perception; it is a visuomotor area, whereas BA 7b is somatosensory. Area 7a is connected to areas in the occipital (V6) and temporal cortices (in the parahippocampal cortex), as well as inferotemporal cortex (IT, area preoccipital inferotemporal cortex). Area 7a is also connected to other regions in the parietal lobe (both SPL, IPL, and all along the IPS), the frontal cortex (via the frontal and supplementary eye fields, and BAs 46 and 45), and the cingulate. Area 7b is also connected to visual areas (V6) and IT; BA 5, the medial dorsal parietal area (MDP), and BA 7a in the SPL; medial superior temporal area (MST) and superior temporal polysensory area (STP) in the temporal lobe; areas 45, 46, and 12 in the frontal cortex; and the cingulate and insular cortices. *Inferior parietal lobule*: IPL is anatomically and functionally divided into two gyri: the SMG and the angular gyrus (AG; anterior and posterior IPL, respectively). The AG (BA 39) is posterior to the SMG and superior to the temporal lobe. The SMG (BA 40) lies at the posterior end of the lateral fissure, continuous with the postcentral and superior temporal gyri. SMG has connections to the frontal and prefrontal cortices, the caudate nucleus, the parahippocampal gyrus, the hippocampus, the precuneus, the superior frontal gyrus, and the SMG.

Anteroposteriorly, the IPS divides the PPC into the anterior intraparietal (AIP), the medial intraparietal (MIP), the lateral intraparietal (LIP), and the ventral intraparietal (VIP) areas. AIP receives input from the caudal IPS, and projects to area F5 in the ventral premotor cortex, caudal (PMc). MIP, in the posterior half of the medial wall of the IPS, connects to other areas in the parietal lobe (including area 5, V6, V6a, and VIP) and projects to the premotor cortex. The LIP runs along the lateral wall of the IPS and is important for planning saccades. As such LIP is connected to many areas in the visual (occipital) cortex, somatosensory and sensorimotor (parietal) cortices, and the frontal eye fields and area 46 in the frontal cortex. It also has subcortical connections to the superior colliculus (SC) and pulvinar nucleus. Finally, VIP is roughly the middle third of the bottom trough ('fundus') of the IPS. It receives input from somatosensory (BAs 2, 5, and 7), visual (LIP, V6, and MST), and polysensory cortices (MIP, MDP, and area temporal – parietal – occipital area). VIP projects to area F4 in the PMc.

Functions

Generally speaking, the PPC is responsible for visually guided movements, as well as keeping track of 'mental space' for the processing of numerous spatial tasks. PPC neurons are very active when an action is performed toward an object, or even if attention is shifted toward that object. The neurons in the PPC are feature-sensitive; some react to size, others to shape, or orientation. These neurons are therefore thought to be responsible for determining appropriate hand posture for grasping. The PPC also directs saccadic eye movements, such that the fovea centers on an object of interest. PPC neurons are particularly sensitive to 'rewarding' stimuli; that is, stimuli that have behavioral relevance. PPC also contains neurons that are part of the mirror neuron system (*see* Premotor Cortex). Patients with PPC lesions are impaired at mental manipulation and mental rotation tasks, particularly after lesions to the PPC in the right hemisphere. PPC lesions can also result in impairments in the ability to perform mental arithmetic, syntax in language, movement sequence coordination, and/or recognition of familiar objects when viewed from novel angles. PPC lesions could also lead to Balint's syndrome. This syndrome consists of deficits in orienting the eyes to a certain location, spatial neglect, and optic ataxia (explained below).

Perhaps one of the most well-studied functions of the PPC is that of visually guided actions such as reaching and grasping. The dorsal visual stream (as described by Goodale and Milner) runs along the PPC. In their account of how the primate visual system is organized, Goodale and Milner proposed that the dorsal 'action' stream, which projects from early visual areas to the PPC, provides the flexible control required to act upon objects in the world. Neuropsychological and imaging research has supported this proposal. Using fMRI, research has shown that the IPS is activated during reaching movements, with stronger activation of the lateral and medial IPS during pointing movements, and the anterior IPS becoming activated when viewing or manipulating graspable objects, especially tools. Furthermore, recent fMRI studies have shown that regions within the PPC are selectively activated when planning reach-to-grasp movements. Specifically, these studies have revealed that the superior parieto-occipital cortex is involved in determining whether an object is within reach, whereas the anterior intraparietal sulcus becomes active when one is planning to grasp an object. Progress has been slower in mapping the visuomotor functions of the PPC and the parietal lobe in

general, compared to other areas because of the technical challenges of designing neuroimaging studies involving action. The next 10 years, however, should be filled with exciting discoveries as more ingenious protocols and sophisticated analysis tools become available.

Other functions of the PPC include attention (*see* Attention, Neural Basis of), language (*see* Speech Processing, The Cortical Organization of), spatial cognition (*see* Spatial Cognition), and number processing (*see* Numbers and Number Sense). Regarding PPC's role in attention, fMRI studies tell us that the IPS is activated when visual attention is directed toward a specific location, and the right temporoparietal junction becomes activated when visual targets are detected in unattended locations. The PPC is also known to be involved in processing the meaning and phonology of written words. Damage to the IPS, particularly the SMG, leads to phonological errors in naming (e.g., spaghetti named as 'pasghetti'). Additionally, PPC handles quasi-spatial processing, such as the spatial processing involved in mathematics and reading. Damage to this area can result in acalculia (inability to perform simple mental arithmetic), impaired cross-modal matching (being unable to recognize an object across two or more sensory modalities), right–left confusion, drawing disorders, and other spatial ability impairments. Damage to the PPC (especially the IPS) can also result in object manipulation deficits and apraxia. fMRI studies reveal that the human IPL (specifically the posterior part of the right-side SMG) contains a network of mirror neurons, homologous to that found in the macaque ventral premotor cortex and IPL. Finally, PPC (the right IPL) is also critical for self-recognition; if this area is stimulated by repetitive transcranial magnetic stimulation, a person's ability to recognize their own face from an array of other familiar faces is selectively disrupted.

Deficits Associated with Posterior Parietal Damage (Most Common)

Apraxia (Lesions Including but Not Limited to IPS)

Apraxia (*see* Apraxia) is a disorder of voluntary movement that cannot be explained by weakness, spasticity, rigidity, akinesia, sensory loss, or cognitive impairment including language comprehension. Patients with apraxia are unable to carry out learned, skilled motor acts despite preserved motor and sensory systems, coordination, comprehension, and cooperation. Although apraxic patients understand the meaning of gestures and objects, they often display difficulties imitating these same gestures, performing meaningful gestures on command, and using tools and objects. Apraxia is seen more frequently after left hemisphere damage, but unlike other parietal lobe disorders (e.g., hemispatial neglect), unilateral lesions often affect both sides of the body. This suggests that the left PPC plays a special role in the planning and control of skilled voluntary movements.

Oculomotor Apraxia (Lesions Including but Not Limited to LIP)

The PPC, specifically the LIP, is crucial in the planning, generation, and control of saccadic eye movements (*see* Eye Movement, Control of (Oculomotor Control)). Bilateral damage to LIP leads to oculomotor apraxia: an inability to shift gaze to objects in the visual periphery. Patients with oculomotor apraxia present with random eye movements. Patients may present with the inability to maintain fixation on a stimulus or, conversely, they may display difficulties in moving gaze away from a specific focus of fixation.

Optic Ataxia (Lesions Including but Not Limited to SPL, IPS)

Patients with optic ataxia show great difficulty reaching and grasping for objects particularly when reaching in the periphery. This impairment is accompanied by inappropriate hand orientation and grip formation. Optic ataxia appears more severe after left hemisphere damage when compared to right hemisphere damage. Patients with left hemisphere damage show a 'hand effect' and have great difficulty reaching toward objects and shaping their grasp appropriately with their right hand anywhere in space. With their left hand, however, the deficit is apparent only in the right visual field (the field contralateral to the lesion). Patients with right hemisphere damage do not show a hand effect and their deficit in visuomotor control is limited to the field contralateral to their lesion. That is, they can acquire objects successfully when reaching with either hand into the right visual field, but show deficits with both hands when reaching for objects in the left visual field. This finding suggests, again, that the left PPC plays a special role in visually guided actions.

Hemispatial Neglect (Lesions Including but Not Limited to IPS, AG, STG)

Hemispatial neglect (*see* Hemispatial Neglect, Neural Basis of) is characterized by the inability to orient toward and attend to stimuli in contralesional space. Neglect is most frequently seen after unilateral brain injury, especially the right hemisphere. Patients with neglect often present with loss of awareness in the left side of space (or right side for patients with left hemisphere lesions), responding slowly or completely ignoring visual, auditory, and/or somatosensory information from that hemifield. In extreme examples of neglect a person may not be aware of the deficit, so they might not eat food on the left side of their plate or they may ignore the left side of their body. For example, they may not shave or apply makeup to the left side of their face. Because patients with neglect often present without any visual field deficits, hemispatial neglect has long been considered an attentional problem, rather than a sensory one. In fact, there is evidence that posterior parietal neurons participate in the control of attention. Imagine the earlier scenario of reaching and grasping for a glass at the dinner table; visual awareness will be necessary to selectively direct one's attention to the target (i.e., glass) and not to the other many objects on the table.

Acalculia (Lesions Including but Not Limited to the Left IPS, AG)

Patients with acalculia (acquired calculation deficit) demonstrate difficulties comprehending and producing Arabic numerals (*see* Numbers and Number Sense). It is characterized by more severe deficits when reading three-digit numbers as

opposed to one- or two-digit numbers. Because acalculia can present with the inability to recognize one's own fingers (finger agnosia), it has been speculated that acalculia may originate from impairments in relating numbers and fingers by using finger-counting strategies.

Working Memory Deficits (Lesions Including but Not Limited to SPL)

Evidence is accumulating that the PPC plays an important role in working memory (*see* Working Memory, Neural Basis of), particularly in object working memory. Patients with lesions to the right PPC display impairments in remembering sequentially presented objects or object locations even when a few objects are used. Deficits in working memory have also been documented for both auditory – verbal and visual – spatial stimuli, especially in the mental manipulation of such information in working memory.

Conclusion

In summary, the primary role of the parietal lobe is to transform sensory input into motor output. This transformation does not occur in a vacuum but rather within a rich context of cognitive computations including attention, spatial representation, number processing, and memory. Ultimately, the parietal lobe integrates this sensory, motor, and cognitive information for the production of purposefully guided behavior.

See also: Apraxia; Attention, Neural Basis of; Eye Movement, Control of: Oculomotor Control; Hemispatial Neglect, Neural Basis of; Numbers and Number Sense; Premotor Cortex; Spatial Cognition; Speech Processing, The Cortical Organization of; Working Memory, Neural Basis of.

Bibliography

Behrmann, et al., 2004. Parietal cortex and attention. Current Opinion in Neurobiology 14, 212–217.

Berryhill, M.E., Chein, J., Olson, I.R., 2011. At the intersection of attention and memory: the mechanistic role of the posterior parietal lobe in working memory. Neuropsychologia 49 (5), 1306–1315.

Capotosto, P., Tosoni, A., Spadone, S., Sestieri, C., Perrucci, M.G., Romani, G.L., Della Penna, S., Corbetta, M., 2013. Anatomical segregation of visual selection mechanisms in human parietal cortex. Journal of Neuroscience 33 (14), 6225–6229.

Carminiti, et al., 2010. Understanding the parietal lobe syndrome from a neurophysilogical and evolutionary perspective. European Journal of Neuroscience 31 (12), 2320–2340.

Culham, J.C., Valyear, K.F., 2006. Human parietal cortex in action. Current Opinion in Neurobiology 16 (2), 205–212.

Dick, A.S., Tremblay, P., 2012. Beyond the arcuate fasciculus: consensus and controversy in the connectional anatomy of language. Brain 135 (Pt 12), 3529–3550.

Freund, H.J., 2001. The parietal lobe as a sensorimotor interface: a perspective from clinical and neuroimaging data. Neuroimage 14 (1 Pt 2), S142–S146.

Goldberg, M.E., 2001. Parietal lobe. In: International Encyclopedia of Social and Behavioral Sciences, first ed. Elsevier Science Ltd.

Goldenberg, G., 2009. Apraxia and the parietal lobes. Neuropsychologia 47 (6), 1449–1459.

Goodale, M.A., Milner, A.D., 1992. Separate visual pathways for perception and action. Trends in Neurosciences 15, 20–25.

Hubbard, E.M., Piazza, M., Pinel, P., Dehaene, S., 2005. Interactions between number and space in parietal cortex. Nature Reviews Neuroscience 6 (6), 435–448.

Kaas, J.H., 2008. The evolution of the complex sensory and motor systems of the human brain. Brain Research Bulletin 75 (2–4), 384–390.

Koenigs, M., Barbey, A.K., Postle, B.R., Grafman, J., 2009. Superior parietal cortex is critical for the manipulation of information in working memory. Journal of Neuroscience 29 (47), 14980–14986.

Kolb, Whishaw, 2015. Fundamentals of Human Neuropsychology, seventh ed. Worth Publishers.

Mars, R.B., Jbabdi, S., Sallet, J., O'Reilly, J.X., Croxson, P.L., Olivier, E., Noonan, M.P., Bergmann, C., Mitchell, A.S., Baxter, M.G., Behrens, T.E., Johansen-Berg, H., Tomassini, V., Miller, K.L., Rushworth, M.F., 2011. Diffusion-weighted imaging tractography-based parcellation of the human parietal cortex and comparison with human and macaque resting-state functional connectivity. Journal of Neuroscience 31 (11), 4087–4100.

Orban, G.A., Claeys, K., Nelissen, K., Smans, R., Sunaert, S., Todd, J.T., Wardak, C., Durand, J.B., Vanduffel, W., 2006. Mapping the parietal cortex of human and non-human primates. Neuropsychologia 44 (13), 2647–2667.

Rafal, R.D., 2006. Oculomotor functions of the parietal lobe: effects of chronic lesions in humans. Cortex 42 (5), 730–739.

Sack, A.T., September 14, 2009. Parietal cortex and spatial cognition. Behavioural Brain Research 202 (2), 153–161.

Schwartz, M.F., Faseyitan, O., Kim, J., Coslett, H.B., 2012. The dorsal stream contribution to phonological retrieval in object naming. Brain 135 (Pt 12), 3799–3814.

Simon, O., Mangin, J.F., Cohen, L., Le Bihan, D., Dehaene, S., 2002. Topographical layout of hand, eye, calculation, and language-related areas in the human parietal lobe. Neuron 33 (3), 475–487.

Van Essen, D.C., Lewis, J.W., Drury, H.A., Hadjikhani, N., Tootell, R.B., Bakircioglu, M., Miller, M.I., 2001. Mapping visual cortex in monkeys and humans using surface-based atlases. Vision Research 41 (10–11), 1359–1378.

Parliamentary Government

Klaus von Beyme, University of Heidelberg, Heidelberg, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 11054–11058, © 2001, Elsevier Ltd.

Abstract

Parliamentary democracy is the product of the twentieth century in most countries. Before 1918 there were waves of democratization, which mostly did not exceed parliamentarization of constitutional regimes. Criteria for the notion of parliamentary democracy are developed and various types (including semi-presidential systems) are differentiated. Functions of parliaments in parliamentary systems and the policy performance of this type of government compared to others are evaluated.

Development of Parliamentary Regimes and Democracy

Parliamentary democracy is the product of the twentieth century and in many countries developed only after 1918. There were some attempts to create such a system during the European revolutions of 1848 without leading to consolidated systems. What Huntington has called the first long wave of democratization (1828–1926) in the light of research on parliamentary systems was only parliamentarization. Parliamentary government is characterized by a government *responsible to the majority in parliament*. Democracy was added to the parliamentary regimes only later. The minimal criterion of democracy is *universal suffrage*. In most countries with a regime of constitutional monarchy, such as Britain, Scandinavia, or the Benelux countries, democracy was introduced later than parliamentary regime. Only in some rather authoritarian monarchies was democratic universal suffrage introduced or preserved without granting a parliamentary regime with responsible government (France, 1851 in the Second Empire Germany 1871 in the Empire of Bismarck) (see Table 1).

Only in a few countries (France, Finland, Sweden) was the process of parliamentarization and democratization synchronized. A continuously democratic parliamentary regime existed only in France. Even Britain – the mother of 'parliamentary government' – in the nineteenth century could not be called a 'parliamentary democracy' (cf. Table 1). There was a divergence of *parliamentarization* (mostly implemented by liberal majorities which were reluctant to add universal suffrage to the system) and *democratization* (frequently introduced under the pressure of radicals and socialists). These groups were initially unwilling to accept the checks and balance of the liberal parliamentary regime and advocated a democratic regime with many features of plebiscitarian instead of representative democracy. Only when they came to power did they slowly accept the rules of the game which were mostly developed in former constitutional monarchies.

Democratization in some countries has ruined the parliamentary system (Italy, Austria, Germany, Spain) when counter-regime forces gained power. In other established parliamentary systems the regime in the 1930s came to the brink of a collapse (Belgium, France) under the pressure of right-wing extremism. Only after 1945 and the disastrous experiences with authoritarian or even totalitarian dictatorship, were parliamentary government and democracy combined in peaceful coexistence. Some countries (such as the Federal Republic of Germany) even then, however, showed a tendency to emphasize the legal state and *representative* parliamentary *government* much stronger than *democratic participation*. The theory of constitutional and representative government did not easily accept parliamentary democracy but interpreted the preponderance of the parliamentary majority in terms of a rather mechanistic system of checks and balances and 'mixed government.' Because responsible cabinet government developed first in Britain and Belgium, the jurists for quite a time used the term '*Coburg government style*.' Only since the 1830s was the term 'parliamentary government' accepted in Britain and France and the regime was recognized not only as a mode of governance but as a *regime sui generis* which made the old typologies of regime types according to the number of rulers rather obsolete.

Types and Criteria of Parliamentary Democracy

There are many types of parliamentary government, but most of them have developed common *institutional* criteria.

a. *Compatibility of parliamentary mandate and ministerial office* (exceptions are the French Fifth Republic, Luxembourg and The Netherlands) in order to guarantee close cooperation between the parliamentary majority and the executive.
b. Prime ministers are normally *members of parliament*.
c. The government has to have the confidence of the parliamentary majority. *Ministerial responsibility* is not defined in terms of legal responsibility as in dualistic constitutional monarchies. The government has to resign in the event of a vote of no-confidence unless it chooses to ask the head of the state to dissolve parliament in order to give the electorate the opportunity to resolve the conflict.
d. Parliament controls the *government by raising* questions, exercises the right to interpellate and set up committees of enquiry, which facilitates the decision about whether the strongest sanction or a vote of no-confidence should be used. This right is shared by dualistic presidential systems, apart from the possibility of toppling the head of government and his or her cabinet.

Table 1 Parliamentary democracy: the merger of parliamentarization and democratization

	Continuity of a system of estates transformed into a constitutional regime						*New regimes*		*Parliamentarization of neo-absolutist regimes*				
	UK	*Sweden*	*Norway*	*Finland*	*The Netherlands*	*Luxembourg*	*Belgium*	*France*	*Italy-Piedmont*	*Spain*	*Germany*	*Austria*	*Denmark*
General male suffrage	1918	1921	1898	1906	1917	1919	1919	1848 1875	1919	1869 1907	1871 1919	1907	1915
Time of creation of a government responsible to parliament	1688ff 1832–35 Final conflict	1917	1884	1906	1868	1868ff	1833ff	1821ff 1837 1848–51 1871ff	1860	1874ff	1918	1918	1901

e. Some analysts have postulated as an essential feature a formal vote of the government at its first meeting with parliament (investiture), as in the French system under the Third and Fourth Republic and in Italy, or even demanded the formal election of the prime minister, which was developed in the German Basic Law of 1949. Most parliamentary regimes do not accept this condition.
f. More widespread is acceptance that *parliament* should be *dissolved* if the prime minister has lost the confidence of the parliamentary majority. Sometimes this has not been instituted, for example in Norway and Israel. In other cases the legal possibility of dissolution has become inapplicable because it has been abused by unparliamentary presidents, as in the Third French Republic (when President MacMahon dissolved the Chamber of Deputies in 1877).

Besides the minimal institutional criteria, certain social structural features are essential for the consolidation of parliamentary government:

1. *Organized parties* to facilitate the building of parliamentary majorities.
2. Party-building to facilitate the development of *cabinet solidarity*.
3. Development of the *office of prime minister*. In early systems where the estates dominated, and even in certain absolutist monarchies, the power centers – the estates or the crown – impeded the activities of first minister. A certain *hierarchization of the ministerial council* also strengthens cabinet solidarity.
4. Development of a *political culture* favorable to appropriate parliamentary behavior and alternating government.

Not all these institutional criteria can be found in every type of parliamentary system. Therefore subtypes have been proposed. Most ideological is the distinction between *authentic British parliamentarism* and the allegedly *inauthentic parliamentary system of the French type* – close to revolutionary government. Latecomers to parliamentarization, such as Germany (1918) and Sweden (1917), have often denounced the dangers of the French type, especially the German school of Redslob.

Oddly enough, France itself has reacted strongly against the shortcomings of the parliamentary system of the Third and Fourth Republics. De Gaulle's constitutional ideas, close to those of Carré de Malberg and later developed by Michel Debré were first pronounced in 1946 in the famous speech of Bayeux but failed to obtain majority when the Fourth French Republic was developed in 1946/47. But De Gaulle got his revenge when the Fourth Republic collapsed and he was able to shape the institutions of the Fifth Republic according to his ideas. A new system was born that since Duverger has been called the *semipresidential system*.

In spite of its name this was clearly a subtype of parliamentary government though the double responsibility before parliament and the head of the state – which existed in constitutional monarchies such as that of Louis Philippe – re-emerged. But in the long run the dominance of parliament in executive-legislative relations is re-established. De Gaulle's new system was not new at all: it had developed as a Republican type of constitutional regime from 1848 (the Second French Republic, 1848–51) and re-emerged in the Weimar Republic (1919) and Finland. Semipresidential systems are not a mixed type, not even the 'three-quarter presidential system' that Yeltsintried to develop. The backlash of a humiliated legislature was felt in Russia in 1998 when Yeltsin was forced to renounce Chernomyrdin as prime minister and accept the parliamentary candidate Primakov.

Functions and Performance in Parliamentary Democracy

Parliaments were more powerful in times of parliamentary sovereignty without universal manhood suffrage than in modern parliamentary democracies. The main actor in parliamentary democracy is no longer parliament or government, but the two powers are linked together. This linkage has strengthened the prime minister in most parliamentary regimes, unless high party fragmentation and consociational practices of conflict resolution have weakened the head of the government. Parliamentary democracy seems to be misnamed because of much discussion on the 'decline of parliament.' Therefore, *cabinet government, prime ministerial government, or chancellor's democracy* became more popular to characterize the bigger parliamentary regimes.

Parliamentary systems have to perform various functions. There is no general decline of parliament but rather a change of the main functions of the system. *Parliamentary control* is weakened in systems with close relations of parliament and government, interlaced by party government. The *controlling functions* declined or shifted to other institutions from constitutional courts and courts of account to the mass media. *Representation* and *articulation of interests* in the systems has changed over time. The *legislative function* is still important because more social problems are regulated in modern democracies than in old constitutional regimes. The *recruiting function* of the system, however, is strengthened after democratization of the parliamentary regimes.

Parliamentarians have become less removed from their voters in terms of *social structure*. What has been called *virtual representation* in early parliaments for those not eligible to vote in *postparliamentarism* is true of the social ties between deputies and voters. But people have adapted to this development and demand less strongly to be represented by members of the same region, the same sex, the same profession, or the same social class. The influence of voters on the selection of representatives is still mediated by the parties and their formal members. Primary elections can undermine the selecting function of the parties among other disadvantages, such as a lower voter turnout in the general election. Modern democracies are not ruled according to President Lincoln's Gettysburg formula of government *of* the people, *by* the people, and *for* the people. The distance of the representatives is compensated for by increasing *responsiveness* and sensibility to the wishes of the voters. Government *for* the people in the modern democracy of fluctuation of mass public moods has increased rather than decreased.

Nonetheless the departliamentarization of democracy is lamented. This has a *domestic political* side. At the stage of policy formulation, the practice of expert committees working together with the government is growing stronger. In the decision-making arena, the government is the most important initiator of legislation. Informal initiatives behind the formal initiatives of advisory boards, constitutional courts, or international

organizations are further undermining the part played by parliamentarians in legislation. The penetration of interest groups into the work of ministers and parliament has shifted the preparatary work to a round tables of extraparliamentary committees and working teams. But parliament is not just a rubber stamp to ratify the decisions of networks outside parliament. Parliaments, formally the institutional seat of the people's sovereignty, represent the framework for the coordination of networks of actors from parliamentarians, party strategists, interest groups, ministry officials and in federate states the federate units. The cosy triangles of early American parliamentary studies, composed of parliamentarians, bureaucrats, and interest groups, have become uncosy squares, since party steering was rediscovered in the 'legislative Leviathan.' In federal states such as Germany, deparliamentarization at the Länder level has led to greater participation of the Länder in the national decision-making process via intergovernmental decision-making. This further actor has been added, making an uncosy pentagon. The Maastricht Treaty has increased the number of areas that are subject to EU rules and legislation. With growing competence at the EU level, lobbying at that level is likewise growing. Meanwhile, national parliaments – more than the executive – have been downgraded to the position of lobbyists. Paradoxically the *democratization* of the European Union has led to *deparliamentarization* at the national level. Until 1979, Members of the European Parliament were tied to the national parliaments, but since the introduction of direct elections this has no longer been the case. However, the European Parliament does not yet pose a real threat to the national parliaments as it can only draw up laws when asked to do so by the Commission. Formally, EU legal acts are passed by the Council of Ministers and not by the European Parliament.

There are three threats to parliamentary democracy:

1. In the Process of *regionalization*, regional assemblies will try to strengthen themselves via regional politics at the cost of national parliaments.
2. In the process of *Europeanization*, more and more domains of the national parliaments will be regulated by the EU. This process will be supported by the activities of the European Court and the gradual standardization of the national legal systems.
3. In the process of *globalization*, a consolation at the national level is the fact that worldwide organizations such as the World Trade Organisation and GATT are starting to retaliate to the EU for what it did to the national decision-makers because the room for maneuver of the European organizations is increasingly limited by the global level, although still strongly limited in some arenas.

See also: Democracy, History of; Democratic Theory; History: Representation Within and Functions of; Parliaments, History of; Political Representation.

Bibliography

Blondel, J., Müller-Rommel, F. (Eds.), 1988. Cabinets in Western Europe. Macmillan, Basingstoke, UK.

Budge, I., Keman, H.J., 1995. Parties and Democracies. Oxford University Press, Oxford, UK.

von Beyme, K., 2000. Parliamentary Democracy. Democratization, Destabilization, Reconsolidation 1789–1999. Macmillan, Basingstoke, UK.

Döring, H. (Ed.), 1995. Parliaments and Majority Rule in Western Europe. St. Martin's Press, New York.

Laver, M., Shepsle, K.A. (Eds.), 1994. Cabinet Ministers and Parliamentary Government. Cambridge University Press, Cambridge, UK.

Norton, P., 1993. Does Parliament Matter? Harvester Wheatsheaf, New York.

Schütt-Wetschky, E., 1984. Grundtypen Parlamentarischer Demokratie. K. Alber, Freiburg, Germany.

Shugart, M.S., Carey, J.M., 1992. Presidents and Assemblies. Cambridge University Press, Cambridge, UK.

Strom, K., 1990. Minority Government and Majority Rule. Cambridge University Press, Cambridge, UK.

Parliaments, History of

Estevao C de Rezende Martins, University of Brasilia, Brasilia, Brazil

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A Parliament is a central institution of the democratic state under the rule of law. A Parliament is a representative institution, usually elected by direct universal suffrage. Political scientists and historians study parliaments from closely related viewpoints. The functions of parliaments differed according to the system of government. Contemporary parliaments are classified as bicameral or unicameral, as 'active' or 'deliberative.' Modern parliaments stem from two roots: first, the corporative assemblies of the Estates that were summoned under the Ancien Regime in Europe, and the second, the broad-based popular representation originating from revolutionary movements. The role of parliaments evolved during the course of the nineteenth century from being initially consultative to being deliberative. Central to an understanding of the evolution of parliaments is the concept of popular sovereignty. There are five principal trends in the historical research on modern parliaments: recovery and systemization of sources; sociological analysis of the composition of parliamentary representation; a prosopographic examination of parliamentary speeches; analysis of the role and impact of parliamentary speeches; and analysis of the role of parliaments in public administration.

Nowadays, a Parliament is a central institution of the democratic state under the rule of law. The term Parliament comes from French and describes a forum in which social understandings are reached by verbal negotiation, through the act of speaking. A Parliament is a representative institution, usually elected by direct universal suffrage. It plays an indispensable role in the workings of the State, as an active participant in the coalescing of political will, in public management, and essentially through conduct of the legislative process and concentration of the ruling elite. The decisive function of policy legitimization performed by parliaments is central to representative democracy, which is the predominant system of political organization in contemporary states. Experience of direct democracy, which dispenses with the representative parliamentary system or acts in parallel to it, is generally limited to referenda, Switzerland being among the few countries that practice direct democracy on a regular basis. A Parliament is an assembly of public delegates who receive a mandate for a limited period, are voted to office through elections based upon a pluralistic political party system, and serve as a channel for the expression of the opinions of society.

The approaches of political scientists and historians to the study of parliaments are closely related. Most contemporary researchers apply the insights of both disciplines in their efforts to analyze and explain the institutional structures and workings of parliaments. Comparative studies of the trajectory of societies, formation of states, and composition of their social and economic groups are the components of research on representation, participation, and decision-making processes conducted through and by means of parliaments. The achievement that parliamentary institutions represent, in terms of the development of democracy within society throughout the nineteenth and twentieth centuries, provides the social sciences with ample scope for research. Political science, sociology, and history continually delve into the origins, composition, work methods, social relevance, political prominence, and social roles of parliaments.

Functions and Types

The functions of parliaments have differed according to the system of government. Various terms may designate a parliamentary institution according to the different systems, such as: parliament, court, assembly, council. Under a parliamentary system, executive power emanates from the legislative branch invested in parliament, and governments hold office and remain in power so long as they enjoy the confidence of the parliament. Under a presidential system, parliament serves as a counterpoint to the political legitimacy bestowed upon the president through direct presidential elections. Parliaments also exist in nondemocratic regimes, such as the fascist dictatorships and 'democratic peoples' republics' of the twentieth century. Under such systems they tend to be mere puppets or facades since they exercise no real power but rather serve as stamps to endorse the power of the government, devoid of the legitimacy bestowed by a pluralistic party system and in the absence of representative electoral processes.

Depending on the process leading to their formation, parliaments can be classified according to three criteria: composition, action, and structure. Composition is subject to two variants: the first of these relates to criteria based on noble birth or propertied status, whereas the second is based on universal suffrage. Functionally speaking, contemporary parliaments can also be classified as bicameral or unicameral, depending on the constitutional framework of the State, which may have opted for a single chamber or two chambers. In any case, the chamber (or at least one of the chambers) ought to be elected directly by the people in free elections.

In functional terms, parliaments may be classified as 'active' or 'deliberative.' Active parliaments are those that, within a parliamentary system, have the power to appoint and support the government, as is the case in most modern parliamentary democracies. The so-called deliberative parliaments are those that have no direct responsibility for the business of the government, and include the pro forma parliaments tolerated by authoritarian regimes. Under a presidential system, as

a counterpoint to the legitimacy of the 'working' Executive Branch, the parliament assumes a predominantly deliberative role. These characteristics apply equally to unicameral and bicameral parliaments, in both federative and centralized states.

It is the decisive role performed by parliaments in the process of forming governments and providing them with support that has given currency to the term 'Parliamentary System' to describe the form of government in which empowerment and support of the Executive Branch are dependent on a vote of confidence from the parliament. This definition applies to the form of government, a term that expresses the legitimacy that stems from the legislative branch in its executive form. Under a presidential system in which the Executive Branch is the predominant branch of the government, as is the case in the United States or in Brazil, the Head of State is voted to office by direct elections and consequently his legitimacy competes with that of the parliament. In both parliamentary and presidential systems, the parliament is said to embody the Legislative Branch and its political legitimacy stems from the ballot.

Origins

Modern parliaments stem from two roots. The first of these is the corporative assemblies of the Estates that were summoned under the *Ancien Regime* in Europe, and the second, the broad-based popular representation originating from revolutionary movements. The first model of corporative representation brought together the three Estates – namely, the nobility, the clergy, and the 'Third Estate' that evolved gradually into a form of parliamentary representation as sovereign rulers required assent to levy taxes and make war. This first version of the dependence of the sovereign ruler on an institution of a parliamentary nature did not lead directly to democracy or recognition of popular sovereignty. The origin of these lies in the second root of modern parliaments – the popular revolutionary root – from which stem: the origins of the process of definition of popular sovereignty, the concept of citizenship, the legitimacy of power based on representation and the delegation of sovereign power, the rule of impersonality of the law within the context of absolute civil equality among citizens, the extension of such rights to all adult citizens within society and adoption and consolidation of liberal democracy, the expansion of suffrage and political parties, the constitution of government and of its control (under both the parliamentary and presidential systems), and the expression of popular will through the ballot. Both roots reflect antiabsolutist sentiments and it is from them that a representative democratic model, with elections at predetermined intervals, developed.

Both these origins of parliamentary power are clearly preserved and apparent in the structure and functioning of the British Parliament. The House of Commons, based upon the weight it amassed in the Civil War of 1642–49 and the upon manifestation of its social power in the Glorious Revolution of 1688–89, assumed its role of expressing the will as a source of power detrimental to that of the Crown. The result of the revolution, which established the right of the parliament to close a new sovereign, is reflected in the jusnaturalist and contractual arguments presented in Locke's *Two Treaties on Civil Government* (1690) – one of the principal theoretical works in support of the legitimacy and controlling powers of parliament.

The House of Lords, with its elitist composition, preserves the corporativist tradition of the Estates. Modern French tradition, established with the Jeu de Paume Oath of 20 June 1789 at the advent of the French Revolution, signals a break with the system of Estates General of the *Ancien Regime* and use of the Parliament to designate the judicial and notarial functions, as it had since the twelfth century. The parliamentary assembly installed in the very earliest days of the revolution enshrined the principal of popular sovereignty. In both the British and the French cases, affirmation of the principle of popular sovereignty arose from revolutionary movements, whereby the political absolutism of European states was substituted by constitutionalism. The role performed by a representative system in the formation of the United States is unique. The affirmation of the will of the Americans against unconditional rule by the British Crown led to the establishment of a community-based version of the decision-making process, similar to the community practices of the Protestant congregations, which served as a collegiate political model for the provincial and continental congresses, in which no form of corporative representation was present.

Evolution

At the onset of the nineteenth century, there were three types of parliaments: mixed-popular and aristocratic-bicameral under the English model; the French unicameral, popular model; and the American bicameral, popular and federalist model. In Western Europe and Latin American republics, an era of constitutionalization began. Both monarchies and republics adopted political systems dependent on representative legitimacy. Under parliamentary systems (in which the Government is dependent on investiture by the Legislative Branch), and under presidential systems (in which the Executive is chosen by direct elections), parliaments assumed increasing importance in political life. Parliament became the forum for representation and expression of social and economic forces, where the repercussions of rights claimed took on resonance and the extension of political and civil rights was played out.

The role of parliaments evolved during the course of the nineteenth century from being initially consultative to being deliberative. The legitimacy of parliaments became independent of the legitimacy of the sovereign upon emerging from the electoral process. As a consequence of electoral and census systems, which, throughout the 1800s, maintained cultural and economic distinctions among citizens, despite their all being held (theoretically) equal before the law, parliaments maintained an indirect corporative dimension until the institution of universal suffrage. Such distinctions were based on the systematic exclusion of women and the preference for the rich, the wellborn, and the literate.

The itinerary of consolidation of the power of the Legislative Branch was implemented in different ways under the North American presidential, the British parliamentary, and the French and German parliamentary models. The United States

Congress reflects the federative structure of state and the fact that the United States spent the nineteenth century engaged in a process of expanding its frontiers and acquiring new territory. Furthermore, the political model adopted by the American Congress implies that each state of the federation, through actions taken by its own political spheres, has joined the Union. The principle of legally joining a federation of states, together with that of popular sovereignty, is enshrined in the bicameral parliamentary systems predominant in federative states (such as the United States, Canada, Australia, Brazil, India, Austria, Mexico, Argentina, Germany, Japan, and Indonesia). In centralized states, bicameralism is the result of the continuity of mixed forms of representation, with roots stemming back to the Estates and to popular sovereignty. Both the British and the French parliamentary systems provide examples of such evolution. In both cases, tradition has been partially preserved through the upper house (the House of Lords and the Senate), whereas, popular representation is ensured by the Lower House (the House of Commons and the National Assembly). This model was also adopted by the Scandinavian monarchies and also, at specific times, by such states as the German Empire (1871–1918) and the multinational Austrian-Hungarian Empire (1848–1918).

Parliaments emerged upon the European scene as a counterpoint to monarchic autocracy. In their contemporary form as 'parliaments of citizens,' they emerged from an extensive line of development going back to the principle of egalitarian representation first put forward in the eighteenth century. The parliamentary model of the late twentieth century is the fruit of the maturing of philosophical concepts with respect to the human rights and the separation of powers. As the twenty-first century evolves, there is widespread social awareness that there can be no exercise of power without political representation, and that this resides in the legislative bodies. The transformation of political regimes into full-fledged democracies has led to an expansion and diversification of the roles of parliaments. To the role of authorizing tax collection has been added deliberation of public budgets, supervision of government actions, and full participation in decisions regarding public policies.

The reform of electoral systems and the extension of suffrage have placed on the stage of public and social life, the professional politician, a party member who represents specific interests. The principal interest that became enshrined over the length of the nineteenth century – from the French Revolution to the end of World War I – was to force the acceptance of, respect for, and practice of popular sovereignty as the source of the legitimacy of all power.

Central to an understanding of the evolution of parliaments is the concept of popular sovereignty. It was the new element that entered the political scene in the eighteenth century, whereby monarchic authority – though still enjoying wide support – was no longer accepted as being absolute, nor independent of social support. It found expression in parliamentary democracy that essentially signifies a system in which the government is made up of a nationally elected assembly to which it is accountable. Acceptance of the concept was consolidated during the course of the nineteenth century when the ideal of democracy became associated with the parliamentary system of government.

'Active' parliaments – the importance of which has increased remarkably since 1945 and is still increasing at the onset of the twenty-first century – have taken on a direct responsibility for furthering the public interest. This is in line with a historic trend, initiated in the 1790s, of delegating sovereignty to elected representative bodies. In such parliaments, the house assumes not only the role of deliberating themes of interest to society but also of making decisions relating to public policies stemming from them. Under presidential systems, as is the case of the United States, the powers of the Executive are legitimized by popular vote, independently of parliament. Although this is true in terms of the origin of legitimacy (the ballot), it does not mean that the concrete exercise of power is free of controls. On the other hand, the extension of suffrage and the continuous rate of development of political participation since the 1830s have brought about changes in the exercise of power that have not excluded the presidential systems. Though it may even be possible to govern without parliament – against parliament it is not possible to govern. Though a strong presidential government might tend to ignore collective assemblies, nonetheless, the democratic framework of the state serves to counterbalance the relationship. Parliamentary democracy is an achievement of the Western European States and increasingly of a number of Latin American and former communist Eastern European States. It is also the result achieved by certain nations of the British Commonwealth, which, since becoming independent of Britain, have incorporated the parliamentary system of government. Such is the case in Canada, Australia, India, and New Zealand. Democratic and parliamentary systems are rarer and less stable in Africa, the Middle East, and the Far East.

The expansion of suffrage, resulting in mass democracy and recourse to regular popular participation (formalized through political parties or indirectly through lobbies or special interest groups), have caused parliaments to assume the roles of formulation and assessment of public policies (and of decision making under parliamentary systems of government), particularly in the financial area (through the preparation of the budget and the exercise of control over its execution). The rate at which parliamentary institutions have assumed a leading role in the political process varies from one region to another. In periods of economic expansion or of war, the active role of parliaments may go into retreat, as occurred in the *Belle Epoque* and during World Wars I and II. Under authoritarian regimes, of the left or of the right, it is not uncommon for parliaments to be merely irrelevant decorative trimmings to the institutional architecture (as was the case in Fascist Italy, Nazi Germany, in the communist countries, and under the military dictatorships of Latin America in the twentieth century). Beginning in the 1990s, successful experiences of systems of government centered upon parliaments have been adopted in various regions of the world. In Eastern European countries, the West European parliamentary model of political organization has been widely adopted, though it runs up against the tendency to fall back on a powerful Head of State with a presidential profile that is a legacy of the centralizing tendency of a communist past that lasted from 1945 to 1990. In Latin America, with the restoration of democracy that began in the 1980s, parliaments have taken on importance in a context clearly dominated by the presidential models of the North American type. In Asia, Japan,

and India, active parliaments have been the rule since the late 1940s. In Africa, although practically all countries have parliamentary institutions, the parliamentary model, whether active or deliberative, remains incipient and the role of parliaments largely irrelevant, in a context in which the growth of religious movements strongly affects the neutrality of the parliamentary institution.

Future Prospects

Issues discussed when examining the history of parliaments fall largely into three groups: the external environment; domestic characteristics; and the political characteristics of parliaments. The external environment includes efforts to articulate four trends: the constitutional structure of parliaments, the administrative structure, the (political or electoral) system upon which they are based, and the interest groups with which they must interact. Their domestic characteristics comprise five factors that are decisive for the process of constituting parliaments: their members (their social and professional background, political experience and aspirations, and their degree of commitment to the legislative branch); parliamentary parties; committees; and houses as operational institutions; relations between members of parliament and their constituents as a reflection of social, economic, and cultural structures. The characteristics of their politics open up the possibility of studying parliaments as forums where the repercussions of short-, medium-, and long-term structures are felt and the way in which they reflect the dynamics of political decisions.

There are five principal trends in the historical research on modern parliaments: recovery and systemization of sources; sociological analysis of the composition of parliamentary representation; a prosopographic examination of parliamentary speeches; analysis of the role and impact of parliamentary speeches; and analysis of the role of parliaments in public administration (the formation and control of governments, monitoring and assessment of public policies, and taking of measurements among the voters and public institutions). A great portion of these trends is typically European. In Great Britain and in Germany, for example, recovery of sources is an important approach when placing parliaments within a historical context. The same is true in Brazil, Canada, and Australia. In all these cases, this monumental work has been linked directly to the consolidation of their respective identities over time and to the legitimization of the democratic structures of the States. However, as also occurs in other walks of life, historical research trends evolve.

Not so long ago it was possible to write on the history of parliaments by concentrating solely upon formal texts. Causes and developments would be carefully taken into account, accounts taken down in writing in accordance with the level of detail required, and the consequences for political parties and other significant players in the political processes analyzed. In many of the older texts, an implicit (and sometimes explicit) premise recurring throughout the explanation of the causes and consequences of parliamentary life was that it was possible to chart their progress. History reveals the emergence of parliamentary democracy, as legislation expands the electorate (in accordance with the results of the census, especially in the nineteenth century), to a certain degree making the vote more equal and elections more 'honest.' It was perfectly possible to explore democracy through reference to certain key political dates in each country. Thus the British system came to be regarded as a paradigm by others, as it ensures mechanisms that guarantee the peaceful transfer of power.

Such emphases have been radically altered. First, historical research, in a more skeptical and less self-confident age, no longer seeks to chart the progress from a smaller to a larger state, not least because society at the end of the twentieth century has become much more ambivalent about the building of progress. Second, because much more emphasis had been devoted to continuity than to change (especially with respect to representative democracy). Third, research efforts (polling) at the constituency level have greatly increased. It is hardly surprising that such work produces revelations that do not support the progressive view. It could be argued that the nature of the political process has undergone changes for the worse, from the viewpoint of the lower classes. The way in which governments manipulate electoral laws and the rules defining constituencies and how majorities are defined (for example, proportional systems, simple majorities, or mixed systems) has become relevant research themes in the field of historical research, the results of which are only recently coming to light.

Furthermore, historians have begun to question the importance of the vote for participation in the political process. Indeed, a well-substantiated demand addressed to the administration is likely to be more effective than heated political speeches supported by a few votes in a legislative body. The parallel with nineteenth-century Brazilian politics (and the process of overthrowing dictatorships in the twentieth century, particularly in Latin America) are cases in point. As O'Gorman and Vernon showed for Great Britain, lack of voter participation has not been an obstacle to political participation. Members of Parliament must take a broad spectrum of interests into account if they wish to remain on good terms with their constituents. The right to vote does not necessarily bring with it the access to political power. The suffragettes of the early twentieth century produced brilliant political propaganda, showing that educated, intelligent women with an interest in public affairs were denied the access to vote, whereas ignorant, vulgar, and drunken men were free to participate in it. In reality, such campaigns had a reach far beyond the mere exercise of the ballot: they were an affirmation of the importance of women in a stable political system.

Interest in electoral policies has been reinforced by detailed studies of elections, supported by long-term statistics. Such statistics have used materials that historians immediately recognize as being prime raw material for such studies. Voter registration records prior to the introduction of the secret ballot in Great Britain in 1872, for example, show who voted for whom. Such surveys make it possible to monitor changes in voting habits and to identify which candidates or parties lost favor with different social groups. The 'social geography' of elections and the 'sociological premises' of a given society thus become important instruments for historical analysis. Historical research brings together both the quantitative and qualitative techniques of the social sciences to obtain relevant results for analysis of parliaments.

Political studies relating directly to social and occupational structures also touch upon two other lines of work carried out over the past 40 years. 'History from below' was in vogue in the 1960s and 1970s, when many researchers began to devote themselves increasingly to social structures. Attempts to discover the origins of the 'working class' and, somewhat later, of the 'middle class' involved deep social analyses. Historians more sensitive to issues of social class admit that political awareness plays a part in the issue. Indeed, class consciousness is related directly to the exercise and pursuit of power and in turn requires the existence of political structures that need to be changed. Nonetheless, in Europe, as of the early nineteenth century, debate on parliamentary reform became enmeshed in discussions on social class and on changes in the 'political culture.'

Another important intellectual development has been the emergence of the 'linguistic turn' that began to influence historians in the 1980s. Detailed studies of language as a vehicle for conveying ideas, conducted by historians, have been inspired by radically new ideas borrowed from literary criticism. The relevance, and even the validity, of such an approach has engendered much debate among historians versed in the empirical tradition. Many are reluctant to accept that language can build its own realities, rather than having been invented to serve economic, social, or cultural needs. Unquestionably, extreme versions of poststructuralism, which claim that a text has no validity except in its own context, undermine any type of historical investigation in that they do not accept the notion of relative importance and suggest that change cannot be explained by intentional 'privilege' attributed to particular parts of historical documents. The 'linguistic turn' causes historians to be even more strict in the control of their premises on parliamentary discourse as a historical source. Language may effectively be used to mold ideas. Studies on demagoguery, populism, and *caudillismo* and their influence upon the formation of parliaments in such countries as Brazil, Mexico, and Argentina, are pioneers in this field. The study of the language used by the Chartist reformers of parliamentary democracy in Great Britain, for example, shows that their real goal was political change.

A third theme relating to the expansion of the democratic process is treated under the term 'progress.' Research shows clearly that, when it comes to political change, progress is neither simple nor linear. Parliamentary history is related to, but not defined by, the history of democratization. 'Democratization' of itself is a highly problematic and contested concept. Nonvoters are not necessarily excluded from politics. No parliamentary reform ever 'grants power' to another social group – at any rate, not directly. Women do not all of a sudden hold power simply as a consequence of a scrap of paper called a 'law.' Current research trends show that the law is more likely to ratify a political order already in place than to introduce sweeping change. The development of supranational parliaments in the second half of the twentieth century (the European Union, Mercosur) provides an entirely new field, into which investigative research is just beginning.

Historians have become more involved with continuities. Much of the historical research carried out in the 1980s and 1990s was accompanied by warnings about the dangers of economic reductionism, or of any other type of reductionism. However, the determination of historians to attempt to see things 'within their context' has contributed to the clarification of a series of relevant conclusions on the subject of parliamentary reform. They are now more aware that neither the 'top-down' explanations based upon the higher policies of central parliaments, nor the 'bottom-up' approaches that concentrate on extraparliamentary forms of pressure, are of themselves sufficient. The emphasis has swung to the need to understand the intricate interrelations of the complex factors, which created the political culture of the twentieth century and, above all, in the sequel of three major events of contemporary times, the World War II, the fall of the Berlin Wall and the revival of religious militancy. The changing methods and interpretations nonetheless confirm the central role of parliaments in this culture.

See also: Citizenship, Historical Development of; Democracy, History of; Parliamentary Government; Political History: History of Politics; Political Parties, History of; Public Sphere: Nineteenth and Twentieth-Century History.

Bibliography

Adamietz, H., 1978. Parlamentsgeschichte. Hauschild, Bremen.
Agirreazkuenaga, J., Urquijo, M. (Eds.), 1999. Contributions to European Parliamentary History. Bizkaiko Batzar Nagusiak, Bilbao.
Bélanger, R., Jones, R., Vallières, M., 1994. Les Grands Débats Parlementaires. Presses Université Laval, Laval, QC.
Bisset, A., 1882. A Short History of English Parliament. Williams and Norgate, London.
Chacon, V., 1997. História Institucional Do Senado. Senado Federal, Brasília.
Copeland, G.W., Patterson, S.C. (Eds.), 1994. Parliaments in the Modern World: Changing Institutions. University of Michigan Press, Ann Arbor, MI.
Das Österreichische Parliament, 1984. Nationalrat, Vienna.
Franks, C.E., 1987. The Parliament of Canada. University of Toronto Press, Toronto, ON.
Fraser, A., Masson, R.H.P., Mitchell, P., 1995. Japan's Early Parliaments 1890–1905. Routledge, New York.
Goldsworthy, J., 2001. The Sovereignty of Parliament: History and Philosophy. Oxford, USA.
International Commission for the History of Representation and Parliamentary Institutions, 2000. Taylor and Francis, London.
Interparliamentary Union (IPU), 2011. World Directory of Parliaments. Interparliamentary Union, Geneva.
The Journal of Legislative Studies, 1995. Frank Cass, London.
Kluxen, K., 1983. Geschichte und Problematik des Parlamentarismus. Suhrkamp, Frankfurt am Main.
Kühne, T., 1998. Parlamentsgeschichte in Deutschland: Probleme, Erträge, Perspektiven. History of Parliamentarianism in Germany Geschichte and Gesellschaft 24 (2), 323–338.
Kurian, G.T. (Ed.), 1998. World Encyclopedia of Parliaments and Legislatures. Congressional Quarterly, Washington, DC.
Laundy, P., 1989. Parliaments in the Modern World. Gower, Dartmouth, UK.
Loewenberg, G., Patterson, S.C., 1988. Comparing Legislatures. University Press of American, Lanham, MD.
Olivier-Martin, F., 1997. L'absolutisme français; suivi de Les parlements contre l'absolutisme traditionnel. Librairie générale de droit et de jurisprudence, Paris.
Parliaments of the World: A Comparative Reference Compendium, 1986. IPU, Geneva.
Parlement[s]. Revue D'histoire Politique (since 2003). L'Harmattan, Paris.
Parliaments, Estates, and Representation (since 1980). Rouledge, London.
Ritter, G.A. (Ed.), 1974. Gesellschaft, Parlament und Regierung: zur Geschichte des Parlamentarismus in Deutschland. Kommission für Geschichte des Parlamentarismus und der politischen Parteien, Düsseldorf.
The History of Parliament, 1998. CD-Rom Edition: House of Commons. Cambridge, UK.
Rodolico, N., Sardo, G., Novacco, D., 1963–1984. Storia del Parlamento Italiano. S.F. Flaccovio, Palermo.

Parsons, Talcott (1902–79)

David Sciulli, Texas A&M University, College Station, TX, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 11063–11068, © 2001, Elsevier Ltd.

Abstract

Talcott Parsons' social theory rivals the 'classic' theories of Karl Marx, Emile Durkheim, and Max Weber in empirical suggestiveness and analytical rigor. His social theory – popularly called action theory, functionalism, and structural–functionalism – evolved through three periods or phases: voluntaristic action, pattern variables, and the AGIL schema. Throughout his career, Parsons endeavored to ground the social sciences against relativism by basing them on analytical distinctions rather than on empirical generalizations or ideal types. In the last two decades of his life, his publications were subjected to increasing criticism, much of it rhetorical and personal. By the late 1980s, a 'neofunctionalism' emerged in the United States and Germany to press for a reconsideration of Parsons' publications and for new empirical applications of his theory.

From the late 1940s to the mid-1960s, Talcott Parsons was the single most influential social theorist in the world. A developer of what is now popularly called 'action theory,' 'functionalism,' and 'structural–functionalism,' Parsons spent his entire career at Harvard University, which helped considerably in institutionalizing his ideas and also in providing him access to talented graduate students. By the 1950s, his publications became a major part of what was literally a 'common curriculum' for graduate students in all major sociology departments – as well as in many of the best political science departments. By being institutionalized in this way, Parsons' publications elevated the rigor of American graduate training in sociology in particular. More than anyone else, he also defined the 'classic' theoretical tradition for sociology as a discipline and then added a more abstract, arguably more conceptually sophisticated contemporary rival to it. With this, he raised the bar for social theory worldwide, from Germany, France, and Great Britain to Japan and even the former Soviet Union. A strong case can be made today that every major social theory since the mid-1960s has been developed in direct or indirect dialog with Parsons' functionalism.

Having become literally an icon in the discipline in the first three decades of his career, in the last two decades he attracted a considerable share of the iconoclasm that more generally characterized the student movement and academia of the late 1960s and early 1970s. Parsons was subject across two decades to far greater criticism than that directed to any other theorist of his generation or since. Given the tenor of the times, most criticisms were rhetorical, often personal. Today, they leave a strong impression that few critics had bothered to read his works with the care typically afforded to basic empirical studies, let alone with the rigor and dispassion that any complicated social theory demands. Yet, the cumulative effect of the broadsides he received in the late 1960s and early 1970s was to leave the collective memory of the discipline with an understanding of Parsonian functionalism that is difficult to reconcile with often-repeated, explicitly stated positions in his publications. What Parsons endeavored to accomplish in his day, providing the social sciences with a common language, and how he went about accomplishing it, developing a 'conceptual framework' comprising analytical distinctions, now eludes the discipline's collective memory.

Talcott Parsons was born on 13 December 1902 in Colorado Springs, the youngest of six children of Edward S. Parsons, a congregational minister and college administrator, and Mary Augusta Ingersoll Parsons, a housewife and active suffragist (on biographical details, I rely primarily on Nielsen, 1991). Both parents were New England WASPs who traced their ancestry to mid-eighteenth century settlers, Mary to theologian Jonathan Edwards. A graduate of Yale Divinity School, Edward Parsons holds fundamentalist views of Christianity but leftist views in politics, particularly regarding the issue of economic inequality. He is an active participant in the Social Gospel movement that supported organized labor. Both parents inculcate into Talcott and his older siblings the importance of finding a mission in life and then pursuing it assiduously. Many observers comment later on the mature Talcott's remarkable energy and perseverance – including on the final 2 days of his life in Munich in 1979. He published over 150 sole-authored articles and 14 books or collections.

In 1917, the Trustees of Colorado College force Edward to resign as Dean after he supports two women who charged the College's President with sexual harassment. The family then moves to New York City where Talcott enters Horace Mann High School, a laboratory school for boys operated by Teacher's College of Columbia University. Talcott earns an undergraduate degree at Amherst College (as had his father and two older brothers) from 1920 to 1924. At the time Amherst was also an experimental, progressive school. Parsons at first focuses on biology and chemistry but in his junior year becomes 'converted' to the social sciences under the influence of an unorthodox institutional economist, Walton Hamilton.

Parsons' graduate training was remarkably brief: one academic year (1924–25) at the London School of Economics (LSE) as a nondegree candidate, then another (1925–26) as a degree candidate at Heidelberg University. At LSE, Parsons was influenced by Morris Ginsberg, L.T. Hobhouse, and the methodological functionalism of anthropologist Bronislaw Malinowski – as well as by fellow-student E. E. Evans-Pritchard. Parsons also met his future wife and mother of three children,

Helen Banerott Walker, a Bryn Mawr College student studying banking. Her family of conservative white- and blue-collar workers consider Talcott a leftist, a 'pinky.' At Heidelberg, Parsons first learned of the recently deceased German social theorist Max Weber and discussed his work with his wife Marianne, his younger brother Alfred, and Alexander von Schelting. Taking coursework from Edgar Salin and Emil Lederer (on economics), Karl Mannheim (a friend of Georg Lukacs) and Karl Jaspers (on Kant), Parsons selected how sociology and economics differ in portraying 'capitalism' as his dissertation topic.

During his year at Heidelberg, Parsons was offered a 1-year teaching appointment at Amherst in the Economics Department which also allowed him to teach an independent course in sociology. In 1927, Parsons was offered an instructorship in economics at Harvard, thus beginning a life-long career there. Harvard's Department of Sociology was a relative latecomer in the discipline, beginning in the fall 1931 under the chairmanship of Pitrim Sorokin, a 42-year-old emigre from Russia recruited from University of Minnesota. Parsons, then 29, was appointed as an instructor. Disagreements and misunderstanding between the two theorists became common knowledge in the Department and across the University. Yet, throughout Parsons' career, both critics and proponents often commented on his low-key demeanor when grappling with ideas and interacting with colleagues and students. It is difficult to find accounts of him being particularly effusive, let alone dominating. A short, stocky, already balding young man who sports a mustache and always smokes cigarettes, critics also comment often on his charm when interacting with other major national and international academicians at Harvard, in the American government, and overseas.

During his earliest years at Harvard, Parsons was impressed with philosopher Alfred North Whitehead and his notion of the 'fallacy of misplaced concreteness.' Consistent with Kant, Whitehead held that scientists apprehend 'reality' or scientific 'truth' only through analytical distinctions, not more immediately – whether experientially or directly through empirical findings. A few years later in correspondence with phenomenologist Alfred Schutz from 1940 to 1941, Parsons defended the notion of 'analytical realism' – as opposed to ontological realism or empirical realism – against Schutz's position that social scientists can somehow gain more direct access to the 'lifeworld' of their subjects of study (Grathoff, 1978).

By contrast to Schutz and then also to ethnographers at University of Chicago and American empirical researchers more generally, Parsons endeavored across his 50-year career to identify the most irreducible analytical distinctions unique to sociology, the scientific study of social life. In pursuing this project, Parsons arrived at three successive 'theoretical syntheses' and along the way trains four remarkably talented and productive cohorts of graduate students. All three of Parsons' theories, which he called 'frameworks of concepts,' involve a 'functionalist' approach to the study of social life. This is an approach much maligned today, and yet a basic premise of functionalism is hardly controversial. Any social scientist who poses the following question is operating within a functionalist approach: What is the relationship, if any, between the substantive area of social life I am studying and the direction of social change? Here, then, is a common ground shared by critical theorists and Parsonian functionalists but not by more historicist researchers, those who study particular events or periods in isolation and are loathe to identify any direction of social change.

While exceedingly abstract and complex, there are two ways to grasp the significance of Parsons' publications across his career. In general, he endeavored to account for the place and purpose of norms in maintaining social order both historically and cross-nationally. More particularly, he endeavored to account for the place and larger significance of professions, a unique set of occupations, in contemporary societies. Professions are a pivotal subject for sociological inquiry because they are important nongovernmental bodies in all modern societies. Yet, their prestige or status in any society hinges as much on whether they are valued culturally, supported by generally recognized social norms, and centrally situated in a social structure as on whether practicing professionals compete effectively in self-regulating markets. In all modern societies, professionals are simultaneously driven by economic pressures to maximize profits, like other workers, and also by normative pressures to conform to certain extraeconomic standards of conduct that typically do not bind most other workers. Parsons was interested in accounting at a theoretical level for the rise, evolution, and institutionalization of these extraeconomic restraints on self-interested behavior, and then also to explore whether and how they contribute to social order.

In his earliest publications, from the 1920s to the mid-1940s, Parsons approached professions by first distinguishing their behavior from that of economic enterprises, on one side, and religious organizations, on the other. Corporate managers and shareholders tend to act in strictly utilitarian or market-mimicking ways as they endeavor to maximize either growth or profit, a quantitative end that can be recognized interpersonally. Clerics and religious believers, however, tend to exhibit ritualistic fidelity to norms as they – presumably – seek spiritual salvation. This is a qualitative end that is transcendental or metaphysical; as such, its attainment is a matter of faith, not something that can be recognized interpersonally. Professionals, Parsons held, also exhibit fidelity to norms, but more 'voluntaristically.' They do not act in strictly market-mimicking ways, but they also do not conform to norms ritualistically with any transcendental end in view. Rather, they exhibit fidelity to norms as a means to attain qualitative ends that are worldly or empirical, and thus capable of being recognized interpersonally – such as their patients' physical or mental health, or their clients' legal innocence, or scientific truth.

Parsons came to appreciate the significance of 'voluntaristic action' after a careful reading of two economic theorists, Alfred Marshall and Vilfredo Pareto, and two social theorists, Emile Durkheim and Max Weber. This reading forms the core of *The Structure of Social Action* (1937) which, along with Parsons's, 1930 translation of Weber's *The Protestant Ethic and the Spirit of Capitalism*, established Weber and Durkheim as two of the new discipline's 'classic' theorists. During the 1930s, Parsons trained a first cohort of graduate students in the classics and his own theory of 'voluntaristic action,' including: Robert Merton, Kingsley Davis, John and Mathilda

Riley, Edward Devereux, Marion Levy, Wilbert Moore, and Florence Kluckhohn (along with Edward Hartshorne who was killed by a sniper while on active service in postwar Germany).

In spring 1944, Parsons was promoted to full professor and the chairmanship. By January 1946, he transformed the Department of Sociology into the Department of Social Relations, with the explicit aim of linking structural and social psychological approaches in the scientific study of social life. In the same year, Parsons also began formal training in psychoanalysis as a Class C candidate at the Boston Psychoanalytic Institute. He had been lecturing informally on Freud since fall 1938.

As Parsons assumed his leadership role in Harvard's Department, his theory went through a brief mid-career change from the late 1940s to the mid-1950s as he considered findings from Robert Bales's study of small groups. Parsons began thinking about professions (and how to identify norms in social life more generally) in terms of six basic decisions that people performing any role or task either make explicitly or have imposed on them by cultural values, institutionalized norms, or social structures. For example, certain roles (including those performed by professionals) mandate treating others (clients or patients) in universalistic ways. By contrast, other roles (including those performed by parents) encourage more particularistic treatment of others (children). Parsons called these decisions and requirements 'pattern variables.' By his account, researchers can use six such variable-pairs (1) to distinguish professions from other occupations and (2) to identify changes in the behavior of professions both historically and cross-nationally. Parsons' major publications during this period are a collection, *Essays in Sociological Theory Pure and Applied* (1949), and three monographs: *The Social System* (1951), *Towards a General Theory of Action* (1951, with Shils), and *Working Papers in the Theory of Action* (1953, with Robert Bales and Edward Shils).

During this period, Parsons trained a second cohort of graduate students in the pattern variables, including: Bernard Barber, Harry Johnson, Robin Williams, Jr., Jesse Pitts, Harold Garfinkel, Francis Sutton, and Robert Bales. More generally, as part of the buoyancy and optimism that pervaded the US after the war, many sociology departments reevaluate their curricula. The discipline is open collectively to receiving a new theoretical synthesis, and Parsons' notion of pattern variables, while preliminary and as it turned out fleeting, is a prime candidate. Another is Robert Merton's notion of 'middle range theory,' an idea he first proposed at the annual meeting of the then American Sociological Society in 1947 in opposition to Parsons' 'general theory.'

Finally, beginning slowly in the mid-1950s and then with more rapid developments from the early 1960s to the end of his career, Parsons approached professions within what he called the 'AGIL schema.' This is Parsons' single most significant contribution to social theory. During his transition from the pattern variables to the AGIL schema, Parsons' publications, lectures, and presentations create an unusually kinetic intellectual excitement among collaborators and graduate students. Everyone working in his circle operated with a palpable sense that social theory is in the midst of a fundamental breakthrough, a sense that American sociology has not witnessed since.

This transition period to Parsons' mature social theory is marked in particular by *Economy and Society* (1956, with Neil Smelser), and Parsons' training of a third cohort of graduate students, including: Albert Cohen, Renee Fox, Joseph Berger, Norman Birnbaum, Neil Smelser, James Olds, Jackson Toby, and Miriam Johnson. In addition, many earlier graduate students whose training had been interrupted by wartime service returned to Harvard. Talented undergraduates who come into contact with Parsons during this period include Robert Bellah and Charles Tilly. Also in transition during this period is the leadership of sociology at Harvard. Parsons resigns as chair of the Department of Social Relations in 1956, and within the next 2 years learns that he has diabetes. Sorokin retires in 1959.

From 1956 to 1961, Lewis Coser and Dennis Wrong in the US, David Lockwood in Great Britain, and Ralf Dahrendorf in Germany challenged Parsons' functionalism in relatively measured tones, but with uneven rigor. They were also shooting at a moving target in that Parsons was developing his mature social theory. The high-point of scholarly criticism during this period comes with a collection of essays by philosophers and sociologists edited by Max Black, to which Parsons responded with a long concluding chapter. In April 1964, Parsons was a central figure at a famous session of the German Sociology Association in Heidelberg devoted to reconsidering Max Weber's contributions. He found himself at odds with Frankfurt School critical theorists, most notably the young Jürgen Habermas and the more established Max Horkheimer and Herbert Marcuse (see Stammer, 1965/1972 for a collection of these exchanges).

From this point forward, criticisms of Parsons become harsher, more personal, and ideological than analytical and scholarly. Parsons was attacked, particularly in the US, for: his religious background, his WASP ethnic and cultural heritage, his writing style, his amazing productivity, his putatively strategic rather than scholarly decision to make sociology as rigorous conceptually as economics, his putative mistranslation of Weber, the putative conservative nature of his own family life including his putative responsibility for his daughter's suicide (in June 1964, 2 weeks after the Heidelberg debate), his putative political conservatism, and his putative general obeisance to 'American capitalism' or 'the establishment.'

Parsons unveiled the AGIL schema explicitly in 1963, with major articles on the concepts of power and influence, and then in 1964, with his first statement about 'evolutionary universals' (see the Turner, 1999 collection for the 1964 article and references to the others). As fully developed, the AGIL schema isolates analytically those aspects of behavior in any role or position that contribute, respectively, to each of four general social 'functions': economic efficiency ('adaptation'), administrative effectiveness ('goal attainment'), fidelity to social norms ('integration'), and fidelity to cultural values or a society's most basic institutional arrangements ('latency'). Parsons also proposed that four 'media of interchange' circulate between these analytical subsystems – money, power, influence, and value commitments – thereby bringing a certain orderliness or predictability to the whole. He added notions of 'systems-theory,' 'pattern maintenance,' and 'hierarchy of control' to this basic four-function breakdown of social life.

In 1973, Parsons and Platt published what is arguably his most important single work since 1937, *The American University*. In this work, Parsons rethinks his approach to professions by incorporating two major points into his social theory. First, he proposes that professions and sites of professional practice (such as universities, hospitals, and research institutes) are organized in a 'collegial form,' not in a bureaucratic form or a democratic form. Second, he proposes that professions are distinguished from other occupations by their members' willingness to bear 'fiduciary responsibilities' to clients and the larger society. These two points provide the foundations for an account of the place and purpose of professions in contemporary societies that differs radically from any effort, whether by rational choice theorists or social critics on the left, to treat the professions analytically as interchangeable with other occupations or other corporate entities. By 1973, however, the collective memory of the discipline is so dominated by criticisms of the AGIL schema that this major work, arguably Parsons' single most important substantive contribution to social theory, goes largely unread by critics and proponents alike.

Through the 1960s and 1970s, Parsons trained a fourth and final cohort of graduate students in the AGIL schema and related theoretical developments, including: Victor Lidz, Jan Loubser, Leon Mayhew, Rainer Baum, Dean Gerstein, John Akula, and Willy DeCramer. Parsons also influences undergraduate Jeffrey Alexander. Only in the mid-1970s, a few years before his death, did Parsons begin methodically to respond to at least a few criticisms of his work, most notably the informed analyses of Bershady (1973) and Warner (1978). Bershady's *Ideology and Social Knowledge* in particular marks a major sea change in commentary on Parsons' works. Bershady is the first to draw attention to Parsons' effort to ground the social sciences against relativism with analytical concepts. More than anyone else Bershady appreciates that Parsons pitched his social theory at an unusually abstract level of analysis because he sought literally a common language for the social sciences. Prior to Bershady, both proponents and critics generally failed to convey the power and sweep of Parsons' social theory in form, its aspiration to unite the social sciences. Parsons' project was essentially to give structuralists and symbolic interactionists the means, the common language, by which to demonstrate to each other the central import of their respective findings.

Parsons demonstrated his capacity to 'translate' others' findings and ideas in his own scholarship. He coauthored, team-taught, or otherwise collaborated closely with a remarkable range of theorists and researchers across disciplines, including: Edward Shils, Joseph Schumpeter, Samuel Stouffer, Florence Kluckhohn, Robert F. Bales, James Olds, Morris Zelditch, Philip Slater, Neil Smelser, Lon Fuller, Kaspar Naegele, Jesse Pitts, Winston White, S.N. Eisenstadt, Robert Bellah, Victor Lidz, Gerald Platt, Edward Laumann, Erwin Scheuch, Andrew Effrat, Mark Gould, and Dean Gerstein. In addition, he engaged in private and public correspondence and debates with Frank Knight, Crane Brinton, Robert Bierstedt, Albert Blumenthal, Alfred Schutz, Chester Barnard, Philip Selznick, Eric Voegelin, Robert Merton, Kenneth Boulding, C. Wright Mills, Robert Bellah, Bennett Berger, David Riesman, Jurgen Habermas, Herbert Marcuse, Max Horkheimer, and Ernst Mayr. Finally, an impressive set of scholars were influenced heavily by Parsons himself, his students, or his theory (some of whom also influence Parsons and his students): Niklas Luhmann, Seymour Martin Lipset, Amitai Etzioni, Robert Marsh, Daniel Bell, Joseph Ben-David, Benjamin Nelson, Gabriel Almond, James S. Coleman, Karl Deutsch, David Apter, Lucian Pye, Sidney Verba, Chalmers Johnson, S.N. Eisenstadt, Ken'ichi Tominaga (Japan's leading postwar sociologist), Lawrence Brownstein, Martin Martel, Adrian Hayes, and Frank Lechner.

Parsons and Helen traveled frequently in the 1970s, including three separate trips to Japan. In early May 1979 they returned to Heidelberg, then proceeded to Munich, on the occasion of the fiftieth anniversary of Talcott's degree from Heidelberg. After Parsons' typical full day of presentations and scholarly exchanges, including a lecture on 'The Declining Significance of Social Class' attended by host Horst Helle as well as Jurgen Habermas, Niklas Luhmann, Richard Munch, and Wolfgang Schluchter, Parsons died of a stroke in the early hours of May 8. At a memorial service 10 days later in Harvard Chapel, Victor Lidz notes in eulogy: "No sociologist of recent times has had to endure more bitter criticism than Talcott. Although a passionate man, Talcott bore criticism with equanimity as well as courage. Not daunted by even sweeping attack, he held to his own program of research and writing. Moreover, adhering to principles of value-freedom, he declined to reciprocate ideological criticism. He dealt with criticisms, always, as technical matters within the domain of social scientific theory" (Nielsen, 1991).

Parsons' critics inadvertently encouraged the fragmentation of research and theory in sociology and the social sciences that Parsons endeavored to prevent. Today, in the absence of any shared framework of analytical concepts, social scientists routinely talk past each other, as opposed to revealing to each other the general importance of their research findings and theoretical developments. Parsons's 'general theory' has given way not to a finite set of readily identifiable theories of the 'middle range,' as Robert Merton anticipated in 1949, but rather to an ever more finely grained cacophony of research specialties. This trend is then recapitulated, and accelerated, by ongoing changes in graduate training. Graduate students are trained more and more narrowly, and their required survey courses in social theory hardly encourage them to move beyond any research specialty. If there is any 'common language' in the social sciences today, it is that of research methods. But even here there is more internal division among methodologists than they often wish to acknowledge.

In support of Habermas' point that we can expect any new theoretical development in the social sciences to respond directly or indirectly to Parsons' functionalism, French social theorist Pierre Bourdieu in many regards today recapitulates Parsons' AGIL schema. He identifies two poles within what he calls 'fields of power' – the economic field and the field of cultural production – and then distinguishes four fields in between: politics, the higher civil service, the professions, and the university (Bourdieu, 1989/1996). The major difference with Parsons is that the French theorist's categories are more directly descriptive, less analytically rigorous. As a result, it is not clear on what basis he is distinguishing fields, why there are six fields of power rather than four or eight, or on what basis

researchers may categorize particular groups or activities as components of one field or another.

By the early 1980s, a new generation of social theorists, including members of Parsons' last cohort of graduate and undergraduate students at Harvard, initiated the effort to have Parsons' publications reconsidered on scholarly grounds. Led initially by Munch (1981) in Germany and Alexander (1983, 1985) in the US, today's 'neofunctionalism' finds support from theorists whose political positions range from radical left to liberal reformist to conservative republican (see the collections edited by Hamilton, 1992; Colomy, 1990). What they share, with Parsons himself, are two general principles (which Alexander, 1988 now explicitly rejects). One is that it is vitally important to the social sciences to continue Parsons' effort to ground social theory on the basic analytical distinctions of the AGIL schema rather than to rely more directly on descriptive concepts (whether empirical generalizations or ideal types). The other principle is that it is incumbent today to demonstrate the empirical potential of Parsonian functionalism, to identify lines of empirical inquiry that this social theory uniquely presents to view (Sciulli, 1992). A strong case can be made that Parsons' social theory has been mined only superficially (Takagi, 1992). An equally strong case can be made that theorists and researchers today can draw fruitfully on Parsons' basic AGIL schema, and then the first set of subdivisions he drew within each 'function,' but that they need not follow Parsons' effort to draw further distinctions within each subdivision (Mouzelis, 1995).

Social theorists outside the US today who support the Parsonian tradition, even if critical of certain parts of it, include: Munch, Hans Joas, Uta Gerhardt, Horst Helle, and Harald Wenzel in Germany; Helmut Staubmann in Austria; Jens Kaalhauge Nielsen in Denmark; Pierpaolo Donati, Andrea Maccarini, Matteo Bortolini, and Riccardo Prandini in Italy; Ken'ichi Tominaga, Kiyomitsu Yui, and Kazuyoshi Takagi in Japan; and Nicos Mouzelis, Bryan Turner, and Jeremy Tanner in Great Britain.

Editor's Note

Parsons' work, theories, and influence continue to be discussed in contemporary social science discourse. He has been identified as one of the fifty most important social theorists in the history of sociology in Scott (2007). His continuing relevance to economic sociology is discussed in Tittenbrun (2014) and Swedberg (2014) discusses the relationship and mutual influence between Parsons and his contemporaneous Harvard colleague, the economist Joseph Schumpeter. Other current commentators still turn to Parsons as a theorist of community (Lidz, 2014), a complexity theorist (Zwick, 2011), a sociologist of religion (Vanderstraeten, 2012), an analyst of the 'sick role' (Varul, 2010), and a great deal more. His works are still being reprinted decades after his death in 1979 including *The Social System*, which was reissued by Routledge in 2013. Book-length exegeses of his thinking and contributions to social theory continue to proliferate (e.g., Gerhardt, 2011).

See also: Action, Theories of Social; Control, Social; Conventions and Norms: Philosophical Aspects; Durkheim, Emile (1858–1917); Emergent Properties; Functionalism, History of; Institutions; Norms; Political Economy, History of; Professions in Organizations; Professions, Sociology of; Realism/Neorealism; Shared Belief: Philosophical Aspects; Social Influence, Sociology of; Social Structure; Social System; Sociological Theory; Sociology, History of; Status and Role: Structural Aspects; Symbolic Interaction: Methodology; Value Pluralism; Values, Sociology of; Weber, Max (1864–1920).

Bibliography

Alexander, J.C., 1983. Theoretical Logic in Sociology. In: The Reconstruction of Classical Thought: Talcott Parsons, vol. 4. University of California Press, Berkeley, CA.
Alexander, J.C. (Ed.), 1985. Neofunctionalism. Sage, Beverly Hills, CA.
Alexander, J.C., 1988. Neofunctionalism and After. Blackwell, Malden, MA.
Bershady, H.J., 1973. Ideology and Social Knowledge. Wiley, New York.
Black, M. (Ed.), 1961. The Social Theories of Talcott Parsons. Prentice-Hall, Englewood Cliffs, NJ.
Bourdieu, P., 1989/1996. The State Nobility: Elite Schools in the Field of Power. Stanford University Press, Stanford, CA.
Colomy, P. (Ed.), 1990. Neofunctionalist Sociology. Elgar, Aldershot, UK.
Grathoff, R. (Ed.), 1978. The Theory of Social Action: The Correspondence of Alfred Schutz and Talcott Parsons. Indiana University Press, Bloomington, IN.
Gerhardt, Uta, 2011. The Social Thought of Talcott Parsons: Methodology and American Ethos. Ashgate, Surrey, UK.
Habermas, J., 1981/1987. Theory of Communicative Action. In: Lifeworld and System: A Critique of Functionalist Reason, vol. 2. Beacon Press, Boston.
Hamilton, P. (Ed.), 1992. Talcott Parsons: Critical Assessments, 4 vols. Routledge, London.
Lidz, Victor, July 18, 2014. On the development of Talcott Parsons' conception of societal community. In: Paper Presented at the 18th ISA World Congress of Sociology, Yokohama, Japan.
Mouzelis, N., 1995. Sociological Theory: What Went Wrong? Diagnosis and Remedies. Routledge, London.
Munch, R., 1981. Talcott Parsons and the theory of action, I: the structure of the kantian core. American Journal of Sociology 86, 709–739.
Nielsen, J.K., 1991. Talcott Parsons' Life, Theory and Political Orientation, 5 vols. Unpublished Manuscript.
Parsons, T., 1937/1968. The Structure of Social Action, 2 vols. Free Press, New York.
Parsons, T., Platt, G.W., 1973. The American University. Harvard University Press, Cambridge, MA.
Sciulli, D., 1992. Theory of Societal Constitutionalism: Foundations of a Non-Marxist Critical Theory. Cambridge University Press, New York.
Scott, John (Ed.), 2007. Fifty Key Sociologists: The Contemporary Theorists. Routledge, London.
Stammer, O. (Ed.), 1965/1972. Max Weber and Sociology Today. Harper and Row, New York.
Swedberg, Richard, 2014. Schumpeter and Talcott Parsons. Journal of Evolutionary Economics (published online June 1, 2014).
Takagi, K., 1992. Talcott Parsons and American Intellectual Society. Cambridge University Press, Cambridge, UK.
Turner, B.S. (Ed.), 1999. The Talcott Parsons Reader. Blackwell, Malden, MA.
Tittenbrun, Jacek, 2014. Talcott Parsons' economic sociology. International Letters of Social and Humanistic Sciences 2, 20–40.
Varul, Matthias, 2010. Talcott Parsons, the sick role and chronic illness. Body and Society 16 (2), 72–94.
Vanderstraeten, Raf, 2012. Talcott Parsons and the enigma of secularization. European Journal of Social Theory (published online July 4, 2012).
Warner, R.S., 1978. Toward a redefinition of action theory: paying the cognitive element its due. American Journal of Sociology 83, 1317–1349.
Zwick, Martin, 2011. Complexity theory and political change: Talcott Parsons occupies Wall Street (Chapter 7). In: Youngman, Paul A., Hadzikadic, Mirsad (Eds.), Complexity and the Human Experience: Modeling Complexity in the Humanities and Social Sciences. CRC Press, Boca Raton, FL.

Partial Orders

Reinhard Suck, Universitat Osnabrueck, Osnabrueck, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 16, pp. 11068–11073, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

The article expounds the definition and basic properties of partially ordered sets. Related notions such as quasiorders, weak, and linear orders are introduced. Semiorders and interval orders are discussed with particular emphasis on their numerical representations. The order theoretic and algebraic properties of lattices are summarized. The question of the possibility of linear extensions of posets is answered by presenting Szpilrajn's theorem. This topic leads to the concept of order dimension. The discussion of the 'MacNeille completion' of an order ends the article.

The idea of ordering captures a basic faculty of the human mind: to decide for two given elements from a set of objects which one 'dominates' the other with respect to an attribute of interest, or, alternatively, which one is 'preferred' to the other.

Formally, a partially ordered set (a poset for short) consists of a pair $(P, \leq)$ where P is a nonempty set and $\leq$ a binary relation on P satisfying for all $x, y, z \in P$:

1. reflexivity: $x \leq x$,
2. asymmetry: $x \leq y$ and $y \leq x$ implies $x = y$,
3. transitivity: $x \leq y$ and $y \leq z$ implies $x \leq z$.

Transitivity is, in a sense, the most important property; it is shared by all variants of order relations, though it might be a consequence of other conditions. However, asymmetry also contributes in an essential way to the meaning of a partial order. Reflexivity, on the contrary, is merely a matter of taste or convenience. In some cases its opposite, irreflexivity, is postulated. Such an order is called *strict*. If the asymmetry condition is replaced by symmetry, i.e., for all $x, y \in Px \leq y$ implies $y \leq x$, then the resulting structure is an equivalence relation which lacks the interpretation of 'dominance' or 'preference.' Such a structure is more akin to classification or grouping with respect to similarity.

Related Notions and Diagrams

Sometimes asymmetry is simply dropped from the conditions. A relation satisfying reflexivity and transitivity is called a quasiorder. The investigation of quasiorders can be split into the respective properties of a partial order and an equivalence relation in the following way: Let $(Q, \lesssim)$ denote a quasiorder. Define $x \sim y$ if $x \lesssim y$ and $y \lesssim x$ both hold true. Then it is readily shown that $\sim$ is an equivalence relation on Q. Let P be the set of equivalence classes and define for $p, q \in P$ the relation $\leq$ by $p \leq q$ if $x \lesssim y$ for some $x \in p$ and $y \in q$. Then, one demonstrates two things: first, that the definition of $\leq$ is independent of the particular elements $x \in p$ and $y \in q$; and second, that $(P, \leq)$ is a poset. Thus, the 'dominance-aspect' of a quasiorder is contained in $\leq$, and its 'classification-aspect' is hidden in $\sim$.

One property of an intuitive meaning of dominance is not captured in the notion of a partial order: the idea that any two elements can be compared with respect to $\leq$. The defining conditions of a poset do not imply connectivity, i.e.,

$$x \leq y \text{ or } y \leq x \text{ for all } x, y \in P.$$

(Note: often this property is called 'completeness,' but later in this article another notion of completeness is introduced.) If connectivity is added to the properties of a partial order $(P, \leq)$, then $\leq$ is called a linear or total order. A quasiorder satisfying connectivity is called a weak order. Applying the above procedure on a weak order results in a linear order on the equivalence classes of the weak order.

There is a tendency to associate the property of connectivity automatically with an order given in some field of application, e.g., if one rank orders stimuli or subjects. Doing this amounts to representing the order numerically, i.e., to assuming a function

$$f : P \to \mathbb{R} \text{ such that } x \leq y \text{ if } (x) \leq f(y). \qquad [1]$$

If such a representation is possible, the order is necessarily a weak order. However, in social science, numerical representations are used widely even for orders which cannot easily be justified to be connected. For example, if the items of an achievement test are 'Rasch-scaled' (see the discussion following eqn. [3]), a difficulty coefficient is attached to each item. This numerical scale orders the items weakly and ascribes the often made observation that subjects solve a difficult item correctly and fail at an easier one to influences of chance and error. This interpretation appears so obvious that its proponents tend to overlook the *petitio principii* which it entails, that the difficulty of an item is defined by the numerical parameter, whereas the weak ordering of the items according to a material notion of difficulty is an empirical property which will in many cases not be fulfilled.

For example, an expert judging this might well regard some items as incomparable. In such a situation a partial- (or quasi-) ordering, which allows for incomparable items, is more flexible. Thus, the assignment of numerical values, albeit ubiquitous in psychological tests, restricts drastically the range of possible models. Even partial orders can be too restrictive because transitivity is also a strong condition. The theory of knowledge spaces takes account of this aspect (*see* Cognition: Knowledge Spaces for a deeper treatment of these issues).

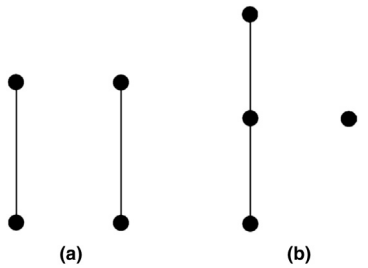

Figure 1 The forbidden substructures of a semiorder.

The most convenient way of representing a poset is by its (*Hasse*) *diagram* (see Figures 1 and 2). To construct it, one first defines $x < y$, in words 'y covers x,' to mean: $x < z \leq y$ implies $z = y$. Next, one draws a graph in which the vertices correspond to the elements of P, two elements are connected by an edge if one of them covers the other, and furthermore the covering element is placed above the horizontal line determined by the covered element. Hasse diagrams are easy to draw and are of great help in theoretical considerations as well as in interpreting empirical findings.

Special Classes of Orders

There are numerous specializations, that is, additions of further properties, such that the resulting structure is something in between a poset and a linear order. In this section, two classes are considered which have useful numerical representations, although not the strong one of eqn. [1]. The third specialization considered here, namely lattices, is important because it tends to bridge the gap between ordering relations and algebraic operations.

Semiorders

In psychophysics one has to deal with the phenomenon of subliminal changes in intensity of physical stimuli. One stimulus may be (physically) slightly more intense than another, but a subject does not perceive the difference. For example, one might imagine a coffeetester confronted with an extended series of cups of coffee, each with one more grain of sugar than its immediate predecessor. He/she will probably not notice the difference in taste between successive cups, but surely the difference between the first and the last cup will be clearly noticeable. In other words, indifference with respect to $\leq$ is not transitive. Similar observations can be made in decision theory and other branches of psychology. One way of describing such facts is to introduce sensory thresholds, thereby creating a scaling problem where one wants to scale both the subjective intensity and the magnitude of the threshold. Luce (1956) considered two conditions which seem plausible in this context (P is the set of stimuli and $x \leq y$ denotes the observation x is less intense than y):

$$\text{(S1)} \quad \text{If } x \leq y \text{ and } u \leq v, \text{ then } x \leq v \text{ or } u \leq y.$$

$$\text{(S2)} \quad \text{If } x \leq y \leq z \text{ and } W \in P, \text{ then } W \leq z \text{ on } x \leq W.$$

One easily shows that $(P, \leq)$ is a quasiorder, i.e., without losing generality, one can assume that it is a partial order. Condition (S1) says that it does not contain Figure 1(a) as a suborder. Similarly, condition (S2) is equivalent to the fact that it does not contain Figure 1(b) as a suborder.

From the viewpoint of measurement the most important property of semiorders is the possibility of a numerical representation with a constant threshold. More precisely, one can show that for a finite semiorder $(P, \leq)$ there is a function f

$$f : P \to \mathbb{R} : \text{ such that } x < y \text{ if } f(x) + 1 < f(y). \qquad [2]$$

The condition of finiteness is far too restrictive for eqn. [2], but without further restrictions it does not obtain. The height of the threshold is arbitrary, that is, the number one in eqn. [2] can be replaced by any other positive number (with a rescaled f). More on this kind of representation theorem can be found in Roberts (1979: Chap. 6.1) or Suppes et al. (1989: Chap. 16). Clearly, the converse of this theorem is true: a structure with a representation, eqn. [2] is necessarily a semiorder.

Although the semiorder concept seems to have its origin in the psychological context of the possibility of measurement of subjective intensity described above, it is now a well-established notion in the mathematical theory of posets and several remarkable results are connected with it. Some of them are mentioned in the following text.

Interval Orders

A structure satisfying only condition (S1) but not necessarily (S2) is called an interval order. This notion derives from the

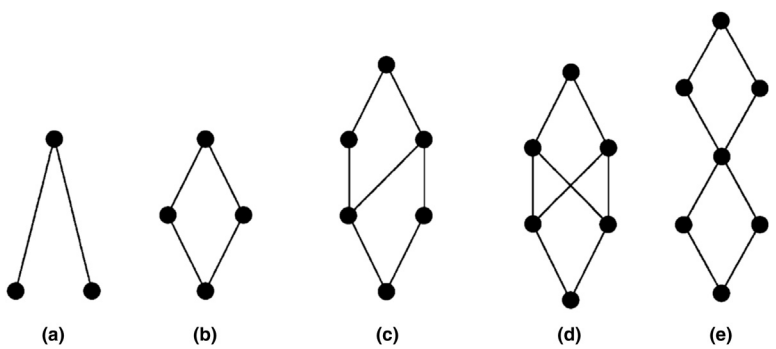

Figure 2 Lattices (b, c, e) and non-lattices (a, d).

numerical representation which is characteristic of this kind of order. Each element can be represented by an interval on the real line and $x < y$ or $y < x$ is tantamount to the nonoverlapping of the respective intervals. Put another way, a scale value and a threshold are attached to each element, but the height of the threshold can vary in contrast to eqn. [2]. The reader will find many interesting properties and proofs in Fishburn (1985).

Historically, it is of interest that, although the representation theorem is due to Fishburn (1970), the investigation of this structure goes back to papers of Norbert Wiener, early, in the twentieth century. What is even more remarkable is that in this work he mentioned possible applications in experimental psychology (see Fishburn and Monjardet, 1992).

The theory of interval orders has applications in archeology in connection with the seriation problem. In this case P is a set of artifacts, each element of which was some time in use. From excavations in various historical sites it is known for any two artifacts whether the period of use of one follows that of another with or without overlap. If this relation satisfies (S1) then, by the interval order representation theorem, it follows that it is possible to assign a time interval to each artifact, preserving the empirical observations.

A related notion is that of a biorder. The primitives of a biorder consist of two sets P and Q and a relation ρ between P and Q, i.e., a subset $\rho \subseteq P \times Q$ satisfying for all $x, y \in P$ and $u, v \in Q$

$$x\rho u,\ y\rho v,\ \text{implies } x\rho v \text{ or } y\rho u. \qquad [3]$$

To understand this condition, it is helpful to think of P as a set of subjects and of Q as the items of an achievement test. $x\rho u$ means 'subject x fails to solve item u.' The premise of eqn. [3] says that x fails u and y fails u: if one assumes that x solves v correctly, then y has to fail u. With this interpretation, the biorder concept is the basis of Guttman scaling. Probabilistic modifications of it underly Rasch scaling, which is well-established in the theory of psychological tests. Thus, with the biorder condition eqn. [3], a testable property is formulated which, when satisfied, justifies a numerical difficulty value for items because a numerical representation of a biorder consists of two functions $f : P \to \mathbb{R}$ and $g : Q \to \mathbb{R}$ such that:

$$x\rho u \text{ if } f(x) < g(y).$$

This, in turn, implies eqn. [3]. In the terminology of the above interpretation of a biorder $f(x)$ is the ability of subject x and $g(u)$ is the difficulty of item u. If in a biorder, the two sets coincide, $P = Q$, then eqn. [3] reduces to (S1), i.e., the structure is an interval order.

Lattices

In a partial order $(P, \leq)$ subsets of P may or may not have upper or lower bounds, where upper bounds are elements which dominate all elements of the subset and lower bounds are elements which are dominated by all its elements. If each two-element subset has a unique smallest upper bound – called the supremum – and a unique greatest lower bound – called the infimum – then the structure is called a lattice. The infimum of x and y is denoted by $x \wedge y$ and the supremum by $x \vee y$. Examples of lattices and nonlattices are given in **Figure 2**.

Further examples are the power set of a set ordered by set inclusion, any subset of the reals naturally ordered, and the natural numbers $\{1,2,3, \ldots\}$ ordered by divisibility, i.e., $m \leq n$ if m divides n. In the last example, the least common multiple and the greatest common divisor are the supremum and the infimum, respectively. It follows easily by induction that any finite subset of a lattice P has an infimum and a supremum. However, infinite subsets do not necessarily have this property. If all subsets of P have it, then the lattice is called complete. Thus, finite lattices are always complete, but the rationals with the natural ordering are not. The question of completion of lattices similar to the construction of the reals by Dedekind cuts is discussed in Section Embeddability and Completion of Orders.

There is a more algebraic way of characterizing lattices, where infimum and supremum are regarded as binary operations on P:

Theorem 1 Let $(P, \leq)$ be a lattice. Then for all $x, y, z \in P$:

1. $x \vee (y \vee x) = (x \vee y) \vee z$ and $x \wedge (y \wedge z) = (x \wedge y) \wedge z$
2. $x \vee y = y \vee x$ and $x \wedge y = y \wedge x$
3. $x \wedge (x \vee y) = x$ and $x \vee (x \wedge y) = x$

Conversely, if a structure $(P, \vee, \wedge)$ is given with operations $\vee$, $\wedge$ satisfying the conditions (1), (2), and (3), then $\leq$ is defined by

$$x \leq y, \text{ if } x \wedge y = x$$

is a partial order on P such that $(P, \leq)$ is a lattice.

Lattices play an important role in pure and applied mathematics. John von Neumann, one of founders of lattice theory, developed his formulation of quantum mechanics in terms of lattices. Today they are used in many branches of mathematics – for instance, in combinatories, probability theory, logic, and functional analysis. The wide applicability of lattices in technical and social science contexts is best demonstrated by the development of 'concept analysis' by Rudolf Wille (see Ganter and Wille, 1996).

Extension of Orders; Szpilrajn's Theorem

Is it possible to extend a partial order by adding (x, y) or (y, x) to $\leq$, when x and y are incomparable in $\leq$ and at the same time keeping the extended relation a partial order? Is it possible to extend $\leq$ to a linear order in this way? Both questions are answered in the affirmative by the following theorem.

Theorem 2 (Szpilrajn's Theorem) Let $(P, \leq)$ be a poset and $x, y \in P$ such that neither $x \leq y$ nor $y \leq x$. Then there are two linear orders $\leq'$ and $\leq''$, each of which contains $\leq$ as a subset, and such that $x \leq' y$ and $y \leq'' x$.

The proof of this theorem is not difficult for finite P, but entails a set theoretic twist in the infinite case. If x and y are incomparable in $\leq$ then one defines $x \leq' y$, adds all pairs to $\leq'$ which this choice and transitivity forces and goes ahead until no incomparable pairs are left over. The latter step of this argument is a bit of a problem for an infinite P.

Sometimes one can regard a partial order as ensuing from incomplete data collection, incomplete in so far as a linear order can be expected, but the available data give only a part of the full order information. Then the question of a linear

extension is relevant, in particular, one can ask how many additional pairwise comparisons must be performed to obtain the full linear order, which is the most parsimonious way to find such an extension.

For these problems the number of linear extensions is of importance. It was shown by Fishburn and Trotter (1992) that for all partial orders with fixed $|P| = n$ and k pairs in $\leq$ those with a maximal number of linear extensions are semiorders. For $k \leq n$ these semiorders are characterized, but for larger K there seem to remain several unsolved problems.

It can be shown that a linear extension can be found with at most K questions of the form 'is x larger than y,' where K is of the magnitude of the logarithm of the number of linear extensions (see Trotter, 1995 for details). In connection with this problem it seems that a pair x, y for which many extensions with $x \leq y$, as well as many with $y \leq x$ exist is best to investigate first because, whatever the answer, it considerably reduces the number of data which remain to be collected. In this context the so called 1/3–2/3 conjecture was made. It says that each poset which is not a linear order contains a pair (x, y) for which the relative frequency of extensions with $x \leq y$ is between 1/3 and 2/3. This conjecture seems still open. However, bounds are known, e.g., the relative frequency is between 3/11 and 8/11. The proof of this result draws on a geometric inequality and is extremely beautiful. For semiorders the 1/3–2/3 conjecture is true. More on this topic and some of the proofs can be found in Trotter (1995).

Order Dimension

A corollary of Szpilrajn's Theorem is that each partial order is the intersection of its linear extensions. One can ask for the minimal number of linear extensions with this property and regard this number as a measure of complexity of the order. It is called the dimension of $(P, \leq)$.

All orders on three, four, and five elements are two-dimensional (provided they are not linear). There are two orders with six elements of dimension three. One can group the posets in classes of increasing dimension. The semiorders have a dimension less than or equal to three, but the interval orders can have arbitrary large dimension.

One classical result on the dimension is that it is always smaller than or equal to the maximal number of pairwise incomparable elements. For $|P| \geq 4$ ($|P|$ denotes the number of elements in P) the dimension is bounded above by $|P|/2$ and this is the best possible solution. It is an active area of research to find inequalities such as these which relate the dimension to other characteristics of the poset (see Trotter, 1992).

There are other variants of the dimension concept: one can replace linear orders by semiorders (and obtain the semiorder dimension) or by interval orders or by any other class of partial orders. For biorders, see Doignon et al. (1984) and Koppen (1987). These two papers attempt to link this dimension concept to techniques of extracting factors to explain data. In this way, an ordinal factor analysis is outlined. Factor analytic procedures used to be widely spread in social science research, but until now they seem to lack a sound axiomatic foundation. The order theoretic dimension concept can provide a weaker concept of a factor which is, on the other hand, much easier to justify, or empirically to test for adequacy.

Embeddability and Completion of Orders

Sometimes the nonexistence of suprema and infima is a nuisance; just think of the inconvenience caused when a sequence of rational numbers 'converges,' but the limit does not exist (in the rationals). One tries to avoid this situation by embedding a given poset into a complete lattice which contains the original order as a suborder. In contrast to the techniques in Section Extension of Orders; Szpilrajn's Theorem, one adds new elements to P instead of new pairs to $\leq$. Thus, the aim is to find a superstructure which resembles the given order as closely as possible, but which can be handled more conveniently because of the existence of suprema and infima. The prototype of this kind of extension is the time-honored construction of the reals from the rational numbers.

To begin with, some notation must be introduced. For $X \subseteq P$ let:

$$X^{\nabla} := \{y \in P : y \leq x \text{ for all } x \in X\}$$

$$X^{\Delta} := \{y \in P; x \leq y \text{ for all } x \in X\}.$$

Thus X^{∇} is the set of lower bounds of X and X^{Δ} the set of upper bounds of X. The sets $\{x\}^{\nabla}$ and $\{x\}^{\Delta}$ for $x \in P$ are abbreviated by x^{∇} and x^{Δ}, respectively. One simple way of embedding consists of considering

$$\nabla_{(P, \leq)} := \{x^{\nabla}; x \in P\}$$

the so called principal ideal representation. Note that $(\nabla_{(P, \leq)}, \subseteq)$, is a suborder of the complete lattice of the power set, isomorphic to $(P, \leq)$. However, this embedding is not sufficient for the purposes of this section; the superstructure carries too little information from the original one. Therefore, a more parsimonious extension must be constructed.

To this end, a pair (X, Y) is called a MacNeille cut if and only if $X = Y^{\nabla}$ and $Y = X^{\Delta}$. The set of all MacNeille cuts is a better candidate for an embedding. Indeed, one can prove:

Theorem 3 Denote by $C_{(p, \leq)}$ the set of all MacNeille cuts of a partial order $(P, \leq)$. Define for two cuts (X, Y) and (X', Y') the relation

$$(X, Y) \sqsubseteq (X', Y') \text{if } X \subseteq X'.$$

Then $(C_{(p, \leq)}, \sqsubseteq)$ is a complete lattice. Moreover, it is the unique (up to isomorphism) smallest complete lattice which contains the original poset $(P, \leq)$ as a supremum dense subset, i.e., each element of $C_{(p, \leq)}$ is supremum of elements of $(P, \leq)$.

A few remarks should illustrate this theorem. First, the set corresponding to $(P, \leq)$ in $(C_{(p, \leq)}, \sqsubseteq)$ is the set of cuts (x^{∇}, x^{Δ}). Second, if this theorem is applied to the rationals with the usual (linear) order, then the reals with added $\{-\infty\}$ and $\{\infty\}$ result as the MacNeille completion. The element $-\infty$ is the cut $(\emptyset, P)$ while ∞ is the cut $(P, \emptyset)$. Third, although the notion of a complete lattice only bears meaning for infinite base sets P, the MacNeille completion makes sense also for finite posets. **Figure 2(b)** shows the MacNeille completion of (a) and **Figure 2(e)** of (d). Finally, it is worth mentioning that the mappings $X \to X^{\nabla}$ and $X \to X^{\Delta}$ form a pair of functions that are known as a Galois connection. This construction is often used when dealing with posets; it is applicable in other contexts, for example, formal concept analysis (see Ganter and Wille (1996) is based on a Galois connection). In this respect, concept lattices are a generalization of a MacNeille completion.

See also: Majorization and Stochastic Orders; Measurement, Representational Theory of; Order Statistics; Ordered Relational Structures.

Bibliography

Bruggemann, R., Carlsen, L., 2011. An improved estimation of averaged ranks of partial orders. MATCH: Communications in Mathematical and in Computer Chemistry 65, 383–414.

Doignon, J.-C., Ducamp, A., Falmagne, J.-C., 1984. On realizable biorders and the biorder dimension of a relation. Journal of Mathematical Psychology 28, 73–109.

Fishburn, P.C., 1970. Intransitive indifference with unequal indifference intervals. Journal of Mathematical Psychology 7, 144–149.

Fishburn, P.C., 1985. Interval Orders and Interval Graphs. Wiley, New York.

Fishburn, P.C., Monjardet, B., 1992. Norbert Wiener on the theory of measurement (1914, 1915, 1921). Journal of Mathematical Psychology 36, 165–184.

Fishburn, P.C., Trotter, W.T., 1992. Linear extensions of semiorders: a maximization problem. Discrete Mathematics 103, 25–40.

Ganter, B., Wille, R., 1996. Formale Begriffsanalyse. Springer, Berlin.

Koppen, M.G.M., 1987. On finding the bidimension of a relation. Journal of Mathematical Psychology 31, 155–178.

Luce, R.D., 1956. Semiorders and a theory of utility discrimination. Econometrica 24, 178–191.

Mašulović, D., Order November 2007. Homomorphism-Homogeneous Partially Ordered Sets 24 (4), 215–226.

Roberts, F.S., 1979. Measurement Theory with Applications to Decisionmaking, Utility, and the Social Sciences. Addison- Wesley, Reading, MA.

Suppes, P., Krantz, D.H., Luce, R.D., Tversky, A., 1989. Foundations of measurement. In: Geometrical, Threshold, and Probabilistic Representations, vol. 2. Academic Press, New York.

Trotter, W.T., 1992. Combinatories and Partially Ordered Sets: Dimension Theory. Johns Hopkins University Press, Baltimore, MD.

Trotter, W.T., 1995. Partially ordered sets. In: Graham, R.L., Grötschel, M., Lovász, L. (Eds.), Handbook of Combinatories. Elsevier, Amsterdam, pp. 433–480.

Volkov, M.V., September 1, 2009. Synchronizing automata preserving a chain of partial orders. Theoretical Computer Science 410 (37), 3513–3519.

Participant Observation

Gary Alan Fine, Northwestern University, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Participant observation is one of several qualitative field methodologies used to understand the culture and structure of communities. Used primarily by sociologists, participant observation involves the researcher engaging with the community being examined. This article describes the benefits and drawbacks of this form of ethnographic observation, discussing the historical development of this methodology and describing the variety of roles that a researcher can take in gathering data from a community, and comparisons with other related methodologies. The article concludes by discussing several of the ethical issues (deception, informed consent, confidentiality, and precision of descriptions) that practitioners of this methodology must confront.

Participant observation has been part of the arsenal of methodological techniques employed by social scientists over the past century, notably within the discipline of sociology. This style of data collection involves the active engagement of the researcher with the members of the community that he or she wishes to study, typically as an equal member of the group.

This methodology falls under the general rubric of field research. It is important to recognize that there exist a variety of forms of 'field research,' a range of possibilities that this article will not cover. However, in contrast to the manner in which observational research has traditionally been conducted in anthropology, in sociology much ethnography is based on the assumption that the researcher is not an outsider, but at least nominally a participant in the group being studied. Yet, the line that separates participant observation and other forms of ethnographic observation is uncertain and hazy, and for some scholars the terms 'participant observation' and 'ethnography' have been used as synonyms. Today the term 'ethnography' is the more general of the two, found in wider circulation, such as in the journal titles *Ethnography* and the *Journal of Contemporary Ethnography*.

Since the mid-1970s, there has been a notable increase in the legitimacy of field methodologies within the social sciences. These qualitative, interpretive strategies, once considered 'unscientific,' akin to journalism, have demonstrated their utility in a number of disciplines. Whereas once it was rare to find research grounded on direct observation published in leading journals, this is no longer the case today. That this occurred during a period in which social scientific funding was being decreased is surely no coincidence, as participant observation is not capital intensive, even though it requires much labor on the part of the observer. Although qualitative field methods still do not have equal standing with quantitative methods (especially in fields such as psychology and economics), their growth and acceptance are impressive.

Advantages

Several distinct advantages justify participant observation; among these are the benefits of rich data, validity, *verstehen* (or interpretive understanding), and economy.

Richness

This methodology, in contrast to most research techniques that do not involve personal witnessing, provides for the gathering of rich and detailed data. To investigate *action in situ*, direct observation is essential. When observation is coupled with the questioning of actors about their choices, tied to some measure of active involvement by the researcher, the value of participant observation is known through the detail provided. This style of data collection permits an intense depiction that produces a fullness of understanding.

Validity

A second benefit is analytical validity. Because the observations are of behavior *in situ*, a researcher can rely upon the claim that the findings are close to the scene as it is known by others. If the conclusions drawn from participant observation research are not always reliable, given that other investigators might reach different conclusions, the actions described did occur and descriptions reflect the belief that some aspect of the world, as it 'truly' is, is being described.

Interpretive Understanding

Participant observation supports the demand of Max Weber to produce research that is characterized by *verstehen*, or a personal, knowing understanding. In this way, participant observation with its emphasis on both participation and observation adds to lived knowledge. By directly involving the researcher in the activity, one can understand from an immediate level the dynamics and motivations of behavior. The observer crossed the boundary between outsider and insider. While research projects differ significantly on this dimension of direct experience, involvement has advantages for a *verstende* analysis that other approaches cannot match.

Economy

Participant observation research is typically inexpensive. While this research strategy is labor intensive, it is not as capital intensive as techniques such as survey research or experimentation. In many cases, the researcher works alone and, as

a result, can set the terms of his or her own involvement. Increasingly, however, team research projects are used to collect data, and in these cases wages and benefits for hired observers may involve considerable costs, even if they multiply the number of sites observed. Yet, in the traditional participant observation project, the key resource necessary is a technology for inscribing observations (e.g., notepads or word processors). Perhaps for reasons of cost, connected to the time and energy involved, participant observation is a style of research that is most frequently engaged in by scholars at the early stages of their academic careers.

Disadvantages

Just as significant advantages characterize this methodology, several disadvantages are evident. Among the key problems that have been described are those relating to proof, generalizability, bias, and time commitments.

Proof

Participant observation relies upon a single case study: the examination of one place, one scene, and one group. This raises questions about the nature of proof, or, put another way, about reliability. Will two researchers examining similar social scenes reach the same conclusions? Often because of different perspectives upon entering the field and different experiences while observing, conclusions vary greatly. While the observations and the interpretations of those observations may be compelling, a reader can reasonably wonder whether any set of conclusions is definitive.

Generalizability

Even if we accept the legitimacy of analyzing one scene, on what grounds can we generalize beyond that setting? How far can our conclusions be pushed? Participant observation research has a problem in this regard because of the absence of scientific control that characterizes experimental research and produces confidence in the claim that important variables of social life have been adequately captured. As a result, the extent to which generalizability is legitimate will inevitably be problematic in participant observation research. Participant observers must present a theoretical model that helps readers judge the legitimacy of their broader claims in light of the audience's own experiences. Do the findings apply only to one setting at one moment, or can they be extended?

Bias

A strength of participant observation methodology is that the researcher's insight and perspective are taken into account, but this strength has a downside. It is difficult to distinguish between perspective and bias – the former with a positive implication, and the latter with a negative tone. The personal background and set of experiences that the researcher brings to a scene will inevitably be distinctively different from those of other researchers and, for that matter, from the perspectives of the participants in the setting. To the extent that the researcher's perspectives differ significantly from those of the participants – possibly because of the generally progressive values and upper-middle-class status of academics – the understanding of a particular scene may be systematically biased.

Time

Just as participant observation research is relatively inexpensive, it is also highly labor intensive. This style of research requires that the researcher be present in the observed social scene. One cannot visit quickly, but must spend extended periods of time so that the full range of activities in which participants engage is seen and understood. Much participant observation depends upon the vagaries of chance – what happens to occur at the moment of observation, and, as a result, a significant investment of time is needed. While there is no definitive rule for the appropriate length of time necessary for observation, most projects require months, if not years, to complete. Coupled with a modest requirement for capital equipment, this means that, as noted, this methodology is feasible for younger scholars. As a result, participant observation studies often do not have the depth of theoretical understanding that is more likely to characterize the work of senior scholars.

The Growth and Development of Participant Observation

Although participant observation methodology has deep roots in ethnographic traditions (including in travelers' accounts and missionary tales), participant observation within the academy has its origins in the discipline of sociology at the turn of the twentieth century, particularly with scholars affiliated with the Department of Sociology at the University of Chicago. Notable at this early stage were the participant observation studies of Annie Marion MacLean, a student at Chicago and later a professor at Adelphi University. Maclean worked and observed in department stores, sweatshops, and coal fields in order to gain an appreciation of conditions in these scenes. By the 1920s, participant observation was well established at the University of Chicago, as evidenced in studies by Frances Donovan on the lives of waitresses and Frederic Thrasher's observations of gang life. Although members of the 'Chicago school' of sociology employed numerous methodological techniques, these scholars were probably best known for their observational research and depiction of the life stories of informants. By midcentury, participant observation had expanded well beyond the bounds of Chicago. Classic works such as Whyte's (1943) *Street Corner Society*, Hollingshead's (1949) *Elmtown's Youth*, Vidich and Bensman's (1958) *Small Town in Mass Society*, and Gans's (1963) *The Urban Villagers* collectively demonstrated that through intense involvement in a community, a researcher could gain a perspective on a social domain that was both empirically and theoretically powerful. The fact that many of these research projects covered life in city environments eventually led some to describe this approach as 'urban ethnography,' an approach that has become more prominent in the past 20 years, as exemplified in Anderson's (1992) *Streetwise*, Duneier's (1999) *Sidewalk*, and Pattillo's

(1999) *Black Picket Fences*. Another stream of participant observation research was grounded in the study of deviant groups and activities. Researchers in such domains have the dilemma of determining the degree to which participation should bolster observation, a problem that is simultaneously strategic, ethical, and legal. Exemplifying such research are studies of the mentally ill, gang members, and drug dealers, notably the influential works by Goffman (1961) *Asylums*, Liebow (1967) *Talley's Corner*, Horowitz's (1983) *Honor and the American Dream*, Snow and Anderson's (1993) *Down on Their Luck*, and Venkatesh's (2008) *Gang Leader for a Day*. These themes – gritty urban life and troubling social deviance (and their control) – have been integral to participant observation. The increased interest in these themes over the past few decades has created a 'new urban ethnography' that focuses on broader issues of urban poverty and racial injustice, along with the analysis of group dynamics and local cultures.

Not all researchers focus on social problems. In the last few decades, beginning with the inspiration of the Hawthorne studies, and continuing the research of Donald Roy, Melville Dalton, and Robert Jackall, increasing attention has been paid to observations of workplaces. The participant observation of organizational life has developed rapidly, creating an active focus on the dynamics of organizational domains, both in traditional social science departments and within schools of management. Factories, airlines, restaurants, and high-tech engineering firms have all been the foci of research. Yet, every social world in which individuals with a common interest meet and to which they hold affiliation can be analyzed through participant observation, whether clubs, sports teams, or bars.

Comparison with Other Field Methods

It is easy to assert that participant observation is synonymous with qualitative field research, and, in reality, the lines are blurred. There are now so many different, overlapping approaches, which have been given different names, even if the styles and traditions of research do not differ much. However, participant observation can be contrasted with several other techniques with which it shares some commonalities. Some scholars engage in 'ethnographic immersion.' In such investigations, the researcher becomes a full member of the group and, in effect, studies his or her own group. In this case, it is claimed that the researcher has 'gone native.' This strategy has the advantage that the researcher comes by his or her 'member's knowledge' naturally, but simultaneously the immersion may diminish the ability to examine the group life from the critical, outsider perspective that being marginal sometimes brings. Most forms of participant observation involve a distinction between being a researcher and being a full member of the group studied. Typically, members are told that they are being observed by an outsider, who, for purposes of the study, may participate actively in the life of the group, even though she or he is not a full member and is not recognized as such.

A second technique, related to ethnographic immersion, has been labeled 'auto-ethnography,' akin to the methodology of introspectionism of the early years of the twentieth century. In these studies, the researcher mines his or her life experiences and emotions, hoping to demonstrate the embodied characteristics of social forces. One such example is Ellis's (1995) *Final Negotiations*, an emotional and analytical account of her own reaction to the lengthy dying of her significant other. This writing, while occasionally powerful, insightful, and moving, can be faulted for its reliance on subjective experience. Auto-ethnography might be seen as an extreme instance of total immersion, with nothing outside the researcher's experience; given this personal source of data, these provocative arguments become difficult for others to evaluate. The author proposes a singular account of emotions that, because it is his or her account, leaves no room for alternative interpretations.

A third approach is to collect data by means of in-depth interviews. Most participant observation involves extended questioning of informants. However, some research relies on this questioning to the exclusion of direct observation. This method has the advantage of permitting the researcher to gain multiple perspectives of the same scene and permits the investigation of domains to which it might be impossible to gain direct access. However, in-depth interviews also limit the researcher to the presentations of interested parties. The claims that are revealed in interviews may be systematically different from the private thoughts of informants and may also differ substantially from their behavior. Observations provide a check on what one is told for reasons of impression management. As always, triangulation of data provides advantages, curbing the weakness of any one methodology.

Fourth, some researchers deemphasize participation and are, in practice, pure observers. One group of observers makes precise statistical measurements of public behaviors, but this approach is distant from participant observation. More relevant are those scholars who examine public behavior, particularly behavior within open urban spaces. This distinguished tradition of scholarship includes such researchers as William H. Whyte, Erving Goffman, Lyn Lofland, and, more recently, Carol Brooks Gardner. A close reading of street activity, parallel to certain forms of animal ethology, permits the recognition of patterns of human activity.

Within participant observation itself, researchers select particular roles or strategies. The key dimension on which roles can be differentiated is the extent to which researchers choose to be involved in group life. Adler and Adler (1987) distinguish participant observation methods according to the extent of active involvement: peripheral membership, active membership, and complete membership. In general, the peripheral member observes as an outsider to the group and its culture, whereas the active member is directly engaged in the life of the group, while making it clear to the group that she or he is not fully a member and is conducting research. In the model of the Adlers, the complete observer engages in ethnographic immersion, sometimes informing the group of the research goals.

Strategies of Participation Observation

Unlike many techniques of quantitative research and statistical analysis, there are no firm and fast rules that determine how participant observation research should be conducted. Such a reality has proven frustrating to generations of students who desire definite directions. Traditionally, a novice learned the

skills of participant observation through what amounted to an apprenticeship relation with a senior researcher. However, because of the increasing numbers of field researchers, many social science departments, particularly those in sociology and anthropology, have instituted field methods courses in which some rudimentary – and flexible – strategies of the method are taught.

Access

The first issue that must be confronted in a research project is the question of access. How does one convince a group that they should willingly open their activities to a stranger whom they do not know and whom may criticize the group? This is especially applicable to groups that are generally perceived as deviant (criminal gangs or politically marginal movements), but it also recognizes that any group, striving to maintain boundaries that differentiate it from others, may engage in some activities that members wish to shield from outsiders and that could be viewed negatively. Participant observers must, within the limits of their ethical standards, persuade the group that they can be trusted not to violate shared expectations. Through this process, they set the rules of engagement. Such rules, explicit or implicit, are uniquely formulated for each project. Further, this access emerges and changes over time. One is rarely accepted all at once, but often the group maintains a cautious vigilance until they become satisfied that the researcher can be trusted. The researcher is continually being tested. The participant observer must be sensitive to this process of judgment.

Socialization

Access is only a first step. Once the researcher gains entree to a setting, he or she must determine what 'behavior' means. In effect, the participant observer must become socialized to the environment with its norms, cultural traditions, and jargon. The ease of this process varies as a consequence of how well acquainted the researcher is with the setting prior to the observation, but unless the researcher begins as a full member, a process of acculturation will typically have to be mastered. At times, the group will become frustrated by the lack of competence of the observer, but more often this process of learning is valuable as a methodological strategy, permitting the researcher to ask questions that would be denied to full members as being too obvious.

Developing Relationships

Participant observation is a methodology that depends upon the establishment of relationships. It relies upon sociability. One essential skill is the ability to recruit informants from the field. The most important of these are termed 'key informants,' such as Doc in Whyte's (1943) classic *Street Corner Society*. These individuals serve both as guides for the participant observer and as a means of vouching for the researcher's legitimacy within the setting, providing an insider's stamp of approval. The researcher who lacks the ability to make connections will have difficulty collecting credible research data.

Field Notes

The researcher not only needs to observe activity, but also must subsequently inscribe that material in what are labeled 'field notes,' documents that serve as the evidentiary basis of the published report. Considerable discussion has addressed what constitutes appropriate field notes and how these notes should be presented (Van Maanen, 2011). These notes provide the 'facts' that are observed, including actions, statements, and the feelings of the observer. Participant observers are enjoined to write as much of what transpires as possible, particularly in the early stages of the research when one is determining what issues are most salient. This emphasizes that participant observation is often an inductive methodology, with the field data intuitively revealing the important issues, rather than deductively testing preestablished hypotheses.

Exit Strategies

Much research simply ends. Once completed, the project closes down. Exiting is rarely as simple as this in participant observation because the researcher and informants have developed intense social ties. Leaving the field involves the termination of a set of powerful relationships, and sometimes produces loneliness or even betrayal among either party. As a result, the process of disengagement often involves a longer period, and some connection with the field setting often continues, even if the research is officially completed. Researchers are often advised – for both practical and ethical reasons – to share their writing with their informants, either to gain additional feedback or to permit those informants to use the analysis for their own ends. The linkages that are established during the course of the observation can be a continuing source of strength, both for researchers and for informants.

Ethical Concerns

Although all forms of research that involve human subjects raise ethical issues, the variety of ethical concerns in participant observation is particularly great because of the degree of interpersonal dynamics involved. Four ethical issues have been central to participant observation research: deception, informed consent, confidentiality, and the accuracy of depiction. Although these by no means exhaust potential ethical concerns, each one raises important issues.

Deception

Most participant observers agree that it is both pragmatically and ethically necessary to announce to one's informants that one is conducting research. In part, this is because participant observation depends on personal relations among individuals, who, for the period of time of the study, are status peers. To claim to be something that one is not – to pretend to be a full member of the group – without revealing one's purpose could undermine the entire enterprise if it becomes discovered. Lies are often difficult to maintain for long periods. Further, deception suggests that one's interests take priority over the interests of another: can we justify such trickery? While research on groups that hope to keep their activities hidden from

a wider public – either deviant groups or, occasionally, elite groups – motivates deceptive practices, such strategies are troublesome.

Informed Consent

Deception involves the active misleading of others, but how should we view the absence of full disclosure? Do group members have the right to know that they are being investigated? For instance, one can easily join a leisure group without announcing that one is taking notes or planning to publish. Is this morally proper? As noted, good reasons may exist for announcing one's presence, permitting one's informants to serve, in effect, as research assistants. But should researchers be required to gain the informed consent of their informants; further, how much information is necessary for informed consent, particularly in a methodology that depends on inductive methods in which the researcher may truly not be aware of what issues will eventually become central to the analysis? One's topics depend on what one observes. During the past few decades, American universities have instituted *ethics* committees (Human Subjects Committees and Institutional Review Boards) that are designed to protect the rights of research subjects. Frequently, researchers are required to gain some measure of informed consent, but how much information is required remains a matter of debate.

Confidentiality

One element that most (although not all) researchers agree on is that informants' names and clearly identifying information should not be included in any published research report, although some who critically examine elites feel that such forms of anonymity are not essential. In general, researchers believe that publishing the names or identifiers of particular informants or settings is neither necessary nor warranted. At times this decision poses a problem in that field notes may describe actions or statements that can be recognized, certainly by others within the scene. As a result, participant observers must choose how much information to alter to preserve confidentiality, even when this shades the full 'facts' of the case.

Precision of Description

Although it is clearly desirable to present data that are precisely accurate – direct quotations and exact descriptions of behavior – participant observation methodology and the limits of human ability make these values unattainable goals, as argued in my *'Ten Lies of Ethnography'* (Fine, 1993). In part we are imperfect in our ability to depict reality, but in part we may recognize that the details of human activity may actually obscure our crucial analytical points. For instance, the speech of informants, like the speech of researchers themselves, is marked by false starts, errors, and grammatical infelicities. Further, so much behavior is happening at any given moment that the researcher can easily become overwhelmed with a surfeit of trivial observations. The ethical balance that must be struck is to do justice to the events that are transpiring, while simultaneously making the account as clear as possible to permit communication with readers.

Summary

This brief overview examines central features of the participant observation methodology. In doing so, I have deemphasized both the strategies of analysis (the extended case study method and grounded theory) used by practitioners and the modes of presentation (realist tales and impressionist tales). No method of data collection can exist without related techniques by which data are analyzed and presented.

Throughout the history of modern social science, participant observation has proven to be a powerful technique by which researchers can understand a diverse set of social worlds in a way that does justice to the complexity of the activities of group members and to their own perspectives and understandings. While this style of research benefits by the triangulation with methodologies in order to create adequate theoretical understanding, participant observation provides the basis by which researchers can understand the richness of the social settings that they – and others – inhabit.

See also: Field Observational Research in Anthropology and Sociology; Fieldwork in Social and Cultural Anthropology; Fieldwork: Ethical Aspects; Observational Studies: Outcome-Based Sampling; Observational Studies: Overview.

Bibliography

Adler, P.A., Adler, P., 1987. Membership Roles in Field Research. Sage, Thousand Oaks, CA.
Anderson, E., 1992. Streetwise. University of Chicago Press, Chicago, IL.
Duneier, M., 1999. Sidewalk. Farrar Straus Giroux, New York.
Ellis, C., 1995. Final Negotiations. Temple University Press, Philadelphia, PA.
Fine, G.A., 1993. Ten lies of ethnography. Journal of Contemporary Ethnography 22, 267–294.
Gans, H., 1963. Urban Villagers. Free Press, New York.
Goffman, E., 1961. Asylums. Anchor, New York.
Hollingshead, A., 1949. Elmstown's Youth. John Wiley, New York.
Horowitz, R., 1983. Honor and the American Dream. Rutgers University Press, New Brunswick, NJ.
Liebow, E., 1967. Talley's Corner. Little Brown, Boston, MA.
Pattillo, M., 1999. Black Picket Fences. University of Chicago Press, Chicago, IL.
Snow, D., Anderson, L., 1993. Down on Their Luck. University of California Press, Berkeley, CA.
Van Maanen, J., 2011. Tales of the Field, second ed. University of Chicago Press, Chicago, IL.
Venkatesh, S., 2008. Gang Leader for a Day. Penguin, New York.
Vidich, A., Bensman, J., 1958. Small Town in Mass Society. Princeton University Press, Princeton, NJ.
Whyte, W.F., 1943. Street Corner Society. University of Chicago Press, Chicago, IL.

Participatory Action Research in Environmental and Ecological Studies

Frances Fahy, School of Geography and Archeology, National University of Ireland Galway, Galway, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Often used in attempts to merge theory with practice and to collaboratively generate solutions to practical problems, participatory and action research have emerged as leading approaches within the social and behavioral sciences. While there is no one definitive form of action or participatory research, this article focuses on participatory action research (PAR) and highlights its key features and functions, drawing on examples of its application. An overview of some of the studies and settings within which PAR has been frequently used is presented. The article concludes by reviewing significant challenges and opportunities for researchers employing these participatory and action research methods within the social and behavioral sciences.

Introduction: What is Participatory and Action Research?

'Participation,' 'research,' and 'action' are often perceived as the bedrock for action research methods and the merits of these approaches to research have been identified as their ability to generate both rigorous scientific evidence and solutions to practical and policy problems. Moreover, their potential to empower communities and practitioners has been recognized. In the past, participatory and action orientated approaches were concerned primarily with collecting and sharing the views of interested people as an end in itself. Increasingly, the democratic and collective nature of knowledge creation is generating more intensive study and interest. While participatory and action research approaches are not new, developments in the employment of these approaches have resulted in a more critical participatory action research (PAR) agenda, which has garnered a recent revival in interest, earning it a spot as one of the leading approaches within the social and behavioral sciences. The following sections briefly review the origins of participatory and action research before examining the form and features of PAR.

Participatory Approaches

While there are numerous participatory approaches, for the purposes of this article it is important to distinguish participatory research as a research method in contrast to research surrounding participatory planning and decision making.

Participatory approaches in planning in particular gained prominence in the 1970s and the early 1980s. Interest in these approaches has grown as researchers began to realize that the opinions and priorities of individuals should be incorporated into policy development and project implementation, paralleling a shift from top-down, autocratic interventions toward a new emphasis on bottom-up, democratic empowerment. Since the 1990s the latter view is dominant, and indeed increasingly expected, among policy makers. The often-stated policy goal of involving all stakeholders, especially those who previously did not have a voice, reflects this emphasis on bottom-up approaches in which the participants in the research are involved in decision making and the selection of priorities. Within planning and policy-making arenas, there are frequent debates not only about the type of the engagement that is required by participating groups but also in the extent to which that engagement is linked to decision-making practices. Arnstein's (1969) classic 'ladder of citizen participation' is often employed to categorize varying degrees of participation, from low, limited, or nonparticipation, through tokenism (consultation, information provision, and placation), to high levels of participation (involving *participants as coresearchers* and forms of citizen control). Early participatory approaches were often critiqued for centering on the lower end of Arnstein's ladder of participation, focusing, for example, on limited consultation through opinion poll surveys and/or public meetings. The growth in participatory approaches has resulted in a plethora of new tools and techniques, including citizens' juries, visioning, world cafes, consensus conferencing, appreciative inquiries, forum theaters, and planning for real.

While participatory research is frequently utilized by researchers concerned with planning and policy making, participatory research is distinct from participatory planning. It is essentially a method where the central aim is to develop a process where locally based knowledge emerges to develop theory that is sensitive to local contexts. Participatory research involves conducting research *in conjunction with the people whose lives and activities are under investigation*. Moreover, priority is given to the views and attitudes of local people, recognizing them as *the* experts within *their own* environments. Participatory research approaches have a long history and are frequently associated with social transformation in the developing world. Often viewed as an alternative philosophy of social research, participatory research has its origins in liberation theology and neo-Marxist approaches to community development, for example in Latin American contexts (Freire, 1972). But it also has roots in human rights activism, for example in Asia (Kemmis and McTaggart, 2008).

Participatory research strives to be inclusive (involve all, or representatives of all, groups who will be affected by the outcomes of a process regardless of their status); transparent (with open dialogue and communication); and empowering (should promote mutual learning and empowerment). Cooperation and sharing of responsibility of power are seen as essential elements within participatory research. However, there are varying degrees of participation and engagement that can be evident in any participatory research project. Cornwall

and Jewkes (1995) contend that researchers primarily use four modes of participation within participatory research. The first and weakest form is *contractual*, a process whereby people essentially act as informants in a research study; they are contracted onto the project of researchers. The second, *consultative*, is a form of participation where people are asked for their opinions and consulted by researchers prior to actions being taken. The third form, *collaborative*, involves researchers and local people working jointly on projects formulated, designed, and coordinated by researchers. The final form is termed *collegiate* where researchers and local people work collectively as colleagues, combining their different skills in the cocreation of knowledge and learning. While in theory the goal of many participatory researchers is to adopt this *collegiate* form of participation, in practice very few researchers implement this form throughout the entire research process.

Overall, participatory research offers opportunities to critically explore 'how' data are gathered, 'what sort' of new information and knowledge is gained, and 'what impacts' it has on a community. Three particular attributes of participatory research which often serve to distinguish it from conventional research are community-based analysis of social problems, shared ownership of research projects, and an orientation toward community action (Kemmis and McTaggart, 2008).

Action Research

Action research involves relevant parties engaging in current action with the objective of altering and improving it. In this way this research approach is similar to the commonsensical concept of 'learning by doing.' Action research is not just research which we anticipate will be followed by action; i.e., changes to behavior do not just happen at the end of the research, they can occur throughout the process as the research develops and as participants contemplate and progress their understandings about their behavior. Social psychologist Kurt Lewin is credited with devising the term 'action research' in the 1940s. Drawing on his work with community action programs in the US at that time, Lewin (1946) coined the term to describe the cycles of reflection, action, and problem solving in research.

A review of the literature points toward three vital elements of action research: its participatory character; its democratic impulse; and its contribution to both social science and social change (Meyer, 2000). With regard to participation, an essential element is that participants must be willing to play an active role in the research process. Here it is action that is important, as while the majority of research requires willing participants, the level of commitment required for action research often requires far more than agreeing to answer questions in a survey or to simply be observed. In action research, participants can play an important role in shaping the research process. For example in many action research projects the researcher together with the participants highlight actions to take and collectively they analyze results, critically consider the actions taken and results obtained, and formulate new courses of action. Consequently, this approach is often deemed more democratic than other methods because often the researcher and participants act jointly to realize results for change. In addition to ongoing consultation with participants throughout the research process, a critical feature within many action research projects is that findings can be fed back to the participants during the research as well as at its conclusion. Subsequently, it has been argued that the outcomes of the research are more closely grounded in the reality of day-to-day practice than other methods and have been deemed to be more meaningful to practitioners. Finally, it has been suggested that through action research "academics can learn from what takes place on the ground in order to investigate their own theoretical endeavours and overcome some of the distance between theory and practice" (Halfacree, 1999, p. 209) to generate solutions to practical problems while maintaining attention to broader conceptual practices (Kitchen and Tate, 2000).

Views of action research are constantly changing and developing. While there are many differences between action research and participatory research, researchers are increasingly focusing on their points of convergence. As a result the past two decades have seen the emergence of a plethora of hybrid approaches, including cooperative inquiry, participatory learning and action, emancipatory action research, community-based participatory research, and participatory learning research (Kemmis and McTaggart, 2008; Kindon et al., 2007). Changing a problematic situation for the better by using the insights into everyday activities as voiced directly by the people concerned is a common aim of all these approaches.

Drawing on the strengths of both action research and participatory research approaches, PAR is another approach specifically employed to enhance collaboration, empower participants, and advance knowledge acquisition and social change. As it has an extensive history in many fields of social and behavioral sciences, including social-scientific and interdisciplinary sustainability research, the remainder of this article will focus on PAR.

Guiding Principles and Key Features of PAR

Although definitions vary, most commentators agree that PAR involves practical problem solving in real-life situations where researchers work with, and for, people, as opposed to undertaking research about them. Essentially PAR is guided by four key principles: collaboration through participation; empowerment of participants; garnering of knowledge; and initiation of social change.

PAR is committed to ensuring that the research will benefit the community and contribute to the well-being of the participants, rather than considering them as simply objects of investigation. PAR involves 'research with' as opposed to 'research on' participants/communities providing practical investigation of real people in real contexts, which results in real and concrete applications. A core objective of PAR is to improve the circumstances and well-being of the participants through the research process. PAR is a learning process that strives to take account not only of the actions of people but also the interactions of people among themselves and with the world about them. This research approach explores the history of the people under investigation, in addition to embracing their everyday lifestyles, behaviors, and practices set in their social and cultural context. Power sharing is another central concept underpinning PAR. By reflecting on the information gathered and by understanding the social context and

interaction of the people under investigation, PAR aims to empower people to take control of their futures and to enhance them for the better.

Thus, PAR is committed to the 'opening up' and demystification of research and the destabilization of hierarchical relationships between researchers and participants. It also can help to create a perception of field research as a learning process and an educational tool.

PAR Methods and Outputs

As should be evident from this article thus far, PAR is not simply about data collection. While there is no one single methodology that typifies a PAR approach, a general guideline for methods employed within PAR is the adoption of a structure which adheres to a cyclical process of action and reflection. When designing a PAR project, considerable time is dedicated to planning the process in a constructive collaborative manner among all participants. This initial phase aims to identify the issue in need of change and recognizing the strengths and assets of all the participants. This is then followed by an action stage: utilizing the capability and assets within the group to instigate change and improvement of a situation. The changes outlined in the initial plan are observed on the ground and recorded. All research participants reflect on and evaluate the action process.

PAR methods should be context specific and developed collaboratively between researchers and participants. Consequently a PAR project may employ a range of qualitative and quantitative methods to suit specific objectives. Methods frequently used in PAR can include traditional and innovative approaches such as interviews, surveys, video and photography, mapping (paper maps as well as outputs from GIS), oral narratives and storytelling, participant observation recordings, diaries or journal entries, field-trip excursions, focus groups, and collaborative data analysis.

As it involves a holistic approach to problem solving, research tools and techniques may be combined in innovative ways as a PAR project progresses. Outputs from PAR often represent a diverse range of both traditional and innovative outputs (e.g., results are often disseminated through public workshops and social media platforms in addition to traditional academic journal articles and formal presentations). The coproduction of outputs, such as presentations, video clips, and written reports, is increasingly common within this research approach.

As a primary goal of any PAR process is social change, the importance of producing action-orientated outputs cannot be overstated. Nevertheless, the actual impact of the outputs is equally critical to the success or failure of a PAR process. In fact, PAR could serve to reproduce and even exacerbate existing inequalities unless it actually translates stated goals into tangible actions that contribute to the improvement of participants' lives (Kindon et al., 2007).

PAR Applications

Although used extensively for the past two decades in the field of health sciences, social work, and education, PAR is only emerging as a prominent feature in other fields such as human and social geography, (environmental) sociology, or environmental studies. Within these disciplines participatory research approaches were originally employed primarily to 'give voice' to previously marginalized participants, rather than to generate change in behavior either of the participants or in policy circles (Hoggart et al., 2002). The growth of PAR (with an emphasis on action) in these research fields is evident in research conducted by the Global Action Plan *Eco-teams* project, which was designed with the primary objective of changing household actions with regard to a number of behaviors including energy and water conservation (for more information see www.globalactionplan.org.uk). Similarly, in the field of waste management Fahy and Davies (2007) worked with householders to develop practical ways in which they could change their waste management practices while, at the same time, garnering information from them about the feasibility of such practices for their particular lifestyles.

This research approach was utilized by Mark Garavan (2013) in a project which reveals his experiences in the early- to mid-2000s researching and representing a local community in County Mayo in the west of Ireland whose members have resisted the construction of a large-scale gas pipeline project. His research highlights that local responses to the pipeline development involve culture-specific cognitive and linguistic efforts as well as visceral reactions by people who feel that the future of their place, community, and livelihoods is under threat. This shows how local people's concepts of sustainability are often very different from rational-scientific discourses that dominate most sustainability debates. Throughout the research Garavan argues for a participatory approach to research that takes the concerns and voices of the participants seriously and adopts a long-term view of human life and the research process. There are other numerous calls for the application of PAR to pressing sustainability challenges such as climate change and biodiversity loss (see Reason, 2007).

Considering current and future trends for PAR, technological innovations such as smart phones and other personal technological devices as well as the explosion of social media platforms are opening up new forms of knowledge creation. These innovations have the potential to empower and facilitate people to more actively engage with research activities. This increasing coproduction of knowledge is particularly evident in a number of recent geographical information systems-mapping projects which amalgamate data gathered by individual citizens of their local environments (see for example the Open Street Maps project (Huckley, 2010). Also see Howe (2009) for an overview of crowdsourcing and Goodchild and Glennon (2010) for crowdsourcing applications in the field of disaster responses.

Critiques of PAR

While PAR has the potential to give marginalized groups the ability to generate knowledge and power through research activities, there are also critical voices who view as exaggerated assumptions surrounding the use of the research process as an empowerment tool (Cooke and Kothari, 2001). Indeed, the approach has the potential to reproduce existing inequalities i.e., replicate preexisting power relations within a participating

community. Many projects employing the PAR approach attempt to dispense with traditional roles and titles of researcher and participant. However, this can result in confusion over the role of those assisting others on how to conduct PAR. The illusion of occupying a neutral role while remaining in control of the research project, for example making the decision to cast specific communities or participants as subjects requiring research, is clearly problematic.

Furthermore, detractors of PAR see the potential for new hierarchies to emerge. For example, certain local knowledge could be prioritized simply because of the participatory process that produced it. Similarly if researchers become established as experts in PAR, or indeed if they classify other methods – that are not participatory – as less valid, PAR researchers could be accused of replicating the very hierarchies they aim to challenge. As with any research approach, it is important to adopt a self-reflexive position, acknowledging the limitations of PAR and recognizing that although power is inherent within it and the research process can shape people's behavior, it does so with the aim of providing them with the means to challenge and change more dominant power structures and relations in their lives (see Kindon et al., 2007 for a full discussion on the negative power effects of PAR).

In addition to broader concerns over the use of their techniques and terms and the dangers of becoming co-opted into top-down policy and research initiatives, participatory action researchers face a number of practical challenges when undertaking research.

Challenges and Opportunities Involved in Undertaking Participatory Action Research

As with all research approaches, PAR has a series of practical limitations, including demands on researchers' and participants' time, issues of trust, (lack of) coordination of different skills and levels of participation, and attempts to meet a number of divergent goals (for an in-depth discussion see Breitbard, 2010). PAR approaches can be extremely resource intensive when compared with many other research approaches. In addition to the potential direct costs associated with many intervention and action studies, a PAR approach requires more time and effort on behalf of both researcher and participant than research methods such as one-off interviews or focus group discussions. This research approach is demanding of participants and if individuals are not interested in participating then this may affect the kinds of people willing to do the research and therefore skew results. There are differing degrees of 'action' in PAR and there is a need for balance between how much 'action' you can require of participants and whether they will be willing to participate.

Another possible limitation is the problem of extension. Frequently PAR is conducted on a small scale incorporating case studies. Scaling up these studies may thus prove to be impractical. For example, when Fahy and Davies (2007) utilized household waste minimization exercises to explore waste management behavioral change in Ireland, almost all of the participants in the sample of 10 households recommended that regular home visits were one of the most successful ways of raising awareness about waste and improving waste management behavior. However, conducting such an exercise on a larger scale would require significant investment of time and effort that is impossible without adequate financial support.

Reiterating the importance of action-orientated outputs and the significant role of power processes within PAR, uncertainty surrounding the potential transformative power of the research presents an inherent challenge. Cooke and Kothari in their book *Participation: The New Tyranny* (2001) posit that the theoretical ideal of participation often does not live up to its rhetoric as an instrument for social transformation and a redivision of power. In this book they contend that many initiatives embracing participation result in the maintenance of existing power relationships, i.e., this power is disguised behind the methods and rhetoric of participation.

While social change may be one of the goals of a specific project, it may be impossible from the outset to gauge the impact of the research produced. Even when policy makers and regulators are active coresearchers alongside publics and academics, key decision makers are rarely formally obliged to adhere to any of the recommendations resulting from the research. For example Fahy and O'Cinneide's (2009) research on community mapping in Galway City in the west of Ireland was instigated through a PAR project funded by a national-level agency and mediated by research staff of a local university, with community groups throughout the city actively collaborating with the municipal authority. The research process in addition to generating high levels of public participation and opening up spaces for shared reflections on local areas, resulted in the creation of a map providing grassroots perspectives of the city. It details aspects of life that various publics have identified as important to them. However, even though the municipal authority were central to this project and actively engaged in dialogue with citizens and communities about their local environments, there was no obligation on the council to utilize the data collected in future city planning.

Despite these potential limitations, PAR offers researchers a very rewarding and exciting approach to researching pressing social and environmental sustainability issues. In addition to providing nuanced understandings of lived contexts and everyday actions, the data produced have the potential to initiate actions that address people's real needs and desires.

Conclusion

To conclude, PAR potentially offers a range of opportunities for those involved in social and behavioral research, not least by enhancing learning opportunities for all participants. Overall, this style of research involves more engagement between researcher and participant relative to other methods of research, such as questionnaire surveys. It is not simply another form of consultation; instead projects based on PAR aim to be exemplars of active coresearch, *for* and *by* the researched. This contrasts with more traditional research methods that clearly differentiate between researcher and researched and that place the generation of 'objective' scientific knowledge over and above the (research) needs of those being studied. In addition, PAR can offer a more nuanced understanding of information gathered from other methodological sources. Within the social

and behavioral sciences, investigating understandings of the contextual variables which influence behavior is critical and PAR approaches can assist in providing such understandings.

See also: Attitudes, Political and Civic Culture; Communities of Practice; Deliberation and Democracy; Democratic Transitions.

Bibliography

Arnstein, S.R., 1969. A ladder of citizen participation. Journal of the American Institute of Planners 35 (4), 216–224.
Baumfield, V., Hall, E., Wall, K., 2008. Action Research in the Classroom. Sage, Thousand, Oaks, CA.
Berg, A.M., Eikeland, O. (Eds.), 2008. Action Research and Organisation Theory. Peter Lang, Frankfurt.
Breitbard, M., 2010. Participatory research methods. In: Clifford, N., French, S., Valentine, G. (Eds.), Key Methods in Geography. Sage, London.
Bryant-Lukosius, D., DiCesno, A., 2004. A framework for the introduction and evaluation of advanced practice nursing roles. Journal of Advanced Nursing 48 (5), 530–540.
Cammarota, J., Fine, M. (Eds.), 2008. Revolutionizing Education: Youth Participatory Action Research in Motion. Routledge, New York.
Chandler, D., Torbert, B., 2003. Transforming inquiry and action. Action Research 1, 133–152.
Cooke, B., Kothari, U. (Eds.), 2001. Participation: The New Tyranny?. Zed Books, New York.
Cornwall, A., Jewkes, R., 1995. What is participatory research? Social Science & Medicine 41 (12), 1667–1676.
Fahy, F., Davies, A., 2007. Home improvements: household waste minimization and action research. Resources, Conservation and Recycling 52, 13–27.
Fahy, F., O'Cinneide, M., 2009. Re-constructing the urban landscape through community mapping: an attractive prospect for sustainability? Area 41 (2), 165–175.
Freire, P., 1972. Pedagogy of the Oppressed. Penguin, Harmondsworth.
Garavan, M., 2013. Local lives and conflict: towards a methodology of dialogic research. In: Fahy, Rau (Eds.), Methods of Sustainability Research in the Social Sciences. Sage, London, pp. 72–87.
Goodchild, M.F., Glennon, J.A., 2010. Crowdsourcing geographic information for disaster response: a research frontier. International Journal of Digital Earth 3 (3), 231–241.
Greenwood, D., Levin, M., 2006. Introduction to Action Research: Social Research for Social Change, second ed. SAGE Publications, Thousand Oaks, CA.
Gustavsen, B., 2003. New forms of knowledge production and the role of action research. Action Research 1, 153–164.
Hall, B.L., 2005. In from the cold: reflections on participatory action research from 1970–2005. Convergence 38 (1), 5–24.
Halfacree, K., 1999. Anarchy doesn't work unless you think about it: intellectual interpretation and DIY culture. Area 31, 209–220.
Hinchey, P.H., 2008. Action Research Primer. Peter Lang, Frankfurt.
Hoggart, K., Lees, L., Davies, A., 2002. Researching Human Geography. Arnold, London.
Howe, J., 2009. Crowdsourcing. Random House Business, New York.
Huckley, M., 2010. How good is volunteered geographic information? A comparative study of OpenStreetMap and Ordinance Survey datasets. Environment and Planning B: Planning and Design 37 (4), 682–703.
Kemmis, S., McTaggart, R., 2008. Participatory action research. In: Denzin, N.K., Lincoln, Y.S., Smith, L.T. (Eds.), Handbook of Critical and Indigenous Methodologies. Sage, Thousand Oaks, CA.
Kindon, S., Pain, R., Kesby, M. (Eds.), 2007. Participatory Action Research Approaches and Methods: Connecting People, Participation and Place. Routledge, London.
Kitchen, R., Tate, N.J., 2000. Conducting Research in Human Geography: Theory, Methodology and Practice. Prentice Hall, New York.
Lewin, K., 1946. Action research and minority problems. Journal of Social Issues 1-2, 34–36.
McIntyre, A., 2008. Participatory Action Research. Sage, Thousand Oaks, CA.
Meyer, J., 2000. Using qualitative methods in health related action research. British Medical Journal 320, 178–181.
Reason, P., Bradbury, H. (Eds.), 2013. Handbook of Action Research, second ed. Sage, London.
Reason, P., 2007. Education for ecology: science, aesthetics, spirit and ceremony. Management Learning 38 (1), 27–44.
Reeb, R.N. (Ed.), 2006. Community Action Research: Benefits to Community Members, and Service Providers. Haworth Press, Binghamton.
Stringer, E., 2013. Action Research, fourth ed. Sage, London.
Whitehead, D., Taket, A., Smith, P., 2003. Action research in heath promotion. Health Education Journal 63 (4), 347.

Relevant Websites

Action Learning, Action Research Association (ALARA), available at: www.alarassociation.org, ALARA is a network of people interested in using action research and action learning to generate collaborative learning, research and action to transform workplaces, schools, colleges, universities, communities, voluntary organisations, governments and businesses.
Action Research Journal (ARJ), blog available at: www.arj.sagepub.com, The ARJ blog is the web log for the print publication Action Research, which is a forum for the development of the theory and practice of action research.
Global Action Plan, available at: www.globalactionplan.org.uk.

Relevant Online Journals

Action Learning published by Taylor & Francis, available at: www.tandf.co.uk.
Educational Action Research published by Taylor & Francis, available at: www.tandf.co.uk.
International Journal of Action Research published by Hampp, available at: www.hampp-verlag.de.
Participatory Learning and Action published by IIED, available at: www.iied.org.
Systematic Practice and Action Research published by Springer, available at: www.springerlink.com.

Participatory Action Research in Social Research

María Elena Torre, The Graduate Center, City University of New York, New York, NY, USA
Caitlin Cahill, Pratt Institute, Brooklyn, NY, USA
Madeline Fox, Brooklyn College, New York, NY, USA

Published by Elsevier Ltd.

Abstract

Participatory approaches to doing research have gained critical attention from across the behavioral and social sciences, as scholars, activists, and practitioners committed to social change work closely with communities to investigate their concerns, develop proposals for transformative change, and identify new questions to investigate. Rooted in principles of justice and democracy, participatory action research (PAR) is an inclusive, collaborative approach to research defined both by participation and a determination to produce knowledge in the interest of social change. Within the social sciences, PAR is a stance committed to engaging knowledge and expertise beyond the 'ivory tower' by involving those who are most intimately affected by the research in shaping the research questions, framing interpretations, and designing meaningful research products and actions.

> The silenced are not just incidental to the curiosity of the researcher but are the masters of inquiry into the underlying causes of the events in their world. In this context research becomes a means of moving them beyond silence into a quest to proclaim the world.
>
> Paulo Freire, 1982: 30–31

Definition

Rooted in principles of justice and democracy, participatory action research (PAR) is an inclusive, collaborative approach to research defined both by participation and a determination to produce knowledge in the interest of social change. Within the social sciences, PAR is a stance committed to engaging knowledge and expertise beyond the 'ivory tower' by involving those who are most intimately affected by the research in shaping the research questions, framing interpretations, and designing meaningful research products and actions. Marked by an inclusive orientation, PAR expands traditional notions of expertise, repositioning those who have been traditionally 'researched' and excluded from the academy (historically oppressed groups including people of color, young people, people with disabilities, LBGTQ, undocumented, incarcerated, poor, working class) as research partners. More than simply a method, PAR is an epistemological position grounded in everyday life experience and a commitment to the 'right to research' (Appadurai, 2006). In this way PAR reflects a critical challenge to positivist epistemology and academic research traditions that assume knowledge to be produced by credentialed experts who are qualified to measure and analyze an objective reality. In contrast, PAR is founded upon a belief that all knowledge is collectively produced. Decentering individualized expertise, PAR draws from situated knowledge and assumes that the subjects and agents of knowledge are multiple, heterogeneous, contradictory, embodied, and socially constructed.

Participatory approaches to doing research have gained critical attention from across the behavioral and social sciences, as scholars, activists, and practitioners committed to social change work closely with communities to investigate their concerns, develop proposals for transformative change, and identify new questions to investigate (Kindon et al., 2007). At the same time critical scholars have pointed to the ways that broad applications of the term 'participation' may mask tokenism and provide an illusion of consultation (Cooke and Kothari, 2001; Arnstein, 1969) when participatory research is presented as a set of techniques rather than a commitment to working *with* communities, as this may result in the reproduction, rather than the challenging, of social inequities.

The important debates and questions within the field of PAR revolve around the intertwined relationship of knowledge and power. More recently, there have been moves to reassert the critical underpinnings of PAR and highlighting its liberatory history and connection to feminist, queer, critical race, neo-Marxist, and indigenous theories (Torre et al., 2012). Critical participatory action research (CPAR) centers this analysis, joining social movements as a transformative social justice project committed to documenting the grossly uneven structural distributions of opportunities, troubling ideological categories projected onto communities, demonstrating how science has been recruited to legitimate dominant policies and practices. To this end, CPAR has profound implications for rethinking the politics of representation and challenging what Foucault (1980) identified as the 'subjectifying social sciences.' Taking seriously the intertwining of 'the global and the intimate' (Pratt and Rosner, 2006), CPAR reflects a commitment to map out the relationships between social structures and injustice in everyday life experiences. Critical PAR considers how questions of political economy, geopolitical context, and identity (race, class, gender, sexual orientation, and status) are made meaningful in a particular context and raise up questions of vulnerability and power.

Epistemological and Ontological Orientation of PAR

Drawing upon situated knowledge and life experiences, the epistemology and ontology of PAR challenge traditional approaches to research through its emphasis on collaboration and the integration of theory and practice. In this way, a participatory approach takes seriously what it means to do

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10554-9

social science (c kinpaisby-hill, 2011). By working collectively with coresearchers in participatory ways, PAR shifts the often hierarchical social relations within research that produce knowledge in isolation from the very communities who are most affected by the research. Recognizing that all knowledge is socially produced, a participatory epistemology is based on an ontology that affirms the capacity of human beings for creative thinking, analysis, and creativity. Recognizing the connectedness of knowing, doing, and being, participatory praxis destabilizes dualisms between theory and practice offering 'an alternative ontology of theorizing' that moves between social constructions and embodied and emotional experiences (ibid). As c kinpaisby-hill explains (2011: 223), PAR "is a citational process, a relational bricolage of one's own and other's ideas, and it is an embedded and constantly ongoing praxis of coming to knowing through iterative cycles, moving between experiences of everyday social life and individual and collective analysis and reflection."

Historical Roots

Building upon the long-standing traditions of grassroots social movements, critical pedagogy and activism, and legacies of critical race, decolonizing, and feminist theories (Smith, 1999; Kelley, 1997; hooks, 1990; Torre, 2009; Freire, 1970; Fals Borda and Rahman, 1991), PAR's historical roots traces different interdisciplinary lineages from around the world, from liberation theology, to critical psychology, sociology, and geography (Martin-Baro, 1994; Lewin, 1946; Wormser and Selltiz, 1951; Du Bois, 1898; Bunge, 1971).

Within the United States, the theory and methods of action-oriented research in the social sciences is often attributed to the work of the social psychologist Kurt Lewin. His seminal article, "Action research and minority problems" published in 1946 insisted that 'minority problems' were indeed 'majority problems' and that social science conducted through collective research should be aimed at changing communities and social institutions. Lewin's description of action research involved a spiral of self-reflective cycles of fact-finding, action, observation, evaluation, and then replanning and more fact-finding, action, and evaluation. Convinced of the important role of social science in informing social change, Lewin established Commission on Community Interrelations, an early university partnership with an advocacy organization, dedicated to developing research and methodologies wherein community groups joined researchers as 'fact-finders' to study 'real-life' situations and produce knowledge that would affect social change. One such method, formalized through the leadership of Margot Wormser and Claire Seltiz, was the Community Self-Survey that brought together diverse teams of community members and researchers to engage in large-scale participatory research on discriminatory practices in housing, employment, education, and public services in their communities (see Torre and Fine, 2011). These surveys echoed the early sociological work of Jane Addams and the residents of Hull House in Chicago and W.E.B. Du Bois in Philadelphia and Atlanta that took place in the 1890s. Addams and Du Bois rejected contemporary frames centered on the 'immigrant' or 'Negro problem' and positioned individuals and groups as pathological or problematic. Instead they each argued for structural analyses of race- and class-based oppression, asking what are the policies, institutions, and social arrangements that enrich and limit human development? With this new framing, the 'problem' became one of classed and racialized patterns of housing, sanitation, education, criminal justice, and access to health care and healthy environments.

Other roots of PAR that have nourished present practices of PAR lie within liberatory popular education movements, particularly those in the southern United States and Latin America. Working with communities in the Appalachian Mountains, Miles Horton, a co-founder of the Highlander Research and Education Center used participatory education and research to organize communities around worker and civil rights beginning in the 1930s. With inspiration from the Hull House, Horton's work emphasized community participation, arguing that effective institutional change would come about only with solutions determined by people who were directly experiencing the problem and who would be most affected by any actions taken (Lewis, 2006). Further south in Colombia, the sociologist Orlando Fals Borda (1977) long argued for the revolutionary potential of a 'people's science' – one that weaved together popular education, community organizing, and social science. Drawing on Lewin's action research, Fals Borda was the first to expressly articulate a praxis of *participatory* action research in his work with villagers in Boyaca, Colombia. Committed to the liberatory potential of PAR, Fals Borda led the organizing effort that resulted in the world's first conference explicitly about PAR in Cartagena in 1977.

At the heart of PAR's South American lineage is Paulo Freire's (1970) participatory approach to social change through conscientization, a process wherein people develop critical consciousness through a reflexive process of collective inquiry, reflection, and action on the economic, political, and social contradictions they are embedded in. Deeply committed to what Antonio Gramsci (1971) called 'organic intellectuals,' Freire, Fals Borda, and their colleagues argued that a research of the people by the people, that encouraged critical reflection on one's experiences, had the power to generate theory, design research, and engage action that could interrupt the oppressive conditions of the status quo.

PAR has continued to be developed in and across universities and communities in South America and around the world in countries such as Australia, Bangladesh, Canada, England, India, and the United States, among others.

Ethics

The epistemological stance of PAR has profound implications for rethinking ethical commitments which in turn, raise new questions for theory and practice. While many researchers choose to engage in PAR for ethical reasons, doing so does not circumvent ethical dilemmas (Manzo and Brightbill, 2007; Cahill et al., 2007; Tuck and Guishard, 2013). Indeed participatory research, informed by principles of inclusive democratic participation, explicitly engages the politics of collaboration, positionality, power, and accountability that are too often not addressed in traditional research practices. Participatory ethics are necessarily collective and take seriously questions of

collaboration in practice. Rather than top down, ethics are emergent and negotiated with participants. Guided by principles of inclusive democratic participation, a participatory ethics is characterized by mutual respect, relationality, humility, reflexivity; dignity, and collective action (cf Tuck and Guishard, 2013; Banks et al., 2011).

Motivated by a vision for 'what could be,' and the commitment to negotiating asymmetries of power, privilege, and knowledge production, participatory praxis throws into sharp relief the tensions between institutionalized and participatory ethics. Highlighting an epistemological shift, ethical questions that are traditionally founded upon the assumption of an individualized autonomy of institutional review processes are reassessed through a collective negotiation that reframes issues of consent, protection, and vulnerability within larger sociopolitical contexts. In this vein, participatory ethics are understood to be culturally and contextually specific, situated, and accountable to place and community (in all of its heterogeneity).

Ethics are traditionally framed around discussions of the overall harms and benefits of research, the rights of participants to information, privacy, confidentiality, and the integrity of researchers (Banks et al., 2011). PAR, however, opens up new and expansive ways of thinking about ethics, guided by its commitments to collaboration, transformative change, addressing unequal relations of power, and a retreat from a dispassionate stance. Beyond doing no harm, "a PAR inspired understanding of social justice suggests that it is in fact unethical to look in on circumstances of pain and poverty and yet do nothing" (Manzo and Brightbill, 2007: 35). Put another way, PAR recognizes research as a site of contestation in itself, and therefore participatory ethics are a commitment to disrupt the status quo, to challenge hegemonic practices and discourses, and to 'theorize back' (Tuck, 2009). In this sense, a participatory ethics can be understood as an ethical social scientific stance against neutrality. Informed by an 'ethic of care' in its most profound sense, participatory ethics reflect deep a commitment to justice and respect for relationships and humanity (Ellis, 2007).

Methodological Considerations and Commitments

The epistemological and ethical commitments of PAR fundamentally shape methodological and praxis decisions throughout the research process. PAR projects engage a wide range of methodologies for data collection and analysis, including both qualitative and quantitative approaches. At best, PAR designs include multiple and triangulated methods to capture interconnected individual, social, institutional, and cultural layers of the phenomena, being studied. Participatory methodologies include (but are not limited to) surveys, interviews, focus groups, archival data, maps, oral histories, digital stories, participatory video, photo-voice, participant observation, cross-site visits, 'stats in action,' and 'sidewalk science' (cf Stoudt and Torre, 2014).

While PAR projects are not restricted to particular methodologies, they share an orientation that informs decision-making throughout the research process. As a research practice *with*, rather than *for* or *on* communities, PAR revolves around a shared set of commitments for creating conditions of collaboration within research collectives. Articulations of these conditions differ depending on the situation and context but they all underscore the following commitments (cf 'Principles' www.publicscienceproject.org):

- to include knowledges and expertise that have been historically marginalized and delegitimized;
- to exchange expertise and share resources within the research collective;
- to an ongoing negotiation of issues of power and attention to vulnerability within the participatory collective and created by the research and associated actions;
- to creation of a research space wherein individuals and the collective can express their multiplicity and use this multiplicity to inform research questions, design, and analyses;
- to excavate and surface disagreements in the interest of consensus as a lens into larger social/political dynamics informing the research;
- to conduct intersectional analyses at the individual, social, historical, cultural, and institutional levels;
- to conceive multiple forms of action over the course of the PAR project;
- to an ongoing negotiation of conditions of collaboration.

The epistemological, methodological, and political commitments of PAR raises and complicates the questions of objectivity and validity that are arguably relevant to all social research, but that PARs make explicit. PAR challenges, redefines, and transforms traditional understandings of these concepts. To begin, approaching research methods through a PAR framework adopts what feminist scholar Sandra Harding (1993) calls 'strong objectivity,' which honors the rigor of situated knowledge and committing to exploring competing perspectives through democratic participation.

Building off of traditional conceptualizations, PAR reframes validity revising metrics of expertise and offering new definitions. First, engaging the iterative exchange between critical theory and research findings, PARs are concerned with *construct validity* that is the extent to which an analysis recognizes the impact of history and structures of injustice on individual and collective levels. *Ecological validity* is a new construct that theorizes the way a 'problem' can be understood at multiple scales and levels of analysis, from the intimate to the global, or the micro to the macro. *Expert validity* reflects deep participation, contesting and opening up traditional views of expertise, by engaging diverse standpoints while recognizing the relationship between situated knowledges and systems. Finally, *impact validity*, attending to what type of impact the research has on the injustice the research is centered around, is designed to be accountable and in solidarity with movements for social change (Torre et al., 2012; Fine, 2008). Breaking the monopoly on who holds knowledge and for whom social research should be undertaken, PAR offers new ways of thinking about methodologies, science, and expertise.

Purpose, Publics, and Products

While all research must ask questions of purpose, PAR features these questions explicitly from the start. Intertwined questions of purpose, publics, and products are revisited and

collaboratively interrogated throughout the PAR process reflecting the commitment to produce research that addresses urgent questions as defined by those most impacted. Indeed, it is the purpose of the PAR project that determines the design of the research and brings co-researchers together in the first place. Questions of purpose, publics, and products encouraged within PAR push the research collective to ask: What is the research trying to accomplish? What audiences is the research trying to engage? And, how might the research best communicate to a specific audience?

Participatory collectives attend to and negotiate these questions – often recognizing multiple perspectives on the goals of the project – that demand diverse publics, and therefore the development of a variety of research products. Multiple products not only provide different ways of speaking to different publics (be they academics, policymakers, young people, educators, etc.), they also create a way of attending to the 'thickness' of research and the different levels of analysis that were explored.

PARs are generally committed to making a meaningful contribution to social change beyond an 'armchair revolution' (Freire, 1970). This means that PAR collectives regularly attempt to reach publics beyond the academy. Products typically include: scholarly articles that speak to (or back to) ongoing academic debates; policy reports; popular journalism pieces; data posters, postcards, and infographics; performances and documentaries; websites and social media campaigns. Regardless of form, research collectives take seriously the political nature of knowledge production and the question of impact: How might the analysis inform a sense of responsibility and incite the possibility for change? What is the research asking audiences to do? Will interpretations be individually authored or collective voice? Who is made vulnerable by the research? What might be too sacred or dangerous to make public? And, finally, how might the research effectively provoke actions that feed organizing campaigns, theoretically reframe the issues, nudge those in power, and/or motivate audiences to shift how they think and act in the world? The contextually specific ways these questions are answered speak to PAR's commitments to blend systematic inquiry, democratic engagement, and liberatory organizing in order to engage research for justice (Cahill and Torre, 2007).

Future of the Field

Critical issues of funding, tenure, and promotion must be addressed if the field of PAR is to continue its growth within the academy. While disciplines such as public health have embraced PAR with robust funding streams dedicated for what is often called 'community-based participatory research' the CBPR, this is not true in the social sciences more generally. Despite a burgeoning interest across the disciplines, within the university questions remain in evaluation committees about 'impact factors' and how to value multiple authorship and research products that are not in the form of peer-reviewed journal articles. There are some positive recent developments however as there is a push toward civically engaged scholarship within the academy and for institutions of higher education to contribute to the communities they are located within – rethinking the role and moral responsibility of the academy in desperately unequal times. Communities are increasingly demanding reciprocity with university researchers and clarity with regard to the questions of purpose and publics issues that are well addressed by PAR projects. In addition, research centers are beginning to spring up that bridge universities and communities, some offering training institutes on PAR methodologies as well as supporting PAR projects. As a result of these trends, it is now standard to find PAR theory and methods included in volumes on research methodology, and feminist, critical, and community-based research approaches across the social sciences. While there is a way to go before PAR is fully embraced by the academy, there is a growing body of scholarship, and increasing interest in creating critical opportunities for PAR as public scholarship, since it offers a strategic and thoughtful approach to strengthening the social compact between the academy and larger society.

See also: Addams, Jane (1860–1935); Applied Social Research, History of; Collective Action; Critical Psychology; Critical Race Theory; Du Bois, William Edward Burghardt (1868–1963); Lewin, Kurt (1890–1947); Mixed Methods; Postcoloniality; Reflexivity; Research Ethics: Research; Social Justice: Historical and Theoretical Considerations.

Bibliography

Addams, J. (Ed.), 1896. Hull-House Maps and Papers: A Presentation of Nationalities and Wages in a Congested District of Chicago, Together with Comments and Essays on Problems Growing Out of the Social Conditions.

Appadurai, A., 2006. The right to research. Globalisation, Societies and Education 4 (2), 167–177.

Arnstein, S.R., 1969. A ladder of citizen participation. Journal of the American Institute of Planners 35 (4), 216–224.

Banks, S., et al., 2011. Community-Based Participatory Research: Ethical Challenges. AHRC Connected Communities.

Bunge, W., 1971. Fitzgerald: Geography of a Revolution. Schenkman, Cambridge, MA.

Cahill, C., Sultana, F., Pain, R. (Eds.), 2007. Participatory Ethics. ACME: An International E-Journal for Critical Geographies. 6 (3) (special issue).

Cahill, C., Torre, M.E., 2007. Beyond the journal article: representations, audience, and the presentation of participatory research. In: Kindon, Sara, Pain, Rachel, Kesby, Mike (Eds.), Participatory Action Research Approaches and Methods: Connecting People, Participation and Place. Routledge.

Cooke, B., Kothari, U., 2001. Participation: The New Tyranny? Zed Books, London.

Du Bois, W.E.B., 1898. The Study of the Negro Problems. American Academy of Political and Social Science, Philadelphia.

Ellis, C., 2007. Telling secrets, revealing lies: relational ethics in research with intimate others. Qualitative Inquiry 13 (1), 3–29.

Fals Borda, O., 1977. Por la praxis: El problema de cómo investigar la realidad para transformarla. In: Fals-Borda, En O. (Ed.), Crítica y Política en Ciencias Sociales. Punta de Lanza, Bogotá, Colombia.

Fals Borda, O., Rahman, M.A., 1991. Introduction. In: Action and Knowledge: Breaking the Monopoly with Participatory Action-Research. The Apex Press, New York, pp. 3–34.

Fine, M., 2008. An epilogue, of sorts. In: Cammarota, J., Fine, M. (Eds.), Revolutionizing Education: Youth Participatory Action Research in Motion. Routledge, New York, pp. 213–234.

Foucault, Michel, 1980. In: Colin, Gordon (Ed./Trans.), Power/Knowledge: Selected Interviews & Other Writings 1972–1977. Pantheon Books, New York.

Freire, P., 1970. Pedagogy of the Oppressed. Herder and Herder, New York.

Freire, Paulo, 1982. Creating alternative research methods. Learning to do it by doing it. In: Hall, Budd, Gillette, Arthur, Tandon, Rajesh (Eds.), Creating Knowledge: A Monopoly. Society for Participatory Research in Asia, New Delhi, pp. 29–37.

Gramsci, A., 1971. Prison Notebooks. International Publishers, New York.

Greene, M., 1988. The Dialectic of Freedom. Teachers College Press, New York.
Harding, S., 1993. Strong objectivity: a response to the new objectivity question. Synthese 104 (3), 331–349. http://dx.doi.org/10.1007/BF01064504.
hooks, bell, 1990. Yearning: Race, Gender, and Cultural Politics. Between-the-Lines, Toronto, Canada.
Kelley, R., 1997. Yo' Mama's Disfunktional! Fighting the Culture Wars in Urban America. Beacon Press, Boston, MA.
Kindon, S., Pain, R., Kesby, M. (Eds.), 2007. Participatory Action Research Approaches and Methods: Connecting People, Participation and Place. Routledge, London.
kinpaisby-hill, c., 2011. Participatory praxis and social justice: towards more fully social geographies. In: Del Casino Jr., V., Thomas, M., Panelli, R., Cloke, P. (Eds.), A Companion to Social Geography. Wiley-Blackwell, Oxford.
Lewis, H.M., 2006. Participatory research and education for social change: highlander research and education center. In: Reason, P., Bradbury, H. (Eds.), Handbook of Action Research, Concise Paperback ed. Sage Publications, London, pp. 262–268.
Lewin, K., 1946. Action research and minority problems. Journal of Social Issues 2 (4), 34–46.
Manzo, Lynn, Brightbill, Nathan, 2007. Towards a participatory ethics. In: Kindon, Sara, Pain, Rachel, Kesby, Mike (Eds.), Connecting People, Participation and Place: Participatory Action Research Approaches and Methods. Routledge, London, pp. 33–40.
Martín-Baró, I., 1994. Writings for a liberation psychology. Harvard University Press, Cambridge, MA.
Pain, R., 2008. Ethical possibilities: towards participatory ethics. Children's Geographies 6, 104–108.
Pratt, G., Rosner, V., 2006. Introduction: the global and the intimate. Women's Studies Quarterly 34 (1, 2), 13–24.
Reason, P., Bradbury-Huang, H., 2007. Handbook of Action Research: Participative Inquiry and Practice. Sage, London.
Stoudt, B.G., Torre, M.E., 2014. Criminal justice and participatory action research. In: The Encyclopedia of Action Research. Sage.
Smith, L.T., 1999. Decolonizing Methodologies: Research and Indigenous Peoples. Zed Books, Ltd, New York.
Torre, M.E., 2009. Participatory action research and critical race theory: fueling spaces for nos-otras to research. The Urban Review 41 (1), 106–120.
Torre, M.E., Fine, M., Stoudt, B., Fox, M., 2012. Critical participatory action research as public science. In: Camic, P., Cooper, H. (Eds.), The Handbook of Qualitative Research in Psychology: Expanding Perspectives in Methodology and Design, second ed. American Psychological Association, Washington, DC, pp. 171–184.
Torre, M.E., Fine, M., 2011. A wrinkle in time: tracing a legacy of public science through community self-surveys and participatory action research. Journal of Social Issues 67 (1), 106–121.
Tuck, E., 2009. Re-visioning action: Participatory action research and indigenous theories of change. The Urban Review 41 (1), 47–65. http://dx.doi.org/10.1007/s11256-008-0094-x.
Tuck, E., Guishard, M., 2013. Uncollapsing ethics: scientifically based research and settler coloniality: an ethical framework of decolonial participatory action research. In: Kress, T., Malott, C., Porfilio, B. (Eds.), Challenging Status Quo Retrenchment, pp. 3–27.
Wormser, M.H., Selltiz, C., 1951. How to Conduct a Self-Survey of Civil Rights. Association Press, New York.

Relevant Websites

http://www.acme-journal.org/volume6-3.html – ACME Special Issue on Participatory Ethics.
https://www.dur.ac.uk/beacon/socialjustice/toolkits/ – Center for Social Justice & Community Action (Toolkits, Guides & Case Studies).
http://www.datacenter.org/ – The Data Center.
http://jgc.stanford.edu/resources/tools.htm – The John Gardner Center for Youth and their Communities.
http://www2.bc.edu/~lykes/research.htm – M. Brinton Lykes's Participatory Action Research Resource Website (Tools, Websites, and Links to Journals).
http://www.publicscienceproject.org/ – The Public Science Project.
http://www.pygyrg.co.uk/ – Participatory Geographies Research Group.

Partnership Formation and Dissolution in Western Societies

Brienna Perelli-Harris, School of Social Sciences, University of Southampton, Southampton, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by K. Kiernan, volume 16, pp. 11092–11099, © 2001, Elsevier Ltd.

Abstract

Partnership formation and dissolution in Western countries has changed dramatically over the last decades. Marriage has declined while cohabitation and childbearing in cohabitation has increased. Union dissolution, living apart together, and repartnering have become common. Despite the nearly universal increase in these behaviors, patterns of union formation continue to differ across Europe and the United States. This article briefly describes the correlates and consequences of these behaviors and the difficulties in disentangling selection effects and causality. It outlines the main explanations for developments in union formation and suggests historical, cultural, and legal factors that may account for country differences.

Over the past few decades, the process of partnership formation has become much more complex in most Western nations. No longer do individuals simply enter marriage and remain married for life. Instead, individuals are likely to enter into unmarried, coresidential relationships that may persist, convert to marriage, or end in dissolution. Individuals may experience a single union that lasts a lifetime or have multiple unions over the lifecourse. Marriage may occur early, be delayed until after childbearing, or never occur at all. Thus, the timing, duration, and repetition of cohabiting and marital unions have become very complex. Here, these developments and their implications are examined for a range of Western nations with a particular focus on Europe and the United States.

Partnership Formation

Marriage Trends

Since the 1970s, marriage rates in most Western countries have declined dramatically. In 27 countries of the European Union, the crude marriage rate fell from 7.9 marriages per 1000 inhabitants in 1970 to 4.4 marriages per 1000 inhabitants in 2010, equivalent to a 36% decline in the absolute number of marriages (Eurostat, 2013). Crude marriage rates in the US were considerably higher in 2010 at 6.8 per 1000 inhabitants (Centers for Disease Control, 2013). Although marriage rates have been declining across Europe, the US, and Australia, the rates of decline have differed substantially by region. Marriage rates first started declining in Sweden and Denmark in the late 1960s, declined throughout most of Western Europe in the early part of the 1970s, and began declining in the southern European countries (Spain, Italy, Portugal, and Greece) around the mid-1970s. Central and Eastern Europe, however, had only gradual declines in marriage rates until the collapse of socialism in the 1990s, when marriage rates plunged. Recently, marriage rates across Europe have stabilized at lower levels or even experienced small increases (Eurostat, 2013). In fact, in the late 1990s and 2000s, Sweden experienced a 'marriage revival,' with an upswing in marriage rates.

Some of the decline in marriage has been due to the postponement of marriage to older ages. In 1975, the average ages of first-time brides in Western Europe were clustered in the 22–24 years range; by the 2000s they were between 27 and 29 years. In Central and Eastern Europe, the mean age at marriage for women was between 20 and 24 years until the 1990s, when the mean age at marriage rose by about 2 years. The pattern of change was similar for men, but on average, men marry 2 years later than women, an age difference that has been fairly stable over time. As couples marry later in life, marriage rates may flatten or even recuperate. However, the decline in marriage rates is also due to the increase in unmarried partnerships that may be replacing marriage or never entering a partnership at all (*see* Transition to Adulthood).

Cohabitation

One of the most striking changes to partnership formation has been the rapid and nearly universal increase in cohabiting unions throughout Europe and the United States. A small percent of couples did live together without marrying before the 1970s, particularly after divorce, but the recent increase suggests that cohabitation is transforming the process of union formation. Over the past few decades, researchers have struggled to understand and classify cohabitation using a variety of comparison groups and criteria (Smock, 2000). Depending on criteria analyzed, cohabitation can be a prelude to marriage, a stage in the marriage process, an alternative to being single, an alternative to marriage, or even indistinguishable from marriage (Heuveline and Timberlake, 2004). Some research investigates whether cohabitation is an alternative to being single or married by analyzing individual characteristics and resources (Rindfuss and VandenHeuvel, 1990), or the duration of the union and whether it dissolves or transitions into marriage (Heuveline and Timberlake, 2004). Others have investigated the role of pregnancy (Raley, 2001) and childbearing (Perelli-Harris et al., 2012) in decisions to cohabit or marry. It is important to keep in mind that the meaning of cohabitation differs considerably based on social, cultural, religious, and policy context.

The Incidence of Cohabitation

Unlike marriage, which must be officially registered with state agencies, cohabitation is rarely officially recorded, although in

some countries partners can register their unmarried partnerships, which would be enumerated in official records. Within the past 10–20 years, however, cohabitation has also been recorded in housing registers in some Scandinavian countries. Most of the data on cohabitation in Europe come from surveys, which can be subject to recall error and vague definitions of what it means to be living with a partner. Nonetheless, survey data do provide general information about this behavior. The European Social Survey, which provides cross-sectional information on living arrangements in 2010, asked respondents living with a partner whether they are married, in a legally recognized partnership, or in a partnership not legally recognized (they could also report being separated or divorced). The category 'legally recognized partnership' differs by country. Some countries, such as Belgium, the Netherlands, and France, require partners to officially register their partnerships in order to gain rights. Other countries have no registration system, but respondents may perceive that their partnership is legally recognized. Whether the relationship is actually recognized in law differs by country and depending on policy area (Perelli-Harris and Sanchez Gassen, 2012). In some countries, respondents may think their partnership is legally recognized, even though they may not have the same rights and responsibilities as married couples. In addition, the law may differ for heterosexual and homosexual couples, with same-sex couples having access to legally recognized unions, but not marriage (Table 1 does not distinguish between hetero- and homosexual partnerships).

Table 1 shows the distribution of types of unions for respondents aged 15–45 currently living with a partner in 26 European countries in 2010. First note that the majority of respondents in most countries are married; only Sweden and Estonia have fewer than 50% of couples not married. Second, in most countries, few respondents report being in legally recognized unions, except in the Netherlands, where registered partnerships have become much more prevalent (20.55%) and Sweden, where over half (54.41%) report that their unions are legally recognized (nonetheless, not all laws in Sweden are the same for cohabiting and married respondents – see Perelli-Harris and Sanchez Gassen, 2012). A substantial proportion in most countries reports living in unions that are not legally recognized, ranging from over 40% in Scandinavia, around a third in Northern and Central Europe, to about a fifth in Eastern Europe, although Poland and Ukraine only have about 12% of respondents living in cohabitation. Some parts of Southern Europe, particularly Greece and Croatia, have fewer than 10% of couples living together without being married, indicating very low uptake of these new living arrangements. Overall, however, it is striking at how common cohabitation has become across Europe.

In order to better show the increase in cohabitation over time, Figure 1 shows the proportion of women who have ever cohabited for birth cohorts of women from 1930 to 1979 for a selected set of European countries and the US. The data come from nationally representative surveys in each country conducted between 2003 and 2010. Figure 1 demonstrates the diversity of behaviors across countries and differences in the rate of change; for example, the increase was very fast in some countries such as Norway and the UK, while much slower to increase in countries such as Bulgaria and the Netherlands.

Table 1 Relationship status of men and women age 15–45 currently living with a partner in 2010

	Married	*Legally recognized union*	*Not legally recognized union*	*N*
Belgium	63.72	17.91	18.37	430
Bulgaria	70.09	0	29.91	446
Cyprus	87.72	1.09	11.19	274
Czech Republic	78.09	0	21.91	546
Denmark	65.75	1.37	32.88	365
Estonia	49.64	14.84	35.52	411
Finland	57.92	0.9	41.18	442
France	57.58	9.87	32.55	383
Germany	74.24	0.73	25.03	635
Greece	92.05	1.03	6.92	656
Croatia	94.48	1.29	4.23	323
Hungary	66.39	5.51	28.1	395
Ireland	74.83	4.76	20.41	582
Italy	90.44	9.56	0	632
Lithuania	86.4	0	13.6	288
The Netherlands	63.51	19.59	16.91	454
Norway	52.09	0.85	47.06	428
Poland	89.35	0	10.65	477
Portugal	76.83	10.68	12.49	362
Russian Federation	82.48	0	17.52	616
Sweden	46.45	53.55	0	366
Slovenia	63.46	13.95	22.59	301
Slovakia	80.19	0	19.81	375
Spain	72.64	6.28	21.08	481
Switzerland	76.9	2.82	20.28	355
Ukraine	87.56	0	12.44	441
The UK	67.99	1.06	30.95	522

Note: Results do not include respondents who reported being separated/divorced from their partner.
Weights applied where available.
Source: European Social Survey 2010.

Note that the percent of women ever cohabiting is higher than those currently living outside of marriage (shown in Table 1), suggesting that many women cohabit and then marry or experience cohabitation after divorce.

Type of First Partnership

Cohabitation is becoming increasingly prevalent as the way in which unions begin; couples often move in together before marriage rather than directly marrying. Column 1 of Table 2 shows the proportion of women who started their first partnership with cohabitation for cohorts born in the 1960s and 1970s. In most of Northern and Central Europe, direct marriage is becoming outdated: over 75% of unions start with cohabitation, especially those of the younger cohorts. In the United States for the 1970 cohort, nearly 70% of unions begin with cohabitation. In parts of Southern and Eastern Europe which traditionally had low levels of cohabitation, for example, Spain and Russia, nearly half of all unions now start with cohabitation. Even in parts of Europe where the diffusion of cohabitation was said to have stalled (i.e., Italy and Romania), a quarter of unions start with cohabitation. Note, however, that the 1970s cohort includes women who would have only been in their early twenties at the time of the interview, which may overestimate the percent cohabiting.

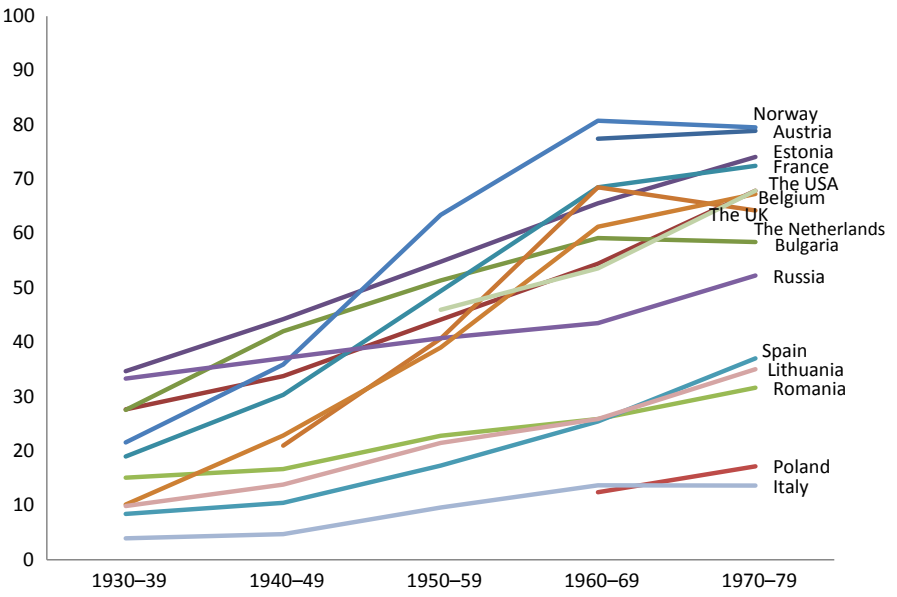

Figure 1 Proportion of women who ever cohabited by birth cohort. Note: weights applied where available. Sources: Generations and Gender Surveys in Austria (2008–09), Belgium (2008–10), Bulgaria (2004), Estonia (2004–05), France (2005), Italy (2003), Lithuania (2006), Norway (2007–08), Romania (2005), and Russia (2004); Fertility and Family Survey in the Netherlands (2003); British Household Panel Survey (2005–06); Employment, Family, and Education Survey in Poland (2006); Spanish Fertility Survey (2006); U.S. National Survey of Family Growth (2006–08).

Women who delay marriage may be more likely to directly marry, but we cannot know for sure until more recent data become available.

Duration of Cohabiting Unions

Although cohabitation is becoming the normative way of starting a union, marriage does still occur for the majority of couples in nearly all countries. Column 2 of Table 2 shows the proportion of women who marry their first partner within 5 years of starting to live together. For the 1960s cohort, more than half of women married their first partners within 5 years in all countries except France, Norway, and Spain. For the 1970s cohort, far fewer women had married their partners in most countries, but again it is important to note that results may not reflect ever marrying, because this younger cohort had less time to marry before the interview. In any case, in most countries, cohabitation continues to precede marriage rather than being a long-term living arrangement. In addition, first cohabiting unions also frequently dissolve within 5 years, as can be seen in column 3 of Table 2. This is particularly the case in the UK and the USA, where over a quarter of unmarried partnerships dissolve; these 'fragile' relationships have been frequently discussed in the U.S. literature (e.g., Carlson et al., 2004). However, around a third of cohabiting unions also dissolve in Norway, where cohabitation is usually considered a long-term relationship, and a quarter dissolve in Italy, where cohabitation is less frequently practiced. Hence, dissolution can be common in a country where cohabitation is widely practiced or in one where cohabitation is more select. To reiterate, this table demonstrates the diversity of union formation practices across Europe and the United States. Although direct marriage is clearly declining everywhere, what happens after entering cohabitation differs greatly by country.

Childbearing within Cohabitation

Along with the general increase in cohabitation, childbearing within cohabitation has been on the rise in nearly every Western country. Figure 2 shows the percent of first births in cohabitation for 14 countries of Europe and the United States. As with the overall trends in cohabitation, the largest increases have occurred in Norway, France, and Estonia. Sweden (not shown) also has had a large increase in the percent of births within cohabitation (over 50% of all births were in cohabitation in the 1990s: Kennedy and Thomson, 2010), as has eastern Germany (50% of first births occurred in cohabitation for the 1971–73 cohort: Perelli-Harris et al., 2012). Other countries in Europe have experienced moderate increases, so that in many countries one-quarter to one-third of births are in unmarried partnerships. Even countries such as Poland and Italy have experienced small increases. For example, in Italy in 1995–2003, 9% of births were to cohabiting women (Perelli-Harris et al., 2012).

The increase in childbearing within cohabitation indicates that the meaning and function of marriage is changing rapidly. Having children outside of marriage is increasingly accepted and most Western governments grant unmarried fathers the same rights and responsibilities as married fathers (Perelli-Harris and Sanchez Gassen, 2012). Nonetheless, cohabitation is not necessarily chosen as a long-term relationship type for raising children. The majority of women across Europe who

Table 2 Proportions of first unions starting with cohabitation and first cohabiting unions that had converted to marriage or dissolved (derived from cumulative incidence functions) within 5 years by birth cohort of woman

Country	*First unions beginning with cohabitation*	*First cohabiting unions married within 5 years*	*First cohabiting unions dissolved within 5 years*
Austria			
1960	79	57	16
1970	86	37	25
Belgium			
1960	68	51	6
1970	81	31	10
Bulgaria			
1960	62	90	1
1970	73	74	5
Estonia			
1960	61	69	10
1970	85	40	18
France			
1960	78	47	15
1970	90	39	20
Italy			
1960	14	53	19
1970	25	52	26
Lithuania			
1960	21	75	10
1970	40	68	11
The Netherlands			
1960	69	49	16
1970	83	30	29
Norway			
1960	84	45	21
1970	90	27	32
Poland			
1960	13	57	7
1970	20	50	13
Romania			
1960	22	76	10
1970	32	73	7
Russia			
1960	33	72	12
1970	51	63	19
Spain			
1960	26	43	4
1970	47	32	10
The UK			
1960	61	51	27
1970	82	40	29
The USA			
1960	51	59	26
1970	67	50	31

Note: Weights applied where available.
Sources: Generations and Gender Surveys in Austria (2008–09), Belgium (2008–10), Bulgaria (2004), Estonia (2004–05), France (2005), Italy (2003), Lithuania (2006), Norway (2007–08), Romania (2005), and Russia (2004); Fertility and Family Survey in the Netherlands (2003); British Household Panel Survey (2005–06); Employment, Family, and Education Survey in Poland (2006); Spanish Fertility Survey (2006); U.S. National Survey of Family Growth (2006–08).

give birth within a union do marry at some point around the birth of a first child (Perelli-Harris et al., 2012). This finding suggests that marriage is not disappearing, but is being postponed to a later stage in family formation.

Who Cohabits?

Union formation not only varies across countries, but also within countries. Patterns of union formation are becoming more destandardized, with increasing heterogeneity in whether individuals cohabit, marry, dissolve their partnerships, or repartner. However, some trends are more likely to occur among certain subgroups of the population. For example, cohabitation is more likely to be practiced by those who are less religious and those who experienced parental divorce during childhood. In the United States, evidence shows that cohabitation is more likely to be practiced by African-Americans, as well as those who are less educated and have lower incomes (Kennedy and Bumpass, 2008). The association between cohabitation and education for other countries is mixed; some countries in Eastern Europe have a negative educational gradient of ever having cohabited, while some countries in Western and Central Europe have a positive educational gradient. The educational gradient for cohabiting at the time of birth, however, is more consistent: in eight countries of Europe women with lower education have higher rates of having a birth within cohabitation than women with higher education (Perelli-Harris et al., 2010). This finding suggests that although women may not differ by education upon entrance into cohabitation, more highly educated women are more likely to marry before the birth of a child.

Besides cohabitation becoming more common among certain groups in society, studies also show differences between cohabiting and married individuals. For example, cohabitors are more likely to keep their finances separate (Lyngstad et al., 2011) and have lower-subjective well-being (Soons and Kalmijn, 2009), lower relationship quality and less commitment to each other (Wiik et al., 2009). Cohabitors also have lower fertility, both first- and second-birth rates (Perelli-Harris, 2014). It is important to keep in mind, however, that differences between cohabiting and married couples could be due to the characteristics of those who cohabit. In addition, as discussed above, many cohabitors eventually marry, and their behaviors may become more similar to married couples over time. For example, women who have a first birth in cohabitation and then marry have similar second-birth rates to those who were married at first birth (Perelli-Harris, 2014).

Living Apart Together

Besides an increase in relationships in which unmarried couples live together, relationships in which unmarried partners live in separate households, or 'Living Apart Together' (LAT) relationships, are gaining visibility in Western countries. In the past, these relationships may have been considered 'going steady' or 'dating' relationships, but today LATs are usually considered to be more stable and may be facilitating the postponement of cohabitation or marriage. The reasons for couples to live apart are numerous, for example some young couples live apart because they attend school or have jobs in

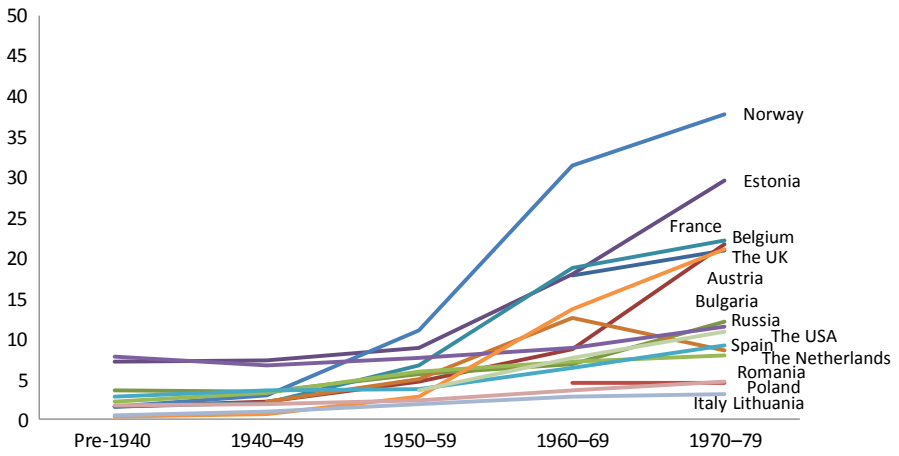

Figure 2 Proportion of women with a first birth in cohabitation by birth cohort. Note: Weights applied where available. Sources: Generations and Gender Surveys in Austria (2008–09), Belgium (2008–10), Bulgaria (2004), Estonia (2004–05), France (2005), Italy (2003), Lithuania (2006), Norway (2007–08), Romania (2005), and Russia (2004); Fertility and Family Survey in the Netherlands (2003); British Household Panel Survey (2005–06); Employment, Family, and Education Survey in Poland (2006); Spanish Fertility Survey (2006); U.S. National Survey of Family Growth (2006–08).

different places or they may be unable to afford housing separate from their parents. Older couples may live apart after having been in previous unions to maintain autonomy and separate finances, or due to relationships with adult children. LATs can be difficult to study due to the ambiguity of the definition and meaning of these relationships. However, some researchers have analyzed survey questions to determine the prevalence of LATs. For example, in the UK in 2002–03 about 3 in every 10 of adults aged 16–59 who were neither cohabiting nor married reported having a regular partner living in another household (Haskey, 2005). This figure is very similar for those aged 23–70 in the United States in 1996 and 1998 (Strohm et al., 2009) and in Spain for those aged 15–49 in 1999 (Castro-Martín et al., 2008).

Partnership Dissolution

Divorce Trends

Over the past few decades, divorce has become much more common in Western countries (*see* Divorce, Sociology of). In most Western countries divorce has increased since the mid-1960s, but in some European countries (e.g., Malta and Ireland) divorce has only recently become legal. In the EU-27, crude divorce rates nearly doubled from 1.0 divorce per 1000 per inhabitants in 1970 to 1.9 divorces by 2009 (Eurostat, 2013). In the 1970s the highest divorce rates among the EU-27 were in the Baltics, especially Latvia (4.6) and Estonia (3.2), but Russia and the US also had high-divorce rates (3.0 and 3.5. divorces per 1000 respectively) (UN, 2009). The divorce rates in these countries increased even further throughout the 1980s, but declined in the 1990s and 2000s. By the late 2000s, however, the divorce rate had increased again slightly. While many other countries experienced steep increases, few had rates approaching those found in these high-divorce countries. Since the 1980s, Northern Europe has had crude divorce rates roughly between 2.0 and 3.0, while Western and Central Europe started out having more moderate divorce rates between 1.0 and 2.0, but with steep increases to over 2.0 in the 1990s and 2000s. Southern Europe and Ireland continued to have divorce rates lower than 1.0 until 2000, when Spain's divorce rate increased to 2.1. The most recent data from 2009 to 2011 show relatively high crude divorce rates throughout Europe, possibly due to the financial crisis. Note, however, that crude divorce rates are calculated per 1000 inhabitants and not per marriage and therefore may be influenced by changes in marriage rates and the age structure of the population.

Many studies have examined the correlates of divorce, but these tend to differ across countries and change over time. For example, Härkönen and Dronkers (2006) found that women with higher education had a higher risk of divorce in France, Greece, Italy, Poland, and Spain, but that there was no relationship in Estonia, Finland, West Germany, Hungary, Latvia, Sweden, and Switzerland and a negative relationship in Austria, Lithuania, and the United States. In the US, divorce rates are diverging between higher and lower educated women, with college-educated women experiencing a decline in divorce rates while women with less than a college degree continuing to have high-divorce rates (Martin, 2006).

Other correlates of divorce include early marriage, poverty, having no religious affiliation, living in an urban area, parents' divorce, previous divorce, and having no common children (see Amato, 2010; Amato and James, 2010; Lyngstad and Jalovaara, 2010 for reviews). Wagner and Weiß (2006) found that in many European countries, premarital cohabitation was associated with divorce; however, this relationship may change over time and space. Liefbroer and Dourleijn (2006) found that couples who premaritally cohabited had higher risks of union dissolution only in countries where premarital cohabitation is relatively rare or very common. In addition, research has found that wives' employment and earned income are positively associated with divorce in the US and Europe, however this relationship may be moderated by gender attitudes,

marital quality, and division of household labor (Amato and James, 2010).

Many studies have shown that divorce has negative consequences, both on individuals and their children (see Amato and James, 2010). Research from Europe and the United States shows that divorced individuals have worse mental and physical health than their married counterparts and often die earlier. Children with divorced parents, on average, score worse on a number of emotional, behavioral, and academic outcomes and have weaker ties to their parents. As adults, they have poorer physical health and experience more problems with their own relationships. However, it is difficult to tell whether divorce causes these negative consequences or whether these outcomes are due to selection effects, such as other characteristics or factors that may cause both the divorce and the poor outcomes. Thus, the outcomes may have been similar if the couple stayed together rather than divorcing.

Union Dissolution

Not only has there been an increase in marital unions ending in divorce, relationship breakdown in general has increased, due to the frequent dissolution of cohabiting unions, as discussed above. As seen in **Figure 3**, the percent of women who have ever experienced separation from either a marital or nonmarital relationship has been increasing throughout Europe. In Norway, Estonia, and the UK, over 40% of women born in 1960–69 have experienced union dissolution. In Spain, Romania, and Italy, the proportion is substantially lower, but still around 10%. Thus, declines or the stabilization in divorce rates may be due to increases in the number of cohabiting couples that separate before marrying.

Remarriage and Repartnering

Along with the increase in union dissolution, entering into second and higher order partnerships has become more common. These partnerships are more likely to be cohabiting unions, rather than direct marriages. The partners frequently have children from previous relationships resulting in stepfamilies and blended families. Often the meaning of stepfamily relationships is less certain than for first marriages, especially when children are involved (Sweeney, 2010). Hence stepfamilies are considered to be 'incompletely institutionalized,' due to the ambiguity in the role of stepparents toward children, rights and obligations of stepfamily members, and even the names that stepchildren should call their stepparents. A growing number of studies have provided information on the characteristics, determinants, and outcomes associated with repartnering (see Sweeney, 2010 for a review). Often this research is complex, because it can investigate characteristics of the individual, the couple – in which either or both are remarried – or the incidence of multiple or serial partnerships. More research has been conducted on remarriage and stepfamilies in the US, but research in Europe is increasing and sometimes has different outcomes. For example, Sweeney (2002) found a positive association between educational attainment and remarriage for U.S. women, but studies in Europe and the US found either no relationship or a negative association with repartnering in general (see Sweeney, 2010). In most countries, second and higher order marriages are more likely to dissolve than first marriages. As with divorce and cohabitation, it is important to examine the selection processes that lead to remarriage and repartnering to better understand how family structure influences outcomes.

Explanations for Changes in Partnership Behavior

Over the past few decades, several explanations have emerged for the changes in union formation. The most common explanation is the Second Demographic Transition (Van de Kaa, 1987; Lesthaeghe, 1983). Broadly speaking, the Second Demographic Transition describes changes in family

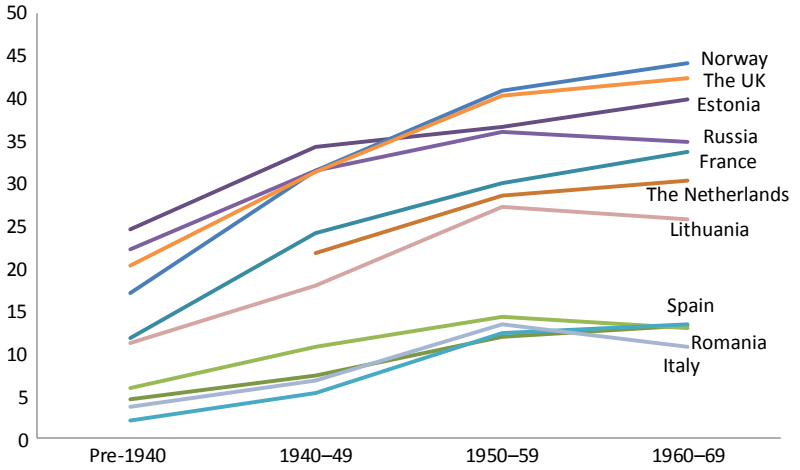

Figure 3 Proportion of women ever separated or divorced from marital or nonmarital unions, by birth cohort for selected countries of Europe. Note: Weights applied where available. Sources: Generations and Gender Surveys in Austria (2008–09), Bulgaria (2004), Estonia (2004–05), France (2005), Italy (2003), Lithuania (2006), Norway (2007–08), Romania (2005), and Russia (2004); Fertility and Family Survey in the Netherlands (2003); British Household Panel Survey (2005–06); Employment, Family, and Education Survey in Poland (2006); Spanish Fertility Survey (2006).

behavior such as the postponement of marriage, the increase in cohabitation, and the gradual decoupling of marriage and fertility. Proponents of the Second Demographic Transition explanation argue that changes in behavior are the result of ideational change, including shifts in values toward greater autonomy, individualism, and rejection of authority and institutions such as marriage (Lesthaeghe, 2010). This explanation also suggests that family change is a result of the increase in female labor force participation and women placing a greater emphasis on their career, an idea also posed by Becker (1981) as an economic explanation for the 'erosion of marriage.' Implicit in these explanations is that the more highly educated are the forerunners of these new behaviors.

Other explanations have proposed that changes in family formation are due to different social and economic factors, such as increases in economic uncertainty, risks in the labor market, or inequality. Oppenheimer et al. (1997) critiqued the idea that women's independence was responsible for declines in marriage and proposed that young men's uncertain economic position was responsible for the shift from marriage to cohabitation. A range of evidence from the US and Europe supports this theory by showing that men's unemployment lowers the probability of marriage and vice versa (see Kalmijn, 2011). In addition, the negative educational gradient of childbearing within cohabitation shows that highly educated women are more likely to marry before having a birth than lower educated women (Perelli-Harris et al., 2010). These findings suggest that individuals with less education and/or facing economic difficulties may be choosing temporary, flexible arrangements rather than committing to marriage. In general, these changes may be a sign that family formation behavior is diverging between social strata, with the highly educated more likely to postpone marriage and childbearing and the lower educated more likely to have early births, bear children outside of marriage, and divorce (McLanahan, 2004). Evidence from the United States supports McLanahan's concept of 'diverging destinies,' indicating, for example, that cohabitation is associated with lower education, violence, and abuse, and that the children born into these unions tend to have lower well-being and worse psychological and academic outcomes. The evidence for these diverging trends in other countries, however, is far less extensive.

As described above, partnership formation and dissolution differs dramatically across Europe and the United States. Some studies have examined the underlying reasons for these differences; for example, Kalmijn (2007) found that countries with a higher proportion of women in the labor force had lower marriage rates and higher divorce and cohabitation rates, while countries with a higher percent belonging to a church had higher marriage rates and lower divorce rates. However, these simple relationships mask complex differences across countries, differences that reflect entrenched historical and cultural patterns. For example, Reher (1998) argued that in northern European countries 'weak family ties' lead to early home-leaving and high levels of cohabitation, while in southern European countries 'strong family ties' encourage marriage and oppose divorce. Western and Eastern Europe also have historical differences in household formation patterns, with Eastern Europe characterized by early, universal marriage, and Western Europe typified by later age at marriage and a substantial proportion never marrying (Coale, 1992). These historical norms and values continue to influence partnership formation today. In addition, welfare states and legal systems can shape union formation; Western European countries fall along a continuum from countries with laws that treat cohabiting and married couples the same to those that privilege marriage and the breadwinner model (see Perelli-Harris and Sanchez Gassen, 2012). While it is difficult to know to what extent lawmakers are responding to changes in demographic behavior, the laws themselves may encourage citizens to choose one type of union over another. All in all, understanding differences across countries requires a multifaceted approach that incorporates social, economic, legal, and historical explanations.

Conclusion

Changes in partnership formation and dissolution have been occurring rapidly in nearly every Western country. It is difficult to know, however, what these changes imply. Are cohabitors simply postponing marriage to later in the lifecourse or is cohabitation replacing marriage altogether? Will marriage rates level off, or will marriage disappear? Will unions become increasingly more unstable, or will there be a return to greater stability? Are all countries on a similar trajectory toward relationship transformation or will some maintain their traditional, stable patterns? And finally, will patterns of family formation become markers and reproducers of inequality, with the more advantaged choosing marriage and the less advantaged cycling through less stable relationships? Future research needs to shed light on the answers to these and many related questions.

See also: Cohabitation: United States; Divorce, Sociology of; Family Theory: Competing Perspectives in Social Demography; Family Theory: Economics of Marriage and Divorce; Family as Institution; Fertility of Single and Cohabiting Women; Repartnering and Stepchildren; Second Demographic Transition; Transition to Adulthood.

Bibliography

Amato, P.R., 2010. Research on divorce: continuing trends and new developments. Journal of Marriage and Family 72 (3), 650–666.

Amato, P.R., James, S., 2010. Divorce in Europe and the United States: commonalities and differences across nations. Family Science 1 (1), 2–13.

Becker, G., 1981. A Treatise on the Family. Harvard University Press, Cambridge, MA.

Carlson, M., McLanahan, S., England, P., 2004. Union formation in fragile families. Demography 41 (2), 237–261.

Castro-Martín, T., Domínguez-Folgueras, M., Martín-García, T., 2008. Not truly partnerless: non-residential partnerships and retreat from marriage in Spain. Demographic Research 18 (16), 443–468.

Centers for Disease Control, 2013. http://www.cdc.gov/nchs/fastats/divorce.htm.

Coale, A.J., 1992. Age of entry into marriage and the date of the initiation of voluntary birth control. Demography 29 (3), 333–341.

Eurostat, 2013. Marriage and Divorce Statistics. http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Marriage_and_divorce_statistics.

Härkönen, J., Dronkers, J., 2006. Stability and change in the educational gradient of divorce. A comparison of seventeen countries. European Sociological Review 22 (5), 501–517.

Haskey, J., 2005. Living arrangements in contemporary Britain: having a partner who usually lives elsewhere and living apart together (LAT). Population Trends 122, 35.

Heuveline, P., Timberlake, J.M., 2004. The role of cohabitation in family formation: the United States in comparative perspective. Journal of Marriage and the Family 66 (5), 1214–1230.

Kalmijn, M., 2007. Explaining cross-national differences in marriage, cohabitation, and divorce in Europe, 1990–2000. Population Studies 61 (3), 243–263.

Kalmijn, M., 2011. The influence of men's income and employment on marriage and cohabitation: testing Oppenheimer's theory in Europe. European Journal of Population 27 (3), 269–293.

Kennedy, S., Bumpass, L., 2008. Cohabitation and children's living arrangements: new estimates from the United States. Demographic Research 19, 1663–1692.

Kennedy, S., Thomson, E., 2010. Children's experiences of family disruption in Sweden: differentials by parent education over three decades. Demographic Research 23 (17), 479.

Lesthaeghe, R., 1983. A century of demographic and cultural change in Western Europe: an exploration of underlying dimensions. Population and Development Review 3, 411–435.

Lesthaeghe, R., 2010. The unfolding story of the second demographic transition. Population and Development Review 36 (2), 211–251.

Liefbroer, A.C., Dourleijn, E., 2006. Unmarried cohabitation and union stability: testing the role of diffusion using data from 16 European countries. Demography 43 (2), 203–221.

Lyngstad, T.H., Jalovaara, M., 2010. A review of the antecedents of union dissolution. Demographic Research 23 (10), 257–292.

Lyngstad, T.H., Noack, T., Tufte, P.A., 2011. Pooling of economic resources: a comparison of Norwegian married and cohabiting couples. European Sociological Review 27 (5), 624–635.

Martin, S.P., 2006. Trends in marital dissolution by women's education in the United States. Demographic Research 15 (20), 537–560.

McLanahan, S., 2004. Diverging destinies: how children are faring under the second demographic transition. Demography 41 (4), 607–627.

Oppenheimer, V.K., Kalmijn, M., Lim, N., 1997. Men's career development and marriage timing during a period of rising inequality. Demography 34 (3), 311–330.

Perelli-Harris, B., Sigle-Rushton, W., Lappegård, T., Keizer, R., Berghammer, C., Kreyenfeld, M., 2010. The educational gradient of nonmarital childbearing in Europe. Population and Development Review 36 (4), 775–801.

Perelli-Harris, B., Sanchez Gassen, N., 2012. How similar are cohabitation and marriage? Legal approaches to cohabitation across Western Europe. Population and Development Review 38 (3), 435–467.

Perelli-Harris, B., 2014. How similar are cohabiting and married parents? Second conception risks by union type in the United States and across Europe. European Journal of Population. http://link.springer.com/article/10.1007/s10680-014-9320-2.

Perelli-Harris, B., Kreyenfeld, M., Sigle-Rushton, W., Keizer, R., Lappegård, T., Jasilioniene, A., Giulio, P.D., 2012. Changes in union status during the transition to parenthood: an examination of 11 European countries. Population Studies 66 (2), 167–182.

Raley, R.K., 2001. Increasing fertility in cohabiting unions: evidence for the second demographic transition in the United States. Demography 38 (1), 59–66.

Reher, D.S., 1998. Family ties in Western Europe: persistent contrasts. Population and Development Review 24 (2), 203–234.

Rindfuss, R.R., VandenHeuvel, A., 1990. Cohabitation: a precursor to marriage or an alternative to being single? Population and Development Review 16 (4), 703–726.

Smock, P.J., 2000. Cohabitation in the United States: an appraisal of research themes, findings, and implications. Annual Review of Sociology 26, 1–20.

Soons, J.P.M., Kalmijn, M., 2009. Is marriage more than cohabitation? Well-being differences in 30 European countries. Journal of Marriage and Family 71 (5), 1141–1157.

Strohm, C.Q., Seltzer, J.A., Cochran, S.D., Mays, V.M., 2009. "Living apart together" relationships in the United States. Demographic Research 21, 177.

Sweeney, M., April 2002. Remarriage and the nature of divorce: does it matter which spouse chose to leave? Journal of Family Issues 23, 410–440.

Sweeney, M., 2010. Remarriage and stepfamilies: strategic sites for family scholarship in the 21st century. Journal of Marriage and Family 72 (3), 667–684.

United Nations. 2009. www.un.org/esa/population/…/UNPD_WMD_2008_DIVORCES.xls.

Van de Kaa, D.J., 1987. Europe's second demographic transition. Population Bulletin 42 (1), 1–57.

Wagner, M., Weiß, B., 2006. On the variation of divorce risks in Europe: findings from a meta-analysis of European longitudinal studies. European Sociological Review 22 (5), 483–500.

Wiik, K.A., Bernhardt, E., Noack, T., 2009. A study of commitment and relationship quality in Sweden and Norway. Journal of Marriage and Family 71 (3), 465–477.

Relevant Websites

http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Marriage_and_divorce_statistics – Eurostat.

www.nonmarital.org – Information on a Project on Childbearing in Cohabitation.

http://www.cdc.gov/nchs/fastats/divorce.htm – U.S. Centers for Disease Control.

Parts of Speech

Walter Bisang, Johannes Gutenberg-Universität Mainz, Mainz, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M. Haspelmath, volume 24, pp. 16538–16545, © 2001, Elsevier Ltd.

List of Abbreviations

1, 2, 3 First, Second, Third Person Marker
A Agent
ABS Absolutive
ART Article
AT Agent Trigger (Tagalog)
CAUS Causative
HUM Human
IND Indicative
LK Linker (Nahuatl)
LOC Locative
LT Locative Trigger (Tagalog)
MOM Momentaneous Aspect
N Noun
OBJ Object Role
P Patient
P/A Patient/Agent Marker
PFV Perfective Aspect
PL Plural
POSS Possession Marker
PST Past
PT Patient Trigger (Tagalog)
SG Singular
T Trigger Marker (Tagalog)
V Verb

Abstract

After a short traditional discussion of parts of speech, the present article will start out from a definition of four prerequisites that are relevant for the definition of parts of speech. These prerequisites will be used for comparing the two most comprehensive typological approaches to parts of speech by Croft (1991, 2000, 2001) and Hengeveld (1992, 2013). These approaches are then checked against three languages whose parts-of-speech systems have been described as typologically remarkable in the literature (classical Nahuatl, Tagalog, Tongan). A short conclusion summarizes these findings.

Introduction: Traditional Word Classifications

The classification of words has a long tradition. Dionysios Thrax (second century BC, Lallot, 1998), who was the first writer of a grammar in Europe and who summarized several centuries of Greek reflections on language before his time, already had a term for 'parts of speech' (The Greek term used by Dionysios Thrax was *mér-ē lóg-ou* [part-NOM:PL:NEUTR speech-GEN:SG:MASC] 'parts of speech.'). In his *téxnē grammatikē* 'Art of Grammar,' he defined the following eight parts of speech: nouns, verbs, participles, articles, pronouns, prepositions, adverbs, and conjunctions. With the exception of participles, these parts of speech are still relevant in most traditional grammars of English and many other languages. The following 10 parts of speech are the ones that are most commonly used in contemporary grammatical descriptions (Haspelmath, 2001). Each part of speech is illustrated by a few words from English.

1. Noun: *tiger, book, wind, departure, movement, carelessness*
   Verb: *know, come, push, give*
   Adjective: *good, red, happy, Russian*
   Adverb: *well, quickly, desperately, fortunately*
   Pronoun: *I, you, he, she*
   Adposition (preposition, postposition): *in, on, for, because of*
   Conjunction: *and, or, if, because*
   Numeral: *one, two, twenty, five hundred*
   Article: *the, a*
   Interjection: *ouch, tut-tut, shh, yeah*

The parts of speech listed in (1) do not all have the same size. There are open word classes that accept new members and comprise large numbers of words, and there are closed word classes consisting only of a small and fixed number of words. Verbs, nouns, and in some languages also adjectives and adverbs are open-class categories; the remaining parts of speech are closed-class categories. The individual members of the open-class categories occur comparatively rarely in discourse, while the members of the closed-class categories are rather frequent, first of all if they express a grammatical function (pronouns, adpositions, conjunctions, articles). The distinction between content words and function words covers a similar range of parts of speech. Verbs, nouns, adjectives, and adverbs denote individual concepts and are thus called 'content words,' while the 'functional words' stand for grammatical functions. Due to processes of grammaticalization, defined as the development of grammatical markers out of full words, the distinction between content and functional words is sometimes unclear. The class of auxiliaries is not part of the Greek and Roman tradition and this may well be

the reason why it is usually not listed as a separate word class, even though the auxiliaries *be* and *have* (*She **is** swimming. He **has** come.*) and their equivalents in other languages are very important function words.

The traditional criteria for distinguishing parts of speech are a combination of semantic and formal (morphological and syntactic) criteria. Attempts to define word classes purely in terms of semantics are too simplistic and can easily be falsified. Thus, a definition like 'nouns denote things, persons, and places' is disproved by a word like *movement*, which is a noun but refers to an action. The definitions of 'verbs denote actions and processes' and 'adjectives denote properties' run into similar problems (cf *know*, *exist* for the former and *urban*, *Bavarian* for the latter). The traditional formal criteria are based on distribution, i.e., the morphosyntactic environments in which an individual word with its semantics can occur. The formal criteria for defining verbs are morphological criteria like the compatibility with markers for tense, aspect and mood, markers of polarity (negative vs affirmative), of agreement (with subject and sometimes other categories like direct and indirect objects for person, number, and gender) and of valency change (passive, antipassive, causative, anticausative, middle, reciprocals). Syntactically, verbs can form the head of verbal phrases and they are characterized by their elaborate argument system. While other parts of speech can have maximally one argument, verbs can have one (intransitives: *sleep*), two (transitives: *hit*), or three arguments (*give*). Nouns are often characterized by morphological markers of case (nominative, accusative, ergative, absolutive, dative, etc.), number (singular, plural, dual, trial, paucal), gender (masculine, feminine, neuter), definiteness/indefiniteness, and possessor (cf Turkish *ev-im* [house-POSS:1.SG] 'my house'). Syntactically, nouns can take the head position of noun phrases and the argument positions of verbs. In addition, they are often combined with words like demonstratives or articles and they are the antecedents of pronouns and reflexives. Adjectives are characterized in some languages by special morphological forms or by specific syntactic constructions. English uses both expression formats, depending on the individual adjective. With adjectives like *happy*, the comparative and the superlative are formed morphologically (e.g., *happier*, *happiest*), while adjectives like *intelligent* express the same grammatical functions syntactically (e.g., *more intelligent*, *most intelligent*). In addition, adjectives cannot only occur in comparative constructions but also as modifiers of nominal heads. Finally, adverbs are the most heterogeneous category. In a strict definition, they modify verbs as *quickly* in *he runs quickly*. Morphologically, they are often unmarked (German: *Er singt schön* [he sing beautiful:ADV] 'He sings beautifully') but they can also have a specific marker (cf English *-ly*). A closer look at English *-ly* reveals that there are many other words marked by *-ly* whose scope is beyond the verb in its function as a predicate. Ramat and Ricca (1998) discuss the following types in their comparative study of adverbs in European languages: predicate adverbs that follow the strict definition, event adverbs (e.g., *obligatorily*, *rarely*, *yestereday*, *often*), propositional adverbs (e.g., *probably*, *unfortunately*, *cleverly*), and speech-act adverbs (e.g., *briefly*, *frankly*).

Parts-of-speech definitions of the above type are problematic for at least two reasons. (1) The generality problem is due to the language specificity of formal criteria based on expression formats for individual grammatical categories. Many languages have no articles or verb agreement. Due to differences like these, the formal criteria can be used at best for individual languages and are of little use for cross-linguistically valid universal definitions. (2) The subclass problem arises from the difficulty of deciding which of the language-particular criteria should be used for word-class definition. As was discussed above, adjectives can be defined by the presence of comparative morphology or by a syntactic comparative construction. In the case of English, one would have to distinguish between two subclasses of adjectives because the language uses both expression formats. Even though it is generally agreed that this is not sufficient for postulating two subclasses, the theoretical foundations for such a conclusion are anything but clear.

Criteria for Defining Parts of Speech

The current discussion on universally applicable criteria for the cross-linguistic definition of parts of speech is based on the following four prerequisites (Sasse, 1993b, pp. 196–201; Bisang, 2011, p. 281):

- semantic criteria;
- pragmatic criteria/criteria of discourse function;
- formal criteria (morphology, syntax);
- distinction between lexical and syntactic levels of analysis.

As will be seen in the Approaches to Parts of Speech and Short Case Studies and Problems sections, all four prerequisites are relevant for a universal assessment of parts of speech. The present section will only discuss each prerequisite individually from the perspective of the open-class categories of nouns and verbs.

Langacker's (1987) notional description of nouns and verbs is probably the most elaborate semantic approach to parts of speech. His definition of nouns is based on the concept of the REGION, which is conceived of as 'a set of interconnected entities' (Langacker, 1987, p. 62). Since nouns designate REGIONS, they are stable through time (static) and holistic. Verbs are defined as processes. They evolve through time state by state in "successive transformations which derive each component state from its predecessor" (Langacker, 1987, p. 74). In spite of its thoroughness, this approach cannot avoid the shortcomings of purely semantic definitions and it does not offer discovery procedures for parts-of-speech identification (Sasse, 1993a, p. 649). Another problem that concerns almost all definitions of semantic criteria with the exception of Dixon (2004) was pointed out by Wierzbicka (2000). Semantic criteria like object versus action or region versus process are too general because one cannot just take any object-denoting or action-denoting lexeme that belongs to these concepts (e.g., *parrot* or *grovel*) for testing word-class membership across languages. Many languages simply do not have a corresponding word. For that reason, the cross-linguistic assessment of parts-of-speech distinctions must be based on 'genuinely universal linguistic prototypes' (Wierzbicka, 2000, p. 288).

Hopper and Thompson (1984, p. 708) see the semantic properties of nouns and verbs as "derivative of (and perhaps even secondary to) their discourse roles." From such

a perspective, the prototypical discourse function of nouns consists in the introduction of participants and their deployment in the ongoing text. For that purpose, nouns must show continuity of identity and this, in turn, depends on time stability. Verbs prototypically "assert the occurrence of an event" (Hopper and Thompson, 1984, p. 708) and thus reply to the question 'What happened?' They can occur in many contexts in which their categoriality is reduced. Hopper and Thompson (1984) discuss stativity, irrealis, negation, serial verbs, compound verbs, and dependent clauses. From a broader perspective, nouns and verbs are associated with information structure and its categories of topic (nouns) and comment (verbs).

The importance of the morphological and the syntactic form in which the relevant semantic and pragmatic criteria are expressed is undisputed. It plays a crucial role in the traditional approach (Introduction: Traditional Word Classifications section) as well as in the more recent typological approaches discussed in the Approaches to Parts of Speech section. Even Dionysios Thrax clearly mentions morphological criteria (the expression of case, number, and gender), a fact that is missed by his critics who state that his definitions are purely semantics-based (Bisang, 2011, p. 280).

Most linguistic theories take it for granted that there is a perfect match between the lexicon and syntax. The lexicon consists of lexemes that are specified for certain word classes (N, V, etc.) and the syntactic structures are derived from these properties (N projects into NP, V projects into VP; cf Sasse, 1993a,b; Broschart, 1997). This generalization seems to hold for the majority of languages but there are two ways in which languages may violate this perfect match. Either they have only one open word class which basically encompasses all content words (cf the Extremely Rigid section on classical Nahuatl and Tagalog) or their lexical items lack parts-of-speech specification in the lexicon (cf the Contentives: The Case of Tongan section on Tongan). The latter type is sometimes discussed under the term of 'precategoriality' (Bisang, 2008, 2011).

Approaches to Parts of Speech

Croft (1991, 2000, 2001)

Croft (1991, 2000, 2001) offers a solution that significantly minimizes various shortcomings of earlier approaches. These shortcomings are due to general problems associated with the distributional analysis and the two perspectives of 'lumping' and 'splitting.' The distributional method has to face at least two problems. One of them is that a construction for determining word class in one language is not available in another language (cf the Introduction: Traditional Word Classifications section). The other problem is that parts-of-speech behavior in one construction is not sufficient. Individual words may well obscure parts-of-speech distinctions in the compatibility with one construction, but that does not imply the pervasiveness of that property across other constructions. Croft (2000, pp. 73–75) illustrates this fact with data from Makah as it is described by Jacobsen (1979). In the following example from this Nootka language, the inflectional marking of predicates is characterized by obligatory person marking (1.SG/3.SG) and by modality (IND = indicative). Even though inflectional verb morphology is quite complex in Makah and in Nootka in general, the glosses to the clause-initial predicate in the second line of examples (2a–d) indicate that this morphology is compatible with words denoting actions like 'to point' (2a), objects like 'white man' (2b), properties like 'to be big' (2c), and with adverbs like 'still' (2d):

2. Makah (Jacobsen, 1979, p. 110–111; from Croft, 2000, p. 73 and 2001, p. 30):
   a. k'upšil baʔas ʔuyu·q.
      point:MOM:IND:3 house OBJ
      'He's pointing at the house.'
   b. babałdis.
      white.man:IND:1.SG
      'I'm a white man.'
   c. ʔi·ʔi· X̣^{W} ʔi.
      big:IND:3
      'He's big.'
   d. hu ʔaxis haʔuk^{w} ap.
      still:IND:1.SG eat:CAUS
      'I'm still feeding him.'

As Croft (2000, pp. 73–76; 2001) points out, the data in (2) do not support the conclusion that Makah lacks any major parts-of-speech distinction. As he illustrates, the same words show clear distinctions in other constructions. For that reason, he concludes that it is possible to identify nouns, verbs, adjectives, and even adverbs and auxiliaries in Makah (for examples with significantly different marking patterns, cf the case studies in the Short Case Studies and Problems section).

A complete distributional analysis that integrates all possible compatibilities "provides the complete picture of the grammatical patterning of a language" (Croft, 2000, p. 76). This picture is extremely complex and fine-grained. Depending on the perspective of the individual approach to the definition of parts of speech, one can either ignore important distributional patterns and thus lump various subclasses together into a more general class (e.g., by defining Makah word classes only on the basis of the inflectional marking in (2)) or one can split the more general classes into subclasses by integrating data on more distributional environments. Ultimately, both perspectives are unsatisfactory. To avoid the problems of lumping and splitting, Croft offers his own construction-based universal theory.

Croft's theory understands the word classes of nouns, verbs, and adjectives as universal categories rather than as language-particular categories and thus takes a view that is diametrically opposed to most other approaches. In his view, nouns, verbs, and adjectives are typological prototypes that can be situated within a conceptual space that consists of three semantic classes and three pragmatic functions. The three semantic classes are objects, properties, and actions. The semantic classes form the vertical dimension of the conceptual space and are combined with the three pragmatic functions of reference, modification, and predication that take the horizontal dimension (cf **Table 1**). The pragmatic categories are defined as follows (also cf Bisang, 2011, p. 289). Predication is "what the speaker intends to *say about* what he is talking about (the referent)" (Croft, 1991, p. 52). The function of reference "is to get the hearer to identify an

Table 1 Croft's conceptual space for parts of speech

	Reference	*Modification*	*Predication*
Objects	*Object reference:* UNMARKED NOUNS	*Object modifier:* genitive, adjectivalization, prepositional constructions (cf Engl. *of N*)	*Object predication:* predicate nominals, copulas
Properties	*Property reference:* de-adjectival nouns	*Property modifier:* UNMARKED ADJECTIVES	*Property predication:* predicate adjectives, copulas
Actions	*Action reference:* nominals, complements, infinitives, gerunds	*Action modifier:* participles, relative clauses	*Action predication:* UNMARKED VERBS

Adapted from Croft, W.A., 1991. Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information. Chicago University Press, Chicago, 53/67; Croft, W.A., 2000. Parts of speech as typological universals and as language particular categories. In: Vogel, P.M., Comrie, B. (Eds.), Approaches to the Typology of Word Classes. Mouton de Gruyter, Berlin, 89; and Croft, W.A., 2001. Radical Construction Grammar. Syntactic Theory in Typological Perspective. Oxford University Press, Oxford, 88/92.

entity as what the speaker is talking about" (Croft, 1991, p. 52). Modification is defined as an 'accessory function to reference and predication' because it is either a more specific instrument for identification in the case of restrictive modification or an instrument that provides a secondary comment in the case of unrestrictive modification (Croft, 1991, p. 52).

Each of the nine cells within the conceptual space stands for a type of construction. The terms for these construction types are given in italics. For each type, some individual constructions are listed in normal print.

Parts of speech manifest themselves within this space through markedness patterns of function-indicating morphosyntax in individual languages. In this context, markedness is seen as a universal property of conceptual categories (cf Croft, 2003, pp. 87–101, 110–117 on 'typological markedness'). Thus, a certain grammatical category is expressed with a certain degree of overt morphosyntactic markedness relative to other categories. In the above conceptual space for parts of speech, three combinations of pragmatic function and semantic class are prototypically unmarked. These same combinations also constitute parts of speech universally:

3. Croft (2000, p. 88):

noun = reference to an object
adjective = modification by a property
verb = predication of an action

This can be illustrated by a few examples from English. An object-denoting lexeme like *boy* tends to be less marked if it is in the pragmatic function of reference (*boy*) than if it is in the function of modification (*of a boy, the boy's, boyish*) or in the function of predication (*is a boy*). Similar statements are possible for property-denoting lexemes in the function of modification and action-denoting lexemes in the function of predication.

In Croft's (2000, p. 89) typological view on marking, "the structural coding criterion specifies only that the marked member is encoded by *at least as many* morphemes as the unmarked member." Thus, Croft's theory only excludes cases in which an unmarked member is expressed by more morphemes than a marked member. Equal marking and zero marking across neighboring cells are possible. In other words, the conceptual space of parts of speech predicts that unmarked combinations cannot be marked by more morphosyntactic material than the typologically marked combinations. It excludes a language in which predication of actions is expressed by a copula, while predication of properties and/or objects has no morphosyntactic marking.

From the perspective of the four prerequisites (Criteria for Defining Parts of Speech section), Croft's theory is characterized by the absence of the distinction between lexical and syntactic levels of analysis (fourth prerequisite). This is in line with his perspective of Radical Construction Grammar (Croft, 2001), which rejects the existence of universal categories (e.g., syntactic categories like N and V) and takes constructions as the basic primitive units of syntactic representation.

Functional Grammar (Hengeveld, 1992, 2013)

Hengeveld's (1992, 2013) approach to parts of speech is situated in the framework of Functional Grammar. He classifies basic and derived lexemes according to their formal behavior within a four-slot space formed by two parameters with two values each. One parameter involves the opposition between predication and reference, the other between heads and modifiers (cf **Table 2**).

The definition of parts of speech within the above table depends on markedness and on the basic question of whether a lexeme requires lexical or syntactic derivation for being accepted in the individual cells. Individual lexemes can be either rigid or flexible. A rigid parts-of-speech system is characterized by classes of lexemes that are tied to a single cell and require additional marking to be used in other functions. In Hengeveld's (1992, 2013) analysis, English is such a language. It displays separate lexeme classes of verbs, nouns, adjectives, and manner adverbs (this approach is only concerned with manner adverbs or MAdv). Thus, a verb like *sing* in (4) is used as the head of a predicate phrase. If it takes another functional slot, it is more marked. For the function of a referential head, it needs ablaut (*song*). In the two modifier functions, *sing* gets the form of *sing-ing* if it modifies a referential phrase and (*by*) *singing* if it modifies a predicate phrase.

Table 2 Lexemes and functions

	Head	*Modifier*
Predicate phrase	V	MAdv
Referential phrase	N	Adj

Hengeveld, K., 2013. Parts-of-speech systems as a basic typological determinant. In: Rijkhoff, J., Van Lier, E. (Eds.), Flexible Word Classes. Typological Studies of Underspecified Parts of Speech. Oxford University Press, Oxford, p. 32.

4. The tall$_{A}$ girl$_{N}$ sings$_{V}$ beautifully$_{MAdv}$.

Nouns are heads of referential phrases, adjectives are modifiers of referential phrases, and manner adverbs are modifiers of predicate phrases. Languages whose lexemes are rigid in all four functions are called 'differentiated' by Hengeveld (1992, 2013). A flexible parts-of-speech system is characterized by classes of lexemes that can be used in more than one cell without any difference in markedness. In the most extreme case, there is only one class of lexemes that covers all four cells. As will be seen in the Contentives: The Case of Tongan section, Tongan comes close to this type, which is called 'contentive.' Classical Chinese (Bisang, 2008) also belongs to this type.

The four functions in **Table 2** form a hierarchy, the parts-of-speech hierarchy as claimed by Hengeveld (1992) and Hengeveld et al. (2004):

5. Head of predicate phrase > Head of referential phrase > Modifier of referential phrase > Modifier of predicate phrase

The hierarchy determines the likelihood with which a language develops specialized classes of lexemes. The more to the left a function is in (5), the more likely it is that a language has a specialized class of lexemes. The hierarchy is implicational. Thus, it predicts for example, that a language with a specialized class for the expression of 'modifier of referential phrase' (adjective) also has specialized classes for 'head of referential phrase' (noun) and 'head of predicate phrase' (verb).

This hierarchy plus the distinction between flexible, differentiated, and rigid languages yields seven possible parts-of-speech systems summarized in **Table 3**. There are three different types of flexible systems (types 1–3), three different types of rigid systems (types 5–7), and the maximally rigid type called 'differentiated' (type 4).

In Croft's (2000, 2001) analysis, this system clearly sides with the lumpers that disregard various facts. This can be illustrated by the following example from Tongan that is discussed by Hengeveld (1992, p. 66) and Croft (2000, p. 70; 2001, p. 68):

6. Tongan (Tchekhoff, 1981, p. 4):
   a. na'e **si'i** 'a e akó.
      PST small ABS ART school
      'The school was small.'
   b. 'i 'ene **si'í.**
      in POSS:3:SG childhood
      'in his/her childhood'
   c. na'e ako 'a e tamasi'i **si'i** iate au.
      PST study ABS ART child small LOC 1.SG
      'The little child studied at my home.'

As the above example shows, the lexeme *si'i* 'small' can take the verbal position (6a), the nominal position (6b), and the adjectival position (6c) without any change in markedness. Thus, it counts as a highly flexible lexeme. As Croft (2000, 2001) points out, a closer look reveals that the meaning of *si'i* in these individual functions cannot be regularly derived at least in the case of 'be small' (6a) versus 'childhood' (6b). Since it has underivable meanings in the individual functions, Croft argues that one should still analyze it in terms of different lexemes from a more splitting perspective (but cf the Contentives: The Case of Tongan section).

Table 3 Parts-of-speech systems

PoS system		*Head of predicate phrase*	*Head of referential phrase*	*Modifier of referential phrase*	*Modifier of predicate phrase*
Flexible	1	Contentive			
	2	Verb	Nonverb		
	3	Verb	Noun	Modifier	
Differentiated	4	Verb	Noun	Adjective	Manner adverb
Rigid	5	Verb	Noun	Adjective	
	6	Verb	Noun		
	7	Verb			

Hengeveld, K., 2013. Parts-of-speech systems as a basic typological determinant. In: Rijkhoff, J., Van Lier, E. (Eds.), Flexible Word Classes. Typological Studies of Underspecified Parts of Speech. Oxford University Press, Oxford, p. 36.

Hengeveld's (1992, 2013) lumping tendency has the positive effect of leading to a number of predictions concerning correlations between the parts-of-speech system of a language and other properties of its grammar. Hengeveld (2013) groups them under four headings of (1) identifiability, (2) integrity, (3) unity, and (4) pervasiveness. Some of these generalizations will be briefly sketched here (cf Hengeveld, 2013 for a comprehensive summary):

Identifiability

In a rigid system, the functional slot to which a class of lexemes belongs is clearly determined. In flexible systems, the lexical information is insufficient for uniquely identifying the function of a lexeme. Thus, in a language of type 1 (contentive), the lexical information provided by a lexeme is underspecified for the categorical status of the phrase in which it occurs as a predicate phrase or as a referential phrase within a sentence. This lack of identifiability in flexible systems is compensated by less freedom of word order in flexible systems. More specifically, languages of type 1 are verb-first or verb-final languages, they do not have predicate medial basic word order. Only with these word orders is it possible to clearly associate that position with the predicate phrase (but cf Conclusion).

Integrity

The flexibility of a lexeme class would be significantly restrained if its shape would depend on specific functional environments. Thus, "flexible lexemes may be expected not to show morphologically conditioned stem alternation, such as morphophonological variation, irregular stem formation, or suppletion" (Hengeveld, 2013).

Unity

The functional elasticity of flexible lexemes crucially depends on the phonological, morphological, and semantic unity or homogeneity of a lexical class. This homogeneity is hampered if flexible classes are divided into subclasses. Thus, systems of

subclassification like declension and conjugation classes are unlikely to occur in flexible languages. An example of semantic subclassification is number. Flexible lexemes are not intrinsically specified for singular (or plural) number.

Pervasiveness

The predictions discussed under this heading deal with the pervasiveness of flexibility through different levels of analysis. The general prediction is that flexibility decreases from lower to higher levels. In languages with flexible systems, the parts-of-speech system used for the expression of subordinated clauses (e.g., the use of a single construction type for relative clauses and complement clauses) tends to be less flexible than the parts-of-speech system at the level of the matrix clause (Van Lier, 2009).

Hengeveld's (1992, 2013) approach includes all four prerequisites. The opposition between predication and reference covers the semantic and the pragmatic criteria. The formal criteria are reflected in the markedness patterns associated with the lexeme classes. The lexemes with their specific forms are mapped onto the head or the modifier positions of the predicate phrase and the referential phrase. Thus, the distinction of the two levels of lexicon and syntax is a crucial component of this approach.

Short Case Studies and Problems

The approaches of Croft (section Croft (1991, 2000, 2001)) and Hengeveld (section Functional Grammar (Hengeveld, 1992, 2013)) differ with regard to the fourth prerequisite (Criteria for Defining Parts of Speech section). From Croft's (2001) perspective of Radical Construction Grammar, there are no universal syntactic categories onto which lexical items have to be mapped, while Hengeveld's approach integrates the mapping of lexemes onto the positions of heads and modifiers. As was mentioned at the end of the Criteria for Defining Parts of Speech section, there are two types of parts-of-speech systems which can only be fully understood if the fourth prerequisite is integrated. In one type, there is only one open word class that covers all content words. This type is captured by the extremely rigid type 7 in Hengeveld's terminology. The other type is characterized by the absence of parts-of-speech specification in the lexicon and is reflected in Hengeveld's contentive type with its extremely flexible system.

In both types, the noun/verb distinction does not work on the syntactic level. In the extremely rigid type, there is only one open word class, while there is no noun/verb distinction in the lexicon of the contentive type. The first type will be illustrated by examples from classical Nahuatl (Omnipredicativity in Classical Nahuatl section) and Tagalog (Tagalog And The Parts-Of-Speech Hierarchy section), the second type by examples from Tongan (Contentives: The Case of Tongan section). Classical Chinese as another example of the second type will not be discussed here (Bisang, 2008). As will be seen, the languages that belong to these types still offer quite a few challenges to current approaches to parts of speech. Most remarkably, the morphology of most of them still shows noun/verb distinctions even if the lexicon/syntax mapping disregards them (On the Role of Morphology section).

Extremely Rigid

Omnipredicativity in Classical Nahuatl

Classical Nahuatl, an Uto-Aztecan language spoken in Mexico and adjacent areas at the time of the conquest of America by Spain, is claimed to be a language whose content words are all mapped onto predicates (Launey, 1994). Thus, its lexical inventory of content words consists exclusively of verbs.

In classical Nahuatl morphology, the person markers for intransitive arguments (S) and for agents (A) are identical to the person markers for predicative constructions with lexemes that denote physical objects (cf **Table 4**).

Thus, the prefix *ni-* '1SG' can mark an action-denoting lexeme (*ni-chōca* [1SG-cry] 'I cry') as well as an object-denoting lexeme (*ni-ticitl* [1SG-doctor] 'I am a doctor'). Since the prefix for 3SG is zero, a seemingly simple sentence like (7) that would be translated as 'The child cries' can be analyzed as consisting of two predications: 's/he cries' and 'it is a child':

7. Launey (1994, p. 29):
   ϕ-chōca in ϕ-piltōntli.
   3.SG-cry LK 3.SG-child
   'S/he cries – it is a child'

Similarly, a transitive clause is expressed by three predications:

8. Launey (1994, p. 37):
   ϕ-qui-cua in ϕ-piltōntli in ϕ-nacatl.
   3.SG.A-3.SG.P-eat LK 3.SG-child LK 3.SG-meat
   'S/he eats it – it is a child – it is meat' for 'The child eats meat.'

The word *in* (glossed as LK for 'linker'; cf Launey, 1994, pp. 122–132) is generally used as a demonstrative or as a relative-clause marker. In its function as a linker (LK), both of these functions are used simultaneously. It is the demonstrative argument of one predication and the relative marker of the other. In (7), it is a demonstrative that takes the argument position of 'cry' (*ϕ-chōca* 'That one/he/she cries') and the relative marker of the second predication 'be a child' (*in ϕ-piltōntli* 'the one who is a child'). Thus, (7) may be translated literally as 'That one cries, the one who is a child.'

This analysis consistently accounts for person marking across action-denoting and object-denoting lexemes. It also integrates the functions of *in* as a demonstrative and as a relative-clause marker that produces cleft-like

Table 4 Person agreement markers on action-denoting lexemes and on object-denoting lexemes

	Action-denoting lexemes			*Object-denoting lexemes*	
	S/A	*P*	*Reflexive*	*S*	*Possessive*
1.SG	*n(i)-*	*-nēch-*	*-n(o)-*	*n(i)-*	*-n(o)-*
2-SG	*t(i)/x(i)-*	*-mitz-*	*-m(o)-*	*t(i)-*	*-m(o)-*
3.SG	*ϕ-*	*-c-/-qu(i)-*	*-m(o)-*	*ϕ-*	*-i-*
1.PL	*t(i)-*	*-tēch-*	*-t(o)-*	*t(i)-*	*-t(o)-*
2.PL	*aM-/x(i)-*	*-amēch-*	*-m(o)-*	*aM-*	*-am(o)-*
3.PL	*ϕ-*	*-quiM-*	*-m(o)-*	*ϕ-*	*-iM*

Adapted from Launey, M., 1994. Une grammaire omniprédicative. Essai sur la morphosyntaxe du nahuatl classique. CNRS Éditions, Paris, 10–11.

constructions. Because of free word order in classical Nahuatl, clefting can be used for marking information structure.

Tagalog and the Parts-of-Speech Hierarchy

A simple declarative clause in Tagalog starts with an action-denoting lexeme. What is crucial for the discussion of parts of speech is that this lexeme is marked for the semantic role of the constituent that is marked by the particle *ang*. Other categories marked on the first lexeme are aspect and kind of action (indicative, distributive, aptative, etc.). In (9a), the infix *-um-* of the action-denoting lexeme *bili* 'buy' marks the element *lalaki* 'man' in the *ang*-phrase as agent (A). The element in the *ang*-phrase has often been described under the terms of 'focus' or 'topic,' a terminology that is misleading, since *ang* rather functions as a marker of specificity (Himmelmann, 1997, p. 103). To avoid this problem, Schachter (1993) has introduced the term 'trigger.' The example in (9a) thus illustrates the function of agent trigger (AT). In (9b), the ϕ-suffix stands for patient trigger (PT). Thus, the element in the *ang*-phrase is marked as the patient on the clause-initial predicate (*b<in>ili-ϕ*). In addition to the roles of agent and patient, the predicate can also be marked for locative (locative trigger, LT) as in (9c) and for other roles like instrumental and benefactive. The elements that are not in the trigger function are introduced by other particles, among them *ng* for agents, patients, and some instrumentals and *sa* for locatives.

9. Tagalog (Foley and Van Valin, 1984, p. 135):
   a. B<um>ili **ang lalaki** ng isda sa tindahan.
      PFV:/AT/-buy T man P/A fish LOC store
      'The man bought fish at the store.'
   b. B<in>ili-ø ng lalaki **ang isda** sa tindahan.
      /PFV/-buy-PT P/A man T fish LOC store
      'The man bought *the fish* at the store.'
   c. B<in>il-han ng lalaki ng isda **ang tindahan.**
      /PFV/-buy-LT P/A man P/A fish T store
      'The man bought fish *at the store.*'

As was shown by Himmelmann (1987), trigger morphology can be analyzed as a certain type of morphological derivation that determines the orientation of a nominalized action to a participant with its specific semantic role. Thus, *b<um>ili* in (9a) can be analyzed as a nominal with agent orientation (cf English 'buyer'). Analogously, *b<in>ili-ϕ* in (9b) corresponds to 'the object that is bought' and *b<in>il-han* in (9c) to 'the place of buying.' In such an analysis, the clause-initial position has the function of a nominal predicate that is followed by the subject marked by the trigger *ang* (on subjects in Tagalog, cf Kroeger, 1993). Thus, the declarative construction as a whole takes the function of an equational construction.

This analysis is further supported by the marker *ng*, which is also used as a marker of possession:

10. lapis ng bata
    pencil POSS child
    'the pencil of the child'

With this background, example (9a) can be analyzed as 'The man is a buyer of fish at the store.' Similarly, (9c) corresponds to 'the store is the buying place for fish of the man,' etc. If this analysis is correct, all the slots in the equational constructions discussed in (9) are nouns. Thus, Tagalog can be described as a language whose constructional inventory is characterized exclusively by reference-related structures. From such a perspective, one may think of describing Tagalog as an extremely rigid language in terms of Hengeveld (2013; cf type 7 in **Table 3**). The problem with this analysis is that Tagalog only consists of nouns, a fact which falsifies the parts-of-speech hierarchy in (5).

Contentives: The Case of Tongan

Tongan (Austronesian: Central-Eastern Malayo-Polynesian: Oceanic: Polynesian) is described as a language whose lexicon does not preclassify content words for syntactic categories. Irrespective of their semantic class (object, action, property) and their morphology, lexical items of Tongan can occur in the verbal slot characterized by tense marking and in the nominal slot characterized by an article. Thus, the action-denoting lexeme *lele* 'run' takes the V-slot in (11a) and the N-slot in (11b). Similarly, the object-denoting lexeme *fefine* 'woman' is in the nominal position in (11a) and in the verbal position in (11b). Notice that Tongan is a verb-initial language:

11. Tongan (Broschart, 1997, p. 134):
    a. na'e **lele** e kau **fefiné.**
       PST run ART PL.HUM woman
       'The women were running.'
    b. na'e **fefine** kotoa e kau **lelé.**
       PST woman all ART PL.HUM run
       'The ones running were all female.'

This example differs from (6) inasmuch as the meaning of the lexemes can be predicted from their position. The lexeme *fefine* in (11b) means 'to be a woman/women,' while the lexeme *lele* means 'the one(s) who run(s)' (the diacritics on the clause-final words in (11) mark clause-final accents). Thus, Croft's (1991, 2000, 2001) argument against the analysis of Tongan as a highly flexible language in terms of Hengeveld (1992, 2013) does not work with (11). The fact that there may be a certain percentage of lexemes which show unpredictable semantic differences in different syntact functions as in (6) is not an argument against the overall analysis of a language as highly flexible. A limited number of instances of lexical specification as a separate process of lexicalization is always to be expected (cf Bisang, 2013 on classical Chinese).

Evans and Osada (2005) in their paper on Mundari (Dravidian) discuss three criteria that claims of the lack of noun/verb distinction must meet. Since they concentrate on the contentive type, their paper will be discussed here. Compositionality is the first criterion. Differences in meaning of the individual lexemes must be fully derivable from the meaning of the syntactic slots in which they occur. It seems to hold for most lexical items of Tongan but there are exceptions (cf example (6)) and it holds for classical Chinese (Bisang, 2008). Bidirectionality requires that if an object-denoting lexeme can take the V-position, an action-denoting lexeme

should also be able to occur in the N-position. As can be seen from (11), this condition is met. Finally, the lack of noun/verb distinction must hold for all the relevant words of the lexicon (exhaustiveness; cf my above remark on the acceptability of a limited number of exceptions).

On the Role of Morphology

Even though the noun/verb distinction is irrelevant for the lexicon/syntax mapping in extremely rigid languages like classical Nahuatl and Tagalog and contentive languages like Tongan, this distinction remains important at the level of morphology in each of these three languages. In none of them, the noun/verb distinction is obsolete.

In *classical Nahuatl*, the noun/verb distinction shows up in tense/aspect marking and in the number of arguments (Launey, 1994, p. 201). Quite a few tense-aspect markers only occur with action-denoting (and property-denoting) lexemes (e.g., *-z* for future, *-c* for perfective, *ō* for completed action, etc.; Launey, 1994, p. 28). The differences in terms of argument structure can be seen from **Table 4**. Object-denoting lexemes differ with regard to their relational structure from action-denoting lexemes. Object-denoting lexemes have only one relation (S-argument) plus a possessor argument (last column of **Table 4**), while lexemes denoting transitive actions have an additional secondary argument (P in **Table 4**). Finally, there are a considerable number of morphemes which are clearly associated with nouns. The most prominent marker which occurs exclusively with object-denoting lexemes is *-tl/-tli* [*-ɬ/-ɬi*] (Launey, 1994, p. 207–208).

In Tagalog, the noun/verb distinction can be seen from the compatibility with trigger morphology. Object-denoting lexemes like *lalaki* 'man,' *bata* 'child,' and *isda* 'fish' (cf (9)) cannot take trigger morphology, while action-denoting (and property-denoting) words like *bili* 'buy' can (Himmelmann, 2008). Thus, one can conclude that Tagalog is a language with distinct morpholexical categories but no distinct terminal syntactic categories that are associated with the syntactic slots for nouns and verbs.

In Tongan, there are lexical 'paradigms' (Broschart, 1997, p. 143) in the sense of sets of lexemes that can be combined with certain morphemes but these paradigms cannot be used for mapping into syntax. The morpheme *-Canga* (marker of domain) can be suffixed to action-denoting lexemes (*pule* 'govern' –> *pule'anga* 'government'), to property-denoting lexemes (*motu'a* 'old' –> *motu'a'anga* 'reason for having aged'), and to object denoting lexemes (*api* 'home' –> *apit'anga* 'homestead'). In spite of the nominal properties of these words, they can also take the verbal position characterized by its TAM marking.

Conclusion

The definition of parts of speech depends on four prerequisites (Criteria for Defining Parts of Speech section). Croft's (1991, 2000, 2001) approach only includes semantics, pragmatics, and form (Croft (1991, 2000, 2001) section), while Hengeveld's (1992, 2013) approach also includes lexicon/syntax mapping (Functional Grammar (Hengeveld, 1992, 2013) section). The discussion of some language-specific and typologically remarkable cases can only be understood if the fourth prerequisite of lexicon/syntax mapping is included (cf classical Nahuatl, Tagalog, and Tongan in the Short Case Studies and Problems section). From such a perspective, Hengeveld's approach is more comprehensive.

Both approaches offer typological generalizations. Croft's conceptual space for parts of speech (**Table 1**) offers a universal framework for the cross-linguistic comparison of parts-of-speech systems. Since this conceptual space does not include lexicon/syntax mapping, it cannot distinguish extremely rigid systems (type 7) from contentive systems. Hengeveld's parts-of-speech hierarchy in (5) can fully integrate this distinction but it may be empirically falsified by Tagalog (Tagalog and the Parts-Of-Speech Hierarchy section), a language with an extremely rigid parts-of-speech system that is based exclusively on nouns and reference-related structures. Hengeveld's approach additionally offers interesting claims of universal correlations between the seven parts-of-speech types and other grammatical domains of language (Functional Grammar (Hengeveld, 1992, 2013) section). To what extent they hold needs to be seen. At least the generalization that contentive-type languages should be either verb-initial or verb-final (Functional Grammar (Hengeveld, 1992, 2013) section) is problematic from the perspective of classical Chinese, which is a precategorial language with SVO word order.

Croft's and Hengeveld's approaches are the most thorough accounts of parts of speech from a universal perspective. A look at individual languages and the discussions on their parts-of-speech systems by specialists reveals that there still are some problems that require more research and theoretical thinking. What is interesting is the fact that the noun/verb distinction shows up in the morphological properties even of those languages which are sometimes claimed to have no noun/verb distinction from the perspective of lexicon/syntax mapping (cf the On the Role of Morphology section).

See also: Comparative Method in Linguistics; Functional Approaches to Grammar; Iconicity in Linguistics; Information Structure in Linguistics; Lexicon; Markedness in Linguistics; Morphology in Linguistics; Semantics; Syntax (General).

Bibliography

Bisang, W., 2008. Precategoriality and syntax-based parts of speech: the case of late Archaic Chinese. Studies in Language 32, 568–589.

Bisang, W., 2011. Word classes. In: Song, J.J. (Ed.), The Oxford Handbook of Linguistic Typology. Oxford University Press, Oxford, pp. 280–302.

Bisang, W., 2013. Word class systems between flexibility and rigidity: an integrative approach. In: Rijkhoff, J., Van Lier, E. (Eds.), Flexible Word Classes. Typological Studies of Underspecified Parts of Speech. Oxford University Press, Oxford, pp. 275–303.

Broschart, J., 1997. Why Tongan does it differently: categorial distinctions in a language without nouns and verbs. Linguistic Typology 1, 123–165.

Croft, W.A., 1991. Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information. Chicago University Press, Chicago.

Croft, W.A., 2000. Parts of speech as typological universals and as language particular categories. In: Vogel, P.M., Comrie, B. (Eds.), Approaches to the Typology of Word Classes. Mouton de Gruyter, Berlin, pp. 65–102.

Croft, W.A., 2001. Radical Construction Grammar. Syntactic Theory in Typological Perspective. Oxford University Press, Oxford.

Croft, W.A., 2003. Typology and Universals, second ed. Cambridge University Press, Cambridge.

Dixon, R.M.W., 2004. Adjective classes in typological perspective. In: Dixon, R.M.W., Aikhenvald, A.Y. (Eds.), Adjective Classes. A Cross-Linguistic Typology. Oxford University Press, Oxford, pp. 1–49.

Evans, N., Osada, T., 2005. Mundari: the myth of a language without word classes. Linguistic Typology 9, 351–390.

Foley, W.A., Van Valin Jr, R.D., 1984. Functional Syntax and Universal Grammar. Cambridge University Press, Cambridge.

Haspelmath, M., 2001. Word classes and parts of speech. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Elsevier Science Ltd., Oxford, pp. 16538–16545.

Hengeveld, K., 1992. Non-verbal Predication. Theory, Typology, Diachrony. Mouton de Gruyter, Berlin.

Hengeveld, K., 2013. Parts-of-speech systems as a basic typological determinant. In: Rijkhoff, J., Van Lier, E. (Eds.), Flexible Word Classes. Typological Studies of Underspecified Parts of Speech. Oxford University Press, Oxford, pp. 31–54.

Hengeveld, K., Rijkhoff, J., Sievierska, A., 2004. Parts of speech systems and word order. Journal of Linguistics 40, 527–570.

Himmelmann Jr., N.P., 1987. Morphosyntax und Morphologie—Die Ausrichtungsaffixe im Tagalog. Fink, München.

Himmelmann Jr., N.P., 1997. Deiktikon, Artikel, Nominalphrase. Zur Emergenz Syntaktischer Strukturen. Niemeyer, Tübingen.

Himmelmann Jr., N.P., 2008. Lexical categories and voice in Tagalog. In: Austin, P., Musgrave, S. (Eds.), Voice and Grammatical Relations in Austronesian Languages. Center for the Study of Language and Information (CSLI) Publications, Stanford: California, pp. 246–293.

Hopper, P.J., Thompson, S.A., 1984. The discourse basis for lexical categories in universal grammar. Language 60, 703–752.

Jacobsen Jr., W.H., 1979. Noun and verb in Nootkan. In: Efrat, B.S. (Ed.), The Victoria Conference on Northwestern Languages. British Columbia Provincial Museum, Victoria, BC, pp. 83–155 (British Columbia Provincial Museum HeritageRecord No 4).

Kroeger, P., 1993. Phrase Structure and Grammatical Relations in Tagalog. Center for the Study of Language and Information (CSLI) Publications, Stanford, California.

Lallot, J., 1998. La grammaire de Denys le Thrace. CNRS Éditions, Paris.

Langacker, R.W., 1987. Nouns and verbs. Language 63, 53–94.

Launey, M., 1994. Une grammaire omniprédicative. Essai sur la morphosyntaxe du nahuatl classique. CNRS Éditions, Paris.

Ramat, P., Ricca, D., 1998. Sentence adverbs in the languages of Europe. In: van der Auwera, J., Ó Baoill, D.P. (Eds.), Adverbial Constructions in the Languages of Europe. Mouton de Gruyter, Berlin, pp. 187–275.

Sasse, H.-J., 1993a. Syntactic categories and subcategories. In: Jacobs, J., von Stechow, A., Sternefeld, W., Vennemann, T. (Eds.), Syntax: An International Handbook of Contemporary Research. Mouton de Gruyter, Berlin, pp. 646–686.

Sasse, H.-J., 1993b. Das Nomen – eine universale Kategorie? Sprachtypologie und Universalienforschung 46, 187–221.

Schachter, P., 1993. Tagalog. In: Jacobs, J., von Stechow, A., Sternefeld, W., Vennemann, T. (Eds.), Syntax: An International Handbook of Contemporary Research. Mouton de Gruyter, Berlin, pp. 1418–1430.

Tchekhoff, C., 1981. Simple Sentences in Tongan. Australian National University, Canberra.

Van Lier, E., 2009. Parts of Speech and Dependent Clauses: A Typological Study (PhD dissertation), University of Amsterdam; LOT Dissertation Series no. 221. Netherlands Graduate School of Linguistics, Utrecht. http://www.lotpublications.nl/index3.html.

Wierzbicka, A., 2000. Lexical prototypes as a universal basis for cross-linguistic identification of 'parts of speech'. In: Vogel, P.M., Comrie, B. (Eds.), Approaches to the Typology of Word Classes. Mouton de Gruyter, Berlin, pp. 285–317.

Party Identification and Nonpartisanship

Russell J Dalton, Center for the Study of Democracy, University of California, Irvine, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Sanders, volume 16, pp. 11099–11103, © 2001, Elsevier Ltd.

Abstract

Party identification (PID) is defined as an early-socialized, enduring, affective, psychological identification with a specific political party. It was detected by some of the first empirical studies of electoral choice in the United States. PID receives such importance because it can structure a person's view of the political world, provide cues for judging the candidates and issues of the day, shape voting choice, influence patterns of political participation, and promote stability in voting behavior. One can arguably claim that PID is the most important concept in modern electoral behavior research.

Origin and Evolution of the Concept

Angus Campbell and his colleagues (1960) based at the University of Michigan directed the first American national election surveys in the 1950s. As they examined how voters made their choices during an election, they noted that many people entered the campaign with their decision already made. Often people attributed their votes to long-standing family traditions, or values they had held across many elections. These partisan loyalties appeared to be a central part of individual belief systems, serving as a source of political cues for other attitudes and behaviors. The Michigan researchers described these partisan attachments as a sense of *party identification* (PID): a long-term, affective, psychological identification with one's preferred political party. Social identity theory argues that these partisan ties are similar to identifications with a social class, religious denomination, or other social group. Party attachments are distinct from voting preferences, which explains why some Americans vote for the presidential candidate of one party while expressing loyalty to another party. Indeed, the conceptual independence of voting and PID gives the latter its significance. PID emerged as a central element in these researchers' sociopsychological model of electoral behavior.

After the discovery of party ID in the United States, the concept was exported to other democracies where the Michigan scholars helped to establish new electoral surveys. In several cases, researchers had problems finding an equivalent measure of partisanship in multiparty systems or in nations where the term 'partisanship' holds different connotations for the voters. In Britain, for example, partisan loyalties run deep, and with a single parliamentary vote there are not the ticket-splitting choices that separate long-term loyalties from short-term vote choices as in the United States. In addition, the concept of a partisan 'independent' is not as common in other nations as it is in the United States, so there might not be a neutral nonpartisan category for people who did not feel close to any party. Researchers could not simply translate the American party ID question into French or German; they had to find a functional equivalent for measuring partisan attachments.

For nearly a decade in the 1970s and early 1980s researchers debated whether PID was a distinctly American phenomenon. Different surveys experimented with ways to ask about partisanship, and tracked the stability of measures over time. Eventually most election experts agreed that voters in established democracies hold some form of party allegiances that endure over time and shape their views of politics. These identities strongly influence perceptions of parties and electoral politics, judgments about the government of the day, and affect patterns of political participation.

How Partisanship Develops

Partisanship is similar to other personal identities; it reflects early life learning and the social conditions of the individual. Studies of young children found that many develop basic partisan ties at a very early age, often during the primary school years. Children learn party loyalties before they can understand what the party labels stand for – a process similar to the development of many other social identities such as religion or class. These early party attachments provide a reference structure for future political learning (which often reinforces early partisan biases).

In established democracies, parents play a central role in the socialization of these values. Research that compares the party identifications of parents and their children routinely finds high levels of partisan agreement in stable party systems. Parents successfully transmit their partisanship to their children because party loyalties are formed when parents are the dominant influence in a child's life. Parents commonly expose their children to partisan cues since political discussion frequently includes some partisan content. Consequently, it does not take long for a child to figure out his or her parents' partisan leanings from their reactions to television news and statements in family discussions.

Similarly, the parents' social situation provides politically relevant cues on what party is supported by others in one's church, social class, ethnicity, or region. Adolescents and young adults see which party their friends and coworkers support, which often reinforce their initial party ties. Thus, party identities become enmeshed in a web of other social identities that tend to persist through life. As a result, studies routinely find that party ties tend to strengthen with age – or more precisely with continued experience in supporting one's

early preferred party. For example, the 2012 American National Election Study reported that 48% of Americans under age 30 had a party identification, compared to 60% among those over age 60. The same age pattern appears in most established democracies.

The situation is different in many of the new democracies that have emerged as part of the democratization wave that began in the late 1980s and early 1990s. In these nations, such as the new democracies of Central-Eastern Europe, East Asia, and Africa, the democratic experience is new to parents and their children. Indeed, parents might struggle to adapt to the change in political regimes. Consequently, party ties lack the element of family and social traditions that reinforce these identities in established democracies. Parents cannot socialize partisan values that they lack themselves. The political parties running in elections are also new, and changes in the partisan lineup over time are likely. Research on partisanship in new democracies suggests that party ties will initially be weak, and will be learned slowly over subsequent elections if the party system stabilizes and families and social groups develop firm party loyalties. Thus, new democracies may require a generation or more to develop to levels of partisanship observed in established democracies.

Levels of Partisanship

One of the contentious issues involving partisanship is how to measure these orientations. PID means a long-term psychological attachment to a political party. To tap into these identities, the Michigan team developed what is now the standard question assessing PID in the United States:

> Generally speaking, do you think of yourself as a Republican, a Democrat, an Independent, or what?

If the respondents say they are a Republican or a Democrat, they are asked if they would call themselves strong (Republican/Democrat) or not very strong (Republican/Democrat). If they say they are independent, they are asked if they think of themselves as closer to the Republican or to the Democratic Party. American researchers use this question to create a seven-point party ID scale running from being a strong Democrat at one end, an independent close to neither party in the middle, and a strong Republican at the other end.

Cross-national research found that this question could not be used in some party systems because specific elements of the American question did not capture partisanship in multiparty parliamentary systems. After much debate, researchers developed questions to tap these allegiances that reflect national traditions and party system conventions. The now standard question on partisanship in Germany illustrates one nation's alternative:

> Many people in the Federal Republic lean toward a particular party for a long time, although they may occasionally vote for a different party. How about you: Do you in general lean toward a particular party? Which one?

This is an example of a functional equivalent to the American question even though the wording is different to reflect conditions in another nation. Other nations adopted their own measure (or alternative measures) of partisanship. Questions on PID are now asked in the election studies of virtually all contemporary democracies. There are predictable variations in the strength of these identities across nations, especially as a function of the longevity of the party system and the characteristics of the party system. In most nations, many people approach a new election with a predisposition toward a favored party already in place.

Because of these variations in political context and the phrasing of the partisanship question, it is complicated to compare levels of partisan cross-nationally. The best evidence on the extent of partisanship in a nation comes from the respective national election studies that ask a question suited to national conditions. But this is little help in comparing levels of partisanship cross-nationally if each nation uses a differently worded question. So instead we turn to the 2012 European Social Survey that asked an alternative party attachment question in about two dozen European democracies:

> Is there a particular political party you feel closer to than all the other parties? Which party is that? How close do you feel to this party? Do you feel that you are very close, quite close, not close, or not at all close?

This question exchanges the idea of a long-term partisan identity for the concept of closeness to a party. Closeness should produce a 'softer' measure of partisanship, which might make it easier to express a party attachment asked separately from immediate vote choice. Another advantage is that this question has been asked in recent surveys in the United States.

Figure 1 displays the percentage saying they are close to any party across several European nations and the United States. There is considerable variation in the level of partisanship across nations, ranging from about two-thirds in Denmark, Sweden, and Norway to less than a third in the nations at the bottom of the figure. Some of this cross-national variation might come from the inevitable differences in how a standard question is translated into each language or the specific events occurring in the nation. Overall, multiple waves of the European Election Survey find patterns similar to the evidence presented here.

Partisanship in the United States ranks above the European average. Given the past emphasis on the strong partisanship of Europeans, especially in proportional representation parliamentary systems, this is somewhat surprising. However, it may reflect the greater functional value of partisanship in the United States because of the long ballots and multiple choices that voters face. A Briton who casts 5–6 votes in a 5-year electoral cycle may feel less need to psychologically identify with a party than an American partisan who casts more than a hundred votes for local, state, and federal offices over this same period.

The other notable feature of the figure is the contrast between established democracies and the new democracies in Central-Eastern Europe. Four of the five nations with the lowest percentage feeling close to a particular party are from the East. Ireland is the only Western democracy in this group, and its low

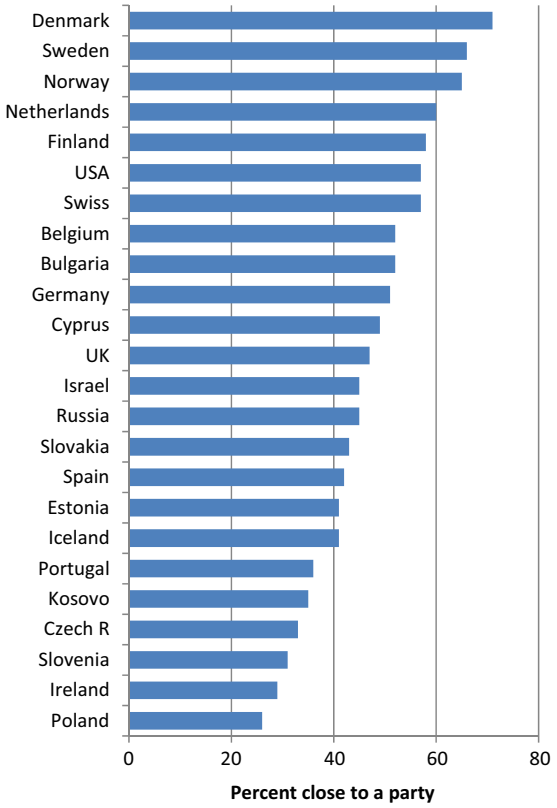

Figure 1 Close to a particular political party. Sources: European Social Survey 2012; the U.S. results are from the 2008 American National Election Studies.

level of partisanship probably reflects a popular backlash against parties because of the economic collapse in 2009. Conversely, only one of the 10 nations with the highest level of party closeness is from the East. All else being equal, one might expect party ties to slowly grow in these new democracies as party choices stabilize, voting experience grows, and people develop a record of repeatedly voting for a preferred party. But even two decades after democratization, this East–West gap is still substantial.

Consequences of Partisanship

Politics can be a complicated undertaking for citizens; it takes considerable time and effort to follow the details of elections, the policy choices offered by the parties, and the veracity of campaign pronouncements. Especially in a parliamentary system with a large number of parties, making a decision can be a daunting task. PID first gained prominence as a method to simplify this process, to provide a cue or heuristic to guide political behavior. PID helps to make politics 'user friendly.' When the political parties take clear and consistent policy positions, the party label provides an information shortcut on how 'people like me' should decide. Once voters decide which party generally represents their interests, this single piece of information can act as a perceptual screen – guiding how they view events, issues, and candidates. A policy advocated by one's party is more likely to meet with favor than one advocated by another party.

Compared to social group cues such as class and religion, PID is a more valuable heuristic. Party cues are relevant to a broader range of phenomena because parties are so central to democratic politics. Issues and events frequently are presented to the public in partisan terms, as the parties take positions on the issues of the day or react to the statements of other political actors. People vote for parties or party candidates at elections. Governments are managed by partisan teams. So, reliance on partisanship may be the ultimate example of the satisficing model of politics.

There are many examples of how people use partisan cues to fill in voids in their political knowledge. For instance, if they are unfamiliar with a specific public policy, knowing which party favors or opposes the policy is a strong cue on what position to take. Similarly, party ties shape evaluations of new political candidates, because their party generally signals a broad political philosophy. Party cues also shape one's evaluations of the past and expectations for the future. When one's party wins office, for example, its partisan supporters are more likely to improve their predictions about the nation's political and economic future – even before the new government is installed. Many voters view politics through this perceptual screen.

Party ties also mobilize people to become politically active. Just like loyalty to a sports team, attachment to a political party encourages an individual to become active in the political process to support his or her side. The 2012 American National Election Studies found that turnout was 26% higher among strong partisans than among independents. In addition, strong partisans are more likely to try to influence others, to display campaign materials, to attend a rally, or to give money to a candidate during the campaign. Partisanship functions in a similar way in other established democracies. Strong partisans voted at a higher rate in the 2009 German Bundestag elections, and they were several times more likely to participate in campaign events and twice as likely to try to persuade others how to vote. Similar patterns exist for most established democracies.

Party identifications also structure voting choices. Partisanship means that a voter has a predisposition to support his or her preferred party. Partisanship has been described as the basis for a 'normal vote' – the vote expected when other factors in the election are evenly balanced. If other factors come into play, such as issue positions or candidate images, their influence can be measured by the change in preferences from initial partisan predispositions. For the unsophisticated voter, a long-term partisan loyalty and repeated experience with one's preferred party provides a clear and low-cost cue for voting. Even for the sophisticated citizen, a candidate's party affiliation normally signifies a policy program that serves as the basis for reasonable electoral choice.

Generally, there is often a very close relationship between PID and voting in parliamentary elections. In the 2010 British House of Commons elections, for example, 84% of partisans voted for their own party. The small number of elected offices in most established democracies and the party-centered nature of parliamentary campaigns tend to narrow the separation between partisanship and vote.

The American citizen, in contrast, has to cope with a vast number of candidates seeking a variety of offices at federal, state, and local levels. This encourages Americans to become conscious of a generalized tie to one party. The separation between attitudes and behavior is therefore the most noticeable in American elections, especially when voters are asked to make serial choices for local, state, and federal offices. In highly visible and politicized presidential elections, candidate images and issue appeals have the potential to counteract partisan preferences. The success of Republican presidential candidates from Ronald Reagan to George W. Bush, for example, occurred because they attracted defectors from the Democratic majority as well as independents. Similarly, the two-candidate runoff in French presidential elections is decided by the size of the vote the candidates can attract from parties other than their own.

At the national level, strong partisan ties can be a stabilizing influence on electoral politics. If most people identify with one of the established parties, then this lessens the potential for voters to be attracted to new parties or demagogic leaders. More generally, widespread partisan ties dampen the impact of short-term political events on election outcomes and limit the potential electoral appeal of new political personalities. Extensive partisanship among the electorate thus works to stabilize party alignments and lessen electoral change.

In summary, partisanship is the ultimate heuristic because it

- creates a basis of political identity
- provides cues for evaluating political events, candidates, and issues
- mobilizes participation in campaigns and election turnout
- provides cues on voting preferences, and
- stabilizes voting patterns for the individual and the party system

Research Controversies

Because PID is so central to understanding citizen political behavior, there is continuing research on the nature of partisanship and its correlates. One of the most visible research controversies focuses on potential changes in the levels of partisanship. American election studies noted a decline in the percentage of party identifiers starting in the mid-1960s. With some variability, this trend has continued to the present. Where once only a quarter of Americans lacked a party identity, in 2012 roughly 40% were independents; this means nonpartisans are a plurality compared to either Republican or Democratic identifiers.

Eventually, similar declines in partisanship became apparent for other established democracies in Europe, North America, and the Pacific Rim. Long-term election studies for these nations broadly show a shrinking number of partisans or a weakening of the strength of party identities. There are parallel declines in the number of people who are members of a political party, and favorable opinions of political parties have broadly declined. These patterns provide evidence for what Dalton (2012) describes as *partisan dealignment* – a persisting decline in the public's level of partisanship.

If partisanship structures the political behavior of average citizen, as many experts argue, then the broad cross-national evidence of dealignment over the last several decades may be a cause of concern. Significantly fewer people now approach politics with a fixed party loyalty – although many people do still have party allegiances. When a plurality of citizens lacks party identities, this has the potential to reshape electoral behavior and the nature of citizen politics.

Several electoral experts have questioned the existence of these trends in the United States and other democracies. The most common criticism in the United States is that many people claim to be nonpartisans because it is fashionable to be independent, but then they actually behave like hidden partisans. Other research demonstrated that partisans in the United States remain loyal voters, so the overall impact of party ID on vote choice has changed only slightly. Similar doubts were raised about the dealignment trends in European democracies. The centrality of partisanship is so great that many scholars doubted that these ties were really weakening.

With additional evidence across time and across nations, the evidence of dealignment has mounted. It is not a phenomena linked to the political history of any one nation, such as a rejection of American political parties because of the issue controversies or scandals of the past several decades. Moreover, there are behavioral manifestations of weakening party ties. For example, there is cross-national evidence that fewer voters enter elections with standing party preferences that endure to Election Day, and more voters are deciding later in the campaign based on the events of the election. With weaker party ties, turnout in elections has also generally declined in established democracies. Other research indicates that contemporary publics are more likely to shift their party preferences between elections, because fewer have firm party loyalties. Thus, the volatility in electoral results has increased. And since party ties systemically stabilize party systems, the erosion of party ties has been accompanied by the emergence of new political parties of all colors that can capture these floating voters. Because of its sources in the changing nature of contemporary societies and citizen values, dealignment is likely to be a continuing feature of electoral politics in established democracies.

This evidence of dealignment stimulated concerns about the consequences for electoral politics and democracy overall. Previous studies had found that nonpartisans were largely disengaged from politics, ill-informed on the issues of the day, and seldom voted. More citizens like this would not benefit the democratic process. However, the rise of nonpartisans is disproportionately concentrated the young and better educated, who follow politics and participate despite lacking party loyalties. Thus the growth in nonpartisans has the potential to increase the number of voters who more closely follow the theoretical model of an informed, rational voter.

Much of the dealignment literature focuses on established democracies, attributing these trends to broad forces of social and political change – such as changing norms of citizenship, increased education, and a more assertive media. But similar forces are at work in many emerging democracies. There is some evidence that partisanship is not developing in many new democracies, as prior theorizing would predict. Instead, citizens are hesitant to develop party allegiances, and some elites react by stressing personality over party. While established democracies are undergoing dealignment, many new

democracies are failing to develop party alignments in the first place, which has generated research on this irregular learning process and its consequences.

Another research controversy focuses on whether PID actually fits the assumptions of the initial sociopsychological model of electoral behavior. The original formulation of partisanship defined it as an affective psychological identity, which implied it would show relatively high stability through the life span and structure perceptions of the political world to generate this persistence. In contrast, Fiorina's (1981) rational choice approach suggests that partisanship can be treated as a 'running total' of accumulated partisan experience. Or in Bayesian terms, individuals may begin with a prior partisanship based on early life socialization, but then this partisanship is constantly updated by successive experience. If one voter supports his preferred party in an election, then his party ID might strengthen. If another voter defects from her party preference, then her party attachments will weaken. This is a much more dynamic view of partisanship.

The rational choice approach was a positive step forward in understanding partisanship, but it is not an either/or choice. Repeated interviews with the same people demonstrate that partisanship is one of the most stable of political attitudes, more resistant to change than major economic or cultural policy positions. Yet party ties do sometimes change. The issue is the relative balance between stability and change, the force needed to produce lasting change in partisan loyalties, and how this varies across different population subgroups. For individual citizens, the stable element of party identity is typically stronger than the forces for change. Thus, changes in party vote shares between elections often come from nonpartisans or new voters who lack strong partisan loyalties.

A more recent critique by Lavine et al. (2012) questions the value of PID as a heuristic for making informed democratic choices. They argue that partisanship might not just color political views, but may act as blinders that shield partisans from discordant information that is necessary for making informed choices. If friends of my friends are always friends, and friends of my enemies are always enemies, then partisans are unlikely to see alternatives for discourse and compromise – and the making of better choices. Indeed, other scholars have stressed the vulnerability of heuristics as a basis of informed democratic choice. Even if it is right most of the time, blind heuristics cannot identify when the wrong choice is being presented. This is especially problematic when citizens have limited political sophistication to make informed choices without partisan cues.

In summary, research clearly suggests that if one wanted to understand the political behavior of individuals with a single question in a survey, then PID would be the best question to ask. It provides a conceptual framework for many citizens, shaping their views of politics, and their political activities. Thus even if levels of partisanship are dropping as a result of dealignment, partisanship provides the touchstone for understanding political behavior. After reviewing four decades of electoral research on partisanship, Weisberg and Greene (2003) concluded "Party identification is the linchpin of our modern understanding of electoral democracy, and it is likely to retain that crucial theoretical position."

See also: Voting, Explanations of: Social Class; Voting: Turnout.

Bibliography

Bartle, John, Bellucci, Paolo (Eds.), 2009. Political Parties and Partisanship: Social Identity and Individual Attitudes. Routledge, London.

Campbell, Angus, Converse, Philip, Miller, Warren, Stokes, Donald, 1960. The American Voter. Wiley, New York.

Dalton, Russell, 2012. The Apartisan American: Dealignment and Electoral Change. CQ Press, Washington, DC.

Dalton, Russell, Wattenberg, Martin (Eds.), 2000. Parties without Partisans: Political Change in Advanced Industrial Democracies. Oxford University Press, New York.

Fiorina, Morris, 1981. Retrospective Voting in American National Elections. Yale University Press, New Haven, CT.

Green, Donald, Palmquist, Bradley, Schickler, Eric, 2002. Partisan Hearts and Minds: Political Parties and the Social Identities of Voters. Yale University Press, New Haven.

Hajnal, Zoltan, Lee, Taeku, 2011. Why Americans Don't Join the Party: Race, Immigration, and the Failure of Parties to Engage the Electorate. Princeton University Press, Princeton.

Lavine, Howard, Johnston, Christopher, Steenbergen, Marco, 2012. The Ambivalent Partisan: How Critical Loyalty Promotes Democracy. Oxford University Press, Oxford.

Lewis-Beck, Michael, Jacoby, William, Norpoth, Helmut, Weisberg, Herbert, 2008. The American Voter Revisited. University of Michigan Press, Ann Arbor.

Thomassen, Jacques (Ed.), 2005. The European Voter. Oxford University Press, Oxford, UK.

Weisberg, Herbert, Greene, Steve, 2003. The political psychology of party identification. In: MacKuen, Michael, Rabinowitz, George (Eds.), Electoral Democracy. University of Michigan Press, Ann Arbor.

Party Systems

Peter Mair, Leiden University, Leiden, The Netherlands

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 11106–11108, © 2001, Elsevier Ltd.

Abstract

A party system is the system of interactions between political parties that results from their mutual competition or cooperation. For party systems to exist there must therefore be more than one party involved. Beyond the requirement for a plurality of parties, party systems can be said to exist only when these interactions between the parties are patterned, familiar, and reasonably predictable. A plurality of parties is therefore a necessary but not sufficient condition for the existence of a party system. When it comes to the classification of party systems, the most common and traditional approach is based simply on the number of parties in competition, usually distinguishing between two-party systems, on the one hand, and multiparty systems, on the other. Greater insight is offered by the more elaborate classification developed by Sartori, which takes into account not only the number of parties, but also the ideological distance that separates them. The core of all party systems, and their most defining feature, involves the structuring of competition for executive office. Party systems as systems are strongest and are most institutionalized when this structure of competition is closed. They are weakest, and more or less nonexistent, when it is wholly open.

A party system is the system of interactions between political parties that results from their mutual competition or cooperation (see Sartori, 1976, p. 44). For party systems to exist there must therefore be more than one party involved. One-party states do not have party systems. Beyond the requirement for a plurality of parties, party systems can be said to exist only when these interactions between the parties are patterned, familiar, and reasonably predictable. A plurality of parties is therefore a necessary but not sufficient condition for the existence of a party system. Parties that exist alongside one another, but that do not interact in any patterned or structured fashion, do not make for a party system as such.

While it may seem difficult to conceive of situations in which a plurality of parties exists without at the same time embodying a set of patterned interactions, two types of limiting cases can be identified. The first is where each of the parties involved exists in an entirely self-sufficient world, based on a closed electoral constituency, with neither its survival nor performance being in any way effected by the other parties in the polity. Even in this limiting case, however, the absence of interactions at the electoral level might well be accompanied by the need to interact – through competition or cooperation – at the parliamentary or governmental level. Thus, while no electoral party system would exist, it might well prove possible to speak of a party system in parliament. The second type of limiting case is where a new multiparty democracy is being established more or less from scratch, such as in the wake of a democratic transition, and where neither the new parties themselves, nor their interactions, have stabilized sufficiently to allow the identification of a distinct pattern. In these situations, a party system may be in the process of being structured or institutionalized but it cannot yet be said to have definitely emerged. In their survey of party systems in Latin America, Mainwaring and Scully (1995) adopt a similar distinction in specifying party systems as either 'institutionalized' or 'inchoate.' Strictly speaking, however, an inchoate or noninstitutionalized party system is a contradiction in terms: to be a system is to be institutionalized.

Classifying Party Systems

Although scholars have paid relatively scant attention to what defines a party system as such, tending instead to assume the existence of a party system in all polities where there exists a plurality of parties, they have devoted considerable effort to distinguishing between different types of party system. The most conventional and frequently adopted approach to distinguishing party systems is based simply on the number of parties in competition, and the most common distinction involved here, which goes back to Duverger (1954), is very straightforward – that between two-party systems, on the one hand, and multiparty (i.e., more than two) systems, on the other. This particular classification was also believed originally to reflect a more fundamental distinction between more or less stable and consensual democracies, such as the United Kingdom and the United States, which were seen as typical of the two-party variety, and more or less unstable or conflictual democracies, such as Fourth Republic France, Italy, or Weimar Germany, which were seen as typical of the multiparty type. Although this simple association of numerical categories of party system with levels of political stability and efficacy was later undermined by reference to a host of smaller democracies that were characterized by both a multiplicity of parties and a strong commitment to consensual government, the core distinction between two- and multi-party systems has continued to be widely employed within the literature on comparative politics, although it is sometimes modified by taking into account not only the sheer numbers of parties in competition, but also their relative size. Thus, Blondel (1968), for example, uses the relative size of the parties to distinguish four types of party system: two-party systems, two-and-a-half-party systems, multiparty systems with a dominant party, and multiparty systems without a dominant party.

The most substantial attempt to move away from a primary reliance on the simple numbers of parties in competition was that of Sartori (1976, pp. 117–323), who combined counting the parties with a measure of the ideological distance that

separated them. Sartori's typology was the first that focused directly on the interactions between the parties – the 'mechanics' of the system – and hence on the differential patterns of competition and cooperation. Following this approach, party systems could be classified according to the number of parties in the system, in which there was a distinction between systems with two parties, those with up to some five parties (limited pluralism) and those with some six parties or more (extreme pluralism); and according to the distance that separated the parties lying at either extreme of the ideological spectrum, which would either be small ('moderate') or large ('polarized'). These two criteria were not wholly independent of one another, however, in that Sartori also showed that the format of the system, that is, the number of parties, contained 'mechanical predispositions,' that is, it could help determine the ideological distance, such that extreme pluralism could lead to polarization. When combined, the two criteria resulted in three principal types of party system: two-party systems, characterized by a limited format and a small ideological distance (e.g., the UK); moderate pluralism, characterized by limited or extreme pluralism and a relatively small ideological distance (e.g., Denmark or the Netherlands); and polarized pluralism, characterized by extreme pluralism and a large ideological distance (e.g., Weimar Germany, early postwar Italy, or pre-Pinochet Chile). Sartori also noted the existence of a 'predominant-party system,' in which one particular party, such as, for example, the Congress party in India, or the old Unionist party in Northern Ireland, consistently won a winning majority of parliamentary seats.

Party Systems and Structures of Competition

At the core of any party system is the competition for executive office. It is this that structures the party system in the first place, and facilitates its institutionalization. Moreover, albeit often implicitly, the competition for executive office is also the most important criterion employed in many of the various classifications of party systems. Two-party systems, for example, are referred to as such not because there only two parties present themselves to the voters – indeed, this is rarely if ever the case – but rather because only two parties matter when it comes to forming a government, be this in a presidential or a parliamentary system. In two-and-a-half or multiparty systems, by contrast, there are more than two parties that enjoy the potential access to executive office. Even within the more complex classification developed by Sartori, it is the competition for office that proves most important. Sartori's moderate pluralism, for example, involves competition between alternative coalition governments, whereas polarized pluralism involves a patterns in which a center party or parties is more or less permanently in office, with one or more extreme parties being permanently excluded from government.

It follows from this that party systems might best be understood not only in terms of some all-embracing classification, within which the question of the numbers of parties in competition obviously plays an important role, but also in terms of contrasting patterns of government formation (Mair, 1997, pp. 199–223). Three related criteria are important here. The first concerns the ways in which governments alternate with one another, where systems in which there is a regular process of wholesale alternation, in which incumbents are replaced wholly by former nonincumbents, may be separated from those in which alternation is only partially complete. Systems in which two alternative parties or coalitions alternate with one another (the USA, the UK, Fifth Republic France) may therefore be distinguished from those in which one or more parties leaves office, while another remains, albeit often with a new coalition partner (Belgium, the Netherlands). The second, relevant criterion here is familiarity, where systems in which government is always contested by the same parties or sets of parties may be contrasted with those in which patterns of government formation prove more innovative or promiscuous. In postwar Ireland, for example, voters had long become accustomed to facing a choice at election time between, on the one side, the powerful *Fianna Fáil* party, which always governed alone, or, on the other side, a coalition of all of the smaller parties. Once *Fianna Fáil* decided to opt for a coalition strategy, however, in 1989, government formation became more innovative, and the formerly familiar patterns broke down. The third criterion involves the degree of access to the government formation process, and the extent to which new parties can hope to win a place in the executive. New parties have always found it relatively easy to join innovative government coalitions in the Netherlands, for example. In the UK, by contrast, no party other than the Conservatives or Labour has gained access to government since 1945.

Putting these three criteria together enables party systems to be ranged along a notional continuum according to the degree to which the structure of competition is open or closed. At one extreme lie the wholly closed systems, such as the UK or the US, where alternation in government is always wholesale, where government formation processes are wholly familiar, and where new or third parties are always excluded. At the other extreme, it is difficult to speak of a party system at all: there is no discernible pattern in how governments alternate, the potential alternatives themselves are unfamiliar and may never have governed before, and access to office is, in principle, open to all. To travel from this latter extreme to the other is therefore to witness the progressive closure of the structure of competition, which is simply another way of saying that systemness itself increases. Party systems as systems are strongest and are most institutionalized when the structure of competition is closed. They are weakest, and more or less nonexistent, when this structure is wholly open.

To understand party systems in this way, and hence to treat the systemness of a party system, or its degree of institutionalization, as a variable, also affords an insight into the question of the persistence and change of party systems. Strongly institutionalized party systems will inevitably enjoy a bias towards persistence, not least because the parties at the core of such systems will have a vested interest in maintaining the existing structure of competition. It is for this reason, for example, that Lipset and Rokkan (1967) could speak of the 'freezing' of west European party systems in the wake of full democratization in the 1920s. Less institutionalized systems, on the other hand, will prove more susceptible to change. This approach also affords an insight into the relationship between party systems and individual parties, for it is not just the individual parties that together create the party system, but also the party system

that acts to reinforce the role and standing of the parties themselves.

See also: Dictatorship; Electoral Systems; Party Identification and Nonpartisanship; Political Parties, History of; Politics and Participation.

Bibliography

Blondel, J., 1968. Party systems and patterns of government in western democracies. Canadian Journal of Political Science 1, 180–203.

Duverger, M., 1954. Political Parties: Their Organization and Activity in the Modern State. Methuen, London.

Lipset, S.M., Rokkan, S., 1967. Cleavage structures, party systems and voter alignments: an introduction. In: Lipset, S.M., Rokkan, S. (Eds.), Party Systems and Voter Alignments. Free Press, New York.

Mainwaring, S., Scully, T.R., 1995. Introduction: party systems in Latin America. In: Mainwaring, S., Scully, T.R. (Eds.), Building Democratic Institutions: Party Systems in Latin America. Stanford University Press, Stanford, CA.

Mair, P., 1997. Party System Change: Approaches and Interpretations. Clarendon Press, Oxford, UK.

Sartori, G., 1976. Parties and Party Systems: A Framework for Analysis. Cambridge University Press, Cambridge, UK.

Wolinetz, S.B., Brookfield, U.T. (Eds.), 1998. Party Systems. Ashgate, Aldershot, UK.

Passion and Motivation

Julia Moeller, University of Helsinki, Helsinki, Finland
Jacquelynne S Eccles, University of California, Irvine, Irvine, CA, USA
Katariina Salmela-Aro, University of Helsinki, Helsinki, Finland
Julia Dietrich, University of Jena, Jena, Germany
Barbara Schneider, Michigan State University, MI, USA
Robert Grassinger, University of Augsburg, Augsburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article gives an overview about a decade of passion research (2003–2013). First, definitions and measurements of passion are introduced, differentiating between general models on the one hand and domain-specific definitions on the other hand. Second, the relations of passion to other constructs are addressed and the research findings concerning similar constructs, predictors, correlates, and outcomes of passion are summarized. Finally, open research questions and directions for future research are discussed.

Introduction

Some individuals persist in activities such as playing the piano or doing research over a long time, and some of them practice these activities continuously and intensely despite occurring obstacles, frustrations, and other aversive experiences (Ericsson et al., 1993). Yet, many of the individuals report being passionate about the respective activity and desire to engage and persist in their effortful practice (Bonneville-Roussy et al., 2011). It is known that persistent practice is needed during the process of becoming an expert (Ericsson et al., 1993), but little is known about the motivation that keeps some individuals on track while others drop out. One explanation for such motivation to persist in effortful activities is the construct passion. This article introduces definitions of passion and describes how passion has been measured, reviews existing research approaches and findings, and finally discusses open questions that should direct future research on passion.

Defining and Measuring Passion

There are many different definitions and measures of passion. The term passion has different connotations in everyday life, different cultures and disciplines and its connotations have changed much across the centuries (Dixon, 2003). Therefore it is complicated to reach a clear, consensual, and scientific definition of passion. Even within psychology, there is no consensual concept but a variety of passion definitions. Many authors use the term implicitly, presupposing that readers understand the meaning without further explication. This has led to terminological ambiguities: Everybody understands what passion means, but different people understand different things. Implicit definitions often use passion as a synonym to strong interest (e.g., Achter and Lubinski, 2005; Pressley et al., 1992; Gagné, 2007), consistently pursued interests (Duckworth et al., 2007), and strong and persistent motivation in learning and achievement contexts (e.g., Achter and Lubinski, 2005; Mills, 2003; Moraes et al., 2004; Noble et al., 2008; Wai et al., 2009). In addition, several explicit psychological passion definitions were proposed (Cardon et al., 2009b; Fredricks et al., 2010; Renzulli et al., 2006; Vallerand et al., 2003). Most explicit definitions include intense, mostly positive emotions, approach motivation, and long-term commitment as passion components. Some describe general aspects of passion (Fredricks et al., 2010), while others refer to specific types or components, or contexts of passion (Vallerand et al., 2003), some conceptualize passion as a hedonic experience (Vallerand et al., 2003; Fredricks et al., 2010), while others emphasize more the experience of approach motivation. The core characteristics of the different explicit passion definitions are described below.

Domain-Unspecific Models of Passion

The Dual Model of Passion: Harmonious versus Obsessive Passion

Very influential is the dual model of passion (Vallerand et al., 2003). The authors define passion as "a strong inclination toward an activity that people like, that they find important, and in which they invest time and energy" (Vallerand et al., 2003, p. 756). According to the dual model, passionate individuals identify with the activity. Different forms of identification supposedly cause two different types of passion. The adaptive harmonious passion (HP) describes the identification with an autonomously chosen activity, whereas the maladaptive obsessive passion (OP) describes the compulsion to engage in an activity that results from intra- or interpersonal pressure.

The dual model of passion has been studied with self-report questionnaires among individuals from child age to high-age adults, in different countries and languages, and in many different contexts and activities (e.g., sports, music, dance, political engagement, work contexts, teaching, and gambling; for an overview see Marsh et al., 2013). Vallerand et al. (2003) measure three aspects of passion: general passion criteria, HP, and OP. The general passion criteria (liking, finding important, investing time and energy, and labeling oneself as being passionate) were used in some studies to

distinguish between passionate and nonpassionate individuals (e.g., Bonneville-Roussy et al., 2011; Philippe et al., 2009b; Rip et al., 2012). The authors asked individuals about an activity, which they intensively practiced (play of music instruments, political engagement, and the individuals' favorite activity) and categorized individuals as passionate if they obtained above-scale midpoint scores on the mean-composite of the above-mentioned passion criteria. According to this cutoff, the vast majority of the investigated individuals was categorized as 'passionate' for the corresponding activity. HP and OP are strongly correlated with the general passion criteria, and the correlation between HP and OP ranges from low (e.g., Marsh et al., 2013) over moderate (e.g., Rip et al., 2006; Vallerand et al., 2007) to high (e.g., Séguin-Lévesque et al., 2003; Vallerand et al., 2006, 2003, 2008b). Most studies concerning the dual model of passion investigated differences between HP and OP. In contrast, differences between generally passionate and nonpassionate individuals are rarely investigated in these studies.

The Com.pass Model: Passion as Coincidence of Commitment and Desire

A definition of specific components of general passion was suggested by Moeller and Grassinger (2014a) as a result of a review and integration of the previous research on passion and commitment. In their commitment and passion (com.pass) model, the authors define passion as a coincidence of desire and commitment, where desire is defined as affect-intense approach motivation and commitment as motivation to persist, according to the investment model of commitment (Rusbult et al., 1998) and the related sport commitment model (Scanlan et al., 2009). In line with these models, commitment is subdivided into the subcomponents 'continuous intent to engage and persist,' 'long-term goals,' and 'identification with the activity.' These three components of commitment and the fourth passion component 'desire' can be measured with a self-report questionnaire developed by Moeller and Grassinger (2014b). These passion components and their relations to measures of commitment, personality traits, and other constructs have been investigated in samples of adolescents and young across different activity contexts (including soccer, poker, and dance) and in different countries and languages.

The com.pass model differs from the dual model of passion in its emphasis on the general degree rather than types of passion and its distinction between specific components of general passion. Furthermore, the com.pass model does not distinguish between HP and OP. A particularity is moreover the integration of the previously separated research lines on passion and commitment within the com.pass model, and the findings that established passion and commitment scales both measure the same construct. In contrast to the dual model of passion (Vallerand et al., 2003), the com.pass model defines the experience of affect-intense approach motivation ('wanting') as an essential component of passion.

Passion as Intrinsic, Self-Relevant Absorption

In contrast to the previously described approaches, Fredricks et al. (2010) used qualitative interview data from a long-term longitudinal study to identify passionate individuals. The authors investigated passion among (1) adolescents who were perceived by themselves, their teachers, and parents as 'talented' for extracurricular activities such as, e.g., dance, instrumental music, sports, or drama and (2) adolescents who were identified as 'gifted' by their elementary schools and belonged to the 25% highest performing students among the gifted students. These students had been high performing and intensely engaged in these respective activities for years and were interviewed about a variety of themes including their experiences and aspirations in the activities they were intensely engaged in. The authors defined adolescents to be passionate if they (1) reported that they wanted to do the activity all the time, (2) got intense involvement in the activity and experienced emotional release through the activity, (3) invested much time and energy into doing the activity, and (4) identified with the activity. Fredricks and colleagues found that passion was a rare, exceptional experience, eventually reported by some individuals with intense engagement in extracurricular activities, but virtually never by individuals high performing and intensely engaged in academic activities.

Fredricks and colleagues' definition of passion resembles the above-described definitions of the dual model and the com.pass model in its description of a person–activity relationship and its inclusion of desire, commitment, and self-relevance as characteristics of passion. The main differences to other passion definitions are the applied qualitative measurement instrument, the finding that passion in terms of the authors' criteria was a rare, exceptional phenomenon, and the authors' definition of involvement and absorption as essential components rather than just correlates of passion.

Domain-Specific Models of Passion

Passion as Part of the Triangular Theory of Love

A domain-specific definition of passion was suggested by Sternberg (1986), who described passion as one of three love components in his triangular theory of love. He defined passion as a short-term sexual attraction of one person to another, including intense longing for the other person, frequent thoughts about the other person, and arousal (Sternberg, 1997). This definition differs from the above described in the understanding of passion as a quickly fading state, distinct from commitment, and in its reference to persons (or objects, see Albert et al., 2008) instead of activities as objects of passion. A corresponding measure is the passion subscale of the triangular love scale (Sternberg, 1997, p. 329). Sternberg's passion concept influenced the research about love in romantic relationships (e.g., Azevedo and Aron, 2009) and the research about a consumer's love for brands and products (e.g., Albert et al., 2008), but it has not been applied in the research on passion for activities.

Entrepreneurial Passion

Entrepreneurial passion is another domain-specific concept, defined as "consciously accessible, intense positive feelings experienced by engagement in entrepreneurial activities associated with roles that are meaningful and salient to the self-identity of the entrepreneur" (Cardon et al., 2009b). This definition resulted from a thorough review of commonalities

and differences of different passion concepts in the entrepreneurship literature. The authors distinguish between different entrepreneurial activities and link them to three different role-identities and corresponding types of entrepreneurial passion: (1) the inventor identity, (2) the founder identity, and (3) the developer identity. There are two different instruments for the assessment of entrepreneurial passion: a self-report questionnaire with subscales assessing passion for inventing, founding, and developing (Cardon, 2008; Chen et al., 2009; Cardon et al., 2009a), and an instrument to assess how the entrepreneurial passion of one person is perceived by others, distinguishing affective, cognitive, and behavioral aspects of entrepreneurial passion (Cardon et al., 2009a). The instruments are described as reliable and valid by the authors, but detailed reports about their psychometric properties have not been published to date. Cardon (2008) states that her context-specific measures of entrepreneurial passion are important innovations and cannot be substituted by the dual-model passion scale (Vallerand et al., 2003) applied to entrepreneurial activities, because an empirical comparison of these measures showed that they assess different phenomena.

Relations of Passion to Other Constructs

The relation of passion to other experiences and behaviors is addressed in many studies (e.g., Fredricks et al., 2010; Vallerand et al., 2003). Several motivational constructs overlap with passion in some aspects, which can create confusion and redundancies. In order to remedy terminological ambiguities, we therefore map similarities and differences between passion and related constructs in the following, and then summarize the existing research about the predictors, correlates, and outcomes of passion below.

Theoretical Mappings of Overlapping Constructs

Fredricks et al. (2010) provided the first systematic comparison between passion and similar, overlapping constructs. According to the authors, passion includes the experience flow, which is an exceptional and rewarding state of intense absorption, concentration, and loss of self- and time-monitoring (Csikszentmihalyi, 1996). Furthermore, Fredricks et al. (2010) described similarities between the passion-component identification with an activity, and the concept of attainment value, i.e., the subjective importance to perform well on a given task because of its relation to the individuals' self-concept (see Wigfield and Eccles, 1992).

Another overlapping construct is the 'well-developed personal interest' (Hidi and Renninger, 2006). Passion and personal interest describe pleasant experiences, whose content is well known and valued by the individual, and in which the individual desires to persist. Another overlapping is Winners' (1996, 2000) construct 'rage to master.' This construct and Fredrick et al.'s notion of passion both describe intrinsic and obsessive motivation to acquire skills in a given domain and include experiences of obsessive interest, loss of interest for the 'outside world,' and the sharp focus on the area of expertise. One of the most similar concepts to passion is commitment, which shares all of its components with passion (see above). Commitment is a part of passion, according to the com.pass model, and the other above-mentioned constructs are partially similar but distinct. In sum, passion is a multifaceted construct that overlaps in some of its components with other constructs. The discriminant and incremental validity of passion over these overlapping constructs is understudied at the moment this article was written.

Predictors of Passion

Several empirical studies have investigated the predictors of passion, most of them focusing on differential predictors for HP and OP in terms of Vallerand et al. (2003). According to the dual model of passion, HP is supposedly predicted by autonomous identification with an activity, whereas controlled identification with an activity is supposed to predict OP. This was supported by Mageau et al.'s (2009) findings that parental autonomy supports positively predicted HP, whereas parental activity valuation predicted OP. Similarly, Bonneville-Roussy et al. (2013) found that musical students' perception of getting autonomy support by their teacher predicted HP, whereas the perception of being controlled by the teacher predicted OP, while identification with playing music predicted both types of passion. Similar findings were reported by Vallerand et al. (2008a), who found that an autonomous personality orientation predicted HP, a controlled personality orientation predicted OP, and activity valuation predicted both. Concerning the predictors of the general degree of passion, Mageau et al. (2009) found that children's identification with the activity, their preference to specialize in that activity, and the autonomy support provided to them by their parents predicted the development of general passion for music 5 months later.

Several studies have tried to disentangle person-specific and activity- or situation-specific preconditions of passion (e.g., Amiot et al., 2006; Balon et al., 2013; Rip et al., 2012). Most authors conceptualize passion as a relationship of a person toward an activity and assume that the degree of passionate individual experiences is determined by both personal- and situational-specific characteristics and their interactions. A simultaneous study of both person-specific stable and situation-specific fluid components of passion was reported by Moeller et al. (under review), who found in a longitudinal Experience Sampling Approach that situational determinants accounted for 80% of variance in passion, whereas about 20% of the variance was due to person-specific determinants. Particularly, specific facets of extraversion (dispositional approach motivation and positive trait affect) are related to several passion measures (Moeller et al., 2014). Moreover, several studies on links between HP and OP and the big five personality traits found small and inconsistent correlations (Balon et al., 2013; Tosun and Lajunen, 2009; Wang and Yang, 2008).

Regarding activity-specific preconditions, Fredricks et al. (2010) found that extracurricular activities are far more often experienced in a passionate way than academic activities. Concerning the interaction of personal- and activity-specific characteristics, Amiot et al. (2006) found that an interaction between the type of passion and the competitiveness of the team sport predicted the psychological adjustment of elite hockey players. Obsessively passionate individuals were better

adjusted in highly competitive teams, whereas in less competitive teams the harmoniously passionate individuals were better adjusted.

Correlates of HP and OP

Many studies investigated correlations of HP and OP with other constructs. HP is positively correlated with intrinsic experiences and adaptive behavior, such as, e.g., flow (Philippe et al. 2009a; Vallerand et al., 2003; Carpentier et al., 2012), psychological and physical well-being (Vallerand, 2012; Vallerand et al., 2006), deliberate practice (Bonneville-Roussy et al., 2011), positive affect during and after activity engagement (Philippe et al., 2010), and maintained control over the activity (Vallerand et al., 2003; Mageau et al., 2005). In contrast, OP is correlated with maladaptive and undesirable experiences such as loss of control and symptoms of dependency (Ratelle et al., 2004; Philippe and Vallerand, 2007), negative affect during activity engagement and when the individual is prevented from doing the activity (Philippe et al., 2010; Vallerand et al., 2003), sport injuries (Rip et al., 2006), ruminative thoughts about the activity (Carpentier et al., 2012), aggression under identity threat (Rip et al., 2012), impaired concentration (Ratelle et al., 2004), problematic persistence (Vallerand et al., 2003), and relationship conflicts (Séguin-Lévesque et al., 2003).

Outcomes

There are huge expectations regarding the outcomes of passion. Passion is expected to make "life worth living" (Marsh et al., 2013; Vallerand et al., 2003, p. 757; Vallerand, 2008) and to increase well-being, positive affect, and motivation in general (Fredricks et al., 2010). HP is expected to cause desirable outcomes such as high achievement (Bonneville-Roussy et al., 2011), whereas OP is supposed to cause problematic behavior and suffering. In a short-term longitudinal study, Bonneville-Roussy et al. (2011) found that HP predicted mastery goals, deliberate practice, and performance in musical instrument play, whereas OP predicted performance approach and avoidance goals and predicted lower performance (for similar findings see Vallerand et al., 2007, 2008a). In contrast, Bélanger et al. (2013) found in experimental studies that OP predicted greater performance in the condition in which the individuals received the information that they had failed in a previous task, whereas HP was unrelated to the information about success or failure.

Directions for Future Research

Many questions remain unanswered to be investigated in future research. In their literature review about entrepreneurial passion, Cardon et al. (2009b) identified the following important desiderata and research gaps that are equally relevant for the general research about passion: (1) how (entrepreneurial) passion can be defined clearly and theoretically based, (2) whether (entrepreneurial) passion could be understood as a personality trait or rather an activity- or situation-specific experience, and (3) which conceptual model could explain the influence of (entrepreneurial) passion on relevant outcomes. Additional specific questions are discussed below together with directions for future research.

Problems Regarding the Conceptualization of Passion

Need for Consensus about Definitions, Components, and Measures of Passion

The disparate definitions and measures of passion still contribute more confusion than insights to the psychology of motivation. First, reviews have tried to remedy the existing ambiguities by integrating similar definitions (e.g., Cardon et al., 2009b). However, most publications refer to only one definition and measure of passion but interpret their findings as valid insights about passion in general. In order to ensure that findings are interpretable and generalizable beyond the applied measure, we need more knowledge about the (in-)variance between different definitions and measurement instruments of passion. Open questions are, for instance: Which operationalization of general passion is more useful: a dichotomous or continuous variable? Which would be a useful cutoff score for dichotomous passion definitions? The currently applied cutoff at the scale midpoint (e.g., Bonneville-Roussy et al., 2011; Philippe et al., 2009b; Rip et al., 2012) seems arbitrary, so future research should determine better criteria for the distinction between passionate and nonpassionate individuals.

Person-Oriented Analyses: General Degree versus Specific Types of Passion

The psychological literature about passion has often stated assumptions about differences between harmoniously passionate and obsessively passionate individuals (e.g., Vallerand et al., 2003). In line with this assumption, numerous variable-centered studies have shown differences between the scales for HP and OP. However, the frequent assumption that individuals are either harmoniously or obsessively passionate has been challenged by findings of several cluster analytical studies, which did not find any group of predominantly harmoniously or obsessively passionate individuals, but instead groups of individuals with mainly high, or low, scores in all subscales of passion (Moeller et al., 2014; Wang and Yang, 2008). These findings together with the often moderate to high correlation between HP and OP (see above) raise the question whether HP and OP are different types in terms of homogeneous groups of individuals or rather aspects of passion that are measured by distinct scales but which can co-occur within individuals. A useful framework for corresponding future studies would be the distinction between variable- and person-oriented approaches (Bergman and Magnusson, 1997). If individuals shall be described properly, then it is probably more insightful to analyze the existence of homogeneous subgroups of individuals and the person-specific development (person-oriented approach), instead of the previous praxis to rely on analyses of (cor-) relations between variables across all individuals (variable-oriented approach).

The above-mentioned cluster analyses indicate that the general degree of passion accounts for group differences. Therefore, future studies should shed more light on the so far neglected difference between high and low passion.

Particular Functioning and Interactions of Specific Components

The explicit definitions of passion conceptualize passion as a multifaceted construct, but most studies do not address the particular functioning and interaction of these facets. Consequently, the contribution of the specific components to the explanation of the relevant outcomes of passion is unclear and should be studied more extensively. One of the strengths of the construct passion is its integration of short-term motivational intensity with aspects of long-term commitment, which is an important research desideratum (Katz, 1999; Duckworth et al., 2007). This potential to bridge previously separated research lines within a joined framework is lost if the mechanisms through which the different components influence the outcomes are not explained in testable models and empirical studies. The distinction between specific components of passion promises new insights, given that studies assessing specific facets (e.g., of personality) are generally more informative and better interpretable than those relying on higher order factors (Armstrong and Anthoney, 2009; Beauducel et al., 2007; Bipp et al., 2008; Schimmack et al., 2004).

The distinction between specific components could also remedy redundancies in relation to other constructs, because it is easier to identify overlaps and differences between two constructs if their specific components are comparably laid out.

The measurement model underlying integration of the specific components of passion to one composite score should be theoretically sound and empirically insightful. Current measures average and weight all passion components equally. For example, the passion-criteria score of the dual model of passion is computed as the mean score of the items that assess liking, finding important, spending time and energy, and declaring an activity to one's passion (Vallerand et al., 2003). However, the extent to which the different components contribute to the explanation of relevant criteria might be different, and this should be reflected in the applied measurement models. More research is needed to determine which measurement model best represents the specific functioning and interaction of the components and their specific contribution to the prediction of relevant outcomes.

State versus Trait Concepts of Passion: Related Problems

Need for Integrative Use of State and Trait Measures; Pleading for a Larger Variety of Research Methods and More Objective Measures

Several studies have addressed trait-aspects or situation-dependent aspects of passion separate of each other, but both approaches are rarely integrated in a joined study or model. In order to disentangle person- and situation-specific determinants of passion, future studies should apply trait and state measures together.

Most studies applied self-report assessments and situation-unspecific measures. Such measures are known to be prone to biases, because the individuals commit errors when they retrospectively remember, evaluate, and mentally aggregate their physical arousal across several situations (e.g., Goetz et al., 2013); moreover, some experiences like physical arousal are not necessarily consciously aware and therefore not validly reportable. Future studies should apply more objective and in-the-moment assessments of passion. Particularly, the strongly situation-dependent components like the affect intensity (emphasized, e.g., in the definition of Cardon et al., 2009b) could be more validly assessed with situation-specific measures such as the Experience Sampling Method or with psychophysiological assessment of arousal such as electrodermal resistance or pupil dilatation.

The validity of the trait measures could be enhanced by more objective measures such as reports by teachers, or peers, which would also help to find out how individual passion is perceived by significant others (Cardon et al., 2009a). A limitation of the previous research of person-specific determinants of passion is that most studies missed all those individuals who dropped out at earlier time points of the learning process before the highly engaged samples were assessed (e.g., Bonneville-Roussy et al., 2011). More representative samples and dropout analyses could help to understand why some individuals persist passionately, while others drop out of the respective activities.

Problems Related to the Aim to Explain Relevant Criteria

Incremental Validity

The previous research left unanswered whether passion contributes incrementally to the explanation of relevant phenomena. Incremental validity is proven when passion contributes additional explanation of the relevant criterion after *all* similar constructs are taken into account. Eventual findings of incremental validity of passion over one or two related constructs would be only one piece of the puzzle. If the incremental validity of passion could be shown, it would be interesting to know which specific components account for it and whether there is also incremental validity in the *prospective* prediction of relevant criteria by passion in comparison to other constructs.

Need for Process Models Explaining the Stability and Long-Term Development of Passion, Its Predictors, and Outcomes

Most authors conceptualize passion as a long-term stable experience (e.g., Vallerand et al., 2003; but see Sternberg, 1986, 1997). However, the stability of passion is rarely investigated. There also is a lack of theoretical models to explain how passion and its long-term predictors and outcomes develop. Furthermore, there is a lack of appropriate longitudinal studies on the development, predictors, and outcomes of passion. The existing studies are mostly cross-sectional. The few longitudinal studies do not sufficiently describe the long-term development of passion because most of them only investigated periods of few months (e.g., Amiot et al., 2006; Bonneville-Roussy et al., 2013; Mageau et al., 2009; Philippe et al., 2010; Vallerand et al., 2006, 2007, 2008b) and often did not assess whether passion *changed* in dependence of certain predictors, thus not allowing one to test the assumed causal effects of passion on its hypothesized outcomes.

Prospective longitudinal studies assessing the change in predictors, outcomes, and in passion itself are required in order to test assumptions about predictors of passion. For instance, a central assumption of the dual model of passion states that HP results from the autonomous internalization of an activity and OP from the controlled internalization (Vallerand et al.,

2003). However, very few studies assessed the *preceding* determinants of passion *before* the passion developed (e.g., Mageau et al., 2009; Vallerand et al., 2006), so it is not known conclusively to what extent the type of passion is determined by preceding forms of internalization. Further investigation is also needed in regard to the unknown long-term development of HP and OP.

Importantly, there is a need for testable models explaining the long-term development of passion, its predictors, and outcomes. The current research is inconsistent: In different studies the same construct (e.g., flow, meaning intense absorption and concentration on current tasks) figures as predictor of passion (Fredricks et al., 2010), correlate of passion (Vallerand et al., 2003), or essential component of passion (Fredricks et al., 2010). Possibly, iterative process models such as dynamic systems approaches (e.g., Lichtwarck-Aschoff et al., 2009) could explain how a construct such as flow can be both antecedent and result of an individual's passionate inclination toward an activity, and how such experiences can reinforce each other in iterative cycles. However, currently, there is no such theoretical model. The specification and empirical testing of process models is a task for future research to complete.

In sum, passion is a relatively new construct that has inspired a rapidly growing number of studies since the publication of the dual model of passion in the year 2013. Several different definitions were proposed. Most authors conceptualize passion as a multifaceted construct that describes the inclination of a person to engage in particular activities. It is often assumed that passion remains stable across long time periods and that passion explains long-term persistence in the corresponding activities. Trait and state components of passion were discussed here. Variable-oriented approaches have shown important differences between harmonious and obsessive subscales of passion, whereas cluster-analytic approaches showed that within individuals, HP and OP are mostly aligned, meaning either similarly high or low.

See also: Affect-Regulation Motivation; Avoidance and Approach Motivation: A Brief History; Burnout, Psychology of; Control Behavior: Psychological Perspectives; Flow in Motivational Psychology; Gender and Academic Motivation; Interest, Psychology of; Motivation and Actions, Psychology of; Optimism and Pessimism; Personal Projects; School Burnout and Engagement: Lessons from a Longitudinal Study in Finland; Self and Emotional Development in Adulthood and Later Life; Self-Regulation in Adulthood; Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model.

Bibliography

Achter, J.A., Lubinski, D., 2005. Blending promise with passion: best practices for counseling intellectually talented youth. In: Brown, S.D., Lent, R.W. (Eds.), Career Development and Counseling: Putting Theory and Research to Work. John Wiley & Sons, Hoboken, NJ, pp. 600–624.

Albert, N., Merunka, D., Valette-Florence, P., 2008. When consumers love their brands: exploring the concept and its dimensions. Journal of Business Research 61, 1062–1075. http://dx.doi.org/10.1016/j.jbusres.2007.09.014.

Amiot, C., Vallerand, R.J., Blanchard, C., 2006. Passion and psychological adjustment: a test of the person-environment fit hypothesis. Personality and Social Psychology Bulletin 32, 220–229. http://dx.doi.org/10.1177/0146167205280250.

Armstrong, P.I., Anthoney, S.F., 2009. Personality facets and RIASEC interests: an integrated model. Journal of Vocational Behavior 75 (3), 346–359. http://dx.doi.org/10.1016/j.jvb.2009.05.004.

Azevedo, B.P., Aron, A., 2009. Does a long-term relationship kill romantic love? Review of General Psychology 13 (1), 59–65. http://dx.doi.org/10.1037/a0014226.

Balon, S., Lecoq, J., Rimé, B., 2013. Passion and personality: is passionate behavior a function of personality? European Review of Applied Psychology 63 (1), 59–65. http://dx.doi.org/10.1016/j.erap.2012.06.001.

Beauducel, A., Liepmann, D., Felfe, J., Nettelsnstroth, W., 2007. The impact of different measurement models for fluid and crystallized intelligence on the correlation with personality traits. European Journal of Psychological Assessment 23, 71–78. http://dx.doi.org/10.1027/1015-5759.23.2.71.

Bélanger, J., Lafrenière, M.-A., Vallerand, R.J., Kruglanski, A.W., 2013. Driven by fear: the role of failure in passionate individuals' performance. Journal of Personality and Social Psychology 104, 180–195. http://dx.doi.org/10.1037/a0029585.

Bergman, L.R., Magnusson, D., 1997. A person-oriented approach in research on developmental psychopathology. Development and Psychopathology 9, 291–319.

Bipp, T., Steinmayr, R., Spinath, B., 2008. Personality and achievement motivation: relationship among big five domains and facet scales, achievement goals, and intelligence. Personality and Individual Differences 44, 1454–1464. http://dx.doi.org/10.1016/j.paid.2008.01.001.

Bonneville-Roussy, A., Lavigne, G.L., Vallerand, R.J., 2011. When passion leads to excellence: the case of musicians. Psychology of Music 39 (1), 123–138. http://dx.doi.org/10.1177/0305735609352441.

Bonneville-Roussy, A., Vallerand, R., Bouffard, 2013. The role of autonomy support and harmonious and obsessive passions in educational persistence. Learning and Individual Differences 11, 22–31. http://dx.doi.org/10.1016/j.lindif.2012.12.015.

Cardon, M., 2008. The discriminant validity of entrepreneurial passion (Summary). Frontiers of Entrepreneurship Research 28 (4). Retrieved from: http://digitalknowledge.babson.edu/fer/vol28/iss4/3.

Cardon, M.S., Sudek, R., Mitteness, C., 2009a. The impact of perceived entrepreneurial passion on angel investing. Frontiers of Entrepreneurship Research 29 (2), 1–15.

Cardon, M.S., Wincent, J., Singh, J., Drnovsek, M., 2009b. The nature and experience of entrepreneurial passion. Academy of Management Review 34 (3), 511–532. http://dx.doi.org/10.5465/AMR.2009.40633190.

Carpentier, J., Mageau, G., Vallerand, R.J., 2012. Ruminations and flow: why do people with a more harmonious passion experience higher well-being? Journal of Happiness Studies 13, 501–518. http://dx.doi.org/10.1007/s10902-011-9276-4.

Chen, X.-P., Yao, X., Kotha, S., 2009. Entrepreneur passion and preparedness in business plan presentations: a persuasion analysis of venture capitalist's funding decisions. Academy of Management Journal 52 (1), 199–214. http://dx.doi.org/10.5465/AMJ.2009.36462018.

Csikszentmihalyi, M., 1996. Creativity and the Psychology of Discovery and Invention. HarperCollins, New York.

Dixon, T., 2003. From Passions to Emotions. The Creation of a Secular Psychological Category. Cambridge University Press, Cambridge, UK.

Duckworth, A.L., Peterson, C., Matthews, M.D., Kelly, D.R., 2007. Grit: perseverance and passion for long-term goals. Journal of Personality and Social Psychology 92 (6), 1087–1101. http://dx.doi.org/10.1037/0022-3514.92.6.1087.

Ericsson, K., Krampe, R.T., Tesch-Römer, C., 1993. The role of deliberate practice in the acquisition of expert performance. Psychological Review 100, 363–406. http://dx.doi.org/10.1037/0033-295X.100.3.363.

Fredricks, J.A., Alfeld, C., Eccles, J.S., 2010. Developing and fostering passions in academic and nonacademic domains. Gifted Child Quarterly 54 (1), 18–30. http://dx.doi.org/10.1177/0016986209352683.

Fredricks, J.A., Blumenfeld, P.C., Paris, A.H., 2004. School engagement: potential of the concept, state of the evidence. Review of Educational Research 74 (1), 59–109.

Gagné, F., 2007. Ten commandments for academic talent development. Gifted Child Quarterly 51 (2), 93–118. http://dx.doi.org/10.1177/0016986206296660.

Goetz, T., Bieg, M., Lüdtke, O., Pekrun, R., Hall, N.C., 2013. Do girls really experience more anxiety in mathematics? Psychological Science 24, 2079–2087. http://dx.doi.org/10.1177/0956797613486989.

Hidi, S., Renninger, K.A., 2006. The four-phase model of interest development. Educational Psychologist 41, 111–127. http://dx.doi.org/10.1207/s15326985ep4102_4.

Katz, L.D., 1999. Dopamine and serotonin: integrating current affective engagement with long-term goals. Behavioral and Brain Sciences 22 (3), 539–540. http://dx.doi.org/10.1017/S0140525X99312048.

Lichtwarck-Aschoff, A., Kunnen, S.E., van Geert, P.L.C., 2009. Here we go again: a dynamic systems perspective on emotional rigidity across parent–adolescent conflicts. Developmental Psychology 45 (5), 1364–1375. http://dx.doi.org/10.1037/a0016713.

Mageau, G.A., Vallerand, R.J., Charest, J., Salvy, S.-J., Lacaille, N., Bouffard, T., et al., 2009. On the development of harmonious and obsessive passion: the role of autonomy support, activity specialization, and identification with the activity. Journal of Personality 77 (3), 601–646. http://dx.doi.org/10.1111/j.1467-6494.2009.00559.x.

Mageau, G., Vallerand, R., Rousseau, F., Ratelle, C., Provencher, P., 2005. Passion and gambling: investigating the divergent affective and cognitive consequences of gambling. Journal of Applied Social Psychology 35, 100–118. http://dx.doi.org/10.1111/j.1559-1816.2005.tb02095.x.

Marsh, H.W., Vallerand, R.J., Lafreniere, M.A.K., Parker, P., Morin, A.J.S., Carbonneau, N., et al., 2013. Passion: does one scale fit all? Construct validity of two-factor passion scale and psychometric invariance over different activities and languages. Psychological Assessment 25 (3), 796–809. http://dx.doi.org/10.1037/a0032573.

Mills, C.J., 2003. Characteristics of effective teachers of gifted students: teacher background and personality styles of students. Gifted Child Quarterly 47 (4), 272–281. http://dx.doi.org/10.1177/001698620304700404.

Moeller, J., Dietrich, J., Eccles, J.S., Schneider, B. On the variability and long-term stability of passion: a longitudinal experience sampling approach. Manuscript submitted for publication, under review.

Moeller, J., Grassinger, R., 2014a. Opening a can of worms: a review of passion concepts and their overlaps with commitment. In: Moeller, J. (Ed.), Passion as Concept of the Psychology of Motivation. Conceptualization, Assessment, Inter-Individual Variability and Long-Term Stability, pp. 13–52. Dissertation published online at:http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-29036/DissJuliaMoeller.pdf.

Moeller, J., Grassinger, R., 2014b. Measuring passion and commitment with one joint scale: psychometric properties and validity of the com.pass scale. In: Moeller, J. (Ed.), Passion as Concept of the Psychology of Motivation. Conceptualization, Assessment, Inter-Individual Variability and Long-Term Stability, pp. 53–111. Dissertation published online at: http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-29036/DissJuliaMoeller.pdf.

Moeller, J., Keiner, M., Wächter, D., 2014. A disposition for passion? Linking passionate approach motivation to facets of extraversion and non-substance related dependency. In: J. Moeller (Ed.), Passion as Concept of the Psychology of Motivation. Conceptualization, Assessment, Inter-Individual Variability and Long-Term Stability, pp. 159–178. Dissertation published online at: http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-29036/DissJuliaMoeller.pdf.

Moraes, L.C., Rabelo, A.S., Salmela, J.H., 2004. Papel dos Pais no Desenvolvimento de Jovens Futebolistas (The role of parents in the development of young soccer players). Psicologia: Reflexão e Crítica 17 (2), 211–222. http://dx.doi.org/10.1590/S0102-79722004000200009.

Noble, K.D., Childers, S.A., Vaugham, R.C., 2008. A place to be celebrated and understood: the impact of early university entrance from parents' points of view. Gifted Child Quarterly 52 (3), 256–268. http://dx.doi.org/10.1177/0016986208319976.

Philippe, F., Vallerand, R.J., 2007. Prevalence rates of gambling problems in montreal, canada: a look at old adults and the role of passion. Journal of Gambling Studies 23, 275–283. http://dx.doi.org/10.1007/s10899-006-9038-0.

Philippe, F.L., Vallerand, R.J., Andrianarisoa, J., Brunel, P., 2009a. Passion in referees: examining their affective and cognitive experiences in sport situations. Journal of Sport & Exercise Psychology 31 (1), 77–96.

Philippe, F.L., Vallerand, R.J., Houlfort, N., Lavigne, G.L., Donahue, E., 2010. Passion for an activity and quality of interpersonal relationships: the mediating role of emotions. Journal of Personality and Social Psychology 98 (6), 917–932. http://dx.doi.org/10.1037/a0018017.

Philippe, F., Vallerand, R.J., Lavigne, G., 2009b. Passion does make a difference in people's lives: a look at well-being in passionate and non-passionate individuals. Applied Psychology: Health and Well-Being 1, 3–22. http://dx.doi.org/10.1111/j.1758-0854.2008.01003.x.

Pressley, M., El-Dinary, P., Marks, M.B., Brown, R., Stein, S., 1992. Good strategy instruction is motivating and interesting. In: Renninger, K.A., Hidi, S., Krapp, A. (Eds.), The Role of Interest in Learning and Development. Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp. 333–358.

Ratelle, C.F., Vallerand, R.J., Mageau, G.A., Rousseau, F.L., Provencher, P., 2004. When passion leads to problematic outcomes: at look at gambling. Journal of Gambling Studies 20 (2), 105–119. http://dx.doi.org/10.1023/B:JOGS.0000022304.96042.e6.

Renzulli, J.S., Koehler, J.L., Fogarty, E.A., 2006. Operation Houndstooth intervention theory: social capital in today's schools. Gifted Child Today 29 (1), 14–24.

Rip, B., Fortin, S., Vallerand, R., 2006. The relationship between passion and injury in dance students. Journal of Dance Medicine & Science 10 (1,2), 14–20.

Rip, B., Vallerand, R.J., Lafrenière, M.-A.K., 2012. Passion for a cause, passion for a creed: on ideological passion, identity threat, and radicalization. Journal of Personality 80, 573–602. http://dx.doi.org/10.1111/j.1467-6494.2011.00743.x.

Rusbult, C.E., Martz, J.M., Agnew, C.R., 1998. The investment model scale: measuring commitment level, satisfaction level, quality of alternatives, and investment size. Personal Relationships 5, 357–391. http://dx.doi.org/10.1111/j.1475-6811.1998.tb00177.x.

Scanlan, T.K., Russell, D.G., Magyar, T.M., Scanlan, L.A., 2009. Project on elite athlete commitment (PEAK): III. An examination of the external validity across gender, and the expansion and clarification of the sport commitment model. Journal of Sport & Exercise Psychology 31, 685–705.

Schimmack, U., Oishi, S., Furr, R.M., Funder, D.C., 2004. Personality and life satisfaction: a facet-level analysis. Personality and Social Psychology Bulletin 30 (8), 1062–1075. http://dx.doi.org/10.1177/0146167204264292.

Séguin-Lévesque, C., Laliberté, M.L., Pelletier, L.G., Vallerand, R.J., Blanchard, C., 2003. Harmonious and obsessive passions for the Internet: their associations with couples' relationships. Journal of Applied Social Psychology 33, 197–221. http://dx.doi.org/10.1111/j.1559-1816.2003.tb02079.x.

Sternberg, R.J., 1986. A triangular theory of love. Psychological Review 93 (2), 119–135. http://dx.doi.org/10.1037/0033-295X.93.2.119.

Sternberg, R.J., 1997. Construct validation of a triangular love scale. European Journal of Social Psychology 27, 313–335. http://dx.doi.org/10.1002/(SICI)1099-0992(199705)27:3<313::AID-EJSP824>3.0.CO;2–4.

Tosun, L.P., Lajunen, T., 2009. Why do young adults develop a passion for Internet activities? The associations among personality, revealing ''true self'' on the Internet, and passion for the Internet. Cyberpsychology and Behavior 12 (4), 401–406. http://dx.doi.org/10.1089/cpb.2009.0006.

Vallerand, R.J., 2008. On the psychology of passion: in search of what makes people's lives most worth. Canadian Psychology 49 (1), 1–13. http://dx.doi.org/10.1037/0708-5591.49.1.1.

Vallerand, R.J., 2012. The role of passion in sustainable psychological well-being. Psychological Well-Being: Theory, Research, and Practice 2, 1–21.

Vallerand, R.J., Blanchard, C.M., Mageau, G.A., Koestner, R., Ratelle, C., Léonard, M., et al., 2003. Les passions de l'âme: on obsessive and harmonious passion. Journal of Personality and Social Psychology 85 (4), 756–767. http://dx.doi.org/10.1037/0022-3514.85.4.756.

Vallerand, R.J., Mageau, G.A., Elliot, A.J., Dumais, A., Demers, M.-A., Rousseau, F., 2008a. Passion and performance attainment in sport. Psychology of Sport and Exercise 9 (3), 373–392. http://dx.doi.org/10.1016/j.psychsport.2007.05.003.

Vallerand, R.J., Ntoumanis, N., Philippe, F., Lavigne, G.L., Carbonneau, N., Bonneville, A., et al., 2008b. On passion and sports fans: a look at football. Journal of Sport Sciences 26, 1279–1293. http://dx.doi.org/10.1080/02640410802123185.

Vallerand, R.J., Salvy, S.-J., Mageau, G.A., Elliot, A.J., Denis, P.L., Grouzet, F.M., et al., 2007. On the role of passion in performance. Journal of Personality 75 (3), 505–534. http://dx.doi.org/10.1111/j.1467-6494.2007.00447.x.

Vallerand, R.J., Rousseau, F.L., Grouzet, F.M., Dumais, A., Grenier, S., Blanchard, C.M., 2006. Passion in sport: a look at determinants and affective experiences. Journal of Sport and Exercise Psychology 28, 454–478. http://dx.doi.org/10.1016/j.psychsport.2007.05.003.

Wai, J., Lubinski, D., Benbow, C.P., 2009. Aligning potential and passion for promise: a model for educating intellectually talented youth. In: Renzulli, J.S. (Ed.), Systems and Models for Developing Programs for the Gifted and Talented. Creative Learning Press, Inc., Mansfield Center, CT.

Wang, C.C., Yang, H.W., 2008. Passion and dependency in online shopping activities. Cyberpsychology and Behavior 10 (2), 296–298. http://dx.doi.org/10.1089/cpb.2006.9954.

Wigfield, A., Eccles, J.S., 1992. The development of achievement task values: a theoretical analysis. Developmental Review 12, 265–310. http://dx.doi.org/10.1016/0273-2297(92)90011-P.

Willms, J.D., 2003. Student Engagement at School. A Sense of Belonging and Participation. Results from PISA 2000. OECD Publishing, Paris.

Winner, E., 1996. The rage to master: the decisive role of talent in the visual arts. In: Ericsson, K.A. (Ed.), The Role to Excellence: The Acquisition of Expert Performance in the Arts and Sciences, Sports, and Games. Lawrence Erlbaum, Hillsdale, NJ, pp. 271–301.

Winner, E., 2000. The origins and ends of giftedness. American Psychologist 55, 159–169. http://dx.doi.org/10.1037//0003-066X.55.1.159.

Relevant Website

http://www.er.uqam.ca/nobel/r26710/LRCS/rjvaller_en.htm – Research Laboratory on Social Behavior.

Pastoralism in Anthropology

John G Galaty, McGill University, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Pastoralism is a mode of subsistence that involves raising domestic animals in grassland environments using herd and household mobility. Combined with nomadism, pastoralism has allowed humans to inhabit the world's vast dry lands. Pastoral systems are adaptations to diverse political and ecological conditions, but today are influenced by national and global societies. Livestock are objects of pastoralist identification used for subsistence, market sale, social exchange, and symbolic expression and were the world's first currency. Against the tragedy of the commons argument, pastoral use of rangelands is sustainable, and should continue as pastoralists increasingly participate in global institutions and cultural modernity.

What Is Pastoralism?

Pastoralism is a mode of subsistence that involves raising domestic animals on natural pastures, a livelihood practice mainly pursued in dry grassland environments that support few competing forms of land use. Pastoralism requires herd and household mobility as herders continually seek fresh grazing. Indeed, the nomadic element has made possible successful human habitation of vast areas of the world's grasslands and semideserts using locally appropriate breeds: cattle, small stock, and dromedaries in Africa, dromedaries, sheep, and goats in the Middle East, camels, cattle, and sheep in Central Asia, yaks in mountainous areas of the Tibetan Plateau, llama and alpaca in the Andean highlands, and reindeer in the circumpolar Tundra. But the frequency of pastoral mobility increases as a function of the low level, unpredictability, and dispersed nature of rainfall received in various grassland regions. Pastoralism usually relies on family or community-based labor for animal care, achieving economies of scale when relatively few pastoralists herd large numbers of domestic stock across wide expanses of pasture, though sometimes herds are contracted out. It thus involves using labor intensively but land extensively, in contrast to intensive land use in mixed peasant farming or capital intensive ranching (Galaty and Johnson, 1990). As a livelihood strategy, pastoralism as an enterprise focuses mainly on milk production for family use, but also relies on domestic animals for meat and traction, to the end not of maximizing livestock numbers but of optimizing the human population supported in dry lands (Dyson-Hudson and Dyson-Hudson, 1980).

The theoretical concerns of the anthropology of pastoralism have changed over time, from querying how pastoralism fitted into evolutionary sequences, to asking what functions livestock play within an integrated cultural pattern, to exploring how the social institutions of pastoral societies relate to herding livelihoods, to investigating what type of society is characterized by nomadism and pastoral specialization, and how these societies work (Ingold, 1984). Today, it is recognized that all pastoral groups reflect the influences of national and global societies, while representing adaptations to particular political and ecological conditions. Key questions now concern how those with strong engagement with animal husbandry respond dynamically to climate change and environmental stresses, the spread of national and global institutions, challenges to their security of land holding, expansion of market relations, the growing influence of governments, their use of international law as indigenous peoples, and more diffuse processes of social change and modernization (Hodgson, 2011). A new anthropology of the contemporary does not render earlier questions irrelevant, but contextualize them within national and global trends and encroachments on pastoral autonomy.

The Many Functions of Livestock

Although many livestock-keeping peoples engage in numerous other productive activities, such as hunting, fishing, horticulture, arboriculture, and smuggling, still pastoralism looms large when practiced by entire societies. Such societies are often organized around sociopolitical institutions, beliefs, and practices that reflect pastoral preoccupations: political organizations given structure by territorial sections or localized clans, segmentary lineages or age-set systems, bodies of indigenous knowledge about livestock and range ecology, and symbolic practices that reflect the centrality of domestic animals and their husbandry. Livestock are both the 'products' that pastoralists produce (as live animals, milk, meat, skins, etc.) and the 'means' of production, since animals both beget products and more animals, used for subsistence and sale. Livestock also are used as stores of value, objects of investment, and units of exchange (i.e., 'money') in terms of which the value of other goods can be denominated (Schneider, 1979). Indeed, a 'head' (*caput*) of domestic animals represented the original form of financial 'capital' among the Romans. Exchange of domestic animals also serves as means of *social* reproduction, of marital ties via bride-wealth, networks of friendship through gifts, and solidarity among social groups through sacrifice. These social and economic functions underpin the intense personal and affective attachments herders feel toward domestic animals, as well the elaborated systems of classification and poetic and religious expression, and complex bodies of ethnozoological and ethnoecological knowledge that are

part of pastoralists' cultural endowment. Since pastoralists raise each animal from birth, their familiarity with individual stock is intensely personal, and on the basis of this intimacy people identify with and define themselves through their domestic animals.

Theories of Pastoralism

Morgan and after him Marx proposed, biblically, that pastoralism evolved from hunting and foraging economies, with stock representing the first forms of property. But the archaeology of southwestern Asia suggested that animal and plant domestication arose together, with protodomesticate animals attracted to grains found near human habitations. Pastoralism was a specialized form of husbandry that arose when herders and domestic stock moved into drier lands distant from cultivated fields. However, more recent studies show that the first cattle were likely domesticated around 9000–8000 BCE by wild grass-gathering and pottery-making Northern Sudanians profiting from the southern spread of *Bos taurus* into the Sahara during a wetter period, about a 1000 years before cattle were domesticated in the eastern Mediterranean and the Middle East and later in India (Ehret, 2002). So animal domestication may in fact have evolved out of hunting and grass gathering in northeastern Africa about a 1000 years before plants were domesticated in the same region.

The modern anthropological study of pastoralism was initiated with Herskovits' (1926) study of the 'cattle complex' in East Africa that described the cultural elaboration of livestock-centered ideas and practices but misguidedly suggested that domestic stock were objects of just religious rather than also economic value. Evans-Pritchard's (1940) pathbreaking studies of the Nuer of the Sudan emphasized the centrality of animal husbandry to Nuer subsistence, but also to understanding Nuer ecology and the territorial politics based on the segmentary lineage system, kinship, and religion. Scholarly effort was spent trying to typologize livestock keepers, as 'pure' or 'semi'-pastoralists, nomads and seminomads, belying the fact that diverse livestock keepers adopted different degrees of pastoral specialization and mobility according to changing conditions, and both produced and consumed agricultural produce as needed and as available. Detailed ethnographies of pastoral societies of Africa and the Middle East conceptualized and modeled pastoral systems from ecological, economic, and political perspectives.

The ecological approach asked to what extent pastoralism and forms of social organization represented adaptations to the distinct environments in which pastoralism is found. Certainly dry land environments influence the territorial systems that regulate herd mobility, but many key social and cultural institutions in pastoral societies reflect regional cultural patterns that cut across both sedentary and nomadic communities, and thus demand historical rather than ecological explanations. Careful analysis of dynamics of herd growth, the demographics of human fertility and mortality under constraints of pastoral nutrition and labor and the productive potential of various grassland regimes produced models of ecological balance between herd and human populations and pasture dynamics, the latter providing protein and caloric resources for humans through the mediation of ruminants (Dahl and Hjort, 1976; Homewood, 2008).

The economic perspective, which drew from formalist (classical), substantivist, and Marxian approaches, has examined the strategic decisions pastoralists make in the context of subsistence husbandry and market production of livestock and has theorized the linkage between the material processes of mobile husbandry and social and cultural systems. From the Marxian perspective, it was asked whether there was a distinctive 'pastoral mode of production' (most say 'no'), while from a classical economic view whether herders sought to 'maximize' herd sizes or to 'optimize' them in relation to herders' actual objectives of supporting human communities, human needs being traded off against herd growth through the division of milk between calves and children. Rather than utility maximizers, some proposed that pastoralists more closely resembles the domestic-based family firms of early capitalism, described by Weber, where hard work by all family members together with low personal consumption, savings, and reinvestment within a strict religious-based ethic, encouraged capital growth and economic expansion by tightly knit domestic units. In France, scholars contributing to the development of Marxian theory in Anthropology insisted that the technical practices of dry land husbandry could not explain pastoral social and ideological systems, since institutions and ideologies themselves serve as historical determinants of the social relations of production that shape how material practices unfold (Lefébure, 1979). Everywhere today, even remote pastoral communities have been incorporated into a broad system of capitalist relations, reflected in market networks and new forms of authority, though 'articulated' with precapitalist or customary social practices, ideologies, and economic orientations.

The political perspective, inspired by the work of Evans-Pritchard and Frederick Barth, proposed that pastoral political systems were not only shaped by the exigencies of animal husbandry but also by power relations that related herders to a hierarchy of authority within their own groups, to pastoral and sedentary neighbors with whom they interacted, competed, and conflicted, and to states, which persistently sought to pacify, capture, or 'encapsulate' nomadic pastoralists within their own structures of authority and control, including the right of taxation. Nomadism came to be seen not simply as ecological but also political adaptations, aimed at avoiding and evading state control, since centralized authorities could exert less control over peoples who are able to exercise the 'exit' option (Salzman, 2004). While pastoralists were able to exercise bristly autonomy, they interacted judiciously with their neighbors through multilingual trade, exchange and often intermarriage, both defending ethnic boundaries and flexibly sliding across lines of identity. Indeed, while some pastoralist individuals and families exercise ethnic fluidity, some entire groups moved into cultural and ethnic boundary areas where cultural and linguistic practices were exchanged and shared, with the essential identities of some groups evolving over time (Schlee, 1989). The binding together of pastoral polities, such as the Maasai, depended not just on leadership or centralizing forms of organization but shared languages, cultural exchanges, and rituals and emergent forms of identity (Spear and Waller, 1993).

Pastoralism: A Type of Society or a Continuum?

The notion that pastoralism is an adaptive livelihood strategy leads us to ask whether pastoralism is a 'type' of society or economy, or rather a vantage point for understanding how livestock husbandry affects livelihoods, regional relationships, and cultural commitments of communities that are in fact quite diverse. Pastoralists primarily use domestic livestock for subsistence, market sale, and social exchange, while agropastoralists use them for traction, transport, and manure, to support the agricultural economy, secondarily for subsistence. Where the money economy is weak, because currency is unavailable in remote, rural areas or too fungible among networks of kin and neighbors, investments from agriculture into livestock, or vice versa, may be viable household strategies for accumulating capital. Pastoral households usually hold more than one head of cattle per person, often much more, while agropastoralists usually possess less, so the value of livestock among agropastoralists may be higher than among pastoralists, due to proportionate weight of animals in small versus large herds. The social and ritual value of livestock, especially cattle, may be quite high in both systems. But agropastoralists must dedicate relatively large amounts of labor to herding small numbers of livestock along the margins of fields, so fail to enjoy the efficiencies of scale that pastoralists profit from, where a boy or man alone can manage a herd of 50–100 beasts or more grazing across an expanse of grasslands. Should we consider as pastoralists only those who specialize in extensive, mobile animal husbandry, or any who endow livestock with significant social, economic, symbolic, and ritual weight, independent of whether they are also engaged in cultivation, hunting, fishing, market sales, or other pursuits in a complex economy? The 'pastoral continuum' stretches from deserts to well-watered lands, where agriculture is also viable, from highly mobile to sedentary husbandry, from highly specialized to highly diversified systems, with animal husbandry, crop production, hunting, foraging, fishing, and trade combined within a single domestic economy (Spencer, 1998). Some largely produced for subsistence (e.g., the Baluch, Karamojong, Nuer), while others were strongly market-oriented (e.g., the Basseri and Bedouin), and over time change has occurred in what is being produced (e.g., The Bedouin of the Arabian peninsula have switched from camels to sheep, northern Samburu from cattle to dromedaries, Inner Mongolians to goats for cashmere) and by what means (Bedouin adopting trucks!).

What Comes First: Commitments of Herders or Needs of Animals?

Though most pastoral systems value a particular predominant species (witness the 'cattle culture' in East Africa), most pastoralists keep a complex of domestic animal species. Since domestic species thrive in different microenvironments, find different grasses palatable, need water on varying schedules, and can cover different amounts of terrain in a herding day, pastoralists are often faced with decisions about whether to herd species together by compromising on how well their various needs are fulfilled or splitting the herd into smaller management units by species. The Turkana of East Africa, for instance, may keep all domestic species together in the rainy season, but as the dry season progresses keep the camels and goats together in the lowlands, and the cattle and sheep together in the highlands, and in severe drought keep each species separate. As a result, families may cooperate among themselves in providing labor for herding subgroups, or a family may divide its members among several sub-homesteads pursuing specialized herding (McCabe, 2004). Family labor allocation may be made even more complicated if cultivation is considered, since some members may move to prepare fields in the late dry season, while herds circulate between sparse pastures and dwindling water resources. Across a region controlled by the same pastoral group, such as the Somali, drier areas are primarily used by dromedaries, less dry areas for small stock and cattle, resulting in certain subgroups being associated with different species mixes. Do pastoralists with commitments to certain valued livestock seek out the most suitable environmental niches for those species, or do they adopt the species that are most adapted to the environments they come to politically control? Fulani historically have traversed Africa from west to east along the savanna corridor compatible with their cattle and small-stock herding, and camel keepers from Central Asia have migrated throughout the cold arid zone where Bactrian camels thrive, but Oromo speakers migrating into drier regions in the Horn of Africa have adopted dromedaries as the most adaptable animal for the drier ecozones they have come to inhabit (Schlee, 1989). History, culture, and ecology all influence the shape of pastoral systems.

From Studying Pastoralism to Studying Pastoralists

During the last 25 years, the study of pastoralism has extended from the integrated perspective of Anthropology to more specialized studies in the social and natural sciences; at the same time, Anthropology has assimilated questions and approaches from diverse other disciplines, from Economics and Development Studies to Geography and Range and Animal Sciences. Focus has moved from theorizing pastoral systems as such to the analysis of emergent conditions that affect livestock keepers and grassland societies, including ecology and climate change, property rights and land fragmentation, the spread of livestock markets, state-directed programs of development, international conservation, and the spread of global institutions, like schools, clinics, churches, and vehicles. Further advances have been made in the three founding areas mentioned earlier, of pastoral ecology, economics, and politics. Light has also been shed on other problems that currents of social change in pastoral and rangeland communities have stimulated.

Current Perspectives on Pastoralism and the Environment

A significant influence on thinking about how grazing interacts with dry land environments stemmed from an article by Garret Hardin (1968) called 'The Tragedy of the Commons,' which used the desire to put one more head of cattle on a piece of commonly held pastures as a metaphor for how any resource

without designated owners, from which benefits could be realized without concomitant costs, would in principle be overused. Without reference to empirical work on institutions that do regulate use of common pool resources (Ensminger, 1992), the tragedy of the commons argument held that pasture degradation could only be avoided if common holdings were transformed into either private or state property, policy prescriptions for land reform that guided, and still guide, the design of rangeland development programs. With the aim of studying human–livestock–environmental interactions, the South Turkana Ecosystem Project pursued long-term interdisciplinary research with the highly mobile Ngisonyoka Turkana of northwestern Kenya, with collaboration among anthropologists, ecologists, nutritionists, and range scientists. Analyses of pastoralist mobility demonstrated how household labor could be mobilized to carry out multiple strategic moves involved splitting and recombining of households and herds in order to respond to microclimatic conditions and momentary grazing opportunities. Ecological study, led by James Ellis, revealed that the long-term empirical effects of grazing pressure on pasture quality under 'non-equilibrium conditions' were in fact minimal, since the grassland vegetation systems – having evolved over millennia in which rainfall was low, unreliable, and unpredictable – rejuvenated after rainfall (Galvin et al., 2008).

The theoretical and policy implications of these findings are great, since the 'density independence' of many rangeland ecosystems essentially vindicated the major principles of pastoral land use, which involved building up herds in good seasons to optimally exploit standing pastures, while accepting herd die-downs in times of deprivation, and using rapid and opportunistic herd movement both to access areas where pastures were replenished and to relieve the pressure that is exerted on local pastures under conditions of sedentarization or herd fixity (Scoones, 1994). In short, rather than customary pastoral range use practices being seen as the source of long-term rangeland degradation, they were seen as adaptive in the face of intrinsic constraints on maintaining herds and human populations in the dry lands, and to reflect sound understanding of range dynamics, which required herd mobility to 'track' temporally unpredictable and spatially dispersed grazing resources (Homewood, 2008). In the work of Bollig (2006), pastoral institutions were viewed as mechanisms for avoiding or mitigating the risks inherent to dry land husbandry, whether avoiding risks of pasture exhaustion through herd movement, the protection of grazing and livestock through territorial political systems, the spreading of risk through livestock exchange and building social networks, or the forging of age-set organization to ensure rapid response to threats to territories. Rather than aiming to achieve static 'equilibrium' between livestock and pastures, difficult under conditions that constantly varied, the approach to rangeland 'complexity' emphasizes the importance of 'resilience' of land use practices and pastoral husbandry in response to dramatic variations in ecological and political conditions.

Pastoralism within Regional Economies

The evolving understanding of how pastoral economies work was also influenced by the tragedy of the commons argument, in its view that perverse economic motives led pastoralists to pursue herd growth without considering its effect on rangeland ecologies. The misunderstanding that pastoralists sought herd growth for reasons of 'prestige' or religious value only (associated with Herskovits but still repeated by development theorists) exacerbated misperceptions regarding pastoral range degradation. Careful study of domestic pastoral livestock production strategies focused on the multiple functions of domestic stock, the careful optimizing of milk allocation both for family subsistence and calf nourishment, and strategies of allocating pasture labor to best care for herds in the context of accessing pasture, water, and mineral resources. McPeak and Little (2006) studied the judicious use of livestock markets in Eastern Africa, demonstrating the high degree of market integration of pastoral producers and their important contribution to the regional food and export economy. The transnational flow of livestock from pastoral areas feeds great capital cities and provides significant exports, which should vindicate pastoralists against the charge that they are not contributing to national economies. The cash needs of households have grown as pastoralists participate in complex national and global institutions, including schools, clinics, trade in clothing, tools and foodstuffs, and commodities that support animal husbandry, including veterinary drugs (Catley et al., 2013). Two common herding strategies involve moving the community along with herds or moving herds out from a settled community, the latter facilitating agropastoral pursuits. As pointed out by Barth (1961), both rich and poor pastoralists tend to settle, but, whether sedentarized or not, pastoral groups are increasingly diversifying their sources of household revenues, whether through the buying and selling of livestock at market, cultivation where possible, business or professions for the more highly educated and relatively wealthy, and wage labor and trade for the less wealthy and poorer. Diversification for some households is a poverty response as their herds diminish, but for other households it is a strategy for growth built on the base of animal husbandry. One outcome of economic stratification in both West and East Africa is the increasing investment in livestock by members of the urban elite, whether or not of pastoralist origin, who become absentee herd owners with stock tended by hired herders (Little, 1992). In the work of Homewood and collaborators, Maasai livelihood diversification is seen as an enlargement of economic opportunities within a complex economy, not a turning away from animal production (Homewood et al., 2009). For some, wildlife conservation offers opportunities for cash incomes from land leasing and tourism, but these revenues tend to be captured by local elites rather than being shared among pastoral communities.

The Politics of Pastoralism in National and Global Systems

Increasingly the Middle Eastern and Central Asian political model of nomadic herders integrated in complex relations with states and markets (Khazanov, 1984), the latter providing crucial food supplements and supplies on which husbandry depends, seems applicable to sub-Saharan Africa. In this regard, markets are an arm of government, responsible for

pulling pastoralists toward sites where they access social services and trade and participate in national and international exchanges of goods, information, and identities. Orbits of pastoral movement take into account the need to access towns to find social and political security, so semisedentarized homesteads are often established near towns where children can attend schools and everyone can access town life, most importantly health services and retail markets (Fratkin et al., 1994). Yet the centripetal force of sedentarization is constantly countered by the centrifugal force of pastoral movement to access pastures; in the context of semisedentarization near town, livestock are often moved by young men to more distant grazing, thus dividing the community between semipermanent settlements and mobile camps. The pattern of increasingly large settlements out of which livestock herds are moved to access distant grazing is replicated across the dry lands in response to insecurity and modernization, leaving some areas less intensively occupied but opening up opportunities for more diversified occupations in growing small towns that relieve environmental pressures that might otherwise be exerted by growing populations.

Though pastoralists are often characterized by urban dwellers in pejorative terms that associate them with their rural proclivities, as 'conservative,' resistant to change, traditionalists, and so on, they increasingly participate in globalized institutions, attending schools, accessing modern medicine, using markets, discussing national politics, tuning into radio and even television programs, and using mobile phones, the latter used to share information on available pastures and market prices, not to mention planning of gatherings, ceremonies, and martial pursuits. The social cohesion of many pastoral groups, involving strongly defended social boundaries, has in many instances been bolstered by resilient forms of local religion, involving shared beliefs and ritual practices. Schlee points out that to be a Borana or a Gabra was long associated with a refusal to embrace Islam. Yet increasingly pastoralists reflect a larger trend toward conversion to one of the world's great religions, notably Islam or Christianity, the result of which is to link often remote societies to much larger communities of belief, assuring them of participation in the globalized world. Though conversion is often more of a collective than an individual experience, a political choice, as it were, it often reflects deep discord within pastoral societies, between those who convert and those who resist conversion, the educated and noneducated, and those who vehemently affirm customary practices – not least important being those related to gender expectations – and those who call for change. Pastoral societies are increasingly socially diversified so resist singular characterizations.

The Land Factor

Rangeland Enclosures and Global Land Grabbing

In many areas, the most significant intervention of the state into rangeland affairs has been through the pursuit of 'pastoral development' carried out by means of land reforms that have sought to formalize pastoral land tenure, allocating identifiable lands to certain pastoral groups and over time to individuals. In some countries like Botswana, land reform has resulted in a pastoral elite gaining legal authority over vast lands via the 'borehole syndicates' described by Peters (1994), relegating small herders and hunters to proletarian status. Many early attempts in Niger, Tanzania, and Kenya to create formal pastoral groups with land rights over specified herding territories were largely unsuccessful, both because of the near impossibility of restricting grazing to defined territories under conditions of low and unpredictable rainfall and the difficulty of managing internal transactions in the absence of customary precedents for coordinating husbandry between hundreds of households managing thousands of livestock across dry land terrain (Sandford, 1983). Important studies of rangeland privatization demonstrate not that collectively held pastures lead to the 'tragedy of the commons' but that breaking trust when some individual 'enclose' and 'privatize' collectively held land initiates the 'tragedy of the enclosure,' as individuals inexorably seek their own piece of land, no matter how nonviable, if they fear a 'run' on the commons that might leave them with nothing (Galaty, 1994). Increasingly smaller units of land holding have been progressively demarcated among the Maasai of Kenya throughout the twentieth century, dividing pastoral commons into the Masai Reserve, into two and then three districts, into pastoral 'sections,' then Group Ranches, and finally individual family parcels under title deeds (Rutten, 1992). Maasai rangeland privatization has encouraged marginal agricultural practices, land speculation, and land sales, as land buyers often seek title deeds to use as collateral more than land itself, leading to land loss by herders, to the benefit of rural and urban elites and investors.

Socialist land reforms in the developing world often involved forced sedentarization and consolidation of pastoralists in formally constituted 'villages.' Despite economic liberalization many postsocialist governments claim radical title to all land, making pastoralists subject to arbitrary displacements when their land is appropriated for mining, wildlife conservation, or allocation to national or international commercial enterprises (Shivji, 1998). Given the alienation of mobile livestock keepers from government, leaders often view grasslands as unoccupied or used inefficiently, so ripe for allocating to outside investors (Galaty, 2013a). The Ethiopian government, for example, argues that its land lease policy makes otherwise unused lands available to multinational agribusinesses, but in fact it is engaged in what international observers call 'land grabbing' in the most fertile regions occupied by pastoralists, the 'wetlands in the drylands,' along the Awash River of the Afar, or the Omo River Valley, long the home of diverse Omotic, Sudanic, and Eastern Cushitic speaking pastoralists, such as the Dassanetch, Mursi, Bodi, and Arbore. Elsewhere, pastoralists engaged in community conservation and ethnotourism, often near National Parks, Game Reserves, or Conservation Areas such as Ngorongoro, find themselves elbowed aside by national elites or by tourist entrepreneurs allocated local lands by national governments. In the name of economic globalization, investors being allocated lands for commercial exploitation ignore property rights that in their own countries would require agreements with local landholders. In China, Tibetan, and Mongolian herders are being forcibly moved off their grazing lands under the principle of 'ecological resettlement,' a policy that

uses environmental degradation as a pretext for appropriating pastoral lands for both political and economic reasons. What seems clear is that claims by postsocialist governments to radical title to all national lands is a stratagem for depriving rangeland pastoralists, among others, of the lands that they require for animal husbandry or alternative uses emerging with livelihood diversification and social change. What national elites call 'development' is too often purchased with poverty for many pastoralists, among other rural dwellers.

Borders and Conflicts: How Pastoralists Resist the State

The state has become the strongest force controlling pastoral polities, and the imposition of national borders is one crucial mechanism in its doing so. Most national borders that cross dry lands (and the logic of national integrity most often puts wet and fertile regions at the center, low drylands at the borderlands!) slice across pastoral territories, dividing herding communities between contiguous countries. Especially in pastoral regions, this inconvenience disrupts social life and rational patterns of land use, since borders often divide markets from customers and sellers, dry season from wet season pastures, ritual sites from the ceremonial celebrators, and wife-givers from wife-takers. Pastoralists today do cross some borders with impunity, but are rarely unaware that they are doing so and as a result risk sanctions. But borders also afford opportunities and in this sense represent 'resources' that mobile people exploit (Feyisse and Hoehne, 2010). Pastoralists are some of the world's greatest black-market profiteers, moving goods back and forth between countries; Bedouin in Gaza and Sinai, Afar to and from Djibouti, Somalis into Kenya, are examples of pastoralists who have come to play major roles in regional economies, illicitly but profitably moving goods and guns acquired from distant producers to remote customers, as quintessential middle-men. They also move themselves between countries, acquiring multiple citizenships or identity cards, establishing credential as refugees in one place, while maintaining rights in sites of origin. In Moyale, between Kenya and Ethiopia, electoral competition draws ethnic compatriots from one country to tip the balance of voting in another (Fekadu, 2010).

Pastoral culture has long been associated with the potential for violence, because raiding and retaliating is intrinsic to an economy of valuable yet removable goods (i.e., domestic animals) in the absence of adequate political security, making the need to protect herds from predators and thieves culturally ingrained. Borders represent virtual curtains that hide raiding parties preparing to cross into a neighboring region, and a haven for thieves to retreat to after carrying out a successful foray. Alliances between contiguous cross-boundary segments of the same ethnic group provide resources for carrying out transborder transactions, but their occupying different nations can also begin to divide them into mutually hostile groups. Many insurgencies are carried out by pastoralists, often with the assistance of cross-border allies. The Tuareg insurgencies in Mali and Niger, the most recent associated with Islamist motives, have recurred throughout the twentieth century, and the Somali insurgencies, in Ethiopia, Kenya, Djibouti, as well as Somalia, have been so successful that they have completely eliminated the state! The now successful southern Sudan civil war for secession pitted southern pastoralists against the northern government, while the Darfur conflicts have seen pastoralists from Chad and northwest Sudan resist the formalization of tenure in highland areas that would prevent them from accessing seasonal pastures. The proposition that pastoralists are egalitarians because of their ethic of independence and mutual autonomy has some merit, but ignores that pastoral honor is often purchased at the price of the subjugation of other peoples in regional systems of stratification, whether among Bedouin or Baluch of the Middle-East or the many Saharan societies (Salzman, 2004).

Pastoral Futures

Is there a future for nomadic pastoralism? Constraints on the continuing practice of pastoralism include progressive land loss that diminishes the availability of pastures, pressures on pastoral families to settle, and the growing pull away from agrarian life for pastoral youth. Some argue that pastoralism is an archaic mode of subsistence unsuited to the demands of the modern world. Yet the world's great rangelands will continue to be dedicated to animal husbandry, all the more important as global populations and the demand for livestock products inexorably increase. Even if parts of the world's grasslands are diverted to wildlife conservation, mineral exploitation, irrigated or dry land agriculture, urban development or small-scale industry, much of the grasslands remote from urban centers will have no more productive use than extensive animal husbandry (Catley et al., 2013). The great alternative to pastoralism, which, recall, involves intensive use of labor to exploit land extensively, is ranching, which uses similar pastures to raise livestock for commercial ends using less labor due to the investment of capital in fencing, upgraded stock, fixed water points, feed supplements, veterinary medicine, motor vehicles, airplanes, modern communications, and so forth. In Botswana and Namibia, pastoralism has in many quarters given way to commercial ranching, under conditions of strict veterinary controls and restricted livestock movement, with the assurance of export agreements with the UK and the European Union. But in Kenya, land was taken from pastoralists at the beginning of the twentieth century, but commercial ranching enterprises there have never supplied more than a fraction of the national meat supply, that aimed at high-end restaurants and the urban bourgeoisie. In fact, the productivity of pastoral land use has proven higher than that of ranching, given the intensive investment of pastoral labor in raising and marketing livestock.

Where ranching thrives – in the US, Canada, Australia, and Argentina – settlers established enterprises on lands taken from indigenous peoples. What pastoralism has historically succeeded in doing is to enable the productive human habitation of the world's great grasslands, through an economic pursuit that supports an optimal number of people on land with few alternative uses, while efficiently providing growing local, national, regional, and global markets with livestock and meat. At present, a balance is maintained between available land, the number of livestock that can be supported on that land, and humans who can be productively engaged and employed in pastoral livelihoods. Population growth in the world's rangelands is largely absorbed in the diversification of rural

livelihoods and in the growth of small towns across savanna lands. Although many from pastoral communities will gain education, will come to appreciate national social services and urban amenities, and will grow distant from pastoral pursuits, some will remain on the land, changing to husbandry practices based on commercial models, using new skills, techniques, and technologies in adapting animal husbandry to new conditions. Pastoral futures will undoubtedly include mobile animal rearing in the grasslands, pursued by communities that bear the imprint of modernity while continuing to occupy dry regions of the globe (Galaty, 2013b).

See also: Age: Anthropological Aspects; Borders, Anthropology of; Community Studies: Anthropological; Conflict: Anthropological aspects; Development: Social-Anthropological Aspects; East Africa: Sociocultural Aspects; Economic Anthropology; Ethnography; Exchange in Anthropology; Fieldwork in Social and Cultural Anthropology; Food Production, Origins of; Friendship, Anthropology of; Households in Anthropology; Human Rights and Social Work; Human Rights, Anthropology of; Identity in Anthropology; Indigeneity; Land Tenure; Law: Anthropological Aspects; Markets: Anthropological Aspects; Migration: Anthropological Perspectives; Modernity; Nongovernmental Organizations: Anthropological and Historical Aspects; Pastoralism in Anthropology; Property: Anthropological Aspects; Sacrifice; Tourism; Violence in Anthropology.

Bibliography

Adugna, F., 2010. Making use of kin beyond the international border: inter-ethnic relations along the Ethio-Kenyan border. In: Feyisse, D., Hoehne, M. (Eds.), Borders and Borderlands as Resources in the Horn of Africa. James Currey Press, Suffolk, pp. 45–59.

Barth, F., 1961. Nomads of South Persia: The Basseri Tribe of the Khamseh Confederacy. Allen & Unwin, London.

Bollig, M., 2006. Risk Management in a Hazardous Environment: A Comparative Study of Two Pastoral Societies. Springer Publications, New York.

Catley, A., Lind, J., Scoones, I. (Eds.), 2013. Pastoralism and Development in Africa: Dynamic Change at the Margins. Earthscan from Routledge, Oxon.

Dahl, G., Hjort, A., 1976. Having herds: pastoral herd growth and household economy. In: Studies in Social Anthropology, vol. 2. Stockholm.

Dyson-Hudson, N., Dyson-Hudson, R., 1980. Nomadic pastoralism. Annual Review of Anthropology 9, 15–61.

Ehret, C., 2002. The Civilizations of Africa: A History to 1800. University of Virginia Press, Charlottesville.

Ensminger, J., 1992. Making a Market: The Institutional Transformation of an African Society. Cambridge University Press, Cambridge.

Evans-Pritchard, E.E., 1940. The Nuer. Oxford University Press, Oxford.

Feyisse, D., Hoehne, M. (Eds.), 2010. Borders and Borderlands as Resources in the Horn of Africa. James Currey Press, Suffolk.

Fratkin, E., Galvin, K., Roth, E. (Eds.), 1994. African Pastoralist Systems: An Integrated Approach. Lynn Rienner Publishers, Boulder.

Galaty, J., 1994. Rangeland tenure and African pastoralism. In: Fratkin, E., Galvin, K., Roth, E. (Eds.), African Pastoralist Systems: An Integrated Approach. Lynne Rienner, Boulder, London, pp. 185–204.

Galaty, J., 2013a. Land grabbing in the Eastern African rangelands. In: Catley, A., Lind, J., Scoones, I. (Eds.), Development at the Margins: Pathways of Change in the Horn of Africa. Earthscan, at Routledge, London, pp. 143–153 (Chapter 12).

Galaty, J., 2013b. The indigenization of pastoral modernity: territoriality, mobility, and poverty in dryland Africa. In: Bollig, M., Wotzka, H.-P. (Eds.), African Pastoralism: Past, Present, Future. The Emergence, History and Contemporary Political Ecology of African Pastoralism. Berghahn Press, Oxford, pp. 473–510.

Galaty, J., Johnson, D. (Eds.), 1990. The World of Pastoralism: Herding Systems in Comparative Perspective. Guilford Press, New York.

Galvin, K., et al. (Eds.), 2008. Fragmentation in Semi-arid and Arid Landscapes. Springer Publications, New York.

Hardin, G., 1968. The tragedy of the commons. Science 162, 1243–1248.

Herskovits, M., 1926. The cattle complex in East Africa. American Anthropologist 28.

Hodgson, D., 2011. Being Maasai, Becoming Indigenous: Postcolonial Politics in a Neoliberal World. Indiana University Press, Indianapolis.

Homewood, K., 2008. Ecology of African Pastoralist Societies. James Currey, Oxford.

Homewood, K., Kristjanson, P., Trench, P.C., 2009. Staying Maasai: Livelihoods, Conservation, and Development in East African Rangelands. Springer, New York.

Ingold, T., 1984. Hunters, Pastoralists and Ranchers: Reindeer Economies and Their Transformations. Cambridge University Press, Cambridge.

Khazanov, A., 1984. Nomads and the Outside World. Cambridge University Press, Cambridge.

Lefébure, C., 1979. Introduction: the specificity of nomadic pastoral societies. In: Pastoral Production and Society, Équipe ecologie et anthropologie des societes pastorales. Maison des Sciences de l'Homme, Paris and Cambridge University Press, Cambridge.

Little, P., 1992. The Elusive Granary: Herder, Farmer, and State in Northern Kenya. Cambridge University Press, Cambridge.

McCabe, J.T., 2004. Cattle Bring Us To Our Enemies: Turkana Ecology, Politics, and Raiding in a Disequilibrium System. University of Michigan Press, Ann Arbor.

McPeak, J., Little, P., 2006. Pastoral Livestock Marketing in Eastern Africa: Research and Policy Challenges. Intermediate Technology Publications, Bourton on Dunsmore.

Peters, P., 1994. Dividing the Commons: Politics, Policy and Culture in Botswana. University of Virginia Press, Charlottesville and London.

Rutten, M.M., 1992. Selling Wealth to Buy Poverty: The Process of Individualization of Landownership among the Maasai Pastoralists of Kajiado District, Kenya, 1890–1990. Verlag Breitenbach Publishers, Saarbrücken, Fort Lauderdale.

Salzman, P.C., 2004. Pastoralists: Equality, Hierarchy, and the State. Westview Press, Boulder.

Sandford, S., 1983. Management of Pastoral Development in the Third World. Overseas Development Institute, John Wiley & Sons, Chichester.

Schlee, G., 1989. Identities on the Move: Clanship and Pastoralism in Northern Kenya. Manchester University Press, Manchester.

Schneider, H.K., 1979. Livestock and Equality in East Africa: The Economic Basis for Social Structure. Indiana University Press, Bloomington and London.

Scoones, I. (Ed.), 1994. Living with Uncertainty: New Directions in Pastoral Development in Africa. Intermediate Technology Publications, London.

Shivji, I.G., 1998. Not yet democracy: reforming land tenure in Tanzania. International Institute for Environment and Development, HAKIARDHI, and the Faculty of Law, University of Dar es Salaam, London.

Spear, T., Waller, R. (Eds.), 1993. Being Maasai: Ethnicity and Identity in East Africa. James Currey, London.

Spencer, P., 1998. The Pastoral Continuum: The Marginalization of Tradition in East Africa. Clarendon Press, Oxford.

Patenting in the Biological Sciences

Myles W Jackson, New York University, New York NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A. Webster, volume 11, pp. 7612–7615, © 2001, Elsevier Ltd.

Abstract

This article traces the history of patenting products of nature with a view to explaining how it is that genes were deemed as patent eligible material by all major patent offices worldwide. The article concludes with a current discussion of the recent legal challenge mounted against gene patenting by the American Civil Liberties Union against Myriad Genetics' patents on the so-called breast cancer genes.

Introduction

In an interview after the announcement in 1955 of the success of the polio vaccine, Edgar R. Murrow asked Jonas Salk, "Who owns the patent on this vaccine?" Salk famously replied: "Well, the people, I would say. There is no patent. Could you patent the sun?" (As cited in Johnson, 1990.) Interestingly, Salk actually tried to patent the vaccine, but the application was rejected. Similarly, as Robert Kohler has demonstrated in his *Lords of the Fly*, the of the early twentieth century had a network of credit based on the sharing of mutant flies, techniques, and data. Personal credit and commercialism were deemed anathema to the collective spirit of scientific research (Kohler, 1994). At that time, geneticists enjoyed the openness of a scientific republic. By the 1970s, that was beginning to change, as intellectual property issues began to invade the biological sciences: the future of start-up biotech firms such as Genentech was predicated upon successful patenting of biological entities.

Commercialization of the biological sciences began in the 1970s with the rise of biotech firms (Shapin, 2008). Sheila Jasanoff has argued that the biotechnology industry in the United States could not exist as one now knows it without patent protection (Jasanoff, 2005). Biotech patents extended ownership rights to entities that had previously been considered to be outside the realm of patentability, rendering these objects as marketable commodities. Typically, in the early stages of a biotech company's existence, there were no marketable products. This was particularly true of the sequencing companies. Hence, intellectual property was the only source of credibility for a fledgling company looking to attract both corporate venture capitalists and less aggressive investors. Patents provided some assurance that there would be returns from future marketable products.

The theme of profiting from another person's body with a view to commodifying biological entities has of course been fraught with controversy and was the theme of a landmark California State Supreme Court Case, *Moore v. the Regents of California* (1990; see Rabinow, 1996; Boyle, 1996). It dealt with issues of commercialization and ownership of biological materials. A patient, John Moore, suffered from hairy-cell leukemia and was being treated by Dr David W. Golde of the University of California, Los Angeles Medical Center. Golde suggested that Moore's spleen be removed, and Moore signed the medical consent form, which stated that the hospital would dispose of, or cremate, any tissue removed from his body. Unbeknownst to Moore, Golde created a cell line from Moore's T-cells, clearly aware of the financial potential of such a research tool. The Regents of California received a patent on the cells in 1984. In 1983, Moore had been presented with another consent form, asking him to waive his rights to any profits from his cell line or any other type of tissue removed from his body. He refused to sign and showed the form to his lawyer, who then investigated and found out about the patent. Moore sued. The court opined that while Dr Golde was legally required to inform his patient that he had a financial interest in any samples removed from Moore's body, Moore did not have property rights to his T-cell line (Rabinow, 1996). In a similar, more tragic, example, Henrietta Lacks' family never received financial compensation for the immortal cell line (HeLa cells) derived from a tumor in her cervix (Skloot, 2010).

This desire for patent protection in the age of biocapitalism and biocolonialism, as numerous STS scholars have labeled it, however, raises an interesting conundrum (Pottage, 1995; Rajan, 2006). According to the product-of-nature doctrine, natural products are not patentable. So, how can one patent biological materials, the cornerstone of the biotech sector? The product-of-nature doctrine, while never explicitly invoked in a particular case, represents the accumulation of numerous US Supreme Court decisions disallowing the patenting of organisms or substances as they occur in nature. Simply stated, they are not inventions, and therefore are not patentable subject material. That said, as Beauchamp has recently argued, nineteenth-century Supreme Court rulings deemed some 50 years later as being relevant to the product-of-nature doctrine were not obviously so when they were originally decided (Beauchamp, 2013). The first Supreme Court decision relevant to the patenting of a natural product was the *American Wood Paper Company vs. Fibre Disintegrating Company* case (1874) (henceforth *American Wood-Paper Patent*), which revolved around the question of whether paper pulp (cellulose) extracted from wood via chemical processes and from a vegetable substance by means of chemical and mechanical processes could be patented. The mechanical process involved a machine that shreds the wood or vegetable substance into shavings, and the requisite chemical processes included a myriad of reactions, such as boiling in a strong alkali solution under high temperature and pressure,

chlorination, and washing with hydrochloric acid. The product, cellulose, was not patent eligible because it was well known at the time. That part of the patent failed due to prior art (Beauchamp, 2013).

Ten years later, the Supreme Court ruled on another case relevant to the patenting of a natural substance, *Cochrane v. Badische Anilin & Soda Fabrik* (1884). The reissued letters patent No. 4321 had been granted to C. Graebe and C. Liebermann of BASF for both the processes of producing alizarin, a compound found naturally in the root of the madder plant, and the composition of matter itself, which they called artificial alizarin. The Court ruled against composition-of-matter patent, concluding along the same lines as the *American Wood-Paper Patent* case that such a patent could not be awarded for artificial alizarin, since it is chemically equivalent to the naturally occurring substance, which was well known at the time. Again, it failed on grounds of prior art. The new method of production, however, was indeed deemed patentable.

The United State Patent and Trademark Office (USPTO) followed the rule established by these Supreme Court cases. In 1888, William Latimer applied for a process patent to extract the fibrous core from the center of the pine needles of Longleaf Pines (*Pinus australis*). He also applied for a composition-of-matter patent on the pine needles themselves as their centers could be used in the burgeoning textile industry. Citing these earlier cases – namely, *American Wood-Paper Patent* and *Cochrane* – the examiner argued that while the process was patentable, the product was not, as it was a product of nature, and therefore not an invention. The US Patent Commissioner affirmed the examiner's rejection in his *Ex parte Latimer* ruling (Harkness, 2011).

The Twentieth Century

Things were about to get interesting with respect to the patentability of natural products. *Parke-Davis & Co. v. H. K. Mulford Co.* (28 April 1911, 189 *Federal Register* 95), although not a Supreme Court case, was a patent-infringement case that raised the issue of patentability of biological entities. Indeed, this decision, written by Judge (Billings) Learned Hand, has been often referred to in the legal literature as the precedent for the early cases patenting of genes (Andrews, 2002; Doll, 1998). The case commenced with the affirmation of the patentability of a substance isolated from animal tissue that had therapeutic value, even though it differed from its natural homolog only in degree of purity. The key question was whether Parke-Davis' biochemist Jokichi Takamine's purification of adrenaline was sufficient to render it new therapeutically and commercially, and therefore patentable (Harkness, 2011).

Patent examiner James B. Littlewood, who was responsible for adjudicating the original patent application for Adrenalin – the commercial name for the mostly purified, patented adrenaline – in 1900 denied the patent multiple times after various revisions by the patentee on the grounds that it was a product of nature. Littlewood finally granted the patent on the seventh review on 2 June 1903. The critical point, however, was that Takamine's lawyers argued that 'Adrenalin,' the medical product, was something different from the naturally existing hormone (Harkness, 2011).

Adrenalin, of course, proved to be a rather important medical product. Not surprisingly, Parke-Davis & Co. jealously enforced their patent, and in 1905 engaged in a priority dispute with H. F. Mulford. The case now became one of patent infringement, and precious little was said about Adrenalin being a product of nature. The experts brought in by both sides did not address this issue of patenting natural products; they were interested only in matters of priority.

Priority was critical, since during the early years of the twentieth century, a number of adrenaline-based products were being sold in the United States and Europe. Specifically, Hoechst of Germany was marketing Suparenin, but Parke-Davis & Co.'s Adrenalin was relatively pure, safe, and efficacious (Dutfield, 2009). Interestingly, chemical and pharmaceutical substances were not patentable in Germany at that time. The first pan-German patent law, created after unification passed in May 1877, explicitly stated that the following inventions were excluded from patent protection: those "that, if developed, are against [the] law or good morals" (a sentiment that is also present today in European Patent Convention clause 53), and those "consisting of food, stimulants or therapeutic agents, as well as the invention of substances produced by means of chemistry if the invention does not include a specific production process" (as cited in Gaudillière, 2008). Unlike Britain, or the United States, chemical substances (including drugs) could not be patented in late nineteenth- and early twentieth-century in Germany. Schering was able to patent hormones in the 1920s and 1930s thereby establishing the precedents in Germany for subsequent patenting of purified biological substances (Gaudillière, 2008). Such a decision was seen to be in the interest of the public good and as such a monopoly on drugs was deemed both to be dangerous to the public welfare and to be deleterious to innovation. Patenting also seemed superfluous to the organic chemists who synthesized such substances. A cadre of top-notch German chemists could invent a myriad of processes, which of course could be patented, to produce a given chemical; hence, companies could create a wall of patent protection around a particular chemical substance. German chemists from August Wilhelm von Hoffmann to Paul Ehrlich and Carl Duisberg were renowned for their abilities in synthetic organic chemistry. German governmental officials worked closely with chemical company officials in order to draft and tailor the nation's patent law to the interests of the chemical industry (Dutfield, 2009). At the time, the American Pharmaceutical Association lobbied the federal government for weaker patent protection, as they argued that they were facing unfair competition from German companies. And they advocated in favor of changing the law such that only patents for chemicals processes and not the chemicals themselves could be patented in the United States. The government, however, did not change its stance.

Lacking a strong tradition in chemical synthesis, the burgeoning US and British pharmaceutical industries were centered on the extraction and purification of natural substances, rather than the organic synthesis of artificial substances (Lesch, 2007). Not surprisingly, in the United States this industry turned to patent laws, which starting with Judge Learned Hand's decision permitted the patenting of natural substances as long as they were isolated and purified.

Purification was crucial, as the United States introduced drug regulations in 1902 that required the establishment of a governmental laboratory to test and monitor new drugs for their safety and efficacy (Dutfield, 2009). American scientists were able to combine their practices with natural-product chemistry, fermentation, and endocrinology, with synthetic organic chemistry techniques gleaned from their research with German university chemists. Natural product-based companies, such as Parke-Davis, began to compete with the well-established German pharmaceutical companies. Indeed, from World War II to 1986, the patenting of products of nature in the form of antibiotics, hormones, and vitamins was dominated by the United States and, to a lesser extent, the United Kingdom (Achilladelis, 1993). In short, two different patent regimes helped to foster two different research traditions. In the Anglo-American world, which included the United States, United Kingdom, and Canada, pharmaceutical research was centered on physiological chemistry and the isolation and purification of natural products from plants, animals, and humans, whereas in Germany and Switzerland, such research was predicated on synthetic organic chemistry being carried out not only in universities but also in research laboratories of the dyestuff industries. These chemists synthesized the active principles found naturally in plants and animals and improved upon their natural analogs for use as medications.

On 28 April 1911, in his landmark ruling in *Parke-Davis*, Judge Hand of the District Court for the Southern District of New York upheld Takamine's patents, famously arguing:

> But, even if it [Adrenalin] were merely an extracted product without change, there is no rule that such products are not patentable. Takamine was the first to make it available for any use by removing it from the other gland-tissue in which it was found, and, while it is of course possible logically to call this a purification of the principle, it became for every practical purpose a new thing commercially and therapeutically. That was a good ground for a patent. … The line between different substances and degrees of the same substance is to be drawn rather from the common usages of men than from nice considerations of dialectic.
>
> (28 April 1911, 189 *Federal Register* at 103)

The patenting of the first generation of hormones in general, and adrenaline in particular, proved to be a challenge to what was now in the 1920s seen as the well-established product-of-nature doctrine in the United States, since the hormones were not created artificially (i.e., synthetically), but were isolated and purified from nature. While Hand's ruling undoubtedly benefited the US pharmaceutical companies in gaining the lead over their German counterparts, it also increasingly blurred the boundary between natural and artificial, nonpatentable and patentable (Dutfield, 2009). This case paved the way for the subsequent patenting of sex hormones, steroids, vitamins, and antibiotics in the United States (Achilladelis, 1993; Dutfield, 2009). Although penicillin was never patented (the processes for its mass production of course were), most antibiotics thereafter were patented despite being products of nature.

In 1930, Congress passed the Plant Patent Act, which permitted patents to be issued to persons who had "invented or discovered and asexually reproduced any distinct and new variety of plant, other than a tuber-propagated plant" (as cited in Kevles, 2007; see also Bugos and Kevles, 1992). Congress did query whether or not this act would contradict the product-of-nature doctrine, and whether a new variety of plant would be considered an invention or discovery, and consequently if the breeder should be considered an inventor. The House Committee on Patents argued that, although a new plant variety was indeed a product of nature, it was created through human intervention. The plants had characteristics, which differed from those that grew in the absence of human agency. The link between biological and chemical entities was established early on, as the House Committee insisted upon the analogy between "the part played by the plant originator in the development of new plants and the part played by the chemist in the development of new compositions of matter" (cited in Bugos and Kevles, 1992). The House Committee ruled that a new variety of plant that was cultivated (rather than found in the field) was a type of discovery worthy of patent protection, and its cultivator was the discoverer. Patent protection at this point did not extend to seeds, as those skilled in the art at the time believed that 'invented' seed plant varieties were too unstable across generations. The topic of patenting subsequent generations via seeds and genes became highly controversial decades later.

While decisions of the USPTO seemed confused and contradictory, no such ambiguity seemed to exist at the level of the Supreme Court, whose decisions of the 1920s through 40s paralleled the cracking down on patents by the Antitrust Division of the Department of Justice (Beauchamp, 2013). In 1948, the Supreme Court ruled on the case, *Funk Brothers Seed Co. v. Kalo Inoculant Co*. Kalo Inoculant sued Funk Brothers for patent infringement, which claimed the product and process of creating a mixture of root-nodule bacteria for inoculating legumes. Funk Brothers argued that the patent was invalid. Kalo's patent included six strains of naturally occurring bacteria, each of which could be used to fertilize some legumes without inhibiting the actions of the others. The patent holder combined these strains so that a farmer could purchase one fertilizer for various plants, rather than one for each type of plant (McGarrigle and Norviel, 2008). The bacterial strains did not inhibit one another. The Seventh Circuit ruled that the inventor had produced new and different composition of noninhibitive strains, which possessed utility (253 F.2d 156; 116 U.S.P.Q. 484). The Supreme Court disagreed, arguing that the implementation of a natural principle or phenomenon was not patentable. Indeed, the properties of each strain of bacteria were exactly those exhibited in nature, even though the combination had utility.

The ruling of the Supreme Court not withstanding, mixed signals were being sent from the lower courts. *Merck & Co., Inc. v. Olin Mathieson Chemical Corporation* of 1958 was the relevant case for patentability tried after the passing of the Patent Act of 1952 (253 F.2d 156; 116 U.S.P.Q. 484). In 1926, scientists discovered that patients suffering from pernicious anemia could benefit from adding a large quantity of cattle liver to their diets. Some 20 years later, new liver extracts could be administered to patients; however, they were expensive, and not all patients could tolerate them. At that time, biochemists could not isolate or identify the specific agent responsible for the improvement in the patients' health. Dr Mary S. Shorb,

a bacteriologist at the Department of Agriculture and later at the University of Maryland, developed an assay with a microorganism, *Lactobacillus lactis Dorner* (or LLD), to discover this unknown agent. In 1947, she entered into a collaboration with scientists at Merck, one of whom was Thomas R. Wood, who had been working on this unknown compound since 1938. Wood hired a chemist, Edward L. Rickes, and in October 1947 they isolated a pinkish compound, which upon further purification and concentration turned out to be a red crystalline material, which they subsequently named vitamin B_{12}.

In 1957, the district court ruled that the patent of vitamin B_{12} was invalid on the grounds that it was a product of nature and that there was a lack of invention. In reversing the decision a year later, the US Court of Appeals for the Fourth Circuit specifically addressed the 'product of nature' defense. The court interpreted the Patent Act of 1952 as not precluding products of nature for patent eligibility as long as they were new and useful. The court ruled that, contra the *American Wood-Paper Patent*, the vitamin B_{12} produced was new, and it possessed unique properties vis à vis the naturally occurring substance. The US pharmaceutical industry lobbied to ensure that the 1952 Patent Act would permit the patenting not just of vitamins but also of discovered antibiotics as well as the techniques of their mass production and screening (Kingston, 2004; Dutfield, 2009).

Throughout the 1960s and 1970s, more rigorous agricultural intellectual property statutes were being sought in the United States due to the Green Revolution shepherded by Norman E. Borlaug, as well as the rising domestic seed market and seed industry (Bugos and Kevles, 1992). On 2 December 1961, an International Conference for the Protection of New Plant Products in Paris created the Union International pour la Protection des Obtentions Végétales, or UPOV. Initially, UPOV sought intellectual property protection of plant varieties by means other than patents. Since the United States declined to join unless patents were included in such protection, the UPOV acquiesced, and in 1978, patents were granted on patent-eligible plants internationally.

The key to this more recent legislation, then, is the relationship between the patented entity and its natural homolog. Products of nature are only rendered patentable if, by virtue of human hand bringing about the isolation and purification, the new substance is different in kind from the natural product, echoing Learned Hand's sentiment (Sweet, 2010).

1980s and 1990s: The Post *Diamond v. Chakrabarty* World

The 1980 *Diamond v. Chakrabarty* ruling is the only Supreme Court case cited as precedent for the patenting of genes (447 U.S. 303 (1980)). While working at General Electric, the microbiologist A. M. Chakrabarty created a new bacterium of the *Pseudomonas* genus. The initial utility claimed by Chakrabarty for this artificial bacterium was the production of proteins from the degradation of hydrocarbons (Kevles, 1994). Later, it became clear that this bacterium possessed the coveted attribute of being able to break down crude oil into biodegradable materials. Clearly, such a quality was applicable to oil spills. At the time of his patent application, there were four known species of bacteria that could metabolize different components of crude oil. These species, however, competed with each other, thereby limiting the amount of oil they could convert. Chakrabarty took the genes coding for the proteins that degrade the oil and cloned them in bacterial vectors. Irradiating the bacteria with ultraviolet light after introducing the genes into the recombinant DNA vectors, he discovered a method for cross-linking that resulted in all four genes being present in one bacterium. This newly created bacterium could degrade two times more oil than the four original species of bacteria. We are speaking here of a bacterium that clearly was the product of human hands, as it did not exist before Chakrabarty's invention, and it was unlikely that the natural bacteria could break down more components of the oil.

A significant amount of human intervention was required, and such intervention needed to render the artful product sufficiently different from other types of naturally existing *Pseudomonas* bacteria (Palombi, 2004). In short, while it cannot be accurately claimed that it serves as a legal precedent for gene patenting, it can be argued that *Diamond v. Chakrabarty* set the milieu for the patenting of genes. This case marks the first of many victories of the biotech industry to alter biotech patent rules in ways that suited their financial interest (Dutfield, 2009).

Their next round of biotech's successes occurred 2 years later. Intellectual property (IP) lawyers persuaded the USPTO and the courts that genes were chemical compounds; hence, chemical IP could be applied to biotech patents. With this reasoning, one of the first gene patents – the Adrenocorticotropin-lipotropin precursor gene, US Patent 4 322 499 – was awarded in March 1982 (Calvert and Joly, 2011). The analogy between genes and chemicals seemed a perfect one for the USPTO, since patent law in the United States had by this time reached a stage whereby the patenting of isolated and purified natural products was well established (Conley and Makowski, 2003; Demaine and Fellmeth, 2002). This decision went unchallenged until the end of the 1990s, when scientists from leading research institutions began to query the sagacity of reducing genes to chemical compounds (Calvert and Joly, 2011).

Diamond v. Chakrabarty paved the way for the patenting of life forms. In 1988, the USPTO granted a patent on living animal for the first time in the office's history: Harvard's oncomouse, a genetically engineered rodent that was predisposed to cancer (Kevles, 2002). DuPont Corporation, which had given Harvard $6 million to fund research on the oncomouse, enjoyed an exclusive license for its use (Kevles, 2002; Haraway, 1997). Much like Chakrabarty's *Pseudmonas* bacterium, the transgenic mouse did not exist in nature. Although it was clearly a living mammal, it was the creation of human hands in the laboratory. As Fiona Murray astutely points out, the oncomouse changed the so-called cycles of credit in the academy (Murray, 2011). DuPont's commercial economy based on its exclusive license seemed to many to be antithetical to the scientific, moral economy of knowledge sharing. That moral economy is now often regarded as an antiquated model. Many academic scientists own patents in this new age of the academic economy of credit. Patents establish a new type of credit distinct from and in addition to the published paper. Not all scientists with patents, however, enter into exclusive licensing agreements with biotech firms.

Canada, Europe, and the United States

The oncomouse case is an illustrative one, as it shows the differences among three major patent offices despite attempts by Trade-Related Aspects of Intellectual Property Rights to harmonize patent law worldwide. In 2002, the Supreme Court of Canada opined that higher forms of life were not patent eligible since they were not manufactured within the meaning of invention of the Patent Act of 1869. While microorganisms might be the result of a human mixture and therefore patent eligible, the body of a mouse is not. Unlike the USPTO, the European Patent Office (EPO) considered issues of morality, which have played a much larger role in European patent deliberations (Jasanoff, 2005). Indeed, article 53 of the European Patent Commission cites two areas where patents are to be negated: first, if they are deemed to be contrary to 'ordre public' or morality, and second, if they are applied to plant or animal varieties, or essentially biological processes for the production of plants and animals. In 2004, the EPO ruled that the oncomouse patent was indeed valid. They did not consider the oncomouse an animal variety, but rather a product of human intervention. With respect to the ordre public clause, they reckoned that the benefits humans gained from the creation of the oncomouse outweighed the transgenic rodent's suffering.

Genes are considered patentable by all major patent offices. Article 5(1) of the European Union (EU) biotech directive states: "The human body, at the various stages of its formation and development, and the simple discovery of one of its elements, including the sequence or partial sequence of a gene, cannot constitute patentable inventions" (as cited in Jasanoff, 2005). Articles 5(2) and 5(3), however, state that isolated gene sequences may be patented, even if their sequences are identical to those occurring in nature, if they are derived by a technological procedure and they possess an identifiable industrial application. France initially objected to the patenting of human genes, and would not support the EU biotech directive until it finally acquiesced and signed on in 2004. Germany and France have provisions in their law narrowing the scope of biotech patents to the recited utility. These provisions apply only to biotech patents, and are highly controversial.

Between January 1980 and December 2003, 15 603 patent families claiming human DNA sequences were patented. Of those sequences, 94% received at least one patent from the USPTO, while less than 5% (750) were granted at least one patent by the EPO (Hopkins et al., 2007). The larger health care market the United States offers, the longer time it takes for the EPO to review patents, and the EPO's higher bar of patentability all contribute to those skewed numbers. The three major US sequencing companies (Human Genome Sciences (HGS), Incyte, and Millennium Pharmaceuticals) have found it particularly difficult to have their patent applications approved by the EPO. As of 2005, Incyte had only two European patents granted out of 485 applications, while Millennium had no patents out of 94 applications, and HGS had no patents out of 378 applications. The EPO's more stringent requirements for specific utility claims and the insufficiency of sequence homology to establish functional/utility claims combine to explain those paltry figures. Indeed, the EPO's stance reflects a much greater awareness of, and concern with, gene patenting among Europeans.

American Civil Liberties Union vs. Myriad Genetics

Most recently, the question of patenting natural products was the key issue in the controversial ruling in the United States of Judge Robert W. Sweet of the US District Court of Southern New York. The case, *Association for Molecular Pathology et al. vs. USPTO et al.* (henceforth, the Myriad Case), is an attempt by numerous scientific and medical organizations headed up by the American Civil Liberties Union (ACLU) to render Myriad Genetics' breast cancer (BRCA) 1 and 2 gene patents invalid by arguing, among other things, that genes are products of nature and therefore are not patentable as defined in section 101 of US patent law (Parthasarathy, 2007). In March 2010, Judge Sweet ruled that the Supreme Court precedent declares that products of nature are not patentable unless "a change [is made] that results in the creation of a fundamentally new product" (Sweet, 2010). In short, "markedly different characteristics" are, according to Sweet's interpretation of previous Supreme Court rulings, critical to satisfy section 101 of US patent law (447 U.S. 303 [1980] at 310).

Myriad's attorneys responded to Sweet's ruling by appealing to the US Court of Appeals for the Federal Circuit (CAFC) (Brief for the Appellants, No. 2010-1406.) The Acting Solicit General, Neal Katyal, intervened on 4 April 2011, marking the first time that the top legal advocate of the executive branch appeared before the CAFC (Frankel, 2011). He did not support the USPTO's position that isolated DNA molecules are patentable, rather he urged a compromise via the so-called magic microscopic test. If an imaginary microscope could focus on a claimed DNA sequence as it exits in the human body, then the DNA is a product of nature, and therefore not patentable (Lourie, 2011). Accordingly, since BRCA 1 and 2 genes could, in principle, be seen with this magical microscope, they should not be patentable. This microscope, however, could not find cDNA sequences in the human genome, since cDNA is the product of the human hand and is, in his view, therefore patentable.

The CAFC did not seem to be persuaded by this analogy. By a vote of two to one, they ruled on 29 July 2011, that Myriad's patents on isolated sequences of DNA were patent eligible, thereby reversing Judge Sweet's decision of March 2010. Judge Alan D. Lourie of the CAFC concluded that genes are indeed patentable, as they "have a distinctive chemical identity and nature from molecules that exist in nature" (Lourie, 2011). DNA sequences are not contained within a physical mixture, and unlike uranium or adrenaline they cannot be purified; they must be chemically cleaved from their linkage with other genetic materials. Or, as Lourie explains, "[i]n other words, in nature, isolated DNAs are covalently bonded to such other materials. Thus, when cleaved, an isolated DNA molecule is not a purified form of a natural material, but a distinct chemical entity" (Lourie, 2011).

Federal Circuit Court Judge Kimberly A. Moore concurred with Lourie; however, her reasoning was slightly different (Moore, 2011). Pace Lourie took *Funk Brothers* and *Chakrabarty* as the precedent cases for establishing the

criteria of patentability; however, she argued that a mere purification of a substance is normally insufficient grounds for patentability. Moore turned to the chemical structure of DNA fragments to argue for their uniqueness and artificiality. Fragments, which include DNA isolates, have different chemical properties than the parent molecule from which they were removed.

In addition, cDNA does not exist in nature: it is the product of the laboratory after adding reverse transcriptase to mRNA. cDNA does not have the same sequence as mRNA (indeed, it has the complementary sequence), and because cDNA contains thymine while mRNA contains uracil and deoxyribose sugar as opposed to ribose, cDNA is more stabile than mRNA. Therefore, cDNA is, in Moore and Lourie's views, patentable. Moore never discussed the nonobviousness criterion, however, which (ironically) becomes an obvious problem if cDNA is patentable but isolated DNA is not.

What about DNA sequences that have the same sequence as the DNA of the gene found in nature? For Moore this was a different, and more complex, issue. She reasoned the fact that these smaller isolated DNA sequences are a part of 'the larger chromosomal polymer' does not render them unpatentable. These molecules are not quite analogous to naturally occurring minerals; rather, isolated DNA molecules possess different chemical ends than the naturally occurring DNA of the chromosome (Moore, 2011).

The last portion of Moore's opinion is rather disquieting. There, she expressed the fear that since USPTO had granted tens of thousands of patents on isolated nucleic acid sequences and purified natural products for decades,

> Changing course years after the fact will only serve to punish those companies who made the reasonable decision to invest large amounts of time and money into the identification, isolation, and characterization of genes. Unsettling the expectations of the biotechnology industry now, based on nothing more than unsupported supposition, strikes me as far more likely to impede the progress of science and useful arts than advance it. Given the complicated technology and conflicting incentives at issue here, any change must come from Congress.
>
> (Moore, 2011: p. 28 of her ruling, p. 83 for the pdf document.)

She concluded: "We cannot, after decades of patents and judicial precedent, now call human DNA fruit from the poisonous tree, and punish those inquisitive enough to investigate, isolate, and patent it" (Moore, 2011: pp. 31–32 of her ruling, pp. 85–86 of the pdf.).

The third Federal Judge hearing the appeal was William Curtis Bryson. Although he concurred with the majority ruling that Myriad's cDNA and methods claims are patentable (except for the diagnostic method claim, which all three judges thought was invalid), he was rather adamant that isolated genes are not patent-eligible (Bryson, 2011). He was particularly unconvinced that an isolated human gene differed from a 'native gene.' He feared that the majority ruling would have serious adverse effects on research and treatment. Pace Moore and Lourie, and Bryson drew upon *Chakrabarty* and *Funk Bros.* for defining patentable substances; however, unlike them he felt that although gene extraction might be difficult, isolated genes "are analogous to the 'new mineral discovered in the earth,' or the 'new plant found in the wild' that the Supreme Court referred to in *Chakrabarty*" (Bryson, 2011).

In December 2011, the ACLU appealed the CAFC's ruling to the US Supreme Court, which granted a writ of certiorari on 26 March 2012. The case was remanded back to the CAFC for further consideration in light of *Mayo Collaborative Services* v. *Prometheus Laboratories, Inc.*, 566 U.S. (2012). The opinion was written by Justice Breyer who in *Labcorp v. Metabolite* wrote a dissent to a decision by the Court majority not to take the case. In *Mayo*, the US Supreme Court ruled 9–0 that natural laws may not be patented standing alone or in connection with processes that involve "well-understood, routine, conventional activity." The case involved Prometheus' method claims involving administering a drug to a subject, determining a metabolite concentration in the subject's blood, and inferring the need for a change in dosage based on concentration. The Court opined that the methods were not patent-eligible subject matter since they contributed nothing inventive to the law of nature that formed the basis of their claimed invention.

On 20 July, the CCAF reconsidered the case, and on 16 August, they offered their ruling. The original three judges' opinions remained unchanged, confirming their original decision that genes are patent eligible by a vote of two to one. Lourie, once again representing the majority, argued that isolated DNA molecules have chemically different ends from naturally occurring DNA; therefore, they are patent eligible. An interesting aspect of Judge Lourie's opinion is his attempt to distance himself from the public policy implications of the decision:

> [i]t is important to state what this appeal is not about. It is not about whether individuals suspected of having an increased risk of developing breast cancer are entitled to a second opinion. Nor is it about whether the University of Utah, the owner of the instant patents, or Myriad, the exclusive licensee, has acted improperly in its licensing or enforcement policies with respect to the patents. The question is also not whether is it desirable for one company to hold a patent or license covering a test that may save people's lives, or for other companies to be excluded from the market encompassed by such a patent — that is the basic right provided by a patent, *i.e.*, to exclude others from practicing the patented subject matter. It is also not whether the claims at issue are novel or nonobvious or too broad. Those questions are not before us. It is solely whether the claims to isolated BRCA DNA, to methods for comparing DNA sequences, and to a process for screening potential cancer therapeutics meet the threshold test for patent-eligible subject matter under 35 U.S.C. § 101 in light of various Supreme Court holdings, particularly including *Mayo*. The issue is patent eligibility, not patentability.
>
> We would further note, in the context of discussing what this case is not about, that patents on life-saving material and processes, involving large amounts of risky investment, would seem to be precisely the types of subject matter that should be subject to the incentives of exclusive rights. But disapproving of patents on medical methods and novel biological molecules are policy questions best left to Congress, and other general questions relating to patentability and use of patents are issues not before us.
>
> (Lourie, 2012: pp. 1324–1325)

In light of the *Mayo* decision, Lourie argued against the assumption that Myriad's composition claims are merely reflections of a law of nature.

Pace Lourie and Moore argued that *Funk Brothers* and *Chakrabarty* were more relevant to the case than *Mayo*; however, *Mayo* was informative regarding the scope of the law of nature exception. Using all three cases, Moore reckoned that there were two principles involved, namely, that laws or manifestations of nature are not patentable and whether the patented DNA fragments were (as stated by Bryson in his dissent) compositions of matter with "markedly different characteristics from that found in nature with the potential for significant utility is directed to patentable subject matter" (Bryson, 2012: p. 1340). To the first principle, she reiterated her stance that isolated DNA fragments possess truncations with different ends than naturally occurring DNA, and hence constitute an artificial construction of the human hand. To the second point, she concluded that these fragments do indeed possess markedly different characteristics from their natural homolog, and they do possess utility. Hence, they are patentable subject matter.

Bryson also applied *Mayo* to the case, but came to a different conclusion, stressing that the changes made to the DNA were not inventive, but merely incidental, and thus insufficient to warrant a patent. The majority stressed the uniqueness of patented DNA sequences based on structure, whereas Bryson turned to the informational content, as Judge Sweet did in the District Court ruling. He reasoned that such a move was an appropriate one in the light of *Mayo*, and asserted that DNA's informational content was a critical characteristic. Since the information encoded in the claimed isolated sequences was identical to that found in nature, the sequences were not markedly different from those found in nature (689 F.3d 1355). The ACLU appealed the CAFC's second decision to the Supreme Court, which on 30 November 2012, decided that they would hear the case in the spring of 2013.

On 13 June 2013, the Supreme Court decided by a vote of nine to zero that genes merely excised from the genome were not patent eligible; they were products of nature and not a result of invention. They did, however, rule that cDNA was an artificial construct created in the laboratory and therefore was patentable.

Ever since the 1980s, the USPTO, Congress, and the CAFC have ceded much ground to certain members of the biotech industry by lowering the bar for patent eligibility, utility, and nonobviousness. What makes this case so interesting is that the Biotechnology Industry Organization is split over the issue of gene patenting. While sequencing companies have obviously benefitted from gene patenting, a number of pharmaceutical companies have opposed such IP protection, since they need to pay royalties or licensing fees to patent holders in order to develop diagnostic tests and therapeutic drugs relevant to gene products.

Times might just be changing. Certainly the major patent offices have become more stringent with adjudicating gene patents over the past decade. While the number of gene patents increased considerably during the 1990s, that number has decreased significantly during the first decade of the new millennium (Hopkins et al., 2007). The proportion of gene-patent applications granted by the EPO by September 2005 decreased from 45% of those submitted in the 1980s, to 8% of those submitted between 1996 and 2000. And some commentators conjecture that a number of patents will be abandoned prematurely, as either technical problems prove insurmountable or commercial viability becomes limited. For example, only 70% of the USPTO DNA patents awarded in the early 1990s were still maintained by 2005. Gene-patent filing at all three major patent offices declined sharply after the Human Genome Project made the human genome sequence public in 2001, thereby rendering the novelty criterion problematic. Many of the post-2001 applications are still in the examination phase. Numerous patentees interviewed felt that they were now faced with significant challenges, such as more stringent criteria for demonstrating utility, novelty, nonobviousness, full written description, and unity of invention (i.e., the parts of the invention are intrinsically linked). And some – but certainly not all – commentators argued that the USPTO's Revised Utility Guidelines of 2001 had raised the bar on patent eligibility. In short, the publication of the human genome and stricter requirements placed on gene patent applications have resulted in a decreased number of patent applications and patent awards awarded during the first decade of the century. Speculative, broad utility claims such as the ones granted by the USPTO in the 1990s just might be a thing of the past.

See also: Bioethics; Biotechnology; Courts and Adjudication; European Union Law; Genetic Engineering; Genetic Screening for Disease; Genetics and Society; Genetics, Disease, and Reproduction; Genetics: Legal Aspects; Genomics, Ethical Issues In; Health Politics and Policy; Human Genome Project: History and Assessment; Innovation; Intellectual Property: Law in Context; Pharmaceutical Industry: Political Economies of Drugs and Knowledge; Property Ownership; Science and Law; Science, History of.

Bibliography

Achilladelis, B., 1993. The dynamics of technological innovation: the sector of antibacterial medicines. Research Policy 22, 279–308.

American Wood Paper v. Fibre Disintegrating, 90 U.S. 566 (1874). http://supreme.justia.com/cases/federal/us/90/566/case.html (accessed 18.01.13.).

Andrews, L.B., 2002. Genes and patent policy: rethinking intellectual property rights. Nature Reviews: Genetics 3, 803–807.

Beauchamp, C., 2013. Patenting nature: a problem of history. Stanford Technology Law Review 16, 257–311. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2152105 (accessed 07.09.13.).

Boyle, J., 1996. Shamans, Software & Spleens: Law and the Construction of the Information Society. Harvard University Press, Cambridge, MA.

Bryson, W., 2011. U.S. Court of appeals for the Federal Circuit, The Association for Molecular Pathology, et al., v. the U. S. Patent and Trademark Office and Myriad Genetics, Inc. and et al., decided on 29 July 2011, opinion for the court filed by Circuit Court Bryson. http://www.genomicslawreport.com/wp-content/uploads/2011/07/Decision-in-USPTO-vs-MYGN.pdf (accessed 18.01.13.).

Bryson, W., 2012. Association for Molecular Pathology, et al., v. U.S. P.T.O., et al. 689 F.3d 1303. http://www.leagle.com/xmlResult.aspx?page=12&xmldoc=In%20FCO%2020120816212.xml&docbase=CSLWAR3-2007-CURR&SizeDisp=7 (accessed 18.01.13.).

Bugos, G., Kevles, D., 1992. Plants as Intellectual Property: American Practice, Law, and Policy in World Context. In: 2nd Series, vol. 7. Osiris, 75–104.

Calvert, J., Joly, P.-B., 2011. How did the gene become a chemical compound? The ontology of the gene and the patenting of DNA. Social Science Information 50, 1–21.

Cochrane v. Badische Anilin & Soda Fabrik, 111 U.S. 293 (1884). http://scholar.google.com/scholar_case?case=18061865851672104499&hl=en&as_sdt=2&as_vis=1&oi=scholarr (accessed 18.01.13.).

Conley, J.M., Makowski, R., 2003. Back to the future: rethinking the product of nature doctrine as a barrier to biotechnology patents, parts I and II. Journal of the Patent and Trademark Office Society 85 (301–334), 371–398.

Demaine, L.J., Fellmeth, A.X., 2002. Reinventing the double helix: a novel and nonobvious reconceptualization of the biotechnology patent. Stanford Law Review 55, 303–462.

Doll, J., 1998. The patenting of DNA. Science 280, 689–690.

Dutfield, G., 2009. Intellectual Property Rights and the Life Science Industries: Past, Present and Future, second ed. World Scientific, Hackensack, NJ.

Frankel, A. 2011. U.S. Solicitor General to make unprecedented federal circuit appearance in Myriad case. The American Lawyer (24 February). http://www.law.com/jsp/cc/PubArticleCC.jsp?id=1202483155902&US_Solicitor_General_to_Make_Unprecedented_Federal_Circuit_Appearance_in_Myriad_Case (accessed 18.01.23.).

Gaudillière, J.-P., 2008. Professional or industrial order? Patents, biological drugs, and pharmaceutical capitalism in early twentieth century Germany. History and Technology 24, 107–133.

Haraway, D., 1997. Feminism and Technoscience. Routledge, NY. Modest_Witness@Second_Millennium.FemaleMan_Meets_OncoMouse.

Harkness, J.H., 2011. Dicta on adrenalin(e): Myriad problems with Learned Hand's product-of-nature. Journal of the Patent and Trademark Office Society 93, 363–399.

Hopkins, M., et al., 2007. DNA patenting: the end of an era? Nature Biotechnology 25, 185–187.

Jasanoff, S., 2005. Designs on Nature: Science and Democracy in Europe and the United States. Princeton University Press, Princeton, NJ.

Johnson, G., 1990. Once again: a man with a mission. The New York Times, (25 November). http://www.nytimes.com/1990/11/25/magazine/once-again-a-man-with-a-mission.html?pagewanted=all (accessed 18.01.13.).

Kevles, D., 1994. Ananda M. Chakrabarty wins a patent: Biotechnology, Law, and Society. Historical Studies on the Physical and Biological Sciences 25, 111–135.

Kevles, D., 2002. Of mice and money: the story of the world's first animal patent. Daedalus 131, 78–88.

Kevles, D., 2007. Patents, protections, and privileges: the establishment of intellectual property in animals and plants. Isis 98, 323–331.

Kingston, W., 2004. Removing some harm from the World Trade Organization. Oxford Development Studies 32, 309–320.

Kohler, R., 1994. Lords of the Fly: *Drosophila* Genetics and the Experimental Life. University of Chicago Press, Chicago IL.

Lesch, J., 2007. The First Miracle Drugs: How the Sulpha Drugs Transformed Medicine. Oxford University Press, NY.

Lourie, A., 2011. U.S. Court of appeals for the Federal Circuit, the Association for Molecular Pathology, et al., v. the U. S. Patent and Trademark Office and Myriad Genetics, Inc. and et al., decided on 29 July 2011, opinion for the court filed by Circuit Court Lourie. http://www.genomicslawreport.com/wp-content/uploads/2011/07/Decision-in-USPTO-vs-MYGN.pdf (accessed 18.01.13.).

Lourie, A., 2012. Association for Molecular Pathology, et al., v. U.S. P.T.O., et al. 689 F.3d 1303. http://www.leagle.com/xmlResult.aspx?page=12&xmldoc=In%20FCO%2020120816212.xml&docbase=CSLWAR3-2007-CURR&SizeDisp=7 (accessed 18.01.13.).

McGarrigle, P., Norviel, V., 2008. Laws of nature and the business of biotechnology. Santa Clara Computer and High Technology Law Journal 24, 275–334.

Moore v. Regents of the University of California. 793 P.2d 479 at 490 (Cal. 1990).

Moore, K., 2011. U.S. Court of appeals for the Federal Circuit, the Association for Molecular Pathology, et al., v. the U. S. Patent and Trademark Office and Myriad Genetics, Inc. and et al., decided on 29 July 2011, opinion for the court filed by Circuit Court Moore. http://www.genomicslawreport.com/wp-content/uploads/2011/07/Decision-in-USPTO-vs-MYGN.pdf (accessed 18.01.13.).

Murray, F., 2011. Patenting life: the oncomouse. In: Biagaioli, M., Jaszi, P., Woodmansee, M. (Eds.), Making and Unmaking Intellectual Property: Creative Production in Legal and Cultural Perspective. The University of Chicago Press, Chicago, IL, pp. 399–411.

Palombi, L., 2004. The Patenting of Biological Materials in the Context of TRIPS. Ph.D. dissertation. Australian National University.

Parthasarathy, S., 2007. Building Genetic Medicine: Breast Cancer, Technology, and the Comparative Politics of Health Care. MIT Press, Cambridge, MA.

Pottage, A., 1995. The inscription of life in law: genes, patents, and bio-politics. Modern Law Review 61, 740–765.

Rabinow, P., 1996. Severing the ties: fragmentation and dignity in late modernity. In: Rabinow, P. (Ed.), Essays in the Anthropology of Reason. Princeton University Press, Princeton, NJ, pp. 129–152.

Rajan, K.S., 2006. Biocapital: the Politics of Postgenomic Life. Duke University Press, Rayleigh-Durham, NC.

Shapin, S., 2008. The Scientific Life: a Moral History of a Late Modern Victorian. University of Chicago Press, Chicago, IL.

Skloot, R., 2010. The Immortal Life of Henrietta Lacks. Crown Publishing, NY, New York.

Sweet, R., 2010. United States District Court, Southern District of New York. The Association for Molecular Pathology, et al. Plaintiffs-Appellees v. United States Patent and Trademark Office, Defendant, Myriad Genetics, Inc., et al., Defendant-Appellant. Appeal of case no. 09-CV-4515-RWS, Document 255, Senior Judge Robert W. Sweet, 29 March 2010, case 1:09-cv-04515-RWS Document 255, filed on 29 March 2010. http://www.aclu.org/files/assets/2010-3-29-AMPvUSPTO-Opinion.pdf (accessed 18.01.13.).

Pathological Gambling and Gambling Disorder

David M Ledgerwood, Wayne State University School of Medicine, Detroit, MI, USA
Nancy M Petry, University of Connecticut Health Center, Farmington, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Pathological gambling is a psychiatric disorder that affects 0.4–1.8% of the adult population and is associated with debt, bankruptcy, illegal behaviors, family/social conflict, vocational/educational problems, and psychiatric and substance abuse comorbidity. Pathological gambling has been reclassified in the Substance-Related and Addictive Disorders category of the *Diagnostic and Statistical Manual for Mental Disorders*, Fifth Edition, and has been renamed gambling disorder. This article provides an overview of pathological gambling diagnostic classification, epidemiology, etiology, and treatment.

Pathological gambling (PG), recently renamed gambling disorder (GD), is a psychiatric disorder that has been characterized as an impulse control disorder in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM) until its recent placement as a non-substance-related disorder in the Substance-Related and Addictive Disorders section of the 5th edition (DSM-5, American Psychiatric Association, 2013). Briefly, it is characterized by "persistent and recurrent maladaptive gambling behavior that disrupts personal, family, or vocational pursuits." GD is associated with many debilitating problems including financial indebtedness, bankruptcy, legal and vocational difficulties, and interpersonal conflict (Petry and Armentano, 1999). Individuals with GD also are more likely to experience several other co-occurring psychiatric and substance abuse conditions including mood, anxiety, and personality disorders; heightened suicide risk; and alcohol and drug abuse (Petry et al., 2005). In this article, GD symptoms, classification, epidemiology, etiology, and treatments are reviewed.

Classification of Pathological Gambling and Reclassification for DSM-5

PG was first introduced as a Disorder of Impulse Control, Not Elsewhere Classified in the DSM-III, and the disorder remained in the section for impulse control disorders in the DSM-IV. Based on increasing research evidence revealing that PG may have several characteristics in common with substance use disorders, however, PG has been reclassified for the DSM-5 alongside alcohol and other substance use disorders in the section titled Substance-Related and Addictive Disorders (APA, 2013).

As of May 2013 with the publication of the DSM-5, PG has been renamed gambling disorder. According to the DSM-5, to be diagnosed with GD, an individual must meet at least four of nine symptoms. These symptoms include (1) preoccupation with gambling or ways to get money with which to gamble; (2) needing to gamble with increasing amounts of money to experience excitement; (3) restlessness or irritability when not gambling; (4) repeated unsuccessful attempts to cut down or stop gambling; (5) using gambling as a way to escape from problems or relieve painful affect; (6) chasing gambling losses; (7) lying to close others to conceal the extent of gambling involvement; (8) jeopardizing important relationships, jobs, or educational opportunities because of gambling; and (9) relying on others for financial bailouts from desperate financial situations caused by gambling. One symptom, engaging in illegal activities to support gambling, was dropped from the DSM with the 5th edition, and the number of symptoms needed to meet diagnostic criteria was decreased from five to four. The gambling symptoms must not exclusively occur during a manic episode. For the remainder of this article, the authors primarily use the term PG because nearly all of the current research literature is based upon this older formulation of GD.

Epidemiology

Prevalence rates of PG have been established in several general population surveys conducted in many countries. The National Gambling Impact Study Commission (Gerstein et al., 1999) conducted a survey of 2417 randomly selected US adult residents. In this study, lifetime rate of PG was 0.8% and lifetime rate of problem gambling (subthreshold PG reflected in meeting three or four PG symptoms) was 1.3%. Past-year rates of PG and problem gambling were lower at 0.1 and 0.4%, respectively. Another national telephone survey around that same time of 2638 randomly selected US adults found somewhat higher lifetime prevalence rates of 2.0 and 2.8% for PG and problem gambling, respectively (Welte et al., 2001).

Two additional US surveys utilized data from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) and the National Comorbidity Survey Replication (NCS-R). The largest survey to date, the NESARC, included in-person interviews of over 43 000 adults. A lifetime PG rate of 0.4% was established in this survey (Petry et al., 2005). The NCS-R survey found a similar lifetime PG rate of 0.6% among 9282 adult respondents (Kessler et al., 2008). This survey also reported higher lifetime problem gambling rates of 2.3%.

Other countries with legalized gambling have similar prevalence rates to the United States. One recent study compared prevalence rates from several countries, including Australia, Canada, China, Italy, New Zealand, Norway, Singapore, Sweden, Switzerland, United Kingdom and the United States (Stucki and Rihs-Middel, 2007). Weighted mean prevalence rates for both pathological and problem gambling were similar to the above studies from the United States, at 0.8–1.8% for PG and 1.2–2.4% for problem gambling. Thus, prevalence rates tend to be similar across countries with legalized gambling.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.21007-6

Prevalence rates also tend to be higher for adolescents and young adults than they are for adults. One large-scale telephone survey that examined both populations in the United States found that the prevalence of problem gambling among respondents age 14–21 years was 2.1% (Welte et al., 2008).

Etiology

Demographic Correlates

Several demographic factors appear to be associated with a higher risk of PG (see Petry, 2005 for a review). Age, for example, appears to be inversely related to gambling problems, as indicated by prevalence studies cited above that find higher rates of PG among younger individuals. Gender is also a contributing factor, as men have traditionally been at substantially higher risk for developing PG than women, although there is also some evidence that the gap between men and women has been narrowing with the proliferation of legal gambling venues.

Other demographic factors that may incur some increased risk for PG include race, socioeconomic status, and marital status. Across countries with legalized gambling, PG tends to be more prevalent among racial and ethnic minorities. In the United States, for example, the current literature consistently shows that PG is more prevalent among African Americans, Hispanics, and Native Americans than it is among European Americans (e.g., Kessler et al., 2008; Welte et al., 2001). Across countries with legal gambling and majority populations of European heritage, higher rates of PG have been found among all nonwhite groups in the population, including those of Asian, Hispanic, Pacific Islander, and Indigenous background (see Petry, 2005 for review).

Lower socioeconomic status has also consistently been found to be associated with PG (Petry, 2005). The effect of socioeconomic status on the rates of PG is somewhat difficult to interpret because it may be confounded by other factors that may also correlate with poverty such as nonmajority racial/ethnic status or co-occurring psychopathology. However, most studies show that lower income earners are overrepresented among pathological gamblers, and there are some studies that show lower education and disability status to be associated with higher rates of PG (e.g., Morasco and Petry, 2006; Welte et al., 2001).

Marital status also appears to be associated with PG status (Petry, 2005). Specifically, studies have often found that pathological gamblers are more likely to be divorced, separated, and/or never married relative to nonproblem gamblers. There are several possible reasons for the relationship between PG and marital status (e.g., being single may place one at greater risk of PG because there are fewer external controls over gambling behavior; pathological gamblers may be less able to form stable relationships because of underlying psychopathology). However, the etiological role of being single is not fully understood.

Psychiatric Comorbidity

Psychiatric and substance use disorders are highly comorbid with PG. In the United States, population surveys have demonstrated that individuals with PG have very high rates of mood, anxiety, substance use, and personality disorders (Petry et al., 2005). A recent meta-analysis examined comorbidity across 11 population-based studies of problem and/or pathological gamblers conducted in Canada, United States, Korea, and Switzerland (Lorains et al., 2011). Although there was significant heterogeneity across studies, this meta-analysis revealed that globally pathological gamblers experience high rates of nicotine dependence (60.1%), substance use disorder (57.5%), mood disorder (37.9%), and anxiety disorder (37.4%).

Etiological models of PG suggest that there may be different subtypes of PG, and that the subtype may depend in part on the extent to which co-occurring psychiatric conditions contributed to the development of the disorder. The most frequently discussed model is the Pathways model, which presumes that the etiology of PG takes one of three primary courses: (1) behaviorally conditioned gambling in which the pathological gambler develops problems primarily as a result of the contingencies associated with the games themselves, rather than any underlying psychopathology; (2) escape gambling where the PG develops primarily as a maladaptive coping strategy to alleviate painful affect associated with depression and/or anxiety disorders; or (3) antisocial-impulsive gambling, which includes elements of behaviorally conditioned and escape gambling, and is also characterized by greater underlying impulse control problems that often manifest themselves in the form of illegal activities, antisocial personality, and impulse control problems (Blaszczynski and Nower, 2002). For the most part, studies support the possibility that these three underlying subtypes of PG may exist (see Milosevic and Ledgerwood, 2010 for review). However, the question of whether underlying depression, anxiety, or antisocial disorders precede the development of PG, or impact its treatment, is not yet settled.

Impulsivity and Neurobiological Factors

Impulsivity is a personality construct that is often broadly defined as a tendency to act without forethought. Impulsivity has been operationally defined in many ways including as inattention, response disinhibition, rate of discounting of delayed rewards (in favor of smaller, immediate rewards), risk-taking, sensation seeking, and others (Evenden, 1999). It is a trait that is almost universally found to be elevated in PG (see Van Holst et al., 2010 for review). Indeed, some results show that impulsivity in childhood may predict later problem gambling (Pagani et al., 2009). Further, neurodevelopmental changes that occur during adolescence in pathways associated with reward and impulsivity may place some at greater risk for gambling later in life (see Chambers and Potenza, 2003 for review). These recurring findings regarding the relationship between gambling and impulsivity have led to recent attempts to understand the neurobiological underpinnings of impulse control problems experienced by pathological gamblers.

Neurobiological studies of pathological gamblers have found similarities with individuals who have substance use disorders in frontal and limbic/striatal system functioning (e.g., Reuter et al., 2005). Imaging research with substance-abusing individuals has revealed deficits in activation of the orbitofrontal cortex and dorsolateral prefrontal cortex (Bolla et al., 2005). Functional magnetic resonance imaging studies of pathological gamblers and healthy controls reveal that

pathological gamblers experience deficits in functioning in frontal areas of the brain associated with executive function and decision making (e.g., orbitofrontal cortex, caudate/basal ganglia, thalamus, ventromedial prefrontal cortex, and ventral striatum; e.g., Potenza et al., 2003). Exposure to gambling cues may also be associated with increased activity in executive function areas of the brain such as the dorsolateral prefrontal cortex in pathological gamblers relative to controls (Crockford et al., 2005). Furthermore, other studies have found significant differences between pathological gamblers and controls in function in the ventral striatum and limbic system structures (Miedl et al., 2010; Reuter et al., 2005). Indeed, Reuter et al. (2005) found reduced activation in the ventral striatum and ventromedial prefrontal cortex among pathological gamblers during a gambling task, suggesting hypoactivation of mesolimbic reward pathways relative to healthy controls.

Nonimaging studies have also found reason to suspect neurological deficits among pathological gamblers. For example, it appears that problem gamblers exhibit differential neurotransmitter activation relative to healthy controls. Moreno et al. (1991) demonstrated that pathological gamblers show blunted prolactin response indicating 5-hydroxytryptamine (5-HT) hypoactivity. Similar findings of differential functioning between problem gamblers and controls have also been found in studies of the dopaminergic system (Bergh et al., 1997). Research has also demonstrated that variants of the DRD2 dopamine receptor gene may also be a risk factor for the development of GDs (Commings et al., 1996).

There is further evidence that pathological gamblers experience impaired decision making and planning and cognitive flexibility deficits indicative of impaired executive function (Goudrian et al., 2006; Ledgerwood et al., 2012). Goudriaan et al. (2006) also reported an absence of physiological arousal (e.g., increased heart rate) in pathological gamblers exposed to win scenarios on the Iowa Gambling Task. Taken together, these neurobiological findings point to likely deficits in executive function and decreased reward sensitivity among pathological gamblers that serve to maintain problematic gambling behavior. For many pathological gamblers, these deficits may also contribute to the development of the disorder, but longitudinal research is needed to establish whether these deficits precede problem gambling, or develop in response to excessive gambling.

Behavioral Treatment

Behavioral treatments for PG include fellowships, such as Gamblers Anonymous (GA), and evidence-based psychotherapies such as cognitive behavioral therapy (CBT) and motivational interviewing. Below, the authors describe these therapies and provide some of the evidence for their effectiveness. For thorough reviews of the effectiveness of behavioral treatments for PG, the reader is referred to recent reviews (e.g., Gooding and Tarrier, 2009).

Gamblers Anonymous

GA is a mutual-help fellowship based on a 12-step model that is also the basis of other addiction fellowships including Alcoholics Anonymous and Narcotics Anonymous. Established in 1957, GA is the most frequently utilized intervention for PG in the United States, and chapters exist in many other countries (Petry, 2003).

Because the GA model is based on anonymity and mutual support rather than a therapist-driven treatment approach, study of the efficacy of GA is difficult. Consequently, there are few studies that examine the effectiveness of GA, but the few that do exist provide important information about the usefulness of this intervention. An early Scottish study followed individuals ($N = 232$) who attended at least one GA meeting to determine the extent to which they remained involved with the organization and engaged in recovery behaviors (Stewart and Brown, 1988). The study revealed that 22% of participants attended only one meeting, and 70% had stopped attending GA by their 10th meeting. Moreover, only 8% were abstinent from gambling a year after starting in GA and 7% were abstinent after 2 years.

Another study found more positive outcomes for GA, particularly when it is combined with professional treatment. This clinical trial, which included 342 pathological gamblers, reported that 48% of patients who attended treatment and GA concurrently abstained from gambling for at least 2 months, compared with 36% of those who only attended professional treatment (Petry, 2003). Thus, there may be some benefits of participating in GA for gambling abstinence. However, without randomized trials, one cannot rule out self-selection biases (e.g., greater motivation for treatment) as potential confounding factors that account for the apparent benefits of GA.

Cognitive and Cognitive-Behavioral Therapies

The most frequently studied treatments for PG include cognitive and CBT approaches. These treatments, consequently, also have the most empirical support for their efficacy. Cognitive and CBT approaches differ in that cognitive treatments focus primarily on gambling-related cognitive distortions such as biased memories about wins and losses, and illusion of control over gambling outcomes. CBT approaches, on the other hand, focus somewhat less on distortions, and place more emphasis on behavioral components including reinforcement of non-gambling behaviors, encouragement of problem solving, social skills building, and relapse prevention. In practice, the differences between cognitive and CBT treatments tend to be a matter of degree of emphasis rather than philosophical differences between these approaches. Thus, in this article, the authors refer to these approaches as 'CBT' regardless of whether the emphasis is more on cognitive or behavioral components.

Early studies by Ladouceur et al. revealed evidence for the efficacy of CBT approaches. In one study, participants ($N = 40$) were randomly assigned to twice-weekly (total 30 h) CBT or a wait-list control condition (Sylvain et al., 1997). Although the study experienced some retention issues ($n = 29$ were assessed at follow-up), 8 of 14 gamblers treated with CBT reported significant reductions in gambling severity compared with only 1 of 15 control participants.

Other research groups have similarly found beneficial effects of CBT for PG, and have also addressed some of the retention issues found in earlier studies. In one, 231 pathological gamblers were randomly assigned to one of three treatment

groups: referral to GA; GA referral plus a self-directed CBT manual; or GA referral plus eight sessions of individual CBT (Petry et al., 2006). CBT-treated participants exhibited better outcomes (reduced measures of problem gambling severity, frequency of gambling, and amount of money gambled) than did participants who only received a GA referral.

A number of other studies have found similarly positive outcomes of CBT for PG. Several thorough reviews have chronicled the efficacy of CBT approaches for PG. One meta-analysis examined published CBT clinical trials and found large effects for treatment outcome (e.g., abstinence from gambling) during treatment, as well as at 6-, 12-, and 24-month follow-up assessments (Goodling and Tarrier, 2009). Thus, the evidence for the efficacy of CBT for treating PG appears strong.

Motivational Enhancement and Brief Interventions

Brief interventions may be appropriate with problem and pathological gamblers who have less severe gambling problems, or who would otherwise not attend treatment. These interventions have often taken the form of a small number of therapy sessions (often 1–4) using brief advice about behavior change, or motivational enhancement therapy alone or in combination with CBT. In some cases, self-help manuals are employed as an adjunct to therapy.

In one study, problem gamblers ($N = 102$) were randomly assigned to one of three treatment conditions: (1) self-help workbook alone, (2) self-help workbook plus one motivational interview, or (3) wait-list (Hodgins et al., 2001). The workbook plus motivational interview treatment resulted in the greatest improvement in gambling symptoms. The workbook alone did not result in improvements in gambling symptoms compared to the waitlist condition. Another more recent study compared problem gamblers ($N = 314$) randomly assigned to receive (1) brief telephone motivational interview plus a self-help workbook, (2) brief motivational interview and workbook plus six booster telephone calls over a 9-month period, (3) workbook only, or (4) wait-list assignment (Hodgins et al., 2009). The two brief motivational interview conditions reported less gambling at 6 weeks and 6 months than those assigned to the workbook only and waitlist conditions. Booster calls did not appear to improve long-term outcomes.

Two other published brief intervention studies focused on motivational enhancement, CBT, and brief advice approaches (Petry et al., 2008, 2009). Both studies used a similar design, randomly assigning participants to one of four treatment conditions: (1) assessment only control, (2) 10 min of brief advice (to reduce gambling), (3) one 50-min session of motivational enhancement therapy (MET; based on the principles of motivational interviewing), or (4) one 50-min session of MET and three sessions of CBT. One study recruited problem gamblers from the community ($N = 180$) and found that brief advice resulted in significant reductions in gambling at the 6-week point and at 9 months. MET plus CBT significantly reduced gambling severity on one outcome measure (Addiction Severity Index Gambling score) between the 6-week and 9-month follow-up assessments, but MET alone did not result in significant changes compared with assessment only (Petry et al., 2008). The second study recruited problem gambling college students ($N = 117$) and revealed that the MET condition showed significant decreases in problem gambling severity over time compared with the assessment-only condition (Petry et al., 2009). Long-term benefits of MET for reducing gambling were also apparent at a 9-month follow-up. Brief advice and MET plus CBT also resulted in some positive gambling outcomes compared with assessment only. Taken together, these studies show that brief interventions can be effective in reducing gambling problems among gamblers for whom more intensive treatments are either not available or not wanted.

Medication Treatments

The most frequently studied pharmacological treatments for PG have been from three categories: (1) opioid antagonists, (2) antidepressants, and (3) mood stabilizers. To date, there are no medications that are approved by the American Food and Drug Administration (FDA) for treatment of PG. Thus, all medications prescribed for PG in the United States must be prescribed as off-label. Some of the research evidence for the efficacy of these medications is discussed below. The reader is referred to recent reviews for a more thorough summary of the cumulative evidence for medication efficacy (e.g., Pallesen et al., 2007).

Opioid Antagonists

Medications such as naltrexone and nalmefene hydrochloride are full opioid antagonists, and are typically used to treat opioid and alcohol dependence. There is evidence that addictive behaviors, such as gambling, may be associated with dysregulation of the mesolimbic dopamine system. Furthermore, endogenous opioids may serve to modulate dopamine release from the nucleus accumbens (Nacc). Modulation of the dopamine system may serve to reduce the excitement and urges associated with gambling for individuals with PG (Kim et al., 2001).

Several studies have examined the efficacy of opioid antagonists (most frequently naltrexone), and most have found that these medications reduce gambling symptoms and urges (see Pallsen et al., 2007). In one example of a double-blind, controlled study, participants with PG ($N = 83$) were randomly assigned to receive naltrexone (titrated to 250 mg) or placebo (Kim et al., 2001). Data analyzed from 45 participants who continued in the study past week six revealed that 75% of those who received naltrexone were 'much improved' or 'very much improved,' compared with only 24% of placebo-treated patients. Other double-blind trials have found similar results (e.g., Grant et al., 2008).

Antidepressant Medications

Selective serotonin reuptake inhibitors (SSRIs) and norepinephrine/dopamine reuptake inhibitors (NDRIs) that are typically used as treatments for major depression have been studied with gambling disordered individuals, in part because of the high rates of comorbidity between PG and depression, and also because of findings that suggest hypoactivation of 5-HT among problem gamblers (Moreno et al., 1991).

One meta-analysis of pharmacological interventions for PG found medications (opioid antagonists, antidepressants, and mood stabilizers) were generally effective at reducing problem gambling symptoms, and there were no differences in the effect sizes found for different medications (Pallesen et al., 2007). However, examination of individual controlled double-blind studies suggests that the effectiveness of antidepressant medications may be mediated by other factors including the specific medication used and characteristics of the patient population. For example, one double-blind study testing the effectiveness of paroxetine (an SSRI) among pathological gamblers ($N = 45$) revealed that it led to greater reductions in GD severity and gambling urges than did placebo (Kim et al., 2002). Another double-blind trial testing fluvoxamine (also an SSRI) against placebo ($N = 32$) showed that fluvoxamine did not result in significant benefits (Blanco et al., 2002). Similarly, a placebo-controlled, double-blind study of bupropion (an NDRI) for PG ($N = 39$) failed to find any significant improvements in GD symptoms relative to the control condition (Black et al., 2007). Thus, more study of antidepressants is needed to determine their efficacy in reducing gambling symptoms.

Other Medications

Mood stabilizers such as lithium have primarily been studied for treating PG among gamblers with concurrent bipolar disorder. A double-blind, placebo-controlled study of lithium carbonate for gambling among individuals with co-occurring PG and bipolar spectrum disorder ($N = 40$) revealed that lithium was associated with reduction in problem gambling symptoms, thoughts, and urges relative to placebo (Hollander et al., 2005) suggesting its potential efficacy in this population.

Glutamate is a neurotransmitter that has been linked to reward processes, and there is evidence that glutamate homeostatis imbalance may affect changes in corticostriatal neuroplasticity, which may impair communication between brain regions that are essential for processing rewards (e.g., Nacc and prefrontal cortex). Thus, there is some interest in pursuing medications that modulate glutamate in the brain. In the only open-label study conducted to date, 27 pathological gamblers were treated with *N*-acetyl cysteine (NAC), an amino acid that may enhance extracellular glutamate levels in the Nacc (Grant et al., 2007). In total, 60% of those treated with NAC were considered treatment responders. More rigorous, double-blind studies are needed to assess whether NAC is efficacious for treatment of PG, as well as whether other medications may be useful.

Non-Treatment-Assisted Recovery

Although there are several treatments that are efficacious for reducing PG severity, it is important to also acknowledge that the majority of individuals who recover from this disorder do so without the assistance of any intervention (Slutske, 2006). The largest study identified 206 lifetime pathological gamblers from two large household surveys and found that 36–39% (depending on the survey) reported having no past-year gambling problems despite the fact that only 7–10% of those surveyed had ever sought treatment (Slutske, 2006).

Summary

PG is a psychiatric disorder that affects 0.4–1.8% of the adult population across countries with legalized gambling, and frequently results in serious negative consequences including debt, bankruptcy, illegal behaviors, family and social conflict, and vocational/educational problems. PG is also associated with high rates of psychiatric and substance abuse comorbidities including relatively high rates of mood, anxiety, alcohol, drug use, and personality disorders, as well as elevated risk of suicidal ideation and attempt. PG was included in the DSM-IV as an Impulse Control Disorder Not Elsewhere Classified, but is now (as of May 2013) classified as an addictive disorder along with substance use disorders in the DSM-5, and has been renamed Gambling Disorder. Demographic factors, co-occurring psychopathology, and neurological vulnerabilities have been studied as potential risk factors of PG. GA is the most frequently utilized behavioral treatment for PG, but research evidence is the most abundant for CBT and motivational interviewing approaches at this time. Several medications have been examined in the treatment of gambling, but no medication has been approved by the FDA for this indication. Future research will be needed to ascertain the optimal treatment approaches for individuals suffering from gambling problems, but there is excellent support for some psychosocial interventions and growing support for some pharmacological ones.

See also: Behavior Therapy: Background, Basic Principles, and Early History; Concepts and Methods of Cognitive Therapies; Control Behavior: Psychological Perspectives; Counseling Psychology; Defiant Behavior During Adolescence across Cultures; Evidence-Based Psychotherapy; Exposure Therapies and Stress Inoculation: A Brief Overview; Forensic Psychology; Personality and Risk Taking; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychological Treatments: Randomized Controlled Trials; Relapse Prevention Training; Therapist–Patient Relationship.

Bibliography

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Press, Arlington, VA.

Bergh, C., Eklund, T., Sodersten, P., Nordin, C., 1997. Altered dopamine function in pathological gambling. Psychological Medicine 27, 473–475.

Black, D.W., Arndt, S., Coryell, W.H., Argo, T., Forbush, K.T., Shaw, M.C., Perry, P., Allen, J., 2007. Bupropion in treatment of pathological gambling: a randomized, double-blind, placebo-controlled, flexible-dose study. Journal of Clinical Psychopharmacology 27, 143–150.

Blanco, C., Petkova, E., Ibanez, A., Saiz-Ruiz, J., 2002. A pilot placebo-controlled study of fluvoxamine for pathological gambling. Annals of Clinical Psychiatry 14, 9–15.

Blaszczynski, A., Nower, L., 2002. A pathways model of problem and pathological gambling. Addiction 97, 487–499.

Bolla, K.I., Eldreth, D.A., Matochik, J.A., Cadet, J.L., 2005. Neural substrates of faulty decision-making in abstinent marijuana users. Neuroimage 26, 480–492.

Chambers, R.A., Potenza, M.N., 2003. Neurodevelopment, impulsivity, and adolescent gambling. Journal of Gambling Studies 19, 53–84.

Comings, D.E., Rosenthal, R.J., Lesieur, H.R., Rugle, L.J., Muhleman, D., Chiu, C., Dietz, G., Gade, R., 1996. A study of the dopamine D2 receptor gene in pathological gambling. Pharmacogenetics and Genomics 6, 223–234.

Crockford, D.N., Goodyear, B., Edwards, J., Quickfall, J., el-Guebaly, N., 2005. Cue-induced brain activity in pathological gamblers. Biological Psychiatry 58, 787–795.

Evenden, J.L., 1999. Varieties of impulsivity. Psychopharmacology 146, 348–361.

Gerstein, D., Volberg, R.A., Toce, M.T., Harwood, H., Johnson, R.A., Buie, T., et al., 1999. Gambling Impact and Behavior Study: Report to the National Gambling Impact Study Commission. National Gambling Impact Study Commission, Chicago, IL.

Gooding, P., Tarrier, N., 2009. A systematic review and meta-analysis of cognitive-behavioural interventions to reduce problem gambling: hedging our bets? Behaviour Research and Therapy 47, 592–607.

Goudriaan, A.E., Oosterlaan, J., DeBeurs, E., Van Den Brink, W., 2006. Psychophysiological determinants and concomitants of deficient decision making in pathological gamblers. Drug and Alcohol Dependence 84, 231–239.

Grant, J.E., Kim, S.W., Odlaug, B.L., 2007. N-acetyl cysteine, a glutamate-modulating agent in the treatment of pathological gambling: a pilot study. Biological Psychiatry 62, 652–657.

Grant, J.E., Kim, S.W., Hartman, B.K., 2008. A double-blind, placebo-controlled study of the opiate antagonist naltrexone in the treatment of pathological gambling urges. Journal of Clinical Psychiatry 69, 783–789.

Hodgins, D.C., Currie, S., el-Guebaly, N., 2001. Motivational enhancement and self-help treatments for problem gambling. Journal of Consulting and Clinical Psychology 69, 50–57.

Hodgins, D.C., Currie, S.R., Currie, G., Fick, G.H., 2009. Randomized trial of brief motivational treatments for pathological gamblers: more is not necessarily better. Journal of Consulting and Clinical Psychology 77, 950–960.

Hollander, E., Pallanti, S., Allen, A., Sood, E., Rossi, N.B., 2005. Does sustained-release lithium reduce impulsive gambling and affective instability versus placebo in pathological gamblers with bipolar spectrum disorders? American Journal of Psychiatry 162, 137–145.

Kessler, R.C., Hwang, I., LaBrie, R., Petukhova, M., Sampson, N.A., Winters, K.C., 2008. DSM-IV pathological gambling in the National Comorbidity Survey Replication. Psychological Medicine 38, 1351–1360.

Kim, S.W., Grant, J.E., Adson, D.E., Shin, Y.C., 2001. Double-blind naltrexone and placebo comparison study in the treatment of pathological gambling. Biological Psychiatry 49, 914–921.

Kim, S.W., Grant, J.E., Adson, D.E., Shin, Y.C., Zaninelli, R., 2002. A double-blind placebo-controlled study of the efficacy and safety of paroxetine in the treatment of pathological gambling. Journal of Clinical Psychiatry 63, 501–507.

Ledgerwood, D.M., Kaploun, K., Milosevic, A., Orr, E., Frisch, G.R., Rupcich, N., Lundahl, L.H., 2012. Executive function deficits in pathological gamblers compared with healthy non-problem gambling controls. Journal of Gambling Studies 28, 89–103.

Lorains, F.K., Cowlishaw, S., Thomas, S.A., 2011. Prevalence of comorbid disorders in problem and pathological gambling: systematic review and meta-analysis of population studies. Addiction 106, 490–498.

Miedl, S.F., Fehr, T., Meyer, G., Herrmann, M., 2010. Neurobiological correlates of problem gambling in a quasi-realistic blackjack scenario as revealed by fMRI. Psychiatry Research Neuroimaging 181, 165–173.

Milosevic, A., Ledgerwood, D.M., 2010. The subtyping of pathological gambling: a comprehensive review. Clinical Psychological Review 30, 988–998.

Morasco, B.J., Petry, N.M., 2006. Gambling problems and health functioning in individuals receiving disability. Disability and Rehabilitation 28, 619–623.

Moreno, I., Saiz-Ruiz, J., Lopez-Ibor, J.J., 1991. Serotonin and gambling dependence. Human Psychopharmacology: Clinical and Experimental 6, S9–S12.

Pagani, L.S., Derevenski, J.L., Japel, C., 2009. Predicting gambling behavior in sixth grade from kindergarten impulsivity. Archives of Pediatrics and Adolescent Medicine 163, 238–243.

Pallesen, S., Molde, H., Arnestad, H.M., Laberg, J.C., Skutle, A., Iversen, E., Stoylen, I.J., Kvale, G., Holsten, F., 2007. Outcome of pharmacological treatments of pathological gambling: a review and meta-analysis. Journal of Clinical Psychopharmacology 27, 357–364.

Petry, N.M., 2003. Patterns and correlates of gamblers anonymous attendance in pathological gamblers seeking professional treatment. Addictive Behaviors 28, 1049–1062.

Petry, N.M., 2005. Pathological Gambling: Etiology, Comorbidity, and Treatment. American Psychological Association, Washington, DC.

Petry, N.M., Armentano, C., 1999. Prevalence, assessment, and treatment of pathological gambling: a review. Psychiatric Services 50, 1021–1027.

Petry, N.M., Stinson, F.S., Grant, B.F., 2005. Comorbidity of DSM-IV pathological gambling and other psychiatric disorders: results from the National Epidemiological Survey on alcohol related conditions. Journal of Clinical Psychiatry 66, 564–574.

Petry, N.M., Ammerman, Y., Bohl, J., Doersch, A., Gay, H., Kadden, R., et al., 2006. Cognitive-behavioral therapy for pathological gamblers. Journal of Consulting and Clinical Psychology 74, 555–567.

Petry, N.M., Weinstock, J., Ledgerwood, D.M., Morasco, B., 2008. Brief interventions for problem gamblers. Manuscript submitted for publication. Journal of Consulting and Clinical Psychology 76, 318–328.

Petry, N.M., Weinstock, J., Morasco, B.M., Ledgerwood, D.M., 2009. A randomized trial of brief interventions for problem and pathological gambling college students. Addiction 104, 1569–1578.

Potenza, M.N., Leung, H.-C., Blumberg, H.P., Peterson, B.S., Fulbright, R.K., Lacadie, C.M., Skudlarski, P., Gore, J.C., 2003. An fMRI stroop task study of ventromedial prefrontal cortical function in pathological gamblers. American Journal of Psychiatry 160, 1990–1994.

Reuter, J., Raedler, T., Rose, M., Hand, I., Gläscher, J., Büchel, C., 2005. Pathological gambling is linked to reduced activation of the mesolimbic reward system. Nature Neuroscience 8, 147–148.

Slutske, W.S., 2006. Natural recovery and treatment-seeking in pathological gambling: results from two U.S. national surveys. American Journal of Psychiatry 163, 297–302.

Stewart, R.M., Brown, R.I., 1988. An outcome study of Gamblers Anonymous. British Journal of Psychiatry 152, 284–288.

Stucki, S., Rihs-Middel, M., 2007. Prevalence of adult problem and pathological gambling between 2000 and 2005: an update. Journal of Gambling Studies 23, 245–257.

Sylvain, C., Ladouceur, R., Boisvert, J.M., 1997. Cognitive and behavioral treatment of pathological gambling: a controlled study. Journal of Consulting and Clinical Psychology 65, 727–732.

Van Holst, R.J., van den Brink, W., Veltman, D.J., Goudriaan, A.E., 2010. Why gamblers fail to win: a review of cognitive and neuroimaging findings in pathological gambling. Neuroscience and Biobehavioral Reviews 34, 87–107.

Welte, J., Barnes, G., Wieczorek, W., Tidwell, M.C., Parker, J., 2001. Alcohol and gambling pathology among U.S. adults: prevalence, demographic patterns and comorbidity. Journal of Studies on Alcohol and Drugs 62, 706–712.

Welte, J.W., Barnes, G.M., Tidwell, M.-C.O., Hoffman, J.H., 2008. The prevalence of problem gambling among U.S. adolescents and young adults: results from a national survey. Journal of Gambling Studies 24, 119–133.

Patient Adherence to Health Care Regimens

Tricia A Miller and M Robin DiMatteo, University of California, Riverside, Riverside, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.R. DiMatteo, volume 16, pp. 11111–11115, © 2001, Elsevier Ltd.

Abstract

A significant barrier to effective medical treatment involves patients' failure to follow treatment recommendations given by their health care provider, a term known as nonadherence (or noncompliance). More than 40% of patients take medications incorrectly, and almost 80% fail to follow dietary or exercise recommendations; some continue to engage in compromising health behaviors (e.g., smoking and alcohol abuse). Improving adherence depends upon effective communication between patients and providers, trust in the therapeutic relationship, and a realistic assessment of patients' knowledge and understanding of treatment. Such collaboration fosters improvements in health outcomes and reduces nonadherence.

Introduction

Nonadherence to prescribed medications and lifestyle treatments is an ever-present and complex problem that imposes a considerable economic burden upon modern health care systems (DiMatteo et al., 2012). Yearly U.S. expenditures for the consequences of patient nonadherence are estimated to be between 290 and 300 billion dollars (DiMatteo, 2004a; New England Healthcare Institute, 2009). Nonadherence can take many forms: treatment advice given to patients can be misunderstood, carried out incorrectly, forgotten, or completely ignored (Martin et al., 2005). Rates of treatment adherence vary widely across different diseases and treatment regimens, and can be extremely low even for treatments that are highly effective (DiMatteo et al., 2002; DiMatteo, 2004a; DiMatteo et al., 2007). For some long-term chronic medical conditions, adherence can be worse than 50% (DiMatteo, 2004a; Haskard-Zolnierek and DiMatteo, 2009). Moreover, nonadherence has been strongly associated with poor health outcomes, including more than 125,000 deaths each year (Christensen and Ehlers, 2002; Martin et al., 2005).

The medical literature on patient adherence is extensive, and continuous efforts to improve the relationship between adherence and treatment outcomes remain an ongoing challenge. Current research on the physician–patient relationship suggests that effective communication can enhance patient adherence through various mechanisms, including collaborative partnerships, provider empathy, and patient-centered communication (Haskard-Zolnierek and DiMatteo, 2009).

Patient nonadherence can be intentional or unintentional. 'Unintentional nonadherence' is used to describe situations in which patients incorrectly believe they are adherent, or try to be adherent but fail, whereas 'intentional nonadherence' describes the purposeful choice to dismiss treatment or change the regimen without consulting the provider (DiMatteo et al., 2012). Whether purposeful or accidental, nonadherence contributes to suboptimal health outcomes.

Adherence Terminology

Historically the term 'compliance' was first used to describe the behavior of following physicians' recommendations (Lutfey and Wishner, 1999; Vermeire et al., 2001; McNicholas, 2012). Recent studies have argued, however, that 'compliance' suggests an 'authoritative provider to submissive patient' model of care (Lutfey and Wishner, 1999). 'Adherence' has been proposed as a more accurate term that implies a physician–patient relationship based on collaborative partnerships and mutual respect (McNicholas, 2012). The shift in terminology reflects a fundamental change in the way health professionals view the process and delivery of care. Such change enables physicians to more accurately identify potential patient constraints (e.g., social and economic), and provides patients with more autonomy in managing their treatments (Lutfey and Wishner, 1999). The term 'adherence' further characterizes patients as independent individuals who pursue active and voluntary roles in efforts to achieve specific health goals.

Factors that Affect Patient Adherence

Adherence to medical treatment is a multifaceted process by which many factors influence the degree to which patients engage in disease management behaviors. Such factors include doubt about the expected benefits and efficacy of treatment (including side effects and perceived severity), and the unique demands of the treatment regimen itself (e.g., complex medication schedules) (DiMatteo et al., 2002; DiMatteo, 2004b; Martin et al., 2005). Socioeconomic factors, financial constraints (e.g., cost of medications), and patient characteristics can also negatively affect adherence. DiMatteo (2004b) found that the lack of social support strongly predicted patient nonadherence, in that adherence rates were 1.74 times higher in patients from cohesive families and 1.53 times lower in patients from families in conflict (DiMatteo, 2004b). Social support can benefit patients by helping them adjust and live with illness by buffering stress, influencing health behaviors, and changing patients' affective state (DiMatteo, 2004b).

In the following subsections, health literacy, the interpersonal dynamics of the physician–patient relationship, and depression will be discussed as major factors that affect patient adherence to medical treatment.

Health Literacy

The ability to adhere to medical directives requires patients' clear understating of their disease and treatment regimen. Health literacy involves the ability to obtain, process, and understand basic health information needed to make appropriate health decisions (Schillinger et al., 2003; Baker, 2006). The Institute of Medicine (2004) has estimated that more than 90 million people in the United States lack the necessary literacy skills needed to understand and act on health information given by their health care providers. Evidence to date suggests that limited health literacy negatively impacts health outcomes and decreases patient adherence (Gazmararian et al., 2003; Zaghloul and Goodfield, 2004). For example, in a study by Wagner et al. (2009), HIV patients with low health literacy had the highest rate of nonadherence, had less knowledge of their HIV treatment, and had lower self-efficacy for taking prescribed medications. Patients with low health literacy often use fewer preventive services, incur higher medical costs, have limited understanding of their treatment regimen, and have poorer health status (Gazmararian et al., 2003; Schillinger et al., 2003, 2004). Thus, low health literacy increases health disparities and negatively impacts patients' self-management and collaborative care (Schillinger et al., 2004).

While associations between health literacy and patient outcomes are evident, current literacy interventions aimed at improving adherence have shown inconsistent results (Wagner et al., 2009; Paasche-Orlow et al., 2006). Further research is needed to elucidate the underlying mechanisms in which health literacy influences health outcomes, including potential factors that may mediate or moderate the health literacy–adherence relationship.

Interpersonal Factors

The interpersonal dynamics of the physician–patient relationship are important factors that influence rates of adherence. For example, patients who feel that their physicians communicate well with them and actively encourage them to be involved in their own care are more motivated to adhere (O'Malley et al., 2002). Such agreement fosters a cohesive partnership and allows physicians and patients to work together toward a mutually agreed upon treatment plan (Jahng et al., 2005; Martin et al., 2005). In a meta-analysis conducted by Haskard-Zolnierek and DiMatteo (2009), researchers found that training physicians in communication skills resulted in significant improvements in patient adherence. The odds of patient adherence were 1.62 times higher for physicians who were trained in communication compared to the physicians who did not receive communication training (Haskard-Zolnierek and DiMatteo, 2009). Effective communication gives patients the opportunity to openly discuss the risks and benefits of recommended treatments and offers patients the ability to build rapport and trust in their physician, thus increasing disease management skills and ultimately improving adherence (Haskard-Zolnierek and DiMatteo, 2009).

Patients' trust in their physician is a crucial component of the physician–patient relationship. Some researchers suggest that trust has the largest effect on patient satisfaction and adherence to treatment (Shilling et al., 2003). Patients must believe that their physician is someone who can provide them with reliable and honest advice. O'Malley et al. (2002) found that adherence rates were three times higher in primary care relationships characterized by high levels of patient trust. Physicians who promote trust through collaborative partnerships and effective communication and who express compassion and 'bedside manner' for their patients, often succeed in fostering adherence and cooperation for a variety of treatments regimens (Martin et al., 2005).

Depression

Research to date suggests that one of the strongest predictors of nonadherence is patient depression. Depression has been cited as the most prevalent mental illness that physicians are called upon to treat and diagnose (Martin et al., 2005). Depression has been associated with devastating impairments equal to or greater than chronic diseases such as diabetes and hypertension (Hays et al., 1995; Wells and Sherbourne, 1999; Noël et al., 2004). Patients who are depressed experience pessimism and severe cognitive impairments in their daily functioning and often withdrawal from their social support, which diminishes their willingness and ability to follow health care directives (DiMatteo et al., 2000; Martin et al., 2005). DiMatteo et al. (2000) found that the odds of patient nonadherence to medical treatment were three times higher for depressed patients compared to nondepressed patients. Furthermore, researchers estimate that between 50 and 70% of patients with depression often go undiagnosed by primary care physicians (Martin et al., 2005). Recognizing depression as a risk factor for poor health outcomes fosters the potential for improvements in patients' commitment to care and can contribute to increases in subsequent adherence behaviors (DiMatteo et al., 2000).

Assessment of Patient Adherence

The delivery of effective medical care depends on the accurate assessment of patient adherence. Knowledge of patients' health behaviors can offer valuable insight into treatment efficacy and diagnosis (DiMatteo et al., 2013). No single method for the measurement of adherence has proven to be most effective; however, research does suggest that a multifaceted, personalized approach has the most potential for success (DiMatteo et al., 2012). The reasons for patient nonadherence are many, and the factors affecting the accuracy of patient reported adherence vary; thus, methods for improving the measurement of adherence must also vary (DiMatteo et al., 2013).

Patient adherence can be measured in a variety of ways depending on the disease or condition type (chronic or acute). Such methods include self-reports, patient diaries, collateral reports (reports by others), and pill counts. Electronic measures (e.g., medication event monitoring system) can also be used; however, such methods are quite expensive and often do not guarantee actual adherence behaviors. For example, the electronic event monitoring system (also known as MEMS) records the time and date of when patients open or remove medications from pillboxes; while such mechanisms can offer

reliable information, opening or removal of medication from MEMS systems do not necessarily mean that patients actually took their medications. Pharmacy refill assessment can be helpful, and direct observation can also be used to measure patient adherence. The latter, however, is difficult and costly (McNicholas, 2012). Direct observation is used in the treatments of infectious diseases like tuberculosis and HIV, where there are clear medical and economic advantages (McNicholas, 2012). Last, biological markers, using saliva or blood levels can also be an effective and accurate, although costly, way of assessing patient adherence. Thus, understanding patient adherence requires a multimethod approach, where multiple tools are used simultaneously to accurately identify how medical directives are being followed (DiMatteo et al., 2013).

Improving Patient Adherence

Improving patient adherence depends upon trust in the therapeutic relationship, and a realistic assessment of patients' knowledge and understanding of their treatment through effective physician–patient communication. Although the precise estimation of patient adherence is often elusive, physicians' full understanding of whether and why patients choose to adhere is essential (Martin et al., 2005). Moreover, nonadherence, despite its prevalence, often goes unrecognized (or is overestimated) by even the most clinically experienced health professionals (DiMatteo, 2004a; DiMatteo et al., 2012). As mentioned above, the accurate assessment of adherence depends largely on the development of a physician–patient relationship based on trust, where patients are able to admit difficulties in their ability to adhere without fear of criticism. When patients are given the opportunity to express their opinions in a nonthreatening environment, they are more likely to offer the most honest and accurate responses in regard to their treatment behaviors (Martin et al., 2005). From this, vital information about the patient can be obtained (e.g., patient attitudes and cultural beliefs, availability of social support, and any health or emotional challenges), allowing physicians to establish and further explore treatment options that are conducive to the specific needs of their patients.

Research to date suggests that there are numerous adherence-enhancing interventions; however, fewer than half have actually proven to be effective (McDonald et al., 2002; DiMatteo et al., 2012; McNicholas, 2012). Small changes, such as a simpler dosage of medication, may improve adherence. However, such small changes in complex treatment regimens may not be enough to achieve and maintain adherence over time (DiMatteo et al., 2012). As a result, no single intervention strategy can significantly improve adherence; instead a multifaceted approach offers the greatest opportunity for success.

Interventions that have successfully improved patient adherence have utilized a combination of patient education programs and structural support systems, such as patient reminders, increases in patients' clinical visits, or telephone calls from medical teams. Other interventions have incorporated the use of providing reinforcement or incentives, increasing patients' motivation, and providing counseling or family therapy and supportive care (McNicholas, 2012; Miller and DiMatteo, 2013). For example, patients with diabetes face many challenges in efforts to manage their chronic condition. In addition to patient education, continuous reinforcements, and long-term management are often required to maintain adherence over time (Norris et al., 2002). Additionally, several meta-analyses have targeted the importance of tailoring behavioral interventions to increase effectiveness. In Gibson et al. (2000), a review of 36 studies showed that training patients in asthma self-management reduced the risk of hospitalizations, and the number of medical visits and work days missed. Thus, adherence-enhancing interventions must be tailored to meet the unique needs of patients in order to obtain optimal effectiveness (DiMatteo et al., 2012).

Information-Motivation-Strategy Model

The elements of the *Information-Motivation-Strategy* (IMS) model (DiMatteo et al., 2012) offer health professionals the ability to assist patients in building effective disease management strategies to improve health outcomes and maximize treatment adherence overtime. The first component of the IMS model, *information*, outlines the importance of patients becoming more knowledgeable of their disease or treatment condition through effective provider–patient communication. Patients must be able to clearly understand their treatment recommendation in order to be adherent. The second component, *motivation*, highlights the importance of helping patients to believe in the potential efficacy of their recommended treatment and to make a commitment to follow it (Haskard-Zolnierek and DiMatteo, 2009). The final component, *strategy*, suggests that patients must be equipped with all the necessary resources to carry out their adherence behaviors. Providers and medical teams must be willing to help their patients overcome practical barriers that inhibit their ability to adhere (e.g., cost of medication and lack of transportation to scheduled appointments). Additionally, social support is a key element in patents' ability to adhere (DiMatteo, 2004a; DiMatteo et al., 2012). The IMS model offers physicians and health care professionals a practical heuristic that emphasizes the importance of targeting patient needs to promote and achieve optimal health outcomes.

Incentive-Based Engagement Strategies

Current research has explored incentives for physicians and other health professionals to counsel and monitor their patients for various health actions (e.g., smoking). Behavioral economic efforts that include incentive-based engagement strategies, informed collaborative choice, and health decision making, have the potential to help solve the most complex problems that underlie patient nonadherence. As discussed above, elements of the IMS model can be incorporated into provider counseling behaviors to influence patient adherence. As such, patients would be provided with *information*, encouraged to develop

individual *motivation*, and offered *strategies* to motivate and maintain (long-term) desired health habits through reinforcement (Martin et al., 2010).

The framework for incentive-based strategies offers some promise for the future of adherence-enhancing interventions. Worksite studies have already begun to examine incentive design, incentive value, communication strategies, and worksite culture in risk assessment participation. For example, in Volpp et al. (2008), a daily lottery, or financial incentive, demonstrated significant enhancements in patient adherence to medication use. Other incentive-based engagement strategies for various health behaviors, health outcomes, and patients populations have included value-based benefit design in health plans and electronic self-monitoring. In Martin et al. (2010), the results suggest that incentive-based strategies may be the first step in motivating patient adherence and engagement over time. However, further research is needed to determine the relative effects of various types of reinforcements and incentives, and their long-term effectiveness. A full understanding of the mechanisms that encourage health behaviors and the conditions that contribute to patient success will be required in order to implement these strategies in the most cost-effective ways (Volpp et al., 2008).

Conclusion

Patient adherence is essential for the success of quality health care outcomes. Provider–patient communication and the accurate assessment of adherence can increase patient satisfaction, improve overall quality of life, and ultimately reduce the risk of nonadherence. Nonadherence occurs for a variety of reasons, and depends upon the disease condition, the characteristics of the patient, and the type of treatment regimen. Providing patients with *information*, *motivation*, and helpful *strategies* can improve health outcomes and increase patient adherence. Behavioral economics approaches suggest that incentive-based strategies offer future advancements in the delivery of care. Incentive-based strategies can be used as a starting point to help motivate and move patients toward achieving long-term treatment goals (Martin et al., 2010).

See also: Behavioral Medicine; Cancer: Psychosocial Aspects; Cardiac Disease, Coping with; Chronic Illness, Psychosocial Coping with; Co-morbidity of Mental and Physical Conditions; Diabetes: Psychosocial Aspects; Doctor–Patient Interaction in the West: Psychosocial Aspects; Gender and Physical Health; Health Psychology; Illness Behavior and Care-Seeking; Personality and Health; Personalized Medicine; Rehabilitation (Cardiac): Psychosocial Aspects.

Bibliography

Baker, D.W., 2006. The meaning and the measure of health literacy. Journal of General Internal Medicine 21, 878–883.

Christensen, A.J., Ehlers, S.L., 2002. Psychological factors in end-stage renal disease: an emerging context for behavioral medicine research. Journal of Consulting and Clinical Psychology 70, 712–724.

DiMatteo, M.R., 2004a. Variations in patients' adherence to medical recommendations: a quantitative review of 50 years of research. Medical Care 42, 200–209.

DiMatteo, M.R., 2004b. Social support and patient adherence to medical treatment: a meta-analysis. Health Psychology 23, 207–218.

DiMatteo, M.R., Giordani, P.J., Lepper, H.S., Croghan, T.W., 2002. Patient adherence and medical treatment outcomes a meta analysis. Medical Care 40, 794–811.

DiMatteo, M.R., Haskard, K.B., Williams, S.L., 2007. Health beliefs, disease severity, and patient adherence a meta analysis. Medical Care 45, 521–528.

DiMatteo, M.R., Haskard-Zolnierek, K.B., Martin, L.R., 2012. Improving patient adherence: a three-factor model to guide practice. Health Psychology Review 6, 74–91.

DiMatteo, M.R., Lepper, H.S., Croghan, T.W., 2000. Depression is a risk factor for noncompliance with medical treatment. Archives of Internal Medicine 160, 2101–2107.

DiMatteo, M.R., Miller, T.A., Martin, L., 2013. The importance of effective measurement for fostering change. In: Martin, L.R., DiMatteo, M.R. (Eds.), The Oxford Handbook of Health Communication, Behavior Change and Treatment Adherence. Oxford University Press, Oxford, United Kingdom, pp. 370–384.

Gazmararian, J.A., Williams, M.V., Peel, J., Baker, D.W., 2003. Health literacy and knowledge of chronic disease. Patient Education and Counseling 51, 267–275.

Gibson, P.G., Coughlan, J., Wilson, A.J., Hensley, M.J., Abramson, M., Bauman, A., Walters, E.H., 2000. Limited (information only) patient education programs for adults with asthma. The Cochrane Database of Systematic Reviews 2, CD001005.

Haskard-Zolnierek, K.B., DiMatteo, M.R., 2009. Physician communication and patient adherence to treatment: a meta-analysis. Medical Care 47, 826–834.

Hays, R.D., Wells, K.B., Sherbourne, C.D., Rogers, W., Spritzer, K., 1995. The functioning and well-being outcomes of patients with depression compared with chronic general medical illnesses. Archives of General Psychiatry 52, 11–19.

Institute of Medicine, 2004. Health Literacy: A Prescription to End Confusion. http://www.nap.edu/openbook.php?isbn=0309091179 (accessed 25.05.13.).

Jahng, K.H., Martin, L.R., Golin, C.E., DiMatteo, M.R., 2005. Preferences for medical collaboration: patient-physician congruence and patient outcomes. Patient Education and Counseling 57, 308–314.

Lutfey, K.E., Wishner, W.J., 1999. Beyond compliance is adherence. Diabetes Care 22, 635–639.

Martin, L.R., Haskard-Zolnierek, K.B., DiMatteo, M.R., 2010. Health Behavior Change and Treatment Adherence: Evidence-Based Guidelines for Improving Health Care. Oxford, New York.

Martin, L.R., Williams, S.L., Haskard, K.B., DiMatteo, M.R., 2005. The challenge of patient adherence. Therapeutics and Clinical Risk Management 1, 189–199.

McDonald, H.P., Garg, A.X., Haynes, R.B., 2002. Interventions to enhance patient adherence to medicine prescriptions: scientific review. Journal of the American Medical Association 288, 2868–2879.

McNicholas, F., 2012. To adhere or not, and what we can do to help. European Child & Adolescent Psychiatry 27, 657–663.

Miller, T.A., DiMatteo, M.R., 2013. Importance of family/social support and impact on adherence to diabetes therapy. Diabetes, Metabolic Syndrome and Obesity 6, 421–426.

New England Healthcare Institute, 2009. Thinking outside the Pillbox: A System-Wide Approach to Improving Patient Medication Adherence for Chronic Disease. http://www.nejm.org/doi/full/10.1056/NEJMp1002305 (accessed 26.05.13.).

Noël, P.H., Williams, J.W., Unützer, J., Worchel, J., Lee, S., Cornell, J., Katon, W., Harpole, L.H., Hunkeler, E., 2004. Depression and comorbid illness in elderly primary care patients: impact on multiple domains of health status and well-being. Annals of Family Medicine 6, 555–562.

Norris, S.L., Lau, J., Smith, S.J., Schmid, C.H., Engelgau, M.M., 2002. Self management education for adults with type 2 diabetes: a meta-analysis of the effect on glycemic control. Diabetes Care 25, 1159–1171.

O'Malley, A.S., Forrest, C.B., Mandelblatt, J., 2002. Adherence of low-income women to cancer screening recommendations. Journal of General Internal Medicine 17, 144–154.

Paasche-Orlow, M.K., Cheng, D.M., Palepu, A., Meli, S., Faber, V., Samet, J.H., 2006. Health literacy, antiretroviral adherence, and HIV-RNA suppression: a longitudinal perspective. Journal of General Internal Medicine 21, 835–840.

Schillinger, D., Bindman, A., Wang, F., Stewart, A., Piette, J., 2004. Functional health literacy and the quality of physician-patient communication among diabetes patients. Patient Education and Counseling 52, 315–323.

Shilling, V., Jenkins, V., Fallowfield, L., 2003. Factors affect patient and clinician satisfaction with the clinical consultation: can communication skills training for clinicians improve satisfaction? Psychooncology 12, 599–611.

Schillinger, D., Piette, J., Grumbach, K., Wang, F., Wilson, C., Daher, C., Leong-Grotz, K., Castro, C., Bindman, A.B., 2003. Closing the loop physician communication with diabetic patients who have low health literacy. Archives of Internal Medicine 163, 83–90.

Vermeire, E., Hearnshaw, H., Royen, V.P., Denekens, J., 2001. Patient adherence to treatment: three decades of research a comprehensive review. Journal of Clinical Pharmacy and Therapeutics 26, 331–342.

Volpp, K.G., Loewenstein, G., Troxel, A.B., Doshi, J., Prince, M., Laskin, M., Kimmel, S.E., 2008. A test of financial incentives to improve warfarin adherence. BMC Health Services Research 28, 206–214.

Wagner, C.V., Steptoe, A., Wolf, M.S., Wardle, J., 2009. Health literacy and health actions: a review and a framework from health psychology. Health Education & Behavior 36, 860–877.

Wells, K.B., Sherbourne, C.D., 1999. Functioning and utility for current health of patients with depression or chronic medical conditions in managing primary care practices. Archives of General Psychiatry 56, 897–904.

Zaghloul, S.S., Goodfield, M.,J., 2004. Objective assessment of compliance with psoriasis treatment. Archives of Dermatology 140, 408–414.

Patron–Client Relations, Social and Anthropological Study of

Luis Roniger, Wake Forest University, Winston-Salem, NC, USA; Hebrew University of Jerusalem, Jerusalem, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Patron–client relations (PCR) involve asymmetric but mutually beneficial, open-ended transactions based on the differential control by individuals or groups over the access and flow of resources in stratified societies. Patrons provide selective access to resources they control, or place themselves or the clients in positions from which they can divert resources in their favor. In exchange, the clients are supposed to provide resources and services, while sometimes boosting the patron's reputation. The study of PCR has revealed the persisting impact of hierarchical networks in contemporary societies, contributing to the reevaluation of paradigms in social science. Debates continue about the institutional viability and significance of PCR and clientelism in late modernity.

Background

Patron–client relations (PCR) involve asymmetric but mutually beneficial, open-ended transactions based on the differential control by individuals or groups over the access and flow of resources in stratified societies. Whereas in antiquity, PCR formed part of the normative framework of society and could hardly be thought as conflicting with legal institutions and the social ethos, in modern societies PCR are built around such a conflict. While in principle and by law the clients have access to power centers and the ability to convert resources autonomously, they do not actualize this potential because it is neutralized by the patrons' control of the avenues to resource conversion and access to political centers. Yet, as in antiquity, mediation is projected into the institutional arena. Such mediation is contingent on the clients' entering into a relationship of exchange that necessarily limits the scope and convertibility of resources freely exchanged.

PCR originate in settings where social interaction is based on nonascriptive criteria and there are opportunities for mobility and yet there is structural inequality. Even though there are PCR in castelike and fully hierarchical societies, they are not a full-fledged phenomenon, as found in societies of the Mediterranean, Latin America, Southeast Asia, and the Middle East throughout the nineteenth to twenty-first centuries, and earlier on also in Western Europe. In such societies, PCR are motivated by the search for access to scarce resources, be they land, water sources, employment, a loan, schooling opportunities, or medical services. They may be rooted in the search for political protection, in bureaucrats providing preferential access to petitioners in return for future favors, in poor people sharing the social visibility of men or rank, while the latter use these links to consolidate their status. Professionals may seek prestige and a following of 'friends' and 'friends of friends' by offering services without demanding an immediate return. Union leaders, political activists, and government officials may use PCR to build networks of supporters, instrumental for gaining higher level positions, office, and control of institutional resources.

Terminology

The terms used by anthropology and the other social sciences to define such relationships originated in ancient Republican Rome, where relationships of *patrocinium* and *clientelae* proliferated during the Republic and into the Empire. These terms found their way into the vernaculars of Mediterranean and Latin American societies, whereas other terms are used for PCR elsewhere.

The Study of PCR

Expanding in the 1960s and 1970s, by the late 1970s and 1980s, it engaged in general and comparative analyses. In the 1990s, in-depth studies renewed interest in such phenomena in polities under transition to market economies and democracy, with a reevaluation of some of the premises of earlier approaches, especially regarding the connection between PCR and underdevelopment and their symbolic dimension. In the 2000s, the interest in PCR has been connected with the spread of research from anthropology to other social sciences, particularly political science and to a lesser extent sociology. From early dyadic emphases, research has increasingly revealed a more complex range of PCR analyzed in terms of networks, factions, and coalitions.

The study of PCR has been connected and contributed to analytical shifts in the social sciences, especially as it involved dealing with phenomena for which neither the group corporate model nor the paradigms anchored in the formal predicaments of modern institutions could provide a satisfactory guide. In parallel, research revealed the complexities of studying PCR, due to their informal character and to their cross-institutional insertion that requires interdisciplinary knowledge, bridging anthropology and sociology, political science and economics.

Increasingly, however, research has moved to normative assessments geared to eradicate PCR, often conflating it with corruption. In spite of several waves of research, the study of the conditions favoring or undermining clientelism is, perhaps surprisingly, still in its beginnings. Anthropological studies have provided great insight into the perspectives of politicians, brokers, voters, and others involved in PCR, yet their findings have been hard to generalize. Conversely, there are serious problems of reliability in the use of statistical proxies adopted by other social scientists, primarily political scientists, to establish to what extent a system is clientelistic. Given the informal and covert character of PCR, only systematic and long-term fieldwork can corroborate the reliability of the latter. One of the greatest challenges for future studies of PCR is to find

ways to integrate anthropological and sociological perspectives with approaches of economics and political science.

A Logic of Social Exchange

Research has identified a specific logic of social exchange and reciprocity beneath the wide gamut of PCR. Both the control of material and human resources and the presence of trust or its absence play a crucial role in it. This logic has been defined in the literature as *clientelism*. A related term is *patronage*, used interchangeably at times and alluding to the support granted by the patron to his followers and protégés.

The logic of clientelism involves a series of traits:

1. Hierarchy. Whereas friendship is an idiom of equality and similarity between human beings, PCR are built on an idiom of reciprocity and social exchange colored by various grades of inequality and hierarchy. (Although, in some cases, the latter are constrained by the potentially egalitarian images of individual self-fulfillment.)
2. The selective provision of benefits and resources to only limited segments of the population, in areas such as literacy, public health, defense, crime prevention, and infrastructure. Under such conditions, individuals and institutional actors are able to gather followings because they can dispense such access. They do so whether acting as *patrons* (in a strict sense), or as *brokers* with those who control the goods and services, or through a combination of both roles, in what has been called *patron brokerage*.
3. A relationship shaped according to the particular traits of the partners rather than on the basis of what they would be entitled to as members of social categories or in terms of formal institutional definition of roles.
4. PCR are characterized by the simultaneous exchange of two different types of resources and services: instrumental (e.g., economic and political) and 'sociational' or expressive (e.g., promises of loyalty). A package deal allowing for more than a restricted balanced, or market exchange is built-in within PCR, which has been identified as connected to generalized exchange. This determines expectations of entering more embracing attachments, often couched in terms of commitment, even though some PCR are very ambivalent in this respect. It brings agents to value or decry clientelism both as an instrumental means to a calculated end and for their own sake.
5. PCR are said to undermine group (i.e., class and strata) solidarity among both clients and patrons, especially among clients, and to exacerbate resentment and social antagonisms.
6. The favoritism involved implies that many people are excluded from PCR or are related to them only indirectly or intermittently. Where clientelism is highly valued, it generates 'inflationary' expectations that will be disappointed. This adds to the pressures exerted in PCR and to the fragility of clientelistic commitments, even where clientelism is a relatively persistent systemic trait.
7. PCR are neither fully contractual nor legal. They are based on informal, though tightly bound understandings. Most PCR are vulnerable to systemic pressures and to the changed contextual and market situation of the partners who can break them off accordingly.

These traits create an inherently contradictory situation, in which:

1. Asymmetrical power and inequality are combined with promises of solidarity.
2. Potential and/or actual coercion and exploitation coexist with an ideological emphasis on the voluntary character of the attachment on the one hand and on its obligatory nature on the other.
3. PCR structure of limiting payoffs can be only maintained by making payoffs.
4. The covered, informal, and unregulated character of PCR is used to project public claims over social resources and to bolster and propagate public images of power and reputation.
5. Even where clientelism is persistent, PCR are characterized by instability and by a perpetual contest, resource manipulation, and instability.

Exchange Strategies

The agent performing as client is not only expected to provide his patron with specific resources but must also accept the patron's control over his access – and that of other clients – to markets and public goods, and over his ability to convert resources fully. In a parallel manner, the patron's position is not as solid as it may seem, nor guaranteed, for instance by kinship or other ascriptive criteria. Although patrons invest much time and energy in gaining and retaining control over clients, their control is never fully legitimized. It is vulnerable to attack by social forces committed to formal, universalistic principles of social organization and exchange, by the competition of other patrons and brokers, potential and actual, and by social forces excluded from clientelistic relations. Owing to these constant threats, patrons are compelled to rely on their followers to solidify their position. The patron must also relinquish some of the short-term material gains that might accrue from a position of preeminence in order to earn the right to determine the basic rules of the social relationships. In return, the client is protected from social or material insecurity, and is provided with goods, services, and social advancement. Certain types of trust (focalized rather than generalized) are associated with such dynamics.

Complementary exchange strategies are built, which signal what Vincent Lemieux defined as a 'double transformation.' That is, an abdication of autonomy on the client's part and a relaxation of hierarchical controls on the patron's part, through which the client's lack of power becomes dominated power and the latter's lack of domination becomes dominating authority. These exchange strategies are not only affected by immediate, mostly technical considerations of power and instrumentality, but also often encompass mutual, relatively long-term, compromises based on commitments as the prerequisite for the maintenance of ongoing social relationships.

The Context of PCR

The contradictions and instability inherent in PCR and clientelism are symptomatic of the macrosocial context in which they arise. Unlike societies in which hereditary ascriptive principles predominate, full-fledged clientelism flourishes

where markets are no longer controlled through primordial units and allow for an open flow of resources and opportunities for mobility. (This is what distinguishes clientelism from *feudalism*, as pointed out already by Max Weber in *Economy and Society* (vol. II, chapters 12–13) and later on by historians of European nations.) This trend, however, goes hand in hand with a strong tendency toward unequal access to markets and sociopolitical spheres. As they affect distribution and redistribution, clientelism and PCR remain subject to the dynamics of the political economy. Marketing economies, accelerated urbanization, and the expansion or reduction of state administrations affect PCR. Research has shown that the spread of market forces and/or the establishment of forceful administrations eroded patrimonial patterns of social exchange in rural Third World settings, making the terms of trade increasingly oppressive for rural clienteles. Similarly, the impact of world economic trends, fluctuations in the international price of commodities, the complexity of international trade, banking, and aid affected the pool of patronage resources available to states and other agencies, and influenced patterns of control, distribution, and redistribution. Moreover, as these arrangements are not fully legitimized, they remain vulnerable to the challenge of countervailing social forces.

The Comparative Analysis of PCR

Until the early 1980s, it was dominated by the contrast between traditional dyadic patronage and modern party-directed clientelism, a distinction that derived from the dominant paradigm of modernization. This contrast was made classic by Alex Weingrod, who traced it in the different terminology of anthropology and political science and related it to the degree of segmentation or integration of local sectors within nation-states. A clear-cut developmental emphasis was identified as effecting greater network complexity and eroding the seemingly personalized and intimate relationships of traditional settings. This approach focused almost exclusively on the organizational differences of PCR, deriving other aspects from them in a functionalist manner.

In the 1980s and 1990s, comparative analysis has moved away toward approaches that pay attention also to other aspects of PCR. Greater attention was paid first to the modes of agents' role making. Second, the styles of setting a relationship were studied, following what Robert Paine very early on called a generative transactional model of patronage. Third, symbolic and discursive elements, including images and trust, were analyzed as important for the development of PCR. Fourth, the fundamental distinction between PCR as an *addendum* to the core institutional framework in most human societies and clientelism as a strategy of social exchange at the core of institutional life in certain societies and periods. And last but not the least, the issues related to the institutional viability of PCR under democratization, liberalization, and institutional change of late modernity.

Debates on Institutional Viability

Researchers have differed in their interpretation of the institutional viability and significance of PCR in late modern settings. From one perspective, PCR have been analyzed as neutralizing the system of representation and general entitlements, by placing 'friends' in the strategic positions of power and control. Clienteles have been depicted as inimical to the institutionalization of public accountability and therefore stand in contradistinction to a politics open to generalization and participation, and to a discourse aimed at the protection of individual and collective rights. Others have emphasized the pragmatics of social action, stressing that PCR have been an important mechanism for obtaining transactional benefits, in resource allocation and in providing local–regional–national mechanisms of interrelation and integration. While PCR run counter to universalistic standards, they are sensitive to local sentiment, instrumental for solving problems, and provide ways of incorporating new sectors of the population such as immigrants and of advancing political actors with an entrepreneurial ability. From this perspective, it is claimed that as long as clienteles maintain some balance of reciprocity, the participation in political and economic markets by the 'capi-clientele' (to borrow Gaetano Mosca's expression) – be he or she a broker, a patron or a patron–broker – constitutes a means for individuals in their entourage to influence public decisions. In this sense, the literature alludes to PCR as reconciling public and private authority and formal and informal rules of the game.

PCR may be resented, criticized, and opposed by social forces and coalitions wishing to curtail its presence alongside bureaucratic universalism and market rationality. In parallel, sectors benefiting from patronage may see it as a pragmatic avenue of controlled freedom, useful for advancing in social, economic, and political domains that are regulated by competition for access to power, resources, and services. This duality reflects a major tension of modern democratic polities. As part of the Western political tradition, modern democracies are built on citizenship that implies political equality but leaves the economic domain open to inequalities.

This explains the paradoxical development of clientelistic networks in societies that have adopted policies of macroeconomic adjustment and restructuring. Promotion of liberalization, reduction of state intervention in favor of market mechanisms, privatization of state-owned and state-supported services, and curtailment of union power further fragment society and heighten the need for support networks. When available, clientelism has remained as important as ever throughout political revamping in Poland, Russia, Hungary, Romania, Brazil, Argentina, and Turkey.

In centralized polities, PCR have been depicted as a 'lubricant' to bypass the inefficient and rigid official system, constituting highly trusted informal problem-solving networks that provide instrumental advantages and emotional support in situations of low institutional trust. In representative democracies, clientelism has been sometimes highly effective in encouraging and rewarding party activists and in molding incumbency to public offices in accordance with politically sensitive stances. Clientelism continues to be useful in the competition for power and, once power is achieved, for effectively formulating and implementing policies. Within such a framework, patronage has become a restricted but legitimate procedure related to the acknowledged right of high office incumbents to appoint their followers to positions of public responsibility. In other respects, clientelism remains

controversial and is open to allegations of corruption, due to its informal, unregulated character, and the particularistic use of public resources. In other situations, it is ignored as long as possible or is disguised as friendship, which is more acceptable in terms of the proclaimed ethos of modern equality.

Contemporary studies suggest that in the environment of postindustrial societies, patronage develop among professional and upper strata, rather than being a phenomenon restricted to the lower classes. In addition, PCR are not confined to politics in the narrow sense, but proliferate as well in the arts, the academia, the church, the media, and even in business – whenever we are dealing with the power of appointment and the granting of access to benefits, goods, services, influence, and honors.

Changes in the perception of PCR are of systemic consequence if they result in the institutionalization of mechanisms through which citizens can press for their rights and entitlements without personal mediation. Such changes include: civil service reforms, nonpartisan public systems, recognized charters of rights, controls over party fund raising, and nonpartisan comptrollers as a prestigious and trustworthy branch of government. The functioning of these institutional mechanisms hinge on public support for a configuration of the public sphere structured around public accountability and formal responsiveness to turn more and more discrete issues into publicly negotiable and politically consequential.

See also: Brokers and Brokerage, Anthropology of; Development: Social-Anthropological Aspects; Friendship, Anthropology of; Honor and Shame; Mafia, Anthropology of; Modernity: Anthropological Aspects; Political Anthropology.

Bibliography

Auyero, J., 1998. Poor People's Politics. Peronists' Survival Networks and the Legacy of Evita. Duke University Press.

Berenschot, W., 2010. Everyday mediation: the politics of public service delivery. Development and Change 41 (5), 883–905.

Boissevain, J., 1974. Friends of Friends. Basil Blackwell, Oxford.

Brinkerhoff, D.W., Goldsmith, A., 2002. Clientelism, Patrimonialism and Democratic Governance: An Overview and Framework for Assessment and Programming. Abt Associates' report for the US Agency for International Development.

Clapham, C. (Ed.), 1982. Private Patronage and Public Power. St. Martin's Press.

Eisenstadt, S.N., Roniger, L., 1984. Patrons, Clients, and Friends. Cambridge University Press.

Ganev, V.I., 2007. Preying on the State: The Transformation of Bulgaria after 1989. Cornell University Press.

Gay, Robert, 1998. Rethinking clientelism: demands, discourses and practices in contemporary Brazil. European Review of Latin American and Caribbean Studies 66, 7–24.

Gellner, E., Waterbury, J. (Eds.), 1977. Patrons and Clients in Mediterranean Societies. Duckworth.

Hilgers, T. (Ed.), 2012. Clientelism in Everyday Latin American Politics. Palgrave.

Keefer, P., Vlaicu, R., 2008. Democracy, credibility and clientelism. Journal of Law, Economics and Organization 24 (2), 371–406.

Kitschelt, H., Wilkinson, S. (Eds.), 2007. Patrons, Clients and Policies. Cambridge University Press.

Lemarchand, R., 1990. The state, the parallel economy, and the changing structure of patronage systems. In: Rothchild, D., Chazan, N. (Eds.), The Precarious Balance. State and Society in Africa. Westview, pp. 149–170.

Lemieux, V., 1987. Le sens du patronage politique. Journal of Canadian Studies 22 (2), 5–18.

Penfold-Becerra, M., 2006. Clientelism and Social Funds: Empirical Evidence from Chávez's 'Misiones' Programs in Venezuela. Advanced Institute in Administrative Studies.

Rocha Menocal, A., 2001. Do old habits die hard? A statistical exploration of the politicisation of progresa, Mexico's latest federal poverty-alleviation programme, under the Zedillo administration. Journal of Latin American Studies 33, 513–538.

Roniger, L., 2004. Political clientelism, democracy and market economy. Comparative Politics 36 (3), 353–375.

Roniger, L., Güneş-Ayata, A. (Eds.), 1994. Democracy, Clientelism and Civil Society. Lynne Rienner.

Shefter, M., 1977. Patronage and Its Opponents. Cornell University Western Societies Program, Ithaca. Occasional Papers 8.

Stoica, C.A., 2004. From good communists to even better capitalists? Entrepreneurial pathways in post-socialist Romania. East European Politics and Society 18, 236–277.

Weingrod, A., 1968. Patrons, patronage and political power. Comparative Studies in Society and History 7, 377–400.

Willerton, J.P., 1992. Patronage and Politics in the USSR. Cambridge University Press.

Pavlov, Ivan Petrovich (1849–1936)

Daniel A Powell, VA Medical Center, University of South Carolina, Columbia, SC, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 11127–11130, © 2001, Elsevier Ltd.

Abstract

Ivan Petrovich Pavlov is recognized as one of the most renowned of Russian scientists. He was born in 1849 in a small Russian village and attended the local ecclesiastical schools. However, he was heavily influenced by Darwin, as well as the famous Russian physiologist Sechenov and entered St. Petersburg State University to study physiology. He received a medical diploma in 1879 and successfully defended his doctoral dissertation in 1883. Pavlov later became the Chair of the Physiology Department at the Military Medical Academy and Head of the Department of Physiology in the Institute of Experimental Medicine. He received the Nobel Prize in 1905 for his work on the nervous control of the digestive system by its autonomic input. Early in the twentieth century Pavlov's work changed to the study of conditioned reflexes, using salivation as the conditioned response. Pavlov was able to demonstrate in an objective manner through the development of a salivary fistula that salivation could be elicited not only by food but also by neutral stimuli that came to be associated with feeding. The use of these conditioned reflex techniques has come to be one of the major methods used in studying brain-behavior relationships in modern psychology.

Ivan Petrovich Pavlov is recognized as one of the most renowned of Russian scientists and is clearly the best known Russian physiologist of his era. His work on the physiology of the digestive system was well respected by his colleagues worldwide during his lifetime. This respect earned him the first Nobel Prize awarded to a physiologist and the first to a Russian scientist in 1905. Although the excellence of his work on the physiology of the digestive system is well recognized, Pavlov is even better known for his investigations into what he termed 'conditioned reflexes.' This work has had a worldwide impact on the fields of learning and memory, psychology, neuroscience, etc. Pavlov's life can be divided roughly into three phases. The first, including his early years, was a phase in which Pavlov received his early training in physiology and medicine. A second investigational phase was one in which his work on the physiology of the digestive system was done. The third includes his work, for which he became most famous, on conditioned reflexes.

Pavlov's Early Life

Ivan Petrovich Pavlov was born on 14 September 1849, in the small town of Ryazan, about 100 miles southeast of Moscow. He was the eldest of five children. His father was a priest in one of the poorer parishes. Thus, during Pavlov's early life he lived in extreme poverty. Nevertheless, the life of Pavlov and his siblings was apparently intellectually rich and their father instilled a love of learning in each of his children. Pavlov entered the ecclesiastical high school of Ryazan, and from there to the local seminary, obstentially to pursue an ecclesiastical career, as his father and other members of his family had done. Pavlov thus entered school during the Russian Cultural Revolution of the 1860s and therefore was exposed to the leaders of this revolution, such as Turgeanyev, Doskovsky, Tolstoy, etc. However, he was apparently most heavily influenced by Pisarev, from whom he learned of Darwin's theory of natural selection and the famous Russian physiologist Sechenov, who had just published *Reflexes of the Brain* in 1896. It is clear that these kinds of influences on Pavlov steered him in the direction of science and consequently he left the ecclesiastical seminary and entered the Mathematics–Physics Department at St. Petersburg State University.

In 1875 Pavlov graduated from the University with a degree in natural sciences. However, due to his interest in physiology he started work as an assistant in the Department of Physiology in the Medical Surgery Academy (now the Military Medical Academy) and in 1879 received a medical diploma. He was then recommended for an appointment with Andre Botkin, who was the most renowned clinician in St. Petersburg at the time. Botkin assigned Pavlov the task of heading his recently established experimental research laboratory. Since Botkin had little interest in experimental work, Pavlov was able to work pretty much independently on problems of his choice. He worked here until 1890. He was heavily influenced by Botkin's theories on the role that the nervous system plays in the regulation of physiological function, and it was here that Pavlov began his studies of the peripheral nerves of the heart and where in 1883 he successfully defended his doctoral dissertation in this area.

During this period of time he also became acquainted with a small circle of friends including another university student, Seraphima Vasilievna, who was in her last year at the Pedagogical Institute. They were married in 1881. During the first years of their married life they lived in extreme poverty. After a first miscarriage, Seraphima gave birth to a son but he also died the next summer. They did, however, later have four children (three boys and a girl) and, based on Seraphima's memoirs published after Pavlov's death, apparently enjoyed a happy family life. In 1883 Pavlov spent 2 years working abroad in the laboratories of Heidenhain and Ludwig in Germany. On his return to St. Petersburg, he resumed his work in Botkin's laboratory and in 1890 was appointed Professor in the Pharmacology Department at the Military Medical Academy. Five years later he joined the Physiology Department there, which he headed for the rest of his career. He was also later appointed the Head of the Department of Physiology in the newly established Institute of Experimental Medicine. Pavlov

worked in his laboratory there until his death at the age of 87 in Leningrad on 27 February 1936.

Pavlov's Work on the Digestive System

Pavlov did most of his research on the physiology of digestion during his early years at the Institute for Experimental Medicine. Here he was able to develop his surgical skills with chronic experiments, which he advocated over the acute experiments that were typically done in physiology at that time. Pavlov believed that the study of the entire organism in as normal a state as possible was crucial to understanding physiological systems. Moreover, he was an excellent surgeon and became quite experienced in developing surgical procedures that allowed his dogs to fully recover. In such animals the digestive processes could be studied under normal conditions. As a consequence of this approach to physiology he developed a 'stomach pouch,' which involves isolating a portion of the dog's stomach, so that it can be directly observed by the experimenter. Unlike previous researchers, the nervous input to the isolated portion of the stomach was intact in the 'Pavlov Pouch.' Using this preparation Pavlov could observe the digestive juices and enzymes that occurred as a result of normal food intake, or after direct placement of food into the stomach. It was the comparison of these two states which led to Pavlov's interesting discovery that food placed in the mouth as a normal consequence of eating, produced a significantly larger amount of gastric secretion than when food was placed directly in the stomach. Moreover, when animals were sham fed, viz. when the normal intake of food through the mouth was allowed to take place but the food was diverted externally, a significant amount of gastric secretion occurred, even though the food never reached the stomach.

It was these types of experiments, which led to Pavlov's pioneering discoveries regarding the control of the digestive system by its autonomic nervous input, primarily the vagus nerve. He showed for example that a major consequence of this neural control involves both gastric and pancreatic secretions before any food actually reaches the stomach, by the sight of the food bowl for example, or other stimuli previously associated with eating. These 'psychic secretions' turned out to be of great importance for digestion and led to Pavlov's major conclusions regarding what came to be known as the doctrine of nervism, in which the major tenet states that the physiological control of bodily systems is primarily through its nervous input. This work culminated in the publication of *Lectures on the Work of the Digestive Glands* in 1897, later translated into English in 1910 (Pavlov, 1910). However, it was during this same period of time that two British physiologists, W. Bayliss and E.J. Starling, demonstrated that the secretin of the pancreatic enzymes was primarily due to release of a substance by the intestines during eating, which came to be called secretion. These investigators thus claimed that the doctrine of nervism was invalidated by their discoveries. However, as is typically the case with two opposing theories of biological function, both experimental outcomes were eventually confirmed. The new science of neuroendocrinology thus came to be established, and it is now widely accepted that nervous control over hormonal output is a ubiquitous aspect of physiological function.

Pavlov's Work on Conditioned Reflexes

As a result of his work on the digestive processes, as noted above, Pavlov also observed that nonfood stimuli that became associated with food were able to elicit salivary secretions, and to a lesser extent digestive secretions, even though no food was actually present in the mouth. These 'psychic secretions' came to form the basis for the remainder of Pavlov's work and resulted in his experiments moving in a new direction, which caused an almost complete cessation of work on the digestive system proper. However, Pavlov believed that by studying 'psychic secretions,' which he referred to as 'conditioned reflexes,' he would be able to demonstrate in an objective manner how the brain controls adaptive behaviors. Pavlov developed a salivary fistula, similar to the stomach pouch, in which the release of salivation in the mouth was directed through a tube to a container outside the mouth. Through this technique he was able to discover the amount and kinds of salivary secretions that were produced by different kinds of sensory stimuli.

Using these techniques Pavlov began the study of a new kind of learning, which has come to be known as 'classical' or 'Pavlovian' conditioning. Classical conditioning occurs whenever a neutral stimulus acts as a signal for a forthcoming significant event. Thus, for example, in one of Pavlov's original experiments a pure tone, a metronome, or some other equally neutral stimulus signaled to a dog that it was about to receive an appetitive stimulus, i.e., meat powder (see Pavlov, 1927). In Pavlov's experiments this neutral auditory stimulus was termed the conditioned stimulus (CS) and the meat powder the unconditioned stimulus (UnS). The UnS always elicits what Pavlov referred to as an unconditioned response (UR), viz. salivation in response to the meat powder. The behavior in which Pavlov became most interested, however, was not increased salivation in response to the meat powder, which he had previously studied intensively, but the new learned response to the neutral CS. This response also consisted of salivation, but in this case salivation in response to the initial neutral auditory stimulus, which resulted from its consistently preceding the meat powder over many CS/UnS presentations. This new learned response was referred to as a conditioned response (CR) and was thought to be only temporary and required reinforcement by the UnS for its maintenance. The UnS is thus often referred to as a reinforcer, since it reinforces the new response to the CS.

It should be noted that the CR does not, however, always resemble the UR as is the case with the original salivation experiments, which Pavlov reported. Indeed often the CR appears to be opposite to the UR. For example, the autonomic changes associated with the contextual cues that signal drug administration, an often-studied type of classical conditioning, are opposite in direction to those produced by the drug itself (Siegel, 1979) and the heart rate CR to CSs that signal aversive UnSs consist of bradycardia, whereas the UR to these same aversive stimuli alone consists of tachycardia. There are many other experimental operations that have been studied since the time of Pavlov, which produce similar new responses to an original neutral stimulus that is always followed by either a noxious or appetitive event. Pavlov thus developed a new

experimental methodology for studying the role of the brain as it initiates new behaviors enabling animals to adapt to their environmental circumstances.

One of Pavlov's greatest contributions to physiology was to emphasize that the nature of the 'psychic secretions,' which were previously thought to be in the realm of psychology, could be objectively studied by physiologists through the conditioned reflex method. He thus believed that the true road to understanding brain function and therefore human behavior was through the objective physiological techniques, which he had developed. In fact Pavlov went on to consider a range of higher level functions such as thinking, reading, emotional reactivity, etc., in terms of conditioned reflexes. Using the ideas of cortical excitation and inhibition, which he studied extensively using the conditioned reflex methodology, he was thus able to explain many complicated behavioral phenomena, such as the neuroses and other psychiatric disturbances. Although Pavlov's influence on experimental psychology has been extensive and dramatic, as detailed in Section Pavlov's Influence on Modern Psychology, many of the details of his theoretical interpretations have since been found to be lacking. Nevertheless the basic operational procedures for determining the effects of a signal on brain processing has been extremely influential in studying brain function.

Pavlov's Influence on Modern Psychology

During his lifetime and immediately thereafter, Pavlov had a tremendous influence on physiology and the study of brain function. Many students were drawn to his laboratory during the early part of the twentieth century to study the basic laws governing the activity of the brain, and during this period of time Pavlov received worldwide acclaim and recognition. Pavlov's work was little affected by the Russian Revolution, which was also ongoing during this time. He maintained a skeptical attitude regarding politics and government, but nevertheless his worldwide recognition led the Communists to continue to fund his research at high levels. Thus, the Soviet Union became known for its support of the study of physiology, and it was during this time that a great center for the study of physiology with many distinguished workers was developed in the Soviet Union, primarily under Pavlov's leadership.

It was his influence outside the Soviet Union, however, that resulted in Pavlov's most notable successes in science, primarily in the field of psychology, which Pavlov had previously rejected. Thus the conditioned reflex methodology was instrumental to the development of the behavioristic movement in psychology in the early part of the twentieth century. John Broadus Watson, one of the major pioneers in the development of behaviorism, utilized the conditioned reflex methodology to explain the entire subject field of psychology, which consisted, according to behaviorists, in the study of overt behavior without reference to subjective phenomena such as cognitions, learning, feelings, etc. The publication of *Psychology as the Behaviorist Views It* (1913) by Watson thus became instrumental in informing American and English scientists of Pavlov's work. The translation of Pavlov's *Conditioned Reflexes* (1927) by Anrep made his work available to English speaking scientists. As a result, experimental psychologists began to explore the new conditioned reflex methods for studying behavior at several academic centers in the USA. As was noted by Babkin (1949), however, much of the work done by the American researchers focused on skeletal reflexes, as opposed to the visceral reflexes, to which Pavlov and his students had previously devoted most of their work.

One of the basic experiments developed by American psychologists was the classical eyeblink-conditioning paradigm. Using this methodology human or animal subjects are presented with a corneal airpuff, which causes reflexive closure of the eyelids. However, when this unconditioned stimulus is preceded by a conditioned stimulus such as a light or pure tone, eventually subjects began to show anticipatory eyeblinks during the conditioned stimulus, which occur even though the corneal airpuff is not presented. This paradigm came to be used extensively in the early 1940s and 1950s as behaviorism began to gather momentum. Its success in the hands of the American psychologists led to the de-emphasis on visceral changes as conditioned responses in the USA. However, Horsley Gantt, who had previously studied with Pavlov, demonstrated in his Johns Hopkins laboratory that both visceral and skeletal responses could be studied in the same organisms, and that the two responses differed greatly in their acquisition and the conditioning parameters required to elicit learning (Gantt, 1960). Much of this early work on classical conditioning was, however, overshadowed by the emphasis given to operant conditioning by other behaviorists, such as B.F. Skinner and Clark Hull. Thus, it was only during the early 1960s that classical conditioning became a popular technique for studying behavior. This was due to the development by I. Gormezano of the classically conditioned nictitating membrane response in the rabbit (Gormezano, 1966). The nictitating membrane is a third laterally moving eyelid found in the rabbit and some other mammals (e.g., cat). Using this animal preparation Gormezano and his students were able to demonstrate the parametric circumstances under which this kind of learning takes place and what kinds of visceral changes accompany them, e.g., conditioned changes in heart rate, blood pressure, etc. The popularity of this technique became even greater when integrated with the simultaneously developing field of behavioral neuroscience. Classical conditioning offers several advantages for studying concomitant brain function as well as new learned behaviors. For example, concomitant electrophysiological recording from single neurons in specific parts of the brain have led to the discovery that different brain structures are involved in visceral vs skeletal learning (Thompson, 1991). Moreover, more recent brain scanning techniques in humans have demonstrated that even during simple classical eyeblink conditioning activation of several specific areas of the brain, which are known to be involved in learning and memory processes, occurs (e.g., Blaxton et al., 1996).

Thus the use of the conditioned reflex techniques originally developed by Pavlov has come to be one of the major methods used in studying brain-behavior relationships in modern psychology. Pavlov's contribution to this new technology was at a basic level and much of his earlier conclusions regarding the results of his manipulations were, of course, erroneous.

However, his contributions we now know were instrumental to the development of modern psychology.

See also: Experimentation in Psychology, History of.

Bibliography

Babkin, B.P., 1949. Pavlov: A Biography. University of Chicago Press, Chicago, IL.

Blaxton, T.A., Zeffiro, T.A., Gabrieli, J.D.E., Bookheimer, S.Y., Carrillo, M.C., Theodore, W.H., Disterhoft, J.F., 1996. Functional mapping of human learning: a positron emission tomography activation study of eyeblink conditioning. Journal of Neuroscience 16, 4032–4040.

Gantt, W.H., 1960. Cardiovascular component of the conditional reflex to pain, food, and other stimuli. Physiological Review 40, 266–291.

Gormezano, I., 1966. Classical conditioning. In: Sidowski, J.B. (Ed.), Experimental Methods and Instrumentation in Psychology. McGraw Hill, New York.

Pavlov, I.P., 1910. Lectures on the Work of the Digestive Glands (W.H. Thompson, Trans.), second ed. Charles Griffin and Co., London.

Pavlov, I.P., 1927. Conditioned Reflexes (G.V. Anrep, Trans.). Oxford University Press, London.

Sechenov, I.M., 1866/1965. Refleksy Golovnogo Mozga (Reflexes of the Brain). MIT Press, Cambridge, MA.

Siegel, S., 1979. The role of conditioning in drug tolerance and addiction. In: Keehn, J.D. (Ed.), Psychopathology in Animals. Academic Press, New York.

Thompson, R.F., 1991. Are memory traces localized or distributed? Neuropsychologia 29, 571–582.

Watson, J.B., 1913. Psychology as the behaviorist views it. Psychological Review 20, 158–177.

Pay, Compensation, and Performance, Psychology of

Joseph J Martocchio, University of Illinois at Urbana-Champaign, Champaign, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H. Thierry, volume 16, pp. 11130–11134, © 2001, Elsevier Ltd.

Abstract

Pay-for-performance stands for a variety of compensation practices that are designed to reinforce past performance and encourage future performance. The purpose of this article is to review some of the more important issues pertaining to pay-for-performance concepts and research. First, pay-for-performance concepts and practices are introduced in the context of total compensation practices. Second, consideration is given to a longstanding vexing concern about whether rewarding performance with tangible rewards may actually undermine performance. Third, job performance constructs in the context of pay-for-performance are considered. Fourth, executive compensation determination and its implications for pay-for-performance are reviewed. Finally, the importance of recognizing national culture in pay-for-performance research is addressed.

The pay-for-performance topic is ubiquitous throughout all types of organizations. In recent decades, the role of pay-for-performance was expanding beyond some of the traditional topics that have driven past psychological research. Questions such as 'Are alternative pay-for-performance plans equally effective in influencing employee motivation and job performance?' were the focus. Increasingly the term pay-for-performance was being used to convey a much broader and very different set of activities than it once did. For instance, researchers are asking questions such as 'Are executives paid commensurately with organizational performance such as changes in stock prices?' Whereas traditional research focused largely on rank-and-file and managerial employee categories, more recent research has focused on the chief executive officers. The former question fits with traditional pay-for-performance research that has been based in psychological theories of motivation; the latter represents a more contemporary approach of compensation practices in a business strategy context, often referred to as executive compensation that has been largely based on theories in economics, finance, and sociology.

Compensation represents both the intrinsic and extrinsic rewards employees receive for performing their jobs (Martocchio, 2015). Together, both intrinsic and extrinsic compensation describe an organization's total compensation. In practice, organizations pay employees to perform the jobs for which they were hired. By the end of 2013, the US private sector labor force totaled, approximately 115 million individuals and the costs of pay, wages, and salaries have increased dramatically in recent years ($20.55 in hourly wages, which increased 3.8% in an 18-month period through late 2013). It would seem reasonable, then, to expect there to be a strong continuing stream of research on topics such as pay-for-performance to better understand how increasing wage and salary costs relate to employee productivity. However, this has not been the case.

According to Gupta and Shaw (2014), of more than 1200 sessions at the 2013 annual meetings of the Society for Industrial and Organizational Psychology, only three are related to compensation, rewards, or benefits. Similarly, the program of the HR division of the Academy of Management featured only three sessions on compensation, among hundreds devoted to other topics, at the 2012 annual meetings. In addition, they cited an analysis which revealed that of 111 articles published in *Personnel Psychology* in 2003–07, only two focused on compensation issues; the parallel proportion was seven of 457 in *Journal of Applied Psychology* (Cascio and Aguinis, 2008). For the period 2008–13, only three additional articles on employee compensation were published in the *Journal of Applied Psychology* (of 602 total articles published in this period) and one article on employee compensation was published in *Personnel Psychology* (of 153 articles). Research on executive compensation is plentiful, and it appears in scholarly journals that publish strategic management research, including *Academy of Management Journal*, *Journal of Management*, and *Strategic Management Journal*.

Compensation Components

Compensation represents both the intrinsic and extrinsic (monetary) rewards or returns employees receive for performing their jobs, and compensation professionals typically refer to both types of compensation as total compensation (Martocchio, 2015). Intangible compensation includes recognition, status, employment security, challenging work, and unplanned learning opportunities. Tangible rewards include pay (hourly wage or annual or annual salary) and employee benefits (such as paid time-off). Whereas organizational development professionals promote intrinsic compensation primarily through effective job design and employee feedback mechanisms, compensation professionals are responsible for extrinsic compensation.

Intrinsic Compensation

Intrinsic compensation represents employees' critical psychological states (that is, intrinsic motivation) that result from performing their jobs. Job characteristics theory (Hackman and Oldham, 1976) describes these critical psychological states. According to this job theory, employees experience enhanced psychological states (that is, intrinsic compensation) when

their jobs rate high on five core job dimensions: skill variety, task identity, task significance, autonomy, and feedback.

Extrinsic Compensation

Extrinsic compensation includes both pay and employee benefits. Employees receive base pay, or money, for performing their jobs. Base pay is recurring; that is, employees continue to receive base pay as long as they remain in their jobs. Organizations disburse base pay to employees as hourly pay or (wage) as an annual salary. Employees earn hourly pay for each hour worked. They earn salaries for performing their jobs, regardless of the actual number of hours worked. Organizations measure salary on an annual basis. The Fair Labor Standards Act established criteria for determining whether employees should be paid on an hourly or salary basis for the purposes of determining who qualifies for overtime pay when working longer than the standard work week. In general, managerial, professional, and executive-level jobs are exempt from the overtime pay provision while most other jobs are not exempt.

In either case, most organizations implement pay-for-performance plans to determine whether pay increases should be awarded, and, if so, by how much. Organizations embrace the importance of capital, which refers to the factors that enable themselves to generate income and economic value. Human capital (Becker, 1975), as defined by economists, refer to sets of collective skills, knowledge, and ability that employees can apply to create economic value for their employers. There are a variety of different types of capital that organizations rely on to create value, including financial capital (cash) and capital equipment (state-of-the-art robotics used in manufacturing). Unlike other forms of capital, employers cannot own human capital. Organizations purchase the use of human capital by paying employees an hourly wage, salary, or bonuses and providing benefits such as paid vacation and health insurance. Also, organizations help develop human capital to their advantage by offering training programs aimed at further boosting employee productivity, and rewarding better performance increases with boosts to pay.

Pay-for-Performance

Merit pay and incentive pay are the primary mechanisms for providing performance-based pay. The intent of pay-for-performance plans is to reward excellent job performance, motivate future performance, and help employers retain valued employees. Under *merit pay*, employees earn permanent increases to base pay according to their performance. Most often, an employee's supervisor or manager provides an evaluation of past performance (oftentimes for periods ranging between 6 months and 1 year) based on subjective judgment. The amount of a merit pay increase should reflect prior job performance levels according to the performance evaluation judgment and the amount ideally motivate employees toward striving for exemplary performance. The pay raise amount should be meaningful to employees in an absolute and relative sense. In an absolute sense, the concept of just-meaningful pay increase refers to the minimum pay increase that employees will see as making a meaningful change in compensation (Krefting and Mahoney, 1977). The basic premise of this concept is that a trivial pay increase for average or better employees is not likely to reinforce their performance or to motivate enhanced future performance. Equity theory (Adams, 1963) tells us that individuals who judge that their pay or pay increases are commensurate with their inputs such as effort, skills, and knowledge that manifest in job performance, relative to the pay or pay increases of comparison others (coworkers) who perform similar or identical jobs. Perceptions of underpayment inequity (Adams, 1965) may intensify because budget reductions make it more challenging to recognize differences in employee performance. Exemplary performers may receive pay raises that are only slightly higher than average performers. Also, overall lower pay increases may fall below the threshold of just-meaningful pay differences (Zedeck and Smith, 1968), then, possibly exacerbating perceptions of inequity.

Incentive pay is defined as compensation (other than base wages or salaries) that fluctuates according to employees' attainment of some standard based on a *preestablished* formula, individual or group goals, or organization earnings. Incentive pay or variable pay rewards employees for partially or completely attaining a predetermined work objective based on objective criteria such as sales volume, safety records, and team performance such as manufacturing quality.

Effective incentive pay systems are based on three assumptions. First, individual employees and work teams differ in how much they contribute to the organization, both in what they do as well as in how well they do it. Second, the organization's overall performance depends to a large degree on the performance of individuals and groups within the organization. Third, to attract, retain, and motivate high performers and to be fair to all employees, an organization needs to reward employees on the basis of their relative performance.

Goal setting is a motivational technique that involves employees' participation in setting precise job performance goals for themselves. Locke (1968) first established the theoretical basis for how goal setting can impact motivation. Latham and Yukl (1975) reviewed a multitude of goal setting research from which they created a theoretical framework for how goal setting should be used in organizations to motivate employee performance. Rewards are those tangible things (pay, benefits, recognition, promotion opportunities, and so forth) that an organization provides to its employees in return for their performance.

Incentive pay augments employees' base pay, but incentive pay appears as one-time payments. Employees usually receive a combination of recurring base pay and incentive pay, with base pay representing the greater portion of monetary compensation. More employees are presently eligible for incentive pay than ever before, as organizations seek to control costs and motivate employees continually to strive for exemplary performance.

Is Monetary Compensation a Nonstarter for Discussing Employee Performance?

Notwithstanding the ubiquity of pay-for-performance practices, a body of literature has given pause to speculate whether pay may diminish performance. Cognitive evaluation theory

(CET) maintains that pay is likely to reduce an individual's interest in the job at hand. Consequently, reduced intrinsic motivation may be more likely to diminish job performance.

At center stage is the concept of intrinsic motivation. Deci and Ryan (1985) describe intrinsic motivation as "based in the innate, organismic needs for competence and self-determination" (p. 33), and argue that it occurs in its purest form when "a person does an activity in the absence of a reward contingency or control" (p. 35). In the CET framework, pay undercuts intrinsic motivation because it exerts a governing quality. Employers expect employees to work diligently when on the job. In rare instances, employees do have the freedom not to perform their jobs whenever they desire if they expect to be paid. For most, pay provides for basic needs including food and shelter without which individuals would not survive. According to Rynes et al. (2005), this perception of being controlled is assumed to be "demotivational" and to work against the potential incentive effects of extrinsic rewards such as pay.

The number of experiments on the effects of extrinsic rewards on intrinsic compensation is substantial; however, strong evidence to support or disconfirm the main principle of CET is mixed. Two metaanalytic reviews did not clarify this issue. For example, Eisenberger and Cameron (1996) examined dozens of studies that compared a rewarded group (verbal or tangible rewards) with a no-reward control group on two traditional measures of intrinsic interest: amount of free-time devoted to a task when the tangible or verbal reward is subsequently withdrawn or self-reported attitude toward the task (e.g., interest, enjoyment, or satisfaction). The results revealed no significant effect of reward condition and free-time, but did show that the relationship between reward condition and task-related attitudes was actually positive, which is inconsistent with CET. Deci et al. (1999) disagreed with the Eisenberger and Cameron (1996) conclusion and addressed them with a metaanalysis on the same issues. Still, they found no overall effect for performance-contingent rewards on the attitude measure of intrinsic interest (p. 644).

Fang and Gerhart (2012) conducted a field examination of CET to check whether Deci and Ryan (1985) may have focused too much on the potentially controlling aspects of rewards and not enough on the informational aspect. Consistent with this hypothesis, Fang and Gerhart (2012) found that employees covered by pay-for-performance plans reported higher intrinsic interest than employees not covered by pay-for-performance plans. Their finding is consistent with the fact that people often report pay as a motivator when it is perceived as recognition for good performance (Lawler, 1971).

As noted previously in this article, equity theory, and goal setting theory help explain the motivational impact of pay-for-performance practices. Expectancy theory also adds to the discussion. According to expectancy theory, an individual will choose to behave in a particular manner because they are motivated to select a specific behavior rather other behavior. The choice is primarily based on the desirability or valence of the outcome of the chosen behavior. Contrary to the premise of CET, a metaanalysis conducted by Locke et al. (1980) concluded: "Money is the crucial incentive … no other incentive or motivational technique comes even close to money with respect to its instrumental value" (p. 379). Subsequent research has continued to support their conclusion.

Ultimately, the body of evidence and thinking about the influence of tangible rewards (pay) on motivation and subsequent job performance is mixed. Reasonably, there is some merit to both perspectives on theoretical rationale and evidence. Perhaps the issue is not whether pay undermines motivation, but, rather, the conditions that influence that relationship. Undoubtedly, continued research will help shed further light on this important matter.

Job Performance

According to Motowidlo (2003), *job performance* is defined as the total expected value to the organization of the discrete behavioral episodes that an individual carries out over a specified time period. Those researchers emphasize two key issues in this definition. First, performance is an aggregated property of multiple, discrete behaviors that occur over time. Second, the property of behavior to which performance refers is its expected value to the organization.

Two performance criteria – mean performance and performance variation – are among the most commonly studied indicators in pay-for-performance research. According to Reb and Cropanzano (2007), average performance evens out employee's contribution to an organization. Average performance evens out variations from the mean that might be due to passing influences outside the control of the employee.

Following this rationale, past research found that average performance strongly predicts variance in pay and reward allocation (Barnes and Morgeson, 2007; Zhou and Martocchio, 2001), largely because typical, or average, performance represents the dominant conceptualization of performance (Rushton et al., 1981). In the pay-for-performance context, it is eminently reasonable to expect that higher mean performance will lead to positive changes in compensation level.

Performance variation also plays a role in compensation award decisions. Sturman (2007) maintains that long-term or directional changes over time are differentiated from short-term fluctuations, or unsystematic variation of performance. Within-person performance variation can be due to a variety of factors, including affective state (Weiss and Cropanzano, 1996). Between persons, some individuals may show large variation in performance, that is, are inconsistent, whereas others show little variation, that is, perform consistently around the mean level or a long-term trend.

According to Barnes et al. (2012), there are several reasons to expect that performance variation of greater magnitude is associated with smaller compensation. By definition, it is easier to predict the performance of employees who show little performance variability as compared to those who show high variability. Organizations tend to value predictability. Employees performing inconsistently can create uncertainty and disruptions for team members and other parties dependent on the employee, sometimes making it difficult to plan and perform interactively. This increases the risk of performance failures, coordination problems, and disrupted activities for other employees who are downstream in interdependencies.

Moreover, inconsistent performance has been found to lead to attributions of negative traits (Fox et al., 1995) such as undependable. The word undependable, which is one of the

hallmarks of the personality trait conscientiousness (Barrick and Mount, 1991), could be linked to high variability. Empirically, at least one study has found high performance variability to be associated with lower pay (Barnes and Morgeson, 2007). Organizations may pay higher compensation to more consistent performers in an effort to reward and retain these valued employees.

Executive Compensation

From an economic standpoint, the chief executive officer (CEO) is the seller of his or her services, and the compensation committee is the buyer of these services. Under classic economic theory, a reasonable price is obtained through negotiations that are at arm's length between an informed seller and an informed buyer. An awkward situation can result when the CEO hires a professional compensation director or compensation consultant. In this case, the compensation consultant who makes the recommendation to the compensation committee works for the CEO. In theory, the CEO hires the consultant to perform an objective analysis of the organization's executive pay package and to make whatever recommendations the consultant feels are appropriate. This relationship has the potential to promote a conflict of interest because of the perceived pressure for the consultant to protect the CEO's financial interests. The irony is that the consultant is often viewed as representing the shareholders' interests. In a sense, the buyers of the CEO's services are the shareholders and their representatives, the compensation committee of the board of directors. They tend to act upon the compensation consultant's recommendation (Walters et al., 1995).

Different individuals and groups participate in setting executive compensation. They include compensation consultants, compensation committees, and boards of directors. Each plays a different role in setting executive compensation.

Executive compensation consultants usually propose several recommendations for alternate pay packages.

Consultants make recommendations about what and how much to include in executive compensation packages based on strategic analyses, which entail an examination of an organization's external market context and internal factors. Examples of external market factors include industry profile, information about competitors, and long-term growth prospects. Financial condition is the most pertinent internal factor regarding executive compensation. Strategic analyses permit compensation consultants to see where their client organization stands in the market based on external and internal factors. Strong organizations should be able to devote more financial resources to fund lucrative executive compensation programs than weaker organizations. According to Crystal (1991), more often than not, executive compensation consultants find themselves in conflict-of-interest situations: "Ostensibly, compensation consultants were hired by the CEO to perform an objective analysis of the organization's executive pay package and to make whatever recommendations the consultant felt were appropriate. In reality, if those recommendations did not cause the CEO to earn more money than he was earning before the compensation consultant appeared on the scene, the latter was rapidly shown the door" (p. 12).

Executive compensation consultants' professional survival may depend on recommending lucrative compensation packages. Recommending the most lucrative compensation packages will quickly promote a favorable impression of the consultant among CEOs, leading to future consulting engagements.

Since 2008, the US Securities and Exchange Commission rulings require organizations to include the identity of the consulting firm in public disclosure statements. This ruling has created concerns about possible conflicts of interest for consulting firms that also provide consulting services in other areas (e.g., performance management and change management) for the same client firms. Executive compensation consulting represents just one of many possible areas of management consulting. The conflict potentially arises when a consultant intentionally recommends a more-lucrative-than-warranted executive compensation package in the hope of gaining management favor and additional other management consulting opportunities. Of course, compensation consulting firms are concerned about the public image of possible conflicts of interest and have considered a variety of tactics to ensure the integrity of their recommendations to client firms. For example, there has been some speculation that executive consulting practices will be spun off into independent businesses. Many organizations are concerned about their own public image and have instituted policies to prevent compensation consultants from conducting other work for management.

The board of directors represents shareholders' interests by weighing the pros and cons of top executives' decisions. Boards of directors have approximately 15 members. These members include CEOs and top executives of other successful organizations, distinguished community leaders, well-regarded professionals (e.g., physicians and attorneys), and possibly a few top-level executives of the organization.

Boards of directors give final approval of the compensation committee's recommendation. Some critics of executive compensation have argued that CEOs use compensation to coopt board independence. CEOs often nominate candidates for board membership, and their nominations usually lead to candidates' placement on the board. Board members receive compensation for their service to the boards. It is not uncommon for a board member to earn more than $50 000 per year plus a fee ($10 000 or more) for each board meeting attended. Along with monetary and stock compensation, organizations are using such benefits as medical insurance, life insurance, and retirement programs to attract top-notch individuals to join boards of directors. In general, board members' failure to cooperate with CEOs may lead either to fewer benefits or their removal. "The board determines the pay of the CEO. But who determines the pay of the outside directors? Here, a sort of formal Japanese Kabuki has developed. The board of directors determines the pay of the CEO, and for all practical purposes, the CEO determines the pay of the board of directors. Is it any accident, then, that there is a statistical relationship between how highly the CEO is paid and how highly his outside directors are paid?" (Crystal, 1991: pp. 20–21).

Board of directors members within and outside the organization make up an organization's compensation committee. Outside board members serve on compensation committees to

minimize conflict of interest. Thus, outside directors usually are the committee's membership majority.

Compensation committees perform three duties. First, compensation committees review consultants' alternate recommendations for compensation packages. Second, compensation committee members discuss the assets and liabilities of the recommendations. The complex tax laws require compensation committees to consult compensation experts, legal counsel, and tax advisors. Third, based on these deliberations, the committee recommends the consultant's best proposal to the board of directors for its consideration.

Although pay-for-performance is the public rationale for setting executive compensation, reality is often quite different. Three alternative theories explain the principles and processes for setting executive compensation: agency theory, tournament theory, and social comparison theory.

Agency Theory

Ownership is distributed among many thousands of shareholders in large organizations such as Microsoft, General Electric, and the International Business Machines Corporation (IBM). For example, owning at least one share of stock in IBM bestows ownership rights in IBM. Each shareholder's ownership is quite small, with most owning far less than 1% of the total shares. Inability to communicate frequently or face to face to address business concerns is a major disadvantage of thousands of shareholders.

According to *agency theory*, shareholders delegate control to top executives to represent their ownership interests; however, top executives usually do not own majority shares of their organizations' stocks. As a result, executives usually do not share the same interests as the collective shareholders. These features make it possible for executives to pursue activities that benefit themselves rather than the shareholders. The actions of executives on behalf of their own self-interest are known as the agency problem (Jensen and Meckling, 1976). Executives may specifically emphasize the attainment of short-term gains (e.g., increasing market share through lower costs) at the expense of long-term objectives (e.g., product differentiation). Boards of directors may be willing to provide executives generous annual bonuses for attaining short-term gains.

Shareholders negotiate executive employment contracts with executives to minimize loss of control. Executive employment contracts define terms of employment pertaining to performance standards and compensation, specifically current and deferred compensation and benefits. The main shareholder objective is to protect the organization's competitive interests. Shareholders use compensation to align executives' interests with shareholders' interests. As discussed earlier, boards of directors award organization stock to align executives' interests with shareholders' interests.

Tournament Theory

Tournament theory casts lucrative executive compensation as the prize in a series of tournaments or contests among middle- and top-level managers who aspire to become CEOs (Lazear and Rosen, 1981). Winners of the tournament at one level enter the next tournament level. In other words, an employee's promotion to a higher rank signifies a win, and more lucrative compensation (e.g., higher base pay, incentives, enhanced benefits, and perks) represents the prize. The ultimate prize is promotion to CEO and a lucrative executive compensation package. The chance of winning competitions decreases dramatically as employees rise through the ranks: There are fewer positions at higher levels in corporate hierarchical structures.

Social Comparison Theory

According to *social comparison theory*, individuals need to evaluate their accomplishments, and they do so by comparing themselves to similar individuals (Festinger, 1954). Demographic characteristics (e.g., age or race) and occupation are common comparative bases. Individuals tend to select social comparisons who are slightly better than themselves (Festinger, 1954). Researchers have applied social comparison theory to explain the processes for setting executive compensation (O'Reilly et al., 1988).

As we discussed earlier, compensation committees play an important role in setting executive compensation, and compensation committees often include CEOs from other organizations of equal or greater stature. Based on social comparison theory, compensation committee members probably rely on their own compensation packages and the compensation packages of CEOs in organizations of equal or greater stature to determine executive compensation.

National Culture as Context

Cultural values are probably a key factor in the effectiveness of pay-for-performance systems, or, for that matter, whether pay-for-performance is even compatible in particular cultural contexts. Culture is defined as "the human-made part of the environment (Herskovits, 1955). It has both objective elements- tools, roads, appliances- and subjective elements- categories, associations, beliefs, attitudes, norms, roles, and values" (Triandis, 1994: p. 113). Cultural values shape work-related attitudes and behaviors (Triandis, 1994). Culture also influences the domain of normative behavior (e.g., behavior that is desirable versus condemned for members of the culture), defines roles for individuals in the social structure, and prescribes guiding principles and values in one's life. As a result, culture specifies how things in the environment, including an organization's practices, policies, and procedures, are to be evaluated and subsequent reactions to such procedures (Robert et al., 2000).

These cultural values are apparent in compensation systems. The predominant bases for pay in the United States are performance and knowledge, which represents equity (Heneman and Werner, 2005). According to these cross-national studies, there are three primary rules governing individuals' compensation award decisions: equity, equality, and need. The equity rule specifies that individuals should be rewarded on the basis of their contributions such as task performance. The equality rule specifies that individuals should be rewarded equally, regardless of their contributions to the group or the organization. And the need rule states that

individuals should be rewarded on the basis of their personal needs. In general, previous studies, which mostly use undergraduate student samples, have found that individuals in collectivism-dominated nations tended to prefer the equality rule or the need rule, because individuals who hold the collectivist cultural values are concerned with interpersonal relationships. In contrast, individuals in individualism-dominated nations tended to prefer the equity rule, because they are less concerned with relationships and more concerned with task contributions (Bond et al., 1982; Hui et al., 1995; Kim et al., 1990). However, there is at least one notable exception to these findings. Specifically, using a sample of actual employees, Chen (1995) concluded that individuals in China – a collectivism-dominated nation – chose the equity rule, yet individuals in the US – an individualism-dominated nation – chose the equality rule.

For example, just one (among many) noteworthy cultural contrast exists between the US and the People's Republic of China (PRC). The differences between the US market economy and the PRC's centralized government-controlled economy sets the stage for cultural clashes. For decades, the Chinese government owned and operated virtually all business organizations. The Communist party places substantial emphasis on equal contributions to society, group welfare, and the concern for interpersonal relationships. In addition, the Communist party calls for greater emotional dependence of Chinese citizens on their employers. Further, it expects employers to assume a broad responsibility for their members. Zhou and Martocchio (2001) demonstrated support for the difference in ideals when Chinese (employed in the People's Republic of China) and US managers made decisions about monetary and nonmonetary compensation awards.

These ideals are evident in the Chinese workplace and in compensation practices. Employers provide housing and modest wages for food and clothing. The Chinese receive health care under government-sponsored protection programs. Because of the Communist ethic, the Chinese do not identify very well with pay-for-performance programs though that is slowly changing as a capitalist presence creates for-profit organizations.

See also: Job Analysis and Work Roles, Psychology of; Job Satisfaction; Money and Finances, Psychology of; Program Evaluation.

Bibliography

Adams, J.S., 1963. Toward an understanding of inequity. Journal of Abnormal and Social Psychology 67, 422–436.

Adams, J.S., 1965. Inequity in social exchange. In: Berkowitz, L. (Ed.), Advances in Experimental Psychology, vol. 2. Academic Press, San Diego, CA, pp. 267–299.

Barnes, C.M., Morgeson, F.P., 2007. Typical performance, maximal performance, and performance variability: expanding our understanding of how organizations value performance. Human Performance 20, 259–274.

Barnes, C.M., Reb, J., Ang, D., 2012. More than just the mean: moving to a dynamic view of performance-based compensation. Journal of Applied Psychology 97, 711–718.

Barrick, M.R., Mount, M.K., 1991. The big five personality dimensions and job performance: a meta-analysis. Personnel Psychology 44, 1–26.

Becker, G., 1975. Human Capital. National Bureau of Economic Research, New York.

Bond, M.H., Leung, K., Wan, K.C., 1982. How does cultural collectivism operate? The impact of task and maintenance contributions on reward distributions. Journal of Cross-Cultural Psychology 13, 186–200.

Cascio, W.F., Aguinis, H., 2008. Research in industrial and organizational psychology from 1963 to 2007: changes, choices, and trends. Journal of Applied Psychology 93, 1062–1081.

Chen, C.C., 1995. New trends in rewards allocation preferences: a Sino-US comparison. Academy of Management Journal 38, 408–428.

Crystal, G.S., 1991. Why CEO compensation is so high. California Management Review 34, 9–29.

Deci, E.L., Ryan, R.M., 1985. Intrinsic Motivation and Self-determination in Human Behavior. Plenum, New York.

Deci, E.L., Koestner, R., Ryan, R.M., 1999. A meta analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin 125, 627–668.

Eisenberger, R., Cameron, J., 1996. Detrimental effects of rewards: reality or myth? American Psychologist 51, 1153–1166.

Fang, M., Gerhart, B., 2012. Does pay for performance diminish intrinsic interest? The International Journal of Human Resource Management 23, 1176–1196.

Festinger, L., 1954. A theory of social comparison processes. Human Relations 7, 117–140.

Fox, S., Bizman, A., Hoffman, M., Oren, L., 1995. The impact of variability in candidate profiles on rater confidence and judgments regarding stability and job suitability. Journal of Occupational and Organizational Psychology 68, 13–23.

Gupta, N., Shaw, J.D., 2014. Employee compensation: the neglected area of HRM research. Human Resource Management Review 24, 1–4.

Hackman, J.R., Oldham, G.R., 1976. Motivation through the design of work: test of a theory. Organizational Behavior and Human Performance 16, 250–279.

Heneman, R.L., Werner, J.M., 2005. Merit Pay: Linking Pay to Performance in a Changing World. Information Age Publishers, Greenwich, CT.

Herskovits, M.J., 1955. Cultural Anthropology. Knopf, New York.

Hui, H., Yee, C., Eastman, K.L., 1995. The relationship between individualism–collectivism and job satisfaction. Applied Psychology 44, 276–282.

Jensen, M.C., Meckling, W.H., 1976. Theory of the firm: managerial behavior, agency costs, and ownership structure. Journal of Financial Economics 3, 305–360.

Kim, K.I., Hun-Joon, P., Nori, S., 1990. Reward allocations in the United States, Japan, and Korea: a comparison of individualistic and collectivistic cultures. Academy of Management Journal 33, 188–198.

Krefting, L.A., Mahoney, T.A., 1977. Determining the size of a meaningful pay increase. Industrial Relations 16, 83–93.

Latham, G.P., Yukl, G.A., 1975. A review of research on the application of goal setting in organizations. Academy of Management Journal 18, 824–845.

Lawler, E.E., 1971. Pay and Organizational Effectiveness: A Psychological View. McGraw-Hill, New York.

Lazear, E., Rosen, S., 1981. Rank-order tournaments as optimum labor contracts. Journal of Political Economy 89, 841–864.

Locke, E.A., 1968. Toward a theory of task motivation and incentives. Organizational Behavior and Human Performance 3, 157–189.

Locke, E.A., Feren, D.B., McCaleb, V.M., Shaw, K.N., Denny, A.T., 1980. The relative effectiveness of four methods of motivating employee performance. In: Duncan, K.D., Gruenberg, M.M., Wallis, D. (Eds.), Changes in Working Life. Wiley, New York, pp. 363–388.

Martocchio, J., 2015. Strategic Compensation: A Human Resource Management Approach. Pearson Higher Education, Upper Saddle River, NJ.

Motowidlo, S.J., 2003. Job Performance. Handbook of Psychology, second ed. Wiley, New York. pp. 82–103.

O'Reilly III, C.A., Main, B.G., Crystal, G.S., 1988. CEO compensation as tournament and social comparison: a tale of two theories. Administrative Science Quarterly 33, 257–274.

Reb, J., Cropanzano, R., 2007. Evaluating dynamic performance: the influence of Salient Gestalt characteristics on performance ratings. Journal of Applied Psychology 92, 490–499.

Robert, C., Probst, T.M., Martocchio, J.J., Drasgow, F., Lawler, J.J., 2000. Empowerment and continuous improvement in the United States, Mexico, Poland, and India: predicting fit on the basis of the dimensions of power distance and individualism. Journal of Applied Psychology 85, 643–658.

Rushton, J.P., Jackson, D.N., Paunonen, S.V., 1981. Personality: nomothetic or idiographic? A response to Kenrick and Stringfield. Psychological Review 88, 582–589.

Rynes, S.L., Gerhart, B., Parks, L., 2005. Personnel psychology: performance evaluation and pay for performance. Annual Review of Psychology 56, 571–600.

Sturman, M.C., 2007. The past, present, and future of dynamic performance research. Research in Personnel and Human Resources Management 26, 49–110.

Triandis, H.C., 1994. Cross-cultural industrial and organizational psychology. In: Triandis, H.C., Dunnette, M., Hough, L. (Eds.), Handbook of Industrial and Organizational Psychology, second ed., vol. 4. Consulting Psychologists Press, Palo Alto, CA, pp. 103–172.

Walters, B., Hardin, T., Schick, J., 1995. Top executive compensation: equity or excess? Implications for regaining American competitiveness. Journal of Business Ethics 14, 227–234.

Weiss, H.M., Cropanzano, R., 1996. Affective events theory: a theoretical discussion of the structure, causes and consequences of affective experiences at work. Research in Organizational Behavior 18, 1–74.

Zedeck, S., Smith, P.C., 1968. A psychophysical determination of equitable payment: a methodological study. Journal of Applied Psychology 52, 343–347.

Zhou, J., Martocchio, J.J., 2001. Chinese and American managers' compensation award decisions: a comparative policy-capturing study. Personnel Psychology 54, 115–145.

Peace

Johan Galtung, TRANSCEND Peace University

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Peace is seen as a transnational, transdisciplinary, and translevel – from inner peace to civilizations – concept. There are two aspects: negative peace as absence of violence, and positive peace as cooperation and harmony. A formula is presented based on equity and harmony for positive peace, and reconciliation of traumas and resolution of conflicts to avoid violence for negative peace.

Summary and Overview

The point of departure for peace is the broad, basic suffering fulfillment distinction, for all life. The double task of peace is to reduce the suffering by violence – negative peace – and increase the fulfillment – positive peace. The two can be pursued together.

One basic diagnosis for violence is underlying, *unresolved conflicts at present*, simple, complex, or structural (see Johan Galtung, *A Theory of Conflict*, TRANSCEND University Press, 2010; Parts Two, Three, and Four respectively). The parties are coupled; the task is to change this coupling from incompatibility to compatibility of goals, or to no coupling at all.

Another basic diagnosis for violence is *unconciliated traumas from the past*, inflicted by the parties in the coupling, or by some other party. The task is to clear the past, closure, and create a future together. But reconciliation should not serve as a pretext to neglect the conflict resolution job lest it becomes pacification.

In addition, there is the equity job to take care of, for not only mutual, but not too unequal fulfillment, harmony, and emotional resonance. But these two positive peace tasks may come to nothing if traumas and conflicts are not adequately handled.

Peace politics is the transformation from direct and structural violence to negative and positive peace by peaceful means. Nonviolent struggle, mediation, and reconciliation can be substituted for war, guerilla-terrorism, and negotiation to win. The goals are negative and positive peace, working on both together, not one after the other.

That opens for four stages beyond negative peace: symbiosis, transcendence, synthesis, and absorption. How far we want to go depends on the concrete case, but as a minimum there should be some *project*, some *peace construction*, with *cooperation* for mutual and equal benefit and *harmony*, attitudinal resonance. Transcendence and synthesis may come when old identities are waning and the capacity to handle equity and new conflicts is waxing. Peace is a never-ending process.

What Is Peace? An Operational Definition

Peace cuts across history, geography, and levels of human organization; from inner and interpersonal peace via states and nations to regional and civilizational relations. Definitions should reflect positive peace as cooperation and harmony, and negative peace as absence of violence; the latter not only as absence of direct, intended violence, acts of commission, but also of indirect structural violence by acts of omission, and by cultural violence, justifying one or the other. Peace is incompatible with massive, lethal, and social inequality – like health is incompatible with tumors and circulation problems with no overt pain or fever. Both are ideals, sums of all good things.

Definitions should serve thought, speech, and indicate action. This formula defines peace operationally, in terms of four tasks (the formula logic is structural peace – equitable cooperation – and cultural peace – empathic harmony knowing and feeling the reality of other – threatened by direct violence from unreconciled trauma and unsolved conflicts. For more details, see Johan Galtung, *A Theory of Peace*, TRANSCEND University Press, 2013, Part One. For more about the formula, see Johan Galtung and Dietrich Fischer, *Peace Mathematics*, TRANSCEND University Press, 2012, pp. 28–32.):

$$\text{Peace} = \frac{\text{Equity} \times \text{Harmony}}{\text{Trauma} \times \text{Conflict}}$$

The formula applies across time, space, level, from peace with nature, and inner peace to, for some, peace with their God. Needless to say, no academic discipline, or civilization, has any monopoly on 'peace.'

There are two factors in the numerator: *the more the better*; and two in the denominator leading to violence, *the less the better*. The definition relates to the reality of peace, not to literature about peace and is operational, neither semantic, nor 'essentialist.'

The four factors or tasks can be spelt out:

1. *More equity*: cooperative behavior for mutual and equal benefit
2. *More harmony*: empathic attitudes of emotional resonance, in the daoist sense of enjoying the joy, and suffer the suffering, of other
3. *Less trauma*: reconciliation, clearing the past, acknowledging wrongs, wishing them undone, and dialogs to create a future together
4. *Less conflict*: resolution, acceptable and sustainable, making incompatible goals compatible, and changing negative attitudes and behavior

Using marriage as an example, parity and empathy are basic, but unreconciled traumas from the past and unresolved conflicts may cause verbal or physical violence and cannot be left unattended.

Political examples of the four tasks carried out successfully:

Equity: the 1950–58–92 French–German, then as European Economic Community (EEC), then as European Union (EU), European cooperation on reasonably equal terms; now (2013) threatened by the economic inequality between Germany and the EU periphery;

Harmony: the feminist revolution has made many men understand how the man–woman divide or fault line looks from a woman's point of view; women being familiar with the prejudices associated with patriarchy.

Reconciliation: the South African Truth and Reconciliation Commission exchanging confession of political crimes (and contrition and compensation, if possible) for amnesty.

Resolution: black–white relations in the United States, moving toward racial equality of opportunity.

How good EU is at the other three is now being tested; the same applies to the feminist revolution, race in South Africa and the United States. Peace as defined operationally by the formula is a demanding agenda.

Major insights in *how to* are provided by worldviews, religion, and philosophy, but they are not unproblematic. Thus, reconciliation in South Africa built on the Zulu *ubuntu*, "I exist because you exist, we are in each other," defining a strong spiritual coupling in a *we-culture*. The unit of suffering and joy is a we, not two separate, parallel I's; problematic to an individualistic West that may prefer negative peace to being coupled together in positive peace.

Compare this to the *Golden Rule*, positive (or negative):

Do (do not) unto others what you (do not) like them to do unto you. The subject for violence avoidance and peace is an individual "you"; the ethical budget is egocentric I-culture, not collective we-culture. In addition, be careful: "their tastes may be different" (GB Shaw).

The same applies to the *Kantian categorical imperative*: "Act so that the principle of your intention always could serve as the basis for world law" (be generalizable, universalizable). This is a traffic rule for coexistence: Kant wanted a multilateral normative umbrella, not only *Golden Rule* bilateralism. Both are useful rules for positive coexistence in an I-culture, but do not produce the resonance of a we-culture. Or Mo Tzu's universal love as opposed to partiality, (Mo Tzu (one of the 'hundred philosophers' 551–233 BC) argues universal love, also of one's enemies, as opposed to the partial love only of one's friends. The rulers must start: "Now universal love and mutual benefit are both profitable and easy beyond all measure. The only trouble, as I see it is that no ruler takes any delight in them. If the rulers really delighted in them, promoted them with rewards and praise, and prevented neglect of them with punishments, then I believe that people would turn to universal love and mutual benefit as naturally as fire turns upward or water turns downward, and nothing in the world could stop them." *Mo Tzu*, *Basic Writings*, New York, London: Columbia University Press, 1963, p. 49.) long before the universalism of Christianity, Kant, or human rights.

Being well together by fulfilling each other is *positive peace*; with mutual assured bliss (MAB), as the extreme – an abstraction found in religious discourses but also in love defined as the union of the unions of body, mind, and spirit. Inflicting suffering on each other is typical of *direct violence and war*, with the mutual assured destruction (MAD) of the Cold War nuclear arms race, as the extreme; not an abstraction. One doing well and the other doing badly is typical of *structural violence*. Neither fulfilling each other, nor inflicting suffering, is typical of *negative peace*: there is passive peaceful coexistence, neither violence, nor positive peace; mutual irrelevance.

Peace politics can be seen as intended efforts to reduce direct and structural violence for negative peace – not accepting exploitation of one by the other as 'peace' even if there is no verbal or physical violence – moving on to positive peace with ever more cooperation for mutual and equal benefit, and more empathy for harmony; feeling each other's suffering and fulfillment as parts of one's own. Two traps on the way: ignoring exploitation and repression as justified by cultural violence – economic and cultural powers are seen as 'soft power' – and being satisfied with the negative peace of mutual irrelevance. (Mutual isolation is problematic in a globalizing world with the means of transportation–communication reaching every corner of the world; sometimes as means of structural violence like means of destruction are the means of direct violence; as important in human history as Marx' means of production.) If past traumas and present conflicts are not mediated, negative peace will be lost, and unless the coupling is not equitable and harmonious, there will be no positive peace, and sooner or later acts of violence.

Peace and Related Concepts

Security stands for the absence of direct violence, the threat thereof, and absence of efforts to change the 'normal' interaction and division of labor between persons and between states. Security stands for *stability*, *status quo*, law and order; generally favored by those who benefit from that concept. Positive peace with equity is too revolutionary, standing both for equality and a togetherness close to fusion, like in a good marriage or community. Absence of positive peace opens for freedom, for exploitation, and for freedom to exploit.

The Japanese concept of *kyosei* (see Yoichiro Murakami and Thomas J. Schoenbaum (eds.) *A Grand Design for Peace and Reconciliation: Achieving Kyosei in East Asia*. Cheltenham, UK, and North Hampton, Massachusetts, USA: Edward Elgar, 2008), conviviality, comes closer to positive peace with less emphasis on equity; mutual fulfillment, but not necessarily equally much fulfillment. There is much of that in the classical Japanese company; less in concrete Japanese foreign policy, without reconciliation, resolution, and equity.

Article 9 in the constitution, depriving Japan of the right to offensive war, is compatible with no joint projects like an East Asian Community, no harmony, no reconciliation with the victims of Japanese imperialism. The preamble mentions trade, but Japanese trade is not known for equity. The peace concept underlying *A9* was only negative.

Human security is more problematic. 'Human' has positive connotations, but 'security' has the negative connotations mentioned, including military intervention. 'National security' also conveys an image of a hedgehog turning inward with noninviting quills outward, attending to the inside at the neglect of the excluded outside. But inside seven types of human security may be met: food, health, environment, personal, economic, community, and political security. (See, for instance, the *Philippine Human Development Report 2005*,

p. 5.) However, there is no *kyo*, con, sym, cocreation with others; it comes from the inside and from above. That is where 'mutuality and equality' enters the peace concept, like in the *mutual and equal benefit* of the *Panchshila* of Zhou Enlai and Nehru, for China–India, in 1954.

The military, extending a violent arm to violent arms: depends on how far the arms reach. If only within the country, the arms are defensive of Self; if also outside they are offensive, violent to Other. Even if not used the threat of use is direct psychological violence to the mind of Other as fear; and to the spirit of both by impeding positive peace. Defensive defense neither offends, nor provokes; is neither violence, nor positive peace, but the negative peace of mutual security.

Justice is ambiguous, divided into *criminal* and *social* justice, the former into *punitive* and *restorative*, and the latter into *distributive* and *equitative* justice. Depends on what is meant.

Under *punitive justice* an actor is arraigned into court for adjudication. If found guilty, punishment, meaning violence, follows.

Under *restorative justice* an actor who has broken a law will be attempted restored to society, for instance through community work. This concept could be extended to guilty states, doing 'world work.'

Under *distributive justice* the focus is on 'equality' as equal opportunity; no correlation with gender, race, etc.; low dispersion of wealth; lifting the social bottom up. All contribute to equality.

Under *equitative justice* the focus is on patterns of interaction (like parity in marriage) that generate mutual and equal benefit.

Punitive justice as victor's justice are continuations of war by legal means; no healing traumas, solving conflicts, and building projects.

Restorative justice may help reconciliation, solve conflicts, and open for projects, like inviting Germany into the European Community.

Distributive justice will promote equality, which facilitates dialogue between equals, empathy, nonviolence, equity, and harmony.

Equitative justice: builds peace directly into the structure.

Punitive justice may prolong violent struggle to deny the other side victor's justice, and be counterproductive when used against nonviolent struggle. The other three may be peace-productive.

Rule of law is also ambiguous, like the Ten Commandments strong on forbidding acts of commission, but permissive of acts of omission, all those doing nothing. This opens for huge structural violence justified by cultures weak on prescribing the good acts of commission. Thus, "Thou shalt not lie" changes character when adding "but share generously your truths with others" (as opposed to patents). For world rule through world law a more positive, less punitive, rule of law is needed; with rewards, not only punishment. Moreover, social ills like slavery and colonialism were not so much outlawed as being overtaken by other structures, some of them liberating, others not.

Human rights are better, but to the individualist focus peoples' rights have to be added, like the right of villages to survive as human habitats, of traditional crafts to survive, and of clans and tribes to be recognized as juridical persons. The right of (some) nations to (some) self-determination is not enough. A better balance between individual and collective rights is needed for human rights to serve, as they 1 day probably will, as a world constitution.

Democracy matters, but to the individualist one-person one-vote focus and majority rule – after debate to win that majority – dialog to consensus, more satisfactory to we-cultures, should be added. For world democracy add a democratic UN without big power veto, and an elected UN Peoples' Assembly with the member states as constituencies.

Ho'o pono pono is an ingenious Polynesian reconciliation and resolution method, focusing less on perpetrator and victim and more on the community as the basic victim, unable to prevent violence because of too many doing nothing. These acts of omission are identified by the process, and then turned into positive acts of commission.

Thus, there are many peace-related concepts. The purpose of such exercises is to identify what there is of peace in a concept, and to identify what is missing; in a never-ending concept dialog.

Negative Peace: The Absence of Violence

A year with one war between states is often seen as a war-year, not counting all the peaceful relations, the days not at war, and the people on both sides not participating. We have become so mesmerized by war and blind to the prevalence of peace as absence of violence – even if only because of mutual irrelevance and distance – as the general rule for humans that we do not even see it. But such errors keep the lie about innate violence by 'human nature' alive, giving humans a bad name, and a worse reputation than humans deserve. (The UNESCO *Sevilla Declaration* is one way of doing this, declaring that there is no scientific basis for the idea that violence-war is innate. Missing in the declaration is the idea that there are positive ways of handling trauma and conflict, may be not innate either, but like war easily learnt.)

Unresolved conflict has been identified as a key cause of violence, and solution as the remedy. Violence is to hurt and harm to the point of incapacitating Other(s); conflict is incompatibility in the means-ends complexes pursued and used by the parties. From "Other pursues his goals with his means and impedes me from doing the same" does not follow that "Other stands in my way," but that "Other's goals stand in the way of my goals." A puzzle in search of a solution, and there seem to be five approaches, starting with the violent ones.

No. 1, *modern war*, to win and dictate the solution, is now waning between states: destructive power is too high and invulnerability too low; hence the use of state terrorism, unleashing destruction on soft targets far away, often unable to hit back.

No. 2, *guerilla, terrorism*, to win and dictate the solution, less lethal than state terrorism, takes more time; but the asymmetry seems to favor guerrilla as judged by Viêt Nam (The classical exchange when levels 1 and 2 Vietnam war turned into level 3, negotiation between United States and Viêt Nam: "You were never able to beat us in open battle!", "Correct. But it is immaterial." It is worth remembering that at level 2 there is no beginning and no end, nor formal declaration of war nor capitulation. That plays in their favor, exhausting Other's resilience.), Afghanistan, Iraq, etc. It is the continuation of war by informal means.

No. 3, *negotiation*, is the continuation of war by verbal means, *debate*, ideally so that the strongest arguments based on data, values, and theory win. There is polarization between two sides of the table, and winners and losers, like in Nos. 1 and 2.

No. 4, *negative nonviolence*, like the multiseason Arab spring, is the continuation of war by such nonviolent means as noncooperation and civil disobedience, with winners and losers. As revolt effective, as revolution, changing the system, too shallow, and too negative.

No. 5, *positive nonviolence*, conversion, not coercion, bridging polarization (Polarization in "us vs them" glorifies us and demonizes, dehumanizes them to objects not protected by human empathy with each other. Depolarization is an important part of the work for negative peace, and cooperation on conflict solving is already a first step for positive peace. The Tajfel–Turner and Tropp–Pettigrew traditions in social psychology have much to say about social identity and intergroup relations, but less on identifying and mediating conflict and trauma.) in *dialogue* searching together for solutions, without or with mediator, *mapping* the conflict – who are the parties, what are their goals and means – *testing* the goals for legitimacy – legal? human rights? basic needs? – and *bridging* between legitimate goals, jointly searching for a new reality. It calls for peace education in schools, peace journalism in the media, peace in the arts; like No. 1 calls for training to kill, militarized universities, No. 2 for camps, etc.

No. 4 for struggle and No. 5 for solutions are normal human ways; the others are aberrations, and should be known as such. (The rest is a question of knowledge and skills. This author's contributions are in the two books *Transcend & Transform*, 2003, and *50 Years: 100 Peace & Conflict Perspectives*, 2008.)

Unconciliated trauma; wounds from a conflict, even forgotten, have also been identified as cause of violence. Five approaches (And the book corresponding to the preceding paragraph is Joanna Santa Barbara, Johan Galtung, Diane Perlman, *Reconciliation*, TRANSCEND University Press, 2012.):

No. 1, *revenge* according to the classical rule of *lex talionis*, an eye for an eye, etc., carried out by the victim, or by the state.

No. 2, *displaced aggression*, violence in any direction, also against self. Violence leads to new trauma leads to violence leads to trauma *ad inf.*, like family feuds, *vendetta*; no end, no closure.

No. 3, *negotiation*, extracting compensation from the perpetrator also in monetary terms. Superficial, but may lead to shallow closure.

No. 4, *visibility*, "see what you have done to me"; museums, parks, monuments for traumas already suffered, may lead to shallow closure.

No. 5, *conciliation*, opens for the whole complexity, conscious and subconscious. If the perpetrator is traumatized by traumatizing the victim, they may have a joint interest in clearing the past by wishing it undone, working for an acceptable future, and for closure.

No. 5 one way or another is the normal human way; the other four are aberrations, and should be known as such (see Joanna Santa Barbara, Johan Galtung and Diane Perlman, *Reconciliation: Clearing the Past, Opening the Future*, TRANSCEND University Press, 2012).

Positive Peace: The Presence of Cooperation and Harmony

Level I: *Symbiosis = Equity + Harmony*. There is cooperation for mutual and equal benefit, and there is emotional resonance. Another term would be active peaceful coexistence, as opposed to the passive peaceful coexistence of negative peace.

Level II: *Transcendence: A new actor is born*. The parties are still there, but something new has emerged. In the integration language for states they start "talking with one voice," "having a joint foreign and security policy," etc. The term often used for level I is community, confederation (German: *Staatenbund*); and for level II the terms are union, federation (German: *Bundesstaat*).

Transcendence means something above, beyond; among states, multistate conferences, a permanent secretariat, a commission that becomes more than the sum of the actors. The process can be seen as follows:

1. Negative peace: separate goals, separate means; avoiding violence.
2. Positive peace I: separate goals, shared means-resources.
3. Positive Peace II: shared goals, shared means-resources.

In cohabitation the parties share their resources, but preserve the autonomy of having separate goals. In a marriage they share a goal, building a happy marriage, beyond the sum of the two parties.

That brings us to the next level:

Level III: *Synthesis*. Feudal counties, duchies, etc. die; a state is born. States gradually die, regions are born. Regions gradually die, a world is born. We are right now living the drama of all these processes in different parts of the world.

The resources, the means, have already been pooled; the goals have been aligned. They are not only compatible, but identical. Out of numerous I-ness's a new we-ness, of a new kind, has emerged. The old actors, like old people, have had their days, but the new actors are not like a younger generation merely replicating the old; but a transition to a new era in human history.

Level IV: *Fusion, absorption*, found in the religious discourses, but often reserved for the beyond, the afterlife, for souls only. The ultimate absorption-union is with God in Christianity and Islam, with each other in the Hindu nirvana concept, with the network of all life in Buddhism. But the human individual spirit is capable of level IV without dying, as absorbed, fully devoted, in the nation, in the civilization, in humanity; in a marriage with a fusion of partners.

There is ambiguity in positive peace. At levels I and II between states, for instance, war has become 'unthinkable' if the integration manages to handle, constructively, complex conflicts with high numbers of parties and goals. But that is a huge order. If unable to handle conflicts, the coupling may break up, parties may decouple, hopefully only down to level 0, negative peace, not further down to violence.

The real crisis may come at level III, let alone level IV. As the process proceeds, the identity of the actors is eroding. Thus, for the United Kingdom, France, and Germany to be absorbed into Europe, with no identity of their own, will be hard, let alone the United States under UN.

The solution lies in seeing the levels less as linear progress, more as parallel processes. A couple will delight in the union of

unions of body, mind, and spirit as bliss, but find it exhausting and proceed with one at the time. A political union may be needed in crisis, but revert to level I under more normal conditions.

The levels are modes of coexistence, not life sentences.

Negative peace as mutual assured survival (I am indebted to Diane Perlman for this felicitous term, a common sense between MAB and MAD) is attractive in its simplicity, but positive peace has much more to offer. Active peaceful coexistence imposes mutual rights and obligations on us, the trick to make the relation peaceful lies in the equity and harmony. The parties will be well advised to train in both and not go for higher levels of positive peace than they can master. And Western individualism and state sovereignty may impede positive peace.

Massive Category Killing as the Extreme Peace Negation

Concepts are better understood contrasted with their negations. The Black Plague stimulated health studies, and twentieth-century genocides stimulated peace studies; not only *Shoa*, but also MAD, by the logic of the Cold War nuclear arms race. The nuclear arms race as structure was blamed, not an evil perpetrator against innocent victims. This discourse switch from evil party to bad relation-orientation was among the factors ending the Cold War.

1. *Definition*: Genocide = Massive Category Killing, MCK, meaning by.
2. Massive = All of a kind available, till exhaustion–saturation.
3. Category = The same ascribed or achieved characteristics.
4. Killing = Avoidable life deprivation, by commission or omission.

"There is no good German but a dead German" are words typical of wars. The target is 'all of a kind,' defined by ascribed or achieved attributes; with massive air raids, year after year. Or massive ethnic cleansing, year after year. By the MCK definition both are genocides. Lemkin's (Raphael Lemkin, 1900–59, Polish lawyer) focus on *genus*, people, opens for that. But the massive killing of Germans was done by World War II (WWII) winners, of Jews by WWII losers; ethnic cleansing of Moors and Jews in Spain from 1492 was by victors, Serbian acts in Yugoslavia by losers. Losers commit genocide, winners not; 'massive category killing' covers both.

The 'category' is a set of humans, violence – like peace – is a relation, the bridging concept is the *fault line* between categories:

1. *Definition*: Fault line = division between categories of humans.

Across fault lines massive structural violence may grow, exploiting, repressing, alienating; with sparks of massive direct violence. The genocide concept focuses on the effect, less on the fault line cause.

The atrocious effect comprises all three types of violence:

1. Direct violence = Intended, by *actors* hurting–harming–killing.
2. Structural violence = Unintended, by *structures* hurting–harming (exploitation, repression, and alienation).
3. Cultural violence = by *cultures* legitimizing either violence.

The cultural violence justifies the direct and structural violence that causes each other.

There are at least eight fault line causes with such effects: humans/nature, men/women, middle-aged/young/old, white/colored, 'normal'/deviants, high/low class (with four types of class: economic, military, political, cultural; generally powerful vs powerless); dominant/dominated nations; center/periphery states.

If we combine the fault lines with the two forms of violence, direct actor violence and indirect structural violence (cultural violence is not itself violence, but very important as it legitimizes the other two), we get:

A basic paradigm for massive category killing

Private no.	*Fault lines*	*A: actor violence*	*S: structural violence*
1	Nature	Slaughter of 'beasts, savages, wilderness'	Depletion pollution
2	Gender	Killing women: abortion, infanticide, witch-burning	Patriarchy as prison of women, putting them 'in their place'
3	Generation	Privileging middle-aged	Schools as ghetto Homes as ghetto
4	Race	Controlling dangerous races	Eradication slavery Colonialism slavery
5	Deviance ● Criminal ● Mental ● Somatic	Controlling dangerous persons	 Euthanasia exclusion Euthanasia exclusion Euthanasia exclusion
6	Class ● Military ● Economic ● Political ● Cultural	Controlling dangerous classes Elimination	 Exploitation of body Repression of mind Alienation of spirit
7	Nation Culture Ideology	Genocide narrowly defined	State as prison of nations ruled by 'majority'
8	State Country Territory	War (killing for food, sacrifice, conquest)	Imperialism Isolating 'Pariah' states

Here are 16 types of MCK by pure logic, not by ideology. To focus on 7A only is selective, ideological.

The extreme peace negation, the total Table 7A, is atrocious; but 6-7-8S, massive structural killing of the poorest in periphery countries, is worse in numbers. And 2A, killing unborn and newborn girls, has by far outdone 7A. The focus in the table is moving over time, but one factor remains: 1A and

1S against animals may serve as training in taking life, an argument for vegetarianism.

The dominating world elite is human, male, middle-aged, white, 'normal,' powerful, from dominant nations (like Anglo-American), from center states (like United Kingdom–United States).

Today that elite is challenged by all its fault line opposites, including Mother nature. But the discourse is still focused on 7A, with cases like Armenia, Nanjing, Gulag-KZ, *Shoa*, Hiroshima, Indonesia, Cambodia, Rwanda; making actors visible as perpetrators and victims, but structures invisible as any focus on their victims – women, colored, etc. – might lead to deeper scrutiny of that elite.

To restore law and order they bring perpetrators to punitive justice – tribunals – and the guilty are punished, even killed. But adjudication also serves to acquit the judges, the non-indicted, structures and cultures, and future generations. The focus is narrow.

Some waning discourses still legitimize omitting other victims:

Nature: because animals are seen as soulless, feeling no pain and no bereavement, having no families.

Gender: because women were seen as born inferior to bear children, as emotional not intellectual, as temptresses, as being of Satan.

Generation: because the old generation has to make way for the next, and the young generation has to wait and first be adequately trained.

Race: because primitives are seen in the same way as animals or women.

Deviance: because criminals have excluded themselves, the mentally ill are unfit, the somatically ill are a burden to themselves and society.

Class: because class sorts the fittest from the less fit.

Nation: because some nations are seen as up front in human history, others as lagging behind, having to wait to be adequately trained.

States: because states (except Japan) have the right to war, and wars identify the fittest states to rule the world from the less fit.

There is also cultural violence built into academic disciplines making some of the 16 invisible: international studies covers only 8A, are often structure blind, like international (actually interstate) Law; best at 6A and 8A, but only as security studies. Lemkin's 7A insistence was at the time a breakthrough, and resisted.

The micro- and mesolevel social sciences, psychology–sociology, have made basic contributions studying prejudice and discrimination as cultural and structural violence; but rarely touching the MCK drama.

This selective approach to violence leads to a selective approach to peace, focused on actor, not structural violence, with a decreasing priority in the order $8 > 7 > 6 > 5 > 4 > 3 > 2 > 1$.

Peace studies, spanning all levels of human organization and all types of violence and peace, must cover all 16 (and not only 7A and within 7A *Shoa* only, excluding the suffering of other peoples (Thus, *nakba*, the horrors suffered by the Palestinians at the hands of the Israelis, from 1948 up till today, was massive and directed against a category, but not MCK. What was killed was and is a society; not genocide, but *sociocide*.)).

And peace politics has to dampen all fault lines through equity and harmony to overcome discrimination and prejudice, reconciliation for past atrocities, and creative conflict resolution for the present.

This is what peace theories and practice are up against, like health theories and practice are up against *massive lethal diseases*, contagious from the outside as epidemics, even pandemics; structural from the inside as cancers, circulation diseases, and mental disorders with an 'epidemiological shift' from the former to the latter; for health, as for peace from direct to structural violence. States are waning in salience, nations not; interstate wars are less frequent (like killing of monarchs with monarchy waning or like wars of liberation from colonialism as classical colonialism no longer matters. Interstate wars show signs of decline because the states, except for the biggest, and with the states the state system, decline in salience as they get absorbed into regions. Steven Pinker, *The Better Angels of Our Nature: Why Violence Has Declined*, London: Penguin, 2011 adds important psychology to historiography, but his violence and peace concepts do not reflect contemporary reality.) but massive inequality is increasing, changing the focus in the peace formula to equity.

See also: Conflict and Conflict Resolution, Social Psychology of; Conflict and War, Archaeology of: Weapons and Artifacts; Genocide and War; Peace Movements; Peace Processes; Peace Promotion, Psychology of; Peace and Nonviolence: Overview of Anthropological Aspects; Peace-Making in History; Slavery and Violence; War, Social Causes and Consequences; Warfare in History; Youth, Music, and Peace Building.

Bibliography

Adolf, Antony, 2009. Peace: A World History. Polity Press, Cambridge.
Brewer, John D., 2010. Peace Processes. Polity, Cambridge.
de Figueiredo, Fidelino, 1966. Ideais de Paz. Portugalia, Lisboa.
Galtung, Johan, 1996. Peace by Peaceful Means. SAGE, London.
Galtung, Johan, 2003. Transcend & Transform. PLUTO, Paradigm.
Galtung, Johan, 2008. 50 Years: 100 Peace & Conflict Perspectives. TRANSCEND University Press, Oslo.
Galtung, Johan, 2010. A Theory of Conflict. TRANSCEND University Press, Oslo.
Galtung, Johan, 2013. A Theory of Peace. TRANSCEND University Press, Oslo.
Galtung, Johan, Fischer, Dietrich, 2012. Peace Mathematics. TRANSCEND University Press, Oslo, pp. 28–32.
Giorgi, P.P., 2008. La Violenza Inevitabile: Una Menzogna. JACA, Milano.
Kant, Immanuel, 1992. Perpetual Peace. Hackett, Indianapolis/Cambridge.
Muñoz, Martínez (Ed.), 2000. Historia de la Paz. Eirene, Granada.
Murakami, Yoichiro, Schoenbaum, Thomas J. (Eds.), 2008. A Grand Design for Peace and Reconciliation: Achieving Kyosei in East Asia. Edward Elgar, Cheltenham, UK, North Hampton, USA.
Pinker, Steven, 2011. The Better Angels of Our Nature: Why Violence Has Declined. Penguin, London.
Santa Barbara, Joanna, Galtung, Johan, Perlman, Diane, 2012. Reconciliation. TRANSCEND University Press, Oslo.
Tzu, Mo, 1963. Basic Writings. Columbia University Press, New York, p. 49.

Peace and Nonviolence: Anthropological Aspects

Leslie E Sponsel, University of Hawaii, Honolulu, HI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The accumulating scientific evidence proves beyond any doubt that nonviolent and peaceful societies not only exist, but are actually the norm throughout human prehistory and history. This article demonstrates and elucidates this scientific fact through a historical inventory of the most important empirical documentation and a summary of the most seminal cases. Next it reveals the significance for anthropology of the provocative new perspective and possibilities of the conceptual framework of nonkilling and of the transformation toward that ideal. Finally, it briefly discusses the obstacles to further advancing this fertile and crucial research frontier while refuting the spurious assertions of opponents.

Introduction

Nonviolence and peace are natural, ubiquitous, and normal in the human species throughout its evolution and adaptation. In fact, for well over a million years evolution has selected for a human nature that is naturally inclined toward nonviolence, peace, cooperation, reciprocity, empathy, and compassion in the context of nomadic forager societies that comprise 99% of human existence. Otherwise the species would not have survived and flourished.

Serial and mass murder, and especially in the form of warfare justified by whatever rationalizations, are relatively recent aberrations in human cultural evolution and prehistory, and are uncommon in history as well. Normal refers to the statistical sense of nonviolence and peace as common. Normal also refers to the moral sense in that nonviolence is valued as an ideal, what is right. All of the major religions have core principles proscribing killing and prescribing treating others as one would like to be treated. Accordingly, nonviolence and peace are not rare, although they are rarely acknowledged, this despite the fact that there has been a progressively accumulating record of publications about them since at least the nineteenth century and especially over the last three decades. Biases in Western culture, and especially in American culture, tend to view humans as innately competitive and aggressive, thus neglecting if not completely ignoring nonviolence and peace.

Anthropology of Peace Founder

Ashley Montagu (1905–99) may be considered the founder of peace studies in anthropology. A humanist as well as a scientist specializing in biocultural anthropology, he was a prolific writer and prominent public intellectual. Montagu's (1976) *Nature of Human Aggression* is one of the earliest rebuttals of the genre of biological reductionism and determinism reflected in such pseudoscientific stereotypes as man the hunter, man the warrior, killer apes, beast within, and demonic males. He dismissed such stereotypes as 'the myth of innate depravity.'

Montagu compiled the first set of ethnographic case studies on nonviolent and peaceful cultures in the edited book *Learning Non-Aggression* with chapters about the Fore, !Kung San, Inuit, Semai, Australian Aborigines, Mbuti, and Tahitians. In the Introduction Montagu (1978: p. 5) lists the Punan, Hadza, Birhor, Todas, Veddahs, Arapesh, Yamis, Tikopia, Land Dayaks, Lepchas, Papago, Hopi, Zuni, and Ifaluk as additional nonaggressive societies. He observes that given the existence of such nonaggressive societies plus the wide range of aggression among other societies it is clear that aggression is not a biologically determined innate drive, but learned cultural behavior. Furthermore, he mentions that many nonliterate peoples have readily abandoned fighting, raiding, and warfare (p. 9).

Accumulating Documentation

Since Montagu's pioneering anthology of more than three decades ago, a substantial literature has been accumulating and advancing further documentation, including at least five sets of ethnographic case studies (Evans Pim, 2010; Fry and Bjorkqvist, 1997; Howell and Willis, 1989; Kemp and Fry, 2004; Sponsel and Gregor, 1994), four major monographs (Dennen, 1995; Fry, 2006; Kelly, 2000; Melko, 1973), an annotated bibliography (Bonta, 1993), and an encyclopedic Web site (Bonta, 2005). By now several hundred nonviolent and peaceful societies have been documented, and it is likely that many times more exist.

Beyond anthropology, especially noteworthy is the diachronic perspective provided by historian Matthew Melko (1973: pp. 34–35) who has identified 52 societies, mostly states, in the history of the world with periods of peace without any warfare extending over a century or more, and some as long as seven centuries. Melko (1973: p. 8) notes "It strikes me that any society chosen at random for any hundred years would experience far more peace than war. The examples of war and revolution are usually experiences of a few years at most in any century. And in those years, war is intermittent, and usually does not involve the whole society. If history is a history of war and revolution, that is because historians tend to focus on periods of violence and rapid change …. When we have finished observing man at his most violent and dynamic, the fact remains that generations have lived in peace, and people

and their leaders have managed to deal with their internal hostilities without recourse to war. Peace is a fact, not a vision. It is ubiquitous, incessant, normal. Peace prevails in most places at most times."

Returning to anthropology, beyond Montagu there are several outstanding pioneers in researching and documenting nonviolence and peace, and thereby in various ways refuting the unfounded presumption that there is some kind of biological drive inevitably generating aggression among humans universally and externally. David Fabbro (1978) wrote the early seminal article called 'Peaceful Societies' for the *Journal of Peace Studies*. Therein he asserted that absent in such societies is violence, war, external threat, social stratification, regular political leadership, police, and the military. Present is a small intimate community, egalitarian social structure, generalized reciprocity, group consensus in decision-making, and nonviolent values and conditioning.

Through more extensive cross-cultural research Bruce D. Bonta (1996: pp. 404–405) independently confirms Fabbro's attributes of nonviolent and peaceful societies and further elaborates on them. In their worldview, ethos, values, attitudes, practices, institutions, and customs such societies are fundamentally different from violent and warlike societies. This includes their perceptions of human nature, conflicts, dispute resolution, and intolerance of violence. Bonta asserts that peacefulness is a condition of human society characterized by a relatively high degree of interpersonal harmony; little if any physical violence; workable strategies for nonviolent resolution of conflicts and for averting violence; a commitment to avoiding violence with other societies; and strategies for the socialization and enculturation of children to adopt and continue nonviolent ways. He identified six strategies for the nonviolent resolution of conflicts: self-restraint, negotiation, separation for avoidance, intervention, meetings to reduce tensions, and humor. Bonta (1996: p. 414) observes "The basic reason for peacefulness in these societies is that people are strongly opposed to actual physical violence and firmly in favor of nonviolence…. Violence is never acceptable…. Peacefulness is an absolute commitment for them."

Semai

These attributes are illustrated in the specific case of the Semai of Malaysia. The Semai were studied first for more than 2 years by Robert Knox Dentan (1997) and subsequently for more than 2 years by Clayton A. and Carole J. Robarchek (1998). Independently they reached the same conclusions about the nonviolent and peaceful character of this society. The Semai themselves consider their own society to be nonviolent and peaceful, and likewise, every individual thinks of themselves in that way. Nonviolence is the ethos of their society, the pivotal organizing principle and focus. Physical violence of any kind terrifies any Semai person, they considered it to be senseless and stupid. Force is not returned with force, but only with passivity or withdrawal. The Semai say that they do not get angry and do not hit another person or fight. Nonviolence and self-restraint are pivotal social values learned from childhood. Children are never hit and see only nonviolent adults as their role models. Yet naturally some tensions and conflicts arise within a small community. If a dispute is pursued, then it may be resolved through compensation, by the community talking it through until concerns and emotions subside, or by requesting adjudication by respected elders. Social solidarity is maintained by recognition of mutual dependence, through cooperation and sharing, and by fear of dangers from external forces.

For more than a 100 years various outsiders have mentioned the abhorrence of violence by the Semai. Suicides and homicides are extremely rare, although higher estimates of homicide rates by authors other than the above ethnographers are bogus. However, outside of their own society a few Semai were recruited by the British during the early 1950s in counterinsurgency warfare against Communists wherein they were ordered to kill non-Semai and did so, but this is the only known exception of killing. Thereafter the Semai fighters returned to their village where they were reintegrated as nonviolent and peaceful individuals (Dentan, 1997: pp. 55–59, 133).

Natural History of Peace

In a chapter titled "The Natural History of Peace: A Positive View of Human Nature and Its Potential" Leslie E. Sponsel (1996) marshals evidence and arguments from the four subfields of anthropology to present the opposite perspective from innate depravity. His conclusions are as follows: (1) While conflict is inevitable and ubiquitous, violence is not. (2) Human nature has the psychobiological potential to be either nonviolent/peaceful or violent/warlike. (3) Nonviolence and peace prevailed in most prehistoric and nonstate societies. (4) War is not a cultural universal. (5) The potential for the development of a more nonviolent and peaceful world is latent in human nature.

In *Warless Societies and the Origin of War*, Raymond C. Kelly (2000) provides a most sophisticated and detailed analysis. He aims to develop a general model for the evolution of war that is grounded in cross-cultural comparison and tested against the archaeological record. Kelly (2000: pp. 121–124, 160) importantly observes "The origin of war is a question of enduring interest because the conclusions reached are of central relevance to our conceptions of human nature, and such conceptions inform the political philosophies that shape and legitimize our social institutions."

In Kelly's (2000: p. 10) view, the hallmark of the origin of warfare is not just capital punishment or blood revenge. Several generalizations emerge from Kelly's (2000: pp. 147–148) analysis: almost all early Upper Paleolithic societies lacked war; the earliest conclusive evidence of war is from Nubians around 14 000–12 000 BC; warfare was rare to absent until very late in human prehistory, around 7500–7000 BC; it was related to the development of farming with sedentary villages, thereafter competition over trade routes, and finally to states as centralized and hierarchical political organizations.

In Kelly's (2000: pp. 44–45, 56–57) analysis, nomadic foraging or unsegmented societies are warless. They have minimal social groups like the universal nuclear family and local community; no higher level of social organization nor any subunits; lack concepts of group liability; crucially avenge

murder does not lead to blood revenge feuding; group membership is fluid rather than fixed; and there is no food storage. In contrast, more sedentary or specialized forager societies have a segmented hierarchy and concepts of group injury, liability, and responsibility that combine to provide the basis for feuding or warfare (Kelly, 2000: pp. 46–48).

One of the foremost anthropological researchers on the subjects of nonviolence and peace is Douglas P. Fry. Through his numerous authored and edited publications for more than a decade he proves that humans are not innately warlike, but instead are fully capable of living in peace. Also he demonstrates that much of past scholarship has been biased by a presumption of a 'beast within.' Fry identifies the 'man the warrior myth' as the vacuous and unfounded assertion that war is ancient, universal, natural, normal, and inevitable. He argues that this assertion is irrefutably contradicted by reason and evidence. It reflects a cultural bias including Western assumptions about human nature; is an unconscious and unexamined assumption; results in selectivity and advocacy in the representation of war; and exaggerates aggression. It is careless thinking to define war so broadly as intergroup aggression and to conflate it with interpersonal violence and feuding.

Two antithetical models of human nature are identified by Fry. In the man the warrior model male competition for reproductive fitness generates belligerent militarism. His detailed critical analysis reveals the cultural bias, selective use of evidence, and muddled thinking involved in this model. The second model is the simple hunter-gatherer pacifist. Here natural selection generates nonviolent conflict management and helps maintains peace. The result of Fry's extensive critical analysis, his model is objective, systematic, holistic, and comparative.

Fry identifies factors that in various combinations facilitate conflict management in different societies: avoidance, tolerance, negotiation, self-redress, friendly peacemaking, mediation, arbitration, adjudication, and repressive peacemaking. He emphasizes that human adaptive plasticity and international governance can help generate more peace in the future. He recommends that peace may be advanced through extending nonviolent conflict management from within democracies to international relations; recognizing interdependence (economic, security, and environmental); cultivating crosscutting ties; and strengthening the United Nations and regional alliances (Fry, 2006). (For elaborations on aspects of his work and related subjects see Fry, 2013.)

Reexamining Tribal Warfare

One of the foremost anthropological students of warfare, R. Brian Ferguson, is among those who have called for a serious reconsideration of ethnographic representations of some cultures as engaged in chronic endemic warfare. In a book Ferguson coedited with Neil L. Whitehead (2000) ethnohistory and other lines of evidence demonstrate that much of what has been considered to be traditional 'primitive' or tribal warfare was actually generated at least in part by the external forces of colonialism. Even if warfare existed prior to colonial contact, often it was modified and intensified by such forces. These assertions potentially undermine much of the ethnological record of war along with the presumption that human nature is inevitably warlike.

In the particular case of the Yanomami Ferguson (1995) extended this provocative perspective in a most exhaustive, meticulous, and penetrating systematic study of the historical, ethnohistorical, and ethnological literature. He demonstrates that for centuries Yanomami have been influenced by the political economy of regional, national, and international systems, first directly on the periphery of their territory, and second indirectly spreading inward. He focuses mostly on intervillage conflict and violence generated by competition in situations of differential access to metal trade goods.

Another reconsideration of Yanomami warfare by Sponsel (1996, 2010b) systematically questions the scientific validity and utility of Napoleon A. Chagnon's (2012) depiction of their aggression. Furthermore, the analysis even questions whether Yanomami aggression can be realistically characterized as warfare rather than blood feuding with raiding as one tactic, a phenomenon more reminiscent of the famous historical case of the Hatfields and McCoys in Appalachia. This reexamination of Chagnon's ethnography also reveals that within his own description there is considerable information about nonviolence and peace among the Yanomami, even though it is mostly overshadowed by his obsessive fixation on aggression, the latter quite unlike any of the other more than four dozen anthropologists who have worked with the Yanomami, several for many more years longer. Thus, one of the commonly presumed classic documentations of tribal warfare is called into question, thereby also problematizing the stereotypical demonization of so-called 'primitive' peoples and human nature more generally.

Transformation from Killing to Nonkilling

From another angle, the Waorani in the Ecuadorian Amazon provide a most fascinating and significant case of the transition from a violent to a nonviolent society. Clayton and Carole Robarchek (1998a) conducted fieldwork among the Waorani who have long had a reputation for intergroup violence through a cycle of blood feuds with revenge raids. However, after the Waorani had been exposed by missionaries to the idea that a society could be nonviolent and peaceful, they adopted such ways surprisingly rapidly. This transition from a killing to a nonkilling society reveals the plasticity of human nature and domination of culture over biology. Also it emphasizes a society's cultural construction of its worldview, values, and attitudes, and how their practical consequences are manifest in behavior. In addition, the Robarchek's (1998b) have compared the traditional Waorani and Semai, noting that they are similar in their culture and ecology, except for their worldview which makes all of the difference. (The Dani of New Guinea made a similar transition as documented by Karl Heider, 1997: p. 172.)

Nonkilling

During recent years a largely new perspective has emerged in peace studies, one that is potentially revolutionary in its

ramifications. Political scientist Glenn D. Paige (2009: p. 127) in a pioneering book argues "The time has come to set forth human killing as a problem to be solved rather than to accept enslavement by it as a condition to be endured forever." He challenges the assumption that killing is an inescapable part of the human nature, the human condition, and social life. He goes beyond peace and nonviolence to focus sharply and objectively on killing because it is a more concrete and measurable attribute. Furthermore, Paige recognizes killing as the prime source and sustainer of other forms of violence, and he calls for liberation from killing through a shift from the politics of killing to the those of nonkilling.

Paige poses the pivotal and provocative question: Is a nonkilling society possible? He defines a nonkilling society as a human community from the local to the global levels characterized by the absence of killing of humans, threats to kill, weapons designed to kill, justifications for using them, and conditions of society dependent upon the threat or use of lethal force for maintenance or change (Paige, 2009: p. 21).

From the previous discussion in the present article the answer to Paige's primary question is obviously in the affirmative. Indeed, clearly not only is a nonkilling society possible in theory, it exists in reality in several hundred societies as documented by numerous and diverse anthropologists, some as in the case of the Semai, independently affirming through their fieldwork with the same society that it is nonkilling. However, this does not mean that interpersonal violence is totally absent in all of these cases, but that it is so rare as to be virtually so.

Paige's underlying thesis is that humans are capable of both killing and nonkilling; most humans are not killers (99%); nonkilling capabilities already work in a wide range of social institutions; and the capability exists for nonkilling transformative interventions. Accordingly, he problematizes the very foundational acceptance of violence in political science, politics, government, and society. He calls for objective scientific and scholarly investigation of the possibilities and actualities of nonkilling.

Another aspect of Paige's (2009: pp. 73–74, 79, 88–89) book is the four-part logic of analysis: What are the conditions, causes, and consequences of killing, nonkilling, and the transition from one to the other in either direction? Within the anthropology of peace studies, the ethnographic cases previously mentioned in the present article of the Yanomami, Semai, Waorani, and Dani are seminal examples coinciding with components of this mode of analysis. Far more systematic and objective empirical research is sorely needed to document in detail the varieties of nonviolent and peaceful societies throughout the world as well as transitions from violence and war to nonviolence and peace, or the converse. Bonta's (2005) Web site on peaceful societies is a prime resource to commence such research.

Expanding on the idea of a peace-to-war continuum of Karolina Baszarkiewicz and Fry (2008: pp. 1560–1561) can help to transcend some of the problems that arise from identifying a society as either killing or nonkilling, as, for example, when killing occurs but rarely in an otherwise nonkilling society, or when most individuals in a killing society do not kill. It may help to envision both intrasocietal and intersocietal variation along this continuum between the polar extremes of nonkilling and killing recognizing that most societies fall somewhere along this continuum, but that their position may shift in either direction. This idea of a continuum surmounts problematic absolutist and essentialist thinking in terms of either-or, all-or-none, always-or-never.

Apologists for War

In various ways some anthropologists and others appear to be apologists for war in the sense that ultimately they blame some ethereal human nature as an excuse rather than recognizing the primacy of the specific political and moral choices that are decisive. Archaeologist Lawrence H. Keeley (1996) in his book *War Before Civilization*, primatologist Richard Wrangham with Dale Peterson (1996) in *Demonic Males: Apes and the Origins of Human Violence*, and archaeologist Steven A. Le Blanc with Katherine E. Register (2003) in *Constant Battles: The Myth of the Peaceful, Noble Savage* are anthropological examples of a genre arguing that war is ubiquitous in time and space as a manifestation of human nature (cf Sponsel, 2012: pp. 21–30). Furthermore, they assert that the psychobiology of *Homo sapiens* inevitably generates territoriality, competition, hostility, violence, and warfare within and between human societies. In the guise of science and scholarship this political ideology celebrates the man the warrior model wherein the predator naturally turns inward to prey on its own kind thereby becoming the premier killer ape.

Keeley, for example, aims to expose and rebut the supposed 'pacification of the past' by liberal anthropologists. However, he carelessly defines war so broadly as to be meaningless; neglects the substantial literature irrefutably documenting nonviolent and peaceful societies; ignores the very serious limitations of the archaeological evidence from prehistory; and constructs blatantly biased interpretations of 'data.'

This mind-set has persisted since journalist Robert Ardrey's books *African Genesis* (1961) and *Territorial Imperative* (1966), and for that matter, with Thomas Hobbes (1651) *Leviathan* and long before. Psychoanalyst Sigmund Freud (1930) in *Civilization and Its Discontents* and ethologist Konrad Lorenz (1966) in *On Aggression* contributed to this mind-set as well in both the scientific and public arenas.

Most recently the essence of this perspective is exhibited in biologist Edward O. Wilson's (2012) *The Social Conquest of Earth* in the chapters revealingly titled "Tribalism Is a Fundamental Human Trait" and "War as Humanity's Heredity Curse." He equates tribalism with universal 'groupist behavior' which is some vague kind of presumed hardwired or inborn tendency for in-group social identity inexplicably automatically correlated with its territoriality, hostility, and aggression toward any out-group. Wilson (2012: p. 62) asserts "Our bloody nature, it can now be argued in the context of modern biology, is ingrained because group-versus-group was a principal driving force that made us what we are. In prehistory, group selection lifted hominids that became territorial carnivores to heights of solidarity, to genius, to enterprise. And to *fear*. Each tribe knew with justification that if it was not armed and ready, its very existence was imperiled. Throughout history, the escalation of

a large part of technology has had combat as its central purpose." Then Wilson (2012: p. 65) reveals "It should not be thought that war, often accompanied by genocide, is a cultural artifact of a few societies. Nor has it been an aberration of history, a result of the growing pains of our species' maturation. Wars and genocide have been universal and eternal, respecting no particular time or culture."

Wilson's reasoning may be a logical construction, but it is a speculative 'just so story' reminiscent of nineteenth century evolutionists devoid of any convincing proof and acknowledgment of the substantial contrary scientific evidence sampled in this article. This is simplistic and reductionistic biological determinism. Such gross generalizations are contradicted by the tremendous variation within and among societies from prehistory to the present. If some kind of an instinctive drive for aggression actually existed as imagined by Wilson, then far more violence including warfare would exist as well, yet 99% of humanity has never killed anyone.

It is especially relevant here that there is significant evidence from psychology contradicting the demonic male and related speculative models. Military personnel have to be trained systematically to kill, it does not come naturally. In spite of training, including desensitizing the soldier and dehumanizing the enemy, most military do not kill, even on the battlefield. Only about 20% of combat soldiers even fire their weapons. Those who do kill often suffer lingering emotional problems thereafter, in many cases posttraumatic stress disorder. Shooting can be more traumatizing than being shot. Surprisingly, even operators of drones (unmanned aircraft) that kill humans from a distance may suffer emotionally. Ordinarily there is a natural inhibition against killing another human. Indeed, the case can be made that selective pressures facilitated nonkilling in the course of human evolution as a successful adaptation (Evans Pim, 2009: pp. 327–340, 359; Fry, 2006, 2013).

Fry very usefully discusses in detail seven ways that evidence of nonviolent and peaceful societies is disputed, denied, or ignored: "l. Stridently emphasize violence and simply ignore all evidence to the contrary …. 2. Set an impossible standard, such as requiring a peaceful society to be absolutely and totally free of all forms of aggression across time, find a few real or apparent exceptions to the impossible standard, and then dismiss the society in question as clearly NOT peaceful …. 3. Count any type of conflict as if it were aggression …. 4. Cite an inflated homicide rate to exaggerate violence …. 5. Regarding warfare in particular, make clever use of terminology to exaggerate the number of societies that engage in war …. 6. Apply cases of violence or war from a certain place or time to a totally different place or time …. 7. Create a bogus peaceful society, then destroy your creation" (Kemp and Fry, 2004: pp. 185–204). (Also see the chapters by Piero P. Giorgi, Leslie E. Sponsel, and Robert W. Sussman and Donna Hart in Evans Pim, 2010.)

When the very substantial evidence for the natural history of peace that has been accumulating for well over three decades as highlighted in the present article is totally ignored by apologists for war and others, then the only logical conclusion is that either the author is entirely unaware of the relevant literature and thus has failed to pursue adequate research to provide a fair and balanced approach, or the author is disingenuous in avoiding contradictory data and arguments. Either of these two deficiencies is irresponsible and unworthy of the designation science or scholarship. Moreover, in some situations privileging the study of war and other forms of violence to the complete exclusion of any consideration of nonviolence and peace might tend to legitimize and celebrate violence and war, and thereby might indirectly facilitate them even if inadvertently.

More Obstacles

There is a systemic bias privileging attention to, and often pursuit of, violence and war in Western culture, and especially in America. It is a product of the industrial–military–scientific–academic–media complex, especially in the U.S.A. It reflects some variants of Abrahamic religions as well. Science and scholarship are not necessarily entirely apolitical and amoral, but can be biased by culture, religion, ideology, politics, and/or personality.

This systemic bias is an empirical fact that can be demonstrated quantitatively in the public, academic, peace studies, and anthropology arenas. For instance, on 16 October 2012, a search of Google.com for these key words revealed the following: violence 387 000 000; nonviolence 5 380 000; war 2 480 000 000; and peace 922 000 000. In a survey of articles in *the Journal of Peace Research*, Hakan Wiberg (1981) found that from 1964 to 1980 there were 400 articles, but only one on peace, the previously mentioned one by Fabbro (1978). As a more recent example, in the *Encyclopedia of Violence, Peace, and Conflict* (Kurtz, 2008), out of a total of 289 entries, only 10 (3.5%) are on nonviolence while only 29 (10%) are on peace, the rest focused on various aspects of violence and war. In the case of anthropology, a search on 16 October 2012 of the Anthropology Index Online revealed the following number of citations for journal articles from 1950 to 2012 for these key words: violence 2789; nonviolence 2; war 10 000; and peace 608. It should be cautioned that not all of the citations found are relevant to the key words. Furthermore, the citations that appear under peace are overwhelmingly about conflict, violence, and/or war.

This striking discrepancy is troubling because it is impossible to achieve a more nonviolent and peaceful society by ignoring these phenomena. To use an analogy, health and illness are necessarily complementary concepts. Exclusively studying only disease is insufficient, if the concern is with health, and the converse holds as well. Furthermore, when the components of nonviolence and peace in a society are neglected, then reality may be distorted to the point of being misleading and that may have serious practical consequences (Sponsel, 1998, 2010a,b).

This systemic bias can discourage concerns for peace and peace activism. If war is an inherent tendency in human nature, then it seems to be inevitable, unalterable, and unavoidable. Such a mind-set may contribute to some government leaders privileging military over diplomatic resolutions of conflicts, pursuing an arms race, spending excessive amounts of money on weapons instead of other matters like medical research and education, considering preemptive strikes, and so on (cf Fry, 2006: pp. 251, 262).

Another obstacle is that many authors writing about war and peace pursue the negative concept of peace as no more than the absence of war. The focus is on direct violence, violent resolution of conflict, war as natural, defense of nation, and superficial symptoms rather than ultimate causes, a realist purview. In contrast, the positive concept of peace is far more inclusive. It encompasses the presence of human rights, defense of humanity, indirect or structural violence, the ultimate causes of violence and war, nonviolent conflict resolution, and peace as natural, a liberal purview. Peace is more than simply the absence of war, just as life is more than simply the absence of death (Sponsel and Gregor, 1994: pp. 1–36).

Ultimately much of the above also reflects a fundamental difference in political ideology despite the overt pretense by most scientists of being apolitical and amoral. The realist or conservative tends to pessimistically view humans as evil and driven by selfish instincts; international politics as competition for power and national interests; war as inevitable; everything as subordinated to national interest; defensive strength as necessary to subdue or deter enemies; and violence as the means to resolve conflicts. On the other hand, the liberal or leftist tends to optimistically view humans as good; altruism as natural; international politics as cooperation for mutual interests and principles; peace as inevitable; national and international interests as often coinciding; international associations as pivotal; and nonviolent resolution of conflicts as the most desirable and effective (Kegley and Raymond, 1999: pp. 20–21, 245). Of course, individuals and groups may not neatly fall into one category or the other depending, on the specific issues involved.

Conclusions

Nonviolence and peace are ubiquitous, violence and war are not. There are societies that do not have warfare, and where homicide and other forms of violence are extremely rare or totally absent. Furthermore, there is tremendous variation in the manifestation of violence, nonviolence, war, and peace among some 7000 cultures in the world today which renders facile generalizations about human nature untenable. Even those societies wherein interpersonal or intergroup aggression is manifest in various ways and degrees, it is not ubiquitous simply because for any society to survive there are indispensable attributes like sociability, cooperation, reciprocity, altruism, and empathy that strongly discourage any regular violence and war. If nonviolence and peace are not taken into account, then considerations about human evolution, human nature, the human condition, cultures, regions, culture history, and conflicts may be incomplete and may misinform and mislead.

The Semai, Yanomami, Waorani, and Dani cases are especially pertinent to the anthropology of peace and nonviolence. All four are small-scale indigenous societies pursuing a subsistence economy in mountainous tropical forest ecosystems in three widely separate geographical areas. Such societies may be considered to be among the most basic expressions of human nature, yet they are very different with respect to violence, nonviolence, war, and peace. Their differences are not explicable in terms of biology or other aspects of 'nature' but only in terms of 'nurture' which means culture. For instance, if their differences were attributable to some fixed biological aspect of human nature, then this would not allow the abrupt, willing, and definitive transition from violence and war to nonviolence and peace in the case of the Waorani and Dani among many others.

Many problems and issues are discussed and debated in the anthropological literature on peace studies, among them: the antiquity, universality, and inevitability of war versus peace; biological versus cultural explanations of the causes and effects of war and peace, the biological including sociobiology and its successor evolutionary psychology, the cultural with materialistic versus mentalistic explanations; tribal warfare compared with that of states; and the impact of states on aggression within and among tribes. In any case, nonviolence and peace surely merit just as much serious, systematic, and empirical attention as violence and war, if the aforementioned biases can be surmounted by those anthropologists and others who identify themselves as objective scientists and scholars. Paige (2009:126) asserts that every individual can be a center for global nonkilling to facilitate the transition to a nonkilling world. This applies to anthropologists as well as to both their teaching and their research.

See also: Colonialism, Anthropology of; Conflict and War, Archaeology of: Behavior and Social Organization; Conflict and War: Anthropological Aspects; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Peace; Violence in Anthropology; Warfare in History.

Bibliography

Baszarklewicz, K., Fry, D., 2008. Peaceful societies. In: Kurtz, L. (Ed.), Encyclopedia of Violence, Peace, & Conflict, vol. 2. Elsevier, San Diego, CA, pp. 1557–1570.

Bonta, B.D., 1993. Peaceful Peoples: An Annotated Bibliography. The Scarecrow Press, Inc., Metuchen, NJ.

Bonta, B.D., 1996. Conflict resolution among peaceful societies: the culture of peacefulness. Journal of Peace Research 33 (4), 403–420.

Bonta, B.D., 2005. Peaceful Societies: Alternatives to Violence and War. http://www.peacefulsocieties.org.

Chagnon, N.A., 2012. Yanomamo, sixth ed. Harcourt Brace Jovanovich, Fort Worth, TX.

van der Dennen, J.M.G., 1995. The Origin of War, vol. 2. Origin Press, Groningen, The Netherlands.

Dentan, R.K., 1997. The Semai: A Nonviolent People of Malaya, Fieldwork ed. Holt, Rinehart and Winston, New York.

Evans Pim, J. (Ed.), 2009. Toward a Nonkilling Paradigm. Center for Global Nonkilling, Honolulu, HI. http://www.nonkilling.org.

Evans Pim, J. (Ed.), 2010. Nonkilling Societies. Center for Global Nonkilling, Honolulu, HI. http://www.nonkilling.org.

Fabbro, D., 1978. Peaceful societies: an introduction. Journal of Peace Research XV (1), 67–83.

Ferguson, R.B., 1995. Yanomami Warfare: A Political History. School for Advanced Research, Santa Fe, NM.

Ferguson, R.B., Whitehead, N.L. (Eds.), 2000. War in the Tribal Zone: Expanding States and Indigenous Warfare, second ed. School of American Research, Santa Fe, NM.

Fry, D.P., 2006. The Human Potential for Peace: An Anthropological Challenge to Assumptions about War and Violence. Oxford University Press, New York.

Fry, D.P. (Ed.), 2013. War, Peace, and Human Nature. Oxford University Press, New York.

Fry, D.P., Bjorkqvist, K. (Eds.), 1997. Cultural Variation in Conflict Resolution: Alternatives to Violence. Lawrence Erlbaum Associates, Publishers, Mahwah, NJ.
Heider, K., 1997. Grand Valley Dani: Peaceful Warriors, third ed. Harcourt Brace College Publishers, Fort Worth, TX.
Howell, S., Willis, R. (Eds.), 1989. Societies at Peace: Anthropological Perspectives. Routledge, New York.
Keeley, L.H., 1996. War before Civilization. Oxford University Press, New York.
Kegley Jr., C.W., Raymond, G.A., 1999. How Nations Make Peace. St. Martin's Press, New York.
Kelly, R.C., 2000. Warless Societies and the Origin of War. University of Michigan Press, Ann Arbor, MI.
Kemp, G., Fry, D.P. (Eds.), 2004. Keeping the Peace: Conflict Resolution and Peaceful Societies Around the World. Routledge, New York.
Kurtz, Lester (Editor-in-Chief), 2008. Encyclopedia of Violence, Peace, and Conflict, second ed., vol. 4. Academic Press, San Diego, CA.
LeBlanc, S.A., Register, K.E., 2003. Constant Battles: The Myth of the Peaceful, Noble Savage. St. Martin's Press, New York.
Melko, M., 1973. 52 Peaceful Societies. Canadian Peace Research Institute, Oakville, ON.
Montagu, A., 1976. The Natural History of Aggression. Oxford University Press, New York.
Montagu, A. (Ed.), 1978. Learning Non-aggression: The Experience of Non-literate Societies. Oxford University Press, New York.
Paige, G.D., 2009. Nonkilling Global Political Science. Center for Global Nonkilling, Honolulu, HI. http://www.nonkilling.org.
Robarchek, C.A., Robarchek, C.J., 1998a. Waorani: The Contexts of Violence and War. Harcourt Brace College Publishers, Fort Worth, TX.
Robarchek, C.A., Robarchek, C.J., 1998b. Reciprocities and realities: worldviews, peacefulness, and violence among the semai and waorani. Aggressive Behavior 24, 123–133.
Sponsel, L.E., 1996. The natural history of peace: a positive view of human nature. In: Gregor, T.A. (Ed.), The Natural History of Peace. Vanderbilt University Press, Nashville, TN, pp. 95–125. http://www.peacefulsocieties.org/Archintr/spo96int.html.
Sponsel, L.E., 1998. Yanomami: an arena of conflict and aggression in the amazon. Aggressive Behavior 24 (2), 97–122. http://nku.edu/~humed1/darkness_in_el_dorado/documents/0603.pdf.
Sponsel, L.E., 2010a. Reflections on the possibilities of a nonkilling society and a nonkilling anthropology. In: Evans Pim, J. (Ed.), Nonkilling Societies. Center for Global Nonkilling, Honolulu, HI, pp. 17–52. http://www.nonkilling.org/pdf/nksocieties.pdf.
Sponsel, L.E., 2010b. Into the heart of darkness: rethinking the canonical ethnography of the yanomamo. In: Evans Pim, J. (Ed.), Nonkilling Societies. Center for Global Nonkilling, Honolulu, HI, pp. 197–240. http://www.nonkilling.org/pdf/nksocieties.pdf.
Sponsel, L.E., 2012. Spiritual Ecology: A Quiet Revolution. Praeger, San Diego, CA. http://www.spiritualecology.info.
Sponsel, L.E., Gregor, T. (Eds.), 1994. Anthropology of Peace and Nonviolence. Lynne Rienner Publishers, Inc., Boulder, CO.
Wiberg, H., 1981. JPR 1964–1908 – what have we learned about peace. Journal of Peace Research XVIII (2), 111–148.
Wilson, E.O., 2012. The Social Conquest of Earth. W.W. Norton & Company, New York.
Wrangham, R., Peterson, D., 1996. Demonic Males: Apes and the Origins of Human Violence. Houghton Mifflin Company, Boston, MA.

Peace Keeping

Franz Kernic, ETH Zurich / Swiss Federal Institute for Technology Zurich, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Peacekeeping can be understood as a measure to limit and manage armed conflict and political violence. Its main goal is to assist societies and states in the transition from war to peace and to prevent a potential escalation toward war. Since 1948, the United Nations and other actors have conducted more than 100 peacekeeping operations around the globe. In the 1990s, the concept of peacekeeping was further developed and expanded. New missions were launched leading to a tremendous increase in the number of operations and peacekeepers deployed and a wide array of different operations in terms of mandate, scope, size, duration, and troop composition. New theories and concepts started to emerge at the turn of the millennium, contributing to a broader understanding of peacekeeping and linking it closer to ideas of liberal peace, democracy, and development.

Introduction

Peacekeeping is commonly understood as a tool to manage and resolve armed conflict and political violence and to assist societies in the transition from war to peace. Although the basic idea can be traced back further in history, the term itself came into use in the 1950s in association with the United Nations' conflict management efforts in global politics. At the early stage, peacekeeping was regarded as a measure to be taken in the aftermath of war in order to assist the implementation of a peace agreement and to lay the groundwork for conflict resolution and a peaceful settlement of the conflict. Since the 1990s, a few peacekeeping operations have also been set up specifically for the purpose of conflict prevention. Over the last 60 years, the concept of peacekeeping has been modified and expanded significantly. Changes in global environment and international system have stimulated both the emergence of new types of peace operations and attempts to conceptualize and design them. This development has led to today's wide spectrum of definitions and concepts, as well as different practical arrangements for peacekeeping.

Reviewing the existing definitions and concepts of peacekeeping, one quickly comes to the conclusion that there is in fact little agreement among scholars and theorists as to what peacekeeping is beyond the fact that the term itself refers to activities that help create conditions for conflict resolution and sustainable peace. Every scholar attempting to define peacekeeping is confronted with the fact that there are a number of commonly used terms with blurry boundaries: peace-building, peacemaking, peace-enforcement, peace-support, etc.; terms that, in common language, are frequently used interchangeably to refer to very similar activities and phenomenon. An additional difficulty emanates from the fact that, due to its specific connotations, the term peacekeeping is often used by governments to label and legitimize political action that in fact has nothing to do with the original peacekeeping idea and frequently only helps to cover up conventional military intervention. Finally, it must also be noted that, in recent years, the term peacekeeping has been incorporated into, or even merged with, the broader term of 'peace operations.'

Since the early years of the Cold War, the UN and other actors have conducted more than 100 peacekeeping operations in the field, ranging from large-scale troop deployment to small-size missions and significant variations in terms of scope, mandate, duration, and composition. The UN in particular has taken on a key role in conceptualizing peacekeeping as an instrument for the settlement of armed conflict, the construction and maintenance of peace in war-torn societies, and in establishing and deploying peacekeeping forces around the world. As of July 2012, the UN is the world's largest multilateral contributor to international peacekeeping and postconflict peace-building, deploying approximately 97 000 uniformed personnel (troops, military observers, and police) from 119 countries in 16 UN-led peace operations on four continents (UN Peacekeeping Fact Sheet, 2012). This impressive number represents a tremendous increase in both UN operations and UN peacekeepers since the year 1989. During the Cold War era, the number of both UN peacekeeping missions and peacekeepers was significantly lower. Since the first Cold War UN missions, both theory and practice of peacekeeping have changed dramatically, and the nature and scope of the operations, their specific role in global politics, and their impact on conflict management and resolution in modern societies have been significantly altered too.

Theories and Concepts

Of course, all theories and concepts of peacekeeping need to be seen as embedded in the much broader context of our prevailing political and social theories, particularly our theories of international relations and peace and conflict. Major changes and paradigm shifts in those theories have a significant impact on how peacekeeping is defined and conceptualized. In the last few decades, a number of new theoretical approaches in the social sciences and behavioral studies have contributed to a comprehensive reevaluation of our prevailing concepts of peacekeeping and their respective key assumptions, thus enriching the scientific study of peace operations from multidisciplinary perspectives.

To begin with, it is important to note that the UN's attempts to develop a conceptual framework for its peacekeeping efforts and for systematically dealing with security and transition issues in the aftermath of war and violent conflict have always suffered from the fact that the term 'peacekeeping' is not found in the UN Charter. Nevertheless, references have often been made to a nonexisting 'Chapter six-and-a-half,' meaning that the Charter still needs to be seen as providing a legal and normative framework for such operations that could, theoretically, be placed right in between the arrangements outlined in Chapters VI and VII. The UN has never overcome this deficit and even today the organization is still struggling with the problem of clearly stipulating the meaning of peacekeeping. Not surprisingly, member states have continued to express different views on how to translate the basic idea of peacekeeping into concrete action in the field and to define precisely what one could write down and commonly agree on in such a Chapter Six-and-a-half.

Nevertheless, the UN did manage to quickly achieve a minimum consensus about a few basic principles of peacekeeping, thus opening the door for launching and managing a number of UN operations and making the necessary practical arrangements. First, there was general agreement among member states about the utility of peacekeeping for conflict management, and that it was desirable to use this tool within the UN system and under specific circumstances to limit the incidence of armed conflict, to implement peace accords, and to assist societies and states in the transition from war to peace (particularly if the conflicting parties request such an assistance). Second, there was also a broad consensus that peacekeepers could and should carry out some essential tasks needed to facilitate such a peace process (e.g., monitor ceasefire, observe peace processes and report violations of peace accords, apply confidence-building measures, provide administrative and logistic support, etc.). Third, close collaboration with diplomatic efforts toward conflict resolution seemed highly desirable, thus also prompting the idea of peacekeepers as 'diplomats in uniform.' Finally, there was widespread acceptance of the following three core principles of peacekeeping: consent, impartiality, and minimum use of force. In this respect, Canadian diplomat Lester B. Pearson and Dag Hammarskjöld from Sweden, who served as UN Secretary-General from 1953 to 1961, are generally given much credit for coining the term and for conceptualizing peacekeeping, together with promoting ideas of preventive diplomacy and collective security.

Most common definitions of peacekeeping are either linked to this original idea and the three outlined key principles (UN DPKO, 2008: 31–35) – frequently focusing on the 'functions' of peacekeeping – or they tend to emphasize more the specific 'tasks' that peacekeepers – the military, police, and civilian personnel – have to accomplish (e.g., UN DPKO, 2003: 1–2). Frequently, nonengagement in combat, i.e., the use of force being strictly limited to self-defense, is stressed as an important characteristic. But precisely this issue has always been very controversially discussed, particularly due to the fact that it is extremely difficult to draw a clear line between peacekeeping and peace-enforcement. As early on as the 1960s, the UN became very much aware of the potential for escalation and the danger that a peacekeeping operation suddenly might be forced to exceed its original mandate and engage in combat (e.g., the experience of UN peacekeeping mission to the Congo, 1960–64).

The big differences in composition and conduct of peacekeeping operations over the last 60 years have encouraged scholars to develop certain 'typologies.' For example, Bellamy and Williams (2010) developed a typology based "on the type of actors that conduct them (i.e. individual states, coalition of the willing, or formal regional arrangements and other international organizations) and their relationship to the UN (i.e. whether they are UN 'blue helmet' operations, UN-authorized operations or non-UN operations)" (Bellamy and Williams, 2010: 43f). Others tended to differentiate between various 'generations' of peacekeeping, each generation representing a different type of operation and nature of peacekeeping: First-generation peacekeeping (or traditional, prior to 1988), followed by second-generation peacekeeping (multidimensional and more complex operations, large in size, far-reaching objectives), third-generation peacekeeping (since the late 1990s, large-scale and robust 'peace support operations'), and, finally, fourth-generation peacekeeping (since the mid-1990s, subcontracting the military component). However, if seen in strict chronological order, this approach is misleading. It appears more appropriate to see a general widening of peacekeeping over the years, thus allowing a parallel existence of all generations, although with shifting priorities and weights.

This changing nature of peacekeeping was already addressed in Boutros-Ghali's *An Agenda for Peace* (1992), stating that "the established principles and practices of peacekeeping have responded flexibly to new demands of recent years, and the basic conditions for success remain unchanged: a clear and practicable mandate; the cooperation of the parties in implementing that mandate; the continuing support of the Security Council; the readiness of Member States to contribute the required military, police and civilian personnel, including specialists; effective United Nations command at Headquarters and in the field; and adequate financial and logistic support" (paragraph 50). This document also provided a definition of peacekeeping, distinguishing it from other terms such as 'preventive diplomacy' and 'peace-making': "*Peace-keeping* is the deployment of a United Nations presence in the field, hitherto with the consent of all the parties concerned, normally involving United Nations military and/or police personnel and frequently civilians as well. Peace-keeping is a technique that expands the possibilities for both the prevention of conflict and the making of peace" (paragraph 20).

Paradigm shifts in international relations theory have heavily influenced the way peacekeeping is conceptualized. While realism tended to emphasize the role of states in global politics and, consequently, with respect to peacekeeping, to focus primarily on the possibilities and conditions for a peaceful settlement of disputes between states (taking a specific state-centered or 'Westphalian view' of the international system), the emergence of new ('post-Westphalian') concepts of liberal peace gave birth to a quite different perspective on the purpose and utility of peacekeeping. In the 1990s, many scholars and politicians saw the spread of liberal democracy as the best path to global security and peace. Consequently, new concepts of peacekeeping emerged that

were significantly broader than before, focusing much more on how to create, spread, and maintain liberal democracies and promote an agenda for sustainable peace than on limiting and managing armed conflicts between or inside states. This paradigm shift brought peacekeeping much closer to the political agendas of development, humanitarian assistance, rule of law, and public security.

In the early twenty-first century, these concepts of a 'widened' or 'enlarged peacekeeping' were increasingly met with skepticism and sometimes radically called into question. Some emerging new theories (for example, critical theory and globalization theory) took on a rather critical stand when evaluating peace operations and their long-term impact on modern societies. In this context, new questions were raised, such as whether it is the international community that decides autonomously where and how peace operations are conducted or autonomous processes of 'globalization' and/or a specific 'global culture' (Paris, 2003) that force states to act in a specific manner and to apply the available peacekeeping instrument according to certain 'structural needs.' Critical theory, in turn, has drawn our attention to issues of control, disparity, and social struggle in the global economic system, and the problem of economic hegemony, opening the door, for example, for Michael Pugh's argument that "peace support operations serve a narrow, problem-solving purpose – to doctor the dysfunctions of the global political economy within a framework of liberal imperialism" (Pugh, 2004: 39). According to him, the limits of the post-Cold War forms of peacekeeping lie in "the inherent contradictions of a capitalist hegemony trying to keep the lid on instability" (Pugh, 2004: 53). In recent years, such critical voices have given rise to a number of new proposals for a radically reconceptualized peacekeeping. Among them, the idea of 'cosmopolitan peacekeeping' has gained considerable influence (Woodhouse and Ramsbotham, 2005).

UN and Non-UN Peacekeeping

Although the UN has become one of the primary peacekeeping actors in global politics since the early 1950s, it has never been the only actor conducting peacekeeping operations. In their study, Heldt and Wallensteen (2006) even argue that "in terms of number of operations and personnel," non-UN "activities are fully comparable to the UN's efforts" (39): "At the aggregate level there is a strong relationship over time between the number of ongoing peacekeeping operations carried out by the UN and non-UN actors, and an even stronger one for these actors' operations in intrastate conflicts, while virtually none exists for deployment in interstate conflicts" (Heldt and Wallensteen, 2006: 39). In particular, regional organizations, such as the Organization of American States, Economic Community of West African States, Organization of African Unity (OAU) and African Union (AU), North Atlantic Treaty Organization (NATO), and the European Union (EU), have played a much more active role in peacekeeping around the globe in the last two decades; but also single states and state coalitions ('coalitions of the willing') have become important peacekeeping actors, sometimes explicitly authorized by the UN Security Council. The UN Charter gives the UN Security Council the power to authorize regional organizations or other coalitions to conduct peacekeeping operations on its behalf, but it is important to note that the record of peacekeeping operations during the Cold War and post-Cold War era includes several actions that obviously lack such authorization.

From the early 1950s onward, one central problem of UN peacekeeping has been related to an obvious lack of resources and force capabilities to carry out these tasks and being permanently at the disposal of the organization. The UN still has no standing military forces – despite the existence of respective provisions in UN Charter Articles 42 and 43 – so that ad hoc UN troops for every mission and task are to be established. This, in turn, has made UN peacekeeping highly dependent on the willingness of its member states (and others) to provide the necessary troops, material resources, and funds. In this context it is worth noting that some of the weaknesses of the UN – with its unique attempt to balance power among major powers and member states – have had a negative and hampering impact on the way the organization organizes and conducts its daily business, including peacekeeping. Several shortfalls have been identified over time, thus leading to various internal reforms and attempts at strengthening both the conceptualization and design as well as the practical handling of peacekeeping operations. Two milestones in this respect, the Brahimi Report 2000 and the Peace Operations 2010 Reform Agenda, have stimulated an intense public debate about the future direction of UN peace operations and the proper use of this instrument in global conflict management.

The UN has an impressive record of peacekeeping activities: As of July 2012, the organization has conducted a total of 67 peacekeeping operations since 1948, with 16 operations ongoing; and more than 100 countries have provided military, police, and civilian personnel (UN Peacekeeping Operations Fact Sheet, 2012). The numbers of UN peacekeeping operations and UN peacekeepers deployed have significantly changed over time, with a first peak in the early 1990s (78 444 troops, police, and military observers in July 1993) and a second peak in the year 2010 (101 393 troops, police, and military observers in March 2010) (UN DPKO Statistical Material, 2012a). The total number of fatalities in those missions comes to slightly more than 3000 (1948–2012) (UN DPKO Situation Center, 2012) (for additional statistical material and data, see Heldt and Wallensteen, 2006; Bellamy and Williams, 2010).

In addition to UN peacekeeping, non-UN actors have conducted a number of similar operations that can be considered peacekeeping in terms of their mandate and scope. The list is lengthy and includes both large-scale (for example, the 'Indian Peace Keeping Force,' in Sri Lanka, 1987–90, had a size of more than 50 000 soldiers) and quite small operations. Since the mid- and late 1990s, several regional organizations have taken on an active role in planning and setting up new peacekeeping operations; among them the EU, which incorporated the so-called 'Petersberg tasks' of the Western EU (1992) into its Common Foreign and Security Policy, thus identifying peacekeeping as one of the Union's key security and military tasks. In recent years, the AU has started to build up its own peacekeeping capabilities and to conduct its own operations in Africa (e.g., African Union Mission in Sudan, 2004–07; African Union Mission in Somalia, since 2007). The EU has financially contributed to this development and concluded

a comprehensive 'Africa-EU Peace and Security Partnership' in order to strengthen regional security in Africa and to enhance the AU's capabilities to lead and conduct simultaneously multiple peacekeeping operations on the African continent.

The historical perspective allows us to identify a few interesting patterns and trends in UN peacekeeping:

- First, a significant shift in terms of troop-contribution has occurred since the first UN peacekeeping operation was set in motion. During the Cold War, European states dominated UN peacekeeping. However, since the early 1990s, South Asian (Pakistan, Bangladesh, India, Sri Lanka, and Nepal) and African states (Ghana, Nigeria) have become the largest troop providers to UN peace operations – a trend that has continued until today. Countries such as Bangladesh, Pakistan, India, Ethiopia, and Nigeria currently (July 2012) shoulder the greatest burden and provide a huge number of personnel to the UN, while the troop-contribution of Western countries – Europe and the United States – are on decline, counting only for about 5% of the personnel deployed in UN peacekeeping operations (UN DPKO Statistical Material, 2012b; compare Bellamy and Williams, 2010: 147).
- Second, the composition of UN peacekeeping troops has always tended to be multinational and comprising soldiers, police officers, and civilian personnel, although with a strong majority of military personnel. In this respect, the available data (UN DPKO Statistical Material, 2012c) point to an increase in the number of police officers deployed in UN missions, thus also indicating the increasing importance of policing and rule-of-law tasks in those operations.
- Third, specific procedures and arrangements were established in order to avoid some of the pitfalls and risks of peacekeeping. Many of them have proved crucial with respect to mission success and maintaining political control over the practical management in the field, including, for example, the provision that UN peacekeeping operations are mandated for a short period of time only (in general a 6-month period).
- Fourth, despite fluctuation in the number of peacekeeping operations and peacekeepers deployed, it is important to note two important trends: First, there are significantly more peacekeepers deployed in the post-Cold War period than before during the Cold War era. Second, the number of peacekeeping operations – both UN and non-UN – has also significantly increased after 1989 (Human Security Report, 2009/10: 70, Figure 4.3).
- Finally, it is remarkable that the peacekeeping operations have never been equally distributed among the continents and regions. There was a strong regional bias in the Middle East and only few UN missions were deployed in Europe, Africa, and the Americas during the Cold War. In the post-Cold War era this focus shifted, showing a significant increase in UN peace operations in Africa and Asia.

The Historical Development

Although instances of peacekeeping can be traced further back in history, particularly the nineteenth and early twentieth centuries (e.g., the deployment of a Swedish-Norwegian peace force to Schleswig 1849–50; the deployment of a League of Nations peacekeeping force to the Saar region), peacekeeping as a political instrument for global conflict management started to systematically develop with the UN's activities during the early years of the Cold War. The design, setup, and conduct of the first UN missions need to be seen as an important step toward establishing peacekeeping as a measure to be taken by the international community to limit and manage armed conflict in the bipolar and highly confrontational international environment of the Cold War period. In 1948, the UN had to deal with two major conflicts where diplomatic efforts were deemed highly necessary in order to maintain peace and security: Palestine and Kashmir. Interestingly, the two UN peacekeeping operations established at that time to manage those two conflicts and to oversee ceasefire agreements turned out to be not only the organization's first such operations, but also those which have lasted the longest (UN Truce Supervision Organization, 1948–present; United Nations Military Observer Group in India and Pakistan, 1949–present).

The first UN operations also served as role models for other similar arrangements, i.e., they helped to institutionalize certain mechanisms and procedures for putting together a peacekeeping mission. When confronted with the need to deal successfully with the Suez crisis, the UN further developed its peacekeeping concept and gradually adopted a common view on how ad hoc missions could be integrated into a broader concept of preventive diplomacy, thus allowing the UN from then on to take on a more decisive and coherent role in global conflict management. The UN operation in Sinai (United Nations Emergency Force I, 1956–67) turned out to be crucial in this respect. Despite the fact that different views and ideas about the proper use of this instrument continued to exist, the UN managed to authorize a number of additional peacekeeping operations until the end of the Cold War. Not surprisingly, numerous modifications and adaptations had to be made, resulting in a wide spectrum of different types of operation with sometimes quite different characteristics.

There were a total of 15 UN peacekeeping missions conducted from 1948 to 1988 (UN DPKO Statistical Material, 2012d) with a strong regional bias in and focus on the Middle East. However, tensions during the Cold War sometimes made consensus among major powers and member states extremely difficult. The role of the UN in global conflict management therefore remained highly contested. In the 1980s, the utility of peacekeeping operations as a tool for conflict management was increasingly called into question, thus leading to a period of stagnation in terms of practical arrangements and scope of UN peace operations. The largest and most complex operation during the Cold War period was conducted in the former Belgian colony of the Congo (United Nations Operation in the Congo, 1960–64) where troops engaged in combat.

A tremendous boost to UN peacekeeping happened immediately after the end of the Cold War. A number of new, larger, and more complex missions were authorized by the Security Council, increasingly also incorporating civilian tasks (like monitoring elections) into the mandate. Africa became a region with an increased demand for peacekeeping, with peacekeeping operations to be established in Angola, Western

Sahara, Somalia, Mozambique, and Uganda–Rwanda, but also Asia (e.g., in Cambodia) and Central America entering the spotlight of the UN's conflict management. This period also witnessed an increased participation of Western great powers in peace operations, e.g., the U.S. engagement in Somalia. The experiences in those operations led to new debates about the minimum use of force principle and the proper boundaries between peacekeeping and peace-enforcement.

Bellamy and Williams (2010: 93) speak about a 'triple transformation' that occurred with respect to peace operations during the 1990s, consisting of a quantitative, normative, and qualitative dimension. The authors emphasize that there was not only a radical increase in numbers of missions and troops deployed, but also a gradual normative shift and a new quality of peacekeeping in the nature of how these operations were conducted and which tasks were assigned (Bellamy and Williams, 2010: 100). In fact, the numbers of peacekeepers deployed around the globe peaked in the early 1990s and the number of missions that were authorized between 1988 and 1993 came to a total of 20 (Bellamy and Williams, 2010: 95). However, a tremendous setback happened only a few years later as a consequence of two major failures of UN peacekeeping, the Rwandan genocide (1994), and the massacre in Srebrenica and Bosnia-Herzegovina (1995). Both events – in combination with other failures such as in Somalia and Angola – led to a very controversial international debate about the utility of peacekeeping for limiting and managing conflict in the new post-Cold War world order, i.e., a period of reevaluation and general decline in peacekeeping efforts.

This development cumulated in a number of new ideas on how to strengthen peacekeeping and to avoid the pitfalls of past experiences. The obvious failures of the mid-1990s immediately led to a reduction in the number of UN peacekeepers. The number of troops deployed dropped from about 70 000 in 1993 to approximately 20 000 in 1996 (UN DPKO Statistical Material, 2012a), the UN facing a certain reluctance of its member states with respect to troop-contribution. However, this trend did not continue and, on the contrary, the numbers of operations and troops significantly increased again after the turn of the millennium. The first decade of the twenty-first century witnessed a 'limited rebirth' of UN peacekeeping and increased peacekeeping efforts of regional organizations. Two major operations were launched around 1999–2000: one in Kosovo (the NATO-led international peacekeeping mission of the 'Kosovo Force,' under UN mandate and the United Nations Interim Administration Mission in Kosovo) and the other in East Timor (the International Force for East Timor, a multinational peacekeeping operation under UN mandate, and the United Nations Transitional Administration in East Timor, which was succeeded by the United Nations Mission of Support in East Timor, in May 2002); and two smaller missions were authorized (Sierra Leone and Democratic Republic of the Congo). The number of peacekeepers doubled again and Africa became a region with increased and intensified peacekeeping efforts from both United Nations and other regional organizations. The following years saw a steady increase in the number and size of UN peace operations, peaking in 2010 with slightly more than 100 000 uniformed personnel being deployed and a slight decline in the following years (UN DPKO Statistical Material, 2012e).

The Study of Peacekeeping Operations

Since the 1950s, a wide array of theories and concepts has been developed in order to understand peacekeeping and to explain certain peculiarities of specific missions. The growing number and range of peace operations during the 1990s has stimulated further empirical research, as well as a number of new theoretical approaches. The scientific study of peacekeeping comprises all academic disciplines; in particular, political science, history, sociology, and anthropology have shown great interest in studying its dynamics. The growing bulk of research has led to countless case studies and comparative analysis, several of them addressing the question of success of such operations with respect to the goal of conflict settlement. When Diehl studied six major peacekeeping operations conducted during the Cold War, he concluded with a very pessimistic view stating that they were in the end not very successful (Diehl, 1994). Later studies came to more positive insights, arguing that "peacekeeping after civil wars does indeed make an important contribution to the stability for peace" (Fortna, 2004: 269) and that "it does tend to make peace more likely to last, and to last longer" (Fortna, 2004; Fortna, 2008). Regarding this debate, it is also important to note that, regardless of the conclusions drawn by researchers and scholars, public and elite perceptions about the success or failure of peacekeeping operations often have a significant impact on how much states are willing to engage in such efforts.

Another important factor for conducting comprehensive studies of peacekeeping has been a general desire to learn from previous experiences ('lessons learned') and to improve the performance of ongoing and future missions. As a result, there is a bulk of literature that focuses on the social and political dynamics of peacekeeping and aims at identifying major practical problems and shortfalls of past operations. In this context, a number of 'hot issues' and unintended consequences of peacekeeping are being addressed, including the UN's frequently quoted failure to establish suitable forces for their operations and assigned tasks, and to give them a proper mandate with achievable goals; organizational dilemmas within the UN system leading to critical events (in this respect compare, for example, the respective biographical notes and memoirs of military commanders of UN troops such as Carl Van Horn and Roméo Dallaire); general constraints due to the existing rules of engagement; the potential (psychological and physical) harm to troops and peacekeepers due to stress and the risky character of peacekeeping ('psychology of peacekeeping'); and the involvement of peacekeepers themselves in criminal activities such as human trafficking and forced prostitution.

Within the discipline of military sociology, a number of important societal and social issues of peacekeeping have been investigated thoroughly: recruitment and composition of a peacekeeping force with its practical consequences; behavior and different roles of soldiers in operations; variations in the way units conduct their day-to-day standard military activities; multinationality and complexity and the dynamics of social interaction (Soeters and Manigart, 2008); civil–military relations in peacekeeping operations; perceptions of the operational environment; postdeployment issues; and much more. The theater of operations of peacekeeping, where contingents from many different countries work side by side, has often been the main focus of contemporary sociological studies of

peacekeeping, studying in particular cultural aspects and existing points of tension in such 'complex cultural encounters' (Callaghan and Schönborn, 2004).

Outlook

Today, peacekeeping operations are confronted with a series of new challenges. A general crisis in global economy and increasing economic and financial pressure on Western states make it difficult for the UN, other regional organizations, and single states to commit themselves to new large-scale and long-term peacekeeping operations. This coincides with a general downsizing of standing military forces in many European countries, thus gradually reducing the pool for peacekeeping and shifting the burden even more from Europe to Asia and Africa. Although the UN's role in global conflict management has not radically been called into question in the early 2010s, the debate about a future regionalization of security and of conflict management seems to continue. This discussion is accompanied by a new debate about the proper design and use of peacekeeping forces. Of course, the interplay between great powers and international organizations still has a significant impact on when, where, and how peacekeeping operations are conducted; and this will continue into the future.

Two key questions have always been in the center of all debates about peacekeeping regarding its utility for the prevention and settlement of conflict in contemporary global politics: Which qualities and skills make a 'good peacekeeper'? And what is the best composition of a peacekeeping force in order to guarantee success in the field? The answers given to these questions usually do not differ much; although there has been a trend over time to increasingly stress the importance of the civilian component of peacekeeping and a significant shift with respect to gender, emphasizing much more the enormous importance of women in such endeavors and their specific situations, roles, and needs (see UN Resolution 1325). Frequently, answers have focused primarily on the aspect of 'soldiering' and whether a soldier trained for combat is in fact the right person to be deployed in a peacekeeping operation carrying out completely different tasks than needed for war-fighting. In this respect, former UN Secretary-General Dag Hammarskjöld has often been quoted and is believed to have said that "peacekeeping is not a job for soldiers, but only soldiers can do it." In fact, the military has always been the backbone of peacekeeping operations, frequently counting for more than 80% of all deployed personnel, thus contributing to the creation of new concepts and ideas of a 'soldiering for peace.'

See also: Arms Control; Conflict and Conflict Resolution, Social Psychology of; Military Sociology; Peace Movements; Peace Processes; Peace; War, Sociology of.

Bibliography

Bellamy, A.J., Williams, P.D., 2010. Understanding Peacekeeping, second ed. Polity, Cambridge.
Boutros-Ghali, B., 1992. An Agenda for Peace: Preventive Diplomacy, Peacemaking and Peacekeeping. UN Department of Public Information, New York.
Callaghan, J., Schönborn, M. (Eds.), 2004. Warriors in Peacekeeping. Points of Tension in Complex Cultural Encounters. LIT Muenster.
Diehl, P.F., 2008. Peace Operations. Polity, Cambridge.
Diehl, P.F., 1994. International Peacekeeping. Johns Hopkins University Press, Baltimore.
Doyle, M., Sambanis, N., 2006. Making War and Building Peace: United Nations Peace Operations. Princeton University Press, Princeton.
Durch, W., 2006. Twenty-first Century Peace Operations. US Institute of Peace, Washington.
Findlay, T., 2002. The Use of Force in UN Peace Operations. Oxford University Press, Oxford.
Fetherston, A.B., 1994. Towards a Theory of United Nations Peacekeeping. St. Martin's Press, New York.
Fortna, V.P., 2004. Does peacekeeping keep peace? International intervention and the duration of peace after civil war. International Studies Quarterly 48, 269–292.
Fortna, V.P., 2008. Does Peacekeeping Work? Shaping Belligerents' Choices after Civil War. Princeton University Press.
Heldt, B., Wallensteen, P., 2006. Peacekeeping Operations: Global Patterns of Intervention and Success, 1948–2000. Research Report, second ed. Folke Bernadotte Academy, Stockholm.
Human Security Report, 2009/2010 (2011). The Causes of Peace and the Shrinking Costs of War. Oxford University Press, New York.
Kondoch, B., 2007. International Peacekeeping. Ashgate, Aldershot.
Morjé Howard, L., 2008. UN Peacekeeping in Civil Wars. Cambridge University Press, Cambridge.
Paris, R., 2003. Peacekeeping and the constraints of global culture. European Journal of International Relations 9, 441–473.
Pugh, M., 2004. Peacekeeping and critical theory. International Peacekeeping 11, 39–58.
Schmidl, E.A. (Ed.), 2000. Peace Operations between War and Peace. Frank Cass, London.
Soeters, J., Manigart, P. (Eds.), 2008. Military Cooperation in Multinational Peace Operations. Managing Cultural Diversity and Crisis Response. Routledge, London.
UN, 1996. The Blue Helmets: A Review of United Nations Peace-keeping. United Nations, New York.
UN, 2000. Report of the Panel on United Nations Peace Operations ('Brahimi Report'). United Nations, General Assembly, New York.
UN, 2012. UN Peacekeeping Operations Fact Sheet 2012. Prepared by the Peace and Security Section of the United Nations Department of Public Information and dated 31 July 2012.
UN DPKO (United Nations Department of Peacekeeping Operations), 2003. Handbook on United Nations Multidimensional Peacekeeping Operations. DPKO, New York.
UN DPKO (United Nations Department of Peacekeeping Operations), 2008. United Nations Peacekeeping Operations Principles and Guidelines. DPKO, New York.
UN DPKO Situation Center, 2012. UN Peacekeeping: Fatalities by Year up to 31 August 2012. document dated 11 September 2012.
UN DPKO Statistical Material, 2012a. Surge in Uniformed UN Peacekeeping Personnel from 1991–Present. Prepared by the Peace and Security Section of DPI in Consultation with the Office of Military Affairs of DPKO DPI/2444/Rev. 18 August 2012.
UN DPKO Statistical Material, 2012b. Ranking of Military and Police Contributions to UN Operations. Monthly report, 31 July 2012.
UN DPKO Statistical Material, 2012c. Monthly Summary of Military and Police Contribution to United Nations Operations, 1995–2004. July 2012. Prepared by Cdr Waranya Prasoppen, FGS/MD/DPKO/UNHQ/NY.
UN DPKO Statistical Material, 2012d. List of Peacekeeping Operations 1948–2012.
UN DPKO Statistical Material, 2012e. Monthly Summary of Military and Police Contribution to United Nations Operations, 2005–2012. July 2012. Prepared by LtCol Florin Stanciu, FGS/OMA/DPKO/UNHQ/NY.
Wallensteen, P., 2012. Understanding Conflict Resolution, third ed. SAGE, London.
Woodhouse, T., Ramsbotham, O., 2005. Cosmopolitan peacekeeping and the globalization of security. International Peacekeeping 12, 139–156.

Relevant Websites

www.peacekeepingbestpractices.unlb.org – Peacekeeping Resource Hub.
www.un.org/en/peacekeeping/ – United Nations Peacekeeping.

Journals

International Peacekeeping. (Taylor & Frances, since 1994).
Journal of International Peacekeeping. (Brill, since 2009; In continuation of the 'International Peacekeeping: The Yearbook of International Peace Operations' Series).

Peace-Making in History

Peter J Lineham, School of Humanities, Massey University, North Shore, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Peacemaking in its contemporary sense is identified within the paradigms of theories of international relations and moral responses to war, and within this definition, examples and precursors are identified and analyzed within the ancient world, medieval and modern history. A particular focus is the growth of peacemaking in the nineteenth and twentieth centuries, first as a system of international pacification and secondly as pacification within civil society.

Introduction

Peacemaking is necessary because of the prevalence of war. It is a new term coined to label and identify activities that work to restore peace in situations of conflict. It is focused primarily on civil war and international conflict, although its methods are closely related to counseling models used to restore harmony within families and local communities.

Peacemaking is carried out by agents, individual and corporate and national, which seek means to lead nations and communities to cease armed conflict. Usually these agents are a third party, which acts as a broker, although sometimes peacemakers may emerge from within one or other of the warring parties. Peacemaking is thus an active strategy to limit war and violence. Processes include mediation, arbitration, and adjudication. Peacemaking may involve the process of seeking resolution to a conflict while the conflict is ongoing, but it may also involve some kind of arbitration of issues after armed conflict has ceased (Ceadel, 2000).

To study a modern concept in a historical context is to analyze changing concepts and practices over the centuries. The historian can identify changes in the practice of war and armed conflict over the centuries and in the same way can trace changes in patterns of reconciliation that led to the modern form of peacemaking.

The changes in warfare have not just been technological but also conceptual. A key aspect of change is the way in which the state has changed its definition over the years. Our definitions of the nation and our idealized concept of the democratic state bear little resemblance to the ancient or medieval state.

Theory Underlying Peacemaking

Notions of peacemaking arise from the application of modern social science theory to debates over the application of moral frameworks to warfare.

Contemporary international relations theories are typically divided into 'liberal' and 'realist' approaches. Realist and idealist theories are both useful in explaining international relations. Peacemaking seeks to adapt the insights of liberal theory so that it can function within the realist tradition. Rational choice theory is the underlying basis of realist theories. Such approaches see war as an inevitable tool used by states in order to strengthen their interests. Realists therefore tend to dismiss Just War theories as mere window dressing. Yet the idea that wars are commenced and ended solely for rational reasons often seems to be contradicted by the practices of 'warmongers.' Some protagonists clearly seek to prolong the struggle in order to inflict maximum damage on their opponents. Longstanding bitterness frequently contributes to the causes and conduct of warfare (Ripson, 2007).

The advocates of the realist analysis sought to curtail the application of morality to the affairs of nations. They analyzed the goals and strategies of armies and states in nonmoral language and focused primarily on the ways in which states furthered their own interests and their success. Yet they were ill-equipped to respond to the deep sense of moral outrage in populations who had lost territory and status through defeat.

The liberal theory of international relations seeks to apply moral language to international warfare, so that wars can be evaluated. Thus it represents a current form of a traditional ethical system, which debated the application of morality to warfare. The Just War theory in its various manifestations sought to constrain and limit warfare.

The theory of peacemaking draws to some extent on *jus in bello* theories that warfare should be just in its conduct and also the *jus ad bellum* that a war should have just causes and likelihood of success. Recent suggestions (notably by Michael Walzer) that war should also be just in the outcome (*jus post bellum*) would also fit with these theories. The criteria of Just War – especially the combination of *jus ad bellum* and *jus in bello* – place so many moral demands on the proponent of war that few wars seem likely to pass the test. Just War theory is thus an incentive to peacemaking. If war is a last and dubious resort, alternatives to war are both moral and practical necessities.

However, since Just War theory is a supposed guide as to when war is appropriate, it is relatively silent on the task of peacemaking. The pacifist traditions have been more aware of the value of working for peace, although only when they have been prepared to risk their traditional absolute opposition to moral compromise.

Peacemaking thus offers a discourse and more important a practice that speaks relevantly to both realists and liberal theorists. Peacemaking moreover brings the two moral responses to war – Just War theory and pacifism – into practical cooperation.

Secular and Religious Approaches to Peacemaking

Religion clearly has been a significant player is some of the most intractable conflicts. Religious wars are always more difficult to settle. Issues of bitterness and grievance can last for

very long periods. So it is appropriate that the practices of peacemaking should come from the same sources.

There has historically been a major issue in getting people to realize that they need to be peacemakers. Peacemaking is often most effectively provided by committed individuals and groups. In the nineteenth century peace societies emerged, which sought to promote missions. In many cases these were organized by Quakers and others who were supporters of peaceful solutions to global problems. William Lloyd Garrison, the antislavery campaigner took this approach in his New England Non-Resistance Society formed in 1838. Tolstoy and Gandhi were influenced by the same trends, sharing a vision of religious harmony, internationalism, and free trade. A peace congress held in 1871 promoted this vision. In World War I internationalist socialists refused to fight as did conscientious objectors to war. In the interwar years further organizations – the Fellowship of Reconciliation and the Peace Pledge Union – also encouraged people to decline to fight. However, none of this effectively made peace; it simply opposed war and almost invariably adopted unilateral actions. This is a typical pacifist approach that did not necessarily make peace.

A strong religious rhetoric is often a feature of moral responses to war. The religious traditions of Buddhism abhorred anger as a motivation in human affairs. Confucianism strongly condemned barbarianism and promoted a healthy secular discourse about the skills of diplomacy.

The practice of peacemaking has also developed strong support from recent moral and religious peace campaigners, many of whom speak of 'just peacemaking' to characterize their work. Distinguished religious leaders including the Dalai Lama, the Pope, and the Archbishop of Canterbury have deployed their own delegates for example Terry Waite. Other scholarly practitioners in various academic centers of peace studies have used psychological and counseling models for their peacemaking and have advocated systems of economic pacification. Mahatma Gandhi and Martin Luther King are example of the impact of high-profile practitioners of peacemaking with their own distinctive strategies. Non-governmental organizations (NGOs) and state agencies have also sought reconciliation between leaders, and former world leaders like President Jimmy Carter have played a significant role.

Part of the problem of peacemaking organizations is that they are based on the liberal rather than the realist tradition in international relations. New strategies have been needed to facilitate peace, shaped within the realist tradition, and this has been the focus of much recent centers of peace.

Religious discussions of war only occasionally acknowledge the influence of realism, but when they have done so their influence has broadened. Reinhold Niebuhr, for example, abandoned pacifism in favor of 'Christian realism,' which argues that the only effective way to limit war is to recognize its prevalence and to find practical ways to control it. Peacemaking fits well within this framework.

Peacemaking in the Ancient World

The ultimate test of the relevance of the concept of peacemaking as a tool of history is whether it applies in the ancient world. The evidence is mixed.

The constant warfare between the Greek states to some extent invited peacemaking, partly because of the role of democratic instruments in choosing to fight or to sue for peace. The Peloponnesian War between Athens and Sparta was a war between a league of states led by Sparta and the modest Athenian empire. Thucydides' account suggests that there were people working for peace in the midst of the struggle. Greek tradition did value peace as well as military virtues. During the ancient Olympics there was some kind of peace that was expected to dominate.

The Romans emerged from the same kind of citizen republic and in the republican period there were examples of peacemaking, but there was also increasing evidence of Roman reluctance to consider any concessions to their enemies. The famous Pyrrhic victory as Asculum in 231 BCE against the Romans seemed to some to be a price too high. The Romans instituted the *Pax Romana*, abolishing war within the empire, and policing its frontiers. They were not uniformly tough on their defeated opponents. Rome had historically effectively shown both a harsh face and a generous face in the aftermath of its victories. Conquered enemies were enslaved, but conquered cities were rarely destroyed. The Romans talked about a punitive peace, a *deballatio*, which generally involved occupation of a state until some kind of penalty or fine was exacted.

The three Punic Wars between Rome and Carthage are a critical turning point in this trend. They made Rome's reputation and brought it to a position of dominance in the Western Mediterranean. In the first two Punic Wars, Rome limited its revenge on Carthage. Then in the Third Punic War, the constant cry of Cato the elder, 'Carthago delenda est,' insisting on the destruction of Carthage urged for a harsh settlement of a territorial threat. At his insistence genocide was pursued, leading to half a million deaths, and many more placed into slavery. Romans pondered the meaning of these events as part of their lessons in statecraft.

Overall, the Roman Empire was a harsh opponent, although Julius Caesar was famous for a generous approach to enemies, and latterly enemies were frequently incorporated as mercenaries into the state.

Theories of peacemaking were guided by the teachings of Aristotle and Cicero and their notion that the moral criteria that justified war should guide the return to peace. The ancient notion of the cycle of history required that there be a return to calm after war. War served as a tool for the righting of wrongs. Cicero in *De Officiis* urged that the end of warfare was a moral victory, where the vanquished acknowledged their moral failure, and then peace was truly attained. Thus Cicero placed a high value upon peace.

Contemporary peacemaking can thus find few precedents in ancient history, except from the Greek age, where democratic states were so frequently at war and in search of peace.

Peacemaking in the Medieval World

In the medieval period, there are many examples that have some similarities with modern peacemaking, although there are also profound differences. One factor that would be unthinkable in the modern approach to peacemaking was the discomfort with religious diversity. Religious diversity certainly

existed in the medieval period but was never regarded as a satisfactory in the long term. This principle explains the crusading tradition that is so antithetical to peacemaking.

Feudal society lacked concentration of power. Authority was dispersed and armed knights were not fully accountable for their actions. Consequently conflicts over lost rights and offended rights were frequent. Many conflicts in the medieval period were the consequence of feudal offences, but this also meant that within a hierarchical society, feudal communities sought arbitrators for warfare.

The key moral authorities were feudal overlords and the church. Their priority lay in reducing the level of violence. The church attempted to curtail violence by regulating the days on which soldiers could resort to warfare under the Peace of God and the Truce of God. The empire and church leaders frequently sought to negotiate peace with honor, sometimes using bribes or other inducements. Church penitentials identified warfare as sinful. The empire believed it was responsible to preserve the Pax. Dependent cities even sent emissaries for peace to their own capital.

The Crusades left a reputation for the brutality of religious and ideological warfare, and they significantly disrupted the Eastern Empire where peacemaking with the new forces of Islam had been remarkably successful. In Western Europe, the Crusades were in part an outlet for the desire for knightly valor, which had been significantly curtailed by attempts to regulate warfare between fellow Christians, which had restricted both the theory of justified war and the days on which knights were permitted to practice it.

As a general rule, medieval theories of the Just War looked to Augustine's theory, in which war was permitted (in contrast to the Christian caution about the vocation of the soldier). Thomas Aquinas spelled this out in his theory in which just cause, right agents, and just goals were all required in a morally justified war. While the concept of peacemaking is not spelt out in this doctrine, the application of Just War doctrine forced kings to consider.

Peacemaking in Modern European History

A significant feature of the nation-state was the placement of formal diplomatic representation in neighboring kingdoms, increasingly on a permanent basis. The first of these were between the small states of Italy at the end of the fifteenth century, but the pattern rapidly spread to the new kingdoms. These diplomatic agents were not usually entrusted with negotiating peace. This required special emissaries in neutral locations.

The interests of the monarch were of central importance in diplomacy and had to be guarded in peacemaking. And consequently status factors were central in war and in peace. Kings were cautious about any hint of independence in the development of foreign offices and the bureaucratization of foreign affairs, although this did occur in the eighteenth century.

While hegemonic forces dominated the European state system, the imperial force was usually endowed with the power of coercion, which could be used aggressively or in the maintenance of peace. Both the Habsburg Empire and the French Bourbon hegemony worked hard to reduce tensions. The advent of a multipolar system in the eighteenth century made the maintenance of peace more precarious. States combined together to block the emergence of a hegemonic power.

This was coupled with the rise of the nation-state and it made the resort to war in the pursuit of national interests much more calculated. The loss of hegemony in the national structure of Europe produced a tradition of alliances and by the eighteenth century a 'balance of power.' Realists saw this balance of power as a way to maintain the peace, whereas others held that a thoroughly developed theory of nations should provide a settlement of grievances.

The rise of Prussia is often cited as a classic example of calculated risk taking in the pursuit of national aggrandisement. In this case, the balance of power and the diplomatic revolution of 1756 were strategies of Austria and France designed to thwart the aggressive stance of Prussia.

General settlements after periods of war offered opportunities for adjustments. The various settlements at the end of the religious wars in 1648, the Peace of Utrecht (1714), the Vienna Treaties at the end of the Napoleonic wars, and the Versailles Settlement at the end of World War I, have each been scrutinized for the their ability to sustain peace.

The processes of peacemaking in the first secular polemical war, the wars of the French Revolution and the wars of Napoleon, are a useful model for the model of peacemaking. There was widespread debate over the nature of the war and the ways to end it, which produced sharply different outlooks from states and from the supporters of the French. The supporters believed in the justice of the cause of the peoples. A party of 'Friends of Peace' emerged among English Whigs in the 1790s who conceived of the issue in geopolitical terms. Immanuel Kant in his book 'Perpetual Peace' (1795) proposed a vision for a peaceful world of nations living in harmony. However, the Napoleonic conflict showed the difficulties of realizing this vision.

The Congress of Vienna and associated treaties were the first to develop models seeking to ensure ongoing peace. Scholars have noted that by addressing broader issues of geopolitics the Congress recognized the deeper dimensions of maintaining the peace. They worked carefully to realign national borders in order to balance state interests and avoid border disputes. Congresses were to be held when necessary to deal with issues threatening war. This vision was a conservative anticipation of twentieth-century internationalism. Countries that had experienced 20 years of war wanted a stable Europe. They were willing to subscribe to a new system of the Balance of Power doctrine, expressed in the Quadruple Alliance. This positively rehabilitated France into European power politics, after a short period of occupation. It also laid down a system of congresses to deal with new issues threatening war, and congresses were held at Aix la Chappelle in 1818, Troppau and Laibach in 1820–21, and at Verona in 1822. However, the heavy-handed conservatism of Prince Clemens Metternich (expressed in the 'Holy Alliance') eventually antagonized the English tradition of constitutionalism and isolationism, and this brought the congresses to an end.

The pattern of consultations between the great powers to deal with international issues survived. The 'concert of powers' frequently consulted to remove potential causes of

warfare. Thus, in the nineteenth century unification nationalism was not wholly a negative force, but the large imperial powers were the critical agencies for peacemaking. The concert accommodated the emergence of various national movements and had a number of signal successes, including the Treaty of London, which guaranteed Belgian independence. The role of Louis Napoleon in brokering peace in Italy to his own political advantage in domestic politics showed an interesting combination of realism and idealism, as he supported the Risorgimento dream of a united national state. Other great power negotiations focused on Schleswig Holstein, the Crimea, Syria, Luxembourg, and Crete. There were other failures, signaled by the Crimean war. It has been argued that after 1848 the resort to war became more common as the conservative tone of Metternich collapsed. It did not succeed in preventing the wars of 1866 and 1870 because when one of the powers determined on warfare, the concert broke down.

The theory of peacemaking at that time was a theory of diplomatic responsibility. Warfare was seen as an instrument of state and politics. Clausewitz, the great military strategist, appreciated the links between war and politics. International relations were based on the assumption that states would act in their national interests and that war was rarely in the national interest. There were also particular factors that affected the peacemaking role of different states. Metternich was preoccupied by the risk of violent revolution. Britain under Palmerston was pressured by a growing democratic system, which idealized moral intervention. Prussia under Bismarck tightly integrated internal and external policy for maximum advantage.

Meanwhile, voluntary organizations committed themselves to a more peaceful international community, some from religious motives and some working from an enlightenment vision of peace and harmony between nations. Peace Societies, sometimes with Quaker links, also contributed. The first was founded in Britain in 1816 and this gradually developed what has been called a pacificist tradition, which stopped short of the pacifists' moral condemnation of all war. A series of national and international congresses gained popular support for ideas to limit warfare. In 1863, the Red Cross movement was founded in Geneva by Henri Dunant and Gustave Moynier after the wars between Austria and Sardinia in Northern Italy. The goal of the organization was the protection and aid for injured soldiers. This was followed at the instigation of the Swiss Government by the first Geneva Convention in 1864, revised in 1906. The Hague Convention was agreed in 1907, which provided laws and customs for land wars. The Permanent Court of Arbitration founded in 1899 was the first permanent international body. The tradition of humanitarianism assisted understanding of the abiding consequences of war and thus laid a foundation for peacemaking. Meanwhile, the Olympic Movement was founded in 1894 with a vision of bringing nations together. The Papacy under Leo XIII having withdrawn from active involvement in Italian affairs gained a reputation as a peace broker. There were also regional peace organizations notably in the Americas as a result of the Hague Conference of 1900. The Nobel Peace Prize was established in 1901 and showed the growing sense of value placed on negotiated peace.

Peacemaking in Non-European History

There are significant different traditions toward the responsibility for peace in the Chinese worldview of itself as the middle kingdom, in the struggles for control of the Indian subcontinent and in African tribal traditions. Some of these diverse approaches became critically important with European colonial expansion. European states intervened in global affairs, disturbing existing structures by exploiting their superior fire power. This asymmetrical relationship made peacemaking difficult, and the only balancing factor at the time was where different European states competed for influence. All such intervention created deep grievances.

Vitoria in his application of Just War theory to the Americas insisted that difference in religion and extension of empire were not acceptable justifications for war. There had to be a wrong committed, and therefore he held that the Spanish could not justify their colonial occupation of South America.

Today post-colonialism has led to a search for indigenous traditions of peacemaking, but so far this has contributed little to the resources of peacemaking.

Twentieth-Century Emergence

It is in the twentieth century that these various factors slowly assembled into a more coherent peacemaking tradition.

After the total warfare of the period 1914–18, the World War I peace negotiations at Versailles reverted to the early ideas of a general settlement. Different ideas of that settlement by the victorious powers left a sense of grievance behind among the defeated German and Austrian powers. The territorial penalties exacted by the French left profound grievances behind. Moreover, the sense of the failure to create a just peace was widespread in the victorious powers. The League of Nations was founded. These factors provided moral grounds for the policy of appeasement, although this was also based on a realistic appraisal of the inadequate preparations for war by the powers. Chamberlain's famous visit to Hitler and the Munich agreement could indeed be regarded as one of the first significant examples of peacemaking.

The Just War tradition profoundly shaped the allied case for fighting World War II. Attempts to negotiate an end to the war were impossible in the light of the absolutist stance of both Germany and the allies. Consequently, the allied war aims called for unconditional surrender of Germany and Japan and justified war trials of German and Italian leaders. Yet the aftermath of World War II did also include attempts to avoid repeating Germany's sense of grievance after World War I, and the Marshall Plan was a remarkably successful attempt to continue peacemaking while the United Nations created the basis of much of the modern discourse on peacekeeping and peacemaking.

The nuclear age and the Cold War restricted the development of peacemaking in the subsequent 40 years. Just War language struggled in the face of the threat of nuclear war. The CND campaigns created a great sensibility nuclear pacifism. East and West in the Cold War developed their positions on realist lines and found on frontiers in Vietnam and other parts of the third world. The ideological rift between the two sides

was profound. Western protestors against the legitimacy of these wars did sponsor some peacemaking missions to Vietnam. Peacemaking thus developed as an attempt to find positive ways to overcome the ideological rift. The United Nations became a crucial setting for negotiations, but some peacemaking avoided this forum.

The formation of the state of Israel and the profound rift with Arab states that resulted was another critical flashpoint during the Cold War and led to a range of great power and local initiatives. Notably in this case the leaders of Egypt and Israel were involved in significant negotiations to settle outstanding issues.

Internationalism

Conflict resolution at an international level has often borrowed the tools of personal and industrial conflict resolution. Such approaches seek each side to understand the point of view of the opponent and analyze both the rational basis of grievances and the feelings that lie behind the others' viewpoint before offering solutions. Some of the means used may indeed be highly relevant, although in foreign affairs this is even more difficult to attain.

The Liberal peacemaking tradition works on the theory that democracies do not easily go to war. Formalized systems and procedures check the advance into war and curtail personal grievances. Yet democracy is sometimes a major obstacle to peacemaking initiatives, as populist voices oppose the righting of grievances. External aggression has proved to be one way to win democratic support.

The conventions of diplomacy have increasingly provided tools to settle issues. Occasionally nations have refused to accept the binding authority of treaties (*pacta sunt servanda*) but these conventions have been highly significant. Some neutral states, notably Switzerland, have been able to play a critical role as intermediaries between other states.

The formation of the League of Nations was very significant as a focus of internationalism. Its goal was collective security and disarmament, and it sought to regulate the conduct of states including regulating against illegal cession of states, but it lacked any military force and was undermined in the 1930s by the withdrawal of belligerent powers. The Locarno Treaties of 1925 sought to normalize relations between Germany and Western Europe. The Permanent Court of International Justice and the International Labor Organization, the Disarmament Commission, the Health Organization, and a number of other bodies were associated with it. The League did play a part in the solution of a number of territorial disputes, but its most ambitious schemes, disarmament and the Kellogg Briand Pact of 1928 and attempts to intervene in disputes between China and Japan, between Italy and Abyssinia and in the Spanish Civil War were massive failures.

The United Nations marked a further step in the evolution of peacemaking structures and rhetoric. Acute awareness of the failures of the League led it to be more closely related to a realistic appraisal of the problems of making and keeping the peace. Its peacemaking and peacekeeping role required the supervision of the Security Council and the agreement of the great powers, each of which was given a veto. This curtailed the powers of the General Assembly, and during the Cold War it rendered most attempts at peacemaking ineffective.

Peace movements particularly focused on the nuclear threat. The Baruch Plan was tabled at the United Nations in 1946, inviting the United States to renounce the use of nuclear weapons so long as no others would develop them. This proved an impossible bargain for those who did not trust the benevolence of the United States. Instead the USSR tested its own atomic weapons in 1949. The Stockholm Peace Appeal of 1950 in fact was one of the steps toward the growth of the Cold War.

Peacemaking work was based on chapter 6 article 33 and chapter 7 of the United Nations Charter. Resolutions have strengthened the powers under this role, but left them condition on the endorsement of the Security Council. Although this has been much criticized, there have been some successes, especially after the end of the Cold War period and the inevitable boycotts of that period. Many studies have shown that the organization is particularly successful over a long period for example in Sierra Leone, where through trial and error solutions are identified and resourced. More immediate flashpoints like Bosnia and Syria and Iraq have proved incapable of great power agreement or the effective neutralization of the issues essential for an internationally agreed solution. The failures of peacemaking can also be linked to the problems in sustaining effective peacekeeping.

The United Nations is however only one aspect of the apparatus of peacemaking. Many other international agencies play an active role. The World Court at the Hague has sought to strengthen peacemaking by treating human rights crimes in wartime as subject to criminal prosecution. War Crimes Tribunals may, however, contribute to the sense of grievance, because of their retroactivity and selectivity. Recent Truth Commissions have included investigations and involving processes of forgiveness as well offer a significant modification of the model.

Other NGOs and think tanks have also sought to foster international understanding and reduce the buildup of hostilities in specific parts of the world, but they all require crucial agreement from sovereign nations at critical points.

Contemporary Peacemaking in Historical Perspective

The argument of this article is that peacemaking is largely a contemporary phenomenon, but it is profoundly affected by the historical legacy of its precursors. There was a rise in civil wars after the end of the Cold War. Also, since the end of the Cold War the problems of peacemaking have been exacerbated by the emergence of terrorism and by the extreme responses of nations to it. The very language of terror makes any action exceedingly difficult.

Contemporary peacemakers have also sought to address internal disruptions and civil disorder to a greater extent than in many of their historical predecessors, and internal and international disputes are more than ever interconnected in an age of terrorism and undeclared warfare. There has been a remarkable institutionalization of peace researchers, supported by government and international contracts and student idealism. The study of peace has also proved highly relevant to the military of many countries.

Peacemaking has a history, but present models of peacemaking are in such distinctive circumstances that historical precedents have limited relevance. However, the realistic development of peacemaking requires that mixed heritage to be taken seriously.

See also: Arms Control; Conflict Mediation; International Arbitration; International Law and Treaties; Militarism; Nation-State and War; Peace Keeping; Peace Movements; Peace Processes; Peace Promotion, Psychology of; Peace and Nonviolence: Overview of Anthropological Aspects; Peace; Territorial Conflicts; War and Democracy; War and Nationalism; War, Sociology of; Warfare in History; Wars among Nation-States: Patterns and Causes; Youth, Music, and Peace Building.

Bibliography

Anderson, M.S., 1993. The Rise of Modern Diplomacy 1450–1919. Longman, London.

Barkan, E., 2009. Introductions: historians and historical reconciliation, AHR Forum: truth and reconciliation in history. American Historical Review 114 (4), 899–913.

Ceadel, M., 1996. The Origins of War Prevention: The British Peace Movement and International Relations 1730–1854. Clarendon Press, Oxford.

Ceadel, M., 2000. Semi-detached Idealists: The British Peace Movement and International Relations 1854–1945. Oxford University Press, Oxford.

Carter, A., 1992. Peace Movements: International Protest and World Politics since 1945. Longman, London.

Clausewitz, C. von, 1989. In: Howard, M., Paret, P. (Eds.), On War. Princeton University Press, New Haven.

Cortright, D., 2008. Peace: A History of Movements and Ideas. Cambridge University Press, Cambridge.

Darby, J., MacGinty, R., 2008. Contemporary Peacemaking: Conflict, Peace Processes and Post-war Reconstruction, second ed. Palgrave Macmillan, New York.

Kegley, C.W., Raymond, G.A., 1999. How Nations Make Peace. St Martin's Press & Worth Publishers, New York.

Ripson, N.M., 2007. Peacemaking and democratic peace theory: public opinion as an obstacle to peace in post-conflict situations. Democracy and Security 3 (1), 89–113.

Ryan, S., 2000. The United Nations and International Politics. Macmillan Press, Basingstoke.

Smith-Christopher, D.L. (Ed.), 1998. Subverting Hatred: Peace and Nonviolence in Confucianism and Daoism. Boston Research Centre for the 21st Century, Boston.

Stassen, G.H. (Ed.), 2008. Just Peacemaking: The New Paradigm for the Ethics of Peace and War. Pilgrim Press, Cleveland.

Thistlethwaite, S. (Ed.), 2011. Interfaith Just Peacemaking Jewish, Christian, and Muslim Perspectives on the New Paradigm of Peace and War. Palgrave Macmillan, New York.

Tikhonov, V., Brekke, T. (Eds.), 2012. Buddhism and Violence: Militarism and Buddhism in Modern Asia. Routledge, London.

Van den Dungen, P., Wittner, L., 2003. Peace history: an introduction. Journal of Peace Research 40 (3), 363–373.

Relevant Websites

http://www.incore.ulst.ac.uk/ – International Conflict Research Institute, University of Ulster.

http://justpeacemaking.org/ – Just Peacemaking Initiative (G.H. Stassen, Fuller Theological Seminary).

http://kroc.nd.edu/ – Kroc Institute of Notre Dame University.

http://scar.gmu.edu/node/8824 – School for Conflict Analysis and Resolution, George Mason University.

Peace Movements

Marco Giugni, University of Geneva, Geneva Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The origin of peace movements can be traced back to the early nineteenth century, with the foundation of the first peace societies in the Anglo-Saxon world. Issues addressed by the movements include the general fight against war and promotion of peace (including internationalism), antiwar mobilization, nuclear disarmament (including nuclear test ban), mobilization against military infrastructures, and for civil service. Different phases can be discerned in the Western context: the rise of pacifism as a collective and public issue during the nineteenth and early twentieth century; the Cold War era; peace movements as part of the new social movements from the late 1960s to the late 1980s; and the post-Cold War era. The strength and specific features of peace movements vary both across time and across space depending on the specific features of each national context. Today, peace movements are seen as part of the broader family of the new social movements. Scholarly works have characterized the profile of participants in these movements as being rooted in the new middle class, displaying left-libertarian values, and sharing a common concern over social issues, but have also stressed important difference across countries in their social bases. Peace movements find their most important effects at the societal and cultural level rather than at the political level.

Peace movements can be tackled from different angles. Three such angles deserve mention. First, one may look at their underlying ideology and values. Most obviously, this involves fighting war and all the means leading to it (such as weapons or more generally armies) as well as promoting peace in a variety of ways (such as opposing war efforts, sensitizing the public opinion, and educating the younger generations). Second, one may focus on the actors involved. Since social movements are a collective endeavor, this means mainly examining the organizations – both formal and informal – involved in collective efforts to fight war and promote peace. Third, one may stress the actions carried out to fight war and promote peace. Since social movements can be seen as public displays of worthiness, unity, numbers, and commitment (Tilly, 1999), this implies looking at the public expression of the struggle against war and the promotion of peace.

Peace movements, in fact, are all three things at the same time: ideas, people, and actions. In order to get a grasp on their rise and development over time, we then need to consider a broader definition. In this vein, Della Porta and Diani (1999) argue that social movements are informal networks, based on shared beliefs and solidarity, which mobilize about conflictual issues, through the frequent use of various forms of protest. This qualifies peace movements and a special instance of contentious politics, which is a broader analytical category. As stressed among others by Tarrow (1998: 2), "[c]ontentious politics occurs when ordinary people, often in league with more influential citizens, join forces in confrontation with elites, authorities and opponents … When backed by dense social networks and galvanised by culturally resonant, action-oriented symbols, contentious politics leads to sustained interaction with opponents. The result is the social movement." This also distinguishes peace movements from pacifism, which is best confined to the realm of ideas.

From a thematic and substantial point of view, we may distinguish between different branches or, perhaps more accurately, thematic foci of peace movements. The most prominent are probably general fight against war and promotion of peace (including internationalism), antiwar mobilization, nuclear disarmament (including nuclear test ban), mobilization against military infrastructures, and for civil service. To that, we might add the nonviolent movement, which is, however, both narrower and larger than peace movements themselves. The relative strength of each thematic focus varies both across time and across space depending on the specific features of each national context.

From a historical point of view, peace movements have gone through a number of phases, at least in the Western context. Here we address four such phases: (1) the rise of pacifism as a collective and public issue during the nineteenth and early twentieth century; (2) the Cold War era; (3) peace movements as part of the new social movements from the late 1960s to the late 1980s; and (4) the post-Cold War era. Peace movements display specific features in each of these historical phases. The latter are not neatly delimited and sometimes overlap. For example, the peace mobilizations for nuclear disarmament that took place in the late 1970s and early 1980s were at the same time conducted in the context of an exacerbated Cold War climate.

Early Peace Efforts

Peace movements are undoubtedly among the major societal forces that have characterized the twentieth century. Their roots, however, can be traced back to the early nineteenth century. The first attempts to create an organized effort to promote peace emerged in the Anglo-Saxon world, in particular in Britain and the U.S. For example, the New York Peace Society and the Massachusetts Peace Society were founded in 1815, while the London Peace Society was created only 1 year later, in 1816. This early peace reformism, however, has little to do with the peace movements of the mid- and late twentieth century. It is more a matter of movements of ideas led by a small number

of intellectual elites, rather than popular movements involving thousands of middle-class citizens. We can thus talk of pacifism more than peace movements.

Some prominent pacifist associations were founded in the following decades in most Western countries, both nationally, such as for example, the American Peace Society (1928), and internationally, such as for example, the *Ligue internationale et permanente de la paix* (Paris, 1863) and the *Ligue international de la paix et de la liberté* (Geneva, 1867). As their names often suggest, many of these associations worked in the frame of an internationalist approach to peace, that is, based on the assumption that the latter could be reached only through a dialog across national actors and governments. Other transnational associations emerged from wartime pacifism, such as the British and US Quaker service committees (by 1917), the Women's International League for Peace and Freedom (1919), and the International War Resister League (1921). Although many efforts by peace movements have later focused on influencing national governments and public opinions, such an internationalist approach remains alive within the movement until nowadays, and has in fact found new vigor with the rise of the global justice movement in the 2000s.

An important feature of early efforts to fight war and promote peace consists in the strong religious background of peace associations. Often such associations were created that have a strong religious background, in particular of the Christian religion. More generally speaking, often national peace movements divide in two main orientations: a Christian (either Catholic, Protestant, or both) orientation and a leftist (either Socialist, Communist, or both) orientation. The relative strength and importance of each stream depends among other on the cleavage structure in each specific country. Other streams have emerged later on, in particular with the rise of the new social movements in the late 1960s and early 1970s.

Peace Movements under the Cold War

During and in the aftermath of World War II, peace activism and movements continued to flourish within a changed context. This is the context of the Cold War, where the Soviet Union and the U.S. engaged in a tug of war largely based on the nuclear arms race. As a result, peace activism in this period has often focused on nuclear disarmament not only in the U.S., but in Europe as well. At the same time, a peace coalition was formed during World War II around the United Nations ideal, thus continuing the strong internationalist perspective of peace movements. In this context, peace issue and human rights issues were put together in a broader effort to promote peace.

Yet peace activism during this phase mainly focused on the nuclear arms issue. Particularly in the U.S., a number of campaigns were launched which targeted nuclear weapons. This includes the nuclear test ban movements, led by a coalition of organizations that included the Committee for a Sane Nuclear Policy (Katz, 1986). More generally, a grassroots transnational movement formed around opposition to nuclear tests and nuclear arms more generally. This opposition also got the support of leading intellectual figures, such as for example, the British philosopher Bertrand Russell.

A major phase of peace activism occurred during the U.S. intervention in Vietnam and the war that stemmed from it between the mid-1960s and the mid-1970s. Obviously, the American peace movement was at the forefront of this cycle of protest. Indeed, opposition to the Vietnam War represents the peak of the U.S. peace movement, and the Vietnam War also spurred a wealth of protest activities in other countries. In the U.S. the antiwar movement gained momentum starting from February 1965, when American troops began Operation Rolling Thunder in North Vietnam. Opposition started with teach-ins in colleges and then evolved in a wide range of activities involving a broad coalition of congressional critics, liberal intellectuals, radical pacifists, New Left students, as well as disillusioned war veterans. Opposition escalated along with the war, both in terms of the number of events and the number of participants and in terms of the radicalization of the protest. The protest radicalized especially among students within university campuses, leading to what some have qualified of 'campus wars' (Heineman, 1993). At the same time, large mass demonstrations were staged across the country, such as for example, two demonstrations held in Washington: one on 15 November 1969 attended by nearly 500 000 people and another one on 24 November 1969 mobilizing between 200 000 and 500 000 people.

The 'New' Peace Movements

The late 1960s and early 1970s witnessed the emergence of the so-called new social movements. This term refers to a specific type of movements that emerged in Europe after World War II as a product of the shift to a postindustrial society and post-material age, stressing demands that moved away from instrumental issues relating to the class conflict toward post-materialist issues concerning the quality of life broadly speaking (Inglehart, 1977). As such the new social movements are seen as qualitatively different from 'old' movements, most notably the labor movement (Melucci, 1981), in terms of their social basis, ideology, value orientations, and organizational structure, but also in terms of their tactics and action repertoires.

Although some scholars have pointed the environmental movement as being the new social movement par excellence (Touraine, 1978), peace movements are seen as major component of this movement family. In this sense, the advent of the new social movements brought a new impetus to peace movements, bringing a younger generation of activists as well as novel forms of protest. Moreover, the mobilization of the 'new' peace movements reflected the changing structure of social and political cleavage that crossed the European societies in the decades after War World II.

Peace movements have indeed been a protagonist of the wave of protest that has characterized the Western world in the early 1980s. In particular between 1981 and 1983, pacifists across Europe as well as in the U.S. have gathered around the issue of nuclear disarmament, an issue that had been on the agenda of the movements for a long time, but that was took on a new dimension following NATO's decision in December 1979 to base 572 cruise and Pershing II missiles in five West European countries (Belgium, Great Britain, Italy, the

Netherlands, and West Germany). This decision, but above all its adoption by the governments of the countries concerned, led to the mobilization of an antinuclear weapons movement not only in those countries, but also in other countries, including in Eastern Europe. The result was perhaps the largest wave of protest in Western Europe since World War II, with hundreds of thousands of people getting into the streets to protest against NATO's decision and more generally against the nuclear arms race and for peace.

In the U.S., this opposition to nuclear arms took crystalized around the Nuclear Weapons Freeze Campaign. The idea of the freeze as an arms control proposal and a strategy for action was first proposed by peace activist Randall Forsberg in December 1979. It was asked "[t]o improve national and international security, the U.S. and the Soviet Union should stop the nuclear arms race. Specifically, they should adopt a mutual freeze on the testing, production, and deployment of nuclear weapons and of missiles and new aircraft designed primarily to deliver nuclear weapons. This is an essential, verifiable first step toward lessening the risk of nuclear war and reducing the nuclear arsenals" (quoted in Meyer, 1990: 160). The proposal was immediately adopted by pacifists and the movement gained momentum. Protest took several forms: from mass demonstrations and to disruptive actions, to teach-ins at colleges and universities, and to public education tactics. While the Nuclear Weapons Freeze Campaign can be situated between 1979 and 1985 (Chatfield, 1992), 1982 represented the peak of the mobilization. A number of important protest actions took place in that year, including the largest demonstration ever in the U.S. history at that point, when nearly 1 million people filled the street of New York City (12 June).

At the same time, scholars have shown that the mobilization capacity of peace movements – and of all social movements, for that matter – varies very much across countries as a result of different political contexts and opportunity structures (Kriesi et al., 1995; Rochon, 1988). For example, peace mobilization in the 1970s and 1980s was particularly strong in those countries where traditional cleavages such as the class and religious cleavage have been largely pacified, leaving room for the emergence of new cleavages based on postmaterialist issues, and where the configuration of power was favorable for mobilization around these issues. Thus, for example, Germany and the Netherlands witnessed a much stronger wave of protest than France (Kriesi et al., 1995). The ground was particularly fertile for peace mobilization in the Netherlands, as shown by the fact that the largest demonstrations in 1981 and 1983 perhaps took place on Dutch soil, as well as by the huge success of the 'people's petition' of the Dutch peace movement in 1985, which was signed by 3.8 million people (Kriesi, 1989).

Peace Movements in the Post-Cold War Era

The context for the mobilization of peace movements changed dramatically at the end of the 1980s with the fall of the Berlin Wall and the end of the Cold War. This implied a modified frame for pacifism and peace movement mobilization, from a focus on the East–West axis of conflict to an increasing focus on the South–North axis. Also, peace activism shifted away from the nuclear arms issue – with some notable exceptions such as the opposition to France's decision to run a nuclear test series at Mururoa in 1995, which, however, remained quite limited in scope – to a broader range of issues. At the national level, the latter varied depending on the country, addressing military infrastructures, military spending, arms sales, civil service, and so forth. At the international level, the most important events have dealt with opposition to military interventions.

One such events was the U.S. intervention in the Persian Gulf in 1990–91, aimed at driving Iraq out of Kuwait and called Operation Desert Storm, but which became known as Gulf War. This military intervention raised intense opposition from peace activists and more generally citizens. Such an opposition, however, the protest was as short-lived as the war. Quite understandably given the direct involvement of this country, the protest was stronger in the U.S., but peace movements in other countries mobilized as well. Opposition was manifested both before and after the military intervention, but the largest demonstrations as well as the most disruptive protest occurred once intervention began in January 1991. The protest waned as quickly as it rose after cease-fire was called on 28 February 1991.

Another event that spurred an international mobilization by peace movements was the military intervention in the Serb province of Kosovo in 1999 to terminate Serbian oppression in that province of the country made under the banner of the NATO. In part because other armies than the U.S. one were involved in this mission, but in part also because the intervention was done on European soil, this time peace movements in Europe were more active than American pacifists. Perhaps due to the geographical proximity of the country, the protest was particularly strong in Italy.

A few years later another military intervention in the Middle East raised strong opposition from peace movements worldwide. This is the invasion by the U.S. and its allies (Australia, Poland, and the UK) of Iraq between March and May 2003 in what become known as the Iraq War. Among a number of other activities, shortly before the intervention millions of people in more than 600 cities worldwide took the streets to protest against the invasion. In what has been depicted as the largest 1-day protest in human history (Walgrave and Rucht, 2010), pacifists from various countries staged demonstrations on 15 February to cry out loud 'The World Says No to War.' While some of the events were small-scale, others are among the largest demonstrations ever, such as for example, the estimated 3 million people gathered in Rome, Italy. But other large events took place in many other European cities as well as in the U.S.

These three examples illustrate quite well the main focus of peace movements on the global level in the post-Cold War era: protesting against military interventions (mostly by the U.S.) in foreign countries made in the name of freedom and democracy, but which is often seen by pacifists as a neo-imperialist move. Parallel to that, peace movements at the national level have addressed specific issues that may vary from one country to the other. However, these more specific issues only seldom have been addressed publicly, for example, by staging mass demonstrations, and have most often been the object of more 'hidden' advocacy activities by peace movement organizations.

Starting in particular from the late 1990s, peace issues were also addressed as part of a broader agenda of the global justice movement, along concerns around environmental protection, social justice poverty, and inequalities, both nationally and on a global scale. Of course, peace movements have always been strongly internationalized, since the early stages and even more so in the postwar period. However, during the first years of the third millennium peace came to be seen, even more than before, as a global issue and one which needs to be dealt with along other, related issues, most notably human rights and social justice on a planetary scale.

Participants

The literature on social movements can roughly be divided into two main subfields: one dealing with the movements as collective actors, including examining movement organizations (macro and meso levels), and another addressing individual participation (micro level). The same distinction may be applied to the study of peace movements.

Studies focusing specifically on peace activists or more broadly on individual participants in peace movements are rare. Most often, micro level analyses look at participation in social movements in general or, at best, in a given movement family. Scholarly works have stressed the specific profile of participants in the new social movements. Although no consensus exists on this matter, most accounts seem to agree on saying that the ideal-typical new social movement participants are rooted in the new middle class, display left-libertarian values, and share a common concern over social issues (Cotgrove and Duff, 1981; Eder, 1993; Kriesi, 1989). They are thus both structurally and culturally distinguished from 'old' movements such as the labor movement.

This characterization also applies to activists and participants in the 'new' peace movements. However, a number of studies in different countries show a variegated picture of peace movement participants (Parkin, 1968; Kaltefleiter and Pfaltzgraff, 1985; Walgrave and Rucht, 2010). In one of the rare broad comparative analyses of peace movements, the authors have examined a sample of participants in the 15 February 2003 protest against the war on Iraq (Walgrave and Rucht, 2010: 261) with the aim of "analyzing who those demonstrators were, why they took to the streets, and how they were mobilized." The book is based on a comparative protest survey of participants in 11 demonstrations against the (threatened) U.S. intervention in Iraq in eight countries (Belgium, Germany, Italy, the Netherlands, Spain, Switzerland, the UK, and the U.S.).

As the authors stress in the concluding article, "[r]egarding the sociodemographic profiles, of course, the average demonstrator, in the aggregate of all countries, resembled the typical new social movements' activists with high levels of education, a relatively large proportion of whom were women, belonged to the younger age cohorts, and predominantly worked in the human service sector" (Walgrave and Rucht, 2010: 163). At the same time, however, the authors of this research found that there are substantial differences among protesters in the eight countries if the study, attesting to the importance of the context in explaining not only the mobilization of peace movements at the macro and meso levels, but also the characteristics and motivations of individual participants at the micro level. Thus, protesters in the 15 February 2003 demonstrations displayed important cross-national differences in their sociodemographic profile, attitudes, and behaviors.

Outcomes

The study of the outcomes of social movements is at the same time the most neglected and perhaps the most difficult aspect to address, the former probably being a consequence of the latter. Studying the outcomes of peace movements is not an exception to this rule, and students of peace movements have faced the same obstacles as scholars interested in the effects of social movements in general, most notably the problem of establishing a causal nexus between protest and its alleged effects.

Scholars often make a distinction between three main areas of influence of social movements: political, biographical, and cultural outcomes (Giugni, 2008). How do peace movements score on these three counts? Most existing works address political, or even more narrowly, policy effects, founding mixed evidence. For example, while some have found that the U.S. congressional action on the Vietnam War was influenced, at least in part, by antiwar protests (Burstein and Freudenburg, 1978), others show that peace movements have little leverage on policy change (Giugni, 2004a). This certainly holds for the massive mobilizations that occurred in the early 1980s in various European countries as well as in the U.S. and aimed at fighting the nuclear arms race engaged by the two superpowers of the time. In spite of the millions of people involved and their prominence in the mass media, these mobilizations did not prevent the governments of the five countries who agreed with the NATO's decision from actually hosting the nuclear missiles on their soil. Thus, at least in the short term, what was at the time the largest wave of protest in the postwar period was unsuccessful.

Biographical outcomes refer to effects on the life-course of individuals who have participated in movement activities, effects that are at least in part due to involvement in those activities. Scholars have not addresses directly such effects of participation in peace movements, but have looked more generally at the biographical impact of activism in the American New Left during the cycle of protest of the 1960s and 1970s (see Giugni, 2004b for a review). In general, these follow-up studies of New Left activists quite consistently point to a strong and durable impact on the political and personal lives of activists. This suggests that involvement in peace movements, at least if through intense commitment, leaves a strong imprint on the lives of those who take parts in those protest activities.

The effects of peace movements, however, are perhaps most deeply felt at the broader societal and cultural level. Sensitizing national and international public opinions to peace issues, changing cultural orientations with regard to peace and war, affecting the values of the younger generations: this is perhaps where peace movements may have their strongest and most durable effects. That is to say, both pacifism as an ideology and peace movements as the combination of ideas, people, and

actions aimed at fighting war and promoting peace can have their deeper and lasting effects.

See also: Arms Control; Civil Wars; Cold War, The; Conflict and War, Archaeology of: Weapons and Artifacts; Conflict and War: Anthropological Aspects; Deterrence; Genocide and War; Insurgency; Media and Social Movements; Militarism; Military Sociology; Military, War, and Politics; Nation-State and War; National Security Studies and War Potential of Nations; Peace Keeping; Peace Processes; Peace and Nonviolence: Anthropological Aspects; Peace-Making in History; Peace; Political Protest and Civil Disobedience; Public Interest; Social Movements and Political Violence; Social Protest and New Media; Social Protest; War Propaganda; War and Democracy; War and Nationalism; War, Political Violence, and Effects on Children; War, Social Causes and Consequences; War, Sociology of; Warfare in History; Wars among Nation-States: Patterns and Causes.

Bibliography

Burstein, Paul, Freudenburg, William, 1978. Changing public policy: the impact of public opinion, anti-war demonstrations and war costs on senate voting on Vietnam War motions. American Journal of Sociology 84, 99–122.

Chatfield, Charles, 1992. The American Peace Movement. Twayne Publishers, New York.

Cotgrove, Stephen, Duff, Andrew, 1981. Environmentalism, values, and social change. British Journal of Sociology 32, 92–110.

Eder, Klaus, 1993. The New Politics of Class: Social Movements and Cultural Dynamics in Advanced Societies. Sage, London.

Giugni, Marco, 2004a. Social Protest and Policy Change: Ecology, Antinuclear, and Peace Movements in Comparative Perspective. Rowman and Littlefield, Lanham, MD.

Giugni, Marco, 2004b. Personal and biographical consequences. In: Snow, David A., Soule, Sarah, Kriesi, Hanspeter (Eds.), Blackwell Companion to Social Movements. Blackwell, Oxford, pp. 489–507.

Giugni, Marco, 2008. Political, biographical, and cultural consequences of social movements. Sociology Compass 2, 1582–1600.

Heineman, Kenneth J., 1993. Campus Wars: The Peace Movement at American State Universities in the Vietnam Era. New York University Press, New York.

Inglehart, Ronald J., 1977. Silent Revolution: Changing Values and Political Styles among Western Publics. Princeton University Press, Princeton, NJ.

Kaltefleiter, Werner, Pfaltzgraff, Robert L. (Eds.), 1985. The Peace Movements in Europe and the United States. Croom Helm, London.

Kriesi, Hanspeter, 1989. New social movements and the new class in the Netherlands. American Journal of Sociology 94, 1078–1116.

Kriesi, Hanspeter, Koopmans, Ruud, Duyvendak, Jan Willem, Giugni, Marco G., 1995. New Social Movements in Western Europe. University of Minnesota Press, Minneapolis.

Katz, Milton S., 1986. Ban the Bomb: A History of SANE, the Committee for a Sane Nuclear Policy, 1957–1985. Praeger, New York.

Melucci, Alberto, 1981. Ten hypotheses for the analysis of new movements. In: Pinto, Diana (Ed.), Contemporary Italian Sociology. Cambridge University Press, Cambridge.

Meyer, David S., 1990. A Winter of Discontent: The Nuclear Freeze and American Politics. Praeger, New York.

Parkin, Frank, 1968. Middle Class Radicalism: The Social Bases of the British Campaign for Nuclear Disarmament. Praeger, New York.

Porta, Donatella D., Diani, Mario, 1999. Social Movements: An Introduction. Blackwell, Oxford.

Rochon, Thomas R., 1988. Mobilizing for Peace: The Antinuclear Movements in Western Europe. Princeton University Press, Princeton, NJ.

Tarrow, Sidney, 1998[1994]. Power in Movement: Social Movements and Contentious Politics, second ed. Cambridge University Press, Cambridge.

Tilly, Charles, 1999. From interactions to outcomes in social movements. In: Giugni, M. (Ed.), How Social Movements Matter. University of Minnesota Press, Minneapolis, pp. 253–270.

Touraine, Alain, 1978. La voix et le regard. Seuil, Paris.

Walgrave, Stefaan, Rucht, Dieter (Eds.), 2010. The World Says No to War: Demonstrations against the War on Iraq. University of Minnesota Press, Minneapolis.

Peace Processes

John D Brewer, Institute for the Study of Conflict Transformation and Social Justice, Queen's University Belfast, Belfast, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Peace processes are defined as the series of actions that lead up to the resolution of conflict and the negotiation and signing of a peace accord, as well as the series of actions that implement and consolidate the agreement afterward. A critical distinction is made between political and social peace processes and the series of actions that address these different types are outlined and discussed. The tension that can exist between them is emphasized and reasons are given to explain why peace processes can be so fragile.

To understand what we mean by 'peace process,' we first have to deconstruct what is meant by the word 'process.' The term 'process' normally refers to a series of actions directed toward some end, which need not necessarily be linear but which have some goal as the conclusion. This means that the term 'peace process' describes the often uneven and discontinuous series of actions directed toward the establishment of peace as an end point. We can leave aside here the thorny question of what we mean by peace to concentrate on the series of actions by which it is established and consolidated. Establishment and consolidation are different goals and it is important to distinguish between the actions necessary to establish peace, some of which refer to conflict resolution and peace negotiations, and those which consolidate the conflict transformation to make it more secure after the accord has been signed and implemented. A peace process is thus the series of actions that lead up to the resolution of conflict and the negotiation and signing of a peace accord, as well as the series of actions that implement and consolidate the agreement afterward.

This definition, however, raises another important distinction which will be the main focus here, namely between the political and social peace processes (this was first developed in Brewer, 2010; Brewer et al., 2011). Political peace processes refer to the actions by which peace settlements are negotiated by political representatives and third parties, which involve the negotiation of an accord which brings an end to the violence by which the underlying conflict was pursued, as well as the series of actions established within the compromise agreement by which the terms of the agreement are monitored and the political parties are kept to their obligations under it. It is referred to as the political peace process because these actions involve political management (see Darby and MacGinty, 2000), which includes the establishment of new forms of politically representative institutions, fairer systems of voting, the introduction of human rights law and associated accountability mechanisms, the introduction of a free press, and the creation of procedures by which international observers monitor the accord. Emphasis is placed on a series of actions that introduce structures and institutions of good governance in the view that stable statebuilding is the necessary foundation on which all effective peacebuilding occurs (emphasized most strongly by Call, 2008; see Paris and Sisk, 2009 for some of the differences between the concepts of statebuilding and peacebuilding). In this way, the political peace process constitutes the necessary actions for achieving conflict transformation.

The social peace processes are concerned with the difficult question of societal healing and relationship rebuilding after conflict, and thus refers to the series of actions by which reconciliation in society occurs (on the idea of healing between nations rather than within society see Amstutz, 2004). These actions include truth and reconciliation procedures, forgiveness and atonement strategies, policies that facilitate and encourage public tolerance and compromise, new forms of memory work, memorialization and remembering, public apologies, new cultural symbols, such as national flags, anthems, and the like, and the reassessment and reevaluation of identity. In conflict societies where social inequality is high and was part of the causes of the conflict, as occurs often in settler and divided societies and with wars of decolonization, societal healing is greatly assisted by policies of social redistribution, for people's senses of justice in these types of postconflict society get wrapped up and confused with their experiences of inequality; the demand for justice is a surrogate for demands for social redistribution. When this confusion occurs, the social peace process is affected by the forms of economic policy adopted in the postconflict phase (see Boyce and O'Donnell, 2007) and must concern itself with social policies that address inequality as much as injustice. In all these ways, the social peace process involves a process of social transformation that is much wider than the process of conflict transformation and deals with issues that are deeply emotional and well beyond the usual terrain of the political peace process.

The series of actions through which societal reconciliation is addressed can occur in the preagreement phase as part of the political peace process, as attempts at relationship rebuilding are encouraged in the lead-in to the negotiations or the possibilities of compromise are explored as part of the political debate to enhance the success of the negotiations. Mostly, however, the social peace process builds on the political peace process as conflict transformation opens up spaces for exploring the possibilities of restoration and rebuilding once the violence has stopped or downscaled. The social peace process can be a precursor to the political peace process therefore, but tends to proceed apace only once the political peace process is underway.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.96043-4

The distinction between these two forms of peace process is important for understanding the tension that often occurs between them in the postconflict phase. It is often assumed that a peace process is successfully accomplished and comes to an end when conflict transformation results in a negotiated political settlement that introduces good governance structures and stable statebuilding. Fairer systems of voting, more representative forms of political institutions, legal systems that enshrine human rights law or bills of rights, a free press independent of regime interference and workable institutions for decision-making and policy implementation are among the main features of statebuilding as the route to successful peacebuilding. These actions form the main concerns of the political peace process. However, it is naive to assume that social healing and reconciliation automatically follow once problematic politics are negotiated away and good governance structures are implemented.

For one thing, political peace processes are mostly fragile (see Brewer, 2010, pp. 29–43), since accords go through various iterations and parties to the agreement often abrogate their responsibilities under it. There are many reasons for this. Some of these are obviously political. Weak governments often result in the new institutions becoming unworkable. Peace accords reflect the decision of all parties to opt for a second-best preference (the compromise arrangement) and to give up their first-choice option (victory) for the sake of peace. However, dissidents who oppose the settlement can continue with a campaign of violence to undermine the new institutions, escalate the costs to the economy of the peace process, and force the continuance of militarization in a way that inhibits both statebuilding and societal healing. This 'spoiler violence' (a term used in Stedman, 1997) destabilizes the political and social peace processes, such that ongoing violence is dangerous to the consolidation of peace and can perhaps call into question to capacity of the new governance structures to deliver or respect human rights, fairness, and justice (on the general negative effects of violence on peace processes see Darby, 2001). It is for this very reason that those who wish to see the continuance of conflict, make military targets of those moderates who are prepared to cut a deal, often reserving the worst barbarism for their erstwhile colleagues, and why levels of violence can actually increase during the interregnum when a peace accord is being negotiated (in South Africa's case see Du Toit, 2001), although the latter can also happen as a deliberate ploy to increase militants' bargaining power at the negotiating table.

Some reasons for the fragility of the political peace processes are also economic. Spoiler violence can be motivated by the wish to continue to benefit materially from a war economy, and war economies are often made all the weaker by the costs of militarization involved in pursuing dissidents and spoilers (on war economies see Collier, 1999; Keen, 1998). The descending cycle of failed economy, failed government, and failed statebuilding puts political peace processes to a severe test. War economies permit corruption to go relatively unchecked, often involving the fraudulent exploitation of foreign aid, as well as encouraging a system of patronage that facilitates rich war lords to retain large armies for personal as well as political and economic gain. The ending of war economies can be hugely problematic to the patrons as much as the war lords, especially where the economy is weak and does not offer economic opportunities that improve much on the level of patronage, giving them all a material interest in the continuance of violence. Where the patrons are children, in the form of either child soldiers or young girls used for sexual exploitation by adult male soldiers, the use of which has dramatically increased as forms of organized violence have intensified in late modernity (see Honwana, 2006), their economic reintegration is all the harder, especially where their education was interrupted by the conflict. War economies weaken states and make them open to further destabilization from dissidents, resulting in failed political deals in an endless cycle of violence.

Political peace processes can be fragile also for psychological reasons. This can be misunderstood in the mistaken view that the end goal of peace has to be universally desirable. Yet violence can be all that some people have known and peace comes with its own ontological costs. Peace processes require people to reevaluate themselves in relation to their former enemies. People are required to reassess who they are and who is their enemy, perhaps having to overcome centuries of cultural socialization and stereotyping of 'the other,' to now see their erstwhile enemies in government, and to question their own identities and former privileges. Peace can sometimes come at too high a price for some people, with the ontological insecurities of continued violence posing less of a challenge to their identity than peace itself. It is for this reason that, counterintuitively, some people can look back on political transitions through glasses darkly, seeing in the past greater security, stability, and safety than in newly democratic times. These people form a residual constituency of support for deposed antidemocratic regimes and parties not for political reasons but because they experience political change as increasing their ontological insecurity. It constitutes a form of transition stress that encourages romanticized views of the security found in failed regimes in the past.

However, another of the recurring weaknesses in political peace processes is that they leave unresolved the difficult question of societal healing and relationship rebuilding, or assume that this occurs naturally once problematic politics are resolved. The processes for social transformation are often neglected in the political peace process as parties to the negotiation concentrate on bringing warring factions to the table, keeping them there, and eking out the basis of an agreement through trying to establish workable governance structures and institutions. Peace accords therefore rarely discuss the measures by which postviolence societal healing is to be effected. Peace agreements, for example, rarely specify in detail how truth recovery is to be conducted, what policies are to be implemented to aid victims or socially reintegrate former prisoners, or how postconflict social redistribution should be addressed. But it could be argued that the social peace process only really begins once conflict transformation has begun and that peacebuilding requires as much attention to the issue of social reconciliation as it does to statebuilding. Paris and Sisk (2009), for example, have recently analyzed some of the contradictions between statebuilding and peacebuilding and highlight that effective statebuilding presupposes progress on issues within the social peace process like dealing with the past (for the contradictions between statebuilding and peacebuilding in Afghanistan see Rubin, 2006).

The social peace process requires policy makers to focus on a broader set of issues than is normally captured in our understanding of political peace processes. Some of these concerns are dealt with under the notion of transitional justice (on transitional justice in postconflict societies see Braithwaite, 2002; Brewer and Hayes, 2011; Elster, 2004), which is now understood to mean more than the criminal processing of conflict-related crime in either retributive criminal courts or restorative justice procedures. Transitional justice these days covers the whole process of social reconciliation and societal healing after conflict. Issues around amnesty for excombatants or early release of prisoners are particularly difficult to resolve given the preference many victims have for retributive justice through the criminal courts (for elaboration see Mallinder and McEvoy, 2011), but restorative justice procedures have been deployed in many cases where traditional notions of authority survived the conflict and where they already contained strong restorative elements, notably the famous *gacaca* courts in Rwanda (on which see Cobban, 2006). Where these traditional systems of authority have not survived, advocates of restorative justice policies had tended to prioritize truth recovery as the nearest form of restoration between victim and perpetrator (for example, Braithwaite, 2002). It is rare for truth commissions *not* to be part of postconflict recovery ever since they were popularized in Latin America (for a review see Hayner, 2002) and made universal as a result of the high-profile activities of the South African Truth and Reconciliation Commission (on which see Rotberg and Thompson, 2000; Shore, 2009; Wilson, 2001). They have become almost the panacea for societal reconciliation.

However, the social peace process is much broader than truth recovery. It involves policies for managing victimhood, assisting in the social reintegration of former combatants and child soldiers, citizenship education courses to address the past, new forms of memory work and memorialization, the various atonement strategies that deal with public apologies and forgiveness, policies of economic restructuring and equality of opportunity, and new identity formation processes and practices in art and performance (for a review of the role of art in peacebuilding see Shank and Schirch, 2008), literature and writing, peace journalism (for an introduction to this field see Lynch and McGoldrick, 2005), sport (sport and peace is a growing interest; see Giulianotti, 2011; Hoglund and Sandberg, 2008; Sugden, 2008), museums (on museums and peacebuilding see Crooke, 2005; Shearing and Kempa, 2004), gender mainstreaming (for a summary of the huge literature on gender and peacebuilding see Brewer, 2010, pp. 68–102), the school curriculum, among other things.

All of these things rightly come under the rubric now of 'peace process,' making it an interdisciplinary interest that extends beyond political science, international relations, and human rights law, the traditional disciplines that focused on statebuilding in the political peace process. Peace processes are now as much the domain of psychology, economics, memory studies, sociology, sports studies, drama and the performing arts, cultural studies, and theology and religious studies. Theologians, for example, write as much on political forgiveness as hard-nosed political scientists (for reflections on political forgiveness within theology and politics see Digeser, 2001; Griswold, 2007; Jankélévitch, 2005; Shriver, 1995; Torrance, 2006), psychologists on victim issues as well as medics (see Hamber, 2009), sociologists on atonement strategies as much as human rights lawyers (see Brewer, 2006; Misztal, 2003, 2011a).

Interdisciplinarity is not the only advantage that follows from this new way of approaching the conceptual field in peace processes. There are practical policy benefits. When the emphasis was placed entirely on the political peace process, the domain for its operation and implementation was correspondingly political. Peace processes were the responsibility of governments, and the political actors that made up or aspired to be governments, as well as the political parties who once formerly opposed them militarily. Political actors negotiated the settlements, they occupied the new governance structures and oversaw the new institutions, and they either endorsed or collapsed them, making political decisions about whether to abrogate or stick with it. International political actors often assisted in this responsibility, in the form of third party governments or international actors connected to regional power blocs or the United Nations. In this view, peace processes were political affairs, commanded over by politicians, the future of which was the responsibility of political actors.

Once we give recognition to the idea of the social peace process, peace processes become everyone's responsibility and the domain in which they function and are consolidated widens to include civil society (for an analysis of the role of civil society in peace processes see Brewer, 2010; Brewer et al., 2011; Kaldor, 2003; Van Leeuwen, 2009). It has become fashionable to emphasize the importance of civil society in peace processes, with several philanthropic foundations funding initiatives that promote the growth of civil society in postconflict societies (see Anheier and Daly, 2005). This recognition, however, is not just a consequence of the huge number of civil society groups that are involved in peacebuilding or the sizable funds now available to fund them, for civil society is a crucial domain in which the social peace process is enacted. While only a minority of peace agreements make explicit reference to civil society (see Bell and O'Rourke, 2007), as befits the orthodoxy that peace processes are essentially political, civil society has come to be seen as critical in several respects when dealing with the aftermath of conflict.

This usefulness includes the role of civil society groups in the allocation and distribution of resources and humanitarian aid, the monitoring of parties' obligations under peace accords, the provision of participative forums to widen civic engagement and deliberation, and direct involvement in societal reconciliation. In this regard social peace processes benefit from the involvement of, for example, faith-based organizations in dealing with issues that fall within their compass and competence, like public engagement with the meaning of forgiveness, hope, and healing (on the phenomenon of religious peacebuilding see Brewer et al., 2011; Coward and Smith, 2004; Herbert, 2003; Shore, 2009; Schlack, 2009), or women's groups with respect to encouraging gender empowerment, trade unions with pressing questions of inequality, or peasant groups with land redistribution, universities, schools, and youth groups in pushing for educational and curriculum reform and wider access to

education, medical aid agencies with respect to victims and trauma or in managing the medical problems left from gendered violence, or human rights groups with respect to continued abuses or in advancing the application of international norms of justice in whatever localized restorative justice procedure is adopted, and so on.

This fundamentally shifts responsibility for social peace processes to make them practitioner based not just politician led. Religious actors are not politicians, peasant leaders mostly do not want to build constitutions, or members of women's groups rule the country. Yet within spheres of competence, members of civil society groups can become effective peace agents as practitioners within the grassroots. Recognition of peacemaking as a practitioner activity has spurred the growth of civil society in postconflict societies as well as given impetus to the emergence of an industry in training manuals, training workshops, peace studies courses and summer schemes, and direct public engagement through forms of civic or peace education (the practitioner nature of civil society is emphasized in Van Leeuwen, 2009). This industry is not to be undermined, for some of the main weaknesses postconflict societies face lie with the feelings of disempowerment in the grassroots arising from the view that the peace process is the politicians' responsibility not theirs, as well as the tendency of civil society groups to reproduce the preagreement divisions and cleavages as they split along traditional lines (in Sri Lanka's case see Orjuela, 2008; on problems with the 'darker side' of civil society see Chambers and Kopstein, 2001). This means that civil society groups play a significant role in trying to persuade people to feel they can make a difference in the social peace process. In so doing they should not themselves reproduce the old lines of division but work together to deal with the pressing issues that define the problem of social reconciliation.

It is important, however, not to set up the political and social peace processes as mutually exclusive and competing series of actions. They have, after all, the same end goal, the conclusion of peace, and its consolidation. The relationship between them is also recursive. Progress in the social peace process prior to a political settlement, in the form of cross community and civil society initiatives that challenge the terms of the conflict and bring enemies together, can assist the negotiation of a peace accord, facilitating the success of the political peace process. On the other hand, the consolidation of the settlement and the bedding down of the new governance structures and institutions is greatly assisted if the problem of societal healing is broached and systematic attempts are made to facilitate social reconciliation. In other words, the social and political peace processes complement each other and problems arise in postconflict societies only if the policy emphasis and reconstruction and aid money neglects one in favor of the other.

They do, nonetheless, operate on different time frames. Governance structures and institutions can be quickly devised and implemented. Statebuilding in the political peace process can advance quickly, if not necessarily always successfully. Foreign aid money can bolster weak economies rapidly, so long as it is sustained, and new systems of voting can result quickly in fairer elections. This compressed time frame is partly the reason why peace processes can be wrongly perceived to have reached their conclusion when the new institutions are up and running. There is, of course, acceptance that new governance structures take time to institutionalize and become embedded but this is seen as just a matter of time. Institutions carry their own institutionalization in time. Social reconciliation, however, has to operate on a much longer time frame. This is sometimes a source of great frustration as outsiders – as well as some insiders – in the global publics as well as international politicians get exasperated when ancient conflicts erupt at flash points or when people in societies emerging from conflict seem not to have learned the lessons of the past. The social peace process, however, is of long duration. Some changes to cultural and national symbols can be quickly implemented – changes to flags and national anthems need little time (on the problem of managing culture in postconflict societies see McEvoy, 2011) – but shifts in culture, identity, socialization patterns, gender relations, and the rest take generations. This is one of the reasons why institutional reform in the political peace process can be seen to be progressing well while interpersonal antagonisms and group conflicts continue, and societal reconciliation seems as far away as ever.

Two things follow from these temporal considerations. Social peace processes require patience. They also demand sustained attention. If trust is a casualty of war, patience is of peace. Time frames of expectation are very short as people want peace on their terms and want it now. But forgiveness, truth, reintegration, and healing are for the long term, and patience is required because reconciliation proceeds slowly, stumblingly, and sometimes falters. The contradiction is clear. On the one hand, the social peace process poses people challenges that make it difficult to embrace former enemies, while conversely, people get frustrated and exasperated when old resentments and divisions resurface. This is the paradox of social peace processes. They have within them an inherent contradiction provoked entirely by the problem of expectations. There are two dimensions to the problem of expectations. Most people want societal healing but without themselves having to change. 'The other' has to make the compromises; it is 'them' not 'us' who have to reform. Most people expect accommodation in others, not themselves; they do not consider they have to move an inch. The second dimension to the problem of expectations is that most people want the benefits of peace to accrue to themselves and their group not to 'the other.' They do not want to see former enemies benefitting; if it seems that they are being rewarded, their commitment to the social peace process weakens in the view that it is one sided. For these reasons, the problem of expectations can stymie the social peace process, and the flash point outbreaks of renewed conflict the problem of expectations provoke can infect the political peace process and cause strain in the working of the new governance structures.

The problem of expectations can be resolved, however, by recognizing the importance of patience – patience toward others when they seem uncompromising and in ourselves when we feel the same. And the time frame for meeting expectations can be compressed if the social peace process is given the serious attention it deserves by practitioners in civil society and politicians and policy makers in the new

governance structures. Discharging this responsibility is critical, getting recognition of it as an obligation even more so. This means faith-based organizations have to become actively involved in managing the tensions within the social peace process as patience and compromise breaks down, so too women's groups, trade unions and peasant groups, businesses, universities, schools and youth groups, victim support groups, and the rest: it is not just the responsibility of politicians. In short, empowering civil society as practitioner peacebuilders accelerates the social peace process.

There is one further advantage to the new conceptual field by which we understand the nature and development of peace processes. Peace processes are not solely political processes anymore but moral ones (see Lederach, 2005 for the strongest statement of this). This has significant implications for how we conceptualize them. To understand this, it is first necessary to appreciate the changing nature of war and organized violence in late modernity (on which see Malesevic, 2010). Organized violence has escalated in this period and transformed into what Kaldor calls 'new wars' (1999), dominated by intrastate conflict, without a set battle zone and involving large numbers of noncombatants: violence of a kind that no longer involves professional armies but engages the whole of society, whose resolution equally prompts a broad response that involves civil society (a point emphasized by Van Leeuwen, 2009, p. 38). This new form of organized violence also involves the human body being turned into a battle site as barbaric atrocities come to be inflicted on enemies, sometimes with premodern forms of weaponry like the machete, often with sophisticated weaponry that is now more easily accessible, like land mines (as this affects women's bodies in particular see Hynes, 2004). Turning the body into a battleground reflects the moral enervation of the enemy, in which they are stripped of moral worth and dignity and not recognized as human or deserving respect as humans. This is most obviously the case in instances of genocide (on which see Shaw, 2007), which is designed to destroy a whole people, but is also the case in other forms of organized violence (see Halpern and Weinstein, 2004). Paradoxically, this expansion in the barbarity of organized violence is matched simultaneously with the increased regulation of the conduct of war, with the growth in institutional norms that monitor war crimes, the introduction of United Nations' regulatory mechanisms of accountability and enhanced international intervention in local and regional wars, all justified on moral grounds to protect the human rights of innocent civilians. This has increased the tendency of the international community to intervene in local wars not only for reasons of *realpolitik* but now also for humanitarian purposes. Reconciliation is seen as a moral imperative (see Walker, 2006).

The combination of these two processes – greater levels of organized violence and its increased international regulation – has made peace processes and postconflict reconciliation burgeoning intellectual interests for a wide range of practitioners and academic disciplines because peace is recognized as raising deeply moral questions that touch on the nature of the normative values that define late modern ideas of humanitarianism (or what is increasingly called cosmopolitanism, on which see Beck, 2012; Delanty, 2009). There is a profound cultural shift occurring globally into which this moral interest in peace processes has to be embedded. Many sociologists advance the claim that feelings of responsibility toward strangers and their 'distant suffering' are beginning to mark the human condition (see Boltanski, 1999). Globalization has compressed time and space in a way that brings distant others into the daily experience of late modern life. Television brings distant suffering into people's living rooms and new forms of social networking technology like Twitter and YouTube collapse time and space, giving distant suffering immediacy to the point where images of distant suffering bombard late modernity. Global digital media give visual form to the material conditions of political economy that cause the distant suffering and which provoke a humanitarian response (see Sayer, 2000 for the link between the moral economy of late modernity and its political economy). This is in part due to an emerging sense of the shared vulnerability of humankind. Sociology is beginning to pick up on this shift, affecting sociological treatments of human rights (Turner, 2006), the emergence of what Wilkinson calls the sociology of suffering (2005), Mizstal's (2011b) attempt to chart the sociology of vulnerability, in Delanty's outline of late modernity's 'cosmopolitan imagination' (2009) and in notions of justice that prioritize human dignity over socioeconomic redistribution (for example, Wolterstorff, 2010).

This moral sensibility, of course, should not be exaggerated. Antiforeigner sentiments and aggressive xenophobia are equally evident and exist alongside ethical cosmopolitanism. As the German sociologist Ulrich Beck (2012) argues however, the world is in a process of cosmopolitanization, of arriving toward rather than yet having achieved cosmopolitanism. Cross-national comparative research by Pichler (2012) on cosmopolitanism at the interpersonal level (rather than the usual macro-level attention to cosmopolitanism) shows it to be uneven across its various dimensions (also see Norris and Inglehart, 2009), such as in politics (global governance), identity (notions of global citizenship), and ethics (empathy and trust toward others). Using World Value Survey data from 49 countries between 2005 and 2008, Pichler (2012, pp. 28–31) shows ethical cosmopolitanism to be more prevalent than political cosmopolitanism, especially in the global first world and among urban dwellers, the highly educated, and professionals. A cultural shift is thus unevenly occurring, but in significant sections of the population in many late modern societies, moral sensibility is motivating an interest in peace processes. It is for this reason that moral philosophers (for example, Margalit, 2009), theologians (for example, Brudholm, 2008; Brudholm and Cushman, 2009), cultural sociologists (for example, Alexander, 2012), and the like concern themselves with the nature of societal healing after trauma. Moral reflections like these on the boundaries and limitations of memory (for example, see Ricoeur, 2004; Volf, 2006), forgiveness (for example, Brudholm, 2008), and truth (for example, Braithwaite, 2011; Rotberg and Thompson, 2000) are turning the study of peace processes into a modern form of an ancient concern with the nature of moral sensibility.

See also: Civil Wars; Conflict Mediation; Conflict and Conflict Resolution, Social Psychology of; Cosmopolitanism; Justice,

Conflicts, and the Justice of Conflict Resolution; War Crimes Tribunals, Crimes against Humanity and Transitional Justice.

Bibliography

Alexander, J., 2012. Trauma: A Social Theory. Polity Press, Cambridge.
Amstutz, M., 2004. The Healing of Nations. Rowan and Littlefield, London.
Anheier, H., Daly, S., 2005. Philanthropic foundations: a new global force? In: Anheier, H., Glasius, M., Kaldor, M. (Eds.), Global Civil Society 2004/5. Sage, London.
Beck, U., 2012. Redefining the sociological project: the cosmopolitan challenge. Sociology 46, 7–12.
Bell, C., O'Rourke, C., 2007. The people's peace? International Political Science Review 28, 293–324.
Boltanski, L., 1999. Distant Suffering. Cambridge University Press, Cambridge.
Boyce, J., O'Donnell, M., 2007. Peace and the Public Purse. Lynne Rienner, Boulder.
Braithwaite, J., 2002. Restorative Justice and Responsive Regulation. Oxford University Press, Oxford.
Braithwaite, J., 2011. Partial truth and reconciliation in the longue durée. Contemporary Social Science 6, 129–146.
Brewer, J.D., 2006. Memory, truth and victimhood in post-trauma societies. In: Delanty, G., Kumar, K. (Eds.), The SAGE Handbook of Nations and Nationalism. Sage, London.
Brewer, J.D., 2010. Peace Processes: A Sociological Approach. Polity Press, Cambridge.
Brewer, J.D., Hayes, B.C., 2011. Post-conflict societies and the social sciences: a review. Contemporary Social Science 6, 5–18.
Brewer, J.D., Higgins, G., Teeney, F., 2011. Religion, Civil Society and Peace in Northern Ireland. Oxford University Press, Oxford.
Brudholm, T., 2008. Resentment's Virtue. Temple University Press, Philadelphia.
Brudholm, T., Cushman, T., 2009. The Religious in Responses to Mass Atrocity. Cambridge University Press, Cambridge.
Call, C., 2008. Building States to Build Peace. Lynne Reiner Publishers, London.
Chambers, S., Kopstein, J., 2001. Bad civil society. Political Theory 29, 837–865.
Cobban, H., 2006. Amnesty after Atrocity? Paradigm, Boulder.
Collier, P., 1999. On the economic consequences of civil war. Oxford Economic Papers 15, 168–183.
Coward, H., Smith, G., 2004. Religion and Peacebuilding. SUNY Press, New York.
Crooke, E., 2005. Dealing with the past: museums and heritage in Northern Ireland and Cape Town. International Journal of Heritage Studies 11, 131–142.
Darby, J., 2001. The Effects of Violence on Peace Processes. United States Institute of Peace Press, Washington.
Darby, J., MacGinty, R., 2000. The Management of Peace Processes. Macmillan, London.
Delanty, G., 2009. The Cosmopolitan Imagination. Cambridge University Press, Cambridge.
Digeser, P., 2001. Political Forgiveness. Cornell University Press, Ithaca.
Du Toit, P., 2001. South Africa's Brittle Peace. Palgrave, Basingstoke.
Elster, J., 2004. Closing the Books. Cambridge University Press, Cambridge.
Giulianotti, R., 2011. Sport, peacemaking and conflict resolution. Ethnic and Racial Studies 34, 207–228.
Griswold, C., 2007. Forgiveness: A Philosophical Exploration. Cambridge University Press, Cambridge.
Halpern, J., Weinstein, H., 2004. Rehumanizing the other: empathy and reconciliation. Human Rights Quarterly 26, 561–583.
Hamber, B., 2009. Transforming Societies after Political Violence. Springer, New York.
Hayner, P., 2002. Unspeakable Truths. Routledge, London.
Herbert, D., 2003. Religion and Civil Society. Ashgate, Aldershot.
Hoglund, K., Sandberg, R., 2008. Reconciliation through sports: the case of South Africa. Third World Quarterly 29, 805–818.
Honwana, A., 2006. Child Soldiers in Africa. University of Pennsylvania Press, Philadelphia.
Hynes, H.P., 2004. On the battlefield of women's bodies. Women Studies International Forum 27, 431–445.
Jankélévitch, V., 2005. Forgiveness. University of Chicago Press, Chicago.
Kaldor, M., 1999. New and Old Wars. Polity Press, Cambridge.
Kaldor, M., 2003. Global Civil Society. Polity Press, Cambridge.
Keen, D., 1998. The Economic Functions of Violence in Civil Wars. Oxford University Press, Oxford.
Lederach, J.P., 2005. The Moral Imagination. Oxford University Press, Oxford.
Lynch, J., McGoldrick, A., 2005. Peace Journalism. Hawthorne Press, Sydney.
Malesevic, S., 2010. The Sociology of War and Violence. Cambridge University Press, Cambridge.
Mallinder, L., McEvoy, K., 2011. Rethinking amnesties: atrocity, accountability and impunity in post-conflict societies. Contemporary Social Science 6, 107–128.
Margalit, A., 2009. On Compromise and Rotten Compromises. Princeton University Press, Princeton.
McEvoy, J., 2011. Managing culture in post-conflict societies. Contemporary Social Science 6, 55–72.
Misztal, B., 2003. Theories of Social Remembering. Open University Press, Buckingham.
Misztal, B., 2011a. Forgiveness and the construction of new conditions for a common life. Contemporary Social Science 6, 39–54.
Misztal, B., 2011b. The Challenges of Vulnerability. Palgrave, Basingstoke.
Norris, P., Inglehart, R., 2009. Cosmopolitan Communications. Cambridge University Press, Cambridge.
Orjuela, C., 2008. The Identity Politics of Peacebuilding. Sage, London.
Paris, R., Sisk, T., 2009. The Dilemmas of Statebuilding. Routledge, Abingdon.
Pichler, F., 2012. Cosmopolitanism in a global perspective. International Sociology 27, 21–50.
Ricoeur, P., 2004. Memory, History, Forgetting. University of Chicago Press, Chicago.
Rotberg, R., Thompson, D., 2000. Truth versus Justice: The Morality of Truth Commissions. Princeton University Press, Princeton.
Rubin, B.R., 2006. Peace building and state-building in Afghanistan. Third World Quarterly 27, 175–185.
Sayer, A., 2000. Moral economy and political economy. Studies in Political Economy 61, 79–104.
Schlack, A., 2009. The Role of Religion in Peacebuilding and Conflict Transformation. VDM Verlag, Saarbrucken.
Shank, M., Schirch, L., 2008. Strategic arts-based peacebuilding. Peace and Change 33, 217–242.
Shaw, M., 2007. What Is Genocide? Polity Press, Cambridge.
Shearing, C., Kempa, M., 2004. A museum of hope: a story of Robben Island. Annals of the American Academy of Political and Social Sciences 592, 62–78.
Shore, M., 2009. Religion and Conflict Resolution. Ashgate, Farnham.
Shriver, D., 1995. An Ethic for Enemies: Forgiveness and Politics. Oxford University Press, Oxford.
Stedman, S., 1997. Spoiler problems in peace processes. International Security 22, 5–53.
Sugden, J., 2008. Anyone for football for peace? Soccer and Society 9, 405–415.
Torrance, A., 2006. The Theological Grounds for Advocating Forgiveness in the Sociopolitical Realm. Centre for Contemporary Christianity, Belfast.
Turner, B., 2006. Vulnerability and Human Rights. Pennsylvania State University Press, Philadelphia.
Van Leeuwen, M., 2009. Partners in Peace: Discourses and Practices of Civil-Society Peacebuilding. Ashgate, Farnham.
Volf, M., 2006. The End of Memory. Eerdmans, Grand Rapids.
Walker, M., 2006. Moral Repair: Reconstructing Moral Relations after Wrongdoing. Cambridge University Press, Cambridge.
Wilkinson, I., 2005. Suffering. Polity Press, Cambridge.
Wilson, R., 2001. The Politics of Truth and Reconciliation in South Africa. Cambridge University Press, Cambridge.
Wolterstorff, N., 2010. Justice: Rights and Wrongs. Princeton University Press, Princeton.

Peace Promotion, Psychology of

Michael G Wessells, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Psychological processes such as enemy images, ideology, and obedience to authority mix with political, economic, and social factors to fuel destructive conflict. Following armed conflict, psychological processes can help to heal the emotional wounds, promote nonviolent conflict resolution, and improve intergroup relations. Psychologists promote social justice and prevent violence through work on transitional justice, policy advocacy, and peace education.

Introduction

As far back as William James, who argued in 1906 that societies need to develop morally appropriate yet exciting alternatives to war, psychologists have been active in promoting peace, defined broadly to include both negative and positive peace. Negative peace includes efforts to reduce or contain violence, whereas positive peace includes efforts to reduce social injustice and the structural violence caused by inequalities and denial of human rights (Christie et al., 2008). Attention to structural violence and positive peace is often achieved through work not only on militarism but also racism, sexism, classism, and corporate greed. Taking a multidisciplinary perspective, this article examines the psychological origins of war and destructive intergroup conflict, psychological approaches to postconflict peacebuilding, and psychologically informed work on building social justice.

Causes of War and Destructive Intergroup Conflict

Armed conflicts, including the intrasocietal conflicts that are now the dominant form of war, have complex political, historic, economic, religious, and cultural roots. Much destructive conflict owes to a divergence of objective interests such as competition over scarce resources such as land, oil, diamonds, and water, and worsening life conditions. The multidecade series of armed conflicts in Afghanistan, for example, have been in part struggles over land and contested political power, including the power sought by the Taliban to impose an extremist form of Islam on the country. Contestation over objective interests has also animated conflicts in Iraq, Palestine, Sudan, Colombia, Democratic Republic of Congo, Sudan, Sri Lanka, and Rwanda, among others.

Enemy Images and Cognitive Biases

In addition to these objective sources of armed conflict, there are a host of subjective, psychological sources that interact with the objective sources and that need to be analyzed and addressed in building peace. For example, as conflict over resources escalates, negative psychological dynamics stimulate spirals of hostile perception and behavior, and become self-perpetuating parts of the conflict process. The parties to a conflict often experience mutual fears, which may include the existential fear that the other group seeks to annihilate or exterminate one's own group. Accompanying and inflaming these fears are enemy images, which are demonic stereotypes of the Other that develop alongside of a heightened sense of the 'Good Us' (White, 1984). The dehumanized nature of these images, which may be expressed in popular media, political rhetoric (e.g., the 'axis of evil'), and political cartoons encourages moral exclusion, the psychological placement of the enemy outside the moral universe.

These images may also bias perceptions by leading adversaries to make negative attributions or inferences about the Other's behavior or motives. For example, a group that already fears an impending attack may view troop movement by the adversary as offensively motivated when in fact the adversary had moved the troops for defensive purposes. Harboring such images and fears and having biased perceptions, adversaries may view violence as both justified and necessary. Indeed, political leaders frequently manipulate these images to stir up war fever and prepare the public for genocide and other mass violations of human rights such as the use of torture against terrorists and the creation of black sites that operate outside the dictates of international law.

Subjective influences are visible also in the biased decision-making of policy elites. Many leaders' aversion to losses encourages excessive risk-taking and departure from rational decision making. In intense conflicts, leaders may experience powerful fears, which motivate misperceptions regarding their adversaries' motives, strength, and willingness to fight. As international tensions escalate, leaders may experience a cognitive constriction evidenced in reduced complexity of views of the adversary expressed in speeches. Leaders may also use inappropriate historical analogies to guide thinking about current crises. Under pressure of crises, small leadership groups having a strong *espirit de corps* and a charismatic leader may make flawed decisions guided by premature consensus, a sense of invulnerability, failure to weigh moral concerns, internal suppression of dissent, and poor contingency planning (Janis, 1982).

In a variety of contexts, obedience to authority is a powerful enabler of armed conflict and of mass violations of human rights. Obedience is particularly strong when the person giving the order to fight is seen as the legitimate leader of the country and as facing a diabolical enemy who cannot be dealt with through peaceful means. Obedience, however, seldom acts alone but combines with diverse factors such as progressive

 http://dx.doi.org/10.1016/B978-0-08-097086-8.22013-8

exposure to violence, the failure to meet basic needs, the lack of bystanders who intervene to prevent violations, and ideologies of good and evil, among others (Staub, 2011).

Social Identity, Societal Narratives, and Ideology

Armed conflict also has subjective sources such as social identity, societal narratives, and ideology. Even when no objective divergence of interest exists, people tend to categorize themselves as belonging to different groups and to show in-group favoritism in judging people, thereby attaining self-esteem and positive self-identity from identification with the in-group (Tajfel, 1982). Typically, they judge others relative to the standards set by the in-group, leading to ethnocentric attitudes. In normal circumstances, such attitudes may lead to provincial thinking but not to violence.

Although identity differences do not themselves cause violence, they may become part of a system of violence in settings of intense fear, mistrust, and real or perceived wrongdoings by the other group. As seen in the Northern Ireland conflict, people who have been attacked by another group may see their identity group – Catholics or Protestants, respectively – as under assault and may experience existential fear as a result. As members of a group are attacked, members of the group create narratives and communal memories about the trauma and damage that the other group has inflicted on them. They form 'chosen traumas' (Volkan, 1997) that they integrate into their social identity and transmit to their families and communities through a mixture of stories, memorials, and other means. Over time, the attitudes and perceptions of each group harden, and identity issues become central to the conflict. These dynamics have been visible in pregenocide Rwanda and also in intractable conflicts such as the Israeli-Palestinain conflict and the Cyprus conflict.

Views of the out-group as threatening and implacably hostile may become sufficiently widespread that they form a social ethos (Bar-Tal, 2007) or informal ideology. Such ideologies are self-sustaining sources of the conflict and are highly resistant to change. The power of ideology, however, becomes most apparent when it is constructed and legitimated by a State, which uses its control over the media and ability to stir the fires or nationalism and whip up public hysteria to instigate mass violence, including genocide. For example, before the Rwandan genocide, President Habyarimana, a Hutu, used mass media to create an ideology of hatred toward Tutsis. Similarly, Hitler used an ideology of racial superiority and global Jewish conspiracy to fuel antisemitism and enable the Holocaust. As these and other examples suggest, key tasks of the psychology of promoting peace are to discourage such ideologies of hatred, racial superiority and violence, and nurture instead peaceful social identities and ideologies of nonviolence.

Postconflict Healing and Peacebuilding

In the aftermath of war, two key psychological tasks are to heal the psychosocial wounds of war and to support peacebuilding processes such as nonviolent conflict resolution and transforming systems of structural violence and oppression. Conceptually, there is a reciprocal relation between psychosocial support and peacebuilding. A significant impediment to peace following armed conflict is the ongoing fear, trauma, and distress that people experience and weave into their social identities. Healing the trauma and distress of war makes it easier to build positive relations between previously hostile groups and to build social and political unity. Conversely, the improvement of relations between groups helps to reduce the intense fear and hostility that stress people and impede healing. For ease of exposition, these are discussed separately below.

Mental Health and Psychosocial Support

Armed conflict has diverse and potentially profound effects on people's mental health and their psychosocial well-being. Mass exposure to attack, violence, and death typically increases above their baseline rate the levels of mental disorders such as depression, anxiety, and posttraumatic stress disorder (PTSD). PTSD, which frequently receives extensive attention following armed conflict, is characterized by reexperiencing horrible events, emotional numbing, and hypervigilance. However, it cannot be assumed that PTSD is the most significant psychological issue or that everyone has it. Consensus practice guidelines indicate that the effects should be identified through a systematic assessment of diverse clinical issues and the use of contextually appropriate measures (IASC, 2007).

In addition, many psychological effects of war are not individual but relate to the damage of relationships and to social divisions and destruction. These interacting psychological and psychosocial effects – referred to collectively as the mental health and psychosocial effects of war – may not reach clinical intensity yet cause significant suffering and distress. Prominent among the psychosocial effects of war are loss of loved ones, family separation, destruction of one's home and community, inability to meet the basic needs of one's family, unemployment and lack of a livelihood, and fear of further violence. A significant source of distress is the sexual violence that occurs frequently not only during war but in communities and in refugee and displacement settings following armed conflict.

The psychosocial effects of armed conflict vary by gender, age, and culture. For example, the threat of sexual violence is often higher for girls than for boys. Similarly, the lack of education and employment may be a larger problem for teenagers than for children who are 7–12 years of age. People's culturally and socially constructed beliefs and practices also influence the psychosocial effects of armed conflict (Wessells, 2006). For example, young people who had been recruited into armed forces or groups and who had killed people may believe they are haunted by the angry spirits of the people they had killed. They and their family and community members may see this as a collective affliction if the local beliefs are that angry spirits wield enormous power and can cause crop failure, deaths, and even war. Such culturally situated beliefs are not static but continually undergo change as new ideas and practices enter the social system. Like mental health effects, the psychosocial effects of armed conflict vary according to the context and should be identified through systematic assessment (IASC, 2007).

The field of mental health and psychosocial support in humanitarian crises has a weak base of evidence regarding which interventions are most effective and sustainable. Evidence from diverse settings suggests that key steps are to support survivors' sense of self-esteem and self-efficacy, particularly in a context of building positive social relations and community mobilization. Also of high importance are cultural supports such as the conduct of burial rites for the dead and the conduct of spiritual cleansing rituals to appease or reduce the influence of the angry spirits discussed above.

Practitioner expertise indicates the need for multiple levels and kinds of mental health and psychosocial support, which include a mixture of naturally occurring supports and those organized by psychologists, governments, or humanitarian agencies. These levels can be envisioned as an intervention pyramid. The base of the pyramid, which applies to the greatest number of people, consists of meeting basic needs such as those for food, shelter, water and sanitation, and health care. When these needs have been met, people benefit from natural supports such as those provided by family, friends, and social interaction with peers, women's groups, youth groups, and religious groups, among others. The second level – family and community supports – includes steps to reunite children separated from their parents and to reestablish normal processes such as education, which is a significant source of psychosocial support for war-affected children. The third level consists of focused interventions for people such as survivors of rape and gender-based violence who may face specific issues such as stigma and need legal and health support. The top level, which affects a smaller number of people yet is of considerable importance, consists of specialized interventions such as therapeutic counseling for the most severely affected people, who may be overwhelmed or unable to function in everyday tasks such as working or, for children, going to school. In war zones, all four of these levels of support are needed, yet it has proven difficult in practice to establish and coordinate them or to develop the referral mechanisms needed in such a systemic approach.

Postconflict Peacebuilding

Peacebuilding includes a wide spectrum of steps, processes, and means to limit and prevent violence and social injustice and to address destructive conflict in a nonviolent manner. During and following armed conflict, the central focus of peacebuilding is typically on nonviolent conflict management and resolution and reconciliation, which aims to improve relations between people and groups who had formerly been locked in destructive conflict. Conflict can be constructive when it is managed in ways that promote communication, healthy relationships, and a sense of positive interdependence (Deutsch et al., 2006). To handle destructive conflict nonviolently, psychologists have developed or refined numerous tools designed to resolve conflict and to transform or repair damaged relationships.

Negotiation, the handling of conflict through communication and bargaining by the parties themselves, has important psychological dimensions. If, as often occurs in official diplomacy between hardened adversaries, opponents view negotiation as a win–lose affair, they may dig into entrenched positions and use damaging, coercive tactics such as threats. Psychologically, a more useful alternative is principled negotiation, which reframes negotiation as a win–win affair. In this approach, parties collaborate to solve their problems and seek not to beat the opponent but to enable all parties to meet their underlying needs, thereby contributing to improved communication, empathy, and relationship. Negotiators vary considerably in which approaches they take and in when and how they express and use emotion. Which approaches to negotiation are likely to succeed depend not only on the situation, the kind of issues at stake, the amount of fear and relational damage but also on culture. For example, US negotiators favor direct communication and emphasize what is said, while Asian negotiators often communicate indirectly, relying on context and maintaining group harmony by avoiding direct confrontation.

Strong mistrust, fear of appearing weak, and poor communication can make negotiation infeasible. In this situation, a preferred approach is mediation, in which a third party helps the parties to negotiate a settlement. By proposing alternatives that the parties may not be willing to suggest themselves, mediators enable face-saving. They may also use social influence processes such as 'carrots and sticks,' a mixture of promised rewards for conflict reducing steps and promised punishments for intransigent behavior. Skilled mediators manage the conflict process, separating the parties when they are likely to say or do damaging things.

Official diplomacy, however, is limited since hardened adversaries may be unwilling to meet publicly. Also, public agreements can evoke backlashes, and official treaties may not change the polarization, fear, and hostility present in divided societies and communities. Herbert Kelman (2005) has pioneered the use of interactive problem-solving workshops as a tool of unofficial diplomacy and relational improvement. Regarding the Israeli–Palestinian conflict, a typical workshop brings together three to six Israelis with an equal number of Palestinians for 3 days of private dialog facilitated by social scientists. The participants are people of rising influence in their respective societies who may subsequently move into official positions where they can influence policy directly. Aside from influencing rising leaders, however, the workshops have a dissemination function in which the participants share their learning with others in their communities, thereby influencing civil society and attitudes.

In analytic, problem-solving discussion, the participants examine the main issues at stake in the conflict, explore their concerns and fears, and identify possible steps and solutions that could reduce the political and psychological barriers on both sides. Empathy, problem-solving, and group exploration of 'what-if' questions are encouraged. By conducting the workshops in private and in academic settings where norms of truth-telling and political nonalignment are prominent, the workshops permit reflection outside participants' ordinary contexts, which are highly politicized, and give them freedom to think 'off the record' and to build productive relationships with participants on the other side. Afterward, participants take their new learning about the other side back into their communities, beginning the wider process of community transformation. By building positive relationships and infusing

useful ideas into conflict-torn communities, this method provides a useful complement to official negotiation.

Positive contact between members of polarized, isolated groups can also reduce destructive conflict (Pettigrew and Tropp, 2011). Positive contact is likely to occur in a context of equal status that enables individuals to get to know each other personally and with steps having been taken to gain leaders' support and prevent damaging behavior. To reduce perceptions that the positive individuals from the out-group are exceptions, one may use cognitive strategies that enable generalization. One useful strategy is to build superordinate group identities, as witnessed in postapartheid South Africa when blacks and whites forged larger identities as 'South Africans' and championed the national rugby team.

The most powerful form of cross-group contact involves cooperation toward the accomplishment of shared goals that neither group could achieve on its own. This approach is called the superordinate goals strategy. Following the brutal, decade-long war in Sierra Leone that ended in 2002, this strategy was used to enable the community reintegration of young men who had been part of the rebel group that had attacked and looted villages. Groups of ex-combatants and other vulnerable youth worked together to build schools, health posts, and other community-selected projects, learning valuable economic and life skills in the process. Their cooperation reduced their enmity, built positive relations with community members, and enabled the social acceptance of the former fighters (Wessells, 2006). Global problems such as nuclear proliferation, environmental destruction, and the HIV/AIDS pandemic provide many possibilities for cooperation on shared goals that mitigates and prevents destructive conflict.

Work on postconflict peacebuilding has recently evolved away from a concern with the management or resolution of specific conflicts toward conflict transformation (Lederach, 1995), which seeks to change the relationships and societal beliefs, norms, and structures within which particular conflicts arise. The emphasis on conflict transformation is appropriate following armed conflict since the goal is not merely to stop the fighting and return the society to the prewar condition, which contained the seeds of conflict. Conflict transformation recognizes that particular conflicts are often manifestations of underlying causes of conflict that, left unaddressed, will likely spawn new conflicts. An important underlying source of destructive conflict is social injustice.

Social Justice and Prevention

Psychologists' promotion of peace frequently entails work on diverse aspects of justice. Distributive justice has to do with social equity and the equal distribution of and access to resources such as money, health care, education, voting privileges, and political power. Restorative justice entails making amends for social transgressions or repairing the harm caused by objectionable behavior such as criminal actions and human rights violations. Restorative justice contrasts with retributive justice, which seeks retribution for harms done, typically through the denunciation of crimes and the imposition of stiff penalties. This section emphasizes distributive and restorative justice since retributive justice frequently involves an 'eye for an eye' approach that sits uneasily with values of peace. Also, retributive justice often models the violence that peace promoters seek to prevent.

Transitional Justice

Although emergencies such as armed conflicts typically take a heavy toll in individuals and societies, they may also present opportunities for constructive social change. To promote justice and peace following mass atrocities and injustice, many societies have created transitional justice processes that may include truth processes and means of establishing justice (Hamber, 2009). Transitional justice processes have been constructed in diverse postconflict societies such as South Africa, Rwanda, Sierra Leone, Liberia, Cambodia, Philippines, Chile, and El Salvador, among others. Although these processes vary significantly across contexts, they seek in general to uncover the truth about the past and to end the impunity that encourages perpetrators of human rights violations. Psychologically, truth telling aims to bring hidden violations against people to light, give survivors a voice, and remind them that they are not alone, and, enable the families of those who had been detained, tortured, attacked, raped or killed to achieve closure by learning what happened to their loved ones. Justice mechanisms vary yet often enable public hearings that put perpetrators on the stand, invite public apology and expressions of remorse for wrongs that had been done, and provide reparations for people who had suffered or been harmed.

A useful example of both the promise and the complexity of such processes is the Truth and Reconciliation Commission (TRC) of South Africa. The postapartheid government of South Africa initiated the TRC as a means of dealing with the previously hidden pain and injustice inflicted by the state terrorism of the apartheid regime, correcting falsifications of history, and enabling forward movement toward social equality. Chaired by Archbishop Desmond Tutu, the TRC worked through three committees: the Human Rights Violation Committee, the Reparation and Rehabilitation Committee, and the Amnesty Committee. With hearings conducted in diverse places, thousands of victims of rights abuses that occurred between 1960 and 1994 told their stories and in some cases confronted the perpetrators, who were granted amnesty in exchange for telling the full truth. The plan was to provide reparations for the survivors of injustices as a means of enabling their rehabilitation.

The evidence regarding the effectiveness of the TRC has been mixed. Supporters have noted that the process helped to uncover the injustices and wounds of the past and to enable national healing and movement toward a new South Africa in which racism was unacceptable. Also, as survivors told their harrowing stories and cried, they were supported and affirmed by the members of the TRC. Some survivors who testified said that they felt better having told their story, brought the violation of their right to light, and publicly confronted the perpetrator. The public nature of the hearings, which were televised, was seen as beginning a process of healing and reconciliation that would enable equal participation of people of diverse ethnic backgrounds in the new South Africa. Through the work of the Reparation and Rehabilitation Committee, which was headed by a psychologist, the TRC organized a mix of communal and individual reparations, thereby reducing problems of favoritism, stigmatization of survivors, and social

divisions. Internationally, the TRC was instrumental in legitimating the new order and transforming the images of South Africa, which had been denigrated during the apartheid era as a highly racist, unethical society that was best dealt with through severe sanctions.

The effectiveness of the TRC, however, was limited in numerous respects. Some of the survivors who testified reported that the perpetrators may have told the truth but showed no remorse, without which amnesty seemed inappropriate. Doubts were also expressed about the accuracy of some of the perpetrators' testimonies, particularly since amnesty was not granted for the most severe offenses and it was in perpetrators' interest to not disclose their most egregious crimes. Although the TRC had granted amnesty sparingly to perpetrators on both sides, the South African military stood outside the process, despite evidence that the military had committed significant violations. In addition, some survivors who testified reported that they felt better having done so, whereas others said they did not or had felt used. Large numbers of survivors were unable to testify through the formal TRC process and participated instead in nonformal community processes organized through groups such as Kulumani. Many survivors who had testified in the TRC were frustrated and upset that although they had told their stories, they and their families continued to live in conditions of extreme poverty and deprivation. The inability to provide extensive economic reparations on a wide scale also limited the effectiveness of the TRC.

Despite these limits and complexities, processes such as the TRC are best viewed as springboards for additional work on social reconstruction for peace. In many societies that have engaged in transitional justice processes, the gains have come less from the specific outcomes of the formal process than from the societal reflection, dialog, and transformation that had accompanied and followed it.

International Criminal Court

The International Criminal Court (ICC) was established in 2002 as a permanent tribunal under international law that prosecutes perpetrators of genocide, crimes against humanity, and war crimes. In contrast to the restorative justice approach of many transitional justice processes, the ICC takes a retributional approach that not only punishes individuals for their crimes but also seeks to build norms of justice by ending the impunity that is frequently associated with human rights violations. The ICC engages with forensic psychologists on a regular basis. Also, clinical psychologists have helped to provide emotional support for survivors who have taken the stand, where they may be forced to relive very painful experiences and where their testimony may place themselves or their family members at risk.

In the case against Thomas Lubanga, the former commander of an armed group in the Democratic Republic of Congo, the ICC made in 2012 its first conviction of a person on charges of having recruited children under the age of 15 years as soldiers. Psychological expert testimony supported the case against Lubanga, arguing that children's participation in an armed group may have powerful psychological effects.

There is debate, however, about whether this type of conviction and work supports justice and peace. Proponents point out that the case draws a line against child recruitment and may deter potential child recruiters in the future. Also, they note that children's exploitation as soldiers is globally one of the enablers of armed conflict. Critics, however, have argued that the ICC has focused excessively on child recruiters in Africa, whereas child recruitment occurs in many different regions. Also, some analysts have expressed concerns that the retributive justice meted out by the ICC sets a bad example and in some contexts, undermines the peace that might have been achieved through the pursuit of less formal mechanisms. For example, the ICC had also brought charges against Joseph Kony in northern Uganda for crimes against humanity committed via his so called Lord's Resistance Army. Significant sentiment in northern Uganda held that such legal prosecution by outsiders made it difficult to pursue the deeper, indigenous peace process that was needed in northern Uganda, where the conflict had taken place. It remains for future research to determine the long-term effects of the ICC process, which are influenced by multiple rulings, and the most roles and the effective balance of formal and informal processes in contributing to peace.

Advocacy for Social Justice and Human Rights

In many countries, psychologists have been active in advocating policies aimed at ending social injustice in the form of 'isms' such as racism, sexism, and militarism. For example, in the US, racism was deeply entrenched in public life long after slavery had ended. In the South, blacks experienced systematic discrimination through Jim Crow laws that were backed by the ongoing threats of violence. Discrimination was prominent in public education, which was segregated into white and black schools, respectively, under guidance of an 1896 Supreme Court ruling (*Plessy vs Ferguson*) that accepted the idea of 'separate but equal.' In opposition, distinguished psychologists such as Kenneth Clark gave expert legal testimony that segregated schools imparted a sense of inferiority to black children. This testimony is widely seen as having helped to motivate the Supreme Court ruling in 1954 (*Brown vs Board of Education*) that prohibited separate schools and mandated the desegregation of public education. It is noteworthy that this policy has not been backed by consistent social mobilization in favor of integrated schools. As a result, the implementation of the policy has faced numerous obstacles and has been uneven in its effectiveness.

Psychologists in diverse countries have also been active in addressing policy and ethics issues associated with torture and related human rights violations. Following the 9/11 terrorist attacks, for example, the US government enacted policies of rendition, detention without basic rights such as *habeas corpus*, and the use of harsh methods such as waterboarding, stress positions, and sleep deprivation in hopes of extracting intelligence information from people who were being detained under charges of being suspected terrorists. Reports of psychologists having advised on the interrogation of some detainees provoked much discussion and moral reflection in the profession of psychology. Organizations such as Psychologists for Social Responsibility contributed to efforts to change the ethics codes and policies of psychology organizations in ways that would prohibit psychologists from working in sites that operated outside of international law and from advising on or participating in the use of coercive interrogation methods that the International Committee of the Red Cross has

indicated are tantamount to torture. To prevent the misuse of psychology, psychologists worldwide are active in promoting international human rights standards as the basis for codes of professional ethics.

Peace Education

Social psychological analyses indicate that in different countries social learning processes teach values and behaviors that support violence in a mixture of intentional and unintentional ways. For example, parents who spank their children for disobedient behavior may not only suppress that behavior but also teach their children that physical punishment and violence is acceptable, even appropriate. At school, children may observe and imitate the violence seen in arguments, fights, and bullying, and boys in particular may be rewarded for behaving in ways regarded as masculine or tough. Through violent video games, films, and electronic media, young people may learn violent means of responding to conflict. Through war memorials and nationalistic media, children may learn to fear and hate national enemies and to see violence conducted by the military as honorable and necessary.

Peace education is a means of socializing children for peace and of preventing violence and social injustice. Peace education may occur formally through activities in schools, yet also occurs informally at home, among peers, in the community, and in the wider society. Broadly, peace education enables learning that supports values and behavior that promote nonviolence and social justice. Peace education nurtures empathy, compassion, and a sense of social interdependence; builds specific skills of handling conflict through nonviolent means; and promotes social justice.

Psychologists contribute to peace education by being positive role models, using methods of constructive controversy to enable critical thinking, stimulating reflection and learning that undermines stereotypes of other groups, and analyzing effective ways of teaching and learning the values and skills that support nonviolence. Extensive psychological research has focused on cooperative learning, in which students collaborate in school to solve problems or complete assignments. Compared to competitive efforts by individual students, cooperative learning typically produces higher levels of productivity and achievement, increased cohesion and positive relationships among students, and increased self-esteem, psychological well-being, and ability to cope with adversity (Johnson et al., 2008). Similarly, psychologists have helped to develop means of measuring the psychological impact of various peace education activities, programs, and curriculi.

Viewed in this manner, peace education is both transformational and systemic. It is transformational because it seeks to prepare people to promote peace and social justice rather than to reproduce socially constructed patterns of violence and injustice. It is systemic since it seeks to promote ongoing learning for peace at levels that engage individuals, families, communities, mass media, and entire societies. Through peace education and the other activities discussed in this article, psychologists seek to do their part to transform current systems of violence into systems of peace.

See also: Child Soldiers as Victims and Actors: Lessons from Ethiopia; Child Soldiers; Conflict and Conflict Resolution, Social Psychology of; Ethnic Identity, Psychology of; Group Processes, Social Psychology of; Human Rights and Social Work; Mass Trauma: Psychopathological Effects across the Life Span; Mediation, Arbitration, and Alternative Dispute Resolution (ADR); Mental Health and Social Work; Peace; Social Identity in Sociology.

Bibliography

Bar-Tal, D., 2007. Sociopsychological foundations of intractable conflicts. American Behavioral Scientist 50, 1430–1453.
Christie, D., Tint, B., Wagner, R.V., Winter, D.D., 2008. Peace psychology for a peaceful world. American Psychologist 63 (6), 540–552.
Deutsch, M., Coleman, P., Marcus, E. (Eds.), 2006. Handbook of Conflict Resolution, second ed. Jossey-Bass, San Francisco, CA.
Hamber, B., 2009. Transforming Societies after Political Violence: Truth, Reconciliation, and Mental Health. Springer, New York.
Inter-Agency Standing Committee, 2007. IASC Guidelines on Mental Health and Psychosocial Support in Emergency Settings. IASC, Geneva.
Janis, I., 1982. Groupthink. Houghton Mifflin, Boston, MA.
Johnson, E., Johnson, R., Holubed, E. (Eds.), 2008. Cooperation in the Classroom, seventh ed. Interaction Book Company, Edina, MN.
Kelman, H.C., 2005. Interactive problem solving in the Israeli-Palestinian case: past contributions and present challenges. In: Fisher, R.J. (Ed.), Paving the Way: Contributions of Interactive Conflict Resolution to Peacemaking. Lexington Books, Lanham, MD, pp. 41–63.
Lederach, J.P., 1995. Preparing for Peace: Conflict Transformation across Cultures. Syracuse University Press, Syracuse.
Pettigrew, T.F., Tropp, L.R., 2011. When Groups Meet: The Dynamics of Intergroup Contact. Psychology Press, New York.
Staub, E., 2011. Overcoming Evil: Genocide, Violent Conflict, and Terrorism. Oxford University Press, New York.
Tajfel, H., 1982. Social Identity and Intergroup Relations. Cambridge University, New York.
Volkan, V., 1997. Bloodlines. Farrar, Straus and Giroux, New York.
Wessells, M.G., 2006. Child Soldiers: From Violence to Protection. Harvard University Press, Cambridge.
White, R.K., 1984. Fearful Warriors. Free Press, New York.

Pearson, Karl (1857–1936)

John Aldrich, University of Southampton, Southampton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Karl Pearson had a successful career as a physical scientist before turning to the biological sciences. He published comparatively little on the social and behavioral sciences but his influence was considerable for he was the first modern statistician. Although some of the aspirations and even some of the methods of modern statistics can be traced back centuries, the modern discipline took off at the end of the nineteenth century. Karl Pearson contributed to the take-off by identifying problems amenable to statistical treatment, by developing appropriate techniques and by laying the foundations of a discipline. Some of his contemporaries did some of these things better, but none combined them so effectively.

Karl Pearson published extensively on the physical and biological sciences and comparatively little on the social and behavioral sciences, but to the latter he contributed the modern role of *statistician*. While some of the aspirations and even some of the methods of modern statistics can be traced to the seventeenth century, it is generally held that the modern discipline took off at the end of the nineteenth century: see Stigler (1986) and Hald (1998). To the take-off Karl Pearson contributed by identifying problems amenable to statistical treatment, by teaching influential practitioners and by creating the infrastructure needed for a discipline including university departments, courses, laboratories, journals, statistical tables; for Pearson's influence on economists see Aldrich (2010) and, for infrastructure, Magnello (1999) and Aldrich (2013). Some of his contemporaries did some of these things better, but none combined them so effectively. There is an annotated bibliography of Pearson's writings by Morant (1939) and there is now an extensive secondary literature surveyed by Aldrich (2011). The best overall account of Karl Pearson's life remains the *Appreciation* by his son E.S. Pearson (1936–38), but there is a rich study of the early 'KP' by Porter (2004).

Life

Karl Pearson was born in London, the son of a successful lawyer, but he was always proud that his ancestors were of Yorkshire 'yeoman stock.' Karl was a delicate child but he thrived at Cambridge University, where he studied mathematics. Success in the examinations brought him a fellowship and for a few years the financial freedom to pursue his very wide interests. Pearson qualified as a barrister and studied social, philosophical, and historical issues. He developed his own view of man's place in a post-Christian world, expressing this in a novel and a play as well as through essays, some of those were collected in the volume *Ethic of Freethought* (1888).

In 1884, the essayist and freelance lecturer became professor of applied mathematics and mechanics at University College London, where he would remain for the rest of his life. In his early days at the College, Pearson took on two books left unfinished when their authors died. W.K. Clifford's *Common Sense of the Exact Sciences* explained the nature of mathematics and for Pearson it would serve as a bridge from his early general essays to his *Grammar of Science* (1892). The *Grammar of Science* presented the scientific method as "the orderly classification of facts followed by the recognition of their relationship and recurring sequences." In later editions, the positivist ideas about space and force were joined by an exposition of statistical ideas. The other book, Isaac Todhunter's *History of the Theory of Elasticity*, was the history of Pearson's own specialism and Pearson actually wrote much more of this vast work than Todhunter himself. Pearson's later historical researches – a biography of Galton and lectures on the history of statistics – would combine Todhunter's thoroughness with his own interest in broad intellectual currents.

In his mid-30s, the elastician with philosophical interests took up biometry and statistics under the influence of his colleague W.F.R. Weldon. Weldon had joined the College in 1890 as professor of zoology and was already applying Francis Galton's statistical methods to the formation of species. Pearson helped Weldon extend the methods, but he was not just a mathematical consultant, he became a contributor to the biological literature in his own right. The partnership with Weldon was central to Pearson's development and he made his most important contributions to statistical thinking in these early years with Weldon. Pearson wrote over 300 pieces after 1900 but they were on lines he had already established. These works demonstrated what statistical methods could achieve and contributed to making statistics a recognized discipline.

In 1901, Pearson, Weldon, and Galton founded *Biometrika*, a 'journal for the statistical study of biological problems,' which published empirical studies, tables, and some statistical theory. From 1903, a grant from the Worshipful Company of Drapers funded the Biometric Laboratory. The laboratory drew visitors from all over the world and from many disciplines. Pearson wrote no textbook or treatise but the visitors passed on what they had learned when they returned home and they could keep up to date by reading *Biometrika*. In 1907, Pearson took over a research unit founded by Galton and reconstituted it as the Eugenics Laboratory. This laboratory researched human pedigrees, but it also produced controversial reports on the role of inherited and environmental factors in tuberculosis, alcoholism, and insanity – great topics of the day. One of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61105-4

Pearson's projects was to create a scientific basis for eugenics; see Kevles (1985). In 1911, a bequest from Galton enabled Pearson finally to give up responsibility for mechanics and applied mathematics and become professor of eugenics and head of the department of applied statistics. During World War I, normal activities virtually ceased but afterward expansion was resumed. However, Pearson was no longer producing important new ideas and Ronald Fisher was supplanting him as the leading figure in statistics. Pearson retired in 1933 but he continued to write and, with his son Egon S. Pearson, to edit *Biometrika*.

Pearson made strong friendships, notably with Weldon and Galton, but he also thrived on controversy. Some of the encounters produced lasting bitterness, those with William Bateson and Ronald Fisher being particularly unforgiving. Following the 'rediscovery' of Mendel's work in 1900, Bateson questioned the point of investigating phenomenological regularities such as Galton's law of ancestral heredity. The biometricians replied that such regularities had been established and that any theory must account for them. The quarrel about the value of biometric research faded with Weldon's death in 1906 and Pearson's gradual withdrawal from research into inheritance; the controversy is reviewed by Provine (1971). Fisher, like Bateson, was a Mendelian, but he agreed with Pearson on the importance of the statistical method. Fisher found fault with the execution of Pearson's statistical work, but he also judged that Pearson abused his commanding position in biometry.

Contributions to Statistics

For Pearson, empirical research and methodological innovation went hand in hand and his empirical work was always serious: he wanted to make a contribution to knowledge in the field to which he was applying his methods as well as showing that they 'worked.' However, his contributions to statistical methodology overshadowed his contribution to any substantive field and his scientific reputation rests upon them. It is convenient to divide his statistical contributions into ways of arranging numerical facts – univariate and multivariate – and methods of inference – estimation and testing.

Univariate Statistics: The Pearson Curves

Pearson's first major statistical paper (1894) was on extracting normal components from 'heterogeneous material.' Pearson was not, however, an admirer of the normal distribution or of the 'theory of errors' for which it provided the basis. For him the usual phenomenon was 'skew variation in homogeneous material' and this called for additional distributions, and Pearson (1895) presented a system of frequency curves, which would be of use in 'many physical, economic, and biological investigations.' Systems were also developed by Edgeworth, Charlier, Kapteyn, and others, but Pearson's was the most widely used. Other systems have been devised, inspired by his, but none has been as central to the statistical thought of its time.

From probabilistic considerations – which now seem contrived – Pearson obtained a basic differential equation governing the probability density function, or frequency curve, $y = y(x)$:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{y(x+a)}{b_0 + b_1 x + b_2 x^2}$$

The constants determine the form of the curve. Pearson developed methods for choosing an appropriate curve and for estimating the constants. From his laboratory came tables to ease the burden of calculation.

While Pearson used these curves as data distributions, he assumed that the estimators of the constants would be approximately normal, certainly in the large samples he generally worked with. In the work of 'Student' (W.S. Gosset of Guinesss) and especially Fisher the data distributions are normal but the exact distributions of the sample statistics are usually not, though *they* may follow a Pearson curve. 'Student' did his distribution work in Pearson's laboratory, but for many years Pearson thought that such 'small sample' work was off the main track and of interest only to 'naughty brewers.'

In the Pearson system the normal distribution is just one type. In Fisher's analysis of variance, which appeared in the 1920s, it is the *only* type. Workers schooled in the Pearson system did the earliest robustness studies and one of the continuing roles of the system has been in this area. While Pearson's skepticism toward normality pervades much modern work, the modern approach to inference is so different from his that he is not a direct inspiration.

Multivariate Statistics: Correlation and Contingency

In 1888, Galton introduced correlation and soon Edgeworth was developing the theory and Weldon was applying the method in his research on crustaceans. Pearson, however, produced the definitive formulation of normal correlation and had the vision that correlation would transform medical, psychological, and sociological research.

Pearson (1896) set out to frame hypotheses about inheritance in terms of the multivariate normal distribution but these were of less enduring interest than his account of correlation and multiple regression. The familiar partial and multiple correlation and regression coefficients appeared at this time either in Pearson's work or that his assistant, G. Udny Yule. Soon, Pearson and those around him were applying correlation in meteorology, sociology, and demography, as well as in biometry.

The 1896 ideas formed the basis of a definitive theory, especially when completed by the exact distribution theory of 'Student' and Fisher. But Pearson soon moved on to qualitative, nonnormal, and time series data, though not with the same success. Pearson's scheme for contingency, the qualitative analogue of correlation, has latent continuous variables – surprisingly jointly normal ones. Yule, no longer working with him, rejected the scheme; the controversy has been studied from a social studies of science perspective by MacKenzie (1981). Pearson toiled unsuccessfully to create a theory of skew surfaces, uniting correlation and skew curves.

Time series correlation analysis was another disputed field. Pearson thought of the true correlation between series as obscured by time trends superimposed upon them. The scheme was widely used, but in the 1920s Yule criticized it effectively without putting anything in its place.

From the beginning Pearson encountered problems with correlation that have lived in the underworld of statistics, occasionally surfacing in fields such as path analysis and structural modeling. Pearson formalized Galton's notion of correlation as measuring the extent to which variables are governed by 'common causes.' He was soon aware of cases of 'spurious correlation,' where the natural interpretation is not sustainable. He and Yule – who developed the other standard interpretation of a direct causal relationship between the correlated variables – found numerous pathological cases. They devised treatments but developed no systematic way of thinking about the problem.

Pearson – following Galton – established multivariate analysis. Yet, by a curious twist the multivariate data for which Pearson struggled to create a theory of frequency surfaces are often treated by Fisher's 'regression analysis,' which was a development from Gauss's (univariate) theory of errors rather than from Pearson's correlation theory.

Estimation: The Method of Moments and Inverse Probability

Pearson usually found the constants of frequency curves by the method of moments. He introduced this in his first major paper (1894) for he could find no other method for estimating a mixture of normals. Pearson went on to apply the method quite generally. It seemed particularly appropriate for the Pearson curves where the constants determine the first four moments. However, the method was put on the defensive after Fisher criticized it for not being 'efficient' – even when applied to the Pearson curves.

Probable errors (scaled standard errors corresponding to the 50 percent point) for method of moments estimators were derived by Pearson and Filon (1898) using a modification of an argument in the 1896 correlation paper. Pearson had obtained the product-moment estimate from the posterior distribution (based on a uniform prior) and its probable error from a large sample normal approximation to the posterior. In 1898, he assumed that the same technique would always yield probable errors for the method of moments. He seems to have dropped the original technique when he realized that the formulae could be wrong but it was reshaped by Fisher into the large sample theory of maximum likelihood.

In 1896–98, Pearson used Bayesian techniques quite mechanically but he could also be a thoughtful Bayesian. Over a period of 30 years he wrote about the 'law of succession,' the standard Bayesian treatment of predicting the number of successes in future Bernoulli trials given the record of past outcomes – 'the fundamental problem of practical statistics' he called it. Pearson held that a prior must be based on experience; a uniform prior *cannot* just be assumed. Pearson pressed these views in estimation only once when he criticized Fisher's maximum likelihood as a Bayesian method based on an unsuitable prior – essentially Pearson's own method of 1896–98. Fisher was offended by this misinterpretation of his Bayes-free likelihood analysis.

There was plenty of improvisation in Pearson's estimation work; it was eclectic, even inconsistent. He was not the consistent Bayesian that Harold Jeffreys would have liked. Pearson used sampling theory probable errors, but he did not follow up the results on the properties of estimators in the old theory of errors. Pearson was not unusual in his generation; the theory of errors he found to hand mixed frequentist and Bayesian arguments. Some of his contemporaries were more acute but foundational clarity and integrity only became commanding values toward the end of his career when Fisher, Jeffreys, and Neyman set about eliminating the *wrong* kind of argument.

Testing: Chi-Squared

Pearson's earliest concerns were with estimation – curve fitting and correlation – though their testing could be done using the estimates and their probable errors and making the assumption of large sample normality. However, his chi-squared paper of 1900 introduced a new technique and gave testing a much bigger role.

The chi-squared test made curve fitting less subjective. One could now say: "In 56 cases out of a hundred such trials, we should on a random selection get more improbable results than we have done. Thus, we may consider the fit remarkably good." The tail area principle was adopted without any examination. Naturally, Pearson, the critic of the theory of errors, enjoyed showing that the fit of the normal distribution to errors in astronomical measurements was *not* good.

Pearson brought chi-squared into his treatment of contingency and by 1916 he had established the main applications of the chi-squared distribution – a test for goodness of fit of distributions, a test for independence in contingency tables, a test for homogeneity of samples, a goodness of fit test for regression equations.

Pearson's distribution theory involved the exact distribution of the exponent in the multinormal density, this density appearing as a large sample approximation to the multinomial. Pearson slipped in the mathematics and used the same approximation whether the constants of the frequency curves are given a priori or estimated from the data. Fisher realized the importance of the difference and reconstructed the theory, finding that the chi-squared approximation was valid but the number of degrees of freedom had to be altered.

Fisher continued Pearson's chi-squared work and gave significance testing an even bigger role in applied statistics. Pearson's methods were more or less vindicated by the later work of Fisher and then of Neyman and E.S. Pearson but, as in estimation, his reasons did not pass critical examination.

Summary

Of the many statistical techniques Pearson devised, only a few remain in use and, though his ideas sometimes find reexpression in more sophisticated form, such as the correlation curve or the generalized method of moments, there is little to

suggest that Pearson continues to directly inspire work in statistics. Pearson broke with the theory of errors but in the next (Fisher) generation the theory was restored through the analysis of variance and regression to stand beside and even overshadow Pearsonian statistics. From a modern perspective, Pearson's inference theory seems desperately superficial. Yet, the problems he posed have retained their importance and the ground he claimed for the discipline of statistics has not been given up. Appropriately, the sesquicentenary of Pearson's birth was marked by both the Royal Statistical Society and the International Statistical Institute – for the latter, see Seneta and Stamhuis (2009). Ironically, Pearson never belonged to either body for he scorned their lack of scientific credentials and it is a mark of his effect on statistics that the modern bodies should recognize *him* as an honored ancestor.

See also: Bayesian Statistics; Eugenics as an International Movement; Eugenics in Europe; Fisher, Ronald A (1890–1962); Galton, Sir Francis (1822–1911); Linear Hypothesis: Regression (Basics).

Bibliography

Aldrich, J., 2010. The econometricians' statisticians 1895–1945. History of Political Economy 42, 111–154.

Aldrich, J., 2011. Karl Pearson: A Reader's Guide. Website: http://www.economics.soton.ac.uk/staff/aldrich/kpreader.htm.

Aldrich, J., 2013. Karl Pearson's biometrika: 1901–36. Biometrika 100, 2–15.

Hald, A., 1998. A History of Mathematical Statistics from 1750 to 1930. Wiley, New York.

Kevles, D.J., 1985. In the Name of Eugenics: Genetics and the Use of Human Heredity. Knopf, New York.

MacKenzie, D.A., 1981. Statistics in Britain 1865–1930: The Social Construction of Scientific Knowledge. Edinburgh University Press, Edinburgh, UK.

Magnello, M.E., 1999. The non-correlation of biometrics and eugenics: rival forms of laboratory work in Karl Pearson's career at University College London, (in two parts). History of Science 37, 79–106, 123–150.

Morant, G.M., 1939. A Bibliography of the Statistical and Other Writings of Karl Pearson. Cambridge University Press, Cambridge, UK.

Pearson, E.S., 1936/38. Karl Pearson: an appreciation of some aspects of his life and work, in two parts. Biometrika 28, 193–257, 29, 161–247.

Pearson, K., 1888. The Ethic of Freethought. Fisher Unwin, London.

Pearson, K., 1892. The Grammar of Science (Further editions in 1900 and 1911). Walter Scott, London (A & C Black, London).

Pearson, K., 1894. Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society A 185, 71–110.

Pearson, K., 1895. Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. Philosophical Transactions of the Royal Society A 186, 343–414.

Pearson, K., 1896. Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. Philosophical Transactions of the Royal Society A 187, 253–318.

Pearson, K., Filon, L.M.G., 1898. Mathematical contributions to the theory of evolution IV.On the probable errors of frequency constants and on the influence of random selection on variation and correlation. Philosophical Transactions of the Royal Society A 191, 229–311.

Pearson, K., Filon, L.M.G., 1900. On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine 50, 157–175.

Pearson, K., Filon, L.M.G., 1914/24/30. The Life, Letters and Labours of Francis Galton, vols. I, II, IIIA, IIIB. Cambridge University Press, Cambridge, UK.

Porter, T.M., 2004. Karl Pearson: The Scientific Life in a Statistical Age. Princeton University Press, Princeton, NJ.

Provine, W.B., 1971. The Origins of Theoretical Population Genetics. Chicago University Press, Chicago.

Seneta, E., Stamhuis, I.H., 2009. Preface to Karl Pearson issue. International Statistical Review 77, 1–2.

Stigler, S.M., 1986. The History of Statistics: The Measurement of Uncertainty before 1900. Belknap Press of Harvard University Press, Cambridge, MA.

Peasantry in the Twenty-First Century

Jan Douwe van der Ploeg, Wageningen University, Wageningen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The deregulation of markets and the increasingly interlinked global markets make it very difficult for small farmers to survive. Thus, small primary agricultural producers must search for new ways to continue farming and earn a reasonable income. In many places, this has triggered a process of repeasantization that is reversing the once seemingly inevitable processes of modernization and urbanization. This article argues that this process occurs through six mechanisms that together have a strong impact on the nature and dynamics of the process of agricultural development.

Both the social and academic conceptions of what a peasant is extend beyond that of someone just involved in agricultural activities. Being a peasant is more than just being a producer of food or other agricultural products. True, peasants are involved in agriculture, but they do so because it offers them a livelihood that provides them with employment, income, an identity, a place to belong to (which is also a relatively safe and clean space for their children to grow up), social networks, and, probably, some level of dignity and the feeling that they are part of a more comprehensive whole. It is difficult to find such a livelihood in the slums that abound in most of the big cities of the world. So for many, peasantry is the only realistic alternative. And when peasants engage in labor migration (often crossing national borders), this form of livelihood remains a critical and essential point of reference. Migrant peasant workers send much of what they earn back home in the form of remittances and cherish the hope and dream that someday they will return, even if only after many years, and invest their savings in their land. For them, their farm represents an important psychological and economic fallback position.

For peasants, their livelihood (that often includes faraway places) offers a starting point for improving their lives, albeit slowly and in a step-by-step way. It is the space in which they struggle for autonomy and improvements (van der Ploeg, 2008; Langthaler, 2012). They do so, mostly because this is the only way they have to respond to the dependency and deprivations they experience. Such struggles involve resistance, sturdiness, doggedness, and distrust of the outside world.

Peasants have specific and characteristic patterns of involvement in agricultural activities that differ, sometimes subtly, sometimes more obviously, from corporate or entrepreneurial ways of farming. Ecological capital and labor are central to peasant farming. Labor is used to convert living nature into valuable products. Peasant farming is labor intensive and is developed through labor-driven (as opposed to capital-driven) intensification. It aims to improve the value added that can be gotten out of the scarce available resources and to reproduce and develop these resources over time. Labor investments (in the further improvement of the resource base) are a powerful mechanism for growth and development in peasant agriculture.

Yet at the same time, peasant livelihoods are also vulnerable and are sometimes nearly suffocated by outside forces. Since peasant farms operate in at least four different markets (for inputs, agricultural produce, labor, and consumption goods), they can run considerable economic risks. In addition, their farming practices may be prone to uncontrollable climatic risks and a range of natural disasters. Finally, their power relations with the surrounding society are often weighted against them and can pose the threat of dispossession (through, for example, land or water grabbing or legal suits about property rights to genetic material).

For decades it has been assumed (and continuously argued) that farming-as-a-livelihood (i.e., peasant farming) is completely at odds with the entrepreneurial attitudes, growth, investments, and rationality of modern-day society. In short, farming-as-a-livelihood was equated with backwardness. And although some persist in making such arguments, it is now widely accepted among scholars of the rural (IAASTD, 2009; HLPE, 2013) that this is not the case. Peasant agriculture has the potential to be highly productive and efficient. Its yields are often considerably higher than those from other forms of agricultural production, and its output-to-input ratios reflect high levels of efficiency. In addition, peasant farming creates much employment and generates considerable incomes, thus contributing to the enlargement of the domestic (and particularly rural) market (which in turn may trigger other forms of economic growth, Haggblade et al., 2007). It can be a safeguard for sustainability and contribute to the emancipation of downtrodden social groups in society. Equally it plays a strong role in safeguarding biodiversity, scenic landscapes, and the attractiveness and accessibility of the countryside as a whole.

These potentials have been realized in many different places. This most often occurs when the political, economic, social, and cultural conditions – or "space" as Halamska (2004) phrased it – are favorable for peasant farming. If such space is lacking, peasant agriculture may well regress and perform poorly, and rural livelihoods become a synonym for poverty and an absolute lack of prospects.

During the last decades, peasant agriculture has faced a severe and long-lasting market squeeze. It has simultaneously suffered from high levels of market volatility, the effects of a decades-long neglect by states and multilateral organizations (as recently admitted by the World Bank; see McMichael, 2009), highly unequal terms of competition, lack of access to credit, and unequal power relations that create general insecurity and favor land and water grabbing. It is becoming

increasingly difficult for peasants to reproduce the farm and farm household solely by producing for the general markets for food and other agricultural products. However, peasants are not passive in the face of these many threats. On the contrary, they have, and are developing, an impressive range of responses to these many threats and finding new ways to (re)enlarge autonomy and (re)construct otherwise wrecked livelihoods. Analytically, we can identify six mechanisms that peasants have developed to reinvent their farms. These will be briefly discussed later. Such mechanisms are being deployed in both the Global South and in the Global North. The article also draws attention to the new modes of agricultural development that result from these reinventions. In this respect, it is important to note that overall agricultural development is always the aggregated outcome of different, and often mutually contrasting, development trajectories that coexist alongside each other. Hence, when peasant trajectories show positive results and large-scale capitalist farming show negative ones, the overall outcome might be zero. Overall tendencies might very well veil the value of peasant modes of agricultural development.

Reducing the Use of External Inputs

Contemporary agriculture can be characterized as having high (and growing) input levels. This is largely due to the introduction of technological models that built on the use of such *external* inputs. Most of these inputs replace internal resources: chemical fertilizers replace manure, green fertilizers, and a carefully nurtured soil biology; industrial concentrates replace feed and fodder grown on or near the farm; self-selected seeds are replaced by bought seeds, produced elsewhere; weeding is replaced by herbicides; etc. This also contributes to a kind of standardization: While internal inputs (i.e., produced and reproduced within the farm) are mostly highly adapted to local ecosystems, local ways of farming and in line with local knowledge, external inputs are far less aligned with local conditions, needs, and possibilities.

The increased use of external inputs (often combined into a 'package,' as was the case with the green revolution in the Global South and the *Holsteinization* of dairy herds in Europe) has undoubtedly helped to raise levels of productivity considerably – at least initially. Yet over time, these packages gave rise to new problems. The cost of acquiring these inputs was increasingly beyond the reach of many farming households. Hence, they needed credit, creating new financial and transaction costs, which continued to spiral. And, as production increased, farm-gate prices for produce started to decline. As a result, the costs of the external inputs utilized came to represent a growing proportion of the farm's total earnings.

This situation has triggered the first mechanism for reinventing peasant farming. Many farmers are returning to their neglected internal resources, trying to upgrade their quality, and simultaneously reduce the use of costly external inputs (Domínguez García, 2007; Paredes, 2010). This has the potential of bringing considerable cost reductions (while levels of production may remain the same or even increase). Beyond that, for many farmers, this change represents a return of the 'joy of farming.' Instead of simply applying the scripts developed by the agricultural sciences and brought to them in the form of external inputs and manuals (that specify how, when, in what sequence, by whom, and in which combinations these inputs should be used), farmers are again trying out things for themselves, experimenting and developing new local knowledge.

A good European example is the rediscovery of manure. This internal input became, during the period of modernization, a waste product that stank and even had the potential to badly pollute the soils and waterways. Thus, increasing amounts of chemical fertilizers became necessary. Now, facing both an environmental and an economic crisis, many farmers are once again rebuilding manure into a highly valuable internal input, to replace chemical fertilizers. This is done through a range of ingenious adaptations, including the dietary regime of the animals, the storage of the manure, the use of straw in the cowsheds, and the way manure is spread on the fields (Reijs, 2007).

More generally, the reduction in the use of external inputs (and the technologies that make them necessary) is strongly linked to the (re)emergence of what European farmers know as the style of farming economically, in which the farmer seeks to minimize monetary expenses (Kinsella et al., 2002). Thus, peasant farms are reemerging and, while their form may be new, their dynamics and rationale are, without doubt, very much peasant-like.

In Central and South America a similar trend is occurring, although the key word here is agroecology (Altieri et al., 2011; Rosset and Martinez-Torres, 2012). Agroecology embodies three strongly interrelated elements: a new scientific view on farming, a new practice (characterized by low levels of external inputs), and a strong social movement that seeks more autonomy for peasants and new ways of farming (Toledo, 2011). Recently it seems that a fourth element is emerging: increased government support for agroecology (MDA, 2013).

In terms of aggregate agricultural development, these processes and practices are making a large contribution to new forms and levels of sustainability. These differ markedly from approaches that involve the imposition of end-of-the-pipeline solutions, the conventional way of addressing environmental problems.

Developing Multifunctionality

A second mechanism used by peasants to consolidate and strengthen their farms is the production of *new* services and *new* products (and the subsequent development of new market channels) that bring new income flows to their farms and households. The magnitude of these new income flows can compensate for, or even exceed, the decrease in income derived from conventional production.

The production of new products and services is best captured with the notion of multifunctionality. This refers to the farm having more than just one function, i.e., producing agricultural products (which, in conventional patterns, are mostly raw materials for agro-industries). The other functions of a farm can include energy production, water retention, the management of nature and landscape, increasing or safeguarding biodiversity, agritourism, afforestation, fish ponds, care facilities, and a range of classical forms of

diversification (such as organizing transport, leasing storage facilities, etc.). These nonagricultural activities (*sensu stricto*) are combined with the typical agricultural activities within the farm. Both types of activities use the same resource base. In addition to creating additional income flows, this also creates synergies that generate considerable additional economic benefits (Saccomandi, 1998), which strengthen overall household income.

The pursuit of multifunctionality may also involve converting conventional agricultural production into new forms that result in an *enlarged* value added (VA) per unit of end product, a shift which also strengthens the overall income at household level. Examples here are organic farming, the production of high-quality products and regional specialties (which, especially in Europe, encompass a wide and increasingly legally protected spectrum), on-farm processing, and direct marketing. Multifunctionality is now strongly developed throughout Europe. Chinese agriculture has always been multifunctional, but a bewildering array of new expressions is currently being constructed (Ye et al., 2010). Brazil has strong pockets of newly developing multifunctionality (Schneider et al., 2010), and the same approach has always been present in the Andean mountains. More recently, forms of multifunctional farming are also being constructed in parts of the USA (Joannides et al., 2001).

The development of multifunctional farms is a clear expression of resistance (Schneider and Niederle, 2010). The desire to remain in farming translates into highly visible forms of reinvention: multifunctionality involves replacing the monotony of standard farming practices by often surprising new rural enterprises. It contributes to the emergence of new modes of agricultural development, by reorienting agriculture to new social needs that originate in the cities, making agriculture more attractive again and, especially through agritourism and peasant markets, (re)linking it to consumers (Low and Vogel, 2011). Multifunctionality also helps to enlarge resilience and to slow down the rural exodus. It is especially important to note that multifunctionality is often used to support investments in food production *sensu stricto*, thus contributing to its resilience (de Rooij et al., 2013).

Embedding Farming in Multiple Job Holding

Another increasingly important mechanism for reinventing peasant farming is the growing engagement in what the Chinese call 'multiple job holding' (MJH). Peasants not only work within their own farm (be it in agriculture or other activities related with the multifunctional nature of the farm), they increasingly also have a job elsewhere (van der Ploeg and Ye, 2010).This might be far away (giving rise to complex movements through time and space) or nearby. In Europe, this phenomenon is also widespread. In 85% of Dutch dairy farms, the man or the wife (or both) has another job that provides them with at least one-third, and sometimes more than a half, of the household income. Thus, farming is increasingly becoming a 'part-time' phenomenon (as it is now often seen in the Netherlands).

For many decades it was thought that part-time farming or MJH is the first step in the process of a farm disappearing. It is becoming clear, though, that it is exactly the other way around. By embedding the farm in a larger portfolio of economic activities, within and beyond the farm, farm continuity and development become feasible. MJH assures a flow of resources (savings, remittances, but also new knowledge, new lifestyles, new urban contacts) from the urban toward the rural and thus helps to sustain farming. It strengthens the resilience of agricultural and rural development and slows down the rural exodus. Sometimes the effects go even further. For example, in China, migrant workers from Yi County started to bring glass noodles (processed sweet potatoes) as gifts for their bosses, friends, and relatives in the cities after every spring festival (which they spend in their villages). This led to the glass noodles becoming a well-known and highly appreciated product that now is central to multifunctional agriculture in several villages in Yi County (Wu et al., 2014).

The Retreat of Entrepreneurial Farming

For several decades entrepreneurial farming has been presented as the promise for the future, while peasant farming was understood as a remnant of the past. Thus, the transition to 'modern agriculture' became central to agricultural policies such as the Common Agricultural Policy in Europe, the Farm Bill in the USA, and many programs for 'integrated rural development' throughout Latin America and Africa. In Asia, the many green revolution programs carried a similar rationale.

One aspect of entrepreneurial farming is that much of its growth basically occurs through the takeover of other farms (their resources, quota, environmental rights, shares in subsidies, market share, etc.), which is financed through the banking system. Entrepreneurial farming makes less use of ecological capital and is highly reliant on the use of external inputs. This means entrepreneurial farming is basically a financial operation. It is about operating with money in order to make more money. With higher financial and transaction costs, the costs of production are increased and the margin (or VA per unit of end product) is relatively low. Because of the sheer size of these farms, the final income might be considerable. But this is only the case when prices remain more or less stable and at the required level.

In the past 15 years or so, deregulation and liberalization have created a new reality characterized by an accentuated volatility in prices. As empirical studies now demonstrate, the sudden price falls have produced a hitherto unexpected consequence, i.e., negative cash flows, which are particularly evident in large entrepreneurial farms. Farm-gate prices have fallen below the costs of production, and the high financial costs and the high levels of external input use (tied to previous choices on technology use) make it almost impossible to react to this dangerous situation. On the other hand, the smaller peasant farms have higher margins and are more flexible; so, in this respect, they are far less vulnerable. Initially, the agricultural banks refinanced the entrepreneurial farms that entered into difficulties, but since they have recently been obliged to increase their own liquidity (under the Basel III Accords), this is no longer the case. Consequently, many places of the world are witnessing

a demise in entrepreneurial farming. It is notable that it is the largest, most modernized, and highly specialized farms that have experienced the greatest difficulties, which have often proved insurmountable (Oostindie et al., 2013). These farms are mainly located in Denmark, the Netherlands, and other parts of northwestern Europe. Similar difficulties are being experienced elsewhere by those who attempt to continue with the green revolution approach. Indebtedness, high levels of environmental pollution, stagnating yields, and the unwillingness of banks to continue their credit lines are but some of the telltale signs that give rise to frustration and disappointment among producers. The entrepreneurial model is 'running out of steam.' Many farmers who still have the 'space' to do so are opting out of this approach and rebuilding their farms along more peasant-like lines. Others are far too entrapped in path dependency. They mostly go bankrupt or accept a very low income (selling the farm often is not a solution since debts often exceed the decreased value of the assets). All this contributes to making the aggregate agricultural sector in many countries more peasant-like, (see for Russia Mamonova et al., 2013) and, at the symbolic level, the entrepreneurial model has lost much of its appeal and legitimacy. Meanwhile, the peasant way of farming is coming to be understood as a strong institution and possibly the best guarantee for food security and food sovereignty (HLPE, 2013).

In terms of agricultural development, the demise of the entrepreneurial model means that agro-industry and the state have lost at least part of their authority to shape agricultural development. At the same time, self-regulation, the development, the use of common pool resources (Ostrom, 1990), and new institutional arrangements that link producers, consumers, and local authorities (Morgan and Sonnino, 2008; Perez, 2012) are on the rise.

Enlarging the Ranks

While the farming population of all OECD countries has significantly decreased in recent decades, in the Global South the number of peasant producers keeps growing. It is currently estimated that on a global level there are some 550 million peasant units of production. This is more than ever before in world history. Within this global panorama some telling episodes can be discerned. In Brazil, the Movement of Landless People (*Movimento dos Sem Terra*) is leading a massive flow of people from the *favelas* of the big cities or overcrowded rural areas who go and occupy large tracts of extensively used land, creating new peasant farms. Thus, between the last two agricultural censuses (1996 and 2006) the number of peasant farms grew by more than 400 000 units (MDA, 2009). Currently, the rural areas of Europe's Mediterranean countries are also witnessing an inflow of young people, partly supported by European and national policies. In China, a noticeable feminization of agriculture is occurring (Meng, 2014).

In brief, the composition of the populations economically involved in agriculture is changing in many parts of the world and, in some places, their absolute number is increasing. In terms of overall agricultural development this is highly important, not least because it helps to counterbalance rural aging (White, 2011).

Countering Big Capital

From a sociological point of view it is noteworthy that, for the first time in history, the world now has an organization, *La Via Campesina* (LVC, which translates as 'the peasant trajectory'), which has developed to become a truly global peasant federation. This is a well-coordinated agglomeration of many organizations and movements that operate at national, regional, and local levels (Borras et al., 2008). LVC gives a voice to those who would otherwise remain neglected in this world. It is one of the main forces resisting the processes of land and water grabbing around the globe. LVC is a strong advocate of agroecology (discussed above), which it considers as one of the main defense lines against big capital groups and their endeavors to conquer spaces and create new conditions for capital accumulation in the countryside. One of the themes that LVC has developed is the concept of 'food sovereignty,' often loosely defined as the right of people to decide on how they want the production, processing, distribution, and consumption of food to be organized. This represents a significant reintroduction of democratic principles and rights into the realm of food production.

Some of the main themes championed by LVC partly coincide with policies that are now emerging in different places, including the rural development policies that have been put in place in regions as different as the European Union, China, and Brazil (van der Ploeg et al., 2012). These policies seek to correct the negative, often destructive tendencies of the large commodity markets for food and agricultural products and provide the space for the different forms of peasant reinvention discussed in this article, which hold the key to generating new modes of agricultural development.

See also: Agricultural Sciences and Technology; Agroecology and Agricultural Change; Food in Anthropology; Geographies of Food: Politics, Place, Power Reprised; Peasants and Rural Societies in History (Agricultural History); Peasants in Anthropology; Rural Geography; Rural Planning: General; Rural Sociology.

Bibliography

Altieri, Miguel A., Funes-Monzote, Fernando R., Petersen, Paulo, 2011. Agroecologically Efficient Agricultural Systems for Smallholder Farmers: Contributions to Food Sovereignty. INRA and Springer-Verlag, Paris and Berlin.

Borras, Jun, Edelman, Marc, Kay, Cris, 2008. Transnational agrarian movements: origins and politics, campaigns and impact. In: Borras, S.M., et al. (Eds.), Transnational Agrarian Movements Confronting Globalization, Special Issue of Journal of Agrarian Change, vol. 8, 1/2.

Domínguez García, M.D., 2007. The Way You Do It Matters: A Case Study on Farming Economically in Galician Agroecosystems in the Context of a Cooperative. Wageningen University, Wageningen.

Haggblade, S., Hazell, P., Dorosh, P., 2007. Sectoral growth linkages between agriculture and the rural nonfarm Economy. In: Haggblade, S., Hazell, P., Reardon, T. (Eds.), Transforming the Rural Nonfarm Economy. John Hopkins University Press, Baltimore, pp. 141–182.

Halamska, M., 2004. A different end of the peasants. Polish Sociological Review 3 (147), 205–268.

HLPE (High Level Panel of Experts), 2013. Investing in Smallholder Agriculture for Food Security, a Report by the High Level Panel of Experts on Food Security and Nutrition, CFS (Committee on World Food Security). FAO, Rome.

IAASTD (International Assessment of Agricultural Knowledge, Science and Technology for Development), 2009. Agriculture at a Crossroads, Global Report. Island Press, Washington.

Joannides, J., Bergan, S., Ritchie, M., Waterhouse, B., Ukaga, O., 2001. Renewing the Countryside. Institute for Agriculture and Trade Policy, Minnesota, MN.

Kinsella, J., Bogue, P., Mannion, J., Wilson, S., 2002. Cost reduction for small-scale dairy farms in County Clare. In: van der Ploeg, J.D., Long, A., Banks, J. (Eds.), Living Countrysides, Rural Development Processes in Europe: The State of Art. Elsevier, Doetinchem, The Netherlands.

Langthaler, Ernst, 2012. Balancing between autonomy and dependence: family farming and agrarian change in Lower Austria, 1945–1980. In: Bischof, Günter, Plasser, Fritz (Eds.), Austrian Lives (Contemporary Austrian Studies XXI), pp. 383–402. New Orleans.

Low, Sarah A., Vogel, Stephen, 2011. Direct and Intermediated Marketing of Local Foods in the United States, ERR-128. U.S. Department of Agriculture, Economic Research Service.

Mamonova, N., Visser, O., Spoor, M., 2013. The Re-emerging Peasantry in Russia: 'Peasants-against-their-own-will', 'Summertime peasants' and 'Peasant-farmers'. Working Paper. ISS, The Hague.

McMichael, P., 2009. Banking on agriculture: a review of the world development report 2008. Journal of Agrarian Change 9, 235–246.

MDA (Ministério do Desenvolvimento Agrário), 2009. Agricultura Familiar no Brasil e O Censo Agropecuário 2006. MDA, Brasília.

MDA (Ministério do Desenvolvimento Agrário), 2013. Brasil Agroecológic, Plano Nacional de Agroecología e Produçao organica-planapo. MDA, Brasilia.

Meng, X., 2014. Feminization of Agricultural Production in Rural China. Sociological Analysis (Ph.D. manuscript). Wageningen University, Wageningen.

Morgan, K., Sonnino, R., 2008. The School Food Revolution: Public Food and the Challenge of Sustainable Development. Earthscan, London.

Oostindie, H., van der Ploeg, J.D., van Broekhuizen, R., 2013. Buffercapaciteit: Bedrijfsstijlen in de melkveehouderij, volatiele marketen en kengetallen. Wageningen University, Wageningen.

Ostrom, E., 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, New York.

Paredes, M., 2010. Peasants, Potatoes and Pesticides: Heterogeneity in the Context of Agricultural Modernization in the Highland Andes of Ecuador. Wageningen University, Wageningen.

Perez, Julian, 2012. A construção social de mecanismos alternativos de mercados no âmbito da Rede Ecovida de Agroecologia. MADE-UFPR, Paraná.

van der Ploeg, J.D., 2008. The New Peasantries, Struggles for Autonomy and Sustainability in an Era of Empire and Globalization. Earthscan, London.

van der Ploeg, Jan Douwe, Ye, Jingzhong, 2010. Multiple job holding in rural villages and the Chinese road to development. Journal of Peasant Studies 37 (3), 513–530.

van der Ploeg, Jan Douwe, Ye, Jingzhong, Schneider, Sergio, 2012. Rural development through the construction of new, nested, markets: comparative perspectives from China, Brazil and the European Union. Journal of Peasant Studies 39 (1), 133–173.

Reijs, J., 2007. Improving Slurry by Diet Adjustments: A Novelty to Reduce N Losses from Grassland Based Dairy Farms (Ph.D. thesis). Wageningen University, Wageningen, The Netherlands.

de Rooij, S.J.G., Ventura, F., Milone, P., van der Ploeg, J.D., 2013. Sustaining food production through multifunctionality: the dynamics of large farms in Italy. Sociologia Ruralis 54 (3), 303–320. http://dx.doi.org/10.1111/soru.12025. Article first published online: November 4, 2013.

Rosset, Peter M., Martinez-Torres, Maria Elena, 2012. Rural social movements and agroecology: context, theory, and process. Ecology and Society 17 (3), 17.

Saccomandi, V., 1998. Agricultural Market Economics: A Neo-institutional Analysis of Exchange, Circulation and Distribution of Agricultural Products. Royal van Gorcum, Assen, The Netherlands.

Schneider, S., Niederle, P., 2010. Resistance strategies and diversification of rural livelihoods: the construction of autonomy among Brazilian family farmers. The Journal of Peasant Studies 37 (2), 379–405.

Schneider, S., Shiki, S., Belik, W., 2010. Rural development in Brazil: overcoming inequalities and building new markets. Rivista di Economia Agraria, LXV 2, 225–260.

Toledo, Victor M., 2011. La agro-ecología en Latinoamercia: tres revoluciones, una misma transformación. Agroecología 6, 37–46.

White, Ben, 2011. Who Will Own the Countryside? Dispossession, Rural Youth and the Future of Farming. ISS (International Institute of Social Studies), The Hague.

Wu, Huifang, Ding, Baoyin, Ye, Jingzhong, 2014. The construction of new nested markets and rural development in China. In: Hebinck, P., Schneider, S. (Eds.), Rural Development and the Construction of New Markets. Routledge, London.

Ye, Jingzhong, Rao, Jing, Wu, Huifang, 2010. Crossing the river by feeling the stones: rural development in China. Rivista di Economia Agraria, Anno LXV 2, 261–294.

Peasants and Rural Societies in History (Agricultural History)

Janken Myrdal, Swedish University of Agricultural Sciences, Uppsala, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Studies of peasants in history, rural history, have expanded in three waves. The first spanned the late nineteenth and first half of the twentieth centuries and was dominated by ethnology–anthropology, whilst the second was driven by economic and social history during the 1950s and 1960s. The third wave began in the 1990s and involved different disciplines, eventually developing into transnational studies. Some overarching discussions have concerned: ownership and the community, the market and commercialization, and productivity and farming systems. During recent decades, peasants as agents in historical change have been recognized.

The concept of 'rural societies' is based on the division between rural and urban. There is a historical dimension as well since the concept refers to societies dominated by rural areas, and thus relates to rural history. In fact, there have been some major advances in the historiography of rural history since the turn of the twenty-first century (Thoen and van Molle, 2006; Alfonso, 2007).

Terminology

Discussing the terminology is an important part of theory building in the social sciences. An obvious case concerns the term 'peasant.' From the 1950s there were intense discussions about its meaning, and a large body of literature was published in the decades that followed. The debate was largely limited to the Anglo-Saxon world since most languages other than English do not make a clear distinction between peasants and farmers. In German, for instance, the term *Bauer* can mean both peasant and farmer (Rösener, 1993). The most important contributions to the discussion were published from the 1950s to the early 1970s (Redfield, 1956; Wolf, 1966; Potter et al., 1967; Shanin, 1971). Peasant society was defined on two levels: (1) chronologically in relation to earlier and later types of social systems and (2) socially in relation to the overall social structure within which it existed.

The peasant society was seen as a partial society (i.e., a part of the larger social structure). This formed the basis of the chronological separation of peasants from what was called the 'tribal society,' or, in a more precise terminology, a prestate society. The emergence of the town was identified as a decisive change. Peasants either were obliged or chose to send their produce to the town. The formation of the state and state authorities also had a major impact since peasant society became more obviously subordinated and even alienated from a larger social system as such. To some extent, both these criteria can be said to represent the same change. Peasant society has also been described as part of an overarching socio-economic feudal system (*see also* Feudalism). The other chronological distinction that separates peasants and farmers has also been discussed. Not only do modern farmers produce for the market, but also their actions are profit driven and farms are managed in a commercial way.

In 1971, Teodor Shanin (in his *Introduction* to Shanin, 1971) summarized the discussion about peasant society and highlighted some of its distinguishing characteristics: the family farm as the basic unit of multidimensional social organization; land husbandry as the main means of livelihood directly providing for the major part of the consumption needs; specific traditional culture related to the way of life of small communities; and the underdog position, specified as the domination of the peasantry by outsiders.

Different aspects were emphasized in the discussion. Wolf regarded the relationship to the market as decisive, whilst Redfield emphasized peasant tradition. Marxists later joined the discussion by stating that when peasants act as a group with class consciousness, they become a class in themselves. The Russian peasant researcher Alexander Chayanov stressed that the preservation of the family was of fundamental importance to the peasant (Chayanov, 1966).

It also became increasingly clear that the specific historical situation had to be taken into account. Rodney Hilton noted that his definition of the English medieval peasant was specific to a particular part of Europe (Hilton, 1975). This has connotations with the so-called Brenner debate, which was brought about by Robert Brenner's article in 1976 and followed by a discussion in which many leading European historians took part. It started with comparisons between French and English peasantries, but gradually came to include more regions. Class structure and the relation to the market were key issues (Aston and Philpin, 1985; Hoppenbrouwers and van Zanden, 2001).

The concept of the peasant nowadays tends to be used in a comprehensive sense to include all those who grow crops or raise cattle, on a partially nonprofitable base, whilst being integrated into a larger social structure. The definition includes peasantries in societies where the production of foodstuffs requires the major part of the society's labor.

Another important conceptual distinction involves the gradual transition from agricultural history to agrarian history and rural history. The first journals and publications considered 'agricultural history' to be the most natural term. This placed farming and production in the center, although other topics were also addressed. In the mid-nineteenth century, it became more common to refer to 'agrarian history,' broadening the scope of the subject to all aspects involving the local, often agriculture-dominated, community. In the first volume of the

Agricultural History Review in 1953, G.E. Fussel presented a list of books and articles on 'agricultural history.' Next time such a list was presented, by Joan Thirsk in 1955, it was on agrarian history. She also in 1956 took the initiative to the project *The Agrarian History of England and Wales*. The German term *Agrargeschichte* is quite broad, whilst the French concept of *études rurales* is even more inclusive.

In 1990 a new journal entitled *Rural History* was launched, to a degree in opposition to what was labeled 'cow and plow history.' It sought to link social, economic, and even cultural history to achieve a broader perspective. This partly reflected the 'cultural turn' in history writing as such, and partly it was just one stage in the long process of broadening this area of research.

Before World War II

Systematic research into peasants and rural history began at the end of the nineteenth century, and tended to focus on peasants in their own countries. Citizens of the United States perceived their country as an agricultural nation, and hence it followed that agricultural history was considered a subject of national interest. This soon crystallized into a discipline in its own right. The oldest journal on the subject, entitled *Agricultural History*, was founded in the United States in 1927. Whilst articles published in this journal have included subject matter from outside the United States right from the very beginning, the focus has largely been on American agricultural history.

In Europe, agricultural history was linked with the formation of national identity, as demonstrated by various national 'agricultural museums.' In Hungary, an agricultural museum was founded in conjunction with the country's millennial celebration in 1898. Similar museums were established in Denmark and the former Czechoslovakia at the same time. The second oldest journal in the agricultural history was associated with the National Museum of Agriculture in Czechoslovakia (*Věstník Československého Zemědělského Muzea* (1928–43) and Vědecké Práce Zemědělského Muzea (1961-onwards)).

Widening the perspective from rural history as such, folk-life studies (ethnology) focused on peasants, who were regarded as representing the real heart of the nation. Folk-life studies became common in Eastern Europe, Scandinavia, Germany, France, Portugal, Ireland, Scotland, and other countries and regions. The research focused on both spiritual and material culture, including folksongs and riddles as well as tools and houses. Even if the original goal was to use ethnology as a way to understand national identity, the opposite was increasingly found to be true. Instead of a folk culture separating and uniting the nation, it was found that every region and nation existed in a flow of influences. Cultural phenomena moved freely across borders. When a national identity was built on ethnology, often referring to the peasants, it had to be invented; Germany in the 1930s is the typical case (though we also have others). A number of national atlases on peasant society and its manifestations were published, yet unfortunately no all-encompassing European atlas was produced.

In countries with colonies, in particular the United Kingdom, the study of peasant societies outside Europe became increasingly popular at the end of the nineteenth century. Originally, these studies were part of the colonial administration (e.g., the genre of studies focused on the Indian countryside that developed during the nineteenth century). Eventually, social anthropology developed into a discipline in its own right. The purpose of these studies was to understand the local community, with a focus on a village or a region. However, they gradually pointed to a different result. It was increasingly realized that local society was integrated with, and subordinate to, larger social structures. This sparked the aforementioned debate on the definition of the peasant concept.

In Europe, the first wave of interest in the history of agriculture, ethnology, and anthropology focused on the immediate past. At that time history, as a discipline, was still mainly oriented toward the nation state and the elite classes. Rural history studies first developed in France under the influence of geography, where a dominating group of scholars favored all-embracing studies where every aspect of a region should be described, including history. The historian who succeeded in giving rural history a solid base was Marc Bloch with his publication *Les caractères originaux de l'histoire rurale française* (*French Rural History*), which appeared in 1931 (Bloch, 1966). It was based on a series of lectures that he had held at the Institute for Comparative Research in Human Culture in Oslo in 1929, which had been established to promote folk-life studies.

Bloch's work still provides inspiration and guidelines to a number of areas: the role of the village community, peasant uprisings, the links between technology and social change, and so on. However, it was not until after World War II that his books garnered widespread recognition.

Breakthrough after World War II

The subsequent success of Bloch's work needs to be explained in terms of the changes that took place in the study of history. During the two decades following the end of World War II, agricultural history became a discipline in its own right in Western Europe. In the 1950s agrarian or agricultural history groups were established at several universities across Europe: at the University of Wageningen (Netherlands), the University of Hohenheim (Germany), and the University of Reading (United Kingdom).

A number of national European journals were also founded at this time. In the major languages we have the *Agricultural History Review* (1953–) in the United Kingdom, *Zeitschrift für Agrargeschichte und Agrarsoziologie* (ZAA) (1953–) in Germany, and *Études rurales* (1961–) in France. Other journals were published in the Netherlands (1953–), Denmark (1956–), Hungary (1957–), and Italy (1961–). The aforementioned Czech journal at the agricultural museum was relaunched in 1961, and a similar journal came in Bratislava (Slovakia) from 1962. The German journal *ZAA* in 1954–57 published historiographical essays on agricultural history in France, the United Kingdom, Scandinavia, and Poland, written by (for instance) Georges Duby and Joan Thirsk.

In time, national syntheses comprising several volumes were published. The German *Deutsche Agrargeschichte* consists of six volumes and was published between 1962 and 1984. The *Histoire de la France rurale* was produced between 1975 and 1978,

which comprises four volumes. *The Agrarian History of England and Wales* was published in several volumes from 1967 to 2000, and it remains one of the largest national history projects ever conducted, comprising a total of over 9000 pages.

This wave of research had an increased focus on long-term change, not least with respect to the Middle Ages. Two important works were published in the early 1960s. One was George Duby's *L'économie rurale et la vie des campagnes dans l'Occident medieval* (it was published in French in 1962, and the English translation appeared in 1968). In 1963, B. H. Slicher van Bath published *The Agrarian History of Western Europe A.D. 500–1850*, which had originally appeared in Dutch in 1960.

The agrarian crisis of the late Middle Ages became a central field of research in economic and social history. Although this economic decline had already been observed prior to the World War II (in Germany, Norway, and the United Kingdom), it did not gain general acceptance until the 1950s. The discovery was theoretically significant since it showed that the development of production and population had not always been positive.

Another important theoretical change was that agriculture was seen as a dynamic sector in medieval and postmedieval Europe, which rejected earlier theories about the countryside having a restraining effect on economic development.

There are many underlying reasons for the breakthrough of rural history in the 1950s. One factor was the increased political importance of peasants on a worldwide scale and in conjunction with colonies attaining their independence. The immediate conclusion was that peasants and rural societies were identified as essential elements of the social structure. Yet the dominant premise remained that peasant culture had filtered down to the peasants from the elite. Foster (1967) was one of several scholars to make this interpretation in a rather explicit form. Yet another conclusion gradually emerged. Redfield (1956) drew a distinction between a 'great tradition' and a 'little tradition' – between the classical, learned elite culture, and the peasant tradition – and he emphasized the existence of a mutual exchange.

The view of a conservative peasant society unwilling to accept changes that may upset traditions is similar to what is sometimes called 'the dumb peasant theory.' It implied that peasants had to be forced to develop, an assumption that largely guided Western relations with the Third World in the 1950s and 1960s, and to some extent since.

The 1970s and 1980s

From the 1970s onward, studies of peasants and rural societies became more diversified. Two partially linked phenomena began to change the picture – one was the challenge that Marxism presented to academia and the other was the success of peasant uprisings, not least in the Vietnam War. These developments challenged old thought structures, but did not lead to a new hegemony. After the major theoretical battles of the 1970s, the field was instead left open to a number of different theories.

A common question that many of the new interpretations struggled with was the relationship between agency and structure, with peasants being increasingly regarded as real actors in the historical process. One expression of this new and deeper interest in peasant studies was the creation of the *Journal of Peasant Studies* (1974–). Its editorial policy stated that the journal would be concerned with political economics and that it would attempt to understand the role of peasants worldwide, in particular outside Europe and North America.

The previous focus of anthropological research on material cultural products weakened. Instead, it was the social and cultural structure that attracted increasing attention, and this was not least evident in discussion about peasant societies. Clifford Geertz was a key figure. In the 1960s, he made important contributions to the theory on agrarian technology, and then gained even more influence with his *The Interpretations of Cultures* from 1973.

The classical study of material culture continued for some time, especially in the exchange between Eastern and Western European researchers. The Danish journal *Tools & Tillage* was established (1968–95), and for a while it was regarded as the leading international journal dealing with the development of agricultural technology in peasant societies. Of special importance as a meeting place was the Association International des Musées d'Agriculture. It was formed in 1974, formally under the jurisdiction of UNESCO. Meeting venues alternated between Eastern and Western Europe and the United States. The conference reports were published as *Acta Museorum Agriculturae* (1966–88).

There was a growing interest in rural history outside Europe, as exemplified by the publication of the multivolume *A History of Agriculture in India* in the 1980s, and the series *Science & Civilisation in China*, edited by Joseph Needham, with a remarkable volume on *Agriculture* from 1984 written by Francesca Bray.

Land reform and class conflict were intensely debated, and studies of peasant revolts and uprisings became a special genre. The geographical focus of these studies was on Europe, Asia, and Latin America, whilst studies on Africa begun early on to be more focused on the environment.

New interpretations developed in the study of long-term changes in European history as the class perspective was introduced (e.g., the Brenner debate). At the same time, there was a shift in interest from pure agrarian history to general cultural history. Here we can draw a parallel between anthropology and history, although historians had a much more recent interest in agrarian matters. By 1990, it seemed as though studies about peasants and rural society were stagnating and even decaying. However, in reality the opposite was true: rural history stood on the verge of a new leap forward.

Reorganization from the 1990s

The first wave of rural history that spanned the late nineteenth and first half of the twentieth centuries had been dominated by ethnology–anthropology, whilst the second was driven by the considerable advance of economic and social history during the 1950s and 1960s. The third wave began in the 1990s and involved different disciplines.

One such group of disciplines was social and economic history. New journals were launched: *Rural History* (1990–) in the United Kingdom; *Noticiario de Historia Agraria* (1991–), from 1998 renamed *Historia Agraria* in Spain (in Barcelona, an agrarian history journal had been published since 1983); and

Histoire et Sociétés Rurales (1994–) in France. The Royal Swedish Academy of Agriculture and Forestry (KSLA) published a series on agrarian history from 1991 onward. Higher education institutions and courses were organized in a number of countries (e.g., Austria, Belgium, Hungary, and Sweden). National multivolume syntheses were published in several European countries: in Italy during the 1980s and in 2002, in Germany during the 1990s, and in Denmark during the 1980s and in the rest of the Nordic countries (Sweden, Norway, and Finland) from 1998 to 2004. This quasi-joint Nordic endeavor came to cover, in 16 volumes, almost as many pages as *The Agrarian History of England and Wales* mentioned above. The Norwegian and the Swedish projects were also later summarized in English in two separate volumes.

The formation of three networks in economic and social history was a key development. The first, based in Belgium, was the Comparative Rural History of the North Sea Area founded in 1996 by Erik Thoen. An impressive series of books have been published, now summarized in four volumes covering northwestern Europe. The Rural History Network was formed in 2000 as part of the biannual European Social Science History Conference. The third network to be formed was the Program for the Study of European Rural Societies. It came into being in 2005 thanks to funding from the European Union's research program Cooperation in Science and Technology (COST). Hundreds of European scholars have made contacts and worked together through these networks. In 2010, the first European (and in reality international) conference for rural and agrarian historians was held in the United Kingdom and resulted in the formation of the European Rural History Organisation. The second was held in Switzerland in 2013.

Some main topics in these three networks have included property rights, the role of the state, commercialization, and food processing. Although peasant culture proper has not been extensively researched, topics such as gender and family structures have attracted much attention. Landscapes as well as agrosystems have been among other favored topics. This new rural history had new questions and answers, but largely remained in the area of research where the inclusive agrarian history had been for decades. There was, however, a shift in chronological focus from the Middle Ages to the nineteenth and twentieth centuries, which now dominates syntheses and conferences.

Medieval rural history reappeared in another context, as a parallel movement that evolved in archaeology and came to include the remnants of classical rural ethnology. There were conferences on medieval rural archaeology, and papers were subsequently published in the biannual journal *Ruralia* (1996–). Prehistorical archaeology is by definition rural, whilst antiquity is a major field of research in itself. Thus rural archaeology came from medieval archaeology. Agrarian issues had long been developing as new methods merged with new questions. New scientific methods such as pollen analysis, osteology, and dendrochronology provided answers to questions scholars had never before been able to even ask about landscapes, cultivation, and animal husbandry. Archaeological excavations at the same time moved from mounds and castles to fields and villages.

The rural archaeology conferences began with an inventory of rural settlements across Europe, and then addressed a number of different themes, often corresponding to those seen in social and economic history. Examples include marginal lands, landscapes, and even food processing (including markets and commercialization). Not surprisingly, archaeologists focused on buildings, which resulted in a direct interconnection with classical ethnology. Chronologically the focus was on the Middle Ages, although ethnological comparisons stretched right into the nineteenth century. With new excavation methods and ambitions, rural archaeology is also moving far into the modern period.

In recent decades, environmental issues have acquired an increasingly historical dimension with the recognition of long-term human impact. Journals were launched. The American *Environmental History* is the most influential. It was created in 1996 by merging two established journals: the *Environmental History Review*, founded in the 1970s, and *Forest & Conservation History*, which dated back to the 1950s. Forest history has generally tended to be included as part of environmental history research. In 1995 a European journal entitled *Environment and History* was founded. Both this journal and the aforementioned *Environmental History* have strong links with science. The issues covered have included long-term sustainability in relation to current problems such as soil erosion, salinization, and climate change.

This renewed interest in rural history did not just include the Western world, as exemplified by the formation of the Indian journal *Asian Agri-History* (1997–). Anthropology also enjoyed a renewed interest in peasant societies. An example is that Yale University in 1991 launched its Agrarian Studies Program, where James Scott was a central figure. He had already published important works about peasants in Southeast Asia in the 1970s. In 2009, he released a groundbreaking book about peasants and the state, in which he argued that many peasants actively avoided intensification of agriculture to escape taxes and control.

These movements in social and economic history, in archaeology, and in anthropology have rather few interdisciplinary contacts so far. In **Figure 1** I have tried, in a generalized way, to describe in what other major topics this research was conducted. I also tried to describe major flows of influence, though of course a web of contacts in fact existed between scholars in these different disciplines doing rural history.

A possible merger in the future could be the growing interest in continental and global syntheses. The increasing number of regional studies and syntheses from conferences and books published in English provide the material for such major projects (Rasmussen, 2010). However, new methods will be required to handle diversified and at times overwhelming information. An academic overview by Marcel Mazoyer and Laurence Roudart was published in 2006 (it first appeared in French in 1997).

Some Major Discussion

The discussion about the definition of 'peasant' meant that peasants came to be regarded as a part of a larger society. Once the crisis of the late Middle Ages was taken into account, the long waves of history came into focus. During the 1970s, research often emphasized social conflict, although it also expanded into new fields and this formed the basis for the

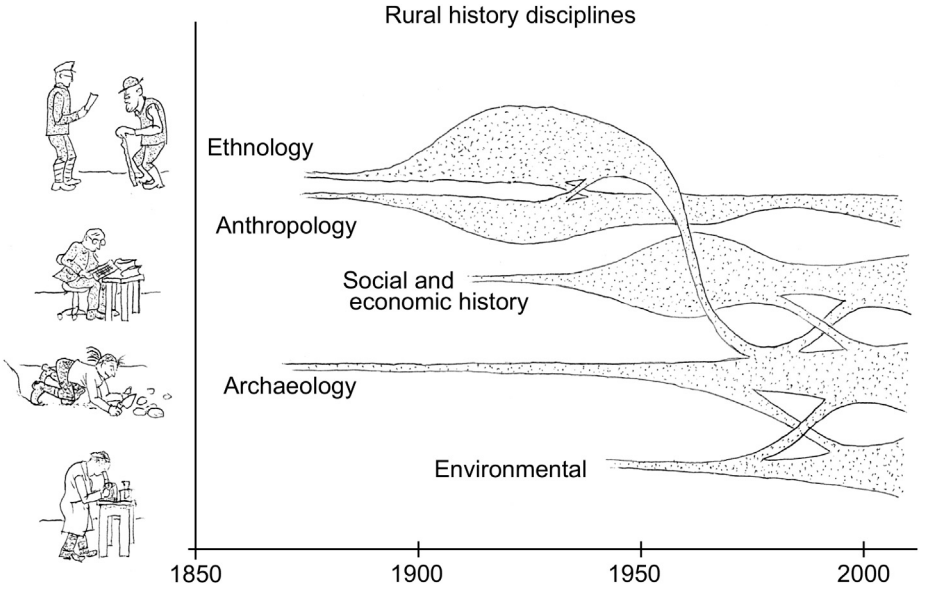

Figure 1 A sketch of major academic disciplines in which the subfield rural history was conducted. There are no strict relative proportions. Major flows of influence have been indicated, but a web of other contacts in reality existed between scholars in different disciplines. Sketch made by J. Myrdal and P. Thunström.

revival in the decades that followed. Yet it was the recognition of peasants' agency that most changed perspectives on a general level. The old idea of elite-dominated history was a kind of history where the majority of the people were seen as the boards of a stage on which the actors – kings and nobility – performed their play. This view has gradually been abandoned, and a fuller and richer drama appears before us.

A few constellations of themes will be mentioned, where discussions have been intense and advances considerable. One such concerns productivity, and it is rooted in the works of Slicher van Bath, Duby, and other early publications on agrarian history. Much of the later arguments are technical in nature, concerning problems of source criticism and different ways of measuring productivity.

The discussion about productivity prompts some overarching issues. One is the reconstruction of historical per capital production. Agricultural production is the single most important factor, besides demographics, prior to the industrial revolution when calculating total production. In 2005, Giovanni Federico compiled national data from the last two centuries. Now research is pushing even further back in time. If reliable data sets can be created, this will be of importance for some of the major debates. One is the ongoing debate on: Why the West? Another is the question about why and how historical expansion of production and population occurred. The Malthusian pessimistic and Boserupian optimistic views about ongoing intensification have been challenged, leading us to the next overarching issue: farming systems (Boserup, 1965).

Ideas about farming systems begin with the view that agriculture is an integrated system of interlinking elements. Scale is of importance, as we have systems from the farm level to the global level. Emphasis has often been placed on regional systems, but these must be conceptualized as subgroups when the discussion shifts to global farming systems. This change of perspective may well pose the greatest challenge to European rural history, since it is so well developed and thus has identified a number of regional systems and differences (Grigg, 1974: 152).

Another part of the farming system discussion concerns technology. The technological complex can be regarded as such, separate, and then compared and related to the social structure, or one can already from the start regard an integrated agrosystem, where social structure is inseparable from technology (Astill and Langdon, 1997; van Bavel and Thoen, 1999). As in so many discussions, the somewhat unexciting answer is that we have to do both kinds of analysis.

One problem is that research into technology was already languishing prior to the revival of rural history in the 1990s, and it has not recovered since. Tangible material will always remain essential to archeological research, and archaeology has adopted some of the earlier ethnological issues, such as buildings and construction. Classical questions about plows, harrows, and shepherding have not, however (bar some important exceptions) been intensively researched and discussed in recent decades.

The growing importance of science is another aspect of the discussion about production. Agrarian history could be an interdisciplinary field that straddles natural science and the humanities. Whilst simple facts (e.g., the nitrogen cycle) have been known for a long time, more complex ecological models are now gradually being used. An example is biologists using historical data to explain prevailing biodiversity, not only in Europe but also, for instance, in the Amazonas.

Major clusters of discussions often are like branches of an enormous tree, stretching out from a common trunk of basic themes. Discussion about themes such as property and the community is an example, where two large and interlinking branches are ownership and social conflict, whilst other

branches deal with questions about village communities and collective action. At first glance, it may seem that there are two opposing views about the relationship between different classes – one that points to conflict and tension as the essential driving force, and another that emphasizes cohesion and social symbiosis as the foundation of a society. A closer look reveals that both viewpoints often merge into a single approach, which shows that societies in which peasants (and the general public) have more influence seem to be less troubled by harsh conflict (e.g., Blickle, 2000). This change of perspective can also be seen as a shift of interest from serfs to free peasants (van Bavel and Hoyle, 2010).

A fast-growing branch of rural history is the discussion about how peasants handled common resources. An earlier, rather negative view of the village community and their common land has come to be replaced by a more positive stance. The peasant community was earlier seen as a hindrance to technological development, and it was also perceived as a threat to the environment. The idea of 'the tragedy of the commons' implied that everyone would take as much as possible from a common pool of resources. To counter this, the American political economist Elinor Ostrom and others showed that a functioning community could in fact manage common resources sustainably, and the establishment of such a community must be regarded as a historical process (Ostrom, 1990; De Moor et al., 2002). This not only had an impact on the environmental discussion but also prompted renewed research into village communities.

Another constellation of themes and discussions concerns the market. The peasant-farmer discussion is obviously rooted in this, yet the market was now no longer seen as alien or even hostile to a peasantry striving for self-sufficiency. Instead, the market came to be seen as a driving force and even a primary goal for production in rural societies. This links in with the aforementioned discussion about farming systems and productivity. Several important case studies have been produced, including Bruce Campbell's research on medieval London (Campbell et al., 1993). This also feeds into other key discussions conducted in recent decades. If peasants, who made up the bulk of the population, had goals other than just producing food for themselves, for instance if they strived for 'gadgets,' this could incite them to work harder. This is connected with the theory of the 'industrious revolution' (De Vries, 2008).

There are of course numerous other major discussion threads, such as gender. This has been very much steered into a discussion about marriage patterns and the division of labor, on both European and global levels (as is true of all discussions mentioned here). Environmental issues could also be examined in more detail, not least with respect to climate change. The antiquity has hardly been mentioned, though research has been very successful, highlighting subjects like water management and grain trade. The most important subject not given its space is probably global farming systems. Research is making great strides in exploring the agrarian and rural histories of Asia, Africa, and the Americas. The next leap forward will certainly be based on this new knowledge base.

The overarching themes mentioned here – the property and community, the market and commercialization, and productivity and farming systems – have one common denominator: They all basically regard peasants as actors in their own right. In coming decades, it can be expected that this will be counterbalanced by growing global research, which will show large systems and path dependency within these systems. Rural history will be more globally interconnected because scholars now are publishing material in a common language: English. Historical knowledge about land use will presumably be much in demand, as such knowledge is of great use in helping to solve such global problems as climate change. One interesting question is, then, how this grand-system approach will work with the peasant-actor approach that has become so strong in rural history over recent decades.

See also: Agricultural Sciences and Technology; Agriculture, Economics of; Bloch, Mark Léopold Benjamin (1886–1944); Boserup, Ester (1910–99); Crowds in History; Economic History; Environmental History; Feudalism; Food in Anthropology; Labor History; Nomads and Nomadism in History; Peasants in Anthropology; Rural Geography; Work, History of.

Bibliography

Alfonso, I. (Ed.), 2007. The Rural History of Medieval European Societies: Trends and Perspectives. Brepols, Turnhout.

Astill, G.G., Langdon, J. (Eds.), 1997. Medieval Farming and Technology: The Impact of Agricultural Change in Northwest Europe. Brill, Leiden.

Aston, T.H., Philpin, C.H.E. (Eds.), 1985. The Brenner Debate: Agrarian Class Structure and Economic Development in Pre-industrial Europe. Cambridge University Press, Cambridge.

Blickle, P., 2000. Kommunalismus: Skizzen einer gesellschaftlichen Organisationsform 2: Europa. Oldenbourg, München.

Bloch, M., 1966. French Rural History. Routledge & Kegan Paul, London.

Boserup, E., 1965. The Conditions of Agricultural Growth: The Economics of Agrarian Change Under Population Pressure. Allen & Unwin, London.

Bray, F., 1984. Science and Civilisation in China 6:2 Agriculture. Cambridge University Press, Cambridge.

Campbell, B., et al., 1993. A Medieval Capital and Its Grain Supply: Agrarian Production and Distribution in the London Region c. 1300. Institute of British Geographers, London.

Chayanov, A.V., 1966. In: Thorner, D., Kerblay, B., Smith, R.E.F. (Eds.), The Theory of the Peasant Economy. American Economic Association Translation Series, Homewood, IL.

De Moor, M., Shaw-Taylor, L., Warde, P. (Eds.), 2002. The Management of Common Land in North West Europe, c. 1500–1850. Brepols, Turnhout.

De Vries, J., 2008. The Industrious Revolution: Consumer Behavior and the Household Economy, 1650 to the Present. Cambridge University Press, Cambridge.

Duby, G., 1968. Rural Economy and Country Life in the Medieval West. Edward Arnold, London.

Federico, G., 2005. Feeding the World: An Economic History of Agriculture, 1800–2000. Princeton University Press, Princeton, NJ.

Foster, G., 1967. What is a peasant? In: Potter, J.M., Diaz, M.N., Foster, G.M. (Eds.), Peasant Society, A Reader. Little, Brown, Boston, MA.

Geertz, C., 1973. The Interpretation of Cultures: Selected Essays. Basic Books, New York.

Grigg, D.B., 1974. The Agricultural Systems of the World: An Evolutionary Approach. Cambridge University Press, London.

Hilton, R.H., 1975. The English Peasantry in the Later Middle Ages. Clarendon Press, Oxford, UK.

Hoppenbrouwers, P., van Zanden, J.L. (Eds.), 2001. Peasants into Farmers?: The Transformation of Rural Economy and Society in the Low Countries (Middle Ages – 19th Century) in Light of the Brenner Debate. Brepols, Turnhout.

Mazoyer, M., Roudart, L., 2006. A History of World Agriculture: From the Neolithic Age to the Current Crisis. Monthly Review Press, New York.

Ostrom, E., 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, Cambridge.

Potter, J.M., Diaz, M.N., Foster, G.M. (Eds.), 1967. Peasant Society, A Reader. Little, Brown, Boston, MA.
Rasmussen, R.K. (Ed.), 2010. Agriculture in History (vols. 3). Calif.: Salem Press, Pasadena.
Redfield, R., 1956. Peasant Society and Culture: An Anthropological Approach to Civilization. The University of Chicago Press, Chicago, IL.
Rösener, W., 1993. Die Bauern in der europäischen Geschichte. CH Beck, Munich, Germany.
Scott, J.C., 2009. The Art of Not Being Governed: An Anarchist History of Upland Southeast Asia. Yale University Press, New Haven, CT.
Shanin, T. (Ed.), 1971. Peasants and Peasant Societies. Penguin, Harmondsworth, UK.
Slicher van Bath, B.H., 1963. The Agrarian History of Western Europe AD 500–1850. Edward Arnold Limited, London.
Thoen, E., van Molle, L. (Eds.), 2006. Rural History in the North Sea Area: An Overview of Recent Research, Middle Ages – Twentieth Century. Brepols, Turnhout.
van Bavel, B., Hoyle, R., 2010. Social relations: property and power. In: Thoen, E. (Ed.), Rural Economy and Society in North-Western Europe, 500–2000. Brepols, Turnhout.
van Bavel, B., Thoen, E. (Eds.), 1999. Land Productivity and Agro-Systems in the North Sea Area: (Middle Ages – 20th Century): Elements for Comparison. Brepols, Turnhout.
Wolf, E., 1966. Peasants. Prentice-Hall, Englewood Cliffs, NJ.

Peasants in Anthropology

Enrique Mayer, Yale University, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Kearney, volume 16, pp. 11169–11171, © 2001, Elsevier Ltd.

Abstract

Peasants, an anthropological approach, begins with Kroeber's definition of them as part societies and part cultures, describing a segment of national societies as a class of people who are family-based agriculturalists who produce food for themselves and for the market. They also provide (under various forms of coercion) labor, rent, taxes, and manpower for armies or pharaonic constructions such as temples, roads, or dams. Their culture is folk based but related in interesting ways to high-cultured learning practiced by elites. Politically, peasants are often subjects of landowners and even in free communities, they are subservient to the dictates of members of the state and to ecclesiastical control. Regularly exploited and overworked, they occasionally rise in rebellion and participate in revolutionary wars, but more frequently resist exactions in devious ways, driven by attempts at self-defense that appeals to an ethic, named by James C. Scott, the moral economy of the peasant. It asserts the higher moral right to food and shelter and to family social relationships in the face of constant expropriation of their crops, labor, animals, and land. Increases in land productivity are achieved through labor-intensive ways and environmental management seeking sustainability as a way of life rather than the single-minded pursuit of profit. Their participation in markets as traders and sellers of labor and crops is an important aspect of their life, even though part of their production is consumed directly. In modern times, peasants, in alliance with other political actors, resist changes in agricultural technologies such as mechanization, chemical fertilizers and pesticides, genetically modified crops, and the extraction of mineral or oil resources that adversely affect them. Peasants also extend their lifestyles to urban areas where, as casual labor migrants, they re-create zones of peasant life reminiscent of their rural home base, even if they now travel to other countries to find work and send money back to sustain the peasant base.

Peasants are primarily rural dwellers who, with their traditions, labor, tools, and land, produce food for themselves and for nonfarm populations. Although they are contained within the boundaries of countries, their cultural patterns are set apart and therefore aptly defined by Alfred Kroeber (1948) as "part societies and part cultures." In common English, the word peasant also carries an offensive overtone referring to an uncouth person, ill mannered and uneducated, an opinion originally held by the mannered aristocrats of renaissance times in Europe. Other, more positive and romantic, meanings have been added and these have shifted over time. Intellectuals like the Brothers Grimm the German folklorists, Vincent van Gogh the Dutch painter, Nicolai Rimski Korsakoff the Russian musician, or Amalia Hernández the Mexican ballet choreographer found sources for inspiration in portraying aspects of peasant life with admiration that enhanced the identity of their nations.

In Europe, the musings of Karl Marx, Friedrich Engels, and Vladimir Lenin on the defeudalizing process brought about by the bourgeois revolution and the role played in it by the peasantry have always provided a source of inspiration for peasant studies (Marx, 1971; Lenin, 1982). Theodor Shanin's (1973, 1971) *The Awkward Class* and *Peasants and Peasant Societies* on the East European peasantry are important references for its complete treatment of serfs and independent peasants that bolster Immanuel Wallerstein's (1974) general point that it is in the semiperipheries of the world capitalist system, like Eastern Europe, where in agricultural estates or in small villages, a subordinate class of peasants were the food producers, the tax and tribute payers, and the suppliers of soldiers.

Peasant Cultures

Sociocultural anthropologists doing field work in Europe concentrated their work in marginal regions of Europe such as Italy (Banfield, 1958), Ireland (Arensberg and Kimball, 1968), Greece (Peristiany, 1966), or Spain (Pitt-Rivers, 1966; Kenny, 1962) looking for particular regional cultural patterns such as aspects of envy, honor, and shame or forms of self-government and lack of cooperation because they felt peasant cultures there had remained more ingrained than elsewhere in that continent.

Latin America, particularly Mexico and the Andes where the Aztec and Inca civilizations had established a stable tribute paying agricultural peasantry that the Spanish empire took over and transformed, is another Wallersteinian semiperiphery. Anthropologists specialized in community studies where the emphasis was strongly cultural and local, in that, Mexican peasant culture and personal character traits seemed to form distinct patterns worthy of note. Much of this work was inspired by the pioneering work of Robert Redfield (1973) whose description of a 'folk society' in Tepotzlán revealed the positive value systems he thought were at the heart of the peasant ethos. Oscar Lewis (1963) restudied the village and found much that was negative and nasty, raising the question whether anthropologists find in the field that which they are looking for. With the 'folk urban continuum' Redfield (1956) developed from field work in the village of Chan Kom, Yucatan, a mixed Maya and Spanish language population practicing a syncretic religion that blended Catholicism with Mayan beliefs, he pointed out how this folk society was neither tribal, nor urban, but in between, sharing characteristics of both and yet unique and whole. He also

showed how cultural patterns derived from dominant urban centers of power, learning, and cultural production of mostly Hispanic (the 'big tradition') tended to be whittled down and be reinterpreted to make them compatible with the villager's oral folk practices (of the 'little traditions'). This generalization is applicable to other stratified cultural contexts such as India or China.

Eric Wolf and Sydney Mintz (1955) outlined the conditions under which rural proletarians who worked in capitalist plantation systems were distinguishable from traditional rural estates (*haciendas*) with peasant serfs bound to the landlords through other forms than wages. Elsewhere 'free' peasants were of two types, 'closed corporate communities' (Wolf, 1955) with collective land holdings and civil religious self-managed local authorities were the result of the colonial policies of indigenous reservations, ecclesiastic control, and taxation, contrasted with the more 'open' market oriented and mestizo colonizers of peripheral areas as producers of mixed subsistence and exportable cash crops, such as coffee production in Venezuela (Roseberry, 1983) or even on the banks of the Amazon river in Brazil (Harris, 2000). Wolf's short textbook *Peasants* (1966) summarized extensive scholarship and is widely used by scholars because he wrote with clarity and a perspective of a peasantry as a type that is an integral part of the world.

Peasant Politics

The fact that Mexican peasants had participated extensively in the 1910 Mexican Revolution that culminated in an agrarian reform that created modern collective land holding units for them (*ejidos*) and that the ruling party relied on the support of peasants (*campesinos*) adds a necessary political component to the conceptualization of peasants in social sciences. Beyond liberating themselves from landlords, peasants have developed complicated relationships with national states. In Latin America, peasant rebellions, peasant movements, and oppositional stances are extensively covered in the literature by scholars and protagonists (Landsberger, 1969; Handelman, 1975; Zamosc et al., 1997; Blanco, 1972; Marcos, 1995; Menchú, 1964).

James C. Scott working in Malaysia, outlining the *Moral Economy of the Peasant* (1976) described how a collective ethos of the right to subsistence, and its defense against its encroachment, underwrites the political stances peasants will take when confronting exploitative landlords, traders, or members of the state. Not always violent and only partially effective, as he points out in *Weapons of the Weak* (1985). Resistance has underscored the complex confrontational aspects as a characteristic practice of peasant politics in the defense of their modes of livelihood. Indeed, rather than allies, Bernstein (1994) characterizes the peasants as victims of rapacious states reminding us of murderous land collectivization of the Soviet and Chinese regimes, repression campaigns in Central America, Southeast Asia, or India. The state against the peasants also manifests itself in less strong terms, as cooperativization schemes in Peru (Mayer, 2009) even in supposedly benevolently intended development programs (Escobar, 1995) but nonetheless imposed and counterproductive. One can observe most recently, peasant resistance against multinational companies whose extractive industries like mining, oil, logging, and other environmentally damaging resource extractions are opposed by affected populations (Bebbington, 2011). In the changed politics of these times, these movements often clothe themselves in ethnic terms as Indigenous rather than as peasants, although they are both. Wolf's *Peasant Wars of the Twentieth Century* (1973) goes beyond resistance, because he studied active peasant participation in the Mexican, Russian, Chinese, Vietnamese, Algerian, and Cuban revolutions, and in *Europe and the People without History* (1982) used Marxism to point to the expansion of capitalism that generated the peasantries of the world, while at the same time ignoring their historical role in this narrative.

Peasant Farm Ecology

Clifford Geertz (1963) using ecological approaches described how peasant agricultural intensification in Javanese irrigated rice terraces *sawa* 'involuted' as increasing population and rising taxation requirements squeezed ever more products out of their environmentally amazingly well-managed agroecosystems. Family-based labor-intensive agricultural systems produce surpluses and generate inward-looking self-managed culturally distinct ecological arrangements controlled by village governance such as the Mexican maize *milpa* (Alcorn, 1984), the Andean maize and potato *chacras* (Brush, 2004), or the Swiss Alpine dairy systems (Netting, 1981, 1993). In all of them, it is the autonomous family farm that is part of a collectively self-governing village that often holds parts of the land collectively and manages a 'commons' (Ostrom, 1990) that guarantees production, manages the internal politics, and ensures sustainability.

The farming family peasant household has certain characteristics that differ in degree from what is a basic assumption in economics. Because there is no separation between the resource-holding household and the income-producing household, the efficiency models of the single-minded profit-seeking small firm in economic theory does not fully apply to peasant households. Given that the prime responsibility is to provision family members, the needs of the household often loom larger than the profit-seeking objective. Moreover, as Stephen Gudeman points out in *Conversations in Colombia*, material flows of labor expenditures and harvested farm products flowing over time through the peasant farm ecosystem allow it to become more diversified, to grow in biomass, and increase in members' well-being without any clear cutoff point: "The house model has no goal or end, it is a mode of life, a set of practices" (Gudeman and Rivera, 1990: p. 116). An Indigenous formulation with Quechua and Aymara concepts of similar subsistence-oriented ideas has recently been enshrined as the *buen vivir* goal in the constitutions of Bolivia and Ecuador (Gudynas, 2011).

Western technological revolutions in agriculture, many of them associated with capitalist agriculture and multinational corporations, although responsible for vastly increased output, nonetheless have negative consequences for peasants who have been displaced, expropriated, marginalized, proletarianized, injured, or starved in various regions of the world.

The process of expropriation to make room for modern agriculture is often violent and legal redress slow or absent, creating a political dimension to this process. Mechanization comes with consolidation of land holdings, provoking land invasions and calls for agrarian reforms. Chemical fertilizers and pesticides bring with it soil and environmental degradation, pollution, and illnesses, raising the call for regulation and protection as well as the implementation of more environmentally benign forms of production. The plant genetic revolution for high-yielding varieties and genetically modified seeds of food crops causes the loss of biodiversity with it fears of future unintended, uncontrollable biological phenomena. It also concentrates ownership over the seeds in the hands of private multinational corporations, raising ethical and practical concerns about who owns seeds and whether they can be commoditized. Concentration of food-processing and distribution networks kills small businesses on which peasants rely. Industrialization of foodways has consequences for the health of the people it feeds.

In many parts of the world, peasants and their allies are resisting these trends in various ways, by forming new social movements, joining the environmentalists and consumers, appealing to international organizations for rights and protection, and inventing alternatives. Examples are the international alliance of organizations *Via Campesina* (peasants way) and the *Slow Food* movement which defend peasant production and processing strategies. *Via Campesina* states in its brochure that "small farmers, including peasant fisher-folk, pastoralists and indigenous people, who make up almost half the world's people, are capable of producing food for their communities and feeding the world in a sustainable and healthy way" (http://www.viacampesina.org).

The Dutch anthropologist Jan Doeve van der Ploeg (2008) studied several instances of 'repeasantization' around the globe describing ways in which people are readopting some of the 'peasant principles' to resist against the forms that agribusinesses have been developing. Repeasantization fights against the depredation of uncontrolled free enterprise as embodied in the neoliberal concept of the commodity-specialized profit-seeking commercial firms who accumulate profits. Many of the strategies are derived in interesting ways from classic peasant cultural, social, political, and environmental strategies adapted to modern contexts.

Peasant Economics

Not pursuing the profit motive in the way that an enterprise would, does not mean, as Mayer and Glave (Mayer and Glave, 1999) have shown, that peasant households do not seek to gain money from market transactions. When not required for farm work, a favorite activity is to engage as traders in peasant marketplaces to make money as *Penny Capitalists* (Tax, 1963). Mayer's *Articulated Peasant* (2002, 2005) shows that households are nodal points in extensive networks of reciprocal, barter, and cash transactions that extend outward from the village and region to capital cities within their countries, and since the 1980s, beyond Peru.

Off-farm peasant rural labor power concentrates in centers of casual employment, principally cities and extractive industries to supplement the farm base. The economist Alain de Janvri (1981) in *The Agrarian Question and Reformism in Latin America* calls this characteristic 'functional dualism' of Latin American peasantries in that it operates a subsistence farming sector and it is an active participant in the market, selling labor, handicrafts, trading, building, and investing. It is functional because the costs of creating the labor force and its partial maintenance are not borne by the employer but by the peasant/laborer. When this labor is urban based, it creates a dynamic that Bryan Roberts (1978) calls cities of peasants. Many aspects of peasant ways of life, such as foodways, cultural institutions, kinship patterns (Lobo, 1982), forms of mutual aid, reciprocity patterns (Lomnitz, 1977), and marketing strategies (Seligmann, 2004; Clark, 1994) are re-created in cities. Scholars like Alejandro Portes et al. (1989), Marie Estelie Smith (1989), and Hernando de Soto (1989) have labeled it the 'informal economy' and point out that informality is very dynamic and pervades ever larger in sectors of city neighborhoods, segments of the economy, and financial markets. The informal economy is an outgrowth of the peasant economy.

International labor migration without completely cutting ties to the rural home base entails, as Michael Kearney (1996) so nicely argues in *Reconceptualizing the Peasantry*, a triple cultural transformation. The *Mixe* Indian from the highlands of Oaxaca becomes a generic poor Mexican on reaching Tijuana on the United States–Mexican Border, and once crossed to the other side becomes a *Chicano* able to manage multilingual multicultural situations, a member of an underclass and an illegal who nonetheless works, lives, and saves, and, occasionally, even thrives. When he does so, he sends money back home. In 2007, remittances sent back home by migrating Latin Americans amounted to US $66.5 billion in more than 100 million small international transactions by about 10 million people (García Zamora and Orozco, 2009). That amount of money exceeds foreign investment and international development aid in many countries and, therefore, has become of crucial importance in the management of the flow of currencies, their national economies, and financial operations.

Therefore, contrary to previously held commonplaces, peasant societies are an intricate component of today's capitalisms and not a disappearing social segment that is held to be tradition bound that will not modernize according to the designs of the current neoliberal ideology (Brycerson et al., 2000).

See also: Development: Social-Anthropological Aspects; Irrigation Societies; Markets: Anthropological Aspects; Modes of Production.

Bibliography

Alcorn, J., 1984. Huastec Mayan Agriculture. University of Texas Press, Austin.
Arensberg, C., Kimball, S., 1968. Family and Community in Ireland. Harvard University Press, Cambridge.
Banfield, E., 1958. The Moral Basis of a Backward Society. The Free Press, New York.
Bebbington, A. (Ed.), 2011. Social Conflict, Economic Development and the Extractive Industry: Evidence from South America. Routledge, Milton Park.
Bernstein, H., 1994. Agrarian classes in capitalist development. In: Sklair, L. (Ed.), Capitalism and Development. Routledge, London, pp. 40–71.

Blanco, H., 1972. Land or Death: The Peasant Struggle in Peru. Pathfinder Press, New York.
Brush, S.B., 2004. Farmers' Bounty: Locating Crop Diversity in the Contemporary World. Yale University Press, New Haven.
Brycerson, D., et al. (Eds.), 2000. Disappearing Peasantries? Rural Labor in Africa, Asia and Latin America. Intermediate Technology Publications, London.
Clark, G., 1994. Onions Are My Husband: Survival and Accumulation by West African Market Women. University of Chicago Press, Chicago.
de Janvri, A., 1981. The Agrarian Question and Reformism in Latin America. The Johns Hopkins University Press, Baltimore.
de Soto, H., 1989. The Other Path: The Invisible Revolution in the Third World. Harper & Row, New York.
Escobar, A., 1995. Encountering Development: The Making and Unmaking of the Third World. Princeton University Press, Princeton.
García Zamora, R., Orozco, M., 2009. Migración Internacional, Remesas y Desarrollo Local en América Latina y el Caribe, Universidad Autónoma de Zacatecas, Inter-American Dialogue. Miguel Angel Porrúa, Mexico, DF.
Geertz, C., 1963. Agricultural Involution: The Process of Ecological Change in Indonesia. University of California Press, Berkeley.
Gudeman, S., Rivera, A., 1990. Conversations in Colombia: The Domestic Economy in Life and Text. Cambridge University Press, Cambridge.
Gudynas, E., 2011. Buen vivir: today's tomorrow. Development 54, 441–447.
Handelman, H., 1975. Struggle in the Andes: Peasant Political Mobilization in Peru. The University of Texas Press, Austin.
Harris, M., 2000. Life in the Amazon: The Anthropology of Brazilian Peasant Village. Oxford University Press, Oxford.
Kearney, M., 1996. Reconceptualizing the Peasantry: Anthropology in Global Perspective. Westview, Boulder.
Kenny, M., 1962. A Spanish Tapestry; Town and Country in Castile. Indiana University Press, Bloomington.
Kroeber, A., 1948. Anthropology: Race, Language, Culture, Psychology, Prehistory, new ed. Harcourt, New York.
Landsberger, H., 1969. Latin American Peasant Movements. Cornell University Press, Ithaca.
Lenin, V., 1982/1899. The differentiation of the peasantry. In: Harriss, J. (Ed.), Rural Development: Theories of Peasant Economy and Agrarian Change. Hutchinson, London.
Lewis, O., 1963. Life in a Mexican Village: Tepoztlán Re-studied. University of Illinois Press, Urbana.
Lobo, S., 1982. A House of My Own: Social Organization in the Squatter Settlements of Lima, Peru. University of Arizona Press, Tucson.
Lomnitz, L., 1977. Networks and Marginality: Life in a Mexican Shantytown. Academic Press, New York.
Marcos, Subcomandante, 1995. Shadows of Tender Fury: The Letters and Communiqués of Subcomandante Marcos and the Zapatista Army of National Liberation. Monthly Review Press, New York.
Marx, K., 1971. Peasantry as a class. In: Shanin, T. (Ed.), Peasants and Peasant Societies. Penguin, Harmnondsworth.
Mayer, E., 2002. The Articulated Peasant: Household Economies in the Andes. Westview Press, Boulder.
Mayer, E., 2005. Households and their markets. In: Carrier, J. (Ed.), Handbook of Economic Anthropology. Berg Publishers, Oxford, pp. 405–422.
Mayer, E., 2009. Ugly Stories of the Peruvian Agrarian Reform. Duke University Press, Durham.
Mayer, E., Glave, M., 1999. Alguito para ganar (a little something to earn): profits and losses in peasant economies. American Ethnologist 26, 344–369.
Menchú, R., 1964. I, Rigoberta Menchú: An Indian Woman in Guatemala. Verso, London.
Netting, R.McC., 1981. Balancing on an Alp: Ecological Change and Continuity in a Swiss Mountain Community. Cambridge University Press, Cambridge.
Netting, R.McC., 1993. Smallholders, Householders: Farm Families and the Ecology of Intensive, Sustainable Agriculture. Stanford University Press, Palo Alto.
Ostrom, E., 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, Cambridge.
Peristiany, J.G., 1966. Honour and Shame: The Values of Mediterranean Society. Weidenfeld and Nicolson, London.
Pitt-Rivers, J.A., 1966. The People of the Sierra. University of Chicago Press, Chicago.
Ploeg, J van der, 2008. The New Peasantries: Struggles for Autonomy and Sustainability. Earthscan, London.
Portes, A., Castells, M., Benton, L., 1989. The Informal Economy: Studies in Advances and Less Developed Countries. Johns Hopkins University Press, Baltimore.
Redfield, R., 1956. Peasant Society and Culture. University of Chicago Press, Chicago.
Redfield, R., 1973/1946. Tepotzlan, a Mexican Village: A Study of Folk Life. University of Chicago Press, Chicago.
Roberts, B.R., 1978. Cities of Peasants: The Political Economy of Urbanization in the Third World. Arnold, London.
Roseberry, W., 1983. Coffee and Capitalism in the Venezuelan Andes. Texas University Press, Austin.
Scott, J.C., 1976. The Moral Economy of the Peasant: Rebellion and Subsistence in Southeast Asia. Yale University Press, New Haven.
Scott, J.C., 1985. Weapons of the Weak: Everyday Forms of Peasant Resistance. Yale University Press, New Haven.
Seligmann, L., 2004. Peruvian Street Lives: Culture, Power, and Economy among Market Women in Cuzco. University of Illinois Press, Urbana.
Shanin, T. (Ed.), 1971. Peasants and Peasant Societies. Harmondsworth.
Shanin, T., 1973. The Awkward Class: Political Sociology of Peasantry in a Developing Society: Russia 1910–1925. Oxford University Press, London.
Smith, M.E., 1989. The informal economy. In: Plattner, S. (Ed.), Economic Anthropology. Stanford University Press, Stanford, pp. 253–317.
Tax, S., 1963. Penny Capitalism: A Guatamalan Indian Economy. University of Chicago Press, Chicago.
Wallerstein, I., 1974. The Modern World System: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century. Academic Press, New York.
Wolf, E.R., 1955. Types of Latin American peasantry: a preliminary discussion. American Anthropologist 57, 452–471.
Wolf, E.R., 1966. Peasants. Prentice Hall, Englewood Cliffs.
Wolf, E.R., 1973. Peasant Wars of the Twentieth Century. Harper and Row, New York.
Wolf, E.R., 1982. Europe and the People without History. University of California Press, Berkeley.
Wolf, E.R., Mintz, S., 1955. Haciendas and plantations in Middle America and the Antilles. Social and Economic Studies VI (33), 380–412.
Zamosc, L., et al., 1997. Estructuras Agrarias y Movimientos Campesinos en América Latina (1950–1990). Ministerio de Agricultura, Pesca y Alimentación, Madrid.

Relevant Website

http://www.viacampesina.org (accessed 14.06.13.) – La Via Campesina.

Peer Review and Quality Control

Silvio O Funtowicz, Centre for the Study of the Sciences and the Humanities (SVT), University of Bergen, Bergen, Norway
Jerome R Ravetz, Institute for Science, Innovation and Society, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Peer review is the recognized standard for quality control in science. There is a growing awareness about the need to introduce modifications to the traditional peer review process in order to meet the challenges of a continuously expanding scientific enterprise, the emergence of new forms of science, and the impact of new technologies. The growing application of science and science-based technologies to society requires innovative quality control mechanisms and the inclusion of new stakeholders. Collegial peer review is being rapidly transformed to review by an 'extended peer community,' raising important issues to the governance of science.

Background

Peer review is a mechanism for quality control in science, including the assessment of proposed projects and of completed work. Its roots go back to the emerging science of seventeenth century, when novelty in natural knowledge became distinguished from technical invention, and the ideals of reproducibility and publicity became established. Peer review was acknowledged by the mid-twentieth century as the unquestioned norm in academic science.

Recent experience has shown the need for important modifications in the process of quality assurance, such as changes to the traditional, largely informal, procedures of collegial peer review, in order to take into account the emergence of new forms of science, the increased competitiveness of the research enterprise, the impact of new technologies, and the inclusion of new stakeholders. Collegial peer review is being rapidly transformed to review by an 'extended peer community.'

Among contemporary forms of scientific practice, curiosity-driven research with no regard for applications has been increasingly marginalized. A diversification has occurred so that quality assurance must also be considered in such areas as mission-oriented and issue-driven research, technoscience and innovation research, and the provision of scientific advice for policy.

As the unquestioned norm in academic science, it was assumed that quality was unproblematic, owing to some special ethical disposition of scientists and their methods. With subsequent transformations in the social practice, embeddedness, and pervasiveness of science in society – which now extends into the economy, technology, policy, and life forms and styles – corresponding changes in quality assurance, are urgently needed. The crucial requirement is for a more aware understanding of the encompassing community of those producing, evaluating, sharing, profiting from, and even 'suffering' the products and processes of science (see, e.g., Stampa, 1997).

In most other spheres of institutional activity, the formalization of quality assurance has become the norm, as, for example, through the wide-ranging standard-setting procedures of the International Standards Organization. In academic science, however, along with cultural pursuits like the arts, the methods have been and are, still, largely informal. Science has been almost unique in having self-assessment performed by accredited practitioners rather than by users or external critics. To what extent and in what ways this must change to keep pace with science's expanding role in public and private life has become an urgent question in the conduct of science, in the governance of research and in the coevolution of science and society.

The assurance of quality is not a straightforward task. This has been known since the time of the Romans, as indicated by the Latin motto *quis custodiet ipsos custodes?* (Who will guard the guardians themselves?). This motto implies an indefinite iteration. It is a reminder that, however, routine may be the tasks of quality control, full quality assurance demands yet higher levels of supervision at which informality, extended participation, and explicit value judgments are necessary (Funtowicz and Ravetz, 1990).

These insights were introduced in industrial quality control by the pioneering work by William Edwards Deming at the end of the 1940s. His concepts of profound knowledge and total quality assurance were successfully applied by Japanese car manufacturers before being imported back, much later, to the US. Edwards Deming's ideas were quite easily absorbed in the Japanese cultural landscape. These ideas included the establishment of quality circles, in which the assembly line was transformed into a kind of participatory process. Two aspects of this process were quite revolutionary: the first was the extension of the community of quality controllers and evaluators to the work force, relying on practical knowledge, experience, and commitment; the second was the encouragement of whistle blowing, whereby any member of the community, including the work force, was entitled to stop the process if they believed that quality was violated. On Edwards Deming's philosophy of quality, see *Out of the Crisis* (Deming, 1986).

As long as science remained mainly academic, problems of quality were assumed to be resolved by self-correction, anticipating or responding to peer criticism; a process conceived as part of the very nature of the scientific endeavor. The informal systems of checking by peers seemed a rational and efficient response to the problem, rather than a culturally contingent mechanism characteristic of a particular epoch. Scientific facts

 http://dx.doi.org/10.1016/B978-0-08-097086-8.85016-3

were believed to be discovered by a reliable method, and scientists themselves were viewed as being endowed with certain superior moral qualities that protected them and their work from ordinary human failure or error. This self-correcting property of science could be explained in sociological terms, as in the four norms of scientific institutional practice expounded by Robert K. Merton in 1942 (Merton, 1973[1942]): Communalism, Universalism, Disinterestedness, and Organized Scepticism defined the ethos of modern science. Philosophically, it could be justified in the committed, determined attempts at self-refutation supposed by Karl Popper to be normal scientific practice (Popper, 1959), where the refutation potential becomes the hallmark of a scientific hypothesis.

Science as the Endless Frontier

With the onset of the industrialization of science after World War II, the self-conscious study of science as a social activity, including the methods of quality assurance, became inevitable. With the adoption of Vannevar Bush's report, Science the Endless Frontier (Bush, 1945), the stakes of the scientific enterprise became higher, inviting scrutiny of the quality of processes and products. Growth in size, capital investment, scale, and social differentiation within science created divisions between managers and research workers, as well as between researchers and teachers in universities. A *Gemeinschaft* (community) of scholars could no longer realistically be assumed. The earliest disciplined analyses of the quality of production in science were quantitative. Derek J. de Solla Price, who devised some measures of quality and provided analyses of its distribution, did the pioneering work. He noticed that at the leading British science reference library, only a minority of journals was ever requested. The contents of the others could be inferred to have no interest, and hence to be of very low scientific quality (Price de Solla, 1963). This phenomenon is a reminder that 'quality' is a relational attribute, and its definition 'fitness for purpose or function' depends on whose purposes are dominant; not always perhaps those of a community devoted to the advancement of learning and achievement of reliable knowledge, but possibly only of those scientists working under constraints of 'publish or perish.'

Price's studies were continued in two directions. At the Institute for Scientific Information, Eugene Garfield produced more searching and sophisticated measures of quality, using citations rather than mere number of publications. Such attempts at quantification were bound to become controversial (Brooks, 1982; Garfield, 1970, 1972). It was impossible to avoid bias in the selection of the relatively small set of journals used for citations; those in conventional mainstream English-language research science were inevitably privileged at the expense of all others. Further, when quantitative measures of citations came to be used as indicators of academic merit, manipulative practices, including reciprocal citations, inevitably developed. The deep problems of a quantitative proxy for quality suddenly became acute.

In a more reflective vein, Jerome R. Ravetz applied the *quis custodiet* principle to analyze the vulnerability of the quality assurance system in science. He observed that the processes of peer review are conducted largely informally and (unlike research) are not themselves normally submitted to open scrutiny and review. They require different competences, which are not part of the formal training of scientists; and they also are more open to a variety of abuses, ranging from bias to plagiarism. One can understand in these terms the phenomena of low quality, both in scientific research and in technological development. Thus, while denying that the practice of science automatically produces a higher morality, Ravetz agrees that moral standards are necessary for the successful practice of science. On this basis, he stresses the importance of morale and morality (and ultimately idealism and leadership) in science (Ravetz, 1996[1971]).

This analysis offers a background for the increasing interest in trust as an essential element of practice in science, in society, and in their interactions. The broader society has provided resources to the esoteric activities of science because it trusts the scientific community to make good use of them. There has always been an undercurrent of distrust, based on evidence either of pointless research or of malign applications. Once science became deeply involved in technology and related policy problems that crucially affect public health and welfare, the traditional relations of trust could no longer be assumed. It appeared to be necessary for the principles and practices of accountability to be extended from the institutions of political and economic governance (as, e.g., representative democracy) to those institutions that govern science and its applications.

Quality control in research science has become more difficult as the relatively inflexible technical requirements of the traditional printing process have first been relaxed, and then made obsolete by novel information and communication technologies and economic restructuring. There no longer is a well-defined gateway to publication through the institutions that control reproduction of, and hence access to, documents. First through inexpensive photocopying and now through the Internet, it has become easy for anyone to distribute knowledge wares to an unrestricted audience. It is like a reenactment of the period that followed the invention of the printing press and the publication of Gutenberg's Bible around 1450. Now, as then, the monopoly of knowledge dissemination is collapsing and new and expanded interested audiences are being created.

In addition, the presence of the global media tends to bypass the traditional processes of evaluation that were conducted personally among colleagues. The demystification of scientific practice enables such events to become news, and, at the same time, is fostered by their being exposed. Top scientists become like media celebrities – needing the media for advertising themselves yet simultaneously hating them for their unwanted intrusions.

Isolated scientific results can and have become media events (Close, 1991). Seralini and colleagues' rat study, relating cancer to genetically modified maize published in a peer-reviewed journal (Seralini et al., 2012), created not only great controversy among scientists (Butler, 2012) and stakeholders (see, for example, http://www.euractiv.com/innovation-enterprise/commission-science-supremo-endor-news-514072), but a new form of science-related reality TV show as Russian scientists planned to reproduce and broadcast the genetically modified organism (GMO) experiment. The scientist in charge of the project told the news agency: "This is a unique

experiment …. There hasn't been anything like this before – open, public research by opponents and supporters of GMO" (RT.com, 2012).

All those with an interest in the outcome, such as researchers, consumers, politicians, regulators, and the stock market, become potential stakeholders in the evaluation and promotion of the results. Thus, science arguably becomes accountable to a drastically extended peer community in the quality assurance process. The criteria of quality applied by these heterogeneous actors need not be identical to those of public knowledge generated within the tightly knit scientific networks of academic science.

Pathologies of Science in Society

These developments may be judged in different ways. While they may seriously disrupt the procedures of quality assurance in normal science, they can also bring needed public scrutiny to bear on innovation, controversies and, even, scandals. In 1999, the Netherlands Environmental Assessment Agency was accused by an employee, senior statistician Hans de Kwaadsteniet, of lying and deceiving; his claim was that figures released by the agency, based on virtual reality computer models were essentially unreliable. The whole episode, which lasted several months, ended with the transformation from a technocratic model of quality assurance to a more accountable, transparent, and participatory paradigm (Petersen et al., 2011).

The 'Baltimore affair,' centering on the US Nobel laureate David Baltimore's laboratory at MIT, was perhaps in its time the most notorious case in which a dispute about scientific misconduct was blown up into a lengthy, visible, political saga that unquestionably damaged all the individuals and institutions involved (Kevles, 1998). The episode was symptomatic of an increasingly recognized problem of 'deviance' in science, which carries the unspoken danger that, without timely correctives, misconduct might become the norm.

There has been a growing awareness and concern, expressed in studies and articles not only in the scientific press (Jha, 2012; Zimmer, 2012), about the quantity and quality of fraud and misconduct cases (Fanelli, 2009; Corbin, 2012). Several initiatives have been proposed to tackle misconduct and related issues, such as retractions (see, for example http://retractionwatch.wordpress.com) and the reproducibility (see, for example, http://www.reproducibilityinitiative.org) of research results (Nature Biotechnology, 2012). The problem does not affect a single discipline but it is felt across a wide range of research areas; thus, for example, the Nobel Laureate Daniel Kahneman, has challenged psychologists to "clean up their act," according to Nature (Yong, 2012).

All these developments affect the maintenance of trust, which is necessary for ordinary scientific practice and even more so for quality assurance. As in other professional domains, the normal tendency in science has been for misconduct to be covered up by the responsible institution (not necessarily by the community of scientists). In such situations, ultimate exposure does even more damage and further erodes the basis for mutual trust. Grigory Perelman, perhaps the most gifted mathematician of his generation, is quoted to have said, after refusing the European Mathematical Society top prize in 2007 that the award committee was unqualified to assess his work (Paulos, 2010). Perelman's mistrust toward his colleagues and the organized mathematical community is also illustrated by his decision to publish his proofs of the geometrization conjecture in the nonpeer reviewed archive for electronic preprints, arXiv. Nasar and Gruber report that Perelman had not mentioned the proof or shown it to anyone. "I didn't have any friends with whom I could discuss this," he said in St. Petersburg. "I didn't want to discuss my work with someone I didn't trust" (Nasar and Gruber, 2006).

Attempts to circumvent the need for trust by increasing bureaucratic surveillance are likely to be counterproductive in their own way, by erecting impediments to free inquiry and communication among colleagues.

The relations between social science and natural science have also been transformed during the last decades, with implications for quality control. Starting with the acceptance of natural science as the ideal of knowledge, essentially independent of social forces, there has been a gradual but accelerating shift toward recognizing all sciences as incorporating social constraints and biases. An early critical interaction was in connection with the astronomical community's management of the eccentric Velikovsky (De Grazia, 1966). Later, the social science community embraced Thomas Kuhn's disenchanted picture of 'normal science' (Kuhn, 1970). Finally, post-Feyerabend studies of science reexamined the whole institution of scientific inquiry without presupposing any privileged status in relation to either virtue or natural knowledge (Bloor, 1991; Barnes et al., 1995; Collins and Pinch, 1993; Fuller, 1993).

When natural scientists, led by physicists, eventually confronted the emerging socialized picture of their discipline, the reaction was so strident that 'science wars' became an appropriate label (Gross et al., 1997; Nelkin, 1996; Ross, 1996). Sociologists of science and postmodernists were indiscriminately blamed for all the ills of science, including decline of public trust, budget cuts, resurgent Creationism, and even poor teaching of science. A physicist whose hoax article (Sokal, 1996) was accepted by a leading cultural studies journal, *Social Text*, helped to crystallize the attack (Bricmont and Sokal, 1998). The implication was that the critics of science had no real quality control of their productions. The science warriors' assumption was that within real science, such problems are prevented from occurring because of the verifiable empirical content of scientific research. In the ensuing debate, there was little mention of the ease of publication of erroneous or vacuous research in the standard scientific literature. Historical episodes, like Millikan's manipulation of his oil-drop results in the course of a controversy on the charge of the electron, were discounted as mere embarrassments (Segerstale, 1995).

It has been presupposed thus far that 'science' refers primarily to traditional basic research. However, among contemporary forms of scientific practice, curiosity-driven research with no regard for applications is giving way to new forms of science. Quality assurance must also be considered and performed in such areas as mission-oriented, issue-driven and innovation research, forensic science (Foster and Huber, 1997; Jasanoff, 1995), and the provision of scientific advice for policy (Jasanoff, 1990; Salter, 1988). In addition, the products themselves and the media through which they are diffused increasingly are diversified. For example, patents

are now a common outcome of a research process, and this form of intellectual property is radically different from traditional published papers (Myers, 1995). Also, results are reported in nonpeer reviewed electronic Web sites, unpublished consultancy reports and gray literature, or kept confidential within institutions or even totally sealed under lawyer–client confidentiality and legal settlement agreements.

Peer Review Extended and Democratized

With traditional peer review as the norm, the challenges of quality assurance for these new products and processes are nearly unrecognizable. A genre of critical literature has developed, with some authors directing anger at the new contexts of scientific production (Huber, 1991), and others more clearly appreciating the problems they present (Crossen, 1994; Jasanoff, 1990, 1995).

A parallel diversification has occurred in the types of knowledge production that are accepted as legitimate. The democratization of knowledge now extends beyond the juries who assess the quality of technical evidence in courts (Jasanoff, 1998) to include those who master previously esoteric aspects of their predicament (e.g., illness, contamination, pollution, oppression, discrimination, exploitation) through special-interest groups or Internet-based social media. In addition, claims of specialized or local knowledge are present in even more diverse contexts, as among indigenous peoples, and in systems of complementary or 'traditional' medicine. These claims are commanding increasing commercial and political support among various publics, as well as gaining explicit recognition in numerous international treaty regimes. As a result, a new philosophy of knowledge appears to be emerging, based on a new disciplined awareness of complexity, in which a plurality of legitimate perspectives is taken for granted (Funtowicz and Ravetz, 1992). Modern science, with its characteristic methodology and social location, is part of this enriched whole, but not coextensive with it. The criteria and tasks of quality assurance must explicitly involve additional values and interests, incorporating even the ontological commitments of groups other than scientists. This new configuration has been termed postnormal science.

There are now many initiatives, increasing in number and significance all the time, for involving wider circles of people in science, decision making and implementation on policy issues. The contribution of all the stakeholders in cases of postnormal science is not merely a matter of broader democratic participation. Novel and emerging technoscientific issues, promoted and produced as innovation research, are in many ways different from those of traditional research science, professional practice, or industrial development. Each of those has its established means for quality assurance of the products of the work, whether through peer review, professional associations, or the market. However, for these new forms of knowledge production, the maintenance of quality depends on open dialogue between all those affected. In postnormal science, it is called an 'extended peer community,' consisting not merely of persons with some form or other of institutional accreditation, but rather of all those with a desire to participate in the resolution of the issue. Since this context of science involves politics and policy, we might see this extension of peer communities as analogous to earlier extensions of the franchise in other fields, such as women's suffrage and patients' rights.

Extended peer communities are already being created, in increasing numbers, either when the authorities cannot see a way forward, or when they know that without a broad base of consensus, no policy can succeed. They are called citizens' juries, focus groups, consensus conferences, or any one of a great variety of other names; and their forms and powers are correspondingly varied. But they all have one important element in common: they assess the quality of policy proposals, including a scientific element, on the basis of the science they master combined with their commitment and the knowledge of the ways of the world. In addition, their verdicts all have some degree of moral force and hence political influence.

These extended peer communities will not necessarily be passive recipients of the materials provided by experts. They will also possess, or create, their own 'extended facts.' These may include craft wisdom and community knowledge of places and their histories, as well as anecdotal evidence, neighborhood surveys, investigative journalism, and leaked documents. Such extended peer communities have achieved enormous new scope and power through the Internet. Activists scattered among large cities or rain forests can engage in mutual learning and coordinated activity, providing themselves with the means of engagement with global vested interests on less unequal terms than previously. Along with the regulatory, evaluative function of extended peer communities, another, even more intimately involved in the policy process, is springing up. Particularly at the local level, the discovery is being made, again and again, that people not only care about their own environment and health but can also become quite ingenious and creative in finding practical, mixed social, and technological means for its improvement; and local people can imagine solutions and reformulate problems in ways that the accredited experts, with the best will in the world, do not find 'normal.' This is most important in the phases of policy formation, and also in the implementation and monitoring of policies. Thus in addition to extending the traditional processes of quality assessment, participants can enhance the quality of the problem-solving processes themselves.

Thanks in large part to the new powers created by information and communication technologies and the new social media, the extended peer community can on occasion inject itself into quality debates of a traditional sort. A notable instance of this tendency is the activity of the so-called deniers in the debate on global warming science. Some have focused on the evidence adduced for the phenomenon of global warming, independently of whether it is actually happening. The most famous of these is Steve McIntyre, a retired mining engineer, who has criticized the tree-ring analysis conducted by Michael Mann and others. Part of his criticism was about the persistent refusal of Mann to supply him with his data, which were obtained on public funds and so would be presumed to be in the public domain. His more serious criticism related to the statistical techniques underlying the 'hockey stick' picture of global temperatures in recent centuries. Although his arguments were never formally recognized by the academic scientists he criticized, he did gain important respect from key science journalists (Harrabin, 2011). McIntyre has his own site

in the 'blogosphere,' but he and others became a powerful presence through popular Web sites like "Watts Up With That?" (http://www.wattsupwiththat.com, accessed 10.09.13.), which have attracted millions of visits and also won prizes for quality. In this way, the new social media have given strength to the extended peer community in science in a way reminiscent of the contribution of printing to the Reformation.

Quality assurance, thus, can be seen as a core commitment of postnormal science, replacing 'truth' as science's ultimate regulative principle (Funtowicz and Ravetz, 1993). Defined in terms of uncertainties and decision stakes, quality assurance encompasses public interest, citizen, and vernacular sciences. In a period of domination by globalized corporate science (Gibbons et al., 1994), this effort to make scientists accountable to interested groups presents a coherent conceptual alternative for the survival of the public knowledge tradition of science. Collegial peer review is thereby transformed into review by an extended peer community. This new form of quality assurance will be given its institutional structure and routines by those heterogeneous actors who put it into practice.

See also: Citizen Science; Ethical Practices: Institutional Oversight and Enforcement; Intellectual Property: Law in Context; Norms in Science; Patenting in the Biological Sciences; Science Communication; Science and Law; Science and Technology Studies: Experts and Expertise; Science, New Forms of; Science, Public Engagement with; Scientometrics; Social Work Ethics; Trust, Political; Trust, Sociology of; Universities and Science and Technology: United States.

Bibliography

Barnes, B., Bloor, D., Henry, J., 1995. Scientific Knowledge: A Sociological Analysis. University of Chicago Press, Chicago.

Bloor, D., 1991. Knowledge and Social Imagery. University of Chicago Press, Chicago.

Bricmont, J., Sokal, A.D., 1998. Fashionable Nonsense: Post-modern Intellectuals' Abuse of Science. Picador, New York.

Brooks, H., 1982. Science indicators and science priorities. In: La Follette, M.C. (Ed.), Quality in Science. MIT Press, Cambridge, MA, pp. 1–32.

Bush, V., 1945. Science, the Endless Frontier: A Report to the President. US Government Printing Office, Washington, DC.

Butler, D., 2012. Rat study sparks GM furore. Nature 489 (7417), 484.

Close, F.H., 1991. Too Hot to Handle: The Race for Cold Fusion. Princeton University Press, Princeton, NJ.

Collins, H., Pinch, T., 1993. The Golem: What Everyone Should Know about Science. Cambridge University Press, Cambridge.

Corbin, Z., 2012. Misconduct is the main cause of life-sciences retractions. Nature 490, 21.

Crossen, C., 1994. Tainted Truth: The Manipulation of Fact in America. Simon & Schuster, New York.

De Grazia, A. (Ed.), 1966. The Velikovsky Affair: The Warfare of Science and Scientism. University Books, New York.

Deming, W.E., 1986. Out of the Crisis. MIT Press, Cambridge, MA.

Fanelli, D., 2009. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS ONE 4 (5), e5738. http://dx.doi.org/10371/journal.pone.0005738.

Foster, K.R., Huber, P.W., 1997. Judging Science: Scientific Knowledge and the Federal Courts. MIT Press, Cambridge, MA.

Fuller, S., 1993. Philosophy, Rhetoric and the End of Knowledge: The Coming of Science and Technology Studies. University of Wisconsin Press, Madison.

Funtowicz, S.O., Ravetz, J.R., 1990. Uncertainty and Quality in Science for Policy. Springer, Berlin.

Funtowicz, S.O., Ravetz, J.R., 1992. Three types of risk assessment and the emergence of post-normal science. In: Krimsky, S., Golding, E. (Eds.), Social Theories of Risk. Greenwood Press, Westport, CT, pp. 251–273.

Funtowicz, S.O., Ravetz, J.R., 1993. Science for the post normal age. Futures 25 (7), 739–755.

Garfield, E., 1970. Citation indexing for studying science. Nature 227, 669–671.

Garfield, E., 1972. Citation analysis as a tool in journal evaluation. Science 178, 471–479.

Gibbons, M.C., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., Trow, M., 1994. The New Production of Knowledge. Sage, Beverly Hills, CA.

Gross, P.R., Levitt, N., Lewis, M.W. (Eds.), 1997. The Flight from Science and Reason. Johns Hopkins University Press, Baltimore, MD.

Harrabin, R., March 19, 2011. Harrabin's Notes: Getting the Message. BBC News. http://www.bbc.co.uk/news/10178454 (accessed 27.11.12.).

Huber, P.W., 1991. Galileo's Revenge: Junk Science in the Courtroom. Basic Books, New York.

Jasanoff, S., 1990. The Fifth Branch: Science Advisors as Policy as Policymakers. Harvard University Press, Cambridge, MA.

Jasanoff, S., 1995. Science at the Bar: Law, Science and Technology in America. Harvard University Press, Cambridge, MA.

Jasanoff, S., 1998. The eye of everyman: witnessing DNA in the Simpson trial. Social Studies of Science 28 (5/6), 713–740.

Jha, A., 13 September 2012. False Positives: Fraud and Misconduct Are Threatening Scientific Research. The Guardian. http://www.guardian.co.uk/science/2012/sep/13/scientific-research-fraud-bad-practice (accessed 27.11.12.).

Kevles, D.J., 1998. The Baltimore Case: A Trial of Politics, Science and Character. Norton, New York.

Kuhn, T.S., 1970. The Structure of Scientific Revolutions, second ed. University of Chicago Press, Chicago.

Merton, R.K., 1973[1942]. The normative structure of Science. In: Storer, N.W. (Ed.), The Sociology of Science. The University of Chicago Press, Chicago, pp. 267–278 (Originally: Merton, R.K., 1942. A note on science and democracy. Journal of Legal and Political Sociology 1, 115–126.).

Myers, G., 1995. From discovery to invention: the writing and rewriting of two patents. Social Studies of Science 25 (1), 57–105.

Nasar, S., Gruber, D., 28 August 2006. Manifold Destiny. A Legendary Problem and the Battle over Who Solved It. The New Yorker. http://www.newyorker.com/archive/2006/08/28/060828fa_fact2?currentPage=all (accessed 27.11.12.).

Nature biotechnology. Further Confirmation Needed, Editorial, Nature Biotechnology 30, 2012, 806.

Nelkin, D., 26 July 1996. What are the science wars really about? The Chronicle of Higher Education, A52.

Paulos, J.A., 29 April 2010. He Conquered the Conjecture. The New York Review of Books. http://www.nybooks.com/articles/archives/2010/apr/29/he-conquered-the-conjecture/?pagination=false (accessed 27.11.12.).

Petersen, A.C., Cath, A., Hage, M., Kunseler, E., van der Sluijs, J., 2011. Post-normal science in practice at the Netherlands Environmental Assessment Agency. Science, Technology, & Human Values 36 (3), 362–388.

Popper, K., 1959. The Logic of Scientific Discovery. Basic Books, New York.

Price de Solla, D.J., 1963. Little Science, Big Science. Cambridge University Press, Cambridge.

Ravetz, J.R., 1996[1971]. Scientific Knowledge and its Social Problems. Transaction Publishers, New Brunswick, NJ (Clarendon Press, Oxford, 1971).

Ross, A. (Ed.), 1996. Science Wars. Duke University Press, Durham, NC.

RT.com, 29 September 2012. Rat Reality Show: Russian Scientists to Broadcast GMO Experiment. http://rt.com/news/gmo-experiment-online-rats-240/ (accessed 27.11.12.).

Salter, L., 1988. Mandated Science. Kluwer, Dordrecht.

Segerstale, U., 1995. Good to the last drop? Millikan stories as "canned" pedagogy. Science and Engineering Ethics 1, 197–214.

Seralini, G.-E., Clair, E., Mesnage, R., Gress, S., Defarge, N., Malatesta, M., Hennequin, D., Spiroux de Vendomois, J., 2012. Long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize. Food and Chemical Toxicology 50 (11), 4221–4231.

Sokal, A.D., 1996. Transgressing the boundaries. Social Text 14, 217–252.

Stampa III, A.E., 1997. Advances in peer review research. Science and Engineering Ethics 3 (1), 1–104.

Yong, E., 3 October 2012. Nobel Laureate Challenges Psychologists to Clean up Their Act. Nature. http://www.nature.com/news/nobel-laureate-challenges-psychologists-to-clean-up-their-act-1.11535 (accessed 27.11.12.).

Zimmer, C., 16 April 2012. A Sharp Rise in Retractions Prompts Calls for Reform. The New York Times. http://www.nytimes.com/2012/04/17/science/rise-in-scientific-journal-retractions-prompts-calls-for-reform.html?pagewanted=all&_r=0 (accessed 27.11.12.).

Peer Review: Organized Skepticism

Katri Huutoniemi, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article approaches the evolution of peer review from a social scientific perspective, and looks at the various modes of the practice as instantiations of the norm of organized skepticism. Coined by Merton, organized skepticism is both the methodological and institutional mandate for scientific knowledge production: it guarantees good quality while preserving the social contract between science and society. Recent institutional changes in knowledge production and evaluation, however, imply an increasing ambiguity in the purposes and beneficiaries of peer review, and raise a question of whether and how it still grants intellectual authority.

Introduction

Science produces knowledge that is based on evidence, and it is thought to work best in conditions of unrestrained scrutiny. Epistemic scrutiny is, indeed, regarded as one of the glories of science as a systematic pursuit of knowledge. Peer review is the most wide-ranging institutional mechanism for realizing this scrutiny. Peer review is in some sense the *sine qua non* of the academy, but there exists little research on its historical development within the social and behavioral sciences, specifically. However, it appears that peer review developed in this field in ways very similar to how it developed in natural sciences and medicine, even if the origins of peer review can be traced earlier than the institutionalization of social science disciplines themselves.

I approach the question of peer review in historical and social scientific terms, and address questions raised about peer review by social scientists. I begin in the most obvious place, with Robert Merton's norm of organized skepticism, and look at the evolution of peer review through that lens. By drawing on existing historical accounts (e.g., Abbott, 1999; Biagioli, 2002; Burnham, 1990; Kronick, 1990; Rennie, 2003; Spier, 2002), I sketch out how the different modes and practices of peer review have (or have not) exhibited this ideal. I then discuss the changing contexts of peer review, which imply an increasing ambiguity in the purposes and beneficiaries of the practice. Today the social mechanisms that authorize and channel peer review are shaped as much by considerations of public policy as by the technical criteria of the scientific community (Chubin and Hackett, 1990). In conclusion, I identify some unresolved questions about peer review, among them the validity of the norm of organized skepticism itself. Throughout the article, I touch upon key conceptual issues, including the definition and scope of 'peer,' the range of peer review from *ex post* to *ex ante* forms, the tension between anonymity and openness, as well as the alleged conformism of the practice.

Peer Review as Organized Skepticism

'Organized skepticism' was coined by Robert K. Merton, America's pioneering sociologist of science, in 1942. It is the shared intellectual attitude among professional scientists that ensures detached scrutiny of beliefs and claims in terms of objective criteria. Although Merton's account arose out of an idealized view of the social relations characterizing the scientific enterprise, rather than of empirical and historical analysis of scientists' actual practices, it provides a viewpoint from which to look at the institution of peer review. Indeed, even in criticizing and expanding Merton's account of scientific norms, practicing scientist John Ziman (2000: 246) reinforces the importance of the norm of organized skepticism: "The active, systematic exercise of this norm by individual researchers is what, above all, makes science a communal enterprise. 'Peer review' is the key institution of the scientific culture." To put the point slightly differently, peer review is not only the institutional expression of the scientific norm of organized skepticism, but also a defining characteristic of what makes something 'really' scientific.

Peer review is cast in two crucial and related roles: it is assumed to guarantee good science while preserving the social contract between science and society. These roles were also identified by Merton in his original definition of organized skepticism as "both a methodological and institutional mandate" for scientific knowledge production (Merton, 1973 [1942]: 277). Merton framed this ideal as one of the four academic norms that together constitute what he called the 'ethos of science,' i.e., the institutional logic that scientific communities use for distinguishing appropriate and inappropriate values, beliefs, attitudes, and behaviors. The other three norms include (intellectual) communism, universalism, and disinterestedness.

The methodological mandate of organized skepticism is to provide an epistemic safeguard for reliable knowledge. As an opposite of dogmatism, skepticism means that all knowledge claims must be open to critical scrutiny by means of further research. Many philosophers of science (most notably, perhaps, Karl Popper, 1959) have emphasized the vital role of criticism in the advancement of learning and knowledge acquisition. This principle might be dated back to the Socratic method of asking and answering questions to stimulate critical thinking. In most general sense, peer review may we viewed as an institutionalization of this method into the conduct of science. It protects our knowledge base from contamination by error and poor argument and affords us contributions to attaining truth (Shatz, 2004).

The institutional mandate of organized skepticism is to legitimate scientific knowledge in society. It is the principle

through which science regulates itself in order to construct value, and is put into practice by a series of rational judgments and decisions. For example, Kathleen Fitzpatrick (2011) argues that the primary purpose that peer review serves today is not quality control per se, but that of institutional warranting, of conveying to e.g., university administrators that the work of their employees is appropriate and well-regarded in its field. The history of peer review can be understood as scientists' various attempts to discipline themselves according to the tacit social contract between science and society. Internalized by academics across intellectual traditions, ideals of democracy and accountability may provide pragmatic rules for peer review even in the absence of shared epistemological standards (Lamont, 2009).

Peer review unites the two mandates. As an instrument of knowledge governance in the 'republic of science' (Polanyi, 1962), peer review is claimed to be analogous to deliberative democracy. Many journal editors, for example, trying to justify the complex system, regularly make this analogy (Rennie, 2003). But it also underlies sociological and philosophical ideals of organized skepticism. If Popper's philosophy of science is read alongside his political philosophy, it is easy to see that openness to criticism is essential to democratic politics (Fuller, 2003). Similarly, Merton's ethos of science can be portrayed as an antidote to the political and ideological tutelage prevailing in the East (Elzinga, 2012).

Institutionalization of Peer Review

The Role of Royal Academies

The origin of peer review is traced to the royal academies in London and Paris in the seventeenth century. Members of the academies circulated letters among themselves or read papers in society meetings, reporting the results of their investigations with the explicit intention of eliciting response. It would be difficult to draw the line between authoring and reviewing in the context of early practices of the academies. Peer review thus emerged as a distinctive procedure only with the introduction of individual authorship, and its first instantiation was book review.

According to Mario Biagioli (2002), peer review of books developed as the result of royal privileges attributed to the very few academics to become part and parcel of the book licensing and censorship systems in the eighteenth century. It seems that the first purpose of review was to filter out books whose topic fell outside of the licensing jurisdiction of the royal academies – one that was limited to 'natural philosophy.' Skepticism was thus targeted to demarcate between what was found to be 'suitable' and what was not; this practice had little to do with quality in an epistemological or methodological sense. Biagioli's analysis shows that early peer review was anything but 'blind,' and it was precisely by being tied to specific authoritative institutions and their memberships that peer review could emerge to begin with.

Beginning in the seventeenth century, books began to be systematically reviewed also after their publication. Book reviews were published in classic 'republic of letters' periodicals like the *Journal de Savants* (established in 1665) with a view to providing the audience information about their contents and quality. These reviews had similarities to the royal 'imprimaturs' included in licensed books – they were often written by people who also worked as the royal censors – but they had clearly different status. While both kinds of reviews commented on the style, relative novelty, relevance, and other substantive qualities of the book (Biagioli, 2002), only book reviews in periodicals served organized skepticism in the sense of constructive criticism, not censorship.

The transition of peer review from an in-house disciplinary technique for specific academies to the evaluation of the work of nonmembers was connected to the development of academic journals, which brought about important changes in the role of peer review and the very meaning of 'peer.' The publishing protocols of the eighteenth century republic of letters were the first step toward redefining 'peer' in nonlocal terms, as well as toward establishing a broad scientific publication system hinged on review by academic peers. However, the republic of letters was still far from being an open system; rather, it was understood as a community having a more or less circumscribed membership (Biagioli, 2002; Broman, 2000).

The Rise of Academic Publishing

The first academic journal, *Philosophical Transactions* (established by the Royal Society of London in 1665), did not use what today would be considered proper review practices for the first 100 years. Only from 1752 were materials sent for publication subjected to inspection to a select group of members who were knowledgeable in such matters. The *Royal Society of Edinburg* seemed to have a formalized, anonymous peer review system in its *Medical Essays and Observations* starting from 1731 (Kronick, 1990). Editorial peer reviewing, however, did not evolve as a discernible movement, but as individual editors' attempts to adapt to changes in their professional environments. Historical accounts of medical publishing indicate that early peer review was meant to augment editorial expertise rather than to exercise more conventionally understood modes of quality control (Burnham, 1990). For a long time, the urgent need to fill journal pages discouraged editors from adopting any rigorous process of weeding out poor material (Abbott, 1999; Rennie, 2003; Spier, 2002).

At first, the system of refereeing presupposed a network of personal connections. The key to this system was the editor's trust in the referees, since judgments were accepted with minimal evidence (Abbott, 1999). However, it was appreciated from the start that the peer review process could not authenticate knowledge because the reviewers were not in a position to verify the claims of authors (Rennie, 2003). For authors, the newly instituted journal provided an opportunity of having competent appraisals of their work by other authoritative scientists. According to Harriet Zuckerman and Robert Merton (1971; see also Whitley, 1984), this concern of scientists with having recognition of their work by peers is central to the workings of science as a social institution. Although it no longer licenses texts, peer review has maintained the aura of the imprimatur, though one that is now about technical accuracy, not legal approbation (Biagioli, 2002).

The necessity for rigorous review has become greater as journals have had to compete on the apparent quality of their content. High rejection rate has become an indicator of

journal's quality, and journals with highest rejection rates are those that set the standards (Rennie, 2003). Even in humanities, 90% of journal editors used peer review in 1966 (Lavelle, 1966). During the latter half of twentieth century when peer review became more sophisticated and systematic, the process itself began to be investigated (Weller, 2001). While the majority of peer review processes since 1750s have been anonymous or 'single-blind' (Spier, 2002) – authors do not know who the reviewers are – the practice of blinding reviewers to the authors' identity stems from more recent concerns about bias and fairness of peer review. In sociology, for instance, the *American Sociological Review* initiated 'double-blind' reviewing in 1956, shortly after the editor had conducted a test of the affiliations and backgrounds of authors in the major sociology journals and found that "some of the reviewing was of persons rather than of submissions" (Abbott, 1999: 145). Double-blind reviewing has become standard practice across many disciplines, but the depersonalization and decontextualization of knowledge inherent in the ethic of anonymous reading remain controversial questions.

As with journals, the editorial policies of book publishing in early university presses relied on largely passive methods of manuscript selection. As a result of university presses' gradual movement from serving the university community toward a business-oriented model, review policies became more strict and substantive. The early relationship with a particular university's mission has been replaced by the goal to promote 'excellence' for the service of an international community on a field-by-field basis. Under the current system, researchers tend to seek to publish their books with the most prestigious press in the field, regardless of affiliation. The importance of unbiased, organized skepticism is visible in many academics' preference to publish at presses other than their 'own,' as institutional distance avoids suspicion of favoritism and provides 'external validation' (Fitzpatrick, 2011).

Ex ante Modes of Peer Review

As research funds became available in the nineteenth century, a new mode of peer review emerged: the process of awarding research grants on a competitive basis on the recommendation of regular or ad hoc committees of scientists, who could be considered peers of the grant competitor. Beginning in 1831, the *British Association for the Advancement of Science* developed a system of expert sectional committees to recommend expenditure of funds for specific research projects. In the United States, the *Carnegie Institution of Washington* set a similar system in 1902, followed by other granting agencies (Burnham et al., 1987). As noted by John Burnham (1990), grant peer reviewing developed independently of editorial peer reviewing, and in contrast to the latter, it seemed to spread by diffusion through personal contact and example within the scientific community.

In the post-World War II period, when a scaled-up scientific enterprise came to depend on the state for funding, the peer review of grants became firmly established. The implicit social contract underwriting public support for science was most famously institutionalized by Vannevar Bush's report for the US president in 1945. As national research funding throughout the Western world reached a steady-state dynamic in the 1970s, competition on resources intensified, and peer review procedures to select projects for funding became more strongly formalized (Elzinga, 2012; Ziman, 1994). However, the process still takes a variety of forms, and is shaped by considerations of public policy, national prosperity, and industrial needs along with scientific quality control.

Peer review has traditionally worked best *ex post* – that is, in making judgments about past performance, typically about completed work that one wishes to publish. Specific issues arise when peer reviewers are asked to assess research *ex ante* – before the research actually takes place. According to Steve Fuller (2002), special measures are needed to ensure that judgments that are fundamentally about the future – such as the direction of the research agenda and its larger public import – are not held entirely captive to the past. Indeed, the tendency for decisions to hold on peer consensus, or what psychologists call 'interrater reliability,' brings about a conservative bias of peer review, as controversial and truly novel ideas may become filtered out. Peer review may also be inherently conservative due to its reliance on existing knowledge and its gatekeepers (e.g., Horrobin, 1990). This applies to both *ex ante* and *ex post* reviewing.

As the assessment of performance has become more extensive in higher education and research policy making, peer review is being used more widely. It is increasingly employed to conduct both retrospective and forward looking evaluations of people, practices, and organizations. It is used, for example, in considerations of tenure and academic posts, to assess the performance of university departments or research programs, or to rank the relative merits of individual scholars. Because these reviews typically involve assessors examining in a summary way publications and grant-aided research activities that have already been peer-reviewed, they are sometimes referred as second-order or secondary peer review (British Academy, 2007). Whether these procedures still carry the meaning of organized skepticism can be doubted; instead of scrutinizing knowledge in a critical manner, they seem to have turned academic excellence into an ideology in itself (Readings, 1996).

Changing Contexts of Peer Review

The dominant status of peer review in modern science has resulted from the rise of academic disciplines. Between 1850 and 1900, Western science was transformed from a loose collection of local societies, research institutes and academic programs into a series of densely professionalized disciplines (Burke, 2012: 258). Following this transformation, academic disciplines have defined peers (Holbrook, 2010). During the recent decades, however, the proliferation of inter- and transdisciplinary research programs, institutes and other movements of resistance to increasing specialization have posed problems for peer review as the legitimator of knowledge. In this context, a fundamental question arises: who should count as a peer? The most obvious answer to this question thus has an obvious flaw: Disciplinary experts are often asked to serve as peer reviewers of research that may extend beyond the bounds of their expertise.

The definition of peer on disciplinary basis is insufficient also in its assumption of the universality of knowledge. After the age of positivism in social sciences, it is increasingly

recognized that domains of knowledge are not uniform across the world, nor are the means of scholarly communication. In social and behavioral sciences, in particular, cultural specificities play an important role, and contextual knowledge may be necessary for one to act as a competent peer reviewer. Moreover, despite the rise of English as the lingua franca of academic world, increasingly employed in journals, language boundaries limit the actual scope of peers.

The decline of discipline as the sole basis of defining peers is accompanied by another trend, that of democratizing expertise. It favors the extension of peer review from narrow academic elites to the full range of higher education institutions, contract researchers, as well as nonacademic members of society (e.g., Funtowicz, 2001; Fuller, 2002). This trend has questioned the democratic ideals associated with the 'ethos of science,' because it may obscure how some parts of the scientific community tend toward collusion with power elites in the political and state bureaucratic spheres.

Not only the scope of peers, but also the organizational forms of peer review are in upheaval. Open peer review, in which the reviewer and the author are known to each other, has emerged as a response to criticism of the incumbent anonymous process. Its major advantage is that it renders reviewers more accountable for their judgments. Moreover, open review through the internet provides space for new forms of community inspection. Through direct 'peer-to-peer review' practices, online communities not only bring in more voices, but make the review an open conversation. A key activity of such a system is the review of reviewers, which takes the idea of organized skepticism to a step further. However, online open review has not gained much popularity yet. An experiment undertaken by *Nature* in 2006, for example, failed: Few authors agreed to have their papers opened to public comment, and the comments that the articles received were not thorough enough to elicit editorial decisions (Fitzpatrick, 2011).

As the social system of science has become increasingly complicated, unregulated, and dispersed, peer review has been elevated to a unifying principle for a remarkably fragmented field (Biagioli, 2002). Rather than being an unambiguous medium of organized skepticism, peer review is increasingly used to serve corporate, governmental, or other objectives defined outside the autonomous sphere of science. Research is scrutinized not only for its academic quality, but also for its strategic importance and socioeconomic impacts. At the same time, measures adopted from the economic realm, like cost-benefit ratios, performance indicators, rankings, and benchmarks are used to evaluate scientific activities. Peer review is but one means among many of judging quality; rather than holding a central place in the process, peers are often asked to assist in applying evaluative indicators and interpreting their meanings in specific scholarly contexts. The meritocratic reward system of science has turned into a mixed economy of worth, in which attractiveness to external capital is valued along with the intellectual authority granted by peer review.

Conclusion

The need of critical judgment on scholarly work remains the *raison d'être* of peer review, but the purpose and beneficiaries of this practice are more ambiguous than ever before. As historians have shown, the original interests protected by peer review were primarily those of the state and its academies, not those of the broader scholarly community. Peer review underwent a slow transformation from state disciplinary techniques into academic ones, and became a symbol of the relationship between science and liberal societies (Biagioli, 2002). In the latter half of twentieth century, the expansion, dispersal, specialization, and fragmentation of scholarly knowledge have deeply changed the definition and role of peer review again, narrowing its meaning to a mode of 'gate keeping' that is not simply about quality but about policing the limits of a discursive field. Current push for inter- and transdisciplinarity, greater accountability, and profitable knowledge have left the status of peer review contentious.

An optimal peer review procedure can hardly be put on place without some trade-offs between the various values the system is asked to serve (Hackett and Chubin, 2003). From the perspective of organized skepticism, not every critical voice is constructive to the knowledge enterprise. What exactly makes intellectual critique relevant enough to require further inspection? And under what conditions does reviewing improve the outcome, and when does it do more harm than good? On the other hand, it is claimed that peer review is already overused, as scholars spend more and more time in reviewing others instead of conducting research. Questions of this kind have begun to challenge the previously unquestioned status of peer review and raised serious discussions of its future.

Perhaps the most important conceptual and practical problem of peer review is the degree to which it exhibits a conservative bias toward what is already known, and if so, should that bias be celebrated or overcome? There is a necessary creative tension between innovation and quality control, and peer review can be placed at the center of this tension (Horrobin, 1990). Ziman (1994) suggests that organized skepticism is meant as a governor on knowledge. The idea is that organized skepticism serves to slow us down so that we do not make outrageous leaps that cannot be supported. Ziman's conception of peer review might be characterized as a sort of precautionary principle for epistemology. There are, however, also 'proactionaries' (e.g., Fuller, 2002; Holbrook and Briggle, 2014), who emphasize the accountability of knowledge producers for achieving real world outcomes rather than accountability to one's disciplinary peers. It can be claimed that future wisdom relies not on certified knowledge, but on our ability to act in the face of uncertainty, ambiguity, and the unknown.

There may, in fact, be a historical shift to the contract between science and society, which undermines the original mandate of organized skepticism. The ethos of organized skepticism arose against the unquestioned 'truths' of political and religious authorities. In today's world of multiple values, beliefs – and truths – it is increasingly claimed that a new contract must ensure that scientific knowledge is not just reliable, but 'socially robust,' and that its production is seen by society to be both transparent and inclusive (e.g., Gibbons, 1999). In this context, peer review is not enough to legitimate academic work outside of the academy.

Acknowledgment

I thank J. Britt Holbrook for his help and constructive comments in the preparation of this article.

See also: Discipline Formation in the Social Sciences; Funding of Social Science, History of; Journals, Scholarly; Norms in Science; Peer Review and Quality Control; Policy Knowledge: Universities; Publishing: Academic and University Presses; Science and Politics: Value Neutrality; Science, History of; Science, Public Engagement with; Scientific Disciplines, History of; Scientific Misconduct, Plagiarism, and Institutional Control of Misconduct.

Bibliography

Abbott, A., 1999. Department and Discipline: Chicago Sociology at One Hundred. The University of Chicago Press, Chicago; London.

Biagioli, M., 2002. From book censorship to academic peer review. Emergences 12 (1), 11–45.

British Academy for the Humanities and Social Sciences, 2007. Peer review: the challenges for the humanities and social sciences. A British Academy Report. http://www.britac.ac.uk/policy/peer-review/contents.cfm (accessed 28.10.13.).

Broman, T., 2000. Periodical literature. In: Frasca-Spada, M., Jardine, N. (Eds.), Books and the Sciences in History. Cambridge University Press, Cambridge, pp. 225–238.

Burke, P., 2012. A Social History of Knowledge II: From the Encyclopédie to Wikipedia. Polity Press, Cambridge.

Burnham, J., 1990. The evolution of editorial peer review. Journal of the American Medical Association 263 (10), 1323–1329.

Burnham, J., Sauer, J., Gibbs, R., 1987. Peer-reviewed grants in U.S. trade association research. Science, Technology, & Human Values 12 (2), 42–51.

Bush, V., 1945. Science, the Endless Frontier. United States Government Printing Office, Washington, DC.

Chubin, D., Hackett, E., 1990. Peerless Science: Peer Review and U.S. Science Policy. State University of New York Press, Albany.

Elzinga, A., 2012. Features of the current science policy regime: Viewed in historical perspective. Science and Public Policy 39, 416–428.

Fitzpatrick, K., 2011. Planned Obsolescence: Publishing, Technology, and the Future of the Academy. New York University Press, New York.

Fuller, S., 2002. Knowledge Management Foundations. Butterworth-Heinemann, Boston.

Fuller, S., 2003. Kuhn vs. Popper: The Struggle for the Soul of Science. Columbia University Press, New York.

Funtowicz, S., 2001. Peer review and quality control. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopaedia of the Social and Behavioural Sciences. Elsevier, pp. 11179–11183.

Gibbons, M., December 2, 1999. Science's new social contract with society. Nature 402, C81–C84.

Hackett, E., Chubin, D., 2003. Peer Review for the 21st Century: Applications to Education Research. National Research Council, National Academy of Sciences, Washington, DC.

Holbrook, J.B., 2010. Peer review. In: Frodeman, R., Klein, J.T., Mitcham, C. (Eds.), The Oxford Handbook of Interdisciplinarity. Oxford University Press, Oxford, pp. 321–332.

Holbrook, J.B., Briggle, A., 2014. Knowledge kills action – why principles should play a limited role in policy-making. Journal of Responsible Innovation 1 (1), 51–66.

Horrobin, D., 1990. The philosophical basis of peer review and the suppression of innovation. Journal of the American Medical Association 263 (10), 1438–1441.

Kronick, C., 1990. Peer review in 18th century scientific journalism. Journal of the American Medical Association 263 (10), 1321–1322.

Lamont, M., 2009. How Professors Think: Inside the Curious World of Academic Judgment. Harvard University Press, Cambridge.

Lavelle, J., 1966. Facts of journal publishing, IV. Publications of the Modern Language Association of America 81 (6), 3–12.

Merton, R.K., 1973[1942]. The normative structure of science. In: Storer, N.W. (Ed.), The Sociology of Science: Theoretical and Empirical Investigations. The University of Chicago Press, Chicago.

Polanyi, M., 1962. The Republic of Science: its political and economic theory. Minerva 1, 54–73.

Popper, K., 1959. The Logic of Scientific Discovery. Basic Books, New York.

Readings, B., 1996. The University in Ruins. Harvard University Press, Cambridge.

Rennie, D., 2003. Editorial peer review: its development and rationale. In: Godlee, F., Jefferson, T. (Eds.), Peer Review in Health Sciences. BMJ Books, London, pp. 1–13.

Slaughter, S., Leslie, L.L., 1997. Academic Capitalism: Politics, Policies and the Entrepreneurial University. Johns Hopkins University Press.

Shatz, D., 2004. Peer Review: A Critical Inquiry. Rowman & Littlefield Publishers, Lanham.

Spier, R., 2002. The history of the peer review process. Trends in Biotechnology 20 (8), 357–358.

Weller, A., 2001. Editorial Peer Review: Its Strengths and Weaknesses. Information Today, Medford, NJ.

Whitley, R., 1984. The Intellectual and Social Organization of the Sciences. Clarendon Press, Oxford.

Ziman, J., 1994. Prometheus Bound: Science in a Dynamic Steady State. Cambridge University Press, Cambridge.

Ziman, J., 2000. Real Science: What It Is, and what It Means. Cambridge University Press, Cambridge.

Zuckerman, H., Merton, R., 1971. Patterns of evaluation in science: institutionalisation, structure and functions of the referee system. Minerva 9, 66–100.

Pension Policy

Karen M Anderson, Radboud University Nijmegen, Nijmegen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines why and how the first pension policies were adopted in North America, Europe, and the antipodes; identifies what the most salient dimensions of pension policy are; and discusses national policy responses to the challenges of aging, changing employment patterns, and fiscal austerity. The article also analyzes the most important explanations for cross-national differences in pension policy. The discussion centers on the affluent democracies, with references to developing countries where relevant.

Introduction

Pension policy refers to the mix of public and private programs that provide income to an individual or his/her survivors during retirement or incapacity. Pensions take myriad forms: they may be flat-rate, earnings related, lump sum, or annuities. In most affluent democracies, state provision dominates pension policy, usually supplemented by private pensions that are part of employment contracts or individual private pension products. This mix of public and private provision is central to understanding the historical development of pension policy and national responses to the current dilemmas presented by aging and fiscal austerity.

Today, pension spending is the largest item in government budgets in most Organisation for Economic Co-operation and Development (OECD) countries. In 2010, the 28 richest OECD countries spent an average of 9.3% of gross domestic product (GDP) on public pensions, with Italy topping the spending at 15.3% of GDP and Iceland spending the least at 4% of GDP. Expenditure on private pensions is also substantial, accounting for 40–50% of total pension spending in Denmark, Switzerland, and the Netherlands (figures for 2006; OECD, 2008: 42), and the proportion is rising in the OECD countries dominated by public provision. Affluent democracies thus spend considerable resources on pension policy, and the amount will continue to grow as populations age. Developing countries also devote considerable resources to pension policy, but given these countries' lower level of affluence, pensions are usually insufficient to finance complete exit from the labor force. This article examines why and how the first pension policies were adopted in North America, Europe, and the antipodes; identifies what the most salient dimensions of pension policy are; and discusses national responses to the challenges of aging, changing employment patterns, and fiscal austerity. The discussion centers on the affluent democracies with references to developing countries where relevant.

Origins and Development

Pensions have a long history. For more than a 1000 years, rulers, church officials, and governments have provided pensions to favored advisers, high-ranking officials, and military officers to reward them for loyalty and to ease them out of active service. *Pension policy*, in contrast, is a quite different thing, because it entails purposive state policies that shape the mix between public, private, and informal forms of support during old age or disability. It is only since the late 1800s and the early 1900s that a conscious pension policy in Europe, North America, and the Antipodes began, and it is only since the postwar period that pension policy emerged as a central component of state policy in the OECD countries and the developing world. Indeed, full exit from the labor force – retirement – was scarcely attainable for most groups until after the Second World War.

Until well into the twentieth century, old age and disability usually meant poverty for most people. Privileged groups, in contrast, have had access to different forms of support in old age since the Middle Ages, and the less fortunate had to make do with informal forms of support when they could no longer work. In the early industrial craft economy, older workers could move to less-demanding tasks, they could work at home, or they would simply be supported by family members. Industrialization changed all of that because it forced most workers into the wage economy, and few workers could save enough to support themselves in old age. The result was a surge in old-age poverty.

Some groups of workers did have access to pension provision, albeit usually at low levels. In North America, railroad companies were among the first to set up pension plans for their employees. In Germany, miners and railroad workers were the first to receive pension coverage, in the early- to mid-1800s. The introduction of pensions encouraged efficient retirement and promoted labor peace. Employers later used pensions for recruitment, retention, and workforce management, and occupational pension coverage spread to public servants like teachers and public utilities employees. Banking and insurance companies in the United States introduced occupational pensions in the 1920s, and the practice spread to industries like textiles, steel, paper, and printing firms.

Innovation in public pensions had to wait until the last decades of the nineteenth century when Germany introduced contributory pensions for private sector employees in 1898. Other countries were also quick to introduce public pensions, opting for poverty-preventing, flat-rate pensions: Denmark introduced means-tested pensions in 1891, New Zealand in 1898, the United Kingdom in 1908, and Sweden in 1913. Italy followed with contributory pensions for industrial workers in

1919 and France in 1932. The United States introduced contributory, earnings-related pensions for all workers in 1936.

The postwar period saw the rapid expansion of public pension schemes; benefits were upgraded and extended to more categories of workers, and several Beveridgean pioneers (New Zealand, Sweden, the United Kingdom) added earnings-related schemes to their existing flat-rate pensions. Other countries, such as the Netherlands, Denmark, and Switzerland, added or expanded mandatory or quasi-mandatory private occupational pensions. By the 1960s, most workers in the affluent democracies had access to a retirement pension that paid at least 40–60% of average wages after 40 years of service.

Analyzing Pension Policy: The Nuts and Bolts of Pensions

Pensions are complex, and understanding the interplay of different forms of public and private provisions requires some discussion of the mechanics of pension arrangements. The point of departure for most analyses of pension policy is to distinguish between state provision (the first pillar); private, employment-based pensions (the second pillar); and voluntary, private provision (the third pillar). The three pillar classification scheme was popularized by the World Bank's 1994 publication, *Averting the Old Age Crisis*. For the World Bank, the three pillars were not merely a classification scheme, but a model to be emulated (more on this below). The OECD and other international organizations have adopted the three pillar classification system, and the World Bank recently amended it to include means-tested social pensions (the 'zero pillar') and informal forms of support (the fourth pillar; Holzmann et al., 2008). The discussion here is based on the three pillar model.

Political and economic actors have been fighting over the minutiae of pension plans and the division of labor between public and private schemes since the first public and private schemes were introduced. These details of policy structure are important because they generate resources like market access for private providers, organizational resources for unions and employers charged with pension plan administration, and above all, a social right to retirement income for large groups of citizens. Moreover, the details of policy design determine the extent to which pensions, especially public ones, are redistributive or actuarially neutral.

There are four crucial aspects of pension design in all three pillars: *coverage*, *benefits*, *financing*, and *administration*. That is, pension programs (both public and private) differ concerning the groups that they provide benefits to, the generosity and type of benefit that is offered, the way in which these benefits are financed, and how they are administered. The most important dimension of variation concerning the *coverage* of public pension schemes concerns the degree of universalism. Are pensions universal in the sense that they constitute a right associated with citizenship or residence? Or are pensions selectively organized in that they are awarded on the basis of occupational status? If the example of basic pension schemes offering flat-rate benefits (for example, the state pensions in the Netherlands or Denmark) is taken, universalism means that all citizens or residents receive the same package of benefits, regardless of employment history. If public, earnings-related schemes like Sweden's are taken as examples, universalism means that all occupational groups participate in the same system – there are no special schemes for civil servants or the self-employed as there are in occupationally fragmented systems like Germany's. This means that an occupationally fragmented system like the German one provides near-universal coverage to the working population, but it does so within the confines of a system with different schemes for dependent employees, civil servants, miners, and the liberal professions.

The *coverage* of second pillar schemes depends critically on whether national regulations require or encourage employers to offer occupational pensions and employees to participate in them. In terms of compulsion, governments have three options: require employers to provide occupational pension coverage, delegate this issue to collective bargaining, or rely on voluntary provision. Switzerland is an example of the first strategy, Denmark and the Netherlands of the second, and the United Kingdom of the third.

The structure of *financing* refers to the choice between the use of taxes or contributions to fund pensions and the extent of prefunding. In a pure pay-as-you-go (PAYGO) system, the working-age population pays taxes or pension contributions that finance current pension payments. Several variations of this model are possible. For example, PAYGO schemes may include 'buffer funds' to cover fluctuations in costs caused by varying numbers of beneficiaries, contributions, and changes in benefits. The design of buffer funds differs across systems, but the basic idea is to have a financial reserve for meeting a portion of current or future pension payments. Few contributory public pension schemes are fully contribution financed. Governments often provide a subsidy financed from general revenues to the public pension scheme to cover deficits, lower contribution rates, or pay for the contributions of those who are unemployed, sick, or caring for children. Tax incentives are an important driver of second and third pillar provision. Employers can typically deduct the cost of occupational pension contributions from their taxable income, and similar provisions typically exist for individual contributions to second pillar and often, third pillar schemes.

Second pillar schemes are often fully funded in the sense that financial assets are held in trust to cover pension liabilities. The term 'full funding' is deceptive, however, because its meaning varies enormously across countries. The first key parameter that determines the definition of full funding is the distinction between group life insurance contracts and pension funds. Simplifying greatly, capital requirements are higher for insurers than for pension funds. A second key parameter is the specific nature of funding requirements: the required level of funding (as percent of liabilities), the method for calculating the value of future liabilities, and the rules governing over- and underfunding, including the ownership of surpluses and liability for deficits. Given that life insurance is regulated by European Union (EU) legislation and pension funds less so, national authorities have much leeway in tailoring pension fund legislation to national needs than they do for insurance-based occupational pension schemes.

In general, there are two methods for designing pension *benefits* in the first and second pillar. In a *defined benefit* (DB) scheme, the pension plan sponsor (the government, an

employer, or a pension fund) promises to pay an employee a specific benefit at retirement. These schemes may be prefunded or paid out of current payroll costs, but the size of the pension benefit is fixed in advance, and contribution levels are adjusted as necessary to finance pension benefits. Typical DB occupational schemes pay some percentage (often 60%) of average earnings or the final salary after 40 years of employment. Obviously, the number of years required for a full pension, the annual rate of pension accumulation, and the reference salary (final salary, average salary) used to calculate the pension level vary greatly across schemes.

In its purest form, a *defined contribution* (DC) scheme involves an individual account in which participants save for retirement. The contributions are often cofinanced or fully financed by employers. The principal and income earned by the investments in the account are then converted to an annuity at retirement or simply withdrawn at regular intervals after retirement. There is no insurance or compensation for bad financial decisions or market turmoil; returns completely depend on the level of contributions and investment performance.

These two 'pure' types obscure the ways in which the DB–DC distinction is really a continuum rather than two mutually exclusive categories. DC schemes may include elements associated with DB schemes, such as guaranteed rates of return and collective, rather than individual, membership. When these parameters apply to DC schemes, the individual bears much less risk than in a pure DC scheme. Similarly, DB schemes often incorporate features more commonly associated with DC schemes, such as making a portion of the pension dependent on investment returns.

The *administration* of pension schemes is also an important aspect of pension design. The state administers first pillar schemes, sometimes in cooperation with representatives of employers and unions (in France, for example). Second pillar schemes vary enormously in terms of administrative structure. Broadly speaking, occupational pensions may be organized as company schemes, sectoral (multiemployer) schemes, or professional schemes. This dimension determines the size and structure of the risk pool. The next key dimension is the composition of administrative boards: are employers solely responsible for pensions and entitled to make all relevant decisions, as in the United Kingdom? Or do corporatist boards with representatives of both unions and employers run things, as in Denmark, Switzerland, and the Netherlands? Are pensioners represented on these boards? As Myles and Pierson (2001) argue, these administrative boards are key actors in occupational pension policy because they make the key decisions about benefit formulas, contribution rates, eligibility, and investment policy.

Typologies of Pension Policy

Comparative analyses of pension policy typically use Esping-Andersen's typology of pension regimes as their point of departure. Esping-Andersen argues that there are 'worlds of welfare' among the affluent democracies, and pension policy is one of the core programs that the typology is based on. Esping-Andersen's classification is based on the division of labor between the state, market, and family in the division of welfare. For pensions, the core issues are the public–private mix and the quality of social rights. Universal, tax-financed, generous public pensions that crowd out private provision and reduce status differentials are the hallmarks of the social democratic pension regime. Occupationally fragmented, contribution-financed, status-preserving public pensions are the core of the corporatist/conservative regime. Market-based pensions combined with state pensions aimed at poverty prevention are the key elements of the liberal regime (Esping-Andersen, 1990). Several researchers argue that there is a fourth regime type, the Southern European regime, characterized by fragmented earnings-related pension provision, clientelistic pension policy, and informal forms of support in old age (Ferrera, 1996).

An alternative to Esping-Andersen's three regimes is to classify pension systems based on the distinction between Beveridge and Bismarck (Bonoli, 2000; Hinrichs, 2000; Ebbinghaus, 2011). The Bismarckian approach emphasizes first pillar, earnings-related pensions financed by earmarked contributions that are usually divided evenly between employers and employees. Germany is the prototypical case: pensions are part of the social insurance system that insures the incomes of wage earners. Benefits are based on 'equivalence' – the principle that benefits are strictly related to contributions. The state steps in to finance contributions for periods of nonemployment (education, child rearing) that earn pension entitlement. Bismarckian systems emphasize status maintenance because the income ceiling for pension entitlement is typically high (1.5 or 2 times the average wages) and dependent spouses are usually covered. Besides Germany, the classic Bismarckian countries are France, Austria, Belgium, and Italy (see Hinrichs and Lynch, 2010, for more).

The Beveridgean approach – named for William Beveridge, author of the influential 1942 report that provided the blueprint for the postwar British welfare state – emphasizes universal, tax-financed, flat-rate provision aimed at poverty alleviation. The English-speaking countries (except for the United States) and the Nordic countries exemplify Beveridgean principles because the state pension, either alone or in combination with means-tested supplements, is sufficient to lift the poorest pensioners out of poverty. Beveridgean countries such as the United Kingdom, Australia, Denmark, Switzerland, and the Netherlands later added extensive private occupational coverage to the basic security provided by their state pensions. According to Hinrichs (2000), Sweden, Finland, Norway, and Canada began as Beveridgean systems because they introduced basic security pensions in the first half of the twentieth century, but later added public earnings-related pensions that moved them to the Bismarckian category.

The Bismarckian–Beveridge classification system has gained ground in recent scholarship, informing several important analyses of pension policy (see for example, Bonoli, 2000; Ebbinghaus, 2011; Häusermann, 2010). The chief virtue of this approach is that it emphasizes the institutional goals of public pension institutions (basic security or status maintenance) and the political coalitions that constructed them and continue to maintain them. Esping-Andersen's threefold typology has drawn criticism for being too static and too focused on the outcomes of decommodification and social stratification, yet

the move toward the Bismarck–Beveridge dichotomy in recent research often obscures more than it illuminates. One of the strengths of Esping-Andersen's typology is that it distinguishes between conservative and social democratic goals in welfare policy. Center-right parties preferred status-maintaining social policies that also reinforced traditional family structures, so they adopted Bismarckian principles. Nordic social democrats usually wanted something else: universal social policies that would create cross-class solidarities and promote individual autonomy. These distinctions between social democratic and conservative/corporatist pension policies disappear in the Bismarck–Beveridge dichotomy.

Explaining Institutional Variation

Scholarly attempts to classify pension systems according to twofold or threefold typologies impose order on tremendous institutional diversity. The key questions animating research in pension policy concern the timing and content of early public pension programs (Beveridge vs Bismarck), the contours of the expansion and consolidation of public programs, and the nature of the public–private mix in pension policy. Since the 1990s, researchers have been preoccupied with the issue of *pension reform*, investigating the causes and consequences of rolling back or adjusting pension arrangements that have been in place for several, if not many, decades.

The Modernization Thesis

Some of the most influential early analyses of the development of pension policy attributed the development of pension policies in the developed countries to the pressures arising from industrialization and democratization (Flora and Alber, 1981). Industrialization forced laborers out of agriculture and small-scale, craft-based industries and into wage labor in large factories. With industrialization, existing informal forms of support for older workers disappeared. In the early capitalist period, older workers performed less-demanding tasks within small, family-owned enterprises, or they were supported by family members. Craftsmen also usually owned their enterprises, and this capital could be used as a source of support in old age. Employment in large-scale industries at low wages meant that few could afford to save for their old age; indeed, most people worked until they no longer could, because of disability, old age, or both. Thus industrialization created a growing class of elderly poor, and existing poor relief policies could not keep pace. The earliest public pension schemes, whether Beveridgean or Bismarckian, were created to address the problem of old-age poverty, at least in part. Bismarck's pioneering pension reform in 1889 aimed not only to reduce poverty among the aged but also to pacify the growing industrial labor force.

Other researchers working within this perspective highlight the role of economic growth and demographic change in driving the expansion of social protection after the introduction of the first programs. Economic growth generates additional resources that can be spent on social policies like pensions; a high-level of socioeconomic development is thus a prerequisite for the introduction of social insurance schemes like old-age pensions, and it is the growing wealth of the country that is one of the most important factors driving their expansion. Another variation of this perspective emphasizes demographic change, pointing to the correlation between the growing size of the elderly population in affluent democracies and spending on programs targeted at the elderly like old-age pensions. Wilensky (1975) was the first to argue that the relationship between economic developments and welfare state effort was mediated by the size of the elderly population. Recent variants of this argument include Tepe and Vanhuysse (2009).

The modernization perspective has been widely criticized for its emphasis on more or less automatic state policy responses to objective pressures (see, for example, Orloff, 1993; Lynch, 2006). Explanations of pension policy development based on modernization theory are by their very nature functionalist arguments; they emphasize the fact that nearly all affluent democracies introduced public pension schemes by the first half of the twentieth century and expanded them after the Second World War. However, the content of these policies and their effects on income inequality and social stratification are left unexamined. This means that political ideologies and distributional conflict are largely absent in this approach.

Politics Matters

There is a large and influential literature that builds on modernization theory by arguing that political processes mediate the pressures caused by industrialization. Whereas modernization theorists emphasize the similar responses of democratizing states to the social problems created by industrialization, proponents of politics-centered explanations of pension policy development highlight the crucial role played by political mobilization in shaping outcomes. Thus the extension of the suffrage in Europe, North America, and the antipodes created opportunities for the political representation of groups favoring state-led social reform. At the same time, workers and employers began to organize, building powerful organizations to represent their interests in workplace negotiations and electoral politics. Korpi (1983), for example, argues that the structures of welfare state policies, including pension policy, are the result of the 'democratic class struggle.' The working class, represented in the workplace by trade unions and in the political arena by social democratic parties, drew on its 'power resources' to press for social policy reforms in the democratic political process. Organized business is generally viewed to be hostile to such reform efforts, preferring to retain as much control as possible over working conditions (see Swenson, 2002; Mares, 2003 for arguments that challenge this view). The character of welfare state policies like pensions is thus the outcome of the balance of power between organized labor and capital. In countries where the working class was powerful, the state introduced generous public pensions that crowded out private provision. Esping-Andersen's (1990) typology of pension regimes is an influential application of power resources theory. As discussed above, it is the influence of class actors working within the democratic political system, as well as their ability to build coalitions across class lines, that

is the most important cause of variations in the structure of pension regimes.

Institutionalism

The renaissance of scholarship concerning the role of political institutions in shaping political processes and outcomes has been an important focus of recent research concerning the development of pension policy. This literature is large and varied, but contributions to it share the assumption that stable governance arrangements structure political outcomes. These governance arrangements, whether they are constitutional structures, industrial relations systems, or party systems, form the institutional context within which political actors formulate their interests and choose their political strategies. Understanding the development of pension policy thus requires investigation of the ways that institutional context shapes both the goals and strategies of advocates and opponents of pension reform.

Research concerning the role of constitutional design on welfare state outcomes has generated important insights about the constraints and opportunities presented by formal political institutions on policy making. Immergut's (1992) analysis of the role of veto points in the development of health care policy in Western Europe is the basis for a large literature that emphasizes the role of veto points in shaping the contours of pension policy (see, for example, Bonoli, 2000; Immergut et al., 2007). Veto points refer to the opportunities created by constitutional rules for reform opponents to block or weaken legislation. Federal systems like the United States, Switzerland, and Germany offer more veto opportunities than Westminster systems like the United Kingdom. The more hurdles a legislative proposal must pass through, the more difficult it is for the executive to pass a reform. Other scholars working within the institutionalist perspective stress the role of political competition in the making of pension policy (Lynch, 2006; Orloff, 1993).

A particularly influential variant of the institutionalist perspective is the policy feedback approach advanced by Pierson (1994, 2001) and others (Orloff, 1993). The central assumption of this approach is that policies have political effects and thus shape future political processes. Applied to the welfare state and pension policy, policy feedbacks include incentives for interest group activity, the emergence of strong support coalitions, coordination effects, and changing expectations about the role of government in social protection. Extensive public pensions generate their own sources of support because interest groups are created to protect the beneficiaries of these programs, individuals plan their retirement around the existence of these programs, and over time, the public accepts the role of government in old-age provision. Pensions are everyone's favorite example of path dependence, because the feedbacks that create policy stability are often very strong. Indeed, researchers point to the strong potential for 'lock-in' effects in pension policy (Myles and Pierson, 2001). Because pension policies entail long-term commitments affecting individual and family incomes, they are particularly stable and resistant to retrenchment.

Pension Policy Outside the OECD

The developing world is not immune to the challenges associated with creating pension policies that address poverty and income maintenance in old age, but the nature of these challenges is much different than in the OECD. Whereas the OECD countries must reform their mature pension systems to deal with aging and fiscal austerity, the developing countries face the challenge of expanding policies that offer incomplete and/or inadequate protection from old-age poverty. There is enormous variation in pension policy in the developing world. In general, however, they face four common challenges. First, mandatory coverage is fairly low, whether in the first or second pillar. Second, benefits are usually insufficient to permit complete exit from the labor force. Third, informal means of support in old age play an important role in ensuring the livelihood of the elderly. Finally, privatization has progressed further than in the OECD.

The former communist countries inherited pension systems that were not well suited to the market economy. The transition to a market economy undermined full employment and state-sponsored social benefits like pensions. State pensions were retained in most of the Central and East European countries, but spending exploded in the wake of high unemployment and labor shedding. Most of the transition countries privatized large portions of their retirement provision, resulting in lower benefits for future retirees and an increase in individual exposure to investment risks (see Orenstein, 2008; Müller, 2001).

The Asian/Pacific countries did not introduce pension schemes until the 1950s, and the expansion of coverage and benefit adequacy has been uneven. Public provision dominates, and benefits are generous in China, Vietnam, and Pakistan. However, coverage is low: 13% of the economically active in South Asia and 35% in East Asia. (see OECD, 2009), and many systems allow participants to withdraw savings before retirement or to receive their pension as a lump sum. Like Latin America (see below) the Asian/Pacific countries are experiencing a rapid demographic transition. As the old-age dependency ratio rises, it will be even more difficult for these incomplete pension systems to prevent poverty in old age.

The first pension schemes were introduced in Latin America in the 1920s (Chile and Ecuador), and the rest of the region followed between 1930 and the 1950s. Since the 1980s, the region has been known for far-reaching privatization. Chile transformed its public, PAYGO scheme into a system of private, individual accounts in 1982. Nine other countries have implemented some form of privatization, including Mexico, Bolivia, El Salvador, and Peru. Despite these reforms, coverage is patchy, and benefits will be inadequate for large groups of pensioners. Indeed, Latin America faces its own set of challenges caused by aging: falling birth rates and rising longevity mean an increasing old-age dependency ratio. In one generation, Latin America will face demographic constraints similar to those currently experienced in Europe (Jackson et al., 2009).

The Future of Pension Policy

Two challenges dominate debates about the future direction of pension policy. First, the dramatic aging of populations creates

unprecedented pressure on social protection institutions that were designed for a high-fertility, high-employment, high-growth socioeconomic context. According to the most recent projections of the European Union, the proportion of people aged 65 years and older in the EU-27 will increase from 17% in 2007 to 30% in 2060. Moreover, the ratio of elderly persons to working-age persons (age 15–64 years) will increase from 1:4 today to 1:2 in 2060. Population aging coincides with a second challenge: rapidly changing labor markets and employment patterns. Full employment for standard, full-time workers seems to be a thing of the past. Instead, 'dual' or 'segmented' labor markets have emerged in many European economies (see, for example, Palier and Thelen, 2010), characterized by high levels of youth unemployment, the expansion of part-time and atypical work, and persistent long-term unemployment.

The heart of the 'pension problem' lies in the fact that fewer and fewer workers will be financing more and more pensioners. The more that pensioners believe their pension rights to be earned (on the basis of contributions on earnings) and therefore to have the status of rights, the less likely they will be to support pension reductions. This dependency of older generations on younger ones often generates severe distributional conflict when revenues are not sufficient to cover pension costs, because neither pensioners nor workers are likely to want to cover the funding shortfall. Thus the 'contract between the generations' is just as likely to unleash intergenerational conflict as it is to foster intergenerational solidarity.

A second challenge related to pension reform concerns growing income inequality among pensioners and between pensioners and the working population (obviously these two trends take different forms in different countries). The OECD reports that in the mid-2000s, people aged 65 years and older had an average income that was 82.4% of average population income. Those aged 66–75 years had higher average incomes than those aged 76 years and older. Women are more likely to be represented in the over 75 group because of their higher life expectancy, and their nonstandard earnings biographies (part-time, career breaks, etc.) result in lower earnings and therefore lower pensions. The OECD also reports that in 13 of the 25 OECD countries studied, the incomes of the elderly grew more quickly than that of the working population between the mid-1980s and the mid-2000s (OECD, 2011: 146).

These statistics point to growing pensioner affluence, even if average pensioner incomes remain below the average income for the entire population. In other words, pensioners have an increased capacity to maintain an autonomous household and to enjoy security in terms of health care and housing. At the same time, however, the numbers of frail elderly are growing, as are the numbers of females older than 75 years with low incomes. In order to avoid either cutting benefit levels significantly or substantially increasing contributions, many policy makers look to some form of funding as a way to make future pension payments more affordable. However, the shift from a PAYGO to a fully or partially funded system is difficult, because current contributors in a PAYGO system are already paying for the pensions of current retirees. Thus, in order to introduce funding, current workers would have to pay higher contributions in order to fund both current PAYGO pensions and their own future funded pensions. This double burden of the transition cohort of pensioners is known as the 'double payment problem' (see Pierson, 1994) and it is one of the primary mechanisms that 'locks' policy makers into maintaining PAYGO financing.

Policy makers in the OECD have responded to reform pressures by adopting parametric reforms that modestly reduce benefits and are sometimes accompanied by the introduction of funded private provision. One important policy measure has been to try to tighten the link between benefits and contributions by changing benefit formulas so that benefits depend more directly on contributions. Examples include increasing the reference period for the calculation of benefits and requiring longer qualifying periods. Other ways to reduce pension costs include changing the basis for the benefit calculation from gross to net wages, lowering the replacement rate, changing the formulas for benefit indexation or the tax status of benefits, or even simply skipping or delaying benefit indexation.

A different approach to addressing the financial sustainability of public pension systems has been to increase the revenues of pension systems, for example, by increasing the contribution made by general tax revenues or special taxes (value-added tax or energy taxes, for example) to pension systems – although these approaches, obviously, increase the general tax burden. Governments have also proposed fusing separate special pension schemes in order to eliminate inequities, reduce administrative costs, and increase the coverage of the system to include greater proportions of higher income persons, such as by rolling civil servants into the general pension schemes or eliminating special privileges for public sector workers.

In the multipillar systems of the United Kingdom, the Netherlands, and Switzerland, the financial crises of the 2000s and 2010s have prompted reductions in private occupational pensions. In the United States and United Kingdom, there has been a rapid movement from DB to DC plans over the past decades (Bridgen and Meyer, 2005). In the public sector, however, efforts to change the rules of the game for pension benefits are more explicitly political and tend to involve more overt political conflict. Regardless of whether the decision to convert a benefit scheme from DB to DC is made by governmental authorities is an outcome of collective bargaining, or is the result of what the private market is willing to offer in terms of individual pension policies, the shift from DB to DC exposes individuals to more market risk. If returns to investments are high, such a risk may prove beneficial; if not, pensioners will experience lower retirement income (see also Anderson, 2011).

See also: Retirement, Economics of; Welfare Economics; Welfare State, History of; Welfare and Work; Welfare.

Bibliography

Anderson, K.M., 2011. Occupational pensions in the Netherlands: adapting to demographic and economic change. In: Ebbinghaus, B. (Ed.), The Varieties of Pension Capitalism. Oxford University Press, Oxford, pp. 292–317.

Bonoli, G., 2000. The Politics of Pension Reform. Cambridge University Press, Cambridge.

Bridgen, P., Meyer, T., 2005. When do benevolent capitalists change their mind? Explaining the retrenchment of defined-benefit pensions in Britain. Social Policy and Administration 39, 764–785.

Ebbinghaus, B. (Ed.), 2011. The Varieties of Pension Governance. Oxford University Press, Oxford.
Esping-Andersen, G., 1990. The Three Worlds of Welfare Capitalism. Princeton University Press, Princeton.
Ferrera, M., 1996. The "Southern model" of welfare in social Europe. Journal of European Social Policy 6, 17–37.
Flora, P., Alber, J., 1981. Modernization, democratization, and the development of welfare states in Western Europe. In: Flora, P., Heidenheimer, A.J. (Eds.), The Development of Welfare States in Europe and America. Transaction Books, New Brunswick, NJ, pp. 37–80.
Häusermann, S., 2010. The Politics of Welfare State Reform in Continental Europe. Modernization in Hard Times. Cambridge University Press, Cambridge.
Hinrichs, K., 2000. Elephants on the move: patterns of public pension reform in OECD countries. European Review 8, 353–378.
Hinrichs, K., Lynch, J., 2010. Old-Age Pensions. In: Obinger, H., Pierson, C., Castles, F., Liebfried, S., Lewis, J. (Eds.), The Oxford Handbook of the Welfare State. Oxford University Press, Oxford.
Holzmann, R., Hinz, R.P., Dorfman, M., 2008. Pension systems and reform. Conceptual Framework. World Bank SP Discussion Paper no. 0824.
Immergut, E.M., 1992. Health Politics. Interests and Institutions in Western Europe. Cambridge University Press, Cambridge.
Immergut, E.M., Anderson, K.M., Schulze, I. (Eds.), 2007. The Handbook of West European Pension Politics. Oxford University Press, Oxford.
Jackson, R., Strauss, R., Howe, N., 2009. Latin America's Ageing Challenge. The Global Aging Initiative, CSIS.
Korpi, W., 1983. The Democratic Class Struggle. Routledge and Kegan Paul, London.
Lynch, J.F., 2006. Age and the Welfare State. The Origins of Social Spending on Pensioners, Workers, and Children. Cambridge University Press, Cambridge.
Mares, I., 2003. The Politics of Social Risk. Cambridge University Press, Cambridge.
Müller, K., 2001. The political economy of pension reform in Eastern Europe. International Social Security Review 54, 57–79.
Myles, J., 1994. Old Age in the Welfare State. The Political Economy of Public Pensions. Little, Brown and Company, Boston.
Myles, J., Pierson, P., 2001. The comparative political economy of pension reform. In: Pierson, P. (Ed.), The New Politics of the Welfare State. Oxford University Press, New York, pp. 305–333.
OECD, 2008. Private Pensions Outlook 2008. OECD, Paris.
OECD, 2009. Pensions as a Glance. Special Edition: Asia/Pacific. OECD, Paris.
OECD, 2011. Pensions at a Glance. OECD, Paris.
Orenstein, M.A., 2008. Out-liberalizing the EU: pension privatization in Central and Eastern, Europe. European Journal of Public Policy 15, 899–917.
Orloff, A.S., 1993. The Politics of Pensions. A Comparative Analysis of Britain, Canada, and the United States, 1880–1940. University of Wisconsin Press, Madison, WI.
Palier, B., Thelen, K., 2010. Institutionalizing dualism: cmplementarities and change in France and Germany. Politics & Society 38, 119–148.
Pierson, P., 1994. Dismantling the Welfare State. Cambridge University Press, Cambridge.
Pierson, P. (Ed.), 2001. The New Politics of the Welfare State. Oxford University Press, New York.
Rein, M., Rainwater, L. (Eds.), 1986. Public/Private Interplay in Social Protection: A Comparative Study. M.E. Sharpe, Armonk, NY.
Swenson, P., 2002. Capitalists against Markets. Oxford University Press, Oxford.
Tepe, M., Vanhuysse, P., 2009. Are aging OECD welfare states on the path to Gerontocracy? Evidence from 18 democracies, 1980–2002. Journal of Public Policy 29, 1–28.
Thane, P., 2006. Old Age in English History. Oxford University Press, Oxford.
Wilensky, H., 1975. The Welfare State and Equality. University of California Press, Berkeley.

People in Organizations

Jennifer A Chatman and Jack A Goncalo, University of California, Berkeley, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 11183–11189, © 2001, Elsevier Ltd.

Abstract

This article offers a critical review of research in organizational behavior from an interactionist perspective: we consider the reciprocal influence between personal characteristics and the organizational context. Specifically, some situations are stronger than others, some people's behavior is more consistent across situations, some people can influence situations more than others, and people select themselves into situations that suit their preferences. We illustrate the value of this perspective by identifying person–situation interactions that are relevant to a set of currently vibrant theories: organizational culture, demography, leadership, and creativity. Emerging from this review is a greater appreciation for the richness of behavior in organizations and the challenge of understanding and managing its complexities.

This article offers a critical review of research in organizational behavior from an interactionist perspective: we consider the reciprocal influence between personal characteristics and the organizational context. Specifically, some situations are stronger than others, some people's behavior is more consistent across situations, some people can influence situations more than others, and people select themselves into situations that suit their preferences. We illustrate the value of this perspective by identifying person–situation interactions that are relevant to a set of currently vibrant theories: organizational culture, demography, leadership, and creativity. Emerging from this review is a greater appreciation for the richness of behavior in organizations and the challenge of understanding and managing its complexities.

Organizational researchers study how individuals and organizations interact with their environment to accomplish both individual and collective goals. People are able to accomplish goals in organizations that they would be unable to accomplish alone, and they spend large portions of their lives interacting in organizational contexts. The complexities that emerge from bringing people together in organizational settings, therefore, provide a fascinating lens into the human experience. Organizational research brings together psychology, sociology, economics, and anthropology to examine both individual-level processes at work, as well as the organizational-level processes that dictate how firms relate to each other and their environment.

Reflecting this focus on both individuals and organizations, the field is split between 'macro' organizational behavior, drawing upon organizational sociology, and 'micro' organizational behavior, drawing upon social, personality, and industrial-organizational psychology. Macro-organizational behavior is concerned with the various strategies organizations use to adjust to their environment, such as the development of networks between firms. In contrast, micro-organizational behavior is concerned with the way personal characteristics combine with contextual features such as task, group, and organizational characteristics to predict work outcomes. Though the strength and uniqueness of organizational behavior rests on the equal emphasis of both individual and organizational processes and the utilization of a variety of research methods that cross disparate levels of analysis, our focus in this article is on micro-theory and research. Excellent reviews of macro-organizational theory can be found elsewhere in this collection (*see* Organization: Overview; Administration in Organizations; Organizational Control; Organizations and Culture; Organizations: Authority and Power) and in Carroll and Hannan (2000) and Scott (1998).

Micro-theory: Person–Situation Interactions

Micro-organizational behavior examines both personal and situational characteristics and, as in the field of psychology, researchers debate the relative utility of each in explaining behavior. Some have emphasized the stability of attitudes and behaviors over time. For example, a person's satisfaction with his or her job remains relatively stable over years and even decades (Staw and Ross, 1985). From this perspective, individual characteristics are the best predictors of behavior since they derive from personal dispositions that remain stable over time and across situations. Others have criticized this view and posited that organizations should be conceptualized as 'strong' situations that are powerful enough to shape individual behavior (e.g., Davis-Blake and Pfeffer, 1989). In strong situations individual differences are unlikely to be expressed. Instead, people learn appropriate attitudes and behaviors from their co-workers, established norms, and organizational practices and procedures; these social influence processes are presumed to predict individual behavior better than are personal characteristics.

Researchers have typically considered personal and situational factors in isolation from one another, but a complete understanding of organizational behavior requires their simultaneous consideration. An interactional approach is more complex than a mere additive melding of personal and situational characteristics because it attempts to represent both personal and situational factors and their reciprocal influence.

Interactions between personal and situational characteristics may take at least four forms. First, as specified above, some situations are stronger than others, leading to different levels of behavioral uniformity. Second, work situations do not affect everyone in the same way; some people's behavior is more consistent across varying situations. Third, certain people,

such as those exhibiting 'charismatic' leadership, can influence situations more than others. Finally, people do not select into situations randomly, but rather, into situations in which they think their attitudes and behaviors will be appreciated. Developing a complete theory of behavior in organizations, then, requires moving from considering personal and situational factors in isolation to considering the complexity and diversity of possible person–situation interactions and their effects on work outcomes. We illustrate this by identifying the types of person–situation interactions that are relevant to a set of vibrant research domains within organizational behavior: organizational culture, demography, leadership, and creativity.

Organizational Culture: Selection, Socialization, and Person–Culture Fit

Research on organizational culture has demonstrated that norms and values can shape individual behavior. Recent research has also demonstrated that behavior can be predicted by understanding how personal characteristics interact with cultural context. Below we discuss the four types of person–situation interactions as they relate to organizational culture.

Organizational Culture as Strong Situation

Though researchers agree that organizational culture is meaningful and important, they do not agree about how to define and measure it. Organizational culture research typically draws upon theories in anthropology or cross-cultural psychology and uses ethnographic or quantitative methodologies. Some researchers have emphasized shared values and meaning (O'Reilly et al., 1991) while others have emphasized the ambiguity of cultural values and the existence of subcultures (Martin, 1992). Most agree that the existence of different perceptions among members does not preclude the existence of shared assumptions. Organizational culture can be understood as a system of shared values and norms that define what is important and how organizational members ought to feel and behave. Through members' clarity about organizational objectives and their willingness to work toward these objectives, culture influences the attainment of valued organizational goals by enhancing an organization's ability to execute its strategy. This conceptualization focuses on two primary aspects of culture: first, the extent to which agreement and intensity exist about values and norms (culture strength), and, second, the extent to which these norms and values differ across settings in important ways (culture content).

Culture content refers to shared values and norms that define desirable behavior in the organization. For example, some cultures may stress the value of being a 'team player' while others may emphasize independence. For a culture to be strong, members must agree about which values are important and be willing to enforce them strenuously. Members of an organization may agree that they value, for example, being cooperative, but unless unequivocal and salient consequences result from compliance (e.g., cooperating with co-workers) and non-compliance (e.g., competing with or undercutting co-workers), the culture cannot be characterized as strong. A strong culture is characterized by social control in that members both agree about and are willing to enforce values and norms, even if such enforcement violates hierarchical lines of authority. Further, greater behavioral homogeneity among members, for better or worse, should be observed in stronger organizational cultures.

Cultural values are conveyed to members through the socialization processes that new recruits experience when they enter an organization. Though socialization often takes the form of training and contributes to increased task knowledge, information about the norms of the organization is also transmitted through training and other types of socialization (Morrison, 1993). Interestingly, such normative information appears to be transmitted very early in new recruits' membership since cultural fit changes little after members' first year (e.g., Rynes and Gerhart, 1990).

Behavioral Consistency Across Various Organizational Cultures

Although culture influences members' behavior, they are not merely passive recipients of social control. Individual characteristics may interact with the organization's culture to predict important behavioral and work outcomes. Research on person–organization fit, defined broadly as the compatibility between people and the organizations for which they work (Kristof, 1996) has focused primarily on congruence between patterns of organizational values and patterns of individual values. New employees whose values more closely match those of the organization view it as more attractive, are more likely to join when made an offer, are able to adjust to it more rapidly, perform better, are more satisfied, and remain with it longer (e.g., Chatman, 1991).

But, just as organizational cultures may differentially affect behavior, people may also differ in the extent to which their behavior is shaped by an organization's culture. For instance, compared to individualists, cooperative people were more responsive to the individualistic or collectivistic norms characterizing their organization's culture, and exhibited greater behavioral adaptation to each across the two types of organizational cultures (Chatman and Barsade, 1995). Thus, congruence models, which presume an additive equivalence of person and situation factors and assume that greater congruence is always better, cannot fully explain behavior in organizations. Instead, a focus on mismatches between person and organization characteristics that challenge people to either act in accordance with the culture and thereby contradict enduring personal tendencies, or vice versa, might generate insight into such interactions.

Future research might focus on the set of characteristics, such as cooperation, self-monitoring, and self-esteem, that contribute to people's flexibility across situations. Identifying such characteristics could improve predictions of the behavioral expression of person characteristics both across time and across situations, and in particular, the extent to which an organization's culture and processes will influence member behavior. Research might investigate other organizationally relevant matches and mismatches to understand how different situations influence person–situation interactions. For example, examining mismatches between honest people

and dishonest organizations may help to identify if and when good people 'turn bad.'

Individuals' Influence Over Organizational Culture

Founders and senior executives, who have legitimacy and authority, may be the most influential individuals in an organization. The person who creates the organization has significant impact on the strategies that the group develops to survive, and these are often preserved in the organization's values, practices, and promotion patterns well past the individuals' actual presence. Ironically, newcomers, who are at the other end of the spectrum in terms of authority and legitimacy from founders, may also exert a great deal of influence on culture. This influence may be indirect; research has shown that the process of recruiting new members, including emphasizing an organization's attractions, strengthens current members' connection to their organization (Sutton and Louis, 1987), and can exert mutual learning of the organizational culture (March, 1991). An organization's culture may also be transformed by the entrance and exit of its members. Thus, the strength of an organization's culture may depend on the individual mobility of its members. Using simulation methods, organizations' length of service distributions have been examined as indicators of the extent to which members have been socialized into the culture's norms and values. This research has shown that variations in service distributions are positively associated with heterogeneity in organizational culture due to three distinct cultural processes: socialization, group cohesiveness, and common historical experiences (Carroll and Harrison, 1998).

Situation Selection Based on Organizational Culture

Schneider (1987) developed the Attraction–Selection–Attrition (ASA) model, which posits that the key factor influencing the relationship between people and organizations is the fit between individual personality and the modal personality represented in the organization. People are differentially attracted to organizations on the basis of the anticipated fit between their own characteristics, such as personality traits, and organizational characteristics, such as culture (e.g., Cable and Judge, 1997). Job seekers take an active role in the recruitment process and are attracted to careers and organizations reflecting their personal values and interests. Further, organizations have developed formal and informal strategies to identify and select recruits who will be compatible with their goals and working environment, elements that are strongly influenced by an organization's cultural values. For instance, rather than focusing on job-related criteria, selection appears to be based on such socially based criteria as 'personal chemistry,' and fit with the organization's values.

Organizational Demography

People may differ from each other in many ways that are both observable and unobservable. Members' observable demographic characteristics constitute a context for every other individual in the organization. For instance, a lone newcomer in an organization of established members may have a very different experience than a newcomer entering an organization characterized by members with varying tenure. Given the dramatic changes in US labor force demography over the past decade, relational demography, or the distribution of demographic characteristics within organizations, has become an active area of research.

Research on organizational demography improves upon past research on employee turnover. Turnover typically indicates a discrete departure event while demography focuses on length of service and rate of departure by taking into account the distribution of people by the length of time they have spent in the organization. This distribution can be influenced by a host of factors including personnel policies, technology, and the degree of unionization in the workforce (Pfeffer, 1983). Most research on observable differences has examined the effect of demographic heterogeneity versus homogeneity on performance and has yielded mixed results (see Williams and O'Reilly, 1998 for a comprehensive review). Some studies have demonstrated the positive effects of demographic heterogeneity for increasing the number of novel perspectives that can be used to solve difficult problems, increasing an organization's network of contacts, and facilitating organizational change. However, demographic heterogeneity may also lead to communication problems, less social integration in workgroups and greater turnover (e.g., Tsui et al., 1992). Demographic distribution among members across various attributes is an important situational factor that deserves further research since it can influence behavior differently depending on an individual's own demographic profile.

Behavioral Consistency Across Demographically Heterogeneous Workgroups

Research suggests that an organization's culture may influence the relationship between demographic diversity and work outcomes. For example, the purported benefits of a demographically diverse workforce are more likely to emerge in collectivistic organizations that make organizational membership more salient than membership in a demographic category (Chatman et al., 1998). An organization's culture may dictate the extent to which members view certain demographic characteristics as valuable and others as associated with lower status within an organization's informal social system (Spataro, 2000). Furthermore, each attribute, such as tenure, race, or sex heterogeneity within a group, may differentially influence individual behavior and the combinations of various attributes can result in 'fault lines' which become stronger as more attributes align themselves in the same way (Lau and Murnighan, 1998).

One explanation for the lack of clarity about the benefits and detriments of diversity is that researchers have neglected to consider key mediating processes between demographic composition and performance. As shown in one recent study, heterogeneous groups initially created norms fostering independence rather than cooperation among members, but cooperative norms subsequently mediated the relationship

between group demographics and performance (Chatman and Flynn, 2001). Similarly, a group's level of conflict influenced the impact of demographic heterogeneity on performance (Jehn et al., 1999).

Leadership

Some Individuals Can Effect Change More Than Others

Early leadership research focused on the physiological and psychological traits thought to be associated with exceptional leaders. These 'great man' theories of leadership examined the effects of personal characteristics such as height, physical appearance, and intelligence on leaders' emergence and effectiveness. This stream of research has its counterpart in more current studies examining the effects of self-confidence, extraversion, and energy level (e.g., House, 1988). The aim of this approach has been to identify a leadership personality. However, it leaves many crucial questions unanswered, such as whether certain personal characteristics become more important than others depending on the organizational context, and why, regardless of formal authority, followers perceive some people as leaders and not others.

Contingency theories of leadership were advanced to explain how certain personal characteristics made a leader effective in certain situations (e.g., House and Baetz, 1979). For example, leaders who initiated structure raised the productivity and satisfaction of a group working on a boring or simple task but lowered the productivity and satisfaction of a group working on a complex task, while a considerate leader raised the satisfaction and productivity of a group engaged in a boring task but had little effect on a group engaged in a task they found intrinsically interesting. Additionally, research showed that allowing members to participate in decision making increased commitment but depended on the amount of trust the leader had in his or her subordinates as well as the urgency of task completion (Vroom and Jago, 1978). Thus, contingency theories of leadership were more comprehensive than trait theories; however, they still did not account for the interactive effects of leader characteristics and their situational contexts.

Recent research has focused on charismatic and transformational leadership, demonstrating that some individuals influence situations more than others. This research takes an interactional approach by conceptualizing leadership as a personal relationship between the leader and his or her followers. A leader must have certain interpersonal skills in order to inspire followers to set aside their goals and to pursue a common vision. Charismatic leaders are thought to have the ability to change their circumstances by increasing followers' motivation and commitment and, sometimes, to change the direction of the entire organization (e.g., Meindl et al., 1985). However, a leader is only charismatic if followers recognize him or her as such; followers must identify with the vision articulated by the leader. In one particularly exhaustive laboratory study of charismatic leadership (Howell and Frost, 1989), confederates were trained to display qualities of a charismatic leader, such as projecting a dominant presence, articulating a large overarching goal, and displaying extreme confidence in followers' ability to accomplish this goal. In addition, norms were created in each group for either high or low productivity. In contrast to participants working under a considerate or structuring leader, participants working under the charismatic leader displayed higher task performance regardless of the group productivity norm. This finding suggests that leaders mold their styles in response to the situation. Moreover, some leaders are capable of changing the situation itself by changing followers' perceptions and motivation.

Leadership as a Function of the Strength of the Situation

Some researchers have been skeptical of a leader's ability to change situations, and have suggested that leadership is far more situationally determined than might have been assumed. The attributional theory of leadership suggests that because people tend to associate certain behaviors with those of a leader, leadership qualities will be attributed to a person displaying these behaviors (Calder, 1977). Various biases emerge from this attribution, however. For instance, individuals tend to overestimate the amount of control a leader has over events that are, in fact, random or uncontrollable (Pfeffer, 1977). Furthermore, a leader will be given more credit when situations are unfavorable (Meindl et al., 1985). Individuals' lay conceptions of leadership can be used or misused for the purposes of organizational impression management (Ginzel et al., 1993). In sum, leadership research has focused more on contexts and followers rather than on the characteristics of a focal leader. Future research might examine whether leaders reflect the personal characteristics of their followers or complement their weaknesses, if some followers have a greater psychological need for leadership than others, and the various substitutes for leadership, or how people can be compelled to lead themselves.

Creativity and Innovation

Research on creativity, like leadership, has moved from emphasizing traits, to considering the organizational context as well as the interaction between the two. Creativity is generally viewed as distinct from innovation. Creativity occurs at the individual level, and refers to the generation of ideas that are both novel and useful. Innovation refers to the process of implementing these ideas at the organizational level.

Some People Are More Creative Across Situations Than Others

Early creativity research focused on the personality traits associated with individuals who had made creative contributions in their respective fields. People who have a wide variety of interests, are attracted to complex tasks, tolerant of ambiguity, and self-confident perform more creatively (e.g., Barron and Harrington, 1981). However, trait research ignores how organizational contexts influence people's ability to perform creatively.

Some Situations Foster More Creativity Than Others

The degree to which one's job encourages intrinsic versus extrinsic motivation affects one's creative performance. Early

studies suggested that when people worked on tasks that they found intrinsically interesting, adding an extrinsic reward lowered their interest in performing the task for its own sake (Deci, 1971). More recent studies in organizational settings showed that individuals were most creative when they were intrinsically motivated. This intrinsic interest led them to stay focused, persist longer to complete difficult tasks, to 'play' with ideas, and suggest novel solutions (Amabile, 1988). This suggests that the situational factors that are associated with decreases in intrinsic motivation, such as a controlling supervisory style, and an emphasis on external rewards, may indirectly diminish people's creative potential.

Organizational culture also influences creativity and innovation. Organizations that have mechanisms to express confidence in members, and communicate this confidence through the culture's core values, increase creativity among members (Kanter, 1988). These findings are supported by a recent ethnography of a product design firm, IDEO, which creates new products by taking technologies from one industry and applying them in other industries where these technologies are unknown (Hargadon and Sutton, 1997). At IDEO, employees are encouraged to create analogies between past technological solutions and current problems and to share them in brainstorming sessions. Further, employees are selected who have unique hobbies or skills that can be used to solve design problems.

Employees who have traits associated with creativity are more likely to thrive in organizations such as IDEO, which place an emphasis on creative performance. For example, employees produced the most creative work when they possessed the appropriate personal characteristics, were working on complex assignments, and were supervised in a supportive, noncontrolling fashion (Oldham and Cummings, 1996). While the possibility that organizations can manage creativity through the use of a strong culture appears promising (Flynn and Chatman, 2001), some worry that mechanisms of social control will stifle, not encourage creativity (Nemeth and Staw, 1989). Future researchers might examine how organizations can achieve harmony and cohesion without sacrificing the flexibility and constructive conflict necessary for creativity and innovation. Many believe that creativity and innovation are the last sustainable competitive advantages and as such these issues will continue to generate a great deal of interest.

Conclusion

As theorists endeavor to develop a complete understanding of behavior in organizations, the analysis of both personal and situational factors, as conjoined units of behavior, will become increasingly fundamental to organizational studies. Person–situation interactions are much more complex than the simple addition of personal and situational characteristics, and these interactions may take a variety of forms. Some people are more responsive to situations than others, some situations can shape behavior to a greater degree than others, and some people have the unique capability to shape situations to their advantage or that of the organization. Furthermore, group members' personal characteristics may constitute the situational context as every individual responds to the personal or demographic characteristics of every other individual in the organization (Carroll and Harrison, 1998; Tsui et al., 1992). These interactions are both complex and diverse: hence the field of organizational behavior must necessarily reflect this diversity. It will become increasingly important to observe these interactions over time. Since people and situations adjust to each other, cross-sectional research will not capture the dynamic interplay between the two. By pursuing interactional research over time, organizational behavior will increasingly evolve into a field that is as vibrant as the organizations it seeks to understand.

See also: Innovation; Leadership in Organizations, Sociology of; Marx, Karl (1818–83); Organization: Overview; Organizational Behavior, Psychology of; Organizational Climate; Organizational Culture, Anthropology of; Organizational Decision Making; Organizations and Culture; Organizations, Sociology of; Organizations: Authority and Power; Schumpeter, Joseph A. (1883–1950); Weber, Max (1864–1920).

Bibliography

Amabile, T., 1988. A model of creativity and innovation in organizations. In: Staw, B.M., Cummings, L.L. (Eds.), Research in Organizational Behavior, vol. 10. JAI Press, Greenwich, CT, pp. 169–212.

Barron, F., Harrington, D.M., 1981. Creativity, intelligence and personality. Annual Review of Psychology 32, 439–476.

Cable, D.M., Judge, T.A., 1997. Interviewers' perceptions of person–organization fit and organizational selection decisions. Journal of Applied Psychology 82, 546–561.

Calder, B.J., 1977. An attribution theory of leadership. In: Staw, B., Salancik, G. (Eds.), New Directions in Organizational Behavior. St. Clair Press, Chicago, pp. 179–204.

Carroll, G., Hannan, M., 2000. The Demography of Corporations and Industries. Princeton Press, Princeton, NJ.

Carroll, G., Harrison, J.R., 1998. Organizational demography and culture: Insights from a formal model and simulation. Administrative Science Quarterly 43, 637–667.

Chatman, J.A., 1991. Matching people and organizations: Selection and socialization in public accounting firms. Administrative Science Quarterly 36, 459–484.

Chatman, J.A., Barsade, S.G., 1995. Personality, organizational culture, and cooperation: Evidence from a business simulation. Administrative Science Quarterly 40, 423–443.

Chatman, J., Flynn, F., 2001. The influence of demographic heterogeneity on the emergence and consequences of cooperative norms in work teams. Academy of Management Journal 44 (5).

Chatman, J.A., Polzer, J.T., Barsade, S.G., Neale, M.A., 1998. Being different yet feeling similar: the influence of demographic composition and organizational culture on work processes and outcomes. Administrative Science Quarterly 43, 749–780.

Davis-Blake, A., Pfeffer, J., 1989. Just a mirage: The search for dispositional effects in organizational research. Academy of Management Review 14, 385–400.

Deci, E.L., 1971. The effects of externally mediated reward on intrinsic motivation. Journal of Personality and Social Psychology 18, 105–115.

Flynn, F., Chatman, J., 2001. Strong cultures and innovation: Oxymoron or opportunity. In: Cartwright, S., Chatman, J., Cummings, T., Earley, P.C., Holden, N., Sparrow, P., Starbuck, W. (Eds.), International Handbook of Organizational Culture and Climate. Wiley, Chichester, UK, pp. 263–288.

Ginzell, L., Kramr, R., Sutton, R., 1993. Organizational impression management as a reciprocal influence process. In: Staw, B.M., Cummings, L.L. (Eds.), Research in Organizational Behavior, Vol. 15. JAI Press, Greenwich, CT, pp. 227–266.

Hargadon, A., Sutton, R.I., 1997. Technology brokering and innovation in a product design firm. Administrative Science Quarterly 42, 716–749.

House, R.J., 1988. Power and personality in complex organizations. In: Staw, B.M., Cummings, L.L. (Eds.), Research in Organizational Behavior, vol. 10. JAI Press, Greenwich, CT, pp. 305–357.

House, R., Baetz, M., 1979. Leadership: Some empirical generalizations and new research directions. In: Staw, B.M., Cummings, L.L. (Eds.), Research in Organizational Behavior, Vol. 1. JAI Press, Greenwich, CT, pp. 341–423.

Howell, J., Frost, P., 1989. A laboratory study of charismatic leadership. Organizational Behavior and Human Decision Processes 43, 243–269.

Jehn, K.A., Northcraft, G.B., Neale, M.A., 1999. Why differences make a difference: A field study of diversity, conflict and performance in workgroups. Administrative Science Quarterly 44, 741–763.

Kanter, R.M., 1988. When a thousand flowers bloom: Structural, collective and social conditions for innovation in organizations. In: Staw, B.M., Cummings, L.L. (Eds.), Research in Organizational Behavior, vol. 10. JAI Press, Greenwich, CT, pp. 169–211.

Kristof, A.L., 1996. Person–organization fit: An integrative review of its conceptualizations, measurement, and implications. Personality Psychology 49, 1–49.

Lau, D., Murnighan, J.K., 1998. Demographic diversity and faultlines: The compositional dynamics of organizational groups. Academy of Management Review 23, 325–340.

March, J.G., 1991. Exploration and exploitation in organizational learning. Organization Science 2, 71–87.

Martin, J., 1992. Culture in Organizations: Three perspectives. Oxford University Press, New York.

Meindl, J., Ehlrich, S., Dukerich, J., 1985. The romance of leadership. Administrative Science Quarterly 30, 78–102.

Morrison, E.W., 1993. Longitudinal study of the effects of information seeking on newcomer socialization. Journal of Applied Psychology 78, 173–183.

Nemeth, C.J., Staw, B.M., 1989. The tradeoffs of social control and innovation in small groups and organizations. Advances in Experimental Social Psychology 22, 175–210.

Oldham, G.R., Cummings, A., 1996. Employee creativity: Personal and contextual factors at work. Academy of Management Journal 39, 607–634.

O'Reilly, C.A., Chatman, J.A., Caldwell, D.M., 1991. People and organizational culture: A Q-Sort approach to assessing person–organization fit. Academy of Management Journal 34, 487–516.

Pfeffer, J., 1977. The ambiguity of leadership. Academy of Management Review 2, 104–112.

Pfeffer, J., 1983. Organizational demography. In: Staw, B.M., Cummings, L.L. (Eds.), Research in Organizational Behavior, vol. 5. JAI Press, Greenwich, CT, pp. 299–357.

Rynes, S., Gerhart, B., 1990. Interviewer assessments of applicant 'fit': an exploratory investigation. Personnel Psychology 43, 13–35.

Schneider, B., 1987. The people make the place. Personnel Psychology 40, 437–453.

Scott, R.W., 1998. Organizations: Rational, Natural and Open Systems, fourth ed. Prentice Hall, Englewood Cliffs, NJ.

Snyder, M., 1987. Public Appearances, Private Realities: The Psychology of Self-monitoring. WH Freeman, New York.

Spataro, S., 2000. Not all differences are the same: The role of status in predicting reactions to demographic diversity in organizations. Unpublished doctoral dissertation, University of California, Berkeley, CA.

Staw, B.M., Ross, J., 1985. Stability in the midst of change: The dispositional approach to job attitudes. Journal of Applied Psychology 70, 469–480.

Sutton, R.I., Louis, M.R., 1987. How selecting and socializing newcomers influences insiders. Human Resource Management 26, 347–361.

Tsui, A.S., Egan, T.D., O'Reilly, C.A., 1992. Being different: Relational demography and organizational attachment. Administrative Science Quarterly 37, 549–579.

Vroom, V.H., Jago, A.G., 1978. On the validity of the Vroom–Yetton model. Journal of Applied Psychology 63, 151–162.

Williams, K.Y., O'Reilly, C.A., 1998. Demography and diversity in organizations: a review of 40 years of research. In: Staw, B.M., Cummings, L.L. (Eds.), Research in Organizational Behavior, vol. 20. JAI Press, Greenwich, CT, pp. 77–140.

Perception of Action Space: Using Multiple Frames of Reference

Carl Gabbard, Texas A&M University, College Station, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C.L. Colby, volume 16, pp. 11205–11209, © 2001, Elsevier Ltd.

Abstract

Our ability to move through and act on the environment around us is driven by, seemingly effortless, yet quite complex systems. The focus of this article is on the visuospatial frames of reference that are used to plan and execute actions in peripersonal and extrapersonal space. The particular representation of space in use at any time depends on the task the individual is trying to perform. Spatial locations of objects in the surrounding environment can be perceived based on observer-centered coordinates and on object-centered coordinates known, respectively, as egocentric and allocentric frames of reference. The discussion includes basic terms and a review of selected research on neural underpinnings, and typical and atypical behaviors.

Most of us take for granted our ability to navigate through the environment and when needed or intending, to interact with objects. For example, the seemingly effortless way in which we reach for an object belies the complex computations that go into determining its location. The psychological and neural processes that underlie our visuospatial abilities have been extensively explored in humans and animals, but to this day remain somewhat of a mystery. This article describes selected psychological and neural aspects of spatial perception with particular focus on how we use visual frames of reference to guide our actions.

Basic Terms Associated with Perceiving and Understanding the Space around Us

Spatial Cognition

Underscoring the perception and action on objects in space involves a degree of *spatial cognition.* Although several definitions have been suggested, the general assumption is that it involves representing, organizing, understanding, and interacting with stimuli in the environment.

Spatial cognition involves not one but many specific abilities. Within the visual domain these include locating points in space, determining the orientation of lines and objects, assessing location in depth, appreciating geometric relations between objects, and processing motion, including motion in depth. These spatial skills can be applied to mentally representing (imagining) objects as well as acting on external stimuli (*see* Mental Imagery, Visual Cognition).

According to Burgess (2008), spatial cognition embraces an ensemble of representations supporting visuospatial perception, mental imagery, spatial memory, and navigation. *Visuospatial perception* refers to the ability to process and interpret visual information about the locations of objects in space. It underlies the ability to navigate in an environment, orient appropriately, and plan actions, such as reaching for objects. Visuospatial information can be either categorical or coordinate; categorical representations are abstract and propositional, whereas coordinate relations specify metric information on the relative location and distance of objects (Ruggiero et al., 2012; van der Ham and Borst, 2011) (also *see* Human Spatial Orientation, Neural Basis of, Visual Perception, Neural Basis of).

Peripersonal and Extrapersonal Space

Effective action planning and execution, such as when intending to reach for an object, requires an integrated neural representation of the body and of the space surrounding the body. *Peripersonal* (near) space is behaviorally defined as the space within the arm-reaching distance, whereas *extrapersonal* (far) space represents the area outside the arm-reaching distance. The coding of space as near and far is not only determined by the arm-reaching distance, but it is also dependent on how the brain represents the extension of the body space (Berti and Frassinetti, 2000). Closely associated with the notion of space representation are the findings from animal research on specialized visual neurons coded for the detection of near space (e.g., Iriki et al., 1996; Làdavas, 2002). In humans, a functionally homologous coding of peripersonal space is largely supported by behavioral studies, showing stronger visual–tactile interaction in near than far space in brain damaged patients (Brozzoli et al., 2006) and healthy individuals (see review by Macaluso and Maravita, 2010).

One line of research that has tested the perceived boundaries of peri- and extrapersonal space is the examination of tool use. There is a general consensus from research that a tool extends an individual's peripersonal space. Complementing this modulation is a 'remapping' of extrapersonal to peripersonal space. For example, Iriki et al. (1996) found neurons in the intraparietal sulcus of monkeys that fired when a raisin was presented within arm's reach but not beyond. The monkeys were then taught to reach with a rake, which extended their reach. The so-called 'reachability neurons' adapted to this change and responded to raisins that were presented further away but within reach with the rake. This research suggests that there exists visual neurons that code for what is within reach and that these neurons adapt to changes in reachability resulting from tool use. This general finding of remapping of far space as near space via tool use has also been demonstrated

with healthy adults and with patients showing spatial neglect (e.g., Cardinali et al., 2009; Neppi-Mòdona et al., 2007).

Space as Represented in Visual Frames of Reference

Egocentric and Allocentric Frames

In an excellent review by Wang (2012), the writer notes that the issue of reference frames is central to our understanding of spatial representations. Research findings are clear that there is not a single, explicit, topographic representation of space suitable for incorporating every kind of sensory input and generating every kind of motor output. On the contrary, evidence points to multiple representations of space, in a variety of coordinate frames, and linked to separate output systems designed to guide specific motor effectors.

The particular representation of space in use at any time depends on the task the individual is trying to perform. Spatial locations of objects in the surrounding environment are not only perceived based on an observer-centered (egocentric) coordinate, but also on an object-centered (allocentric) coordinate. For example, if you were to draw out a route for a hiking trip on a map, the route would be in the coordinates of the map, the piece of paper. If you were then going to walk along that route, you would have to begin by locating your current position within that coordinate frame and constructing a representation of the route with respect to your starting point. The first map-based representation is an example of an *allocentric representation* (framework), in which locations are represented in reference frames extrinsic to the observer. Allocentric representations include those centered on the object(s) of interest (*object-centered*) and those in environmental (room-centered or world-centered) coordinates.

The second representation, the one in the coordinates of the hiker's current position, is an example of an *egocentric representation*, in which locations are represented relative to the observer. Egocentric representations include those in eye-centered, head-centered, hand-centered, and body-centered coordinates. Most scientists would agree that multiple reference frames can be activated simultaneously when performing action tasks. For example, when reaching, allocentric cues are used to code specific characteristics of the object (e.g., shape, size, and location) and its surroundings. Then using an egocentric frame of reference the individual maps his or her body including the effector (hand) information with that of the reaching object.

Visual Processing Streams

Human brain imaging studies confirm that different cortical regions are active when participants process ego- or allocentric information. It has been proposed that egocentric cues associated with intended action are processed 'primarily' in dorsal stream brain areas, whereas allocentric representations are found mainly along the ventral processing stream. Moreover, specific propositions of the two-visual-system hypothesis (visual-for-perception and vision-for-action; Goodale et al., 2005) and behavioral data support the idea that for estimation judgments (as with reaching actions), the dorsal stream via egocentric reference is 'metrically' more accurate than the ventral stream (allocentric reference). Recent work also supports the idea that each hemisphere preferentially processes and integrates contralateral egocentric and allocentric stimuli (Zhou et al., 2012).

It should be acknowledged that, although the consensus of research leads to the notion that the two visual pathways have distinct roles, for action process, the two systems are not hermetically sealed from one another – perception and action are intimately linked. "After all, many actions, which are mediated by mechanisms in the dorsal stream, are conditional upon the presence of complex stimuli that can be interpreted only by mechanisms in the ventral perceptual stream" (Goodale et al., 2005: p. 274).

Moreover, the perception and action streams are speculated to operate under very different temporal constraints. That is, real-time (no-delay) movements depend on pathways from the early visual areas via relatively encapsulated visuomotor mechanisms in the dorsal stream. These dedicated visuomotor mechanisms, together with motor centers in the premotor cortex and brainstem, compute the absolute metrics of the target object and its position in the egocentric coordinates of the effector used to perform the action. In contrast, it is suggested that memory-driven actions make use of a perceptual representation of the target object generated by the ventral stream. Research testing the temporal aspects of the visual systems is supportive (Graham et al., 1998; Hu et al., 1999). Furthermore, extensive neuroimaging and neurophysiology work is supportive of the separation of the two streams (e.g., Cohen et al., 2009) (*see* Visual Streams: Dorsal and Ventral; Visual Perception, Neural Basis of).

Impairments Associated with Spatial Perception

Much of our knowledge about the neural basis of spatial perception comes from observations in patients with spatial deficits following brain damage. These include a wide range of perceptual and motor deficits, such as: poor localization of visual, auditory, or tactile stimuli; inability to determine visual or tactile line orientation; impaired performance on mazes; impairment on tests of mental spatial transformations; right–left confusion; poor drawing; impaired eye movements to points in space; misreaching; defective locomotion in space; and amnesia for routes and locations. As can be seen from this partial list, spatial behavior involves many kinds of skills and, not surprisingly, a number of brain regions have been implicated in spatial perception and performance. Broadly speaking, the parietal lobe is responsible for spatial perception and representation of immediate extrapersonal space, while temporal and parahippocampal cortices are more involved in topographic memory and navigation (*see* Navigation in Virtual Space: Psychological and Neural Aspects). The frontal lobe receives input from both parietal and temporal cortex and is responsible for generating actions.

An important point about deficits in spatial perception is that they are far more common after right hemisphere damage than left. While patients with left hemisphere damage may also exhibit spatial deficits, it is clear that the right hemisphere in humans has a superordinate role in spatial processing and behavior (Heilman et al., 1985). Two kinds of spatial deficits

following brain damage are particularly illuminating. First, a common sensorimotor deficit following parietal lobe damage is a difficulty in using visual information to guide arm movements, referred to as optic ataxia. Patients with optic ataxia have difficulty with everyday tasks that require accurate reaching under visual guidance, such as using a knife and fork. They both misdirect the hand and misorient it with respect to the object, and are most impaired when using the contralesional hand to reach for an object in the contralesional half of space (*see* Visual Streams: Dorsal and Ventral).

A second, classic disorder of spatial perception in humans is the tendency to ignore one half of space, called hemispatial neglect (*see* Hemispatial Neglect, Neural Basis of). The most common form of neglect arises from damage to the right parietal lobe and is manifested as a failure to detect objects in the left half of space. Neglect is more than just a visual deficit, however. It can occur separately or jointly across many sensory modalities. Moreover, neglect occurs with respect to many different spatial reference frames. A patient with right parietal lobe damage typically neglects objects on the left but left may be defined with respect to the body, or the line of sight, or with respect to an attended object. Further, this neglect is dynamic, changing from moment to moment with changes in body posture and task demands (Behrmann, 2000). Neglect is apparent even in the purely conceptual realm of internal images. Patients exhibit neglect when asked to imagine a familiar scene, such as a city square, and describe the buildings in it. The portion of space that is neglected changes when individuals are asked to imagine the scene from a different viewpoint. As this example illustrates, neglect can occur with respect to an internal image constructed by the individual. Patients with neglect show evidence of using multiple reference frames, just as intact individuals do.

One of the lesser-known disorders associated with difficulty with visuospatial tasks including spatial navigation is Williams syndrome. Recent work by Bernandino et al. (2013) found that these individuals also had significant impairment in making egocentric and allocentric spatial judgments.

Neural Aspects

Parietal Neurons Encode Spatial Information in Multiple Reference Frames

The standard approach for investigating the role of parietal neurons in spatial perception is to record electrical activity from individual neurons while the monkey performs a spatial task. Because brain tissue itself has no sensory receptors, fine microelectrodes can be introduced into the brain without disturbing the animal's performance. By recording neural responses during carefully designed tasks, neural activity can be related directly to the sensory and representational processes that underlie spatial behavior. The general conclusion from these studies is that the function of parietal cortex is to transform spatial information from sensory coordinates into the motor coordinates that are necessary for the guidance of action (Andersen et al., 1997; Colby and Goldberg, 1999) (*see* Human Spatial Orientation, Neural Basis of) (*see* Parietal Lobe).

Head-centered spatial representation in the ventral intraparietal area (area VIP). The area VIP is located in the floor of the intraparietal sulcus, where inputs from high-order visual and somatosensory cortex converge. In the visual domain, VIP neurons are characterized by direction and speed selectivity, and thus resemble neurons in other dorsal stream visual areas that process stimulus motion (*see* Neural Representation of Direction (Head Direction Cells)).

In the somatosensory domain, these same neurons respond to light touch on the head and face. The somatosensory and visual receptive fields of individual neurons exhibit strong spatial correspondence: they match in location, in size, and even in their preferred direction of motion. The existence of spatially matched receptive fields raises an interesting question: what happens when the eyes move away from primary position? If the visual receptive fields were simply retinotopic, they would have to move in space when the eyes do and so would no longer match the location of the somatosensory receptive field. Instead, for some VIP neurons, the visual receptive field moves to a new location on the retina when the eyes move away from the straight-ahead position. For example, a neuron that has a somatosensory receptive field near the mouth and responds best to a visual stimulus moving toward the mouth will continue to respond to that trajectory of motion regardless of where the monkey is looking. These neurons have head-centered receptive fields; they respond to stimulation of a certain portion of the skin surface and to the visual stimulus aligned with it, no matter what part of the retina is activated. Neurons in area VIP send projections to the region of premotor cortex that generates head movements. Area VIP neurons thus contribute to the visual guidance of head movements and may play a special role in hand, eye, and mouth coordination. They operate in a head-centered reference frame in order to generate appropriate signals for a particular motor effector system, namely that which controls head movements.

Eye-centered spatial representation in the lateral intraparietal area (area LIP). In contrast to area VIP, neurons in area LIP construct an eye-centered spatial representation. Individual neurons become active when a salient event occurs at the location of the receptive field. This can be a sensory event, such as the onset of a visual or auditory stimulus, or a motor event, such as a saccade toward the receptive field, or even a cognitive event, such as the expectation that a stimulus is about to appear, or the memory that one has recently appeared. The level of response reflects the degree to which attention has been allocated to the site of the receptive field (Colby et al., 1995).

Again, as we saw for area VIP, the animal's ability to make eye movements raises an interesting question about spatial representation. Neural representations of space are maintained over time, and the brain must solve the problem of how to update these representations when a receptor surface moves. With each eye movement, every object in our surroundings activates a new set of retinal neurons. Despite this constant change, we perceive the world as stable. Area LIP neurons contribute to this perceptual stability by using information about the metrics of the eye movement to update their spatial representation of salient locations (Duhamel et al., 1992). For example, LIP neurons are activated when the monkey makes an eye movement that brings a previously illuminated screen location into their receptive field. These neurons respond to the

memory trace of the earlier stimulus: no stimulus is ever physically present in the receptive field, either before or after the saccade. The proposed explanation for this surprising finding is that the memory trace of the stimulus is updated at the time of the saccade. Before the saccade, while the monkey is looking straight ahead, the onset of the stimulus activates a set of LIP neurons whose receptive fields encompass the stimulated screen location, and they continue to respond after the stimulus is extinguished, maintaining a memory trace of the stimulus. At the time of the saccade, a corollary discharge, or copy of the eye movement command, containing information about the metrics of the saccade, arrives in parietal cortex. This corollary discharge causes the active LIP neurons to transmit their signals to the new set of LIP neurons whose receptive fields will encompass the stimulated screen location after the saccade. The representation of the stimulated location is thus dynamically updated from the coordinates of the initial eye position to the coordinates of the final eye position. The significance of this observation is in what it reveals about spatial representation in area LIP. It demonstrates that the representation is dynamic and is always centered on the current position of the fovea.

Instead of creating a spatial representation that is in purely retinotopic (sensory) coordinates, tied exclusively to the specific neurons initially activated by the stimulus, area LIP constructs a representation in eye-centered (motor) coordinates. The distinction is a subtle one but critical for the ability to generate accurate spatial behavior. By representing visual information in eye-centered coordinates, area LIP neurons tell the monkey not just where the stimulus was on the retina when it first appeared but also where it would be now if it were still visible. The result is that the monkey always has accurate spatial information with which to program an eye movement toward a real or remembered target. The transformation from sensory to motor coordinates puts the visual information in its most immediately useful form. Compared to a head-centered or world-centered representation, an eye-centered representation has the significant advantage that it is already in the coordinates of the effector system that will be used to acquire the target. Humans with unilateral parietal lobe damage fail on an eye movement task that requires an eye-centered representation of a remembered target position. This failure presumably reflects an impairment of the updating mechanism in parietal cortex.

In sum, posterior parietal cortex plays a critical role in spatial perception and representation. Physiological studies in monkeys show that parietal neurons represent spatial locations relative to multiple reference frames, including those centered on the head and the eye. Individual neurons (in area VIP) combine spatial information across different sensory modalities, and specific spatial reference frames are constructed by combining sensory and motor signals (in area LIP). In accord with the physiology, human neuropsychological studies show that neglect can be expressed with respect to several different reference frames.

Frontal Lobe Mechanisms of Spatial Representation

The parietal lobe transforms sensory representations of attended objects into the motor coordinate frames most appropriate for action. They do not actually generate those actions. Dorsal stream outputs to frontal cortex provide the sensory basis, in the correct spatial framework, for producing specific motor outputs. The following sections describe specific spatial representations found in three regions of frontal cortex.

Neurons in premotor cortex have head-centered and hand-centered visual receptive fields. Two different forms of spatially organized visual responsiveness have been described in ventral premotor cortex. First, in subdivisions representing facial movements, neurons respond to visual stimuli at specific locations relative to the head, much like those described above in area VIP. These neurons have been characterized by recording activity while objects approach the monkey's head along various trajectories. The preferred trajectory is constant with respect to the head, and is not affected by changes in eye position (Fogassi et al., 1996). This specific zone of premotor cortex receives inputs from area VIP, and uses the same form of spatial representation. In both cases, stimuli are represented in the motor coordinate frame that would be most useful for acquiring or avoiding stimuli near the face by means of a head movement.

A different kind of spatial representation has been observed in subdivisions of premotor cortex involved in generating arm movements. Here, neurons respond to visual stimuli presented in the vicinity of the hand. When the hand moves to a new location, the visual receptive field also moves with it. Moreover, the visual receptive field remains fixed to the hand regardless of where the monkey is looking, suggesting the existence of a hand-centered representation (Graziano et al., 1994). A fascinating observation indicates that some neurons in this area are capable of even more abstract forms of representation. Rizzolatti and coworkers have described neurons in premotor cortex that are activated both when the monkey grasps an object, such as a raisin, and when the monkey observes the experimenter performing the same action (Gallese et al., 1996). These 'mirror' neurons encode not just the monkey's own motor actions but also the meaning of actions performed by others (*see* Premotor Cortex).

Object-centered spatial representation in the supplementary eye field (SEF). Actions are directed toward objects in the environment and toward specific locations on an object. Picking up your coffee cup requires that you locate both the cup in egocentric space and the handle in relation to the cup. The spatial reference frame that guides such movements is not limited to the egocentric representations described above. Evidence from the SEF demonstrates that single neurons can encode movement direction relative to the object itself (Olson and Gettner, 1995). The SEF is a division of premotor cortex with attentional and oculomotor functions. Neurons here fire before and during saccades. In monkeys that are trained to make eye movements to particular locations on an object, SEF neurons exhibit a unique form of spatial selectivity: they encode the direction of the impending eye movement as defined relative to an object-centered reference frame. For example, a given neuron may fire when the monkey looks toward the right end of a bar placed at any of several different locations on the screen, regardless of whether the eye movement itself is a rightward or leftward saccade. Moreover, the same neuron will fail to respond when the monkey makes a physically identical eye movement toward the left end of the

bar stimulus. This striking result indicates that single neurons can make use of quite abstract spatial reference frames. Object-centered spatial information could potentially guide arm movements as well as eye movements. Moreover, neuropsychological evidence indicates that an object-centered reference frame can be used to direct attention: some patients exhibit object-centered neglect after parietal lobe damage (Behrmann, 2000). Parietal lobe lesions in monkeys likewise produce impairments on tasks that require an object-centered spatial representation (*see* Neural Representations of Objects).

In sum, behavioral and neuropsychological studies indicate that we make use of multiple spatial representations in the perception of action space. Neurophysiological studies are beginning to uncover the neural mechanisms underlying the construction of these egocentric and allocentric spatial representations.

See also: Cortical Areas Engaged in Movement: Neuroimaging Methods; Eye Movement, Control of: Oculomotor Control; Human Spatial Orientation, Neural Basis of; Mental Imagery: Visual Cognition; Motor Cortex; Navigation in Virtual Space: Psychological and Neural Aspects; Neural Representations of Objects; Parietal Lobe; Premotor Cortex; Spatial Navigation; Visual Perception, Neural Basis of; Visual Streams: Dorsal and Ventral.

Bibliography

Andersen, R.A., Snyder, L.H., Bradley, D.C., Xing, J., 1997. Multimodal representation of space in the posterior parietal cortex and its use in planning movements. Annual Review of Neuroscience 20, 303–330.

Behrmann, M., 2000. Spatial reference frames and hemispatial neglect. In: Gazzaniga, M. (Ed.), The Cognitive Neurosciences. MIT Press, Cambridge, MA, pp. 651–666.

Bernandino, I., Mouga, S., Castelo-Branco, M., Van Aselen, M., 2013. Egocentric and allocentric, spatial representation in Williams syndrome. Journal of the International Neuropsychological Society 19 (1), 54–62.

Berti, A., Frassinetti, F., 2000. When far becomes near: remapping of space by tool use. Journal of Cognitive Neurscience 12 (3), 415–420.

Brozzoli, C., Dematte, M.L., Pavani, F., Frassinetti, F., Farne, A., 2006. Neglect and extinction: within and between sensory modalities. Restorative Neurological Neuroscience 24 (4/6), 217–232.

Burgess, N., 2008. Spatial cognition and the brain. Annals of the New York Academy, of Sciences 1124, 77–97.

Cardinali, L., Frassinetti, F., Brozzoli, C., Urquizar, C., Roy, A.C., et al., 2009. Tool-use induces, morphological updating of the body schema. Current Biology 19 (12), R478–R479.

Colby, C.L., Duhamel, J.R., Goldberg, M.E., 1995. Oculocentric spatial representation in parietal cortex. Cerebral Cortex 5, 470–481.

Colby, C.L., Goldberg, M.E., 1999. Space and attention in parietal cortex. Annual Review of Neuroscience 22, 319–349.

Cohen, N.R., Cross, E.S., Tunik, E., Grafton, S.T., Culham, J.C., 2009. Ventral and dorsal stream, contributions to the online control of immediate and delayed grasping: a TMS approach. Neuropsychologia 47 (6), 1553–1562.

Duhamel, J.R., Colby, C.L., Goldberg, M.E., 1992. The updating of the representation of visual space in parietal cortex by intended eye movements. Science 255, 90–92.

Fogassi, L., Gallese, V., Fadiga, L., Luppino, G., Matelli, M., Rizzolatti, G., 1996. Coding of peripersonal space in inferior premotor cortex (area F4). Journal of Neurophysiology 76, 141–157.

Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G., 1996. Action recognition in the premotor cortex. Brain 119, 593–609.

Goodale, M.A., Kroliczak, G., Westwood, D.A., 2005. Dual routes to action: contributions of the dorsal and ventral streams to adaptive behavior. Progress in Brain Research 149, 269–283.

Graham, J., Bradshaw, M.F., Davis, A., 1998. The effect of pre-movement delays on pointing accuracy in middle childhood. Perception 27, 1379–1389.

Graziano, M.S.A., Yap, G.S., Gross, C.G., 1994. Coding of visual space by premotor neurons. Science 266, 1054–1056.

Heilman, K.M., Watson, R.T., Valenstein, E., 1985. Neglect and related disorders. In: Heilman, K.M., Valenstein, E. (Eds.), Clinical Neuropsychology. Oxford University Press, Oxford, UK, pp. 243–293.

Hu, Y., Eagleson, R., Goodale, M.A., 1999. The effects of delay on the kinematics of grasping. Experimental Brain Research 126, 109–116.

van der Ham, I.J.M., Borst, G., 2011. Individual differences in spatial relation processing:, effects of strategy, ability, and gender. Brain and Cognition 76, 184–190.

Iriki, A., Tanaka, M., Iwamura, Y., 1996. Attention-induced neuronal activity in the monkey somatosensory cortex revealed by pupillometrics. Neuroscience Research 25 (2), 173–181.

Làdavas, E., 2002. Functional and dynamic properties of visual peripersonal space. Trends in Cognitive Science 6 (1), 17–22.

Macaluso, E., Maravita, A., 2010. The representation of space near the body through touch and vision. Neuropsychologia 48, 782–795.

Neppi-Mòdona, M., Rabuffetti, M., Folegatti, A., et al., 2007. Bisecting lines with different tools in right brain damaged patients: the role of action programming and sensory feedback in modulating spatial remapping. Cortex 43 (3), 397–410.

Olson, C.R., Gettner, S.N., 1995. Object-centered direction selectivity in the macaque supplementary eye field. Science 269, 985–988.

Ruggiero, G., Ruotolo, F., Iachini, T., 2012. Egocentric/allocentric and coordinate/categorical, haptic encoding in blind people. Cognitive Processing 13 (Suppl. 1), S313–S317.

Wang, R.L., 2012. Theories of spatial representations and reference frames: what can, configuration errors tell us? Psychonomic Bulletin Review 19, 575–587.

Zhou, Y., Liu, Y., Zhang, M., 2012. Asymmetric influence of egocentric representation onto allocentric perception. Journal of Neuroscience 24 (3), 8354–8360.

Relevant Websites

http://www.nmr.mgh.harvard.edu/mkozhevnlab/?page_id=308 – Allocentric vs. Egocentric Spatial Processing. Harvard University Imagery Lab.

http://www.sciencemag.org/content/324/5925/403 – Mirror Neurons Differentially Encode the Peripersonal and Extrapersonal Space of Monkeys.

http://www.shockmd.com/2009/03/27/the-neuroscience-of-interpersonal-space/ – The Neuroscience of Interpersonal Space.

http://www.sciencedaily.com/releases/2010/01/100106193434.htm – Seeing with Your Hands? Hand Amputation May Result in Altered Perception around the Hands.

http://psychology.uwo.ca/faculty/goodale/research/ – Visual Processing Streams, Goodale Lab.

Performance: Anthropological Aspects

Felicia Hughes-Freeland, London University, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Anthropological approaches to performance include the analysis of behavior in social interactions; structured movement and speech language in rituals, acting, artistry, and aesthetics in staged theatrical performances; public events and their mediation; and political symbolism. Anthropologists use performance heuristically to understand social and cultural practices, drawing on concepts from ritual analysis, sociolinguistics, and philosophy. Key questions concern the relationship between text and performance, problems presented by the illusory aspects of performance in Western traditions for understanding social realities, and how to conceptualize the relationship between performance and ritual.

In anthropology, performance is a concept used to analyze a distinctive event or aspects of behavior and identity in all kinds of social contexts and situations. Performance was subsumed into ritual until the concepts of dramaturgical models were developed in the United States by Erving Goffman and led to the concept of 'cultural performances.' Two members of the British 'Manchester School,' Victor Turner and Abner Cohen, were influential in promoting performance as a means to understand the expressive and political aspects of social life. Communicative performativity developed in sociolinguistics has also been influential and has also shaped Judith Butler's account of gender as performative, which in turn has influenced anthropological thinking. Performance has produced more embodied forms of research, and Turner's theory in particular has stimulated interdisciplinary and intercultural dramatic projects. Anthropologists also produce in-depth ethnographic studies of performance to better understand the social processes of community building in religious and secular contexts, politics, resistance, conflict resolution, healing, creativity, and aesthetics. The key problem remains the question of boundaries. Is performance a discrete category of behavior or a fundamental aspect of all behavior? Is performance fundamentally deceitful? Is it determined by preexistent scripts and texts, or is it creative and emergent? Is it even applicable or relevant cross-culturally? What, if anything, does 'performativity' add to an account of social action? Anthropology uses performance to understand the varieties of human imagination and artistry in a wide range of different contexts, latterly including tourism and heritage. In the future, it is likely to be applied to the social engagement in the fields of information communication and technology and to understanding any form of professional practice, including medicine, law, finance, and education.

Anthropologists apply the concept of performance to distinctive events or aspects of behavior to contribute to the understanding of human behavior in any social context. These range from the most extraordinary and stylized forms of human behavior to the fundamental routines of everyday life, and address different areas of specialized professionalism and expertise, in ritual, commercial, economic, medical, technological, dramatic, artistic, creative, and sportive spheres. Initially driven by a theoretical concern with social function and meaning, contemporary research uses the performance concept to address issues of identity and the relationship between performance, self, researcher, time, and place, explicitly in relation to ritual, theater, and cultural politics, and implicitly in general ethnographic analysis.

Development of the Concept

Performance as an anthropological concept developed after the 1950s. Previously, anthropologists subsumed performance into the analysis of ritual, although researchers such as Gregory Bateson and Terence Ranger were already highlighting its analytical and empirical significance. British social anthropology had built its reputation as a science of total functioning systems and structures centered on kinship networks, and embodied forms of human expression, such as acting or dancing in theatrical and commercial venues, were mostly considered marginal to the discipline. By the mid-1960s, the broader humanist concepts of culture that characterized European and American anthropology gave the concept of performance a more central role in both social and cultural anthropological analyses.

The American sociologist Erving Goffman developed a model in which social interaction was explicable as acts and performances: "The world, in truth, is a wedding," he declared (Goffman, 1956: p. 23), having defined performance as "all of the activity of a given participant on a given occasion which serves to influence in any way any of the other participants" (1956: p. 8). Goffman coined the phrase 'impression management' and demonstrated the centrality of etiquette, tact, evasion, and white lies in micro-interactions. He employed theatrical metaphors to explain different aspects of behavioral presentation and 'multiple versions of reality.' All behavior occurs in stage-like social contexts: the public face of social behavior is 'frontstage' and more informal interactions are characterized as 'backstage.' This model was intended as a fifth perspective to complement existing technical, political, structural, and cultural analyses, subsequently integrated into his work on frame analysis. His dramaturgical and transactional vision remains influential in anthropological analyses of social interaction.

Milton Singer, Dell Hymes, Clifford Geertz, and Victor Turner have also shaped the ways in which anthropologists use

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12127-0

performance notions metaphorically to explain social action as 'cultural performance,' a complex system of symbolic and meaningful behavior that produces an ordered understanding of the world. Performance in these terms is not limited to representation but is an active and creative precondition for social life.

Victor Turner's work has been particularly influential on current uses of performance. Turner's approach is collectivist in contrast to Goffman's individualism; he moved from a functionalist study of ritual to cultural performance models and finally to theatrical performance itself. Yet performance was never separated completely from ritual, defined as a "transformative performance revealing major classifications, categories, and contradictions of cultural process" (Ronald L. Grimes, cited in Turner, 1987: p. 75). Turner's theory of ritual can be summed up by three key explanatory concepts. 'Ritual process' consists of three phases, separation, transition or liminality, and reintegration. Conceived by the Dutch structural anthropologist Van Gennep in the early twentieth century as a universal theory to explain rites of passage, Turner applied it to male initiation rites among the Ndembu. In this ethnographic study he also developed the second concept, 'social drama,' which explain social conflicts as having a threefold narrative dramatic structure – problem, confrontation, resolution – with people's actions being determined by their 'roles.' Thirdly, after having been influenced by Gluckman's functionalist theory that conflict ultimately gives way to harmony, Turner developed the argument that ritual action has the potential to release humans from the determining structure of their everyday life into a creative and liberating 'anti-structure,' or 'communitas.' This was the liminal phase of the ritual process that contrasted with the structure of everyday life. Before his untimely death in 1983, Turner extended liminality beyond the process of a specific ritual event and reconceptualized it as 'liminoid' to describe artistry and creativity in society, including theatrical performances, such as Noh in Japan (Turner, 1987: pp. 99–122). In this way, his conceptual framework came to merge the borders between initiation rituals and social events such as carnivals, films, and spectacles into cultural performances (Turner, 1987: pp. 21–74).

Turner has been one of the most widely influential of Max Gluckman's Manchester School of anthropology, partly because he worked in the United States. Other anthropologists of this background have also produced actor-centered transactional and conflict models using performance perspectives on more political topics, often drawing on Marxist theory. In his study of London's Notting Hill Carnival, Abner Cohen (1993: p. ix) reaffirms his interest in performance as part of a wider analysis of the relationship between cultural forms and political formations. Performance in political symbolism is interesting because it expresses the contradictions between the two dimensions of social interaction: the unintentional and moral nature of cultural action, and the amoral self-interest of political action (Cohen, 1993: p. 8). Cohen's work is recognized for highlighting the ritual dimension of political systems in general, but it has also influenced the analysis of performance as a political tool in postcolonial societies dominated by processes of retribalization and issues of ethnic identity; it is also relevant to the analysis of multiculturalism and cultural politics.

Other salient influences have come from sociolinguistics and linguistic anthropology. Theories that characterized communication as performative as well as informative were inspired by the Russian linguist Roman Jakobson's concept of 'poetic function' – "the way in which a communication is carried out" (Bauman, 1986: p. 3) – and J.L. Austin and John Searle's attention to performativity in speech act theory. For instance, Richard Bauman (1986) has analyzed American folk narratives using a framework of story, performance, and event, and has greatly influenced research into oral tradition and oratory as performance. This specialized use of performance has also been very influential in establishing the performative dimensions of other kinds of communication, particularly in the study of competence and play. Anthropologists such as James Fernandez have also developed communicative approaches to the performance of metaphor in various social contexts. Such work on verbal performance has influenced general theories of performance in human behavior. The concept of performativity has also shaped Judith Butler's influential analysis of gender, which in turn has fed back into some anthropological accounts of the performance of gender and sexuality in everyday life.

Methodological Innovations

Anthropology has always been a discipline that attracts people with an experience of life outside academe, and the 1980s saw the emergence of performer-anthropologists, scholars who had previously worked on the stage professionally as dancers and actors. Anthropologists who have not been professional actors or dancers may choose to include training in performance as part of their research, using their bodies to explore familiar or unfamiliar techniques self-consciously by learning alongside aspiring professionals and experts. Increasing attention is being paid to different performance skills and genres, not simply as objects of study but also methodologically. This has contributed to critiques of Cartesian dualism in accounts of human action and the long-overdue recognition of the importance of embodied behavior and what Marcel Mauss conceptualized in the 1930s as 'techniques of the body.'

Turner's Legacy in Theater

The anthropological interest in ritual that Turner extended to theater resulted in an intellectual collaboration with Richard Schechner, a theater director and theorist of comparative performance who researches both practically and theoretically the common foundation of human creativity in ritual and theater. It is largely through him that Turner's last writings on performances as liminoid phenomena were published, promoting an interest in theatrical and other forms of contemporary performance in performance studies and cultural studies as well as anthropology. Indeed, in his preface to the book *Between Theatre and Anthropology*, Turner wrote that Schechner "might just be the catalyst anthropologists need to get them thinking about what Dilthey [the German hermeneutic philosopher] called 'lived-through experience'" (Turner in Schechner, 1985: p. xii; see also Turner, 1987: pp. 84–90).

The continuum between performance with ritual efficacy and theatrical performance rests on the theory that performance is 'restored behavior,' strips of behavior with 'a life of their own,' which can be rearranged or reconstructed independently of the various causal systems that brought them into existence (Schechner, 1985: p. 35). The 'machine' for the restoration of behavior is the liminal "workshop-rehearsal process … [which] is the same in theater as in ritual," as both function to restore behavior (Schechner, 1985: p. 113). Schechner has been a dominant theorizer of performance drawing on anthropological interests, but other theater director-researchers, from James Brandon to John Emigh and Phillip Zarrilli, also produce work informed by an anthropological sensitivity to fundamental cultural differences that provides inspiration for anthropologists who do research on performance in different societies across the world.

The encounter between theater and anthropology became a well-established interdisciplinary field in the United States during the 1980s and in the United Kingdom and Europe a decade later. In 1980, Eugenio Barba founded the International School of Theatre Anthropology; "Turner did for anthropology what Barba is doing for theater," declared Schechner (1985: p. 30; see also p. 257). Theater anthropology is not the study of performative phenomena by anthropologists, but of pre-expressive behavior by theater professionals in organized performance situations. It operates collaboratively with performers from societies with long-established theatrical traditions, including Japan, India, Indonesia, and Brazil, and in some cases with anthropologists. In 1990, Kirsten Hastrup's life story was 'restored' (to use Schechner's term) by Barba's Odin Teatret as a drama, *Talabot* (Hastrup, 2004: pp. 35–36). In the past decade, Performance Studies has become stronger with funding since the recognition of practice-based research for funding. It is also the case that performance practitioners who once might have become anthropologists are more likely to turn to performance studies.

Aesthetics and Creativity

Performer-anthropologists are often inspired by Turner's approach, including the former ballet dancer Anya Peterson Royce. She analyzes performance as 'the performing arts' cross-culturally to develop a general explanatory model of virtuosity and artistry based on "aesthetic principles that are common across genres, space and time" (Royce, 2004: p. 237). This approach brings together various dance forms (including Tewa ritual dances), mime, and cello music, and challenges the more common anthropological caution about aesthetics as a cross-cultural category. This study is among a number that are extending the range of embodied practices analyzed as performance by anthropologists, including the martial arts, performance art (see multimedia annex), literary writers, and others in interdisciplinary research in sports science.

More generally, anthropologists analyze creativity as a social phenomenon that is subject to structural constraints, both inside and outside the domain of the performing arts, and apply their theory to cross-cultural creative performance collaborations. In the twenty-first century, cross-cultural performances have increasingly been analyzed with reference to hybridity, most recently in dance studies in Asia. Anthropologists have critiqued hybridity because it reintroduces cultural reification by implying an authenticity or cultural purity in the elements that are combined and conceals substantial power differentials, including abuses of intellectual property. This critical awareness of hybridity may impact on performer creativity and feed back into anthropological analyses of performance. Anthropological approaches can also examine conflicts about ownership that are more frequent in transcultural theatrical productions.

Theater and Social Action

Anthropological analyses and uses of performance are not limited to the technicalities or aesthetics in theater and acting. Anthropologists also explore theoretical questions about social action through the analysis of stage acting in theatrical contexts. Johannes Fabian's pioneering ethnography with the theatrical Troupe Mufwankolo in Shaba, Zaire, marked a shift from "informative to performative ethnography" and a more "dialogical" form of anthropological practice (1990: pp. 5, 19). Performance is approached like an ethnography of speaking, not as the enactment of a script but by "giving form to … making, fashioning, creating" (1990: pp. 11–13). Fabian explored local concepts of power through drama and television, and demonstrated that performance must not focus on the aesthetic at the expense of power (1990: p. 17). Fabian is critical of the application of performance in a theatrical sense to societies with a different understanding of theater (1990: p. 14), and the conflation of performance and a general model of action, but he recognizes the real, not ritualized, redressive power of performance as social praxis (Fabian, 1990: p. 286).

Fabian has inspired a study of theatrical actors by Kirsten Hastrup, who explores the implications of action to understand human agency in general:

> Theatre is no metaphor for social life in the present context it *is* social life … To study theatre, therefore, is to study human actions in the real world as experienced by agents who are their own centers of attention to the world. (Hastrup, 2004: p. 19, emphasis in the original)

Building on longstanding interests in theater and philosophies of human action and knowledge as a reflexive situated practice validated by experience, Hastrup interviews Shakespearean actors about their experiences of performing in productions on stage and screen. Her account of the plays and their performance reveal an unfolding and mysterious world shaped by the power of language: "As thoughts in action words do change the world; nothing is ever quite the same after the words have been spoken" (2004: p. 319). Separating agency from intentionality, and noting that our actions are followed by a long coming to terms of how they bear on actions that affect others, Hastrup plots a schema for theater to explain the sociality of individuals (2004: pp. 322–331). This approach to individual human experience and actions through theater is very different from the rational calculation of Goffman's social actors through dramaturgical metaphors, and raises different implications for thinking about individual human choice and agency.

In considering anthropological approaches to the relationship between theater and action, a brief comment about gender identity and staged performances is necessary. Since James Peacock's (1968) early analysis of male-to-female cross-dressing in popular Javanese *ludruk* theater, accounts of theatrical cross-dressing have been complicated by Butler's theoretical analysis of gender-as-performance, where performance refers to action and identity. When theatrical gendering is discussed, for instance in relation to cross-dressing, there is a risk of performative skill being conflated with an expression of sexual preference. Anthropological approaches to cross-gendered stage performances are now engaging more rigorously with cultural and historical factors in cross-cultural analyses that recognize specific approaches to skillful theatrical gender impersonations by women as well as men (see multimedia annexes).

Ritual, Performance, and Media

Anthropological analyses of performance in ritual contexts continue to produce conceptually innovative research. For example, studies of performance within the social processes of healing have combined transactional and conflict models with analyses of the aesthetic and expressive dimensions of performance in the face of political repression or disenfranchisement. Michael Taussig (1993) invokes Marx and the Frankfurt School to explore mimetic behavior that lies at the heart of shamanic healing and social imagery, simultaneously representing and falsifying the world, and demonstrates how irrationality is being performed at the heart of the everyday. 'Mimetic excess' creates 'reflexive awareness,' which makes connections between history, make-believe, and otherness. More radically, anthropologists have applied healing processes from other cultures to help the sick at heart in their own societies by means of "a theatre model of dramatherapy" (Jennings, 1995: p. xxvii). While Aristotle's concept of catharsis may seem universal, different theaters produce forms of what Sue Jennings (1995: p. 185) calls 'abandonment,' and these vary greatly in their effects and forms of control. Ultimately it is this variation that must concern the anthropologist, not a universal explanation that is generalized out of one particular cultural perspective. Such explorations are a reminder of how the study of ritual performance promotes an understanding of how humans express their possession and dispossession, and of the permeability of the boundary between ritual performance and artistic performance.

Anthropological approaches to performance are shaped by whether the methodological approach is one that privileges analytical intersections or boundaries. As human interaction intensifies across time and space, anthropological approaches to performance have addressed the limits of dichotomizing 'everyday' vs 'ritual' realities in the complexity of contemporary life. Attention to media (television, film, and video) has opened up wider contexts for debates about performance being a specialized form of behavior or part of emergent everyday lives. The concept of performativity has been combined with Bell's definition of ritualization ("a way of acting that is designed to distinguish and privilege what is being done in comparison to other usually more quotidian, activities"), which could equally apply to performance, to question the division between ritual and everyday life, and to propose that ritual is not a genre of performance but that "performance in part constitutes ritual and ritualization" (Hughes-Freeland, 1998: pp. 3–6). The three domains were all deemed "in some sense virtual realities, mediated by technologies … in contradistinction to sensate everyday reality," although Goffman's penetration of everyday life by ritual undermines any fixed formulation (Hughes-Freeland, 1998: pp. 10–11). A consideration of these intersections affirmed that anthropologists achieve an understanding of their own categories through their encounter with different models, and that any convergences that arise will help to move forward debates about "the relationship of various expressions of imagination to other forms of life" (Hughes-Freeland, 1998: p. 15).

In the past decade, the analytical relationship between performance and ritual is being disentangled. This is due to a number of reasons. The concept of secular ritual made possible an exploration of intersections, but the term *ritual* is now being employed metaphorically across so many different domains that it risks losing analytical salience. For example, liminality is routinely applied to ceremonial events covered extensively by the print and broadcast media. In some approaches, technological mediation replaces ritual with spectacle, while in others, television broadcasts extend the ritual into the secular domain. The question to ask now is whether there is a limit to which the ritual process can be stretched across time and space (Hughes-Freeland, 2006: pp. 612–614). Just as ritual events may need to be distinguished from media ones, the relationship of performance to ritual, and the relationship of both to action, is also being reevaluated as a central issue (see following).

Performance and Cultural Politics

Some of the previously discussed approaches have included the political dimension of performance, but there is also evidence of Abner Cohen's somewhat neglected approach (outside the domain of African studies) to political symbolism and ethnicity in studies of performance. This is illustrated in two ethnographic studies about Southeast Asian societies.

The first study explains how social and cultural values are embodied and enacted in everyday social interactions in Javanese court dance movement and aesthetics, and how these become transformed into official Indonesia culture (Hughes-Freeland, 2008). In this analysis, performance is caught up in the contradictions between the two dimensions of social interaction: the unintentional and moral nature of cultural action, and the amoral self-interest of political action (Cohen, 1993: p. 8). Classicized court dance is a form of social production that plays a symbolic and an instrumental role in the construction and continuation of local and national communities (2008: pp. 11–13). Cultural politics include court dance in the domain of official national culture but exclude other performances practices with ritual and healing powers. For example, *tayuban* dancing occurs at annual community events in certain highland villages in Java due to the support of the local leader. As part of the occasion, people bring their children and animals (represented by their tethering ropes) to be cured by the face powder of women singer-dancers that carry the power of the village spirit (Figure 1). A more

Figure 1 Female dancer in Java protects a baby with face powder (as in first edition).

political approach to performance is evident in other research into the relationship of embodiment as sensorial experience and citizenship.

Cohen's influence is more implicit in a second analysis of the dynamics of ethnicity and elitism in cultural politics in Chinese street opera in multicultural Singapore. Lee's (2009) cultural analysis begins with changes in the social organization and troupe membership following the first migrations from China in the 1840s. By the 1970s, Singaporean cultural politics was privileging amateur troupes that perform a few times a year, mostly in indoor theaters (2009: pp. 75–95). Meanwhile professional troupes, despite their decline after theater closures in the 1960s, perform almost daily at temple festivals (2009: pp. 117–137). These two forms of Chinese opera have distinct audiences and value systems that differ from Chinese opera performances for tourist audiences at Clarke Quay. These provide a Geertzian "metacommentary upon the role of cultural tradition and processes of modernization in Singapore … tourism and culture are constitutive of each other" (2009: p. 137). This ethnomusicological research is also informed by business studies, providing further evidence for the potential for interdisciplinary forms of contemporary social analysis of performance and cultural politics. The recognition of internal differentiation and heterogeneity of socio-aesthetic and social values within these anthropological approaches to performance run parallel to those of the French philosopher Jacques Rancière, who has lately become fashionable, but differ from him because his theories about aesthetics and power derive entirely from Western art and epistemology. Both these studies also address the way in which governments and performers respond to local and international tourism, another field for new analytical perspective on performance.

Place, Tourism, and Heritage

For the past 30 years, the analysis of performance in pilgrimage has created a link between ritual and tourism, but in the past decade, the 'topological' or 'topographic' turn in anthropology has produced new approaches to performance in relation to place and movement across space.

In anthropological accounts of tourism, performance applies to performances for tourists, and to tourism itself as a distinct sphere of action. Earlier accounts of tourism that privileged the gaze and spectatorship have been challenged by concepts from performance theory, and specific tourist behaviors are being analyzed as performances. This performative approach gives greater weight to participation as "a dynamic sense of embodied and performed, as well as visualized and textualized, engagement with places and tourist activities," and go "beyond the visual to performance" (Coleman and Crang, 2002: pp. 1, 11).

Approaches to mobility simultaneously generate a revived interest in place and in site specificity. Anthropologists also collaborate with archaeologists and art historians to research public interactions with heritage sites, as in Blain and Wallis's sacred sites project (see multimedia annex). Another recent framework for the practical exploration of performance in space, theater archaeology, developed in Wales by Pearson and Shanks, could potentially produce interesting new collaborative approaches to place and performance.

The interest in place and mobility is linked to the enormous field of heritage studies. Here, performance is discussed in relation to its classification by the United Nations Educational, Scientific, and Cultural Organization (UNESCO) as intangible heritage. Critiques of the criteria for classifying cultural heritage have resulted in heritage researchers moving away from performance as event to performance as action. Smith and Waterton (2009: p. 291) note a common ambivalence about intangible heritage in "authorised heritage discourse," and provocatively suggest that all heritage is intangible. Heritage "is the performance and negotiation of identity, values, and a sense of place …. One of the key consequences of the heritage performance is the way in which it is used to legitimise and authorise identity" (Smith and Waterton, 2009: pp. 292–293). As performance itself involves these

processes of negotiation and legitimization, there is a theoretical complication here that needs to be untangled in the future. This is also an area in which anthropologists are researching conflicts about cultural ownership and the control of representation in performance.

Central Issues

Performance is productive both of individuals, with gendered, ethnic, or other specified markers of social identity, and of social formations of different scales. There has been a tendency to explore performance as ritual, or in theater, or in art, or to frame certain specialist behaviors, such as tourism, as performance. Exceptions are the concerns with the relationship between performance, action, and ritual. Performance is also integrated into wider ethnographic analysis, but in the past decade it has been subsumed by a range of theories and methods and has had a lower profile as a key anthropological concept, even as anthropological approaches to performance are being highlighted in sociology and 'performance geography.'

An important debate about performance is the extent to which human action is scripted or emergent. Is performance best understood as mimesis and imitation, or as *poesis*, creative, emergent, and transformative? In the 1980s, performance was addressed by linguistic anthropologists to counter textual approaches to oral traditions. This generated a reaction emphasizing performance in anthropology, and debates became polarized as performance models vs textual models, with a link to process models vs structural ones. One party stressed the determinacy of form; the other gave primacy to emergence, praxis, and experience. By the start of the twenty-first century this insight was recognized belatedly in the humanities (for instance, in Beyond Text, a research program funded by the British Arts and Humanities Research Council).

However, in anthropology there had already been doubts about the performative emphasis and the ensuing polarization between text and performance:

> performance theory provided wonderful conceptual tools for thinking about emergence …. [but] It offered inadequate resources for understanding how the fluid is consolidated, and why stunningly creative oral performers so often claim that their texts have never changed by so much as a syllable. (Barber, 2007: p. 29)

The necessary complement to performativity is entextualization, a concept developed by anthropologists working on performance in the 1990s (Kuipers, Urban, and Silverstein), because it allows an oral performance to be detached from the time and place of its enactment. It also explains how orally transmitted forms, as well as written ones, endure or 'stick' (Barber, 2007: pp. 28–31). This is one example of a reintegration of the relationship between text and performance, which are no longer necessarily mutually exclusive or oppositional, though there are variations in the analysis of script and enactment and how performances generate textual meanings (see multimedia annex).

This debate situates performance centrally within current polemics about anthropological knowledge and whether it should represent itself as operating within a system of models and rules or performatively and dialectically. Anthropologists still approach performance as being risky and interactive, rather than determined and predictable, using active and participatory models of spectatorship, not passive ones, on a continuum of general varieties of human interaction.

Another central concern is the cross-cultural application of performance, given its fictive theoretical foundations. Schieffelin (1998) has warned against exporting Western concepts to analyze other societies. Goffman's performance model draws on Western notions of theatrical performance in which illusion is central: theater is about make-believe, not real life. However, for Kaluli audiences at séance healings and the Gisalo ritual in Papua, New Guinea, the effects of performance are real, not illusory; audiences respond with real, not conventional, emotion. The anthropology of performance should move away from the representation of reality to the social construction of reality: "the concepts 'performance'/'performativity' have the potential to open a considerable domain for anthropological exploration of the way that cultures actively construct their realities" (Schieffelin, 1998: pp. 199–200). This resonates with approaches by Fabian and Hastrup that use different epistemological frameworks. Performativity in Goffman's account necessitates questioning what, if ever, is authentic action or representation, although a focus on experience and embodiment has moved analysis away from such questions – but they refuse to go away. Interestingly, it has recently been argued that Schechner's distinction between ritual and performance as a distinction between efficacy and instrumentality in contrast to entertainment suggests that, despite his association with Turner and his ritual theory, he is closer to Goffman (the ritual skeptic) than to Turner (the ritual enthusiast) because Schechner's emphasis is on the "fictive, contrived nature" of liminoid restored behavior (Grimes, 2012: pp. 32–34).

There have also been some challenges to the intersection between performance and ritual. Bruce Kapferer, formerly a proponent of performance in an analysis of healing and personhood inspired by G.H. Mead, has revised his performative approach to ritual, arguing that ritual is constituted by virtuality (from Langer to Deleuze and Guattari): "Insofar as ritual can be conceived as a technological virtualizing practice, a virtual reality machine, it is a device for entering into human actualities and is an opening and slicing into actuality" (Kapferer, 2006: p. 675). Ritual accommodates performance and mimesis, but should not be defined by them. This argument here is based on the premise that ritual performance is representational, but as most of the approaches to performance discussed here are nonrepresentational, this criticism may not result in ritual being uncoupled from performance. Kapferer's technical idiom is intriguing, but even more significantly it is the reminder that different orders of reality are socially constructed; there are a multiplicity of forms of experience that provide scope for further analytical differentiation.

A second challenge to the relationship between ritual and performance occurs in a recent statement by a pioneer of performance theory in ritual analysis. In a critique of the trajectory of 'cultural performances' within anthropology, including Turner's own contribution, Ronald L. Grimes stresses

that "*ritual* and *performance* are different only two children of the parent term, *action*. They have other siblings, such as *play*, *sport*, *dance* and *music* …. When speaking technically, I prefer to say that ritualists 'enact' rituals, whereas actors 'perform' plays" (Grimes, 2012: pp. 34–35, emphasis in the original). The differences between performance and ritual are not essential, but can be worked out using family resemblance theory (polythetic classifications), and with due recognition to "*degrees* of ritualization" (Grimes, 2012: p. 39). This qualification allows for a less categorical opposition between ritual and performance, but it is a warning against routinizing analytical hybrids such as performance-as-action or ritual-as-performance, and to be alert to dramaturgical metaphors coming into explanatory frameworks through the back door, and a tendency to use 'performing' as a synonym for a phrase involving certain kinds of action in contexts such as gender and tourism, for example. Grimes's challenge would also certainly require a revisiting of performativity, and an answer to a question that the late Raymond Firth once asked (at the ASA conference in Swansea in 1996): "Do we need performativity?"

More fundamentally 'ritual,' like 'theater' and 'performance,' should not be assumed to be culturally ubiquitous. As a path-breaking critic of the essentialization of dance as a category has recently observed, she researches 'structured movement' within ritual, but the viability of ritual itself as an analytical category is still being assessed through her ongoing comparative research (Kaeppler, 2010: p. 271). Once again, the anthropological impulse is to test, rather than to theorize a priori.

This heuristic attitude also applies to performance. It is important to specify whether performance is treated as universalizing meta-theory, a metaphor, event/action context, or a locally grounded concept. In anthropology, there are different strategies toward performance and its assessment, and these vary according to attitudes to concept formation and theoretical preferences. Some of the innovations noted previously appear to reassert boundaries and situate performance in empirical theatrical contexts: Zaire theater troupes, Shakespearean plays, Chinese street opera, or Javanese court dance. But they do this to explore broader social issues. By critically questioning the presuppositions supporting categorical oppositions implied in performance, such as pretense vs authenticity, and by going against the grain of their own cultural presuppositions, anthropologists are able to construct descriptions of alternative ways of seeing and being in the world, and to celebrate the diversity of human cultural experience despite its shared biophysical nature. Anthropologists in principle thus differ in their treatment of performance from those working in other disciplines, insofar as they recognize regional frames of reference and produce detailed ethnographic research. This may refer to what performers do and how they explain their work in contexts that nonanthropologists may identify as theatrical or staged; or it may refer to other contexts and events, or to the processes of everyday interactions.

Future Directions

Anthropologists have used performance as a means to understand social life, social interaction, and the social contexts for the formation and expression of human creativity. By avoiding definitions and using the performance concept in a heuristic manner, they are able to develop insights into the acquisition of competence, as general socialization or as training to acquire particular skills, and to further their understanding of social interactions and formations. In the future, it is to be hoped that they will be able to continue to do so.

By the end of the twentieth century, questions about function and meaning that had framed the anthropological project were being superseded by concerns with identity and experience.

Media and information and communication technologies facilitate virtual realities, and complicate the different ways in which performance may be transmitted and received, beyond the defined space–time continua of theater, ritual arena, or event. Anthropologists research all performance arenas, including cinema, television, and the Internet, and now Web 2.0, social media to the latest mobile mass communications, offer further permutations of sociality that would have been unimaginable 20 years ago. The proliferation of performance and ritual metaphors in the analysis of media events, for instance, has raised questions about analytical intersections, and some have asked for a rethinking of conceptual boundaries between action and performance as one of its other domains.

Visual technologies give access to all kinds of performance, rather than simply verbal accounts. Although these moving images are not pure imitations of the performances as they occur, it could be argued that the researcher is freer to respond directly to the performance than when reading about it. For example, Maya Deren's film of *vodun* possession in Haiti, *Divine Horsemen* (1947–51), and Jean Rouch's film of the *hauka* possession cult in Ghana in 1953, *Les Maîtres Fous* (*The Mad Masters*), presents images that language might fail to represent as convincingly real (see multimedia annex). Experimental ethnographic explorations of the relationship between embodied performance and text suggest innovative and exciting approaches to issues of analysis and representation in anthropology.

In the past decade, there has been a radical change in access to images of performance on YouTube and Vimeo, but the ethnographic context is not always clear, and such images can produce fragmentary impressions and endorse prejudice, rather than anthropologically founded understanding. The anthropological analysis of performance in cyberspace and the knowledge it produces is potentially valuable within the much wider field of transmission, learning, and social continuity. Performativity meanwhile continues to have relevance in the analysis of social media, particularly in relation to processes of self-invention and the creation of false identities. Important questions for anthropologists to examine concern the kinds of reality constructed in between public enactments, on the walls and pages of social media and in virtual worlds in cyberspace, and the relationships of these spheres of activity to off-line behavior.

The potential fields for the development of anthropological approaches to performance are numerous. The study of linguistic performance has already been employed in communication analyses in the field of artificial intelligence,

a development that extends to understandings of embodied performance. Some of the innovations discussed before will continue to develop. Performance will be part of anthropological engagements with public culture and issues of social relevance, such as migrant experiences, claims to cultural ownership, and the role of social media. The Occupy Movement has brought back performance to the streets as an expression of political dissent. This will add to the use of performance action in participatory methods and to applied theater in development and community projects. The anthropology of performance is among the most ostensibly humanist of the more recent subareas of anthropology, and its qualitative methodologies will be of value in balancing quantitative approaches in analyzing any form of professional practice, including medicine, law, finance, education, and the creative industries, in order to understand the processes that have come to give it the meaning of 'meeting targets' in neoliberal discourses.

Despite living in a globalized world, the requirement for national relevance in research is, to a certain extent, impeding attention to cultural diversity in performance, and less international research is being done away from the anthropologist's country of employment. An unintended consequence of this trend is for a new 'armchair anthropology' to be developing, in front of computer screens. To forestall the very real possibility of a return to a new ethnocentrism as a result of this lack of cross-cultural experience in research, it is crucial that in the immediate future, anthropologists continue to work against the grain of the new parochialism lurking behind the seeming borderlessness of information and communications technology (ICT), and to sustain a less culturally entrenched view of human performance activities. To achieve this, anthropologists need to engage in intracultural or cross-cultural performances to promote wider understanding about the different ways of acting human within the constraints of the group at whatever level and the challenges performance presents to those constraints which allow the expression of human nature in its different forms.

See also: Body: Anthropological Aspects; Cultural Heritage; Dance, Anthropology of; Discourse, Anthropology of; Hybridity; Music: Anthropological Aspects; Personhood, Anthropology of; Play, Anthropology of; Ritual; Shamanism; Spirit Possession, Anthropology of; Tourism.

Appendix A Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/B978-0-08-097086-8.12127-0.

Bibliography

Barber, K., 2007. Improvisation and the art of making things stick. In: Hallam, E., Ingold, T. (Eds.), Creativity and Cultural Improvisation. Oxford, Berg, pp. 25–41 (ASA monographs 44).

Bauman, R., 1986. Story, Performance, and Event. Cambridge University Press, Cambridge, UK.

Cohen, A., 1993. Masquerade Politics. Berg, Oxford, UK.

Coleman, S., Crang, M., 2002. Grounded tourists, travelling theory. In: Coleman, S., Crang, M. (Eds.), Tourism: between Place and Performance. Berghahn, Oxford, UK, pp. 1–17.

Fabian, J., 1990. Power and Performance. University of Wisconsin Press, Madison, WI.

Goffman, E., 1956. The Presentation of Self in Everyday Life. University of Edinburgh Social Sciences Research Centre, Edinburgh, UK (Monograph No. 2).

Grimes, R.L., 2012. Religion, ritual and performance. In: Gharavi (Ed.), Religion, Theatre and Performance: Acts of Faith. Routledge, New York and Abingdon, pp. 27–41.

Hastrup, K., 2004. Action. Anthropology in the Company of Shakespeare. Museum Tusculanum Press, University of Copenhagen, Copenhagen.

Hughes-Freeland, F., 1998. Introduction. In: Hughes-Freeland, F. (Ed.), Ritual, Performance, Media. Routledge, London and New York, pp. 1–28 (ASA monographs 35).

Hughes-Freeland, F., 2006. Media. In: Kreinath, J., Snoek, J., Stausberg, M. (Eds.), Theorizing Rituals: Issues, Topics, Approaches, Concepts. Brill, Leiden and Boston, MA, pp. 595–614.

Hughes-Freeland, F., 2008. Embodied Communities: Dance Traditions and Change in Java. Berghahn Books, New York and Oxford, UK.

Jennings, S., 1995. Theatre, Ritual and Transformation: The Senoi Temiars. Routledge, London.

Kaeppler, A.L., 2010. Interpreting ritual as performance and theory Association for Social Anthropology in Oceania 2010 Distinguished Lecture. Oceania 80 (3), 263–271.

Kapferer, B., 2006. Virtuality. In: Kreinath, J., Snoek, J., Stausberg, M. (Eds.), Theorizing Rituals: Issues, Topics, Approaches, Concepts. Brill, Leiden and Boston, MA, pp. 671–684.

Lee, Tong Soon, 2009. Chinese Street Opera in Singapore. University of Illinois Press, Urbana and Chicago, IL.

Peacock, J., 1968. Rites of Modernization. Chicago University Press, Chicago, IL.

Royce, A.P., 2004. Anthropology of the Performing Arts. AltaMira Press, Lanham, MD.

Schechner, R., 1985. Between Theater and Anthropology. University of Pennsylvania Press, Philadelphia, PA.

Schieffelin, E.L., 1998. Problematizing performance. In: Hughes-Freeland, F. (Ed.), Ritual, Performance, Media. Routledge, London and New York, pp. 194–207.

Smith, L., Waterton, E., 2009. The envy of the world? Intangible heritage in England. In: Smith, L., Akagawa, N. (Eds.), Intangible Heritage. Routledge, London and New York, pp. 289–302.

Taussig, M., 1993. Mimesis and Alterity. Routledge, New York.

Turner, V.W., 1987. The Anthropology of Performance. Preface by R. Schechner. Performing Arts Journal, New York.

Performance Appraisal and Evaluation

Dina Van Dijk and Michal M Schodl, The Guilford Glazer Faculty of Management, Ben-Gurion University of the Negev, Israel; and Faculty of Health Sciences, Ben-Gurion University of the Negev, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Performance appraisal (PA) refers to the methods and processes used by organizations to assess the level of performance of their employees and to provide them with a feedback. This process can be used for both developmental and administrative purposes. The research on PA includes examination of the psychometric aspects of the appraisal tools, the cognitive process and the biases involved, and the social context. Although PA is an important tool for managing employees, managers, and workers are rarely satisfied with it. Therefore, clearly defining the purposes of the PA and addressing the employees' reactions are essential to the success of the PA.

Definition of Performance Appraisal

Performance appraisal (PA) plays a central role in managing human resources in organizations (e.g., Boswell and Boudreau, 2002; Cardy and Dobbins, 1994; Judge and Ferris, 1993). The term performance appraisal (or performance evaluation) refers to the methods and processes used by organizations to assess the level of performance of their employees. This process usually includes measuring employees' performance and providing them with feedback regarding the level and quality of their performance (DeNisi and Pritchard, 2006). The main goal of the PA in organizations is to improve employee performance (DeNisi and Pritchard, 2006; Murphy and Cleveland, 1991). This goal could be achieved through three possible mechanisms: (1) the information provided by the PA can be used for administrative decisions linking the evaluated performance to organizational rewards or punishments such as a pay raise, promotion, or discharge (Cleveland et al., 1989; Landy and Farr, 1980; Raynes, Gerhart, & Parks, 2005); (2) the PA process involves providing performance feedback (i.e., information regarding the level of performance) to the employees who were evaluated, allowing them to adjust their performance strategies to match the desired performance (e.g., Erez, 1977; Kluger and DeNisi, 1996; Locke and Latham, 2002); and (3) the PA is a process that raises employee awareness to the fact that they are being measured. As has been shown since the Hawthtorne studies (Roethlisberger and Dickson, 1939) and is expressed in the aphorism "what gets measured gets done," the mere fact of knowing that one is being observed or measured increases performance and fosters cooperative behavior (Bateson et al., 2006; Haley and Fessler, 2005; Keller and Pfattheicher, 2011). While the first and second mechanisms of PA have been widely explored, the third has received little attention in the PA literature.

Criteria for PA Effectiveness

Assuming that the main purpose of the PA process is to increase performance, an effective PA would be one that achieves this purpose. However, it is not as obvious as it seems to establish the criteria for effective PA that lead to performance improvement: Does effective PA mean that the scales in use are accurate and free of rater biases? Or does it mean that the workers accept the PA results and are willing to change their manners accordingly? While early research focused primarily on the accuracy of ratings as the only criterion for PA effectiveness, recent studies have suggested focusing on motivational aspects such as ratees' reactions to the PA (e.g., satisfaction, feedback acceptance). In order for workers to improve their performance following an appraisal, they must accept the appraisal rating and be willing to change their performance accordingly. Ensuring workers' reactions such as satisfaction, commitment, acceptance of the appraisal, and trust in management could help organizations achieve the primary PA purpose of improving performance (Keeping and Levy, 2000; DeNisi and Pritchard, 2006; Mayer and Davis, 1999). In order to yield positive reactions among ratees, the PA process should be perceived as reliable, accurate, and free of political interests on the one hand, and allow the ratee to participate in the process and express his or her voice on the other. The shift from measuring the accuracy of the appraisals to measuring the ratees' reactions and motivations reflects a significant change in the research of PA (Levy and Williams, 2004). This shift will be discussed in the following section.

Changes in the Focus of Performance Appraisal Literature

Until the 1980s, the focus in the PA literature was on measurement development and the psychometric characteristics of the different tools used to evaluate performance (Arvey and Murfey, 1998; Landy and Farr, 1980). Specifically, most research was concerned with improving these tools, constructing rating scales, and examining the advantages and disadvantages of different types of ratings. This psychometric tradition changed dramatically in the early 1980s when the focus of research shifted to the cognitive characteristics of the raters (e.g., Feldman, 1981; Landy and Farr, 1980). Drawing on social cognition and cognitive psychology, researchers began to explore how raters create impressions and judgments of their subordinates (Arvey and Murfey, 1998). The PA literature has become more concerned with attribution errors, categorizations, stereotyping, and other biases involved in information processing (e.g., DeNisi et al., 1984; Dobbins et al., 1988; Feldman, 1981).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.22034-5

Another historical change in the PA literature has occurred since the 1990s. At the beginning of the decade, several keynote reviews (e.g., Bretz et al., 1992; Ilgen et al., 1993; Murphy and Cleveland, 1991) argued that cognitive process models had failed to narrow the gap between research and practice, and since PA is an applied subject, there should be increased effort to make the research more applicative. The cognitive approach assumed that the cognitive process is universal and, therefore, the different content domains in which PAs practically occur (Ilgen et al., 1993) were overlooked. Researchers at this time suggested that in order to contribute to the practice of PA in organizations, it is important to understand the specific context in which an appraisal takes place, especially the social context (Levy and Williams, 2004). As the appraisal process takes place in a social context, it is, therefore, crucial to understand this context for appraisal success. The PA process is doomed to fail if the ratees perceive it as unfair or manipulative, even when the psychometric aspects of the process are valid (Cawley et al., 1998; Keeping and Levy, 2000; Levy and Williams, 2004). Some prominent components of the social context that influence the PA process will be reviewed further in this article, but first we introduce a key element in every PA system: feedback.

Performance Feedback

Providing employees with feedback regarding their performance is a common reality in most organizations. This type of feedback is defined as "actions taken by (an) external agent (s) to provide information regarding some aspect(s) of one's task performance" (Kluger and DeNisi, 1996: p. 255). This definition reflects the idea that feedback is 'information' or 'knowledge' delivered to an employee in order to make him or her aware of the level of their work outcomes. Performance feedback could serve as an important motivational source for workers if it is conducted effectively (Peiperl, 2001). Also, workers are interested in performance feedback in order to know how close they are to fulfilling their job assignments. However, giving (and receiving) feedback seems to be a challenging task. In the mid-1990s, Kluger and DeNisi (1996) published an extensive review and meta-analysis on the impact of feedback interventions. Their meta-analysis of feedback literature includes 131 articles and 607 feedback effects published between 1905 and 1996. Kluger and DeNisi found that in contrast to the common view that governed the feedback literature for nearly a hundred years, feedback does not always achieve its desired effect. Specifically, in more than a third (38%) of the studies reviewed, feedback was actually followed by a decrease in performance. Moreover, they found that the level of positivity or negativity of the feedback (i.e., feedback sign) does not explain why some feedback interventions were effective, while others were not. Lastly, the authors suggested and demonstrated that the level of the feedback (i.e., task details, task performance, and meta-level – the self) moderates the feedback–performance relationship. Specifically, effective feedback should help the receiver concentrate on the level of task performance, namely, the specific ways to improve his or her performance. Feedback often shifts the attention of the receiver from this level to a higher or a lower level; for example, the feedback could shift receiver attention to a higher level of the self (e.g., what this feedback says about me), or to a lower level of the task details (e.g., how I am holding my hands while teaching). The authors argued that feedback that either shifts attention to too high or too low a level, may be ineffective. This is mainly because people's cognitive resources are limited and if too much attention is invested on specific details or even worse, in one's self, there may not be enough resources available for performance improvement.

Recently, it was suggested (Van Dijk and Kluger, 2004, 2011) that the distinction between promotion and prevention goals or tasks (Higgins, 1997, 1998) could help defining the situations under which feedback will be effective. Specifically, tasks that are performed to satisfy one's wishes and aspirations (promotion tasks) benefit from positive feedback, while tasks performed to meet duties and obligations benefit from negative feedback. Practically, it suggests that one size does not fit all; thus, managers should tailor the feedback according to worker and task characteristics.

In light of the complexity of executing effective feedback processes, some researchers have suggested alternative interventions to feedback that focus on identifying conditions for success (e.g., Bouskila-Yam and Kluger, 2011; Kluger and Van Dijk, 2010). However, research on alternative interventions to feedback is only in its beginning stages, and the findings are not yet conclusive. Finding effective interventions to improve the feedback process remains one of the challenges for development in this field.

Multisource (360-Degree) Appraisal Technique

Multisource appraisal is one of the most widely used PA techniques in today's organizations (Antonioni, 1996). With this technique, the appraisal of an employee does not rely solely on one source (e.g., manager) but on multiple sources such as managers, peers, subordinates, and clients. Using the multisource technique helps evaluators collect a range of views regarding an employee's performance, with each source adding a unique perspective. For example, a manager can evaluate a nurse on professionalism and compliance to rules, a peer could evaluate how well he or she gets along with colleagues, and a patient could add information on how this nurse treats patients. With multisource feedback, the ratee usually receives the results along with normative data and self-ratings, providing the employee with comparative information (London and Smither, 1995). The purpose of the multisource appraisal is developmental (Brett and Atwater, 2001; DeNisi and Kluger, 2000), namely, to develop and nurture employees and to help them achieve their goals and aspirations, which eventually leads to a promotion focus. As such, use of multisource appraisals assumes to improve employee performance.

Some studies have reported improvements in overall performance following multisource appraisals or upward feedback (e.g., Atwater et al., 1995; Reilly et al., 1996). However, other studies (e.g., Brett and Atwater, 2001; Smither et al., 2005) questioned the effectiveness of this type of appraisal. Specifically, in a meta-analysis, Smither and colleagues (2005) found overall multisource feedback had a small effect on performance improvement. Furthermore, in a longitudinal study on

multisource feedback, Brett and Atwater (2001) found that when ratings were low, employees reacted negatively, expressed anger, and rejected the feedback results. Even when ratings were high, employees did not show positive reactions, as would be expected, but rather, an absence of negative reactions. These findings are especially disappointing because the specific 360-degree feedback that was used in this study was developmental – it was designed to give useful information to the employees without any intention (or capacity) to use the data for other administrative purposes. Thus, despite some positive reported outcomes of the multisource appraisals, "the available data on the effectiveness of these programs is extremely limited, and the conclusions that we can draw from these studies are rather limited" (DeNisi and Kluger, 2000: p. 135).

The most difficult issue in PA is delivering negative feedback to an employee (Ilgen and Davis, 2000). Managers are not enthusiastic to give negative feedback, and employees find it very difficult to accept such feedback (especially if it contradicts their expectations). As a result, workers may become less cooperative and their work performance may suffer, preventing further work improvement and behavioral change. Reactions to negative feedback are influenced by individual differences in self-esteem, regulatory focus, goal orientation, and performance history (e.g., Brett and Atwater, 2001; Ilgen and Davis, 2000); therefore, these factors should be taken into account when negative feedback is given. In addition, the appraisal data should be used only for developmental purposes, and if other uses are considered, this fact should not be hidden from the employees because no PA system can work without mutual trust (DeNisi and Kluger, 2000; Mayer and Davis, 1999).

In view of the challenges and difficulties involved in the feedback process, it is clear that giving feedback demands a high degree of skill and knowledge. However, mastering interpersonal or communication skills is only one part in constituting an effective evaluation system. As mentioned earlier, the social context also is of great importance. This will be addressed next.

The Social Context of the PA Process

The social context in which the PA takes place is crucial to its success. The social context of the PA process consists of issues related to the rater, the ratee, and the relationship between them, as well as to global factors that indirectly affect the PA process such as the organizational culture, legal climate, human resource strategies, and organizational goals (Levy and Williams, 2004). In the following sections, we will review two of the social context factors that are essential to the success of the PA process: procedural justice and leadership.

Perceived Justice of Ratees

Perceived justice appears to be an essential mechanism through which appraisals affect employees' reactions (e.g., Erdogan, 2002; Greenberg, 1986); thus, this has received considerable attention in the PA literature. So how can appraisals be conducted in a fair manner? To answer this question, we need to examine two types of justice described in the literature of justice and fairness that are involved in the PA process: distributive justice and procedural justice. Distributive justice relates to the fairness of the appraisal relative to the exerted effort. Procedural justice relates to the fairness of the procedures that were used to determine the appraisal ratings (Greenberg, 1986). Levels of perceived procedural justice are positively related to important organizational outcomes such as organizational citizenship behavior (Ball et al., 1994; Moorman et al., 1998; Organ et al., 2006; Skarlicki and Latham, 1996, 1997; Whiting et al., 2008), trust in leadership (Barling and Philips, 1993; Folger and Konovsky, 1989; Whiting et al., 2008), organizational commitment, job satisfaction, and performance (Colquitt et al., 2001; Korsgaard et al., 1995).

Folger et al. (1992) have offered a comprehensive model that presents the key variables for designing a procedurally just PA system – the due process model. This model consists of three elements: adequate notice, fair hearing, and judgment based on evidence. Accordingly, the standards for evaluation should be evidence-based, employees should receive early notice about the evaluation standards, and consistent periodical feedback should be given to employees regarding their performance. In addition, employees should be given an opportunity to influence the process during evaluation meetings and to present their opinions (voice).

Several studies have found support for the due process model. Specifically, Taylor and colleagues (Taylor et al., 1998) examined employee–manager pairs that were randomly assigned to two types of evaluations: the customary existing appraisal process versus the due process appraisal. Employees involved in the due process appraisals displayed more positive reactions (e.g., perceived fairness, evaluation of the manager, intention to remain in the organization) than employees in the customary appraisals, even though the evaluations in the due process procedure were actually lower. Further support for the due process model has been found in other studies (e.g., Erdogan et al., 2001; Poon, 2004). Erdogan and colleagues found that characteristics of the due process appraisal (evidence-based criteria and fair hearing) were associated with perceived procedural justice. In addition, Poon (2004) found that when employees perceived the PA process as manipulative and skewed by the political interests of the raters (as opposed to the due process), they demonstrated less satisfaction and higher intention to quit their jobs.

Though the due process guidelines are clear, building a PA system accordingly is a significant challenge, as the performances of different workers are not always comparable and cannot be tested in a similar manner. Unlike student evaluations in which all students receive one standard test at the same time under the same conditions, managers attempting to evaluate their workers face a completely different situation. They do not always have the opportunity to observe all their workers for the same amount of time, and workers are not always performing comparable tasks in terms of difficulty and complexity. Moreover, with today's globalization, managers often supervise their workers from a distance without sufficient opportunity to observe their work, which makes this process even more challenging. Moreover, biases and impression management tactics seem to impact the ratings (e.g., Dulebohn and Ferris, 1999).

Therefore, research in recent years has striven to understand the factors affecting the perceived procedural fairness of PA systems (Erdogan, 2002; Erdogan et al., 2001; Folger et al., 1992) including factors concerning the rater's personality and perceptions (Heslin and VandeWalle, 2011; Mayer et al., 2007), which affect perceived procedural justice. Specifically, managers who exhibit high levels of conscientiousness and agreeableness, in combination with low levels of neuroticism (Mayer et al., 2007), and believe that workers are capable of change (Heslin and VandeWalle, 2011) are perceived as more procedurally just.

Leadership Effect on PA Effectiveness

Another contextual factor influencing the effectiveness of a PA is the manager's leadership behavior (Elicker et al., 2006; Gabris and Ihrke, 2001; Purcell and Hutchinson, 2007; Waldman et al., 1987). More specifically, the exchange relationship between the managers (who perform the evaluation) and their employees (those being evaluated) strongly affects the reactions of the employees to the PA process. The PA session is a 'contact point' (Holbrook, 2002) that is influenced by the existing exchange relationship (Elicker et al., 2006) and by leadership credibility (Gabris and Ihrke, 2001).

The literature on leader–member exchange (LMX) has contributed to our understanding of supervisor–subordinate relationships (Gerstner and Day, 1997; Graen et al., 2006) and their impact on various organizational outcomes (e.g., Fairhurst, 1993; Kacmar et al., 2003; Kark and Van Dijk, 2008). Specifically, the LMX model identifies the supervisor–subordinate relationship as a dyadic social exchange process that is unique to each supervisor–subordinate pair (Graen and Uhl-Bien, 1995). When LMX is high (in-group), the worker receives more attention and support from his or her leader, their relationship is close and warm, and they trust each other. As a result, the worker demonstrates a higher level of performance and has a positive attitude compared with a worker with low LMX (out-group). Given the initial warm and supportive relationship between managers and their in-group workers, it is obvious that during PA sessions, the workers with high LMX will be more confident in their ability to communicate with their managers and, therefore, will achieve more positive outcomes than the out-group workers.

Elicker and colleagues (Elicker et al., 2006) have found that the opportunity for an employee's voice and his or her perception of justice are important mechanisms for the effect of LMX on his or her reaction to a PA. Specifically, they found that the quality of the exchange relationship between leaders and workers affects the opportunity of the workers to voice opinions during a PA session, and as a result, increases the perception of justice, and this perception, in turn, influences workers' attitudes toward the PA process (i.e., motivation, satisfaction, perceived accuracy, and perceived utility). Besides LMX, other leadership behaviors are also found to influence the effectiveness of the PA process. Specifically, transformational leadership has been related to satisfaction with the PA process, whereas transactional leadership has been related to lower satisfaction (Waldman et al., 1987).

Both leadership and procedural justice are important contextual factors influencing the PA success; therefore, attention should be given to these issues while conducting PAs in an organization.

Summary

In this article, we have pointed out some landmarks in the PA literature and discussed the factors threatening the PA process success. Raters' judgment biases and lack of accuracy of the ratings threaten the PA validity. The organizational conditions, which often limit opportunities to observe all workers sufficiently, threaten the procedural justice. Also, the rater does not always master the delicate skills needed for providing effective feedback, and lastly, leadership biases, such as preferring in-group members, are also a possible threat to this process.

PA processes present numerous difficulties and obstacles (e.g., DeNisi et al., 1984; DeNisi and Peters, 1996; Feldman, 1981; Kluger and DeNisi, 1996), and it is safe to say that managers and workers are generally not very satisfied with it (e.g., Bouskila-Yam and Kluger, 2011; Coens and Jenkins, 2002; Peiperl, 2001; Smither et al., 2005). Yet, organizations do not seem to be in a hurry to cancel these evaluations, and existing alternatives are not being adopted. Is this an unfortunate mistake? Possibly, but it is also likely that one of the reasons why the PA process persists is that during it, workers are told who has been successful and who has not, who has reached the top and who is left behind, who has excelled, who is appreciated, and who has achieved something meaningful. By doing this, the PA process relates to the most basic motivational processes that drive people to work in the first place.

See also: Human Resource Management, Psychology of; Industrial–Organizational Psychology: Science and Practice; Job Design and Evaluation: Organizational Aspects; Organizational Behavior, Psychology of; Personnel Selection, Psychology of.

Bibliography

Antonioni, D., 1996. Designing an effective 360-degree appraisal feedback process. Organizational Dynamics 25, 24–38.

Arvey, R.D., Murfey, K.R., 1998. Performance evaluation in work settings. Annual Review of Psychology 49, 141–168.

Atwater, L., Roush, P., Fischithal, A., 1995. The influence of upward feedback on self-ratings and follower-ratings of leadership. Personnel Psychology 48 (1), 35–59.

Ball, G.A., Trevino, L.K., Sims Jr, H.P., 1994. Just and unjust punishment: influences on subordinate performance and citizenship. Academy of Management Journal 37 (2), 299–322.

Barling, J., Phillips, M., 1993. Interactional, formal and distributive justice in the workplace: an exploratory study. Journal of Psychology 127 (6), 649–656.

Bateson, M., Nettle, D., Roberts, G., 2006. Cues of being watched enhance cooperation in real-word setting. Biology Letters 2 (3), 412–414.

Boswell, W.R., Boudreau, J.W., 2002. Separating the developmental and evaluative performance appraisal uses. Journal of Business and Psychology 16 (3), 391–412.

Bouskila-Yam, O., Kluger, A.N., 2011. Strength-based performance appraisal and goal setting. Human Resource Management Review 21 (2), 137–147.

Bretz, R.D., Milcovich, G.T., Read, W., 1992. The current state of performance-appraisal research and practice: concerns, directions and implications. Journal of Management 18 (2), 321–352.

Brett, J.F., Atwater, L.E., 2001. 360 degrees feedback: accuracy, reactions and perceptions of usefulness. Journal of Applied Psychology 86 (5), 930–942.

Cardy, R.L., Dobbins, G.H., 1994. Performance Appraisal: A Consideration of Alternative Perspectives. South-Western Publishing, Cincinnati, OH.

Cawley, B.D., Keeping, L.M., Levy, P.E., 1998. Participation in the performance appraisal process and employee reactions: a meta-analytic review of field investigations. Journal of Applied Psychology 83 (4), 615–633.

Cleveland, J.N., Cropanzano, R., Hautaluoma, J., Murphy, K.R., Thornton, G.C., 1995. Industrial organizational psychology program, Colorado State University, USA. International Journal of Selection and Assessment 3 (4), 242–244.

Cleveland, J.N., Murphy, K.R., Williams, R.E., 1989. Multiple uses of performance appraisal: prevalence and correlates. Journal of Applied Psychology 74 (1), 130–135.

Coens, T., Jenkins, M., 2002. Abolishing performance appraisals: why they backfire and what to do instead. Berrett-Koehler Publishers, San Francisco.

Colquitt, J.A., Conlon, D.E., Wesson, M.J., Porter, C.O.L.H., Ng, K.Y., 2001. Justice at the millennium: a meta-analytic review of 25 years of organizational justice research. Journal of Applied Psychology 86 (3), 425–445.

Dulebohn, J.H., Ferris, G.R., 1999. The role of influence tactics in perceptions of performance evaluations' fairness. Academy of Management Journal 42 (3), 288–303.

DeNisi, A.S., Kluger, A.N., 2000. Implicit theories of performance as artifacts in survey research: replication and extension. Organizational Behavior and Human Performance 21 (3), 358–366.

DeNisi, A.S., Peters, L.H., 1996. Organization of information in memory and the performance appraisal process: evidence from the field. Journal of Applied Psychology 81 (6), 717–737.

DeNisi, A.S., Cafferty, T.P., Meglino, B.M., 1984. A cognitive view of performance appraisal process: a model and research propositions. Organizational Behavior and Human Performance 33 (3), 360–396.

DeNisi, A.S., Pritchard, R.D., 2006. Performance appraisal, performance management and improving individual performance: a motivational framework. Management and Organization Review 2 (2), 253–277.

Dobbins, G.H., Cardy, R.L., Truxillo, D.M., 1988. The effects of purpose of appraisal and individual differences in stereotypes of women on sex differences in performance ratings: a laboratory and field study. Journal of Applied Psychology 73 (3), 551–558.

Elicker, J.D., Levy, P.E., Hall, J.R., 2006. The role of leader-member exchange in the performance appraisal process. Journal of Management 32 (4), 531–551.

Erdogan, B., 2002. Antecedents and consequences of justice perceptions in performance appraisals. Human Resource Management Review 12, 555–578.

Erdogan, B., Kraimer, M.L., Linden, R.C., 2001. Procedural justice as a two-dimensional construct. Journal of Applied Behavioral Science 37 (2), 205–222.

Erez, M., 1977. Feedback: a necessary condition for the goal setting-performance relationship. Journal of Applied Psychology 62 (5), 624–627.

Fairhurst, G.T., 1993. The leader-member exchange patterns of women leaders in industry: a discourse analysis. Communication Monographs 60 (4), 321–351.

Feldman, J.M., 1981. Beyond attribution theory: cognitive-processes in performance appraisal. Journal of Applied Psychology 66 (2), 127–148.

Folger, R., Konovsky, M.A., 1989. Effects of procedural and distributive justice on reactions to pay raise decisions. Academy of Management Journal 32, 115–130.

Folger, R., Konovsky, M.A., Cropanzano, R., 1992. A due process metaphor for performance appraisal. In: Staw, B.M., Cummings, L.L. (Eds.), Research in Organizational Behavior, vol. 14. JAI press, pp. 129–177.

Gabris, G.T., Ihrke, D.M., 2001. Does performance appraisal contribute to heightened levels of employee burnout? The results of one study. Public Personnel Management 30 (2), 157–172.

Gerstner, C.R., Day, D.V., 1997. Meta-analytic review of leader–member exchange theory: correlates and construct issues. Journal of Applied Psychology 82 (6), 827–844.

Graen, G.B., Hui, C., Taylor, E.A., 2006. Experience-based learning about LMX leadership and fairness in project teams: a dyadic directional approach. Academy of Management Learning & Education 5 (4), 448–460.

Graen, G.B., Uhlbein, M., 1995. Relationship-based approach to leadership-development of leader-member exchange (LMX) theory of leadership over 25 years-applying a multilevel multidomain perspective. Leadership Quarterly 6 (2), 219–247.

Greenberg, J., 1986. Determinants of perceived fairness of performance evaluation. Journal of Applied Psychology 71 (2), 340–342.

Haley, K.J., Fessler, D.M.T., 2005. Nobody's watching? Subtle cues affect generosity in anonymous economic game. Evolution and Human Behavior 26 (3), 245–256.

Heslin, P.A., VandeWalle, D., 2011. Performance appraisal procedural justice: the role of manager's implicit person theory. Journal of Management 37 (6), 1694–1718.

Higgins, E.T., 1998. Promotion and prevention: regulatory focus as a motivational principle. In: Zanna, M.P. (Ed.), Advances in Experimental Social Psychology, vol. 30. Academic Press, New York, pp. 1–46.

Higgins, E.T., 1997. Beyond pleasure and pain. American Psychologist 52 (12), 1280–1300.

Holbrook Jr, R.L., 2002. Contact points and flash points: conceptualizing the use of justice mechanisms in the performance appraisal interview. Human Resource Management Review 12 (1), 101–123.

Ilgen, D.R., Barnes-farrell, J.L., Mckellin, D.B., 1993. Performance appraisal process research in the 1980s: what has it contributed to appraisals in use. Organizational Behavior and Human Decision Processes 54 (3), 321–368.

Ilgen, D.R., Davis, C.A., 2000. Baring bad news: reactions to negative performance feedback. Applied Psychology: An International Review 49 (3), 550–565.

Judge, T.A., Ferris, G.R., 1993. Social Context of performance evaluation decisions. Academy of Management Journal 36 (1), 80–105.

Kacmar, K.M., Witt, L.A., Zivnuska, S., Gully, S.M., 2003. The interactive effect of leader-member exchange and communication frequency on performance ratings. Journal of Applied Psychology, 88 (4), 764–772.

Kark, R., Van Dijk, D., 2008. Birds of a feather flock together: the relationship between leader-follower self-regulation congruency, LMX and outcomes. In: Graen, G.B., Graen, J.A. (Eds.), Knowledge Driven Corporation: A Discontinuous Model. LMX Leadership: The Series, vol. VI. Information Age Publishing Inc., Charlotte, NC.

Keeping, L.M., Levy, P.E., 2000. Performance appraisal reactions: measurement, modeling and method bias. Journal of Applied Psychology 85 (5), 708–723.

Keller, J., Pfattheicher, S., 2011. Vigilant self-regulation, cues of being watched and cooperativeness. European Journal of Personality 25 (5), 363–372.

Kluger, A.N., DeNisi, A., 1996. The effects of feedback intervention on performance: a historical review, a meta-analysis and a preliminary feedback intervention theory. Psychological Bulletin 119 (2), 254–284.

Kluger, A.N., Van Dijk, D., 2010. Feedback, the various tasks of the doctor, and the feedforward alternative. Medical Education 44 (12), 1166–1174.

Korsgaard, M.A., Schweiger, D.M., Sapienza, H.J., 1995. Building commitment, attachment and trust in strategic decision-making teams: the role of procedural justice. Academy of Management Journal 38 (1), 60–84.

Landy, F.J., Farr, J.L., 1980. Performance rating. Psychological Bulletin 87 (1), 72–107.

Levy, P.E., Williams, J.R., 2004. The social context of performance appraisal: a review and framework for the future. Journal of Management 30 (6), 881–905.

Locke, E.A., Latham, G.P., 2002. Building a practically useful theory of goal setting and task motivation: a 35 year odyssey. America Psychologist 57 (9), 705–717.

London, M., Smither, J.W., 1995. Can multi-source feedback change perceptions of goal accomplishment self-evaluations, and performance-related outcomes? Theory-based applications and directions for research. Personnel Psychology 48 (4), 803–839.

Mayer, R.C., Davis, J.H., 1999. The effect of the performance appraisal system on trust for management: a field quasi-experiment. Journal of Applied Psychology 84 (1), 123–136.

Mayer, D., Nishii, L., Schneider, B., Goldstein, H., 2007. The precursors and products of justice climates: group leader antecedents and employee attitudinal consequences. Personnel Psychology 60, 929–963.

Moorman, R.H., Balkely, G.L., Neihoff, B.P., 1998. Does perceived organizational support mediate the relationship between procedural justice and organizational citizenship behavior? Academy of Management Journal 41 (3), 351–357.

Murphy, K.R., Cleveland, J.N., 1991. Performance Appraisal: An Organizational Perspective. Allyn and Bacon, Boston.

Organ, D.W., Podsakoff, P.M., Mackenzie, S.B., 2006. Organizational Citizenship Behavior: Its Nature, Antecedents and Consequences. Sage, Thousand Oaks, CA.

Peiperl, M.A., 2001. Getting 360-degree Feedback Right. Harvard Business Review 79 (1), 142–147.

Poon, J.M.L., 2004. Effects of performance appraisal politics on job satisfaction and turnover intention. Personnel Review 33 (3), 322–334.

Purcell, J., Hutchinson, S., 2007. Front-line managers as agents in the HRM performance causal chain: theory, analysis and evidence. Human Resource Management Journal 17 (1), 3–20.

Reilly, R.R., Smither, J.W., Vasilopoulos, N.L., 1996. A longitudinal study of upward feedback. Personnel Psychology 49 (3), 599–612.

Roethlisberger, F.J., Dickson, W.J., 1939. Management and the Worker. Harvard University Press, Cambridge, MA.

Rynes, S.L., Gerhart, B., Parks, L., 2005. Personnal psychology: performance evaluation and pay for performance. Annual Review of Psychology 56, 571–600.

Skarlicki, D.P., Latham, G.P., 1996. Increasing citizenship within a union: a test of organizational justice theory. Journal of Applied Psychology 81 (2), 161–169.

Skarlicki, D.P., Latham, G.P., 1997. Leadership training in organizational justice to increase citizenship behavior within a labor union: a replication. Personnel Psychology 50 (3), 617–633.

Smither, J.W., London, M., Reilly, R.R., 2005. Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. Personnel Psychology 58 (1), 33–66.

Taylor, M.S., Masterson, S.S., Renard, M.K., Tracy, K.B., 1998. Managers' reactions to procedurally just performance management systems. Academy of Management Journal 41 (5), 568–579.

Van Dijk, D., Kluger, A.N., 2011. Task type as a moderator of positive/negative feedback effects on motivation and performance: a regulatory focus perspective. Journal of Organizational Behavior 32, 1084–1105.

Van Dijk, D., Kluger, A.N., 2004. Feedback sign effect on Motivation: is it moderated by regulatory focus? Applied Psychology: An International Review 53 (1), 113–135.

Waldman, D.A., Bass, B.M., Enstein, W.O., 1987. Leadership and outcomes of performance appraisal processes. Journal of Occupational Psychology 60 (3), 177–186.

Whiting, S.W., Posdakoff, P.M., Pierce, J.R., 2008. Effects of task performance, helping, voice and organizational loyalty on performance appraisal ratings. Journal of Applied Psychology 93 (1), 125–139.

Perinatal Psychiatry

Keiko Yoshida, Yumi Nishikii, and Hiroshi Yamashita, Fukuoka, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Perinatal psychiatry covers a wide range of issues that are related to maternal mental health and infant development. The maternal issues include unplanned pregnancy, lack of emotional support by their partners, and domestic violence during pregnancy. Onset of psychiatric disorders is more common in the perinatal period. Postnatal depression is more prevalent than other disorders. More or less 10% of new mothers suffer from depression, and the Edinburgh Postnatal Depression Scale is a well-known screening. The impact of interaction between a mother and a baby on later infant development should be considered, and baby care and family support are important for family well-being.

Introduction

Perinatal mental health is an important issue not only for pregnant and postnatal women but also for their infants and families. Problems widely range from anxiety concerning the delivery, which is commonly experienced, to severe chronic psychiatric disorders, which need formal psychiatric treatment.

Therefore support and treatment for postnatal women is carried out based on women's situations; however, emotional support is essential for all women. Issues to be considered include the mother's emotional bonding to her baby, mother and baby interaction, baby's attachment, and impact of perinatal mental problems on long-term prognosis.

Historically, perinatal mental illness has been recognized since the time of Hippocrates. In the nineteenth century, medical interest in perinatal mental illness increased, in current scope of a general interest in mental illness. Marce, who was devoted to the study of perinatal mental illness, published an important series of case studies of women suffering from various forms of perinatal mental illness over 150 years ago. During the twentieth century, clinical accounts and research on prevalence, risk factors, and consequences to the mother, fetus, and infant expanded dramatically. Recent studies on perinatal mental health have also been influenced by the fetal programming hypothesis. The dynamic mix of biological, socioenvironmental, and psychological factors affects the expression of mental health and illness across the life span. Pregnancy can also be a time of increased stress for women, and the impacts of domestic abuse and relationship strain are related to psychosocial vulnerability not only for women but also for the fetus and developmental prognosis of children. The environment in the womb, during different sensitive periods for specific outcomes, can alter the development of the fetus, permanently affecting the child.

Psychiatric Disorders in the Perinatal Period

Mental illness among women is prevalent in the perinatal period, which starts from pregnancy to postpartum. In a narrow definition, postpartum covers a few weeks after delivery and many studies on postpartum illness consider up to three postnatal months. However, in terms of maternal subjective negative feelings, lack of confidence as a mother, and disturbed mother and infant interactions that last longer even after their symptoms improved, the postnatal period should covered the period of one year after delivery. Some illnesses have new onset in particular during the perinatal period due to drastic psychosocial and relationship changes for women's roles as well as biological changes including rapid hormonal changes. Postnatal depression in particular has a high incidence of more than 10% of all new mothers. Some illnesses may relapse in this period; for example, bipolar illness tends to reoccur during the perinatal period partially because women stop medication in fear of adverse effects on their fetus and obstetric outcome. Also chronic illnesses such as schizophrenia need to be considered as illnesses not only for the possibility of relapse but also for their insufficient maternal skill and less active interaction with their babies. The professionals who care for and treat women with perinatal mental illness should consist of a multidisciplinary team of obstetric, pediatric, and psychiatric medical staff as well as health visitors and social workers in the community.

Anxiety Disorders

A wide range of anxiety disorders are prevalent in the perinatal period, including generalized anxiety, obsessive–compulsive, panic, and social anxiety disorders, as well as depression. Rates of anxiety disorders in the perinatal period vary widely; however, these disorders are less common compared to depression disorders. However, women with anxiety are also depressive at some perinatal point. In a large-scale postpartum screening study with follow-up diagnostic evaluations (Wisner et al., 2013) 66% of women with major depression in the postpartum period had comorbid anxiety disorders. Anxiety and depression are common and they are more prevalent in the perinatal than postnatal period.

Recent findings show that cortisol and pregnancy-specific anxiety independently predicted child anxiety. Children exposed to elevated prenatal maternal cortisol and pregnancy-specific anxiety are at increased risk for developing anxiety problems during the preadolescent period (Davis and Sandman, 2012).

Depression

The incidence and prevalence of prenatal depression has been reported to be from 6% to 38% (Field, 2011); compared to

 http://dx.doi.org/10.1016/B978-0-08-097086-8.27082-7

postnatal depression, prenatal depression was almost twice that of postpartum depression, and the incidence rate was 29% and 17%, respectively (Andersson et al., 2006). Prenatal depression enhances the risk of preterm delivery, which will be related to significant behavioral, cognitive, and emotional problems in children (Li et al., 2009). High prenatal cortisol levels associated with depression may result in elevated fetal cortisol and delayed fetal growth and prematurity (Van den Bergh et al., 2005; Davis et al., 2007; Lundy et al., 1999). Also prenatal depression significantly predicts postpartum depression (Milgrom et al., 2008); therefore these women should be monitored continuously throughout their pregnancies. The risk factors of prenatal depression are past history of depression, lack of emotional and social support from their partner or others, poverty, marital difficulties, increased life stress, substance abuse, a history of previous abortion and ambivalence toward the pregnancy, and anxiety about the fetus. Women with low self-esteem, prenatal anxiety, negative cognitive style, major life events, and a history of abuse are also at risk (Leigh and Milgrom, 2008). Of these items, past history of depression, lack of emotional and social support from a woman's partner or others, and severe life events are also risk factors of postnatal depression. Other relevant risk factors that have smaller associations with postpartum depression include low socioeconomic status, being single, unwanted pregnancy, obstetrical stressors, and difficult infant temperament (O'Hara, M.W. and Wisner, K.L., 2014).

The mood of women after delivery is labile and most of them can be weepy, especially in the first 10 days after delivery. Women tend to cry for unexplained sadness or feel a little bit high. This phenomenon is called postpartum blues. It is very common, and more than half of postnatal women experience crying dysphoric mood or sometimes an elevated high feeling, insomnia, anxiety, irritability, loss of appetite, and headaches. These phenomena usually disappear within a couple of days and nonspecific treatment is necessary.

Postnatal depression is common among women, and while they are celebrated for a baby's birth, their depressed feelings are not disclosed easily. Incidence is higher than the one in the non-postnatal period, and diagnosis of depression and antidepressant prescriptions are increased to more than 14% (Davé, S. et al., 2010). Apart from the risk factors mentioned earlier, the relevant risk factors that have smaller associations with postpartum depression include low socioeconomic status, being single, unwanted pregnancy, obstetrical stressors, and difficult infant temperament (O'Hara, M.W. and Wisner, K.L., 2014).

Screening for Postnatal Depression

Most mothers with postnatal depression are unlikely to access psychiatric care, even though their symptoms are serious (McIntosh, 1993). Therefore, neonatal home visits by health visitors are regarded as a potentially useful opportunity for detecting mothers with postnatal depression. A home visit system by community health visitors has been organized in many countries. In the past, the purpose of the visit was to promote infant growth and development. Health visitors tend to focus on the baby's physical health rather than on any potential psychological problems that the mothers may have. Underestimating mothers' mental health seemed to be common and mental problems were easily overlooked even by those professionals unless a mother asked for their help.

The Edinburgh Postnatal Depression Scale is most commonly used for screening by midwives and health visitors. Women are often asked to complete this 10-item questionnaire that asks about depression and anxiety-related symptoms over the recent 7 days. Scores can range from 0 to 30, and women whose scores exceed a specified threshold (usually between 10 and 13 in European and American countries, but it depends on their culture and countries) are often referred for further evaluation or treatment for a perinatal depressive or anxiety disorder (Cox, J. et al., 1987).

Puerperal Psychosis

Postpartum psychosis occurs rapidly after birth, usually within a few days after delivery. The symptoms are mood fluctuation, confusion, delirium, bizarre behavior, insomnia, and visual and auditory hallucinations.

Recent studies showed that bipolar disorder is very important to be recognized and treated properly, because half of the suffering women are undiagnosed (Sharma, V. et al., 2009). Mothers could kill their babies due to their delusions or distorted beliefs. Women are vulnerable to affective psychoses after birth more than at any other time during their lives (Munk-Olsen, T. et al., 2009).

Childbirth was associated with an increased risk of readmission during the first postpartum month, and a previous diagnosis of bipolar affective disorder was the strongest predictor of readmission. Postnatal women with current chronic psychosis need to be monitored early in the postnatal period, especially in the first month after delivery. With regard to bipolar disorders, Munk-Olsen et al. (2009) reported a population-based study of 28 124 women, of which they reported that the prevalence of incident admissions among new mothers was 0.1% at up to 4 months postpartum and readmission due to recurrence of a mental disorder occurred even more often.

Suicide and Infanticide

The United Kingdom Confidential Enquiry into All Maternal Deaths between 1997 and 1999 revealed that 10% of postpartum deaths resulted from suicide, and 12% of maternal deaths resulted from psychiatric causes such as accidental overdose, deaths from medical conditions due to substance use, and violent deaths (Oates, 2003). Furthermore, maternal suicide can be associated with a risk of infanticide (Brockington, 2001). It is a rare event; therefore intervention and treatment should be carried out.

Mother and Infant Interaction and Related Problems

Maternal Bonding Disorder toward Her Baby

Maternal emotional involvement with the baby in the perinatal period has been recognized and named 'bonding' and has been mainly discussed in relation to postnatal depression. Although some depressed mothers maintain positive maternal attitudes toward their newborn babies, a proportion of depressed

mothers experience severe negative affection toward their infants (Kumar 1997). Preliminary work by Kumar (1997), using items selected from mothers' narrative accounts, resulted in the development of a self-report screening scale, the Mother–Infant Bonding Questionnaire (MIBQ). This scale has nine items: loving, resentful, neutral or felt nothing, possessive, joyful, dislike, protective, disappointed, and aggressive. The MIBQ was later modified by Kumar's colleague Marks. Compared with the original questionnaire, the wording was changed in three items and the item 'scared or panicky' was newly added, resulting in a total of 10 items (Yoshida et al., 2012).

Based on interview responses, Brockington et al. (2006a) interviewed severe disorders of the mother–infant relationship in over 200 mothers. They found that depressed mothers had normal or disturbed mother–infant relationships and nearly one-third of the mothers had pathological anger toward the baby, and more or less the same rate of mothers had rejection or threatened rejection of the baby. They developed a questionnaire regarding bonding disorder and reported its validation (Brockington et al., 2006b).

Maltreatment and Poor Mothering by Mentally Ill Mothers

The multiple psychosocial difficulties experienced by women with mental disorders can have adverse effects on the development of mother–infant attachment and the child.

Kumar et al. (1995) investigated 100 consecutive admissions to a mother-and-baby unit. They reported that the proportion separated from their infant on discharge was 50% among the group with schizophrenia, and the proportion was much higher compared to the group with affective psychosis or nonpsychotic disorders (8% and 4%, respectively). The babies born to the mothers with schizophrenia had lower birth weights compared to the babies born to mothers with affective disorders, and with poor parenting, the developmental outcomes of those infants are worse compared to those of the other groups of mothers (Yoshida et al., 1999).

Adverse Impact of Maternal Psychosocial Vulnerability and Mental Illness on Child Development and Outcome

Maternal mental illnesses and various psychosocial difficulties during the perinatal period affect both the physical and mental development of a newborn. Prenatal depression has been associated with excessive activity and growth delays in the fetus as well as prematurity, low birth weight, disorganized sleep, and less responsiveness to stimulation in the neonate (Field, 2011). Over the past 15 years, several studies have examined associations between postnatal depression and child outcome longitudinally (Sutter-Dallay, A.L. et al., 2011) Depressed mothers lacked responsiveness to infant cue, that leads to difficulties in mother–infant relationship in the first year, as well as insecurity in the mother–infant relationship at 18 months (Murray, L., 2009). The association between postnatal depression and children's cognitive outcome may be related to the severity, persistence, or recurrence of maternal depressive symptoms beyond the postpartum period (Sutter-Dallay, A.L. et al., 2011). Long-term support and intervention for both mothers and infants can be ideal and effective.

Treatment and Service Plan

Perinatal mental illness refers to psychiatric disorders that are prevalent during pregnancy and as long as one year after delivery. Therefore, a multidisciplinary approach is necessary. The obstetric staff should be involved first in evaluation, detection, and preventive care for women at risk or with mental illness. Obstetricians and midwives should be aware of risk factors for perinatal mental illness, care and treatment, and referral options. However, most women at risk or who already have mental disorders would not actively ask for support. Therefore screening and early intervention are necessary. Quite often the different specialists have to collaborate with each other from the time of pregnancy and hand over their evaluations and care reports to other staff.

Psychosocial Interventions

Obstetricians and hospital or community midwives can identify women with past or present mental illness including schizophrenia, other psychoses, depression, substance abuse, a family history of perinatal psychiatric disorder, as well as psychosocially vulnerable women suffering domestic violence. Women identified at risk need to be referred to specialist perinatal mental health professionals for further assessment, specific interventions, and the monitoring of their mental health both during pregnancy and postpartum.

Psychological therapies are usually recommended as the first line of treatment for mild-to-moderate postnatal depression for women with no previous history of mental health conditions. Home visits by community-based health visitors can be carried out as a psychosocial intervention of psychiatric illness. Psychoeducation and emotional support for the partner and other family members is important, and family members should be kept informed of the patient's condition, treatment plan, and expected outcome. Help with nocturnal infant care should be recommended to minimize the mother's sleep disruption. Once the mother's condition is stable, issues such as parenting skills and maternal–infant bonding should be addressed. Cognitive behavioral therapy (CBT) and interpersonal therapy (IPT) (Dennis, C.-L. and Dowswell, T., 2013; Milgrom, J. et al., 2005; Kopelman, R. and Stuart, S., 2005) are effective for depression and maybe for other psychological and psychiatric disorders. CBT is a type of therapy based on the idea that unhelpful and unrealistic thinking leads to negative behavior. CBT aims to break this cycle and find new ways of thinking that can help a woman behave in a more positive way. IPT aims to identify whether her relationships with others may be contributing toward feelings of depression.

Pharmacological Management

Selective serotonin reuptake inhibitors (SSRIs) are advised for treatment for pre- and postnatal depression. Rapid and abrupt hormonal changes over the course of delivery also contribute to onset of depression.

Estradiol levels rapidly decrease due to placenta delivery. Monoamine oxidase A levels increase and neurotransmitters such as serotonin, dopamine, and norepinephrine are depleted, which are the triggers of the onset of depression (Sacher et al.,

2010). Three neurotransmitters have been implicated in prenatal depression, including elevated norepinephrine and low levels of dopamine and serotonin, which were noted in depressed pregnant women (Lundy et al., 1999). SSRIs are used for pharmacological treatment. Very few adverse effects of SSRIs are reported such as neonatal withdrawal syndrome (Sanz et al., 2005) and infant pulmonary hypertension (Chambers et al., 2006). Discontinuation of antidepressants or antipsychotic drugs can often lead to relapse. If women are very reluctant to take medication in pregnancy, it is recommended to take medication immediately after delivery.

In case of bipolar depression, treatment with antidepressants resulted in a highly unstable illness course, and discontinuation of antidepressants and initiation of mood stabilizers and atypical antipsychotics resulted in sustained improvement (Sharma et al., 2009). The significant impacts of prophylactic treatment with mood stabilizers and antipsychotics on postpartum mood episodes in women with bipolar I or II disorder have been reported in clinical studies. Treatment during lactation requires minimizing infant exposure and adverse effects while maintaining optimum maternal mental health. In women with depressive disorder or bipolar disorder, the benefits of breast-feeding should also be carefully balanced against the deleterious effect of sleep deprivation in triggering mood episodes.

The effectiveness of the use of lithium, carbamazepine, divalproex, and olanzapine was reported in the reduction of postpartum recurrences. Carbamazepine and valproic acid are reported to be teratogenic and should be avoided in only early pregnancy. Women on lithium should have a slow reduction of lithium to reduce the risk of relapse. In these women, the risk of untreated psychiatric illness outweighs any risk to the fetus from exposure to psychotropic medications (Oyebode et al., 2012). The most important strategy is to evaluate maternal mental problems and symptoms and try to reduce their problems.

Treatment might also include complementary and alternative treatments, such as exercise, bright light therapy, and yoga. Massage therapy and psychotherapy reduce cortisol levels and they have been tried for women with postnatal depression. Treatment is usually on an outpatient basis, although psychiatric emergencies, such as suicidal thoughts with a plan and intent, or postpartum psychosis, warrant hospitalization.

Conclusion

Maternal prenatal and postnatal psychiatric disorders are related to their psychosocial and biological vulnerability, which disturb mother and infant interaction and baby care function, and those are a negative impact on infants' later development and their well-being in adolescence. This long-term influence should be noted when caring for and treating women with perinatal psychiatric disorders because those adolescents will be the parents at risk in the next generation. Psychiatry is a cultural issue as well as medical one (Bhugra and Bahl, 1999). Strategies to deal with perinatal psychiatry in particular need to be focused on patients' cultural backgrounds. A life cycle and cultural perspective is important to addressing the challenges of this complex period.

Bibliography

Andersson, L., et al., 2006. Depression and anxiety during pregnancy and six months postpartum: a follow-up study. Acta obstetricia et gynecologica Scandinavica 85 (8), 937–944. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16862471.

Bhugra, D., Bahl, V., 1999. Ethnicity: An Agenda for Mental Health. Royal College of Psychiatrists, London. ISBN 1-901242-15-3.

Brockington, I., 2001. Suicide in women. International Clinical Psychopharmacology 16 (2), S7–S19.

Brockington, I.F., Aucamp, H.M., Fraser, C., 2006a. Severe disorders of the mother-infant relationship: definitions and frequency. Archives of Women's Mental Health 9 (5), 243–251. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16699838.

Brockington, I.F., Fraser, C., Wilson, D., 2006b. The postpartum bonding questionnaire: a validation. Archives of women's mental health 9 (5), 233–242. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16673041.

Chambers, C.D., et al., 2006. Selective serotonin-reuptake inhibitors and risk of persistent pulmonary hypertension of the newborn. Obstetrical & Gynecological Survey 61 (6), 370–371. Available at: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00006254-200606000-00009.

Cox, J.L., Holden, J.M., Sagovsky, R., 1987. Detection of postnatal depression. Development of the 10-item Edinburgh Postnatal Depression Scale. The British Journal of Psychiatry: The Journal of Mental Science 150, 782–786.

Davé, S., et al., 2010. Incidence of maternal and paternal depression in primary care: a cohort study using a primary care database. Archives of Pediatrics & Adolescent Medicine 164 (11), 1038–1044. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20819960.

Davis, E.P., et al., 2007. Prenatal exposure to maternal depression and cortisol influences infant temperament. Journal of the American Academy of Child and Adolescent Psychiatry 46 (6), 737–746. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17513986.

Davis, E.P., Sandman, C.A., 2012. Prenatal psychobiological predictors of anxiety risk in preadolescent children. Psychoneuroendocrinology 37 (8), 1224–1233. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3356791&tool=pmcentrez&rendertype=abstract.

Dennis, C.-L., Dowswell, T., 2013. Psychosocial and psychological interventions for preventing postpartum depression. Cochrane Database of Systematic Reviews (Online) 2, CD001134. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23450532.

Field, T., 2011. Prenatal depression effects on early development: a review. Infant Behavior & Development 34 (1), 1–14. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20970195.

Kopelman, R., Stuart, S., 2005. Psychological treatments for postpartum depression. Psychiatric Annals 64 (1), 103–118. Available at: http://psycnet.apa.org/psycinfo/2006-04164-005.

Kumar, R., et al., 1995. Clinical survey of a psychiatric mother and baby unit: characteristics of 100 consecutive admissions. Journal of Affective Disorders 33 (1), 11–22. Available at: http://linkinghub.elsevier.com/retrieve/pii/016503279400067J.

Kumar, R.C., 1997. "Anybody's child": severe disorders of mother-to-infant bonding. The British Journal of Psychiatry 171 (2), 175–181. Available at: http://bjp.rcpsych.org/cgi/doi/10.1192/bjp.171.2.175.

Leigh, B., Milgrom, J., 2008. Risk factors for antenatal depression, postnatal depression and parenting stress. BMC Psychiatry 8, 24. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2375874&tool=pmcentrez&rendertype=abstract.

Li, D., Liu, L., Odouli, R., 2009. Presence of depressive symptoms during early pregnancy and the risk of preterm delivery: a prospective cohort study. Human Reproduction (Oxford, England) 24 (1), 146–153. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18948314 (accessed 28.07.14.).

Lundy, B.L., et al., 1999. Prenatal depression effects on neonates. Infant Behavior and Development 22 (1), 119–129. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0163638399800095.

McIntosh, J., 1993. Postpartum depression: women's help-seeking behaviour and perceptions of cause. Journal of Advanced Nursing 18 (2), 178–184. Available at: http://doi.wiley.com/10.1046/j.1365-2648.1993.18020178.x.

Milgrom, J., et al., 2005. A randomized controlled trial of psychological interventions for postnatal depression. The British Journal of Clinical Psychology/The British Psychological Society 44 (Pt 4), 529–542. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16368032.

Milgrom, J., et al., 2008. Antenatal risk factors for postnatal depression: a large prospective study. Journal of Affective Disorders 108 (1–2), 147–157. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18067974.

Munk-Olsen, T., et al., 2009. Risks and predictors of readmission for a mental disorder during the postpartum period. Archives of General Psychiatry 66 (2), 189–195. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19188541.

Murray, L., 2009. The development of children of postnatally depressed mothers: evidence from the Cambridge longitudinal study. Psychoanalytic Psychotherapy 23 (3), 185–199. Available at: http://www.tandfonline.com/doi/abs/10.1080/02668730903227289.

O'Hara, M.W., Wisner, K.L., 2014. Perinatal mental illness: definition, description and aetiology. Best Practice & Research: Clinical Obstetrics & Gynaecology 28 (1), 3–12. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24140480.

Oates, M., 2003. Suicide: the leading cause of maternal death. The British Journal of Psychiatry: The Journal of Mental Science 183, 279–281.

Oyebode, F., et al., 2012. Psychotropics in pregnancy: safety and other considerations. Pharmacology & Therapeutics 135 (1), 71–77. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22483705.

Paschetta, E., et al., 2014. Perinatal psychiatric disorders: an overview. American Journal of Obstetrics and Gynecology 210 (6), 501–509.e6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24113256.

Sacher, J., et al., 2010. Elevated brain monoamine oxidase A binding in the early postpartum period. Archives of General Psychiatry 67 (5), 468–474. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20439828.

Sanz, E.J., et al., 2005. Selective serotonin reuptake inhibitors in pregnant women and neonatal withdrawal syndrome: a database analysis. Lancet 365 (9458), 482–487. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15705457.

Sharma, V., Burt, V.K., Ritchie, H.L., 2009. Bipolar II postpartum depression: detection, diagnosis, and treatment. The American Journal of Psychiatry 166 (11), 1217–1221. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19884236.

Sutter-Dallay, A.L., et al., 2011. A prospective longitudinal study of the impact of early postnatal vs chronic maternal depressive symptoms on child development. European Psychiatry: The Journal of the Association of European Psychiatrists 26 (8), 484–489. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20621453.

Van den Bergh, B.R.H., et al., 2005. Antenatal maternal anxiety and stress and the neurobehavioural development of the fetus and child: links and possible mechanisms. A review. Neuroscience and Biobehavioral Reviews 29 (2), 237–258. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15811496.

Wisner, K.L., et al., 2013. Onset timing, thoughts of self-harm, and diagnoses in postpartum women with screen-positive depression findings. JAMA Psychiatry 70 (5), 490–498. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23487258.

Yoshida, K., et al., 2012. A Japanese version of Mother-to-Infant Bonding Scale: factor structure, longitudinal changes and links with maternal mood during the early postnatal period in Japanese mothers. Archives of Women's Mental Health 15 (5), 343–352. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3443344&tool=pmcentrez&rendertype=abstract.

Yoshida, K., et al., 1999. Sensorimotor and cognitive development of infants of mothers with schizophrenia. The British Journal of Psychiatry: The Journal of Mental Science 175, 380–387. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10789307.

Relevant Websites

Beyond Blue, 2011. Clinical Practice Guidelines for Depression and Related Disorders: Anxiety, Bipolar Disorder, and Puerperal Psychosis in the Perinatal Period. A Guideline for Primary Care Health Professionals. The National Depression Initiative, Melbourne. http://www.beyondblue.org.au/resources/health-professionals/clinical-practice-guidelines/perinatal-clinical-practice-guidelines.

http://marcesociety.com/ – Marcé Society for Perinatal Mental Health.

National Institute for Health and Clinical Excellence (NICE), 2007. Antenatal and postnatal mental health: clinical management and service guidance. In: NICE Clinical Guidelines, No. 45. British Psychological Society, Leicester, UK. Available at: http://guidance.nice.org.uk/cg45 (accessed 12.03.13).

http://www.waimh.org/ – World Association for Infant Mental Health.

Period and Cohort Analysis in Demography

Yang C Yang, Department of Sociology and Lineberger Cancer Center, Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
Ryan K Masters, Department of Sociology and Institute of Behavioral Science, University of Colorado at Boulder, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the long-standing analytic problem in period and cohort analysis in demography about the conceptualization, estimation, and interpretation of the differential contributions of distinct time-related changes attributed to period- and cohort-based sources of change in demographic phenomena. It introduces innovative approaches that emerged in recent methodological and substantive research on this problem using three common research designs and the generalized linear mixed effects modeling framework. It also outlines some promising areas of study that employ and extend these approaches in various important topics in contemporary demography.

Introduction

Demography is fundamentally concerned with time-varying phenomena. Key components of population changes such as fertility, mortality, and migration are always measured as age- and time-specific events. While age is by far the most important source of temporal variation in vital rates, historical time and birth cohort are the other two dimensions of time that have preoccupied much attention in statistical studies of human populations. Indeed, the vast majority of quantitative demographic research accounts for period-based or cohort-based temporal processes, with researchers making implicit assumptions about each temporal dimension's role in driving population change. Such assumptions usually translate to fitting analytic models that favor one dimension over the other, with the other excluded outright. A question that cannot be escaped, however, is the tenability of such assumption, which can only be tested in analyses that simultaneously include measures of both period-based and cohort-based changes. In this article, we review the long-standing analytic problem in demography about the conceptualization, estimation, and interpretation of the differential contributions of distinct time-related changes in demographic phenomena, especially period- and cohort-based sources of change. We introduce innovative approaches that emerged in recent methodological and substantive research on this problem. We refer interested readers to a recent book that introduces new guidelines for cohort analysis by Yang and Land (2013) for greater details. Finally, we outline some promising areas of study that employ and extend these approaches, and breathe new life into period and cohort analysis in demography.

Conceptualizations

A fundamental characteristic of demographic research is the study of age-specific data recorded at different points in time and from different birth cohorts. As illustrated in **Table 1**, a typical data set takes the form of an age (a30, a40, a50, a60) by period (p1970, p1980, p1990, p2000) array of outcomes where birth cohorts fall on the diagonals (c1910, c1920, … c1970). What often demands demographers' attention is the depiction of variation in such data along one or more of these temporal dimensions: (1) the ages of the individuals at the time of observation, termed 'age' (A) effects, (2) the time periods of observation or measurement of the outcome, termed 'period' (P) effects, and (3) the year of birth or another shared life event for a set of individuals, termed 'cohort' (C) effects.

Since the discovery of the Gompertz law, studies of the A effects have long existed in the history of population science. Studies that jointly consider P and C variations as distinct entities from A effects, termed as the 'age-period-cohort' (APC) analysis, appeared relatively recently (see, e.g., Hobcraft et al., 1982, for a review). Of central focus in the APC analysis is the distinction of P and C effects as the other two components of temporal variation in demographic outcomes. The meanings of P and C effects can be most differentiated by contrasting them in specific demographic contexts. Here we focus on their conceptualizations in the context of health, chronic disease, and mortality.

Period (P) effects are variations over time periods or calendar years that simultaneously influence all individuals. They subsume a complex set of historical events and environmental factors such as world wars, famine, and pandemics of infectious diseases, rapid adoption of technological advances, or economic expansions and contractions. Period effects are often evident from a correspondence in the timing of changes in outcomes and conditions that induced these outcomes. They can also arise from changes in disease classification or diagnostic techniques. For example, increases in breast cancer death rates from 1970 to 2000 were found to result from increases in mammography usage, and hence detections of malignant cases during this period, whereas the subsequent downturn in death rates starting around year 2000 was due to the sudden reductions of hormone replacement therapy usage (Yang and Land, 2013: Figures 8.7

Table 1 Hypothetical data arrayed by age, period, and cohort

	p1970	*p1980*	*P1990*	*p2000*
a30	c1940	c1950	c1960	c1970
a40	c1930	c1940	c1950	c1960
a50	c1920	c1930	c1940	c1950
a60	c1910	c1920	c1930	c1940

and 8.9). Similar period-based changes in mortality from heart disease have been observed in the U.S. population following the introduction and widespread use of pharmacological technologies to manage cholesterol (e.g., statins), hypertension (e.g., beta blockers and other antihypertensive drugs), and myocardial infarction (e.g., stents) (Lloyd-Jones et al., 2009).

Cohort (C) effects are also observed in many processes related to population health and mortality. A birth cohort ages together, encountering the same historical and social events at the same stage of the life course. The exit of an old cohort through death is constantly being replenished by the entry of a new cohort through birth, thereby continuously shaping and reshaping a population's composition. This continual process of cohort succession and replacement has long been recognized as an essential component of 'demographic metabolism' driving social change (Ryder, 1965). Further, variation across cohorts can arise from differential exposure to early life conditions and life-long accumulation of exposures to physiological, socioeconomic, behavioral, and environmental risk factors from gestation to old age (Fogel and Costa, 1997; Ben-Shlomo and Kuh, 2002; Finch and Crimmins, 2004). Importantly, cohort membership operates as a contextual characteristic for the understanding of the heterogeneous experiences of aging. Indeed, the 'principle of cohort differences in aging' posits that each cohort ages across the life course in unique and identifiable ways (Riley, 1987).

Challenges to Estimation and Interpretation

The conceptualizations of period and cohort effects suggest that these two are distinct sources of demographic variation driven by different population processes. However, such distinctions are often difficult to discern in empirical studies. As a result, period and cohort effects are frequently confounded with each other and also confounded with the age effects. We now briefly explain challenges to the estimation and interpretation of period and cohort effects in applications of conventional methods of APC analysis.

Descriptive Analysis

The vast majority of demographic studies that are concerned with time trends rely on descriptive techniques that provide qualitative impressions about temporal patterns. One common practice uses summary measures that are independent of age composition, such as standardized indices (e.g., age-standardized death rates) arrayed by time periods. An alternative device is a graphical display of the two-way table of age-period-specific rates or age-cohort-specific rates. Using **Table 1** as an example, one can plot the age-specific data for a30 to a60 for each of the time period, p1970 to p2000, and compare how the age patterns differ across the periods. Plotting the age-specific data for each of the birth cohort, c1910 to c1970, would similarly enable the comparison of cohort differences in the age patterns.

Although comparisons of trends based on these descriptive measures almost always have implications for the P and C effects underlying the observed data, neither approach explicitly adopts a formal modeling framework for the attribution of differential contributions of each to the phenomena under study (Hobcraft et al., 1982). These types of descriptive analyses share two disadvantages. First, they only describe variation in the demographic outcome associated with period or birth cohort, thereby making implicit assumptions about the lack of influence of the omitted dimension. Thus, the possibility of confounding of the two factors cannot be tested. Second, neither approach summarizes the corresponding data table satisfactorily. Standardization ignores different trends at different ages and graphical methods represent all available data, providing no summary at all (Osmond, 1985). Moreover, each approach has its own limitations. Standardized rates by definition depend on the age composition of a particular period that is used as the standard, and give more weight to rates at older ages. The substantive conclusion about time changes in these rates can thus be sensitive to the choice of the standard, which might have changed drastically over the past few decades due to the substantial population aging of many populations. The use of a more recent year as the standard would yield higher rates than using a younger population in a much earlier year.

The graphical presentation of all data can also produce inconsistent conclusions. For example, **Figure 1** shows the age-specific rates of lung cancer incidence for white females for years 1969–2007. The age curve in any given time period, say 2000–2004, cuts across a number of birth cohorts, including 1920, 1925, and 1930. Therefore, the shape of the period curve is affected by both varying age effects and cohort effects. It is not possible to determine how these effects operate simultaneously to shift period curves. To obtain a quantitative assessment of the source of change in the data, one needs to use statistical regression modeling (Kupper et al., 1985).

Linear Regression Models for APC Analysis

Linear regression models have been long used in quantitative social science research to account for the variation in demographic phenomena. The standard models for the APC analysis of age-by-time period data in the form of **Table 1** were known as the APC accounting/multiple classification models, which were introduced by Mason et al. (1973) as a general methodology for analysis when age, period, and cohort variables are all potentially of interest. The general form of the model can be written in linear regression form as:

$$f(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k \qquad [1]$$

where y_{ijk} denotes the observed outcome for the ith age group for $i = 1, \ldots, a$ age groups at the jth time period for $j = 1, \ldots, p$ time periods and of the kth cohort for $k = 1, \ldots, (a + p - 1)$ cohorts, with $k = a - i + j$; $f(.)$ is a simple transformation (linear, log, or logit); μ denotes the intercept or adjusted mean; α_i denotes the ith row age effect or the coefficient for the ith age group; β_j denotes the jth column period effect or the coefficient for the jth time period; γ_k denotes the kth diagonal cohort effect or the coefficient for the kth cohort.

Methodological Problems

In spite of its theoretical merits and conceptual relevance, conventional APC analysis suffers from a major methodological problem, which is inevitably a function of applying the

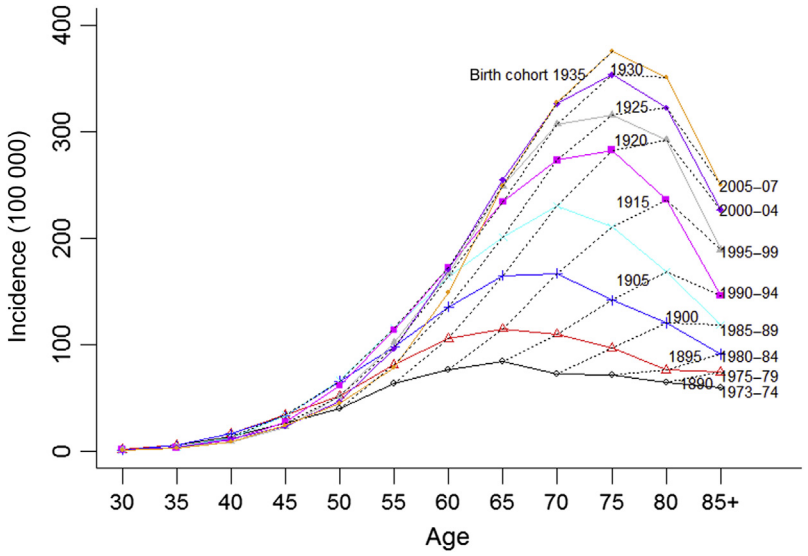

Figure 1 Age-specific lung cancer incidence rates for white females by time period and birth cohort.

linear regression model specified in eqn [1]. The "identification problem" occurs when both of the following conditions are satisfied (Mason et al., 1973): (1) A, P, and C are linearly related to each other, that is, A + C = P; and (2) each variable is postulated to be linearly related the outcome y_{ijk}. In the application of the linear models to data organized in the particular structure as shown in **Table 1**, this problem can be viewed as a special case of collinear regressors that produces a singular matrix (of one less than full rank) used in the statistical estimation process. Since a singular matrix produces multiple estimators of the three effects, it is statistically impossible to separately estimate the unique, true effects of each temporal dimension. Greater algebraic detail of this problem is available elsewhere (Yang and Land, 2013). The linear models dominated APC analysis for so long that analysts seldom reflect upon their deficiencies. There is confusion about the origin of the problem that it is inherent in the data because the A, P, and C variables are almost always exactly linear when measured in the same interval length, as is the case for major demographic data on vital rates. As a result, previous studies were almost exclusively focused on finding solutions to break condition (1) while paying little attention to condition (2). This will become clear in our discussion of conventional solutions below. A new insight into the old problem, therefore, is that the problem is model specific rather than data specific.

Conventional Approaches and Limitations

One commonly adopted approach is to estimate a reduced age and period or age and cohort two-factor model. This strategy does not incur the identification problem and simplifies the estimation considerably. The greatest challenge to the validity of this approach is the potential spuriousness of the estimated period or cohort effects due to the other effect being omitted from the model. For example, early studies of mortality infrequently tested cohort effects. More recent studies have found results strongly suggesting its presence, implying that certain assumptions previously used by demographers to analyze mortality trends may have been incorrect. Indeed, the recent evidence suggests that rates of U.S. mortality declines were likely associated with period-specific conditions – such as economic advance and health care technology – as well as being differentially experienced by birth cohorts. The latter finding implicates cohort-related changes in early-life conditions, cohort differences in intergenerational transfers of health capital, and/or cohorts' disparate exposures to health-enhancing technologies in recent historical declines in U.S. mortality levels (Finch and Crimmins, 2004; Fogel and Costa, 1997; Masters, 2012; Masters et al., 2012, in press; Yang, 2008a). Consequently, the omission of cohort effects from mortality models concerned with trends would likely be misspecified and inconsistent with accumulating evidence for cohort changes in population health.

Another widely adopted approach to estimating eqn [1] is the constrained generalized linear models (CGLIMs) approach. Here, one or more additional identifying constraint (usually an equality constraint) can be placed on the parameter vector to just identify or overidentify the model. Using the data in **Table 1** as an example, letting the effects of a30 and a40 be equal would just identify the model and yield unique estimates for A, P, and C effects. Grouping multiple ages, periods, or cohorts into less-refined categories would be one kind of overidentifying constraint that would also yield unique estimates of these effects. Discussions of problems with this CGLIM approach can be found elsewhere (Mason et al., 1973; Yang et al., 2004). The major limitations included that the choice of constraints relies on external information that usually cannot be found or validated and hence can only be arbitrary, that different choices can also produce different estimates of P and C effects, and that all just-identified models will produce identical model fit statistics. The latter makes it impossible to discriminate among different constrained models.

There are other approaches that focus on the linear dependency problem in the data such as nonlinear parametric

transformation of one or more of the A, P, and C variables (Mason and Feinberg, 1985), the proxy variables approach that substitute one of these variables with measured characteristics (O'Brien et al., 1999; Winship and Harding, 2008), and those used frequently in biostatistics such as estimable functions of nonlinear components of trends and the minimization of penalty function for the derivation of linear constraint (Robertson et al., 1999; Holford, 2005). A detailed review of these methods is included in Yang and Land (2013). The most important limitation shared by all the conventional approaches discussed above is that they are confined to the linear models framework that necessarily incurs the identification problem.

Theoretical Problems

The lack of a proper understanding of the origin of the model identification problem also contributed to conceptual limitations. While the problem has always been conceived as a methodological one, we believe it is also a theoretical one. The APC accounting models induce this problem by imposing the key assumption of additivity and treating the three variables as fixed effects that are independent of each other. That is, the outcome is modeled as a function of $A + P + C$. This is a simple approximation to the actual processes underlying social change and may not accurately describe period- and cohort-related phenomena. A new framework is needed to accommodate nonadditive and random processes in addition to additive and fixed effects; to better characterize the cohort and period effects; to stimulate new conceptualizations of processes of change at multiple levels (individual, historical time, and their intersection); and to address new theoretical questions about change.

New Approaches to Old Problems

In recent decades, new statistical models and computationally intensive estimation methods have been developed. Also, data sets with research designs that invite or even require accurate period and cohort analysis have become more available. Together, new possibilities have emerged for sophisticated analyses of period- and cohort-based demographic processes. In a recent book, Yang and Land (2013) presented guidelines for conducting APC analysis that provide new avenues to approaching the old problems outlined above. Using three research designs, they showed the application of a generalized linear mixed models framework to the analyses of time related dynamics and population heterogeneity therein.

Three Useful Research Designs

While the use of cross-section design leads to the confounding of age and cohort effects, and the use of single-cohort panel design leads to the confounding of age and period effects, three other research designs are particularly useful for period and cohort analysis.

The repeated cross-sectional data design subsumes two subcategories that are similar in form but different in unit of analysis. The first is the age-by-time period tabular array of population-level rates or proportions, the classical data structure fit to the linear models in eqn [1]. As depicted in **Table 1** and **Figure 1**, age-specific observations of aggregate rates can be summarized in an A by P table for age groups and time periods that are of equal interval lengths. The diagonal elements of the matrix correspond to observations of birth cohorts. Only a longitudinal panel study design provides data on true birth cohorts, which follow identical individuals over time. That said, repeated cross-sectional designs are far more common and rely on the synthetic-cohort approach that repeatedly samples the same birth cohort over a long period of time. Differential temporal groupings for the A, P, and C variables are also possible using this design. When the data used to construct the A by P array are in such a form that different interval widths are used for the variables, the same design can be considered alternatively as a multilevel data design. Some prototypical data sets that used this design include death rates in the Human Mortality Database, and cancer incidence and mortality rates in Surveillance, Epidemiology, and End Results.

The second design is a repeated cross-sectional sample survey that also contains microdata on individual respondents. Pooling data by birth cohorts and survey years, one can obtain the cohort-by-period cross-classification data structure, in which multiple individual observations are nested within the cross-classifications of two higher level units – birth cohort and time period. Although this design shares characteristics with tabular data at the aggregate population level, it also possesses an additional and important feature: it provides multilevel data on both the individual-level responses and a wide range of covariates. These features enable much finer-grained analyses in these data than those structured entirely from aggregated cross sections. Some prototypical data sets that used this design include the General Social Survey, the National Health and Interview Survey, and the National Health and Nutritional Examination Survey.

The third design is the prospective cohort panel and the accelerated longitudinal design that follows multiple cohorts forward over multiple points in time. Such a structure provides cross-time linkages within individuals in true cohorts. This design has been an important advance in aging and life course research because it allows a more rapid accumulation of aging information from multiple cohorts. It is thus especially useful for testing aging- and cohort-related hypotheses because it allows the distinction between intraindividual change with age within cohorts and intercohort differences in age trajectories. Some prototypical data sets using this design include the Americans' Changing Lives Study and Health and Retirement Study.

The Intrinsic Estimator for Linear Regression Models

A number of recent methodological studies have introduced an innovative approach to the model identification and estimation problem concerning period and cohort analysis within the linear/generalized linear models framework. Incorporating recent methodological developments on estimable functions in the context of the APC accounting model, the Intrinsic Estimator (IE) was proposed as a new method of estimation for distinct age, period, and cohort effects. The uniqueness of the IE

is its property of estimating the unique estimable function of all components of the model (1) determined by the Moore–Penrose generalized inverse. It achieves model identification with minimal assumptions about the individual effects included in the model. Further, it can be estimated using principal component regression. The IE is another constrained estimator, but the special constraint used by the IE is to remove the influence of the design matrix (which is fixed by the number of age and period groups and not related to the outcome y_{ijk}) on coefficient estimates, thus allowing the intrinsic trends of variation in the data to emerge. This constraint produces an estimator that has desirable statistical properties, including estimability, unbiasedness, relative efficiency, and asymptotic consistency (Yang et al., 2004). The IE method can be implemented in multiple statistical packages such as R and Stata. Interested readers should refer to the book website for Yang and Land (2013) for sample codes.

To date, the IE has been empirically applied in a number of studies in epidemiology, demography, and social science, and has yielded sensible results on period and cohort trends in a variety of topics, such as cause-specific mortality rates (Yang, 2008a; Masters et al., in press; Powers, 2013), rates of religious activities and beliefs (Schwadel, 2011), and mortality due to accidental poisoning (Miech et al., 2011). Additional methodological studies of its properties have also continued. For example, Powers (2013) couched the IE in a longer statistical tradition of estimable functions, showing "a number of equivalent numerical methods (i.e., principal components, singular value decomposition, generalized inverse, etc.) may be employed in various ways to provide solutions." And Powers's (2012) work provides a flexible extension (ie_rate) of Schulhofer-Wohl and Yang's (2006) APC IE Stata program (apc_ie). In addition, Yang and Land (2013) have further illustrated the robustness of the statistical properties of the IE via Monte Carlo simulation and empirical model validation analyses. They demonstrated numerically the finite-time-period and asymptotic properties of the IE. They also showed that the IE results are completely consistent with coefficient estimates of age, period, and cohort effects from an entirely different family of models (hierarchical APC models, which are discussed below) that do not use any identifying constraints. Similarly, Masters et al. in press found estimates of temporal trends in official U.S. adult mortality rates from models using the IE to be entirely consistent with coefficient estimates from Hierarchical Bayesian Models using Gibbs sampling (Yang, 2006).

Hierarchical Age-Period-Cohort Models

We noted before that there are theoretical limitations of the conventional linear models for analysis of period and cohort effects. And these limitations call for a better modeling framework that can generalize various processes underlying these effects. The mixed effects models are one family of models that satisfy this criterion, which we term as the 'hierarchical APC' or 'Hierarchical Age-Period-Cohort' (HAPC) models. We emphasize two breakthroughs of HAPC models compared to the linear fixed effects models. First, because individual-level observations are nested in higher-level social contexts defined by time period as well as birth cohort, we can contextualize individual lives within periods and cohorts by employing hierarchical linear models that both address multilevel heterogeneity and avoid the model identification problem. Second, the HAPC models allow for the testing of mechanism-based explanatory hypotheses through the inclusion of additional covariates at both the individual and period- or cohort-levels. Hence, this new family of models can also effectively address the criticism lodged against the general purpose APC analysis: namely, that it is not useful because it does not reveal any underlying mechanisms driving the observed period-based and cohort-based trends.

Recalling the two conditions inducing the identification problem, (1) the use of multilevel data resolves the first condition such that no exact linear dependency exists between A, P, and C variables, and (2) the new HAPC models can be specified so that the three variables are not linearly related to the outcome. The combination of both approaches is the most effective way of accounting for change. In the APC accounting model, the assumption of fixed period and cohort effects ignores the level-two heterogeneity. In the HAPC models, one can model the level-two heterogeneity directly to assess the possibility that individuals within the same periods and cohorts might share unobserved random variance. Specifically, one specifies the cross-classified random effects model (CCREM) as the following:

Level-1 Within-cell model:

$$f(y_{ijk}) = \beta_{0jk} + \sum_{p=1}^{P} X_p \beta_p \qquad [2]$$

Level-2 Between-cell model:

$$\beta_{0jk} = \gamma_0 + u_{0j} + v_{0k} \qquad [3]$$

In the level-1 within-cell model, β_{0jk} is the intercept or 'cell mean,' that is, the mean outcome value of individuals who belong to birth cohort j and surveyed in year k; the outcome variable of respondent i within each birth cohort j and survey year k is assumed as a function of a vector of individual-level covariates, X_p, for $p = 1, \ldots, P$, where P indicates the maximum number of covariates and $\beta_1, \ldots, \beta_p$ are the level-1 fixed effects. In the level-2 between-cell model, the intercept β_{0jk} is allowed to vary randomly from cohort to cohort and from period to period. γ_0 is the model intercept, or grand-mean; u_{0j} is the cohort effect, or residual random effect of cohort j, that is, the contribution of cohort j averaged over all periods, on β_{0jk}, and v_{0k} is the period effect, or residual random effect of period k, that is, the contribution of period k averaged over all cohorts.

A two-step approach can be used to assess the statistical significance of random period and cohort effects. In step , one studies the patterns and statistical significance of the individual estimated coefficients for birth cohorts (u_{0j}) and time periods (v_{0k}). The random coefficients and their standard errors can be obtained by the restricted maximum likelihood-empirical Bayes estimation (REML-EB) available in most statistical packages such as SAS and Stata that also

reports their ratios as asymptotic *t*-ratios for the statistical test. In step 2, one further assesses whether the birth cohort and/or time period effects, taken as a whole, are also statistically significant. This question is one of whether these effects together contribute to explained variance in the model. A first approach to this question is to examine the model deviance statistics, variance components, and associated *z*-scores using the REML-EB method. When the numbers of level-two units (cohorts and periods) are small, however, the large sample properties of random effect estimators may not be appropriate and the REML estimates of the variance components are only proximate. In this case, one can use an *F* test to test the hypothesis of the presence of random effects. The specific procedures are available in Yang and Land (2013: Chapter 7).

The basic model shown above in eqns [2] and [3] is a *random-intercept model* that allows only the level-1 intercept to have random variations. It can be extended to become a *random-coefficient model* wherein level-1 slope coefficients also have random variations across cohorts and periods. The random-intercept model is useful for assessing the period and cohort trends in mean outcomes. And the random-coefficient model is more general and can be useful for assessing period and cohort trends in not only the mean but also slopes of level-1 fixed effects that may define social inequalities (e.g., sex, race, education) in the outcomes. For example, a recent study of social disparities in happiness in the United States used the random-coefficient CCREM to show that racial gaps in happiness changed significantly over time, but were constant across birth cohorts (Yang, 2008b). One can further add level-two covariates to explain the cohort or/and period variances. An example is that cohort patterns of newspaper reading and TV watching were found to explain part of the recent cohort-based declines in U.S. verbal ability (Yang, 2006). In all, the HAPC models provide a much more flexible framework than the linear models, move far beyond the identification problem, and can accommodate a wide variety of specifications to address complex questions related to demographic phenomena.

New Areas of Period and Cohort Analysis

Period and cohort analysis, together with the model identification problem, have intrigued and frustrated demographers for decades. Developments of new estimation methods – and the extensions of more general models summarized above – have brought renewed interest to such analysis (Smith, 2008). With increasing applications of new models and methods, the scope of period and cohort analysis is quickly expanding. It is beyond the scope of this article to offer an exhaustive review of all newly emerged directions of research. We briefly provide some examples that highlight new areas of methodological and substantive research relevant to demography.

APC Model-Based Demographic Forecasting

When period and cohort factors are both implicated in demographic patterns and trends, the forecasting of future rates of change needs to account for both effects. Indeed, omitting either period-based or cohort-based variation from trend models will likely lead to inaccurate forecasts. By design, an APC framework for analyzing trends in demographic outcomes can be easily extended to forecasting future population trends. For example, a simple application of the APC model-based forecasting analysis shows that it was able to predict the slowdown of coronary heart disease mortality decline among men ages 45–49 and increases among men ages 25–29 after 2000. Such trends can likely be attributed to changing health conditions in more recent cohorts, such as rising obesity prevalence and increasing duration of time spent obese across early life (Reither et al., 2011). Previous work relying on traditional forecasting methods was inattentive to these changes, and inaccurately forecasted coronary U.S. heart disease mortality among these age groups.

Yang and Land (2013) further compared the performance of traditional forecasting methods with the APC model-based projections in predicting future cancer mortality rates. The conventional models tested were (1) simple linear extrapolation and (2) time series methods such as the autoregressive integrated moving average (ARIMA). The difference between these two kinds of forecasting is that the simple method produced future rates by (1) fitting linear or (2) ARIMA models to mortality rates, whereas the APC method forecasted future rates by fitting the same models to APC model coefficients. The results of internal validation analysis showed unequivocally that the APC approach was able to depict actual mortality rates much better by their accounting for cohort changes in trends. Such cohort-based changes were missed by both traditional methods, leading to less accurate predictions.

Thus, the APC model based demographic forecasting methods have shown substantial advantages over the simple data extrapolation methods in accuracy of point forecasts and assessments of uncertainty in these forecasts through prediction intervals. Therefore, they hold much promise for demographic studies of cohort processes and their implications for future trends of population change.

Temporal Dynamics in Population Heterogeneity

Period and cohort analysis have traditionally been utilized to document temporal trends and dynamics in population-wide processes. Recent studies have paid increasing attention to systematic temporal variations in health behaviors and health inequalities across subgroups. We briefly mention three categories of research that focuses on heterogeneity in terms of either variations by specific sociodemographic groups or overall dispersion.

Origins and Trends in Social Inequalities in Health

Health inequalities have long been the subject of demographic analysis. Changes in health disparities have important scientific, policy, and social justice implications, and demographers have recently employed the APC framework to gain insight into a number of relevant questions. One line of research, for example, applied HAPC-CCREM analyses to the repeated cross-sectional design of the National Health Interview Survey and individual-level records of mortality status to investigate

temporal trends in race/ethnic and socioeconomic differences in U.S. adult mortality rates (Masters et al., 2012). Findings revealed important race and socioeconomic disparities in recent U.S. adult mortality trends that were largely related to cohort-based changes in the population.

Another line of research employed HAPC-growth curve models on a prospective cohort design from longitudinal surveys. The models systematically examined both inter- and intracohort variation in aging-related health trajectories, and implicated both processes in the formation of social inequalities in health over the life course. The model specification is similar to that of HAPC-CCREM, except that within-individual repeated observations are now the level-1 units and the individuals are the level-2 units. Cohort membership can then be modeled as the individual-level attribute conditioning within-individual change with age. In this regard, the growth curve models have been used to test aging- versus cohort-related hypotheses. For instance, research shows strong support for the intercohort change hypothesis that posits aging trajectories of health significantly vary by birth cohort. For instance, more recent cohorts showed higher mean levels of physical functioning and self-assessments of health (Yang and Lee, 2009) and experienced slower increments in frailty with age (Yang and Lee, 2010). Conversely, there is no strong support for the intracohort heterogeneity hypothesis. That is, several lines of research have repeatedly shown health inequalities to persist across the life course after conditioning on cohort membership. Such findings suggest little within-cohort variation in health-related aging processes. However, birth cohorts differ substantially in within-cohort inequalities in aging trajectories. For example, it was found that racial gaps in frailty converged with age only in the oldest cohort born before 1924 but remained constant in other cohorts, and socioeconomic status (SES) disparities in frailty either diverged or converged depending on indicators used and cohorts under study (Yang and Lee, 2010).

Together with other recent studies on aging and social inequalities in health, the newly emerged research above has illustrated the relevance of cohorts' disparate exposures to biological and sociohistorical contexts to our understanding contemporary inequalities in health and longevity. We, therefore, strongly encourage researchers concerned with health inequalities to incorporate cohort considerations into their models and/or formally adopt cohort analysis to better understand current patterns of and temporal trends in social inequalities in a wide-range of health outcomes.

Overall Dispersion: Between-Group and Within-Group Variance

We know that varying gender, race, and SES inequalities across the life course, birth cohorts, and time periods conceivably contribute to overall inequality across these temporal dimensions. An extension to the HAPC models is to combine them with the variance function regression (VFR), which can facilitate the study of both between-group inequality (i.e., variations in the conditional mean of an outcome) and within-group inequality (i.e., variations in the conditional variance or dispersion of an outcome). As exposited in Zheng et al. (2011), the core idea of the HAPC-VFR model is to estimate mixed effects regressions for the two equations in the VFR model, whereby the first mixed effects regression models how age, period, and cohort affected mean estimates of an outcome net of individual-level covariates (sex, race, education, etc.). The second regression models how within-group disparities changed across the three temporal dimensions. Applying the HAPC-VFR to measures of self-rated health in the National Health Interview Survey data, Zheng and colleagues found that changes in health disparities in the last two decades have predominantly reflected cohort processes over period processes. The findings are consistent with the studies summarized above on other health outcomes, lending additional support for the cohort perspective on health disparities. And compared to the studies we mentioned above that focused on changes in health inequality defined by specific aspects of social stratification system such as sex, race, and SES, the HAPC-VFR model offers an analytic tool to encapsulate the general dispersion of health across age, period, and cohort. This additional information provides the basis for further research to examine the contribution to the overall dispersion by each specific subpopulation.

Inequality or disparity in statuses occurs in many domains of population processes and social life. The HAPC-VFR model provides a powerful framework and lens through which to identify and study the temporal evolution of variations in any given outcome. Therefore, the model should be broadly applicable to the study of population heterogeneity in many different substantive contexts.

Cohort Variation in Associational Studies

We finish by briefly reiterating the importance of considering cohorts' disparate exposures to biological and sociohistorical contexts when investigating demographic outcomes, even in cases wherein the outcome of interest does exhibit significant cohort-based variation. That is, cohort analysis is not only useful to demonstrate trends in certain outcomes themselves, but also useful as an explanatory factor for the associations between other variables. To highlight an historical case, the Jones Criteria in 1942 was a period-based introduction to health knowledge and technology, wherein Jones advanced an approach to identifying strep throat and administering the use of Penicillin to prevent the subsequent onset of rheumatic fever. This is a classic example of a period-based intervention of health technology to immediately improve population health. Two points, however, are especially worth noting that made the Jones Criteria an important cohort-based intervention as well. First, the age group most susceptible to the emergence of rheumatic fever is between 5 and 15 years of age. Second, rheumatic heart disease later in life has been shown to be directly influenced by rheumatic fever early in life. Thus, cohort-specific rates of rheumatic heart disease overwhelmingly reflected cases of rheumatic fever among these same cohorts early in their lives. As a result of the Jones Criteria, the immediate reduction of rheumatic fever across the 1940s and 1950s directly led to subsequent reductions in case of rheumatic heart disease several decades later. Moreover, it was largely in those cohorts that were born after the implementation of the Jones Criteria that fully benefited from. Those cohorts that

had already aged out of the most critical age range (5 to 15) had already been 'scarred' by their experiences with rheumatic fever, and thus didn't reap the full, long-term benefits of the period-based intervention.

A contemporary example of cohort processes indirectly shaping mortality trends is Lynch's (2003, 2006) and Masters et al.'s (2012) work showing recent educational differences in U.S. self-rated health (Lynch, 2003, 2006) and mortality risk (Masters et al., 2012) to be significantly influenced by, first, cohort changes in early life survival and, second, cohort changes in the qualitative meaning of education. That is, changes in early life survival across the twentieth century have made the adult environment relatively more important in shaping U.S. longevity. This, in turn, coupled with the educational revolution across the twentieth century has made socioeconomic resources more important in navigating the adult environment. Together, both processes help to fuel the observed increases in education-based health disparities in the United States.

In conclusion, research to date utilizing period and cohort analysis has gone well beyond the model identification problem, which was a major obstacle and point of contention in the field in the 1980s. New ways of conceptualizing the research problems, new techniques for analyzing these problems, and the collection and availability of superior data sources, set the stage for a new era of period and cohort analysis. We hope this emerging era will enhance our understanding of the complex interplay of human life course, social and epidemiologic conditions, and historical events and technological advances that shape demographic processes.

See also: Demographic Techniques: LEXIS Diagram; Early Life Influences on Health and Mortality in Adulthood; Life Table; Multistate Transition Models in Demography.

Bibliography

Ben-Shlomo, Y., Kuh, D., 2002. A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. International Journal of Epidemiology 31, 285–293.

Finch, C.E., Crimmins, E.M., 2004. Inflammatory exposure and historical change in human life-spans. Science 305, 1736–1739.

Fogel, R.W., Costa, D.L., 1997. A theory of technophysio evolution, with some implications for forecasting population, health care costs, and pension costs. Demography 34, 49–66.

Frost, W.H., 1939. The age selection of mortality from tuberculosis in successive decades. American Journal of Hygiene 30, 91–96.

Hobcraft, J., Menken, J., Preston, S., 1982. Age, period, and cohort effects in demography: a review. Population Index 48, 4–43.

Holford, T.R., 2005. Age-period-cohort analysis. In: Armitage, P., Colton, T. (Eds.), Encyclopedia of Biostatistics. John Wiley & Sons, Hoboken, pp. 82–99.

Kupper, L.L., Janis, J.M., Karmous, A., Greenberg, B.G., 1985. Statistical age-period-cohort analysis: a review and critique. Journal of Chronic Diseases 38, 811–830.

Lloyd-Jones, D., Adams, R., Carnethon, M., et al., 2009. Heart disease and stroke statistics – 2009 update: a report from the American heart association statistics committee and stroke statistics subcommittee. Circulation 119, 480–486.

Lynch, S.M., 2003. Cohort and life-course patterns in the relationship between education and health: a hierarchical approach. Demography 40 (2), 309–331.

Lynch, S.M., 2006. Explaining life course and cohort variation in the relationship between education and health: the role of income. Journal of Health and Social Behavior 47 (4), 324–338.

Mason, K.O., Mason, W.M., Winsborough, H.H., Poole, W.K., 1973. Some methodological issues in cohort analysis of archival data. American Sociological Review 38, 242–258.

Mason, W.M., Feinberg, S.E. (Eds.), 1985. Cohort Analysis in Social Research: Beyond the Identification Problem. Springer-Verlag, New York.

Masters, R.K., 2012. Uncrossing the U.S. black-white mortality crossover: the role of cohort forces in life course mortality risk. Demography 49 (3), 773–796.

Masters, R.K., Hummer, R.A., Powers, D.A., 2012. Educational differences in U.S. adult mortality risk: a cohort perspective. American Sociological Review 77 (4), 548–572.

Masters, R.K., Hummer, R.A., Lin, S., Beck, A.N., Finch, B.K. Long-term trends in adult mortality for U.S. blacks and whites: an examination of period-and cohort-based changes. Demography, In press.

Miech, R., Koester, S., Dorsey-Holliman, B., 2011. Increasing US mortality due to accidental poisoning: the role of the baby boom cohort. Addiction 106, 806–815.

O'Brien, R.M., Stockard, J., Isaacson, L., 1999. The enduring effects of cohort characteristics on age-specific homicide rates, 1960–1995. American Journal of Sociology 104, 1061–1095.

Osmond, C., 1985. Using age, period and cohort models to estimate future mortality rates. International Journal of Epidemiology 14, 124–129.

Powers, D.A., 2012. IE_RATE: Stata module to conduct age, period, and cohort (APC) analysis of tabular rate data using the intrinsic estimator. Statistical Software Components S457445, Boston College Department of Economics (revised 12 February 2014). IDEAS. http://ideas.repec.org/c/boc/bocode/s457445.html.

Powers, D.A., 2013. Black-white differences in maternal age, maternal birth cohort, and period effects on infant mortality in the US (1983–2202). Social Science Research 42 (4), 1033–1045.

Reither, E.N., Olshansky, S.J., Yang, Y., 2011. New forecasting methodology indicates more disease and earlier mortality ahead for today's younger Americans. Health Affairs 30, 1562–1568.

Riley, M.W., 1987. On the significance of age in Sociology. American Sociological Review 52, 1–14.

Robertson, C., Gandini, S., Boyle, P., 1999. Age-period-cohort models: a comparative study of available methodologies. Journal of Clinical Epidemiology 52, 569–583.

Ryder, N.B., 1965. The cohort as a concept in the study of social change. American Sociological Review 30, 843–861.

Schwadel, P., 2011. Age, period, and cohort effects on religious activities and beliefs. Social Science Research 40, 181–192.

Smith, H., 2008. Advances in age-period-cohort analysis. Sociological Methods & Research 36 (3), 287–296.

Winship, C., Harding, D.J., 2008. A mechanism-based approach to the identification of age-period-cohort models. Sociological Methods & Research 36 (3), 362–401.

Yang, Y., Fu, W.J., Land, K.C., 2004. A methodological comparison of age-period-cohort models: the intrinsic estimator and conventional generalized linear models. Sociological Methodology 34, 75–110.

Yang, Y., 2006. Bayesian inference for hierarchical age-period-cohort models of repeated cross-section survey data. Sociological Methodology 36, 39–74.

Yang, Y., 2008a. Trends in U.S. adult chronic disease mortality: age, period, and cohort variations. Demography 45, 387–416.

Yang, Y., 2008b. Social inequalities in happiness in the United States, 1972 to 2004: an age-period-cohort analysis. American Sociological Review 73, 204–226.

Yang, Y., Lee, L.C., 2009. Sex and race disparities in health: cohort variations in life course patterns. Social Forces 87, 2093–2124.

Yang, Y., Lee, L.C., 2010. Dynamics and heterogeneity in the process of human frailty and aging: evidence from the U.S. older adult population. Journal of Gerontology: Social Sciences 65B, 246–255.

Yang, Y., Land, K.C., 2013. Age-Period-Cohort Analysis: New Models, Methods, and Empirical Applications. Chapmen & Hall/CRC Press, New York.

Zheng, H., Yang, Y., Land, K.C., 2011. Variance function regression in hierarchical age-period-cohort models: applications to the study of self-reported health. American Sociological Review 76, 955–983.

Person and Self: Philosophical Aspects

Alessandro Ferrara, Università di Roma Tor Vergata, Roma, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A brief historical reconstruction of the shifting meanings of the two terms 'person' and 'self' is offered. Then four specific challenges are highlighted, to which all conceptions of person or self must somehow respond: (1) What accounts for identity over time? (2) Is self-consciousness an inborn capacity or the product of socialization and recognition? (3) Is the person or self a unified source of agency or a decentered field of forces? (4) Is there continuity or radical discontinuity between persons and animals or machines?

Finally, a contrastive account of major conceptions of authenticity of the self or person is outlined.

Few terms like 'person' and 'self' in the course of history have proven to be more difficult to pin down to a univocal meaning. Furthermore, both are among the terms of most common usage, yet their exact meaning can prove quite elusive. More than offering an exhaustive definition and an exhaustive reconstruction of the history of the notions of 'person' and 'self,' it is sensible to attempt here to offer an account of the scope of variation in their usage as well as an account of the different problems that different authors and traditions have associated with them.

Person and Self Contrasted

To complicate things further, there is no settled account of the relation *between* 'person' and 'self.' The term 'person' has for a long time been the prevailing one for referring to a single human being. The onset of modernity marks the beginning of its relative obsolescence and its gradual replacement by the term 'self,' and sometimes 'individual,' at the center of the debates within the social sciences and social and political philosophy. Over the past few centuries, the term 'person' has almost disappeared from the vocabulary of the social sciences, only to undergo a sort of revival, sometimes tinged with religious overtones, in the existentialist and so-called personalist philosophical currents of the twentieth century, both religiously oriented and secular. Modern social theory has rather operated with the notion of the individual since the time of Hobbes and Locke; and then later, starting with Cooley's and Mead's symbolic interactionist perspective, it has widely adopted the concept of the self. Today a further evolution is perhaps underway: the concept of the person is regaining center stage. From Renouvier's (1903) contribution to Maritain's and Mounier's diverse versions of personalism at mid-twentieth century, the notion of the person has become pivotal in a number of more specialized fields of inquiry. In the political philosophy of the late twentieth century, for example, in *Political Liberalism* John Rawls had based his view of liberalism on a political conception of the *person*, not of the self or of the individual, in order to signal his taking distance from the traditional 'atomistic' orientation of liberalism. In order to grasp these vicissitudes of the two terms, one has to look at some of the main turning points in their evolution.

Etymologically, the term 'person' originated at the crossroads of the ancient Greek and Roman civilizations. In ancient Greek, *prós-opon* (literally 'placed in front') designated the artificial face, the mask that actors used to wear on stage when performing tragedies. The Latin verb *personare* also alluded to the realm of the performing arts: it meant to 'sound through,' to let the voice reach out to the watching and listening public through the mask. Metaphorically, then, in its origins the term person carries a trace of the willful appearance for the sake of others. But the Roman civilization also grafted a *legal* connotation on the term 'person': the legal person was another sort of mask, the individual human being in her capacity as bearer or rights, privileges, prerogatives, but also duties and obligations. In that sense, to be a self without being a legal person would be tantamount to being nothing. Typically, in these public minded cultures the sphere of the private was designated by terms such as *idiotes* and *privatus*, both of which carried a diminishing connotation – they presupposed that something was somehow missing.

Sometime during the Hellenistic period and at the apex of the Imperial Age, a shift of meaning began to occur. The increase in complexity and the somewhat more anomic quality taken by social life in the Hellenistic monarchies and in Imperial Rome, the increase in complexity of the political apparatus alongside the more despotic and arbitrary styles of government, and the increasing religious eclecticism of the Roman public world and the appearance of the Christian ethos of holiness and personal sanctity all contributed to the rise of a cleft between *public* and *private* life and to a new distribution of value across that divide. Stoic philosophers like Seneca, Epictetus, and Marcus Aurelius on the one hand, and the early Christians on the other, were at one in regarding with suspicion what now appeared as a realm of merely public, artificial, contrived personal masks or roles. Inwardness and depth belonged to the *self*, whereas the person stood now for convention, tradition, custom, and superficial appearance. This new relation between the two terms continued to dominate throughout the Middle Ages (Augustine and Erasmus), the Renaissance, and the Reformation (Luther).

Another major shift occurred with the rise of modern political philosophy, the rise of the modern notion of rights, and the modern conception of citizenship. Whether the focus was on the capacity for autonomous reasoning or autonomous

agency, whether the human being was regarded primarily as the possessor of rationality or the bearer of inalienable rights, throughout early modernity and the Enlightenment interest was primarily directed to the human self as such – the rights were typically predicated as universal rights of 'man' as such – and the notion of the person began to look inextricably entangled with the particularism of estates and the privileges of the order of ranks typical of the absolutist monarchies. As Hannah Arendt has pointed out, the French revolutionaries had no place for the notion of the person, or thought much of the legal person as a category protected by the rule of law.

A new reversal of the connotations of these two terms takes place in the twentieth century, when the term 'person' becomes synonymous with a *nonatomistic* view of the human self. The thick web of relations of recognition that supposedly constitute the person is now understood within as diverse philosophical traditions as existentialism, Mounier's personalism, and Habermas's intersubjective view of subjectivity, as a powerful alternative to the modern atomistic self. This motif is at work in John Rawls's above-mentioned 'political conception of the person' as endowed with two moral powers, namely, the capacity to have a sense of justice and the capacity to form a conception of the good.

The Challenges of Conceiving Person and Self

Aside from the issue of their reciprocal relation and their specific difference, 'person' and 'self' refer to a single human being considered under the aspect of his or her distinctively human capacities. Here we immediately reach a point where the overlap between different conceptions ends and a host of problems arises to which different solutions are offered, which distribute without a clear pattern across the usage of the two terms.

Identity over Time

The first issue on which different conceptions of the self and the person diverge concerns identity over time. What allows us to consider a person or self that obviously undergoes continuous changes over time still 'the same person'? In what sense as 50 year olds are we still the same person as the 20 year old of decades ago? We can contrast four different answers to this question. The first answer, typical of ancient and medieval philosophy, links identity over time with the sameness of an (differently conceived) essence – a personal essence underlying also the body – which manifests itself through the changes undergone by the person or self (Plato, Aristotle, Augustine, and Aquinas). The second position connects identity with the psychological continuity of its possessor: it is *memory* and not the physical continuity of the body that allows us to consider the person of 50 years the same as the 20 year old she once was (Locke). The third answer is Hume's and Parfit's answer: psychological continuity is largely an illusory attribution of an identity, on the part of a mental agency, to episodes that only in a vague sense are similar and related. The self or the person, so conceived, is indeed a gallery of selves and persons whose relatedness is quite tenuous and precariously maintained in the way that habits and customs are maintained. By the end of the nineteenth century, Nietzsche will give another twist to the same line of reasoning: we seek to put a construction of unity upon the diversity of our self-experiences for the sake of feeling secure and reassured. Finally, Kant's answer transforms identity into a transcendental feature inherent in all human beings.

Self-Consciousness: Inborn Capacity or Product of Socialization and Recognition?

Another bone of contention has been the nature and scope of that capacity for self-reflectiveness and self-consciousness that traditionally all conceptions of the self and of the person have presupposed. The various conceptions have polarized around two major options. One view, often called 'atomism' by its critics, takes this capacity for self-awareness as a sort of *natural* endowment of the human being. Whether because it belongs to her essence, because it is given to her as a potential of the soul, or because it emerges as a result of neurobiological processes, self-awareness and self-consciousness constitute inborn capacities that need the world only as a sounding board on which the person or self tries self-initiated action-projects and, in so doing, becomes aware of herself as a center of agency. This view comes in many versions. Augustine, Descartes, and Husserl, under different premises, all favor introspection as a privileged path to inner reality. Others, like Hume and Nietzsche, are much more skeptical about it, but all these authors share a view of the self as basically self-sufficient. Within political and social philosophy, Hobbes and Locke are the most illustrious representatives of an atomistic understanding of the self.

The other view starts from the assumption that the capacity for self-awareness is formed in the context of the person's or self's relation to the social world. Hegel and Mead from very different premises do converge on the idea that self-sufficiency vis-à-vis the world is largely an illusion, but in fact the self-awareness distinctive of personhood and selfhood emerges in the process of interacting with other persons and selves as a product of either seeking and obtaining recognition (Hegel) or learning to see oneself through the eyes of the other (Mead). From Mead's perspective, thinking becomes a special case of the activity of carrying out a conversation mediated by symbols with others: indeed, thought is best conceived as a conversation with oneself. From a quite different angle, in his thoughts on the impossibility of a private language and on the notion of following a rule, Wittgenstein also emphasizes the dependency of individuality and subjectivity on shared practices.

This divide is reproduced in almost all social sciences. Within the sociological tradition, for example, theorists such as Spencer, Pareto, Homans, and Blau understand the self of the social actor as coming to the scene of social interaction with a preformed rational capacity for reflection, whereas not just Mead and Cooley but also theorists such as Durkheim, Schutz, and Parsons understand that capacity for reflection as the inner sedimentation of processes of interaction. Within the psychoanalytic tradition, to take another example, theorists like Freud and Klein understand the capacity for ego strength and rational self-awareness as the product of internal vicissitudes of the drives, while object relations theorists such as Winnicott or self-psychologists like Kohut understand it as the product of the self's embedment in a relation with an empathic caretaker.

A Unified Source of Agency, or a Decentered Field of Forces?

These considerations lead to yet another major issue with respect to which theories of the person and of the self have taken sides. Some of these theories tend to portray the person as a more or less unified source of autonomous initiatives. Certainly what we call a person or a self contains several distinct aspects – cognitions, beliefs, desires, propensities, dispositions, capacities, virtues, and so on – but the views of authors like Plato, Aristotle, Augustine, Aquinas, Descartes, Hobbes, Locke, Kant, Hegel, and Sartre, whom we find disagreeing on other aspects, converge in depicting these aspects as being somehow hierarchically subjected to a central agency of sorts: the soul, reason, the will, and self-consciousness are all different ways in which this central agency has been identified. On the other hand, authors again as diverse as Hume, Diderot, Schopenhauer, Nietzsche, Freud, Bataille, Foucault, Goffman, Rorty, Barthes, Deleuze, and Derrida all share the suspicion that this impression of centeredness is largely illusory and the product of 'impersonal' or 'intrapersonal' forces that create it and make it a compelling appearance for the person.

Continuity and Radical Discontinuity between Persons and Animals or Machines

But what is it for a being to be a person or self, and what kind of entity can be called a person or a self? Here the opposition is between conceptions that radically sever personhood or selfhood from nonhuman forms of consciousness and conceptions that take a more nuanced view of the matter, and depict the difference in terms of a gradient of capacities. Starting from the latter, typically we find here conceptions of the person or the self that posit the capacity for representing the world, in order to then act on it, as the crucial defining ability. Humans are then singled out from the rest of beings by self-representation or self-consciousness – which belongs to them alone – but that capacity in turn is only a reflective enhancement of a representational capacity that as such can also be found, though to a lesser degree, in certain animal species and in advanced artificial intelligence machines. Over limited territories – for example, the kind of representational and practical competence involved in playing chess – machines can even outdo humans. The conceptions that introduce a more radical discontinuity between humans and nonhumans define the former basically as 'subjects of significance' (Taylor, Heidegger, and Gadamer) – beings for whom things *matter*. From their point of view, what distinguishes persons and selves from other beings capable of representation is not just a greater share of 'strategic power,' 'the capacity to deal with the same matter of concern more effectively' or a superior capacity for planning (Taylor, 1985: p. 104). The central defining feature is responsiveness, openness to certain strong evaluations, principles or values in light of which not only single action plans but even one's own life is conceived and assessed. From this standpoint, the sense of the self is the sense of where one stands in relation to this matrix of meaning. In one of its most famous versions, this view posits that as persons or selves, human beings are beings for whom their own being is a matter of concern (Heidegger).

The Person, the Self, and Fulfillment

Finally, conceptions of the person and of the self do take quite different views of what it means for a person or a self to *flourish* or to attain *authenticity*. Some conceptions understand authenticity as being primarily linked with breaking free from the constrictions of an entrenched social order. For example, some authors affiliated to aesthetic modernism (Baudelaire) or to *Lebensphilosophie* (notably Nietzsche) tend to understand authenticity as something to be attained '*in opposition* to the demands of society' and of culture, whereas in the writings of authors such as Rousseau, Schiller, Herder, Kierkegaard, and others we find a recognition (with very different degrees of awareness and explicitness) of the fact that social expectations, roles, and institutions cannot merely be understood as playing a constraining, 'disciplinary' or repressive role but also somehow constitute the symbolic material out of which fulfillment can be generated. In the psychological realm, too, we find representatives of these two versions of authenticity. Laing's understandings of schizophrenia as a response to the imposition of an inauthentic life and of 'madness' as resistance to the irregimentation of mental life, and Lacan's understanding of the unconscious, are views of the first kind. On the other hand, Kohut's psychology of the self and Winnicott's notions of 'creative living' and of the 'true self' are the most obvious representatives of an integrative understanding of authenticity within the psychoanalytic tradition.

From another point of view, we can contrast coherence-oriented and decentered views of authenticity. Far from considering fulfilled a life course that can be summed up as a coherent narrative, some consider *narratability* as the epitome of failure. In his reflections on the notion of 'limit-experience,' for example, Foucault, continuing a tradition that he associates with the names of Nietzsche and Bataille, equates the paradigmatic kind of fulfilled subjectivity with the one undergoing a 'limit-experience.' From a variety of *lebensphilosophisch*, aesthetic modernist, poststructuralist, therapeutic, and postmodernist vocabularies, the advocates of a *decentered* understanding of authenticity oppose all attempts to restore an internal hierarchy between what is central and what is peripheral to a life project, a personality, or an identity. By contrast, all authors who propound a *centered* notion of authenticity do wish (1) to maintain some kind of orderly stratification of the layers of an identity, (2) to continue to speak of a core and a periphery, and (3) to make sense of the plurality of experiences, detours, and side narratives to a life history as variations on a unique theme.

Finally, we can contrast *consequentialist* and *exemplary* views of fulfillment. According to the former (what Ronald Dworkin has called the 'model of impact'), a person's life is all the more fulfilling the greater its impact on the external world, or the larger the difference it has made. Hegel and Marx certainly share such a view. For both of them, the full flourishing of a person's life is connected with the person's objective contribution to advancing a differently conceived historical process in which the person is immersed. According to the exemplary view (called by Dworkin the 'model of challenge'), the measure of fulfillment is provided by success in meeting a challenge that carries a unique significance for the person, independently of the impact of such success on the external world. Rousseau and

Kierkegaard offer representative versions of such a view. For both of them, the full flourishing of a person's life is connected with the person's capacity to reconcile 'being and appearance' (Rousseau) or to 'choose her personality' (Kierkegaard), in other words to bring outer conduct and deeper inner motives into alignment.

Concluding Remark

To conclude this review of the ways in which the person and the self have been conceived, there could not be a sillier question to raise then 'Which of these views comes closest to grasping the reality of the human self or the person?' For these views are in a way not to be understood as representations of an independent reality in the empiricist sense of the word – a realm of facts about the constitution of the human being that we may grasp or fail to grasp. Rather, they are to be seen as competing vocabularies, as alternative types of lenses – indeed, as sets of premises – which we use in order to make sense of something as crucial for our experience of the world as the idea of personhood. Evidence for and against each of them can be and certainly is usually considered, but is unlikely to be ultimately decisive. Ultimately decisive has generally been our sense of the contribution that adopting one or the other can give to shaping an overall view of the world in which we find reflected what most matters to us.

See also: Autonomy, Philosophy of; Identity and Identification; Personal Identity: Philosophical Aspects; Self: Philosophical Aspects.

Bibliography

Carrithers, M., Collins, S., Lukes, S. (Eds.), 1985. Anthropology, Philosophy, History. Cambridge University Press, Cambridge, MA.

Heller, Th. C., Sosna, M., Wellbery, D.E. (Eds.), 1986. Reconstructing Individualism. Autonomy, Individuality, and the Self in Western Thought. Stanford University, Stanford, CA.

Maritain, Jacques, 1947. The Person and the Common Good. Translated by John J. Fitzgerald. University of Notre Dame Press, Notre Dame, IN, 1985.

Mounier, Emmanuel, 1950. Personalism. Translated by Philip Mairet. University of Notre Dame Press, Notre Dame, IN, 1952.

Rawls, John, 1993. Political Liberalism. Columbia University Press, New York.

Renouvier Charles, B., 1903. Le personnalisme. F. Alcan, Paris.

Ricoeur, P., 1992. Oneself As Another. Translated by Kathleen Blamey. University of Chicago Press, Chicago and London.

Taylor, Ch, 1985. The concept of a person. Philosohical Papers. In: Human Agency and Language, vol. 1. Cambridge University Press, Cambridge, MA, pp. 97–114.

Person Perception, Accuracy of

David C Funder, University of California, Riverside, Riverside, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The accuracy of person perception concerns the connection between judgments people make of the psychological attributes of others, usually personality traits, and the actual status of those attributes in the people who are judged. Such judgments are correct roughly two times out of three and the degree of accuracy likely to be attained depends upon four moderator variables: properties of the judge, the target, the trait that is judged, and the information upon which the judgment is based.

The *accuracy* of person perception (also called *personality judgment*) refers to the connection between judgments people make of the psychological attributes of others and the actual status of those attributes in the persons who are judged. The psychological attributes may be another person's personality traits, or his or her thoughts and feelings at a given moment. The ability to judge personality has concerned psychological researchers, off and on, for more than a half century (see Allport, 1937; Funder, 1999; Kenny, 1994). The ability to guess what someone is thinking and feeling, called *empathic accuracy*, is a newer topic that is receiving increasing attention (Rollings et al., 2011). Research on either topic entails special difficulties because the notion of 'accuracy' raises thorny questions concerning how accuracy is to be conceptualized and evaluated. Despite these difficulties, research since 1980 can report important empirical and theoretical progress.

The Importance of Accuracy

The accuracy of person perception has obvious practical importance. Employers and other people in everyday life rely heavily on their perceptions of other persons' attributes such conscientiousness, intelligence, friendliness, dominance, and many others. It is an everyday occurrence to make decisions concerning other people as a function of beliefs about what those people are like. Moreover, clinical assessment typically relies heavily on the impressions of a psychiatrist or psychologist concerning a person of interest (Block, 2008). Interpersonal judgments and sociometric ratings are also often used as data for psychological research. The accuracy of these perceptions is critical for the quality of the data and the research that is based on them.

The accuracy of personality judgment is also important for theoretical reasons. Accurate person perception implies knowledge about the connections between observable behaviors, on the one hand, and personality traits, on the other hand. Thus, to understand the accuracy of person perception would be to understand the ways in which personality affects behavior. Person perception is a traditional topic of social psychology (*see* Social Psychology) and the connections between personality and behavior are traditionally studied by personality psychology (*see* Personality Psychology). The topic of accuracy in person perception transcends the boundaries between these traditionally separate subfields of research.

Finally, the oldest topic in philosophy, and one of humankind's most ancient concerns, is the relationship between perception and reality. Research on the accuracy of person perception is driven by curiosity concerning just this issue.

Difficulties in Accuracy Research

In the light of the importance of the accuracy of person perception, it may be surprising to learn that researchers have neglected the topic for long periods of time (e.g., the years 1955–85), and some psychologists with relevant interests (e.g., in person perception) continue to shy away. This avoidance has been caused by several considerations. The most important consideration is the topic's inherent difficulty. At a general level, the topic of accuracy is difficult to conceptualize or to operationalize, because it requires the development of some criterion by which interpersonal judgments can be evaluated as right or wrong. Some psychologists have considered the development of such criteria to be either impossibly difficult or in principle impossible, and therefore turned their attention to other topics. At a more specific level, an important series of articles by Lee Cronbach (e.g., 1955) revealed complex and hitherto-unsuspected technical difficulties inherent in evaluating accuracy in terms of agreement among different judges. Rather than leading to improved methods, this article and related critiques managed to shut down a then-lively field of research, and appeared to intimidate a subsequent generation of researchers from addressing the topic (see Funder, 1999). A further obstacle to accuracy research was the widespread influence of a set of attitudes that seemed to discourage interest in the topic. The 'person–situation debate' in personality psychology (Kenrick and Funder, 1988) led many researchers to believe that the connections between personality and behavior were few and weak. If this were true, it would make little sense to attempt to investigate the accuracy of person perception. At about the same time, the increasing popularity of research on 'errors' in person perception led to a widespread belief that human judgment is characteristically, even fundamentally mistaken. If this were true, investigations of accuracy would seem to hold little hope.

Resolving the Difficulties in Accuracy Research

Research on accuracy was able to move forward only after these difficulties began to be resolved.

First, views concerning the existence and importance of personality traits, and the abilities of human judgment, began to change during the 1980s. The person–situation debate was resolved as reiterating the importance of stable attributes of personality (Kenrick and Funder, 1988). Research on the errors of human judgment was reinterpreted in a broader context in which it was less often seen as implying that human judgment is always or fundamentally wrong (e.g., Swann, 1984). Studies of error identify important shortcuts ('heuristics') and other processes of human cognition. However, they do not imply that judgment is inept, because the processes that produce errors in the laboratory often lead to correct judgments in real life. Moreover, research on error is fundamentally limited because it can illuminate only how judgments fail to achieve perfection, rather than how correct judgments are ever achieved (Krueger and Funder, 2004). The methodological issues identified by Cronbach and others also received a reexamination during the 1980s. This reexamination yielded more sophisticated analytic techniques that allowed the methodological complications to be bypassed or specifically incorporated into analytic designs. David Kenny's 'Social Relations Model,' a data-analytic technique that accounts for effects of judge, target, and their interaction on interjudge agreement, was an important breakthrough in this regard (Kenny, 1994).

The Realistic Conceptualization of Accuracy

Historically, the thorniest issue in accuracy research has been the criterion problem. By what standard can a judgment of personality be evaluated as right or wrong? The answer to this question depends on how accuracy is conceptualized.

Although alternative conceptualizations have been proposed, current research on the accuracy of personality judgment is generally based on a *realistic* approach, consistent with the writings of Egon Brunswik (1956). The approach assumes that psychological attributes of people really exist, though they are only probabilistically knowable through cues of uncertain validity (Funder, 1995). The real nature of a person must then be ascertained, as well as possible but always with less-than-certain validity, through multiple sources of evidence including the person's self-judgment, judgments by others, and his or her relevant behavior. The index of whether a judgment is accurate, in this view, is the degree to which it converges with these multiple criteria.

The Degree of Accuracy

When research on accuracy in person perception revived during the early 1980s, one of its first orders of business was to dispel the impression left by research on error, that human social judgment is always or fundamentally mistaken.

A large number of studies documented the accuracy of personality judgment in two ways. The most common method was to show that different judgments of the same person tend to agree. Research examined self–other agreement and agreement among judges and found that significant agreement was the rule rather than the exception across a wide band of personality traits, even after brief periods of acquaintance. Evidence concerning the ability of personality judgments to predict behavior is more difficult to gather and therefore was slower in coming, but eventually showed that layperson's personality judgments have an impressive ability to predict independent observations of an individual's behavior (Funder, 1999).

The overall degree of accuracy in person perception is difficult to characterize, because it depends upon a number of factors including the ability of the judge, the judgability of the target, the trait that is judged, and the amount and kind of acquaintance between target and judge (see next section). But while it is for these reasons an oversimplification, it is also a fair characterization to state that the usual correlation among criteria for accuracy is in the range 0.30–0.40. This translates to the statement that one person's perception of the personality of another is correct roughly two times out of three.

Moderators of Accuracy

Research since 1980 has found that the degree of accuracy in person perception depends upon four moderator variables.

The Good Judge

The search for the 'good judge of personality' is one of the oldest concerns in the accuracy literature. This search motivated many investigators during the 1930s to 1950s and it was their failure to find consistent results that contributed to the three-decade hiatus that followed. Part of the reason for this failure may be the presence of methodological flaws in many early studies, and another part may be that individual differences are not very large because personality judgment is an essential social skill. The typical person is described in positive terms, and so people who describe others favorably tend to be more accurate – curmudgeons are usually wrong. Perhaps for this reason, good judges of personality are generally described by others as pleasant, happy, and easy to be with (Letzring, 2008).

The Good Target

As Gordon Allport noted years ago, some people are relatively open books whereas others are difficult to understand even after years of acquaintance. According to research by Colvin (1993), the judgable person is one who has an integrated, well-adjusted personality. The consistency among such a person's actions, thoughts, and personality leads them to be coherent overall, and relatively easy to judge. Moreover, most people can accurately identify who, at first acquaintance, they can judge most accurately (Biesanz et al., 2011).

The Good Trait

Traits differ strongly in their judgability. Relatively observable traits, like sociability and talkativeness, are much easier to

judge accurately than more hidden traits like ruminativeness and anxiety. Some evidence also suggests that traits that are extremely desirable or undesirable to possess are more difficult to judge accurately than traits more neutral in evaluative connotation. Vazire's self–other knowledge asymmetry model theorizes that less visible traits should be judged more accurately by the self, whereas highly desirable or undesirable traits should be judged more accurately by others (Vazire, 2010).

Information

The accuracy of a judgment also depends critically upon the information upon which it is based. The information variable has two facets, quantity and quality.

The quantity variable pertains to the *acquaintance effect*, the tendency to judge someone more accurately the longer one has known him or her. The straightforward expectation would be that more information is likely to produce more accurate judgment, an expectation that has been repeatedly confirmed. The quality variable pertains to the kind of context in which one person observes another. Even holding amount of acquaintance constant, different contexts may produce different degrees of accuracy (Letzring et al., 2006).

Theoretical Development

Several theories describe how accurate judgments come about. One theory, Kenny's 'PERSON' model, conceptualizes personality judgments as the result of multiple influences including the judge's response proclivities, his or her knowledge of people in general, the actual personality of the target, and other factors. Another theory is Funder's Realistic Accuracy Model (RAM, Funder, 1995). RAM conceptualizes an accurate personality judgment as the result of a process that begins with the target emitting relevant information (e.g., performing a relevant behavior), which then must become available to a judge, who then detects it and utilizes it correctly in reaching an accurate judgment. Both the PERSON and RAM models can be used to rationalize the moderators of accuracy that have been found to date, and suggest new ones for investigation in future research.

Conclusion

The accuracy of person perception is an important topic that, after periods of neglect, has become a mature field of research and a mainstream concern of personality and social psychology, which is becoming even more important in the twenty-first century.

See also: Decision Making, Psychology of; Heuristics in Social Cognition; Personal Intelligence and Competencies; Personality and Economics; Personality, Cognitive Models of; Personality, Trait Models of.

Bibliography

Allport, G.W., 1937. Personality: A Psychological Interpretation. Holt, New York.
Biesanz, J.C., Human, L.J., Paquin, A., Chan, M., Parisotto, K.L., Sarracino, J., Gillis, R.L., 2011. Do we know when our impressions of others are valid? Evidence for realistic accuracy awareness in first impressions of personality. Social Psychological and Personality Science 2, 452–459.
Block, J., 2008. The Q-sort in Character Appraisal: Encoding Subjective Impressions of Persons Quantitatively. American Psychological Association, Washington, DC.
Brunswik, E., 1956. Perception and the Representative Design of Psychological Experiments. University of California Press, Berkeley, CA.
Colvin, C.R., 1993. 'Judgable' people: personality, behavior, and competing explanations. Journal of Personality and Social Psychology 64, 861–873.
Cronbach, L.J., 1955. Processes affecting scores on 'understanding of others' and 'assumed similarity'. Psychological Bulletin 52, 177–193.
Funder, D.C., 1995. On the accuracy of personality judgment: a realistic approach. Psychological Review 102, 652–670.
Funder, D.C., 1999. Personality Judgment: A Realistic Approach to Person Perception. Academic Press, San Diego, CA.
Kenny, D.A., 1994. Interpersonal Perception: A Social Relations Analysis. Guilford Press, New York.
Kenrick, D.T., Funder, D.C., 1988. Profiting from controversy: lessons from the person-situation debate. American Psychologist 43, 23–34.
Krueger, J.I., Funder, D.C., 2004. Towards a balanced social psychology: causes, consequences and cures for the problem-seeking approach to social behavior and cognition. Behavioral and Brain Sciences 27, 313–327.
Letzring, T.D., 2008. The good judge of personality: characteristics, behaviors, and observer accuracy. Journal of Research in Personality 42, 914–932.
Letzring, T.D., Wells, S., Funder, D.C., 2006. Quantity and quality of available information affects the realistic accuracy of personality judgment. Journal of Personality and Social Psychology 91, 111–123.
Rollings, K.H., Cuperman, R., Ickes, W., 2011. Empathic Accuracy and Inaccuracy. John Wiley & Sons Inc, Hoboken, NJ, US, pp. 143–156.
Swann, W.B., Jr., 1984. Quest for accuracy in person perception: a matter of pragmatics. Psychological Review 91, 457–477.
Vazire, S., 2010. Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. Journal of Personality and Social Psychology 99, 174–190.

Personal Identity: Philosophical Aspects

Gertrud Nunner-Winkler, University of Munich; Max Planck Institute for Human Cognitive and Brain Sciences, Munich, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The two main classes of philosophical theories of numerical identity are described: substance-based theories focus on the soul, body, and brain, and relational theories on psychological continuity and connectedness. Social science analyses deal with the subjective sense of identity. Postmodern theories renounce the unity of the person. Modern theories assume its indispensability, yet differ in the grounding of identity, the mechanisms stipulated for stabilizing identity, and the description of the stances that actors take toward themselves. Empirical data are presented that support the value commitment model of identity stabilization.

Identity has two meanings: it denotes category membership (e.g., X is a man), or it refers to a specific individual (e.g., X is Mr. A). Goffman (1963) refers to the former as social and to the latter as personal identity. Personal identity is synonymous with numerical identity. There are formal properties of the concept of identity, reflexivity, and congruence, which are not relevant to the practical problem of verifying the truth of statements such as 'A and B are the same person.' This requires material conditions of correctness and a clear explanation of the way the terms 'same' and 'person' are used (Garrett, 1998).

Personal identity is highly relevant to practical issues – to moral as well as self-regarding prudential concerns, for example 'Is A the wanted murderer?', 'Does a relation of personal identity obtain between the fetus that B once was and B as an adult (e.g., can B as an adult claim compensation for a medical malpractice she suffered as a fetus?)?' If this is so: 'Is abortion justifiable?', 'Are my brain transplant and I the same person?', 'On noticing the onset of Alzheimer – can I issue an advanced directive to abstain from life-saving measures, thus binding my later totally Alzheimer-stricken self who might desire to keep on living?', 'What about immortality?' (Shoemaker, 2012).

Numerical identity concerns objects as well as persons. Persons are objects – they have material bodies that are located at specific space–time points. But whereas the identity of objects can be judged only from an external or third-person perspective, persons have an internal or first-person perspective as well. From this internal perspective, the identity question attains an additional meaning: 'Do I experience to be the same person today as I was yesterday?' This pertains to the 'sense of identity' which can be based on the same criteria used to determine someone else's identity. But for sociostructural reasons discussed later in this article, identity criteria applied from an external perspective and those from an internal perspective increasingly come to diverge.

In the following, I will first present different approaches to personal identity (Section Personal Identity Theories), compare them, and discuss differing meanings of the core terms 'person' and 'same' (Section Some Interim Considerations). In the second part, I will describe several theories dealing with the sense of identity (Section Sense of Identity Theories), then present some relevant empirical data and discuss their implications for sense of identity theories (Section Sense of Identity – Some Empirical Findings).

Personal Identity Theories

Theories of personal identity can be divided into two main classes with each of them divided into further subcategories. Substance-based theories are immaterialist (soul) or materialist (body and brain). Relational theories refer to psychological continuity or hold the 'identity does not matter' view.

Substance-Based Theories

Immaterialist Theories

Dualist approaches assume that persons are composed of two entities – body and mind – which are only contingently related to each other. Thus, Plato maintained that persons are essentially immaterial souls and therefore will survive death and the destruction of their bodies. Christian creed, too, implies that mortal flesh and eternal soul are separable. Also, beliefs in reincarnation or in an immersion in some transcendental consciousness assume an ego substance that is present at any stage of life and even after one's current life has ended. Today, dualist assumptions are referred to as the Cartesian ego view since it was Descartes who had anchored the only truth beyond doubt in our thinking, thus specifying mental activity as the essence of personhood.

Several objections have been raised against the dualist claim. (1) It provides no answer to the mind–body problem. Descartes' mechanistic conceptualization of the mind as an inner ghost operating a body understood as an automat or clockwork is no longer tenable. Modern brain sciences have shown that neither emotions nor complex thinking processes are separable from the structure and functioning of the biological organism (Singer, 2002; Roth, 2003). (2) The concept of a soul does not unequivocally solve the problem of numerical identity – especially beliefs about reincarnation or immersion disregard the issue of individualized existences. (3) Souls are unobservable. The assumption that they extend over time and have more than a momentary existence rather than being continuously exchanged cannot be verified (Perry, 1978). Finally, (4) the belief in unobservable souls does not meet the practical concerns underlying the debate of identity.

Such objections do not refute dualism. Its metaphysical assumptions withstand empirical verification. This shows that criteria for the nature of persons and evidence for personhood

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63057-X

may diverge. In any case, modern philosophical debates pay little attention to dualistic positions.

Materialist Theories

On the body theory, a person is identical to a living human body. Persons exist as long as their bodies exist, and identity is guaranteed by sameness of body. Sameness allows the gradual replacement of matter that is functionally absorbed. The paradox of Theseus' ship illustrates the basic idea: "The planks of the ship corrode and gradually are replaced; the old planks are piled up in a shed. Finally all planks are exchanged. What is Theseus' ship – the renovated ship in the harbor or the pile of original planks stored in the shed?" Consistently, the renovated ship is considered Theseus' ship: despite exchanged materials, it conserves continuity of form and function over time.

In the normal case, same body and same person coincide. Yet in thought, experiments philosophers seek to disentangle variables. 'Brain transplant' devises the following situation: 'A's body is devastated in an accident, and B's brain destroyed by a tumor. A skillful surgeon succeeds in putting A's brain in B's body. Is the resulting person A or B?' Most would agree: the survivor is A – the person is where the brain is. A further thought experiment – 'Bionic replacement' – tests for acceptance of matter exchange: 'Cancer is disintegrating A's brain. A new technology allows gradually replacing A's brain by silicon parts. Finally, A ends up with an entirely bionic brain. Is postoperative A the same person he was before?' Again, agreement is assumed: the new brain is continuous with and serves the same function as the old one (i.e., it will sustain self-conscious mental life distinctive of the person). This answer marks a transition to relational theories. But first some objections to equating brain and identity will be presented.

These objections are as follows: (1) as noted above, emotions are psychophysiological processes (i.e., they essentially involve the body); (2) long-term stress or intense traumata are stored in a bodily memory and can produce permanent bodily disorders, and healings attained in psychosomatic medicine and trauma therapies evidence that body and mind are inseparably intertwined; and (3) early childhood experiences that happened before conscious memory functions have evolved are stored in the body (Buchholz, 2010). Altogether, exchanging bodies may well jeopardize identity.

Relational Theories

Psychological Continuity

John Locke (1632–1704) rejected the assumption that some substance is necessary for the retention of one's identity. He defined person as a thinking and intelligent being with reason and reflection whose identity reaches as far as his consciousness extends backward to past actions or thoughts (i.e., I am the same today as yesterday if I remember having had an experience as that person). Thus, the core criterion of identity is memory, a psychological relation connecting temporally distinct person stages.

Already his peers voiced objections. First, memory fades. Thomas Reid (1710–96) gives an example: a 40-year-old officer stealing the enemy's standard recalls stealing apples from the neighbor's orchard. As an old man, he remembers stealing the standard, but not stealing the apples. On Locke's account, the officer would have to be both identical to the apple-stealer (because of the transitivity relation that identity implies) and not identical to the apple-stealer (because he has no direct memory of the boy's experience). Second, Joseph Butler (1692–1752) notes that rather than constituting identity, memory presupposes identity, for I can only remember experiences as my own if I have a self-understanding. Remedies have been suggested. Chains of overlapping memories are substituted for the more rigorous requirement of direct memory continuity. And Locke's criterion of memory has been extended to include other psychological aspects (e.g., beliefs, goals, desires, and resemblance of character). In short, personal identity is taken to be constituted by psychological continuity.

Identity Does Not Matter

Parfit (1982; 1995) distends the conceptual strategy of equating personal identity with psychological continuity. He argues from the thought experiment 'fission': "My brain hemispheres are functional duplicates of each other. They are transplanted into the bodies of my two triplet brothers. After the transplantation both are psychologically continuous with me – they are both me. What has happened to me? I cannot be both and there exists no non-arbitrary criterion for deciding among the two. The only option left is: I do not survive. Yet this differs from normal death. The survivors seem to remember my thoughts and experiences, will have the same beliefs and desires, they will realize my intentions and their character will be like mine – in fact it will almost be as if I had survived." Parfit concludes: What matters for practical concerns is not identity but psychological connectedness. This relation to my replica contains all that matters – even though they are not me. Hence, my special concern for their future is not annulled. But what is normally understood as a self-interested prudential concern is now turned into a moral obligation – the obligation to care for others.

Some Interim Considerations

Comparing the Different Identity Theories

Exclusive Positions

Some theories assume only one constitutive criterion. (1) Immaterialist positions assert that persons essentially are souls. This belief, however, cannot be empirically verified. (2) Materialist positions maintain that persons are essentially biological organisms. This allows imputing numerical identity between a patient in a permanent vegetative state (PVS) and the adult, infant, or fetus he once was. It disagrees, however, with common responses to thought experiments like brain transplant. (3) Relational theories hold that persons are essentially conscious minds. This claim does not, however, accommodate the inseparability of body and mind noted above. (4) The identity does not matter view, too, neglects body-based aspects of persons. In fact, the easy acceptance of the brain solution in the transplant cases might reflect latent dualist preconceptions still prevalent in our everyday thinking habits.

Composite Positions

Some positions assume that persons require both body and mind, yet differ in the relative weight attributed to physical

versus psychological criteria. (1) The claim that the body criterion is more fundamental is backed by pointing to its greater usefulness in determining numerical identity: it is more extensive (the person is present whenever the body is there) and more varied (there are more and more reliable ways of determining physical sameness, e.g., direct and genetic fingerprints, blood tests, photographs, and witnesses) (Penelhum, 1972). (2) Some authors deem psychological criteria as more important. They see persons as constituted by consciousness. Body and brain are necessary to sustain mental life, yet this cannot be reduced to material conditions: Although I am made of matter, I am distinct from my body and brain (Garrett, 1998). Although biological continuity preserves continuity over time, it is the relation of psychological continuity that grounds our practical concerns (Shoemaker, 2012). These considerations show that criteria for the nature of persons and for determining sameness over time may diverge.

Renouncing a Univocal Identity Concept

Locke declared 'person' a 'forensic' term, a moral and legal concept denoting a responsible agent endowed with rights and duties. This understanding is highly relevant to many ethical issues. However, as Shoemaker suggests, there is no guarantee that there is one single correct conceptualization: "Why not think that some practical concerns bear a relation to one criterion or sense of identity, while other practical concerns bear a relation to a different criterion or sense of identity" (2012: 39). Thus, the biological criterion grounds claims to being compensated as an adult for what happened to him as a fetus. In contrast, signing advanced directives relies on psychological criteria. Also, the differing definitions of death in use hinge on different criteria for personal identity: the 'higher brain standard' according to which death occurs at the point of irreversible loss of the capacity of consciousness is based on psychological criteria, whereas the actual practice of medicating PVSs relies upon the biological criterion – death occurs when the organism dies.

Different Uses of the Basic Terms

Same

The word 'same' is used in a comparative and a numerical sense. In the comparative sense, "'same' if applied to two distinct things … means 'alike' and, if applied to one thing at different times … means 'unaltered' … (In) the numerical sense, in which two things are said to be the same are said not to be two, but one" (Penelhum, 1972: 99). These two senses correspond to different meanings of the identity question – the reidentification and the characterization question. Reidentification allows for a multitude of 'normal' changes. Thus, the fetus can easily be numerically identified with the grown adult and the demented patient if there is bodily continuity that is independently verifiable by various tests. What are considered 'normal' changes depend, of course, upon the way the object under scrutiny is conceptualized. A gold statue and the lump into which it is transformed when melted are the same if analyzed in terms of matter (e.g., in terms of the chemical properties of gold). They are not the same if considered in terms of a piece of art (Garrett, 1998). Analogously, attesting sameness of a person depends upon what 'person' is taken to mean.

Person

The term 'person' (cf. Danto, 1972; Bohlken, 2009) may refer to any human being, just as the word 'thing' refers to any object whatsoever. In fact, 'person' often is contrasted with 'thing' in order to point to the special treatment that human beings deserve (i.e., never to be used merely as a means) (Kant, 1959/1787). In a more restricted sense, 'person' specifies a legal term denoting a being with rights and duties held to be accountable. By implication, not all human beings are legally persons (e.g., children and idiots), and not all legal persons are human beings (e.g., a cooperation considered a juridical person). Equating person with accountability marks self-awareness and rationality as constitutive for personhood.

Substance-based theories equalize 'person' with some essential entity – the soul, body, or brain – which warrants identity across time. Relational theories refer to psychological features – memory, beliefs, desires, and similarity of character – thus denoting a specific state or phase of human beings. Infants, fully demented individuals, or PVSs are not considered persons. Both approaches, however, draw on criteria that can be ascribed from a third-person perspective for determining numerical identity. In contrast, theories of a sense of identity take a first-person perspective for dealing with the question: How do individuals achieve and stabilize a sense of inner unity and sameness?

Sense of Identity Theories

In the social sciences, this question has become topical. Thus, just in psychology the number of identity-related publications rose from about 8000 to almost 22 000 between 1984 and 1993 (Ashmore and Jussim, 1997: 5). There are sociostructural reasons for this heightened interest. In traditional societies, criteria determining social as well as personal identity converged across contexts, life cycles, and perspectives. To illustrate, the identification of X as 'tenant of the Z farm' simultaneously denotes X's social (e.g., man and farmer) and personal identities (e.g., the stout man living at the south end of the village). These identity markers are valid in all contexts (in his family, the village, and the pub, everybody knows him as the stout man running the Z farm) and across his life span (from birth to death, he is linked to the Z farm as its future, present, or former tenant). And descriptions from first- and third-person perspectives agree: his farmhands, wife, and other villagers know him as the tenant of Z, and he himself identifies with this attribution – in fact, so much so that he will have inscribed his tombstone, 'Here lies the tenant of Z farm.' With modernization, these dimensions diverge. With urbanization, social and personal identities separate: townspeople can only vaguely guess each other's social identity, let alone know their personal identity. Social differentiation and cultural pluralization produce consistency problems: individuals partake in diverse subsystems or subcultures, each operating according to their own code (Luhmann, 1985) or claiming specific values. Thus, across different contexts they often face contradictory expectations. And rapid social change jeopardizes experiencing personal continuity over the life course. From a third-person perspective, this is unproblematic. Members of modern societies are used to ignoring each other as strangers. When

someone's identity needs to be known, there are mechanisms for determining social identity (e.g., titles and uniforms) and personal identity (e.g., passport, genetic analysis, and witnesses). Problems may, however, arise from a first-person perspective: "Am I a unitary person – given I act so differently as a father, manager, and church member?" "Am I still the same person I was – after I changed my job, left the church, and got divorced?" Reference to unparalleled fingerprints does not help – individuals do not conceive of themselves in terms of physical markers.

Theories addressing this problem diverge. The main split concerns acceptance or rejection of the modern identity concept based on the criteria advanced by Erikson – consistency across situations and continuity across one's life cycle. Postmodernists claim that there is no consistent self. Others adhere to the modern concept, yet differ in the way they see identity grounded, in the mechanisms they assume will stabilize identity, and in the way they describe the stance that agents take toward themselves.

Postmodern Approaches

There are precursors to today's questioning of consistency and continuity. Thus, Montaigne (1533–92) notes: The self is "unstable and fluctuating from one minute to the next" (2005: 296). Hume (1711–76), too, attacks the spiritual-substance assumption of a unique and simple self. People are "a bundle or collection of different perceptions." Given their constant changes, identity does not exist.

Postmodern authors agree. The classical requirements of reliability and constancy no longer match modern living conditions (Gergen, 1990). These necessitate "a mindset able to leave contradictions and not dominated by an obsessive need for identity" (Keupp, 1997: 17). New constructs are invented, for example, 'tinker mentality' (Gross, 1985), 'patchwork identity' (Keupp, 1988), 'the plural subject' (Welsch, 1993), and 'identity as collage' (Pazzini, 1986). Such revisions are influenced by French poststructuralists debunking classical views of the subject as essentialist, naturalizing, and totalitarian. Thus, Derrida (1983; cf. Zirfas, 2010) deconstructs the idea of identity as selfsameness: given multiperspectivity and mutability, identity no longer aims at concordance. Instead, identity is constituted by difference to others; it is fragile, in a constant state of flux, unequivocal, and plurivalent. Adapting to 'The Age of the Internet,' Turkle (1995) replaces the modern identity question 'who am I?' by the postmodern question 'who am we?' and responds by noting, 'I am many.' The Internet is like a vast social laboratory. People experiment with ego constructions and reconstructions and leave with patchwork identities fitting in with postmodern life conditions.

Mostly, these authors applaud renouncing the unity of the person. Replacing unity and sameness with difference and heterogeneity, discontinuity and flexibility, and plurality and fragmentation is praised as liberation from obsessive identity constraints one-sidedly fixed to rationality and reason.

Objections have been raised. (1) Resemblances to schizophrenia or multiple personality disorders have been noted (Straub, 2000). (2) In fact, findings from dissonance theory and attribution theory, studies on dogmatism and split personality, clearly evidence that at least some consistency is indispensable for psychological well-being and mental health. (3) The fitting of a multiple decentered self-system with neoliberal demands for flexibility is criticized (cf. Münte-Goussar, 2010: 287f) and (4) Turkle's net user is denounced as "digital reincarnation of the modern Cogito-subject … freed from all bodily and sensual contaminations" (ibidem: 292f).

In response, mediating concepts have been introduced: a 'transversal subjectivity' is postulated, pervading subjects' differentiated parts and preventing the decomposition of plurality into 'multiphrenia' (Welsch, 1996: 847; cf. Klika, 2010). A 'linking competence' is devised, affording individuals a sufficient degree of coherence (Keupp, 1999: 271).

Modern Theories

These approaches assume the indispensability of a sense of psychic unity: subjects can only act if fleeting and contradictory ideas, inner multiplicity and heterogeneity, are somehow organized and structured. Yet, theories differ in their assumptions about how this is to be accomplished. These concern the grounding of identity – in material contents, in formal competences, or in both; the mechanisms stabilizing identity across time – consistent narratives and commitment to values; and, most importantly, the stance subjects are assumed to take toward themselves, whether a theoretical-observing or a practical one.

Material Content or Formal Competence

Traditional approaches deal with the establishment of identity and make content the basis for individuals' sense of unity and continuity: identity is socially determined by positions held and is experienced as unquestionable inasmuch as individuals have developed "need dispositions toward the fulfillment of role expectations" (Parsons, 1964: 32). This role identity is prevalent in societies in which positions are ascribed and social expectations internalized from early on.

For Erikson, ego-identity encompasses content and competences. Content consists in lifelong substantive commitments (e.g., to a partner, occupation, or worldview) incurred at the successful resolution of the adolescence crisis. Formal competences are then required for maintaining identity. The core is "accumulated trust that the unity and continuity one enjoys in the eyes of others corresponds to a competence to uphold inner unity and continuity" (1973: 107). Nowadays, however, people keep experiencing ever new crises (e.g., midlife crises and retirement crises) in which former commitments are revoked (e.g., by divorce, religious or ideological conversions, or loss of employment). Giddens takes this into account. He no longer speaks of commitments entered once and for all but of a multiplicity of choices: "Each of the small decisions a person makes every day – what to wear, what to eat, how to conduct himself at work, whom to meet with later in the evening … all such choices (as well as the larger and more consequential ones) are decisions not only about how to act but who to be … Lifestyle concerns the very core of self-identity in making and remaking" (1991: 81). Nevertheless, a formal competence – narrative competence – is required as well (discussed below). The model 'identity qua autonomous commitment' (Nunner-Winkler, 2002), too, integrates content and competence. Commitment needs an object (i.e., content is required). And

autonomous commitments presuppose competences (e.g., self-distancing and reasoning powers).

Krappmann defines identity exclusively on formal terms, debunking any substantive commitments as signs of rigidity, alienation, and reification. Identity is created in situated processes of balancing divergent expectations of different interaction partners and one's own strivings for "uniqueness and irreplaceability" (1969: 48). This involves "the efforts of an artist constantly juggling and balancing" (p. 56) and presupposes formal ego resources such as tolerance of ambiguity, role distance, and empathy.

Mechanisms for Stabilizing Identity

How do individuals uphold an inner sense of consistency and continuity despite changes? Some theories rely upon narrative competence: self-narrations "transform experienced contingency into something meaningful and significant … (integrating) different, even heterogeneous material into one coherent gestalt" (Straub, 2000: 132). Giddens, too, assumes that individuals who keep altering their lifestyle decisions, and thus their identity, need narrative competence to maintain coherence. "Self-identity becomes a reflexively organized endeavor … which consists in the sustaining of coherent, yet continuously revised, biographical narratives" (1991: 5).

Narrative competence, however, does not provide a convincing solution. People may unwittingly distort or repress unwelcome facts or experiences or construe rationalizations. This has implications for identity: upholding defenses consumes strength, and it may amplify distortions and entail unsuitable decisions.

Shoemaker notes the attractiveness of the concept of narrative identity given its good fit with the practical issue of moral accountability. Yet – so his philosophical objection – it does not answer the reidentification question – the proper question to ask for clarifying numerical identity. Instead, it answers the characterization question 'Who am I really?' Indeed, narrative identity presupposes numerical identity: "one can't be a person, on the narrative view, unless one gathers up the various experiences one has as a subject of experiences into a coherent narrative, but then the identity of that subject of experiences must be preserved across time for its experiences to be so gathered up" (2012: 12).

Another mechanism for stabilizing identity across time lies in autonomous commitment. This device, however, depends on a different description of the self-relation underlying the identity question.

Theoretical-Observing versus Practical Stance

Identity can be judged from two attitudes. In classical philosophy, identity is seen to arise "from comparing the content of consciousness with itself at different times and places" (Eisler, 1904: 482). Tugendhat criticizes this subject–object model and its epistemological presupposition that "all direct empirical knowledge must be based on perception … even knowledge about oneself" (1979: 34). Referring to Heidegger, he argues for conceptualizing agents' relationship to themselves as a practical one: "In saying 'I am' I do not in an 'aesthetically' observing mode refer to my being, I do not state my existence as a given, rather I take a stance in the mode of self-caring … I behave towards my being as to one I have to be" (p. 36). This is not "a theoretical comment on beliefs but a volitional practical stance" which implies "saying 'yes' or 'no' to being this or that way or to acting this or that way" (p. 37). It requires the ability to distance oneself from and reflect upon one's spontaneous desires. Tugendhat's analysis agrees with Frankfurt's conception of 'person': "A person is … prepared to … organize the preferences and priorities by which his choices are ordered … to consider whether what attracts him is actually important to him" (1988: 23). Thus, commitments to specific 'ideals' is a necessary constituent of personhood: if someone had nothing he cares so much about that he cannot bring himself to betray it, he is not a person but an object at mercy of his inner drives, fleeting whims, external constraints, or contingent circumstances (Frankfurt, 1993). Koorsgaard (1989), too, advocates a practical stance: we need to see ourselves as practical agents. This requires conceiving of oneself as a unified person at any time and across time, given we have only one body with which to act.

Postulating autonomous commitment as a core mechanism for achieving and stabilizing a sense of identity relies on the practical stance model. In the following, I will present some empirical evidence supporting this conception.

Sense of Identity – Some Empirical Findings

In philosophical debates, thought experiments are often used to clarify the structure of concepts. Such reflections draw their persuasiveness from the assumption that there is general agreement – about the best description of the thought experiments and about the intuitions guiding the responses (Garrett, 1998: 13ff). Thus, theoretical conclusions are based on what variously is described as "our considered judgments" (ibidem 1989: 50), "our core intuitions" (p. 51), "general agreement" (p. 52), "our core set of common-sense judgments" (p. 56), and "(the answer) most of us would give" (p. 91). Social sciences can test such assumptions. This is one of the many ways in which empirical research and philosophical reflection interact.

Findings about three topics will be reported: everyday understanding of the sameness of persons, constitutive facets of a sense of identity, and the stability of value commitments across a lifetime.

Everyday Understanding of the Sameness of Persons

Representative samples (of adolescents, university students, and elderly people) were presented in the following hypothetical scenarios in which a person experienced change (Nunner-Winkler, 2002):

1. Alzheimer's: by a grave disease (e.g., Alzheimer's), a man is reduced to the state of an infant (Nozick, 1981).
2. Brainwashing: a convinced democrat by brainwashing turns into a communist (ibidem).
3. 68 adherent: A formerly convinced adherent of the 1968 movement often protested against war and force of arms. Today he argues for military interventions in conflict areas.
4. Autobiographical memory: by an accident, a man suffers loss of his autobiographical memory. His powers of reasoning are intact.

In scenarios 1–4, the question was: 'Does … remain the same person? Why?/Why not?'

5. Nuclear power plant: Y from firm conviction got involved in Greenpeace. At the same time, he holds a leading position in a nuclear power plant. Does Y have identity problems? Why or why not?

The identity relevance of reasoning power (scenario 1) and autobiographical memory (scenario 4) was controversial (with about one-third in each case affirming or denying the identity or remaining undecided). The other items drew high agreement. Subjects ascribed most problems to persons who experienced a radical change in beliefs due to external manipulation (brainwash) or who acted against their beliefs (nuclear power plant). Few problems were expected in case of a 'normal' change in attitudes (68 adherent).

Thus, in general, beliefs are ascribed high identity relevance – an interpretation that is confirmed by justifications provided (e.g., "Principles are part of one's personality – someone turning around 180° is not the same man as before"). This is true, however, only of subjects who deem the specific beliefs important – either from their own perspective or from the assumed perspective of the protagonist (e.g., brainwash – different person: "Political beliefs are an important part of one's personality" and "Given he cares much about politics, these convictions are a significant aspect of his personality") – or believe they were betrayed (e.g., nuclear power plant – "Yes, he has big identity problems – after all, his beliefs are contradictory"). Subjects expect no identity problems, if they either do not care much about politics (e.g., brain wash – same person: "He only changed political attitudes") or believe that basic convictions have not really been betrayed (e.g., nuclear power plant – same person: "Maybe he accepts the NPW but fights against marine pollution or CO_2 emission").

Most subjects in referring to beliefs explicitly commented on the way the change came off. Manipulation (brainwashing) is seen to destroy identity (e.g., "Only if he changed on his own will, will he be the same person"). However, no problem arises if changes are seen to result from learning processes. These can be implicit: attitude change is viewed as a normal part of individual development (e.g., in the case of the 68 adherent, one could opine, "He has matured" or "Beliefs can change – children believe in Santa Claus"). Or they can result from explicit learning processes, i.e., by choice and with good reasons (e.g., "Beliefs have genuinely changed – of his own free will" or "He changed by rational thinking").

Facets of a Sense of Identity

Subjects were also given the following question: 'Imagine you would have different parents/different ideas about right and wrong/different looks/a lot more money/different interests resp. a different field of study resp. a different occupation/different friends/you were born and raised in a different country/belonged to the other sex. How different a person would you be?'

Across generations very few subjects gave high importance to 'money' and 'looks', whereas 'ideas about right and wrong' figured among the three most highly rated items (e.g., "If I had different principles, I'd be a different person"). The other two important items were parents and sex membership. As justifications indicate, these mostly stand for values, too (e.g., "Ideals are taken on from parents" and "Gender roles inform attitudes").

Despite these overall agreements, there are also clear interindividual differences. Some subjects accord high importance to 'looks' (e.g., "I define myself by my appearance"), but most don't (e.g., "Even if I had different looks, I'd still be the same person"). Some care a lot about their field of study (e.g., "If I got the place in psychology, I'd have a goal"), whereas others don't (e.g., "I don't define myself by university or by work. That's merely necessary for making a living"). Some see sex membership as important (e.g., "I am very feminine. I would not want to miss these weapons"), but others don't (e.g., "Different sex would not change me much; my basic attitudes toward life would remain the same").

Stability of Value Commitments across a Lifetime

A representative sample of young, middle-aged, and elderly adults were asked to evaluate 25 possibly morally relevant scenarios (e.g., homosexuality, working mother, and leaving the church) (Nunner-Winkler, 2005). Large differences were found between the generations. This change in moral understanding can be attributed to a change in the grounding of morality. Traditionally, moral norms were derived from authorities – God, the church, and nature. Today, they are justified in terms of a hypothetical consensus of all concerned. This change has led to a minimal morality that forbids harming others directly (e.g., do not kill, lie, steal, etc.) or indirectly (e.g., do your duty) and attributes all other behavior to the personal realm. To illustrate, most of the older subjects condemned homosexuality (e.g., by stating that 'it is sinful', 'unnatural', 'pathological', 'sickening'), whereas most of the youngest accepted it ('whom love befalls' and 'that's their decision'). Interestingly, most older subjects – though well aware that the zeitgeist has changed (e.g., 'younger people are more liberal') – held on to attitudes that they acquired in adolescence.

Implications for Sense of Identity Theories

The disagreement found in border cases (Alzheimer's and loss of autobiographical memory) matches philosophical disputes about the relative importance of physical or psychological criteria. In normal cases, however, everyday judgments rely on psychological criteria backing the concept of identity qua autonomous commitment: most subjects saw their convictions as decisive in defining their personhood. Most of the older subjects held on to their former beliefs – maybe as a way of stabilizing identity in times of rapid changes. Alterations of beliefs, however, are not excluded. Individuals changing their mind of their own free will and with good reasons (i.e., by learning processes) do not jeopardize their identity.

This model disagrees with other approaches. (1) Gidden's claim that all lifestyle choices are relevant to identity does not hold up. Individuals decide which choices they deem important – and they differ in these decisions. (2) Krappmann's rejection of content as alienating is not shared – subjects understand their value commitments as constitutive of their personhood. (3) Narrative competence may not be so relevant. The (assumed)

truth of essential beliefs counts more than the way they are acquired. To illustrate: being sure that Jesus is God's son presumably means more to Paul than the story of his conversion at Damascus. (4) The value commitment model questions the common assumption that identity requires difference: if I identify with specific values, I want them not to be distinctive of me but to be shared and propagated by others, too. The emphasis on difference seems to reflect a confusion of identity and uniqueness. Thus, people strive toward unique interests, original hobbies, and special lifestyles to secure their irreplaceability. Frustrated, they then experience that all newly invented identity markers are socially absorbed and commercially coopted (Cohen and Taylor, 1976). However, sense of uniqueness does not result from public displays of singular features, but from authentic and genuine commitment (Keniston, 1968; Nunner-Winkler, 1985): it is a necessary by-product, and it will inevitably be missed if directly accessed (Elster, 1987).

Conclusion

The identity questions – what constitutes and stabilizes personal identity? And how do people attain a sense of identity? – have found different answers. Numerical identity in substance-based theories rests upon some essence that remains constant (soul, body, and brain), in relational theories on psychological connectedness (memories, beliefs, and desires). The relative weight of physical versus psychological criteria is disputed. This theoretical dissent corresponds to disagreements in everyday understanding (e.g., about how to deal with loss of reasoning power) and in common practices (e.g., medically supporting PVSs yet honoring advanced directives).

Sense of personal identity is addressed by theories taking a first-person perspective. Postmodernists claim that personal unity is neither existent nor necessary. Instead, they rejoice the emancipation from identity constraints. In contrast, classical theories assume that individuals need some consistency and continuity, yet differ in how they see this requirement fulfilled. Empirical research supports the assumption that individuals attain a sense of identity by committing themselves to values that they hold despite outer changes but revise in the light of new insights.

The two discourses are interrelated: the criteria used in philosophical reflections for defining identity must have some counterpart in everyday understanding. And social science analyses of components of a sense of identity are relevant for the constitutional debate inasmuch as they accommodate the fact that persons – in contrast to things – have a first-person perspective.

See also: Autonomy, Philosophy of; Free Will and Action; Person and Self: Philosophical Aspects; Self: Philosophical Aspects.

Bibliography

Ashmore, R.D., Jussim, L., 1997. Self and Identity. Fundamental Issues. Oxford UP, NY, Oxford.
Bohlken, E., 2009. Person. In: Bohlken, E., Thies, C. (Eds.), Handbuch Anthropologie. Stuttgart & Weimar, J.B. Metzler, pp. 391–395.
Buchholz, M.B., 2010. Über den Individualismus hinaus. Die Entwicklung des Selbstempfindens bei Daniel N. Stern und einige Befunde der Säuglingsforschung. In: Jörissen, B., Zirfas, J. (Eds.), Schlüsselwerke der Identitätsforschung. VS, Wiesbaden, pp. 69–86.
Cohen, S., Taylor, L., 1976. Escape Attempts. The Theory in Practice of Resistance to Everyday Life. Allen Lane, London.
Danto, A.C., 1972. Persons. In: Edwards, P. (Ed.), The Encyclopedia of Philosophy, vol. 6. Macmillan Publ. Co., Inc. & The Free Press, NY, London, pp. 110–114.
Derrida, J., 1983. Grammatologie. Suhrkamp, Frankfurt a.M.
Eisler, R., 1904. Identität. In: Eisler, R. (Ed.), Wörterbuch der philosophischen Begriffe, second ed. vol. 1. Mittler, Berlin, pp. 482–485.
Elster, J., 1987. Subversion der Rationalität. Campus, Frankfurt/New York.
Erikson, E.H., 1973. Identität und Lebenszyklus. Suhrkamp, Frankfurt a.M.
Frankfurt, H.G., 1988. The Importance of What We Care About. Philosophical Essays. Cambridge University Press, Cambridge/New York.
Frankfurt, H.G., 1993. On the necessity of ideals. In: Noam, G., Wren, T.E. (Eds.), The Moral Self. MIT Press, Cambridge, MA, pp. 16–27.
Garrett, B., 1998. Personal Identity and Selfconsciousness. Routledge, London.
Gergen, K.J., 1990. Die Konstruktion des Selbst im Zeitalter der Postmoderne. Psychologische Rundschau 41, 191–199.
Giddens, A., 1991. Modernity and Self-Identity: Self and Society in the Late Modern Age. Polity Press, Cambridge.
Goffman, E., 1963. Stigma. Notes on the Management of Spoiled Identity. Prentice Hall, Englewood Cliffs.
Gross, P., 1985. Bastelmentalität: ein postmoderner Schwebezustand. In: Schmid, T. (Ed.), Das pfeifende Schwein. Wagenbach, Berlin, pp. 63–84.
Kant, I., 1787/1959. Kritik der praktischen Vernunft. Felix Meiner Verlag, Hamburg.
Keniston, K., 1968. Young Radicals – Notes on Committed Youth. Harcourt, Brace and World, New York.
Keupp, H., 1988. Auf dem Weg zur Patchwork-Identität? Verhaltenstherapie und psychosoziale Praxis 4, 425–438.
Keupp, H., 1997. Diskursarena Identität: Lernprozesse in der Identitätsforschung. In: Keupp, H., Höfer, R. (Eds.), Identitätsarbeit heute. Klassische und aktuelle Perspektiven der Identitätsforschung. Suhrkamp, Frankfurt/Main, p. 17.
Keupp, H., 1999. Identitätskonstruktionen. Das Patchwork der Identitäten in der Spätmoderne. Rowohlt, Reinbek.
Klika, D., 2010. Wolfgang Welsch: Identität im Übergang. In: Jörissen, B., Zirfas, J. (Eds.), Schlüsselwerke der Identitätsforschung. VS, Wiesbaden, pp. 297–314.
Koorsgaard, C.M., 1989. Personal identity and the unity of agency: a Kantian response to Parfit. Philosophy & Public Affairs 18, 101–132.
Krappmann, L., 1969. Soziologische Dimensionen der Identität. Klett, Stuttgart.
Luhmann, N., 1985. Soziale Systeme. Suhrkamp, Frankfurt.
de Montaigne, M., 1953. Essais. Auswahl und Übertragung von Herbert Lüthy. Manesse, Zürich.
Münte-Goussar, S., 2010. Ich ist viele. Sherry Turkles Identitätstheorie. In: Jörrisen, B., Zirfas, J. (Eds.), Schlüsselwerke der Identitätsforschung. VS, Wiesbaden, pp. 275–296.
Nozick, R., 1981. The identity of the self. Philosophical Explanations. Harvard University Press, Cambridge, MA, pp. 27–114.
Nunner-Winkler, G., 1985. Identität und Individualität. Soziale Welt 36, 466–482.
Nunner-Winkler, G., 2002. Identität und Moral. In: Straub, J., Renn, J. (Eds.), Transitorische Identität. Der Prozesscharakter des modernen Selbst. Campus, Frankfurt a.M., New York, pp. 56–84.
Nunner-Winkler, G., 2005. Changes in moral understanding – an intergenerational comparison. In: Edelstein, W., Nunner-Winkler, G. (Eds.), Morality in Context. Elsevier, Amsterdam.
Parfit, D., 1982. Personal identity and rationality. Synthese 53, 227–241.
Parfit, D., 1995. The unimportance of identity. In: Harris, H. (Ed.), Identity. Oxford UP, Oxford.
Parsons, T., 1964. The Social System. The Free Press of Glencoe, London.
Pazzini, K.J., 1986. Ein bildlicher Zugang zum Identitätskonzept. Loccumer Protokolle 58/'85: Jugendarbeit und Identität, pp. 61–80.
Penelhum, T., 1972. Personal identity. In: Edwards, P. (Ed.), The Encyclopedia of Philosophy, vol. 6. Macmillan Publ. Co., Inc. & The Free Press, NY, London, pp. 95–107.
Perry, J., 1978. A Dialogue on Personal Identity and Immortality. Hackett Publishing Company, Indianapolis, IN.
Roth, G., 2003. Fühlen, Denken, Handeln. Suhrkamp, Frankfurt.
Shoemaker, D., 2012. Personal Identity and Ethics. Stanford Encyclopedia of Philosophy, Standford, USA, pp. 1–51.
Singer, W., 2002. Der Beobachter im Gehirn. Suhrkamp, Frankfurt/Main.
Straub, J., 2000. Identitätstheorie, empirische Identitätsforschung und die 'postmoderne' armchair psychology. In: Kulturwissenschaftliches Institut im Wissenschaftszentrum N. R. W, Jahrbuch 1999/2000. Stattwerk e. G. Essen, Essen, pp. 125–156.
Tugendhat, E., 1979. Selbstbewußtsein und Selbstbestimmung. Sprachanalytische Interpretationen. Suhrkamp, Frankfurt.

Turkle, S., 1995. Life on the Screen. Identity in the Age of the Internet. Simon & Schuster, New York.

Welsch, W., 1993. 'ICH ist ein anderer'. Auf dem Weg zum pluralen Subjekt? In: Reigber, D. (Ed.), Frauen-Welten. Marketing in der postmodernen Gesellschaft – ein interdisziplinärer Forschungsansatz. Econ-Verlag, Düsseldorf u.a., pp. 282–317.

Welsch, W., 1996. Vernunft. Die zeitgenössische Vernunftkritik und das Konzept der transversalen Vernunft. Suhrkamp, Frankfurt a.M.

Zirfas, J., 2010. Jacques Derrida: Das andere Kap. Die vertagte Demokratie. Zwei Essays zu Europa. In: Jörrisen, B., Zirfas, J. (Eds.), Schlüsselwerke der Identitätsforschung. VS, Wiesbaden, pp. 241–258.

Personal Intelligence and Competencies

John D Mayer, University of New Hampshire, Durham, NH, USA
David R Caruso, Yale University, New Haven, CT, USA
AT Panter, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Personal intelligence concerns the ability to reason about personality and personality-related information; it includes both self-knowledge and knowledge about the personalities of other people. Personal intelligence encompasses a wide range of areas of reasoning from perceiving cues to personality to planning one's life. There are individual differences in such reasoning, and some people are much better at such reasoning than others. The article examines the nature of intelligences, personality, and personal intelligence. The skills involved are examined as well as how they are measured and the relation of the intelligence to life criteria.

We vary as to how well we understand personalities – or, at least, how well we *think* we do. Some of us claim great skill in this area; others less so. Oprah Winfrey claims great empathic ability – and she argued that her success as a daytime television host was because she relied on her skill in drawing out others:

> I think the show's been successful because … I relate to the core of everyone's pain and promise because I've known pain and promise … Having that understanding and that connection has really given me wings to fly…
>
> Kantrowitz et al., 24 October 2005: p. 48

The best-selling author Tucker Max – the author of the rowdy, *I Hope They Serve Beer in Hell*, by contrast, has represented himself as 'not very adept at picking up subtle social cues' – although whether this is part of a satirical façade or an honest self-appraisal is difficult to discern (Max, 2009: p. 58).

The idea that there exist skills to understand personality and that people vary in those skills is an ancient one. In 2500 BCE, the Egyptian sage Ptahhotep instructed leaders to interview those around them in private to gauge their character, and by 200 BCE, Theophrastus had outlined a score of personalities to watch for, from the miserly person to the talkative one (Ptahhotep, 2003/*c*.2450 BCE; Theophrastus, 1970/*c*.319 BCE).

Personal intelligence, the topic of this article, concerns the ability to reason about personalities – both one's own and the personalities of other people. According to a theory of personal intelligence, there are important forms of reasoning about personality that can be described, and some people are much better at such reasoning than others (Mayer, 2008). In this article, we examine the nature of intelligences and of personality, personal intelligence as a theory, the skills involved and how they are measured, and the relation of the intelligence to life criteria.

A Contemporary View of Intelligences

Personal intelligence is concerned with thinking clearly and accurately both about one's own personality and the personalities of others. To understand personal intelligence, it helps to begin with an understanding both of contemporary intelligences and of personality itself.

Intelligence Theories Today

Perhaps the most influential theories of today view intelligence as a three-tiered hierarchy of mental abilities from the most general at the top to highly specific mental abilities at the bottom – sometimes referred to as the three-strata or Cattell–Horn–Carroll model (e.g., Carroll, 1993; McGrew, 2009). At the top of the hierarchy is *g* or general intelligence. Charles Spearman first proposed *g* as a general capacity to reason abstractly: to see similarities and differences, to follow the logic of an argument, and to draw correct conclusions from information (Spearman, 1904). And, as its name implies, *g* is general – a set of capacities that are applicable to many diverse domains of human problem-solving (Spearman, 1904; Sternberg and Grigorenko, 2002). Commensurate with its uppermost position in the three-stratum hierarchy, *g*'s position distinguishes it from the broad areas of problem-solving people carry out at the second level immediately beneath it. These still-broad areas of problem-solving – 'broad intelligences' – are estimated at between 8 and 14 in number. They include fluid reasoning (the ability to deduce principles from information), verbal intelligence, visual-spatial thinking (the ability to store and recall visual representations), long-term memory and retrieval, auditory processing (including the ability to discriminate speech sounds), processing speed, performance (mechanical) intelligence, and working memory capacity (Carroll, 1993).

At the bottom of the hierarchy are the nitty-gritty specific abilities on which our cognition rests: in the verbal–comprehension area, these include decoding words, spelling ability, writing speed, and reading comprehension, among others. As a second example, in the 'long-term storage and retrieval' area of broad intelligence, the skills include associative memory, associational fluency, and free-recall memory. Seventy-three such specific skills have been identified according to an estimate, and perhaps there are upwards of 100 (Carroll, 1993; McGrew, 2009).

The 'Felt' or 'Hot' Intelligences

A group of intelligences that are not yet included in the three-stratum models are the 'hot' or 'felt' intelligences – personal, emotional, and social intelligence – so-called because when we use them we are thinking about information that especially

matters to us and is deeply relevant to our relationships with others. Robert Abelson first suggested that people engage in 'hot cognition' when their attention shifts from traditional problem-solving involving language, objects, and spaces to matters that are of particular personal concern to them (Abelson, 1963). In the 1980s, researchers uncovered a host of effects of such personal information on thought, creating a new field called 'cognition and affect' along the way (Abelson, 1963; Blanchette and Richards, 2010; Bower, 1981; Isen et al., 1978).

For example, people who are in good moods often see positive events as more likely and positive instances as better examples of categories than those in negative moods (Blanchette and Richards, 2010). Additionally, psychobiologists have detailed how memory about the self, for example, involves different brain regions than memory more generally, and information that challenges one's self-knowledge is processed differently from other information (Klein et al., 2004; Vaillant, 2000). The idea that some broad intelligences might focus on such personally relevant hot or 'felt' intelligences was a natural development of this line of thinking (Mayer, 2014; Salovey and Mayer, 1990). That said, some psychologists continue to debate whether hot intelligences exist and can be measured (e.g., Locke, 2005; Maul, 2011; Zeidner et al., 2001). And because measures of these felt intelligences have been developed only recently, they have yet to be integrated into the hierarchical model of intelligence.

Personality and Personal Intelligence

Among those felt intelligences, personal intelligence is unique in its focus on personality itself. Just as verbal intelligence concerns a focus on language and spatial intelligence concerns three-dimensional objects in space, personal intelligence involves an understanding of how personality operates.

According to one contemporary view, personality is closely connected to four surrounding systems: First is our body (and especially our brain), from which personality emerges; personality must meet the needs of those biological systems. Second, personality navigates the setting in which we find ourselves – on a train, at work, walking down the street, or in a sports arena. These settings provide a foundation, in turn, for the psychological situations (third) that emerge from them. These situations include such meaningful sets of actions as getting up in the morning, going to work, asking a colleague for help, or painting ourselves blue and screaming at the top of our lungs in a basketball arena. And fourth, there are groups to which personality belongs – political parties, religious or secular groups, artistic movements from which we draw, in part, our identity (Mayer, 2005; Mayer and Korogodsky, 2011).

These four surrounding systems tell us what personality must do: our personalities negotiate our personal needs – those wants and desires that are directed in part by our body, brain, and psychological systems. To meet our needs, we act in the outer environment: our settings, situations, and groups – through negotiation, reciprocation, love, and sometimes through conflict or self-deprivation (Buss, 2008; McAdams and Pals, 2006).

To negotiate these relationships and attain benefits from them, each of us has at our disposal our major psychological subsystems – our motives, emotions, intelligences, and models of ourselves and other people. According to contemporary psychologists, personality, in essence, is the system that organizes the rest of an individual's psychological systems – those same motives, emotions, capacities for self-management, and other areas of functioning (Mayer, 2007). Although personal intelligence begins with an intuitive understanding of who we and other people are, it may proceed to a more formal understanding that may approximate to some degree what psychologists have learned about personality. Personal intelligence is personality's guidance system – the part of personality that helps us to motivate ourselves, understand our social and emotional traits and the traits of others, and to choose our goals and understand other people's goals (Mayer, 2014).

Personal Intelligence: Theoretical Bases, Precursors, and Related Concepts

Theoretical Bases and Precursors

Theoretical Bases

The contemporary approach to personal intelligence is an outgrowth of many previous ideas. One theme that wends its way through earlier works is that throughout human evolutionary history, the ability to reason about one's own and others' personalities conferred an adaptive advantage. According to Dunbar (2009), for example, evolution selected for larger human brains so that our evolutionary ancestors could keep track of the behaviors of those in their group – through observing them directly, but also through gossiping about them. For example, Haselton and Funder have observed that judging others' personalities has the characteristic of an instinct in that the phenomenon happens automatically, quickly, and universally (Haselton and Funder, 2006).

Converging evidence for this innate interest in personality comes from our earliest written documents – in which can be found a concern with the issues of personality from around the world (Mayer et al., 2011). Already during the Great Transformation – the period from 800 BCE to 200 BCE – written records frequently addressed the best way to live, and how to guide one's personality and to understand the personalities of others. In ancient China, Confucius spoke of the different personalities of his students and how they required different handling. In the Middle East, Kohelet (Ecclesiastes) concerned himself with how best to live a life. In ancient Greece, Plato extolled the virtues of self-knowledge, and Aristotle's successor in the Greek Lyceum, Theophrastus, enumerated over a dozen character types mostly recognizable still today, including the talkative type, the stingy type, and even distinguishing the city slicker from the country bumpkin (Theophrastus, 1970/*c*.319 BCE). Writers took up the continued description of personalities in their works, and engaged in 'characterology,' developing highly nuanced literary depictions of different types of people (Roback, 1927). Physicians sought to understand differences in personality and how a person's physical health might bring those about (Ellenberger, 1956).

Precursors

More recent precursors of personal intelligence emerged during the past century of psychological research. For example, in 1947, Wedeck sought to discover whether there was an 'ability

to understand psychology,' that was independent of Spearman's *g*; Wedeck's results were somewhat disappointing as the findings provided weak evidence at best for such a new area of reasoning (Wedeck, 1947). A few years later, at the Menninger Clinic in Kansas, a group of psychologists began studying 'psychological mindedness.' Their conception was that certain patients had an ability to 'think psychologically' about themselves and others (Appelbaum, 1973). Psychological mindedness, as those mental health practitioners saw it, involved "A person's ability to see relationships among thoughts, feelings, and actions with the goal of learning the meanings and causes of his experiences and behaviour" (Appelbaum, 1973: p. 36). And, it included:

> …the easily observed talent that some people have, and others lack, to 'see', 'know', 'feel' covert and inexplicit psychological events … [with] … the goal of learning the meanings and causes of … behaviour…
>
> Appelbaum, 1973: p. 36

Gardner (1983) drew on such earlier ideas when he suggested there might be two personal intelligences, as part of his vision of multiple intelligences (MIs). The first of these was 'intrapersonal' and the second, 'interpersonal.' The core of intrapersonal intelligence involved deciphering inner-feeling states – overlapping in part what, today, is called emotional intelligence – and this intrapersonal intelligence, as Gardner viewed it, also concerned a sense of identity (Erikson, 1950). For Gardner, people also drew on the second, 'interpersonal' intelligence, to understand other people, their personalities, the rules of society, and adherence to social customs (Gardner, 1983). Yet, Gardner actively discouraged anyone interested in his theories from developing intelligence tests based on them (Gardner, 1999: p. 80). He may have succeeded in part in dampening such empirical investigations: as one researcher recently opined, "Despite the popularity of the MI idea in some educational circles, there is virtually no evidence for the theory" (Hunt, 2010: p. 119).

Alongside psychological mindedness, and the intra- and interpersonal intelligences, some researchers examined specific abilities that were (and are) highly relevant to a possible personal intelligence. Funder (2001) speculated on the characteristics of a 'good judge' – an individual who is particularly accurate at evaluating a target-individual's personality, and he outlined characteristics of a good judge, including extraversion and general intelligence. Pillemer (1998) concluded in his book on personal memories that there might exist people with an 'autobiographical intelligence' who employed recollections from their past to motivate and guide themselves.

Drawing on sources that included ability conceptions of emotional intelligence, Funder's concept of the good judge, and studies by Jackson (1972) of inferential accuracy, Christiansen and colleagues proposed a theory of a 'dispositional intelligence' that drew on "Three declarative knowledge structures …: (a) knowledge of trait-behavior linkages, (b) understanding of situation-trait relevance, and (c) proficiency at trait concepts" (Christiansen et al., 2005: p. 126); and the authors developed a scale we will review in our section on measurement.

A Theory of Personal Intelligence

A pair of theoretical articles in 2007 and 2008 drew together many of these earlier works and described a new theory of personal intelligence. The theory divided personal intelligence into four areas of problem-solving: (1) identifying cues to personality, (2) forming models of oneself and of other people, (3) guiding choices using personally relevant information, and (4) systematizing one's goals and life plans. These areas are shown in **Figure 1** (Mayer, 2008, 2009). Before the theory, the term 'personal intelligence' had been used in a number of ways, from the title of a gossip column to a shorthand way to remark upon Gardner's intrapersonal intelligence (see, for example, Park and Park, 1997).

The characteristic attributes of people high in these mental abilities were outlined in a second article developing the theory. For example, those who are especially good at 'identifying cues to personality' might often express an unstable sense of self as new information was drawn inward; they also were better at seeking feedback, often through asking nonbiasing questions about themselves, such as 'how can I improve at this job,' as opposed to the more biasing, 'am I doing a good job?' Individuals who were especially good at guiding their own choices were expected to better fit their personalities with their occupations, and those who could systematize their own plans would exhibit "goals that are well chosen and mostly consistent with each other" (Mayer, 2009: Table 2, p. 52). With the theory itself laid out, we turn next to empirical work in the area.

Measuring Personal Intelligence and Related Competencies

Intelligence, Competencies, and Ability Measurement

The optimal approach to measuring an intelligence or a competency is to employ an ability approach – that is, test-takers are given problems to solve in the area and their answers are compared against a standard of correctness. Because people are generally unable to give accurate self-reports in the area, there is little point to asking people how smart or competent they are. There are many reasons for this, including that many people lack the understanding necessary to evaluate their own performance (Dunning, 2005).

The terms 'intelligence' and 'competency' both refer to people's capacity to successfully carry out certain tasks, but the two concepts differ somewhat in their emphasis. Intelligence refers centrally to a general ability to think – a person's ability to function well in defined, broad area of problem-solving (Carroll, 1993). Competencies, on the other hand, refer to attained, demonstrable levels of performance in a specific area, often as judged against a socially defined criterion. In the case of personal intelligence, aptitude would be assessed by asking test-takers to solve a number of general problems believed to reflect their ability in the domain, and comparing their answers to those of other people. By contrast, competencies are more often assessed in a specific context. For example, a human resources professional who selects candidates for various positions needs to be competent to do so – meeting or exceeding the organization's standards for correctly placing individuals in their jobs. That specific performance in a given

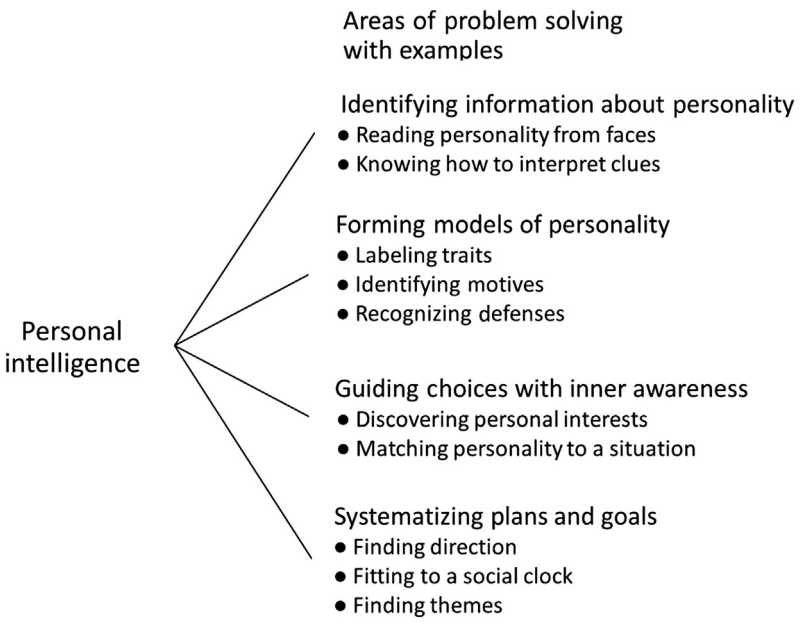

Figure 1 An overview of personal intelligence. Personal intelligence can be thought of as drawing on four broad areas of problem-solving depicted in the figure. By carrying out reasoning across the areas, people are better able to understand and anticipate their own and other people's actions. Illustration adapted from Personal Intelligence by John Mayer. Copyright © 2014 by John Mayer. Used by permission of Farrar, Straus and Giroux, LLC.

context is a competency. Most attention to personal intelligence so far has been on measuring the intellectual ability.

Two Classes of Ability Scales

Ability scales or tests of personal intelligence can be divided into two classes: those that focus on a specific ability or context, and those that measure the overall intelligence. Many of the first specific measures were not designed with personal intelligence in mind, but plainly tap one or another part of the concept. Existing specific measures fall into areas of being 'a good judge' in relation to evaluating other people's characteristics accurately, and of understanding traits. Another relevant group of measures, arguably, are ability measures of emotional intelligence, because understanding emotions are part of understanding personality. Ability measures of emotional intelligence are not reviewed here as the Encyclopedia contains a separate article on emotional intelligence.

Specific Area Scales

This section examines several measures of specific areas of personal intelligence.

- The Good Judge: The key set of studies of being a good judge involve, not an ability scale, but rather a laboratory index of performance that examines how well participants assess the personalities of other people in their groups. For example, Letzring (2008) examined the best judges among participants in the Riverside Accuracy Study, conducted at the University of California, Riverside. Letzring indexed accuracy according to how well a judge's perception of a person matched several criteria including the participant's self-description, other participants' impressions of the target person, and clinical interviewers' impressions. The best judges were high in social skills, agreeableness, and positive adjustment – as independently rated by undergraduate coders from videotapes of the judges' interactions in small groups.
- Measurement Scales of Good Judges: Moving to actual measurement scales, Vogt and Colvin (2005) collected ratings of the personalities of 143 undergraduates from their parents, observers, and behavioral ratings by judges who viewed them on videotape. They developed indices of how closely participants evaluated themselves relative to each of those observers and demonstrated moderate correlations among the different indices. Overall accurate self-knowledge (i.e., smaller discrepancies) correlated with well-being.
- Trait Understanding: Christiansen et al. (2005) developed a measure of 'dispositional intelligence,' a test-taker's knowledge of traits including how traits relate to behaviors and to situations. For example, the following item from their test asks about trait–behavior relationships:

> Lucy's coworkers all describe her as efficient, thorough, and persistent.
> Most likely Lucy also
> A. feels the need to be around a lot of people
> B. has a great deal of sympathy for those less fortunate
> C. does not often give in to her impulses
> D. enjoys fantasizing and daydreaming (Christiansen et al., 2005: p. 131).

The researchers identified "C. does not often give in…" as the correct answer because self-control is related to the overall trait of conscientiousness, about which the item inquires. As a second example from their test, an item that related traits and situations asked:

Which of the following situations are most relevant to the trait of sociability?

A. A week after taking a final exam, you go to the professor's office to find out your final grade and you run into a classmate there. While you both are waiting for your grades, your classmate tells you he found the course difficult and is concerned about his performance.
B. You have just heard that your supervisor received a promotion that he/she has wanted for a long time.
C. Over the last 2 years, you have been employed at a job that entails working by yourself. Your boss offers you a chance to do essentially the same thing, but in a group of coworkers (Christiansen et al., 2005: p. 148).

Here, the authors identified "C. Over … in a group of coworkers" as the correct answer, because the new job directly impacts the ability to fulfill one's need for sociability. In Christiansen et al.'s research, an expert panel of judges (two graduate students and three doctoral-level psychologists) chose the correct answers to each of the 45 questions and exhibited high levels of agreement in doing so. The authors reported that the finished scale of 45 items exhibited a reliability of $r = 0.82$ for the main sample of 157 participants in their larger study. Test-takers who scored high were able to better predict, relative to low-scorers, the personality attributes of an interview candidate who had been videotaped.

General Scales of Personal Intelligence

The Test of Personal Intelligence (TOPI) was designed as a general scale of personal intelligence that measures a broad range of problem-solving abilities in the area. The TOPI asks questions that tap the labeling of inner states such as fatigue and dreams, and understanding one's reputation; it asks about trait knowledge using items with formats similar to those on the dispositional intelligence scale, and it asks questions about planning and goal setting, among other areas. The most recent version of the scale, the TOPI 1.2, contains 16 different item clusters, each asking about a different problem-solving area of personal intelligence (Mayer et al., 2012).

Omnibus scales of personal intelligence such as the TOPI 1.2 serve as *de facto* experiments in whether the construct exists. If personal intelligence is indeed a broad intelligence, then performance on the wide range of problem-solving measured by the scale ought to be reasonably highly correlated and represented well as a single global factor with subsidiary areas. Mayer, Panter, and Caruso were able to model the scale in that fashion using confirmatory factor analysis (Mayer et al., 2012). Reliability for the 134 item TOPI 1.2 was reported at 0.89, although in a recent unpublished study its reliability was 0.83 (Mayer et al., 2012). A short form, the TOPI MINI-12 has reliability variously calculated between $r = 0.52$ and 0.82 depending upon the sample, and correlates reasonably well with the overall TOPI 1.2 (Mayer et al., 2013).

Competencies of Personal Intelligence

People who develop their personal intelligence employ a wide range of competencies to understand themselves and those around them, as reflected by the tasks of the specific and general measures of the abilities related to the intelligence. A full-length review of many of these specific skills is available elsewhere (Mayer, 2014). Here, it is worth pointing out that people may use their personal intelligence to develop competencies that are specifically related to their individual needs. For example, a high-school student hoping to understand his own and his friends' behaviors may seek to focus, for example, on skills at obtaining feedback, on eliciting information about people, and on integrating information into a balanced picture of himself and others (avoiding naive optimism or an overcritical attitude). A new employee training in a human resources department of an organization, on the other hand, may want to hone both her interviewing skills and ability to make predictions about an applicant's performance on the job. And, a psychotherapist will want to build on his skills of empathy and setting limits. The new university president who previously was uninterested in college athletics may realize that he should attend student rallies clamoring for the popular basketball coach to remain at the university rather than accepting a rival school's job offer. Although there are shared factors across these contexts – the person's overall personal intelligence – there are also unique motivational and training requirements in each case so as to reach a level of competency.

The Relation of Personal Intelligence to Assorted Criteria

Relation to Socioemotional Traits, Interests, and Other Self-Evaluations

Self-Estimated Personal Intelligence and Psychological Mindedness

A growing number of findings have related personal intelligence to a variety of criteria. For example, intelligences are generally uncorrelated with people's estimates of their own mental abilities, and personal intelligence is apparently no exception to the rule. Using an online sample, Moore (2012) reported a correlation consistent with past research – of just $r = 0.11$ – between a measure of self-estimated personal intelligence and the TOPI MINI-12 (Only a portion of the TOPI MINI and SEPI were used. See Moore, 2012). People with greater self-reported interpersonal competency on the Interpersonal Competency Scale (Buhrmester et al., 1988) likewise show no significant relationship with the TOPI 1.2 overall (Mayer et al., 2012). Personal intelligence is, however, related $r = 0.30$–0.40 to self-reported psychological mindedness (Mayer et al., 2012), where such psychological mindedness represents a personal interest in understanding personality coupled with a belief that such understanding is of importance (Conte et al., 1990).

Measures of Normal Personality Traits, Work Behavior, and Clinical Symptoms

Studies with the TOPI 1.2 and TOPI MINI indicate that personal intelligence has low positive correlations – roughly between $r = 0.10$ and 0.20 – with three of the big five personality traits: Openness, Conscientious, and Agreeableness (Mayer et al., 2013; Mayer et al., 2012). A study with the TOPI MINI-12 and the Counterproductive Work Behavior Checklist (Spector et al.,

2006) found that those lower in personal intelligence engaged in significantly more abusive relations ($r = -0.16$), sabotaging behaviors ($r = -23$), and thefts at work ($r = -0.15$), all $p < 0.05$ (two-tailed), using a sample of 336–379 anonymous online responders (Moore, 2012). And in a study using measures of psychiatric symptoms, people who scored lower on the full TOPI 1.2 also exhibited more symptoms of narcissism and dependent personality (Mayer et al., 2012).

Relationship to Mental Abilities

The TOPI 1.2 correlates with verbal intelligence at about the $r = 0.40$ level. Although the TOPI asks mostly about cues to personality, traits, motives, and goals, and has just a few items concerning emotion, the TOPI 1.2 correlated $r = 0.68$ with the strategic areas (understanding and managing emotions) of the Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT). In the same study, the TOPI 1.2 also correlated $r = 0.53$ with the Reading the Mind in the Eyes Test – a measure of skill at inferring inner states from the eyes, inversely related to autism (Mayer et al., 2012).

People who score high on Christiansen et al.'s (2005) scale of dispositional intelligence are better able than low scorers at identifying how job applicants in a video described their personalities. Finally, people who draw accurate conclusions from reading paragraph-length descriptions of a person's behavior, as opposed to misreading the situation or drawing black-and-white or biased perceptions, also score higher on the TOPI MINI-12 (Allen and Mayer, in preparation).

Discussion and Conclusion

The a priori theoretical argument for personal intelligence is evolutionary and adaptive: people who developed skills at understanding their own and others' personalities had an adaptive advantage over others, and as a consequence, our evolutionary ancestors developed a capacity to reason in this domain.

Personal intelligence is, relatively speaking, a newly introduced intelligence. In the principal theory we discuss, it divides into a number of problem-solving areas from identifying personality-related information to systematizing life plans. Psychologists have developed measures of some of the specific abilities involved, such as knowledge of traits, and of the overall intelligence (Christiansen et al., 2005; Mayer et al., 2012). They also have developed scales that measure many different kinds of reasoning in the area. Factor analyses indicate that these specific areas of problem-solving form a unitary area of ability. Measures of personal intelligence are related to measures of verbal intelligence at a moderate range, and somewhat higher with 'hot' or 'felt' intelligences (depending upon one's preferred terminology), including the MSCEIT and the Reading the Mind in the Eyes Test (Baron-Cohen et al., 2001). Preliminary findings indicate that personal intelligence has promise for predicting important outcomes in people's lives. That said, much research remains to be done to further develop the theory, demonstrate that it exists independently of general and hot intelligences, and to improve the measures of this ability.

See also: Emotional Intelligence and Competencies; Personality Assessment: Overview; Personality: Historical and Conceptual Perspectives; Self-Concept: From Unidimensional to Multidimensional and Beyond; Social Intelligence and Competencies.

Bibliography

Abelson, R.P., 1963. Computer simulations of "hot cognition." In: Tompkins, S.S., Messick, S. (Eds.), Computer Simulations of Personality. Wiley, New York, pp. 277–302.

Allen, J., Mayer, J.D. Learning about a person: a functional view, in preparation.

Appelbaum, S.A., 1973. Psychological-mindedness: word, concept and essence. The International Journal of Psychoanalysis 54 (1), 35–46.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., Plumb, I., 2001. The 'reading the mind in the eyes' test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. Journal of Child Psychology and Psychiatry 42 (2), 241–251. http://dx.doi.org/10.1111/1469-7610.00715.

Blanchette, I., Richards, A., 2010. The influence of affect on higher level cognition: a review of research on interpretation, judgement, decision making and reasoning. Cognition and Emotion 24 (4), 561–595. http://dx.doi.org/10.1080/02699930903132496.

Bower, G.H., 1981. Mood and memory. American Psychologist 36 (2), 129–148. http://dx.doi.org/10.1037/0003-066X.36.2.129.

Buhrmester, D., Furman, W., Wittenberg, M.T., Reis, H.T., 1988. Five domains of interpersonal competence in peer relationships. Journal of Personality and Social Psychology 55 (6), 991–1008. http://dx.doi.org/10.1037/0022-3514.55.6.991.

Buss, D.M., 2008. Human nature and individual differences: evolution of human personality. In: Pervin, L.A. (Ed.), Handbook of Personality Psychology: Theory and Research, third ed. Guilford Press, New York, USA, pp. 29–60.

Carroll, J.B., 1993. Human Cognitive Abilities: A Survey of Factor-analytic Studies. Cambridge University Press, New York, USA.

Christiansen, N.D., Wolcott-Burnam, S., Janovics, J.E., Burns, G.N., Quirk, S.W., 2005. The good judge revisited: individual differences in the accuracy of personality judgments. Human Performance 18 (2), 123–149. http://dx.doi.org/10.1207/s15327043hup1802_2.

Conte, H.R., Plutchik, R., Jung, B.B., Picard, S., 1990. Psychological mindedness as a predictor of psychotherapy outcome: a preliminary report. Comprehensive Psychiatry 31 (5), 426–431. http://dx.doi.org/10.1016/0010-440X(90)90027-P.

Dunbar, R.I.M., 2009. The social brain hypothesis and its implications for social evolution. Annals of Human Biology 36 (5), 562–572. http://dx.doi.org/10.1080/03014460902960289.

Dunning, D., 2005. Self-Insight: Roadblocks and Detours on the Path to Knowing Thyself. Psychology Press, New York, USA.

Ellenberger, H., 1956. The ancestry of dynamic psychotherapy. Bulletin of the Menninger Clinic 20, 288–299.

Erikson, E.H., 1950. Childhood and Society. W.W. Norton, New York.

Funder, D.C., 2001. Accuracy in personality judgment: research and theory concerning an obvious question. In: Hogan, R. (Ed.), Personality Psychology in the Workplace. American Psychological Association, Washington, DC, USA, pp. 121–140. http://dx.doi.org/10.1037/10434-005.

Gardner, H., 1983. Frames of Mind: The Theory of Multiple Intelligences. Basic Books, New York, USA.

Gardner, H., 1999. Intelligence Reframed. Basic Books, New York.

Haselton, M.G., Funder, D.C., 2006. The evolution of accuracy and bias in social judgment. In: Kenrick, D.T. (Ed.), Evolution and Social Psychology. Psychosocial Press, Madison, CT, USA, pp. 15–37.

Hunt, E., 2010. Intelligence. Cambridge University Press, New York.

Isen, A.M., Shalker, T.E., Clark, M., Karp, L., 1978. Affect, accessibility of material in memory, and behavior: a cognitive loop? Journal of Personality and Social Psychology 36 (1), 1–12. http://dx.doi.org/10.1037/0022-3514.36.1.1.

Jackson, D.N., 1972. A model for inferential accuracy. Canadian Psychologist/Psychologie Canadienne 13 (3), 185–195. http://dx.doi.org/10.1037/h0082183.

Kantrowitz, B., Peterson, H., Wingert, P., October 24, 2005. How I got there. Newsweek 146, 48–62.

Klein, S.B., Cosmides, L., Murray, E.R., Tooby, J., 2004. On the acquisition of knowledge about personality traits: does learning about the self engage different mechanisms than learning about others. Social Cognition 22 (4), 367–390. http://dx.doi.org/10.1521/soco.22.4.367.38295.

Letzring, T.D., 2008. The good judge of personality: characteristics, behaviors, and observer accuracy. Journal of Research in Personality 42 (4), 914–932. http://dx.doi.org/10.1016/j.jrp.2007.12.003.

Locke, E.A., 2005. Why emotional intelligence is an invalid concept. Journal of Organizational Behavior 26 (4), 425–431. http://dx.doi.org/10.1002/job.318.

Maul, A., 2011. The factor structure and cross-test convergence of the Mayer–Salovey–Caruso model of emotional intelligence. Personality and Individual Differences 50 (4), 457–463. http://dx.doi.org/10.1016/j.paid.2010.11.007.

Max, T., 2009. I Hope They Serve Beer in Hell. Citadel, New York.

Mayer, J.D., Panter, A.T., Caruso, D.R., 2013. Test of Personal Intelligence MINI MARKER-12 (TOPI MINI-12), second ed. Brief manual, Technical report. http://www.unh.edu/personalitylab/Assets/measures-PI/TOPI%20MINI%2012%20Manual%202nd%20ed%202013-07-09.pdf.

Mayer, J.D., Panter, A.T., Caruso, D.R. Test of Personal Intelligence MINI MARKER-12 (TOPI MINI-12), second ed. Brief manual, unpublished manuscript.

Mayer, J.D., 2005. A tale of two visions: can a new view of personality help integrate psychology? American Psychologist 60 (4), 294–307. http://dx.doi.org/10.1037/0003-066X.60.4.294.

Mayer, J.D., 2007. Asserting the definition of personality. P: The Online Newsletter for Personality Science 1 (1), 8. June 2012.

Mayer, J.D., 2008. Personal intelligence. Imagination, Cognition and Personality 27 (3), 209–232. http://dx.doi.org/10.2190/IC.27.3.b.

Mayer, J.D., 2009. Personal intelligence expressed: a theoretical analysis. Review of General Psychology 13 (1), 46–58. http://dx.doi.org/10.1037/a0014229.

Mayer, J.D., 2014. Personal Intelligence: The Power of Personality and How It Shapes Our Lives. Scientific American/Farrar Strauss & Giroux, New York.

Mayer, J.D., Korogodsky, M., 2011. A really big picture of personality. Social and Personality Psychology Compass 5, 104–117. http://dx.doi.org/10.1111/j.1751-9004.2010.00336.x.

Mayer, J.D., Lin, S.C., Korogodsky, M., 2011. Exploring the universality of personality judgments: evidence from the great transformation (1000 BCE–200 BCE). Review of General Psychology 15, 65–76. http://dx.doi.org/10.1037/a0020711.

Mayer, J.D., Panter, A.T., Caruso, D.R., 2012. Does personal intelligence exist? Evidence from a new ability-based measure. Journal of Personality Assessment 94, 124–140. http://dx.doi.org/10.1080/00223891.2011.646108.

McAdams, D.P., Pals, J.L., 2006. A new Big Five: fundamental principles for an integrative science of personality. American Psychologist 61 (3), 204–217. http://dx.doi.org/10.1037/0003-066X.61.3.204.

McGrew, K.S., 2009. CHC theory and the human cognitive abilities project: standing on the shoulders of the giants of psychometric intelligence research. Intelligence 37 (1), 1–10. http://dx.doi.org/10.1016/j.intell.2008.08.004.

Moore, M., 2012. Personal Intelligence in the Workplace and Relationships. University of North Carolina at Chapel Hill, unpublished B.S. thesis with Honors in Psychology.

Park, L.C., Park, T.J., 1997. Personal Intelligence. In: McCallum, M., Piper, W.E. (Eds.). Lawrence Erlbaum Associates Publishers, Mahwah, NJ, USA, pp. 133–167.

Pillemer, D.B., 1998. Momentous Events, Vivid Memories. Harvard University Press, Cambridge, MA, USA.

Ptahhotep, 2003/c.2450 BCE. The maxims of Ptahhotep. In: Simpson, W.K. (Ed.), The Literature of Ancient Egypt (V.A. Toben, Trans.). Yale University Press, New Haven, pp. 129–148.

Roback, A.A., 1927. The Psychology of Character. Kegan Paul, London, England.

Salovey, P., Mayer, J.D., 1990. Emotional intelligence. Imagination, Cognition and Personality 9 (3), 185–211.

Spearman, C., 1904. 'General intelligence,' objectively determined and measured. The American Journal of Psychology 15 (2), 201–293. http://dx.doi.org/10.2307/1412107.

Spector, P.E., Fox, S., Penney, L.M., Bruursema, K., Goh, A., Kessler, S., 2006. The dimensionality of counterproductivity: are all counterproductive behaviors created equal? Journal of Vocational Behavior 68 (3), 446–460. http://dx.doi.org/10.1016/j.jvb.2005.10.005.

Sternberg, R.J., Grigorenko, E.L., 2002. In: Sternberg, R.J., Grigorenko, E.L. (Eds.), The General Factor of Intelligence: How General Is It?. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, USA.

Theophrastus, 1970/c.319 BCE. Theophrastus Characters. Kent State University Press, Kent State (W. Anderson, Trans.).

Vaillant, G.E., 2000. Adaptive mental mechanisms: their role in a positive psychology. American Psychologist 55 (1), 89–98. http://dx.doi.org/10.1037/0003-066X.55.1.89.

Vogt, D.S., Colvin, s R., 2005. Assessment of accurate self-knowledge. Journal of Personality Assessment 84 (3), 239–251. http://dx.doi.org/10.1207/s15327752jpa8403_03.

Wedeck, J., 1947. The relationship between personality and 'psychological ability.'. British Journal of Psychology 37, 133–151.

Zeidner, M., Matthews, G., Roberts, R.D., 2001. Slow down, you move too fast: emotional intelligence remains an 'elusive' intelligence. Emotion 1 (3), 265–275. http://dx.doi.org/10.1037/1528-3542.1.3.265.

Relevant Websites

http://www.intelltheory.com/index.shtml – Human Intelligence.

http://www.unh.edu/personalitylab/lab-theoretical-work/xpersintell.htm – The Personality Laboratory at the *University of New Hampshire.*

Personal Projects

Brian R Little, University of Cambridge, Cambridge, UK and Carleton University, Ottawa, ON, Canada
Simon Coulombe, Université du Québec à Montréal, Montréal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Personal projects are extended sets of personally salient action in context. They range from the incidental tasks of typical Thursdays to the overarching commitments of a lifetime. Little (1983) provided a rationale and methodology for the study of personal projects with emphasis on their consequences for well-being and human flourishing. Personal projects analysis (PPA) is a multimodular, flexible assessment suite based on 12 explicit measurement criteria. PPA provides modules for assessing the content, appraisal, impact, dynamics, and social ecology of personal projects. Research broadly confirms the proposition that well-being comprises the sustainable pursuit of core projects in our lives.

Personal projects are constituent elements of daily lives. They can range from the seemingly inconsequential acts of Thursday mornings ('put out the dog,' 'trim my toenails') to the overarching commitments of our lives ('help my mother through Alzheimer's,' 'liberate my people'). Little (1983) introduced a rationale and methodology for the study of personal projects, in part as a contribution to the person–environment debate that was raging between personality and social psychology at the time (e.g., Mischel, 1968). The methodology, personal projects analysis (PPA), has been subsequently adopted in various fields of social and behavioral sciences, from psychology (e.g., personality, social, developmental, health, environmental, organizational) to occupational therapy, rehabilitation science, environmental planning, gerontology, management, and public policy (Little et al., 2007).

Little introduced personal projects as units of analysis that uniquely capture features of both individuals and the contexts in which they act. Personal projects can be formally defined as "extended sets of personally salient action in context" (Little, 2007, p. 25). Each term can be delineated as follows:

Extended: A project is not a momentary act, but typically a sequence that is extended in space and time (minutes, hours, days, or years).

Set: A project encompasses a series of actions that are considered interrelated by the project pursuer.

Personally salient: A project is a noteworthy aspect of the life of the person who pursues it, and can be defined as such only by that person.

Action: A project is not a passive response to external forces but an intentional sequence of behavior involving cognitive, affective, conative, and volitional processes.

In context: A project is enacted in physical, social, cultural, and temporal contexts that can facilitate or thwart project pursuit.

The Social Ecology of Personal Project Pursuit

From the outset, personal projects research explored the impact of project pursuit on well-being and the quality of lives (Little and Ryan, 1979; Palys and Little, 1983). It was a central theoretical element of a social ecological model of human development, depicted in **Figure 1** (Little, 1999). The model posits that well-being is influenced by stable and dynamic features of both persons and their contexts and that each of these sources of influence is integrated in the personal projects individuals pursue.

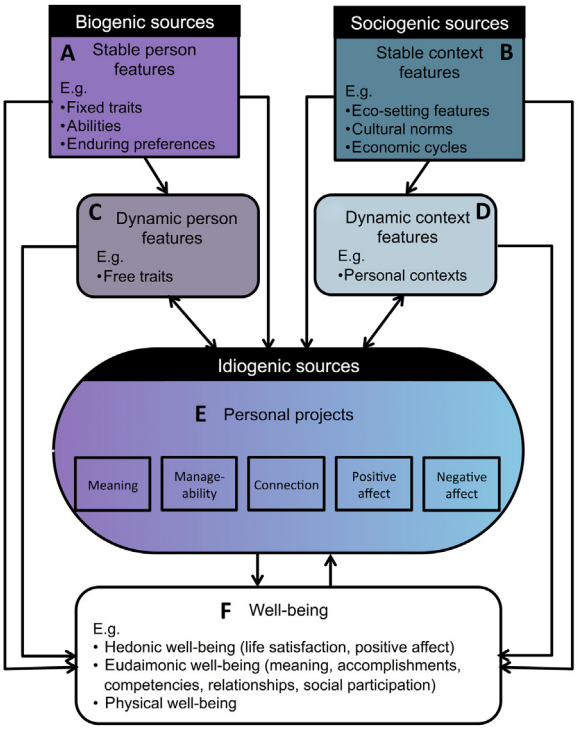

Note. The saturation of colors refers to the degree of fixedness of the element. Greater saturation (more intense color) indicates that the element is more resistant to change; lower saturation (less intense color) indicates that the element is more flexible. The colors blend together in the *Personal projects* box to symbolize that individual (purple) and contextual (turquoise) features merge through personal projects.

Figure 1 Social ecological model of well-being (Little, B.R., 1999. Personal projects and social ecology: themes and variations across the life span. In: Brandtstädter, J., Lerner, R.M. (Eds.), Action & Self-Development: Theory and Research through the Life Span. Sage Publications, Thousand Oaks, CA, pp. 197–221).

Personal projects share features with other personal action constructs (PAC units) that arose in the 1970s and 1980s (Little et al., 1992), and differ from traditional units of analysis in personality psychology such as traits or narratives. Among PAC units, personal projects occupy a middle-level position, with personal strivings and related goal units focusing mainly on the internal motivation behind actions, and life tasks being more externally oriented, focusing on age-related normative expectations. A detailed account of the development of personal project theory and how personal projects differ from other constructs such as personal goals appears in Little (2007).

PPA Methodology

PPA is a multimodular instrument designed to measure personal projects in research, clinical, and other applied settings. PPA is typically presented in written form. A computer-enhanced version has been developed (see the Excel spreadsheet in Appendix). PPA can also be administered through an oral interview (Little and Gee, 2007) and this is particularly valuable for use in applied settings.

Measurement Criteria

PPA complies with 12 measurement criteria that are regarded as a propaedeutic to more conventional psychometric criteria (Little, 2000b). These criteria are presented in **Table 1**.

These 12 criteria are implemented in PPA through four standard modules: Project Elicitation, Project Appraisal, Hierarchical Analysis, and Cross-Impact Matrices.

Project Elicitation

Individuals are presented with a description and examples of personal projects, and asked to list their personal projects in 10–15 min. As an illustration, **Figure 2** shows the spreadsheet form completed by Peter, a graduate student in psychology. This module generates information about the number and content of a person's projects and about the idiosyncratic language through which they are expressed. Typically participants list around 15 personal projects although this will vary as a function of such factors as age and openness (Lawton et al., 2002; Little and Gee, 2007; Little et al., 1992).

The listed projects can be analyzed by traditional content analytic procedures. The standard categorization framework shows high interrater reliability and comprises eight main categories based on domains of activity: interpersonal (e.g., phone my parents more often), academic or work (e.g., complete my PhD), intrapersonal (e.g., become a better person), recreational/leisure (e.g., have fun with my friends this week-end), health (e.g., lose 15 pounds), maintenance (e.g., organize my time to work less during week-ends), and others. The most frequent categories are interpersonal and academic or work projects (depending on whether the participants are students or in the work force). Participants themselves can also be asked to classify their projects into these categories and this allows for greater precision (e.g., "work on Suzanne's car" might be primarily a work project or an interpersonal one). Project categories are related to well-being outcomes.

Table 1 Methodological criteria underlying personal projects analysis

Criteria	*Description: PPA…*
Reflexive	Provides a theoretical framework that is jointly applicable to the scientist's actions and those of the participants being studied.
Personally salient	Focuses on noteworthy aspects of the participants' existence, as defined by themselves and expressed in their idiosyncratic language.
Evocative	Gives participants an opportunity for self-reflection and new insights about themselves.
Ecologically representative	Documents the influence of objective and subjective features (e.g., physical, social) of the participants' ecosystem.
Temporally sensitive	Documents the dynamic nature of personal projects through their stages of progression.
Social indicator potential	Provides policy-relevant data on the quality of the social ecological settings where the participants evolve.
Systemic	Documents the interrelations between the multiple projects pursued concomitantly by participants, and which, aggregated, comprise their personal project systems.
Middle-level	Provides access to both lower order (e.g., acts) and higher order (e.g., values and aspirations) aspects of human experience.
Modular	Provides modules that can be selected, adapted, and supplemented to offer the best answer to the study question, in the specific research setting.
Conjointly idiographic and normative	Allows analysis of relationships between projects dimensions at the participant level (across each participant's different projects) and at the sample level (across all the participants' project systems).
Integrative	Provides a single instrument to document cognitive, behavioral, affective, and social aspects of the participants' experience, integrated around their personal projects.
Directly applicable	Focuses on a tractable unit of analysis that can be targeted in individual and social interventions to promote well-being.

Note. Content adapted from Little, B.R., 2000b. Persons, contexts, and personal projects: assumptive themes of a methodological transactionalism. In: Wapner, S., Demick, J., Yamamoto, T., Minami, H. (Eds.), Theoretical Perspectives in Environment-Behavior Research: Underlying Assumptions, Research Problems, and Methodologies. Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 79–88.

For example, in a sample of seniors (Lawton et al., 2002), recreational/leisure projects were positively associated with positive affect, and interpersonal projects were associated positively with positive affect and negatively with depression. Salmela-Aro (1992) has shown that students seeking psychological counseling frequently report intrapersonal projects, which she found negatively associated with sense of coherence, self-esteem, mental health, and life satisfaction. Given the preponderance of intrapersonal projects (4 out of 15) in Peter's list, we might expect him to experience limited levels of well-being. However, this might be compensated by the positive affect that he derives from his three interpersonal projects and his two recreational/leisure projects.

Project elicitation

To start, please take 10–15 minutes and type in the following cells as many personal projects and activities you can that you are currently engaged in or considering – remember these need not be formal projects or even important ones – we would prefer you to give us more of the everyday kinds of activities or concerns that characterize your life at present.

Please be completely honest in your answers, as they will not be connected to your name. Also, if possible, please fill out this survey in one sitting.

Project #	Project description
1	Become a better person
2	Complete my PhD
3	Exercise 4 times a week
4	Move in with Emily
5	Be less anxious
6	Lose 15 pounds
7	Have fun with my friends this week-end
8	Go at the beach for a couple of days
9	Organize my time to work less during week-ends
10	Walk the Way of St.James
11	Phone my parents more often
12	Be happy
13	Write the first part of my dissertation
14	Write that research report that has been due for one month
15	Adopt a child

Figure 2 Project elicitation module.

Linguistic aspects of project content have also been analyzed. For example, Chambers (2007) presents a convincing argument that the language we use to communicate about our projects is strongly related to their likelihood of successful pursuit. Phrasing projects as 'tryings' or as things to avoid for example was found to be related to lower well-being (Chambers, 2007).

Project Appraisal

After having elicited their projects, participants are asked to select i projects ($i =$ usually 10) that are more important for understanding them. Then, they appraise each project on a scale from 0 to 10 along j dimensions, thus filling in an $i \times j$ matrix.

Standard Dimensions

At the time of PPA publication (Little, 1983), the matrix comprised 17 standard dimensions in addition to two open-ended columns asking where and with whom the project is carried out. These dimensions were selected to cover and differentiate between a range of major theoretical constructs in psychology and the social sciences. In the 1990s, with the growing interest in affective processes in psychological research, a set of affective dimensions was added to explore the feelings that peoples experience while doing or thinking about their projects (Little and Gee, 2007). Consequently, the current version of PPA comprises two appraisal matrixes, the first comprising 17 primarily cognitive or cooler dimensions and the second nine affective or hotter dimensions (see the matrixes in **Figures 3** and **4** for the complete list of dimensions, and example data from Peter's appraisal of his projects).

Based on theoretical postulations and empirical factor analysis, five themes have emerged from these dimensions: (1) meaning (e.g., importance, self-identity, value congruency), (2) manageability (e.g., control, likelihood of successful completion, progress), (3) connection (e.g., visibility, importance as viewed by others, support from others), (4) positive affect (e.g., enjoyment, hope, love), and (5) negative affect (e.g., stress, angry, depressed). Positive and negative measures of well-being (e.g., respectively life satisfaction and depression) have been consistently associated with these project themes (Little and Gee, 2007; see also Little, 1989, 2011). People experience more well-being when their projects are more meaningful, manageable, positively connected with others, and associated with more positive affect relative to negative affect.

Ad Hoc Dimensions

In the first publication about the method, Little (1983) stipulated that PPA is designed to be a flexible measurement approach. Thus, standard dimensions should be complemented by dimensions that are pertinent for the specific group of persons studied and their particular eco-setting. These ad hoc dimensions refer to aspects of the individuals and their transactions with their social ecological contexts. For example, researchers have asked participants to appraise their projects in terms of flow (Omodei and Wearing, 1990), organizational support (Little et al., 2007), and supportiveness of neighborhood environments (Sugiyama

Matrix 1: What do you think about what you are doing?

Please rate each project below from 0 – 10 on the series of dimensions listed above them.

For example, Michelle might rate her project 'Get the car muffler replaced' as an 8 on importance, 3 on difficulty, 5 on visibility, 7 on control, 9 on responsibility, 2 on time adequacy and so on.

	Your projects:	Importance	Difficulty	Visibility	Control	Responsibility	Time Adequacy	Outcome / Likelihood of Success	Self Identity	Other's View	Value Congruency	Progress	Challenge	Absorption	Support	Competence	Autonomy	Stage
1	Become a better person	8	8	5	7	10	3	7	7	5	9	7	8	3	7	7	9	6
2	Complete my PhD	10	7	10	8	10	8	10	10	7	10	8	9	10	10	8	10	7
3	Move in with Emily	9	6	6	7	7	4	10	8	5	10	5	6	5	7	9	8	4
4	Be less anxious	10	8	7	7	10	3	7	7	8	8	9	9	3	10	6	10	7
5	Lose 15 pounds	9	8	6	7	10	3	6	7	9	7	4	8	3	7	6	8	3
6	Have fun with my friends this week-end	8	7	9	5	6	2	7	10	9	10	7	9	5	9	7	9	7
7	Walk the Way of St.James	5	9	0	5	8	5	3	3	5	7	0	10	2	5	3	10	1
8	Phone my parents more often	8	5	0	10	10	0	6	7	7	6	2	5	2	3	5	6	2
9	Be happy	10	10	7	5	10	7	7	8	9	9	6	10	3	10	5	10	7
10	Adopt a child	6	8	5	6	6	5	7	6	5	8	1	9	2	5	3	9	1

Figure 3 Project appraisal matrix (cognitive dimensions).

Matrix 2: How do you feel about what you are doing?

Please rate from 0 to 10 the extent to which you feel each emotion while engaged or thinking about each project.

Use 10 if you experience the emotion very strongly, and 0 if you don't feel it at all.

In the 'Other Emotion' column, you have the opportunity to write in any specific emotion that you feel characterizes your project, but may not have been mentioned.

	Your projects:	Sad	Fearful / Scared	Full of Love	Angry	Happy / with Enjoyment	Hopeful	Stressed	Uncertain	Depressed	Other Emotion
1	Become a better person	3	6	7	4	7	7	7	7	4	
2	Complete my PhD	2	7	7	4	8	9	9	2	3	
3	Move in with Emily	0	3	10	0	9	10	5	1	0	eager
4	Be less anxious	1	7	2	3	8	7	7	4	3	
5	Lose 15 pounds	0	2	0	4	8	7	4	5	2	energized
6	Have fun with my friends this week-end	0	0	7	0	8	7	5	4	0	excited
7	Walk the Way of St.James	0	10	2	0	6	3	9	9	0	confused
8	Phone my parents more often	0	0	8	3	6	5	1	2	0	
9	Be happy	3	6	8	3	9	7	7	7	2	
10	Adopt a child	0	10	8	0	7	7	9	8	0	

Figure 4 Project appraisal matrix (affective dimensions).

and Ward Thompson, 2007). Results from multiple studies suggest that ad hoc dimensions often emerge as key and distinctive predictors of well-being (Little and Gee, 2007). For example, in a sample of Southeast Asian immigrants to Canada, the absence of language difficulties posed by their personal projects was an excellent predictor of well-being (see Little and Gee, 2007).

Once the appraisal module has been completed, person-centered (idiographic) analyses can be performed. Within a single case, it is possible to compute the correlations

between two appraisal dimensions across the person's projects. Alternatively, scores from the appraisal matrix can be averaged across projects in order to provide indices reflecting the global state of the person's project system on each dimension. The mean scores are then used in variable-centered (normative) analyses to study how between-person variation at the level of the project system is associated with variation in other characteristics (e.g., sociodemographic characteristics, personality traits, well-being, etc.). Hierarchical statistical procedures can also be applied to PPA data (e.g., Nurmi et al., 2009) to simultaneously consider both within- and between-persons variation and it is likely that such an analytic strategy will become widely adopted.

Extrapolation of research findings from one level to the other must be made with caution because the results obtained at one level might not be true at the other level, as a particular case of Simpson's paradox. For example, data from Peter's example (Figures 3 and 4) reveal a strong positive correlation ($r = 0.93$) between the dimensions *Responsibility* and *Angry*. In contrast, data from a group of participants ($N = 184$) in research conducted in Little's laboratory, analyzed at the normative level, reveal a modest negative correlation ($r = -0.22$). Interestingly, Peter's means on these dimensions (respectively 8.70 and 2.10 for *Responsibility* and *Angry*) are relatively similar to the group means (8.45 and 1.68). Such diverging and complex results might be of particular interest for a clinical psychologist or a counselor who wants to stimulate discussions with Peter. However, on the whole, examination of individual and normative project space suggests that there is a high degree of isomorphism between idiographic and nomothetic project spaces (see Little, 2000b).

Other Modules: Cross-Impact and Hierarchical Analysis Modules

People are typically engaged in multiple projects, particularly if the focus is on daily pursuits rather than hourly activity, and these projects may facilitate or frustrate each other. The Cross-Impact Matrix of PPA addresses the relationships between projects. One way of assessing this is to ask participants to indicate if the impact is positive or negative, using plus or minus signs. The relationship between two projects is not necessarily reciprocal as shown in Peter's example in Figure 5. He estimates that working on completing his PhD has a strong positive impact on his project to be happy. However, being happy has a null impact on his PhD completion. The projects to lose 15 pounds and have fun with his friends this week-end have negligible effect on other projects. Only the project to adopt a child seems to frustrate more than one of the other

Matrix 3: How do your projects impact *each other*?

Now for each project separately, starting with project 1 on the left, decide whether the project will have (or is having) a positive (+) very positive (++), negative (-) very negative (- -) or neutral impact (0) on the remaining projects (across the top).

Do the same for project 2, 3 etc. all the way to 10, filling in the blocks on both sides of the diagonal.

Remember, evaluating the impact of project 2 on project 3 is not necessarily the same as evaluating the impact of project 3 on project 2. That is why we want you to complete the entire matrix.

	Your projects:	Become a better person	Complete my PhD	Move in with Emily	Be less anxious	Lose 15 pounds	Have fun with my friends this week-end	Walk the Way of St.James	Phone my parents more often	Be happy	Adopt a child
1	Become a better person		0	0	++	+	0	+	++	++	+
2	Complete my PhD	+		+	++	++	0	+	+	++	++
3	Move in with Emily	+	0		0	-	0	0	0	++	++
4	Be less anxious	++	+	0		++	0	-	+	++	++
5	Lose 15 pounds	0	0	0	+		0	0	0	+	0
6	Have fun with my friends this week-end	0	0	0	+	0		0	0	+	0
7	Walk the Way of St.James	++	0	0	++	+	-		0	++	0
8	Phone my parents more often	++	0	0	+	0	0	0		+	0
9	Be happy	++	0	+	++	+	++	+	+		+
10	Adopt a child	++	0	+	-	-	0	0	++	+	

Figure 5 Cross-impact matrix.

projects according to Peter's point of view. In contrast, Peter's project to complete his PhD is the one that affects the most his other projects, and in a positive way. Little (2007, 2011) suggests that human flourishing is contingent upon the sustainable pursuit of core projects. Core projects are operationalized as projects that are more central to the project system, i.e., those which are strongly linked to other projects and which would have the most impact if abandoned (Little, 2007, 2011). Peter's project to complete his PhD satisfies this criterion and is probably crucial to his well-being.

Taking into account the social ecological context in which personal projects are enacted, the Joint Cross-Impact Matrix (optional) aims to explore the impact of a significant other person's projects on the participant's projects. The matrix follows the same logic as the individual Cross-Impact Matrix. Global indices of positive and negative impact can be computed, with greater negative impact indicating more relational difficulties between the two persons (Little and Gee, 2007).

The last PPA module (optional) focuses on the hierarchy of personal projects. This module derives from the conceptualization of personal projects as middle-level units located in a hierarchy between schedulable acts (at the molecular level) and superordinate goals or values (at the molar level). PPA provides a way of studying the linkage between these three levels by an adaptation of a *laddering* technique used in personal construct research. In Why-laddering, individuals are asked *why* they are engaged in each of their projects. To each of these answers they are again asked why they are engaged in them. This iterative process continues until the answer is regarded as an end in itself – a terminal value. How-laddering proceeds in the same way but asks *how* each project is going to be carried out. Again, for each answer the 'how' question is repeated iteratively until a schedulable act is reached. This module has mainly received interest from clinicians. However, research suggests that the personal project hierarchy could be better represented as a lattice than a ladder, with multiple acts accomplished for a single project and multiple reasons while pursuing it (see Little and Gee, 2007).

Figure 6 presents a diagram that exemplifies the hierarchical structure of Peter's projects, starting with his core project of completing his PhD. Peter is completing his PhD for the terminal reason of being happy. In order to complete his PhD, he will engage in a variety of actions, themselves with multiple ramifications: finish his thesis, do a credited internship, complete his data analysis, submit a scholarship application, etc. Although primarily useful as a clinical or counseling tool, the hierarchical analysis has relevance to broader theoretical issues. For example, it is postulated that projects are sustainable to the extent that they are closer to terminal values and therefore high in a sense of meaning. But it is also postulated that projects are sustainable to the extent that they are closer to schedulable acts and therefore high in manageability. This poses the real possibility of a meaning–manageability trade-off in the pursuit of daily projects (Little, 1989, 2011).

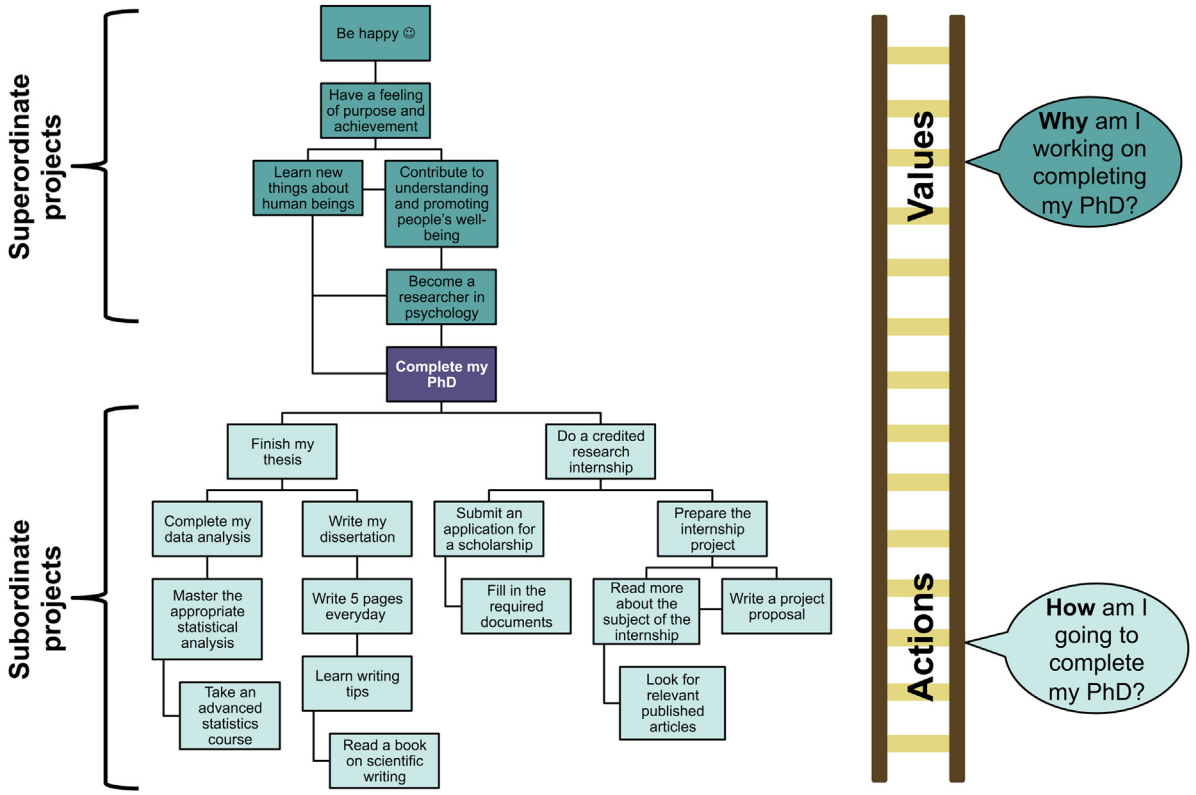

Figure 6 Laddering and latticing diagram.

SEAbank

The Social Ecological Assessment Databank (SEAbank) integrates data collected since the 1970s in the studies conducted by Little and his students and colleagues in the Social Ecological Research Groups at Carleton, Harvard, and Cambridge universities. Data from 22 studies that took place since 1978 are included in the electronic database and these are updated periodically. This represents more than 12 000 personal projects and 3500 research participants from a diversity of groups and eco-settings. Data in SEAbank comprise project content in a participant's own words, project categorization, appraisals on standard PPA dimensions, sociodemographic variables, personality scores (e.g., Big Five dimensions), and diverse measures of well-being. The aggregation of data from multiple studies over several decades offers a unique opportunity to explore such issues as generational effects in project appraisals across decades and meta-analyses on the relation between personal project appraisals and well-being (see Little and Gee, 2007). The following questions are currently being explored using SEAbank: At the system and project levels, do hedonic (i.e., positive affect) and eudaimonic (i.e., meaning) aspects of personal projects represent distinct constructs as recently debated by researchers in the positive psychology movement? In multilevel modeling, when taking all pertinent variables into account (each project category, characteristics of the project system, personality measures, gender, age), which variables best predict project appraisal? The diversity of SEAbank participants' background also permits the investigation of commonalities and specificities in the pursuit of their personal projects.

Psychometric Findings

Psychometric qualities of PPA have been the object of a detailed examination, using the initial standard dimensions, which were mainly cognitive (Little and Gee, 2007). Test–retest stability (24 h and 2 weeks) proved to be moderate, whether considering all projects or only those that were identified as the same at both measurement times. Stability was higher for dimensions reflecting more stable aspects of personality (e.g., meaning dimensions). Considering that PPA was designed to take into account the fluctuating ecosystemic aspects of project pursuit, this moderate stability is satisfying. Indeed, this suggests a desirable level of sensitivity to change. Concerning internal consistency, a moderate alpha coefficient has been obtained for each appraisal dimension when considering the scores given by participants across their projects (Little et al., 1992). This moderate alpha is expected due to the dynamic nature of the project system. Also concerning reliability, high interrater agreement is generally reached when coding project categories according to domains of activity (Little and Gee, 2007).

The factorial structure of personal project appraisals has varied as new dimensions have been added to the original set, especially those more affective dimensions. Factor analyses performed on data collected with the 17 cognitive and nine affective dimensions support a five-factor structure: meaning, manageability, connection, positive affect, and negative affect (Little, 2011; Little and Gee, 2007). It is noteworthy that the positive and negative affective dimensions are orthogonal. Research has shown independent associations between each of these factors and measures of quality of life and clinical concerns like depression (Little, 2011), adding evidence to PPA construct validity. More details on the psychometric and statistical aspects of PPA research can be found in Little and Gee (2007).

Representative Research

In this section, some research findings will be presented in order to exemplify the variety of studies conducted on personal projects. To organize these findings, the social ecological model of human well-being (Little, 1999) will be used. As shown in **Figure 1**, personal projects constitute the central part of the model. The main relationships between the elements of the model can be summarized as follows. Stable characteristics of the person (box A) and stable contextual features (box B) both influence personal projects (box E). The first emanate from evolutionary and genetic sources, and the second from social scripts, rules, and norms. The third originate from idiosyncratic concerns and values. Dynamic aspects of the person (box C) and the context (box D) also have an impact on personal projects. For example, a biogenic extravert may act as a pseudointrovert for the sake of his/her project. A person's perceived affordances of his/her living environment might influence his/her decision to pursue or not a project. Such environmental perceptions are considered dynamic contextual features because they are more flexible than aspects like macrolevel economic or political conditions, which are more resistant to change. Finally, personal projects contribute to human well-being (box F).

Much of the research that has been conducted on personal projects fits into this model, covering one or more of the arrows in **Figure 1**. Here is a selection of representative research findings for each arrow connected to personal projects in the model.

A ➔ E. Emerging from the field of personality science, personal projects have been examined in relation to personality traits in many studies (e.g., Albuquerque et al., 2013; Little et al., 1992). Considering the Big Five dimensions, openness to experience, conscientiousness, extraversion, and agreeableness have been associated with a positive project appraisal across many PPA dimensions, with conscientiousness having the strongest pattern of correlations (Little et al., 1992). Neuroticism is linked to negative project appraisal (Little et al., 1992). The associations vary depending on the categories of projects: the associations of project appraisal with extraversion and agreeableness were particularly important in the case of interpersonal projects. A recent study using structural equation modeling has demonstrated that personality dimensions, except extraversion, have indirect effects on well-being through their impact on the perception of project efficacy (Albuquerque et al., 2013). However, supporting the relative distinctiveness between personality traits and projects, Salmela-Aro et al. (2012) have shown

that genetic effects contribute to the content of personal projects, over and above genetic influence shared with personality traits. The relationships between character strengths and personal projects have also currently been examined in a study exploring the strengths that people with physical disabilities perceive as most useful to the accomplishment of their projects (Jutras et al., 2012).

B ➔ E. The impact of stable contextual features (i.e., objective environmental characteristics) on project pursuit has received limited attention. As mentioned earlier, a study conducted with SEAbank has focused on generational effects on personal projects, providing insights concerning the eventual influence of societal transformations that have occurred across decades since the development of PPA (see Little and Gee, 2007). The study revealed high levels of stability across decades. Recent results reported in Little and Frost (2013) also exemplify the kind of influence that macrosocial aspects of eco-settings have on personal projects. This study examined the intimacy-related projects of lesbian, gay, and bisexual individuals in comparison with heterosexuals. Although, the projects were rated as equally meaningful by the two groups, the first perceived more macrosocial devaluation (e.g., from media, religious groups, government) and structural barriers (e.g., laws and policies) to their projects. It should be noted that this influence is likely to be reciprocal. By publically pursuing intimacy projects, members of sexual minorities collectively act as change agents in their eco-settings: in the long term, they might have a positive impact on the macrosocial devaluation and barriers that impede their intimacy projects (Little and Frost, 2013).

C ➔ E. Dynamic person features have received less empirical attention than stable person features. At the theoretical level, Little's free trait theory (Little, 2000a; Little and Joseph, 2007) postulates that individuals may act out of character when pursuing some of their core personal projects. A biologically introverted person, for example, may find that the successful pursuit of a professional aspiration may entail acting as a (pseudo)extravert. However, such acting out of character, while advancing a person's projects, might have a toll on physical and mental well-being in the long term. The theory proposes that the availability of *restorative niches* where people can freely act in concordance with their first nature can mitigate the negative consequences of acting out of character. Empirical evidence providing initial support for free trait theory has been reported by Little and Joseph (2007).

D ➔ E. Contrary to the limited knowledge on stable (or objective) contextual features, many studies in environmental psychology and urban planning demonstrate that people's perception of their proximal environment plays an important role in the pursuit of their personal projects. For instance, Wallenius (2004) has shown that noise annoyance in the dwelling (due to traffic outside) is correlated with the perception of more stressful projects. In another study (Wallenius, 1999), adult participants from Finland rated the extent to which the places where they pursue their personal projects (mainly home, workplace, and sports facilities) were supportive of their projects. The higher the degree of supportiveness, the higher the level of expected accomplishment. Roe and Aspinall (2012) have shown that, in comparison with local environment, faraway and citywide settings are niches where adolescents from Scotland experience more positive affect and less negative affect while pursuing their projects. Sugiyama and Ward Thompson (2007) have examined the perceived supportiveness of neighborhood environments for older people's outdoor activity projects. In the same vein, in a study conducted in assisted living facilities for people with severe physical disabilities, Jutras et al. (2012) have explored the qualitative perceptions of residents and staff about social and environmental support to residents' projects.

E ➔ F. As stated earlier, well-being is promoted by projects that are more meaningful, manageable, positively connected with others, and that are associated with more positive affect and less negative affect (Little, 1989, 2011). Various aspects of well-being have been examined: hedonic (e.g., satisfaction with life, Palys and Little, 1983), eudaimonic (e.g., purpose in life, McGregor and Little, 1998), and physical well-being (e.g., diabetes metabolic control, Karoly and Bay, 1990). Salmela-Aro and her colleagues have developed a major research program concerning the role of personal projects in adjustment to critical life transitions. For example, longitudinal data suggest that it is beneficial for women to actively adjust their projects during the transition to parenthood: after childbirth, an increase in family projects was associated with a diminution of depressive symptoms (Salmela-Aro, 2012). Highlighting the tractability of personal projects for well-being promotion, Salmela-Aro (2012) has shown that a group intervention can be effective in increasing family projects and have a positive impact for nulliparous women with a severe fear of childbirth.

Conclusion

The study of personal projects has been steadily growing since the introduction of PPA methodology in 1983. It offers an alternative to trait and narrative theories of personality, a methodology that meets the twin challenges of humane and rigorous assessment and a wide-ranging research program that focuses on the diverse ways in which the pursuit of projects enhances well-being. Its adoption in diverse fields of basic and applied research continues to stimulate interdisciplinary research on those forces, both internal and external, that help explain and enhance the quality of lives.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/B978-0-08-097086-8.26100-X.

See also: Affect-Regulation Motivation; Big Five Factor Model, Theory and Structure; Counseling Psychology; Eudaemonism; Extraversion; Factor Analysis and Latent Variable Models in Personality Psychology; Five Factor Model of Personality, Assessment of; Five Factor Model of

Personality, Facets of; Five Factor Model of Personality, Personality Disorder; Five Factor Model of Personality, Universality of; Health Self-Regulation, Motivational and Volitional Aspects of; Motivation and Actions, Psychology of; Motivation in Clinical Interventions; Motivation: Life Course and Sociological Perspectives; Openness to Experience; Passion and Motivation; Personality Assessment, Forced-Choice; Personality Assessment: Overview; Personality Assessment; Personality and Adaptive Behaviors; Personality and Self; Personality, Trait Models of; Personality: Historical and Conceptual Perspectives; School Burnout and Engagement: Lessons from a Longitudinal Study in Finland; Self-Determination Theory; Self-Regulation in Adulthood; Subjective Wellbeing, Psychology of; Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model; Vocational Interests, Values, and Preferences, Psychology of.

Bibliography

Albuquerque, I., Pedroso de Lima, M., Matos, M., Figueiredo, C., 2013. The interplay among levels of personality: the mediator effect of personal projects between the Big Five and subjective well-being. Journal of Happiness Studies 14 (1), 235–250.

Chambers, N.C., 2007. Just doing it: affective implications of project phrasing. In: Little, B.R., Salmela-Aro, K., Phillips, S.D. (Eds.), Personal Project Pursuit: Goals, Action, and Human Flourishing. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, pp. 145–169.

Jutras, S., Coulombe, S., Taupier, K., August 21–25, 2012. Contribution of Character Strengths, Personal Projects, and Well-being to Social Participation of Physically Disabled Persons, Presentation at the 26th Conference of the European Health Psychology Society. Prague, Czech Republic.

Karoly, P., Bay, R.C., 1990. Diabetes self-care goals and their relation to children's metabolic control. Journal of Pediatric Psychology 15 (1), 83–95.

Lawton, M.P., Moss, M.S., Winter, L., Hoffman, C., 2002. Motivation in later life: personal projects and well-being. Psychology and Aging 17 (4), 539–547.

Little, B.R., 1983. Personal projects: a rationale and method for investigation. Environment and Behavior 15 (3), 273–309.

Little, B.R., 1989. Personal projects analysis: trivial pursuits, magnificent obsessions, and the search for coherence. In: Buss, D., Cantor, N. (Eds.), Personality Psychology: Recent Trends and Emerging Directions. Springer-Verlag, New York, NY, pp. 15–31.

Little, B.R., 1999. Personal projects and social ecology: themes and variations across the life span. In: Brandtstädter, J., Lerner, R.M. (Eds.), Action & Self-Development: Theory and Research through the Life Span. Sage Publications, Thousand Oaks, CA, pp. 197–221.

Little, B.R., 2000a. Free traits and personal contexts: expanding a social ecological model of well-being. In: Walsh, W.B., Craik, K.H., Price, R.H. (Eds.), Person-Environment Psychology: New Directions and Perspectives. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, pp. 87–116.

Little, B.R., 2000b. Persons, contexts, and personal projects: assumptive themes of a methodological transactionalism. In: Wapner, S., Demick, J., Yamamoto, T., Minami, H. (Eds.), Theoretical Perspectives in Environment-Behavior Research: Underlying Assumptions, Research Problems, and Methodologies. Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 79–88.

Little, B.R., 2007. Prompt and circumstance: the generative contexts of personal projects analysis. In: Little, B.R., Salmela-Aro, K., Phillips, S.D. (Eds.), Personal Project Pursuit: Goals, Action, and Human Flourishing. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, pp. 3–49.

Little, B.R., 2011. Personal projects and motivational counseling: the quality of lives reconsidered. In: Cox, W.M., Klinger, E. (Eds.), Handbook of Motivational Counseling: Goal-Based Approaches to Assessment and Intervention with Addiction and Other Problems. Wiley-Blackwell, Chichester, UK, pp. 73–87.

Little, B.R., Frost, D.M., 2013. Aspects of love: connecting, romancing, and caring. In: Hojjat, M., Cramer, D. (Eds.), Positive Psychology of Love. Oxford University Press, New York, NY, pp. 162–176.

Little, B.R., Gee, T.L., 2007. The methodology of personal projects analysis: four modules and a funnel. In: Little, B.R., Salmela-Aro, K., Phillips, S.D. (Eds.), Personal Project Pursuit: Goals, Action, and Human Flourishing. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, pp. 51–94.

Little, B.R., Joseph, M.F., 2007. Personal projects and free traits: mutable selves and well beings. In: Little, B.R., Salmela-Aro, K., Phillips, S.D. (Eds.), Personal Project Pursuit: Goals, Action, and Human Flourishing. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, pp. 375–400.

Little, B.R., Lecci, L., Watkinson, B., 1992. Personality and personal projects: linking Big Five and PAC units of analysis. Journal of Personality 60 (2), 501–525.

Little, B.R., Ryan, T.J., 1979. A social ecological model of development. In: Ishwaran, K. (Ed.), Childhood and Adolescence in Canada. McGraw-Hill Ryerson, Toronto, Ont., Canada, pp. 273–301.

Little, B.R., Salmela-Aro, K., Phillips, S.D. (Eds.), 2007. Personal Project Pursuit: Goals, Action, and Human Flourishing. Lawrence Erlbaum Associates Publishers, Mahwah, NJ.

McGregor, I., Little, B.R., 1998. Personal projects, happiness, and meaning: on doing well and being yourself. Journal of Personality and Social Psychology 74 (2), 494–512.

Mischel, W., 1968. Personality and Assessment. Wiley, New York, NY.

Nurmi, J.-E., Salmela-Aro, K., Aunola, K., 2009. Personal goal appraisals vary across both individuals and goal contents. Personality and Individual Differences 47 (5), 498–503.

Omodei, M.M., Wearing, A.J., 1990. Need satisfaction and involvement in personal projects: toward an integrative model of subjective well-being. Journal of Personality and Social Psychology 59 (4), 762–769.

Palys, T.S., Little, B.R., 1983. Perceived life satisfaction and the organization of personal project systems. Journal of Personality and Social Psychology 44 (6), 1221–1230.

Roe, J.J., Aspinall, P.A., 2012. Adolescents' daily activities and the restorative niches that support them. International Journal of Environmental Research and Public Health 9 (9), 3227–3244.

Salmela-Aro, K., 1992. Struggling with self: the personal projects of students seeking psychological counselling. Scandinavian Journal of Psychology 33 (4), 330–338.

Salmela-Aro, K., 2012. Transition to parenthood and positive parenting: longitudinal and intervention approaches. European Journal of Developmental Psychology 9 (1), 21–32.

Salmela-Aro, K., Read, S., Nurmi, J.-E., Vuoksimaa, E., Siltala, M., Dick, D.M., Pulkkinen, L., Kaprio, J., Rose, R.J., 2012. Personal goals and personality traits among young adults: genetic and environmental effects. Journal of Research in Personality 46 (3), 248–257.

Sugiyama, T., Ward Thompson, C., 2007. Older people's health, outdoor activity and supportiveness of neighbourhood environments. Landscape and Urban Planning 83 (2–3), 168–175.

Wallenius, M.A., 1999. Personal projects in everyday places: perceived supportiveness of the environment and psychological well-being. Journal of Environmental Psychology 19 (2), 131–143.

Wallenius, M.A., 2004. The interaction of noise stress and personal project stress on subjective health. Journal of Environmental Psychology 24 (2), 167–177.

Relevant Website

http://www.brianrlittle.com/Topics/research/personal-projects-analysis/ – Online Resources on Personal Projects Analysis.

Personality and Adaptive Behaviors

Gian V Caprara, La Sapienza University, Rome, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Personality is viewed as a self-regulatory system and adaptive behaviors as modes of person–environment interactions that maximize individuals' well-being and development. Over the course of life, people acquire competencies, knowledge about themselves and the social world, and standards for evaluating the worth of own actions. Self-reflective and self-directive processes enable individuals to pursue optimal adjustment by taking advantage of environmental opportunities and selecting activities which optimize the actualization of their potentials.

Personality is the psychological system that mediates the relationship of the individual with the environment in the service of adaptation. It is the complexity of structures, processes and patterns of affect, cognition and behavior that we call 'personality' that conveys to others the sense of unity and distinctiveness that distinguishes individuals from one another and that grants to each individual a sense of identity, of wholeness, and uniqueness. Because the term encompasses a great variety of phenomena, it is not surprising that different ideas appear in various combinations in the traditional and common usage of the word personality: human being, person, the quality of being a person, the quality or fact of being a particular person, patterns of habits and qualities of any individual as expressed through physical and mental activities, the sum of such properties as impressing or likely to impress others, and what makes an individual conspicuously different from others. Nor is it surprising that, given the complex nature of personality, different assumptions, foci, and strategies of research have marked the study of personality since the beginning of the discipline (Allport, 1937; Murray, 1938; Stagner, 1937; Stern, 1935). The resolution of these different definitions and assumptions has considerable implications for the conceptualization and study of adaptation and adaptive behaviors.

Common Assumptions Regarding Personality

Before outlining these different assumptions, let us first consider some areas of commonality in the contemporary field. Most investigators in personality psychology share an 'interactionist view.' In other words, they view personality as an open system that develops and functions through continuous and reciprocal interactions with the environment (Bandura, 1986; Hettema, 1979; Magnusson and Stattin, 1998). Personality, then, refers to a complexity of psychological structures and processes that operate in concert in these person–situation interactions. The personality system underlies the distinctive patterns of affect, cognition, and behavior that foster coherence in individual conduct and experience over time and across settings. Molar psychological mechanisms, such as the goals, beliefs, and skills through which individuals regulate their experiences and actions, result from the synergism of underlying biological and psychological subsystems. Multiple biological systems, cognitive and affective structures, behavior, and the environment all operate as interacting determinants of what personality is at any moment within a network of reciprocal causation. The relative influence of one or another personality determinant varies in different individuals, and across activities and circumstances. Individual development and functioning are processes of continuous restructuring of the whole system and its subsystems within the boundaries set by biological and social constraints (Caprara and Cervone, 2000).

Competing Arguments Regarding Personality

Despite the widespread adherence to an interactionist viewpoint, continuing debates in the field attest to the fact that many fundamental issues are still pending, with different investigators adhering to different theoretical and metatheoretical principles and assigning different importance to nature, nurture, or the individual in the shaping of their personalities (Cervone, 1991). Throughout much of its history, personality psychology has been concerned with individual differences in observable variations in styles of behavior, affect, and cognition. The variations have been organized according to, and traced back to, simple systems of dispositional constructs.

Dispositional constructs are designed to encompass the variety of phenotypic expressions of individuality and to capture the consistent individual differences that are observed. Trait constructs have been posited to account for stable patterns of experience and action that people exhibit, and that distinguish them from one another. As different traits are differentially relevant to different settings and their expression is differentially sensitive to social contexts, different traits attest to the consistency of personality in various degrees depending upon their generalizability (across individuals), pervasiveness (across situations), and stability (across time). This has led dispositional theorists to conceptualize personality as a hierarchical organization and to focus on high-level traits (e.g., extroversion) that organize lower level tendencies (e.g., sociability) which, in turn, supervise lower level behavioral habits (e.g., talkativeness) (Eysenck, 1970).

In the past, alternative taxonomies have been in conflict with each other. However significant consensus recently has been reached on the importance of five factors (so-called Big

 http://dx.doi.org/10.1016/B978-0-08-097086-8.26045-5

Five) that are obtained robustly: extroversion, agreeableness, conscientiousness, emotional stability, and openness to experience. This five-factor model has become a point of convergence between the psycholexical tradition, mostly concerned with personality description (Goldberg, 1993), and the tradition, mostly concerned with the identification of the primary structures of personality (McCrae and Costa, 1996).

Significant differences are found among five-factors investigators, who may disagree on the exact nature and conceptual status of the trait variables (see Wiggins, 1996). However, far more profound differences separate those who view traits as the ultimate determinants of personality from those who conclude that clustering descriptors related to habitual behaviors (as resulting from factor analysis) cannot lead too far in understanding either the regularities that explain personality functioning or the uniqueness which distinguishes persons from one another (Bandura, 1999). The issue in contention is not whether people have personal dispositions, nor their determinative role in personality functioning, but rather how dispositions are conceptualized and operationalized. Dispositions may correspond to habitual styles of behavior rooted in genotypes, or to self-regulatory structures (as internal standards, aspirations and goals, and efficacy beliefs) resulting from the organization of affect, cognition, and behavior which emerges over the course of life.

In conceiving of personality as a hierarchical architecture in which basic traits are at the origins of any other individual disposition or behavior, most convinced advocates of traits as genotypes of personality derive their arguments from genetic, longitudinal, and cross-cultural studies. Genetic studies have found a consistent component of heritability for each of the Big Five (Lohelin, 1992), longitudinal studies have shown a significant stability of the same traits over the life span (Costa and McCrae, 1997), and cross-cultural studies have shown that the multiple descriptors used to describe personality in different cultures can be traced back to few common dimensions broadly overlapping with the Big Five (De Raad, 1998). The trait psychologist converges with the evolutionary psychologist in viewing adaptation as largely predetermined by the basic dispositional equipment of the species selected through generations. According to this view, adaptive behaviors (in mating, fighting, solving problems, and coping with adversities) largely reflect the influence of evolved mechanisms whose nature and functioning has been determined over the eons through processes of natural selection (Buss, 1997).

In contrast, in conceiving of personality as an open, dynamic, unifying and integrating system, social cognitive and motivational theorists (Cervone and Shoda, 1999) point to the emerging properties of the mind and focus on the processes and mechanisms conducive to knowledge structures which enable personality to function as a proactive self-regulatory system. Thus, they contest the generalizability to individuals of correlational findings derived from studying populations, and thus leaving large portions of variability unexplained. Rather, they focus on the construction of personality as an integrative and coherent system as it takes place over the course of life, on the processes that enable the system to function proactively with the environment and on the structures that orchestrate these processes. Ultimately they emphasize the flow of reciprocal determination among biological, psychological, and social factors which sustain the whole construction, and from which the agentic properties of personality derive. In this view of personality, 'adaptation' is eminently relational and contextual. Adaptive behaviors are not ascribed to any preformed, endogenous dispositions, but to psychological systems which themselves develop through dynamic interactions with the social environment. These systems enable the person to cope in a flexible, innovative manner with the diverse challenges of life.

Personality as a Self-Regulatory System

Personality can be viewed as both a construction and an agentic system. Over time, people gradually construct a sense of self through their interactions with the social world. People's capacity for forethought, performance control, and self-reflection contributes to their ability to regulate their own experiences and actions. Self-regulatory capacities enable people to contribute proactively to the course of their development. Personality, then, reveals itself as a construction endowed with agentic properties. These properties enable people to exert a broad influence on their environment and their own psychological and physical selves (Caprara and Cervone, 2000).

Agentic personality functioning involves a number of distinct processes. People select the environments they encounter and thereby create opportunities for themselves, guide their behavior purposively, appraise their circumstances, and chose levels of mental and physical challenge that match their perceived capabilities. Person–situation transactions underscore the fact that personality functioning is not reactive but proactive. Individuals do not only react to the environments, but actively contribute to their transformation and creation. Individuals' distinctive ways of functioning reflect more than the unfolding of a genetic blueprint or the interactions which have taken place between the organism and the environment over the course of development. They also reflect individuals' active role in shaping their own personality.

Thus, one should look at adaptation as a dynamic, proactive process. People do not 'adapt' merely by altering themselves to fit environmental constraints and pressures. They select and transform environments in order to actualize their full potentials and maximize their own well-being. Adaptation relies more on individuals' capacities to devise strategies and to implement tactics able to match successfully the requirements and affordances of multiple situations, than on the fixed pattern of responses. The capacity to expand knowledge of oneselves and of situations is no less important than the capacity to master such knowledge with the flexibility that is required by the variety and continuous change of person–situation transactions.

Over the course of life, people acquire competencies, knowledge about themselves and the social world, and standards for evaluating the worth of own actions. Self-reflective and self-directive processes enable individuals to meet environmental opportunities and constraints as well as to maintain personally valued courses of action in the relative absence of tangible rewards or social support. While people show a striking ability to adapt their behavior to physical,

interpersonal, and social circumstances, and also to select and create the circumstances that best fit with their purposes and abilities, successful adjustment across the life span corresponds to the best allocation of personal and social resources within the set of temporally bound opportunities and constraints.

Adaptive strategies include the realistic adjustment of personal aspirations and goals in accordance with accessible resources and opportunities. This may lead either to develop capacities that allow to alter one's social and physical environment, and to modify one's life views, habits, and internal states. Thus selection, optimization, and compensation have been regarded as main strategies which attest the property of personality to capitalize upon experience in the achievement of higher levels of functioning and well-being. People must select life goals, activities, and pathways that are manageable; they must optimize the allocation of internal and external resources that are made available to them; they must develop strategies to compensate for lacks or losses of goal relevant means (Baltes, 1997; Heckausen and Schultz, 1998).

As people are causal agents who actively contribute to the course of their life in selecting and transforming the environment they encounters, adaptive behaviors are those behaviors that promote the conditions for the full expression of individual capacities and potentials. As individuals' course of life is closely interdependent with the course of life of others, ultimately adaptive behaviors are those behaviors that serve for both the betterment of individual and collective welfare.

Among behaviors fostering successful adaptation, many findings attest to the positive influence of altruistic and prosocial behaviors such as sharing, donating, caring, comforting, and helping (Eisenberg et al., 2006; Penner et al., 2005). Findings from developmental research show that prosocial behavior is positively correlated with well adjustment in children and adolescents while contributing to their accomplishments in social and academic domains (Caprara et al., 2000). The positive impact of prosocial behaviors extends over the entire life span, serving as protective factors that foster self-acceptance, rewarding relationships with others, and life satisfactions, while contrasting antisocial behaviors (Caprara and Steca, 2007; Midlarsky, 1991).

Among knowledge structures fostering successful adaptation across domains of functioning a vast body of research has documented the unique contribution of beliefs of personal efficacy beyond dispositions and skills. Beliefs of personal efficacy influence level of aspirations and goals, what self-regulative standards people adopt, how much efforts they invest in their endeavors, and how resilient they are to adversity (Bandura, 1997). While self-efficacy can be properly developed by proper interventions aimed to strengthen people capacities to reflect and to capitalize upon their experiences, agentic self-beliefs in the domain of affect and interpersonal relations may represent fundamental psychological mediators and moderators that significantly contribute to turn largely inherited tendencies and predispositions into proper views, feelings, goals, and performances (Bandura et al., 2003; Caprara et al., 2010).

In viewing adaptation as a dynamic process that engages the entire personality, future research should address its multiple expressions and structures, how they develop and operate in concert and within their continuous and reciprocal interactions with the environment over the course of life. In particular, it should focus on mechanism and experiences that can strengthen people's capacity to adapt to a rapidly changing world. To this aim, theories and findings are mostly needed to identify the strategies more suitable to promote individual development by enabling people to make choices and to engage in pursuits that mostly serve their growth, health, and well-being.

See also: Coping across the Lifespan; Dimensional Comparisons and Their Consequences for Self-Concept, Motivation, and Emotion; Grit; Health Self-Regulation, Motivational and Volitional Aspects of; Optimism and Pessimism; Personality and Risk Taking; Self-Efficacy; Self-Regulation in Adulthood; Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model; Temperament and Motivation.

Bibliography

Allport, G.W., 1937. Personality: A Psychological Interpretation. Holt, Rinehart & Winston, New York.

Bandura, A., 1986. Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice-Hall, Englewood Cliffs, NJ.

Bandura, A., 1997. Self-efficacy. The Excise of Control. Freeman, New York.

Bandura, A., 1999. Social cognitive theory of personality. In: Cervone, D., Shoda, Y. (Eds.), The Coherence of Personality: Social-Cognitive Bases of Consistency, Variability, and Organization. Guilford Press, New York, pp. 185–241.

Bandura, A., Caprara, G.V., Barbaranelli, C., Gerbino, M., Pastorelli, C., 2003. Role of affective self regulatory efficacy in diverse sphere of psychosocial functioning. Child Development 74, 769–782.

Baltes, P.B., 1997. On the incomplete architecture of human ontogeny: selection, optimization, and compensation as foundation of developmental theory. American Psychologist 52, 366–380.

Buss, D.M., 1997. Evolutionary foundations of personality. In: Hogan, R., Johnson, J., Briggs, S. (Eds.), Handbook of Personality Psychology. Academic Press, San Diego, CA, pp. 318–344.

Caprara, G.V., Alessandri, G., Barbaranelli, C., 2010. Optimal functioning: contribution of self efficacy beliefs to positive orientation. Psychotherapy and Psychosomatics 79, 328–330.

Caprara, G.V., Barbaranelli, C., Pastorelli, C., Bandura, A., Zimbardo, P., 2000. Prosocial foundation of children's academic achievement. Psychological Science 11, 302–305.

Caprara, G.V., Cervone, D., 2000. Personality: Determinants Dynamics, and Potentials. Cambridge University Press, New York.

Caprara, G.V., Steca, P., 2007. Prosocial agency: the contribution of values and self efficacy beliefs to prosocial behavior across ages. Journal of Social and Clinical Psychology 26, 218–239.

Cervone, D., 1991. The two discipline of personality psychology. Psychological Science 2, 371–377.

Cervone, D., Shoda, Y., 1999. The Coherence of Personality. Guilford Press, New York.

Costa, P.T., McCrae, R.R., 1997. Longitudinal stability of adult personality. In: Hogan, R., Johnson, J., Briggs, S. (Eds.), Handbook of Personality Psychology. Academic Press, San Diego, CA, pp. 269–290.

De Raad, B., 1998. Five big five issues: rationale, structure, status and cross-cultural assessment. European Psychologist 3, 113–124.

Eisenberg, N., Fabes, R., Spinrad, T.L., 2006. Prosocial development. In: Damon, W., Lerner, R. (Eds.), Handbook of Child Psychology, vol. 3. Wiley, New York, pp. 646–718.

Eysenck, H., 1970. The Structure of Personality, third ed. Methuen, London.

Goldberg, L., 1993. The structure of phenotypic personality traits. American Psychologist 48, 26–34.

Heckausen, J., Schultz, R., 1998. Developmental regulation in adulthood: selection and compensation via primary and secondary control. In: Heckausen, J., Dweck, C.S. (Eds.), Motivation and Self-regulation across the Life Span. Cambridge, New York.

Hettema, J., 1979. Personality and Adaptation. North-Holland, Amsterdam.

Lohelin, J.C., 1992. Genes and Environment in Personality Development. Sage, Newbury Park, CA.

Magnusson, D., Stattin, H., 1998. Person-context interaction theories. In: Damon, W., Lerner, R.R.M. (Eds.), Handbook of Child Psychology: Theoretical Models of Human Development, fifth ed., vol. 1. Wiley, New York, pp. 685–760.
McCrae, R.R., Costa, P.T., 1996. Toward a new generation of personality theories: theoretical contexts for the five-factor model. In: Wiggins, J.S. (Ed.), The Five-Factor Model of Personality. Theoretical Perspectives. Guilford Press, New York, pp. 51–87.
Midlarsky, E., 1991. Helping as coping. In: Clark, M.S. (Ed.), Prosocial Behavior. Sage Publications, Thousand Oaks, CA, pp. 238–264.
Murray, H.A., 1938. Explorations in Personality. Oxford University Press, New York.
Penner, L.A., Dovidio, J., Piliavin, J.A., Schroeder, D.A., 2005. Prosocial behavior. Multilevel perspectives. Annual Review of Psychology 56, 365–392.
Stagner, R., 1937. Psychology of Personality. McGraw Hill, New York.
Stern, W., 1935. Allgemeine Psychologie auf Personalischer Grundlage. Nijhoff, Dordrecht.
Wiggins, J.S. (Ed.), 1996. The Five-Factor Model of Personality. Theoretical Perspectives. Guilford Press, New York.

Personality and Creativity

Jacob B Hirsh, Institute for Management & Innovation, University of Toronto, Mississauga, ON, Canada; and Rotman School of Management, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Personality plays an important role in the creative process. The search for the 'creative personality' has a long history, and psychologists have converged on the personality dimension of Openness/Intellect as the most important trait predictor of creative achievement. High levels of Openness/Intellect are associated with an increased tendency toward cognitive exploration, supported by divergent thinking, implicit learning, working memory, and decreased latent inhibition. Collectively, variation in these cognitive processes helps to explain the link between personality and creativity.

Creativity is commonly defined as the generation of mental products (e.g., ideas, mathematical formulas, music, visual art, behaviors, technological innovations, etc.) that are both novel and contextually valuable (Runco, 2004; Sternberg and Lubart, 1999). Creative individuals are those who demonstrate an aptitude for generating such creative products. The study of the 'creative personality' has a long scientific history, with researchers attempting to identify the key psychological characteristics that differentiate highly creative individuals from their less creative counterparts.

History of Research on Creativity and Personality

The scientific study of the creative personality began in late-nineteenth century England with Francis Galton's attempts to understand the heritability of genius. Galton, a cousin of Charles Darwin, observed that prominent families tended to produce a large number of eminent individuals, suggesting to him that intellectual and creative abilities were directly inherited from one's parents and ancestors (Galton, 1869). Galton argued that creative individuals were distinguished by their greater intellectual capacity, which he thought was dependent upon underlying biological differences. No clear differentiation between creativity and intelligence was made at the time, as both were considered to be aspects of the broader dimension of 'genius' or 'eminence.'

Lewis Terman continued Galton's research tradition in America during the 1920s. Like Galton, Terman believed that creative genius was driven by biological factors that produced intellectual superiority. Superior intelligence was thus regarded as the primary factor in predicting creative achievement. Terman tested this hypothesis with a longitudinal study of 1300 intellectually gifted children, all of whom were identified as high scorers on one of the first modern IQ tests, the Stanford–Binet Intelligence Scales (Terman, 1925). As the study progressed, however, Terman observed that only a small fraction of the sample of intellectually gifted children ended up producing significant creative accomplishments as adults (Terman and Oden, 1959). Terman concluded that creative achievement must depend on factors beyond intelligence.

In his 1950 presidential address to the American Psychological Association, J.P. Guilford argued that the study of creativity had been largely neglected by psychologists but was of critical importance (Guilford, 1950). Recognizing that intelligence tests did not account for differences in creative achievement, Guilford argued that a broader range of personality characteristics needed to be considered in order to identify creative individuals. He also suggested that creativity is related to 'divergent thinking,' a pattern of thought in which multiple alternative possibilities are generated, in contrast to 'convergent thinking,' which attempts to identify a single correct solution to a problem (Guilford, 1967). Guilford's research helped to advance the psychometric assessment of creativity as distinct from what is captured by traditional intelligence tests.

Although the psychometric assessment of creativity did not fully emerge prior to Guilford's work, creativity was still regarded as an important process in many theories of personality function. Sigmund Freud believed that creativity involved the sublimation of repressed sexual desires into products of the imagination (Freud, 1908). Carl Jung argued that creativity is driven by powerful unconscious forces that can overwhelm the conscious will of the creative artist who experiences them (Jung, 1933). Alfred Adler regarded creativity as central to an individual's ability to take control of his or her life and transform it for the better (Adler, 1927). Otto Rank viewed creativity as an adaptive process of 'unlearning' socially prescribed beliefs and behavioral patterns, with neurosis reflecting the failure to adapt creatively to life (Rank, 1932). George Kelly suggested that creativity involves the temporary loosening of one's conceptual system, which is then reformed around a novel idea that seems promising (Kelly, 1955). Abraham Maslow regarded creativity as a central component of healthy psychological development and personal growth, naturally emerging from a self-actualizing individual's engagement with the world (Maslow, 1962). Carl Rogers also regarded creativity as a natural human process, emphasizing that it is disrupted when the individual feels threatened, flourishing only under conditions of psychological safety (Rogers, 1961). B.F. Skinner, meanwhile, attributed creativity to the accidental 'mutation' of previously learned behavioral patterns, which are only preserved if they are reinforced by the environment (Skinner, 1970).

Since Guilford's call to scientific action, a large number of personality psychologists have examined the character traits associated with creative achievement, both by examining

creativity in the normal population and by studying the personalities of highly eminent creative people. Some of the personality traits that are most commonly associated with creativity include greater aesthetic sensitivity, broad interests, attraction to complexity, independence of judgment, autonomy and self-direction, tolerance of ambiguity, and a clear sense of one's self as 'creative' (Barron and Harrington, 1981). Many of these characteristics are captured in the 'creative personality scale,' which was designed to assess the personality traits that most clearly distinguish creative individuals from less creative ones (Gough, 1979).

Openness/Intellect and the Five-Factor Model of Personality

Although researchers in the 1950s, 1960s, and 1970s identified numerous personality traits that were associated with creativity, integrating these findings into a coherent framework proved to be difficult. In particular, the diversity of personality taxonomies and scales used by different investigators resulted in a proliferation of overlapping trait constructs with unclear boundaries. Beginning in the late 1980s, however, the five-factor model of personality emerged as an integrative taxonomy for categorizing human personality traits (Digman, 1990). According to this framework, personality varies along five broad trait dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness/Intellect (Goldberg, 1990). Studies employing the five-factor model have consistently associated greater creative abilities with higher scores on the Openness/Intellect trait dimension, using both self- and peer-rated assessments of personality (McCrae, 1987). While other traits, including high Extraversion and low Agreeableness, have also been related to creative accomplishments, these effects tend to disappear when statistically controlling for Openness/Intellect levels (e.g., King et al., 1996).

As a personality dimension, Openness/Intellect describes the extent to which an individual is prone to aesthetic absorption and cognitive exploration (DeYoung et al., 2005). The broad trait domain is comprised of the two lower-order traits of 'Openness,' which reflects aesthetic sensitivity and imaginative tendencies, and 'Intellect,' which reflects intellectual engagement and interest in abstract ideas (DeYoung et al., 2007). A strong biological basis exists for Openness/Intellect, with approximately 61% of the variance in this trait being driven by genetic factors (Jang et al., 1996). Genetic variation in the mesocortical dopamine system, which supports cognitive exploration and flexibility, has been identified as contributing to an individual's standing on this trait dimension (DeYoung et al., 2011).

Cognitive Processes Linking Openness/Intellect with Creativity

Openness/Intellect has been associated with a number of cognitive processes that appear to support its relationship with enhanced creativity, including decreased latent inhibition and superior divergent thinking, implicit learning, and working memory performance.

Decreased Latent Inhibition

Individuals with higher levels of Openness/Intellect have been observed to display decreased latent inhibition (Peterson and Carson, 2000). Latent inhibition describes the cognitive process by which stimuli with no apparent motivational relevance are marked as insignificant and not worthy of further attention. Those with high levels of latent inhibition have difficulty learning new associations with previously encountered objects, relying instead upon existing memories derived from previous experience. Decreased latent inhibition, on the other hand, is associated with the ability to examine familiar objects and ideas as if they were being encountered for the first time. This ability to reexamine and transform familiar meanings into novel patterns is in turn associated with higher levels of creative achievement (Carson et al., 2003).

Divergent Thinking

Openness/Intellect is also the best personality predictor of divergent thinking (McCrae, 1987; Silvia et al., 2009). Since Guilford's seminal work on the topic, divergent thinking has been considered to be one of the core cognitive processes that support creative idea generation because it involves the proliferation of creative possibilities. Tests of divergent thinking, such as asking people to think of as many unusual uses for a familiar object as they can (Torrance, 1966), are able to predict a variety of real-world creative outcomes (Runco, 1991). Openness/Intellect appears to support the generation of novel possibilities and cognitive associations.

Implicit Learning

The subtrait of Openness has been associated with the facilitation of implicit pattern learning (Kaufman et al., 2010). Although we learn many things using conscious, or explicit, processing, we also learn a great deal implicitly, or without conscious awareness (Reber, 1989). Implicit learning has long been theorized to play an important role in the creative process. In particular, it has been argued that creative ideas begin as implicit or unconscious representations, which need to be explicitly articulated in a symbolic form during the creative process (Policastro, 1995). Unconscious processing has indeed been identified as an important aspect of creativity, due to its predominantly divergent and associative nature (Dijksterhuis and Meurs, 2006). Implicit learning appears to facilitate the generation of creative insights during the so-called 'incubation' phase of creative problem-solving, when no conscious efforts are being made toward a solution (Dorfman et al., 1996). Higher levels of Openness thus appear to support creativity through the facilitation of implicit learning and intuitive information processing (McCrae, 1993). This heightened ability to learn implicit patterns may also facilitate the use of analogical reasoning, which supports creative idea generation (Holyoak and Thagard, 1995).

Working Memory

The subtrait of Intellect is the only personality trait that shows consistent relationships with standard measures of

intelligence, and it has been uniquely associated with brain activity in regions that support working memory (DeYoung et al., 2009). Working memory capacity describes the amount of information that can be simultaneously apprehended by an individual, and reflects one of the core components of intelligence. Working memory capacity varies tremendously from one person to the next, with some people being able to attend to and process a much larger amount of information (Engle et al., 1999). Being able to process more information simultaneously is likely to allow a broader range of creative idea combinations to emerge into consciousness (Campbell, 1960).

It should be pointed out that although an individual's Openness/Intellect scores can predict creative achievement beyond the effects of intelligence alone, there is still a significant positive relationship between intellectual ability and creative achievement. According to the threshold hypothesis, however, the relationship between intelligence and creativity tends to drop substantially once a critical level of intelligence is reached (an IQ of approximately 120; Batey and Furnham, 2006).

Collectively, the psychological processes associated with Openness/Intellect reflect the tendency to explore and revise perceptual patterns and cognitive models of the world (DeYoung, 2014). The heightened 'permeability of consciousness' associated with this personality dimension appears to play a critical role in facilitating the creative process (McCrae, 1994).

Changes in Openness/Intellect over Time

Although Openness/Intellect has a large genetic component, it can also change over time as a result of experience. On average, scores on this trait dimension tend to rise sharply during adolescence and then begin to decline during old age (Roberts et al., 2006). It may be possible to preserve levels of Openness/Intellect into old age, however, using cognitive exercise. An intervention with older adults found that a 16-week inductive reasoning training program, which included weekly Sudoku and crossword puzzles, resulted in a significant increase in Openness/Intellect as compared to a wait-list control group (Jackson et al., 2012). Other life events, such as marriage, appear to hasten the age-related declines in this trait (Specht et al., 2011). Pharmacological interventions also appear to have an influence, with a single high dose of psilocybin producing spiritual experiences that led to increased Openness/Intellect scores being observed a full year later (MacLean et al., 2011).

It is also worth noting that modern personality theories recognize that there are substantial within-person fluctuations in personality characteristics from one situation to the next. What this suggests is that a relatively 'creative person' will not necessarily be creative at all times. A person's overall score on a standard personality test in fact represents his or her 'average' standing on a given trait dimension, which can vary depending upon the situational context (Fleeson and Gallagher, 2009). An individual's degree of Openness/Intellect, and thus his or her 'creative potential,' can therefore change substantially from one situation to another.

Creativity and Psychopathology

Openness/Intellect is also relevant for understanding the fact that creativity has long been associated with psychopathology in the popular imagination, with a thin line separating madness from genius. Indeed, Hans Eysenck proposed that creativity is facilitated by subclinical levels of psychotic tendencies (Eysenck, 1993). In particular, he argued that although schizophrenia has often been regarded as a categorical disorder, it should more appropriately be regarded from a dimensional perspective as a spectrum of cognitive and emotional patterns that vary throughout the entire population. Recent work has confirmed that the positive symptoms of schizotypy are associated with extremely high levels of the personality trait of Openness (Edmundson et al., 2011; Piedmont et al., 2009). Excessive Openness (but not Intellect) appears to be related to an increased tendency toward magical beliefs and unusual perceptual experiences (DeYoung et al., 2012). While the ability to detect and develop novel sensory and ideational patterns is associated with enhanced creativity, it can also result in delusional thinking and a loosened grip on social reality. The extent to which Openness is associated with creativity rather than schizotypy may depend upon the individual's ability to integrate the ongoing flow of novel cognitive and sensory patterns within a coherent worldview. High levels of intelligence are in fact a protective factor against schizotypal diagnosis (David et al., 1997), suggesting that strong convergent thinking abilities are needed to 'reign in' the tendencies toward divergent thinking that are associated with very high levels of Openness. Traits related to schizotypy do not appear to predict creative success beyond the effects of Openness and Intelligence (Miller and Tal, 2007), suggesting that any of the creative benefits of schizotypy are mediated through the same processes as those associated with Openness/Intellect.

Personality and Domain-Specific Creativity

Although Openness/Intellect is associated with creativity across a variety of domains, the importance of other personality traits appears to vary depending upon the specific domain of creative endeavor. A meta-analysis that examined the personality characteristics associated with scientific and artistic creativity found that higher levels of Openness/Intellect predicted creativity in both domains (Feist, 1998). Creative scientists, however, had substantially higher levels of Conscientiousness (the tendency to be organized and industrious), whereas creative artists displayed lower levels of Emotional Stability (the tendency to experience negative emotion). A recent study examining how scientific and artistic creativity map onto the lower-order personality traits within the five-factor model found that while higher levels of Openness were associated with creativity in artistic domains, higher levels of Intellect were associated with creativity in scientific domains (Kaufman, 2009).

An important point to take from the research on personality and creativity is that there are in fact many different varieties of 'creative personality.' While Openness/Intellect appears to be the strongest and most reliable trait predictor of creativity across different domains, creative success may also require other traits depending upon the specific field (Baer, 1998). Creative individuals are thus likely to be united by their high

levels of Openness/Intellect, but may vary broadly in terms of their other personality characteristics.

See also: Big Five Factor Model, Theory and Structure; Five Factor Model of Personality, Assessment of; Five Factor Model of Personality, Personality Disorder; Intelligence: Central Conceptions and Psychometric Models; Intelligence: Historical and Conceptual Perspectives; Openness to Experience; Personality, Biological Models of; Personality, Cognitive Models of; Personality, Trait Models of; Personality: Historical and Conceptual Perspectives; Schizophrenia.

Bibliography

Adler, A., 1927. Understanding Human Nature. Greenberg, Oxford, England.

Baer, J., 1998. The case for domain specificity of creativity. Creativity Research Journal 11, 173–177.

Barron, F., Harrington, D.M., 1981. Creativity, intelligence, and personality. Annual Review of Psychology 32, 439–476.

Batey, M., Furnham, A., 2006. Creativity, intelligence, and personality: a critical review of the scattered literature. Genetic, Social, and General Psychology Monographs 132, 355–429.

Campbell, D.T., 1960. Blind variation and selective retentions in creative thought as in other knowledge processes. Psychological Review 67, 380–400.

Carson, S., Peterson, J., Higgins, D., 2003. Decreased latent inhibition is associated with increased creative achievement in high-functioning individuals. Journal of Personality and Social Psychology 85, 499–506.

David, A.S., Malmberg, A., Brandt, L., Allebeck, P., Lewis, G., 1997. IQ and risk for schizophrenia: a population-based cohort study. Psychological Medicine 27, 1311–1323.

DeYoung, C.G., 2014. Openness/intellect: a dimension of personality reflecting cognitive exploration. In: Cooper, M.L., Larsen, R.J. (Eds.), The APA Handbook of Personality and Social Psychology: Personality Processes and Individual Differences, Vol. 4. American Psychological Association, Washington, DC, pp. 369–399.

DeYoung, C.G., Cicchetti, D., Rogosch, F.A., Gray, J.R., Eastman, M., Grigorenko, E.L., 2011. Sources of cognitive exploration: genetic variation in the prefrontal dopamine system predicts openness/intellect. Journal of Research Personality 45, 364–371.

DeYoung, C.G., Grazioplene, R.G., Peterson, J.B., 2012. From madness to genius: the openness/intellect trait domain as a paradoxical simplex. Journal of Research in Personality 46, 63–78.

DeYoung, C.G., Peterson, J.B., Higgins, D.M., 2005. Sources of openness/intellect: cognitive and neuropsychological correlates of the fifth factor of personality. Journal of Personality 73, 825–858.

DeYoung, C.G., Quilty, L.C., Peterson, J.B., 2007. Between facets and domains: ten aspects of the big five. Journal of Personality and Social Psychology 93, 880–896.

DeYoung, C.G., Shamosh, N., Green, A., Braver, T., Gray, J., 2009. Intellect as distinct from openness: differences revealed by fMRI of working memory. Journal of Personality and Social Psychology 97, 883–892.

Digman, J.M., 1990. Personality structure: emergence of the five-factor model. Annual Review of Psychology 41, 417–440.

Dijksterhuis, A., Meurs, T., 2006. Where creativity resides: the generative power of unconscious thought. Consciousness and Cognition 15, 135–146.

Dorfman, J., Shames, V.A., Kihlstrom, J.F., 1996. Intuition, incubation, and insight: Implicit cognition in problem solving. In: Implicit Cognition. Oxford University Press, New York, pp. 257–296.

Edmundson, M., Lynam, D.R., Miller, J.D., Gore, W.L., Widiger, T.A., 2011. A five-factor measure of schizotypal personality traits. Assessment 18, 321–334.

Engle, R.W., Kane, M.J., Tuholski, S.W., 1999. Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In: Models of Working Memory: Mechanisms of Active Maintenance and Executive Control, pp. 102–134.

Eysenck, H.J., 1993. Creativity and personality: suggestions for a theory. Psychological Inquiry 4, 147–178.

Feist, G.J., 1998. A meta-analysis of personality in scientific and artistic creativity. Personality and Social Psychology Review 2, 290–309.

Fleeson, W., Gallagher, P., 2009. The implications of Big Five standing for the distribution of trait manifestation in behavior: fifteen experience-sampling studies and a meta-analysis. Journal of Personality and Social Psychology 97, 1097–1114.

Freud, S., 1908. Creative writers and daydreaming. In: The Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 9. Hogarth Press, London, England, pp. 142–152.

Galton, F., 1869. Hereditary Genius: An Inquiry into its Laws and Consequences. Macmillan and Co, London, England.

Goldberg, L.R., 1990. An alternative 'description of personality': the Big-Five factor structure. Journal of Personality and Social Psychology 59, 1216–1229.

Gough, H.G., 1979. A creative personality scale for the Adjective Check List. Journal of Personality and Social Psychology 37, 1398–1405.

Guilford, J.P., 1950. Creativity. American Psychologist 5, 444–454.

Guilford, J.P., 1967. The Nature of Human Intelligence. McGraw-Hill, New York.

Holyoak, K.J., Thagard, P., 1995. Mental Leaps: Analogy in Creative Thought. The MIT Press, Cambridge, MA.

Jackson, J.J., Hill, P.L., Payne, B.R., Roberts, B.W., Stine-Morrow, E.A.L., 2012. Can an old dog learn (and want to experience) new tricks? Cognitive training increases openness to experience in older adults. Psychology and aging 27, 286–292.

Jang, K.L., Livesley, W.J., Vemon, P.A., 1996. Heritability of the big five personality dimensions and their facets: a twin study. Journal of Personality 64, 577–592.

Jung, C.G., 1933. Modern Man in Search of a Soul. Harcourt, Brace, Oxford, England.

Kaufman, S.B., 2009. Beyond General Intelligence: The Dual-Process Theory of Human Intelligence (Ph.D.). Yale University, New Haven, CT. Retrieved from: http://search.proquest.com//docview/305042238.

Kaufman, S.B., DeYoung, C.G., Gray, J.R., Jiménez, L., Brown, J., Mackintosh, N., 2010. Implicit learning as an ability. Cognition 116, 321–340.

Kelly, G., 1955. The Psychology of Personal Constructs. Norton, New York.

King, L., Walker, L., Broyles, S., 1996. Creativity and the five-factor model. Journal of Research in Personality 30, 189–203.

MacLean, K.A., Johnson, M.W., Griffiths, R.R., 2011. Mystical experiences occasioned by the hallucinogen psilocybin lead to increases in the personality domain of openness. Journal of Psychopharmacology 25, 1453–1461.

Maslow, A., 1962. Toward a Psychology of Being. D Van Nostrand, Princeton, NJ.

McCrae, R.R., 1987. Creativity, divergent thinking, and openness to experience. Journal of Personality and Social Psychology 52, 1258–1265.

McCrae, R.R., 1993. Openness to experience as a basic dimension of personality. Imagination, Cognition and Personality 13, 39–55.

McCrae, R.R., 1994. Openness to experience: expanding the boundaries of Factor V. European Journal of Personality 8, 251–272.

Miller, G.F., Tal, I.R., 2007. Schizotypy versus openness and intelligence as predictors of creativity. Schizophrenia Research 93, 317–324.

Peterson, J.B., Carson, S., 2000. Latent inhibition and openness to experience in a high-achieving student population. Personality and Individual Differences 28, 323–332.

Piedmont, R.L., Sherman, M.F., Sherman, N.C., Dy-Liacco, G.S., Williams, J.E.G., 2009. Using the five-factor model to identify a new personality disorder domain: the case for experiential permeability. Journal of Personality and Social Psychology 96, 1245–1258.

Policastro, E., 1995. Creative intuition: an integrative review. Creativity Research Journal 8, 99–113.

Rank, O., 1932. Art and Artist; Creative Urge and Personality Development. Knopf, Oxford, England.

Reber, A.S., 1989. Implicit learning and tacit knowledge. Journal of Experimental Psychology: General 118, 219–235.

Roberts, B.W., Walton, K.E., Viechtbauer, W., 2006. Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. Psychological Bulletin 132, 1–25.

Rogers, C.R., 1961. On Becoming a Person. Houghton Mifflin, Oxford, England.

Runco, M.A., 1991. Divergent Thinking. Ablex Publishing, Westport, CT.

Runco, M.A., 2004. Creativity. Annual Review of Psychology 55, 657–687.

Silvia, P.J., Nusbaum, E.C., Berg, C., Martin, C., O'Connor, A., 2009. Openness to experience, plasticity, and creativity: exploring lower-order, high-order, and interactive effects. Journal of Research in Personality 43, 1087–1090.

Skinner, B.F., 1970. Creating the creative artist. In: Toynbee, A.J. (Ed.), On the Future of Art. Viking Press, New York.

Specht, J., Egloff, B., Schmukle, S.C., 2011. Stability and change of personality across the life course: the impact of age and major life events on mean-level and rank-order stability of the Big Five. Journal of Personality and Social Psychology 101, 862–882.

Sternberg, R.J., Lubart, T.I., 1999. The concept of creativity: prospects and paradigms. In: Handbook of Creativity. Cambridge University Press, New York, pp. 3–15.

Terman, L.M., 1925. Genetic Studies of Genius. Mental and Physical Traits of a Thousand Gifted Children. Stanford University Press, Palo Alto, CA.

Terman, L.M., Oden, M.H., 1959. Genetic Studies of Genius. In: The Gifted Group at Mid-life, vol. V. Stanford University Press, Palo Alto, CA.

Torrance, E., 1966. The Torrance Test of Creative Thinking: Norms-technical Manual. Personnel Press, Lexington, MA.

Personality and Crime

David P Farrington, Institute of Criminology, Cambridge, UK
Darrick Jolliffe, University of Greenwich, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Personality traits are persisting underlying tendencies to behave in particular ways in particular situations. Temperament is the childhood equivalent of personality, and a difficult, undercontrolled temperament in infancy predicts later offending. The most crucial personality trait that predicts offending is impulsiveness, which may reflect deficits in the executive functions of the brain. Low empathy is probably related to offending, but the evidence is less convincing. Psychopathic personality (a cold, callous personality, including low empathy, low guilt, and low remorse) is more clearly correlated with offending. The Eysenck theory links extraversion, neuroticism, and psychoticism to offending, but analyses of Eysenck's personality inventories show that the most significant relationships reflect items measuring impulsiveness. Currently, the most widely accepted personality system is the five-factor model or 'Big Five' personality dimensions. Of these, agreeableness and conscientiousness are most strongly correlated (negatively) with offending. More research is needed on the theoretical linkages between personality constructs and offending and on their biological bases. Longitudinal studies are especially needed in which changes in personality constructs are related to changes in offending. The main policy implication of research on personality and crime is that it is desirable to implement cognitive-behavioral skills training techniques to try to counteract high impulsiveness and low empathy.

It is a basic assumption of psychology that behavior arises out of the interaction between the individual and the environment. Studies have shown that behavior is generally consistent over time or, more accurately, the relative ordering of individuals is remarkably consistent over time. It is hypothesized that this behavioral consistency depends primarily on the persistence of underlying tendencies to behave in particular ways in particular situations. These tendencies are termed personality traits, such as impulsiveness, sensation seeking, assertiveness, modesty, and dutifulness. Personality dimensions, such as extraversion, are clusters of personality traits that are typically found to be associated with one another (e.g., impulsiveness and sensation seeking). Roberts and Del Vecchio (2000) found that average test–retest correlations increased from .31 in childhood to .54 in the college years and .74 in adulthood (for an average time interval of 7 years).

Many constructs that are assumed to underlie behavior are not included under the heading of personality. This is true, for example, of intelligence and moral reasoning, and of cognitive processes such as thinking, reasoning, and decision making. However, the boundaries between personality and all of these constructs are not well defined. For example, it has been suggested that neuropsychological deficits in the executive functions of the brain (e.g., reasoning, planning, sustaining attention, and concentration) can cause both low IQ and high impulsiveness (Moffitt, 1990). This article will focus on topics conventionally included under the heading of personality; there is not space here to discuss underlying biological processes in any detail (see e.g., Beaver and Walsh, 2010).

The most common types of offending studied are burglary, theft, violence, vandalism, fraud, and drug offenses. Since crimes are a subset of antisocial acts, any potential to commit crime is probably part of a broader potential to commit antisocial acts, but the focus here is on crime. The focus is also on normal personality, and as far as possible discussions of related psychiatric conditions such as antisocial personality disorder, psychopathy, and attention deficit-hyperactivity disorder are avoided. However, the boundary between normal and pathological behavior is not always clear, depending on judgments about the extent to which normal functioning has been impaired. There will also be some discussion of psychopathic personality.

In understanding the potentially causal relationship between personality and crime, tautological statements should be avoided. Any construct that is presumed to cause crime should be measuring something different from crime. Thus, it is useful to investigate the extent to which low empathy causes crime and less useful to investigate the extent to which high aggressiveness causes violent crime because both are probably behavioral manifestations of the same persisting underlying theoretical construct. Early research using the psychopathic deviate scale of the Minnesota Multiphasic Personality Inventory and the socialization scale of the California Psychological Inventory was essentially tautological in showing that antisocial behavior was correlated with offending (Wilson and Herrnstein, 1985: pp. 186–194).

A bewildering number of personality traits and dimensions have been postulated, measured, and related to offending. In some cases, different terms are probably used to describe the same underlying construct. Only the most important topics can be reviewed here: temperament, impulsiveness, empathy, the Eysenck theory, and the 'Big Five' personality dimensions.

Temperament

Temperament is basically the early childhood equivalent of personality, although there is more emphasis in the temperament literature on constitutional predisposition and genetic and biological factors that influence temperament. The modern study of child temperament began with the New York Longitudinal Study of Chess and Thomas (1984). Children in their first 5 years of life were rated on temperamental dimensions by their parents, and these dimensions were combined into

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45032-4

three broad categories of easy, difficult, and 'slow to warm up' temperament. Having a difficult temperament at ages 3–4 years (frequent irritability, low amenability and adaptability, irregular habits) predicted poor adult psychiatric adjustment at ages 17–24.

Remarkably, Bates (1989) found that mothers' ratings of difficult temperament as early as age 6 months (defined primarily as frequent, intense expressions of negative emotions), predicted mothers' ratings of child conduct problems between the ages of 3 and 6 years. Similar results were obtained in the Australian Temperament Project, which found that children who were rated as irritable, not amenable, or showing behavior problems at ages 4–8 months tended to be rated as aggressive at ages 7–8 years (Sanson et al., 1993). When information at each age comes from the same source it is possible that the continuity lies in the rater, rather than the child. However, in a sample of 103 children followed up from age 2 to age 17, Guerin et al. (2003) found that difficult temperament in infancy, rated by mothers, predicted a greater frequency of attention problems, aggressive problems, and thinking problems in the elementary school years, as reported by teachers.

Other researchers have investigated more specific dimensions of temperament than the rather vague 'difficult temperament.' For example, Kagan et al. (1988) in Boston classified children as inhibited (shy or fearful) or uninhibited at age 21 months on the basis of their observed reactions to a strange situation, and concluded that about 10–20% of children were inhibited, while about 30–40% were uninhibited. They found that children were significantly stable on this classification up to age 7 years. Furthermore, the uninhibited children at age 21 months significantly tended to be identified as aggressive at age 13 years, according to self and parent reports (Schwartz et al., 1996).

There are several systems for classifying childhood temperament. Evans and Rothbart (2007) identified three important dimensions: (1) effortful control (including planning ahead and the suppression of inappropriate responses), which was related to the 'Big Five' dimension of conscientiousness in adulthood (see later); (2) negative affect (including frustration, fear, and sadness), which was related to the 'Big Five' dimension of neuroticism; and (3) extraversion (including impulsiveness and low shyness), which was related to the 'Big Five' dimension of extraversion. Lahey et al. (2006) also identified three important dimensions (low prosociality, negative emotionality, and high daring) and showed that all three, measured at age 7, predicted delinquency up to age 17 in the Pittsburgh Youth Study, which is a prospective longitudinal survey of over 1500 Pittsburgh boys.

The most important results on the link between childhood temperament and later offending have been obtained in the Dunedin longitudinal study in New Zealand, which has followed up over 1000 children from age 3 years into their thirties. Temperament at age 3 years was rated by observing the child's behavior during a testing session involving cognitive and motor tasks. The most important dimension of temperament was being undercontrolled (restless, impulsive, with poor attention), which predicted aggression, self-reported delinquency, and convictions at ages 18–21 years and problem gambling at age 32 years (Caspi, 2000; Slutske et al., 2012).

Impulsiveness

The most important study of childhood temperament in relation to later offending essentially identified impulsiveness as the key dimension, and it is generally true that impulsiveness is the most crucial personality dimension that predicts offending. There are a large number of constructs that refer to a poor ability to control behavior. These include low self-control, hyperactivity, inattention, a poor ability to delay gratification, short time horizons, risk taking, sensation seeking, and not considering consequences before acting. There are differences among the specific definitions of many of these terms, but all these variables (variously measured) are consistently related to offending (Ellis and Walsh, 2000: p. 217). Impulsiveness is genetically influenced and relatively stable over time (Niv et al., 2012).

High impulsiveness might contribute to an increased likelihood of offending either directly or indirectly. By the direct pathway, impulsiveness would have a fundamental impact on the way in which an individual would act in any given situation. With less time and cognitive energy available for contemplation of potential responses, an impulsive individual would, on balance, select the option that provides the most immediate perceived benefit. Numerous researchers and theorists have detailed how criminal offending tends to satisfy immediate urges at the risk of later, uncertain, and possibly negative consequences (e.g., Farrington and Welsh, 2007: p. 52).

Other researchers have suggested that high impulsiveness may contribute to an increased likelihood of offending indirectly through a person–environment interaction (e.g., Neumann et al., 2010). Impulsiveness may have different effects in different families and neighborhoods. Also, an impulsive child may prove a challenge to even the most conscientious parent, over time eliciting poor parenting, which is in itself an established risk factor for later offending. Thus, there may be child effects on parenting that lead to offending. Similarly, impulsiveness may lead to offending by reducing the likelihood of school success. Typically, formal education is delivered in a restrictive environment that might prove very difficult for children who are impulsive. Thus, impulsiveness may lead to low achievement, which in turn leads to offending. Children who have low school achievement are more likely to drop out of school and more likely to become socioeconomically disadvantaged. With reduced opportunity to succeed in conventional ways, these individuals may be more likely to turn to antisocial or criminal methods to obtain money and status (see Agnew, 2006).

In the Cambridge Study in Delinquent Development, in which 411 males have been followed up from ages 8 to 48, boys identified by teachers as lacking in concentration or restless, those identified by parents, peers, or teachers as the most daring or risk taking, and those who were the most impulsive on psychomotor tests at ages 8–10, all tended to become offenders later in life. Later self-report measures of impulsiveness were also related to offending. High daring or risk taking, poor concentration, and restlessness all predicted both official

convictions and self-reported delinquency, and high daring was consistently one of the best independent predictors (e.g., Farrington et al., 2009).

A multi-method, multi-source assessment of impulsivity was carried out in the Pittsburgh Youth Study (White et al., 1994). The measures that were most strongly related to self-reported delinquency at ages 10 and 13 were teacher-rated impulsiveness (e.g., acts without thinking), self-reported impulsiveness (e.g., does things without planning), self-reported undercontrol (e.g., unable to delay gratification), motor restlessness (from videotaped observations), and psychomotor impulsiveness (from the Trail Making Test). Generally, the verbal behavior rating tests yielded stronger relationships with offending than the psychomotor performance tasks, suggesting that cognitive impulsiveness (based on thinking processes) was more relevant to delinquency than was behavioral impulsiveness (based on psychomotor test performance). The future time perception and delay of gratification tasks were only weakly related to self-reported offending.

Jolliffe and Farrington (2009) undertook a systematic review and meta-analysis of the relationship between early measures of impulsiveness and later violence using only data from prospective longitudinal studies. The results suggested that early measures of impulsiveness (measured as early as age 5) were quite strongly related to later violence (measured as late as age 40). While all measures of impulsiveness were associated with later violence, measures of high daring and risk taking were particularly strongly related to later violence. Jolliffe and Farrington suggested that high daring and risk taking might influence offending directly (because risk taking leads to offending), whereas the other forms of impulsiveness (e.g., poor concentration) might have a more muted effect through mediating variables such as school failure. Meta-analyses have also shown the link between low self-control and offending (e.g., Pratt and Cullen, 2000), and between attention-deficit hyperactivity disorder and offending (e.g., Pratt et al., 2002).

There have been many theories put forward to explain the link between impulsiveness and offending. One of the most popular suggests that impulsiveness reflects deficits in the executive functions of the brain, located in the frontal lobes (Moffitt, 1990). Persons with these neuropsychological deficits will tend to commit offenses because they have poor control over their behavior, a poor ability to consider the possible consequences of their acts, and a tendency to focus on immediate gratification (Morgan and Lillienfeld, 2000). There may also be an indirect link between neuropsychological deficits and offending that is mediated by hyperactivity and inattention in school and the resulting school failure. In discussing links between executive functions and offending, impulsiveness may be difficult to disentangle from intelligence, although Lynam and Moffitt (1995) argued that they were different constructs.

Empathy

There is a widespread belief that low empathy is an important personality trait that is related to offending. A distinction has often been drawn between cognitive empathy (i.e., understanding the emotions of another person) and affective empathy (i.e., actually experiencing another person's emotions). Therefore, those with high empathy should be less likely to offend because they are sensitized to the emotional impact of their actions on others. Those who lack empathy, on the other hand, can victimize others free from these internal cognitive and affective inhibitions. The theoretical relationship between low empathy and offending is more logical for interpersonal offenses such as violence and sex than for offenses where the victim is less obvious (e.g., shoplifting or drug possession).

Support for the inverse relationship between empathy and offending also comes from the study of psychopathy. Psychopathic personality is a constellation of psychological and behavioral traits that are linked to an increased likelihood of criminal and antisocial behavior (Skeem and Cooke, 2010). Those classified as psychopaths tend to be glib, superficial, and manipulative, but it is the lack of appropriate emotionality and empathy that especially characterizes this condition. It is well established that psychopaths are more violent and more likely to recidivate than nonpsychopaths, and it is the conspicuous lack of empathy of psychopaths that is often used to explain these behavioral outcomes. The childhood equivalent of psychopathic personality is callous-unemotional personality (Frick and White, 2008).

While there is clearly a strong theoretical relationship between low empathy and offending, the empirical evidence for this relationship is less impressive. For example, studies comparing empathy scores of offenders and nonoffenders have not produced consistently convincing results. This is true of studies comparing general offenders to nonoffenders (e.g., Goldstein and Higgins-D'Alessandro, 2001) as well as studies comparing sex offenders to nonoffenders (e.g., Hayashino et al., 1995). These findings are contrary to expectation, especially as it has long been assumed that sex offenders have significant deficits in empathy that directly contribute to their offending.

In an attempt to address this inconsistency, Jolliffe and Farrington (2004) carried out a systematic review and meta-analysis that included 35 studies comparing questionnaire measures of empathy with official records of offending. The results showed that low cognitive empathy was strongly related to offending, but low affective empathy was only weakly related. The findings of the systematic review also suggested that younger offenders and violent offenders were especially likely to have low empathy. In particular, however, the relationship between low empathy and offending was greatly reduced after controlling for intelligence or socioeconomic status, suggesting either that these might be more important risk factors for offending or that low empathy might mediate the relationship between these risk factors and offending.

Jolliffe and Farrington (2006) created the 20-item Basic Empathy Scale to measure cognitive and affective empathy. Jolliffe and Farrington (2007) showed that boys who self-reported offenses in the previous year had lower empathy, and this was especially the case for affective empathy. Girls who self-reported offenses did not have low empathy, but both boys and girls who self-reported violence did have low empathy. Also, boys and girls who were high-rate offenders also had significantly lower empathy (especially affective) compared to low-rate male and female offenders. These results provide more

convincing evidence of the importance of low empathy in relation to offending.

Eysenck's Theory

Before 1990, the most influential theory of personality and crime was that of Eysenck (1996). He viewed offending as natural and even rational, on the assumption that human beings were hedonistic, sought pleasure, and avoided pain. He assumed that delinquent acts, such as theft, violence, and vandalism, were essentially pleasurable or beneficial to the offender. In order to explain why everyone was not a criminal, Eysenck suggested that the hedonistic tendency to commit crimes was opposed by the conscience, which was viewed as a conditioned fear response.

Eysenck proposed that the conscience was built up in childhood. Each time a child committed a disapproved act and was punished by a parent, the pain and fear aroused in the child tended to become associated with the act by a process of classical (automatic) conditioning. After children had been punished several times for the same act, they felt fear when they next contemplated it, and this fear tended to stop them committing it. According to the theory, this conditioned fear response was the conscience, and it would be experienced subjectively as guilt if the child committed a disapproved act.

On the Eysenck theory, the people who commit offenses are those who have not built up strong consciences because either their conditioning was poor (i.e., they had not been taught right from wrong) or because they have inherently poor conditionability. Poor conditionability is linked to Eysenck's three dimensions of personality: extraversion (E; sensation-seeking, sociability), neuroticism (N; anxiety, depression), and psychoticism (P; tough-mindedness, low empathy).

According to Eysenck, people who are high on E build up conditioned responses less well, because they have low levels of cortical arousal. People who are high on N also condition less well, because their high resting level of anxiety interferes with their conditioning. Also, since N acts as a drive, reinforcing existing behavioral tendencies, neurotic extraverts should be particularly criminal. Eysenck also predicted that people who are high on P would tend to be offenders, because the traits included in his definition of psychoticism (emotional coldness, low empathy, high aggression) were typical of criminals. However, the meaning of the P scale is unclear, and it might perhaps be more accurately labeled as psychopathy. Zuckerman (1989) suggested that it should be termed "impulsive unsocialized sensation seeking."

A review of studies relating Eysenck's personality dimensions to official and self-reported offending concluded that high N (but not E) was related to official offending, while high E (but not N) was related to self-reported offending (Farrington et al., 1982). High P was related to both, but this could have been a tautological result, since many of the items on the P scale are connected with antisocial behavior, or were selected in light of their ability to discriminate between prisoners and nonprisoners. In the prospective longitudinal study of London boys, those high on both E and N tended to be juvenile self-reported offenders, adult official offenders, and adult self-reported offenders, but not juvenile official offenders. These relationships held independently of other criminogenic risk factors, such as low family income, low intelligence, and poor parental child-rearing behavior.

When individual items of the personality questionnaire were studied, however, it was clear that the significant relationships were caused by the items measuring impulsiveness (e.g., doing things quickly without stopping to think). Therefore, it seems likely that research inspired by the Eysenck theory essentially confirms the link between impulsiveness and offending. More recent empirical studies (e.g., Carrasco et al., 2006) and meta-analyses (e.g., Cale, 2006) that have compared Eysenck's dimensions to crime and antisocial behavior have given further support to this conclusion.

The 'Big Five'

In recent years the Five Factor Model (FFM) of personality has replaced Eysenck's three-factor theory as the primary model of personality traits (McCrae and Costa, 1997). The FFM includes the dimensions of E and N (defined similarly to those of Eysenck), as well as Agreeableness (A; altruism, modesty), openness (O; imagination, aesthetic sensitivity), and conscientiousness (C; self-discipline, competence). Research has suggested that Eysenck's P is inversely related to both A and C (Costa et al., 1991). Controversially, McCrae et al. (2000) argued that these personality dimensions are biologically based tendencies that follow intrinsic developmental pathways independently of environmental influences. McCrae and Costa (1997) also found that the same five factors emerged in several different countries.

Relatively few studies have related the 'Big Five' personality dimensions to offending. In the Pittsburgh Youth Study, John et al. (1994) found that the most delinquent boys scored significantly lower on agreeableness, conscientiousness, and openness and significantly higher on extraversion. In Australia, Heaven (1996) reported that agreeableness and conscientiousness were most strongly (negatively) correlated with delinquency. Similarly, Van Dam et al. (2005) in the Netherlands reported that students scored significantly higher on agreeableness and openness than juvenile delinquents.

In their meta-analysis relating a number of personality scales to antisocial behavior, Miller and Lynam (2001) concluded that low agreeableness and low consciousness were significantly related to antisocial behavior, while neuroticism was only weakly related and extraversion was not related. In the most recent meta-analysis, Jones et al. (2011) concluded that agreeableness and conscientiousness were negatively related to antisocial behavior and aggression, while neuroticism was positively correlated. The most important facets that were related to antisocial behavior were hostility and impulsiveness (N), straightforwardness and compliance (A; negatively), and dutifulness and deliberation (C; negatively).

Future Directions

More research is needed on the theoretical linkages between personality constructs and offending, and on the biological bases of personality constructs. Prospective longitudinal

studies are especially needed in which personality traits are measured early in life and related to later offending, in order to investigate the relative importance of biological, individual, family, peer, and school influences on personality and crime. Such studies should especially investigate whether changes in personality are followed by changes in offending.

Research on childhood temperament suggests that potential offenders can be identified early in life and that efforts should be made to change young children who have uninhibited and uncontrolled temperaments. The evidence from prospective longitudinal studies is clear in showing that high impulsivity is related to later offending. This implies that interventions that target impulsiveness, such as child skills training (e.g., Augimeri et al., 2007; Lösel and Beelman, 2006), should be effective in reducing later offending. Empathy has yet to be included in such prospective longitudinal studies, so it is difficult to determine how, or indeed if, low empathy in childhood is related to adult offending. Research on personality and crime should be useful in matching types of people to types of interventions because it seems likely that different personality types would interact differently with different criminological risk factors, and therefore that interventions should be tailored to personalities. Ideally, personality assessment should be included in any intervention program.

There is no doubt that personality influences behavior, including offending behavior. However, there is still the danger that measures such as a 'difficult temperament' or 'low agreeableness' are symptoms of the same underlying construct (e.g., an antisocial personality) as is offending, so that these personality traits cannot be regarded as causes of offending. The one personality trait that probably causes offending is high impulsiveness, and it seems likely that low empathy is also an important cause. If high impulsiveness and low empathy have causal effects on offending, targeting them in intervention programs should have desirable effects in reducing crime and associated types of social problems.

Bibliography

Agnew, R., 2006. Pressured into Crime. Roxbury, Los Angeles.

Augimeri, L.K., Farrington, D.P., Koegl, C.J., Day, D.M., 2007. The SNAP under 12 outreach project: effects of a community based program for children with conduct problems. Journal of Child and Family Studies 16, 799–807.

Bates, J.E., 1989. Applications of temperament concepts. In: Kohnstamm, G.A., Bates, J.E., Rothbart, M.K. (Eds.), Temperament in Childhood. Wiley, Chichester, UK, pp. 321–355.

Beaver, K.M., Walsh, A. (Eds.), 2010. Biosocial Theories of Crime. Ashgate, Farnham, UK.

Cale, E.M., 2006. A quantitative review of the relations between the "Big 3" higher order personality dimensions and antisocial behavior. Journal of Research in Personality 40, 250–284.

Carrasco, M., Barker, E.D., Tremblay, R.E., Vitaro, F., 2006. Eysenck's personality dimensions as predictors of male adolescent trajectories of physical aggression, theft and vandalism. Personality and Individual Differences 41, 1309–1320.

Caspi, A., 2000. The child is father of the man: personality continuities from childhood to adulthood. Journal of Personality and Social Psychology 78, 158–172.

Chess, S., Thomas, A., 1984. Origins and Evolution of Behavior Disorders: From Infancy to Early Adult Life. Brunner-Mazel, NY.

Costa, P.T., McRae, R.R., Dye, D.A., 1991. Facet scales for agreeableness and conscientiousness: a revision of the NEO personality inventory. Personality and Individual Differences 12, 887–898.

Ellis, L., Walsh, A., 2000. Criminology: A Global Perspective. Allyn and Bacon, Boston, MA.

Evans, D., Rothbart, M.K., 2007. Developing a model for adult temperament. Journal of Research in Personality 41, 868–888.

Eysenck, H.J., 1996. Personality and crime: where do we stand? Psychology. Crime and Law 2, 143–152.

Farrington, D.P., Biron, L., LeBlanc, M., 1982. Personality and delinquency in London and Montreal. In: Gunn, J., Farrington, D.P. (Eds.), Abnormal Offenders, Delinquency, and the Criminal Justice System. Wiley, Chichester, UK, pp. 153–201.

Farrington, D.P., Coid, J.W., West, D.J., 2009. The development of offending from age 8 to age 50: recent results from the Cambridge Study in Delinquent Development. Monatsschrift fur Kriminologie und Strafrechtsreform (Journal of Criminology and Penal Reform) 92, 160–173.

Farrington, D.P., Welsh, B.C., 2007. Saving Children From a Life of Crime. Oxford University Press, Oxford, UK.

Frick, P.J., White, S.F., 2008. Research review: the importance of callous-unemotional traits for developmental models of aggressive and antisocial behavior. Journal of Child Psychology and Psychiatry 49, 359–375.

Goldstein, H., Higgins-D'Allessandro, A., 2001. Empathy and attachment in relation to violent vs. nonviolent offense history among jail inmates. Journal of Offender Rehabilitation 32, 31–53.

Guerin, D.W., Gottfried, A.W., Oliver, P.H., Thomas, C.W., 2003. Temperament: Infancy through Adolescence. Kluwer Academic/Plenum, New York.

Hayashino, D.S., Wurtele, S.K., Klebe, K.J., 1995. Child molesters: an examination of cognitive factors. Journal of Interpersonal Violence 10, 106–116.

Heaven, P.C.L., 1996. Personality and self-reported delinquency: analysis of the 'Big Five' personality dimensions. Personality and Individual Differences 20, 47–54.

John, O.P., Caspi, A., Robins, R.W., Moffitt, T.E., Stouthamer-Loeber, M., 1994. The 'Little Five': exploring the nomological network of the five-factor model of personality in adolescent boys. Child Development 65, 160–178.

Jolliffe, D., Farrington, D.P., 2004. Empathy and offending: a systematic review and meta-analysis. Aggression and Violent Behavior 9, 441–476.

Jolliffe, D., Farrington, D.P., 2006. Development and validation of the Basic Empathy Scale. Journal of Adolescence 29, 589–611.

Jolliffe, D., Farrington, D.P., 2007. Examining the relationship between low empathy and self-reported offending. Legal and Criminological Psychology 12, 265–286.

Jolliffe, D., Farrington, D.P., 2009. A systematic review of the relationship between childhood impulsiveness and later violence. In: McMurren, M., Howard, R. (Eds.), Personality, Personality Disorder and Violence. Wiley, Chichester, UK, pp. 41–61.

Jones, S.E., Miller, J.D., Lynam, D.R., 2011. Personality, antisocial behavior, and aggression: a meta-analytic review. Journal of Criminal Justice 39, 329–337.

Kagan, J., Reznick, J.S., Snidman, N., 1988. Biological bases of childhood shyness. Science 240, 167–171.

Lahey, B.B., Loeber, R., Waldman, I.D., Farrington, D.P., 2006. Child socioemotional dispositions at school entry that predict adolescent delinquency and violence. Impuls: Tidsskrift for Psykologi 3, 40–51.

Lösel, F., Beelman, A., 2006. Child social skills training. In: Welsh, B.C., Farrington, D.P. (Eds.), Preventing Crime: What Works for Children, Offenders, Victims and Places. Springer, New York, pp. 33–71.

Lynam, D.R., Moffitt, T.E., 1995. Delinquency and impulsivity and IQ: a reply to Block. Journal of Abnormal Psychology 104, 399–401.

McCrae, R.R., Costa, P.T., 1997. Personality trait structure as a human universal. American Psychologist 52, 509–516.

McCrae, R.R., Ostendorf, F., Angleitner, A., Hrebickova, M., Avia, M.D., Sanz, J., Sanchez-Bernardos, M.L., Kusdil, M.E., Woodfield, R., Saunders, P.R., Smith, P.B., 2000. Nature over nurture: temperament, personality, and life span development. Journal of Personality and Social Psychology 78, 173–186.

Miller, J.D., Lynam, D., 2001. Structural models of personality and their relation to antisocial behavior: a meta-analytic review. Criminology 39, 765–798.

Moffitt, T.E., 1990. The neuropsychology of juvenile delinquency: a critical review. In: Tonry, M., Morris, N. (Eds.), Crime and Justice, vol. 12. University of Chicago Press, Chicago, IL, pp. 99–169.

Morgan, A.B., Lillienfeld, S.O., 2000. A meta-analytic review of the relation between antisocial behavior and neuropsychological measures of executive function. Clinical Psychology Review 20, 113–136.

Neumann, A., Barker, E.D., Koot, H.M., Maughan, B., 2010. The role of contextual risk, impulsivity, and parental knowledge in the development of adolescent antisocial behavior. Journal of Abnormal Psychology 119, 534–545.

Niv, S., Tuvblad, C., Raine, A., Wang, P., Baker, L.A., 2012. Heritability and longitudinal stability of impulsivity in adolescence. Behavior Genetics 42, 378–392.

Pratt, T.C., Cullen, F.T., 2000. The empirical status of Gottfredson and Hirschi's general theory of crime. Criminology 38, 931–964.

Pratt, T.C., Cullen, F.T., Blevins, K.R., Daigle, L., Unnever, J.D., 2002. The relationship of attention deficit hyperactivity disorder to crime and delinquency: a meta-analysis. International Journal of Police Science and Management 4, 344–360.

Roberts, B.W., Del Vecchio, W.F., 2000. The rank-order consistency of personality traits from childhood to old age: a quantitative review of longitudinal studies. Psychological Bulletin 126, 3–25.

Sanson, A., Smart, D., Prior, M., Oberklaid, F., 1993. Precursors of hyperactivity and aggression. Journal of the American Academy of Child and Adolescent Psychiatry 32, 1207–1216.

Schwartz, C.E., Snidman, N., Kagan, J., 1996. Early childhood temperament as a determinant of externalizing behavior in adolescence. Development and Psychopathology 8, 527–537.

Skeem, J.L., Cooke, D.J., 2010. Is criminal behavior a central component of psychopathy? Conceptual directions for resolving the debate. Psychological Assessment 22, 433–445.

Slutske, W.S., Moffitt, T.E., Poulton, R., Caspi, A., 2012. Undercontrolled temperament at age 3 predicts disordered gambling at age 32: a longitudinal study of a complete birth cohort. Psychological Science 23, 510–516.

Van Dam, C., Janssens, J.M.A.M., De Bruyn, E.E.J., 2005. PEN, Big Five, juvenile delinquency and criminal recidivism. Personality and Individual Differences 39, 7–19.

White, J.L., Moffitt, T.E., Caspi, A., Bartusch, D.J., Needles, D.J., Stouthamer-Loeber, M., 1994. Measuring impulsivity and examining its relationship to delinquency. Journal of Abnormal Psychology 103, 192–205.

Wilson, J.Q., Herrnstein, R.J., 1985. Crime and Human Nature. Simon and Schuster, New York.

Zuckerman, M., 1989. Personality in the third dimension: a psychobiological approach. Personality and Individual Differences 10, 391–418.

Personality and Economics

Daniel D Santos, Universidade de São Paulo, São Paulo, Brazil

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mainstream modern economics studies the decision behavior of rational agents. In a prototypical model, individuals take four sets of *primitives* (preferences, information, technology, and incentive constraints) as given, and choose the best feasible action in each situation. This article summarizes how economic theories have incorporated findings and ideas from personality psychology in each of these four types of primitives. In the discussion, it will become clear that the introduction of some aspects of personality into standard mainstream economic models has seriously challenged their scientific validity, it being one of the most controversial topics in the field.

Introduction

Modern economics aims to study the decision behavior of individuals (and more generally, of economic agents). The foundation of the field is the 'theory of rational choice,' which states that agents are able to order their possible actions into a choice set according to a unidimensional criterion. (This criterion is called utility, when the decision-maker is the consumer or the household.) In a typical decision-making problem, four sets of objects are taken as given by the agent when choosing the best action to take: (1) Individual *Preferences*, which comprise tastes, attitudes toward risk, and the intertemporal discount factor (the rate at which the individual is willing to trade present pleasure for future pleasure), among others; (2) *Information and Beliefs*, meaning that agents are assumed to act on what they think is best for them (instead of on what is truly best); (3) *Technology*, that determines what individuals can do in a given situation; and (4) *Incentives*, which modify the payoffs of specific actions and contribute, together with preferences, to forming the individual ranking of actions coherent with the aforementioned criterion. These four objects are called the *primitives* of the decision behavior in an economic model.

Psychology, in turn, investigates the mental processes that determine behavior in general. As we will discuss in this article, personality traits may affect the decision problem in different ways, sometimes being central to the determination of the observed economic behavior. Whenever possible, this article attempts to make connections between the ongoing discussion in economics and the personal–situational debate that dominated the field of psychology of personality throughout the twentieth century. Although the debate is grounded in great differences about how people conduct themselves, it was propelled by empirical evidence that suffers from a fundamental identification problem: there are two causes for a single effect. One must assume that one of the sources is of little relevance in order to learn about the other from observed effect. Understanding the motivations behind each theoretical view and how they were adapted to have empirical content is crucial to further interpret how individual intrinsic characteristics (which I call personality traits) affect behavior, and how their importance compares to that of environmental variables. The next section briefly exposes the challenges involved in empirically identifying causal relations between personality attributes and observed decision behavior.

The third section of this article discusses how economists have included personality traits in the *primitives* of economic behavior, and the related empirical evidence. The last section concludes and speculates about the future of the field.

The Fundamental Identification Problem in the Search for Causal Relations between Personality Traits and Individual Economic Outcomes

Empirical attempts at estimating the structure of personality and the role of its elements in the determination of observed behavior typically face two types of obstacles. First, personality is a multidimensional concept that involves multiple categories of objects, each of them with multiple elements. In an influential study, Roberts (2006) concludes that personality is composed of traits (e.g., the 'Big Five' factors), motives and values (such as goals and interests), abilities (e.g., factor 'g,' verbal and spatial), and narratives (stories, significant memories, and scripts). Other approaches use different taxonomies, but usually describe personality in a multidimensional fashion as well. Existent heterogeneity in observed behavior is generally the result of heterogeneity in personality and the prevalent context/situation at the moment the behavior was registered. If one could measure all relevant aspects of personality and context, the causal effect of an element of personality (let us call it 'P') in the determination of a given behavior or outcome (y) could be computed by the direct comparison between the distribution of the respective behavior or outcome in subsamples of the population characterized by different levels of this element of personality, keeping other components of personality (Q) and environmental conditions (C) constant (i.e., the comparison between $f(Y|P=p;\ Q=q,\ C=c)$ and $f(Y|P=p';\ Q=q,\ C=c)$). In reality many relevant aspects of personality and context are not measurable, and only subsets Q^o of Q and C^o of C are effectively observed. Thus, the comparison between $f(Y|P=p;\ Q^o=q^o,\ C^o=c^o)$ and $f(Y|P=p';\ Q=q,\ C=c)$ can only be interpreted as a causal consequence of changes in P on Y if one assumes that the unobserved parts of Q and C are constant or statistically unrelated to P. To

illustrate this point, suppose that Y is the final GPA obtained in high school and P is a measure of IQ. The fact that people with low IQ have lower average GPA can only be interpreted as a causal effect if other personality elements and the context were similar across people with different IQ levels. Conversely, if IQ were systematically related to self-esteem, it would not be clear if the differences in GPA as measured by the IQ level were a consequence of differences in IQ or in self-esteem.

As stated above, economics aims to explain the decision behavior through four sets of *primitives*. Each of these *primitives* contains elements of environment (context/situation) and personality. Theories and models that emphasize the importance of variations in the environment in rationalizing heterogeneity in behavior tend to use the hypothesis that personality is constant or displays only minor differences among agents, so that differences in behavior could be mostly attributed to differences in the environment (and the opposite holds for models that explain economic behavior through individual differences in personality).

The second identification problem holds that most of the personality components of interest are not directly measurable. Both classical test theory and item response theory psychometrics consider behavior to be a function of a low-dimensional set of latent personality traits, and use statistical filters to compute the individual value of such traits from a set of observed behavior measurements. Two conditions are necessary for this strategy to work. On the one hand, behaviors must be an invertible function of elements of personality, so that individuals with identical observed behaviors have the same personality structure (ideally, once controlled for environmental conditions). On the other hand, if two or more personality components determine the same observed behavior, identification of the individual personality structure depends on the number of measured behaviors being at least as large as the number of personality elements one wishes to estimate. Since few theorists are willing to postulate ex-ante the number of relevant components of the personality structure, estimation of these components relies only on empirical regularities of the data revealed through the chosen statistical filters (e.g., the Big Five factors of personality, revealed by linear factor analysis applied to inventories based on the lexical hypothesis (see Allport and Odbert, 1936)).

Relevant for our discussion is to note that elements of personality were incorporated into mainstream economics in different ways, depending on the underlined economic question. Sometimes, personality traits are assumed to be stable, in order to allow for a deeper analysis of the relationship between context variables on the decision behavior. In other circumstances, personality plays a central role in determining behavior, being even treated as subject to modifications.

The Inclusion of Personality in Economic Models

Research in psychology generally seeks to explain the relationship between mental functioning and (observed) behavior. An economic model that realistically describes the individual decision behavior would likely be incomplete if it ignored the findings of personality psychologists. In fact, aspects of personality are an important ingredient of all *primitives* of economic models. This section shows how personality relates with each of the four *primitives* described in the Introduction.

Personality and Preferences

Individual preferences are a system of relations governed by intrinsic characteristics of economic agents, which inform how the outcomes of actions translate into payoffs (be it utility, satisfaction, or happiness). In a mainstream economic model, preferences determine how an individual ranks the existing actions in a given situation. Aspects of preferences include *tastes*, *attitudes toward risk*, and *intertemporal rate of discount* (the rate at which an individual is willing to trade present satisfaction for future satisfaction), among others. In principle, all personality traits and abilities that affect the way individuals order their actions are also part of preferences.

Although the importance of tastes in determining the allocation of scarce resources in an economy has been noticed since Adam Smith's work (see Rosenberg, 1968), it was after the utilitarian revolution that it became a key feature of decision behavior models. The statements that "agents do what is best for them" and that "aggregate behavior is no more than the result of a set of individual actions" are central in both microeconomic and macroeconomic models.

It is important to mention, however, that the supremacy of the rational choice theory (RCT) in economics did not come without controversy. As simple as it is, the RCT departs from two basic axioms (*Completeness*: agents can compare any pair of existing actions, A and B, and rank their respective payoffs in terms of their utilities; and *Transitivity*: for every triple (A, B, and C), if A's payoff is at least as high as B's, and B's payoff is at least as high as C's, then A's payoff is at least as high as C's.) to guarantee the existence of an individual ranking of rewards, each of which correspond to an existing action. In a given situation, a subset of such actions is available, and individuals choose the best action in this subset. The problem with this approach is that it is tautological without further restrictions (see Sen, 1973; Marx, 1867), meaning that every observed behavior could be explained by some structure of preferences in every specific context. The empirical content of economic models typically comes from additional assumptions on preferences, which have been subject to debate.

In fact, most economic models assume that preferences are to some extent *stable*, so that differences in the decision behavior are mostly due to differences in the environment (e.g., technology and incentives). Although stability has been implicitly assumed since the first empirical economic models, it was with Stigler and Becker's (1977) influential paper that the formal role of this hypothesis was established. The authors advocated not only that individual preferences are fully stable, both throughout the life cycle and across state contingencies, but also that heterogeneity of preferences among economic agents is small and negligible in terms of explaining the observed heterogeneity in behavior. The authors respond to common criticisms that addiction, advertisements, cultural changes, fashion, and the invention of new goods would transform tastes, by saying that it is not the tastes that are modified by these factors but rather the information and learning about the actions and their payoffs.

The discussion around *stability* that followed the Stigler–Becker argument (and much of Becker's subsequent work) resembles to a large extent the personal–situation debate in psychology. In both cases, heterogeneity in individual traits and in context generates heterogeneity in behavior, but as discussed in the previous section, one must impose restrictions in one type of heterogeneity in order to recover the other from the observed behavior. An economic theory that stated individual decision behavior is mostly driven by tastes is of no use for policymaking or for understanding the role of economic variables (such as prices and income) in the process. In general, economists study decision behavior in order to investigate which type of interventions could improve welfare (at the individual or aggregate level). Almlund et al. (2011) use a similar argument to suggest that "The debate over the relative importance of person and situation in the 1960s and 1970s reflected deeper currents in psychology and social science more generally, that still run strong. (…) Behaviorism (…) posited that experience explains all aspects of behavior. There was the widely held notion that situation and experience were all powerful—that people were born as blank slates. This captured the interventionist spirit of the times."

Although the textbook hegemony and the Nobel Prize (to both Stigler – in 1982 – and Becker – in 1992) confer a privileged position upon their point of view, their antagonists are no less influential. Behavioral economists and psychologists such as Daniel Kahneman (also a Nobelist, 2002) and Amos Tversky documented the existence of a type of cognitive bias called *framing effect*, i.e., the fact that the agent's decision over the same alternatives depends on the way these alternatives are offered to the agent. (Many *paradoxes* of preferences studied in the literature are the result of a framing effect. See Tversky and Kahneman, 1981.) In a similar vein, Kahneman, Tversky, Robert Thaler, and coauthors investigated the *endowment effect* (see Thaler, 1980; Kahneman and Tversky, 1979; Kahneman et al., 1990). In a context of consumption, the endowment effect implies that a person is more willing to pay to retain a good than to acquire it, suggesting that attachment and loss aversion are important aspects of consumer behavior.

The axiom of transitivity is a second major source of disagreement between mainstream economists and their critics. May (1954), and especially Tversky (1969) exposed individuals to specific series of risky gambles in a laboratory experiment to show that individuals seemed to systematically make choices in an intransitive way. Tversky's paper inaugurated a prolix research agenda in the field of behavioral economics in which many researchers tried to prove transitivity is unrealistic, methodically replied by mainstream economists (mostly on the basis of the weak external validity and adequacy to reality of such laboratory experiments).

Tversky rationalized his findings by proposing an alternative decision theory for risky actions. According to him, agents first order such actions by their probabilities of positive outcomes and expected pleasure, and afterward by expected loss (lexicographic semiorder interpretation). This was a sharp contrast with the prevailing expected utility (EU) theory, first proposed in 1738 by Daniel Bernoulli (1954) (as a response to the St. Petersburg Paradox (posed by his uncle, Nicolas, in 1713)), and broadly used to model decisions under uncertainty. The EU hypothesis states that agents compare the expected payoff of every possible action in order to decide what to do, so that if two states of nature deliver a gain and a loss of the same magnitude with equal probability, they should cancel each other out. Indeed, risk aversion and decision under uncertainty are a third major source of controversy about preferences. Tversky's original explanation was generalized in a joint work with Kahneman (1979) who originated the *Prospect Theory*, within which agents have reference points and behave in order to minimize expected loss (instead of maximizing EU). Typically, losses hurt more than the pleasure associated to gains (loss aversion, which was also formulated by Akerlof and Yellen (1985), but in a slightly different way). Prospective theorists also seek to explain how agents set the reference utility point, a topic in which *focusing effects* and *anchoring effects* (giving disproportional importance to a specific piece of information and to the first information when setting the reference point, respectively) play important roles.

Behavioral economics has not only challenged the mainstream view about decision-making in situations involving *risk* (i.e., when actions' payoffs are unknown ex-ante, but agents know their respective probabilities), but also proposed a new theory about choice involving *uncertainty* (when even probabilities are unknown). Tversky and Fox (1995), for instance, show that if two lotteries are equally valuable to agents when evaluated in isolation but the underlined probabilities of one of them are unknown, then individuals systematically choose the other when both alternatives are offered. After this work, risk (or loss) aversion and ambiguity aversion were considered two different traits of preferences in many papers in economics.

A related but different cause of divergence in terms of preferences is the description of *intertemporal decision behavior*. The mainstream view usually assumes additive discounted intertemporal preferences, meaning that: (1) future satisfaction does not depend upon current satisfaction and depends on the current state of nature only to the extent that it affects feasible actions in the future; and (2) the rate at which individuals are willing to trade present satisfaction for satisfaction in the next period is the same as in any two subsequent periods of life. (Some authors put it as a requirement for preferences to be *dynamically consistent*, i.e., for the optimal decision in a contingency decision plan not to depend on the time or history of realizations of the contingencies. Lack of dynamic consistency (such as that produced by hyperbolic discounting) may generate opportunities of *arbitrage* in the credit and financial markets.) Critics of this view generally argue that people are *present-biased* (see Thaler and Shefrin, 1981), and that the willingness to trade the present for next-period pleasure is much lower than the respective rate between two subsequent future periods. This fact was mathematically formalized through the *hyperbolic discounting utility* theory (Laibson, 1997; Akerlof, 1991), and has been related to psychological concepts of delay of gratification, self-regulation, and self-control (Akerlof, 2002).

Finally, economists have progressively adapted the pure rational choice formulation to account for social aspects of decision. The study of social preferences can be divided into two major groups: (1) situations in which the individual's attitudes affect others' welfare and actions (which may alter the individual's behavior, if he/she cares about the other's welfare, or if the other's actions have consequences for his/her satisfaction); and (2) situations in which the other's (established) status

directly affects the individual's preferences (see Chapters 1 (by Kolm) and 8 (by Fehr and Schmidt) of the *Handbook of the Economics of Giving, Altruism and Reciprocity* (2006), for a recent and broader overview).

In the real world, individuals often give gifts and help people with no expectations of receiving something in return. In an early formulation, Becker (1974) says individuals acquire a psychic good ('a warm glow'), so that gifts and help are no evidence of altruism. Levine (1998) proposes a model in which other people's satisfaction directly affects individual welfare, introducing altruism directly into the structure of preferences. Akerlof (1982) investigated why some firms would pay above-market wages to their workers, and introduced the idea of reciprocity, fairness, and attachment to social norms to modern economic models. According to him, "a worker's effort depends upon the norms determining a fair day's work." Firms would then pay higher wages in order to affect those norms, and receive higher effort in exchange. (Other authors (e.g., Shapiro and Stiglitz, 1984) examined the same fact, which is now known as the 'wage-efficiency theory.' In his Nobel Prize lecture (2001), Akerlof advocates the inclusion of *reciprocity*, *fairness*, and *adherence to group norms* in other contexts involving apparently *irrational* decision behavior.)

In a different direction, Simon (1947) and Akerlof (1991) studied the role of *authority and obedience* in decision-making at the organizational level. The higher-ranked professional usually expects obedience, and the subordinate is usually willing to obey. Authority and obedience therefore influence the actions of all actors, and is the cause and effect of the establishment of goals and values in an organizational environment, as well as of the socioprofessional relationships and reward systems.

The role of distributive issues in shaping preferences has also been an active topic of research with interactions with the psychology of personality. On the one hand, there is an old tradition of economists (going back at least to Veblen, 1899) who argue that an individual's welfare depends not only on his/her income (and more generally, resources and opportunities), but also on relative income (or relative position in society), bringing the concept of *envy* to the heart of the decision problem. On the other hand, Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), among others, suggest that inequality hurts an individual's welfare regardless of his/her social position, proposing *inequality aversion* should also be included in the structure of individual preferences.

Many papers have surveyed the empirical evidence concerning the validity of mainstream and alternative assumptions about preferences, and their relation with personality traits. Most of the work on this matter was developed through laboratory experiments, among which Rabin (1998), Sendhil Mullainathan and Thaler (2001), and Camerer (2006) are good references. DellaVigna (2009) provides new evidence coming from nonlaboratory data sets.

Beliefs and Information

Economic theory has truly experienced a revolution over the last 40 years, in which the availability of information, learning, and belief formation processes was definitely incorporated into the description of decision behavior. Succinctly speaking, researchers have realized that individuals make decisions based on what they know about the world and think is optimal for them (even though this may not in fact be true). Knowledge of the world depends on restrictions on information disclosure and processing, on the beliefs that agents have about the future and about how other people react to their actions, and both available information and beliefs are subject to learning over time.

Herbert Simon (also a Nobelist, 1978) has argued that an individual's capacity to process and order the payoffs of all possible actions in a given context is limited, and agents therefore develop strategies to overcome such difficulties. Simon's contribution became what is now known as the *Bounded Rationality* agenda. In dealing with information processing and availability limitations, individuals build strategies that usually involve subjective beliefs about the result of their actions; the basis of the *Subjective Expected Utility* theory.

The inability to process all available information is also the focus of the literature on *Myopic Behavior*, which explores the fact that agents do not forecast or compute all possible consequences of their actions. A well-studied example of myopic behavior is the monetary illusion (the fact that in some circumstances agents decide based on the nominal value – instead of the real value – of economic variables). Both bounded rationality and myopic behavior have a strong component of cognitive limitation, and may be reinforced by characteristics such as self-control and anxiety (see Fisher (1928), Fehr and Tyran (2001), and Shafir et al. (1997) for both theoretical end empirical discussions on myopic behavior).

On a different front, behavioral economists document and theorize about the systematic formation of incorrect beliefs. Overconfidence (part of self-perception and the locus of control) distorts expected payoffs and directly affects decisions. Excessive reliance on anecdotic situations and the extrapolation of its consequences to other contexts is a second source of biased beliefs. Underestimation of one's ability to adapt to new circumstances (which sometimes justify overreaction to adverse shocks) also changes expectations about the payoffs of many actions. These three types of incorrect beliefs are all surveyed in DellaVigna (2009). Overall, many of the documented cognitive biases are potential sources of incorrect beliefs, with consequences on decision making.

Misinterpretation of the information may also produce incorrect beliefs, being also related to a difficulty in information processing (related therefore to cognitive limitations). A popular topic in empirical work in economics is self-selection and survivorship (see Heckman, 1979; Heckman and Sedlacek, 1985; Elandt-Johnson and Johnson, 1999). People choose what to do based on the observation of a world of winners. Observed distributions of wages and profits, for example, do not reflect the respective ex-ante payoffs of prospective workers and firms in their decisions on supplying labor or starting a business. The outcomes of the losers were not there, because they have been eliminated from the game by market forces. Researchers should be able to separate the statistical overoptimism, produced by the misinterpretation of reality, from the true overoptimism, determined by a personality trait. Other examples of statistical biases (studied in economics) that may distort beliefs are statistical discrimination (Becker, 1993) and regression to the mean (Galton, 1869; Samuels, 1991).

Technology and Incentives

In many situations, the individual decision problem is constrained by incentive restrictions imposed by others' responses to his/her actions. Such responses, in their turn, are also the result of the rational decisions of other agents, and therefore depend on preferences, beliefs, and technologies that characterize their decision problems. Since others' preferences and beliefs are deeply influenced by their own personality traits, incentive constraints are influenced by them as well. Bowles et al. (2001) go a step further to study how certain specific personality attributes make people more susceptible and reactive to changes in the structure of incentives. In their paper, they argue that the workers with external locus of control and low self-efficacy expend less effort in their jobs and do not react to incentives, since they do not believe that their actions are effective. Individuals who value future pleasure and have a sense of shame in being without a job are also more sensitive to incentives. In a different approach, Tversky and Kahneman (1974) highlight the importance of social pressure and persuasion in workplace performances and voting decisions.

Labor earnings are the main portion of personal income, an important technological constraint (to the extent that available income strongly restricts the actions one can choose in many contexts). Besides being affected by incentive constraints (which are subject to the influence of personality), wages are usually taken as a function of the set of individual productive attributes, which may include personality traits as well. In fact, a promising recent branch of economics literature has explored this argument, and shown that the so-called *noncognitive* abilities and traits are at least as important as the *cognitive* ones in the formation of wages (see Heckman et al., 2006).

In a comprehensive survey of the empirical evidence of the predictive power of personality traits, Almlund et al. (2011) show that conscientiousness and its facets, as well as some facets of neuroticism (the locus of control and self-esteem) are important determinants for job performance and wages.

Previous literature about individual outcomes related to the stock of human capital also investigated its impact on health, education, family formation, and involvement with violence. Being part of human capital, it is not surprising that many researchers documented the relationship between personality traits and these outcomes. Almlund et al. (2011) summarized this evidence as follows:

- *Education*: Openness to experience and conscientiousness consistently appear to be important predictors of educational attainment, although most empirical studies cannot convincingly prove this relationship to be causal. The locus of control (sometimes taken as a facet of neuroticism) is also related to attainment, and this evidence appears in different countries. Regarding grades, conscientiousness (especially self-control) is the most robust predictor, but the authors warn against the risk of teachers using grades to directly punish/reward attributes not related to academic achievement. Studies that used standardized test scores roughly confirmed the relevance of conscientiousness, and stressed the role of attention, possibly one of its facets.
- *Health*: The facets of openness to experience, conscientiousness, and agreeableness are associated with longer lives, whereas the opposite happens with aspects of neuroticism. Agreeableness and conscientiousness also correlate with involvement with drugs, tobacco, and alcohol.
- *Violence*: Self-control seems to be the main individual trait that determines crime and involvement with violent episodes.

When understood as part of the individual stock of human capital, personality attributes are assumed to be subject to investments in their improvement. In sharp contrast to the way mainstream economics treats these characteristics when talking about their influence in other primitives, here personality is not only treated as changeable, but also people have a certain amount of control over their evolution across the life cycle. Cunha et al. (2010) built and estimated an economic model of human capital accumulation in which individuals could invest in both cognitive and noncognitive characteristics. They conclude that both types of attributes are relevant determinants of many individual outcomes, and that people with higher initial levels of (cognitive and noncognitive) skills benefited more from an investment in human capital ('skills beget skills'). Public and private policies and programs have also proved to be effective in changing personality traits (Heckman et al., 2010; Gottschalk, 2005; Diamond et al., 2007).

Conclusions

The goal of this article was to summarize the economic theories and empirical evidence that dedicate to personality a predominant role in the determination of decision behavior and economic outcomes. Most of the discussion cited herein has emerged from and developed predominantly among the so-called 'behavioral economists,' who have systematically challenged the mainstream view in economics, in which economic behavior is mainly driven by environmental variables. The debate is old, but recent research has boosted the field and brought new empirical evidence.

While the textbook economic models assume that preferences (intrinsic personal characteristics, which include personality traits) are stable over time and display little heterogeneity across agents, behavioral economics relativize this scenario by showing that circumstances often affect the way individuals order their preferred actions. While mainstream economics suggests that individuals share common beliefs about the world, its critics argue in favor of subjective beliefs and an imperfect ability of processing information. In spite of the existing divergences, most economists seek to identify environmental variables that affect behavior, and my interpretation of the controversy concerns whether the environment only affects decision through technology (restrictions that constrain individuals' sets of choice), or whether it can also directly influence the way agents rank their available actions' payoffs in order to decide what to do.

A third agenda relating personality to economics has emerged in the last 20 years, incorporating personality traits into the well-established human capital theory. Human capital is usually defined as the set of productive traits and abilities which are then used to produce goods (with wages as the

respective reward), health, noninvolvement with violence, family stability, and well-being in general. Contrary to other economic views of personality, families and the state can invest in the improvement of the personality traits of their members, and do so in a rational way, aiming to maximize intertemporal satisfaction.

In the present context, we know that noncognitive attributes as a whole contribute approximately as much as cognitive aspects to academic achievement, as measured by grades, dropout rates, and educational levels achieved. Noncognitive characteristics are also rewarded in the labor market in the form of higher wages and shorter unemployment spells. In other investigated dimensions (health, involvement with crime and violence), the role of noncognitive attributes surpasses that of abstract intelligence, providing higher levels of satisfaction. Some researchers add to these facts the interdisciplinary evidence that cognitive traits crystallize earlier in life than the noncognitive ones, to advocate that public policy oriented toward equality should prioritize investments in noncognitive skills, if the target group is composed of adolescents and adults.

Finally, history has shown that many theoretical disagreements can survive only while data are not available to test their respective predictions. Measuring personality traits has always been a difficult task, but the intensification of the interdisciplinary dialogue may bring new methods and empirical content to the existing theories that help to overcome these difficulties.

Acknowledgments

I am thankful to Fernanda Estevan for her helpful comments on the first version of this article. I also benefited from thoughtful discussions with Ricardo Primi, Mirela Carvalho, Chris Soto and Richard Roberts.

See also: Behavioral Economics; Consumer Psychology; Decision Making, Psychology of; Expectations, Economics of; Information, Economics of; Money and Finances, Psychology of; Organizational Behavior, Psychology of; Pay, Compensation, and Performance, Psychology of; Personality and Educational Outcomes; Personality and Risk Taking; Personality and Values at Work; Personality, Trait Models of; Rational Choice and Organization Theory.

Bibliography

Akerlof, G., 2002. Behavioral macroeconomics and macroeconomic behavior. American Economic Review 92 (3), 411–433.

Akerlof, G.A., Yellen, J.L., 1985. Can small deviations from rationality make significant differences to economic equilibria? American Economic Review 75 (4), 708–720.

Akerlof, G., 1982. Labor contracts as partial gift exchange. Quarterly Journal of Economics 97 (4), 543–569.

Akerlof, G., 1991. Procrastination and obedience. American Economic Review 81 (2), 1–19.

Allport, G.W., Odbert, H.S., 1936. Trait-Names: A Psycho-Lexical Study. Albany, Psychological Review Company, New York.

Almlund, M., Duckworth, A.L., Heckman, J., Kautz, T.D., 2011. Personality psychology and economics. In: Hanushek, E.A., Machin, S., Woessmann, L. (Eds.), Handbook of the Economics of Education. Elsevier, Amsterdam, pp. 1–181.

Becker, G., 1993. Nobel lecture: the economic way of looking at behavior. Journal of Political Economy 101 (3), 385–409.

Becker, G., 1974. A theory of social interactions. Journal of Political Economy 82, 1063–1093.

Bernoulli, D., 1954. Exposition of a new theory on the measurement of risk. Econometrica 22, 23–36. Translation of Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. Papers Imperial Academy of Sciences of St. Petersburg 5, 175–192.

Bolton, G., Ockenfels, A., 2000. ERC: a theory of equity, reciprocity, and competition. American Economic Review 90 (1), 166–193.

Bowles, S., Gintis, H., Osborne, M., 2001. The determinants of earnings: a behavioral approach. Journal of Economic Literature 39 (4), 1137–1176.

Camerer, C., 2006. Behavioral economics. In: Blundell, R., Newey, W., Persson, T. (Eds.), Advances in Economics and Econometrics: Theory and ApplicationsNinth World Congress, second ed. Cambridge University Press, Cambridge and New York, pp. 181–214.

Cunha, F., Heckman, J., Schennach, S., 2010. Estimating the technology of cognitive and noncognitive skill formation. Econometrica 78 (3), 883–931.

DellaVigna, S., 2009. Psychology and economics: evidence from the field. Journal of Economic Literature 47 (2), 315–372.

Diamond, A., Barnett, S., Thomas, J., Munro, S., 2007. Preschool program improves cognitive control. Science 318 (5855), 1387–1388.

Elandt-Johnson, R., Johnson, N., 1999. Survival Analysis (Biometry); Failure Time Data Analysis; Mortality; Medical Statistics; Statistical Methods. Wiley, New York.

Fehr, E., Schmidt, K.M., 2006. The economics of fairness, reciprocity and altruism–experimental evidence and new theories. In: Kolm, S.C., Mercier Ythier, J. (Eds.), Handbook of the Economics of Giving, Altruism and Reciprocity, vol. 1, pp. 615–691.

Fehr, E., Tyran, J.R., 2001. Does money illusion matter? American Economic Review 1239–1262.

Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Quarterly Journal of Economics 114 (3), 817–868.

Fisher, I., 1928. The Money Illusion. Adelphi, New York.

Galton, F., 1869. Hereditary Genius. Macmillan and Company, New York.

Gottschalk, P., 2005. Can work alter welfare recipients' beliefs? Journal of Policy Analysis and Management 24, 485–498.

Heckman, J., 1979. Sample selection bias as a specification error. Econometrica 47 (1), 153–161.

Heckman, J., Sedlacek, G., 1985. Heterogeneity, aggregation, and market wage functions: an empirical model of self-selection in the labor market. Journal of Political Economy 93 (6), 1077–1125.

Heckman, J., Stixrud, J., Urzua, S., 2006. The Effects of Cognitive and Non-Cognitive Abilities on Labor Market Outcomes and Social Behavior. National Bureau of Economic Research, Working Paper, Age: Anthropological Aspects.

Heckman, J.J., Humphries, J.E., Mader, N., 2010. The Ged. In: Hanushek, E.A., Machin, S., Wößmann, L. (Eds.), Handbook of the Economics of Education. Elsevier, Amsterdam.

Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47 (2), 263–291.

Kahneman, D., Knetsch, J.L., Thaler, R.H., 1990. Experimental tests of the endowment effect and the Coase theorem. The Journal of Political Economy 98 (6), 1325–1348.

Kolm, S.C., 2006. Introduction to the economics of giving, altruism and reciprocity. In: Handbook of the Economics of Giving, Altruism and Reciprocity, vol. 1, pp. 1–122.

Laibson, D., 1997. Golden eggs and hyperbolic discounting. Quarterly Journal of Economics 112 (2), 443–477.

Levine, D.K., 1998. Modeling altruism and spitefulness in experiments. Review of Economic Dynamics 1, 593–622.

Marx, K., 1867. Das Kapital: A Critique of Political Economy. Gateway Editions, 1961.

May, K., 1954. Intransitivity, utility, and the aggregation of preference patterns. Econometrica 22, 1–13.

Mullainathan, S., Thaler, R., 2001. Behavioral economics. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Oxford University Press, New York and Oxford, pp. 1094–1100.

Rabin, M., 1998. Psychology and economics. Journal of Economic Literature 36 (1), 11–46.

Roberts, B., 2006. Personality development and organizational behavior. In: Staw, B. (Ed.), Research on Organizational Behavior, vol. 27. Elsevier Science/JAI Press, pp. 1–41.

Rosenberg, M.J., 1968. Hedonism, inauthenticity, and other goals toward expansion of a consistency theory. In: Abelson, R.P. (Ed.), Theories of Cognitive Consistency: A Sourcebook. Rand McNally, Chicago.

Samuels, M.L., 1991. Statistical reversion toward the mean: more universal than regression toward the mean. The American Statistician 45 (4), 344–346.
Sen, A., 1973. Behaviour and the concept of preference. Economica 40 (159), 241–259.
Shafir, E., Diamond, P., Tversky, A., 1997. Money illusion. Quarterly Journal of Economics 112 (2), 341–374.
Shapiro, C., Stiglitz, J.E., 1984. Equilibrium unemployment as a worker discipline device. American Economic Review 74 (3), 433–444.
Simon, H.A., 1947. Administrative Behavior, third ed. Macmillan, New York.
Simon, H.A., 1978. Rationality as process and as product of thought. American Economic Review 68 (2), 1–16.
Stigler, G.J., Becker, G., 1977. De gustibus non est disputandum. American Economic Review 67 (2), 76–90.
Thaler, R., 1980. Toward a positive theory of consumer choice. Journal of Economic Behavior and Organization 1, 39–60.
Thaler, R., Shefrin, H., 1981. An economic theory of self-control. Journal of Political Economy 89 (2), 392–406.
Tversky, A., 1969. Intransitivity of preferences. Psychological Review 76 (1), 31–48.
Tversky, A., Fox, C.R., 1995. Weighing risk and uncertainty. Psychological Review 102 (2), 269–283.
Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. Science 185 (4157), 1124–1131.
Tversky, A., Kahneman, D., 1981. The framing of decisions and the psychology of choice. Science 211 (4451), 453–458.
Veblen, T., 1899. The Theory of the Leisure Class. Macmillan, New York.

Personality and Educational Outcomes

Arthur E Poropat, School of Applied Psychology, Griffith University, Mount Gravatt, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Among the most important outcomes for individuals and countries are those linked with education, and recent research has established that these outcomes are reliably correlated with personality measures based upon the five-factor model. Conscientiousness measures are the most strongly associated with educational outcomes at all educational levels, apparently due to conscientiousness being based upon effortful control and links with self-regulation of learning activities. Openness is next most important, resulting from associations with curiosity and motivation to learn. Correlations of academic performance with self-rated agreeableness, extraversion, and emotional stability decline at higher educational levels. Other-rated personality measures typically provide stronger, more consistent correlations with educational outcomes. Although intelligence affects some of these relationships, educational outcomes are independently linked with personality with effect sizes comparable with correlations with intelligence. Consequently, personality has a crucial role in education.

Educational outcomes are among the most important factors in advanced economies, both for individuals and nations. Better-educated populations result in greater economic growth (Hanushek and Woessmann, 2012), while an individual level, achievement on grades and grade point average (GPA) reflects not only learning achievement but also predicts later work performance (Roth et al., 1996), and occupational prestige and status (Strenze, 2007). Consequently, recent work that has established an important role for personality and related factors with respect to GPA and other educational outcomes assumes much more than academic interest.

Cognitive and Noncognitive Predictors of Educational Outcomes

Early research on predictors of educational outcomes led to the delineation of the preeminent feature of individual differences research within both psychology and education, namely intelligence. The pioneering work of Binet and Simon (1905) aimed at predicting students' potential for succeeding at school, the end-result of which were the Stanford–Binet Intelligence Scales (Wolf, 1973), while Spearman identified his *g* factor of intelligence by factor-analyzing educational performance measures (Spearman, 1927). These cognitively based individual difference factors have long been recognized as important within educational settings.

Systematic research aimed at identifying "noncognitive skills" (Heckman and Rubinstein, 2001) relevant to education has continued for at least as long as have explorations of intelligence. Yet although early research by Webb (1915) produced substantial associations between personality ratings and educational performance, most reviewers reported inconsistent results (Harris, 1940; Margrain, 1978), largely due to the lack of coherent organizing frameworks for personality measurement, effectively precluding comparison of results from different studies. In response, De Raad and Schouwenburg (1996) argued for the use of comprehensive, reliable personality factor structures for systematically reviewing this research. This recognition, together with comparable research into links between personality and performance in other settings, especially the workplace (Barrick and Mount, 1991), provided the inspiration for recent reviews that have established consistent links between personality and educational outcomes.

The currently dominant model of personality, which has enabled recent integrative reviews, is the five-factor model (FFM). Norman (1963) is often credited with proposing the FFM, but hints of this model were provided as early as Webb (1915). The factors within the FFM are conscientiousness (organized, responsible, and self-disciplined); agreeableness (sympathetic, warm, and kind); emotional stability (relaxed and patient) but often identified by its opposite pole labeled as neuroticism (moody, irritable, and envious); extraversion (talkative and sociable); and openness (intellect and creativity, smart, innovative, and inquisitive) (Saucier and Goldberg, 1996).

Although other models of personality have been developed, and there are significant concerns about its cross-cultural validity, the FFM remains the dominant model for personality research generally, and within educational research specifically, with the overwhelming majority of studies of personality and educational outcomes using FFM-based measures (Richardson et al., 2012).

FFM and Educational Outcomes

The first systematic review of correlations with educational outcomes that was organized on the basis of the FFM was presented by Hough et al. (1990), but this has recently been followed by more rigorous and exhaustive reviews (Poropat, 2009, 2014a,b; Richardson et al., 2012). Collectively, these have demonstrated that personality has reliable and important associations with educational outcomes.

Conscientiousness

Using a precursor of factor analysis, Webb (1915) was the first researcher to identify a 'factor' approximating conscientiousness, by proposing the existence of a variable reflecting

'persistence of motives,' or alternatively, "*consistency of action resulting from deliberate volition, or will*" (p. 60, italics in the original): in other words, a *w* factor as a complement to Spearman's *g*, and incorporating a rating scale labeled *conscientiousness*. The FFM factor of the same name has been reported as having the strongest correlation with educational outcomes, ranging from grades (Hough et al., 1990; Poropat, 2009, 2014a,b; Richardson et al., 2012), to workplace-learning achievements (Barrick et al., 2001), and within educational settings there has been more research on conscientiousness than any other personality variable (Poropat, 2009, 2014a,b; Richardson et al., 2012).

Conscientiousness has been shown to have reliable correlations with academic performance at primary, secondary, and tertiary levels of education (Poropat, 2009), and this pattern has been observed across cultures (Kaufman et al., 2008; Poropat, 2009). Typically, research on personality uses self-rated measures, yet correlations of academic performance with conscientiousness increase substantially when someone other than the student rates their personality. Notably, observed correlations (i.e., uncorrected for measurement artifacts) of academic performance with other-rated conscientiousness exceed comparable correlations with intelligence at secondary and tertiary levels of education (Poropat, 2014a,b).

At the behavioral level, conscientiousness is associated with compliance with social norms, setting and pursuing goals, and operating in a purposeful manner (Roberts et al., 2009). Goal-setting reliably predicts school achievement, as do a range of other conscientiousness-related behaviors and activities, including self-monitoring and self-management, planfully and persistently following a course of action regardless of distractions, and noticing and correcting errors (Hattie, 2009). So, in many ways the actions of the conscientious student mirror those that educators should encourage in their students in order to achieve better learning outcomes.

At a deeper level, conscientiousness has been closely associated with *effortful control*, a component of individual temperament (Rothbart, 2007). Effortful control has been posited to encompass elements such as the ability to control attention and inhibit inappropriate responses, and consequently to self-regulate and persist with activities. The capacities associated with effortful control are apparently based upon executive attention activities within the brain (Posner and Rothbart, 2009), which have been explicitly linked with conscientiousness (Rothbart, 2007). These abilities play an important role in the relationship between conscientiousness and educational outcomes (Bidjerano and Dai, 2007; Eilam et al., 2009; Richardson et al., 2012). For example, more conscientious students are more likely to stay enrolled (Komarraju and Karau, 2005; Sitzmann, 2012), and are better at managing and allocating their time to enable them to study more (MacCann et al., 2012a; Poropat, 2011b; Poropat and Jones, 2009). But more conscientious students also study more effectively (Bidjerano and Dai, 2007), adapting their activities to specific learning requirements (Swanberg and Martinsen, 2010), resulting in better outcomes (Zhang et al., 2011).

Conscientiousness is also linked with intrinsic motivation to learn (Clark and Schroth, 2010), both initially and subsequent to formal educational activities, when more conscientious students are more likely to focus on activities that will assist them in applying their learning beyond educational settings (Yamkovenko and Holton, 2010). At the same time, conscientiousness is valuable in assisting students to control negative responses to learning, with more conscientious students being more able to manage stress and to refocus from stressors to their learning-based tasks (Saklofske et al., 2012). This is aided by the use of problem-focused coping (MacCann et al., 2012b), including shifting focus from performance goals to learning goals, which in turn reduces levels of stress and tension for students who are more conscientious (Cianci et al., 2010) resulting in these students feeling better within themselves (MacCann et al., 2012b). Thus, conscientiousness aids learning in many ways, which helps to explain why it is the FFM dimension that has the strongest associations with education outcomes.

Agreeableness

Agreeableness is associated with the ability of students to develop and maintain positive relationships with others, which seems to account for links between agreeableness and learning (Saklofske et al., 2012), because the adaptability associated with agreeableness makes it easier for students to make extra efforts outside of formal teaching, including making greater efforts with homework (Lubbers et al., 2010). Yet more agreeable students also appear to be generally better at time management and independent effort regulation, potentially contributing to later academic performance (Bidjerano and Dai, 2007). More agreeable students have a greater ability to internalize the interests of others, enabling them to more readily recognize and behaviorally comply with extrinsic demands even when they conflict with personal interests (Clark and Schroth, 2010). These factors assist students to learn more effectively in groups, which is important given that group learning is often more effective than individual learning (Michael, 2006).

Consistent with these associations, agreeableness has reliable correlations with academic performance in primary education, although there is reason to think that this may be partly due to methodological issues associated with self-ratings among younger children (Poropat, 2014a). When agreeableness is rated by someone other than the student, reliable but modest correlations are found (Poropat, 2014a,b), suggesting that agreeableness has persistent but limited impact on educational outcomes.

Emotional Stability

Emotional stability is often referred to by the label for its opposite pole, namely *neuroticism*, a close correlate of the component of temperament called negative affectivity (De Pauw, 2012; Rothbart, 2007). Negative affectivity interferes with learning by prompting task interruption and goal-blocking in response to distress (Rothbart, 2007). More emotionally stable people have weaker reactions to negative stimuli and hence are less readily discouraged, less distractible, and more confident in their own abilities (Ormel et al., 2013). Consistent with this, more emotionally stable students are less anxious and pay less attention to errors, which facilitates learning from errors (Zhao, 2011). This appears to be partly due to the ability of emotionally stable students to switch

focus away anxiety-prompting information, thereby reducing the degree to which they are distracted by ruminative thoughts that actively interfere with learning (Kircanski et al., 2008). Likewise, students who are higher on emotional stability are more able to stay focused on learning activities such as homework, and have better academic outcomes as a consequence (Lubbers et al., 2010).

The consequences of emotional stability extend to motivation and involvement in learning and education. Emotional stability is negatively correlated with discouragement, withdrawing from studies, or general avoidance motivations in education (Komarraju and Karau, 2005; Payne et al., 2007). Unlike agreeableness, emotional stability is negatively correlated with attention to extrinsic demands, such as those imposed by educators (Clark and Schroth, 2010; Komarraju et al., 2009), and is negatively correlated with motivation to learn (Zhao, 2011), possibly due to links with fear of failure. So, the association between emotional stability and academic motivation is complicated.

The association between self-rated emotional stability and academic performance in primary schooling is practically as well as statistically significant, but becomes virtually nonexistent in higher levels of education (Poropat, 2009). However, when other-ratings are used, there is no corresponding decline in this association (Poropat, 2014a,b), suggesting that what changes is the extent to which self-raters accurately report their level of emotional stability, possibly because low levels of emotional stability are socially undesirable.

Extraversion

Extraversion has long been recognized as a major component of individual differences in interpersonal behavior (Wiggins, 1991), and has been linked with aspects of temperament such as surgency and sociability (Mervielde and Asendorpf, 2000), as well as reduced behavioral inhibition and positive affectivity (McCrae and Costa, 1989). Poropat (2009) reported an association between extraversion and academic performance at primary education, but this association disappeared in secondary and tertiary education. Unlike emotional stability, use of self- or other-raters does not affect this pattern. Consequently, the differences in correlations between extraversion and academic performance do not appear to be due to changes in validity of self-ratings.

Poropat (2009) suggested that interaction levels may increase the level of attention that teachers give to extraverted students in primary education, making it easier to identify behavioral examples that warrant better grades. Consistent with this, teachers in primary schools tend to rate shyer, less-extraverted children as lower in intelligence and academic ability (Coplan et al., 2011). It appears that changes in educational assessment methods and reduced levels of interaction with students in secondary and tertiary education explain the reduced correlations with extraversion (Poropat, 2009).

Nonetheless, the impulsivity associated with extraversion can assist educational processes, with students who impulsively provide answers to questions showing greater learning (Tymms and Merrell, 2011), apparently because the simple act of providing an answer increases comprehension (Graesser, 2009). Similarly, extraversion is linked with novelty-seeking (Bernard, 2010), resulting in greater levels of experience and subsequent ability (Dragoni et al., 2011), especially in situations where learners are able to explore their learning environment independently (Orvis et al., 2011). So it appears that extraversion may be useful for learning, but the results are unlikely to affect grades.

Openness

There are strong theoretical arguments for expecting openness to be linked with education (De Raad and Schouwenburg, 1996), and openness has the strongest association with academic performance of any FFM factor apart from conscientiousness (Poropat, 2009), especially when rated by someone other than the student (Poropat, 2014a,b). Much of the reason for the link between openness and educational outcomes appears to be due to motivational effects, with students high on openness having greater curiosity, more interest in developing their ability, enjoying thinking, and generally wanting to improve (Bernard, 2010; Clark and Schroth, 2010; Komarraju and Karau, 2005). Consistent with this, openness is the FFM factor that has the strongest correlations with learning goal-orientation (Payne et al., 2007), and learning goal-orientation appears to mediate the association between openness and learning, and hence educational outcomes (Steinmayr et al., 2011). This focus on learning for its own sake operates independently of formal learning, with students higher on openness demonstrating greater learning in fields independent of their studies (Hambrick et al., 2008).

Students, who are higher on openness also adopt more effective learning strategies, such as greater use of meta-cognitive learning, elaboration on learned information, and effective time management and effort regulation (Bidjerano and Dai, 2007), as well as adopting a deep learning approach to study (Chamorro-Premuzic and Furnham, 2009), all of which reliably affect learning (Graesser, 2009). Taken together, these observations support the conclusion of De Raad and Schouwenburg (1996) that students, who are high on openness appear to reflect the 'ideal' student.

Alternative Personality Models

The advent of the FFM reduced the use of personality measures that fall outside that framework. Although it once was one of the dominant tools for personality assessment, the Eysenckian personality scales (Eysenck and Eysenck, 1975) are rarely used in educational research when compared with FFM measures. Two of the Eysenckian scales, extraversion and neuroticism, correspond closely to scales in the FFM, but although the Eysenckian psychoticism scale has overlaps with conscientiousness it is substantially less strongly correlated with educational outcomes, and further appears to be psychometrically unsound in educational environments (Poropat, 2011a).

The model of personality based upon Reinforcement Sensitivity Theory (RST) was explicitly linked to behavioral learning theory (Corr, 2004). Despite this, much remains to be done to establish RST as a compelling account of learning (Matthews, 2007), and surprisingly little research has been reported that

has tested RST in educational settings, so its relevance to learning and education beyond the psychological laboratory remains to be established.

Moderating Factors

The relationship between personality and educational outcomes has been shown to systematically vary depending on a range of factors. As with correlations with intelligence, correlations of academic performance with self-rated measures of most of the FFM dimensions decline from primary through to tertiary levels (Poropat, 2009). The declines in correlations with intelligence are often attributed to range restriction resulting from reductions in enrollments at higher levels of education, and these range restriction effects seem to also apply to correlations with personality. This makes the lack of significant reduction in correlations with self-rated conscientiousness notable, suggesting that its true relationship with academic performance may actually increase at higher educational levels.

Measurement also has a substantial effect on the size of correlations between personality and academic performance, with other-rated personality typically having significantly larger correlations. In the case of conscientiousness, other-rated measures have correlations with academic performance at both primary (50: Poropat, 2014a) and secondary/tertiary level (38: Poropat, 2014b) that are of similar magnitude to some of the largest effect sizes for influences on educational outcomes, those for class attendance (Credé et al., 2010) and feedback (Hattie and Timperley, 2007). Likewise, openness has substantially higher correlations with academic performance when other-rated, both at primary (43: Poropat, 2014a) and higher educational levels (28: Poropat, 2014b). The reasons for these moderating effects are complex but are likely to include the extent to which different raters use different sets of observations as the basis for their assessments and the extent to which they are differentially affected by desires to present the target of their ratings based on social desirability concerns (Poropat, 2014a,b).

Conclusions

It is a long time since the first systematic study aimed at identifying noncognitive predictors of academic performance, during which personality has mostly been treated as of lesser concern than have cognitive variables, especially intelligence. Recent reviews have clearly demonstrated that this neglect of personality within educational settings has been misplaced, with conscientiousness especially, but also openness and to a lesser extent emotional stability, having important consequences for students' outcomes. These consequences extend beyond the classroom, as has been shown by the now substantial body of research on the consequences of personality for work, health, and social outcomes (Roberts et al., 2007). Careful consideration of the role of personality within educational settings is therefore likely to provide similar benefits to students and society as have earlier efforts inspired by an appreciation of the strengths and limitations of other measures of individual differences.

See also: Academic Self-Concept and Achievement; Agreeableness; Conscientiousness; Extraversion; Five Factor Model of Personality, Assessment of; Intelligence: History of the Concept; Intrinsic Motivation: A Cultural Perspective; Motivation, Learning, and Instruction; Neuroticism; Openness to Experience; Person Perception, Accuracy of; Personality Development and Temperament; Personality and Adaptive Behaviors; Personality and Economics; Procrastination; Self-Concepts: Educational Aspects; Self-Regulated Learning: Theories, Measures, and Outcomes; Self-Regulated Learning; Systematic Reviewing and Meta-Analysis; Temperament and Motivation; Temperament.

Bibliography

Barrick, M.R., Mount, M.K., 1991. The Big Five personality dimensions and job performance: a meta-analysis. Personality Psychology 44, 1–26.

Barrick, M.R., Mount, M.K., Judge, T.A., 2001. Personality and performance at the beginning of the new millennium: what do we know and where do we go next. International Journal of Selection and Assessment 9 (1–2), 9–30. http://dx.doi.org/10.1111/1468-2389.00160.

Bernard, L.C., 2010. Motivation and personality: relationships between putative motive dimensions and the five factor model of personality. Psychological Reports 106 (2), 613–631. http://dx.doi.org/10.2466/pr0.106.2.613-631.

Bidjerano, T., Dai, D.Y., 2007. The relationship between the big-five model of personality and self-regulated learning strategies. Learning and Individual Differences 17 (1), 69–81. http://dx.doi.org/10.1016/j.lindif.2007.02.001.

Binet, A., Simon, T., 1905. Méthodes nouvelles pour le diagnostic du niveau intellectual des anormaux. L'Année psychologique 11, 191–336.

Chamorro-Premuzic, T., Furnham, A., 2009. Mainly openness: the relationship between the Big Five personality traits and learning approaches. Learning and Individual Differences 19 (4), 524–529. http://dx.doi.org/10.1016/j.lindif.2009.06.004.

Cianci, A.M., Klein, H.J., Seijts, G.H., 2010. The effect of negative feedback on tension and subsequent performance: the main and interactive effects of goal content and conscientiousness. Journal of Applied Psychology 95 (4), 618–630. http://dx.doi.org/10.1037/a0019130.

Clark, M.H., Schroth, C.A., 2010. Examining relationships between academic motivation and personality among college students. Learning and Individual Differences 20, 19–24. http://dx.doi.org/10.1016/j.lindif.2009.10.002.

Coplan, R.J., Hughes, K., Bosacki, S., Rose-Krasnor, L., 2011. Is silence golden? Elementary school teachers' strategies and beliefs regarding hypothetical shy/quiet and exuberant/talkative children. Journal of Educational Psychology 103 (4), 939–951. http://dx.doi.org/10.1037/a0024551.

Corr, P.J., 2004. Reinforcement sensitivity theory and personality. Neuroscience and Biobehavioral Reviews 28 (3), 317–332. http://dx.doi.org/10.1016/j.neubiorev.2004.01.005.

Credé, M., Roch, S.G., Kieszczynka, U.M., 2010. Class attendance in college: a meta-analytic review of the relationship of class attendance with grades and student characteristics. Review of Educational Research 80 (2), 272–295. http://dx.doi.org/10.3102/0034654310362998.

De Pauw, S.S.W., 2012. How are temperament and personality related in children and adolescents from 1.5 to 15 years?. In: Paper Presented at the 16th European Conference on Personality, Trieste, Italy.

De Raad, B., Schouwenburg, H.C., 1996. Personality in learning and education: a review. European Journal of Personality 10, 303–336. http://dx.doi.org/10.1002/(SICI)1099-0984(199612)10:5<303::AID-PER262>3.0.CO;2–2.

Dragoni, L., Oh, I.S., Vankatwyk, P., Tesluk, P.E., 2011. Developing executive leaders: the relative contribution of cognitive ability, personality, and the accumulation of work experience in predicting strategic thinking competency. Personnel Psychology 64 (4), 829–864. http://dx.doi.org/10.1111/j.1744-6570.2011.01229.x.

Eilam, B., Zeidner, M., Aharon, I., 2009. Student conscientiousness, self-regulated learning, and science achievement: an explorative field study. Psychology in the Schools 46 (5), 420–431. http://dx.doi.org/10.1002/pits.20387.

Eysenck, H.J., Eysenck, S.B.G., 1975. Manual of the Eysenck Personality Questionnaire. Hodder and Stoughton, London.

Graesser, A.C., 2009. Inaugural editorial for Journal of Educational Psychology. Journal of Educational Psychology 101 (2), 259–261. http://dx.doi.org/10.1037/a0014883.

Hambrick, D.Z., Pink, J.E., Meinz, E.J., Pettibone, J.C., Oswald, F.L., 2008. The roles of ability, personality, and interests in acquiring current events knowledge: a longitudinal study. Intelligence 36, 261–278. http://dx.doi.org/10.1016/j.intell.2007.06.004.

Hanushek, E.A., Woessmann, L., 2012. Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. Journal of Economic Growth 17 (4), 267–321. http://dx.doi.org/10.1007/s10887-012-9081-x.

Harris, D., 1940. Factors affecting college grades: a review of the literature, 1930–1937. Psychological Bulletin 37 (3), 125–166.

Hattie, J.A., 2009. Visible Learning: A Synthesis of Meta-analyses Relating to Achievement. Routledge, Abingdon, Oxon, UK.

Hattie, J.A., Timperley, H., 2007. The power of feedback. Review of Educational Research 77 (1), 81–112. http://dx.doi.org/10.3102/003465430298487.

Heckman, J.J., Rubinstein, Y., 2001. The importance of noncognitive skills: lessons from the GED testing program. The American Economic Review 91 (2), 145–149. http://dx.doi.org/10.1257/aer.91.2.145.

Hough, L.M., Eaton, N.K., Dunnette, M.D., Kamp, J.D., McCloy, R.A., 1990. Criterion-related validities of personality constructs and the effect of response distortion on those validities. Journal of Applied Psychology 75 (5), 581–595. http://dx.doi.org/10.1037/0021-9010.75.5.581.

Kaufman, J.C., Agars, M.D., Lopez-Wagner, M.C., 2008. The role of personality and motivation in predicting early college academic success in non-traditional students at a hispanic-serving institution. Learning and Individual Differences 18, 492–496. http://dx.doi.org/10.1016/j.lindif.2007.11.004.

Kircanski, K., Craske, M.G., Bjork, R., 2008. Thought suppression enhances memory bias for threat material. Behaviour Research and Therapy 46, 462–476. http://dx.doi.org/10.1016/j.brat.2008.01.009.

Komarraju, M., Karau, S.J., 2005. The relationship between the Big Five personality traits and academic motivation. Personality and Individual Differences 39, 557–567. http://dx.doi.org/10.1016/j.paid.2005.02.013.

Komarraju, M., Karau, S.J., Schmeck, R.R., 2009. Role of the Big Five personality traits in predicting college students' academic motivation and achievement. Learning and Individual Differences 19, 47–52. http://dx.doi.org/10.1016/j.lindif.2008.07.001.

Lubbers, M.J., Van der Werf, M.P.C., Kuyper, H., Hendriks, A.A.J., 2010. Does homework behavior mediate the relation between personality and academic performance? Learning and Individual Differences 20 (3), 203–208. http://dx.doi.org/10.1016/j.lindif.2010.01.005.

MacCann, C., Fogarty, G.J., Roberts, R.D., 2012a. Strategies for success in education: time management is more important for part-time than full-time community college students. Learning and Individual Differences 22, 618–623. http://dx.doi.org/10.1016/j.lindif.2011.09.015.

MacCann, C., Lipnevich, A.A., Burrus, J., Roberts, R.D., 2012b. The best years of our lives? Coping with stress predicts school grades, life satisfaction, and feelings about high school. Learning and Individual Differences 22 (2), 235–241. http://dx.doi.org/10.1016/j.lindif.2011.08.004.

Margrain, S.A., 1978. Student characteristics and academic performance in higher education: a review. Research in Higher Education 8 (2), 111–123.

Matthews, G., 2007. Reinforcement sensitivity theory: a critique from cognitive science. In: Corr, P.J. (Ed.), The Reinforcement Sensitivity Theory of Personality. Cambridge University Press, Cambridge, pp. 482–507.

McCrae, R.R., Costa Jr., P.T., 1989. The structure of interpersonal traits: Wiggins' circumplex and the five-factor model. Journal of Personality and Social Psychology 57, 586–595.

Mervielde, I., Asendorpf, J.B., 2000. Variable-centred versus person-centred approaches to childhood personality. In: Hampson, S.E. (Ed.), Advances in personality psychology, vol. 1. Taylor & Francis, Philadelphia, pp. 37–76.

Michael, J., 2006. Where's the evidence that active learning works? Advances in Physiology Education 30, 159–167. http://dx.doi.org/10.1152/advan.00053.2006.

Norman, W.T., 1963. Toward an adequate taxonomy of personality attributes: replicated factor structure in peer nomination personality ratings. Journal of Abnormal and Social Psychology 66, 574–583.

Ormel, J., Bastiaansen, A., Riese, H., Bos, E.H., Servaas, M., Ellenbogen, M., Aleman, A., 2013. The biological and psychological basis of neuroticism: current status and future directions. Neuroscience and Biobehavioral Reviews 37 (1), 59–72. http://dx.doi.org/10.1016/j.neubiorev.2012.09.004.

Orvis, K.A., Brusso, R.C., Wasserman, M.E., Fisher, S.L., 2011. E-Nabled for e-learning? The moderating role of personality in determining the optimal degree of learner control in an e-learning environment. Human Performance 24 (1), 60–78. http://dx.doi.org/10.1080/08959285.2010.530633.

Payne, S.C., Youngcourt, S.S., Beaubien, J.M., 2007. A meta-analytic examination of the goal orientation nomological net. Journal of Applied Psychology 92 (1), 128–150. http://dx.doi.org/10.1037/0021-9010.92.1.128.

Poropat, A.E., 2009. A meta-analysis of the five-factor model of personality and academic performance. Psychological Bulletin 135 (2), 322–338. http://dx.doi.org/10.1037/a0014996.

Poropat, A.E., 2011a. The Eysenckian personality factors and their correlations with academic performance. British Journal of Educational Psychology 81 (1), 41–58. http://dx.doi.org/10.1348/000709910X497671.

Poropat, A.E., 2011b. The role of citizenship performance in academic achievement and graduate employability. Education + Training 53 (6), 499–514. http://dx.doi.org/10.1108/00400911111159467.

Poropat, A.E., 2014a. A meta-analysis of adult-rated child personality and academic performance in primary education. British Journal of Educational Psychology 84 (2), 239–252. http://dx.doi.org/10.1111/bjep.12019.

Poropat, A.E., 2014b. A meta-analysis of other-rated personality and academic performance. Learning and Individual Differences 34 (1), 24–32. http://dx.doi.org/10.1016/j.lindif.2014.05.013.

Poropat, A.E., Jones, L.S., 2009. Development and validation of a unidimensional measure of citizenship performance. Journal of Occupational and Organizational Psychology 82 (4), 851–869. http://dx.doi.org/10.1348/096317908X374139.

Posner, M.I., Rothbart, M.K., 2009. Toward a physical basis of attention and self-regulation. Physics of Life Reviews 6 (2), 103–120. http://dx.doi.org/10.1016/j.plrev.2009.02.001.

Richardson, M., Abraham, C., Bond, R., 2012. Psychological correlates of university students' academic performance: a systematic review and meta-analysis. Psychological Bulletin 138 (2), 363–387. http://dx.doi.org/10.1037/a0026838.

Roberts, B.W., Jackson, J.J., Fayard, J.V., Edmonds, G., Meints, J., 2009. Conscientiousness. In: Leary, M.R., Hoyle, R.H. (Eds.), Handbook of Individual Differences in Social Behavior. Guilford, New York, pp. 369–381.

Roberts, B.W., Kuncel, N.R., Shiner, R.L., Caspi, A., Goldberg, L.R., 2007. The power of personality: the comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. Psychological Science 2 (4), 313–345. http://dx.doi.org/10.1111/j.1745-6916.2007.00047.x.

Roth, P.L., BeVier, C.A., Schippman, J.S., Switzer III, F.S., 1996. Meta-analysing the relationship between grades and job performance. Journal of Applied Psychology 81 (5), 548–556.

Rothbart, M.K., 2007. Temperament, development, and personality. Current Directions in Psychological Science 16 (4), 207–212. http://dx.doi.org/10.1111/j.1467-8721.2007.00505.x.

Saklofske, D.H., Austin, E.J., Mastoras, S.M., Beaton, L., Osborne, S.E., 2012. Relationships of personality, affect, emotional intelligence and coping with student stress and academic success: different patterns of association for stress and success. Learning and Individual Differences 22 (2), 251–257. http://dx.doi.org/10.1016/j.lindif.2011.02.010.

Saucier, G., Goldberg, L.R., 1996. Evidence for the Big Five in analyses of familiar English personality adjectives. European Journal of Personality 10, 61–77. http://dx.doi.org/10.1002/(SICI)1099-0984(199603)10:1<61::AID-PER246>3.0.CO;2-D.

Sitzmann, T., 2012. A theoretical model and analysis of the effect of self-regulation on attrition from voluntary online training. Learning and Individual Differences 22 (1), 46–54. http://dx.doi.org/10.1016/j.lindif.2011.11.004.

Spearman, C., 1927. The Abilities of Man. MacMillan, London.

Steinmayr, R., Bipp, T., Spinath, B., 2011. Goal orientations predict academic performance beyond intelligence and personality. Learning and Individual Differences 21 (2), 196–200. http://dx.doi.org/10.1016/j.lindif.2010.11.026.

Strenze, T., 2007. Intelligence and socioeconomic success: a meta-analytic review of longitudinal research. Intelligence 35, 401–426. http://dx.doi.org/10.1016/j.intell.2006.09.004.

Swanberg, A.B., Martinsen, O.L., 2010. Personality, approaches to learning and achievement. Educational Psychology 30 (1), 75–88. http://dx.doi.org/10.1080/01443410903410474.

Tymms, P., Merrell, C., 2011. ADHD and academic attainment: is there an advantage in impulsivity? Learning and Individual Differences 21 (6), 753–758. http://dx.doi.org/10.1016/j.lindif.2011.07.014.

Webb, E., 1915. Character and Intelligence: An Attempt at an Exact Study of Character. Cambridge University Press, Cambridge, UK.

Wiggins, J.S., 1991. Agency and communion as conceptual coordinates for the understanding and measurement of interpersonal behavior. In: Grove, W., Cicchetti, D. (Eds.), Thinking Clearly about Psychology: Essays in Honor of Paul E. Meehl, vol. 2. University of Minnesota Press, Minneapolis, pp. 89–113.

Wolf, T.H., 1973. Alfred Binet. University of Chicago Press, Chicago.

Yamkovenko, B., Holton, E., 2010. Toward a theoretical model of dispositional influences on transfer of learning: a test of a structural model. Human Resource Development Quarterly 21 (4), 381–410. http://dx.doi.org/10.1002/hrdq.20054.

Zhang, X., Nurmi, J.-E., Kiuru, N., Lerkkanen, M.-K., Aunola, K., 2011. A teacher-report measure of children's task-avoidant behavior: a validation study of the behavioral strategy rating scale. Learning and Individual Differences 21 (6), 690–698. http://dx.doi.org/10.1016/j.lindif.2011.09.007.

Zhao, B., 2011. Learning from errors: the role of context, emotion, and personality. Journal of Organizational Behavior 32, 435–463. http://dx.doi.org/10.1002/job.696.

Personality and Health

Christian Hakulinen and Liisa Keltikangas-Järvinen, Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H.S. Friedman, volume 16, pp. 11264–11270, © 2001, Elsevier Ltd.

Abstract

Research on the association between personality and health has a long and notable history. Recently, remarkable advances have been made and thus there is considerably more understanding of why personality is important for health. The development of personality theories has contributed to the improvement in current knowledge and understanding of the association between personality and health, e.g., in terms of health behaviors and stress. In this context, temperament has been shown to be of importance. In the future, more multidisciplinary research is required to better understand the association between personality and health.

Personality has a central role in all-round well-being and health. It is one of the factors, which partly explains variances in health; and to a certain degree, why – even under the same circumstances – some individuals fall ill, whereas others do not. Interest in the connection between personality and health started very early, although coherent personality concepts, which are in line with current personality theories, have only been developed recently. Previously, personality researchers strongly criticized personality concepts, which were used in health psychology. Nowadays a better balance has been reached and the most recent personality theories are used in research on the complicated association between personality and health. Further, sophisticated statistical methods and large follow-up samples are seen as prerequisites for tenable research.

History of the Research

The relationship between personality and health has been investigated for over 2000 years. In Ancient Greece, individuals were already divided into four groups according to predominating bodily humors. Those bodily humors were blood, yellow bile, black bile, and phlegm; and each of these humors was thought to be connected with a certain personality trait or a certain 'temperament,' as the personality traits were called. For instance, an individual whose predominating bodily humor was blood, was believed to have a sanguine temperament that made the person socially active and a seeker of pleasure. In turn, the domination of yellow bile was believed to lead to a choleric temperament that was characterized by high levels of energy and ambition. Further, the predominating bodily humors referred to specific somatic weaknesses too, and consequently, temperaments were thought to be associated with specific illnesses. For example, individuals who were predominated by blood and thus had sanguine temperaments, were seen to be more prone to heart-related problems; whereas individuals with a combination of yellow bile and choleric temperament were thought to be at risk for fatty liver problems.

The current research on the role of personality in all-round health originates in the concept of 'Psychosomatics' as conceived in the 1930s. According to this concept, psychological factors are causative agents for particular illnesses; and certain personalities are predisposed to certain diseases. Personality did not play a role in all conditions but there was a selected category of diseases whose etiology was believed to implicate personality factors. Consequently, personality was not involved in all-round health, but was a special feature in some sensitive 'psychosomatic' individuals. Due to the presumed link between a certain personality and a specific disease, this concept was called the 'Specificity Theory' (Alexander, 1950).

Even in the 1950s and 1960s, this concept provoked growing criticism and it was not widely accepted outside psychoanalytic circles. However, related lines of research have evolved, focusing on the relations between personality, disposition or specific behavior patterns, and the predisposition to specific diseases. A current variation on this specificity theme is the Type A concept, proposed in 1953 to describe people expressing certain behavior traits, such as aggression, impatience, hard-driving effort and hostility; and this pattern of behavior was presumed to be linked to risk for coronary heart disease (Friedman and Rosenman, 1959). This behavior pattern was considered to reflect personality; and during the next years, several very different personality features were named as Type A behavior or were suggested to underlie this behavior pattern.

The Type A concept was rapidly and widely accepted, not because of its solid evidence but because of its practicality. It was an easy way to cope with the complicated concept of personality. It also divided personality into three clear dimensions: attitude, affect, and behavior.

Even though the Type A concept was vague and reflected the specificity theory, it became a landmark concept of modern health psychology in the sense that it raised a wide interest in the role of personality in somatic health. It also brought about a great number of medical and epidemiological studies where psychological factors were included for the first time.

In the 1960s and 1970s, there was a large and growing body of studies on personality and health. Research widened into personality factors to such a degree, that it is not an overstatement to say that personality was associated with onset and progress of all known diseases. Research on psychological risk factors of coronary heart disease, however, maintained its

dominating role. Despite all efforts, however, the results were meager. Personality was shown to be of significance; and always, when personality was assessed in context of health and disease, some associations were found. Nevertheless, the findings were more likely to confirm the existence of personality than demonstrate its role in health. Study designs were mostly correlative and they offered no explanation for the mechanisms through which personality might have effect. In the 1960s and 1970s, searching for this mechanism was not even the focus of the research; demonstrating a correlation between personality and health was seen to be enough.

Long-term, population-based, follow-up studies to examine the development and progress of disease among healthy people and experimental designs to identify mechanisms underlying the association between personality and disease, were typical of the 1980s. Instead of 'psychosomatics,' the term 'behavioral medicine' was used to refer to this branch of science that linked psychology and medicine.

Emotions and Health

Type A Behavior

The Type A behavior pattern was a landmark of modern behavioral medicine. The beginning of the Type A behavior research era was promising and in the well-known longitudinal studies, i.e., the Framingham Study and the Western Collaborative study, an association between Type A behavior and the pathogenesis of coronary heart disease was established (Haynes et al., 1980; Rosenman et al., 1975). Nevertheless, the original finding that the Type A behavior pattern is a risk factor for coronary heart disease could not be replicated. From the vast amount of evidence, it was concluded in 1999 that Type A behavior has only small etiologic relevance for coronary heart disease among healthy adults; and that it is not associated with the prognosis of patients with coronary heart disease (Hemingway and Marmot, 1999).

Nevertheless, instead of the theoretically vague global Type A score, some of its constituent components were shown to be relevant for health, in particular its affective components, i.e., anger and hostility. This development turned the focus of behavioral medicine to the role of affects and emotions in health, i.e., to the very topics which early psychosomatics started from.

Anger and Hostility

From the large amount of Type A behavior-related research, anger and hostility emerged as the most important components affecting health. Up until now, a vast amount of evidence has shown that anger and hostility not only predict physical health but also a range of mental health and social problems. A recent meta-analytic review concluded that anger and hostility are associated with risk of coronary heart disease risk in healthy populations and also in clinical populations with coronary heart disease; the total hazard ratio was estimated to be around 1.2 (Chida and Steptoe, 2009). Currently, it is, however, not clear whether the risks of hostility and anger are equal or whether they differ. Most studies have combined measures of anger and hostility, even though the evidence suggests that there might be important differences between these concepts. In any case, findings on anger and hostility confirmed an old statement that emotional processing is very relevant to health.

Type D Personality

The identification of the role of anger and hostility within the Type A concept contributed to the conceptualization of the Type D personality. Instead of behavior and attitudes, which were emphasized by the Type A concept, negative emotions and low capacity to process emotions characterize the Type D personality. Thus, the Type D personality can be defined as the individual's tendency to experience the combination of negative emotions and social inhibition. Individuals with Type D personality are thus more prone to experience stress and to have poor health behavior-related habits. Early findings suggested that the Type D personality was a clear risk factor for coronary heart disease (Denollet et al., 1996). In a recent review it was concluded that Type D personality was associated with a three-fold increased risk of poor coronary heart disease prognosis (Denollet et al., 2010). Nevertheless, there are also negative reports suggesting that the original estimates might be too high (Smith, 2011).

Alexithymia

Emphasizing a role for emotional processing on health did not start from the identification of Type A behavior but emerged earlier in the 1930s and 1940s, with physicians reporting on patients with severe deficiencies in verbalizing emotions. In the 1970s, Sifneos termed this deficiency as 'alexithymia' (Sifneos, 1977), which means 'no words to emotions.' The term refers to the inability of alexithymic individuals to understand other people's emotional expressions, as well as their inablility to verbalize and interpret their own emotions. In addition to a deficiency of interpreting and verbalizing emotions, alexithymic persons are presumed to lack imagination too, which results in difficulty to benefit from therapy. Alexithymia has been associated with increased mortality, poor physical outcomes, and poor health behavior (Ferguson, 2013).

Toward Comprehensive Personality Concepts

A strong development of personality theories is characteristic of the 1980s and 1990s. Especially in the 1990s, molecular genetics significantly contributed to the development of the concept of personality. More advanced concepts that viewed personality from genetic, neurobiological, and physiological perspectives, also offered better tools to study the link between personality and health through the mechanisms via which personality is likely to have its effect. The focus shifted more toward modern personality theories while pathogenesis, health behavior, and coping with illness and its precursors (e.g., stress) were seen as mechanisms through which personality operates.

Personality Theories

Starting from the 1980s there has been a rapid development in personality research. Nevertheless, consensus about personality

traits has been reached only slowly. Currently, it is widely acknowledged that personality can be seen to consist of a small number of general traits that are formed from a wide number of specific personality dispositions (Digman, 1990). The Five-Factor Model, i.e., the 'Big Five,' is currently the most widely adopted concept to organize these general characteristics into a single model. The Five-Factor Model follows five broad traits: extraversion (e.g., sociability and sensitivity to positive emotions), neuroticism (e.g., proneness to anxiety), conscientiousness (e.g., self-control and allegiance to social norms), agreeableness (e.g., cooperation and trust toward other people), and openness to experience (e.g., curiosity and open-mindedness).

Another widely accepted approach to personality involves differentiation between 'personality' and 'temperament' (Buss, 1991) or 'temperament' and 'character' (Cloninger et al., 1993), in the individual's personality. Temperament, especially, has been shown to be very relevant in the context of health.

Temperament traits are seen as relatively stable behavioral dispositions and tendencies, which are partly heritable and connected to certain biological mechanisms. Possibly due to their neurobiological and physical correlates, temperamental traits have been documented as playing an important role in disease and well-being. At least, Gray's, Buss and Plomin's, and Cloninger's theories can be highlighted in this context.

Gray's (1991) temperament theory postulates two major temperament systems: the Behavioral Approach System (BAS) and the Behavioral Inhibition System (BIS). The BIS is hypothesized to be related to sensitivity to punishment, whereas BAS is seen to be related to sensitivity to reward. Thus, BAS can be seen as the 'GO' or motivation system and BIS as the 'STOP' or alert system. According to Buss (1991), there are three major temperament traits: negative emotionality, activity, and sociability. Negative emotionality is related to arousal and ease in experiencing negative emotions; activity relates to individual differences in physical energy; and sociability is defined as the ability to enjoy other people's company versus being alone. Cloninger's psychobiological theory of temperament includes four temperament dimensions, which are novelty seeking (NS), harm avoidance (HA), reward dependence (RD), and persistence (P) (Cloninger et al., 1993). NS is related to tendency to respond strongly to novelty, whereas HA is defined as tendency to respond intensely to signals of aversive stimuli. RD relates to the tendency to respond strongly to rewards such as social approval, and P is defined as resistance to extinction despite intermittent reinforcement. In addition, Cloninger (1987) suggested that NS is influenced by dopaminergic activity, HA by serotonergic activity, and RD by noradrenergic activity.

Pathogenesis

During the last 20 years, quite a large amount of evidence has been gathered suggesting that personality is related to the pathogenesis of diseases. First findings from the prospective Young Finns Study, which includes 3596 children and adolescents who have been followed from the 1980s up to date, demonstrated that associations between temperament and health are already present at childhood, in that the temperament trait negative emotionality has been found to be associated with the metabolic syndrome in children (Ravaja and Keltikangas-Järvinen, 1995). Later findings from the Young Finns Study have demonstrated that high NS, high RD, and low HA are associated with preclinical atherosclerosis (Hintsanen et al., 2009); and high NS with higher body-mass index (Hintsanen et al., 2012). In addition to evidence from the prospective studies, findings from experimental studies suggest also that there is a connection between temperament and physiological systems. For example BAS-sensitivity has been associated with heart rate reactivity and parasympathetic withdrawal during stressful laboratory tasks (Heponiemi et al., 2004). Temperament trait harm avoidance, which is closely related to the concept of neuroticism, has been associated with higher resting heart rate, which in turn reflects higher for coronary heart disease (CHD) risk (Puttonen et al., 2008).

More recently research has focused on the association between the Five-Factor Model of personality and health. Although associations between almost all of the five personality traits and health were found, conscientiousness could be the most relevant personality trait for health-related outcomes (Bogg and Roberts, 2004). Recent studies using large prospective cohort data from the United Kingdom, the United States, Germany, and Australia have further shown that low conscientiousness is related to obesity (Jokela et al., 2013a), diabetes (Jokela et al., 2013b), and all-cause mortality (Jokela et al., 2013c).

Health Behavior

From the Five-Factor Model personality traits, low conscientiousness has been found to be related to almost all health-risk behaviors, such as smoking and drinking (Bogg and Roberts, 2004). In addition, it has also been hypothesized that individuals who are highly conscientious are more likely to follow social norms and public health recommendations that are related to health behaviors. It is widely acknowledged that most of the associations between temperament or personality traits and somatic problems are at least partially explained by health behaviors. Nevertheless, as demonstrated in the recent study by Jokela et al. (2013c), health behaviors do not fully explain the association between low conscientiousness and mortality. Thus, it seems that personality also affects health through pathways other than health behaviors.

Coping with Illness and Stress

Associations between acccumulated stress and health have been numerously demonstrated. Research has specially focused on work stress such as job strain, job demands, and job control. Numerous studies have shown that personality is related to stress and also to stress perception. Studies using the Young Finns Study prospective data have shown that personality traits, such as high neuroticism, low openness, and low conscientiousness, are associated with high job strain; high neuroticism has also been shown to be associated with high demands and low control (Törnroos et al., 2013). In addition, high negative emotionality and low sociability have been associated with high job strain and imbalance between effort and reward

(Hintsanen et al., 2011). There is also some evidence from experimental studies. For example, Cloninger's temperament trait, high harm avoidance, has been associated with high 'anticipated threat' before a stressful situation and also high 'perceived stress' after the situation (Ravaja et al., 2006).

Other Explanations

Of course, there are also other possible pathways between personality and health. For example personality traits such as high neuroticism, low conscientiousness, and low extraversion have been associated with major depressive disorder (Kotov et al., 2010), which in turn has been shown to predict mortality (Cuijpers and Smit, 2002). Thus, it is possible that individuals first get depressed and after that their physical problems start.

There are also a number of biological processes that can explain the association between personality and health. It is possible that the same genes that contribute to certain personality disposition are also related to certain disease outcomes. Nevertheless, quantitative evidence for these kinds of associations is currently quite scarce. Personality traits are complex traits, which are affected by a vast number of different genes. In addition, it is well-known that genes do not function in isolation but interact with environments. This further adds complexity to the picture.

Conclusion

The evidence suggests that personality plays a major role in explaining why some individuals are more likely to get ill than others. During the last two decades there has been rapid progress in identifying such associations. Nevertheless, there is still much to be done in explaining the association between personality and health. One challenge is to move away from the deterministic single-trait approach that was characteristic of the past and to develop a theoretically solid framework to better understand the link between health and personality. In addition, previous research has been quite disease-focused, and in the future it will be important also to examine the association between personality and good health. Important advances have also been recently achieved in the field of positive psychology.

Future research should especially concentrate on further explaining how personality is associated with the illness process over the life span. It is also important to consider an evolutionary point of view, i.e., personality traits can be seen as evolutionary mechanisms that have different costs and benefits; and thus individual behavior could be adaptive in one environment and maladaptive in another (Ferguson, 2013). It is also important to continue multidisciplinary research, which makes it possible to investigate the role of genes and the environment in the context of personality and health. Taken together, much can be expected from research in the field of personality and health in the future.

See also: Behavioral Medicine; Chronic Pain: Models and Treatment Approaches; Control Beliefs: Health Perspectives; Coping with Job Stress; Coronary Heart Disease, Psychosocial Aspects; Epigenetics in Social and Behavioral Sciences; Health Behaviors; Health Risk Appraisal and Optimistic Bias; Hypertension: Psychosocial Aspects; Occupational Health; Resilience; Startle Reflex and Health; Stress and Cardiac Defense; Stress, Coping, and Health.

Bibliography

Alexander, F., 1950. Psychosomatic Medicine – Its Principles and Applications. W.W. Norton and Company Inc., New York.

Bogg, T., Roberts, B.W., 2004. Conscientiousness and health-related behaviors: a meta-analysis of the leading behavioral contributors to mortality. Psychological Bulletin 130, 887–919.

Buss, A.H., 1991. The EAS theory of temperament. In: Strelau, J., Angleitner, A. (Eds.), Explorations in Temperament: International Perspectives on Theory and Measurement, Perspectives on Individual Differences. Plenum Press, New York, pp. 43–60.

Chida, Y., Steptoe, A., 2009. The association of anger and hostility with future coronary heart disease: a meta-analytic review of prospective evidence. Journal of the American College of Cardiology 53, 936–946.

Cloninger, C.R., 1987. A systematic method for clinical description and classification of personality variants – a proposal. Archives of General Psychiatry 44, 573–588.

Cloninger, C.R., Svrakic, D.M., Przybeck, T.R., 1993. A psychobiological model of temperament and character. Archives of General Psychiatry 50, 975–990.

Cuijpers, P., Smit, F., 2002. Excess mortality in depression: a meta-analysis of community studies. Journal of Affective Disorders 72, 227–236.

Denollet, J., Rombouts, H., Gillebert, T., Brutsaert, D., Sys, S., Stroobant, N., 1996. Personality as independent predictor of long-term mortality in patients with coronary heart disease. The Lancet 347, 417–421.

Denollet, J., Schiffer, A.A., Spek, V., 2010. A general propensity to psychological distress affects cardiovascular outcomes evidence from research on the type D (distressed) personality profile. Circulation: Cardiovascular Quality and Outcomes 3, 546–557.

Digman, J.M., 1990. Personality structure: emergence of the five-factor model. Annual Review of Psychology 41, 417–440.

Ferguson, E., 2013. Personality is of central concern to understand health: towards a theoretical model for health psychology. Health Psychology Review 7, S32–S70.

Friedman, M., Rosenman, R.H., 1959. Association of specific overt behavior pattern with blood and cardiovascular findings: blood cholesterol level, blood clotting time, incidence of arcus senilis, and clinical coronary artery disease. The Journal of the American Medical Association 169, 1286–1296.

Gray, J.A., 1991. The neuropsychology of temperament. In: Strelau, J., Angleitner, A. (Eds.), Explorations in Temperament: International Perspectives on Theory and Measurement. Perspectives on Individual Differences. Plenum Press, New York, pp. 105–128.

Haynes, S., Feinleib, M., Kannel, W.B., 1980. The relationship of psychosocial factors to coronary heart disease in the Framingham Study. III. Eight-year incidence of coronary heart disease. American Journal of Epidemiology 111, 37–58.

Hemingway, H., Marmot, M., 1999. Evidence based cardiology: psychosocial factors in the aetiology and prognosis of coronary heart disease: systematic review of prospective cohort studies. British Medical Journal 318, 1460–1467.

Heponiemi, T., Keltikangas-Järvinen, L., Kettunen, J., Puttonen, S., Ravaja, N., 2004. BIS–BAS sensitivity and cardiac autonomic stress profiles. Psychophysiology 41, 37–45.

Hintsanen, M., Hintsa, T., Widell, A., Kivimäki, M., Raitakari, O.T., Keltkangas-Järvinen, L., 2011. Negative emotionality, activity, and sociability temperaments predicting long-term job strain and effort–reward imbalance: a 15-year prospective follow-up study. Journal of Psychosomatic Research 71, 90–96.

Hintsanen, M., Jokela, M., Cloninger, C.R., Pulkki-Råback, L., Hintsa, T., Elovainio, M., Josefsson, K., Rosenström, T., Mullola, S., Raitakari, O.T., Keltikangas-Järvinen, L., 2012. Temperament and character predict body-mass index: a population-based prospective cohort study. Journal of Psychosomatic Research 73, 391–397.

Hintsanen, M., Pulkki-Råback, L., Juonala, M., Viikari, J.S., Raitakari, O.T., Keltikangas-Järvinen, L., 2009. Cloninger's temperament traits and preclinical atherosclerosis: the cardiovascular risk in Young Finns Study. Journal of Psychosomatic Research 67, 77–84.

Jokela, M., Hintsanen, M., Hakulinen, C., Batty, G., Nabi, H., Singh-Manoux, A., Kivimäki, M., 2013a. Association of personality with the development and persistence of obesity: a meta-analysis based on individual–participant data. Obesity Reviews 14, 315–323.

Jokela, M., Batty, G.D., Nyberg, S.T., Virtanen, M., Nabi, H., Singh-Manoux, A., Kivimäki, M., 2013b. Personality and all-cause mortality: individual-participant meta-analysis of 3,947 deaths in 76,150 adults. American Journal of Epidemiology 178, 667–675.

Jokela, M., Elovainio, M., Nyberg, S.T., Tabák, A.G., Hintsa, T., Batty, G.D., Kivimäki, M., 2013c. Personality and risk of diabetes in adults: pooled analysis of 5 cohort studies. Health Psychology. Epub ahead of print. http://psycnet.apa.org/index.cfm fa=buy.optionToBuy&id=2013-29655-001.

Kotov, R., Gamez, W., Schmidt, F., Watson, D., 2010. Linking "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. Psychological Bulletin 136, 768–821.

Puttonen, S., Elovainio, M., Kivimäki, M., Koskinen, T., Pulkki-Råback, L., Viikari, J.S., Raitakari, O.T., Keltikangas-Järvinen, L., 2008. Temperament, health-related behaviors, and autonomic cardiac regulation: the cardiovascular risk in Young Finns Study. Biological Psychology 78, 204–210.

Ravaja, N., Keltikangas-Järvinen, L., Kettunen, J., 2006. Cloninger's temperament dimensions and threat, stress, and performance appraisals during different challenges among young adults. Journal of Personality 74, 287–310.

Ravaja, N., Keltikangas-Järvinen, L., 1995. Temperament and metabolic syndrome precursors in children: a three-year follow-up. Preventive Medicine 24, 518–527.

Rosenman, R., Brand, R., Jenkins, D., Friedman, M., Straus, R., Wurm, M., 1975. Coronary heart disease in western collaborative group study. Final follow-up experience of 8½ years. The Journal of the American Medical Association 233, 872–877.

Sifneos, P., 1977. Problems in psychotherapy of patients with alexithymic characteristics and physical disease. Psychother Psychosom 26, 65–70.

Smith, T.W., 2011. Toward a more systematic, cumulative, and applicable science of personality and health: lessons from type D personality. Psychosomatic Medicine 73, 528–532.

Törnroos, M., Hintsanen, M., Hintsa, T., Jokela, M., Pulkki-Råback, L., Hutri-Kähönen, N., Keltikangas-Järvinen, L., 2013. Associations between five-factor model traits and perceived job strain: a population-based study. Journal of Occupational Health Psychology 18, 492–500.

Personality and Marriage

Benjamin R Karney, University of Florida, Gainesville, FL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 11270–11274, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The success or failure of a marriage is likely to depend in part on the personality (i.e., the enduring traits and characteristics), of the individual spouses. To review research that has explored this idea, this article is organized into three sections. The first section addresses which personality traits have demonstrated consistent associations with marital outcomes. Although over 200 separate personality traits have been examined for their effects on marriage, only neuroticism, or negative affectivity, has demonstrated unique effects across multiple studies. The second section examines the role of similarity and complementarity in accounting for marital outcomes. Although some research supports the beneficial effects of similarity, these effects must be qualified by a number of methodological complexities in the literature. There has been no support for the complementarity hypothesis. The third section describes research on two kinds of intervening processes through which individual personality traits affect marital outcomes. Cumulative continuity suggests that the personality of each spouse shapes the environment of the marriage by giving rise to the stressful circumstances that couples face. Interactional continuity suggests that the personality of each spouse affects the nature of the interpersonal processes occurring within the marriage. Both processes have been demonstrated empirically.

The success or failure of a marriage is likely to depend, at least in part, on the personality of the individual spouses, broadly defined as the set of enduring traits and characteristics that each spouse brings to the relationship. Three questions have motivated most research on the role of personality in marriage. First, what personality traits are most relevant to predicting marital outcomes? Second, does similarity or dissimilarity between partners' personalities affect marital outcomes? Third, through what mechanisms does the personality of each spouse affect the relationship? This article reviews each of these areas of research in turn.

Which Personality Traits Affect Marital Outcomes?

The earliest efforts to predict marital outcomes were guided by the belief that 'a large proportion of incompatible marriages are so because of a predisposition to unhappiness in one or both of the spouses' (Terman, 1938: p. 110). To explore the support for this belief, the first marital researchers conducted two kinds of studies: cross-sectional studies, in which lengthy personality inventories were administered to samples that included happily married and distressed couples, and longitudinal studies, in which personality inventories were administered to engaged couples whose marital quality was assessed several years later. Early marital researchers were more interested in determining whether spouses' personalities accounted for marital outcomes than in identifying which aspects of personality were associated with marital outcomes. Thus, the results of these initial studies were frequently reported in terms of a global personality score rather than divided into specific personality traits. Nevertheless, when Burgess and Wallin (1953), in the first review of this literature, concluded that success and failure in marriage could be predicted by the personalities of each spouse, they noted that personality traits associated with neuroticism (e.g., irritability, moodiness, ill-temper) appeared to be especially predictive. Spouses who reported an enduring tendency to experience negative affect across a variety of contexts seemed to be at greater risk of developing marital distress.

As these results accumulated, the same researchers noticed that other enduring characteristics of the spouses (e.g., childhood family environment, sexual history, demographic variables), were also significant predictors of marital outcomes. Once statistical techniques for conducting multiple regression became widely available, a second wave of research examined the ability of personality traits to account for marital outcomes relative to these other sources of influence. Perhaps the most impressive of these studies is that of Kelly and Conley (1987), who followed a sample of 300 couples for over 40 years. At the initial phase of data collection, engaged partners provided self-reports of their early family environment, attitudes towards marriage, sexual history, and stressful life events. In addition, the researchers contacted five acquaintances of each partner and asked these individuals to rate the personality of the subject, a substantial improvement over the self-report inventories relied on by most prior (and subsequent) studies of personality. When all of these variables were examined at the same time, only the personality ratings emerged as significant predictors of marital stability. Specifically, compared to marriages that remained intact, couples whose marriages dissolved over the course of the study were characterized by higher levels of neuroticism in both spouses and higher levels of impulsivity in husbands.

In the late 1980s, the development of the five-factor model of personality briefly promised to expand the focus of marital research beyond the negative aspects of spouses' personalities. According to this model, the many traits that had been studied in prior research on personality could be adequately summarized by five relatively independent dimensions: neuroticism, impulsivity, extraversion, agreeableness, and conscientiousness. Emerging support for this model lead marital researchers to ask whether dimensions of the 'Big Five' besides neuroticism

and impulsivity may also be associated with marital outcomes. Thus, for example, Kurdek (1993), in a five-year study of newlywed couples, examined the ability of each of the five personality dimensions to predict the quality and stability of marriages over time. This study did find that wives' conscientiousness was weakly but positively associated with the stability of the marriage over five years. However, echoing the results of previous research, far larger effects were obtained for the neuroticism of both spouses. Karney and Bradbury (1995), in a meta-analysis of longitudinal research on personality and marriage, reported that of all of the Big Five personality dimensions, only neuroticism had demonstrated consistent effects on marriage across multiple studies. As the first marital researchers suggested, a spouse's 'predisposition to unhappiness' appears to predict decline and instability in marriage more reliably than other personality traits.

Similarity and Complementarity

As research on personality in marriage has developed, some have argued that the content of each spouse's personality may be less important to understanding marital outcomes than the degree of similarity between spouses' personalities. The assumption of this perspective is that spouses who share similar personalities should experience fewer opportunities for conflict and thus decreased risk of marital dissolution. Bentler and Newcomb (1978) offered some support for this view in a longitudinal study of newlywed couples. Although each spouse's level of extraversion had no independent association with their marital outcomes, the correlation between husbands' and wives' extraversion scores was significantly higher among couples who remained married than among couples who divorced.

Despite the intuitive appeal of this perspective, the accumulation of further support has been limited by a number of difficulties in assessing the effects of similarity in couples. First, the effects of similarity on a given trait are rarely independent of the effects of the trait itself. For example, Kurdek (1993) found that the difference between husbands' and wives' levels of neuroticism was greater in marriages that dissolved than in marriages that remained intact. This finding may support the role of personality similarity in marriage, but it may also result from the fact that differences between spouses are likely to be larger in couples where levels of neuroticism in either partner are higher overall. Estimating the unique effects of similarity between spouses requires that the effects of each spouse's personality be accounted for first. This is a control that has rarely been implemented.

A second difficulty with this kind of research is the fact that, as Kenny and Acitelli (1994) pointed out, 'A couple may appear to be very similar in their responses, yet they may be no more similar than a man and a woman who are totally unacquainted' (p. 419). Because there may be a typical way that most people tend to respond to personality inventories, the personalities of randomly matched people are likely to be similar to some degree. Estimating the unique degree of similarity in a given marriage requires that the 'typical cultural response' be controlled, but again this is a control that has rarely been implemented.

When similarity between spouses has been estimated accurately, support for beneficial effects of similar personalities has been sparse. For example, Russell and Wells (1991) reported that after controlling for the independent effects of each spouse's responses to a personality inventory, the association between marital satisfaction and the differences between spouses' responses was rarely significant. Furthermore, the overall association between spouses' personalities tends to be weak (Buss, 1984), even though spouses tend to be highly similar to each other on many other dimensions. In general, this research suggests that it may be spouses' similarity on those other dimensions, and in particular the degree to which spouses are similar in their values and attitudes, that may be more relevant to account for marital outcomes. With respect to personality, the nature of each spouse's personality appears to be more important than the degree of similarity between them.

In contrast to the similarity perspective, a competing view suggests that it is the degree of complementarity in spouses' personalities that is key to the successful marriage. For example, Winch (1958) argued that spouses select each other to resolve inadequacies in their own characters. Thus, extraverted people should be happiest with introverted partners and domineering people should be happiest with submissive partners. Although this is also an intuitively appealing idea, to date all attempts to obtain empirical support for the complementarity hypothesis have failed.

How Does Personality Affect Marital Outcomes?

Having established that personality traits, especially those associated with neuroticism, predict marital outcomes, the current focus of research on personality and marriage is to explain how this association comes about. Caspi et al. (1989), in their work on the effects of personality throughout the life course, proposed a broad framework that may be applied to understanding this issue. These researchers described two kinds of processes through which personality may affect development throughout the life course. 'Cumulative continuity' refers to all of the ways that individuals' personalities channel them towards certain environments and opportunities. 'Interactional continuity' refers to all of the ways that individuals' personalities affect their behavior towards and cognitions about other people. There is substantial evidence that both of these processes may account for the effects of spouses' personalities on their marriages.

Cumulative Continuity in Marriage

From early childhood, the personality of an individual affects not only the environments selected by the individual but also the opportunities available to the individual. Thus, for example, children who are prone to negative moods may choose career paths that are compatible with their temperament, but their choices may also be restricted by the behavioral consequences of their temperament (e.g., poorer educational attainment, lower professional accomplishment). In this way, individuals' personalities may have cumulative consequences that greatly shape the context of their adult lives. The nature of their adult environments may in turn affect the quality of

individuals' marital relationships. For example, women who experience chronic depression as adolescents tend to marry at a younger age compared to women who do not experience chronic depression (Gotlib, et al., 1998). Age at marriage is associated with risks of divorce, such that people who marry at a younger age are at higher risk for divorce than people who marry later in life. Thus, the stable tendency to experience negative affect in childhood appears to be associated, at least for women, with a life course that is detrimental to marriage.

In a similar vein, the personality characteristics of each spouse may give rise to stressful events and circumstances that impact the relationship. For example, individuals who score higher on measures of neuroticism describe events in their lives as more stressful than individuals scoring lower on such measures (Marco and Suls, 1993). Furthermore, controlling for these reporting effects, higher levels of neuroticism appear to be associated with greater numbers of stressful life events (Poulton and Andrews, 1992). To explain such findings, some researchers have suggested that individuals who are prone to experiencing negative affect frequently generate stressful circumstances that then serve to maintain their negative orientations. As a result of these processes, the marriages of individuals scoring higher in neuroticism may develop in more challenging circumstances than the marriages of less neurotic individuals. Because challenging circumstances, such as unemployment and financial difficulty, have been shown to predict divorce, this may be another route through which the cumulative consequences of each spouse's personality contribute to marital success and failure.

Interactional Continuity in Marriage

In addition to its role in shaping the circumstances of the marriage, personality may also affect interpersonal processes occurring within the marriage. Spouses' expectations for each other, the behaviors they exchange during their interactions, and the way they interpret those behaviors, have all been associated with the stable dispositions of each spouse. Through these mechanisms, the personality of each member of the couple may give rise to self-fulfilling prophecies, such that stable personality traits predict spouses' feelings towards each other and thus lead spouses to act in ways that confirm or enhance those feelings.

Support for these ideas has been especially strong with regard to negative affectivity, that is, neuroticism. For example, spouses who are higher in negative affectivity tend to make less charitable attributions for their partners' negative behaviors. That is, these spouses are more likely to hold their partners responsible for negative behaviors, whereas spouses who are lower in negative affectivity are more likely to excuse their partners for such behaviors (Karney, et al., 1994). Perhaps as a consequence, among newlywed couples, negative affectivity is associated with marital satisfaction, such that spouses who are higher in negative affectivity are likely to evaluate the marriage as a whole more poorly than spouses who are lower in negative affectivity (Karney and Bradbury, 1997). The stable tendency to experience negative affect thus appears to affect spouses' interpretations of the marriage at multiple levels, leading to partners that are perceived as less sensitive and marriages that are perceived as less rewarding.

The effects of each spouse's personality are not merely intrapsychic, however. In established marriages, the negative affectivity of husbands is associated with the marital satisfaction of their wives, such that the wives of husbands who are higher in negative affectivity tend to be less satisfied with the marriage, controlling for their own level of negative affectivity (Karney, et al., 1994). One way that the personality of one spouse may affect the satisfaction of the other is through the effects of personality on the behaviors that spouses exchange during their interactions with each other. Indeed, in established marriages, neuroticism is associated with couples' problem-solving behavior such that more neurotic spouses report feeling more negative than less neurotic spouses immediately after an attempt to resolve a marital problem (Geist and Gilbert, 1996). Furthermore, observational coding of spouses' behavior indicates that more neurotic wives express more negative emotions during conflict discussions. Given that the quality of couples' problem-solving behaviors predicts stability and change in the quality of the marriage over time (Karney and Bradbury, 1997), the link between neuroticism and marital interactions may be one way that the personality of an individual may influence the development of marital quality for a couple.

Future Directions

Although research on the interactional consequences of personality has focused almost exclusively on behavior during marital problem-solving interactions, personality may play an even more prominent role in the way spouses solicit and provide each other with social support. To the extent that personality, and in particular negative affectivity, affects the way spouses treat each other and the nature of the stressful events they encounter, then research on the provision of social support in response to stressful events would be a direction for future research that promises to integrate the cumulative and interactional consequences of personality in marriage.

Editor's Note

A great deal of new research on marriage and personality has appeared since Karney's entry was first published, some of it by the author himself. For a recent statement from the author, see Schoebi et al. (2012).

An important line of research in this area that is not explored in the original is the effects of personality traits on sexual and reproductive behavior. A recent examination of this question is Jokela et al. (2011).

As Karney notes, personality convergence in married couples may result from spousal selection processes (people seek out for marriage partners people who are similar to themselves) or from true convergence (people become more similar in personality traits the longer they are married). The bulk of current research points to 'assortative mating' and not convergence over time as the likeliest explanation. See Humbad et al. (2010); also Le Bon et al. (2013).

Another recent line of inquiry focuses on marital processes (satisfaction, quality of marriages and the like) as the

mechanisms through which personality traits influence health outcomes, for example, coronary arterial disease. For an interesting examination of such possibilities, see Smith et al. (2013).

For an economic analysis of marriage and personality; see Lundberg (2012), which explores the effects of personality traits on the formation and dissolution of domestic partnerships from an economist's perspective. And for an extension of many of the 'marriage and personality' themes to the case of same-sex couples, see the commentary by Cherlin (2013); and the several papers that follow.

See also: Attributional Processes: Psychological; Lifespan Development, Theory of; Marriage; Neuroticism; Personality Theory and Psychopathology; Personality and Adaptive Behaviors; Temperament and Human Development.

Bibliography

Bentler, P.M., Newcomb, M.D., 1978. Longitudinal study of marital success and failure. Journal of Consulting and Clinical Psychology 46, 1053–1070.

Burgess, E.W., Wallin, P., 1953. Engagement and Marriage. Lippincott, Chicago.

Buss, D.M., 1984. Marital assortment for personality dispositions: Assessment with three different data sources. Behavior Genetics 14, 111–123.

Caspi, A., Bem, D.J., Elder, G.H., 1989. Continuities and consequences of interactional styles across the life course. Journal of Personality 57, 376–406.

Cherlin, A.J., 2013. Health, marriage and same-sex partnerships. Journal of Health and Social Behavior 54 (1), 64–66.

Geist, R.L., Gilbert, D.G., 1996. Correlates of expressed and felt emotion during marital conflict: satisfaction, personality, process, and outcome. Personality and Individual Differences 21, 49–60.

Gotlib, I.H., Lewinsohn, P.M., Seeley, J.R., 1998. Consequences of depression during adolescence: Marital status and marital functioning in early adulthood. Journal of Abnormal Psychology 107, 686–690.

Humbad, M.N., Brent Donnellan, M., Iacono, W.G., McGue, M., Alexandra Burt, S., 2010. Is spousal similarity for personality a matter of convergence or selection? Personality and Individual Differences 49 (7), 827–830.

Jokela, M., Alvergne, A., Pollet, T.V., Lummaa, V., 2011. Reproductive behavior and personality traits of the five factor model. European Journal of Personality 25 (6), 487–500.

Karney, B.R., Bradbury, T.N., 1995. The longitudinal course of marital quality and stability: A review of theory, method, and research. Psychological Bulletin 118, 3–34.

Karney, B.R., Bradbury, T.N., 1997. Neuroticism, marital interaction, and the trajectory of marital satisfaction. Journal of Personality and Social Psychology 72, 1075–1092.

Karney, B.R., Bradbury, T.N., Fincham, F.D., Sullivan, K.T., 1994. The role of negative affectivity in the association between attributions and marital satisfaction. Journal of Personality and Social Psychology 66, 413–424.

Kelly, E.L., Conley, J.J., 1987. Personality and compatibility: A prospective analysis of marital stability and marital satisfaction. Journal of Personality and Social Psychology 52, 27–40.

Kenny, D.A., Acitelli, L.K., 1994. Measuring similarity in couples. Journal of Family Psychology 8, 417–431.

Kurdek, L.A., 1993. Predicting marital dissolution: A five-year prospective longitudinal study of newlywed couples. Journal of Personality and Social Psychology 64, 221–242.

Le Bon, O., Hansenne, M., Amaru, D., Albert, A., Ansseau, M., Dupont, S., 2013. assortative mating and personality in human couples: a study using cloninger's temperament and character inventory. Psychology 4 (1), 11–18.

Lundberg, S., 2012. Personality and marital surplus. IZA Journal of Labor economics 1 (3), 1–21.

Marco, C.A., Suls, J., 1993. Daily stress and the trajectory of mood: Spillover, response assimilation contrast, and chronic negative affectivity. Journal of Personality and Social Psychology 64, 1053–1063.

Poulton, R.G., Andrews, G., 1992. Personality as a cause of adverse life events. Acta Psychiatrica Scandanavica 85, 35–38.

Russell, R.J.H., Wells, P.A., 1991. Personality similarity and quality of marriage. Personality and Individual Differences 12, 407–412.

Schoebi, D., Karney, B.R., Bradbury, T.N., 2012. Stability and change in the first 10 years of marriage: Does commitment confer benefits beyond the effects of satisfaction? Journal of Personality and Social Psychology 102 (4), 729–742.

Smith, T.W., Baron, C.E., Grove, Jeremy L., 2013. Personality, emotional adjustment, and cardiovascular risk: Marriage as a mechanism. Journal of Personality. http://dx.doi.org/10.1111/jopy.12074. Wiley Periodicals, online prepublication.

Terman, L.M., 1938. Psychological Factors in Marital Happiness. McGraw-Hill, New York.

Winch, R.F., 1958. Mate Selection: A Study of Complementary Needs. Harper and Brothers, New York.

Personality and Religion

Vassilis Saroglou, Université catholique de Louvain, Louvain-la-Neuve, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Traditionally, religion has been perceived to shape individuals' personality (traits, but also values and social attitudes). I review here recent research describing the associations between personality and general religiosity, as well as modern spirituality, fundamentalism, and intense religious change. Additionally, longitudinal, genetic, and experimental studies help us to clarify the causal processes. Religion shapes, to some extent, personality but, inversely, genetic and environmental influences also affect the personality–religiosity links. Personality interacts with environmental factors to predict religiosity and with religion to predict relevant psychological outcomes.

For centuries, probably thousands of years, and across all human societies, people vary significantly on whether and to what extent they are religious or nonreligious. In other words, there exists important variability in 'religiousness.' Additionally, people differ in the specific ways (motivations, trajectories, cognitive styles, emotional quality) they are religious or irreligious. In other words, there exists important variability in religious forms. Finally, research in the behavioral and social sciences has shown that believers and nonbelievers, and the different subtypes of each, differ in many aspects of their individual and social lives (cognition, emotions, self-concept, interpersonal relations, social behavior, intergroup relations, family, work, economy, and politics). In other words, religiousness seems to have implications for human psychological functioning at the intraindividual, interpersonal, and group levels (Paloutzian and Park, 2013; Saroglou, 2014).

Personality–Religion Links: Theory

From a personality psychology perspective, it is reasonable to expect that individual differences in religiousness should be followed, to some extent, by personality differences. Religion, that is, beliefs, emotions and rituals, norms, and community, and their related positive or negative attitudes (religiousness), should influence the typical way people think, feel, and behave, that is, their personality. However, given that the basic 'deep' personality traits, also called 'first-level' personality, are mostly stable and enduring, as they are heavily dependent on genetic influences, biology, and temperament, religion's influence on personality can more easily be conceived for characteristics that are at the 'surface' level, also called 'second-' and 'third-level' of personality. These latter characteristics (e.g., social attitudes, identity, self-concept, values, meaning, and ultimate concerns) are more culturally sensitive and malleable and emerge later in life.

However, the opposite causal direction from personality to religiousness is equally if not even more plausible. People with certain basic personality traits, that is, personality predispositions, may be more or less likely to be interested and invest in religious beliefs, practices, and groups, because the latter may either correspond or not to their personality. Unbelievable but true, 'blind' observers are capable of guessing the religiosity of others based only on photographs of their face (Naumann et al., 2009). This causal direction may be direct, from personality to religiousness, or more complex. For instance, individual differences on both personality and religiousness may result from common underlying factors such as common genetic and environmental influences. Alternatively, religiousness, like social attitudes, value hierarchies, and political orientation, may be one of many other narrow characteristics, cultural adaptations of some broader personality predispositions, the latter importantly interacting with specific contextual and social factors (e.g., salience of religion during socialization or key life events).

Note that most scholars in the area of personality and religion agree that religiousness is a sui generis construct of individual differences that is closer, because of its ideological and moral components, but not reducible to values and social attitudes rather than to personality traits (e.g., Ashton, 2013; but see Piedmont, 1999, for an alternative view of spirituality as a basic personality dimension). Basic personality traits are transsituational, importantly heritable, enduring, and descriptive, whereas religiosity, values, and social attitudes are domain-specific, importantly subject to environmental influences, malleable, and rather normative. Empirical studies confirm stronger associations of religiousness with value hierarchies and relevant social attitudes than with personality traits (Roccas et al., 2002; Saroglou and Muñoz-García, 2008).

Research Trends

The study of the personality characteristics of believers, nonbelievers, and their different subtypes has been constant for many decades in the research at the intersection of personality psychology and psychology of religion. That research has often relied on self-reported measures for both personality and religiousness and has been carried out predominantly among Westerners of Christian tradition. Nevertheless, that research has included most of the different theories and models of personality of the last 50 years.

Since the late 90s, research on religion and personality has intensified. This research has increasingly used the five-factor model (FFM) of personality whose strengths, among others, are a structure with relative cultural universality and the capacity to

encompass and integrate into five major dimensions most of the personality traits that were identified in previous personality models. Moreover, recent research has started to investigate religion and personality in non-Christian and non-Western contexts. It has also used meta-analytic techniques, measures of personality other than self-reports (peer-ratings, behavioral measures, or text content analysis), and research designs beyond the simple collection of cross-sectional correlational data (longitudinal studies, laboratory experiments, and multilevel analyses of international data).

Personality–Religion Links: Empirical Evidence

After a short presentation of the FFM and two alternative modern models of similar structure (taxonomy of many traits under a few broad dimensions) that have been used in recent research on personality and religion, a synthesis of the main findings will be presented.

Major Personality Models and Traits

According to the FFM of personality, people differ from each other by being predisposed to typically think, feel, and behave (1) prosocially or egoistically when interacting interpersonally with others (agreeableness); (2) energetically or not in everyday life and when in group (extraversion); (3) orderly and methodically or without self-control when dealing with thoughts, feelings, and goals (conscientiousness); (4) stressfully or with emotional stability when facing challenges from the external and the internal world (neuroticism); and (5) openly and flexibly or conservatively in front of ideas and experiences marked by novelty, variety, and complexity (openness to experience).

Each of these five broad personality factors encompasses many distinct, lower-level, and thus narrower personality traits (facets) that typically point to two more specific tendencies within each factor. Agreeableness includes prosocial, other-oriented tendencies and social politeness; conscientiousness includes prohibitive and proactive tendencies; neuroticism combines emotional volatility and emotional withdrawal; extraversion points to energy but also to assertiveness; and, finally, openness to experience combines intellectual and strictly experiential openness. Some traditional personality traits are disseminated through their subcomponents into more than one factor within the FFM. For instance, the classic trait of impulsivity has its different components under three FFM factors. Other traits (e.g., humorousness or sexiness) are not necessarily captured by the FFM.

Older than the FFM, but still in use, the three-dimensional personality model of Eysenck includes extraversion and neuroticism, but also psychoticism. The latter is a global dimension that combines the low poles of agreeableness and conscientiousness and also includes impulsivity. However, while the two FFM factors point to the prosocial personality, psychoticism is a broader dimension indicating an antisocial and aggressive personality. Furthermore, HEXACO (Ashton, 2013), a newer personality model than the FFM, maintains the five dimensions overall (with minor changes in subcomponents between some factors), but adds a cross-culturally important sixth factor, that is, honesty/humility.

Personality Traits and Religiosity: Main Links

The vast majority of the more than 100 studies of the last 15–20 years have been carried out among Western believers and nonbelievers of Christian (Protestant and Catholic) tradition, but an increasing number of studies in Muslim, Jewish, and Buddhist religious and cultural contexts seem to extend the main lines of findings to non-Christian and non-Western contexts (see for two meta-analyses: Lodi-Smith and Roberts, 2007; Saroglou, 2010).

Indeed, rather consistently across the many studies, religiosity (intrinsic religiosity, global positive religious attitudes, religious beliefs and practice) first reflects prosocial tendencies when interacting with others: it is related to high agreeableness, usually to most of its facets except often modesty, and social warmth, a facet belonging – though marginally – to the extraversion factor. Second, religiosity implies high orderliness, methodicalness, and a tendency for self-control: it is associated with high conscientiousness, in its both inhibitive and proactive aspects, and low impulsivity, a trait whose subcomponents are located under three factors of the FFM and under psychoticism in the Eysenck's model. Third, religiosity is related to the single facet of (low) openness to values (ideological/moral conservatism) but not necessarily to the other components of the openness to experience factor (Saroglou, 2010; see **Figure 1** for the mean effect sizes). In line with the above, across the many dozens of studies that have used Eysenck's model, religiousness is typically associated with low psychoticism, the dimension of antisocial personality defined by coldness and distance in interpersonal relationships and high impulsivity (Francis and Penny, 2014; Lodi-Smith and Roberts, 2007). Finally, also in line with the above, studies on religion and the HEXACO personality traits confirm that religiosity also reflects honesty/humility, a factor

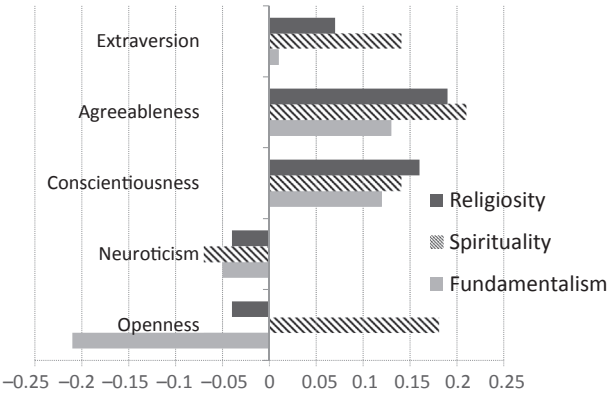

Figure 1 Mean effect sizes (weighted mean correlation coefficients), depicted in bars, of the associations of personality factors with religiosity, spirituality, and fundamentalism. Adapted from Saroglou, V., 2010. Religiousness as a cultural adaptation of basic traits: a five factor model perspective. Personality and Social Psychology Review 14, 108–125.

that is partly a blend of agreeableness and conscientiousness (Ashton, 2013).

The above findings indicate that a major underlying tendency of a religious personality is stability, in terms of both social harmony and internal self-control. Indeed, agreeableness and conscientiousness are known to constitute a higher-order factor denoting personal and interpersonal stability. Beyond these personality traits, religiosity is unrelated overall to the core aspects of extraversion (energy), neuroticism (negative emotionality), and experiential and intellectual openness. In other words, religiosity is overall orthogonal with respect to the traits that constitute the second higher-order factor of personality denoting plasticity and growth. These three personality traits are connected with religiosity only in some contexts, for some people, and for specific forms of religiousness (see a later section).

Robustness and Implications of the Links

Beyond occasional exceptions of studies with null findings, more than a hundred studies to date suggest that the above personality 'profile' of religious people may imply some universality. This profile is present in a large number of countries from different continents, across the major religions, different age groups, both genders, and seemingly various cohorts from the last 70 years. More importantly, ratings by peers (parents, teachers, siblings, colleagues, and 'blind' observers), behavioral studies (see a next section), text content analyses, and studies that have controlled for social desirability confirm that although the later may inflate the size of the associations, the religious personality 'profile' is not a simple artifact of social desirability (Saroglou, 2010).

Personality Differences beyond Traits

Moreover, studies that measure second- and third-level personality constructs (e.g., values, social attitudes, identity), which are known to adapt the basic personality traits of the first level (the big five factors) in specific life domains, typically confirm the above main lines of findings. Religious people typically tend to attribute (1) high importance to values denoting conservation (tradition and conformity) instead of openness to change (self-direction); (2) high value to self-transcendence in relationships with proximal others (benevolence, compassion, honesty, but not necessarily universalism); and (3) low importance to values that oppose self-restraint and self-control (hedonism, and to some extent, stimulation) (Saroglou, 2014). Similarly, religious adolescents and young adults prefer identity styles/statuses that emphasize engagement, possibly but not necessarily after exploration, that is, achievement or foreclosure (Saroglou, 2012, for review). Along the same line, studies have shown moderate links of general religiosity (and strong links of fundamentalism) with conservatism (political, ideological, and moral, but not necessarily economic) and authoritarianism (Paloutzian and Park, 2013; Saroglou, 2014), but overall no relation with social dominance (e.g., Hirsh et al., 2013). These social attitudes are known to reflect low openness to experience, but whereas authoritarianism and conservatism additionally relate to high conscientiousness, social dominance additionally reflects low agreeableness.

Cognitive Preferences

The personality characteristics of religiosity tend to translate into specific patterns and preferences in cognition, emotions, and behavior. For instance, studies among nonbelievers and believers socialized as Catholics and Protestants in Europe and the United States show a typical association between religiosity and, more strongly, fundamentalism, with the epistemic need for closure, and in particular with order and predictability – not necessarily closed-mindedness or an urgency to make decisions (e.g., Brandt and Reyna, 2010). This means that, although religiosity does not necessarily reflect simplistic thinking or dogmatism in beliefs and an indifference to opposite beliefs, it is marked by the need for order and structure in the inner world and thus the need for answers, as opposed to keeping existential and other questions open. Similarly, several studies across different countries and religions confirm that religious people tend to endorse just-world beliefs (like 'everybody gets what s/he deserves') and are prone to system justification.

Stereotypes and Meta-Stereotypes

Believers and nonbelievers hold reciprocal stereotypes (what ones think about others) and meta-stereotypes (what ones believe others think about them) in line with the findings on the religion-personality links. High versus low prosociality and conservatism, as well as low versus high hedonism seem to be associated with believers versus nonbelievers, respectively.

One may thus suspect that peer-validation and indirect, though proximal constructs, paper-and-pencil validation of religious self-reported personality tendencies may still suffer from various bias such as stereotypes ('s/he is religious, so s/he must be prosocial'), impression management ('I am religious, so I am prosocial'), or spurious accuracy (peers of religious targets, especially if selected by the latter, may just conform to stereotypes when evaluating the target) (see Galen, 2012). However, existing studies provide behavioral confirmation of the religious personality, at least to some extent and within some contexts.

Behavioral Implications

When results are significant, studies confirm that, under specific conditions (e.g., no conflicting principles, no threat of values) and with some limitations (in targets, motivations, and cost), religious people of various religious traditions tend to behave prosocially, showing for instance, nonaggression, generosity, reciprocity, and assistance (Saroglou, 2013). Religious people also tend to show many conscientiousness-related and control of impulsivity-related behaviors, like low tobacco and drug use, sexual infidelity, and crime, and high educational achievement as well as the capacity to delay gratification (McCullough and Carter, 2013). Moreover, an analysis of the European Social Survey data showed that people who chose fields of study and corresponding jobs involving care for others (teaching, education, personal care services, and medical/health services) had the highest mean scores on religiosity. The fields of study where mean religiosity was the lowest were engineering, sciences, mathematics, and computing, domains independent from prosocial concerns. Finally, religious people do not appreciate disparaging jokes regarding vulnerable targets (disabled people) and jokes that transgress social conventions (death- and disgust-related jokes) and also tend to inhibit spontaneous

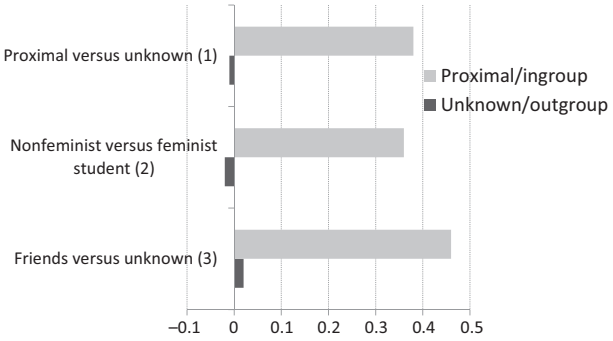

Figure 2 Correlation coefficients, depicted in bars, between religiosity and willingness to help proximal/ingroup versus unknown/outgroup targets in three studies: respectively, proximal versus unknown (1), nonfeminist versus feminist (2), and friends versus unknown people (3). Adapted from (1): Saroglou, et al., 2005. Prosocial behavior and religion: new evidence based on projective measures and peer ratings. Journal for the Scientific Study of Religion (Study 2); (2) and (3): Blogowska, Saroglou, 2011. Religious fundamentalism and limited prosociality as a function of the target. Journal for the Scientific Study of Religion (Studies 1 and 2).

creation of humor as a response to hypothetical hassles (Saroglou, 2014, for review). These behaviors clearly reflect high agreeableness and/or conscientiousness.

Note that the mean effect sizes of the associations of religiosity with personality traits and related attitudinal and behavioral outcomes are typically small, most often within the range of 0.15–0.25. Nevertheless, when focusing on many subtler personality traits/facets (instead of the broad personality factors), the explained variance of religiosity may increase up to 15–25% (Kandler and Riemann, 2013; Paunonen, 1998); and when adding values to personality traits, the explained variance may reach 17–45% (Roccas et al., 2002; Saroglou and Muñoz-García, 2008). Finally, as far as the behavioral implications of religious personality are concerned, note that it is the interaction between individual differences and situational factors that predict behavior better than either of these factors alone. For instance, across several studies, the religiosity-helping association ranges between 0.22 and 0.45 when the target is proximal or an ingroup member, but is around zero when the target is unknown or perceived to threaten religious values (see **Figure 2**).

Spirituality, Fundamentalism, and Intense Religious Change

Agreeableness and conscientiousness not only accompany general religiosity, but, as shown in many studies, also characterize – possibly in different strengths – specific religious forms. As shown in Saroglou's (2010) meta-analysis, modern spirituality, which may be independent from institutionalized religion, and mature (reflective) forms of religiosity, and also religious fundamentalism, are all associated with high scores on the above two personality factors. However, other personality traits importantly characterize these religious forms. Spirituality, an autonomous and nontraditional form of reference to some kind of transcendence, denotes growth/plasticity, that is, high extraversion and, more importantly, openness to experience. Religious fundamentalists, on the contrary, tend to be aversive to novelty, variety, and complexity: they tend to be low in openness to experience (see **Figure 1** for the mean correlations across studies).

Moreover, people differ in the way they are, become, or remain religious, spiritual, or atheist. An extensive body of research attests that people who have been religiously socialized, especially by parents to whom they have a secure attachment, develop secure working models of self and others (Granqvist et al., 2010). Their adult religiosity tends to be marked by positive emotionality, positive God representations, and the use of positive religious coping. On the contrary, people who have experienced 'dramatic,' intense conversions and people who have joined marginal small-size religious groups often tend to be characterized by an insecure attachment to parents, insecure adult attachment, a high need for closure, and neuroticism-related traits. They tend to use negative religious coping and have a negative God representation. Finally, atheists and deconverts, or at least some subtypes of them (note that research on personality and specific forms of irreligion is only emerging), tend to be high in openness to experience. Nevertheless, an area still open for future investigation is the extent to which dogmatic believers may resemble or differ from dogmatic, militant atheists.

There is some behavioral validation of the personality characteristics of these specific religious forms. Whereas traditional religiosity typically predicts prosociality toward proximal targets but not necessarily unknown targets and certainly not people perceived to threaten values, spirituality, which combines agreeableness with openness to experience, predicts altruism extended to outgroups, thus being universal. On the contrary, fundamentalism combining low openness with some agreeableness typically predicts outgroup prejudice and derogation, an effect partially explained by authoritarianism and the need for closure, but not by social dominance, a non-empathetic orientation (Saroglou, 2013, for review).

Age, Gender, and Culture

In his meta-analysis, Saroglou (2010) also investigated the potential moderating role of age, gender, and culture on the religion-personality traits associations. The associations of religiosity with agreeableness and conscientiousness, although consistently present across three age groups (adolescents, young adults, and adults), were stronger among older adults compared to younger adults. One interpretation is that religious personality becomes more 'personal' and intrinsic in adulthood, more strongly reflecting prosocial and order-related motives. Moreover, spirituality was more strongly associated with positive emotionality (high extraversion and low neuroticism) in adult samples relative to younger samples, which can be understood in terms of faith maturity: religious doubt reflects less stress and instability in adulthood than in earlier stages of life.

Women's higher prevalence of religiosity – across religions, especially in beliefs and private practice rather than public practice – may result from gender differences in personality (Francis and Penny, 2014). Women, or individuals of either gender who have a high femininity orientation, are higher in prosociality and orderliness, lower in risk-taking, and higher in

dependability and susceptibility to irrational ideas. Beyond these mean differences on both personality and religiosity, the existing evidence suggests that the associations of religiosity with the big five personality traits are similar for men and women.

Agreeableness and conscientiousness accompanied higher religiosity not only in studies from the United States, Canada, and Europe but also in studies from some other parts of the world – though underrepresented in the meta-analysis. Beyond these similarities, differences in the magnitude of effects were found when comparing the first three cultural zones, all of which have a predominantly Christian tradition. Compared to Europe, in the United States, religiosity reflected higher emotional stability (lower neuroticism), what possibly indicates more positive emotionality of religious expressions in today's American Protestantism compared to Europe, more historically marked by Catholicism's guilt. It also reflected higher agreeableness and no low openness to experience, what could be attributable to a stronger multicultural and multireligious tradition in the United States, comparatively with European countries marked by dominant mono-religious traditions. Finally, evidence from large international data (Gebauer et al., 2014) indicates that the associations of religiousness with agreeableness and conscientiousness (a *communion* orientation) are more pronounced in religious countries (religious people seek to assimilate into the predominant norms of these societies). In more secular societies these associations attenuate, and openness to experience and agency (competence, uniqueness, and ambition) become more important characteristics of religiousness (religious people may seek to be unique from the many secular others). These findings are in line with the idea of a shift, in secular societies, from traditional religiosity to modern spirituality.

Personality, Religion, and Behavior: Causal Processes

As stated earlier, theoretically, both directions of causality are conceivable: religion and religiousness may shape personality, and personality may contribute to individual differences on religiousness. Alternatively, personality and religiousness may both depend on third, common causes, or personality may interact with situational factors, thus influencing religiousness. Finally, religion and personality may interact to predict social behavior. Below, the empirical evidence for the above pathways will be presented.

Religious Influences on Personality

There is weak but not dismissible evidence that religiousness longitudinally predicts personality and personality changes and that religious stimulation affects to some extent personality, at least surface-level personality characteristics.

The maintenance of, or increase in, religious involvement among US adolescents was found to longitudinally predict, across late adolescence and emerging adulthood, commitment – but not exploration – in identity status, that is, an increase in foreclosure and a decrease of moratorium and diffusion (Hardy et al., 2011). Religious values of middle adolescents in Australia predicted an increase in agreeableness and a decrease in psychoticism 2 years later (Huuskes et al., 2013); a similar longitudinal influence of adolescent religiosity on agreeableness in late adulthood was found among US women (Wink et al., 2007).

Moreover, evidence from studies on conversion suggests that conversion has an impact on second- and third-level personality constructs (narrow traits, values, social attitudes, meaning, identity) rather than first-level basic personality traits, which are less changeable and more stable and enduring (Paloutzian and Park, 2013). Finally, a series of laboratory experiments show that implicit or explicit exposure to religious ideas and symbols activates agreeableness- and conscientiousness-related social attitudes, cognitions, and behaviors, such as (ingroup) prosociality, self-control, honesty, submission/conformity (Galen, 2012; Saroglou, 2013, for reviews), and decreased social dominance orientation (Hirsh et al., 2013).

Personality Influences on Religiousness

Longitudinal studies from the United States, Australia, and Belgium indicate that personality has chronological priority and an impact on religiousness (studies by Heaven, McCullough, Wink, and colleagues cited in Saroglou, 2010; see also Duriez et al., 2008). Across these studies, religiousness in late adolescence, adulthood, and late adulthood was predicted by high conscientiousness or low psychoticism, measured when the participants were children or adolescents. In addition, high versus low agreeableness in adolescence predicted respectively an increase versus decrease of religiousness throughout adulthood; and openness to experience in adolescence predicted spirituality in late adulthood. Finally, high versus low exploration in adolescent identity status later influenced, respectively, the symbolic versus literal way one approaches religious ideas.

Therefore, from a five-factor theory perspective, religiosity can be seen as one of many other characteristic adaptations, that is, later in life, and narrower in focus (like identities, values, social attitudes, and ideologies), contextual, and cultural adaptations of early basic personality predispositions – here a combination of agreeableness with conscientiousness and low impulsivity. Of course, there is no direct, mere transformation of personality traits to religiosity: value hierarchies and specific social attitudes, themselves being characteristic adaptations of basic personality traits, are stronger predictors of religiosity; additional individual differences (see below) may complement personality and value influences on later religiosity; and individual differences do not predict religiosity alone, but in addition to or in interaction with environmental factors, in particular religious socialization.

Indeed, in addition to basic personality traits, values, and social attitudes, religiosity is associated with (1) other personality characteristics that are located beyond the five factors such as honesty, conservatism, low sexiness, and low humorousness (Saroglou, 2014, for review), as well as (2) individual differences, beyond personality, that are more cognitive in nature, such as low intelligence (Zuckerman et al., 2013) and holistic and intuitive rather than analytic thinking (Gervais and Norenzayan, 2012).

More importantly, there is substantial evidence that an important, if not the major, predictor of religiosity is parents' attitudes toward religion. Religious socialization in general (mainly parents' influence, but also broader socialization) importantly undermines future religious attitudes. Finally, negative life events and also some self-transcendent positive experiences play a role in amplifying interest and investment in religion and spirituality (Saroglou, 2014, for review). Consequently, an integrative theoretical model that better identifies the role of personality in religiousness is one that puts personality in interaction with global environmental factors and personal life events (see Figure 3). Siblings' differential trajectories with regard to religion is the typical example attesting the influence of individual differences beyond, or in interaction with, environmental influences such as religious or atheist family socialization.

Therefore, people who are agreeable and conscientious, if religiously socialized (family and society's influences), and especially if they had developed a secure attachment to their parents, will tend overall to remain religious because religious ideas, emotions, and norms solidify their prosocial and conscientious tendencies. Religion emphasizes ideals of compassion, social harmony, social cohesion, personal stability, and self-mastery. Even if these people experience religious doubts and no longer believe, for example because of their cognitive and social development, their personality predispositions may prevent them from totally abandoning religion due to their desire to minimize conflict and maintain harmony with their religious families and peers.

Agreeable and conscientious people who are nonreligiously socialized, in principle, would tend to develop attitudes and endorse secular ideologies in favor of secular humanism, social justice, and pro-environmental concerns to express and solidify their personality dispositions and respective values. If however they have experienced insecure attachment to parents or in adulthood, or if they are confronted with significant negative life events, their agreeableness and conscientiousness may make them more predisposed than their nonagreeable and nonconscientious peers to be interested and invest in spirituality and religion in order to (re)-establish personal stability, meaning, trust, and faith to their values.

Beyond the above pathways, people who, in addition to agreeableness and conscientiousness, are also characterized by low openness to experience may turn to conservative and fundamentalism religious forms, whereas those with high openness and extraversion, especially in secular countries, will be more in osmosis with modern forms of spirituality.

Genetic and Environmental Influences on Religiousness

Several genetic studies have confirmed that individual differences in religiosity and specific forms of it, like religious fundamentalism, are due to both environmental (shared and nonshared, i.e., unique) and genetic influences (see also Figure 3). New studies helped to clarify the underlying processes. Genetic influences on religiosity are very small during adolescence; in that age period religiosity is heavily influenced by shared environmental factors, mainly family education and broad religious socialization. However, genetic influences become much more important in early adulthood, probably because young adults gain autonomy from family, allowing more fully for the expression of their genetic predispositions (Button et al., 2011; Koenig et al., 2005, 2008). Across studies, heritability estimates for religiosity range between 2 and 29% in adolescence, but between 27 and 65% in adulthood.

A reasonable hypothesis is that the genetic influences on religiosity can be explained through personality influences or that the links between personality and religiosity are due to common genetic influences. This was confirmed in three recent

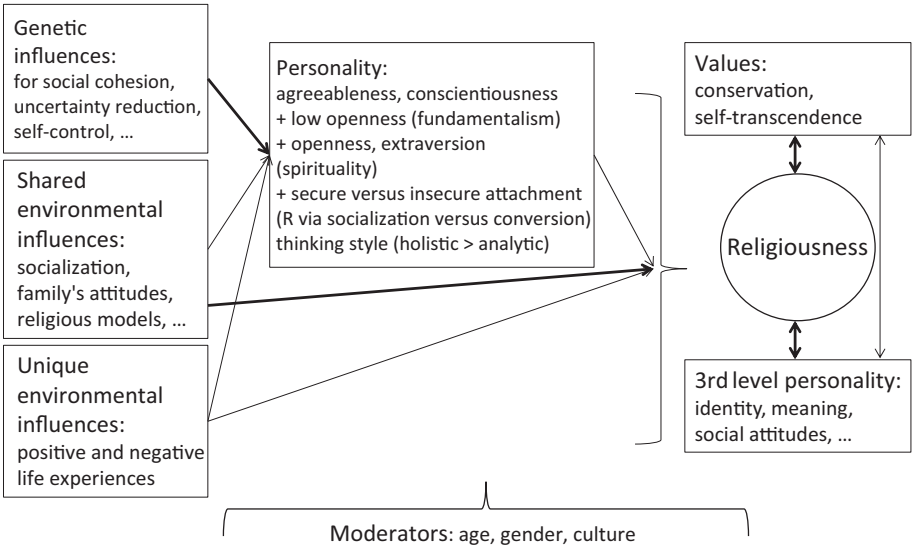

Figure 3 Theoretical model describing the genetic and environmental influences on religiousness and the role of personality characteristics: basic traits and other individual differences.

studies. In the first, the relationship between religiousness and antisocial behavior measured among 265 male twin pairs (US) was due to both genetic and shared environmental effects. Altruistic behavior also shared almost all of its genetic influence, but only half of its shared environmental influence, with religiousness (Koenig et al., 2007). In the second study on 993 pairs of twins (US), heritable effects underlying community integration and existential uncertainty strongly overlapped with the heritable influences on religiosity (Lewis and Bates, 2013). In the third study on 394 twin pairs (Germany), the associations between religiosity and personality traits in an older and a younger group (agreeableness and low openness, in both groups; and, additionally, in the younger group, conscientiousness) were entirely attributable to genetic influences in the older group and to genetic influences and shared environmental effects in the younger group (Kandler and Riemann, 2013).

There is also evidence that change in religiosity in late adolescence and young adulthood is mostly genetic in origin, with increased nonshared, unique environmental effects (e.g., personal experiences, significant negative or positive life events); stability is determined by both genes and family (shared) environment (Button et al., 2011; Koenig et al., 2008). Shared environmental effects were found to importantly mediate the parent-offspring correspondence in religiosity, but these decreased with age. On the contrary, the self-peers' (friends and spouse) correspondence on religiosity increased with age and was partly explained by genetic and nonshared environmental effects (Kandler and Riemann, 2013). Finally, there seems to exist a genotype-environment correlation in the personality characteristics of religiosity: parents who are agreeable and less open to values transmit both genetically and environmentally their personality dispositions as well as their religiosity to their offspring (Kandler and Riemann, 2013). Note also that, because of the assortative mating (future spouse-similarity on religiousness), the influence of heritability on offspring religiosity (possibly through personality) may have been underestimated.

Personality–Religion Interaction as Predicting Psychological Outcomes

A final way in which personality is involved in religion is that personality characteristics may interact with religious ideas, emotions, norms, and identities or with individual religiousness, to predict various psychological outcomes at all levels: intraindividual, interpersonal, and social.

Here are some illustrative examples coming from recent experimental studies (Saroglou, 2014, for review). In a series of four studies, priming religious concepts increased accessibility of submission-related thoughts, submissive behavior to the experimenter requesting retaliation, conformity to informational social influence, and deontological nonempathetic moral choices, but only among people who were high in dispositional submissiveness or authoritarianism. Religious priming was also found to increase prosocial behavior but only among participants with DRD4 (dopamine D4 receptor gene) susceptibility. Similarly, the priming of Buddhist concepts among Westerners decreased implicit ethnic prejudice, but only among participants who highly valued universalism. Finally, individuals with avoidant attachment style were found to be reluctant to use the compensatory mechanism of God as a safe haven in response to attachment threats.

From a personality psychology perspective, these findings confirm the idea that personality tendencies better predict relevant behavior in contexts and environments (here those involving religious stimulation) that correspond to, and encourage the expression of, such tendencies. From a psychology of religion perspective, these findings underline the fact that religious beliefs, rituals, norms, and groups do not necessarily have universal psychological potential but often seem to be relevant particularly for individuals with specific personality characteristics.

Conclusion

To conclude, personality is, to some extent and rather at the surface level, shaped by religion. More importantly, a specific pattern of personality dispositions (basic personality traits and corresponding values and social attitudes, as well as specific cognitive styles) influence later attitudes toward, specific forms of, and life span changes relative to, religion. Personality, mostly through genetic influences, in early adulthood 'clarifies' or removes the earlier family effects on religious attitudes, in addition to, and possibly in interaction with, significant life events. Finally, religion's influence on individuals' psychological functioning may also pass through, or interact with, personality characteristics. Search for quality in interpersonal relationships, social harmony and cohesion, and personal order and control, especially in the face of uncertainty, seem to be the key mechanisms beneath the personality-religion links. Motives for growth and plasticity do not seem to be the core business of religious personality, but their strength versus weakness importantly colors various forms and expressions of religion, mainly modern spirituality versus fundamentalism.

See also: Agreeableness; Attitudes and Behavior; Authoritarianism; Belief in a Just World; Conscientiousness; Five Factor Model of Personality, Facets of; Gender Differences in Personality and Social Behavior; Honesty and Humility; Openness to Experience; Personality Differences and Development: Genetic and Environmental Contributions; Personality, Trait Models of; Prejudice and Discrimination; Spirituality; Values, Social Psychology of.

Bibliography

Ashton, M.C., 2013. Individual Differences and Personality, second ed. Academic Press, San Diego, CA.

Brandt, M.J., Reyna, C., 2010. The role of prejudice and the need for closure in religious fundamentalism. Personality and Social Psychology Bulletin 36, 715–725.

Button, T.M., Stallings, M.C., Rhee, S.H., Corley, R.P., Hewitt, J.K., 2011. The etiology of stability and change in religious values and religious attendance. Behavior Genetics 41, 201–210.

Duriez, B., Smits, I., Goossens, L., 2008. The relation between identity styles and religiosity in adolescence: evidence from a longitudinal perspective. Personality and Individual Differences 44, 1022–1031.

Francis, L.J., Penny, G., 2014. Gender differences in religion. In: Saroglou, V. (Ed.), Religion, Personality, and Social Behavior. Psychology Press, New York, pp. 313–337.

Galen, L.W., 2012. Does religious belief promote prosociality? A critical examination. Psychological Bulletin 138, 876–906.

Gebauer, J.E., Bleidorn, W., Gosling, S.D., Rentfrow, P.J., Potter, J., 2014. Cross-cultural variations in Big Five relations with religiosity: a sociocultural motives perspective. Journal of Personality and Social Psychology. Advanced online publication.

Gervais, W.M., Norenzayan, A., 2012. Analytic thinking promotes religious disbelief. Science 336, 493–496.

Granqvist, P., Mikulincer, M., Shaver, P., 2010. Religion as attachment: normative processes and individual differences. Personality and Social Psychology Review 14, 49–59.

Hardy, S.A., Pratt, M.W., Pancer, S.M., Olsen, J.A., Lawford, H.L., 2011. Community and religious involvement as contexts of identity change across late adolescence and emerging adulthood. International Journal of Behavioral Development 35, 125–135.

Hirsh, J.B., Walberg, M.D., Peterson, J.B., 2013. Spiritual liberals and religious conservatives. Social Psychological and Personality Science 4, 14–20.

Huuskes, L., Ciarrochi, J., Heaven, P.C.L., 2013. The longitudinal relationships between adolescent religious values and personality. Journal of Research in Personality 47, 483–487.

Kandler, C., Riemann, R., 2013. Genetic and environmental sources of individual religiousness: the roles of individual personality traits and perceived environmental religiousness. Behavior Genetics 43, 297–313.

Koenig, L.B., McGue, M., Iacono, W.G., 2008. Stability and change in religiousness during emerging adulthood. Developmental Psychology 44, 532–543.

Koenig, L.B., McGue, M., Krueger, R.F., Bouchard Jr., T.J., 2007. Religiousness, antisocial behavior, and altruism: genetic and environmental mediation. Journal of Personality 75, 265–290.

Koenig, L.B., McGue, M., Krueger, R.F., Bouchard Jr., T.J., 2005. Genetic and environmental influences on religiousness: findings for retrospective and current religiousness ratings. Journal of Personality 73, 471–488.

Lewis, G.J., Bates, T.C., 2013. Common genetic influences underpin religiosity, community integration, and existential uncertainty. Journal of Research in Personality 47, 398–405.

Lodi-Smith, J., Roberts, B.W., 2007. Social investment and personality: a meta-analysis of the relationships of personality traits to investment in work, family, religion, and volunteerism. Personality and Social Psychology Review 11, 68–86.

McCullough, M.E., Carter, E., 2013. Religion and self-regulation. In: Pargament, K.I., Exline, J.J., Jones, J.W. (Eds.). APA Handbook of Psychology, Religion, and Spirituality, vol. 1. American Psychological Association, Washington, DC, pp. 123–138.

Naumann, L.P., Vazire, S., Rentfrow, P.J., Gosling, S.D., 2009. Personality judgments based on physical appearance. Personality and Social Psychology Bulletin 35, 1661–1671.

Piedmont, R.L., 1999. Does spirituality represent the sixth factor of personality? Spiritual transcendence and the five-factor model. Journal of Personality 67, 985–1013.

Paloutzian, R.F., Park, C.L. (Eds.), 2013. Handbook of the Psychology of Religion and Spirituality, second ed. Guilford Press, New York.

Paunonen, S.V., 1998. Hierarchical organization of personality and prediction of behavior. Journal of Personality and Social Psychology 74, 538–556.

Roccas, S., Sagiv, L., Schwartz, S.H., Knafo, A., 2002. The big five personality factors and personal values. Personality and Social Psychology Bulletin 28, 789–801.

Saroglou, V. (Ed.), 2014. Religion, Personality, and Social Behavior. Psychology Press, New York.

Saroglou, V., 2013. Religion, spirituality, and altruism. In: Pargament, K.I., Exline, J.J., Jones, J.W. (Eds.). APA Handbook of Psychology, Religion, and Spirituality, vol. 1. American Psychological Association, Washington, DC, pp. 439–457.

Saroglou, V., 2012. Adolescents' social development and the role of religion: coherence at the detriment of openness. In: Trommsdorff, G., Chen, X. (Eds.), Values, Religion, and Culture in Adolescent Development. Cambridge University Press, Cambridge, UK, pp. 391–423.

Saroglou, V., 2010. Religiousness as a cultural adaptation of basic traits: a five factor model perspective. Personality and Social Psychology Review 14, 108–125.

Saroglou, V., Muñoz-García, A., 2008. Individual differences in religion and spirituality: an issue of personality traits and/or values. Journal for the Scientific Study of Religion 47, 83–101.

Wink, P., Ciciolla, L., Dillon, M., Tracy, A., 2007. Religiousness, spiritual seeking, and personality: findings from a longitudinal study. Journal of Personality 75, 1051–1070.

Zuckerman, M., Silberman, J., Hall, J., 2013. The relation between intelligence and religiosity: a meta-analysis and some proposed explanations. Personality and Social Psychology Review 17, 325–354.

Personality and Risk Taking

Bernd Figner, Behavioural Science Institute, Radboud University Nijmegen, Nijmegen, The Netherlands; and Center for the Decision Sciences, Columbia University, New York, NY, USA
Elke U Weber, Center for the Decision Sciences, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

There is a long-standing and persistent belief that risk-taking is a stable personality trait, often referred to as risk attitude or risk preference. This belief implies that a given individual will take similar risks across a range of situations. The article reviews different definitions of risk attitude that show cross-situational consistency to varying degrees. Additionally, several important person-related and situation-related factors affecting risk-taking and its underlying mechanisms are reviewed, including risk perception, risk–value tradeoffs, affective and deliberative processes, and emotion regulation.

There is a long-standing and persistent belief that risk-taking is a stable personality trait, often referred to as risk attitude or risk preference. The belief implies that a given individual will take similar risks across a range of situations and that some people will be more risk-averse (or more risk-seeking) across situations than others. The article reviews different definitions of risk attitude that show cross-situational consistency to varying degrees and highlights several factors related to individual differences that are known to affect risk-taking. Section 'Domain and Framing Effects on Risk-Taking' shows that risk attitude defined within the expected utility (EU) framework varies greatly across situations as a function of decision content and outcome framing. Section 'Risk-Taking and Risk Perception' describes a more promising conceptualization of risk-taking, within a risk–value framework. It models risk-taking as a function of (1) decision-makers' perception of the riskiness and value or return of different courses of action, and (2) their attitude toward perceived risk (PRA), i.e., their willingness to trade off (perceived) risk for return. Two individuals might differ in their recreational pursuits, for example, either because they assess the relative risks of skydiving, bungee jumping, and playing poker very differently (based on past experience, person A may perceive playing poker to be riskier than skydiving and thus choose to go skydiving out of risk aversion – a negative attitude toward risk, while person B may perceive playing poker as the less risky option and engage in it, also out of risk aversion) or, whether their risk perceptions agree or not, because they have different attitudes toward risk as they see it (with persons C and D agreeing on the greater risk posed by skydiving, but person C being attracted by this risk and thus taking it on, and person D being repelled by it and thus choosing to play poker instead).

When modeled within this framework (as described in the section 'Perceived-Risk Attitude as a Stable Trait'), situational differences in risk-taking turn out to result from differences in the perception of risk in different situations rather than differences in willingness to take on (perceived) risk, thus restoring credibility to the notion of PRA as a stable trait. Individual differences in PRA exist, but are smaller and less systematic than individual and group differences in risk perception (with some notable exceptions, e.g., adolescent risk-taking). While the determinants of risk perception are relatively well known at this point not much is known about the determinants of PRA. As described in the section 'Personality, Risk Perception, and Perceived-Risk Attitude,' personality differences in variables known to be related to risk-taking seem to have their effect often via differences in risk perception. In the section 'Multiple Processes in Risk-Taking: Affect, Deliberation, Cognitive Control' we describe an important factor influencing risk-taking that is related to both situation and person, namely the type of the decision processes involved in the decision, in particular more 'hot' affective–motivational processes versus more 'cold' deliberative processes (Figner et al., 2009; Figner and Weber, 2011; Metcalfe and Mischel, 1999). Affective processes have received growing attention in the last two decades. It has become clear that they can influence risk-taking by different routes, for example, in the form of concrete emotions such as anger, fear, or sadness or by way of more subtle and persistent moods (Loewenstein et al., 2001; Peters et al., 2006). Further, it has been shown that (individual or experimentally induced) differences in the *regulation* of these affective phenomena also influence risky decisions (e.g., Heilman et al., 2010; Panno et al., 2013).

Domain and Framing Effects on Risk-Taking

In the EU framework, the dominant normative model of risky decision-making, the term 'risk-taking' is used to characterize choice patterns. Choice of a sure outcome over a lottery with equal expected value is modeled by a concave utility function and described as risk-averse; choice of the lottery is modeled by a convex utility function and described as risk-seeking. Despite the fact that risk-taking simply describes the curvature of the utility function that is derived from a series of choices, "those who coined the term *risk aversion* had in mind the psychological interpretation that someone who prefers the expected value of a gamble over playing the gamble *does not like to take risks*" (von Winterfeldt and Edwards, 1986: 256). Popular as well as managerial folklore tends to interpret risk-taking as a personality trait.

Bromiley and Curley (1992) review the evidence for risk-taking as a personality trait, i.e., as a preference for risk that is stable across situations, and find it lacking (see also Fox and

Tannenbaum (2011) for a more recent overview coming to a similar conclusion). Risk-taking seems to be influenced jointly by the situation and characteristics of the decision-maker, as well as by the interaction of the two (Figner and Weber, 2011). Decision domains in which the same person often shows different degrees of risk-taking include games of chance/gambling, financial investing, business decisions, health decisions, recreational choices, social choices, and ethical decisions (MacCrimmon and Wehrung, 1986; Blais and Weber, 2006; Weber et al., 2002). Further, even within the same domain, individuals can differ in their risk-taking depending on whether they make the decisions in an affect-charged versus a deliberative state (Figner et al., 2009; Figner and Weber, 2011; Gladwin et al., 2011). Modeling risk-taking within EU theory and defining risk attitude as the curvature of a utility function thus is clearly problematic for the notion of risk attitude as a personality trait. Attempts to restore cross-situational consistency to the construct of risk attitude by factoring differences in marginal value (e.g., the incremental value of an additional dollar or an additional life saved) out of the utility function were not successful (Keller, 1985; Weber and Milliman, 1997).

Prospect theory (Kahneman and Tversky, 1979) generalizes EU by postulating different utility functions for outcomes that are framed as gains as opposed to losses. For moderate to high probabilities, when outcomes are framed as gains, choices tend to be risk-averse; when the same outcomes are framed as losses (relative to a higher reference point), choices tend to be risk-seeking, further complicating the interpretation of risk-taking in the EU sense as a stable trait.

Risk-Taking and Risk Perception

In the risk–value framework, risk-taking is a compromise between greed (value) and fear (risk). Risk–value models in finance equate 'value' with the expected value of a risky option and 'risk' with its variance. Generalized risk–value models allow for a broader range of risk measures. Situational differences such as outcome framing also result in different risk perception (Mellers et al., 1997). As a result, apparent differences in risk-taking may be the result of differences in the perception of the riskiness of the choice options, and not of differences in attitude toward (perceived) risk. Cooper et al. (1988) report, for example, that – contrary to managerial folklore – the characteristic that differentiates entrepreneurs from other managers is not a more positive attitude toward risk, but instead an overly optimistic perception of the risks involved. For an outside observer who perceives risk more realistically, entrepreneurs will thus appear to take great risks. However, when differences in risk perception are factored out, entrepreneurs – just as other managers – demonstrate a preference for tasks that they see as only moderate in risk (Brockhaus, 1982).

Perceived-Risk Attitude as a Stable Trait

PRA is a measure of the degree to which individuals find perceived risk attractive (or unattractive) and therefore will choose alternatives that carry greater (or less) risk, all other things being equal. Weber and Milliman (1997) examined its cross-situational consistency by asking commuters to choose between pairs of trains that had risky arrival times (that depended on making a connection that had a stated probability) and to judge which of the two trains was the riskier one. The two trains in each pair had arrival times with equal expected value but different variance. Some pairs of trains had only positive arrival times (faster or equal to current travel times), others had only negative arrival times (slower or equal to the status quo). There was little consistency in people's risk-taking across the gain and the loss domain when risk-taking was defined in the EU sense. Few commuters had preferences that resulted in utility functions that were either both risk-seeking (convex) or both risk-averse (concave). However, consistency across the two domains was very high when PRAs were compared. The majority of commuters were risk-averse in both domains, i.e., consistently chose the train in a given pair that they had judged to be the less risky of the two.

In another study, MBA students participated in two sessions of an investment game where they had to pick one of six stocks (described by standard financial indicators) in each of 10 investment periods, and had to rate the riskiness of the stocks at different points throughout each session (Weber and Milliman, 1997). In one session, participants lost money in most of the 10 periods, whereas in the other session they mostly made money. Choices were very different across sessions (with more switching between stocks in the failure session), as were the ratings of the riskiness of the six stocks. However, over 80% of investors had the same PRA in both sessions, with three-quarters consistently investing in stocks that they perceived to be less risky and one-quarter consistently investing in stocks that they perceived to be more risky.

In a cross-national study, Weber and Hsee (1998) obtained risk judgments as well as minimum buying prices for risky financial investment options from respondents in the USA, Germany, the People's Republic of China, and Poland. Both risk judgments and buying prices showed significant cross-national differences, with Americans perceiving the most risk and Chinese paying the highest prices. However, after differences in risk perception were taken into consideration, the proportion of individuals who were perceived risk-averse or perceived risk-seeking were not significantly different in the four countries, with the majority again being perceived risk-averse, and only a small percentage in each country being perceived risk-seeking.

Personality, Risk Perception, and Perceived-Risk Attitude

Some psychologists have questioned the assumption of finance models that people will and should strive to minimize risk, arguing instead that people's ideal point for risk or uncertainty could differ, either as a personality difference (Lopes, 1987) or as a situational difference (Weber and Kirsner, 1997). Ideal-point models (Coombs, 1975) assume a person will perceive the riskiness of an alternative as the deviation between the alternative's level of uncertainty or unpredictability and the person's ideal point on the

uncertainty continuum. Perceived risk of an alternative with a high objective level of uncertainty would be high for a person with a low ideal point, but low for a person with a high ideal point. Individual differences in ideal points for risk and uncertainty have been measured by the construct of sensation seeking (Zuckerman, 1979), which seem to have some biological basis (Zuckerman et al., 1988) and vary with age and gender. Bromiley and Curley (1992) reported evidence linking sensation seeking to behavioral correlates that include greater risk-taking, especially in the health/safety and recreational domain. Weber et al. (2002) also reported high positive correlations between sensation seeking and its subscales in several content domains, with especially high correlations between the thrill-and-adventure-seeking subscale and recreational risk-taking and the disinhibition subscale and ethical risk-taking. Consistent with the predictions of ideal-point models, the path by which differences in sensation seeking seem to affect risk-taking appears to be differences in the perceptions of risk, rather than differences in PRA. For example, Weller and Tikir (2011) found both domain-*general* and domain-*specific* relationships between perceived risks and benefits and risk-taking on the one hand and HEXACO personality traits on the other hand: The personality trait emotionality was associated with heightened risk perception while the trait conscientiousness was associated with lower perceived benefits, both in a domain-general manner. In contrast, the trait openness was associated with risk-taking and perceived benefits in the social and recreational domain, while the trait honesty/humility was associated with greater risk-taking in the health/safety and ethical domain. These results show that personality traits can be related to both domain-specific and domain-general aspects of risk-return tradeoffs and risk attitudes.

Multiple Processes in Risk-Taking: Affect, Deliberation, Cognitive Control

Recent decision research has highlighted the role of different types of processes involved in risky decisions. Some risky decisions are made in emotionally charged contexts while other decisions are made in a cold deliberative state and these states can affect individuals' risk-taking levels (Figner and Weber, 2011). Figner, Weber, and colleagues developed a 'hot' affective and a 'cold' deliberative version of a risky choice task (the Columbia Card Task, CCT; Figner et al., 2009; Figner and Weber, 2011) and showed that age and other individual differences are not uniform across these types of risk-taking situations, highlighting both that context and involved processes are crucial to understanding risk-taking and that the assumption of a simple single unitary concept of risk-taking is problematic.

For example, adolescents, compared to adults, showed increased risk-taking levels only in the hot, but not the cold, CCT (Figner et al., 2009). This result points to the crucial role of affective processes as well as the (relative lack of) cognitive control processes in adolescent risk-taking and is consistent with currently popular neurodevelopmental models of adolescent risk-taking. This model explains adolescents' risk-taking tendency as a consequence of a developmentally transient potential for an imbalance between strong (subcortical) bottom-up affective–motivational processes versus still immature (prefrontal) top-down cognitive control processes (Gladwin et al., 2011; Somerville et al., 2010; see also Blakemore and Robbins, 2012; Crone and Dahl, 2012; Ernst and Fudge, 2009; Steinberg, 2010; Reyna and Farley, 2006). Thus, adolescent risk-taking is assumed to originate from an interaction between person (adolescents' potential imbalance between strong active affective–motivational processes and relatively weak cognitive control processes) and situation (only in situations in which substantial affective processes are triggered is increased adolescent risk-taking expected).

Further highlighting the role of contextual characteristics and involved processes, personality traits have been found to show differential associations with 'hot' versus 'cold' risky decisions. As expected, risk-taking levels in the hot, but not the cold CCT, were correlated with the motivational construct 'need-for-arousal' (Figner et al., 2009); this construct is related to Zuckerman's sensation seeking construct which, in a separate study, showed qualitatively the same associations with hot and cold risk-taking as need-for-arousal (Penolazzi et al., 2012). In contrast, risk-taking in the cold CCT are related to individual differences in executive functions such as planning and reasoning (Figner et al., 2009) as well as scores on the I7 scale that measures trait impulsivity (Penolazzi et al., 2012). An advantage of the CCT is that it allows for the decomposition of overt risk-taking levels into underlying processes, namely sensitivity to gain, loss, and probability as well as risk-return tradeoffs, which also show differential associations with personality variables. Individuals' reward responsiveness was correlated with increased gain sensitivity when losses were high in the hot CCT whereas individuals' sensitivity to punishment was related to participants' loss sensitivity in the cold CCT (Penolazzi et al., 2012).

Highlighting that not only affect, but also its *regulation* affects risk-taking, recent work has shown that individual (as well as experimentally induced) differences in the strategies people adopt to regulate emotions while making risky decisions lead to differences in risk-taking. For example, Panno et al. (2013) have shown that habitually increased reliance on cognitive reappraisal as emotion regulation strategy is associated with increased risk-taking levels in the cold CCT, accompanied by a reduced sensitivity to both the magnitude and the probability of a possible loss (see Heilman et al., 2010 for similar risk-taking results using experimentally induced emotion regulation strategies). Finally, clinical phenomena related to affective processes and/or their (relative lack of) regulatory processes have been shown to be related to risk-taking: for example, Di Rago et al. (2012) have shown that anxiety, but not depression, is related to decreased risk-taking (see also Maner et al., 2007; Giorgetta et al., 2012). Taken together, these results highlight the importance of both age-related and other individual differences such as personality traits (which also extend into the domain of clinically relevant phenomena) when investigating risk-taking levels; further, the differential associations with 'hot' versus 'cold' forms of risky decisions underscore the need to take into account the context and type of involved processes and the crucial role of affective processes.

Summary, Caveats, and Future Directions

The current research consensus suggests an interactional model of risk-taking (e.g., Sitkin and Weingart, 1995; Figner and Weber, 2011) in which situational characteristics and person-centered characteristics, as well as their interaction (e.g., in adolescents) jointly influence risk-taking. Situational constraints include the content domain of the risky decision, contextual variables such as outcome framing and aspiration levels (Lopes, 1987), and which psychological processes are involved in making risky decisions. Person-centered characteristics include age, gender, culture, and personality. Often, these variables influence risk-taking by changing people's perception of the riskiness of decision alternatives, rather than by affecting their willingness to take on more or less risk. However, these characteristics, e.g., age, can also interact with situational characteristics such that, e.g., excessive adolescent risk-taking is only likely to occur when affective processes are strongly involved and overwhelm adolescents' still immature cognitive control abilities that otherwise might counteract impulsive risk-taking, as it is usually the case in adults (Figner et al., 2009; Figner and Weber, 2011; Gladwin et al., 2011; Somerville et al., 2010).

Because of the domain specificity of risk-taking, measures of risk attitude that employ choice situations across a range of content domains (e.g., the Choice Dilemmas Questionnaire of Kogan and Wallach, 1964) have little predictive validity. Domain-specific scales of risk-taking that help to diagnose apparent differences in risk-taking into differences in either risk perception and/or PRA have been developed (Blais and Weber, 2006; Weber et al., 2002). Similarly, the involved processes matter, e.g., whether risky decisions are made using mainly deliberative calculus or 'hot' affective processes; tasks that systematically vary the involvement of these processes have been recently developed as well (Figner et al., 2009; Figner and Weber, 2011). Future research will provide additional insights into the complex interactions between personality and situation that have been explored for a range of other traits (Mischel, 1999) with respect to risk-taking. A combination of task analysis and theory about the reasons for risk-taking and its cognitive and emotional constraints should lead to the development of gender-, culture-, and domain-specific risk-taking profiles, that predict level of risk-taking in a situation- and person-contingent fashion.

See also: Decision Making, Psychology of; Decision Making: Nonrational Theories; Defiant Behavior During Adolescence across Cultures; Emotional Regulation; Flow in Motivational Psychology; Personality Changes During Adolescence Across Cultures; Personality and Adaptive Behaviors; Personality and Economics; Self-Regulated Learning; Self-Regulation During Adolescence: Variations Associated with Individual–Context Relations; Sensation Seeking: Behavioral Expressions and Biosocial Bases; Social Psychology; Stress in Adolescence: Effects on Development; Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model.

Bibliography

Blais, A.R., Weber, E.U., 2006. A domain-specific risk-taking (DOSPERT) scale for adult populations. Judgment and Decision Making 1, 33–47.

Blakemore, S.J., Robbins, T.W., 2012. Decision-making in the adolescent brain. Nature Neuroscience 15, 1184–1191.

Brockhaus, R.H., 1982. The psychology of the entrepreneur. In: Kent, C.A., Sexton, D.L., Vesper, K.G. (Eds.), The Encyclopedia of Entrepreneurship. Prentice Hall, Englewood Cliffs, NJ, pp. 92–94.

Bromiley, P., Curley, S.P., 1992. Individual differences in risk-taking. In: Yates, J.F. (Ed.), Risk-Taking Behavior. Wiley, New York, pp. 87–132.

Coombs, C.H., 1975. Portfolio theory and the measurement of risk. In: Kaplan, M.F., Schwartz, S. (Eds.), Human Judgment and Decision. Academic Press, New York, pp. 63–68.

Cooper, A.C., Woo, C.Y., Dunkelberg, W.C., 1988. Entrepreneurs' perceived chances for success. Journal of Business Venturing 3, 97–108.

Crone, E., Dahl, R.E., 2012. Understanding adolescence as a period of social–affective engagement and goal flexibility. Nature Reviews Neuroscience 13, 636–650.

Di Rago, V., Panno, A., Lauriola, M., Figner, B., 2012. All negative affects are not equal: anxiety but not depression predicts risk-taking in behavioral tasks. Psychology & Health 27 (Suppl. 1), 195.

Ernst, M., Fudge, J.L., 2009. A developmental neurobiological model of motivated behavior: anatomy, connectivity, and ontogeny of the triadic nodes. Neuroscience and Biobehavioral Reviews 33, 367–382.

Figner, B., Weber, E.U., 2011. Who takes risks when and why? Determinants of risk taking. Current Directions in Psychological Science 20, 211–216.

Figner, B., Mackinlay, R.J., Wilkening, F., Weber, E.U., 2009. Affective and deliberative processes in risky choice: age differences in risk taking in the Columbia Card Task. Journal of Experimental Psychology: Learning, Memory, and Cognition 35, 709–730.

Fox, C.R., Tannenbaum, D., 2011. The elusive search for stable risk preferences. Frontiers in Psychology 2, 298. http://dx.doi.org/10.3389/fpsyg.2011.00298.

Giorgetta, C., Grecucci, A., Zanon, S., Perini, L., Balestrieri, M., Bonini, N., Sanfey, A.G., Brambilla, P., 2012. Reduced risk-taking behaviour as a trait feature of anxiety. Emotion 12, 1373–1383.

Gladwin, T.E., Figner, B., Crone, E.A., Wiers, R.W., 2011. Addiction, adolescence, and the integration of control and motivation. Cognitive Developmental Neuroscience 1, 364–376.

Heilman, R.M., Crisan, L.G., Houser, D., Miclea, M., Miu, A.C., 2010. Emotion regulation and decision making under risk and uncertainty. Emotion 10, 257–265.

Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47, 263–291.

Keller, L.R., 1985. An empirical investigation of relative risk aversion. IEEE Transactions on Systems, Man, and Cybernetics 15, 475–482.

Kogan, N., Wallach, M.A., 1964. Risk-Taking: A Study in Cognition and Personality. Holt, New York.

Loewenstein, G.F., Weber, E.U., Hsee, C.K., Welch, N., 2001. Risk as feelings. Psychological Bulletin 127 (2), 267–286.

Lopes, L.L., 1987. Between hope and fear: the psychology of risk. Advances in Experimental Social Psychology 20, 255–295.

MacCrimmon, K.R., Wehrung, D.A., 1986. Taking Risks: The Management of Uncertainty. Free Press, New York.

Maner, J.L., Richey, J.A., Cromer, K., Mallott, M., Lejuez, C.W., Joiner, T.E., et al., 2007. Dispositional anxiety and risk-avoidant decision-making. Personality and Individual Differences 42, 665–675.

March, J.G., Shapira, Z., 1992. Variable risk preferences and focus of attention. Psychological Review 99, 172–183.

Mellers, B.A., Schwartz, A., Weber, E.U., 1997. Do risk attitudes reflect in the eye of the beholder? In: Marley, A.A.J. (Ed.), Choice, Decision, and Measurement. Erlbaum, Hillsdale, NJ, pp. 57–71.

Metcalfe, J., Mischel, W., 1999. A hot/cool-system analysis of delay of gratification: dynamics of willpower. Psychological Review 106, 3–19.

Mischel, W., 1999. Implications of person–situation interaction: getting over the field's borderline personality disorder. European Journal of Personality 5, 455–461.

Panno, A., Lauriola, M., Figner, B., 2013. Emotion regulation and risk taking: predicting risky choice in deliberative decision making. Cognition & Emotion 27, 326–334.

Penolazzi, B., Gremigni, P., Russo, P.M., 2012. Impulsivity and reward sensitivity differentially influence affective and deliberative risky decision making. Personality and Individual Differences 53, 655–659.

Peters, E., Vastfjall, D., Garling, T., Slovic, P., 2006. Affect and decision making: a "hot" topic. Journal of Behavioral Decision Making 19, 79–85.

Reyna, V.F., Farley, F., 2006. Risk and rationality in adolescent decision making – implications for theory, practice, and public policy. Psychological Science in the Public Interest 7, 1–44.
Sitkin, S.B., Weingart, L.R., 1995. Determinants of risky decision making behavior: a test of the mediating role of risk perceptions and risk propensity. Academy of Management Journal 38, 1573–1592.
Somerville, L.H., Jones, R.M., Casey, B.J., 2010. A time of change: behavioral and neural correlates of adolescent sensitivity to appetitive and aversive environmental cues. Brain and Cognition 72, 124–133.
Spear, L.P., 2011. Rewards, aversions and affect in adolescence: emerging convergences across laboratory animal and human data. Developmental Cognitive Neuroscience 1, 390–403.
Steinberg, L., 2010. A dual systems model of adolescent risk-taking. Developmental Psychobiology 52, 216–224.
von Winterfeldt, D., Edwards, W., 1986. Decision Analysis and Behavioral Research. Cambridge University Press, Cambridge, UK.
Weber, E.U., Hsee, C.K., 1998. Cross-cultural differences in risk perception but cross-cultural similarities in attitudes towards risk. Management Science 44, 1205–1217.
Weber, E.U., Kirsner, B., 1997. Reasons for rank-dependent utility evaluation. Journal of Risk and Uncertainty 14, 41–61.
Weber, E.U., Milliman, R.A., 1997. Perceived risk attitudes: relating risk perception to risky choice. Management Science 43, 122–144.
Weber, E.U., Blais, A.R., Betz, N.E., 2002. A domain-specific risk-attitude scale: measuring risk perceptions and risk behaviors. Journal of Behavioral Decision Making 15, 263–290.
Weller, J.A., Tikir, A., 2011. Predicting domain-specific risk taking with the HEXACO personality structure. Journal of Behavioral Decision Making 24, 180–211.
Zuckerman, M., 1979. Sensation Seeking: Beyond the Optimal Level of Arousal. Erlbaum, Hillsdale, NJ.
Zuckerman, M., Simons, R.F., Como, P.G., 1988. Sensation seeking and stimulus intensity as modulators of cortical, cardiovascular, and electrodermal response: a cross-modality study. Personality and Individual Differences 9, 361–372.

Relevant Websites

http://columbiacardtask.org – For the Columbia Card Task (CCT).
http://dospert.org – For the Domain-Specific Risk-Taking (DOSPERT) Scale.

Personality and Self

Michael D Robinson and Clay Routledge, North Dakota State University, Fargo, ND, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article seeks to integrate personality and social psychology in a manner that enriches both, particularly with reference to the self and its processes. It is suggested that the typical approach to personality assessment must somewhat necessarily be assessing variations in people's self-concepts and that the beliefs assessed are likely to be generalized (i.e., not very contextual) in nature. Such generalized views of the self will have predictive value, but also predictive limitations. It is further suggested that social psychological views of how personality traits operate may have considerable merit, although there are several ways in which the personality trait literature counters or at least qualifies assumptions made in the social psychology literature. Overall, the personality–self interface appears to be a particularly generative one in understanding individual differences and how they function.

Personality psychology seeks to describe and potentially understand how people differ from each other. It tends to focus on the trait construct – presumably stable, cross-situational tendencies to be a particular type of person. For example, we might state that Suzy is extraverted and Ann is introverted. If this is true, Suzy should be more talkative, dominant, and gregarious than Ann across a wide variety of contexts or situations. Social *psychology*, in contrast to personality psychology, typically views behavior as more malleable and as a consequence of particular situational factors. From this perspective, it may be a mistake to attribute extraversion to Suzy and introversion to Ann in that both individuals may be relatively extraverted when interacting with friends but relatively introverted when in class (Ross and Nisbett, 1991).

A watershed event for both personality and social psychology was the publication of Mischel's (1968) book on personality and assessment. Mischel concluded that traits are rather modest in their ability to predict behaviors in specific situations, that the trait construct may be flawed for this reason, and that it might be more profitable to understand behavior in terms of its situational elicitors. Such conclusions gave rise to what is termed the *person–situation debate*: Is behavior better predicted by knowing about the person or about the situation? In a classic study of this type, Darley and Batson (1973) found that aspects of the person (e.g., intrinsic religiosity) were nonsignificant predictors of helping a potentially-ill confederate, whereas a manipulated situational factor – available time – had a significant influence. The person–situation debate was a source of considerable friction for many years, but it is now apparent that reconciliation is possible. For example, personality psychologists might improve their predictions by attending to situational context, and individuals may differ in their susceptibility to situational influence (Kenrick and Funder, 1988).

Nonetheless, personality and social psychologists continue to pursue different themes while operating on different premises. It is still true that personality psychologists generally adopt correlational designs, whereas social psychologists generally adopt experimental designs (Kenrick and Funder, 1988). And it is still true that there are questions concerning the malleability of behavior from situation to situation (Mischel, 2009). Of particular pertinence to this article, personality and social psychologists also differ in their preferred individual difference variables: Personality psychologists have embraced the trait construct, whereas social psychologists are much more likely to focus on the self-concept and on self-related motives (e.g., self-enhancement and self-protection), oftentimes in the context of transient experiences. The purpose of this article is to highlight ways in which personality and social psychology can inform each other, especially with respect to the personality–self interface. We start by asking a series of questions.

What Do We Assess When We Assess Personality Traits?

Conceptually, personality traits define the major ways in which people differ from each other. For example, the trait of neuroticism encompasses tendencies toward negative emotional experiences, self-consciousness, and stress-related impulsivity. One could potentially assess personality traits in a number of ways (e.g., clinical observations), but in fact one method predominates. People are asked to characterize themselves in relation to adjectives (e.g., 'nervous') or short phrases (e.g., 'worry a lot') consistent with possessing the trait in question (here, neuroticism). There are good reasons to assess personality traits in this manner, perhaps the most important one being that people have considerable knowledge concerning their thoughts, feelings, and behaviors, knowledge that typically outstrips that available to observers.

In assessing personality traits in this way, though, it must be recognized that we are more or less asking people about their *self-concepts* – that is, how the self is viewed. It should, therefore, not be surprising that there are strong relationships between personality trait assessments and assessments of aspects of the self. This is true, for example, for measures seeking to assess the certainty of the self-concept (Campbell et al., 1996), the extent to which one engages in self-deception (Paulhus and John, 1998), and the extent to which one is

high in self-control (Tangney et al., 2004). Moreover, there are strong relationships between *self-esteem* – whether the person views the self as worthy or unworthy – and personality traits as they are typically assessed. In fact, it is now clear that self-esteem may actually constitute a sort of higher-order glue that pervades all five of the 'Big 5' traits – extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience (e.g., Erdle et al., 2010). We do not wish to suggest that personality assessments are synonymous with self-esteem, but we do wish to highlight the importance of the self-concept to what is captured by self-reports of the Big 5.

What Type of Self-Knowledge Is Assessed by Personality Trait Measures?

There are different levels of personality. McAdams (2012) distinguishes personality traits from life narratives, the latter being more fleshed out, intimate, and closely tied to the ongoing goals of the individual. Mischel and Shoda (1995) contrast personality traits, which are situation-general, from a personality-in-context approach contending that consistency can be better understood in situation-specific ('bottom-up') terms. Robinson and Clore (2002a) contrast assessments of current or recent behaviors and experiences with assessments that ask people to characterize themselves over long time frames such as several months, years, or in general.

What is common to all of these frameworks is the idea that personality traits, as typically assessed, are capturing *generalized beliefs about the self*. The focus is on abstract self-knowledge (McAdams, 2012) and long rather than short time frames (Robinson and Clore, 2002b). Along the latter lines, Robinson and Clore (2002b) asked people to report on their emotions over seven time frames varying from short (e.g., last few hours) to medium (e.g., last few months) to trait (i.e., in general). For time frames shorter than the last few months, longer time frames resulted in longer latencies of judgment, consistent with a retrieve and aggregate reporting strategy – that is, recalling many distinct experiences and averaging across them in one's head. For time frames longer than the last few weeks, response times became faster, suggesting abandonment of a retrieve and aggregate strategy in favor of the consultation of more generalized beliefs about the self. When people report on their emotions in general (or over long time frames), that is, they do not appear to retrieve specific information about their actual emotional experiences in real time. **Figure 1** presents data of this type.

While Robinson and Clore (2002b) focused on the effects of time frame, Schell et al. (1996) focused on the influence of context. Participants in their study were asked to judge the self's traits *in college* versus *in general*. After making a self-judgment in relation to a particular adjective (e.g., helpful), participants had to recall a relevant exemplar behavior that they had engaged in (e.g., the time I helped Lucy move in to her new apartment). A control condition was also used in which people merely defined each trait term as it appeared. Very different findings occurred for the self in college versus in general. For the college self, recall was facilitated, suggesting that people recalled relevant behaviors when making these self-judgments. For the self in general, no such facilitation occurred,

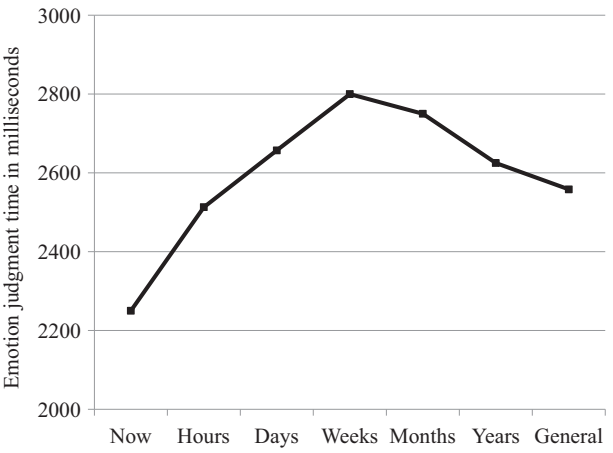

Figure 1 Emotion judgment time as a function of time frame (right now, last few hours, last few days, last few weeks, last few months, last few years, and in general).

suggesting that people made these judgments without recalling any specific information that could support them. Again, it appears that trait judgments (i.e., what one is generally like) are not based on retrieving specific information about what one has done, thought, or felt in a real-world context.

Klein et al. (1996) extended this analysis in a very interesting way. They studied the trait judgments of a woman who had been in a terrible motorcycle accident that produced pronounced retrograde amnesia – that is, the inability to recall particular experiences or events from the past. Despite her inability to recall specific experiences, or specific behaviors that she had engaged in, the woman's trait judgments were reliable and valid. Her conceptualization of her own personality traits agreed strongly with knowledgeable informants, for example. In understanding these findings, Klein et al. (1996) contrasted two memory systems, one episodic (i.e., event specific) and centered in the hippocampus and the other semantic and centered in the frontal cortex. Although this woman's hippocampal memory system was severely compromised, she could nonetheless make reliable and valid trait judgments because doing so involves a frontal memory system responsible for much more general self-views – that is., those not tied to any specific memories, past experiences, or behaviors.

Can Such an Analysis Explain Why Trait Judgments Are So Stable Over Time?

Although personality changes over time to some extent, McCrae and Costa (1994) made the rather convincing case that personality trait stability is the more robust phenomenon. In this research, the authors examined test–retest stability for each of the Big 5 personality traits over periods of 5–10 years. The stability coefficients ranged from 0.6 to 0.8. These stability coefficients do not approach 1, the highest value possible, but they are quite high. Moreover, McCrae and Costa (1994) note that such stability coefficients are necessarily limited by the reliability of the assessment scales, which ranged from 0.7 to 0.9.

Thus, personality traits are almost as stable over time as they are reliable at a given time point, a rather remarkable set of findings. The authors further noted that the stability coefficients did not vary by the trait in question, sex or race, or as a function of whether people thought that their personalities had changed over time or not.

In the preceding section, we made a case for the idea that people retrieve what may be termed *crystalized* (i.e., relatively fixed) self-knowledge when making judgments about their personality traits. This perspective can account for the results of McCrae and Costa (1994). Specifically, if people develop crystalized notions of what the self is like, and if this is the sort of knowledge that people retrieve when making trait self-judgments (Klein et al., 1996), then one would expect a high degree of stability of such self-judgments over time. This perspective can explain other phenomena as well. Eder and Mangelsdorf (1997) report that the trait self-judgments of children are quite a bit less reliable than those of adults, differences that can be ascribed (at least in part) to a lack of crystalized notions of what the self is generally like. Relatedly, Watson (2000) reports that trait emotion reports are more stable over time than state (i.e., more recent or current) reports of emotion, precisely as should be expected on the basis of the idea that trait judgments are differentially made by retrieving crystalized self-knowledge.

McCrae and Costa (1994) note that their analysis is somewhat pessimistic on the idea that psychotherapy could change a person's neuroticism level substantially. We agree with this skepticism on the basis of the idea that psychotherapy would be unlikely to greatly affect crystalized notions concerning the self if they are somewhat fixed, as we have contended. On the other hand, it is certainly possible to alleviate and even eliminate more state-related experiences of anxiety and depression on the basis of psychotherapy. Such points were made by Widiger et al. (1999): Successful psychotherapy does not alter trait levels of neuroticism much if at all, but does reduce state-related experiences of distress to a considerable extent. Thus, and as will be further discussed, trait assessments are not particularly sensitive to the ebb-and-flow of our actual emotions and behaviors.

To What Extent Is Self-Esteem Malleable?

Following the influential statements of early sociologists (e.g., Cooley, 1902), social psychologists often view self-esteem as quite malleable. It should rise and fall by successes and failures, whether in achievement-related or social realms. Contrary to this perspective, when self-esteem is assessed in trait-related terms, it does not exhibit these properties. The test–retest stability of self-esteem is remarkably high over time periods as long as 75 years (Trzesniewski et al., 2003). Moreover, the heritability of self-esteem (i.e., the percentage of variance attributed to genetic factors) is quite comparable to that of the Big 5 personality traits (Neiss et al., 2006). Finally, we know of very few studies that have shown that trait reports of self-esteem can be altered by situational feedback. A personality trait perspective can explain why self-esteem does not possess the malleability that was initially proposed in the social psychology literature.

Can the Present Analysis Help Us Understand Trait–State Relations?

When we assess traits, we assess general, crystalized self-knowledge. By contrast, when we assess states or momentary behaviors, we do not. Therefore, personality traits should somewhat necessarily be imperfect predictors of everyday experiences and behaviors. In the emotion realm, Robinson and Clore (2002a) review evidence for dissociations of this type. For example, women report themselves to be more emotional than men on a number of trait emotionality scales, but when emotionality is assessed in state-related terms, sex differences are far less evident. A particularly compelling study was that of Barrett et al. (1998), who asked men and women to complete an experience-sampling protocol and also report on their trait emotionality. Although women were more emotional on the trait scales, they were not more emotional in their everyday lives. On the basis of findings of this type, Robinson and Clore (2002a) suggested that women believe themselves to be more emotional than men to a much greater extent than is actually the case.

A related question is whether trait self-knowledge can mislead us concerning past experiences that we have had. In speaking to this question, Larsen (1992) had participants complete a trait neuroticism measure, report on somatic symptoms (e.g., fatigue, headache, and nausea) in everyday life, and then recall their overall levels of somatic symptoms during this time period. The trait of neuroticism predicted daily somatic symptoms, but it also biased the recall of such symptoms independent of their actual occurrence. That is, as levels of trait neuroticism increased, people were more likely to recall symptoms as having been more pronounced than they actually had been. More work of this type should be performed. For example, would extraversion (agreeableness) bias the extent to which participants thought that they had engaged in extraverted (agreeable) behavior in daily life? Based on extensions of the Robinson and Clore (2002a) model, such retrospective biases should be evident.

There is further support for the Robinson and Clore (2002a) model. A premise of it is that personality traits, as they are typically assessed, constitute a type of generalized self-knowledge that people can default to when they lack more specific information (e.g., when they are asked to recall experiences they may have forgotten). If so, people who have difficulty ascribing meaning to events as they occur should tend to make judgments that are more trait consistent. In two investigations, abilities to assign meaning to events as they occur were assessed in terms of the speed with which people could assign words to different meaning categories. Fast categorizers can assign meaning quickly, whereas slow categorizers cannot. We predicted that trait–outcome relations would be pronounced among slow categorizers, but potentially absent among fast categorizers. In support of these ideas, Robinson and Oishi (2006) found that the trait of extraversion predicted life satisfaction to a much greater extent for slow categorizers than fast categorizers (see **Figure 2** for such a pattern), and Robinson and Clore (2007) found conceptually identical results for relations between neuroticism and somatic symptom reports (see also Robinson et al., 2006). These results encourage us to think of trait assessments in terms of generalized self-knowledge

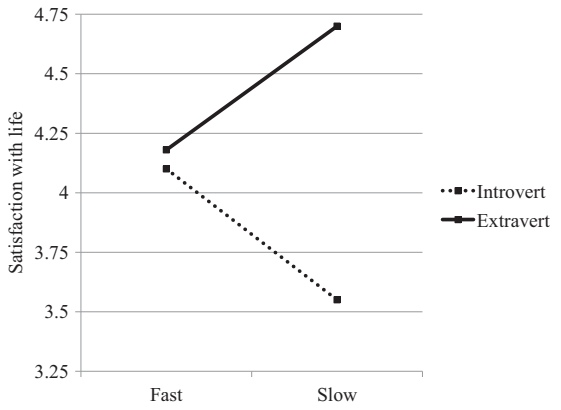

Figure 2 Life satisfaction as a function of categorization speed (fast vs slow) and extraversion level (introvert vs extravert).

that can 'contaminate' more state-related judgments among some people.

Are the latter sorts of effects merely about self-reports, though? We would like to suggest that they extend beyond them. In support of this idea, Robinson and Oishi (2006) found that slow extraverts (relative to fast extraverts or introverts) displayed greater happiness in pictures, a finding that cannot be ascribed to merely *self-reporting* greater happiness (see also Robinson and Cervone, 2006). Similar considerations apply to the results of Robinson et al. (2006). Thus, we would like to suggest that some people (e.g., slow categorizers) live lives that are more influenced by their personality traits than others (e.g., fast categorizers), an idea that extends the analysis of Robinson and Clore (2002a).

Are There Limitations to the Trait Assessment Approach?

Following the analysis above, it may be useful to recognize that people are not always very 'traited' in their everyday experiences and behaviors. For example, nearly all people engage in extraverted behavior in some situations, but introverted behavior in others (Fleeson, 2007). In addition, the situations that activate extraverted behavior are likely to differ across individuals (Cervone, 2004). These sorts of considerations have given rise to an alternative perspective on individual difference consistency that views it in terms of if–then situation–behavior profiles (Mischel and Shoda, 1995). The idea is that people are consistent in how they respond to particular situations over time (e.g., Cathy is almost always talkative at home), but not in how they respond to different situations at a given time (e.g., Cathy is rarely talkative at school, though she is at home). The trait approach will miss such nuances.

Another observation worth making is that personality trait measures assess explicit or conscious views of the self. Such views have their limitations in that people may lack awareness concerning some important aspects of the self and its tendencies (Vazire and Wilson, 2012). Supporting this point, there is a long line of research suggesting that motives can often operate implicitly or unconsciously. Implicit measures of motivation do not tend to correlate with explicit measures, yet predict spontaneous and long-term outcomes consistent with possessing the motive in question (e.g., argumentativeness as a function of power motivation: McClelland et al., 1989).

There is another body of work showing that Implicit Association Tests can be used to assess the implicit self-concept. Such assessments are reliable (Schnabel et al., 2006), tend to be only weakly correlated with their explicit (trait) counterparts (Egloff and Schmukle, 2002), and yet predict relevant behavioral outcomes. For example, implicit views of oneself as shy predict global ratings of anxiousness in stressful situations (Schnabel et al., 2006), dysfluent speech (Egloff and Schmukle, 2002), and a variety of other nonverbal indicators of distress in social situations (Back et al., 2009). The trait assessment approach, then, can miss out on some important process-based predictors of relevant outcomes (Robinson and Wilkowski, 2014).

Can Social Cognitive Views of the Self Speak to How Traits Operate?

The trait approach is primarily descriptive – that is, it describes how individuals differ from each other. In understanding how traits function, though, social cognitive views of the self, and particularly its motives, may represent a useful addition to the literature. Paulhus and John (1998) make the case that *self-enhancement* motives (i.e., to view oneself as positively as possible) may account for why people generally characterize themselves in socially desirable terms. Individual differences in operations of a self-enhancement motive have been most convincingly established for the trait of narcissism (Paulhus and Williams, 2002), but can be extended to the traits of the Big 5 (Paulhus and John, 1998).

In addition, people are motivated to confirm existing views of the self – a motive termed *self-verification* (Swann, 1992). This important motive can explain why personality traits are remarkably stable over time (McCrae and Costa, 1994). It can also explain other features of how traits function. For example, people high in the trait of neuroticism often seem to create distress for the self, whether due to their tendencies to appraise events negatively, habitual worry, poor coping choices, or negative interpersonal behaviors (Suls and Martin, 2005). Tamir (2005) has shown that people high in neuroticism will *choose* to worry in some performance contexts and Tamir and Robinson (2004) have shown that negative emotional experiences seem to facilitate affective understanding for neurotic (but not non-neurotic) individuals. At some paradoxical level, then, people high in neuroticism seem to prefer to be distressed – not because it is pleasurable, but because it fits existing conceptions of the self. Such self-verification perspectives and processes should be extended to the other Big 5 traits.

Turning to a different self-related motive, terror-management theory (TMT) posits a *self-transcendence* need (i.e., the need to perceive the self or part of it as symbolically transcending physical death) that compels people to identify with their own cultural beliefs while discriminating against out-group members who do not share such beliefs, among other dependent measures (Solomon et al., 2000). Terror-management processes are typically examined by randomly assigning some people to

write about what would happen with the self's death versus a control topic, a manipulation termed mortality salience (MS). Initial research on TMT examined the effects of MS in normative terms – that is, effects that presumably characterize all people equally (Solomon et al., 2000).

It is increasingly apparent, however, that the effects of MS manipulations can inform our understanding of how personality traits operate. Harmon-Jones et al. (1997) first reported that people low, but not high, in trait self-esteem were more affected by such manipulations, results consistent with the idea that higher levels of self-esteem serve a terror-management function. Indeed, Routledge et al. (2010) observed that MS increased distress and decreased well-being only among individuals low in trait self-esteem. Another body of work has shown that 'creatureliness' (i.e., having bodies that eat, sweat, have sex, and so on) appears to be more threatening as neuroticism levels increase. Under conditions of MS, people high (but not low) in neuroticism seek to avoid activities that remind them of their bodies (e.g., Goldenberg et al., 2006). TMT perspectives of attachment styles (Cox et al., 2008) and personal need for structure (Landau et al., 2004) have also been productive. Such findings are not just informative to the TMT literature, but potentially also to the personality trait literature.

Another body of work has highlighted the value of attending to *self-regulation* processes in understanding the personality trait of agreeableness. Agreeableness is a strong inverse predictor of anger and aggression, but there are several alternative reasons why this might be the case. To disentangle these alternatives, Meier and Robinson (2004) measured the accessibility of hostile thoughts because hostile thoughts are a major contributor to anger and aggression. Agreeable people did not have less accessible hostile thoughts, but such thoughts were less likely to result in anger and aggression at high (relative to low) levels of agreeableness. See **Figure 3** for an idealized pattern of this type. In other words, it appears that agreeable people self-regulate the influence of their hostile thoughts, essentially remaining agreeable despite them.

Subsequent studies have provided direct evidence for agreeable self-regulation. Wilkowski et al. (2006) found that agreeable people shifted spatial attention away from hostile stimuli, whereas disagreeable people did not. Meier et al. (2006) found that agreeable people recruited prosocial thoughts when primed by hostile ones, whereas disagreeable people did not. More recently, Bresin et al. (2012) found that hostile thoughts facilitated motor control among agreeable people, but not disagreeable people, and that this 'agreeable' response to hostile primes predicted lesser anger reactivity in daily life. Other studies have shown that agreeable people (in these studies, in terms of low-trait anger) recruit cognitive control when primed by hostile thoughts (Wilkowski and Robinson, 2008) and pause following their negative evaluations (Robinson et al., 2012). Altogether, the studies converge on the idea that agreeable people possess remarkable self-regulatory mechanisms that can explain why they are less predisposed to anger and aggression. Process-based analyses of this type can be applied to other personality traits as well (Robinson and Wilkowski, 2014).

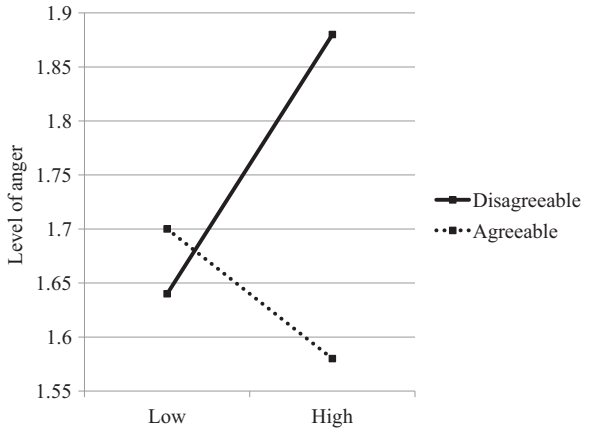

Figure 3 Level of anger as a function of hostile thought accessibility (low vs high) and agreeableness level (disagreeable vs agreeable).

Conclusions

The interest in behavior and experience is shared by personality and social psychologists, but their favored constructs, methods, and assumptions differ. Despite such differences, these subdisciplines of psychology can inform each other, as this article makes clear. We first suggest that personality traits, as typically assessed, must somewhat necessarily be assessing aspects of the self-concept. The fact that there are five relatively orthogonal dimensions along which self-concepts differ is a contribution to the social psychology literature. There are different levels of self-knowledge, perhaps most clearly articulated in Klein's (e.g., Klein et al., 1996) work. From the perspective of this work, it is apparent that personality traits capture generalized or crystalized beliefs about the self that are independent of particular contexts or time frames. Such assessment methods can explain why trait judgments are so stable over time, a contribution to the personality literature. Although several theorists contend that self-esteem is quite malleable, this appears not to be the case, a contribution to the social psychology literature.

Because personality traits assess generalized self-knowledge, we should expect systematic divergences between trait and state assessments and some divergences of this type (e.g., concerning sex differences in emotionality) were reviewed. Moreover, this perspective on what is assessed by personality traits possesses implications for understanding variability in trait–state or trait–outcome relations across people. In specific terms, and following Robinson and Clore's (2002a) model, it was proposed that traits should be predictive of outcomes or states particularly among people less capable of assigning meaning to events as they occur. Some evidence of this type was reviewed, work that highlights a role for cognitive processes in understanding trait–state relationships. The final two sections suggested potential limitations to the trait assessment approach and called for complementing social-cognitive analyses.

In sum, we wish to emphasize the following points. The typical approach to assessing personality traits must somewhat

necessarily be assessing what people think of themselves. That such assessments are useful in outcome prediction may be understood in terms of self-related processes to a greater extent than previously recognized. If so, self-related perspectives of personality traits possess considerable merit in understanding how they operate.

See also: Core Self-Evaluations; Five Factor Model of Personality, Facets of; Personality, Trait Models of; Self and Identity Development During Adolescence across Cultures; Self in Culture: Early Development; Self-Development in Childhood and Adolescence.

Bibliography

Back, M.D., Schmukle, S.C., Egloff, B., 2009. Predicting actual behavior from the explicit and implicit self-concept of personality. Journal of Personality and Social Psychology 97, 533–548.

Barrett, L.F., Robin, L., Pietromonaco, P.R., Eyssell, K.M., 1998. Are women the 'more emotional' sex? Evidence from emotional experiences in social context. Cognition and Emotion 12, 555–578.

Bresin, K., Fetterman, A.K., Robinson, M.D., 2012. Motor control accuracy: a consequential probe of individual differences in emotion regulation. Emotion 12, 479–486.

Campbell, J.D., Trapnell, P.D., Heine, S.J., Katz, I.M., Lavallee, L.F., Lehman, D.R., 1996. Self-concept clarity: measurement, personality correlates, and cultural boundaries. Journal of Personality and Social Psychology 70, 141–156.

Cervone, D., 2004. The architecture of personality. Psychological Review 111, 183–204.

Cooley, C.H., 1902. Human Nature and the Social Order. Charles Scribner's Sons, New York.

Cox, C.R., Arndt, J., Pyszczynski, T., Greenberg, J., Abdollahi, A., Solomon, S., 2008. Terror management and adults' attachment to their parents: the safe haven remains. Journal of Personality and Social Psychology 94, 696–717.

Darley, J.M., Batson, C.D., 1973. 'From Jerusalem to Jericho': a study of situational and dispositional variables in helping behavior. Journal of Personality and Social Psychology 27, 100–108.

Eder, R.A., Mangelsdorf, S.C., 1997. The emotional basis of early personality development: implications for the emergent self-concept. In: Hogan, R., Johnson, J.A., Briggs, S.R. (Eds.), Handbook of Personality. Academic Press, San Diego, CA, pp. 209–240.

Egloff, B., Schmukle, S.C., 2002. Predictive validity of an implicit association test for assessing anxiety. Journal of Personality and Social Psychology 83, 1441–1455.

Erdle, S., Irwing, P., Rushton, J.P., Park, J., 2010. The general factor of personality and its relation to self-esteem in 628 640 internet respondents. Personality and Individual Differences 48, 343–346.

Fleeson, W., 2007. Situation-based contingencies underlying trait-content manifestation in behavior. Journal of Personality 75, 825–862.

Goldenberg, J.L., Hart, J., Pyszczynski, T., Warnica, G.M., Landau, M., Thomas, L., 2006. Ambivalence toward the body: death, neuroticism, and the flight from physical sensation. Personality and Social Psychology Bulletin 32, 1264–1277.

Harmon-Jones, E., Simon, L., Greenberg, J., Pyszczynski, T., Solomon, S., McGregor, H., 1997. Terror management theory and self-esteem: evidence that increased self-esteem reduced mortality salience effects. Journal of Personality and Social Psychology 72, 24–36.

Kenrick, D.T., Funder, D.C., 1988. Profiting from controversy: lessons from the person-situation debate. American Psychologist 43, 23–34.

Klein, S.B., Loftus, J., Kihlstrom, J.F., 1996. Self-knowledge of an amnesic patient: toward a neuropsychology of personality and social psychology. Journal of Experimental Psychology: General 125, 250–260.

Landau, M.J., Johns, M., Greenberg, J., Pyszczynski, T., Martens, A., Goldenberg, J.L., et al., 2004. A function of form: terror management and structuring the social world. Journal of Personality and Social Psychology 87, 190–210.

Larsen, R.J., 1992. Neuroticism and the selective encoding and retrieval of symptoms: evidence from a combined concurrent-retrospective study. Journal of Personality and Social Psychology 62, 480–488.

McAdams, D.R., 2012. Meaning and personality. In: Wong, P.T.P. (Ed.), The Human Quest for Meaning: Theory, Research, and Applications, second ed. Routledge, New York, pp. 107–123.

McClelland, D.C., Koestner, R., Weinberger, J., 1989. How do self-attributed and implicit motives differ? Psychological Review 96, 690–702.

McCrae, R.R., Costa, P.T., 1994. The stability of personality: observations and evaluations. Current Directions in Psychological Science 3, 173–175.

Meier, B.P., Robinson, M.D., 2004. Does quick to blame mean quick to anger? The role of agreeableness in dissociating blame and anger. Personality and Social Psychology Bulletin 30, 856–867.

Meier, B.P., Robinson, M.D., Wilkowski, B.M., 2006. Turning the other cheek: agreeableness and the regulation of aggression-related primes. Psychological Science 17, 136–142.

Mischel, W., 1968. Personality and Assessment. John Wiley & Sons, Hoboken, NJ.

Mischel, W., 2009. From personality and assessment (1968) to personality science, 2009. Journal of Research in Personality 43, 282–290.

Mischel, W., Shoda, Y., 1995. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. Psychological Review 102, 246–268.

Neiss, M.B., Sedikides, C., Stevenson, J., 2006. Genetic influences on level and stability of self-esteem. Self and Identity 5, 247–266.

Paulhus, D.L., John, O.P., 1998. Egoistic and moralistic biases in self-perception: the interplay of self-deceptive styles with basic traits and motives. Journal of Personality 66, 1025–1060.

Paulhus, D.L., Williams, K.M., 2002. The dark triad of personality: narcissism, Machiavellianism, and psychopathy. Journal of Research in Personality 36, 556–563.

Robinson, M.D., Cervone, D., 2006. Riding a wave of self-esteem: perseverative tendencies as dispositional forces. Journal of Experimental Social Psychology 42, 103–111.

Robinson, M.D., Clore, G.L., 2002a. Belief and feeling: evidence for an accessibility model of emotional self-report. Psychological Bulletin 128, 934–960.

Robinson, M.D., Clore, G.L., 2002b. Episodic and semantic knowledge in emotional self-report: evidence for two judgment processes. Journal of Personality and Social Psychology 83, 193–215.

Robinson, M.D., Clore, G.L., 2007. Traits, states, and encoding speed: support for a top-down view of neuroticism/state relations. Journal of Personality 75, 95–120.

Robinson, M.D., Oishi, S., 2006. Trait self-report as a 'fill in' belief system: categorization speed moderates the extraversion/life satisfaction relation. Self and Identity 5, 15–34.

Robinson, M.D., Wilkowski, B.M., 2014. Personality processes and processes as personality: a cognitive perspective. In: Mikulincer, M., Shaver, P.R. (Eds.), APA Handbook of Personality and Social Psychology. American Psychological Association, Washington, DC, pp. 129–145.

Robinson, M.D., Wilkowski, B.M., Kirkeby, B.S., Meier, B.P., 2006. Stuck in a rut: perseverative response tendencies and the neuroticism-distress relationship. Journal of Experimental Psychology: General 135, 78–91.

Robinson, M.D., Wilkowski, B.M., Meier, B.P., Moeller, S.K., Fetterman, A.K., 2012. Counting to ten milliseconds: low-anger, but not high-anger, individuals pause following negative evaluations. Cognition and Emotion 26, 261–281.

Ross, L., Nisbett, R.E., 1991. The Person and the Situation: Perspectives of Social Psychology. Mcgraw-Hill Book Company, New York.

Routledge, C., Ostafin, B., Juhl, J., Sedikides, C., Cox, C., Liao, J., 2010. Adjusting to death: the effects of self-esteem and mortality salience on well-being, growth motivation, and maladaptive behavior. Journal of Personality and Social Psychology 99, 897–916.

Schell, T.L., Klein, S.B., Babey, S.H., 1996. Testing a hierarchical model of self-knowledge. Psychological Science 7, 170–173.

Schnabel, K., Banse, R., Asendorpf, J.B., 2006. Assessment of implicit personality self-concept using the implicit association test (IAT): concurrent assessment of anxiousness and angriness. British Journal of Social Psychology 45, 373–396.

Solomon, S., Greenberg, J., Pyszczynski, T., 2000. Pride and prejudice: fear of death and social behavior. Current Directions in Psychological Science 9, 200–204.

Suls, J., Martin, R., 2005. The daily life of the garden-variety neurotic: reactivity, stressor exposure, mood spillover, and maladaptive coping. Journal of Personality 73, 1485–1510.

Swann, W.B., 1992. Seeking 'truth,' finding despair: some unhappy consequences of a negative self-concept. Current Directions in Psychological Science 1, 15–18.

Tamir, M., 2005. Don't worry, be happy? Neuroticism, trait-consistent affect regulation, and performance. Journal of Personality and Social Psychology 89, 449–461.

Tamir, M., Robinson, M.D., 2004. Knowing good from bad: the paradox of neuroticism, negative affect, and evaluative processing. Journal of Personality and Social Psychology 87, 913–925.

Tangney, J.P., Baumeister, R.F., Boone, A.L., 2004. High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. Journal of Personality 72, 271–322.

Trzesniewski, K.H., Donnellan, M.B., Robinson, R.W., 2003. Stability of self-esteem across the life span. Journal of Personality and Social Psychology 84, 205–220.

Vazire, S., Wilson, T.D., 2012. Handbook of Self-knowledge. Guilford Press, New York.

Watson, D., 2000. Mood and Temperament. Guilford Press, New York.

Widiger, T.A., Verheul, R., van den Brink, W., 1999. Personality and psychotherapy. In: Pervin, L.A., John, O.P. (Eds.), Handbook of Personality Psychology: Theory and Research, second ed. Guilford Press, New York, pp. 347–366.

Wilkowski, B.M., Robinson, M.D., 2008. Guarding against hostile thoughts: trait anger and the recruitment of cognitive control. Emotion 8, 578–583.

Wilkowski, B.M., Robinson, M.D., Meier, B.P., 2006. Agreeableness and the prolonged spatial processing of antisocial and prosocial information. Journal of Research in Personality 40, 1152–1168.

Personality and Values at Work

Ryan Klinger and Mark Mallon, Old Dominion University, Norfolk, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Individuals' personality and values have important implications in the workplace and can be used in conjunction with other metrics to predict outcomes relevant to organizations, such as job performance. Instruments for assessing personality and values can thus be potent tools in hiring and promotion processes. In this article, we review the history of research surrounding personality and values in the workplace, discuss recent contributions to the literature, and identify existing issues and areas of future research.

Two decades ago, noted organizational scholar Adrian Furnham lamented that "organizational behavior theorists and management scientists have neglected to examine individual differences in any systematic way … while personality theorists have been eager to examine clinical, educational, medical and social correlates of individual differences/personality dimensions, they have consistently ignored occupational/organizational correlates" (Furnham, 1992: p. 2). A surge of research over the last 20 years has addressed this oversight. Hundreds of empirical studies and dozens of meta-analyses have explored relationships between individual differences, such as personality and values, and a vast array of important workplace attitudes and behaviors.

In this article, we explore the roles of personality and values in the workplace, their impact on work outcomes, and implications concerning organizational policies, practices, and systems. We define personality as the relatively enduring structures and propensities within individuals that explain their generalized affective, cognitive, and behavioral tendencies (Hogan et al., 1996). Values reflect preferences for certain situations, objects, goals, or behaviors over others (Schwartz, 1999). A subtle yet important distinction between personality and values is that the former describes one's social reputation, whereas the latter describes one's preferences. In the following sections, we review the literature regarding these two individual difference variables, focusing on empirical findings, recent developments in paradigmatic focus, and practical implications. We also highlight existing questions, emerging issues, and areas in need of future research.

Personality

The History of Personality in the Workplace

The above-mentioned dearth of empirical research relating personality to workplace outcomes through the late 1980s could be attributed to several causes. Critics of personality (with roots traced to behavioralism) argued that individual behavior was a function of the situation rather than any internal forces, that researchers overestimated behavioral consistency, and that relationships between personality traits and behaviors were illusory in nature. In addition, reviews of what little empirical evidence existed offered rather disappointing conclusions. For instance, Guion and Gottier (1965) analyzed the literature regarding the use of personality tests in organizational contexts and concluded, "It is difficult in the face of this summary to advocate, with a clear conscience, the use of personality measures in most situations as a basis for making employment decisions about people" (p. 160).

At least three factors have accounted for increased optimism and research regarding the role of personality in the workplace. First, longitudinal studies revealing the stability of personality over time and the impressive relationships between personality and outcomes measured years apart put to rest several criticisms regarding the illusory nature of this construct. For instance, Finn (1986) found considerable stability in personality traits across a 30-year span. Moreover, Judge et al. (1999) found childhood measures of personality to predict career success assessed 50 years later.

Second, a taxonomy of personality traits – referred to as the 'Big Five' or 'Five Factor Model' for its aggregation of personality into five broad trait domains (see **Table 1**) – gained widespread acceptance in the field of personality psychology (*see* Five Factor Model of Personality, Facets of). Before the introduction of the Big Five to the organizational sciences, the accumulation of

Table 1 The Big Five personality traits

Personality trait	*Basis*	*Facets*
Conscientiousness	Need for achievement and commitment to work	Competence; order; dutifulness; achievement striving; self-discipline; deliberation
Agreeableness	Quality of social interaction	Trust; straightforwardness; altruism; compliance; modesty; tendermindedness
Emotional stability	Propensity toward negative emotions and anxiety	Anxiety; hostility; depression; self-consciousness; impulsiveness; vulnerability to stress
Openness to experience	Appreciation for art, emotion, and new things	Fantasy; aesthetics; feelings; actions; ideas; values
Extroversion	Preferred quantity of social interaction	Warmth; gregariousness; assertiveness; activity; excitement-seeking; positive emotion

personality trait findings was hampered by the vast number of different traits being investigated. Consensus regarding the nomenclature and hierarchical organization of traits allowed researchers to better aggregate and communicate empirical findings. This also led to a shift of focus from very narrow, specific personality traits optimized for predicting narrow behavioral facets to broad personality domains seemingly better suited for predicting broad, complex work outcomes like job performance and turnover.

Finally, accompanying the emergence of the Big Five taxonomy was the introduction of meta-analytic techniques to the organizational literature (*see* Systematic Reviewing and Meta-Analysis). By organizing the vast array of personality traits into broad factors, quantitatively integrating findings, and accounting for study artifacts such as sampling and measurement errors, researchers attained stable patterns of relationships undiscovered by early narrative literature reviews. The following section highlights relationships between Big Five personality traits and important workplace attitudes and behaviors uncovered by meta-analytic reviews conducted in the time since the publication of the Encyclopedia's previous edition.

The Big Five Personality Traits in the Workplace

In terms of all Big Five traits, perhaps none is as relevant to workplace criteria as conscientiousness. Reflecting an individual's industriousness and orderliness, this trait is the strongest Big Five predictor of overall job performance, and the importance of conscientiousness generalizes across cultures, occupations, and performance criteria (Barrick et al., 2001). Conscientious employees are more likely to engage in positive citizenship behaviors by going above and beyond their job roles to help out their employer and fellow coworkers (Chiaburu et al., 2011). They are also less likely to engage in deviant behaviors at work, such as theft and property damage (Berry et al., 2007); be involved in workplace accidents (Clarke and Robertson, 2008); and voluntarily leave their jobs (Zimmerman, 2008). Conscientiousness has been shown to be an important trait in the context of leadership, though it more strongly predicts the likelihood that an individual emerges into a leadership position than the likelihood that an individual is effective in that role (Judge et al., 2002a).

Emotional stability – or the absence of neuroticism – is generally regarded as the second most important personality trait predictor of workplace outcomes. Like conscientiousness, this trait predicts overall job performance across occupations, though to a slightly lower degree (Barrick et al., 2001). Emotional stability is also negatively related to workplace deviance, accidents, and turnover. Furthermore, emotionally stable individuals are more likely to emerge into leadership positions and tend to be more effective in these roles (Judge et al., 2002a). The importance of the remaining Big Five traits tends to be much more dependent on the type of job and/or workplace criterion. For instance, extroversion is a valuable predictor of success in managerial jobs, agreeableness in jobs that require interpersonal interactions, and openness in contexts requiring training and learning as opposed to routine task accomplishment (Barrick et al., 2001; Mount et al., 1998).

Accounting for the above-mentioned relationships with workplace behaviors, personality impacts several critical work attitudes and processes. Conscientiousness, emotional stability, and extroversion all positively predict overall job satisfaction (Judge et al., 2002b), and individuals with these traits also tend to be more motivated in the workplace (Judge and Ilies, 2002), as well as more resilient in the face of stress (Alarcon et al., 2009). Likewise, team performance is affected by member personalities, though the optimal configurations depend on the particular trait (Bell, 2007). For agreeableness, team performance is heavily influenced by the team's least agreeable member, as one 'bad apple' can spoil the bunch. Concerning extroversion, it may be desirable to have a heterogeneous mix of this trait so that introverted members can balance out the extroverts and vice versa (Neuman et al., 1999). However, for conscientiousness, too much heterogeneity can be a suboptimal combination (Barrick et al., 1998).

In sum, recent meta-analytic investigations have uncovered meaningful relationships between personality and a vast array of important workplace attitudes and behaviors. However, whether and how organizational decision makers can best capture and utilize personality information has been the subject of intense scrutiny and several critical questions remain. Below, we explore existing debates and future research streams.

Current Concerns and Future Research

A poll conducted by the Society for Human Resource Management in 2011 revealed that less than 20% of organizations use personality tests for hiring or promotion decisions (SHRM, 2011). Despite considerable evidence in support of statistically significant relationships between personality traits and workplace outcomes, some researchers and practitioners question whether the magnitudes of effects are practically meaningful and strong enough to justify their use in personnel selection contexts. Moreover, some scholars argue that the effect sizes of personality traits are just as disappointingly low as they were in Guion and Gottier's review, and that even with methodological improvements, individual personality traits only account for a very small portion of overall job performance.

In response, advocates of personality testing point out that meta-analytic investigations may offer conservative predictive validity estimates since they often aggregate empirical research without any theoretical reasoning to suspect that a given personality trait would be predictive in a certain context. When only specific and theoretically justified relationships between personality and performance are meta-analyzed, personality traits tend to have much greater explanatory power (Hogan and Holland, 2003). In addition, organizations rarely assess just one personality trait in testing contexts, so the composite validities of multiple traits may be a more appropriate metric. Research shows that simultaneous analysis of all Big Five traits greatly improves the prediction of job-related outcomes, making personality about as predictive as other important job-selection constructs, such as cognitive ability and job experience (Barrick and Mount, 2005). In addition, the Big Five traits generally do not differ systematically across racial and ethnic groups, which can help promote hiring practices that foster a diverse workforce (Barrick and Mount, 2005).

A second obstacle to the widespread acceptance of personality testing in the workplace is the issue of responder faking. In high-stakes contexts, where hiring or promotion decisions depend on the outcome of personality tests, applicants may be especially motivated to distort their responses in order to appear more attractive to employers. Organizations have several options to combat faking. First, they might attempt to reduce applicants' desire to fake by requiring applicants to elaborate on their responses (e.g., by providing specific examples of situations where they exhibited a trait) or simply by warning applicants not to fake. Second, organizations might reduce applicants' ability to fake. Forced-choice inventories require applicants to select from equally desirable options (e.g., "do you describe yourself as ambitious or helpful?"), making it difficult for test-takers to ascertain the 'correct' response. In addition, conditional reasoning tests require applicants to justify their actions in certain situations; personality trait scores are inferred based on the applicants' underlying reasoning processes. Finally, techniques exist to identify individuals who fake. One such technique relies on 'lie scales' – items included in the assessments that are designed to measure the likelihood that respondents are purposefully distorting responses on the actual personality tests. More advanced techniques involve measuring individual item's response times or tracking test-taker's eye movements, and are currently being developed with preliminary success (van Hooft and Born, 2012).

Some researchers contend that the issue of faking is much ado about nothing, as the opportunity to fake does not necessarily make it impossible to extract meaningful information from personality test scores. For starters, they argue that even though faking is possible, the actual prevalence is lower than presumed (Hogan et al., 2007). Moreover, the occurrence of faking alone is not necessarily harmful to personality test validity. For instance, if all applicants fake (to a similar degree), then the rank-ordering of applicants by personality test scores will remain consistent even though response distortion occurs. Alternatively, if different people fake to a different extent, but the factors that impact the magnitude and direction of faking are themselves organizationally desirable, then response distortion may not necessarily invalidate assessment results. For instance, motivated individuals may be more likely to engage in impression management and intelligent individuals may be better at distorting responses in the organizationally desired direction. In regard to these issues, empirical evidence shows that the predictive validity of the personality tests remains substantial despite the threat of faking (Ones et al., 2007).

The current state of research has experienced a shift in focus from examining if personality is worthwhile to examining how to maximize its predictive power. One stream of research is revisiting the notion that the Big Five represents the optimal operationalization of personality traits. Recent meta-analytic evidence suggests that assessing personality at a level of specificity narrower than the Big Five might increase criterion-related validity (Judge et al., 2013). In addition, several broader, compound traits – integrity, proactive personality, core self-evaluations – increase predictive validity in certain contexts (Ones et al., 2005). Then there are other traits, for example, Machiavellianism and Narcissism, that do not quite fit within the construct domain of the Big Five and yet deserve increased empirical attention.

A second approach to increasing the effectiveness of personality testing involves contextualizing such assessments. Respondents are asked to rate not their general or global personality but their personality specifically in the workplace. The evidence to date suggests that providing respondents with a frame of reference increases both the reliability and predictive validity of personality assessments (Lievens et al., 2008).

Finally, recognizing the limitations of self-report assessments of personality (such as faking or a lack of self-awareness), researchers continue to explore alternative means of collecting applicant personality information. Some fruitful avenues of research entail acquiring personality ratings from coworkers, spouses, and independent raters. Moreover, independent raters might evaluate personality from information collected via observation, interviews, application materials, or even publicly available online social profiles. Oh et al. (2011) meta-analyzed the predictive validity of observer ratings of personality and found that they were between 30 and 340% higher than self-ratings, depending on the traits being rated.

Summarizing the literature, Judge et al. (2008: p. 1983) remarked: "From the vantage point of today, that personality has shown itself relevant to individual attitudes and behaviors as well as team and organizational functioning seems an incontrovertible statement." Despite the progress made since publication of this Encyclopedia's previous issue, additional work is needed to understand the dynamic processes through which personality shapes workplace attitudes and behaviors and vice versa. In addition, future research will continue to refine and innovate assessment techniques to overcome current limitations and maximize their predictive validity.

Values

Values are "conceptions of the desirable that guide the way social actors (e.g., organisational leaders, policy-makers, individual persons) select actions, evaluate people and events, and explain their actions and evaluations" (Schwartz, 1999: p. 24–25). Like personality, values shape how people perceive and respond to the external environment. However, two features distinguish values from personality: (1) personality reflects behavioral reputations, whereas values describe preferences and intentions; and (2) personality traits vary in terms of how much of a particular characteristic is present/absent in an individual's reputation, whereas values vary in terms of the relative importance of a given preference over other values (Bilsky and Schwartz, 1994). Like early personality trait research, the accumulation of knowledge regarding the importance of work values is hampered by the multitude of values and value typologies being examined. In the following section, we outline two distinct streams of research. The first explores the role of employee–organization value congruence – that is, the match between the values employees consider most important and those that are able to be satisfied by the organization. The second explores the role of national culture on shaping the values of a workforce.

Value Congruence

A well-established maxim in the organizational literature is that employees are most engaged in their work when there is a sense of compatibility between their values and the work environment. Whether conceptualized as congruence with one's job, work unit, supervisor, or organization, value congruence has been meta-analytically linked to outcomes such as job satisfaction, organizational commitment, organizational identification, performance, and turnover (Kristof-Brown et al., 2005). More recently, scholars have begun integrating parallel streams of congruence research. For instance, one paradigm conceptualizes value congruence in terms of complementary fit, where the organization provides what the individual most values. A second paradigm conceptualizes congruence in terms of supplementary fit, where the employee's values are consistent with those of the organization (or existing organizational members). Empirical evidence suggests that both conceptualizations of congruence contribute independently – and with similar magnitudes – to work outcomes (Cable and Edwards, 2004).

A second recent advancement within this literature is a deeper examination of the impact of misfit. Whereas traditional theoretical and empirical approaches treated all types of misfit as equally suboptimal, recent developments suggest more complex models. Relying on polynomial regression and surface plot analysis methodologies, several scholars have uncovered asymmetrical misfit effects. An oversupply of particular values – getting more than one might desire – is less detrimental than an undersupply and, in many cases, leads to more positive outcomes than exact correspondence between values (Kristof-Brown and Guay, 2011).

Cultural Values

Whereas personality research is predominately concerned with individual differences, there is an extensive and growing body of literature on the impact of national culture on the values of employee populations. Broadly speaking, cultural values shape how people prefer to manage and be managed, as well as their attitudes toward work. Several common frameworks exist for understanding cultural values. Schwartz's (1999) theory of cultural values suggests that all cultures are confronted with three social issues and the institutions developed to resolve these issues impact member values. The first issue is the need to define the nature of the relationship between the individual and the group. How cultures resolve this affects values concerning conservatism, intellectual autonomy, and affective autonomy. The second need is to ensure responsible behavior that will maintain social order. Resolution of this issue influences member values concerning hierarchy and egalitarianism. Finally, cultures must address the relationship between humankind and the natural and social world. The institutions designed to address this need are reflected in mastery and harmony values.

Perhaps the most well-known and widely used framework for studying national culture was developed by Geert Hofstede and published in 1980. Hofstede analyzed the values of thousands of IBM employees from dozens of countries and came up with a framework based on four dimensions: individualism–collectivism, power distance, uncertainty avoidance, and masculinity–femininity. Individualism–collectivism reflects the extent to which people prefer to act on their own or as part of a group. In individualistic societies, organizations can expect higher levels of competition and self-interested actions, whereas harmony and loyalty may be the norms in collectivistic societies.

Power distance refers to what extent a society accepts the unequal distribution of power. In low power-distance cultures, managers socialize with subordinates and often involve them in decision-making processes. In high power-distance cultures, superior–subordinate relations are very formal; subordinates are expected to obey their superiors without questioning. Uncertainty avoidance reflects the extent to which people feel threatened by situations that are ambiguous or have unknown outcomes and subsequently act to avoid or control such circumstances. In cultures with high uncertainty avoidance, there is a strong preference for formal rules and guidelines; deviance from established behaviors is regarded negatively. In cultures with low uncertainty avoidance, employees are more apt to take risks and be more tolerant of behaviors and opinions that are different from theirs. Hofstede's masculinity–femininity dimension refers to certain sets of values preferred in a society. Masculine societies favor assertiveness, performance, success, money, and materialism. Feminine societies emphasize a friendly environment, job security, cooperation, quality of life, warm personal relationships, and caring for the weak. Masculine cultures are more likely to divide work and social tasks along gender lines, differentiating between jobs meant for men and those meant for women.

A third typology for distinguishing the impact of national culture on individual values is the GLOBE (Global Leadership and Organizational Behavior Effectiveness Research) project, which defines nine distinct cultural dimensions theorized to impact how leaders manage and structure their workforce. Building on Hofstede's typology, the GLOBE dimensions include uncertainty avoidance, power distance, and collectivism. However, they distinguish between two forms of collectivism: societal collectivism and in-group (e.g., family and/or organization) collectivism. Hofstede's masculinity/femininity is likewise split into a gender egalitarianism dimension and an assertiveness dimension. Finally, GLOBE adds dimensions for future orientation, performance orientation, and humane orientation, representing the degree to which individuals engage in future-oriented behaviors, reward performance improvement, and value fairness, respectively.

Cultural Values and Work-Related Behaviors

The effects of national culture on work-related behaviors are well documented. A meta-analysis by Taras et al. (2010) found Hofstede's cultural dimensions to predict a variety of organizational outcomes including job performance, absenteeism, turnover, organizational commitment, citizenship behaviors, organizational identification, and team attitudes. While cultural values accounted for relatively low amounts of variance in specific behavioral outcomes, their predictive power concerning emotions, attitudes, and perceptions rivaled – and in some cases surpassed – those of personality and general mental ability.

In their review of empirical studies using Hofstede's framework, Kirkman et al. (2006) summarized findings regarding culture's influence on individual styles of change management, conflict management, human resource management, negotiation, reward allocation, decision making, and leadership. In addition, culture was examined as an antecedent of organizational justice, work-related attitudes, and employee motivation.

Of the four dimensions, not only was individualism–collectivism the most frequently investigated, but comparatively, it tended to be the most powerful predictor of workplace outcomes. Collectivist values are associated with increased cooperation and positive attitudes regarding teams. Additionally, individuals from collectivist cultures tend to prefer nondirective leadership that emphasizes participation and teamwork. On the other hand, individualism is associated with increased conflict, and those from individualist cultures tend to use personal experience rather than formal rules to manage conflict. Individualism is positively associated with avoiding unethical behavior and preferring paternalistic leadership, and negatively associated with concern for others' interests in conflict management (Kirkman et al., 2006; Taras et al., 2010).

Individuals from high power-distance cultures tend to have higher organizational commitment, person–organization fit, and a preference for directive leadership; however, they are less likely to avoid unethical behavior and seek feedback (Kirkman et al., 2006; Taras et al., 2010). The value of uncertainty avoidance is strongly associated with an individual's team commitment and a preference for directive leadership, and negatively associated with innovation and a preference for participative leadership (Taras et al., 2010). Finally, masculinity is positively associated with a preference for a compromising conflict management style and social avoidance, and negatively correlated with a preference for an avoidance conflict management style and the value of individual equality (Taras et al., 2010).

Cultural values can also shape individual personality. In their literature review, Heine and Buchtel (2009) discussed the differences in personality across cultures, especially between Eastern and Western cultures. In Eastern societies, perceptions of the self tend to be more fluid, context-specific, and dependent on others. Easterners are more likely to have contradictory views of themselves, and they may change the way they act based on the situation. Because of the collectivist nature of many Eastern cultures, Easterners tend to act in ways that promote harmony with others. They are less likely than Westerners to engage in self-enhancing activities, perhaps because collectivist cultures may punish such actions more than individualist cultures. Western, individualist cultures, on the other hand, may provide benefits of self-enhancement that outweigh the costs, leading to its greater prevalence as a personality trait in such countries. Additionally, Westerners tend to view their personalities as stable and believe they act somewhat consistently across different situational contexts.

Despite extensive research, there remains much to do to better understand the influence of values at work. While a considerable amount of research has accumulated regarding some cultural values (e.g., individualism), others have been less explored. Moreover, paralleling the personality literature, researchers have begun to call for an exploration of more narrowly defined value facets than the broad dimensions conceptualized in the dominant typologies. In addition, process models accounting for how values develop, change over time, and impact attitudes and behaviors are needed to clarify our existing empirical findings.

See also: Big Five Factor Model, Theory and Structure; Cross-Cultural Psychology; Culture: Contemporary Views; Five Factor Model of Personality, Assessment of; Five Factor Model of Personality, Facets of; Organizations and Culture; Personality Assessment, Faking And; Personality Assessment, Forced-choice; Personality Assessment: Overview; Personality Assessment; Personality, Trait Models of; Personnel Selection, Psychology of; Systematic Reviewing and Meta-analysis; Values, Psychology of; Values, Social Psychology of; Vocational Interests, Values, and Preferences, Psychology of.

Bibliography

Alarcon, G., Eschleman, K.J., Bowling, N.A., 2009. Relationships between personality variables and burnout: a meta-analysis. Work & Stress 23, 244–263.

Barrick, M.R., Mount, M.K., 2005. Yes, personality matters: moving on to more important matters. Human Performance 18, 359–372.

Barrick, M.R., Stewart, G.L., Neubert, M.J., et al., 1998. Relating member ability and personality to work-team processes and team effectiveness. Journal of Applied Psychology 83, 377–391.

Barrick, M.R., Mount, M.K., Judge, T.A., 2001. Personality and performance at the beginning of the new millennium: what do we know and where do we go next? International Journal of Selection and Assessment 9, 9–30.

Bell, S.T., 2007. Deep-level composition variables as predictors of team performance: a meta-analysis. Journal of Applied Psychology 92, 595–615.

Berry, C.M., Ones, D.S., Sackett, P.R., 2007. Interpersonal deviance, organizational deviance, and their common correlates: a review and meta-analysis. Journal of Applied Psychology 92, 410–424.

Bilsky, W., Schwartz, S.H., 1994. Values and personality. European Journal of Personality 8 (3), 163–181.

Cable, D.M., Edwards, J.R., 2004. Complementary and supplementary fit: a theoretical and empirical investigation. Journal of Applied Psychology 89 (5), 822–834.

Chiaburu, D.S., Oh, I.-S., Berry, C.M., et al., 2011. The five-factor model of personality traits and organizational citizenship behaviors: a meta-analysis. Journal of Applied Psychology 96, 1140–1166.

Clarke, S., Robertson, I., 2008. An examination of the role of personality in work accidents using meta-analysis. Applied Psychology 57, 94–108.

Finn, S.E., 1986. Stability of personality self-ratings over 30 years: evidence for an age/cohort interaction. Journal of Personality and Social Psychology 50, 813–818.

Furnham, A., 1992. Personality at Work: The Role of Individual Differences in the Workplace. Routledge, London.

Guion, R.M., Gottier, R.F., 1965. Validity of personality measures in personnel selection. Personnel Psychology 18, 135–164.

Heine, S.J., Buchtel, E.E., 2009. Personality: the universal and the culturally specific. Annual Review of Psychology 60, 369–394.

Hogan, J., Holland, B., 2003. Using theory to evaluate personality and job-performance relations: a socioanalytic perspective. Journal of Applied Psychology 88, 100–112.

Hogan, R., Hogan, J., Roberts, B.W., 1996. Personality measurement and employment decisions: questions and answers. American Psychologist 51, 469–477.

Hogan, J., Barrett, P., Hogan, R., 2007. Personality measurement, faking, and employment selection. Journal of Applied Psychology 92, 1270–1285.

van Hooft, E.A., Born, M.P., 2012. Intentional response distortion on personality tests: using eye-tracking to understand response processes when faking. Journal of Applied Psychology 97 (2), 301–316.

Judge, T.A., Ilies, R., 2002. Relationship of personality to performance motivation: a meta-analytic review. Journal of Applied Psychology 87, 797–807.

Judge, T.A., Higgins, C.A., Thoresen, C.J., et al., 1999. The big five personality traits, general mental ability, and career success across the life span. Personnel Psychology 52, 621–652.

Judge, T.A., Bono, J.E., Ilies, R., et al., 2002a. Personality and leadership: a qualitative and quantitative review. Journal of Applied Psychology 87, 765–780.

Judge, T.A., Heller, D., Mount, M.K., 2002b. Five-factor model of personality and job satisfaction: a meta-analysis. Journal of Applied Psychology 87, 530–541.

Judge, T.A., Klinger, R.L., Simon, L.S., et al., 2008. The contributions of personality to organizational behavior and psychology: findings, criticisms, and future research directions. Social and Personality Psychology Compass 2, 1982–2000.

Judge, T.A., Rodell, J.B., Klinger, R.L., et al., 2013. Hierarchical representations of the five-factor model of personality in predicting job performance: integrating three organizing frameworks with two theoretical perspectives. Journal of Applied Psychology 98 (6), 875–925.

Kirkman, B.L., Lowe, K.B., Gibson, C.B., 2006. A quarter century of culture's consequences: a review of empirical research incorporating Hofstede's cultural values framework. Journal of International Business Studies 37, 285–320.

Kristof-Brown, A.L., Zimmerman, R.D., Johnson, E.C., 2005. Consequences of individuals' fit at work: a meta-analysis of person-job, person-organization, person-group, and person-supervisor fit. Personnel Psychology 58, 281–342.

Kristof-Brown, A., Guay, R.P., 2011. Person–environment fit. In: Zedeck, S. (Ed.), APA Handbook of Industrial and Organizational Psychology. American Psychological Association, Washington, DC, pp. 3–50.

Lievens, F., De Corte, W., Schollaert, E., 2008. A closer look at the frame-of-reference effect in personality scale scores and validity. Journal of Applied Psychology 93 (2), 268–279.

Mount, M.K., Barrick, M.R., Stewart, G.L., 1998. Five-factor model of personality and performance in jobs involving interpersonal interactions. Human Performance 11, 145–165.

Neuman, G.A., Wagner, S.H., Christiansen, N.D., 1999. The relationship between work-team personality composition and the job performance of teams. Group & Organization Management 24, 28–45.

Oh, I.S., Wang, G., Mount, M.K., 2011. Validity of observer ratings of the five-factor model of personality traits: a meta-analysis. Journal of Applied Psychology 96 (4), 762–773.

Ones, D.S., Viswesvaran, C., Dilchert, S., 2005. Personality at work: raising awareness and correcting misconceptions. Human Performance 18, 389–404.

Ones, D.S., Dilchert, S., Viswesvaran, C., Judge, T.A., 2007. In support of personality assessment in organizational settings. Personnel Psychology 60 (4), 995–1027.

Schwartz, S.H., 1999. A theory of cultural values and some implications for work. Applied Psychology 48, 23–47.

Society for Human Resource Management, 2011. Personality Tests for the Hiring and Promotion of Employees SHRM Poll. http://www.shrm.org/research/surveyfindings/articles/pages/shrmpollpersonalitytestsforthehiringandpromotionofemployees.aspx (accessed 20.03.14.).

Taras, V., Kirkman, B.L., Steel, P., 2010. Examining the impact of culture's consequences: a three-decade, multilevel, meta-analytic review of Hofstede's cultural value dimensions. Journal of Applied Psychology 95, 405–439.

Zimmerman, R.D., 2008. Understanding the impact of personality traits on individuals' turnover decisions: a meta-analytic path model. Personnel Psychology 61, 309–348.

Personality Assessment

Johannes Hewig, Uta Kraus, Patrick Mussel, Roman Osinsky, and Marko Paelecke, Julius Maximilians University Würzburg, Würzburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The present article provides an overview on personality assessment in basic research and applied areas of psychology. Both descriptive and explanatory models of personality and the derived methods of assessment are presented. Further personality assessment in the applied fields of psychology is introduced focusing on clinical psychology, and touching on the topic in organizational and educational psychology.

Definitions

The necessary formal aspect of a definition of personality assessment is not easy due to the nonuniformity of the personalist terminology. The terms 'person' and 'personality,' derived from the Latin words *persona* and *personalitas*, are among the most abstract expressions of our language. It is therefore not surprising to find many approaches to define the personality of a person; Allport (1937) had already compiled 50 definitions. Within everyday language, personality denotes the individual characteristics of a person, but can also have a meritorious connotation. Within recent psychology, definitions of personality are neutral and focus on stable individual differences in behavior. From an evolutionary perspective, McAdams and Pals (2006: p. 204) defined personality as "(a) an individual's unique variation on the general evolutionary design for human nature, expressed as a developing pattern of (b) dispositional traits, (c) characteristic adaptations, and (d) self-defining life narratives, complexly and differentially situated (e) in culture and social context." In the present context we focus on personality assessment as the measurement of relatively stable, traitlike interindividual differences in temperament, which become manifest in trait-relevant situations.

Backgrounds

Descriptive Approaches to Personality

Descriptive approaches to personality date back to ancient literature. A well-known example is the characterology of Theophrastus of Eresus, published around 319 BC. The rather humorous than moralizing character sketches contain typical personality traits that seem to have survived more than two millennia. Scientific personality assessment within clinical psychology started in the twentieth century. The methodological groundwork for the study of individual differences was laid by Sterns (1911) research strategies for personality psychology. After World War I, Woodworth (1919) developed the earliest self-report personality questionnaire in order to measure signs of maladjustment to military life in soldiers – the Woodworth Personal Data Sheet (PDS). It contained rationally devised items describing several symptoms of emotional distress. The PDS spurred the development of further uni- and multidimensional questionnaires for the assessment of adjustment problems in the 1920s and 1930s, with items usually constructed in a similar, i.e., rational or deductive way (for a historical overview, see Butcher, 2010). A well-known exception was the criteria-orientated development of the Minnesota Multiphasic Personality Inventory (Hathaway and McKinley, 1943), where item selection was purely based on discrimination ability between a clinical group patients and a nonclinical control group. It is still available in several versions and remains one of the most used personality measures in clinical psychology.

After World War II, exploratory factor analysis became a widely used construction method. With regard to the underlying dimensions of personality, the results of factor analytic personality research appear quite inconsistent at first glance. For instance, Cattell (1965) or Guilford (1975) proposed rather different factorial description systems of individual differences. Nevertheless, a five-factor model of basic dimensions of personality has emerged, which is represented by various authors and groups of authors (for a comprehensive account of the historical development of the five-factor model, cf Digman, 1994). Lexical analysis of personality description terms and joint factor analyses of existing personality measures repeatedly revealed five broad factors, 'neuroticism,' 'extraversion,' 'openness to experiences,' 'agreeableness,' and 'conscientiousness.' Costa and McCrae (1995), major proponents of the model, argue that this relatively small number of dimensions captures the major variance in the personality descriptive terms: "What the model hypothesizes is that almost every personality trait is substantially related to one or more of the five factors, and that any remaining traits … form a miscellaneous category rather than covarying to define a sixth or subsequent factor" (p. 218).

The five-factor model is currently considered the standard in the field of personality theories for global and comprehensive assessment of personality. Several measures exist, the most prominent being the NEO-Personality Inventory (NEO-PI), which is available in a revised as well as a shortened version (Costa and McCrae, 2008). Although primarily a measure of normal personality variation, the NEO questionnaires can also be used in clinical settings (Costa and McCrae, 2008). Similar to other personality inventories, response patterns of clinical and nonclinical subjects in the Revised NEO Personality Inventory (NEO-PI-R) were structurally invariant (O'Connor, 2002b).

In addition to the five-factor model for a global and comprehensive assessment of personality, a variety of constructs and associated measures exist for assessing specific personality traits. Within personality research it is still controversial to what extent variance observed in specific constructs is not already covered by the five factors or if, on the other hand, relevant dimensions are missing in the five-factor model (e.g., O'Connor, 2002a).

Explanatory Personality Models

Several psychobiological personality models emerged in the second half of the twentieth century, aiming to explain mechanisms underlying individual differences in personality. The theories of Eysenck (1967) and Gray (1982) combined basic personality dimensions with differences in arousal, in the sensitivity to cues of reinforcement, and the sensitivity to cues of punishment. In these theories, individual differences in cognition and motivation are based on underlying biological characteristics and refer to basic psychological constructs.

Eysenck's temperament theory (e.g., Eysenck, 1967) includes basic orthogonal dimensions of personality and underlying hypothetical neurobiological models. Basis of the measurement and classification of personality traits were factor analyses with which the theoretically assumed personality structures should be uncovered. In its early conception Eysenck's personality model comprised two basic dimensions, which he described as 'introversion–extraversion' and 'neuroticism.' Later Eysenck and colleagues added a third dimension, 'psychoticism,' linked to behaviors such as lack of empathy, aggression, and impulsivity (see Eysenck et al., 1985). The success of this dimensional classification system lay in the fact that the two dimensions of extraversion and neuroticism are found in virtually all factor analyses of trait judgments that are based on a sufficient number of traits. The differentiating aspect of Eysenck's approach (compared to the aforementioned factorial description systems) is the anchoring of the personality dimensions in neurobiological systems. Eysenck (1967) suspected the biological bases of individual differences in neuroticism and extraversion in interindividual differences in neurophysiological excitation and inhibition processes. To assess the three dimensions, Eysenck and colleagues devised the Eysenck Personality Questionnaire (EPQ) in a revised version (see Eysenck et al., 1985; EPQ-R).

Gray (1982) proposed a modification of Eysenck's theory, based on animal learning theory. In his two-dimensional temperament system, two behavioral systems play a crucial role that come into play in emotional situations. A Behavioral Activation System (BAS) organizes the response to conditioned stimuli that signal positive reinforcement (reward) or negative reinforcement (nonpunishment), and leads to approach or active avoidance behavior. A Behavioral Inhibition System (BIS) organizes the response to stimuli that are unknown or signal nonreward or punishment. This leads to behavioral inhibition and to an increase in limbic arousal and attention. Individual differences of these two systems yield two orthogonal dimensions that were described by Gray as 'anxiety' and 'impulsivity.' In Gray's theory extraversion and neuroticism represent derivations of these more fundamental factors. Anxiety ranges from 'low extrovert and highly neurotic' (equivalent to high anxiety) to 'highly extroverted and low neurotic' (equivalent to low anxiety). High impulsivity corresponds to 'strongly extroverted and highly neurotic,' low impulsivity to 'low extraverted and low neurotic.' With increasing anxiety, susceptibility to nonreward, punishment, and unknown stimuli increases. Increased impulsivity is associated with susceptibility to reward and nonpunishment. In a modification of Eysenck's theory extraverts are no harder to condition than introverts, but rather more sensitive to punishment or the threat thereof, as well as more sensitive to reward and positive affirmation. In 2000, Gray and McNaughton presented the revised Reinforcement Sensitivity Theory (r-RST). Similar to Gray's original theory, they assumed three biological systems with similar names (a BAS, a BIS, and a Fight-Flight-Freeze-System (FFFS)), which are anatomically separated but functionally related.

To assess the two dimensions sensu Gray's original theory, the most widely used measure to date are the BIS/BAS scales developed by Carver and White (1994). Jackson (2009) developed scales to separately tap the BAS, BIS, FFFS of the r-RST. Recently, Corr and McNaughton (2012) provided another refinement of this line of thinking reviewing evidence on approach and avoidance personality. Further psychobiological models of personality partially influenced by Gray were developed by Cloninger (1994) and Zuckerman (1991).

Clinical Psychology

The assessment of individual differences in personality has a long tradition in the fields of clinical psychology and psychiatry. Today it is widely accepted that personality plays a crucial role in several clinical issues, including, for example, etiology of mental disease, pathoplasticity, treatment efficiency, or risk for disorder relapse. For instance, several dispositional traits (e.g., neuroticism) have been shown to contribute to an individual's capability to cope with both daily stressors as well as severe negative life events. Personality is therefore closely linked to the concepts of vulnerability and psychological resilience, adding to a person's risk to develop mental illness. Thus, the consideration of personality traits is not only helpful but often crucial to understand the pathogenesis of certain psychopathological disorders. Based on personality measures, at-risk individuals may be identified and introduced to early prevention treatment. Generally, a comprehensive knowledge about a patient's personality can help clinicians in choosing appropriate therapeutic interventions, since some psychotherapeutic techniques and/or pharmacological treatments may work out well in some persons but not in others. Accordingly, personality traits often have a high prognostic value with regard to therapeutic effects or even disorder relapse. Finally, personality measures are frequently used as outcome measures, providing important information about the achievement of therapeutic goals. As personality is not absolutely stable but may substantially change during and/or after a pathological episode, one should generally distinguish between premorbid, morbid, and postmorbid personality.

In most of the available studies on the relation between personality and psychopathology self-report questionnaires

have been used for personality assessment. This practice is probably fostered by the low temporal and economic expenses of such questionnaire measures. However, this approach also entails several potential problems as the ability to correctly assess the own personality can be strongly affected by the current state (e.g., negative mood, limited self-insight) and, therefore, by the disorder itself. Moreover, self-report measures of personality and psychopathology often share a considerable number of similar items, resulting in high common method variance. This may lead to an overestimation of the personality–psychopathology relationship. For some individuals (e.g., preschool children) questionnaires are even totally unsuitable. To overcome some of the problems of self-report inventories, future research may therefore more frequently use other types of personality measures, such as interviews, informants' reports, or observational measures.

As mentioned above the relation between personality and psychopathology is manifold and any attempt to provide a complete overview here could only fail. Below, we therefore exemplarily describe some aspects of the personality–psychopathology interrelation, also considering the issue of personality assessment.

Affective-Motivational Personality Constructs and Psychopathology

Over the last century several dimensional models have been proposed, in which personality has been mainly described in terms of affective-motivational dispositions. Most traits described in these models can be subsumed under or are related to the broad concepts of negative emotionality/neuroticism (NE/N) and positive emotionality/extraversion (PE/E) (e.g., Markon et al., 2005). In adults and adolescents, traits in the spectrum of NE/N and PE/E are typically assessed using self-report questionnaires like the EPQ (Eysenck et al., 1985), the NEO-PI-R (see Costa and McCrae, 2008), or the Temperament and Character Inventory (Cloninger, 1994). Child temperamental traits related to NE/N (e.g., behavioral inhibition) have been assessed in many ways using, for example, retrospective self-reports or behavioral observations (e.g., Caspi et al., 2003). For obvious reasons, a vast number of studies have investigated whether and how NE/N is related to mood disorders. With regard to unipolar depression the role of NE/N seems to be manifold, with NE/N and depression sharing common etiological influences, NE/N predicting the onset and course of depressive episodes, and NE/N interacting with negative life events in increasing the risk for depression (for an overview see Klein et al., 2011). NE/N is also increased in individuals with a bipolar disorder. However, unipolar versus bipolar mood disorders also strongly differ with regard to their relation to low-order personality facets (e.g., Quilty et al., 2013). Using multifacet inventories like the NEO-PI-R may therefore provide important differential diagnostic information. Besides its' role in the context of mood disorders, NE/N is linked to other forms of mental disease (especially, anxiety disorders) as well as general health problems, pointing to a rather unspecific relation between NE/N and (psycho-) pathology. This, however, also indicates that NE/N contributes to comorbidity among many mental disorders (for a general overview of the role of NE/N in psychopathology, see Lahey, 2009). With regard to PE/E and its role in psychopathology the existing literature is less consistent. For instance, while some studies report low PE/E in patients with unipolar depression others did not report such an association. Facets of PE/E that are closely related to appetitive motivation (e.g., novelty seeking and sensation seeking) have been repeatedly found to be associated with a higher risk for substance abuse, or problematic gambling (e.g., Zuckerman and Kuhlman, 2000).

Personality Disorders

When it comes to the topic of personality and psychopathology the probably most obvious link, by definition, lies in the field of personality disorders. There is vast number of available tools for the assessment of maladaptive personality functioning, including symptom checklists (e.g., the Psychopathy Checklist Revised Hare, 2003), self-report inventories (e.g., the Millon et al., 2009), or diagnostic interviews (e.g., the Structured Clinical Interview for DSM-IV Axis II Disorders First et al., 1997). In the classification systems for mental diseases (International Classification of Diseases (ICD) and Diagnostic and Statistical Manual of Mental Disorders (DSM)), personality disorders have traditionally been conceptualized as distinct categories, i.e., a personality disorder is diagnosed when a predefined threshold is exceeded. This practice has received much criticism, and calls have been made for a more dimensional conceptualization and multitool-based assessment of personality disorders (e.g., Clark, 2007). However, the categorical model of personality disorders was kept in the official classification system (Section II) of the most recent 5th edition of the DSM, while a hybrid categorical-dimensional model has been introduced only in DSM-V Section III under the label of 'Emerging measures and models.' Although this categorical-dimensional model was also criticized because of its low practicality in daily clinical work and lack of empirical support (e.g., Verheul, 2012) it might be, at least, a starting point for a more dimensional conceptualization of personality disorders in future editions of DSM and/or ICD.

Dispositions of Self-Regulation

The term self-regulation is a very broad one, referring to an individual's capabilities and strategies to regulate one's emotions, behaviors, and cognitions in order to achieve certain goals (e.g., completing a certain task, avoiding distress, or maximizing pleasure). It comprises several other concepts such as emotion regulation, effortful-control, coping, or self-efficacy. In the field of clinical psychology, habitual tendencies in self-regulation have been investigated, for example, with regard to their role in stress vulnerability/reactivity and therapeutic outcomes. For instance, the personality trait of conscientiousness (i.e., the individual's tendency to set and keep long-range goals and inhibiting impulses that may interfere with these goals) is negatively associated with depression, thereby also interacting with NE/N. High conscientiousness combined with low NE/N is also linked to increased treatment efficiency (Klein et al., 2011). NE/N, PE/E, and conscientiousness are also related to coping styles, with PE/E and conscientiousness being positively correlated with engagement coping (i.e., dealing with the stressor or the resulting negative emotions) and NE/N being positively correlated rather with disengagement coping

(i.e., avoiding to deal with stressor or resulting negative emotions; Connor-Smith and Flachsbart, 2007). Conscientiousness (like NE/N and PE/E) is typically assessed with self-report questionnaires like the NEO-PI-R. In children, effortful-control (i.e., the ability to inhibit a dominant response) is a frequently investigated temperament that is closely related to the concept of executive control/attention. It has been shown that low effortful-control in childhood predicts divers forms of psychopathology and maladjustments later on (Nigg, 2006). Individual differences in children in effortful-control have often been assessed using behavioral observations in attentional tasks (e.g., Rothbart et al., 2003). Finally, perceived self-efficacy is related to diverse aspects of health behavior (O'Leary, 1985). Thus, stable tendencies in the attribution and perceived control of behavioral outcomes influences the individual's stress reactivity and health status (e.g., Roddenberry and Renk, 2010). Several self-report scales are available to assess such tendencies in perceived self-efficacy and locus of control (see, e.g., Furnham and Steele, 1993).

Organizational Psychology

Assessing personality in organizational settings is an important application of personality psychology, and has been a driving force in the development of personality theory and assessment. In organizational practice, personality assessments are mostly utilized for personnel selection; additionally, other important areas include the assessment of personality for personnel development, placement, and career counseling. Hundreds of thousands of assessments are conducted each year around the world, which provide valuable information for both test takers and organizations, for example, by providing valuable information to take well-informed career decisions or to staff individuals who will perform well and contribute to the success of the team and the organization. Despite these applied purposes, investigating personality in organizational settings also advances our knowledge regarding the personality constructs themselves, as convergent and discriminant validities with theoretically associated criteria give important information regarding the construct domain of a personality trait.

Broad Dimensions of Personality

The significance of personality traits for predicting job performance has been investigated so often that, in the meanwhile, second-order *meta*-analyses have been conducted (Barrick et al., 2001). Regarding the Big Five, the factor most strongly associated with job performance is conscientiousness. Corrected for unreliability in the predictor and the criterion, range restriction, and lack of construct validity an average coefficient of 0.27 has been obtained. Across studies, the validity of conscientiousness has been shown to generalize across different criteria (supervisory ratings, objective performance indicators, teamwork) and across occupations (sales performance, managerial performance, professionals, police, skilled, or semiskilled). A reason for the consistent, even though only moderate correlation might be that individuals with high, compared to low levels on conscientiousness are more likely to be on time, to keep their workplace tidy, to perform their tasks in a structured manner, to deliver orders within the prescribed time limit, to maintain a high standard of quality, to engage in the task, and to reassure the quality of their work. Interestingly, the relatively high correlation of conscientiousness with job performance is mirrored by correlations with other criteria of intellectual performance, such as school and academic performance (Poropat, 2009).

Additionally to conscientiousness, generalized validity for job performance has also been found for emotional stability, even though the coefficient of 0.13 is rather small. Contrary, openness to experience, agreeableness, and extraversion are only marginally correlated with general job performance. At most, correlations can be found for specific criteria or occupations. For example, extraversion and openness for experience have been found to predict training performance, and extraversion has been found to predict performance in managerial positions and, to a smaller degree, for police officers.

In light of the relatively small correlation coefficients, the relevance of personality traits for predicting job performance has been criticized (see Tett and Christiansen, 2007, for a discussion). For its application in high stakes situations like personnel selection, an additional problem is social desirability, i.e., the tendency to euphemize self-reports of personality to maximize the probability for a job offer.

Narrow Facets of Personality

While the investigation according to the Big Five model of personality has largely helped to compare results across studies, there are hints that the Big Five might be too broad for predicting job performance (Tett and Christiansen, 2007). For example, Mussel et al. (2011) found the facet openness to ideas to significantly predict job performance, but found negligible correlations for the remaining facets as well as the Big Five factor openness to experience. Judge et al. (2013) compared the predictive validity of facets, compared to Big Five traits *meta*-analytically. They found higher validities for facets, compared to the Big Five, which explained approximately twice the variance. Differential validity for facets was especially pronounced for openness to experience, emotional stability, and extraversion.

Additional evidence for the significance of personality for job performance provides so-called Criterion Orientated Personality Scales (COPS, Ones and Viswesvaran, 2001). COPS are personality traits, which were specifically developed for predicting job performance. For example, significant prediction of job performance around 0.40 has been found for integrity tests, which were initially developed for predicting deviant behavior such as stealing, or mobbing. Similarly, personality scales assessing customer service orientation have been found to significantly predict job performance.

A final view regarding the role of personality in organizational settings takes a look from the dark side. Therefore, the question is not whether certain personality traits predispose the likelihood of succeeding at work, but rather at predictors of failure. Compared to the studies discussed so far, this approach has not so much been guided by the Big Five personality traits. Recently, the so-called dark triad of personality, consisting of narcissism, machiavellianism, and psychopathy, has received increasing attention (e.g., Furnham

et al., 2013). In the work context, high levels on the dark triad has been found to be associated with counterproductive work behavior, the tendency to manipulate individuals in pursuit of selfish gains ('snakes in suits'), and criminal activities, such as shoplifting. However, individuals with high levels on the dark triad also have a strong desire for power and a strong motivation to be very rich, which should be advantageously related to career success, especially when combined with other factors, like intelligence or physical attractiveness (Furnham et al., 2013). Disentangling these opposing effects for diverse organizational outcome variables will be an important endeavor for the future.

Educational Psychology

Personality assessment in educational contexts comprises all temperamental and personality traits, habits, and states that are relevant to academic learning (Braden, 2003). Temperament and personality measures can be applied to single persons during psychological school service or in connection with educational research or achievement tests in school studies, cross-national, or large-scale school studies. The main focus of personality assessment lies on the child or young adult as the individual learner. Personality characteristics are relevant to educational contexts because they can serve as vulnerabilities and protective factors in achievement situations and influence the development of learning disorders (e.g., Hasselhorn, 2008). Learning-related questions and problems comprise a wide range of aspects from learning disabilities and learning disorders to high achievers and giftedness. In the last decade the focus of personality assessment in educational contexts was extended to personality characteristics of peers and teachers that can influence successful learning in various ways. Personality assessment in educational contexts deals with the problem that personality traits develop during childhood and adolescence (e.g., attribution stile, volitional control abilities) and that a broader and more differentiated construct of personality does not come up until late childhood. The choice of personality assessment in educational contexts thus depends on the cognitive development of the learner and can vary between younger and older children.

In educational contexts broad personality inventories are used less often because various broad personality traits are not strongly related to learning outcomes. However, consciousness and emotional stability seem to be important for learning to some degree and they appear to go along with higher academic performance (e.g., Poropat, 2009). It is interesting to note that the latter two traits are also highly relevant for both clinical and organizational assessment (see above). Yet, in light of the weak relations for broad personality traits (Big Five, etc.), research focused on more narrow, single personality characteristics that are central for learning. These specific personality characteristics are assessed frequently in educational contexts and are relevant to most of the school service problems and questions. The selection of possible personality characteristics always depends on the concern of the school service and on the age or cognitive ability level of the child (e.g., learning disorder, learning disability, achievement drop, giftedness, aggression, anxiety). The most important specific learning-related personality characteristics are motivation and volition (see section below and cf Hasselhorn, 2008).

Domains and Methods of Assessment

Temperament characteristics, in particular more narrow aspects, are of particular interest for educational contexts because they can be applied early in life, and they are relevant for various learning-related behavioral outcomes (e.g., inhibition). The most commonly used temperament dimensions are activity level, anger/frustration, behavioral inhibition/fear, effort control, and positive affect (Zentner and Bates, 2008). These temperament dimensions do not represent a single theory rather they are a consensus of researchers about the most important dimensions (Zentner and Bates, 2008). Temperament dimensions can be measured by self-report interviews and questionnaires, other-report questionnaires, behavioral observation, and psychobiological measures. Depending on the concern, a multimethod approach should be preferred, to validate the single measures and to cope with the weakness of each method. In educational contexts temperament characteristics are measured mainly by self-report (child interview, questionnaire) and other-report instruments (parent questionnaire, teacher questionnaire) that often contain behavioral observations (child in school and/or family context; for an overview see Zentner and Bates, 2008).

Motivation and Volition

Motivation and the achievement motive system play a central role for learning. It is the first relevant precondition for successful learning (Hasselhorn, 2008), it should be part of personality assessment in an educational context, and its measurement has been debated intensively in the past (e.g., Hulleman et al., 2010). The quality of the individual achievement motive system comprises the three aspects (1) success orientation or fear of failure, (2) attribution style, and (3) self-concept of competences (Hasselhorn, 2008). The development of achievement motive is observable in the attributions children employ to explain their success or failure and becomes relatively stable in early adolescence (Hasselhorn, 2008). Overoptimism, unrealistic wishful thinking, and a lack of social comparisons are defense mechanisms of the achievement motive system that play a major role in the development of competency beliefs and are characteristic for young children (Hasselhorn, 2008). The excessively optimistic self-concept of children in their first two school years descents and a more realistic school-related self-concept is developed. In this process, using the individual reference norm instead of social reference or criteria norms supports a positive development of learning and the achievement motive system. In contrast, if protective mechanisms of self-esteem are not available, external attributions to the own success may develop and can become a risk factor of a failing individual achievement motive. Children with ongoing experiences of failure are at high risk to develop a weak belief in their own competences and may lose the ability to use cognitive resources and metacognitive abilities available to them. All of the risk mechanisms in turn increase the probability to develop psychological symptoms like anxiety,

depression, aggression, psychosomatic symptoms, or learning disorders (Vahia et al., 2011).

Volition is the second critical precondition for successful learning (Hasselhorn, 2008). In addition to motivation, volitional control of behavior is necessary to realize learning intentions and learning goals. For example, Shoda et al. (1990) showed that reward gratification increases with increasing age and the presence of this capacity at preschool age goes along with more self-confidence, higher frustration tolerance, and achievements in school. Besides these findings, delay gratification remains difficult for some adolescents and adults indicating that a temperamental component might be critical for initiating control. Nevertheless in demanding contexts volitional control of behavior necessitates strategies of metacognitive regulation that develop relatively late (cf Hasselhorn, 2008).

See also: Academic Achievement Motivation, Development of; Big Five Factor Model, Theory and Structure; Dark Triad: The "Dark Side" of Human Personality; Emotional Regulation; Five Factor Model of Personality, Assessment of; Minnesota Multiphasic Personality Inventory; Neuroticism; Personality Assessment, Faking And; Personality Assessment, Forced-choice; Personality Assessment: Overview; Personality Development: Systems Theories; Personality Disorders; Personality and Values at Work; Personality, Biological Models of; Personality, Trait Models of; School Achievement: Motivational Determinants and Processes; Temperament and Human Development.

Bibliography

Allport, G.W., 1937. Personality. Henry Holt, New York.

Barrick, M.R., Mount, M.K., Judge, T.A., 2001. Personality and performance at the beginning of the new millennium: what do we know and where do we go next? International Journal of Selection and Assessment 9 (1–2), 9–30.

Braden, J.B., 2003. Psychological assessment in school settings. In: Weiner I.B., (Editor-in-chief), Graham J.R., Naglieri J.A. (Eds.), Handbook of Psychology, vol. 10. Assessment Psychology, Wiley & Sons, Hoboken, NJ, pp. 261–290.

Butcher, J.N., 2010. Personality assessment from the nineteenth to the early twenty-first century: past achievements and contemporary challenges. Annual Review of Clinical Psychology 6, 1–20.

Carver, C.S., White, T.L., 1994. Behavioral-inhibition, behavioral activation, and affective responses to impending reward and punishment – the BIS/BAS scales. Journal of Personality and Social Psychology 67, 319–333.

Caspi, A., Harrington, H., Milne, B., Amell, J.W., Theodore, R.F., Moffitt, T.E., 2003. Children's behavioral styles at age 3 are linked to their adult personality traits at age 26. Journal of Personality 71, 495–513.

Cattell, R.B., 1965. The Scientific Analysis of Personality. Penguin, London.

Clark, L.A., 2007. Assessment and diagnosis of personality disorder: perennial issues and an emerging reconceptualization. Annual Review of Psychology 58, 227–257.

Cloninger, C.R., 1994. The Temperament and Character Inventory (TCI): A Guide to Its Development and Use. Center for Psychobiology of Personality, Washington University, St. Louis, MO.

Connor-Smith, J.K., Flachsbart, C., 2007. Relations between personality and coping: a meta-analysis. Journal of Personality and Social Psychology 93, 1080–1107.

Corr, P.J., McNaughton, A.J., 2012. Neuroscience and approach/avoidance personality traits: a two stage (valuation-motivation) approach. Neuroscience and Biobehavioral Reviews 36 (10), 2339–2354.

Costa, P.T., McCrae, R.R., 1995. Solid grounds in the wetlands of personality: a reply to block. Psychological Bulletin 117, 216–220.

Costa, P.T., McCrae, R.R., 2008. The NEO inventories. In: Archer, R.P., Smith, S.R. (Eds.), Personality Assessment. Taylor & Francis, New York, pp. 213–245.

Digman, J.M., 1994. Historical antecedents of the five factor model. In: Costa, P.T., Widiger, T.A. (Eds.), Personality Disorders and the Five Factor Model of Personality. APA, Washington.

Eysenck, H.J., 1967. The Biological Basis of Personality. Thomas, Springfield.

Eysenck, S.B.G., Eysenck, H.J., Barrett, P., 1985. A revised version of the psychoticism scale. Personality and Individual Differences 6, 21–29.

First, M.B., Gibbon, M., Spitzer, R.L., Williams, J.B.W., Benjamin, L.S., 1997. Structured Clinical Interview for DSM-IV Axis II Personality Disorders, (SCID-II). American Psychiatric Press, Inc., Washington, DC.

Furnham, A., Steele, H., 1993. Measuring locus of control: a critique of general, children's, health- and work-related locus of control questionnaires. British Journal of Psychology 84, 443–479.

Furnham, A., Richards, S.C., Paulhus, D.L., 2013. The dark triad of personality: a 10 year review. Social and Personality Psychology Compass 7 (3), 199–216.

Gray, J.A., McNaughton, N., 2000. The Neuropsychology of Anxiety. Oxford University Press, Oxford.

Gray, J.A., 1982. The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-hippocampal System. Oxford University Press, Oxford.

Guilford, J.P., 1975. Factors and factors of personality. Psychological Bulletin 82, 802–814.

Hare, R.D., 2003. Manual for the Revised Psychopathy Checklist, second ed. Multi-Health Systems, Toronto, ON, Canada.

Hasselhorn, M., 2008. Competencies for successful learning: developmental changes and constraints. In: Hartig, J., Klieme, E., Leutner, D. (Eds.), Assessment of Competencies in Educational Contexts. Hogrefe & Huber Publishers, Cambridge, MA, pp. 23–43.

Hathaway, S.R., McKinley, J.C., 1943. The Minnesota Multiphasic Personality Inventory. University of Minnesota Press, Minneapolis, MN.

Hulleman, C.S., Schrager, S.M., Bodmann, S.M., Harackiewicz, J.M., 2010. A meta-analytic review of achievement goal measures: different labels for the same constructs or different constructs with similar labels? Psychological Bulletin 136 (3), 422.

Jackson, C.J., 2009. Jackson-5 scales of revised reinforcement sensitivity theory (r-RST) and their application to dysfunctional real world outcomes. Journal of Research in Personality 43, 556–569.

Judge, T.A., Rodell, J.B., Klinger, R.L., Simon, L.S., Crawford, E.R., 2013. Hierarchical representations of the five-factor model of personality in predicting job performance: integrating three organizing frameworks with two theoretical perspectives. Journal of Applied Psychology 98 (6), 875–925.

Klein, D.N., Kotov, R., Bufferd, S.J., 2011. Personality and depression: exploratory models and review of evidence. Annual Review of Clinical Psychology 7, 269–295.

Lahey, B.B., 2009. Public health significance of neuroticism. American Psychologist 64, 241–256.

Markon, K.E., Krueger, R.F., Watson, D., 2005. Delineating the structure of normal and abnormal personality: an integrative hierarchical approach. Journal of Personality and Social Psychology 88 (1), 139–157.

McAdams, D.P., Pals, J.L., 2006. A new big five: fundamental principles for an integrative science of personality. American Psychologist 61, 204–217.

Millon, T., Millon, C., Davis, R., Grossman, S., 2009. MCMI-III Manual, fourth ed. Pearson Education, Inc, Minneapolis, MN.

Mussel, P., Winter, C., Gelléri, P., Schuler, H., 2011. Explicating the openness to experience construct and its subdimensions and facets in a work setting. International Journal of Selection And Assessment 19 (2), 145–156.

Nigg, J.T., 2006. Temperament and developmental psychopathology. Journal of Child Psychology and Psychiatry 47, 395–422.

O'Connor, B.P., 2002a. A quantitative review of the comprehensiveness of the five-factor model in relation to popular personality inventories. Assessment 9, 188–203.

O'Connor, B.P., 2002b. The search for dimensional structure differences between normality and abnormality: a statistical review of published data on personality and psychopathology. Journal of Personality and Social Psychology 83, 962–982.

O'Leary, A., 1985. Self-efficacy and health. Behaviour Research and Therapy 23 (4), 437–451.

Ones, D.S., Viswesvaran, C., 2001. Integrity tests and other criterion-focused occupational personality scales (COPS) used in personnel selection. International Journal of Selection and Assessment 9, 31–39.

Poropat, A.E., 2009. A meta-analysis of the five-factor model of personality and academic performance. Psychological Bulletin 135 (2), 322–338.

Quilty, L.C., Pelletier, M., DeYoung, C.G., Bagby, R.M., 2013. Hierarchical personality traits and the distinction between unipolar and bipolar disorders. Journal of Affective Disorders 147, 247–254.

Roddenberry, A., Renk, K., 2010. Locus of control and self-efficacy: potential mediators of stress, illness, and utilization of health services in college students. Child Psychiatry and Human Development 41 (4), 353–370.

Rothbart, M.K., Ellis, L.K., Rueda, M.R., Posner, M.I., 2003. Developing mechanisms of temperamental effortful control. Journal of Personality 71 (6), 1113–1144.

Shoda, Y., Mischel, W., Peake, P.K., 1990. Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification. Developmental Psychology 26, 978–986.

Stern, W., 1911. Die Differentielle Psychologie in Ihren Methodischen Grundlagen [Methodological Foundations of Differential Psychology]. Barth, Leipzig.

Tett, R.P., Christiansen, N.D., 2007. Personality tests at the crossroads: a response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt. Personnel Psychology 60, 967–993.

Vahia, I.V., Chattillion, E., Kavirajan, H., Depp, C.A., 2011. Psychological protective factors across the lifespan: implications for psychiatry. Psychiatric Clinics of North America 34 (1), 231–248.

Verheul, R., 2012. Personality disorder proposal for DSM-5: a heroic and innovative but nevertheless fundamentally flawed attempt to improve DSM-IV. Clinical Psychology and Psychotherapy 9, 369–371.

Woodworth, R.S., 1919. Examination of emotional fitness for war. Psychological Bulletin 15, 59–60.

Zentner, M., Bates, J.E., 2008. Child temperament: an integrative review of concepts, research programs, and measures. European Journal of Developmental Science 2, 7–37.

Zuckerman, M., Kuhlman, D.M., 2000. Personality and risk-taking: common biosocial factors. Journal of Personality 68 (6), 999–1029.

Zuckerman, M., 1991. Psychobiology of Personality. Cambridge University Press, Cambridge.

Personality Assessment, Faking and

Matthias Ziegler and Doreen Bensch, Institut für Psychologie, Humboldt-Universität zu Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The present article summarizes key findings from faking research. The article starts off by comparing different definitions for the phenomenon. Afterward faking behavior is placed in a nomological net of social desirable responding. The article continues by presenting empirical findings for the main research questions, i.e., the existence of faking, its impact on a test score's validity, and ways of measuring or preventing faking. Throughout the presentation of these issues, a broad perspective is taken including a variety of even contrasting findings. The article concludes with a summary and an outlook.

What do academic success, job performance, getting a divorce, and life expectancy have in common? The answer to this general question is that each one of these outcomes is predicted by interindividual differences in personality traits, especially the Big 5 and its facets (Roberts et al., 2007). These and other meta-analytical findings are the reason why personality differences are considered important in all kinds of selection settings. Moreover, there are also longitudinal studies showing that personality traits are predictive above and beyond general mental ability for academic success (Ziegler et al., 2009, 2010a), vocational training (Ziegler et al., 2014), job success (Lindqvist and Vestman, 2011), and lifelong learning (Ziegler et al., 2012).

Despite the overwhelming empirical evidence supporting the usefulness of personality traits for predicting important life outcomes, there is a long-lasting debate based on criticism regarding the assessment method of choice: self-reports. Cattell (1958) differentiated three media for personality assessment: life record observations or life data (L-data), questionnaire data (Q-data), and (objective) test data (T-data). In spite of the early description of these three ways, especially Q-data are used today to assess interindividual personality trait difference by administering self-report questionnaires. Regardless of the above-mentioned success of such questionnaires in terms of test criterion correlations, two major criticisms have been raised again and again.

McClelland (1987: 11) stated "A scientist cannot believe what the people say about their motives." This quote most of all was due to the idea that motives should be differentiated in self-attributed motives (can be measured with self-reports) and implicit motives (cannot be measured with self-reports because they are not accessible with introspection). While McClelland meant to scourge the use of questionnaires to assess motivation, his famous quote has been used to depreciate self-reports in general. Much could be said about the more or less fruitful attempts to capture personality traits with so-called implicit measures. In addition to the theoretical differentiation between implicit and explicit traits (Hofmann et al., 2005; McClelland et al., 1989), the varied operationalizations for implicit traits ranging from picture story tests to implicit association tests have all met considerable measurement problems (Ziegler et al., 2010b). Thus, this criticism against self-reports continues to be weakened by the lack of empirical evidence strongly supporting alternative measurement approaches.

The second criticism brought forward against self-reports is that the answers people give can easily be distorted. Especially, response behavior that systematically distorts self-reports from reality toward a more favorable impression, so-called *social desirable responding* (SDR), is regarded as problematic. Paulhus (2002) wrote a seminal chapter in which he summarized the history of research on SDR. His summary finishes with an integrative model of SDR. Within this model, Paulhus describes two general themes (an egoistic and a moralistic bias) that are crossed with two different addressees (self and other) resulting in four kinds of SDR: self-deceptive enhancement, agency management, self-deceptive denial, and communion management. The use of personality self-reports for selection purposes is especially endangered if the test-taker aims his or her distortion at someone else (the potential employer) to appear in a certain way. This kind of behavior is often called faking. Thus, faking should not be confused with SDR. Rather than that faking can be seen as a subform within the SDR framework by Paulhus (2002). Nevertheless, faking is potentially a serious threat for the utility of self-reports, especially in high-stakes situations (i.e., situations in which the test-taker can gain something based on the self-report results).

Definition of Faking

In terms of a nomological net faking might therefore best be positioned within the SDR framework provided by Paulhus (2002) as agentic or communion management. However, this broad categorization has to be complimented with a sound definition of the actual phenomenon in order to avoid confusion. MacCann et al. (2011: 309) summarized definitions from experts in the field of faking research and concluded that all experts agreed upon four key elements: "(〈〈〈) faking is a behavior rather than a trait, (嶋) faking requires motivation (i.e., it is goal-oriented), (嶇) faking behavior results in an inaccurate or enhanced impression, and (帙) faking involves the interaction of person (traits) and situational factors (environmental influences)." Most experts also saw faking as an intentional and conscious act. Based on these key characteristics of faking, the following chapter adopts a definition provided by Ziegler et al. (2011: 8) who define

faking as "a response set aimed at providing a portrayal of the self that helps a person to achieve personal goals. Faking occurs when this response set is activated by situational demands and person characteristics to produce systematic differences in test scores that are not due to the attribute of interest."

Faking as a Response Set

Jackson and Messick (1958) distinguished between response styles (those phenomena that occur across time and different questionnaires) and response sets. Paulhus (2002) also uses these terms and sees response sets as short-lived response biases. More importantly, such response sets are caused by temporary distractions or motivation. In other words, faking behavior does not occur always. Instead, test-takers fake due to some situational characteristics.

Faking as a Means to Achieve Personal Goals

One important aspect of the definition provided above is that faking does not occur without a specific goal. This means test-takers have a specific personality profile in mind when faking, believing that portraying this profile will help them to gain something out of the situation they are in. Such a profile could be the profile of the ideal employee for the offered position (Bowen et al., 2002) or a very specific personality profile for a specific job (Pauls and Crost, 2005). In both cases, matching the profile will increase the chances of getting the job in a selection scenario.

This goal directedness bears important implications for faking research. Whenever effects of faking are to be researched, samples need to be used that have a highly similar goal in mind. In that sense, combining applicants for different jobs into one sample potentially blurs faking effects (Birkeland et al., 2006). Using think aloud protocols, Ziegler (2011) could show that test-takers match their answer in a self-report to their knowledge and implicit theories about the job they are applying for. Thus, if a sample contains test-takers who have different job profiles in mind, the faking results will be different and potentially even reversed.

Faking as an Interaction between Person and Situation

As mentioned above, faking behavior occurs when situational demands call for it. However, not everyone fakes and those who fake do not all end up with the same test score. Thus, individual differences in faking behavior exist. Ellingson (2011) states capability to fake, believing in the necessity to fake, and the belief that the opportunity to fake is valued as important interindividual differences determining if and how much a person fakes.

Finally, this definition implies that faking occurs most often in so-called high-stakes situations. Especially, selection procedures such as job selections constitute the natural setting for researching faking and its effects. Therefore, within the rest of this article, faking behavior will always mean that test-takers distort their responses to a self-report measure of personality in order to achieve a personal goal in a specific (high stakes) situation. Whereas this specification helps to mark the playground for faking research, there is the question whether faking behavior actually exists.

Does Faking Really Exist?

There is a lot of evidence that faking behavior can be elicited in the lab (e.g., Ziegler et al., 2007). However, lab studies have been criticized for their lacking ecological validity. Moreover, results from empirical studies using real applicant data seem to suggest that faking does not exist outside of the laboratory. Hogan et al. (2007) published results from an impressive data set of applicants. The findings show only little differences (Cohen's $d < |.15|$) in the Big 5 domains between applicants and a research sample. While the size of the data set itself is impressive, the results are difficult to interpret without knowing the specific jobs the applicants applied for. As mentioned above, in order to prevent different job expectations from blurring faking effects, job homogeneous samples are necessary. Whether the applicants in the paper presented by Hogan et al. all applied for the same job remains unclear, though. The idea that job homogenous samples are needed is supported by meta-analytical findings from Birkeland et al. (2006). Those authors showed that job characteristics (sales vs nonsales and management vs nonmanagement) moderate extend and direction of faking behavior.

Griffith et al. (2007) introduced a method based on the ideas of true scores and test score reliability from classical test theory. Within this framework it is assumed that each person possesses a so-called true score, i.e., the true standing of that person on a latent variable. However, assessing this true score is virtually impossible because nonsystematic errors distort single measurements. Multiple measurements of the same trait for one person would yield a normal distribution with the true score as a mean and the error determining the variance. Therefore, when administering a questionnaire, a random sample out of this intraindividual distribution is drawn. In order to estimate the true score, a confidence interval around the measured score can now be estimated based on the test score's reliability and a set error probability (usually 5%). Griffith et al. have applied this general approach to individual assessments in order to estimate the amount of people faking in an applicant setting. They used a data set of applicants who had filled out a personality questionnaire twice, once in an applicant setting and the second time in an anonymous research condition. Griffith et al. estimated a confidence interval around the score from the anonymous condition. Depending on the kind of interval used, 30 to 50% of the applicants had scores under applicant conditions that were outside of their confidence interval. Thus, faking behavior occurs in reality and can be verified using methods based on ideas from classical test theory.

Does Faking Matter?

Whether faking matters or not is harder to answer. First of all, it has to be noted that most research conducted in order to answer this question used some kind of real or simulated job applicant condition. There is research within the clinical area,

looking at faking (i.e., malingering) within a clinical context (Hall and Hall, 2011) which is not discussed here.

Usually, when the impact of faking behavior is discussed, three different outcomes are looked at construct validity, test criterion correlation, and selection outcome. Whereas the first two outcomes generally use group statistics such as correlations, the latter one looks at the actual selection results and therefore, might be the most important in terms of practical implications.

Faking and Construct Validity of a Test Score

Important aspects of a test score's construct validity are the correlations with other test scores close by in the nomological net. Moreover, the factorial structure of the questionnaire used oftentimes is a key issue. Using structural equation modeling early work in this area suggested that faking behavior did not affect construct validity (Ellingson et al., 2001; Smith and Ellingson, 2002). However, two issues raise questions with regard to this interpretation. First of all, in order to identify a subsample of fakers, high scores in so-called social desirability scales were used. As will be outlined below, by now it seems clear that such scales are no reliable indicators of faking behavior. Secondly, in order to fit the models, the authors had to include a number of correlated residuals. Based on the definition given above, such correlated errors might be exactly the manifestation of faking behavior (Schmidt et al., 2003).

Other research suggests that correlations between test scores increase when the test-takers fake with a similar faking goal (Pauls and Crost, 2005). Moreover, Ziegler and Bühner (2009) also used structural equation modeling in combination with an experimental design and could show that test scores that were faked have stronger intercorrelations. More importantly though, a multiple group structural equation model was used to show that faking can be modeled as a common method variable. This approach decreased the intercorrelations between faked scores almost to zero. The large mean differences between a faking and an honest condition also disappeared. Thus, the approach allowed to remedy the effects of faking behavior on the mean score as well as construct validity of a test score and causally attribute them to a faking instruction. It has to be noted though that these findings can be criticized due to the fact that all this research was conducted in a lab.

Thus, there is strong empirical evidence that suggests effects of faking behavior on a test score's construct validity. However, this needs to be tested outside of the laboratory.

Faking and Test Criterion Correlations

As before, there is mixed evidence regarding the effect of faking on test criterion correlations. An early meta-analysis suggested that faking does not harm or influence test criterion correlations (Ones and Viswesvaran, 1998). However, faking in this meta-analysis was operationalized using social desirability scales which are problematic indicators at best (see below). The study by Ziegler and Bühner (2009) also showed no negative effects of faking behavior on test criterion correlations. But, as was criticized before, these findings were derived from lab data. Again using lab data, Ziegler et al. (2010a) could show that faking does distort test criterion correlations. The new perspective in this study was an aim on personality facets and not domains as was the case before.

Thus, as was the case with construct validity, further research using real applicant data from job homogeneous samples is needed.

Faking and Selection Outcomes

No one seems to argue against the fact that if faking occurs, it should change a test score's group mean into the desirable direction. Considering the evidence provided above that faking actually occurs in real job applicant settings, such mean changes seem inevitable. However, if all test-takers faked to the same amount, this mean-level change would not affect a selection procedure. Considering the definition of faking presented above as well as the moderating variables suggested by Ellingson (2011) and others, it seems likely that people differ in their amount of faking. This in turn would greatly affect selection outcomes. In fact, a study by Converse et al. (2009) revealed the effects of faking on top-down selection processes using a Conscientiousness scale only or multiple predictors. Depending on the selection ratio and other criteria, the percentage of applicants getting selected because they faked in the Conscientiousness measure ranged between 5% and 14% in a multiple predictor selection process. Using only the Conscientiousness scale could yield up to almost 30% of applicants getting selected because they had faked.

Thus, this research shows practical consequences. Whenever faking seems likely to occur in a selection process, multiple predictors (e.g., ability tests or interviews in addition to self-reports) should be used. It seems reasonable to apply a self-report as a first step to select out before other selection tools are applied.

Can We Measure and Correct for Faking?

If it were possible to measure faking, it would also be possible to correct potential detrimental effects. Therefore, numerous different ways have been tried to measure faking behavior. The most important seem to be social desirability scales, over-claiming, and item response theory (IRT)-based approaches.

Social Desirability Scales

Using embedded items within a questionnaire or whole scales within a test battery in order to assess faking is anything but new. The general idea is that such items ask for behavior that is common but highly undesirable (e.g., littering) or behavior that is highly desirable but rather uncommon (e.g., always telling the truth). The most important problem with this operationalization is that it is possible that test-takers actually possess the trait level that is considered highly unlikely. They would mistakenly be flagged as fakers. Moreover, test-takers often do not feel that 'never' really means 'never.' In that sense, some might think of themselves as someone who only lies in order not to hurt someone. This, however, can be seen

as positive. Therefore, the test-taker might think that such behavior is not meant, giving the statement never to lie a high agreement rating. Again, this person would erroneously be flagged as a faker. Tellegen et al. (2006) wrote "The most important and sophisticated program of inventory development to have been predicated on the idea of a general and separately measurable social desirability dimension ended up abandoning that construct" (p. 150). Paulhus (2011) also came to a generally negative conclusion: "Unfortunately, some researchers continue to indict the validity of self-report instruments if they show high correlations with SDR measures" (p. 152). Thus, all in all, it has to be said that social desirability scales are not useful to assess faking (see Griffith and Peterson, 2008 for a more detailed review).

Overclaiming

Overclaiming is a technique that was first used in marketing and brought to psychologists' attention through the work of Paulhus (Paulhus, 2011; Paulhus et al., 2003). The general idea of this technique is to have test-takers rate their familiarity with specific terms (e.g., historical events, artists, vocabulary). Some of these terms, the so-called foils, do not exist, though. Thus, test-taker claiming to be familiar with these terms exaggerate the extend of their knowledge. Using formulas from signal detection theory it is then possible to estimate a bias score as well as an accuracy score. The former is used as an indicator of response distortion. However, research so far has placed overclaiming into the category of self-enhancement. Applying the definition from above, such a behavior might not qualify as faking behavior if it occurred unintentionally.

Bing et al. (2011) used an overclaiming measure to capture faking in an applicant setting. Moreover, those authors could show that overclaiming behavior actually suppressed unwanted error variance in personality test scores. Ziegler et al. (2013) could show that a short measure of overclaiming consisting only of 15 items might be used to identify SDR in an interview.

Thus, overclaiming, despite appearing to be conceptually too far away from faking behavior, has shown promise as a means of actually capturing faking variance. In how far this might be useful on an individual level, i.e., in selection contexts, remains an open question, though.

IRT-Based Approaches

Zickar and colleagues applied so-called mixed Rasch models (MRMs) to identify fakers (Zickar et al., 2004; Zickar and Robie, 1999). MRMs are a mixture of Rasch models and latent class analysis. This allows to identify qualitatively distinct response patterns, representing faking sets while maintaining the assumption of a continuous trait measured at the same time (Rost et al., 1997). Results derived from applying this method to applicant data repeatedly have suggested two distinct faking response sets: slight and extreme faking. Recently, this interpretation has been challenged due to the overlap of these response sets with middle and extreme point responding, response sets found in anonymous situations (Ziegler and Kemper, 2013). Moreover, Zickar and Sliter (2012) summarized that IRT models so far have shown only limited success in identifying individual fakers.

Other Techniques

There are a number of other approaches to measuring faking, e.g., word use (Ventura, 2011), decision trees (Lukoff, 2011), the Bayesian truth serum (MacCann et al., 2011), or anchoring vignettes (Buckley, 2009), to name just a few. All of these approaches have some merits but also substantial problems.

Therefore, a summary at this point must come to the conclusion that there is still much to learn about faking behavior before a successful measurement method can be designed. Consequently, corrections for faking are equally elusive.

Can We Prevent Faking?

There are two general approaches trying to prevent faking. The first approach tries to create a fake-proof questionnaire. The second approach tries to reason with the test-taker.

Fake Resistant Measures

Bäckström et al. (2009) have suggested the use of neutralized items. Such items are phrased in a way that makes them neither socially desirable nor undesirable. Their empirical evidence shows that a general evaluative factor underlying all personality test scores can dramatically be reduced using neutralized items. However, if we apply the definition for faking provided above, the neutralized items seem to tackle the issue of self-enhancement or denial. Research showing that such items could be used to reduce faking is missing. Moreover, the authors themselves did not construct the items in order to reduce faking but refer to social desirability in general.

A technique clearly designed to reduce faking is the pairwise preference model (Stark et al., 2011). Questionnaires using this approach present the test-takers with item quadruples. Each of those usually contains four items and test-takers have to choose which item they agree with the most and which item the least. The items are matched on their desirability making it impossible to achieve an overall positive impression simply by using high ratings for all items as is the case in a typical Likert scale questionnaire. Such forced choice items have been criticized because of their inability to capture interindividual differences. However, recent statistical developments allow to overcome these problems (Brown and Maydeu-Olivares, 2011; Stark et al., 2011). Thus, this approach currently seems to be the most promising. In how far the coaching of test-takers might affect this technique remains to be seen, however.

Reasoning with the Test-Taker

Incorrect selection decisions based on personality self-ratings bear not only important implications for the person selecting but also the test-taker. Imagine an applicant who desperately wants a job and therefore, portrays a picture that is in large

agreement with the job profile but does not resemble the person's true personality. If the person gets selected, a behavior is expected that is atypical for the person. Thus, performing in the job will be either impossible or at least attached to high costs for the person. Arguments like these are often used to reason with test-takers not to fake. While first empirical evidence has shown that this might actually reduce faking on group level, the individual effects are mixed. Whereas some individuals do not fake at all, others fake just as much as without warning. Moreover, some test-takers are afraid of being accused of faking and actually lower their scores. Thus, on average the test score mean does not increase. The rank order, however, is greatly affected (Griffith et al., 2005). Thus, warnings seem to be no proper way to handle faking.

Conclusion

Within this article we first defined faking and put it in a nomological net of SDR. According to this, faking is an intentional and goal-directed behavior occurring due to an interaction between person and situation. Despite the mixed findings, recent research convincingly shows that faking does exist in real applicant situations. Its effects on a test score's validity are controversial and more research using job homogeneous applicant samples is needed. Nevertheless, the results so far at least suggest an effect of faking on ranking which in turn negatively affects selection decisions. A prereject selection procedure including multiple selection tools might offer a solution here. However, this approach might only work in situations where a large applicant pool can be drawn from. Measuring faking is also rather problematic so far. Overclaiming might be a promising approach considering existing research. However, correcting for faking on an individual level seems to be a long jump considering the current knowledge. More positive is the conclusion regarding the prevention of faking. Using pairwise preference models potentially offers a way to assess personality with self-reports free of faking influences.

On the whole the current overview provides an idea of a broad research field. The main ideas as well as conclusions are outlined and it should be clear that faking research still offers many potential breakthroughs and should continue to attract a large number of researchers.

See also: Dark Triad: The "Dark Side" of Human Personality; Personality Assessment, Forced-Choice; Personality Assessment: Overview; Personality Assessment; Personnel Selection, Psychology of; Situational Judgment Test.

Bibliography

Bäckström, M., Björklund, F., Larsson, M.R., 2009. Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. Journal of Research in Personality 43, 335–344.

Baldwin, S.A., Fellingham, G.W., 2012. Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. Psychological Methods. http://dx.doi.org/10.1037/a0030642. No Pagination Specified.

Bing, M.N., Kluemper, D., Kristl Davison, H., Taylor, S., Novicevic, M., 2011. Overclaiming as a measure of faking. Organizational Behavior and Human Decision Processes 116, 148–162.

Birkeland, S.A., Manson, T.M., Kisamore, J.L., Brannick, M.T., Smith, M.A., 2006. A meta-analytic investigation of job applicant faking on personality measures. International Journal of Selection and Assessment 14, 317–335.

Bowen, C.C., Martin, B.A., Hunt, S.T., 2002. A comparison of ipsative and normative approaches for ability to control faking in personality questionnaires. International Journal of Organizational Analysis 10, 240–259.

Brown, A., Maydeu-Olivares, A., 2011. Item response modeling of forced-choice questionnaires. Educational and Psychological Measurement 71, 460–502. http://dx.doi.org/10.1177/0013164410375112.

Buckley, J., 2009. Cross-national Response Styles in International Educational Assessments: Evidence from PISA 2006. Retrieved from: https://edsurveys.rti.org/PISA/.

Cattell, R.B., 1958. What is "objective" in "objective personality tests?". Journal of Counseling Psychology 5, 285–289.

Converse, P., Peterson, M., Griffith, R., 2009. Faking on personality measures: implications for selection involving multiple predictors. International Journal of Selection and Assessment 17, 47–60.

Ellingson, J.E., 2011. People fake only when they need to fake. In: Ziegler, M., MacCann, C., Roberts, R. (Eds.), New Perspectives on Faking in Personality Assessment. Oxford University Press, New York, pp. 19–33.

Ellingson, J.E., Smith, D.B., Sackett, P.R., 2001. Investigating the influence of social desirability on personality factor structure. Journal of Applied Psychology 86, 122–133.

Griffith, R.L., Chmielowski, T., Yoshita, Y., 2007. Do applicants fake? An examination of the frequency of applicant faking behavior. Personnel Review 36, 341–357.

Griffith, R.L., Peterson, M.H., 2008. The failure of social desirability measures to capture applicant faking behavior. Industrial and Organizational Psychology 1, 308–311.

Griffith, R.L., Yoshita, Y., Gujar, A., Malm, T., Socin, R., 2005. The Impact of Warnings on Applicant Faking Behavior, submitted for publication.

Hall, R.C.W., Hall, R.C.W., 2011. Plaintiff s who malinger: impact of litigation on fake testimony. In: Ziegler, M., MacCann, C., Roberts, R. (Eds.), New Perspectives on Faking in Personality Assessment. Oxford University Press, New York, pp. 255–281.

Hofmann, W., Gschwendner, T., Nosek, B.A., Schmitt, M., 2005. What moderates implicit-explicit consistency? European Review of Social Psychology 16, 335–390.

Hogan, J., Barrett, P., Hogan, R., 2007. Personality measurement, faking, and employment selection. Journal of Applied Psychology 92, 1270–1285.

Jackson, D.N., Messick, S., 1958. Content and style in personality-assessment. Psychological Bulletin 55, 243–252.

Lindqvist, E., Vestman, R., 2011. The labor market returns to cognitive and noncognitive ability: evidence from the Swedish enlistment. American Economic Journal: Applied Economics 3, 101–128. http://dx.doi.org/10.1257/app.3.1.101.

Lukoff, B., 2011. Is faking inevitable? Person-level strategies for reducing faking. In: Ziegler, M., Maccann, C., Roberts, R. (Eds.), New perspectives on Faking in Personality Assessment, vol. 1. Oxford University press, New York, pp. 240–253.

MacCann, C., Ziegler, M., Roberts, R., 2011. Faking in personality assessment: reflections and recommendations. In: Ziegler, M., MacCann, C., Roberts, R. (Eds.), New Perspectives on Faking in Personality Assessment. Oxford University Press, New York, pp. 309–329.

McClelland, D.C., 1987. Human Motivation. Cambridge University Press, New York.

McClelland, D.C., Koestner, R., Weinberger, J., 1989. How do self-attributed and implicit motives differ? Psychological Review 96, 690–702.

Ones, D.S., Viswesvaran, C., 1998. The effects of social desirability and faking on personality and integrity assessment for personnel selection. Human Performance 11, 245–269.

Paulhus, D.L., 2002. Socially desirable responding: the evolution of a construct. In: Braun, H.I., Jackson, D.N., Wiley, D.E. (Eds.), The Role of Constructs in Psychological and Educational Measurement. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 49–69.

Paulhus, D.L., 2011. Overclaiming on personality questionnaires. In: Ziegler, M., MacCann, C., Roberts, R.D. (Eds.), New Perspectives on Faking in Personality Assessment. Oxford University Press, New York, pp. 151–164.

Paulhus, D.L., Harms, P.D., Bruce, M.N., Lysy, D.C., 2003. The over-claiming technique: measuring self-enhancement independent of ability. Journal of Personality and Social Psychology 84, 890–904.

Pauls, C.A., Crost, N.W., 2005. Effects of different instructional sets on the construct validity of the NEO-PI-R. Personality and Individual Differences 39, 297–308.

Roberts, B.W., Kuncel, N.R., Shiner, R., Caspi, A., Goldberg, L.R., 2007. The power of personality: the comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. Perspectives on Psychological Science 2, 313–345. http://dx.doi.org/10.1111/j.1745-6916.2007.00047.x.

Rost, J., Carstensen, C.H., Von Davier, M., 1997. Applying the mixed Rasch model to personality questionnaires. In: Rost, J., Langeheine, R.E. (Eds.), Applications of Latent Trait and Latent Class Models in the Social Sciences. Waxmann, New York.

Schmidt, F.L., Le, H., Ilies, R., 2003. Beyond alpha: an empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. Psychological Methods 8, 206–224.

Smith, D.B., Ellingson, J.E., 2002. Substance versus style: a new look at social desirability in motivating contexts. Journal of Applied Psychology 87, 211–219.

Stark, S., Chernyshenko, O., Drasgow, F., 2011. Constructing fake-resistant personality tests using item response theory: high-stakes personality testing with multidimensional pairwise preferences. In: Ziegler, M., MacCann, C., Roberts, R. (Eds.), New Perspectives on Faking in Personality Assessment. Oxford University Press, New York, pp. 214–239.

Tellegen, A., Ben-Porath, Y.S., Sellbom, M., Arbisi, P.A., McNulty, J.L., Graham, J.R., 2006. Further evidence on the validity of the MMPI–2 Restructured Clinical (RC) scales: addressing questions raised by Rogers, Sewell, Harrison, and Jordan and Nichols. Journal of Personality Assessment 87, 148–171.

Ventura, M., 2011. The detection of faking through word use. In: Ziegler, M., MacCann, C., Roberts, R. (Eds.), New Perspectives on Faking in Personality Assessment. Oxford University Press, New York, pp. 165–176.

Zickar, M.J., Gibby, R.E., Robie, C., 2004. Uncovering faking samples in applicant, incumbent, and experimental data sets: an application of mixed-model item response theory. Organizational Research Methods 7, 168–190.

Zickar, M.J., Robie, C., 1999. Modeling faking good on personality items: an item-level analysis. Journal of Applied Psychology 84, 551–563.

Zickar, M.J., Sliter, K.A., 2012. Searching for unicorns: item response theory-based solutions to the faking problem. In: Ziegler, M., MacCann, C., Roberts, R. (Eds.), New Perspectives on Faking in Personality Assessment. Oxford University Press, New York, pp. 113–130.

Ziegler, M., 2011. Applicant faking: a look into the black box. The Industrial and Organizational Psychologist 49, 29–36.

Ziegler, M., Bensch, D., Maaß, U., Schult, V., Vogel, M., Bühner, M., 2014. Big Five facets as predictor of job training performance: the role of specific job demands. Learning and Individual Differences 29, 17.

Ziegler, M., Bühner, M., 2009. Modeling socially desirable responding and its effects. Educational and Psychological Measurement 69, 548–565.

Ziegler, M., Danay, E., Heene, M., Asendorpf, J., Bühner, M., 2012. Openness, fluid intelligence, and crystallized intelligence: toward an integrative model. Journal of Research in Personality 46, 173–183. http://dx.doi.org/10.1016/j.jrp.2012.01.002.

Ziegler, M., Danay, E., Schölmerich, F., Bühner, M., 2010a. Predicting academic success with the big 5 rated from different points of view: self-rated, other rated and faked. European Journal of Personality 24, 341–355. http://dx.doi.org/10.1002/per.753.

Ziegler, M., Schmukle, S., Egloff, B., Bühner, M., 2010b. Investigating measures of achievement motivation(s). Journal of Individual Differences 31, 15–21. http://dx.doi.org/10.1027/1614-0001/a000002.

Ziegler, M., Kemper, C., 2013. Extreme response style and faking: two sides of the same coin? In: Winker, P., Porst, R., Menold, N. (Eds.), Interviewers' Deviations in Surveys: Impact, Reasons, Detection and Prevention (Schriften Zur Empirischen Wirtschaftsforschung). Peter Lang Gmbh, Frankfurt a.M., pp. 221–237.

Ziegler, M., Knogler, M., Bühner, M., 2009. Conscientiousness, achievement striving, and intelligence as performance predictors in a sample of German psychology students: always a linear relationship? Learning and Individual Differences 19, 288–292.

Ziegler, M., MacCann, C., Roberts, R.R., 2011. Faking: knowns, unknowns, and points of contention. In: Ziegler, M., MacCann, C., Roberts, R.R. (Eds.), New Perspectives on Faking in Personality Assessment. Oxford University Press, New York, pp. 3–16.

Ziegler, M., Schmidt-Atzert, L., Bühner, M., Krumm, S., 2007. Fakability of different measurement methods for achievement motivation: questionnaire, semi-projective, and objective. Psychology Science 49, 291–307.

Personality Assessment, Forced-Choice

Anna Brown, School of Psychology, University of Kent, Canterbury, Kent, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Instead of responding to questionnaire items one at a time, respondents may be forced to make a choice between two or more items measuring the same or different traits. The forced-choice format eliminates uniform response biases, although the research on its effectiveness in reducing the effects of impression management is inconclusive. Until recently, forced-choice questionnaires were scaled in relation to person means (ipsative data), providing information for intraindividual assessments only. Item response modeling enabled proper scaling of forced-choice data, so that interindividual comparisons may be made. New forced-choice applications in personality assessment and directions for future research are discussed.

Personality assessment is largely reliant on respondent–reported information. Objectively scored personality tests are perhaps the most widely used assessment mode. Such tests typically consist of multiple items, each intended to measure one personality trait. The most popular way of gathering responses on test items is by presenting them one at a time, as single stimuli ('single-stimulus' response format). Alternatively, two or more items at a time may be presented, forcing respondents to make a choice between them ('forced-choice' response format). Although appealing in many applications due to their ability to elicit finer differentiation between stimuli, forced-choice questionnaires have been controversial due to substantial scaling problems they pose. While it is easy to infer relative standings on different personality traits within one person, it has proven difficult to infer absolute trait standings for interpersonal comparisons from forced-choice data.

In this article, I consider first the different types of forced-choice formats, followed by a short discussion of their potential advantages in reducing response biases. I then turn to the measurement models that have been used to scale forced-choice responses, starting with the classical scoring model leading to ipsative scores, and followed by new model-based approaches based on item response theory (IRT). Models that have been applied to real-world forced-choice personality assessments are briefly described. Next, I evaluate the merits of different forced-choice questionnaire designs in providing accurate assessments of personality traits. Finally, I discuss the direction for future research with forced-choice measures in cross-cultural, educational, and industrial–organizational psychology.

Methods for Collecting Forced-Choice Questionnaire Data

Forced-choice data may be collected in many ways. Forced-choice questionnaires typically consist of multiple 'blocks,' presenting two or more items at the same time, and respondents are asked to indicate their preferences for the items within each block. Items within blocks may measure the same traits ('unidimensional' forced choice or UFC) or different traits ('multidimensional' forced choice or MFC). Table 1 summarizes the main types of forced-choice designs; each briefly described in this section.

Table 1 Methods for collecting forced-choice data

Block size	*Binary preference*	*Graded preference*	*Proportional preference*
$n = 2$	Paired comparison	Graded paired comparison	Proportional paired comparison
$n \geq 3$	Ranking; Partial ranking; Ranking with ties (Q-sort)		'Proportion-of-total-amount' task

Binary Preference

The simplest forced-choice task involves a pair of items, or block of size $n = 2$. Respondents are asked to indicate their preference for one item, for instance, to indicate which of the two statements is most true of them.

	Most true of me
A. I pay attention to details	X
B. I change my mood a lot	

When a block consists of three or more items ($n \geq 3$), respondents may be asked to rank statements according to the extent they are true of respondents, or to indicate the top rank only, or the top and bottom ranks, for example.

	Rank order	*Most/least true*	*Most true*
A. I pay attention to details	2		
B. I change my mood a lot	4	Least	
C. I have a good word for everyone	3		
D. I catch on to things quickly	1	Most	Most

Examples of assessments using forced-choice pairs are the Tailored Adaptive Personality Assessment System (TAPAS), the Navy Computerized Adaptive Personality Scales (NCAPS), the Edwards Personal Preferences Schedule (EPPS), and the Myers-Briggs Type Indicator (MBTI). Examples of tests using rankings are the Occupational

 http://dx.doi.org/10.1016/B978-0-08-097086-8.25084-8

Personality Questionnaire (OPQ32i and OPQ32r), the Personality and Preference Inventory (PAPI), the Customer Contact Styles Questionnaire (CCSQ 7.2), the Gordon Personal Profile Inventory (GPP-I), the Survey of Interpersonal Values (SIV), the DiSC Classic, and the Kolb Learning Style Inventory.

In ranking with ties, also known as Q-sort (Block, 1961), respondents have to assign items to categories, complying with a predefined distribution. For example, respondents are asked to sort 10 items into five piles, according to the extent to which the items describe respondents' personality, and only allowing a certain number of items in every pile:

Very untrue of me	*Somewhat untrue of me*	*Neutral*	*Somewhat true of me*	*Very true of me*
1	2	4	2	1
Number of items in pile.				

California Adult Q-Sort, Riverside Behavioral Q-sort, and Riverside Situational Q-sort are all examples of this format. Q-sorts are not rating tasks but pure forced choice, because assignments to categories are constrained so that respondents must perform direct comparisons between items. Even if no items are 'very true' of the respondent, he/she must assign the prescribed number of items to that category. The obtained responses are essentially rankings with ties – a group of items ranked first, a group of items ranked second, etc.

Assessments using paired comparisons and ranking tasks yield binary preference data, since the only information collected for any given pair of items is whether one item was preferred to the other or not. For full rankings, binary choices are known for every pairwise comparison; that is, in the example above, it is known that item A was preferred to items B and C, but not to item D. For partial rankings, such as those resulting from 'most/least' choices, some outcomes of pairwise comparisons are not known; for instance, it is not known whether item A was preferred to C or not. For rankings with ties, preferences for items within same category are not known.

Graded Preference

It is also possible to gather information on the extent of preference for one item over another. For example, we may ask respondents to indicate to what extent one statement is more (or less) true of them than the other using ordered categories.

	Much more true	*Slightly more true*	*Slightly more true*	*Much more true*	
A. I pay attention to details				X	B. I change my mood a lot

With this format graded (ordinal) preference information is available for every pairwise comparison.

Proportional Preference

To gather further quantitative information about the relative merits of items, respondents may be asked to distribute a fixed number of points between them. For instance, we may ask respondents to distribute 100 points among $n = 4$ statements according to the extent the statements describe them.

	Points (100 in total)
A. I pay attention to details	40
B. I change my mood a lot	0
C. I have a good word for everyone	10
D. I catch on to things quickly	50

This format captures yet more information about the extent of preference than graded comparisons. It is not only known, for example, that A was preferred to B more than it was preferred to C, but how much more. Proportions of the total amount are collected and therefore it may be inferred, for instance, that preference for A (40 points) was 4 times stronger than preference for C (10 points).

Advantages of the Forced-Choice Formats

Comparative judgments employed in forced-choice questionnaires can have substantial advantages over absolute judgments. First, forced choice makes it impossible to endorse all items indiscriminately (the so-called 'acquiescence' bias). It is also impossible to elevate or reduce ratings across all items ('leniency/severity' effects), or provide uniformly extreme or middle ground ratings ('extremity/central tendency' responding). Overall, the forced-choice formats eliminate any systematic response sets that apply uniformly across items (Cheung and Chan, 2002).

Second, forced choice tackles the problem with lack of differentiation in ratings (the so-called 'halo' effects). Halo effects are particularly problematic in personality assessments involving external raters (such as spouses, colleagues, or bosses) who often have overgeneralized perceptions of different characteristics of the assessment target based on one important dimension. Forcing choice between various characteristics of the assessment target facilitates finer nuances of judgment and reduces halo effects, enhancing the quality of data.

Third, binary preferences do not require any rating scales since items are compared directly. This is an advantage since test takers do interpret verbal and nonverbal anchors provided with the rating scale differently. Furthermore, Maydeu-Olivares and Böckenholt (2008) argue that comparing items directly may be cognitively simpler than rating them, particularly when there are many rating categories with few or poor verbal anchors.

Finally, the use of forced-choice formats in personality assessments has been largely motivated by attempts to reduce socially desirable responding. It has been thought from conception of forced-choice personality measures that

combining equally desirable items in the same block would reduce socially desirable responding compared to single-stimulus formats, where all desirable items can be easily endorsed and all undesirable ones can be rejected. Extreme forms of socially desirable responding, often referred to as 'faking good,' are particularly concerning in high-stakes personality assessments, where interest in the use of forced-choice questionnaires has been growing. Over the years, evidence for superiority of forced choice in high stakes (e.g., Christiansen et al., 2005; Jackson et al., 2000) as well as against it (Feldman and Corah, 1960; Heggestad et al., 2006) has been published. Findings are inconclusive for many reasons; including lack of control for differences in questionnaire designs and testing contexts as well as technical challenges in modeling forced-choice data (see Section Scaling of Forced-Choice Responses). Good methodology is essential to move this research forward, but most importantly, good understanding of test takers' cognitions when completing personality assessments in high stakes. While test takers' cognitions have been studied with single-stimulus measures (e.g., Robie et al., 2007), there is a clear gap in our understanding of such cognitions in forced-choice assessments.

Scaling of Forced-Choice Responses

While the forced-choice formats have been shown to eliminate or reduce many types of response biases, their use in personality assessment has been very controversial. The reason is that classical scoring methods, when applied to forced-choice questionnaires, yielded test scores inappropriate for interindividual comparisons. Recently, this problem has been overcome by the use of item response modeling.

Classical Scoring Model and Ipsative Scores

In the classical scoring scheme for single-stimulus items, more points are awarded to items that respondents endorse to a higher degree. The same logic has been applied to the forced-choice questionnaires. If item A is preferred to item B, it ought to be awarded more points. For instance, in a 'percentage-of-the-total-amount' task, the points that respondents award to items become item scores. Because all respondents have the same number of points to distribute, their total score for each block is the same (for instance, 100).

	Scale	*Person X*	*Person Y*
A. I pay attention to details	Conscientiousness	40	5
B. I change my mood a lot	Neuroticism	0	25
C. I have a good word for everyone	Agreeableness	10	60
D. I catch on to things quickly	Openness	50	10
Total		100	100

When the item scores are then added to make up scale scores, the total score (the sum of all scale scores) on a classically scored forced-choice test is the same for everyone. The same logic applies to classical scoring of all forced-choice designs. For example, in a 'most/least' partial ranking of four items, the most preferred item is given 2 points, the next two items are given 1 point each, and the least preferred item is given 0 points. The total amount of 4 points per block is the same for everyone, and the total score on the test is a constant.

This type of data is called 'ipsative' (from Latin 'ipse' – he, himself), because the scale scores are relative to self. Indeed, Person X in the example above obtained 50 points for Openness in one block, and this is his highest score in that block. However, because respondents have to indicate their preferences even if they agree with all items or disagree with all of them, this high score is the only high relative to other scores in the same block. It is possible that the absolute standing on Openness for Person X is in fact lower than that for Person Y, who only obtained 10 points for the same item. Ipsative scores are useful for intraindividual assessments because they provide valuable information of relative standings on personality traits for one person. They can be used in personal development, feedback, counseling, etc. They, however, cannot be used for interindividual comparisons.

The fact that the total score on the test is constant, despite different compositions of scale scores within that total for different people, causes substantial psychometric problems well described in the literature. As Clemans (1966), Hicks (1970), and others showed, ipsative scores cannot be analyzed in the same way as normative scores. The problems can be summarized as follows:

1. Ipsative scale scores always correlate negatively on average, even if the personality traits they measure correlate positively.
2. Ipsative scores cannot be factor analyzed using maximum likelihood method, because their variance–covariance matrix is of a reduced rank. Principal components analysis can be applied to ipsative scores; however, the results are difficult to interpret since the components tend to contain conflicting rather than convergent traits (Cornwell and Dunlap, 1994).
3. Ipsative scores distort criterion-related validity estimates since their correlations with any external measure must sum to 0, creating spuriously positive and negative correlations.
4. Internal consistency reliability cannot be applied since the assumption of consistent coding is not met in ipsative scores (Brown and Maydeu-Olivares, 2013).

Different remedies have been suggested to alleviate the impact of ipsative constraints in classical scoring. One such remedy is increasing the number of measured traits in forced-choice questionnaires using the MFC format. It has been argued that when 30 or so traits are assessed, interpersonal comparisons can be performed meaningfully (Baron, 1996). Another way of releasing the ipsative constraint is to present items indicating negative trait standings (for example, 'I miss important details') as well as items indicating positive trait standings (for example, 'I pay attention to details'). Selecting the former statement will take points away from the Conscientiousness score, and selecting the latter will add points. Hence, the total score will show some variation. However, because both types of items indicating Conscientiousness are scored relatively to items indicating other traits, variation of the

total score is still partially constrained (hence the name, 'partially ipsative' data). Partially ipsative data and ipsative data arising from MFC questionnaires measuring many traits are less problematic; however, their psychometric problems are not eliminated but merely reduced.

The troubles with ipsative data stem from the fact that the implicit scoring model bears no relation to the psychological process used in forced-choice judgments (Meade, 2004) – namely, relative positions on items are treated as if they were absolute positions. Adding relative positions of items together cannot possibly constitute a scale score that reflects person's absolute standing on a trait.

Item Response Modeling of Forced-Choice Questionnaire Data

To overcome the problems of ipsative scores, any scoring protocol for forced-choice data must consider the true meaning of item responses. To this end, the response process involved in comparative judgments must be modeled. There are many theoretical models for individual choice behavior. The oldest and the best known is the law of comparative judgment of Louis Thurstone (1927). Other influential models are Coombs' (1950) unfolding preference model, and Luce's (1959) choice axioms. These theories have been adopted for modeling of forced-choice questionnaire data; in addition, Luce's choice axioms have been applied by Andrich (1989) to provide explicit probability of binary choice.

IRT models for forced-choice questionnaire data can be classified according to three criteria: (1) block size that can be modeled, (2) dimensionality of comparisons (whether items measuring different traits can be modeled), and (3) the measurement model thought to underlie absolute judgments about questionnaire items.

Two types of measurement models have been utilized in forced-choice modeling: linear factor analysis models and ideal-point models. Linear factor analysis models assume that person p's tendency to endorse item i (psychological value or item 'utility,' as named by Thurstone) measuring personality trait a is a linear function of his/her trait score. (LFA models allow items measuring multiple traits; a special case of factorially simple items is given here for simplicity.)

$$\text{utility}_{pi} = \text{mean}_i + \text{loading}_i \text{trait}_p^a + \text{error}_{pi} \qquad [1]$$

The factor analysis model assumes that for items indicating positive trait standings (for example, 'I pay attention to details,' indicating Conscientiousness), the item utility will increase as the trait score increases. For items indicating negative trait standings (for example, 'I miss important details'), decrease in item utility is expected with an increase in trait score. This so-called 'dominance' model is a fair representation of response process for most personality items that tend to represent extreme positive or negative standings on personality traits.

Ideal-point models assume that person p's utility for item i measuring trait a is a function of distance between his/her trait score and the item location. (There are many forms of ideal-point models; an expression using the unweighted Euclidean distance is given here without loss of generality.)

$$\text{utility}_{pi} = \text{mean}_i - \left|\text{trait}_p^a - \text{location}_i\right| + \text{error}_{pi} \qquad [2]$$

This family of models assumes that items and people can be placed on the same trait continuum. Each item represents some position on the trait and can be thought of as an 'ideal' item describing characteristics of persons with this level of the trait (hence the name 'ideal point'). For instance, item 'My paperwork is always in order' represents a very high Conscientiousness score, and item 'My attention to detail is about average' represents an average score. For the latter item, the relationship between the item utility and Conscientiousness trait is clearly not linear, because the utility for this item will peak around the average Conscientiousness score, and be lower for respondents with either high or low scores. The ideal-point response process was originally proposed for attitude measurement, but recently has been suggested for use in personality assessments (Drasgow et al., 2010).

Thurstonian IRT Model

Scope

Brown and Maydeu-Olivares (2011) developed the Thurstonian IRT model to enable analysis of data arising from forced-choice tests measuring multiple traits with ranking blocks of any size. Item parameters, correlations between the latent traits, and person trait scores can be estimated using this IRT model.

Items in each block may measure the same trait (UFC) or different traits (MFC), or any combination of the two. A linear factor analysis model is assumed to describe the relationship between the items and traits they measure (i.e., the model utilizes the most common type of personality items, dominance items).

Origins

This model is based on Thurstone's (1927) law of comparative judgment. Thurstone postulated that preference judgments are determined by utilities of items under comparison. Each item elicits a utility judgment (judgment of the item's psychological value), and person p will rank item i above item k if his/her utility for i is higher than for k:

$$\text{prefer item} \begin{cases} i & \text{if} \quad \text{utility}_{pi} \geq \text{utility}_{pk} \\ k & \text{if} \quad \text{utility}_{pi} < \text{utility}_{pk} \end{cases} \qquad [3]$$

Observed rank orders of items within a block, therefore, can be seen as manifestations of the order of corresponding item utilities. To describe the observed ranks in a block, all pairwise comparisons between items are considered. There are $n(n-1)/2$ nonredundant pairwise comparisons in a block of size n. For example, any given rank order of four items, A, B, C, and D, is fully described by knowing outcomes of six paired comparisons: {A, B}, {A, C}, {A, D}, {B, C}, {B, D}, and {C, D}. For any of these six pairwise comparisons, the utility maximization rule [3] applies.

Model

Because preference for item i over item k is determined by the difference of their utilities, the utility difference is the central unit of analysis in Thurstonian modeling. To model responses to ranking blocks in relation to personality traits the items measure, Brown and Maydeu-Olivares (2011) assumed that

person's utilities for items relate to the traits via a linear factor analysis model. Thus, the utility differences can be expressed as linear combinations of personality traits using [1]. Observed binary preferences are thought to be dichotomized expressions of unobserved utility differences. With this, the probability $P(y_{p\{i,k\}} = 1)$ of preferring the first item in a pair $\{i,k\}$ of items measuring different traits a and b, conditional on person p's trait score is given by the cumulative standard normal function.

$$P(y_p\{i,k\} = 1) = \Phi\left(\frac{-\text{threshold}_{\{i,k\}} + \text{loading}_i \text{trait}_p^a - \text{loading}_k \text{trait}_j^b}{\sqrt{\text{var}(\text{error}_i) + \text{var}(\text{error}_k)}}\right) \quad [4]$$

For items with positive factor loadings, the probability of preferring item i to item k increases when the score on the trait measured by item i increases and the score on the trait measured by item k decreases. The response function [4] defines a surface, an example of which is presented in **Figure 1**, where the probability of preferring one item to another is plotted against two traits. The surface illustrates that the dominance process of evaluating individual items results in a monotonic relationship between the person's trait scores and the choice probability.

When items measuring the same trait are being compared in a forced-choice block, the conditional probability of preferring item i to item k by person p [4] simplifies to

$$P\left(y_{p\{i,k\}} = 1\right) = \Phi\left(\frac{-\text{threshold}_{\{i,k\}} + (\text{loading}_i - \text{loading}_k)\text{trait}_p^a}{\sqrt{\text{var}(\text{error}_i) + \text{var}(\text{error}_k)}}\right) \quad [5]$$

It follows from [5] that two items with similar factor loadings measuring the same trait will yield a pairwise comparison with a near-zero pairwise factor loading; thus it provides very little information for measurement of the latent trait. In other words, items that are likely to have very similar valuations by the same person are ineffective as a pair for determining the absolute standing on the trait. This is a fundamental property of comparative judgments rather than of a specific model; this feature is shared by all forced-choice models, as the reader shall see.

The unidimensional response function is a curve, examples of which are presented in **Figure 2**. The figure illustrates that two items with very different factor loadings yield a highly informative comparison, whereas two items with very similar factor loadings (albeit highly discriminating in their own right) yield an uninformative comparison.

To enable parameter estimation, the pairwise preference model [4] is embedded in a generalized Structural Equation Modeling (SEM) framework. Responses to all ranking blocks (coded as binary outcomes of pairwise comparisons using [3]) are modeled simultaneously, each outcome serving as an indicator of common factors (latent personality traits), resulting in a single measurement model with binary outcomes (IRT model). The model takes care of dependencies that exist in blocks of size $n > 2$, where pairwise comparisons involving the same item share the same factor loading and the same error term. The model estimates item parameters according to [4] – factor loading and error variance for every item, and threshold for every pairwise comparison. In addition, correlations between the latent traits are estimated.

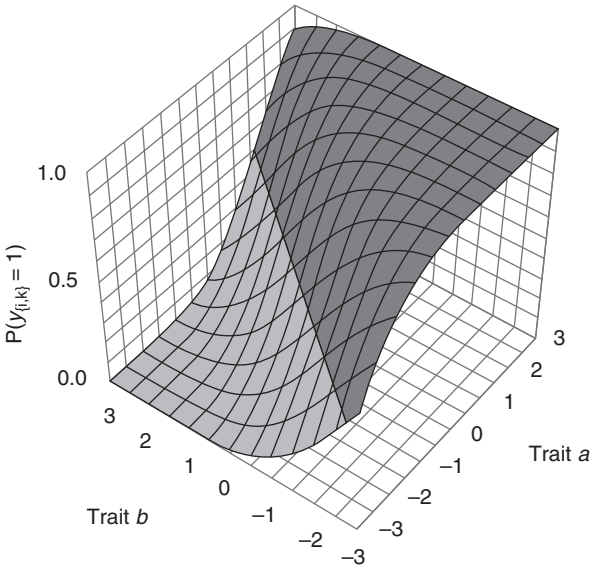

Figure 1 Example Thurstonian IRT response function for a pair of items measuring two different traits.

After the item parameters have been estimated, person trait scores can be estimated by either maximum likelihood method or Bayesian estimation. It has been shown that Thurstonian item response modeling approach overcomes the problems of ipsative data (Brown and Maydeu-Olivares, 2013).

Applications

Applications of this flexible model have included reanalysis of existing forced-choice questionnaire data and development of new measures. Reanalysis of the CCSQ demonstrates the advantages of IRT modeling for personality assessment; specifically, interpersonal comparability of person trait scores estimated by the IRT method as opposed to the classical method resulting in ipsative scores (Brown and Maydeu-Olivares, 2013). The development of a new IRT scored version of the Occupational Personality Questionnaire (OPQ32r) illustrates how Thurstonian IRT modeling may be applied to reanalyze and redevelop an existing assessment tool, enhancing its strong features and transforming its scoring protocol (Brown and Bartram, 2009). The Thurstonian IRT model was also used to inform the development of a new measure, the Forced-Choice Five Factor Markers (Brown and Maydeu-Olivares, 2011).

Zinnes–Griggs Model for Unidimensional Pairwise Preferences

Scope

Zinnes and Griggs (1974) developed an IRT model to enable analysis of data arising from forced-choice tests consisting of simple paired comparisons (block size $n = 2$) of items measuring the same trait (UFC). Item parameters and trait scores for individuals can be estimated using this IRT model. An ideal-point process is assumed to describe the relationships between test items and the trait they measure.

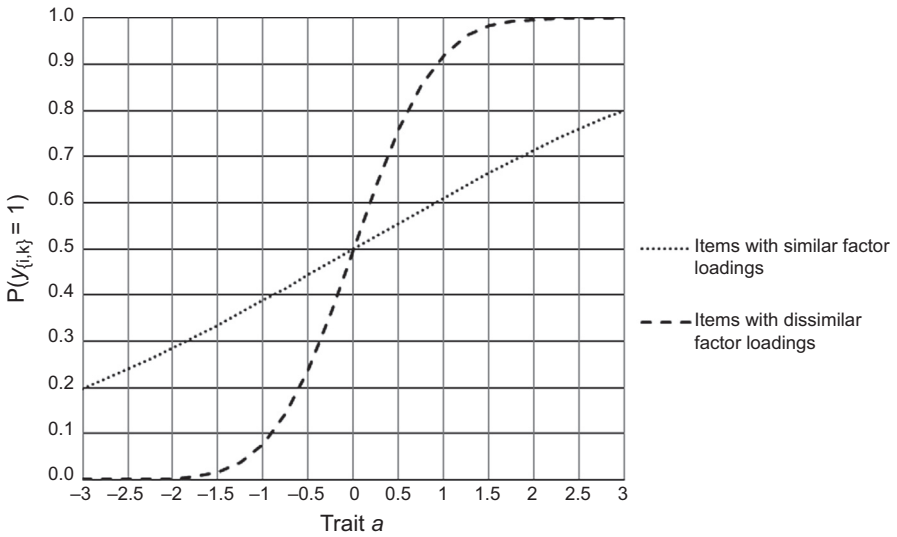

Figure 2 Example Thurstonian IRT response functions for two pairs of items measuring the same trait.

Origins

This model is based on Coombs' (1950) preference decision theory. According to Coombs, when facing a choice between two items, the person will prefer the item closer to an 'ideal' item representing own position on the trait ('ideal point'). Formally, given a choice between items i and k, each with their own location on the trait, person p will

$$\text{prefer item}\begin{cases} i & \text{if} \quad |\text{trait}_p - \text{location}_i| \leq |\text{trait}_p - \text{location}_k| \\ k & \text{if} \quad |\text{trait}_p - \text{location}_i| > |\text{trait}_p - \text{location}_k| \end{cases} \qquad [6]$$

Coombs' model implicitly assumes an ideal-point response process for every item involved in comparison; that is, the psychological value or utility of an item to a person equals the inverse of the distance between the person trait score and the item location. Because of this inverse relationship between the distance to an item and the person's utility for it, Zinnes and Griggs called the item-person distance 'disutility.'

Model

Zinnes and Griggs (1974) modified the original deterministic version of Coombs' decision model to make it probabilistic. To this end, they consider 'noisy perceptions' of item locations and the person's own ideal point at the time of comparison – three random variables distributed normally around their expected values. They showed that the probability $P(y_{p\{i,k\}} = 1)$ of preferring the first item in a pair $\{i,k\}$ of items measuring the same trait, conditional on person p's trait score is a one-dimensional IRT model, where $\Phi(\cdot)$ is a cumulative standard normal distribution function:

$$\begin{aligned} &P\left(y_{p\{i,k\}} = 1\right) \\ &\quad = 1 - \Phi\left(a_{p\{i,k\}}\right) - \Phi\left(b_{\{i,k\}}\right) + 2\Phi\left(a_{p\{i,k\}}\right)\Phi\left(b_{\{i,k\}}\right) \end{aligned} \qquad [7]$$

where $a_{p\{i,k\}} = (2 \cdot \text{trait}_p - \text{location}_i - \text{location}_k)/\sqrt{3}$; $b_{\{i,k\}} = \text{location}_i - \text{location}_k$.

In this simple model, the conditional probability depends only on the person's trait score and the item locations. It is therefore assumed that items vary only in their locations on the trait continuum, and thus they are equally good measures of the trait (equally discriminating).

An example response function is illustrated in Figure 3, where the conditional probabilities of preferring the first item in a pair are plotted for two pairs of items, with different combinations of item location parameters. This figure shows that comparing items with similar locations results in a very 'flat' function with a shallow slope; and comparing items with very dissimilar locations results in a function with a steep slope. Therefore, comparisons between items located closely on the same trait are noninformative. The same effect is observed when items with similar factor loadings are compared under the Thurstonian IRT model (see Figure 2).

Applications

This straightforward unidimensional model has been applied to create the NCAPS.

Multi-Unidimensional Pairwise-Preference Model

Scope

Stark et al. (2005) developed the Multi-Unidimensional Pairwise-Preference (MUPP) model to estimate person trait scores from binary paired comparisons (block size $n = 2$) of items measuring the same trait (UFC) or different traits (MFC). Currently, item parameters cannot be estimated from the actual forced-choice data and are assumed known. The MUPP model assumes an ideal-point response process for the items involved in comparisons.

Origins

The MUPP model adopts an approach to explaining preference judgments originally suggested by Andrich (1989). Andrich's motivation was to write an explicit expression for the

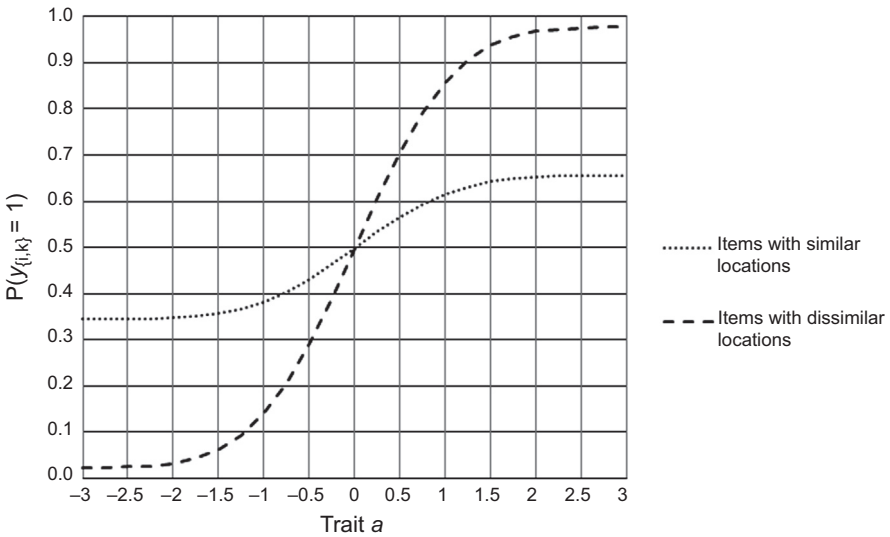

Figure 3 Zinnes-Griggs response functions for two pairs of items measuring the same trait.

probability of preferring one item to another through probabilities of accepting and rejecting the individual items. First, he postulated that the event of picking item i from the set $\{i,k\}$ has the same probability as the event of picking the first of two alternatives: (1) endorsing i and rejecting k, or (2) endorsing k and rejecting i.

Second, Andrich assumed that acceptances and rejections of individual items are independent events, conditional on the traits the items measure; therefore the joint probability of endorsing i and rejecting k is the product of two probabilities, the probability of endorsing i and the probability of rejecting k. With this, the probability $P(y_{\{i,k\}} = 1)$ of preferring the first item in a pair $\{i,k\}$ of items measuring different traits a and b, conditional on person p's trait score is given by

$$P\left(y_{p\{i,k\}} = 1\right) = \frac{P\left(\text{endorse } i \middle| \text{trait}_p^a\right) P\left(\text{reject } k \middle| \text{trait}_p^b\right)}{P\left(\text{endorse } i \middle| \text{trait}_p^a\right) P\left(\text{reject } k \middle| \text{trait}_p^b\right) + P\left(\text{reject } i \middle| \text{trait}_p^a\right) P\left(\text{endorse } k \middle| \text{trait}_p^b\right)} \quad [8]$$

If items measuring the same trait are compared, the MUPP gives a response curve similar to those depicted in Figure 3. Just like under the Zinnes–Griggs model, comparisons between unidimensional items with similar locations are noninformative, and the same effect is observed when items with similar factor loadings are compared under the Thurstonian IRT model (see Figure 2).

Applications

The MUPP model was used in the development of the TAPAS, a comprehensive and customizable assessment system measuring any subset of 23 personality facets deemed important for predicting job performance in civil and military organizations.

Model

Stark et al. (2005) use an ideal-point model to substitute the probabilities of accepting and rejecting individual items in general expression [8]. Specifically, they advocate the use of a binary version of the Generalized Graded Unfolding Model or GGUM (Roberts et al., 2000). The GGUM is very flexible and allows items to differ in discrimination, locations, and even in maximum probability of endorsement.

If items measuring different traits are compared, the item response function [8] in conjunction with the binary GGUM defines a surface, an example of which is presented in Figure 4. This surface is rather complex, with the ideal-point process of evaluating individual items resulting in a nonmonotone relationship between the latent traits and the pairwise preference.

McCloy–Heggestad–Reeve Unfolding Model for Multidimensional Ranking Blocks

Scope

McCloy et al. (2005) sketched a model for the process of responding to MFC blocks of any size, compiled from ideal-point items, to inform item selection and estimate person trait scores. Item parameters cannot be estimated from the actual forced-choice data and are assumed known.

Origins

This approach adopts an extension of Coombs' original one-dimensional unfolding model to the multidimensional case. The model predicts that a person will prefer an item that is located nearer to the person in multidimensional trait space than another item located further from the person in that

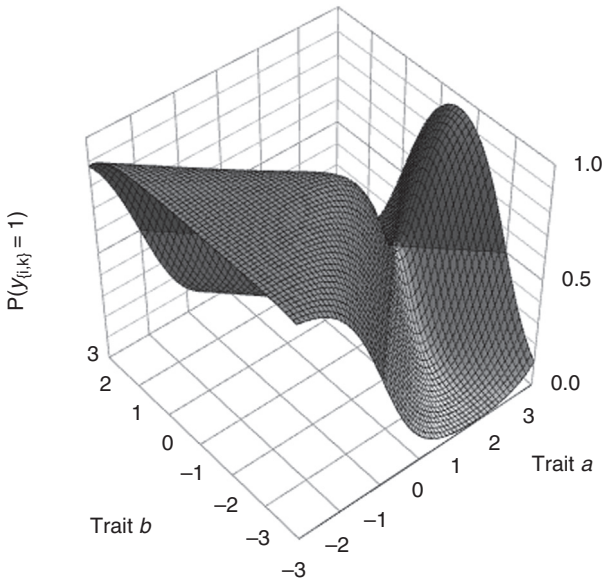

Figure 4 MUPP response function for a pair of items measuring different traits.

space. The use of Coombs' unfolding decision model implicitly assumes the ideal-point response process for every item involved in comparisons, each reflecting changes in its own personality trait.

Model

Items are assumed to be equally good measures of the traits (equally discriminating); therefore the only differentiating feature for items is their locations. This and another simplifying assumption – that of no random influences on preference decisions – are used to inform creation of forced-choice designs that are effective for accurate estimation of trait scores. McCloy and colleagues show that assuming item locations known, an effective forced-choice measure can be assembled from blocks of items with locations that vary across the attribute space.

Discussion

Since conception of forced-choice personality measures, the biggest technical challenge was scaling of person's relative preferences to reflect their absolute standing on traits of interest. Classical scoring methods that simply counted the relative responses as absolute did not provide adequate measurement, making the forced-choice assessment method controversial. Modern scoring methods have the potential to provide proper scaling by taking to account the comparative nature of judgments. These methods link observed comparative decisions to unobserved personality traits, controlling for properties of items.

Despite the breakthroughs in scaling forced-choice questionnaire data, not every forced-choice design is guaranteed to deliver good measurement properties once an IRT model has been applied to it. Just like a single-stimulus questionnaire with poor items that do not provide information on target traits or with a poor rating scale, a forced-choice questionnaire may be incapable of providing accurate measurement of target traits whichever modern scoring model is applied. This is because the forced-choice format dictates its own rules of measurement.

One such rule was demonstrated in this article – comparisons between items that are likely to elicit similar value judgments in the same person are not informative for establishing the absolute trait position. For example, two items, 'My paperwork is always in order' and 'I am always on top of my paperwork,' may be good measures of Conscientiousness in their own rights, but when put in a paired comparison, they will provide no information on the Conscientiousness trait. This is because high scorers will agree with both items and will be forced to make a random choice between the two. Low scorers will disagree with both items and will again make a random choice. The same will happen for any given standing on the trait – the trait score will have no effect on the preference decision, with the random error playing the largest part. No IRT model, whether utilizing dominance items or linear factor analysis items, will be able to recover the absolute position on the trait in this situation. This is the fundamental property of comparative judgments, not of any specific model.

More generally, items that elicit similar utilities, using either the dominance process or the ideal-point process, are ineffective in forced-choice blocks. This is why comparing items measuring highly correlated traits is not effective, and combining items with factor loadings of opposite sign (measuring the positive and the negative extremes of positively correlated traits) may be necessary (Brown and Maydeu-Olivares, 2011). The same conclusions follow when ideal-point items are utilized in multidimensional forced choice – items with very different locations on traits are required (McCloy et al., 2005).

There has been much debate about relative merits of dominance and ideal-point items as utilized in forced-choice measurement (e.g., Drasgow et al., 2010; Brown and Maydeu-Olivares, 2010). However, it is becoming clear that all models for forced-choice are essentially similar; they model the same process, which has its own fundamental properties. The successful applications of both types of models in personality assessment show that neither dominance nor ideal point family model is inherently superior for forced-choice measurement. Rather, the choice of model must be dictated by the nature of items used. If items that elicit the ideal-point response process are used (because they deemed more appropriate on conceptual grounds), an ideal-point forced-choice model is appropriate. If items that elicit dominance responses are used, a linear factor analysis forced-choice model is appropriate.

Finally, it is worth noting that all the existing IRT models are choice models; that is, they are suitable for modeling binary preference data (i.e., paired comparisons, rankings, and partial rankings). Although extensions to ordinal or proportional preference data are possible within the generalized SEM framework adopted by Thurstonian modeling, these are yet to establish themselves in forced-choice personality assessment. Such developments would open doors for proper scaling of Q-sorts and questionnaire employing 'proportion-of-total-amount' tasks.

Conclusions and Directions for Future Research

The main objective of personality assessment – differentiating people in relation to traits of interest – can be achieved by collecting either absolute judgments or comparative judgments. In respondent-reported assessments, absolute judgments can be subject to numerous response biases, such as idiosyncratic uses of rating scales, unconditional agreement with statements as presented (acquiescence), lack of differentiation (halo effects), and others. Forcing choices between questionnaire items can help eliminate any effects acting uniformly across items. Reducing nonuniform effects, such as differential inflation of scores on traits deemed important in specific contexts ('faking good'), with forced-choice formats is proving more challenging. More research is needed to understand the motivated distortions to forced-choice questionnaires in high-stakes personality assessments. With rapid development of item response modeling approaches, we are well placed to tackle this challenge.

Even before the advent of item response modeling in the forced-choice assessment space, ipsative and partially ipsative scores demonstrated similar or even enhanced criterion-related validities compared to single-stimulus measures (Bartram, 2007; Salgado and Táuriz, 2014). This improvement is presumably due to reduction in response biases detrimental to validity. The psychometric problems of ipsative scores, however, should not be underestimated as they can result in spuriously significant validity coefficients (Brown and Maydeu-Olivares, 2013). It is my belief that conclusive validity evidence for forced-choice assessments can only be gained by using model-based measurement. With increasing repertoire of properly scaled forced-choice measures, there is much to look forward to in this area of research.

Recent developments in item response modeling have enabled the use of forced-choice personality assessments without the downsides of ipsative data. The potential reduction in response biases that forced-choice measures can provide, while maintaining interpersonal comparability of trait scores, is an exciting prospect for many applications in personality assessment. For instance, cross-cultural personality research, where culturally specific response sets present a challenge for score comparability, could benefit from the use of direct comparative judgments. Assessments by external raters, whether in workplace, health, or education, could also benefit from the use of carefully designed forced-choice questionnaires to enhance validity by reducing rater effects such as halo and leniency/severity. Provided that appropriate methods are used to design and score such assessments, the forced-choice formats can be a viable alternative to the single-stimulus formats.

See also: Personality Assessment, Faking and.

Bibliography

Andrich, D., 1989. A probabilistic IRT model for unfolding preference data. Applied Psychological Measurement 13, 193–216.

Baron, H., 1996. Strength and limitations of ipsative measurement. Journal of Occupational and Organizational Psychology 69, 49–56.

Bartram, D., 2007. Increasing validity with forced-choice criterion measurement formats. International Journal of Selection and Assessment 15, 263–272.

Block, J., 1961. The Q-sort Method in Personality Assessment and Psychiatric Research. Charles C. Thomas, Springfield, IL.

Brown, A., Bartram, D., 2009. Doing Less but Getting More: Improving Forced-choice Measures with IRT. http://www.shl.com/assets/resources/Presentation-2009-Doing-less-but-getting-more-SIOP.pdf (accessed 28.02.14.).

Brown, A., Maydeu-Olivares, A., 2010. Issues that should not be overlooked in the dominance versus ideal point controversy. Industrial and Organizational Psychology 3, 489–493.

Brown, A., Maydeu-Olivares, A., 2011. Item response modeling of forced-choice questionnaires. Educational and Psychological Measurement 71, 460–502.

Brown, A., Maydeu-Olivares, A., 2013. How IRT can solve problems of ipsative data in forced-choice questionnaires. Psychological Methods 18, 36–52.

Cheung, M.W.L., Chan, W., 2002. Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. Structural Equation Modeling 9, 55–77.

Christiansen, N.D., Burns, G.N., Montgomery, G.E., 2005. Reconsidering forced-choice item formats for applicant personality assessment. Human Performance 18, 267–307.

Clemans, W.V., 1966. An Analytical and Empirical Examination of Some Properties of Ipsative Measures. Psychometric Monograph 14. Psychometric Society, Richmond, VA.

Coombs, C.H., 1950. Psychological scaling without a unit of measurement. Psychological Review 57, 145–158.

Cornwell, J.M., Dunlap, W.P., 1994. On the questionable soundness of factoring ipsative data: a response to Saville & Willson. Journal of Occupational and Organizational Psychology 67, 89–100.

Drasgow, F., Chernyshenko, O.S., Stark, S., 2010. 75 years after Likert: Thurstone was right! Industrial and Organizational Psychology 3, 465–476.

Feldman, M.J., Corah, N.L., 1960. Social desirability and the forced choice method. Journal of Consulting Psychology 24, 480–482.

Heggestad, E.D., Morrison, M., Reeve, C.L., McCloy, R.A., 2006. Forced-choice assessments of personality for selection: evaluating issues of normative assessment and faking resistance. Journal of Applied Psychology 91, 9–24.

Hicks, L.E., 1970. Some properties of ipsative, normative, and forced-choice normative measures. Psychological Bulletin 74, 167–184.

Jackson, D.N., Wroblewski, V.R., Ashton, M.C., 2000. The impact of faking on employment tests: does forced choice offer a solution? Human Performance 13, 371–388.

Luce, R.D., 1959. Individual Choice Behavior: A Theoretical Analysis. John Wiley, New York, NY.

Maydeu-Olivares, A., Böckenholt, U., 2008. Modeling subjective health outcomes: top 10 reasons to use Thurstone's method. Medical Care 46, 346–348.

McCloy, R.A., Heggestad, E.D., Reeve, C.L., 2005. A silk purse from the sow's ear: retrieving normative information from multidimensional forced-choice items. Organizational Research Methods 8, 222–248.

Meade, A.W., 2004. Psychometric problems and issues involved with creating and using ipsative measures for selection. Journal of Occupational and Organizational Psychology 77, 531–552.

Roberts, J.S., Donoghue, J.R., Laughlin, J.E., 2000. A general item response theory model for unfolding unidimensional polytomous responses. Applied Psychological Measurement 24, 3–32.

Robie, C., Brown, D.J., Beaty, J.C., 2007. Do people fake on personality inventories? A verbal protocol analysis. Journal of Business and Psychology 21, 489–509.

Salgado, J.F., Táuriz, G., 2014. The Five-Factor Model, forced-choice personality inventories and performance: a comprehensive meta-analysis of academic and occupational validity studies. European Journal of Work & Organizational Psychology 23, 3–30.

Stark, S., Chernyshenko, O.S., Drasgow, F., 2005. An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: the Multi-Unidimensional Pairwise-Preference Model. Applied Psychological Measurement 29, 184–203.

Thurstone, L.L., 1927. A law of comparative judgment. Psychological Review 34, 273–286.

Zinnes, J.L., Griggs, R.A., 1974. Probabilistic, multidimensional unfolding analysis. Psychometrika 39, 327–350.

Personality Assessment: Overview

Adrian Furnham, University College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Cervone, G.V. Caprara, volume 16, pp. 11281–11287, © 2001, Elsevier Ltd.

Abstract

This article looks at different essential methods to assess personality and individual differences. It considers a simple but robust selection model, which includes selecting *out* and selecting *in* people at work. It then goes on to consider current trends and recent developments in the area. The main part of this article considers, in more detail, the advantages and disadvantages of the standard methods like interviews, references, biodata, and tests. The final part of the article concerns addressing 12 myths concerning psychometric testing in general.

Introduction

There is a great deal of interest and research into assessing people at work. Assessment can be formal and informal; based on test data or personal intuitive judgments; with more or less serious consequences. Many people invest a large amount of time and energy in assessing others which they believe is a good investment.

Assessing people at work is important for many reasons. The most important is the cost benefit analysis: *the benefits of the right decision* over *the costs of getting it wrong*. In other words, there are great financial benefits for hiring a positive and productive person; and many costs, especially financial, to hiring an unhappy, unproductive, and difficult person.

Using good assessment tools has other benefits: it can 'up-skill' managers who use various tools and techniques and *increase their psychological mindedness*. It also has a major benefit for the interviewee as feedback can considerably *increase their self-awareness*. Further, if a test is used widely in an organization people can often have a *useful shared language* to discuss issues in terms of psychological profiles and preferences.

The question for assessors is essentially *what to assess, who is best suited to do it, when and why*. To some extent the 'what' can neatly be divided into three areas.

What a person can do? This refers to their *ability*. It is about their capacity to do various tasks efficiently given that they have the desire to do so. It also refers to their ability to learn new tasks. Assessing what a person can do is more often measured by cognitive ability (intelligence) and skills tests though it may also be useful to assess creativity as well as the ability to lead others.

What a person will do? This refers to a person's *motivation* or what they want to do. Motivation refers to a person's values and drives. Everyone can be persuaded to do things as a function of rewards and punishments but this refers to what a person will do on an everyday basis without strong rewards or punishments trying to shape behavior.

What a person wants to do? This refers to preferences/*personality* for certain activities over others. It is about what a person likes to do and will do so freely without any form of coercion. It is about their values and personality and motivation which pushes them in one direction or another.

Assessors need to know all three things about a person they are assessing.

The Essential Methods

There are, in essence, mainly five different methods to collect data on people. Of these the first three are most commonly used.

Self-Report

This is essentially what people say about themselves in interviews (both structured and unstructured); in self-report personality and other preference tests; and on their CV (personal statement or application form). These are very common ways of assessing people and they nearly always want and expect an interview when they can answer questions and talk about themselves.

There are however two major problems with self-report. *The first is called under various names like dissimulation, faking, or lying*. It concerns people giving false information, or embellished information about themselves. This behavior has been broken down by psychologist into two further types of behavior.

Impression Management

This is where the person attempts to create a good impression by leaving out information, adding untrue information (errors of omission and commission) as well as giving answers that are not strictly correct but, they hope, create a good impression in the interviewer's mind. This is done consciously and is very common. Indeed, it is expected in the answer to some questions but it can be very serious when, for instance, people claim to have qualifications or experiences they have not had, or leave out important information (about their health, criminal past).

Self-Deception

This occurs when a person, in their own view answers honestly, but what they say is untrue because they lack self-awareness. Thus they might honestly believe that they are a 'good listener' *whereas* all the evidence from reliable sources is that

this is not true. This can occur for both good qualities (cognitive and emotional intelligence) as well as weaknesses (impulsivity, depression). People with low self-awareness often self-deceive.

The way personality and other preference tests attempt to deal with this issue is to use Lie Scales in the test. These go under various names (validity, dissimulation) and many exist. They are generally known as measures of response bias.

The second is about self-insight. This is primarily concerned with what people cannot say about themselves even if they wanted to. This is best seen with issue around motivation where people cannot, rather than will not, give honest answers about the extent to which they are motivated by power, recognition, money, or security. Indeed motivation is one of the most difficult topics to accurately assess, and yet, for business people, though as among the most important.

One way psychologists have tried to deal with this issue is projective techniques such as the Thematic Apperception Test, or various sentence completion tasks. For instance, a person may be asked to complete the following sentence: "My greatest regret is…"; "At work I often appear to be…"; "My parents would be most proud of my…" The idea is to content analyze these responses to detect themes. These methods are expensive in terms of time and unreliable meaning too many minor factors (like their mood and where the interview takes place) affects their response. They remain 'unproven' for most psychometricians.

Observation Data

This is what other people say about an individual in references and testimonials; 360 ratings (multisource feedback) and appraisal and other performance management data. Most organizations attempt to get reliable reports from other people who know the candidate that they are attempting to assess. Many application processes ask candidates to list people who know them well in some salient setting and may be called upon.

There are also problems with this type of data. The first is *the* 'data bank' *of the observer*. This essentially means what information the observer has about the candidate. Thus a boss has a different data set than a colleague or a subordinate. A school teacher or university lecturer will have a different data set than an employer. The question is *what they know*: the quality, and quantity of data on a person's ability, motivation, work style, etc.

The second issue is the extent to which they are prepared to *tell the truth about an individual.* Some organizations refuse staff to write references because of litigation. They can be taken to court for what they did say or did not say. They are told all they can say is that "X worked here from dates A to dates B." People sometimes have a vested interested in 'talking up' a person they want to leave, or else they do not disclose important aspect about an individual's work history such as accusations of inappropriate behavior.

People also choose their specific, favorite referees because they hope that they will be very positive. There seems to be an etiquette with respect to what people write or rate on references. Many know the power of negative information and therefore try strongly to resist providing any negative information. It is therefore rare to get very accurate or useful data on a person's weaknesses or challenges from references.

Test Performance

This refers to how well people do on tests: power, timed, ability tests. These are of *maximum performance*. There are also preference, untimed, personality tests which are *of typical performance*. There are also behavioral tests, which measure actual observed and measured behavior often in groups.

There are thousands of tests of many types to choose from. Most professionals only know about a smaller number. They also are often not clear about why one test maybe psychometrically better than another. This is nearly always a question of evidence of the psychometric properties of the test. Tests differ enormously. Some require a one-to-one administration, others can be easily and effectively be run in largish groups. Some are 'objective' versus open-ended tests: the former requires the choice of several responses; the latter means one has to generate the response.

Physiological Evidence

This is probably the newest and most disputed of all measures. Thus for some jobs, employees have to go through a '*medical checkup*' which they may have to do on a regular (i.e., annual) basis simply to keep their job. This would be true of such jobs as being an airplane pilot. In other jobs, for instance, working in the alcohol industry, it maybe a requirement that people go through a *liver function test.* Simple blood tests and saliva samples can be used for various diagnoses including drug taking and stress levels. Everyday it seems new simple physical measures are being devised that are claimed to be able to detect such things as whether a person is more likely to get a debilitating mental or physical disease.

Personal History/Biography

This refers to person's personal history, for instance, where they were born and educated; the family from which they came;' and their present family and address. Some information is thought to be very important such as what was the social class of the parents; does the person come from a minority race or religious group; how many brothers and sisters do they have and what is their place in the birth order; what was their schooling like and how successful were they at it.

This area is called biodata. It aims to determine, through empirical methods the biographical markers of success in very particular jobs. It has limitations which will be discussed later.

A Simple Selection Model

Below is the simplest selection model. The aim is to select the good and reject the bad candidates. Through job analysis selectors usually have a list of competencies that they are looking for. However there are two problems with this model. First, the assumption of linearity: this is the idea that more is better. The more you have of a quality (intelligence, creativity, integrity) the better. For most jobs you need an optimal

amount rather than a maximal amount. Second, the failure to select out: that is to actively look for things that you do not want in the person being assessed.

	Good	*Bad*
Select	A	B
Reject	C	D

In this model, B and C are errors. Everyone is concerned with getting A but nobody with D.

The more important the job and the more the consequences of failure or derailment count, the more important it is to seek assess potential derailers.

Recent Developments: What Is Hot and What Is Not

There are a number of new definable products, trends, and developments in the area. These are driven by various different forces.

First, there are *changes in the law*: legal changes and litigation risks have driven some issues (e.g., integrity testing, diversity training). Second, there are *changes in business*: as many organizations attempt become more flexible and competitive they become concerned with specific assessment-related issues, e.g., in spotting and managing talent; in strengthening leadership; in managing restructuring, acquisitions, and mergers. Third, there are the *ideas of gurus (business writers)*: popular books highlight various concepts, issues, and methods, some of which are enthusiastically embraced by business people (e.g., emotional intelligence). Fourth, there are the *recommendations of consultants/academics*: they often have their own agenda, which might or might not be related to the above.

Thus it is possible to see the emergence of various concepts, products, and measures used to assess people for jobs and to develop them. Further just as some of these products begin to emerge others tend to wane. Thus, 360° evaluation/feedback is currently probably at its zenith but it now much less popular than it has been. Equally, outward bound/outdoor training to strengthen teamwork seems on the wane.

Other factors have also played a part in developments in this area. First, *labor market shortages*: Shortages across Europe have led many companies to rethink their strategy and process to ensure it is fair to all concerned and that they can attract and correctly identify talented individuals. Second, *technological developments* and two issues are relevant here: The administration of tests via computers as well as recruitment and testing online via the Internet. Third, *applicant perceptions* which refers to the perceptions of applicants as to the fairness and validity of the assessment process. This is, in part, not only an impression management task for all organizations, but also refers to the effect of assessment methods on job acceptance and subsequent performance. Fourth, *construct-driven approaches* which refers to being clear about what one is trying to assess and why. That is having a theory and evidence for the factors/constructs that one is trying to assess and show their predictive validity.

There have been very specific developments and various trends are noticeable. There have been many advances in new data tools. For instance, there are now new *interactive tests* where people respond to carefully prepared video/acted scenes on a computer. These can even be delivered using mobile devices. They can be adapted for very specific company uses. One of the newest, but as yet untested ideas is the use of Avatars in assessment situations. There is now more and more interest in samples of real-life behavior sometimes called 'thin slides' whereby people make accurate judgments of real-life behavior perhaps recorded by video. There is an increased interest on the test-takers experience and reactions. Most organizations are aware of the limitations of the traditional interview and are making efforts to introduce well structured, panel interviews. Poor validity response rates and litigation issues have meant the traditional reference is being replaced by structured telephone calls asking specific questions of targeted interviewees (often peers).

There is now great emphasis on the power and usefulness of peer and subordinate ratings for assessment purposes. This has been a major consequence of an interest in 360° multirater feedback work. Organizations are being very sensitive to the possibility of bias (sex, race, language) in ability tests. They are also particularly concerned that they have face validity for candidates. There is always the development of tests to measure new (faddish) constructs, e.g., spiritual intelligence, practical intelligence. Psychometric tests continue to be developed in large numbers and put on the market. Some clearly 'chime with the spirit of the time' like tests of *emotional intelligence*. Biodata, psychomotor tests, and tests of job knowledge are *not* frequently used. They have however never disappeared and remain the favorite of particular clients in very particular settings.

The concept of using work-samples and probation periods appears to be gaining more attention. Probation periods if strictly enforced give both parties (employer and employee) a chance to revisit their decisions. Assessment Centers remain the basis of many organizations' preferred methods. They are very expensive and time consuming but recognized as the best. Most are designed very specifically for each company and there particular requirements. There is a great interest in assessing 'people with potential' or 'highfliers' but people are not sure how to do it. Organizations differ widely to whether they should bring assessment expertise in-house or outsource all or some of it. Much less interest and expertise is devoted to assessment for development purposes. This is done through mentoring and coaching which may or may not involve some formal assessment procedures. Organizations are beginning to realize that they can cost the effectiveness of successful assessment/selection in monetary terms which can have important implications.

Assessment Methods

A primary issue in the field of assessment is the wide choice of method to use – ranging from the informal to the highly technical.

The Interview

There are many variations on the job interview: How long they last; how many interviewers they have; whether they are

panel/board interviews; and most importantly whether they are structured or not?

Do different interviewers agree? That is what is the interrater/observer/judge's reliability? The answer is about $r = 0.50$ and better as to their reliability and validity in *structured* interviews. Validity is however much lower. For structured interviews it is about the same for one-to-one versus board interviews ($r = 0.35$) but for unstructured interviews the one-to-one ($r = 0.11$) is about half as good as board interviews ($r = 0.21$). This means that interviewers of the same person do not agree very much on their assessments *and* that these assessments are not very useful predicting success (or failure) on the job. Only planned, structured interviews offer good data to really be useful in assessment.

Why are interviews so poor at predicting performance? There is a long and growing list of factors that collectively explain this. Interviewers differ in insight, skills, preferences. They differ in motives, attention, and need for justification of their decisions. Interviewees often try hard to manage a positive (not totally realistic) impression by self-promotion and self-enhancement. Interviewees are also increasingly being self-coached on how to behave in interviews. Put simply, interviewees (and interviewers and referees) tell lies of omission and commission. There are many and systematic variations in how interviewers use the rating scale.

There is also range restriction in ratings meaning that raters never use the full scale particularly the two ends. They do not discriminate/differentiate clearly enough between the different candidates. Too often interviewers make up their mind before the interview. Interviewers, like everyone else are susceptible to forming a first impression and ignoring later data. More importantly we know that reasons to reject (i.e., select out) factors have disproportionate weight compared to select-in factors. Finally many interviewers have their own (wrong, unproved, bizarre) implicit personality theories (i.e., red heads are intelligent, rugby players are good team workers).

Interviews can be improved to increase their reliability and validity by various relatively simple steps. These are very important to make the ever-popular interview a useful assessment method. Select, where possible good, insightful, natural interviewers. Always use more than one interviewer in panel or board interviews and use the same interviewers throughout the process. It is important to train interviewers in the fundamental skills of interviewing and ensure that they take notes to help memory and settle arguments.

Personal References

References are usually free response or ratings that an observer reports on another. That is they are observation by an individual known to the candidate. Though cheap and popular they are thought of as unreliable and of poor validity. There are three main reasons given for their poor reliability. First, *leniency*: Most references are indiscriminately positive. This is for three reasons. Candidates choose themselves those they know are most likely to be very positive; respondents worry negative evaluations may result in a libel suit; respondents have no incentive to take the time or tell the truth. Second, *idiosyncrasy:* People can and do use idiosyncratic language, examples and criteria to describe and evaluate others. Third, *free-form references*: Reference writers are often offered no guidelines in what they are required to do. They can offer long/short, descriptive/evaluative, and relevant/irrelevant data.

It is possible to improve the validity of the reference by explaining fully the purpose of the reference, using rating scales or a forces choice format. Using well-designed rating scales also helps a great deal.

The data also suggest that *peer* ratings are the most useful, valid, and reliable. Referees are best when the employer chooses the referees; those referees are peers who know the candidate well; the referees are asked specific questions and are guaranteed anonymity. Switching to the Internet has been argued to provide quicker, cheaper selection and assessment with wider access for many people. It saves time and travel costs, even leading to the possibility of 'same day offers.'

Organizations thus now make job offers (recruitment advertisements) on their Web site. Electronic application forms can be used to collect data and do a simple first-filter. Those without certain qualifications may quickly be rejected; software can be written to do a matching task between answers to questions and the 'ideal' profile. However, still not everyone has access to the Internet. There are age, educational, ethnicity, income, and gender correlates of those with and without access. This has legal implications. Electronic media do not bypass paper-based problems like faking (social desirability) both impression management as well as omissions and commissions. Also, sifting is fast but many still only look for keywords, thus it may be no more reliable and accurate.

Overall the new assessment technologies (predominantly the Web) have specific goals: improve efficiency, enable new screening tools, reduce costs, standardize the HR system, expand the applicant pool, promote the organizational image, increase applicant convenience.

However there are also unintended effects. The use of the Internet does expand the applicant pool but also increases the number of underqualified and out-of-country applicants. It is easy to be *flooded with inappropriate applicants*. There is also the *loss of personal touch* that both assessor and assessee value and respect. There are also concerns about *cheating* if tests are used. Finally there are still concerns about *adverse impact* for organizations when certain groups simply do not have access to the technology.

HR technology remains a challenge. The hopes of individuals have been both very high and not all the experiences have been positive. But it seems to be the future. Internet advertising and recruitment seems very cost-effective. Young people expect it.

Biodata

This is a method of scoring biographical factors. It makes the assumption that the past predicts the future and that certain life experiences at and before work have predictive validity. Biodata analysis is nearly always very specific. Instruments are devised for, and scored in, organizations and they have to be regularly updated.

It has many obvious advantages. First, *objectivity:* The same questions are asked of everyone who completes the form, and the answers given are assessed in a consistent way. Second, *cost*: Although there may be fairly extensive R&D

costs, once criteria are known, biodata are very cheap. Biodata forms can be developed in multiple-choice formats, which are amenable to machine scoring or direct entry to a computer terminal. Third, *checkability*: It is possible, in theory at any rate, to check up on the answers of the respondents. This fact may reduce faking. Fourth, *validity*: Nearly all the studies have shown that biodata provides a valid and reliable way of selecting individuals. Questionnaires, are as good as ability tests. Fifth, *self-presentation*: Most people prefer, and expect, to be asked about their biographies rather than their personalities. For many, biodata is more preferable and acceptable than personality tests.

There are also well-known disadvantages to using this method. First, there is the issue of *homogeneity versus heterogeneity*: If many biographical items are used in selection, the organization inevitably becomes more homogeneous, which has both advantages and disadvantages. Heterogeneity may occur across divisions (with different criteria) but not within them. Second, there is the problem of *cloning the past*: Biodata work on the idea that past behavior predicts current performance, but if current criteria are very unstable (say in a rapidly changing market) one is perpetually out of date. Biodata may be best in stable organizations and environments. Third, there is the ever common problem of *faking*: Biodata has been shown to be fakeable. Checking that a person has not faked can be difficult and extremely expensive. Next, *fairness in the law*: If biodata items show major biographical correlates such as sex, race, religion, and age, one may want to select or reject particular individuals, which is illegal. Items such as age, sex, and marital status may in fact be challenged by the courts if such items are included in inventories for the purpose of personnel selection. The fifth issue is that possibly *minorities cannot easily be identified and treated fairly*: In setting up a biodata form, the link between the information on the form and suitability for the job needs to be established. This in itself tends to ensure that any other elements are excluded from the selection process. Biodata forms, especially when computer-scored, are completely blind to incidental items such as a person's name, which might indicate ethnic background.

Next, there is the critique that *biodata do not travel well*: The same criteria do not have the same predictability across jobs, organizations, countries, or time periods. Because criteria have to be established every time, the development of biodata can be expensive. Seventh, it should be recognized that this method is very *time consuming*: For example, it might take at least 12 months to obtain reliable, meaningful job-related data on new employees. If an organization does not have a regular intake of new staff, there could be 2 or 3 years delay between sending the draft biodata form to applicants and obtaining a sample of employees large enough to warrant further development work. Last, there is the issue of *shrinkage over time*. Biodata scoring keys do not appear to hold up indefinitely. There is evidence that the validity of biodata shrinks over time, and periodic revalidation and reweighting may be necessary.

Biodata has never become very popular nor has it ever really disappeared. There are probably three reasons why it remains a relatively little used assessment method. It takes investment of time and expertise to devise a biodata measure. Candidates do not like biodata and may think it is unfair. Organizations worry about litigation. However it is used as part data collection for many Assessment Centers. And of course, simply employing application forms to decide who to, and who not to, interview is of course, represents a use of biodata.

Cognitive and Mental Ability Tests

This refers to achievement aptitude or mental ability tests. Many terms cover the same areas: ability achievement, cognitive ability, mental ability, and intelligence tests. These are distinct from social intelligence tests, creativity, or divergent thinking tests.

Academic research has shown that quite consistently, cognitive ability accurately predicts job performance across *all* jobs but particularly in *complex* jobs. Many believe that intelligence is the *single best* predictor of (senior, managerial) work performance. All recent research points to the predictive power of cognitive ability and hence the importance of using these tests in selection.

The publication of a recent highly controversial book on intelligence – *The Bell Curve* by Hernstein et al. – led over 50 of the world's experts on intelligence to write to the *Wall Street Journal* on 15 December 1994. Their summary is an excellent and clear statement on what psychologists think about intelligence. However they do also note five points of what they call practical importance. First, IQ is strongly related, probably more so than any other single measurable human trait, to many important educational, occupational, economic, and social outcomes. Its relation to the welfare and performance of individuals is very strong in some arenas in life (education, military training), moderate but robust in others (social competence), and modest but consistent in others (law-abidingness). Whatever IQ tests measure, it is of great practical importance.

Second, a high IQ is an advantage in life because virtually all activities require some reasoning and decision-making. Conversely, a low IQ is often a disadvantage, especially in disorganized environments. However, a high IQ no more guarantees success than a low IQ guarantees failure yet the odds for success in our society greatly favor individuals with higher IQs. Third, the practical advantages of having a higher IQ increase as life settings become more complex (novel, ambiguous, changing, unpredictable, or multifaceted). For example, a high IQ is generally necessary to perform well in highly complex professional jobs; it is a considerable advantage in moderately complex jobs (crafts, clerical, and police work); but it provides less advantage in settings that require only routine decision-making or simple problem solving (unskilled work).

Fourth, differences in intelligence are certainly not the only factor affecting performance in education, training, and highly complex jobs, but intelligence is often the most important. When individuals have already been selected for high (or low) intelligence – and so do not differ much in IQ – as in graduate school (or special education), other influences on performance loom larger in comparison. Fifth, certain personality traits, special talents, aptitudes, physical capabilities, experience, and the like are important (sometimes essential) for successful performance in many jobs, but they have narrower (or unknown) applicability or 'transferability' across tasks and settings compared with general intelligence. Some scholars choose to refer to these other human traits as 'intelligences.'

The recent research evidence is very clear. The single best predictor of success in complex, changing managerial jobs is intelligence. Brighter people learn faster; they have a greater store of knowledge; and they tend to be intellectually more self-confident. They analyze problems more efficiently and are less threatened by change. One can, of course, be too bright for the job if it is unskilled or unchallenging. However, all mental ability tests are the focus of much controversy.

Mental ability tests predict work performance fairly well – as well as any other method available. The relationship between mental ability and work performance is continuous and linear – the brighter someone is, the more likely their work performance is to be good. Research on mental ability and team (rather than individual) performance suggests the existence of more complex relationships. Low scorers may hold the whole team back. Research on why mental ability tests predict work performance is less well developed, but suggests that mental ability leads to improved job knowledge, which in turn leads to better work performance. Mental ability tests create a large adverse impact on some ethnic minorities in the United States. There is a dearth of information on this issue elsewhere. Attempts to solve the adverse impact problem in the US include trying to identify features of mental ability testing that will reduce adverse impact. Attempts to deal with the adverse impact problem in the US include score banding, which defines a range of scores as equivalent, thus allowing selection within the band to be based on achieving diversity.

Personality Tests

The bottom line in this complex and very hotly debated area is this. The validity of personality tests in predicting future work performance depends on the test, the criteria, and the population sample. The tests are at their most useful/accurate when a good, well-validated test is used and the criteria are clear, well measured and theoretically related to the trait is measured.

As a result of some hundred years of research and disagreement researchers have (for the most part) agreed on what is the underlying structure (i.e., personality, periodic table) of traits. The Big Five factor model is widely accepted, but may not give the best prediction of work behavior, though other models exist. Recent studies have shown that two of these are relevant to all jobs, namely trait neuroticism and conscientiousness. Successful managers have a clear profile: they tend to be low on neuroticism (i.e., very stable), low on agreeableness, average on openness, high on extraversion, and very high on conscientiousness. Different jobs require different profiles but there are none where high neuroticism and low conscientiousness are an advantage.

Personality is a vaguer and more diffuse concept than ability, and there are many models of personality. Personality is most conveniently assessed by inventory or questionnaire. It can also be assessed by interview, Assessment Center, references, or peer ratings. Personality inventories are compiled using a variety of methods which generally overlap to some extent, including internal analysis by statistics, and external analysis comparing groups of people with known characteristics.

Personality inventories have limited value in predicting how well a person can do a job. Personality inventories may be more successful in predicting effort and drive. Personality inventories may also be more successful in predicting avoidance of deviant or problematic behavior at work. Personality inventories are self-reports and not tests, and can be faked. Faking appears to affect the outcome of selection. Personality tests are also widely used to screen out people with problem characteristics, such as violent tendencies, deviant sexual impulses, or dishonesty. Screening tests for employee honesty are surprisingly successful. Gender differences are often found with personality tests, but ethnicity differences do not seem to be a problem. Personality inventories have also been criticized on the grounds that some of their questions are unduly intrusive.

Among the main advantages of personality testing are first, that tests provide *numeric information*, which means individuals can more easily be compared on the same criteria. In interviews, different questions are asked of different candidates, and the answers are often forgotten or not properly analyzed. Next, tests give *explicit and specific results* on temperament and ability rather than the ambiguous coded platitudes that are often found in references. A percentage (provided of course that it is valid) makes for much clearer thinking about personal characteristics than terms such as 'satisfactory,' 'sufficient,' or 'highflier.' Third, it could be argued that tests are *fair* because they eliminate corruption, favoritism, old-boy, mason – or Oxbridge – networks from self-perpetuating. That is, if a person does not have the ability, or has a 'dangerous' profile, they will not be chosen, irrespective of their 'other assets.' Next, it is often asserted that tests are *comprehensive* in that they cover all the basic dimensions of personality and ability from which other behavior patterns derive. Finally psychometricians assert that tests are *scientific* in that they are empirically based on theoretical foundations; that is, they are reliable, valid, and able to discriminate the good from the mediocre and the average from the bad.

Many, often related, objections are made to psychological testing. Among the most common and sensible objections include the idea that many of these tests can be *faked*; that is, people like to put themselves in a good light and receive a good score so that they may be accepted, but this in a way reflects their real personality (some tests have lie scores to attempt to overcome this). Tests are unreliable in that all sorts of temporary factors, such as test anxiety, boredom, weariness, a headache, period pains, all lead people to give different answers on different occasions (although this is partly true, it is only a minor factor).

Most importantly, some people claim that most, if not all tests are *invalid*; they do not measure what they say they are measuring and those scores do not predict behavior over time. For many tests, this is indeed the Achilles heel and they are often lamentably short of robust proof of their validity. It has also been asserted that while tests might be able to measure all sorts of dimensions of behavior, but *not the crucial ones* to the organization such as trustworthiness and likelihood of absenteeism. Buying personality tests is like having a set menu, and what many managers want is an *a la carte* menu where they can select only factors or dimensions that they want.

People have to be *sufficiently literate* or articulate to do these tests, not to mention sufficiently familiar with North American jargon. Many organizations therefore believe their workforce

could not complete them properly, would take up too much time or would cause needless embarrassment. Others claim that there are not *good norms* (normative population data), at least for the population they want to test, and comparing them to White US students can be dangerously misleading. One old assertion is that *tests are unfair and biased* toward WASPS. Also, males tend to do better or get a more attractive profile and therefore get selected – which although ostensibly meritocratic, can conflict with antidiscriminatory legislation. It is also true that the *interpretation of the tests* takes skill, insight, and experience, and this is either too expensive or not available. In the wrong hands they are dangerous, because profiles are either given inaccurate or too literal interpretations. Finally *freedom of information legislation* may mean that candidates would be able to see, and hence, challenge either the scores themselves or their interpretation of the decisions made on them. The less objective recorded data the better. Finally, as tests become known, people could buy copies and practice, so that they know the correct or *most desirable answers*. This happens extensively with ability and personality testing and, when it does happen, results may have more to do with preparation and practice than actually *reflecting* true ability.

All assessment methods have advantages and disadvantages. Personality tests can greatly facilitate assessment. But the user needs to have some knowledge of test construction and theory both to choose tests wisely but how and when to implement them in the process. We do know that the two of the 'Big Five' factors seem to be important for all jobs – neuroticism which is a select-out factor because it is correlated negatively with success at work, and conscientiousness which is a select-in factor because it predicts work success. Test users would do wisely then to select tests which measure *both* these important traits.

But if tests do not predict behavior at work very well, the question is why not. First, there may be problems with the theoretical formulation in the first place, in that the personality theory or system has been poorly or inaccurately conceived. Quite simply, the theory upon which the test is based is wrong. Next, the measurement of the personality trait is poor for a variety of reasons, for example, it measures a single trait which is actually multidimensional. Also the measure has poor reliability and validity.

See also: Intellectual Functioning, Assessment of; Intelligence: Assessments of; Intelligence: Central Conceptions and Psychometric Models; Personality Assessment, Faking and; Personality Assessment; Psychometrics.

Bibliography

Alimo-Mettcalfe, B., Alban-Metcalfe, R., 2001. Development of the TLQ. Journal of Occupational and Organisational Psychologists 74, 1–24.

Anderson, N., Cunninham-Snell, N., 2000. Personnel selection. In: Chmiel, N. (Ed.), Introduction to Workand Organizational Psychology. Blackwell, Oxford, pp. 69–99.

Anderson, N., Lievens, F., van Dam, K., Ryan, A., 2004. Future perspectives on employee selection: key directions for future research and practice. Applied Psychology 2004 (53), 487–501.

Arnold, J., et al., 2005. Work Psychology: Understanding Human Behaviour in the Workplace. Prentice-Hall, Harlow.

Bartram, D., 2004. Assessment in organisations. Applied Psychology 53, 237–259.

Barrett, P., 2003. Beyond Psychometrics: measurement, non-qualitative structive and applied numerics. Journal of Managerial Psychology 18, 421–439.

Bray, D., 1982. The assessment centre and the study of lives. American Psychologist 19, 419–429.

Brown, R., 1999. The use of personality tests: a survey of usage and practice in the UK. Selection and Development Review 15, 3–8.

Bywater, J., Thompson, D., 2005. Personality questionnaires in a redundancy/restructuring setting – what do we know now? Selection and Development Review 21, 7–13.

Chamorro-Premuzic, T., Furnham, A., 2010. The Psychology or Personnel Assessment. Cambridge University Press, Cambridge.

Chamorro-Premuzic, T., von Stumm, S., Furnham, A., 2011. The Wiley-Blackwell Handbook of Individual Differences. Wiley-Blackwell, Oxford.

Cohen, R., Montague, P., Nathanson, L., Swerdlek, M., 1988. Psychological Testing. Manfield, Mountain View.

Cook, M., 2008. Personnel Selection: Adding Value through People. Wiley, Chichester.

Cox, L., Cooper, C., 1988. High Flyers. Oxford, Blackwell.

Chapman, D., Webster, J., 2003. The use of technologies in the recruiting, screening, and selection process for job candidates. International Journal of Selection and Assessment. 11, 113–120.

Doyle, C., 2003. Work and Organisational Psychology. Psychology Press, Hove.

Dulewicz, V., Higgs, M., Slaski, M., 2003. Measuring emotional intelligence. Journal of Managerial Psychology 18, 405–420.

Dulewicz, V., Higgs, M., 2004. Design of a new instrument to assess leadership dimensions and styles. Selection and Development Review 20, 7–12.

Fletcher, C., Bariley, C., 2003. Assessing self awareness. Journal of Managerial Psychology 18, 395–404.

Furnham, A., 2005. Gender and personality differences in self – and other ratings in business intelligence. British Journal of Management 16, 91–103.

Furnham, A., Taylor, J., 2004. The Dark Side of Behaviour at Work. Palgrare MacMillan, Basingstoke.

Furnham, A., 2008a. Personality and Intelligence at Work. Routledge, London.

Furnham, A., 2008b. HR Professionals' beliefs about, and knowledge of, assessment techniques and psychometric tests. International Journal of Selection and Assessment 16, 301–306.

Furnham, A., Chamorro-Premuzic, T., 2010. Consensual beliefs about the fairness and accuracy of methods at university. International Journal of Selection and Assessment 18, 417–424.

Furnham, A., Jackson, C., 2010. Practitioner reactions to work-related psychological tests. Journal of Managerial Psychology 26, 549–565.

Goodge, P., 2004. Capabilities versus competencies. Selection and Development Review 20, 7–10.

Greenberg, J., Baron, R., 2003. Behaviour in Organisations. Pearson Educational, New York.

Groth-Marnet, G., 1984. Handbook of Psychological Assessment. Van Nostrand, New York.

Hambleton, R., Oakland, T., 2004. Advances, issues and research intensity practises around the world. Applied Psychology 53, 155–156.

Hausknecht, J., Day, D., Thomas, S., 2004. Applicant reactions to selection procedures. Personnel Psychology 57, 639–683.

Hogan, H., Hogan, J., 2004. Assessing Leadership: a vision form the dark side. Selection and Development Review 20, 3–15.

Hogan, R., Hogan, J., Trickey, G., 1999. Goodbye mumbo-jumbo: the transcendental beauty of a validly coefficient. Selection and Development Review 15, 3–16.

Jeanneret, R., Silzer, R., 2000. An overview of individual psychological assessment. In: Jeanreret, R., Silzer, R. (Eds.), Individual Psychological Assessment. Jossey Bass, San Francisco, pp. 3–26.

Klehe, U.-C., 2004. Choosing how to choose: Institutional pressures affecting the adoption of personnel selection procedures. International Journal of Selection and Assessment 12, 327–342.

Kwaitkowski, R., 2003. Trends in organisations and selection. Journal of Managerial Psychology 18, 382–394.

Lievens, F., van Dam, K., Anderson, N., 2002. Recent trends in challenges in personnel selection. Personnel Review. 31, 580–601.

Lievens, F., DeCorte, W., Brysse, K., 2003. Applicant perceptions of selection procedures. International Journal of Selection and Assessment 11, 67–71.

McCall, M., 1998. High Flyers. Harvard University Press, Harvard.

Oakland, T., 2004. Use of educational and psychological tests internationally. Applied Psychology 53, 157–172.

Ones, D., Andreson, N., 2002. Gender and ethnic group differences on personality scales in selection. Journal of Occupational and Organisational Psychology 75, 255–276.

Ones, D., Viswuaran, C., 1998. Integrity testing in organisations. In: Griffin, R., O'Leary-Kelly, A., Collins, J. (Eds.), Dysfunctional Behaviour in Organisations. Al Press, Greenwich, CT.

Potosky, D., Bobko, P., 2004. Selection testing via the internet. Personnel Psychology 57, 1003–1034.

Ryan, A.M., Sackett, P., 1988. Individual assessment: the research base. In: Jeanreret, R., Sulzer, R. (Eds.), Individual Psychological Assessment. Jessey Bass, San Francisco, pp. 54–87.

Salgado, J., 2001. Some landmarks of 100 years of scientific personal selection at the beginning of the new century. International Journal of Selection and Assessment 9, 3–8.

Silzer, R., Jeanneret, R., 2000. Anticipating the future; assessment strategies for tomorrow. In: Jeanreret, R., Silzer, R. (Eds.), Individual Psychology Assessment. Jessy Bass, San Francisco, pp. 445–477.

Smith, M., Robertson, I., 1989. Advances in Selection and Assessment. John Wiley, Chichester.

Spector, P., 2006. Industrial and Organisational Psychology: Research and Practice. John Wiley.

Spreitzer, G., McCall, M., Mahoney, J., 1997a. Early identification international executive potential. Journal of Applied Psychology 82, 6–29.

Spreitzer, G., McCall, M., Mahoney, J., 1997b. Early identification of international executive potential. Journal of Applied Psychology 82, 6–29.

Te Nijenhuis, J., Voskuijl, O., Schijive, N., 2001. Practice on coaching on IQ tests. International Journal of Selection and Assessment 9, 302–306.

Truxillo, D., Steiner, D., Gilliland, S., 2004. The importance of organisational justice in personnel. International Journal of Selection and Assessment 12, 39–53.

van de Vijver, F., Phalet, K., 2004. Assessment in multicultural groups. Applied Psychology 53, 215–236.

Wareing, B., Fletcher, C., 2004. Ethnic minority differences in self-assessment. Selection and Development Review 20, 7–10.

Whiddett, S., 2004. Whose job is it anyway? Selection and Development Review 20, 7–10.

Woodruffe, C., 1995. Assessment Centres: Identifying and Developing Competencies. IPD, London.

Young, D., Hamilton, C., Kirk, A., 2004. Web-based coaching – lessons from a practical experience. Selection and Development Review 20, 18–23.

Personality, Biological Models of

Philip J Corr, City University London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Personality theories attempt to describe systematic differences between people in affect, cognition, and behavior, across situations and over time. Some theories attempt also to explain the causal roots of these individual differences, and biologically oriented ones focus on genetic and neurophysiological processes that underlie fundamental emotion and motivation processes. Approach-avoidance personality theories have dominated this literature: These are summarized along with the relevance of genetic and neuroimaging research methodologies. Conceptual issues in personality psychology are discussed to show how biologically oriented theories relate to other theoretical perspectives (e.g., social and psychometric).

Theories of personality aim to describe systematic differences between people in affect, cognition, and behavior, across situations and over time. A second major aim is to account for the causal roots of these individual differences, either in social, psychological, or neurological terms. Biological models of personality are specifically concerned with explaining the genetic and neurophysiological bases of these differences.

The phenomena of interest to the personality psychologist include both person and environmental factors, and their interactions. For this reason, there is a need to recognize multiple levels of analyses. Therefore, in a strict scientific sense, there cannot exist purely 'biological' or 'social' theories of personality. What 'prefix' theories of personality can claim to achieve is to emphasize the relative role played by a specific level of explanation. However, in order for any such theory to relate to personality factors (descriptions) and processes (causes), it cannot escape the fact that it is part of a broader conceptual framework. For example, the vast majority of personality theories start with a statistical description of systematic differences between people, typically relying upon psychometric principles and questionnaires (or some other form of self/other report). Data from these forms of measurement are neither purely 'biological' nor 'social' but entail 'method variance.' From a specific theoretical perspective, although such variance may be considered 'nuisance,' it is nonetheless real and reflects the multitude of influences on all forms of personality description. In addition, self/other ratings are about something (e.g., openness to experience of *what*?): From the respondent's point of view, test items are imbued with meaning. These observations do not undermine the attempt to construct a scientific theory of personality, but they do complicate matters and render single-level explanations problematic in terms of offering a comprehensive description and explanation of personality.

Once these complexities are recognized, it is possible to develop a 'social' theory of personality that emphasizes the relative effects of environmental factors, although such theories encounter significant problems in accounting for the full range of variation observed between people when tested in identical environments. In the case of a 'biological' theory of personality, it is simply not possible to ignore the role of the environment and broader social context: Biological effects are shaped by and expressed through the environment. In this strict sense, there cannot exist a purely 'biological' model of personality.

Far from being a counsel of despair, the above considerations highlight the important issues that need taking into account in understanding the biological approach to personality.

Surface Traits and Source Processes

In any personality theory, especially biological ones that lay stress upon underlying causal factors, there is an important distinction between two major levels of explanation: Surface traits and source processes.

In everyday life, it is typical to describe people's personalities in lexical terms. For example, we may describe John's characteristic ways of thinking, feeling, and behaving as 'smart,' 'fearful,' and 'cautious,' and these adjectives are readily understood and serve as a convenient means of communication, especially when they have an evaluative component (e.g., 'lazy,' 'ambitious,' 'aggressive'). We also talk about people's expected future behavior in the same manner; for example, John is 'trustworthy' and a 'regular' guy, and this allows us to predict the likely reaction of John to a range of situations.

Research personality psychologists have adopted essentially the same approach to describing personality, although their descriptions are couched in statistical terms. There is a long tradition of analyzing the ratings of a large number of trait adjectives to produce descriptive taxonomies, the most popular being the Big Five: Extraversion, Neuroticism, Conscientiousness, Agreeableness, and Openness to Experience/Intellect. Meta-analyses agree that these Big Five dimensions emerge from many existing personality questionnaires not designed specifically to measure them (e.g., Markon et al., 2005). These five dimensions describe systematic individual differences at the population level, and they are typically called 'traits' (or more generally, 'factors').

As a 'surface' level description of variation at the population level, the Big Five does a good (but not uncontested) job. But it should be obvious that statistical description at this high level of abstraction is bound to be influenced by a broad range of factors, ranging from biological (DNA) to societal. There are also likely to be significant cross-cultural differences in the formation and expression of these traits.

What the Big Five and other population-wide statistical descriptions of personality do not readily provide is information regarding the causal 'sources' of these surface traits: Scientific

explanation is about explaining, not just naming. What we observe at this surface level of description is the end product of a chain of interacting processes, and we can have little confidence in the assumption that these interactions are not complex. Since at least the time of Sir Francis Bacon, it is appreciated in science that it is difficult, sometimes seemingly impossible, to infer causes from observed effects. This is a major problem in personality psychology. To cut this Gordian knot, it is necessary to propose causal theories of underlying processes. Biological theories have been central in this scientific enterprise.

One important question that cries out for an answer is why do people feel, think, and act in the ways they do? One model is to assume that the environment (or situation) determines behavior, and that a history of reinforcement shapes individual propensities to respond to situations in specific ways. This behaviorist account has fallen out of favor for several reasons. First, it is tautological and circular in argument, and for a scientific theory it is remarkably immune to refutation. Second, there is now little doubt that the traits of personality are under heavy genetic influence (typically, heritability hovers somewhere between 40 and 60%). Third, even if the behaviorist account had merit, we would still have a legitimate right to want to know how observed effects are instantiated and processed by the brain. But these conclusions are not to discount the value of either a learning account or a reinforcement-based one. As shown below, both of these accounts have proved fruitful in the biological approach to personality.

Biosocial Perspective

If it is not possible to have a pure theory, what can be labeled 'biological' in personality psychology? For a start, such a theory would need to account for endocrine and physiological processes in explaining individual differences in response to the same environment. In this way, it lays emphasis upon internal agency, and not merely reactivity to external stimuli. We witness these features most clearly articulated in the first major biological approach to personality, namely H.J. Eysenck's arousal-activation (1967) theory.

The seeds of this major biological theory of personality were planted in Eysenck's 1944 seminal factor analysis of a medical checklist of 700 'war neurotics' invalidated out the British Army due to a failure of adaptation to the environment (in this case, military training during World War II). It will repay the student of personality psychology to consider the outlines of this theoretical approach as it reveals the assumptions of all biological theories of personality that took inspiration from it – and most did.

Eysenck's general approach was not so much 'biological' as 'biosocial.' It explicitly recognized that personality traits are the end product of the interaction of biological and social factors. In common with other biological theories of personality, Eysenck's theory points to the crucial role played by genetics and physiology, but he also pointed to the role of the *interplay* of genes/brain and the environment. Eysenck's starting point was to assume that human behavior and variations between people (i.e., 'personality') can be measured and classified in much the same manner as other biological phenomena – to him, there was nothing special about the 'mind,' no *élan vital*. His was a monist view of the mind/body (brain) problem.

A crucial assumption is that to the extent that environmental influences have lasting effects, they *must* go through the brain. By simple inference, individual differences in brain structure and functioning (heavily genetically influenced) affect the perception and analysis of the 'environment' and, thus, to a large extent, both define and determine its influence. For example, it is known that some people see events, as it were, through rose-tinted glasses, while for other people the hue of their glasses is blue. That is, given the same (objectively defined) event (the 'situation') there are different psychological consequences (e.g., emotional experience) depending on pre-existing dispositions (i.e., personality). At university, it is common for one student to perceive an examination as an opportunity to prove themselves, while for another student it is something to be dreaded – and these different psychological perspectives lead to different outcomes (emotional, cognitive, and behavioral, and in this example academic attainment). Contrary to purely social theories of personality, the environmental event (e.g., university examination) does not determine the psychological outcome; instead, biologically based individual differences in the construction and perception of the environment are the important ingredient. Today, this brain–environment interplay is a very fashionable topic of psychological research, seen in the guise of such effects as 'gene–environments interactions' (see below).

In Eysenck's personality theory, the conceptual bridge between the brain and the environment was conditioning (or more broadly speaking, learning). As there are significant individual differences in brain processes, so too there are differences between people in their rates of learning. This basic premise underlies many biological personality theories, which tend to focus on the role played by basic emotion and motivation systems, and how they lead to different forms of learning (i.e., modification of behavior by experience).

The biological approach also assumes that genetics provide a blueprint for the development of a specific phenotype (e.g., extraversion) but, critically, it is the interplay with environmental factors that determines what is observed – in much the same way that there must be a blueprint (or 'preparedness') for language acquisition but we all learn specific languages with local dialects.

Important questions are raised by this literature. How do biological processes affect perception and reactions to the environment? How do environmental factors affect biological processes? In other words, 'biological' and 'social' are bidirectional influences. Put another way, although nothing is more apparently 'biological' than DNA, this is merely a blueprint for potential development of the brain, which is proximally responsible for thoughts, feelings, and behavior. The brain is, indeed, a 'necessary' condition, but psychological phenotypes need 'sufficient' environmental conditions in order for them to develop and to be expressed.

A Functional Perspective

The search for the biological bases of personality traits must ask a number of questions, most important of which are What processes, and where to look for them? This problem is more real than apparent, and its resolution requires a conceptual

model of personality processes. The famous neuropsychologist Donald Hebb (1955) made the vital distinction between a 'conceptual nervous system' and the real 'central nervous system.' This distinction has proved important in suggesting how the scientific investigation of personality psychology might best proceed.

In order to understand why people react in different ways to the same stimulus or situation, we must first discover what drives people's actions and reactions. In the past, inferring motivation from observed behavior (and consistencies in behavior, 'personality') was something of a dark (and often murky) art. Psychoanalysis and a variety of 'projective' approaches may have enlivened psychology, but to many people they failed to throw proper scientific light on the true roots of motivation. Personality psychologists have looked elsewhere for inspiration.

One approach that has yielded theoretical fruit is based on the biology of motivational control systems. This approach operates on the premise that stable individual differences in behavior (personality traits) must be due to relatively stable individual differences in the operation of brain systems that produce (state) behavior from moment-to-moment. From this perspective, population-level traits reflect the operations of brain systems that have evolved to respond to different classes of functional requirements (Denissen and Penke, 2008). Accordingly, the search for biological underpinnings of personality should entail the search for answers to two questions: (1) What are the functions of specific reactions, and systems of reactions? and then (2) What is the neurophysiology of coordinated systems that mediate these reactions?

This functional approach has proved informative, and it has been augmented by two other approaches: The correlational approach (which biological processes are empirically associated with traits?); and the instantiation approach (what are the structures and workings of neurophysiological processes?). Most biological theories of personality combine different elements of these three approaches.

These approaches also help to differentiate the 'conceptual' (e.g., identifying the existence of causal systems by the construction of theories of behavior and learning) and 'real' (the 'wetware' of brain processes) nervous systems. In order to search for brain processes underlying the major systems of personality, it is necessary first to identify the existence of such systems by extracting separable factors of personality (traits) and building behavior-based accounts of such systems (e.g., classical and instrumental learning, and reward and punishment systems). One of the most comprehensive accounts of this whole literature is provided by Zuckerman (2005).

Functional accounts of personality have gained widespread popularity in recent years, and have afforded important insights in to human behavior. As discussed below, this approach provides a taxonomy of behaviors not in terms of the naming of traits and not even their specific behavioral outputs but, rather, in terms of the functions they serve. For example, different behavioral strategies (e.g., cooperation or coercion) may be identical in function (e.g., social influence) but their expression is influenced by context (e.g., power relationship). Contrariwise, the same behavior (e.g., cooperation) may serve different functions (e.g., altruism or submission), again depending on the context. Functions cannot simply be 'read-off' from behaviors; there is need for some form of conceptual nervous system to guide the search for the main systems of the real nervous system.

The functional perspective also throws light on the adaptive nature of, so-called, pathological behaviors and traits (e.g., anxiety and paranoia). The basic assumption is that personality traits – to the extent that they are systematic, genetically influenced, and pervade daily life – evolved by Darwinian natural selection, and thus have adaptive value – but this may not always be obvious. Adopting Darwin's hypothesis that psychological attributes are shaped by natural selection in much the same way as anatomical characteristics, modern personality theorists have come to favor accounts couched in terms of the functions of central states of emotion and motivation – an area of study suggested by Darwin himself in his work on the expressions of emotion in man and animals. These 'approach-avoidance theories' aim to describe the major systems that motivate behavior in reaction to classes of appetitive (rewarding) and aversive (punishing) stimuli, and to explain consistent patterns of individual differences in these behaviors. This literature is summarized by Corr (2013).

Approach-Avoidance Theories of Personality

Approach-avoidance theories of personality are based on the assumption that important classes of motivational stimuli can be grouped into 'rewards' and 'punishments.' Animals can be seen as cybernetic systems with attractors and repulsors (positive and negative goals) that have evolved to promote survival and reproduction. Without a tendency to approach beneficial stimuli (e.g., food, drink, and sexual mates) and to avoid aversive stimuli (e.g., predators and poisons) individuals and a whole species would simply not survive and pass on their DNA blueprint – the only thing that matters in biological inheritance.

Current approach-avoidance personality theories can trace their origins to early behavioral-learning researchers (e.g., Mowrer, 1960), who posited that two motivation/emotion processes underlie behavior: One related to reward (approach behavior and positive emotions), the other to punishment (avoidance behavior and negative emotions). This animal-based work migrated into personality psychology in the 1970s through the work of Jeffrey A. Gray (e.g., 1975), who argued that the major traits of personality reflect long-term stabilities in systems that mediate reactions to different classes of reinforcing stimuli, generating emotion and shaping ('motivating') approach and avoidance behavior. Neuroscience measures, including pharmacological manipulation, recording of neural activity, and neuroanatomical studies have been used to investigate the neuropsychological systems that underlie these reactions. A wealth of empirical research provides confirmation of the hypothesis that distinct and separable systems in the brain mediate reward and punishment-related emotion, motivation, and learning (Gray and McNaughton, 2000).

Personality Traits as Probabilistic Constants

To be of scientific value, a personality trait must be more than a score on a questionnaire: It must have predictive validity. In

biological theories, a personality trait is defined as a probabilistic constant in equations that predict the frequency and intensity with which individuals exhibit various motivational states, as well as the behavioral, emotional, and cognitive states that accompany these motivational states (Fleeson and Gallagher, 2009).

The biological approach aims to account for both the neuropsychological systems that are responsible for the states associated with any given trait and the parameters of those systems that cause them to differ across individuals. The systems themselves will be present in every intact human brain, but the values of their parameters will vary from person-to-person. Thus, for example, all people have brain systems that respond to punishments, but across individuals these systems respond differently to a given stimulus. It is the *typical* level of response of such a system in any given individual, averaged across different situations, that is associated with that individual's score on the personality trait in question. This is *not* to imply that an individual will respond the same way in all situations; rather, it implies that knowing the strength of the individual's trait predicts how he or she is likely to respond in a certain situation and, in particular, predicts variation in such responding across a set of individuals experiencing the same situation. This perspective means that the same person may respond quite differently in different situations (e.g., introverts under low and high level of arousal; Eysenck, 1967).

Many personality researchers have embraced this basic notion, and a number of personality models postulate traits reflecting sensitivity to reward and punishment (e.g., Elliot and Thrash, 2002). With the upsurge of neuroscience in psychology, there has been a marked increase of such theories, sometimes incorporated as part of larger theoretical schemes, associated with names such as Charles Carver, Robert Cloninger, Richard Davidson, and Jaak Panksepp. DeYoung and Gray (2009) summarize this large literature.

Reinforcement Sensitivity Theory of Personality

As Jeffrey Gray's early research was the principal impetus for reward/approach and punishment/avoidance concepts in mainstream personality psychology, and his work led to the formulation of the most elaborated biological theory of personality, it would be appropriate to use the current version of his theory, the reinforcement sensitivity theory (RST), to illustrate this broader scientific landscape.

Gray identified three primary systems that control reactions to punishment, reward, and conflict: The *fight-flight-freeze system* (FFFS), the *behavioral approach system* (BAS), and the *behavioral inhibition system* (BIS) (Gray, 1982; Gray and McNaughton, 2000; for summary, see McNaughton and Corr, 2008).

The one positive system, the BAS, is responsible for mediating reactions to all rewarding stimuli (which includes relief from nonpunishment) and it activates exploratory approach behavior. It is related to the emotion of hope and anticipatory pleasure and is measured by various personality scales, described below.

Turning to the defensive systems, the FFFS is responsible for mediating reactions to all punishing stimuli (which includes omission/termination of expected reward, and frustration), and is involved in active avoidance and escape behavior. It is related to the emotion of fear. In contrast, the BIS is activated by stimuli that indicate conflict between goals, especially between FFFS- and BAS-related ones. Upon activation, it inhibits the actions of the BAS and FFFS and generates cautious approach in potentially dangerous situations (i.e., passive avoidance). Also once activated, it generates risk assessment behavior, rumination, and increased arousal. Of interest, the BIS can be activated by the conflict of two opposing approach–approach goals, for example, receiving two equally appealing graduate school offers (the aversive component resides in the potential of making the wrong decision and incurring a relative loss). The BIS is related to the emotion of anxiety.

RST and Personality Traits

The most recent version of RST (Corr and McNaughton, 2008) proposes that the level of the trait Neuroticism reflects sensitivity to punishment and threat in general (both FFFS and BIS). However, within Neuroticism, there are two traits/emotions, each of which maps on to one of the two major systems for defensive behavior: *Fear* and *trait fearfulness* arise from the functioning of the FFFS; and *anxiety* and *anxiety-proneness* arises from variations in the sensitivity of the BIS.

Research evidence confirms that the personality traits associated with FFFS and BIS sensitivity are differentiable in measurement terms. Psychometric measures of fear and anxiety have been distinguished through confirmatory factor analysis (CFA), predictive validity studies, and associated facial expressions. Although Carver and White's (1994) widely used BIS scale was developed with only one avoidance system in mind, recent CFA studies show that this 'BIS' scale can be divided into separate FFFS (fear) and BIS (anxiety) components. There are now also studies using antianxiety and antipanic drugs to probe the FFFS and BIS, and these too support the separation of these systems. These studies are described by Corr et al. (2013)

Turning to BAS-related approach trait, Gray (1982) originally speculated that the trait associated with BAS sensitivity was 'impulsivity' largely because impulsive people are more likely to be sensitive to cues of the immediate possibility of reward. Although BAS sensitivity does play a role in impulsivity, researchers have since concluded that impulsivity is not its purest manifestation because it is determined not only by individual differences in the strength of impulses to pursue immediate reward but also by individual differences in the ability of top-down control systems to restrain and control those impulses (DeYoung, 2010). Indeed, Extraversion rather than impulsivity appears to represent the primary manifestation of BAS sensitivity in personality and various measures of reward sensitivity fall within the broad Extraversion factor (e.g., Gable et al., 2003).

Breaking down reward sensitivity into subfactors has not been as systematic as the approach to identifying traits associated with BIS and FFFS, largely because Gray elaborated only a single reward system. However, the most commonly used measure of BAS sensitivity has three subscales: Drive, Reward Responsivity, and Fun Seeking (Carver and White, 1994). Whereas Drive and Reward Responsivity both appear to characterize sensitivity to reward primarily, Fun Seeking appears to be equally related to impulsivity, and thus may not be as pure

an indicator of BAS sensitivity. How this biological model of personality is related to the population-level factors identified by the Big Five research tradition is summarized by Corr et al. (2013).

Genetics of Personality

One of the major correlational approaches to understanding the biology of personality factors and their underlying processes is genetics, which is seen in two forms: (1) Statistical/behavioral, and (2) molecular. This literature is reviewed by Munafo (2009).

Statistical/Behavioral Genetics

Using twin and adoption designs, this approach attempts to decompose the variance in some phenotype (e.g., Neuroticism) into genetic and nongenetic (i.e., 'environmental') components. By this means, heritability estimates are calculated. The literature shows that for almost all personality factors there is a substantial heritability (40–60%). Increasingly elaborate designs are being used to tease part genetic and environmental influences (e.g., new families after divorce which often have children with different biological fathers).

Molecular Genetics

In the case of molecular genetics, the approach is to find statistically significant associations between the frequencies of specific genetic variants (i.e., alleles) and scores on personality questionnaires. There are two main techniques used. The first is 'candidate gene' association technique, which first identifies specific genes thought to be important in the causal processes.

Candidate gene studies take as their starting point what is already known about the neurobiology of the trait of interest. This is used to identify genetic 'candidates,' in other words, genes that encode products involved in relevant neurotransmitter pathways. For example, when studying anxiety-related traits, such as neuroticism, genes involved in the serotonin pathway are the likely candidate, while for approach-related traits, such as extraversion or novelty seeking, genes involved in the dopamine pathway are the focus. As well as identifying a candidate gene in this way, it is necessary to identify a polymorphism within this gene (i.e., a region that can exist in multiple forms, 'alleles'). Genetic variation at this locus should, therefore, confer biological individual differences, which in turn should result in behavioral (phenotype) differences between people (e.g., anxiety).

The second approach is 'genome-wide association studies,' which takes a blind atheoretical approach and scans the entire genome for 'hot spots' of association – often revealed by Manhattan plots which highlight 'tall' associations. A very large number (500 000+) of genetic markers are examined to determine if any are related to the phenotype of interest and, if so, to what extent. Then, once the associated genetic markers are reliably identified, the process of exploration of the function of the related genes can start in earnest. Once again, the aim is to look for a significant correlation between genetic variation and phenotypic variation. To achieve reliable results, very large sample sizes are needed as the genetic associations are typically very small (usually less than 1% of phenotypic variance).

Although it is well-established that heritability for personality traits is substantial, the search for specific genes has not yet proved successful. This result may be due to several factors: The very small effect sizes of individual genes; the interaction of genes and the environment; and a basic lack of knowledge concerning how genes confer their effects on brain development and the expression of phenotypes. Certainly, the candidate gene literature is characterized by a pattern of early excitement followed by disappointment (Ebstein, 2006) as results have failed to replicate on independent samples.

Another reason for this lack of success may be the possibility of interactions between genes. This possibility is suggested by the work of Reuter et al. (2006) who found a significant DRD2 by COMT interactions (i.e., epistasis) for the Carver and White (1994) total BAS scale as well as for the subscales Drive (D) and Fun Seeking (FS). The details of this interaction are much less interesting than its existence: Genes may work together in influencing specific phenotypes.

Gene-by-Environment Interactions

There are now new approaches to exploring just how genes and the environment work together. This work throws new light on the role played by genes, but also what is meant by 'the environment.' To illustrate this approach, Poulin et al. (2012) found that genes act in tandem with a person's perceptions of the world (e.g., whether it is perceived to be threatening or not) to predict their level of generosity. Individuals who believe their surroundings are threatening are less likely to help others. However, this was found only in those individuals who lacked the gene variant that enable them to overcome their negative feelings. Specifically, genes related to oxytocin and vasopressin (hormones known to be related to maternal and prosocial behaviors) modify the effects of perceived threat (a measure of the environment) to predict engagement in volunteer work, charitable activities, and commitment to civic duty.

Neuroimaging and Personality

The biological approach to personality was given great momentum by the availability of functional neuroimaging (principally, magnetic resonance imaging, MRI), which permits direct investigation of the brain processes associated with an experimentally controlled performance task – this functional MRI is distinct from structural MRI, which is valuable in its own right (and comes as a 'freebie' with fMRI). It is now possible to relate personality traits to these brain processes. Although this literature is still in its infancy, it is already making progress in showing the brain instantiation of personality processes. This literature is summarized by Canli (2009).

Recent fMRI studies have demonstrated remarkably powerful and expected associations between personality traits, measured by a simple questionnaire, and brain activity during a number of cognitive and affective tasks (e.g., Mobbs et al., 2005).

Before the availability of functional brain imaging the only way to measure brain activity was with the use of electroencephalograms and event-related potentials – these remain valuable methods and are still used, especially because of their high temporal resolution (in milliseconds). The other main technique used by early researchers of personality and brain

function was lesions sustained through accidents or disease (or experimentally induced in laboratory animals, typically rats). These more traditional techniques still play a role, albeit a relatively declining one, in biological personality psychology.

As a specific example of utility of fMRI in relation to RST, Cunningham et al. (2010) provide evidence for different associations between FFFS- and BIS-related traits on amygdala activity (which is a brain region crucially involved in the detection of motivational salience). Their pattern of findings is consistent with the idea that the FFFS responds to all punishing stimuli, whereas the BIS responds to conflict associated with concurrent approach tendencies.

Conclusion

Biological theories of personality have made significant progress, in both theoretical elaboration of conceptual systems as well as uncovering the genetic and neurophysiological bases of systematic individual differences in thinking, feeling, and behaving. This research is embedded in a much broader conceptual framework of personality description and social-environmental influences. Gene-by-environment interaction studies are proving especially insightful, including showing how 'the environment' is, in many respects, a psychological construction of the brain and, itself, influenced by genetic factors.

See also: Agreeableness; Behavioral Inhibition and Social Withdrawal across Cultures; Conscientiousness; Emotions, Psychological Structure of; Extraversion; Facial Emotion Expression, Individual Differences in; Five Factor Model of Personality, Assessment of; Imaging Methods in Clinical Psychology; Motivation in Clinical Interventions; Neuroticism; Openness to Experience; Panic Disorder and Agoraphobia Across the Lifespan; Pathological Gambling and Gambling Disorder; Personality Assessment; Personality Changes During Adolescence Across Cultures; Personality Differences and Development: Genetic and Environmental Contributions; Personality Theory and Psychopathology; Personality, Evolutionary Models of; Personality, Trait Models of; Phobias Across the Lifespan; Projective Methods in Psychology; Reinforcement, Principle of; Social Phobia across the Lifespan; Temperament Development, Theories of; Temperament and Human Development; Temperament.

Bibliography

Canli, T., 2009. Neuroimaging of personality. In: Corr, P.J., Matthews, G. (Eds.), The Cambridge Handbook of Personality Psychology. Cambridge University Press, Cambridge, pp. 305–322.

Carver, C.S., 2005. Impulse and constraint: perspectives from personality psychology, convergence with theory in other areas, and potential for integration. Personality and Social Psychology Review 9, 312–333.

Carver, C.S., White, T.L., 1994. Behavioral inhibition, behavioural activation, and affective responses to impending reward and punishment: the BIS/BAS scales. Journal of Personality and Social Psychology 67, 319–333.

Corr, P.J., 2013. Approach and avoidance behavior: multiple systems and their interactions. Emotion Review 5, 286–291.

Corr, P.J., DeYoung, C.G., McNaughton, N., 2013. Motivation and personality: a neuropsychological perspective. Social and Personality Psychology Compass 7, 158–175.

Corr, P.J., McNaughton, N., 2008. Reinforcement sensitivity theory and personality. In: Corr, P.J. (Ed.), The Reinforcement Sensitivity Theory of Personality. Cambridge University Press, pp. 155–187.

Cunningham, W.A., Arbuckle, N.L., Jahn, A., Mowrer, S.M., Abduljalil, A.M., 2010. Aspects of neuroticism and the amygdala: chronic tuning from motivational styles. Neuropsychologia 48, 3399–3404.

Denissen, J.J.A., Penke, L., 2008. Motivational individual reaction norms underlying the five-factor model of personality: first steps towards a theory-based conceptual framework. Journal of Research in Personality 42, 1285–1302.

DeYoung, C.G., 2010. Mapping personality traits onto brain systems: BIS, BAS, FFFS, and beyond. European Journal of Personality 24, 404–407.

DeYoung, C.G., Gray, J.R., 2009. Personality neuroscience: explaining individual differences in affect, behavior, and cognition. In: Corr, P.J., Matthews, G. (Eds.), The Cambridge Handbook of Personality Psychology. Cambridge University Press, Cambridge, pp. 323–346.

Ebstein, R.P., 2006. The molecular genetic architecture of human personality: beyond self-report questionnaires. Molecular Psychiatry 11, 427–445.

Elliot, A.J., Thrash, T.M., 2002. Approach-avoidance motivation in personality: approach and avoidance temperaments and goals. Journal of Personality and Social Psychology 82, 804–818.

Eysenck, H.J., 1944. Types of personality: a factorial study of seven hundred neurotics. Journal of Mental Science 90, 851–861.

Eysenck, H.J., 1967. The Biological Basis of Personality. Charles C. Thomas, Springfield, IL.

Fleeson, W., Gallagher, P., 2009. The implications of Big Five standing for the distribution of trait manifestation in behavior: fifteen experience-sampling studies and a meta-analysis. Journal of Personality and Social Psychology 97, 1097–1114.

Gable, S.L., Reis, H.T., Elliot, A.J., 2003. Evidence for bivariate systems: an empirical test of appetition and aversion across domains. Journal of Research in Personality 37, 349–372.

Gray, J.A., 1975. Elements of a Two-Process Theory of Learning. Academic Press, London.

Gray, J.A., 1982. The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-Hippocampal System. Oxford University Press, Oxford.

Gray, J.A., McNaughton, N., 2000. The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-Hippocampal System, second ed. Oxford University Press, Oxford.

Hebb, D.O., 1955. Drives and the C.N.S. (conceptual nervous system). Psychological Review 62, 243–254.

Markon, K.E., Krueger, R.F., Watson, D., 2005. Delineating the structure of normal and abnormal personality: an integrative hierarchical approach. Journal of Personality and Social Psychology 88, 139–157.

McNaughton, N., Corr, P.J., 2008. The neuropsychology of fear and anxiety: a foundation for reinforcement sensitivity theory. In: Corr, P.J. (Ed.), The Reinforcement Sensitivity Theory of Personality. Cambridge University Press, Cambridge, pp. 44–94.

Mobbs, D., Hagan, C.C., Azim, E., Menon, V., Reiss, A.L., 2005. Personality predicts activity in reward and emotional regions associated with humor. Proceedings of the National Academy of Sciences (PNAS) of the United States of America 102, 16502–16506.

Mowrer, O.H, 1960. Learning Theory and Behavior. Wiley, New York.

Munafo, M.R., 2009. Behavioural genetics: from variance to DNA. In: Corr, P.J., Matthews, G. (Eds.), The Cambridge Handbook of Personality Psychology. Cambridge University Press, Cambridge, pp. 287–304.

Poulin, M.J., Holman, E.A., Buffone, A., 2012. The neurogenetics of nice: receptor genes for oxytocin and vasopressin interact with threat to predict prosocial behavior. PsychologicalScience 23, 446–452.

Reuter, M., Schmitz, A., Corr, P., Hennig, J., 2006. Molecular genetics support Gray's personality theory: the interaction of COMT and DRD2 polymorphisms predicts the behavioral approach system. International Journal of Neuropsychopharmacology 9, 155–166.

Zelenski, J.M., Larsen, R.J., 1999. Susceptibility to affect: a comparison of three personality taxonomies. Journal of Personality 67, 761–791.

Zuckerman, M., 2005. Psychobiology of Personality, second ed. Cambridge University Press, New York.

Relevant Websites

http://www.philipcorr.net – Professor Philip Corr Psychological Research.

http://www.psychol.cam.ac.uk/personalityproject – University of Cambridge Department of Psychology.

http://www.psychometriclab.com – London Psychometric Laboratory at UCL.

Personality Changes During Adolescence Across Cultures

Boris Mlačić and Goran Milas, Institute of Social Sciences Ivo Pilar, Zagreb, Croatia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of personality development in adolescence is a relatively new enterprise, with most of the research in this field having been conducted in the last 10 years. Reasons for the delayed interest in personality changes during adolescence are in the early views on personality development, the lack of a shared model and consensus on measuring stability versus change, and the focus of trait psychology on the adult personality. There are two broad views on the Big-Five personality development in adolescence: one which advocates adolescence as a key period of change, and another, which considers young adulthood to be more important. Regarding the changes in personality dimensions, research findings are mixed; but there are indications that after adolescence Extraversion decreases, Openness (Intellect) increases, while Agreeableness and Conscientiousness follow a curvilinear trend, decreasing in adolescence and increasing in early adulthood. The direction of change for Emotional Stability is still unclear. Personality traits in adolescence are related to important life outcomes; however, cross-cultural studies are still scarce.

Introduction

The study of personality changes in adolescence has only recently captured the interest of researchers, mostly over the past 10 years. There are several reasons for the scarcity of studies on personality development in general, and specifically during adolescence, prior to the last decade. The first reason was associated with the early views dominant in personality psychology which considered personality development to be mostly finalized by the ages of six or seven, i.e., in childhood. The second reason was the absence of a shared model in personality psychology, which hampered the whole field until the rise of the Big-Five model (Goldberg, 1990). The third reason was the lack of consensus regarding the measurement of personality development and methods for assessing change versus stability, not just during adolescence, but throughout the course of life. Finally, the fourth reason for the relative lack of studies on personality changes in adolescence was the focus on the adult personality in such studies, particularly those devoted to the structure of personality or the trait approach, which were conducted mostly on student samples.

Early Views on Personality Changes in Adolescence

Adolescence is often viewed as a period of 'storm and stress' (Arnett, 1999); it is therefore surprising that the studies on personality changes in adolescence emerged so late in the day. One of the reasons outlined above was associated with the view of the early personality theorists, especially Freud (1966), who claimed that the structure of personality with its main components of id, ego and superego, as well as the relations among those structures, are already determined by the start of the latency period, approximately at the age of six. Freud's views were shared by most of his fellow psychoanalysts, including Adler, Horney, Jung, and Anna Freud, who either emphasized the role of childhood as being crucial in personality development or just generally commented on the importance of other periods in life for personality change – as was, for example, Jung's insistence on middle age as a crucial period for full personality development.

The most elaborated view on personality development throughout the life cycle, within the field of psychoanalysis and probably the whole of personality psychology at that time, was delivered by Erik Erikson (1950) who posited the psychosocial theory of development. Erikson's theory was prominent for three major elements. The first was a detailed account of eight stages of development that spanned from birth till death, with each stage consisting of a basic crisis and the solution of that crisis (or lack thereof). The second was a notion that each crisis is salient during one stage, but has roots in a previous stage and consequences in the next stages. The notion highlights the continuity of personality development and could be translated to recent studies in the sense that each phase of personality change and stability is rooted in the previous stage, and reflects in the following stage. The third was an emphasis on adolescence as a key period in personality development with an emerging sense of identity. Postulation of Erikson's theory marked the end of the first era regarding views on personality development, where childhood was deemed the most important. It opened two doors for subsequent developments – the first being research on personality development throughout the whole life cycle and the other being the research on adolescence as a keystone for the formation of adult personality.

The Development of a Shared Model of Personality

The history of research on personality changes in adolescence was not only marred by the early disregard of other developmental periods besides childhood but also by the lack of a shared model which would organize and subsume the main findings. Personality psychology, probably partly due to developments in other disciplines in psychology and partly due to the complexity of its own subject matter, witnessed at least five major approaches, namely the psychoanalytical,

humanistic, social–cognitive, learning, and structural (dispositional), whose influences waxed and waned throughout the last century (cf Pervin et al., 2005). Each of these approaches had not just a unique view on what was most important for the study of personality, supplemented with specific research methods, but more importantly, a unique view on personality development. Even within the structural–dimensional approach to personality (frequently referred to as the 'trait approach'), the dominant model switched between Cattell's 16 factors and Eysenck's three dimensions, finally settling on the Big-Five (Goldberg, 1993) or the Five-Factor-Model (FFM) model (McCrae and Costa, 1997a). Personality traits reflect individual differences or dispositions in behaving, thinking, and feeling in consistent ways (Pervin et al., 2005). Since traits are both organizational constructs in the way they influence an individual in organizing his or her behavior, and developmental constructs in the sense that they show changes across the life course (Roberts et al., 2001), it is not surprising that the recent surge of interest in research for personality changes in adolescence, came after the consolidation of a shared model in the dispositional (or the trait) approach to personality.

The shared model of personality was built upon a lexical approach to personality with a simple premise that all significant individual differences can be found in language, since personality traits are central to human functioning and interactions. Working in the psycholexical tradition, Lewis Goldberg promoted at the beginning of 1990s the popularity of the Big-Five model with the main dimensions of Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Intellect. Costa & McCrae's FFM model, built upon the success of the lexical approach, differs from Goldberg's model only in the fifth dimension (Intellect), which in the FFM model is named 'Openness to Experience.' The first, broad, personality dimension of Extraversion pertains to traits of sociability, talkativeness, assertiveness, liveliness, impulse expression, boldness, and confidence. Most succinctly, Extraversion reflects an energetic approach of the individual to the social and material world (John and Srivastava, 1999). The dimension of Agreeableness refers to characteristics of understanding, warmth, compassion, trusting, helping others, peacefulness, and sincerity. Agreeableness is an interpersonal dimension mostly related to prosocial behavior (Graziano and Eisenberg, 1997). The Conscientiousness dimension comprises traits such as organization, work, responsibility, orderliness, efficiency, reliability, and impulse control. Since Conscientiousness refers to conformity and socially prescribed impulse control (Hogan and Ones, 1997), older models of personality used names such as 'character' or 'superego' to describe the core content of this dimension. Emotional Stability pertains to aspects of stability, even temper, calmness, happiness, tranquility, moderation, and unemotionality. Traditionally, personality psychologists were interested in the negative pole of this dimension, 'Neuroticism,' characterized by anxiety, depression, impulsiveness, vulnerability, and generally poor ability to cope with stress (cf Pervin et al., 2005). The fifth Big-Five dimension of Intellect refers to cleverness, ingenuity, imagination, depth, creativity, and reflection – thus concentrating on core aspects of being smart, whereas Costa & McCrae's fifth dimension of 'Openness to Experience' differs in content and apart from intellectual curiosity and imagination includes fantasy, aesthetics, feelings, ideas, actions, and values (McCrae and Costa, 1997b). Although the validity of the Big-Five model was confirmed in many languages other than English, the replication of the model was not universal, with studies yielding different numbers of factors; recent cross-cultural comparisons showed three factors (Extraversion, Agreeableness, and Conscientiousness) to be most salient and comparable in different cultures (De Raad et al., 2010). Nevertheless, the personality model with five factors became dominant in the discipline and spurred the research on personality development in subsequent years.

Measurement of Personality, Change, and Stability

After a consensus regarding the model on which to base the research was reached, the scientific study of personality changes still had to resolve one essential element: i.e., the issue of measurement. The choice of a measurement strategy can fundamentally affect the conclusions of a study on personality development. For instance, following the seminal publication of Costa and McCrae (1994), a common belief was that by the age of 30, personality would become 'set like plaster' and therefore, no personality change was expected after a person had entered the adult age. However, Costa and McCrae's main conclusions were based on rank-order consistency of personality or the relative position of an individual in a group through time. When researchers started to base their conclusions on mean-level consistency, i.e., on increases or decreases in personality dimensions over time, it became evident that there can be changes in personality even after the age of 30 years, especially reflected in increases in Conscientiousness and Agreeableness (Srivastava et al., 2003).

In trying to tackle the problem of stability versus change, the research on personality traits agreed that change is a multifaceted construct (Caspi and Roberts, 1999), and that conclusions regarding personality change in any given age could be faulty or even misleading if all (or at least the most important) facets of change are not addressed. Publications that followed (De Fruyt et al., 2006; Roberts and DelVecchio, 2000) identified five types of personality continuity, especially important in childhood and adolescence: the differential, mean-level, structural, individual-level, and ipsative. Differential and mean-level stability have been studied the most, the former describing the degree of stability of relative differences among individuals through time, and the latter describing the change of personality scores across time. Differential stability requires longitudinal studies and compares trait correlations over time; whereas mean-level stability can be measured in longitudinal studies but also in cross-sectional studies comparing age cohorts (De Fruyt et al., 2006). Meta-analyses are also valuable for both differential stability and mean-level stability studies (Roberts and DelVecchio, 2000; Roberts et al., 2006). The other three types of stability have been studied far less frequently. Structural stability seeks for invariance of covariance structures across time and is a prerequisite for mean-level stability studies since the conclusions about mean-level changes cannot be valid if the covariance structure changes over time (De Fruyt

et al., 2006). Individual-level stability pertains to a degree of change of a certain trait in a certain individual across time. Finally, ipsative stability examines the relative salience of attributes within an individual over time (Roberts and DelVecchio, 2000). The latter two stabilities are related to individual-level information and importantly complement the former two stabilities related to the group-level data.

Having fulfilled the final criterion necessary to reliably address the issue of personality change through the life course, (i.e., the comprehensive view on change vs. stability), the trait approach, over the past several years, witnessed a rising number of studies devoted to personality development, especially in childhood and adolescence.

Personality Changes and Stability in Adolescence

The Big-Five Dimensions and Lower-Level Constructs

Broadly speaking, there are two views on Big-Five personality development in adolescence. The first view considers that important changes in Big-Five dimensions occur during the period of adolescence (e.g., Branje et al., 2007; Soto et al., 2011), while the other view believes that adolescence is simply a preparation phase that lays the groundwork for change, and that most of the changes actually happen during young adulthood (e.g., De Fruyt et al., 2006; Roberts et al., 2006). Moreover, important studies with intermediate results (e.g., Allik et al., 2004), added to the cacophony of findings on personality change in adolescence.

Drawing on data from 92 studies, 113 samples and over 50 000 participants, Roberts et al. (2006) undertook a large meta-analysis and concluded that most of the changes in personality traits do not occur during adolescence, but rather during young adulthood, specifically from ages 20 to 40 years. Moreover, they concluded that adolescence was not a critical period for shaping personality traits, but a period of lull for trait development. Roberts et al. (2006) also argue that maturation is the major process behind changes in personality, reflected in the fact that most of the change occurs in young adulthood when people face major life transitions. However, Roberts et al. (2006), noted one important change in personality traits that does occur during adolescence, namely they reported a mean-level change in social dominance (a facet of Extraversion) that increased significantly in adolescence.

The study of De Fruyt et al. (2006) stands out as a rare example, as it examined all the five types of personality continuity described above in a single study. Working with a representative population sample and a twin and sibling sample of children and adolescents, the authors found evidence for structural stability at the level of five personality factors, stable differential or rank ordering of individuals in that period; and also reported that at the individual level, changes can be observed more frequently in young adulthood rather than in adolescence (De Fruyt et al., 2006). Regarding mean-level changes, De Fruyt et al. (2006) found some small-magnitude differences at the facet level in late adolescence and generally concluded that the level of personality continuity in adolescence is higher than usually expected.

The research of Allik et al. (2004), working with a large sample of 2650 Estonian adolescents found evidence at the mean-level score, for discontinuity of two Big-Five factors in adolescence, namely, Agreeableness and Conscientiousness and interpreted it as a developmental gap. In their study, Neuroticism and Extraversion smoothly transitioned from adolescence to young adulthood, while Openness tended to slightly increase during late adolescence. However, Allik et al. (2004) concluded, that when compared to the great changes in intellectual capabilities in adolescence, the development in personality traits was very small. Notable from this study is also the conclusion that a person is never more extraverted than at the age of 16 years, and that self-reported personality structure at the age of 16 years is the same as adult structure.

The study of Branje et al. (2007) is particularly interesting for findings on different mean-level changes in adolescence, depending on gender. While adolescent girls increased in self- and other-reported Conscientiousness, other-reported Agreeableness and Extraversion, and also self-reported Openness, adolescent boys decreased in self- and other-reported Extraversion and in other-reported Openness (Branje et al., 2007). The authors made a bold claim that in early adolescence, cerebral cortical development is faster in girls, and that this is parallel to earlier personality development. Branje et al. (2007) also concluded that there is evidence of increasing maturity even in adolescence, since adolescents participate increasingly in roles outside their homes and immediate family.

Finally, a recent study by Soto et al. (2011) indicated that regarding mean-level differences in Big-Five factors, unlike the conclusions outlined above, late childhood and adolescence are key periods for understanding personality differences throughout the life cycle. Working with a massive cross-sectional sample of over a million and a quarter participants from the world wide web, ranging in age from 10 to 65 years, Soto et al. (2011) concluded that some traits substantially differed only in late childhood, and some showed curvilinear trends, decreasing from late childhood to early adolescence and then rising as persons enter young adulthood. Soto et al. (2011) interpret these findings by paralleling the negative personality trends in adolescence with biological and social challenges of the same age. The strength of the study by Soto et al. (2011), apart from the large sample, is the fact that although they worked within the confines of a cross-sectional study, the authors managed to escape the pitfalls of cohort effects inevitably associated with such studies, by collecting data over a 7 year period, ensuring generalizability over cohorts.

Although the above section indicates that among researchers there still is no consensus as to whether adolescence is the key period for personality change, or whether young adulthood has taken over the role of being a cornerstone for personality maturation, some conclusions regarding the Big-Five dimensions and lower-level constructs in adolescence can be made.

Extraversion

Each of the Big-Five domains represents a broad construct at the top-level of abstraction in the personality hierarchy, and

consists of several lower-level specific traits. McCrae and Costa (1997a) posit that each FFM dimension comprises exactly six facets. The work on the Big-Five, on the other hand, was not univocal about the subdomains, with the number of facets ranging from only a few and up to nine per dimension, and finally settled on a two-level hierarchy with two broader aspects and 15 narrower facets for each of the Big-Five factors (DeYoung et al., 2007).

Just as there are different views on the change versus stability of Big-Five factors in adolescence in general, the research findings regarding the changes in the broad domain of Extraversion are mixed. Some deem that Extraversion shows mostly stability during adolescence and that there are no consistent changes at the level of the broad factor and associated facets (Asendorpf and Van Aken, 2003; McCrae et al., 2002). However, the study of Akse et al. (2007) found that the mean-level of Extraversion decreased from early to middle adolescence while the rank-order stability of Extraversion increased at the same age. Soto et al. (2011) found that overall, Extraversion showed a modest negative trend from late childhood to adolescence and then flattened through early adulthood. We can, therefore, conclude that if there is a change in Extraversion during adolescence, it will manifest in a decrease, at least after the age of 16 years (Allik et al., 2004).

However, all the Extraversion facets do not follow this general trend since Roberts et al. (2006) found that people increase in social vitality (marked by gregariousness, positive affect, and level of energy) in adolescence. Roberts et al. (2006) also reported an increase in the second facet of Extraversion, social dominance (marked by self-confidence and assertiveness) during the same age, although moderated by cohorts – with younger cohorts (born after 1940) showing a greater increase in the measured facet. Soto et al. (2011) found that the facet of activity showed a marked negative trend from late childhood to adolescence, while the facet of assertiveness showed only a modest negative trend in the same period.

Agreeableness

The findings on changes in the broad factor of Agreeableness in adolescence almost parallel the previous dimension (i.e., Extraversion) with mixed results. While a group of researchers found that there were no changes either at the factor or the facet level (McCrae et al., 2002), and that Agreeableness showed stability in adolescence (Asendorpf and Van Aken, 2003; Roberts et al., 2006), Allik et al. (2004) found that the level of Agreeableness decreased from 12 to 18 years. Soto et al. (2011) found a negative trend for Agreeableness from late childhood to early adolescence and then a positive trend from adolescence to young adulthood. To add to the variety of results, Akse et al. (2007) reported that Agreeableness increased from early to middle adolescence. We can summarize that if there is a change in Agreeableness in adolescence, it will follow a curvilinear trend. However, more research is needed to establish firmer conclusions.

Regarding the facets constituting Agreeableness, Soto et al. (2011) reported that altruism and compliance followed the same trends as broad Agreeableness, which would be a negative trend from late childhood to early adolescence and then a positive trend. The exception is compliance in boys which did not show a negative trend, since boys were low in compliance even at the age of 10 years.

Conscientiousness

Research on the development of Conscientiousness in adolescence again shows inconsistencies, with a group of researchers indicating no substantial change (McCrae et al., 2002; Roberts et al., 2006), and another group presenting evidence for change. De Fruyt et al. (2006) found evidence for a small decrease in Conscientiousness in the age group from 12 to 13 years. Similarly, Allik et al. (2004) found a slight decline of Conscientiousness from ages 12 to 18 years, from which there was a considerable increase. Soto et al. (2011) came to the same conclusions for Conscientiousness as for Agreeableness; both showed a negative trend from late childhood to early adolescence and then a positive trend to young adulthood. As outlined earlier, Branje et al. (2007) found an increase of Conscientiousness in adolescent girls, but no change in adolescent boys. Therefore, as with Agreeableness, we can conclude that if there is a change in Conscientiousness during adolescence, it will follow a curvilinear, nonmonotonic age trend. Still, more research on this topic is needed.

Regarding the facets, the study of Soto et al. (2011) showed that both self-discipline and order followed the same trends in adolescence as overall Conscientiousness, first expressing a negative trend from late childhood to adolescence and then a positive trend into young adulthood.

Developmental gaps for both Agreeableness and Conscientiousness in adolescence can be interpreted with a leap in socialization at the age of graduation from high school (Allik et al., 2004; McCrae et al., 2002).

Emotional Stability (Neuroticism)

Unlike the three previously described Big-Five factors, there is almost an universal consensus that Emotional Stability does change in adolescence. However, McCrae et al. (2002) interpreted the changes as modest, and noticed that girls increased in Neuroticism from age 12 to 18 years. McCrae et al. (2002) offered both a biological explanation for this effect (in girls' hormonal changes) and a cultural explanation (in society-imposed sex roles). However, Roberts et al. (2006) found the opposite effect: a small increase in Emotional Stability during teenage years, with no gender effects. A similar effect was found by De Fruyt et al. (2006): small mean-level decrease in emotional instability for age groups 10–11 years and 12–13 years. Allik et al. (2004) reported that Neuroticism showed a smooth transition from adolescence to young adulthood, meaning that the mean-level of adolescents resembles the mean-level of adults. Soto et al. (2011) found very different age trends for Neuroticism depending on gender. For females, Neuroticism showed positive trends from childhood to adolescence; while for men Soto et al. (2011) reported the opposite effect, a negative trend from late childhood. We can conclude that there are indications for changes in Emotional Stability (Neuroticism) during adolescence, but for final conclusions on the direction of change and gender differences, more research is needed.

Regarding the facets, Soto et al. (2011) found similar trends to those of overall Neuroticism: females showed positive trends

for anxiety in adolescence, while the opposite was found for men. Depression showed two peaks for females in adolescence and early adulthood, while for men there was a positive trend from late childhood into early adulthood.

Intellect (Openness to Experience)

The final dimension of the Big-Five or FFM should change in adolescence, judging by the available studies. Even McCrae et al. (2002), the main advocates noting no substantial changes in other dimensions during adolescence, report an increase in Openness for both genders and interpret such growth as being parallel with the growth of cognitive capacities and interest in various aspects of experience. The increase in Openness during adolescence was also supported by studies of Akse et al. (2007), Allik et al. (2004), Branje et al. (2007), and Roberts et al. (2006) who also reported a small increase in Openness in adolescence. However, the study of De Fruyt et al. (2006) found a negative trend for Openness in early adolescence while Soto et al. (2011) found negative trends from late childhood into early adolescence and further negative trends across adolescence for females. After that, there were positive trends for both genders into early adulthood (Soto et al., 2011). Therefore, we can conclude that the results indicating a positive change in Openness during adolescence prevail, however with contrary findings; the claim by Soto et al. (2011), of a curvilinear trend from childhood to young adulthood with negative trends at younger ages, deserves further investigation.

Regarding the facets of Openness, McCrae et al. (2002) found evidence for increase in four facets in adolescence: aesthetics, feelings, ideas, and values. However, De Fruyt et al. (2006) found a slight decrease for curiosity and creativity in the age group 12–13 years. Soto et al. (2011) found a negative trend from childhood to early adolescence for openness to ideas, and a positive trend for male openness to aesthetics from late childhood to young adulthood versus a flat trend for females.

Correlates of Personality Traits in Adolescence

Research on personality changes in adolescence is not only important per se, but also because there are findings linking the adolescent personality to important life outcomes. An early study of John et al. (1994) found that adolescent boys with externalizing behavior problems such as aggression, stealing, inattention, and hyperactivity had the pattern of low Agreeableness and Conscientiousness but also high Extraversion. Boys with internalizing behavior problems such as anxiety, somatic complaints, and social withdrawal were more neurotic and less conscientious, while delinquent boys were low on Agreeableness and Conscientiousness (John et al., 1994).

Shiner (2000) summarized the findings on childhood personality traits as important predictors of adaptation in adolescence: Extraversion was a predictor of academic achievement in children of school-going age but introversion was a predictor of achievement in late adolescence. 'Surgent engagement' (related to Extraversion) predicted late adolescent academic achievement negatively, while children high on 'mastery motivation' (related to Conscientiousness) showed less antisocial behavior and better social competence during adolescence. Moreover, children high on Agreeableness were more socially competent in late adolescence with less agreeable adolescents reporting higher levels of everyday conflicts and destructive conflict tactics; Academic Conscientiousness became more relevant for scholastic achievement over time, etc. (Shiner, 2000). The study of McCullough et al. (2003) showed that adolescent Conscientiousness was related to higher religiousness in early adulthood while the most emotionally stable adolescents showed the weakest relation between their religious upbringing and subsequent adult religiousness. The findings of both studies corroborate the notion of continuity of personality development, where childhood personality predicted the outcomes in adolescence; and adolescent personality was related to adult outcomes.

Cross-Cultural Variations in Personality during Adolescence

If the study of personality changes and stability in adolescence is itself in its adolescence, with most of the research having been conducted over the last 15 years, then the research on cross-cultural similarities and differences of personality in adolescence is in its true infancy. Most of the studies conducted so far used samples of participants either from the United States, from English-speaking countries or from the western world. The studies on adolescent personality outside of the westernized world are even more scarce than cross-cultural studies on adult personality, where the most research so far has been done on personality structure and assessment (Church and Lonner, 1998). McCrae et al. (2002) compared American and Belgian adolescents and found more evidence for similarity than for differences, concluding that there are indications for generalizability of developmental patterns. McCrae et al. (2000) compared British, German, Spanish, Czech, and Turkish adolescents with their adult counterparts and found that adolescents were lower then adults on Conscientiousness, again indicating generalizability. The study of Knyazev et al. (2008) generally found evidence for the five factor structure in children and adolescents both in Russia and Slovenia. Regarding mean-level differences, Slovenian adolescents scored higher on Extraversion and Openness, somewhat higher on Conscientiousness and Neuroticism, while Russians scored higher on Agreeableness, with economic, social, and climate factors as potential explanations for the differences (Knyazev et al., 2008).

However, most of the cross-cultural studies on adolescent personality, if not all of them, were made within the confines of etic strategy (Berry, 1969), where existing personality measures are imported, translated, and adapted to a different culture, and true emic strategies (Berry, 1969) are still missing in the study of adolescent personality. The universality and sufficiency of the Big-Five or FFM models in Asian contexts has been questioned and the etic approach was frequently criticized by Fanny M. Cheung and her coworkers (e.g., Cheung, 2004; Cheung et al., 2001, 2003). Specifically, Cheung and her colleagues came to the conclusion that 'openness to experience'

seems like an imported construct to Chinese culture and they discovered an indigenous dimension of 'interpersonal relatedness,' beyond the Big-Five, emphasizing interdependence in collectivistic cultures, such as in China (Cheung et al., 2001, 2011). Moreover, Cheung et al. (2011) pleaded for a combined etic–emic approach in personality psychology. Those studies, trying to identify indigenous personality constructs for any given culture and comparing them to other cultures are starting to appear within the lexical approach to personality (De Raad et al., 2010). These are conducted using adult samples, but would be useful to translate into research on adolescent personality as well.

Future Directions

The overview of research on personality changes in adolescence outlined above, shows that this is a new, vibrant, and promising field that can enhance our understanding of personality development and maturation, as well as underscore the importance of each developmental period, the important outcomes, and personality continuity through the life cycle. Although numerous important findings have stemmed from just a decade of research, many questions still remain to be answered. As noted here, rare are the studies that investigate all types of personality continuity and change (e.g., De Fruyt et al., 2006). Two types of continuity, differential and mean-level continuity, are still being studied the most. There are important findings at the level of broad Big-Five factors; however, the research on facets is still far from being conclusive, partly due to different conceptions of what lies below the broad factors. Integrative longitudinal studies on personality in adolescence and the whole course of life, should be conducted far more often in concert with other important measures of development, such as cognitive reasoning and moral judgment.

Finally, the lack of studies on personality changes in adolescence in cultures other than western societies is almost an embarrassment for the field, and future research should take a giant step toward that direction.

See also: Cross-Cultural Psychology; Erikson's Theory of Psychosocial Development; Factor Analysis and Latent Variable Models in Personality Psychology; Indigenous Psychology; Motivational Development, Systems Theory of; Neuroticism; Personality Assessment: Overview; Personality Differences and Development: Genetic and Environmental Contributions; Personality, Trait Models of; Personality: Historical and Conceptual Perspectives; Prosocial Behavior During Adolescence; Self and Identity Development During Adolescence across Cultures; Stress in Adolescence: Effects On development; Wisdom, Psychology of.

Bibliography

Akse, J., Hale, W.W., Engels, R.C.M.E., Raaijmakers, Q.A.W., Meeus, W.H.J., 2007. Stability and change in personality type membership and anxiety in adolescence. Journal of Adolescence 30, 813–834.

Allik, J., Laidra, K., Realo, A., Pullmann, H., 2004. Personality development from 12 to 18 years of age: changes in mean levels and structure of traits. European Journal of Personality 18, 445–462.

Asendorpf, J.B., Van Aken, M.A.G., 2003. Personality-relationship transaction in adolescence: core versus surface personality characteristics. Journal of Personality 71, 629–666.

Arnett, J.J., 1999. Adolescent storm and stress, reconsidered. American Psychologist 54, 317–326.

Berry, J.W., 1969. On cross-cultural comparability. International Journal of Psychology, 119–128.

Branje, S.J.T., van Lieshout, C.F.M., Gerris, J.R.M., 2007. Big Five personality development in adolescence and adulthood. European Journal of Personality 21, 45–62.

Caspi, A., Roberts, B.W., 1999. Personality change and continuity across the life course. In: Pervin, L.A., John, O.P. (Eds.), Handbook of Personality Theory and Research. Guilford Press, New York, pp. 300–326.

Cheung, F.M., 2004. Use of Western- and indigenously-developed personality tests in Asia. Applied Psychology: An International Review 53, 173–191.

Cheung, F.M., Cheung, S.F., Wada, S., Zhang, J.X., 2003. Indigenous measures of personality assessment in Asian countries: a review. Psychological Assessment 15, 280–289.

Cheung, F.M., Leung, K., Zhang, J.-K., Sun, H.-A., Gan, Y.-Q., Song, W.-Z., Xie, D., 2001. Indigenous Chinese personality constructs: is the five-factor model complete? Journal of Cross-cultural Psychology 32, 407–433.

Cheung, F.M., van de Vijver, F.J.R., Leong, F.T.L., 2011. Toward a new approach to the study of personality in culture. American Psychologist 66 (7), 593–603.

Church, A.T., Lonner, W.J., 1998. The cross-cultural perspective in the study of personality: rationale and current research. Journal of Cross-Cultural Psychology 29, 32–62.

Costa Jr., P.T., McCrae, R.R., 1994. Set like plaster: evidence for the stability of adult personality. In: Heatherton, T.F., Weinberger, J.L. (Eds.), Can Personality Change? American Psychological Association, Washington, DC, pp. 21–40.

De Fruyt, F., Bartels, M., Van Leeuwen, K.G., De Clercq, B., Decuyper, M., Mervielde, I., 2006. Five types of personality continuity in childhood and adolescence. Journal of Personality and Social Psychology 91, 538–552.

De Raad, B., Barelds, D.P.H., Levert, E., Ostendorf, F., Mlačić, B., Di Blas, L., et al., 2010. Only three factors of personality description are fully replicable across languages: a comparison of 14 trait taxonomies. Journal of Personality and Social Psychology 98, 160–173.

DeYoung, C.G., Quilty, L.C., Peterson, J.B., 2007. Between facets and domains: 10 aspects of the Big Five. Journal of Personality and Social Psychology 93, 880–896.

Erikson, E., 1950. Childhood and Society. W. W. Norton & Company, New York.

Freud, S., 1966. Introductory Lectures on Psychoanalysis. Norton, New York.

Goldberg, L.R., 1990. An alternative "description of personality": the Big-Five factor structure. Journal of Personality and Social Psychology 59, 1216–1229.

Goldberg, L.R., 1993. The structure of phenotypic personality traits. American Psychologist 48, 26–34.

Graziano, W.G., Eisenberg, N., 1997. Agreeableness: a dimension of personality. In: Hogan, R., Johnson, J.A., Briggs, S.R. (Eds.), Handbook of Personality Psychology. Academic Press, San Diego, pp. 795–824.

Hogan, J., Ones, D.S., 1997. Conscientiousness and integrity at work. In: Hogan, R., Johnson, J.A., Briggs, S.R. (Eds.), Handbook of Personality Psychology. Academic Press, San Diego, pp. 849–870.

John, O.P., Caspi, A., Robins, R.W., Moffitt, T.E., Stouthamer-Loeber, M., 1994. The "little five": exploring the nomological network of the five-factor model of personality in adolescent boys. Child Development 65, 160–178.

John, O.P., Srivastava, S., 1999. The Big Five trait taxonomy: history, measurement, and theoretical perspectives. In: Pervin, L.A., John, O.P. (Eds.), Handbook of Personality Psychology: Theory and Research, second ed. Guilford Press, New York, pp. 102–138.

Knyazev, G.G., Zupančič, M., Slobodskaya, H.R., 2008. Child personality in Slovenia and Russia: structure and mean level of traits in parent and self-ratings. Journal of Cross-Cultural Psychology 39 (3), 317–334.

McCrae, R.R., Costa Jr., P.T., 1997a. Personality trait structure as a human universal. American Psychologist 52, 509–516.

McCrae, R.R., Costa Jr., P.T., 1997b. Conceptions and correlates of openness to experience. In: Hogan, R., Johnson, J.A., Briggs, S.R. (Eds.), Handbook of Personality Psychology. Academic Press, San Diego, pp. 269–290.

McCrae, R.R., Costa, P.T., Ostendorf, F., Angleitner, A., Hrebícková, M., Avia, M.D., Sanz, J., et al., 2000. Nature over nurture: temperament, personality, and life span development. Journal of Personality and Social Psychology 78, 173–186.

McCrae, R.R., Costa Jr., P.T., Terracciano, A., Parker, W.D., Mills, C.J., De Fruyt, F., Mervielde, I., 2002. Personality trait development from age 12 to age 18: longitudinal, cross-sectional and cross-cultural analyses. Journal of Personality and Social Psychology 83, 1456–1468.

McCullough, M.E., Tsang, J.-A., Brion, S., 2003. Personality traits in adolescence as predictors of religiousness in early adulthood: findings from the Terman longitudinal study. Personality & Social Psychology Bulletin 29, 980–991.

Pervin, L.A., Cervone, D., John, O.P., 2005. Personality: Theory and Research, ninth ed. John Wiley & Sons, New York.

Roberts, B.W., Caspi, A., Moffitt, T.E., 2001. The kids are alright: growth and stability in personality development from adolescence to adulthood. Journal of Personality and Social Psychology 81, 670–683.

Roberts, B.W., DelVecchio, W.F., 2000. The rank-order consistency of personality traits from childhood to old age: a quantitative review of longitudinal studies. Psychological Bulletin 126 (1), 3–25.

Roberts, B.W., Walton, K., Viechtbauer, W., 2006. Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. Psychological Bulletin 132, 1–25.

Shiner, R.L., 2000. Linking childhood personality with adaptation: evidence for continuity and change across time into late adolescence. Journal of Personality and Social Psychology 78, 310–325.

Soto, C.J., John, O.P., Gosling, S.D., Potter, J., 2011. Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. Journal of Personality and Social Psychology 100, 330–348.

Srivastava, S., John, O.P., Gosling, S.D., Potter, J., 2003. Development of personality in early and middle adulthood: set like plaster or persistent change? Journal of Personality and Social Psychology 84, 1041–1053.

Personality, Cognitive Models of

Gerald Matthews, Institute for Simulation and Training, University of Central Florida, Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cognitive models of personality assume that people differ in their styles of understanding the world and themselves. More specifically, people differ in core self-beliefs and in the processes that shape and maintain the self. Contemporary research has focused on stable individual differences in cognitive content and process that may be associated with major personality traits. Studies of self-regulation and coping link traits to cognitive theories of these constructs. Studies using objective performance metrics allow traits to be further related to individual differences in information-processing. Cognitive neuroscience and cognitive-adaptive theory provide contrasting perspectives on the place of processing differences in personality.

Historical Overview

Cognitive psychology sees the person as an active information-processor. People are motivated to construct internal models of the physical and social world around them and their place in that world. These mental models both direct attention to personally relevant stimuli and guide response. People differ in the models they construct, and in the information-processing that supports model construction. The humanistic psychologist Abraham Maslow noted that if all you have is a hammer, everything looks like a nail. In relation to personality, people differ in the mental toolkit they possess for understanding the world. Some may define the world primarily in terms of its threats and dangers, so that their personalities are colored by strategies for identifying, forestalling, and coping with threat. Others may be more tuned into social affordances, so that the world is construed in relation to the social network around the individual.

The idea that personality reflects a style of personal understanding is simple and appealing, but growing in the shade of, first, psychoanalysis, and, subsequently, reductionistic neurobiology, it took some time for modern cognitive theories to develop. In the 1920s and 1930s, Gordon Allport defined traits as neuropsychic structures that render many stimuli functionally equivalent for the individual, e.g., in perceiving a multiplicity of stimuli as threats. The 'neuropsychic structure' was in part cognitive, in that it included general attitudes. Allport also recognized the metacognitive element of personality, in that individuals form beliefs about which of their traits are important to their sense of self and identity. Later theorists elaborated cognitive perspectives by emphasizing the roles of personal construction of the social world (George Kelly) and of representing events in relation to their importance to the self (Carl Rogers). Kelly's view of people as naïve scientists trying to understand their place in the world through reasoning and evaluation of events captures the spirit of theorizing in the 1950s, prior to the 'cognitive revolution' in experimental psychology.

Several developments of the 1960s helped to guide these early and somewhat prescientific insights toward modern cognitive theories of personality. First, Aaron Beck's schema theory of emotional disorder pinpointed faulty thinking and irrational self-beliefs as the basis for depression. It is a small further step to propose that normal variation in traits stems from milder biases in cognition. Second, social learning theorists, notably Albert Bandura, introduced cognitive constructs such as self-efficacy into traditional learning theory, suggesting paths for personality development in childhood. Third, the birth of modern cognitive psychology started to provide experimental paradigms that could be used to measure individual differences in information-processing, and so freeing personality research from the constraints of the armchair and the clinician's couch.

A fourth development was less obviously cognitive: Hans Eysenck's arousal theory of extraversion and neuroticism (Eysenck and Eysenck, 1985). Although skeptical of cognitive psychology, Eysenck's theory proved to be essential for later cognitive work. Eysenck's work established traits as 'real' personal qualities, expressed in individual differences in brain functioning. Researchers could now investigate cognitive bases for traits on a nomothetic basis (conforming to universal principles), without being confined to the idiographic qualities of the individual. Eysenck also emphasized the importance of experimental tests of predictions from theory. Relating traits to individual differences in arousability supported tests of how traits relate to perception, attention, memory, and response selection under different levels of environmental stress. Cognitive psychologists, of course, can infer relationships between traits and information-processing from such experimental paradigms. Most recently, the wheel has turned full circle in the form of cognitive neuroscience, which integrates classic psychobiological perspectives with information-processing theory.

The remainder of this article is structured as follows. The foundational place for trait theory in modern cognitive theory will be elaborated. Two variants of cognitive personality theory will then be differentiated. One is the cognitive self-theory built on the work of Rogers, Bandura, and, especially, Beck: the explanation of traits in terms of individual differences in core self-beliefs, and in the cognitive processes that shape and maintain those beliefs. The second looks more to experimental studies of information-processing and seeks to explain traits in relation to processing differences. The article will finish with perspectives from cognitive neuroscience and an adaptive account of personality.

Traits as Cognitive Constructs

Traits are stable personal qualities that moderate perception and action across a range of different situations (Matthews et al., 2009). They are often seen as basic temperamental characteristics, such as emotionality, that are rather closely tied to brain functions and influenced by heredity as well as environment. However, without denying the importance of neurobiology, the stability of traits may also be seen in cognitive terms. Bandura's developmental studies suggest that children internalize stable self-beliefs in long-term memory that influence emotional and behavioral response. In Beck's theory (Beck and Clark, 1988), an organized set of core self-beliefs is described as the self-schema. While the self is somewhat plastic in childhood, it becomes increasingly stable in the adult years (though not 'set in concrete'). Thus, extraversion may express a set of beliefs that one is socially dominant and that one is admired by others. Similarly, neuroticism may be associated with beliefs surrounding personal vulnerability. Further, traits can also be linked to action tendencies and response inclinations: beliefs about the value of social interaction in the case of extraversion, and about lack of effectiveness in coping in the case of neuroticism.

Contemporary research is often, though not exclusively, framed in terms of the five-factor model (FFM), which adds traits of conscientiousness, agreeableness, and openness to the Eysenck extraversion and neuroticism factors. The three additional factors may also be framed in cognitive terms. Conscientiousness is associated with persistence and effort in goal-directed behavior, as evidenced in numerous studies of academic and occupational achievement. A key cognitive factor associated with conscientiousness is self-efficacy, i.e., in believing that one can accomplish the tasks necessary to reach a goal, overcoming obstacles in the process. Conversely, the unconscientious student, lacking academic self-efficacy, may give up on assignments because of beliefs that the task is too hard. On the action side, conscientious individuals use task-focused coping strategies that promote systematic effort.

Agreeableness has been linked to social beliefs and behavioral inclinations. The more agreeable person is more likely to view other people as basically helpful and supportive, and to believe that cooperation is more successful than competition. Aggression is not identical to low agreeableness, but aligns with the lower pole of the FFM trait. Aggression may stem from distorted beliefs about others, such as falsely believing them to be hostile, and from tendencies to respond confrontively. The last factor in the FFM, openness, is typically defined in relation to interest in intellectual matters and activities, and belief in one's own intelligence. It may also encompass interests in the arts and in the unusual. Openness also correlates with cognitively defined traits, such as need for cognition, the tendency to engage in, and value effortful cognitive processing.

Thus, we could define each of the 'Big Five' (and other) traits in cognitive terms, as a set of characteristics beliefs about self and others, attitudes to key domains of life (e.g., work, socializing, study), and preferred options for taking action and coping with challenges and stressful events. Together, such characteristics define a self-schema that underpins the trait concerned. Evidence on relationships between the traits and relevant psychological constructs, such as appraisal, social attitudes, and coping, supports this conceptual picture. Nevertheless, three significant challenges to this viewpoint should be mentioned.

The first challenge comes from cognitive-social learning theorists such as Walter Mischel and Daniel Cervone (e.g., Caprara and Cervone, 2008), working in the tradition of Bandura. They question whether the beliefs that define personality are really general in nature, preferring instead to link beliefs to specific situations. A person might be socially self-confident on one setting but not in another, so that they would variously appear both extraverted and introverted, according to the environment. Situational influences on cognition are certainly important. The trait theorist would counter that people differ in general biases in thinking (e.g., in social self-efficacy), biases that may be differently expressed, or even reversed, across different contexts. As Allport (1931) stated, "Acts, and even habits, that are inconsistent with a trait are not proof of the non-existence of the trait." However, like Allport, social learning theorists are also prone to emphasize the idiographic beliefs of the individual over more general nomothetic principles.

The second challenge derives from neuroscience. Whereas a cognitive theorist might see the contents of the self-schema as a causal influence on personality, the neurobiologist see beliefs as an output of the brain system that is the true basis for the trait concerned. Thus, socially anxious people do not shun parties because they believe that others will mock them. An oversensitive brain punishment system simultaneously activates behavioral avoidance tendencies as well as thoughts of social threat. At the limit, the person's thoughts may be epiphenomenal froth with no causal influence on behavior. The issue here is the extent to which behavior is subject to neural reductionism.

The third challenge is accommodating emotion and motivation into trait theory. The recent popularity of 'emotional intelligence' is in part a reaction against seeing people as advanced biological robots for whom everything is merely an intellectual exercise, like the character of Spock in 'Star Trek.' Personality traits have defining emotional and motivational qualities that cannot necessarily be reduced to cognition. Notably, extraversion is often associated with positive emotion and approach motivations, and neuroticism with negative emotion and avoidance. Two resolutions are possible. On the one hand, it may be asserted that cognitive processes are indeed primary, but cognition should not be identified only with conscious, deliberate problem-solving; emotions are shaped in part by rapid, unconscious appraisal of stimulus meaning, for example. Alternatively, its cognitive accounts of personality are incomplete, with traits being higher-order qualities that bind together cognition, emotion, and motivation. There remain difficult and unresolved causal questions about how these distinct psychological domains influence one another.

Traits and Self-Beliefs

This section addresses further the role of the cognitive self in personality. At this point, it is important to distinguish content and process in cognitive theory. 'Content' refers to the nature of self-beliefs as they may be represented in long-term memory.

From a clinical perspective, abnormal personality would be a reflection of faulty data. 'Process' refers to the transformations of data that influence the contents of the self, and the behavioral control exerted by the self. In this case, abnormality is dependent on, as it were, faults in the mental subroutines for data processing.

Cognitive Content and the Self-Schema

Both content and process feature in Beck's schema theory (Beck and Clark, 1988). The content of thinking in clinical depression typically reflects a 'negative cognitive triad,' of a negative demeaning view of oneself, the world, and the future, e.g., "I am a worthless person, others despise me, and the future is hopeless." Beck also identified faulty processing that contributes to depression, such as overgeneralization, e.g., assuming one is universally denigrated because of a critical remark. Beck's later theory differentiated depression from anxiety primarily in terms of content (the content-specificity hypothesis). Whereas depression is characterized by themes of worthlessness, loss, and hopelessness, thoughts in anxiety revolve around beliefs that threats are imminent and the person is ill-equipped to deal with them. In fact, tests of the content-specificity hypothesis have had mixed outcomes, and there is some general negative self-referent content common to both emotional disorders.

Cognitive content may play a corresponding role in trait variation in normal populations free of clinical disorder, as shown in studies of neuroticism and trait anxiety. Dispositionally anxious individuals typically express a range of negative self-beliefs including overestimation of risks and underestimation of personal competence and success. Beliefs may be tied to a specific context; individuals high in test anxiety may believe they will fail at even mildly challenging tests, those high in computer anxiety believe they cannot cope with information technology, and so forth (Zeidner, 1998).

Judge et al. (2004) pointed out that, although neuroticism is typically defined as an emotional trait, it correlates substantially and negatively with three important cognitive constructs: self-esteem, locus of control, and generalized self-efficacy. Locus of control refers to the person's style of attribution toward important events. An internal locus signifies personal control of destiny, and is related to emotional stability. They showed that these three belief-based constructs could be subsumed together with neuroticism under a common factor of core self-evaluations. Whereas neuroticism is strongly negatively correlated with the factor, extraversion and conscientiousness are both associated with positive self-evaluations. Core self-evaluations predict job performance, via motivational variables such as goal setting and goal commitment.

Cognitive Process Factors

Process factors are important in part because of individual differences in self-regulation. This term refers to the management of thoughts, feelings, and actions as the person pursues personal goals. Self-regulation implies monitoring one's own internal states to assess the state of progress; indeed emotional responses may be tied to goal progress. It also implies development, application, and monitoring of strategies for maintaining progress. Such strategies may be directed toward regulating personal behavior and action, the external environment, and internal states, including emotion regulation. Similar cognitive models have been applied in other domains of psychology. Lazarus' (1999) transactional theory of stress and emotion describes a cycle of appraising environmental demands, initiating coping, and then responding to feedback as the interaction between person and environmental demands unfolds.

Personality traits correlate with the process variables described by theories of self-regulation and coping. Assuming a cycle of monitoring and 'action' (broadly defined), traits might influence both self-perceptions and choice of regulative or coping strategy. Studies reviewed by Robinson and Sedikides (2009) link traits to processes that shape and maintain the self, such as self-enhancement and self-verification (confirmation of strongly held views of the self). People may use the generalized self-knowledge associated with traits to 'fill in' appraisals of events when lacking detailed information. Thus, among individuals who have trouble making fine discriminations between stimuli, neuroticism is especially predictive of distress. Those high in neuroticism presumably draw upon their negative self-schema in evaluating incompletely encoded stimuli.

Many studies of traits and appraisals investigate both demanding real events, and controlled laboratory scenarios, such as failing on a performance task. Individuals high in neuroticism and trait anxiety tend to appraise events as being more threatening, and themselves as lacking self-efficacy and control over events. Because appraisals are a direct driver of emotional response (Lazarus, 1999), such findings help to explain why dispositionally neurotic individuals are prone to negative emotion. Traits for negative emotionality also tend to be associated with self-consciousness, elevating the salience and immediacy of negative feeling states, thoughts, and images. Conversely, the tendencies toward happiness and positive affect of extraverted individuals may reflect a bias toward appraising events as being challenging, i.e., offering benefits even if there are obstacles to be overcome.

Traits also predict coping preference. There are various classificatory schemes for coping, most simply Lazarus' (1999) distinction between problem-focused (externally directed, action-oriented) and emotion-focused coping (internally directed, regulation of feelings and thoughts). Neuroticism is associated with potentially maladaptive forms of emotion focus, such as self-criticism, and extraversion with more constructive, problem-focused strategies. Associations between these traits and coping style may also contribute to mediating their impacts on emotion. Conscientiousness is also associated with strategies oriented toward controlling the situation including problem focus and cognitive restructuring.

Another important type of process is metacognition. In monitoring their mental landscape, people make judgments about the importance and significance of their thoughts and feelings, which in turn may drive strategic efforts to regulate them. In emotional disorders, people are often disturbed by thoughts and images that seem to intrude automatically into consciousness. Such intrusions reflect cognitive contents as previously described, e.g., images of some feared disaster in

the case of anxiety. However, the impact of intrusions is accentuated by metacognitive processing (Wells, 2011). Patients with generalized anxiety disorder tend to worry about their own worries (metaworry), a tendency that prolongs worry and serves to maintain the disorder. Similarly, unproductive rumination is characteristic of depression, and both worry and rumination feature in subclinical neuroticism. A variety of metacognitive styles can be identified that are often maladaptive (Wells, 2011). More constructive forms of emotion regulation may involve reappraisal of the disturbing thought or image to understand its origins and lack of real significance.

Wells and Matthews (1994) developed a cognitive theory of normal and abnormal negative emotionality that pulls together several of the issues discussed above. The self-referent executive function (S-REF) theory describes a cognitive architecture with three levels. A level of self-knowledge defines the stable generic self-knowledge (content) that shapes processing of self-relevant information. Self-knowledge may be declarative (accessible to consciousness) or procedural (specifying a processing routine that may not be conscious). An executive system accesses self-knowledge in appraising personally relevant events, including internal 'events' such as intrusive thoughts, and in regulating coping. It also initiates metacognitive processing that may direct attention toward thoughts and images. Intrusions are generated by a third level of automatic processing networks that often operates unconsciously but may activate concepts and images to such an extent that they break through into awareness and initiate executive processing.

In 'normal' anxiety and dysphoria, negative self-knowledge may activate the executive with some frequency, but, typically, appraisal and coping are adequate to prevent major life disruption. In clinical patients, however, several factors may come together to maintain extended cycles of perseverative worry or rumination, including unrealistically negative appraisals, choice of ineffective emotion-focused coping strategies, and metacognitions that serve to lock attention onto the internal worry cycle. The person cannot effectively refocus attention on the external world, impairing effective coping. Furthermore, episodes of worry may strengthen the dysfunctional self-knowledge that contributes to the syndrome so that a vicious cycle develops. The S-REF model emphasizes that personality dysfunction is not just a matter of abnormality in content and/or process but an outcome of disordered dynamic interaction between person and environment.

Theories of personality and emotion based on cognitive theories of self-regulation or stress have several strengths. They provide integrated accounts of traits such as anxiety that accommodates evidence from a variety of sources, including evidence on clinical patients and normally functioning individuals. Evidence is supportive of a variety of specific associations between traits and cognitive constructs, such as appraisal, coping and emotion-regulation strategies, and metacognition. Theories are also capable of relating personality to the dynamic interactions between person and environment that feature in contemporary cognitive theory. That is, personality is more than a set of static descriptive characteristics. In clinical psychology, cognitive models provide the basis for an array of cognitive-behavioral therapies (Beck and Clark, 1988; Wells, 2011).

Limits of Cognitive Theory

However, there are also limitations to cognitive theory. Perhaps the most severe is that much of the evidence on which theory is based relies on self-reports of various kinds, including questionnaire measures of constructs such as coping, and patient reports of symptoms. Recent work on implicit aspects of personality shows that much 'cognitive' processing is unconscious and cannot be assessed validly using self-report. Reporting of explicit cognitive content and process may be possible, but also subject to bias, as a consequence of lack of personal insight and awareness, or of deliberate self-distortion (also a feature of self-regulation). Self-verification, for example, may lead to people exaggerating the extent to which they cope consistently with different types of stressor. It may be informative to look at how individuals cope with specific challenges rather than relying on general self-assessments of coping style.

Studies of implicit personality (Schnabel et al., 2008) show the importance of objective measures of cognition. One of the classic techniques is the implicit association test (IAT). Originally developed to assess unconscious social biases, such as prejudice, it has been adapted to assess personality. The test uses reaction time data to determine how easily the person associates various personality characteristics with the self. Implicit measures such as IAT may predict nonconscious behaviors better than explicit scales do. There are typically positive but modest correlations of around 0.3 between corresponding explicit and implicit measures. Thus, while questions remain about the reliability, stability, and validity of implicit trait measures, there appear to be cognitive aspects of personality that do not register on standard questionnaires.

Information-Processing Models for Personality

Evidence for implicit cognitive traits highlights the need to integrate theories of personality with cognitive psychological models that provide fine-grained accounts of information-processing based on objective data on speed and accuracy of performance. Two broad currents of research have been influential in this area. The first starts from the experimental studies of Eysenck (Eysenck and Eysenck, 1985), which showed that traits correlated with performance on tasks such as choice reaction time, vigilance, and short- and long-term recall from memory. In keeping with the stress research of his day, Eysenck attributed these associations to individual differences in arousal of the cortex. In the 1970s and 1980s, stress researchers rejected arousal theory as being overly simplistic, and unable to account for the variation of stressor effects with the cognitive demands of the task. Indeed, arousal did not seem, in general, to play the mediating role stated by Eysenck's personality theory (Matthews and Gilliland, 1999). Rather than investigating performance in some global sense, it is more informative to test how traits relate to different cognitive functions, and the specific processes that support these functions.

An influential cognitive theory was proposed by Humphreys and Revelle (1984) based on 'processing resources' as mediating constructs. In attentional theory, a resource is a general-purpose reservoir of 'mental energy' that facilitates certain processes (controlled processing). Humphreys and

Revelle (1984) differentiated a resource used for sustained information transfer (attention) from a resource for short-term memory. The resource concept is metaphorical, and open to criticism as lacking in precision, but it has proved useful in explaining and predicting how attention varies with task demands. Humphreys and Revelle also retained arousal as a mediating construct. Details of their theory are beyond the scope of this article, but their causal assumptions were that personality traits interact with environmental factors to influence arousal, which influences availability of the two resources, which in turn influences performance on sufficiently demanding tasks. They used the theory to explain why extraversion–introversion effects on intelligence test performance varied with arousing agents such as caffeine and time of day.

A second longstanding line of research originated with one of the first cognitive theories of trait expression, Charles Spielberger's (1979) theory of anxiety traits and states. Consistent with worry being the element of anxiety most harmful to attention and performance, Spielberger attributed anxiety effects to disruption of information-processing. More specifically, anxiety may be associated with diversion of the focus of attention from external to internal processing, with loss of attentional resources to processing personal concerns, or to loss of working memory space (Zeidner, 1998). The conclusion that anxiety disrupts attention is generally sound, but cognitive psychological research shows that 'attention' is not a unitary process. Prioritizing information, multitasking and sustaining attention over time all depend on multiple processes that may be differentially sensitive to personality.

The attentional control theory (ACT) of Eysenck and Derakshan (2011) focuses on the cognitive control or executive processes that afford voluntary control of attention. Various specific processes may be distinguished, including inhibition of task-irrelevant stimuli, switching between different task sets, and updating the contents of working memory. In general, anxiety effects on attentional tasks are somewhat inconsistent. Sometimes anxiety impairs performance, sometimes not. According to ACT, anxiety impairs inhibitory processes most strongly, so that anxiety is most damaging when the task requires the person to ignore distracting stimuli. Fine-grained information-processing accounts may help to explain the sometimes elusive effects of traits and emotional states on task performance.

Another important area of research concerns selective attention. In anxiety, and other emotional states, there appears to be a bias toward attending to congruent emotive stimuli (Cisler and Koster, 2010). That is, in anxiety, attention is drawn toward threats or sources of threat, even when the stimulus is irrelevant to the task. Bias is demonstrated with a version of the Stroop test, in which the person must name the ink colors of words that include threatening words such as murder and failure. Anxious individuals (patients and normal samples) are slow to name the colors of threat words, even though the task requires them to ignore word meaning. There appears to be a robust, moderate effect for threat-related attentional bias in anxiety, demonstrated using a variety of different experimental paradigms. The effect works for subliminal presentation of stimuli, where, on the Stroop test, for example, the person could not consciously identify the word stimulus. This finding suggests a bias in early encoding processes operating prior to conscious attention (although the bias may be modified by strategic factors). In ACT (Eysenck and Derakshan, 2011) threat-related bias is attributed not only to the attention-engaging effect of threat stimuli, but also to weak inhibitory control.

Studies show reliable associations between traits and information-processing, provided that care is taken with methods and tasks. However, questions remain about the broader significance of these findings. The major traits relate to a variety of individual differences in information-processing, which cannot be reduced to any single underlying mechanism (Matthews and Gilliland, 1999). The trait relates to a 'cognitive patterning' of typically modest processing biases (Matthews, 2008). It is debatable what such findings mean for the view of traits as coherent, unitary characteristics for which personality theory typically strives. This article finishes with two contrasting perspectives on the theoretical relevance of information-processing studies. First, from a 'bottom-up' perspective, variation in processing may be informative about individual differences in brain functioning that shape personality, as specified in contemporary cognitive neuroscience. Second, from a 'top-down' perspective, the performance data may be understood in the context of traits as adaptive factors.

Cognitive Neuroscience Perspectives

Traditionally, neurobiological theories of personality have referred to broad constructs such as arousal or sensitivity to reward and punishment. However, cognitive neuroscience suggests that a more fine-grained account of biological bases for personality may be attained. The neural networks that control attention are increasingly understood. For example, Michael Posner differentiated networks for executive control, spatial orienting, and alertness. Use of tasks informed by such theories affords inferences about individual differences in brain function, inferences that can be substantiated using methods including brain imaging. For example, one expression of the threat bias in anxiety is spatial orienting toward threat (see Cisler and Koster, 2010). Performance studies show that anxiety relates most reliably to one of the several attentional functions supported by the spatial orienting network: disengagement of attention from a location associated with threat. Brain-imaging studies localize the effect in the prefrontal cortex and functionally related structures. Other mechanisms contributing to observed affective biases can be triangulated similarly (Cisler and Koster, 2010). An additional step is to specify the dynamics of the neurocognitive architecture thus identified using connectionist models of processing.

Similarly, at least some extravert–introvert differences in performance may be attributed to dopaminergic functioning. Simplifying considerably, ascending dopaminergic afferents from the midbrain innervate a variety of structures including prefrontal areas that support executive control. Brain-imaging studies support the neurobiological hypothesis that extraversion is associated with greater activity in these pathways, in rewarding conditions (Kennis et al., 2013). To the extent that dopaminergic activity can be linked to functioning of the prefrontal cortex, we have explanations for extraversion effects on tasks dependent on this brain area, including working

memory and executive control tasks. It remains challenging, however, to show that observed performance effects are directly mediated by specific brain processes.

Thus, progress is being made in integrating neurobiological and cognitive psychological perspectives on personality traits. Nevertheless, there remains a concern that identifying traits with sets of discrete neurological biases may be missing the forest for the trees, especially in relation to explaining variation in the 'cognitive self' and the role of cognitive content. Neurobiological understanding of traits requires supplementation with psychological accounts that specify at a more abstract level how variation in self-content and process translates into different styles of self-regulation and adaptation to environmental challenges.

Cognitive-Adaptive Perspectives

Matthews (2008) has proposed a cognitive-adaptive theory of traits that poses the question of the functional significance of the cognitive correlates of traits. Extraverts tend to be superior in selective attention, verbal working memory, and speech fluency, but poorer than introverts at vigilance and reflective problem-solving. How does this package of cognitive characteristics influence extraverts' adaptation to social and physical environments? The answer may be that the cognitive patterning supports an adaptive specialization toward challenging social environments that extraverts are known to favor. However, the role of information-processing is indirect, in affecting the speed of acquisition of the contextualized skills needed for specific environments. The cognitive characteristics of extraversion facilitate the acquisition of conversational skills, for example, that allow the extravert to navigate through demanding scenarios such as high-stakes financial deal-making and standing out at a party. Conversely, the introvert, while vulnerable to social overload, has the advantage in contexts where social reinforcement is infrequent, such as working as a programmer.

A similar analysis can be applied to other traits. For example, level of neuroticism may represent the person's place on a spectrum of strategies for adapting to social threat, ranging from anticipation of threat and avoidance (high neuroticism) to stoic direct management (low neuroticism). As with extraversion–introversion, the characteristic strategies of those high and low in neuroticism may be effective in some social environments but not others. The cognitive-adaptive perspective allows cognitive correlates of traits to be understood in terms of underlying purpose, and so provides a means for integration findings on a wide range of trait attributes, at different levels of abstraction from the neural substrate.

See also: Agreeableness; Big Five Factor Model, Theory and Structure; Concepts and Methods of Cognitive Therapies; Conscientiousness; Coping, Personality and; Core Self-Evaluations; Emotion in Cognition; Extraversion; Implicit Association Test; Neuroticism; Openness to Experience; Personality and Self; Personality, Biological Models of; Personality, Trait Models of; Self-Efficacy; Self-Regulated Learning.

Bibliography

Allport, G.W., 1931. What is a trait of personality? Journal of Abnormal and Social Psychology 25, 368–372.
Beck, A.T., Clark, D.A., 1988. Anxiety and depression: an information processing perspective. Anxiety Research 1, 23–36.
Caprara, G.V., Cervone, D., 2008. Personality: Determinants, Dynamics and Potentials. Cambridge University Press, New York.
Cisler, J.M., Koster, E.H., 2010. Mechanisms of attentional biases towards threat in anxiety disorders: an integrative review. Clinical Psychology Review 30, 203–216.
Eysenck, H.J., Eysenck, M.W., 1985. Personality and Individual Differences. Plenum Press, New York.
Eysenck, M.W., Derakshan, N., 2011. New perspectives in attentional control theory. Personality and Individual Differences 50, 955–960.
Humphreys, M.S., Revelle, W., 1984. Personality, motivation and performance: a theory of the relationship between individual differences and information processing. Psychological Review 91, 153–184.
Judge, T.A., Van Vianen, A.E., De Pater, I.E., 2004. Emotional stability, core self-evaluations and job outcomes: a review of the evidence and an agenda for future research. Human Performance 17, 325–346.
Kennis, M., Rademaker, A.R., Geuze, E., 2013. Neural correlates of personality: an integrative review. Neuroscience and Biobehavioral Reviews 3, 73–95.
Lazarus, R.S., 1999. Stress and Emotion: A New Synthesis. Springer, New York.
Matthews, G., 2008. Personality and information processing: a cognitive-adaptive theory. In: Boyle, G.J., Matthews, G., Saklofske, D.H. (Eds.), Handbook of Personality Theory and Testing, Personality Theories and Models, vol. 1. Sage, Thousand Oaks, CA, pp. 56–79.
Matthews, G., Deary, I., Whiteman, M., 2009. Personality Traits, third ed. Cambridge University Press, Cambridge.
Matthews, G., Gilliland, K., 1999. The personality theories of H.J. Eysenck and J.A. Gray: a comparative review. Personality and Individual Differences 26, 583–626.
Robinson, M.D., Sedikides, C., 2009. Traits and the self: toward an integration. In: Corr, P.L., Matthews, G. (Eds.), Cambridge Handbook of Personality. Cambridge University Press, Cambridge, pp. 457–472.
Schnabel, K., Asendorpf, J.B., Greenwald, A.G., 2008. Using implicit association tests for the assessment of implicit personality self-concept. In: Boyle, G.J., Matthews, G., Saklofske, D.H. (Eds.), Handbook of Personality Theory and Testing, Personality Theories and Models, vol. 1. Sage, Thousand Oaks, CA, pp. 508–528.
Spielberger, C.D., 1979. Understanding Stress and Anxiety. Harper and Row, New York.
Wells, A., 2011. Metacognitive Therapy for Anxiety and Depression. Guilford Press, New York.
Wells, A., Matthews, G., 1994. Attention and Emotion: A Clinical Perspective. Lawrence Erlbaum, Hove, Sussex.
Zeidner, M., 1998. Test Anxiety: The State of the Art. Plenum, New York.

Personality Development and Temperament

Yiyuan Xu and Alexander W Krieg, University of Hawaii at Manoa, Honolulu, HI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.S. Kagan, volume 16, pp. 11287–11290, © 2001, Elsevier Ltd.

Abstract

This article first defines and compares the concepts of temperament and personality. Second, it discusses contextual variations of temperament and personality, and the distinction between continuous and categorical views of these two concepts. Third, the article focuses on development of temperament and how temperament contributes to development of personality. Finally, the possible influence of culture on temperament and personality development is discussed.

The Concepts of Temperament and Personality

Temperament can be defined as enduring individual differences in reactivity and self-regulation that appear early in life and have a presumed biological basis. Reactivity refers to individuals' responsiveness to changes in their external and internal environment, and is manifested in latency, duration, intensity, and frequency of motor, affective, and attentional reactions (Rothbart and Bates, 2006). For instance, infants vary in their response to a new moving object; some infants become aroused easily and thrash or cry within a few seconds while others show limited arousal to the same stimulation (Kagan and Snidman, 2004).

Self-regulation refers to neural and behavioral processes that function to modulate reactivity and is apparent in individual differences in effortful control, or the capacity to focus and shift attention, to inhibit a dominant response and/or to activate a subdominant response, to plan, and to detect errors (Rothbart and Bates, 2006). Some children are better than others in sustaining their attention, delaying gratification, regulating their emotional arousal, and suppressing and initiating activity to signals or instructions.

While temperament is limited to basic processes of reactivity and self-regulation in affect, attention, and activity, personality refers to a broader range of enduring individual differences in thoughts, feelings, and behavior and includes goals, motives, identities, habits, coping styles, values, beliefs, and other processes that are not typically considered aspects of temperament (McAdams and Pals, 2006). For instance, individual difference in the view of the self in relation to others represents an aspect of personality but not temperament.

Despite of the distinctions between temperament and personality, there are also similarities between these two domains of individual differences (Caspi and Shiner, 2008). Both temperament and personality traits appear to be partially but not uniquely heritable, and are shaped by maturation and experience. Both are relatively stable within similar contexts but are less consistent in distinct situations. The core aspects of temperament and personality, particularly during childhood, are similar; both involve individual variations in effect and self-regulation.

It is helpful to consider early individual differences as temperament, which then sets constraints on the later development of personality. Children's maturation and experience permits the development and expression of new aspects of personality that are not captured by early differences in temperament.

Temperament and Personality in Context

Temperament and personality both describe individual differences that are not continually expressed but require eliciting conditions (Rothbart and Bates, 2006). Temperament and personality may be viewed consisting of a pattern of 'if-stimulus-then-response' dispositions that are not expected to show situational invariance (Caspi and Roberts, 2001), but rather to reflect what (Murray, 1938) referred to as 'environmental press'; where a particular trait is elicited in some contexts but not in others, and a particular context may evoke behavioral expression of some but not other traits. For instance, highly reactive children are typically distressed by novelty, discrepancy in stimulation, or cues for punishment, but they may be at ease when interacting with siblings or familiar peers. Moreover, a given context will differ in its effects on individuals who differ in temperament or personality. Thus, going to a new place may be exciting for sociable children who enjoy and seek stimulation, but may be unpleasant for shy children who are afraid of novelty.

Continuous and Categorical Views

There is disagreement as to whether temperament or personality should be viewed as a continuous dimension or as a qualitative category (Kagan and Snidman, 2004; Shiner et al., 2001). While there appears to be a preference for the continuous view of temperament, the emphasis on dimensions that are shared by all individuals, could mask configurations of multiple traits within a particular individual that may represent a qualitatively unique temperament or personality category. An integration of both continuous and categorical approaches of temperament has been useful. In their pioneering work, Thomas and Chess (1977) first identified nine primary dimensions that can be used to rank order temperament in all children. Based on the within-individual configurations of these nine dimensions, they later categorized children into three types: difficult, easy, and slow to warm up. Children who

were classified as difficult were more likely than those classified as easy and slow to warm up to develop behavioral and psychological problems.

Development of Temperament

An infant is not born with a full complement of temperamental characteristics; many aspects of temperament emerge later due to a complex interplay among heredity, maturation, and environmental influence. While individual differences in reactivity emerge as early as in the first year of life (Fox et al., 2005), temperamental self-regulation, particularly effortful control, does not emerge until toddlerhood, and continues to develop in preschool years (Rothbart, 1989). The late developing nature of effortful control and its possible role of regulating reactivity may partly explain why many young children who were initially identified as 'highly reactive,' i.e., being fearful and inhibited, became less so or even uninhibited years later (Fox et al., 2005).

The development of effortful control is susceptible to not only maturation, but also individual experiences. For example, the integration of threatening or challenging environmental information into a child's mental representations may facilitate the development of effortful control. With increasing effortful control, the child can develop goal-oriented coping strategies to modulate negative emotions, to anticipate novelty or punishment cues by recognizing contextual information, and to evaluate the potential consequences of their actions and voluntarily decrease attention to unpleasant stimuli while shifting attention to relieving information (Rothbart and Bates, 2006). In addition, socialization, particularly parents' child rearing practices, facilitates the development of children's cognitive representation of what constitutes normative social standards and how personal actions may affect others' psychological experiences. Thus, effortful control increasingly coordinates with internalized goals, and operates to regulate a child's performance in a means–end fashion that is consistent with societal norms.

Given the evolving interaction among heredity, maturation, and environmental influences, temperament changes across age. Temperament during infancy appears to include three broad dimensions: two dimensions of reactivity, surgency/extraversion, and negative affectivity, and the third dimension of orienting/regulation. Surgency/extraversion is defined by approach tendency, activity level, simulation seeking, positive affect, and high intensity pleasure, whereas negative affectivity refers to negative emotions such as sadness, frustration, and fear. Orienting/regulation comprises of low intensity pleasure, cuddliness, duration of orienting, and soothability. While infants are viewed as largely reactive and lacking regulation, their reactive emotional systems do have self-regulatory components. For instance, negative affectivity may regulate/inhibit infants' approach of unfamiliar objects that may be potentially dangerous.

With the development of motor, cognitive, emotional, and language abilities, temperament in childhood and adulthood includes not only three broad temperament dimensions of surgency/sociability, negative affectivity and effortful control, but also a fourth dimension of affiliativeness. These four temperamental dimensions are remarkably similar to four of the Big Five factors, that is, extraversion, neuroticism, constraint/conscientiousness, openness, and agreeableness, found in studies of child and adult personality (Ahadi and Rothbart, 1994; Goldberg, 1993). For instance, surgency/sociability, defined mainly as positive emotionality, sociability, and activity level, resembles the personality factor of extroversion; negative affectivity, defined primarily by the expression of fear, frustration, and sadness, is similar to the personality factor of neuroticism; effortful control, which includes mainly attention regulation and volitional control, is similar to the personality factor of constraint/conscientiousness; affiliativeness has been characterized by prosocial tendencies including empathy or sympathy, and significantly overlaps with the personality trait of agreeableness. With the exception of openness to experience, which has been considered the fifth personality trait in the Big Five factors model, but not a temperamental dimension, there appears to be a strong one-to-one correspondence between temperamental and personality traits in childhood and adulthood (Rothbart et al., 2000). It should be noted that, however, while it may be possible to link affiliativeness to the personality factor of agreeableness, the temperament roots of affiliativeness remain speculative (Nigg, 2006).

Temperamental Contributions to Personality

The underlying biological basis of early temperament, such as the reactivity of the neural system, is relatively stable, and may lead to personality attributes that share similar biological processes. For instance, highly reactive infants, who were easily distressed by novel stimuli, were more likely than their peers to become fearful and inhibited between 24 and 36 months, and to develop anxious symptoms by school age (Kagan and Snidman, 2004). Even a reactive school age child who no longer shows inhibited behavior toward strangers, may still exhibit elevated heart rate during an encounter with a stranger, a physiological response characteristic of temperamental reactivity and the personality trait of shyness (Xu et al., 2009).

Temperament may also contribute to personality through a process of developmental elaboration by which temperamental attributes, as part of individual genetic heritage, accumulate response strength through repeated reinforcement and become elaborated into personality traits, or cognitive and affective representations that are quickly and frequently activated (Caspi and Shiner, 2008). This view emphasizes the transaction between the individual and the environment in maintaining and reinforcing individual differences in how contexts are selected, interpreted, and evoked, and thus producing coherent patterns of continuity between early temperament and later personality. For instance, an inhibited child may actively seek out a nonthreatening environment that does not require intensive social participation, or elicit parents' overly intrusive and protective behavior that inadvertently reinforces the child's shyness or social withdrawal over time (Rubin and Asendorpf, 1993).

While temperament may represent a starting point in development and provides a reaction range for later development, it does not determine the destiny for personality development. Temperamental contributions to personality are

clearly contingent upon socialization experiences, such as parenting and peer relationships. For instance, an early temperamental inclination to negative affectivity might predict later internalization only in children who had a responsive, nonpower assertive mother (Kochanska et al., 1997). Given the importance of socialization, it is not surprising that children with similar temperamental traits develop different personalities, whereas children with distinct temperament show similar personality later in life. Thus, only a small number of inhibited infants remain inhibited over time (Kagan and Snidman, 2004), and children who are initially inhibited could develop either anxiously shy behavior or modest and unassuming behavior, depending on children's effortful control that is susceptible to parenting styles (Zhou et al., 2004).

The Role of Culture

Temperament sets a range of opportunities for development, but culture defines the limit of what is desirable and normal within a particular setting. The transaction between the child and the environment may lead to different developmental outcomes, including personalities, contingent upon the 'goodness of fit' (Thomas and Chess, 1977) between the child's temperament and the environmental demands and challenges that are shaped by cultural values, norms, and socialization goals. Thus, American children with an inhibited temperament often develop anxiously shy behavior, whereas their Chinese peers may exhibit modest and unassuming behaviors that reflect Chinese cultural norms for interdependence and humility (Xu et al., 2009).

Within the goodness of fit framework, personality development is not simply viewed as individual histories of growth, but as forms of mutual adaptation, or the fit between the behavioral expression of early temperament, expectations of significant others such as parents or peers, and the limits of their tolerance for deviance from the behavioral norms (Lerner et al., 1989). Individual attributes that are most strongly associated with negative developmental outcomes are generally those that are inconsistent with local cultural norms.

See also: Five Factor Model of Personality, Facets of; Five Factor Model of Personality, Universality of; Genetics of Personality; Neonatal Temperament: From Phenomenology to Mechanism; Personality Assessment: Overview; Personality Development: Systems Theories; Personality and Educational Outcomes; Temperament and Motivation; Temperament.

Bibliography

Ahadi, S.A., Rothbart, M.K., 1994. Temperament, development, and the Big Five. In: Halverson Jr., C.F., Kohnstamm, G.A., Martin, R.P. (Eds.), The Developing Structure of Temperament and Personality from Infancy to Adulthood. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, England, pp. 189–207.

Caspi, A., Roberts, B.W., 2001. Target article: personality development across the life course: the argument for change and continuity. Psychological Inquiry 12 (2), 49–66. http://dx.doi.org/10.1207/s15327965pli1202_01.

Caspi, A., Shiner, R., 2008. Temperament and personality. In: Rutter, M., Bishop, D., Pine, D., Scott, S., Stevenson, J., Taylor, E., Thapar, A. (Eds.), Rutter's Child and Adolescent Psychiatry, fifth ed. Wiley-Blackwell, pp. 182–198.

Fox, N.A., Henderson, H.A., Marshall, P.J., Nichols, K.E., Ghera, M.M., 2005. Behavioral inhibition: linking biology and behavior within a developmental framework. Annual Review of Psychology 56, 235–262. http://dx.doi.org/10.1146/annurev.psych.55.090902.141532.

Goldberg, L.R., 1993. The structure of phenotypic personality traits. American Psychologist 48 (1), 26–34. http://dx.doi.org/10.1037/0003-066x.48.1.26.

Kagan, J., Snidman, N., 2004. The Long Shadow of Temperament. Belknap Press/Harvard University Press, Cambridge, MA.

Kochanska, G., Murray, K., Coy, K.C., 1997. Inhibitory control as a contributor to conscience in childhood: from toddler to early school age. Child Development 68 (2), 263–277. http://dx.doi.org/10.2307/1131849.

Lerner, J.V., Nitz, K., Talwar, R., Lerner, R.M., 1989. On the functional significance of temperamental individuality: a developmental contextual view of the concept of goodness of fit. In: Kohnstamm, G.A., Bates, J.E., Rothbart, M.K. (Eds.), Temperament in Childhood. John Wiley & Sons, Oxford, England, pp. 509–522.

McAdams, D.P., Pals, J.L., 2006. A new Big Five: fundamental principles for an integrative science of personality. American Psychologist 61 (3), 204–217. http://dx.doi.org/10.1037/0003-066x.61.3.204.

Murray, H.A., 1938. Explorations in Personality. Oxford University Press, Oxford, England.

Nigg, J.T., 2006. Temperament and developmental psychopathology. Journal of Child Psychology and Psychiatry 47 (3–4), 395–422. http://dx.doi.org/10.1111/j.1469-7610.2006.01612.x.

Rothbart, M.K., 1989. Temperament and development. In: Kohnstamm, G.A., Bates, J.E., Rothbart, M.K. (Eds.), Temperament in Childhood. John Wiley & Sons, Oxford, England, pp. 187–247.

Rothbart, M.K., Ahadi, S.A., Evans, D.E., 2000. Temperament and personality: origins and outcomes. Journal of Personality and Social Psychology 78 (1), 122–135. http://dx.doi.org/10.1037/0022-3514.78.1.122.

Rothbart, M.K., Bates, J.E., 2006. Temperament. In: Eisenberg, N., Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology: Vol. 3, Social, Emotional, and Personality Development, sixth ed. John Wiley & Sons Inc, Hoboken, NJ, pp. 99–166.

Rubin, K.H., Asendorpf, J.B., 1993. Social Withdrawal, Inhibition, and Shyness in Childhood. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, England.

Shiner, R.L., Tellegen, A., Masten, A.S., 2001. Exploring personality across childhood into adulthood: can one describe and predict a moving target? Psychological Inquiry 12 (2), 96–100.

Thomas, A., Chess, S., 1977. Temperament and Development. Brunner/Mazel, Oxford, England.

Xu, Y., Farver, J.A.M., Yu, L., Zhang, Z., 2009. Three types of shyness in Chinese children and the relation to effortful control. Journal of Personality and Social Psychology 97 (6), 1061–1073. http://dx.doi.org/10.1037/a0016576.

Zhou, Q., Eisenberg, N., Wang, Y., Reiser, M., 2004. Chinese Children's effortful control and dispositional anger/frustration: relations to parenting styles and children's social functioning. Developmental Psychology 40 (3), 352–366. http://dx.doi.org/10.1037/0012-1649.40.3.352.

Personality Development: Systems Theories

Katarína Millová and Marek Blatný, Institute of Psychology of Academy of Sciences of the Czech Republic, Brno, Czech Republic

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Developmental systems theories are an important part of modern research of human development. These theories see individuals as a part of a developmental system. The system is composed of the individual and also of the environment. The theories emphasize the holistic principle in research and reject any form of reductionism. Developmental systems theories include dynamic systems theory, developmental contextualism, person–context interaction theory, and holistic, developmental, systems-oriented perspective. Unlike the dynamical systems theory, which is a mathematical construct, the *dynamic* systems theory of Esther Thelen and Linda Smith is primarily nonmathematical and comprises qualitative theoretical propositions in behavioral biology and is the broadest and most encompassing of all development theories.

Systems Theories of Personality Development

Developmental systems theories are an important part of modern research of human development. Their philosophy follows up on the general system theory by Ludwig von Bertalanffy (1968) and is applied to the area of life-span development. They see individuals as being part of a developmental system in their life course. The system is not only composed of the individuals alone but also of the environment in which they develop. These theories emphasize the holistic principle in research and reject any form of reductionism and the traditional input–output model. The system is considered as a whole, which must be examined inseparably, i.e., we examine the individual as well as the environment in which they develop in a gestalt conceptualization. The whole system cannot be broken down into parts; therefore persons cannot be examined in isolation, separately from their environment. If we did that, we would obtain biased and incomplete results. Instead of isolated characteristics (e.g., cognitive development, social development) the theories explore the relationship between system components. By emphasizing the holistic principle they follow the ideas of W. James and J. Dewey, as well as W. Stern, K. Lewin, and G.W. Allport.

In an effort to ensure the most complex possible view of human development they place an emphasis on multidisciplinary collaboration using knowledge not only from developmental psychology but also from other scientific disciplines, such as sociology, evolutionary biology, medicine, and anthropology. Developmental systems theories include dynamic systems theory (Thelen and Smith, 2006), developmental contextualism (Lerner, 1991), person–context interaction theory (Magnusson and Stattin, 2006), and holistic, developmental, systems-oriented perspective (Wapner and Demick, 2003). However, we can find developmental systems perspectives also in other modern theories of development (e.g., Bronfenbrenner, Elder, Brandtstädter).

Probabilistic Epigenesis

The work of Gilbert Gottlieb (1992, 1998) who introduced the concept of developmental systems whereby the changing gene–context or organism–context *relations* are of key importance became an inspiration and a great contribution to the elaboration of systems theories. He applied his theory both in the development and in the evolution analyses. He was inspired by the works of Garstang (1922), de Beer (1958), and Goldschmidt (1933) and points out that the phylogenesis is not the cause but the product of a succession of various ontogeneses, i.e., that variations in the development produce evolution, while evolution does not produce development. "Ontogeny in each generation is a consequence of the coaction of hereditary or genetic factors and many different local environmental circumstances that determine the expression of the phenotype during the course of development" (Gottlieb, 1992, p. 95).

Gottlieb saw the development process as an interrelated, fully coactional system in which the activity of genes can be affected through the cell cytoplasm by events occurring at any other level of the system, including the external environment. He proposed a new definition of epigenesis: "Individual development is characterized by an increase of complexity of organization – i.e., the emergence of new structural and functional properties and competencies – at all levels of analysis (molecular, subcellular, cellular, organismic) as consequence of horizontal and vertical coactions among its parts, including organism–environment coactions" (Gottlieb, 1992, pp. 159–160).

Horizontal coactions are those taking place at the same level (gene–gene, cell–cell, tissue–tissue), whereas vertical coactions occur at different levels (cell–tissue, behavioral activity–nervous system) and are reciprocal. Genes, which are part of the development system like other components (i.e., cells, tissue, organism), are susceptible to influence from other levels in the course of individual development.

To support his theory, Gottlieb gathered substantial evidence proving that the action of genes (gene expression) was influenced by events at other levels of the development system and that all levels of the system could thus be considered as potentially equal. Genetic activity in itself does not produce the final state of the anatomic and physiological differentiation but requires further intervention from the level above cellular level.

Thus, Gottlieb's developmental systems framework indicates that all organism characteristics (genes, cells, organs) as well as the whole organism operate in bidirectional or

reciprocal 'dynamic interaction' relationship with the context in which the organism is embedded. If the time dimension is added, the result is a multilevel, integrated operation that determines the individual course of development.

Dynamic Systems Theory

Esther Thelen and Linda Smith (2006) derived their dynamic systems theory both from studying the complex and nonlinear systems in physics and mathematics and from the traditions of systematic thinking in biology and psychology. In the most general sense the term 'dynamic systems' refers to systems of elements that change over time. In a narrower, more technical sense, the term 'dynamical systems' refers to a class of mathematical equations that describe time-based systems with particular properties. Unlike the dynamical systems theory, which is a mathematical construct, the dynamic systems theory of Esther Thelen and Linda Smith is primarily nonmathematical and comprises qualitative theoretical propositions in behavioral biology and is the broadest and most encompassing of all development theories.

The theory is founded on two central ideas that are expressed not only in the dynamic systems theory itself but also in the history of research into the development process:

1. Development can only be understood as a multiple, mutual, and continuous interaction of all levels of the developing system, from the molecular to the cultural.
2. Development can only be understood as nested processes that unfold over many timescales from milliseconds to years (Thelen and Smith, 2006, p. 258). Developmental systems theory can be seen as a metatheory, as it can be applied to various animal species, age levels, and areas of development. Equally, it can be used as a specific theory that deals with how people acquire knowledge from their daily activities.

According to Thelen and Smith, the defining feature of development is the increasing complexity of the organism, namely the creation of new forms. Qualitative discontinuity, emergence, epigenesis, some kind of novelty are the essence of developmental changes. Each of developmental transitions involves the emergency of new patterns of behavior from precursors that themselves do not contain those patterns.

In the quest for answers about the sources of developmental change Thelen and Smith reject the 'nature and nurture' explanation but also the interaction position/view. They presume the tradition of systems theories of biological organization that explains the creation of new forms through the self-organization process. By self-organization Thelen and Smith (2006, p. 259) mean a process where the pattern or order emerges from interactions among components of a complex system without explicit instructions from the organism itself or from the environment.

The system levels are conceptualized as more than merely interacting with each other; instead they are seen as integrally fused together. The behavior and its development are interrelated as an ever-changing set of relations and the history of these relations over time (Thelen and Smith, 2006, p. 266). Due to this fusion Thelen and Smith replace the linear system of causality with the nonlinear system in which bidirectional relations within and across the fused levels of organization mutually conditionally change over time.

A key feature of the dynamic system is that many heterogeneous system components (cells, tissues, and organs within the individual and the various aspects of an individual's external context) can be freely combined in a virtually unlimited number of ways. In reality, however, the number of relational patterns is smaller. Order emerges from the system complexity through the process of self-organization. Order emerges from the disparate/heterogeneous parts because the human development is an open system, i.e., a system where energy is consumed by the system and used to establish order within it.

Such integration enables the description of the system by using a smaller number of dimensions or parameters. Such integrative variables that emerge in the open system with the aim to reduce its dispersion and to increase its organization are referred to as the collective variables or order parameters.

The emergence of such collective variables provides for continuity and stability within the system. The system settles into or prefers only a few behavior modes. In the dynamic terminology this behavioral mode is an attractor state, as the system has an affinity to this state (2006, p. 272). But even highly stable attractors are dynamic and changeable. Thus, the system – even though it may seem as such – is not fixed, hardwired or programmed, rather it is softly assembled.

The existence of soft assembly means that the development of a person includes a dynamic link between the system stability on the one hand and the fluctuations around the stability states, i.e., changes that are a source of functional novelty within the system, on the other. The system is reiterative, i.e., each state is dependent on the previous state and this reiterative process occurs at all timescales.

Developmental Contextualism

The theory by Richard M. Lerner (1991, 1998) has its roots not only in nonreductionist systematic theories of biological evolution (e.g., von Bertalanffy, 1968; Gottlieb, 1992) but also in sociologically oriented theories concerned with the influence of social and historical changes on human development (Elder and Giele, 2009). From the perspective of development contextualism, development is a result of changes in the entire development system. It is based on the changing relationships between different levels of organization. These consist of the organism and its biological characteristics that exist in a particular sociocultural and historical environment (Lerner, 1991). The developmental system is (1) plastic and open to change, and (2) the relationship between different levels of the system is reciprocal and, at the same time, they are integrated. Each developmental system is (3) hierarchically organized and embedded in a particular historical period. All systems have (4) a certain degree of generalizability, diversity, and individual differences (Lerner, 1998). All of the described components are closely linked. When examining their development, we cannot view them separately because they are interdependent. This is why Lerner in his theory of developmental contextualism emphasizes an integrated, multidisciplinary approach to the study of life-span development.

Another key idea is the inseparability of the organism from the environment. Individuals and their environment are intertwined in the life course. Therefore, we should not examine them separately because human development always occurs in a specific environment. Developmental contextualism applies these general findings to individuals who live and develop in a variety of family, social, cultural, and historical conditions (Lerner, 1991). The developmental model takes into account individual characteristics (e.g., relations between parent and child) and the characteristics of the society or culture that form the largest part of the environment where the development takes place. Developmental contextualism rejects any dichotomy in developmental theories and social sciences (e.g., nature vs nature, person vs context). On the contrary, it emphasizes their integration (Lerner, 1998).

Holistic–Interactionistic Approach

David Magnusson (1999) in his theory puts an emphasis on the integration of an individual and the holistic approach to examining development. He was not only influenced by other developmental systems theories but also by Gestalt psychology, for example. According to the holistic interactionism human development is based on four pillars: interactionism, holism, interdisciplinarity, and longitudinal study. A person does not exist in the environment merely as a passive object; on the contrary, the individual rather exists as an active and conscious part of the context system. Within the system a person develops and operates as an integrated, comprehensive and dynamic unit. We cannot understand the functioning of the environment without taking into consideration the person in it. At the same time, it is not possible to understand an individual's functioning and development without understanding the environment (Magnusson and Stattin, 2006).

The holistic perspective of development is a consequence of gaps in the research so far. Criticism applies especially to fragmentation (excessive specialization without integration), emphasis on variables, prediction (the principal objective of examining personality), unidirectional causality, prevalence of statistics over analysis of the phenomenon and an insufficient link between theory and empirical research. On the other hand, the holistic view describes personality as a living, active, and purposeful organism, functioning and evolving as a comprehensive integrated being. The modern holistic interactionism follows the classical interactionism, an approach that emerged in the 1970s (Endler and Magnusson, 1976).

According to the interaction theory the developmental system functions as a whole. The features of the totality are a result of interactions between all parts of the system. Simpler processes taking place in the system (such as emotional changes) take less time than more complex processes (such as changes in the relationship between the person and the environment). In the life course the system undergoes constant restructuring at all levels. Any process taking place within the system, however, is directed by certain laws. Activities of all parts of the system are coordinated in such a way that each subsystem works toward achieving the goals of the whole system. In the context of a developing system the holistic interactionism describes the principle of so-called amplification of minimal effects. This principle is related to the long-term influence of marginal deviations from the norms of behavior or appearance of a person. Such deviations can cause reactions from the proximal social environment, for example, social isolation in the group as a result of an aggressive behavior. These reactions from the environment may be one of the sources of recurring malfunction in the interaction between an individual and the environment, which can potentiate the maladaptive behavior to a pathological level (Magnusson and Stattin, 2006).

We can analyze the human behavior and experience from three different perspectives: current (synchronic), developmental (diachronic), and evolutionary. The holistic interactionism focuses on the first two perspectives: current and developmental. The model for current functioning is concerned with why people behave in a certain way, in terms of their current psychological and biological dispositions. Developmental models explain the current functioning on the basis of individual developmental history (Magnusson, 1999). In order to understand the current functioning completely we need to understand the conditions of development.

Environment is a key term in models that deal with the current functioning and development. The holistic interactionism sees the environment as a holistic, integrated, and organized system hierarchically structured (Magnusson and Stattin, 2006). However, the objective environment is not so important; it is more important how it is perceived and interpreted. Individual perception and interpretation of the external world is shaped by the actual (objective) environment. This approach does not evaluate an individual and the environment as isolated elements. It explores the relationship between people and their environment, which they helped to create to some extent (Magnusson, 1999). In the integrated person–context system the person acts as an active agent. This does not mean that we cannot examine its specific characteristics, such as aspiration level, physical performance, or current genesis of emotion. We should, however, include them in the overall functioning of the system and take into account their role in its activity (Magnusson and Stattin, 2006).

Holistic, Developmental, Systems-Oriented Perspective

The theory developed by Seymour Wapner (2000) connects organismic and contextual approach for examining human development. It integrates individuals and their environment into a single unit which it calls the person-in-environment system. It is based on extending the theory of Heinz Werner to life-span development (Valsiner, 2005). In it we find inspiration by the developmental task theory by R.J. Havighurst (1972). It rejects any reductionism, particularly biological. Individual levels of the system differ in their qualities and cannot be reduced to other levels. This approach focuses not only on the development of the child but also on the development in the adult age, social relationships, health problems, changes in the physical environment, cross-cultural psychology, psychopathology, and psychology of organizations (e.g., Wapner, 2000; Wapner and Demick, 2003; etc.).

The three key areas of theory are the holistic approach, assessing development as progressive and regressive changes

in which the orthogenetic principle applies (that describes development as a change from a dedifferentiated to differentiated and hierarchically integrated state of the system; Wapner, 2000), and focus on the entire person-in-environment system. The theory further focuses on the relationship between an individual and an environment (transactionalism) and the active operation of the system in making its experience with the environment (constructivism). It also deals with the diversity of the system, the nature of time and space experience, it takes into account the planning processes and focuses both on the description (structural analysis), and the process (dynamic analysis). Although the approach is distinctly macroscopic, it considers specific aspects of a person and an environment. It describes three personal characteristics: (1) physical/biological (e.g., health, physical performance), (2) intrapsychological (e.g., self-respect, self-esteem, stress), and (3) social and cultural (e.g., adoption of a child, social role). Similarly, the environment contains three characteristics: (1) physical (e.g., natural disasters, buildings), (2) interpersonal (e.g., the nature and development of friendship), and (3) social and cultural (e.g., rules, laws) (Wapner and Demick, 2003).

The process of development takes place in such way that the system configuration changes. As a person gains experience and new skills, the structure and function of the system changes. With new experience and skills the system differentiates (becomes more complex), becomes arranged hierarchically and, at the same time, is being integrated. Some system changes are almost imperceptible; others are very important and crucial for development. This theory deals only with those transitions that are critical for an individual. Critical life transitions are of key importance, as they are an opportunity for a change in the behavior and experience of the individual. They come at any age. Many events are significant subjectively, so what is critical for one person is not necessarily critical for another. On the other hand, there are situations, which generally form the basis for critical transitions. These are changes such as the birth of the first child, divorce, adoption, functioning of the family, loss of a partner, retirement or going back to school (Wapner, 2000).

Future Directions

An important initiative of developmental systems theories authors is an effort to interlink knowledge at the level of theory and empirical research but also to find their use in social policies and programs (e.g., Lerner, 1998; Magnusson and Stattin, 2006). These researchers also promote the extending of research boundaries into human development because nowadays the knowledge limited to developmental psychology is not sufficient for a complete understanding and explanation of the developmental process. It is crucial to obtain knowledge in other sciences, such as evolutionary biology, developmental psychology, physiology, anthropology, social psychology, and other related disciplines that make up the so-called developmental science (Magnusson, 1999; Hartup, 2002). A considerable expansion of developmental science took place in the mid-1990s and it is no coincidence that many researchers exploring developmental systems are notable promoters of these ideas. They are those who promote the need for multidisciplinary research. It is only natural that a systematic research into the theoretical and empirical levels of developmental science is associated with a remarkable progress in the quantitative statistical approach, which is connected mainly with the longitudinal methods. They are important for assessing the changing relationship between people and their environment. In addition, the importance of qualitative methods is growing as a valuable analytical tool for assessing an individual life course and as a means of triangulation of the quantitative evaluation of human development. In connection with progress in human development research a growing number of scientists prefer to use the term developmental science instead of developmental psychology (Lerner, 2002; Damon and Lerner, 2006; Lerner et al., 2011).

See also: Cross-Cultural Research Methods in Psychology; Motivation: Life Course and Sociological Perspectives.

Bibliography

de Beer, Gavin R., 1958. Embryos and Ancestors, third ed. Clarendon Press, Oxford.
von Bertalanffy, Ludwig, 1968. Organismic Psychology and Systems Theory. Clark University Press, Worchester.
Damon, William, Lerner, Richard M. (Eds.), 2006. Handbook of Child Psychology Theoretical Models of Human Development, sixth ed., vol. 1. Wiley, Hoboken.
Elder Jr., Glen H., Giele, Janet Z., 2009. Life course studies: an evolving field. In: Elder Jr., , Glen, H., Giele, Janet Z. (Eds.), The Craft of Life Course Research. Guilford Publications, New York, pp. 1–24.
Endler, Norman S., Magnusson, David, 1976. Toward an interactional psychology of personality. Psychological Bulletin 83, 956–974.
Garstang, Walter, 1922. The theory of recapitulation: a critical re-statement of the biogenetic law. Journal of the Linnean Society of London Zoology 35, 81–101.
Goldschmidt, Richard, 1933. Some aspects of evolution. Science 78, 539–547.
Gottlieb, Gilbert, 1992. Individual Development and Evolution: The Genesis of Novel Behavior. Oxford University Press, New York.
Gottlieb, Gilbert, 1998. Normally occurring environmental and behavioural influences on gene activity: from central dogma to probabilistic epigenesis. Psychological Review 105, 792–802.
Hartup, W.W., 2002. Growing points in developmental science: a summing up. In: Hartup, W.W., Silbereisen, R.K. (Eds.), Growing Points in Developmental Science: An Introduction. Psychology Press, Hove and New York, pp. 329–344.
Havighurst, Robert J., 1972. Developmental Tasks and Education, third ed. McKay, New York.
Lerner, Richard M., 1991. Changing organism–context relations as the basic process of development: a developmental contextual perspective. Developmental Psychology 27, 27–32.
Lerner, Richard M., 1998. Theories of human development: contemporary perspectives. In: Damon, William, Lerner, Richard M. (Eds.), Handbook of Child Psychology, Theoretical Models of Human Development, fifth ed., vol. 1. Wiley, New York, pp. 1–24.
Lerner, Richard M., 2002. Concept and Theories of Human Development, third ed. Lawrence Erlbaum Associates, Mahwah.
Lerner, Richard M., Lewin-Bizan, Selva, Alberts Warren, Amy Eva, 2011. Concepts and theories of human development. In: Bornstein, Marc H., Lamb, Michael E. (Eds.), Cognitive Development. Psychology Press, New York, pp. 19–65.
Magnusson, David, 1999. Holistic interactionism: a perspective for research on personality development. In: Pervin, Lawrence A., John, Oliver P. (Eds.), Handbook of Personality: Theory and Research, second ed. Guilford Press, New York, pp. 219–249.
Magnusson, David, Stattin, Håkan, 2006. The person in context: a holistic–interactionistic approach. In: Damon, William, Lerner, Richard M. (Eds.), Handbook of Child Psychology, Theoretical Models of Human Development, sixth ed., vol. 1. Wiley, Hoboken, pp. 400–464.
Thelen, Ester, Smith, Linda B., 2006. Dynamic systems theories. In: Damon, William, Lerner, Richard M. (Eds.), Handbook of Child Psychology, Theoretical Models of Human Development, sixth ed., vol. 1. Wiley, Hoboken, pp. 258–312.

Valsiner, Jaan (Ed.), 2005. Heinz Werner and Developmental Science. Kluwer Academic, New York.

Wapner, Seymour, 2000. Person-in-environment transitions: developmental analysis. Journal of Adult Development 7, 7–22.

Wapner, Seymour, Demick, Jack, 2003. Adult development: the holistic, developmental and systems-oriented perspective. In: Demick, Jack, Andreoletti, Carrie (Eds.), Handbook of Adult Development. Kluwer Academic, New York, pp. 63–84.

Personality Differences and Development: Genetic and Environmental Contributions

Christian Kandler, Bielefeld University, Bielefeld, Germany
Wiebke Bleidorn, Tilburg University, Tilburg, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the twenty-first century, behavioral genetic research has broadened our knowledge about the origins of personality differences and development. On average, genetic factors account for more than 50% of the variance in accurate measures of personality traits. However, heritability estimates of personality traits steadily decrease with age. Genetic factors represent the primary source of long-term continuity of individual differences in personality, but also account for change – particularly in younger ages. On the other hand, environmental factors represent the primary source of personality change in every period of life, but also contribute to the relatively high stability of personality differences throughout the adult life span.

Introduction

Behavioral genetic research on the nature and nurture of personality differences and development has made important progress during the last decades. This review article presents major findings on the influences of genetic and environmental factors on individual differences, stability, and change in personality traits. For that purpose, we first give a short definition of personality traits. We then present an overview of findings from recent behavioral genetic studies on individual differences in personality traits. Third, we review findings from longitudinal behavioral genetic studies on personality development in different periods of life. In addition, we illustrate how these findings can inform different theories on the nature and the nurture of personality development. Finally, we outline implications and provide an outlook on present trends in the fields of quantitative and molecular genetics.

Personality Traits

In this review article, we focus on traits as basic characteristics for the description of personality differences. Personality traits are commonly defined as relatively stable patterns of thoughts, feelings, and actions in which each individual is different from the others. That is, personality traits are fairly stable over time and across situations but still open to change. Personality traits primarily represent economic units of analysis of individual differences in stylistic and regulatory aspects of cognitions, emotions, motivations, and behavior. They do not include broad-sense individual differences, such as physical features (e.g., attractiveness) or abilities (e.g., intelligence).

The Big Five trait taxonomy is the most influential conceptual model that captures personality traits along five dimensions of individual differences (John et al., 2008). Although the five dimensions have been labeled in many different ways, a common characterization employs the following labels: (1) *neuroticism* versus *emotional stability* or *negative emotionality*; (2) *extraversion* versus *introversion* or *positive emotionality*; (3) *openness to experiences* or *intellect*; (4) *agreeableness*; and (5) *conscientiousness* or *constraint*. The Big Five trait taxonomy appears to capture the trait landscape as portrayed in many different languages, societies, and cultures (McCrae and Costa, 2008). It is useful to integrate various systems and major inventories of personality description in terms of temperament and personality traits in a common framework (John et al., 2008). Each of the five traits hierarchically subsumes a set of more specific traits (e.g., anxiety, impulsiveness, gregariousness, activity, altruism, straightforwardness, dutifulness, achievement striving, and openness to feelings or ideas). This hierarchical model is called five-factor model (FFM) of personality.

The Big Five (or FFM) traits predict several specific behaviors, such as tobacco consumption, as well as important life outcomes, such as occupational success, divorce, or even mortality (Roberts et al., 2007). Therefore, the Big Five traits are typically conceptualized as core characteristics – the essential (i.e., the genetically anchored) basis of personality (McAdams and Pals, 2006; McCrae and Costa, 2008).

Genetic and Environmental Influences on Personality Differences

According to the theory that Big Five personality traits reflect genetically anchored core characteristics of personality, several studies have found that individual differences in personality traits are substantially genetically influenced (Johnson et al., 2008). Moreover, the hierarchical structure of the five-factor model "has a solid biological basis and may represent a common heritage of the human species" (Yamagata et al., 2006: p. 987). In their meta-analysis, Johnson et al. (2008) reported that genetic factors accounted for about 50% of the variance in primarily self-rated Big Five personality traits. The remaining variance was primarily due to environmental influences that were not shared by individuals and by family members raised within the same family (see **Figure 1**).

After correction for variance due to random error of measurement and rater-specific perspectives (e.g., self-perspective or well-informed peers), estimates of heritability for personality traits (i.e., estimates of the degree to which individual differences in traits are due to genetic differences) are often considerably higher (Kandler et al., 2010b; Riemann

 http://dx.doi.org/10.1016/B978-0-08-097086-8.25011-3

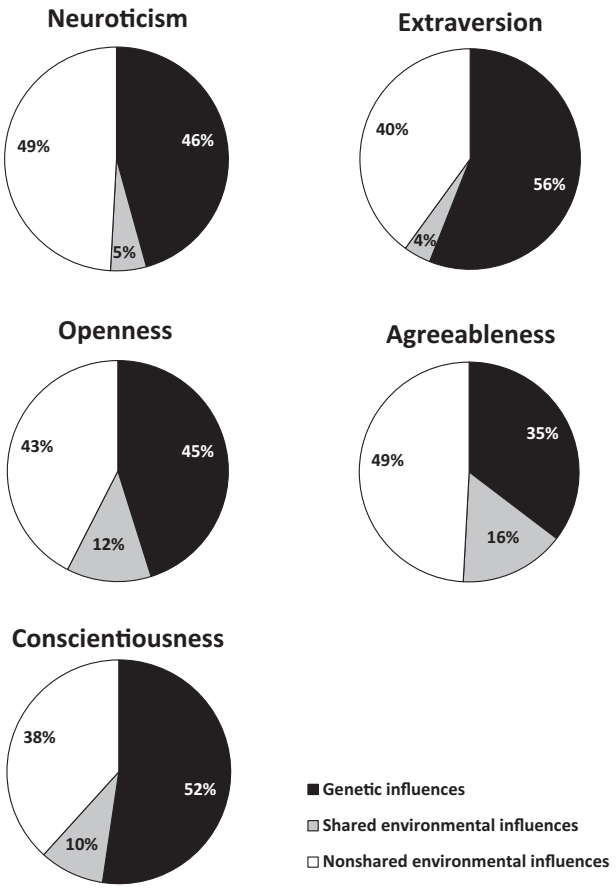

Figure 1 Standardized estimates of genetic and environmental influences (in percent) on individual differences in Big Five personality traits based on findings from the meta-analysis by Johnson, A.M., Vernon, P.A., Feiler, A.R., 2008. Behavioral genetic studies of personality: an introduction and review of the results of 50+ years of research. In: Boyle, G.J., Matthews, G., Saklofske, D.H. (Eds.), The SAGE Handbook of Personality Theory and Assessment, vol. 1. SAGE, London, pp. 145–173. Personality Theories and Models. Estimates of genetic influences include both additive and nonadditive genetic factors, whereas environmental influences are disentangled into factors shared and not shared by family members raised within the same family.

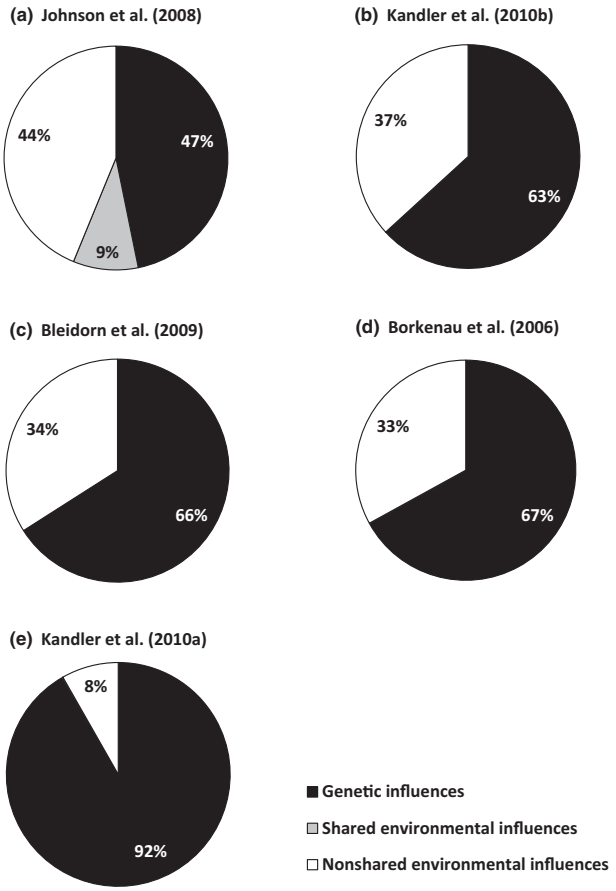

Figure 2 Standardized estimates of genetic and environmental influences (in percent) on individual differences in personality traits averaged across the Big Five traits. Estimates are based on (a) the meta-analysis by Johnson, A.M., Vernon, P.A., Feiler, A.R., 2008. Behavioral genetic studies of personality: an introduction and review of the results of 50+ years of research. In: Boyle, G.J., Matthews, G., Saklofske, D.H. (Eds.), The SAGE Handbook of Personality Theory and Assessment, vol. 1. SAGE, London, pp. 145–173. Personality Theories and Models, in which more than 90% of considered studies on adult personality relied on self-reports; (b) the multiple-rater twin study by Kandler, C., Riemann, R., Spinath, F.M., Angleitner, A., 2010b. Sources of variance in personality facets: a multiple-rater twin study of self-peer, peer-peer, and self-self (dis-) agreement. Journal of Personality 78, 1565–1594; (c) the longitudinal twin study by Bleidorn, W., Klimstra, T.A., Denissen, J.J.A., Rentfrow, P.J., Potter, J., Gosling, S.D., 2013. Personality maturation around the world: a cross-cultural examination of social-investment theory. Psychological Science. http://dx.doi.org/10.1177/0956797613498396 published online; (d) the twin study of Person × Situation profiles by Borkenau, P., Riemann, R., Spinath, F.M., Angleitner, A., 2006. Genetic and environmental influences on person × situation profiles. Journal of Personality 74, 1451–1479; and (e) the longitudinal multiple-rater twin study by Kandler, C., Bleidorn, W., Riemann, R., Spinath, F.M., Thiel, W., Angleitner, A., 2010a. Sources of cumulative continuity in personality: a longitudinal multiple-rater twin study. Journal of Personality and Social Psychology 98, 995–1008.

and Kandler, 2010). Genetics factors accounted for more than 50% of individual differences in these more accurate personality trait scores (compare **Figure 2**(a) with **2(b)**). In a similar vein, the proportion of individual differences in personality traits that was stable over time (Bleidorn et al., 2009) and across situations (Borkenau et al., 2006) showed larger genetic influences that accounted for about two-thirds of the variance in consistent trait scores (see **Figure 2(c) and 2(d)**). Moreover, after correction for random and nonrandom error of measurement as well as occasion-specific and short-term stable influences (<7 years), genetic factors accounted for nearly 100% of the variance in long-term stable (>12 years) personality trait scores (see **Figure 2(e)**), particularly in young adulthood (Kandler et al., 2010a). These findings are in line with the conceptualization of Big Five personality traits as genetically anchored core characteristics of personality.

Genetic and Environmental Influences on Personality Development

Although individual differences in the Big Five traits are fairly stable over the adult life span, recent studies have found

a reversely U-shaped pattern for the degree of stability of individual differences (i.e., rank–order continuity) from young to old age (e.g., Lucas and Donnellan, 2011). That is, the rank–order continuity tends to be relatively low in childhood (4-year test–retest correlations < 0.60), but rises across adolescence and young adulthood reaching a plateau between the ages of 40 and 70 (4-year test–retest correlations range between 0.60 and 0.80). Beyond age 70, the rank–order continuity appears to decay, again (4-year test–retest correlations < 0.60). Moreover, the continuity declines as the time interval between measurements of personality increases. Most longitudinal studies with at least four times of measurement suggest a gradual decline in rank–order continuity (Fraley and Roberts, 2005). However, the rank–order continuity does not decline to zero but reaches a positive asymptote of about 0.40 (Fraley and Roberts, 2005). These patterns of stability (or instability) of individual differences in personality traits over time suggest that (1) personality can change at every age throughout the life span and (2) the most pronounced changes occur during young and old ages. But what are the sources of these changes?

Genetic Set-Points and Environmental Fluctuation

The classic genetic set-point theory that is also known as the dynamic equilibrium model (see Ormel et al., 2012 for a review) attributes long-term rank–order continuity of personality traits to individual differences in genetically driven immutable individual set-points. Environmental influences are assumed to represent situational or short-term stable factors affecting reversible changes or fluctuations around these set-points. That is, on the long run, individuals' personality traits will always regress to their enduring set-points.

According to this theory, an enhanced frequency of environmental fluctuations can be expected during early and late adulthood that can explain the reduced rank–order continuity at these ages (see **Figure 3**). In fact, adolescents and young adults experience more positive and negative life events than do middle-aged people (Kandler et al., 2012). Life events are linked to major changes in life circumstances and social roles (e.g., moves, several graduations from school, apprenticeship, or university, and start an own family). These changes in life circumstances may be associated with reversible changes in personality traits. Life events may change the daily routine of individuals and require new behavioral responses and individual adaptations. Older adults also report more changes in life circumstances (e.g., retirement, death, and illness of close relatives or peers) than mid-adults (Plomin et al., 1990). Yet, on the long run, personality scores of individuals should return to their dispositional set-points accounting for the long-term rank–order continuity of personality traits.

The classic theory of genetically driven immutable set-points and environmental fluctuation, however, cannot explain the steady decline in rank–order continuity over time (Fraley and Roberts, 2005). In addition, this theory cannot explain systematic mean-level age trends; that is, increases in agreeableness and conscientiousness or decreases in neuroticism from adolescence to middle adulthood (Roberts et al., 2006). These findings contradict the assumption of genetically driven set-points that are immutable.

Genetic Continuity and Biological Maturation

Similar to the dynamic equilibrium model, McCrae and Costa (2008) attributed continuity in personality traits to genetic sources. In their five-factor theory (FFT), they explain age trends and rank–order change in FFM personality traits by genetically predisposed maturation processes, in particular in the first decades of life. They do not negate environmentally driven long-term changes, but they claim that maturation is largely independent of environmental influences and that external factors only manifest through biological changes (see **Figure 4**). This theory can explain long-term mean-level changes as well as changes in individual ranks.

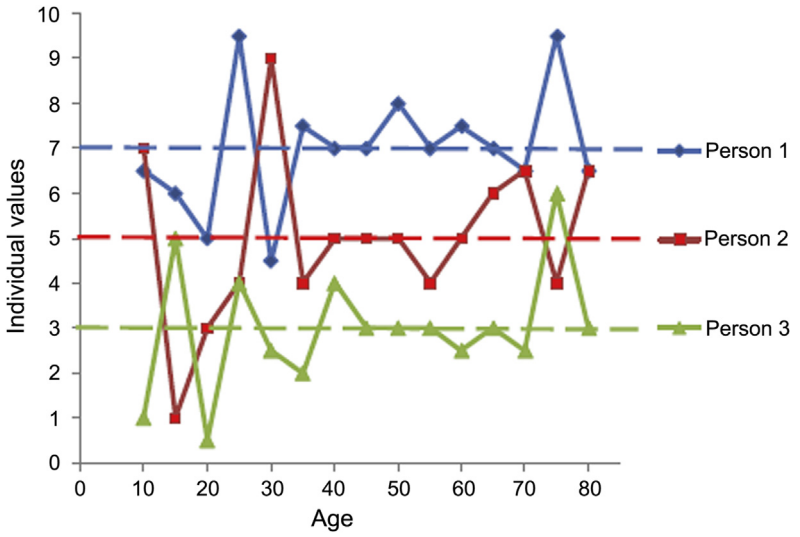

Figure 3 This figure illustrates how the genetic set-point theory can explain that rank–order continuity is lower in younger and older ages. Dotted lines represent the immutable set-points of three persons. Continuous lines reflect environmental fluctuations round these individual set-points that are more pronounced in younger and older ages.

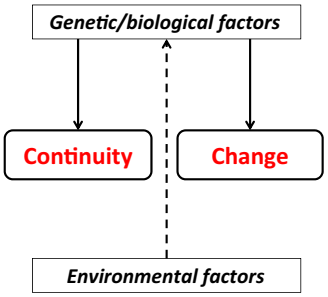

Figure 4 This figure illustrates the major claims of the theory by McCrae, R.R., Costa, P.T., 2008. The five-factor theory of personality. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research, third ed. Guilford, New York, pp. 159–181. Regarding the sources of continuity and change in personality traits. While genetic and biological factors contribute to both continuity and change, environmental factors can only unfold their influences through biological changes.

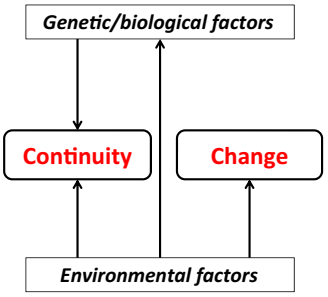

Figure 5 This figure illustrates the major claims of the socio-genomic theory by Roberts, B.W., Jackson, J.J., 2008. Sociogenomic personality psychology. Journal of Personality 76, 1523–1544. Regarding the sources of continuity and change in personality traits. Both genetic and environmental factors contribute to personality continuity, whereas environmental factors contribute to personality change both directly and mediated by biological changes.

In line with the FFT, mean-level trends are largely similar across several nations and cultures (McCrae et al., 2005). Also in line with the hypothesis of genetically driven maturation, several longitudinal behavioral genetic studies on personality stability and changes in childhood, adolescence, and young adulthood have found a non-perfect genetic continuity of personality differences (e.g., De Fruyt et al., 2006; Gillespie et al., 2004; Hopwood et al., 2011; Spengler et al., 2012). This reflects genetically mediated changes in individual ranks during this period of life (Kandler, 2012).

Even though rank–order continuity in personality traits are mainly due to genetic sources, environmental change, however, can also become manifest in long-term rank–order continuity ($\geq$10 years), in particular in middle adulthood (Bleidorn et al., 2009; Kandler et al., 2010a). Environmental influences may accumulate throughout the life span leading to a decline in heritability of personality traits with age (Kandler, 2012). A recent study also found systematic cross-cultural differences in age trends in personality traits that can be linked to the normative duration of education in the respective countries (Bleidorn et al., 2013). These findings contradict the assumption of internally driven maturation of personality traits that are largely independent of environmental influences.

Genetic and Environmental Continuity and Social Maturation

In contrast to FFT, many other theories stress the importance of environmental influences on personality stability and change over the life span. For example, socio-genomic theory of personality by Roberts and Jackson (2008) states that the observed continuity of individual differences in personality can be explained by both genetic as well as environmental factors, in particular social factors (e.g., continuity in social roles). This theory explains personality change by environmental processes (e.g., changes in individual life conditions due to major life transitions or changes in social roles) (see **Figure 5**).

Consistent with this perspective is the fact that major familial and occupational roles are largely stable in middle adulthood. This accounts for the relatively high stability of personality differences in this period of life. Furthermore, there are normative biosocial transitions (e.g., puberty, leaving parental home, graduation from school, apprenticeship, or university, starting a family, menopause, or retirement) in each period of life and there may occur life events with positive (e.g., birth of own child or raise of salary) and negative consequences (e.g., accident or layoff) that may lead to changes in social roles, individual living conditions, and new persistent demands on the individual that can explain enduring personality changes in every age (Roberts et al., 2008). The increase of personality stability from childhood to adulthood may reflect a process of social maturation – the development of, committing to, and maintaining an identity that is functional and integrated in its social community, society, and culture. In line with this position is the finding that the increase of rank–order continuity of personality traits in middle adulthood is, in fact, attributable to stabilizing environmental influences (Kandler et al., 2010a).

Social maturation alone, however, cannot explain the consistent findings of non-perfect genetic continuity of personality differences from childhood to young adulthood (Kandler, 2012). But instable or changing genetic influences to personality differences in younger ages are explainable by the hypothesis of genetically driven maturation.

An Integrative Perspective on Genetic and Environmental Development

The above mentioned theoretical models provide different perspectives on the sources of personality differences, stability, and change. Whereas the dynamic equilibrium model and FFT attribute long-term rank–order continuity in personality traits to genetically anchored immutable set-points, the socio-genomic model allows for experience-dependent long-term personality stability. Whereas FFT mainly attributes personality change to genetic and biological sources, the socio-genomic model primarily explains change by environmental sources that become manifest in individual states or social roles and have to be stable or reoccur for a while before they will be permanently anchored in personality traits.

As discussed above, it seems as every single model cannot explain the entire range of empirical findings. This calls for an

integrative model that allows for genetic and environmental continuity as well as genetic and environmental change as a function of age. Such a model can be described in terms of the following five principles that are mainly derived from the empirical findings described above.

1. Genetic factors represent the primary source of long-term continuity of individual differences in personality.
   This principle is consistent with findings that suggest a larger genetic continuity compared to environmental continuity across age (Kandler, 2012). Moreover, it explains the findings of larger genetic contributions to long-term ($\geq$10 years) rank–order continuity of personality traits (Bleidorn et al., 2009; Kandler et al., 2010a). Moreover, a stable genetic core may account for the positive (non-zero) asymptote to which rank–order continuity declines as the time interval between measurement occasions get longer (Fraley and Roberts, 2005).

2. Environmental conditions and related personality-relevant life experiences are more stable in middle adulthood than in younger and older periods of life.
   This principle is in accord with the finding that environmental continuity of personality differences is larger in middle adulthood than in other life phases (Kandler, 2012). The high environmental continuity (e.g., in familial and working contexts) accounts for the relatively high plateau of rank–order continuity of personality traits between the ages 40 and 70 years (Lucas and Donnellan, 2011).

3. Genetic (or biological) maturation can account for personality change in childhood, adolescence, and young adulthood.
   This principle explains the non-perfect but increasing genetic continuity in the first decades of life until perfect genetic continuity is reached in middle adulthood (Kandler, 2012). Universally similar maturation processes in personality may account for similar age trends across different cultures and societies (McCrae et al., 2005).

4. Life experiences (social maturation) can affect personality change in every period of life.
   According to this principle, environmental influences are the primary source of changes in personality traits (Bleidorn et al., 2009). Moreover, environmental contributions (compared to genetic contributions) to individual differences in personality traits increase across the life span leading to a decline of heritability across age (Kandler, 2012). Processes of identity formation that leads to a socially functional and well-adapted personality (i.e., social maturity) may play a major role in adolescents and young adulthood accounting for the increase in environmental continuity across adolescence, young, and middle adulthood (Hopwood et al., 2011; Kandler et al., 2010a). It lies in the human nature to use its natural potentials of learning, self-reflection, and personal enhancement to develop a functional identity that is socially accepted and culturally integrated. Thus, personality maturation may also depend on social and cultural influences (Bleidorn et al., 2013).

5. Complex interplays between genetic and environmental factors contribute to personality development during all phases in life.
   Emerging adults are typically more motivated to choose or change environments that match their genetic predispositions (i.e., niche picking). This may explain the larger heritability estimates for personality traits within the first decades of life (Kandler, 2012). On the contrary, processes of personality adaptation to maintain the created niches and the functional identity are more important in middle and old adulthood. This can account for the shift from the importance of genetic influences to environmental influences on personality differences across adulthood (Kandler, 2012). Most of the reported behavioral genetic findings have relied on twin data. In twin studies, however, interactions between genetic factors and environmental factors are often not taken into account in the analyses of twin similarity (Eaton et al., 2012). If interactions between genetic factors and environmental factors shared by twins reared together are present but not taken into account, these effects would act like genetic influences. Given this interaction plays a major role in the first decades of life when twins share their parental home and other common experiences (e.g., school or cliques), then the contribution of genetic influences to individual differences in personality should be larger in younger ages. This, again, is in line with larger heritability estimates for personality traits within the first decades of life. The other way round, if interactions between genetic factors and environmental factors not shared by twins reared together are present but not taken into account, these effects would be confounded with estimates of non-shared environmental influences. Given this interaction plays a major role in adulthood when twins have left their common parental home and cut their own paths, then heritability should decrease in this period of life. This is in accord with the increasing importance of environmental influences on personality differences across adulthood.

In summary, a model that integrates ostensibly contradicting claims can account for the whole empirical findings. Personality development is a very complex affair. We can only understand the complete picture, if we focus on both genetic and environmental sources as well as the complex interplay between them.

Outlook on Present and Future Trends

Future Studies on Personality Differences and Development

Many of the findings we referred to above refer to only a few broad personality traits, primarily neuroticism and extraversion. It remains an open question whether we can find the same patterns and sources of continuity and change across the life span for other personality traits, such as openness, agreeableness, conscientiousness, or more specific personality facets.

With respect to broad-sense personality, there are interesting differences between personality traits and abilities regarding the relative contributions of genetic and environmental influences to individual differences over the life span. Whereas heritability increases for abilities across age (Briley and Tucker-Drob, 2013), heritability of traits appears to decrease in adulthood. Maybe, there are different processes

involved in the development of traits and abilities (Kandler et al., 2013). This needs to be addressed directly in future studies.

A related question is which patterns of genetic and environmental continuity and change we can find for other personality characteristics that are associated to some degree with the Big Five traits and cognitive abilities, but represent other attributes with more motivational content (e.g., self-related and social attitudes, values, motives, needs, and interests). Future studies should invest more efforts to study the nature and nurture in the development of these characteristics.

Bridging of Quantitative and Molecular Genetic Studies

There is still a large discrepancy between the quantitative genetic findings of large heritability estimates for personality traits and the inconsistent and less replicable findings of very small effects of gene variants accounting for personality differences from the field of molecular genetics. This is often called the missing heritability problem. How can we overcome this large discrepancy? In the following, we provide several hypotheses that need to be investigated by future research.

First, developmental behavioral genetic studies suggest that some genetic factors may emerge in specific ages and that heritability appears to vary across age. Therefore, age differences have to be taken into account in molecular genetics' association studies.

Second, most large genome-wide association scan (GWAS) studies on personality differences rely on short and unreliable measures of personality traits. Since heritability estimates increase with accuracy of measurement (Kandler et al., 2010b; Riemann and Kandler, 2010; Kandler et al., 2013), molecular genetic studies should be able to use more reliable and valid measures of personality traits. For sure, this may be useful to catch more robust gene effects on the level of candidate genes or even single nucleotide polymorphisms (SNPs).

Third, personality traits may be affected by multiple gene variants with very small and statistically insignificant effects. In line with this consideration, a recent study has found that common SNPs – that is, the additive combination of the smallest gene units that vary among individuals – can explain 12% of the variance in extraversion and 6% of the variance in neuroticism (Vinkhuyzen et al., 2012).

Fourth, heritability estimates often stem from studies on within-generational relatives (e.g., twins and other siblings). These estimates include variance due to the nonadditive combination of gene variants that vary among individuals within gene loci (i.e., allelic dominance) or between gene loci (i.e., epistatic gene interactions). Since twin studies have often found higher heritability estimates on the basis of twin similarity compared to adoption or pedigree studies on personality similarity among other family relatives (Vinkhuyzen et al., 2012), this speaks for epistatic gene interactions that are only shared by monozygotic (genetically identical) twins. Epistatic interactions are hard to identify and its effects are even harder to replicate. From these four points of view, the discrepancy between quantitative and molecular genetics regarding the amount of genetic influences on individual differences in personality is not surprising.

Manifestations of Environmental and Epigenetic Factors

Behavioral genetic research suggests that nurture is primarily individual-specific. That is, environmental influences are primarily not shared by family members raised within the same family. Effects of specific environmental variables (e.g., life events) are typically relatively small, accounting for less than 2% of individual differences in personality traits (e.g., Kandler et al., 2012). This displays a parallelism to the missing heritability problem we, therefore, call the missing environmentality problem.

As a parallelism to multiple gene variants with very small and statistically insignificant effects, specific environmental variables alone may have negligible effects. In line with this perspective aggregates of several environmental variables can account for about 13% of the variance in individual attributes (Turkheimer and Waldron, 2000). In addition, as a parallelism to the role of epistatic gene interactions, several environmental factors may affect some but not all individuals and they may interact with other experiences.

Finally, twin studies on epigenetic effects indicate that epigenetic imprinting (e.g., DNA-methylation or histone-modification) can account for differences between monozygotic twins' gene expressions (see van Dongen et al., 2012 for a review). Not surprisingly, epigenetic differences show a small heritability of less than 10%. That is, epigenetic effects primarily reflect individual-specific environmental influences. Epigenetic effects may increase over the life span (epigenetic drift). This epigenetic drift can mirror the increase of environmental contributions to the variance in personality traits across the life span on the level of gene expression.

Conclusions

Behavioral genetic research on personality has made enormous progress within the last decade. We have learnt a lot about the genetic and environmental contributions to personality differences, stability, and change. Genetics factors account for about 50% of the variance in personality self-reports. However, heritability estimates are often considerably larger for more accurate, cross-time stable, and cross-situations consistent with measures of personality traits. Genetic factors represent the primary source of long-term continuity of individual differences in personality, whereas the stability of environmental factors increases until middle adulthood. Environmental sources affect personality change in every period of life, but genetic factors can also explain change in childhood, adolescence, and even in young adulthood. Finally, heritability of personality traits steadily declines in adulthood that is completely different to the well-known trend for cognitive abilities. These findings have helped to integrate ostensibly contradicting theoretical models by formulating five principles of the nature and the nurture in personality development that fit the empirical results. There are still many open questions that need to be answered by future research. Two broad areas of for future research are the missing heritability and environmentality problems that need to be addressed by future behavioral genetic studies.

See also: Big Five Factor Model, Theory and Structure; Epigenetic Inheritance; Five Factor Model of Personality, Assessment of; Five Factor Model of Personality, Facets of; Five Factor Model of Personality, Universality of; Genetics and Anthropology; Genetics of Personality; Genetics: The New Genetics.

Bibliography

Bleidorn, W., Kandler, C., Riemann, R., Angleitner, A., Spinath, F.M., 2009. Patterns and sources of adult personality development: growth curve analyses of the NEO-PI-R scales in a longitudinal twin study. Journal of Personality and Social Psychology 97, 142–155.

Bleidorn, W., Klimstra, T.A., Denissen, J.J.A., Rentfrow, P.J., Potter, J., Gosling, S.D., 2013. Personality maturation around the world: a cross-cultural examination of social-investment theory. Psychological Science 24, 2530–2540.

Borkenau, P., Riemann, R., Spinath, F.M., Angleitner, A., 2006. Genetic and environmental influences on person × situation profiles. Journal of Personality 74, 1451–1479.

Briley, D.A., Tucker-Drob, E.M., 2013. Explaining the increasing heritability of cognitive ability across development: a meta-analysis of longitudinal twin and adoption studies. Psychological Science 24, 1704–1713.

De Fruyt, F., Bartels, M., Van Leeuwen, K.G., de Clercq, B., Decuyper, M., Mervielde, I., 2006. Five types of personality continuity in childhood and adolescence. Journal of Personality and Social Psychology 91, 538–552.

van Dongen, J., Slagboom, P.E., Draisma, H.H.M., Martin, N.G., Boomsma, 2012. The continuing value of twin studies in the omics era. Nature Reviews Genetics 13, 640–653.

Eaton, N.R., Krueger, R.F., South, S.C., Gruenewald, T.L., Seeman, T.E., Roberts, B.W., 2012. Genes, environments, personality, and successful aging: toward a comprehensive developmental model in later life. Journal of Gerontology 67A, 480–488.

Fraley, C., Roberts, B.W., 2005. Patterns of continuity: a dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. Psychological Review 112, 60–74.

Gillespie, N.A., Evans, D.E., Wright, M.M., Martin, N.G., 2004. Genetic simplex modeling of Eysenck's dimensions of personality in a sample of young Australian twins. Twin Research 7, 637–648.

Hopwood, C.J., Donnellan, M.B., Krueger, R.F., McGue, M., Iacono, W.G., Blonigen, D.M., Burt, S.A., 2011. Genetic and environmental influences on personality trait stability and growth during the transition to adulthood: a three-wave longitudinal study. Journal of Personality and Social Psychology 100, 545–556.

John, O.P., Naumann, L.P., Soto, C.J., 2008. Paradigm shift to the integrative big five trait taxonomy. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research, third ed. Guilford, New York, pp. 114–158.

Johnson, A.M., Vernon, P.A., Feiler, A.R., 2008. Behavioral genetic studies of personality: an introduction and review of the results of 50+ years of research. In: Boyle, G.J., Matthews, G., Saklofske, D.H. (Eds.), The SAGE Handbook of Personality Theory and Assessment, vol. 1. SAGE, London, pp. 145–173. Personality Theories and Models.

Kandler, C., 2012. Nature and nurture in personality development: the case of neuroticism and extraversion. Current Directions in Psychological Science 21, 290–296.

Kandler, C., Bleidorn, W., Riemann, R., Angleitner, A., Spinath, F.M., 2012. Life events as environmental states and genetic traits and the role of personality: a longitudinal twin study. Behavior Genetics 42, 57–72.

Kandler, C., Bleidorn, W., Riemann, R., Spinath, F.M., Thiel, W., Angleitner, A., 2010a. Sources of cumulative continuity in personality: a longitudinal multiple-rater twin study. Journal of Personality and Social Psychology 98, 995–1008.

Kandler, C., Riemann, R., Spinath, F.M., Angleitner, A., 2010b. Sources of variance in personality facets: a multiple-rater twin study of self-peer, peer-peer, and self-self (dis-) agreement. Journal of Personality 78, 1565–1594.

Kandler, C., Riemann, R., Angleitner, A., 2013. Genetic and environmental continuity and change of energetic and temporal aspects of temperament in adulthood: a longitudinal twin study of self- and peer reports. Developmental Psychology 49, 1739–1753.

Lucas, R.E., Donnellan, M.B., 2011. Personality development across the life span: longitudinal analyses with a national sample from Germany. Journal of Personality and Social Psychology 101, 847–861.

McAdams, D.P., Pals, J., 2006. A new big five: fundamental principles for an integrative science of personality. American Psychologist 61, 204–217.

McCrae, R.R., Costa, P.T., 2008. The five-factor theory of personality. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research, third ed. Guilford, New York, pp. 159–181.

McCrae, R.R., Terracciano, A., 78 Members of the Personality Profiles of Cultures Project, 2005. Universal features of personality traits from the observers perspective: data from 50 cultures. Journal of Personality and Social Psychology 88, 547–561.

Ormel, J., Riese, H., Rosmalen, J.G.M., 2012. Interpreting neuroticism scores across the adult life course: immutable or experience-dependent set points of negative affect? Clinical Psychology Review 32, 71–79.

Plomin, R., Lichtenstein, P., Pedersen, N.L., McClearn, G.F., Nesselroade, J.R., 1990. Genetic influence on life events during the last half of the life span. Psychology and Aging 5, 21–30.

Riemann, R., Kandler, C., 2010. Construct validation using multitrait-multimethod-twin data: the case of a general factor of personality. European Journal of Personality 24, 258–277.

Roberts, B.W., Jackson, J.J., 2008. Sociogenomic personality psychology. Journal of Personality 76, 1523–1544.

Roberts, B.W., Kuncel, N.R., Shiner, R., Caspi, A., Goldberg, L.R., 2007. The power of personality: the comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. Perspectives of Psychological Science 2, 313–345.

Roberts, B.W., Walton, K.E., Viechtbauer, W., 2006. Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. Psychological Bulletin 132, 1–25.

Roberts, B.W., Wood, D., Caspi, A., 2008. The development of personality traits in adulthood. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research, third ed. Guilford, New York, pp. 375–398.

Spengler, M., Gottschling, J., Spinath, F.M., 2012. Personality in childhood: a longitudinal behavior genetic approach. Personality and Individual Differences 53, 411–416.

Turkheimer, E., Waldron, M., 2000. Nonshared environment: a theoretical, methodological, and quantitative review. Psychological Bulletin 126, 78–108.

Vinkhuyzen, A.A.E., Pedersen, N.L., Yang, J., Lee, S.H., Magnusson, P.K.E., Iacono, W.G., McGue, M., Madden, P.A.F., Heath, A.C., Luciano, M., Payton, A., Horan, M., Ollier, W., Pendleton, N., Deary, I.J., Montgomery, G.W., Martin, N.G., Visscher, P.M., Wray, N.R., 2012. Common SNPs explain some of the variation in the personality dimensions of neuroticism and extraversion. Translational Psychiatry 2, e102.

Yamagata, S., Suzuki, A., Ando, J., Ono, Y., Kijima, N., Yoshimura, K., Ostendorf, F., Angleitner, A., Riemann, R., Spinath, F.M., Livesley, W.J., Jang, K.L., 2006. Is the genetic structure of human personality universal? A cross-cultural twin study from North America, Europe, and Asia. Journal of Personality and Social Psychology 90, 987–998.

Personality Disorders

Peter Tyrer, Imperial College, London, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by H. Saß, volume 16, pp 11301–11308, © 2001, Elsevier Ltd.

Abstract

Personality disorders are the most common disorders in the general population and also show overlap (comorbidity) with other mental disorders. As a consequence between 40% and 90% of patients seen in psychiatric practice have personality disorder. The term 'personality disorder' has a long history. It developed from the concept of psychopathy as a degenerative disorder from the opinions of French, German, and Anglo-American psychiatric traditions. Now it is realized that sociocultural factors are much more important than previously suspected, even though around half of normal personality characteristics are genetically determined. Modern classification systems (the proposed ICD-11 and DSM-5) have tried to break out from the somewhat unproductive categorical system of pejorative named disorders that have no external evidence of support for their separation. Only borderline and antisocial personality disorders are used commonly in practice, and this is probably because they are prominent in more severe personality disorders. In this article the current state of classification, treatment, and course of personality disorders is outlined; there is ground for some optimism.

General Considerations

Until the 1980s, both the scientific and the therapeutic aspect of personality disorders were neglected in both practice and research despite their great significance for clinic and practice. Since the introduction of separate Axis II personality registration used in DSM-III (American Psychiatric Association, 1980) and the subsequent editions, DSM-III-R and DSM-IV (American Psychiatric Association, 1987, 1994), research and publication worldwide have prospered, with the introduction of professional journals and scientific societies devoted solely to the field of personality disorders. But there is also much dispute in the field. Now, at the beginning of the twenty-first century, there is a veritable deluge of theoretical concepts of personality and therapeutic interventions, each vying with each other for attention and support despite little evidence that one or more is superior to any other. The alternative more positive way at looking at the field is to congratulate the participants as there are many runners and conclude with Lewis Carroll – "Everyone has won and all shall have prizes."

More than in any other area of psychiatry, this has been a field dominated by opinion and dogma, and hard data on nosology, etiology, and pathogenesis of the alleged individual personality disorders remaining elusive. There is indeed a wealth of isolated facts about the etiological and pathogenetic factors that influence personality traits, but a comprehensive theoretical or empirical model remains as elusive as ever. It is only possible to address certain points and make some qualified references, as the exceedingly complex nature of the subject still does not allow anymore definite statements. In this account there are some contentious points made; the author considers there is more evidence in their support than the alternatives.

'Personality disorder' remains the only term that is widely used and which encompasses all those deviations from a normal personality development that require treatment. There are, necessarily, fluid borders to mental health on the one hand and mental disorder on the other.

Concepts with meaning similar or identical to personality disorder are psychopathy, abnormal personality, psychopathic development, psychopathic or dissocial personality, sociopathy, and personality difficulty.

History of the Concepts of Abnormal Personality

The concept of personality disorder is closely linked to 'psychopathy' (a confusing term as it now refers to the antisocial spectrum only) and comes from French, the German, and the Anglo-American psychiatric traditions.

French Concepts

Pinel's (1809) concept of *manie sans délire* includes instances of 'deranged personality' and can be looked upon as the beginning of the scientific study of personality disorder as a nosological entity. In the eighteenth century, all mental diseases were regarded as fundamental disturbances of the intellect. Pinel was one of the first to point out that in some disorders, the emotions were primarily involved. Nevertheless, the early nineteenth-century definitions of madness remained centered mostly around intellectual function.

Esquirol (1838) developed the concept of monomania, a diagnostic category based on the partial, primary – and independent – involvement of the intellectual, emotional, and/or volitional functions of the mind, including states where a single behavioral disturbance became the sole diagnostic criterion for the condition (e.g., pyromania or kleptomania). Not surprisingly, monomania was one of the sources of the concept of psychopathy. Since its inception, the concept of monomania has been criticized on clinical and medicolegal grounds both in France and abroad.

Morel's (1857) idea of degeneration was a pseudobiological account strongly tinged with the idea of original sin. He proposed that (1) degenerative alterations are

pathological deviations from the normal; (2) mental illness is mostly hereditary; and (3) degeneration is both quantitative and qualitative, that is, new disorders may appear. According to Morel's model, all types of mental and neurological disorders can be traced back to one common hereditary origin. His idea of progressive and polymorphic degeneration was accepted as an explanation of mental illness.

Anglo-American Concepts

Prichard (1835) defined moral insanity as "madness consisting of a morbid perversion of the natural feelings, affections, inclinations, temper, habits, moral dispositions and natural impulses without any remarkable disorder or defect of the interest or the reasoning faculties, and particularly without any illusion or hallucinations." During the early nineteenth century, the word 'moral' had many uses, in the psychological sense chiefly to denote the affective and the conative (rather than the purely intellectual) functions.

The British concept of psychopathy was further shaped by D.K. Henderson (1939), who saw 'psychopathic states' as a condition of 'constitutional abnormality.' In contrast to other (especially German) psychiatrists, he thought of constitution as deriving from both heredity and environment. There were three psychopathic states: the predominantly aggressive, the inadequate, and the creative. The former two were characterized by antisocial traits and soon became part of the Anglo-American concept of personality disorder.

Rush (1812) was the first American psychiatrist to study individuals whose disturbances were primarily characterized by irresponsibility and aggressiveness; who showed, as he put it, a "moral alienation of the mind." He believed that these reprehensible acts were the manifestations of mental disease that they were unmotivated and driven by 'a kind of involuntary power.'

Partridge (1930) defined the concept of psychopathic personality as a persistent maladjustment which cannot be corrected by ordinary methods of education or by punishment. From Partridge on, the emphasis has been on description, and etiological speculation has taken the back seat. To this day, a view of psychopathy as sociopathy dominates Anglo-American psychiatry. A parallel concept, that of the antisocial personality disorder, appeared in DSM-III (APA, 1980) and was retained in DSM-III-R (APA, 1987) and DSM-IV (APA, 1994).

German Concepts

In Germany, the concept of psychopathy embraces most forms of abnormal personality. The current German definition can be traced back to Koch, whose monograph *Psychopathische Minderwertigkeiten* [Psychopathic Inferiorities] (1891–93) was as decisive in shaping the concept of abnormal personality in Germany as Pinel's work was in France, Rush's in the USA, and Prichard's in Great Britain. German ideas of psychopathy also influenced French and Anglo-American views, especially after the 1930s, when many German-speaking psychiatrists and psychoanalysts emigrated to these countries.

Kraepelin's concept of psychopathy, influenced by the French theory of degeneration, formed the basis of Kurt Schneider's typology, and via the latter, of today's German doctrine of psychopathy. In successive editions of his textbook, Kraepelin elaborated on his concept of 'psychopathic states' as abnormal personality. He employed the term 'psychopathic personalities' in a predominantly social sense, including also innate delinquents, unstable liars, swindlers, and pseudo-querulous individuals.

During this period, there also appeared various 'systematic' typologies to draw 'psychopathic' categories from prototypic personality theories. Foremost among these is Kretschmer's *konstitutionstypologisches Modell* (1919). He suggested a specific correlation between body type and personality. Kretschmer (1921) and Ewald (1924) also introduced the concept of *Reaktionstypologien* to account for specific styles of dealing with experience. Systematic typologies lost influence after the publication of Schneider's monograph in 1923.

Kurt Schneider is known especially for his famous monograph *Die psychopathischen Persönlichkeiten* (The Psychopathic Personalities) (1923). He moved the area away from a notion that personality disorder was primarily concerned with neurological degeneration or constitutional inferiority and instead defined it as a problem of behavior and relationships. Schneider coined the memorable phase that those with personality disorder 'suffered from their disorder and also caused society to suffer' which emphasized the necessary problematic interaction with others before the diagnosis could be made.

Schneider's influence has been considerable. It is sometimes said that the eminence of a scientist can be measured by the number of years that he is able to hold up progress, and in Schneider's case this has now almost passed its century. Schneider described a set of 'psychopathic' personalities, although this term is confusing because the word 'psychopath' now has a different construction. His main 10 categories were:

- Hyperthymic (now regarded as within the bipolar spectrum)
- Depressive
- Insecure (now termed avoidant or anxious)
- Fanatic (showing some overlap with narcissistic)
- Attention-seeking (directly translated in the word 'histrionic')
- Labile (now called borderline)
- Explosive (equivalent to impulsive in the UCD-10 classification)
- Affectionless (equivalent to antisocial)
- Weak-willed (no obvious equivalent)
- Asthenic (equivalent to dependent)

Only the current paranoid (whom Schneider described as sensitive personalities (*der* sensitive Beziehungswahn), schizoid, and obsessive–compulsive personalities were not described by Schneider. It is clear that he was an excellent observer and describer of personality and its variation.

Borderline Syndrome

The disorder that perhaps lacked good description, and certainly has poor reliability (Standage, 1979) (from Schneider's list) is labile, now borderline, personality disorder. In many ways this is the odd one out as its subsequent development after Schneider's description came from psychoanalysis

rather that descriptive psychopathology. In the famous paper by Hoch and Polatin (1949) the term 'pseudo-neurotic schizophrenia' was used to describe a condition that many now regard as an example of the admixture of anxiety associated with the development of a schizophrenic psychosis. This paper was published at the time (1949) when schizophrenia was a very loose diagnosis in the United States in which any dysfunctional thought could be interpreted in a way to lead to the diagnosis. It was quickly realized that many patients with pseudo-neurotic or latent schizophrenia did not have any psychosis in the conventional sense, but were extremely disturbed in their mental function, with rapid changes in their affective state that were to some extent different from bipolar disorder, were unable to control their mood and emotions (subsequently called emotional dysregulation), had periods of hopelessness and utter despair often leading to suicidal behavior, acting impulsively with subsequent regret, and only occasional brief psychotic episodes. If this condition had been described at any other time, or by any other psychiatric group other than a psychoanalytic one, it would probably have been classified as yet another mental state disorder, but it was instead incorporated into the personality spectrum.

Much of the credit for placing it with the personality grouping came from the pioneering work of John Gunderson, who first described the cardinal features of the borderline syndrome with his colleague Dr Singer in 1975 (Gunderson and Singer, 1975). Otto Kernberg (1985) also reformulated the psychodynamic interpretation of borderline personality disorder as a diffuse system that had few boundaries and allowed people to register their feelings as primitive emotions without apparent constraint or control. The phenomenology of borderline personality disorder was not very different from other mood disorders, but its impact on others was more like that of personality disorders.

DSM-III

By the time of the DSM-III revision of personality disorders in 1980 it was difficult to know where borderline personality disorder, and indeed the other personality disorders, might be placed. In a bold move, that many have seen as political rather than scientific, Robert Spitzer, the architect of DSM-III, offered personality disorders to the psychodynamic lobby of American psychiatrists as a separate axis of classification on the grounds that no other group of psychiatrists would be particularly interested in this diagnosis, and by giving it a separate axis its full implications could be explored independently of any other disorder. This was essentially a trade-off, the resistance of many psychodynamic psychiatrists to DSM-III could be reduced if they then had a condition in which their expertise was recognized to be highly important. However, in what may be a personal minority view, of the personality disorders, it is only borderline personality disorder that is really attractive to psychodynamically orientated therapists, and, in terms of its phenomenology, this disorder is really more of a complex affective state rather than a personality disorder (Tyrer, 2009).

DSM-III finished with a new set of personality disorders and a new classification axis for them (Axis II). But, despite this great achievement all was not exactly well with the new classification. Axis 1 disorders were diagnosed using the operational criteria that had been so influential in defining schizophrenia, depression, and bipolar disorder and to a lesser extent the non-psychotic disorders. This procedure did not work so well with Axis II. Clinicians were not being asked to identify an abnormal mental state with a fixed-time onset and often a short duration, but a continuous problem of relationships and behavior that could not easily be described and accessed in a single interview. An attempt was made to describe typical (prototypic) behaviors that brought out the main features of the personality disorder. For example, one of the criteria for paranoid personality disorder "reads hidden meanings or threats into benign remarks or events, for example, suspects that neighbours put out trash early to annoy them." This is a culturally derived concept that is very difficult to extrapolate to the worldwide stage; how would a paranoid Eskimo or Bedouin nomad score on this particular item? Despite this, DSM-III has dominated personality disorder classification since this time, and the current classification and description are shown in Tables 1 and 2.

Focusing on specific behaviors shows how difficult it is for an assessor to decide whether someone is genuinely paranoid simply by picking on one or two single types of problem and hoping that these can be melded together into an overall condition that will allow the diagnosis of paranoid personality disorder. Short cuts to the diagnosis of personality problems are inherently flawed – rather like guessing the nature of a picture from a thousand-pieces jigsaw puzzle by seeing only twenty of the pieces in their proper position. What is needed is a better overall perspective of the subject, not the exact contours of a few specific parts.

Furthermore, when detailed studies were made of the distribution of the different criteria among the nine personality disorders in DSM-III the real structure was found to be completely different from that of the architects of the individual disorders (Livesley et al., 1994). Clinicians were

Table 1 General diagnostic criteria for personality disorder

1. A lasting pattern of inner experiences and behavior which deviates markedly from the expectations of the culture the individual lives in. This pattern is manifested in two or more of the following areas:
   a. Cognition (i.e., ways of perceiving and interpreting oneself, others, and events)
   b. Emotiveness (i.e., range, intensity, stability, and appropriateness of emotional response)
   c. Interpersonal functioning
   d. Impulse control
2. The pattern is inflexible and pervasive across a broad range of personal and social situations.
3. The pattern leads to clinically significant distress or impairment in social, occupational, or other important areas.
4. The pattern is stable and long-lasting, and its onset can be traced back to adolescence or early adulthood.
5. The pattern is not better accounted for as a manifestation or consequence of another mental disorder.
6. The pattern is not due to direct physiological effects of a substance (e.g., drug or medication) or a general medical condition (e.g., head trauma).

Table 2 Classification of personality disorders

DSM-IV	*ICD-10*
Paranoid	Paranoid
Schizoid	Schizoid
Schizotypical	–
Antisocial	Dissocial
Borderline	Impulsive
Histrionic	Histrionic
Narcissistic	–
Avoidant	Anxious
Obsessive–compulsive	Obsessive–compulsive
Not otherwise specified	Not otherwise specified and others

also being asked to make the diagnosis of personality disorder in two stages: (1) the identification of the presence or absence of the personality disorder, followed by (2) the description of the disorder by category. The trouble with this advice was that people were so fascinated by the individual labels of personality disorder that they went straight to the second phase without really considering whether or not the person had a generic personality disorder and satisfied or did not satisfy the descriptions for this. This was also complicated by the consequence, entirely predicted by the empirical evidence, that because the factor structure of the individual categories overlapped, the disorders themselves seemed to overlap to a similar degree. It was therefore fairly rare for a person to have a 'pure' categorical personality disorder without any features of any other. This almost completely nullifies the advantages of a specific categorical label.

Likely Changes in the Classification DSM-5 and ICD-11

The current situation of personality classification can be summarized as the 'Hong Kong problem.' Hong Kong was a part of China but was leased to the British government because it was highly productive and allowed trade that would otherwise have not taken place between China and the rest of the world. It could always be only temporary and now Hong Kong is part of China. Similarly, allowing personality disorder to have its own special axis of classification led to expansion of research and activity in this field but the separate classification was always likely to be 'on loan,' and now it appears that the main body of mental state disorders wants to reclaim personality disorder as its own, and all axes are proposed to be abolished.

This will mean that all mental disorders will be classified in the same group and personality dysfunction may sometimes be forgotten about when it is necessary to make a primary diagnosis only. One of the reasons for providing a second axis of classification, in addition to satisfying the psychodynamic lobby, was to give more attention to conditions that might otherwise be forgotten. There is no doubt that this has been a benefit, as least to researchers, and abandoning it might reduce the profile of personality disorder in the future by relegating it to a secondary position. On the positive side it would allow personality disorder to benefit from the process of nosological inclusion – no longer on the outside of the main system, and possibly reduce the stigma to which this diagnosis is subject, particularly among psychiatrists. But it has to earn its place there and will have great competition. It would help greatly if more than one diagnosis was encouraged as the influence of personality disorder on the outcome of other conditions, as outlined in the following section, can be considerable. What is needed is a simple way of diagnosing personality disturbance that is socially inclusive rather than stigmatizing. This is being attempted in the revision of personality disorder in ICD-11 by a working group on behalf of the World Health Organization.

Introduction and Summary of Current ICD-11 Personality Disorder Proposal

The fundamental changes proposed for the revision of the classification of ICD-11 personality disorders can be summarized as follows:

1. to make the primary classification based on *severity* of personality disturbance;
2. a *secondary classification* of *five trait domains* that qualify the nature of the personality disturbance but which is not essential to the classification;
3. a *simple rating procedure* that allows diagnosis to be made easily both in primary and secondary care (Tyrer et al., 2011).

The DSM-5 proposals for the revision of the DSM-IV classification of personality disorders also regard severity as part of the classification. However, the primary classification is again by category, but into a somewhat smaller group than existed previously. Currently, borderline, schizotypal, narcissistic, avoidant, obsessive–compulsive, and antisocial personality disorders are in the DSM-5 diagnostic frame. The DSM-5 proposal is also likely to be considerably more complex than the ICD-11 proposal, taking up to 40 min or longer to assess properly, and hence there are questions about its clinical utility. Unfortunately, the American Psychiatric Association regarded the complexity overcame its clinical usefulness and currently it is in a new group, that is, Section 3 of the classification (for further study).

The new ICD-11 system abolishes categories of personality disorder and the secondary domains of personality only serve to qualify the classification of severity. Personality status will be classified on five-dimensional levels:

1. No personality disturbance.
2. Personality difficulty (some personality problems in certain situations but not universally) (coded as a Z-code so it does not become a formal psychiatric disorder).
3. Mild personality disorder (definite well-demarcated personality problems across a range of situations).
4. Moderate personality disorder (definite personality problems usually covering several personality domains and across all situations).
5. Severe personality disorder (definite personality problems usually covering several personality domains and across all situations which carry significant risk to the individual or to society).

The adoption of a severity classification does not overcome the need to have a division between those who have personality disorder and those who do not, although the dimensional range makes the point of separation less critical. The critical diagnostic point is reached when there is clear evidence that interpersonal dysfunction is pervasive and persistent, a behavioral and attitudinal fingerprint that shows itself in all situations. The distinction between interpersonal and general social dysfunction is an important one. I personally think that the best definition of interpersonal dysfunction is 'a pattern of general impairment in human relationships that prevents mutual understanding' as it is the lack of mature connection with other people that is the main problem here. This is an attempt to make a clear dividing line between social malfunction occurring as a consequence of other mental disorder and that associated with personality difficulties.

The domains of personality have not yet reached final agreement but are likely to be five: dissocial, detached, disinhibited, negative emotional, and anankastic domain traits (Tyrer et al., 2014) – the externalizing, internalizing, schizoid, and obsessive–compulsive domains. Thus, in describing someone as suffering from a personality disorder currently within the borderline emotionally unstable domain there will only be one diagnosis indicating the severity, and additional diagnostic information such as 'with externalizing and internalizing features.' Thus, the clinician will only be pointing out the features that are most prominent within the severity framework, not fixing on 'borderline' as a descriptor. In ordinary psychiatric practice many of the finer points about the subdivision of the categories of personality disorder – affective instability and impulsivity in *borderline personality disorder*; callousness and irresponsibility in *antisocial disorder*; and dependence and anxiousness in *avoidant personality disorder* – show the lack of sharp boundaries between each of the individual categories. Despite the wish to give them rock-like diagnostic status, the current classification are fundamentally unstable. Trait domains are more secure, but they still overlap, particularly when personality disorder gets more complex and severe, when they end to accumulate the features of other domains.

Mixed and Unspecified Personality Disorders

One of the most serious deficiencies in the current classifications of personality disorder is the very large number of diagnoses that indicate personality disturbance but do not fit the existing categories. These can account for up to a third of all diagnoses.

The proposed ICD-11 classification should help greatly in avoiding the need to use mixed categories. At higher severity levels, mixed personality disorders are the norm. Both complex and severe disorders are likely to exhibit characteristics from more than one domain, and the proposed system offers the opportunity to record the domains that are most prominent. By making the severity level the primary rubric of classification, more complex formulations are kept within the overall structure; everybody with personality disturbance has to be classified at one of the severity levels. In this way, the domains of personality disturbance are essentially qualifiers for personality disorders at a given level of severity. The only time that 'unspecified' would be used is when the severity level cannot be determined.

Why Does the Present System Need to be Changed?

Most clinicians agree that the current system is unsatisfactory in practice (Bernstein et al., 2007) and the criticisms boil down to three issues: (1) the frequent difficulty of making a good fit between patients seen in clinical practice and the formal categorical descriptions; (2) the heterogeneous and shifting nature of personality disturbance, even though personality is meant to be one of the most enduring of human characteristics; (3) the artificial dichotomy between pathology (personality disorder) and health (no personality disorder). Both DSM-5 and ICD-11 agree on this last point, and they introduce dimensions in various ways into the classification.

Clinical Utility – Making a Better Fit

The categories of personality disorder are good for office parties but not for office psychiatry. Some of the more amusing discussions at social occasions where the conversation begins to flag are about the personality characteristics of well-known individuals. Like a moth to a flame, the adjectives 'paranoid,' 'borderline,' 'avoidant,' 'narcissistic,' and 'antisocial' are just too enticing, particularly when they can be applied to politicians, doctors, administrators, and bureaucrats who create annoyance in their different ways. The curious thing is that when clinicians go back to their office they do everything possible to avoid using these terms, even though the pathology is obvious for all to see. There is now good evidence that somewhere between 10% and 20% of patients in primary care and between 30% and 40% of patients attending secondary care clinics in both psychiatry and general medicine have significant personality disturbance when assessed by researchers (Tyrer et al., 2014).

Not only are we very chary of using the diagnosis of personality disorder but we mainly confine our descriptions of personality disorder to three groups: borderline, antisocial, and (in the USA) schizotypal personality disorder. The others are hardly ever used in practice or in research. In the USA there is concern over the possible loss of narcissistic personality disorder and, while the gossipers will doubtless predict this concern as self-fulfilling, since the diagnosis is accompanied by a strong sense of entitlement, this alone is not be a good reason to retain it!

Treatment

One of the reasons why personality disturbance in any form was, and to some extent still is an unpopular diagnosis is that treatment options are limited. But there has been increasing optimism over the past 20 years, even though the gains are still relatively small, as one might expect after a long, and until recently justified, reputation of untreatability.

There are now over 20 treatments competing for treatment of personality disorder.

The biggest impact has been made by dialectical behavior therapy (DBT) in a paper published 20 years ago (Linehan et al., 1991). This unashamedly positive paper, with a well-promoted treatment protocol, was a major influence in persuading clinicians and managed care companies that at least one personality disorder, borderline, was responsive to treatment, particularly in reducing episodes of self-harm. (It is also fair to add that Marsha Linehan, the originator of DBT, does not like the term borderline personality disorder, and prefers the more accurate 'emotional dysregulation disorder'.)

Following the success of DBT, a treatment originally described as a form of psychodynamic psychotherapy linked to day hospital treatment (partial hospitalization), but later relabeled as mentalization-based treatment (MBT) (Bateman and Fonagy, 1999), has been almost as influential. MBT offers a way forward for psychodynamic therapy, then held back by lack of evidence, and this treatment has now been manualized. Transference-focused therapy, based on Kernberg's model, has also shown similar benefit. Cognitive-behavior therapy adapted for personality disorders, and its related treatment, schema-focused therapy, also have published evidence of their efficacy, but despite the high quality of these trials, have not had as much impact as MBT and DBT trials, partly because they were published later and partly because their results are not quite as impressive (Davidson et al., 2006; Giesen-Bloo et al., 2006). More recently, Systems Training for Emotional Predictability and Problem Solving (STEPPS) (Blum et al., 2008) has also yielded good results comparable with other trials. Cognitive analytical therapy has also been widely used, but with limited evidence of efficacy when compared with good standard care. Indeed, when all these treatments have been compared with good standard care their impact, is less impressive and shows that generic aspects of psychological treatments are more important than their specific differences (Bateman et al., 2014).

Other forms of personality disturbance have attracted reduced therapeutic interest. The main reason for this is simple one: a majority of those with any form of personality disturbance do not want to have treatment, irrespective of any value it may have. Type R (treatment resisting) personalities are three times commoner than Type S (treatment seeking) personalities in clinical populations (Tyrer et al., 2003) and nine times commoner in other populations (Gardiner et al., 2010), so it is no wonder that people with these disorders are not flocking to take part in randomized trials or promoting their pathologies. The exceptions are avoidant personalities, mainly because they overlap with chronic social and generalized anxiety, which are associated with Type S, and antisocial (dissocial) ones, who may not want treatment themselves but who have treatment thrust upon them because society demands it. Here the results are equivocal, with the only robust finding being that programs based on punishment are generally less effective than those based on reward (Gibbon et al., 2010). Randomized trials of treatment in this condition are rare but are feasible, and deserve more attention.

There is a form of treatment for the Type R personality disorders too. Nidotherapy (nest therapy), coined and developed by the author, describes the 'treatment,' not of the patient but of the environment, and is named because a bird's nest is an excellent example of an environmental adjustment to suit whatever is placed within it. For people with personality disturbance who have no wish to change themselves, but are only too willing for things to change around them, nidotherapy is an ideal option and has already been shown to be of value in those with comorbid personality disturbance and severe mental illness (Ranger et al., 2009). The aim of nidotherapy is to explore with the patient the opportunities for environmental change and then to plan a course of action with a timetable in order to achieve this. Examples of successful nidotherapy in those with schizoid personality disturbance can include placement in a job that involves no people contact (e.g., night security guard), in those with antisocial personality disturbance (certain roles within armed services), and in those with obsessive–compulsive personality disturbance (all predictable activities requiring careful checking). In all these instances the aim is not to change the pathology but to make it less discordant with the environment.

Outcomes

The discerning reader will realize that the choice of treatment for those with personality disturbance is not a properly informed one. Almost all the treatments that have demonstrated at least some evidence of effectiveness have been compared with no treatment or standard care. When comparisons have been made between active treatments they have shown discordant findings (e.g., Giesen-Bloo et al., 2006), suggesting that the differences are due to the enthusiasm and orientation of the main investigators rather than any intrinsic superiority of one treatment over any other. In complex trials such as these, reporting bias is considerable and has a major impact on the way results are presented. For this reason NICE guidelines (the guidelines of the National Institute of Health and Clinical Excellence in England) for borderline personality disorder recommended *any* well-structured, theoretically driven, psychological program, of long-term duration, for borderline personality disorder. They do not specify any particular treatment, with the exception of dialectical behavior therapy which is undeniably successful in preventing repeated self-harm in patients with this condition (National Collaborating Center for Mental Health, 2009).

We must remember that the public health implications of personality dysfunction are enormous. It is also clear that treatment should ideally be given early in the course of the disorder. Unfortunately, the current diagnostic systems only allow for the diagnosis of personality disorder when a person reaches the age of 18 years, although this is not backed by any empirical evidence. Fortunately, in ICD-11 the age limits have been removed (Tyrer et al., 2014). Many existing diagnostic boundaries (e.g., conduct disorder/dissocial personality disorder and disruptive mood dysregulation disorder/borderline personality disorder) are artificially set by age. Clinicians should have the option of diagnosing personality disorder at any age, including later in life if the problems are sufficiently persistent. There is bound to be some reluctance to use the diagnosis of personality disorder in young people because of its stigmatizing implications, but if we can bring the attribution of personality disturbance away from prejudice and discrimination there is no reason why it cannot be discussed openly and constructively early in life.

Importance of Personality Disturbance in General Psychiatric Management

One of the advantages of having two axes of classification in DSM was the encouragement to make separate mental state and personality assessments when evaluating patients. In practice this is constantly in clinicians' minds when they choose treatment. Therefore, for example, a clinician would rarely give a potentially dangerous drug such as a monoamine oxidase inhibitor to a person with borderline personality disorder, as the risks would be deemed too great, yet would be willing to give a drug such as amitriptyline in gradually increasing dosage involving the monitoring of adverse effects to an obsessional person who liked to keep a constant check on symptoms and monitor them closely. The presence of personality disturbance could, and should, be a guide to both treatment duration and prognosis. There is now good evidence that the outcome of depression is worse in the presence of personality disorder compared with patients with depression and no personality disorder (Newton-Howes et al., 2014), irrespective of the type of personality disorder. If personality disturbance is postulated to be a diathesis in which sufferers are more vulnerable than others to stresses (Tyrer, 2007), it is not surprising that this vulnerability shows itself in repeated episodes of ill health.

The artificial dichotomy in our current classification that requires clinicians to decide who does or does not have a personality disorder also leads many to avoid using the diagnosis altogether. Many service users feel this diagnosis is unhelpful indeed there is a feeling of collective anger against it. One advantages of having a dimensional range of personality disturbance in which the initial diagnosis is as much one of *current* personality function is that both correction and change are implicit in the diagnostic structure. In addition, the capacity for improvement without losing all personality pathology is a better way of assessing progress than the crude attribution of the presence or absence of a personality disorder.

If we abandon the assessment of personality altogether we would lose a large part of the overall diagnostic view of patients, which is tending more and more to diagnose short-term conditions that can be ephemeral and which have little connection with real life-history and continuing problems in living. 'Pack 'em in, treat 'em quick, get 'em out' is often the covert message given to the young psychiatrist of today. The compression of time frames is accentuated by the fragmentation of services, so that it is now rare for patients to have the chance, even if it is wished, to develop a proper relationship with a psychiatrist. The notion of a personality diathesis, in which the mental state pathology of an individual can be predicted in the short and longer term by knowledge of continuing personality structure, is not easy to foster in this climate.

The Future of Personality Disturbance in Diagnostic and Clinical Practice

Whenever a patient presents to a medical or psychiatric service there should be an assessment of personality status. All current forms of assessment are very crude and rely on self-report or the opinions of others, but it is likely over the next few years that simple biological or psychological measures will be found to give an independent assessment of personality status. Why should this not be relevant in all parts of medicine? In my first junior job as an assistant to an orthopedic surgeon I asked my consultant one day why he was recommending surgery to one lady who had the same problem as another one in the adjacent bed, but for whom he was recommending a slow rehabilitation strategy. 'Because she is a tiger,' he replied, 'the other is a lamb.' His personality assessment, rightly or wrongly, had a major impact on his choice of intervention. The same should happen in all parts of medical and psychiatric assessments. We should not only assess current symptoms and function, but also place the person in a wider context so whatever we decide will fit in well with our acute plan of management.

Thus, for example, if a patient presents in a highly distressed and depressed state to an emergency center with suicidal symptoms the treatment should not always focus on immediate suicide prevention. If the patient is a 65-year-old widow whose son has just been killed in a motor accident, and whose previous mental health has been free of concern, the management should be very different from that of a woman of similar age who has a history of recurrent mental illness and four previous suicide attempts. The evaluation of personality is the key to management, not the number of boxes ticked on a DSM depressive symptom checklist.

At the end of the assessment the formulation of any medical or psychiatric problem would include a statement about personality status and its relevance to treatment. It will not only include aspects of personality dysfunction as described above but also personality strengths and advantageous attributes. Even the most skeptical reader will recognize that, whatever their views about personality disorder as a diagnosis, everyone has a personality, a set of characteristics that make them more or less reliable to adhere to treatments, to follow complex instructions, to act with discretion or on impulse, and to collaborate with their fellows or do things in isolation. All complex and persistent diseases such as hypertension, diabetes, and arthritis have a course and morbidity that is determined as much by their personality as by the specific treatment given. By knowing both personality status and the physical or mental pathological state it is possible to combine them to produce the maximum benefit.

See also: Borderline (Patient) Personality; Obsessive-Compulsive Disorder; Personality Assessment; Personality Theory and Psychopathology; Personality and Health; Personality: Historical and Conceptual Perspectives.

Bibliography

American Psychiatric Association, 1980. Diagnostic and Statistical Manual of Mental Disorders, third ed. American Psychiatric Association (DSM-III), Washington, DC.

American Psychiatric Association, 1987. Diagnostic and Statistical Manual of Mental Disorders, third ed. American Psychiatric Association revised (DSM-III-R), Washington, DC.

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders, fourth ed. American Psychiatric Association (DSM-IV), Washington DC.

Bateman, A.W., Gunderson, J., Mulder R. Treatment of personality disorder. Lancet, in press.

Bateman, A.W., Fonagy, P., 1999. The effectiveness of partial hospitalization in the treatment of borderline personality disorder - a randomised controlled trial. American Journal of Psychiatry 156, 1563–1569.

Bernstein, D.P., Iscan, C., Maser, J., Boards of Directors of the Association for Research in Personality Disorders, 2007. Opinions of personality disorder experts regarding the DSM-IV personality disorders classification system. Journal of Personality Disorders 21, 536–551.

Blum, N., St John, D., Pfohl, B., et al., 2008. Systems training for emotional predictability and problem solving (STEPPS) for outpatients with borderline personality disorder: a randomized controlled trial and 1-year follow-up. American Journal of Psychiatry 165, 468–478.

Bohus, M., Stieglitz, R.D., Fiedler, P., Berger, M., 1999. Persönlich-keitsstörungen. In: Berger, M. (Ed.), Psychiatrie und Psychotherapie. Urban und Schwarzenberg, Munich, pp. 771–846.

Cloninger, C.R., Pryzbeck, T.R., Surakic, D.M., et al., 1994. The Temperament and Character Inventory (TCI): A Guide to its Development and Use. Washington University, St Louis, MO.

Costa, P.T., McCrae, R., 1990. Personality disorders and the five-factor model of personality. Journal of Personality Disorders 4, 362–371.

Davidson, K., Norrie, J., Tyrer, P., et al., 2006. The effectiveness of cognitive behavior therapy for borderline personality disorder: results from the borderline personality disorder study of cognitive therapy (BOSCOT) trial. Journal of Personality Disorders 20, 450–465.

Esquirol, E., 1838. Die Geisteskrankheiten in Beziehung zu Medizin und Staatsarzneikunde. Voss, Berlin.

Ewald, G., 1924. Temperament und Charakter. Springer, Berlin.

Gardiner, C., Tsukagoshi, S., Nur, U., Tyrer, P., 2010. Associations of Type R (treatment resisting) and Type S (treatment seeking) personalities in medical students and their clinical implications. Personality and Mental Health 4, 59–63.

Gibbon, S., Duggan, C., Stoffers, J., et al., 2010. Psychological interventions for antisocial personality disorder. Cochrane Database of Systematic Reviews 6, CD007668.

Giesen-Bloo, J., van Dyck, R., Spinhoven, P., et al., 2006. Outpatient psychotherapy for borderline personality disorder: randomized trial of schema-focused therapy vs transference-focused psychotherapy. Archives of General Psychiatry 63, 649–658.

Gunderson, J.G., Singer, M.T., 1975. Defining borderline patients: an overview. American Journal of Psychiatry 132, 1–10.

Henderson, D., 1939. Psychopathic States. Norton, New York.

Hoch, P.H., Polatin, P., 1949. Pseudoneurotic forms of schizophrenia. Psychiatric Quarterly 23, 248–276.

Kernberg, O.F., 1985. Borderline Conditions and Pathological Narcissism. Jason Aronson, New York.

Koch, J.L.A., 1891–93. Die Psychopathischen Minderwertigkeiten. Maier, Ravensburg, Germany.

Kraepelin, E., 1903–04. Psychiatrie. Ein Lehrbuch für Studierende und Ärzte. Barth, Leipzig, Germany.

Kretschmer, E., 1921. Körperbau und Charakter. Springer, Berlin.

Linehan, M.M., Armstrong, H.E., Suarez, A., Allmon, D., Heard, H.L., 1991. Cognitive-behavioral treatment for chronically parasuicidal borderline patients. Archives of General Psychiatry 48, 1060–1064.

Livesley, W.J., Schroeder, M.L., Jackson, D.N., Jang, K.L., 1994. Categorical distinctions in the study of personality disorder: implications for classification. Journal of Abnormal Psychology 103, 6–17.

Loranger, A.W., Sartorius, N., Andreoli, A., Berger, P., Buchheim, P., Channabasauanna, S.M., Coid, B., Dahl, A., Diekstra, R.F., Ferguson, B., 1994. The international personality disorders examination. Archives of General Psychiatry 51, 215–224.

Morel, B.A., 1857. Traité des dégénérescences physiques, intellectuelles et morales de l'espèce humaine et des causes qui produisent ces variétés maladives. Baillière, Paris.

National Collaborating Centre for Mental Health, 2009. Borderline Personality Disorder, Treatment and Management. The British Psychological Society and The Royal College of Psychiatrists, London.

Newton-Howes, G., Tyrer, P., Johnson, T., Mulder, R., Kool, S., Dekker, J., Schoevers, R., 2014. Influence of personality on the outcome of treatment in depression: systematic review and meta-analysis. Journal of Personality Disorders 28, 577–593.

Partridge, G.E., 1930. Current conceptions of psychopathic personality. American Journal of Psychiatry 10, 53–99.

Pinel, P., 1809. Traité médico-philosophique sur l'aliénation mentale, second ed. Brosson, Paris.

Prichard, J.C., 1835. A Treatise on Insanity and Other Disorders Affecting the Mind. Gilbert & Piper, London, Sherwood.

Ranger, M., Tyrer, P., Milošeska, K., Fourie, H., Khaleel, I., North, B., Barrett, B., 2009. Cost-effectiveness of nidotherapy for comorbid personality disorder and severe mental illness: randomized controlled trial. Epidemiologia e Psichiatria Sociale 18, 128–136.

Rush, B., 1812. Medical Inquiries and Observations upon the Diseases of the Mind. Kimber & Richardson, Philadelphia.

Schneider, K., 1923. Die Psychopathischen Persönlichkeiten. Deuticke, Vienna, Austria (1. Aufl. 1923, Thieme, Leipzig Germany 1950).

Spitzer, R.L., Williams, J.B., 1985. Structured Clinical Interview for DSM-iii, Patient Version. Biometrics Research Department, New York State Psychiatric Institute, New York.

Standage, K.F., 1979. The use of Schneider's typology for the diagnosis of personality disorders–an examination of reliability. British Journal of Psychiatry 135, 238–242.

Tyrer, P., 2007. Personality diatheses: a superior description than disorder. Psychological Medicine 37, 1521–1525.

Tyrer, P., 2009. Why borderline personality disorder is neither borderline nor a personality disorder. Personality and Mental Health 3, 86–95.

Tyrer, P., Mitchard, S., Methuen, C., Ranger, M., 2003. Treatment-rejecting and treatment-seeking personality disorders: type R and type S. Journal of Personality Disorders 17, 265–270.

Tyrer, P., Crawford, M., Mulder, R., on behalf of the ICD-11 Working Group for the Revision of Classification of Personality Disorders, 2011. Reclassifying personality disorders. Lancet 377, 1814–1815.

Tyrer, P., Reed, G.M., Crawford, M. Personality disorder: classification, assessment, prevalence and impact. Lancet, in press.

Widiger, T.A., Costa, P.T., 1994. Personality and personality disorders. Journal of Abnormal Psychology 103, 78–91.

World Health Organization (WHO), 1992. International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10). World Health Organization, Geneva.

Personality, Evolutionary Models of

Hannah J Haysom, University of Queensland, Brisbane, QLD, Australia
Karin JH Verweij, Department of Developmental Psychology, EMGO Institute for Health and Care Research, VU University, Amsterdam, The Netherlands
Brendan P Zietsch, University of Queensland, Brisbane, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Behavioral genetic studies have shown that around a third to a half of the between-individual variation in personality traits can be accounted for by genetic differences between individuals. There is rapidly growing interest in understanding the evolutionary basis of this genetic variation. In this article we review the major explanations that have been proposed to explain the evolutionary maintenance of genetic variation in personality. We cover selective neutrality, various forms of balancing selection, and mutation–selection balance, as well as the possibility that personality trait levels are calibrated to the level of other traits that are themselves under mutation–selection balance.

While evolutionary psychology has had success in identifying and explaining many putative species-typical psychological adaptations, a similar understanding of individual differences remains more elusive (Buss and Hawley, 2011; Penke et al., 2007). Many basic human characteristics are universal in the population: Barring accidents, virtually everyone has the same number of eyes and limbs, for example. Such characteristics may have been historically vital to survival and reproduction, or they might be subject to strong developmental constraints. However, the color of eyes or length of limbs does vary between individuals, and virtually all of the 'universal' traits that interest evolutionary psychologists actually differ considerably between individuals. For example, cross-cultural research shows a 'universal' effect whereby men place greater value than do women on physical attractiveness, whereas women place greater value than do men on a partner's earning potential (Buss, 1989). However, despite this consistent mean difference between men and women's preferences, the within-sex variation between individuals is considerably greater than the between-sex difference. Furthermore, this within-sex variation is not just due to random environmental experiences – genetic variation accounts for a substantial proportion of individual-differences in mate preferences (Verweij et al., 2012a; Zietsch et al., 2012) and virtually all other individual-differences that have been investigated (Bouchard and McGue, 2003).

A key evolutionary question is how and why this genetic variation is maintained in the population despite countless generations of winnowing by natural selection. Some individual differences, such as fingerprint ridge count, have negligible fitness consequences. Variation in such traits is easily explained because they are selectively neutral: there is no winnowing effect. But behavioral traits such as personality (i.e., consistent patterns of behavior (Pervin et al., 2005)), pose a conundrum because they seem at face value to comprise behavioral patterns of high relevance to fitness: social boldness versus shyness (extraversion), tendency to cooperate versus compete (agreeableness), emotional volatility versus stability (neuroticism), and so on. Nevertheless, Tooby and Cosmides (1990) made the early and influential argument that the substantial heritability of personality traits is evidence against personality serving an adaptive function, primarily because natural selection works to deplete variation. Given that virtually all quantitative traits (i.e., traits that vary continuously) are substantially heritable, this position suggests that heritable individual differences in general are evolutionarily unimportant; indeed, Tooby and Cosmides (1990) make the argument that understanding the evolutionary basis of species-typical characteristics is far more important than the study of individual differences. The prominence of Tooby and Cosmides in the nascent field at the time meant that their position probably discouraged evolutionary research on personality and individual differences for much of the field's development.

However, the importance of understanding personality and individual differences from an evolutionary perspective has recently been emphasized in both human (Buss and Hawley, 2011; Penke et al., 2007; Verweij et al., 2012b) and animal (Bell, 2007; Dingemanse and Wolf, 2010; Reale et al., 2010) literatures. Understanding the evolutionary basis of individual differences is crucial to understanding the operation of evolution, because genetic variation is the substrate upon which selection acts. For example, mate choice drives sexual selection, and mates are chosen on their *relative* merits based on individual differences in personality, intelligence, physical attractiveness, etc., while species-typical characteristics are invisible (Buss, 2009).

All personality traits exhibit substantial heritability (heritability is the proportion of variation attributable to genetic variation); behavioral genetic studies over more than 50 years indicate that around a third to a half of the variation in personality traits can be accounted for by genetic differences between individuals (Johnson et al., 2008). In contrast, the family environment has a negligible effect on personality variation (Johnson et al., 2008). The persistent heritability of personality traits presents a theoretical challenge, because natural selection typically works to deplete genetic variation (Barton and Keightley, 2002; Fisher, 1930; Merila and Sheldon, 1999). Alleles that increase fitness tend to become fixed in the population, while those that decrease fitness tend to be eliminated. Given this winnowing effect of selection, the question arises as to why heritable variation remains in the population. We review the major explanations that have been proposed to explain this apparent paradox, outlining some of the evidence

for and against each. We cover selective neutrality, various forms of balancing selection, and mutation–selection balance. We also look at the possibility that personality trait levels are calibrated to the level of other traits ('facultative calibration') that are themselves under mutation–selection balance.

Selective Neutrality

Selectively neutral genetic variants are not subject to natural selection because they have no net effect (positive or negative) on fitness. A simple explanation of why heritable variation in personality is maintained is that personality traits do not affect fitness and that the underlying alleles are therefore not winnowed by selection.

However, genetic variants underlying personality traits seem unlikely to be selectively neutral. Personality traits have been shown to influence a number of fitness-relevant outcomes, including mating success (Nettle, 2005), sexual behavior (Heaven et al., 2000; Zietsch et al., 2010), health (Neeleman et al., 2002), longevity (Friedman et al., 1995), mortality (Mosing et al., 2012), parenting (Spinath and O'Connor, 2003), and reproductive success (Bailey et al., 2013; Eaves et al., 1990). Personality also guides behavioral responses to adaptive problems, such as securing resources, selecting mates, or forming cooperative social alliances (Michalski and Shackelford, 2010). Given that personality is closely linked to fitness outcomes, it seems unlikely that it has remained invisible to natural selection.

However, it is possible that personality traits are related to several fitness components in such a way that positive effects on one component of fitness are canceled out by negative effects on other components of fitness, making the effect on overall fitness approximately uniform (selectively neutral) across the normal personality range (MacDonald, 1995). Expanding on this view, Nettle (2005) proposed concrete cost-benefit trade-offs associated with the Big-Five dimensions of personality variation in humans. High extraversion is predicted to confer benefits of increased status and the associated access to resources and mates, but also the costs of an increased propensity for accidents, disease, and social conflict. Nettle (2006) further extended this line of thinking to the other Big-Five traits. He proposed that high neuroticism may provide increased vigilance to threat at the cost of increased stress and mental illness, while conscientiousness may provide the benefit of increased health and longevity at the cost of rigidity and difficulties adapting to a fluctuating environment. Agreeableness provides advantages in social interaction and cooperative ventures while increasing the risk of one being taken advantage of or cheated, and increased openness to experience may provide the benefit of creativity (an intersexually attractive trait) at the cost of susceptibility to disorganized or potentially psychotic thinking (Nettle, 2006).

As such, although personality traits are demonstrably associated with *components* of fitness, it remains theoretically possible that variations in personality traits have no *net* fitness effects once multivariate relationships are factored in. This idea is concordant with the geometric view of adaptation proposed by Walsh and Blows (2009), in which genetic variants affecting multiple traits can be invisible to selection when multivariate genetic constraints result in little or no variation in fitness effects, even when the individual traits correlate with fitness and have substantial genetic variation.

However, the genetic architecture of selectively neutral traits is expected to match the distribution of allele frequencies that arises from pure genetic drift (i.e., no selection). For a trait under no selection, the distribution of the additive genetic variance explained as a function of minor allele frequency (i.e., the frequency of the less common of the two alleles) is expected to be flat, so that every minor allele frequency range of the same width accounts for the same proportion of variance in the trait (Eyre-Walker, 2010). As such, the vast majority of the genetic variance in a selectively neutral trait should be accounted for by common variants (i.e., variants in which the minor allele is not rare in the population). This is not the case for personality traits, of which most of the genetic variation is not accounted for by common variants (Verweij et al., 2012b; Vinkhuyzen et al., 2012). Further, Verweij et al. (2012b) show that personality traits are impacted by inbreeding, which is predicted of a trait that has been under directional selection, but is not expected for a selectively neutral trait. These findings suggest that genetic variation in personality traits is unlikely to have been maintained due to selective neutrality.

Balancing Selection

An alternative possibility is that selection actively maintains (rather than depletes) variation in personality. This can occur when selection pressures vary across different contexts (Hedrick, 2007). It should be well noted that the presence of context-dependent selection pressures does not necessarily imply that balancing selection is maintaining genetic variation – balancing selection can only maintain genetic variation under quite restrictive conditions (Turelli and Barton, 2004). There are a number of mechanisms of balancing selection, including spatiotemporal environmental heterogeneity, sex-dependent selection, frequency-dependent selection, and heterozygote advantage.

Spatiotemporal Heterogeneity

An intuitively appealing form of balancing selection is caused by selection pressures that vary across space and time – this is called spatiotemporal heterogeneity (Buss, 2008). It is easy to imagine that optimum personality trait values would vary with the environmental conditions, such that the fitness value of a given personality trait will be higher or lower depending on the prevailing niche. For example, relatively safe and stable environments may favor a bold and exploratory personality, while high-risk environments may favor a more conservative and cautious type (Reale et al., 2010). Variation in personality can in theory be maintained when there is no one optimum – different alleles are favored in different environments and so multiple alleles can be maintained in the population.

Several empirical studies have indicated that differential environmental selection pressures produce different personality profiles in both humans and nonhuman animals. Dingemanse et al. (2004) tested the impact of fluctuating

selection pressures on the 'personalities' of great tits, comparing 'fast' exploring, aggressive, and bold birds with 'slow' exploring, nonaggressive, and shy birds. Dingemanse et al. (2004) showed that fluctuating selection pressures influenced male and female birds differently. Rich years favored slow exploring females, while poor years favored fast exploring females. Conversely, fast exploring males benefited in rich years, while slow exploring males were favored in poor years. This suggests that the selection pressures produced by a fluctuating environment could potentially maintain variation, as no one personality phenotype is at a clear advantage or disadvantage to any other type.

Camperio-Ciani et al. (2007) conducted one of the few empirical tests of the effect of varying socioecological niches on the personality profile of a human population. They compared the personality of individuals living on a series of culturally and geographically similar islands off the coast of Italy with their corresponding mainland populations. When compared to mainlanders, the island inhabitants demonstrated higher conscientiousness and emotional stability, but less extraversion and openness to experience. Emigrants who left the island for the mainland scored higher on measures of extraversion and openness than those who had never left the island. Camperio-Ciani et al. (2007) concluded that the significant personality differences between islanders and mainlanders were indicative of genetic adaptation to the particular socioecological niche of the islands.

In a similar vein, Schaller and Murray (2008) investigated the effect of variations in pathogen prevalence on the personality profiles of various cultures worldwide. Schaller and Murray (2008) theorized that in environments with a high risk of disease transmission, natural selection would favor alleles that are associated with lower levels of extraversion, openness, and sociosexuality. They found that cultures who experienced a historically high level of pathogen prevalence exhibited a personality profile characterized by lower levels of sociosexuality, extraversion, and openness to experience. In circumstances where pathogen risk is high, it is likely that selection favors a cautious personality type.

These studies suggest that personality traits are subject to selection pressures that vary across time and space, which is necessary (but not sufficient) for balancing selection via environmental heterogeneity to maintain heritable variation in the population. Both Camperio-Ciani et al. (2007) and Schaller and Murray (2008) describe between-group personality differences that are consistent with expectations from different selection pressures, but no strong inferences can be drawn. First, rather than the observed between-group differences being due to between-group genetic differences resulting from differential selection pressures, the group differences could instead be environmentally caused. There is no evidence to distinguish these possibilities. Second, even if the between-group differences are due to selection-driven genetic differences, this does not imply that within-group heritable individual variation (which is primarily what we are trying to explain) are also due to the same selective processes. Further, while the effect of environmental variation found in these studies may be statistically significant, it only accounts for a very small proportion of the total observed variance in personality.

As well as these limitations, mathematical modeling indicates that balancing selection via environmental heterogeneity can only maintain genetic variants at high frequency in the population – once a variant becomes rare, the balancing mechanism becomes unstable and the rare variant tends to be lost (Turelli and Barton, 2004). In contrast to this prediction, Verweij et al. (2012b) found that most of the genetic variation in personality traits is due to rare variants. As such, it is unlikely that balancing selection via environmental heterogeneity actively maintains much of the variation in personality traits. However, different selection pressures across time or space may still decrease the average strength of directional or stabilizing selection and hence slow the reduction of genetic variation by selection.

Sex-Dependent Selection

Sex-dependent selection occurs when the fitness of an allele depends on the sex of the individual it resides in. As with environmental heterogeneity, this means that there is no one optimum allele at a given locus, and selection can in theory act to maintain genetic variation. There are many examples of sex-dependent selection at the phenotypic level; that is, where certain characteristics or trait levels are more advantageous in one sex than the other (or advantageous in one sex and disadvantageous in the other). Physical examples are the most obvious – for example, height affects attractiveness and reproduction differently in men and women (Stearns et al., 2012; Stulp et al., 2013), and facial masculinity similarly has very different effects on men's and women's attractiveness (Lee et al., 2013; Mitchem et al., 2013). Twin studies have shown for height and masculinity that the same genes have the same effects in men and women; this creates a situation called intralocus sexual conflict, which can maintain genetic variation in the population because the same alleles have different fitness effects in males and females. While the tools exist to test for intralocus sexual conflict in psychological characteristics such as personality traits, little work has been done on this.

Frequency-Dependent Selection

Frequency-dependent selection occurs when the selection on an allele varies depending on frequency of that allele in the population. In negative frequency-dependent selection, selection favors an allele when it is rare, while more common variants are disfavored (Cockerham et al., 1972). The advantage of a rare characteristic relies upon exploiting an environmental niche with few competitors (Burger, 2005). As the characteristic becomes more common and competition increases, its advantage decreases. Frequency-dependent selection also applies to traits that allow individuals to occupy specific niches in the social environment, as well as the physical environment (Bergmuller and Taborsky, 2010). Negative frequency-dependent selection may be a viable mechanism for maintaining personality variation in humans, as it can theoretically maintain substantial amounts of genetic variation (Burger, 2005). Note that positive frequency-dependent selection cannot maintain variation because it disfavors rarer variations, speeding their elimination.

Heterozygote Advantage (Resulting from Antagonistic Pleiotropy)

Heterozygote advantage, or overdominance, can maintain genetic variation when the heterozygote (*Aa*) at a genetic locus has higher fitness than either homozygote (*AA* or *aa*) (Ewens and Thomson, 1970; Fisher, 1922; Hedrick, 2012). Heterozygote advantage can result from antagonistic pleiotropy, which occurs when a gene influences multiple traits such that one trait benefits fitness, while another trait is detrimental to fitness (Hedrick, 1999; Roff and Fairbairn, 2006). The best-known example of heterozygote advantage in nature is at the gene that causes sickle-cell anemia in humans. Individuals with an *AA* genotype are susceptible to malaria, while those with an *aa* type have an increased risk of mortality from sickle cell anemia. Those with the heterozygote genotype, *Aa*, however, do not develop anemia and are more likely to survive a malarial infection, which gives these individuals a significant fitness advantage (Hedrick, 2011). As the alleles, and not genotypes, are passed on to subsequent generations, variation can be maintained when the average fitness of each allele remains equal. The rarer an allele becomes in the population, the more likely it will become part of a heterozygote genotype, at which point its fitness, and eventually frequency, will increase. However, while this mechanism can demonstrably successfully maintain genetic variation (Ewens and Thomson, 1970), only a handful of genes with polymorphisms maintained by heterozygote advantage have ever been identified (Endler, 1986; Hedrick, 2012).

Summary

Overall, balancing selection seems an unlikely mechanism for maintaining genetic variation in personality. Balancing selection can only maintain genetic variation when highly restrictive conditions are met (Burger and Gimelfarb, 2002; Curtsinger et al., 1994; Hedrick, 1999; Turelli and Barton, 2004), and is often evolutionarily unstable (Curtsinger et al., 1994; Turelli and Barton, 2004). Mathematical modeling suggests that most forms of balancing selection – sex-dependent selection, frequency-dependent selection, and all forms of heterozygote advantage (including from antagonistic pleiotropy) – can only maintain variation in a few genes. Furthermore, all forms of balancing selection predict that genetic variation in personality should be due to common variants (Burger, 2000; Curtsinger et al., 1994; Hedrick, 1999; Kopp and Hermisson, 2006; Mani et al., 1990; Turelli and Barton, 2004).

In a study comparing the predictions of selective neutrality, balancing selection, and mutation–selection balance, Verweij et al. (2012b) found that the genetic architecture of personality traits is not consistent with the predictions of balancing selection. For example, the study showed that only a small proportion of the genetic variance in personality was accounted for by common additive variants, which is predicted by mutation–selection balance, but not by balancing selection (Verweij et al., 2012b). Verweij et al. (2010) also showed that personality is underlain by a very large number of genes of very small effects. Combined, these results are inconsistent with the predictions of any form of balancing selection. This conclusion is further reinforced by Bubb et al. (2006), who conducted a scan of the human genome to look for signals of balancing selection (regions of the genome with greater-than-chance densities of single nucleotide polymorphisms), and found no evidence for ancestral balancing selection beyond the few previously known examples. These various sources of evidence suggest that balancing selection is not a major cause of extant genetic variation in humans.

Mutation–Selection Balance

Under mutation–selection balance, genetic variation in a trait is maintained because the loss of variation from selection is offset by the accumulation of new mutations, so that selection and mutation exist in a state of equilibrium (Barton and Keightley, 2002; Zhang and Hill, 2005). Most mutations are at most mildly harmful to the fitness of an individual, and so are not immediately removed by selection. As a result, substantial amounts of genetic variation can potentially be maintained in mutation–selection balance (Gangestad, 2010). Keller and Miller (2006) argued that mutation–selection balance is the most likely mechanism of maintaining heritable mental disorders, which otherwise would be rapidly eliminated from the population due to their deleterious effects on fitness.

Selection will not usually eliminate new mutations immediately unless they are catastrophically deleterious (e.g., cause unviability or sterility). Because mildly deleterious mutations can accumulate over generations, each individual carries a load of mutations of variable number and severity. Alleles with large and/or dominant deleterious effects are rapidly eliminated, so deleterious alleles of small and recessive effects are more likely to be retained over many generations (Barton and Keightley, 2002). A mutation causing a 1% decrease in overall fitness will be carried by an average of around 100 individuals before being eliminated (Keller and Miller, 2006). Individuals pass on around half of their mutation load to their offspring, so those who exhibit a lower mutation load (perhaps indicated by traits such as physical attractiveness and intelligence) are more desirable as mates (Gangestad, 2010; Penke et al., 2007). Recent estimates indicate that new mutations in humans are introduced at a rate of 1.20×10^{-8} per nucleotide per generation (Kong et al., 2012). In the absence of selection or random drift the heritability of a trait would increase by approximately .1–1% each generation due to mutation (Barton and Keightley, 2002).

Zhang and Hill (2005) investigated whether mutation–selection balance could account for substantial heritable trait variation, in response to previous work which suggested that mutation–selection balance alone may not generate a sufficient amount of genetic variation to be a plausible mechanism of trait variation. Under mutation–selection balance, selection does not necessarily work in a directional fashion, but rather may favor the average phenotype over either extreme variation (i.e., stabilizing selection). Zhang and Hill (2005) analyzed a joint model of mutation–selection balance, where stabilizing selection was assessed in conjunction with 'apparent stabilizing selection,' in which mutations affecting a focal trait pleiotropically influence the fitness of other traits. Where previous studies had determined that neither mechanism alone was sufficient to sustain high levels of genetic variation, Zhang and Hill (2005) determined that their joint model of mutation–selection

balance could plausibly maintain genetic variation at high levels.

Using high-density genotype data and personality scores from thousands of individuals, Verweij et al. (2012b) directly compared mutation–selection balance, selective neutrality, and balancing selection as alternative explanations of personality trait variation. The study compared the genetic similarity of individuals to their personality similarity to determine how much of the variance in personality was accounted for by common causal variants. Verweij et al. (2012b) determined that common additive causal variant effects explained only a small proportion (~25%) of the genetic variation in personality, and that most of the genetic variation was due to rare variants and nonadditive genetic effects. From the genetic data, distant inbreeding could also be quantified for each individual, allowing detection of directional dominance of genetic effects. This analysis indicated that greater inbreeding was associated with the low-fitness ends of personality traits, suggesting that deleterious alleles were more likely to be recessive, which is a signature of mutation–selection balance. Combined, these results are consistent with a mutation–selection balance explanation of personality, but not with the explanations of selective neutrality or balancing selection (Verweij et al., 2012b).

Facultative Calibration

Consistent with the idea that personality variation has been maintained by mutation–selection balance is the suggestion by some researchers that an individual's personality trait levels have evolved to be calibrated to the level of other condition-dependent traits that are themselves under mutation–selection balance (Buss, 2008; Lukaszewski, 2013; Lukaszewski and Roney, 2011). For example, Lukaszewski and Roney (2011) proposed a 'facultative calibration' model of the personality trait extraversion: they propose that extraverted behavior is adaptive for attractive individuals (and strong individuals, in the case of males), but maladaptive for unattractive individuals (and weak males). Strong and/or attractive individuals may be better able to exploit the benefits of extraversion, such as social status or access to resources and mates, while minimizing the risks or costs, such as accidents, disease, and social conflict (Lukaszewski and Roney, 2011). Facultative calibration of extraversion therefore predicts that the optimum level of extraversion will vary between individuals, based on their other phenotypic qualities. According to this model, personality exhibits heritability because it is calibrated to variation in other condition-dependent heritable traits, a concept termed 'reactive heritability' (Tooby and Cosmides, 1990). Lukaszewski and Roney (2011) found that higher levels of extraversion were correlated with higher levels of strength and attractiveness, consistent with a facultative calibration model. Lukaszewski (2013) recently expanded the idea to other personality traits and a central construct to which these personality traits were calibrated: 'relative bargaining power.' Again, correlational evidence was largely consistent with the facultative calibration hypothesis.

Although examining phenotypic correlations is an important first step in testing the facultative calibration hypothesis, genetically informative studies are required to properly test these ideas. Findings from identical and nonidentical twins (Haysom et al., Manuscript submitted for publication) cast doubt on key aspects of the facultative calibration hypothesis: Most notably, only a very small proportion of the heritability of extraversion can be accounted for by physical attractiveness and other traits related to relative bargaining power.

Conclusions

While there has been substantial theoretical exploration of the possible mechanisms underlying heritable personality variation, there is a strong need for more empirical research. The empirical research that does exist suggests that selection pressures vary across time and space (Camperio-Ciani et al., 2007; Dingemanse et al., 2004; Schaller and Murray, 2008), but there is no direct evidence that this spatiotemporal heterogeneity actively contributes to maintaining heritable variation within populations. Indeed, genetic evidence is inconsistent with the genetic architecture that would be expected if selection were actively maintaining variation in personality traits (Bubb et al., 2006; Verweij et al., 2012b). In contrast, genetic evidence is consistent with predictions from mutation–selection balance, where the multigenerational accumulation of random, mildly deleterious mutations is offset by selection against those mutations (Verweij et al., 2012b). Further theoretical and empirical work is needed to understand how random mutational variation manifests in a dimensional personality structure, and how dimensions of individual differences are related to species- or sex-typical behavioral adaptations (see Zietsch, 2009: pp. 85–99).

See also: Extraversion; Five Factor Model of Personality, Facets of; Personality Differences and Development: Genetic and Environmental Contributions; Personality and Adaptive Behaviors; Personality, Biological Models of; Personality, Trait Models of.

Bibliography

Bailey, D.H., Walker, R.S., Blomquist, G.E., Hill, K.R., Hurtado, A.M., Geary, D.C., 2013. Heritability and fitness correlates of personality in the ache, a natural-fertility population in Paraguay. PLoS ONE 8 (3), 1–10.

Barton, N.H., Keightley, P.D., 2002. Understanding quantitative genetic variation. Nature Reviews Genetics 3 (1), 11–21.

Bell, A.M., 2007. Evolutionary biology: animal personalities. Nature 447 (7144), 539–540.

Bergmuller, R., Taborsky, M., 2010. Animal personality due to social niche specialisation. Trends in Ecology and Evolution 25 (9), 504–511.

Bouchard, T.J., McGue, M., 2003. Genetic and environmental influences on human psychological differences. Journal of Neurobiology 54 (1), 4–45.

Bubb, K.L., Bovee, D., Buckley, D., Haugen, E., Kibukawa, M., Paddock, M., Palmieri, A., Subramanian, S., Zhou, Y., Kaul, R., Green, P., Olson, M.V., 2006. Scan of human genome reveals no new loci under ancient balancing selection. Genetics 173 (4), 2165–2177.

Burger, R., 2000. The Mathematical Theory of Selection, Recombination and Mutation. Wiley, Chichester, UK.

Burger, R., 2005. A multilocus analysis of intraspecific competition and stabilizing selection in a quantitative trait. Journal of Mathematical Biology 50, 355–396.

Burger, R., Gimelfarb, A., 2002. Fluctuating environments and the role of mutation in maintaining quantitative genetic variation. Genetical Research 80, 31–46.

Buss, D.M., 1989. Sex-differences in human mate preferences: evolutionary hypotheses tested in 37 cultures. Behavioral and Brain Sciences 12 (1), 1–14.

Buss, D.M., 2008. Human nature and individual differences: evolution of human personality. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality, third ed. Guilford Press, New York.

Buss, D.M., 2009. How can evolutionary psychology successfully explain personality and individual differences? Perspectives on Psychological Science 4 (4), 359–366.

Buss, D.M., Hawley, P.H. (Eds.), 2011. The Evolution of Personality and Individual Differences. Oxford University Press, New York.

Camperio-Ciani, A.S., Capiluppi, C., Veronese, A., Sartori, G., 2007. The adaptive value of personality differences revealed by small island population dynamics. European Journal of Personality 21, 3–22.

Cockerham, C.C., Burrows, D.M., Young, S.S., Prout, T., 1972. Frequency-dependent selection in randomly mating populations. The American Naturalist 106 (950), 493–515.

Curtsinger, J.W., Service, P.M., Prout, T., 1994. Antagonistic pleiotropy, reversal of dominance, and genetic polymorphism. The American Naturalist 144 (2), 210–228.

Dingemanse, N.J., Both, C., Drent, P.J., Tinbergen, J.M., 2004. Fitness consequences of avian personalities in a fluctuating environment. Proceedings of the Royal Society of London Series B - Biological Sciences 271 (1541), 847–852.

Dingemanse, N.J., Wolf, M., 2010. Recent models for adaptive personality differences: a review. Philosophical Transactions of the Royal Society B – Biological Sciences 365 (1560), 3947–3958.

Eaves, L.J., Martin, N.G., Heath, A.C., Hewitt, J.K., 1990. Personality and reproductive fitness. Behavior Genetics 20 (5), 563–568.

Endler, J.A., 1986. Natural Selection in the Wild. Princeton University Press, Princeton, NJ.

Ewens, W.J., Thomson, G., 1970. Heterozygote selective advantage. Annals of Human Genetics 33 (4), 365–376.

Eyre-Walker, A., 2010. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proceedings of the National Academy of Sciences of the United States of America 107, 1752–1756.

Fisher, R.A., 1922. On the dominance ratio. Proceedings of the Royal Society of Edinburgh 42, 321–341.

Fisher, R.A., 1930. The Genetical Theory of Natural Selection. Oxford University Press, Oxford.

Friedman, H.S., Tucker, J.S., Schwartz, J.E., Martin, L.R., Tomlinson-Keasey, C., Wingard, D.L., Criqui, M.H., 1995. Childhood conscientiousness and longevity: health behaviors and cause of death. Journal of Personality and Social Psychology 68 (4), 696–703.

Gangestad, S.W., 2010. Evolutionary biology looks at behavior genetics. Personality and Individual Differences 49 (4), 289–295.

Haysom, H.J., Mitchem, D.G., Lee, A.J., Wright, M.J., Martin, N.G., Keller, M.C., Zietsch, B.P., Manuscript submitted for publication. No support for the reactive heritability model of extraversion.

Heaven, P.C.L., Fitzpatrick, J., Craig, F.L., Kelly, P., Sebar, G., 2000. Five personality factors and sex: preliminary findings. Personality and Individual Differences 28, 1133–1141.

Hedrick, P.W., 1999. Antagonistic pleiotropy and genetic polymorphism: a perspective. Heredity 82 (2), 126–133.

Hedrick, P.W., 2007. Balancing selection. Current Biology 17 (7), R230–R231.

Hedrick, P.W., 2011. Population genetics of malaria resistance in humans. Heredity 107 (4), 283–304.

Hedrick, P.W., 2012. What is the evidence for heterozygote advantage selection? Trends in Ecology & Evolution 27 (12), 698–704.

Johnson, A.M., Vernon, P.A., Feiler, A.R., 2008. Behavioral genetic studies of personality: an introduction and revivew of the resutls of 50+ years of research. In: Boyle, G., Matthews, G., Saklofske, D. (Eds.), Handbook of Personality Theory and Assessment. Sage, London, pp. 145–173.

Keller, M.C., Miller, G., 2006. Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? Behavioral and Brain Sciences 29, 385–452.

Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Wong, W., Sigurdsson, G., Walters, G.B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D.F., Helgason, A., Magnusson, O.T., Torsteinsdottir, U., Stefansson, K., 2012. Rate of *de novo* mutations, father's age, and disease risk. Nature 488 (7412), 471–475.

Kopp, M., Hermisson, J., 2006. The evolution of genetic architecture under frequency-dependent disruptive selection. Evolution 60 (8), 1537–1550.

Lee, A.J., Mitchem, D.G., Wright, M.J., Martin, N.G., Keller, M.C., Zietsch, B.P., 2013. Genetic factors increasing male facial masculinity decrease facial attractiveness of female relatives. Psychological Science electronic publication ahead of print.

Lukaszewski, A.W., 2013. Testing an adaptationist theory of trait covariation: relative bargaining power as a common calibrator of an interpersonal syndrome. European Journal of Personality 328–345.

Lukaszewski, A.W., Roney, J.R., 2011. The origins of extraversion: joint effects of facultative calibration and genetic polymorphism. Personality and Social Psychology Bulletin 37 (3), 409–421.

MacDonald, K., 1995. Evolution, the five-factor model, and levels of personality. Journal of Personality 63 (3), 526–567.

Mani, G.S., Clarke, B.C., Shelton, P.R., 1990. A model of quantitative traits under frequency-dependent balancing selection. Proceedings of the Royal Society B 240 (1297), 15–28.

Merila, J., Sheldon, B.C., 1999. Genetic architecture of fitness and nonfitness traits: empirical patterns and development of ideas. Heredity 83, 103–109.

Michalski, R.L., Shackelford, T.M., 2010. Evolutionary personality psychology: reconciling human nature and individual differences. Personality and Individual Differences 48, 509–516.

Mitchem, D.G., Purkey, A.M., Grebe, N.M., Carey, G., Garver-Apgar, C.E., Bates, T.C., Arden, R., Hewitt, J.K., Medland, S.E., Martin, N.G., Zietsch, B.P., Keller, M.C., 2013. Estimating the sex-specific effects of genes on facial attractiveness and sexual dimorphism. Behavior Genetics 1–12.

Mosing, M.A., Medland, S.E., McRae, A., Landers, J.G., Wright, M.J., Martin, N.G., 2012. Genetic influences on life span and its relationship to personality: a 16-year follow-up study of a sample of aging twins. Psychosomatic Medicine 74 (1), 16–22.

Neeleman, J., Sytema, S., Wadsworth, M., 2002. Propensity to psychiatric and somatic ill-health: evidence from a birth cohort. Psychological Medicine 32 (05), 793–803.

Nettle, D., 2005. An evolutionary approach to the extraversion continuum. Evolution and Human Behavior 26, 363–373.

Nettle, D., 2006. The evolution of personality variation in humans and other animals. American Psychologist 61 (6), 622–631.

Penke, L., Denissen, J.J.A., Miller, G.F., 2007. The evolutionary genetics of personality. European Journal of Personality 21, 549–587.

Pervin, L.A., Cervone, D., John, O.P., 2005. Personality Theory and Research, ninth ed. Wiley, Hoboken, NJ.

Reale, D., Dingemanse, N.J., Kazem, A.J.N., Wright, J., 2010. Evolutionary and ecological approaches to the study of personality. Philosophical Transactions of the Royal Society B 365, 3937–3946.

Roff, D.A., Fairbairn, D.J., 2006. The evolution of trade-offs: where are we? Journal of Evolutionary Biology 20, 433–447.

Schaller, M., Murray, D.R., 2008. Pathogens, personality, and culture: disease prevalence predicts worldwide variability in sociosexuality, extraversion, and openness to experience. Journal of Personality and Social Psychology 95 (1), 212–221.

Spinath, F.M., O'Connor, T.G., 2003. A behavioral genetic study of the overlap between personality and parenting. Journal of Personality 71 (5), 785–808.

Stearns, S.C., Govindaraju, D.R., Ewbank, D., Byars, S.G., 2012. Constraints on the coevolution of contemporary human males and females. Proceedings of the Royal Society B: Biological Sciences.

Stulp, G., Buunk, A.P., Pollet, T.V., 2013. Women want taller men more than men want shorter women. Personality and Individual Differences 54 (8), 877–883.

Tooby, J., Cosmides, L., 1990. On the universality of human nature and the uniqueness of the individual: the role of genetics and adaptation. Journal of Personality 58 (1), 17–67.

Turelli, M., Barton, N.H., 2004. Polygenic variation maintained by balancing selection: pleiotropy, sex-dependent allelic effects and G x E interactions. Genetics 166, 1053–1079.

Verweij, K.J.H., Burri, A.V., Zietsch, B.P., 2012a. Evidence for genetic variation in human mate preferences for sexually dimorphic physical traits. PloS One 7, e49294.

Verweij, K.J.H., Yang, J., Lahti, J., Veijola, J., Hintsanen, M., Pulkki-Raback, L., Heinonen, K., Pouta, A., Pesonen, A.K., Widen, E., Taanila, A., Isohanni, M., Miettunen, J., Palotie, A., Penke, L., Service, S.K., Heath, A.C., Montgomery, G.W., Raitakari, O.T., Kahonen, M., Viikari, J., Raikkonen, K., Eriksson, J.G., Keltikangas-Jarvinen, L., Lehtimaki, T., Martin, N.G., Jarvelin, M.-R., Visscher, P.M., Keller, M.C., Zietsch, B.P., 2012b. Maintenance of genetic variation in human personality: testing evolutionary models by estimating heritability due to common causal variants and investigating the effect of distant inbreeding. Evolution 66, 3238–3251.

Verweij, K.J.H., Zietsch, B.P., Medland, S.E., Gordon, S.D., Benyamin, B., Nyholt, D.R., McEvoy, B.P., Sullivan, P.F., Heath, A.C., Madden, P.A., Henders, A.K., Montgomery, G.W., Martin, N.G., Wray, N.R., 2010. A genome-wide association study of Cloninger's temperament scales: implications for the evolutionary genetics of personality. Biological Psychology 85, 306–317.

Vinkhuyzen, A.E., Pedersen, N.L., Yang, J., Lee, S.H., Magnussen, P.K.E., Iacono, W.G., McGue, M., Madden, P.A.F., Heath, A.C., Luciano, M., Payton, A., Horan, M., Ollier, W., Pendleton, N., Deary, I.J., Montgomery, G.W., Martin, N.G., Visscher, P.M., Wray, N.R., 2012. Common SNPs explain a significant proportion of the 'missing heritability' in the personality dimensions of neuroticism and extraversion. Translational Psychiatry e102.

Walsh, B., Blows, M.W., 2009. Abundant genetic variation plus strong selection = multivariate genetic constraints: a geometric view of adaptation. Annual Review of Ecology Evolution and Systematics 40, 41–59.

Zhang, X., Hill, W.G., 2005. Genetic variability under mutation selection balance. Trends in Ecology and Evolution 20 (9), 468–470.

Zietsch, B.P., 2009. The Genetic Etiology of Human Sexuality. University of Queensland, Brisbane.

Zietsch, B.P., Verweij, K.J.H., Bailey, J.M., Wright, M.J., Martin, N.G., 2010. Genetic and environmental influences on risky sexual behaviour and its relationship with personality. Behavior Genetics 40, 12–21.

Zietsch, B.P., Verweij, K.J.H., Burri, A.V., 2012. Heritability of preferences for multiple cues of mate quality in humans. Evolution 66, 1762–1772.

Personality: Historical and Conceptual Perspectives

Frank Dumont, McGill University Montreal, Hudson, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article traces the 24-century evolution of embryonic concepts of personality psychology into a fully effloresced scientific, contemporary domain. This analysis links the rise of Western individualism to the empirical study of differences in temperament and personality. Focus is placed on the biological, empirical substrate of personality development sketched in Hellenic science, the negative psychology of early Middle East and European religious movements in the West, and their gradual mutation into the positive psychology of a secular, evidence-based personality and developmental psychology.

Defining Personality

The concepts that are useful today in advancing personality psychology have evolved over thousands of years. In the West, psychology in all its guises was long subsumed under the rubric 'philosophy.' This rubrical issue does not invalidate the idea that the study of personality had been sluggishly maturing over millennia before it effloresced in the nineteenth century as a formal science. The term personality is now typically used to label consistent behavioral, cognitive, and emotional reaction patterns that individuals experience and predictably exhibit. In this context one can take it as axiomatic that even a spontaneous thought is a stimulus to one's brain and internal milieu. Further, it is assumed that categorical terms such as 'personality,' 'traits,' 'factors,' or even, say, 'species,' are mental constructs; they do not exist independently in nature, though they do have a foundation in reality. Thus one can only be said to actually possess a personality in the sense that one is assessed by others and viewed by oneself as disposed to behave consistently and predictably in a variety of situations.

Given the complexity of human beings and the data-dense web of relationships in which they promote their well-being as they perceive it, personality psychology was progressively split into numerous specialties, for example, trait psychology, self psychology, emotion psychology, gendered personality, motivational psychology, disordered personality, neuroethology, cultural psychiatry, and many others. Given the scope of personality psychology as a field of study, one finds, not surprisingly, that it implicates many disciplines beyond psychology – some broad-scale such as anthropology, sociology, evolutionary biology, ethology, and neuroscience. Others, such as cognitive neuropsychology or social role theory, are narrower in scope, though intrinsically complex in themselves.

Conversely, as Moghaddam (1998), a social psychologist, and Pinker (2002), a cognitive psychologist, have cogently explained, there are hundreds of specific behavior patterns (and schemas for calibrating and valuing relationships) that are universal, that is, common to the human species. (Linguists and cultural anthropologists have developed the notions of *emic* and *etic*, neologisms that are highly useful to designate culture-specific and cross-cultural traits, respectively. Psychologists have borrowed the terms to designate analogous concepts in their respective disciplines.) Even more broadly, certain personality theories have been useful in comparative psychology, exploring the characters, say, of canines, bulls, octopi, rats, (For example, Jaak Panksepp and Jeff Burgdorf authored 'Laughing' rats and the evolutionary antecedents of human joy? *Physiology & Behavior*, 79 (2003) 533–547.) or even microbes, ("Motile microorganisms of the same species are like solitary eccentrics in their swimming behavior. When they are searching for food, some tumble in one direction for precisely so many seconds before quitting, while others tumble differently and for different, but characteristic, periods of time. If you watch them closely…you can tell them from each other by the way they twirl, as accurately as though they had different names" (Thomas, 1979: 2).) and their interrelationships with humans as well as among themselves.

In this perspective, the term personality designates not only the distinctive social and behavioral differences we observe between individuals, but also distinctive cultural patterns among nations – and, indeed, some behavioral norms that are found in all societies. Thus personality can be construed as a multidimensional, continuous variable stretching between molar and molecular poles, from patterns of personal behavior that are characteristic of the human species (see Brown, D.E., 1991) to those of a single human being. Near the molar pole this evokes etic, anthropological schemas; toward the molecular pole emic, personological ones.

Principles and Disciplines

Early Biological Substrates of Personality

Hippocrates (born *c*.460 BC) was late pre-Socratic in inspiration. He understood that "from nothing else but the brain come joys, delights, laughter and sports, and sorrows, griefs, despondency, and lamentations" (1992: 336) – not to mention madness, immobilizing anxieties, and delirium. He deplored attributing mental illnesses to divine causes and averred that we must seek answers in natural causes – and remedies in human resources. Plato noted in *Timaeus* (86e ff: 474) that humans' varying physiological conditions can cause turmoil and disease in their spirits. He endorsed the Hippocratean notion that bodily humors (blood, phlegm, yellow bile, and black bile) influence population variability of temperaments. The legacy of this archaic but useful formulation of personality traits (sanguine, choleric, phlegmatic, melancholic) will be evident to students of Carl Gustav Jung (and of the 'Myers–Briggs Type Indicator,' a popular

assessment instrument that is a derivative of Jungian personality theory). This Hippocratean schema conforms roughly to Hans Eysenck's four-factor model of trait psychology (alternatively compacted into a three-factor model). Although analogous Chinese and Indian schemas can be cited here, their impact on contemporary personality psychology is not as clear or immediate.

Personality in Later Hellenism

Many other concepts that are useful today in exploring the science of personality psychology can be traced back to the Hellenic era. For example, the Aristotelian principle that the mind is, as it were, a blank slate destined to be written upon was reinvented in the later Renaissance and has since been extensively modified in the modern sciences of developmental psychogenetics. Notable contributions were made by Plato and Aristotle, the 'twin stars' of Hellenic thought as Arthur Koestler called these philosophers.

Among countless personological examples, one can cite Plato's (370 BCE/1991) observations about the character of young boys. In *Laws* he wrote: "Now of all wild young things a boy is the most difficult to handle. Just because he more than any other has a fount of intelligence in him that has not yet run 'clear', he is the craftiest, most mischievous, and most unruly of brutes. So the creature has to be held in check, as we may say, by more than one bridle" (Book 7, Section 808d). Yet Plato cautions that the privacy of the home itself conceals (presumably abusive) familial reactions to children's passions and outbursts that can bring on 'a medley of incongruities' into their characters. This and numerous other astute observations can be endorsed by all contemporary developmental psychologists.

Plato insisted there should be compulsory schooling for all, both males and females, as the care, instruction, and nurturance of all children will shape their adult character. Tilting against a prevalent misogynism, to which he was not totally immune, he asked, what can be more absurd than depriving women the opportunity to follow the "same pursuits [as men] with all their strength and with one mind" – and here note the economic kicker – "for thus the state, instead of being a whole is reduced to a half, but has the same imposts to pay and the same toils to undergo" (Plato, Laws. *c*.350 BCE/1997, [804]: 1473). He noted moreover "We are not all alike; there are diversities of natures among us which are adapted to different occupations" (Plato, 4th *c*.BCE/1952a. *The Republic* Book II, Section 370: 317). Here we have not only a precursor to the career psychology principle, 'aptitude–occupation fit,' but a salient recognition of individual differences. Aristotle provided equally cogent views on individual differences and on the influence of learning and social forces that shape personality. Attention to personality types can be found in his *Rhetoric* and his *Ethics*. His student Theophrastus (*c*.4th *c*.BCE/1929), many of whose writings have disappeared, left arresting portraits of certain personality types.

Augustine in the Ascendant

Gnostic beliefs of numerous religious sects, freighted with mystic conceptions of the world and negative views of human nature, contested the brilliant science of the Pythagoreans and later Hellenes (see Jonas, 1958; Robinson, 1977). Certain elements of Gnostic belief systems infiltrated early Judaic traditions and a nascent Christianity, and powerfully shaped the evolution of the concept of the Human in the first millennium CE. Though there were many varieties of these Gnostic sects, traceable to Persian and other Middle Eastern cultures that preceded Roman hegemony, they all shared beliefs in domains of light and dark, truth and falsehood, spirituality and materiality, good and evil supernatural agents, and, in some sects, the radically flawed nature of 'Man.' (Scholars note that certain sects declaimed that "the very origin of the world was attribut[able] to a terrible fault" (e.g., Robinson, 1977: 4) and that Evil, destined to rule, was abundantly evident in secular societies. In sharp contrast to this mystical ethos Plato, in *Timaeus* (4th *c*.BCE/1952b, [86, 87]), affirmed the contrary that humans are not voluntarily bad but become bad by virtue of physiological disorders and a bad education. Further, these entail psychological disorders that in varying degrees of intensity are aggravated by evil discourse uttered in public or private; and by evil government, which can worsen the character of the state's citizens (474).) These dualistic systems shaped several streams of thought bearing on the nature of the Human. In the West they dominated up to the Renaissance – and in attenuated form even into the nineteenth century when they spawned pathognomonic schemas of human nature.

The seepage of select eremitical and gnostic ideas and values into the Christian systems that mushroomed during the first millennium (CE) had a stultifying impact on the pursuit of Hellenic science. The most potent influence was St Augustine's, whose conversion to Christianity from Manicheism allowed the infusion of a quasignostic ethos, no doubt indeliberately, into the structure of what had become Christian orthodoxy. Those attitudes were enmeshed with a prevalent Neoplatonism, which, while admitting the goodness of the material world, nevertheless denigrated humans' carnal nature. Augustine's ascetic counsels may seem egregious to readers of our times, yet they were prevalent in the anguished and despair-ridden epoch that ushered in the European Dark Ages. ("To escape temporal encapsulation in the present, hermeneuticists have provided exegetes and historians with a scientific methodology for understanding the thinking of individuals who were immersed in a historical and cultural context different from their own. Initially developed to interpret Sacred Scripture, hermeneutics protects scholars from anachronistic perspectives and is no less important for historical texts bearing on psychology" (Dumont, 2010: 1–2).)

No doubt the antitheses of theocentric asceticism and geocentric science played out to the detriment of the latter. Archimedes' *The Sand Reckoner*, which investigated out of sheer curiosity the number of grains of sand that it would take to literally fill the universe, must have stupefied Augustine in his mature years – committed as he was to exploring exclusively the mysteries of the supernatural and to approaching the Divine Object of his devotion. Indeed, he considered studying the course of the stars and other phenomena of the natural world as morbid vanity. Such an important personality trait as curiosity was in his estimation, an illness. This view prevailed for centuries.

Augustine's views on other human proclivities would have a profound influence on personological approaches to daily life in the Dark Ages that followed upon the decline of the Roman Empire. He viewed our senses as bridges to seductive worldly pleasures (Book X, Chapters 31–35 of *The Confessions*). He deplored the enjoyment of ecclesiastical music unless attention was given more to the meaning of the words than the beauty of the sound. Comparing 'enjoyment' to 'attention' raises a scaling problem he did not address. Bach's *Matthew Passion* or Handel's *Messiah* would have presented him with challenging temptations. He extolled the value of fasting and deplored taking pleasure in eating even when famished – indeed he regarded food and drink only as medicine to repair the wastage of the body (*The Confessions*: 103). A twenty-first-century psychologist would demur. The delights taken in a family dinner tastefully prepared by its members is a bonding process for the entire family. Enjoying our food and drink is no more separable from the conviviality of a shared meal than is separating the butter from the batter.

The Dark Ages

The centers of learning in the Dark Ages were dedicated more to religious than philosophical pursuits – though they were somewhat intermeshed. The monasteries were the epiphenomena of a mushrooming Christianity more than academies to engage in Greek and Roman intellectual pursuits. Although the works of Lucretius, Marcus Aurelius, Plutarch, and Cicero, among others, were studied by individual scholars, the copying of religious manuscripts and the evangelization of the local populace were of greater importance to monastics. The cognitive structures of the peasantry, no less than monastic communities, were suffused with the ideals of the Church Fathers, most notably Jerome, Augustine, and Gregory the Great. Even rulers like Charlemagne took inspiration from principles from Augustine's *City of God*. There is little evidence this emperor read philosophy and the work of the Neoplatonists.

Five centuries elapsed between the Augustinian *oeuvre* and the appearance of an equally great thinker, the brilliant John Scotus Eriugena, who revived a dormant Neoplatonism, and seven centuries before Moses Maimonides, the Spanish-born Judaic, Arabist, Hellenist polymath, appeared. In the earlier interval only Boethius, a sixth-century politician of uncertain religious beliefs and author of the *Consolations of Philosophy* (and other philosophical works) illuminated the post-Patristic period. Little information survives suggesting people were viewed as distinctly individualized. Our modern concept of an individual who is documented as such with a birth certificate and later a social insurance number, endowed perhaps with a medical record, a coded driver's license, a Medicare card, and a (typically) inalienable passport would have struck the denizens of Europe in the Dark Ages as preposterous. Individuals are now concerned that their identity can be stolen – and that they can be impersonated in a way that diminishes their rights, status, and property. Though impersonations no doubt happened, there is little evidence that such even entered the theater of, say, a serf's waking state. The emergence of a post-classicist individualism and the notion of individual differences lay in the future.

The Sluggish Rise of Individualism

In our modern developed world each of us normally finds a space that affords us solitude. In the first millennium, life was predominantly communal. Private bedrooms and baths were not the norm nor are they today in many parts of the world; neither were looking glasses in plentiful supply, to aid in personal self-grooming. The individual was less a person than an idea as Philippe Ariès (1962) suggests. Neoplatonic essentialism dominated the existentialism that was later to appear in the music, poetry, and theater of popular entertainment. Portraiture in the later Middle Ages depicts members of a family of varying ages in such wise that they are hardly distinguishable, except by size and sex. One cannot of course imagine individualism without evoking the very differences that distinguish individuals from each other. In a later humanist era the depiction of individuals with unique needs, aspirations, values, and habits would inevitably give rise to concepts of personality.

These views would only gain momentum in the second millennium. Slowly people began to appear as free actors, each distinct from everyone else, with unique physical and characterological features. Artisans, a merchant class, educators, poets, and dramatists progressively began to see themselves and others, not as Platonic knock-offs from a universal template, but as differentiated and unreplicable individuals – embodying a personal history, particular tastes, occupational aptitudes, evolving capabilities – displaying and enjoying a panoply of virtues and faults. The extent to which an evolving ethos and economic order generated the concept of individuality (and vice versa) is not yet known. Fully understanding what generated the nascent individualism of the Middle Ages that served as the seedbed for a renewed concept of personality is still an object of study.

Passions and the Individual in Love

People had been regarded in the Dark Ages as copies struck from a broken mold, 'conceived in sin,' 'shapen in iniquity,' and burdened with the legacy of Original Sin. Taking any pleasure in sexual experiences was regarded as intrinsically sinful (but only venially so within the framework of marriage, according to Pope Gregory I (Finn, 2011: 51)). Consistent with this notion, the celibacy of Jesus of Nazareth and the Virgin Mary was lofted to a Christian's preferred state of life. In the wake of the moral tsunami unleashed by Augustine, the cult of the Virgin flourished. That ethos dominated Europe for eight centuries, which pulled in tow the Carolingian dynasty, suffused European societies with a penitential mindfulness, and filled the cultural landscape with magnificent architectural and artistic testaments to the Virgin's virtues and intercessory power. In twelfth-century Europe, the iridescent love affair between a renowned theologian, Peter Abelard who was in holy orders, and his talented paramour, Héloïse Garlande (who in due course became pregnant) erupted in disgrace, genital mutilation, and tragedy. The epic proportions of this flagrant yet beautiful love story attest to the scandal it generated in the hoi polloi, the power élite and crown, and to some extent the clergy.

The Minstrelsy of Medieval Europe

The drama of Peter and Héloïse was a prequel to twelfth century minstrelsy, comprising Provençal troubadours, German-language *Minnesinger*s, French *chansonniers*, and other professional performers who traveled widely and enchanted their audiences with epic as well as mundane accounts of romantic love, resonances of which can be found in Orff's Carmina Burana – musical renditions of racy Medieval poems. Falling in love – what St Augustine deplored as a satanic snare to which he succumbed – was celebrated in the *canzoni* and epic poems that were composed and sung by these minstrels. These were arguably a powerful dynamic in shaping the individualism of that time. Two people who fall passionately in love find each other highly individualized to say the least. Lovers who only have eyes for one another do not regard their partners as fungible love objects.

Personality in Renaissance Literature

As the culture of courtly and romantic musical performances gained popularity and acceptance in Europe, the appreciation of individual differences increased – with lovers singing 'ballads to their mistresses' eyebrow.' The works of Dante, Petrarch, Boccaccio, and Chaucer, among others, depict the subtle as well as salient personality traits of both eminent and ordinary folk. Chaucer's *The Canterbury Tales* display fine characterizations of pilgrims as well as haunting descriptions of their paramours and family members. The richly embroidered personalities of the miller, the prioress, the physician, the merchant, the oft-widowed wife of Bath – are described with humor, subtlety, and fine detail by Chaucer. These are brilliantly illumined characters that enchant us to this day. One can view the intrinsic logic of the society of the day that brazenly overreached the religious constraints of earlier centuries – and then permitted amends of sorts to be made. The cook, for example, describes

> A prentice living in our town
> Worked in the victualing trade, and he was brown,
> Brown as a berry: spruce and short he stood,
> As gallant as a goldfinch in the wood.
> Black were his locks and combed with fetching skill;
> He danced so merrily, with such a will,
> That he was known as Revelling Peterkin.
> He was as full of love, as full of sin
> As hives are full of honey, and as sweet.
> Lucky the wench that Peter chanced to meet.
> At every wedding he would sing and hop,
> And he preferred the tavern to the shop.
>
> p. 324

The tide of humanism had begun to rise with these creative and daring thinkers and is epitomized in the works of William Shakespeare. He held up a mirror to the human throngs that have through the ages attended and studied his plays. Like lay psychologists, we have caught glimpses of ourselves, our strengths, foibles, aspirations, joys, terrors, depressions, and tragic flaws. The 'vasty deep' of the Bard's mind reflected the complexities of the human race in diverse roles and connected us to our most deeply sourced feelings. In his plays he limned the personalities of over 120 characters – among the most notable, Falstaff, Prince Hal, King Lear, Cordelia, Iago, Rosalind, Lady Macbeth, and that crazed depressive, Hamlet – the study of whom continues to reward the personality theorist. This genius not only gave us a language for describing human personality, but also sketched beautifully the crises and stages of a human's life.

Shakespeare is the peerless example of a class of dramatists, portraitists, essayists, novelists, sculptors, and painters that emerged in that humanist era (see Dumont, 2010). Among the greatest were Cervantes, who is numbered among the novelists of the Western canon, and Michel Eyquem de Montaigne (1533–92) an essayist who likely (but unknowingly) spirited some insights into the plays and sonnets of his contemporary, Shakespeare.

Approaches to Contemporary Personology

Stages of Human Development

Personality psychology is founded more on a sound developmental psychology than on any other psychological discipline. Science, as such, is a vast interlocking system in which ideas get sifted from one discipline to another. The attitudes and values that shape them individually have implications for all the others. An example is evolutionary psychology, which reverse engineers human nature, such as we understand it and, abetted by a fossil record, infers the conditions and causes that brought about our current condition, thus suggesting new developmental hypotheses. That there are stages in personality development of a lawful and etic character has always been evident, as has the knowledge that all humans play many parts during the course of their life. A classic literary example of developmental stage theory is Jaques' soliloquy in Shakespeare's *As You Like It*.

> All the world's a stage,
> And all the men and women merely players:
> They have their exits and their entrances;
> And one man in his time plays many parts,
> His acts being seven ages. At first the infant,
> Mewling and puking in the nurse's arms.
> And then the whining school-boy, with his satchel
> And shining morning face, creeping like snail
> Unwillingly to school. And then the lover,
> Sighing like furnace, with a woeful ballad
> Made to his mistress' eyebrow. Then a soldier,
> Full of strange oaths and bearded like the pard,
> Jealous in honor, sudden and quick in quarrel,
> Seeking the bubble reputation
> Even in the cannon's mouth. And then the justice,
> In fair round belly with good capon lined,
> With eyes severe and beard of formal cut,
> Full of wise saws and modern instances;
> And so he plays his part. The sixth age shifts
> Into the lean and slipper'd pantaloon,
> With spectacles on nose and pouch on side,
> His youthful hose, well saved, a world too wide
> For his shrunk shank; and his big manly voice,
> Turning again toward childish treble, pipes
> And whistles in his sound. Last scene of all,
> That ends this strange eventful history,
> In second childishness and mere oblivion,
> Sans teeth, sans eyes, sans taste, sans everything.
>
> Act II, Scene vii.

Three centuries later, the principle of 'multiple personas' was luminously described again by William James (1890/1990: 189–190).

Developmental concepts emerged in abundance in the works of Montaigne, Francis Bacon, Hobbes, Locke, and other observers of the human scene. Newtonian and Leibnizian science shaped the European mind. The loom of history, on which the lives of individuals and their societies are spun like so many endless tapestries, began to depict a materialist and mechanist vision of humans in the seventeenth and eighteenth centuries. People were imagined as (Pierre-Simon) Laplacian puppets driven by causes, noble and ignoble, over which they had no control (cf Menand, 2001: 177–200).

Group as well as individual differences were amply described in the literature of the classicist and Renaissance eras of the West. Divergences of manners, temperaments, and fashionable illnesses were evidenced not only among individuals in cross-sections of societies, but also longitudinally in societies and subcultures over generations. Some patterns remained constant, an example of which is seventeenth-century aggressiveness. Philippe Ariès, for example, noted that in 1588 the king of France sent magistrates, police superintendents, and some sergeants-at-arms to disarm students who during the St Germain fair in Paris "had gone there armed to behave insolently." In 1680 at the Collège de Bourgogne a regulation was enacted: "Neither firearms nor swords are to be retained in pupils' rooms and those who possess such weapons must hand them over to the Principal." In Jesuit colleges students, upon entering, were in return for a receipt relieved of their weapons; their arms were placed in the custody of the Master. 'Jealous in honor, sudden and quick in quarrel' youths and men-at-arms were quick to engage in dueling and other forms of combat. These deadly encounters, endemic to Europe and England (Ariès, 1962: 315–328), eventually gave way to fist fights and wrestling matches.

The manners and mores of Western societies changed gradually, almost imperceptibly, in premodern Europe compared to changes in our rapidly evolving contemporary times. In the post-Cartesian age that ushered in brilliant scientific and technological achievements, Platonic/Cartesian dualism began to suffer continuous attacks from theistic, deistic, and atheistic currents of thought. Intellectual revolutions like Keplerian and Newtonian science spread with alacrity among the cognoscenti. The Enlightenment, spawned by eighteenth-century thinkers, became an intellectual flood tide against which established religions have to this day fought a rearguard action. A cleavage opened between 'believers' and secularists, idealists and empiricists, traditionalists and laboratory-bench methodologists. The scientific *Zeitgeist* seeped into the nooks and crannies of the academies and the markets – by stealth as it were. Urbanized Europeans, simply by 'breathing the common air,' experienced a transmutation of mores that psychologists could not evade. The mathematicization of cognition on which Johann Friedrich Herbart labored in the late-1700s began a process that would adumbrate into the twentieth century, when Allport and Odbert catalyzed the search for underlying personality factors and Raymond B. Cattell helped refine factor analysis, the privileged instrument for the advancement of trait psychology.

Personality Theories and Nineteenth-Century Science

Although personality psychology was not yet regarded in the nineteenth century as a formal discipline (nor was any psychology), there were several streams of scientific investigation that effectively generated models for it. Building on the science of such thinkers as Gottfried Wilhelm Leibniz (1646–1716) and Johann Friedrich Herbart (1776–1841) who did seminal work on human cognition, experimentalists Gustav T. Fechner and Hermann von Helmholtz, among others, conducted research on the neurophysics of human motivation. Helmholtz, indeed, discovered in 1861 the dynamic of unconscious inference. He regarded this "as a kind of instantaneous and unconscious reconstruction of what our past [had] taught us" (Ellenberger, 1970: 313). The implications of this for personality psychology are profound as it indicates how adaptive behaviors in novel situations can intuitively and instantaneously surface.

We fabricate memories to explain our behavior when we cannot otherwise plausibly explain it to ourselves or others. By using posthypnotic suggestion, Jean-Martin Charcot, a renowned neurologist at the Salpétrière (Hospital) in Paris, demonstrated that we tell more than we can know. And Nietzsche, who unmasked many defense mechanisms we use to rationalize our behaviors, wrote that when consciously speaking we produce "a more or less fantastic commentary on an unconscious perhaps unknowable, but felt text" (cited in Ellenberger, 1970: 273) – lying to ourselves even more than we do to each other. The implication is that unconscious schemas, often ancient and inaccessible, direct our current conscious behaviors – and the ideational traffic goes in both directions. Thus, the constituents of our personality reside on both sides of a fuzzy liminal divide.

Arthur Schopenhauer and Natürpsychologie

Schopenhauer (1819) generated one of the dominant motifs of nineteenth-century psychology in *The World as Will and Idea*, a masterpiece of the Western canon. As Henri Ellenberger noted, he presented a view of humans that was pansexual, pessimistic, irrationalist, and determinist. Like Nietzsche whom he influenced, his ideas on the role of tacit cognitions in human behavior set the epistemological stage for psychodynamicists of a later age. His influence was pervasive. Thomas Mann (a Nobel laureate in literature) observed that Schopenhauer's views can be mapped on the Freudian personological schemas that adumbrated into the twentieth century.

Romantics in the Natürpsychologie tradition in which Schopenhauer was embedded placed great emphasis on beauty, virtues of the Volk, nature, homeland, the life of the mind, sentiments of affection, the spiritual, and related values that were vectored by poetry, music, belles-lettres, and a bucolic life style. Goethe and Schiller were venerated pillars of this tradition that served as counterpoise to the rationalism of the Enlightenment *philosophes*. Tension arose of course with the natural scientists, which was finessed by Gustav Fechner, a psychophysicist, experimental estheticist, and philosopher, and other scientists who shared both traditions. Fechner frequently conducted studies of a philosophical character, and Wilhelm Wundt did extensive research in historical, folk

psychology (Völkerpsychologie). The struggle to define personality psychology eventually tilted in the first third of the twentieth century in favor of the psychiatrists and clinician researchers.

Clinician Researchers

The most influential personality theorists in the nineteenth century were medically trained psychotherapists. Their clinical discoveries made during a professional lifetime spent treating, in large part, upper-class female patients as well as others within the context of asylums and mental hospitals, inevitably influenced their modeling of 'normal' human beings. A pathognomonic bias was structured into their theories of personality development and conceptions of the Human. Although they studied with many great 'brain scientists' and neuropathologists such as Ernst Wilhelm von Brücke and Theodor Meynert, or neurologists such as Jean-Martin Charcot, they abandoned the nerve sciences and ensconced themselves in their offices and hospitals to conduct talk and behavioral therapy. The reason, as Emil Kraepelin observed at the turn of the century, was that the meticulous work of the lab-based neuroanatomists had given the world precious little insight into methods for healing mental disorders. Théodore Flournoy, Auguste Forel, Eugen Bleuler, Paul Dubois, Sigmund Freud, Pierre Janet, and Carl Gustav Jung, among many others, predominantly explored the world of psychodynamics. Alfred Adler, a promoter of a community-based, positive psychology (and the ideology of community interest – *Gemeinschaftsgefühl*) was a notable exception (see Henri Ellenberger, 1970, for a peerless historical analysis of nineteenth-century psychiatry, and Shorter, 1997; for a more concise review).

Psychoanalysis

The most salient of the theories of personality that emerged from late-nineteenth-century psychiatry was developed by Sigmund Freud. It began in earnest by the publication with Josef Breuer of *Studies on Hysteria* in 1895 and continued to evolve into Freud's advanced years. The archaic developmental psychology on which Freud built his psychosexual stage theory had become seriously outdated by mid-twentieth century. Radical revisions of his male-centered perspectives were necessitated by an evolving sociopolitical climate and advances in various human sciences. There was a vigorous push-back by liberal, women's movements, keynoted most notably in the works of Karen Horney.

Critical to the Freudian paradigm is the hypothesis that human development is shaped by the drama of son and father in a (the so-called Oedipal) conflict for the affections of the central female figure in their lives. The incestuous overtones in this rivalry, purported to be universal, raised serious anthropological difficulties. That a male child normally has an incestuous appetition for his mother has not been demonstrated. (Freud acknowledged that his theories were not evidence based. Indeed, in recounting the case of 'Little Hans' he stated he had never met a patient who willingly admitted to have ever experienced castration anxiety, a concept integral to the Oedipal conflict.) Indeed the *Westermarck Effect* implies the contrary: that those who are raised in close familial relationship with each other are typically averse to sexual partnering. If the Freudian vision of an Oedipal conflict is true, it raises a serious issue for females. As the development of a moral sense depends on the resolution of the Oedipal conflict in a univocal sense, the development of a 'superego' is compromised among women. This is nettlesome to those who, like Carl Jung, take it as axiomatic that women's moral development is at least on a par with men's. Jung attempted to defuse the problem for Freudian psychodynamicists by inventing the Electra complex – a ploy that Freud always rejected. Following World War I, Freud added the earlier Adlerian aggressive drive to the sexual. He also formally supplemented his psychosexual stage theory with a topographical model positing the hypothetical constructs, id, ego, and superego.

This psychiatric phase in an evolving personality psychology was redirected by the science of developmental psychology and other related social and behavioral sciences.

From Illness Models to Wellness Models of Human Nature

The pathognomonic theories of personality that proliferated in nineteenth-century psychiatry began to yield to powerful currents of positive psychology following World War I. The modern origins of wellness models of human nature are arguably anchored in the work of Wertheimer, Koffka, and Köhler, a team that developed Gestalt psychology – and provided a rich perspective on the individuality of each person. An influential book *Holism and Evolution*, written by Jan Christiaan Smuts, a South African naturalist, soldier, politician, and philosopher, appeared in 1926 and enriched this perspective. Positive psychology was given further impetus by Alfred Adler's educational reforms and his principle that mental health depended in large part on living in sanitized communities. A former member of Freud's Wednesday Psychoanalytic Society from which he was dismissed, Adler is often mistakenly thought by the popular press to have been a Freudian. Teleologist and contrarian that he was, he averred that probing a patient's personal history was less important than planning his or her future. He rejected the pessimism, pansexualism, and determinism of Freud, affirming that humans have only one drive: self-actualization.

The groundswell of positive psychology that was launched by these pioneers and advanced by Henry Murray, B.F. Skinner (an eminent behaviorist), Gordon Allport (among the first factor analysts), Abraham Maslow, Carl Rogers, and many others, is now being furthered by such notables as Martin Seligman and Mihaly Csikszentmihalyi.

Select Positive Developments

Personality psychology is so complex that it is not possible to examine all aspects in a single text or handbook, impressive as the depth of coverage may be. Health enhancing self-grooming, playfulness, attitudes to food and cooking, altruism, curiosity, sense of humor, dignity, empathy, ticklishness, creativity, self-transcendence, fairness, and intercorrelations of many such facets will long be research domains for personologists. Many

are still given short shrift in graduate-level textbooks. This complexity is multiplied when one stirs in multicultural variables (e.g., see Marsella et al., 2000). Over the past century several others of a life span and gerontological character have been klieg-lighted.

Reduction of Pre- and Postnatal Stressors

Innumerable prenatal stressors (not least, say, maternal adrenocortical stress hormones and other fetus-toxic substances) and early childhood traumatic experiences have now been shown to seriously affect for the worse the normal development of personality (as Mary Ainsworth and John Bowlby, among many others, during the past 75 years have shown). The nurturant bond of an infant with its primary caregivers is ruptured by long hospitalization, divorce, parental death, or other traumatizing experiences. As with severe stimulus deprivation (both social and material) those events can result in enduring harm. Though complex and multidimensional, later positive experiences and healthful environments can mitigate their effects as Michael Rutter, Jerome Kagan, Robert Plomin, and other recent developmentalists have demonstrated. As Arnold Gesell noted, reducing childhood stress by scheduling educational experiences to fit actual rather than anticipated levels of neurological development is of cardinal importance.

Epigenetics

The strength and duration of environmental variables and their interaction with one's genome produces, for better or for worse, far-reaching effects that are possibly transgenerational. Much of the plasticity in our neuroemotional systems depends on epigenetic effects, that is, the expression of dormant genes (which can be positive or negative) triggered by significant environmental events. Positive experiences can do much to repair developmental damage and effect healthful neuronal alterations. Indeed, Nobelist Eric Kandel intimates that personality change through psychotherapy is achieved through gene expression at the neuronal level. Such findings strengthen the foundations for furthering positive psychology, which could hardly have been imagined by the grand-system builders of the late-nineteenth century.

Neoteny

In the context of positive personality development, the principle of neoteny comes into focus. Humans have a comparatively lengthy childhood. Neoteny affirms that the evolutionary prolongation of juvenile stages into adulthood has favored increased neural plasticity, rich behavioral opportunities permitting optimal life-altering choices, and the partial if not total reversal of the effects of destructive childhood experiences. Neoteny allows for the remediation of serious mistakes and provides affordances of redemption for adolescents or their caregivers. The developmental advantage of this is that as children become neurologically more mature, they can better compute the risks and the unforeseen consequences to themselves and others that their immature behaviors can generate. And they can avail themselves of future benefits in behavioral repertoires that are more prosocial – as well as career options that are more beneficial for them. A collateral benefit arises in that the neurological maturity of malefactors and their level of intelligence (fluid as well as crystallized), have a bearing on legal judgments of their guilt. This is a significant educational, developmental, and personological issue.

Overturning Homuncularism

Homuncularism is a multifaceted, now outdated construct positing that an infant is an integrally preformed human being whose infantile impulses and reaction potentials are the same as an adult's. The infant male is thus regarded as a 'little man' who can enter into the same dramatic sexual struggles – incestuous, hetero/homosexual, polygynous (and variants of these), which his father and other adults engage in. That allowed personologists to hypothesize theories of personality development that range from the most rigidly preformationist to the modestly predeterminist. Ausubel et al. (1980) affirmed that a modern example of homuncularism is the classic psychoanalytic notion "that infantile and adult sexuality are qualitatively equivalent" (15).

A mitigated variant of preformationism is predeterminism. Arnold Gesell, for example, a renowned developmentalist posited that we are biologically programmed, such that our temperamental traits as well as more specific manners are instinctive. Studies of twins separated at birth have given some support to a current genomic view. In a greatly attenuated form, the predeterminist position is mainstream science today. Though environmentally accommodative, this view accounts for as much as 60% or more of trait variance in the Eysenck, Five-Factor, or Cattellian models of personality types. On the other hand, the advice of neuroscientist Joseph LeDoux (2002) relative to the influence of the human genome on our personality seems apt. He acknowledges that our human nature is relatively constant, but he notes that we are not born preassembled. In our individual lifetimes we can only be 'glued together by life.'

Summary

The history of psychological sciences does not have a beginning, middle, and end, or an Aristotelian *telos* that drives it – unless one transmutes it into a theology. This is particularly true of the science of personality. Michael Mahoney (1991) similarly averred that history cannot be construed as proceeding toward a particular end-state. Using a fiction writer's techniques to narrativize the sweep of human history risks imposing cognitive schemas and plot lines on a panorama that has no logic or terminal climacteric, except in the historian's mind. One cannot in short, impose a teleological, still less gnostic, directionality that suggests there is a covert plot or story line to be uncovered.

In the light of current science, it appears that theories of personality will continue to evolve in directions that are genetics-based yet less instinct-driven, life-span developmental, holistic and sociobiology based, gender-equal yet gender-differentiated, cross-culturally conditioned, and Heraclitean, that is, fashioned by the changing societies and multiple cultures in which one is bred and educated. Clearly, personality

theories that predate the Second World War are not, by and large, consistent with these descriptors. It appears that the era of the grand normative systems is past.

For readers, finally, who want a more detailed and comprehensive analysis of personality psychology and its evolution over time, I suggest they begin with Corr and Matthews (2009), John et al. (2008), and Dumont (2010). These will also lead them to many other valuable sources. To follow the development of this discipline, readers are urged to regularly consult the Annual Review of Psychology, Journal of Personality and Social Psychology, European Journal of Personality, Developmental Psychology, and other respected periodicals that publish articles in this domain.

See also: Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis; Factor Analysis and Latent Variable Models in Personality Psychology; Personality Assessment; Personality Development: Systems Theories; Personality, Trait Models of; Social Psychological Theory, History of; Wisdom, Psychology of.

Bibliography

Ariès, P., 1962. Centuries of Childhood: A Social History of Family Life. Vintage Books, New York.

Augustine, 4th c.C.E./1952. The confessions. In: Great Books, vol. 16. Encyclopædia Britannica, Chicago.

Ausubel, D.P., Sullivan, E.V., Ives, S.W., 1980. The Theory and Problems of Child Development, third ed. Grune & Stratton, New York.

Brown, D.E., 1991. Human Universals. McGraw-Hill, New York.

Chaucer, G., 1386/1990. The Canterbury tales. In: Encyclopædia Britannica. Great Books of the Western World, vol. 19. University of Chicago Press, Chicago, pp. 275–476.

Corr, P.J., Matthews, G. (Eds.), 2009. The Cambridge Handbook of Personality Psychology. Cambridge University Press, Cambridge, UK.

Dumont, F., 2010. A History of Personality Psychology: Theory, Science, and Research from Hellenism to the 21st Century. Cambridge University Press, Cambridge, UK.

Ellenberger, H.F., 1970. The Discovery of the Unconscious: The History and Evolution of Dynamic Psychiatry. Basic Books, New York.

Finn, T., 2011. Sex and marriage in the sentences of Peter Lombard. Theological Studies 72 (1), 41–69.

Hippocrates, c.5th c.BCE/1990. Hippocratic writings. In: Great Books, vol. 9. Encyclopædia Britannica, Chicago.

John, O.P., Robins, R.W., Pervin, L.A. (Eds.), 2008. Handbook of Personality: Theory and Research, third ed. Guilford Press, New York.

Jonas, H., 1958. The Gnostic Religion. The Beacon Press, Boston.

Koestler, A., 1989. The Sleepwalkers: A History of Man's Changing Vision of the Universe. Arkana Penguin Books, London.

LeDoux, J., 2002. Synaptic Self: How Our Brains Become Who We Are. Viking/Penguin, New York.

Löwith, K., 1949. Meaning in History. University of Chicago Press, Chicago.

Mahoney, M., 1991. Human Change Processes: The Scientific Foundations of Psychotherapy. Basic Books, New York.

Marsella, A., Dubanoski, J., Hamada, W.C., Morse, H., 2000. The measurement of personality across cultures: historical, conceptual, and methodological features and considerations. American Behavioral Scientist 44, 41–62.

Menand, L., 2001. The Metaphysical Club: A Story of Ideas in America. Farrar, Straus and Giroux, New York.

Moghaddam, F.M., 1998. Social Psychology: Exploring Universals across Cultures. W.H. Freeman, New York.

Pinker, S., 2002. The Blank Slate: The Modern Denial of Human Nature. Penguin, New York.

Plato, 370 BCE/1991; Book VII, Section 808. Laws. In: Adler, M.J. (Ed.), Plato, Great Books of the Western World. Encyclopædia Britannica, Chicago, pp. 640–799.

Plato, 4th c.BCE/1952; Book II, Section 370. The Republic. In: Adler, M.J. (Ed.), Plato, Great Books of the Western World. Encyclopædia Britannica, Chicago, pp. 295–441.

Plato, 4th c.BCE/1952. Timaeus. In: Adler, M.J. (Ed.), Plato, Great Books of the Western World, Great Books, vol. 6. Encyclopædia Britannica, Chicago, pp. 295–441, pp. 442–477.

Robinson, J.M., 1977. Introduction. In: Meyer, M. (Ed.), The Nag Hammadi Library in English. Harper and Row, San Francisco, pp. 1–25.

Schopenhauer, A., 1819/1969. The World as Will and Representation (Trans. E. F. J. Payne). General Publishing, Toronto.

Shorter, E., 1997. A History of Psychiatry from the Era of the Asylum to the Age of Prozac. John Wiley, New York.

Theophrastus, c.4th c.BCE/1929. The Characters. Heinemann, London (J.M. Edmonds, Trans.).

Thomas, L., 1979. The Medusa and the Snail: More Notes of a Biology Watcher. Viking Penguin, New York.

Personality Theory and Psychopathology

Kate E Walton and Stephanie R Pavlos, St. John's University, Jamaica, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The connection between personality and psychopathology has been studied for centuries. In the current article, we provide an overview of contemporary research in the field. We begin with a review of trait models of personality and the current psychopathology classification system in use. We discuss the link between normal personality traits and personality disorders and other types of psychopathology and conclude with a discussion of different theoretical perspectives explaining the personality–psychopathology connection.

The link between personality and psychopathology has been long-recognized, dating back to ancient Greece and Hippocrates' discussions of the four humors. The four essential fluids of the body – phlegm, blood, bile, and black bile – were thought to determine temperament. Depending on the dominant humor, one could be phlegmatic, sanguine, choleric, or melancholic, and each type had an accompanying set of attributes. An imbalance in these humors led to symptoms of illness, and therefore temperament was thought to be connected to all disease, physical or mental. Later, in the nineteenth century, Darwinian-influenced perspectives on the personality–psychopathology link arose. These were evolutionary perspectives, depicting mental illness as a genetically based character deficiency. Stemming from this was Freud's influential theory of the stages of character development, from the primitive to the more complex, and associated psychopathology (Maher and Maher, 1994).

Contemporary psychologists and psychiatrists continue to explore this association, profiting from reliable and valid taxonomies of both personality and psychopathology. Trait models are contemporary models of personality that have gained consensus. These models emphasize traits as dimensions, covering the range from normal to disordered personality (vs a categorical distinction between the two). We begin with an overview of these trait models of personality. Next, we review the status of personality traits and disorders in the most current nosology put forth by the American Psychiatric Association, the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5; American Psychiatric Association, 2013). We note that the World Health Organization provides a classification system separate from the American Psychiatric Association's DSM, the International Classification of Diseases, which is currently being revised. Preliminary proposed changes to the eleventh edition, due for release in 2015, have not been made public; therefore, we limit our focus to the DSM. After reviewing personality traits and disorders in the DSM, we then review literature linking personality traits and other mental disorders. Finally, we conclude with a discussion of different theoretical perspectives explaining the personality–psychopathology connection.

Personality Trait Models

Allport offered a coherent definition of personality, writing, "Personality is the dynamic organization within the individual of those psychophysical systems that determine his characteristic behavior and thought" (1961: 28). Allport went on to elaborate on the key features of this definition, including the emphasis on individual differences. At the root of these individual differences are traits, dimensional tendencies to show consistent patterns of thoughts, feelings, and actions across time and situations. Further, personality traits are thought to be hierarchical, with broad traits subsuming more specific subfacets. A long-standing goal of personality psychologists has been to develop a taxonomy to characterize these broad trait dimensions. Although there is recognition that these dimensional traits cover a broad spectrum from moderate (i.e., normal) to extreme (i.e., abnormal) personality, some models focus more on the normal range, while others focus on the abnormal range of personality variation.

Models of Normal-Range Personality

Though many have fallen out of favor, multiple normal personality trait taxonomies have been put forth with the number of key personality traits ranging from two (Block and Block, 1980) to as many as twenty (Gough, 1987). Today two models, the Big Three and the Big Five, receive the most attention in the literature. The three-factor model varies somewhat across researchers and measures in terms of trait labels, however the essence of the constructs is the same. The models include positive temperament, negative temperament, and constraint (vs disinhibition; Tellegen and Waller, 2008; Watson and Clark, 1993) or parallel constructs such as extraversion, neuroticism, and psychoticism (Eysenck and Eysenck, 1976).

The Big Five, born out of the lexical tradition, represents an expanded taxonomy. Beginning with Allport's efforts (Allport and Odbert, 1936) and continuing for a quarter of a century (Tupes and Christal, 1992), trait terms were amassed from the English language dictionary and were subsequently factor analyzed. Across different samples, five factors emerged. These include extraversion (positive affectivity), agreeableness (vs antagonism), conscientiousness (constraint), emotional stability (vs neuroticism or negative affectivity), and openness to experience (or intellect and unconventionality; Lynam and Widiger, 2001). Tremendous empirical support for the Big Five has surfaced in the past few decades, as the same five-factor structure is found in numerous populations and settings (Costa and McCrae, 1988, 1994; De Raad et al., 1998;

Halverson et al., 1994; McCrae and Costa, 1990, 2003; Van de Vijver and Valchev, in press).

Another strength of the Big Five model is that it captures commonalities among most existing trait taxonomies (for an overview of how the Big Five maps onto other popular personality trait models, see John et al., 2008: 115). This hierarchical nature of personality traits allows us to reconcile seemingly distinct trait taxonomies, such as the Big Three and Big Five. For example, positive emotionality/temperament from the three-factor models is akin to extraversion, and negative emotionality/temperament is akin to neuroticism, while constraint (vs disinhibition) subsumes agreeableness and conscientiousness (openness does not appear in the three-factor models reviewed here). The three-factor models emphasize what one might call superfactors, while the Big Five is an extension of these broad domains (Markon et al., 2005). Though some have criticized the Big Five model for diminishing the complex schema of personality to only five traits, the purpose of the model is to "provide a scientifically compelling framework in which to organize the myriad individual differences that characterize humankind" (Goldberg, 1993: 27).

We can garner a more nuanced view of one's personality by considering the lower-order facets of the broad traits (Hong and Paunonen, in press). According to Costa and McCrae's (1995) model of the Big Five, there are six facets within each of the five domains. For example, they suggest that the domain of neuroticism consists of the more specific facets of anxiety, angry hostility, depression, self-consciousness, impulsiveness, and hostility. In this model, conscientiousness consists of competence, order, dutifulness, achievement striving, self-discipline, and deliberation (Costa and McCrae, 1995). Different models may identify a different lower-order structure. In particular, much work has been carried out on the lower-order structure of conscientiousness. In one study integrating lexical and questionnaire studies, five facets of conscientiousness emerged including industriousness, orderliness, impulse control, reliability, and conventionality (Roberts et al., 2005). Developing a consensus of the lower-order structure of each of the Big Five is critical because facets have better predictive validity than broad traits (Paunonen and Ashton, 2001), they show different developmental patterns (Jackson et al., 2009) and they can facilitate a better understanding of the client in clinical practice.

Models of Disordered Personality

Dimensional models of disordered personality have also been put forth. Two widely used measures of disordered personality traits are the *Schedule for Nonadaptive and Adaptive Personality* (Clark, 1993) and the Dimensional Assessment of Personality Pathology (Livesley and Jackson, 2009), consisting of 12 and 18 scales, respectively. Findings suggest these scales make up four broad, higher-order traits, which are similar, yet not identical, to those found in models of normal personality. Emotional dysregulation, dissocial behavior, inhibition, and compulsivity resemble neuroticism, (dis)agreeableness, (lack of) extraversion, and conscientiousness, respectively (Livesley et al., 1998; O'Connor and Dyce, 1998; Widiger, 1998). There is a parallel to each of the Big Five except openness. In addition to these four pathological traits, it has been argued that a distinct fifth factor is necessary to account for the variance in pathological personality functioning. This fifth trait has been labeled oddity or peculiarity (Watson et al., 2008). Some have posited that models of normal and abnormal personality should be integrated into a single hierarchy (Markon et al., 2005). For example, Watson et al. (2008) proposed a Big Six model that includes both normal and abnormal traits, including the Big Five and oddity.

Personality Traits and Disorders in the DSM

As discussed, these models highlight the idea that traits are dimensional, with abnormal personality representing extreme variants of normal-range traits. This counters the long-held categorical conceptualization of abnormal personality in the DSM. This perspective regards personality disorders (PDs) as "qualitatively distinct clinical syndromes" (American Psychiatric Association, 2013: 646). Ten specific PDs are defined and these are split into three clusters based on descriptive similarities. Cluster A includes paranoid, schizoid, and schizotypal PDs; Cluster B includes antisocial, borderline, histrionic, and narcissistic PDs; and Cluster C includes avoidant, dependent, and obsessive-compulsive PDs. A specific number of criteria must be met in order for a PD to be diagnosed. For example, antisocial PD is diagnosed if a patient meets at least three of seven criteria, including failure to conform to social norms with respect to lawful behaviors, deceitfulness, impulsivity, irritability and aggressiveness, reckless disregard for safety, consistent irresponsibility, and lack of remorse.

Each of these 10 PDs has been linked with normal personality traits (Crego and Widiger, in press). Saulsman and Page (2004) offered a meta-analysis documenting the relationship between each PD and the Big Five. Some similarities among PDs were uncovered (e.g., high neuroticism and low agreeableness are characteristic of most PDs) yet they also concluded that most PDs have a distinct personality trait profile. For example, antisocial PD, characterized by a pervasive pattern of disregard for the rights of others, is marked by low agreeableness and conscientiousness. Extraversion also shows strong relationships with most PDs, though the direction of that relationship varies across PDs. Extraversion is positively related to histrionic and narcissistic PDs, as individuals with histrionic PD crave attention and are often flirtatious and dramatic and those with narcissistic PD often demonstrate grandiosity in their beliefs and behavior. Extraversion is negatively related to schizoid, schizotypal, and avoidant PDs, on the other hand, as individuals with these PDs demonstrate a lack of interest in, or anxiety regarding, interpersonal contact and relationships. Conscientiousness shows some relations with PDs, though this finding is not as robust as for neuroticism, agreeableness, and extraversion, and openness shows no strong relationship with any PD. Samuel and Widiger (2008) extended this meta-analysis by

concentrating on the specific facets of the broad traits and largely replicated the findings of Saulsman and Page.

Numerous problems with the categorical perspective have been cited repeatedly in the literature (e.g., Widiger, 2003). These include a lack of qualitative distinction between normal and abnormal personality, heterogeneity within diagnostic categories, and excessive comorbidity. Critics of the categorical perspective argue that normal-range personality traits and those specific to PDs overlap considerably and share a parallel structure in genetic and environmental components (Livesley et al., 1998). They also contend that the categorical system's practice of using a subset of items to meet diagnostic threshold often results in incongruence between individuals who share the same diagnosis (Trull and Durrett, 2005). Two individuals meeting the diagnosis for antisocial PD, for example, may share all or no symptoms. Proponents of the dimensional system argue that the categorical system's difficulty with excessive comorbidity between PD diagnoses could be reproduced and better explained with a dimensional model. Clinicians would be better able to conceptualize their patients without the constraints of the DSM-defined PDs, and researchers would not need to unnecessarily exclude a majority of patients targeted to a study due to comorbidity (Krueger, 2013).

Given these criticisms and mounting support for integrative hierarchical models of personality dimensions, the authors of DSM-5 spent a great deal of time debating how to best conceptualize PDs and reflect that conceptualization in the diagnostic criteria. Many proposed replacing the categorical system with a dimensional model, but the authors were not able to arrive at a consensus approach on how best to do this (Krueger, 2013). Instead, DSM-5 offers an additional dimensional perspective of PDs in addition to the preexisting categorical one from DSM-IV-TR (American Psychiatric Association, 2000). Inclusion of both models was done to maintain current clinical practice while offering an alternative that addresses the numerous problems with the prior diagnostic system (American Psychiatric Association, 2013: 761). Section II contains the same diagnostic criteria for 10 specific PDs from DSM-IV-TR, while Section III contains the alternative dimensional model. In this model, "personality disorders are characterized by impairments in personality *functioning* and pathological personality *traits*" (p. 761; emphasis in original). While there are a few similarities across Sections II and III, such as criteria including stability and pervasiveness across personal and social situations, two key changes are emphasized in Section III. First is the evaluation of disturbances in self and interpersonal functioning on a five-point continuum ranging from no to extreme impairment. The second key change is the use of 25 specific pathological personality trait facets organized into five broad domains including negative affectivity, detachment, antagonism, disinhibition, and psychoticism. These five domains align, respectively, with the opposite poles of emotional stability, extraversion, agreeableness, and conscientiousness from the Big Five model, as well as the oddity/peculiarity trait from models of abnormal personality. These can be assessed with the Personality Inventory for DSM-5 (Krueger et al., 2013) with self- or informant-reports. Section III recognizes 6 of the 10 specific Section II disorders, including antisocial, avoidant, borderline, narcissistic, obsessive-compulsive, and schizotypal PDs. These six are differentiated according to the manner in which their disturbances in self and interpersonal functioning are manifest, as well as their defining pathological personality traits. That is, each disorder has a unique personality trait profile. Antisocial PD, for example (in addition to impairment in personality functioning), is characterized by six out of seven pathological personality traits, including manipulativeness, callousness, deceitfulness, and hostility (all facets of antagonism), in addition to risk taking, impulsivity, and irresponsibility (all facets of disinhibition). Section III offers the first step toward adopting a complete dimensional perspective of PDs.

Personality and Nonpersonality Disorder Psychopathology

Although personality traits connect most obviously with PDs, there is a plethora of research indicating a link between personality and other types of psychopathology as well. Before delving into a review of this research, it is important to first discuss the organization of (non-PD) mental disorders in the DSM-5, which will provide a framework to guide the discussion on the personality–psychopathology link. Unlike DSM-IV-TR, DSM-5 recognizes spectra of disorders that share common neural substrates, genetic and environmental risk factors, biomarkers, and temperamental antecedents, among other things (Andrews et al., 2009). Disorders within each spectra are situated in adjacent chapters. The spectra are also positioned to reflect developmental course, with disorders beginning in childhood (i.e., neurodevelopmental disorders, the schizophrenia spectrum, and other psychotic disorders) positioned first, and neurocognitive disorders, which have an onset of adulthood, falling at the end. Two clusters, internalizing and externalizing disorders, with adolescent or early adult onset lie in between. Internalizing disorders are those characterized by anxiety, depressive, and somatic symptoms and externalizing disorders are those characterized by impulsive, disruptive conduct, substance use, and other addictive symptoms (American Psychiatric Association, 2013: 13). The ensuing discussion involves the connection between personality traits and internalizing and externalizing disorders.

There are a number of published studies documenting the relationship between specific mental disorders and personality traits in specific clinical samples. Given space limitations and our objective of presenting a comprehensive overview of the personality–psychopathology connection, we limit our review to meta-analyses and epidemiological studies of multiple disorders. Trull and Sher (1994) provided a first look at the connection between the Big Five and multiple (non-PD) disorders in a nonclinical sample. Individuals were assessed for multiple lifetime diagnoses including substance use disorders, several anxiety disorders, and major depression. In general, neuroticism, extraversion, and conscientiousness were the traits most highly related to the various disorders. Most substance use disorders were characterized by high neuroticism and openness, and low extraversion,

agreeableness, and conscientiousness. Anxiety disorders showed a similar personality profile to substance use disorders. Any lifetime anxiety disorder showed elevated neuroticism and openness, and low extraversion, agreeableness, and conscientiousness. More specific anxiety disorders showed a similar pattern, though in some cases, the effects were not significant (i.e., the diagnosed group was not significantly different from the nondiagnosed group). The one exception was simple phobia where no differences were observed between the diagnosed and the nondiagnosed group. Major depression showed the same personality profile again; the diagnosed group had elevated levels of neuroticism and openness and lower levels of extraversion and conscientiousness. They were lower on agreeableness, though this difference was not statistically significant. In addition to documenting the personality profiles of diagnostic groups relative to those with no diagnoses, Trull and Sher confirmed that the Big Five provides clinically useful information by showing that Big Five scores accounted for a significant amount of variance in all diagnoses except simple phobia beyond that accounted for by a measure of psychopathology.

Krueger and colleagues extended the work of Trull and Sher (1994) with a longitudinal study of a representative birth cohort (Krueger et al., 1996). Age 18 Multidimensional Personality Questionnaire (MPQ; Tellegen and Waller, 2008) personality profiles of members of four diagnostic groups (affective disorder, anxiety disorder, substance dependence disorder, and conduct disorder) were compared with a control group, individuals with no diagnoses. The disorders were assessed at ages 15, 18, and 21. At the broad-factor level, compared with the control group, the affective disorder, substance dependence, and conduct disorder groups at age 18 scored low on positive emotionality and constraint, and high on negative emotionality (there was a greater distinction among the diagnostic groups at the level of the more specific scales). The anxiety disorder group scored low on positive emotionality and high on negative emotionality. These correlations between the age 18 personality traits exhibited a consistent pattern with age 15, 18, and 21 mental disorders. Using the same longitudinal data set, Krueger (1999) followed up by examining whether the age 18 personality traits predicted age 21 diagnostic membership above and beyond age 18 diagnostic membership. Greater negative emotionality at age 18 increased the odds of being diagnosed with affective disorder at age 21. Greater negative emotionality and constraint increased the odds of an anxiety disorder diagnosis. Increased negative emotionality and decreased constraint increased the odds of substance dependence disorder as well as antisocial PD.

Krueger and colleagues also examined the relationship between mental disorders and personality traits, again as measured by the MPQ, in a large population-based study (Krueger et al., 2001). In men and women, negative emotionality was positively correlated with an internalizing factor (including major depressive episode, panic disorder, simple phobia, and social phobia), and constraint was negatively correlated with an externalizing factor (including adult antisocial behavior, conduct disorder, alcohol dependence, and drug dependence). Positive emotionality had a negative correlation with the internalizing factor for women only.

Kotov and colleagues carried out a comprehensive meta-analysis of studies on the association between personality and internalizing and externalizing disorders, providing effect sizes to indicate the strength of relationship between them (Kotov et al., 2010). Specifically, they considered six broad personality traits, including extraversion/positive emotionality, agreeableness, conscientiousness, disinhibition, neuroticism/negative emotionality, and openness. Neuroticism showed the strongest association with mental disorders, with some effect sizes exceeding 2.0, followed by low conscientiousness. Low extraversion was fairly consistent across disorders, though the magnitude of the effect was smaller than for neuroticism and conscientiousness. Disinhibition better distinguished among disorders; the strongest and most consistent association was with substance use disorders. Low agreeableness was also related to substance use disorders, though to a lesser degree than anticipated. Openness showed no strong pattern of association with mental disorders. Further, Kotov and colleagues examined the personality profiles of clusters of disorders, examining externalizing (substance use disorders) and two clusters of internalizing disorders, distress (depressive disorders) and fear (anxiety disorders). The distress and fear profiles were markedly similar, though the distress cluster scored slightly higher on all traits other than openness. The substance use disorder cluster exhibited a distinct profile with elevated levels of disinhibition and extraversion and lower levels of neuroticism and agreeableness.

In sum, high neuroticism (negative emotionality) and low conscientiousness and extraversion (positive emotionality) are overwhelmingly associated with most mental disorders. Externalizing disorders are typically set apart from internalizing disorders by increased disinhibition (low conscientiousness coupled with low agreeableness). These findings are robust, and therefore the importance of personality traits in mental disorders cannot be ignored.

Explaining the Personality–Psychopathology Link

As evidenced by the collection of studies reviewed here, personality traits are closely linked with multiple types of mental disorders. In this next section, we consider six explanations for this relationship. Four models (vulnerability, pathoplasty, scar, and complication) consider personality and psychopathology as distinct constructs, while two models (common cause and spectrum) treat them as having common underpinnings. We use the well-established connection between low conscientiousness (and its facets of industriousness, impulse control, orderliness, reliability, and conventionality; Roberts et al., 2005) and increased substance use to illustrate each (Kotov et al., 2010; Walton and Roberts, 2004).

The vulnerability model is a causal model, suggesting that premorbid personality traits can act as risk factors, predisposing an individual to mental disorder. For example, preexisting low conscientiousness puts one at risk for developing substance use disorders. Those low on impulse control might be more likely to engage in binge drinking. The pathoplastic relationship hypothesis suggests that personality traits influence mental

disorder once the disorder is already developed. That is, the course of illness, its expression over time, and its response to treatment are susceptible to the influence of personality traits. For example, substance abuse problems might continue to worsen in individuals with low conscientiousness.

While the vulnerability and pathoplasty models consider the influence of personality on mental disorders, the scar and complication models consider the influence of mental disorders on personality. Like the vulnerability model, the scar model is causal, positing that mental disorders have a direct effect on personality traits. For example, those engaging in excessive substance use might exhibit decreased levels of key characteristics of conscientiousness, such as industriousness and reliability, as a result of their misuse. That is, they might be less like to complete tasks, follow through with responsibilities, etc. The complication model follows from the scar model, but the effect of the disorder on personality traits is hypothesized to be temporary, lasting only the duration of the disorder. For example, individuals abusing alcohol or drugs would be much more likely to lack impulse control when their disorder expression is heightened, but their impulse control levels would become more normative once the disorder was treated.

Finally, the common cause and spectrum models suggest that personality traits and psychopathology are not distinct constructs but instead share a common underpinning. The common cause model posits that personality and psychopathology share a single basis. A common root, such as a common genetic risk factor, would explain the overlap in conscientiousness and substance use (Vrieze et al., 2013). According to the spectrum model, personality and psychopathology are different manifestations of a common process. Some have argued that conscientiousness and substance use are part of an externalizing spectrum (Eaton et al., 2010).

It has been argued that these six hypotheses are not mutually exclusive and that more than one may appropriately describe the personality–psychopathology relationship (Andersen and Bienvenu, 2011; South et al., 2010). The cause of the relationship could vary from person to person or from disorder to disorder. While there is some evidence supporting each model, further longitudinal data are necessary to distinguish among them. For example, we cannot appreciate the interplay between conscientiousness and substance use without tracking the lifetime course of one in light of the developmental course of the other. Nevertheless, we argue that the common cause and spectrum models are in line with the research reviewed in the sections above, demonstrating that there is no distinction between normal and abnormal functioning but rather they fall along a common continuum with abnormal functioning representing extreme expressions of normal personality traits (Widiger, 2011).

Conclusion

There is a large body of research describing the associations between personality and psychopathology, which has been facilitated by a growing consensus regarding personality and psychopathology taxonomies. Understanding these associations is important for several reasons. Research on the personality–psychopathology link has helped us understand comorbidity, and we have seen changes to classification systems as the result of this understanding. For example, in the DSM-5, we no longer see a divide between Axis I and Axis II disorders, there is greater appreciation for developmental continuity, and there is recognition of spectra of disorders. Continued research will not only help us refine our classification systems, but will help us gain insight into the etiology of psychiatric conditions and improve prevention and intervention efforts.

See also: Five Factor Model of Personality, Personality Disorder; Personality Disorders; Personality, Trait Models of.

Bibliography

Allport, G.W., 1961. Pattern and Growth in Personality. Holt, Reinhart and Winston, Oxford, England.

Allport, G.W., Odbert, H.S., 1936. Trait-names: a psycho-lexical study. Psychological Monographs 47 (1), i-171.

American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders, fourth ed. text revision. American Psychiatric Publishing, Arlington, VA.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Publishing, Arlington, VA.

Andersen, A.M., Bienvenu, O.J., 2011. Personality and psychopathology. International Review of Psychiatry 23 (2), 234–247.

Andrews, G., Goldberg, D.P., Krueger, R.F., Carpenter, W.T.J., Hyman, S.E., Sachdev, P., Pine, D.S., 2009. Exploring the feasibility of a meta-structure for DSM-V and ICD-11: could it improve utility and validity? Psychological Medicine 39 (12), 1993–2000.

Block, J.H., Block, J., 1980. The role of ego-control and ego resiliency in the organization of behavior. In: Collins, W.A. (Ed.), Minnesota Symposia on Child Psychology. Erlbaum, Hillsdale, NJ, pp. 39–101.

Clark, L.A., 1993. The Schedule for Nonadaptive and Adaptive Personality (SNAP). University of Minnesota Press, Minneapolis, MN.

Costa Jr., P.T., McCrae, R.R., 1988. Personality in adulthood: a six-year longitudinal study of self-reports and spouse ratings on the NEO personality inventory. Journal of Personality and Social Psychology 54 (5), 853–863.

Costa Jr., P.T., McCrae, R.R., 1994. Set like plaster? Evidence for the stability of adult personality. In: Heatherton, T., Weinberger, J.L. (Eds.), Can Personality Change? American Psychological Association, Washington, DC, pp. 21–40.

Costa, P., McCrae, R., 1995. Domains and facets: hierarchical personality assessment using the revised NEO personality inventory. Journal of Personality Assessment 64 (1), 21–50.

Crego, C., Widiger, T. Five factor model of personality, personality disorder, in press.

De Raad, B., Perugini, M., Hrebickova, M., Szarota, P., 1998. Lingua franca of personality. Taxonomies and structures based on the psycholexical approach. Journal of Cross-Cultural Psychology 29 (1), 212–232.

Eaton, N.R., South, S.C., Krueger, R.F., 2010. The meaning of comorbidity among common mental disorders. In: Millon, T., Krueger, R.F., Simonsen, E. (Eds.), Contemporary Directions in Psychopathology: Scientific Foundations of the DSM-V and ICD-11. Guilford Press, New York, pp. 223–241.

Eysenck, S.G., White, O., Eysenck, H.J., 1976. Personality and mental illness. Psychological Reports 39 (3), 1011–1022.

Goldberg, L.R., 1993. The structure of phenotypic personality traits. American Psychologist 48 (1), 26–34.

Gough, H.G., 1987. California Psychological Inventory Administrator's Guide. Consulting Psychologists Press, Palo Alto, CA.

Hong, R., Paunonen, S. Five factor model of personality, facets of, in press.

Halverson, C.F., Kohnstamm, G.A., 1994. In: Martin, R.P. (Ed.), The developing structure of temperament and personality from infancy to childhood. Lawrence Erlbaum Associates, Publishers, NJ.

Jackson, J.J., Walton, K.E., Harms, P.D., Bogg, T., Wood, D., Lodi-Smith, J., Roberts, B.W., 2009. Not all conscientiousness scales change alike: a multimethod, multisample study of age differences in the facets of conscientiousness. Journal of Personality and Social Psychology 96 (2), 446.

John, O.P., Naumann, L.P., Soto, C.J., 2008. Paradigm shift to the integrative big five trait taxonomy: history, measurement, and conceptual issues. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research, third ed. Guilford Press, New York, pp. 114–158.

Kotov, R., Gamez, W., Schmidt, F., Watson, D., 2010. Linking "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. Psychological Bulletin 136 (5), 768–821.

Krueger, R.F., 2013. Personality disorders are the vanguard of the post-DSM-5.0 era. Personality Disorders: Theory, Research, and Treatment 4 (4), 355–362.

Krueger, R.F., 1999. The structure of common mental disorders. Archives of General Psychiatry 56 (10), 921–926.

Krueger, R.F., Caspi, A., Moffitt, T.E., Silva, P.A., McGee, R., 1996. Personality traits are differentially linked to mental disorders: a multitrait-multidiagnosis study of an adolescent birth cohort. Journal of Abnormal Psychology 105 (3), 299–312.

Krueger, R.F., Derringer, J., Markon, K.E., Watson, D., Skodol, A.E., August 2013. The Personality Inventory for DSM-5 (PID-5) – Adult, American Psychiatric Association. Available from: http://www.psychiatry.org/File%20Library/Practice/DSM/DSM-5/ThePersonalityInventoryForDSM5FullVersionAdult.pdf.

Krueger, R.F., McGue, M., Iacono, W.G., 2001. The higher-order structure of common DSM mental disorders: internalization, externalization, and their connections to personality. Personality and Individual Differences 30 (7), 1245–1259.

Livesley, W.J., Jackson, D.N., 2009. Manual for the Dimensional Assessment of Personality Pathology – Basic Questionnaire. Sigma Press, Port Huron, MI.

Livesley, W.J., Jang, K.L., Vernon, P.A., 1998. Phenotypic and genetic structure of traits delineating personality disorder. Archives of General Psychiatry 55 (10), 941–948.

Lynam, D.R., Widiger, T.A., 2001. Using the five-factor model to represent the DSM-IV personality disorders: an expert consensus approach. Journal of Abnormal Psychology 110 (3), 401–412.

Maher, B.A., Maher, W.B., 1994. Personality and psychopathology: a historical perspective. Journal of Abnormal Psychology 103 (1), 72–77.

Markon, K.E., Krueger, R.F., Watson, D., 2005. Delineating the structure of normal and abnormal personality: an integrative hierarchical approach. Journal of Personality and Social Psychology 88 (1), 139–157.

McCrae, R.R., Costa Jr., P.T., 1990. Personality in Adulthood. Guilford Press, New York.

McCrae, R., Costa, P., 2003. Personality in Adulthood: A Five-factor Theory Perspective, second ed. Guilford Press, New York.

O'Connor, B., Dyce, J., 1998. A test of models of personality disorder configuration. Journal of Abnormal Psychology 107 (1), 3–16.

Paunonen, S.V., Ashton, M.C., 2001. Big five factors and facets and the prediction of behavior. Journal of Personality and Social Psychology 81 (3), 524–539.

Roberts, B.W., Chernyshenko, O.S., Stark, S., Goldberg, L.R., 2005. The structure of conscientiousness: an empirical investigation based on seven major personality questionnaires. Personnel Psychology 58 (1), 103–139.

Samuel, D.B., Widiger, T.A., 2008. A meta-analytic review of the relationships between the five-factor model and DSM-IV-TR personality disorders: a facet level analysis. Clinical Psychology Review 28 (8), 1326–1342.

Saulsman, L.M., Page, A.C., 2004. The five-factor model and personality disorder empirical literature: a meta-analytic review. Clinical Psychology Review 23 (8), 1055–1085.

South, S.C., Eaton, N.R., Krueger, R.F., 2010. The connections between personality and psychopathology. In: Millon, T., Krueger, R.F., Simonsen, E. (Eds.), Contemporary Directions in Psychopathology: Scientific Foundations of the DSM-V and ICD-11. Guilford Press, New York, pp. 242–262.

Tellegen, A., Waller, N.G., 2008. Exploring personality through test construction: development of the multidimensional personality questionnaire. In: Boyle, G.J., Matthews, G., Saklofske, D.H. (Eds.), The SAGE Handbook of Personality Theory and Assessment, Personality measurement and Testing, vol. 2. Sage Publications, Thousand Oaks, CA, pp. 261–292.

Trull, T.J., Durrett, C.A., 2005. Categorical and dimensional models of personality disorder. Annual Review of Clinical Psychology 1, 355–380.

Trull, T.J., Sher, K.J., 1994. Relationship between the five-factor model of personality and axis I disorders in a nonclinical sample. Journal of Abnormal Psychology 103 (2), 350–360.

Tupes, E.C., Christal, R.E., 1992. Recurrent personality factors based on trait ratings. Journal of Personality 60 (2), 225–251.

Van de Vijver, F., Valchev, V. Five factor model of personality, universality of, in press.

Vrieze, S.I., McGue, M., Miller, M.B., Hicks, B.M., Iacono, W.G., 2013. Three mutually informative ways to understand the genetic relationships among behavioral disinhibition, alcohol use, drug use, nicotine use/dependence, and their co-occurrence: twin biometry, GCTA, and genome-wide scoring. Behavior Genetics 43 (2), 97–107.

Walton, K.E., Roberts, B.W., 2004. On the relationship between substance use and personality traits: abstainers are not maladjusted. Journal of Research in Personality 38 (6), 515–535.

Watson, D., Clark, L.A., 1993. Behavioral disinhibition versus constraint: a dispositional perspective. In: Wegner, D.M., Pennebaker, J.W. (Eds.), Handbook of Mental Control. Prentice-Hall, Englewood Cliffs, NJ, pp. 506–527.

Watson, D., Clark, L.A., Chmielewski, M., 2008. Structures of personality and their relevance to psychopathology: II. Further articulation of a comprehensive unified trait structure. Journal of Personality 76 (6), 1545–1586.

Widiger, T.A., 1998. Four out of five ain't bad. Archives of General Psychiatry 55 (10), 865–866.

Widiger, T.A., 2003. Personality disorder diagnosis. World Psychiatry 2 (3), 131–135.

Widiger, T.A., 2011. Personality and psychopathology. World Psychiatry 10 (2), 103–106.

Personality, Trait Models of

Peter Borkenau and Nancy Tandler, Martin-Luther-Universität Halle-Wittenberg, Halle, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Traits have been defined as dimensions of individual differences in tendencies to show consistent patterns of thoughts, feelings, and actions. Such consistent patterns are supported by evidence, and robust dimensions of personality differences have been identified. Traits are best measured by multiple indicators, and as they reflect general trends of thinking, feeling, and acting in trait-consistent ways, they predict socially important criteria including job performance, educational attainment, physical and mental health, longevity, and subjective well-being.

What is a trait? McCrae and Costa (2003) define traits as "dimensions of individual differences in tendencies to show consistent patterns of thoughts, feelings, and actions" (p. 25). *Dimensions of individual differences* means that traits refer to differences between persons, assuming that these differences are continuous. For example, by saying that Anne is sociable, we mean that Anne's sociability is above average in a population of persons who differ continuously in their sociability. The deviations from the population average may be large, small, or nil, the distributions of most traits approximating the familiar normal curve. In psychometrics, trait models are distinguished from class models. Whereas trait models imply continuous differences between persons, class models imply qualitatively different personality types (De Boeck et al., 2005). An example of a class model having received considerable empirical support (Chapman and Goldberg, 2011) is Block's (1971) distinction between resilients, overcontrollers, and undercontrollers.

Patterns of thoughts, feelings, and actions means that traits do not describe single and specific habits like smoking or talking in a loud voice, but broad individual differences in experience and behavior. Moreover, *tendencies* implies that traits influence but do not determine persons' experience and behavior. Traits matter, but situations matter too (Fleeson, 2004). Traits are therefore conceived as dispositions implying probabilities: Persons with a high trait level are more likely to show trait-consistent behavior than persons with a low trait level, but probabilities of exactly 1.0 do never occur. For example, extraverts are more likely to seek the company of others than introverts, but even an extremely extraverted person may sometimes prefer to be alone.

Consistent means that differences between persons reflecting traits are stable across time and situations. That Anne has the trait of being gregarious implies that she is more extraverted than others not only today, but will also be so next year. Nevertheless, some long-term change in personality occurs: First, there are age trends. For example, extraversion tends to decline with age. Second, the rank order of personality differences is largely but not perfectly stable (Lucas and Donnellan, 2011). If Anne is more extraverted than her classmate Bill at age 15, there is a slight probability that Bill will be more extraverted than Anne 20 years later. Moreover, rank-order stabilities change across the life span: Personality differences tend to stabilize until the sixth decade of life, but afterward stability decreases (Lucas and Donnellan, 2011; Roberts and DelVecchio, 2000). Finally, personality differences are manifested across multiple situations. For instance, extraverts seek the company of others not only during leisure but also in the workplace. It is an empirical issue, however, across which situations trait-relevant behaviors generalize. Among others, this is linked to the breadth of the trait. There are narrower traits predicting behavior in a more restricted range of situations than do broader traits (Hampson et al., 1986). An example is *helpful* versus *kind*: Whereas helpful persons tend to support others being in need of help, kind persons tend to please others in a broader range of situations.

Personality psychologists distinguish between traits and states. Whereas traits are largely (but not perfectly) stable, states are transient. Traits can be understood as frequencies that relevant states occur. Measuring personality states repeatedly, Fleeson (2001) compared the variation of personality states within individual persons with the variation between persons in their mean state levels, averaged across multiple occasions: For most traits, more variance was found within than between persons. For example, the participants' conscientiousness varied more from one occasion to the next than that the averaged levels of conscientiousness varied from one participant to the other. This shows that situations influence behavior, but it is useful to distinguish between two varieties of situational influences: First, variability within persons may reflect main effects of situations leaving the persons' rank order unchanged. A hypothetical example would be that most persons are more anxious before than after an exam, and those being more anxious before are so after the exam as well. As the rank order of persons remains stable across situations here, such a pattern would strongly support a trait model of personality, exemplifying that situational influences are entirely compatible with a trait model of individual differences. Second, the rank order of persons may vary from one situation to the next, that is, there may be Person × Situation interactions, implying that the individual differences are not perfectly consistent across situations.

As a matter of fact, cross-situational consistencies are moderate if behavior in a single setting is compared with behavior in another single setting. Across 15 different single settings (e.g., introducing oneself, telling a joke), Borkenau et al. (2004) obtained an average reliability-corrected cross-situational consistency coefficient of 0.43. But cross-situational consistencies tend to be much higher if the average level of trait-relevant behavior in a *set* of settings is compared with the average level of trait-relevant behavior in another *set* of settings

(Epstein, 1979; Fleeson, 2001, 2004). Thus traits are not particularly useful for predicting single behaviors in specific situations, but they are highly useful for predicting general trends of thinking, feeling, and acting in trait-consistent ways. For example, a conscientious person will act, on average, in more organized ways than a less conscientious person. Such tendencies strongly influence important outcomes in various domains of life. And because these tendencies are also quite stable across time, traits predict numerous socially important criteria (Ozer and Benet-Martínez, 2006), among them life satisfaction (Steel et al., 2008), longevity (Bogg and Roberts, 2004; Hampson and Friedman, 2008), academic achievement (Poropat, 2009), and career success (Connelly and Ones, 2010).

But the view that personality should be conceptualized in terms of traits has not remained undisputed. Mischel and Shoda (1995, 2008) suggest studying intraindividual patterns of variability instead of cross-situationally consistent patterns of behavior. For example, whereas Anne may be more talkative in large groups than in dyadic settings, the reverse may be true of Bill. As far as such situation–behavior interactions are stable across time, Mischel and Shoda (2008) argue that they are highly diagnostic of persons' cognitive and affective processes. This is frequently referred to as the *social-cognitive approach* to personality.

Measuring Traits

Measuring tendencies for showing consistent patterns of thoughts, feelings, and actions via behavior observations would be extremely laborious: multiple behaviors by numerous persons would have to be observed and coded repeatedly in multiple situations. Yet even then many thoughts and feelings would escape their observation. Therefore, most researchers of personality have relied on self-reports, harnessing that persons are aware of their own thoughts, feelings, and actions in many situations, and remember them over long periods of time. But self-reports are not without problems: People may be unaware of their unconscious motives, and their memories may be distorted. Moreover, persons may misrepresent themselves knowingly or unknowingly, to protect their reputation and self-esteem (Paulhus, 1984; Vazire, 2010). Furthermore, as the only persons who can provide self-reports are the to-be-described persons themselves, their objectivity cannot be ascertained (Hofstee, 1994).

Researchers have therefore looked for convenient additional methods for measuring personality traits, and have identified descriptions by knowledgeable informants like peers and spouses as a useful complement (Vazire, 2006). Reports by knowledgeable informants seem to be less valid measures of neuroticism-related but more valid measures of intellect-related traits than self-reports (Vazire, 2010). Moreover, other-ratings seem to be more valid predictors than self-reports of academic achievement and job performance, and they predict incrementally to self-reports (Connelly and Ones, 2010). Some authors therefore argue that aggregates of self-reports, informant reports, and maybe additional measures provide the best possible assessment of personality traits (Letzring et al., 2006).

Most measures of personality are scales consisting of multiple items being correlated with each other, each item alluding to a somewhat different facet of the trait. For example, a scale measuring neuroticism may comprise items addressing feelings of inferiority and items addressing feelings of anxiety. The responses to these items are aggregated (usually summed up). There are several advantages to such multiitem scales over one-item measures of traits: First, multiitem scales allow for more fine-grained distinctions between trait levels. For example, if responses are provided on five-point rating scales, and a scale consists of 10 items, there are 41 possible scores. Second, the reliability (precision) of the measures increases with the number of their items, as random errors cancel each other out. Third, the content being common to the items and reflected in their covariances is highlighted, whereas item-specific variances are suppressed. That is desirable as traits are quite general patterns of thoughts, feelings, and actions, not the quite specific behaviors and experiences that single items ask for.

Sources of Traits

Why do persons differ in their traits? Ultimately, genetic as well as environmental factors are influential, and quantitative behavior–genetic research has gone for estimating their relative contributions. Most of this kind of research relies on the classical twin design, comparing the similarity of monozygotic (MZ) twins with the similarity of dizygotic (DZ) twins reared in the same household. This design allows partitioning the phenotypic (observed) variance into three components: (1) variance accounted for by the effects of genes, (2) shared environmental variance contributing to the similarity of persons reared together independent of their genetic relatedness, and (3) nonshared environmental variance not contributing to the similarity of persons. Higher correlations between MZ twins than DZ twins suggest genetic, similar and positive twin correlations for MZ twins and DZ twins suggest shared-environmental, and less-than-perfect correlations between MZ twins reared together suggest nonshared-environmental influences.

Most twin studies on personality traits relied on self-reports, some of them studying very large samples. Put in a nutshell, the core finding of these studies has been that genes and nonshared environment each account for approximately 50% of the individual differences in personality traits, whereas shared environmental influences seem to be very small (Jang et al., 1996). Thus being raised in the same household seems to contribute negligibly to similarity in personality. But twin studies relying on self-reports do not allow separating nonshared environmental influences from measurement error. Measurement error can be appropriately controlled, however, if each of the twins under study is described by at least two informants. Running such a study and controlling for measurement error, Riemann et al. (1997) found that genes and nonshared environment account for approximately 70 and 30% of the personality differences, respectively.

Finding genetic influences on personality traits is interesting but raises a lot of questions. One question is whether specific genes can be located in the genome that account for the observed individual differences. This is the field of molecular behavior genetics. Here, the frequencies of different gene variants, so-called alleles, are compared between groups of persons scoring high versus low on a trait. One early finding of this line

of research has been that a sequence-length polymorphism (the occurrence of either a short or a long allele) at a gene locus in the regulatory region of the serotonin transporter gene (5-HTTLPR) accounted for 3–4% of the variance in anxiety-related traits (Lesch et al., 1996). Later on, it was reported that this genotype interacts with environmental stressors to the effect that the number of stressful life events raises the risk of a major depressive episode only among persons with at least one short variant of this allele (Caspi et al., 2003). Such findings are potentially important as they may help explain why some persons seem to be invulnerable to the effects of environmental stressors whereas others are not. In a meta-analysis of 14 relevant studies, however, Risch (2009) found a main effect of the number of stressful life events only, but neither a main effect of the genetic polymorphism as reported by Lesch et al. (1996), nor a gene–environment interaction as reported by Caspi et al. (2003).

Such failures to replicate are not uncommon in allelic association studies, probably reflecting that (1) personality traits are influenced by thousands of genes, the effect of each single gene being very small; and (2) there are millions of genetic polymorphisms in the human genome that might be related to human behavior, resulting in a high probability to occasionally find false-positive gene–behavior associations (McCrae et al., 2010). But much molecular-genetic research on personality and psychopathology is going on, and some genes that exert a strong influence on behavior might yet be identified.

Another issue is how genes affect behavior. After all, genes code for proteins, or they regulate the coding of proteins by other genes. Thus their effects on behavior are indirect. There are several possible pathways, many of them involving neurotransmitters, i. e., chemical substances transmitting information in the brain. Cloninger (1987) suggested three major brain systems modulated by different neurotransmitters: A behavioral activation system (BAS) modulated by dopamine and accounting for individual differences in novelty seeking (a trait resembling extraversion), a behavioral inhibition system (BIS) modulated by serotonin and accounting for individual differences in harm avoidance (a trait resembling neuroticism), and a behavioral maintenance system modulated by norepinephrine and accounting for individual differences in reward dependence (a trait resembling agreeableness). Likewise, Gray and McNaughton (2000) suggested a BAS organizing responses to signals of reward, a fight-flight-freezing system organizing responses to signals of punishment, and a BIS being activated in cases of conflict, that is, if a stimulus signals potential reward as well as potential punishment. Smillie (2008) discusses possible genetic sources of individual differences in these three systems.

Which Traits Should Be Measured and Studied?

In major languages like English, there are thousands of personality-descriptive terms, many of them describing traits. Measuring all of them would be extremely laborious, unacceptable to research participants, and redundant because discrete terms may denote very similar (e.g., friendly and kind) or opposite (e.g., friendly and hostile) traits. Personality researchers have therefore sought to identify 'basic dimensions' of personality that research should focus on. The data-analytic method mostly used here is factor analysis that accounts for correlations between observed variables, for example scores on multiple personality scales, by their loadings on common latent variables called *factors*. For example, that scales measuring anxiety are substantially correlated with scales measuring depression has been interpreted as evidence that anxiety and depression are facets of a common latent factor labeled either Neuroticism or – its opposite – Emotional Stability. Several factor models of personality have been suggested by authors like Guilford, Cattell, Eysenck, and Tellegen, but the currently most popular model is the Five-Factor model of personality advocated by several authors, most prominently Goldberg (1990) and McCrae and Costa (1987).

The Five-Factor model is an outcome of the lexical approach to personality that is based on the lexical hypothesis. As the factors being found depend on the variables being entered into a factor analysis, identifying the basic factors of personality requires that the traits in the analysis are representative of all traits. The lexical hypothesis posits that most of the socially relevant and salient personality characteristics have been encoded in the natural language. Thus, the personality vocabulary contained in the dictionaries of a natural language provides an extensive, yet finite, set of attributes that the people speaking that language have found important in their daily interactions (Goldberg, 1981). Therefore, researchers have culled thousands of trait-descriptive terms from comprehensive lexica, and grouped them into clusters of similar terms. These clusters of trait-descriptive terms were then used for describing actual persons, and the descriptions were correlated. Five recurrent factors were identified, and 'nothing else of any consequence' (Tupes and Christal, 1992). Moreover, very similar factors were found in descriptions by acquaintances on adjective scales, in self-reports on such scales, and in self-reports on personality questionnaires (McCrae and Costa, 1987).

Usually, these factors are labeled *Neuroticism*, *Extraversion*, *Openness to Experience*, *Agreeableness*, and *Conscientiousness*. This Five-Factor model is extremely useful by providing a common language for personality researchers thus helping them to coordinate their research efforts. In recent research on issues like life-span development and rank-order stability of personality (Lucas and Donnellan, 2011), interjudge agreement on personality traits (Connelly and Ones, 2010), prediction of academic achievement (Poropat, 2009), prediction of job performance (Connelly and Ones, 2010), and the relationship between personality and subjective well-being (Steel et al., 2008), personality differences have been framed in terms of the Five-Factor model.

That does not imply that traits not covered by the Five-Factor model are unimportant. First, personality traits are ordered hierarchically, the Five-Factor model describing personality differences at a broad level. There are also more specific traits describing personality at a narrower level, like anxiety and depression that may both be understood as facets of neuroticism. Second, some researchers are less interested in the basic dimensions of personality and their facets, than in traits they consider as useful for their special purposes. Thus Watson and Tellegen (1985) were particularly interested in current and long-term affect. Studying the structure of

mood, they identified two dimensions of core affect, Positive Activation contrasting excitement at the high pole with depression at the low pole, and Negative Activation contrasting distress at the high pole with calmness at the low pole (Watson et al., 1999). Individual differences in these dimensions were found to be quite stable across time (Watson and Walker, 1996).

And they are not independent of the dimensions of the Five-Factor model: Whereas Positive Activation is strongly associated with Extraversion, Negative Activation is strongly associated with Neuroticism. These findings gave rise to a distinction suggested by Elliot and Thrash (2002) between approach motivation subsuming Extraversion, Positive Activation, and a strong BAS, and avoidance motivation subsuming Neuroticism, Negative Activation, and a strong BIS.

Personality Traits and Socially Relevant Outcomes

That traits describe general trends to think, feel, and act in trait-consistent ways, makes them highly useful predictors of socially important criteria in several life domains. Ones and Viswesvaran (2011) summarized several meta-analyses concerning occupational outcomes, demonstrating that conscientiousness is a consistent predictor of overall job performance. This holds even across heterogeneous sets of professions and countries with validity coefficients ranging from 0.19 to 0.22. Furthermore, it was shown that conscientiousness predicts job satisfaction positively ($r = 0.26$), whereas it is negatively related to counterproductive work behavior ($r = -0.38$) and turnover ($r = -0.22$). According to a meta-analysis by Barrick et al. (2001), the other traits in the Five-Factor model are important for more specific job performance criteria: Whereas extraversion is positively related to managerial performance and training proficiency, neuroticism is negatively linked to successful social interacting, whereas agreeableness predicts successful teamwork, and openness to experience is associated with training proficiency. The importance of these traits is not limited to predicting performance. Neuroticism is additionally linked to satisfaction at work ($r = -0.29$) and to turnover ($r = 0.20$), whereas agreeableness is additionally related to counterproductive work behavior ($r = -0.28$) and turnover ($r = -0.27$; Barrick et al., 2001; Ones and Viswesvaran, 2011). Interestingly, the prediction of occupational outcomes from personality traits seems to be unique as it persists if socioeconomic status and cognitive abilities are controlled for (Ones and Viswesvaran, 2011; Roberts et al., 2007). Furthermore, turnover is predicted more strongly from conscientiousness, neuroticism, or agreeableness than from characteristics of the job like its complexity (Zimmerman, 2008).

Considering academic performance, the associations resemble those in the occupational domain. Conscientiousness is the best predictor of academic performance too. In a meta-analysis by Poropat (2009), conscientiousness correlated with Grade Point Average (GPA) at 0.19 quite independently of the level of intelligence. Agreeableness and openness to experience predicted GPA as well, although with smaller validity coefficients of 0.07 and 0.10, respectively. The correlation of GPA with openness to experience diminished marginally if intelligence was controlled for (Poropat, 2009). The latter result, combined with the positive relation of openness to experience to training proficiency (Ones and Viswesvaran, 2011), underlines the importance of openness to experience for being a successful student in various life domains.

Within the health domain, personality traits show consistent and unique effects on various outcomes like longevity, or physical and mental health processes. They seem to support or to hamper recovering from various diseases. Neuroticism was shown to predict higher overall mortality (Wilson et al., 2005), and to be associated with a higher risk of dying from particular illnesses such as cardiovascular diseases (Denollet et al., 2006; Shipley et al., 2007). Among patients suffering from chronic renal insufficiency, high neuroticism is associated with higher mortality (Christensen et al., 2002). In contrast, extraversion predicts a longer life (Wilson et al., 2005), and high conscientiousness is associated with lower mortality among renal patients (Christensen et al., 2002). Personality traits predict mortality uniquely even if socioeconomic (Roberts et al., 2007) and initial health status (Shipley et al., 2007) are controlled. In terms of risk factors for specific diseases, for example, cardiovascular diseases, studies showed that neuroticism and hostility (low agreeableness) are positively, whereas conscientiousness is negatively, associated with weight gain and increase of body mass index (Brummett et al., 2006; Niaura et al., 2000). Furthermore, conscientiousness is inversely related to adverse health behaviors (e.g., excessive alcohol and drug use, tobacco use, unhealthy eating) which are important contributors to poor health outcomes such as cardiovascular diseases (Bogg and Roberts, 2004).

Personality traits are associated with mental disorders as well. According to a meta-analysis by Kotov et al. (2010), neuroticism is directly related to depressive disorders, anxiety disorders, and substance abuse that are quite common in the adult population. The average effect size of neuroticism on all mentioned disorders was $d = 1.65$. In contrast, conscientiousness showed strong and consistently negative connections with mental disorders ($d = -1.01$), and the effect of extraversion was also negative but limited to fewer disorders (e.g., social phobia, dysthymic disorder). The same held for agreeableness and openness to experience that showed a few negative correlations to some kinds of substance abuse and to the dysthymic disorders, respectively.

Finally, personality traits predict subjective well-being. In a meta-analysis by Steel et al. (2008) neuroticism and extraversion were strong predictors of subjective well-being, with neuroticism being directly related to negative affect and inversely related to happiness and quality of life. In contrast, extraversion was directly related to positive affect, happiness, and quality of life.

To summarize, across various life domains such as occupation, education, health, and subjective well-being, personality traits, that is individual differences in tendencies to show consistent patterns of feelings, thoughts, and actions, predict numerous and socially important real-life criteria.

See also: Agreeableness; Big Five Factor Model, Theory and Structure; Conscientiousness; Extraversion; Factor Analysis and Latent Variable Models in Personality Psychology; Five Factor Model of Personality, Assessment of; Five Factor Model

of Personality, Facets of; Neuroticism; Openness to Experience; Personality Differences and Development: Genetic and Environmental Contributions; Personality and Educational Outcomes; Subjective Wellbeing, Psychology of.

Bibliography

Barrick, M.R., Mount, M.K., Judge, T.A., 2001. Personality and performance at the beginning of the new millennium: what do we know and where do we go next? International Journal of Selection and Assessment 9, 9–30.

Block, J., 1971. Lives through Time. Bancroft Press, Berkeley.

Bogg, T., Roberts, B.W., 2004. Conscientiousness and health-related behaviors: a meta-analysis of the leading behavioral contributors to mortality. Psychological Bulletin 130, 887–919.

Borkenau, P., Mauer, N., Riemann, R., Spinath, F.M., Angleitner, A., 2004. Thin slices of behavior as cues of personality and intelligence. Journal of Personality and Social Psychology 86, 599–614.

Brummett, B.H., Babyak, M.A., Williams, R.B., Barefoot, J.C., Costa, P.T., Siegler, I.C., 2006. NEO personality domains and gender predict levels and trends in body mass index over 14 years during midlife. Journal of Research in Personality 40, 222–236.

Caspi, A., et al., 2003. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science 301, 386–389.

Chapman, B.P., Goldberg, L.R., 2011. Replicability and 40-year predictive power of childhood ARC types. Journal of Personality and Social Psychology 101, 593–606.

Christensen, A.J., Ehlers, S.L., Wiebe, J.S., Moran, P.J., Raichle, K., Ferneyhough, K., et al., 2002. Patient personality and mortality: a 4-year prospective examination of chronic renal insufficiency. Health Psychology 21, 315–320.

Cloninger, C.R., 1987. A systematic method for clinical description and classification of personality variants. Archives of General Psychiatry 44, 573–588.

Connelly, B.S., Ones, D.S., 2010. An other perspective on personality: meta-analytic integration of observers' accuracy and predictive validity. Psychological Bulletin 136, 1092–1122.

De Boeck, P., Wilson, M., Acton, G.S., 2005. A conceptual and psychometric framework for distinguishing categories and dimensions. Psychological Review 112, 129–158.

Denollet, J., Pedersen, S.S., Vrints, C.J., Conraads, V.M., 2006. Usefulness of type D personality in predicting five-year cardiac events above and beyond concurrent symptoms of stress in patients with coronary heart disease. The American Journal of Cardiology 97, 970–973.

Elliot, A.J., Thrash, T.M., 2002. Approach-avoidance motivation in personality: approach and avoidance temperaments and goals. Journal of Personality and Social Psychology 82, 804–818.

Epstein, S., 1979. The stability of behavior: I. On predicting most of the people much of the time. Journal of Personality and Social Psychology 37, 1097–1126.

Fleeson, W., 2001. Towards a structure- and process-integrated view of personality: traits as density distributions of states. Journal of Personality and Social Psychology 80, 1011–1027.

Fleeson, W., 2004. Moving personality beyond the person-situation debate: the challenge and the opportunity of within-person variability. Current Directions in Psychological Science 13, 83–87.

Goldberg, L.R., 1981. Language and individual differences: the search for universals in personality lexicons. In: Wheeler, L. (Ed.), Review of Personality and Social Psychology, vol. 2. Sage, Beverly Hills, pp. 141–165.

Goldberg, L.R., 1990. An alternative "description of personality": the Big-Five factor structure. Journal of Personality and Social Psychology 59, 1216–1229.

Gray, J.A., McNaughton, N., 2000. The Neuropsychology of Anxiety. An Inquiry into the Functions of the Septo-hippocampal System, second ed. Oxford University Press, Oxford.

Hampson, S.E., Friedman, H.S., 2008. Personality and health: a lifetime perspective. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research. Guilford Press, New York, pp. 770–794.

Hampson, S.E., John, O.P., Goldberg, L.R., 1986. Category breadth and hierarchical structure in personality: studies of asymmetries in judgments of trait implications. Journal of Personality and Social Psychology 51, 37–54.

Hofstee, W.K.B., 1994. Who should own the definition of personality? European Journal of Personality 8, 149–162.

Jang, K.L., Livesley, W.J., Vernon, P.A., 1996. Heritability of the big five personality dimensions and their facets: a twin study. Journal of Personality 64, 577–591.

Kotov, R., Gamez, W., Schmidt, F., Watson, D., 2010. Linking "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. Psychological Bulletin 136, 768–821.

Lesch, K.P., et al., 1996. Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. Science 274, 1527–1531.

Letzring, T.D., Wells, S., Funder, D.C., 2006. Quantity and quality of available information affect the realistic accuracy of personality judgment. Journal of Personality and Social Psychology 91, 111–123.

Lucas, R.E., Donnellan, M.B., 2011. Personality development across the life span: longitudinal analyses with a national sample from Germany. Journal of Personality and Social Psychology 101, 847–861.

McCrae, R.R., Costa, P.T., 1987. Validation of the five-factor model of personality across instruments and observers. Journal of Personality and Social Psychology 52, 81–90.

McCrae, R.R., Costa, P.T., 2003. Personality in Adulthood: A Five-factor Theory Perspective. Guilford Press, New York.

McCrae, R.R., Scally, M., Terracciano, A., Abecasis, G.R., Costa, P.T., 2010. An alternative to the search for single polymorphisms: toward molecular personality scales for the five- factor model. Journal of Personality and Social Psychology 99, 1014–1024.

Mischel, W., Shoda, Y., 1995. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. Psychological Review 102, 246–268.

Mischel, W., Shoda, Y., 2008. Toward a unified theory of personality: integrating dispositions and processing dynamics within the cognitive-affective processing system. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research. Guilford Press, New York, pp. 208–241.

Niaura, R., et al., 2000. Hostility and the metabolic syndrome in older males: the normative aging study. Psychosomatic Medicine 62, 7–16.

Ones, D.S., Viswesvaran, C., 2011. Individual differences at work. In: Chamorro-Premuzic, T., von Stumm, S., Furnham, A. (Eds.), The Wiley-Blackwell Handbook of Individual Differences. Blackwell Publishing Ltd, Chichester, pp. 379–407.

Ozer, D.J., Benet-Martínez, V., 2006. Personality and the prediction of consequential outcomes. Annual Review of Psychology 57, 401–421.

Paulhus, D.L., 1984. Two-component models of socially desirable responding. Journal of Personality and Social Psychology 46, 598–609.

Poropat, A.E., 2009. A meta-analysis of the five-factor model of personality and academic performance. Psychological Bulletin 135, 322–338.

Riemann, R., Angleitner, A., Strelau, J., 1997. Genetic and environmental influences on personality: a study of twins reared together using the self- and peer report NEO-FFI scales. Journal of Personality 65, 449–475.

Risch, N., 2009. Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression. Journal of the American Medical Association 301, 2462–2471.

Roberts, B.W., DelVecchio, W.F., 2000. The rank-order consistency of personality traits from childhood to old age: a quantitative review of longitudinal studies. Psychological Bulletin 126, 3–25.

Roberts, B.W., Kuncel, N.R., Shiner, R., Caspi, A., Goldberg, L.R., 2007. The power of personality: the comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. Perspectives on Psychological Science 2, 313–345.

Shipley, B.A., Weiss, A., Der, G., Taylor, M.D., Deary, I.J., 2007. Neuroticism, extraversion, and mortality in the UK Health and Lifestyle Survey: a 21-year prospective cohort study. Psychosomatic Medicine 69, 923–931.

Smillie, L.D., 2008. What is reinforcement sensitivity? Neuroscience paradigms for approach-avoidance process theories of personality. European Journal of Personality 22, 359–384.

Steel, P., Schmidt, J., Shultz, J., 2008. Refining the relationship between personality and subjective well-being. Psychological Bulletin 134, 138–161.

Tupes, E.C., Christal, R.C., 1992. Recurrent personality factors based on trait ratings. Journal of Personality 60, 225–251.

Vazire, S., 2006. Informant reports: a cheap, fast, and easy method for personality assessment. Journal of Research in Personality 40, 472–481.

Vazire, S., 2010. Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. Journal of Personality and Social Psychology 98, 281–300.

Watson, D., Tellegen, A., 1985. Toward a consensual structure of mood. Psychological Bulletin 98, 219–235.

Watson, D., Walker, L.M., 1996. The long-term stability and predictive validity of trait measures of affect. Journal of Personality and Social Psychology 70, 567–577.

Watson, D., Wiese, D., Vaidya, J., Tellegen, A., 1999. The two general activation systems of affect: structural findings, evolutionary considerations, and psychobiological evidence. Journal of Personality and Social Psychology 76, 820–838.

Wilson, R.S., Krueger, K.R., Gu, L., Bienias, J.L., de Leon, C.F.M., Evans, D.A., 2005. Neuroticism, extraversion, and mortality in a defined population of older persons. Psychosomatic Medicine 67, 841–845.

Zimmerman, R.D., 2008. Understanding the impact of personality traits on individuals' turnover decisions: a meta-analytic path model. Personnel Psychology 61, 309–348.

Personalized Medicine

Robert A Phillips, University of Toronto, Toronto, ON, Canada
Claus Vögele, Institute for Health and Behaviour, University of Luxembourg, Walferdange, Luxembourg

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The rapid advances in medical science over the past two decades have already changed the way medicine is practiced, but the acceleration of advances suggests that health care systems worldwide are facing a tsunami of new advances in understanding and in technology that will require radical reorganization of the health care system. The improved possibility of personalizing health care is one of the major drivers of change. Unfortunately, health care systems respond very slowly to innovation, and radical changes are almost impossible. In this article, we outline various changes that are expected to happen in the future in relation to personalized medicine, and discuss why behavioral scientists must participate in the reshaping of health care systems and the successful delivery of personalized care at the individual level.

What Is Personalized Medicine?

Historical Perspective

In its most basic sense, personalized medicine refers to the provision of the right drug to the right patient at the right time. Physicians, of course, correctly argue that they have practiced medicine this way for hundreds of years. Diagnosis of disease involved a family history, physical examination, and simple diagnostic tests (e.g., blood test or X-ray). For most diseases, only one or two therapies were available and physicians made their choice of treatment depending on various guidelines and their intuition based on previous experience with other patients. Very often, only a small proportion of patients with a specific disease responded to treatment, but the reasons for the successes and failures of these treatment methods were not clear.

New Advances Enhance Feasibility of Improving Personalized Care

Advances in understanding and technology have greatly changed this historical scenario. Physicians now often understand why some patients respond to treatment and others do not. New molecular understanding of disease sometimes allows for the categorization of patients into subgroups, each with different treatment options. Because of the increased precision of diagnosis and the stratification of patients into groups, the new medicine is often called 'precision medicine' or 'stratified medicine' (Allison, 2008). Nevertheless, in this article we will use the term personalized medicine to refer to these new approaches.

In the following sections, we will give examples of some of the biological and technological discoveries that are driving changes in medicine toward increasing personalized treatments. Many articles have discussed these changes and discussed their implications (Brücher et al., 2014). However, few articles have outlined the potentially important contributions that behavioral scientists must play in the new paradigms of medical practice.

Breadth of Changes

The onset of disease is a complex process involving interactions between the genetic makeup, environmental influences (including diet and gut bacteria), and behavior. Studies on cancer have led to the revolution of personalized medicine, and this article will use mainly examples from cancer. Nevertheless, personalized medicine is affecting all diseases, from metabolic diseases (e.g., diabetes mellitus) to autoimmune diseases (e.g., rheumatoid arthritis) and mental disorders (e.g., anxiety and depression). The challenges that face implementation of personalized treatment for cancer apply equally to other diseases.

Advances in genomics have created the greatest opportunities for personalized treatment. The development of specific biomarkers of disease, such as hormone receptors in breast cancer, and genetic profiles of the entire genome of patients are leading to increased stratification of cancer patients (Dancey et al., 2012). However, genomics goes beyond just diagnosing disease; genomic profiles can sometimes predict who will develop the disease, or who will be likely to respond (or not respond) to a specific therapy. The predictive potential of genomic profiles will require a higher level of interaction between health care professionals from several disciplines and patients and families in managing their health and health care.

Diagnosis and Screening

Precise diagnosis is essential in order to categorize individuals into the most appropriate group for treatment. The accuracy of the diagnosis depends on high sensitivity (i.e., very few false negative tests) and high specificity (i.e., few false positives) of the diagnostic tests. As interventions to improve health become increasingly complex and expensive, the feasibility and sustainability of the intervention depends on the accuracy and feasibility of the diagnostic test. In addition, since many interventions are only suitable for a small proportion of individuals with unique genetic or physiological traits, tests must be cost effective. For example, a test costing $100 becomes very expensive if only 1% of patients meet the criteria for receiving the intervention; detecting each responsive patient costs $10 000, a factor that must be included in cost–benefit analyses of the intervention.

Similarly, effective treatments may become uneconomic if the diagnostic test lacks specificity (i.e., has many false positive

results) and results in substantial numbers of individuals receiving an unnecessary, expensive intervention. The prostate-specific antigen (PSA) test for prostate cancer is an example of a test with poor specificity (Bunting et al., 2002). While a high proportion of individuals with early prostate cancer have this biomarker in their blood, many nonmalignant conditions also lead to an increase in PSA. Most countries no longer recommend routine PSA testing because it leads to many unnecessary biopsies and surgeries.

Treatment

Cancer results when an individual cell in the body acquires mutations in several pathways that regulate cell growth and survival (Hanahan and Weinberg, 2011). These mutations disrupt the regulation of cell division resulting in uncontrolled growth and cancer. Recent studies in gene sequencing and in understanding the regulation of cell division show that there are many routes to cancer, and that understanding the precise defects can sometimes lead to a novel therapy that targets the mutated pathway. For example, understanding that the unique Philadelphia chromosome abnormality in all patients with chronic myelogenous leukemia (CML) created an abnormal *abl* oncogene, which led to the development of Gleevec, a novel chemical that acts specifically on this abnormal protein and one other, the *kit* oncogene, which is mutated in the rare gastrointestinal stromal tumors (GIST). This therapy results in long-term survival of both CML and GIST and is the best example in cancer of how understanding the molecular basis of disease can lead to effective new therapies (Druker, 2009).

Another recent example is a new treatment for lung cancer. A small proportion of patients with lung cancer (5–15%) have alterations in a gene called ALK-1, and these patients undergo significant remissions in the disease following treatment with crizotinib, a specific inhibitor of ALK-1 (Gandhi and Janne, 2012).

In the United States, the FDA has approved over 30 targeted therapies for cancer. In most instances, the agencies that fund the therapies require a diagnostic test measuring a biomarker to ensure that the patient has the potential to respond to the therapy. Most biomarkers are either specific mutations, such as ALK-1 or abl, or for unique growth factor receptors, such as the HER2 receptor in some breast cancers (National Cancer Institute, 2013).

Pharmacogenetics will play a major role in developing personalized treatments. Most injectable or oral medications are metabolized and ultimately degraded and/or modified and excreted. However, individual patients metabolize drugs at different rates. Although metabolic pathways may be the same for the vast majority of people, there are interindividual differences in the activity of the enzymes that make up these pathways. Given the same dose of drug, this variation can result in marked differences between individuals in drug response (i.e., concentrations), depending on the rate of drug metabolism. Rapid metabolizers may get little benefit if the blood level of the drug never reaches the concentrations needed for a therapeutic effect. Conversely, a very slow metabolizer may have a higher than desired blood level leading to toxic side effects, such as for Warfarin, a commonly used drug for preventing thrombosis (Wheeler et al., 2013).

An additional diagnostic approach to identify those at risk of toxic side effects because of slow metabolic rates may involve behavioral and psychophysiological assessments, such as for Efavirenz, a commonly used nonnucleoside reverse transcriptase inhibitor, frequently prescribed in combination with other antiretroviral regimens (e.g., protease inhibitors) for the treatment of HIV (Sütterlin et al., 2010).

Prevention

According to a recent report from the World Health Organization, the two major causes of mortality in the developed world are cardiovascular disease and cancer. **Figure 1** shows the data for the 10 leading causes of death in high-income countries (The World Health Organization, n.d.); in developing countries, the same diseases are prevalent, but infectious diseases become more important. There are preventive strategies applicable to all of these causes. However, in almost all instances, the preventive intervention requires behavior modification: increased physical activity, eating a healthy diet (five portions of fruit and vegetables daily), smoking cessation, and participation in regular screening for diseases.

Tobacco use is the best example of the difficulty of changing behavior to improve health. The first warnings about the dangers of tobacco appeared in the 1950s with the report from the United Kingdom by Doll and Hill in 1954 on the high incidence of lung cancer in doctors who smoked, and finally the 1964 Surgeons General Report in the United States that identified smoking as the major cause of lung cancer in men (for review, see (Proctor, 2012)). These reports stimulated many efforts to reduce smoking, but in the past 50 years, smoking prevalence in North America has only decreased from 45 to 20% despite widespread acceptance that smoking is bad for health. In many parts of the Far East, rates exceed 50% and show no signs of decreasing (World Health Organization, 2008).

Palliative Care and End-of-Life Care

Palliative care is any treatment or procedure to alleviate symptoms or to slow the progression of disease rather than to attempt cure. Improving quality of life is the primary goal of palliative care processes. Palliative care can involve relief of

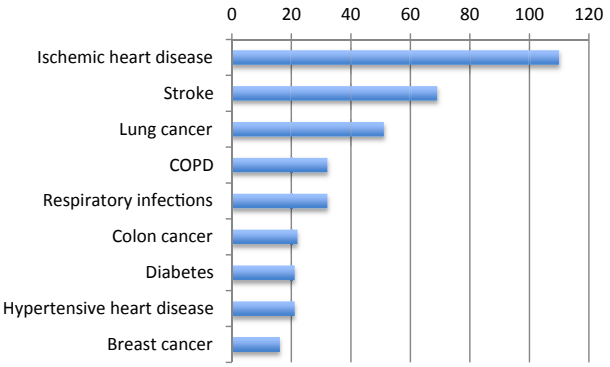

Figure 1 Leading causes of death in high-income countries (mortality/100 000). COPD, Chronic obstructive pulmonary disease.

pain, nausea, physical disabilities, and stress, and taking care of psychological and spiritual issues.

The advances in genomic characterization of neural receptors involved in pain and in the understanding of the physiological and psychological aspects of pain have increased the potential for personalized pain management. Approaches to chronic pain management involve a combination of analgesics, psychological support, and physical therapy, but the algorithms for pain management remain a work in progress (Bruehl et al., 2013).

End-of-life care refers to the support to patients and their families during and after death. Caregivers often fail to discuss death with patients and families dealing with life-threatening illnesses. The end result is that patients often die in situations that are not comfortable and compassionate, and family members do not receive adequate support for bereavement. Providing personalized information to terminally ill patients, i.e., information that strikes the balance between ensuring that patients and their families are fully informed and fostering adaptive coping that allows for hope has been found to be crucial for successful communication in medical settings, and in alleviating stress and improving well-being, even in those facing imminent death (Cavers et al., 2013; McLean et al., 2011).

Scotland, through the Scottish Partnership for Palliative Care, has created one of the most innovative programs to deal with end-of-life care. The program called 'Good Life, Good Death, Good Grief' focuses on individuals and their families in the recognition that there are many community and social and cultural aspects to end-of-life care (Scottish Partnership for Palliative Care, n.d.).

Implications for Health Care Systems

The advances associated with personalized medicine will have multiple effects both on the health care system and on the way individuals manage their own health.

The impending evolution from traditional health care approaches to a system that focuses on personalized, precision medicine requires a radical change to health care. Historically, health care has evolved into a slow evolutionary process with many new innovations in health care taking up to 15 years to go from confirmation in a clinical study to widespread adoption by health care professionals. This rate of evolution must accelerate dramatically for the health care system to provide the many benefits of recent research in the medical and behavioral sciences. The following sections outline a few of the most challenging issues.

Health Systems Facing a Tsunami of Change

Risk of Increased Costs

Many of the newly targeted therapies have very high costs, sometimes in excess of $100 000 per year. Funders of health care, usually governments or insurance companies, can only justify costs if these interventions are limited to those likely to benefit. A major challenge to health care systems will be to implement the new personalized approaches in a cost-neutral or minimal-cost process. This restructuring of health care systems will require elimination of noneffective procedures (e.g., routine health exams), and restriction of medications only to those most likely to benefit. The widespread off-label use of drugs must decrease.

Cost–benefit studies are already becoming standard in many countries before a new innovation is introduced and covered by the health system. The National Institute for Health and Care Excellence (NICE) system in the United Kingdom is leading the way in this process. NICE is developing a process to evaluate both health and social value benefits and the costs of any benefit (Shah et al., 2012). To help make the decision objective and not subjective, the expert teams from NICE calculate the cost of health benefits in terms of dollars (pounds) required to achieve 1 year of quality-adjusted life years (QALY). The calculation of a QALY considers increase in both life years and quality of life expected by the new treatment. Generally, if a new intervention costs more than £20 000–30 000/QALY, it will not be considered to have sufficient benefit to be recommended for routine use. Other countries are using a similar approach and a similar benchmark for implementation. However, funders sometimes make exceptions for expensive drugs used for very rare diseases. For example, Cerezyme (imiglucerase) is an enzyme replacement for individuals with Gaucher's disease, a hereditary disease with a deficiency in this enzyme (Doneda et al., 2013). Treatment costs $200 000 per year and patients must take the supplement for their entire life. Nevertheless, many countries have opted to cover the cost because of the benefits to patients and because the disease is relatively rare (~1/50 000 in the U.S. population).

Electronic Records and Data Management

A recent U.K. report, 'Realising the potential of stratified medicine' emphasized that collection of data is important and will require standardization of data collection, enhanced bioinformatics to analyze the data, and strong policies to protect privacy while allowing sharing of data with physicians, patients, and researchers (The Academy of Medical Sciences, 2013).

All health care systems of the world currently struggle to create electronic health records that will allow physicians anywhere to have access to basic clinical data on patients. However, the new era of personalized medicine will generate literally billions of data points for each patient. This volume of data will create three major challenges. The first concerns the infrastructure to generate the data and to store them securely and confidentially but yet allow easy access to qualified caregivers. The second challenge will be the need to create sophisticated analytical tools to mine the data for the key information that will inform patients and caregivers about the nature of their illness and suggestions for future studies and even possible treatments. The third challenge will be to develop regulatory frameworks that allow sharing of patient data across jurisdictional barriers. At present, regulations and technical challenges make it difficult to even share data between hospitals. But the future will require data sharing between regions within a country and between countries.

Health research will also require access to large international data sets in order to identify critical correlations between biological biomarkers, clinical symptoms, behavior traits, and

environmental exposures. As stratification becomes more precise, it will be difficult for a single institution to have enough patients to completely characterize subgroups. Some recent studies in neuropsychiatric diseases have involved analyses of tens of thousands of patients.

Complexity of Delivery of Health Care by Nontraditional Means

The benefits of the personalized approach to health care require the interaction and collaboration of large groups of professionals and other support staff, as well as active participation by patients in managing their health care. These interactions will require new administrative and governance structures to maintain both quality and smooth operation of these complex systems. It is very likely that health care will involve the incorporation of nontraditional (i.e., nonmedical) caregivers, such behavioral scientists, personal trainers, bioinformaticians, and so on, into the care team. In addition, it will likely become impossible for a single institution to have all the types of skills needed to deliver personalized care; multi-institutional care teams create many problems in coordination and governance.

Consequences of Predicting Outcomes

One of the challenges for patients in the increasing use of genetic profiles to predict disease is to decide what to do with this information. Prediction creates many problems for patients and caregivers. For some diseases such as hereditary breast cancer, where having a genetic mutation greatly increases the risk of breast cancer, women can choose more frequent screening to detect early, more easily treatable cancer, or they can choose more radical approaches such as preventive mastectomy. The decisions to pursue preventive strategies will require input from the individuals involved as well as other experts and support persons. Other diseases such as Huntington's disease have no known preventive strategies. In these instances, individuals obtaining genetic profiles will need to specify before the screen whether or not they wish to receive information about risks of incurable diseases.

In other instances, how to deal with prediction becomes highly stressful and difficult. Individuals receiving a prediction of increased susceptibility to diabetes or heart disease can choose lifestyle modifications that can reduce the risk. However, at present, both the prediction and the lifestyle changes are subjected to significant variation. Thus, a genetic profile may indicate a 30% increase in risk of heart disease. However, development of heart disease is a complex process involving many variables, only one of which is genetic, and an individual with a high genetic risk may not get heart disease. In addition, because of the variability in factors influencing the onset of a disease, the recommended behavior modification may not prevent the disease. For these reasons, it will become increasingly necessary to provide expert support to individuals who seek to use these genetic profiles to improve their health.

Another area where prediction of outcome creates challenges for patients and caregivers is in the use of genetic profiles to predict which patients will respond to a specific course of therapy. In some of the examples described above, the predictions are highly reliable and not challenging to implement. Thus, patients with lung cancer will only receive crizotinib if their tumors contain an alteration in the ALK-1 gene. There are much data showing that the majority of patients lack this change and do not benefit from the drug (Gandhi and Janne, 2012).

A more challenging situation exists for breast cancer, where a genetic profiling kit is being used to predict the chance of recurrent metastatic breast cancer in women who initially had localized cancer (node-negative). The commercial Oncotype DX test measures the expression of 21 genes in the patient's tumor. Based on the results, patients are categorized into low, intermediate, or high risk. Several clinical trials have shown that although some women in the low or intermediate risk develop metastatic disease, there is no benefit in receiving chemotherapy as a preventive strategy (Medical Advisory Secretariat, 2010). In contrast, women in the high-risk group have 40% chance of developing metastatic cancer if they receive standard therapy, but the risk reduces to 12% if women receive extra chemotherapy in addition to standard therapy. While the test provides clear benefit to some women, it is important to note that even in the high-risk group 60% of women will do well with standard therapy and do not benefit from the additional chemotherapy.

Need New Systems for Introducing Innovations in Health Care

The magnitude of changes on the horizon for health care systems is causing great angst among the policy- and decision-makers in these systems (Downing, 2009). As described above, the move to personalized health care will require radical change in health care systems, but no one has a blueprint for the change. In fact, it is likely impossible to establish a formula for change of this magnitude, as it is unlikely that one size will fit all. Health care is very contextually sensitive and is greatly influenced by history, culture, and social pressures.

The Basque Country in Spain has developed a model, which others might emulate. The Ministry of Health in the Basque Country has established the Basque Foundation for Health Innovation and Research "to promote innovation and research in the Basque Health Service, Osakidetza, to achieve continuous development and improvement in the capacity of the service to care for the health of the people in the region." The Foundation has established several programs to foster innovation in health care by promoting pilot studies on the system to identify innovations that bring great health benefit in a cost-effective manner (e.g., Orueta et al., 2014).

Roles for Behavioral Scientists

As mentioned several times above, behavioral scientists need to help shape the evolving new health care systems that will focus on individualized therapy. In the sections below, we will describe a few areas that offer unique opportunities for behavioral scientists to contribute both to research and innovation in new health care systems.

The future will see patients taking increased responsibility for their own health care. For example, the increasing availability of commercially available diagnostic kits (glucose level, pregnancy, HIV, etc.) allows individuals to self diagnose and decide their own course of treatment. There will be increasing numbers of nontraditional (i.e., nonmedical) health care professionals, and health care will involve teams of caregivers

working collaboratively with patients to deliver increasingly complex, but more effective, interventions.

Interventions Requiring Behavior Change

Some of the greatest opportunities and challenges for behavioral scientists lie in developing new strategies for modifying behaviors for health benefits. Everyone accepts that behavior is a mixture of nature and nurture with important influences coming from genetic, environmental, psychological, and social factors. Recent work has also shown that epigenetic factors such as methylation likely play important roles in developing behavioral patterns (Champagne, 2010). A major challenge will involve dealing with the complexity, variability, and plasticity of epigenetic changes over time. Investigating these interactions will require large data sets on genetic composition of patients and family members, a social history of the patient, and extensive health records over time. One can envisage complex, multidisciplinary teams that will analyze these large data sets and optimize individual plans for patients to achieve a desired behavior change such as smoking cessation, loss of weight, or increased fitness.

For example, psychologists will shape smoking cessation programs based on the knowledge of the nicotine receptors that influence the strength of the addiction; recent studies show the strong correlation between the genetic type of nicotinic acid receptor and response to cessation therapy (Chen et al., 2012). The increased knowledge of the role of genetics in behavior is already sparking many new research topics, and these opportunities will only increase if knowledge of the links between genetics and behavior increases (Munafò et al., 2004; Schnoll et al., 2007).

This optimization of individual plans for behavior change has seen increased attention in the literature over the recent years, and has been referred to as 'tailoring,' i.e., "any combination of strategies and information intended to reach one specific person, based on characteristics that are unique to that person, related to the outcome of interest, and derived from an individual assessment" (Kreuter et al., 2013: p. 277). As opposed to targeted interventions, which are designed to be effective with all segments of the population, tailored communications are uniquely individualized to one person. Results from a *meta*-analysis including 57 studies (Noar et al., 2007) suggest that tailored health messages are effective and identify numerous factors that moderate the effects of tailoring. The strongest (print) tailored health message interventions were those that intervened on preventive or screening behaviors, generated pamphlets, newsletters, or magazines (including visual elements), utilized more than one intervention contact, recruited participants from households rather than clinics or health centers (probably due to differences in socioeconomic status), and involved psychological concepts such as attitudes, self-efficacy, stage of change, and social support.

Even though the studies included in this *meta*-analysis involved those using mainly health behavior information, it is clear that the case for tailoring health behavior interventions becomes even more convincing (and probably successful) when it includes genetic information. Nevertheless, as of now, no genotype-tailored interventions, i.e., interventions matched to an individual's specific genetic variation, have been evaluated in terms of increasing adherence to or favorable outcomes of behavior change interventions. One small study (Arkadianos et al., 2007) investigated whether the inclusion of genetic information to personalize a patient's diet (nutrigenetics) could improve long-term weight management. The results suggest that the addition of nutrigenetically tailored diets leads to better adherence, longer-term BMI reduction, and improvements in blood glucose levels. Although this study had a number of methodological limitations (e.g., nonrandomized design), it provides a first example of research that should be followed up in the future.

As this example shows, the inclusion of genetic information in personalized health interventions may increase their efficacy (and effectiveness). Nevertheless, it is also likely that the factors known to reduce adherence in general, e.g., forgetting (retrospective and prospective memory impairments), and perceptual and practical barriers (motivation and beliefs, capacity and resource limitations) will also reduce the efficacy of personalized medicine approaches. In addition to these factors, it needs to be taken into account that patients may need (personalized) treatment of comorbid mental disorders such as depression and anxiety (Prince et al., 2007), which in turn may reduce the effects of the intervention through nonadherence and other behavioral processes. It will be crucial for the success of personalized medicine interventions to develop approaches that consider these psychological factors.

The evidence concerning the use of genetic information to motivate risk-reducing behaviors has been reviewed by McBride and coauthors (McBride et al., 2010). Most research in this area has focused on the effects of information related to rare genetic variants on screening behaviors, in which genetic risk feedback has been demonstrated to be significantly associated with improved screening adherence. Much less is currently known about the effects of information on common genetic variants; however, the limited evidence available suggests that genetic information based on single-gene variants with low-risk probabilities has little impact on behavior. In the case of smoking cessation, for example, findings suggest that genetic risk feedback will not add value to existing smoking cessation programs when (1) very small increases in risk are conveyed, (2) personalized risk feedback is based on a single genetic variant, or (3) smokers have high preexisting levels of motivation to quit.

Change Management

The evolution (revolution) of health care systems will create enormous problems in change management: the recent attempts to implement general health insurance in the United States illustrate this problem (Blank, 2012). Job descriptions will change, and caregivers will have to work in new environments. Patients will have to accept new responsibilities and navigate increasingly complex health care systems. Successful evolution will likely happen best in those systems that effectively consider and plan for all the change management issues that will certainly arise. Because of the complexity and extent of the anticipated changes, change management teams will be essential for the successful introduction of new innovations.

The complexity of the anticipated innovations will also create many opportunities for research on change management. Change management is highly dependent on the context in which the change occurs. Cultural, social, and historical factors can greatly influence the ease or difficulty of implementing a new innovation. A desired outcome of research in change management will be the development of flexible strategies and processes that many can follow to facilitate change in many different situations. A unique feature of the Basque Country program is the recognition that implementation of new innovations may involve different pathways in different areas of the Country.

Creating Teams

The success of personalized medicine will depend greatly on the creation of new multidisciplinary, multi-institutional teams to develop and manage health care. Many of these teams will likely consist of disciplines that do not frequently interact and whose different approaches to problem solving will make interactions difficult. Understanding how to encourage and reward teams will become increasingly important. One challenge that many institutions already face is how to reward and recognize participation in teams. The traditional reward mechanisms tend to recognize only the leaders, and this process will not work in the future.

Risk Perception

Many of the diagnostic tests for personal medicine will give probabilities of various health outcomes. For patients and caregivers to use such information, they must have some understanding of risk and what the numbers actually mean. Scientists have studied risk perception for many years and generally found that most people either underestimate risk or overestimate actual risk. Recent studies have shown that in the United States, only 13% of the people are proficient in numeracy (O'Rourke, 2013), so perhaps it is not surprising that accurate understanding is rare.

Some experts have suggested that absolute risk is an easier concept to understand. Consider the following hypothetical situation. A patient is told that a certain medication has a 50% increased risk of heart attack over the next 5 years. Many patients would be frightened by this number and assume that 50% of people taking the new drug would experience a heart attack. The patients need additional information to interpret this risk. Assume that among 1000 people who are not taking the drug, two will have heart attacks. Among those taking the drug, three will have heart attacks, a 50% increase. The absolute risk is only one extra death per 1000 people taking the drug. This example is a simple case, but illustrates the need to express risk in terms easily understood by patients. Understanding health risks posed by genomic testing will be much more complicated because many factors in an individual profile will influence both the risk and the perception of risk. Clearly the area of risk perception will require new strategies to help individuals understand these new sources of health information.

A quantitative review of studies investigated the effects of communicating DNA-based disease risk estimates on risk-reducing behavior (Marteau et al., 2010). Mindful of the weak evidence of a total of only 14 studies of limited quality, the authors conclude that communicating DNA-based disease risk estimates has little or no effect on smoking and physical activity, but may have a small effect on self-reported diet and on intentions to change behavior.

Largely unexplored are the effects on psychological and behavioral outcomes of risk feedback based on the more realistic scenario in which genetic risk is based on numerous genetic variants. Lastly the evidence reviewed by McBride et al. (McBride et al., 2010) suggests that health literacy will affect the understanding of and engagement with genomic information in ways that diminish its potential to induce health behavior change. This underlines the need for providing expert support to individuals who seek to use these genetic profiles.

It is likely that the categorization of patients will not be deterministic in terms of prognosis. Rather, the diagnostic protocol will provide probabilities in terms of disease risk or response to treatment. In general, patients do not deal well with risk assessment in terms of changing their health behaviors or accepting treatment (e.g., adherence) as a consequence of risk information. This may be due to (nongenetic) risk information often being too general (i.e., not personalized enough) to allow perceived vulnerability to sufficiently develop and, therefore, to increase motivation to change (McBride et al., 2010). Even though the highly personalized nature of genetic information may have an advantage over other risk information for influencing health behaviors, a large body of evidence from the social and behavioral sciences suggests that the perception of threat is not sufficient to motivate behavior change (Bowen et al., 2004). It will, therefore, be important to develop tools and protocols to support patients, their social networks (e.g., families, friends), and caregivers in specific health behavior change intervention programs, and to allow them to understand the treatment options, and most importantly why specific patients may be denied a treatment (Vernon, 1999).

Conclusions

Research in medical and behavioral sciences is leading to increased personalization of treatments based on improved understanding of the genetic, biological, environmental, psychological, and social factors affecting each individual. The need for the health care systems to adopt changes represents a level of unprecedented change and could lead to disruption and dysfunction in health care provision unless a widespread spectrum of expertise helps shape new systems of health care. At present, there is little interaction between the medical and behavioral sciences, and we urge that the groups increase their interaction in order to shape a radical new paradigm of health care that incorporates both medical and behavioral elements of personalization.

See also: Behavioral Medicine; Cancer Screening; Doctor–Patient Interaction in the West: Psychosocial Aspects; Epigenetics; Health Behavior, Psychosocial Theories of; Health

Behaviors; Health Education and Health Promotion; Health Insurance: Economic and Risk Aspects; Health Psychology; Health Risk Appraisal and Optimistic Bias; Health Risk Perception; Patient Adherence to Health Care Regimens.

Bibliography

Allison, M., 2008. Is personalized medicine finally arriving? Nature Biotechnology 26 (5), 509–517. Nature Publishing Group.

Arkadianos, I., Valdes, A.M., Marinos, E., Florou, A., Gill, R.D., Grimaldi, K.A., 2007. Improved weight management using genetic information to personalize a calorie controlled diet. Nutrition Journal 6 (1), 29. BioMed Central Ltd.

Blank, R.H., 2012. Transformation of the US Healthcare System: why is change so difficult? Current Sociology 60 (4), 415–426.

Bowen, D.J., Alfano, C.M., McGregor, B.A., Andersen, M.R., 2004. The relationship between perceived risk, affect, and health behaviors. Cancer Detection and Prevention 28 (6), 409–417. Elsevier.

Brücher, B.L., Lyman, G., van Hillegersberg, R., Pollock, R.E., Lordick, F., Yang, H.-K., Ushijima, T., Yeoh, K.-G., Skricka, T., Polkowski, W., Wallner, G., Verwaal, V., Garofalo, A., Ugo, D.D., Roviello, F., Steinau, H.-U., Wallace, T.J., Daumer, M., Maihle, N., Reid III, T.J., Ducreux, M., Kitagawa, Y., Knuth, A., Zilberstein, B., Steele, S.R., Jamall, I.S., 2014. Imagine a world without cancer. BMC Cancer 14 (1), 1–8.

Bruehl, S., Apkarian, A.V., Ballantyne, J.C., Berger, A., Borsook, D., Chen, W.G., Farrar, J.T., Haythornthwaite, J.A., Horn, S.D., Iadarola, M.J., Inturrisi, C.E., Lao, L., Mackey, S., Mao, J., Sawczuk, A., Uhl, G.R., Witter, J., Woolf, C.J., Zubieta, J.-K., Lin, Y., 2013. Focus article. The Journal of Pain 14 (2), 103–113. Elsevier Ltd.

Bunting, P., Brundage, M., Goel, V., Klotz, P., Iscoe, N., Morash, C., Paszat, L., Rosser, W., Shapiro, J., 2002. Prostate-Specific Antigen (PSA) Screening in Asymptomatic Men, pp. 1–72. Available at: http://www.ices.on.ca/file/prostate-specific%20antigen%20(psa)%20screening%20in%20asymptomatic%20men.pdf (accessed 11.12.13.).

Cavers, D., Hacking, B., Erridge, S.C., Morris, P.G., Kendall, M., Murray, S.A., 2013. Adjustment and support needs of glioma patients and their relatives: serial interviews. Psycho-Oncology 22 (6), 1299–1305.

Champagne, F.A., 2010. Early adversity and developmental outcomes: interaction between genetics, epigenetics, and social experiences across the life span. Perspectives on psychological science. Perspectives on Psychological Science 5 (5), 564–574.

Chen, L.-S., Baker, T.B., Piper, M.E., Breslau, N., Cannon, D.S., Doheny, K.F., Gogarten, S.M., Johnson, E.O., Saccone, N.L., Wang, J.C., Weiss, R.B., Goate, A.M., Bierut, L.J., 2012. Interplay of genetic risk factors (CHRNA5-CHRNA3-CHRNB4) and cessation treatments in smoking cessation success. American Journal of Psychiatry. American Psychiatric Association 169 (7), 735–742.

Dancey, J.E., Bedard, P.L., Onetto, N., Hudson, T.J., 2012. The genetic basis for cancer treatment decisions. Cell 148 (3), 409–420.

Doneda, D., Netto, C.B., Moulin, C.C., Schwartz, I.V.D., 2013. Effects of imiglucerase on the growth and metabolism of Gaucher disease type I patients: a systematic review. Nutrition & Metabolism 10 (1), 34. BioMed Central Ltd.

Downing, G.J., 2009. Policy perspectives on the emerging pathways of personalized medicine. Dialogues in Clinical Neuroscience 11 (4), 377–387.

Druker, B.J., 2009. Perspectives on the development of imatinib and the future of cancer research. Nature Medicine 15 (10), 1149–1152.

Gandhi, L., Janne, P.A., 2012. Crizotinib for ALK-rearranged non-small cell lung cancer: a new targeted therapy for a new target. Clinical Cancer Research 18 (14), 3737–3742.

Hanahan, D., Weinberg, R.A., 2011. Hallmarks of cancer: the next generation. Cell 144 (5), 646–674. Elsevier Inc.

Kreuter, M.W., Farrell, D.W., Olevitch, L.R., Brennan, L.K., 2013. Tailoring Health Messages. Lawrence Erlbaum Associates, Inc, Mahwah, NJ, USA.

Marteau, T.M., French, D.P., Griffin, S.J., Prevost, A.T., Sutton, S., Watkinson, C., Attwood, S., Hollands, G.J., 2010. Effects of communicating DNA-based disease risk estimates on risk-reducing behaviours. The Cochrane Database of Systematic Reviews 10, CD007275. John Wiley & Sons, Ltd, Chichester, UK.

McBride, C.M., Koehly, L.M., Sanderson, S.C., Kaphingst, K.A., 2010. The behavioral response to personalized genetic information: will genetic risk profiles motivate individuals and families to choose more healthful behaviors? Annual Review of Public Health 31 (1), 89–103.

McLean, M., Cleland, J.A., Worrell, M., Vögele, C., 2011. 'What am I going to say here?' The experiences of doctors and nurses communicating with patients in a cancer unit. Frontiers in Psychology 2, 339.

Medical Advisory Secretariat, 2010. Gene expression profiling for guiding adjuvant chemotherapy decisions in women with early breast cancer. Ontario Health Technology Assessment Series Internet, 1–57.

Munafò, M., Clark, T., Johnstone, E., Murphy, M., Walton, R., 2004. The genetic basis for smoking behavior: a systematic review and meta-analysis. Nicotine & Tobacco Research 6 (4), 583–598.

National Cancer Institute, 2013. Targeted Cancer Therapies. Available at: http://www.cancer.gov/cancertopics/factsheet/Therapy/targeted (accessed 19.03.14.).

Noar, S.M., Benac, C.N., Harris, M.S., 2007. Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. Psychological Bulletin 133 (4), 673–693.

Orueta, J.F., García-Álvarez, A., García-Goñi, M., Paolucci, F., Nuño-Solinís, R., 2014. Prevalence and costs of multimorbidity by deprivation levels in the Basque country: a population based study using health administrative databases. PLoS ONE 9 (2), e89787.

O'Rourke, P.P., 2013. Genomic medicine: too great expectations? Clinical Pharmacology & Therapeutics 94 (2), 188–190.

Prince, M., Patel, V., Saxena, S., Maj, M., Maselko, J., Phillips, M.R., Rahman, A., 2007. No health without mental health. Lancet 370 (9590), 859–877. Elsevier.

Proctor, R.N., 2012. The history of the discovery of the cigarette-lung cancer link: evidentiary traditions, corporate denial, global toll. Tobacco Control 21 (2), 87–91.

Schnoll, R.A., Johnson, T.A., Lerman, C., 2007. Genetics and smoking behavior. Current Psychiatry Reports 9 (5), 349–357.

Shah, K.K., Cookson, R., Culyer, A.J., Littlejohns, P., 2012. NICE's social value judgements about equity in health and health care. Health Economics, Policy and Law 8 (02), 145–165.

Scottish Partnership for Palliative Care, n.d. Good Life, Good Death, Good Grief. Available at: http://www.goodlifedeathgrief.org.uk/ (accessed 19.03.14.).

Sütterlin, S., Vögele, C., Gauggel, S., 2010. Neuropsychiatric complications of efavirenz therapy: suggestions for a new research paradigm. The Journal of Neuropsychiatry and Clinical Neurosciences 22 (4), 361–369.

The Academy of Medical Sciences, 2013. Realising the Potential of Stratified Medicine, pp. 1–106.

Vernon, S.W., 1999. Risk perception and risk communication for cancer screening behaviors: a review. Monographs Journal of the National Cancer Institute 25, 101–119. Available at: http://jncimono.oxfordjournals.org/content/1999/25/101.full.pdf.

Wheeler, H.E., Maitland, M.L., Dolan, M.E., Cox, N.J., Ratain, M.J., 2013. Cancer pharmacogenomics: strategies and challenges. Nature Reviews. Genetics 14 (1), 23–34.

The World Health Organization, n.d. WHO – The Top 10 Causes of Death. WHO. World Health Organization. Available at: http://www.who.int/mediacentre/factsheets/fs310/en/index1.html (accessed 19.03.14.).

World Health Organization, 2008. WHO Report of the Global Tobacco Epidemic, 2008: The MPOWER Package. Geneva. pp. 1–342.

Personhood, Anthropology of

Nigel Rapport, University of St. Andrews, St. Andrews, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The anthropology of personhood encompasses the definition and study of three conceptual terms: person, self, and individual. It explores the identity of the individual actor and the relationship between that identity and the symbolic forms and material and moral practices of various sociocultural milieus. Recognizing the *individual* in the *role player*, the anthropologist comes to portray sociocultural milieus as composed of and constituted by a diversity of individual consciousnesses: human actors making a diversity of meaningful worlds by way of a commonality of cultural forms. Recognizing that *any* human individual may exist within any sociocultural role, the anthropologist moves from particular environments of conventional symbolic exchange toward an appreciation of personhood as an issue of moral universality.

The anthropology of personhood encompasses the definition and study of three conceptual terms: person, self, and individual. It is the exploration of the nature of the identity of the individual actor and the relationship between that identity and the symbolic forms and material and moral practices of differing sociocultural milieus.

On neither the meaning nor the implication of these conceptual terms, however, can significant disciplinary consensus be found; a number of writers have addressed definitional issues directly (Harris, 1989; Morris, 1994), but strong disagreements remain. The dissension is perhaps understandable when one considers that what is fundamentally at issue is the nature of individual consciousness and its manifestations – emergence, development, and construal – in different times and places. What is the relationship between 'nature' and 'nurture' in this context? Between consciousness as a biological given and as a sociocultural construct? Between consciousness as privately experienced and as publicly validated? Between consciousness as uniquely embodied and as collectively elicited and known?

The position taken in this article is that an anthropological study of personhood must be careful to distinguish between what may be termed 'individualism' and 'individuality': that is, between the conventional conceptualization of individual actors in particular settings ('individualism'), on the one hand, and the universal condition of our human-individual embodiment ('individuality'), on the other. Much of the conceptual difficulty the discipline has experienced arises from a conflating of these two, a confusion that goes back to Durkheim's formulation of a 'social' as against a 'psychological' science.

Social Structure and System

The Durkheimian Individual

Durkheim conceived of human beings as *homo duplex*; on one side there was the biological and personal (comprising the individual body, with its material senses and appetites) and on the other side the social and moral (the conceptual and conscientious). The individual thus led a double existence: one rooted in the physical organism and one (morally, intellectually, and spiritually superior) in a social organism of collectively and uniformly held ideas and practices. Between the two there was ongoing antagonism and tension, but through inculcation into a public language and culture, humankind was capable of rising above mean (animal) 'individuality' and becoming part of a collective consciousness and conscience in which the (sacred) traditions of a society were enshrined. If individuals were indeed conscious of themselves as individuals, then this was equally a product of their socialization in a collective consciousness; 'individualism' was a social product, like all moralities and all religions.

From Durkheim's (structuralist) descendants, a collectivist narrative was elaborated that conceptually subsumed the individual actor within sociocultural contexts. Marcel Mauss (1985) took it upon himself to detail how experience was constituted by cultural categories and thus how society exerted its force on the physiological individual. He outlined a purported evolution in individual identity that could be tied to the particular forms of social structuration. First comes the Tribal stage of *personnage*, in which individuals are conceived of as ephemeral bearers of a fixed stock of names, roles, and souls in clan possession; having no existence independent of the clan, individuals possess no inner conscience. Next comes the Classical stage of *persona*, in which individuals are conceived of as independent and autonomous citizens of a state; they are responsible, legal persons but still they possess no inner life or identity beyond the civic. With the rise of Christianity comes the stage of *personne*; conceived of as indivisible and rational, possessing a conscience, indeed, a unique sacred soul, the individual now serves as the foundation of all political, economic, and legal institutions. Finally comes the peculiar Western stage of *moi*, the individual as a 'self,' with self-interest and self-knowledge, as validated by modern institutions of psychological science.

Beginning with the assumption that sociocultural holism represents the paradigmatic form of human experience, Louis Dumont (1986) set out to plot more precisely the origination and progression of the singular Western idea of the autonomous individual as bearer of supreme moral value by comparing it with the Oriental 'archetype.' The Christian *personne*, he suggests, is reminiscent of the Hindu figure of the 'world renouncer,' someone who seeks ultimate truth by

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12128-2

forgoing the world in favor of his own individual spiritual progress. In the world renouncer, one finds a Western-like individualist for whom society has lost its absolute reality and who, despite the constraining interdependence Indian society ubiquitously imposes on its members, has become self-conscious and self-sufficient. The evolution of the Western individual and the 'substitution' of self for society as absolute value marks the triumph of a (Christian) religious ideology in which worldly renunciation becomes the collective norm. Christ's teaching that a human being possesses a soul that absolutely transcends the world of social institutions and powers (a soul that is absolutely individual in relation to God and meets others' souls only in God) engenders a community of 'outworldly' individuals who meet on Earth but have their hearts in heaven. Nonetheless, Dumont concludes, the evolution need not end there; in the wake of such essentialist movements as nationalism, multiculturalism, fascism, and Islamic fundamentalism, the cultural future of individualism is, to say the least, unpredictable.

Characterizing the above, the 'Durkheimian' line of thought, in short, is the idea that the individual actor in Western society is the result of a recent and exceptional historicocultural development. The concept of 'the individual' and its moral and social significance is absent elsewhere, with no a priori differentiation being made between individual and role, between self and society.

Versed in this paradigm, moreover, it is not surprising that in much anthropological reportage about the person, one finds a denial of individualism and little discussion of individuality. Inasmuch as 'the individual' exists, it is as a particular manifestation of 'the person'; individuals become collective constructs within the contexts of specific cultural ideologies and social practices, their psyches emanations of certain pre-given and prestructured life-worlds of socialization. It is not individual human beings who are seen to possess features and characters – such as uniqueness or self-awareness or loneliness – so much as the working parts of the complex social systems of which they are representative components and conduits.

For instance, Marilyn Strathern (1988) describes Melanesian personhood not in terms of individuals – distinct actors possessing discrete emotions, awareness, and agency – but in terms of 'dividuals,' beings constituted by properties, goods, and substances that are exchanged with others in an ongoing manner. The Gahuku-Gama of New Guinea, Kenneth Read (1955) elaborates, are conscious of themselves only as members of common categories of relationships: rather than, say, 'friendships' between unique individuals there are only relationships between socially defined positions. Similarly, Fred Myers (1986) records Australian Aboriginal 'emotions' not as pertaining to deep psychological experiences of the individual but as cultural evaluations of people and circumstances that persons must make in order to live in fluxional social settings. Rane Willerslev (2007) describes how, for the Yukaghirs of Siberia, humans, animals, and spirits are mimetic doubles of one another, with the consequence that in maintaining their identity 'individuals' also transcend difference; here are human beings who inhabit a betwixt-and-between state in which their souls are both their own and reincarnated others, human and animal. Finally, Geertz (1973, p. 360 ff) relativizes Western conceptions of the autonomous individual by emphasizing the collective role playing fundamental to the Balinese habitus of the person. Culture, Geertz suggests, translates into symbolic 'control mechanisms' by which the breadth and indeterminate qualities of actors' potential social lives are programmatically reduced to the specificity and narrowness of their actual ones.

In sum, a Durkheim-inspired anthropology of personhood has eventuated in comparative inquiries into the diversity of sociocultural conceptions of the human individual, how these conceptions connect with other sociocultural 'institutions' (family, law, exchange), and how these different ways of conceptualizing determine differences in experience. 'Personhood,' in this view, amounts to "those culturally constituted and socially conferred attributes, capacities and signs that mark a moral career and its jural entitlements in a particular society" (Poole, 1998, p. 842). 'Selfhood' refers to a human being as a locus of experience, as this experience is culturally allocated and defined – most often as relational, socially embedded, and sociocentric. The term 'Individual' encompasses an actor's uniqueness to the (rare) extent that the latter is realized in certain collectively structured life-worlds.

The Non-Durkheimian Individual

There have been exceptions to the Durkheimian line, nevertheless: ethnographies and analyses that recognize the individual to be always 'the anthropological concrete' (Auge, 1995, p. 20) from which social analysis must begin. Not only is it always individuals whom the anthropologist meets as informants (Miller, 2009), then, but "any history or description of a social situation is an exercise of psychological knowledge" (Simmel, 1971, p. 32). It is simply not good enough for social science to claim a social-factual priority (ontological, developmental, historical) of the social structural so as to assert an overwriting (causing and conditioning) of individual experience.

For instance, tracing the 'origins of English individualism,' Alan Macfarlane (1978) sets out specifically to refute those theorists who would see individualism as a recent sociocultural development – whether epiphenomenal upon religious or Renaissance or Enlightenment or capitalistic or industrialistic thinking. For "the majority of ordinary people in England from at least the thirteenth century were rampant individualists," Macfarlane writes (1978, p. 163), "highly mobile both geographically and socially, economically "rational," market-oriented and acquisitive, ego-centred in kinship and social life." In New Guinea, meanwhile, Kenelm Burridge (1979) describes how most people are 'individuals' and 'persons' in different respects and at different times, where the word *persons* may be understood to mean those who realize given sociocultural categories prescribed by traditional morality and the social order, whereas *individuals* are those who use their perception and intuition to create anew. If persons are products of current sociocultural conditions, then individuals exist in spite of them. Each 'spatially bounded organism' – or human being – is able to switch between these two modalities; Burridge concludes that individuality represents an existential imperative that pre-exists culture. This also accords with Unni Wikan's (1990) reappraisal of the Balinese, where the institution of cultural norms of

personal composure, grace, and placidity do not obviate the individual emotional effort that is recognized as being necessary to effect and sustain their appearance.

An emphasis on individual actors, their personal experience and agency, has similarly found expression in anthropological theorists of consciousness (from Gregory Bateson and Anthony Wallace onward), in the Transactionalism with which Fredrik Barth (also F.G. Bailey and Robert Paine) developed Edmund Leach's insights into instinctually 'interested' individuals, in the work of network analysts (such as John Barnes and Ulf Hannerz), and in the burgeoning of studies within interpretive or symbolic-interactionist anthropology more generally, which focus on the constructions of experience made by particular individuals in interaction. These accord more with Weberian sentiments that collectivities must be approached solely as the resultants and modes of organization of the particular acts of individual persons because it is they who remain the upper limit and the sole carriers of meaningful conduct.

Existence and Experience

Form and Meaning

Structuralist and poststructuralist discourses still denigrate anthropological sensitivity of the kind discussed immediately above as 'methodological individualism': as an erroneous couching of explanation in terms of individuals' characteristics, behaviors, and interests, such that the foundational conditions of sociocultural reality are obscured. The centerground of anthropology, it may be true to say, remains the preserve of 'methodological collectivism,' positing social phenomena as determined by factors that bypass individual rationality, hence envisaging cultural development quite independent of individual consciousness. There is a continuing insistence, in other words, that the distinction between the individual and the societal is specific to the West and must be collapsed in favor of the latter – or at least of 'social relations' – for anthropology persuasively to encounter cultural others.

On the other hand, there is a growing insistence that it is a peculiar ethnocentrism for anthropologists to fail to allow, in the 'others' we study, the personal complexity we recognize in ourselves, as both individuals and persons, role players and rebels, conventional and creative beings. The anthropology that has sought most deliberately to keep this truth in view may be termed 'existential' or 'phenomenological,' after the philosophical traditions associated with such writers as Kierkegaard, Nietzsche, Husserl, and Sartre (Jackson, 2005). It sets out with very different tenets and intentions from the 'sociological' orientation toward social structures and systems predominant above, privileging instead individual existence and experience and their moral demands.

To become human is to become individual, Geertz admits (1973, p. 52), even though we become individual in environments of sociocultural forms and patterns in whose terms we give order, direction, and point to our lives. The important question this raises is the precise relationship between sociocultural forms and the individual lives lived by them. Becoming human and individual in an environment of common sociocultural forms need not eventuate in becoming 'the same' as others or even necessarily comparable, nor in becoming according to a deterministic or directly influenced fashion, nor in becoming only once-and-for-all. For in this human becoming, energy, agency, intention, and interpretation can be seen to remain properties of embodied individual subjects. It is human beings as individual actors who are responsible for the animation, deployment, and ongoing development of systems of symbolic forms (Rapport, 1993). Forms may possess an environmental objectivity, but they come to be imbued with subjective significance and deployed in contexts whose depths of subtle distinction are likely to be more private than public.

Individualism and Individuality

An intrinsic dichotomy between individual and world is often eschewed as a peculiarity of the 'Western' sociocultural milieu, as we have heard, and hence as methodologically and analytically inapplicable. This is hardly defensible. Human beings may be socioculturally situated, but they are also, always, interpretively independent – responsible for a 'personalization' of the world – and inexorably and inevitably (by) themselves. It is not good enough simply to say that only Western culture valorizes the concept of the individual – as 'individualism' – and therefore only in Western society do individuals act distinctively with 'individuality.' For although the former might be said to describe a particular sociocultural form of behavior (the pursuit of self-distinguishment), the latter concerns a condition that is a human universal: by virtue of a unique, embodied consciousness, each of us perforce engages with the world as other and is possessed of distinct perspectives upon it.

Individualism, more precisely, pertains to a specific historicocultural conceptualization of the person that includes such notions as the ultimate value and dignity of human individuals; their moral and intellectual autonomy; their rationality and self-knowledge; their spirituality; their right to privacy, self-sovereignty, and self-development; and their voluntary contracts with a society, market, and polity. *Individuality*, by contrast, refers to that unique awareness, and awareness of awareness, that is the mark of human embodiment. It is the universal nature of existence that human beings engage with others by virtue of discrete sense-making apparatuses (nervous systems and brains); they possess discrete centers of perceptual activity in discrete bodies. Furthermore, the human body (in particular, the brain) gives rise to knowledge of the world, to a perspective upon the world, that is inherently individual. Human beings come to know themselves in the world by way of cognitions and perceptions, thoughts, feelings, and imaginings that are unique to them (Edelman, 1992).

There is an individuality intrinsic to human consciousness and also to human agency. For it is in terms of his or her perceptions that an individual comes to act, each individual becoming a discrete center of motivation and intention, of 'existential power' (Rapport, 2003). Not only is the individual's being-in-the-world universally mediated by very particular interpretative prisms that distance him or her from it, but while intrinsically 'of the world,' the individual inexorably comes to

know the world and act toward it, as 'other.' Individuals experience and interpret (and experience themselves interpreting) and therefore, they are.

An individuality of consciousness and agency is existent – an empirical truth – whatever the provenance of individualism as a cultural norm. Individuality is the human a priori, the physical–psychical basis on which all human knowledge of the world, all creativity within it, and all activity upon it, rests (including the creation and representation of sociocultural milieus). Individuality remains consequential, that is, whether or not individual consciousness is an item of collective discourse, whether or not individual reflection is publicly eschewed, and whether or not individual distinctiveness is disparaged through the institutionalization of a fund of common behaviors (concerning, for instance, notions of the person). The experience of self is essentially distinct from sociocultural representations of self (Spiro, 1993). Finally, an existential anthropology of personhood can hope to 'decolonize' (Cohen, 1994) the individual human subject, both from overdetermining cultural conditions and overweening social institutions (discourse, language-game, collective representation, social relationship, habitus, praxis) and from holistically minded social-scientific commentators. Recalling Simmel, the anthropologist offers an account both of the individual agency that brings sociocultural milieus to life alongside the common sociocultural forms and practices by which individuals coordinate their activities within these milieus (Amit and Dyck, 2006).

The interface between form and meaning is ambiguous, but recognizing the distinction is vital both epistemologically and ethically. Recognizing the individual within the role player, the anthropologist portrays sociocultural milieus as being composed of and constituted by a diversity of individual consciousnesses – human actors making a diversity of meaningful worlds by way of a commonality of cultural forms. Recognizing that *any* human individual may exist within *any* sociocultural role, the anthropologist moves from particular environments of conventional symbolic exchange toward an appreciation of personhood as an issue of moral universality (Rapport, 2012). The individual actor is warranted a space to inhabit worldviews and life-projects of his or her own making and esteeming and in distinction from the conventional conceptualizations of personhood.

See also: Authenticity, Anthropology of; Body: Anthropological Aspects; Citizenship: Political; Cognitive Anthropology; Cosmopolitanism; Cultural Critique, Anthropological; Cultural Relativism, Anthropology of; Human Rights and Social Work; Identity in Anthropology; Liberalism, Anthropology of; Psychological Anthropology; Worldview.

Bibliography

Amit, V., Dyck, N. (Eds.), 2006. Claiming Individuality. Pluto, London.
Auge, M., 1995. Non-places. Verso, London.
Burridge, K., 1979. Someone, No One. Princeton University Press, Princeton.
Cohen, A.P., 1994. Self Consciousness. Routledge, London.
Dumont, L., 1986. Essays on Individualism. University of Chicago Press, Chicago.
Edelman, G., 1992. Bright Air, Brilliant Fire. Penguin, Harmondsworth.
Geertz, 1973. The Interpretation of Cultures. Basic, New York.
Harris, G., 1989. Concepts of individual, self and person in description and analysis. American Anthropologist 91, 599–612.
Jackson, M., 2005. Existential Anthropology. Berghahn Books, Oxford.
Macfarlane, A., 1978. The Origins of English Individualism. Blackwell, Oxford.
Mauss, M., 1985. A category of the human mind: the notion of person, the notion of self. In: Carrithers, M., Collins, S., Lukes, S. (Eds.), The Category of the Person. Cambridge University Press, Cambridge.
Miller, D. (Ed.), 2009. Anthropology and the Individual. Berg, Oxford.
Morris, B., 1994. Anthropology of the Self. Pluto, London.
Myers, F., 1986. Pintupi Country, Pintupi Self. Smithsonian, Washington.
Poole, F., 1998. Socialization, enculturation and the development of personal identity. In: Ingold, T. (Ed.), Companion Encyclopaedia of Anthropology. Routledge, London.
Rapport, N., 1993. Diverse World-Views in an English Village. Edinburgh University Press, Edinburgh.
Rapport, N., 2003. I am Dynamite: An Alternative Anthropology of Power. Routledge, London.
Rapport, N., 2012. Anyone, the Cosmopolitan Subject of Anthropology. Berghahn, Oxford.
Read, K., 1955. Morality and the concept of the person among the Gahuku Gama. Oceania 25 (4), 233–282.
Simmel, G., 1971. On Individuality and Social Forms. University of Chicago Press, Chicago.
Spiro, M., 1993. Is the Western conception of the self peculiar within the context of the world cultures? Ethos 21, 107–153.
Strathern, M., 1988. The Gender of the Gift. University of California Press, Berkeley.
Wikan, U., 1990. Managing Turbulent Hearts. University of Chicago Press, Chicago.
Willerslev, R., 2007. Soul Hunters. University of California Press, Berkeley.

Personnel Selection, Psychology of

Dan Ispas, Illinois State University, Normal, IL, USA
Walter C Borman, Personnel Decisions Research Institutes, Inc. (PDRI), Tampa, FL, USA; and University of South Florida, Tampa, FL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by H. Schuler, volume 16, pp. 11343–11348, © 2001, Elsevier Ltd.

Abstract

Personnel selection is a very important component of human resource management toward creating and sustaining effective organizations. In this article, we argue that organizations can best build a successful workforce by using systematic selection procedures that employ approaches and methods from the field of industrial and organizational psychology. Thus, we describe job analysis to identify the performance requirements of jobs, as well as the knowledge, skills, and other personal characteristics necessary to be successful in each of these target jobs. This, in turn, drives the choice of assessments to evaluate applicants most likely to be effective employees, and we describe these kinds of assessments.

When an organization has a certain number of positions to fill and more applicants than positions, some type of selection process is necessary to place applicants in the jobs. Of course, this can be handled quite casually by conducting unstructured interviews and selecting people who seem most qualified for each position. Or, we can take a more systematic approach by conducting a job analysis for each of the target jobs, choosing or developing assessments to evaluate each applicant's predicted performance on one of these jobs, and then selecting an applicant who performs very well on the assessments and is thus presumably well qualified for the position. This more systematic approach typically involves an additional step in which the validity of the assessment process is evaluated to ensure the process is selecting individuals who do in fact perform effectively on the job.

The following sections discuss this more systematic approach to personnel selection. We first describe basic job analysis procedures and then one result of these procedures, the development of performance requirements or criteria. These criteria can inform the choice of predictors. Measurement of these criteria can also help in the predictor validation process to be described toward the end of this article.

Next, we present the variety of types of tests (e.g., personality, vocational interests) and other predictors (e.g., structured interviews). And finally, we discuss the validation process to evaluate the effectiveness of the predictors' identification of applicants who turn out to be high-performing employees.

Job and Work Analyses

Job analysis is a process of systematic inquiry aimed at discovering and documenting the essential nature of work (Brannick et al., 2007). Job analysis can be used for a variety of purposes including personnel selection, training, job design, and compensation. For personnel selection purposes, job analysis is concerned with identifying work behaviors and the knowledge, skills, abilities, and other characteristics (KSAOs) needed to perform those work behaviors. On the predictor side, job analysis is concerned with identifying the KSAOs and developing or using existing tests to measure them. On the criterion side, job analysis involves identifying job performance measures that tap the important performance requirements of the job and inform on KSAOs needed to perform effectively on the job.

Criterion Development and Measurement

It is now widely accepted that job performance is a multidimensional construct (Sackett and Lievens, 2008). The most commonly studied dimensions include task performance, citizenship performance, adaptive performance, and counterproductive work behaviors.

Task Performance

Task performance refers to the technical proficiency part of job performance (Borman et al., 2010). Task performance can be measured objectively or subjectively. Although objective measures of job performance look appealing, for most jobs they only tap into part of the criterion space. Objective criteria include production rates, sales, work samples, and job knowledge tests. Subjective criteria are the typical performance ratings usually provided by a supervisor. Performance ratings are the most commonly used criterion measures in industrial and organizational psychology research and practice.

At least six taxonomies examining the dimensionality of task performance have been proposed (e.g., Borman and Brush, 1993; Borman et al., 1994; Hunt, 1996). Borman et al. (2010) summarized the common content across the six taxonomies. All of the taxonomies involve content related to communication and productivity/proficiency. Personal qualities and skills are reflected in five taxonomies. Problem solving, organizing/planning, and leadership/supervision are represented in four taxonomies. Accordingly, these are some of the exemplars we expect to see representing task performance dimensions such as those just identified. Task performance is usually measured by asking supervisors to rate their subordinates on various behaviors relevant for the job. Several rating formats are

available including Behaviorally Anchored Rating Scales (Smith and Kendall, 1963), Behavior Summary Scales (Borman, 1979), Behavior Observation Scales (Latham and Wexley, 1981), and Computer Adaptive Rating Scales (Borman et al., 2001). To improve the quality of the ratings, raters can be trained using rater error training (which shows raters the common psychometric and perceptual errors commonly made) and frame-of-reference training (which familiarizes raters with the content of each performance dimension).

Citizenship Performance

In addition to task performance, employees engage in activities that support the broader work environment. These activities constitute the contextual or citizenship performance domain (Borman and Motowidlo, 1993). Since the introduction of organizational citizenship behavior (OCB) in the early 1980s (e.g., Smith et al., 1983), researchers proposed several taxonomies and Podsakoff et al. (2000) identified more than 30 unique dimensions for OCB or contextual performance. Coleman and Borman (2000) integrated the various dimensions into a three-dimensional interpretation consisting of personal support, organizational support, and conscientious initiative. A general and commonly used distinction is made between citizenship performance activities targeted at individuals and performance activities targeted at the organization. Citizenship performance can be measured using supervisor ratings; however, peer and self-ratings are also commonly used, with some authors (e.g., Allen et al., 2000) suggesting that peers are the best source for the ratings.

Adaptive Performance

Adaptive performance refers to behaviors that meet the demands of a new or changed situation or event. Pulakos et al. (2000) developed an eight-dimension taxonomy of adaptive performance (e.g., solving problems creatively, learning new tasks and procedures). Adaptive performance is distinct from task and citizenship performance (Hesketh et al., 1996). Adaptive performance is predicted by cognitive ability, self-efficacy, and achievement motivation (Chen et al., 2005; Pulakos et al., 2000, 2002). Jobs differ in their adaptability requirements (Pulakos et al., 2000) with most of the adaptive performance research being conducted on military samples.

Counterproductive Work Behaviors

Counterproductive work behaviors (CWB) are behaviors that are intended to harm the organization and its stakeholders (Spector and Fox, 2005). The most common CWB typology distinguishes between CWB targeted at the organization and CWB targeted at the individuals (Robinson and Bennett, 1995). Spector et al. (2006) proposed a more refined taxonomy consisting of five dimensions: abuse, production deviance, sabotage, theft, and withdrawal. Due to their nature (CWB are behaviors most likely performed without observers), CWB are primarily assessed using self-reports. Alternative sources for the assessment of CWB include coworker reports (Fox et al., 2007) or objective indicators retrieved from organizational records (e.g., Ilie et al., 2012).

These and other criterion measures are used to evaluate the validity of predictors in a personnel selection context. A wide variety of tests and other predictors have been used to aid in selecting employees.

Types of Predictors

Cognitive Ability

Cognitive ability is defined as a general mental capability involving reasoning, problem solving, planning, abstract thinking, complex idea comprehension, and learning from experience (Gottfredson, 1997). Probably the most comprehensive taxonomy for cognitive abilities is the three strata model derived by Carroll (1993). The first stratum consists of specific and narrow abilities, the second includes group factors and broad abilities, and the third stratum is general intelligence or g (Ones et al., 2012).

Cognitive ability is widely considered the best predictor of job performance (Schmidt and Hunter, 1998). *Meta*-analytic reviews and primary studies link cognitive ability to job performance in both United States and European countries (Ispas et al., 2010; Ones et al., 2012). The predictive validity of cognitive ability depends on the complexity of the job with the strongest validity coefficients observed for highly complex jobs (Ones et al., 2012). Cognitive ability impacts job performance through job knowledge acquisition (Borman et al., 1991, 1993); high cognitive ability individuals are better equipped to acquire the knowledge needed to perform their jobs at the highest levels. The *meta*-analytic estimate of the general cognitive ability–overall job performance relationship is 0.51.

A small number of studies examined citizenship performance or CWB as criteria. Alonso et al. (2008) *meta*-analyzed 13 studies linking cognitive ability and citizenship performance and found an uncorrected correlation of 0.05. When using CWB as the criterion, the results show validity coefficients around −0.15 to −0.20 (Ones et al., 2012). Dilchert et al. (2007) reported a corrected correlation of −0.33 between cognitive ability and objectively measured CWB.

Personality

Until the early 1990s, the general attitude in the field of industrial and organizational psychology was that personality is not an important predictor of job performance. The emergence of the Five Factor Model (FFM) of personality and the publication of several *meta*-analytic reviews (e.g., Barrick and Mount, 1991; Tett et al., 1991) led to a renewed interest in personality in personnel selection. According to the FFM, personality can be organized into five major domains: emotional stability/neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. The FFM personality domain has been linked with all of the performance domains: task, citizenship, adaptive, and CWB.

The FFM traits most consistently related to task performance across jobs are conscientiousness and emotional stability (Hurtz and Donovan, 2000; Barrick et al., 2001). Contextualized personality measures (created when scale items include 'at

work' to specify the intended context for the item rather than at home or in general, for example) show a higher predictive validity compared with noncontextualized measures (Shaffer and Postlethwaite, 2012).

For predicting citizenship performance, early *meta*-analyses showed that the FFM traits of conscientiousness and agreeableness had the strongest correlations (Borman et al., 2001). However, the most recent *meta*-analysis (Chiaburu et al., 2011) found that conscientiousness, agreeableness, and openness are the best predictors, and even the FFM traits of extraversion and emotional stability are important by showing incremental validity over conscientiousness and agreeableness. Overall, the FFM shows stronger associations with citizenship performance (multiple $R = 0.28$) than with task performance (multiple $R = 0.19$; Chiaburu et al., 2011).

For CWB, conscientiousness, agreeableness, and emotional stability emerge as the most consistent predictors (Berry et al., 2007; Dalal, 2005). Conscientiousness is more strongly related to CWB targeted at the organization (−0.34) than CWB targeted at individuals (−0.19), whereas Agreeableness shows the opposite pattern (−0.36 for individual CWB and −0.25 for organizational CWB). Emotional stability shows similar relationships with both interpersonal CWB (−0.20) and organizational CWB (−0.19).

In addition to the FFM domains, research has also examined the predictive validity of the facets that constitute the domains (Judge et al., 2013). The achievement striving facet of conscientiousness emerged as the best predictor with an estimated corrected correlation of 0.20 for task performance and 0.29 for citizenship performance.

Although the validities of personality tests have been criticized as being small (e.g., Morgeson et al., 2007), personality has incremental validity over cognitive ability (Schmidt and Hunter, 1998) and allows organizations to achieve diversity goals, that is, hiring more qualified minorities (Hough et al., 2001). Additionally, the validities of personality are larger when using all the FFM traits together (Barrick and Mount, 2012).

Several investigations also showed that the relationship between personality and job performance depends on the context. Meyer et al. (2009), in a *meta*-analytic review, found that the relationship between conscientiousness and job performance was stronger in weak situations (situations low in constraints, e.g., more autonomous). Additionally, some studies found evidence for a curvilinear effect of personality. For example, Grant (2013), in examining the relationship between extraversion and sales performance, found that the most successful employees in terms of sales were 'ambiverts' – those with average levels of extraversion.

In addition to the FFM of personality, several other traits have been linked to job performance. Proactive personality has been related to both task and citizenship performance (Greguras and Diefendorff, 2010). In some cases, non-FFM traits have shown stronger links with performance outcomes. One such example is trait anger which shows stronger links with CWB than any of the FFM traits (Herschovis et al., 2007; Ilie et al., 2012).

Although personality for selection purposes is primarily assessed using self-report inventories, research suggests that other reports (e.g., coworkers) can be more valid predictors of job performance (Connelly and Ones, 2010; Oh et al., 2011). Using other reports can alleviate some of the concerns such as applicant faking raised against the use of personality assessments for personnel selection.

Vocational Interests

Vocational interests are noncognitive constructs defined as relatively stable individual preferences for certain types of work (Harmon et al., 1994). An early *meta*-analysis (Hunter and Hunter, 1984) ranked interests second to last of 11 predictors relative to their predictive potential (only age was less predictive) with a mean validity of 0.10. However, the Hunter and Hunter (1984) *meta*-analysis was based on only three studies. Van Iddekinge et al. (2011a) conducted an updated *meta*-analysis across 74 studies and found mean validities of 0.14 against job performance, 0.26 for training performance, –0.19 for turnover intentions, and −0.15 for actual turnover. Furthermore, in a primary study, Van Iddekinge et al. (2011b) found that vocational interests show incremental validity over cognitive ability and personality for predicting job performance factors. Another line of research focused on using a different conceptualization of vocational interests based on profile coefficients. Following this approach, Nye et al. (2012) found even higher predictive validity ($r = 0.36$) between a congruence index and task performance. In a primary study, Iliescu et al. (in press) found that the congruence index shows incremental validity over personality for the prediction of CWB. Taken together, results of these studies should lead to a reassessment of the role vocational interests can play in personnel selection.

Emotional Intelligence

Emotional intelligence (EI) refers to the mental processes involved in recognizing, using, understanding, and managing one's own and others' emotional states to solve problems and regulate behavior (Mayer and Salovey, 1997; Salovey and Mayer, 1990). There are two major conceptualizations of EI: ability-based EI and trait-based EI (Lievens and Chan, 2010). EI can be measured using self-reports (where the respondents indicate their agreement with descriptive statements, similar to personality tests), and ability-based tests that ask questions about emotions and have right and wrong answers. The link between EI and job performance has been examined in two recent *meta*-analytic investigations (Joseph and Newman, 2010; O'Boyle et al., 2010). In both *meta*-analyses, EI was found to predict job performance. Furthermore, EI showed incremental validity over cognitive ability and personality (O'Boyle et al., 2010), particularly for jobs high in emotional labor, that is, requiring effort to get along with others (Joseph and Newman, 2010). More recently, a primary investigation by Iliescu et al. (2012) found that an ability-based test of EI showed incremental validity over cognitive ability and FFM for jobs where social interactions are important (such as customer service and managerial jobs) and linked CEO EI to institutional performance.

Interviews

Unlike cognitive ability, personality, interests, or EI, interviews are a method not a construct (Arthur and Villado, 2008). Interviews are considered the most commonly used method for

personnel selection in all setting (Ryan et al., 1999). Huffcutt et al. (2001) examined the constructs assessed by employment interviews. They found that 35% of the questions measured personality characteristics, 26% measured social skills, 16% mental ability, 10% knowledge and skills, 4% interests and preferences, and 3% organizational fit.

An important factor that can impact the outcomes of employment interviews is their structure. Structured interviews are more valid on average in *meta*-analyses than unstructured interviews ($r = 0.51$ vs 0.38).

Validation of Predictors

Recall that in order to demonstrate the usefulness of our predictors in a performance selection context, we often empirically validate the relationship between scores on the predictors associated with applicants or job incumbents and their criterion scores assessing their job performance. A significant correlation indicates the predictors are useful in the selection process.

Until fairly recently, the process of establishing validity was conceptualized as consisting of three approaches: content validation, criterion-related validation, and construct validation. Content validation is focused on establishing that the test items are representative of the behaviors performed on the job. Criterion validation is about establishing a link between test items and the performance domain, as we have been highlighting. Construct validation is concerned with establishing that a psychological construct has meaning for job performance.

Following the publication of the most recent Standards for Educational and Psychological Tests (1999), validity is now seen as a unitary concept that requires several sources of evidence. The first source of evidence is concerned with the relationship between the predictors and other related variables. This involves convergent validity evidence (relating the predictor with other tests measuring the same construct) and discriminant validity evidence (relating the predictor with other constructs not expected to be related to the target construct). For personnel selection purposes this refers mostly to criterion-related validity evidence. (Are test scores on the predictor related to job performance measures?) Finally, we should also examine the consequences of using tests for making decisions. For personnel selection, this refers to examining whether the use of tests results in disproportionate hiring of groups (based on age, sex, race, disability level, etc.). This is usually accomplished by assessing subgroup differences in mean predictor scores.

The Society for Industrial and Organizational Psychology published a set of Principles for the Validation and Use of Personnel Selection Procedures (2003). The purpose of the principles is to "established scientific findings and generally accepted professional practice in the field of personnel selection psychology in the choice, development, evaluation, and use of personnel selection procedures designed to measure constructs related to work behavior with a focus on the accuracy of the inferences that underlie employment decisions." The main idea from these Principles is that selection predictor measures should be related to performance on the job.

Conclusion

Personnel selection can be very important for screening applicants out of and *into* jobs. Getting people into jobs where they are qualified and can be successful is obviously desirable for both the hiring organization and the employee. Personnel selection thrives in today's organization under the umbrella of talent management. This concept recognizes that selection is a critical part of building organizational human capital capabilities, in concert with training and performance management initiatives. In sum, personnel selection has a highly important role in human resource management, and, more broadly, in enhancing organizational effectiveness.

See also: Behavior Analysis, Applied; Emotions and Work; Industrial–Organizational Psychology: Science and Practice; Job Analysis and Work Roles, Psychology of; Job Satisfaction; Organizational Citizenship Behavior; Organizational Commitment; Performance Appraisal and Evaluation; Personality and Values at Work; Program Evaluation; Work Motivation.

Bibliography

Allen, T.D., Barnard, S., Rush, M.C., Russell, J., 2000. Ratings of organizational citizenship behavior: does the source make a difference? Human Resource Management Review 10, 97–114.

Alonso, A., Viswesvaran, C., Sanchez, J.I., 2008. The mediating effects of task and contextual performance. In: Deller, J. (Ed.), Research Contributions to Personality at Work. Rainer Hampp, Munich, Germany, pp. 3–17.

American Educational Research Association/American Psychological Association, 1999. Standards for Educational and Psychological Testing. American Educational Research Association, Washington, DC.

Arthur Jr., W., Villado, A.J., 2008. The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. Journal of Applied Psychology 93, 435–442.

Barrick, M.R., Mount, M.K., 1991. The big five personality dimensions and job performance: a meta-analysis. Personnel Psychology 44, 1–26.

Barrick, M.R., Mount, M.K., 2012. Nature and use of personality in selection. In: Schmitt, N. (Ed.), The Oxford Handbook of Personnel Assessment and Selection. Oxford University Press, New York, pp. 225–251.

Barrick, M.R., Mount, M.K., Judge, T.A., 2001. Personality and performance at the beginning of the new millennium: what do we know and where do we go next? International Journal of Selection and Assessment 9, 9–30.

Berry, C.M., Ones, D.S., Sackett, P.R., 2007. Interpersonal deviance, organizational deviance, and their common correlates: a review and meta-analysis. Journal of Applied Psychology 92, 409–423.

Borman, W.C., 1979. Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology 64, 410–421.

Borman, W.C., Ackerman, L.D., Kubisiak, U.C., 1994. Development of a Performance Rating Program in Support of Department of Labor Test Validation Research (Contract Nos. 93-2 and 93-3). Cooperative Personnel Services, Sacramento, CA.

Borman, W.C., Brush, D.H., 1993. More progress toward a taxonomy of managerial performance requirements. Human Performance 6, 1–21.

Borman, W.C., Bryant, R.H., Dorio, J., 2010. The measurement of task performance as criteria in selection research. In: Farr, J.L., Tippins, N. (Eds.), Handbook of Employee Selection. Psychology Press, New York.

Borman, W.C., Buck, D.E., Hanson, M.A., Motowidlo, S.J., Stark, S., Drasgow, F., 2001. An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. Journal of Applied Psychology 86, 965–973.

Borman, W.C., Hanson, M.A., Oppler, S.H., Pulakos, E.D., White, L.A., 1993. The role of early supervisory experience in supervisor performance. Journal of Applied Psychology 78, 443–449.

Borman, W.C., Motowidlo, S.J., 1993. Expanding the criterion domain to include elements of contextual performance. In: Schmitt, N., Borman, W.C. (Eds.), Personnel Selection in Organizations. Jossey-Bass, San Francisco, CA, pp. 71–98.

Borman, W.C., White, L.A., Pulakos, E.D., Oppler, S.H., 1991. Models of supervisory job performance ratings. Journal of Applied Psychology 76, 863–872.

Brannick, M.T., Levine, E.L., Morgeson, F.P., 2007. Job and Work Analysis: Methods, Research, and Applications for Human Resource Management, second ed. Sage Publication, Thousand Oaks, CA.

Carroll, J.B., 1993. Human Cognitive Abilities. Cambridge University Press, Cambridge, United Kingdom.

Chen, G., Thomas, B.A., Wallace, J.C., 2005. A multilevel examination of the relationships among training outcomes, mediating regulatory processes, and adaptive performance. Journal of Applied Psychology 90, 827–841.

Chiaburu, D.S., Oh, I.-S., Berry, C.M., Li, N., Gardner, R.G., 2011. The five-factor model of personality traits and organizational citizenship behaviors: a meta-analysis. Journal of Applied Psychology 96, 1140–1166.

Coleman, V.I., Borman, W.C., 2000. Investigating the underlying structure of the citizenship performance domain. Human Resource Management Review 10, 25–44.

Connelly, B.S., Ones, D.S., 2010. Another perspective on personality: meta-analytic integration of observers' accuracy and predictive validity. Psychological Bulletin 136, 1092–1122.

Dalal, R.S., 2005. A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. Journal of Applied Psychology 90, 1241–1255.

Dilchert, S., Ones, D.S., Davis, R.D., Rostow, C.D., 2007. Cognitive ability predicts objectively measured counterproductive work behaviors. Journal of Applied Psychology 92, 616–627.

Fox, S., Spector, P.E., Goh, A., Bruursema, K., 2007. Does your coworker know what you're doing? Convergence of self- and peer-reports of counterproductive work behavior. International Journal of Stress Management 14, 41–60.

Gottfredson, L.S., 1997. Mainstream science on intelligence: an editorial with 52 signatories, history, and bibliography. Intelligence 24, 13–23.

Grant, A.M., 2013. Rethinking the extraverted sales ideal: the ambivert advantage. Psychological Science 24, 1024–1030.

Greguras, G.J., Diefendorff, J.M., 2010. Why does proactive personality predict employee life satisfaction and work behaviors? A field investigation of the mediating role of the self-concordance model. Personnel Psychology 63, 539–560.

Harmon, L.W., Hansen, J.C., Borgen, F.H., Hammer, A.L., 1994. Applications and Technical Guide for the Strong Interest Inventory. Consulting Psychologists Press, Palo Alto, CA.

Hershcovis, M.S., Turner, N., Barling, J., Arnold, K.A., Dupré, K.E., Inness, M., et al., 2007. Predicting workplace aggression: a meta-analysis. Journal of Applied Psychology 92, 228–238.

Hesketh, B., Allworth, E., Considine, G., 1996. Preliminary Report on Phase One of the Selection Project for the Hilton Hotel (Unpublished manuscript).

Hough, L.M., Oswald, F.L., Ployhart, R.E., 2001. Determinants, detection and amelioration of adverse impact in personnel selection procedures: issues, evidence and lessons learned. International Journal of Selection and Assessment 9, 152–194.

Huffcutt, A.I., Conway, J.M., Roth, P.L., Stone, N.J., 2001. Identification and meta-analytic assessment of psychological constructs measured in employment interviews. Journal of Applied Psychology 86, 897–913.

Hunt, S.T., 1996. Generic work behavior: an investigation into the dimensions of entry-level, hourly job performance. Personnel Psychology 49, 51–83.

Hunter, J.E., Hunter, R.F., 1984. Validity and utility of alternative predictors of job performance. Psychological Bulletin 96, 72–98.

Hurtz, G.M., Donovan, J.J., 2000. Personality and job performance. The big five revisited. Journal of Applied Psychology 85, 869–879.

Ilie, A., Penney, L., Ispas, D., Iliescu, D., 2012. The role of trait anger in the stressor-CWB relationship: convergent findings from multiple samples and methodologies. Applied Psychology: An International Review 61, 415–436.

Iliescu, D., Ilie, A., Ispas, D., Ion, A., 2012. Emotional intelligence in personnel selection: applicant reactions, criterion, and incremental validity. International Journal of Selection and Assessment 20, 347–358.

Iliescu, D., Ispas, D., Sulea, C., Ilie, A. Vocational fit and counterproductive work behaviors: a self-regulation perspective. Journal of Applied Psychology, in press.

Ispas, D., Iliescu, D., Ilie, A., Johnson, R.E., 2010. Examining the criterion related validity of the general mental ability measure for adults: a two sample investigation. International Journal of Selection and Assessment 18, 224–227.

Joseph, D.L., Newman, D.A., 2010. Emotional intelligence: an integrative meta-analysis and cascading model. Journal of Applied Psychology 95, 54–78.

Judge, T.A., Rodell, J.B., Klinger, R.L., Simon, L.S., Crawford, E.R., 2013. Hierarchical representations of the five-factor model of personality in predicting job performance: integrating three organizing frameworks with two theoretical perspectives. Journal of Applied Psychology 98, 875–925.

Latham, G.P., Wexley, K.N., 1981. Increasing Productivity through Performance Appraisal. Addison-Wesley, Reading, MA.

Lievens, F., Chan, D., 2010. Practical intelligence, emotional intelligence, and social intelligence. In: Farr, J.L., Tippins, N.T. (Eds.), Handbook of Employee Selection. Routledge, New York, pp. 339–360.

Mayer, J., Salovey, P., 1997. What is emotional intelligence? In: Salovey, P., Sluyter, D. (Eds.), Emotional Development and Emotional Intelligence: Educational Implications. Basic Books, New York, pp. 3–31.

Meyer, R.D., Dalal, R.S., Bonaccio, S., 2009. A meta-analytic investigation into situational strength as a moderator of the conscientiousness-performance relationship. Journal of Organizational Behavior 30, 1077–1102.

Morgeson, F.P., Campion, M.A., Dipboye, R.L., Hollenbeck, J.R., Murphy, K., Schmitt, N., 2007. Reconsidering the use of personality tests in personnel selection contexts. Personnel Psychology 60, 683–729.

Nye, C.D., Su, R., Rounds, J., Drasgow, 2012. Vocational interests and performance: a quantitative summary of over 60 years of research. Perspectives on Psychological Science 7, 384–403.

O'Boyle, E.H., Humphrey, R.H., Polack, J.M., Hawver, T.H., Story, P.A., 2010. The relation between emotional intelligence and job performance: a meta-analysis. Journal of Organizational Behavior 31, 788–818.

Oh, I.-S., Wang, G., Mount, M.K., 2011. Validity of observer ratings of the five-factor model of personality: a meta-analysis. Journal of Applied Psychology 96, 762–773.

Ones, D.S., Dilchert, S., Viswesvaran, C., 2012. Cognitive abilities. In: Schmitt, N. (Ed.), Oxford Handbook of Personnel Assessment and Selection. Oxford University Press, New York, pp. 179–224.

Podsakoff, P.M., MacKenzie, S.B., Paine, J.B., Bachrach, D.G., 2000. Organizational citizenship behaviors: a critical review of the theoretical and empirical literature and suggestions for future research. Journal of Management 26, 513–563.

Pulakos, E.D., Arad, S., Donovan, M.A., Plamondon, K.E., 2000. Adaptability in the workplace: development of a taxonomy of adaptive performance. Journal of Applied Psychology 85, 612–624.

Pulakos, E.D., Schmitt, N., Dorsey, D.W., Arad, S., Hedge, J.W., Borman, W.C., 2002. Predicting adaptive performance: further tests of a model of adaptability. Human Performance 15, 299–323.

Robinson, S.L., Bennett, R.J., 1995. A typology of deviant workplace behaviors: a multidimensional scaling study. Academy of Management Journal 38, 555–572.

Ryan, A.M., McFarland, L., Baron, H., Page, R., 1999. An international look at selection practices: nation and culture as explanations for variability in practice. Personnel Psychology 52, 359–391.

Sackett, P.R., Lievens, F., 2008. Personnel selection. Annual Review of Psychology 59, 1–32.

Salovey, P., Mayer, J., 1990. Emotional intelligence. Imagination, Cognition and Personality 9, 185–211.

Schmidt, F.L., Hunter, J.E., 1998. The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. Psychological Bulletin 124, 262–274.

Shaffer, J., Postlethwaite, B.E., 2012. A matter of context: a meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. Personnel Psychology 65, 445–494.

Smith, C.A., Organ, D.W., Near, J., 1983. Organizational citizenship behavior: its nature and antecedents. Journal of Applied Psychology 68, 653–663.

Smith, P.C., Kendall, L.M., 1963. Retranslation of expectations: an approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology 47, 149–155.

Society for Industrial and Organizational Psychology, 2003. Principles for the Validation and Use of Personnel Selection Procedures. Society for Industrial and Organizational Psychology, Bowling Green, OH.

Spector, P.E., Fox, S., 2005. A model of counterproductive work behavior. In: Fox, S., Spector, P.E. (Eds.), Counterproductive Workplace Behavior: Investigations of Actors and Targets. American Psychological Association, Washington, DC, pp. 151–174.

Spector, P.E., Fox, S., Penney, L.M., Bruursema, K., Goh, A., Kessler, S., 2006. The dimensionality of counterproductivity. Journal of Vocational Behavior 68, 446–460.

Tett, R.P., Jackson, D.N., Rothstein, M.G., 1991. Personality measures as predictors of job performance: a meta-analytic review. Personnel Psychology 44, 703–742.

Van Iddekinge, C.H., Roth, P.L., Putka, D.J., Lanivich, S.E., 2011. Are you interested? A meta-analysis of relations between vocational interests and employee performance and turnover. Journal of Applied Psychology 96, 1167–1194.

Van Iddekinge, C.H., Putka, D.J., Campbell, J.P., 2011. Reconsidering vocational interests for personnel selection: the validity of an interest based selection test in relation to job knowledge, job performance, and continuance intentions. Journal of Applied Psychology 96, 13–33.

Persuasion Theories

Antonis Gardikiotis, Aristotle University of Thessaloniki, Thessaloniki, Greece
William D Crano, Claremont Graduate University, Claremont, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article outlines a series of persuasion theories in social psychology. Persuasion may affect both attitudes and behaviors. The following basic approaches are considered: the learning approach, which emphasizes the process of learning attitudes; the consistency approach, which focuses on the importance of consistency between attitudes and behaviors; the attitude–behavior approach, which is concerned with the conditions under which attitudes predict behavior; and the cognitive approach, which is concerned primarily with the cognitive processes that inhibit or promote persuasion.

Persuasion is a fundamental human process that affects almost all aspects of social interaction. In interpersonal relations, mass communication, politics, economics, and even international relations, people attempt to influence the thoughts and behaviors of others. Persuasion usually is concerned with changing the attitudes and consequently (under specific conditions) the behavior of people in a context of relative freedom (for detailed reviews see Crano and Prislin, 2008; Dillard and Shen, 2012; O'Keefe, 2002; Perloff, 2013). In this article, we present various theoretical depictions of the processes implicated in persuasion. Studies of the persuasion process concern the ways people learn to appreciate social environment, maintain equilibrium among attitudes and behaviors, predict behaviors from attitudes, and how they think, more or less attentively, about messages designed to influence. The article presents a number of theories of persuasion that overtime have helped us understand the processes and factors affecting persuasion.

The Learning Approach

The learning approach advanced the idea that to be influenced, the persuasion target needed first to understand and learn the ideas conveyed by the message. Influence in this context was concerned mainly with the way attitudes were developed and formed (*see* Attitudes and Attitude Change). According to this approach, attitudes are nonobservable responses that are learned and thus amenable to reinforcement processes (Allport, 1935; Doob, 1947). These processes may involve Pavlovian classical conditioning, that is, the repeated coappearance in time and space of a neutral stimulus with another stimulus that elicits a strong response, which in turn leads to a perceived connection between the two stimuli (i.e., learning), so that the initially neutral stimulus alone comes to induce the response even in the absence of the original (unconditioned) stimulus. For example, in an exemplar demonstration of this approach, Staats and Staats (1958) repeatedly exposed participants to pairs of words consisting of (1) adjectives with an explicit evaluative tone (e.g., beautiful, kind, pleasant), that is, the *unconditioned stimuli* that produce an *unconditioned response* of liking or disliking, and (2) names of nationalities toward to whom participants presumably did not hold well-formed attitudes (e.g., Swedish, Dutch), the *conditioned stimuli*. The attitudes of participants formed after the conditioning process toward the nationalities significantly mirrored the valence of the adjectives with which the nationalities had been presented, thus eliciting the conditioned response. Similarly, Zanna et al. (1970) found that participants formed a less favorable attitude toward a previously neutrally rated word (conditioned stimuli) when this word preceded the onset of an electric shock (unconditioned stimulus). Although such research on attitude formation has attracted criticism because of its susceptibility to alternative explanations based on demand effects (i.e., participants may infer that researchers expect them to report congruent evaluations to those that were presented in the adjective nationality pairing: see Page, 1969; Fishbein and Middlestadt, 1995), it cannot easily be ruled out in studies that employed more subtle research designs. For example, Berkowitz and Knurek (1969) followed Staats and Staats' design but instead of nationalities, paired two persons' names ('Ed' and 'George') with the same evaluative adjectives. In an ostensibly unrelated experiment, participants met two confederates with the same names and rated their behavior in a way that reflected their previous conditioning (but see Brewer, 1974, for an alternative explanation).

Another proposed underlying learning process responsible for persuasion is operant conditioning. In this case, a behavior is learned because it is *reinforced*, that is, it is followed by a positive outcome or is avoided because it is *punished* (i.e., the behavior has a negative outcome). For example, Hildum and Brown (1956) reinforced their participants' positive or negative attitudes on an issue by providing positive verbal or nonverbal feedback by saying 'good' or nodding 'mm-hmm' after they reported their attitudes. Significant differences were found: negatively reinforced attitudes became less favorable than positively reinforced attitudes. Explanation based on possible demand characteristics of the design were overruled when it was shown that participants, reinforced for expressing either positive or negative attitudes, believed the experimenter was favorably disposed to the issue in question.

The most prolific programmatic research in persuasion based on learning processes is found in reports of the Yale School of Communication and Persuasion, founded by Carl Hovland, which grew from U.S. applied attitude research conducted during World War II (see Johnson and Nichols,

1998). Their research followed a conceptualization of persuasive communication based on Lasswell's (1948) scheme of "who says what, in which channel, to whom, with what effect" (p. 37). In a wide-ranging series of studies, each of these variables was manipulated while holding the rest constant, to assess the effect of each variable on persuasion. For example, the persuasive effects of the characteristics of the source ('who says') were examined (e.g., its credibility: trustworthiness or expertise), as was message characteristics ('what') such as length, structure, order of argument presentation, recipients' characteristics ('to whom') such as self-esteem or intelligence, the channel of communication ('in which channel') such as written or orally conveyed messages, and persuasive outcomes ('with what effect') such as behavior or attitude changes. It was hypothesized that the characteristics of the communication variables would affect the degree to which a recipient would attend a message, comprehend its content, reiterate its arguments, and comply with its position (Hovland et al., 1953). Although the Yale group followed a theoretically eclectic approach (McGuire, 1996), the most prominent underlying process was that of reinforcement. In this theory, attitude change depended on the association of positive or negative attributes of the communication variables with the attitude object. A logical implication of this theory is a correlation between attitude change and the retention of message arguments; however, empirical demonstration of this association was not without problems, and memory of the persuasive message proved a poor predictor of persuasion (Eagly and Chaiken, 1993). Even so, research from the Yale group significantly affected subsequent research on persuasion, although the group's results often are seen as conflicting and not easily integrated (Petty and Cacioppo, 1981).

The Consistency Approach

While the learning approach to persuasion developed within the zeitgeist of behaviorism (and the idea of the determination of behavior by external stimuli), the consistency approach drew from Gestalt theorizing and its ideas of completeness, equilibrium, and coherence. The consistency approach essentially consists of a number of theoretical accounts all based on the common underlying assumption that people prefer coherent and harmonious associations among their beliefs, attitudes, behaviors, and interpersonal relationships (Abelson et al., 1968). A prominent theory among these accounts, cognitive dissonance (e.g., Festinger, 1957; for reviews see Cooper, 2007; Harmon-Jones, 2002), focuses on the associations among a person's cognitive elements (e.g., attitudes, beliefs) and the degree to which these cognitions are consonant with each other and with attitudinally implicated behaviors. For example, a smoker who believes (a cognition) that smoking (a behavior) is unhealthy is thought to be in a state of cognitive dissonance. The individual's attitude is not consistent with the behavior of smoking. Conversely, when people recycle (a behavior) their garbage on a regular basis and believe (a cognition) that protecting the environment is beneficial for the survival of the planet, they hold consonant (i.e., consistent) cognitions. People holding dissonant cognitions, or cognitions that are inconsistent with their behaviors, experience psychological discomfort, or dissonance. The degree of discomfort depends on the importance of the dissonant cognitions (e.g., 'I really like smoking' and 'I am certain that there is a strong relationship between smoking and health problems') or on the number of the dissonant cognitions (when there are more dissonant than consonant cognitions, greater dissonance is experienced). The greater the dissonance people experience, the stronger their attempts to reduce it, as dissonance was viewed as an unpleasant state with drive-like qualities.

The experience of discomfort will motivate people to reduce dissonance (cognitive inconsistencies) and to avoid information that is likely to increase it. People will attempt to avoid dissonance either by adding consonant cognitions (e.g., 'smoking helps me focus on my work') or by minimizing dissonant cognitions (e.g., 'a 90-year-old relative of mine still smokes with no health problems'). When there is a chance for people to be exposed to dissonant information and therefore to experience the discomfort of dissonance, they become selective about what they hear or see: the selective exposure hypothesis proposes that people will selectively expose themselves to information that does not contradict their prior attitudes (Frey, 1986; for a review see Smith et al., 2008).

Festinger and Carlsmith's (1959) *forced compliance study* expanded beyond cognitive conflicts to show how inconsistencies between attitudes and behavior could lead to attitude change. In this study, participants were paid ($1 or $10) to tell others that an experiment they had just completed was interesting and enjoyable (it was not). Surprisingly and clearly inconsistent with reinforcement-based expectations, participants paid only $1 reported more positive attitudes toward the experiment than those paid $10. Festinger and Carlsmith explained these nonobvious findings by postulating that greater dissonance was generated in the low-incentive condition, presumably because those participants could not resolve the inconsistency between their lying about the task and their attitudes toward the task. The study provided a strong case on how behavior can influence attitudes (for a recent replication see by Stice et al., 2008) and also strong evidence contrary to the behaviorist hypothesis, according to which the greater incentive was expected to be associated with more positive attitudes toward the task (see Section The Learning Approach).

Inconsistencies between behaviors and attitudes also can be made salient to people (a manipulation called a 'hypocrisy induction'), and consistent with the dissonance model, which motivate them to align their attitudes and behaviors (e.g., Aitken et al., 1994). In a similar vein, people can be made to predict whether they are going to behave in a normatively acceptable way and subsequently behave in that way (called the 'self-prophecy effect,' see Spangenberg et al., 2003), thereby aligning their attitudes with actions.

The Self-Perception Approach

An alternative account of attitude change following attitude–behavior incongruence was proposed in Bem's (1972) *self-perception theory*. According to this account, attitude change does not happen because of the psychological discomfort related with the inconsistency but by the process of self-observation.

People observe their own behavior and infer their attitudes as they do when they observe other people. This is especially true when there are no other cues in the external environment to account for the behavior. Self-perception theory is a purely cognitive approach, and thus could not predict findings based on the arousal that theoretically was generated in response to dissonance. In a series of important and critical studies, Cooper et al. (1978) and Zanna and Cooper (1974) established the arousing properties of dissonance, and thus rendered implausible the basic assumption of Bem's model. Self-perception theory does account for behaviors resulting from attitude–behavior discrepancies, but only when the attitudes are weak or the behaviors either inconsequential or only slightly discrepant from the attitudes (see Taylor, 1975; Visser and Cooper, 2003).

Attitudes and Behavior

Ultimately, persuasion is about influencing another's behaviors, and this focus is seen in many studies, including those concerned with the purchase of a product, a vote for a political candidate, engaging in safe sex, or driving without drinking. It is therefore important to assess the factors that lead to attitude–behavior consistency. Attitudes can predict or cause a person's actions, and considerable research has focused on the conditions that facilitate or retard the consistency between attitudes and behaviors.

The *theory of reasoned action* of Fishbein and Ajzen (1975) focuses on voluntary actions (e.g., voting for a candidate, or buying a product) and proposes that the best predictor of voluntary action is the intention to perform that behavior. Intention is predicted by the attitudes toward the specific behavior (e.g., 'to vote for the specific candidate is desired because I believe he or she will better people's conditions') and by the subjective norms that concern this behavior, that is, what other people, whose opinions are valued, think about it (e.g., 'my family and friends expect me to vote for this candidate'). The model further details the factors that determine the personal (attitudes) and social (norms) factors. Attitudes are determined by the beliefs that concern the behavior (e.g., 'the candidate publicly supports pro-environmental policies,' which I approve of) and the expectancies associated with these beliefs (e.g., 'the candidate will be able to implement these policies'). Norms are determined by the beliefs that concern what significant others think about the behavior (e.g., 'my family wants me to vote for the candidate') and the degree to which the person is motivated to comply with significant others' desires (e.g., 'I would like to follow what my family wants me to do on this issue').

The *theory of planned behavior* (Ajzen, 1991) extends the theory of reasoned action by including, in the predictive model, the belief concerning the degree to which a person consciously controls the critical behavior; that is, whether people think it easy or difficult for them to perform the specific behavior. For example, people may have positive attitudes (e.g., 'quitting smoking is good') and supportive norms (e.g., 'my family would like me to quit smoking') toward a behavior but the important belief about the control over the behavior (e.g., 'I do not think I am able to quit smoking') may determine the predictive power of attitudes and norms on the intention. A number of reviews and quantitative meta-analyses of empirical studies have supported both theories (see Albarracin et al., 2001; Armitage and Conner, 2001; Hardeman et al., 2002).

The Cognitive Approach to Persuasion

The metatheoretical dominance of cognitive approaches across many subfields of psychology during the last decades has been evident in persuasion research as well. In one of the first attempts to bridge learning and cognitive principles in persuasion, Greenwald (1968) pointed to the importance of thinking and elaborating the persuasive message to attitude change. He suggested that persuasion was mediated by "the content of attitude-relevant cognitions elicited (and thus rehearsed and learned) during the persuasion situation" (p. 167). In this view, then, attitude change depends on the valence of thoughts generated when considering a persuasion message. The more favorable the thoughts developed in response to the message, the greater the attitude change in the direction of the message. An important methodological development that was instrumental to the progress of the cognitive approach was the thought-listing technique (see Petty and Cacioppo, 1986) in which participants list the thoughts they had as they were exposed to a message, which are later analyzed for message favorability. This focus on the cognitive responses to the message sparked a new approach to persuasion that dominates today's understanding of persuasion processes.

The persuasion approach proposed that the recipient of a persuasive communication would not necessarily work to draw a logical conclusion regarding the issue presented in the message. Depending on a number of variables of the communication context, recipients will sometimes follow cognitively effortful ways of processing the message, but at other times will follow cognitive shortcuts, which require considerably less cognitive effort to reach an evaluation of the attitude object. Two models that dominated the cognitive approach are the elaboration likelihood model (ELM; Petty and Cacioppo, 1981, 1986; Petty and Wegener, 1998), and the heuristic-systematic model (HSM; Chaiken, 1980; Chaiken et al., 1989; Chen and Chaiken, 1999). Both of these dual-process models propose two qualitatively different cognitive routes to attitude change, one characterized by effortful elaboration of a persuasive message, and the other characterized by near-effortless elaboration.

Elaboration Likelihood Model

In the ELM, the effortful cognitive pathway is called the *central route*. It involves a thoughtful processing of the message, where recipients attend its content and try to evaluate it in terms of their prior knowledge and attitudes. The less effortful path is called the *peripheral route*. It involves a consideration of various aspects of the influence setting (such as the characteristics of the source), with considerably less analysis of the content of the message. Message elaboration within this framework is

perceived as a continuum, a variable from high to low depending on the route recipients follow, which depends on two variables, the receiver's *ability* and *motivation* to process the message. When both ability and motivation are high, that is, when people are able to understand and elaborate, and at the same time are motivated to do so, they are likely to follow the central route of processing and devote significant cognitive resources in responding to the message. Ability is influenced by such factors as people's prior knowledge (the more recipients know about the content of a message, the more they are able to elaborate it) or possible distractions in the influence context, such as noise (the greater the distraction, the less able are recipients to engage in issue-relevant thinking) or cognitive business (*see* Attitudes and Attitude Change).

Motivation also is affected by a number of variables such as personal involvement with the issue at hand, defined as the degree to which recipients perceive relevance between the topic of the message and themselves personally. The model posits that personal involvement is an important variable for issue-relevant thinking. The greater the involvement, the greater the message elaboration. Another factor that affects message processing is recipients' *need for cognition*, which is defined as the degree to which individuals engage in and enjoy thinking. People who enjoy thinking hard are more motivated, and more likely to engage in central processing.

A central assumption of the model is that the effect of message elaboration on attitude change depends on the valence of thinking produced. When favorable thoughts toward the argument exceed unfavorable thoughts, then attitude change toward the advocated position is expected. Valence of thoughts may depend on recipients' prior attitudes, with those already holding positive attitudes toward the attitude object inclined to produce more favorable thoughts as compared to those holding less positive prior attitudes. The qualities of the persuasive arguments, however, are of crucial importance in predicting the outcome of a persuasive attempt. The ELM makes a distinction between strong arguments (characterized by good evidence, use of logic, and powerful reasoning) and weak arguments (e.g., poorly reasoned, weak or ambivalent evidence), with the former leading to more favorable thoughts, assuming that recipients follow the central route. In the empirical examination of the model, and to assess whether a variable of the communication context (e.g., characteristics of the source or the message) instigate central or peripheral processing, the variable is crossed with argumentation quality and is examined at which level of the credibility variable a difference between strong and weak arguments is noticed. Differences in responses to strong and weak arguments, with the former leading to more favorable thinking and greater attitude change, are taken as indicative of central processing. The distinction between strong and weak arguments is central in the model, but it has attracted criticism because of its lack of solid theoretical basis and its reliance solely on empirical pretesting of the arguments (see Eagly and Chaiken, 1993; Stiff, 1986).

When elaboration is low and recipients of an influential communication do not have the ability or the motivation to process a message centrally, the model suggests that peripheral processing will become dominant. Although not a singular type of processing, but rather a host of different kinds of processing modes, peripheral processing is characterized mainly by lack of elaboration of the content of the message. One kind of peripheral processing (others include classical and instrumental conditioning, attributional thinking, etc.) is the use of cognitive heuristics, established rules of thumb used to provide a shortcut to decision-making ("Do I agree with this message or not?"). Such heuristics often concern the source of the influential message: for example, when the source is perceived as expert and trustworthy, recipients who are unlikely to elaborate a message are likely to agree and comply without paying much attention to message's arguments. A similar pattern of results is evident when the source is likable; again at the lower end of elaboration continuum, more likable sources are more persuasive than less likable ones. The consensus heuristic concerns whether other people support the message's position as well. Again at the low end of elaboration continuum, people will agree with a position supported by the majority as compared to one supported by a minority.

The two routes to persuasion are not only different in terms of the quality of thinking connected with them, but also the consequences they have on changed attitudes compared to the attitudes that result after peripheral processing. Attitudes that result after central processing are stronger, more enduring in time, more resistant to counter-persuasion, and more predictive of behavior.

An important characteristic of the ELM is that many variables within the communication context can affect message elaboration and resultant persuasion, and in many different ways (see O'Keefe, 2002). They can influence elaboration, serve as arguments, or as peripheral cues. For example, source credibility may be considered a peripheral cue, but it also can serve as an argument that increases persuasion (Petty et al., 1999).

Another assumption of the ELM is that variables not only can affect the degree of elaboration, and thus the underlying route to persuasion, but also introduce a central route processing bias. This bias may motivate the recipient to process the advocated position so that it can be either verified or falsified. This confirming or disconfirming mode of processing can be affected by variables such as prior knowledge or forewarning. In the former case, recipients elaborate knowledge-consistent information more deeply and produce favorable cognitions; in the latter case, recipients are motivated to counter-argue the message and produce unfavorable thoughts.

The ELM has provided a comprehensive theoretical framework of persuasion phenomena and instigated a prolific program of research. It has attracted criticism, however, on the difficulties of the assessment of elaboration independent of its effects and of the prediction of the specific role a variable can play. Moreover, predicting when central and peripheral processing can co-occur is not without problems, as the conditions under which this can happen are not clearly specified (see Bohner and Schwarz, 2001).

Heuristic and Systematic Processing

The *heuristic-systematic processing model*, similar to the ELM, proposes two different processing modes, one effortful, characterized by issue-relevant thinking called *systematic* and one effortless, called *heuristic* (Bohner et al., 1995; Chaiken et al., 1989; Chen and Chaiken, 1999). Like ELM, HSM describes

message elaboration as a continuum affected by recipients' abilities and motivations. However, HSM is differentiated from the ELM in how it details the effortless mode of processing, the interaction between the processing modes, and the motivational underpinnings of processing modes.

Systematic processing involves recipients' active evaluation of message content and judgment formation or attitude change utilizing prior knowledge (Chaiken et al., 1989). Heuristic processing, a less effortful mode of processing that does not involve strong attention to the content of the message, is defined narrowly as involving the application of heuristics, that is, established cognitive schemata that govern evaluation and decision-making in a communication context (e.g., 'An expert source is right most of the time and thus should be trusted'). This definition is more specific than ELM's peripheral processing, which encompasses a number of different processes (e.g., classical conditioning and so on).

The ways that systematic and heuristic modes interact in the persuasion process are also methodically specified in the HSM (Bohner and Schwarz, 2001). When message receivers are unable or unmotivated to process a message, heuristic processing will predominate, as this processing mode requires minimal effort and cognitive resources. However, at the higher end of the elaboration continuum, systematic and heuristic processing may operate simultaneously, either independently or interactively. Bohner et al. (1995) have developed an empirically supported explication of these singular or interactive processes, which they termed the *co-occurrence hypotheses*. The first two hypotheses concern the case in which both processing modes have independent effects on attitudes: In the *additivity hypothesis* case, heuristic and systematic processing may independently affect attitudes. This is evident only when the two processes do not contradict each other (e.g., when an expert provides very weak arguments). However, according to the *attenuation hypothesis*, because systematic processing "often provides the individual with more, and subjectively more relevant, information any additional effects of heuristic processing may be suppressed" (p. 423).

Bohner et al.'s next hypotheses concern cases in which the two modes of processing interact: The *bias hypothesis* holds that if a message contains both strong and weak arguments, or is of ambiguous valence, it is possible that initial heuristic processing will bias systematic processing so it is assimilated to the valence of the heuristic cue. For example, Chaiken and Maheswaran (1994) manipulated message strength and source credibility and showed that participants highly motivated (and able) to elaborate the message, who received a message of both strong and weak arguments assimilated their attitudes toward the heuristic cue of source credibility. However, when they received messages containing only strong or weak arguments, the former led to greater persuasion than the latter irrespective of source credibility. Hence, the effect of systematic processing suppressed the effects of heuristics, as the attenuation hypothesis would assume. Finally, according to the *contrast hypothesis*, violation of expectancies based on heuristic processing can lead to contrasting systematic processing. More negative thinking is expected when positive expectancies are violated, and when negative expectancies are violated, more positive thinking is expected. This bias was observed in a study by Bohner et al. (2002), where participants rated a message with weak arguments provided by experts less favorably than the same message provided by nonexperts.

The extent of processing motivation can largely be accounted for by the HSM's *sufficiency principle*, which suggests that people attempt to attain sufficient confidence in their judgments and attitudes. Sufficiency is determined by two concepts, the sufficiency threshold (ST: the desired confidence) and actual confidence (AC). When AC is lower than desired, individuals are motivated to systematically process the message. The greater the difference between ST and AC, the greater systematic processing is expected. The less difference between ST and AC, the more is heuristic processing expected. Desired confidence may increase depending on personal relevance and involvement and AC may decrease depending on the difference in the valence of heuristic cues and content (Maheswaran and Chaiken, 1991).

Finally, the HSM describes multiple motives that underlie the extent of processing; these involve *accuracy*, *ego defense*, and *impression motivation*. Individuals may try to attain attitudes that are valid reflections of reality (accuracy-motivated processing is assumed to be more open-minded); they also may try to defend self-defining beliefs and values, and thus may adopt attitudes to look good and be accepted by others (both defense- and impression-motivated processing are assumed as more biased). Differences between desired and AC determine the strategy that will be followed (Bohner et al., 2008).

The Unimodel

A challenge to the dual-process conceptualization of persuasion that has dominated the field during the last decades was advanced by Kruglanski and Thompson (1999). They proposed the *unimodel*, a single process alternative to persuasion accounts. According to this theory, effortful (central or systematic) and effortless (peripheral or heuristic) processing modes incorporate functionally equivalent types of evidence. Evidence is defined as any type of information that can be perceived as pertinent to the judgments an individual will make. The unimodel does not accept the distinction between cues and arguments that has been instrumental to the dual-process models; rather, it suggests that all information relative to the source (e.g., credibility), the message (e.g., quality of argumentation), or the context (e.g., channel modality) can be considered evidence that an individual can take into account to reach a conclusion about an issue. The model has been characterized on the basis of its emphasis on the relevance of evidence to the conclusions drawn by the recipient, as a functional theory of persuasion (see Johnson et al., 2005) that relates the goals of receiver with the persuasive information. It has motivated interesting and confirmatory research, but has not attracted as much attention as the dual-process approaches.

Concluding Remarks

Persuasion theories have provided a wealth of information detailing when and how people are persuaded. The theoretical approaches presented here have focused on different aspects of

persuasion process that have contributed to a greater or lesser degree to our understanding of persuasion. The contemporary dual-process models of persuasion address the importance of ability and motivation to elaboration, and the valence of recipients' thinking about persuasive messages. They have integrated and accounted for a significant number of previous persuasion phenomena and guided contemporary research in the field. However, the challenges these models face are both theoretical (e.g., the explanatory power of the unimodel remains to be seen) and applied regarding significant social domains (e.g., improvement of health behaviors or reduction of prejudice).

See also: Attitudes and Attitude Change; Attitudes and Behavior; Attitude Measurement; Cognitive Dissonance; Heuristics in Social Cognition; Implicit Social Cognition; Social Cognition.

Bibliography

Abelson, R.P., Aronson, E., McGuire, W.J., Newcomb, T.M., Rosenberg, M.J., Tannenbaum, P.H. (Eds.), 1968. Theories of Cognitive Consistency: A Sourcebook. Rand McNally, Chicago.

Aitken, C.K., McMahon, T.A., Wearing, A.J., Finlayson, B.L., 1994. Residential water use: predicting and reducing consumption. Journal Applied Social Psychology 24, 136–158.

Ajzen, I., 1991. The theory of planned behavior. Organizational Behavior and Human Decision Processes 50, 179–211.

Albarracin, D., Johnson, B.T., Fishbein, M., Muellerleile, P.A., 2001. Theories of reasoned action and planned behavior as models of condom use: a meta-analysis. Psychological Bulletin 127, 142–161.

Allport, G.W., 1935. Attitudes. In: Murchison, C. (Ed.), Handbook of Social Psychology. Clark University Press, Worcester, pp. 798–844.

Armitage, C.J., Christian, J. (Eds.), 2003. On the theory of planned behaviour. Current Psychology, Special Issue 22, 187–280.

Armitage, J.C., Conner, M., 2001. Efficacy of the theory of planned behaviour: a meta-analytic review. British Journal of Social Psychology 40, 471–499.

Bem, D.J., 1972. Self-perception theory. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology. Academic Press, San Diego, pp. 1–62.

Berkowitz, L., Knurek, D.A., 1969. Label-mediated hostility generalization. Journal of Personality and Social Psychology 13, 200–206.

Bohner, G., Erb, H.-P., Siebler, F., 2008. Information processing approaches to persuasion: integrating assumptions from the dual- and single-processing perspectives. In: Crano, W.B., Prislin, R. (Eds.), Attitudes and Attitude Change. Psychology Press, New York, pp. 161–188.

Bohner, G., Moskowitz, G., Chaiken, S., 1995. The interplay of heuristic and systematic processing of social information. European Review of Social Psychology 6, 33–68.

Bohner, G., Ruder, M., Erb, H.-P., 2002. When expertise backfires: contrast versus assimilation in the biased processing of persuasive messages. British Journal of Social Psychology 41, 495–519.

Bohner, G., Schwarz, N., 2001. Attitudes, persuasion, and behavior. In: Tesser, A., Schwarz, N. (Eds.), Blackwell Handbook of Social Psychology. Blackwell, Oxford, pp. 413–435.

Brewer, W.F., 1974. There is no convincing evidence for operant or classical conditioning in adult humans. In: Weimer, B., Palermo, David S. (Eds.), Cognition and the Symbolic Processes. LEA, Hillsdale, pp. 1–42.

Chaiken, S., 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. Journal of Personality and Social Psychology 39, 752–766.

Chaiken, S., Liberman, A., Eagly, A.H., 1989. Heuristic and systematic processing within and beyond the persuasion context. In: Uleman, J.S., Bargh, J.A. (Eds.), Unintended Thought. Guilford, New York, pp. 212–252.

Chaiken, S., Maheswaran, D., 1994. Heuristic processing can bias systematic processing: effects of source credibility, argument ambiguity, and task importance on attitude judgment. Journal of Personality and Social Psychology 66, 460–473.

Chen, S., Chaiken, S., 1999. The heuristic-systematic model in its broader context. In: Chaiken, S., Trope, Y. (Eds.), Dual-process Theories in Social Psychology. Guilford, London, pp. 73–96.

Cooper, J., 2007. Cognitive Dissonance: 50 Years of a Classic Theory. Sage, London.

Cooper, J., Zanna, M.P., Taves, P.A., 1978. Arousal as a necessary condition for attitude change following induced compliance. Journal of Personality and Social Psychology 36, 1101–1106.

Crano, W., Prislin, R., 2008. Attitudes and Attitude Change. Psychology Press, New York.

Dillard, J.P., Shen, L. (Eds.), 2012. The Sage Handbook of Persuasion: Developments in Theory and Practice. Sage, Thousand Oaks, CA.

Doob, L.W., 1947. The behavior of attitudes. Psychological Review 54, 135–156.

Eagly, A.H., Chaiken, S., 1993. The Psychology of Attitudes. Harcourt Brace, Orlando, FL.

Festinger, L., 1957. A Theory of Cognitive Dissonance. Stanford University Press, Stanford, CA.

Festinger, L., Carlsmith, J.M., 1959. Cognitive consequences of forced compliance. Journal of Abnormal and Social Psychology 58, 203–210.

Fiske, S., Taylor, S., 1994. Social Cognition, second ed. Addison-Wesley, Reading, MA.

Fishbein, M., Ajzen, I., 1975. Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research. Addison-Wesley, Reading, MA.

Fishbein, M., Middlestadt, S., 1995. Noncognitive effects on attitude formation and change: fact or artifact? Journal of Consumer Psychology 4, 181–202.

Frey, D., 1986. Recent research on selective exposure to information. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology. Academic Press, New York, pp. 41–80.

Greenwald, A.G., 1968. Cognitive learning, cognitive response to persuasion, and attitude change. In: Greenwald, A.G., Brock, T.C., Ostrom, T.M. (Eds.), Psychological Foundations of Attitudes. Academic Press, New York, pp. 147–170.

Hardeman, W., Johnson, M., Johnston, D.W., Bonetti, D., Wareham, N.J., Kinmonth, A.L., 2002. Application of the theory of planned behaviour in behaviour change interventions: a systematic review. Psychology and Health 17, 123–158.

Harmon-Jones, E., 2002. A cognitive dissonance theory perspective on persuasion. In: Dillard, J.P., Pfau, M. (Eds.), The Persuasion Handbook: Developments in Theory and Practice. Sage, Thousand Oaks, CA, pp. 99–116.

Hildum, D.C., Brown, R.W., 1956. Verbal reinforcement and interviewer bias. Journal of Abnormal and Social Psychology 53, 108–111.

Hovland, C.I., Janis, I.L., Kelley, H.H., 1953. Communication and Persuasion: Psychological Studies of Opinion Change. Yale University Press, New Haven, CT.

Johnson, B.T., Maio, G.R., Smith-McLallen, A., 2005. Communication and attitude change: causes, processes, and effects. In: Albarracín, D., Johnson, B.T., Zanna, M.P. (Eds.), The Handbook of Attitudes. Erlbaum, Mahwah, NJ, pp. 617–669.

Johnson, B.T., Nichols, D.R., 1998. Social psychologists' expertise in the public interest: civilian morale research during world war II. Journal of Social Issues 54, 53–77.

Kruglanski, A., Thompson, E., 1999. Persuasion by a single route: a view from the unimodel. Psychological Inquiry 10, 83–109.

Lasswell, H.D., 1948. The structure and function of communication in society. In: Bryson, L. (Ed.), The Communication of Ideas: Religion and Civilization Series. Harper Row, New York, pp. 37–51.

Maheswaran, D., Chaiken, S., 1991. Promoting systematic processing in low motivation settings: the effect of incongruent information on processing and judgment. Journal of Personality and Social Psychology 61, 13–25.

McGuire, W.J., 1996. The Yale communication and attitude-change program in the 1950s. In: Dennis, E.E., Wartella, E. (Eds.), American Communication Research—The Remembered History. LEA, Mahwah, NJ, pp. 39–59.

O'Keefe, D.J., 2002. Persuasion: Theory and Research, second ed. Sage, Thousand Oaks, CA.

Page, M.M., 1969. Social psychology of a classical conditioning of attitudes experiment. Journal of Personality and Social Psychology 11, 177–186.

Perloff, R.M., 2013. The Dynamics of Persuasion, fifth ed. Lawrence Erlbaum, Mahwah, NJ.

Petty, R.E., Cacioppo, J.T., 1981. Attitudes and Persuasion: Classic and Contemporary Approaches. Brown, Dubuque, IA.

Petty, R.E., Cacioppo, J.T., 1986. The elaboration likelihood model of persuasion. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology. Academic Press, San Diego, CA, pp. 123–205.

Petty, R.E., Wegener, D.T., 1998. Attitude change: multiple roles for persuasion variables. In: Gilbert, D.T., Fiske, S.T., Lindzey, G. (Eds.), The Handbook of Social Psychology, fourth ed. McGraw-Hill, New York, pp. 323–390.

Petty, R.E., Wheeler, S.C., Bizer, G.Y., 1999. Is there one persuasion process or more? Lumping versus splitting in attitude change theories. Psychological Inquiry 10, 156–163.

Smith, S.M., Fabrigar, L.R., Norris, M.E., 2008. Reflecting on six decades of selective exposure research: progress, challenges, and opportunities. Social and Personality Psychology Compass 2, 464–493.

Spangenberg, E.R., Sprott, D.E., Grohmann, B., Smith, R.J., 2003. Mass-communicated prediction requests: practical application and a cognitive dissonance explanation for self-prophesy. Journal of Marketing 67, 47–62.

Staats, A.W., Staats, C.K., 1958. Attitudes established by classical conditioning. Journal of Abnormal and Social Psychology 57, 37–40.

Stice, E., Shaw, H., Becker, C.B., Rohde, P., 2008. Dissonance-based interventions for the prevention of eating disorders: using persuasion principles to promote health. Prevention Science 9, 114–128.

Stiff, J.B., 1986. Cognitive processing of persuasive message cues: a meta-analytic review of the effects of supporting information on attitudes. Communication Monographs 53, 75–89.

Taylor, S.E., 1975. On inferring one's attitudes from one's behavior: some delimiting conditions. Journal of Personality and Social Psychology 31, 126–131.

Visser, P., Cooper, J., 2003. Attitude change. In: Hogg, M.A., Cooper, J. (Eds.), Sage Handbook of Social Psychology. Sage, Cambridge, pp. 211–231.

Zanna, M.P., Cooper, J., 1974. Dissonance and the pill: an attribution approach to studying the arousal properties of dissonance. Journal of Personality and Social Psychology 29, 703–709.

Zanna, M.P., Kiesler, C.A., Pilkonis, P.A., 1970. Positive and negative attitudinal effect established by classical conditioning. Journal of Personality and Social Psychology 14, 321–328.

Perversions, Sexual (Paraphilias)

Oyedeji Ayonrinde, Bethlem Royal Hospital, South London and Maudsley NHS Foundation Trust, Beckenham, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.E. Goodman, volume 21, pp. 14002–14007, © 2001, Elsevier Ltd.

Abstract

Sex, sexuality, and human intimacy are universal and strongly influenced by individual and societal factors. Cultures determine abnormality and normality across the range of human behaviors. Historical, religious, legislative, and technological shifts have also influenced the classification, acceptance, attitudes, and distribution of sexual practices.

Paraphilic disorders pose unique challenges to social scientists, clinicians, and the criminal justice system. The private nature of the majority of practices and their legality make study particularly challenging and fraught with limitations. The dearth of knowledge of these disorders is an invitation to researchers of all fields relating to sexual practice.

Introduction

Sex, sexuality, and human intimacy are universal phenomena that are strongly influenced by personal orientations and fantasies. Attitudes toward sex, sexuality, orientation, and procreation are also strongly influenced by societies, their history, and cultures. These identities influence our latent and manifest world views, and invariably how we think and how our attitudes develop. Societies and cultures determine deviancy and normality across the range of human behaviors. This relationship involves a complex interplay of factors, defining norms of acceptable behavior and permissible variants.

Previously known as sexual deviance, perversions, or aberrations the range of abnormal sexual behaviors is now referred to as paraphilias.

The word 'paraphilia' that is derived from Greek means beyond or irregular (*para*), a love of (*philia*), and has less pejorative implications than previously used names.

Brief History

Sexual Orientation and Practices Are Steeped in History

Antiquity

The ancient Greeks often ascribed different forms of sexual behavior to their gods. Zeus was portrayed as having relationships with animal or inanimate objects forms, such as a cloud or a shower of gold, and his objects of desire were of both sexes. On the other hand, Hercules cross-dressed, while Theseus and Achilles also wore women's clothes at times (Licht, 1969; Bullough, 1976). Hippocratic Greeks gave advice on sexual practice and the vitality of life; for example, "sexual intercourse should be more frequent … in winter … and for older men more than the younger."

In the medieval West, the perceived wisdom was that men and women shared the same physiology, the male being perfect version and the female flawed. Female genital organs were believed to be inverted and inferior versions of male genitalia. Menstruation was felt to be the leaking of vital body fluids which men discharged as semen.

History observations and record found paraphilias were present in diverse cultures and across time periods. Sometimes linked with religious practice, rites of passage and cults, a variety of sexual practices flourished. Archaeologists have also found cave paintings and figurines depicting sexual activity from several 1000 years ago.

The more recent development of sexology led to the development of diagnostic and classificatory systems as asylums, jails, and hospitals were faced with managing a range of erotic dispositions and behaviors found unacceptable if not abhorrent to the different societies.

Classifications

Different labels have been applied to sexual behaviors over the centuries. However, it was not till the eighteenth century that attempts to classify and label started in the West. Categories varied with each culture considering, approving or disapproving, behaviors based on the prevailing sexual climate. Classifications offer an insight into the thinking of historical periods.

The Marquis de Sade (1740–1814), while imprisoned in the Bastille, wrote detailed descriptions and short stories of sexual behaviors, proposing a classification a century before his successors, Krafft-Ebing, Havelock Ellis, and Freud – his discourses about sex and morals, *120 Days of Sodom*, being the most memorable work on perversion of the time. De Sade requested people not to mock those with deviant impulses, but to look upon them as one might a cripple, with pity and understanding. De Sade proposed the following classification for sexuality (Gorer, 1963):

(a) People with weak or repressed sexual desires
(b) Natural perverts (born so)
(c) Libertines – people who imitate group (b), but wilfully rather than naturally.

Sade also described sexual acts involving different body fluids, sperm, blood, urine, saliva, and feces.

In the nineteenth century, Darwin in his theory of evolution influenced the thinking of Ernst Haeckel, who suggested that sexual perversions reflected a return to an earlier developmental stage that had become fixated, so that perverts, like other races and women, were at a lower evolutionary point.

Succeeding generations of physicians, sociologists, and therapists continued debating over alternative classifications of perversions. Krafft-Ebing for instance proposed life was

a struggle between instinct and civilization, and mental disorders created no new instincts, but altered existing ones. Like many others he felt excessive masturbation triggered sexual perversion. Krafft-Ebing's *Psychopathia Sexualis* highlighted his classification as below:

Group 1: Deviation of the sex impulse. Too little sexual feeling (impotence, frigidity); too much (satyriasis and nymphomania), and sexual feeling appearing at the wrong time (in childhood or old age) or being directed wrongly (at children, the elderly, or animals).

Group 2: Sexual release through other forms of activity (inflicting or receiving pain or sexual bondage).

Group 3: Inverted sexuality (homosexuality, bisexuality, and transvestism).

Krafft-Ebing's works had many revisions over the 12 editions. Krafft-Ebing encouraged better understanding of 'perverts' referring to them as 'stepchildren of nature.' His belief that most perversions were mainly congenital was challenged by Alfred Binet who argued that fetishism involved, the fetish taking different forms, for example, an obsession with the color of the eyes, bodily odors, clothing, or different inanimate objects. He opined heredity could not determine this choice, and it was more likely a childhood phenomenon. However, only few developed perversions after childhood events. Binet concluded there may well be an innate morbid condition in some individuals.

Freud on the other hand theorized that developmental stages from infancy to adulthood replicated the moral development of humans, from sexual promiscuity, to perversion, and finally heterosexual monogamy. The healthy infant negotiating the Oedipal complex and castration anxiety (a narrative of incestual love) was felt to develop neurosis if unsuccessful with transitions. Freud believed perversions were a consequence of arrested development from trauma culminating in stagnation in sexual pathways. Successful negotiation of childhood difficulties occurred in some due to constitutional factors and moral evolution. Freud classified as deviants individuals with sexual interest other than normal vaginal intercourse. His works are rich in phallic objects and sexual reference.

Current Classification

The modern era has seen a shift in classification of sexual practice. Following centuries of being associated with illness, loss of vital force, and a cause of insanity masturbation is no longer considered a deviant act. Similarly, homosexuality, which had previously been classified a disorder, was removed as a paraphilia in the 1974 edition of the *Diagnostic and Statistical Manual of Mental Disorders*.

The Diagnostic and Statistical Manual of the American Psychiatric Association (DSM-5) defines a paraphilia as "Any intense and persistent sexual interest other than sexual interest in genital stimulation or preparatory fondling with phenotypically normal, physically mature, consenting human partners." Hence, any sexual interest outside normophilic sexual interests. This article explores the range of normophilia and disorders outside its boundaries.

Paraphilic Disorder

A paraphilia that is currently causing distress or impairment to the individual or a paraphilia whose satisfaction has entailed personal harm or risk of harm to others is described as paraphilic disorder.

Preferential sexual interests affect sexual activity and can manifest in erotic activities as well as erotic targets. Paraphilias often come to the attention of mental health professionals and social care in the context of legal complications. Where consenting adults in private situations perform these acts they are unlikely to have clinical contact. In light of this, significant distress or unlawful nature of some activities may trigger referral to legal or medical interventions.

Given the large range and diversity of paraphilic disorders a comprehensive list is beyond the scope of this volume. Table 1 illustrates a number of commoner disorders.

The DSM-5 highlights "knowledge of and appropriate consideration for normative aspects of sexual behaviour are important factors to explore to establish a clinical diagnosis and to distinguish a clinical diagnosis from socially acceptable behaviour."

The role of sexual activity is unavoidably fundamental to the preservation of humanity. However, non-procreational sexual activity is influenced by factors such as the availability of sexual partners, opportunity, individual preferences, and fantasies. These individual factors are set against a backdrop of culture, norms, taboos, religion, mores, and values of the society.

Behaviors that may externally appear similar may have different explanations. For instance, a man may dress in female clothes for a fancy dress party, for sexual pleasure and gratification or part of a therapeutic gender reassignment process. Superficially, each situation involves cross-dressing in a man however the contexts are markedly different. Bhugra and de Silva described the transitional role of uniforms as symbols traversing fashion, fantasy, and fetish.

In some situations (e.g., prisons), absence of heterosexual opportunities may limit heterosexual individuals to engaging in homosexual activity. This sexual behavior may not necessarily reflect individual sexual preference or fantasies.

Furthermore, society may influence perception of the number of partners and acceptable frequency of sexual activity (sexual addiction). In essence, sexual activity should always be considered in context.

While paraphilias are universal phenomena the prevalence, distribution, and legal status may vary in individual societies.

This article explores the nature of paraphilias and the controversies associated with less common sexual behaviors.

Nature and Nurture of Sexuality

Gender roles vary in different societies. While men and women have equal rights and opportunities in some, men dominate in others and rarely do women control the administration of a community. Hofstede described the dimensions of 'masculinity–femininity' in human groups. In high masculinity societies women have restricted opportunities in vocation or societal hierarchy. There are higher degrees of individualism, financial reward, and limited support for the

Table 1 Specific paraphilic disorders

Paraphilic disorder	*Description*	*Prevalence (from DSM-5)*
Voyeuristic disorder	Urge or practice of observing an unsuspecting person who is naked, undressing, or engaging in sexual activity, or in activity deemed to be private.	Most common of potentially law-breaking paraphilias. Population prevalence unknown. Highest lifetime prevalence 12% males; 4% females.
Exhibitionistic disorder	Exposing one's genitals to an unsuspecting person or performing sexual acts that can be observed by others.	Prevalence unknown. Highest population prevalence 2–4% male. Lower in females.
Frotteuristic disorder	Touching or rubbing against a non-consenting person.	Frotteuristic acts in up to 30% adult males in general population.
Sexual masochism disorder	A desire to be humiliated, beaten, bound, or made to suffer for sexual pleasure.	Prevalence unknown. 2.2% male; 1.3% female reported in past 12 months (Australia).
Sexual sadism disorder	Sexual pleasure through the infliction of pain or humiliation on another person.	Prevalence unknown. 2–30% reported.
Paedophilic disorder	Sexual preference for prepubescent children.	Population prevalence unknown. Highest possible male prevalence approximately 3–5%; female lower unknown %.
Fetishistic disorder	Use of inanimate objects to gain sexual excitement.	Unknown.
Transvestic disorder	Arousal from clothing associated with members of the opposite sex.	Unknown but rare.
Other specified paraphilic disorder	Other paraphilias include the use for sexual purposes of corpses (necrophilia), animals (zoophilia), feces (coprophilia), enemas (klismaphilia), and urine (urophilia).	Unknown.

weaker in these groups. Conversely, feminine societies are relationship oriented and focus less on emotional or social differentiation. Individualistic societies with less restricted or hierarchal structures may have a greater expression of paraphilias. Such expression may not indicate the true prevalence however.

Sexual practices may follow set cultural scripts that both promote and discourage different types of sexual behavior (Reiss).

Sociocultural scripts of sexual behavior may be inborn biological (nature) or may be acquired (nurture and environmental). Cultural transmission of sexual roles and behaviors may be age, gender, or position specific through folklore, song, dance, or depiction in artwork. Societies may limit or promote behavior through rites of passage into manhood – for example, affirmation of a female partner's virginity with blood-stained sheets after the consummation of marriage.

Segall et al. (1999) argue that the role of nature (biology) indicating sexual attraction and arousal are biological processes, involving cognitions, nerve transmissions, and physiological changes before sexual activity. These important biological mechanisms play a key role in the process and manifestation of normal and paraphilic behaviors in all societies. Evidence supports the role of biological, cultural, and individual factors in the development of paraphilias. The weight given to each component remains uncertain and will be explored further in this article.

Sexual Behaviors and Attitudes

Standards of what constitute the normal range of sexual attraction, arousal experiences, and behaviors are difficult to define. For instance, this may range from body morphology, skin tone, eye color, hair, to fragrance.

Bullough proposed a classification of societies and cultures as either 'sex positive' or 'sex negative.' Sex-negative cultures see semen loss and even intercourse as a flaw and sexual temperance is encouraged. Sex-positive cultures on the other hand permit the pleasurable, gratifying and non-procreative aspects of sex.

History

Sexual attitudes and values have changed over centuries. Early Greeks and Romans recognized and tolerated alternative sexual behaviors with shifting attitudes toward sexuality. Religious doctrines and belief systems were also promoted by different scriptures and teachings. Over millennia, procreation has been essential to sustaining society and activities such as labor (farming) and military workforce. Non-procreative loss of semen was associated with loss of potent vital energy and the reproductive capability of the individual as well as the society.

Religion

Early Christianity generally disapproved of non-procreative sex. In addition, it disapproved of sex outside marriage and this consequently became a pillar of the Ten Commandments. Over time, competing and parallel religious systems led to different interpretations of the Abrahamic doctrines. However, marriage did not automatically confer the right to sexual intimacy and was seen more as a relationship duty rather than a source of pleasure.

Islamin contrast appeared slightly more accommodating, with the historical appreciation of the sexual traditions of pre-Islamic Arabs. Women were sought for wedlock and not debauchery; however, a contradiction to this was that women were seen as lesser and subservient to men.

Hinduism, as a religion and philosophy, was seen as sex positive as vividly illustrated in the carvings of sex acts in temples and Kama Sutra. The Kama Sutra as numerous erotic

Hindu classics covers numerous features of courtship and sex acts. While superficially sexually liberal, there are sexual restrictions in the faith. For instance, adultery and intercaste relationships were discouraged. However, transsexualism and transvestitism which would otherwise be termed deviance, were tolerated. As societies have evolved, changed, and imbibed new customs, the values and attitudes regarding sexual activity and sexuality have changed. Religion has been a potent vehicle of these attitudes and practice.

Sexual values and attitudes have changed as societies evolve and imbibe new customs. Religion has been a powerful transmitter of attitudes and practice across regions of the world: for example, Pentecostal churches in Africa. Historical record and artistic portrayal of brothels, sex acts, homosexual prostitutes, pornographic media, and the trade in sex toys between nineteenth and twenty-first century England highlight shifting societal trends. On the whole, acceptance of female sexuality is limited and in some parts of the world, deviation from proscribed roles may culminate in death for example Taliban Afghanistan where even female education is deemed immoral.

Prevalence of Paraphilias (see Table 1)

There is a dearth of empirical data on the prevalence and distribution of paraphilias. Investigation into the prevalence of paraphilias and other sexual behaviors face numerous challenges such as embarrassment to acknowledge behaviors and the illegal status of some sexual behaviors even in the most liberal societies. Furthermore, it is difficult to ethically and objectively verify study results. Paraphilias in clinical settings are more likely to be in the context of legal issues than voluntary referral. Research is limited in less permissive (mainly non-Western) societies.

Given the challenges associated with researching paraphilias, proxy measures have been used in some case. Exploring the relative prevalence of some paraphilias, Dietz and Evans monitored the use of pornographic magazines. It was hypothesized that choice of pornography was likely to reflect underlying fantasies. The most common imagery found was bondage and domination. Magazines often presented women wearing lingerie and high heels. A limitation of studies like this is the variation in distribution of pornography and inability to generalize or extrapolate findings to actual behaviors. The growth and variety of pornographic material on the Internet would limit such research currently.

In Sweden, Langstrom and colleagues, in a survey of sexuality and health among 2450 randomly selected 18–60 year olds asked questions regarding paraphilias and problem sexual behaviors: 3.1% of respondents reported sexual arousal by exposing their genitals to a stranger on at least one occasion and 7.7% at least one incident of arousal spying on others having intercourse. There were associations with being male, psychological difficulties, drug and alcohol use, and lower life satisfaction. Voyeurism was nearly three times commoner in men and they were twice as likely to expose than women. A relationship between early childhood experience such as separation from parents and some paraphilias was also observed.

Langstrom and Zucker observed 0.4% of women and about 2.8% men acknowledged at least one episode of transvestic fetishism. These individuals were more likely to have had childhood separation from parents, abuse, same-sex sexual experiences, high-frequency masturbation and sexual arousal with the use of pornography. They also reported higher rates of voyeurism and exhibitionism.

Further review of the sample found an association between hypersexuality and paraphilias. Men in this category were more likely to be young with early sexual activity, experienced separation from parents, and reside in major cities. They were more likely to be involved in same-sex behavior, sadomasochism, voyeurism, exhibitionism, and fund sex. Paradoxically, they were less satisfied with their sex lives and more involved in risk-taking behaviors such as substance misuse, gambling, and alcohol use. These findings suggest some behavioral and social associations with paraphilias which may tenuously contribute to or complement paraphilias.

Gender

Developmentally, boys explore genital pleasure more readily and perhaps younger than girls. It appears societies are more tolerant or permissive of male expression of sexuality while restricting females. The developing girl may become aware of genitalia, menses, and sexuality with embarrassment unlike many of their male counterparts.

The current body of evidence shows a male prevalence in the majority of paraphilias. It may reflect men being more actively in sexual intimacy than women. Men have more frequent sexual thoughts with a desire for physical pleasure while women are more inclined to engage in sexual intimacy when there is an emotional bond.

Paedophilic Disorder

In Ancient Greece and Rome, adult–child sex was permitted and acceptable practice. Greco-Roman art depicting nubile children and adults in poses were common features of the social culture. However, there are also historical accounts of child safety concerns in some societies (e.g., Green).

The developmental stage or age of consent for sex is mainly influenced by the rules of the land, culture, and religious factors as is the case with marriage. There may be variation that may vary for heterosexual and homosexual activity.

In Western societies, paedophilia is socially and legally undesirable with significant criminal justice and clinical resource invested in its management. The global web of the Internet has made this more difficult with the absence of geographical and legislative boundaries as is the case with the 'dark net.'

A number of social deficits have been reported in some paraphilias. A number of paedophiles have for instance been found to have inadequate social skills, while Langevin found them shy, socially withdrawn, passive, and lacking assertiveness. Many paedophiles experienced disturbed childhoods with for example sexual abuse, inadequate parenting, and family dysfunction.

Most research into paedophilia has been in offender populations. A survey by Wilson and Cox observed 71% of male paedophiles reported an attraction to boys aged 12–14 years

and 12% a preference for girls aged 8–10 years with sexual fantasies in about half. While Seto, Cantor, and Blanchard propose child pornography may be a valid diagnostic indicator of paedophilia, this may actually relate to arousal and fantasy rather than actual behaviors. It is very difficult to extrapolate these findings to the general population.

Fetishistic Disorder

Kaplan has written extensively about fetishes from a psychoanalytic perspective suggesting the behaviors are underpinned by the need to transform something unfamiliar and intangible to something familiar and tangible as a feature of the fetishism strategy. The etymological and non-sexual use of the word is related to inanimate items such as carvings, feathers, or artificial material. Substances such as leather, latex, and rubber have also been linked with fetishism in different societies.

The sexualization of these objects may facilitate dealing with internal ambiguous anxieties.

Kaplan suggests that "fetishism permits certain details into the foreground of experience … thereby masking or disguising other features and casting them into the background shadows and margins" (page 6). The purpose of the fetish is masked and the object of fetish becomes more prominent and dramatic.

The Fetish Foot and Shoes

The historical foot binding of Chinese females has been proposed as a cultural fetish (Kaplan). Beginning in the twenty-first century BC, the practice was initially restricted to upper classes – the ruler's mistresses, concubines and dancing girls in specific regions of the country. Over time, foot binding spread across China. Male control in such a manner can be considered fetishsitic control over the object. The modern-day feet continue to draw attention with sexualization of high heel shoes which draw attention to the legs. The role of feet and shoes in sexual fantasy shows variation, for instance, rural, desert, or snow-capped communities with a more practical emphasis on footwear are unlikely to sexualize high heel shoes.

Procreation or Pleasure

Hypothetically, sociocentric cultures that promote sex for reproduction are less likely to develop sexual fetishes, and if so lower rates than egocentric cultures with a greater emphasis on the pleasure of sex.

Increasing globalization, with rural–urban and international migration, may lead to increasing experimentation and changing expectations from sexual partners round the world.

Cultural Behavior or Paraphilia?

Some culturally acceptable behavior may be perceived as paraphilic in other societies. In Papua and New Guinea for instance, the Sambian boys perform fellatio on older tribesmen as a right of passage, swallowing their semen for growth and strength. After about a decade in homosexual relations, the younger males develop a fear of being polluted by women. Once at the stage for heterosexual marriages they discontinue the earlier homosexual practice. These behaviors are limited to tribal rituals and not for pleasure but initiation rituals to assist with the development of masculinity.

Culture modulated reversal from homosexual to heterosexual practice illustrates how behavior considered or attributed paraphilic can be normalized and subsequently reversed according to societal norms.

Technological and digital innovations with the Internet, phones, and camera devices have altered patterns of voyeuristic behavior and proximity is no longer required to intimately observe others.

Similarly, exhibitionism occasionally takes advantage of these technologies over the Internet or televised events. 'Flashing' whereby an individual exposes genitals or breasts in public events is not commonly described in non-Western societies. Nudist colonies on the other hand promote non-sexual total exposure as a way of life.

There has been a growth in industry of sex aids and 'toys' over the years such as vibrators. These items substitute intimacy in some individuals while used as an erotic adjunct in some relationships.

Others achieve sexual fulfillment with telephone 'sex-lines,' 'cybersex,' or virtual reality sexual scenarios.

Frotteurism, a disorder commoner in crowded locations such as congested trains and busy gatherings is likely to be higher in urban than isolated rural settings. In such settings the frotteur and their intentions are not easily identified. It stands to reason that distribution of this disorder is uneven.

Sexual Masochism

Sexual masochism is the deliberate or voluntary participation in an activity that involves being humiliated, beaten, bound, or otherwise abused to experience sexual excitement. Frequently associated with masochism, sadism involves the infliction of physical or mental suffering (pain, humiliation, or terror) on the sexual partner to stimulate sexual excitement and orgasm. Sadistic fantasies linked with complete control of the other are particularly dangerous and may be fatal. Masochistic behaviors that involve self-asphyxiation as part of a masturbatory ritual can also result in accidental death.

The Biology of Sexual Behaviors

Sexual processes and pathways are complex and involve the interplay of biological and environmental factors. So far, social and environmental factors have been explored. Intrauterine development of the human brain is somewhat modulated by the action of the male sex hormone, testosterone. Its presence is key to male development and female development in its absence.

Berridge and Kringlebach present biological and physiological theories. Stating that brain processes of emotion, motivation, and affect are influenced during development with active mechanisms of the brain responding to reward stimuli. It

is suggested that the rewards consist of liking (a conscious cognitive process) and wanting (the motivation for reward through incentive salience) in addition to learning processes. Through this process, different stimuli and experiences produce different degrees of pleasure or displeasure as is the case for preferences in fragrances or body morphology. Noxious stimuli in sexual activity on the other hand are complexly interwoven with emotions.

Behavioral and cognitive ingredients of pleasure may be individually or socially determined. Personal pleasure (whether normal or disordered) may relate to social pleasure or alternative sensory stimuli. In paraphilic disorders, behavioral and cognitive processes seek to enhance pleasure with individual, social, or combined rewards.

Neuroanatomy, Neurochemistry, and Neurophysiology

Various brain structures have been found to mediate reward. For instance, the experience of pleasure has been associated with the orbito-frontol cortex with the support of other subcortical structures and neurochemical transmission such as dopamine. The hypothalamus is largely concerned with the hormonal control of reproduction. Sexual fantasy however requires the cognitive processes of the cerebral cortex that processes complex information, knowledge, and emotion associated with sex. Furthermore, following head injury or the use of psychoactive substances, paraphilic behaviors may manifest in some individuals. This may be associated with the ensuing loss of inhibitory control or personality change.

Frontal and temporal lobe EEG differences have been observed among sex offenders with paraphilias, however they lack specificity.

The neurosciences offer exciting opportunities through neurobiology, neurochemistry, and neuroimaging of normal and abnormal sexual activity. In addition, advances in genetics and epigenetics may help us understand the nature of human sexuality.

Oxytocin, a neurochemical has been found to have a biological modulating influence on emotions, of pleasure and love sensations. The role of oxytocin in the treatment of satiety and emotional dysregulation is being investigated.

Hypothalamic and mesolimbic dopamine pathways have been suggested by Pfaus et al. with endogenous opioid activation in the brain as a basis of sexual reward.

They hypothesize that a 'love map' of components, feelings, movements, and interactions associated with sexual reward occur during a critical period of early sexual experience.

There is also evidence of the impact of neurodevelopmental difficulties and head injuries on the presentation of paraphilias, for example, learning disability and specific regional head injuries for instance frontal lobe with impaired executive function and personality change. Childhood head injury has been reported by Blanchard et al. (2006). Blanchard et al. (2006) also suggested that some deviant sexual behaviors result from pathological or pathogenic developmental variables, which may impede normal sexual development.

In conclusion, emotion, reward, and pleasure appear to be programmed to specific stimuli through experience and exposure. This programming involves certain brain structures as a neurodevelopmental process or in later life (Rahman, 2008). Paraphilias may emerge from these processes with disruption of normal psychosexual social communication and neural development.

Paraphilic disorders pose unique challenges to social scientists, clinicians, and the criminal justice system. Their classification has been molded by history, religion, culture, legislation, and social attitudes. The private nature of the majority of practices, associated embarrassment and the unlawful situations that bring them to attention make studies particularly challenging. Most studies are among offender groups and population prevalence is fraught with limitations. Proxy measures of interests or fantasies (e.g., pornography) may not apply to real-life situations. Paraphilias are also adaptive to technologies and globalization. Our limited knowledge of these disorders is an invitation to researchers of all fields to fill the large gaps in our understanding of sexual practice.

See also: AIDS: Acquired Immune-Deficiency Syndrome; Anxiety and Anxiety Disorders; Culture-Specific Psychiatric Disorders; Family and Psychiatric Illness: Family Therapy.

Bibliography

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. (DSM-5).

Baldwin, J.D., Baldwin, J.I., February 1989. The socialization of homosexuality and heterosexuality in a non-Western society. Archives of Sexual Behavior 18 (1), 13–29.

Baldwin, J.D., Baldwin, J.I., April 1997. Gender differences in sexual interest. Archives of Sexual Behavior 26 (2), 181–210.

Berridge, K.C., Kringelbach, M.L., 2008. Affective neuroscience of pleasure: reward in humans and animals. Psychopharmacology 199 (3), 457–480.

Blanchard, R., Kuban, M.E., Blak, T., Cantor, J.M., Klassen, P., Dickey, R., 2006. Phallometric comparison of pedophilic interest in nonadmitting sexual offenders against stepdaughters, biological daughters, other biologically related girls and unrelated girls. Sexual Abuse: Journal of Research and Treatment 18, 1–14.

Bhugra, D., De Silva, P., 1996. Uniforms: fact, fashion, fantasy or fetish. Sexual and Marital Therapy 11, 393–406.

Bhugra, D., Popelyuk, D., McMullen, Isabel, 2010. Paraphilias across cultures: contexts and controversies. The Journal of Sex Research 47 (2–3).

Bullough, V.L., 1976. Sexual Variance in Society and History. University of Chicago Press, Chicago.

Carroll, J.L., Volk, K., Hyde, J., 1985. Differences between males and females in motives for engaging in sexual intercourse. Archives of Sexual Behavior 14, 131–139.

Casanova, M.H., Mannheim, G., Kruesi, M., 2002. Hippocampal pathology in two mentally ill paraphiliacs. Psychiatry Research Neuroimaging 115, 79–89.

Dietz, P.E., Evans, B., 1982. Pornographic imagery and prevalence of paraphilia. American Journal of Psychology 139, 1493–1495.

Emmers-Sommer, T.M., Allen, M., Bourhis, J., Sahlstein, E., Laskowski, K., et al., 2004. A meta-analysis of the relationship between social skills and sexual offenders. Communication Reports 17 (1), 1–10.

Flor-Henry, P., Lang, R.A., Koles, Z.J., Frenzel, R.R., 1991. Quantitative EEG studies of paedophilia. International Journal of Psychophysiology 10, 253–258.

Gammon, L., Makinen, M., 1994. Female Fetishism: A New Look. Lawrence & Wishart, London.

Garnett, E.S., Nahmias, C., Wortzman, G., Langevin, R., Dickey, R., 1988. Positron emission tomography and sexual arousal in a sadist and two controls. Sexual Abuse: A Journal of Research and Treatment 1 (3), 387–399.

Goodman, R.E., 1997. Understanding human sexuality: specifically homosexuality and the paraphilias in terms of chaos theory and fetal development. Medical Hypotheses 48 (3), 237–243.

Goodman, R.E., 1998. The paraphilias: an evolutionary and developmental perspective. In: Freeman, H., Pullen, I., Stein, G., Wilkinson, G. (Eds.), Seminars in Psychosexual Disorders. Gaskell, London, pp. 142–155.

Gorer, G., 1963. The Life and Ideas of the Marquis de Sade. Peter Owen, London.

Green, R., 2002. Is pedophilia a mental disorder? Journal of Archives of Sexual Behavior 31, 467–471.

Hofstede, G., 1984. In: Abridged (Ed.), Culture's Consequences: International Differences in Work-related Values. Sage, Beverly Hills.

Hendricks, S., Fitzpatrick, D., Hartmann, K., Qauife, M.A., Stratbucker, R.A., Graber, B., 1988. Brain structure and function in sexual molesters of children and adolescents. Journal of Clinical Psychiatry 49, 108–112.

Hucker, S., Langevin, R., Dickey, R., Handy, L., Chambers, J., Wright, S., 1988. Cerebral damage and dysfunction in sexually aggressive men. Annals of Sex Research 1 (1), 32–48.

Hucker, S., Langevin, R., Wortzman, G., Bain, J., Handy, L., Chambers, J., Wright, S., 1986. Neuropsychological impairment in pedophiles. Canadian Journal of Behavioural Science 18 (4), 440–448.

Herdt, G.H. (Ed.), 1984. Ritualized Homosexuality in Melanesia. University of California Press, Berkerly.

Kaplan, L.J., 2006. Cultures of Fetishism. Palgrave Macmillan, New York.

Kirenskaya-Berus, A.V., Tkachenko, A.A., 2003. Characteristic features of EEG spectral characteristics in persons with deviant sexual behaviour. Human Physiology 29, 273–287.

Kringlebach, M.L., 2005. The human orbitofrontal cortex: linking reward to hedonic experience. Nature reviews Neuroscience 6, 691–702.

Kringlebach, M.L., 2008. The hedonic brain: a functional neuroanatomy of human pleasure. In: Kringlebach, M.L., Berridge, K.C. (Eds.), Pleasure of the Brain. Oxford University Press, Oxford, England, pp. 202–221.

Langstrom, N., Zucker, K.J., 2005. Transvestic fetishism in the general population: prevalence and correlates. Journal of Sex and Marital Therapy 31, 87–95.

Langstrom, N., Hanson, R., 2006. High rates of sexual behavior in general population; correlates and predictors. Archives of Sexual Behavior 35, 37–52.

Langstrom, N., Seto, M.C., 2006. Exhibitionistic and voyeuristic behavior in a Swedish national population survey. Archives of Sexual Behavior 35, 427–435.

Laumann, E.O., Gagnon, J.H., Michael, R.T., Michaels, S., 1994. The Social Organisation of Sexuality: Sexual Practices in the United States. University of Chicago Press, Chicago.

Licht, H., 1969. Sexual Life in Ancient Greece. Panther Books, London.

McAnulty, R.D., 2006. PaedophiliaMcAnulty, R.d., Burnette, M.M. (Eds.), 2006. Sex and Sexuality, vol. 3. Praeger, Westport, CT, pp. 81–96.

Ping, W., 2000. Aching for Beauty: Foot Binding in China. University of Minnesota Press, Minneapolis.

Porter, R., 1999. The Greatest Benefit to Mankind – A Medical History of Humanity from Antiquity to the Present. Fontana Press, London.

Pfaus, J.G., Kippin, T.E., Coria-Avila, G.A., Gelez, H., Afonso, V.M., Ismail, N., Parada, M., 2012. Who, what, where, when (and maybe even why)? How the experience of sexual reward connects sexual desire, preference, and performance. Archives of Sexual Behavior 41 (1), 31–62.

Rahman, Q., Symeonides, D.J., 2008. Neurodevelopmental correlates of paraphilic sexual interests in men. Archives of Sexual Behavior 37, 166–172.

Segall, M.H., Dasen, P.R., Berry, J.W., Poortinga, Y.H., 1999. Human Behavior in Global Perspective, second ed. Allyn & Bacon, Boston.

Seto, M.C., Cantor, J.M., Blanchard, R., 2006. Child pornography offenses are a valid diagnostic indicator of pedophilia. Journal of Abnormal Psychology 115, 610–615.

Sprecher, S., 1989. Premarital sexual standards for different categories of individuals. Journal of Sea Research 26, 232–248.

Sprecher, S., McKinney, K., 1993. Sexuality. Sage, Newbury Park, CA.

Stoller, R.J., Herdt, G., 1985. Theories of origin of male homosexuality: a cross-cultural look. Archives of General Psychiatry 42, 399–404.

Swaab, D.F., 2007. Sexual differentiation of the brain and behavior. Best Practice & Research: Clinical Endocrinology & Metabolism 21, 431–444.

Wudarczyk, O.A., Earp, B.D., Guastella, A., Savulescu, J., September 2013. Could intranasal oxytocin be used to enhance relationships? Research imperatives, clinical policy, and ethical considerations. Current Opinion Psychiatry 26 (5), 474–484.

Pestalozzi, Johann Heinrich (1746–1827)

Petra Korte, Dülmen, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11349–11353, © 2001, Elsevier Ltd.

Abstract

The Swiss Johann Heinrich Pestalozzi (1746–1826) was one of the most distinctive writers of his time, and was already in his own lifetime regarded as one of the most famous pedagogs in Europe. He owed his reputation as a writer essentially to his best-selling novel *Lienhard und Gertrud* whose theme was the story of the education and cultivation of an entire village. He achieved his fame as a pedagog through his work in Stans (1799), Burgdorf (1800–03), and Yverdon (1804–25) where for 21 years he directed a teaching and educational institute in which children from all over Europe were taught. Pestalozzi's work deals with a wealth of subjects in the fields of general politics, social policy, constitutional law, jurisprudence, literature, anthropology, philosophy, pedagogy, and social education. His practical and literary commitment was directed against social injustice, oppression, exploitation, dependence, and the lack of political freedom, and against the poverty and neglect of children and adults. His ambition was to free the poor from powerlessness and to help them be able to secure their own subsistence, as well as to abolish the political and social conditions which caused this hardship. Pestalozzi's pedagogical authorship was rooted in social reality. In his active pedagogical work, he developed a theory and practice of a new teaching and training method whose principles were the ability to visualize, simplification, logical progression, and individualization. His method was aimed at the moral elementary education of the child and of the future adult, as well as at imparting the skills with which they would be able to secure their own subsistence.

Introduction

Johann Heinrich Pestalozzi was one of the most distinctive writers of his time, and was already regarded as one of the most famous pedagogs in Europe in his own lifetime. He owed his reputation as a writer essentially to his best-selling novel *Lienhard and Gertrud* (1781), and owed his fame as a pedagog to his work in Stans (1799), Burgdorf (1800–03) and in Yverdon (1804–25) where for 21 years he directed an institute in which children from all over Europe were educated.

Johann Heinrich Pestalozzi's work deals with a wealth of subjects in the fields of general politics, social policy, constitutional law, jurisprudence, literature, anthropology, philosophy, pedagogy, and social pedagogy, and it is documented in 30 volumes in the Critical Edition *Samtliche Werke und Briefe*, 15 of which have hitherto appeared (also available on CD-ROM).

Pestalozzi was, in both his personal and literary life, actively opposed to social injustice, oppression, exploitation, dependence, and lack of political freedom. His ambition was to free the poor and their children from powerlessness, to help them to be able to secure their own individual subsistence, as well as to abolish the very political conditions which caused this hardship. He constantly gave central importance in his work to social justice, the satisfaction of needs, ownership and work, domestic order, and basic needs for all, including children. The subjects he addressed were rooted in social reality. Pestalozzi reflected upon and analyzed the everyday life of the poor masses, the everyday life of children, and the everyday political life of his time.

The special form of this relation to reality Michel Soetard (1981) has convincingly called the 'Principle Reality.'

Life and Work

1746–99: Childhood, Youth, and Early Career

Family Origins and Childhood

Johann Heinrich Pestalozzi was born in Zurich on the 12 January 1746, the son of the surgeon Johann Baptist Pestalozzi and his wife Susanna, nee Hotz. At that time, Zurich was a center of European culture, a city state with an exclusive merchant and entrepreneurial aristocracy. The town was regarded as rich, with a small number of old families holding power in the city council. Neither his father's family nor his mother's – who originally came from the Zurich area where her father was a parish priest – belonged to this group. The married couple produced seven children in 9 years of marriage from which only three grew to maturity. When Pestalozzi – one of seven children – was 5 years old, his father died at the age of 33, leaving his mother to take care of four small children. In order to lead a bourgeois life in spite of the financial predicament, economizing, and making-do became a staple of their dismal everyday life. Pestalozzi had therefore to endure poverty. In the country at his grandfather's house in Höng, where his grandfather was a parish priest, he became acquainted early on with the miserable social condition of the provincial population which was to occupy him for a lifetime.

School and Studies

In Zurich, Pestalozzi – who had the civic rights – grew up with the sons of the patrician families whose carefree material situation was permanently laid out before his eyes. He enjoyed the schooling of a city dweller, and went to the city elementary school from 1751 to 1754; he was then a pupil from 1754 to 1763 at the Schola Abbatissana in Zurich, the Schola Carolina, and the Collegium Humanitatis (also in Zurich), the latter being a Latin school. After some initial difficulties he was

a good pupil, though because of his financial situation he was always an outsider because of his 'extra-ordinary carelessness and uncleanliness.' From 1763 to 1765 he studied at Zurich's theological college, the Carolinum. The appearance of an anonymous pamphlet he had co-written caused the abrupt end to his studies. It was clear to him in any case that he did not wish to become a priest.

In Zurich there were – following the example of other European countries – patriotic and physiocratic societies. From 1764, Pestalozzi belonged to the circle of patriots of the Helvetic Society. The Zurich scholars Bodmer and Breitinger in particular had their circle of influence here. The much-admired model was Rousseau. Republican works were read and written. It was in this context that Pestalozzi's first publications *Agis* (1766) and *Wünsche* (1766) appeared. In 1767, Pestalozzi began a short agricultural apprenticeship with the Bern physiocrat Tschiffeli.

Livelihood

In February 1769, Pestalozzi became a farmer in Müllingen and ran into heavy debt. On 30 September, he married *Anna Schulthess* (1738–1815), and on the 13 August 1770 their only child Hans Jakob, called Jaqueli (1770–1801), was born. Pestalozzi built a manor house, the 'Neuhof auf dem Birrfeld' which he moved into with his young family in 1771. Crop failure, bad purchases of land, and other mistakes led to failure and almost to financial ruin. In 1794, he founded an institution for the poor at the Neuhof. His initial interest was more entrepreneurial than pedagogical; in his *Neuhof-Schriften* (published in 1777), Pestalozzi accompanied his essays with a journalistic advertising campaign for support for his project. In his *Bitte an Menschen-freunde*, he successfully enlisted the support of, among others, Niklaus Emanuel Tscharner (1727–94), the chairman of the Bern Society for Economics. Pestalozzi discovered his interest in all things pedagogical during his time at the Neuhof, motivated by his interaction with the poor children of the Neuhof: "I saw how the overwhelming cruelty of self-interest constantly destroys the body and soul of almost – and I mean almost – all of these children." Pestalozzi's institution failed financially, but he was later to make use of his experiences in the field of national education.

In the middle of the Neuhof crisis, the political work about Zurich *Freyheit meiner Vaterstatt* (1799) – which remained unpublished – was written in which he lamented the wealth of the city and its devastating effects on the constitution and political culture of Zurich.

From 1780 to 1799: Livelihood as an Independent Writer

Out of the necessity of having to finance himself, Pestalozzi became a writer. In 1780 he published *Abendstunde eines Einsiedlers* whose theme is vating people: "Cultivating people toward the truth means cultivating their being and their nature toward a reassuring wisdom." Then he wrote the first part of *Lienhard und Gertrud*, which appeared in 1781. *Lienhard und Gertrud* is a pedagogical novel which its author broadened into a novelistic pedagogical text. The theme of the novel is the story of the education and cultivation of an entire village. With *Lienhard und Gertrud*, Pestalozzi wrote about a fictitious village which he called Bonnal in a novel showing both the development and the education of the characters. The subject of the novel is bound together by a kaleidoscope of multifaceted and fictitious stories of education, cultivation, and socialization.

In the second (1783) and third (1785) books, it becomes clear how many individual fates have been embroiled in the deplorable state of affairs in Bonnal for many years, and that it is thereby not only a question of individual error, but at the same time of failings which lie more deeply in the political structure. Only in the fourth book (1787) does it become possible to continue telling the story of Bonnal and to organize it structurally as a story of reform, with, of course, further novelistic twists and turns and subplots. The village as the scene of the action is increasingly abandoned. Paragraph by paragraph Pestalozzi outlines his ideas for new legislation, the distribution of land, jurisdiction, and the culture of debate and justice. The novel is therefore increasingly interrupted on the level of action and plot, becomes more difficult to read, and thus loses its original novelistic character. The literary code transforms into a politically discursive text. *Lienhard und Gertrud* became Pestalozzi's greatest literary success. Pestalozzi was soon to become known far beyond the Swiss borders, first of all not as a pedagog, but as a writer. The novel brought him the title of Honorary Citizen during the French Revolution.

In *Gesetzgebung und Kindermord* (written in 1780 and printed in 1783), Pestalozzi declared the prevailing legal practice throughout Europe of executing child-murderesses – mostly after extreme humiliation and torture – as inhuman. He mounted an emotive defense of a large number of child-murderesses, whom he viewed as innocent. To him, poor girls and women had the fundamental right to sexuality and marriage, to their children, to protection and support, and to recognition and help. The work was a plea against capital punishment.

In 1782, Pestalozzi published *Christoph und Else*, a novel of instruction, even a 'novel' for the novel *Lienhard und Gertrud*. It was to be used as a guide to how adults should read *Lienhard und Gertrud*. A peasant family talk together over 30 evenings about the plot, the theme, and the constellation of characters in *Lienhard und Gertrud*. On 26 August 1792, Pestalozzi was appointed an Honorary French Citizen and never renounced the honorary title in spite of all the criticism of the Terror, an action which aroused suspicion in conservative circles. The outcome of Pestalozzi's examination of the revolution was the work *Ja oder Nein?* (1793) in which he made the Ancien Régime responsible for the devastating outrages in Europe and thus justified the revolution, though at the same time he sharply denounced the revolutionary Terror.

In 1797, Pestalozzi's main philosophical work *Meine Nachforschungen über den Gang der Natur in der Entwicklung des Menschengeschlechts* appeared. The initial question was 'What am I, and what is mankind?' Pestalozzi expounded a natural state, a social state, and a moral state of humanity, and went on to discuss the individual human being as a 'work of himself,' a 'work of nature,' and a 'work of morality.'

In 1798, the revolution also reached Switzerland. For a short time, a Helvetic Republic was established. Pestalozzi became the editor of the 'Helvetisches Volksblatt' – the publication organ of the new Helvetic government – and at first worked on behalf of the government. He wrote texts, appeals, and pamphlets, as well as *Abhandlung über die Natur des*

Helvetischen Zehnden. Whereas the typical pamphlets on both sides at the time of the revolution constituted obvious works of political agitation in favor of one side or the other, Pestalozzi was a mediator who urged the often very hostile opposing groups toward agreement. He became a soother of passions in his pamphlets. Pestalozzi wanted to promote the unity of the Swiss in the crisis year of 1798, the unity between the poor and rich population, the unity of town and country, and the unity of the Inner-Swiss (Catholic) cantons and the majority Protestant cantons.

1799–1827 the Pedagog and Institute Director: From Stans to Yverdon

Stans

At the age of 53, Pestalozzi received a government commission to run an orphanage in Stans. From 7 December 1798 to 8 June 1799, he lived there together with a housekeeper and as many as 80 children by the end. His pedagogical work and his pedagogical writing were now supported by two pillars. The first was a love of children and a passion for his work. Pestalozzi made himself into the 'father' and 'mother' of all the children; he tried to transpose the notion of the great importance of satisfying basic needs. The second was that the fundamental principle of the 'method' – which was on the one hand an almost technocratic syllabic reading method – was at the same time elevated by the idea of an elementary method of moral education. From then on, Pestalozzi's elementary education and elementary method were central themes. The institute, however, came to a sudden end. Pestalozzi had to hand over the building to French soldiers returning from the Russian campaign for use as a hospital and leave the institute, an action which made him fall into deep depression and despair. During this time, he wrote his strongly influential work, the Stanser Brief (1799), which contained a passionate plea for the malleability of human beings, including, and indeed especially, for the malleability of poor, neglected children.

Burgdorf

Pestalozzi became a teacher in Burgdorf on 23 July 1799. The announcement of a schoolmasters' conference in Burgdorf Palace followed on the 24 October 1800. The appearance of *Wie Gertrud ihre Kinder lehrt* followed in 1801. In the first part, Pestalozzi reported on Stans in autobiographical form, on his first endeavors in Burgdorf, on his young colleagues, some of whom accompanied him for a considerably long time, and on how he conceived method, the unity of knowledge, ability, and feeling. Pestalozzi developed a research program which he continued in Yverdon and which aimed at developing suitable learning aids in the three directions which he soon named 'HeadHeartHand.' The principles of the new teaching method consisted of the ability to visualize, simplification, logical progression, and individualization. The method was aimed at developing the autonomy of the child, at creating an independent personality which could secure its own subsistence. From November 1802 to February 1803, Pestalozzi traveled to Paris as a member of the Consulta. He was, however, unable to push through his concept for national education there. In June 1804, the institute at Burgdorf was moved to Münchbuchsee, and at the end of August the opening of a subsidiary institute followed in Yverdon, a town in French-speaking Switzerland, and on the 6 July 1805, Münchbuchsee was closed.

From 1804 to 1825, Pestalozzi ran one of the largest European educational institutes of the time in the rooms of the palace at Yverdon in French Switzerland. Yverdon was an experimental school on the level of its teaching, education, and teacher training. In order to be able to disseminate the new knowledge acquired from the everyday pedagogical research, a printing works was set up specifically for this purpose. Pestalozzi, who until that time had always written his works alone, now published journals, teaching books, teaching material, discourses on teaching methods, appeals, and speeches with his colleagues. At the center of these works stood Pestalozzi's concept of an elementary education of people which could be applied to different teaching disciplines.

At the height of its impact in 1809, the institute had 150 pupils from 6 to 16 years old, 31 teachers and assistant teachers, and 40 young people as observers from all over Europe who wished to study Pestalozzi's method. Pestalozzi's concept of elementary teaching was already being followed ever more closely in Prussia, Spain, and many other countries. Pestalozzi's most important work from the Yverdon period is *An die Unschuld, den Ernst und den Edelmut meines Zeitalters und meines Vaterlands* (1815), a treatise in the form of an address in which Pestalozzi set out his program for life: "The beginning and the end of my policy is education." Pestalozzi analyzed the approach of the industrial age with its phenomena of concentration, mechanization, and standardization. Only in education did he see the means to prepare people for the Industrial Age. The work caused a great sensation and was a literary success.

The size of the staff and the differences of interest and opinion within the teaching body increasingly led to endless quarrels which finally caused the downfall of the institute. Pestalozzi left Yverdon embittered in 1825, and returned to the Neuhof.

In the last 2 years of his life, he wrote the works *Meine Lebensschicksale als Vorsteher meiner Erzie-hungsinstitute in Burgdorf und Iferten* (1826), the *Schwanengesang* (1826), and finally the *Langenthaler Rede*. During the crisis which industrialization and the reorganization of the states into large administrative states gave rise to, Pestalozzi saw a great opportunity for man to return to himself, and for his inner renewal by means of a comprehensive elementary education.

After a short illness, Pestalozzi died on 17 February 1827. On 19 February, he was buried near the Neuhof in the small village of Birr.

Critical Reception and Research Developments

The research literature on Johann Heinrich Pestalozzi's life and work, amply documented by the research papers and synopses of the Pestalozzi literature of the past 200 years, at first glance appears immensely extensive. This lies in the fact that Pestalozzi, in the history of (German-speaking) pedagogy and educational sciences, has without doubt experienced a considerable *impact* history which has admittedly only to a certain degree coincided with the *reception* history of his numerous written works. Above all, Pestalozzi's teaching method is

prescribed in many countries around the world and his best-known works have been translated into many languages. Should one follow the numerous publications up to the present, it becomes apparent that, whereas in the first half of the twentieth century the 'whole Pestalozzi' became the subject, individual or partial aspects began, however, to dominate more and more markedly within the framework of the differentiation of educational science and its adjacent sciences. In this way, the paradigm exchanges in educational science are mirrored in the respective cognition interests of analyses. If the Pestalozzi from the history of ideas took center stage in scientific literature in the middle of the twentieth century, the research controversy about the *political Pestalozzi* (Rang, 1967) was sparked off in the 1960s and 1970s. In the course of the 1970s with its far-reaching education and school reform, interest was then concentrated rather more on the practitioner and reformer. Within the research history, it furthermore became clear that after a very intensive, bibliographically broad and verifiable acceptance of Pestalozzi the politically reforming pedagog, the research in the 1980s stagnated; indeed the number of papers and studies about Pestalozzi fell off considerably. That only changed in 1996. Since then, the trend has been moving in the direction of an examination of the research into context and reception (the impact history) alongside biographical research using new scientific methods, aimed at producing a differentiated situating of the work as regards history and content. An important impetus comes from the work of the Pestalozzi Research Center which, dividing the work between the Pestalozzianum and the University of Zurich, is pushing forward with the Critical Edition, and publishes the *Neue Pestalozzi-Studien* and the *Neue Pestalozzi-Blätter*.

Pestalozzi's journalistic work is so diverse that it not only touches on areas of social and cultural practice as regards content, or even focuses on them explicitly as a central theme, but also represents groups of different discourse methods. In order to understand the work, interdisciplinary research is required; e.g., into rhetoric, eloquence and discourse, the socioaesthetic culture of sentimentalism, social history, the relationship between the written and the spoken, the problematic nature of text selection, the relationship between truth and fiction, the theory of femininity, and the ensuing mother cult of the eighteenth century.

See also: Educational Institutions, History of; Educational Learning Theory.

Bibliography

Hager, F.P., Tröhler, D., 1996. Pestalozzi—wirkungsgeschichtliche Aspekte. Dokumentationsband zum Pestalozzi-Symposium 1996. Haupt, Bern, Stuttgart, Wien.

Korte, P., 2001. Padagogisches Schreiben um 1800. Der Status von Schriftkultur, Rhetorik und Poetik bei Johann Heinrich Pestalozzi.

Kraft, V., 1966. Pestalozzi oder das Pädagogische Selbst. Eine Studie zur Psychoanalyse pädagogischen Denkens. Julius Klinkhardt, Bad Heilbrunn.

Liedtke, M., 1992. Johann Heinrich Pestalozzi mit Selbstzeugnissen und Bilddokumenten. Rowohlt, Reinbek, Bei Hamburg.

Oelkers, J., Osterwalder, F., 1995. Pestalozzi—Umfeldund Rezeption. Studien zur Historisierung einer Legende. Beltz, Weinheim und Basel.

Osterwalder, F., 1996. Pestalozzi—ein pädagogischer Kult. Pestalozzis Wirkungsgeschichte in der Herausbildung der modernen Pädagogik. Beltz, Weinheim und Basel.

Pestalozzi, J.H., 1946. Sämtliche Briefe. Kritische Ausgabe. NZZ, Berlin, Zurich.

Pestalozzi, J.H., 1994. Sämtliche Werke und Briefe auf CD-ROM. Pestalozzianum, Zurich.

Rang, A., 1967. Der politische Pestalozzi. Frankfurt/Main.

Soetard, M., 1981. Pestalozzi ou la naissance de l'educateur (1746–1827). Peter Lang, Berne, Frankfurt, Las Vegas.

Stadler, P., 1988. Pestalozzi. Geschichtliche Biographie. In: Von der alten Ordnung zur Revolution, Band 1. Verlag Neue Zürcher Zeitung, Zürich.

Stadler, P., 1993. Pestalozzi. Geschichtliche Biographie. In: Von der Umwältzung zur Restauration, Band 2. Verlag Neue Zärcher Zeitung, Zärich.

Tröhler, D., 1997. Neue Pestalozzi-Studien. Haupt, Bern, Stuttgart, Wien.

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 18

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser–Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 18

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 18 *xlv*

VOLUME 18

Phantom Limbs 1
Joel Katz and Samantha Fashler

Pharmaceutical Industry: Political Economies of Drugs and Knowledge 6
Sergio Sismondo

Phenomenology in Sociology 11
Myron Orleans

Phenomenology: Philosophical Aspects 18
Christian Grüny

Philosophy of Law 24
Stefan Magen

Philosophy, Psychiatry, and Psychology 31
Nicholas Deakin

Phobias Across the Lifespan 37
Alfons O Hamm

Pholological Disorders 45
Hugh W Buckingham and Sarah S Buckingham

Phonetics, Articulatory 65
Xiaonong Zhu

Phonology 75
Mary Paster

Photography as a Medium 81
Karin Becker

Physical Activity and Health 87
Reinhard Fuchs

Physical Activity and Health Promotion 91
Gregory W Heath and Gary Liguori

Physical Sciences: History and Sociology 100
Richard Staley

Physicalism and Alternatives 106
Brian P McLaughlin

Piaget, Jean (1896–1980) 114
Jacques Montangero

Piaget's Theory of Cognitive Development 120
Anik de Ribaupierre

Piaget's Theory of Human Development and Education 125
Azwihangwisi E Muthivhi

Pidgin and Creole Languages 133
Salikoko S Mufwene

Pilgrimage 146
Simon Coleman

Place in Geography 149
Andrew Williams

Placebo Effect 153
Fabrizio Benedetti

Placebo Studies (Double-Blind Studies) 157
Harald Walach

Placebo-Controlled Trials, Ethics of 164
Rieke van der Graaf and Annette Rid

Plagues and Diseases in History 174
José L Peset

Planning and Development 180
Vanessa Watson

Planning During Post-Socialism 187
Sonia A Hirt

Planning Ethics 193
Carlos N Silva

Planning Issues and Sustainable Development 198
Peter Newman

Planning Theory: The Good City and Its Governance 202
Patsy Healey

Plato (428 BCE – 347 BCE) 208
Thomas Arnold and Jens Halfwassen

Play, Anthropology of 214
Don Handelman

Play During Infancy and Early Childhood: Cultural Similarities and Variations 218
Hui-Chin Hsu

Plea Bargaining: A Comparative Analysis 226
Jacqueline Hodgson

Pluralism 232
Richard J Ellis

Pluralism and Tolerance 236
Otfried Höffe

Polanyi, Karl (1886–1964) 243
György Lengyel

Police: A Sociology of Knowledge Approach 246
Peter K Manning and Jeffrey Martin

Police, Sociology of 251
Jonathon A Cooper

Police Use of Force 255
Geoffrey P Alpert

Policing 260
Mathieu Deflem and Samantha Hauptman

Policing in Divided Societies 266
Aogán Mulcahy

Policy Analysis 273
Aidan R Vining and David L Weimer

Policy Analysis: Evidence Based Policy-Making 281
Brian W Head

Policy Cycle 288
Michael Howlett and Sarah Giest

Policy Knowledge: Foundations 293
Helmut Anheier

Policy Knowledge: New Liberalism 299
Mary O Furner

Policy Knowledge: Universities 307
Michael Hoelscher

Policy Networks 314
William D Coleman

Policy Process: Nonprofit Institutions in the United States 320
Donald T Critchlow

Politeness and Language 326
Penelope Brown

Political Advertising 331
Stephen Ansolabehere

Political Anthropology 335
Christian Krohn-Hansen

Political Communication 342
Pippa Norris

Political Ecology: Anthropological Perspectives 350
Bengt G Karlsson

Political Economy, History of 356
Margaret Schabas

Political Economy in Anthropology 362
Joel S Kahn

Political Economy of Renewable Energy 367
Volkmar Lauber

Political Geography 374
Alexander B Murphy

Political History: History of Politics 380
James J Sheehan

Political Lawyering 386
Terence C Halliday and Lucien Karpik

Political Money and Party Finance 393
Thomas E Mann and Trevor Potter

Political Parties, History of 398
Estevao C de Rezende Martins

Political Parties: Communist 404
DS Bell

Political Parties: Extreme Right 410
Christopher T Husbands

Political Parties: Manifestoes 417
Ian Budge

Political Protest and Civil Disobedience 421
Dennis Chong

Political Psychology 427
Christian Staerklé

Political Religions: History of the Concept 434
Sabine A Haring

Political Representation 441
Dennis F Thompson

Political Repression 446
Linda C Keith

Political Science and Empirical Implications of Theoretical Models 453
Jim Granato, Melody Lo, and Man Chiu Sunny Wong

Political Science and Polling 459
Robert Y Shapiro

Political Science, History of 466
Erkki Berndtson

Political Sociology 472
Tyson Patros and Judith Stepan-Norris

Political Theory, History of 480
Bjorn Wittrock

Political Thought, History of 486
Rachel Hammersley

Political Tolerance 491
Sandra T Marquart-Pyatt

Political Trials 497
Leora Bilsky

Politics and Participation 504
Carole Jean Uhlaner

Politics and Socialization 509
M Kent Jennings

Politics, Formal Modeling of 512
Jeffrey S Banks

Pooled Cross-Sectional and Time Series Analyses, in Political Science 515
Ellen M Key and Matthew J Lebo

Popper, Karl Raimund (1902–94) 522
Peter Schroeder-Heister

Popular Culture 528
Purnima Mankekar

Popular Sexual Cultures in Western Europe (1930s to the Present) 533
Franz X Eder

Population Aging: Economic and Social Consequences 540
Jacques Légaré

Population Change in the Former Communist States of Europe and Asia 545
Timothy Heleniak

Population Composition by Race and Ethnicity: North America 553
Jenifer Bratter and Ellen Whitehead

Population Cycles, Formal Theory of 559
Shripad Tuljapurkar

Population Dynamics: Classical Applications of Stable Population Theory 562
Thomas W Pullum

Population Dynamics: Mathematic Models of Population, Development, and Natural Resources 566
Alexia Fürnkranz-Prskawetz

Population Dynamics: Momentum of Population Growth 572
Thomas J Espenshade and Jonathan BC Tannen

Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems 579
Peter Jagers

Population Dynamics: Theory of Nonstable Populations 581
Cyrus CY Chu

Population Dynamics: Theory of Stable Populations 584
John H Pollard

Population Dynamics: Two-Sex Demographic Models 588
Robert Schoen

Population Forecasts 593
Juha M Alho

Population Geography 597
Adrian J Bailey

Population Size Estimation and Capture–Recapture Methods 603
Ruth King and Antony Overstall

Populism 609
Koen Abts and Stijn van Kessel

Pornography 613
Gert M Hald and Neil M Malamuth

Positivism, History of 619
Steve Fuller

Positivism, Sociological 625
RW Outhwaite

Possession, Linguistic 630
Jüergen Broschart

Post Traumatic Amnesia 636
Thomas M McMillan

Postal Systems 640
Richard R John

Postcolonial Geography 644
Jane M Jacobs

Postcolonial Law 647
Eve Darian-Smith

Postcolonial Science Studies 652
Warwick Anderson

Postcoloniality 658
John D Kelly

Postindustrial Societies 663
Brian A Hoey

Postmodern Urbanism 670
David B Clarke

Postmodernism in Sociology 676
Alex Preda

Postmodernism: Methodology 680
Pauline V Rosenau

Postmodernism: Philosophical Aspects 684
Paul Patton

Post-Socialism: Political Economy 690
Aleksandra Sznajder Lee

Postsocialist Societies 697
Maya Grekova

Post-Traumatic Stress Disorder 702
J David Kinzie

Post-Traumatic Stress in Social Work 707
Orit Nuttman-Shwartz

Potlatch in Anthropology 714
ME Harkin

Poverty, Culture of 719
Phillippe Bourgois

Poverty, Geography of 722
Anne Bonds

Poverty in History 728
W Fischer

Poverty Law: United States 733
Jeffrey Selbin and Scott L Cummings

Poverty: Measurement and Analysis 741
Robert H Haveman

Poverty Policy 747
Hilary Silver

Poverty, Sociology of 753
Tim Smeeding

Power 760
Andrew S McFarland

Power in Society 765
Stewart Clegg

Power, Politics, and Influence in Organizations 770
John N Harris, Zachary A Russell, Liam P Maher, and Gerald R Ferris

Practical Reasoning: Philosophical Aspects 776
RJ Wallace

Practice: An Anthropological Approach 782
Jane Fajans

Pragmatic Disorders 788
Brigitte Stemmer

Pragmatics, Linguistic 795
Jef Verschueren

Pragmatism: Philosophical Aspects 803
Peter Godfrey-Smith

Pragmatist Social Thought, History of 808
Hans Joas

Prefrontal Cortex 811
Bryan Kolb

Prefrontal Cortex Development and Development of Cognitive Function 817
Robbin Gibb and Bryan Kolb

Prehistoric Art 824
Margaret W Conkey

Prejudice and Discrimination 828
Thomas F Pettigrew

Prejudice in Society: Psychological Perspectives 834
Jacques-Philippe Leyens

Prejudice in Society: Sociological Perspectives 839
Richard T Schaefer

Premotor Cortex 846
Giacomo Rizzolatti and Giuseppe Luppino

Prenatal and Infant Development: Overview, Current Trends, Future Directions 852
Lewis P Lipsitt, Jack Demick, and Carolyn Rovee-Collier

Presidency, United States 861
John Hart

Presidential Government 864
Juan J Linz

Pretend Play and Cognitive Development 870
Mark Nielsen

Prevention of Crime and Delinquency 877
Kristine Levan and David A Mackey

Prevention Research 883
Jean L Hill, Bret Kloos, Elizabeth Thomas, and Abe Wandersman

Primate Behavior: Significance for Understanding Humans 891
Roberto A Delgado, Elizabeth A Fox, and Jeanne Altmann

Primate Groups and Their Correlates 897
Pascal Gagneux

Primates, Evolution of 904
Robert D Martin

Primitive Society 911
Alexander F Robertson

Printing as a Medium 915
Adrian Johns

Prison Life, Sociology of 920
Paul Rock

Prison Life, Sociology of: Recent Perspectives from the United Kingdom 924
Deborah H Drake, Sacha Darke, and Rod Earle

Prisoner's Dilemma, One Shot and Iterated 930
Diego Ríos and Eleonora Cresto

Prisons and Imprisonment 938
Matthew DeMichele

Privacy of Individuals in Social Research: Confidentiality 942
Robert F Boruch and Ellen A Donnelly

Privacy: Theoretical and Legal Issues 947
Kirsty Hughes

This page intentionally left blank

CONTRIBUTORS TO VOLUME 18

Koen Abts
University of Leuven, Leuven, Belgium

Juha M. Alho
University of Helsinki, Helsinki, Finland

Geoffrey P. Alpert
University of South Carolina, Columbia, SC, USA; and Griffith University, Brisbane, QLD, Australia

Jeanne Altmann
Princeton University, Princeton, NJ, USA

Warwick Anderson
University of Sydney, Sydney, NSW, Australia

Helmut Anheier
Hertie School of Governance GmbH, Berlin, Germany

Stephen Ansolabehere
Massachusetts Institute of Technology, Cambridge, MA, USA

Thomas Arnold
University of Heidelberg, Heidelberg, Germany

Adrian J. Bailey
Hong Kong Baptist University, Kowloon Tong, Hong Kong, People's Republic of China

Jeffrey S. Banks
California Institute of Technology, Pasadena, CA, USA

Karin Becker
Konstfack College of Art, Craft and Design, Stockholm, Sweden

D.S. Bell
University of Leeds, Leeds, UK

Fabrizio Benedetti
University of Turin Medical School, and National Institute of Neuroscience, Turin, Italy

Erkki Berndtson
University of Helsinki, Helsinki, Finland

Leora Bilsky
Tel Aviv University, Tel Aviv, Israel

Anne Bonds
University of Wisconsin—Milwaukee, Milwaukee, WI, USA

Robert F. Boruch
University of Pennsylvania, Philadelphia, PA, USA

Phillippe Bourgois
University of California, San Francisco, CA, USA

Jenifer Bratter
Rice University, Houston, TX, USA

Jüergen Broschart
The GEO Magazine, Hamburg, Germany

Penelope Brown
Max Planck Institute of Psycholinguistics, Nijmegen, The Netherlands

Hugh W. Buckingham
University of Oklahoma Health Sciences Center (OUHSC), Oklahoma City, OK, USA

Sarah S. Buckingham
University of Oklahoma Health Sciences Center (OUHSC), Oklahoma City, OK, USA

Ian Budge
University of Essex, Colchester, UK

Dennis Chong
University of Southern California, Los Angeles, CA, USA

Cyrus C.Y. Chu
Academia Sinica, Taipei, Taiwan

David B. Clarke
Swansea University, Swansea, UK

Stewart Clegg
University of Technology, Sydney, NSW, Australia

Simon Coleman
University of Durham, Durham, UK

William D. Coleman
University of Waterloo, Waterloo, ON, Canada

Margaret W. Conkey
University of California Berkeley, Berkeley, CA, USA

Jonathon A. Cooper
Indiana University of Pennsylvania, Indiana, PA, USA

Eleonora Cresto
CONICET – Universidad de Buenos Aires, Buenos Aires, Argentina

Donald T. Critchlow
Arizona State University, Phoenix, AZ, USA

Scott L. Cummings
UCLA School of Law, Los Angeles, CA, USA

Eve Darian-Smith
University of California Santa Barbara, Santa Barbara, CA, USA

Sacha Darke
University of Westminster, London, UK

Nicholas Deakin
Barts Health NHS Trust, London, UK; and Queen Mary University of London, London, UK

Mathieu Deflem
University of South Carolina, Columbia, SC, USA

Roberto A. Delgado
University of Southern California, Los Angeles, CA, USA

Matthew DeMichele
RTI International, Research Triangle Park, NC, USA

Estevao C. de Rezende Martins
University of Brasilia, Brasilia, Brazil

Anik de Ribaupierre
University of Geneva, Genève, Switzerland

Ellen A. Donnelly
University of Pennsylvania, Philadelphia, PA, USA

Deborah H. Drake
The Open University, Milton Keynes, UK

Rod Earle
The Open University, Milton Keynes, UK

Franz X. Eder
Universität Wien, Vienna, Austria

Richard J. Ellis
Willamette University, Salem, OR, USA

Thomas J. Espenshade
Princeton University, Princeton, NJ, USA

Jane Fajans
Cornell University, Ithaca, NY, USA

Samantha Fashler
York University, Toronto, ON, Canada

Gerald R. Ferris
Florida State University, Tallahassee, FL, USA

W. Fischer
Freie Universität Berlin, Berlin, Germany

Elizabeth A. Fox
Duke University, Durham, NC, USA

Reinhard Fuchs
University of Freiburg, Freiburg, Germany

Steve Fuller
University of Warwick, Coventry, UK

Mary O. Furner
University of California Santa Barbara, Santa Barbara, CA, USA

Alexia Fürnkranz-Prskawetz
Vienna University of Technology, Vienna, Austria; Vienna Institute of Demography, Austrian Academy of Sciences, Vienna, Austria; and Wittgenstein Centre of Demography and Global Human Capital (WIC), Vienna, Austria

Pascal Gagneux
University of California San Diego, La Jolla, CA, USA

Robbin Gibb
University of Lethbridge, Lethbridge, AB, Canada

Sarah Giest
Simon Fraser University, Burnaby, BC, Canada

Peter Godfrey-Smith
City University of New York, New York, NY, USA

Jim Granato
University of Houston, Houston, TX, USA

Maya Grekova
Sofia University 'St. Kliment Ohridski', Sofia, Bulgaria

Christian Grüny
Universität Witten/Herdecke, Witten, Germany

Gert M. Hald
University of Copenhagen, Copenhagen, Denmark

Jens Halfwassen
University of Heidelberg, Heidelberg, Germany

Terence C. Halliday
American Bar Foundation, Chicago, IL, USA

Alfons O. Hamm
University of Greifswald, Greifswald, Germany

Rachel Hammersley
Newcastle University, Newcastle upon Tyne, UK

Don Handelman
The Hebrew University, Jerusalem, Israel

Sabine A. Haring
University of Graz, Graz, Austria

M.E. Harkin
University of Wyoming, Laramie, WY, USA

John N. Harris
Florida State University, Tallahassee, FL, USA

John Hart
Australian National University, Canberra, ACT, Australia

Samantha Hauptman
University of South Carolina Upstate, Spartanburg, SC, USA

Robert H. Haveman
University of Wisconsin—Madison, Madison, WI, USA

Brian W. Head
The University of Queensland, St Lucia, QLD, Australia

Patsy Healey
University of Newcastle, Newcastle, UK

Gregory W. Heath
University of Tennessee at Chattanooga and UT College of Medicine, Chattanooga, TN, USA

Timothy Heleniak
George Washington University, Washington, DC, USA

Jean L. Hill
New Mexico Highlands University, Las Vegas, NM, USA

Sonia A. Hirt
Virginia Tech, Blacksburg, VA, USA

Jacqueline Hodgson
University of Warwick, Coventry, UK

Michael Hoelscher
Universität Heidelberg, Heidelberg, Germany

Brian A. Hoey
Marshall University, Huntington, WV, USA

Otfried Höffe
Universität Tübingen, Tübingen, Germany

Michael Howlett
Simon Fraser University, Burnaby, BC, Canada

Hui-Chin Hsu
University of Georgia, Athens, GA, USA

Kirsty Hughes
University of Cambridge, Cambridge, UK

Christopher T. Husbands
London School of Economics and Political Science, London, UK

Jack Demick
Brown University, Providence, RI, USA

Jane M. Jacobs
Yale-NUS College, Singapore

Peter Jagers
Chalmers University of Technology, Gothenburg, Sweden

M. Kent Jennings
University of California, Santa Barbara, CA, USA

Hans Joas
Humboldt University, Berlin, Germany; and University of Chicago, Chicago, IL, USA

Richard R. John
Columbia University, New York, NY, USA

Adrian Johns
University of Chicago, Chicago, IL, USA

Joel S. Kahn
The University of Melbourne, Parkville, VIC, Australia; and La Trobe University, Melbourne, VIC, Australia

Bengt G. Karlsson
Stockholm University, Stockholm, Sweden

Lucien Karpik
École des Mines and École des Hautes Études en Sciences Sociales, Paris, France

Joel Katz
York University, Toronto, ON, Canada

Linda C. Keith
University of Texas at Dallas, Richardson, TX, USA

John D. Kelly
University of Chicago, Chicago, IL, USA

Ellen M. Key
Appalachian State University, Boone, NC, USA

Ruth King
School of Mathematics and Statistics, University of St Andrews, St Andrews, Scotland, UK

J. David Kinzie
Oregon Health and Science University, Portland, OR, USA

Bret Kloos
University of South Carolina, Columbia, SC, USA

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada

Christian Krohn-Hansen
University of Oslo, Oslo, Norway

Volkmar Lauber
University of Salzburg, Salzburg, Austria

Matthew J. Lebo
Stony Brook University, Stony Brook, NY, USA

Aleksandra Sznajder Lee
University of Richmond, Richmond, VA, USA

Jacques Légaré
Université de Montréal, Montréal, QC, Canada

György Lengyel
Corvinus University of Budapest, Budapest, Hungary

Kristine Levan
Plymouth State University, Plymouth, NH, USA

Jacques-Philippe Leyens
Catholic University of Louvain, Louvain-la-Neuve, Belgium

Gary Liguori
University of Tennessee at Chattanooga, Chattanooga, TN, USA

Juan J. Linz
Yale University, New Haven, CT, USA

Lewis P. Lipsitt
Brown University, Providence, RI, USA

Melody Lo
College of Business, University of Texas at San Antonio, San Antonio, TX, USA

Giuseppe Luppino
Università di Parma, Parma, Italy

David A. Mackey
Plymouth State University, Plymouth, NH, USA

Stefan Magen
Ruhr-Universität Bochum, Bochum, Germany

Liam P. Maher
Florida State University, Tallahassee, FL, USA

Neil M. Malamuth
University of California Los Angeles, Los Angeles, CA, USA

Purnima Mankekar
University of California Los Angeles, Los Angeles, CA, USA

Thomas E. Mann
The Brookings Institution, Washington, DC, USA

Peter K. Manning
College of Criminal Justice, Northeastern University, Boston, MA, USA

Sandra T. Marquart-Pyatt
Michigan State University, East Lansing, MI, USA

Jeffrey Martin
University of Hong Kong, Hong Kong, People's Republic of China

Robert D. Martin
The Field Museum, Chicago, IL, USA

Andrew S. McFarland
University of Illinois at Chicago, Chicago, IL, USA

Brian P. McLaughlin
Rutgers University, New Brunswick, NJ, USA

Thomas M. McMillan
University of Glasgow, Glasgow, UK

Jacques Montangero
University of Geneva, Geneva, Switzerland

Salikoko S. Mufwene
University of Chicago, Chicago, IL, USA

Aogán Mulcahy
University College Dublin, Dublin, Ireland

Alexander B. Murphy
University of Oregon, Eugene, OR, USA

Azwihangwisi E. Muthivhi
University of Cape Town, Cape Town, South Africa

Peter Newman
Murdoch University, Perth, WA, Australia

Mark Nielsen
The University of Queensland, Brisbane, QLD, Australia

Pippa Norris
Harvard University, Cambridge, MA, USA

Orit Nuttman-Shwartz
The Israeli National Council for Social Work, School of Social Work, Sapir College, Israel

Myron Orleans
California State University, Fullerton, CA, USA

R.W. Outhwaite
Newcastle University, Newcastle upon Tyne, UK

Antony Overstall
School of Mathematics and Statistics, University of St Andrews, St Andrews, Scotland, UK

Mary Paster
Pomona College, Claremont, CA, USA

Tyson Patros
University of California Irvine, Irvine, CA, USA

Paul Patton
University of New South Wales, Sydney, NSW, Australia

José L. Peset
CCHS — CSIC, Madrid, Spain

Thomas F. Pettigrew
University of California Santa Cruz, Santa Cruz, CA, USA

John H. Pollard
Macquarie University, Sydney, NSW, Australia

Trevor Potter
The Campaign Legal Center, Washington, DC, USA

Alex Preda
King's College London, London, UK

Thomas W. Pullum
University of Texas, Austin, TX, USA

Annette Rid
King's College London, London, UK

Diego Ríos
CONICET — UNTREF — UNSAM, Buenos Aires, Argentina

Giacomo Rizzolatti
Università di Parma, Parma, Italy; and IIT Brain Center for Social and Motor Cognition, Parma, Italy

Alexander F. Robertson
University of California Santa Barbara, Santa Barbara, CA, USA

Paul Rock
London School of Economics, London, UK

Pauline V. Rosenau
University of Texas, Houston, TX, USA

Carolyn Rovee-Collier[†]
Rutgers University, Piscataway, NJ, USA

Zachary A. Russell
Florida State University, Tallahassee, FL, USA

Margaret Schabas
University of British Columbia, Vancouver, BC, Canada

Richard T. Schaefer
DePaul University, Chicago, IL, USA

Robert Schoen
Pennsylvania State University, University Park, PA, USA

Peter Schroeder-Heister
Universität Tübingen, Tübingen, Germany

Jeffrey Selbin
UC Berkeley School of Law, Berkeley, CA, USA

Robert Y. Shapiro
Columbia University, New York, NY, USA

James J. Sheehan
Stanford University, Stanford, CA, USA

Carlos N. Silva
University of Lisbon, Lisbon, Portugal

Hilary Silver
Brown University, Providence, RI, USA

Sergio Sismondo
Queen's University, Kingston, ON, Canada

Tim Smeeding
University of Wisconsin—Madison, Madison, WI, USA

Christian Staerklé
University of Lausanne, Lausanne, Switzerland

Richard Staley
University of Cambridge, Cambridge, UK

Brigitte Stemmer
Université de Montréal, Montreal, QC, Canada; Brock University, St. Catherines, ON, Canada; and CRIUGM, Chemin Queen Mary, Montreal, QC, Canada

Judith Stepan-Norris
University of California Irvine, Irvine, CA, USA

Jonathan B.C. Tannen
Princeton University, Princeton, NJ, USA

Elizabeth Thomas
Rhodes College, Memphis, TN, USA

Dennis F. Thompson
Harvard University, Cambridge, MA, USA

Shripad Tuljapurkar
Mountain View Research, Los Altos, CA, USA

Carole Jean Uhlaner
University of California Irvine, Irvine, CA, USA

Rieke van der Graaf
University Medical Center Utrecht, Utrecht, Netherlands

[†] Deceased.

Stijn van Kessel
Loughborough University, Leicestershire, UK; and Heinrich-Heine-Universität, Düsseldorf, Germany

Jef Verschueren
University of Antwerp, Antwerp, Belgium

Aidan R. Vining
CNABS, Simon Fraser University, Vancouver, Canada

Harald Walach
European University Viadrina, Frankfurt (Oder), Germany

R.J. Wallace
University of California Berkeley, Berkeley, CA, USA

Abe Wandersman
University of South Carolina, Columbia, SC, USA

Vanessa Watson
University of Cape Town, Cape Town, South Africa

David L. Weimer
University of Wisconsin-Madison, Madison, WI, USA

Ellen Whitehead
Rice University, Houston, TX, USA

Andrew Williams
University of Exeter, Exeter, UK

Bjorn Wittrock
Swedish Collegium for Advanced Study in the Social Sciences, Uppsala, Sweden

Man Chiu Sunny Wong
University of San Francisco, San Francisco, CA, USA

Xiaonong Zhu
The Hong Kong University of Science and Technology, Hong Kong, People's Republic of China

Phantom Limbs

Joel Katz and Samantha Fashler, York University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The persisting sensory awareness of a limb after amputation has been termed the phantom limb. Phantom limbs are a normal and common sequela of amputation, occurring in approximately 90% of amputees. The phantom limb is a dynamic, frequently changing perceptual experience dependent upon ectopic activity from amputation stump neuromas, contributions from the sympathetic nervous system, alterations in the functional organization of sensory and motor nuclei after amputation, and the immediate emotional and psychological states of the individual. Phantom limb phenomena range from simple, diffuse sensations of tingling, to perceptually complex experiences of pains and lesions that originally were felt in the limb prior to amputation. While phantom pains and other sensations frequently are triggered by the perception of salient events, thoughts, and feelings, there is no evidence that the painful or painless phantom limb is a symptom of a psychological disorder. The sympathetic nervous system may provide an important link between higher brain centers involved in cognitive and affective processes and phantom limb sensations through its peripheral actions on primary afferent fibers located in stump neuromas by way of a sympathetic-efferent somatic-afferent coupling mechanism.

Many patients awake from the anesthetic after an amputation feeling certain that the operation has not been performed. They feel the lost limb so vividly that only when they reach out to touch it, or peer under the bedsheets to see it, do they realize it has been cut off. This startling realization does nothing to subdue the reality of the limb they experience, and may even intensify the sensations that define it. The first detailed description of this phenomenon was provided by the American Civil War surgeon, Silas Weir Mitchell (1871), who dubbed it the 'phantom limb.' The term so accurately captures the nature and subjective experience of the phenomenon that it has been adopted ever since. This article reviews the nature of phantom limb pain and nonpainful phantom limb sensations which arise after limb amputation.

Phantom Limb Pain

For many amputees, a distressing problem is phantom limb pain. While the distinction between a painful and nonpainful phantom is not always clear-cut, it is useful to separate them for purposes of description, classification, and treatment. Painful phantom experiences vary in a number of dimensions. These include intensity, quality, location, frequency, and duration. Many amputees report a painful intensification of the tingling, 'pins and needles' or paresthetic quality (i.e., dysesthesias) that defines the nonpainful phantom limb. Some sufferers describe bouts of paroxysmal shooting pain that travel up and down the limb. Others report the phantom to be in a cramped or otherwise unnatural posture that gives rise to excruciating pain. Many amputees describe the pain in the phantom limb as indistinguishable from the pain they experienced in the limb before amputation. In still others, the phantom may be immobile or paralyzed so that attempts to move it generate pain. Finally, the phantom is often the seat of an intense burning pain, as if the hand or foot were held too close to an open flame. Frequently amputees suffer from several types of pain. Phantom limb pain, it appears, is equally prevalent among men and women and is not related to side, level, age, or cause of amputation (Nikolajsen, 2013).

Surveys based on several thousand amputees reveal that between 78% and 85% of patients continue to experience significant amounts of phantom limb pain more than 25 years after amputation (Sherman et al., 1997). Prospective studies of the incidence of phantom limb pain are consistent with these estimates (Nikolajsen, 2013). Between 50% and 71% of amputees report phantom limb pain 2 years after amputation. There is a reduction in the frequency and duration of attacks over time. Equally striking is the low success rate of treatments for phantom limb pain: in the long term only 7% of patients are helped by the more than 50 types of therapy used to treat phantom limb pain (Sherman et al., 1997).

The different qualities of phantom limb pain, its high prevalence and incidence, and its low treatment success rate point to the complexity of the problem. Phantom limb pain is not a unitary syndrome, but a symptom class, with each class subserved by different etiologic mechanisms. There is not just one cause of phantom limb pain. The development and severity of this disorder are determined by the interaction of multiple factors that vary across individuals. These include ectopic activity from peripheral nerves that were transected during amputation and have formed a neuroma, contributions from the sympathetic nervous system, sensitization of central nervous system structures due to the effects of amputation, inflammation, and ectopic activity, alterations in the functional organization of sensory and motor nuclei, past experience with pain, and the immediate physiological, motivational, and psychological states of the individual.

Pain Memories in Phantom Limbs

A striking property of phantom limb pain is the presence of a pain that existed in a limb prior to its amputation (Katz and Melzack, 1990). This class of phantom limb pain is characterized by the persistence or recurrence of a previous pain, has the same qualities of sensation, and is experienced in the same region of the limb as the preamputation pain. Case studies of amputees have revealed pain 'memories' of painful diabetic foot ulcers, bedsores, gangrene, corns, blisters, ingrown toenails, cuts and deep tissue injuries, and damage to joints

and bony structures. In addition, the phantom limb may assume the same painful posture as that of the real limb prior to amputation, especially if the arm or leg has been immobilized for a prolonged period.

The proportion of amputees who report similar pain before and after amputation may be as high as 79% (Katz and Melzack, 1990), although according to prospective studies the incidence is approximately 10% at 1 year after amputation (Nikolajsen, 2013). Pain also persists in patients with deafferentation that does not involve amputation. Under these conditions, the involved body part is still present but it is devoid of sensibility due to an interruption in the supply of sensory (afferent) information (i.e., deafferentation). Brachial plexus avulsions, in which the sensory nerve roots supplying the arm and hand are torn from the spinal cord, often produce pain that is felt in the deafferented and anesthetic region. Similarly, patients with spinal cord injuries may complain of pain referred to body parts below the level of the transection, including, for example, a patient who continued to feel the pain of an ingrown toenail after a complete spinal cord break (Katz and Melzack, 2003).

Painful and nonpainful sensations also persist or recur after surgical removal or deafferentation of body structures other than the limbs, such as breasts, teeth, and internal and special sense organs. Ulcer pain has been reported to persist after subtotal gastrectomy with removal of the ulcer. Patients have reported labor pain and menstrual cramps after total hysterectomy, rectal pain and hemorrhoids after removal of the rectum and anus, the burning pain of cystitis after complete removal of the bladder, and the pain of a severely ulcerated cornea after enucleation of an eye.

As noted above, not all phantom limb memories are of painful experiences. Nonpainful preamputation sensations do recur, but they are not as common and tend to include the sensation of objects that were once worn on the limb. These superadded sensations vary in complexity from such simple somatosensory qualities as the sensation of bandages that once wrapped a wound, a plaster cast, finger rings, and wristwatches to the multimodal, perceptually integrated phenomenon of a phantom foot clad in a sock and a shoe of specific type and color (Katz and Melzack, 1990).

Taken together, these case reports and studies of amputees reveal that pain memories are not merely images or cognitive recollections; they are direct experiences of pain that resemble an earlier pain in location and quality. They are perceptually complex experiences that may even involve information from multiple sensory modalities including visual, olfactory, tactile, and motor components that had accompanied the original experience. The precise details of the experiences of pain involve localization, discrimination, affect, and evaluation – that is, all the dimensions of perceptual experience – and these properties are a function of integrated brain activity. It is likely that the outputs of sensitized spinal cells activate the neural structures in the brain that subserve memories of earlier events.

The Nonpainful Phantom Limb

The most salient property of the nonpainful phantom is the tingling, paresthetic quality, but sensations of temperature, posture, length, volume, positional orientation, and movement are also very common (Nikolajsen, 2013). The incidence of the nonpainful phantom is estimated to be between 80% and 100%. In fact, the presence of a nonpainful phantom is so common that it is considered a 'normal' sequela of amputation. It is standard practice to prepare the patient preoperatively for its subsequent occurrence.

The frequency with which a painless phantom occurs does not appear to be related to the extremity, side, level of amputation, age, gender, or reason for amputation (Nikolajsen, 2013). There is considerably more variability in the estimates of the duration of the nonpainful phantom. In some amputees, the phantom may disappear within days or weeks of the amputation. In others, it remains unchanged for up to 60 years.

Fading, Telescoping, and Shrinking

In approximately 30% of cases, the normal phantom undergoes a number of predictable changes within several years of amputation (Katz and Melzack, 2003). These include fading, telescoping, and shrinking. Immediately after amputation the phantom limb usually feels of normal length and size. Over time, the proximal portions begin to fade and soon disappear (**Figure 1**). At this point, the phantom limb consists of only the extremity and major joints. These are felt at a normal distance from the body. The intervening parts have disappeared so that the phantom hand or foot feels as if it were hanging, unattached, in empty space. In the case of an amputation performed at the shoulder, the phantom limb may consist of only the lower arm and hand, with a gap between the shoulder stump and the phantom elbow. In approximately one-third of amputees, a process of 'telescoping' begins. Among upper-extremity amputees, the phantom is perceived gradually to approach the stump so that the hand is located in phenomenal space on a level with the elbow of the other arm. As the process continues, the gap between the stump and phantom continues to close; the amputee may find that the hand is protruding from, or attached to, the end of the stump. Later, the phantom hand may retract into the stump so that only the tips of the fingers jut out. In some cases, eventually even these may disappear completely and permanently into the stump.

There are times when a completely telescoped phantom will temporarily lengthen so that the hand or foot is once again felt at a normal distance from the stump. Less commonly, the converse also occurs; a phantom that is already telescoped may temporarily retract completely into the stump, or one that is of normal or shortened length may become even shorter. The circumstances and stimuli that bring about these changes are extremely varied and range from peripheral stimulation of the stump and other regions of the body to more central influences.

The process of telescoping is sometimes accompanied by a marked shrinking of the phantom to the dimensions of a child's limb. This gradual diminution in perceived size or volume occurs in both upper- and lower-limb amputees. The extent of shrinking that occurs is variable. Adult amputees have reported the phantom to be the size of a doll's or baby's hand, a silver dollar, or even as small as a postage stamp (Katz and Melzack, 2003).

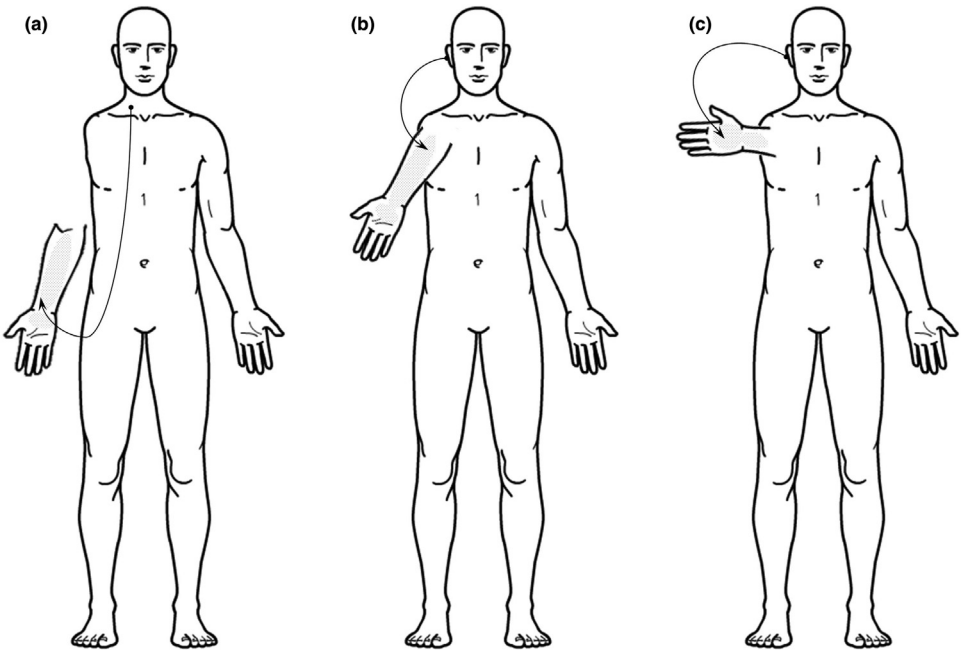

Figure 1 Illustrations of a patient who sustained a complete forequarter amputation of his right shoulder in a work accident and later developed phantom limb pain. The figurines depict the process of telescoping following a series of treatments with transcutaneous electrical nerve stimulation. Dots represent points where stimulation was applied. Arrows indicate the sites of referred phantom limb sensations. (a) The patient reported a phantom limb of normal length that consisted of only the lower arm and hand, with a gap between the shoulder stump and phantom elbow. During stimulation applied at a point above the clavicle, the patient reported that his phantom hand had begun to swell and become warm. (b) In response to stimulation applied at a point on his right ear, he remarked that his phantom arm had 'telescoped' into the stump so that the elbow protruded from the stump. (c) Later, stimulation of the same point resulted in a further retraction of the phantom, leaving only the wrist and hand attached to the stump. Source: Adapted from Katz, J., Melzack, R., 1987. Referred sensations in chronic pain patients. Pain, 28 (1), 51–59, with permission.

Cortical Maps and Phantom Limbs

Factors governing the fading, telescoping, and shrinking of the phantom are not well understood. The gradual fading of the phantom is thought to represent a perceptual correlate of the reestablishment of inhibitory control over hyperactive or spontaneously active cells that subserve the phantom. The gradual telescoping and shrinking have been hypothesized to depend upon input from intact peripheral fibers located in the stump. These experiences appear to be perceptual markers of the short- and long-term reorganizational changes that occur after amputation in central nervous system structures including the somatosensory (Katz and Melzack, 2003; Merzenich et al., 1984). The perceived distance between the phantom and the stump may be a function of the distance separating their respective representations in the cortical map. Cortical restructuring is also evident in other parts of the brain, including the posterolateral thalamus, premotor cortex, and vestibular regions. These findings suggest that the activation of dynamic brain networks is essential for the development and maintenance of phantom limb pain (Moseley and Flor, 2012).

Psychophysical Correlates of Phantom Limb Sensations

Although a normal phantom occurs whenever nerve impulses from the periphery are blocked or otherwise removed, it is also true that direct stimulation of the amputation stump frequently exaggerates the tingling or paresthetic quality of sensation typical of the painless phantom limb. Careful questioning of amputees reveals that the nonpainful phantom limb is not perceived as a static phenomenon. The paresthetic quality of sensation, which defines the phantom limb percept, is in a constant state of flux, with changes occurring in intensity, body part, or both.

One mechanism that accounts for the paresthetic component of the phantom limb is a cycle of sympathetic-efferent somatic-afferent activity (Katz and Melzack, 2003). Changes in the intensity of phantom limb paresthesias reflect the joint activity of cholinergic (sudomotor) and noradrenergic (vasomotor) postganglionic sympathetic fibers on primary afferents located in the stump and stump neuromas. Release of acetylcholine and norepinephrine from postganglionic sympathetic fibers produces transient vasoconstriction and heightened skin conductance responses. In addition, neurotransmitter release onto apposing peripheral fibers trapped in stump neuromas increases primary afferent discharge. This information is transmitted rostrally, where it gives rise to referred phantom sensations upon reaching central structures subserving the amputated parts of the limb. The moment-to-moment fluctuations in the intensity of phantom limb paresthesias reported by many amputees may, in part, reflect a cycle of sympathetic-efferent somatic-afferent activity. Increases in the intensity of phantom limb paresthesias would follow bursts of sympathetic activity, and decreases would correspond to periods of relative sympathetic inactivity. If central sensitization has also developed, through prior injury, trauma during amputation, or

peripheral inflammation, or if the sympathetic–sensory coupling involves nociceptors, the sensation may be one of dysesthesia.

Psychological and Emotional Processes Influence the Phantom Limb

The idea that emotional and psychological processes can cause pain traditionally has been tied to the notion of psychopathology. However, it is becoming increasingly clear that under certain circumstances pain may be triggered by these processes in psychologically healthy individuals as well. The aggravation or alleviation of pain referred to phantom body parts may be mediated in part by psychological processes that alter anxiety levels. Phantom breast pain after mastectomy is provoked by emotional distress in 6% of women 3 weeks after surgery, and in 29% 1 year later. Of lower-extremity amputees, 50% report that attacks of phantom limb pain are triggered by emotional distress as long as 7 years after amputation. In addition, both pain catastrophizing and coping styles have been implicated as important moderators of, and possible risk factors for, phantom limb pain (Vase et al., 2011; Richardson et al., 2007).

Psychological or emotional processes can bring about transient but profound alterations in the quality and intensity of phantom limb sensations. These processes include concentration, distraction, relaxation, fright, forceful reminders of the events that led to amputation, and witnessing cruel and violent acts. One amputee, interviewed by the present writer, described his reaction to an accident involving his wife by reporting "goose bumps and cold shivering down the phantom [leg]. It went through me. Everything emotional will get you that." Another amputee stated, "It's like everything I feel goes there – the good and the bad" (Katz and Melzack, 2003).

A Centrally Triggered Sympathetic-Efferent Somatic-Afferent Coupling Mechanism

The model schematically represented in Figure 2 outlines a mechanism through which cognitive and affective processes associated with higher cortical and limbic centers may alter phantom limb sensations. The reciprocal connections between cortical, limbic, and lateral hypothalamic structures are well documented. The lateral hypothalamus is involved in the control and integration of neural activity associated with

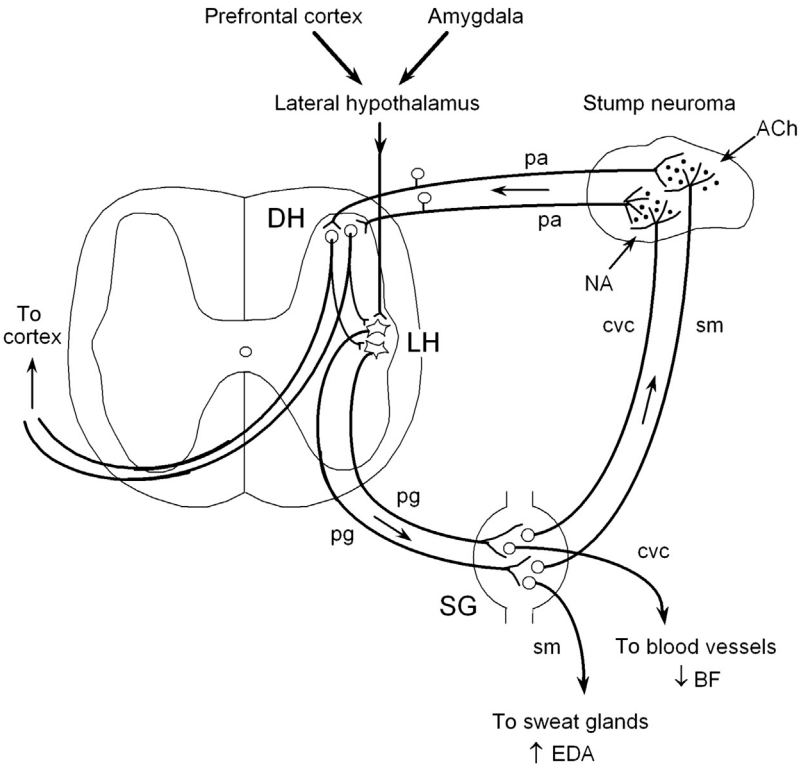

Figure 2 Schematic diagram illustrating a mechanism of sympathetically generated phantom limb paresthesias. Spontaneous sympathetic activity or excitatory inputs descending from the cortex (e.g., due to the perception of a salient event, loud noise, thought, feeling, etc.) increase the discharge rate of preganglionic (pg) sympathetic neurons with cell bodies in the lateral horn (LH) of the spinal cord and terminals in the sympathetic ganglion (SG). These neurons excite postganglionic noradrenergic (NA) cutaneous vasoconstrictor (cvc) and cholinergic (ACh) sudomotor (sm) fibers that impinge on effector organs (vascular smooth muscle and sweat glands) in the stump and on sprouts from large-diameter primary afferent (pa) fibers that have been trapped in a neuroma. The release of ACh and NA on effector organs results in increased electrodermal activity (EDA) and decreased blood flow (BF) to the stump. Release of these chemicals in the neuroma activates primary afferents that project to spinal cord dorsal horn (DH) cells subserving the amputated parts of the limb. These neurons, in turn, feed back to the preganglionic sympathetic neurons and project rostrally where the impulses contribute to the perception of phantom limb paresthesias. If DH cells have been sensitized due to injury, or nociceptive primary afferents are activated, then the perception may be one of dysesthesias. Source: Reproduced from Katz, J., 1992. Psychophysical correlates of phantom limb experience. Journal of Neurology, Neurosurgery, and Psychiatry 55(9), 811–821.

affectively charged behavior and has direct projections to the lateral horn of the spinal cord. The intensity of phantom limb paresthesias and dysesthesias may thus be modulated by higher brain centers involved in cognitive and affective processes via a multisynaptic network of descending inputs, which impinges on preganglionic sympathetic neurons, producing diffuse peripheral autonomic discharge and activation of primary afferent fibers located in stump neuromas.

Treatment for Phantom Limb Pain

Over the years, many different treatments have been tried for phantom limb pain, including surgical, pharmacological, physical, and psychological approaches (Nikolajsen, 2013). Surgical treatments are typically performed on the stump to remove neuromas that contribute to the pain or are directed at more central structures but these too have met with limited success, largely because they produce further deafferentation that contributes to maladaptive reorganization of sensory and motor brain areas. Preemptive analgesic strategies target limb pain prior to amputation with the aim of preventing phantom limb pain (Katz, 1997; Katz et al., 2011) but the evidence for their success is equivocal (Katz and Seltzer, 2009). More recently, several alternative mind–body approaches have been developed to target phantom limb pain including mental imagery training and mirror visual feedback, but it is too early to know whether they will prove to be effective (Moseley and Flor, 2012). Overall, despite the many forms of available treatment, not one has been effective in a majority of cases and there are few, if any, that provide permanent relief. Most amputees continue to suffer the tragedy of phantom limb pain.

Summary and Conclusions

The phantom limb is not perceived as a static entity but as a frequently changing perceptual experience dependent upon the current sensory input, emotional state, and past experience of the individual amputee. Phantom limb phenomena range from simple, diffuse sensations of tingling to perceptually complex experiences of pains and lesions that originally were felt in the limb prior to amputation. While phantom pains and other sensations frequently are triggered by the perception of salient events, thoughts, and feelings, there is no evidence that the painful or painless phantom limb is a symptom of a psychological disorder. The sympathetic nervous system may provide an important link between higher brain centers involved in cognitive and affective processes and phantom limb sensations through its peripheral actions on primary afferents located in stump neuromas. Of the many forms of treatment for phantom limb pain, not one has been effective in a majority of cases and there are few, if any, that provide permanent relief.

See also: Affect-Regulation Motivation; Chronic Illness, Psychosocial Coping with; Chronic Pain: Models and Treatment Approaches; Cognitive Neuroscience; Declarative Memory, Neural Basis of; Emotion, Neural Basis of; Emotional Regulation; Episodic Memory; Health Psychology; Implicit Memory; Neural Plasticity; Neuroprosthetics; Pain, Health Psychology of; Post-Traumatic Stress Disorder; Self and Brain; Stress, Coping and Health; Symptom Perception, Awareness and Interpretation; Virtual Reality and Spatial Cognition.

Bibliography

Katz, J., 1997. Phantom limb pain. The Lancet 350, 1338–1339.

Katz, J., Clarke, H., Seltzer, Z., 2011. Review article: preventive analgesia: quo vadimus? Anesthesia & Analgesia 113 (5), 1242–1253. http://dx.doi.org/10.1213/ANE.0b013e31822c9a59.

Katz, J., 1992. Psychophysical correlates of phantom limb experience. Journal of Neurology, Neurosurgery, and Psychiatry 55 (9), 811–821.

Katz, J., Melzack, R., 1990. Pain "memories" in phantom limbs: review and clinical observations. Pain 43, 319–336.

Katz, J., Melzack, R., 2003. Phantom pain. In: Grafman, J., Robertson, I.H. (Eds.), Handbook of Neuropsychology, second ed., vol. 9. Elsevier Press, Oxford, pp. 205–230.

Katz, J., Melzack, R., 1987. Referred sensations in chronic pain patients. Pain 28 (1), 51–59.

Katz, J., Seltzer, Z., 2009. Transition from acute to chronic postsurgical pain: risk factors and protective factors. Expert Review of Neurotherapeutics 9 (5), 723–744. http://dx.doi.org/10.1586/ern.09.20.

Merzenich, M.M., Nelson, R.J., Stryker, M.P., Cynader, M.S., Schoppmann, A., Zook, J.M., 1984. Somatosensory cortical map changes following digit amputation in adult monkeys. Journal of Comparative Neurology 224, 591–605.

Mitchell, S.W., 1871. Phantom limbs. Lippincott's Magazine of Popular Literature and Science 8, 563–569.

Moseley, G.L., Flor, H., 2012. Targeting cortical representations in the treatment of chronic pain: a review. Neurorehabilitation & Neural Repair 26 (6), 646–652. http://dx.doi.org/10.1177/1545968311433209.

Nikolajsen, L., 2013. Phantom limb. In: McMahon, S.B., Koltzenburg, M., Tracey, I., Turk, D.C. (Eds.), Wall and Melzack's Textbook of Pain, sixth ed. Churchill Livingstone, Edinburgh, pp. 915–925.

Richardson, C., Glenn, S., Horgan, M., Nurmikko, T.J., 2007. A prospective study of factors associated with the presence of phantom limb pain six months after major lower limb amputation in patients with peripheral vascular disease. Journal of Pain 8 (10), 793–801.

Sherman, R.A., Devor, M., Jones, D.E.C., Katz, J., Marbach, J.J., 1997. Phantom Pain. Plenum Press, New York.

Vase, L., Nikolajsen, L., Christensen, B., Egsgaard, L.L., Arendt-Nielsen, L., Svensson, P., Staehelin Jensen, T., 2011. Cognitive-emotional sensitization contributes to wind-up-like pain in phantom limb pain patients. Pain 152 (1), 157–162.

Pharmaceutical Industry: Political Economies of Drugs and Knowledge

Sergio Sismondo, Queen's University, Kingston, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The pharmaceutical industry has become one of the most important agents of social change and contributors to social structure in the developed world, as well as having significant effects in other parts of the world. As a result, the industry also has become a site of research by anthropologists, historians, sociologists, and other scholars in science and technology studies. This article focuses on how: the pharmaceutical industry has been a major force behind international intellectual property agreements, spends large sums on medical research and enrolls a great many subjects in its clinical trials, has a substantial impact on medical knowledge and opinion, and shapes people's perceptions of health and illness.

The Pharmaceutical Industry as a Key Social Actor

Global pharmaceutical sales increase by 5–7% annually, and are projected to total approximately US $1 trillion in 2014. In North America and Western Europe, physicians write more than a dozen prescriptions per capita per year. These include many prescriptions for drugs from which few people will benefit and for conditions they cannot feel, such as drugs to lower cholesterol and blood pressure or to increase bone mineral density. Moreover, increasing numbers of people take drugs for purposes not traditionally seen as medical: to improve their sex lives, reduce their weight, thicken their hair, sleep better, and feel less anxious. Certainly, drugs are a long-standing phenomenon, a form of easy hope in a glass bottle or plastic vial. But this hope's recent growth is extraordinary, expertly nurtured in governments, physicians, and consumers by the scientific marketers of the pharmaceutical industry.

The pharmaceutical industry has become one of the most important agents of social change and contributors to social structure in the developed world, as well as affecting other parts of the world. As a result, the industry also has become the site of significant research by anthropologists, historians, sociologists, and other scholars in science and technology studies. Though there are many other facets to the industry, this article focuses on a few important features: the industry has been a major force behind international intellectual property agreements, spends large sums on medical research and enrolls a great many subjects in its clinical trials, has a substantial impact on medical knowledge and opinion, and shapes people's perceptions of health and illness.

Regulation: From Influence to Partnerships

Even though it has reacted strongly against most governmental attempts at regulation, the pharmaceutical industry has influenced, adapted to, and profited from those regulations. As a result, the industry today is very much a creature of its regulation. The distinction between pharmaceuticals and their lower-status counterparts – such as patent medicines, herbal remedies, supplements, illegal drugs, and traditional medicines – depends on demarcations that require certain medicines to be scientifically tested for their safety and efficacy, approved for particular uses and then prescribed by licensed physicians or sold through licensed pharmacies (e.g., Rodwin, 2011). For the most part, those demarcations are made by government regulators such as the US Food and Drug Administration (FDA), the European Medicines Agency (EMA), or the Japanese Pharmaceutical and Medical Devices Agency (PMDA). The regulations established by the FDA in the 1960s, and widely emulated elsewhere, were sometimes fiercely opposed by the industry (Tobbell, 2012). Although regulation has increased the cost of new drugs, it also has tended to improve their image.

More recently, the industry has attempted to reduce its costs by standardizing requirements internationally; this allows companies to use the same research for their applications for a drug's approval in many jurisdictions. An agreement by the largest pharmaceutical companies and the United States, Europe, and Japan, the International Conference on Harmonization (ICH), established a standardized multiphased clinical research process. In so doing it helped to make the testing of drugs a global phenomenon, even though it has been tempered by local therapeutic cultures (Daemmrich, 2004). It has been argued that the agreement has tended to harmonize standards 'downward,' resulting in overall weaker carcinogenicity testing, weaker reporting of adverse events, and less long-term monitoring of adverse events (Abraham and Reed, 2001).

The ICH is one example of close cooperation between the pharmaceutical industry and drug regulators. However, regulators are usually receptive to pressures to balance interests in safety and effectiveness with the financial interests of a thriving industry that, it is argued, provides long-term health and economic benefits. Thus, the experts who evaluate drug applications are expected to be 'friendly' to the industry even as they apply strict standards, and the regulator as a whole is expected to support the industry even as it upholds laws (Abraham, 2008). Close cooperation is facilitated by the fact that pharmaceutical companies and their lobby groups tend to have coherent and consistent interests, and are sophisticated stakeholders. The cooperation is also facilitated by the increasing reliance of regulators on revenue from the companies themselves: the UK's Medicines and Healthcare Products Regulatory Agency receives 100% of its funding for evaluating new drug applications from the applicants; in the United States the figure is 50%, and the European Medicines Agency collects 70% of its total budget from fees paid by pharmaceutical companies

 http://dx.doi.org/10.1016/B978-0-08-097086-8.85054-0

(Abraham, 2008). National regulators have, provocatively, been characterized as umpires of intellectual property claims, rather than protectors of the public (Mirowski and Van Horne, 2005).

The pharmaceutical industry also affects other aspects of government. In the early 1980s, a group of companies successfully put intellectual property at the top of the US trade agenda and helped to design not only international agreements, but also a very successful strategy for convincing other countries to sign onto those agreements (e.g., Drahos, 1995). Prominent in that group of companies were representatives of the pharmaceutical industry: the Advisory Committee for Trade Negotiations, formed in 1981, was chaired by the CEO of Pfizer, Ed Pratt, and its 1986 successor, the Intellectual Property Committee, consisted of 13 large companies including Pfizer, Bristol-Myers, Johnson & Johnson, and Merck, making pharmaceuticals the largest of the industry sectors represented on the committee; the chemical, computer, automotive, aerospace, and communication industries also were represented. These committees designed a strategy that combined the building of coalitions with non-US companies, forceful bilateral pressure to link intellectual property and other trade, and finally multilateral agreements built on those bilateral pressures (Drahos, 1995). The result was the Agreement on Trade-Related Aspects of Intellectual Property Rights, which has formed the basis for new intellectual property laws in countries around the world.

Clinical Research: From Research and Development to Risk-Managed Production

The pharmaceutical industry sponsors a majority of the 40 000 new trials initiated each year (CenterWatch, 2009). Annually, between 3 and 6 million people are enrolled in industry-initiated clinical trials. The bulk of these trials are conducted by contract research organizations (CROs), companies that perform research and development for profit (Mirowski and Van Horne, 2005). As an indication of their rising importance, CRO revenues increased from US $1 billion in 1992 to an estimated $27 billion in 2010. CROs are involved in nearly 60% of all industry-sponsored studies, and 75% of outsourced R&D expenditures. The rest of the industry's outsourced R&D goes to academic medical centers. It should be noted that contracts with CROs and academic centers may not be comparable: CROs may be winning contracts for trials that otherwise would have been done directly by pharmaceutical companies, and the reasons for a company sponsoring an academic-run trial may have as much to do with establishing or maintaining relations with academic researchers as with other needs for trials.

Contact with subjects may occur at various sites, including academic medical centers that have an intellectual interest in the particular research or that use trial income to support themselves, clinics at which physicians supply small numbers of patients as a supplement to their regular practices, clinics that specialize in trials and earn substantial income from the procedures performed, specialized hospital units that run difficult inpatient trials, and sites that cut costs to a minimum and may house hundreds of subjects – referred to as 'floater sites' in part because the expected lifespan of a low-cost clinic is short, as it is likely to run into regulatory or political problems. Trial income is a factor in almost all of these kinds of clinics, as pharmaceutical companies pay substantial fees per patient and/or per procedure (Fisher, 2009).

Nearly 50% of all trial sites are in North America, and almost 20% are in Western Europe. Those figures correlate well with North America's and Western Europe's shares of the international pharmaceutical market, reflecting various historical and current relationships between trials and drug consumption. Yet the geographical distribution of trials is changing quickly. An increasing number of sites are in Asia, Africa, South America, and Eastern Europe; there are significant West-to-East and North-to-South shifts.

Pharmaceutical companies have better-developed networks of investigators and potential investigators in North America and Western Europe. Moreover, having sites in these places helps to maintain some North American and European investigators as allies and potential authors of papers stemming from clinical research. But the costs of trials – in fees to physicians and subjects, and the cost of medical procedures – are much lower in many other parts of the world, even parts that have sufficient medical infrastructure to serve as good locations for trials.

Recruitment is one driver of clinical trial costs. Many people in the United States and Western Europe are unwilling to participate in trials, and many are ineligible because they are already being treated. Wealthier countries are often 'treatment saturated,' and therefore are poor sources of the more straightforward or even ideal 'treatment naïve' test subjects abundant elsewhere. Moreover, companies are sophisticated about populations; one former chief executive of a CRO says: "Companies can now pick and choose populations … in order to get a most pronounced drug benefit signal as well as a 'no-harm' signal" (quoted in Petryna, 2009: p. 17). This strategy requires access to larger populations, and often populations in multiple countries. India, for example, is well positioned to provide subjects. India's *Economic Times* wrote in 2004: "The opportunities are huge, the multinationals are eager, and Indian companies are willing. We have the skills, we have the people …" (quoted in Shah, 2006: p. 17). And perhaps most obviously, the costs per patient are considerably lower in India than in North America or Western Europe.

Ethical variability is a further reason for the globalization of trials. Different standards apply even between established members of the European Union and newly admitted members in Eastern Europe. When participants in lower-income countries try to apply basic international ethical standards, variability in circumstances leads to substantial variability in application. Thus, there have been cases of enormous differences in protocols between higher- and lower-income countries, and a few extreme cases have sparked considerable debate (e.g., Petryna, 2009: pp. 38–40). CROs take advantage of ethical variability to lessen the cost and increase the efficiency of trials; they also compete with each other and perhaps with academic medical centers, creating downward pressure on the implementation of ethical standards. There may also be public relations and liability pressures to move trials away from North America and Western Europe. Deaths and other severe adverse events are more likely to attract unwanted media attention in

pharmaceutical companies' core markets than if they occur elsewhere. Legal liabilities also tend to be much higher in North America and Western Europe, and especially in the United States, than in the rest of the world.

Professional research subjects, people who earn at least a substantial portion of their income by signing up for Phase I (healthy subject) clinical trials, may see their participation as a kind of risky labor (Abadie, 2010; Sunder Rajan, 2007). From such a perspective, they are exploited for their 'surplus health' (Dumit, 2012), which can be transformed into pharmaceutical company profits. However, from the perspective of the companies, and given the systems built around them, subjects are simply inputs to production, most immediately the production of knowledge that will contribute to the success of the drugs. Subjects are a key, perhaps the key, resource that CROs and clinics can provide.

Clinical trial knowledge is a necessary input for the authorization to market new drugs. It is also a key resource for selling those drugs once on the market. Trials, therefore, are implemented in places and in ways that maximize value – in terms of keeping costs low, increasing the chances of positive results, and leveraging relationships and credibility of researchers.

Illness: From Medicalization to Pharmaceuticalization

Medicalization became a major issue in the 1960s, as critics railed against the authority of physicians and the medical profession as a whole to take control over intimate events and processes, often simply through classification. Criticism of medicalization was especially prominent in challenges to psychiatry by scholars such as Michel Foucault and Thomas Szasz. Their work was later picked up and applied in different realms, especially by feminist scholars. As a result, there have been challenges especially to mental health diagnostic categories ranging from schizophrenia to anxiety and depression, and challenges to the medicalization of ordinary life events and stages, such as childbirth and menopause.

The focus on professions has lost some of its traction, and physicians today appear to be only one set of actors in struggles for control over bodies, health, and illness. Much medicalization, then, looks as though it is merely grease for the wheels of 'pharmaceuticalization' (e.g., Williams et al., 2011). This claim is made even more plausible by the fact that some of the standard objects of critics' attention, such as anxiety, depression, and menopause, are closely associated with new classes of drugs. To increase their sales, pharmaceutical companies engage in 'selling sickness' or 'disease-mongering': expanding awareness of diseases for which their drugs can be prescribed, and increasing the likelihood that people will see themselves as having those diseases. Moreover, at least some pharmaceuticalization is independent of medicalization, when people use drugs as remedies for difficulties that they do not necessarily see in medical terms, such as excess weight, sleepiness, lack of focus, and shyness (Williams et al., 2011).

The category of depression is widely understood as having been affected by the availability of drugs. Before the 1980s, depression was a relatively uncommon diagnosis, and was generally associated with the elderly (Healy, 2004). Beginning with the arrival on the market of Eli Lilly's drug Prozac in 1987, however, ever-increasing numbers of people have been diagnosed with depression and are disabled by depression (Whitaker, 2010). Diagnostic criteria for depression have continually broadened, and thus estimates of the prevalence of depression have risen dramatically. It is now the 'common cold' of mental disorders.

Prozac was the first widely prescribed member of the Selective Serotonin Reuptake Inhibitor (SSRI) family of drugs, now often called simply 'antidepressants.' Companies selling SSRIs have marketed both the drug and the disease. They have invested heavily in research on depression and antidepressants. They have promoted a serotonin-deficiency theory of depression, for which there is little evidence (Whitaker, 2010). They have established close connections with psychiatrists and other physicians who write textbooks, articles, and clinical practice guidelines (Healy, 2004). Through such means and others, the companies appear to have successfully established the disease both medically and culturally, helping physicians see it often and helping patients interpret their feelings and experiences in terms of it – perhaps even shaping their identities around it. The World Health Organization predicts that within 20 years more people will be afflicted with depression than with any other health problem.

Depression may seem like a special case, because it is a mental illness and because the boundaries between the disorder and sadness are malleable. However, many examples of somatic or bodily illness have been shown to have been strongly affected by marketing efforts, including such common chronic diseases as hypertension, diabetes, high cholesterol, and osteoporosis (Greene, 2007; Moynihan and Cassels, 2005). To take one of these examples, when the pharmaceutical company then known as Merck Sharp & Dohme introduced its anti-hypertensive drug Diuril in 1957, high blood pressure was a sign associated with an underlying disease, but was not itself generally viewed as a problem to be controlled. Even the definition of high blood pressure was a difficult matter, since blood pressure had a Gaussian distribution, and any dividing line between high and normal seemed arbitrary. Diuril dramatically lowered blood pressure, apparently only in people with elevated levels, suggesting that it might address root causes of the problem. In conjunction with an unprecedentedly successful marketing campaign, its effects established hypertension as a disease in itself (Greene, 2007). The mildness of side effects of diuretic drugs made it straightforward for companies and their agents to argue for their use to treat ever-broader populations with ever-lower blood pressures. Recommendations followed suit, and blood pressures that were once judged as within a normal range became high. Thus, people came to understand themselves as hypertensive or as disposed to hypertension; in at least one population, hypertensive African-Americans, the condition has become linked to racial identity (Pollock, 2012). The pharmaceuticalization process has continued as new and different antihypertensive drugs have been found and marketed.

Chronic conditions such as hypertension neatly fit a new model of health and illness. Whereas people once generally considered themselves healthy unless they felt ill or had visible symptoms, some changes have resulted in a new model of health (Dumit, 2012). The past half-century has seen the rise of

risk factors: familiar ones like diet, age, and sleep patterns, and unseen and less familiar ones like cholesterol levels and high blood pressure. As a result, we are all at risk, differing only in our degrees of risk. Partly following from the rise of risk factors, we can be normal and unhealthy at the same time, at least when there is some hope for treatment. For example, the results of aging used to be simply unfortunate, but now we look for medical means to stave them off or treat them. There is no contradiction in the thought that people are unhealthy in one or another respect.

Marketing: Moving Medical Knowledge to Physicians and Consumers

Independent evaluations in France, the United States, and Canada indicate that only 10% of newly approved drugs offer significant advantages over existing treatments. Lack of innovation is all the more striking because the pharmaceutical industry portrays itself primarily as doing research and development, and only incidentally involved in marketing, advertising, and sales. This portrayal is necessary, since customers demand that drug decisions be based entirely on good evidence about health benefits and costs, and not on hearsay, fashion, or advertising.

Despite its apparent lack of innovation, the pharmaceutical industry consistently has the highest ratio of profit to investment among industries in the Fortune 500 companies. Paradoxically, then, innovation may not be required for high profits.

Pharmaceutical companies can attempt to structure and restructure medical knowledge in different ways, such as through innovative drug discovery or hegemony over medical knowledge and its communication. The latter, it appears, is a better investment of resources (Gagnon, 2009). Therefore, the industry has accumulated resources for shaping areas of medical knowledge, leading to greater profits than they could reap from investments in traditional drug research and development. Medicine's (and regulators') focus on expensive clinical trials in particular has meant that pharmaceutical companies are in a good position to control scientific knowledge about their products.

To physicians, some of the most visible conduits of both pharmaceutical marketing and communication are sales representatives, who can be important movers of medical knowledge. These 'detailers' beat paths from office to office, trying to earn a minute or so of a physician's time to make a sales pitch. In addition to bringing food for the staff and drug samples to be given away to patients, sales representatives bear reprints of the articles their companies have created to give scientific backing to their pitches. An article is a gift of knowledge, an apparent step away from the normal commercialism of the representatives. The knowledge not only buttresses sales jobs, but its status as a gift provokes a return, in the form of prescriptions (Oldani, 2004).

Sales representatives have limited status and influence, so the pharmaceutical industry often develops and turns to 'key opinion leaders,' typically referred to as 'KOLs,' to disseminate scientific information (Sismondo, 2013). The term 'key opinion leader' derives from research by the sociologist Paul Lazarsfeld and his students on structures of influence in such domains as politics, fashion, culture, and medicine. As the term itself suggests, pharmaceutical companies hire KOLs primarily to influence others, to lead opinions in the directions that the companies prefer. For this reason, relations between the companies and KOLs are ideally, from the point of view of the companies, part of general 'KOL management' plans.

There are different kinds of KOLs. At the low end of the totem pole are ordinary physicians – whether general practitioners or specialists – who are paid to speak to other physicians. They become part of the 'speakers bureaus' for particular drugs, and give talks in clinics over lunchtime to other physicians assembled by sales representatives around a buffet lunch. Or they might be after-dinner speakers at events for physicians, again organized by sales representatives. Higher up on the totem pole, and better paid, are researcher KOLs. These are people who have a significant research profile and whose value to pharmaceutical companies might stem from any number of activities. They might be paid to speak to other researchers, patient groups, or to physicians at continuing medical education sessions. They might be consulted on current medical, marketing, or research issues. They might serve as authors on medical journal articles stemming from company-led research. Or they might contribute to research either by recruiting patients for trials or by initiating their own trials.

One useful way in which researcher KOLs can help establish a pharmaceutical company's preferred knowledge base is by serving as (unpaid) authors on 'ghost managed' manuscripts – manuscripts that pharmaceutical companies control or shape, in ways opaque to readers, through multiple steps in the research, analysis, writing, and publication process (Sismondo, 2009). Although pharmaceutical companies have long used scientific data to convince physicians to prescribe their products, the rise of 'publication planning' over the past three decades has led companies to systematically treat data as an important resource to be managed and marshaled. When potentially major drugs come out, their manufacturers flood medical journals with articles about the disease and the drug.

This is a carefully orchestrated process. Companies propose and design multiple manuscripts around studies by lumping and splitting data. Hired medical writers produce first drafts and edit many papers, and medical education and communication companies expertly shepherd manuscripts through the publication pipeline – not only writing medical manuscripts, but communicating with the formal authors of those manuscripts, submitting them to medical journals where they should have impact, and tracking their progress and effects.

A considerable amount of the medical literature on newly patented drugs is ghost-managed. The CROs hired to run clinical trials have no interest in publishing the results under their own names – they produce data that are wholly owned by their sponsors. Thus, pharmaceutical companies have complete control over roughly half of all clinical trial data. More than 50 agencies advertise publication planning on the Internet. Some boast of having hundreds of employees and handling many hundreds of manuscripts per year. The industry is large enough that there are two competing international associations of

publication planners that organize meetings and seminars. One of those associations, the International Society of Medical Planning Professionals, has over 1000 members. Both ISMPP and its competing association, The International Publication Planning Association, hold annual conferences, and the latter hosts regional conferences. Planners handle dozens of manuscripts per year. From the few cases where there is hard data, it appears that roughly 40% of medical journal articles mentioning major new in-patent drugs are parts of individual publication plans on the drugs (Healy and Cattell, 2003; Ross et al., 2008). This level allows a company to attract interest in a drug and shape perceptions of it under the names of apparently independent authors.

Pharmaceutical companies also provide substantial amounts of funding for continuing medical education (CME) courses in the United States and in some other countries. In the United States, they are legally allowed to provide organization, pay for speakers, help speakers prepare their talks, and provide entertainment for participants. The companies do not control the content of accredited CMEs, but if they have chosen their speakers well, supported those speakers' research, and given speakers templates and slides for earlier talks, these courses will convey the companies' preferred messages.

On the whole, pharmaceutical companies are effective at marketing to physicians, whether via advertisements and sales representatives, or KOLs and other conduits for scientific literature. Marketing directly to consumers is a larger challenge and more complex, because only small portions of the general population are potential consumers for most drugs, and because consumers need prescriptions from physicians for many drugs. In the United States, the world's largest market for drugs, many forms of direct-to-consumer advertising (DTCA) are legal and are widespread. Most other jurisdictions allow only limited forms of DTCA, such as disease state awareness advertisements; these make consumers more aware of symptoms and ailments, and that awareness can lead to increased prescriptions and consumption.

Research, education, and marketing, then, are often fused. When a physician tries to learn about conditions and treatments, she typically turns to agents of pharmaceutical companies, whether they are the local sales reps or researchers and educators sponsored to spread particular messages. This allows the companies to sell without selling, a marketers' ideal.

See also: Biomedical Sciences, History and Sociology of; Depression; Ethics for Biomedical Research Involving Humans: International Guidelines; Medical Profession; Medical Sociology; Medicine, History of Western; Patenting in the Biological Sciences; Research Ethics, Cross-Cultural Dimensions of; Risk, Sociology and Politics of; Risk: Social Psychological Perspectives; Therapeutic Effectiveness Assessment: Approaches in Contemporary Research; Universities and Science and Technology: Europe.

Bibliography

Abadie, R., 2010. The Professional Guinea Pig: Big Pharma and the Risky World of Human Subjects. Duke University Press, Durham, NC.

Abraham, J., 2008. Sociology of pharmaceuticals development and regulation: a realist empirical research programme. Sociology of Health & Illness 30 (6), 869–885.

Abraham, J., Reed, T., 2001. Trading risks for markets: the international harmonisation of pharmaceuticals regulation. Health, Risk & Society 3 (1), 113–128.

CenterWatch, 2009. State of the Clinical Trials Industry: A Sourcebook of Charts and Statistics. CenterWatch, Boston, MA.

Daemmrich, A.A., 2004. Pharmacopolitics: Drug Regulation in the United States and Germany. University of North Carolina Press, Chapel Hill, NC.

Drahos, P., 1995. Global property rights in information: the story of TRIPS at the GATT. Prometheus 13 (1), 6–19.

Dumit, J., 2012. Drugs for Life: How Pharmaceutical Companies Define Our Health. Duke University Press, Durham, NC.

Fisher, J., 2009. Medical Research for Hire: The Political Economy of Pharmaceutical Clinical Trials. Rutgers University Press, New Brunswick, NJ.

Gagnon, M.-A., 2009. The Nature of Capital in the Knowledge-based Economy: The Case of the Pharmaceutical Industry (Ph.D. thesis). York University, Canada, ON.

Greene, J., 2007. Prescribing by Numbers: Drugs and the Definition of Disease. Johns Hopkins University Press, Baltimore, MD.

Healy, D., 2004. Shaping the intimate: influences on the experience of everyday nerves. Social Studies of Science 34 (2), 219–245.

Healy, D., Cattell, D., 2003. Interface between authorship, industry and science in the domain of therapeutics. British Journal of Psychiatry 183, 22–27.

Mirowski, P., Van Horne, R., 2005. The contract research organization and the commercialization of scientific research. Social Studies of Science 35, 503–534.

Moynihan, R., Cassels, A., 2005. Selling Sickness: How the World's Biggest Pharmaceutical Companies are Turning Us All into Patients. Nation Books, New York.

Oldani, M.J., 2004. Thick prescriptions: toward an interpretation of pharmaceutical sales practices. Medical Anthropology Quarterly 18 (3), 325–356.

Petryna, A., 2009. When Experiments Travel: Clinical Trials and the Global Search for Human Subjects. Princeton University Press, Princeton, NJ.

Pollock, A., 2012. Medicating Race: Heart Disease and Durable Preoccupations with Difference. Duke University Press, Durham, NC.

Rodwin, M.A., 2011. Conflicts of Interest and the Future of Medicine: The United States, France, and Japan. Oxford University Press, New York.

Ross, J.S., Hill, K.P., Egilman, D.S., Krumholz, H.M., 2008. Guest authorship and ghostwriting in publications related to rofecoxib: a case study of industry documents from rofecoxib litigation. Journal of the American Medical Association 299 (15), 1800–1812.

Shah, S., 2006. The Body Hunters: Testing New Drugs on the World's Poorest Patients. New Press, New York.

Sismondo, S., 2009. Ghosts in the machine: publication planning in the medical sciences. Social Studies of Science 39 (2), 171–198.

Sismondo, S., 2013. Key opinion leader: and the corruption of medical knowledge: what the Sunshine Act will and won't cast light on. The Journal of Law, Medicine and Ethics 41 (3), 635–643.

Sunder Rajan, K., 2007. Experimental values: Indian clinical trials and surplus health. New Left Review 45, 67–88.

Tobbell, D.A., 2012. Pills, Power, and Policy: The Struggle for Drug Reform in Cold War America. University of California Press, Berkeley, CA.

Whitaker, R., 2010. Anatomy of an Epidemic: Magic Bullets, Psychiatric Drugs, and the Astonishing Rise of Mental Illness in America. Random House, Inc., New York.

Williams, S.J., Martin, P., Gabe, J., 2011. The pharmaceuticalisation of society? A framework for analysis. Sociology of Health & Illness 33 (5), 710–725.

Phenomenology in Sociology

Myron Orleans, California State University, Fullerton, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the analytical approaches to the study of how cognitions and perceptions of social reality are reviewed under the rubric of phenomenological sociology. The roots of this stream of thought are traced through founders, modifiers, and adapters. Analytical and research techniques are presented along with discussions of a variety of applications to substantive areas of study. Recent trends, directions, and potentials in this field of inquiry are assessed.

Introduction

Phenomenology is a movement in philosophy that has been adapted by a number of sociologists and other scholars, and practitioners in the social and behavioral sciences to promote an understanding of the relationship between states of individual consciousness and social life. As an approach within sociology, phenomenology seeks to reveal how human awareness is associated with the production of social action, social situations, and social worlds (Natanson, 1970; Fergeson, 2006). In this presentation, the focus is on sociological phenomenology as an integrative approach linking subjective and objective processes that constitute the social world.

Phenomenology is practiced in sociology, psychology, national, regional and ethnic studies, anthropology, nursing, communications, education, history departments, etc. (Embree et al., 1996; Fergeson, 2006). Perhaps the most remarkable development in the field in the last decade has been the trend of sociological phenomenology to transcend disciplinary borders and expand outward into other related scholarly domains. Social phenomenology is a paradigm within many interacting schools of thought and research approaches evident in many aligned disciplines. A remaining task of sociological phenomenology is to clarify its relationship to philosophy and to other subject areas that apply a variety of strategies in using phenomenological elements (Bird, 2009). There does not yet seem to be a systematic effort to integrate and encompass cross-disciplinary phenomenological studies into a coherent body of investigation. However, phenomenology clearly provides underpinnings for constructionism, constructivism, postmodern and poststructuralist deconstructionism, ethnomethodology, everyday life analysis, conversation and discourse analysis, human sciences, communication studies, routine activities crime theory, and more.

Along with greater specificity and application of methods in focused research, phenomenological sociology is not as much of a debate issue as it had been in the 1970s and 1980s when the field was struggling for recognition. More currently, it is considered a relatively acceptable analytic modality moving away from the margins of scholarly inquiry into a more conventional if not quite a central role. Phenomenological approaches are applied in research methodologies such as ethnography, case study, intensive interviewing, observational study, counseling research, and other qualitative methods.

Background

Phenomenology was initially developed by Edmund Husserl (1859–1938), a German mathematician who posited that the objectivism of science precluded an adequate apprehension of the world (Husserl, 1931, 1970, 2012). He presented various philosophical conceptualizations and techniques designed to locate the sources or essences of reality in human consciousness. Alfred Schutz (1899–1959) brought Husserl's approach into sociology to provide a foundation for Max Weber's theory of action and eventually roused interest in other of the social–behavioral sciences (Schutz, 1967). Schutz distilled from Husserl's rather dense writings a socially relevant approach. Schutz described how subjective meanings give rise to an apparently objective social world (Heiskala, 2011; Schutz, 1962, 1964, 1966, 1970, 1996; Schutz and Luckmann, 1973; Wagner, 1983). Research in this vein has been rich and ongoing (Endress et al., 2011; Langan and Barber, 2011).

Schutz's migration to the United States prior to World War II, along with that of other phenomenologically inclined scholars, resulted in the transmission of this approach to American academic circles and ultimately to render it into its current diverse manifestations. Two fundamental expressions of this approach emerged in the 1960s frequently referred to as 'reality constructionism' and 'ethnomethodology.' Reality constructionism synthesized Schutz's distillation of phenomenology and the corpus of classical sociological thought to account for the possibility of social reality (Berger, 1963, 1967; Berger and Berger, 1972; Berger and Kellner, 1981; Berger and Luckmann, 1966; Potter, 1996; Gergen, 2009). Ethnomethodology integrates the Parsonian concern for social order into phenomenology and examines the means by which actors make ordinary life possible (Garfinkel, 1967; Garfinkel and Sacks, 1970; Francis and Hester, 2004; Pollner, 2012a,b). Reality constructionism and ethnomethodology are recognized to be among the more fertile orientations in the field of sociology (Ritzer, 2008; Nasu et al., 2012).

Additional sources and directions for phenomenological sociology are suggested by a review of Heidegger's approach (Aspers, 2010). Merleau-Ponty's phenomenology has led to a number of interesting studies including one that offers an interesting meditation on dance and digital technology as embodied intentionality (Kozel, 2008). There has been some recent interest in reviving the phenomenological work of

Scheler with his concept of developing 'spirit' as a unifying concept (2008). Lock and Strong (2001) suggest other sources for a diversification of phenomenological thinking in such philosophers as Gadamer (1876–1928), Ricoeur (1913–2005), and Levinas (1906–95).

Phenomenology is used in two basic ways in sociology: (1) to theorize about substantive sociological problems and (2) to enhance the adequacy of sociological research methods. Since phenomenology insists that society is a human construction, sociology itself, and its theories and methods are also examined as constructions (Cicourel, 1964, 1973; Ragin and Amoroso, 2010). Thus, phenomenology seeks to offer a corrective to the field's emphasis on positivist conceptualizations and research methods and measurement practices that may take for granted the very issues that phenomenologists find of interest. Phenomenology presents theoretical techniques and qualitative methods that both illuminate the human meanings of social life and the methods used to engage in sociological study.

Phenomenology has until recently been viewed as a challenger of the more conventional styles of positivistic sociological work and has played a role in emergence of postpositivist research modes. Increasingly, phenomenology is coming to be viewed as an adjunctive or by some as an integral part of the discipline, contributing useful analytical tools to balance objectivist approaches (Aho, 1998; Levesque-Lopman, 1988; Luckmann, 1978; Psathas, 1973; Rogers, 1983). Many research methods' texts in a variety of disciplines now provide extensive explorations of social phenomenological methods lending an increasing degree of scholarly legitimacy to interpretive analysis (Cresswell, 2012; Bernard, 2012).

Techniques

Phenomenological inquiry operates rather differently from conventional social science (Darroch and Silvers, 1982). Phenomenology is a philosophical orientation rather than a conventional theory. It does not generate deductions from propositions that can be empirically tested. It operates more on a metasocial plane, demonstrating its premises through descriptive analyses of the procedures of self, situational, and social constitution. Through its demonstrations, audiences apprehend the means by which phenomena, originating in human consciousness come to be experienced as features of an external world.

Current phenomenological techniques in sociology include the method of 'bracketing' (Bentz, 1995; Hurlburt, 2011; Ihde, 1977). This approach lifts an item under investigation from its meaningful context in the 'common-sense world,' with all judgments and presuppositions suspended. Rather, a reduction is performed in which the item is assessed in terms of how it operates in consciousness: What does the notion do for those who define themselves within its domain? A phenomenological reduction both plummets to the essentials of the notion and ascertains its meanings independent of all particular occasions of its use. The reduction of a bracketed phenomenon is thus a technique to gain new theoretical insight into the meaning and overt products of intentional projections from consciousness. A recent formulation of this technique focused on the bracketing of materiality, technology, and embodiment (Goeminne and Paredis, 2011). Leonardi (2008) explored the implications of deterministic conceptualizations or 'inevitability' on the implementation of new technologies to the benefit of particular actors within an organization.

Phenomenological tools include the use of interpretative, empathic, and introspective methods to offer a detailed description of the external consequences of intentional consciousness (Hitzler and Keller, 1989; Smith et al., 2009). Introspection requires the phenomenologist to use his or her own subjective process as a resource for study, and an empathic effort is needed to apprehend the mind of the other (Helle, 1991; Truzzi, 1974). Interpretation, introspection, and empathic tools of phenomenological analysis are considered to be procedures used both by ordinary individuals to carry out their projects and by researchers to penetrate to their meanings (Horgan, 2011). In this technique, the 'analytic' attitude is systematized as distinct from the 'natural' attitude of everyday life.

Since cognition is a crucial element of phenomenology, some theorists focus on social knowledge as the cornerstone of their technique (Berger and Luckmann, 1966). Frequently, religious thought is given primacy in the study of the sources and legitimations of mundane knowledge, and indirectly social construction (Berger, 1967). It is suggested that common-sense knowledge is produced, disseminated, and internalized. Subsequently, such knowledge is externalized from consciousness to be experienced ultimately as objectified phenomena. Thus, historically embedded meanings and discursive practices enable us to continually create and reinforce taken-for-granted knowledge or common sense. Common sense is then used as a means to pattern and institutionalize behaviors. From this angle, 'deconstructionism' may be viewed as a recursive analysis of this constructive process and we may be able to talk about a 'deconstructive phenomenology' (Watkin, 2009).

Recent work shows how phenomenological analysis extends cognition into the world (Rowlands, 2010). This approach tries to reject the concept of the mind as a purely internal process and seeks to locate the mind as a tied to externalities outside the body, and that mind and activity are reflexively integrated in constructive activity.

Phenomenological concerns are most frequently researched using qualitative methods (Bernard, 2012; Bogdan and Taylor, 1975; Cresswell, 2012; Denzin and Lincoln, 1994, 1998). Phenomenological researchers frequently undertake analyses of small groups, social situations, and organizations using face-to-face techniques of observation and intensive interviewing (Bruyn, 1966; Psathas and Ten Have, 1994; Seidman, 2005; Turner, 1974). Ethnographic research frequently utilizes phenomenological tools (Fielding, 1988). Intensive interviewing to uncover the subject's orientations or his or her 'life world' is also widely practiced (Costelloe, 1996; Englander, 2012; Grekova, 1996; Porter, 1995). Qualitative tools are used in phenomenological research either to yield insight into the microdynamics of particular spheres of human life for its own sake or to exhibit the constitutive activity of human consciousness (Langsdorf, 1995; Lapan et al., 2011).

Techniques particular to the ethnomethodological branch of phenomenology have been developed to reveal the practices

used by people to produce a sense of social order and thereby accomplish everyday life (Cuff, 1993; Eberle, 2010; Leiter, 1980; Mehan and Wood, 1975; Rouncefield and Tomie, 2011). At one time, 'breaching demonstrations' were widely conducted to reveal the essentiality of taken-for-granted routines and the means by which threats to these routines were managed. Since breaching these routines sometimes resulted in serious disruptions of relationships, this technique has been employed more recently only with great caution in research and instruction (Rafalovich, 2006). Frequently, social situations are video- and audiotaped to permit the painstaking demonstration of the actions by which participants produce themselves, their apparent interpretations of the meanings of acts, and to ascertain how they structure a situation (Blum-Kulka, 1994; Jordan and Henderson, 1995; Ten Have, 2004). Conversational analysis has been used widely to describe how people make sense of each other through talk and how they make sense of their talk through their common background knowledge (Psathas, 1994; Schegloff and Sacks, 1974; Silverman, 1998; Wardhaugh, 2009). The interrelations between mundane and abstract reasoning are also examined in great depth as researchers expose, for example, the socially constituted bases of scientific and mathematical practice in common-sense thinking (Dennis, 2004; Knorr-Cetina and Mulkay, 1983; Livingston, 1995; Lynch, 1993).

Theory

The central task in social phenomenology is to demonstrate the reciprocal interactions among the processes of human action, situational structuring, and reality construction. Rather than contending that any aspect is a causal factor, phenomenology views all dimensions as constitutive of all others. Phenomenologists use the term 'reflexivity' to characterize the way in which constituent dimensions serve as both foundation and consequence of all human projects. The task of phenomenology, then, is to make manifest the incessant tangle or reflexivity of action, situation, and reality in the various modes of being in the world (Phillipson, 1972).

Phenomenology commences with an analysis of the 'natural attitude.' This is understood as the way ordinary individuals participate in the world, taking its existence for granted, assuming its objectivity, and undertaking action projects as if they were predetermined. Language, culture, and common sense are experienced in the natural attitude as objective features of an external world that are learned by actors in the course of their lives.

Human beings are open to patterned social experience and strive toward meaningful involvement in a knowable world. They are characterized by a typifying mode of consciousness tending to classify sense data (Kim and Berard, 2009). In phenomenological terms humans experience the world in terms of 'typifications': Children are exposed to the common sounds and sights of their environments, including their own bodies, people, animals, vehicles, and so on. They come to apprehend the categorical identity and typified meanings of each in terms of conventional linguistic forms. In a similar manner, children learn the formulas for doing common activities. These practical means of doing are called 'recipes for action.' Typifications and recipes, once internalized, tend to settle beneath the level of full awareness, which become 'sedimented,' as do layers of rock. Thus, in the natural attitude, the foundations of actors' knowledge, and use of meanings and actions are obscured to the actors themselves.

Actors assume that knowledge is objective and all people reason in a like manner. Each actor assumes that every other actor knows what he or she knows of this world: All believe that they share and all know a common sense. However, each person's biography is unique, and each develops a relatively distinct 'stock' of typifications and recipes. Therefore, interpretations may diverge. Everyday social interaction is replete with ways in which actors create the feeling and belief that common sense is shared, that mutual understanding is occurring, and that everything is OK. Phenomenology emphasizes that humans live within an intersubjective world, yet they at best approximate shared realities. While a paramount reality is commonly experienced in this manner, particular realities or 'finite provinces of meaning' are also constructed and experienced by diverse cultural, social, or occupational groupings.

For phenomenology, all human consciousness is practical – it is always of something. Actors 'intend' projects into the world; they act in order to implement goals based on their typifications and recipes, their stock of knowledge at hand. Consciousness as an 'intentional' process is composed of thinking, perceiving, feeling, remembering, imagining, and anticipating directed toward the world. The objects of consciousness, these intentional acts, are the sources of all social realities that are, in turn, the materials of common sense.

Thus, typifications derived from common sense are internalized, becoming the tools that individual consciousness uses to constitute a 'life world,' the unified arena of human awareness and action. Common sense serves as an ever-present resource to assure actors that the reality that is projected from human subjectivity is an objective reality. Since all actors are involved in this intentional work, they sustain the collaborative effort to 'reify' their projections and thereby reinforce the very frameworks that provide the construction tools.

Social interaction is viewed phenomenologically as a process of reciprocal interpretive constructions of actors applying their stock of knowledge at hand to the occasion. Interactors orient themselves to others by taking into account typified meanings of actors in typified situations known to them through common sense. Action schemes are geared by each to the presumed projects of others. The conduct resulting from the intersection of intentional acts indicates to members of the collectivity that communication, coordination, or something of the like is occurring among them. For these members, conduct and utterances serve as 'indexical' expressions of the properties of the situation enabling each to proceed with the interaction while interpreting others, context, and self. Through the use of certain interpretive practices, members order the situation for themselves in sensical and coherent terms: In their talk they 'gloss' over apparent irrelevancies, fill in innumerable gaps, ignore inconsistencies, and assume a continuity of meaning, thereby formulating the occasion itself.

Ongoing social situations manifest patterned routine conduct that appears to positivist investigators to be normative or rule guided. Phenomenologically, rules are indexical

expressions of the interpretive processes applied by members in the course of their interactions. Rules are enacted in and through their applications. In order to play by the book, the interpreter endeavors to use the rule as an apparent guide. However, he or she must use all sorts of background expectancies to somehow manage the fit between the particular and the general under the contexted conditions of the interaction, and in so doing is acting creatively. Rules, policies, hierarchy, and organization are accomplished through the interpretive acts or negotiations of members in their concerted efforts to formulate a sense of operating in accord with a rational, accountable system. This work of 'doing' structure to the situation further sustains its common-sensical foundations as well as its apparent 'facticity.'

Phenomenologists analyze the ordering of social reality and how the usage of certain forms of knowledge contributes to that ordering. Typified action and interaction become routinely accomplished or 'habitualized' behavior. Through sedimentation in layered consciousness, human authorship of habitualized conduct is obscured and the product is experienced as an external entity, a 'thing.' As meaning-striving beings, humans create theoretical explanations and moral justifications in order to 'legitimate' or justify their habitualized conduct as something other or even more than human products. Some of the more abstract justifications become legitimations and are located in higher contexts of meaning such as religious or patriotic belief. Thus, the conduct becomes 'objectivated.' When internalized by succeeding generations, the conduct becomes fully institutionalized and exerts compelling constraints or normative pressure over individual volition. Periodically, the institutions might be threatened by dissent or aberrant conduct. The response might be then to sustain the institution by explaining the variance as an individual deviation requiring that the individual be 'repaired' through therapy, treatment, punishment, resocialization, or some combination of means primarily designed to restore the sense of institutional normalcy. Thus, straying individuals who cognitively, affectively, or behaviorally 'migrate' or stray from institutional alignment are corrected.

The reality that ordinary people inhabit is constituted by these legitimations of habitualized conduct. Ranging from common-sense typifications of ordinary language explanations to theological constructions to sophisticated philosophical, cosmological, and scientific conceptualizations, these reality justifications or legitimations contribute to the maintenance of the paramount reality of everyday life. Moreover, segmented modern life, with its proliferation of meaning-generating sectors, produces multiple realities, some in competition with each other for adherents. In the current marketplace of realities, individual reality consumers, to varying degrees, may select their meaning sectors or legitimations, as they select their occupation, spouse, lifestyle, and possibly, their religion (Berger, 1967).

Practice

Doing phenomenological sociology involves procedures that are distinct from positivist research. Phenomenological practice is increasingly evident in sociology and related disciplines as more subjectivist work is published. The phenomenological analysis of mass media culture content, for example, applies the elements of the approach to yield an understanding of the reflexive interplay of audience's life worlds and program material (Wilson, 1996; Groys, 2012). Thus, TV talk show discourses or celebrity profiling may be described as social texts that are refracted by programmers from common-sense identity constructs. The visual realization yields narrative images that audiences are seduced into processing using their own experiences. The viewers' life worlds and the TV representations are blended into reality proxies that provide viewers with schema to configure their personal orientations. Subsequently, programmers draw upon these orientations as additional identity materials for a new content development.

Phenomenological work with young children examined how both family interactions and the practices of everyday life are related to the construction of childhood (Davilla and Pearson, 1994). Penetrating the inner world of children requires that the phenomenological practitioners view the subjects in their own terms, from the level and viewpoints of children (Waksler, 1991; Shehan, 1999). Such investigation shuns adult authoritative and particularly scientific perspectives and seeks to give voice to the children's experience of their own worlds. Infants' and children's communicative and interactive competencies are respected and are not diminished by their movement toward higher level functioning (Sheets-Johnstone, 1996).

At the other end of the lifecycle, phenomenologists investigate how aging and its associated challenges are constituted in the consciousness of members and helpers (Starr, 1982). The struggle for meaning, memory, and identity during aging accompanied by deteriorative changes may be facilitated by the availability of constructs that permit the smoother processing of declining capabilities (Li and Orleans, 2002). Members of cultures that stock typifications and recipes for managing aging and pain effectively may well be more likely than others to construct beneficial interpretations in the face of the challenges posed by aging (Encandela, 1997). Phenomenological work encourages the helpers of the elderly to gain empathic appreciation of their clients' life worlds and to enhance the quality of their care through the use of biographical narratives that highlight the individuality and humanity of aging persons (Heliker, 1997). Caregivers for the elderly who were able to adequately manage their feelings of powerlessness and incompetence in relation to their work ideals could be more effective (Juthberg and Sundin, 2010).

The healing professions, particularly nursing, seem to be deeply imbued with a phenomenological focus on the provision of care based on a rigorous emphasis on the patient's subjective experience (Benner, 1995; Edwards, 2003). Substantial attention has been devoted to the ethical implications of various disease definitions, to how language shapes the response to illness, and to how disease definitions and paradigmatic models impact communication between health professionals and patients (Rosenberg and Golden, 1992). Recent applications of social phenomenology in relation to nursing analyzed the construction of bioethical concerns (Shaw, 2010), and the health care consultation process was dissected using video to develop a schema to enhance the quality of communication (Bickerton et al., 2010).

Significant work on the phenomenology of disability has demonstrated how the lived body is experienced in an altered form and how taken-for-granted routines are disrupted invoking new action recipes (Toombs, 1995). Interpretive research with autistic young people has been successfully conducted (Huws, 2008). Nonconventional healing practices have also been examined revealing how embodiment and the actor's subjective orientation reflexively interrelate with cultural imagery and discourse to transfigure the self (Csordas, 1997). Further, phenomenological work has suggested that emotions are best analyzed as interpreted processes both embodied and embedded within experiential contexts (Blum, 1996; Solomon, 1997).

Studies of everyday life activities from a phenomenological foundation are abundant and robust in sociology (Pink, 2012). A number of introductory textbooks focusing on ordinary life have been published for decades with regular new editions (Henslin, 2012; Newman, 2011). Scott's (2009) text introduces students to the social phenomenology of everyday life. Scholarly journals, many of which are included in the references below, such as *Qualitative Research*, *Human Studies*, and *The Journal of Contemporary Ethnography* regularly include articles on the commonplace as constructive human endeavors.

Conclusions

For phenomenology, society, social reality, social order, institutions, organizations, situations, interactions, and individual actions are constructions that appear as suprahuman entities. What does this suggest regarding humanity and sociology? Phenomenology advances the notion that humans are creative agents in the construction of social worlds (Ainlay, 1986). It is from their consciousness that all being emerges.

Sociological phenomenology of the latter part of the twentieth century laid out the philosophical foundations of a program of theorizing and research that has been substantively addressed in the early part of the twenty-first century. Numerous studies employing phenomenological tools have demonstrated the utility of the approach and the steady progress of the field suggests that much more is to come. While remaining a lesser trend within the predominantly objectivist quantitatively oriented discipline of sociology, it has gained more than a measure of legitimacy and acceptance by the broader community of scholars.

While there are not many departments of sociology where phenomenological approaches are predominant, especially in the United States, many if not most departments do include at least one or a few phenomenologically inclined scholars. Some Canadian, European, and some non-Western universities in Asia and Latin America may even include a substantial number of phenomenologists. If further study is desired, readers are advised to review the authors referred to in this paper to consider the range of affiliations and the types of programs offered at various universities.

The future impact of phenomenology is reflected in its resonance with the interests and orientations of a substantial group of sociologists and practitioners in related disciplines. The goal of deriving a penetrating and productive insight into the practical consequences of consciousness with the precision and sensitivity of demanding scholarship is progressing well. The applications of phenomenology to diverse areas, only some of which are identified in this paper, hold great promise of substantial contribution to human understanding.

Among the frontiers of phenomenological analysis that could be more fully examined remain studies of the experiences of virtuality, cyberspace, and computer simulacra, the roots of scientific and technological formulations in everyday meanings and practices, media imagery in relation to explanations of unusual events, representations of celebrities and identity formation, ethnic and racial identity constructions, diverse family patterning, educational assessment practices, crime meanings, and of other as yet unmentioned domains of knowledge.

Due to its lack of presumption, emphasis on subjectivity, its flexible applicability, and its many streams of phenomenological inspiration, this approach has proven beneficial in the opening years of the twenty-first century, and it is likely to continue developing by specialization and focus as we move deeper into the new millennium.

See also: Action, Theories of Social; Conversation Analysis; Knowledge, Sociology of; Neurosociology; Postmodernism in Sociology; Relativism, Cognitive.

Bibliography

Aho, J.A., 1998. The Things of the World: A Social Phenomenology. Praeger, Westport, Conn.

Ainlay, S.C., 1986. The Encounter with phenomenology. In: Hunter, J.D., Ainlay, S.C. (Eds.), Making Sense of Modern Times: Peter L. Berger and the Vision of Interpretive Sociology. Routledge and Kegan Paul, London, pp. 31–56.

Aspers, P., 2010. The second road to phenomenological sociology. Society 47 (3), 214–219.

Benner, P., 1995. Interpretive Phenomenology: Embodiment, Caring and Ethics in Health and Illness. Sage Publications, Thousand Oaks, CA.

Bentz, V.M., 1995. Experiments in Husserlian phenomenological sociology. Studies in Symbolic Interaction 17, 133–161.

Berger, P.L., 1963. Invitation to Sociology: A Humanistic Perspective. Anchor, New York.

Berger, P.L., 1967. The Sacred Canopy: Elements of a Sociological Theory of Religion. Doubleday, New York.

Berger, P.L., Berger, B., 1972. Sociology: A Biographical Approach. Basic Books, New York.

Berger, P.L., Kellner, H., 1981. Sociology Reinterpreted. Doubleday, New York.

Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.

Bernard, H.R., 2012. Social Research Methods: Qualitative and Quantitative Approaches. Sage, Thousand Oaks, CA.

Bickerton, J., Procter, S., Johnson, B., Medina, A., 2010. A video life-world approach to consultation practice: the relevance of a socio-phenomenological approach. Human Studies 33, 157–171.

Bird, G., 2009. What is phenomenological sociology again? Human Studies 32 (4), 419–439.

Blum, A., 1996. Panic and fear: on the phenomenology of desperation. The Sociological Quarterly 37, 673–698.

Blum-Kulka, S., 1994. The dynamics of family dinner talk: cultural contexts for children's passages to adult discourse. Research on Language and Social Interaction 27, 1–50.

Bogdan, R., Taylor, S.J., 1975. Introduction to Qualitative Research Methods: A Phenomenological Approach to the Social Sciences. Wiley, New York.

Bruyn, S.T., 1966. The Human Perspective in Sociology: The Methodology of Participant Observation. Prentice-Hall, Englewood Cliffs, NJ.

Cicourel, A.V., 1964. Method and Measurement in Sociology. Free Press, New York.

Cicourel, A.V., 1973. Theory and Method in a Study of Argentine Fertility. Wiley, New York.

Costelloe, T.M., 1996. Between the subject and sociology: Alfred Schutz's phenomenology of the life-world. Human Studies 19, 247–266.

Csordas, T.J., 1997. The Sacred Self: A Cultural Phenomenology of Charismatic Healing. University of California Press, Berkley, CA.

Cuff, E.C., 1993. Problems of Versions in Everyday Situations. University Press of America, Lanham, Maryland.

Cresswell, J.W., 2012. Qualitative Inquiry and Research Design: Choosing Among Five Approaches. Sage Publications, Thousand Oaks, CA.

Darroch, V., Silvers, R.J. (Eds.), 1982. Interpretive Human Studies: An Introduction to Phenomenological Research. University Press, Washington, DC.

Davilla, R.A., Pearson, J.C., 1994. Children's perspectives of the family: a phenomenological inquiry. Human Studies 17, 325–341.

Dennis, A., 2004. Lynch on Schutz and science: postanalytic ethnomethodology reconsidered. Theory & Science 5 (1).

Denzin, N.K., Lincoln, Y.S., 1994. Handbook of Qualitative Research. Sage Publications, Thousand Oaks, CA.

Denzin, N.K., Lincoln, Y.S. (Eds.), 1998. Collecting and Interpreting Qualitative Materials. Sage Publications, Thousand Oaks, CA.

Eberle, T., 2010. The phenomenological life-world analysis and the methodology of the social sciences. Human Studies 33 (2–3), 123–139.

Edwards, C., 2003. Research into patients' perspectives: relevance and usefulness of phenomenological sociology. Journal of Advanced Nursing 44 (5), 450–460.

Embree, L., (Adapter) et al., 1996. In: Encyclopedia of Phenomenology (Contributions to Phenomenology). Springer, New York.

Encandela, J.A., 1997. Social construction of pain and aging: individual artfulness within interpretive structures. Symbolic Interaction 20, 251–273.

Endress, M., Pathas, G., Nasu, H. (Eds.), 2011. Explorations of the Life-World: Continuing Dialogues with Alfred Schutz (Contributions to Phenomenology). Springer, New York.

Englander, M., 2012. The interview: data collection in descriptive phenomenological human scientific research. Journal of Phenomenological Psychology 43 (1), 13–35.

Fielding, N.G. (Ed.), 1988. Actions and Structure: Research Methods and Social Theory. Sage Publications, London.

Fergeson, H., 2006. Phenomenological Sociology: Experience and Insight in Modern Society. Sage Publications, Thousand Oaks, CA.

Francis, D., Hester, S., 2004. An Invitation to Ethnomethodology: Language, Society and Interaction. Sage Publications, Thousand Oaks, CA.

Garfinkel, H., 1967. Studies in Ethnomethodology. Prentice-Hall, Englewood Cliffs, NJ.

Garfinkel, H., Sacks, H., 1970. The Formal properties of practical actions. In: McKinney, J.C., Tiryakian, E.A. (Eds.), Theoretical Sociology. Appleton-Century-Crofts, New York.

Gergen, K., 2009. An Invitation to Social Construction. Sage Publications, Thousand Oaks, CA.

Goeminne, G., Paredis, E., 2011. Opening up the in-between: Ihde's postphenomenology and beyond. Foundations of Science 16, 101–107.

Grekova, M., 1996. Restructuring of 'the life-world of Socialism'. International Sociology 11, 63–78.

Groys, B., 2012. Under Suspicion: A Phenomenology of Media. Columbia University Press, New York.

Heiskala, R., 2011. The meaning of meaning in sociology. The achievements and shortcomings of Alfred Schutz's phenomenological sociology. Journal for the Theory of Social Behaviour 41 (3), 231–246.

Heliker, D.M., 1997. A narrative approach to quality care in long-term care facilities. Journal of Holistic Nursing 15, 68–81.

Helle, H.J. (Ed.), 1991. Verstehen and Pragmatism: Essays in Interpretative Sociology. Peter Lang, Frankfurt, Federal Republic Germany.

Henslin, J.M., 2012. Essentials of Sociology: A Down-to-Earth Approach. Pearson, Old Tappan, NJ.

Hitzler, R., Keller, R., 1989. On sociological and common-sense verstehen. Current Sociology 37, 91–101.

Horgan, T., 2011. Introspection and the phenomenology of free will: problems and prospects. Journal of Consciousness Studies 18 (1), 180–205.

Hurlburt, R., 2011. Agency: a case study in bracketing presuppositions. Journal of Consciousness Studies 18 (1), 295–305.

Husserl, E., 1931. Ideas: General Introduction to Pure Phenomenology (W.R.B. Gibson, Trans.). Humanities Press, New York.

Husserl, E., 1970. The Crisis of European Sciences and Transcendental Phenomenology (D. Carr, Trans.). Northwestern University Press, Evanston, Ill.

Husserl, E., 2012. Ideas: General Introduction to Pure Phenomenology. Routledge, London.

Huws, J., 2008. Diagnosis, disclosure, and having autism: an interpretative phenomenological analysis of the perceptions of young people with autism. Journal of Intellectual & Developmental Disability 33 (2), 99–107b.

Ihde, D., 1977. Experimental Phenomenology: An Introduction. Putnam, New York.

Jordan, B., Henderson, A., 1995. Interaction analysis: foundations and practice. The Journal of the Learning Sciences 4, 39–103.

Juthberg, C., Sundin, K., 2010. Registered nurses' and nurse assistants' lived experience of troubled conscience in their work in elderly care–a phenomenological hermeneutic study. International Journal of Nursing Studies 47 (1), 20–29.

Kim, K., Berard, T., 2009. Typification in society and social science: the continuing relevance of Schutz's social phenomenology. Human Studies 32 (3), 263–289.

Knorr-Cetina, K., Mulkay, M.J. (Eds.), 1983. Science Observed. Sage Publications, London.

Kozel, S., 2008. Closer: Performance, Technologies, Phenomenology. The MIT Press, Cambridge, MA.

Langsdorf, Lenore, 1995. Treating method and form as phenomena: an appreciation of Garfinkel's phenomenology of social action. Human Studies 18, 177–188.

Langan, R.L., Barber, M., 2011. Phenomenology of the Human Sciences. In: Schutzian Research: A Yearbook of Worldly Phenomenology and Qualitative Social Science, vol. 3. Zeta Books, Bucharest, Hungary.

Lapan, S.D., MaryLynn, T., Quartaroli, M.T., Riemer, F.J. (Eds.), 2011. Qualitative Research: An Introduction to Methods and Designs (Research Methods for the Social Sciences). Jossey-Bass, Hoboken, NJ.

Leiter, K., 1980. A Primer on Ethnomethodology. Oxford University Press, New York.

Leonardi, P.M., 2008. Indeterminancy and the discourse of inevitability in international technology management. Academy of Management Review 33 (4), 975–984.

Levesque-Lopman, L., 1988. Claiming Reality: Phenomenology and Women's Experience. Rowman and Littlefield, Totowa, NJ.

Li, R., Orleans, M., 2002. Personhood in a world of forgetfulness: an ethnography of the self-process among Alzheimer's patients. Journal of Aging & Identity 7 (4), 227–244.

Livingston, E., 1995. The idiosyncratic specificity of the methods of physical experimentation. The Australian and New Zealand Journal of Sociology 31 (3), 1–21.

Lock, A., Strong, T., 2001. Social Constructionism: Sources and Stirrings in Theory and Practice. Cambridge University Press, Cambridge.

Luckmann, T. (Ed.), 1978. Phenomenology and Sociology: Selected Readings. Penguin, New York.

Lynch, M., 1993. Scientific Practice and Ordinary Action: Ethnomethodology and Social Studies of Science. Cambridge U. Press, Cambridge, England.

Mehan, H., Wood, H., 1975. The Reality of Ethnomethodology. Wiley, New York.

Nasu, H., Chaput Waksler, F., Papadimitriou, C., Rehorick, D., 2012. Interaction and Everyday Life: Phenomenological and Ethnomethodological Essays in Honor of George Psathas. Lexington Books, New York.

Natanson, M., 1970. Alfred Schutz on social reality and social science. In: Natanson, M. (Ed.), Phenomenology and Social Reality. Nijhoff, The Hague.

Newman, D.M., 2011. Sociology: Exploring the Architecture of Everyday Life. SAGE Publications, Thousand Oaks, CA.

Phillipson, M., 1972. Phenomenological philosophy and sociology. In: Filmer, P., Phillipson, M., Silverman, D., Walsh, D. (Eds.), New Directions in Sociological Theory. MIT Press, Cambridge, Mass.

Pink, S., 2012. Situating Everyday Life: Practices and Places. Sage Publications, Thousand Oaks, CA.

Pollner, M., 2012a. Reflections on Garfinkel and ethnomethodology's program. American Sociologist 43 (1), 36–54.

Pollner, M., 2012b. The end(s) of ethnomethodology. American Sociologist 43 (1), 7–20.

Porter, E.J., 1995. The life-world of older widows: the context of lived experience. Journal of Women and Aging 7, 31–46.

Potter, J., 1996. Representing Reality: Discourse, Rhetoric and Social Construction. Sage Publications, London.

Psathas, G. (Ed.), 1973. Phenomenological Sociology: Issues and Applications. Wiley, New York.

Psathas, G., 1994. Conversation Analysis: The Study of Talk-in-Interaction. Sage, Thousand Oaks, CA, and London.

Psathas, G., Ten Have, P., 1994. Situated Order: Studies in the Social Organization of Talk and Embodied. University Press of America, Lanham, MD.

Rafalovich, A., 2006. Making sociology relevant: the assignment and application of breaching experiments. Teaching Sociology 34 (2), 156–163.

Ragin, C.C., Amoroso, L.M., 2010. Constructing Social Research: The Unity and Diversity of Method. Sage Publications, Thousand Oaks, CA.

Ritzer, G., 2008. Modern Sociological Theory. McGraw-Hill, New York.

Rogers, M.F., 1983. Sociology, Ethnomethodology, and Experience: A Phenomenological Critique. Cambridge University Press, New York.

Rosenberg, C.E., Golden, J. (Eds.), 1992. Framing Disease: Studies in Cultural History. Rutgers University Press, New Brunswick, NJ.

Rouncefield, M., Tomie, P., 2011. Ethnomethodology at Work (Directions in Ethnomethodology and Conversation Analysis). Ashgate Publishing, Burlington, VT.

Rowlands, M., 2010. The New Science of the Mind: From Extended Mind to Embodied Phenomenology. Bradford Books, Cambridge, Mass.

Schegloff, E., Sacks, H., 1974. Opening up closings. In: Turner, R. (Ed.), Ethnomethodology. Penguin, Baltimore.
Scheler, M., 2008. The Human Place in the Cosmos. Northwestern University Press, Evanston, IL.
Schutz, A., 1962. In: Natanson, M. (Ed.), Collected Papers I: The Problem of Social Reality. Nijhoff, The Hague.
Schutz, A., 1964. In: Brodersen, A. (Ed.), Collected Papers II: Studies in Social Theory. Nijhoff, The Hague.
Schutz, A., 1966. Collected Papers III: Studies in Phenomenological Philosophy. Nijhoff, The Hague.
Schutz, A., 1967. The Phenomenology of the Social World (G. Walsh and F. Lehnert, Trans.). Northwestern University Press, Evanston, Ill.
Schutz, A., 1970. In: Zaner, R. (Ed.), Reflections on the Problem of Relevance. Yale University Press, New Haven.
Schutz, A., Luckmann, T., 1973. The Structure of the Life World. Northwestern University Press, Evanston, Ill.
Schutz, A., 1996. Collected papers IV. In: Wagner, H., Psathas, G., Kersten, F. (Eds.), 1996. Phaenomenologica Series, vol. 136. Kluwer Academic, Boston and London.
Scott, S., 2009. Making Sense of Everyday Life. Polity Press, Cambridge, UK.
Seidman, I., 2005. Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences. Teachers College Press, New York.
Shaw, R., 2010. Phenomenological sociology and the sociology of bioethics: two New Zealand studies. Women's Studies Journal 24 (2), 40–53.
Sheets-Johnstone, M., 1996. An empirical-phenomenological critique of the social construction of infancy. Human Studies 19, 1–16.
Shehan, C.L. (Ed.), 1999. Through the Eyes of the Child: Revisioning Children as Active Agents of Family Life. JAI Press, Greenwich, Conn.
Silverman, D., 1998. Harvey Sacks: Social Science and Conversation Analysis. Oxford University Press, New York.
Smith, J.A., Flowers, P., Larkin, M., 2009. Interpretative Phenomenological Analysis: Theory, Method and Research. Sage Publications, Thousand Oaks, CA.
Solomon, R.C., 1997. Beyond ontology: ideation, phenomenology and the cross cultural study of emotion. Journal for the Theory of Social Behaviour 27, 289–303.
Starr, J.M., 1982. Toward a social phenomenology of aging: studying the self process in biographical work. International Journal of Aging and Human Development 16, 255–270.
Ten Have, P., 2004. Understanding Qualitative Research and Ethnomethodology. Sage Publications, Thousand Oaks, CA.
Toombs, S.K., 1995. The lived experience of disability. Human Studies 18, 9–23.
Truzzi, M., 1974. Verstehen: Subjective Understanding in the Social Sciences. Addison-Wesley, Reading, Mass.
Turner, R. (Ed.), 1974. Ethnomethodology: Selected Readings. Penguin, Baltimore.
Waksler, F.C. (Ed.), 1991. Studying the Social Worlds of Children: Sociological Readings. The Falmer Press, London.
Wagner, H.R., 1983. Alfred Schutz: An Intellectual Biography. University of Chicago Press, Chicago.
Watkin, C., 2009. Phenomenology or Deconstruction?: The Question of Ontology in Maurice Merleau-Ponty, Paul Ricoeur, and Jean-Luc Nancy. Edinburgh University Press, Edinburgh.
Wardhaugh, R., 2009. An Introduction to Sociolinguistics. Wiley-Blackwell, Hoboken, NJ.
Wilson, T., 1996. Television's everyday life: towards a phenomenology of the 'televisual subject'. Journal of Communication Inquiry 20, 49–66.

Relevant Websites

http://www.phenomenologyonline.com/ – Phenomenology Online. A Resource for Phenomenological Inquiry.
www.phenomenologycenter.org/ – The Center for Advanced Research in Phenomenology.
http://www.phenomenology.org/index.html – World Phenomenology Institute, also known as the World Institute for Advanced Phenomenological Research and Learning.
http://sphs.info/ – The Society for Phenomenology and the Human Sciences.
http://societyphenmedia.wix.com/socphenmedia – The Society for Phenomenology and Media.
http://appliedphenom.org/ – The Institute for Applied Phenomenology in Science and Technology.
http://sifp.uoregon.edu/ – The Society for Interdisciplinary Feminist Phenomenology.

Phenomenology: Philosophical Aspects

Christian Grüny, Universität Witten/Herdecke, Witten, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by K. Mulligan, volume 17, pp. 11363–11369, © 2001, Elsevier Ltd.

Abstract

Even though phenomenology had a single founder, Edmund Husserl, it has developed into a diverse tradition that warrants speaking of phenomenologies in the plural. Transforming the objects and actions of everyday experience into phenomena, it aims at clarifying the ways of givenness of the world in relation to an experiencing subject. From the very beginning, sociality has been among its areas of research, ranging from the question of other minds and emotions and of embodiment as the basic human condition, to an investigation of the lifeworld as the ground and horizon of all experience and the other as a fundamental challenge to the autonomous subject. In this way, it has inspired research in various other disciplines including cognitive science.

Phenomenology and Phenomenologies

Phenomenology was one of the most influential philosophical traditions of the twentieth century and continues to influence philosophical thought and various other fields (sociology, political theory, art studies, theory of architecture, pedagogy, among others) today. While its roots lie in the same contexts and discussions in philosophy, logic, and psychology in the nineteenth century that were the foundations of analytic philosophy, this tie was severed long ago. Only recently the debate in the cognitive sciences has revitalized phenomenology in an area that had long been dominated by analytic approaches, promising a new dialogue between the two. Besides this, there have been various parallels to and convergences with developments in the pragmatist tradition.

Unlike many other philosophical traditions, phenomenology can be said to have a single founder: Edmund Husserl. Husserl (1859–1938) studied philosophy, mathematics, and psychology and devoted his first books to the theory of mathematics. In these books, he defended a theory that he was to oppose vehemently in later years, namely a psychologistic explanation of mathematics and logic. The publication of his *Logical Investigations* 1900–1901 amounted to a complete reversion of this earlier position, and it marks the beginning of phenomenology in the modern sense, which has little in common with earlier uses of the word (e.g., in Hegel's *Phenomenology of Spirit*). Husserl's basic insight was inspired by one of his philosophical teachers, Franz von Brentano: the fact of intentionality. Consciousness is always consciousness of something, just as symbolic reference is reference to something. Any investigation of the objects of consciousness has to be complemented by an investigation of the consciousness of objects and vice versa. In this relation, the objects are not constructed by consciousness but rather intuited as independent, and the task of phenomenology is to investigate these objects – among them logical and mathematical sentences – in the way they are given as phenomena (Husserl, 1901: pt. V).

Since the publication of the *Logical Investigations*, phenomenology has undergone several fundamental transformations, and as early as 1953 Paul Ricoeur called its history a history of heresies (Ricœur, 1953: p. 836). Considering this, it might be more appropriate to speak of phenomenologies in the plural. The first of these heretics, however, was Husserl himself, whose transcendental turn in 1913 many of his earlier followers and collaborators (Adolph Reinach, Moritz Geiger, Max Scheler, among others) were not willing to go along with. The 'realist phenomenology' they adhered to did not consider the meticulous analysis of the a priori structures of consciousness Husserl undertook in the first volume of his *Ideas Pertaining to a Pure Phenomenology and Phenomenological Philosophy* (Husserl, 1913) to be warranted or necessary. Instead, they devoted their attention to social and legal phenomena, art, literature, and a variety of other topics that appeared in Husserl's own writings only much later. The major figure after Husserl who even tended to outshine his former teacher appeared on the scene in the 1920s: Martin Heidegger. His existential and ontological transformation of phenomenology paved the ground for subsequent modifications that are associated with names like Jean-Paul Sartre and Maurice Merleau-Ponty who stressed the role of the body and brought into play political topics. Phenomenological heritage can even be detected in those French philosophers whose background lay in phenomenology but who decided to distance themselves from it, including Jacques Derrida, Jean-François Lyotard and others.

In his *Ideas*, Husserl introduced a basic methodical tool he called 'phenomenological' or 'transcendental reduction,' meaning a fundamental change of attitude toward the world (Husserl, 1913: sections 31–32). Without giving up his worldly activities, the researcher was supposed to turn himself into an observer who withdrew from all involvement, methodically transforming things and actions into phenomena and thereby making them accessible in the way they show themselves – a shift from the 'what' to the 'how' of the world, as it were. Even though later theoreticians have expressed doubts about the feasibility of this operation, the necessity of some kind of break between the everyday attitude toward the world and its phenomenological treatment is generally accepted.

Husserl's early writings were primarily concerned with consciousness in a singular sense but the subject of intersubjectivity and social relations figured in his manuscripts and lectures from very early on. The concept of lifeworld that he

introduced in his later texts proved to be a fruitful starting point for sociological analyses. More fundamental, the question of how we can know or have access to others' minds and emotions has been an important phenomenological topic, and phenomenology's contributions to these questions have recently been adopted and developed in cognitive science. In phenomenological theory, the topics of other minds, the body, the lifeworld, and the other are all intertwined.

Minds and Emotions

In the beginning of the twentieth century, Theodor Lipps developed a concept of empathy ('Einfühlung,' literally 'feeling into') to explain not only our knowledge of other persons but also our understanding of art (Lipps, 1903: pp. 107–141). He posited the coupled natural drives of imitation and expression leading to a tendency to imitate perceived expressive movements, thereby reliving the content of their expression and finally 'feeling' this content 'into' the perceived body or work of art. To Lipps this was not an act but an unconscious instinct functioning automatically. Husserl and other phenomenological writers such as Edith Stein (Stein, 1917) drew upon this concept, modifying and criticizing it. According to Husserl's *Cartesian Meditations* (Husserl, 1931: sections 49–62), I perceive others' bodies and their gestures not as symbols but as immediate expressions of an inner life that is 'appresented,' i.e., given along with the visibility of their bodies without ever being truly present to me. Still there is an act of analogy or 'pairing' at the bottom of this which is not based on an actual comparison and judgment of similarity but still makes me the model and 'norm' of the others and lets them appear secondary to and dependent on the sphere of my own self. On the other hand, the objectivity of the world and its objects is based on their being objects not just for myself but for everyone. Even though this reconstruction was meant to counter the alleged solipsistic tendencies of transcendental phenomenology, the problems of reaching an intersubjective and cultural world from a transcendental subjectivity conceptualized as 'monadic' become very apparent.

Before Husserl, Max Scheler had proposed a theory that avoided any reference to 'Einfühlung,' instead relying on the observation of an immediate 'Nachfühlen' – correctly but misleadingly translated as 'empathy' as well – which formed the basis of any act of actual sympathy ('Mitfühlen'), i.e., compassion as an act of not only perceiving the other's sadness and remaining neutral toward it but sympathizing with other (Scheler, 1923: Chapter A 2). He sharply distinguished empathy and sympathy from mere contagion ('Ansteckung') where a person is affected by another's yawn or an infant by another's cries, and which contains no real reference to the other's feelings at all but comes close to a causal reaction. In neither case, however, do we first perceive a physical body which is then provided with an inner life, be it felt into, projected by analogy or inferred by an act of interpretation. To Scheler, this description was a fiction produced by scientific psychology.

As for Lipps and Husserl, emotions are referred to as the most salient instances of human expression while the argument is directed toward the more general question of how we perceive others as agents with an emotional and mental life. Scheler's basic point, which was later taken up by Merleau-Ponty and others and was endorsed by Wittgenstein (1967: sections 220, 225) is that there is no ontological gap between the visible body and the 'inner' life of a person, in a way that Cartesian theory would define it. Even though we do not actually feel the others' emotions, we do not infer them from outward signs or reconstruct them within ourselves either – rather, we see them *in* their behavior. We neither see an animated physical object nor an animated body but a person that exhibits sensible behavior. Theories that posit some kind of act of inference or analogy have difficulties explaining how these mechanisms can be at work even in early infancy where conscious acts of making inferences seem very unlikely.

Recently, this phenomenological position has been revitalized in the discussion on the theory of mind that has involved developmental psychology, philosophy, and cognitive science (Carruthers and Smith, 1996; Davies and Stone, 1995). The two most important theoretical approaches in this debate have been labeled 'theory theory' and 'simulation theory,' the latter of which bears some resemblance to the old theory of 'Einfühlung.' It proposes that we reconstruct others' mental and emotional states by simulating them within our own minds, performing an imaginary representation of what it would be like to be in their position (Gordon, 1986). Recently, the research on mirror neurons has been cited to support this view, providing it with a neuronal basis (Gallese and Goldman, 1998). The 'theory theory,' on the other hand, defends the view that we construct theories or folk psychologies concerning the inner life of others, inferring their beliefs, thoughts, and emotions on the basis of these theories. The strongest and most widely held claim by proponents of both theories is that simulation or theory are our primary means of accessing the mental state of others and that they are acquired early in childhood (Gopnik and Meltzoff, 1997) or are even innate (Baron-Cohen, 1995).

Both these theories have been criticized from a phenomenological point of view, aiming at their very premises (Gallagher, 2005: Chapter 9). The basic assumptions of the whole debate are Cartesian, conceptualizing the mind as a closed entity distinct from the body, and its focus has been explanation and prediction as primary goals of understanding others. All of these presuppositions can be contested on the grounds that they fundamentally misconstrue our basic stance toward the world which is characterized by pragmatic bodily interaction. Hence Shaun Gallagher proposes a third reconstruction which he calls 'interaction theory' (Gallagher, 2005: p. 208). In everyday interaction, our aim is not to theorize about the mind of others or to simulate its content but to make sense of what they do in a particular situation, being part of that situation ourselves, in order to respond to it and enter into interaction and/or communication. There is a primary intersubjectivity of bodily interaction that is accessible even to very young infants who show a pragmatic understanding of the motives and beliefs of others as they are manifested in their actions. This understanding precedes the explicit and rather detached reconstructions that are tested in the classical false belief experiments. Of course this does not mean that we never resort to theory or simulation at all but rather that these modes of reconstruction are not what our interaction with others as persons are based on. What

becomes obvious, however, is the basic role of our bodies in our interaction with and understanding of the world and other human beings.

The Body

In the posthumously published second part of his *Ideas*, Husserl granted the body a much more important position than he had in his previous writings (Husserl, 1952: sections 18, 35–42). He drew on the difference of 'Leib' and 'Körper' that is made in the German language, distinguishing 'Körper' as a physical entity considered from a detached point of view from 'Leib' (usually translated as 'lived body') as the primary way our own body and the bodies of others are given to us experientially. The lived body is that by means of which we experience the world and thus can never be completely objectified, and even though the bodies of others offer the possibility of objectification, we primarily experience them as experiencing the world as well. This entails a shift in the conception of consciousness: the lived body is not an object of consciousness but part of it, as it were. It is our primary mode of dealing with the world. The operation of withdrawal into a region of pure consciousness that characterized transcendental phenomenology thus becomes increasingly implausible. There is a continuity between mind and the lived body rather than an ontological gap, and the distinction between the lived body and the physical body, 'Leib' and 'Körper,' does not distinguish ontologically distinct entities but rather different perspectives on the same thing.

A similar position in a somewhat different theoretical context can be found in the work of Helmuth Plessner who studied medicine and zoology as well as philosophy with Husserl. Plessner's aim was anthropological but he argued from a phenomenological background (Plessner, 1928). On his view, humans are characterized by a dual way of relating to themselves and to the world, which he called eccentric positionality. They relate to the world from a centered position analogous to that of an animal, but they also relate to this relationship from a position that is not centered but situates itself on the border of self and world. This entails a dual relationship to one's body: we *are* a 'Leib' and *have* a 'Körper.' According to this theory, consciousness and reflexivity are firmly rooted in biology.

It was Merleau-Ponty who drew the consequences of these positions within phenomenology. Drawing on contemporary psychology, psychopathology and neuroscience, as well as on Husserlian phenomenology, he outlined a theory of the lived body as a relation to the world, an 'être-au-monde,' which is usually translated with Heidegger's term 'being-in-the-world' but could also be rendered as 'being-towards-the-world' (Merleau-Ponty, 1945). According to him, it's not the Cartesian 'I think' but the Husserlian 'I can' (move, perceive, act, etc.) that must be placed at the center of subjectivity, and the subject must be conceived as an embodied agent.

While Husserl kept his distance toward science in general and contemporary psychology in particular, Merleau-Ponty was an avid reader of psychological, neurological, biological, and later linguistic research, trying to integrate it into a philosophical account of perception. Gestalt psychology and neurology provided him with the insight that the functions of the body cannot be viewed in isolation and have to be understood within the context of its perceiving and acting in the world. From this it follows that an adequate theory of perception must also be a theory of the lived body and vice versa. As a consequence, Merleau-Ponty conceived of mind neither as an entity in itself nor as a mere function of neuronal structures and operations, but rather as the situated activity of bodily beings in interaction with each other and the world. This position is as far from Cartesian dualism as it is from a simple monism. There is a relative independence of higher processes of mind that cannot be reduced to functions of the lived body nor completely severed from it.

Many of the research perspectives Merleau-Ponty relied on and developed were discontinued in the years after the *Phenomenology of Perception* was published, and in the new cognitive science the computational model of the mind was virtually undisputed. Only recently has cognitive science turned toward an account of cognition that has been labeled as 'enactive' or 'embodied,' returning to some of Merleau-Ponty's central theses. In what has been called the 'third wave' in the development of cognitive science, the role of body in cognition is increasingly being acknowledged (Varela et al., 1992; Noe, 2004). From the side of phenomenological philosophy, there have been attempts to follow Merleau-Ponty in critically interpreting the findings of cognitive science and fitting them into a phenomenological framework which in turn has to be modified (Gallagher, 2005; Gallagher and Schmicking, 2010). The question of naturalization is being debated once again but its premises have changed. Various proposals have been made about how phenomenology and neuroscience can mutually inform each other, dropping the idea of a complete convergence (Petitot et al., 2000; Gallagher, 2012: Chapter 1).

The Lifeworld

Husserl first used the term 'Lebenswelt' in a manuscript in 1917 but the topic of the everyday world of the 'natural attitude' was present long before that. Not until 1935, however, did the lifeworld figure in his published writings. Before that, important contributions to a theory of the lifeworld were made by two of his followers, Martin Heidegger and Alfred Schütz. In *Being and Time*, Heidegger introduced the concept of 'Zeug' ('equipment') to describe the primary givenness of things in the everyday world (Heidegger, 1927: sections 14–18). According to Heidegger, our world is not populated by objects of contemplation but by things of practice that form a web or nexus of tools and equipment ('Zeugzusammenhang') where every one thing refers to others and to the practical context it belongs to and can be used in. Consequently, it is not theoretical cognition that marks our primary approach to the world but a practical kind of knowledge of handling things. Primarily the world is that which we find ourselves in, in the horizon of which we act, not a collection of objects; space is characterized by proximity and remoteness, not by a neutral measuring of distance. When he calls the 'Dasein' (i.e., human beings) 'being-in-the-world,' this is not to be understood as the presence of a physical entity within a container-like world but a type of being involved, of handling things, a presence to the world. According to Heidegger, this is

what ontology must take as its starting point, not a neutralized theoretical view.

Husserl's own point of departure in the *Crisis of European Sciences* was similar but somewhat more urgent: he diagnosed a fatal disappearance of the problems of humanity within the prevailing scientific ontology, resulting in the danger of radical skepticism and nihilism, and proceeded to work out a genealogy of natural science from the lifeworld (Husserl, 1954). He did not consider the series of subsequent abstractions in itself to be problematic, but the fact that this genealogy tended to be forgotten and its endpoint taken as a truer reality. A phenomenological 'science' of the lifeworld was supposed to counter these tendencies and ground our self-conception in a deeper understanding of our primary reality. The lifeworld as he described it is a sphere of praxis where science and philosophy must be counted as a particular form of theoretical practice. It is the 'universal field' in which we live, move, communicate, and theorize, and must be considered the ultimate foundation of any theory. It is objective only in that it is intersubjective, relating to the acts of those that inhabit it, but for every one of us it is pregiven, always already there. Even though the world of human life is characterized by great diversity, according to Husserl there are fundamental structures of sense that guarantee its unity – no matter how distant a culture may be from mine, there is always some common ground that lets me interact with its inhabitants. The lifeworld does not exist in the plural.

What was remarkable was Husserl's willingness to take intersubjectivity as a starting point, not as a result to be deduced from individual consciousness: the transcendental must be conceived as collective. Still, the question of mediation between the individual and the social that had not been solved in the *Cartesian meditations* remained a problem. In his *Phenomenology of the social world*, Alfred Schütz had provided a firmly Husserlian account of this mediation, taking individual consciousness as a starting point (Schütz, 1932). Even though the philosophical ambitions of the book are obvious, Schütz understood it as a contribution to sociology, and its scope is somewhat more modest than that of Husserl's *Crisis* where the lifeworld has no natural world or world of science beside it but is meant to encompass every kind of worldly phenomenon. The social world as Schütz analyzed it is a specific field within the world as such, and even when he later adopted the term lifeworld, this understanding remains in place (Schütz and Luckmann, 1973). His aim was to clarify Max Weber's concept of meaningful social action by an analysis of the genesis of meaning or sense in relation to the individual, the other, and society, effectively basing society on the individual. By distinguishing 'in-order-to-motives' and 'because-motives' as the sole types of motivation, Schütz significantly narrowed the range of possible descriptions and explanation of social phenomena. On this view, sense as the fundamental property of social action only comes into play in a retroactive explicit interpretation of one's own and the others' actions.

One might compare this position with the sociological works of Pierre Bourdieu (1971). Even though Bourdieu somewhat aggressively denounced philosophy, he relied heavily on the work of Husserl and especially Merleau-Ponty and developed a theory of the social world that has rather strong philosophical implications. His basic concept of 'habitus' as incarnated social meaning can be read as a sociological elaboration of the latter's notion of the lived body. As elaborated earlier, the body as Merleau-Ponty conceived it is not a given entity but rather our primary means of relating to the world. In this way it is shaped by its interactions just as much as by its innate structure, and the acquisition of skills like manipulating complex tools or playing a musical instrument must be considered as an extension and modification of it. In social interaction, the habitus shapes our way of acting, itself being shaped by previous interactions. The mediation between individual and society does not take place in the constitution and reconstruction of meanings as motives but in the incarnation and reenactment of social meaning by the habitus. Consequently, action should be considered as situated rather than subjective in the traditional sense. According to Bourdieu, social change is always possible because social action is never completely determined, but it is difficult because it is not a matter of mere reinterpretation of social meanings but has to deal with the inertia of the habitus.

The Other

In the phenomenological description of intersubjectivity and sociality presented so far, one aspect was all but absent: the active role of the other. In Husserl's as well as in Scheler's and Schütz' reconstructions, the other appears viewed from the perspective of the self, as someone who must be intuited or understood. Jean-Paul Sartre offered a different perspective. In his *Being and Nothingness*, the gaze of the other opens up a whole new dimension: it objectifies me and pins me down, as it were, transforming me from an essentially indeterminate and free subject into this particular bodily being that can be seen and defined (Sartre, 1943: pt. 3, Chapter 1, IV). Also, the other's gaze upon the world that for Husserl guaranteed its objectivity appears much more dramatic, as a radical reorientation of every object away from me and toward him. Even though Sartre's account of the gaze might seem overdramatized, his insistence on the irreducibility of my being affected by the other to my sphere of consciousness is an important contribution to phenomenology, reminding us that the inaccessibility of other minds might not be the only problem phenomenology has to deal with when it comes to intersubjectivity.

Emmanuel Levinas followed an even more radical path of inquiry. The other with a capital O as he construed him is not primarily an objectifying gaze but a face that issues a demand. In itself, this demand is unconditional, calling me into responsibility for the other in a way that I cannot evade or escape (Levinas, 1961). By describing this relation as primordial and irreversible, Levinas completely reinterprets the relationship between self and other as it was conceived by Husserl. Now it is not the Other who is dependent on me, being a vis-à-vis that I have to intuit or understand, but it is me who is exposed to the demand of his face before I can think of myself as a competent individual who can freely decide what he wants to concern himself with. Even if I am not bound by the demand I have to relate to it in some way, and even if in the end I turn away, I cannot deny it; Levinas speaks of non-indifference (Levinas, 1974: Chapter V 1. e). The primordial

relationship is not one of cognition, and the Other will not be made into an object and cannot be dealt with in the categories phenomenology has to offer – a fact that can only be stated in phenomenological terms.

The situation Levinas describes is one of radical asymmetry: even though he speaks of a face-to-face situation, the relation cannot be turned around so there would be a complementarity of perspectives, and neither is there a common category the other and myself can be subsumed under. The asymmetry in Husserl's version of intersubjectivity ran the other way, with the subject being the norm of the other and the locus, if not the agent of his constitution as other. Even so, there is a process of 'othering' of the self, as Michael Theunissen called it, leading to a situation where I am one among others (Theunissen, 1965: sections 11–14). For Levinas, there is no such operation of leveling myself and the Other, and his alterity is irrevocable. My responsibility stemming from his demand cannot be countered by my rights toward him. It is unconditional and has no natural bounds.

Following the tradition of Husserl and Heidegger, Levinas considered this theory of the other 'first philosophy,' i.e., heir to metaphysics, and contrasted it with ontology which suggests an order of being(s) that can be rationally ordered and comprehended. Ethics in the radical sense he gave it is the true first philosophy because the asymmetrical relation to the other is primordial in the sense that subjectivity and cognition of the world is based on it and it cannot be conceived in neutral terms.

Levinas' position entails some obvious questions concerning the possibility of ethics in the more mundane sense of how we are supposed to act in different situations and of politics as the art of balancing interests where the unconditional has no place. He himself speaks of the 'third party' with whose appearance the question of justice arises, but does little more than state the problem. Others have taken this lead: Burkhard Liebsch calls for a culture of hospitality, calling into question the primacy of rootedness and cultural identity in the face of others whose appeal is urgent and real (Liebsch, 2008). Simon Critchley has offered an interpretation of politics that stresses an ethics based on the Levinasian demand of the other as a potential element of unrest within the political sphere, persistently calling into question any established order, and as an inspiration for resistance and, finally, an anarchist stance (Critchley, 2007). Robert Bernasconi has pursued the topic of race and racism from a Levinasian background (Bernasconi, 2001). Bernhard Waldenfels has been following a different path, drawing on Levinas, Merleau-Ponty, and others to work out a phenomenology of the alien ('der Fremde,' which might also be translated as 'the stranger' or 'the foreigner') (Waldenfels, 1994, 2006). He adheres to the idea of an irrevocable asymmetry but insists on differentiating between different ways of being addressed or appealed to and different registers of responding. Still, there is always a gap between address and response, an element of discontinuity and belatedness or deferment that makes it impossible to simply fulfill a demand or to respond exhaustively. The Other as alien always escapes my attempts to pin him down, to understand him or, most importantly, to do him justice.

Phenomenology in its different varieties has made contributions to various fields of philosophy and related disciplines. Even though aspirations for developing a new version of first philosophy that the tradition has inherited from its founder Edmund Husserl have come to be viewed critically, its productivity remains unbroken. In philosophy, the topics of other minds, the body, the lifeworld, the Other, and others that could not be elaborated here have provided inspirations for thinkers from diverse backgrounds and continue to be a challenge today.

See also: Embodied Social Cognition; Embodiment Theory; Husserl, Edmund (1859–1938); Mind, Theories of; Mirror Neurons, Theory of; Phenomenology in Sociology; Pragmatism: Philosophical Aspects.

Bibliography

Baron-Cohen, Simon, 1995. Mindblindness: An Essay on Autism and Theory of Mind. MIT Press, Cambridge, Mass.

Bernasconi, Robert, 2001. Race. Wiley-Blackwell, London.

Bourdieu, Pierre, 1971. Esquisse d'une théorie de la pratique. Droz, Geneva (Trans. Nice, Richard, 1977. Outline of a theory of practice. Cambridge University Press, Cambridge.).

Carruthers, Peter, Smith, Peter K. (Eds.), 1996. Theories of Theories of Mind. MIT Press, Cambridge, Mass.

Critchley, Simon, 2007. Infinitely Demanding. Ethics of Commitment, Politics of Resistance. Verso, London.

Davies, Martin, Stone, Tony (Eds.), 1995. Folk Psychology: The Theory of Mind Debate. Wiley-Blackwell, Oxford.

Gallagher, Shaun, 2005. How the Body Shapes the Mind. Clarendon Press, Oxford.

Gallagher, Shaun, 2012. Phenomenology. Palgrave Macmillan, Basingstoke.

Gallagher, Shaun, Schmicking, Daniel (Eds.), 2010. Handbook of Phenomenology and Cognitive Science. Springer, Dordrecht.

Gallese, Vittorio, Goldman, Alvin, 1998. Mirror neurons and the simulation theory of mind-reading. Trends in Cognitive Sciences 2, 493–501.

Gopnik, Alison, Meltzoff, Andrew, 1997. Words, Thoughts, and Theories. MIT Press, Cambridge, Mass.

Gordon, Robert M., 1986. Folk psychology as simulation. Mind and Language 1, 158–171.

Heidegger, Martin, 1927 and 1977. Sein und Zeit. Klostermann, Frankfurt am Main (Trans. Macquarrie, John, Robinson, Edward, 1962. Being and time. Harper and Brothers, New York.).

Husserl, Edmund, 1901 and 1984. Logische Untersuchungen 2: Untersuchungen zur Phänomenologie und Theorie der Erkenntnis. Martinus Nijhoff, The Hague (Trans. Moran, Dermot, 2001. Logical investigations, vol. 2. Routledge, London.).

Husserl, Edmund, 1913 and 1976. Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie 1: Allgemeine Einführung in die reine Phänomenologie. Martinus Nijhoff, The Hague (Trans. Kesten, Fred, 1983. Ideas pertaining to a pure phenomenology and to a phenomenological philosophy 1: General introduction to a pure phenomenology. Martinus Nijhoff, The Hague.).

Husserl, Edmund, 1931 and 1950. Cartesianische Meditationen und Pariser Vorträge. Martinus Nijhoff, The Hague (Trans. Cairns, Dorion, 1960. Cartesian Meditations: An Introduction to Phenomenology. Martinus Nijhoff, The Hague.).

Husserl, Edmund, 1952. Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie 2: Phänomenologische Untersuchungen zur Konstitution. Martinus Nijhoff, The Hague (Trans. Rojcewicz, R., Schuwer, A., 1990. Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy 2: Studies in the phenomenology of constitution. Kluwer Academic Publishers, The Hague.).

Husserl, Edmund, 1954. Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie: Eine Einleitung in die phänomenologische Philosophie. Martinus Nijhoff, The Hague (Trans. Carr, David, 1960. The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy. Northwestern University Press, Evanston.).

Levinas, Emmanuel, 1961. Totalité et infini: Essai sur l'extériorité. Martinus Nijhoff, The Hague (Trans. Lingis, Alphonso, 1969. Totality and Infinity: An Essay on Exteriority. Duquesne University Press, Pittsburgh.).

Levinas, Emmanuel, 1974. Autrement qu'être ou au-delà de l'essence. Martinus Nijhoff, The Hague (Trans. Lingis, Alphonso, 1981. Otherwise than Being or Beyond Essence. Martinus Nijhoff, The Hague.).

Liebsch, Burkhard, 2008. Für eine Kultur der Gastlichkeit. Alber, Freiburg and Munich.
Lipps, Theodor, 1903. Ästhetik: Psychologie des Schönen und der Kunst. In: Grundlegung der Ästhetik, vol. 1. Voss, Hamburg.
Merleau-Ponty, Maurice, 1945. Phénomenologie de la perception. Gallimard, Paris (Trans. Smith, Colin. Phenomenology of perception. Humanities Press, New York.).
Noe, Alva, 2004. Action in Perception. MIT Press, Cambridge, Mass.
Petitot, Jean, Varela, Francisco, Pachoud, Bernard, Roy, Jean-Michel (Eds.), 2000. Naturalizing Phenomenology. Stanford University Press, Stanford.
Plessner, Helmuth, 1928. Die Stufen des Organischen und der Mensch. Einleitung in die philosophische Anthropologie. de Gruyter, Berlin.
Ricœur, Paul, 1953. Sur la phénoménologie. Esprit 21, 821–839.
Sartre, Jean-Paul, 1943. L'être et le néant: Essai sur l'ontologie phénoménologique. Gallimard, Paris (Trans. Barnes, Hazel E., 1956. Being and Nothingness. An Essay on Phenomenological Ontology. Philosophical Library, New York.).
Scheler, Max, 1923 and 1973. Wesen und Formen der Sympathie. Francke, Bern (Trans. Heath, Peter, 1954. The Nature of Sympathy. Routledge and Kegan, London.).
Schütz, Alfred, 1932 and 1995. Der sinnhafte Aufbau der sozialen Welt: Eine Einleitung in die verstehende Soziologie. Suhrkamp, Frankfurt am Main (Trans. Walsh, George, Lehnert, Fredrick 1967. Phenomenology of the social world. Northwestern University Press, Evanston.).
Schütz, Alfred, Luckmann, Thomas, 1973. Structures of the Lifeworld (Richard M. Zaner, J. Tristram Engelhardt Jr., trans.). Northwestern University Press, Evanston (German edition (1975). Strukturen der Lebenswelt. Luchterhand, Neuwied.).
Stein, Edith, 1917 and 2008. Zum Problem der Einfühlung. Herder, Freiburg (Trans. Stein, Waltraud, 1964. On the Problem of Empathy. Martinus Nijhoff, The Hague.).
Theunissen, Michael, 1965. Der Andere. Studien zur Sozialontologie der Gegenwart. de Gruyter, Berlin (Trans. Macann, Christopher, 1986. The Other: Studies in the Social Ontology of Husserl, Heidegger, Sartre, and Buber. MIT Press, Cambridge, Mass.).
Varela, Francisco J., Thompson, Evan, Rosch, Eleanor, 1992. The Embodied Mind. MIT Press, Cambridge, Mass.
Waldenfels, Bernhard, 1994. Antwortregister. Suhrkamp, Frankfurt am Main.
Waldenfels, Bernhard, 2006. Grundmotive einer Phänomenologie des Fremden. Suhrkamp, Frankfurt am Main (Trans. Stähler, Tanja, Kozon, Alexander, 2011. Phenomenology of the alien: basic concepts. Northwestern University Press, Evanston.).
Wittgenstein, Ludwig, 1967. Zettel (G.E.M. Anscombe, trans.). Blackwell, Oxford.

Philosophy of Law

Stefan Magen, Ruhr-Universität Bochum, Bochum, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A central task of legal philosophy is to provide a general explanation of the normativity of positive law. Legal positivist theories pursue this aim primarily by philosophical analyses (linguistic, conceptual, or ontological) that explain law by descriptive reference to social practices, mainly of institutional actors. For example, legal validity is explained by a rule of recognition, which is a social norm among legal officials. Naturalistic approaches in legal positivism seek a closer contact to the social sciences. Natural law theories deny that any explanation of law could succeed without reference to (true) political morality.

Since the beginnings of philosophy in ancient Greece, law has been the object of philosophical inquiry. However, classical natural law theories in the tradition of Plato, Aristotle, or Aquinas treated law as part of normative moral or political philosophy. In the modern age, with the emergence of nation-states and market societies, law increasingly became controlled by centralized political authorities and more separated from traditional morals and social customs. In the wake of these developments, a discussion on the concept or nature of 'positive law' emerged in legal philosophy. Jeremy Bentham not only dismissed inquiries into preexisting natural rights as metaphysical 'nonsense on stilts,' as is well known, but also developed a long unpublished analysis of law as commands (Bentham, 1970[1782]). The 'command theory of law' was later elaborated and published by Bentham's disciple and legal philosopher John Austin (1995[1832]). It became a prevalent view in Anglo-Saxon legal positivism and Austin's methodological maxim to keep descriptive analyses of law 'as it is' apart from normative inquiries into law 'as it ought to be,' remains the hallmark of legal positivism. Legal positivism was revived when H.L.A. Hart presented a more sophisticated analysis of law that drew on sociological hermeneutics and Hans Kelsen's 'pure theory of law,' an earlier version of legal positivism in the continental tradition (Kelsen, 1967[1960]; Marmor, 2010). Hart understood legal systems as a union of primary and secondary rules and introduced the distinction between the internal point of view of a participant in the legal system and the point of view of an external observer (Hart, 2012[1961]). Hart's seminal work sparked a debate between legal positivism and natural law theories over the concept or nature of law, which grew into an academic field of its own, with Roland Dworkin's interpretivism arguably being its most influential critique and alternative (Dworkin, 1977, 1986). Today this debate forms a major branch of the philosophy of law (or legal theory or jurisprudence, respectively) on which this article focuses. Normative inquiries into particular legal doctrines or specific areas of law form another branch of legal philosophy, which is not covered here.

Positive Law, Social Facts, and Morality

Positive Law and Social Practice

Legal orders are collections of norms, be it the law of nation-states, supranational entities or international law. All legal orders comprise norms that impose duties to perform or refrain from certain actions on natural persons or legal entities (primary rules). In most cases, legal orders also comprise explicit meta-norms on the creation, change, adjudication and enforcement of primary rules, in particular norms that determine legal powers, procedures, and legal personalities (secondary rules). Legal orders are often, but not necessarily, organized hierarchically, like the legal systems of nation-states, which are united by a common constitution and divided, e.g., into federal law, state law, municipal law. International law, however, with its imperfectly coordinated international regimes, often lacks hierarchical integration. Different to norms endorsed in ideologies or philosophical theories, legal norms are necessarily grounded in and emerge from actual social practices of particular collectivities. Without a social practice supporting it, positive law ceases to exist. Legal norms share this property with informal social norms, and like them, legal systems usually react to noncompliance with some form of explicit sanctions (McAdams and Rasmusen, 2007). But unlike social norms, legal systems are formal institutions administered according to rules by specialized legal officials acting in legal capacities on behalf of legally structured organizations like legislatures and courts. Norms of positive law are thus tied to the actual behavior of legal officials, often subject to deliberate change and confined to a particular jurisdiction. In this regard, positive law exhibits the characteristics of an observable social fact, located in time and space.

Normativity of Legal Norms

Participants in the legal system, however, tend to treat legal rules as genuine norms, not identical with the empirical social facts on which they rest. For example, lawyers often talk about legal norms as if they were entities with an existence of their own. To restrict a legal norm to a mere expression of an actual legislatorial wish combined with a credible threat, seems to reduce the full meaning of the term. Rather, legal obligations also provide intrinsic normative reasons for action. Following Hart, this stance toward the law is often called the internal point of view (Hart, 2012[1961]; Shapiro, 2006). Through this capacity to provide intrinsic normative reasons, legal rules appear to exhibit some characteristics that are usually ascribed to moral norms. Furthermore, lawyers often determine the content of legal norms

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63106-9

in light of normative principles, purposes, or values that are not explicitly stated in authoritative legal sources. And sometimes legal norms or judicial reasoning explicitly refer to standards like 'equality' or 'fairness' that appear to refer to (putatively) true moral norms. Thus, positive law also seems to be inherently interwoven with true morality. Hence, in tying together factual social practices with (possibly) genuine normativity, positive law appears to have a puzzling and seemingly dual nature.

Legal Philosophy, Science, and Political Philosophy

The specific contribution of legal philosophy to the explanation of legal normativity is controversial and should be viewed in contrast to other disciplines: Empirical explanations of the factual side of legal institutions are a matter of the social sciences and the philosophical questions pertaining to it should be addressed by the philosophy of science (Mantzavinos, 2009). The moral merits of positive law should be judged by normative political philosophy and meta-ethics. The content of concrete norms in particular legal orders, as it presents itself from an internal point of view, is discussed by legal doctrine, which may or may not draw on other sources like morality or scientific knowledge in varying ways and degrees. In contrast, the philosophy of law should, as far as this is possible, be concerned with providing general explanations of the normativity of positive law and its different aspects. In any such account, usually an explanation of law's relationship or nonrelationship to morality features prominently. This applies across topics, e.g., with regard to the basic elements and structure of legal orders, legal validity, the nature of legal obligation, legal interpretation and adjudication, etc.

Natural Law Theories versus Legal Positivism

Philosophical positions on these issues are usually perceived against the backdrop of the antagonism between legal positivism and natural law theories. In general, three explanatory strategies, either alone or in combination, are employed in this regard. In natural law theories, the normativity of legal norms is accounted for (1) by reference to normative, moral 'facts,' i.e., some version of a (possibly or purportedly) true political morality. In legal positivism, the normativity of law is (2) either explained by descriptive reference to social facts with 'normative' content like social norms (as in the approach of Hart and most other Anglo-American variants of legal positivism), or (3) the normativity of positive law is regarded as a stand-alone phenomenon that neither lends itself to explanations from social facts or moral norms, as the pure theory of law of Kelsen holds (Kelsen, 1967[1960]). However, despite their fundamental differences, legal positivists and natural law theorists often agree on the existence of many of the aforementioned, ostensibly moral characteristics of law. Real disagreement remains, however, with regard to the interpretation of these phenomena as either being explicable as social facts or as necessarily implying true morality. For example, Hart agrees with natural law theories that obligations stated in valid legal norms provide intrinsic reasons for actions, but he explicates these reasons as normative attitudes and hence mental states. Natural law theorists on the other hand insist that only true morality could actually provide normative reasons for actions. Thus, the main points of contention between legal positivism and natural law theories revolve around the extent to which descriptive or analytical accounts of positive law are methodologically possible, in substance successful, and, if so, also of interest (Dickson, 2012). Put differently, the discussion of the concept of law is to some extent a discussion about separating descriptive from normative questions in the philosophy of law.

The Validity of Law

The concept of validity is critical for any philosophical explanation of positive law. It is also a central issue in the debate between legal positivism and natural law and will be used in the remainder of this article to illustrate their respective positions. The concept of validity is also pivotal for the working mechanisms of positive law. Modern legal systems do not treat the content of law as given by nature, tradition, or religion, but explicitly allow for the creation and change of laws through willful actions of authorized actors as in the conclusion of international treaties, adoption and change of constitutions, legislation, and administrative rule setting or judicial law making via precedents. In consequence legal norms vary over time and between legal systems. The function of the concept of validity is thus to identify those legal norms that actually 'exist' in a particular legal system. Attributed to a norm, legal validity implies its existence as a legal norm, membership in a particular legal system, and normativity or binding force. The concept of legal validity is relevant, i.e., for two different but interconnected problems of positive law: (1) the foundation problem, i.e., what the ultimate grounds of validity are, and (2) the specification problem, i.e., how the actual content of a particular legal system is determined, e.g., by statutes or precedents. Differences between legal positivism and natural law theories over the role of social facts versus true morality play out differently in both problems. In general, with regard to ultimate validity differences between them, the two approaches tend to be more categorical; whereas with regard to specification, they are rather like different ways of accounting for identical features.

Legal Positivism

Separation Thesis

Legal positivist theories hold that the existence and content of law depends on social facts only, not on its merits (Green, 2009). That does not exclude some legal positivists from acknowledging certain contingent or even necessary connections between law and morality (Gardner, 2001). But as far as the validity of legal norms is concerned they endorse the separation thesis. Legal positivism thus strives to account for the normativity of law by reference to social facts. Some legal positivists nevertheless admit certain apparently moral properties in law like the presence of normative principles, purposes, or values in legal reasoning, explicit references to morality, or the perception of legal obligations as intrinsic reasons for actions. But these phenomena are interpreted in a descriptive manner as facts about the law that can be explained by or as social facts. For example, Austin explains the obligation contained in a legal norm with the actual behaviors of threatening sanctions and

habitual obedience. Hart criticizes this reduction as inadequate because, as he contends, from an internal point of view legal obligations are perceived as intrinsic reasons for action. But for Hart the internal point of view does not necessarily rest on good moral reasons to obey the law but on an empirical attitude that accepts law as a standard of conduct. This does not imply that legal positivists always deny the possibility of a true political morality. In fact, Austin believed in Divine law and Hart held moral realist convictions. But legal positivists have to insist that morality is not necessary for such an account, at least as far as the existence and content of law are concerned.

Validity in Legal Positivism

Pure Theory of Law

Most radically perhaps, the pure theory of Hans Kelsen denied that the normative validity of legal systems could be derived from either social facts or from true morality without violating Hume's Law or running into an infinite regress. For Kelsen, validity is ultimately grounded in what he calls a 'basic norm' (Kelsen, 1967[1960]; Marmor, 2010). This basic norm, however, is neither a social norm nor a norm of Divine or natural law; it is not even a real norm but only a presupposition lawyers necessarily make in reasoning about law. Thus, for Kelsen, legal validity is merely hypothetical or 'transcendental.'

Command Theory of Law

Anglo-American legal positivists, however, tend to have no objection, in principle, to locating the ultimate source of legal validity in social facts, namely in social practice with normative content. Rather than tackling the general philosophical problem of how a legal 'ought' could be derived from an 'is,' Anglo-American legal positivism rather focuses on the substantive analysis of the social practices from which legal systems emerge. Before Hart, the prevalent account was Austin's version of the command theory, according to which positive law consists of the commands of a sovereign (Austin, 1995 [1832]). The sovereign is a person or a group that receives habitual obedience from the majority of society but is not in the habit of obeying anybody else. Legal norms are a particular type of command issued by the sovereign, namely general and public orders backed up by the threat of sanctions. The sovereign's capability to produce norms rests in his ability to command the obedience of his inferiors by exercising power. It is widely assumed that Hart demonstrated the inadequacy of this account (see Schauer, 2010). For example, in many legal systems there is no single sovereign with ultimate powers. Also, the concept of sovereignty can hardly account for the restrictions that constitutions impose on law-making powers. But first and foremost, in legal systems, ultimate authority is defined by rules rather than preceding them.

Hart's Concept of Law

In his own account Hart replaced authoritative orders with rules as the essential element of law and introduced the distinction between primary and secondary rules. Primary rules impose duties on people and thus perform the function orders have in Austin's account, whereas secondary rules govern the creation, change, and administration of rules. Thus, legal systems usually comprise secondary legal rules that explicitly regulate the creation of valid legal norms, be it primary rules or lower-level secondary rules. They do so by determining legal procedures and competences which have to be observed for such changes to be acknowledged as actually creating new or changing existing law. Rules granting such powers are often stated explicitly in foundational or supreme legal documents like constitutions or the EU-Treaties or they may be the subject of customary law at this supreme level. Validity is an intrasystemic matter regarding the proper application of power-conferring legal rules, i.e., a legal question to be addressed by legal doctrine.

However, the ultimate source of legal validity, i.e., the validity of the legal system as such, cannot be derived from legal rules without circularity. To account for the ultimate validity of a legal system, Hart introduces the rule of recognition. The rule of recognition is a secondary norm that exhaustively specifies the criteria for validity in a given legal system. Thus a legal norm is valid if it satisfies all the criteria specified in the rule of recognition. However, the rule of recognition itself, as the ultimate source of validity, is not a legal rule, but a social rule among legal officials. It functions, so to speak, as an 'empirical basic norm' that gives positive law a conventional foundation in the social practice of legal officials. This concept is widely accepted in legal positivism, in principle, but considerable debate remains as to the exact nature of the foundational social norm. Hart conceived the rule of recognition as a 'social rule' which is accepted as a common standard and the violation of which is met with criticism. Alternatively it has been proposed to regard the rule of recognition in game theoretic terms as a coordinating convention (Postema, 1982), as a constitutive convention (Marmor, 2009), or in terms of the theory of joint intentionality as a shared plan (Shapiro, 2011). The questions of whether the rule of recognition only addresses judges or also other legal officials, and whether it provides legal officials with authority and/or puts them under a duty to apply the rule of recognition also remain points of contention. There is also a discussion, sparked by Ronald Dworkin's critique, on whether the idea of a rule of recognition can be reconciled with dissent among legal officials over the content or proper application of the rule of recognition.

Inclusive versus Exclusive Legal Positivism

The establishment of positive law as an institution that is clearly set apart from other informal types of social institutions like social customs or traditional morals, is a hallmark of modern legal systems. A major way to bring this about is by adopting secondary rules that organize the formal creation and application of legal norms through specific legal bodies, procedures, and forms, as is the case with parliamentary legislation or judicial precedents. In this way, valid legal norms can be traced back to observable and easily identifiable sources. Put differently, valid legal norms are singled out by originating from a competent body invested with legal authority and following the required procedures and formalities. However, not all legal norms can be traced back to such formal sources. Many national legal systems, as well as international law, adopt customs that are widely observed in practice as customary law if they meet certain requirements specified in the jurisprudence of courts or legal doctrine. Many legal systems also regard some general normative principles as legally binding, even if they are not

explicitly stated in authoritative legal sources but rather inferred as implicitly underlying them. And in many legal systems, judges, under certain circumstances, may make reference to what they regard as morally true without these moral truths being explicitly laid down in positive law. In all these cases positive law seems to include social, implicit, or moral norms that, as such, are not the product of formal law creation and hence lack the institutional pedigree that is typically associated with positive law. On this basis, Dworkin dismissed the rule of recognition as being incapable of accounting for nonsource-based norms and, in particular, for legal principles.

Legal positivism's reaction to this criticism was to split into two camps, exclusive (or hard) and inclusive (or soft) legal positivism. Exclusive legal positivism insists that norms of positive law have to be based on authoritative sources. Nevertheless, the rule of recognition may empower legal officials to treat customary law, legal principles, or moral arguments as part of the law. This, however, does not include social or moral norms among the sources of law, but rather confers the authority (and often the discretion) on judges to make reference to extralegal norms and thereby create new law. The judicial decision is hence to be regarded as the source of positive law, not as the custom, principle, or moral norm. Inclusive legal positivism, on the other hand, takes a more lenient view of the rule of recognition. According to this view, the rule of recognition may include social or moral norms in the sources of law. For example, the rule of recognition could include a proviso that grossly unjust statues are void *ab initio*. Thus, in its inclusive version, practical differences between legal positivism and natural law theories almost vanish. The ultimate explanation conditioning legal validity on morality, however, would still be categorically different: the inclusion of morality rests in the first instance on a contingent consensus among legal officials, in the other on a necessary connection of positive law with true morality.

Natural Law Theories

Moral Realism and the Connection Thesis

Many, but not all theories in normative political philosophy suggest normative standards of political morality that are intended for, or could be applied to the evaluation of positive law. Classical natural law theories find these normative standards in putatively universal moral norms that exist independent of human will and recognition (Aquinas, 2006 [1265–1274]), a view that implies ambitious epistemological and metaphysical assumptions. Modern accounts of political morality tend to avoid them and would rarely be counted as natural law theories in political philosophy. Natural law theories in legal philosophy, however, are defined more specifically as a particular thesis about the relationship between positive law and political morality. Natural law theories in legal philosophy specifically hold that positive law, as such, cannot be properly explained for methodological and/or substantial reasons, without reference to true political morality. Natural law theories thus include two elements: (1) they endorse a substantive position on political morality with regard to the law ('natural law') and thus assume at least the possibility of true political morality (moral realism), and (2) they endorse a thesis about a necessary connection between that morality and the validity and/or content of positive law (connection thesis).

Of course, natural law theories differ widely regarding the kind of 'natural law' they endorse. Some argue on the basis of classical natural law theories and some rely on modern accounts of political morality. For example, John Finnis defends an Aquinian account of natural law that presumes an universal, mind-independent, natural law (Finnis, 2011[1980]). Other modern natural law theories adopt the connection thesis but do not argue for a universal political morality. These approaches are sometimes labeled 'non-positivism' to set them apart form classical natural law theories. By far the most influential non-positivist theory is Ronald Dworkin's interpretivism (Dworkin, 1977, 1986; Stavropoulos, 2014). According to Dworkin, law is an interpretive social practice defined by the aim to legitimate the use of coercion. To fulfill this function, law necessarily includes moral standards that are not mere social facts but, in principle, capable of actually providing the required legitimization. But in contrast to universal natural law these moral standards are not independent from the social practice, but inherent to it (Stavropoulos, 2014). A constructivist natural law theory that is also metaphysically unsuspicious has been proposed by Robert Alexy. It holds that law necessarily claims to be legitimate – a claim that has to fulfil certain criteria of 'correctness,' to be redeemed in political and legal discoursest (Alexy, 2002).

Validity in Natural Law Theories

Morally Defined Function of Law

Natural law theories jointly accuse legal positivism of overlooking the moral assumptions necessarily inherent in positive law. This criticism may be based on purely descriptive–analytical grounds, on grounds of normative political philosophy, or both. Often, it is based on the (descriptive or normative) claim that law is (or should be) an institution that is defined by its function (i.e., a 'functional kind'), and that this function is necessarily moral in nature (Moore, 1992). For example, traditional natural law theories view the purpose of law as the promotion of the common good, modern contract theories tend to find its purpose in guaranteeing individual rights or justice, and Dworkin sees the purpose of US Law in the justification of state coercion. Some natural law theories take functions to be universal, whereas in interpretivist or constructivist versions, moral functions may be restricted to the social practices of a particular community. On both accounts, however, it is regarded as a necessary property of positive law not only to pretend but to actually fulfill its moral functions, at least to a certain degree. If this assumption were true it would seem impossible to explain positive law without reference to true morality. With regard to legal validity, the connection thesis implies that social practices cannot provide sufficient and/or exhaustive grounds for legal validity. Instead, the connection thesis holds that legal validity, in general or in specific contexts, depends on true morality. This is rarely taken to imply the extreme view that legal validity is conditional on full conformity with political morality. Rather, noncontradiction with some fundamental tenets of political morality is required. For example, after World War II, Gustav Radbruch posited what came to be known as the 'Radbruch

Doctrine.' It claimed that grossly unjust laws should not be acknowledged as legally valid; and laws that did not even attempt to bring about justice should be denied the quality of law altogether (Radbruch, 2006[1946]; Alexy, 2002).

Social Facts and the Specification Problem

However, none of these natural law theories denies the relevance of social practice. Rather, they ascribe to social facts the role of fixing the specific content of legal norms within the confines of political morality. Social facts could, for example, serve to adapt universal natural laws to the circumstances obtaining in a particular jurisdiction. In an interpretive account, facts about institutional legal practices are needed in order to infer the moral claims inherent in these practices. Nevertheless, for interpretivism, facts about institutional practices are insufficient to account for the specific content of law, because the content of law is determined by what follows from principles that best explain and justify past institutional practices. On both accounts, social facts are insufficient to account for the validity of positive law.

Methods

Two different kinds of methods (or some combination of them) can be employed in arguing for a concept of law, descriptive methods of philosophical analysis, or normative methods of political philosophy. Although descriptive methods dominate in legal positivism and normative methods are common in natural law theories, methods and substantive positions do not line up perfectly. For instance, legal positivism is sometimes defended on normative grounds as well, e.g., because a positivist concept of law is thought to foster legal certainty and strengthen the rule of law. Proponents of natural law theories, on the other hand, often argue that legal positivist theories fail to account for important features of law even on purely descriptive–analytical grounds. In this vein, Dworkin criticized Hart's notion of a rule of recognition as being incapable of explaining the presence of unwritten legal principles. The methods of normative political philosophy are beyond the scope of this article. A brief discussion of descriptive analyses, however, is warranted because it is not self-evident why explanations of positive law that employ references to social facts are philosophical in nature. All social facts that legal positivist theories draw on could be the object of study in the empirical social sciences, e.g., sociology of law, legal psychology, or economic analysis of law.

Linguistic Analysis

From the publication of Hart's 'Concept of Law' until not long ago, legal philosophy usually has been set apart from social science approaches by its philosophical method of a 'linguistic analysis.' Hart himself described his method as being both 'analytical jurisprudence' and 'descriptive sociology' (Hart, 2012 [1961]: Preface iv) and quoted 'ordinary-language' philosopher John Langshaw Austin, stating that "a sharpened awareness of our words" may be used "to sharpen our perception of the phenomena" (Hart, 2012[1961]: p. 14). Thus it was a common assumption that Hart was committed to 'ordinary language' philosophy and that the reference to sociology rather referred to the empirics of language usage. In 'ordinary language' philosophy, an analysis of the concept of law would imply an analysis of the ways in which the word 'law' functions in rule-governed linguistic practice among competent speakers of a natural language ('language games'), based on observations of that practice and purportedly shared linguistic intuitions (Marmor, 2013). But by now there seems to be broad agreement that few philosophical insights emerge from tracking the meaning of the word 'law' in our linguistic practices and that, contrary to Hart's own insinuations, linguistic analysis was neither the kind of philosophical analysis he actually undertook nor the analysis that legal philosophy should pursue. In any case, as a substantive thesis on legal orders and their relation to morality, legal positivism is not committed to a particular method and is compatible with different semantic, ontological, or epistemological positions.

Conceptual and Ontological Analysis

The method that should replace linguistic analysis is a subject of current debates (Coleman, 2002). The question concerns not only the method but also the proper object of legal philosophy. With regard to the latter, Josef Raz proposed that instead of the 'ordinary language' meaning of law, legal philosophy should study law as part of the 'self-understandings' of a society (Raz, 2005). The concept or nature of law thus does not refer to the institution of law as such but to a society's self-understandings with regard to law (Marmor, 2013). Alternatively, law itself, i.e., the empirical social institution grounded in communication, behavior, and mental states, could be the object of analysis. Since shared beliefs and attitudes may play a vital role in constituting law and thus be part of the ontology of this social institution, analyses of shared self-understandings may also be involved in the analysis of law as a social institution (Leiter, 2013b), although they do not exhaust it. No consensus exists, however, as to what the analyses of either type would imply in terms of philosophical methodology and how these would differ from methods of empirical social sciences. Nonlinguistic, conceptual, or ontological analysis and methodological naturalism are two important candidates, so far.

Appropriate methods of nonlinguistic analyses depend on their subject and aims. If the purpose lies primarily in identifying, clarifying, and/or enlightening the societal concept of law, it might strive to remove ambiguities and contradictions, make implicit assumptions explicit, and/or eliminate implausible metaphysical presuppositions. If conceptual analysis aims at an ontological explanation of the phenomenon itself, it might seek to identify essential, or at least, important general properties of law. Such an analysis could imply, for example, breaking up the complexities of legal orders into simpler conceptual elements (e.g., imperatives, primary and secondary rules, or basic concepts) or into ontologically more fundamental phenomena (e.g., social norms, conventions, or joint plans) or working back to prior principles (e.g., to the separation of law from morality). In this vein, Andrei Marmor proposed a reinterpretation of Hart as a reductionist analysis of law's essential properties (Marmor, 2013). Sometimes intuitions are advocated as the specific kind of evidence philosophical analysis should draw on. But in legal

philosophy it is questionable how intuitions could help to establish descriptive adequacy, once the realm of linguistic analysis is left. Analyses in legal philosophy rather seem to rest on empirical generalizations and can thus hardly claim the self-evident, a priori necessity invoked in other areas of philosophy. Rather than purporting to show necessary or even essential properties of law, legal philosophy might be content with demonstrating general, important, or characteristic features of positive law. Also, legal philosophy might keep closer contact with the social sciences, as Hart was influenced by the sociological hermeneutics of Max Weber.

Naturalism in Legal Philosophy

Reliance on the social sciences is taken one step further in philosophical naturalism. According to methodological naturalism, philosophy should proceed in continuity to the empirical sciences by emulating their methods and/or drawing on their results (Leiter, 2012). Usually this implicates the natural sciences, but in case of legal philosophy the focus lies on the social sciences. Naturalism could mean replacing philosophy with science or locating philosophy at the highly abstract, conceptual, or foundational end of science. Both approaches may, however, be incapable of addressing questions pertaining to the relation between law and true morality, since true morality is beyond the purview of scientific methods. But naturalism may be apt for explaining the relationship between the law's normativity and the social facts it is grounded in. Questions such as these were addressed from a naturalistic point of view in Scandinavian legal realism (Leiter, 2013a). According to Axel Hägerström's austere naturalism, only natural things exist and concepts that do not refer to observable facts are meaningless (Hägerström, 1953). Accordingly, legal concepts like 'duty' or 'right' are dismissed as meaningless and the normativity of law remains a 'mystical force,' since it does not exist in the real world. In this extreme view, legal doctrine – much like theology – is accused of talking nonsense about nonexisting objects. Alf Ross later gave an influential naturalized account of law that acknowledged the normativity of legal norms and explained their binding force in terms of behavior and mental states (Ross, 1959). Normativity, in his view, rests in the fact that legal norms are observed with a psychological sense of obligation. These early naturalist approaches to legal philosophy were heavily criticized by Hart and naturalism largely fell into disfavor until it was brought back to the forefront by Brian Leiter (Leiter, 2007). He argued, for example, that the central tenets of legal positivism, and in particular those of Hart, are best interpreted in naturalist terms (Leiter, 2013b).

Reductionism, Eliminativism, and Emergentism

Leiter and Marmor explain (1) legal norms by reference to underlying social practices and (2) social practices by reference to behavior and mental states. In doing so both offer reductionist theories, because they assert that one set of phenomena can be accounted for by another set of phenomena (Marmor, 2013). But they differ in the methodological justification of their account, i.e., whether they attribute the success of their explanation to philosophical analysis or to compatibility and continuity with social sciences. The second aspect of reducing social phenomena to behavior and mental states concerns the well-known dispute between individualist (reductionist) and holistic (emergentist) theories in social science. Critical with regard to the success of reductionist legal positivist theories, and disputed in legal philosophy, is the possibility of reducing law's normativity to social norms. In general, three possible relations can be drawn (O'Connor and Wong, 2012) between legal norms and social practices, and each could be thought of as either an epistemological or an ontological relation: First, an eliminative reduction would regard the normativity of law (or some aspects of it) as ontologically inexistent or epistemologically inaccessible to rational analysis, as Hägerström's radical legal realism did. Following this reasoning, there is nothing that needs to be explained. Secondly, one could assert that legal norms exhibit properties like 'validity' or 'normativity,' but hold that these properties can be fully explained by underlying social practices (like microeconomics explains aggregate outcomes in terms of individual behavior). This seems to be the kind of reductionism implied by Leiter and Marmor. Thirdly, one could hold that 'validity' or 'normativity,' as properties of legal norms, cannot be fully explained by social facts, either because we lack the knowledge necessary to do so or because normativity is an emergent and hence irreducible property that ontologically supervenes on the respective social practices. Then reductionism would fail. But this would not necessarily mean that legal positivism is wrong and a natural law theory true, because the irreducible property does not have to consist in true morality but could just be the normativity of law. Of course, different aspects of legal norms could require different versions of reductions. For example, it is a plausible view that legal norms do not exist as incorporeal entities but that they do exist as aggregates of behavior and mental states.

See also: Conventions and Norms: Philosophical Aspects; Hermeneutics; Justice: Philosophical Aspects; Law and Morality: A Continental European Perspective; Law and Morality: An Analytical Perspective; Legal Formalism; Legal Positivism; Legal Realism; Natural Law; Naturalism; Social Constructivism.

Bibliography

Alexy, R., 2002. The Argument from Injustice. Oxford University Press, Oxford.
Aquinas, T., 2006[1265–1274]. Summa Theologiae. Cambridge University Press, Cambridge.
Austin, J., 1995[1832]. The Province of Jurisprudence Determined. Cambridge University Press, Cambridge.
Bentham, J., 1970[1782]. On Laws in General. The Athlone Press, London.
Coleman, J., 2002. Methodology. In: Coleman, J., Shapiro, S. (Eds.), The Oxford Handbook of Jurisprudence & Philosophy of Law. Oxford University Press, Oxford, pp. 311–351.
Dickson, Julie, 2012. Legal positivism: contemporary debates. In: Marmor, A. (Ed.), The Routledge Companion to Philosophy of Law. Routledge, New York, pp. 48–64.
Dworkin, R., 1977. Taking Rights Seriously. Duckworth, London.
Dworkin, R., 1986. Law's Empire. Harvard University Press, Cambridge, MA.
Finnis, J., 2011[1980]. Natural Law and Natural Rights. Oxford University Press, Oxford.
Fuller, L., 1969. The Morality of Law. Yale University Press, New Haven.
Gardner, J., 2001. Legal positivism: 5 ½ myths. American Journal of Jurisprudence 46, 199–227.
Green, L., 2009. Legal positivism. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. http://plato.stanford.edu.

Hägerström, A., 1953. Inquiries into the Nature of Law and Morals. Almquist & Wiksell, Stockholm.
Hart, H.L.A., 2012[1961]. The Concept of Law, third ed. Oxford University Press, Oxford.
Kelsen, H., 1967[1960]. Pure Theory of Law (M. Knight, Trans., second ed.German). University of California Press, Berkeley.
Leiter, B., 2007. Naturalizing Jurisprudence. Oxford University Press, Oxford.
Leiter, B., 2012. Naturalism in legal philosophy. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. http://plato.stanford.edu.
Leiter, B., 2013a. Legal realisms, old and new. Valparaiso University Law Review 47, 67–81.
Leiter, B., 2013b. Why Legal Positivism (Again)? Public Law and Legal Theory Working Paper No. 442 University of Chicago, Chicago.
Mantzavinos, C., 2009. Philosophy of the Social Sciences. Cambridge University Press, Cambridge.
Marmor, A., 2009. Social Conventions. Princeton University Press, Princeton.
Marmor, A., 2010. The pure theory of law. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. http://plato.stanford.edu.
Marmor, A., 2013. Farewell to conceptual analysis (in jurisprudence). In: Waluchow, W., Sciaraffa, S. (Eds.), Philosophical Foundations of the Nature of Law. Oxford University Press, Oxford, pp. 209–229.
McAdams, R., Rasmusen, E.B., 2007. Norms and the law. In: Polinsky, A. Mitchel, Shavell, S. (Eds.), Handbook of law and economics, vol. 2. Elsevier, Amsterdam, pp. 1573–1618.
Moore, M.S., 1992. Law as a functional kind. In: George, R.P. (Ed.), Natural Law Theory. Oxford University Press, Oxford, pp. 188–242.
O'Connor, T., Wong, H.Y., 2012. Emergent properties. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. http://plato.stanford.edu.
Postema, G., 1982. Coordination and convention at the foundations of law. Journal of Legal Studies 11, 165–203.
Radbruch, G., 2006[1946]. Statutory lawlessness and supra-statutory law. Oxford Journal of Legal Studies 26, 1–11.
Raz, J., 2005. Can there be a theory of law? In: Golding, M.P., Edmundson, W.A. (Eds.), The Blackwell Guide to the Philosophy of Law and Legal Theory. Blackwell Publishing, Malden, MA, pp. 324–342.
Ross, A., 1959. On Law and Justice. University of California Press, Berkeley.
Schauer, F., 2010. Was Austin right after all? Ratio Juris 23, 1–21.
Shapiro, S., 2006. What is the internal point of view? Fordham Law Review 75, 1157–1170.
Shapiro, S., 2011. Legality. Harvard University Press, Cambridge, MA.
Stavropoulos, N., 2014. Legal interpretivism. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. http://plato.stanford.edu.

Philosophy, Psychiatry, and Psychology

Nicholas Deakin, Barts Health NHS Trust, London, UK; and Queen Mary University of London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H.M. Emrich, W. Paetzold, volume 17, pp. 11376–11381, © 2001, Elsevier Ltd.

Abstract

Since the Ancient Greek origins generally considered as the inception of philosophy, questions relating to self, identity, mental illness, and norms have been the focus of many scholars. Ancient concepts have been refined through the Middle Ages and by contemporary philosophers. Modern psychiatry and psychology can be influenced by philosophical concepts, which can aid patient-centered care, the development of service delivery methods, research, training, and to foster an international community aligned for the advancement of patient care and the sharing of knowledge.

Introduction

An understanding of philosophy is central to insightful practice and appreciation of the work of psychologists and psychiatrists. In fact, independent philosophical concepts are the only way of answering some complex questions in the field. How can we define the nature of the psyche? What is disturbed psychic experience and what do we mean by illness and normalcy? As the mind–body problem purports, does the psychic form part of the material, or is it autonomous, following independent laws of its own? Does inner psychic experience exists and is it accessible by means of introspection, or as behaviorism suggests, should we only describe visible behavior? Is the essence of the psychic found in the active relationship of the 'I' to the 'You,' to the other (intentionality), and is it indeed only comprehension of the other, via *mimesis*, that constitutes one's own psychic experience? These are the sorts of questions posed by philosophical psychology, by psychiatric anthropology, and they can only be satisfactorily tackled if we understand the development of philosophical thinking.

From the very beginning, scientific psychiatry has raised fundamental questions within the fields of philosophy and psychology, namely as regards the nature of the 'psyche' and the concepts as to what constitutes mental illness. The question of what it means to say that something is 'psychical' leads to speculation in psychiatry insofar as special heuristic aspects are involved and this can be explosive if one professional group feels excluded by a certain definition or approach.

While there is on-going debate around psychopathology, pathology and the definitions of key psychiatric concepts, psychiatry has been criticized both internally and externally as being 'unscientific' (Phillips, 2000) and mental health professionals suffer at least part of the stigma that is projected by wider society onto psychiatric patients. Fulford, Stanghellini, and Broome make a compelling argument against these criticisms. They suggest that philosophy can be used to prove that mental health care delivery is both more challenging than other areas of health care and also that psychiatry leads the way among medical fields who are yet to encounter the kind of epistemological and philosophical debates that punctuate daily practice in psychiatry (Fulford et al., 2004). They suggest five key ways in which philosophy can aid practice in mental health care – patient-centered care, supporting new methods of service delivery, advancing research in neuroscience, psychiatric training, and supporting an 'open,' international society of psychiatrists – which shall be expanded on later in this article (Fulford et al., 2004).

Philosophical Psychiatry

In accordance with Brentano's basic idea that the psychical can be identified with the 'intentional' (Brentano, 1874), one might argue that psychiatry is so important to philosophers because the disturbances of intentionality in various domains of the mental world of humans are what might be termed as 'negative images' of the psychic functions, which normally operate in the human mind. The special method by which Karl Jaspers (Jaspers, 1913) elaborated this view of the human mind is the 'phenomenological approach' to various kinds of disturbances of mental functioning in psychiatric patients (deficiency model of psychic dimensions in Jaspers's *General Psychopathology*).

However, this is not the only way in which the relationship between psychiatry and philosophy can be construed. The other aspect is the issue of the potential role of 'normalcy' in social life. As a fluctuating criterion for assigning people to a psychiatric context, normalcy cannot be regarded as a phenomenon of 'average' and represents merely the concept of 'ideal typology' (Blankenburg, 1971). From a philosophical point of view, the subject of psychiatric 'normalcy' constitutes a major problem, since it raises the issue of the existence of an 'individual norm' in the sense of the 'alterity of the other' (Lévinas, 1987). Seen in this way, psychiatric disorders are not simply 'defects' but represent variants of normal psychic life which – when exaggerated – are interpreted as disorders and treated by psychiatrists.

This concept fits in with the approach of evolutionary psychiatry, which considers that psychiatric disorders have evolved because they relate to psychic functions which – in their submaximal expression – yield advantages where natural selection is concerned. Viewed in this way, as Blankenburg (1971) suggests, psychopathology does not constitute a 'statistical, average norm' but an 'individual norm.' In a similar sense, Spaemann and Löw (1981) have argued that health itself contains a teleological notion: "If 99 percent of the population suffered from headaches, headaches still would not be 'normal'." Philosophical concepts in psychiatry, therefore, have

to address the problem of 'normalcy'; they face the difficult challenge of acknowledging deviant cognitive and mental events merely as expressions of the alterity of the other.

Greek Philosophy: Hippocrates, Socrates, Aristotle, and Plato

A full understanding of philosophy is not possible without, at least, a basic awareness of the Greek founding fathers of philosophy – Socrates, Aristotle, and Plato. The famous physician, Hippocrates, is also an important figure who shaped modern medical practice. He was the first thinker to propose that disease was caused naturally by a constellation of environmental and living factors, rather than being a punishment for sin (Adams, 1891); was credited with first propositioning the humoral approach to disease causation incorporating psychological dispositions (Chadwick, 1983) and proposed the seminal Hippocratic Oath, which sought to ensure that doctors practiced in a respectful, professional, and beneficial way toward their patients.

The focus of Greek philosophy more widely is on the individual thinking and acting intentionally, and self-reflection is thus the cornerstone of ancient Greek thought. On the basis of the hypothesis of quasideterminism, the Athenian philosopher Socrates (470–399 BC) asserted that the will responds to the dictates of reason. He was puzzled and perplexed by the question of 'how should one live?' (Robertson, 2009). He developed the Socratic – or 'maieutic' – method, which elicits truths that lie hidden in an individual's unconscious by means of a process of question and answer. This method – in a completely different context – later became the basis of Sigmund Freud's psychoanalysis and the nondirective psychotherapy of Carl R. Rogers. The concepts proposed by Socrates are also central to professional ethics, and the answer to how one should live has evolved to incorporate a set of values. For psychiatrists, there are particular professional values that society both deserves and expects, as elicited by the World Psychiatric Association (World Psychiatric Association, 2011) and national bodies such as the United Kingdom Royal College of Psychiatrists (Royal College of Psychiatrists, 2014).

In works including his *Phaedo*, Socrates' disciple Plato (427–347 BC) differentiated between the spiritual and material worlds and applied this dichotomy to man by postulating that the psyche (soul), together with its logical functions, possesses a preexistence before the birth of the individual and survives into immortality (Plato, 1966). In *Timaios*, Plato defines madness (*manteia*) as a disturbance of the vegetative components of the soul (which he felt were located in the liver) but in the later *Phaidros* he tells us that the Gods can also send madness as an elevating spiritual force. The aim of Plato's philosophical psychology is to create harmony between the different psychic activities. Whereas Plato's idealism defines ideas as the only true reality, his pupil Aristotle (384–322 BC) based his psychology partly on empirical investigation, thus representing a type of *realism*. In his systematic discussion *De anima* (*On the Soul*), Aristotle describes the soul as the *entelechy*, the self-contained purpose of the body. This purpose must be fulfilled in order to achieve *eudaimonia* (human flourishing). In Aristotelian thinking, philosophical self-reflection leads to psychological self-actualization, and in his *Ethica Nikomacheia* Aristotle states that the purpose of ethics is not knowledge but action (Aristotle, 1984).

Although Aristotle was the most systematic Greek philosopher and had an enormous influence on the development of modern science, the treatment of psychological matters in post-Aristotelian Hellenistic philosophy is often described as eclectic. Stoic and Epicurean philosophy, in particular, must be considered as practical psychology. Given that the increasingly urbanized Hellenistic world had a greater resemblance – in some ways – to modern society than the Middle Ages, this practical psychology is still of interest. Epicureans and Stoics tried to find ways to alleviate anxiety, but their anxiety seemed to be more drived by the disturbance caused by a complex, fragile society than fear of God. Epicurean and especially Stoic thinking played an important role in Hellenistic culture and can thus be viewed as the first widespread popular psychological fashions.

After the fall of the Roman Empire, Christian Europe lost sight of some of Aristotle's major works. The philosophical psychology of the dominant neo-Platonism referred back to the static aspects of Platonism, shifting away from rationalism to metaphysical speculation on the relationship between a tangible and an intangible world, between the soul and God. St Augustine (CE 354–430) attempted to synthesize neo-Platonism with Christianity. Following on from Seneca and Plotinus, he leaned toward introspection as his philosophical and psychological method, considering that – alongside objective observation – self-observation of inner mental processes could function as the basis of knowledge. He, therefore, chose self-reflection as the basis of his epistemology, arguing that doubt presupposes the existence of the doubter (Augustine, 1984), an idea that later became extremely influential as Descartes famous *Cogito ergo sum*, more popularly known as the phrase, 'I think therefore I am.'

Metaphysics of Middle Ages and the Mind–Body Problem

The philosophical psychology of the Middle Ages was metaphysical and also deductive. The changeability of the soul and of world were interpreted philosophically as an inevitable but inferior property of the unchangeable, and consequently Christian dogmatism and neo-Platonism showed less interest in the experience of psychological interaction. With the foundation of the first universities from the thirteenth century onward, the process of philosophical debate increased and a more dynamic philosophical psychology developed as Aristotelian thinking was resuscitated through the medium of Arab philosophy. St Thomas Aquinas (1225–74) attempted to integrate Christian dogma with elements of Greek and Arab philosophy. Aristotle differentiated between an active, immortal intellect and a merely passive, understanding one. Aquinas accepted this view, describing the intellect as a passive power, a *tabula rasa*, which had to be filled with experience by the five exterior senses. In addition, there exists a rational power – which makes free will possible – and the soul also has vegetative components. In contrast to the neo-Platonic dichotomy of soul and body, Aquinas attempted to define

man as a psychophysical unity, as a body inspired by a soul. Taking his lead from Aristotle, Aquinas defined the soul as the form of the body and as an immortal spiritual entity. According to this theory, the body of a dead person consists of the same material or substance as that of a living person. As a result of this dichotomy, it is impossible to describe the impact of mental acts on real action without becoming entangled in serious logical difficulties. This is the *mind–body problem*: the issue of whether or not the soul or mind is an incorporeal substance, a nonmaterial entity obeying laws other than those, which govern the material world. Discussion of this problem has continued until the present day.

Modern Philosophical Psychology

Modern philosophical discussions related to psychiatric problems from the outset. René Descartes (1596–1650) reflected on the fundamental nature of illusions and the constitution of reality. He compared the reality of dreams with the consciousness in wakefulness. In his *Meditations on first Philosophy*, Descartes arrived at the irreducibility of the cogito, putting an end to the infinite regress of doubt about possible realities (Descartes, 1996). As a consequence of this concept, the modern subjectivistic understanding of consciousness has evolved, leading to the split between *cogitatio* and *extension* (soul and body), between the mental and the physical world, a split which represents one of the irresoluble problems in psychiatry.

Refuting earlier deductive speculation, the empiricist Locke (1632–1704) again emphasized the Aristotelian view that the human mind is like a blank sheet of paper and that all the material of reason and knowledge comes from observation, experience, and reflection. David Hume (1711–76) described ideas as the copies of sensory impressions. Most of our beliefs go beyond current impressions, but all that we can be certain of is perception, meaning that it is impossible, for example, to comprehend the *self* as a separate identity. Radicalizing this point of view, the positivism of Comte (1798–1857) postulated that the sensible comprises the whole range of knowledge and that any knowledge that cannot be comprehended by the senses must be rejected.

Immanuel Kant (1724–1804) speculated about constitutional aspects of psychiatric disorders. In his *Critique of Pure Reason*, he tried to deduce the a priori conditions of the possibility of experience (Kant, 1974). This type of categorical analysis of mental life has been decisive for all subsequent philosophical considerations about the formation and constitution of subjectivity and consciousness.

In this sense, idealism represents one of the possible forms of reactions to Kantian philosophy. For example, Georg Friedrich Hegel (1770–1831) developed an ontological approach to intellectually reconcile all elements of existence with one another within a systematic structure in which all contradictory moments of reality appear as dialectic mediations of the universal mind (Hegel, 1973). The philosophical psychologist Franz Brentano (1837–1917) elaborated a comparatively modest conceptualization of the difference between the mental and the physical, thereby developing a fundamental concept of philosophy of the psyche, namely the concept of 'intentionality.' In his work *Psychology from an Empirical Standpoint*, he argued: "Each mental phenomenon is characterised by what the scholars of the middle ages called the intentional (also probably mental) inner existence … of an object, and which we … would call the reference to content, direction toward an object" (Brentano, 1874).

One may argue that the main characteristics of psychiatrically ill patients are disturbances of intentionality. Thus psychic functions are 'intentional' insofar as they refer to an ('internal') object; they can 'mean' or refer to something. If this fundamental function of mental life is disturbed, a psychiatric disorder can be anticipated.

The Danish philosopher Kierkegaard (1813–55), deeply aware of Hegel's philosophical concept of the absolute and at the same time vigorously writing against it, developed a concept of subjective reality, prestructuring the later existentialistic conceptualizations of subjectivity, defining it as synthesis of soul and body, carried by spirit. His pioneer work about personal freedom formulated the questions as to *The Concept of Anxiety* (Kierkegaard, 1980). As has been pointed out by Hersch (1989), these basic ideas are represented by 'existence' and 'moment,' in which the subject works through to the moment of the 'reality of freedom,' the condition of 'anxiety.' That means that 'anxiety' cannot be understood primarily as an emotional state, but rather an existential condition, an intensified form of existence, leading to an autonomous subject ('choosing the self'). In this context 'Phenomenological reduction' has to be understood as the attention to the 'nature of the condition of things' and thus the 'things as phenomena.' A great and far-reaching concept in philosophical psychology, the philosophy of the 'language game,' has furthermore been elaborated within the philosophy of language by Wittgenstein (1898–1951). Jaspers has pointed out the importance of philosophy of language as the 'work of man.' The universals of language also play a role in the philosophy of Gehlen (1904–76) who referred to man as a 'nondetermined animal,' whereby Gehlen interprets the human nature as creating its own world as an 'open world' (Gehlen, 1987). Of special importance for psychiatric philosophical psychology is the concept of Helmuth Plessner (1892–1985) who developed the idea of 'positionality' of human subjects. Whereas animals show a 'centered positionality,' human beings live, according to Plessner, in an 'eccentric positionality' (Plessner, 1970). This concept plays a relevant part in psychiatric psychopathology, since disorders like delusions of reference and persecutory ideas may be interpreted as a disordered eccentric positionality (Emrich, 1992). In Jean-Paul Sartre's (1905–80) *Being and Nothingness*, the 'decentering of the world' plays a fundamental role in the phenomenological analysis of the 'glance' (Sartre, 1996). Sartre's analysis of the interpersonal relationship ends up in an intersubject fight of 'glances.' Later Emmanuel Lévinas (1905–95) acknowledged the 'otherness of the other' in the dimension of 'face': "The epiphany of the face is itself a word of honour in a certain way" (Lévinas, 1987). This view of responsibility in the interpersonal encounter is of great importance in psychiatry, stressing the importance of empathy in therapeutic relationships.

As a new academic discipline, psychology was obliged to assert itself against the established materialist sciences, especially medicine and biology. Positivist thinking consequently led to

strongly antimentalist and anti-introspectionist tendencies. The theory that it is only via *objective* observation that scientific method can formulate hypothetical laws from which it can then derive deductive arguments led to the radical behaviorist position that nothing that is the product of *subjective* introspection can be accepted. Positivism thus reduced psychology to a science that aims to predict and control behavior. Watson (1878–1958) and Skinner (1904–90) lead to academic psychology drifting away from philosophy, becoming more practical than theoretical and less concerned with understanding behavior (Watson, 1998; Skinner, 1979). The contemporary philosophy of science, however, would claim that it is an illusion to believe that science can be reduced to a simple, *objective*, and entirely rule-bound methodology. Thomas Kuhn argued that such opinions are based on science textbooks that fail to tell the truth about the history of science (Kuhn, 1996). In contrast to academic psychology, psychiatry has retained a more heterogeneous – even if contradictory – theoretical background. The main question at the start of the twenty-first century is whether psychology and psychiatry can only be governed by certain particular scientific methodologies or if it is possible to accept the existence of different approaches to acquire knowledge. Paul Feyerabend pointed out that "the idea that science can, and should, be run according to fixed and universal rules, is both unrealistic and pernicious" (Feyerabend, 1980). In actual fact, therapeutic practitioners find it unsatisfactory to consider introspective experience and objective behavior as mutually exclusive, as if they were distant continents with no contact between them. Modern therapeutic theories are thus becoming increasingly eclectic, and the gulf that used to exist between introspective psychoanalysis and behavior therapy, for example, is in the process of disappearing.

Self and Identity

Philosophical concepts of 'self' have led to psychological consideration of the nature of 'identity.' Disturbances of the inner equilibrium of the components of the self, which shape identity, are brought about by processes of remembering, making clear the potential for explanation inherent in a theory of identity formation based on a psychology of the self. As Lévinas in particular has shown this cannot be restricted to a theory of the structure of the 'self' (Lévinas, 1987); rather it must take account of interpersonality; the existence of others (the 'trace of the other'). Any attempt to describe the danger to processes of identity formation posed by the existence of the other from the perspective of the psychology of the self must take account of the 'ontology of the in-between' already developed by Martin Buber (1994). For, as Theunissen (1965) demonstrates in his major work on intersubjectivity, taking the other seriously in its own being involves transcending the intentionality mode.

This philosophical/psychological concept is based on the assumption that components of a person's self can manifest themselves in others and can exist further in them.

The basis of this primary interpersonal constitution of the psyche is the mother–child dyad of early childhood. In a way, people unconsciously always retain parts of their childhood dyad 'within them,' although they nevertheless consider themselves to be ego-centered 'individuals.' In a system of ideas in which the *cogito* is identified with the empirical ego, the 'interpersonal self' cannot be conceived. This primary (but extremely repressed or internally never realized) interpersonal constitution of the self has consequences for how people deal with the familiar and unfamiliar, both in the other and within themselves, where aggression is concerned. Christian Scharfetter, referring to Graumann's theory of subjective perspectivity, rightly refers to the fact that precisely the most intimate personal area of a person, his own unconscious, constitutes an alien entity, an unfamiliar area of sinister mystery. Scharfetter (1994) puts this as follows: "The other is not simply an alien entity, however." The other may very well be familiar; it can be close, it can be experienced as analogous (not homologous). Indeed, the alien is able to strengthen and reinforce the self. It is through the 'You' that the 'I' recognizes itself. This is an important factor where psychotherapy is concerned. However, confrontation with the alien within the self is painful; it requires not only cognitive efforts (Hegel's 'strenuous toil of conceptual reflection'), but also psychological effort in the form of crises in identity development.

An extremely effective defense mechanism here is that of 'projective identification.' This means that the unknown within the subject is discovered in others by projection and combated there. Streeck-Fischer says, for example, that 'the confrontation with the alien has become a shock experience, one which hampers successful interaction with the alien. The alien is threatened as a means of stabilizing one's own threatened self' (Streek-Fischer, 1994). As Streeck-Fischer shows, this mechanism becomes particularly effective when one's own position is not only under pressure from the danger posed by crises in the formation of identity but is also reinforced by social pressure. Social devaluation of specific features of the self can significantly increase projective identification in the sense of the development of aggression. Hans-Joachim Maaz, for example, shows how in East Germany "social constraints under a repressive authoritarian system led to the self becoming alien and the alien becoming one's own." Maaz describes the resulting aggression as follows (Maaz, 1994): "violence and hatred of other people are always evidence of internal tensions within someone, of his no longer being able to find and develop the self and the essential but of having to live a life too much of which is imposed by others."

One of the great figures in the theory of *mimesis* is René Girard. In various works, notably *La violence et le sacre*, Girard developed a theory of mimetic occurrences (Girard, 1977). If one accepts his thinking, the main point of the Oedipal drama, which in Freudian theory is said to mold early childhood, is not the Oedipal aspect as such – the desire for the mother in the Freudian sense – but the mimetic rivalry with the father. In an early work of 1961, *Mensonge romantique et vérité romanesque*, Girard elaborated on a discovery central to the entire development of his theory. He was able to identify a phenomenon of 'mimetic desire' in nineteenth- and twentieth-century novels from Stendahl to Flaubert and Proust. The novels he examined shared a peculiar basic structure: the main characters desire someone else, but their feelings are directed toward a third person, a '*mediateur*.' Analyzing Stendhal, Girard says: "in most cases of Stendhalian desire, the *mediateur* himself desires the object, or could desire it. This desire, whether real or assumed,

is precisely what makes the object infinitely desirable in the eyes of the subject."

Desires, needs, and yearnings are thus described as being something not genuinely given, something brought forth internally by drives or makeup, but are 'read,' taken on from a third party because of mimetic dependency. The basic events in the process of identification can thus be described as follows: if we can understand the other in a recreative sense only by putting ourselves into the other's intentions, into that person's world of desires, this means that we have to identify with that other. The amazing consequence of this is that processes of identification are necessary for a proper interpretation of reality. In this sense, one can say that a rational explanation of the world is possible only if there is a fundamental willingness to identify with others.

As a result, one can argue that in psychiatric patients this rational world access is hampered by their impaired interpersonal relations, that is, due to disturbed mimetic processes (Emrich, 1995).

Modern Use of Philosophy in Psychiatry

Fulford, Stanghellini, and Broome suggest five key ways in which philosophy can aid practice in mental health care (2004), which should be considered by all who work in the field – patient-centered care, supporting new methods of service delivery, advancing research in neuroscience, psychiatric training and supporting an 'open,' international society of psychiatrists.

The concept of patient-centered care focuses on the patient and their family to elicit their personal values before involving the wider team and applying specific clinical skills (Fulford, 2004). This could help facilitate engagement and thus improve clinical outcomes. Where values conflict, the authors suggest seeking the advice of the multidisciplinary team before building an awareness of the situation, applying reasoning skills, obtaining knowledge relevant to the decision – including from relevant research – and using communication skills appropriately to reduce and resolve conflicts (Fulford et al., 2004).

The challenges of the general, world-wide move toward multidisciplinary team working in mental health can also be framed from a philosophical point of view. In theory, this allows a variety of professionals who have different training programs, skill sets, and priorities to meet and discuss patient care pathways. On the other, this team work has been criticized for allowing some patients to fall 'through the net' through lack of collaborative decision-making (Onyett, 2003) and many some have suggested that this may be due to the challenge to the doctor's traditional responsibilities for patient care that come from other professional groups (King, 1991). Columbo and colleagues suggest that differences in team member approaches may be framed by unrecognized differences in models of pathology (2003). Fulford, Stanghellini, and Broome suggest that this very conclusion leads to a more complete analysis of the situation, using J. Austin's theory around 'clear thinking' to assert this contribution of philosophy to psychiatry (2004). They suggest that acknowledgment of the different phenomenology is already being utilized to drive service improvements, with examples including Bracken's novel approach to managing post-traumatic stress disorder in traditional societies (2002) and Sabat's use of conversational analysis to improve communication with those who suffer from Alzheimer's dementia (2001).

The development of neuroscience and neuroimaging technology over the past few decades has presented novel challenges to the traditional phenomenological approaches. In particular, research can present new information to contribute to the information on the interplay between genetic, environmental, and psychological factors in leading to mental illness, which can challenge traditional knowledge and definitions.

Andreasen predicted that neuroscience would bring these issues, including issues around self and personal identity, to the fore (1991). The research in this area is evolving rapidly, on both a condition and phenomenological basis – posing challenges to psychiatry and philosophy.

Training for psychiatrists and ideally other health professionals is focused around encouraging exactly the kind of independent thought and contribution valued by the Greek fathers of philosophy and discussed in an earlier part of this article. The concepts central to conflict resolution and communication are a core part of the curricula for health professionals, particularly doctors. Concepts of ethics, including a basis in Greek philosophy, are a key component of undergraduate medical curricula for doctors in training (e.g., Stirrat et al., 2010) and postgraduate practice (World Psychiatric Association, 2011).

The final main contribution of philosophy to psychiatry noted by Fulford, Stanghellini, and Broome is in ensuring that psychiatrists and other professionals are part of an international community that seeks to be supportive and protective of patients (2004). The contributions to psychiatric practice of politics, scholarship, historical context, individual values, professional and institutional standards (Matthews, 2005; Robinson and Walter, 2013) make psychiatry unique among medical disciplines. Episodes where psychiatrists contributed to the widespread euthanasia of those from unfavorable groups in Nazi Germany (Robinson and Walter, 2013) and labeling of political nonconformists as 'schizophrenic' in the USSR (Bloch and Reddaway, 1977) show the vulnerability of psychiatry arising from a failure to maintain a balance of different perspectives and pursuing what Locke described as 'enthusiasms' (Fulford et al., 2004). Maintaining an 'open society' will ensure that a shared discipline, analysis, respect for individual cultures and methodological pluralism protect against future 'enthusiasms' on an international and national level (Birley, 2000). The World Psychiatric Association, European Psychiatric Association, and events such as the United Kingdom Royal College of Psychiatrists annual International Congress are central to the international collaboration, which Birley suggests is essential to safeguard the good standing of the profession in society into the future.

See also: Ethics and Psychiatry; Normality and Mental Health.

Bibliography

Adams, F., 1891. The Genuine Works of Hippocrates. William Wood and Company, New York.

Andreasen, N., 2001. Brave New Brain: Conquering Mental Illness in the Era of the Genome. Oxford University Press, Oxford.

Aristotle, 1984. The Complete Works of Aristotle: The Revised Oxford Translation. Princeton University Press, Princeton, NJ.
Augustine, 1984. City of God. Penguin, Harmondsworth, UK.
Birley, J., 2000. Psychiatric ethics: an international open society. In: Dickenson, D., Fulford, K. (Eds.), Two Minds: A Casebook of Psychiatric Ethics. Oxford University Press, Oxford, pp. 327–335.
Blankenburg, W., 1971. Der Verlust der natürlichen Selbstverständlichkeit. Ein Beitrag zur Psychopathologie symptomarmer Schizophrenien. Ferdinand Enke, Stuttgart, Germany.
Bloch, S., Reddaway, P., 1977. Russia's Political Hospitals: The Abuse of Psychiatry in the Soviet Union. Gollancz, London.
Bracken, P., 2002. Trauma: Culture, Meaning and Philosophy. Whurr, London.
Brentano, F., 1874. Psychologie Vom Empirischen Standpunkt, vol. 1. Duncker & Humblot, Leipzig [1995. Psychology from an Empirical Standpoint. Routledge].
Buber, M., 1994. Ich und Du. Lambert Schneider, Gerlingen [1978. I and Thou. Scribner, New York].
Chadwick, Edited with an introduction by G.E.R. Lloyd; translated from the Greek by J.W.N. Mann et al., (1983). Hippocratic Writings, new ed., with additional material, Reprinted in Penguin classics ed. Penguin, Harmondsworth. p. 262.
Columbo, A., Bendelow, G., Fulford, B., Williams, S., April 2003. Evaluating the influence of implicit models of mental disorder on processes of shared decision making within community-based multi-disciplinary teams. Social Science & Medicine 56 (7), 1557–1570.
Descartes, R., 1996. In: Cottingham, J. (Ed. and Trans.), Meditations on First Philosophy: With Selections from the Objections and Replies. Cambridge University Press, Cambridge, UK.
Emrich, H., 1992. Systems theory of psychosis: 'filtering,' comparison, error correction, and its defects. In: Emrich, H.M., Wiegand, M. (Eds.), Integrative Psychiatry. Springer, Heidelberg, Germany, pp. 81–89.
Emrich, H., 1995. Physiognomy of the psychological – toward a theory of mimesis. Paragrana 4 (2), 126–143.
Feyerabend, P., 1980. Against Method. Verso, London.
Fulford, K., Stanghellini, G., Broome, M., October 2004. What can philosophy do for psychiatry? World Psychiatry 3 (3), 130–135.
Fulford, K., 2004. Ten principles of values-based medicine. In: Radden, J. (Ed.), The Philosophy of Psychiatry: A Companion. Oxford University Press, New York, pp. 205–234.
Gehlen, A., 1987. Der Mensch. Aula Verlag, Wiesbaden [1988. Man. Columbia University Press, Columbia].
Girard, R., 1961. Mensonge romantique et vérité romanesque. Grassett, Paris.
Girard, R., 1977. Violence and the Sacred. Johns Hopkins University Press, London.
Hegel, G., 1973. Phänomenologie des Geistes. Suhrkamp, Frankfurt [1979. Phenomenology of Spirit. Galaxy Books, New York].
Hersch, J., 1989. Das Philosophische Staunen. Piper, München, Germany.
Husserl, E., 1950. Gesammelte Werke. Kluwer, Dordrecht, The Netherlands.
Jaspers, K., 1913. Allgemeine Psychopathologie. Springer, Berlin [1998. General Psychopathology. Johns Hopkins University Press, Princeton, NJ].
Kant, I., 1974. Kritik der reinen Vernunft. Suhrkamp, Frankfurt [1996. Critique of Pure Reason. Hackett, Indianapolis, IN].
King, D., 1991. Moving on from Mental Hospitals to Community Care. The Nuffield Provincial Hospitals Trust.
Kierkegaard, S., 1980. The Concept of Anxiety: A Simple Psychologically Orienting Deliberation on the Dogmatic Issue of Hereditary Sin. Princeton University Press, Princeton, NJ.
Kuhn, T., 1996. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Lévinas, E., 1987. Totalität und Unendlichkeit. Versuch über die Exteriorität. Alber, München [1969. Totality and Infinity: An Essay on Exteriority. Duquesne University Press, Pittsburgh, PA].
Maaz, H., 1994. Das Eigene und das Fremde im deutschen Vereinigungsprozeß. In: Egner, H. (Ed.), Das Eigene und das Fremde: Angst und Faszination. Walter Verlag, Hamburg, pp. 13–27.
Matthews, E., 2005. Philosophy, Psychiatry, & Psychology. In: Moralist or Therapist? Foucault and the Critique of Psychiatry, vol. 2, pp. 19–30.
Onyett, S., 2003. Teamworking in Mental Health. Palgrave McMillan, Basingstoke.
Plato, 1966. The Collected Dialogues of Plato. Pantheon, New York.
Plessner, H., 1970. Philosophische Anthropologie. Fischer, Frankfurt.
Phillips, J., 2000. Conceptual models for psychiatry. Current Opinion in Psychiatry 13, 683–688.
Robertson, M., 2009. An Overview of Psychiatric Ethics. A HETI Resource, Health Education and Training Institute, New South Wales, 14.
Robinson, M., Walter, G., 2013. Ethics and Mental Health: The Patient, Profession and Community. CRC Press, Florida.
Royal College of Psychiatrists, 2014. CR 186: Good Psychiatric Practice. Code of Ethics, London.
Sartre, J., 1996. Being and Nothingness. Pocket Books, New York.
Sabat, S., 2001. The Experience of Alzheimer's Disease: Life through a Tangled Veil. Blackwell, Oxford.
Scharfetter, C., 1994. Im Fremden das Eigene erkennen. In: Egner, H. (Ed.), Das Eigene und das Fremde. Walter Verlag, Hamburg, pp. 13–27.
Skinner, B., 1979. About Behaviorism. Pimlico, London [1999. Random House, New York].
Spaemann, R., Löw, R., 1981. Die Frage Wozu? Geschichte und Wiederentdeckung des teleologischen Denkens. Piper Verlag, München, Germany.
Stirrat, G., Johnston, C., Gillon, R., Boyd, K., 2010. Medical ethics and law for doctors of tomorrow: the 1998 consensus statement updated. Journal of Medical Ethics 36, 55–60.
Streek-Fischer, A., 1994. Haßt du was, dann bist du was'. Über Fremdenhaß und seine selbstreparative Funktion am Beispiel jugendlicher rechtsextremer Skinheads. In: Egner, H. (Ed.), Das Eigene und das Fremde. Walter Verlag, Hamburg, pp. 13–27.
Theunissen, M., 1965. Der Andere. Walter de Gruyter, Berlin.
Watson, J., 1998. Behaviorism. Transaction, New Brunswick, NJ.
World Psychiatric Association, 2011. Madrid Declaration on Ethical Standards for Psychiatric Practice. Consensus Statement. Reasserted 2011.

Relevant Websites

http://inpponline.com/about-inpp/ – International Network of Philosophy and Psychiatry.
http://www.maudsleyphilosophygroup.org/ – Maudsley Philosophy Group.
http://plato.stanford.edu/entries/psychiatry/ – Philosophy of Psychiatry Entry at Stanford Encyclopedia of Psychiatry.
http://www.rcpsych.ac.uk/workinpsychiatry/specialinterestgroups/philosophy.aspx – Philosophy Special Interest Group, Royal College of Psychiatrists (UK).
https://www.press.jhu.edu/journals/philosophy_psychiatry_and_psychology/ – Philosophy, Psychiatry and Psychology Journal.

Phobias Across the Lifespan

Alfons O Hamm, University of Greifswald, Greifswald, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Marked and persistent fear cued by specific objects or situations accompanied by the compelling desire to escape from or avoid the encounter with such objects or events are the key features of specific phobias. Depending on the type of situation that is feared or avoided, specific phobias are further subdivided into five different specifiers: (1) animal; (2) natural environment; (3) blood, injection, or injury; (4) situational; and (5) other subtypes. Specific phobias are frequent in the population and vary across life span. For adults, prevalence rates vary between 11.3% for lifetime and 8.8% for 12-month periods. For children and adolescents, prevalence rates are slightly lower. Females are affected more often than males. The nature of phobic fear varies with the proximity of the phobic object, showing a close interaction between cognitive and emotional systems in the brain. While the anticipation of a phobic object in a certain context is associated with general hypervigilance, selective attention and freezing are engaged once the phobic object is detected. When the phobic object approaches, autonomic arousal increases and the organism prepares for escape. The neural networks organizing these defensive behaviors can be activated by innate threats as well as by learned associations acquired either by direct experience or by observation and communication of other people or even media. Specific phobias can be effectively treated using single-session exposure. During exposure, session patients are confronted with their feared situations and objects in reality and remain in this situation until the fear response subsides. Such exposure treatment is highly effective with effect sizes varying between $d = 1.41$ and 3.80, depending of the primary outcome measure used. Exposure therapy is particularly effective for treating avoidance behavior.

Defining Specific Phobias

Marked and persistent (for 6 months or more) fear cued by specific objects or situations accompanied by a compelling desire to escape from or to avoid such objects or situations – if possible – are the main diagnostic criteria for specific phobia (criterion A and C in the *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition (*DSM-5*); American Psychiatric Association (APA), 2013). Moreover, the fear response is almost always evoked by the phobic object (criterion B in the *DSM-5*; APA, 2013) and increases in intensity with increasing proximity of the external threat (e.g., with the time of entrapment and reduction of oxygen supply, the distance to the spider, and so on) and fades immediately when the situational threat disappears (e.g., the door is opened or the spider is caught and put outside). Thus, the cued nature of the fear is an important defining feature for specific phobias. To meet the criterion for a mental disorder, the fear has to be out of proportion to the actual danger posed by the object and to the sociocultural context (criterion D), and the fear or avoidance has to cause significant distress or impairment in social, occupational, or other important areas of functioning (criterion F). This last so-called impairment criterion is probably the most problematic of all *DSM-5* criteria for specific phobia (see Emmelkamp and Wittchen, 2009). One central issue with regard to the impairment criterion is the question of comorbidity. Several epidemiological studies show very high comorbidity between different anxiety disorders, including specific phobias (see Wittchen et al., 2003). In a sample of more than 1000 anxiety patients, Brown et al. (2001) observed that 70% of patients with a specific phobia diagnosis had a comorbid anxiety or mood disorder. The comorbidity rate, however, dropped to 33% among a subset of patients for whom specific phobia was the principal diagnosis. In other words, whether or not the specific phobia is the principal (i.e., the most severe) disorder strongly predicts the level of functional impairment. Moreover, data from the Early Development Stages of Psychopathology (EDSP) sample suggest that comorbidity increases steadily with increasing duration of the specific phobia (see Bittner et al., 2004). Data from the National Comorbidity Survey Adolescent Supplement (NCS-A) show that 19.3% of all adolescents (ages 13 to 18 years) report a lifetime prevalence of specific phobia, but only 0.6% of all adolescents reported 'a lot' or extreme impairment and severe or very severe distress (Merikangas et al., 2010), suggesting that treatment is needed only for a minority of individuals. Given that increased comorbidity often is associated with stronger functional impairment, however, the requirement of meeting the impairment criterion for specific phobia might foreclose early access to effective treatments when the disorder has not become chronic and is still a pure disorder. This is particularly important because effective treatments are available particularly if the specific phobia is still at an early stage and relatively isolated (Davis et al., 2012). In principle, the criteria for the diagnosis of specific phobias do not differ for adults and children except that children express their marked fear more nonverbally by defensive behavior like clinging, crying, tantrums, or freezing.

Diagnostic Features of Specific Phobias

Because one of the key features of specific phobias is the close link of the fear responses to the eliciting situations, it may not come as a complete surprise that, traditionally, phobic disorders have been categorized according to the feared objects and situations, translated into Greek or Latin preceding the word phobia. Although attractive for the public media who love to call on Friday the 13th to ask phobia experts about patients who suffer from *paraskevidekatriaphobia*, such endless lists of

phobias are scientifically and clinically useless. As a consequence, *DSM-5* has confined the subdivision of specific phobias to five specifiers: (1) animal (fear cued by spiders, snakes, insects, mice, dogs); (2) natural environment (fear cued by heights, thunderstorms, being on or in the water); (3) blood, injection, or injury (fear cued by seeing blood, receiving injections or transfusions, or fear of other medical care, including dental care); (4) situational (fear cued by entrapment in enclosed places like elevators, tunnels, airplanes, buses); and (5) other (fear cued by other stimuli, like choking or vomiting). Although gender distribution, age of onset, and treatment response are similar for these subtypes of specific phobias (LeBeau et al., 2010), the physiological pattern of the fear response, as well as the level of impairment and comorbidity, differs between the subgroups of specific phobia. Specific animal-type phobia is characterized by a sympathetically dominated autonomic physiological fear response pattern preparing the organism for escape (see Hamm et al., 1997; Globisch et al., 1999) accompanied by the central concern of freaking out and loosing control (see Lipsitz et al., 2002). Specific blood-, injection-, or injury-type phobia is marked by a diphasic fear response, an initially sympathetically mediated increase in heart rate and blood pressure followed by a vagally mediated bradycardia and hypotension, which ultimately may lead to a syncope preparing for tonic immobility (Friedman et al., 1993; Hamm et al., 1997), accompanied by the concern of fainting or feel nausea. Specific situational-type phobias (e.g., being entrapped) are characterized by strong cognitive symptoms and concerns like losing control, going crazy, or suffocate accompanied by a very strong urge to escape (e.g., from the enclosed place). Finally, specific natural environment–type phobias are characterized by symptom reports of dizziness (in case of fear of heights) and strong avoidance dispositions that often are accompanied by the central concerns about the potential danger of the situation (the tornado might hit our house; the lightening might set our house in fire). These different subgroups of phobias correspond closely to the clusters of situations that are most feared by children, at least when these fears are assessed with open-ended questions. Asking 394 children between 7 and 12 years old, "What do you fear most?" Muris et al. (1997) found that 10.2% of the entire sample reported fear of spiders as their most intense fear. Besides the fear of animals (fear of snakes was also among the top 10), children reported fears from the natural environment cluster (e.g., darkness) as well as blood- and injury-related fears (e.g., illness), and fear of suffocation (e.g., not being able to breathe) belonging to the cluster coded as situational fears. The prevalence of common fears varies across the life span. During early childhood, intense sensory stimulation often evokes a strong fear response in newborns up to the age of 6 months. Actually, the intense sound produced by striking a steel bar with a hammer was used as the unconditioned stimulus to evoke unconditioned fear in the famous case of Little Albert (Watson and Rayner, 1920). Fear of strangers (including costumed characters, particularly if they produce startling loud sounds) and separation from attachment figures are prominent during the first year of life. Preschoolers often fear ghosts, potential burglars, and monsters from movies, particularly during the night when it is dark. Later (between 5 and 7 years), these fears of annihilation shift to more realistic fears of thunderstorms, disasters, death, and war and also the fear of potentially threatening animals (like snakes, spiders, or dogs) (Muris et al., 2003). In the ages between 8 and 11 years old, children often report fear of poor performance in the academic domain but also in the area of athletic performance (McDonald, 2001). In addition, fear of medical procedures is common during this age. Later, during adolescence, fears of being rejected, being negatively evaluated by other adolescents, or acting in an embarrassing way become prominent. The onset of the different types of specific phobias often can be traced back to these vulnerable phases during the ontogenetic development (Öst, 1987a).

Prevalence of Specific Phobias

Prevalence of specific phobias is high and varies across the life span. According to the results of the National Comorbidity Survey, prevalence estimates for specific phobias in adults vary between 11.3% for lifetime and 8.8% for 12-month periods (Kessler et al., 1994). In adolescents, prevalence rates are higher (19.3% according to the NCS-A; Merikangas et al., 2010) but considerably lower if only specific phobias with strong levels of distress and impairment are included. Data from the EDSP study – a longitudinal study – report prevalence rates for specific phobias of 10.9% (12-month period) for adolescents (age range from 14 to 17 years; Beesdo et al., 2009). For children, prevalence rates are lower and vary between 2.6 and 9.1% with an average of 5% across studies (Ollendick et al., 1997). In the elderly, the prevalence of specific phobias decline. Estimates range from 3.1% up to 7.5% in the NCS (Kessler et al., 2005). The larger estimates in the NCS may be due to the inclusion of diagnoses that have remitted (Wolitzky-Taylor et al., 2010). Females are affected more often than males, at rate of approximately 2:1.

Etiology: Understanding Specific Phobias from a Neuroscience Perspective

According to recent neuroscience-based models, emotions like fear can be conceived as action dispositions associated with efforts to escape or avoid danger or pain (Lang and Bradley, 2010) and are rooted in the survival circuits of the brain (see LeDoux, 2012). These defense circuits of the brain feed back to the sensory systems facilitating the detection of threatening and potentially harmful stimuli and prompting defensive reflexive, autonomic, and motor responses to counter threats. Animal and human data suggest that such defensive behavior is organized dynamically in several stages depending on the physical distance (*proximity*) of the threat (i.e., the phobic object), as outlined in the *threat imminence model* by Fanselow (1994) (see also Adolphs, 2013; Blanchard, 1997; Lang et al., 1997). At the first stage, when the organism is in a context in which a threat (e.g., a phobic object) has been encountered previously, but has not yet been detected (preencounter defense), preemptive behavior including threat-nonspecific

vigilance is engaged. Once the threat is identified (post encounter), the organism freezes and increased selective attention is allocated to the potential threat. Circa-strike defense, characterized by strong autonomic arousal and escape behavior, is activated when the threat is imminent. In a recent experiment, we tested this model by measuring electrical brain responses during the presentation of spider pictures to individuals with spider phobia and control participants (Michalowski et al., 2009). The main results of this study together with a sketch of the model can be seen in **Figure 1**.

Informing individuals with spider phobia that a phobic object (picture of a spider) might occur in the current (experimental) context puts the phobic individual (not the control group) in a state of general hypervigilance toward all visual stimuli that occur in such environment, irrespective of their emotional content – phobic individuals responded with larger P1 amplitudes (peaking 100 ms after the onset of the picture) of the visual evoked potentials over the occipital cortex to both neutral and spider stimuli (see right panel (a) of **Figure 1**). Later (300 ms after stimulus onset), phobic individuals show a larger P3 to the spider pictures relative to controls over the parietal cortex, and this group difference did not occur for the other emotional contents (which were significant for both groups), suggesting that selective attention to spider pictures was enhanced for phobics. Such a dynamic view of defensive reactivity integrates cognitive and emotional processes involved in the expression of phobic fear. If the phobic stimulus is more distant, the organism freezes, selective attention is increased together with increased activation of the amygdala (Etkin and Wager, 2007 for a meta-analysis) and is associated with a potentiation of protective reflexes like the startle response (Hamm et al., 1997). During circa strike, when the feared cue is proximal, autonomic arousal increases and selective attention to external sensory signals are reduced to optimize the escape response. In a recent study, we manipulated the proximity of a stress by increasing the size of the visual picture predicting the occurrence of a painful event. The participants had the opportunity to escape from this painful stimulus by pressing a button as soon as possible. When the anticipated threat was more distal, startle responses were potentiated and the brain responses evoked by the startle probes also were augmented, suggesting that participants were in a freezing mode. When finally arriving at the 'circa-strike zone,' autonomic arousal strongly increased and was associated with an inhibition of the startle responses (Löw and Hamm, submitted for publication). Interestingly, the P3 amplitude to the startle probe decreased dramatically, supporting the view that all attention channels are shut down

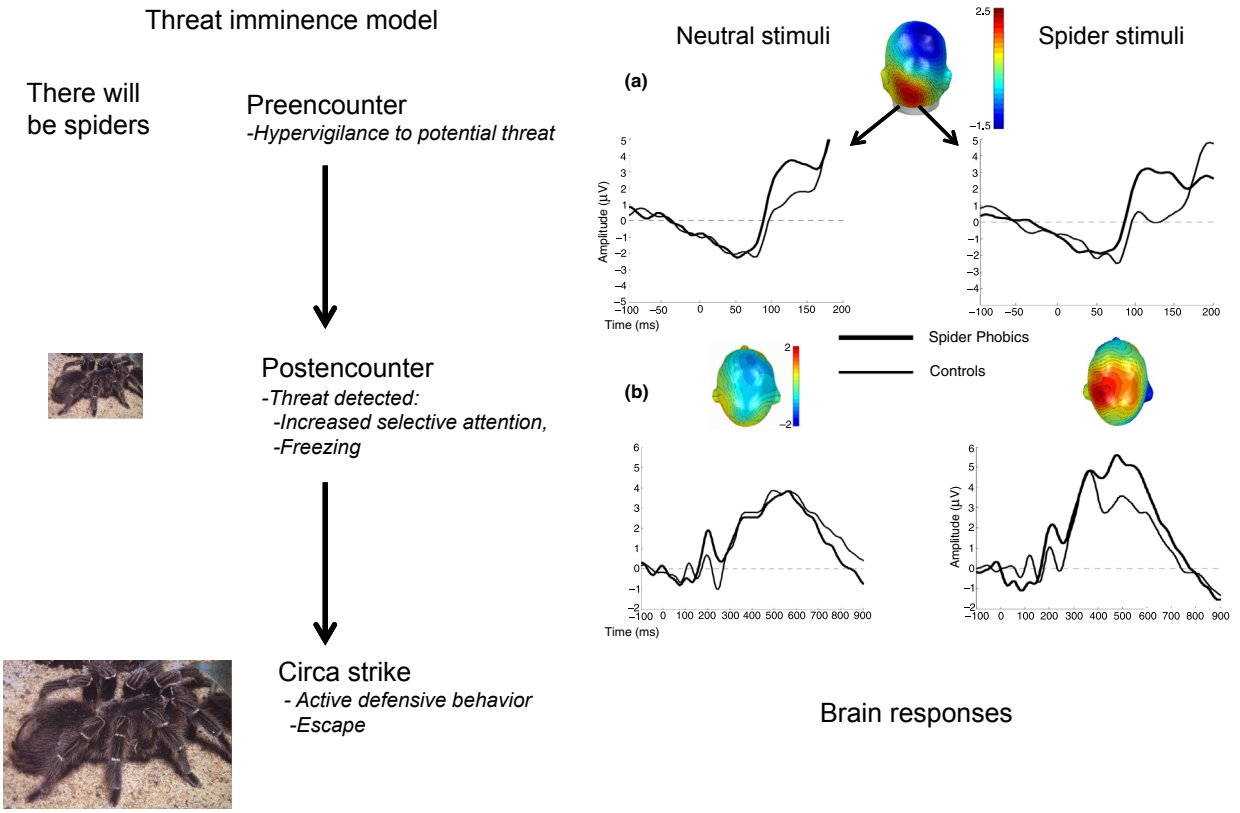

Figure 1 Left panel: The threat imminence model: Defensive behavior in mammals is organized as a cascade of three different stages. Right panel: Brain responses (evoked potentials) to spider and mushroom pictures in spider phobics and healthy controls. (a) Spider-phobic individuals show larger P1 amplitudes over the occipital lobe both to spider pictures (right) and to neutral pictures suggesting a general hypervigilance to all visual cues when phobic individuals are informed that spider pictures might occur. (b) After 300 ms, spider-phobic individuals show larger amplitudes to spider pictures than controls (attention bias) over the parietal cortex, and no group differences were found for neutral stimuli. During postencounter, phobic pictures selectively catch more attention in phobic individuals.

prior to escape. Recent human functional magnetic resonance imaging (fMRI) data also support this interpretation. Using an artificial intelligence predator model to study the dynamics of defensive behavior in humans, Mobbs et al. (2007, 2009) found that postencounter defense activated the amygdala and the subgenual anterior cingulate (ACC), whereas circa strike (when the threat became more imminent) engaged the periaqueductal gray (PAG) and the dorsal ACC, which also is involved in analgesia (Petrovic et al., 2002). From a learning perspective, imminent unconditioned threats (odors from predators, loud noises, pain, lack of oxygen supply) are transmitted from the sensory systems via the amygdala to the hypothalamus, which connects to the dorsal PAG that directs the expression of escape behavior (see LeDoux, 2012). Thus, unconditioned proximal threat activates innate defensive behaviors. More distant threats (e.g., conditioned stimuli that predict the potential occurrence of an innate stress or distant phobic objects) are associated with increased vigilance and freezing behavior. This information also is transmitted via sensory systems to the lateral amygdala and from there via intra-amygdala connections to the central nucleus of the amygdala, which is connected to the ventrolateral PAG (Maren, 2001; Fanselow and Poulus, 2005). With this dynamic model in mind, the vital question remains, how does a specific cue acquire the properties of a phobic object?

According to the model proposed by Rachman (1991), cues can become phobic objects by three different learning experiences: (1) Phobias can be acquired by *direct associative learning experiences* (e.g., if somebody was attacked by a dog). Indeed, numerous fear-conditioning studies show that a previously neutral cue that has been associated with a painful stimulus reliably leads to freezing, potentiation of the startle (Hamm and Weike, 2005, for review), and activation of the amygdala (Etkin and Wager, 2007) after this learning experience. (2) These networks are activated not only during direct conditioning experience but also through *social observation* with no personal experience of the aversive event (e.g., pain). In a study from Olsson et al. (2007), participants observed a movie displaying a male participant receiving painful stimuli in a fear-conditioning experiment. Presenting the fear-conditioned stimuli to the observer (who never received any unconditioned stimulus) also resulted in a clear activation of the amygdala. These data clearly indicate that observational learning relies on the same mechanisms as those involved in learning fear through direct experience. Thus, phobias can be acquired not only by direct aversive experience but also by observing what others have feared before, a phenomenon that has been demonstrated convincingly for the acquisition of snake phobia in primates by Mineka and colleagues (Mineka and Cook, 1993). (3) Fear also can be acquired by *symbolic communication*. Instructing individuals that an aversive event (e.g., a painful stimulus) might follow a certain stimulus (e.g., a blue square) results in robust fear responses to that stimulus (see Grillon et al., 1991). Not surprisingly, the same neural network including the amygdala and the insular cortex is activated during such instructed fear (Phelps et al., 2001) as during direct fear conditioning. These findings clearly demonstrate that specific phobias can be acquired either by (1) direct experience, (2) observing an intense fear response in others, and (3) communicating that some objects or situations might be dangerous. These different learning experiences are mediated by the same neural network including the amygdala and the insula both being core elements of the fear network. Moreover, genetic predisposition seems to modulate how such learning experiences affect the neural network in the brain. Finally, epidemiological evidence from the Dunedin longitudinal study suggests not only that only aversive learning experiences may shape the fear circuit in the brain but also that early exposure to potentially fearful situations might help to immunize the fear networks in the brain. Examining the origin of fear of heights, Poulton and Menzies (2002) found that individuals at age 11 or 18 years with height phobia did not experience serious falls during childhood but rather had less exposure to height stimuli early in life. These findings clearly demonstrate not only that learning experiences are relevant for the development of phobic fears but also that early learning experiences of fear reduction after exposure might help to prevent fear dispositions during childhood from being transferred to phobias that are present during adulthood. Studies asking patients with specific phobias retrospectively about the origins of their phobia using the *Phobic Origins Questionnaire* also find these four pathways. **Table 1** presents the percentage of the answers about the triggers of their phobia for the different types of specific phobias.

Treatment of Specific Phobias

Exposure-Based Treatments

Specific phobias are particularly responsive to exposure-based therapies and therefore are considered the treatment of choice for this disorder (Hood and Antony, 2012). In a meta-analysis, Wolitzky-Taylor et al. (2008) conducted a literature search for treatment studies published between 1966 and 2007 and extracted 33 studies that fulfilled the criteria of using (1) a randomized clinical trial (RCT) with (2) patients meeting full *DSM* criteria and (3) reported phobia-specific

Table 1 Percentage of patients with specific phobias according to the retrospectively recalled origin of their phobia

Origin of phobia	*Animal type*	*Blood/injection injury type*	*Situational type*	*Natural environment type*
Direct conditioning	56%	51%	69%	12%
Vicarious learning	41%	18%	9%	16%
Learning by information	23%	5%	10%	4%
Fear was always there			39%	

The percentages presented in this table are based on results from the following studies: Öst (1987b); Merckelbach et al. (1992); Sartory et al. (1992); Kleinknecht (1994); Menzies and Clarke (1995). The sum of the percentages does not always add up to 100 because multiple answers were allowed in some studies.

outcome measures. Meeting these criteria, there were overall 1193 patients receiving 90 treatments administered. Similarly Choy et al. (2007) did a literature search for treatment studies published between 1960 and 2005 and selected 31 studies that (1) included at least two groups (without the criterion of being an RCT), (2) investigated clinically diagnosed phobia patients (excluding all studies with analogue samples), and (3) reported behavioral or self-report outcome measures. Although using comparable Internet search engines, there was little correspondence (only seven studies were listed in both reviews) between the studies selected for both reviews, thus adding an additional 1527 treated patients to the overall sample. The conclusions drawn from both the meta-analysis and the narrative reviews are clear: Exposure-based therapies result in significantly better outcome at post-treatment irrespective of whether assessed by verbal report measures or by behavioral avoidance tests (BAT) relative to the wait-list control condition (Choy et al., 2007; Wolitzky-Taylor et al., 2008). The overall effect size for this comparison is $d = 1.05$ and is comparable for behavioral measures and questionnaire data. In the group of patients with a specific animal-type phobia, 92% of the patients treated with *in vivo* exposure compared with 0% of the wait-list control were able to perform the terminal task in the BAT (Bandura et al., 1969; Williams et al., 1984). Comparable percentages were reported for patients with either height or driving phobia (87% of the treated patients completed the terminal task of the BAT) and claustrophobia (79% completers; Öst et al., 2001). Choy et al. (2007) found 16 studies in their review, which investigated outcome measures in a follow-up period, ranging from 6 to 14 months and found stable treatment effects for the treatment of animal phobia, height phobia, flight phobia, and claustrophobia. Some studies even found an increase in the treatment effects, whereas other studies did not find any further improvement compared with postmeasures.

Exposure Techniques

Exposure to the feared situation can be performed *in sensu* – that is, the patient is instructed to imagine to be in the feared situation, and often, the intensity of the fearful environment is increased step by step using a technique named 'systematic desensitization' and developed by Wolpe (1958). Wolpe (1958) had suggested that *conditioned inhibition* would be the crucial mechanism for the curative effect of systematic desensitization – for example, if a relaxation response that is antagonistic to the fear response is elicited during the feared cue, it consequently will reduce the fear response so that the fear cue itself will become a conditioned inhibitor (*counter-conditioning*). Numerous studies in the 1960s, however, clearly showed that the induction of relaxation is not a necessary condition for effective exposure treatment (for a critical review, see Tryon, 2005). As a consequence of these developments, systematic desensitization has fallen out of favor and behavior therapy of specific phobias has shifted to exposure *in vivo*, in which patients are confronted with their feared situations or objects in reality (e.g., patients are encouraged to approach and touch the feared animal until the fear subsides, Hellström and Öst, 1995), or patients climb the highest building in the area together with the therapist and practice standing at the top without touching the railings (Williams et al., 1985), or they are entrapped in a narrow dark chamber until the claustrophobic fear subsides (Öst et al., 1982). Such exposure *in vivo* very effectively has been done in one session as described by Öst (1985, 1987b, 1989) for the treatment of animal-type and blood- or injury-type phobias and by Emmelkamp and Felten (1985) for fear of heights. One-session treatment includes the active participation of the therapist in activities, such as explaining the treatment rational, modeling, and collaborating with the patient during the single-session active exposure. The entire therapy session is videotaped, and in some cases, patients are invited to return to the clinic to watch parts of the video 1 or 2 weeks after the exposure session. The active treatment session is preceded by a careful behavioral assessment, including the etiology of the phobia (e.g., conditioning experience), avoidance and escape response patterns, and a detailed analysis of the central concerns and the feared consequences of the exposure (e.g., I might not get enough oxygen when trapped in this chamber). The exposure session is divided into several behavioral exercises that are arranged in hierarchical steps. The patient makes a commitment to remain in the exposure situation until the fear fades away and not to escape from the situation during the treatment. Before moving to the next step, the therapist describes that task, demonstrates how the patient is expected to interact with the feared object, and then has the patient carry out the interaction with the object or situation by himself or herself. Before but also during each behavioral exercise, the therapist prompts the patient for his central concerns and feared consequences to explicitly challenge them during the exposure. The patient is instructed to interact with the feared object or situation at each behavioral step until the fear dissipates. The entire session ends when the final approach goal is met, and the level of fear has been reduced to 50% of its highest value or, ideally, has vanished. On the basis of a thorough review of 25 controlled treatment studies of specific phobias (with 800 patients and 150 control subjects), including many conducted at his own institution, Öst (1997) concluded that one-session *in vivo* exposure is the treatment of choice for specific phobias. Across different types of phobias, such individual single-session treatment results in a clinically significant improvement (e.g., an improvement of two standard deviations relative to the mean symptom score of the patient population) in 77–95% of patients. Also noteworthy is that the average duration of such treatment varies between 1.9 and 3.5 h. One-session treatment of specific phobia has been evaluated empirically for spider phobia, blood and injection phobia, and claustrophobia (see review by Zlomke and Davis, 2008). Significant effect sizes were reported for self-report measures of fear (as assessed by several questionnaires, mean $d = 1.98$; range 0.91–3.05) and reported distress during standardized behavioral avoidance tests (mean $d = 2.20$; range 0.79–3.80). Comparable effect sizes ($d = 1.41$) were reported by Öst et al. (1997) for a 3 h one-session treatment of 14 patients with flight phobia during which the therapist accompanied each patient during the 30 min bus ride to Arlanda airport, during the process of checking and boarding, and during the flight from Stockholm to a domestic destination (overall, 90 min), as well as during the entire procedure on the

return flight to Stockholm. Fear of flying was reduced significantly after this single-session treatment, and 93% of the patients (13 out of 14) took an unaccompanied return flight on their own after treatment. More recently, Haukebo et al. (2008) reported significant improvements in a group of 10 patients with dental phobia after a one-session exposure treatment by a dentist who was trained specifically in this method by Öst.

Exposure in Virtual Reality

With increasing availability of virtual reality (VR) technology, it is now possible to use computer-generated interactive VRs, including visual displays of moving objects, other sensory inputs, and body-tracking devices to conduct exposure therapy. Patients visit these VRs for a prolonged period of times during several sessions. Virtual reality exposure therapy (VRET) is particularly helpful when it is difficult to recreate an environment for exposure in reality. Therefore, in the case of treating specific phobias, VRET has been empirically evaluated primarily for treating flight and height phobia. In their meta-analysis of VRET for anxiety disorders, Powers and Emmelkamp (2008) found 11 studies that tested clinical effectiveness of VRET for treating specific phobias (eight studies investigated treatment effects in flight phobia and four studies investigated treatment effects in height phobia using VRET). Studies that compared VRET either with a wait-list control group or with groups receiving nonspecific interventions, such as relaxation or attention control training in an RCT, found VRET to be more effective than nonspecific interventions. Overall effect sizes were in the range of medium to large across all measures, including verbal report of general distress, cognitive measures, avoidance behavior, and autonomic responding (see Powers and Emmelkamp, 2008; Meyerbröker and Emmelkamp, 2010 for reviews; see also the article on virtual reality in this volume). In the case of height phobia, VRET is as effective as *in vivo* treatment (Emmelkamp et al., 2002). In case of flying phobia, however, results are more mixed. Two studies compared VR with *in vivo* exposure for flight phobia and found VR to be equally effective irrespective of whether verbal report measures or behavioral tests (taking a graduation flight) were taken as outcome measures (Emmelkamp et al., 2002; Rothbaum et al., 2002), particularly at follow-up assessment (see Wolitzky-Taylor et al., 2008). Data of Mühlberger et al. (2006), however, suggest that including *in vivo* exposure (e.g., taking a graduation flight) increases the long-term effects of exposure therapy in VR. Completion of the graduation flight was a strong predictor for long-term treatment efficacy (Mühlberger et al., 2006). Thus, it is still an open question whether VRET of flight phobia is as successful as the 2-day program presented by Van Gerwen and colleagues, showing that 98.6% of 685 treated patients did fly within the next 12 months (Van Gerwen et al., 2002). Santos and coworkers recently reported comparable success rates (Santos et al., 2011).

Summary

Marked and persistent fear cued by specific objects or situations accompanied by a compelling desire to escape from or to avoid such objects or situations – if possible – are the main diagnostic criteria for specific phobia. Although exaggerated fear responses can be elicited in principal by almost each imaginable object (e.g., trypophobia, fear of holes), current diagnostic systems (*DSM-5*) code discriminate specific phobias based on the phobic stimulus into five different specifiers. Specific phobias are quite common and vary across the life span with about 20% prevalence rates in adolescents, 11% in adults, and about 5% in the elderly. Much progress has been made in the recent decades in understanding the neural networks involved during encoding of the phobic stimulus and orchestrating the fear response. The lateral amygdala seems to be a core structure involved in the detection of the emotional significance of and transferring this information to the ACC and the anterior insula. The central nucleus seems to be involved in the organization of the defensive behavior that follows a dynamic pattern from freezing to escape depending on the proximity of the phobic object. Neutral situations or objects can acquire phobic quality and, thus, get access to these networks via direct aversive learning experiences, by observation of others making such experiences, or by the information that some situation might be dangerous. The most effective treatments of specific phobias are exposure-based therapies that can be administered effectively even in a single session.

See also: Behavior Therapy: Background, Basic Principles, and Early History; Evidence-Based Psychotherapy; Manual Based Treatments; Panic Disorder and Agoraphobia Across the Lifespan; Social Phobia across the Lifespan; Virtual Reality in Psychotherapy.

Acknowledgments

The research reported in this article is supported by grants from the German Research Society (Ha 1593/15-1; Ha 1593/18-1) and by the German Federal Ministry of Education and Research (project number 01GV0615).

Bibliography

Adolphs, R., 2013. The biology of fear. Current Biology 23, R79–R93.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Association, Washington, DC.

Bandura, A., Blahard, E.B., Ritter, B., 1969. Relative efficacy of desensitization and modeling approaches for inducing behavioral, affective, and attitudinal changes. Journal of Personality and Social Psychology 13, 177–199.

Beesdo, K., Knappe, S., Pine, D.S., 2009. Anxiety and anxiety disorders in children and adolescents: developmental issues and implications for DSM-V. Psychiatric Clinics of North America 32, 483–524.

Bittner, A., Goodwin, R.D., Wittchen, H.-U., Beesdo, K., Höfler, M., Lieb, R., 2004. What characteristics of primary anxiety disorders predict subsequent major depressive disorder? Journal of Clinical Psychiatry 65, 618–626.

Blanchard, D.C., 1997. Stimulus, environmental, and pharmacological control of defensive behaviors. In: Bouton, M.E., Fanselow, M.S. (Eds.), Learning, Motivation, and Cognition: The Functional Behaviorism of Robert C. Bolles. APA, Washington, DC, pp. 283–303.

Brown, T.A., Campbell, L.A., Lehman, C.L., Grisham, J.R., Mancill, R.B., 2001. Current and lifetime comorbidity of the DSM-IV anxiety and mood disorders in a large clinical sample. Journal of Abnormal Psychology 110, 585–599.

Choy, Y., Fyer, A.J., Lipsitz, J.D., 2007. Treatment of specific phobia in adults. Clinical Psychology Review 27, 266–286.

Davis, T.E., Ollendick, T.H., Öst, L.-G. (Eds.), 2012. Intensive One-Session Treatment of Specific Phobias. Springer, New York.

Emmelkamp, P.M.G., Felten, M., 1985. The process of exposure in vivo: cognitive and physiological changes during treatment of acrophobia. Behaviour Research and Therapy 23, 219–223.

Emmelkamp, P.M.G., Krijn, M., Hulbosch, A.M., De Vries, S., Schuemie, M.J., Van der Mast, C.A.P.G., 2002. Virtual reality treatment versus exposure in vivo: a comparative evaluation of acrophobia. Behaviour Research and Therapy 40, 509–516.

Emmelkamp, P.M.G., Wittchen, H.-U., 2009. Specific phobias. In: Andrews, G., Charney, D.S., Sirovatka, P.J., Regier, D.A. (Eds.), Stress-Induced and Fear Circuit Disorders. American Psychiatric Publishing, Inc., Arlington, VA, pp. 85–110.

Etkin, A., Wager, T.D., 2007. Functional neuroimaging of anxiety: a meta-analysis of emotional processing in PTSD, social anxiety disorder, and specific phobia. American Journal of Psychiatry 164, 1476–1488.

Fanselow, M.S., 1994. Neural organization of the defensive behavior system responsible for fear. Psychonomic Bulletin Reviews 1, 429–438.

Fanselow, M.S., Poulus, A.M., 2005. The neuroscience of mammalian associative learning. Annual Review of Psychology 56, 207–234.

Friedman, B.H., Thayer, J.F., Borkovec, T.D., Tyrrell, R.A., Johnson, B.H., Columbo, R., 1993. Autonomic characteristics of nonclinical panic and blood phobia. Biological Psychiatry 34, 298–310.

Globisch, J., Hamm, A.O., Esteves, F., Öhman, A., 1999. Fear appears fast: temporal course of startle reflex potentiation in animal fearful subjects. Psychophysiology 36, 66–75.

Grillon, C., Ameli, R., Woods, S.W., Merikangas, K.R., Davis, M., 1991. Fear potentiated startle in humans: effects of anticipatory anxiety on the acoustic blink reflex. Psychophysiology 28, 511–517.

Hamm, A.O., Cuthbert, B.N., Globisch, J., Vaitl, D., 1997. Fear and the startle reflex: blink modulation and autonomic response patterns in animal and mutilation fearful subjects. Psychophysiology 34, 97–107.

Hamm, A.O., Weike, A.I., 2005. The neuropsychology of fear learning and fear regulation. International Journal of Psychophysiology 57, 5–14.

Haukebo, K., Skaret, E., Öst, L.-G., Raadal, M., Berg, E., Sundberg, H., Kvale, G., 2008. One- vs. five-session treatment of dental phobia: a randomized controlled study. Journal of Behavior Therapy and Experimental Psychiatry 39, 381–390.

Hellström, K., Öst, L.-G., 1995. One-session therapist directed exposure vs. two forms of manual directed self-exposure in the treatment of spider phobia. Behaviour Research and Therapy 33, 959–965.

Hood, H.K., Antony, M.M., 2012. Evidence-based assessment and treatment of specific phobias in adults. In: Davis, T.E., Ollendick, T.H., Öst, L.-G. (Eds.), Intensive One-Session Treatment of Specific Phobias. Springer, New York, pp. 19–42.

Kessler, R.C., McGonagle, A., Zhao, S., Nelson, C.B., Hughes, M., Eshleman, S., Wittchen, H.-U., Kendler, K.S., 1994. Lifetime and 12-month prevalence of DSM-II-R psychiatric disorders. Archives of General Psychiatry 51, 8–19.

Kessler, R.C., Chiu, W.T., Demler, O., Walters, E.E., 2005. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry 62, 617–709.

Lang, P.J., Bradley, M.M., Cuthbert, B.N., 1997. Motivated attention: affect, activation, and action. In: Lang, P.J., Simona, R.F., Balaban, M.T. (Eds.), Attention and Orienting: Sensory and Motivational Processes. Erlbaum, Hillsdale, NJ, pp. 97–135.

Lang, P.J., Bradley, M.M., 2010. Emotion and the motivational brain. Biological Psychology 84, 437–450.

LeBeau, R.T., Glenn, D., Liao, B., Wittchen, H.-U., Beesdo-Baum, K., Ollendick, T., Craske, M.G., 2010. Specific phobia: a review of DSM-IV specific phobia and preliminary recommendations for DSM-V. Depression and Anxiety 27, 148–167.

LeDoux, J.E., 2012. Rethinking the emotional brain. Neuron 73, 653–676.

Lipsitz, J.D., Barlow, D.H., Mannuza, S., Hofmann, S.G., Fyer, A.J., 2002. Clinical features of four DSM-IV-specific phobia subtypes. Journal of Nervous and Mental Disease 190, 471–478.

Löw, A., Hamm, A.O., 2013. When threat is near get of here: switching from freezing to escape, submitted for publication.

Maren, S., 2001. Neurobiology of Pavlovian fear conditioning. Annual Review of Neuroscience 24, 897–931.

McDonald, A.S., 2001. The prevalence and effects of test anxiety in school children. Educational Psychology 21, 89–101.

Merikangas, K.R., He, J., Burstein, M., Swanson, S.A., Avenevoli, C., Cui, L., Benjet, C., Georgiades, K., Swendsen, J., 2010. Lifetime prevalence of mental disorders in US adolescents: results from the National Comorbidity Study – Adolescent Supplement (NCS-A). Journal of the American Academy of Child and Adolescent Psychiatry 49, 980–989.

Meyerbröker, K., Emmelkamp, P.M.G., 2010. Virtual reality exposure therapy in anxiety disorders: a systematic review of process-and-outcome studies. Depression and Anxiety 27, 933–944.

Michalowski, J.M., Melzig, C.A., Weike, A.I., Stockburger, J., Schupp, H.T., Hamm, A.O., 2009. Brain dynamics in spider-phobic individuals exposed to phobia-relevant and other emotional stimuli. Emotion 9, 306–315.

Mineka, S., Cook, M., 1993. Mechanisms underlying observational conditioning of snake fear in monkeys. Journal of Experimental Psychology: General 122, 23–38.

Mobbs, D., Marchant, J.L., Hassabis, D., Seymour, B., Tan, G., Gray, M., Petrovic, P., Dolan, R.J., Frith, C.D., 2009. From threat to fear: the neural organization of defensive fear systems in humans. Journal of Neuroscience 29, 12236–12243.

Mobbs, D., Petrovic, P., Marchant, J.L., Hassabis, D., Weiskopf, N., Seymour, B., Dolan, R.J., Frith, C.D., 2007. When fear is near: threat imminence elicits prefrontal periaqueductal grey shifts in humans. Science 317, 1079–1083.

Mühlberger, A., Weik, A., Pauli, P., Wiedemann, G., 2006. One session virtual reality exposure treatment for fear of flying: one-year follow-up and graduation flight accompaniment effects. Psychotherapy Research 16, 26–40.

Muris, P., Merckelbach, H., Meesters, C., van Lier, P., 1997. What do children fear most often? Journal of Behavior Therapy and Experimental Psychiatry 28, 263–267.

Muris, P., Bodden, D., Merckelbach, H., Ollendick, T.H., King, N., 2003. Fear of the beast: a prospective study on the effects of negative information on childhood fear. Behaviour Research and Therapy 41, 195–208.

Öst, L.-G., 1985. Single-session exposure treatment of injection phobia: a case study with continuous heart rate measurement. Scandinavian Journal of Behavior Therapy 14, 125–131.

Öst, L.-G., 1987a. One-session treatments for a case of multiple simple phobias. Scandinavian Journal of Behavior Therapy 16, 175–184.

Öst, L.-G., 1987b. Age of onset of different phobias. Journal of Abnormal Psychology 96, 223–229.

Öst, L.-G., 1989. One session treatment for specific phobias. Behaviour Research and Therapy 27, 1–7.

Öst, L.-G., 1997. Rapid treatment of specific phobias. In: Davey, G.C.L. (Ed.), Phobias. Wiley, Chichester, UK, pp. 227–246.

Öst, L.-G., Johansson, J., Jerremalm, A., 1982. Individual response patterns and the effects of different behavioral methods in the treatment of claustrophobia. Behaviour Research and Therapy 20, 445–460.

Öst, L.-G., Brandberg, M., Alm, T., 1997. One versus five sessions of exposure in the treatment of flying phobia. Behaviour Research and Therapy 35, 987–996.

Öst, L.-G., Alm, T., Brandberg, M., Breitholtz, E., 2001. One vs. five sessions of exposure and five sessions of cognitive therapy in the treatment of claustrophobia. Behaviour Research and Therapy 39, 167–183.

Ollendick, T.H., Hagopian, L.P., King, N.J., 1997. Specific phobias in children. In: Davey, G.C.L. (Ed.), Phobias: A Handbook of Theory, Research, and Treatment. Wiley, New York, pp. 201–224.

Olsson, A., Nearing, K.I., Phelps, E.A., 2007. Learning fears by observing others: the neural systems of social fear transmission. Social, Cognitive, and Affective Neuroscience 2, 3–12.

Petrovic, P., Kalso, E., Petersson, K.M., Ingvar, M., 2002. Placebo and opioid analgesia – imaging a shared neural network. Science 295, 1737–1740.

Phelps, E.A., O'Connor, K.J., Gatenby, J.C., Grillon, C., Gore, J.C., Davis, M., 2001. Activation of the left amygdala to cognitive representation of fear. Nature Neuroscience 4, 437–441.

Poulton, R., Menzies, R.G., 2002. Non-associative fear acquisition: a review of the evidence from retrospective and longitudinal research. Behaviour Research and Therapy 36, 17–35.

Powers, M.B., Emmelkamp, P.M.G., 2008. Virtual reality exposure therapy for anxiety disorders: a meta-analysis. Journal of Anxiety Disorders 22, 561–569.

Rachman, S.J., 1991. Neoconditioning and the classical theory of fear acquisition. Clinical Psychology Review 11, 155–173.

Rothbaum, B.O., Hodges, L., Anderson, P.L., Price, L., Smith, S., 2002. Twelve-month follow-up of virtual reality and standard exposure therapies for the fear of flying. Journal of Consulting and Clinical Psychology 70, 428–432.

Santos, E., Palma, H., Napoles, G., Romero, R., Gabriel, R., Grilo, V., Pascoal, C., 2011. Earn your wings program: description and efficacy of an intensive fear of flying treatment program. Aviation Psychology and Applied Human Factors 1, 61–69.

Tryon, W.W., 2005. Possible mechanisms for why desensitization and exposure therapy work. Clinical Psychology Review 25, 67–95.

Van Gerwen, L.J., Spinhoven, P., Diekstra, R.F.W., Van Dyck, R., 2002. Multicomponent standardized treatment program for fear of flying. Description and effectiveness. Cognitive Behavioral Practice 9, 138–149.

Watson, J.B., Rayner, R., 1920. Conditioned emotional reactions. Journal of Experimental Psychology 3, 1–14.

Williams, S.L., Dooseman, G., Kleinfield, E., 1984. Comparative effectiveness of guided mastery and exposure treatments for intractable phobias. Journal of Consulting and Clinical Psychology 52, 505–518.

Williams, S.L., Turner, S.M., Peer, D.F., 1985. Guided mastery and performance desensitization treatments for severe acrophobia. Journal of Consulting and Clinical Psychology 53, 237–247.

Wittchen, H.-U., Lecrubier, Y., Beesdo, K., Nocon, A., 2003. Relationships among anxiety disorders: patterns and implications. In: Nutt, D.J., Ballenger, J.C. (Eds.), Anxiety Disorders. Wiley-Blackwell, Malden, MA, pp. 25–38.

Wolpe, J., 1958. Psychotherapy by Reciprocal Inhibition. Stanford University Press, Stanford, CA.

Wolitzky-Taylor, K.B., Horowitz, J.D., Powers, M.B., Telch, M.J., 2008. Psychological approaches in the treatment of specific phobias: a meta-analysis. Clinical Psychology Review 28, 1021–1037.

Wolitzky-Taylor, K.B., Castriotta, N., Lenze, E.J., Stanley, M.A., Craske, M.G., 2010. Anxiety disorders in older adults: a comprehensive review. Depression and Anxiety 27, 190–211.

Zlomke, K., Davis III, T.E., 2008. One-session treatment of specific phobias: a detailed description and review of treatment efficacy. Behavior Therapy 39, 207–223.

Phological Disorders

Hugh W Buckingham and Sarah S Buckingham, University of Oklahoma Health Sciences Center (OUHSC), Oklahoma City, OK, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article outlines 200 years of neurolinguistic investigations into the nature and causes of phonological disorders in human language. Early empiricist principles such as association frequency, force/strength, and activation/inhibition, forged theoretical relationships among sound types (phonemes), words, and word meanings that drove the understanding of normal and disordered phonology from antiquity through the mid-twentieth century. Throughout this period, researchers in natural philosophy as it evolved into psychology understood sound errors as substitutions of elements sequencing errors within the strictly concatenated links between sound neighbors in the familiar associative sound chains that comprised the phonological forms of words in the British empiricist such as John Locke. Near the turn of the nineteenth century, however, neurologists such as John Hughlings-Jackson and Sigmund Freud challenged associationism, arguing for a more abstract, sensory conceptualization of sounds. Errors, including slips-of-the-tongue, were noted to occur over wider expanses of sound than those permitted by the associationist models. This was a watershed time period in the history of phonology, and it was during the late-nineteenth century that the concept of the phoneme was crafted from within phonetic science. To this was added a deeper phonology from the Prague School, followed upon by the work of Roman Jakobson. By the mid-twentieth century, the cognitive sciences were giving rise to a variety of influential mental models for language processing, including nonassociationist serial modeling as well as later hierarchical, top-down serial models such as the one crafted by Merrill Garrett and later by the nonserial, computer interactive activation model of Gary Dell and colleagues such as Nadine Martin. Today's computer simulations of normal and disordered phonological systems, together with advances in '*in vivo*' neuroimaging technologies, allow a more profound understanding of the real-time contributions of the 'neuro' and the 'linguistic' in understanding phonology. As we can now predict errors in simulated systems and predict patterns of recovery in individuals with aphasia, it is quite clear that as science advances, so does the characterization of phonological disorders.

Introduction

In Buckingham and Christman (2008), we outlined disorders of both phonetics and phonology, almost as if the two levels could be easily dissociated; they cannot. In the present article, we will be sensitive to some sort of interface between the two for the sake of much linguistic and psycholinguistic modeling that keeps them apart.

We will also assume that 'phonological disorders' go as phonology is characterized at the time at the point in history of their description. We will claim that throughout the history of considering 'phonological disorders,' we must consider what phonology could have been taken to be. As a consequence, we will frame our article in terms of the continuity and change of 'phonology,' and how 'the models make the man,' from Jacques Lordat to the connectionism of Dell's interactive activation (IA) to Goldsmith's harmony seeking model, which works off the Hopfield Nets of Smolensky's physics, although we will have relatively little to say about Hebb, Hopfield, or Smolensky. We are essentially adopting the position outlined by Whitaker (1972, p. 431), where he writes that

> Phonological structure is often studied independently from the other components of language and, until recently, it has been studied in greater detail than these other components for obvious reasons: its accessibility to instrumental measurement and the presumed limitation or finiteness of its scope and content.

In this article, Whitaker demonstrated that importantly as goes the phonological brain, so goes phonology. Our article has a fair amount of material, perhaps worth a new text on *all of this from an historical framework, but we will try to keep the train of ideas as distinct* but intellectually and historically integrated to the extent we can. We hope to go beyond Buckingham and Christman (2008), but our new work must integrate some of the issues touch upon there. Moreover, we will consider some fMRI and DTI studies where appropriate for distinguishing 'the where' of distinct types of phonological disorders that may differ along anatomical/physiological dimensions.

Some Phonological Disorders from Jacques Lordat to Hugo Liepmann

Lordat

Jacques Lordat, among his many other efforts, worked with patients with 'apoplexy' in the latter half of the eighteenth century and the early nineteenth at the Montpellier School of Medicine. By 1843, his collection of cases and analyses resulted in the publication of his 'L'aphasie et l'elaboration de la pensee explicite,' where Lordat constructed essentially a 10-stage production model, which Lecours et al. (1987) and more recently Nespoulous (2013) have characterized as one of the earliest, if not *the* earliest, psycholinguistic model of language production scaffolded off data from aphasia (**Figure 1** French (2013) with English translation (1987)).

We pay special attention to Level 9 where previously concatenated successive sounds are fixed to one another 'securing mechanical harmony' ('solidaire' Fr.) as they pass articulatorily from one to the next. Levels 3 and 4 seem to select words

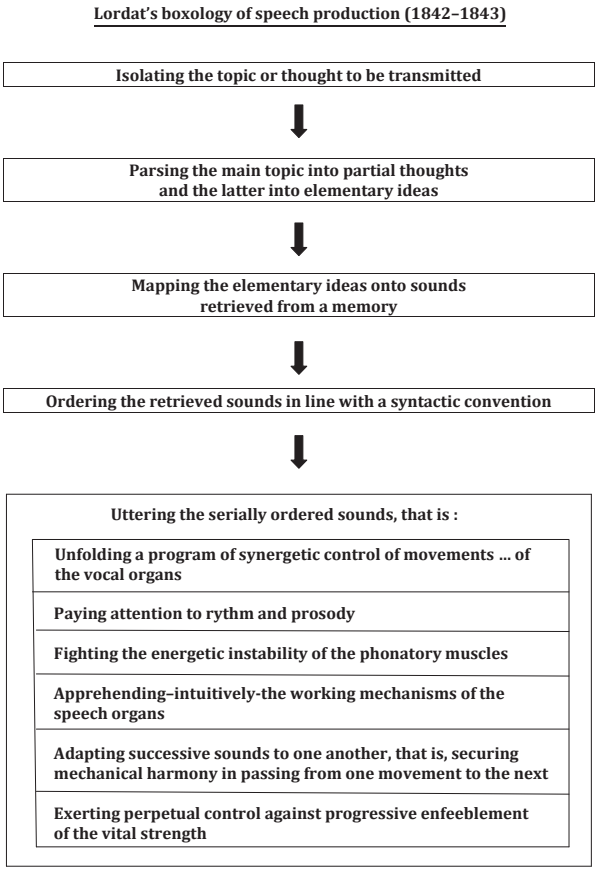

Figure 1 Jacques Lordat's boxology of speech production (1842–43) redrawn by Jean-Luc Nespoulous. Permission granted. This is a series of steps of language/speech production.

containing prearticulatory 'sounds' and ordering those prearticulatory sounds within the word. Apparently, Lordat did not realize that this was a first ordering to be followed by a second ordering at level 9, as is now recognized in most hierarchical and serial production models.

What we wish to point out here regarding Lordat's model is that the articulatory dynamics of the processes at level 9, intriguingly foreshadow the much later work by Wayne Wickelgren (1969) where the 'sounds' and their left and right contexts of Lordat are 'wickelphones,' and thus they mechanically attach to each other, the left and right contextual (coarticulatory) elements are was an automatic and associative account of the stringing together of what we might call today 'articulemes' or 'gesturemes.'

In summary, Lecours et al. (1987, p. 7) write that

> Lordat (1843) teaches his students that 'from one's decision to communicate one's thought, to the embodiment ("corporification," (Nespoulous, 2013: p. 4 – HWB)) of this thought into conventional sounds, one has to accomplish a succession of various acts, (some of them exclusively mental), each of which should be studied separately.

Naturally, by the time level 9 is being computed, we have traversed the level of ideas to the body. Again, Lordat is prescient in his ability to psycholinguistically intuit and of course here we seen his realization the still enigmatic concept of embodiment. And, although Nespoulous (2013, p. 7) claims that higher level of ordering of prearticulatory segments and *errors* allow us to see without the slightest doubt, "… l'ancetre de la 'paraphasie phonemique," he follows with a few examples of phonemic substitutions which he then claims also look like the slips-of-the-tongue of Fromkin (1973) and other sorts of fluent aphasic phonological distortions that would arise at levels 4/5. Recall, however, that later on we will compare the computations at Lordat's level 9 with Wickelgren's notion of the 'context-sensitive elementary motor response' units – the Wickelphones.

Bouillaud, Auburtin, and Broca

The period leading up from Bouillaud to Broca is a well-charted progression (see Schiller, 1994; LaPointe, 2013 and numerous other works). What mostly comes out of the celebrated Anthropological Society meets during this period are descriptions of patients who had nondysarthric articulatory production disruptions, which, although they actually spoke very little, when they did speak most investigators could perceive that those few articulations were not so easily tagged as 'arthric,' which most of the clinical investigators observed had been secondary to lesions subjacent to the cerebral cortex. Since at that time, and with a glaring neglect of work by earlier neurologists, such as Lordat, there was no description of the actual linguistic levels and elements involved in the speech of the adults with apoplexy and most often right hemiplegia. The terminology was largely restricted to 'sensory' versus 'motor,' or 'arthric' versus that other articulatory feature that was perceived to be more 'fluent' and without what were the signposts of the arthric disorder. Aside from vague notions of 'language,' there were no clear parameters with which to seriously describe what that would entail as far as anything linguistic. Theories had abounded since Bouillaud's (1825) claim that the nonarthric speech disorder was a sort of procedural memory of articulatory construction. Assigning that disorder to the realm of memory allowed the investigators to distinguish it from arthric speech, to assume that it was mediated at the cortex, that it was a product of volition. Assuming that the disorder was a faculty of the cortex, motor explanations were *ipso facto* ruled out, since much before the landmark paper by Fritsch & Hitzig (1870/1960) the cerebral cortex was not considered to control motor movement. Broca (1861) finally pinned that cortical zone to be the inferior portion of the third frontal lobe, Marc Dax notwithstanding, that localization was not necessarily placed within the left hemisphere; that came later (Broca, 1865). This point of all this is that so far, there was no available linguistic model to handle much of any of this, much less any other segmental and lexical disruption at an even deeper psychical/mental/abstract level. So, naturally, there were no true and unambiguous phonological disorders. Rather, up to this point it was primarily to figure out the ontological status of this nonarthric articulation disorder. Indeed, since it was specified as a cortical disorder, it was considered simply to be a 'faculty of articulation,' faculties essentially being the work of the cerebral cortex, under the model of the phrenologists. And, since the phrenologists still held to the

'law of symmetry,' they, too, focused only on the frontal lobe for articulation, but viewed the faculty as a bilateral function.

Karl Wernicke

With Wernicke (1874), what slowly emerges is an anatomical account of a more fluent speech output of patients previously most often assigned to the realm of insanity. They had no signs whatsoever of a patient post apoplexy without a unilateral upper limb paralytic dysfunction and a smooth articulatory output, with neither of the previous noted speech problems. They also ambulated in a normal fashion; few seemed to know they had a stroke; Wernicke, however, did. The patients were quickly rerouted from asylums to stroke units. Wernicke carefully chartered their speech outputs and claimed they are those of normal speakers for the most part. They had semantic word substitutions, and many kinds of sound errors, most of which were substitutions and linear transpositions, some so numerous as to render some supposed target word unrecognizable. These quickly were tagged as 'sensory' aphasic errors, and were not only located far from Broca's area in temporal lobe regions, but were also clearly distinguishable clinically from those nonarthric disorders of Broca; there are now three sorts of segmental output disruptions. From a clinical standpoint (not linguistic yet), it was slowly becoming clear that more 'language-like' errors were being produced by these adults with 'sensory' aphasia. Without the necessary linguistics, Wernicke and other physicians such as Bastian (1869) and Bateman (1890), and even earlier with the work of J. Gesner from 1767 to 1788 (see, Benton (2000) for an analysis of Gesner's work on jargon aphasia) began describing the anomias and other sensory/fluent segmental and lexical errors, which in time would unwittingly open up a clearer idea of just what a phonology might be. However, and very unfortunately, Lordat's model could be of no use; no one knew of Lordat, it seemed. Modern day linguists and naïve historians may very well over interpret what Wernicke was saying in 1874 and may talk of phonological disorders in sensory aphasia. This was not at all the case yet with Wernicke's original monograph, although there can be no doubt that anatomically and clinically his work contributed significantly in the slow move to 'phonology.' Paul Eling (2006) laments the fact that most of the Wernicke-like clinical aphasiological physicians in the nineteenth century neglected the work of psychologists such as Chajim Steinthal, and to this extent they have perhaps been overinterpreted as carving the way to a phonology, which would then admit of true 'phonological disorders.' Essentially, two more physicians were needed – Hughlings-Jackson (1979–80) and Freud (1891), perhaps with the addition of Hugo Liepmann (1900/1977) around the turn of the century to help keep 'apraxic' movement disorders dissociated as much as possible from language.

John Hughlings-Jackson and Sigmund Freud

One of the most significant observations in the history of aphasia was John Hughlings-Jackson's (1879–80, p. 194) claim that the fluent errors of his patients were not unlike 'the blunders' that also appear "… in persons whose speech is only very slightly defective: I mean in those who, for the most part, speak well." This passed on to Freud (1891, p. 13) who wrote that, "… the paraphasia observed in aphasic patients does not differ from the incorrect use and the distortion of words which the healthy person can observe in himself in states of fatigue or divided attention or under the influence of disturbing affects…." Thus, 'continuity' was established between persons with pathologies and those without. The 'slip-of-the-tongue' since those years has become an earmark of the description of fluent segmental and lexical 'lapsus linguae.' Accordingly, this 'continuity' between the health and the diseased has played a role in the very notion of phonological disorders, and plays a very important role in the connectionism of Dell et al. (1997a,b). We will have more to say on this later in our article. But, for now, suffice it say that the slip, too, helped pave the way for characterizations of what could be a phonological disorder. See Buckingham (1999, 2006) for extended treatments of slips-of-the-tongue and the 'continuity' thesis.

At this point, we must mention two additional players in the nineteenth-century harbingers of phonological disorders. First, all readers of Freud's monograph on aphasia (1891) will note the influence that the linguist, Berthold Delbruck, had on Freud's growing appreciation of the role of the linguistics of the neogrammarians in the study of aphasia. One piece of irony in all this is that Freud credited Delbruck for his linguistic analyses that well pertained to aphasia, while Delbruck himself admits that many of his early ideas of language production came from Jacques Lordat (Bay, 1969).

Aside from Delbruck, another important influence on Freud of the continuity between the pathological and the nonpathological came from Jean-Marie Charcot's studies of the hysterias. Freud took careful note when visiting Charcot's clinic at the Salpetriere in Paris between September 1885 and February of 1886, of Charcot's 'hysterical' patients who had *no* known nervous system damage, but whose behavior mirrored much of the hysterias of patients who had discernible brain pathologies. Charcot considered that for the 'nondiseased' subjects there was a 'functional' lesion and that as Hughlings-Jackson had noted for slips-of-the-tongue, the disruptions were psychological in nature and therefore needed a descriptive vocabulary at some remove from the physical level. Our point again is that nearing the end of the nineteenth century, a more proper understanding of the mental level of description was coming into vogue that would further delineate a phonological disturbance.

On final player in this is the neurosurgeon, W. Allen Starr (1889), who, in a long paper with many samples of fluent paraphasias, put another nail in the coffin of physical locations for fluent paraphasia. No one had then, or even now for that matter, made a localization claim for the 'functional lesion' that would give rise to slips-of-the-tongue. Starr's patients, and others from the literature he incorporated with fluent, sensory aphasia, had a panoply of posterior lesions that subtended the first and second temporal gyri, the inferior, angular, and supramarginal. He even noted some patients who had additional occipital lesions. Suffice, it says (Buckingham, 2006) that these phonological disturbances were pinned by a 'functional' lesion, and thus, for these cases strict localization was out of the question. Abstract mental levels of descriptions were becoming more fashionable for the physicians in agreement with Hughlings-Jackson and

Freud, and at the same time, the linguistics of the neogrammarians was developing along similar functionalism for historical language change, dialect differences within languages, and child language acquisition. The analyses of aphasia were consequently deepening along the lines of a mentalism, not unrelated to what Herbert Spencer had introduced decades before.

Hugo Liepmann

Ever since Hugo Liepmann (1900/1977) sketched out his initial model for three levels of acquired/skilled movements and their disorders, there has been a certain intractable overlap with language and speech. These movement disorders were characterized by lack of praxis (a+praxia). The hemisphere for these movement skills is normally the left and, as language and speech, are acquired cognitive skill sets. Moreover, there has always been a degree of anatomical propinquity between the two, and consequently both often co-occur in pathologies of the dominant hemisphere. Further, with modern models of embodiment and cognition, the two have become even more intertwined. Despite quite a bit of recent rearranging and subdividing of many apraxic syndromes (e.g., Heilman (2014, in press)), there is still some overlap with the disruptions of phonological processing in Wernicke's and conduction aphasia. These most often are the 'ideomotor' apraxias, and teasing them apart from phonetic/phonological disorders in aphasia is often tricky business. We offer this as another harbinger of phenomena that may color phonological disorders. Lesions in the left temporal lobe causing posterior fluent/sensory aphasia often cause as well interruptions of desired actions, whereby transmission of the willed actions to left premotor areas fail and thus become part of the syndrome of Wernicke's aphasia. Conduction aphasia and Broca's aphasia may also bring along an ideomotor apraxia interfering with the volitional control of limb (upper) movement on either side of the body. Although not necessarily involved in phonological disorders, we would point out the often proposed conflation of the articulatory aspects of Broca's aphasia with an apraxia of speech related most closely to an ideomotor apraxia. In fact, several researchers have likened Broca's 'aphemia' with an apraxia of speech (LaPointe, 2013), and one must admit that Liepmann considered the glosso/labio/pharyngeal muscles of speech to be limbs, although these limbs act in tandem and in 'center line' fashion as opposed to the arms.

Late-Nineteenth Century/Early Twentieth Century

The Phoneme

Although phoneticians such as Henry Sweet and scattered others hit upon the concept of the phoneme, they did not *use* the term as such. It was through the work of Baudouin de Courtenay (1885–86/1972) with his explicit differentiation between the phone (= the sound) and the 'fonema,' from the Russian (Robins, 1967). He did not conflate phonemes with graphemes, as many had done before, and more importantly used the important taxonomic prefix 'allo-' (one of a group of similar items, together which constitutes a structural unit). Therefore, what resulted was a very important three-level distinction between a phone, which, at the lowest level was understood as an element on its own – its individuality. This scatter of phones was later groped into allo-phones, which were then understood as a set of phones that structurally comprised a phoneme. As Robbins (1967, p. 204), the phoneme did not make its true impact until the work and popularity of De Saussure. Subsequently, there was some reluctant to accept this somewhat enigmatic concept – it existed and it did not. Although the early twentieth century phonetician, Daniel Jones, in 1914, made the difference between the allophone and the phoneme, his basis for distinguishing Henry Sweet's terms for 'broad' and 'narrow' transcription, there was some debate as to a kind of 'psychological' reality for the phoneme. Some phoneticians felt the phoneme to be a physiological entity. Others thought it to be a kind of "transcendental entity," while others simply felt it to be "… just a mere descriptive invention" (Robins, 1967, p. 204). Nevertheless, through the further work in the 'Prague School' tradition by the Viennese scholar, Nikolai Trubetzkoy in the years 1923–39, the phoneme because appreciated as a linguistic concept and accordingly was said to have a 'psychological' reality (Trubetzkoy, 1939). Principles of Phonology – (Trans. From the German "Grundzuge der Phonologie").

During this time, Edward Sapir (1925, 1949), as well, published his supporting views of the psychological reality of the phoneme. Without the phoneme, the nature of a phonological disruption could essentially only flourish in terms of a rather vague notion of 'the sound.' The consequences of this discovery had immediate consequences for aphasiology and the deepening of the analysis of the fluent phonological disorders in the sensory aphasias.

Roman Jakobson

Without a doubt, the transition from the Prague School to aphasiological theory in the United States came off the pen of Roman Jakobson, who, together with Gunnar Fant and Morris Halle, influenced the ongoing work at the time, which even extended into Russian, with the work of R. Luria. Early on, however, Nikolai Trubetzkoy, a fellow Russian linguistic, had paved the way for Jakobson to get into the circle in Prague, where he slowly left poetics and transitioned into structural linguistics of the day. From structuralism we see the movement into a more generative framework, where all levels of language were interwoven with a rule system, which described and extended phonology, phonetics, and acoustics into morphology and syntax. The neogrammarians of the latter part of the nineteenth century had tied together, synchronic (at one time) and diachronic (across time) studies to show interrelations between the major languages and their dialects, child language acquisition and historical change. Startling similarities in the processing of linguistic elements at all levels were immediately recognized. The linguists had also become engaged in the search for recognizable patterns in aphasia, as we have seen in the nineteenth century. This was notably extended in the aphasia investigations of Jakobson (1964).

Jakobson not only initiated early statements of syllable phonotactics, but he also utilized the new series of subphonemic features, which then provided a subdefinition of the

phoneme as being a bundle of features, a minimum of which could serve to uniquely identify each phoneme, while additional features could further refine the elements of the phoneme. The features often specified place of articulation in combination with acoustic spectra that could be correlated. So, Jakobson originally used (acute) and (grave) to distinguish the coronals from the labials and the velars. Features of voice, place, and manner were often the major features, marked as binary [+/−] to allow different arrays to minimally distinguish phonemes from one another. The feature itself became extremely useful as a measuring tool of segmental/phonological errors in phonemic paraphasia in fluent phonological disorders. As well, Jakobson borrowed heavily from the structuralism that combined Prague school phonology, the foundations of generative phonology, with notions as well handed down from centuries of Aristotelian association psychology.

In transitionalizing from poetics to linguistics, we glimpse Jakobson's contribution in light of his literary and stylistic studies of meaning to the scaffolding of the artistic scaffolding of meaning onto the study of language. He comfortably navigates tropes and turns them into linguistic dimensions. He made use of frequency of occurrence, the contiguity of elements and the similarity of units. Also, from the association psychology, Jakobson contrasted the 'syntagmatic' dimension with the 'paradigmatic' dimension, where the syntagmatic stemmed from the associative notion of linear order, serial processing, and associative chaining, and the paradigmatic dealt with similarity of form.

Jakobson transposed metaphor and metonymy into his paradigmatic (vertical) and his syntagmatic (horizontal) systems, where the paradigmatic, like metaphor, was based on meaning similarity. His metonymy brought in the most important associative process of Aristotle, which was contiguity – itself not related to meaning similarity but rather to a collocational process whereby reference was based on spatial nearness. Although most would agree that he pushed this distinction a bit too far when he tried to distinguish motor and sensory aphasia in terms of contiguity and similarity – the syntagmatic and the paradigmatic. The phonological disorders, our main theme in this article was one of substitution between features of one value difference (i.e., similar phonemes sharing all but one feature) – the phonemic paraphasias. As well, the sensory/fluent aphasias were replete with substitutions of words bearing similar lines of meaning – the semantic paraphasia.

No one familiar with phonological disruptions in aphasia can fail to note the early contributions of Jakobson as a combination of structural and early generative phonology. From his work on, the frontal lobe motor aphasias disrupted syntagmatic processing, while temporo-parietal lobe lesions gave rise to errors of either phonological similarity or semantic similarity (Buckingham, 2010).

As a modern reflex of Jakobson's model came the publication of Sheila Blumstein (1973), who benefited from his (as well as from that of Harold Goodglass) personal interest in and criticism, assistance, and advice to her while she was a doctoral student at Harvard. Measuring feature counts in the phoneme errors of Broca's, Wernicke's and conduction patients. Across the board, Blumstein found that a great majority of phoneme substitutions different minimally in only one feature. This shows that phonemes of similar bundles of features often substituted for one another, and differing only in one of those features. The feature (sonorant) soon needed to be added for further distinguishing phonemes in substitution errors. Often, an additional redundant feature (not needed in minimally distinguishing some phoneme) was the feature involved in substitution errors. Eventually, groups of features came to serve as specifying 'natural classes' of sounds, such as (+obstruent, – continuant) for the natural class of oral stops, thus separating nasal stops from the oral stops. In this way, we can clearly see that Jakobson's model began to add to what would characterize a phonemic disorder. Admittedly, Jakobson got into trouble restricting his syntagmatic disorder to the motor aphasias, since there he was interesting in serial order articulatory concatenation – not unlike Lordat and Wickelgren. In contradistinction, the phonological disruptions of the presumed fluent aphasias were replete with serial ordering derailments, but at the level of the phoneme, which we have already mentioned were signposts of phonemic paraphasia, where the errors were along both dimensions – for phonological breakdowns: substitution of phonemes and linear transpositions. Moreover, Blumstein's (1973) thesis charted large numbers of substitutions of phonemes in most aphasia syndromes.

Nevertheless, in a span of 2 years, Lecours and Caplan (1975) pointed out several problems in Blumstein's analysis that then reflected on what Jakobson had claimed. Nevertheless, the point here in our article is that Jakobson's phonology went a long way to characterizing phonological disorders – until problems with his model were discovered. As goes phonology, so go phonological disorders.

Slips-of-the-Tongue in the Twentieth Century

Essentially, except for a few nuggets in the first 60 years of the twentieth century, works on slips-of-the-tongue waned. In fact, Fromkin (1973, p. 14), in one of the most scholarly introductions to a volume of reprinted papers on lapsus linguae, writes, "There have been more papers utilizing speech error data written by linguists and psycholinguists in the last five years than in all the previous years of linguistic history." Our article, then, picks up this resurgence of the 1960s, hooking them historically back to Hughlings-Jackson, Freud and the others we mentioned above. Our point then, as is our point now, was that the study of slips not only paved the way for phonology and thus disorders of phonology, but present students have again emphasized the 'continuity thesis,' thus maintaining the connection between health and disease, and therefore in our case between the slips of the healthy and the phonemic paraphasias of the diseased. In the spirit of the present article, Fromkin (1973, p. 35) wrote that most segmental (and lexical for that matter) slips are, "…errors which appear to be solely phonological in nature, thus suggesting partial independence of phonology from other grammatical components." This 'phonological' mature is the modern day reflex of what Hughlings-Jackson and Freud were saying, except, as most know, Freud was quick to fail to notice that most often phonological errors were at the base of the generation of words and phrases that Freud tagged as driven from the troubled unconscious of his

patients – most often with sexual interpretations by him. Nevertheless, we demonstrated the fluent segmental errors of his patients with aphasia, whose 'blunders' were at a higher level of mentality and, if we take Starr's words, "resisted any clear cut location of lesions…as opposed to the motor aphasias of the day." Thus, phonemic paraphasias carried the linguistic stamp of segmental slips, and the higher cognitive levels were allowed to flourish at a mental level, thus holding to Herbert Spencer's theories of mentality and physicality.

However, the papers reprinted in Fromkin (1973) did not concentrate on aphasia, since the early papers there were more of a linguistic and/or psycholinguistic nature. Later, in Fromkin (1980), there were additional more recent studies that linked to the slip modeling of Garrett and Shattuck-Hufnagel in the 1970s and even more interesting to the connectionist models that would take hold a bit later on. We mentioned earlier that essential to all connectionist models was a *sine qua non* acceptance of the continuity theses, since those computation models had to be programmed first to simulate the normal patters, and those patterns were not culled from aphasia, but from the language of nonpathologically compromised speakers. Only then, could the IA models of Dell or the somewhat more cosmic Hopfield net architectures of the physicist Paul Smolensky's and his 'harmony machine' (see Goldsmith's brief but penetrating descriptions of Hopfield's contribution, 1990, p. 11ff). We will have a few more words on connectionist systems, mostly the IA work of Dell and his colleagues as it opens up a myriad of observations, predictions, and simulations of aphasia – mostly the fluent categories. One interesting observation from the Dell and Reich paper in Fromkin (1980) is the prodrome of Dell's connectionist work in from within the Stratificational model, constructed from the work of Sidney Lamb (1966) and explained to linguists by Lockwood (1972). If our readers want an interesting view of how Lamb's model in a way helped pave the way for Dell's later IA models. We mention this simply as another example of the continuity and change in the psycholinguistic modeling of human language and its disorders.

As we close this section, we would suggest that before reading more on spoonerisms, or exchange errors, which we will outline later with the work of Garrett (1984) and Shattuck-Hufnagel (1979, 1987), the reader should consult what we feel to be one of the first sophisticated and in-depth looks at the linguistic and computation nature of the spoonerism by MacKay (1973), reprinted in Fromkin (1973). For a full assessment of Wickelgren's (1969) paper on context-sensitive elementary motor response units and how it fails, the MacKay will make it abundantly clear that data from slips-of-the-tongue can be devastating for any type of context-sensitive, localistic, finite state automaton, even at the level of articulation.

The Serial and Hierarchical Model of Merrill Garrett and Its Sequalae

In Buckingham and Christman (2008, p. 136) in our section for further readings, we wrote of Garrett (1984) that his paper, "…represents one of Merrill Garrett's first forays into aphasiology, where he demonstrated how his serial model of sentence production could serve to characterize paraphasias and locate the principle production levels involved in their generation." There had been one paper published earlier (Garrett, 1982 in A. Ellis (Ed.)), where Garrett wrote on errors of speech, but where he again began to look at speech production data from "normal and pathological language." Our present article has maintained that models based solely on the slips-of-the-tongue in typical speaker–hearers of a language who have sustained no pathological insult are cast at the phonological levels. The continuity thesis assures the reader that models of this sort are cast largely at the level of human language phonology (both the formal and the segmental sublevels) and that they may be incorporated to describe and explain the paraphasias (both lexical and segmental) of the fluent/sensory aphasias. Based as they were originally upon slip data, models such as Garrett's made no commitment as to anatomy at first, and only later when they began to make correlations of the distinct symptomatology for the lesion locations from the clinics of these adults with aphasia. In fact, Garrett's model remained essentially a serial functional one, where lesion location was of secondary concern.

This goes for connectionist models as well; they are functional in nature although many PDF connectionists are mistaken when they claim that they are simply replacing a computer metaphor *for the mind* with a computer metaphor *for the brain* (Fodor and Pylyshyn, 1988, p. 62, ftn 35). Fodor and Pylyshyn write (1988, p.62, ftn 35) that, "… the issue is not at all which metaphor we should adopt; metaphors (whether 'computer' or 'brain) tend to be a license to take one's claim as something less than serious hypotheses …. The claim that the mind has the architecture of a Classical computer is not a metaphor, but a literal empirical hypothesis." We also should bear in mind Pylyshyn's (1980) distinction of the computer models programmed by computer scientists and the computer models programmed by psychologists. The computer scientist's program for playing chess, for example, is successful if the computer can beat a professional human chess player, without worrying whether it does so 'like' the human. The psychologist, on the other hand must go a step further and claim that the computer should behave like the human it simulates. Garrett had none of these commitments, since his model did not even hint at 'Hebbian Learning,' whereby Hebb wrote, "When an axon of Cell A is near enough to excite a Cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased" (cited in Buckingham, 2002, p. 287, see further references there). Garrett's model was a classical psycholinguistic construct for sentence processing, but existing as it did to imply performance and not competence; it had no commitment to computers or the brain. It nevertheless finally allowed linguistic aphasiologists the more closely characterized phonological disorders, where we should include the conceptual level as well as the phonemic form level. Lexical substitutions, accordingly, can be constrained from above by showing semantic similarity and can be constrained from below by showing phonological similarity. Our article, however, is focusing on the segmental rather than the lexical error, but this distinction is often blurred.

Figure 2 shows the overall levels and note the rough agreement with the architecture of J. Lordat's model discussed earlier. His model and its features represent real-language processing examples culled from his and others collected slips-of-the-tongue, produced online by speakers. He also holds to no a priori assumptions on the relations between formal grammar (see Levelt, 2008 and processing structure).

Figure 2 shows the overall set of levels in Garrett's (G) model. We are focusing on the development of modeling that provides functional commitments to phonology and therefore can locate the lexical and phonological processes and disorders in the model nevertheless build upon slips-of-the-tongue on the assumptions of the continuity theses. Note that both the serial models and the connectionist models we will discuss shortly accord with the now established continuum from health to disease, or put another way, from normalcy to pathologically compromised. Again, Buckingham (1999, 2002, 2006) provides more historical evidence for this thesis. A more detailed and succinct summary and comparison over some 20 years of work with serial and connectionist of Garrett, Dell and Levelt, and others is presented in Nickels and Howard (2000).

The top and bottom levels in Figure 2 of Garrett extend his model from the conceptual to the motor control processes. The conceptual and message levels are more pragmatic in nature; less built on slips per se, and, anatomically, would have to cross over to the right hemisphere, where discourse structures are determined by the inferential processes. These high-level processes determine as well as speaker's current perceptual and affective states as well as the ability to access schematic memory for information such as world knowledge. These levels map to the set of logical and syntactic processes. The derailments at these points of pragmatically for discourse, speech acts, implicature, structures more often misaligned in the dementias or in right hemisphere aphasia and thus do not directly affect the phonological disruptions treated in this article.

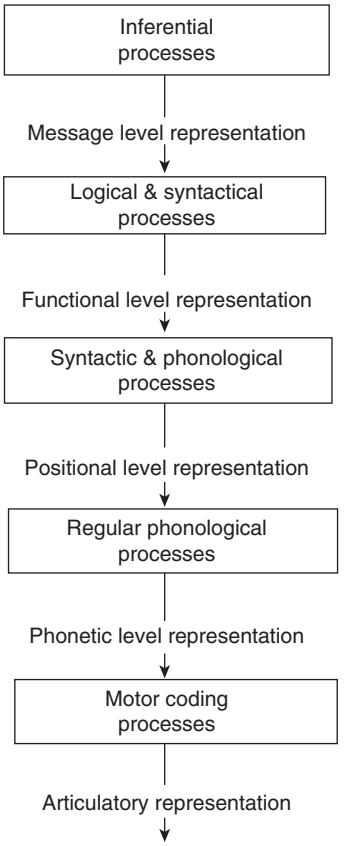

Figure 2 The overall hierarchical scheme of Garrett's model.

At the bottom, Garrett places the motor-coding processes that produce articulatory representations, which subsequently map to subcortical and cranial nerve systems. Slips here do not play into the structures so crafted, which as we claimed is the same case at the very high levels. The processes and levels between these are what we are focusing on for phonological disturbances. Garrett claims that these middle processes and the representations they give rise to are his 'sentence processes,' which in fact are his main focuses. His logical and syntactic processes build the initial linguistic structure. It is prearticulatory argument structure is at issue here. The calculus of $f(x)$ is built of atomic predicates, with a minimum of one predicate and one argument. Much has been written about argument structure at his level, where the argument or arguments are set in their functional/role vis-a-vis the predicate. Some say, as originated by Charles Fillmore (19 68) that some sort of 'subjectivalization' occurs, whereby one of the arguments will transpose around the predicate to become 'the subject' of the sentence, in an SVO language, which is the way English works. Of late, Chomsky has conflated these kinds of 'movements' with others a bit more typical of his earlier work. From Fillmore's case model, we can glimpse the prodromes whereby verb and noun selection must ultimately provide an utterance order where comprehension will involve linking the arguments to their semantic roles (agent, object, source, goal, experiencer, theme, location, dative) depending upon who is constructing the model and upon issues of whether agents overlap with being sources and goals being objects or datives. Often comprehension disabilities of Wernicke's aphasics involve the acquired associations between the predicates, the arguments, and the surface positioning of those arguments. What we have here in Garett's terminology is the functional level representation; it is still preutterance. The syntactic and phonological processes map from the functional- to the positional-level representations, where utterance word order is now in place. The positional level representation contains ordered and labeled frames for the content word lexical items, the frames coming with the fully specified grammatical morphemes – the way most believe. Garett points out that segmental slips-of-the-tongue and segmental phonemic paraphasias in the fluent and sensory aphasias are the derailments of the phonological (not phonetic yet) segmental ordering processes. The positional level in this model still only incorporates abstract forms, words, phonemes, and morphophonemes. Common slips-of-the-tongue here of words in sentences such as "He put the stove on the pot." The common phonological slips and paraphasias most often involve derailments of the phonological processes of segmental ordering, deletion, addition, or substitution. Phoneme ordering errors may be anticipatory, where a preprogrammed phoneme is anticipated

and makes a prior entry into the produced string. Sometimes the moved segment will simply be added to a word, deleted from it, or produced once again – a perseverative error. Note that these added segments come from somewhere in the target word itself, and again come from the nearby context. In slips-of-the-tongue, many exchanges often referred to as the 'spoonerism.' Exchanges, it turns out are rare in phonemic paraphasia, but not in the normal slip. MacKay (1973, p. 186) provides an explanation for why the adult with phonological paraphasia may produce only a few exchange-type errors. There he posits that a fusion of the two phonemes never occurs, which might be expected for a phoneme exchange, but he has suggested an ordering mechanism that metaphorically 'scans' (a mechanism borrowed from visual cognitive systems by MacKay, (p. 183)). In his seminar paper, MacKay writes, "The form of the scanning mechanism was arbitrarily chosen, pending further research." We mention the scanner metaphor at this point in our article because the further research apparently continued to justify the use of this cognitive visual metaphor (see Buckingham, 1986, 1995); it worked well until the connectionists appeared on the scene. It soon became clear that the scanner not only scanned the segmental units in its range (or as MacKay writes (p. 183, it 'sweeps' over the range) of motor engrams in a one-direction anticipatory (it looks 'to the future' in recent terminology), bringing them to threshold. Here, both Garett and MacKay would posit that there must be a kind of 'copy' of the segment. When speakers are not subject to brain damage many forces come to play in the spoonerism. In the uncompromised speakers, the segments scanned are already in a proper serial order. This so-called "first" ordering comes ready-made and ordered with the original selection of the "word" in its prearticulatory form. Subsequently, the phonemes will be scanned and copied during a second serial ordering, unavoidable in these psycholinguistic serial models. The second concatenation will be correct as long as long as the scanner operates in a sufficently slow manner, as MacKay (1973) makes clear. Processed without overarching speech, the normal speaker will produce no phoneme exchanges, i.e. spoonerisms. The slip-of-the-tongue occurs because if there is a relatively close range between the exchanged segments, where the one in 'the future' has tonic accent may cause it to reach threshold sooner than the phoneme preceding it, which is most often at least not the tonic accent – even though it may be stressed. This stress-dependent transpositioning over a relatively short range of items between the exchanged items is quite normal in the slips of nonaphasics. The reason for this is that the spoonerism must depend upon normal. MacKay builds into the scan-copying mechanism a normal reciprocal inhibiting capacity so that the exchange to play out. When this mechanism is not derailed, for example, in a non exchange for a slip, but where the second element is inhibited from moving out front and in this case it simply remains in its position, causing only the anticipatory slip. This then would not be an exchange but rather what some have called a 'doublet' creation (see Buckingham, 1989 in Nespoulous et al. (Eds.)). The point to make here for the purposes of our article is that the spoonerism that is produced must count on the working of this reciprocal inhibitory mechanism. Adults with phonemic paraphasia do not produce nearly as many spoonerisms as do adults without aphasia. The theory is that the aphasic derailments destroy this special normal process and therefore instead of the exchange, we more often see the first movement (the anticipation) but that initial phoneme that was bumped out of its initial slot cannot be held on to long enough for its 'post entry.' This is the crucial process. When the element that is anticipated is activated earlier, the original initial phoneme becomes 'hyper excited' and the expected rebound of excitation is sometimes sufficient to then allow the original displaced segment to then be copied into the vacated slot, thus the exchange occurs. According to Mackay's scheme, the rebound postexcitation discharge, which often occurs during the interaction of similar phonemes, can allow the placement of the earlier phoneme a kind as a 'post entry,' copying it now into the vacated slot of the anticipated segment. Again, the point here is that the phonological aphasic derailment will prevent this normal type of error, and therefore we see only a very few true exchanges of phonemes in paraphasia misordering errors. By the same token, one actually observes many more spoonerisms than 'doublet' creations in adults without aphasia. The Mackay (1973) is a must read for a proper understanding of what would be a more likely slip-of-the-tongue and what would be likely in a paraphasic phonemic serial ordering error. At the risk of being 'parochial' (or 'trite'), we would claim that all the disordered phonemic transposition in adults with phonological paraphasia do, indeed, 'look normal' in terms of their obedience to the 'phonotactic' patterns of the native languages of those adults. Although we will only briefly touch upon the competence model of phonology referred to as 'optimality theory (OT),' and its mappability onto the harmonic Hopfield Nets connectionist architecture of Smolensky, we will touch upon a presumably more profound deepening of just what phonotactic constraints may entail, and they appear to entail it for both optimality competence theories as they might square with the distributive and nonsymbolic connectionist models of psychologists qua physicists, such as Smolensky (1988). These are qualitatively quite distinct from the non-Hebbian, IA architectures of Gary Dell and his colleagues. We will also briefly discuss the latter models, which have had great success in describing and explaining a wide swath of phonological paraphasic errors – many of which deal with segmental ordering, both anticipatory and 'carry over' or 'perseverative.'

Figure 3 in the Garrett model shows the internal structure of mapping from the 'message level' of representation to the 'functional level' of representation.

Here, two processes unfold, much of it in parallel, which is a thorn in a serial model. Nevertheless, at this level, words (the content) are selected on the basis of their amorphous meaning along with the determination of linear slots (frames). The frame and its contents here are abstract and show the logical relations of words in their functional order – which is not truly and order. Here, the amorphic words are 'assigned' to a deeper logical and relational order. Based as these levels are on slips-of-the-tongue, at this level we have either a phrase length frame or a small clause, which may indeed have two deep sentences. At this level, slips look simply like meaning-based lexical substitutions within one slot. The slips here, of course, have no phonological similarity, since we are at

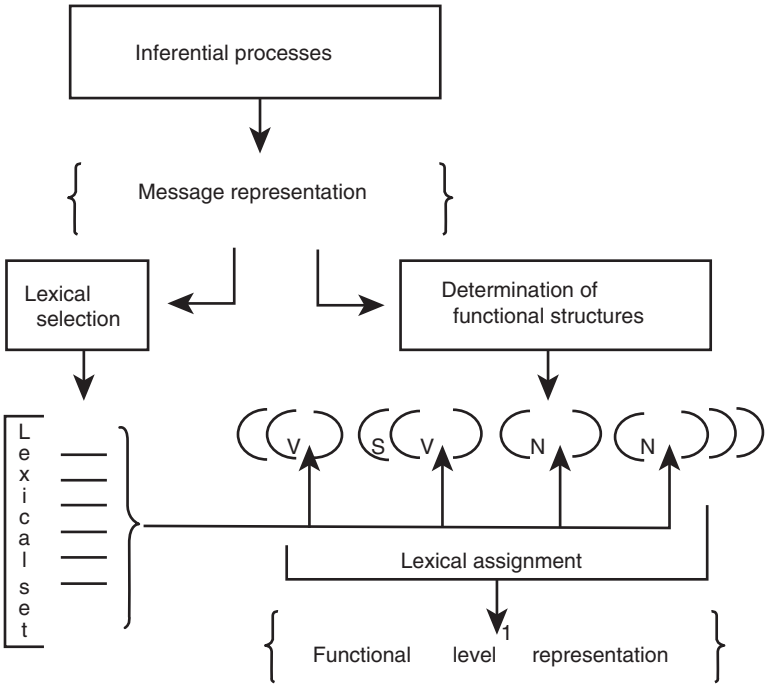

Figure 3 Sketched here are mappings from the inferential processes through the message representation through amorphic semantic lexical selection and the determination of functional frams. Lexical assignments are placed into the non-utterance ordered funcitonal frames. From the M. Garrett Model, Permission to reproduced by MIT Press.

a pre-prearticulatory level. There can also be word-exchange slips, which cross over sentence boundaries, often simply infinitivals such as "[Cat, it's too hungry [for you to be early]], where there is an S(entence) boundary between 'hungry' and 'for.'" These kinds of slips can cross prepositional phrase boundaries as well as a conjunction S boundary. Although we see many semantic substitutions in fluent sensory aphasias, we see fewer word-exchange errors in aphasia, and we would claim that the exchanges of words likely have to have sufficient normal processing to occur as we saw with the spoonerisms of MacKay. Again, all this processing is quite removed from the utterance. It does provide us within Garrett's model of a 'functional level,' which is roughly of the form of deep case or argument structure, where all is relational with no linear order. Needless to say, there is a plethora now in the syntax and semantics of case grammar, thematic roles and the 'syntacticon' of predicate/argument structures and co-occurrences. Here, the 'syntax' is the frame order (non-utterance) and the 'ticon' is the lexicon aspect of it all. For, example, verbs are often said to be stored in the lexicon in terms of the typical arguments they code with. And, of course, in a model like this we would have to have some kind of statement as to which argument would show up in the position of the subject, etc. That is more in line with the pragmatic needs of the sentence producer. But that is quite another story for the purposes of our article.

In Figure 4, we approach the next levels of representation, where now the functional level maps to the positional level. Although we have both words and frames in this mapping as well as above, we are now working into an utterance order – albeit still prearticulatory. This is an extremely important point because the model is readying the words based on their phonological properties and those words are considered to contain their phoneme and morpheme building blocks. Since phonemes are not sounds but abstract sets of sounds, and consequently, we should not blithely conflate the phonemics with the phonetics – although we are entering into somewhat murky waters, since some believe you cannot tease the two completely apart, and some, such as Ohala (1990), deny that there is any 'interface' whatsoever between the two; they are hopelessly intercalated.

The lapsus linguae from which Garrett's original architecture was proposed and the continuity thesis extend once again into the symptom-complexes of phonological disorders, which mirror the processes that map the Functional Level Representation onto the Position Level Representations in his serial model. In the Garrett model, this is the crucial nexus for phonemic paraphasias. We imagine what is going on with the following serial jargon of the model, with an eye toward later considering its worth when confronted with its replacement by connectionist simulation in our historical approach to phonological disorders within neurolinguistic cognitive science, which we would claim is a telescopage of Caplan's (1987) neurolinguistics and linguistic aphasiology.

First notice that there is a retrieval of word forms, which is the second lexical lookup, but this time it is the phonological structure of the word. We do not say that this is after the meaning based word is selected, since the point will be that this are dissociated from one another. They are dissociated when speakers 'know what they want to say' or 'know what the word is,' but cannot retrieve the form despite this. Nonbrain damaged adults do it, adults with aphasia do it, and Alzheimer's patients do it. There is nothing new to this, since the

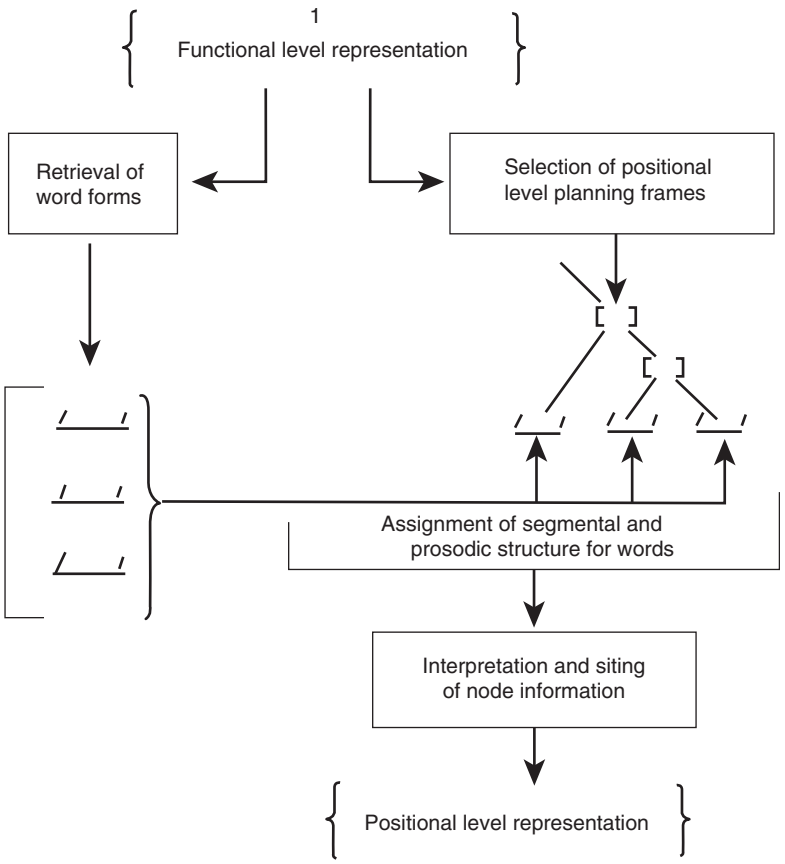

Figure 4 The mapping from the functional level represention (computed in Fig. 3) through the retrieval of word forms and their serial asignment into the positional level frames. Here the segments and the metrical structure of the words are placed into positional level slots in syllable sized chunks along with attachment of the morpho-syntactic and syllabic node information. The output of these processes unfold into the Positional level representation. From the M. Garrett Model. Permission to reproduce from MIT Press.

descriptions of the state of mind of speakers in this disconnected puzzle are legion and go back centuries (Benton, 2000, Chapter 7).

Positional-level planning frames are also encoded. These come with their Fo contours, where the tonic accent has been established at the message level, since, for example, if there has been stress movement, contrastive stress, or where the major pitch change (at least for English) has been relocated for pragmatic purposes. Here it simply comes coded as the sentence is being readied for a closer approximation to utterance order. The frames are coded with grammatical category marks, and affixal morphology, which is nevertheless, in a preaccommodation morphophonemics of sort. This is not only because eventually Garrett want to handle agrammatism as best he can, but also because the morpheme slip-of-the-tongue phenomena show that morphemes can move or be deleted, and the root forms to which they are underlyingly attached accommodate to the morphological alterations. The range of these frames are small clauses, not unlike the message-level logical frames – except that the material here is abstract-form based, since we are at the second lexical lookup for 'words' at this level. An 'abstract' form base should not be a contradiction, because the forms here are phonemic or morphophonemic. At this point, we consider how the model gets the words into the frame slots.

Here, we see a slotting of words: includes situating the segmental units in their normal serial order. This is to say, major processes of serial ordering for utterance order take place here, and as the reader can tell from examining **Figure 4**, there do not seem to be any ways to do this. Naturally, and with the contribution of Shattuck-Hufnagel (1979) and others, three mechanisms were conceived to operate here. In fact, these mechanisms were necessary for the models to describe the actually serial ordering. We have discussed anticipatory slips in spoonerism, with the additional complex intactness of super activation and quick erasure to get exchanges over short ranges through. Recall that the exchanges are most always over short ranges of words. It was in the Mackay paper that we introduced the notion of a 'scanner,' which could look over the range of material to locate the material to be situated into the frames. In its normal function, it looks 'ahead' most often and when it goes awry, the error transposition is one that 'moves from right to left,' which means it gets situated earlier than usual; it comes from the 'future,' and enters the 'present' in new terminology from Dell's work, although we will see shortly that IA is philosophically, computationally, and functionally opposed to serial models. Nevertheless, Lashley (1951), saw the need for anticipatory error in normal slips-of-the-tongue. They are seen in aphasia as well, with one very importation that has come to light from deeper looks at slips. We will discuss this

later in the article when evaluating Dell and Martin's contributions.

Again, for our purposes in this article, phonological disorders are captured by the 'scan' mechanism. However, to finish off the delivery processes into the frames, we must add the obvious fact that the items are 'situated' in these frames, and most feel that this should be part of the scanners operation. Accordingly, the mechanism came to be called 'the scan copier'; it scanned and copied 'the seen' items into their frames. MacNeilage (e.g., 1998) has written at length on slot/frame and their contents in speech production as well in the evolution of human speech. Syllable constituency, too, must be nestled at this junction, since slips-of-the-tongue as well as phonemic paraphasias are sensitive to the constituency. However, it goes without saying that how different models handle syllables may be quite varied, and those models will color how one explains and describes the 'housing unit,' which is the syllable (Meyer, 1990, 1991, 2000; see other contributions in Weldon (Ed.), 2000). Some investigators focus on an absolute necessity of the vowel portion (the NUC of Prince and Smolensky (2004, p. 33), with the universal CV where although not necessary for all languages (thus not absolute) and a CVC with a final coda consonant; these constituencies are much rarer. This view of the syllable stems from 'OT' (McCarthy (Ed.) 2004), which we will briefly mention later. Still others opt for a binary constituency of onset + rime, where the rime consists of the vowel ('nucleus') + coda (Buckingham, 1986). A strict adhesion to one and only one theory of syllable constituency is unwise at this point in our knowledge of speech production. In short, the scan copier must be sensitive to syllabic slots within the utterance, the boundaries between the syllables and the metrics, which would adhere to the overall prosodic structure of the resulting positional level representation.

Admittedly, the scan copier as originally characterized did not have the capacity to appreciate phoneme similarity, since that would have involved subphonetic manipulation and the ability to keep in mind feature sharing counts. The scan copier worked essentially on the segmental level and not on the feature level. Substitutions in adults with aphasia seemed to favor interactions between phonemes differing minimally, more often differing by only one feature. Blumstein (1973) as we claimed earlier found more substitutions with fluent aphasias, but of those that were made by Broca's, Conduction and Wernicke's had similar statistics, favoring the one feature differences. Similarity will turn out to be more easily captured with the continuous numerical values referred to as 'weights,' whereby the weights of /b/ and /p/ would be closer in value that the weights of /b/ and /k/, given that the frequency of occurrence of these phonemes were the same.

Now, the question comes as to how to get rid of the phoneme once situated (copied) in its slot. Thus, a 'checkoff' monitor was conceived to make sure that once a segment was slotted – correctly or incorrectly – this monitor would eliminate that segment (Buckingham, 1986). This, too, would become more easily handled by a manipulable 'decay rate' in connectionism, whereby the segment could be checked off at a rate, which in turn could be fixed to work overly fast, overly slow, or in a normal rate that could be determined by simulating either that would maximally work with upcoming slotting, but could then be fixed to simulate a breakdown whereby the segmental was checked off overly fast or whether it would be checked off overly slowly, which may begin to compete with new stimuli or otherwise be activated a second time. Decay rates as continuous connectionist checking off would become an important addition to connectionist systems.

The phonological disturbances were not at least describable in terms of breakdowns in the mapping computations that structured the positional-level representation from the functional level of representation. Phonological paraphasias could be substitutions, linear ordering errors, both anticipatory and perseverative – looking ahead to the future and bringing an element to the present or holding on to some element in the (nearby) past and copying it again in the present if not checked off by the monitor. Most often, an item will be anticipated but the monitor will fail to check it off, thereby creating a 'doublet' of the segment or segments (Buckingham, 1989). Often the derailments are severe enough to prevent recognition of the target. That is, the error cannot be related to a target, otherwise with less-severe errors, the target can nevertheless be gleaned. The accounts of so-called 'target-related' errors are rather straightforward. On the other hand, nontarget-related errors present a greater challenge to analysis. Often, these nonrelatable errors are categorized as 'neologisms.' Although a strict dictionary definition of a 'newly created form' is a 'neologism,' target-related errors are unfortunately logged as neologisms as well. We refrain from this confusion, and would claim that the less severe error where the target is recognized is not technically in the general aphasiological usage referred to as a 'neologism.' For example, 'neologistic jargon' is not restricted to target-related errors. The neologism has been enigmatic and has always required a good deal of caution in its analysis. Early Boston Diagnostic Aphasia Exams simply considered the neologism to be operationally defined as a form whose phonemic makeup was over 50% off from the target. But, note here there was great difficulty, because the target would have nevertheless had to be recognized in any event. That type of account nevertheless claimed that the neologism was produced all and only by phonemic distortion of a word. Buckingham (1977) questioned the exclusive recourse to phonemic errors, which had come from the work in Kertesz and Benson (1970) and came to be known as 'the conduction theory' of the neologism. An additional proposal then followed such that a neologism may have arisen through a more complex breakdown that included the possibility that there was a first lexical lookup, but that second lexical lookup was either not retrieved or only partially retrieved, whereby other segments would have to be produced in many cases. As per Green (1969), much alliteration and assonance would color severe neology as a result of anticipatory or carryover doublet creation – imparting a 'poetic' hue to the output. Neologistic jargon in the fluent/sensory aphasias took on a life of its own shortly thereafter and different accounts arose.

Finally, with the serial models, there seemed to be no inherent reason why there would be different numbers of anticipations or carryover. Further, and very importantly, rarely if ever do the paraphasic errors result in nonpronounceable words in phonological disorders by disobeying phonotactic constraints.

As we omitted the higher pragmatic message-level processes, we will only note that the 'regular' phonological processes that feed into the phonetic representations in **Figure 2** are largely coarticulatory and general allophonic processes that take us beyond the focus of this article. Garrett was obliged to finish out these lower-level processes, since these final readying phonetic computations ultimately mapped to the actual gestural/articulatory through motor control to the needed sets of cranial nerve innervations. These levels were post slips-of-the-tongue as well as postphonemic paraphasic errors and, in fact, postspeech apraxic gestures. The kinds of breakdowns with 'motor coding processes' were the various dysarthrias described in works such as Duffy (2013).

Neologisms and Neologistic Jargonaphasia

Much has been garnered in the description and analysis of the neologisms in jargon aphasia. First off, we distinguish between 'phonemic' jargon as well as 'semantic' jargon, although both these forms are typical of fluent aphasia. The semantic jargon, where the paraphasias rest largely at the level of *word* selection, has been recently linked to segmental selection processes in that the perseveration involved with derailed phoneme and word selection stems from a common input computational breakdown of weakened activation of the words or the sounds (the phonemes, to be sure) searched for (Martin and Dell, 2007). This then scrambles the postactivation environment with 'unwanted' forms, which then 'compete' with the target on some similarity association and thus have an almost equal footing in their ability to be retrieved in place of the target. So, together with overall weakening ('deafferentation' is used by some) of target words or phonemes, there is a computational/selection misfiring, placing some recently produced error in the 'pre-checkoff' time period (early decay time range) such that it is often perseverated. Moreover, when these error forms are in the recent, postactivated, surround and if they happen to be associated with the sought after target on either a phonological or a semantic basis, the likelihood of being perseverated is enhanced – purely associationistically. Aristotle quite often rears his ancient theories in structuralist aphasia models. It is also not unusual for lexical selection errors to be associatively related to phonologically similar and semantically similar words. Even more important is that the lexical paraphasias often share even stronger associative links, since many errors (often noted in slips-of-the-tongue) link to the target along both phonological and semantic similarity. Aristotle would be quite pleased and vindicated in all this, since so much of computational strength and weakness is due to associative factors, which were discovered centuries ago (Buckingham, 2002). The interactive activational connectivity among the two lexical look ups in the Garrett serial model and in the Dell and Martin IA (symbolic) model in the connectionist paradigm is thus supported even further. The 'repeated phoneme' effect we discussed earlier, the 'lexical effect' observed in phonemic paraphasias where the errors are more likely to be produced if the error turns out to be a word. Connection weight may account for this as well as a late output monitor whose job is to check if the output is a word in the dictionary; that is all. The 'gendarme' metaphor would not go far enough if to the lexical effect constraint, we add semantic and/or phonological feature sharing as well. Weight differences would be explanatory here, since they follow the dictates of Aristotle. The Garrett model does not. This, of course, relates to the deep connections of Donald Hebb and others before him who dared to think about running associationist mathematics and statistics over computers, which could then, with programming finesse, simulate normalcy and abnormalcy in language and speech. Suffice it to say that it is extremely likely that the productions of neologisms may involve similarly tightly yoked connections between Lexical lookups #1 and #2. Since Buckingham (1977) and before, with Luria (1970, p. 296) and Pick's (1931/1973, pp. 56–57) notions of a 'two-stage error' (Buckingham, 1981, p. 48) where some coalescence within a word where the semantic is blended with the phonological, with say, some additional phonemic paraphasia, a nonrecognizable word may appear. The 'two-stage' account obviously stemmed from serial thinking early on, and therefore worded the processing in temporal terms – first a semantic selection error at the functional level and phonological errors applied to at the positional level, thereby rendering the production unrecognizable Lecours and Lhermitte (1972, pp. 304–305).

One could suggest that these neologisms stem from monemic (semantic – HB/SB) paraphasias where the lexical material is subsequently the location of a phonemic transformation.

This scenario has in one form or another come to be known as 'the anomia' account of neology. But, this explanation has more recently run into how we interpret the hesitation and halting behavior of a patient searching the lexicon. This brings us to a recent study of recovery from jargon aphasia to 'anomia.' Note first off, that the anomia end stage was first considered in depth in Kertesz and Benson (1970) for neologistic jargon aphasia evolution/recovery (Buckingham, 1977).

Recovery

Eaton et al. (2011) have recently published an in-depth study of the recovery on a patient who initially had jargon aphasia. To date, this study represents one of the most complete pictures of recovery of a single case of jargon aphasia. The subject was a 67-year-old male, who had sustained three strokes, the most severe being the third in June 2002. CT scanning revealed a large lesion in the left MCA region. The patient had a fluent aphasia with nonwords (phonemic paraphasia) and inappropriate words (often without semantic relations with a target) in paragrammatic, "… sentence-like structures" (p. 1547).

Over a period of 21 months, the patient was tested periodically on single word naming, reading and repetition eight times during this time period. Both correct responses and error responses were noted at testing and later analyzed structurally in terms of whether the words were actual words in an English language dictionary (the Collins Concise Dictionary, 2001, edition) or whether they were nonwords. The first four confrontational tests were in August and September of 2002, 2–3

months post onset. His most severe stroke in June of that year had precipitated the first aphasic disruption. At 7 months post onset, in January 2003, the final three tests were administered.

Monitoring versus Anomia

Very importantly, Eaton and her colleagues went a step further to add to their single word corpus by including three samples of connected and elicited spontaneous speech by providing the subject with pictures, which he was asked to describe – not just name. The first sample was collected within the first 7-month period – actually at 1 month post onset … July 2002. After the single word testing period, two more samples of 'connected speech' were taken at 18 and 30 months post ictus. That brings the range of recovery time closer to their 21-month period. In the conclusions to the study, the single word testing as well as the 'connected' speech samples were coalesced. One important finding was that there was indeed change over time and the authors want to suggest that there is now an ambiguity in the interpretation of the end stage hesitations, struggles, and other 'trouble-indicating behaviors' that seem to the authors to *not* indicate lingering anomia, as claimed by Buckingham (1977), Kertesz and Benson (1970), and Butterworth (1979). This 'strategic adaptation' also seemed to be observed in Kohn and Smith (1994), while two further subjects out of four in Kohn et al. (1996), "… showed no change over time." More studies are needed at the end stages of jargon aphasia to continue closely monitoring the subjects' monitoring abilities to be sure that their stalled behaviors on naming indicate blocked access to the word with egregious error productions during the block, or some periods of no-response, or some internal blockage of felt incipient errors, which are then themselves blocked so as to not produce errors. This is quite another cognitive set of affairs from a 'pure anomia,' and is clearly important in the charting of end-stage recovery abilities. We did not see end-stage testing of comprehension in the Eaton et al. study, while we note a speculation of Nichols and Howard (2000, p. 126) when they were considering the clearing of phonemic paraphasias in a case study being described – but where the patient still produced omissions semantic substitutions. They considered a possible lingering anomia, but reflected that perhaps the patient was, "… using his good comprehension abilities to monitor and edit out, formal paraphasias before they were articulated, but there is no positive evidence in favor of this."

Most of the Eaton et al. study supported other recovery findings, such as (1) more correct responses over time, (2) fewer nonwords; (3) trends where errors slowly proceed from 'nontarget related' to 'target related.' Some use the dictionary definition of 'neologism' and claim that both sets are 'neologisms,' some relatable to a target, others not so. We do not think that this usage of the dictionary definition is felicitous in aphasiology, since 'neologism' has risen to a more technical usage in our field. On the dictionary meaning of neologism, we have *any* new coinage, presumably any slip-of-the-tongue that happens to be a nonword or any phonemic paraphasia, whereby a nonword is produced by one simply change, the so-called 'target-related neologism,' (4) a suggestion that since repetition and reading are computationally easier than naming, those tasks recover better and throughout recovery are generally performed better by the subject. They take Dell et al.'s (1997b) 'global' approach to noise in the system or lack thereof for naming, repetition, and reading, with the last two dissociated from the total recall task of naming. At least both reading aloud and repeating include supplying the subject with elements of the needed response. Naming is total recall – no hints added. We would note that Dell's more recent work with deafferentation and decay rates are much less global in nature, where noise to the system is much more restricted in cognitive processing and local 'lesioning.' And, finally (5) although perseveration featured strongly in this subject's jargonaphasia, they found no *significant* decrease in perseverations over time. They recorded total word perseveration and segmental perseverations that blended with other words, nonwords or previously perseverated words in close range. Buckingham and Christman (2004) described the syllabic constraints obeyed at the splice junctures, where the perseverated material blended with another form. We are not aware of any 'prediction' of a *significant* decrease in perseveration at late states of recovery. Nevertheless, we do agree with Martin and Dell (2004) that a trend in this way is almost always seen, and therefore their 'anticipation ratio' is always higher as the patient recovers (caveat, Eaton, et al., 2011). This is an extremely important metric of recovery to normalcy in an adult with aphasia, since it has been well established that slips-of-the-tongue in nonaphasic speakers favor anticipatory errors, with much fewer carryover errors. This is empirical evidence that recovery in aphasia approaches normalcy and the 'ratio' can serve as a metric of recovery through the evolution of the syndrome. The authors do note, however, that full word perseverations did not even, "… move in the predicted direction" (p. 1554). We do not know what the anticipatory ratio was; perhaps was lower. Their blended perseverations in all tasks, "… showed some reduction" (p. 1554). Perhaps what all this means is that in fact the patients had not truly recovered as well as the authors feel, and that perhaps their 'troubled naming' was still indicative of an anomia and not so much improved monitoring and inhibiting incipient, prearticulatory paraphasias. This presumed ability needs much more investigation as to whether all monitoring behaviors were preutterance and not immediately post utterance. From the study, it would appear that the error monitor was a prearticulatory ability, which on the face of it is of tremendous cognitive significance.

It goes without saying that recovery studies such as Eaton et al. (2011) are the *sine qua non* of an improved understanding not only end stages, but also of a better understanding of the acute stage behaviors. Here, we would point out as well that Jane Marshall, second author of the present study, has contributed much of late to the overall study of jargon aphasia and its neology (cf Marshall, 2006). The reader would do well to consult that work and follow up with the reference to much prior work.

From Serialism to Connectionism

The Lull before the Storm

Before we bring in connectionism, we should mention that up until its dawn, there had been a host of modern technical

neurolinguistic studies of phonology and the brain, which did not ruminate on how the disorders could possibly be simulated on a computer. It would be useful for the reader to consult Whitaker (Ed.) (1988) for investigations on (1) how the nervous might acquire phonology; (2) how phonology may involve not only the cortex, but the subcortex as well; (3) how Alexander Luria's 'afferent motor aphasia' and conduction aphasia overlap to a reasonable degree and how they confront the interleafing of phonology and phonetics; (4) to what degree can phonetics and phonology be separated from motor planning and execution; and (5) clinical/neuropsychological studies of the specific disruptions of oral reading in deep dyslexia.

Wither Simulation

Since connectionist models have had great success at simulating aphasic behaviors that were heretofore accounted for, described, and explained with serials models, such as Garrett's. While we have previously published analyses of phonological disorders in terms of scanners, copiers, erasures of produced elements by a check-off monitor, random generators (Butterworth, 1979; Buckingham, 1981), and final output stage palace guards checking to assure that the output has only words in it. As noted earlier, Buckingham's (1986, 1989) studies of phonemic paraphasia made recourse to all these mechanisms proposed to operate among the processes that mediate the functional representations and the positional representation in serial models such as Garrett's. Connectionist programs properly structured in terms of connectivity strength (mathematical 'weights') between elements on counts of their frequencies of IA, will, with little difficulty when run thousands of times with input 'jolts' to prime the action produce with a good degree of 'faith' the response output. Guided by the continuity thesis, once fairly normal responses can be produced, the associative values of connections and their localization in the system can be manipulated in a vast number of ways to evoke distinct kinds of error patterns. Wheeler and Touretzky (1997) showed this forcefully comparing their simulations to what Buckingham (1986, 1989) described. What Buckingham had detailed using serial mechanism was computationally simulated in a much less cluttered and complicated fashion, suggesting that the serial mechanisms might just as well be eliminated as unnecessary. Further, the mathematical 'lesioning' could be done either locally or globally, with different kinds of error patterns. Altering weights at different points of the system resulted in errors that appeared to be global in nature. Altering them differently would produce 'even' more haphazard 'noise,' again producing errors of a more global nature. The system could be altered to check predictions in many ways. Weight changes with individual classes of phonemes at lexical lookup #2 could produce substitution errors, while slowing down or speeding up of decay rates could simulate perseverations of phonological segments recently activated. Speeded look ahead processing would simulate anticipatory derailments in linear order. If programmed correctly for normalcy, the outputs would often produce spoonerisms, for example, as long as the activation/inhibition parameter switch mechanisms were set according to Mackay's theories of how exchanged items interact to work correctly. Alter these systems with weight change 'lesions' and the runs will not likely produce spoonerism and will produce more simple carryover or anticipatory errors. As far as we can tell, there have not been many studies to date comparing the spoonerism with the many single segmental misorderings, where the anticipated segment remains in its original slot, thus creating the 'doublet'; out of one segment you wind up with two of a kind. Both Wheeler and Touretzky (1997) and Hillis et al. (1999) simulated the production of neologisms, but only used one location of weight lowering (weakening), which was locally manipulated from within the processes mapping words onto their phonological makeup. Small weakening only simulated simpler phonemic paraphasias, but more severe weight weakening produced more severe paraphasias, many of which reduced the input word's recognizability. Some add the modifier 'abstruse' to the frank neologism, but we note that these neologisms arose from only weakening word to phoneme links, and so did not account for a possible anomia underlying the neology, since inputs to the phonemic level in LL 2 had to be properly retrieved words at LL1.

Simulating Pavlov

In Buckingham and Christman (2007) we noted a very interesting finding by Harley and MacAndrew (1992), which seemed to have actually simulated Ivan Pavlov's 'paradoxical state,' later borrowed by Alexander Luria for word retrieval phenomena (1973), where activation dynamics triggered a reversal in stimulus/response such that seeking target behaviors would actually retrieve much weaker or peripherally related items rather than the more frequently and heavily weighted target items. These Pavlovian rules of force were three and are as follows: (1) normal rules, where proper behaviors were elicited at stimulus/response settings, (2) a phase of 'equalization' where because many response types are activated simultaneously, there was no response at all, and (3) the 'paradoxical' phase, were normal rules of force were actually reversed.

Harley and MacAndrew simulated this paradoxical state (see **Figure 5**) in cortical neurodynamics whereby random 'lesioning' with 'noise' superimposed upon a system with an already weakened weight value of the connection strengths within the semantic (conceptual) level to the lexical level in their schema (not unlike the Dell model (1997). The reader of this article will have to consult Buckingham and Christman (2007) for the details of how Pavlov altered the many processes in his 'conditioned' response learning experiments with its 'laws of strength/rules of force.' Unwittingly paraphrasing David Hartley (1746, 1749) (Pavlov (1962, p. 88) in a lecture at the Sorbonnne in 1925 wrote that, "At the present time we have facts which warrant the assumption that the idea of coupling (associating by conditioning – HB/SC) is even an elementary physiological process." This sort of modern day technological simulation of Pavlov's findings attests to the importance of connectionist work. It bears saying in closing that the paradoxical state may indeed be involved at the very least in what was causing 'ottoman' to be easier than 'chair' in the work of Marshall et al. (2001) where they found … an inverse frequency effect in jargon aphasia and hypothesized that our reversal of the rules of force (Pavlov's term) was at least somehow partially caused by the semantic nature

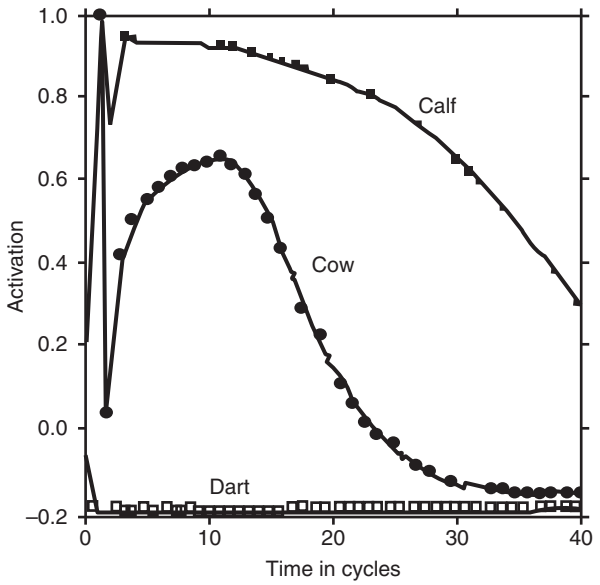

Figure 5 Graph is the result of a combination of connectionist lesioning by "random noise" added to already weakened connectivity of lexical networks. The addition of the random noise throughout the system of prior weakened connections simulated the Pavlov/Luria so-called paradoxical phase of lexical selection whereby a picture of a cow elicited a "weaker" connected element (calf) that the stronger element (cow). From T. Harley and S.B.G. MacAndrew model (1992). Reprinted from their article (1992) and from Buckingham and Christman 2007). Permissions obtained from authors as well as from Psychology Press.

of low frequency words. Recall that Harley and MacAndrew (1992) found the paradoxical effect (simulated it) only when they added random variation ('noise') by a distributed lessening of weights, "… into the weakened weight connections of the semantic to lexical levels." More recently, there is a follow-up on Marshall et al. by Hoffman et al. (2011), where again the semantic component was implicated as the location of the breakdown. They hypothesized that in an altered semantic system, the low-frequency words are much more prone to be perseverated and thus appear to stick in the production program longer than high-frequency words. These authors cleverly tested their patients on 'delayed repetition.' They found a significant advantage for precisely the low frequency, and this they correlated with the perseverative rate of those low-frequency words. This gave rise to the oft observed production of a low-frequency word substituting for a higher frequency in the semantic field on repetition. This would be a plausible explanation for the otherwise paradoxical reversal of expected frequency effects, which is what Pavlov found. It appears that Pavlov as a behaviorist described the physiological phenomenon correlated with the behaviors, but he did not go after any underlying cognitive processing account. Harley and MacAndrew in their connectionist 'lesioning' had weakened weight connections between the semantic and whole word lexical levels in their system, while subsequently adding some random 'noise' (random distribution of further weakening to the already weakened connections) in the same domain (semantic to lexical level). Harley and MacAndrew's simulations are at the very least intriguing. More awaits this intriguing reversal of Pavlovian rules of force.

The Connectionist Model of Gary Dell

In early work by Dell (1986) and Dell et al. (1997a,b) his model is introduced to a wide audience of aphasiologists – theoretical and clinical. Along the way, this model has been evaluated, assessed, criticized, praised with peer approbation, and further tested/expanded, by Dell and others (cf Martin, 2000; Martin and Dell, 2004, 2007). The model has IA architecture as seen in **Figure 6**.

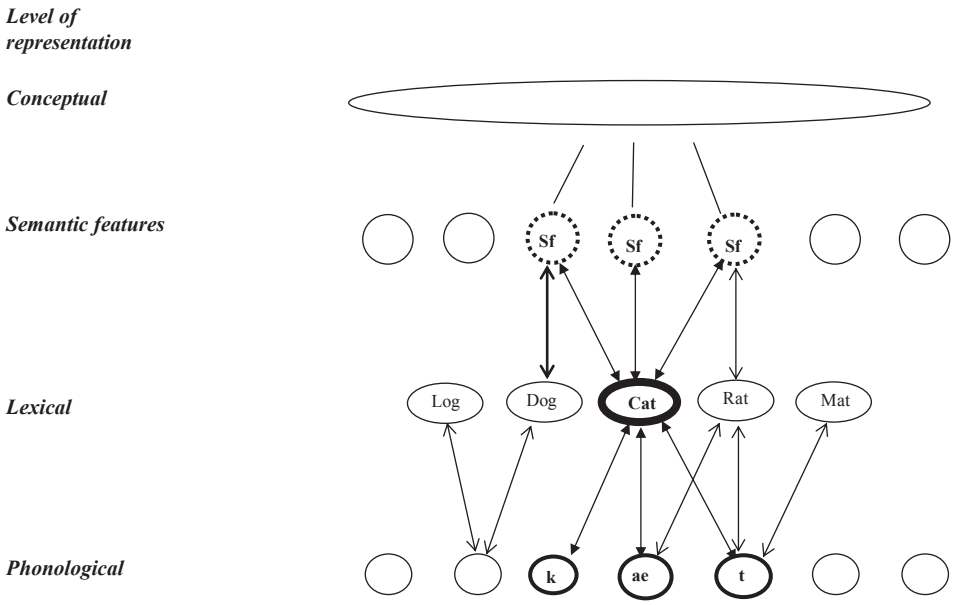

Figure 6 From Martin, N., Dell, G.S., 2007. Common mechanisms underlying perseverative and non-perseverative sound and word substitutions. Aphasiology 21, 1002–1017. Special issue on Verbal Perseveration. In: Stark, J.A. (Ed.), Aphasiology. vol. 21, no. 10/11, Oct/Nov 2007.

As with Harley and MacAndrew's model, Dell does not invoke Hebbian Learning, but rather fixes the connection weight values in assuring that each input (the 'jolt' that gets the system operating) will normally be matched with its output match; word X in, word X out, with the heaviest weight total being the most activated, Dell's system in its initial crafting will simulate nonpathological productions by and large. He can then attempt to simulate the slips-of-the-tongue (of English), which of course constituted the elements and patterns he used in the first place, abiding thusly with the 'continuity thesis' we discussed above – the *sine qua non* of connectionist systems, Hebbian or not. **Figure 6** reveals four levels, with very little now to say specifically about the best way to structure the conceptual levels except to claim that semantic fields and notional categories would be conceived here: animals, furniture, sports items, foods, clothing, and on and on. Each of these would have sets of features, which are referred to as 'semantic features.' Our figure implies that we are treating the conceptual category of 'animal,' whatever that might turn out to be. Evidently, here we have a subset of features that would separate out the human animal – an animal nevertheless. At the lexical level, we can clearly see that the target (to be read aloud, repeated, or named, for example) is 'cat.' Heaviness of print is meant to relate to heaviness of weights – that would produce 'cat' in its semantic sense. Garrett would say that this word comes at the first lexical lookup. The system's weight values set by the builders themselves are first set for the normal case. Lookup 1 worked. This level then spreads downward through the heaviest weighted connection paths, which activate the three phonemes /k/ /ae/, and /t/. Then notice that there are feedback routes that interact with the word level, which itself has two way connectivity with the defining features. The architecture further shows that 'rat' is related to 'cat' by virtue of its shared semantic features, and it clearly shows another word 'mat' that only shares phonemic features – the so-called rime /-aet/. Then, we have the word 'dog,' which shares nothing phonologically, but surely shares some of the semantic features that 'cat' has – though not all, obviously. The architecture therefore 'knows' all the above facts on 'similarity' – either semantic or phonological, and the architecture can be sculpted in order to even favor words like 'rat,' since it associates with 'cat' both semantically and phonologically. Garrett called this the 'lexical effect,' which accrues when LL 1 and 2 interact. Dell works with interaction activation and Garrett's serial model can at least 'talk about' this in terms of the two lexical selections. Within reason, the two architectures can still talk to one another, although we know that the connections are far ahead of the game since they can work the interactions of a mathematical continuum, which squares better with continua of frequency of occurrence, how similar or not the items associated are and ultimately manipulating one and the same system to simulate health or sickness, as we have often shown how this all exists on a health-disease continuum for human language. Again, Dell, et al. (1997) spell out this *sine qua non*, as did Charcot, Hughlings-Jackson, and Sigmund Freud (Buckingham (1999). The beauty of this model is that it is at once localistic and holistic. The massive interaction is qualitatively distinct from its ability to reveal what is going on and where. The system can then be fixed in different ways to answer both theoretical questions as well as more practical clinical questions.

It was this feature that allowed Dell and Martin to be able to dissociate anticipatory versus carryover transpositions. They took the 'continuity thesis' and asked, if normal slips-of-the-tongue are for the most part anticipatory in nature, can we see this dissociation unfold in our model by altering it in ways we predict might be appropriate. Leaving the weights alone and allowing each activated element to return to its 'resting' level, i.e., the normal case, should if we run it too fast, or lower some weight values at distributed locations, or use low-frequency words, perhaps we could get the system to produce more anticipatory errors. If on the other hand, if we weaken the connections from word selection #1 to word selection #2, and also changes the rates at which activated items return to where they came from, we might see more carryover orderings. We do, and in so doing Dell has captured an anticipatory/perseverative ratio that will favor perseverations the more severe the lesioning is to select a word for reading, for repetition or more importantly, for naming. Further and simulating recovery processing, the Dell model can show, by slowly raising the weight values, the perseveration errors tend to wane, as the 'anticipatory ratio' increases. To the extent (but recall the Eaton et al. study) we see improvement run at least in tandem with the reduction of perseveration. This meets the dictates of the continuity thesis, since we are observing the subject approach normality, since the normal slip-of-the-tongue data reveal more anticipatory errors than perseverative. It can serve as a kind of clinical metric for charting the evolution of severe phonemic paraphasia or even the evolution of neologistic jargonaphasia. The evidence we have so far for the success of continuity thesis predictions will now force us to seriously consider findings such as those of Eaton et al. (2011).

Dell's model directly incorporates much of the localistic node symbology, which is later called up first for 'lesioning' the system and then for being able focus upon the exact region and elements that are causal, in that the weight alterations importantly cause the error – or so it can be made to seem. Although only focusing upon Hebbian Learning connectionist models as those of Paul Smolensky (1988), Ramsey et al. (1991) argue that no symbolic sense can be made in terms of what in that system play any causal role at all in the output forms, since the internal processing between input and output are opaque and chaotic. Correct outputs can be obtained although one cannot 'get at' the locations of causes for this or that error in an output, because these models are complexly distributed through the system, more on a 'mass action' analogy that is redundantly equipotential. This architecture for memory systems is not unlike what Lashley (1951) had in mind. The reader of this article would do well to read the Ramsey et al. (1991), which argues that the 'nonsymbolic' nature inherently structured in such self-organizing dynamic systems. There is a full historical account of this exciting era in Bechtel and Abrahamsen (1991), which according to some is, "… a less technical and more prosy introduction," Goldsmith (1993, p. 19, ftn. 5). Suffice it to say here that neither Harley nor MacAndrews's nor Dell's models suffer from these deeply recondite and nonsymbolic infelicities.

And, since their systems do not learn on their own, the weights must set by the architects of the architecture itself.

For Jerry Fodor and Zenon Pylyshyn (1988) and Stephen Stich (Ramsey et al., 1991) the symbolic and localistic models like Dell's at least stand a chance with challenges from symbolic representation theories of the mind and what is often called folk psychology,' folksy, since it is almost parochial to insist on theories of language that have the cognitive labels of human language, with elements that can be recombined in different ways and upon models where we can glimpse what is doing what, causing what, and where it takes place in the architecture. Needless to say, the proposals, the criticisms, and the counter-arguments are legion and not without deep mathematical, logical, philosophical, computational and linguistic complexities, and puzzles. The literature on this is not for the weak at heart. Our final assessment here is that the Dell model and others such as those of Trevor Harley have contributed so much to theory and to its clinical utility in the language and speech sciences they deserve to be further pushed for what is has said and what it can predict. We would also point out that the transparency between the processing architecture of Dell and Garrett is such that we have been able to transform from the serial to the connectionist in some meaningful way, even though connectionism in general has replaced the serial.

Brief Summary with Some More Anatomy

We want to make one thing very clear in our article on phonological disorders. Not only do functional models, psychological or linguistic, determine the underpinnings of how we describe and explain these disorders, but it has now become clear that anatomical modeling adds to the mix. We have seen that connectionist models can uniquely and localistically 'lesion' weights between the semantic levels and the lexical levels of their model, thereby producing many types of whole-word errors, but the alterations at this level only affect whole-word level phenomenon; they do not produce phonemic paraphasia. The 'anomias' are located here, presumably. The connectionist models can do the same, but within the level that maps fully selected, whole words onto their phonological forms. 'Lesions' here will neither simulate anomias nor semantic whole-word substitutions; they will simulate phonological disorders in terms of scrambling or otherwise derailing the phonemic structure of words, and that weakened weight 'lesioning' causes scrambling and derailing within the mapping between the lexical level and the phonemic (sublexical) level. This traditional way of thinking challenges the so-called 'anomia' theory of neology (Buckingham, 1977), Butterworth (1979), Buckingham (1981), and Buckingham and Christman (2008), and appears to allow it no possible second role in explanations of neologistic jargon aphasia. That in turn infiltrates how we view the very nature of neologistic production, and thus it colors our views of a great deal of phonological disturbances.

Furthermore and very importantly, there are anatomical accounts of brain regions and their connections that also, and often as a consequence of, duplicate many connectionist accounts of neology, anomia, semantic, and phonemic paraphasias. In Buckingham and Christman (2008), we assessed the model proposed and utilized in Hillis et al. (1999) for jargon aphasia. The single patient used was an adult with conduction aphasia whose principal disruption was phonological activation, which of course implicates the phonological system as the location of the breakdown. The subject could read passively, match words with pictures, understand and speak with fluency, whereby the authors assume that full connectivity between the semantic and the lexical levels remained uncompromised. They then proceeded to 'lesion' their connectionist system between the lexical and the phonemic level of representation (the 'sublexical level'). They left the decay rate weights alone and did not alter any weights in the semantic to lexical mapping. Their simulations indeed showed phonemic paraphasias and neologisms on oral naming, repetition, and oral reading testing. Anatomically, the patient had brain damage that interfered with the white fiber tracts mediating information exchange between the posterior, superior temporal gyrus in *dorsal* BA area 22 and the more *ventral* regions of BA 22. Crucially, then they propose that dorsal 22 is a site of the phonological processing of words (already correctly selected) and that ventral 22 is 'likely' to be an area for the representation and processing of whole-word forms. It therefore is evident that this study supports 'the conduction theory' of neology, eschewing 'the anomia theory.' All connectionist studies, according to these authors, have simulated neology lesioning the lexical to sublexical levels and that some have globally lesioned their systems and again observed neology. We are not told if those studies saw any difference in degree of severity of the segmental paraphasias that were nonword and unrecognizable. We also do not find out if there was any evidence with the global lesioning simulations. Hillis et al., however, globally lesioned their connectionist system but, alas, could not simulate neology. Consequently, their schema would not allow for or include an 'anomia theory' of neology. Currently, however, there has been increased appreciation of the more ventral regions in the left temporal lobe, in addition to the more classical position for phonological processing in the dorsal regions of Wernicke's area, the Ptp and the supramarginal gyrus (Area 40), processing whole word lexical/semantic functioning the line of recent work on not only classical aphasiology's dorsal system, but now has recognized lexical/semantic processing in the ventral regions in addition to the more phonological processing in dorsal regions.

At the present, the technology of magnetic resonance imaging has led to the even more interesting and revealing diffusion tensor imaging, whereby the biochemical structure of water and its hydrogen atoms react under magnetic conditions to 'line up,' and since the imaging is on fiber tracts and not cortex, the lined up structure highlights, often quite beautifully the actual fiber tract. This technology has opened up impressive studies of already known fiber tracts, plus much that we have not been able to observe. We've seen new tracts, and extensions of tracts, and coursings not yet completely understood.

The focus on ventral function actually started with Hickok and Poeppel (e.g., 2004, 2007) and now with DTI has progressed further with the ventral tract and its extensions from occipital zones (the inferior fronto-occipital fasciculus, lower down and coursing through the external capsule and the extreme capsule, terminating in BA 45 (the pars triangularis).

This anatomical work often brings out functional connectivity differences between BA 44 and 45, often simply considered as a whole as Broca's area; others, though are more careful to constrain the IFG to 44. In even more recent work, the excitement continues with studies such as Friederici (2009) and even more recent ones by Brauer et al. (2013). The point to all this is that, as we saw with the Hillis et al. study, these recent investigations of the ventral regions in the left hemisphere could also color our descriptions of phonological disruptions and systems that might be able to compensate or help ameliorate phonological disorders. These ventral regions have also played a role at the semantic level, as have other ventral areas in posterior middle temporal lobe on the left at the location of the fusiform gyrus that connects with anterior occipital regions. Lesions in this region often gave rise to the production of semantic, whole-word substitutions (either paradigmatically or syntagmatically related in Jakobson's terms, which of course originated in Aristotelian theories of association). There were no sublexical errors found in these patients.

In the end, and as we proposed at the outset of our article, suffice it to say that as neurolinguistic modeling goes, so goes the way we ultimately understand the complex nature of phonological disorders. Here, we would suggest that more work needs to be done on the fluent paraphasias of patients who sustain somewhat larger brain lesions that encompass both ventral and dorsal in the posterior regions of the dominant hemisphere. The paraphasic composition of this group of patients needs close scrutiny.

Future Directions

Most obvious to watch for in future studies of phonological disorders will be further connectionist modeling with close attention to neuroanatomical innovations from imaging technology. We also look for further work in the phonology from OT (McCarthy, 2004) and harmonic phonology (Goldsmith, 1993). These two programs need some disentanglement (see McCarthy, 2004, p. 6), however, since both seem to harmonize in distinct ways. Presumably, OT is being worked out as an exercise in setting up a competence of sorts for Smolensky's connectionism, as we briefly discussed earlier in this article. We are not aware as yet of much language simulation of normality and pathology from these subsymbolic, Hebbian Learning models. Goldsmith's (1993) harmonic phonology has accomplished the simulation of sonority constraints and Turkish vowel-harmony, but certainly not with a subsymbolic connectionism. However, we are not aware of any work in aphasia by him, and only brief connectionist work by Wheeler and Touretzky (e.g., (1997), which we discussed earlier in the article, where they simulated much of the findings on phonemic paraphasia and neology of the work by Buckingham (1986, 1987, 1989, 1991). OT carves out much more complexity in phonotactic constraints. For example, in mapping out sonority constraints, OT would claim that sonority rests not on a general hierarchical degree of "level-true "goodness," such as the focus on overall statements such as that OV's (Obstruent-Vowel) are preferred to NV's (Nasal-Vowel), which are preferred to LV's (Liquid-Vowel), which are preferred to GV's (Glide-Vowel). What has heretofore been left out of the equation is the fact that sonority can only make real sense as it combines with the variant analyses of how the segmental arrays or syllable constituents remain 'faithful' to the extent that their underlying input forms match precisely their output forms. It is no longer strikingly crucial to simply claim that the English word 'quite' abides less by a goodness of sonority fit than the English word 'cry,' since the segment order /kw+V/ is less preferred than the /kr+V/. Otherwise we would be obliged to claim that /kway/ would be a more preferred pronunciation for 'cry.' OT claims that this type of phonotactic statement is 'parochial' (McCarthy, 2004, p. 6). Other phonologists such as Diana Archangeli have called it 'trite.' Repair strategies show this. What if, for whatever reason, the word 'quiet' is uttered as /kayət/? If the repair strategy were based on sonority preference, why would the repair to correctly utter 'quiet' select a 'less preferred' sonority choice kw+V over k+V? To match input to output for 'quiet' one would have to select the less-preferred sonority ranking segment string. In any event, we await future studies in aphasia simulation from systems such as Goldsmith's or Smolensky's. To date, however, it seems that the Dell model has the big lead in the study of paraphasia in brain damaged adults with fluent phonological disorders.

See also: Aphasia; Bayesian Models in Neuroscience; Brain Asymmetry; Cerebral Cortex; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Cognitive Neuroscience; Computational Neuroscience; Functional Brain Imaging of Language processes; Intentionality in Language and Communication, Emergence of; Language and Thought: The Neo-Whorfian Hypothesis; Motor Cortex; Neurolinguistics, History of; Parietal Lobe; Prefrontal Cortex Development and Development of Cognitive Function; Prefrontal Cortex; Prototype Theory in Linguistics; Psycholinguistics: Overview; Serial Order, Representation of; Speech Production, Neural Basis of; Temporal Lobe.

Bibliography

Bastian, H.C., 1869. On the various forms of loss of speech in cerebral disease. British and Foreign Medical and Chirurgical Review 43, 209–236, 470–492.
Bateman, F., 1890. On Aphasia or Loss of Speech and the Localisation of the Faculty of Articulate Language. Cornell University Library Print Collection, Ithaca, NY.
Bay, E., 1969. The Lordat case and its import on the theory of aphasia. Cortex 5, 302–308.
Bechtel, W., Abrahamsen, A., 1991. Connectionism and the Mind. Blackwell, Oxford, UK.
Benton, A., 2000. Exploring the History of Neuropsychology: Johann A. P. Gesner on Aphasia. Oxford University Press, Oxford, UK, pp. 175–182.
Blumstein, S.E., 1973. A Phonological Investigation of Aphasic Speech. Mouton, The Hague, NL.
Bouillaud, J.-B., 1825. Recherches Clinique propres a demontrer que la perte de la parole correspond a la lesion des lobules anterieurs du cerveau. Archives Generale de la medicine 8, 25–45.
Brauer, J., Anwander, A., Perani, D., Friederici, A.D., 2013. Dorsal and Ventral Pathways in Language Development. Brain and Language 127 (2), 289–295. http://dx.doi.org/10.1016/j.bandl (01.03.13.).
Broca, P., 1861. Pert de la parole, ramollissement chronique et destruction partielle du lobe anterieur guache du cerveau. Bulletin Societe d'Anthropologique 2, 235–238.
Broca, P., 1865. Sur le siege de la faculte du langage articule. Bulletin Societe d'Anthropologique 6, 377–393.
Buckingham, H.W., 1977. The conduction theory and neologistic jargon. Language and Speech 4, 174–184.

Buckingham, H.W., 1981. Where do neologisms come from? In: Brown, J.W. (Ed.), Jargonaphasia. Academic Press, New York, NY, pp. 39–62.
Buckingham, H.W., 1986. The scan-copier mechanism and the positional level of language production: evidence from phonemic paraphasia. Cognitive Science 10, 195–217.
Buckingham, 1987. Review: phonemic paraphasias and psycholinguistic production models for neologistic jargon. Aphasiology 1, 381–400.
Buckingham, H.W., 1989. Principle of sonority, doublet creation, and the checkoff monitor. In: Nespoulous, Villiard (Eds.). Op. Cit, pp. 193–212.
Buckingham, H.W., 1995. En torno al copiador-explorador y la metáfora visual cognoscitiva. Neuropsychologia Latina 1, 34–43.
Buckingham, H.W., 1999. Freud's continuity thesis. Brain and Language 69, 76–92.
Buckingham, H.W., 2002. The roots and amalgams of connectionism. In: Daniloff, R.G. (Ed.), Connectionist Approaches to Clinical Problems in Speech and Language. Lawrence Erlbaum, Mahwah, NJ, pp. 265–311.
Buckingham, H.W., 2006. Was Sigmund Freud the first neogrammarian neurolinguist? In: Code, C. (Ed.), The Representation of Language in the Brain: In Honour of John C. Marshall. The Psychology Press, New York, pp. 1085–1104.
Buckingham, H.W., Christman, S.S., 2004. Phonemic carryover perseveration: word blends. Seminars in Speech & Language 25, 363–373.
Buckingham, H.W., 2010. Commentary: Aristotle's functional association psychology. The syntagmatic and the paradigmatic axes in the neurolinguistics of Roman Jakobson and Alexander Luria: an anatomical and functional quagmire. Aphasiology 24, 395–403.
Buckingham, H.W., Christman, S.S., 2007. Independent evidence for the unification of explanatory paradigms in the neurosciences. In: Ball, M.J., Damico, J.S. (Eds.), Clinical Aphasiology: Future Directions. A Festschrift for Chris Code. Psychology Press, Hove, East Sussex, UK, pp. 28–41.
Buckingham, H.W., Christman, S.S., 2008. Disorders of phonetics and phonology. In: Stemmer, B., Whitaker, H.A. (Eds.), Handbook of the Neuroscience of Language. Elsevier, London, pp. 127–136.
Buckingham, Hugh W., 1991. The psycholinguistics of microgenesis: The nature of paraphasia. In: Hanlon, R. (Ed.), Microgenetic Approaches to Cognition: A Neuropschological Perspective. Springer-Verlag, New York, pp. 150–179.
Butterworth, B., 1979. Hesitation and the production of verbal paraphasias and neologisms in jargon aphasia. Brain and Language 8, 133–161.
Caplan, D., 1987. Neurolinguistics and Linguistic Aphasiology. Cambridge University Press, Cambridge, UK.
de Courtenay, J.B., 1885–1886/1972. On pathology and embryology of language. In: Stankiewicz, E. (Ed. and Trans.), A Baudouin de Courtenay anthology. Indiana University Press, Bloomington, IN.
Dell, G.S., 1986. A spreading-activation theory of retrieval in sentence production. Psychological Review 93, 283–321.
Dell, G.S., Burger, L.K., Svec, W.R., 1997a. Language production and serial order: a functional analysis and a model. Psychological Review 104, 123–147.
Dell, G.S., Schwartz, M.F., Martin, N., Saffran, E.M., Gagnon, D.A., 1997b. Lexical assess in aphasic and non-aphasic speakers. Psychological Review 104, 801–838.
Duffy, J., 2013. Motor Speech Disorders: Substrates, Differential Diagnosis, and Management, third ed. Elsevier Mosby, St. Louis, MO. Copyright 2013, 2005, 1995 by The Mayo Foundation for Medical Education and Research.
Eaton, E., Marshall, J., Pring, T., 2011. Mechanisms of change in the evolution of jargon aphasia. Aphasiology 25, 1543–1561.
Eling, P., 2006. The psycholinguistic approach to aphasia of Chajim Steinthal. In: Code, C. (Ed.), The Representation of Language in the Brain: In Honour of John C. Marshall. Psychology Press, New York, pp. 1072–1084. Also a special issue of Aphasiology. 20, nos. 9, 10, 11, 2006.
Fillmore, C.J., 1968. The case for case. In: Bach, E., Harms, R.T. (Eds.), Universals in Linguistic Theory. Holt, Rinehart & Winston, Inc., New York, NY, pp. 1–88.
Fodor, J.A., Pylyshyn, Z.W., 1988. Connectionism and cognitive architecture: a critical analysis. In: Pinker, S., Mahler, J. (Eds.), Connections and Symbols. Elsevier Science Publishers, Amsterdam, NL, pp. 3–71. First MIT Press edition, 1988. Reprinted from Cognition: International Journal of Cognitive Science 28.
Freud, S., 1891/1953. On Aphasia: A Critical Study (E. Stengel, Trans. and Introduction). Imago, London.
Friederici, A.D., 2009. Pathways to language: fiber tracts in the human brain. Trends in Cognitive Sciences 13, 175–181.
Fritsch, G.T., Hitzig, E., 1870/1960. On the electrical excitability of the cerebrum. In: Von Bonin, G., Von, G. (Ed. and Trans.), 1960, Some Papers on the Cerebral Cortex. Charles C. Thomas, Springfield, IL.
Fromkin, V.A. (Ed.), 1973. Speech Errors as Linguistic Evidence. Mouton, The Hague, NL.
Fromkin, V.A. (Ed.), 1980. Errors in Linguistic Performance. Slips of the tongue, ear, pen, and hand Academic Press, New York.
Garrett, M.F., 1982. Production of speech: observations from normal and pathological language use. In: Ellis, A.W. (Ed.), Normality and Pathology in Cognitive Functions. Academic Press, London, pp. 19–76.
Garrett, M.F., 1984. The organization of processing structure for language production: applications to aphasic speech. In: Caplan, D., Lecours, A.R., Smith, A. (Eds.), Biological Perspectives on Language. MIT Press, Cambridge, MA, pp. 172–193.
Goldsmith, J. (Ed.), 1993. The Last Phonological Rule: Reflections on Constraints and Derivations. University of Chicago Press, Chicago.
Green, E., 1969. Phonological and grammatical aspects of jargon in an aphasic patient. Language & Speech 12, 103–118.
Hartley, D., 1746/1959. Various Conjectures on the Perception, Motion, and Generation of Ideas (R.E.A. Palmer, Trans. from the Latin and introduction with notes by M. Kallich). The Augustan Reprint Society, Los Angeles, CA. Publication no. 77–78. University of California, LA, William A. Clark Memorial Library, Los Angeles.
Hartley, D., 1749. Observations on Man, His Frame, His Duties and His Expectations. S. Richardson, London.
Heilman, K., 2014. In press. Apraxia. Continuum.
Harley, T.A., MacAndrew, S.B.G., 1992. Modeling paraphasias in normal and aphasic speech. In: Proceedings of the 14th Annual Conference of the Cognitive Science Society, pp. 378–383. Bloomington, IN.
Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92, 67–99.
Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nature Reviews 8, 393–402.
Hillis, A.E., Boatman, D., Hart, J., Gordon, B., 1999. Making sense out of jargon: a neurolinguistic and computational account of jargon aphasia. Neurology 53, 1813–1824.
Hoffman, P., Jefferies, E., Lambon-Ralph, M.A., 2011. Remembering 'zeal' but not 'thing': reverse frequency effects as a consequence of deregulated semantic processing. Neuropsychologia 49, 580–584.
Hughlings-Jackson, J., 1879–1880. On affections of speech from disease of the brain (III). Brain 2, 325–356.
Jakobson, R., 1964. Towards a linguistic typology of aphasia impairment. In: De Reuck, A.V.S., O'Connor, M. (Eds.), Disorders of Language. Little Brown, Boston.
Kertesz, A., Benson, F., 1970. Neologistic jargon: a clinicopathological study. Cortex 6, 362–386.
Kohn, S.E., Smith, K.L., 1994. Evolution of impaired access to the phonological lexicon. Journal of Neurolinguistics 8, 267–288.
Kohn, S.E., Smith, K.L., Alexander, M.P., 1996. Differential recovery from impairment to the phonological lexicon. Brain and Language 52, 129–149.
Lamb, S.M., 1966. Outline of Stratificational Grammar. Georgetown University Press, Washington, DC.
LaPointe, L.L., 2013. Paul Broca and the Origins of Language in the Brain. Plural Publishing Inc., San Diego, CA.
Lashley, K.S., 1951. The problem of serial order in behavior. In: Jeffress, L.A. (Ed.), Cerebral Mechanisms in Behavior: The Hixon Symposium. Wiley, New York, pp. 112–146.
Lecours, A.R., Lhermitte, F., 1972. Recherches sur le langage des aphasiques: 4. Analyse d'un corpus de neologismes; notion de paraphasie monemique. Encephale 61, 295–315.
Lecours, A.R., Caplan, D., 1975. Blumstein's phonological investigation of aphasic speech. Brain & Language 2, 237–254.
Lecours, A.R., Nespoulous, J.-L., Pioger, D., 1987. Jacques Lordat or the birth of cognitive neuropsychology. In: Keller, E., Gopnik, M. (Eds.), Motor and Sensory Processes of Language. Lawrence Erlbaum, Hillsdale, NJ, pp. 1–16.
Levelt, W.J.M., 2008. An Introduction to the Theory of Formal Languages and Automata. John Benjamins, Amsterdam/Philadelphia.
Liepmann, H., 1900/1977. The syndrome of apraxia (motor asymboly) based on a case of unilateral apraxia (A translation from Monatschrift für Psychiatrie und Neurologie 1900 8, 15-44.). In: Rottenberg, D., Hockberg, F.H. (Eds.), Neurological Classics in Modern Translation. Macmillan, New York.
Lockwood, D.G., 1972. Introduction to Stratificational Linguistics. Harcourt, Brace Jovanovich, New York.
Lordat, J., 1842–1843. Lecons tirees de cours de physiologie de l'annee scolaire 1842–1843: Analyse de la parole pour server a la theorie de divers cas d'alalie et de paralalie que les nosologistes ont mal conus. Journal de la Societe de medecine pratique de Montpellier 7, 333–353, 417–433, and 8, 1–17. (See full bibliographical record for this publication in Lecours, Nespoulous, and Pioger, Op. cit, p. 16).
Luria, A.R., 1970. Traumatic Aphasia. Mouton, The Hague, NL.
Luria, A.R., 1973. The Working Brain: An Introduction to Neuropsychology. Basic Books, New York.

MacKay, D.G., 1973. Spoonerisms: the structure of errors in the serial order of speech. In: Fromkin, V.A. (Ed.), Speech Errors as Linguistic Evidence. Mouton, The Hague, NL, pp. 164–194.

MacNeilage, P.F., 1998. Evolution of the mechanisms of language output: comparative neurobiology of vocal and manual communication. In: Hurford, J., Studdert-Kennedy, M., Knight, C. (Eds.), Approaches to the Evolution of Language. Cambridge University Press, Cambridge, UK, pp. 222–241.

Marshall, J., 2006. Jargon aphasia: what have we learned? Aphasiology 20, 387–410.

Marshall, J., Pring, T., Chiat, S., Robson, J., 2001. When ottoman is easier than chair: an inverse frequency effect in jargon aphasia. Cortex 37, 33–53.

Martin, N., 2000. An interactive activation account of aphasic speech errors: converging influences of locus, type and severity of impairment. In: Hartsuiker, R.J., Bastiaanse, R., Postma, A. (Eds.), Phonological Encoding and Monitoring in Normal and Pathological Speech. Psychology Press, New York, pp. 67–85.

Martin, N., Dell, G.S., 2004. Perseverations and anticipations in aphasia: primed intrusions from the past and future. Seminars in Speech and Language 25, 349–362.

Martin, N., Dell, G.S., 2007. Common mechanisms underlying perseverative and non-perseverative sound and word substitutions. Aphasiology 21, 1002–1017. Special issue on Verbal Perseveration. In: Stark, J.A. (Ed.), Aphasiology. vol. 21, nos. 10/11, Oct/Nov 2007.

McCarthy, J.J. (Ed.), 2004. Optimality Theory in Phonology: A Reader. Blackwell Publishing, Malden, MA.

Meyer, A.S., 1990. The time course of phonological encoding in language production: the encoding of successive syllables of a word. Journal of Memory and Language 29, 524–545.

Meyer, A.S., 1991. The time course of phonological encoding in language production: phonological encoding inside a syllable. Journal of Memory and Language 30, 69–89.

Meyer, A.S., 2000. Form representations in word production. In: Weldon, R.S. (Ed.), Aspects of Language Production. Psychology Press, Hove, East Sussex, UK, pp. 49–70.

Nespoulous, J.-L., Villiard, P. (Eds.), 1990. Morphology, Phonology, and Aphasia. Springer-Verlag, New York, pp. 193–205.

Nespoulous, J.-L., June 2013. Jacques Lordat (1773–1870) De Tournay (Hautes-Pyrenees) a Montpellier: la naissance de la 'neuropsychologie cognitive du langage.'. Paper presented to the l'Academie de Montpellier 3.

Nichols, L., Howard, D., 2000. When the words won't come: relating impairments and models of spoken word production. In: Weldon, L.R. (Ed.), Aspects of Language Production. Psychology Press, Hove, East Sussex, UK, pp. 115–142.

Ohala, J., 1990. There is no interface between phonology and phonetics: a personal view. Journal of Phonetics 18, 153–171.

Pavlov, I., 1962. I. P. Pavlov: Psychopathology and Psychiatry: Selected Works (D. Myshne, S. Belsky, Trans. from the Russian. Compiled by Y. Popov, L. Rokhlin). Foreign Languages Publishing House, Moscow.

Pick, A., 1931/1973. Handbuch der normalen und pathologischen physiologie, vol. 15. Springer-Verlag, Heidelberg, Germany (J.W. Brown, Trans. 1973. Aphasia. Charles C. Thomas, Springfield, IL).

Prince, A., Smolensky, P., 2004. Optimality theory: constraint interaction in generative grammar. In: McCarthy, J.J. (Ed.), Optimality Theory in Phonology: A Reader. Blackwell Publishing, Malden, MA, pp. 3–71.

Pylyshyn, Z.W., 1980. Cognition and computation: Issues in the foundations of cognitive science. Behavioral and Brain Sciences 3, 154–169.

Ramsey, W., Stich, S.P., Garon, J., 1991. Connectionism, eliminativism, and the future of folk psychology. In: Ramsey, W., Stich, S.P., Rumelhart, D.E. (Eds.), Philosophy and Connectionist Theory. Lawrence Erlbaum, Hillsdale, NJ, pp. 199–228.

Robbins, R.H., 1967. A Short History of Linguistics. Indiana University Press, Bloomington, IN.

Sapir, E., 1925. Sound patterns in language. Language 1, 35–51.

Sapir, E., 1933/1949. The psychological reality of the phoneme. In: Mandelbaum, D.G. (Ed.), Selected Writings of Edward Sapir in Language, Culture and Personality. University of California Press, Berkeley, CA. Originally published in 1933 in French as: La realite psychologique des phonemes. Journal de psychologie normale et pathologique 30, 247–265.

Schiller, F., 1994. Paul Broca: Founder of French Anthropology, Explorer of the Brain, second ed. University of California Press, Berkeley, CA.

Shattuck-Hufnagel, S., 1979. Speech errors as evidence for a serial order mechanism in sentence production. In: Cooper, W.E., Walker, E.C.T. (Eds.), Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett. Lawrence Erlbaum, Hillsdale, NJ, pp. 295–342.

Shattuck-Hufnagel, S., 1987. The role of word-onset consonants in speech production planning: new evidence from speech error patterns. In: Keller, E., Gopnik, M. (Eds.), Motor and Sensory Processes of Language. Lawrence Erlbaum, Hillsdale, NJ, pp. 17–51.

Smolensky, P., 1988. On the proper treatment of connectionism. The Behavioral and Brain Sciences 11, 1–74.

Starr, M.A., 1889. The pathology of sensory aphasia, with an analysis of fifty cases in which Broca's centre was not diseased. Brain 12, 82–101.

Trubetzkoy, N.S., 1939/1969. Grundzuge der phonologie. In: Travaux de Cercle Linguistique de Prague, vol. 7. University of California Press, Berkeley, CA. Trans. into English 1969 as Trubetzkoy, Principles of Phonology.

Weldon, L.R. (Ed.), 2000. Aspects of Language Production. Psychology Press, Hove, East Sussex, UK.

Wernicke, K., 1874. The symptom complex of Aphasia. Cohn and Weigert, Breslau. In: Eggert, G.H. (Ed.), 1977. Wernicke's Works on Aphasia: A Sourcebook and Review. Mouton, New York.

Wheeler, D.W., Touretzky, D.S., 1997. A parallel licensing model of normal slips and phonemic paraphasias. Brain and Language 59, 147–201.

Whitaker, H.A., 1972. Brain damage and phonological organization. In: Rigault, A., Charbonneau, R. (Eds.), Proceedings of the Seventh International Congress of Phonetic Sciences. Mouton, The Hague, NL, pp. 431–436.

Whitaker, H.A. (Ed.), 1988. Phonological Processes and Brain Mechanisms. Springer-Verlag, New York.

Wickelgren, W.A., 1969. Context-sensitive coding, associative memory, and serial ordering in (speech) behavior. Psychological Review 76, 1–15.

Phonetics, Articulatory

Xiaonong Zhu, The Hong Kong University of Science and Technology, Hong Kong, People's Republic of China

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P.A. Keating, volume 17, pp. 11381–11386, © 2001, Elsevier Ltd.

Abstract

Articulatory phonetics is concerned with consonants and vowels (segments) as well as tone and phonation types (phonationals). For more than a century, it has been deemed as an applied discipline for practical purposes such as language teaching and description. In recent years, articulatory or linguistic phonetics has developed quickly and general theories are attempted to organize various concepts into a consistent system to account for phonetic phenomena. This article introduces a syllable-centric phonetic theory, which treats tone and phonation types as nonlinear syllabic constituents and consonants and vowels as linear constituents. Phonetic typology is also explored in the article.

Introduction

Phonetics, in its earliest sense, means articulatory phonetics. Although it has now developed into a greater 'phonetic science,' containing dozens of interdisciplines, its base or core part is still articulatory phonetics.

'Articulation' has two meanings. In the narrower sense, it refers to the movements of articulators in the oral tract that produce 'segments' (vowels and consonants). In the broader sense, it also includes 'phonation,' the laryngeal or vocal fold activities, to produce 'suprasegmentals' (cf Suprasegmentals in this volume) or 'phonationals' (register, pitch, and length). Since segments are the main subject to study in traditional phonetics and suprasegmentals are merely secondary features, the two meanings do not create much ambiguity.

In this article 'articulation' (and 'articulator' as well) is used in the narrower sense, as opposed to 'phonation.' However, 'articulatory phonetics' is used in the broader sense, covering the study of both articulation and phonation. It is concerned with the production of speech sounds (cf Speech Production, Psychology of and Speech Production, Neural Basis of): how the vocal cords modify the airstream, when it passes through, to bring about different 'voice qualities' and how the articulators further modify the airstream coming out of the glottis to form different 'sound qualities.'

The interest in the exploration of speech sounds started as early as two thousand years ago in ancient India and ancient Greece, and later in ancient Arab. In ancient China, phonetics emerged in the first century CE (*fanqie* spelling), continued progressing in the sixth century (rhyming books), and flourished around the tenth century (classification of consonants and rhyming tables). Modern phonetics emerged in the late nineteenth century marked with a series of academic events, one of which was the design and promulgation of a set of phonetic symbols, the International Phonetic Alphabet (IPA), by the newly established International Phonetics Association as a standardized representation of all the speech sounds in human languages.

For more than a century phonetics has been deemed as an applied discipline for such practical purposes as language learning and teaching, dictionary making, language/dialect description (cf Dialectology in this volume), speech pathology, singing and acting, translation, and spelling reforms (cf Writing Systems in this volume). Its main task is to describe and classify segments, with suprasegmentals as its subsidiary task.

Since the late 1970s articulatory phonetics has developed quickly in the studies of consonants, vowels, phonation, and general theories.

A great number of new consonants were found and identified, which resulted in a total of five revisions of the IPA chart in a quarter of a century (1978–2005). In the earlier half a century only one revision was made in 1952.

The nature of vowels has been reconsidered. Vowels were previously defined with respect to two physiological criteria: (1) they have a wide enough oral track for the outward airstream so that no noise would be generated, and (2) they are described along three physiological parameters: height (or mouth openness) and backness of the tongue position, and lip rounding. These criteria have been replaced by two new ones. First, the vowel space is defined by the acoustic dimensions of the first two formant frequencies. Second, the nature of the vowel is understood with reference to its phonological status in a syllable.

A new subject, phonation modes, attracted extensive studies first in physiology and acoustics, and later in linguistic phonetics. There are as many as two dozen phonation types in terms of physiology (Catford, 1977; Laver, 1980). In linguistic phonetics, however, only six were identified (Ladefoged and Maddieson, 1996). Recently, in the light of tonal typology and sound evolution, a phonological system of phonation types has been established (Zhu, 2011a), which contains 14 types, grouped into six major types. This finding has amended the classification of consonants, brought about a new tonal model, restructured the syllable, and eventually reshaped the overall phonetic framework.

The most important change happening in phonetics is the change in theoretical conception in recent years when 'linguistic phonetics' (Ladefoged, 1971) is becoming matured. In his linguistic phonetics Ladefoged (2006) conceptualized an overall theoretical framework, in which phonetics was absent. Zhu (2008, 2011a) attempted to refine the framework by setting the syllable as the basic and pivot unit. These theoretical studies have improved the understanding of phonetics.

The remaining part of this article will address some theoretical issues in phonetics before introducing criteria for segment

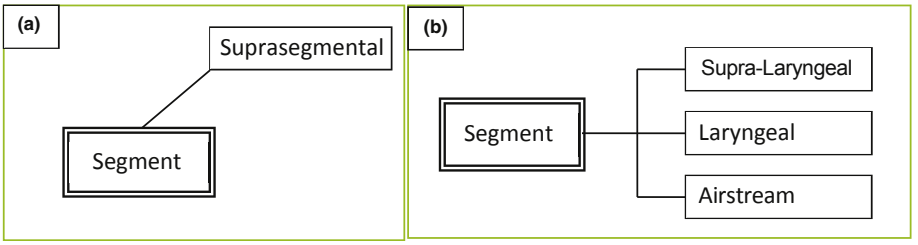

Figure 1 The segment-centric phonetic frameworks: a hypothetical traditional model (a) and Ladefoged's (2006) model (b).

description and classification, followed by a system of phonation types and a set of parameters for tonal representation.

Theoretical Considerations

Linguistic Phonetics Being Phonology

'Pure' phonetics was largely separated from linguistics long ago. Phonologists, from early years' structuralists to the present generativists, include only phonology in linguistics and regard phonetics as part of physics (e.g., Trubetzkoy, 1939; Anderson, 1981). As 'functional phonetics' (Martinet, 1949), "phonology begins where phonetics leaves off" (Sommerstein, 1977; cf Phonology in this volume). Enormous changes have taken place in the past few decades. Phonetics has now become such a large field that it contains dozens of interdisciplines. The phonetics that linguists pursue is what Ladefoged called 'linguistic phonetics,' an interdiscipline of phonetics and linguistics. It developed from articulatory phonetics and remains the core part of the greater phonetic science. As a phonetic discipline, it differs from other disciplines such as engineering phonetics, forensic phonetics, and pathological phonetics in that it is of linguistics, by linguistics, and for linguistics. When facing various linguistic disciplines, however, 'linguistic phonetics' is not an appropriate term because there are no such terms as 'linguistic syntax' or 'linguistic morphology.' In linguistics, 'linguistic phonetics' is just phonology, or the core part of phonology. That is why basic phonetic concepts such as the consonant and vowel should and can be defined with reference to their phonological roles in a syllable, in addition to their articulatory and acoustic criteria and why phonology should and can be conducted based on phonetics.

Basic Framework

There is no explicit general phonetic framework before Ladefoged's proposal. The traditional consensus in phonetics is that the segment is the core subject to study, with suprasegmentals (pitch and length) as secondary features, as schematically represented in Figure 1(a). Ladefoged's framework (Figure 1(b)) is like a diagrammatic combination of the traditional common view and the three processes of speech production (Catford, 1977): initiation (=airstream), phonation (=laryngeal), and articulation (=supralaryngeal).

The segment-centric view has been the mainstream view in phonetics. Yet there is another view held by British linguists (e.g., Abercrombie, 1967) which sets the syllable as the basic unit. This is also a millennia-long tradition in Chinese phonetics/phonology. Following this thread, a new linguistic phonetic framework has been developed (Figure 2).

A syllable has two kinds of components, the linear segments and the nonlinear phonationals. They can be phonologically organized to form linear onset and rime as well as nonlinear register, length, and pitch.

Segments

Segments include consonants and vowels, which, during production, involve (near-) closing and opening of the vocal tract. The most widely used symbols for notating segments are the IPA, which is presented in Figure 3.

Consonants

'Normal' consonants are produced with an egressive airstream. They are arranged in the pulmonic consonant table at the top of the IPA chart (Figure 3) along two dimensions: place and manner of articulation. Places of articulation, also called articulatory targets or passive articulators, refer to the locations in the upper and back areas in the vocal tract where active articulators (the lower lip or the tongue) approach to form various kinds of constrictions to impede airflow. Eleven places of articulation, e.g., labial, alveolar, etc., are established in the table. They are listed on the right side of Figure 4 except one, retroflex, which refers to a type of consonant involving an articulatory target somewhere between postalveolar and palatal and an active articulator of subapical or the tip of the tongue. Note that, according to Jakobson et al. (1952), the 'glottal' sounds

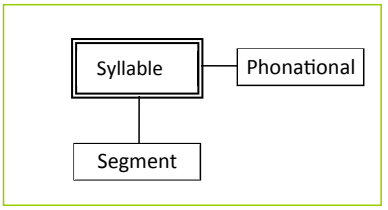

Figure 2 The syllable-centric phonetic framework. Reproduced from Zhu, Xiaonong, 2008. 音節和音節學 Syllable and syllabics. Oriental Linguistics 4, 142–164; Zhu, Xiaonong, 2011a. 語言語音學和音法學：理論新框架 Linguistic phonetics and pan-chronic phonology: a new theoretical framework. Studies in Language and Linguistics 31 (1), 4–85.

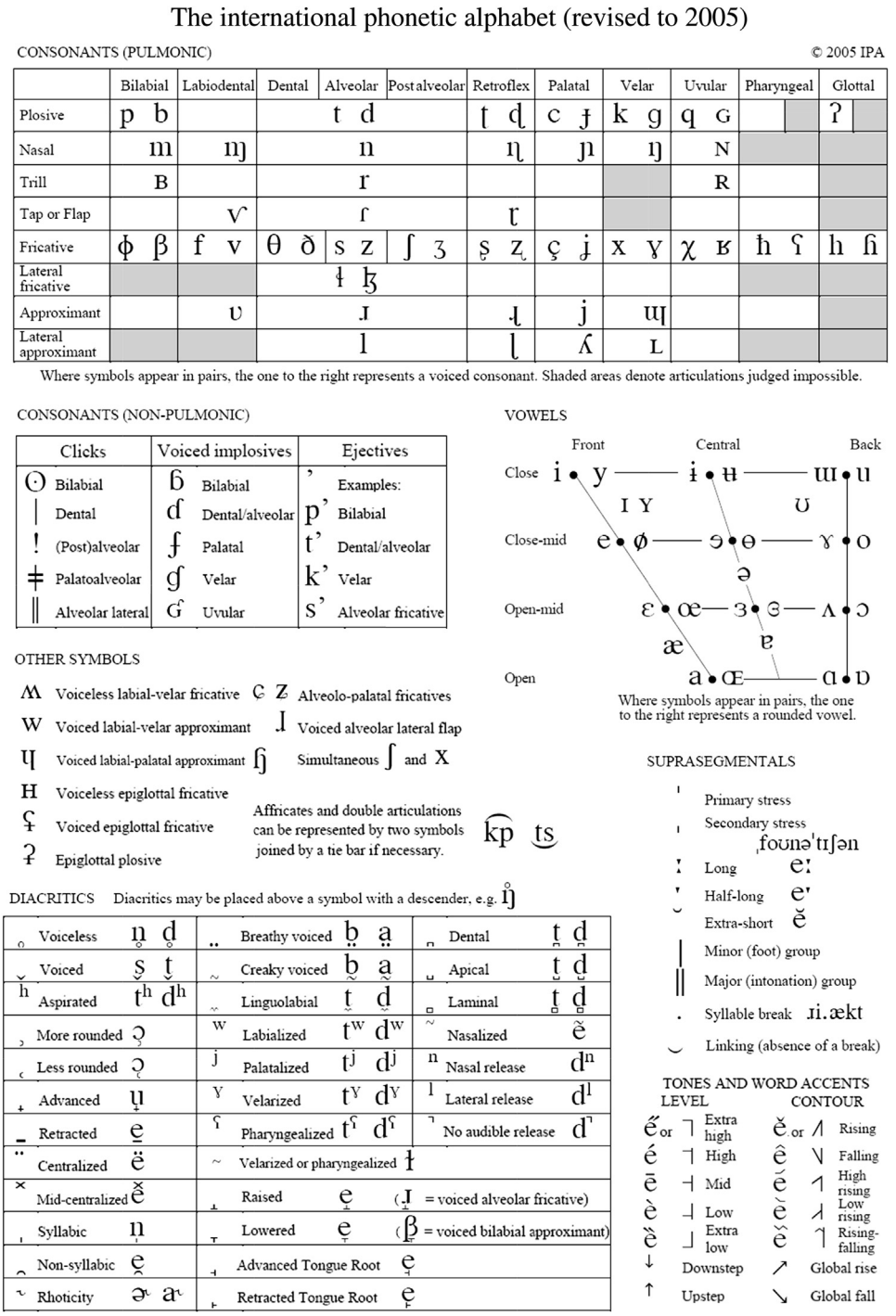

Figure 3 The IPA chart.

are neither consonants nor vowels. Rather, they are phonation types (cf section 'A System of Phonationals' in this article). Active articulators refer to the lower lip and various parts of the tongue, whose terms are given in the middle of the figure. Along the left side are hyphenated terms of active articulators, which, when combining with those for place, form terms of articulatory gesture, e.g., bilabial (labio-labial), labio-dental, apical-alveolar, lamino-dental, etc.

'Manner of articulation' refers to the methods of maneuvering the airstream during consonant production. The airstream can, when going through the vocal tract, be totally blocked before being released (plosives or affricates) or near blocked (fricatives), make little noise (central and lateral approximants), or go out of the nostril (nasals) instead of the mouth. There are two other terms, 'rhotic' and 'liquid,' which are fairly common but not established on the IPA chart. The

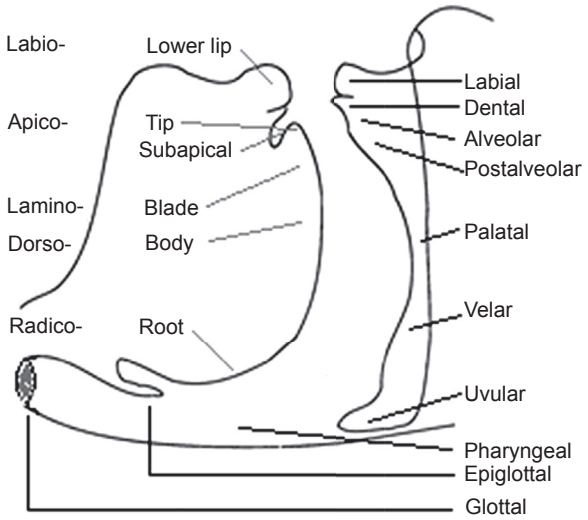

Figure 4 Places of articulation.

former refers to seven *r*-sounds, which could be trills, taps/flaps, or approximants. The latter is a cover term for laterals and rhotics.

In the consonant table there is a third dimension, voicing, implied in the presence of pairs of symbols for obstruents (plosives and fricatives) such as [p ~ b] and [s ~ z]. For example, the symbol at the right, i.e., [b], indicates a voiced sound, and that at the left, [p], voiceless, with the status of aspiration (and breath, etc.) unspecified. This kind of presentation results from the practical and convenience purpose of using the existing Latin alphabet, whose basic function is to transcribe speech sounds in European languages. There are also voicing contrasts occurring in sonorants such as [m · ~m], [n · ~n], and [l · ~l] in other languages but not presented in the table due to the lack of Latin letters. Moreover, the contrast between voiceless and voiced involves only two types of phonation modes; so far up to 14 phonation types have been identified in linguistic phonetics. It is, therefore, better to represent all the contrastive phonation types separately (cf section 'A System of Phonationals').

In the newer versions (1989 onward) of the IPA a nonpulmonic consonant table is included, which contains clicks, implosives, and ejectives. The latter two can also be treated as phonationals because their production and the phonological features that distinguish them from other plosives at the same place of articulation involve glottis movements rather than articulatory gesture.

There is also a list of 'other symbols' on the IPA chart which contains consonants with double articulations of place/manner ([ʍ, w, ɥ, ɰ], etc.) and three rarely seen epiglottal sounds [ʔ, ʕ, ʜ]. Only the pair of prepalatal or alveolo-palatal fricatives [ɕ, ʑ] is an exception. Prepalatals are commonly used in Asian languages such as Sinitic, Tibeto-Burman, Hmong-Mien (Miao-Yao), and Tai-Kadai but rarely seen in European languages (with Polish as an exception). They were previously placed in the consonant table (the 1978 version) next to the palato-alveolar [ʃ, ʒ] (now 'postalveolars'). The two sets of sounds share a similar place of articulation and usually do not contrast with each other phonemically; these might be the reasons to remove the prepalatals out of the consonant table in the later IPA versions. There is, however, an auditory distinction between the two, and a constant articulatory difference is that [ɕ] is more palatalized whereas [ʃ] is more labialized. Perceptually, to a Westerner's ear, Mandarin's [ʂ] rather than [ɕ] sounds more like English's [ʃ]. On the other hand, Mandarin speakers would say [tʂʰɜtʂʰ] rather than [tɕʰɜtɕʰ] when learning to speak the English word 'church.' Even for professional Chinese dialectologists, Cantonese obstruents at that place of articulation are transcribed as [ʧ, ʃ], meaning they are distinguishable from Mandarin [tɕ, ɕ] and sound more similar to those in English.

Vowels

Vowels are placed on a chart according to the positions of the tongue: high–low (or close–open of the mouth) and front–back. When the high-front vowel [i] is being produced, the blade of the tongue is forwarded to approach the alveolo-palatal region. When [u] is being produced, the 'bunched' tongue is pulled back to get close to the velum. While almost all languages in the world have these two high vowels and a low vowel /a/, some have both front [a] and back [ɑ]. There is also a third dimension on the chart: paired symbols like [i ~ y] are distinguished by the shape of the lips, with the symbol on the right indicating a rounded vowel.

The vowel chart defines a physiological quadrilateral space for vowels to reside in. This physiological criterion for vowel description has been changed to an acoustic one. In the newest edition of the *Handbook of the IPA* (1999), the vowel chart in a language is drawn along acoustic parameters of the first two formants' frequencies. For example, the front vowel symbol [ɪ], physiologically defined as between [i] and [e], is placed 'correctly' in the Hindi chart (p. 102), but as low as [e] in the German chart (p. 87) and even lower than [e] in the American English chart (p. 42). That is, a vowel now has its absolute values in terms of formant frequencies, so the IPA phonetic notations share about the same nature as phonemic representations.

In addition, new understanding of the nature of the vowel arose that the vowel is not a solely phonetic concept, but is defined with reference to phonology. This view is reflected in the *IPA Handbook* (1999), in which vowels are defined as "sounds which occur at syllable centers, and … involve a less extreme narrowing of the vocal tract than consonants" (p. 10). Or, more straightforwardly, "the vowel is produced without apparent constriction along the oral tract and serves as the syllabic component of *nucleus*" (Zhu, 2010, p. 231).

Secondary Features

Secondary features are represented by symbols in the 'diacritic' table and the lists of suprasegmentals and tones in the lower part of the IPA chart. Most of the diacritics are used to indicate minor adjustments of articulation. The two for 'syllabic' and 'nonsyllabic' are phonological notations. The remaining five in the top-left corner of the table, voiceless, voiced, aspirated, breathy voiced, and creaky voiced, are symbols for phonation types, which are better grouped with tones and suprasegmentals as phonationals.

Table 1

Major type	*Types*	*Notation*	*Supra-phonational*	
I. Falsetto	1. Falsetto	á	Leng	Long
II. Fortis	2. Preglottalization/ Implosive	ˀp, ˀm/ɓ		Mid-short
	3. Final-glottalization/ Ejective	aʔ/k’		Short
	4. Fortis Creak	ă	P	H/L
III. Clear-1	5. Voiceless (Unaspirated)	p (, m)		H/M/L
	6. Aspirated	pʰ, m̥		4/3/2/1
	7. Weak Aspiration	p‘		
IV.Clear-2	8. Modal Voice	b, m		
V. Stiff	9. Checked Creak	a̰		
	10. Creaky Voice	a̤		
	11. Weak Stiff	a̠		
VI. Breathy	12. Voiced Aspiration(Voiced Breathy)	b̤ʱa̤, m̤a̤		
	13. Slack Voice (Voiceless Breathy)	p̤ a̤ (, m̤ a̤)		
	14. Weak Slack	p a̤		

Phonationals

A System of Phonationals

Phonationals are divided into three groups: phonation types (cf Phonation Types in this volume), major types, and supra-phonationals, which are listed in Table 1.

Traditional phonetics usually deals with three COMMON phonation types: voiceless unaspirated, voiceless aspirated, and modal voice. Ladefoged considered three more (breathy voice, creaky voice, and glottal stop) and regarded all the six as properties of stop consonants. Phonation types are actually equally realized on fricatives, sonorants, and even more on nuclear vowels; they are properties of the syllable rather than of a single segment.

Supraphonationals include length and pitch. Length is subcategorized into long, mid-short, and short, according to the tonal behaviors in Cantonese and other tone languages. Pitch will be discussed in detail later.

Composition of Phonational

Phonational (‘Ph’ in Figure 5), as the first-level nonlinear syllabic constituent, can also be named *tone* (T). Tone consists of three components: *register* (Rg), *length* (Leng), and *pitch* (P); pitch is comprised of *height* (Ht) and *contour* (Cnt). Tone is, therefore, described and represented along four parameters: register, length, height, and contour. These parameters can be subcategorized. *Register*, *length*, and *height* all show a tripartition contrast, i.e., *register* U, M, and L, *length* long, mid, and short, and *pitch height* in 2-way, 3-degree, and 4-level contrasts. Pitch *contour* can be categorized as *level*, *rising*, *falling*, *dipping*, and the less known but rather frequent type *pure low tone*, a low tone without CONTOURICITY specification (Zhu, 2012b). These components are hierarchically structured under the syllable as in Figure 5.

Phonologically, *register*, *length*, and *pitch* are associated with *phonational* or *tone*. Phonetically, “the tone features are sometimes relevant for the initial consonant, sometimes for the nuclear vowel, and sometimes for the final consonants” (Wang, 1967). *Pitch* and *length* are shown on *rime*, and *register* can be on either *onset* or *rime*.

After the nonlinear components have been defined, it is not important which component, the Ph, P, or Cnt, is called ‘tone.’ Or, they could be called ‘tone$_1$,’ ‘tone$_2$,’ and ‘tone$_3$,’ respectively. In the syllable theory (Zhu, 2008), as shown in Figure 5, ‘tone’ equates to *phonational*, dominating all other nonlinear components.

Pitch can be contrastive by height only or by both height and contour. The 2-way contrast in pitch height refers to stress and the accentual tone in Japanese, the 3-degree one to the African register tone and 3-degree stress, and the 4-level one to the Asian contour tone.

In nontonal languages, laryngeal activities are less active and therefore the phonational tier could be saved in their linguistic–phonetic representations. With regard to a few COMMON phonationals such as the two degrees of pitch (strong vs weak), length (long vs short), and phonation (voiceless vs

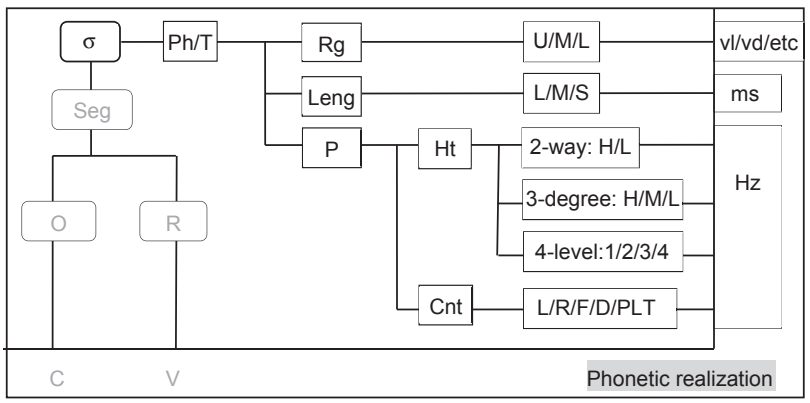

Figure 5 Structure of phonational or tone. Adapted from Figure 4, Zhu, Xiaonong, 2011a. 語言語音學和音法學：理論新框架 Linguistic phonetics and pan-chronic phonology: a new theoretical framework. Studies in Language and Linguistics 31 (1), 64–87.

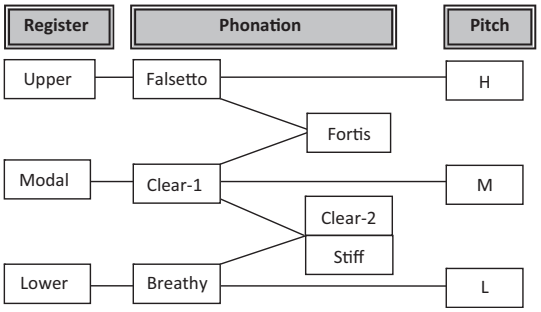

Figure 6 Three tonological registers defined by six major phonation types.

voiced) in these languages, they can be treated, for the sake of convenience, as secondary features of consonants (voicing) or vowels, i.e., suprasegmentals (stress and length). In tone languages, on the contrary, there are more phonationals, usually LESS-COMMON, in the nonlinear tier.

Contour tone is by and large in complementary distribution with stress. In a typical tone language such as Cantonese there is no room for stress, because (1) the language is monosyllabically morphemic and stress resides on polysyllabic words and (2) pitch, the main phonetic cue for stress in nontonal languages, is fully utilized for the distinction of tone in tone languages.

Registers Realized as Phonation Types

The six major types are arranged from top to bottom in Figure 6 roughly in accordance with pitch height. By these phonation types three tonological registers are defined. Not every phonation type can identify a register. Essentially, only three types, falsetto, clear-1, and breathy voice, if occurring in a tone language, can define an independent register. Falsetto defines the highest Upper Register. Breathy voice defines the lowest Lower Register. Clear-1 defines the commonest, default, Modal Register. Every tone language, every language indeed, has this register.

Figure 7 shows the statuses of vocal cords of three phonation types. The vocal cords in the left panel, which is for a falsetto sound, are stretched tightly; they are longer, wider, and thinner than the ones in the middle panel, which is for a modal voice. In the right panel is a slack voice. The vocal cords in this panel do not close entirely. Airflow is continuously leaking out through the aperture of the glottis even during the closing phase of vibration. Slack voice is perceived as having a low pitch and noise on the vowel.

The clear-2 (modal) voice usually does not define a phonological register alone; it is ordinarily accompanied by one or another nonclear voice. For instance, modal voice co-occurs with slack voice in Hmong, Wa, Wu, Xiang, and Gan, and with creaky voice in Thai, Hani, Lolo, and Lahu, etc.

Creaky voice is the commonest among the members of stiff voice. It is associated with the lowest human pitch, and will cause a 'broken tone' in terms of both acoustics and perception. In addition to contrasting with other phonation types in nontonal languages like Mon-Khmer, creak is also found in hundreds of varieties of Chinese and other tonal languages such as Vietnamese and Yi (Lolo). In most tonal cases, creak is a phonetic realization of either M or L Register, serving as an accompanying feature of a low tone. Only in a few linguistic varieties is creak found to play a role of distinctive feature in differentiating two otherwise identical tones (Mai, 2007; Zhu and Yang, 2010). In this case, stiff voice alone defines an L Register.

The last one, fortis voice, is rather subtle in terms of physiological characteristics and acoustic manifestations. Note that fortis and stiff voice are different from the 'stiff voice' suggested by Ladefoged and Maddieson (1996). In short, fortis voice is associated with fortis/strong/long/tense voiceless consonants, e.g., Korean fortis stops and Wu preglottalized consonants, which induce a higher pitch and larger intensity on the following vowel than nonfortis voiceless consonants do. Fortis voice is usually a variant of falsetto or clear-1 and seldom defines a register itself.

Phonation types are not consonants' properties; they define registers, which constitute a syllabic component. Two of them, clear-1 and clear-2, are normal clear voices or COMMON PHONATIONALS and acoustically realized on the initial consonant as well as on the nuclear vowel (a higher pitch after a voiceless than after a model voiced consonant); three, falsetto, slack, and stiff, are nonclear voices or LESS-COMMON PHONATIONALS and are realized on vowels and occasionally on consonants. As for fortis voice, two of its varieties, pre- and final-glottalization, are COMMON PHONATIONALS, and the other two, implosive and ejective, are LESS-COMMON PHONATIONALS. That Register M is defined by clear-1 means Register M is a property of the syllable that begins with an *onset* in clear voice: either a voiceless obstruent, a normal sonorant, or zero *onset* in a clear-voiced syllable.

Figure 7 Sample photographs illustrating three laryngeal statuses: falsetto (left), modal voice (middle), and slack voice (right). Cited in Zhu, Xiaonong, 2012a. Multiregisters and four levels: a new tonal model. Journal of Chinese Linguistics 40 (1), 1–17.

Multiregisters and Four-Level Tonal Model

Based on the syllable model (Figure 5) and the mapping of phonation types onto registers (Figure 6), a new tonal model, the Multi-Register and Four-Level tonal model, is developed (Zhu, 2012a). Basically, it is suggested that (1) a language may have more than one type of phonation which defines up to three tonological registers; (2) each register has its own pitch range, which is divided into four levels; and (3) each register is one notch lower/higher in pitch than an adjacent one. Figure 8 shows a total of three registers and six pitch levels).

In this model the four Beijingese tones are all piled in Register M and numerically transcribed as {55, 35, 22, 52}, and the seven Thai tones, five long and two short, reside in Rg M, too: {33, 22, 52, 45, 25, 45, 32}. Of the six long and two short Vietnamese tones, one is in Rg U, {405}, five in Rg M, {44, 33, 35, 55, 33}, and two in Rg L, {202, 40}.

Phonational Typology

While data of segment distribution (Maddieson, 2005) have been accumulated to an extent that typological studies can be carried out (cf 52009 in this volume), those of phonationals still lack.

Exploration of tone typology started decades ago (Wang, 1967; Cheng, 1973; Maddieson, 1978), based on impressionistically descriptive data. It did not flourish until recently when sufficient amounts of acoustic data were accumulated and a phonational model proposed (Figure 5). There are five basic contour types: level, falling, rising, dipping, and the pure low tone. The falling tone has 11 subtypes (Zhu, 2012b), and the dipping tone, seven (Zhu et al., 2012a). The level tone is the most distinctive one and may have up to six distinct heights (Zhu et al., 2012b). The falling tone is the most frequent type in tone languages. A typological statement can be stated as 'if there is a tone in a language, it tends to be a falling one,' or in terms of implicational universal, 'a level tone implies a falling tone.' Figure 9 shows a typical example from Chálíng Gàn , which has five tones: three circumflex and two falling but no level.

As a LESS-COMMON phonational, tone often does not occur in a language alone, rather it co-occurs with other LESS-COMMON phonationals. The following maps show the geographical coincidence between tones and implosives. Not accidentally breathy voice is also found in the same areas and falsetto in East and Southeast Asia (Zhu, 2010). Geographical locations seem to correlate with linguistic–phonetic patterns: the languages on warmer continents are more phonation-active, while those in colder regions are more segment-prominent

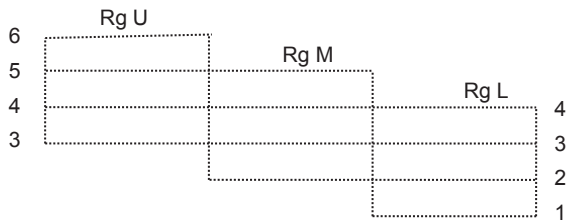

Figure 8 The multiregister and four-level tonal model.

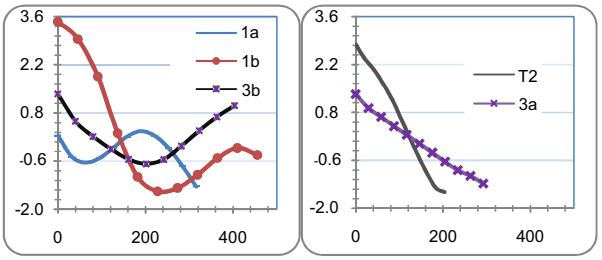

Figure 9 Five tones in Chálíng Gàn: three circumflex: a low dipping {324}, a double circumflex {3232}, and a back dipping {523} (left); and two falling: short {52} versus long {42} (right). Cited in Zhu, Xiaonong, Li, YI, 2012. Double circumflex and back dipping: reports on two newly confirmed types of contour tones. Cahiers de Linguistique Asie Orientale 41 (1), 81–106.

(with more complex phonotactics). A salient example is the Indic languages, which gained implosives and breathy voice after their migration from colder Europe to the warmer subcontinent (Figure 10).

Classification of Some Special Segments

There are some special or strange consonants and vowels whose phonetic properties were previously known to a certain extent but which are difficult to classify phonologically. They are now accountable in the phonational classification (Table 1). Moreover, phonetic properties previously thought to belong to consonant, vowel, or tone separately can be treated uniformly in light of the phonational system.

Three Sets of Voiceless Stops

There are three sets of voiceless stops in Korean, which have been a puzzle in phonetics and phonology for decades. While the aspirated one is considered 'normal,' the other two, unaspirated sets can only be labeled as 'fortis' and 'lenis,' transcribed as, say, **p* and *p* (see the Korean table in Table 2). While the lenis stops have a longer Voice Onset Time (up to 60 ms) than the fortis ones (usually below 30 ms), the perceptual and acoustic properties of the stops are mainly manifested on the nuclear vowels: weaker in intensity and lower in pitch for the vowel after the lenis consonant than for the vowel after the fortis. The problem is how to classify them phonologically or linguistic-phonetically. In terms of the phonational system, they can be classified as different phonation types (Zhu, 2011b): the fortis consonant (the *onset* of a fortis syllable indeed) is of Type 2 preglottalization (like that in French) with an allophonic variant unaspirated (Type 5), and the lenis consonant is of Type 7 weak aspiration with two variants: weak slack (Type 14) and modal voice (Type 8) (Table 2).

There are also three sets of stops in Wu Chinese (see the Wu table in Table 2): a normal aspirated p^h, a 'clear' *p*, and a 'murky' or 'voiced' *b*, as termed in Chinese phonetics and phonology. The murky set was eventually voiceless (Edkings, 1868), but with 'voiced aspiration' (Chao, 1928). In the phonational system the unaspirated *p* is classified as Type 2 preglottalization with an allophone of Type 5 unaspirated, and the murky *b* as Type 13 slack voice with a variant of

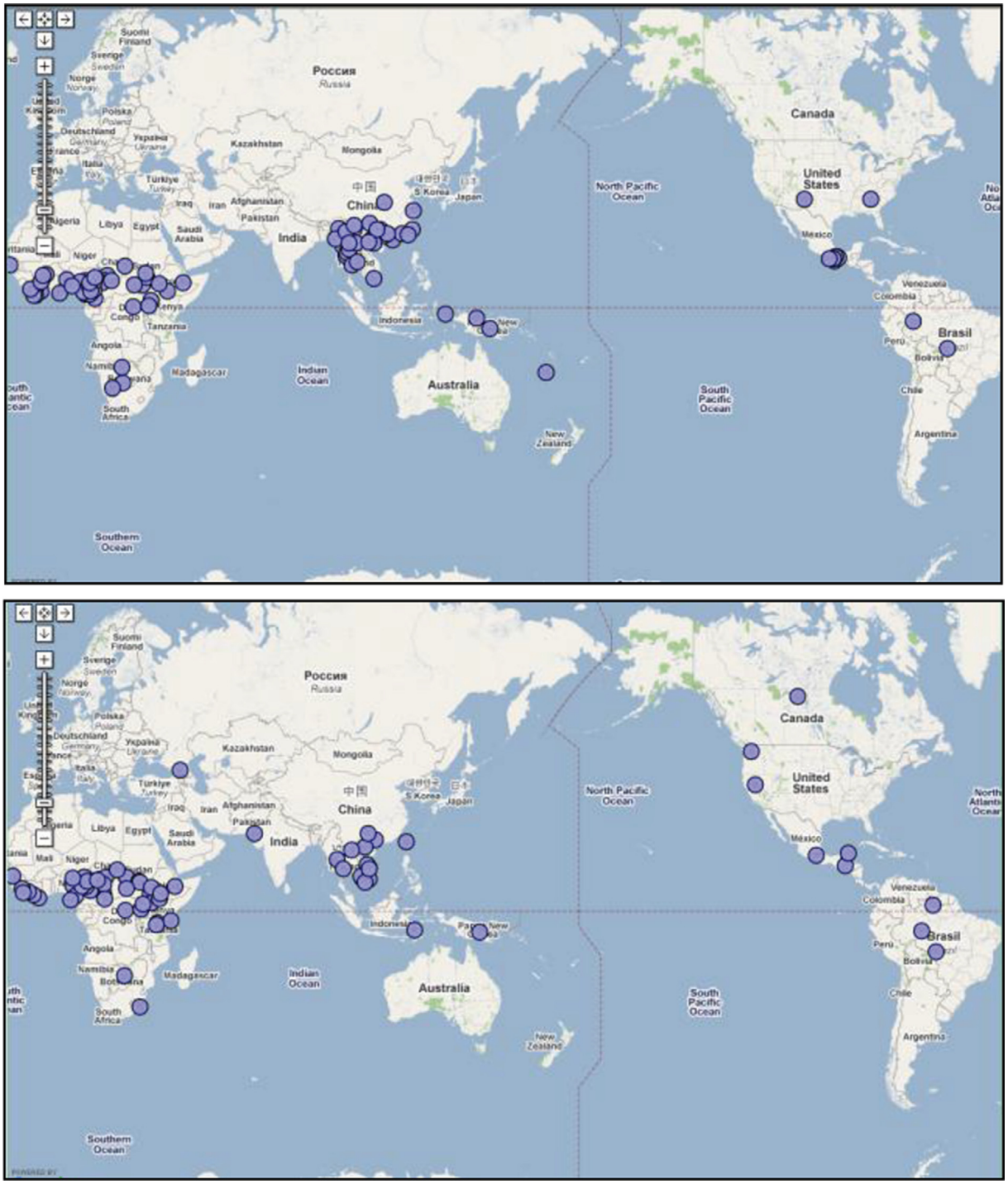

Figure 10 Geographical distributions of tone (top) and implosives (bottom). Cited in Zhu and Guan (2011) .The data are from Maddieson, Ian, 2005. "Tone" and "Glottalized consonants". In: Haspelmath et al. (Eds.), The World Atlas of Language Structures. Oxford University Press, while the finding of the correlation is mine.

Type 8 and Type 14. This is also true for the Hmong language and Xiang and Gan Chinese. Note that there is also a distinction between 'clear' and 'murky' sonorants in Wu, so they are not just properties of stops.

In fact they are not even (syllable-initial) consonants' properties. The perceptual and acoustic properties of these special sounds are mainly manifested on the rime rather than on the onset. Moreover, there is a 'clear' and 'murky' distinction between syllables without initial consonants. It is, therefore, reasonable to assign the phonetic features of these sounds to the syllable rather than to a single segment.

Table 2 Tripartition of phonation types/voiceless stops in Korean (left) and Wu (right)

Kor	trans	Mj type	Type	value
fortis	*p	II.Fts	2.preglt	[ˀp]
		III.Cl-1	5.unasp	[p]
asp.	p^h	III.Cl-1	6.asp	[p^h]
lenis	p	III.Cl-1	7.wk asp	[p‘]
		IV.Cl-2	8.mod	[b]
		VI.Br	14.wk slk	[p̤a̤]

Wu	trans	Mj type	Type	value
clear	p	II.Fts	2.preglt	[ˀp]
		III.Cl-1	5.unasp	[p]
asp.	p^h	III.Cl-1	6.asp	[p^h]
murky	b	VI.Br	13.slck	[p̤a̤]
			14.wk slk	[p̤a̤]
		IV.Cl-2	8.mod	[b]

Tense versus Lax Vowels

The distinction between the tense and lax vowels in English involves segment contrast, with a secondary contrast in length (and other physiological factors that are not phonologically categorized, e.g., wall resistance), but the tense/lax distinction in Austroasiatic and Tibeto-Burman languages involves phonation contrast.

In both Wa of Austroasiatic and Hani of Tibeto-Burman there are two sets of vowels, labeled as ‘tense’ and ‘lax.’ Their phonetic properties and phonological contrast, however, are totally different. In Wa the tense vowels are in normal clear voice while the lax vowels are in slack voice. In Hani, on the contrary, it is the lax vowels that are in clear voice and the tense vowels in creaky voice. So along the ‘tense/lax’ dimension, defined by the adductive tension and medial compression at the glottis, there are at least three values rather than two. As the Hani creaky voice also occurs on the initial consonant sometimes, and the Wa slack voice also affects the initial consonant and makes it ‘murky’ or ‘lenis,’ the phonation types occurring in these languages are not just vowels’ properties, but syllables’.

There has been a puzzle in Wa phonology on why the tense/lax vowel contrast only occurs after an unaspirated consonants. After an aspirated consonant, there are only tense vowels. This is just the case in Korean and Wu: the distinction between fortis/lenis or clear/murky only occurs in unaspirated consonants. So the phonetic nature of the Wa vowels is the same as that of the Korean and Wu consonants: they are all caused by (weak) slack voice (or weak aspiration) at the syllable level, but treated differently in different languages as vowels’ or consonants’ properties.

Moreover, the same slack voice can also be treated as a tonal property in tonal languages: *yang* tones in Wu and even numbered tones in Hmong. Since most of the Austroasiatic languages are nontonal, the slack voice is treated as vowels’ property. Wa is an Austroasiatic language that has four newly emerging tones, three for native words and one for words borrowed from Chinese.

In traditional phonetics and phonology, ‘fortis/lenis’ or ‘clear/murky’ are considered as consonant contrast, ‘tense/lax’ as vowel contrast, and ‘*yin/yang*’ or ‘odd/even numbered’ as tone contrast with unrevealed phonetic properties. They can be analyzed uniformly now in the system of phonation types.

Summary

Phonetics has expanded its territory quickly in the past few decades. Dozens of new disciplines have emerged. By interacting with them, articulatory phonetics has changed enormously so that a split of the discipline has taken place. Half of it remains as articulatory phonetics but is leaving from linguistics. It has become more technology oriented, and many efforts have been made in articulatory modeling, which is more like engineering phonetics than linguistic phonetics.

The other half remains in linguistics but is no longer pure articulatory; it resorts to more parameters than the articulatory parameter. This split is so significant for linguistics that it would reshape linguistic articulatory phonetics and affect disciplines such as phonology, typology, dialectology, and evolutionary/historical phonology. This linguistic articulatory phonetics is becoming more theoretically oriented. It is still part, maybe the core part, of the greater phonetic science, yet differs from other phonetic disciplines in that it is of linguistics, by linguistics, and for linguistics. So it is not ‘pure’ phonetics. In linguistic phonetics, physiological and acoustic parameters are sought after, but they should be phonologically conceptualizable or categorizable, or have linguistic implications for typology or phonological evolution.

Phonetics is to phonology what physics is to chemistry. In phonetics and physics, the basic unit is the smallest analytical element, while in phonology or chemistry, the whole study is established on a structural unit: syllable in phonology and molecule in chemistry. Had chemistry not had its own basic, structural unit, it would not have been independent or autonomous from physics. Phonetics overlapping with linguistics has given rise to the linguistic, articulatory phonetics, which, as part, maybe the core part, of phonology, has the syllable as its basic unit.

See also: Dialectology; Linguistic Typology; Phonology; Speech Production, Neural Basis of; Speech Production, Psychology of; Suprasegmentals; Writing Systems.

Bibliography

Abercrombie, David, 1967. Elements of General Phonetic. Edinburgh University Press.
Anderson, Stephen R., 1981. Why phonology isn’t ‘natural’. Linguistic Inquiry 12 (4), 493–539.

Catford, John C., 1977. Fundamental Problems in Phonetics. Edinburgh University Press.

Chao, Yuen Ren, 1928. Xiandai Wuyu de Yanjiu [Studies in the Modern Wu dialects]. Hsinghua College Press, Peiping.

Cheng, Chin-chuan, 1973. A quantitative study of Chinese tones. Journal of Chinese Linguistics 1 (1), 93–110.

Edkins, Joseph, 1868. A Grammar of Colloquial Chinese as Exhibited in the Shanghai Dialect. Presbyterian Mission Press, Shanghai.

International Phonetics Association, 1999. Handbook of the International Phonetic Association. Cambridge University Press.

Jakobson, Roman, Fant, Gunnar, Halle, Morris, 1952. Preliminaries to Speech Analysis. MIT Press, Cambridge, MA.

Ladefoged, Peter, 1971. Preliminaries to Linguistic Phonetics. The University of Chicago Press.

Ladefoged, Peter, 2006. A Course in Phonetics, fifth ed. Thomson Wadsworth, Boston.

Ladefoged, Peter, Maddieson, Ian, 1996. The Sounds of the World's Languages. Blackwell, Oxford.

Laver, John, 1980. The Phonetic Description of Voice Quality. Cambridge University Press.

Maddieson, Ian, 1978. Universals of tone. In: Greenberg, J. (Ed.), Universals of Human Languages. Stanford University Press, pp. 335–365.

Maddieson, Ian, 2005. "Tone" and "Glottalized consonants". In: Haspelmath, et al. (Eds.), The World Atlas of Language Structures. Oxford University Press.

Mai, Yun, 2007. 广西贺州八步区八都话入声的语音分析 Phonetic analysis of the entering tone in Badu dialect of Hezhou, Guangxi. Journal of Guilin Teachers' College 1, 1–7.

Martinet, A., 1949. Phonology as Functional Phonetics. Publications of the Philological Society XV, 1–27.

Sommerstein, A.H., 1977. Modern Phonology. University Park Press, Baltimore.

Trubetzkoy, Nikolai S., 1939. Principles of Phonology (English translation, 1969).

Wang, William S-Y., 1967. Phonological features of tone. International Journal of American Linguistics 33, 93–105.

Zhu, Xiaonong, 2008. 音節和音節學 Syllable and syllabics. Oriental Linguistics 4, 142–164.

Zhu, Xiaonong, 2010. ≪語音學≫ Phonetics: An Introduction. Commercial Press, Beijing.

Zhu, Xiaonong, 2011a. 語言語音學和音法學：理論新框架 Linguistic phonetics and pan-chronic phonology: a new theoretical framework. Studies in Language and Linguistics 31 (1), 64–87.

Zhu, Xiaonong, 2011b. 輔音類型還是別的什麼類型？－如何在發聲類型框架中安排吳語和韓語的三套塞音? Consonant types or something else: how to classify the three sets of stops in Korean and Wu in a framework of phonation types? Bulletin of Chinese Linguistics 5 (2), 173–186.

Zhu, Xiaonong, 2012a. Multiregisters and four levels: a new tonal model. Journal of Chinese Linguistics 40 (1), 1–17.

Zhu, Xiaonong, 2012b. 降調的種類 A classification of falling tones. Studies in Language and Linguistics 2, 1–19.

Zhu, Xiaonong, Yingwei, Guan, 2010. 全州土話音節的四分發聲活動 Four phonation types in Quanzhou dialects, Guangxi. Dialect 4, 289–300.

Zhu, Xiaonong, Jianfen, Yang, 2010. 嘎裂声作为低调特征 Creaky voice as the feature [+low] in Hebei dialects. Yuyan Yanjiu Jikan 7, 134–147.

Zhu, Xiaonong, Li, Yi, 2012. Double circumflex and back dipping: reports on two newly confirmed types of contour tones. Cahiers de Linguistique Asie Orientale 41 (1), 81–106.

Zhu, Xiaonong, Defu, Shi, Mingying, Wei, 2012. 魚梁苗語六平調和三域六度標調制 Six level tones in Yuliang Hmong and the three-register and six-level tonal model. Minority Languages of China 4, 1–10.

Zhu, Xiaonong, Li, Yi, Ting, Zhang, May 27–29, 2012. Types of dipping tones. In: Paper Presented at the Third International Symposium on Tonal Aspects of Languages. Nanjing, China.

Phonology

Mary Paster, Pomona College, Claremont, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S.R. Anderson, volume 17, pp. 11386–11392, © 2001, Elsevier Ltd.

Abstract

Phonology deals with sound structure in language. While phonetics studies the physical properties of sounds, phonology concerns their mental representations. Phonology originated with the insight that much observable phonetic detail is irrelevant or predictable in language. This led to positing *phonemes* as minimal contrastive units, each comprising a collection of *distinctive features*. Later work went beyond this focus on surface contrast and re-conceived phonology as an aspect of speakers' mental grammars. Theoretical debates have involved the interaction of phonological regularities; the relation of phonology to other components of grammar; and the use of rules versus constraints to encode phonological regularities.

Phonology is concerned with the sound structure of words and utterances, the way distinctions in sound are used to differentiate linguistic items within a language, and the ways in which the sound structure of the 'same' element varies as a function of the other sounds in its context. While both phonology and phonetics are concerned with the role of sound in natural language, they differ in that phonetics deals with the articulatory, acoustic, and perceptual properties of sounds (*see* Phonetics, Articulatory), while phonology studies the ways in which these properties are represented in the mental grammar and how the grammar manipulates these representations.

Goals of Phonology

The study of a language's particular phonological system has a number of aspects. First, it must be able to characterize the language's *inventory*: which phonetically possible sound types occur in utterances in the language? Second, it must characterize matters of *contrast*: which of the phonetic differences that occur in the language can serve to distinguish utterances (words, sentences, etc.) from one another? Third is the matter of *contextual limitation*: even though some property P occurs in language L, are there some environments from which P is excluded? And when P is apparently a property of (some part of) some element but that element occurs in a position from which P is excluded, what other property – if any – appears in its place? Finally, and to some phonologists most importantly, there is the description of *alternation*: when the 'same' linguistic element appears in different overt forms in different environments, what systematic differences occur? What conditions govern the range of phonetically distinct forms that can count as the 'same' word, morpheme, etc.?

It should be noted that the present article is limited to the phonological systems of spoken languages and ignores manual or signed languages (*see* Sign Language). This is misleading in important respects; it has been argued that most of the basic principles of spoken language phonology characterize the organization of signed languages as well (Coulter, 1993). Just as words are composed of sounds, and sounds of component properties, signs are also composed of structured systems of more basic constituent elements. Units such as the syllable have close parallels in signed languages. While there are clear differences that depend on modality, these are argued to be relatively superficial. A comprehensive theory of phonology as a part of the structure of natural language should take these broader issues into account.

Some History

Prior to the early twentieth century, studies of sound in language concentrated on the ways in which sounds are made (articulatory phonetics), often confusing the letters of a language's writing system with its sounds. Toward the end of the nineteenth century, however, increasing sophistication of measurement techniques made it possible to explore a much wider range of differences among sounds, and to lay out the structure of speech in vastly greater detail. Much of what was found involved the observation that speech is continuous, such that whatever is going on at any particular moment is at least a little different from what has gone on just before and what will go on immediately afterward. A full characterization of an utterance as a physical event requires the recognition of a potentially unlimited number of distinct points in time, but it is clear that our understanding of an utterance as a *linguistic* event is hindered, rather than helped, by the recognition of this continuous character of speech. Speech normally is represented as a sequence of a small number of discrete segments, strung out in consecutive fashion like beads on a string; although such a *segmental* representation is fairly abstract, it vastly facilitates the discovery of regularity and coherence in language.

It is clear that the role of particular sound differences varies considerably from one language to another. Thus, in English, the vowel sound in the word *bad* is much longer than that in *bat* (more than half again as long), but such a difference in length is always predictable as a function of the following sound, and never serves by itself to distinguish one word from another. In Tahitian, in contrast, essentially the same difference in length is the only property distinguishing, for example, *paato* 'to pick, pluck' from *pato* 'to break out.' A theory of sound that attends only to physical properties has no way of clarifying the quite different functions these properties may have across various languages.

Saussure and the 'Phonemic Principle'

The great Swiss linguist Saussure (1916) was the first to stress that in order to understand the role of sound in language, it is necessary to focus not (just) on the positive properties of sounds, but on their differences. He suggested that in the study of individual languages, as opposed to general phonetics, utterances should be characterized in such a way that two such representations might differ only in ways that could potentially correspond to a difference between two distinct messages in the language in question. Thus, since long and short vowels never (by themselves) distinguish two distinct utterances in English, the difference should not be indicated in that language, while for Tahitian, it must be. A representation with this property will be called *phonological*; it will obviously be specific to a particular language, and the distinctive elements that appear in it can be called the *phonemes* of that language.

While Saussure enunciated this principle quite forcefully and persuasively, he provided few specific details of just what a phonological representation should look like. There are in fact a variety of ways in which his insight could potentially be realized, and much subsequent discussion in phonology hinges on these differences of interpretation.

The Development of Phonology as a Theory

Various individual investigators arrived at conclusions similar to Saussure's about the importance of attention to language-particular sound contrasts. One of these was the Polish linguist Baudouin de Courtenay (1972), whose work actually antedated Saussure's but attracted less attention. He developed a sophisticated view of the relation between phonetics and phonology both in individual grammars and in linguistic change. As transmitted by his later students, Baudouin's views on the nature of the phoneme constituted an important strand in thinking about language as this developed in Russian linguistics in the early years of the twentieth century. This, in turn, provided the background from which the work associated with the Linguistic Circle of Prague grew in the 1920s and 1930s.

Two of the most prominent members of the Prague Circle were Trubetzkoy (1939) and Jakobson (1941). In their studies of Slavic languages and their histories, they stressed the notion that the collection of potentially contrastive sound types in a language was not simply an inventory, but a highly structured system. This system is organized in terms of a small number of mutually orthogonal dimensions (such as voicing, stop vs continuant, nasality, etc.), each of which serves in parallel fashion as the basis of multiple contrasts. The notion that the fundamental terms of sound structure in language are these properties themselves and not (or at least not only) the complete sounds they characterize has remained an important component of most subsequent theorizing.

American Structuralist Phonology

Early thinking about sound structure in America was dominated by the anthropological interests of Franz Boas and focused on an accurate rendering of the sound contrasts in the comparatively 'exotic' indigenous languages of the new world. Boas's student Edward Sapir, however, was concerned to place the study of language in the broader context of an understanding of the human mind and society. Therefore, he stressed (Sapir, 1925) the notion that the elements of sound contrast in a language should be regarded as having a primarily mental reality, part of the speaker/hearer's cognitive organization rather than as external, physical events.

The rise of positivist views of science in the 1930s, and especially of behaviorist psychology, made Sapir's type of mentalism quite marginal, and replaced it with more rigorous operational procedures for investigating notions of contrast. Especially associated with the ideas of Bloomfield (1933) and later structuralists such as Harris (1951), the result was a theory of the phoneme based exclusively (at least in principle) on a set of mechanical manipulations of corpora of observed linguistic data, from which a set of contrasting minimal elements was to be derived. The central notion of this theory was a phonemic representation related to surface phonetic form in a way that would later be formulated explicitly as a condition of *bi-uniqueness*: the requirement that given either a phonetic or a phonemic representation of an utterance in a given language, that could be converted uniquely into the other (disregarding free variation) without additional information.

Generative Phonology

The phonemic theories of American structuralists provided a way to characterize linguistic contrasts, the inventories of sound types used in a given language, and the ways in which sounds can be combined into larger structures, but other aspects of sound structure were less satisfactorily accommodated within those views. In particular, questions of the ways in which unitary meaningful elements change in shape according to their sound context (or 'allomorphy') failed to receive systematic treatment. Since any difference among sounds that could serve to contrast linguistic elements was *ipso facto* a difference between irreducibly basic terms, there was really no way to express the notion that a single item could take a variety of forms (as in the prefixes in *inefficient*, *imprecise*, *irregular*, *illegal*, etc.) except by simply listing the variants. Such a list is undoubtedly appropriate for cases such as the forms of English to be (*am*, *are*, *is*, *was*, *were*, etc.) which are unrelated to one another in form; but in many other cases, the variation is transparently systematic and a function of the sounds in the element's environment. This sort of variation was recognized by structuralist phonologists, but relegated to marginal status.

Beginning with the work of Morris Halle, a student of Jakobson, linguists began to question the centrality of surface contrasts in sound structure. The result was a new view that allowed morphophonemic regularities as well as more superficial ones to be accommodated within a phonological description. The success of this more abstract notion of sound structure in dealing with hitherto irresolvable problems in the description of stress (*see* Suprasegmentals) contributed greatly to its success, and the resulting theory of generative phonology as developed in the work of Halle together with Noam Chomsky rapidly became the dominant view in the field by the middle of the 1960s.

Phonology as a System of Rules and Representations

A basic insight in the development of generative phonology was the proposal that it is not only the representation of linguistic elements in terms of basic contrasts that matters: an adequate theory must characterize what a speaker knows about

the sound system of the language, and that includes regularities of variation and alternation as well as inventories of basic elements. Combining these two aspects of phonological knowledge required the explicit recognition of a system of rules (expressions of regular patterns of contrast and variation in sound shape) in addition to the theory of representations. Developing an adequate theory of phonological rules, in turn, necessitated a notion of phonological representation that was related to surface phonetic reality in a much more complex and indirect way than the phonemic representations of structuralist linguistics. The central problems of phonological theory came to be articulated in terms of the theory of rules, their nature, and their interaction, and the underlying phonological representations that need to be posited in order to allow them to be expressed in their full generality.

Issues in Phonological Theory

As reviewed in the section 'Some History' above, the problem of phonological description was originally conceived as a matter of discerning the range of contrastive sound units in a language and arranging them in a system that brings out the dimensions of their contrast. When phonology is seen as a form of knowledge, however, as generative phonologists have stressed, the sound structure of natural language takes on quite a different form and presents quite different problems. Among these are (1) the nature of underlying ('phonological' or 'phonemic') representations as well as surface ('phonetic') representations; (2) the ways in which phonological regularities serve to relate phonological to phonetic representation, including the interactions that may obtain among regularities; and (3) the relation between phonological form and other aspects of linguistic knowledge, such as word structure (*see* Morphology in Linguistics) and sentence structure. Various aspects of these questions are addressed below.

The Abstractness of Phonological Representation

Initial reaction to the proposals associated with generative phonology centered on its abandonment of a phonemic representation based on the condition of biuniqueness. Relaxing this defining characteristic of structuralist phonemics led to phonological representations that were considerably more abstract in their relation to phonetically observable properties. The proposals of Chomsky and Halle (1968) concerning the analysis of English, for example, involved positing final 'silent *e*' in words like *burlesque* (phonologically /bVrleske/), geminate consonants in words like *confetti* (/kVnfetti/), a distinction among /k, s, c/ such that *acquiesce* is phonologically /æckwiesce/, etc. Indeed, there were no constraints whatsoever on the relation between phonological and phonetic form, apart from the desire to set up underlying forms from which as much variation as possible in the shapes of particular morphological units could be predicted. Since much of this variation is the residue of earlier sound changes that have affected the same element in different ways in different environments, many early generative analyses resembled historical accounts of the language more than they did the knowledge of current speakers.

The perception of such apparently excessive abstractness led to proposals for constraining the operation of phonological rules in grammars, and the kinds of representation that should be posited. Kiparsky (1973), in particular, suggested a variety of conditions that would have the effect of prohibiting rules of 'absolute neutralization' (by which some posited phonological distinction is eliminated uniformly in surface forms, such that it never corresponds directly to a phonetic distinction) and other perceived abuses of the theory's representational freedom. Other researchers, however, have pointed out cases where absolute neutralization does seem to be called for in the analysis of particular languages, since it allows for an explanation of the different behaviors of sets of surface-identical segments within the phonological system. The notion of *learnability* entered into this debate, with proponents of constraints on abstractness arguing that a highly abstract system is not learnable; others have argued that abstract systems are indeed learnable when there is evidence for the abstract structures in the form of alternations. The tension between a constrained universal theory versus an empirically adequate and/or elegant analysis of specific languages is apparent here as in other theoretical debates in the field. At present there is no real consensus in the field as to what, if any, constraints on abstractness should be hypothesized to exist in grammars.

The Interaction of Phonological Regularities

In the nature of a structuralist phonemic representation, all of the regularities expressed in the grammar are mutually independent. That is, it is only necessary to know the phonemic environment of a phoneme to predict its phonetic realization, and it is only necessary to know the phonetic environment of a phonetic segment to establish its phonemic correlate. When the bi-unique relation between these levels of representation is relaxed, however, more complex possibilities arise for regularities to interact with one another.

Consider the formation of English regular plurals, for example. For the words *cat*, *dog*, *horse* (roughly (kæt), (dog), (hors) respectively, phonetically), the corresponding plurals are *cats*, *dogs*, *horses* ((kæts), (dogz), (horsɪz)). Assume that the phonological representations of the nouns are essentially the same as their phonetic forms, and that the regular plural ending has a single constant phonological form: /z/ for concreteness' sake, though the point to be made is independent of this choice. Now there are two distinct regularities that are involved in determining the pronunciation of the plural ending:

a. If the noun ends in a sibilant ((s, z, ʃ, ʒ, t͡ʃ, d͡ʒ)), a vowel (ɪ) appears between the stem and the ending.
b. If the stem-final sound is voiceless, the ending is voiceless ((s)) as well.

Now consider how these principles interact in determining that the pronunciation of *horses* (phonologically /hors + z/) should be (horsɪz). In this case, the conditions for both of the rules above, *b* as well as *a*, are met, and we might expect both to apply, yielding (incorrect) (horsɪs). The relevant observation is the following: the vowel inserted as a result of rule *a* has the effect that the stem and the ending are no longer adjacent, and so rule *b* is inapplicable. That is, the regularity

represented by rule *b* in this case presupposes that the set of clusters of consonants that will remain adjacent in surface forms (a matter that is potentially affected by applications of rule *a*) has been determined. This logical relation of presupposition between rules is generally expressed by saying that rule *a* applies 'before' rule *b*. In some cases, such rule ordering produces *opacity*, wherein the inspection of a surface form seems to show that a rule should have applied but didn't (*underapplication*) or a rule applied that should not have (*overapplication*). In the theory of rule ordering, these effects are explained by rule *b* creating the surface environment in which rule *a* would have applied if *b* had applied first (in the case of underapplication), or rule *b* eliminating the surface environment in which *a* applies, but after *a* has already applied (in the case of overapplication). Crucial to modeling such effects is the notion that the output of rule *a* is an intermediate form in the derivation, and it is to this intermediate form that rule *b* will then apply.

Much of the abstractness of early generative phonological analyses was made possible precisely by the availability of rule ordering as a descriptive device: the possibility of specifying rule interactions such as the above as an independent parameter within grammars. Although arguments for the necessity of such ordering formed an important part of the literature, and particular accounts sometimes involved apparently crucial depths of ordering as great as 12–15, some linguists felt that (as part of the general campaign to reduce abstractness in phonology) stipulated ordering relations ought not to be permitted as a descriptive device. Despite the existence of cases apparently requiring irreducible specification of relative order (e.g., dialects with the same sets of rules and underlying forms, differing in their surface forms only as a consequence of different interactions of the rules) and the fact that all proposed formulations of principles from which observed rule interactions could supposedly be predicted had obvious and well-established counterexamples, a number of linguists have persisted in their belief that 'extrinsic' rule ordering statements ought to be prohibited. This view leaves a number of kinds of empirically observed interactions unaccounted for, a problem that persists as many phonologists have turned from rules to constraints as descriptive devices.

The Structure of Representations in Phonology

As already noted, the description of speech at either the phonetic or the phonological level as composed of a sequence of discrete segment-sized units is an abstraction from physical reality. Its justification comes not from directly observable acoustic properties of sounds, but from the extent to which it allows the analyst to uncover what is orderly and coherent in linguistic structure. By the mid-1970s, however, it had become apparent that a purely segmental organization of representations impeded the description of linguistic regularity in several respects.

One of the first of these problems to be discussed arose from the analysis of systems of tonal contrasts, common in the languages of Africa and Asia and, in fact, much of the world. Careful analysis revealed multiple ways in which tonal properties were problematic for strictly segmental models: first, what appeared to be a single tonal specification might take as its scope more than a single segment (perhaps the vowels of several consecutive syllables, or an entire word); second, what appeared to be a single segment (a unitary short vowel, for example) might have a tonal specification involving two or even more consecutive tonal levels; and third, there was evidence for tones existing in the phonological representation of a word but not belonging to or being directly phonetically realized on any segment in that word. If it is assumed that each of the consecutive segmental units of which a representation is composed bears exactly one specification for each potentially distinctive property, and the specifications of distinct segments are independent of one another, all of these situations are anomalous.

As a result of these observations, the segmental view came to be replaced with an *autosegmental* notion of representation. On this view, specifications for each potentially distinctive property (or feature) succeed one another discretely, but an additional dimension of representation is the specification of the way these are synchronized or aligned with other properties or features. In the limiting case, where each specification for a given feature is temporally aligned with one and only one specification for each of the other features, the standard segmental picture arises; in other cases, the relations among these features are not one-to-one. Once this view was articulated, it became apparent that many properties other than those of tone were also most appropriately described in this way. Nasality, in particular, behaves similarly in many languages, and the very common phenomenon of assimilation (including vowel harmony, voicing assimilation, and assimilation in place of articulation) is often best treated as an alteration in the scope of some features rather than a change in their values.

A theory of autosegmental representations where each feature stands alone would fail to account for the fact that in many cases, groups of features act as a unit. For example, in the domain of vowel harmony, often two or more features harmonize simultaneously (e.g., the features representing backness and roundness). This observation gave rise to the notion that phonological features are not just an unordered set, but instead are organized into a hierarchical arrangement such that, for example, features of consonantal place of articulation can be manipulated by the grammar as a unit. As of the early 1990s there was some degree of consensus on specific groups of features that should be organized under distinct representational nodes in such a *feature geometry*, but with the advent of constraint-based theories, many phonologists turned to other questions and abandoned the effort to work out the full details of a universal feature geometry. At present, many questions in this area remain unresolved.

Another representational issue concerns the existence of structural units larger than the single segment, such as the syllable. Classical generative phonology, as represented by, for instance, Chomsky and Halle (1968), makes no appeal to syllables (or any other unit above the segment). This was not simply an omission, but rather a systematic claim to the effect that segmental specification could always suffice to express any phonological generalization: that is, that any observation involving syllables (of which there are many in Chomsky and Halle, 1968) could be satisfactorily reformulated in a way that dispensed with any units other than segments. The

boldness of this claim invited immediate attack, and it soon became clear that there were indeed aspects of phonological structure that required reference to syllables and other suprasegmental units. The resulting theory of the *prosodic hierarchy*, by which segments are grouped into syllables, themselves grouped together into *metrical feet*, which in turn form constituents of *phonological words* (perhaps with some additional categories, according to some writers), has become another standard assumption about phonological structure.

Interactions between Phonology and Other Areas of Grammar

The original conception of a generative phonology was as a component of the grammar that served to map lexically specified syntactic structures onto phonetic form, a representation appropriate to serve as the instructions to the language-independent mechanisms of speech production. This picture logically entails the notion that words, phrases, and sentences are already fully assembled in the input to the phonology. This assumption, however, has proven to be problematic.

The notion that phonological properties interact with the syntactic environment in which forms occur is quite a familiar one, especially in the description of accent and intonation (*see* Suprasegmentals). Word formation, too, may depend in some instances on phonological properties of roots and affixes (*see* Morphology in Linguistics) in a way that contradicts the assumption that phonology applies strictly to the output of this part of the grammar.

The most extensively elaborated picture of how phonology interacts with the rest of grammar is presented by the theory of lexical phonology (see papers in Hargus and Kaisse, 1993 for introduction and explication). Briefly, this picture distinguishes between lexical and postlexical regularities, with the former applying in a way that interacts with word formation, and the latter applying to a representation in which fully formed words appear in their syntactic context. The lexical rules, in turn, may (depending on the language) be divided into two or more *strata*, for example, rules applying to *roots*, *stems*, or *words*. A process of word formation takes some lexical unit as its input and yields an output (affixed or otherwise altered in form), with this output then subject to appropriate phonological modification. The resultant representation may serve as the input to further *cycles* of word formation plus phonology, until a complete surface word is arrived at. Words, assembled into larger syntactic constructions, are then subject to the post-lexical rules.

While various forms of lexical phonology generally constitute the background assumption of much phonological discussion, such a view of the overall architecture of grammar has at times been rejected. In particular, constraint-based theories involving a single stage in the conversion of underlying to surface form are, in their strictest form, incompatible with the sequential, derivational character of lexical phonological description. Later work in constraint-based models recognized the need for phonology to be sensitive to the internal structure of words and has attempted to reincorporate this sensitivity into the theory in various ways.

Rules versus Constraints in Phonological Description

Formulations of phonological regularities since the 1950s have drawn on the mechanisms of automata theory: typically, these have been expressed as rewriting rules that map an input representation onto an output, one step at a time. In the early 1990s, however, an alternative possibility was suggested. Instead of converting inputs to outputs in several stages, with each step involving the imposition of a single regularity in the computation of a new representation, one might formulate all of the regularities as elements of a single system of constraints. If these constraints are ranked with respect to one another, it is possible to say that any particular constraint may be violated in the output, provided that such a violation is motivated by the need to avoid violating a more important (higher ranking) constraint. The resulting theory provides an architecture of grammar that is very different from that which previously dominated most of the literature in phonology. For example, the kinds of interaction among regularities treated above as matters of rule ordering can no longer be described in this way, since all of the constraints in the grammar apply (in principle) in a single, simultaneous block. Constraints can have a language-particular ranking among themselves, but this is a matter of priority, not serial ordering. Other differences have been alluded to in the discussion above.

Constraint-based theories along the lines of optimality theory (OT; Prince and Smolensky, 1993) became the most popular mode of phonological theorizing in the US in the 1990s. The earliest versions of OT were argued to be very simple, with a set of universal constraints and a parallel input–output mapping. Attempts to fit known phonological patterns into this theoretical framework posed serious problems, however. It was soon acknowledged that not all constraints can be universal; the theory needs to allow for language-specific constraints to be learned. Cases of opacity were also especially problematic, since a strictly parallel model has no way of incorporating the notion of intermediate forms that were crucial to modeling opacity in rule-based phonology. Various attempts were made to work around this problem with a series of add-ons to the original OT proposal, each of which was criticized as excessively powerful, inadequate to account for all types of opacity, and/or a rejection of the notion of parallelism that had made OT appealing to many phonologists in the first place. No consensus has been reached as to the proper treatment of opacity in constraint-based models. Another significant problem was in the failure of constraint-based theories to take the results of research in lexical phonology into account, since the original OT model was, by design, insensitive to word-internal morphological structure. Later models did incorporate word structure in various ways; for example, stratal OT is essentially OT combined with lexical phonology such that input–output mappings are still done in parallel within each level, but different levels (root, stem, word) constitute separate stages in a derivation. The state of constraint-based theories in the early 2010s appears somewhat fragmented, with some theorists pursuing models that look much more like rule-based theory in incorporating serially ordered steps in a derivation, and others positing numerically weighted constraints as opposed to strict rankings. A number of researchers reject constraint-based theories entirely, arguing against them in favor of rule-based phonology on empirical and conceptual grounds; Vaux and Nevins (2008) give an overview of this debate.

See also: Phonetics, Articulatory; Saussure, Ferdinand de (1857–1913); Speech Perception; Speech Production, Neural Basis of; Speech Production, Psychology of.

Bibliography

Anderson, S.R., 1985. Phonology in the Twentieth Century: Theories of Rules and Theories of Representations. University of Chicago Press, Chicago.

Baudouin de Courtenay, J., 1972. Selected Writings of Baudouin de Courtenay. Indiana University Press, Bloomington, IN.

Bloomfield, L., 1933. Language. Holt, New York.

Chomsky, N., Halle, M., 1968. The Sound Pattern of English. Harper and Row, New York.

Coulter, G.R., 1993. Current Issues in ASL Phonology (Phonetics and Phonology 3). Academic Press, San Diego.

Fischer-Jørgensen, E., 1975. Trends in Phonological Theory. Akademisk Forlag, Copenhagen, Denmark.

Hargus, S., Kaisse, E., 1993. Studies in Lexical Phonology (Phonetics and Phonology 4). Academic Press, San Diego.

Harris, Z., 1951. Methods in Structural Linguistics. University of Chicago Press, Chicago.

Jakobson, R., 1941. Kindersprache, aphasie und allgemeine lautgesetze. In: Jakobson, R. (Ed.), Selected Writings, vol. I. Mouton, The Hague, pp. 328–340 (English Trans. 1968 Child Language, Aphasia and Phonological Universals).

Joos, M., 1957. Readings in Linguistics, vol. 1. American Council of Learned Societies, Washington, DC.

Kenstowicz, M., 1994. Phonology in Generative Grammar. Blackwell, Oxford, UK.

Kiparsky, P., 1973. Phonological representations. In: Fulimura, O. (Ed.), Three Dimensions of Linguistic Theory. The TEC Corporation, Tokyo.

Prince, A., Smolensky, P., 1993. Optimality Theory: Constraint Interaction in Generative Grammar. Rutgers University, New Brunswick, NJ.

Sapir, E., 1925. Sound patterns in language. Language 1, 37–51.

Saussure, F. de, 1916. Cours de linguistique générale. Payot, Paris.

Trubetzkoy, N.S., 1939. Grundzüge der Phonologie. Travaux du Cercle linguistique de Prague 7.

Vaux, B., Nevins, A., 2008. Rules, Constraints, and Phonological Phenomena. Oxford University Press, Oxford.

Photography as a Medium

Karin Becker, Konstfack College of Art, Craft and Design, Stockholm, Sweden

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11395–11402, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

For over a century and a half, photography has been the most popular and widespread means of making images. The first and primary tool of photography is the camera, and changes in the medium and its applications are often attributed to technical developments in making, reproducing, and distributing photographs. Equally if not more significant, however, have been changes in the ways people think about photography, including ideas about what and how photographs represent, and the roles they have played in social, political, and economic life. This article examines developments and conflicts in the idea of photography within a historical framework of changing cultural practice.

Photography's Origins

The invention of perspective in the fifteenth century laid the foundation for the system of visual representation that became photographic. The camera obscura, used by scientists, artists, and magicians since the Renaissance, is the most important precursor to the camera. By the eighteenth century this device had evolved from a large clumsy apparatus for projecting an image onto a plane surface, to a more portable instrument, used for the curiosity and pleasure of looking at images of all sorts of objects and scenes projected by the lens.

A knowledge of light-sensitive materials was popularly available in Europe by the 1720s, and one of the riddles of photography is why nearly a hundred years passed before this knowledge was applied to fixing the camera obscura's image. Despite a broad fascination with display forms, such as the diorama, that created illusions of realistic representations of nature, the idea of the photograph as a permanent copy apparently did not exist. Suddenly, in the 1820s a range of experimenters began simultaneously to compete to be the first to successfully fix the photographic image; the desire to photograph had emerged.

Photography's invention is generally dated to 1839, when breakthroughs were announced nearly simultaneously in France and England. Louis Jacques Mandé Daguerre's induplicable monochrome on a silver-coated copper plate, and William Henry Fox Talbot's reversed image – the negative – were the first successfully publicized and patented processes. The daguerreotype enjoyed an immediate popularity, setting off a craze which by the end of the year the French press had named '*daguerréotypomanie*.' Talbot's calotype system, on the other hand, based on a negative image that could be used to print a potentially infinite number of positives in as many different sizes, laid the foundation for all subsequent developments in photography. The refinement and diffusion of these two systems, alongside many competitors, had an immediate impact on European social life and spread quickly to North America and colonial outposts.

Photographs appeared in many display forms, as cartes-de-visites and stereographs, in albums and as lantern slides, and photography surfaced as a theme in popular music, on the stage, and in artifacts ranging from home furnishings to fashion accessories. Photography as an *idea* permeated mid-century European popular culture.

Early Applications of the Medium

At the same time that photographic tools and processes were being simplified, an agenda was being established for how the new medium was to be used. This included the people, places, and events represented in photographs and, equally important, who was taking the pictures and who were the intended viewers.

Photographs of People

Within months of Daguerre's announcement, professional portrait studios had sprung up in Paris, and soon studios were established in other major European cities, in New York, and in Washington. From 1860 onward, portraiture expanded to provincial towns, to less industrialized European countries, and to Australia, China, Japan, Mexico, and South America.

The photographic portrait represented a democratization of the painted portrait, and was commissioned first for private use within a family. The rising status of certain photographers led to commissions from men in public life whose positions required an official portrait. Most portraits of individuals followed the strict visual conventions of the painted portrait, a head-and-shoulders composition, with the sitter's head slightly turned to avoid a direct gaze into the camera. Portraits of family groups were common, and a custom quickly developed of photographically commemorating important events such as weddings and graduations. Photographs of deceased relatives, lying peacefully in their caskets, were also highly valued.

In the 1860s, studios began to turn out millions of inexpensive *carte-de-visites*, full- or bust-length portraits of workers, tradespeople, members of the bourgeoisie and aristocracy, native people of the colonies. These mass-produced images were often instrumental in establishing the status of a politician or celebrity, and both the French and British royal families permitted the sales of their images as *cartes*. Through its broad

distribution, the *carte-de-visite* also contributed to forming Western ideals of feminine beauty and fashion, as well as ideas about people from the nonindustrialized parts of the world as exotic and other. The *carte* craze further generated a surrounding industry of albums and holders for cataloguing and exhibiting the growing image collections.

In 1888 George Eastman launched the first Kodak camera, with the slogan 'You push the button, we do the rest,' and immediately opened the practice of photography to a much broader segment of the population. Studio portraiture declined as amateurs began to photograph family and friends. The repertoire of photographic events was also expanded as picture-taking became associated with holidays, family celebrations, and vacation trips. Photographic practice had entered the domestic domain, forming a bridge between personal memory and public history. A hundred years after the first Kodak camera, photography was fully integrated into family social life. In the mid-1980s, for example, over 90 percent of US families owned a still camera and amateurs were taking more than 11 billion photographs a year (Chalfen, 1987). The majority of these pictures were of family members and friends and were rarely seen outside that social group. Nonprofessionals' private photography today accounts for the largest single application of the medium. Studies have shown that people attach enormous value to these images, and losing photographs of family and friends, for example, through fire or war, is often experienced as a greater tragedy than the loss of other material possessions.

Distant Worlds

The early years of photography coincided with the period of colonial expansion and exploration of newly won territory for the industrializing nations of Europe and North America. Wonders of nature, the dynamic new cityscape and its architecture, and the exotic cultures of distant lands held a fascination and attraction for many. The camera was an ideal tool for showing what was new or previously unknown to an audience eager to marvel over sights they could not see with their own eyes. Photography also opened up new territories for tourism by people who, having seen a place in pictures, now had the desire and means to experience it themselves. A new aesthetic of the landscape also stimulated tourism, as for example when photographs of the rocky Scandinavian coasts brought British and German tourists north in the 1870s (Rosenblum, 1984: p. 112).

Photographs of popular European tourist sites, such as the city of Rome and its ruins, were marketed as stereographs. Distant lands of North Africa and the Near East became accessible to the armchair tourists of Europe as early as the 1850s through Francis Frith's photographs. Frith's *Egypt and Palestine Photographed and Described* (1858) was the first of 11 volumes and three sets of stereographs (Rosenblum, 1984: p. 122). His work was motivated in part by a concern over growing French influence in the region, and must be seen in the broader context of British expansionist goals of the period. Local Egyptian photographers also opened studios and marketed their work to the growing European fascination with views of the ruins and landscape of the Near East. Similar patterns developed in India and parts of the Far East. By 1884, for example, several thousand native photographers were in business in China, serving both local clients and the Western trade.

In North America, photography was linked to Western expansion. Photographers accompanied survey teams in order to document areas for future mineral exploitation, rail routes, and civilian settlement, at the same time that they were contributing to the visual discourse of the American frontier and its native peoples. The natural wonders of the wilderness, including Niagara Falls, were immensely popular stereographic themes, used in lantern slide shows and collected in albums. This iconography and the photographers who produced it, such as Carleton Watkins and William Henry Jackson, established the grandeur of the mountain landscape as a symbol of transcendent idealism. Their work also contributed to lobbying efforts to establish a national park system in order to save the pristine wilderness from commercial exploitation by private interests.

Nor did the city escape the camera's gaze. In contrast to images of untouched wilderness and ancient unchanging ways of life, here the focus was on physical transformation. New transport systems, bridges, and buildings made possible by new technologies and materials were transforming the city into a modern metropolis. Urban expansion was a source of national pride, and both governments and private interests supported photographic documentation of the major cities of Europe and North America. Cityscapes in panorama format were especially popular, a parallel to the painted panoramas and dioramas that had emerged as mid-nineteenth century forms of entertainment.

Photography's prominent and various positions in the emergent discourse of modernity was never more evident than in the world expositions, the late-nineteenth century spectacles of science and marketing, combining knowledge and entertainment. Beginning with the Crystal Palace in London in 1851 and followed every few years in Britain, France, and the United States, these major events both symbolized and displayed industrial expansion, made possible by new technologies and materials. Displayed side by side with the symbols of modernity were the exotic products and peoples of the underdeveloped nations, the colonies. At this point in the medium's development, photography provided an unproblematized means of expanding knowledge of the visible world, and served 'the general urge in all industrialized societies to measure, describe, and picture the physical substances of all things on earth and in the heavens' (Rosenblum, 1984: p. 140).

Concepts of the Image

Photography's popularity as a democratic form of self-documentation, intertwined with colonial interests and industrial expansion, point to the camera as a tool that linked seeing and knowledge in new ways. Within the sciences, in the arts, and in social reform movements, distinct and often conflicting visual discourses have developed around the photograph as a source of insight and knowledge. We turn now to the ways of seeing with the camera that were generated within each of these broad fields.

Photography and Science

Within the discourses of science, photography was to be a witness to scientific observation with greater accuracy than the human eye was capable of. In medicine, photography provided close visual description of the human body, and was an aid in documentation and diagnosis of medical problems and also psychological reactions and mental aberrations. Examples are portrait photographer Professor John William Draper's book *Human Physiology* (1856), illustrated with woodcuts based on photographs and Dr. Hugh Welch Diamond, also known for his pictorial art photography, who documented women inmates of the Surrey County Asylum. Charles Darwin's *The Expression of Emotions in Man and Animals* (1872) was illustrated with photographs from a variety of sources. Microscope lenses were developed and the discovery of X-rays by Conrad Wilhelm Roentgen (the 1901 Nobel Prize recipient) led to their immediate use as camera images in medical diagnosis. Motion studies and stop-motion experimentation depended heavily on photographic documentation. Eadward Muybridge's 100 000 images analyzing movements involved in animal locomotion and a range of human activities (including walking, curtseying, and bricklaying) provided a new if contested way of seeing that also influenced the arts, in particular Cubist and Futurist painters.

Taxonomy was a unifying concept of the nineteenth century scientific paradigm, and studies of the human species were formed by the delineation and elaboration of physical and cultural distinctions. Anthropology became a dominant discourse of explanation, based on historical and biological notions of culture, and relying heavily on comparative taxonomic methods in order to systematically catalogue the world's peoples (Edwards, 1997). Photography was central to the anthropometric systems developed to produce comparable data for analysis and classification. The system used by the distinguished British biologist Thomas H. Huxley for the Colonial Office, for example, involved posing individuals naked with a measuring rod in order to produce 'a systematic series of photographs of the various races of men comprehended within the British Empire' (cited in Edwards, 1997: p. 56). These studies of human physiogomy, used to expand evolutionary theory, contained a strong moral dimension, concerned with both creating and sustaining the complex system of social knowledge in which the images operated. Variations on the method were applied in hospitals, prisons, and other institutions where photographs of inmates were used to establish and classify social types.

The 1870s and 1880s also brought about an awareness of the contributions photography could make to anthropological fieldwork, by Edward B. Tylor and others. Yet there are few examples of systematic photographic documentation in ethnographic field studies. Early photography of the ways of life of diverse folk groups was most often carried out by missionaries and educators or by artists working in the pictorial tradition. Edward Curtis, Adam Clark Vroman, Robert Flaherty, and Frances Flaherty documented native American tribal life in its positive aspects, romanticizing in particular the close relationship to nature. Nor was it uncommon for ethnographic photographs from the turn of the century of peasant customs made in a pictorialist style to be used to advance the cause of nationalism, as in several Eastern European countries. The most common reference to an ethnographic photography worth emulating is the classic study of *Balinese Character* (1942) by Margaret Mead and Gregory Bateson, as the 'first saturated photographic research in another culture' (Collier and Collier, 1986: p. 12). Visual anthropology at mid-century was more closely associated with documentary film than with photography. Ethnographic photography survived as an *ideal*, however, despite the general absence of large-scale systematic studies. The British 'Mass-Observation' project in the mill towns of northern England in the 1930s, for example, was designed as an 'objective documentation' of working class life, in the manner of an anthropological study.

The archive has worked hand in hand with the scientific endeavors of accumulating and systematizing evidence. Seen as a repository of visual knowledge, the archive is related to the ideal of a comprehensive collection of images that together represent the whole (other examples being the family album or collection of *carte-de-visites*). The structure of the archive and its classification system was intended to mirror the physical and social world that the collection represented. Following Michel Foucault's groundbreaking archeology of the human sciences (1970), museums of natural and cultural history, as well as police, medical, and other archives, can no longer be seen as disinterested institutions that sustain and perpetuate value-free systems of knowledge, but reflect instead the structures of political and economic interests that gave rise to them. The photograph's evidentiary authority is thus inseparable from the archive as a repository of knowledge and power.

Photography and Art

Contrary to the idea of the photograph as an objective visual record, it has been argued that photography is inseparable from the Western pictorial tradition. Particularly in its relationship to landscape and nature, photography employed a set of conventions and aesthetic standards theorized by the European concept of the 'picturesque.' Emulating themes from the fine arts, many early photographers developed techniques that undermined the sharp literalness of the lens in their aspirations to make images that were at once truthful, beautiful, and inspirational. Exhibiting esthetic photographs in an appropriate context was a paramount goal of the turn-of-the-century Secessionist movements in Europe and the United States, and was important to establishing photography's status as art. Pictorialism in photography was a synthesis of the realism of the camera image, with frequently formulaic romantic compositions of peasant or working class life.

A complex and uneasy relationship between photography and the fine arts has continued, even as the repertoire of art photography has expanded. This is generally beyond the scope of this article, apart from several developments relevant to photography's relationship to the social sciences. On the one hand, themes drawn from peasant and working class life as well as of non-Western people integrated scientific inquiry with romantic images of an exotic or primitive Other as legitimate subjects of photographic art. On the other hand, attempts in the 1920s to define and explore the characteristics which distinguish photography from other visual arts laid the foundation for a more radical 'straight' photography, called New

Realism or *Neue Sachlichkeit* in Europe and North America. The nonmanipulated or spontaneous photograph became a means of exploring and exposing familiar subjects in new ways. These different schools and genres of art photography became more or less systematized as museums began to add photographs to their collections. The Museum of Modern Art in New York is a noteworthy example of a leading museum where a series of influential curators established a vocabulary of photographic practice based on the premises of modernism. By the 1940s, a photographic canon was in place, including a history of art photography with specific artists, works, and genres represented in the museum's collections (Newhall, 1949). Of particular importance to the social sciences was the respect this canon established for the genre of documentary photography as a legitimate form of esthetic practice.

The Photographer as Engaged Observer

Parallel to their idealized representations of European peasants, from the 1850s onward artists began to be concerned with the social and ethical consequences of industrialization, preparing the way for documentary photography's role in social reform campaigns. 'Before' and 'after' images became a staple of reform photography in the 1890s, as evidence of the positive effect the reform school or institution had on its residents. Photographs from the slums of New York (by Jacob Riis) and of child labor conditions (by Lewis Hine) were used to raise the fears and sway the consciences of politicians and a middle-class public. The photography of social reform was closely linked to new means of dissemination: the half-tone process for mass-producing photographs had been developed in the 1890s and a range of new periodicals arose in its wake.

Alongside the popular press' increased use of photographs after the turn of the century, social reform journals were among the first to recognize photography's ability to advocate a cause. Here the American *Charities and the Commons* is noteworthy. The *American Journal of Sociology* was also using photographs to illustrate articles, for example, on housing conditions. This visual material, often by women who held administrative positions in social action agencies, later declined under Albion Small's editorship, as the journal aligned itself with sociology's move to establish itself as a scientific discipline (Stasz, 1979).

By the 1920s, documentary photography, as the genre came to be known, was recognized as an effective means of drawing attention to social issues. The best known example of such an effort was the photographers who worked for the US Farm Security Administration (FSA) under the leadership of Roy E. Stryker in the late 1930s, documenting the effects of the depression on rural Americans and the need for government aid. In other cases, photographers used the documentary form to draw attention to marginalized social groups, occasionally drawing on their own experience as members of the culture they documented. One of the clearest ideological expressions of photographers' identity was the Worker Photography movement, most active in Germany between World Wars I and II. The International Communist Party, convinced that photographs unavoidably reflect the class of the person holding the camera, mobilized workers in a variety of documentary projects and publications. In direct contradiction to a claim of impartiality, this photography based its authority on the photographer's membership in the group. It can thus be seen as a precursor to anthropology's interest in indigenous photographers' images as expressions of their specific cultural perspectives.

The authenticity of the documentary style was based on a complex set of factors. Seldom was the work based on a single photograph, but instead was built up through a series of images witnessing the photographer's proximity to the subject over an extended time. Despite many exceptions, the images also tend to follow a specific visual code. Straightforward photographs of people, taken in their everyday environment under available light conditions, are the most typical. If the subject appears to be aware of the camera, he or she is usually gazing directly into the lens. The photograph's claim to truth is based therefore on the seemingly contradictory appeals of the photographer's simultaneous engagement with and objective distance from the subject.

Photography and Mass Society

Closely related to the ways photography has portrayed its subjects is the effect of images on those who see them, the audience. Photography was a mass medium almost from its inception, as the idea of the photograph and the ways it was marketed had a broad and immediate social and cultural impact. The 'very insatiability of the photographing eye' and the 'omnipresence of photographed images' position the medium in a vortex of complex moral, ethical, political, and esthetic issues. Sontag's influential book of essays *On Photography* (1973) brought many of these issues up for debate. Yet the exact nature of the photograph's appeal and its effect on its viewers has remained elusive. Despite extensive research on photography's impact on attitudes and behavior, results have been either inconclusive or too context-specific (regarding the viewer's age and background or the viewing context) to support any general conclusions. What is evident, instead, is a widespread belief in the photograph's influence which has in turn led to specific applications of the medium based on conceptions of the nature of photographic communication as a highly unstable synthesis of factual information and seductive attraction.

From the 1850s photographs began to be used alongside other illustrations in some magazines, and the widespread application of the half-tone process in the 1890s coincided with the emergence of new audiences in Europe and the United States. Photography was thought to appeal in particular to people who were either nonliterate or had not yet mastered the language of their new country. A new inexpensive illustrated press arose that used photography in often sensational ways, and photographs began to appear in advertising. Political campaigns made extensive use of photography from the 1920s onward, in documentary forms described above, but also in photomontage and posters. The reticence of many newspapers to make extensive use of photography (for example, the morning newspapers 'of record' in New York and London) is an exception which nevertheless rests on the same notion of the photograph's communicative power. The photograph, if seen as an unmediated copy of nature, does not require language or reason to be understood, but appeals instead directly to the emotions. Newspapers and other media that wished to base their appeal

on rationality excluded photographs from their pages. Even if an image were manipulated through technical or other means, the message that the audience saw or 'received' was accepted as representing reality.

The model of mass communication that emerged first in the United States and dominated research for several decades was based on the notion of communication as a successful 'transmission' of a message from an active and powerful source to a passive receiver. The origins of the model and the research it generated can also be traced to growing concern over the power of the media, particularly in the service of political propaganda and the rise of fascism in Europe after World War I. Although only sporadically applied in research on photography, the transmission model of communication served as the underlying framework of explanation of photography's influence on its audience. The power of images is assumed to be greater, for example on the relatively unformed psyches of children and youth. 'Moral panics' continue to pursue the image and its influences, giving rise to research on Hollywood films in the 1930s, on comic books in the 1940s, on television's effects on children in the 1950s and 1960s, and video games at the turn of the century.

Countering the transmission model of photography's influence on its audience is the perspective that considers photography as a cultural practice and phenomenon which participates in the construction of power and meaning through the mutual and intertextual establishment of social and political institutions and forms of visual expression. The cultural perspective views the audience not as passive receivers, but as members of a (sub)culture who employ cultural knowledge as they produce, consume, and interpret photographs. The cultural perspective does not deny the power of the media, but rather sets it in a historical, social, or political context, and examines how the media use photographs to address or position the viewer in specific relationships to media content. The perspective shifts the emphasis from an inquiry into photography's impact on individual or group attitudes and behavior to an examination of the ways photography is used in the construction of meaning, often focusing on social values and the use of stereotypes. The medium has been decisive, for example, in the production of visual representations of gender and gender relations, as revealed in Erving Goffman's study of advertising (1979). Other studies have demonstrated how media images of women not only establish standards of feminine beauty, but also link the image of the female body to products and services in a process of commodification that objectifies all women (cf Williamson, 1978; Irigaray, 1985; Bordo, 1993; Solomon-Godeau, 1991). Other analyses have provided insights into the formation and perpetuation of stereotypes of race, in advertising but also in the news media, with groundbreaking work carried out at the Centre for Contemporary Cultural Studies in Britain in the 1970s.

Issues of media power and influence are also linked to photography through a growing concern over the image's role in surveillance. Cameras, in addition to their continually expanding position in private life, are used increasingly in both public and commercial spaces for the ostensible purpose of increasing security and preventing crime. Yet the images taken by these cameras also raise issues of the photograph's potential for restricting, documenting, and controlling individual behavior. As technology extends the capacity to send images over global networks, the specter is raised of expanding spheres of media influence and control that transcend the nation state. At the same time, the interactive aspects of many contemporary media raise the possibility of individual and group autonomy from transnational media content, and the ability of the individual to form and reform images that are sent and received. The cultural impact of these relatively new image technologies is a current topic of debate, particularly with regard to the moral and ethical implications of their distribution and use.

Technology and Revisions of Photographic Meaning

A central question throughout photography's history is whether the meaning of the photograph is a function of the medium *per se* and its technical, chemical, and optical relationship to the subject in front of the camera, or is formed instead through the culture which surrounds it. Benjamin described how photographic reproduction and mass distribution irrevocably altered the individual's encounter with the original, following a more general argument developed within the Frankfort School over the hegemonic power of cultural production in mass society. Yet it was not until the 1970s that a number of theorists began to address the specific problem of photographic meaning, particularly issues related to what and how the photograph represents. Roland Barthes (1977) used a semiotic analysis to unlock the problem of 'the photographic paradox': that is, the photograph represents an analogic reproduction of reality or a message without a code, and simultaneously a connotation or coded message, representing what society thinks about the subject. The major task of the denotative aspect of the image, according to Barthes, is to naturalize the cultural, that is, to transform cultural values into a message that can be 'read' as self-evident, as nature.

The deconstruction of the camera as a cultural artifact and of photography as deeply informed by culture continued on several parallel fronts. The model of the observer that dominates Western science, defined by the Cartesian perspectivalism and the direct gaze, is encoded in the optical system of the camera. The model thus defines a particular relationship between seeing and knowledge that privileges the authority of this gaze and its representations. This 'way of seeing' was critiqued most forcefully in relation to documentary photography for its dehistoricized view of marginalized social groups, a view that in turn established a visual code for representing the Other (Bolton, 1989). Because photography has been constructed as a predominantly male practice, reinforced by the observer's gaze as male, female also becomes Other. Extensions of this critique to the uses of photography in the social sciences have implicated anthropology in particular for constructing its subjects as visually exotic. Related to this phenomenon, as mentioned above, is the institutional gaze, embodied and perpetuated through the archive, the museum, or the press (Tagg, 1988).

In the 1970s, anthropologists also began to criticize the use of film and photography in fieldwork. Instead of being informed by anthropological concepts, the Western anthropologist's visual records reflected 'his' own culture, coded through the practices of family photography. Within the

subfields of visual anthropology and sociology in the United States and Britain and their influential journals, this insight was variously developed into a broader perspective on photography, film, and media in general as sources for studying the cultural values and perspectives of those who create the images. This in turn supported a shift from image analysis to the study of the social and cultural practices in which photography is embedded. This complemented the critique being directed at the field of anthropology as a whole for perpetuating its colonial heritage through techniques of 'writing culture' that represented its subjects as Other (Ruby, 1982). Anthropology's continuing interest in the photography of indigenous peoples and other subcultural groups reflects on the one hand a recognition of photography as a significant source for cultural analysis and on the other, attempts by the field to integrate the anthropological subject's own understandings into research and reports. Current ethnographic studies of media practices frequently include investigations of how images are interpreted and used by ordinary people. The photographic practices of professionals has been largely neglected, however, apart from the many biographies of well-known photographers and analyses of their work.

Digital techniques of making, reproducing, and transmitting photographs have contributed to undermining the authority of the photograph as analog or direct copy of reality. In the mid-1980s, the digitized image was seen by many, particularly within the press, as a direct challenge to established beliefs about how the photograph represents. Gradually the technique has been integrated into professional and private practices of making and distributing images, stimulating the rise of new media networks. The immediacy of image transfer and the portability of the technology have vastly increased the number of photographs in circulation, including those depicting war and human suffering. Some analysts argue that this phenomenon results in 'compassion fatigue,' reducing the moral impact of the image (Moeller, 1999). In the meantime, new applications of interactive image technologies are emerging in the arts, in youth culture, and in diasporic communities. The digital transfer of images globally has given rise in some quarters to a utopian vision with the photograph as a cornerstone of a global visual culture. More realistic analyses point to the diversities of culture worldwide and the continued vast inequalities in access to these technologies as critical impediments to a hegemonic visual culture.

Editor's Note

Since the original publication of this entry, digital photography has replaced traditional film images almost completely and photographic software for the manipulation of digital images (e.g., Photoshop) has become a cultural commonplace. A fine and current study of digital photography is Lister M., 2013. On the larger aesthetic of the photographic medium in the digital age, see Pirenne and Streitberger, 2013. The original entry provides an in-depth discussion of the uses of photography in anthropological research. A fascinating contemporary application of the medium in that discipline is Filip de Boeck et al., 2014.

See also: Mass Communication: Normative Frameworks; Mass Media, Representations in; Mass Media: Introduction and Schools of Thought; Visual Images in the Media; Visualization in Science and Technology.

Bibliography

Barthes, R., 1977. Image Music Text. Hill and Wang, New York.
Bateson, G., Mead, M., 1942. Balinese Character: A Photographic Analysis, 2. Special Publications of the New York Academy of Sciences, New York.
Benjamin, W., 1968. The work of art in the age of mechanical reproduction. In: Arendt, H. (Ed.), Illuminations. Walter Benjamin. Schocken, New York.
Bolton, R. (Ed.), 1989. The Contest of Meaning: Critical Histories of Photography. MIT Press, Cambridge, MA.
Bordo, S., 1993. Unbearable Weight: Feminism, Western Culture, and the Body. University of California Press, Berkeley, CA.
Chalfen, R., 1987. Snapshot Versions of Life. Bowling Green State University Popular Press, Bowling Green, OH.
Collier Jr, J., Collier, M., 1986. Visual Anthropology. Photography as a Research Method. University of New Mexico Press, Albuquerque, NM.
Edwards, E. (Ed.), 1992. Anthropology and Photography, 1860–1920. Yale University Press, New Haven, CT.
Edwards, E., 1997. Ordering others: photography, anthropologies and taxonomies. In: Roberts, R. (Ed.), In Visible Light. Museum of Modern Art, Oxford, UK.
Filip de Boeck, Plissart, Marie-Francoise (Eds.), 2014. Kinshasa: Tales of the Invisible City. Leuven University Press, Belgium.
Goffman, E., 1979. Gender Advertisements. Harper & Row, New York.
Irigaray, L., 1985. This Sex which is not One. Cornell University Press, Ithaca, NY.
Lister, M. (Ed.), 2013. The Photographic Image in Digital Culture, second ed. Routledge, London.
Moeller, S., 1999. Compassion Fatigue. How the Media Sell Disease, Famine, War and Death. Routledge, London.
Newhall, B., 1949/1982. The History of Photography: from 1839 to the Present. Museum of Modern Art, New York.
Pirenne, R., Streitberger, A. (Eds.), 2013. Heterogeneous Objects: Intermedia and Photography After Modernism. Leuven University Press, Belgium.
Rosenblum, N., 1984. A World History of Photography. Abbeville Press, New York.
Ruby, J. (Ed.), 1982. A Crack in the Mirror. Reflexive Perspectives in Anthropology. University of Pennsylvania Press, Philadelphia, PA.
Solomon-Godeau, S., 1991. Photography at the Dock: Essays on Photographic History, Institutions, and Practices. University of Minnesota Press, Minneapolis, MN.
Sontag, S., 1973. On Photography. Farrar, Straus and Giroux, London.
Stasz, C., 1979. The Early History of Visual Sociology. In: Wagner, J. (Ed.), Images of Information. Sage, Beverly Hills, pp. 119–136.
Tagg, J., 1988. The Burden of Representation. Essays on Photographies and Histories. University of Massachusetts Press, Amherst, MA.
Williamson, J., 1978. Decoding Advertisements: Ideology and Meaning in Advertising. Marion Boyars, London.

Physical Activity and Health

Reinhard Fuchs, University of Freiburg, Freiburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article summarizes findings on the relationship between physical activity and various aspects of physical and mental health. Persons who engage in regular physical activity have a lower overall mortality rate as well as a reduced risk of developing coronary heart disease, stroke, colon cancer, breast cancer, diabetes, arthritis, osteoporosis, and obesity. Depression-reducing effects of physical exercise have been found in correlational as well as clinical studies and can be regarded as well confirmed. Regarding anxiety, acute exercise leads to a reduction in state anxiety that lasts for about 1–1.5 h after the exercise is terminated. Finally, evidence from randomized studies suggests that regular physical activity (chronic exercise) improves symptoms in patients suffering from anxiety disorders.

Introduction

The past decades have witnessed a burgeoning of research examining the impact of physical activity on attaining optimal health and preventing disease. Today, overwhelming evidence supports the benefits of exercise for various aspects of physical health. However, exercise has also been shown to be beneficial for maintaining or improving mental health. Unfortunately, our knowledge of the mental health benefits of exercise is still rather limited in comparison to what we know about its physical health benefits. This article provides an overview of the major findings regarding the physical and mental health effects of physical activity.

The Effects of Physical Activity on Physical Health

Overall Mortality

Epidemiological studies suggest that persons who engage in the recommended amounts of physical activity (150 min per week in moderate-intensity activity, 75 min per week in vigorous-intensity activity, or a combination of the two intensities) have overall mortality rates (mortality from all causes) that are 20–30% lower than those of persons who do not reach these amounts (PAGAC, 2008). This finding is supported by longitudinal research examining how changes in the level of cardiorespiratory fitness may affect subsequent mortality risks. For example, in a study by Blair et al. (1995), healthy men with an initially low level of cardiorespiratory fitness participated in two exercise tests about 5 years apart. 5 years after the second exercise test, men in the bottom fifth of the initial fitness distribution who improved their fitness to at least a moderate level had a 44% lower mortality rate than men who remained in the bottom fifth. Similar results had already been found by Paffenbarger et al. (1993) in a cohort of initially sedentary Harvard alumni. The men in this study who started and maintained moderately intense activities extended their longevity by 0.72 years on average.

Cardiovascular Diseases

Cardiovascular diseases (CVDs), in particular coronary heart disease (CHD) and stroke, are major causes of death, disability, and health care expenditures in Western societies. There is clear and strong evidence that, within populations, the incidence of CVD (including CHD, stroke, hypertension, rheumatic fever, congenital heart defects, congestive heart failure, and peripheral vascular disease) is inversely related to habitual physical activity levels (Bauman and Owen, 1999; Janssen, 2012; PAGAC, 2008). Some of the studies in this field demonstrated a dose–response gradient between level of physical activity and risk of CVD outcome. Dose–response relationships were also found in investigations in which measures of cardiovascular fitness, instead of physical activity estimates, were related to CVD mortality.

Coronary Heart Disease

One of the first studies to systematically analyze the association between physical activity and risk of CHD was conducted by Morris et al. (1953), who examined 31 000 employees of the London Transport Executive. A comparison between bus drivers and bus conductors revealed that physically active conductors clearly had a lower risk of a 'first coronary episode' than sedentary drivers. Classifying physically active and inactive people according to job title or occupational activity is typical for studies published before 1978. In the following years, level of physical activity was defined mostly by self-report measures of leisure-time activity or by indices that combined levels of occupational and leisure-time activities. More recent investigations have also emphasized the necessity to control for potentially confounding variables (e.g., age, sex, functional status) and to adjust for factors that are considered causal links between physical activity and CHD, such as serum cholesterol, blood pressure, diabetes, or body mass index.

A vast body of scientific literature indicates that physical activity is causally linked to a reduced risk of CHD. A review conducted on 30 studies published since 1996 suggests that the risk of developing CHD is 19% lower in moderately active (3–6 METs (metabolic equivalents)) men and even 39% lower in vigorously active (>6 METs) men than in men with a low (<3 METs) activity level (PAGAC, 2008). Furthermore, there is strong evidence that this relationship holds true for both men and women. For example, findings from a prospective study among 72 488 female nurses who were 40–65 years old at the beginning of the study suggest that sedentary women who became active in middle adulthood or later had a lower risk of

coronary diseases than their counterparts who remained sedentary. Relative to women who walked infrequently, those who walked at a brisk pace three or more hours per week had a multivariate relative risk (RR) of 0.65 for coronary events (Manson et al., 1999). Physical exertion may temporarily increase the likelihood of an acute coronary event among persons with advanced coronary atherosclerosis who are not used to regular exercise. Although this risk needs to be taken seriously, it does not affect the general fact that physical activity is associated with a substantially lower overall risk for CHD.

Stroke

Today, we know that low physical activity and low cardiorespiratory fitness are also linked to an increased risk of cerebrovascular disease (stroke). Evidence suggests that the protective effects of physical activity and fitness emerge for both ischemic and hemorrhagic strokes. A literature review of 19 studies published since 1996 suggests that the risk of suffering a stroke is roughly 30% lower in moderately (3–6 METs) and vigorously (>6 METs) active men and around 25% lower in moderately and vigorously active women than in their low (<3 METs) active counterparts (PAGAC, 2008).

Biological Mechanisms

Several biological mechanisms may be responsible for the beneficial effects of physical activity on cardiovascular risks. These mechanisms involve the impact on blood pressure, atherosclerosis, plasma lipid/lipoprotein profile, availability of oxygenated blood for heart muscle needs (ischemia), thrombosis, heart rhythm disturbances, overweight, and healthier distribution of body fat.

Cancer

Cancer is one of the leading causes of death in industrialized countries. It is predicted that this proportion will increase in the coming years. Numerous studies have examined the role of physical activity in the prevention of cancer (Sternfeld and Lee, 2009). The relationship between physical activity and reduced cancer risk has been found most consistently for colon and breast cancer (PAGAC, 2008). There is some evidence supporting an association between physical activity and risk for lung cancer (Tardon et al., 2005) and prostate cancer (Lee, 2012).

Colon Cancer

Systematic literature analyses reveal that colon cancer is probably the most often-studied cancer in the field of physical exercise epidemiology. More than 50 studies (cohort and case–control studies) on colon cancer have been published. In sum, these studies indicate that physically active (or fit) persons have a markedly lower risk of developing colon cancer than inactive persons. A meta-analysis by Wolin et al. (2009) reported a risk reduction for colon cancer of 24% for physically active individuals (RR = 0.76 with a 95% confidence interval of 0.72–0.81). Analyses with at least three levels of physical activity suggest the existence of a dose–response relationship: Individuals with the highest level of activity had the lowest risk of colon cancer, those with the lowest level of activity had the highest risk, and those in between showed a risk that was also in between. Such dose–response relationships are a strong argument in support of the hypothesis that physical activity and colon cancer are causally linked. As a biological explanation for the association between physical activity and colon cancer, it has been proposed that physical activity reduces gastrointestinal transit time and thus diminishes the contact duration between the colon mucosa and potential carcinogens in the fecal stream.

Breast Cancer

A literature review by Friedenreich and Cust (2008) included the results of 87 studies on physical activity and the incidence of breast cancer. Overall, the data are fairly consistent with the view that physical activity and the risk of developing breast cancer are inversely connected. The authors reported that the average risk of breast cancer was 25–30% lower in the most physically active group than in the inactive group. A meta-analysis by Monninkhof et al. (2007) yielded a smaller risk reduction for physically active women: Here, the RR was calculated with RR = 0.81 (95% confidence interval: 0.73–0.89), meaning that physically active women had a 19% reduction in the risk of developing breast cancer compared to their inactive counterparts. Dose–response analyses suggest that females need to exercise on a level that is at least equivalent to brisk walking for a minimum of 45–60 min day^{-1} on 5 or 6 days per week (McTiernan, 2008).

Physical Activity among Cancer Patients

The results reported in this overview focus on physical activity (or fitness) in the primary prevention of cancer. In recent years, several studies have been conducted, which look at the effects of physical activity in the treatment and rehabilitation of cancer patients (Sternfeld and Lee, 2009). There is some evidence that being physically active after being diagnosed with cancer may be beneficial to cancer treatment and the patients' quality of life (Speck et al., 2010).

Further Chronic Diseases

Level of physical activity has been related to further chronic diseases and conditions, in particular to diabetes, arthritis, osteoporosis, and obesity. Several prospective cohort studies lend strong support to the hypothesis that regular physical activity offers protection against the risk of non-insulin-dependent diabetes mellitus (Middelbeek and Goodyear, 2012). Reducing obesity through increased physical activity alone is a relatively slow process. Physical activity programs in combination with strategies of dietary management seem to be more effective in this regard (Ross and Janssen, 2012).

The Effects of Physical Activity on Mental Health

Numerous reviews and meta-analyses have summarized the literature dealing with the psychological effects of physical activity (Buckworth et al., 2013). In studies examining the effects of physical activity on mental health, the most frequently observed outcomes include depression, anxiety, mood, self-esteem, perceived stress, and cognitive functioning. In recent years, the focus of research has been extended to include the role

of physical activity in the development and treatment of mental disorders such as schizophrenia, dementia, personality disorders, or eating disorders. Subsequently, we focus on the associations of physical activity with depression and anxiety.

Depression

Results from cross-sectional studies suggest a substantial relationship between physical activity and reduced symptoms of depression in men and women (Buckworth et al., 2013). This relationship holds even when adjusted for potentially confounding factors. A summary of the existing literature led to the conclusion that physically inactive persons are twice as likely to have symptoms of depression than their physically active counterparts. The association between level of physical activity and depression has been found in the general population as well as in clinical subpopulations.

Cross-sectional studies leave the question unanswered whether physical activity has an effect on depression, or whether persons with good mental health are simply more likely to be active. In an attempt to resolve this problem of causality, cohort studies have been conducted in which the relationship between physical activity and depressive symptoms were observed under a longitudinal perspective. One of these cohort studies was conducted by Paffenbarger et al. (1994), which revealed that the level of physical exercise reported at an initial interview was predictive of physician-diagnosed depression approximately 25 years later. In this investigation with 10 201 participants, the RR of depressive symptoms was 27% lower for men who had reported spending more than 3 h engaging in physical activities each week than for men who had reported no such activities. However, in other population studies, the level of physical exercise at baseline was not a risk factor for incidence of depressive symptoms years later (Weyerer, 1992). Thus, the empirical evidence from cohort studies is still limited and inconsistent.

Another approach for resolving the problem of causality is provided by intervention studies conducted in community or laboratory settings. A classic and often-cited work in this field is an experimental study by McCann and Holmes (1984). Altogether, 43 women students who scored higher than the cutoff point on the Beck Depression Inventory were randomly assigned to one of three groups: (1) aerobic exercise group, (2) placebo group, and (3) no treatment group. After 5 weeks, the aerobic exercise group showed a significant reduction in depressive symptoms compared with the other two groups, supporting the hypothesis that increased physical activity results in a reduction in depression. More recently, Blumenthal et al. (2007) compared the efficacy of physical activity to that of pharmacotherapy on the basis of a randomized controlled intervention study. A total of 202 sedentary adults diagnosed with major depression were assigned to 16-week interventions of antidepressant medication (sertraline), placebo pills, home-based aerobic exercise, or supervised exercise. Improvement was defined as no longer meeting the clinical criteria for depression (i.e., remission). At the end of the intervention period, the remission rates were 47% for medication, 45% for supervised exercise, and 40% for home-based exercise, but only 31% for placebo pills, suggesting that the antidepressant effects of regular exercise were comparable to those of pharmacotherapy.

Results on the dosage of physical activity that is needed to elicit an antidepressant effect are provided by Dunn et al. (2005). In this study, 80 adults diagnosed with mild-to-moderate major depressive disorder were randomly assigned to five groups: Four groups participated in 12-week programs of aerobic activity that varied by energy expenditure (low vs high dosage) and frequency (3 vs 5 days a week), and one group (control condition) participated in a 12-week program of flexibility training three times a week. The reduction in depression was significantly greater for the high-energy expenditure condition ($17.5\ \text{kcal} \times \text{kg}^{-1} \times \text{week}^{-1}$) than for the low-energy expenditure condition ($7.0\ \text{kcal} \times \text{kg}^{-1} \times \text{week}^{-1}$) or the control condition (flexibility training), which both showed comparable levels of depression reduction. Whether the exercise was done three or five times per week (frequency factor) did not affect the level of depression. These results suggest that patients with major depressive disorder can be effectively treated with an aerobic exercise program which is based on an energy expenditure of about $17.5\ \text{kcal}\ \text{kg}^{-1}$ of body weight per week (this dosage corresponds to the public health recommendations for physical activity).

Several biological and psychosocial mechanisms have been proposed to explain the antidepressant effects of physical activity. In particular, it has been hypothesized that physical activity may induce changes in brain neuroreceptor concentrations of monoamines (norepinephrine, dopamine, or serotonin) or endogenous opiates (endorphins and enkephalins), which in turn are responsible for changes at the affective level. However, the observed depression-reducing effects of physical exercise may also be mediated by psychological factors such as exercise-induced feelings of mastery and self-efficacy, or by positive social and emotional experiences in the sport context (Buckworth et al., 2013).

Anxiety

Research also suggests that physical activity may be linked to a reduction in anxiety. The psychological assessment of anxiety usually includes state and trait measures. State measures refer to the actual feeling of anxiety, whereas trait measures evaluate the stable personality characteristic of anxiety. In physical activity research, state measures have been used in pre–post designs of intervention studies to explore whether a specific activity episode (acute exercise) has the potential to cause changes in currently experienced anxiety (state anxiety). Results are relatively unequivocal in this respect. Immediately after a session of physical activity, most people report a decrease in state anxiety, which is still detectable 1–1.5 h after the end of the training (Cox et al., 2004; Hale et al., 2002; Knapen et al., 2009). While the short-term effects of acute physical activity on state anxiety are well documented, the long-term effects of chronic (habitual) physical activity on trait anxiety are still subject to debate: A meta-analysis by Herring et al. 2010 confirmed the causal link between regular exercise and trait anxiety, whereas meta-analyses by Long and van Stavel (1995) and Schlicht (1994) did not find convincing evidence in support of this effect.

One of the best randomized studies to examine the therapeutic benefits of regular physical activity in clinically diagnosed anxiety patients is still that conducted by Broocks et al.

(1998). A total of 46 patients with panic disorders with or without agoraphobia were assigned to 10 weeks of (1) a moderate-running program, (2) an anxiolytic pharmacotherapy (clomipramine), or (3) a placebo pill 'treatment.' Compared to the placebo condition, both the running program and pharmacotherapy led to a significant decrease in anxiety symptoms. However, clomipramine improved anxiety symptoms earlier and more effectively than the running program (here it took longer to relieve symptoms). Despite this limitation, the results suggest that regular aerobic exercise alone has the potential to lead to a substantial clinical improvement in patients with panic disorder.

Conclusions

There is strong evidence that physical activity is significantly associated with reduced risk of all-cause mortality, CVD, colon and breast cancer, non-insulin-dependent diabetes, arthritis, osteoporosis, and obesity control. Findings also suggest that regular physical activity may play an important role in the prevention and treatment of mental diseases, including depression and anxiety. Less is known about the psychological and physiological mechanisms behind these antidepressant and anxiolytic effects. The recommended dose for improving physical health is 150 min week^{-1} of moderate-intensity physical activity (3–6 METs) or 75 min week^{-1} of high-intensity physical activity (>6 METs) or any equivalent combination of the two intensities (WHO, 2010). But how much physical activity is needed to improve or maintain psychological well being? The mental health benefits of physical activity are probably not only a matter of expended kilocalories (if at all), but also of activity-related social interactions and emotional experiences. Clarifying the mechanisms by which physical activity may affect mental health dimensions will be a challenge for sport and health psychology in the coming years.

See also: Cancer: Psychosocial Aspects; Coronary Heart Disease: Psychosocial Aspects; Diabetes: Psychosocial Aspects; Gender and Physical Health; Health Behavior, Psychosocial Theories of; Health Behaviors, Assessment of; Health Behaviors; Men's Health; Mental Health; Physical Activity and Health Promotion; Rehabilitation (Cardiac): Psychosocial Aspects; Self-Efficacy and Health; Weight Regulation.

Bibliography

Bauman, A., Owen, N., 1999. Physical activity of adult Australians: epidemiological evidence and potential strategies for health gain. Journal of Science and Medicine in Sport 2, 30–41.

Blair, S.N., Kohl, H., Barlow, C., Paffenbarger, R., Gibbons, L., Macera, C., 1995. Changes in physical fitness and all-cause mortality: a prospective study of healthy and unhealthy men. Journal of the American Medical Association 273, 1093–1098.

Blumenthal, J.A., Babyak, A., Doraiswamy, M., Watkins, L., Hoffmann, B., Barbour, K.A., et al., 2007. Exercise and pharmacotherapy in the treatment of major depressive disorder. Psychosomatic Medicine 69 (7), 587–596.

Buckworth, J., Dishman, R., O'Connor, P., Tomporowski, P., 2013. Exercise Psychology, second ed. Human Kinetics, Champaign, IL.

Broocks, A., Bandelow, B., Pekrun, G., George, A., Meyer, T., Bartmann, U., et al., 1998. Comparison of aerobic exercise, clomipramine, and placebo in the treatment of panic disorder. American Journal of Psychiatry 155, 603–609.

Cox, R., Thomas, T., Hinton, P., Donahue, O., 2004. Effects of acute 60 and 80% VO_2 max bouts of aerobic exercise on state anxiety of women of different age groups across time. Research Quarterly for Exercise & Sport 75, 165–175.

Dunn, A.L., Trivedi, M.H., Kampert, J.B., Clark, C.G., Chambliss, H., 2005. Exercise treatment for depression: efficacy and dose response. American Journal of Preventive Medicine 28, 1–8.

Friedenreich, C.M., Cust, A.E., 2008. Physical activity and breast cancer risk: impact of timing, type and dose of activity and population subgroup effects. British Journal of Sports Medicine 42, 636–647.

Hale, B.S., Koch, K.R., Raglin, J.S., 2002. State anxiety responses to 60 minutes of cross training. British Journal of Sports Medicine 36, 105–107.

Herring, M., O'Connor, P., Dishman, R., 2010. The effect of exercise training on anxiety symptoms among patients: a systematic review. Archives of Internal Medicine 170, 321–331.

Janssen, I., 2012. Physical activity, fitness, and cardiac, vascular, and pulmonary morbidities. In: Bouchard, C., Blair, S., Haskell, W. (Eds.), Physical Activity and Health, second ed. Human Kinetics, Champaign, IL, pp. 185–195.

Knapen, J., Sommerijns, E., Vancampfort, D., Sienaert, P., Pieters, G., Haake, P., et al., 2009. State anxiety and subjective well-being responses to acute bouts of aerobic exercise in patients with depressive and anxiety disorders. British Journal of Sports Medicine 43, 756–759.

Lee, I., 2012. Physical activity, fitness, and cancer. In: Bouchard, C., Blair, S., Haskell, W. (Eds.), Physical Activity and Health, second ed. Human Kinetics, Champaign, IL, pp. 231–244.

Long, B., van Stavel, R., 1995. Effects of exercise training on anxiety: a meta-analysis. Journal of Applied Sport Psychology 7, 167–189.

Manson, J., Hu, F., Rich-Edward, J., Coldit, G., Stampfer, M., Willett, W., et al., 1999. A prospective study of walking as compared with vigorous exercise in the prevention of coronary heart disease in women. New England Journal of Medicine 341, 650–658.

McCann, L., Holmes, D.S., 1984. Influence of aerobic exercise on depression. Journal of Personality and Social Psychology 46, 1142–1147.

McTiernan, A., 2008. Mechanisms linking physical activity with cancer. Nature Reviews Cancer 8, 205–211.

Middelbeek, R., Goodyear, L., 2012. Physical activity, fitness, and diabetes mellitus. In: Bouchard, C., Blair, S., Haskell, W. (Eds.), Physical Activity and Health, second ed. Human Kinetics, Champaign, IL, pp. 215–229.

Monninkhof, E.M., Elias, S.G., Vlems, F.A., van der Tweel, I., Schuit, A.J., Voskuil, D.W., van Leeuwen, F.E., 2007. Physical activity and breast cancer: a systematic review. Epidemiology 18, 137–157.

Morris, J.N., Heady, J., Raffle, P., Roberts, C., Parks, J., 1953. Coronary heart disease and physical activity of work. Lancet 2, 1111–1120.

Paffenbarger, R.S., Hyde, R., Wing, A., Lee, I., Jung, D., Kampert, J., 1993. The association of changes in physical activity level and other lifestyle characteristics with mortality among men. New England Journal of Medicine 328, 538–545.

Paffenbarger, R.S., Lee, I.-M., Leung, R., 1994. Physical activity and personal characteristics associated with depression and suicide in American college men. Acta Psychiatrica Scandinavica Supplementum 377, 16–22.

PAGAC – Physical Activity Guidelines Advisory Committee, 2008. Physical Activity Guidelines Advisory Committee Report, 2008. US Department of Health and Human Services, Washington, DC.

Ross, R., Janssen, I., 2012. Physical activity, fitness, and obesity. In: Bouchard, C., Blair, S., Haskell, W. (Eds.), Physical Activity and Health, second ed. Human Kinetics, Champaign, IL, pp. 197–214.

Schlicht, W., 1994. Does physical exercise reduce anxious emotions? A meta-analysis. Anxiety, Stress, and Coping 6, 275–288.

Speck, R.M., Courneya, K.S., Masse, L.C., Duval, S., Schmitz, K.H., 2010. An update of controlled physical activity trials in cancer survivors: a systematic review and meta analysis. Journal of Cancer Survivorship 4, 87–100.

Sternfeld, B., Lee, I., 2009. Physical activity and cancer: the evidence, the issue, and the challenges. In: Lee, I., Blair, S., Manson, J., Paffenbarger, R. (Eds.), Epidemiologic Methods in Physical Activity Studies. Oxford University Press, New York.

Tardon, A., Lee, W., Delgado-Rodriguez, M., Dosemeci, M., Albanes, D., Hoover, R., Blair, S., 2005. Leisure-time physical activity and lung cancer: a meta-analysis. Cancer Causes and Control 16, 389–397.

Weyerer, S., 1992. Physical activity and depression in the community: evidence from the upper Bavarian field study. International Journal of Sports Medicine 13, 492–496.

WHO, 2010. Global Recommendations on Physical Activity for Health. WHO Press, Geneva.

Wolin, K.Y., Yan, Y., Colditz, G.A., Lee, I.-M., 2009. Physical activity and colon cancer prevention: a meta-analysis. British Journal of Cancer 100, 611–616.

Physical Activity and Health Promotion

Gregory W Heath, University of Tennessee at Chattanooga and UT College of Medicine, Chattanooga, TN, USA
Gary Liguori, University of Tennessee at Chattanooga, Chattanooga, TN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Worldwide it has been estimated that physical inactivity alone accounts for the premature loss of ~5.2 million lives per year. Sedentary and inactive people should be given opportunities to become physically active, and those who are minimally active should be supported in doing more. Through community organization physical activity levels can increase among all individuals living in communities and bring about considerable health benefits. Approaches that demonstrate an acceptable level of effectiveness in increasing physical activity across different ages, social groups, and countries exist. For the whole population, there is much to be gained through policies and environmental supports, particularly through partnerships with the sectors of transport and urban planning that can increase opportunities for physical activity within communities.

Foundations for Physical Activity Promotion

Physical inactivity (PIA) is an independent risk factor for a number of noncommunicable diseases (NCDs) worldwide, such as coronary heart disease, type 2 diabetes mellitus, colon cancer, and breast cancer, where it has been estimated that PIA alone is responsible for the premature loss of ~5.2 million lives per year (Lee et al., 2012). Further evidence demonstrates a positive relationship between participation in regular physical activity (PA) and health improvement (Physical Activity Guidelines Advisory Committee (PAGAC, 2008)). Evidence also exists supporting the quality of life enhancing benefits of regular PA as seen among adults and older adults (Brown et al., 2004).

Despite over a decade of efforts to translate this information and promote PA within communities, physical inactivity remains a serious public health challenge (Pate et al., 1995; Department of Health and Human Services, 1996). The most recent prevalence data summarizing global PA patterns indicate that a third of all adults and four fifths of youth/adolescents are physically inactive (Hallal et al., 2012; **Table 1**). The prevalence of physical inactivity is consistent across high, middle, and low income countries (Hallal et al., 2012). What's even more alarming is that more than half of the world's adults are not participating in the minimal amount of PA necessary to obtain some health benefits (Hallal et al., 2012). Men in general are more likely to meet PA guidelines than women. These levels of participation fall far short of the updated PA guidelines in the United States (PAGAC, 2008) and those issued by the World Health Organization (WHO, 2010) and other organizational bodies (European Union Sport and Health Working Group, 2008; **Table 2**). In addition to moderate- and vigorous-intensity aerobic PA many recommendations/guidelines suggest that every adult should perform physical activities that maintain or increase muscular strength and endurance and skeletal health and function for a minimum of 2 days each week (PAGAC, 2008; WHO, 2010). Participation in aerobic PA and muscle strengthening physical activities above this minimal recommendation provides additional health benefits and results in higher levels of physical fitness (PAGAC, 2008).

Table 1 Prevalence of physical inactivity – global estimates (by age and WHO region)

WHO region	*15 to 29 years*	*30–44 years*	*45–59 years*	*≥60 years*
Africa	23%	25%	30%	38%
Americas	37%	38%	45%	60%
Eastern Mediterranean	38%	39%	43%	55%
Europe	30%	34%	37%	48%
Southeast Asia	14%	13%	18%	30%
Western Pacific	35%	32%	33%	45%

Adapted from: Hallal, P.C., Andersen, L.B., Bull, F.C., Guthold, R., Haskell, W., Ekelund, U., For the Lancet Physical Activity Series Working Group, 2012. Global physical activity levels: surveillance progress, pitfalls, and prospects. Lancet 380, 247–257.

Understanding why some people are more physically active than others is fundamental to developing strategies and/or specific interventions designed to promote PA among individuals. Bauman et al. (2012), in their most recent review of reviews examining both the correlates (factors associated with PA) and determinants (those factors with a causative relationship to PA), identified age, sex, self-efficacy, and motivation as being associated with PA, with the broader influences on PA behavior of social, cultural, and physical environmental factors playing key causative roles. Therefore current evidence suggests that the successful promotion of a physically active lifestyle involves a multitude of complex variables working together, including personal, programmatic, environmental, and related factors. To achieve long-term changes in health-related behaviors, individual, social, environmental, and medical factors must be addressed collectively with the growing realization that the physical environment is key in promoting or inhibiting PA even among the most motivated of persons (Bauman et al., 2012). A future understanding of how each individual's genetic, evolutionary, and personal characteristics interact with her or his physical and social/cultural environment in being physically inactive or active is critical to effective promotion efforts (Bauman et al., 2012).

Table 2 Physical activity guidelines for adults and youth

Physical Activity Guidelines for Americans:

Children and adolescents should do 60 min (1 h) or more of physical activity daily.

- *Aerobic*: Most of the 60 or more minutes a day should be either moderate- or vigorous-intensity aerobic physical activity, and should include vigorous-intensity physical activity at least 3 days a week.
- *Muscle-strengthening*: As part of their 60 or more minutes of daily physical activity, children and adolescents should include muscle-strengthening physical activity on at least 3 days of the week.
- *Bone-strengthening*: As part of their 60 or more minutes of daily physical activity, children and adolescents should include bone-strengthening physical activity on at least 3 days of the week.

All adults should avoid inactivity. Some physical activity is better than none, and adults who participate in any amount of physical activity gain some health benefits.

- For substantial health benefits, adults should do at least 150 min (2 h and 30 min) a week of moderate-intensity, or 75 min (1 h and 15 min) a week of vigorous-intensity aerobic PA, or an equivalent combination of moderate- and vigorous intensity aerobic activity. Aerobic activity should be performed in episodes of at least 10 min, and preferably, it should be spread throughout the week.
- For additional and more extensive health benefits, adults should increase their aerobic PA to 300 min (5 h) a week of moderate intensity, or 150 min a week of vigorous intensity aerobic PA, or an equivalent combination of moderate- and vigorous-intensity activity. Additional health benefits are gained by engaging in PA beyond this amount.
- Adults should also do muscle-strengthening activities that are moderate or high intensity and involve all major muscle groups on 2 or more days a week, as these activities provide additional health benefits.

Global Recommendations for Physical Activity, World Health Organization:

In order to improve cardiorespiratory and muscular fitness, bone health, and cardiovascular and metabolic health biomarkers:

- Children and youth aged 5–17 should accumulate at least 60 min of moderate- to vigorous-intensity physical activity daily.
- Amounts of physical activity greater than 60 min provide additional health benefits.
- Most of the daily physical activity should be aerobic. Vigorous-intensity activities should be incorporated, including those that strengthen muscle and bone*, at least 3 times per week. For this age group, bone-loading activities can be performed as part of playing games, running, turning, or jumping.
- Adults aged 18–64 should do at least 150 min of moderate-intensity aerobic physical activity throughout the week or do at least 75 min of vigorous-intensity aerobic physical activity throughout the week or an equivalent combination of moderate- and vigorous-intensity activity.
- Aerobic activity should be performed in bouts of at least 10 min duration.
- For additional health benefits, adults should increase their moderate-intensity aerobic physical activity to 300 min per week, or engage in 150 min of vigorous-intensity aerobic physical activity per week, or an equivalent combination of moderate- and vigorous-intensity activity.
- Muscle-strengthening activities should be done involving major muscle groups on 2 or more days a week.

In order to improve cardiorespiratory and muscular fitness, bone and functional health, reduce the risk of NCDs, depression and cognitive decline:

- Older adults should do at least 150 min of moderate-intensity aerobic physical activity throughout the week or do at least 75 min of vigorous-intensity aerobic PA throughout the week or an equivalent combination of moderate- and vigorous-intensity activity.
- Aerobic activity should be performed in bouts of at least 10 min duration.
- For additional health benefits, older adults should increase their moderate-intensity aerobic physical activity to 300 min per week, or engage in 150 min of vigorous-intensity aerobic PA per week, or an equivalent combination of moderate-and vigorous-intensity activity.
- Older adults, with poor mobility, should perform physical activity to enhance balance and prevent falls on 3 or more days per week.
- Muscle-strengthening activities, involving major muscle groups, should be done on 2 or more days a week.

When older adults cannot do the recommended amounts of physical activity due to health conditions, they should be as physically active as their abilities and conditions allow.

Sources: http://www.health.gov/paguidelines/ (accessed 01.07.13); http://www.who.int/dietphysicalactivity/factsheet_recommendations/en/ (accessed 01.07.13).

Behavioral and Social Strategies for Promoting Physical Activity

Within the scope of health promotion is the necessity to support individuals attempting to make positive health behavior changes. More specifically, sedentary people should be given opportunity to become physically active, and those who are minimally active should be supported in doing more. However, behavior change for PA, along with many other personal habits, can be quite difficult. Therefore, the field of health promotion typically utilizes theories and models when promoting individual or community-based interventions, including increases in PA. A theory refers to a systematic view of a behavior by specifying relationships between variables and predicting specific behaviors and situations (Glanz et al., 2002). A model is defined as a hypothetical depiction of a behavior or situation (Glanz and Rimer, 1995). Models do not attempt to understand the variables underlying a particular behavior, but rather to represent what is happening with a particular behavior. Theories and models are used in PA promotion as they (1) provide a framework for better understanding adoption of the behavior, (2) can help practitioners understand why an individual has stopped PA participation, and (3) allow the practitioner to identify which types of individuals respond to which types of PA promotion strategies.

Intrapersonal and Interpersonal Concepts and Principles in Promoting Physical Activity

Learning theory is a concrete foundation and strategy for facilitating PA related behavior change (Glanz and Rimer, 1995). Learning theory describes how individuals acquire, enhance, or make changes in knowledge, skills, and values (Nutbeam and Harris, 1999), and also reflects the process through which a PA behavior change occurs. Learning theory

is based on antecedents and consequences: an antecedent is a preceding event or circumstance that prompts PA behavior, while a consequence is the result of the antecedent behavior, or the immediate benefit/consequence. Theory-based behavioral interventions are an effective means for increasing PA among sedentary adults (Greaves et al., 2011), and people who receive learning theory-based strategies increase their exercise adherence considerably more than those who initiate exercise programs with no theory-based strategies (Marcus et al., 2006). Theory-based interventions that are tailored to each individual's specific interests, preferences, and readiness for change can teach behavioral skills that help incorporate exercise into lifelong routines (Kahn et al., 2002; PAGAC, 2008).

Transtheoretical Model

The Transtheoretical Model (TTM) has been used to explain health behaviors including smoking cessation, weight control, high-fat diets, adolescent delinquent behaviors, condom use, mammography screening, and sunscreen use (Prochaska et al., 1992). The TTM incorporates both cognitive theories and behavioral strategies to tailor interventions to the individual. Cognitive processes are used to change the way we think about activity, while behavioral processes are used to change/initiate the actual behavior itself. The TTM has also been used to understand PA behavior and to create PA interventions (Marcus and Forsyth, 2009), which have been shown efficacious in increasing PA among sedentary adults (Marcus et al., 1992). The TTM proposes that individuals move through a series of stages, or 'stages of change,' during PA adoption (Marcus et al., 1992; Marcus and Forsyth, 2009) and specific behavioral and cognitive processes occur as individuals move through each stage. Some of the behavioral processes include enlisting social support, rewarding oneself, committing oneself, and reminding oneself (Lewis et al., 2002). The five stages of change are (1) precontemplation, (2) contemplation, (3) preparation, (4) adoption, and (5) maintenance. According to the TTM, interventions should be tailored to the individual depending on their current stage of change, which can be identified with the following questions:

1. Are you currently accumulating at least 150 min of moderate intensity PA each week?
2. Have you been regularly physically active over the last 6 months?
3. Are you doing any PA?
4. Have you made any actions and/or concrete plans in increasing your PA?
5. Do you plan on becoming more physically active over the next 6 months?

Once you have answers to the above questions, you can better develop a plan of action specific to each client and tailored to their current stage or readiness.

Health Belief Model

The health belief model (HBM) is a conceptual framework outlining an individual's based on his/her health beliefs (Rosenstock, 1974). The model states that as individuals take greater investment in their personal health, they are more likely to make relevant and meaningful behavior changes. The four main components of the HBM are:

1. perception of susceptibility for the identified health threat;
2. perception of the severity of the identified health threat;
3. perception of the benefits from taking action to reduce the identified health threat; and
4. perception of barriers and/or costs of taking action to reduce the identified health threat.

According to the HBM, an individual examines the negative aspects of particular health action and/or behavior and weighs those 'costs' with the benefits of the health action. If benefits outweigh the costs, then a healthy action is more likely. This can also be considered as an equation in which (1) (susceptibility + seriousness) = perceived threat, (2) (benefits + barriers) = outcome expectation, and (3) (threat + expectations) = likelihood of action.

The HBM can be especially relevant when working with individuals and groups who are at higher risk for sedentary related chronic diseases such as cardiovascular disease, diabetes, arthritis, high blood pressure, dyslipidemia, etc.

Social and Cultural Principles in Promoting Physical Activity

Self-Efficacy/Social Cognitive Theory: PA self-efficacy is a term used to describe individuals' specific self-confidence in their ability to engage in PA. More broadly, self-efficacy is part of the Social Cognitive Theory (SCT) (Bandura, 1986) and postulates that the more confident one feels in their capabilities and skills to succeed, the more likely they will engage in that behavior (Bandura, 1977). SCT provides a theoretical framework for understanding PA adoption (McAuley and Blissmer, 2000) and emphasizes reciprocal determinism, or the interaction between individuals and their environments. SCT identifies three main factors that influence behavior and behavioral choices: (1) the environment (e.g., neighborhood, proximity to gym), (2) individual personality characteristics and/or experience (including cognitions), and (3) behavioral factors. Behavior is the product of the interplay between these three factors. In other words, the environment can influence individuals and groups, but individuals and groups can also influence their environments, and in turn, govern their own behaviors. Intervention studies indicate that self-efficacy is like an important component of PA behavior change across a variety of populations (Lewis et al., 2002; White et al., 2011; Haas, 2011). Self-efficacy is a product of four sources of information: (1) enactive mastery experience, (2) vicarious experience, (3) verbal persuasion, and (4) physiological or affective states. Each of these sources can be used individually or collectively to increase self-efficacy (Bandura, 1977). Assessing PA self-efficacy enables the practitioner to tailor feedback and encouragement to each client. To conduct the assessment, ask individuals how confident they are in their ability to regularly engage in PA. Using a PA self-efficacy scale of 1 (not confident) to 5 (very confident), individuals can be categorized as low or high PA self-efficacy, then tailored feedback can be offered, particularly for those with low confidence.

Social Support

Social support research indicates that when support comes from family and friends it can be an important component for exercise adherence (Anderson et al., 2006). Social support refers to anything a friend or family member does to help increase the individual's exercise level. Examples of social support strategies include encouraging exercise, exercising with the individual, watching the individual's children so he or she can exercise, or providing monetary support for attending exercise facilities. In the exercise setting, social support is often achieved via active participation, which creates a health-related win–win situation for both persons. The odds of success in PA are increased when goals are shared with friends and family, which also creates a greater sense of obligation and ownership for the goal-setter. However, the amount and type of social support that is most helpful varies between individuals.

Policy and Environmental Concepts and Principles in Promoting PA

The Social Ecological Model (SEM) is a comprehensive approach to change that integrates intra- and interpersonal factors including community and organizational factors, institutional factors, environmental factors, and public policies (**Figure 1**). All levels of the SEM are equally important, and all can influence individual behavior (Bronfenbrenner, 2009).

Many different versions of the SEM exist in regards to PA, each including a microsystem (personal interactions between family members, work groups, etc.), mesosystem (the physical setting for family, school, and work), and exosystem (the larger social influences of economics, policies, culture, and politics) (Bronfenbrenner, 2009). The SEM is specific to the promotion of health behaviors and focuses on assumptions related to influences of the physical and social environments, multidimensional environments, interactions between individuals, families, and communities, and how individuals influence their surroundings (Stokols et al., 2003).

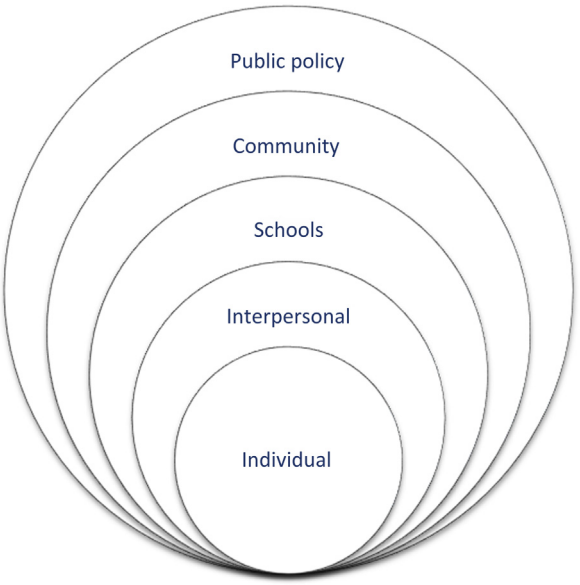

Figure 1 Socioecologic model and the promotion of physical activity.

More recent attention has been on the significant role the environment plays on PA and health behaviors. Fields of transportation, urban planning, school zoning, access to facilities and programs, and supporting social environments that encourage activity may all influence PA behavior (Davison and Lawson, 2006; Heath et al., 2006). Empirical evidence suggests that the social, physical, and policy environments influence PA participation, and the SEM provides the framework to better understanding these factors and barriers. The SEM helps to identify opportunities to promote participation in PA by recognizing the multiple variables that may influence an individual's choices.

Building Community Capacity to Promote Physical Activity

Capacity building is seen as a critical element of health promotion and can be used to address a variety of population-based health risks including PIA (Canadian Cancer Society, 2013). As a community development tool, capacity building occurs on, and creates change at, three separate but related levels – individual, organizational, and social. Capacity building strives to create an increase in the ability of individuals and groups to positively impact the health and well-being of their communities, and to sustain this impact indefinitely. PA occurs in individuals and communities, and when performed at sufficient levels, can bring considerable health benefits, including reduced morbidity and mortality (PAGAC, 2008). Although the prevalence of PA is difficult to measure precisely, it is agreed that considerably less than 50% of the U.S. adult population is regularly active (Carlson et al., 2010) and no more than 35% of students in grades 9–12 meet recommended levels of PA (YRBSS, 2007). Additionally, the body of evidence on community-based PA interventions has grown (CDC, 2001; Kahn et al., 2002). Using strategies based on the socio-ecological perspective, effective PA promotion can occur for individuals, groups, and communities.

Essentials of Community Organization and Community Development

Community-wide campaigns are intended to be sustained efforts with ongoing high visibility. They can be large or small in scope, involve multiple partnerships, and should use various forms of media. A brand or tag line can be especially helpful, as was "VERB, It's What You Do," the CDC's successful youth PA campaign (CDC, 2010). Capacity building for a community typically involves health fairs, community events, policy change, risk factor screening and education at worksites and schools. Additionally, messages can be tailored to fit specific target populations such as school children, worksites, retirees, etc. (Kahn et al., 2002). There are a number of key elements necessary to deliver a successful community wide campaign, but regardless, all campaigns are resource intensive and require a large number of well-trained staff. The initial step in any successful campaign is to conduct the appropriate formative research to develop an effective theme and messaging. Without

a complete and thorough understanding of the community's needs and interests, it is unlikely any campaign can succeed. Concomitant with identifying the community needs and interests is gaining the community's buy-in to the process and intended outcome(s). This may require a considerable amount of time and work, and should include gaining the support of key community leaders, such as the mayor, work-shift leaders, school principals, and others that can positively influence attitudes in the community.

Effective Partnering and Coalition Building

A coalition is an alliance of individuals and organizations that come together to address a specific problem or issue. Goals that focus on system-wide changes and collaborations and require a variety of expertise are particularly well suited for coalitions (Rabinowitz, 2013). Stakeholders invest in a coalition when there are achievable goals and objectives to benefit from, yet unable to achieve on their own. Some inherent coalition benefits include expanding business networks, improving access to contacts, strengthening political position, and increasing access to funding opportunities. For any coalition to be successful there needs to be clear expectations toward the goal, strong leadership, and adequate staffing. At the same time, barriers need to be overcome, including finding the right lead organization, competition between coalition groups, perceived cost–benefit factor, and agreement on a common goal (Cohen et al., 2002). To initiate a community coalition for PA, the recommended steps are grouped under Formation, Management, and Evaluation. Formation includes a diverse group of stakeholders, ample front-end time to prepare, establishing a shared vision, and conducting a needs assessment. The final step of Formation is to create a data-driven plan, using an established strategy like Specific, Measurable, Appropriate, Realistic, and Time specific, and based almost exclusively on the needs assessment. The Management step requires a shared approach to responsibility, effective lines of communication, and a means and schedule for monitoring progress. This last piece should include the flexibility to reassess, revise, and recommit any and all aspects of the goal or plan, but should be based on the available data, not on anecdote or intuition. During Evaluation there is a continuous need to assess the coalition's efforts and to share the lessons learned along the way.

Program Evaluation Tools to Improve Programs

Program evaluation is the final step in determining the outcomes of any organizational efforts. Evaluations can be conducted using many different tools, but regardless, it is best to establish the evaluation plan prior to implementing any aspect of the proposed program, as post-hoc evaluations plans tend to be biased. A vast array of tools exists for evaluating community-based PA programs: The Framework for Program Evaluation in Public Health (CDC, MMWR, 1999), the Physical Activity Evaluation Handbook (USDHHS, 2002), and the Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis (Fine, 2009) are some of the more common. The Community Toolbox (The Community Toolbox, 2013) is an outstanding resource providing comprehensive information for each stage of community health promotion initiatives, including a How-to guide, Toolkits for completing tasks, a Troubleshooting section, and complete information on evidence-based practices.

Evidence-Based Community-Level Strategies for Promoting Physical Activity

In light of the scientific guidelines issued by various international bodies, national centers/institutes, and professional organizations, and the recent international evidence review documenting that regular PA protects against major NCDs (Lee et al., 2012), there appears to be significant health benefits accrued among whole communities by engaging in regular PA (Global Advocacy Council for PA, 2010).

As recently described by Heath et al., 2012, interventions designed to increase PA among whole populations are now much more prominent among public health and health system interventions (Kahn et al., 2005; Heath et al., 2006, 2012). Since PA behaviors are understood by factors that operate concurrently at multiple levels of influence (personal, social/cultural, environmental levels), approaches that operate across various community sectors appear to be the most successful in increasing PA among community residents (Kohl et al., 2012). Therefore health service and public health agencies need to partner with community organizations, including schools, businesses, policy, advocacy, nutrition, recreation, planning, and transport agencies to effectively collaborate, cooperate, coordinate, and communicate promotional efforts to increase PA (Heath et al., 2012).

For convenience, effective community-based PA intervention strategies are presented using the domains established by the Guide to Community Preventive Services (Kahn et al., 2002). These domains conveniently capture the majority of PA intervention strategies delivered throughout the world and consist of descriptors that are also found in other international PA recommendation documents (Heath et al., 2012; **Table 3**).

Physical Activity Interventions that Work

Campaigns and Informational Approaches

Community-wide campaigns are recommended within the campaigns and informational approaches domain

Table 3 Community-based PA intervention domains

1. *Campaigns and Informational approaches* to promoting PA – these are strategies aimed at changing knowledge, atti1tudes, and behavior about the benefits of and opportunities for PA within a community.
2. *Behavioral and social approaches* are aimed at teaching people the behavioral management skills that are necessary for successful adoption and maintenance of behavior change, and working through or creating organizational and social environments that facilitate and enhance behavioral change.
3. *Environmental and policy approaches* that are aimed at structuring the physical and organizational environments so that people have accessible, safe, attractive, and convenient places to be physically active.

(Table 4). These campaigns represent large-scale, high-intensity, high-visibility programming using various media to raise program awareness, disseminate targeted/segmented health messages, and reinforce behavior change. This strategy often employs multicomponent, multisector, and multisite interventions, and are representative of PA promotion efforts directed primarily toward populations living in communities within middle to high income countries (Heath et al., 2012). Mass media campaigns are another intervention strategy that produce PA behavior changes in selected populations, especially when they are linked to specific community programming. An emerging PA intervention strategy which has been employed among Latin American communities is an informational approach using (1) delivery of short PA-related educational and motivational messages, (2) delivered by a health educator/communicator, and (3) to a specific population at key community sites (Table 4). Point-of-decision prompts is an effective single-component informational strategy for promoting PA. This consists of signage designed to remind and motivate people to make acute healthy choices, such as using the stairs instead of the elevator (Table 4).

Behavioral and Social Approaches

The intervention domain of behavioral and social approaches to promoting PA encompasses a number of strategies that are site- and population-specific. Individually-adapted health behavior change programs represent a broad array of interventions that are multi-component approaches, with the goal of having participants incorporate PA into their daily routines (Kahn et al., 2002; Heath et al., 2012). The components of this strategy include goal setting, social support, behavioral reinforcement through self-reward, structured problem solving, and relapse prevention. These interventions are often delivered by email, internet, mail, and/or telephone. Individually-focused interventions usually consist of an assessment of a participant's PA level and readiness to change, a tailored activity plan, and navigation to community interventions by a centralized health provider or promoter. This approach, which focuses on lifestyle PA, has been shown to be cost-effective when compared to supervised PA programs (Roux et al., 2008). Another strategy within the domain of behavioral and social approaches is social support interventions in community settings, which capitalizes on social networks to reinforce PA behavior. Such a strategy can include creating buddy systems, contracting, and forming PA support groups. Community settings where such social support interventions can be carried out include worksites, community centers, health facilities, and parks and recreational facilities. Healthcare provider-based assessment and counseling for PA has undergone systematic review by several different groups (Heath et al., 2012) and still lacks sufficient evidence to be recommended as single component intervention. However, some investigators have demonstrated improved success with this strategy when it is integrated into existing community efforts and has emerged as a promising practice for PA promotion (Patrick et al., 2009). Others have found that the evidence for health-care provider assessment and counseling of patients for the promotion of PA is mixed. Brief stand-alone counseling by providers is not efficacious, but office-based screening and advice followed by telephone or community support for PA is effective in sustaining more long-term PA behavior improvement among patients (van Sluijs et al., 2005). Thus, healthcare provider delivery models that emphasize coordination with clinical and community resources may be the most optimal in promoting PA among healthcare patients (Pavey et al., 2011). An emerging behavioral and social approach to promoting PA is the convening of PA classes in community/public settings for community residents, with no cost to the participants. These programs provide social support, create opportunity where recreational deficits exist, and are relevant for underserved populations that are less likely to achieve PA recommendations. Such populations include women, older adults, and people with limited income. Since participation in such programs usually do not incur payment by the participating public, these strategies may also contribute to reducing social and health disparities (Heath et al., 2012).

Recommended strategies within the behavioral and social domain are school-based interventions such as school-based physical education (PE). These intervention strategies provide great potential to increase levels of PA among children and youth since in many countries PE is mandatory. This increases the likelihood of the least active children participating in some level of regular PA (Heath et al., 2012). School-based interventions also include programs delivered during school time as well as afterschool where core components for the effectiveness of such programs include: increased PA sessions, improved quality of PE classes, increased PA during recess and other times, teacher capacity building and training, changes in PE curricula; provision of equipment and materials; and adjustment of the interventions to specific target populations

Table 4 Overview of effective, promising, and emerging strategies to promote physical activity in communities

Approaches (domain)	*Strategy (intervention)*	*Classification (effectiveness)*
Campaigns and informational	Point-of-decision prompts	Effective
	Community-wide campaigns	Effective/promising
	Mass media campaigns	Promising
	Short informational messages	Emerging
Behavioral and social	School-based strategies	Effective
	Social support in communities	Effective
	Provider-based counseling	Promising
	Community PA classes	Promising
Policy and environmental	Community-scale urban design	Effective
	Street-scale urban design	Effective
	Transportation policy and practice	Emerging
	Community-wide planning/ policy	Emerging

Adapted from: Heath, G.W., Parra, D.C., Sarmiento, O.L., Andersen, L.B., Owen, N., Goenka, S., Montes, F., Brownson, R.C., 2012. Evidence-based intervention in physical activity: lessons from around the world. Lancet 380, 272–281.

(Heath et al., 2012). School-based interventions have been evaluated across outcomes such as PA level, fitness, obesity, other cardiovascular risk factors, and measures of well-being (Heath et al., 2012).

Policy and Environmental Approaches

An effective set of intervention strategies representing the policy and environmental domain is the creation of or enhanced access to places for PA combined with informational outreach activities (Heath et al., 2012). These interventions often consist of creating/enhancing existing walking/biking trails or exercise facilities, and increasing access to existing facilities by reducing structural and environmental barriers (e.g., increasing safety, enhancing affordability) (Heath et al., 2006, 2012; **Table 4**).

Community-scale urban design and land-use policies and practices commonly strive to create more livable communities and have been shown to be effective in promoting PA at the community level (Heath et al., 2006, 2012). These types of interventions use policy instruments such as zoning regulations and building codes, and environmental changes brought about by government policies or developers' practices. The latter includes policies encouraging transit-oriented development, and policies addressing street layouts, density of development, increased store location, jobs and schools that can be walked to (Heath et al., 2012). The NICE review addressing the promotion of PA through built or natural environments provides further evidence in support of this strategy outside of North America (NIHCE, 2008). Given the diversity of countries and populations included in these bodies of evidence, these results should be applicable to diverse settings and populations, provided appropriate attention is paid to adapting the intervention to the target population.

Street-scale urban design and land-use policies to increase PA use policy instruments and practices to support PA in small geographic areas, generally limited to a few blocks. These policies and practices include improved street lighting or infrastructure projects that increase the ease and safety of street crossing, ensure sidewalk continuity, introduce or enhance traffic calming, or enhance the aesthetics of the street area, such as landscaping (Heath et al., 2012). These interventions are designed to enhance the urban environment and/or to increase PA by redesigning streets and sidewalks, creating bike lanes/ways, and improving the perceived environment (Heath et al., 2012). These interventions are internationally representative and cross multiple countries (Heath et al., 2012).

Transportation and travel policies and practices that promote PA include interventions that strive to improve pedestrian, transit and light rail access, increase pedestrian and cyclist activity and safety, reduce car use, and improve air quality (de Nezelle et al., 2011; Heath et al., 2012).

A potentially effective intervention strategy is the use of community-wide policies and planning combined with multicomponent efforts at the community level to promote PA (Reis et al., 2010; Heath et al., 2012). These community action plans and policies are designed to reduce environmental and structural barriers that directly impact PA behaviors and are promoted through media campaigns and incentives at various levels (e.g., individual, corporate, local, and regional) (Reis et al., 2010; Heath et al., 2012).

Summary

As the aforementioned international perspective on PA intervention shows, there are a number of evidence-based approaches that demonstrate an acceptable level of effectiveness in increasing PA across different ages, social groups, and countries. Similarly there are several promising and/or emerging interventions specifically from middle-income countries that deserve attention and more rigorous evaluation. Among children and youth, there is enormous potential to increase PA through school-based initiatives. For the whole population, and particularly for adults, there is much to be gained through developing policies and environmental supports, particularly through partnerships with the sectors of transport and urban planning that can increase opportunities for PA within communities. Interventions are likely to be more effective, to the extent that they can address the determinants of PA at multiple levels. Ideally, PA initiatives should apply the relevant conceptual models and address the determinants at individual, behavioral, social, and environmental and policy levels tailored to specific populations across and within countries.

See also: Adolescent Health and Health Behaviors; Health Behavior, Psychosocial Theories of; Health Behaviors; Health Interventions, Community-based; Health Promotion in Schools; Health Psychology in Childhood; Health Psychology; Physical Activity and Health; Self-Efficacy and Health; Socioeconomic Status and Health; Urban Life and Health.

Bibliography

Anderson, E.S., Wojcik, J.R., Winett, R.A., Williams, D.M., 2006. Social-cognitive determinants of physical activity: the influence of social support, self-efficacy, outcome expectations, and self-regulation among participants in a church-based health promotion study. Health Psychology 25 (4), 510–520.

Bandura, A., 1977. Self-efficacy: toward a unifying theory of behavioral change. Psychological Review 84 (2), 191–215.

Bandura, A., 1986. Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice-Hall, Inc., Englewood Cliffs, NJ, USA.

Bauman, A.E., Reis, R.S., Sallis, J.F., Wells, J., Loos, R.J.F., Martin, B.W., For the Lancet PA Series Working Group, 2012. Correlates of physical activity: why are some people physically active and others not? Lancet 2, 258–271.

Bronfenbrenner, U., 2009. The Ecology of Human Development: Experiments by Nature and Design. Harvard University Press, Cambridge, MA, USA.

Brown, D.W., Brown, D.R., Heath, G.W., Balluz, L., Giles, W.H., Ford, E.S., Mokdad, A.H., 2004. Associations between physical activity dose and health-related quality of life. Medicine and Science in Sports and Exercise 36 (5), 890–896.

Canadian Cancer Society, 2013. Community Capacity Building Strategy. http://cancer.ca.ccsbcy.org/communitycapacitybuildingstrategy/home.

Carlson, S.A., Fulton, J.E., Schoenborn, C.A., Loustalot, F., 2010. Physical activity trends and prevalence estimates based on the 2008 PA Guidelines for Americans, National Health Interview Survey 1998–2008. American Journal of Preventive Medicine 39 (4), 305–313.

Centers for Disease Control and Prevention, 1999. Framework for program evaluation in public health. Morbidity and Mortality Weekly Report 48 (RR-11), 1–40.

Centers for Disease Control and Prevention, 2001. Increasing physical activity: a report on recommendations of the task force on community preventive services. Morbidity and Mortality Weekly Report 50 (RR-18).

Centers for Disease Control and Prevention, 2010. National Center for Chronic Disease Prevention and Health Promotion, Division of Adolescent and School Health. VERB.

Cohen, L., Baer, N., Satterwhite, P., 2002. Developing effective coalitions: an eight step guide. In: Wurzback, M.E. (Ed.), Community Health Education and Promotion: A Guide to Program Design and Evaluation, second ed. Aspen Publishers Inc., Gaithersburg, MD, pp. 144–161.

The Community Toolbox, 2013. Work Group for Community Health and Development, University of Kansas. http://ctb.ku.edu/en/default.aspx.

Davison, K., Lawson, C., 2006. Do attributes in the physical environment influence children's physical activity: a review of the literature. International Journal of Behavioral Nutrition and Physical Activity 3 (1), 19.

European Union Sport and Health Working Group, 2008. European Union Physical Activity Guidelines. European Union, Brussels, Belgium.

Fine, L.G., 2009. The SWOT Analysis: Using Your Strength to Overcome Weaknesses, Using Opportunities to Overcome Threats, first ed. CreateSpace Independent Publishing Platform.

Glanz, K., Rimer, B.K., 1995. Theory at a Glance: A Guide for Health Promotion Practice [NIH Pub. No. 95–3896]. National Cancer Institute, Washington, DC.

Glanz, K., Lewis, F.M., Rimer, B.K., 2002. Theory, research, and practice in health behavior and health education. In: Glanz, K., Rimer, B.K., Lewis, F.M. (Eds.), Health Behavior and Health Education: Theory, Research, and Practice, third ed. Josey-Bass, San Francisco, CA, pp. 279–231.

Greaves, C.J., Sheppard, K.E., Abraham, C., et al., 2011. Systematic review of reviews of intervention components associated with increased effectiveness in dietary and physical activity interventions. BMC Public Health 11, 119–130.

Global Advocacy Council for PA, International Society for Physical Activity and Health, May 20, 2010. The Toronto Charter for PA: A Global Call to Action. http://www.globalpa.org.uk/pdf/torontocharter-eng-20may2010.pdf (accessed 27.07.13.).

Haas, B.K., 2011. Fatigue, self-efficacy, physical activity, and quality of life in women with breast cancer. Cancer Nursing 34 (4), 322–334.

Hallal, P.C., Andersen, L.B., Bull, F.C., Guthold, R., Haskell, W., Ekelund, U., For the Lancet Physical Activity Series Working Group, 2012. Global physical activity levels: surveillance progress, pitfalls, and prospects. Lancet 380, 247–257.

Heath, G.W., Brownson, R.C., Kruger, J., Miles, R., Powell, K.E., Ramsey, L.T., 2006. The effectiveness of urban design and land use and transport policies and practices to increase physical activity: a systematic review. Journal of Physical Activity and Health 1, S55–S71.

Heath, G.W., Parra, D.C., Sarmiento, O.L., Andersen, L.B., Owen, N., Goenka, S., Montes, F., Brownson, R.C., 2012. Evidence-based intervention in physical activity: lessons from around the world. Lancet 380, 272–281.

Kahn, E.B., Ramsey, L.T., Brownson, R.C., Heath, G.W., Howze, E.H., Powell, K.E., Stone, E.J., 2005. Chapter 2, physical activity. In: Zaza, S., Briss, P.A., Harris, K.W. (Eds.), The Guide to Community Preventive Services: What Works to Promote Health. Oxford University Press, pp. 80–113.

Kahn, E.B., Ramsey, L.T., Brownson, R.C., Heath, G.W., Howze, E.H., Powell, K.E., Stone, E.J., Rajab, M.W., Corso, P., 2002. The effectiveness of interventions to increase physical activity: a systematic review. American Journal of Preventive Medicine 22 (4S), 73–107.

Kohl 3rd, H.W., Armstrong, T., Craig, C.L., et al., 2012. The pandemic of physical inactivity: global action for public health. Lancet 380, 294–305.

Lee, I.-M., Shiroma, E.J., Lobelo, F., Puska, P., Blair, S.N., Katzmarzyk, P.T., For the Lancet PA Series Working Group, 2012. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. Lancet 380, 219–229.

Lewis, B.A., Marcus, B.H., Pate, R.R., Dunn, A.L., 2002. Psychosocial mediators of physical activity behavior among adults and children. American Journal of Preventive Medicine 23 (2, Suppl. 1), 26–35.

Marcus, B.H., Forsyth, L.H., 2009. Motivating People to Be Physically Active, second ed. Human Kinetics, Champaign, IL.

Marcus, B.H., Rossi, J.S., Selby, V.C., Niaura, R.S., Abrams, D.B., 1992. The stages and processes of exercise adoption and maintenance in a worksite sample. Health Psychology 11 (6), 386–395.

Marcus, B.H., Williams, D.M., Dubbert, P.M., et al., 2006. What we know and what we need to know: a scientific statement from the American heart association council on nutrition, physical activity, and metabolism (subcommittee on physical activity); council on cardiovascular disease in the young; and the interdisciplinary working group on quality of care and outcomes research. Circulation 114, 2739–2752.

McAuley, E., Blissmer, B., 2000. Self-efficacy determinants and consequences of physical activity. Exercise and Sport Sciences Reviews 28 (2), 85–88.

de Nazelle, A., Nieuwenhuijsen, M.J., Antó, J.M., Brauer, M., Briggs, D., Braun-Fahrlander, C., Cavill, N., Cooper, A.R., Desqueyroux, H., Fruin, S., Hoek, G., Panis, L.I., Janssen, N., Jerrett, M., Joffe, M., Andersen, Z.J., van Kempen, E., Kingham, S., Kubesch, N., Leyden, K.M., Marshall, J.D., Matamala, J., Mellios, G., Mendez, M., Nassif, H., Ogilvie, D., Peiró, R., Pérez, K., Rabl, A., Ragettli, M., Rodríguez, D., Rojas, D., Ruiz, P., Sallis, J.F., Terwoert, J., Toussaint, J.F., Tuomisto, J., Zuurbier, M., Lebret, E., 2011. Improving health through policies that promote active travel: a review of evidence to support integrated health impact assessment. Environment International 37 (4), 766–777.

National Institute for Health and Clinical Excellence, 2008. Promoting and Creating Built or Natural Environments that Encourage and Support Physical Activity. Quick Reference Guide, NIHCE, London.

Nutbeam, D., Harris, E., 1999. Theory in a Nutshell: A Guide to Health Promotion Theory. The McGraw-Hill Companies, Inc., Sydney, Australia.

Pate, R.R., Pratt, M., Blair, S.N., et al., 1995. Physical activity and public health: a recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine. Journal of the American Medical Association 273, 402–407.

Patrick, K., Pratt, M., Sallis, R.E., 2009. The healthcare sector's role in the U.S. National Physical Activity Plan. Journal of Physical Activity and Health 6 (Suppl. 2), S211–S219.

Pavey, T.G., Taylor, A.H., Fox, K.R., et al., 2011. Effect of exercise referral schemes in primary care on physical activity and improving health outcomes: systematic review and meta-analysis. British Medical Journal 343, d6462.

Physical Activity Guidelines Advisory Committee, 2008. Physical Activity Guidelines Advisory Committee Report, 2008. U.S. Department of Health and Human Services, Washington, DC.

Pratt, M., Sarmiento, O.L., Montes, F., et al., For the Lancet Physical Activity Series Working Group, 2012. The implications of megatrends in information and communication technology and transportation for changes in global physical activity. Lancet 380, 282–293.

Prochaska, J.O., DiClemente, C.C., Norcross, J.C., 1992. In search of how people change: applications to addictive behaviors. American Psychologist 47 (9), 1102–1114.

Rabinowitz, P., 2013. Coalition building I: starting a coalition. In: Wolff, Tom (Ed.), Community Tool Box: Part B. http://ctb.ku.edu/tools.

Reis, R.S., Hallal, P.C., Parra, D.C., Ribeiro, I.S., Brownson, R.C., Pratt, M., Hoehner, C.M., Ramos, L.R., 2010. Promoting physical activity through community-wide policies and planning: findings from Curitiba, Brazil. Journal of Physical Activity and Health 7 (Suppl. 2), S137–S145.

Rosenstock, I., 1974. Historical origins of the health belief model. Health Education Monographs 2, 328–335.

Roux, L., Pratt, M., Tengs, T.O., Yore, M.M., Yanagawa, T.L., Van Den Bos, J., Rutt, C., Brownson, R.C., Powell, K.E., Heath, G., Kohl 3rd, H.W., Teutsch, S., Cawley, J., Lee, I.M., West, L., Buchner, D.M., 2008. Cost effectiveness of community-based physical activity interventions. American Journal of Preventive Medicine 35 (6), 578–588.

Stokols, D., Grzywacz, J.G., McMahan, S., Phillips, K., 2003. Increasing the health promotive capacity of human environments. American Journal of Health Promotion 18 (1), 4–13.

van Sluijs, E.M., van Poppel, M.N., Twisk, J.W., et al., 2005. Effect of a tailored physical activity intervention delivered in general practice settings: results of a randomized controlled trial. American Journal of Public Health 95, 1825–1831.

US Department of Health and Human Services, 2002. PA Evaluation Handbook. US Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta, GA.

US Department of Health and Human Services, 1996. Physical Activity Health: A Report of the Surgeon General. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Atlanta, GA.

White, S.M., Wójcicki, T.R., McAuley, E., 2012. Social cognitive influences on physical activity behavior in middle-aged and older adults. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 67 (1), 18–26.

World Health Organization, 2010. Global Recommendations on Physical Activity for Health. WHO Press, Geneva, Switzerland.

World Health Organization, 2007. A Guide for Population-Based Approaches to Increasing Levels of Physical Activity: Implementation of the WHO Global Strategy on Diet, Physical Activity, and Health. WHO Press, Geneva, Switzerland.

Youth Risk Behavior Surveillance System, 2007. National YRBS Overview. www.cdc.gov/HealthyYouth.yrbs.

Relevant Websites

Foundations of Physical Activity Promotion

http://ec.europa.eu/sport/ – Sport – European Commission (accessed 06.01.15).

http://www.health.gov/paguidelines/ – The Office of Disease Prevention and Health Promotion. Physical Activity Guidelines (accessed 03.07.13.).

http://www.lancetphysicalactivityobservatory.com/ – Lancet Physical Activity Observatory (accessed 03.07.13.).
http://www.who.int/dietphysicalactivity/factsheet_adults/en/ – WHO. Physical Activity and Adults (accessed 03.07.13.).

Behavioral and Social Strategies for Promoting Physical Activity

http://activelivingresearch.org/ – Active Living Research (accessed 03.07.13.).
http://www.cdc.gov/physicalactivity/everyone/guidelines/index.html – CDC. Physical Activity for Everyone (accessed 03.07.13.).
http://my.americanheart.org/professional/ScienceNews/AHA-Guide-for-Improving-CV-Health-at-the-Community-Level-2013-Update_UCM_450446_Article.jsp – AHA Guide for Improving CV Health at the Community Level, 2013 Update (accessed 03.07.13.).
http://www.phac-aspc.gc.ca/hp-ps/hl-mvs/pa-ap/index-eng.php – Public Health Agency of Canada. Physical Activity (accessed 03.07.13.).
http://www.who.int/dietphysicalactivity/PA-promotionguide-2007.pdf – WHO. A Guide for Population-Based Approaches to Increasing Levels of Physical Activity (accessed 03.07.13.).

Building Community Capacity to Promote Physical Activity

http://www.cdc.gov/eval/framework/index.htm – CDC. A Framework for Program Evaluation (accessed 03.07.13.).
http://www.cdc.gov/nccdphp/dnpa/physical/handbook/pdf/handbook.pdf – CDC. Physical Activity Evaluation Handbook (accessed 03.07.13.).
http://ctb.ku.edu/en/default.aspx – Community Toolbox, (accessed 03.07.13.).

Evidence-Based Community-Level Strategies for Promoting Physical Activity

http://www.activelivingbydesign.org/ – Active Living by Design (accessed 03.07.13.).
http://www.thecommunityguide.org/pa/index.html – The Community Guide – Increasing Physical Activity (accessed 03.07.13.).
http://www.who.int/dietphysicalactivity/PA-promotionguide-2007.pdf – WHO. A Guide for Population-Based Approaches to Increasing Levels of Physical Activity (accessed 03.07.13.).

Physical Sciences: History and Sociology

Richard Staley, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Krige, volume 17, pp. 11418–11422, © 2001, Elsevier Ltd.

Abstract

Despite substantial differences, at different times in their development the physical sciences have provided a pattern for scientific work in the social and behavioral sciences as well as important theoretical resources and methods of working, and novel forms of social organization. This brief history pays particular attention to the changing social organization of the physical sciences, the development of sociological study of their activities, and the relations that were established with the social and behavioral sciences in the nineteenth century, when present disciplines in these subjects were beginning to form.

Origins of the Physical Sciences

Questions about the historical origins of the sciences have often rested on understandings of the nature of science and their relations to social systems, and answers to them have changed quite radically over time. From the early twentieth century to the 1960s, scholars emphasizing theory and rationalism typically argued for ancient Greek origins and often focused on the geometric (kinematic) models of planetary motion characteristic of Ptolemaic astronomy. More recently scholars have drawn attention to the work of Sumerian and Babylonian astronomers, whose records dating from the seventh century BCE show the use of predictive linear arithmetical sequences to calculate the periodicity of the day and planetary motions, and who related astronomical observations to astrological divination without closely linking astronomy and cosmology as the Greeks did. Babylonian influences on Indian and Greek astronomy had long been recognized. It is now clear that unjustifiably strict dichotomies between theoretical and practical, and rationalist and religious interests had been central to the argument that Babylonian contributions were merely practical, religiously motivated, and prescientific (Rochberg, 2002). Building on specialist studies of the regions concerned, comparative approaches have also begun to bring new complexity to our understanding of the pursuit of astronomy and mathematics in ancient China, as well as Greece and Mesopotamia. Where social and cultural factors are invoked to explain differences, most often they are mentioned in connection with the contrast between a relatively centralized and state-funded endeavor in China (leading to stability but also the potential for conservatism), and the need for philosophers to find fee-paying audiences in Greece – something that promoted an interest in novelty and a more free market competition of ideas (a competition that could also approach chaos and lead to discontinuities) (Lloyd, 2002). In broad outline, a contrast of this kind has been pervasive in studies of the sciences, reaching through to the Cold War. However, in the detailed examination of particular subjects in particular periods there are too many exceptions to ready-made generalizations about the relations between open or closed societies and science to allow those generalizations to stand without nuance and qualification: science has been pursued successfully under many different social systems.

While ancient societies surely offer precedents, and in the case of astronomy a continuous history of concern and development, some historians would date the physical sciences in their present form only to the nineteenth century, which saw the rise of modern experimental disciplines within research universities in Europe and the Americas. Yet the basic concepts and founding institutions in the physical sciences trace back to the scientific revolution. The rise of the sciences in Europe in the early modern period depended in part upon the assimilation of new writings from ancient Greece but also upon existing scholarly traditions (developed in close relations with religious and professional training), rising populations, the expansion of trading networks (and increasing pace of trade, together with the introduction of credit), and the development of seafaring navigation. It also occurred in a context of continual warfare and increasingly democratic societies (Jacob, 2007, reviewing Park and Daston, 2006). Multifactorial understandings of this nature stress the complexity of the social environment in the period in which the institutions of science were formed and have shown convincingly that, despite the fact that the Catholic Church forbade teaching of the Copernican system and that the early societies or academies initiated to cultivate the sciences were formed outside existing universities, new knowledge was far more importantly formed in dialog with religious and educational traditions than in opposition to them. While oppositional tropes and a stress on novelty do pepper some writings of the period, scholars have found they often obscure the nature of earlier contributions (the medieval period was more productively innovative than has often been thought), and it is also important to note that many figures involved, including notably Bacon and Newton, believed they were engaged in the rediscovery of truths long lost rather than the creation of something thoroughly new (Lindberg, 1990).

While structurally similar to Aristotelian understandings in some respects, a view of the world emerged that relied on and gave distinctive emphasis to understandings of the physical sciences that bridged the earth and the heavens, utilized new instruments like the telescope, microscope, and air pump (extending the senses and creating novel environments), pursued an empiricism most sharply represented in private experimental trials and public demonstrations, and could in some instances be closely related to mathematics. The mechanical or corpuscular hypothesis that became widely

shared offered a new assessment of the relations between nature and artifice, emphasizing the possibility of knowing nature in the way that human machines like the mechanical clocks that had been developed since the thirteenth century were known. Intellectual and social tensions between practical and theoretical knowledge and artisanal, commercial, and elite contributions to natural philosophy are abundantly visible in the founding of the early scientific academies and in the development of experimental practices (and such tensions are still felt). Natural philosophy would provide useful knowledge, and this was supposed to be less authoritarian and more open than the scholarly knowledge professed to a passive audience in universities. Yet the audience at Royal Society meetings was circumscribed, Boyle and others routinely neglected to mention the servants and workmen who assisted them in the production of complex experiments, and business interests might be thought to weigh against the credibility of philosophical engagement. Establishing mutually beneficial patronage relations was also important. While the relations of the Royal Society with the British crown were largely symbolic, the Royal Academy of Science in Paris established a hierarchy of salaried and nonsalaried positions. Although many natural philosophers worked primarily in a scholarly setting (like Galileo and Newton), the social norms that most thoroughly informed the new practices being developed were those of the gentleman, and the most common site for early laboratories was the gentleman's home. It was a largely private place (Shapin, 1988, 1989).

Within the physical sciences there were tensions between diverse philosophical stances and methodological approaches, and these often relied upon quite different intellectual and material resources to develop novel claims. Both mathematical and experimental approaches were advanced, and authors such as Descartes, Hooke, and Boyle offered a variety of forms of mechanical philosophy. Traditions in natural magic and alchemy were also important for many participants, such as Boyle and Newton (again, recent scholarship has shown close relations between activities earlier scholars had thought opposed, such as chemistry and alchemy). Although communication of results through the journals founded for this purpose allowed a form of virtual witnessing and had the potential to distribute novel claims widely through the republic of letters, the rarity of some of the resources involved (such as particular forms of optical glass) as well as the necessity of advanced training or personal contact to impart particular techniques (like Newtonian fluxions) meant that diverse communities of practice developed. As a result the methods, aims, and findings of different natural philosophers were often contested, and the wide scope of the debates that took place in the period has provided fertile grounds for linking historical and sociological investigations.

In the early twentieth century, Robert Merton had built on Weber's study of the puritan ethic and capitalism to argue that rather than inhibiting science, English Puritanism had offered cultural values that supported the rise of the experimental sciences (Merton, 1938). He subsequently elaborated a form of sociology that investigated the ethos informing the organization of scientific activity, which he identified in the four values of communism, universalism, disinterestedness, and organized skepticism (*see* Strong Program, Sociology of). Merton paid close attention to those circumstances in which tensions between ethos and institution arose, but did not discuss the content of science in any detail (Merton, 1973[1942]). In contrast, the 'strong program' in the sociology of knowledge that British sociologists articulated in the 1970s opened truth claims to social analysis, combating the assumption that only false beliefs (or the more general norms Merton had analyzed) need to be explained sociologically. Central tenets were that sociological study should be causal, impartial to the truth or falsity of claims, offer a symmetrical form of sociological understanding (giving the same kind of explanation for true and false beliefs), and be reflexive (Bloor, 1991[1976]). Central techniques were to analyze the interests of participants; study controversies, which often emerge around contested attempts to replicate experiments, because they expose the tacit dimensions of scientific work (Collins, 1985); and investigate contemporary laboratories anthropologically, something that led to a form of actor network analysis (Latour and Woolgar, 1979; Latour, 1987). While sociologists and anthropologists developed case studies of recent or current science (Collins notably studied the replication of a laser instrument, and research on gravity waves, while Traweek compared high energy laboratories in the United States and Japan (Traweek, 1988)), historians of early modern science have also drawn significantly on these approaches. Shapin and Schaffer's study of the debates between Boyle and Hobbes was particularly important, showing that diverse understandings of the relations between religion, the state, experimental philosophy, and the status of knowledge derived from the air pump were all at stake in accounts of specific facts; and describing material, literary, and social technologies in the production of knowledge (Shapin and Schaffer, 2011[1985]). Other influential work exploring historical episodes with importantly sociological aims include studies of the relations between instrument and interpretation in the debate around Newton's prism experiments and acceptance of his theory of light and color, and of patronage in Galileo's research and of his trial (Schaffer, 1989; Biagioli, 1993). While 'social constructivism' has sometimes proved controversial, forms of methodological relativism focused on recovering actors' understandings of their activity as a first step in analysis have been widely important in historical work. Perhaps the most general sociological issue underlying changing historiography of the period has been that of boundary work (the historically and situationally relative stipulation of demarcations between science and nonscience), understood across diverse periods. In many cases the sharp distinctions between sciences and other activities pointed to above have been found to reflect later perspectives, rather than contemporary concerns (Gieryn, 1983). One challenging argument has been that Boyle's understanding of mechanism and his gas law embody gendered perspectives, thereby closely relating gender studies to the substance of chemistry and not only the more commonly addressed question of participation in the physical sciences (Potter, 2001; Kovacs, 2012a,b).

Newtonianism over Time and Physics as One Discipline among Others

Patronage, institutional context, and audience have all been shown to play a formative role in the development and propagation of particular research fields, techniques, and knowledge

claims. The fortunes of Newton's work in different contexts illustrate this point, indicating the multifaceted nature of scientific work, and also just how radically understanding of its potential implications can change over time. His *Principia Mathematica* (1687) and *Opticks* (1704) offered rather different models for natural philosophers, the first of a mathematical physics that depended on sophisticated mathematics and needed advanced training to be understood (but also, with book 2 of highly practical pursuits – mechanics rather than astronomy); the second of a speculative experimental physics. In Britain, Newton's views and techniques were propagated in both universities and dissenting academies, and his fluxions were taught, rather then Leibniz's form of differential and integral calculus. Followers advanced Newton's views in philosophical discussions and sermons, where Protestantism was reinvented on the basis of the new sciences and Newton's mechanics might offer a model for good government. Desaguliers, the leading demonstrator and public lecturer at the Royal Society in the 1720s and 1730s applied Newton's mechanics to the practical work of draining mines and swamps, and his spectacular experimental demonstrations offered a form of Newtonianism accessible to less mathematically gifted audiences (Jacob, 1976; Stewart, 1992). In France, where Cartesian models were accepted, reading and propagating Newton relied on translating his language, reworking his mathematics into Leibnitzian form and transforming French cartography to critique existing measures and substitute new values for the length of a degree of arc at different latitudes, showing that the Earth bulged at the equator as Newton predicted. Voltaire and Maupertuis were particularly important in this process, and both had traveled to Britain. Maupertuis addressed a wide range of audiences from the academy to the salon, and studies of Émilie du Châtelet (the author of the first and still standard French translation of the *Principia*) have helped disclose the gendered nature of the spaces in which natural philosophy was pursued (Terrall, 1995, 2002).

Over time, and in success, Newton's work came to stand for quite different things. Similarly, as is best illustrated by the gradual transformation of electricity from being a property of amber to something created, stored, and sometimes tamed in electrical machines, Leyden Jars and lightning, the phenomena explored by the physical sciences were also being chased into diverse aspects of nature and integrated into quite different intellectual and disciplinary constellations. By the end of the eighteenth century, Coulomb had used Newtonian distance laws to mathematize electrostatics and Laplace had extended his mathematics analytically and shown the stability of the solar system; and Newton's work was now being invoked as an example of human rationalism and symbol of the Enlightenment, rather than as providing illustrations of God's active role in the universe. At the same time Galvani's research with frogs and Volta's discovery of the battery were leading other researchers to link chemistry and physics in electrolysis, physics and biology in the study of the electric current in the nervous system, and later electricity and magnetism in a search for the interconversion of different forces that had distinctly romantic aims (Heilbron, 1999). In this period chemistry was the most important linking discipline, and Newtonian models were clearly regarded as limited. For Faraday, gravitation was a force apart that might be brought into a continuum with chemical, electrical, and magnetic forces propagating over time through a medium, rather than instantaneously at a distance. Over time his work helped found field theory, and a widespread concern with conversion processes contributed to what became energy physics; significant cultural and religious dimensions played into its development, and Smith has described a 'spiral of credibility' in the process of addressing wider audiences that helped propagate particular views beyond their origins in small groups of researchers (Smith and Wise, 1989; Smith, 1998).

Faraday both acted as a consultant for industry groups and addressed elite audiences who were intellectually and socially interested in natural philosophy in his Friday evening lectures at the Royal Institution (founded in 1799, initially to bring science to the lower classes). In contrast to his contemporary rival Sturgeon, who put electrical machines on show to model the world and stress his own role as a skilled maker, Faraday worked hard to yield simple experiments that would render natural laws visible to others: the experimenter must show he is a reliable spokesman for nature while representing nature as speaking for herself (Morus, 1998). Faraday's laboratory was a private place known to few, a place of research and not pedagogy. It is important to note that this was also the case in universities, where professors demonstrated phenomena in lectures before students who never entered the laboratory. This changed first in chemical teaching from the late 1820s through the work of Thomas Thomson in Glasgow and Justus Liebig in Giessen, and then in physics from the 1850s with the work of William Thomson (Lord Kelvin) in Glasgow (Gooday, 1990). It is highly significant that chemical industry and the development of electric telegraphy and the power industry around the steam engine provided markets for the kinds of skills that teaching laboratories could provide, and also that by working with their students these scientists created the first sustained schools of advanced research, generating collective work in the sciences of a kind and on a scale previously impossible. The thoroughgoing transformation of mathematical pedagogy from oral to written examinations in Cambridge achieved something similar in mathematical practice (Warwick, 2003). The characteristic scientific collective of the early modern period had been societies in the republic of letters (often looking for an audience); now it was taking the form of research schools in academic disciplines.

While teaching the skills required for advanced industrial society (or the teachers who could convey those skills) was one means for the physical sciences to address industry – and a need for this was highlighted in response to the widespread commercial competition expressed in the international exhibitions of the nineteenth century – figures like Kelvin and Maxwell also sought to determine and calibrate the units to be used in industry. This could be achieved in a variety of ways, and across many different fields of physics, from electrical standards important to the electric telegraphy and power industry through optical standards useful in research on heat and light sources, and standards of length that could be traced from international and national standards down to the gauges controlling machine tools and precision manufacturing. Determining metrological standards provided a second, highly

strategic means through which academic science could be rendered central to industry, and it was particularly significant for the disciplinary development of physics in the period in which physical laboratories first became widespread in the late nineteenth century (Schaffer, 1999). Following the example of the Physikalisch-Technische Reichsanstalt founded in Berlin in 1887, from the turn of the century leading industrial nations founded national standards laboratories to conduct and regulate such metrical work, while also undertaking related research (often in quite close contact with academic research) (Cahan, 1989). Later, military needs stimulated the creation of national laboratories to provide weapons and other critical resources.

A third activity relating the academic laboratory to industry has been the development of commercial instruments and products, where patenting and entrepreneurial activity in partnership with instrument makers or manufacturing firms has been characteristic. This too is notable in the nineteenth century (Kelvin's electrical instruments and relations with James White being a prime example), but it only received greater prominence for the academic sciences as a whole in the later twentieth century, when the distribution of patent profits to the university came to be regarded as an important source of funding, something that has been analyzed more fully in relation to biotechnology than the physical sciences. However, scholars have described the generic forms of instrumentation that often facilitated such innovation, and analyzed the ambiguous social place of the research technologists who have provided it, often working at the interstices of different academic disciplines and industry (Joerges and Shinn, 2001).

There are two important ways in which work in the physical and the social sciences were interrelated through the course of the nineteenth century, both being expressed in the academic foundations of the social sciences as these became disciplines at the turn of the century. The first has received widespread attention in studies of the rise of statistical thinking. In 1809 Gauss had analyzed the errors of astronomical observation, finding them to fit a distribution law; in the 1830s Quetelet moved from astronomy to social science, and applied Gauss's distribution to other phenomena like the chest measurements of Scots, and social statistics like birth, death, and crime rates. Looking for the laws of the average man he sought a social physics that would rest on the characteristics of the social body rather than individuals. It was only after statistics had been taken up in the emerging social sciences in this way that physicists like Maxwell and Boltzmann explored its application in the kinetic theory of gases (Porter, 1986; Stigler, 1986). Some physicists took the kinetic theory of gases to illustrate the increasing power of Newtonian mechanics. Yet many regarded physics to offer models rather than final truth, and by the turn of the century there was a widespread search for alternative foundations as well as antireductionist and phenomenological epistemologies. Mach, Pearson, and others popularized a vision of science as offering simply the most economical description of sensations. Alongside the statistical methods strongly advocated by Galton and Pearson, their epistemological modesty and emphasis on a common method across all forms of science was eagerly taken up by those advancing the social sciences as academic disciplines, especially in the United States (Porter, 1994).

Twentieth-Century Physics: Particles, Projects, and Panels

Physics became a laboratory discipline at the same time that physiology and experimental psychology did, and from the 1860s researchers like Helmholtz and Mach had linked each of these disciplines by combining studies of light, sound, and movement with research on the senses that perceived them. By 1900 X-rays, the electron and radioactivity were offering a vision that could penetrate surfaces, and relativity and quantum theory combined mathematical and experimental physics in an investigation of matter at extreme velocities and in atomic realms previously often thought inaccessible. For many, particle physics became the defining field of physics, and the disciplinary linkages that were closest to home were those with chemistry, through physical chemistry and then the role of the electron in quantum chemistry, and also with molecular biology, especially after the World War II. Such relations have been mediated by instruments such as X-ray diffraction apparatus and the electron microscope, as well as by theories such as quantum chemistry and the informational approach to biology. They also were mediated by people who linked diverse fields. In the case of molecular biology, transfers of instruments, techniques, and practitioners were facilitated by the funding choices of the Rockefeller Foundation, which sought to purge what they regarded as metaphysics from the study of life and encouraged physical approaches to biological research. However, despite hierarchical understandings of the disciplines, interchanges of this kind have also been pursued by antireductionists, and characterized by fruitful engagement between both approaches (Gavroglu and Simões, 2012; Kay, 1993, 2000).

Military needs stimulated closer relations between the physical sciences and the State throughout the Western world from World War I onward, and in World War II helped create a distinctive social form and new collective activity we can describe as project- or problem-oriented science. Different nations created interdisciplinary constellations directed toward the provision of clearly defined material goals – famously in the United States, radar and atomic bombs – bringing members of physical, chemical, engineering, and material science disciplines together with industrial corporations, all under military control. While researchers in industrial firms had earlier experienced somewhat similar constraints and advantages, the experience unsettled established hierarchies, created new funding expectations and involved scientists more extensively in managerial roles than that had been common previously. While some welcomed the changes to previous academic practice, others rued them, and the postwar political environment gave rise to new concerns. Studies of McCarthyism in the United States and of cybernetics in different nations have disclosed a distinctive tension between ideology and science that played out rather differently across Western and Soviet practice, with elements of an interpretation regarded as fundamental in one context sometimes being understood as ideologically distorted in the other (with these judgments also changing over time). They also have shown that, just as the proper social order of science has often been thought to rest preferentially in politically favored structures (whether they be democratic or communist), scientists in the

physical as well as the biological sciences have also found creative resources in social and political analogy. Examples include the elaboration of a 'democratic' bootstrapping form of nuclear physics developed with some success in the United States, and diverse treatments of the concept of information (Wang, 1999; Kaiser, 2002; Mindell et al., 2003).

Project science had major legacies for the conduct of physics after the World War II, with the continuation of weapons laboratories being accompanied by the development of accelerators that concentrated experimental work in a few major sites. In the development of the original atomic bombs, secrecy was not just a question of access but went along with compartmentalization as a managerial strategy; anthropological studies of the postwar weapons laboratory have examined the mix of political silence and technical passion that have sustained life on the other side of the security fence (Gusterson, 1996). When not connected with weapons work, particle physics is of little immediate economical value, so its results could be circulated freely. Yet it requires substantial investment. In the United States, the accelerators were initially understood to provide critical manpower, training the strategic elite required to produce future weaponry. Accordingly, as the Cold War priorities changed or were abandoned, particle physics went through boom and bust. In the boom period, groups of hundreds of researchers and technicians were formed to design, build, and run massive instruments and experiments lasting over decades, often coming together from widely dispersed home institutions to share beam time and floor space around increasingly elaborate detectors. Instruments and disciplinary organization alike were thoroughly reshaped in these extraordinary transformations of scale, and in the course of a career physicists now commonly move through a sharply calibrated hierarchy superimposed on differentiated but interrelated fields of work, transitioning from working as postdocs on defined task (yet needing to challenge expectations to win notice) to perhaps reach the status of project leaders who spend their time in bureaucratic work and searching for funds (Galison, 1997; Traweek, 1988).

In times of bust, physicists were forced to leave the field – sometimes for the financial sector – but they also saw the value of diversifying the training of their graduates. Retrospectively we can see that in at least one case this has helped overcome the costs of subdisciplinary specialization and balkanization. Following the decision to broaden the pedagogy offered prospective particle physicists in the 1970s, the work of savvy students who experienced teaching across both particle physics and cosmology allowed a unification of similar approaches to mass – the Brans–Dicke field in gravitational studies, and the Higgs field in particle physics – that had been published in the same journal while rarely being read by the same people (Kaiser, 2007). Although Europe was slow to gain an accelerator comparable to those developed in the United States, over time the political environment within which the European nuclear agency European Organization for Nuclear Research (CERN) was formed has allowed it to lead a more stable existence. Founded as an instrument of foreign policy by western European nations ready to devote national resources to the endeavour of promoting wider European cooperation, it is one of several examples of collectively international laboratories and organizations, most often achieved in fields with little direct strategic importance, such as space science, molecular biology, or nuclear fusion (Krige and Guzzetti, 1997). While the American physics community saw its hopes for a super conducting super collider dashed at the end of the Cold War, the work of two groups at the Large Hadron Collider at CERN in 2012 may have provided the first experimental evidence for the Higgs Boson, thereby demonstrating the success of the approaches to mass brought together in particle cosmology. Perhaps the best illustration of the unusual social achievement of work in particle physics is that this community has worked out ways for collaborations of thousands to author tightly reasoned papers representing work in common undertaken over many years (Galison, 2003). More than 5000 authors contributed to the two papers published by the Compact Muon Solenoid (CMS) and A Toroidal LHC Apparatus (ATLAS) collaborations at CERN, from several hundred home institutions.

Distributed authorship of a different kind illustrates the second novel social form that has been created by work in the physical sciences through the late twentieth century. Understanding the possibility of climate change over time has relied in part on laboratory studies as early as the 1860s that identified the fact that carbon dioxide was opaque to heat. Such research also relies in part on an understanding of the cycle of carbon through the atmosphere, oceans, and earth – and the potential for feedback loops to drive the climate to warm or cool – that was first stated in the 1890s, and in part on the recognition from the 1930s that emissions of carbon dioxide from industrial activity could be significant on a global scale. Different elements of this argument were offered by physicists, chemists, geologists, engineers, meteorologists, and oceanographers, a form of interdisciplinary collaboration that was developed piece by piece over many decades before it began to be widely persuasive, and which shows that discovery accounts provide a poor understanding of work in such complex fields (Weart, 2003). Demonstrating climate change has also relied upon the assessment of data on temperature, rainfall, barometric pressure, and the like, gathered daily by scientists working at observatories and laboratories to provide reliable weather predictions, with coverage of the globe gradually expanded over the course of the nineteenth and twentieth centuries, and then lifted into the atmosphere and space through the military environmentalism of the Cold War. The physical sciences have thus provided an infrastructure for agriculture, shipping, and air travel through meteorological observations. Always useful but long mundane, these now constitute a record of climate change that has been supplemented by such measures as tree growth and gas capture in ice to extend our understanding over much longer periods (Edwards, 2010). Yet the social and economic interests involved in building an argument that bears so closely on industry required more than interdisciplinary science and long-term climate records. The Intergovernmental Panel on Climate Change was formed by the World Meteorological Organization and the United Nations Environment Program in 1988 in order to provide governments with advice about the state of scientific understanding about climate change, its potential social and economic impacts, and possible response strategies. Recognizing the extent and character of the interests involved, its judgments are provided in reports that are regulated through a process of authorship and review that

includes government representatives as well as subject specialists. They represent the work required to build authority through meticulously distributed procedural authorship, and show that panel science is every bit as novel, challenging, and important a collective form of science as project science has proved to be.

See also: Biomedical Sciences, History and Sociology of; Experiment in Science and Technology Studies; Human Sciences, History of; Norms in Science; Science, History of; Science: Constructivist Perspectives, History of; Scientific Disciplines, History of; Scientific Knowledge, Sociology of; Strong Program, Sociology of.

Bibliography

Biagioli, M., 1993. Galileo, Courtier: The Practice of Science in the Culture of Absolutism. University of Chicago Press, Chicago/London.

Bloor, D., 1991[1976]. Knowledge and Social Imagery. University of Chicago Press, Chicago.

Cahan, D., 1989. An Institute for an Empire: The Physikalisch-technische Reichsanstalt, 1871–1918. Cambridge University Press, Cambridge/New York.

Collins, H., 1985. Changing Order: Replication and Induction in Scientific Practice. Sage, London.

Edwards, P.N., 2010. A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. MIT Press, Cambridge, MA.

Galison, P., 1997. Image and Logic: The Material Culture of Microphysics. University of Chicago Press, Chicago/London.

Galison, P., 2003. The collective author. In: Biagioli, M., Galison, P. (Eds.), Scientific Authorship: Credit and Intellectual Property in Science. Routledge, New York/London, pp. 325–355.

Gavroglu, K., Simões, A., 2012. Neither Physics Nor Chemistry: A History of Quantum Chemistry. MIT Press, Cambridge, MA.

Gieryn, T.F., 1983. Boundary-work and demarcation of science from non-science: strains and interests in professional ideologies of scientists. American Sociological Review 48, 781–795.

Gooday, G., 1990. Precision measurement and the genesis of physics teaching laboratories in Victorian Britain. British Journal for the History of Science 23, 25–51.

Gusterson, H., 1996. Nuclear Rites: A Weapons Laboratory at the End of the Cold War. University of California Press, Berkeley.

Heilbron, J.L., 1999. Electricity in the 17th and 18th Centuries: A Study in Early Modern Physics. Dover, Mineola, NY.

Jacob, M.C., 1976. The Newtonians and the English Revolution, 1689–1720. Gordon and Breach, New York.

Jacob, M.C., 2007. The importance of early modern European science and the state of the field. Isis 98, 361–365.

Joerges, B., Shinn, T. (Eds.), 2001. Instrumentation: Between Science, State, and Industry. Kluwer Academic Publishers, Dordrecht/Boston.

Kaiser, D., 2002. Nuclear democracy: political engagement, pedagogical reform, and particle physics in postwar America. Isis 93, 229–268.

Kaiser, D., June 2007. When fields collide. Scientific American, 62–69.

Kay, L.E., 1993. The Molecular Vision of Life: Caltech, the Rockefeller Foundation, and the Rise of the New Biology. Oxford University Press, New York.

Kay, L.E., 2000. Who Wrote the Book of Life? A History of the Genetic Code. Stanford University Press, Stanford.

Kovács, A.G., 2012a. Gender in the substance of chemistry, Part 1: the ideal gas. HYLE – International Journal for Philosophy of Chemistry 18, 95–120.

Kovács, A.G., 2012b. Gender in the substance of chemistry, Part 2: an agenda for theory. HYLE – International Journal for Philosophy of Chemistry 18, 121–143.

Krige, J., Guzzetti, L., 1997. History of European Scientific and Technological Cooperation. Office for Official Publications of the European Communities, Luxembourg.

Latour, B., 1987. Science in Action: How to Follow Scientists and Engineers through Society. Harvard University Press, Cambridge, MA.

Latour, B., Woolgar, S., 1979. Laboratory Life: The Social Construction of Scientific Facts. Sage Publications, Beverly Hills.

Lindberg, D.C., 1990. Conceptions of the scientific revolution from bacon to butterfield: a reappraisal. In: Lindberg, D.C., Westman, R.S. (Eds.), Reappraisals of the Scientific Revolution. University of California Press, Berkeley.

Lloyd, G.E.R., 2002. The Ambitions of Curiosity: Understanding the World in Ancient Greece and China. Cambridge University Press, Cambridge/New York.

Merton, R.K., 1938. Science, technology and society in seventeenth century England. Osiris 4, 360–632.

Merton, R.K., 1973[1942]. The normative structure of science. In: Storer, N. (Ed.), The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago Press, Chicago, pp. 267–278.

Mindell, D., Segal, J., Gerovitch, S., 2003. From communications engineering to communications science: cybernetics and information theory in the United States, France, and the Soviet Union. In: Walker, M. (Ed.), Science and Ideology: A Comparative History. Routledge, London/New York, pp. 66–95.

Morus, I.R., 1998. Frankenstein's Children: Electricity, Exhibition, and Experiment in Early-Nineteenth-Century London. Princeton University Press, Princeton.

The Cambridge history of science. In: Park, K., Daston, L. (Eds.), 2006. Early Modern Science, vol. 3. Cambridge University Press, New York.

Porter, T.M., 1986. The Rise of Statistical Thinking, 1820–1900. Princeton University Press, Princeton.

Porter, T.M., 1994. The death of the object: Fin de Siècle philosophy of physics. In: Ross, D. (Ed.), Modernist Impulses in the Human Sciences, 1870–1930. Johns Hopkins University Press, Baltimore, pp. 128–151.

Potter, E., 2001. Gender and Boyle's Law of Gases. Indiana University Press, Bloomington.

Rochberg, F., 2002. A consideration of Babylonian astronomy within the historiography of science. Studies in the History and Philosophy of Science 33, 661–684.

Schaffer, S., 1989. Glass works: Newton's prisms and the uses of experiment. In: Gooding, D., Pinch, T., Schaffer, S. (Eds.), The Uses of Experiment: Studies in the Natural Sciences. Cambridge University Press, Cambridge, pp. 67–105.

Schaffer, S., 1999. Late Victorian metrology and its instrumentation: a manufactory of ohms. In: Biagioli, M. (Ed.), The Science Studies Reader. Routledge, New York, pp. 457–477.

Shapin, S., 1988. The house of experiment in seventeenth-century England. Isis 79, 373–404.

Shapin, S., 1989. The invisible technician. American Scientist 77, 554–563.

Shapin, S., Schaffer, S., 2011[1985]. Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life. Princeton University Press, Princeton.

Smith, C., 1998. The Science of Energy: A Cultural History of Energy Physics in Victorian Britain. University of Chicago Press, Chicago.

Smith, C., Wise, M.N., 1989. Energy and Empire: A Biographical Study of Lord Kelvin. Cambridge University Press, Cambridge/New York.

Stewart, L.R., 1992. The Rise of Public Science: Rhetoric, Technology, and Natural Philosophy in Newtonian Britain, 1660–1750. Cambridge University Press, Cambridge.

Stigler, S.M., 1986. The History of Statistics: The Measurement of Uncertainty before 1900. Belknap Press of Harvard University Press, Cambridge, MA.

Terrall, M., 1995. Emilie du Chatelet and the gendering of science. History of Science 33, 283–310.

Terrall, M., 2002. The Man Who Flattened the Earth: Maupertuis and the Sciences in the Enlightenment. University of Chicago Press, Chicago.

Traweek, S., 1988. Beamtimes and Lifetimes: The World of High Energy Physicists. Harvard University Press, Cambridge, MA.

Wang, J., 1999. American Science in an Age of Anxiety: Scientists, Anticommunism, and the Cold War. University of North Carolina Press, Chapel Hill.

Warwick, A., 2003. Masters of Theory: Cambridge and the Rise of Mathematical Physics. University of Chicago Press, Chicago/London.

Weart, S.R., 2003. The Discovery of Global Warming. Harvard University Press, Cambridge, MA.

Physicalism and Alternatives

Brian P McLaughlin, Rutgers University, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Physicalism is the doctrine that our world is fundamentally physical in nature. Substance dualism is an alternative doctrine, according to which there are certain fundamental, irreducible substances (objects) that are not physical – immaterial minds, entelechies, souls, or disembodied spirits. Property (or attribute) dualism, another alternative, maintains that there are fundamental properties that are not physical – spiritual properties or mental properties. Physicalism is a controversial doctrine, even among theorists who deny there is any supernatural realm. Various alternative doctrines to physicalism currently receive some currency, but none receives widespread acceptance. Whether our world is fundamentally physical remains a matter of dispute.

Physicalism

Physicalism is an ontological doctrine. It concerns what there is. There is no consensus about how, exactly, it should be formulated. The strongest formulation is that everything is physical. This formulation is sometimes weakened by restricting the quantifier word 'everything' to contingent existents, that is, to things that exist, but could have failed to exist. That allows, for instance, that there can be nonphysical Platonic universals, which are by hypothesis abstract necessary existents, provided that nothing in our world instantiates them. But many self-professed physicalists regard even this weaker formulation as too strong. They often do so, in part, on the grounds there are certain sorts of nonphysical, functional properties and kinds that have instances in our world, and that this should be compatible with physicalism properly formulated. There are other formulations of physicalism, which receive some currency, that are weaker than the blanket claim that literally everything is physical. Some appeal to supervenience, stating that the nonphysical must supervene on the physical, so that there can be no nonphysical difference without a physical difference (McLaughlin and Bennett, 2011). Some appeal instead to some asymmetric relation that the nonphysical is supposed to bear to the physical – realization, grounding, or constitution. No formulation of physicalism receives widespread acceptance as definitive (Stoljar, 2009). Before examining further formulations of physicalism, however, let us turn to the prior question of what counts as physical.

The Physical

Any doctrine of physicalism must presuppose some conception of what counts as physical. But there is no received view about what counts as physical (Montero, 2009).

Descartes held that the physical is extended in space and that the mental is not; he held the mental is in time, yet nonspatial. If, however, general relativity can be trusted, then nothing can be in time without being in space.

It will not do to say that something is physical if and only if it is in space-time. Many contemporary physicalists reject nominalism, the doctrine that there are no universals or abstract entities of any sort (Rodriguez-Pereyro, 2011). A physicalist can maintain that there are physical universals (e.g., universals cited in physical laws) and that they, as opposed to their instances, are not in space-time, and so that there are physical existents that are not in space-time. Also, physicalists might count mathematical entities such as pure sets (sets that contain only sets as members that contain only sets as members, and so on) as physical existents, despite the fact that pure sets are not in space-time, on the grounds that physics must posit such sets, given its essential use of classical mathematics. Furthermore, some physicists speculate that space-time is not fundamental; though they do not mean to deny, thereby, that what is fundamental is physical. Existing or occurring or obtaining in space-time is not conceptually necessary for being physical. Nor is it conceptually sufficient. Human thinking and feeling, for instance, occur in space-time (wherever the thinker or feeler is), but it does not follow from this that such mental events are physical events, or that the spatiotemporally located thinkers and feelers are physical objects.

Although in ordinary discourse the physical is often contrasted with the mental and the spiritual, it is inadequate to so characterize it. As concerns the spiritual, in a theological sense of 'spiritual' according to which the spiritual is supernatural (rather than mental), naturalism, roughly the view that the natural world is in principle understandable by scientific method and that nothing exists beyond the natural world, denies the existence of the spiritual since it denies that there is any supernatural realm (Papineau, 1993). But not all naturalists are physicalists. Some naturalists maintain, for instance, that although there is no supernatural realm and the world is understandable by scientific method, consciousness exists, is not physical, and moreover fails to be dependent on the physical in the way that nonphysical phenomena must be for physicalism to be true (Chalmers, 1996).

Physicality cannot be adequately characterized by contrast to the mental either. Thus, it will not do to say that something is physical just in case it is not mental. Intuitionists in mathematics maintain that mathematical entities and properties are mental constructs (Lemhoff, 2013), while mathematical Platonists deny that maintaining instead that mathematical entities and properties are mind-independent (Linnebo, 2013). Platonism in mathematics is not thereby committed to the view that mathematical entities and properties are physical.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63060-X

Moreover, the stipulation that something is physical just in case it is not mental excludes a position that some physicalists want to hold, namely, there are mental phenomena, but that everything that is mental is physical. It also excludes idealism, the alternative to physicalism according to which everything, including the physical, reduces to the mental (Robinson, 2009). And it excludes other alternatives to physicalism as well. Proto-panpsychism is the doctrine that everything concrete (everything in space-time) has (yet to be discovered) protomental properties that are themselves neither mental nor physical properties, but that, in combination with physical properties, constitute mental properties, so that having a mental property just consists in having a combination of physical and protomental properties (Nagel, 1979; Seager, 2009). According to neutral monism, yet another alterative to physicalism, both mental and physical properties are ultimately to be explained in terms of neutral properties that are neither mental (nor protomental) nor physical (Stubenberg, 2013). The stipulation that something is physical just in case it is not mental would thus end the debate between physicalists, idealists, proto-panpsychists, and neural monists before it gets interesting. None of these alternatives to physicalism need be true for this objection to the characterization of physicality in question to stand. If any of these alternatives is even just conceptually coherent (i.e., not a priori false), then it is not a conceptual truth that something is physical just in case it is not mental.

One approach to characterizing physicality that has some currency appeals to exemplars of ordinary physical or material objects (e.g., rocks), and characterizes physical properties and relations (which on some views are universals) as follows: a property or relation is a physical property or relation just in case a complete account of the intrinsic nature of such physical objects and their microconstituents would have to appeal to it (Jackson, 1998). This proposal is inadequate. It too ends the physicalist's conversation with idealists, neutral monists, and proto-panpsychists before it gets interesting. Also, panpsychism is typically taken to be the view that every concrete object, no matter at what scale, has mental properties, that mental properties are not physical properties, and that mental properties are required to give a complete account of the intrinsic nature of exemplars of ordinary physical objects and their microconstituents (Nagel, 1979; Seager, 2009). If panpsychism is even just conceptually coherent, then figuring in a complete account of such physical objects and their microconstituents does not conceptually suffice for a property or relation's being physical. (It should be mentioned that at least one panpsychist employs the notion of physicality in question, but drops the claim that mental properties are not physical properties, and so maintains that panpsychism is a version of physicalism (Strawson, 2008). But the vast majority of physicalists would take this version of panpsychism, according to which certain mental properties are fundamental properties, to be an alternative to physicalism, a kind of property or attribute dualism (dualism all the way up and down the micro–macro scale), rather than a version of physicalism.)

A currently leading approach to characterizing the notion of physicality relevant to physicalism appeals to the explicit posits of physics. This appeal is made, in part, because, unlike any of the other sciences, physics aims to be a closed theory in the following sense: it aims to be a theory such that every caused state or event within its domain is such that there is a state or event within its domain that determines its objective probability (Davidson, 1970; Yablo, 1992). If causal determinism is true, that objective probability will be one; if causal indeterminism is true, it will be less than one.

Any approach that ties the notion of physicality to the posits of physics faces 'Hempel's dilemma' (Hempel, 1969; see also Crane and Mellor, 1990). If by 'physics' is meant current physics, then there is a problem if physicalism is characterized in terms of this notion of physicality, namely that current physics is likely false. If, on the other hand, by 'physics' is meant the ideally completed physics that is in fact true of our world, then to characterize the doctrine of physicalism in terms of this notion of physicality, the objection goes, would render the doctrine too obscure to be a subject of rational assessment, for there is no telling now what an ideally completed physics will have to posit.

Although current physics is enormously successful, it no doubt contains errors, and is by no means the final theory of physics; there is at present no adequate quantum theory of gravity, for instance. The first horn of the dilemma seems thus best avoided. What, then, of the second horn?

The problem raised by the second horn is not just that there is no telling (if there is indeed no telling) what an ideally completed physics will have to posit, and so that this notion of physicality renders physicalism too obscure to be a subject of rational debate. It is also the case that there is reason to deny that being posited by an ideally completed physics is a conceptually sufficient condition for being physical in the sense relevant to physicalism. Consider a far out possibility, but a possibility nonetheless. Suppose that it turns out that the physics true of our world is a Cartesian-like physics in that it posits nonspatial, but temporal minds that directly causally influence brains. Although that is incompatible with general relativity and it seems, moreover, preposterous to think that current physics has gone that badly wrong, it is nevertheless not a priori false that the physics that is in fact true of our world is such a Cartesian-like physics. Cartesian physics was not refuted by an a priori proof. It may be preposterous to think that the true, complete physics will be such a Cartesian-like physics, but it is not true by definition that it is not. Suppose, then, that the actual world contains Cartesian minds, substances that are temporal but not extended in space, and that occupy intrinsic, conscious states that exert a fundamental causal influence on the motions of particles in the brains of human beings, and, thereby, a causal influence on the behavior of human beings. Were that so, then the true, complete physics would be a Cartesian-like physics in that it would posit Cartesian minds. The reason is that mechanics, the theory of motion, is a central branch of physics. A true, complete physics would thus include a true, complete mechanics. A true, complete mechanics would take into account all of the fundamental causes of motion. Should states of a Cartesian mind be among those fundamental causes of motion, then a completed mechanics would posit them, and so physics would posit them. In one sense of 'physical', they would thereby count as physical. But that sense of 'physical' is not the one physicalists have in mind. Physicalists do not want to say that were that to turn out to be the

case, then such Cartesian minds are physical. Rather, they would say that were that to turn out to be the case, then physicalism is false since Descartes's substance dualism is true. It thus fails to capture the notion of physicality that typical physicalists are aiming for to say that something is physical if it would be posited by the ideally completed physics true of our world. That does not conceptually suffice for being physical in the sense that they have in mind. What, then, is that sense?

There is no received view about that. But the leading physicalist strategy for dealing with the second horn of Hempel's dilemma is to maintain that the physics that is fact true of our world will be a recognizable descendant of current physics (Pettit, 1993; Lewis, 1994; Jackson, 2007). That is an a posteriori claim that could prove false. Further, it is not preposterous to think that current physics may be very distant from the ideally completed physics that it is true of our world. It could turn out that the ideally completed physics will be as different from current physics as the air, earth, fire, and water chemistry of antiquity is from quantum chemistry. It is generally acknowledged, however, that the doctrine of physicalism should be an empirical doctrine that could prove false. So, it is no objection to this characterization of physicality that a version of physicalism that deploys it could turn out to be false. Moreover, many contemporary physicalists are willing to maintain that physicalism is true only if the physics that is in fact true of our world will be a recognizable descendant of current physics.

They are willing to do so, because they take the empirical case for physicalism to rely, in part, on the success of the atomic theory of ordinary matter (including living matter), the success of quantum mechanics in explaining chemical bonding, the success of organic chemistry in explaining certain biological phenomena, and the fact that it appears that all of the fundamental forces of nature can be found below the level of the atom. They maintain that, on the evidence, we are wholly composed of cells, which are composed of molecules, which are composed of atoms, which are composed of electrons, protons, and (with the exception of hydrogen) neutrons, and that neurons and protons are composed of quarks. They maintain further that, on the evidence, the computational processes that occur in our brains are implemented by neurophysiological processes, which are implemented by chemical processes, which in turn are implemented by quantum mechanical processes involving subatomic particles. Moreover, they maintain that the objective probability of the physical movements of any living organism or sentient creature, and any movement of any of its physical parts, will be wholly determined (or their objective probability will be wholly determined) by prior physical events (perhaps those occurring on some cross section of the backward light cone of the physical movements in question). These considerations do not settle whether physicalism is true. A property or attribute dualist need not deny these claims. But considerations of overall coherence and simplicity of total theory are also invoked in defense of physicalism (Smart, 1959; Hill, 1991; Poland, 1994; Loewer, 2001; Melnyk, 2003; McLaughlin, 2010). Whether physicalism can be adequately defended is not a topic of this article. The important point for present purposes is just that physicalism is defended, in part, by appeal to the successes of current physical theory.

The notion of a recognizable descendant of current physics is of course quite vague. But it is not so vague as to be useless. And it can be made more precise. Thus, it can be claimed that completed physics will be a recognizable descendant of current physics only if all ordinary matter, including the matter that makes us up, is wholly composed of atoms, which are more fundamentally physically constituted, and that all of the fundamental forces of nature are at work at a scale below the level of atoms (though, of course, not only below that scale). No stand need be taken on, for example, whether fundamental physical reality will turn out to be particle-like, or instead a matter of vibrating strings, or a matter of branes. Also, it can be allowed, for instance, that the completed physics may take space-time to be nonfundamental yet still be a recognizable descendant of current physics. Even the condition that all fundamental forces are ones that can be found below the level of the atom need not be insisted upon. Suppose that there turn out to be some fundamental chemical forces exerted only at the scale of certain atoms and above, so that there are fundamental chemical forces in addition to force of gravity, the electromagnetic force, the strong force, and the weak force. (The chemical forces invoked in current chemistry, such as van der Waal forces and dipole–dipole forces, are not fundamental forces; they are electromagnetic in origin.) Some physicalists will maintain that that alone would not be a sufficient break from current physics to render physicalism false. There is thus some flexibility in which family resemblances should be focused on, and some resemblances may be given more weight than others. Although there is such flexibility about how the completed physics that is true of our world should resemble current physics, rational debate about physicalism can proceed so long as it is reasonably clear what family resemblances to current physics are supposed to be relevant in the context of the debate.

To fix our ideas, it will often be assumed in what remains both that the physical is whatever would be posited by the completed physics that is in fact true of our world, and that that physics will be a recognizable descendant of current physics. Although this notion of physicality is vague and context dependent, it is sufficiently determinate in content to allow rational debate about the sorts of matters over which debates about physicalism and its alternatives typically center.

Formulating Physicalism and Distinguishing It from Its Alternatives

'Physicalism' is a technical term of metaphysics. Proponents of physicalism must tell us what the doctrine is. As noted above, there is no consensus about how, exactly, the doctrine should be formulated. Since physicalists are persons who hold the doctrine of physicalism, it is thus somewhat uncertain who, exactly, counts as a physicalist, even given the specific conception of the physical embraced above. Here 'physicalist' will be used to mean self-professed physicalist.

It is generally acknowledged that the doctrine of physicalism should be an empirical doctrine, that is, that it should be both contingent and a posteriori. There is a physicalist doctrine called 'a priori physicalism', but it is misleadingly so called,

because it does not entail that physicalism is a priori. It is, rather, roughly, the doctrine that all truths follow a priori from physical truths. Many physicalists deny that the entailments must be a priori, rather than just metaphysically necessary. They are labeled 'a posteriori physicalists'. But both a priori and a posteriori physicalists agree that physicalism itself is a posteriori (Jackson, 2007; McLaughlin, 2007).

Type physicalism is the thesis that every type or kind of property, and every type or kind of entity (object, event, state, process, or field, etc.) is a physical type. The relevant identity claims, it is maintained, will typically be only a posteriori knowable.

Physicalists typically take type physicalism to be too strong a thesis to require for physicalism. Many argue that the special sciences of biology, psychology, economics, anthropology, and sociology all posit kinds of properties and objects that would not be posited by any physics that is a recognizable descendant of current physics. Yet they want to be realists, rather than eliminitivists (Stich, 1999), about such kinds.

Token physicalism is the thesis that every concrete particular is physical. It is implied by type physicalism, since if every type is physical, any token of any type will be physical. But it does not imply type physicalism. Thus, even if, as seems to be the case, the kind *person* is not a physical kind, if every particular person is a complex physical body, then the existence of persons (instances of the kind *person*) is compatible with token physicalism.

Some physicalists maintain that even token physicalism is too strong a thesis to require for physicalism. The weaker claim is sometimes made that every concrete particular is either physical or wholly constituted by physical things at some level of aggregation (Hellman and Thompson, 1975). Thus, it is sometimes maintained, for instance, that although persons have bodies, they are not identical with their bodies, since persons and their bodies have different persistence conditions (Baker, 2000). A person can cease to exist, even when the person's body continues to exist (e.g., at permanent brain death). Indeed, they typically hold as well that human bodies are not identical with collection of physical particles, since bodies and such collections have different persistence conditions: the same body can lose and acquire physical particles with the result that it is constituted by a different collection. They maintain, however, that persons (at a time) are wholly constituted by their bodies (at the time in question), and so do not have immaterial minds, or immaterial souls, or entelechies. Indeed, they typically claim that persons are wholly constituted (at a time) by their microphysical constituents (at that time), and so that the existence of persons should be compatible with physicalism properly formulated.

Some physicalists maintain that it is too strong even to require for physicalism that every concrete particular is either physical or wholly constituted by physical things. They maintain that there are, for instance, Individual Retirement Accounts (IRAs) (and the like), and that they are neither physical nor constituted by physical things, but that their existence should be compatible with physicalism, provided that what IRAs there are and their pattern of distribution are appropriately supervenient on the physical (Haugeland, 1984). It is not required, they claim, that there are in principle identifiable physical bits of our world that minimally constitute IRAs.

It is also widely held that token physicalism is too weak to characterize physicalism, because it is compatible with the existence of nonphysical properties or kinds that fail to be appropriately ontologically dependent on the physical. Token physicalism is, for instance, compatible with the view that mental properties are nonphysical properties that figure in contingent laws of nature that are also fundamental laws, fundamental in the way that the law of gravity is fundamental in Newtonian mechanics, and that Schrödinger's equation is fundamental in non-relativistic quantum mechanics. No recognizable descendant of current physics would posit contingent fundamental psychophysical laws. If there are such laws, then the mental properties cited in them are contingently, ontologically emergent from the physical (Broad, 1925; McLaughlin, 1992). The thesis that mental properties are contingently ontologically emergent from the physical, a thesis that has some currency for certain mental properties (e.g., qualia, the qualitative aspects of experiences), is compatible with the thesis that every concrete particular is or is wholly constituted by physical concrete particulars. But such a property emergence view is nowadays widely held to be incompatible with physicalism, and so to be an alterative to physicalism. Token physicalism seems both too weak and too strong to count as physicalism in the sense that many self-professed physicalists have in mind.

One of the main reasons given in the literature for not requiring for physicalism that all properties are physical properties is that physicalism should allow that there are functional properties (Van Gulick, 2009). A functional property is a second-order property of having some property or other whose instances occupy or play a certain functional role. A functional role will mainly include a causal role (but may also include certain counterfactual dependencies that are not causal), typically a role both as cause and as effect. For a property to have a causal role is for its instances (its exemplifications) to have or to tend to have certain causes and/or effects. The causal role can be a conditional role as cause and effect: instances of the property may have or tend to have certain kinds of causes under certain kinds of conditions, and certain kinds of effects under certain kinds of conditions. A first-order property that plays the role in question is said to 'realize' the functional property; thus the relation of realization in this functionalist sense is that of role occupancy. (There are other notions of realization in the philosophical literature. See, e.g., Shoemaker, 2007.) If more than one first-order property plays the role in question, then the property is multiply realizable. A functional property may be multiply realizable by physical properties, so fail to be identical with any of its physical realizers. Moreover, a functional property may even be such that it is at least metaphysically possible for it to be realized without being realized by any physical property. That is to say, it may be such that there is a metaphysically possible world in which it is realized without being realized by any physical property. Perhaps there is a metaphysically possible world in which certain functional properties are realized in beings wholly composed of ectoplasm. Such functional properties are not physical properties. But it is fairly widely held that physicalism should nevertheless accommodate some such functional properties.

Whether physicalism can accommodate functional properties is taken to be a crucial issue since many philosophers

embrace the role-functionalist view that mental properties (or at least certain kinds of mental properties such as being a belief, desire, intention, and the like) are functional properties, that is, second-order properties of having some property or other that occupies a certain functional role. Proponents of the view that mental properties are functional properties maintain that their view should be compatible with physicalism properly formulated, even if the functional properties in question are not only multiply realized by physical properties in worlds with our laws of nature, but also such that they can be realized without being physically realized in certain worlds without our laws of nature, worlds that are thus nomologically impossible, yet nevertheless metaphysically possible. They maintain that, in part, because functional properties are not fundamental properties, but rather derivative ones. Something will have a functional property in virtue of having a property that realizes it. In that sense, they are always grounded in other properties, namely their realizers.

Realization functionalism is the doctrine that every instantiated property is physical or realized by a physical property (Melnyk, 2003). This captures instantiated functional properties, provided that they are realized by physical properties whenever they are realized. The doctrine is silent, however, about uninstantiated properties. It is compatible with the existence of uninstantiated, contingent, functional properties that would not be realized by physical properties were they instantiated. Indeed, it is compatible with the existence of contingent properties that are neither physical nor functional, provided that they are never instantiated; for it does not rule out that there are fundamental laws of nature that cite nonphysical, nonfunctional properties, so long as those laws of nature are never instantiated. If there are fundamental, yet uninstantiated nonphysical laws of nature, it might be the case that although every object is physical or wholly composed of physical entities and all instantiated properties are physical properties or realized by physical properties, it is an accident that this is so; for there might be a fundamental law of nature according to which, were certain physical conditions to obtain (that never in fact obtain), wholly disembodied spirits would come into existence. That is, the laws of nature might be such that a certain brew of physical ingredients would cause to come into existence disembodied spirits not constituted by physical entities, even though, by chance, the physical ingredients in question have never come together into that brew (Witmer, 2001). Physicalism should require that even uninstantiated properties, if they are contingent existents, be appropriately ontologically dependent on the physical. Still, though, the fact that it could turn out to be the case that whenever something has a functional property, it has it in virtue of having a physical property that realizes it leads physicalists to maintain that a properly formulated physicalism can countenance functional properties.

How must the nonphysical be ontologically dependent on the physical for physicalism to be true? One leading view is that it is a minimal requirement that the nonphysical appropriately supervene on the physical, so that there can be no nonphysical difference of any sort without a physical difference. The kind of supervenience in question is widely taken to be global supervenience, more specifically, strong global supervenience (McLaughlin and Bennett, 2011). Global supervenience theses invoke possible worlds, that is, complete ways the world might be. The following is a leading formulation of such a supervenience definition of physicalism: physicalism is true if and only if any minimal physical duplicate of our world is a duplicate *simpliciter* of our world (Jackson, 1998). A possible world is a duplicate of our world just in case it is exactly like our world in every physical respect, particular and general: in what kinds of physical entities there are, in how instances of them are distributed in space-time, in what physical properties and relations there are, in how they are distributed over particulars, in what physical laws of nature there are, and so on. A possible world is a minimal physical duplicate of our world just in case it includes nothing other than what it must, of metaphysical necessity (the strongest kind of necessity), include to be a physical duplicate of our world. A duplicate *simpliciter* of our world is any possible world that is exactly like our world in every respect whatsoever. (Lewis (1983) and Chalmers (1996) offer closely related, though nonequivalent, definitions of physicalism that also appeal to global supervenience. Most of what is said below about physicalism in the sense just defined will apply to their definitions of physicalism as well.) On this formulation, the existence of functional properties is compatible with physicalism, provided that any minimal physical duplicate of our world would include exactly the functional properties included in our world, and would have exactly the same worldwide pattern of distribution of those functional properties as in our world. Role functionalists about the mental that are physicalists should maintain that that condition is met for the functional properties that are mental properties.

Our world is trivially a physical duplicate of itself. Our world is the actual world, and so a possible world since what is actual is possible (though not conversely). And our world is a duplicate simpliciter of itself, for the trivial reason that everything is a duplicate simpliciter of itself – that is, everything is like itself in every respect whatsoever. It follows that our world is a physical duplicate of itself. It is, however, by no means trivial whether our world is a minimal physical duplicate of itself; for it is by no means trivial whether our world contains more than it must, of metaphysical necessity, contain to be a physical duplicate of itself. It might, for example, include contingent immaterial minds or disembodied spirits. If it does, then since a world that is a minimal physical duplicate of our world would not include such immaterial minds or disembodied spirits, it is not the case that any minimal physical duplicate of our world will be a duplicate simpliciter of our world. If, then, physicalism is true if and only if any minimal physical duplicate of our world is a duplicate simpliciter of our world, it would follow that physicalism is false. Physicalism is also false if our world includes contingent, fundamental psychophysical emergent laws of nature. For given that the laws are contingent and fundamental, a world that is a minimal physical duplicate of our world will not include them, and so will fail to be a duplicate simpliciter of our world. Indeed, all of the leading alternative doctrines to physicalism entail that it is false that any minimal physical duplicate of our world is a duplicate simpliciter of our world, and so are incompatible with physicalism so understood.

The formulation of physicalism in question does not entail that any physical duplicate of our world is a duplicate simpliciter of our world. Physicalism, so formulated, is consistent with its being the case that there is a physical duplicate of our

world that includes, in addition to what it must include to be a physical duplicate of our world, certain nonphysical entities, properties, relations, or laws of nature. It is taken to be a virtue of the characterization of physicalism in question that it is compatible with the possibility of a physical duplicate of our world that include such 'extras', that is, such extra entities, properties, relations, or laws of nature since that seems to be a metaphysical possibility. Our world's being the complete way it is physicality, it is claimed, is, for instance, compatible with there being a realm of disembodied spirits (or spirits composed of ectoplasm). A physical duplicate of our world can include such a realm of spirits. But the thesis that any minimal physical duplicate of our world is a duplicate simpliciter of our world requires that our world not include such a realm.

This formulation of physicalism is fairly widely taken to state a conceptually necessary condition for physicalism. It is, however, also widely taken to fail to state a conceptually sufficient condition for physicalism. For the condition that any minimal physical duplicate of our world is a duplicate simpliciter of our world is conceptually compatible with there being a necessarily existing God (Jackson, 1998). Any doctrine deserving of the name 'physicalism' should be incompatible with the existence of a God, necessary or contingent. (One exception is the version of pantheism according to which God is just the universe.) Indeed, physicalists must deny that it is even metaphysically possible for there to be a necessarily existing God. For, on a widely held assumption about metaphysical necessity, if it is metaphysically possible for a necessarily existing God to exist, then it is metaphysically necessary that it is possible for a necessarily existing God to exist, and so such a God exists in every metaphysically possible world, and thus exists our in our world. The necessary God objection to this version of physicalism does not rely on the assumption that such a God is indeed metaphysically possible. If a necessarily existing God is even conceptually possible (coherently conceivable), then the condition that any minimal physical duplicate of our world is a duplicate simpliciter of our world fails to be a conceptually sufficient condition for physicalism's being true.

There is no consensus in the literature about how to handle this problem. A natural move, which is to the best of the author's knowledge not made in the literature, is to avoid the problem by defining physicalism as the conjunction of the thesis that any minimal physical duplicate of our world is a duplicate simpliciter of our world and an appropriate statement of naturalism, the doctrine characterized earlier. The idea is that the naturalism conjunct would rule out the possibility of a necessary God. It should be noted in support of this move that it is natural to think of physicalism as a kind of naturalism. A proper formulation of naturalism is itself, however, no trivial matter; the one given earlier was only rough and ready. But the idea is that what di